# InferPloidy: A fast ploidy inference tool accurately classifies cells with abnormal CNVs in large single-cell RNA-seq datasets

Wonjung Sung[1], Jaeyoung Chae[1], Jucheol Moon[2], and Seokhyun Yoon[3]

[1] Dept. of AI Convergence, College of SW Convergence, Dankook University, Yongin-si 16890, Republic of Korea
[2] Dept. of Computer Engineering and Computer Science, College of Engineering, California State University Long Beach, CA 90840, United States
[3] Dept. of Electronics & Electrical Engineering, College of Engineering, Dankook University, Yongin-si 16890, Republic of Korea

Corresponding: Seokhyun Yoon (syoon@dku.edu)

## Abstract

Estimation of copy number variation (CNV) and ploidy inference provides valuable insights into the structural and clonal characteristics of tumor cells in single-cell RNA-seq datasets. Several tools have been developed for these purposes, including CopyKat and SCEVAN, but one limitation is their running times, which hinder their use for large datasets. In this study, we present InferPloidy, a faster and more accurate ploidy inference tool that operates on top of InferCNV. Rather than focusing on precise CNV segmentation or the discovery of intra-tumoral heterogeneity, InferPloidy emphasizes much faster and more accurate classification of aneuploid cells from diploids, which is crucial for identifying diagnostic markers or druggable targets. The accurate identification of malignant cells and the scalability of the tool to handle large datasets with many samples are key aspects of this work. InferPloidy is two orders of magnitude faster than existing tools while providing more accurate classifications, enabling fast and precise downstream analyses, including inter-tumoral heterogeneity studies across different patients.

## Introduction

Single-cell RNA sequencing (scRNA-seq)[1,2,3] is a powerful technique for dissecting cancer tissues and the tumor microenvironment (TME)—a complex ecosystem composed of tumor cells, immune cells, and stromal subsets[4,5,6,7]. A major challenge in scRNA-seq-based TME studies is understanding intra and inter tumoral heterogeneity[8] to get critical insights into tumor evolution and progression[9, 10,11].

A crucial step in scRNA-seq data analysis for studying TME and TH is the precise classification of tumor cells, not only from normal cells (including immune and stromal cells) but also from normal tumor-originating cells. Accurately distinguishing between tumor and normal cells enables researchers to explore tumor heterogeneity, investigate carcinogenesis, and study tumor-normal cell interactions within the TME[12,13,14]. Furthermore, scRNA-seq analysis of cancer tissues has been instrumental in uncovering mechanisms of drug resistance, identifying novel biomarkers, and exploring therapeutic strategies[15,16,17].

Several computational tools have been developed for estimation of copy number variation (CNV) and tumor cell classification at the single-cell level, including CopyKat[18] and SCEVAN[19]. Other related tools, such as InferCNV[20] and sciCNV[21], focus solely on CNV estimation. Both CopyKat and SCEVAN provide CNV estimates along with ploidy classification, distinguishing cells as either diploid or aneuploid[22, 23,24]. Accurate CNV estimation allows researchers to study intra-tumoral heterogeneity and tumor evolution by identifying clonal substructures and their phylogeny[9, 10,11].

However, a major limitation of these tools is their computational inefficiency. While SCEVAN is somewhat faster than CopyKat, both require hours to process moderate-sized datasets (tens of thousands of cells) and days for large datasets (hundreds of thousands of cells). This runtime is significantly longer compared to other scRNA-seq analysis tools used for cell type annotation, cell-cell interaction inference, and differentially expressed gene (DEG) analysis, making CNV-based tumor classification a computational bottleneck and limiting its integration into comprehensive single-cell analysis pipelines.

To address these challenges and enable rapid analysis of large scRNA-seq datasets, we introduce InferPloidy, a high-speed and highly accurate ploidy inference tool that runs on top of InferCNV.

InferPloidy facilitates the identification of tumor heterogeneity (TH) across individuals and cancer types, enabling pan-cancer heterogeneity studies and supporting cancer subtyping for biomarker discovery and drug target identification.

## Results

### Overview of InferPloidy

InferPloidy utilizes pre-computed CNV estimates and a predefined set of normal reference cells. To achieve this, we first use HiCAT[25] for cell-type annotation and then collect T cells, B cells, myeloid cells, and fibroblasts to use as normal reference cells for InferCNV (Figure 1a, b). The core idea is to use CNV estimates and predefined normal reference cells to initially classify cells into three groups: normal, tumoral, and uncertain (others). Using the first two groups, CNV profiles for normal and tumor cells are computed, allowing an initial classification not only for the uncertain cells but also for tumoral cells.

Specifically, we first perform dimensionality reduction (using PCA) and clustering on the CNV estimates to construct a cluster adjacency graph (CAG) (Figure 1c). The CAG can be efficiently generated using a graph-based clustering algorithm, such as Louvain clustering[26]. Based on this CAG, clusters are divided into three categories: (i) the normal reference clusters, (ii) the clusters that are "reachable" from the normal reference clusters, and (iii) those "unreachable" from the normal reference clusters (Figure 1d). To this end, the normal reference cluster is first defined as the cluster where the proportion of normal reference cells (T cells, B cells, and myeloid cells) exceeds a certain threshold. With the normal reference clusters, we next assess reachability from the normal reference clusters to classify the remaining clusters into "reachable" and "unreachable". Clusters unreachable from the normal reference clusters are initially designated as tumor reference clusters while those reachable are set "uncertain". Using these normal and tumor reference clusters, the CNVs of normal and tumor cells are approximated via a Gaussian Mixture Model (GMM) with a diagonal covariance constraint, where model parameters are estimated using the Expectation-Maximization (EM) algorithm (Figure 1e, f).

Ploidy decisions are then made on the cells in the tumor reference clusters and the uncertain clusters by computing the likelihood ratio between the GMMs of normal and tumor cells (Figure 1g, h, i). Steps d to g in Figure 1 are iteratively repeated until either the ploidy classifications remain unchanged or a predefined maximum number of iterations is reached (inner loop for iterative refinement). This entire process, from step c to g (including the inner loop), is also repeated multiple times (outer loop) to obtain a final ploidy classification via majority voting. The outer loop ensures robustness by aggregating results from clustering runs with different random initializations.

A detailed description of the InferPloidy workflow can be found in the Methods section.

