SCYN: Single cell CNV profiling method using dynamic programming

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

[1]Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China
[2]Department of Biomedical Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China

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. Here, we propose SCYN, a CNV segmentation method powered with dynamic programming. SCYN resolves the precise segmentation on two *in silico* datasets. 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. 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`. The visualization tools are hosted on `https://sc.deepomics.org/`.

Keywords: single cell DNA sequencing, CNV segmentation, dynamic programming
†: Equal contributor
Correspondence: ***shuaicli@cityu.edu.hk***

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). Moreover, scientists may leverage bulk genome 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 tumor biology, the data they provide may mask a degree of heterogeneity (19). For instance, if the averaged read-out 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 (scDNA) sequencing delivers a potential solution (20–22). Researchers can overwhelm the deficiencies of bulk approaches 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* experiments, 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, an effecient and effective dynamic programming approach for single cell data CNV segmentation and checkpoint detection. SCYN resolves the precise turning points on two *in silico* datasets, 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 **Figure 1A**. First, the raw scDNA-seq reads of FASTQ format are pre-processed with standard procedures (see **Figure 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 breakpoints on synthetic trials. To evaluate the segmentation power of SYCN against SCOPE, we generated two different combinations of the CNV intensities of blue cell and orange cell along 200 bin regions. In the first simulation, the ground truth segmentation are (1, ..., 49), (50, ..., 99) (100,..., 149), (150,..., 200); and the copy number state alternates between haploid and diploid. Figure 2 shows the SCOPE unable to detect the turning point 100 here, leading to erroneously dropping the loss of heterogeneity event of bin range [100, 123]. In contrast with SCOPE, SCYN accurately detected all turning points and assigned the correct copy number to all bin regions. Then, with fixed copy number turning points (50, 100, and 150) and copy number state alternates between one and four, we simulated the situation where blue cell and orange cell are always heterogeneous. In Figure 2, SCYN successfully categorized all turning points and copy number states with 100% accuracy and uncovered the cell heterogeneity. Even though SCOPE assigned correct copy number to each bin region, we found that it output five turning points 50, 100, 143, 146, and 150. In other words, SCOPE considered there exited consecutive copy number shifts among bin ranges [101, 143], [144, 146], and [147, 150], which opposite against the homogeneous fact. As previously mentioned, the core principle of CNV segmentation is partitioning the genome into 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, these two experiments 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) (32) 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 (32). 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 (see Additional file 1, Supplementary Figure S1A) and Navin *et al.*'s findings. With hierarchical clustering, SCYN categorizes T10 into seven clusters. As illustrated in Figure3 and Additional file 1 Supplementary Figure S2A-3A, 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) (33), here we regarded the CNV profiled from aCGH as golden-standard to examine the SYCN and SCOPE performance. As illustrated in Figure 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 (see Additional file 1, Supplementary Figure S1B) and Navin *et al.*'s findings. Hierarchical clustering characterizes T16 into seven subgroups. As depicted in Figure4 and Additional file 1 Supplementary Figure S2B-3B, 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 (33), 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 Figure 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 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 Figure5A and Additional file 1 Supplementary Figure S4, SCOPE successfully distinguished the two spike-in gastric cancer cells. Furthermore, we visualized the first two principal components of the estimated CNV profiles in Figure5B-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 clustering accuracy of SYCN against SCOPE. The adjusted Rand index (ARI) (34), Normalized mutual information (NMI) (35), and Jaccard index (JI) (36) 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 Table 1 and Table 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 segmentation task 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 time. As illustrated in Table 3 and Figure 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 (see 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 value by adopting the segmentation results of SCYN and SCOPE on real cancer datasets T10, T16P, T16M, and 10x spike-ins. As illustrated in Figure 7A and Supplementary Figure S5A, 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 Figure 7B and Supplementary Figure S5B, the proportions of the simplified mBIC against overall SCOPE-mBICs are overwhelming across all chromosomes, indicating all residual terms actually can be neglected without loss of accuracy. SCYN produced smaller mBIC values than SCOPE on chromosome 16 for 1% spike-in dataset, suggesting that the residual terms take effect on circumstances such as the tiny proportion of cancer cells. However, we believe that the 1% spike case is rare in scDNA sequencing samples and is invalid for downstream analysis, 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 two *in silico* datasets against SCOPE. Then we proved SCYN manifested a more accurate copy number inferring on triple-negative 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.

The implementation of SCYN is wrapped in python packages https://github.com/xikanfeng2/SCYN. It provides the segmented CNV profiles and cell meta-information available for downstream analysis, such as hierarchical clustering and phylogeny reconstruction. Last but not least, the CNV profiles obtained from SCYN can be directly visualized in https://sc.deepomics.org/, which supports real-time interaction and literature-style figure downloading.

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, 37). Incorporating DNA and RNA to profile single cell CNV segmentation might lead to tumor intra-heterogeneity to a higher resolution.

Methods

Data sets.

