

# Inferring tumor absolute copy number and clonal substructure from single-cell chromatin accessibility

Ying Wang<sup>1,2,†</sup>, Yuhao Deng<sup>3,†</sup>, Hang Li<sup>1</sup>, Xinbao Yin<sup>4</sup>, Yanru Zhang<sup>3</sup>, Yurong Chen<sup>5</sup>, Min Zhang<sup>3</sup>, Xin Wang<sup>3</sup>, Zhizhuo Cao<sup>3</sup>, Shaojun Zhang<sup>1,\*</sup>

<sup>1</sup>Guangdong Academy of Medical Sciences and Medical Research Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, No. 106 Zhongshan 2nd Road, Yuexiu District, Guangzhou 510080, China

<sup>2</sup>Center for Endemic Disease Control, Chinese Center for Disease Control and Prevention, Harbin Medical University, No. 157 Baojian Road, Nangang District, Harbin 150081, China

<sup>3</sup>Institute of Precision Medicine, The First Affiliated Hospital, Sun Yat-Sen University, No. 58 Zhongshan Er Road, Yuexiu District, Guangzhou 510080, China

<sup>4</sup>Department of Urology, The Affiliated Hospital of Qingdao University, No.16 Jiangsu Road, Shinan District, Qingdao 266000, Shandong, China

<sup>5</sup>Department of Gynecological Radiotherapy, Harbin Medical University Cancer Hospital, No. 150 Haping Road, Nangang District, Harbin 150081, China

\*Corresponding author. Guangdong Academy of Medical Sciences and Medical Research Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, No. 106 Zhongshan 2nd Road, Yuexiu District, Guangzhou 510080, P.R. China. E-mail: szhang5@outlook.com

<sup>†</sup>Ying Wang and Yuhao Deng contributed equally to this work.

## Abstract

Accurate inference of absolute copy numbers beyond simple gains and losses from single-cell chromatin accessibility (scATAC-seq) data remains challenging, thereby obscuring the distinction between genetic and epigenetically driven oncogenic dependencies. Here, we present TeaCNV, a computational framework that reconstructs clonal absolute copy number profiles and tumor clonal architectures from scATAC-seq data without matched DNA baselines. Through validation both *in silico* and against bulk whole-genome sequencing in renal cell carcinomas, TeaCNV resolved subclonal absolute copy number profiles with less than 10% error and detected copy number variations (CNVs) with 98.6% accuracy, outperforming existing methods. Applied to six cancer types including renal, breast, pancreatic, head and neck, colorectal, and ovarian cancers, TeaCNV delineated polyclonal architectures and revealed distinct chromatin accessibility patterns driven by CNVs in key driver genes, including *AKT2*, *ZNF217*, and *SOX2*. By enabling absolute copy number profiling and clonal deconvolution from epigenomic assays, TeaCNV bridges critical gaps in studying oncogenic dependencies and genotype–phenotype relationships at single-cell resolution.

**Keywords** single-cell, copy number, clonal architectures, tumor heterogeneity, scATAC-seq

## Introduction

Chromosomal copy number variations (CNVs) are fundamental genomic alterations that drive tumor evolution by reshaping transcriptional landscapes [1, 2]. Aneuploidy, defined as the presence of abnormal chromosome numbers, is a hallmark of human cancer and is strongly associated with tumor aggressiveness and poor clinical outcomes [3–6]. Therefore, precise quantification of chromosomal absolute copy numbers (CNs) is crucial for understanding cancer evolution and informing clinical decision-making [7, 8].

Bulk and single-cell whole-genome sequencing (scWGS) remain the gold standards for CNV profiling [9–11], but both techniques have limitations. Bulk WGS enables quantification of tumor ploidy, yet offers limited resolution for clonal substructure deconvolution, particularly in detecting rare clones [8, 12]. Although scWGS improves clonal substructure resolution by providing absolute CN quantification at the single cell level, it cannot characterize the functional consequences of heterogeneous CNVs due to the absence of gene expression data [10,

13]. Moreover, both bulk WGS and scWGS fail to capture the dynamic interactions within the tumor microenvironment [14, 15], and face significant challenges in clinical implementation, including technical complexity, high cost, and limited scalability [16].

Single-cell RNA sequencing (scRNA-seq) has emerged as a powerful alternative, enabling integrated analysis of both microenvironment composition and tumor subpopulation architecture [17, 18]. Despite challenges such as transcriptional noise and posttranscriptional regulation [19], computational tools including inferCNV [20–24] and HoneyBADGER [25] have demonstrated the feasibility of detecting CNVs from scRNA-seq data at large genomic scales. Recent methodological advances, such as CopyKAT's clonal substructure analysis [26] and allele-specific CNV detection by Numbat [27] and XClone [28], further enhance the ability to dissect tumor heterogeneity and link CNVs to their transcriptional consequences. However, scRNA-seq's reliance on transcriptional readouts limits its ability to directly dissect the epigenomic contributions to tumor cell diversity.

**Received:** November 20, 2025. **Revised:** April 12, 2026. **Accepted:** April 29, 2026

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Single cell ATAC sequencing (scATAC-seq), which profiles chromatin accessibility through transposase-accessible DNA sequencing, provides an opportunity to study malignant regulatory programs at single cell level [29, 30]. Characterizing epigenomic regulatory mechanisms in cancer requires careful consideration of CNV effects, as chromosomal alterations can substantially alter chromatin accessibility patterns in affected regions [29]. scATAC-seq offers unique advantages for CNV detection by directly measuring DNA accessibility rather than relying on indirect transcriptional outputs, thereby reducing susceptibility to gene regulatory confounders [31, 32]. Recently, computational methods like epiAneufinder [33] and Copy-scAT [34] have been developed to identify CNVs from scATAC-seq data. However, due to the extreme sparsity of scATAC-seq data which exceeds that of scRNA-seq data, and systematic biases at chromatin regulatory elements (CREs) such as promoters and enhancers, these methods are limited to identifying relative gain/loss at large genomic scales rather than absolute CN quantification [33–35]. This limitation hinders the understanding of clonal substructures and the ability to distinguish between the effects of genomic and epigenomic changes in polyploid tumors.

To address these limitations, we developed TeaCNV, a computational framework that quantifies absolute CN profiles and resolves clonal substructures directly from scATAC-seq data. Validated on both simulated datasets and bulk WGS, TeaCNV achieved 98.6% accuracy of CNVs calling and less than 10% error in absolute CN profiling, while detecting subclonal CNVs undetectable by bulk WGS. Application to multi-cancer types (renal, breast, pancreatic, head and neck, ovarian, and colorectal cancers) revealed lineage-defining amplifications of oncogenic drivers and decoupled genomic from epigenomic contributions within polyclonal architectures.

## Methods

### Human specimens

Four ccRCC patients were collected with informed consent approved by the Institutional Ethics Committee of the Affiliated Hospital of Qingdao University, including two patients for scATAC-seq, two for scATAC&RNA-seq co-assays, and three patients for bulk WGS.

### Experimental methods

#### Nuclei isolation, library preparation, and sequencing

Tissues were dissected, snap-frozen in liquid nitrogen and stored at  $-80^{\circ}\text{C}$ . Nuclei were isolated from frozen tissues following the manufacturer's protocols for scATAC-seq (10× Genomics, CG000212 Rev B) and scATAC&RNA-seq co-assays (10× Genomics, CG000375 Rev C), with RNase inhibitors in the buffer to prevent mRNA degradation during cell lysis.

Nuclei and barcoded beads were then loaded onto the 10× Genomics platform using the Chromium Next GEM Single Cell ATAC Library Kit v2 for scATAC-seq (CG000496 Rev A) and the Single Cell Multiome ATAC + Gene Expression kit for scATAC&RNA-seq co-assays (CG000338 Rev F), according to the manufacturer's instructions. Barcoded libraries were pooled and sequenced with paired-end reads on Illumina NovaSeq 6000.

#### Library preparation and sequencing for bulk WGS

High-quality genomic DNA ( $\sim 0.6\ \mu\text{g}$ ) was sheared with Covaris LE220 Sonicator (Covaris) to  $\sim 350\ \text{bp}$ . The library was constructed using protocol of KAPA Hyper Prep kit (Roche). Fragmented DNA was purified

using sample purification beads, and the product was repaired by the end and A base was added to the 3' end. Then, the adapters are ligated with the specific barcode sequence. The CleanNGS magnetic beads (CleanNA) were used to screen out incomplete connections and self-connecting products. Sequencing libraries were formed by PCR amplification using universal primers complementary to the adaptor sequences. Paired-end sequencing was performed using the NovaSeq 6000 S4 Reagent Kit v1.5 (300 cycles) on Illumina NovaSeq 6000 platform (Illumina, San Diego, USA) by Sequanta Technologies (Shanghai, China).

#### scATAC-seq and scATAC&RNA-seq data processing

Reads from scATAC-seq and scATAC&RNA-seq were aligned to GRCh38 (hg38) and quantified using *cellranger-atac count* (v.1.2.0) and *cellranger-arc count* (v.2.0) pipelines (10× Genomics), respectively. Peaks were called using MACS3 (v.3.0.0) [36] through the *CallPeaks* function in Signac (v.1.14.0) [37]. Peaks on chromosomes X/Y and in the ENCODE Unified GRCh38 Blacklist were removed using “blacklist\_hg38\_unified” in the *subsetByOverlaps* function in Signac. The filtered, sample-specific peaks were used to construct peak-by-count matrices using *FeatureMatrix* in Signac for downstream analyses.

Quality control was performed using Signac. Cells were filtered based on the following criteria: number of fragments in peaks  $> 1000$ , number of peaks in cell  $> 2000$ , and enrichment-score for Tn5-integration events at transcriptional start sites  $> 3$ .

#### Normalization, dimensionality reduction, clustering, and cell classification

The filtered peak-by-count matrices derived from scATAC-seq or scATAC&RNA-seq data were normalized using term frequency-inverse document frequency (TF-IDF) as implemented in Signac. The top 95% of peaks were selected as features for dimensionality reduction. Latent semantic indexing (LSI) was then performed using the RunSVD function, and LSI components 2–30 were used as input for UMAP embedding (RunUMAP, Seurat). Nuclei were subsequently clustered using a graph-based clustering approach in Seurat, also based on LSI components 2–30.

