Feng *et al. BMC Genomics* (2021) 22:651
https://doi.org/10.1186/s12864-021-07941-3

BMC Genomics

METHODOLOGY Open Access

# SCYN: single cell CNV profiling method using dynamic programming

Xikang Feng[1,2†], Lingxi Chen[2†], Yuhao Qing[2], Ruikang Li[2], Chaohui Li[2] and Shuai Cheng Li[2,3*]

*From* The 19th Asia Pacific Bioinformatics Conference (APBC 2021)
Tainan, Taiwan. 3-5 February 2021

## Abstract

**Background:** Copy number variation is crucial in deciphering the mechanism and cure of complex disorders and cancers. The recent advancement of scDNA sequencing technology sheds light upon addressing intratumor heterogeneity, detecting rare subclones, and reconstructing tumor evolution lineages at single-cell resolution. Nevertheless, the current circular binary segmentation based approach proves to fail to efficiently and effectively identify copy number shifts on some exceptional trails.

**Results:** Here, we propose SCYN, a CNV segmentation method powered with dynamic programming. SCYN resolves the precise segmentation on in silico dataset. Then we verified SCYN manifested accurate copy number inferring on triple negative breast cancer scDNA data, with array comparative genomic hybridization results of purified bulk samples as ground truth validation. We tested SCYN on two datasets of the newly emerged 10x Genomics CNV solution. SCYN successfully recognizes gastric cancer cells from 1% and 10% spike-ins 10x datasets. Moreover, SCYN is about 150 times faster than state of the art tool when dealing with the datasets of approximately 2000 cells.

**Conclusions:** SCYN robustly and efficiently detects segmentations and infers copy number profiles on single cell DNA sequencing data. It serves to reveal the tumor intra-heterogeneity. The source code of SCYN can be accessed in https://github.com/xikanfeng2/SCYN.

**Keywords:** scDNA-Seq, CNV segmentation, Dynamic programming

## Background

Numerous studies have shown that copy number variations (CNV) can cause common complex disorders [1–5]. Copy number aberration (CNA), aka, somatic CNV, is also reported to be a driving force for tumor progression and metastasis. For example, George et al. reported the high amplification of oncogene gene *PD-L1* in small-cell lung cancer [6] and amplification of *MYC* is announced prevailing in pan-cancer studies [7]. The loss of tumor suppressor genes like *KDM6A* and *KAT6B* are proclaimed indirectly amplifies harmful cancer-related pathways [8, 9].

Conventional experimental protocols for CNV segmentation lies in the following scenarios. Researchers may infer a coarse CNV profiles utilizing bulk RNA sequencing [10] and single cell RNA sequencing [11–13] [10]. Moreover, scientists may leverage bulk genome techniques such as DNA array comparative genomic hybridization (aCGH) [14], single-nucleotide polymorphism (SNP) arrays [15, 16], and DNA next generation sequencing (NGS) [17, 18] to generate high resolution CNV. Although bulk genome sequencing studies have contributed insights into

*Correspondence: shuaicli@cityu.edu.hk
†Xikang Feng and Lingxi Chen contributed equally to this work.
[2]Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China
[3]Department of Biomedical Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China
Full list of author information is available at the end of the article

tumor biology, the data they provide may mask a degree of heterogeneity [19]. For instance, if the averaged readout overrepresents the genomic data from the dominant group of the tumor cells, rare clones will be masked from the signals. The advent of single-cell DNA sequencing (scDNA-Seq) delivers a potential solution to overwhelm the deficiencies of bulk approaches [20–22]. By assigning a unique barcode to each single cell, scDNA-seq is able to record the minority cell population, thus to address intratumor heterogeneity (ITH) [22], detect rare subclones [19], and reconstruct tumor evolution lineages [20, 23].

In this study, we concentrate on the CNV segmentation and turning points detection approaches customized for single cell DNA sequencing. CNV Segmentation refers to partitioning the genome into non-overlapping segments with the objective of that each segment shares intra-homogeneous CNV profile, and the segment boundaries are often termed to be checkpoints or turning points [24]. Although numerous CNV segmentation tools have emerged leveraging high throughput sequencing data such as Circular Binary Segmentation (CBS) [25, 26] and Hidden Markov Model (HMM) [27, 28], the methods customized for scDNA data is in its infancy. Gingko [29], SCNV [30], and SCOPE [31] applied diverse strategies to normalize the scDNA intensities through simultaneously considering sparsity, noise, and cell heterogeneity, and adopted variational CBS for checkpoint detection. While after in silico experiment, we argue that those CBS approaches might not lead to an optimal segmentation result, some turning points might be masked. Furthermore, with the advance of large scale high throughput technologies, the scale of cells for a single dataset climbs exponentially. For instance, the newly emerged 10x Genomics CNV solution can profile the whole genome sequencing of thousands of cells at one time [22]. Thus, efficiently processing scDNA-seq data is crucial. However, current scDNA CNV segmentation methods are too time-consuming to process thousands of cells.

Therefore, in this paper, we propose SCYN (Single Cell and dYNamic programming), an effecient and effective dynamic programming approach for single cell data CNV segmentation and checkpoint detection. SCYN resolves the precise turning points on in silico dataset, while existing tools fail. SCYN manifested more precise copy number inference on a triple-negative breast cancer scDNA dataset, with array comparative genomic hybridization results of purified bulk samples as ground truth validation. We tested SCYN on two datasets of the newly emerged 10x Genomics CNV solution. SCYN successfully recognizes gastric cancer cells from 1% and 10% spike-ins 10x datasets. Last but not least, SCYN is about 150 times faster than state of the art tool when dealing with thousands of cells.

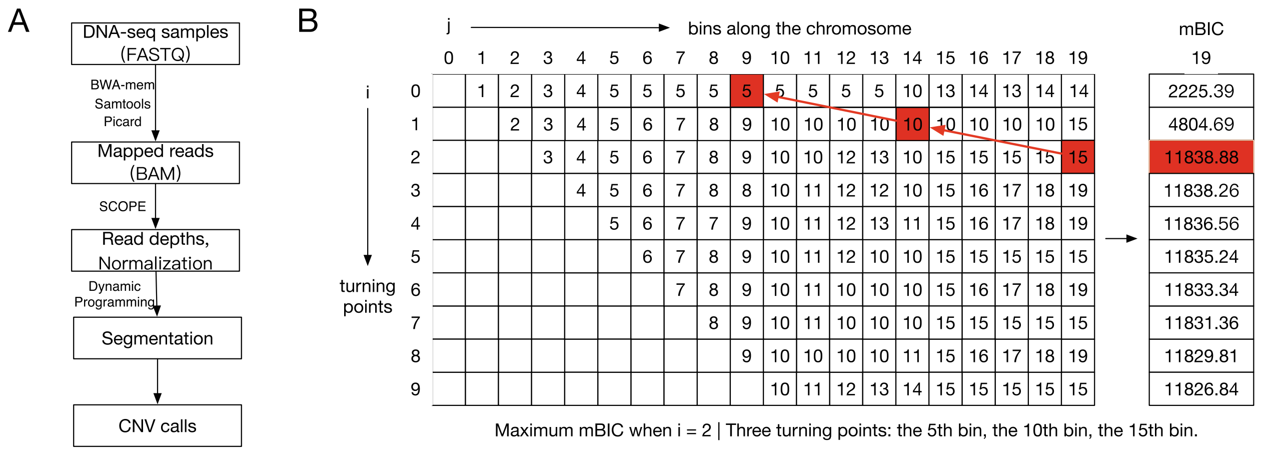
## Results

### Overview of SCYN

We developed an algorithm, SCYN, that adopts a dynamic programming approach to find optimal single-cell CNV profiles. The framework for SCYN displayed in Fig. 1A. First, the raw scDNA-seq reads of FASTQ format are preprocessed with standard procedures (Fig. 1A). SCYN then takes the aligned BAM files as the input. SCYN integrates SCOPE [31], which partitions chromosomes into consecutive bins and computes the cell-by-bin read depth matrix, to process the input BAM files and get the raw and normalized read depth matrices. The segmentation detection algorithm is then performed on the raw and normalized read depth matrices using our dynamic programming to identify the optimal segmentation along each chromosome. The segmentation results are further applied to copy number calculation. Finally, SCYN outputs the cell-by-bin copy number matrix and the segmentation results of all chromosomes for further CNV analysis.