Synthetic data. Two synthetic datasets were generated to evaluate the segmentation power of SCYN. The dimension of each dataset is 400 bins and two cells. The ground truth segmentation is (1, ..., 49), (50, ..., 99) (100,..., 149), (150,..., 200) for both of datasets. For the first dataset, the reads count of two cells for the four segments was designed to around (100, 100), (400, 400), (100, 100) and (400, 400), respectively. For the second dataset, the reads count of two cells for the four segments was designed to around (100, 400), (400, 100), (100, 400) and (400, 100), respectively. Random noise was applied to these reads counts.

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 aligned using BWA-mem (38) to the human hg19 reference genome, and the BAM files were sorted using SAMtools (39). Picard toolkit (40) 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 $1 \leq i \leq m$ and $1 \leq j \leq n$.

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 ℓ segments $s_1,...,s_\ell$, where $s_k = (i_{k-1}+1,...,i_k)$, $k_0 = 0 \leq k_1 < k_2 < ... < k_\ell = n$. 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 $X_{k,j} = \frac{1}{|s_k|}\sum_{i\in s_k} Y_{i,j}$, $\hat{X}_{k,j} = \frac{1}{|s_k|}\sum_{i\in s_k} \hat{Y}_{i,j}$, $1 \leq k \leq \ell$. Given a segmentation $S = (s_1,...,s_\ell)$, its simplified mBIC is calculated as

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

where $\log\frac{L_\tau}{L_0}$ is the generalized log-likelihood ratio, κ_1 and κ_2 are two pre-defined constants and

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

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 **Equation 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 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:

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

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

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

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 β 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 β. Clearly, the values of β and L for one segment can be initialized to equal to l.
The values of β can be stored in a two dimensional array, i.e., a table. The procedure for computing the table β is also displayed in **Algorithm 1**. The table β 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.

Backtracking. The backtracking process of finding the positions of the optimal turning points is demonstrated in **Figure 1B**. Let the table at the left-side of **Figure 1B** as T, where i

Algorithm 1 Computing the table β

```
procedure COMPUTINGTHETABLE$\beta$
    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
```

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:

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

$$T(k-2,j) = T(k-2,T(k-1,m)-1) \tag{7}$$

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 $(L(k1,i') + l(i'+1,...,i))$ 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.

Availability of data and materials

The data and source code included in this study can be found in https://github.com/xikanfeng2/SCYN. The visualisation tools are hosted on https://sc.deepomics.org/.

List of 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

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Funding

This work are funded by the GRF Research Projects 9042348 (CityU 11257316).

Acknowledgements

Not applicable.

Author's contributions

SCL. conceived the idea and supervised the project.
XF, LC, SCL discussed the algorithm and designed the experiments.
XF implemented the code and conducted the analysis.
YQ, RL, CL visualized the CNV profiles.
LC, XF drafted the manuscript.
SCL revised the manuscript.
All authors read and approved the final manuscript.

