Downloaded from genome.cshlp.org on November 14, 2026 . Published by Cold Spring Harbor Laboratory Press



A statistical learning method for simultaneous copy number estimation and subclone clustering with single-cell sequencing data

Fei Qin, Guoshuai Cai, Christopher I. Amos, et al.

Genome Res. 2024 34: 85-93 originally published online January 30, 2024
Access the most recent version at doi:10.1101/gr.278098.123

References This article cites 59 articles, 10 of which can be accessed free at:
http://genome.cshlp.org/content/34/1/85.full.html#ref-list-1

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see https://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**



To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

© 2024 Qin et al.; Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on November 14, 2026 . Published by Cold Spring Harbor Laboratory Press

Method

A statistical learning method for simultaneous copy number estimation and subclone clustering with single-cell sequencing data

Fei Qin,[1] Guoshuai Cai,[2] Christopher I. Amos,[3] and Feifei Xiao[4]

[1]*Department of Epidemiology and Biostatistics,* [2]*Department of Environmental Health Science, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina 29208, USA;* [3]*Department of Quantitative Sciences, Baylor College of Medicine, Houston, Texas 77030, USA;* [4]*Department of Biostatistics, College of Public Health and Health Professions and College of Medicine, University of Florida, Gainesville, Florida 32603, USA*

The availability of single-cell sequencing (SCS) enables us to assess intra-tumor heterogeneity and identify cellular subclones without the confounding effect of mixed cells. Copy number aberrations (CNAs) have been commonly used to identify subclones in SCS data using various clustering methods, as cells comprising a subpopulation are found to share a genetic profile. However, currently available methods may generate spurious results (e.g., falsely identified variants) in the procedure of CNA detection, thereby diminishing the accuracy of subclone identification within a large, complex cell population. In this study, we developed a subclone clustering method based on a fused lasso model, referred to as FLCNA, which can simultaneously detect CNAs in single-cell DNA sequencing (scDNA-seq) data. Spike-in simulations were conducted to evaluate the clustering and CNA detection performance of FLCNA, benchmarking it against existing copy number estimation methods (SCOPE, HMMcopy) in combination with commonly used clustering methods. Application of FLCNA to a scDNA-seq data set of breast cancer revealed different genomic variation patterns in neoadjuvant chemotherapy-treated samples and pretreated samples. We show that FLCNA is a practical and powerful method for subclone identification and CNA detection with scDNA-seq data.

[Supplemental material is available for this article.]

In cancer, a small population of cancer stem cells evolves into a malignant mass of tumor cells, which then diverge and form distinct subclones, contributing to intra-tumor heterogeneity (ITH). Increased ITH levels are associated with tumor progression and resistance to clinical treatments (Stanta and Bonin 2018). Intrinsic mechanistic processes, such as inherent genomic variation, clonal competition, and tumor–host interactions, also contribute to ITH (Yachida et al. 2010; Gerlinger et al. 2012; Vogelstein et al. 2013; Polyak 2014). Therefore, an accurate assessment of ITH and the identification of subclones are essential to understand the mechanisms of tumor progression and resistance to therapy (Dagogo-Jack and Shaw 2018).

Most existing studies characterize clonal diversity using bulk DNA sequencing (Oesper et al. 2013; Ha et al. 2014; Li and Li 2014; Zhang et al. 2014; Deshwar et al. 2015), which is limited to reporting only an average signal from complex populations of cells (Navin 2015). Single-cell sequencing (SCS) technology enables the assessment of ITH on a single-cell basis (Navin et al. 2011; Wang et al. 2014) using either single-cell DNA or RNA sequencing (scDNA/RNA-seq) to reveal cellular evolutionary relationships (Tang et al. 2019). Subclonal populations can be identified within tumor tissues using SCS, allowing for the inference of tumor evolution and providing insights into future developments of targeted therapy (Jiang et al. 2016).

Subpopulations of cancer cells share genetic characteristics (Cariati et al. 2019), with chromosomal copy number aberration (CNA) being one of the most important types, involving the gain or loss of DNA segments. CNAs stimulate the stemness of other tumor cells, leading to the formation of new cancer stem cells and ultimately giving rise to clonal evolution in cancers (Dai and Liu 2021). Consequently, CNAs serve as good biomarkers to assess ITH and identify subclones. Principally, the detection of CNAs has been performed by identifying changes at the boundaries of copy number regions within chromosomal segments. However, it remains challenging to detect CNAs in scDNA-seq data owing to shallow sequencing and uneven depth of coverage (Mallory et al. 2020). The method HMMcopy, which uses a hidden Markov model to detect copy number variations, was designed for data from array comparative genomic hybridization (aCGH) and high-throughput sequencing (Shah et al. 2006; Ha et al. 2012), although it has been extensively applied in SCS data (https://rdrr.io/bioc/HMMcopy/; Laks et al. 2019). A single-cell specific method, SCOPE (Wang et al. 2020), was recently developed for copy number estimation using a Poisson latent factor model for normalization and an expectation–maximization (EM) algorithm. Studies based on these methods have typically used an independent statistical model to first detect the copy number profile, followed by classical clustering methods (e.g., hierarchical) (Bridges 1966) for subclone identification in downstream analyses. However, this

Corresponding author: feifeixiao@ufl.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.278098.123.

© 2024 Qin et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see https://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Downloaded from genome.cshlp.org on November 14, 2026 . Published by Cold Spring Harbor Laboratory Press

two-step framework risks generating spurious results owing to the carry-over of noise in the copy number profiling step into the subclone clustering.

Previously, Rojas and Wahlberg (2014) proposed a fused lasso method for general change-point detection problems; applied here, this approach converts the problems of detecting chromosomal breakpoints to a convex optimization problem whereby we propose to simultaneously identify subclones. We herein developed FLCNA, a CNA detection method based on the fused lasso model, to achieve accurate subclone clustering and copy number profiling without needing an intermediate step to estimate copy number states. Our approach not only yields more precise clustering results but also increases the accuracy of CNA detection in each individual cell by borrowing information across cells within the same subclone. This study represents the first exploration of the fused lasso model in copy number profile–based subclone clustering.

We benchmarked FLCNA against existing copy number profile detection methods (i.e., SCOPE and HMMcopy) (https://rdrr.io/bioc/HMMcopy; Wang et al. 2020), coupled with classical hierarchical (Bridges 1966) and *k*-means (James and others 1967) clustering methods. The performance of each approach at subclone clustering and CNA detection was evaluated through extensive simulations. We also applied FLCNA to a breast cancer data set and successfully clustered subclones based on shared breakpoints of cells in three patients. In conclusion, we have developed a robust and powerful tool for subclone clustering and CNAs detection in SCS data.

Results

Overview of the method

To capture the biological heterogeneity between potential subclones, we developed the FLCNA method based on a fused lasso model (Rojas and Wahlberg 2014), which can simultaneously identify subclones and detect breakpoints in scDNA-seq data. The framework of the FLCNA method is summarized in Figure 1. First, data sets undergo preprocessing procedures, including quality control and normalization. Subclone clustering and breakpoint detection are then achieved simultaneously using a Gaussian mixture model (GMM) combined with a fused lasso penalty term. Finally, based on the shared breakpoints in each cluster, candidate CNA segments for each cell are clustered into different CNA states using a GMM-based clustering strategy (Xiao et al. 2019).

