# HetRCNA: A Novel Method to Identify Recurrent Copy Number Alternations from Heterogeneous Tumor Samples Based on Matrix Decomposition Framework

Jianing Xi, Ao Li, and Minghui Wang

**Abstract**—A common strategy to discovering cancer associated copy number aberrations (CNAs) from a cohort of cancer samples is to detect recurrent CNAs (RCNAs). Although the previous methods can successfully identify communal RCNAs shared by nearly all tumor samples, detecting subgroup-specific RCNAs and their related subgroup samples from cancer samples with heterogeneity is still invalid for these existing approaches. In this paper, we introduce a novel integrated method called HetRCNA, which can identify statistically significant subgroup-specific RCNAs and their related subgroup samples. Based on matrix decomposition framework with weight constraint, HetRCNA can successfully measure the subgroup samples by coefficients of left vectors with weight constraint and subgroup-specific RCNAs by coefficients of the right vectors and significance test. When we evaluate HetRCNA on simulated dataset, the results show that HetRCNA gives the best performances among the competing methods and is robust to the noise factors of the simulated data. When HetRCNA is applied on a real breast cancer dataset, our approach successfully identifies a bunch of RCNA regions and the result is highly correlated with the results of the other two investigated approaches. Notably, the genomic regions identified by HetRCNA harbor many breast cancer related genes reported by previous researches.

**Index Terms**—Cancer genome, copy number aberrations, matrix decomposition, recurrent, bioinformatics

## 1 Introduction

Copy number aberrations (CNAs) are large segments of genome regions, of which the region size ranges from 1 kb to 3 Mb, with copy number amplifications and deletions [1]. In recent years, many researches have reported that there may be a strong association between CNAs and human diseases especially cancers [2], [3], [4], [5], [6]. Distinguishing functional CNAs in cancer genomes which are genetically contributing to the cancer phenotype (driver aberrations) from pathological irrelevant CNAs (passenger aberrations) is a crucial task for both the basis of oncogenesis and the diagnosis and treatment of cancers [5], [7], [8]. Recurrent CNAs (RCNAs) is then defined as CNAs that are altered in at least a set of samples [9]. The common strategy of finding driver aberrations is to find RCNAs, which is based on a wildly-accepted assumption that driver alterations are more likely to be shared by multiple samples while passenger alterations are subject-specific, present randomly among samples [9].

- *J. Xi is with the School of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China. E-mail: xjn@mail.ustc.edu.cn.*
- *A. Li and M. Wang are with the School of Information Science and Technology, and Centers for Biomedical Engineering, University of Science and Technology of China, Hefei, Anhui 230037, China. E-mail: {aoli, mhwang}@ustc.edu.cn.*

*Manuscript received 29 May 2018; accepted 6 June 2018. Date of publication 12 June 2018; date of current version 1 Apr. 2020.*
*(Corresponding author: Jianing Xi.)*
*Digital Object Identifier no. 10.1109/TCBB.2018.2846599*

To efficiently detect RCNAs, several strategies have been proposed [8], [9], [10]. Primitively, some approaches are introduced through straightforward way of observing the minimum common regions in all tumor samples [11], [12]. Afterwards, a bunch of approaches have been proposed with more delicate strategies [8], [10], [13], including approaches based on permutation test [14], [15], peeling-off after scoring [16], [17], [18], [19], [20], [21], correlation based analysis [22], matrix factorization [23], [24], and kernel smoothing frameworks [25], [26]. The application of the RCNA approaches above have successfully discovered many RCNAs, and many oncogenes and tumor suppressors are found to be harbored in these RCNA segments [27], [28], [29].