### Qualitative comparison of InferPloidy and other tools

We first compared the ploidy classification of InferPloidy with that of CopyKat and SCEVAN using the lung cancer dataset GSE131907[27] (Figure 2 a-c). The original dataset consists of 58 samples, including normal lung, tumor lung, lymph node, and brain metastases, totaling 208,000 cells. To reduce the runtime of CopyKat and SCEVAN (but not InferPloidy), we excluded lymph node and brain metastasis samples and randomly selected up to 1,200 cells per sample from the remaining three conditions (distant normal, early tumor, and advanced tumor), analyzing a total of 36,000 cells without splitting them into individual samples. For a fair comparison, all tools used the same set of normal reference cells, including T cells, B cells, myeloid cells, and fibroblasts.

When classifying cells into aneuploid and diploid groups, we observed that CopyKat and SCEVAN misclassified a non-negligible proportion of cells from the normal condition as aneuploid, which is a clear classification error, as all cells from the distant normal condition should be classified as diploid. In contrast, InferCNV + InferPloidy did not exhibit such errors, demonstrating its superiority in ploidy inference. Another drawback of CNV estimation in CopyKat and SCEVAN was that CNVs for a considerable portion of cells were indeterminate, indicated as either 'not.defined' or 'filtered' (Figure 2 a and b). This reduces the number of cells usable not only for ploidy inference but also for downstream analyses, such as marker or druggable target discovery. Similar tendencies could also be observed in another dataset, GSE161529 for breast cancer (Figure 2 d-f), where we have collected 36K cells from 26 samples in a similar way to GSE131907.

The UMAP plots provide a detailed visualization of the ploidy classifications from the three tools

(Figure 3). Note that the cells in the region where normal reference cells are predominant are likely to be normal, as this region encompasses all clusters containing normal reference cells (Figure 3, normal reference indicator), including T cells, B cells, myeloid cells, and fibroblasts. Within this UMAP region, CopyKat and SCEVAN misclassified many cells as aneuploid, whereas InferPloidy did not. These erroneous classifications in CopyKat and SCEVAN (Figure 3a, b) were mostly myeloid cells from the normal condition that were part of the normal reference set (Figure 3d, e, f).

This result may be attributed to either the superior CNV estimation of InferCNV or the superior ploidy inference of InferPloidy. To isolate the contribution of InferPloidy, we applied InferPloidy using the CNV estimates generated by CopyKat and SCEVAN (Supplementary Figure 1). Even in these cases, InferPloidy successfully classified normal cells from the normal condition as diploid, confirming its superiority in ploidy classification.

Notably, the UMAP region corresponding to normal reference cells aligns with clusters exhibiting high sample diversity index values (Figure 3 f and i). Due to the high heterogeneity of tumor cells, their CNV patterns are often unique to individual samples, resulting in a low sample diversity index—i.e., clusters are predominantly occupied by aneuploid cells from a limited number of samples. In contrast, normal cells exhibit cell-type-specific gene expression patterns, and their corresponding CNV estimates remain relatively unchanged across samples, leading to a high sample diversity index—i.e., normal cells from different samples cluster together. Under InferPloidy's classification, we observe that diploid and aneuploid cells can be roughly separated based on their sample diversity index values.

**Quantitative comparison**

We utilized four datasets with available cancer annotations: GSE176078[28], GSE103322[29], GSE72056[30], and GSE131907[27], where, in GSE131907, cancer cell annotations are provided only for advanced tumors; therefore, we excluded samples corresponding to early-stage tumors from our analysis. To show the superiority of InferPloidy, we also performed it with the CNV estimates from CopyKat and SCEVAN.

The results indicated that all three tools performed similarly for GSE103322 and GSE72056. However, for GSE176078 and GSE131907, InferCNV + InferPloidy outperformed CopyKat and SCEVAN, achieving an accuracy of 97.7% and 99.5%, respectively and F1-scores of 98.1% and 99.6, respectively. In contrast, CopyKat and SCEVAN yielded much lower accuracy (< 76% and < 65%, respectively) and F1-scores (< 85% and < 74%, respectively) (Figure 4a, b).

When comparing the proportion of aneuploid, diploid, and 'undecided' cells in each sample, CopyKat and SCEVAN exhibited a considerable fraction of 'undecided' cells (categorized as either not.defined or filtered), whereas this proportion was significantly lower with InferCNV + InferPloidy (Figure 4c, d). As noted previously, in GSE131907, CopyKat and SCEVAN made erroneous classifications and yielded 'undecided' cells in samples from the distant normal condition. In contrast, InferCNV + InferPloidy correctly classified all cells in distant normal samples as diploid (Figure 4d, distant normal).

One possible concern is that InferCNV + InferPloidy might introduce a bias toward classifying cells as diploid. However, this does not appear to be the case, as InferCNV + InferPloidy identified a greater number of aneuploid cells in both datasets compared to the other tools.

**Running time**

One of the most important measures of a software tool is its runtime efficiency in generating final results. While runtime may not be a critical factor if it is relatively short (e.g., on the order of minutes), it becomes a significant concern when the execution time is excessively long. For the down-sampled LUAD dataset from GSE131907, which contains 36,400 cells, CopyKat and SCEVAN required 42 hours and 14 hours, respectively. In contrast, InferCNV and InferPloidy completed the analysis in just 7.5 minutes, respectively—a substantial improvement over CopyKat and SCEVAN.

For a much larger dataset, GSE161529, which comprises 69 samples and 428,000 cells, InferCNV and InferPloidy completed the analysis in less than 2 hours, respectively. In contrast, CopyKat and SCEVAN required 3 days 6 hours and 24 hours, respectively, when each sample was processed separately. In this full dataset, CopyKat and SCEVAN were run one sample at a time, as recommended by the CopyKat developers. However, this approach posed a challenge: some normal samples rarely contained normal reference cells, as immune cells are scarce under some conditions. As a result, CNV estimation tools could not leverage CNV estimates of normal reference cells to accurately infer the

CNVs. To address this issue, we randomly sampled cells across the dataset and processed them together especially to get best results of CopyKat and SCEVAN. In contrast to CopyKat and SCEVAN, we executed InferCNV and InferPloidy without dividing the dataset into individual samples. The results demonstrated that InferPloidy consistently outperformed CopyKat and SCEVAN in ploidy inference, even for the down-sampled dataset.