For scATAC-seq data, cell types were annotated based on gene activity scores of canonical lineage markers: immune cells were identified by high activity of *PTPRC*, endothelial cells by *PECAM1* and *CD34*, fibroblast cells by collagen genes (e.g. *COL1A1*, *COL1A2*, *COL3A1*), and epithelial cells by *EPCAM* and *KRT* family genes. For scATAC&RNA-seq data, gene-count matrices were processed in Seurat (v.4.4.0) [38] for scaling, normalization, and identification of highly variable genes using default parameters. The number of significant principal components (PCs) was determined from the ElbowPlot, and the top 2000 highly variable genes and the first 30 PCs were used for unsupervised clustering. Cell types were then assigned based on the expression of canonical lineage marker genes.

#### Bulk WGS data processing and CNV analysis

Sequencing reads from bulk tumor tissue and matched normal tissues were aligned to GRCh38 using Burrows–Wheeler Aligner (BWA v0.7.17) software [39], and resulting BAM files were sorted and indexed with SAMtools [40]. The “runVarbin” module of CopyKit (v0.1.2) [41] was then used to count reads in 220-kb genomic bins defined on the GRCh38 assembly, with GC correction applied. Consistent with our scATAC-seq preprocessing, we excluded bins on

chromosomes X/Y, as well as ENCODE blacklist regions to avoid sex-specific and artifactual signals. The log ratio of tumor to normal was computed for each bin and used for CNV calling, segmentation, and the inference of absolute CNs. Briefly, we fitted a Gaussian kernel smoothed density to the distribution of segment ratios to identify the diploid peak centered at ratio = 1. The one-copy increment was defined as the distance between the diploid peak and the nearest adjacent peak. Absolute CNs were then obtained by scaling each segment ratio by this increment and rounding to the nearest integer.

## TeaCNV algorithm

### Preprocessing and transformation of scATAC-seq data

To balance sparsity and genomic coverage, we removed peaks detected in <5% of cells. If <10 000 peaks remained, we progressively relaxed this threshold until at least 10 000 peaks were retained. In solid tumors, non-epithelial cells (immune, endothelial, and fibroblast populations) were treated as non-malignant reference cells, while epithelial cells were treated as candidate malignant cells and used as inferred cells for CNV detection. For each group (reference and inferred), we retained cells whose total read counts and numbers of detected peaks fell between the 5th and 95th percentiles. Cells were further excluded if >60% of peaks on any chromosome had zero counts to reduce bias from chromosomal dropout.

We constructed two peak-by-cell matrices:  $X_{n \times m}$  for reference cells and  $Y_{n \times t}$  for inferred cells, where  $n$  denotes the number of peaks ( $n \geq 10\,000$ ),  $m$  represents the number of reference cells, and  $t$  is the number of inferred cells. Values in both matrices were truncated to the range [0,4]. To correct for sequencing-depth variation, we normalized each cell by its total counts and rescaled to the mean library size within each matrix, yielding  $X'$  and  $Y'$ . For each peak  $i$ , we calculated the reference mean  $\bar{X}'_{i\cdot}$ , and defined the ratio matrix as follows:

$$R_{ij} = \frac{Y'_{ij}}{\bar{X}'_{i\cdot}} \quad (1 \leq i \leq n, 1 \leq j \leq t),$$

which encoded relative accessibility in inferred cells compared with reference cells.

### Aggregating cell subpopulations

To capture broad genomic patterns, we sorted peaks by genomic coordinate and, on each chromosome, averaged ratios within windows of five consecutive peaks (user-adjustable), obtaining a genomic-window matrix. Using Seurat, we identified top 2000 highly variable features, performed PCA and applied unsupervised clustering (resolution = 1) on the first 50 PCs yielded an initial set of inferred cell subpopulations (subgroups). For each subgroup  $s$ , we computed subgroup-level ratio for peak  $i$ :

$$\bar{R}_{i\cdot}^s = \text{average}[R_{i,k \in S}]$$

### Genome segmentation of cell subpopulations

For each subgroup  $s$ , we ordered  $R^s$  by genomic coordinate and performed changepoint detection to identify chromosomal segments with homogeneous ratios. We primarily used the pruned exact linear time (PELT) algorithm in *changepoint* R package [42], with FPOP algorithm in *robseg* R package [43] as an alternative. Segments shorter than 2 Mb

were excluded. For each retained segment  $j$ , the segmental ratio was defined as follows:

$$R_j^s = \text{median}\left(\bar{R}_{j \in \text{segment}_j, \cdot}^s\right),$$

which served as the relative copy ratio for subgroup  $s$  at segment  $j$ .

### Inferring absolute copy number

For each subgroup, we obtained segmental ratios  $R = \{R_j | j \in \{1, \dots, l\}\}$  over  $l$  segments and modeled  $R_j$  by a mixture distribution of integer CN states  $Q = \{1, 2, \dots, I\}$ :

$$P(R_j | \mu, \sigma^2, \theta, w_j) = \sum_{q \in Q} P(q | \theta_q, w_j) \cdot N(R_j | \mu_q, \sigma^2)$$

where  $\mu_q$  is the expected relative copy ratio for state  $q$ ,  $\sigma^2$  is a shared variance,  $w_j$  is the genomic fraction of segment  $j$ , and  $P(q | \theta_q, w_j)$  is the prior probability parameterized by  $\theta_q$  ( $q \in Q$ ) and  $w_j$ . Following the maximum entropy principle, as the statistical hypothesis employed in ABSOLUTE [8], we set:

$$P(q | \theta_q, w_j) = \frac{e^{-\theta_q \cdot q^{\#} \cdot w_j}}{\sum_{k \in Q} e^{-\theta_k \cdot k^{\#} \cdot w_j}}$$

where  $q^{\#}$  and  $k^{\#}$  respectively indicate the order of integer CN states  $q$  and  $k$  in  $Q$ , starting from 1 (e.g.  $q = 1$  corresponds to  $q^{\#} = 1$ ). The expected genomic fraction of state  $q$  across the genome is  $\sum_{j=1}^l w_j \cdot P(q_j | \theta_q, w_j)$ .

The unknown parameters  $\theta = \{\theta_q \in Q\}$  were estimated by minimizing the following loss function using the Nelder-Mead optimization algorithm:

$$\hat{\theta} = \arg \min_{\theta} \left[ \sum_{q \in Q} \left( \sum_{j=1}^l w_j \cdot P(q | \theta_q, w_j) - \theta_q \right)^2 \right]^{\frac{1}{2}}$$

With  $\hat{\theta}$ , we computed the prior probabilities  $\hat{P}(q | \hat{\theta}_q, w_j)$ .

Next, we estimated the expected relative copy ratio  $\mu = \{\mu_q | q \in Q\}$  using maximum likelihood estimation. The full likelihood of relative copy ratio  $R$  across the genome is:

$$P(R | \mu, \sigma, \hat{\theta}, w) = \prod_{j=1}^l P(R_j | \mu, \sigma^2, \hat{\theta}, w_j)$$

and the corresponding log-likelihood is:

$$\log L(R | \mu, \sigma, \hat{\theta}, w) = \sum_{j=1}^l \log L(R_j | \mu, \sigma^2, \hat{\theta}, w_j)$$

where

$$L(R_j | \mu, \sigma^2, \hat{\theta}, w_j) = \sum_{q \in Q} P(q | \hat{\theta}_q, w_j) \cdot N(R_j | \mu_q, \sigma^2)$$

For segment  $j$ , the posterior probability of integer CN state  $q \in Q$  is computed as:

$$P(CN = q | \text{segment}_j) = p(q | \hat{\theta}_q, w_j) \cdot \frac{N(R_j | \hat{\mu}_q, \sigma^2)}{\sum_{k \in Q} P(k | \hat{\theta}_k, w_j) \cdot N(R_j | \hat{\mu}_k, \sigma^2)}$$

The final absolute CN state for segment  $j$  is defined as the CN state  $q$  with the highest posterior probability.

To optimize absolute CN estimation, we computed differences between adjacent integer CN states:

$$\Delta_q = \mu_q - \mu_{q-1} \quad (q = 2, \dots, I),$$

which is theoretically constant and  $<1$  (e.g.  $\Delta_q \approx 0.5$  corresponds to a change of 1-copy relative to diploid). However, biases in  $\mu_q$  estimation (e.g. due to uneven sequencing coverage or technique noise) may result in non-constant  $\Delta_q$  values across  $q$  (Supplementary Fig. 17a).

To address this, we constructed a small set of candidate spacings  $\{\Delta_i\}$ , enforced a linear relationship for the expected copy ratio:

$$\mu^{(i)} = \left\{ \mu_q^{(i)} = \mu_1 + (q - 1) \cdot \Delta_i | q \in Q \right\}$$

The final absolute CN configuration for each subgroup was chosen as the candidate  $\mu^{(i)}$  with the minimum Akaike information criterion (AIC) (Fig. 1c). Further details are provided in Supplementary Methods.

### Tumor subclone partitioning and scoring

To avoid potential over-clustering, we constructed an adjacency matrix  $A$  over subgroups. We initialized  $A_{ij} = 1$  if  $i = j$  and  $A_{ij} = 0$  otherwise. For each subgroup  $s_i$ , we computed its mean normalized peak profile  $Y'_{s_i}$ ; for each pair  $(s_i, s_j)$ , we formed the peak signal ratio:

$$R_{s_i, s_j} = \frac{Y'_{s_i}}{Y'_{s_j}}$$

We segmented this profile and compared peak signals in each segment using two-sided t-tests with Benjamini–Hochberg adjustment. If any segment had adjusted  $P$ -value  $< 0.05$  and differed in inferred CN between  $s_i$  and  $s_j$ , we set  $A_{ij} = A_{ji} = 0$ . Otherwise, we updated  $A_{ij} = A_{ji} = 1$ . The final subclonal partitioning was derived from the updated adjacency matrix. For each subclone, we re-estimated integer CN states and computed ploidy as a length-weighted average:

$$\text{ploidy} = \sum_{j=1}^l w_j \cdot \hat{q}_j$$

We evaluated the reliability of integer CN profiles by assuming that, in homogeneous subclones, observed copy ratios follow expected integer-state ratios (Supplementary Fig. 17b) and summarizing their mean squared error (MSE) from the expected values together with the genome fraction well explained by the model into a single composite score. This score additionally penalizes unrealistically small single-copy spacing, excessive segmentation and extreme ploidy (details in Supplementary Methods), such that higher score indicate more homogeneous and reliable subclone CN profiles.