### SCYN effectively identifies all turning points on synthetic trial

To evaluate the segmentation power of SYCN against SCOPE, we conducted one simulation experiment. We first generated a synthetic CNV profile of 100 singe cells on chromosome 22, with 50M bp as one bin, resulting in a 100 ×70 CNV matrix. As illustrated in Fig. 2B, there is a large proportion of normal cells with average diploid copy number and four tumor subclones, which manifests six turning points and seven segments on chr22. Then, we fit the ground truth CNV profiles into single cell sequencing simulator SCSsim [32] to get the synthetic FASTQ reads (Fig. 2A). Figure 2C-D shows the inferred CNV profiles on the simulated reads from SCYN and SCOPE, respectively. Both SCYN and SCOPE able to recognize the normal cells and mask the noises. SCYN did sound work on CNV segmentation to correctly identify all six turning points and uncovered the cell heterogeneity. Nevertheless, SCOPE adds one nonexistent turning point inside segmentation S1, and drops two critical turning points which discriminate S4-S5 and S6-S7. These then lead to erroneous CNV segmentation and CNV estimations. Furthermore, we conducted a series of in silico spike-in experiments for CNV turning points detection with different proportion of normal cells (Additional file 1, Supplementary Figure S1A), different number of cell clusters (Additional file!1, Supplementary Figure S1B), and different number of CNV segments (Additional file 1, Supplementary Figure S1C), respectively. Our results show that SCYN call turning points with 100% accuracy regardless of the cell and segment CNV complexity of ground-truth settings, whilst SCOPE always call false positive and false negative points. As previously mentioned, the core principle of CNV segmentation is partitioning the genome into



| i | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | mBIC 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 |  | 1 | 2 | 3 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 10 | 13 | 14 | 13 | 14 | 14 | 2225.39 |
| 1 |  |  | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 15 | 4804.69 |
| 2 |  |  |  | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 10 | 12 | 13 | 10 | 15 | 15 | 15 | 15 | 15 | 11838.88 |
| 3 |  |  |  |  | 4 | 5 | 6 | 7 | 8 | 8 | 10 | 11 | 12 | 12 | 10 | 15 | 16 | 17 | 18 | 19 | 11838.26 |
| 4 |  |  |  |  |  | 5 | 6 | 7 | 7 | 9 | 10 | 11 | 12 | 13 | 11 | 15 | 16 | 17 | 18 | 19 | 11836.56 |
| 5 |  |  |  |  |  |  | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 10 | 10 | 15 | 15 | 15 | 15 | 15 | 11835.24 |
| 6 |  |  |  |  |  |  |  | 7 | 8 | 9 | 10 | 11 | 10 | 10 | 10 | 15 | 16 | 17 | 18 | 19 | 11833.34 |
| 7 |  |  |  |  |  |  |  |  | 8 | 9 | 10 | 11 | 10 | 10 | 10 | 15 | 15 | 15 | 15 | 15 | 11831.36 |
| 8 |  |  |  |  |  |  |  |  |  | 9 | 10 | 10 | 10 | 10 | 11 | 15 | 16 | 17 | 18 | 19 | 11829.81 |
| 9 |  |  |  |  |  |  |  |  |  |  | 10 | 11 | 12 | 13 | 14 | 15 | 15 | 15 | 15 | 15 | 11826.84 |

Maximum mBIC when i = 2 | Three turning points: the 5th bin, the 10th bin, the 15th bin.

**Fig. 1** Overview of SCYN. (A) The workflow of SCYN. (B) Illustration of SCYN segmentation counting table during dynamic programming. The backtracking process of finding the optimal turning points and corresponding mBIC are highlighted in red

non-overlapping areas with the objective of that each area shares intra-homogeneous CNV profile [24, 30]. SCOPE fails to hit the correct answer as its turning point detection fails. Overall, our experiment on synthetic data suggest that empowered with dynamic programming, SCYN can achieve the correct copy number turning point detection against the segmentation schema SCOPE proposed.

### SCYN successfully identifies subclones in wet-lab cancer datasets

We illustrate the performance of SCYN in cancer single-cell datasets. We collected two cancer data sets, namely the Nature_TNBC (two triple-negative breast cancers) [33] and 10x_Gastric (gastric cancer spike-ins). We illustrated the tumor intra-heterogeneity discovered by SCYN and validated the results of SCYN against the estimation made by SCOPE for ground truth available datasets.

The first benchmark dataset we investigated is Nature_TNBC. 100 single cells were separately sequenced from two triple-negative breast cancer samples, namely, T10 and T16 [33]. For T10, we removed cell SRR054599 as it did not pass the quantity control, resulting 99 single cells from held four subgroups: Diploid (D), Hypodiploid (H), Aneuploid A (A1), and Aneuploid B (A2). We first verified if SYCN could replicate the subclone findings previously reported. Figure 3A demonstrates the genome-wide copy number profiles across the 100 single cells for T10. Overall, the cell subclones recognized by SCYN are concordant with the outputs of SCOPE (Additional file 1, Supplementary Figure S2A) and Navin et al.'s findings. With hierarchical clustering, SCYN categorizes T10 into seven clusters. As illustrated in Fig. 3 and Additional file 1 Supplementary Figure S3A-4A, for T10, cluster 1 matches the diploid (D) cells and cluster 3 represents the hypodiploid (H) group. There are two hyperdiploid subgroups. Cluster 4 corresponds to aneuploid A (A1) and cluster 2,5,6,7 together represents aneuploid B (A2). Navin et al. also separately profiled the four subgroups through array comparative genomic hybridization (aCGH) [34], here we regarded the CNV profiled from aCGH as golden-standard to examine the SYCN and SCOPE performance. As illustrated in Fig. 3B-C, SCYN owns a higher Pearson correlation and a lower root mean squared error (RMSE) of ground-truth against SCOPE.

T16 sample is a mixture of one primary breast tumor (T16P, 52 single cells) and its corresponded liver metastasis (T16M, 48 single cells). Navin et al. identified five cell subpopulations: Primary Diploid (PD), Primary Pseudodiploid (PPD), Primary Aneuploid (PA), Metastasis Diploid (MD), and Metastasis Aneuploid (MA). Figure 4A records T16 genome-wide copy number profiles across the 100 single cells. In all, the cell subclones recognized by SCYN are consistent with SCOPE (Additional file 1, Supplementary Figure S2B) and Navin et al.'s findings. Hierarchical clustering characterizes T16 into seven subgroups. As depicted in Fig.4 and Additional file 1 Supplementary Figure S3B-4B, cluster 1 mates the primary diploid (PD) cells. Cluster 3 represents metastasis aneuploid (MA), and cluster 6,7 together pictures primary aneuploid (PA). As Navin *et al.* only profiled four bulk dissections using of T16 aCGH [34], there lacks the CNV gold standard for 16T *in su* subclones. So we calculated the CNV correlation and RMSE between inferred primary aneuploid (PA) subpopulation and the four dissections, respectively. From Fig. 4B-C, although the association between PA group and four bulk dissections is relatively low, SCYN profiles a closer correlation than SCOPE with higher correlation and lower discrepancy.

We next employed SCYN and SCOPE to the lately published single cell DNA spike-in demo datasets available at