## Discussion

In this study, we present InferPloidy, a significantly faster and more accurate ploidy inference tool that operates on top of InferCNV, a highly efficient CNV estimation tool that lacks ploidy inference capabilities. The improved accuracy of InferPloidy is achieved through a sophisticated algorithm that employs graph search to initially distinguish two key antagonists: normal and tumor reference cells. By further refining the Gaussian mixture model parameters of these two groups through an iterative refinement process, InferPloidy provides more reliable ploidy estimates, surpassing the performance of simpler hierarchical clustering approaches used in tools like CopyKat and SCEVAN.

Even in scenarios where normal reference cells are absent—such as in the normal condition samples in GSE161529, where normal epithelial cells predominate—InferPloidy accurately infers ploidy by identifying large clusters and assessing the reachability of other clusters from these major groups. Notably, the improved accuracy is not merely due to the use of InferCNV. InferPloidy demonstrated superior performance over CopyKat and SCEVAN, even when applied to CNV estimates derived from those tools, further confirming its effectiveness over existing ploidy inference algorithms.

Regarding computational efficiency, reducing runtime may not always be critical if existing tools are manageable or if the reduction is marginal. However, in the case of ploidy inference, the prolonged runtimes of current tools pose a significant bottleneck compared to other analyses, such as cell-type annotation, cell-cell interaction inference, differentially expressed gene (DEG) analysis, and gene set analysis. For large datasets with multiple samples, these inefficiencies limit scalability, hindering rapid downstream analyses and integration into multi-omics pipelines. In contrast, InferPloidy is nearly two orders of magnitude faster than CopyKat and SCEVAN, making it a more practical and feasible solution for large-scale integrative analyses.

Although InferPloidy does not provide functionality for identifying clonal substructures within a sample, its high-speed and accurate ploidy inference enables the rapid downstream analysis of large datasets. This capability is particularly valuable for studying intertumoral heterogeneity, facilitating cancer subtyping, discovery of druggable targets, and the biological investigation of the tumor micro-environment.

## Method

The detailed processing steps of InferPloidy are as follows

### Cell type annotation and CNV estimation

First, HiCAT[25] was used for cell type annotation. Using the cell-by-gene count matrix, the tool generates cell type annotation result in 3 taxonomy levels, i.e., major, minor, and subset, where major cell types were used to collect T cells, B cells, myeloid cells, and fibroblast. Using these as normal reference cell, InferCNV[20] was used to obtain CNV profile of each cell. hg38 genome was used to run InferCNV.

### Clustering and building cluster adjacency graph

Dimension reduction was performed on the CNV profiles using PCA with 15 components. Then, using PCA-transformed CNV patterns, k-neighbors graph, $\tilde{G}$, was built to perform Louvain clustering[26]. In $\tilde{G}$, each node corresponds to a cell and the edge, $e_{ij}$, represents connection between two cells $(i,j)$, which is 1 if $j$ is one of the $k$ nearest neighbors of $i$ and 0 otherwise. Python packages, scikit-learn and scikit-network, were used. With the cluster labels, we next build the cluster adjacency graph, $G$, where a node corresponds to a cluster and an edge to connectivity between them. The connectivity, $c_{ij}$ between a pair of clusters, $(i,j)$, is defined as follows:

$$c_{ij} = \frac{n_{ij}}{k \cdot \min_{l \in \{i,j\}} |C_l|} \quad (1)$$

where $|C_l|$ is the size (number of cells) of cluster $l$ and $n_{ij}$ is the number of edges in $\tilde{G}$ connecting cluster $i$ and $j$. If a cell in cluster $i$ has a nearest neighbor belonging to cluster $j$, it is counted in $n_{ij}$. With the number $k$ in the denominator, $c_{ij}$ satisfies $0 \leq c_{ij} \leq 1$. The connectivity in (1) represents how close the cluster $i$ and $j$ are to each other.

**Grouping clusters into three: Normal reference, reachable, and unreachable**

First, normal reference clusters are defined as those containing at least 25% of normal reference cells. Then, the following successive procedure is used to divide the rest of clusters into the reachable and the unreachable clusters from normal reference clusters. Let $N$ be the set of normal reference clusters and $R$ be the rest. The clusters in $R$ are ordered by choosing cluster one-by-one as follows.

A. Set $R' = \emptyset$
B. Select cluster $j^* \in R$ such that $j^* = \underset{j \in R}{\operatorname{argmax}}\, s_j$ with $s_j = \sum_{i \in N \cup R'} c_{ij}$,
C. Store the cluster index $j^*$ and its connectivity score $s_{j^*}$, and move $j^*$ from $R$ to $R'$
D. do step B and C successively until no cluster is left in $R$.
E. By inspecting the ordered set of connectivity scores $\{s_m : m = 1,2,\dots,|R|\}$ from the first, find the first cluster $m^*$ of which $s_{m^*} < c_{th}$ (some connectivity threshold)
F. Reachable clusters set is defined as $U = \{j_m : m = 1,2,\dots,m^* - 1\}$ while unreachable as $T = \{j_m : m = m^*, \dots, |R|\}$

Note that replacing the summation in step B to maximization, i.e., $s_j = \max_{i \in N \cup R'} c_{ij}$, the problem is equivalent to order nodes in a graph according to their maximum path cost to any one in $N$ (where a path cost is defined as the minimum connectivity along the path to any one in $N$). In this case, one can resort shortest path search procedures, such as Bellman-Ford algorithm. However, instead of using maximum connectivity, sum of all connectivity's to $N$ was used to minimize the variability due to instability of clustering algorithms, i.e., fluctuations stemming from small variance in datasets.

