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], Feifei Xiao[4*]

[1]*Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, SC, 29208, USA*

[2]*Department of Environmental Health Science, Arnold School of Public Health, University of South Carolina, Columbia, SC, 29208, USA*

[3]*Department of Quantitative Sciences, Baylor College of Medicine, Houston, TX 77030, USA*

[4]*Department of Biostatistics, College of Public Health and Health Professions and College of Medicine, University of Florida, Gainesville, FL, 32603, USA*

** To whom correspondence should be addressed*

Corresponding to:

Feifei Xiao, Ph.D.

Department of Biostatistics, College of Public Health and Health Professions and College of Medicine, University of Florida

2004 Mowry Rd., CTRB building Room 5227, Gainesville, FL, 32603

Tel: (352) 294-5917
Email: feifeixiao@ufl.edu

Outlines

Supplemental Methods

Parameter estimation using expectation–maximization algorithm

In FLCNA, the parameter set $\boldsymbol{\Theta} = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}, \pi_k\}_{k=1}^{K}$ are estimated using an expectation–maximization (EM) algorithm. let $\Delta_{j,k}$ be an indicator function of the hidden cluster information for $\boldsymbol{x}_j$, $\Delta_{j,k}= 1$ if $\boldsymbol{x}_j$ is from the k-th cluster, and $\Delta_{j,k}= 0$ otherwise. Assuming $\Delta_{j,k}$ is unobserved, the penalized log-likelihood function for the complete data will be given by

$$Q_P(\boldsymbol{\Theta}) = \sum_{j=1}^{N}\sum_{k=1}^{K}\Delta_{j,k}\{\log(\pi_k) + \log f_k(\boldsymbol{x}_j;\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma})\} - \lambda\sum_{k=1}^{K}\sum_{i=1}^{P-1}\tau_{i,\,i+1}^{(k)}|\mu_{i,k} - \mu_{i+1,k}|. \quad (1)$$

With Eq. (1), the parameter set $\boldsymbol{\Theta} = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}, \pi_k\}_{k=1}^{K}$ can be estimated with the EM algorithm by the following iterative procedure. The EM algorithm iterates between E-step and M-step, and produces a sequence of estimates $\widehat{\boldsymbol{\Theta}}^{(t)}$, $t = 0, 1, 2, ...$

1) Initialization: We first estimate the starting values $\widehat{\boldsymbol{\Theta}}^{(0)} = \{\widehat{\boldsymbol{\mu}}_k^{(0)}, \widehat{\boldsymbol{\Sigma}}^{(0)}, \hat{\pi}_k^{(0)}\}_{k=1}^{K}$ using model without penalty (λ=0).

2) Iteration:

E-step:

We start with the E-step given the current parameter estimates $\widehat{\boldsymbol{\Theta}}^{(t)}$. In this step, we calculate the probability for sample j belongs to k -th cluster with

$$\hat{\Delta}_{j,k}^{(t+1)} = E\left(\Delta_{j,k}\middle|\boldsymbol{X}, \widehat{\boldsymbol{\Theta}}^{(t)}\right) = \Pr\left(\Delta_{j,k}= 1\middle|\boldsymbol{X}, \widehat{\boldsymbol{\Theta}}^{(t)}\right) = \frac{\hat{\pi}_k^{(t)} f_k(\boldsymbol{x}_j; \widehat{\boldsymbol{\mu}}_k^{(t)}, \widehat{\boldsymbol{\Sigma}}^{(t)})}{\sum_{k'=1}^{K}\hat{\pi}_{k'}^{(t)} f_k(\boldsymbol{x}_j; \widehat{\boldsymbol{\mu}}_{k'}^{(t)}, \widehat{\boldsymbol{\Sigma}}^{(t)})}, \quad (2)$$

where the numerator is the density for j-th sample to be clustered into k-th cluster, and the denominator is the sum of densities for j-th sample to be clustered into K different clusters. Then

Eq. (2) will be plugged it into the Eq. (1) about $Q_P(\boldsymbol{\Theta})$ to estimate other parameters, including the cluster "weight" π_k, the variance for i-th marker σ_i^2 and cluster mean $\boldsymbol{\mu}$.

M-Step:

Given $\hat{\Delta}_{j,k}^{(t+1)}$ and $\widehat{\boldsymbol{\Theta}}^{(t)}$, the goal of M-step is to update parameter set $\widehat{\boldsymbol{\Theta}}^{(t+1)}$ by maximizing the log-likelihood function $Q_P\left(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}^{(t)}\right)$. Specifically, the estimate of "weights" $\pi_k's$ can be easily updated by taking the first derivative of $Q_P\left(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}^{(t)}\right)$ w.r.t. π_k with

$$\frac{\partial Q_P}{\partial \pi_k} = 0 \rightarrow \hat{\pi}_k^{(t+1)} = \frac{1}{N}\sum_{j=1}^{N} \hat{\Delta}_{j,k}^{(t+1)}. \quad (3)$$

Given $\hat{\Delta}_{j,k}^{(t+1)}$, $\hat{\pi}_k^{(t+1)}$ and $\hat{\mu}_{i,k}^{(t)}$, we can update the estimate of variance for i-th marker σ_i^2 by taking the first derivative of $Q_P\left(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}^{(t)}\right)$ w.r.t. σ_i^2 with

$$\frac{\partial Q_P}{\partial \sigma_i^2} = 0 \rightarrow \left(\hat{\sigma}_i^{(t+1)}\right)^2 = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{K} \hat{\Delta}_{j,k}^{(t+1)}\left(x_{i,j} - \hat{\mu}_{i,k}^{(t)}\right)^2, 1 \le j \le p. \quad (4)$$

Given $\hat{\Delta}_{j,k}^{(t+1)}$, $\hat{\pi}_k^{(t+1)}$ and $\hat{\sigma}_i^{(t+1)}$, according to Eq. (1), after some transformation, we can update the estimates of mean values $\widehat{\boldsymbol{\mu}}^{(t+1)}$ with

$$\widehat{\boldsymbol{\mu}}^{(t+1)} = \underset{\boldsymbol{\mu}}{\operatorname{argmin}} \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{K}\left\{\hat{\Delta}_{j,k}^{(t+1)}\sum_{i=1}^{p}\frac{\left(x_{i,j} - \mu_{i,k}\right)^2}{\left(\hat{\sigma}_i^{(t)}\right)^2}\right\} + \lambda\sum_{k=1}^{K}\sum_{i}^{P-1}\tau_{i,i+1}^{(k)}\left|\mu_{i,k} - \mu_{i+1,k}\right|. \quad (5)$$

Eq. (5) cannot be solved directly with close form, but $\widehat{\boldsymbol{\mu}}^{(t+1)}$ can be estimated using a local quadratic approximation (LQA) algorithm, which will be discussed in detail next.