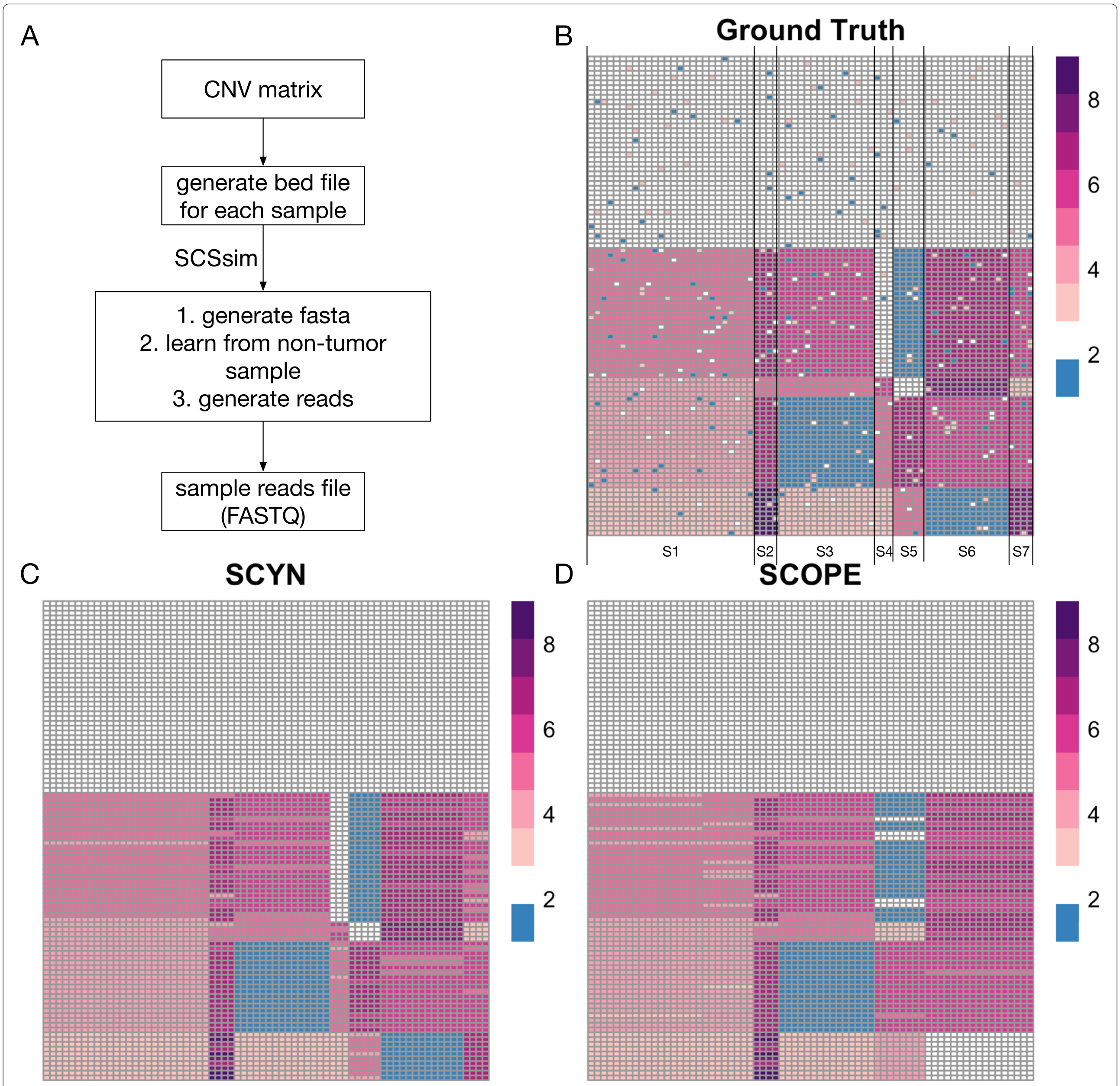


**Fig. 2** SCYN performance on synthetic case. (A) The workflow of synthetic data generation. (B-D) The heatmap of copy number profiles with single cells as row and genomics bins as columns, from ground truth, SCYN, and SCOPE, respectively. The colorbar (blue, white, dark red) represents the value of copy number from 1 to 9, respectively. In (B), vertical lines signify the ground-truth turning points, which divide the genomic bins into seven segments from S1 to S7

the 10x Genomics official website. 10x Genomics mixed BJ fibroblast euploid cell line with 1% and 10% spike-in of cells from MKN-45 gastric cancer cell line. As illustrated in the CNV heatmap Fig.5A and Additional file 1 Supplementary Figure S5, SCOPE successfully distinguished the two spike-in gastric cancer cells. Furthermore, we visualized the first two principal components of the estimated CNV profiles in Fig.5B-C. Cells whose Gini coefficient more massive than 0.12 were highlighted in yellow and regarded as gastric cancer cells from the 1% and 10% spike-ins, respectively. Then, we checked if SYCN produced CNV profiles better preserves the cell subpopulation information against SCOPE. Leveraging Gini 0.12 as the cut-off value, we partitioned cells into normal and cancer subset as benchmark labels. Next, we practiced hierarchical clustering into CNV matrices attained from SYCN and SCOPE, and get two clusters for each spike-in sets. Then, we adopt four metrics to inquire about the

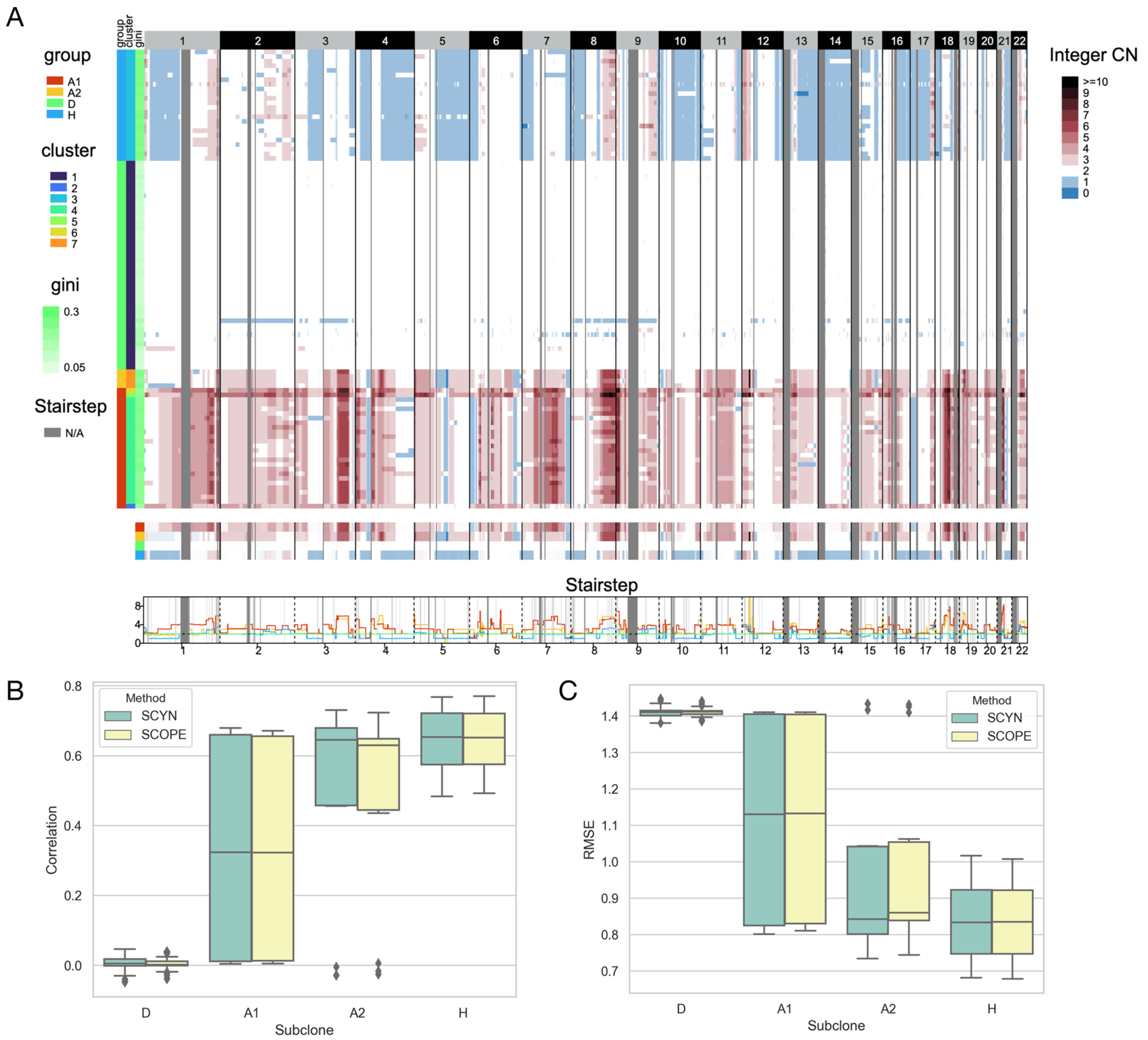


**Fig. 3** Performance of SCYN on T10. (A) Whole genome CNV heatmap, aggregated group CNV heatmap, and aggregated group CNV stairstep of T10. The colorbar (blue, white, dark red) represents the value of copy number from 0 to ≥10, respectively. N/A denotes the missing of copy number in corresponding genomic region. An interactive visualization is available on scSVAS [42] (https://sc.deepomics.org/oviz-project/analyses/view) with demo set "TNBC_T10". (B-C) Pearson correlation and root mean square error (RMSE) as evaluation metrics comparing results by SCYN and SCOPE against golden standard aGCH, respectively

clustering accuracy of SYCN against SCOPE. The adjusted Rand index (ARI) [35], Normalized mutual information (NMI) [36], and Jaccard index (JI) [37] measures the similarity between the implied groups and golden-standard labels; a value approaching 0 purports random assignment, and one reveals accurate inferring. As evidenced in Tables 1 and 2, with ARI, NMI, and JI as measurements, SYCN holds equal clustering accuracy to SCOPE on both 1% and 10% spike-in sets, which indicates SYCN captures substantial interior tumor heterogeneity.