However, all the approaches above are designed without the consideration of CNA heterogeneity of cancer tumor samples. For many types of cancers, the CNA amplifications and deletions among genome display extensive heterogeneity with distinct CNA profiles in different groups of samples [30], [31], [32], [33], [34], [35]. Among the CNA profiles of different groups of samples, there are both CNAs shared publicly by nearly all samples (communal RCNAs) and CNAs which are significantly recurrent in their related subgroup samples (subgroup-specific RCNAs). Nevertheless, most existing methods are only capable for communal RCNAs. When the approached aforementioned are directly applied on heterogeneous tumor samples, some subgroup-specific RCNAs would be regarded as insignificant if only a small proportion of the tested samples belong to the related

subgroup. Consequently, there is an urgent need of methods capable of subgroup-specific RCNA calling.

To illustrate CNA profile with heterogeneity, three scenarios have been proposed to describe different pattern of RCNA regions in tumor samples [18]. Scenario I demonstrates the case of only communal RCNAs. Scenario II displays that the regions of subgroup-specific RCNAs in samples of one subgroup may not be altered in the samples of another subgroup, while Scenario III shows that a region with amplifications in samples of one subgroup may altered with deletions in another subgroup samples. The RCNA regions in Scenario II can also be regarded as the combination of two or more groups of samples with RCNA regions in Scenario I, where the regions are different for these groups (details in Section 4 below and Supplementary Fig. S1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2018.2846599). To detect RCNAs under the three scenarios, Morganella et al. considers within-sample homogeneity into RCNA detection procedure [18]. Zhou et al. approximates the scenarios of RCNAs as low-rank matrices, and decompose a piecewise-constant and low-rank approximation component as RCNAs from the data matrix [24]. Despite the attempt made on finding RCNAs more than communal, calling subgroup-specific RCNAs and their related subgroups is still incapable for the approaches above.

In this article, we introduce HetRCNA, an integrated method for capturing statistically significant RCNAs and their related subgroups. By incorporating biclustering framework [9], [36], HetRCNA identifies subgroups and their related RCNAs simultaneously, and infers significant subgroup-specific RCNAs from each subgroup of the heterogeneous samples through statistical evaluations. We evaluate the performance of HetRCNA by simulated data with a variety of configurations [13], and comparing HetRCNA with two existing methods. A real dataset of breast cancer is then analyzed by HetRCNA, with distinct subgroup-specific RCNAs and their related subgroups found. The following sections of the paper are organized as below: Section 2 introduces three previous RCNA detection approaches with different techniques and their relations with HetRCNA. Section 3 describes our method HetRCNA with detailed techniques. In Section 4, HetRCNA is evaluated by simulated dataset and applied on a real cancer dataset. Finally, we discuss our article with conclusion and some prospects in Section 5. The main contributions of this paper are summarized as below:

- We take into account subgroup-specific RCNAs in RCNA discovery problem. The consideration of subgroup-specific RCNAs can help detect some CNAs recurrent in only a group of samples that might be missed by the methods for only communal RCNAs.
- We introduce an integrated method HetRCNA based on biclustering framework to detect subgroups of samples and their corresponding subgroup-specific RCNAs. HetRCNA is established by incorporating both three weight constraints (details in Section 3.1) on Sparse Singular Value Decomposition (SSVD) [36] and statistical significance test via weighted convolutions.
- We conduct experiments based on both simulated data and real data to evaluate HetRCNA over several existing methods. The validation results show that HetRCNA achieves an improvement over other existing methods averagely.

## 2 Related Works

In previous studies, the frequency of CNAs of a genomic region in all investigated samples is a crucial measurement for detecting communal RCNAs. One of the most widely used RCNA detection approaches is Genomic Identification of Significant Targets in Cancer (GISTIC), which considers both the average CNA amplitudes and the aberration frequencies in samples for the investigated regions of aberrations [16]. In GISTIC, the multiplications of average amplitudes and frequencies are utilized as the scores for the investigated RCNAs, which are called G-scores. Statistical test is then applied to obtain p-values of the investigated regions through permutation test, of which the null hypothesis is that all aberrations are sporadic passenger CNAs [16]. False-discovery rate (FDR) control is then used to correct the p-values and yields the corrected p-values, which are also called FDR q-values. Finally, the significant regions identified by FDR q-values are separated through a greedy peeling-off procedure [16]. GISTIC 2.0 is then proposed in [17] as a revised version of GISTIC. In GISTIC 2.0, the G-scores are redefined as the negative logarithm of the likelihood of each region, of which the value is larger when the related region is expected to be recurrent [17]. Copy number amplifications and deletions are measured separately in GISTIC 2.0. The statistical significance evaluation in GISTIC 2.0 is also based on permutation test and FDR control as in GISTIC. In contrast to GISTIC, The significant RCNAs in GISTIC 2.0 are separated through a revised arbitrated peeling-off strategy [17].

