Benchmarking copy number aberrations inference tools using single-cell multi-omics datasets

Minfang Song[1,2,4], Shuai Ma[2,4], Zhenzhen Yang[3], Bin Xie[1], Tongkun Guo[1], Xingxu Huang[1], Liye Zhang[2,5*]

Affiliations:

[1] Research Center for Life Sciences Computing, Zhejiang Lab, Hangzhou, Zhejiang 311121, China
[2] School of Life Science and Technology, ShanghaiTech University, Shanghai 200032, China; Shanghai Clinical Research and Trial Center, 201210, Shanghai, People's Republic of China
[3] Shanghai Institute for Advanced Immunochemical Studies, School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China
[4] These authors contributed equally
[5] Lead Contact
[*]Correspondence: zhangly@shanghaitech.edu.cn

Abstract
Background
Copy number aberrations (CNAs) are an important type of genomic variation which play a crucial role in the initiation and progression of cancer. With the explosion of single-cell RNA sequencing (scRNA-seq), several computational methods have been developed to infer CNAs from scRNA-seq studies. However, to date, no independent studies have comprehensively benchmarked their performance.

Results
Herein, we evaluated five state-of-the-art methods based on their performance in tumor vs normal cell classification, CNAs profile accuracy, tumor subclone inference and aneuploidy identification in non-malignant cells. Our results showed that Numbat outperformed others across most evaluation criteria, while CopyKAT excelled in scenarios when expression matrix alone was used as input. Additionally, we investigated how referencing settings, inclusion of tumor microenvironment cells, tumor type, and tumor purity impact the performance of these tools.

Conclusions
In summary, our study evaluated five state-of-the-art methods and found that Numbat outperformed others across most evaluation criteria. This study provides a valuable guideline for researchers to select and use the methods appropriately for their datasets.

Keywords
scRNA-seq, copy number aberrations, copy number alterations, copy number variations, inference

Background
Single-cell RNA sequencing (scRNA-seq) is a powerful and popular technology that provides a comprehensive view of the transcriptome of individual cells with high throughput and relatively low cost(1-3). It has been successfully used to dissect the heterogeneous cellular tumor microenvironments, providing important insights into the biological functions of different cell subpopulations and cell-cell interactions, which are valuable for understanding cancer progression and therapy resistance(4, 5). In addition to measuring gene expression levels, scRNA-seq can also reflect genomic aberrations such as copy number alterations (CNAs)(6-9). Characterization of CNAs in tumor cells has significant implications for early tumor detection, delineating tumor heterogeneity, understanding modes of progression, and uncovering mechanisms of therapy resistance(10). Single-cell DNA sequencing (scDNA-seq) is ideal for understanding CNAs at single-cell resolution (11-13). However, the high cost and low coverage of scDNA-seq limit its application in estimation of copy number variations(14, 15). To address this, several state-of the-art computational methods have been developed to infer CNAs from scRNA-seq data, enabling the delineation of intratumoral heterogeneity(16-20).

Current computational methods for inferring CNAs from scRNA-seq data can be broadly classified into two categories. One category is expression-based, where CNAs are inferred directly from gene expression. Examples of such tools include inferCNV(16), CopyKAT(17), and SCEVAN(18). These tools operate on the principle that copy number amplifications (AMPs) or deletions (DELs) will lead to the upregulation or downregulation of genes within the affected genomic regions, respectively. InferCNV utilizes a corrected moving average over gene windows; CopyKAT employs an integrative Bayesian segmentation approach; SCEVAN uses a multi-channel segmentation algorithm. The second category combines gene expression with allele information, with relevant tools such as Numbat(19) and CaSpER(20). Numbat employs a haplotype-aware Hidden Markov Model (HMM) and integrates signals form gene expression, allelic ratio derived from scRNA-seq, and population-derived haplotype to infer allele-specific CNAs. CaSpER utilizes a multiscale signal-processing framework that integrates gene expression and allelic shift signal profiles. Due to the differing principles, parameters, and use scenarios of these methods, it is crucial to evaluate their performance comprehensively.

In this study, we employed single-cell multi-omics datasets that enable simultaneous interrogation of DNA and RNA within the same cell. By using CNAs identified from scDNA-seq as the ground truth, we evaluated the accuracy of CNAs inference tools from scRNA-seq. Our comprehensive performance evaluation encompassed five popular tools: inferCNV, CopyKAT,

SCEVAN, Numbat and CaSpER, from multiple perspectives. We examined the impact of reference settings, inclusion of tumor microenvironment (TME) cells, tumor type, and tumor purity on CNAs inference. Furthermore, we assessed the accuracy of tumor and normal cell classification, CNAs events accuracy, tumor subclone inference, and identification of aneuploid fibroblast cells.

Our analyses suggest that Numbat demonstrated the best performance among the evaluated tools. In cases where only expression matrix is available, CopyKAT is recommended. In addition, our comprehensive study revealed common and specific pitfalls and their potential solutions. Therefore, we believe our benchmark study will serve as a valuable guidance for performing CNAs inference analyses from both single cell RNA-seq and spatial transcriptomics(21).

Results

Benchmarking framework and datasets examined

The benchmark dataset utilized in this study was collected from published studies that involved single cell multi-omics sequencing data, specifically paired scRNA-Seq and scDNA-Seq performed on the same cell. The CNAs inference obtained from scDNA-seq served as the ground truth. Due to the limited number of studies with paired scRNA-seq and scDNA-seq data, we selected samples with a sufficient number of cells from different sources. Specifically, we included eight colorectal cancer (CRC)(22) samples, two acute lymphoblastic leukemia (ALL)(23), one glioma and NPC43 cell line, along with HUVEC cell line samples(24) (**Table S1**).

Five popular CNAs detection tools were benchmarked in this study. These tools can be broadly categorized into two major groups: those that solely require an expression matrix (inferCNV, CopyKAT, and SCEVAN), and those that require both expression matrix and B-allele frequency data (Numbat and CaSpER) **(Fig1**). The performance was assessed by three common applications in tumor cells: (1) classification of tumor and normal cells, (2) accuracy of CNAs events, (3) tumor subclone inference. Additionally, accumulating studies suggest that aneuploidy can also be present in normal cells, and it has been associated with various non-cancer diseases. Several publications have reported clonal expansion of CNAs in benign tissues(21, 25). Furthermore, Zhou et al. identified prevalent CNAs in immune cells, endothelial cells, and fibroblasts within the TME of CRC(22). Therefore, we also evaluated the performance of aneuploidy detection in normal cells.