**CNV profiling of normal and tumor cells using Gaussian mixture model**

Now, our objective is to decide whether normal or tumoral for the cells belonging to reachable and unreachable clusters. Our assumption in doing this is that (1) the cells belonging to clusters in $N$ are certainly normal and (2) the cell in the unreachable group, $T$, are mostly tumoral, even though not all. Based on this assumption, Gaussian mixture model is used to statistically model the CNV of normal and tumor cells. For each genomic spot, the probability density of the two groups is modelled as follows:

$$f_X^{(m)}(x) = \sum_{n=1}^{K} \pi_{X,n}^{(m)} \cdot \mathcal{N}\left(x; \mu_{X,n}^{(m)}, \sigma_{X,n}^{(m)}\right) \quad (2)$$

Where $X$ is the group variable, either $N$ (normal) or $T$ (tumor), $K$ is the number of Gaussian components, $\mathcal{N}(x; \mu, \sigma)$ is the normal density function of mean $\mu$ and standard deviation $\sigma$, and $\pi_{X,n}^{(m)}$, $\mu_{X,n}^{(m)}$ and $\sigma_{X,n}^{(m)}$ are the weight, mean and standard deviation of the $n^{\text{th}}$ component for the $m^{\text{th}}$ genomic position of the group $X$. Given the CNV estimates of cells in the normal reference and tumoral group, the model parameters can be estimated by using the expectation maximization (EM) algorithm. Specifically, GaussianMixture from scikit-learn package was used to obtain the model parameters, $\pi_{X,n}^{(m)}$, $\mu_{X,n}^{(m)}$ and $\sigma_{X,n}^{(m)}$.

**Decision metric**

Given the model parameters of the tumor and normal CNV, our decision metric is defined as the number of genomic spots where the likelihood of the tumor is greater than or equal to the likelihood of the normal, which can be represented as

$$T(\boldsymbol{x}) = \sum_{m=1}^{L} \log f_T^{(m)}(x_m) - \log f_N^{(m)}(x_m) \quad (3)$$

where $L$ is the number of genomic positions where CNV's were evaluated, the vector $\boldsymbol{x} = [x_1, x_2, \dots, x_m, \dots, x_L]$ is the CNV estimates of a cell with $x_m$ being the CNV at the $m^{\text{th}}$ genomic position, $I(A)$ is the indicator function returning 1 if $A$ is true or 0 otherwise. In this work, $L$ was set to 2. The rationale of using Gaussian mixture model, instead of using single-component model, is that, due to

high heterogeneity of CNV's of tumor cells, it is reasonable to assume that there exist multiple states in CNV at each genomic spot. This applies to normal cells too since there exist different normal cell types and some cell types sometime shows distinct CNV patterns at some specific spots stemming from its unique gene expressions.

Note that the model parameters were obtained based on the initial classification of normal cells and tumor cells belonging to the clusters in $N$ and $T$, respectively. Therefore, it would be questionable if the statistical models that were obtained using the cells from $N$ and $T$, respectively, can be used to successfully distinguish normal and tumoral cells from the clusters in another group, $U$, that certainly have different CNV patterns as they were clustered separately from those in $N$ and $T$. They might consist not only of other normal cell types than those we used as normal reference, e.g., endothelial cells or normal epithelial cells, but also tumor cell that has different CNV patterns from those of the clusters in $T$. Therefore, instead of making immediate decisions, an iterative detection and re-estimation approach was devised.

**A cautious approach to make final decision: iterative refinement**

Define $\tilde{N}$, $\tilde{T}$, and $\tilde{U}$ as the sets of cells of normal, tumor, and not-determined-yet, respectively. Initially, they are set as the set of cells that belong to the clusters in $N$, $T$, and $U$, respectively. Let $p_N$ and $p_T$ be the set of model parameters of the Gaussian mixture model in (2), i.e., $p_X = \left\{\pi_{X,n}^{(m)}, \mu_{X,n}^{(m)}, \sigma_{X,n}^{(m)}\ \forall m, n\right\}$. They are updated recursively as follows:

A. Use the three set of clusters $N$, $T$, and $U$, to find $\tilde{N}$, $\tilde{T}$, and $\tilde{U}$, respectively.
B. Compute $p_N$ and $p_T$ using the CNV's of the cells in $\tilde{N}$ and $\tilde{T}$, respectively.
C. Use the parameters, $p_N$ and $p_T$, to compute the decision metric $t_j = T(\boldsymbol{x}_j)$ for all cells.
D. Compute the mean and the standard deviation of $t_j$ for the normal cells in $\tilde{N}$, $(\mu_N, \sigma_N)$, and those for tumor cells in $\tilde{T}$, $(\mu_T, \sigma_T)$.
E. Two decision thresholds are defined: $th_{lower} = \mu_N + \alpha \cdot \sigma_N$ and $th_{upper} = \mu_T - \alpha \cdot \sigma_T$, where $\alpha$ controls the decision margin and set as $\alpha = \beta \cdot \frac{|\mu_T - \mu_N|}{\sigma_N + \sigma_T}$ with $0 < \beta < 1$.
F. Given the upper and lower threshold, the cells with $t_j < th_{lower}$ are decided as normal, those with $t_j > th_{upper}$ as tumor, and others are set 'not-determined-yet'. These decisions are applied only to those cells in $\tilde{T}$ and $\tilde{U}$.
G. Update $\tilde{N}$, $\tilde{T}$, and $\tilde{U}$ as follows:
   i) $\tilde{N} \leftarrow \tilde{N} \cup \{j: t_j < th_{lower}\ \forall j \in \tilde{T} \cup \tilde{U}\}$,
   ii) $\tilde{T} \leftarrow \{j: t_j > th_{upper}\ \forall j \in \tilde{T} \cup \tilde{U}\}$
   iii) $\tilde{U} \leftarrow \{j: th_{lower} \leq t_j \leq th_{upper}\ \forall j \in \tilde{T} \cup \tilde{U}\}$
H. Based on the updated $\tilde{N}$ and the cluster labels previously assigned to each cell, the set of clusters $N$, $T$, and $U$ is renewed, i.e., the clusters that has cells at least 25% or more from (the updated) $\tilde{N}$ is marked as normal reference cluster, those reachable from (the updated) $N$ are grouped into $U$, and those unreachable into $T$. (In this step, only cluster grouping is changed without altering the cell's cluster membership and the reachability between clusters.)
I. The steps A to H are repeated until there are no changes in $N$, $T$, and $U$ or maximum number of iterations is reached.

Note that the number of normal cells in $\tilde{N}$ is non-deceasing, while others can be both increasing or decreasing. Due to this reason, the procedure can be ended up with all cells decided as normal, if the initial grouping is biased more or less to normal. Therefore, the initial setting of $\tilde{N}$, $\tilde{T}$, and $\tilde{U}$ is an important issue to obtain reasonable results from the designed procedure.