To detect RCNAs from tumor samples with heterogeneity, Genomic Analysis of Important Alterations (GAIA) extends the statistical hypothesis framework by taking within-sample homogeneity into account [18]. While the statistical significance procedure applied by the traditional studies [16], [18], [19] is used in GAIA, h-value is also introduced as another important score for measuring the within-sample homogeneity [18]. Since the h-values are based on the Hamming distance between the copy number alterations of each pairs of samples, the within-sample homogeneity considered by GAIA is an attempts to measure samples that are likely to belong to the same subgroups during the RCNA detection procedure. Combining the statistical FDR q-values from significance test and the h-values from within-sample homogeneity into an iterative procedure, GAIA can extract the RCNA regions shared by a set of samples with homogeneity. When compared with the approaches designed for common aberration regions, the detection result of GAIA shows better performance than those of the traditional approaches [18].

From a distinct perspective that the RCNA scenarios may form low-rank matrices, Piecewise-constant and Low-rank Approximation (PLA) utilizes a low-rank approximation strategy to detect RCNAs from a cohort of tumor samples [24]. To decompose the RCNA component, PLA approximates the matrix rank of the RCNA component through the nuclear norm, and formulate a convex optimization problem

that minimize the nuclear norm [24]. Smoothness penalty on each CNA regions is also used to ensure the piecewise-constant of the recovered profiles [24]. An L1-norm penalty is introduced to decompose a sparse component related to passenger CNAs that sparsely appear at different locations for different samples [24]. The RCNAs are then identified by thresholding on the recovered low-rank component. The idea of regarding RCNA regions as low-rank matrix is inspiring in RCNA modeling and detecting problem, and PLA has been proven to outperform the traditional methods that are designed for only common RCNAs [24].

In the RCNA detection task, GISTIC 2.0 and GAIA calculate the scores that represent both abberation amplitudes of the investigated CNA regions and frequencies in all samples. PLA obtatins the abberation amplitudes from a low-rank component which is decomposed from the data matrix of CNA amplitudes of samples. In comparison, HetRCNA computes the abberation amplitudes of CNA regions and frequencies in a group of samples. The abberation scores of GISTIC 2.0 are yielded by the summations of the negative logarithm of the likelihood of the abberation amplitudes of all samples; the abberation scores of CNA regions for GAIA are based on both abberation amplitude summations and homogeneity values; the abberation scores of regions for PLA are the frequencies of abberations contained in the low-rank component; the abberation scores of CNA regions for HetRCNA are obtained from the coefficients of the RCNA profile vectors decomposed from the data matrix. After calculating abberation scores, HetRCNA, GISTIC and GAIA use permutation test to yield the p-values of the investigated CNA regions, and utilize FDR procedure to compute FDR q-values of CNA regions. The significant RCNAs are then obtained by thresholding the q-values for HetRCNA, GISTIC 2.0 and GAIA, while PLA ranks the abberation scores of the regions and selects the top ranked regions as RCNAs. Despite the success achieved by the existing RCNA detection approaches discussed above, all the aforementioned previous approaches [16], [17], [18], [24] cannot find subgroup-specific RCNAs as well as their related subgroup samples when they are applied on heterogeneous tumor samples.