## SCYN segmentation is fast

Recall that efficient processing of scRNA-seq data is essential, especially in today's thousands of single cells throughput. To evaluate the efficiency of SCYN against SCOPE, we measured the checkpoint detection step CPU running time of SCYN and SCOPE on T10, T16M, T16P, 10x 10% spike-in, 10x 1% spike-in, and several simulation data sets (90-1, 90-2, 2000-1, 2000-2, 2000-3, 2000-4, and 2000-5), with the cell number ranging from 48 to around 2000. We respectively ran SCYN and SCOPE on each dataset ten times and calculated the mean CPU running

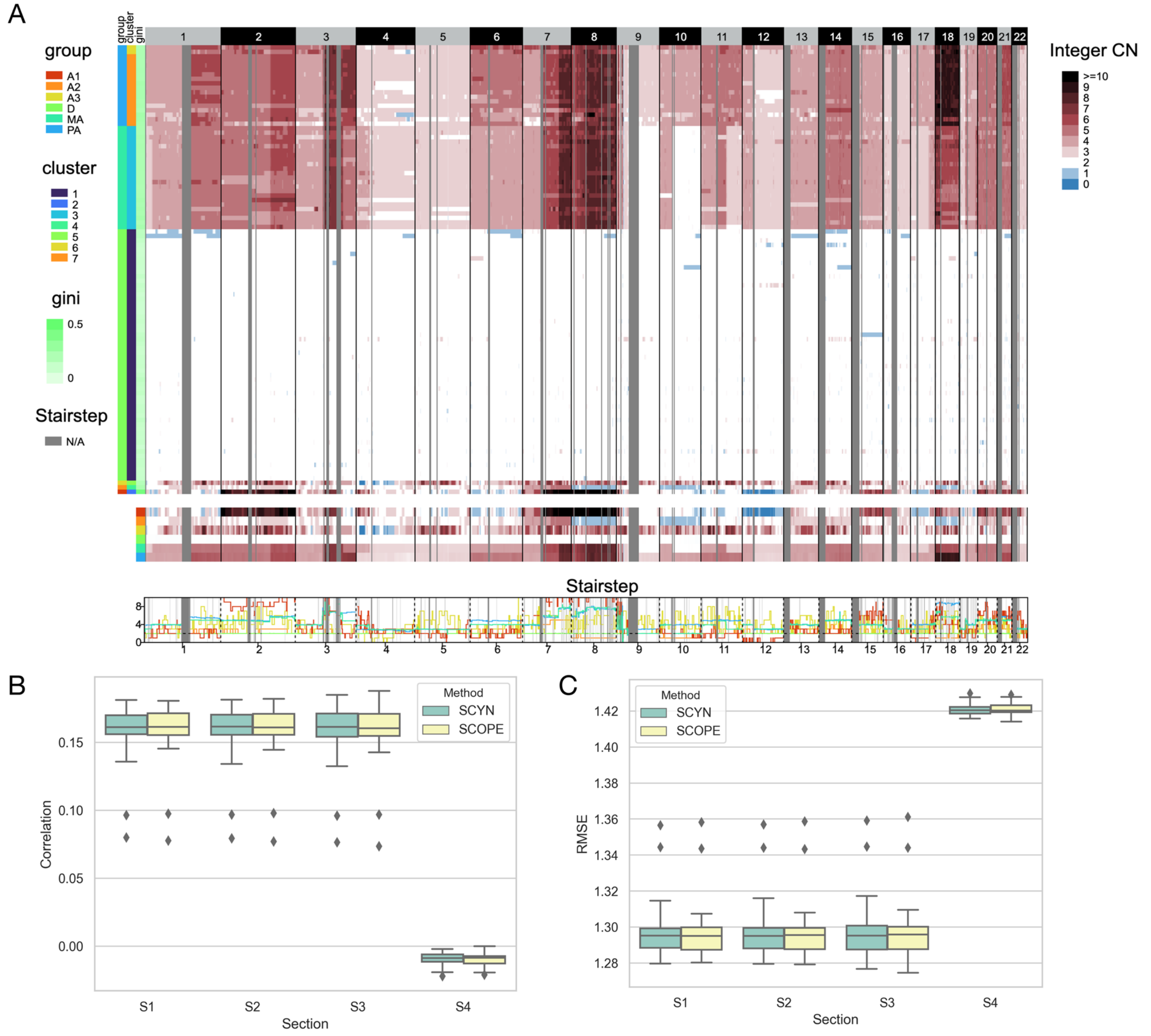


**Fig. 4** Performance of SCYN on T16. (A) Whole genome CNV heatmap, aggregated group CNV heatmap, and aggregated group CNV stairstep of T16. The colorbar (blue, white, dark red) represents the value of copy number from 0 to ≥10, respectively. N/A denotes the missing of copy number in corresponding genomic region. An interactive visualization is available on scSVAS [42] (https://sc.deepomics.org/oviz-project/analyses/view) with demo set "TNBC_T16". (B-C) Pearson correlation and root mean square error (RMSE) as evaluation metrics comparing results by SCYN and SCOPE against golden standard aGCH, respectively

time. As illustrated in Table 3 and Fig. 6, the CPU consuming time of SCYN is almost linear in log scale with the increase of cell number. However, the CPU time of SCOPE rises dramatically when the cell number goes to hundreds or thousands. For instance, for large datasets with 2k cells, SCYN is around 150 times faster than SCOPE, SCYN finished the tasks within eight minutes, while SCOPE is unable to scale 2k cells within 16 hours. In all, SCYN is super fast in respective of datasets scale up to hundreds or thousands.

### SCYN segmentation has better mBIC values

SCYN is fast because we only adopt the simplified version (Equation 1 in Method) of total SCOPE-mBIC [31] as the objective of segmentation and optimize it utilizing dynamic programming. Experiments on synthetic datasets and real cancer datasets successfully validated the tumor intra-heterogeneity exposure efficacy of SCYN against SCOPE. Here we further evaluate SCYN optimization effectiveness against SCOPE in respective of the original SCOPE-mBIC objective. We compared SCOPE-mBIC