Evaluation of FLCNA via spike-in simulations

We first conducted spike-in simulations to assess the clustering performance of FLCNA compared with two other copy number estimation methods (SCOPE and HMMcopy) (https://rdrr.io/bioc/HMMcopy/; Wang et al. 2020) coupled with different clustering methods (hierarchical or *k*-means). Using a mixture of four different copy number states, including deletion of double copies (Del.d), deletion of a single copy (Del.s), duplication of a single copy (Dup.s), and duplication of double copies (Dup.d), FLCNA outperformed the other approaches in all scenarios having different numbers of clusters (five or three) and CNAs in a cluster

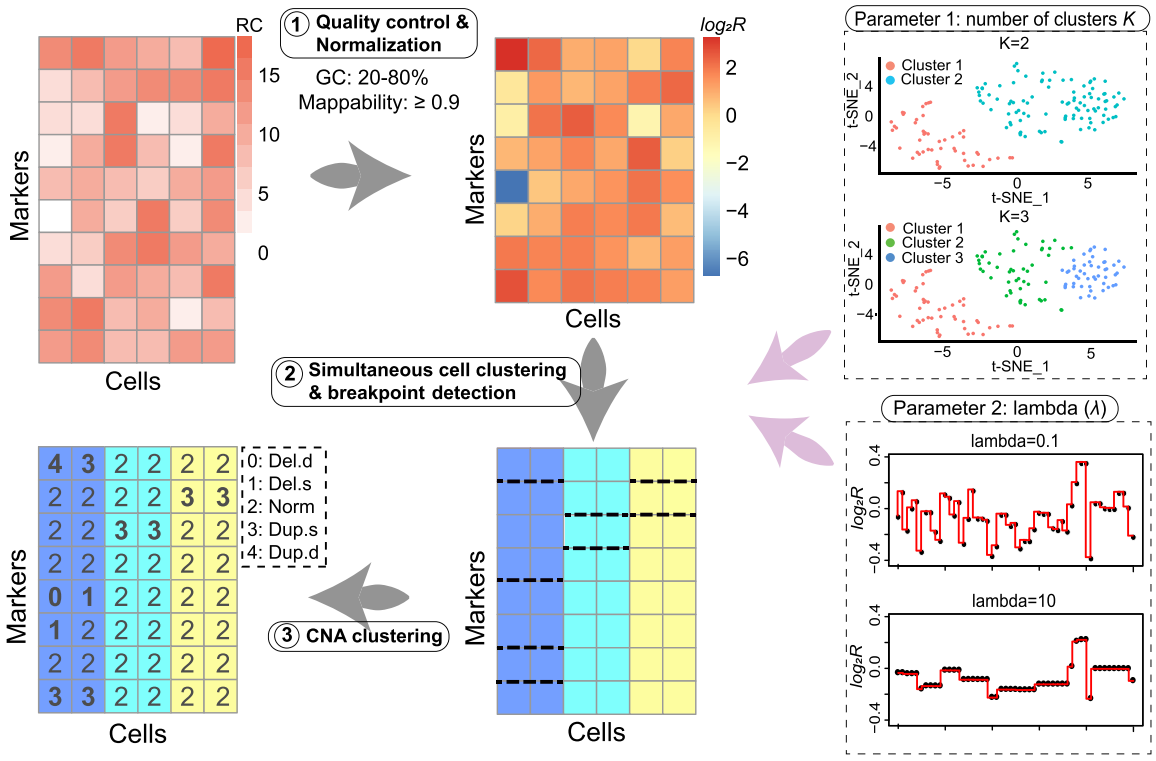


Figure 1. Analysis workflow of FLCNA. FLCNA first implements a quality-control procedure based on GC content and mappability. A two-step median normalization approach sequentially removes the effect of each bias. Data are then scaled by their cell-specific mean and undergo a logarithm transformation (log_2R). With log_2R, FLCNA clusters subclones and simultaneously detects shared breakpoints using a Gaussian mixture model with a fused lasso penalty term. Finally, based on these shared breakpoints in each cluster, segments for each cell are clustered into five different CNA states: deletion of double copies (Del.d), deletion of a single copy (Del.s), normal/diploid (Norm), duplication of a single copy (Dup.s), and duplication of double copies (Dup.d). There are two hyperparameters in the FLCNA model, the number of cell clusters *K* and a tuning parameter λ, which controls the number of breakpoints.

Downloaded from genome.cshlp.org on November 14, 2026 . Published by Cold Spring Harbor Laboratory Press

(Fig. 2; Supplemental Figs. S1, S2). Specifically, with five predefined clusters and a fixed number of CNAs (i.e., 50) in each cluster, FLCNA's clustering performance incrementally improved as the within-cluster CNA sharing proportion increased. FLCNA outperformed the other methods based on adjusted Rand index (ARI) at larger CNA sharing proportions (>40%) (Fig. 2; Supplemental Table S1), although at a CNA sharing proportion of 20%, almost all methods failed to provide desirable clustering performance (ARI < 0.50).

The overall clustering performance of all methods improved in scenarios with three predefined clusters (Supplemental Fig. S1; Supplemental Table S2). For example, with supershort CNAs shared by 60% of samples, the ARI of FLCNA improved from 0.781 in the scenario of five preclusters (Supplemental Table S1) to 0.985 in that of three preclusters (Supplemental Table S2). This improvement was likely because of the increased sample size in each cluster given a fixed total sample size. Compared with the scenario with a fixed number of shared CNAs in each cluster, it was more challenging for all methods to identify subclones when varied numbers of shared CNAs presented, as fewer shared CNAs were generated in some clusters by randomness (Supplemental Fig. S2; Supplemental Table S3). For samples with a single type of copy number state (i.e., Del.d, Del.s, Dup.s, Dup.d), FLCNA still outperformed other methods in clustering subclones (Supplemental Figs. S3–S5; Supplemental Tables S1–S3). All methods resulted in higher ARI for deletions compared with duplications, and for double copy changes compared with single copy, owing to the stronger signals presented in the intensities for the two former cases. FLCNA also clustered cells accurately even for simulations consisting of a single subclone (Supplemental Table S4). Overall, FLCNA had the highest accuracy for identifying subclones in simulated scenarios with commonly shared CNAs.

We also evaluated the accuracy of FLCNA in CNA detection by comparing it to SCOPE and HMMcopy (Fig. 3; Supplemental Figs. S6–S10; Supplemental Tables S5–S7). In comparison, FLCNA showed improved accuracy in identifying CNAs as the proportion of shared variants increased. For example, with a fixed number of CNAs (i.e., 50) shared by all samples in each of five clusters, FLCNA had an *F*1 score of 0.896, whereas SCOPE only had a score of 0.435 and HMMcopy had a score of 0.783 (Fig. 3; Supplemental Table S5). In general, FLCNA outperformed other methods in detecting supershort and short CNAs.

In summary, FLCNA showed overall great performance in clustering subclones with copy number changes shared by a large proportion of samples within a cluster. In copy number profile detection, FLCNA presented its advantage in detecting short CNAs shared by a relatively large proportion of samples within a cluster, benefiting from the information shared among cells within the same subclone.