Distinguishing tumor cells vs. normal cells

To distinguish tumor cells and normal cells based on inferred CNAs profiles, we first applied the aforementioned tools to solid tumors including eight individuals with CRC and one patient with glioma. Cell type annotations for all cells were obtained from corresponding published studies(22, 24). Only epithelial cells for CRC and glial cells for glioma were retained for tumor vs normal cell classification (**Figure 2A**). Overall, Numbat performed the best in distinguishing tumor cells from normal cells (**Figure 2B, Figure S1 and Table S2**). Among the three tools that solely utilize the expression matrix, CopyKAT achieved the overall best performance. SCEVAN exhibited poor performance in samples with a low frequency of tumor cells (less than 40%), such as CRC11, CRC12, CRC16, and CRC15 (**Figure 2B**). InferCNV was reliable except in CRC02, where the majority of cells are tumors (approximately 88%). In CRC02, the copy number gains and losses in tumors cells were incorrectly centered to be the copy number baseline by inferCNV, while normal cells were predicted to harbor the opposite CNAs profile and thus incorrectly classified to be tumor cells based on global copy number alteration levels (**Figure 2C, Figure S1F**). CaSpER also made the same error **(Figure S1F)**. To address this issue, we investigated whether including TME cells (such as immune, endothelial, and fibroblast cells) could improve the performance of inferCNV, CopyKAT, and SCEVAN, in cases with imbalanced tumor vs normal ratios (**Figure 2D, Figure S2A-C**). The inclusion of TME cells significantly improved the accuracy of tumor cell prediction for SCEVAN (**Figure2D, Figure S2C**). This improvement was reasonable because the SCEVAN algorithm incorporates a set of gene signatures from public collections, including cells from the tumor microenvironment such as stromal, immune cells, to identify high-confidence normal cells. Therefore, the addition of TME cells enhanced the ability to distinguish between tumor and normal cells for SCEVAN. Although this overall improvement was not observed in inferCNV, certain samples with a higher number of tumor cells, such as CRC02 (88%), CRC22 (91%), and CRC03 (67%), showed dramatic increases in their F1 scores, from 0, 0.5, and 0.55 to 1, 0.9, and 0.99, respectively (**Figure2D, Figure S2A**). For instance, incorporating TME cells into the CRC02 sample not only improved the ability to distinguish between tumor and normal cells, but also

accurately identified copy number alterations, including copy number amplifications on chr8 and chr20, and copy number deletions on chr15 (**Figure 2E**). This suggests that for inferCNV, when tumor purity is higher, including TME cells can enable more accurate estimation of expression level baselines, thus improving both tumor cell assignment and CNAs profile inference performance. This preliminary conclusion was further supported by the excellent performance of inferCNV in patient CRC15 where the tumor purity was exceptionally low (6%) (**Figure 2F**). To systematically explore the impact of tumor purity on the accuracy of CNA inference, we generated a synthetic dataset spanning a broad range of tumor/normal cell ratios (from 1:100 to 100:1) to examine the influence on the performance using the largest glioma dataset (**Figure 2G**). Numbat consistently outperformed the other tools, even in extreme cases with low tumor/normal cell ratios (**Figure 2H**). Notably, inferCNV exhibited opposite calling between tumor and normal cells when tumor purity was higher, consistent with the results observed in CRC02 (**Figure S3**). To conclude, Numbat demonstrates the best performance when B allele frequency is available, while CopyKAT is recommended for cases where only the expression matrix is accessible.

Accuracy of inferred CNAs profiles

To evaluate the accuracy of inferred CNAs profiles, it was imperative to ensure optimal performance of each software tool. In the process of inferring CNAs from scRNA-seq data, the designation of normal reference cells has a significant impact on the final CNAs results. Numbat and CaSpER software required the input of a reference, preferably derived from the same biological and technical conditions (**See method for more details**). For InferCNV, CopyKAT, and SCEVAN, the use of reference cells is optional. Given the potential problem of incorrect centering in inferCNV (as shown in Figure 2C) when using one-pass mode (single-run without specifying reference cells, **Figure 3A**), we hypothesized that employing a two-pass mode (first-pass run to identify normal cells and second-pass run using these normal cells as reference, **Figure 3A**) could improve the accuracy of CNAs profiles by facilitating baseline estimation. The similarity between CNAs profiles calling form scDNA-Seq and inferred CNAs profiles from scRNA-seq significantly increased in eight of nine cases for inferCNV (**Figure 3B-C**). In contrast, CopyKAT and SCEVAN already incorporated a two pass-like approach in their algorithms where they initially predict normal reference cells and then utilize them as a baseline to correct tumor cell CNAs(17, 18). Thus, adopting a two-pass approach has minimal impact on the performance of these two methods (**Figure S4A-B**). Consequently, when assessing the accuracy of inferred CNAs profiles for inferCNV, the results obtained from the two-pass mode were utilized.

Numbat and CaSpER generated integral copy number profiles, which are easier to interpret and appear cleaner, while inferCNV, CopyKAT and SCEVAN output continuous copy number. No single tool consistently outperformed others in term of consistency between scRNA-Seq inferred profile and scDNA-Seq profile in all cases (**Figure 3D-E**). All tools performed well in cases with sufficient number of both tumor and normal cells, such as glioma, CRC13, CRC11 and CRC12. In samples with an imbalanced number of tumor and normal cells, such as CRC02, CRC22 and CRC15, the overall accuracy of CNAs profiles predicted by the five tools decreased considerably. Notably, in CRC03, although Numbat correctly assigned tumor cells, it accurately identified only a small fraction of deletion CNAs in high CNAs burden tumor cells, resulting in low similarity with CNAs profiles obtained from scDNA-Seq (**Figure 3D, Figure S1I**). We also evaluated the performance on hematopoietic cancer acute lymphoblastic leukemia (ALL). However, due to the unavailability of raw fastq data, we could only evaluate the performance of inferCNV, CopyKAT and SCEVAN (**Figure S5**). Among them, CopyKAT achieved the best performance, although noticeable false positive signals on CNAs were observed for both CopyKAT and SCEVAN (**Figure S5C, S5F**). Furthermore, we examined the performance of these tools on cancer cell lines. All four tools except for Numbat performed poorly when applied to cancer cell line alone (**Figure S6**). The inclusion of normal cells improved the performance of these four tools, but not Numbat (**Figure S7**). Taken together, the CNAs profiles inferred from scRNA-Seq are overall reasonable, but still a large fraction of CNAs were undetected by scRNA-Seq based methods.

Evaluation of subclonal structure inference accuracy

Human cancer exhibits extensive intra-tumoral heterogeneity, constantly evolving through emerging mutations and CNAs in different subclones, which impact their phenotypes and confer fitness advantages(26). Both subclones and the tumor evolutionary history can be inferred from

single-cell copy number profiles(27). To evaluate the accuracy of subclonal structure inference by these tools, we selected three tumors (glioma, CRC11 and CRC03) with clear subclonal structure from scDNA-Seq data (**Figure S1A, S1C, S1I**).