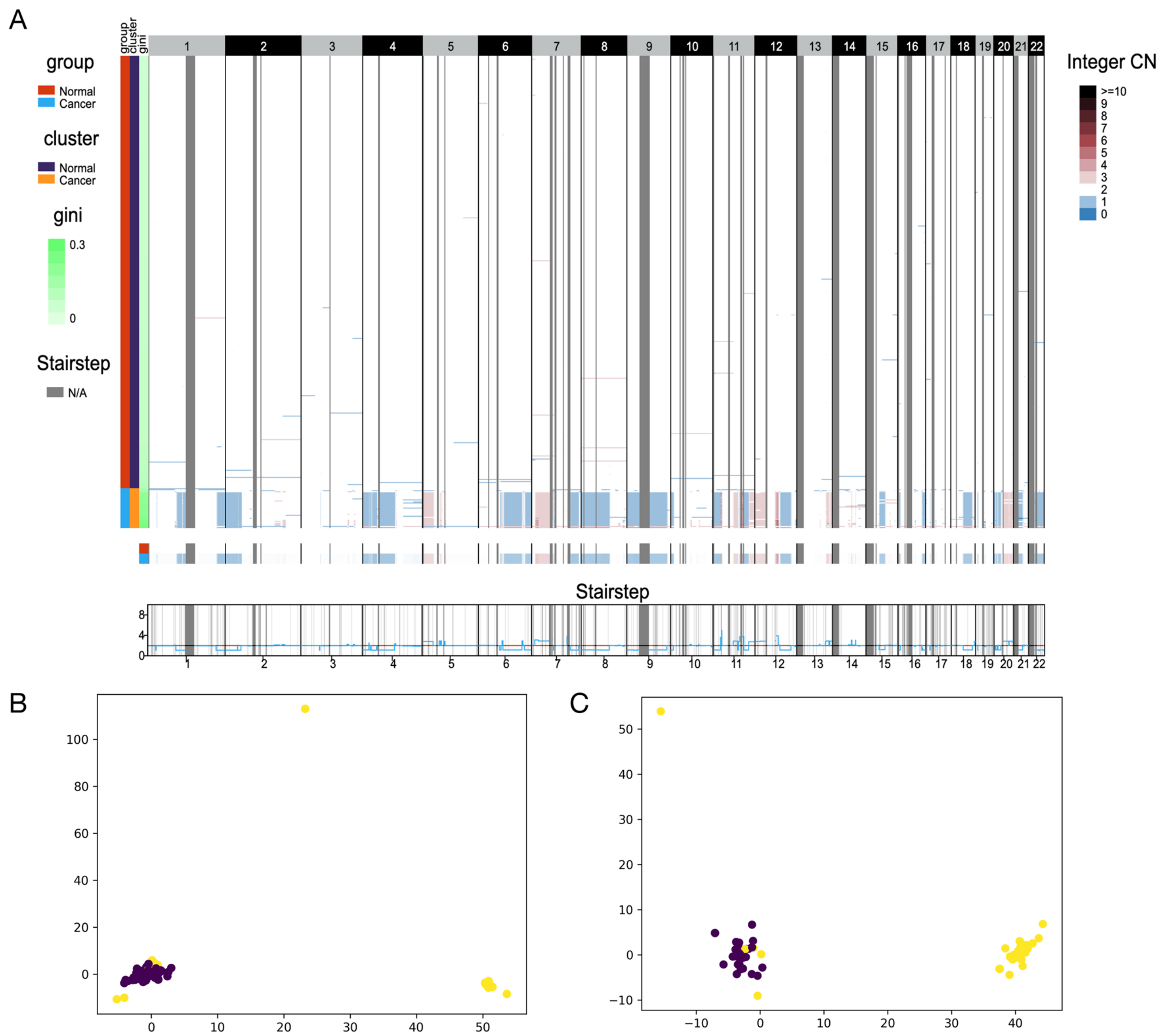


**Fig. 5** Performance of SCYN on 10x spike-ins. (A) Whole genome CNV heatmap, aggregated group CNV heatmap, and aggregated group CNV stairstep of 10% spike-in dataset. The colorbar (blue, white, dark red) represents the value of copy number from 0 to ≥10, respectively. N/A denotes the missing of copy number in corresponding genomic region. An interactive visualization is available on scSVAS [42] (https://sc.deepomics.org/oviz-project/analyses/view) with demo set "10x_10%spike-ins". (B-C) PCA plots on 1% and 10% spike-in datasets respectively. The yellow and purple dots denote cancer and normal cell, respectively

value by adopting the segmentation results of SCYN and SCOPE on real cancer datasets T10, T16P, T16M, and 10x spike-ins. As illustrated in Fig. 7A and Supplementary Figure S6A, the mBICs yielded from SCYN on samples across all chromosomes are always more massive than the mBICs produced by SCOPE, except chromosome 16 of 1% spike-in. Clearly, SCYN achieves better segmentation concerning the tedious SCOPE objective. Furthermore, as illustrated in Fig. 7B and Supplementary Figure S6B, the proportions of the simplified mBIC against overall SCOPE-mBICs are overwhelming across all chromosomes except chr16, indicating the residual terms actually can be neglected without loss of accuracy and the minor fluctuations of mBIC will not affect the ability of SCYN to detect subclones, as proved in the previous section.

## Discussion

In this study, we proposed SCYN, a fast and accurate dynamic programming approach for CNV segmentation and checkpoint detection customized for single cell DNA sequencing data. We demonstrated SCYN guaranteed to resolve the precise turning points on in silico dataset against SCOPE. Then we proved SCYN manifested a more accurate copy number inferring on triple-negative

**Table 1** 10x 1% spike-in datasets clustering evaluation of SCYN and SCOPE on adjusted Rand index (ARI), normalized mutual information (NMI), and Jaccard index (JI), respectively

| Method | ARI | NMI | JI |
|---|---|---|---|
| SCYN | 0.67650 | 0.7623 | 0.5238 |
| SCOPE | 0.67650 | 0.7623 | 0.5238 |

breast cancer scDNA data, with array CGH results of purified bulk samples as ground truth validation. Furthermore, we benchmarked SCYN against SCOPE on 10x Genomics CNV solution datasets. SCYN successfully recognizes gastric cancer cell spike-ins from diploid cells. Last but not least, SCYN is about 150 times faster than state of the art tool when dealing with thousands of cells. In conclusion, SCYN robustly and efficiently detects turning points and infers copy number profiles on single cell DNA sequencing data. It serves to reveal the tumor intra-heterogeneity.

SCYN is user-friendly. The implementation of SCYN is wrapped in python packages https://github.com/xikanfeng2/SCYN and available at PyPI. Users can easily run or call it with one line of command or Python code. For 10x merged BAM (One bam file), SCYN provides the function to split merged bam to cell bams based on the barcodes. SCYN outputs the segmented CNV profiles and cell meta-information for downstream analysis, such as hierarchical clustering and phylogeny reconstruction.

We neglected one crucial issue. Cancer scDNA-seq intensities should be regarded as a mixture of subclone cell signals with confounding of sparsity, GC bias, and amplification bias [31]. The perfect CNV segmentation heavily relies on the cross-cell normalization of intensities in the first place. While we brutely adopt the normalization schema from SCOPE; there lacks a comprehensive evaluation of scDNA intensities normalization. Speaking to further work, inferring CNV profiles from single-cell RNA sequencing (scRNA-seq) is trending [11–13, 38]. Incorporating DNA and RNA to profile single cell CNV segmentation might lead to tumor intra-heterogeneity to a higher resolution.

Copy number variation is crucial in deciphering the mechanism and cure of complex disorders and cancers. The recent advancement of scDNA sequencing technology sheds light upon addressing intratumor heterogeneity, detecting rare subclones, and reconstructing tumor evolution lineages at single-cell resolution. Nevertheless, the current circular binary segmentation based approach proves to fail to efficiently and effectively identify copy number shifts on some exceptional trails.

**Table 2** 10x 10% spike-in datasets clustering evaluation of SCYN and SCOPE on adjusted Rand index (ARI), normalized mutual information (NMI), and Jaccard index (JI), respectively

| Method | ARI | NMI | JI |
|---|---|---|---|
| SCYN | 0.9139 | 0.8770 | 0.8718 |
| SCOPE | 0.9139 | 0.8770 | 0.8718 |

**Table 3** Benchmark for CPU runtimes of checkpoint detection step (in Minutes). 90-1 and 90-2 are two in silico data with around 90 single-cells, and 2000-1, 2000-2, 2000-3, 2000-4, and 2000-5 are five in silico data with approximate 2000 single cells

| Sample | Cell Number | SCYN | SCOPE | Fold change on time |
|---|---|---|---|---|
| T10 | 99 | 2.917 | 46.995 | 16.111 |
| T16M | 48 | 2.566 | 21.94 | 8.55 |
| T16P | 52 | 2.786 | 23.927 | 8.588 |
| 90-1 | 93 | 2.73 | 44.14 | 16.168 |
| 90-2 | 92 | 2.769 | 40.415 | 14.596 |
| 10x-1% spike-in | 1056 | 3.598 | 485.768 | 135.011 |
| 10x-10% spike-in | 462 | 2.615 | 208.854 | 79.868 |
| 2000-1 | 2173 | 6.714 | 1147.658 | 170.935 |
| 2000-2 | 2214 | 7.602 | 1122.881 | 147.709 |
| 2000-3 | 1722 | 6.817 | 947.66 | 139.014 |
| 2000-4 | 1909 | 8.139 | 1122.335 | 137.896 |
| 2000-5 | 2048 | 7.128 | 1118.038 | 156.852 |

## Conclusion

To summarize, we propose SCYN, a CNV segmentation method powered with dynamic programming. Experiments on in silico and wet-lab data demonstrate that SCYN robustly and efficiently detects segmentations and infers copy number profiles on single cell DNA sequencing data. It serves to reveal the tumor intra-heterogeneity.

## Methods

### Data sets

### Synthetic data

The workflow for synthetic data generation is displayed in Fig. 2 A. Firstly, we pre-defined a CNV matrix which contains 100 cells and 70 bins for chromosome 22 and each bin has a fixed length of 500M. Also, 10% random noise is applied to this CNV matrix. Secondly, the bed file of each cell was generated according to the corresponding CNV number. Thirdly, we adopted the tool, SCSsim [32], which accepts the bed file as the input, to generate the DNA sequence of each cell(FASTQ format). Finally, the DNA sequence data of 100 cells were generated.