Application to a TNBC breast cancer single-cell study

Triple-negative breast cancer (TNBC) remains a critical class of breast cancer, constituting 12%–18% of all breast cancer patients (Foulkes et al. 2010), yet owing to chemoresistance, poor overall survival performance is observed in ~50% of TNBC patients (Foulkes et al. 2010). Neoadjuvant chemotherapy (NAC) is the standard therapy for TNBC patients who show a low level of the estrogen, progesterone, and ERBB2 (also known as HER2) receptors and are not eligible for hormone or HER2-targeted therapy (Foulkes et al. 2010). Many studies have shown that patients with TNBC harbor high levels of somatic mutations (Wang et al. 2014; Gao et al. 2016), which partially result in extensive ITH. Identifying subclones and detecting mutations in this subpopulation are critical to untangle the molecular mechanisms of chemoresistance in these patients.

To this end, we applied FLCNA to a single-cell study of three breast cancer patients receiving NAC (i.e., KTN126, KTN129, KTN302), resulting in the identification of three clusters in the KTN126 patient and two clusters in the other two patients (Fig. 4; Supplemental Figs. S11, S12). For the KTN126 patient, 64 cells (17 cells with pretreatment and 47 cells with posttreatment) were clustered in cluster A, with nine cells in cluster B and 20 cells in cluster C, all from the pretreatment group (Fig. 4). Similar copy number profile patterns were observed within clusters. Clusters B and C, which included more pretreatment samples, had higher variation in genetic intensities compared with cells in cluster A, indicating that some copy number aberrations were treatment specific. Alternatively, the treatment of NAC may have led to the extinction of tumor cells with these copy number changes in this patient. Similar patterns were observed in the other two patients (Supplemental Figs. S11, S12), and the locations of these treatment-specific copy number changes were consistent across patients KTN129 and KTN302.

FLCNA identified 264, 154, and 156 shared CNAs within a cluster in the KTN126, KTN129, and KTN302 patients,

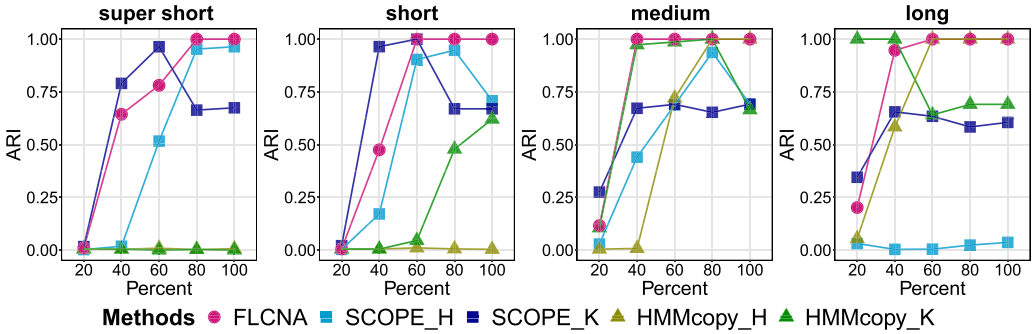


Figure 2. Accuracy of clustering in simulated data with five clusters and mixed CNA states. Clustering results from FLCNA were compared with existing methods (i.e., SCOPE and HMMcopy) coupled with different clustering methods. For each of five clusters, we added 50 CNA segments with varied lengths (supershort: two to five markers; short: five to 10 markers; medium: 10 to 20 markers; and long: 20 to 35 markers) and varied CNA proportions (20%, 40%, 60%, 80%, 100%), respectively, to the background signal. Signals of mixed CNA states (i.e., Del.d, Del.s, Norm, Dup.s, and Dup.d) were spiked in. (ARI) Adjusted Rand index, (SCOPE_H) SCOPE_hierarchical, (SCOPE_K) SCOPE_K-means, (HMMcopy_H) HMMcopy_hierarchical, and (HMMcopy_K) HMMcopy_K-means.

Downloaded from genome.cshlp.org on November 14, 2026 . Published by Cold Spring Harbor Laboratory Press

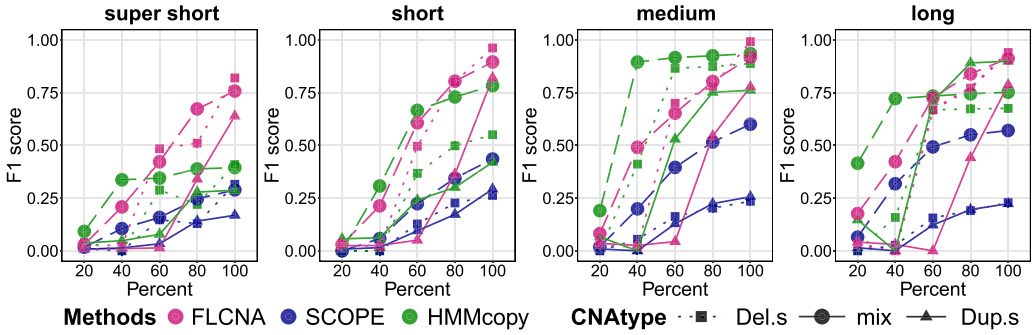


Figure 3. Accuracy of CNA detection in simulated data with five clusters. CNA calls were generated by FLCNA, SCOPE, and HMMcopy, respectively. For each of five clusters, we added 50 CNA segments with varied lengths (supershort: two to five markers; short: five to 10 markers; medium: 10 to 20 markers; and long: 20 to 35 markers) and varied CNA proportions (20%, 40%, 60%, 80%, 100%), respectively, to the background signal. Deletion of a single copy (Del.s), mixed CNA states (mix), and duplication of a single copy (Dup.s) were spiked in separately. *F1* score was used to evaluate the performance of CNA detection for each method.

respectively. Consensus CNA-located genes (e.g., *DAPK1*, *TBX3*, *NCOA3*, *KRAS*) among patients implied the shared common evolutionary path of the tumor cells. Mutations in these genes have been shown to play important roles in evolution (*DAPK1*, *TBX3*) (Fischer and Pflugfelder 2015; Zhao et al. 2015) and breast cancer therapy (*NCOA3*, *KRAS*) (Burandt et al. 2013; Tokumaru et al. 2020). Overall, the shared CNAs were mapped to 436 out of 575 genes associated with breast cancer risk identified from existing genome-wide association studies (GWASs). These mapped genes were enriched in pathways related to cancer, hormones, immunity, and epithelial–mesenchymal transition (EMT) (Supplemental Fig. S13). Most pathways were shared by all three patients and were consistent with findings from existing studies on breast cancer. For instance, PATHWAYS_IN_CANCER was found to be hyperactivated in human tumor tissue (Seryakov et al. 2021). Pathways related to EMT (e.g., ADHERENS_JUNCTION) (Liu et al. 2016) and immune (e.g., TOLL_LIKE _RECEPTOR) (Shi et al. 2020) play crucial roles in the invasion and metastasis of tumor cells. Various hormone-related pathways were also essential in the occurrence and progression of breast cancer (Subramani et al. 2017). We also observed a novel CNA in *13q31.3* (e.g., *LINC01040*) among all three patients, with duplications in the KTN126 and KTN129 patients and deletions in the KTN302 patient. This discovery of a new CNA may provide new insights into understanding the progression of breast cancer.