Estimation of $\widehat{\boldsymbol{\mu}}^{(t+1)}$ using local quadratic approximation

According to LQA, we can approximate

$$\left|\mu_{i,k}^{(s+1)} - \mu_{i+1,k}^{(s+1)}\right| \approx \frac{\left(\mu_{i,k}^{(s+1)} - \mu_{i+1,k}^{(s+1)}\right)^2}{2\left|\hat{\mu}_{i,k}^{(s)} - \hat{\mu}_{i+1,k}^{(s)}\right|} + \frac{1}{2}\left|\hat{\mu}_{i,k}^{(s)} - \hat{\mu}_{i+1,k}^{(s)}\right|, \tag{6}$$

where s is the iteration index used to denote iterations of the LQA within the M-step (different from iteration index t in the EM algorithm), and $\widehat{\boldsymbol{\mu}}^{(s)}$ are the estimates from the previous iteration. Thus, the minimization problem in Eq. (5) has been converted into a generalized quadratic problem which has close form solution. Notably, Eq. (5) can be decomposed into K separate minimization problems. For example, for each k, we can solve (iteratively over s)

$$\min_{\boldsymbol{\mu}_k^{(s+1)}} \frac{1}{2}\sum_{j=1}^{N}\left\{\hat{\Delta}_{j,k}^{(t+1)}\sum_{i=1}^{p}\frac{\left(x_{i,j} - \hat{\mu}_{i,k}^{(s+1)}\right)^2}{\left(\hat{\sigma}_i^{(t)}\right)^2}\right\} + \lambda\sum_{i=1}^{P-1}\tau_{i,i+1}^{(k)}\frac{\left(\mu_{i,k}^{(s+1)} - \mu_{i+1,k}^{(s+1)}\right)^2}{2\left|\hat{\mu}_{i,k}^{(s)} - \hat{\mu}_{i+1,k}^{(s)}\right|}, \tag{7}$$

with close form. To solve Eq. (7), we need to transfer it into matrix form first. Let

- $\hat{\Delta}_k^{(t+1)} = \left(\hat{\Delta}_{1,k}^{(t+1)}, \ldots, \hat{\Delta}_{N,k}^{(t+1)}\right)^T$ be the estimated latent variable for the k-cluster from E-step in the EM algorithm.
- $\boldsymbol{J}_{N\times 1} = (1, \ldots, 1)^T$ is a matrix with all elements to be 1.
- $\tilde{\boldsymbol{\mu}}_k = \left(\tilde{\mu}_{1,k}, \ldots, \tilde{\mu}_{P,k}\right)^T$ is the pre-defined mean vector for the k-th cluster where $\tilde{\mu}_{i,k}$ is estimated from the model without any penalization ($\lambda = 0$).
- $\widehat{\boldsymbol{\mu}}_k^{(s)} = \left(\hat{\mu}_{1,k}^{(s)}, \ldots, \hat{\mu}_{P,k}^{(s)}\right)^T$ is the estimate of mean vector for the k-th cluster from previous iteration in the EM algorithm.
- $\widehat{\boldsymbol{\mu}}_k^{(s+1)} = \left(\mu_{1,k}^{(s+1)}, \ldots, \mu_{P,k}^{(s+1)}\right)^T$ is the estimate of our interest which is the mean vector for the k-th cluster.

- $\boldsymbol{D} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{pmatrix}_{(P-1)\times P}$ is a matrix introduced to calculate the difference of mean values for each pair of consecutive markers in a cluster.

Then Eq. (1) can also be given with

$$\boldsymbol{G}\left(\boldsymbol{\mu}_k^{(s+1)}\right) = \left(\hat{\Delta}_k^{(t+1)}\right)^T \left(\boldsymbol{X} - \boldsymbol{\mu}_k^{(s+1)}\boldsymbol{J}^T\right)^2 \boldsymbol{\Sigma}^{-1}\boldsymbol{J} + \lambda \boldsymbol{D}^{\mathrm{T}}\left(\mathrm{diag}(\boldsymbol{C})\right)^2 \boldsymbol{D}\left(\boldsymbol{\mu}_k^{(s+1)}\right)^2, \qquad (8)$$

where $\boldsymbol{C} = \left[\mathrm{abs}(\boldsymbol{D}\tilde{\boldsymbol{\mu}}_k) \odot \mathrm{abs}\left(\boldsymbol{D}\hat{\boldsymbol{\mu}}_k^{(s)}\right)\right]^{-1/2}$.

Thus, we can easily find the solution for the quadratic equation of $\boldsymbol{G}\left(\boldsymbol{\mu}_k^{(s+1)}\right)$ with respect to $\boldsymbol{\mu}_k^{(s+1)}$,

$$\hat{\boldsymbol{\mu}}_k^{(s+1)} = \mathrm{argmin}(\boldsymbol{G}) = \left(\hat{\Delta}_k^{(t+1)}\right)^T \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left[\left(\hat{\Delta}_k^{(t+1)}\right)^T \boldsymbol{J}\boldsymbol{\Sigma}^{-1} + \lambda \boldsymbol{D}^{\mathrm{T}}(\mathrm{diag}(\boldsymbol{C}))^2\boldsymbol{D}\right]^{-1}.$$

Model selection

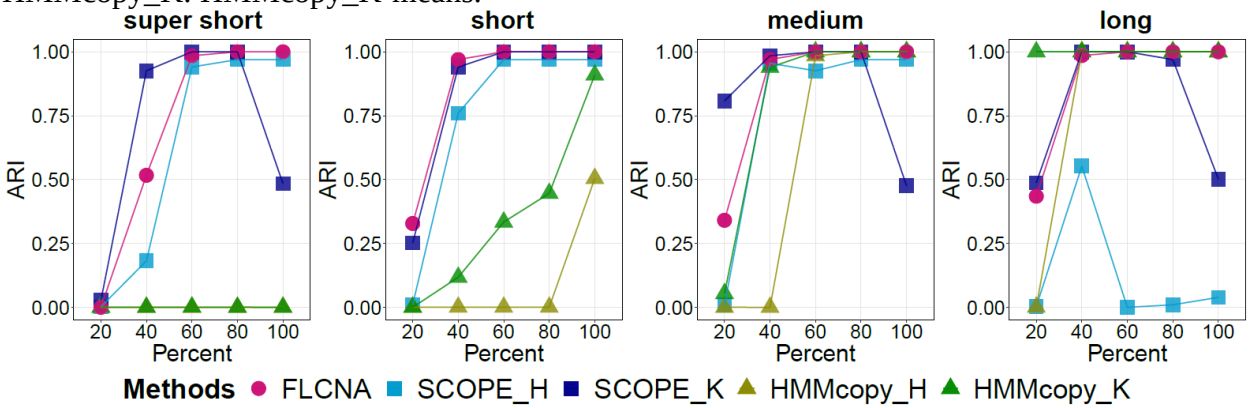
There are two hyperparameters to be pre-defined in the FLCNA method, including the number of clusters K and the tuning parameter λ. To find the optimal values of K and λ, we use a Bayesian information criterion (BIC), defined by

$$\mathrm{BIC}(K,\lambda) = -2\sum_{j=1}^{N} \log\left\{\sum_{k=1}^{K} \hat{\pi}_k\, f_k\left(\boldsymbol{x}_j; \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}\right)\right\} + d\,\mathrm{log}N. \qquad (9)$$

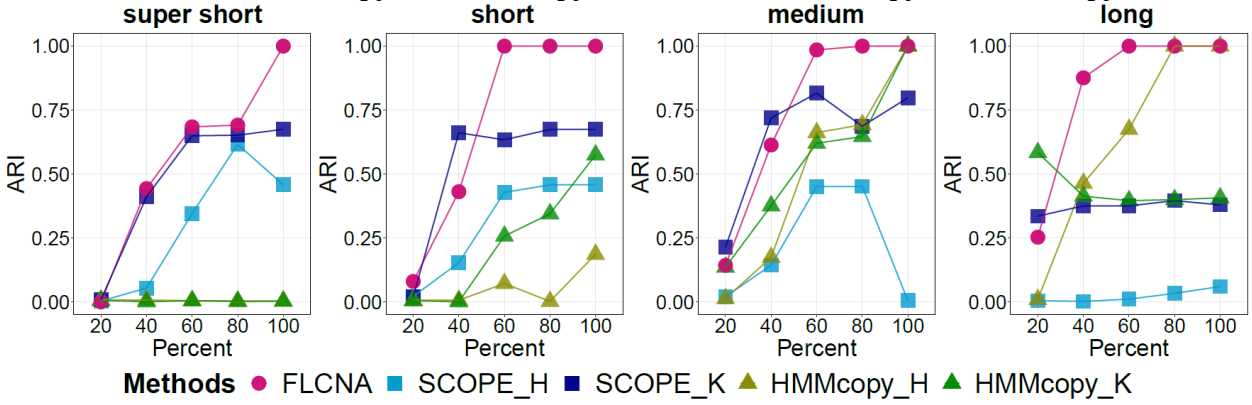
The degrees of freedom $d = K - 1 + P + e(\hat{\boldsymbol{\mu}})$, where $e(\hat{\boldsymbol{\mu}})$ is the number of distinct nonzero elements in $\hat{\boldsymbol{\mu}}$, and was used to adjust the number of breakpoints in degree of freedom. For each pair of parameter values (K, λ), the clustering model with smallest BIC value is selected as the optimal model and the corresponding parameters are estimated.