#### *Single-end real scDNA-seq data*

Two single-end breast cancer scDNA-seq datasets were downloaded from NCBI Sequence Read Archive with the SRA number of SRA018951. The raw fastq files were

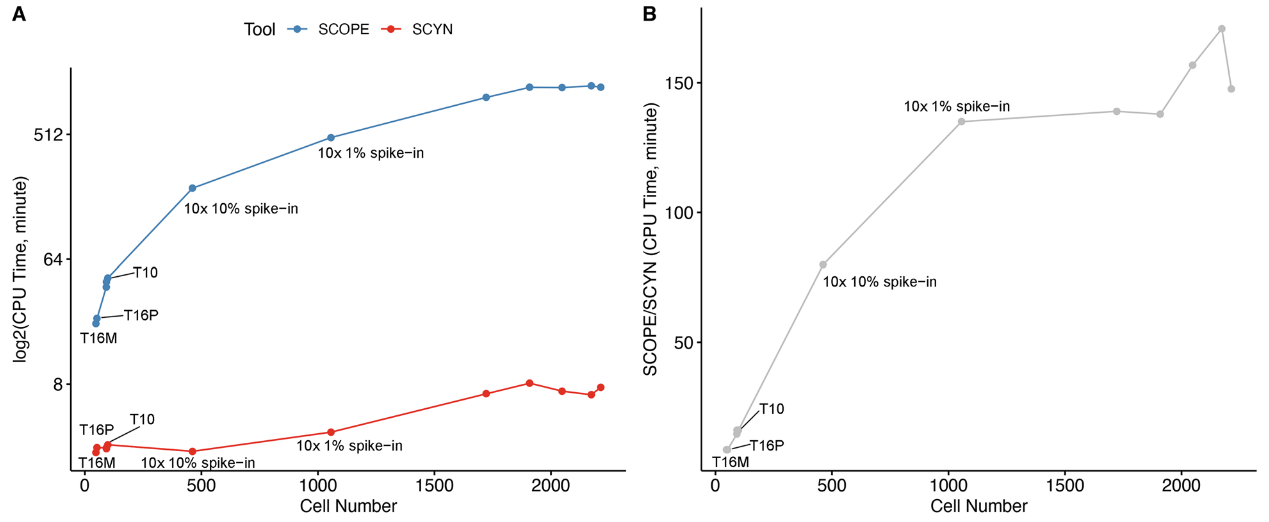


**Fig. 6** Runtime performance of SCYN turning-point detection. (A) CPU time of SCYN and SCOPE turning-point detection step on different cell number scale (48 to around 2000), respectively (Blue for SCOPE, red for SCYN). The CPU time visualized in log2-scale. (B) CPU time fold change of SCOPE against SCYN on different cell number scale (48 to around 2000)

aligned using BWA-mem [39] to the human hg19 reference genome, and the BAM files were sorted using SAMtools [40]. Picard toolkit [41] was used to remove duplicate reads. The clean BAM files were fed as the input of SCYN package.

### *Ten-x (10x) data*

The 10x spike-in scDNA-seq data was collected from the 10x Genomics official dataset with the accession link https://support.10xgenomics.com/single-cell-dna/datasets. The cell-mixed BAM files were demultiplexed to cellular BAMs according to cellular barcodes using Python scripts.

## Notations

To profile the CNV along genomes, first, we partition the genome into fix-size bins. Assume the number of bins as $m$. If the number of cells is $n$, then the input matrices, $Y_{m\times n}$ and $\hat{Y}_{m\times n}$, contain the raw and normalized reads counts, respectively; that is, $Y_{i,j}$ includes the number of raw reads count belong to bin $i$ at cell $j$ and $\hat{Y}_{m\times n}$ contains the number of normalized reads count belong to bin $i$ at cell $j$, where $1 \le i \le m$ and $1 \le j \le n$.

## Segmentation

The first task is to partitioning the bins into segments to optimize an objective function. Here, we choose the objective function to maximize the simplified version of modified Bayesian information criteria (mBIC) proposed by Wang et al. [31].

To calculate the simplified mBIC, we need to partition the sequence of bins into $\ell$ segments $s_1, ..., s_\ell$, where $s_k = (i_{k-1}+1, ..., i_k)$, $i_0 = 0 \le k_1 < k_2 < ... < k_\ell = n$. Denote the number of bins in segment $s_k$ as $|s_k|$ With the partitioning, we can calculate two matrices $X_{\ell\times n}$, $\hat{X}_{\ell\times n}$, where $X_{k,j} = \frac{1}{|s_k|}\sum_{i\in s_k} Y_{i,j}$, $\hat{X}_{k,j} = \frac{1}{|s_k|}\sum_{i\in s_k} \hat{Y}_{i,j}$, $1 \le k \le \ell$.

Given a segmentation $S = (s_1, ..., s_\ell)$, its simplified mBIC is calculated as

$$\beta(S) = \log\frac{L_\tau}{L_0} - \log\binom{m}{\ell-1} - (\ell-1)(\kappa_1 - \kappa_2) \quad (1)$$

where $\log\frac{L_\tau}{L_0}$ is the generalized log-likelihood ratio, $\kappa_1$ and $\kappa_2$ are two pre-defined constants and

$$\log\frac{L_\tau}{L_0} = \sum_{k=1}^{\ell}\hat{X}_k\left(1 - \frac{\lfloor 2X_k/\hat{X}_k\rceil}{2}\right) \quad (2)$$
$$+X_k\log\left(\frac{\lfloor 2X_k/\hat{X}_k\rceil}{2}\right)$$

For more details on the interpretation of the terms in mBIC, we refer the readers to Wang et al. [31]. Our objective here is to find a segmentation $S_{opt}$ such that $\beta(S_{opt})$ is maximized.

## Optimal algorithm

Let $\beta(k,i)$ store the simplified mBIC value for the optimal segmentation which partitions bins $1, ..., i$ into $k$ segments. Associated with $\beta(k,i)$, we also store the corresponding generalized log-likelihood ratio $L(k,i)$, which is the first term in Eq. 1, the log-likelihood ratio $l(i,j)$ for a single segment starting at the $i$-th bin and ending at the $j$-th bin, and