We also evaluated the heterogeneity of CNAs among cells based on the sharing proportion of CNAs per cluster. The mean sharing proportion was 30.3% in the KTN126 patient (18.6% in cluster A, 36.4% in cluster B, and 34.7% in cluster C), 25.7% in the KTN129 patient, and 26.4% in the KTN302 patient (Supplemental Fig. S14). Despite the relatively low sharing percentage for some CNAs, we still observed many CNAs shared by a majority (>60%) of cells within a cluster for the TNBC data set (Supplemental Table S8). For instance, in the case of cluster C in patient KTN126, we found that 18.18% of CNAs were shared by >60% of the cells. Consequently, the priority of our method in data with a large proportion of shared CNAs within clusters will greatly benefit the subclone clustering and CNAs detection in real SCS data. Overall, our method detected CNAs with a wide range of length, varying from two to 1000 bins (Supplemental Fig. S15). In terms of computational speed, on the 93 cells from the KTN126 patient (Supplemental Table S9), FLCNA was significantly faster (1.2 h) than SCOPE (10.5 h) using a high-performance cluster with 12 GB RAM.

Discussion

The importance of copy number changes in modulating human disease is being increasingly recognized, especially concerning response to treatments in cancer. scDNA-seq enables researchers to profile heterogenous tumor tissues at the single-cell level. Accurate detection of CNAs with scDNA-seq data is crucial for identifying copy number profiles at the single-cell resolution to better understand how tumor lineages evolve (Baslan and Hicks 2017; Ren et al. 2018). Numerous scDNA-seq studies (Gao et al. 2016; Hou et al. 2016; Ferronika et al. 2017) have used CNAs to characterize tumor subclones, revealing that most tumors contain multiple subclonal lineages. In our study, we developed FLCNA for simultaneous subclone clustering and copy number estimation based on the fused lasso model. Our method effectively addresses the issue of reduced clustering accuracy caused by false-positive CNAs to which existing two-step methods are susceptible (https://rdrr.io/bioc/HMMcopy/; Wang et al. 2020).

Recent advances in SCS technology have provided emerging tools for computational methods to infer subclones with different genomic variants, including single-nucleotide alterations (SNAs) and CNAs. SAPPH (Xia et al. 2015) estimated CNAs and inferred tumor subclone proportion from paired tumor-normal data. The SCClone (Yu et al. 2022) and SiCloneFit (Zafar et al. 2019) methods clustered subclones using single-cell SNA data to reconstruct the clonal populations and the evolutionary relationship between the clones. Elyanow et al. (2021) overcame the challenges in inferring CNAs from RNA sequencing data by using spatial information of a small group of cells to help identify CNAs and the spatial distribution of clones within a tumor sample. The CONET (Markowska et al. 2022) and SCICoNE (Kuipers et al. 2020) methods were both developed to jointly infer an evolutionary tree on copy number events and copy number profile using scDNA-seq data. A major distinguishing feature of our method is its simultaneous clustering of cells and detection of CNAs, solving the problem of declined clustering performance resulting from falsely identified variants in the stage of variant estimation.

The superior clustering performance of FLCNA for scDNA-seq data has been shown in extensive simulations. Our spike-in simulations provided accurate clustering results from FLCNA for the samples with a large proportion of shared CNAs. Moreover, FLCNA outperformed the other two methods in detecting supershort and short CNAs. Notably, identifying supershort and short CNAs is typically more challenging compared with longer-sized

Downloaded from genome.cshlp.org on November 14, 2026 . Published by Cold Spring Harbor Laboratory Press

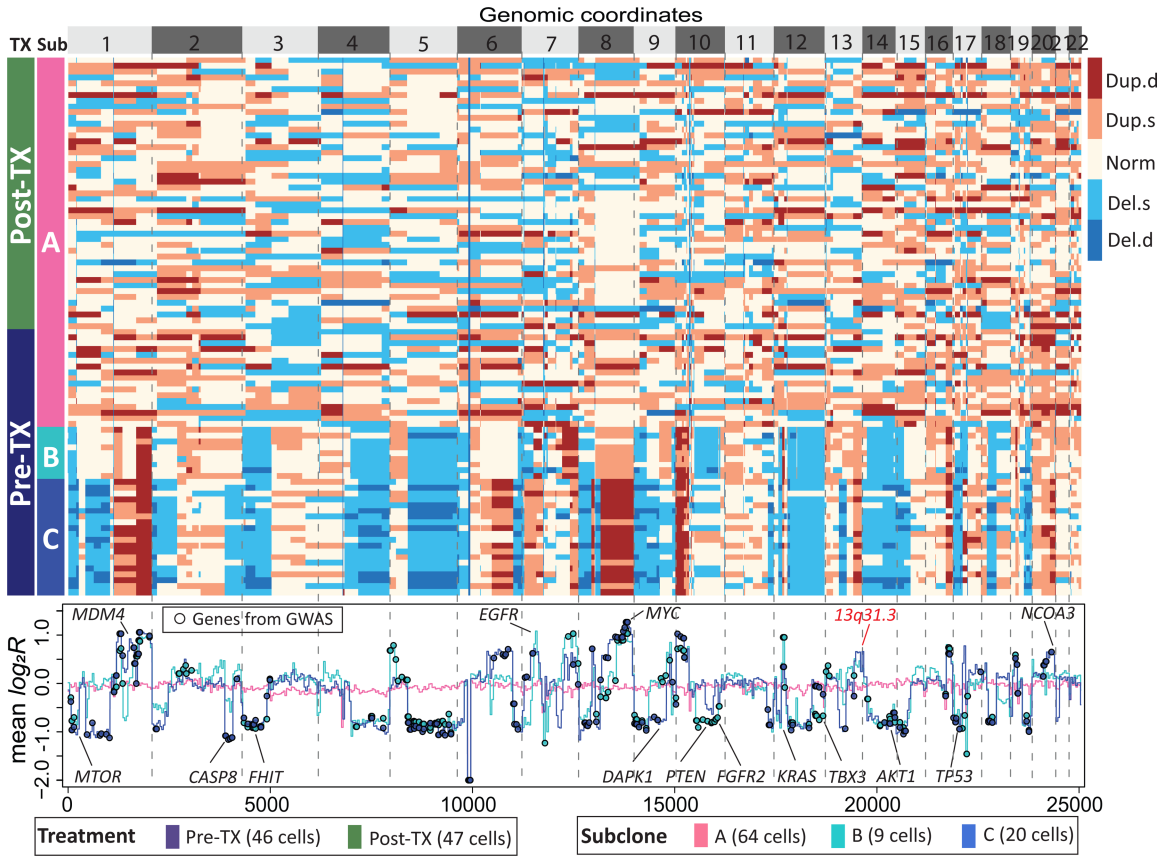


Figure 4. Subclone clustering of the KTN126 patient using FLCNA. Cell clusters and copy number profile with different CNA states (deletion of double copies [Del.d], deletion of a single copy [Del.s], normal/diploid [Norm], duplication of a single copy [Dup.s], and duplication of double copies [Dup.d]) were generated using FLCNA. The mean logarithm transformation of ratio between normalized read counts and its sample-specific mean (log_2R) was provided for each cluster. Shared CNAs identified using FLCNA were matched to significant genes from the genome-wide association studies (GWASs) in the NHGRI-EBI GWAS Catalog. (Pre-TX) Pretreatment, (Post-TX) posttreatment.