Bibliography

1. Dan Levy, Michael Ronemus, Boris Yamrom, Yoon-ha Lee, Anthony Leotta, Jude Kendall, Steven Marks, B Lakshmi, Deepa Pai, Kenny Ye, et al. Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron*, 70(5):886–897, 2011.
2. Christian R Marshall, Abdul Noor, John B Vincent, Anath C Lionel, Lars Feuk, Jennifer Skaug, Mary Shago, Rainald Moessner, Dalila Pinto, Yan Ren, et al. Structural variation of chromosomes in autism spectrum disorder. *The American Journal of Human Genetics*, 82 (2):477–488, 2008.
3. Valentina La Cognata, Giovanna Morello, Velia D'Agata, and Sebastiano Cavallaro. Copy number variability in parkinson's disease: assembling the puzzle through a systems biology approach. *Human genetics*, 136(1):13–37, 2017.
4. Ingo Helbig, Heather C Mefford, Andrew J Sharp, Michel Guipponi, Marco Fichera, Andre Franke, Hiltrud Muhle, Carolien De Kovel, Carl Baker, Sarah Von Spiczak, et al. 15q13. 3 microdeletions increase risk of idiopathic generalized epilepsy. *Nature genetics*, 41(2):160, 2009.
5. J Elia, X Gai, HM Xie, JC Perin, E Geiger, JT Glessner, M D'arcy, R Deberardinis, E Frackelton, C Kim, et al. Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. *Molecular psychiatry*, 15(6): 637, 2010.
6. Julie George, Motonobu Saito, Koji Tsuta, Reika Iwakawa, Kouya Shiraishi, Andreas H Scheel, Shinsuke Uchida, Shun-ichi Watanabe, Ryo Nishikawa, Masayuki Noguchi, et al. Genomic amplification of cd274 (pd-l1) in small-cell lung cancer. *Clinical Cancer Research*, 23(5):1220–1226, 2017.
7. Peter Ulz, Ellen Heitzer, and Michael R Speicher. Co-occurrence of myc amplification and tp53 mutations in human cancer. *Nature genetics*, 48(2):104, 2016.
8. Lian Dee Ler, Sujoy Ghosh, Xiaoran Chai, Aye Aye Thike, Hong Lee Heng, Ee Yan Siew, Sucharita Dey, Liang Kai Koh, Jing Quan Lim, Weng Khong Lim, et al. Loss of tumor suppressor kdm6a amplifies prc2-regulated transcriptional repression in bladder cancer and can be targeted through inhibition of ezh2. *Science translational medicine*, 9(378):eaai8312, 2017.
9. Laia Simó-Riudalbas, Montserrat Pérez-Salvia, Fernando Setien, Alberto Villanueva, Catia Moutinho, Anna Martínez-Cardús, Sebastian Moran, Maria Berdasco, Antonio Gomez, Enrique Vidal, et al. Kat6b is a tumor suppressor histone h3 lysine 23 acetyltransferase undergoing genomic loss in small cell lung cancer. *Cancer research*, 75(18):3936–3945, 2015.
10. Eric Talevich and Alan Hunter Shain. Cnvkit-rna: Copy number inference from rna-sequencing data. *bioRxiv*, page 408534, 2018.
11. Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344 (6190):1396–1401, 2014.
12. Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.
13. Sidharth V Puram, Itay Tirosh, Anuraag S Parikh, Anoop P Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, Christina L Luo, Edmund A Mroz, Kevin S Emerick, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, 171(7):1611–1624, 2017.
14. Daniel Pinkel and Donna G Albertson. Array comparative genomic hybridization and its applications in cancer. *Nature genetics*, 37(6s):S11, 2005.
15. Catherine Emmanuel, Yoke-Eng Chiew, Joshy George, Dariush Etemadmoghadam, Michael S Anglesio, Raghwa Sharma, Peter Russell, Catherine Kennedy, Sian Fereday, Jillian Hung, 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. *Clinical cancer research*, 20(24):6618–6630, 2014.
16. Peter Savas, Zhi Ling Teo, Christophe Lefevre, Christoffer Flensburg, Franco Caramia, Kathryn Alsop, Mariam Mansour, Prudence A Francis, Heather A Thorne, Maria Joao Silva, et al. The subclonal architecture of metastatic breast cancer: results from a prospective community-based rapid autopsy program "cascade". *PLoS medicine*, 13(12):e1002204, 2016.
17. Markus Mayrhofer, Sebastian DiLorenzo, and Anders Isaksson. Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. *Genome biology*, 14(3): R24, 2013.
18. Brett Trost, Susan Walker, Zhuozhi Wang, Bhooma Thiruvahindrapuram, Jeffrey R MacDonald, Wilson WL Sung, Sergio L Pereira, Joe Whitney, Ada JS Chan, Giovanna Pellecchia, et al. A comprehensive workflow for read depth-based identification of copy-number variation from whole-genome sequence data. *The American Journal of Human Genetics*, 102 (1):142–155, 2018.