Supplemental Figures

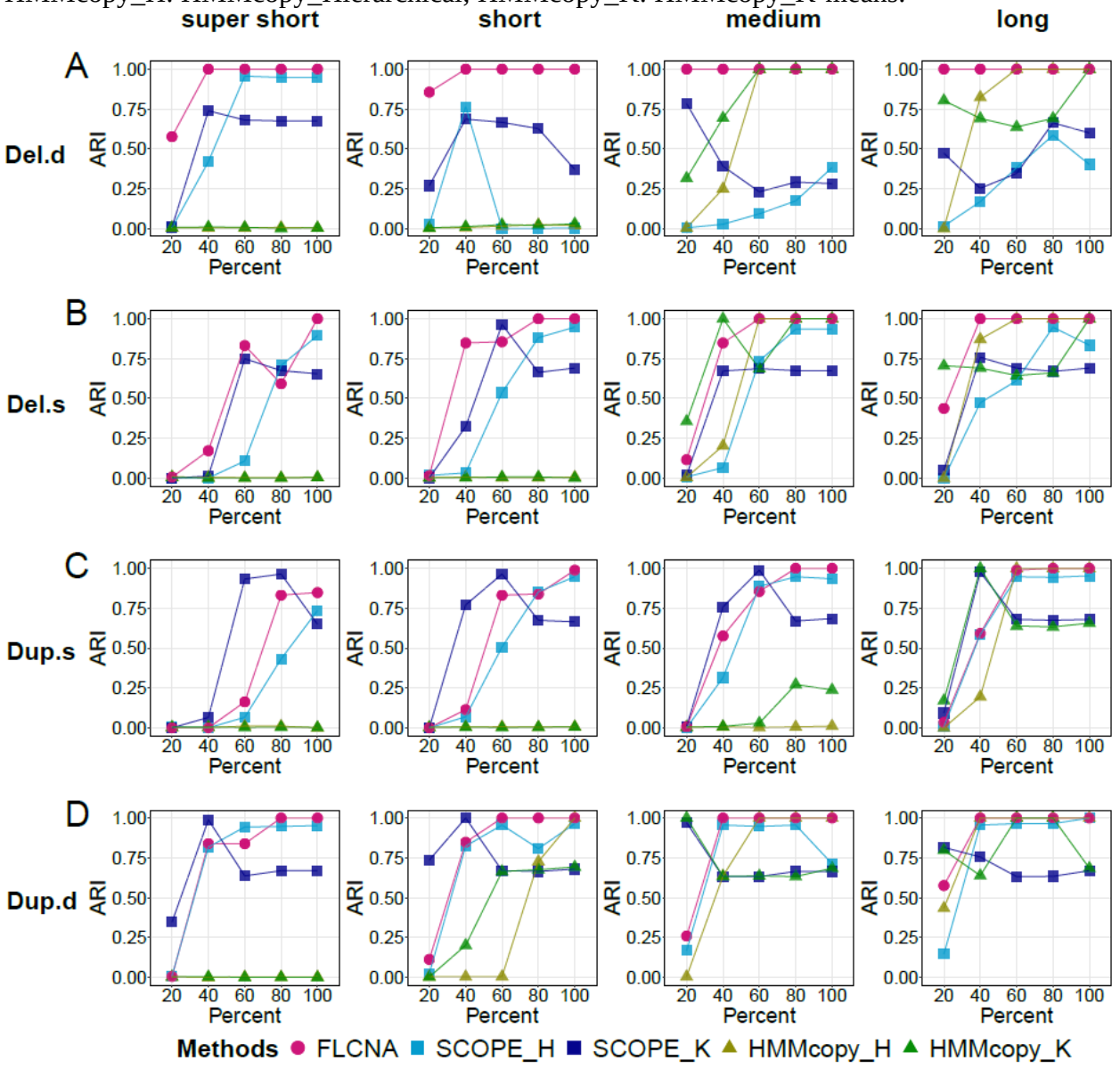
Supplemental Figure S1. Accuracy of clustering in simulated data with three clusters and mixed CNA states. Clustering results from FLCNA were compared to existing methods (i.e., SCOPE and HMMcopy) coupled with different clustering methods. For each of three clusters, we added signals of 50 CNA segments to the background signals with varied lengths (super short: 2~5 markers, short: 5~10 markers, medium: 10~20 markers, and long: 20~35 markers) and varied CNA proportions (20%, 40%, 60%, 80%, 100%), respectively. 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; HMMcopy_K: HMMcopy_K-means.



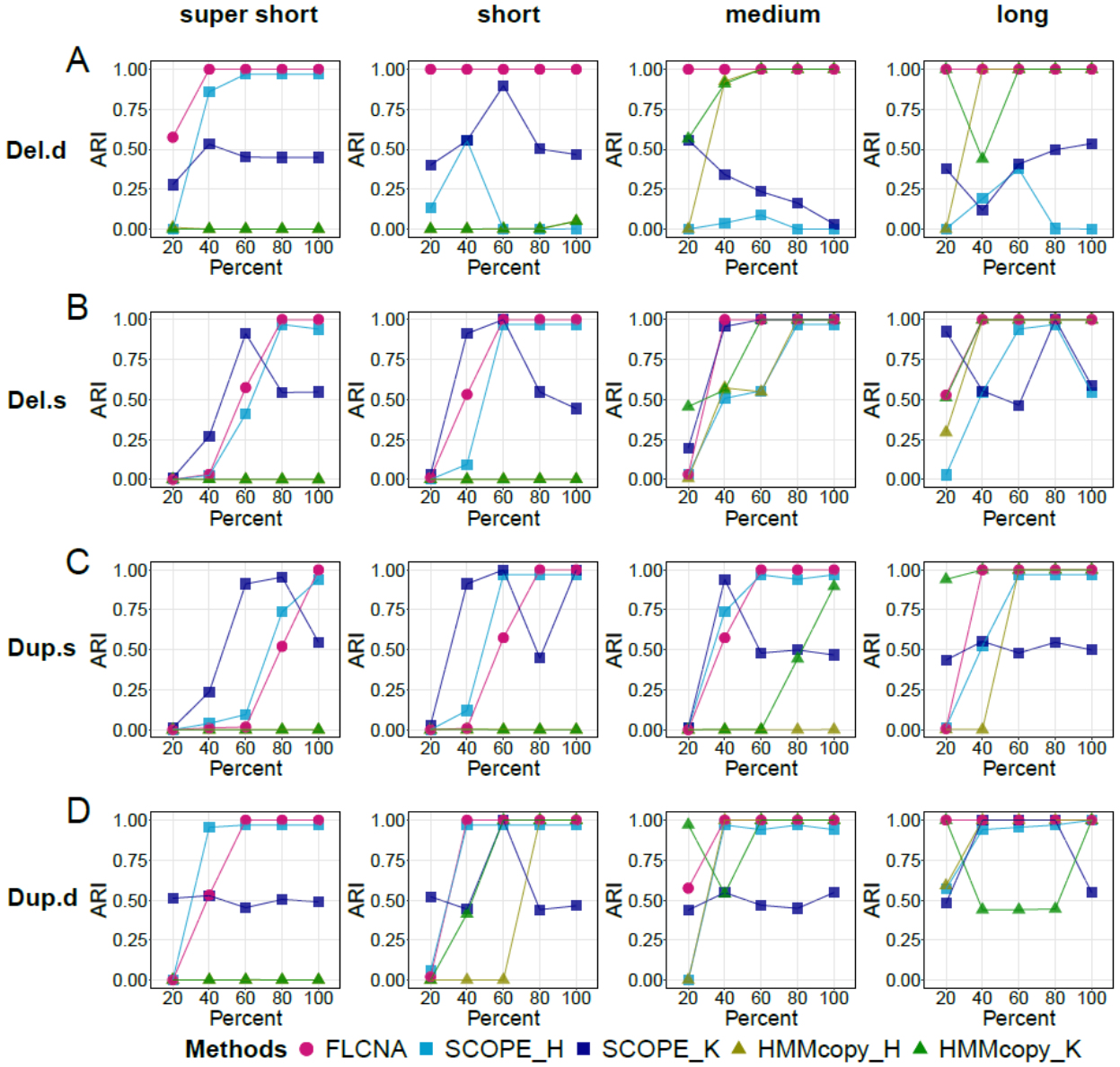
Supplemental Figure S2. Accuracy of clustering in simulated data with five clusters, varied numbers of CNAs and mixed CNA states. Clustering results from FLCNA were compared to existing methods (i.e., SCOPE and HMMcopy) coupled with different clustering methods. For each of five clusters, we added signals of varied numbers of CNA segments (20~80) to the background signals with varied lengths (super short: 2~5 markers, short: 5~10 markers, medium: 10~20 markers, and long: 20~35 markers) and varied CNA proportions (20%, 40%, 60%, 80%, 100%), respectively. 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; HMMcopy_K: HMMcopy_K-means.