CNAs. Using a whole-genome scDNA-seq data set on breast cancer patients, FLCNA successfully clustered subclones containing clearly distinct genomic variation patterns. Our analysis suggests that treatment of NAC may lead to the extinction of tumor cells, despite the possible bias brought by the fact that partial posttreatment samples were collected adjacent to normal tissues. With our method, more CNAs were identified in the previously defined "normal" cells than the original study of this data set (Kim et al. 2018), as well as the SCOPE reanalysis (Wang et al. 2020). In contrast to our approach, these two previous studies predefined "normal" cells as references to calculate signal intensities for change-point detection. FLCNA uses the observed mean value of each cell across bins as a reference, providing increased sensitivity at detecting CNA signals. An advantage of our reference-free normalization method is that it avoids relying on the availability of normal cells, which not all studies contain or which can be mislabeled in cell-type annotation. In addition to prediction accuracy, the high dimensionality of scDNA-seq data sets requires high computational efficiency. Even with 100 kb as the binning size, our method retained computational speed, without compromising the accurate detection of short CNAs.

Like any other studies, there are limitations in our study. First, only shared breakpoints identified in a cluster were used to estimate the underlying copy number profile for each cell, which may not be desirable if the goal is to identify all CNAs at the single-cell level. Additionally, errors owing to doublets in SCS data, in which two or more cells have the same cell barcode label (Zaccaria and Raphael 2021), were not considered in our framework. Although previous studies have shown that doublets can affect the accuracy of copy number estimation (Zaccaria and Raphael 2021), because we only used shared breakpoints for CNA detection, we expect the bias arising from the existence of doublets will be minimal. Additionally, although our method is computationally efficient for a single value of parameters K and λ, selecting the most optimal choice among many candidate value combinations may be computationally intensive.

In conclusion, our FLCNA method offers a novel methodology approach to CNA detection and subclone identification in SCS data. Our method has the potential to extend to other data types, such as scRNA-seq, and to integrate different dimensions of sequencing information (e.g., B-allele frequency, gene expression, spatial information) to improve the identification of subclones and benefit research in cancer outcome-related targeted therapy. For example, we could incorporate spatial information into our model for more comprehensive inference of subclones with scDNA-seq data, as cells located nearby are more likely to share similar genetic patterns and consequently tend to reside within same subclones (Elyanow et al. 2021).

Downloaded from genome.cshlp.org on November 14, 2026 . Published by Cold Spring Harbor Laboratory Press

Methods

FLCNA was developed based on a fused lasso model (Rojas and Wahlberg 2014) to cluster subclones and simultaneously detect CNAs with scDNA-seq data. Note that although our framework was motivated by the detection of somatic copy number change, it can also be naturally extended to germline copy number variation detection. For consistency, our framework and evaluation will use the term CNA in describing and evaluating the methods across the paper.

Quality control and preprocessing of scDNA-seq data

In the first step of FLCNA, the read count data for each marker/bin and each cell undergo a binning strategy (Wang et al. 2020), which divides the genome into equally sized windows, and each window is defined as a marker. These data were subsequently used as input in our model. Next, the GC content and mappability are generated for each marker. To reduce artifacts for copy number detection, a quality-control procedure removes markers with extreme GC content (<20% and >80%) or those with low mappability (<0.9) (Wang et al. 2020). Then a two-step median normalization approach (Magi et al. 2013) is used to further remove the effect of biases from the GC content and mappability. Altogether, the normalized read counts are as follows: $RC_i' = RC_i \frac{m}{m_o}$, where RC_i are the read counts for marker i, m_o is the median read counts of all markers with the same o value (where o = [the GC-content, mappability]) as the ith marker, and m is the overall median read counts of all the markers. In the final preprocessing step, FLCNA computes the ratio of the normalized read counts to their sample-specific mean, the logarithm transformation of which (i.e., log_2R) represents the main signal intensities.

Notations and models

Considering we have N cells in total, let $\boldsymbol{X} = (x_{i,j})_{P\times N}$ be the normalized read count data (i.e., log_2R), where $x_{i,j}$ denotes the value of the ith ($i = 1, \ldots, P$) marker from the jth cell. $\boldsymbol{x_j} = (x_{1,j}, \ldots, x_{P,j})^T$ is the data vector for the jth sample. The samples sharing common biological characteristics (e.g., genomic variation) belong to the same cluster. Starting from a general K-cluster problem with a GMM, the observations $\boldsymbol{x_j}$ are assumed to be independent and are generated from a probability density function $g(\boldsymbol{x_j}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x_j};\ \boldsymbol{\mu_k},\ \boldsymbol{\Sigma_k})$, where the "weights" $\pi_k's$ ($\pi_k \geq 0$ for each cluster, $1 \leq k \leq K$ and $\sum_{k=1}^{K} \pi_k = 1$) are the mixing proportions. For the kth cluster, $f_k(\boldsymbol{x_j}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$ denotes the Gaussian density function with mean vector $\boldsymbol{\mu_k} = (\mu_{1,k} \ldots \mu_{P,k})^T$ and covariance matrix $\boldsymbol{\Sigma_k}$. $\boldsymbol{\mu_k}$ denotes the mean genetic intensity of each marker in the kth cluster, and $\boldsymbol{\Sigma_k}$ captures the correlations among the markers. In this study, we assume that the covariance matrix $\boldsymbol{\Sigma_k} = \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2,\ \sigma_2^2,\ \ldots \sigma_P^2)$ is fixed for all clusters. Parameters including π_k, $\boldsymbol{\mu_k}$, and $\boldsymbol{\Sigma}$ are unknown and need to be estimated.

The problem of copy number profile detection is equivalent to chromosomal breakpoint detection and has been initially explored in the fused lasso model (Rojas and Wahlberg 2014). Fused lasso was extended from the classical LASSO model to select variables and penalize the difference of successive features. Its ability in identifying and quantifying significant features is closely related to our problem of breakpoint detection on locating significant signals from a wide range of constant signals (Rojas and Wahlberg 2014). Using the fused lasso penalty term for change point detection, the penalized log likelihood function in the FLCNA method is given by

$$Q = \sum_{j=1}^{N} \log\left[\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x_j};\ \boldsymbol{\mu_k},\ \boldsymbol{\Sigma})\right] - \lambda \sum_{k=1}^{K} \sum_{i=1}^{P-1} \tau_{i,i+1}^{(k)} |\mu_{i,k} - \mu_{i+1,k}|. \quad (1)$$