**Issue in hyper parameter setting**

Since the initial grouping of cells into $\tilde{N}$, $\tilde{T}$, and $\tilde{U}$, relies on the initial grouping of clusters into $N$, $T$, and $U$, through graph search, it is important to appropriately set the connectivity threshold, $c_{th}$, for initial setting of normal reference clusters. Our background inspection on many datasets suggests $c_{th} = 0.18$ is a good choice for most of the datasets, while some datasets require higher threshold, specifically for better true negatives, i.e., in detecting true normal cells as normal. In this work, the threshold was

set based on the statistics on the connectivity's among clusters in $N$ such that $c_{th} = \max(0.18,\ \mu_C - 1.5\sigma_C)$ with $\mu_C$ is the mean and the standard deviation of the connectivity's among clusters in $N$.

Another thing to note is that the final result can be varied at different clustering run even with the same hyper-parameters. It stems from random initialization of clustering algorithms. Although one can set random seed to a fixed value to get the same result for the same data, it is questionable if the results can be considered as a representative of an ensemble of experiments. This is a critical problem not only in our approach but also in many fields of research that utilizes automatic clustering in anyway. To avoid such variations and to improve reproducibility, the procedure has been performed many times under different clustering run and the one that has the separability index $\frac{|\mu_T - \mu_N|}{\sigma_N + \sigma_T}$ is the smallest was selected for the final result. The rationale of selecting the one with the smallest separability index is that it tends to be more balanced between normal and tumor.

Finally, setting the hyper parameter $\beta$ that controls the decision margin, should also be set appropriately. If it is too high, many cells are decided as not-determined-yet, while, if it is too small, the final classification results can be imbalanced, i.e., ended up with too many normal cells or tumor cells. In this work, it was initially set to 0.4 and reduced by 0.1 at each loop (of step A to H below) until it reaches 0.2.

### Sample diversity index

The sample diversity index is computed for each cluster using the Shannon's entropy as follows

$$SDI_c = -\sum_{k \in S} p_k \log_2 p_k$$

where $S$ is the set of sample index and $p_k$ is the proportion of cells from the $k^{\text{th}}$ sample in the cluster $c$. It is the same as the Shannon diversity index used in microbiome research[31]. From the property of the entropy, $0 \leq SDI_c \leq |S|$. It is zero if a cluster consists of cells from only one sample while close to $|S|$ if a cluster is mixed with cells from many samples.

### Running CopyKat and SCEVAN

The CopyKat and SCEVAN were run with the respective default hyper parameters, except for the normal reference indications. In all tools, including InferCNV and InferPloidy, we have used T cells, B cells, myeloid cells, and fibroblasts as normal reference cells.

## Declaration of Competing Interest

The authors declare that they have no competing interest.

## Authors' contribution

SY and JM designed the overall experiments and devised the key idea. SY developed InferPloidy. WS and JC performed single-cell RNA-seq data analysis. SY and JM guided the analysis and gave feedback on the results. All authors wrote, read, and approved the final manuscript.

## Acknowledgements

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2024-00437102) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation)