## Application of other CNV inference approaches

For comparison, we applied epiAneufinder (v1.1.3) [33] and Copy-scAT (v0.4.0) [34] to scATAC-seq data from three ccRCC samples with matched bulk WGS, using default parameters at single-cell resolution. In addition, for two ccRCC samples with scATAC&RNA-seq, we employed inferCNV (v1.22.0) [20], Numbat (v1.3.2.1) [27], XClone (v0.3.8) [28], CopyKAT (v1.1.0) [26], and sciCNV [44] on the scRNA-seq data, using the same reference cells as in TeaCNV. inferCNV was run with recommended  $10\times$  settings (denoise=TRUE, cutoff=0.1) and we additionally performed “consensus” i6 HMM mode (inferCNV-HMM) to infer integer CN profiles. For Numbat, SNP pileup data were generated following the official documentation, and Numbat was then run with default parameters on allele counts and gene expressions. XClone was run in default mode, combining its RDR and BAF modules to derive CNV profile estimates.

## Evaluating performance of copy number estimation

We evaluated CNV detection using all CNVs derived from bulk WGS as ground truth, regardless of event size. The genome was partitioned into non-overlapping 100-kb bins, and bins overlapping breakpoints in either the ground truth or inferred profiles were excluded. For each method, bins were classified as true positives (TP), false positives (FP), false negatives (FN), or true negatives (TN) according to the presence or absence of CNVs in the ground truth and inferred results. Precision, recall, accuracy, and F1 score were then computed using their standard definitions (see Supplementary Methods).

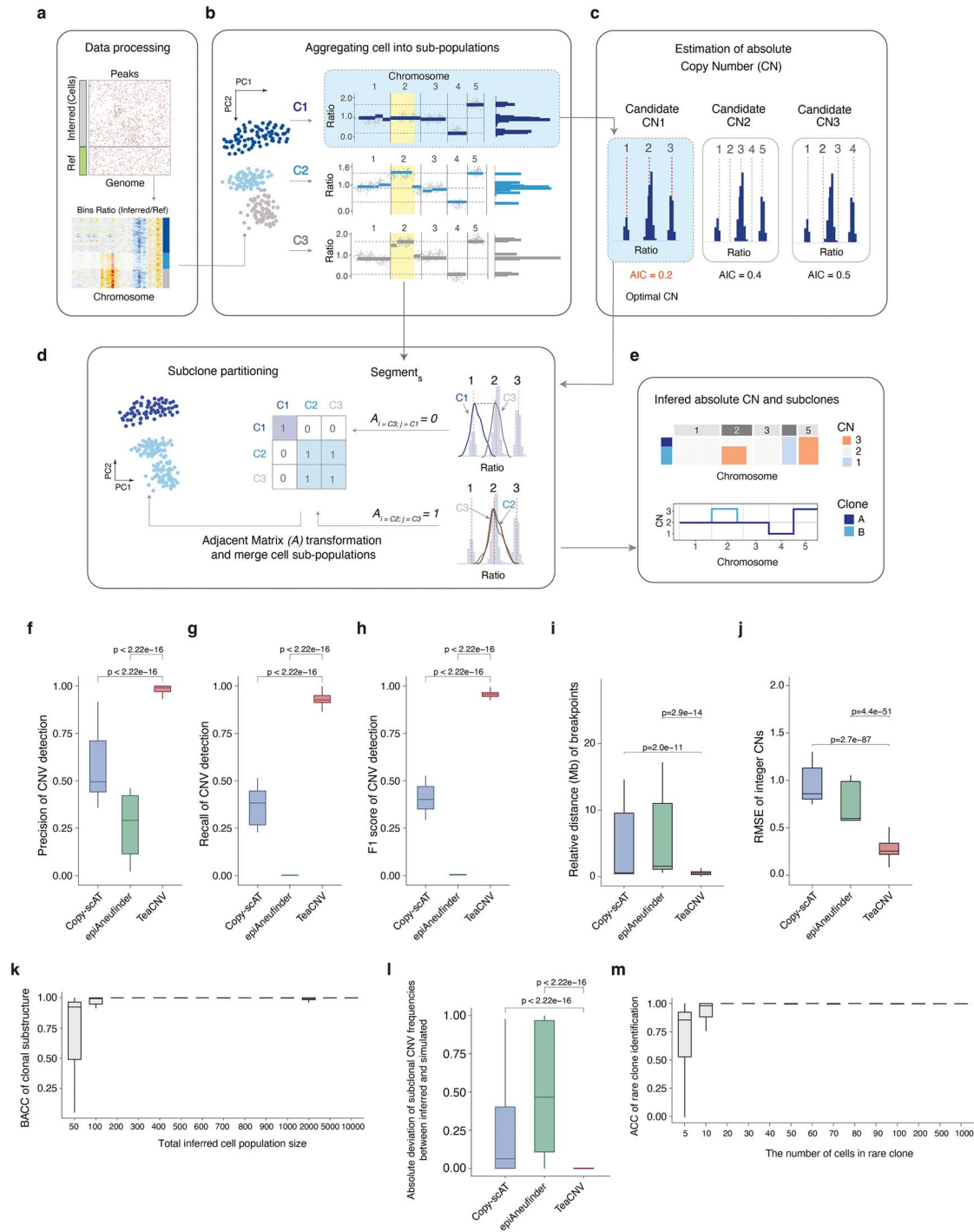
To assess reconstruction of polyclonal substructure, we calculated the balanced accuracy (BACC) [45] for each subclone based on TP, FP, TN, and FN, and summarized the overall performance as weighted average across subclones. Because TeaCNV reports CNVs at clonal level, precision, recall, and F1 score were averaged across the subclones within each sample. For epiAneufinder and Copy-scAT, CNV events were considered detected at the bulk level if present in at least 10%–100% of cells (10% steps), and evaluation metrics were averaged across thresholds.

We further quantified the accuracy of inferred CN profiles using the root mean square error (RMSE) between inferred and WGS-based integer CNs at the bin level. TeaCNV directly provided integer CNs, whereas epiAneufinder and Copy-scAT do not report integer CNs directly, we centered the inferred CNV scores to align with integer CNs for bins with distinct absolute CN states in WGS data. The definition of  $CN_i$  was as follow:

$$CN_i = \begin{cases} CNV \text{ score}_i - \text{average}_{j \in \{j | CN_j^{true} = CN_i^{true}\}} \left( CNV \text{ score}_j \right) & \text{for TeaCNV} \\ CN_i^{true} & \text{for other methods} \end{cases}$$

This adjustment allows for a comparison of inferred profiles against the ground truth.

Finally, to quantify how well distinct integer CN states can be distinguished, we proposed a dispersion score to measure the separation between inferred signal distributions corresponding to different integer CN states, normalized to  $[0, 1]$  (Supplementary Fig. 17c). Higher dispersion score indicates a better ability to distinguish distinct absolute CN states, reflecting a more accurate estimation of CN states (details in Supplementary Methods).



**Figure 1** Overview of the TeaCNV algorithm and performance evaluation. (a) Workflow for preprocessing scATAC-seq data. TeaCNV converts the cell-peak matrix into copy ratio values relative to reference cells (Ref) and merges the ratios of adjacent peaks. (b) Cell clustering and genome-wide segmentation. For each cluster, the mean copy ratio across cells is calculated and subjected to PELT segmentation. The average copy ratio of bins within each segment is used to generate a histogram of genomic segments, in which bar length represents the genomic fraction (x-axis) of each segment. (c) Estimation of absolute CN. Three candidate interpretations of the copy ratio histogram (corresponding to cluster C1 in b) are shown in terms of absolute CNs. Dotted lines indicate copy ratios corresponding to specific absolute CNs. The optimal absolute CN estimation is determined according to the minimum AIC value. (d) Optimization of subclone partitioning. Clusters are merged when no significant differences in either absolute or relative CNs are detected across all genomic segments. (e) Representative TeaCNV output displaying subclonal absolute CN profiles. (f–h) Performance of CNV detection quantified by precision (f), recall (g), and F1 score (h) on simulated datasets. (i) Relative distance between inferred chromosomal breakpoints and the ground truth. (j) RMSE between inferred integer CNs and the ground truth. (k) BACC of subclone identification across datasets with varying population sizes. (l) Absolute deviation between inferred and simulated subclonal CNV frequencies. (m) Classification accuracy (ACC) for rare clone detection in simulated datasets.

## Chromatin accessibility and downstream analysis

We analyzed differential chromatin accessibility in PDAC, BRCA, HNSCC, OV, and CRC samples and classified subclone-specific peaks as CNV-driven or CRE-regulated based on overlap with subclonal CNVs. Genes linked to clonal peaks were subjected to hallmark-based gene set from MSigDB [46] enrichment analysis to characterize subclone-specific pathways. Details are provided in Supplementary Methods.

## Results

### Methodology to infer clonal absolute copy number profile from scATAC-seq

We developed TeaCNV, a quantitative framework for reconstructing clonal absolute CN profiles and clonal substructure from scATAC-seq data without requiring matched DNA sequencing references. The algorithm processes an input peak-by-cell count matrix with cells pre-annotated as reference (normal) or inferred (tumor or epithelial) cells. Peaks are sorted by chromosomal coordinates. The outputs include clonal partitions of single cells and corresponding subclone-level absolute CN states for chromosomal segments (Fig. 1).

The inference pipeline operates through four steps (Methods). First is initial cell clustering. For each inferred cell, relative copy ratios are computed by normalizing peak accessibility signals against the median accessibility in matched reference cells. Dimension reduction via principal component analysis (PCA) on these ratios at the genomic window level enables initial clustering of inferred cells into subpopulations (Fig. 1a). Second is chromosomal breakpoint detection. Cells within each subpopulation are aggregated. Multi-scale chromosome breakpoints are identified using the PELT algorithm [47]. This generates segmentation boundaries for each subpopulation while accommodating scATAC-seq data sparsity (Fig. 1b). Third is joint ploidy-state optimization. For each subpopulation, absolute CNs of genomic segments are inferred through a combined optimization, including Nelder–Mead optimization to model expected relative copy ratios and CN state relationships, maximum likelihood estimation weighted by segment length and ratio variance, optimal integer CN determination via minimum AIC (Fig. 1c). Fourth is clonal architecture refinement. Subpopulations undergo pairwise comparison of their absolute and relative CN profiles. Subpopulations with concordant profiles are merged, followed by breakpoint recalculation and integer CN estimation (Fig. 1d). Last is the consensus profile generation, including clonal absolute CN profiles across chromosomal segments and single-cell clonal assignments (Fig. 1e).

### *In silico* evaluation

To evaluate the performance of TeaCNV, we simulated scATAC-seq data with well-defined ground truth using the simATAC [48] tool (Supplementary Methods). We created a series of simulated datasets with varying inferred cell population sizes (50 to 10 000 cells), clonal substructure complexity (1–5 subclones with prevalence from 10% to 90%), and CNV genomic scales (10% to 60%). Across all conditions and replicates, this yielded 680 simulated datasets (Supplementary Table S1). We randomly selected the subclonal CNV events, and a subclone with 10% frequency was designed as a rare clone. Besides TeaCNV, we also inferred CNV profiles using Copy-scAT [34] and epiAneufinder [33] (Supplementary Methods). The computational efficiency of the TeaCNV algorithm scales linearly with the number of cells (Supplementary Fig. 1a and b).