In the case of glioma, only Numbat correctly assigned the smaller C1 subclone (correctly assigning 14 out of 15 cells) with a lower CNAs burden as tumor cells (**Figure 4A-B**). Therefore, we tested the glioma dataset's subclonal structure inference only using Numbat. We found that the subclone structure inferred by Numbat was largely consistent with scDNA-Seq-based results (**Figure 4C**). In the case of CRC03, a similar tumor calling error occurred only with inferCNV (**Figure 4D**). InferCNV incorrectly assigned the tumor C1 subclone as normal cells, thus failing to detect the two subclones correctly. Adding TME cells to inferCNV corrected the misclassification and improved subclonal structure inference (**Figure S8 A-C**). Most of the other four tools performed well in assigning subclones (**Figure 4 E-F**), although the CNAs profiles for subclones from scDNA-Seq were not fully recapitulated.

In the case of CRC11, where all tools correctly classified tumor cells, all tools achieved favorable performance (ARI >0.8 for all tools, **Figure 4G-I**). Notably, Numbat sometimes misclassified the deleted regions into Copy-Neutral Loss of Heterozygosity (cnLOH) (**Figure 4D and 4G**). To conclude, given the premise of correct tumor cell classification, all five tools achieve desirable performance in delineating the subclonal structures, although a significant fraction of subclonal specific CNAs were undetected.

Aneuploidy inference in normal cells

Accumulating evidence suggest that CNAs can also be found in non-malignant cells and are postulated to exert a crucial impact on cellular fitness(21, 25). Therefore, we also evaluated the performance of these tools in detecting CNAs in non-malignant cells. Unlike tumor cells with pervasive CNAs across the whole genome, most CNAs observed in non-malignant cells are single whole chromosome gain or loss. We focused on the fibroblast cell type, which exhibits the most frequent autosomal aneuploidy events(22). However, the performance of all tools in identifying single-chromosome aneuploidy was generally poor (**Figure S9**). This highlights the importance of developing more sensitive methods for detecting low-burden CNAs in non-malignant cells.

Computational speed performance

To evaluate the computation speed of various methods, we tracked the CPU time used (**Table S3**). Our findings indicated that CopyKAT and SCEVAN exhibited the best performance in terms of runtime. InferCNV and CaSpER demonstrated moderate computation speeds, whereas Numbat required the most CPU time. Additionally, we observed a slight increase in runtime with larger datasets. However, the runtime remained acceptable for every tool tested. Regarding memory, personal computers typically met the requirements necessary to run these software applications, so the assessment of memory usage was not conducted.

Discussion

In this study, we benchmarked five methods used to infer CNAs profiles from scRNA-seq data. Overall, we observed that all five tools could distinguish between tumor and normal cells and accurately infer CNAs profiles to a certain extent. Among these methods, Numbat demonstrated the highest performance across various evaluation criteria. However, we observed a minor issue with Numbat, where it occasionally predicted copy number deletion events as cnLOH (**Figure 4D, 4G, and Figure S1**). For case where only the expression matrix was available, CopyKAT is recommended as the preferred method. Additionally, we have included a comprehensive and informative table (**Figure 5**) that summarizes the characteristics of all the benchmarked methods.

Methods based on gene expression are more susceptible to the influence of tumor purity. Without a reference setting, inferCNV relies on the input cell population to generate a background. When tumor purity is high, actual CNAs signals may be incorrectly considered as background and removed. Consequently, inferCNV occasionally exhibited incorrect centering of CNAs, leading to mis-classification of normal and tumor cells in our benchmark (**Figure 2C and S3**). On the contrary, SCEVAN tended to incorrectly assign non-malignant cells as malignant cells when tumor purity was low. Interestingly, including TME cells in the analysis improved performance in both cases (**Figure 2E, Figure S2C**).

Another observation from our comparative analysis was that the complexity and burden of

CNAs affected each method equally. We found that tools generally performed better on solid tumors than on hematopoietic cancers or cell lines. This could be due to the fewer copy number alterations and lower chromosomal CNAs heterogeneity in hematopoietic cancers. Additionally, the methods performed poorly in identifying aneuploidy in nonmalignant cells such as fibroblasts. This underscored the need to develop more sensitive methods for inferring low-burden CNAs from scRNA-seq to address this gap in the field.

Furthermore, we found that Numbat is more prone to generating false positive results for cnLOH, meaning that when CNA deletions occur in cells, they might be identified as cnLOH by Numbat. This issue may stem from insufficient sequencing depth or inaccurate SNP density in the regions of these CNAs deletions, which hinders Numbat's ability to accurately distinguish between CNAs deletions and true cnLOH. In contrast, other software tends to produce false negatives more frequently; these tools often fail to detect CNAs that are actually present in cells. This problem is largely attributed to the reliance of these methods on gene expression.

One limitation of our benchmark study was the limited availability of single cell multi-omics datasets that simultaneously measured both DNA and RNA in the same cells. Therefore, the number of input cells for CNAs inference in these datasets is relatively small compared to microfluidics-based scRNA-Seq approaches such as 10X Genomics platform. Therefore, we were unable to evaluate how different method perform for large and more complex datasets. In addition, the small number of cells also limits our evaluation of subclonal structure inference.

In summary, this study provides a much-needed independent comparison of CNAs inference tools. Our results will help researchers gain a deeper understanding of the performance and limitations of current methodologies, enabling them to choose and use these tools appropriately. Finally, we believe that many of the lessons learned in this benchmark study may also apply to CNAs inference in spatial transcriptomics.

Conclusions

Collectively, We assessed five state-of-the-art methods and found that Numbat generally outperformed the others across most evaluation criteria. However, CopyKAT demonstrated superior performance specifically when using expression matrices alone as input. Additionally, we examined how various factors such as reference settings, inclusion of tumor microenvironment cells, tumor type, and tumor purity influenced the tools' performance.

Methods

Benchmark datasets collection

The CRC dataset was sourced from the study by Zhou et al.(22). Both raw FASTQ files and processed gene expression count matrices for scRNA-seq were downloaded from the Genome Sequence Archive of the National Genomics Data Center (NGDC) with the accession number HRA000201. The glioma, NPC43 and HUVEC cell line data were obtained from the study by Yu et al(24) with gene expression count matrices for scRNA-seq available from the NCBI Gene Expression Omnibus under accession number GSE185269. Raw FASTQ data were aligned to the genome to generate BAM files, following the processing methods described in the original articles by Zhou et al and Yu et al.(22, 24). Additionally, the processed count matrix for two acute lymphoblastic leukemia samples was obtained from the study by Zachariadis et al with accession number GSE144296(23). Chromosomal copy number variations in scDNA-seq data and cell type annotations for all these datasets was acquired directly from the authors.