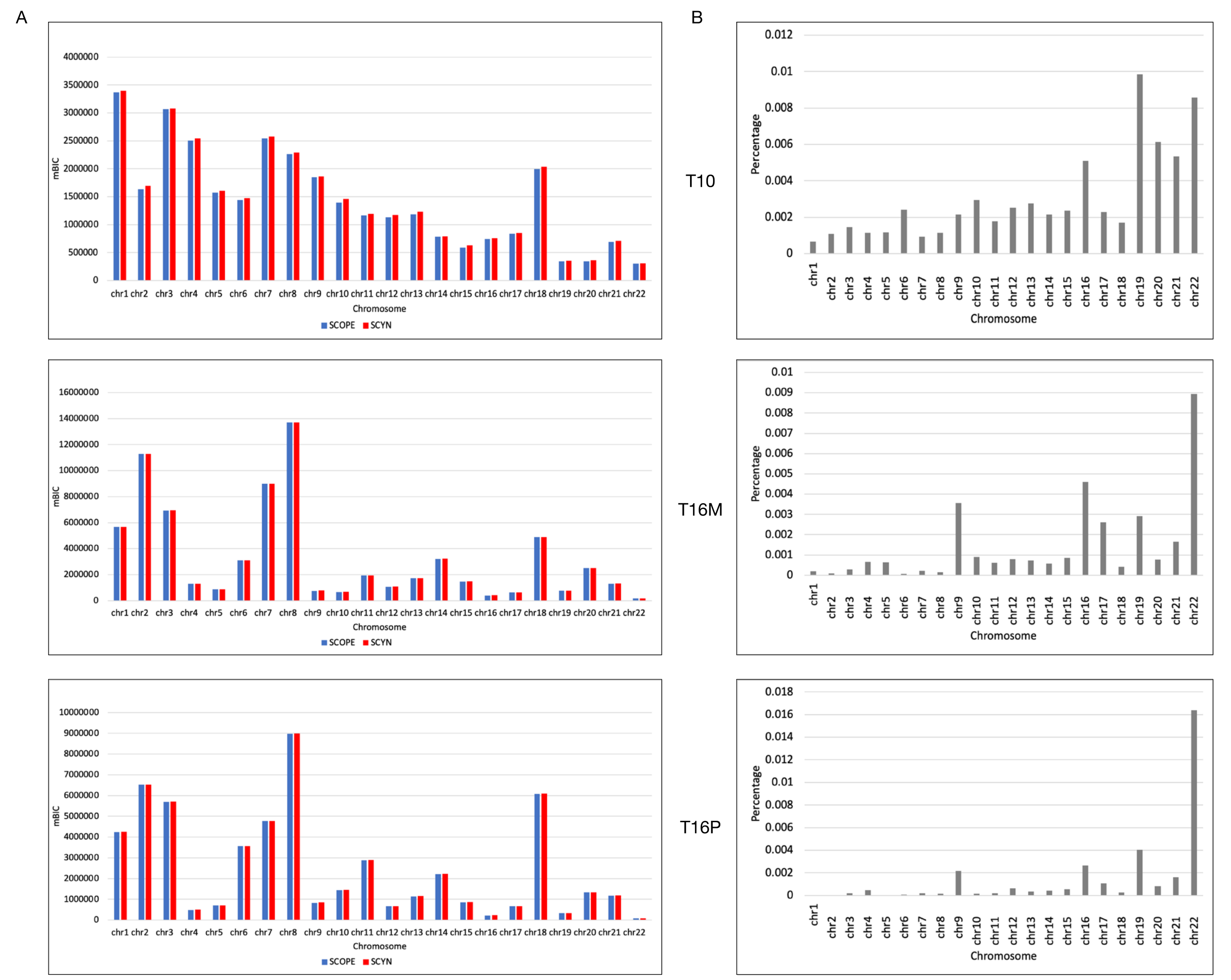


**Fig. 7** mBIC value performance of SCYN. (A) SCOPE-mBIC of T10, T16M and T16P across all chromosomes generated by SCYN and SCOPE, respectively (Blue for SCOPE, red for SCYN). (B) The proportion of residual terms over SCOPE-mBIC across all chromosomes on T10, T16M, and T16P, respectively

the $(k-1)$-th optimal turning point position $T(k-1,i)$ to partition bins $1,...,i$ into $k$ segments.

The $\beta(k,i)$ is calculated by the following recursive formulations:

$$\beta(k,i) = max_{1\leq i'<i}(L(k-1,i') + l(i'+1,...,i) + C) \tag{3}$$

$$L(k,i) = \arg\max_{i'}(\beta(k,i))L(k-1,i') + l(i'+1,...,i) \tag{4}$$

$$T(k-1,i) = \arg\max_{i'}(\beta(k,i)) \tag{5}$$

where $C$ is the sum of last two terms in Equation 1.

As demonstrated in Equation 3, the value of each cell $\beta(k,i)$ in table $\beta$ can be computed based on the earlier store data $L(k-1,i')$ and $l(i'+1,...,i)$. The computed $\beta(k,i)$ is then used to incrementally with $k$ and $i$ to compute the correct values of $\beta$. Clearly, the values of $\beta$ and $L$ for one segment can be initialized to equal to $l$.

The values of $\beta$ can be stored in a two dimensional array, i.e., a table. The procedure for computing the table $\beta$ is also displayed in Algorithm 1. The table $\beta$ will be constructed starting from a single segment $\beta(1,i)$, and moving towards more segments $\beta(k,i)$. The $\beta(1,i)$ and $L(1,i)$ are initialized to $l(1,i)$ and $T(0,i)$ is initialized to 0 when there is only one segment. When computing a cell $\beta(k,i)(k>1)$, we will checks all possible $i'$, $(k \leq i' < i)$ and compute all values of $(L(k-1,i') + l(i'+1,...,i) + C)$ and $\beta(k,i)$ is determined by $\max_{(L(k-1,i')+l(i'+1,...,i))+C}$. Processing the bins form in increasing order on length guarantees that the final optimal segmentation can be detected when $i$ is equal to the total number of bins $m$. At the last, the positions of $k-1$ turning points are stored in table $T$.

**Algorithm 1** Computing the table $\beta$

```
procedure COMPUTINGTHETABLEβ
    for segment number $k$ from 1 to pre-defined $K$ do
        for each bin $i$ from 1 to $m$ do
            if k == 1 then
                $\beta(1,i) = l(1,i)$
                $L(1,i) = l(1,i)$
                $T(0,i) = 0$
            else
                $\beta(k,i) = max_{1\leq i'<i}(L(k-1,i') + l(i'+1,...,i) + C)$
                $L(k,i) = \arg\max_{i'}(\beta(k,i))(L(k-1,i') + l(i'+1,...,i))$
                $T(k-1,i) = \arg\max_{i'}(\beta(k,i))$
            end if
        end for
    end for
end procedure
```

### Backtracking

The backtracking process of finding the positions of the optimal turning points is demonstrated in Fig. 1B. Let the table at the left-side of Fig. 1B as $T$, where $i$ and $j$ are the indexes of turning points and bins respectively. $T(i,j)$ is the position of the $i$-th optimal turning point for a segment $s(0,j)$. The optimal total turning points number is determined by the maximum value of $\beta(i,m)$, where m is the total number of bins. Then the positions of the optimal turning points can be found by the following formulation:

$$T(k-1,m) = \arg\max_{k} \beta(k,m) \qquad (6)$$

$$T(k-2,j) = T(k-2, T(k-1,m)-1) \qquad (7)$$

where k is the total segmentation number ($1 < k \leq K$), j is the index of bin and m is the total number of bins.

### Time complexity

The time complexity of this algorithm is $O(m^2n + m^2k)$, where m is the total bin number, n is the total cell number and k is the total segment number. The time complexity of calculating each $l(i,j)$ is $O(n)$ and we need to go over $O(m^2)$ possible segments for $m$ bins. Therefore we need to $O(m^2n)$ time to construct the table $l$. For a given segments number $k$, we need to calculate $O(m)$ possible $(L(k-1,i') + l(i'+1,...,i))$ values to get the maximum $L(k,i)$ for $m$ possible $i$, total $O(m^2)$ times. The time complexity for calculating the table $L$ is $O(m^2k)$. In conclusion, the time complexity of our algorithm is $O(m^2n + m^2k)$.

### Benchmark settings

SCOPE is a state-of-the-art tool for single cell CNV calling. We followed the steps in SCOPE README tutorial to perform the call CNV tasks in all datasets and the default parameters were used in all experiments. For SCYN, the function 'call()' was used and all parameters were set to default values. For running time analysis experiments, all experiments were run on a Dell server with an Intel(R) Xeon(R) CPU E5-2630 v3 with a clock speed of 2.40GHz. The mean value of 5 independent runs was regarded as the final running time for each tool.

### Abbreviations

CNV: Copy Number Variation; scDNA-seq: Single Cell DNA sequencing; scRNA-seq: Single Cell RNA sequencing; aCGH: array Comparative Genomic Hybridization; CBS: Circular Binary Segmentation; HMM: Hidden Markov Model; ARI: Adjusted Rand Index; NMI: Normalized Mutual Information; JI: Jaccard Index; mBIC: modified Bayesian information criteria

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07941-3.

**Additional file 1:** The PDF file includes all the supporting materials for the manuscript

### Acknowledgements

Not applicable.

### About this supplement

This article has been published as part of BMC Genomics Volume 22 Supplement 5 2021: Selected articles from the 19th Asia Pacific Bioinformatics Conference (APBC 2021): genomics The full contents of the supplement are available at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-22-supplement-5.

### Authors' contributions

S.C.L. conceived the idea and supervised the project. X.F, L.C, S.C.L discussed the algorithm and designed the experiments. X.F implemented the code and conducted the analysis. Y.Q, R.L, C.L visualized the CNV profiles. X.F, L.C, S.C.L drafted the manuscript. S.C.L revised the manuscript. All authors read and approved the final manuscript.

### Funding

This work and publication costs are funded by the GRF Research Projects 9042348 (CityU 11257316) and the Fundamental Research Funds for the Central Universities under the Grant G2020KY05109. The funding body did not play any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

### Availability of data and materials

The synthetic data and source code included in this study can be found in https://github.com/xikanfeng2/SCYN. Two single-end breast cancer scDNA-seq datasets were downloaded from NCBI Sequence Read Archive with the SRA number of SRA018951. The 10x spike-in scDNA-seq data was collected from the 10x Genomics official dataset with the accession link https://support.10xgenomics.com/single-cell-dna/datasets.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

**Author details**
[1]School of Software, Northwestern Polytechnical University, Xi'an Shaanxi, 710072, China. [2]Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China. [3]Department of Biomedical Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China.