In the second term of Equation 1, a tuning hyperparameter λ and predefined adaptive weights (Zou 2006) $\tau_{i,i+1}^{(k)}$ are used to shrink the absolute difference of the mean shift values $|\mu_{i,k} - \mu_{i+1,k}|$ in consecutive markers in the kth cluster, ultimately disclosing change points. The tuning hyperparameter λ is used to control the overall number of change points such that fewer change points tend to be generated with larger λ values (Fig. 1). To shrink each pair of consecutive markers with the same weight, the tuning hyperparameter λ is fixed within each cluster; however, such a strategy may decrease the accuracy of penalization and ultimately affect the clustering performance. To improve the accuracy, an adaptive penalization weight (Zou 2006) $\tau_{i,i+1}^{(k)} = |\tilde{\mu}_{i,k} - \tilde{\mu}_{i+1,k}|^{-1}$ is applied, where $\tilde{\mu}_{i,k}$ is estimated from the same model without any penalization ($\lambda = 0$). This adaptive penalization weight is predefined to dynamically penalize each pair of successive markers. For example, if there is large difference between $\tilde{\mu}_{i,k}$ and $\tilde{\mu}_{i+1,k}$ in the model without penalization, a change point is expected to appear between the ith and $(i+1)$th marker, and this change point tends to be informative for subclone clustering. In this case, according to $\tau_{i,i+1}^{(k)} = |\tilde{\mu}_{i,k} - \tilde{\mu}_{i+1,k}|^{-1}$, $\tau_{i,i+1}^{(k)}$ will be small and consequently, the difference between $\mu_{i,k}$ and $\mu_{i+1,k}$ in Equation 1 will be lightly penalized and will be more informative for the subclone clustering. Otherwise, the difference between $\mu_{i,k}$ and $\mu_{i+1,k}$ will be heavily penalized and will be less informative for the subclone clustering in our FLCNA method. With the focus on subclone identification and change point detection, our goal is to maximize Equation 1 to estimate the parameter set $\boldsymbol{\Theta} = \{\boldsymbol{\mu_k}, \boldsymbol{\Sigma},\ \pi_k\}_{k=1}^{K}$.

Parameter estimation using EM algorithm

In FLCNA, the parameter set $\boldsymbol{\Theta}$ is estimated using the EM algorithm (Dempster et al. 1977). We initialize $\boldsymbol{\Theta}$ with parameters estimated from the model without penalty ($\lambda = 0$) and then update these parameters by alternating between the E- and M-steps. In the M-step, given the starting values of $\boldsymbol{\Theta}$, the probability for the jth sample belonging to the kth cluster is calculated by dividing the density of the kth cluster by the sum of densities from all clusters. Thereafter, an E-step is used to update the estimated values of $\boldsymbol{\Theta}$. Specifically, the "weight" for the kth cluster π_k and the variance for the ith marker σ_i^2 are estimated by taking the first derivative of $Q(\boldsymbol{\Theta})$ w.r.t. π_k and σ_i^2, respectively. The estimates of the cluster means $\hat{\boldsymbol{\mu}}_k$ are computed with a local quadratic approximation algorithm (Fan and Li 2001). These updated $\boldsymbol{\Theta}$ estimation values were iteratively computed between the E- and M-steps until convergence. We refer readers to the Supplemental Methods for a thorough description of this part of algorithm. After this step, all the parameters of interest $\boldsymbol{\Theta} = \{\boldsymbol{\mu_k}, \boldsymbol{\Sigma},\ \pi_k\}_{k=1}^{K}$ are successfully estimated, based on which cells are clustered according to the estimated cluster weights and copy number states are assigned by the estimated cluster means (described below).

Copy number profile identification and hyperparameter estimation

With the estimated cluster means ($\hat{\boldsymbol{\mu}}_k$), we locate and quantify all the shared change points and identify copy number segments within clusters. Based on these identified shared segments, we assign the most likely copy number state for each segment in each

Downloaded from genome.cshlp.org on November 14, 2026 . Published by Cold Spring Harbor Laboratory Press

cell. A GMM-based clustering strategy (Xiao et al. 2019) is implemented for CNA clustering using the normalized read count data (i.e., log_2R). Segments sharing similar intensity levels in a cell are identified as the ones with the same copy number states. Each segment is classified using a five-state classification scheme with Del.d, Del.s, normal/diploid, Dup.s, and Dup.d. We aggregate the states with more than four copies in the copy number state of Dup.d, considering our primary focus is on the clustering of cell subclones. Two hyperparameters will be predefined, including the number of clusters K and the tuning parameter λ. To find the optimal values of K and λ, we use a Bayesian information criterion (BIC) (Guo et al. 2010), and the clustering model with the smallest BIC value is selected as the optimal model.

Data description

We used two publicly available scDNA-seq data sets for illustration and evaluation of FLCNA in simulations and real data analyses. As described below, the BRCA5 data set was used to mimic real data signals in simulations, and the TNBC data set was analyzed for real data applications.

The BRCA5 data set consists of 10,088 cells from an aggregate of five breast tissue nuclei sections in frozen breast tumor tissue. Cells were sequenced using the 10x Genomics platform (https://www.10xgenomics.com/resources/datasets/aggregate-of-breast-tissue-nuclei-sections-10-k-cells-1-standard-1-1-0), which uses microfluidic droplets to barcode cells and perform library construction. We generated a read depth matrix of 28,760 markers and 10,088 cells from BAM files after binning with a 100-kb bin size (Wang et al. 2020). To save computational time, we randomly selected 220 cells from the data set and used the read counts from the entire genome to mimic real data in our simulations. The TNBC data set consists of data from TNBC patients obtained from the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra; accession no. SRP114962) (Kim et al. 2018). TNBC is characterized by extensive ITH and frequently develops resistance to NAC treatment. Three patients (i.e., KTN126, KTN129, KTN302) were used for our analyses, in which tumor cells were only reported in the pretreatment samples. For each patient, cells were sequenced at two time points (pre- and mid/posttreatment) with 93 cells (46 pre- and 47 posttreatment) in the KTN126 patient, 90 cells (46 pre- and 44 posttreatment) in the KTN129 patient, and 92 cells (47 pre- and 45 midtreatment) in the KTN302 patient, respectively. For these samples, FASTQ files were generated with Fastq-dump from SRA-Toolkit (Leinonen et al. 2011) and then aligned to the NCBI hg19 reference genome and converted to BAM files. The raw read depth of the coverage data was generated from the BAM files with a bin size of 100 kb (Wang et al. 2020). We used the hg19 reference genome for alignment in our study because both the original study of this data set (Kim et al. 2018) and the SCOPE paper (Wang et al. 2020) used the same reference genome. Theoretically, our method does not rely on the version of reference genome and can also be applied to data generated with other genome assemblies.