Workflows for inferCNV(16), CopyKAT(17), SCEVAN(18), Numbat(19) and CaSpER(20)

InferCNV: It utilizes a corrected moving average technique to compute expression levels of genes that are adjacent in the chromosomes. This approach helps mitigate gene-specific expression variability and generates chromosomal copy number profiles. The genes are arranged based on their absolute genomic positions, initially organized by chromosome and subsequently by the genomic start position within each chromosome. Leveraging the moving average expression data, InferCNV generates CNV profiles for samples, representing inferred copy number variations across diverse genomic regions. To further refine the CNV profiles of tumor cells, CNAs profiles were adjusted by subtracting the corresponding values from either normal cells or the average of all cells. In this project, we followed the guidelines on the GitHub repository of inferCNV (https://github.com/broadinstitute/infercnv). inferCNV version 1.18.1 was executed with default

parameters, except for the 'tumor_subcluster_partition_method' parameter, which was set to 'random_trees'.

CopyKAT: It combined Bayesian approach with hierarchical clustering to infer copy number profiles of individual cells and delineate clonal substructure from a matrix of unique molecular identifier (UMI) counts. Similar to the inferCNV method, CopyKAT begins by arranging genes based on their genomic coordinates. It then employs integrative hierarchical clustering and Gaussian mixture modeling (GMM) to identify potentially normal cells with high confidence, thereby inferring the copy number (CN) baseline. The subsequent step involves detecting chromosome breakpoints where CNA profiles exhibit changes. To accomplish this, CopyKAT integrates a Poisson-gamma model with Markov chain Monte Carlo (MCMC) iterations to compute posterior means per genomic window and employs Kolmogorov–Smirnov (KS) tests to connect adjacent windows. Ultimately, the final copy number values for each window are determined as the posterior averages of all genes spanning the adjacent chromosome breakpoints in each cell. In this project, we followed the guidelines provided on the GitHub repository for CopyKAT (v 1.1.0) (https://github.com/navinlabcode/copykat). CopyKAT was used with default parameters

SCEVAN: It is capable of automatically distinguishing between tumor and normal cells and inferring copy number breakpoints using an optimization-based joint segmentation algorithm. The analysis begins with a raw count matrix for genes, which is first log-transformed and filtered to remove genes detected in less than 10% of cells. Subsequently, a subset of high-confidence normal cells is then identified using gene signatures through a selection of those from the tumor microenvironment, stromal cells, and immune cells. The median expression of these normal cells serves as the baseline for copy number estimation. Next, a relative gene expression matrix is then derived by subtracting this baseline, followed by edge-preserving nonlinear diffusion filtering to achieve smoothed expression profiles. These smoothed expression profiles are then subjected to multichannel segmentation based on the Mumford and Shah energy principle, resulting in to a copy number matrix. In this project, we followed the guidelines provided in the SCEVAN (v 1.0.1) GitHub repository (https://github.com/AntonioDeFalco/SCEVAN). SCEVAN was executed using the default parameters, with the exception of setting 'SUBCLONES' , 'ClonalCN' and 'plotTree' to 'FALSE' in order to reduce the running time.

Numbat: It integrates expression, allele and haplotype information derived from population-based phasing to comprehensively characterize copy number variations profiles. The workflow of Numbat involves two key steps. First, it extracts allele counts at the cell level from BAM format files. Subsequently, it employs a haplotype-aware Hidden Markov Model (HMM) to infer CNVs within cell population pseudobulk profiles. The assignment of posterior probabilities to cells is used to determine whether a particular cell exhibits specific copy number alterations. In this project, we followed the guidelines provided on the GitHub repository of Numbat: https://kharchenkolab.github.io/numbat/articles/numbat.html. We obtained allele data using the default parameters of the pileup_and_phase.R script. We next used the run_numbat function from the Numbat package (version 1.3.2.1) to obtain CNA profiles with the default parameters. For Non-UMI protocols data, we set the expectation gamma parameter to 5, based on the author's recommendation. However, due to the severe imbalance in the number of tumor cells and normal cells in CRC22 and CRC15, Numbat cannot run successfully with the default parameters. To address this, we made specific adjustments. For the CRC22 dataset, we set the min_LLR parameter to 1, and the max_entropy to 0.6. For the CRC15 dataset, we set the min_cells parameter to 1 in order to obtain more filtered CNAs segments.

CaSpER: It utilizes a multiscale signal processing framework that integrates two streams of information: gene expression and B-allele frequencies (BAF), derived from either bulk or single-cell RNA sequencing data. Initially, CaSpER applies recursive iterative median filtering to smooth both the BAF and expression signals. At each scale, a hidden Markov model is employed to analyze the smoothed expression signal and categorize it into five CNAs states: (1) homozygous deletion, (2) heterozygous deletion, (3) neutral, (4) one-copy amplification, and (5) multi-copy amplification. For segments labeled as states 2 and 4, CaSpER incorporates information from the BAF signal. If a shift in the BAF signal accompanies the segment, states 2 and 4 are adjusted to states 1 and 5,

respectively. BAF shifts are determined by thresholding the smoothed signal, with thresholds calculated using a Gaussian mixture model fit on pooled samples across a segment. Consequently, CaSpER provides discrete CNA classifications: amplification, neutral, or deletion, as a result of these thresholds and corrections. In this project, we followed the guidelines on the GitHub repository of CaSpER: https://github.com/akdess/CaSpER. CaSpER (v 0.2.0) was executed using default parameters. Both required parameters ("cnv.scale" and "loh.scale") were set to 3 based on example codes.

Generation of synthetic datasets by downsampling of full glioma dataset spanning full range of tumor vs normal ratios

We used R (version 4.3.2) to downsample the glioma data based on the cell number ratio between tumor and normal cells of 1:100, 1:20, 1:4, 1:2, 2:1, 4:1, 20:1, and 100:1. Each of these tumor/normal ratios was resampled 50 times. The ratio of 1:1 represents the original full data. For each synthetic dataset, inferCNV, CopyKAT, SCEVAN, Numbat, and CaSpER software were employed to classify tumor and normal cells, and F1 scores were computed. The mean and standard error of F1 scores of 50 resamplings at each condition are reported.

Reference setting for Numbat and CaSpER

Numbat and CaSpER tools must specify reference expression. For Numbat, the parameter 'lambdas_ref' was used to regress out any cell-type-specific expression differences unrelated to CNVs. The reference expression for CRC and fibroblasts data was sourced from non-epithelial cells of corresponding patients. For glioma data, it was originated from astrocyte, oligodendrocyte, GABAergic, and glutamatergic cells. The reference expression for ALL was derived from TME cells. The reference expression for NPC43 cell line data was derived from the HUVEC cell line. For CaSpER, except for CRC data, other data used reference cells consistent with Numbat. Non-epithelial cells were not used in CRC data because we aimed to distinguish potential normal epithelial cells from tumor epithelial cells as reference cells to further improve the accuracy of CaSpER. This approach was more accurate than using non-epithelial cells (such as immune cells, fibroblasts, endothelial cells), which have cell type-specific transcription profiles that compromise the accuracy of CNAs detection. Therefore, we first divided epithelial cells into two categories through the Louvain algorithm. Generally, the transcriptional activity of normal epithelial cells is lower than that of tumor cells, resulting in lower UMIs. Hence, we considered the subset of cells with lower UMIs as the reference. For fibroblasts, we used the mean values across cells as reference for CaSpER.