19. Enrique I Velazquez-Villarreal, Shamoni Maheshwari, Jon Sorenson, Ian T Fiddes, Vijay Kumar, Yifeng Yin, Michelle Webb, Claudia Catalanotti, Mira Grigorova, Paul A Edwards, et al. Resolving sub-clonal heterogeneity within cell-line growths by single cell sequencing genomic dna. *bioRxiv*, page 757211, 2019.
20. Luciano G Martelotto, Timour Baslan, Jude Kendall, Felipe C Geyer, Kathleen A Burke, Lee Spraggon, Salvatore Piscuoglio, Kalyani Chadalavada, Gouri Nanjangud, Charlotte KY Ng, et al. Whole-genome single-cell copy number profiling from formalin-fixed paraffin-embedded samples. *Nature medicine*, 23(3):376, 2017.
21. Dennis J Eastburn, Maurizio Pellegrino, Adam Sciambi, Sebastian Treusch, Liwen Xu, Robert Durruthy-Durruthy, Kaustubh Gokhale, Jose Jacob, Tina X Chen, William Oldham, et al. Single-cell analysis of mutational heterogeneity in acute myeloid leukemia tumors with high-throughput droplet microfluidics, 2018.
22. Noemi Andor, Billy T Lau, Claudia Catalanotti, Vijay Kumar, Anuja Sathe, Kamila Belhocine, Tobias D Wheeler, Andrew D Price, Maengseok Kang, David Stafford, et al. Joint single cell dna-seq and rna-seq of gastric cancer reveals subclonal signatures of genomic instability and gene expression. *bioRxiv*, page 445932, 2018.
23. Yan Gao, Xiaohui Ni, Hua Guo, Zhe Su, Yi Ba, Zhongsheng Tong, Zhi Guo, Xin Yao, Xixi Chen, Jian Yin, et al. Single-cell sequencing deciphers a convergent evolution of copy number alterations from primary to circulating tumor cells. *Genome research*, 27(8):1312–1322, 2017.
24. Min Zhao, Qingguo Wang, Quan Wang, Peilin Jia, and Zhongming Zhao. Computational tools for copy number variation (cnv) detection using next-generation sequencing data: features and perspectives. *BMC bioinformatics*, 14(11):S1, 2013.
25. Adam B Olshen, ES Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572, 2004.
26. ES Venkatraman and Adam B Olshen. A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, 23(6):657–663, 2007.
27. Sohrab P Shah, Xiang Xuan, Ron J DeLeeuw, Mehrnoush Khojasteh, Wan L Lam, Raymond Ng, and Kevin P Murphy. Integrating copy number polymorphisms into array cgh analysis using a robust hmm. *Bioinformatics*, 22(14):e431–e439, 2006.
28. Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan FA Grant, Hakon Hakonarson, and Maja Bucan. Penncnv: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. *Genome research*, 17(11):1665–1674, 2007.
29. Tyler Garvin, Robert Aboukhalil, Jude Kendall, Timour Baslan, Gurinder S Atwal, James Hicks, Michael Wigler, and Michael C Schatz. Interactive analysis and assessment of single-cell copy-number variations. *Nature methods*, 12(11):1058, 2015.
30. Xuefeng Wang, Hao Chen, and Nancy R Zhang. Dna copy number profiling using single-cell sequencing. *Briefings in bioinformatics*, 19(5):731–736, 2017.
31. Rujin Wang, Dan-Yu Lin, and Yuchao Jiang. Scope: a normalization and copy number estimation method for single-cell dna sequencing. *bioRxiv*, page 594267, 2019.
32. Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90, 2011.
33. Nicholas Navin, Alexander Krasnitz, Linda Rodgers, Kerry Cook, Jennifer Meth, Jude Kendall, Michael Riggs, Yvonne Eberling, Jennifer Troge, Vladimir Grubor, et al. Inferring tumor progression from genomic heterogeneity. *Genome research*, 20(1):68–80, 2010.
34. William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.
35. Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
36. Lieve Hamers et al. Similarity measures in scientometric research: The jaccard index versus salton's cosine formula. *Information Processing and Management*, 25(3):315–18, 1989.
37. Jean Fan, Hae-Ock Lee, Soohyun Lee, Da-eun Ryu, Semin Lee, Catherine Xue, Seok Jin Kim, Kihyun Kim, Nikolaos Barkas, Peter J Park, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell rna-seq data. *Genome research*, 28(8):1217–1227, 2018.
38. Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.
39. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.
40. Picard toolkit. http://broadinstitute.github.io/picard/, 2019.