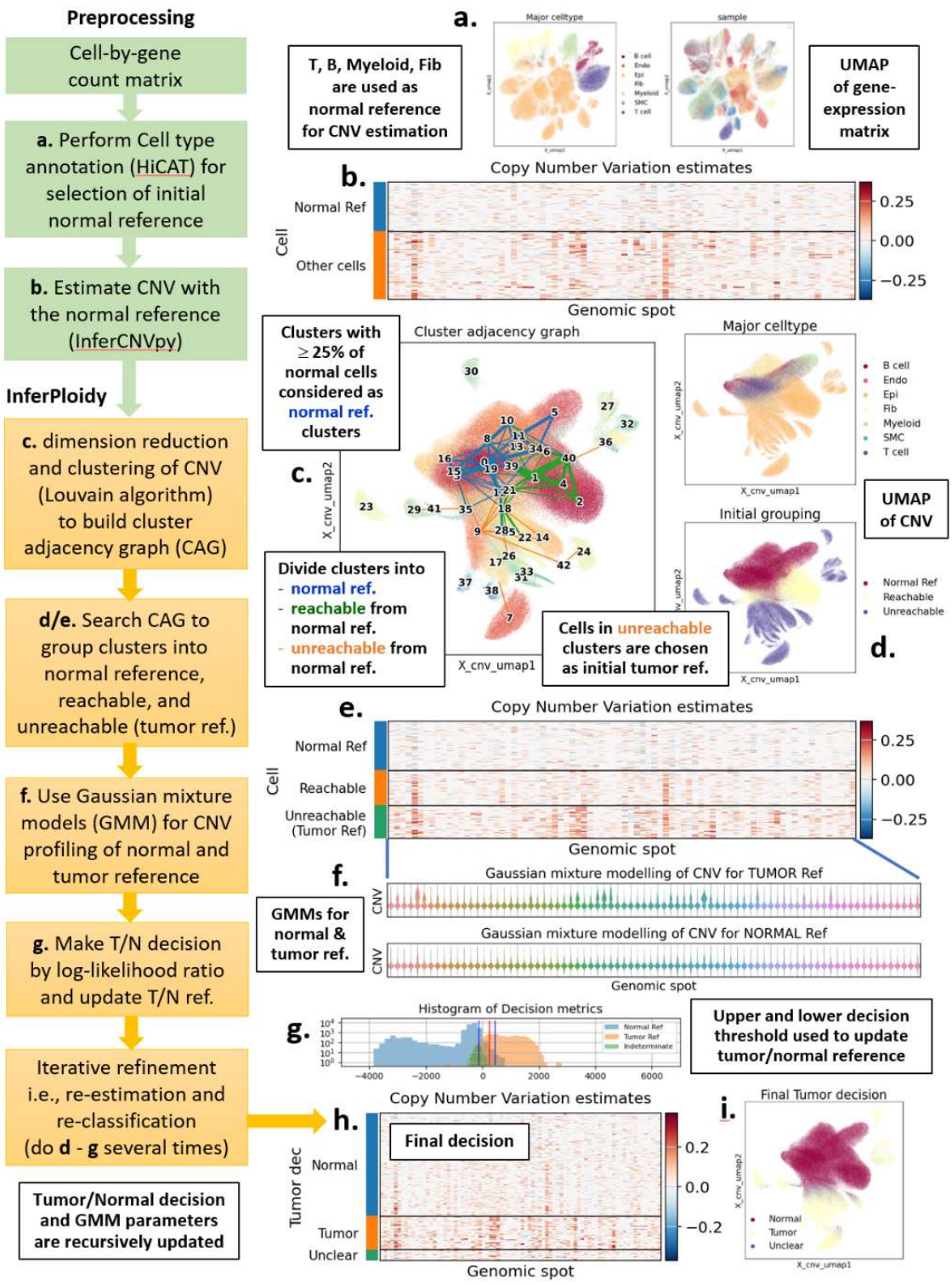


**Figure 1** A brief description of InferPloidy processing steps. **a**, **b:** It starts with CNV estimation using InferCNV, for which HiCAT was used for annotation of cell types to collect normal reference cells. **C:** With the CNV estimates, dimension reduction using PCA and Louvain clustering are performed to find cluster adjacency graph. **d**, **e:** By searching the graph, clusters are divided into three groups, i.e., normal reference clusters, clusters reachable from normal reference, and those unreachable, where the cells in the unreachable clusters are initially used as tumor reference cells. **f:** The CNV patterns of normal and tumor reference cells are modelled by using Gaussian mixture model with diagonal covariance constraint. **g:** The initial ploidy decision is made by the log-likelihood ratio test. Step **d** to **g** is repeated until no changes in ploidy decision is found or the number of repetitions reaches the predefined number.

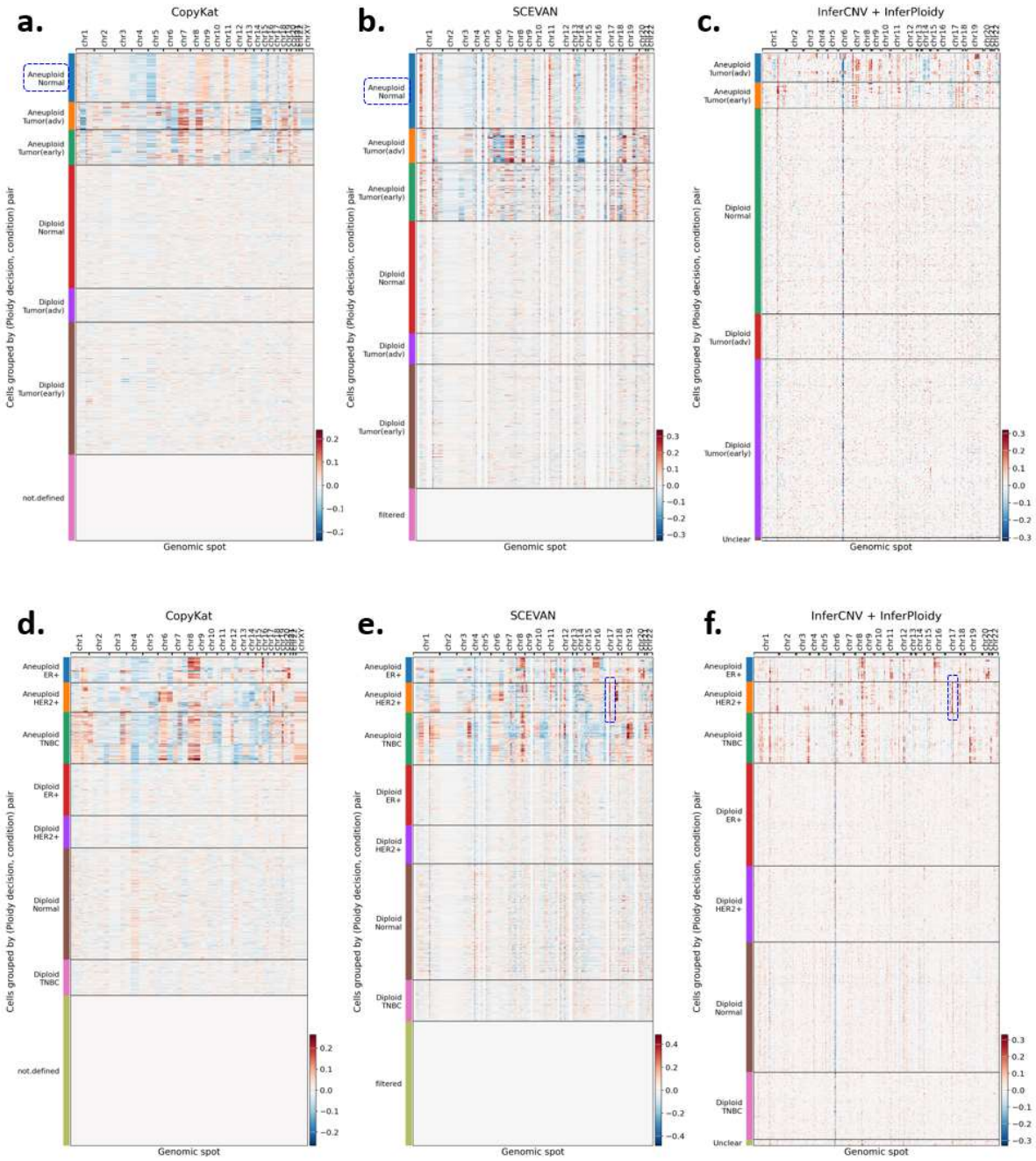


**Figure 2** A comparison of CNV estimates and ploidy decisions for lung cancer dataset GSE131907*, which includes three conditions: distant normal, early tumor, and advanced tumor (**a-c**), and for breast cancer dataset GSE176078*, which includes three conditions: ER+, HER2+, and TNBC (**d-f**). **a,d**: CopyKat, **b,e**: SCEVAN, and **c,f**: InferCNV + InferPloidy. For GSE131907*, many cells were erroneously classified as aneuploid by CopyKat and SCEVAN, whereas no such errors were observed with InferCNV + InferPloidy. In GSE176078*, a copy number gain in the genomic region on chromosome 17, where the ERBB2 gene is located, was identified in the HER2+ condition with SCEVAN and InferCNV (blue boxes in **e** and **f**). In both datasets, a considerable portion of cells were classified as indeterminate with CopyKat and SCEVAN.