Classification of tumor and normal cells

For CopyKAT, SCEVAN, and Numbat, we utilized the software's inherent classification to distinguish between tumor and normal cells. Since inferCNV and CaSpER do not directly classify tumor cells, we conducted hierarchical clustering on individual cell CNAs profiles, aiming for two clusters. We then determined the cluster with the higher standard deviation in mean copy number across cells as tumor cells, with the other cluster being normal cells.

Evaluation CNAs profile accuracy

We evaluated the accuracy of CNA profiles by calculating the Pearson correlation between scDNA-seq and bulk DNA-seq data. The heatmap of CNAs represents the median CAN values from individual cells. As Numbat and CaSpER provide categorical outputs, we assigned numerical equivalents: "deletion" as -1, "neutral" as 0, and "amplification" as 1.

Construction Clonal evolution inference

We performed hierarchical clustering on CNA profiles and identified two subclones from scDNA-seq. Matching scRNA-seq derived subclones (C1 and C2) with those from scDNA-seq with those from scDNA-seq, we calculated the median values for each subclone and computed the Pearson correlation. Subclones were assigned based on the highest correlation.

Aneuploidy fibroblast cells identification

In the CRC data, Numbat detect CNAs from fibroblasts from the CRC15 patient, likely due to their higher CNAs burden compared to other patients. As a result, we present data exclusively from

CRC15 for fibroblast analysis using all five software tools.

Genome-level heatmap

Due to inconsistencies in the start and end coordinates of CNAs segments on chromosomes across different software outputs, we employed the following formula to uniformly compute the average signal by splitting the genome into 1MB windows.

$$v_w = \frac{\sum_i^n x_i w_i}{\sum_i^n w_i}$$

v_w represents the weighted mean of all signal regions, with the weights determined by the width of their intersections in the 1MB window. w_i denotes the interaction width of CNAs segments within the 1MB window, and x_i represents the corresponding CNAs signal value.

Computational speed performance

We evaluated both time and memory usage on a High-Performance Computing platform. Each method was executed using a single-core setting. Runtime was captured for each method using the Sys.time() function in R (v.4.1.2).

Data visualization

Heatmaps were visualized by R package ComplexHeatmap (v2.10.0)(28). All bar plots and box plots were visualized using the R packages ggpubr (v 0.4.0)(29).

Statistics

All statistical analyses and presentations were performed using R (v4.1.2). The methods of statistical test were indicated in the figure legends, and adjustments for multiple comparisons were made using the Benjamini & Hochberg method when necessary.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The scRNA-seq data used in this study are publicly available at the
National Genomics Data Center (NGDC) with accession number HRA000201 (Colorectal cancer)(22), the Gene Expression Omnibus with accession numbers GSE185269 (Glioma and cell line)(24) and GSE144296 (Acute lymphoblastic leukemia)(23). The code and scripts used for the benchmark analysis and figure plotting has been uploaded to GitHub:
https://github.com/bioliyezhang/Benchmark_scRNASeq_by_scMultiomics.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by grants from National Natural Science Foundation of China (No. 31871332) and zhejiang lab Development of Novel Functional Proteins Based on Databases and Artificial Intelligence (No.117005-AC2106/002).

Authors' contributions

L.Z., M.S., Y.Z. and S.M. wrote the paper. L.Z. designed the study with M.S. and S.M. M.S. and S.M. wrote code, prepared the data, analysis and figures. L.Z. and X.H. supervised the work. B.X. and T.G. modified the details of the figures. All authors reviewed and approved the final paper.

Acknowledgements
We thank Prof. Fuchou Tang from Peking University, Prof. Shuhui Bian from Nanjing Medical University, Prof. Angela Ruohao Wu and Dr. Lei Yu from Hong Kong University for providing data. We also thank ShanghaiTech University's high-performance computing platform and Zhejiang lab for support.

Reference

1. Wen L, Li G, Huang T, Geng W, Pei H, Yang J, et al. Single-cell technologies: From research to application. Innovation (Camb). 2022;3(6):100342.
2. Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single-cell RNA sequencing technologies and applications: A brief overview. Clin Transl Med. 2022;12(3):e694.
3. Andrews TS, Kiselev VY, McCarthy D, Hemberg M. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. Nat Protoc. 2021;16(1):1-9.
4. Xue R, Zhang Q, Cao Q, Kong R, Xiang X, Liu H, et al. Liver tumour immune microenvironment subtypes and neutrophil heterogeneity. Nature. 2022;612(7938):141-7.
5. Ren X, Zhang L, Zhang Y, Li Z, Siemers N, Zhang Z. Insights Gained from Single-Cell Analysis of Immune Cells in the Tumor Microenvironment. Annu Rev Immunol. 2021;39:583-609.
6. Sun Y, Wu L, Zhong Y, Zhou K, Hou Y, Wang Z, et al. Single-cell landscape of the ecosystem in early-relapse hepatocellular carcinoma. Cell. 2021;184(2):404-21 e16.
7. Peng J, Sun BF, Chen CY, Zhou JY, Chen YS, Chen H, et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res. 2019;29(9):725-38.
8. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell. 2017;171(7):1611-24 e24.
9. Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun. 2017;8:15081.
10. Bailey C, Black JRM, Reading JL, Litchfield K, Turajlic S, McGranahan N, et al. Tracking Cancer Evolution through the Disease Course. Cancer Discov. 2021;11(4):916-32.
11. Knouse KA, Wu J, Amon A. Assessment of megabase-scale somatic copy number variation using single-cell sequencing. Genome Res. 2016;26(3):376-84.
12. Baslan T, Kendall J, Ward B, Cox H, Leotta A, Rodgers L, et al. Optimizing sparse sequencing of single cells for highly multiplex copy number profiling. Genome Res. 2015;25(5):714-24.
13. Cai X, Evrony GD, Lehmann HS, Elhosary PC, Mehta BK, Poduri A, et al. Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. Cell Rep. 2014;8(5):1280-9.
14. Ruohan W, Yuwei Z, Mengbo W, Xikang F, Jianping W, Shuai Cheng L. Resolving single-cell copy number profiling for large datasets. Brief Bioinform. 2022;23(4).
15. Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal GS, Hicks J, et al. Interactive analysis and assessment of single-cell copy-number variations. Nat Methods. 2015;12(11):1058-60.
16. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, 2nd, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016;352(6282):189-96.
17. Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol. 2021;39(5):599-608.
18. De Falco A, Caruso F, Su XD, Iavarone A, Ceccarelli M. A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data. Nat Commun. 2023;14(1):1074.
19. Gao T, Soldatov R, Sarkar H, Kurkiewicz A, Biederstedt E, Loh PR, et al. Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. Nat Biotechnol. 2023;41(3):417-26.
20. Serin Harmanci A, Harmanci AO, Zhou X. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. Nat Commun. 2020;11(1):89.
21. Erickson A, He M, Berglund E, Marklund M, Mirzazadeh R, Schultz N, et al. Spatially resolved clonal copy number alterations in benign and malignant tissue. Nature. 2022;608(7922):360-7.
22. Zhou Y, Bian S, Zhou X, Cui Y, Wang W, Wen L, et al. Single-Cell Multiomics Sequencing Reveals Prevalent Genomic Alterations in Tumor Stromal Cells of Human Colorectal Cancer. Cancer Cell. 2020;38(6):818-28 e5.
23. Zachariadis V, Cheng H, Andrews N, Enge M. A Highly Scalable Method for Joint Whole-