We compared CNV detection performance using precision, recall, and F1 score (Supplementary Methods; Supplementary Table S2). TeaCNV achieved significantly better CNV detection performance than the other methods (Fig. 1f–h), and its performance remained robust across CNVs of different genomic scales (Supplementary Fig. 1c–e). Additionally, TeaCNV accurately detected CNV boundaries, exhibiting the lowest deviation (mean: 0.57 Mb) and variance, although breakpoint deviation increased with the genomic fraction of CNVs (Fig. 1i and Supplementary Fig. 1f), with detected minimum CNV lengths spanning 2.04 to 4.20 Mb (Supplementary Fig. 1g). Together, TeaCNV achieved an average sensitivity of 0.96 for broad CNVs, including arm- and subarm-level events (Supplementary Fig. 2a). For focal CNVs, sensitivity declined as focal CNV size decreased, with an average sensitivity of 0.47 overall (Supplementary Fig. 2b).

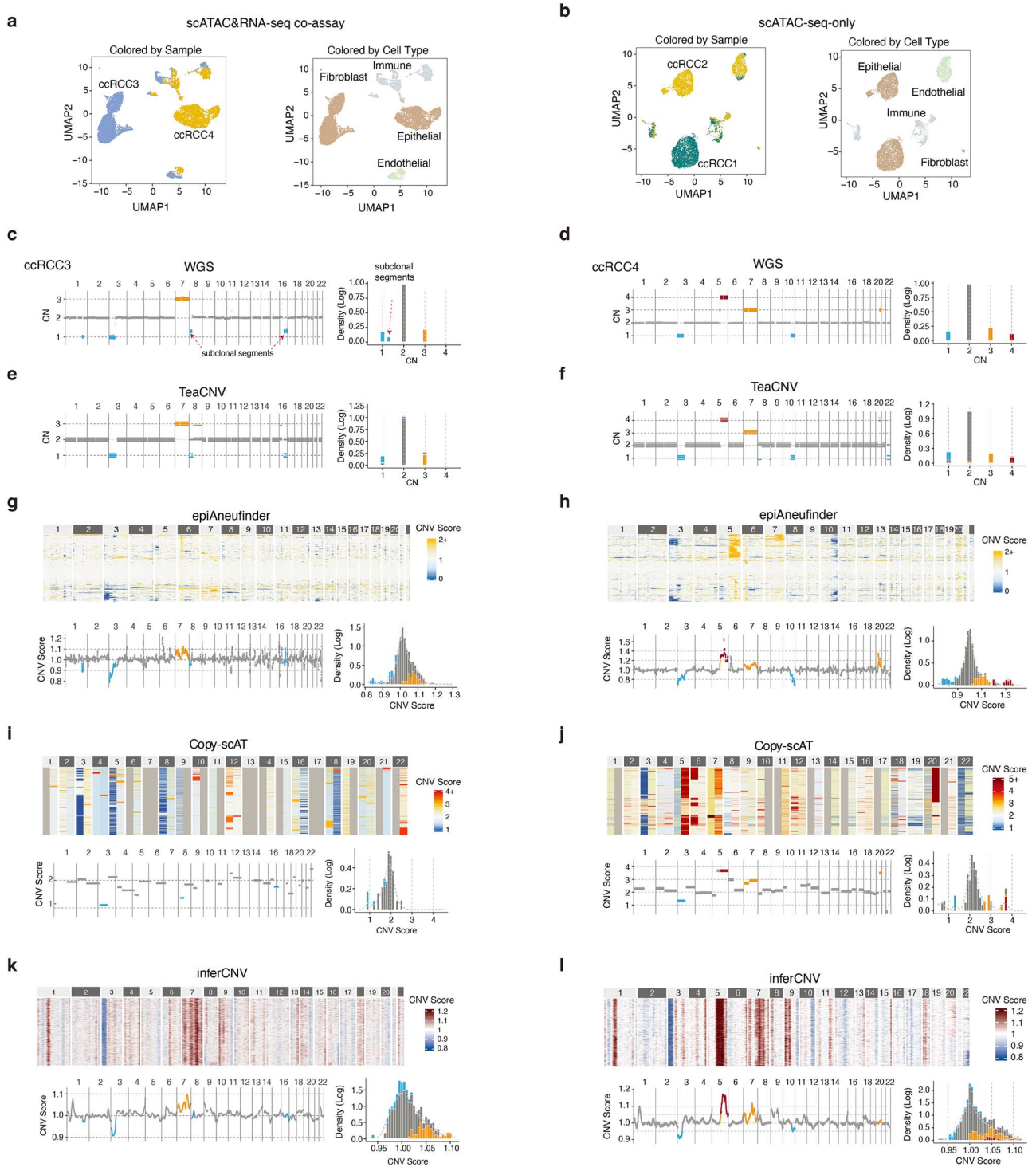
We further calculated the RMSE of inferred integer CNs relative to the ground truth to measure the error of absolute CN quantification. TeaCNV showed the lowest error in integer CN quantification, with the maximum error not exceeding 0.5 (Fig. 1j and Supplementary Fig. 1h). We then evaluated the clonal substructure deconvolution performance across different population sizes using BACC across intra-tumor subclones (Supplementary Methods). TeaCNV accurately and robustly reconstructed clonal substructure with  $\geq 200$  cells (BACC > 0.95, Fig. 1k). Moreover, the prevalence of subclonal CNVs inferred by TeaCNV was closer to the theoretical values (Fig. 1l and Supplementary Fig. 1i). Finally, we evaluated the sensitivity of rare clone detection by TeaCNV and found that at least 20 cells per subclone are required for robust rare clone detection (Fig. 1m and Supplementary Fig. 1j–l).

### Performance evaluation of TeaCNV using matched bulk WGS data in clear cell renal cell carcinoma

To evaluate TeaCNV's ability to reconstruct clonal architecture of CNVs, we performed high-throughput 10× Genomics sequencing on four clear cell renal cell carcinoma (ccRCC) patients. This generated two scATAC&RNA-seq co-assay datasets and two scATAC-seq-only datasets (Fig. 2a and b, Supplementary Fig. 3). Cell type annotation was based on chromatin accessibility signatures of canonical markers (*PTPRC*, *PECAM1*, *CD34*, *COL1A2*, and *EPCAM*), followed by UMAP visualization, which identified 8367 epithelial cells for CNV analysis (Supplementary Figs 4 and 5).

Using non-epithelial cells (immune, endothelial, and fibroblast cells) as references, we applied TeaCNV to infer genome-wide clonal absolute CNs. To validate the performance of TeaCNV, we analyzed three ccRCC patients (ccRCC3, ccRCC4, and ccRCC1) with matched bulk WGS data. For the two scATAC&RNA-seq co-assay samples, the absolute CN profiles from bulk WGS data were considered the ground truth (Fig. 2c and d). The clonal integer CN profiles estimated by TeaCNV showed high concordance with WGS profiles (Fig. 2e and f). In ccRCC3, TeaCNV accurately identified the trunk events (chr3p loss and chr7 gain) and subclonal losses (chr8p and chr16q), which showed absolute CNs between 1 and 2 in bulk WGS data (Fig. 2c and e). Similarly, in ccRCC4, TeaCNV recalled all driver CNVs observed in bulk WGS data, including losses of chr3p/chr10q and gains of chr5q/chr7/chr22q (Fig. 2d and f).

For the scATAC&RNA-seq co-assay samples, we applied epiAneufinder and Copy-scAT to chromatin accessibility data, and inferCNV [20], Numbat [27], XClone [28], CopyKAT [26], and sciCNV [44] to the gene expression data (Fig. 2g–l, Supplementary Fig. 6). The



**Figure 2** Comparison of the estimated copy number profiles with bulk WGS data. (a, b) UMAP plots of cells from ccRCC patients with (a) scATAC&RNA-seq co-assay data and (b) scATAC-seq-only data, colored by sample and cell type. (c, d) Integer CN profiles from bulk WGS data (left) and corresponding genome-wide distributions of absolute CN states (right) for ccRCC3 (c) and ccRCC4 (d). Significantly altered regions are highlighted in blue (loss,  $CN < 2$ ) and orange/red (gain,  $CN > 2$ ). (e, f) Clonal integer CN profiles estimated by TeaCNV (left) and genome-wide distributions of estimated integer CN states (right) for ccRCC3 (e) and ccRCC4 (f). In the left panel of each subfigure, bars of each genomic segment correspond to different clones. CNV segments are highlighted in blue (loss,  $CN < 2$ ) and orange/red (gain,  $CN > 2$ ). (g, h) Heatmaps of relative CN profiles (top) and genome distributions of relative CNV scores (bottom) estimated by epiAneufinder for ccRCC3 (g) and ccRCC4 (h). In the genome distribution plots, blue and orange/red highlights indicate CNV regions identified by bulk WGS data. (i, j) Copy-scAT estimates for ccRCC3 (i) and ccRCC4 (j). (k, l) inferCNV estimates for ccRCC3 (k) and ccRCC4 (l).

scATAC-based approaches either diluted signals from aberrant regions or misidentified neutral regions, leading to ambiguous distributions of inferred CNV scores for segments with distinct CN states (Fig. 2g–j). The scRNA-based approaches exhibited differential biases: inferCNV recalled driver CNVs but also produced false positive signals in focal regions (Fig. 2k and l, Supplementary Fig. 6a), as did CopyKAT and sciCNV (Supplementary Fig. 6d and e), while Numbat and XClone missed positive signals due to their dependence on limited allelic information (Supplementary Fig. 6b and c).

In the scATAC-seq-only sample (ccRCC1), TeaCNV also demonstrated better performance in distinguishing chromosomal segments with distinct CN states compared with other scATAC-based approaches (Fig. 3a–d). We quantified the distinguishability of distinct integer CN states derived from bulk WGS data using dispersion score (Supplementary Methods). A higher score indicates less overlap among the inferred signal distributions of different integer CN states. TeaCNV achieved a higher average dispersion score (0.96 versus 0–0.51 for other approaches), reflecting clear separation of CN states (Fig. 3e). Across all samples, TeaCNV achieved 98.6% overall accuracy (precision: 93.0%, recall: 96.1%, F1 score: 0.92; detailed sample-level confusion matrix metrics are provided in Supplementary Fig. 7a) in CNV event identification, outperforming existing methods (Fig. 3f and g). Additionally, TeaCNV showed the lowest error in absolute CN profiles (average RMSE: 0.07, Fig. 3h, Supplementary Methods), indicating precise quantification of integer CN profiling. Furthermore, analysis of the detected CNV size distribution revealed a minimum event length of 3.2 Mb (Supplementary Fig. 7b).