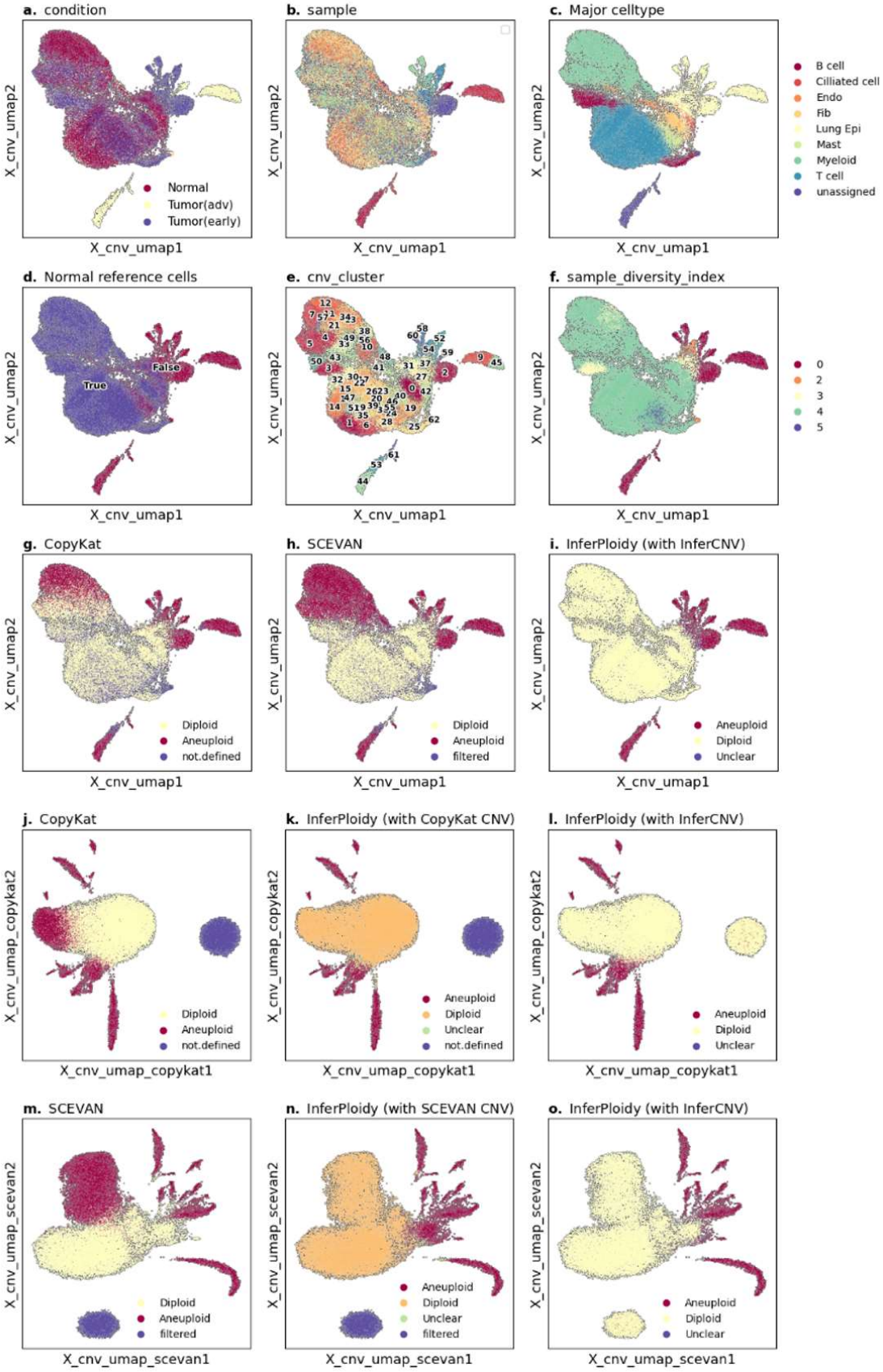


**Figure 3** A summary on UMAP projection based on the CNVs obtained by InferCNV (**a-i**), CopyKat (**j-l**), and SCEVAN (**m-o**) for GSE131907*. InferPloidy (with SCEVAN) and InferPloidy (with CopyKat) are the results of InferPloidy using the CNVs obtained by SCEVAN and CopyKat, respectively.