Figures

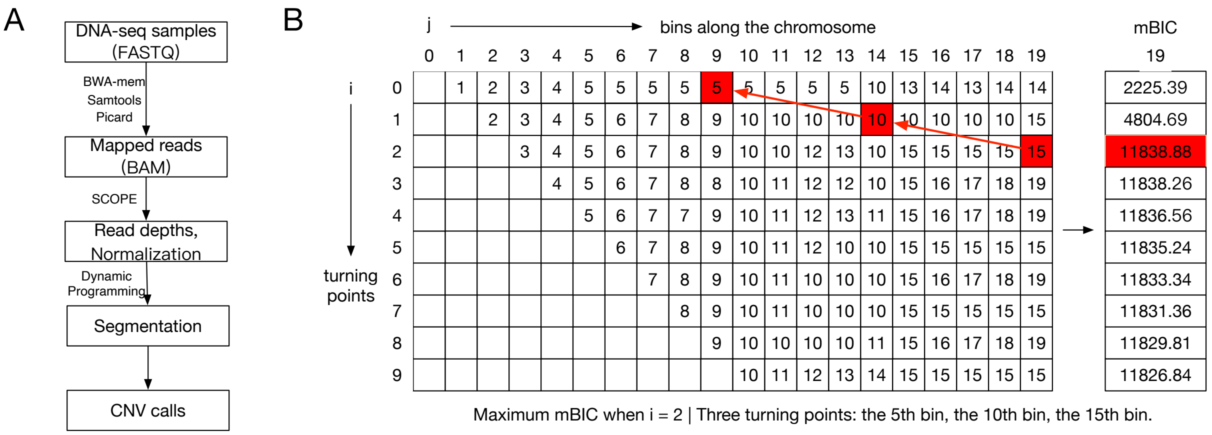
Tables

Table 1. 10x 1% spike-in datasets clustering evaluation

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

Table 2. 10x 10% spike-in datasets clustering evaluation

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



	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

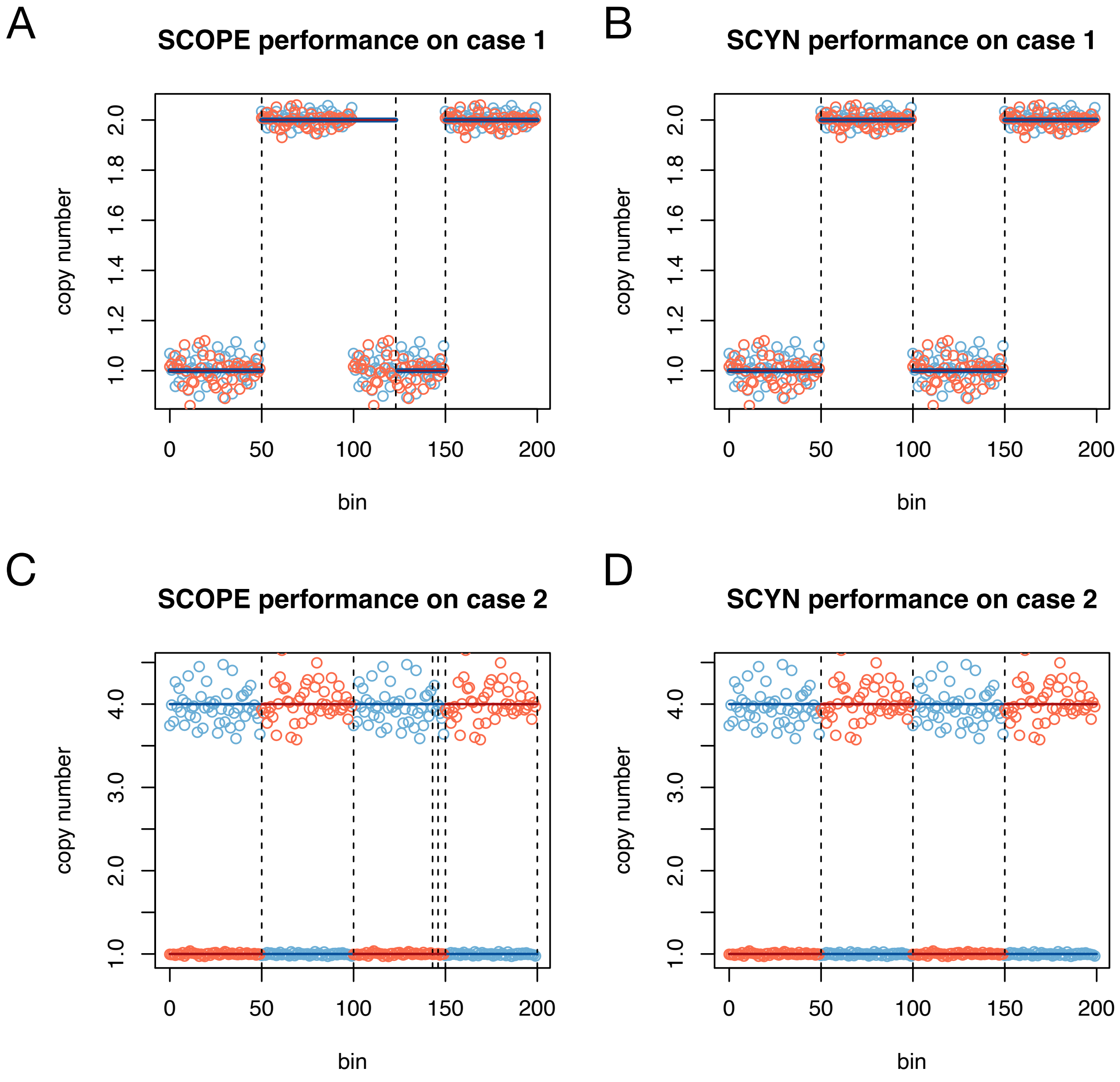


Fig. 2. SCYN performance on synthetic cases
Hollow circles and horizontal lines denote the copy number before and after smoothing respectively. Vertical dashed lines signify the detected turning points. Orange and blue refer to cell 1 and cell 2 respectively.