### Detection of rare clone in clear cell renal cell carcinoma

TeaCNV revealed distinct clonal architectures across the four ccRCC samples. While ccRCC1 displayed a monoclonal architecture, the other three cases showed polyclonal expansion dominated by a major subclone (Fig. 4). The clonal architectures and integer CN profiles remained robust regardless of the reference strategy used, whether using internal reference cell type (immune versus endothelial; Supplementary Fig. 8) or cross-sample references (Supplementary Fig. 9). Reliable reconstruction of clonal substructure required at least 20 reference cells, and performance remained robust across a range of reference cell fractions (Supplementary Fig. 10a and b). Additionally, at least 60% of epithelial cells were required to recapitulate the major population structure (Supplementary Fig. 10c).

Our analysis identified rare clones in samples with polyclonal substructures. In ccRCC2, two clones shared similar epithelial scores but diverged in chromatin accessibility patterns due to distinct absolute CNs at chr14 and chr16q. A rare clonal event involving chr14 loss was detected in 8.8% of tumor cells (Fig. 4b–d). For ccRCC3, TeaCNV resolved three subclones, including two minor subclones (clone2 and clone3: 32.59% combined) characterized by a 1-copy state at chr8p/16q, which is consistent with bulk WGS data (Figs 2c and 4e and f). A rare subclone (clone3: 5.19%) was further distinguished by 3 copies at chr8q/16p, accompanied by corresponding chromatin accessibility changes (Fig. 4g, Supplementary Fig. 11a). These rare clonal CNVs (gains of chr8q/16p) were also supported by inferCNV estimation but were undetectable in bulk WGS data due to limited cell representation (Fig. 2c and k). In ccRCC4, four subclones were stratified by distinct absolute CNs at chr20q and chr22 (Fig. 4h and i). Chr20q showed 4 copies in 66.4% of tumor cells and 2 copies in other

tumor cells, consistent with the observation of an average of 3 copies from bulk WGS data (Figs 2d and f and 4i). Additionally, we identified a minor clonal event (chr22 loss) in 22.9% of tumor cells (clones 3 and 4). A rare clone (clone 4: 10.5%) displayed co-occurrence of chr20q gain (4 copies) and chr22 loss (1-copy) (Fig. 4i, Supplementary Fig. 11b). These events were also supported by inferCNV but undetected in bulk WGS (Fig. 2d and l).

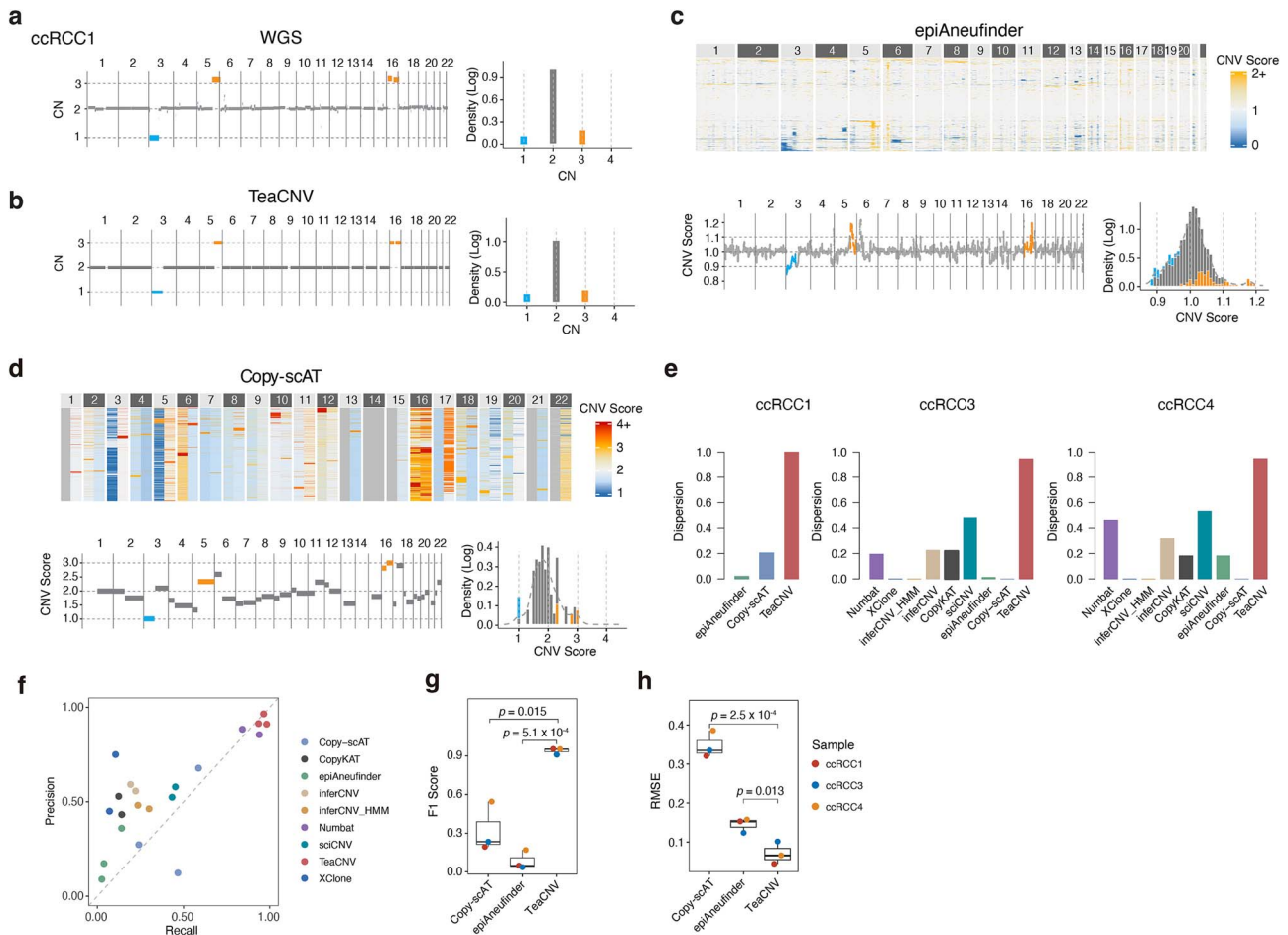
### Dissecting genomic and epigenomic contributions to clonal heterogeneity in solid tumors

We applied TeaCNV to published scATAC-seq datasets [30] spanning pancreatic (PDAC), breast (BRCA), head and neck (HNSCC), colorectal (CRC), and ovarian (OV) cancers (Supplementary Fig. 12).

Analysis of 1694 PDAC epithelial cells revealed 1 aneuploid and 1 diploid subpopulation, with the diploid subpopulation exhibiting lower epithelial scores (Fig. 5a and b). TeaCNV identified PDAC driver events, including *AKT2* amplifications and *SMAD4* deletions, with distinct chromatin accessibility patterns between subpopulations (Fig. 5b) [49]. Differential peak analysis showed that 64% of upregulated peaks in the aneuploid subpopulation were CNV-driven (i.e. peaks located within CNV regions), enriching hallmark pathways including xenobiotic metabolism, apical junction organization, KRAS signaling, and cholesterol homeostasis (Fig. 5c–e). CRE-regulated peaks (i.e. peaks located outside CNV regions) in the aneuploid subpopulation were associated with activation of P53 signaling, apoptosis, and glycolysis (Fig. 5e).

All epithelial cells in the BRCA, HNSCC, OV, and CRC tumor samples were aneuploid and exhibited polyclonal architectures. In the BRCA sample, truncal CNVs involved driver genes (*MCL1*, *MYC*, *CSMD1*, *RB1*, and *WWOX*, Fig. 5f and g) [50]. Subclonal CNV events (chr11q, chr20q13.2/*ZNF217*) delineated three subclones with similar epithelial scores but distinct ATAC signals (Fig. 5g). More than half of the subclonal-specific peaks were CNV-driven, dominating the phenotypic diversity across subclones (Fig. 5h and i). In the HNSCC sample, truncal CNVs involved known driver genes, including *LRP1B*, *FGFR1*, and *MYC* (bottom of Fig. 5j) [51]. Subclonal events harboring cancer genes (*SOX2*, *PDE4D*, and *KLF1*) drove clone-specific accessibility patterns (Fig. 5j). CNV-driven peaks mediated P53 pathway dysregulation in both subclones and specifically proliferation activation in clone 2 (G2M checkpoint, *MYC*, and mitotic spindle), highlighting the major contribution of CNVs to phenotypic divergence across subclones (Fig. 5k and l).

In the OV sample, truncal CNVs involved *MCL1*, *MECOM*, *KRAS*, *METTL17*, and *CCNE1* [52] (Fig. 5m and n). Despite comparable proportions of CNV-driven and CRE-regulated peaks within each subclone, pathway diversity across subclones was primarily CRE-mediated (cholesterol homeostasis in clone 2, adipogenesis/UV response in clone 3, Fig. 5o and p). Similarly, in the CRC tumor sample, subclonal CNVs altered chromatin accessibility patterns, but most enriched pathways were CRE-regulated (Fig. 5q–s) [53]. Moreover, TeaCNV correctly inferred a diploidy copy number profile across epithelial cells using immune cells as the reference population in normal CRC (Supplementary Fig. 12f). It further recovered major CNV patterns and clonal substructures across multiple solid tumors, even when using ccRCC-derived non-epithelial cells as a cross-dataset reference (Supplementary Fig. 13). These findings demonstrate the complex interplay of genomic and epigenomic mechanisms in driving tumor subclonal heterogeneity.



**Figure 3** Performance evaluation of CNV detection methods. (a) Integer CN profiles derived from bulk WGS (left) and the corresponding genome-wide distribution of absolute CN states (right) for ccRCC1. Significantly altered regions are highlighted in blue (CN loss) and orange (CN gain). (b) Clonal integer CN profiles inferred by TeaCNV from chromatin accessibility data (left) and the genome-wide distribution of estimated integer CN states (right) for ccRCC1. (c, d) Relative CN profiles (top) and genome-wide distributions of relative CNV scores (bottom) estimated by (c) epiAneufinder and (d) Copy-scAT for ccRCC1. Blue (CN loss) and orange (CN gain) bars denote CNV regions identified by bulk WGS. (e) Dispersion scores of estimated CNs across genomic segments with distinct absolute CNs defined by bulk WGS. inferCNV-HMM indicates results from inferCNV in copy number state mode. Precision and recall (f), F1 score (g) for CNV detection across methods. Each dot represents one sample. *P*-values were calculated using Wilcoxon test. (h) Average deviation of inferred CN profiles from bulk WGS. RMSE: root mean square error.