Spike-in simulations

We compared FLCNA to existing copy number profile detection methods, including SCOPE (Wang et al. 2020) and HMMcopy (https://rdrr.io/bioc/HMMcopy/), using spike-in simulations. Because these two methods were only designed for the detection of CNAs without cell clustering, they were followed by two commonly used clustering methods, hierarchical (Bridges 1966) and k-means (James and others 1967). The simulation mimicked a scDNA-seq data set of frozen breast tissue, the BRCA5 data set (data description), by randomly selecting 220 cells from the data set and using SCOPE and HMMcopy to remove genetic regions with putative CNAs. Genetic regions with copy number changes detected by either method were excluded from the analysis, and the remaining sequences were treated as copy number–free sequences. Among these cells, 20 cells were randomly selected as reference cells for the SCOPE method, and signals of spiked-in CNAs were added to the remaining cells. To evaluate the robustness of FLCNA, we randomly generated CNAs of varied sizes (supershort: two to five markers; short: five to 10 markers; medium: 10 to 20 markers; long: 20 to 35 markers) and simulated different numbers of clusters (three or five). We evaluated varied copy number states, including Del.d, Del.s, Dup.s, and Dup.d, respectively. Moreover, data with a mixture of the above four different copy number states were generated to mimic real-world data. Because a CNA may not be shared by all the cells in a cluster, different CNA sharing proportions (20%, 40%, 60%, 80%, 100%) were considered. For each cluster, we added 50 CNA segments to the background sequences. Additionally, we evaluated the performance on the scenarios with different numbers of CNAs among clusters by assigning random numbers of CNAs (20 to 80) to each cluster. For each of these scenarios, signals were spiked in by multiplying the background depth of coverage by $c/2$, where c is a normal random variable following $N(0.4, 0.1^2)$ for Del.d, $N(1.2, 0.1^2)$ for Del.s, $N(2.8, 0.1^2)$ for Dup.s, and $N(4.2, 0.1^2)$ for Dup.d (Wang et al. 2020). The ARI (Vinh et al. 2009) was calculated to evaluate the clustering performance of these methods by comparing the identified clusters of each method to the predefined "true" classes. An ARI near one indicates that the clusters identified are completely consistent with the ground truth, whereas values closer to zero indicate random clustering. The performance of can detection for these methods was assessed using the $F1$ score $\left(2*\frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}\right)$, with precision rate defined as $\frac{\text{True Positives}}{\text{True Positives}+\text{False}_{\text{Positives}}}$ and recall rate defined as $\frac{\text{True Positives}}{\text{True Positives}+\text{False}_{\text{Negatives}}}$. Notably, for a CNA to be classified as a true-positive call, it must intersect with regions of "true" CNAs and show copy number states that align with the predefined "true" states.

FLCNA application to the TNBC data set for breast cancer subcloning

With the data set introduced previously, FLCNA was also applied to the TNBC data set of breast cancer with three unrelated patients (i.e., KTN126, KTN129, KTN302) who have been treated with NAC. For the HMMcopy and SCOPE methods, we adhered to their default parameter settings. Three candidate numbers of clusters K (i.e., two, three, four) and the default value for the tuning hyperparameter λ were used for the FLCNA method. We identified shared CNAs using FLCNA and mapped them to 575 significant genes from GWASs with breast cancer curated in the NHGRI-EBI GWAS Catalog (Sollis et al. 2023). Pathway and network analyses were conducted for these genes; gene set enrichment analysis (GSEA) (Subramanian et al. 2005) was conducted with enrichment of Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2016). Further, the summary statistics from GSEA were used to generate connection networks for these three patients using the enrichment map implemented in Cytoscape (Shannon et al. 2003).

Software availability

FLCNA is available as an open-source tool at GitHub (https://github.com/FeifeiXiao-lab/FLCNA) and as Supplemental Code.

Downloaded from genome.cshlp.org on November 14, 2026 . Published by Cold Spring Harbor Laboratory Press

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was supported by the U.S. National Institutes of Health grant R21 HG010925 (F.X.).

References

Baslan T, Hicks J. 2017. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. *Nat Rev Cancer* **17:** 557–569. doi:10.1038/nrc.2017.58

Bridges CC. 1966. Hierarchical cluster analysis. *Psychol Rep* **18:** 851–854. doi:10.2466/pr0.1966.18.3.851

Burandt E, Jens G, Holst F, Jänicke F, Müller V, Quaas A, Choschzick M, Wilczak W, Terracciano L, Simon R, et al. 2013. Prognostic relevance of AIB1 (NCoA3) amplification and overexpression in breast cancer. *Breast Cancer Res Treat* **137:** 745–753. doi:10.1007/s10549-013-2406-4

Cariati F, Borrillo F, Shankar V, Nunziato M, D'Argenio V, Tomaiuolo R. 2019. Dissecting intra-tumor heterogeneity by the analysis of copy number variations in single cells: the neuroblastoma case study. *Int J Mol Sci* **20:** 893. doi:10.3390/ijms20040893

Dagogo-Jack I, Shaw AT. 2018. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol* **15:** 81–94. doi:10.1038/nrclinonc.2017.166

Dai Z, Liu P. 2021. High copy number variations, particular transcription factors, and low immunity contribute to the stemness of prostate cancer cells. *J Transl Med* **19:** 206. doi:10.1186/s12967-021-02870-x

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the *EM* algorithm. *J R Stat Soc Ser B* **39:** 1–22. doi:10.1111/j.2517-6161.1977.tb01600.x

Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. 2015. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol* **16:** 35. doi:10.1186/s13059-015-0602-8

Elyanow R, Zeira R, Land M, Raphael BJ. 2021. STARCH: copy number and clone inference from spatial transcriptomics data. *Phys Biol* **18:** 035001. doi:10.1088/1478-3975/abbe99

Fan J, Li R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* **96:** 1348–1360. doi:10.1198/016214501753382273

Ferronika P, van den Bos H, Taudt A, Spierings DCJ, Saber A, Hiltermann TJN, Kok K, Porubsky D, van der Wekken AJ, Timens W, et al. 2017. Copy number alterations assessed at the single-cell level revealed mono- and polyclonal seeding patterns of distant metastasis in a small-cell lung cancer patient. *Ann Oncol* **28:** 1668–1670. doi:10.1093/annonc/mdx182

Fischer K, Pflugfelder GO. 2015. Putative breast cancer driver mutations in *TBX3* cause impaired transcriptional repression. *Front Oncol* **5:** 244. doi:10.3389/fonc.2015.00244

Foulkes WD, Smith IE, Reis-Filho JS. 2010. Triple-negative breast cancer. *N Engl J Med* **363:** 1938–1948. doi:10.1056/NEJMra1001389

Gao R, Davis A, McDonald TO, Sei E, Shi X, Wang Y, Tsai P-C, Casasent A, Waters J, Zhang H, et al. 2016. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat Genet* **48:** 1119–1130. doi:10.1038/ng.3641

Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, et al. 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366:** 883–892. doi:10.1056/NEJMoa1113205

Guo J, Levina E, Michailidis G, Zhu J. 2010. Pairwise variable selection for high-dimensional model-based clustering. *Biometrics* **66:** 793–804. doi:10.1111/j.1541-0420.2009.01341.x

Ha G, Roth A, Lai D, Bashashati A, Ding J, Goya R, Giuliany R, Rosner J, Oloumi A, Shumansky K, et al. 2012. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res* **22:** 1995–2007. doi:10.1101/gr.137570.112

Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM, Melnyk N, McPherson A, Bashashati A, Laks E, et al. 2014. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res* **24:** 1881–1893. doi:10.1101/gr.180281.114

Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, Wu X, Wen L, Tang F, Huang Y, et al. 2016. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res* **26:** 304–319. doi:10.1038/cr.2016.23

Jiang Y, Qiu Y, Minn AJ, Zhang NR. 2016. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc Natl Acad Sci* **113:** E5528–E5537.