Genome Sequencing and Gene-Expression Profiling of Single Cells. Mol Cell. 2020;80(3):541-53 e5.
24. Yu L, Wang X, Mu Q, Tam SST, Loi DSC, Chan AKY, et al. scONE-seq: A single-cell multi-omics method enables simultaneous dissection of phenotype and genotype heterogeneity from frozen tumors. Sci Adv. 2023;9(1):eabp8901.
25. Iourov IY, Vorsanova SG, Yurov YB. Somatic genome variations in health and disease. Curr Genomics. 2010;11(6):387-96.
26. Dentro SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG, et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. Cell. 2021;184(8):2239-54 e39.
27. Wang F, Wang Q, Mohanty V, Liang S, Dou J, Han J, et al. MEDALT: single-cell copy number lineage tracing enabling gene discovery. Genome Biol. 2021;22(1):70.
28. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32(18):2847-9.
29. Kassambara A. ggpubr: 'ggplot2' Based Publication Ready Plots 2020 [Available from: https://CRAN.R-project.org/package=ggpubr.

Main figure titles and legends

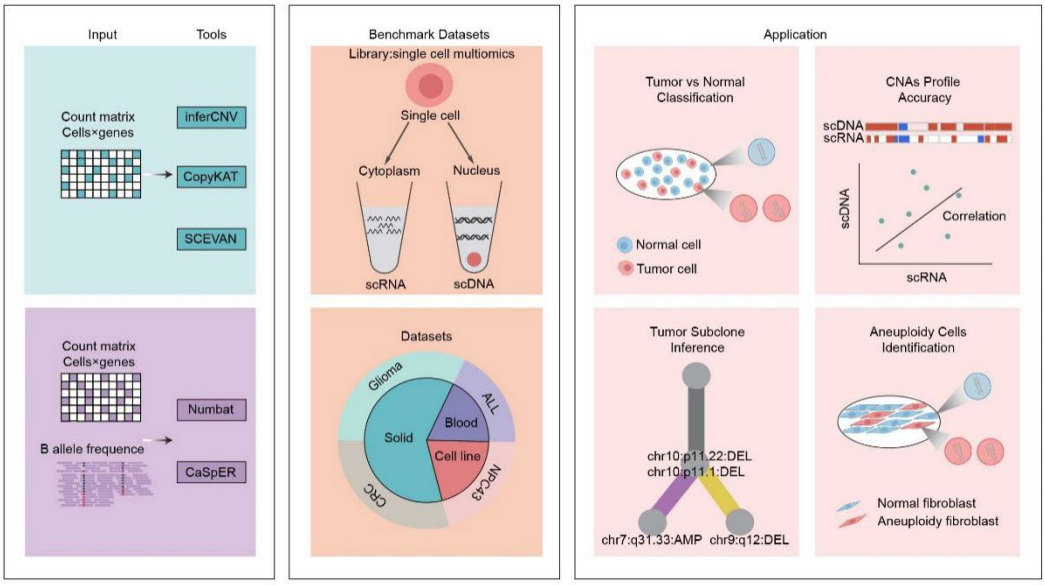

Figure 1: Overview of the benchmarking workflow
We evaluated five tools designed to infer copy number alterations (CNAs) from single-cell RNA-seq data. The methods fall into two categories: three rely solely on gene expression matrices: inferCNV, CopyKAT, and SCEVAN, while the remaining two, Numbat and CaSpER, utilize both gene expression and allele information. Our evaluation utilized single-cell multi-omics datasets that simultaneously interrogate DNA and RNA within the same cell, encompassing solid tumors, hematopoietic malignancies, and cancer cell lines. Copy number alterations (CNAs) detected from scDNA-seq served as the ground truth for evaluating the accuracy of CNAs inference tools based on scRNA-seq data. The performance of these five tools was assessed across several criteria, including tumor-normal classification, CNAs profile accuracy, subclonal structure inference, and aneuploidy cells identification in non-malignant cells.