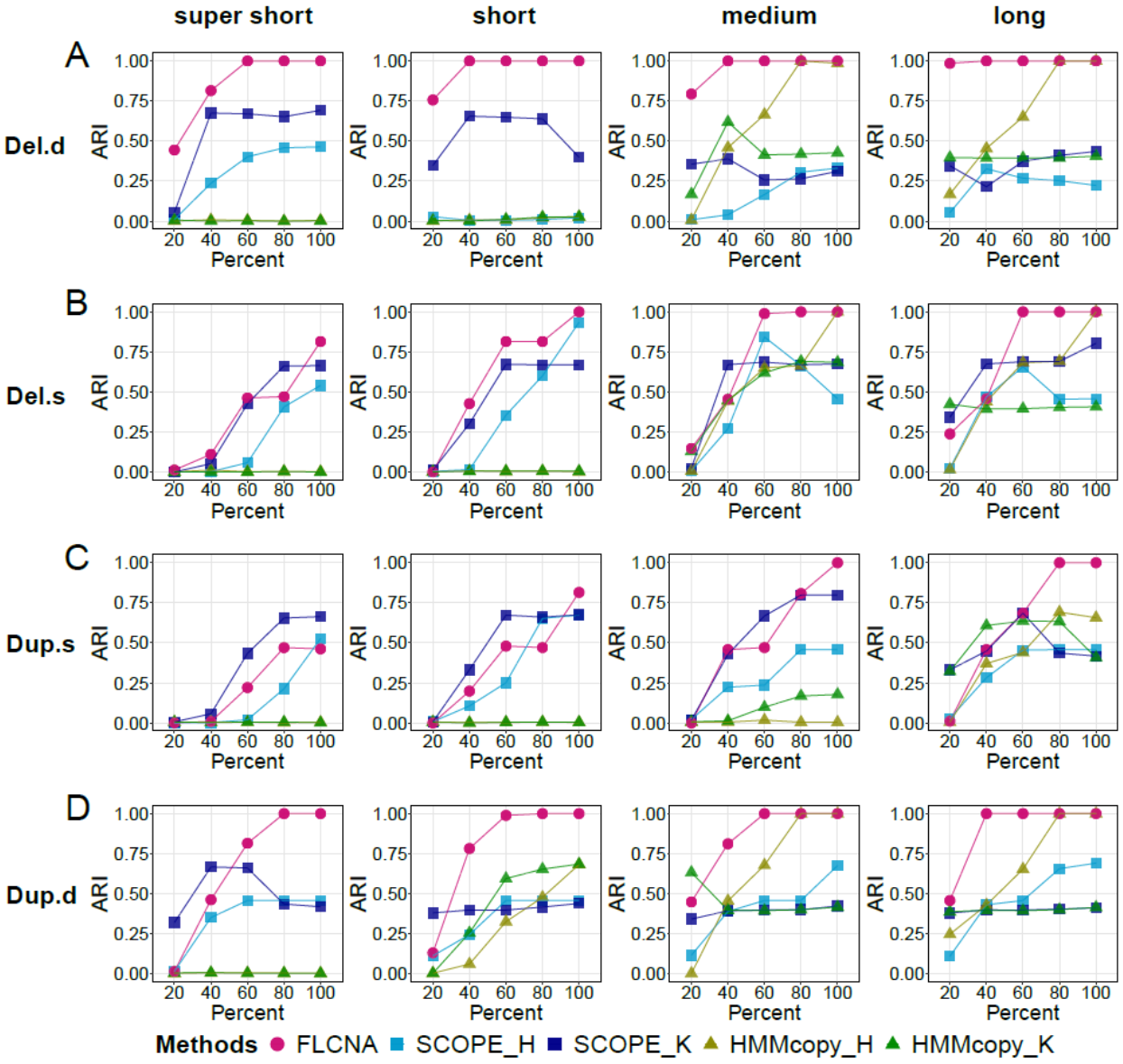
Supplemental Figure S3. Accuracy of clustering in simulated data with five clusters and a single type of CNA state. Clustering results from FLCNA were compared to existing methods (i.e., SCOPE and HMMcopy) coupled with different clustering methods. For each of five clusters, we added signals of 50 CNA segments to the background signals with varied lengths (super short: 2~5 markers, short: 5~10 markers, medium: 10~20 markers, and long: 20~35 markers) and varied CNA proportions (20%, 40%, 60%, 80%, 100%), respectively. Signals of Del.d (A), Del.s (B), Dup.s (C) and Dup.d (D) were spiked in separately. ARI: Adjusted Rand Index; Del.d: Deletion of double copies; Del.s: Deletion of a single copy; Dup.s: Duplication of a single copy; Dup.d: Duplication of double copies; SCOPE_H: SCOPE_Hierarchical; SCOPE_K: SCOPE_K-means; HMMcopy_H: HMMcopy_Hierarchical; HMMcopy_K: HMMcopy_K-means.