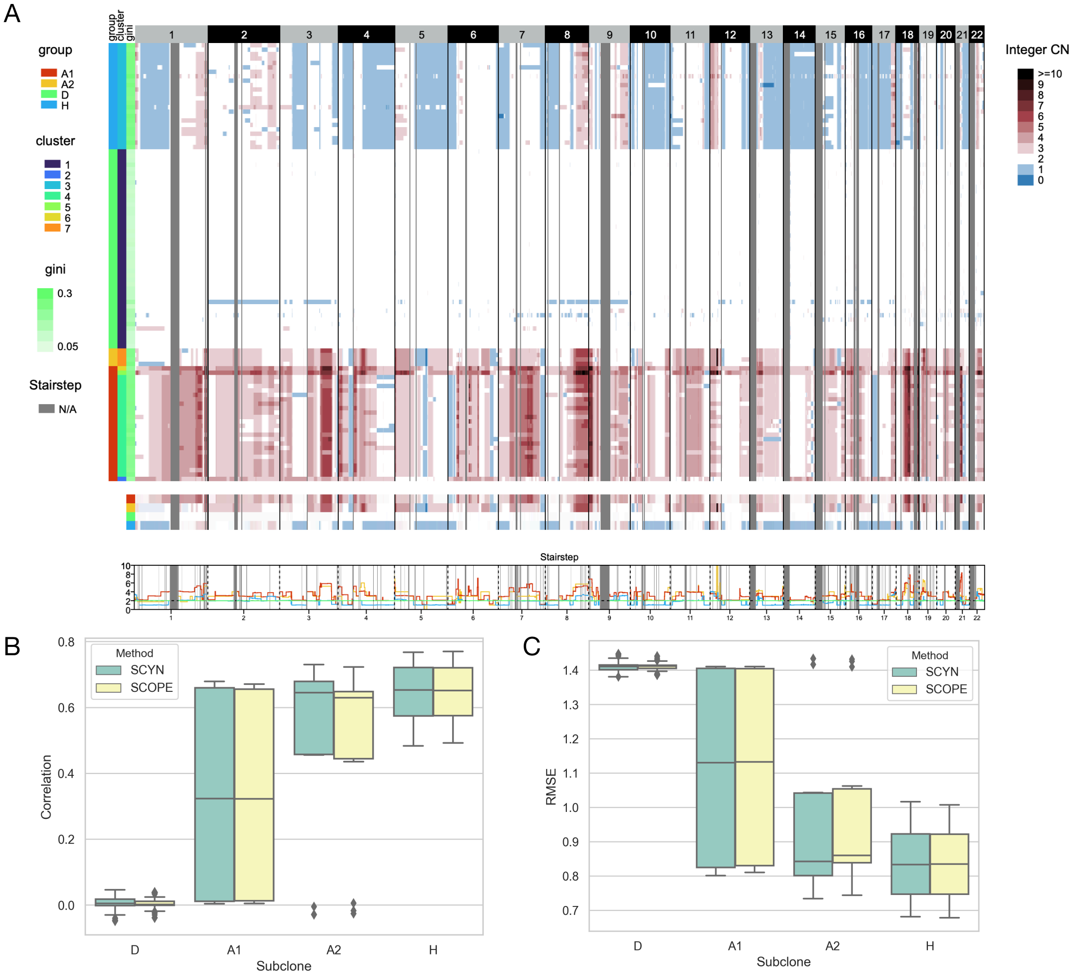


Fig. 3. Performance of SCYN on T10

(A) Heatmap of whole genome CNV profiles (B-C) Pearson correlation and RMSE as evaluation metrics comparing results by SCYN and SCOPE against aGCH.

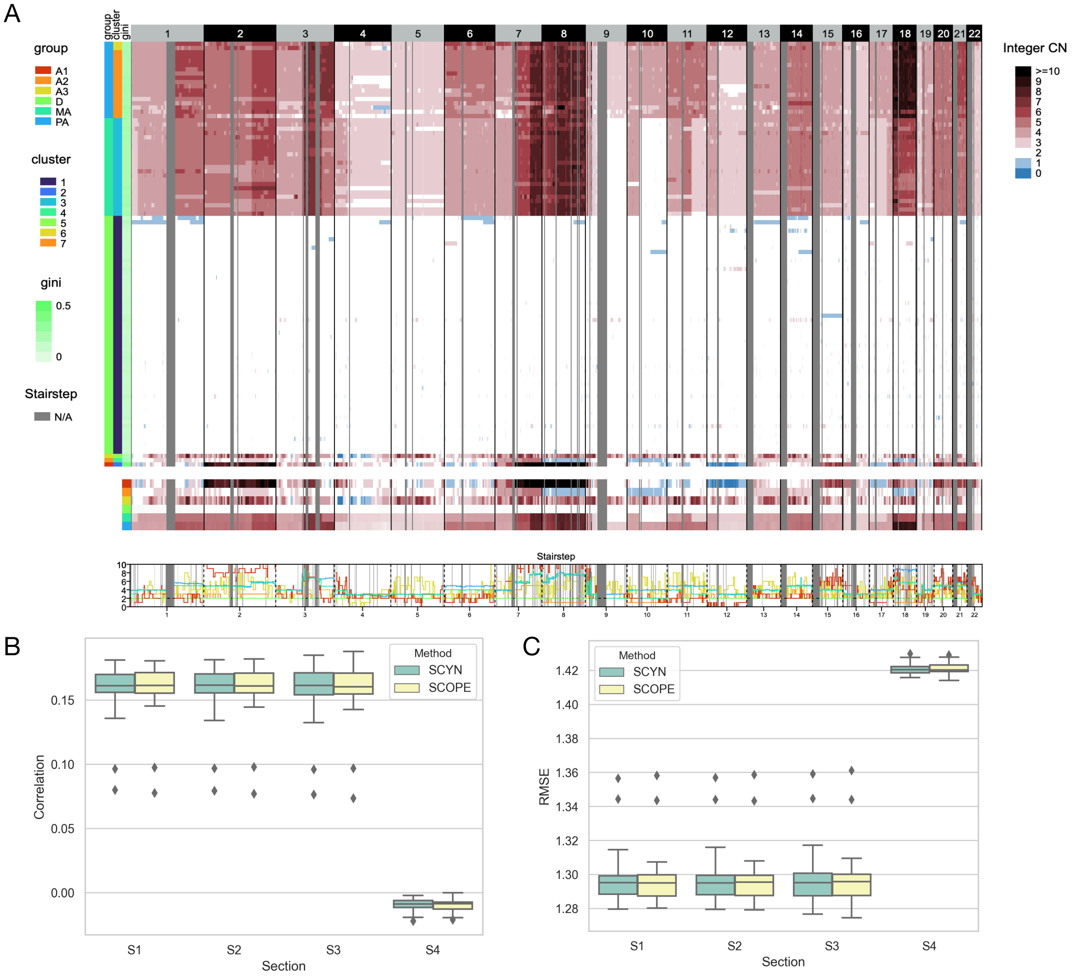


Fig. 4. Performance of SCYN on T16

(A) Heatmap of whole genome CNV profiles (B-C) Pearson correlation and RMSE as evaluation metrics comparing results by SCYN and SCOPE against aGCH.