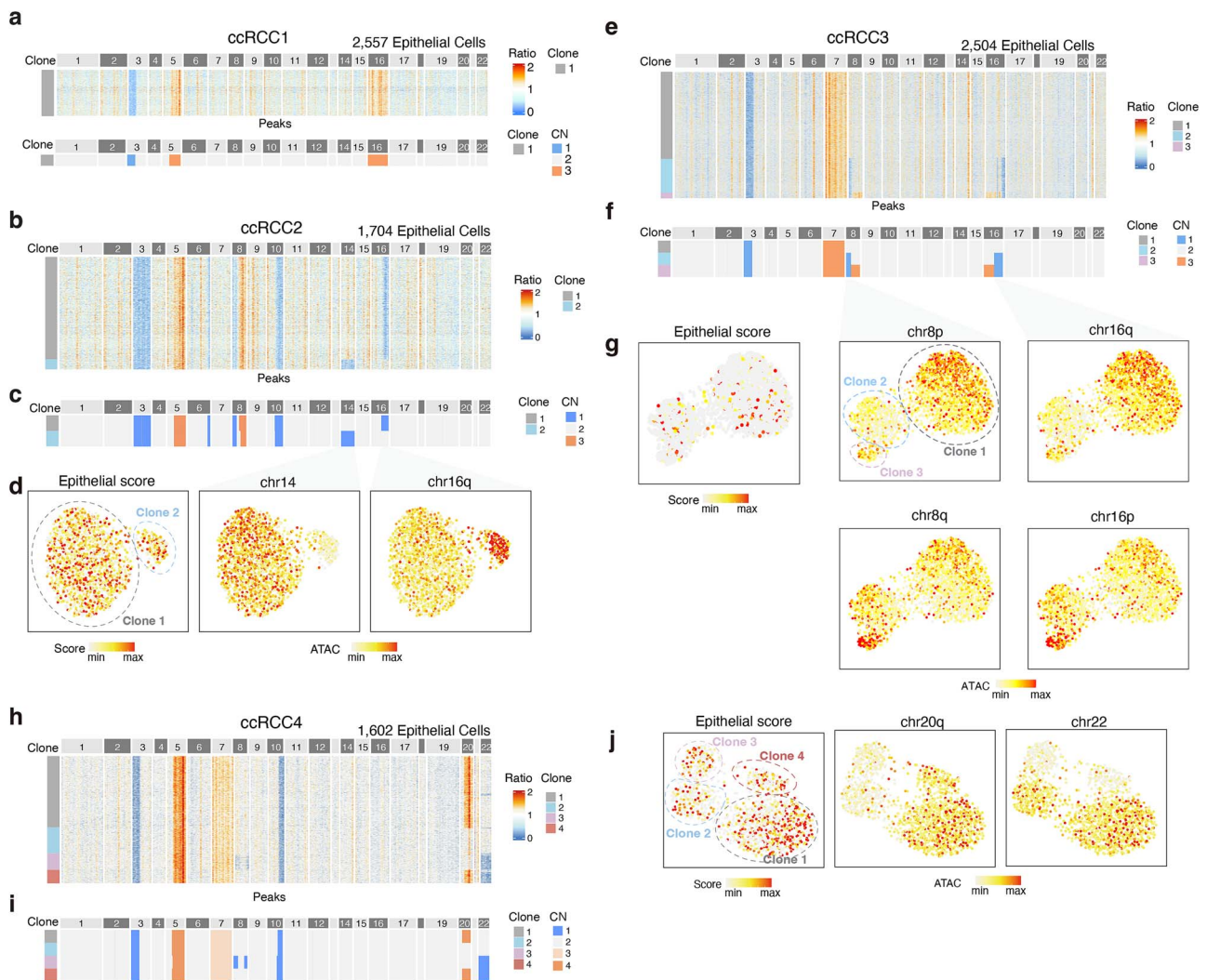
## Discussion

We developed TeaCNV, a computational framework that enables robust estimation of absolute CN profiles and clonal architecture directly from scATAC-seq data, without requiring matched bulk DNA sequencing. TeaCNV resolves CN states beyond simple gain or loss, achieving an absolute CN quantification error of less than 0.5, as validated by simulated and bulk WGS data. Additionally, TeaCNV improves the resolution of tumor heterogeneity in scATAC-seq datasets, particularly for detecting rare clonal CNVs with a prevalence below 10%, which are often undetectable in bulk WGS data (Supplementary Fig. 14). Finally, TeaCNV facilitated the decoupling of genomic and epigenomic contributions to phenotypic heterogeneity, revealing the mechanisms underlying diverse phenotypes across tumor clones.

ScATAC-seq-based CNV detection faces fundamental limitations, including mappability bias, GC content effects, extreme data sparsity, and CRE-related biases [35, 54]. While existing methods address GC content and mappability issues, they remained constrained by sparsity and CRE biases, limiting detection to relative gains or losses

at arm-level genomic scales rather than absolute CNs quantification [33, 34]. Although the scDNA-seq-based method Alleloscope [35] can be applied to scATAC-seq data to estimate allelic copy number, its dependence on matched bulk DNA-seq data poses a major limitation.

TeaCNV addresses potential mappability bias by removing low-quality peaks, a filtering strategy (e.g. excluding ENCODE Blacklist regions) similar to other methods. To minimize inadvertent discarding signals from biologically meaningful regions, we apply this filtering conservatively while aiming to balance signal detection with reduced systematic bias. In addition, TeaCNV calculates inferred-to-reference ratios, which normalize GC-related fluctuations and are independent of GC content (Supplementary Fig. 15). By merging adjacent peaks into genomic windows and aggregating sparse signals within cell sub-populations, TeaCNV reduces the biases introduced by data sparsity and CRE effects. In ccRCC samples, differential CREs explained less than 10% of the CNV variance and over 50% of differential CREs were located outside CNV regions, indicating that TeaCNV effectively reduces confounding effects (Supplementary Fig. 16).

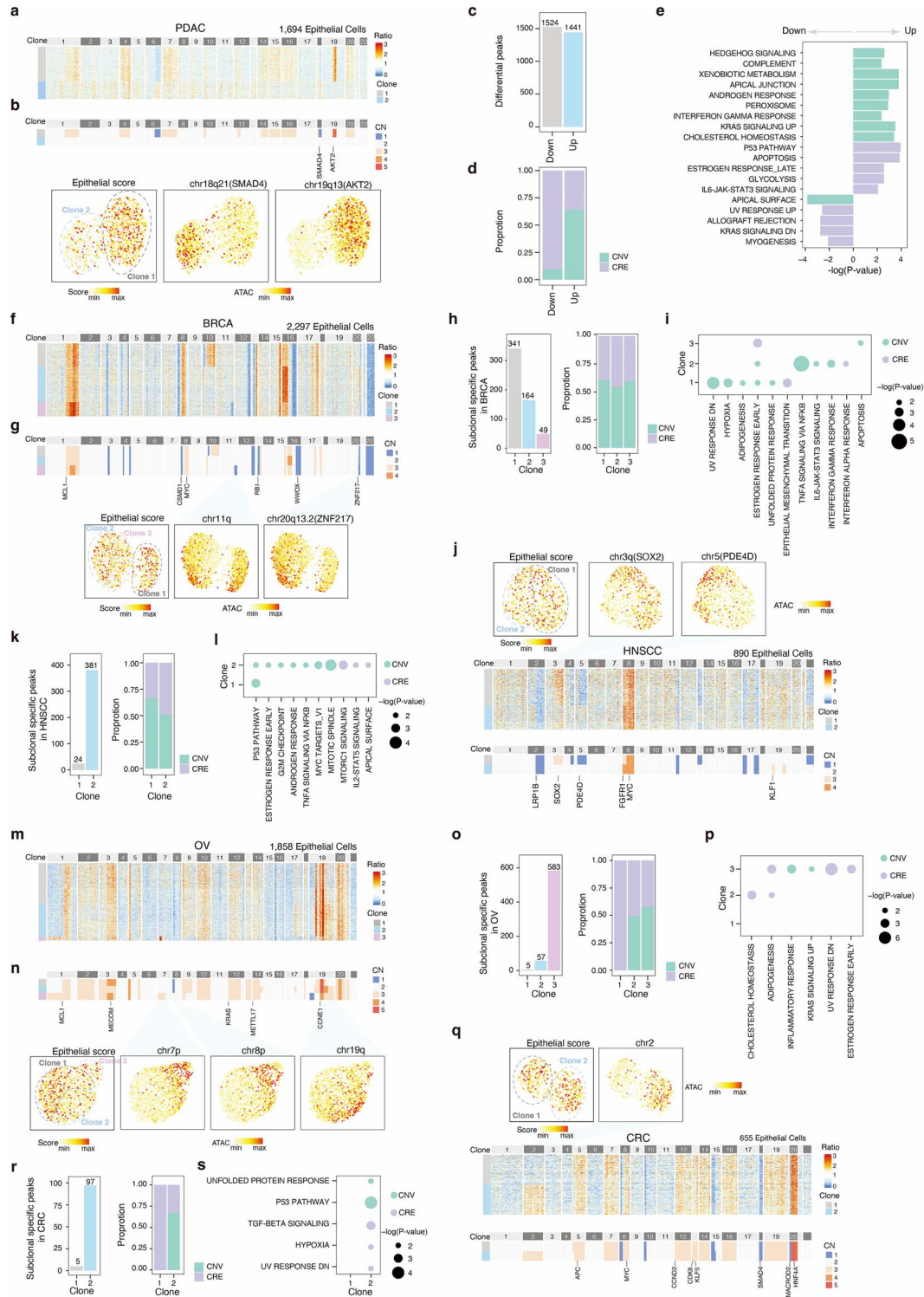


**Figure 4** Resolving clonal substructures in ccRCC patients using scATAC-seq and scATAC&RNA-seq co-assay data. (a) Heatmap of copy ratio profiles in epithelial cells (top) and inferred clonal absolute CN profiles (bottom) for ccRCC1. Rows are annotated by clones. (b, c) TeaCNV results for ccRCC2: (b) copy ratio heatmap and (c) inferred clonal absolute CN profiles. (d) UMAP plots of epithelial cells from ccRCC2 colored by tumor activity scores (*EPCAM*, *KRT18*, *KRT8*, *KRT19*) and by ATAC signals from representative subclonal regions. (e, f) TeaCNV results for ccRCC3: (e) copy ratio heatmap and (f) inferred clonal absolute CN profiles. (g) UMAP plots of epithelial cells from ccRCC3 colored as in (d). (h, i) TeaCNV results for ccRCC4: (h) copy ratio heatmap and (i) inferred clonal absolute CN profiles. (j) UMAP plots of epithelial cells from ccRCC4 colored as in (d).