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44:** D457–D462. doi:10.1093/nar/gkv1070

Kim C, Gao R, Sei E, Brandt R, Hartman J, Hatschek T, Crosetto N, Foukakis T, Navin NE. 2018. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell* **173:** 879–893.e13. doi:10.1016/j.cell.2018.03.041

Kuipers J, Tuncel MA, Ferreira P, Jahn K, Beerenwinkel N. 2020. Single-cell copy number calling and event history reconstruction. bioRxiv doi:10.1101/2020.04.28.065755

Laks E, McPherson A, Zahn H, Lai D, Steif A, Brimhall J, Biele J, Wang B, Masud T, Ting J, et al. 2019. Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell* **179:** 1207–1221.e22. doi:10.1016/j.cell.2019.10.026

Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration. 2011. The sequence read archive. *Nucleic Acids Res* **39:** D19–D21. doi:10.1093/nar/gkq1019

Li B, Li JZ. 2014. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biol* **15:** 473. doi:10.1186/s13059-014-0473-4

Liu F, Gu L-N, Shan B-E, Geng C-Z, Sang M-X. 2016. Biomarkers for EMT and MET in breast cancer: an update. *Oncol Lett* **12:** 4869–4876. doi:10.3892/ol.2016.5369

MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (ed. Lecam L, Neyman J), Vol. 1, pp. 281–297. University of California Press, Berkeley, CA.

Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, Mangano E, Battaglia C, Bonora E, Kurg A, Seri M, et al. 2013. EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol* **14:** R120. doi:10.1186/gb-2013-14-10-r120

Mallory XF, Edrisi M, Navin N, Nakhleh L. 2020. Methods for copy number aberration detection from single-cell DNA-sequencing data. *Genome Biol* **21:** 208. doi:10.1186/s13059-020-02119-8

Markowska M, Cąkała T, Miasojedow B, Aybey B, Juraeva D, Mazur J, Ross E, Staub E, Szczurek E. 2022. CONET: copy number event tree model of evolutionary tumor history for single-cell data. *Genome Biol* **23:** 128. doi:10.1186/s13059-022-02693-z

Navin NE. 2015. The first five years of single-cell cancer genomics and beyond. *Genome Res* **25:** 1499–1507. doi:10.1101/gr.191098.115

Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, et al. 2011. Tumour evolution inferred by single-cell sequencing. *Nature* **472:** 90–94. doi:10.1038/nature09807

Oesper L, Mahmoody A, Raphael BJ. 2013. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol* **14:** R80. doi:10.1186/gb-2013-14-7-r80

Polyak K. 2014. Tumor heterogeneity confounds and illuminates: a case for Darwinian tumor evolution. *Nat Med* **20:** 344–346. doi:10.1038/nm.3518

Ren X, Kang B, Zhang Z. 2018. Understanding tumor ecosystems by single-cell sequencing: promises and limitations. *Genome Biol* **19:** 211. doi:10.1186/s13059-018-1593-z

Rojas C, Wahlberg B. 2014. On change point detection using the fused lasso method. arXiv:1401.5408 [math.ST]. doi:10.48550/arXiv.1401.5408

Seryakov A, Magomedova Z, Suntsova M, Prokofieva A, Rabushko E, Glusker A, Makovskaia L, Zolotovskaia M, Buzdin A, Sorokin M. 2021. RNA sequencing for personalized treatment of metastatic leiomyosarcoma: case report. *Front Oncol* **11:** 666001. doi:10.3389/fonc.2021.666001

Shah SP, Xuan X, DeLeeuw RJ, Khojasteh M, Lam WL, Ng R, Murphy KP. 2006. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics* **22:** e431–e439. doi:10.1093/bioinformatics/btl238

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13:** 2498–2504. doi:10.1101/gr.1239303

Shi S, Xu C, Fang X, Zhang Y, Li H, Wen W, Yang G. 2020. Expression profile of toll-like receptors in human breast cancer. *Mol Med Rep* **21:** 786–794. doi:10.3892/mmr.2019.10853

Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, Groza T, Güneş O, Hall P, Hayhurst J, et al. 2023. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51:** D977–D985. doi:10.1093/nar/gkac1010

Stanta G, Bonin S. 2018. Overview on clinical relevance of intra-tumor heterogeneity. *Front Med* **5:** 85. doi:10.3389/fmed.2018.00085

Downloaded from genome.cshlp.org on November 14, 2026 . Published by Cold Spring Harbor Laboratory Press

Subramani R, Nandy SB, Pedroza DA, Lakshmanaswamy R. 2017. Role of growth hormone in breast cancer. *Endocrinology* **158:** 1543–1555. doi:10.1210/en.2016-1928

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* **102:** 15545–15550. doi:10.1073/pnas.0506580102

Tang X, Huang Y, Lei J, Luo H, Zhu X. 2019. The single-cell sequencing: new developments and medical applications. *Cell Biosci* **9:** 53. doi:10.1186/s13578-019-0314-y

Tokumaru Y, Oshi M, Katsuta E, Yan L, Satyananda V, Matsuhashi N, Futamura M, Akao Y, Yoshida K, Takabe K. 2020. KRAS signaling enriched triple negative breast cancer is associated with favorable tumor immune microenvironment and better survival. *Am J Cancer Res* **10:** 897–907.

Vinh NX, Epps J, Bailey J. 2009. Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, pp. 1073–1080. Association for Computing Machinery, New York. doi:10.1145/1553374.1553511

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LAJ, Kinzler KW. 2013. Cancer genome landscapes. *Science* **339:** 1546–1558. doi:10.1126/science.1235122

Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, Chen K, Scheet P, Vattathil S, Liang H, et al. 2014. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512:** 155–160. doi:10.1038/nature13600

Wang R, Lin D-Y, Jiang Y. 2020. SCOPE: a normalization and copy-number estimation method for single-cell DNA sequencing. *Cell Syst* **10:** 445–452.e6. doi:10.1016/j.cels.2020.03.005

Xia H, Li A, Yu Z, Liu X, Feng H. 2015. A novel framework for analyzing somatic copy number aberrations and tumor subclones for paired heterogeneous tumor samples. *Biomed Mater Eng* **26 Suppl 1:** S1845–S1853.

Xiao F, Luo X, Hao N, Niu YS, Xiao X, Cai G, Amos CI, Zhang H. 2019. An accurate and powerful method for copy number variation detection. *Bioinformatics* **35:** 2891–2898. doi:10.1093/bioinformatics/bty1041

Yachida S, Jones S, Bozic I, Antal T, Leary R, Fu B, Kamiyama M, Hruban RH, Eshleman JR, Nowak MA, et al. 2010. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467:** 1114–1117. doi:10.1038/nature09515

Yu Z, Du F, Song L. 2022. SCClone: accurate clustering of tumor single-cell DNA sequencing data. *Front Genet* **13:** 823941. doi:10.3389/fgene.2022.823941

Zaccaria S, Raphael BJ. 2021. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *Nat Biotechnol* **39:** 207–214. doi:10.1038/s41587-020-0661-6

Zafar H, Navin N, Chen K, Nakhleh L. 2019. SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome Res* **29:** 1847–1859. doi:10.1101/gr.243121.118

Zhang J, Fujimoto J, Zhang J, Wedge DC, Song X, Zhang J, Seth S, Chow C-W, Cao Y, Gumbs C, et al. 2014. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **346:** 256–259. doi:10.1126/science.1256930

Zhao J, Zhao D, Poage GM, Mazumdar A, Zhang Y, Hill JL, Hartman ZC, Savage MI, Mills GB, Brown PH. 2015. Death-associated protein kinase 1 promotes growth of p53-mutant cancers. *J Clin Invest* **125:** 2707–2720. doi:10.1172/JCI70805

Zou H. 2006. The adaptive lasso and its oracle properties. *J Am Stat Assoc* **101:** 1418–1429. doi:10.1198/016214506000000735

Received May 15, 2023; accepted in revised form January 8, 2024.