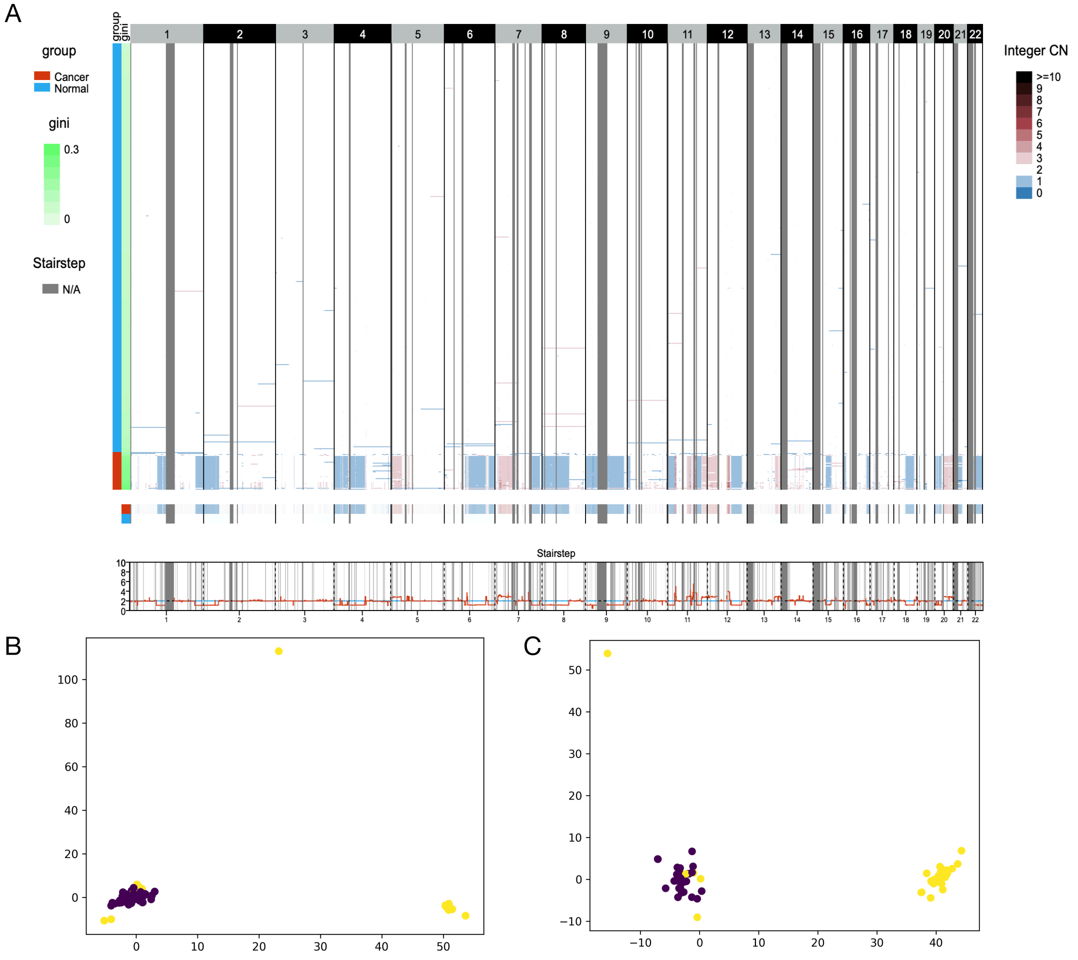


Fig. 5. Performance of SCYN on 10x spike-ins
(A) Heatmap of whole genome CNV profiles of 10% spike-in dataset (B-C) PCA plots on 1% and 10% spike-in datasets respectively. The yellow and purple dots denote cancer and normal cell respectively.