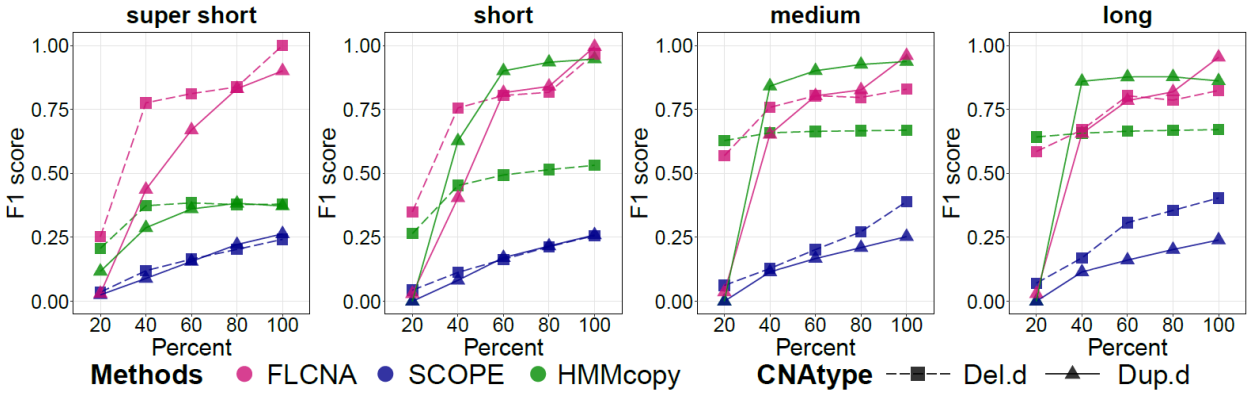
Supplemental Figure S4. Accuracy of clustering in simulated data with three clusters and a single type of CNA state. Clustering results from FLCNA were compared to existing methods (i.e., SCOPE and HMMcopy) coupled with different clustering methods. For each of three clusters, we added signals of 50 CNA segments to the background signals with varied lengths (super short: 2~5 markers, short: 5~10 markers, medium: 10~20 markers, and long: 20~35 markers) and varied CNA proportions (20%, 40%, 60%, 80%, 100%), respectively. Signals of Del.d (A), Del.s (B), Dup.s (C) and Dup.d (D) were spiked in separately. ARI: Adjusted Rand Index; Del.d: Deletion of double copies; Del.s: Deletion of a single copy; Dup.s: Duplication of a single copy; Dup.d: Duplication of double copies; SCOPE_H: SCOPE_Hierarchical; SCOPE_K: SCOPE_K-means; HMMcopy_H: HMMcopy_Hierarchical; HMMcopy_K: HMMcopy_K-means.



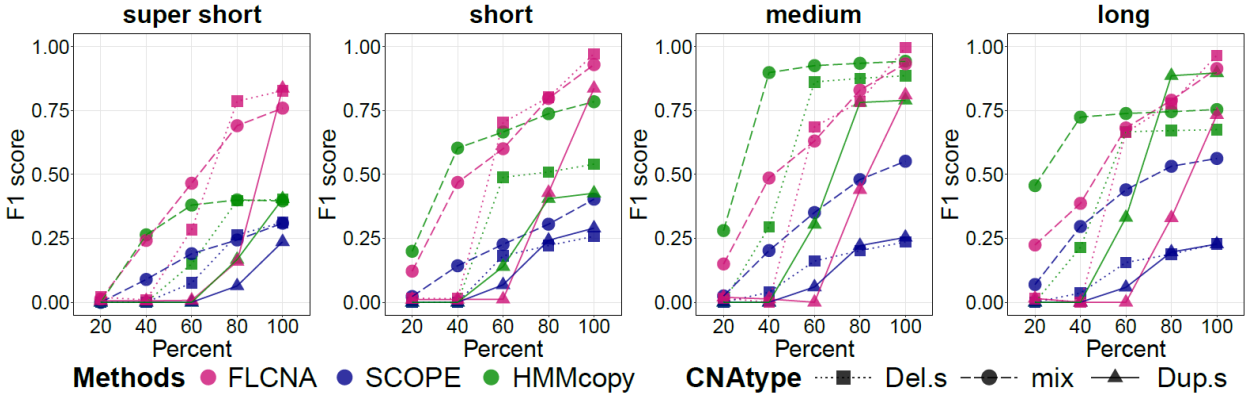
Supplemental Figure S5. Accuracy of clustering in simulated data with five clusters, varied numbers of CNAs and a single type of CNA state. Clustering results from FLCNA were compared to existing methods (i.e., SCOPE and HMMcopy) coupled with different clustering methods. For each of five clusters, we added signals of varied numbers of CNA segments (20~80) to the background signals with varied lengths (super short: 2~5 markers, short: 5~10 markers, medium: 10~20 markers, and long: 20~35 markers) and varied CNA proportions (20%, 40%, 60%, 80%, 100%), respectively. Signals of Del.d (A), Del.s (B), Dup.s (C) and Dup.d (D) were spiked in separately. ARI: Adjusted Rand Index; Del.d: Deletion of double copies; Del.s: Deletion of a single copy; Dup.s: Duplication of a single copy; Dup.d: Duplication of double copies; SCOPE_H: SCOPE_Hierarchical; SCOPE_K: SCOPE_K-means; HMMcopy_H: HMMcopy_Hierarchical; HMMcopy_K: HMMcopy_K-means.



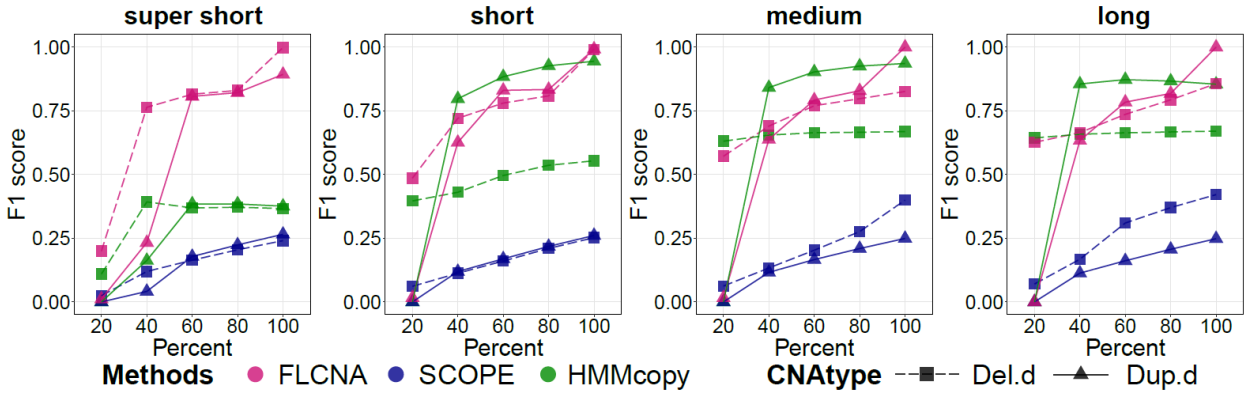
Supplemental Figure S6. Accuracy of CNA detection in simulated data with five clusters and aberration of double copies. CNA calls were generated by FLCNA, SCOPE and HMMcopy, respectively. For each of five clusters, we added signals of 50 CNA segments to the background signals with varied lengths (super short: 2~5 markers, short: 5~10 markers, medium: 10~20 markers, and long: 20~35 markers) and varied CNA proportions (20%, 40%, 60%, 80%, 100%), respectively. Deletion of double copies (Del.d) and duplication of double copies (Dup.d) were spiked in separately. *F*1 score was utilized to evaluate the performance of CNA detection for each method.