Applied to 6 cancer types, TeaCNV revealed polyclonal architectures in 77.78% of cases, demonstrating its broad utility across solid tumors. By enabling precise deconvolution of clonal substructure and absolute CN quantification, TeaCNV supports accurate functional dissection of large-scale CNV regions as well as the contribution to phenotypic diversity [10, 55–57]. We observed numerous differential chromatin accessibility loci across intra-tumor subclones, which underlie phenotypic variations. These differential loci arise through distinct mechanisms: one is driven by CNVs that directly alter chromatin accessibility within affected regions (accounting for a substantial fraction of variance), and another involves CREs located outside CNV regions, which are largely independent of CNVs (Supplementary Fig. 16). Therefore, the complex interplay between genomic and epigenomic mechanisms driving tumor subclonal heterogeneity can be dissected from scATAC-seq using the absolute CN inference. Furthermore, TeaCNV facilitates distinguishing the distinct oncogenic roles of co-altered genes in single-cell multi-omics (scATAC&RNA-seq co-assay) sequencing data,

where chromosomal dosage effects can be discerned from large-scale gene expression changes associated with absolute copy number.

Four considerations guide the application of TeaCNV. First, it performs optimally on subclones composed of sufficiently homogeneous cells, as sparse data limits the ability to call single-cell CNVs—a limitation shared by all scATAC-seq-based approaches. Tumor cell lines or samples with continuous clonal gradients or high subclonal diversity (e.g. those containing numerous rare subpopulations) may yield unstable estimates. Based on benchmarking with simulated data, the minimal subclonal size for reliable detection is 20 cells. Accordingly, TeaCNV is optimized for dissecting clonal architecture and genotype-epigenotype coupling at the subclonal level, rather than at the level of individual cells. Second, robust reference cell selection is critical, particularly for highly malignant samples lacking internal normal cells. TeaCNV ideally requires reference cells from the matched sample with at least 20 cells. Estimation using cross-dataset referencing can also capture the main clonal substructure and CNVs present in the



**Figure 5** Clonal substructure in solid tumors. TeacNV results for PDAC (a, b), BRCA (f, g), OV (m, n) samples: (a, f, m) copy ratio heatmaps and (b, g, n) inferred clonal absolute CN profiles (top) and UMAP plots (bottom) of epithelial cells colored by activity scores of tumor epithelial markers and ATAC-seq signals from representative subclonal regions. Known driver genes are highlighted in the clonal absolute CN profiles. (c) Number of differential accessibility peaks between aneuploid and diploid subpopulations in the PDAC sample. "Down" and "Up" denote peaks with decreased or increased accessibility in the aneuploid versus diploid subpopulation, respectively. (d) Proportion of differential peaks driven by CNVs and CREs in the PDAC sample. (e) Hallmark pathways significantly enriched by CNV-driven and CRE-regulated peak associated genes in PDAC. (j, q) TeacNV results for HNSCC (j) and CRC (q) samples: UMAP plots (top) of epithelial cells colored by activity scores of tumor epithelial markers and ATAC-seq signals from representative subclonal regions, copy ratio heatmap (middle), and inferred clonal absolute CN profiles (bottom). Subclone-specific accessibility analysis for BRCA (h, i), HNSCC (k, l), OV (o, p), CRC (r, s) samples: (h, k, o, r) the number of subclone-specific peaks (left), and the proportion of CNV-driven and CRE-regulated peaks (right); (i, l, p, s) hallmark pathways significantly enriched by CNV-driven and CRE-regulated subclone-specific peak associated genes.

tumor cell population, but may introduce deviations in subclonal CNV events; thus, intra-dataset or batch-matched referencing is preferred over cross-dataset approaches for superior robustness. Third, although TeaCNV performs robustly for broad and large sub-arm-level CNVs, its resolution for focal events is inherently limited by the intrinsic sparsity of scATAC-seq data. Very short genomic intervals often lack sufficient accessible chromatin coverage and signal-to-noise ratio for sensitive CNV detection and precise breakpoint resolution. Consequently, focal amplifications and deletions, including potentially clinically actionable events, may be missed or not resolved as independent CNVs (Supplementary Fig. 7c–e). This limitation becomes particularly evident for smaller focal events, with sensitivity declining markedly below 10 Mb and approaching zero for events smaller than 2 Mb. Lastly, TeaCNV currently infers total CN and does not resolve allele-specific alterations (e.g. loss of heterozygosity or copy-neutral events), which may limit analyses of allelic phenomena such as tumor suppressor loss and immune evasion. Efforts are underway to extend it by incorporating phased genetic variation and/or joint modeling with multi-omic data, which may enable allele-specific CNV inference and further enhance its utility for analyzing complex tumor genomes.

Tumor plasticity arises from aberrant activation of transcriptional programs caused by genetic and non-genetic mechanisms [58–60]. With advancements in single-cell sequencing technologies, simultaneous profiling of the epigenome and transcriptome at single-cell resolution is now possible [61, 62]. TeaCNV bridges this gap in single-cell multiomics by connecting clonal genotypes to epigenomic and transcriptomic phenotypes. By enabling absolute CNV analysis in any scATAC-seq dataset, TeaCNV empowers systematic exploration of how genomic instability shapes epigenetic diversity—a critical factor underlying therapeutic resistance and metastatic progression.

### Key Points

- We developed TeaCNV, a computational framework that infers absolute, genome-wide copy number profiles directly from scATAC-seq, without requiring matched DNA baselines, and robustly determines genome ploidy.
- TeaCNV deconvolves high-resolution clonal substructure, reconstructing CNV-driven polyclonal architectures in most tumors and detecting rare subclones with high sensitivity.
- Across simulated, in-house and public datasets from six cancer types, TeaCNV outperforms existing scATAC- and scRNA-based methods, recovering rare CNVs and clinically relevant driver alterations missed by bulk WGS.
- TeaCNV delineates subclonal CNV architectures and links this genotypic diversity to epigenotypic and transcriptional changes, such as CNV-associated chromatin remodeling, transcriptional heterogeneity, and dysregulated signaling pathways within tumor clones.

### Acknowledgements

We thank Dr. Gui-Hua Zhao for meticulously verifying the entire methodological derivation, ensuring consistency with measure-theoretic probability principles, and validating all statistical formulations.

### Author contributions

Y. W. and S. Z.: Conceptualization, Methodology, Writing – original draft; Y. W.: Software, Formal analysis, with contributions from Y. Z.; X. Y.: Resources; Y. D., Y. C. and Z. C.: Investigation, Data curation; Y. W. performed the Formal analysis and Visualization with the help from M. Z., X. W. and H. L.; S. Z.: Supervision. All authors: Writing – review & editing.

### Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

### Competing interests

The authors declare no competing interests.

### Funding

This work was supported by the National Natural Science Foundation of China (grants 32400528 to Y.W., grants 32570750 and 32170666 to S.Z.).

### Data availability

The scATAC-seq, scATAC&RNA-seq co-assays and bulk WGS validation data of ccRCC can be accessed in Zenodo at <https://doi.org/10.5281/zenodo.14190637>. scATAC-seq data of BRCA, PDAC, HNSCC, CRC, and OV were downloaded from Gene Expression Omnibus (GEO) with accession number GSE240822.

### Code availability

The TeaCNV algorithm is available at <https://github.com/ShaojunLab/TeaCNV>. The analysis scripts used to reproduce results and simulation scripts included in the paper are available at <https://github.com/ShaojunLab/TeaCNVanalysis>.

### References

1. Esteller M, Dawson MA, Kadoch C *et al*. The epigenetic hallmarks of cancer. *Cancer Discov* 2024;**14**:1783–809. <https://doi.org/10.1158/2159-8290.CD-24-0296>
2. Chakravarthi BV, Nepal S, Varambally S. Genomic and epigenomic alterations in cancer. *Am J Pathol* 2016;**186**:1724–35. <https://doi.org/10.1016/j.ajpath.2016.02.023>
3. Friedlander ML, Hedley DW, Taylor IW. Clinical and biological significance of aneuploidy in human tumours. *J Clin Pathol* 1984;**37**:961–74. <https://doi.org/10.1136/jcp.37.9.961>
4. Clark GM, Dressler LG, Owens MA *et al*. Prediction of relapse or survival in patients with node-negative breast cancer by DNA flow cytometry. *N Engl J Med* 1989;**320**:627–33. <https://doi.org/10.1056/NEJM198903093201003>
5. Kallioniemi OP, Punnonen R, Mattila J *et al*. Prognostic significance of DNA index, multiploidy, and S-phase fraction in ovarian cancer. *Cancer* 1988;**61**:334–9. [https://doi.org/10.1002/1097-0142\(19880115\)61:2<334::aid-cnrcr2820610224>3.0.co;2-m](https://doi.org/10.1002/1097-0142(19880115)61:2<334::aid-cnrcr2820610224>3.0.co;2-m)
6. Choma D, Daures JP, Quantin X *et al*. Aneuploidy and prognosis of non-small-cell lung cancer: a meta-analysis of published data. *Br J Cancer* 2001;**85**:14–22. <https://doi.org/10.1054/bjoc.2001.1892>