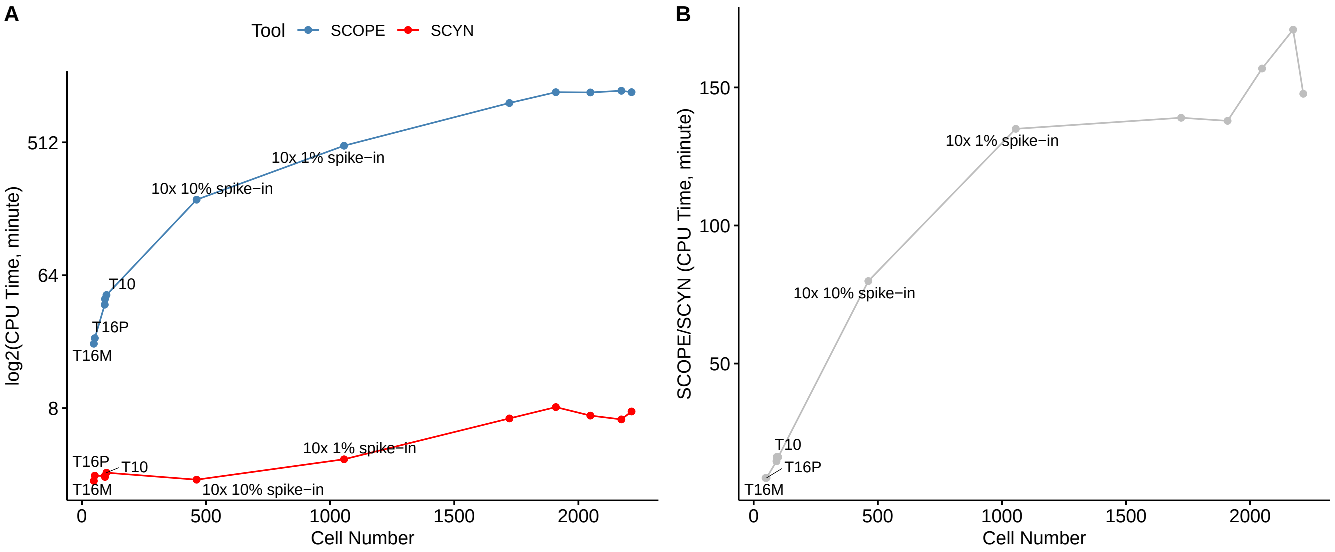


Fig. 6. Runtime performance of SCYN
(A) CPU time of SCYN and SCOPE on different cell number scale, respectively. (B) CPU time fold change of SCOPE against SCYN on different cell number scale.

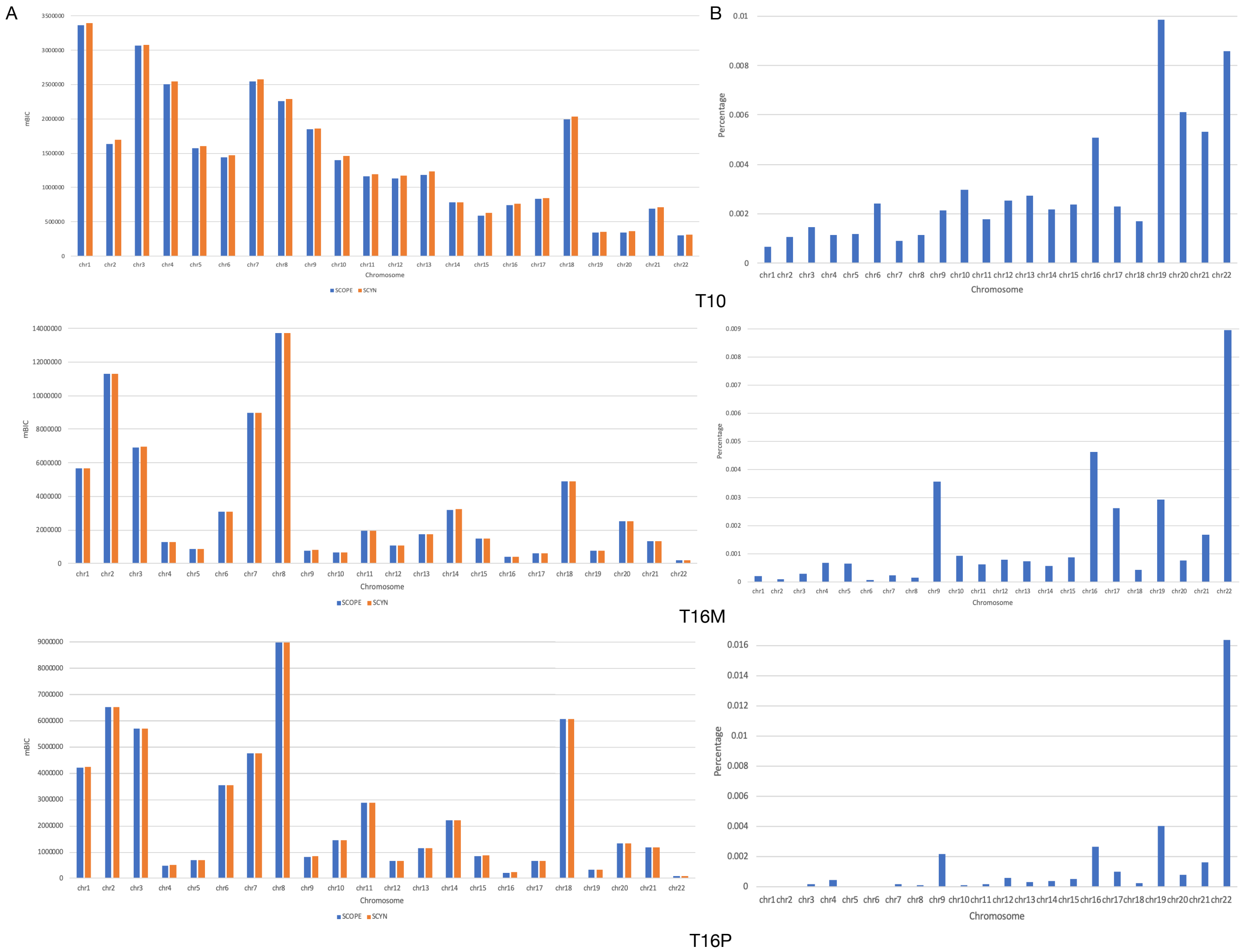


Fig. 7. (A) SCOPE-mBIC of T10, T16M and T16P across all chromosomes generated by SCYN and SCOPE, respectively. (B) The proportion of residual terms over mBIC across all chromosomes on T10, T16M, and T16P, respectively

Table 3. benchmark for runtimes (Minutes)

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