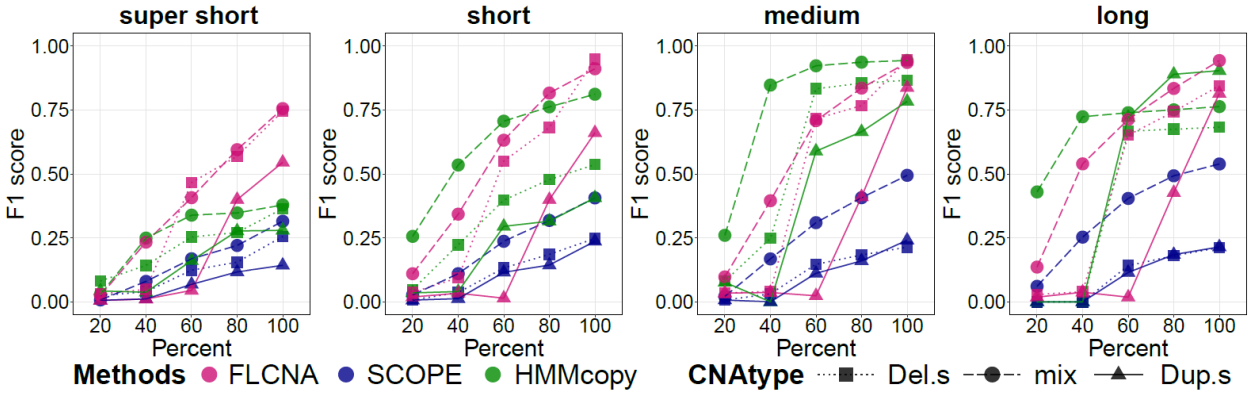
Supplemental Figure S7. Accuracy of CNA detection in simulated data with three clusters. CNA calls were generated by FLCNA, SCOPE and HMMcopy, respectively. For each of three clusters, we added signals of 50 CNA segments to the background signals with varied lengths (super short: 2~5 markers, short: 5~10 markers, medium: 10~20 markers, and long: 20~35 markers) and varied CNA proportions (20%, 40%, 60%, 80%, 100%), respectively. Deletion of a single copy (Del.s), mixed CNA states (mix) and duplication of a single copy (Dup.s) were spiked in separately. *F*1 score was utilized to evaluate the performance of CNA detection for each method.



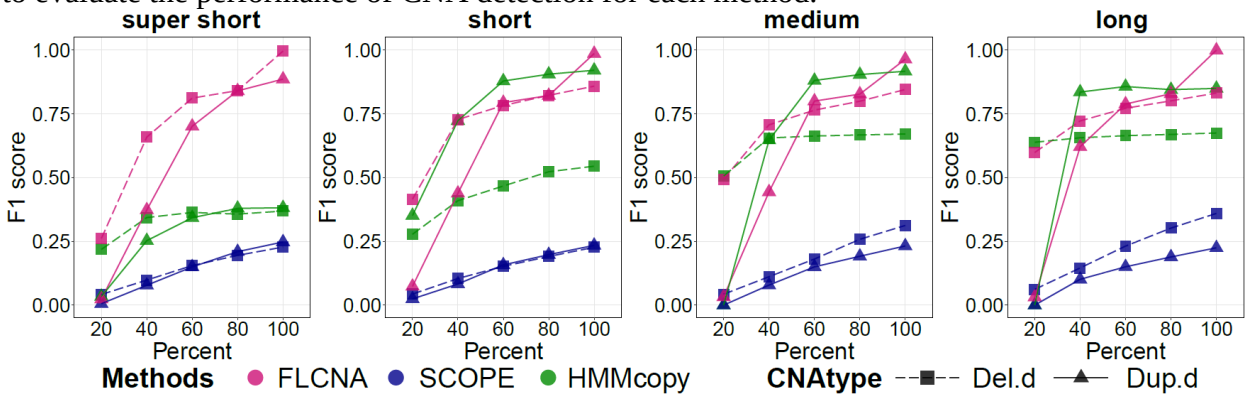
Supplemental Figure S8. Accuracy of CNA detection in simulated data with three clusters and aberration of double copies. CNA calls were generated by FLCNA, SCOPE and HMMcopy, respectively. For each of three clusters, we added signals of 50 CNA segments to the background signals with varied lengths (super short: 2~5 markers, short: 5~10 markers, medium: 10~20 markers, and long: 20~35 markers) and varied CNA proportions (20%, 40%, 60%, 80%, 100%), respectively. Deletion of double copies (Del.d) and duplication of double copies (Dup.d) were spiked in separately. *F*1 score was utilized to evaluate the performance of CNA detection for each method.



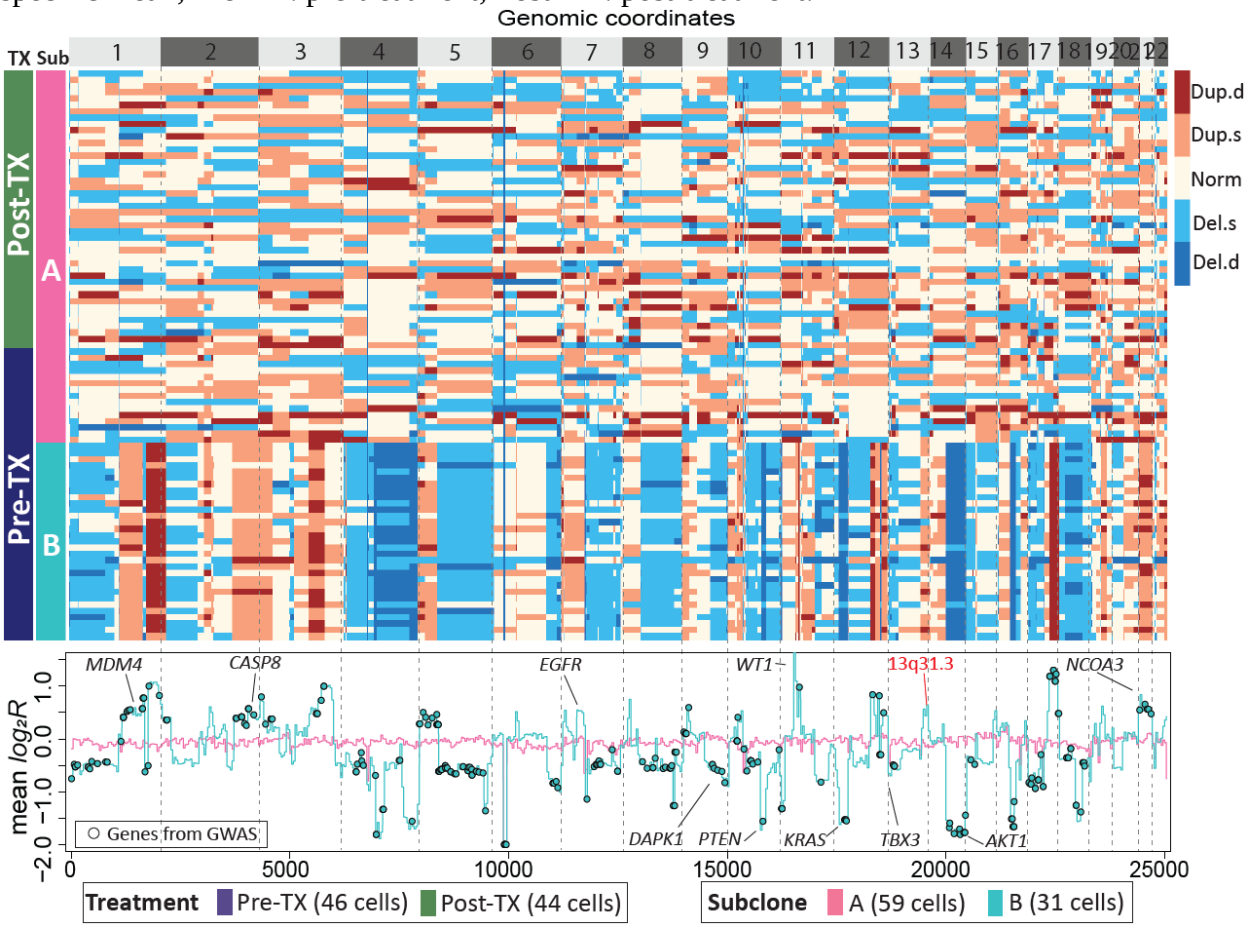
Supplemental Figure S9. Accuracy of CNA detection in simulated data with five clusters and varied numbers of CNAs. CNA calls were generated by FLCNA, SCOPE and HMMcopy, respectively. For each of five clusters, we added signals of varied numbers of CNA segments (20~80) to the background signals with varied lengths (super short: 2~5 markers, short: 5~10 markers, medium: 10~20 markers, and long: 20~35 markers) and varied CNA proportions (20%, 40%, 60%, 80%, 100%), respectively. Deletion of a single copy (Del.s), mixed CNA states (mix) and duplication of a single copy (Dup.s) were spiked in separately. *F*1 score was utilized to evaluate the performance of CNA detection for each method.