Received: 17 August 2021 Accepted: 20 August 2021
Published online: 16 November 2021

## References

1. Levy D, Ronemus M, Yamrom B, Lee Y-h, Leotta A, Kendall J, Marks S, Lakshmi B, Pai D, Ye K, et al. Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron. 2011;70(5):886–97.
2. Marshall C, Noor A, Vincent J, Lionel A, Feuk L, Skaug J, Shago M, Moessner R, Pinto D, Ren Y, et al. Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet. 2008;82(2):477–88.
3. La Cognata V, Morello G, D'Agata V, Cavallaro S. Copy number variability in parkinson's disease: assembling the puzzle through a systems biology approach. Hum Genet. 2017;136(1):13–37.
4. Helbig I, Mefford H, Sharp A, Guipponi M, Fichera M, Franke A, Muhle H, De Kovel C, Baker C, Von Spiczak S, et al. 15q13. 3 microdeletions increase risk of idiopathic generalized epilepsy. Nat Genet. 2009;41(2):160.
5. Elia J, Gai X, Xie H, Perin J, Geiger E, Glessner J, D'arcy M, Deberardinis R, Frackelton E, Kim C, et al. Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. Mol Psychiatry. 2010;15(6):637.
6. George J, Saito M, Tsuta K, Iwakawa R, Shiraishi K, Scheel A, Uchida S, Watanabe S-i, Nishikawa R, Noguchi M, et al. Genomic amplification of cd274 (pd-l1) in small-cell lung cancer. Clin Cancer Res. 2017;23(5):1220–6.
7. Ulz P, Heitzer E, Speicher M. Co-occurrence of myc amplification and tp53 mutations in human cancer. Nat Genet. 2016;48(2):104.
8. Ler L, Ghosh S, Chai X, Thike A, Heng H, Siew E, Dey S, Koh L, Lim J, Lim W, et al. Loss of tumor suppressor kdm6a amplifies prc2-regulated transcriptional repression in bladder cancer and can be targeted through inhibition of ezh2. Sci Transl Med. 2017;9(378):8312.
9. Simó-Riudalbas L, Pérez-Salvia M, Setien F, Villanueva A, Moutinho C, Martínez-Cardús A, Moran S, Berdasco M, Gomez A, Vidal E, et al. Kat6b is a tumor suppressor histone h3 lysine 23 acetyltransferase undergoing genomic loss in small cell lung cancer. Cancer Res. 2015;75(18):3936–45.
10. Harmanci A, Harmanci A, Zhou X. Casper identifies and visualizes cnv events by integrative analysis of single-cell or bulk rna-sequencing data. Nat Commun. 2020;11(1):1–16.
11. Patel A, Tirosh I, Trombetta J, Shalek A, Gillespie S, Wakimoto H, Cahill D, Nahed B, Curry W, Martuza R, et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344(6190):1396–401.
12. Tirosh I, Izar B, Prakadan S, Wadsworth M, Treacy D, Trombetta J, Rotem A, Rodman C, Lian C, Murphy G, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. Science. 2016;352(6282):189–96.
13. Puram S, Tirosh I, Parikh A, Patel A, Yizhak K, Gillespie S, Rodman C, Luo C, Mroz E, Emerick K, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell. 2017;171(7):1611–24.
14. Pinkel D, Albertson D. Array comparative genomic hybridization and its applications in cancer. Nat Genet. 2005;37(6s):11.
15. Emmanuel C, Chiew Y-E, George J, Etemadmoghadam D, Anglesio M, Sharma R, Russell P, Kennedy C, Fereday S, Hung J, et al. Genomic classification of serous ovarian cancer with adjacent borderline differentiates ras pathway and tp53-mutant tumors and identifies nras as an oncogenic driver. Clin Cancer Res. 2014;20(24):6618–30.
16. Savas P, Teo Z, Lefevre C, Flensburg C, Caramia F, Alsop K, Mansour M, Francis P, Thorne H, Silva M, et al. The subclonal architecture of metastatic breast cancer: results from a prospective community-based rapid autopsy program "cascade". PLoS Med. 2016;13(12):1002204.
17. Mayrhofer M, DiLorenzo S, Isaksson A. Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. Genome Biol. 2013;14(3):24.
18. Trost B, Walker S, Wang Z, Thiruvahindrapuram B, MacDonald J, Sung W, Pereira S, Whitney J, Chan A, Pellecchia G, et al. A comprehensive workflow for read depth-based identification of copy-number variation from whole-genome sequence data. Am J Hum Genet. 2018;102(1): 142–55.
19. Velazquez-Villarreal E, Maheshwari S, Sorenson J, Fiddes I, Kumar V, Yin Y, Webb M, Catalanotti C, Grigorova M, Edwards P, et al. Single-cell sequencing of genomic dna resolves sub-clonal heterogeneity in a melanoma cell line. Commun Biol. 2020;3(1):1–8.
20. Martelotto L, Baslan T, Kendall J, Geyer F, Burke K, Spraggon L, Piscuoglio S, Chadalavada K, Nanjangud G, Ng C, et al. Whole-genome single-cell copy number profiling from formalin-fixed paraffin-embedded samples. Nat Med. 2017;23(3):376.
21. Eastburn D, Pellegrino M, Sciambi A, Treusch S, Xu L, Durruthy-Durruthy R, Gokhale K, Jacob J, Chen T, Oldham W, et al. Single-cell analysis of mutational heterogeneity in acute myeloid leukemia tumors with high-throughput droplet microfluidics. AACR. 2018;78(13):.
22. Andor N, Lau B, Catalanotti C, Sathe A, Kubit M, Chen J, Blaj C, Cherry A, Bangs C, Grimes S, et al. Joint single cell dna-seq and rna-seq of gastric cancer cell lines reveals rules of in vitro evolution. NAR Genomics Bioinforma. 2020;2(2):016.
23. Gao Y, Ni X, Guo H, Su Z, Ba Y, Tong Z, Guo Z, Yao X, Chen X, Yin J, et al. Single-cell sequencing deciphers a convergent evolution of copy number alterations from primary to circulating tumor cells. Genome Res. 2017;27(8):1312–22.
24. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (cnv) detection using next-generation sequencing data: features and perspectives. BMC Bioinforma. 2013;14(11):1.
25. Olshen A, Venkatraman E, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based dna copy number data. Biostatistics. 2004;5(4):557–72.
26. Venkatraman E, Olshen A. A faster circular binary segmentation algorithm for the analysis of array cgh data. Bioinformatics. 2007;23(6):657–63.
27. Shah S, Xuan X, DeLeeuw R, Khojasteh M, Lam W, Ng R, Murphy K. Integrating copy number polymorphisms into array cgh analysis using a robust hmm. Bioinformatics. 2006;22(14):431–9.
28. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant S, Hakonarson H, Bucan M. Penncnv: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. Genome Res. 2007;17(11):1665–74.
29. Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal G, Hicks J, Wigler M, Schatz M. Interactive analysis and assessment of single-cell copy-number variations. Nat Methods. 2015;12(11):1058.
30. Wang X, Chen H, Zhang N. Dna copy number profiling using single-cell sequencing. Brief Bioinform. 2017;19(5):731–6.
31. Wang R, Lin D-Y, Jiang Y. Scope: A normalization and copy-number estimation method for single-cell dna sequencing. Cell Syst. 2020;10(5): 445–52.
32. Yu Z, Du F, Sun X, Li A. Scssim: an integrated tool for simulating single-cell genome sequencing data. Bioinformatics. 2020;36(4):1281–2.
33. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472(7341):90.
34. Navin N, Krasnitz A, Rodgers L, Cook K, Meth J, Kendall J, Riggs M, Eberling Y, Troge J, Grubor V, et al. Inferring tumor progression from genomic heterogeneity. Genome Res. 2010;20(1):68–80.
35. Rand W. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66(336):846–50.
36. Cover TM, Thomas JA. Elements of information theory, 2nd ed. New York: Wiley; 2006.
37. Hamers L, et al. Similarity measures in scientometric research: The jaccard index versus salton's cosine formula. Inf Process Manag. 1989;25(3): 315–18.
38. Fan J, Lee H-O, Lee S, Ryu D-e, Lee S, Xue C, Kim S, Kim K, Barkas N, Park P, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell rna-seq data. Genome Res. 2018;28(8):1217–27.
39. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. bioinformatics. 2009;25(14):1754–60.
40. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and samtools. Bioinformatics. 2009;25(16):2078–9.
41. Picard Toolkit. Broad Institute, GitHub Repository: Broad Institute; 2019.

42. Chen L, Qing Y, Li R, Li C, Li H, Feng X, Li S. scsvas: Cnv clonal visualization online platform for large scale single-cell genomics. bioRxiv. 2021.

## Publisher's Note

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