7. Bielski CM, Zehir A, Penson AV *et al*. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat Genet* 2018;**50**:1189–95. <https://doi.org/10.1038/s41588-018-0165-1>
8. Carter SL, Cibulskis K, Helman E *et al*. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 2012;**30**:413–21. <https://doi.org/10.1038/nbt.2203>
9. Salehi S, Kabeer F, Ceglia N *et al*. Clonal fitness inferred from time-series modelling of single-cell cancer genomes. *Nature* 2021;**595**:585–90. <https://doi.org/10.1038/s41586-021-03648-3>
10. Minussi DC, Nicholson MD, Ye H *et al*. Breast tumours maintain a reservoir of subclonal diversity during expansion. *Nature* 2021;**592**:302–8. <https://doi.org/10.1038/s41586-021-03357-x>
11. Casasent AK, Schalck A, Gao R *et al*. Multiclonal invasion in breast tumors identified by topographic single cell sequencing. *Cell* 2018;**172**:205–217.e12. <https://doi.org/10.1016/j.cell.2017.12.007>
12. Tarabichi M, Salcedo A, Deshwar AG *et al*. A practical guide to cancer subclonal reconstruction from DNA sequencing. *Nat Methods* 2021;**18**:144–55. <https://doi.org/10.1038/s41592-020-01013-2>
13. Navin N, Kendall J, Troge J *et al*. Tumour evolution inferred by single-cell sequencing. *Nature* 2011;**472**:90–4. <https://doi.org/10.1038/nature09807>
14. Sharma A, Merritt E, Hu X *et al*. Non-genetic intra-tumor heterogeneity is a major predictor of phenotypic heterogeneity and ongoing evolutionary dynamics in lung tumors. *Cell Rep* 2019;**29**:2164–2174.e5. <https://doi.org/10.1016/j.celrep.2019.10.045>
15. Mo CK, Liu J, Chen S *et al*. Tumour evolution and microenvironment interactions in 2D and 3D space. *Nature* 2024;**634**:1178–86. <https://doi.org/10.1038/s41586-024-08087-4>
16. Lei Y, Tang R, Xu J *et al*. Applications of single-cell sequencing in cancer research: progress and perspectives. *J Hematol Oncol* 2021;**14**:91. <https://doi.org/10.1186/s13045-021-01105-2>
17. Joanito I, Wirapati P, Zhao N *et al*. Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. *Nat Genet* 2022;**54**:963–75. <https://doi.org/10.1038/s41588-022-01100-4>
18. Bischoff P, Trinks A, Obermayer B *et al*. Single-cell RNA sequencing reveals distinct tumor microenvironmental patterns in lung adenocarcinoma. *Oncogene* 2021;**40**:6748–58. <https://doi.org/10.1038/s41388-021-02054-3>
19. Hou Y, Guo H, Cao C *et al*. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res* 2016;**26**:304–19. <https://doi.org/10.1038/cr.2016.23>
20. Patel AP, Tirosh I, Trombetta JJ *et al*. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 2014;**344**:1396–401. <https://doi.org/10.1126/science.1254257>
21. Tirosh I, Izar B, Prakadan SM *et al*. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 2016;**352**:189–96. <https://doi.org/10.1126/science.aad0501>
22. Tirosh I, Venteicher AS, Hebert C *et al*. Single-cell RNA-seq supports a developmental hierarchy in human oligodendrogloma. *Nature* 2016;**539**:309–13. <https://doi.org/10.1038/nature20123>
23. Venteicher AS, Tirosh I, Hebert C *et al*. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* 2017;**355**:eaai8478. <https://doi.org/10.1126/science.aai8478>
24. Puram SV, Tirosh I, Parkh AS *et al*. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* 2017;**171**:1611–1624.e24. <https://doi.org/10.1016/j.cell.2017.10.044>
25. Fan J, Lee HO, Lee S *et al*. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res* 2018;**28**:1217–27. <https://doi.org/10.1101/gr.228080.117>
26. Gao R, Bai S, Henderson YC *et al*. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat Biotechnol* 2021;**39**:599–608. <https://doi.org/10.1038/s41587-020-00795-2>
27. Gao T, Soldatov R, Sarkar H *et al*. Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. *Nat Biotechnol* 2023;**41**:417–26. <https://doi.org/10.1038/s41587-022-01468-y>
28. Huang R, Huang X, Tong Y *et al*. Robust analysis of allele-specific copy number alterations from scRNA-seq data with XClone. *Nat Commun* 2024;**15**:6684, 6684. <https://doi.org/10.1038/s41467-024-51026-0>
29. Sundaram L, Kumar A, Zatzman M *et al*. Single-cell chromatin accessibility reveals malignant regulatory programs in primary human cancers. *Science* 2024;**385**:eadk9217. <https://doi.org/10.1126/science.adk9217>
30. Terekhanova NV, Karpova A, Liang WW *et al*. Epigenetic regulation during cancer transitions across 11 tumour types. *Nature* 2023;**623**:432–41. <https://doi.org/10.1038/s41586-023-06682-5>
31. De Rop FV, Hulselmans G, Flerin C *et al*. Systematic benchmarking of single-cell ATAC-sequencing protocols. *Nat Biotechnol* 2024;**42**:916–26. <https://doi.org/10.1038/s41587-023-01881-x>
32. Hu Y, Jiang Z, Chen K *et al*. scNanoATAC-seq: a long-read single-cell ATAC sequencing method to detect chromatin accessibility and genetic variants simultaneously within an individual cell. *Cell Res* 2023;**33**:83–6. <https://doi.org/10.1038/s41422-022-00730-x>
33. Ramakrishnan A, Symeonidi A, Hanel P *et al*. epiAneufinder identifies copy number alterations from single-cell ATAC-seq data. *Nat Commun* 2023;**14**:5846. <https://doi.org/10.1038/s41467-023-41076-1>
34. Nikolic A, Singhal D, Ellestad K *et al*. Copy-scAT: deconvoluting single-cell chromatin accessibility of genetic subclones in cancer. *Sci Adv* 2021;**7**:eabg6045. <https://doi.org/10.1126/sciadv.abg6045>
35. Wu CY, Lau BT, Kim HS *et al*. Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer. *Nat Biotechnol* 2021;**39**:1259–69. <https://doi.org/10.1038/s41587-021-00911-w>
36. Zhang Y, Liu T, Meyer CA *et al*. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008;**9**:R137. <https://doi.org/10.1186/gb-2008-9-9-r137>
37. Stuart T, Srivastava A, Madad S *et al*. Single-cell chromatin state analysis with Signac. *Nat Methods* 2021;**18**:1333–41. <https://doi.org/10.1038/s41592-021-01282-5>
38. Hao Y, Hao S, Andersen-Nissen E *et al*. Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–3587 e29. <https://doi.org/10.1016/j.cell.2021.04.048>
39. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**:1754–60. <https://doi.org/10.1093/bioinformatics/btp324>
40. Danecek P, Bonfield JK, Liddle J *et al*. Twelve years of SAMtools and BCFtools. *GigaScience* 2021;**10**:giab008. <https://doi.org/10.1093/gigascience/giab008>
41. Wang J, Minussi DC, Davis A *et al*. Delineating the copy-number substructure of metastatic tumors with CopyKit. *Mol Cell* 2026;**86**:407–422.e6. <https://doi.org/10.1016/j.molcel.2025.12.026>
42. Killick R, Eckley IA. Changepoint: AnRPackage for changepoint analysis. *J Stat Softw* 2014;**58**:1–19. <https://doi.org/10.18637/jss.v058.i03>

43. Fearnhead P, Rigai G. Change point detection in the presence of outliers. *J Am Stat Assoc* 2018;**114**:169–83. <https://doi.org/10.1080/01621459.2017.1385466>
44. Mahdipour-Shirayeh A, Erdmann N, Leung-Hagesteijn C *et al*. sci-CNV: high-throughput paired profiling of transcriptomes and DNA copy number variations at single-cell resolution. *Brief Bioinform* 2022;**23**:bbab413. <https://doi.org/10.1093/bib/bbab413>
45. Brodersen KH, Ong CS, Stephan KE *et al*. The balanced accuracy and its posterior distribution. *2010 20th International Conference on Pattern Recognition, Istanbul, Turkey*, 2010, pp. 3121–3124, <https://doi.org/10.1109/ICPR.2010.764>
46. Liberzon A, Birger C, Thorvaldsdottir H *et al*. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst* 2015;**1**:417–25. <https://doi.org/10.1016/j.cels.2015.12.004>
47. Killick R, Fearnhead P, Eckley IA. Optimal detection of changepoints with a linear computational cost. *J Am Stat Assoc* 2012;**107**:1590–8. <https://doi.org/10.1080/01621459.2012.737745>
48. Navidi Z, Zhang L, Wang B. simATAC: a single-cell ATAC-seq simulation framework. *Genome Biol* 2021;**22**:74. <https://doi.org/10.1186/s13059-021-02270-w>
49. Cancer Genome Atlas Research Network. Electronic address aadhe, Cancer Genome Atlas Research N. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* 2017;**32**:185–203.e13. <https://doi.org/10.1016/j.ccell.2017.07.007>
50. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;**490**:61–70. <https://doi.org/10.1038/nature11412>
51. Cancer Genome Atlas N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 2015;**517**:576–82. <https://doi.org/10.1038/nature14129>
52. Cancer Genome Atlas Research N. Integrated genomic analyses of ovarian carcinoma. *Nature* 2011;**474**:609–15. <https://doi.org/10.1038/nature10166>
53. Cancer Genome Atlas Research N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012;**487**:330–7. <https://doi.org/10.1038/nature11252>
54. Satpathy AT, Granja JM, Yost KE *et al*. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* 2019;**37**:925–36. <https://doi.org/10.1038/s41587-019-0206-z>
55. Rice AM, McLysaght A. Dosage sensitivity is a major determinant of human copy number variant pathogenicity. *Nat Commun* 2017;**8**:14366. <https://doi.org/10.1038/ncomms14366>
56. Wang F, Zhang S, Kim TB *et al*. Integrated transcriptomic-genomic tool Texomer profiles cancer tissues. *Nat Methods* 2019;**16**:401–4. <https://doi.org/10.1038/s41592-019-0388-9>
57. Shi H, Williams MJ, Satas G *et al*. Allele-specific transcriptional effects of subclonal copy number alterations enable genotype-phenotype mapping in cancer cells. *Nat Commun* 2024;**15**:2482. <https://doi.org/10.1038/s41467-024-46710-0>
58. Shlyakhtina Y, Moran KL, Portal MM. Genetic and non-genetic mechanisms underlying cancer evolution. *Cancers (Basel)* 2021;**13**:1380. <https://doi.org/10.3390/cancers13061380>
59. Marine JC, Dawson SJ, Dawson MA. Non-genetic mechanisms of therapeutic resistance in cancer. *Nat Rev Cancer* 2020;**20**:743–56. <https://doi.org/10.1038/s41568-020-00302-4>
60. Tellez-Gabriel M, Ory B, Lamoureux F *et al*. Tumour heterogeneity: the key advantages of single-cell analysis. *Int J Mol Sci* 2016;**17**:2142. <https://doi.org/10.3390/ijms17122142>
61. Lee J, Hyeon DY, Hwang D. Single-cell multiomics: technologies and data analysis methods. *Exp Mol Med* 2020;**52**:1428–42. <https://doi.org/10.1038/s12276-020-0420-2>
62. Baysoy A, Bai Z, Satija R *et al*. The technological landscape and applications of single-cell multi-omics. *Nat Rev Mol Cell Biol* 2023;**24**:695–713. <https://doi.org/10.1038/s41580-023-00615-w>