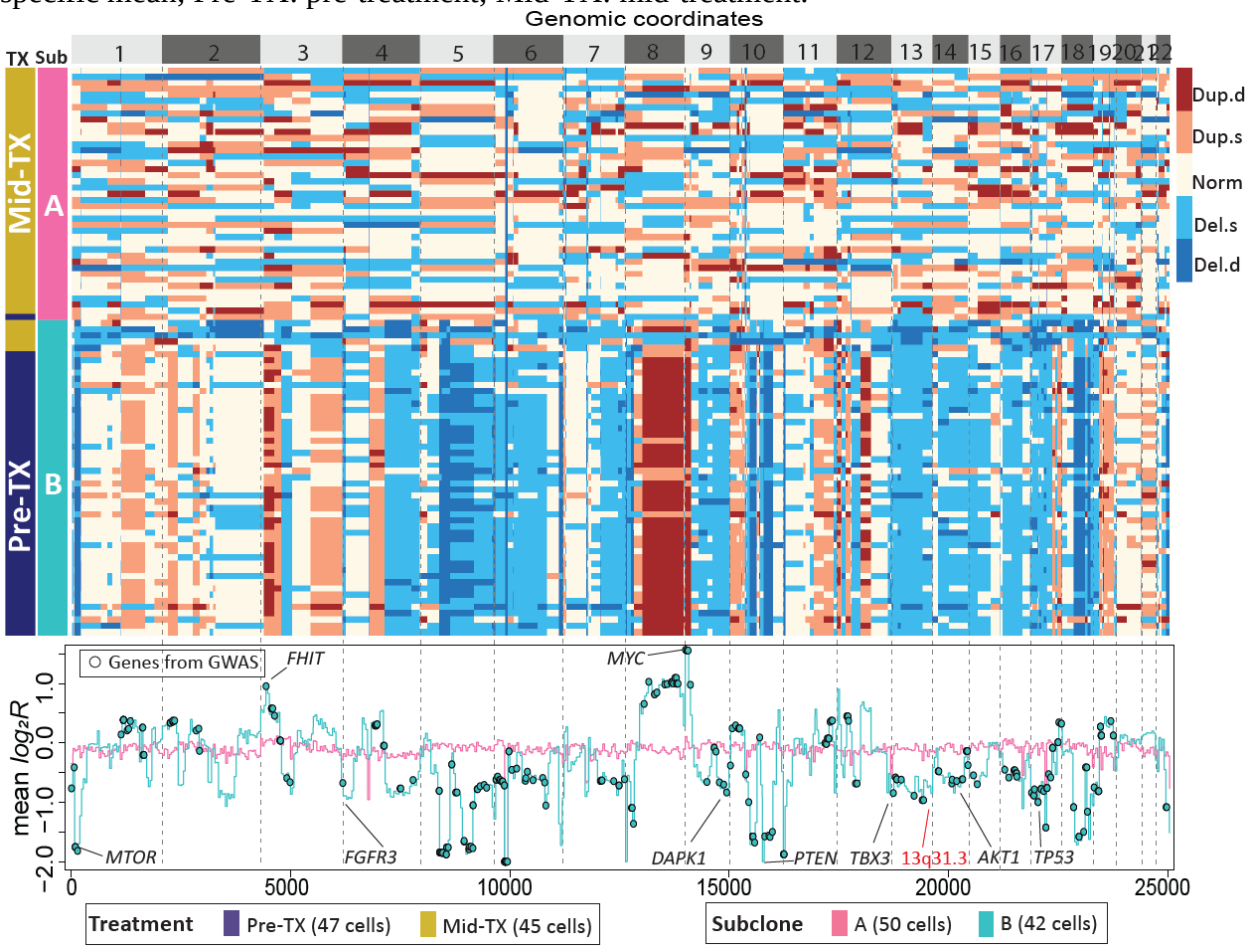
Supplemental Figure S10. Accuracy of CNA detection in simulated data with five clusters, varied numbers of CNAs and aberration of double copies. CNA calls were generated by FLCNA, SCOPE and HMMcopy, respectively. For each of five clusters, we added signals of varied numbers of CNA segments (20~80) to the background signals with varied lengths (super short: 2~5 markers, short: 5~10 markers, medium: 10~20 markers, and long: 20~35 markers) and varied CNA proportions (20%, 40%, 60%, 80%, 100%), respectively. Deletion of double copies (Del.d) and duplication of double copies (Dup.d) were spiked in separately. *F*1 score was utilized to evaluate the performance of CNA detection for each method.



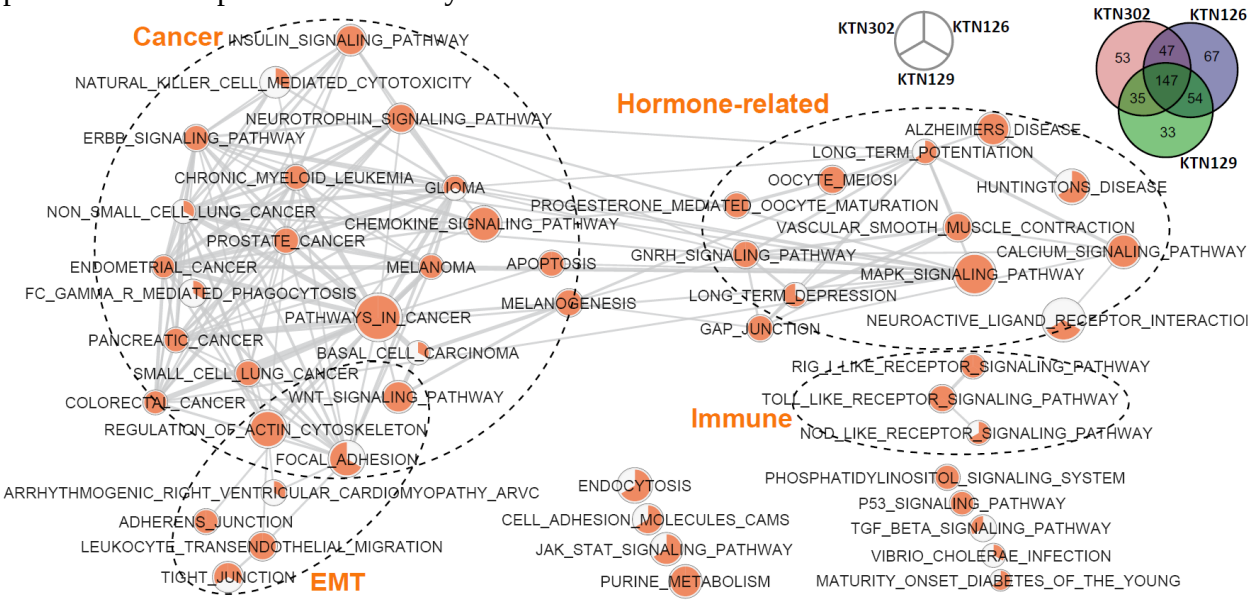
Supplemental Figure S11. Subclone clustering of KTN129 patient using FLCNA. Cell clusters and copy number profile with different CNA states (Del.d, Del.s, Norm, Dup.s and Dup.d) were generated using FLCNA. Mean *log₂R* were provided for each cluster. Shared CNAs identified using FLCNA were matched to significant genes from genome-wide association studies (GWAS) in the NHGRI-EBI GWAS Catalog. Del.d: Deletion of double copies; Del.s: Deletion of a single copy; Norm: Normal/diploid; Dup.s: Duplication of a single copy; Dup.d: Duplication of double copies; *log₂R*: Logarithm transformation of ratio between normalized read counts and its sample specific mean; Pre-TX: pre-treatment; Post-TX: post-treatment.