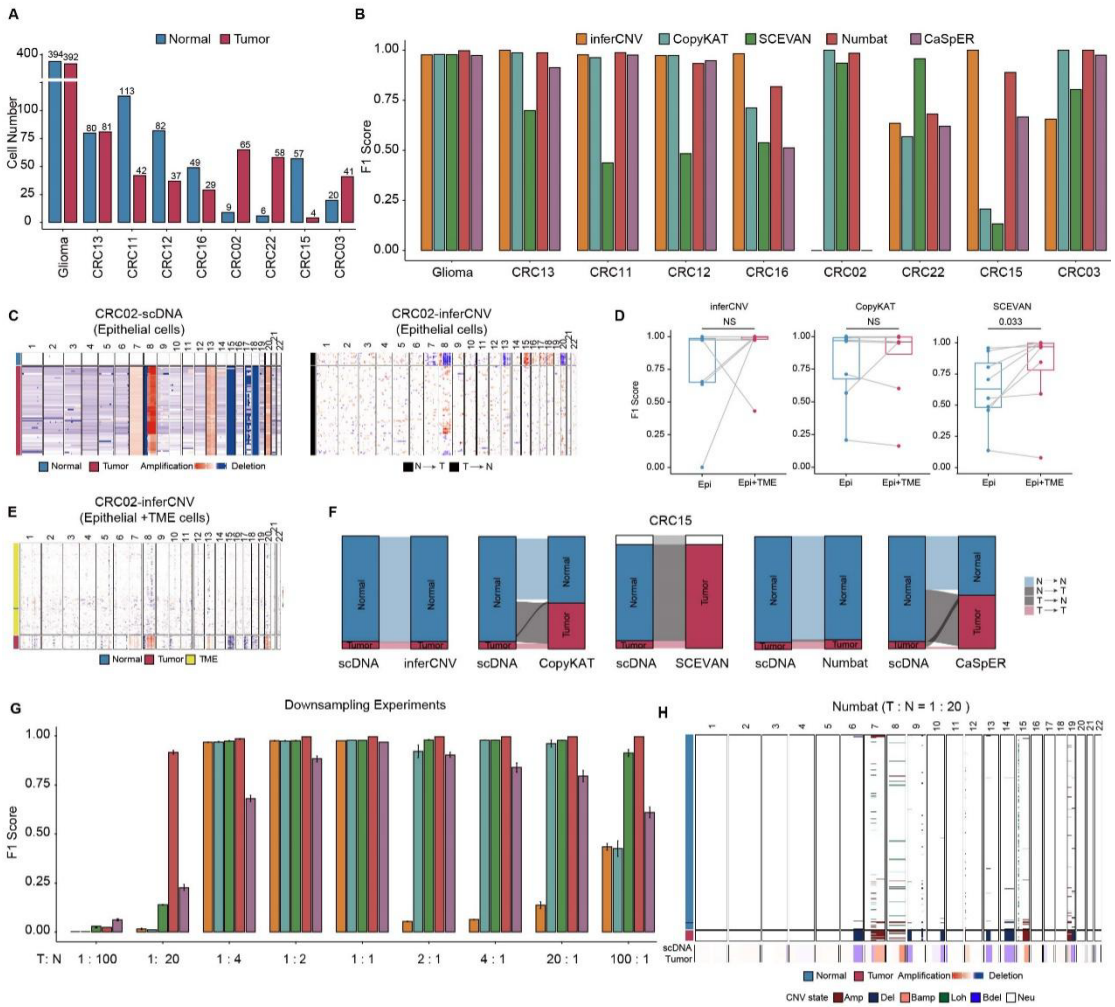

Figure 2: Benchmark of tumor and normal cells classification

A: A bar plot illustrates the counts of tumor and normal cells for each patient.

B: The F1 scores for tumor and normal cell classification for various tools.

C: Heatmap showing the CNAs profiles of CRC02 epithelial cells derived from scDNA-seq (ground truth) (left side) and inferred from scRNA-seq by inferCNV (right).

D: F1 scores between epithelial cells alone and epithelial cells combined with TME cells for CRC patients.

E: Heatmap displaying the CNAs profiles of CRC02 of epithelial cells combined with TME cells estimated by inferCNV.

F: Sankey diagrams illustrating the classification of tumor and normal cells from CRC15, with correct classifications as normal and tumor in both scDNA-seq and scRNA-seq data indicated by N>N and T->T, and misclassifications by N->T and T->N. T represents tumor cells, while N represents normal cells.

G: F1 scores are shown for glioma downsampling experiments, with the x-axis indicating the ratio of tumor and normal cells.

H: The upper heatmap represents the CNAs profiles of glioma from a downsampling experiment with a T : N ratio of 1 : 20, as inferred by Numbat. The lower heatmap represents the average CNA profile of tumor cells based on scDNA-Seq.

p values in D calculated using paired Wilcoxon rank-sum test. T: Tumor cells, N: Normal cells.

Additional data can be found in Figures S1-S3 and Tables S1-2.

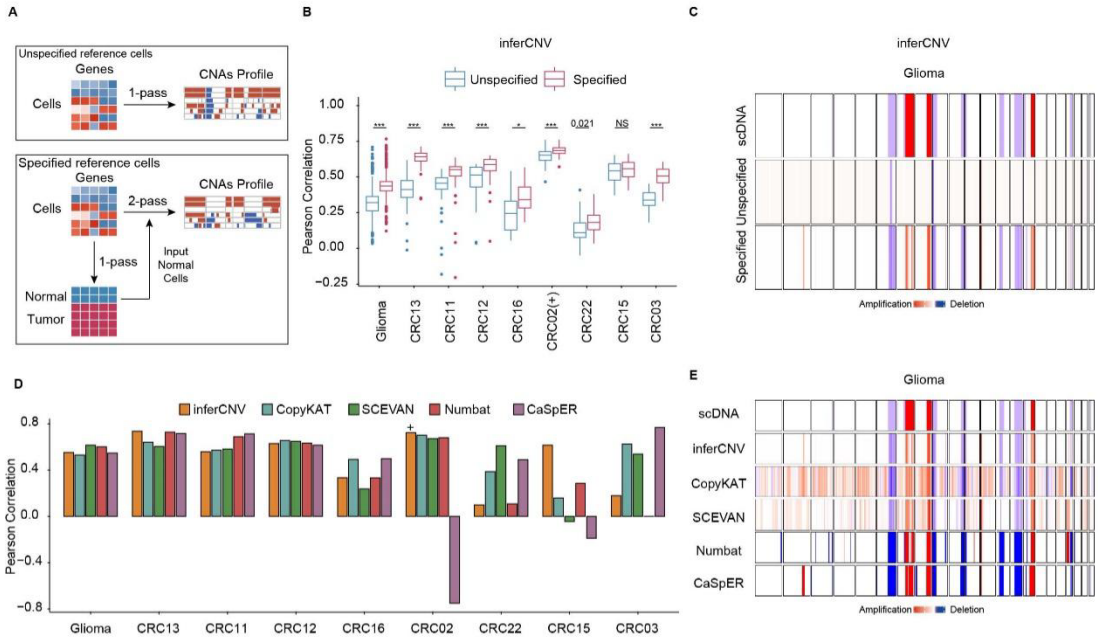

Figure 3: Benchmark of the accuracy of inferred CNAs profiles

A: Diagram illustrating CNAs profiles under unspecified (one-pass) and specified reference cells (two-pass) with inferCNV. In the one-pass scenario, inferCNV processes the whole cell population with the average as the reference. Conversely, in the two-pass approach, "normal" cells identified by inferCNV in the initial run (one-pass) serve as the reference for a second run (two-pass).

B: Pearson correlation between inferred CNAs from scRNA-seq and the CNAs from scDNA-seq for each patient under unspecified (one-pass) and specified reference cells (two-pass) with inferCNV. Each dot represents the Pearson correlation between CNA profiles in scDNA-seq and inferred CNA profiles from paired scRNA-seq from the same cell.

C: Heatmap showing CNAs profiles of glioma tumor cells obtained from scDNA-seq (upper), unspecified reference (middle), and specified reference using inferCNV (bottom).

D: Bar plots illustrate the Pearson correlation of tumor cells between scRNA-seq inferred CNAs from and scDNA-seq for each patient across various tools.