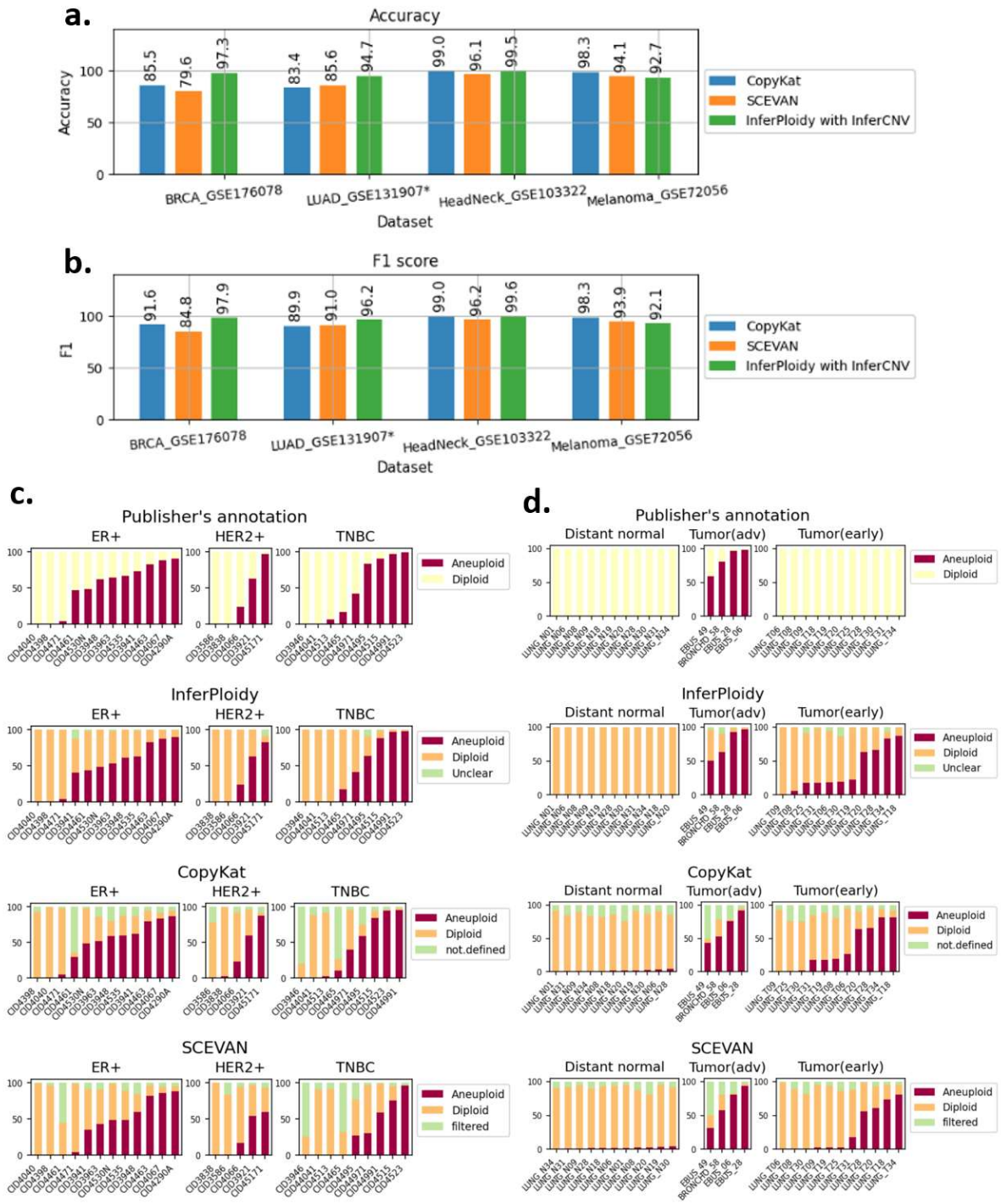


**Figure 4** A comparison of accuracies and F1-scores of the ploidy decisions for the datasets with publisher's ploidy annotations (**a, b**) and a summary of the ploidy decisions per sample for GSE176078* and GSE131907*. The normal reference cells were excluded in the performance evaluation and the summary of per-sample ploidy decisions.

## Reference


1. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-382 (2009).
2. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10**, 1096-1098 (2013).
3. Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C. & Teichmann, S.A. The technology and biology of single-cell RNA sequencing. *Mol Cell* **58**, 610-620 (2015).
4. Sun, G. *et al.* Single-cell RNA sequencing in cancer: Applications, advances, and emerging challenges. *Mol Ther Oncolytics* **21**, 183-206 (2021).
5. Balkwill, F.R., Capasso, M. & Hagemann, T. The tumor microenvironment at a glance. *J Cell Sci* **125**, 5591-5596 (2012).
6. Suva, M.L. & Tirosh, I. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. *Mol Cell* **75**, 7-12 (2019).
7. Chung, W. *et al.* Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat Commun* **8**, 15081 (2017).
8. Liu, J., Dang, H. & Wang, X.W. The significance of intertumor and intratumor heterogeneity in liver cancer. *Exp Mol Med* **50**, e416 (2018).
9. Chen, S., Zhou, Z., Li, Y., Du, Y. & Chen, G. Application of single-cell sequencing to the research of tumor microenvironment. *Front Immunol* **14**, 1285540 (2023).
10. Yeo, A.T. *et al.* Single-cell RNA sequencing reveals evolution of immune landscape during glioblastoma progression. *Nat Immunol* **23**, 971-984 (2022).
11. 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).
12. Zhang, Y. *et al.* Single-cell RNA sequencing in cancer research. *J Exp Clin Cancer Res* **40**, 81 (2021).
13. Zhao, H. *et al.* Single-cell RNA-seq highlights a specific carcinoembryonic cluster in ovarian cancer. *Cell Death Dis* **12**, 1082 (2021).
14. Wu, F. *et al.* Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nat Commun* **12**, 2540 (2021).
15. Nath, A. & Bild, A.H. Leveraging Single-Cell Approaches in Cancer Precision Medicine. *Trends Cancer* **7**, 359-372 (2021).
16. Aissa, A.F. *et al.* Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nat Commun* **12**, 1628 (2021).
17. The, S. *et al.* Integration of Single-Cell RNA-Sequencing and Network Analysis to Investigate Mechanisms of Drug Resistance. *Methods Mol Biol* **2660**, 85-94 (2023).
18. Gao, R. *et al.* Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat Biotechnol* **39**, 599-608 (2021).
19. De Falco, A., Caruso, F., Su, X.D., Iavarone, A. & Ceccarelli, M. A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data. *Nat Commun* **14**, 1074 (2023).
20. Patel, A.P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401 (2014).
21. Mahdipour-Shirayeh, A., Erdmann, N., Leung-Hagesteijn, C. & Tiedemann, R.E. sciCNV: high-throughput paired profiling of transcriptomes and DNA copy number variations at single-cell resolution. *Briefings in Bioinformatics* **23** (2022).
22. Trinh, M.K. *et al.* Precise identification of cancer cells from allelic imbalances in single cell transcriptomes. *Commun Biol* **5**, 884 (2022).
23. Zhong, Z. *et al.* Domain generalization enables general cancer cell annotation in single-cell and spatial transcriptomics. *Nat Commun* **15**, 1929 (2024).
24. Takeuchi, F. & Kato, N. Ploidy inference from single-cell data: application to human and mouse cell atlases. *Genetics* **227** (2024).
25. Lee, J., Kim, M., Kang, K., Yang, C.S. & Yoon, S. Hierarchical cell-type identifier accurately distinguishes immune-cell subtypes enabling precise profiling of tissue microenvironment with single-cell RNA-sequencing. *Brief Bioinform* **24** (2023).
26. Blondel, V.D., Guillaume, J.L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J Stat Mech-Theory E* (2008).
27. Kim, N. *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* **11**, 2285 (2020).
28. Wu, S.Z. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet* **53**, 1334-1347 (2021).
29. Puram, S.V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624 e1624 (2017).
30. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196 (2016).
31. Reese, A.T. & Dunn, R.R. Drivers of Microbiome Biodiversity: A Review of General Rules, Feces, and Ignorance. *mBio* **9** (2018).