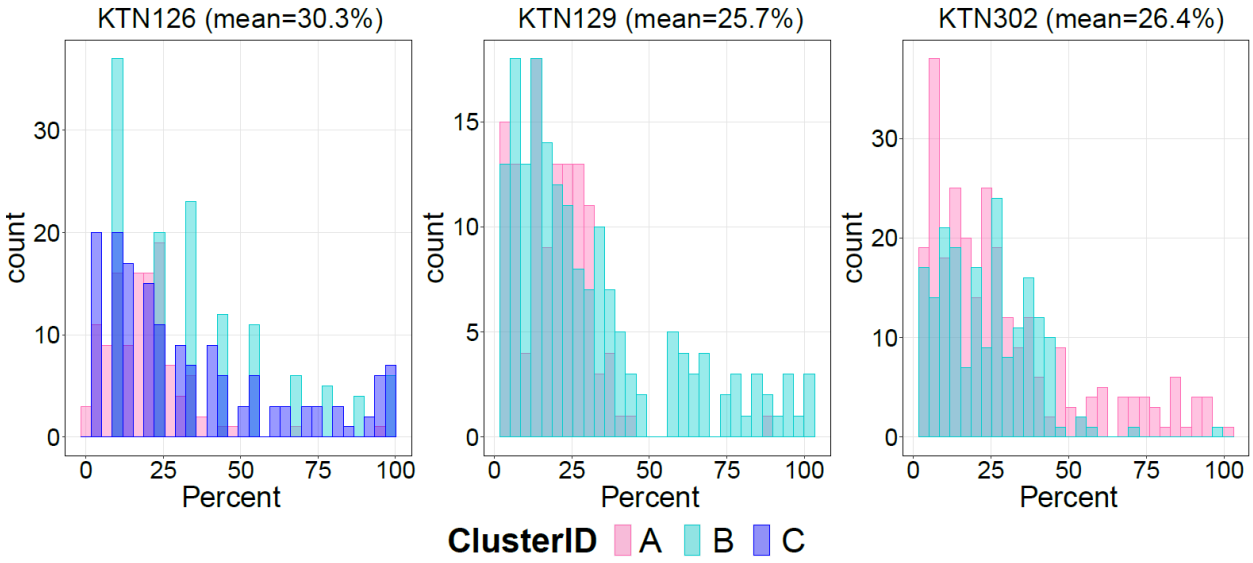
Supplemental Figure S12. Subclone clustering of KTN302 patient using FLCNA. Cell clusters and copy number profile with different CNA states (Del.d, Del.s, Norm, Dup.s and Dup.d) were generated using FLCNA. Mean log_2R were provided for each cluster. Shared CNAs identified using FLCNA were matched to significant genes from genome-wide association studies (GWAS) in the NHGRI-EBI GWAS Catalog. Del.d: Deletion of double copies; Del.s: Deletion of a single copy; Norm: Normal/diploid; Dup.s: Duplication of a single copy; Dup.d: Duplication of double copies; log_2R: Logarithm transformation of ratio between normalized read counts and its sample specific mean; Pre-TX: pre-treatment; Mid-TX: mid-treatment.



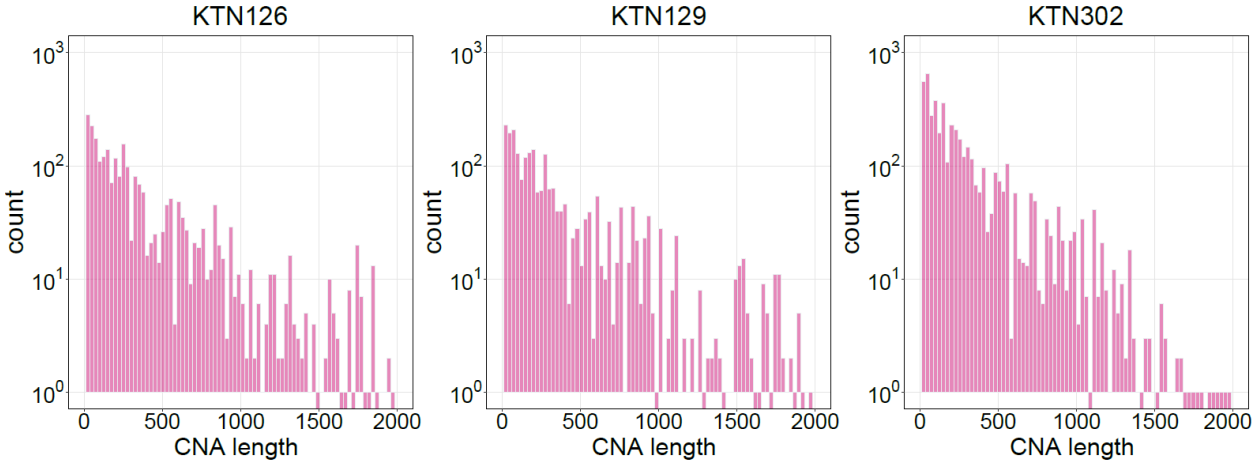
Supplemental Figure S13. Gene expression networks in the TNBC dataset. The shared CNAs identified using FLCNA were mapped into significant genes from the genome-wide association studies (GWAS) with breast cancer. These matched genes were utilized for KEGG pathway enrichment analysis for three patients (i.e., KTN126, KTN129, KTN302). Each node in network is a pie plot showing three patients. Node size corresponds to the number of genes within the pathway. Colors inner the node correspond to the index whether this pathway is identified in this patient. Edge weight corresponds to the number of genes found in both connected pathways. Venn diagrams show the distribution of genes from GWAS which were also detected from above three patients. EMT: epithelial-mesenchymal transition.



Supplemental Figure S14. Distribution of shared percentage for CNAs detected using FLCNA in the TNBC dataset. CNAs were identified from the TNBC dataset with three patients (KTN126, KTN129, KTN302) using the FLCNA method.



Supplemental Figure S15. Distribution of CNAs detected using FLCNA in the TNBC dataset. CNAs were identified from the TNBC dataset with three patients (KTN126, KTN129, KTN302) using the FLCNA method.



Supplemental Tables

Supplemental Table S4. Assessment of FLCNA to cluster cells using simulation data with a single cluster and mixed CNA states. Clustering purity was utilized to evaluate the clustering performance of FLCNA by dividing the number of accurately assigned cells with the total number of cells. We added signals of 50 CNA segments to the background signals with varied lengths (super short: 2~5 markers, short: 5~10 markers, medium: 10~20 markers, and long: 20~35 markers) and varied CNA proportions (20%, 40%, 60%, 80%, 100%), respectively.

Shared CNA proportion	super short	short	medium	long
20	1.000	1.000	1.000	1.000
40	1.000	1.000	1.000	1.000
60	1.000	1.000	1.000	1.000
80	1.000	1.000	1.000	0.795
100	1.000	1.000	1.000	0.795

Supplemental Table S8. Proportion of CNAs with sharing percentage > 60% within clusters. CNAs were identified from the TNBC dataset with three patients (KTN126, KTN129, KTN302) using the FLCNA method.

	Cluster A (%)	Cluster B (%)	Cluster C (%)
KTN126	1.64	16.94	18.18
KTN129	0.84	17.05	
KTN302	12.77	1.05	

Supplemental Table S9. Computational time of different CNA detection methods with scDNA-seq data. A high-performance cluster with 8 cores and 12GB RAM was used for CNA detection with KTN126 patient in the THBC dataset.

Methods	Time (hours)
FLCNA	1.20
SCOPE	10.5
HMMcopy	0.15