E: A heatmap displays the average CNAs profiles of glioma tumor cells derived from scDNA-seq assessed by different tools.

"+" in (B) and (D) indicates tumor cell CNAs profiles estimated when both epithelial and TME cells are analyzed by inferCNV. p-values in (B) are calculated using unpaired two-tailed Wilcoxon rank-sum test and corrected using the Benjamini–Hochberg method.

Additional information is available in Figures S4-S7.

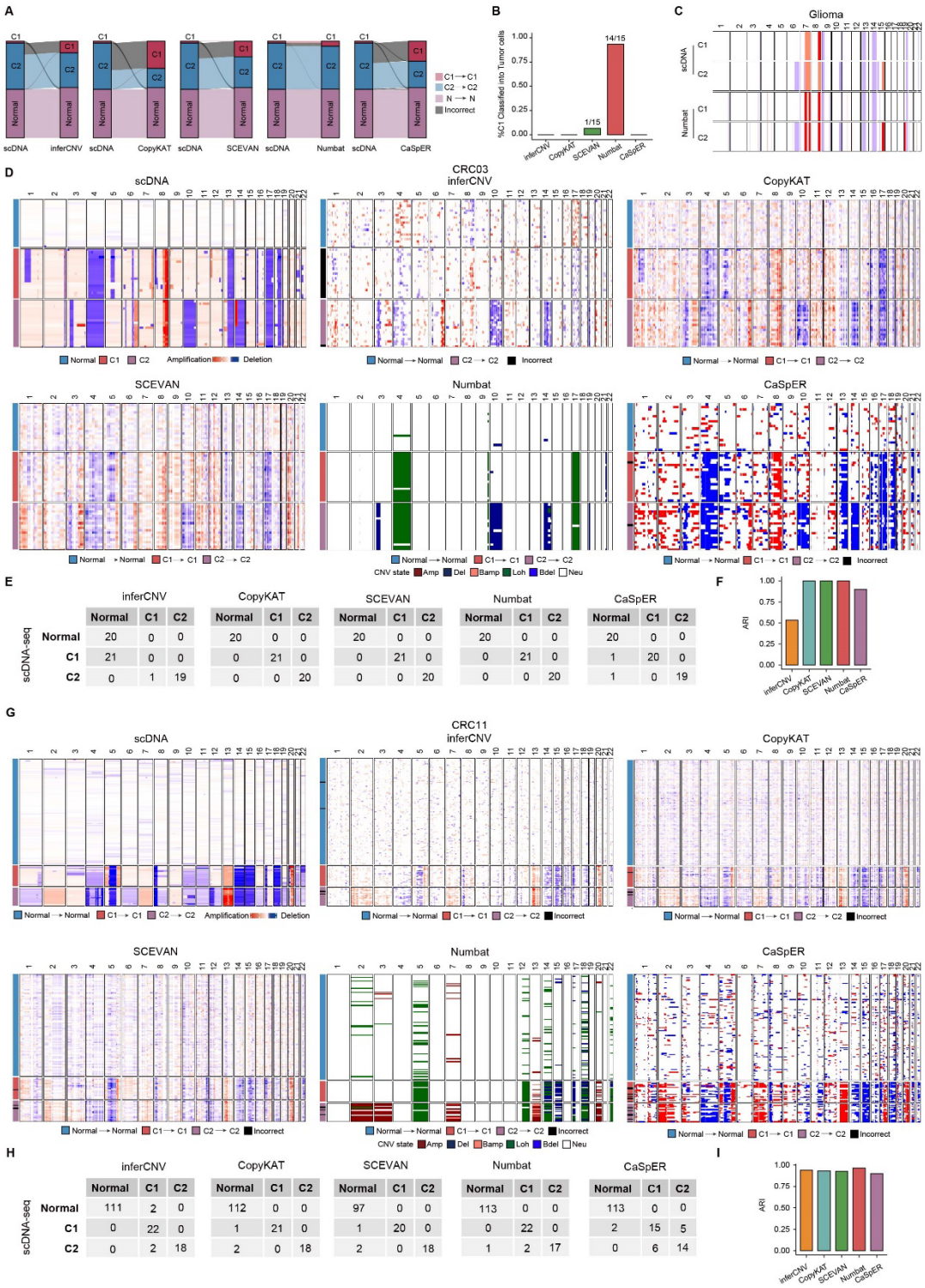

Figure 4: Benchmark of tumor clonal evolution inference

A: Sankey diagrams illustrate the clonal structures inferred from scDNA and scRNA of gliomas.

B: A bar plot indicates the percentage of correct classification of the C1 subclone as tumor cells.

C: Heatmaps compare the CNAs profiles of two subclones in glioma between scDNA-Seq and those inferred by Numbat.

D & G: Heatmaps display the CNAs profiles for CRC03 (D) and CRC11 (G) subclones, obtained from scDNA-seq and scRNA-Seq.

E &H: Tables list the cell numbers associated with subclone labels from scDNA and the predicted subclone labels from scRNA for CRC03 (E) and CRC11 (H).

F & I: Bar plots present the Adjusted Rand Index (ARI) for subclones in CRC03 (F) and CRC11 (I). Additional information is available in Figure S8.

		inferCNV	CopyKAT	SCEVAN	Numbat	CaSpER
	Method	Corrected moving genes average	Bayesian segmentation	Multi-channel segmentation	Haplotype-aware Hidden Markov Model	Multiscale signal-processing
	Programming lanuage	R	R	R	R+Linux	R+Linux
	Version	1.18.1	1.1.0	1.0.1	1.4.0	0.2.0
Input	Gene expression matrix	Required	Required	Required	Required	Required
Input	B allele frequence				Required	Required
Parameter	Reference setting	Optional	Optional	Optional	Mandatory	Mandatory
Parameter	Parameters optimization	Two-pass +TME		+TME		
Output	CNV state	Corrected gene expression	Log (Copy number ratios)	Log (Copy number ratios)	Amp, Del, Bamp Cnloh, Bdel	Amp, Del
Output	CNV break point	No	Yes	Yes	Yes	Yes
Output	Chromosome	Chr1-22 ChrX, ChrY	Chr1-22 ChrX	Chr1-22	Chr1-22	Chr1-22
Output	Tumor prediction	No	Yes	Yes	Yes	No
Output	Subclone predication	Yes	No	Yes	Yes	No
	Speed	Medium	Fast	Fast	Slow	Medium
	Recommendations	★	★★	★	★★★	★

Figure 5: Comparison overview for all methods.
We systematically summarized the methodology, programming language used, tools version, input data formats, parameters settings, output data, computational speed and recommendations for each software. Two-pass represents running inferCNV twice and inputs normal cell inferred from the first run as reference cells, see Figure3A for more details; +TME: including tumor microenvironment cells as input.