## 3 Proposed Approach

### 3.1 Sample Weights and RCNA Regions

The CNA data represent the amplitudes of CNA regions in the tumor samples, which are obtained from the CNA calling procedures on CNA raw data [37], [38], [39], [40], [41], [42], [43], i.e., the logarithm intensity ratio data collected from aCGH or SNP array platform, or the read depth data from the next generation sequencing technique. Each sample represents a patient of the investigated disease. After preprocessing, we align the amplitudes of CNA regions as the sample-CNA matrix $\boldsymbol{X} = (x_{ij})_{n\times p}$, of which the rows denote tumor samples and columns denote the amplitudes of aligned CNA regions.

For finding both RCNA regions and corresponding subgroup samples from the matrix, we considers biclustering framework to align rows to subgroups and columns to CNA regions simultaneously [44], [45]. Motivated by previous research [9], we use biclustering in this study because there is a strong connection between biclustering and finding common regions for different subgroups. The objective of locating sets of alterations which are common only to subsets of samples is similar to identifying sets of columns that show similar patterns across subsets of rows [9]. The property of biclustering framework can also help preserving the within-sample patterns of the samples [9]. The problem then is equivalent to finding a biclustering block from the sample-CNA matrix $\boldsymbol{X}$, which represents a subset of rows (samples) and a subset of columns (CNA regions) with a high similarity abberation scores [44], [45].

According to previous studies [24], [46], the RCNA components can be regarded as sparse components in the sample-CNA matrix. Consequently, we involve the biclustering framework of Sparse Singular Value Decomposition, primarily for analyzing high-dimensional gene expression data [36]. For the sample-CNA matrix $\boldsymbol{X}$, the $k$th subgroup-CNA block $\boldsymbol{X}^{(k)}$ are obtained from $k$th layer of HetRCNA, and the related rank-one layer can be obtained as the outer product of the left vector and the right vector. Here we define the left vector as sample weights vector $\hat{\boldsymbol{u}}^{(k)}$ and the right vector as the weighted average RCNA vector $\hat{\boldsymbol{v}}^{(k)}$ of the $k$th subgroup. The non-zero coefficients of $\hat{\boldsymbol{u}}^{(k)}$ represent the selected samples of the $k$th subgroup, while the non-zero coefficients of $\hat{\boldsymbol{v}}^{(k)}$ are either amplifications or deletions in the CNA regions of the subgroup. Particularly, three constraints are applied on the coefficient of $\hat{\boldsymbol{u}}^{(k)} = (\hat{u}_1^{(k)}, \ldots, \hat{u}_n^{(k)})$ to make them as a group of weights (weight constraints):

- *Non-negativity*: each coefficient of vector are non-negative, i.e., $\hat{u}_i^{(k)} \geq 0$, where $i = 1, \ldots, n$ denotes the sample indices and $k = 1, \ldots, K$ denotes the subgroup indices.
- *Normality*: due to the nonnegativity of the coefficients, the summation of every coefficients in $\hat{\boldsymbol{u}}^{(k)}$ equals one, formulating as $\sum_{i=1}^{n} \hat{u}_i^{(k)} = 1$.
- *Mutual-exclusivity*: one sample mustn't belong to two or more different subgroups. If the $i$th coefficient $\hat{u}_i^{(k)}$ in the $k$th subgroup block is larger than zero, then the $i$th coefficient $\hat{u}_i^{(t)}$ in the other subgroup blocks $t \neq k$ must equals zero strictly.

We introduce the three constraints aforementioned into the procedure of simultaneously detecting subgroups and subgroup-specific RCNAs. We incorporate the layer-by-layer procedure proposed by SSVD [36] for decomposing each biclustering block of subgroup and the related subgroup-specific RCNAs. As reported in previous studies [30], [34], if a tumor sample is assigned to a subgroup of the cancer, this sample cannot be assigned to the other subgroups. Consequently, any two subgroups cannot have overlapping samples, and the samples of subgroups are assumed to be mutual-exclusivity. To detect subgroups from the data matrix, we modify the biclustering framework of SSVD by the assumption of exclusive-rows biclusters into HetRCNA, which is used by Sheng et al. [47] and Tang et al. [48] but not by SSVD [36]. While the coefficients in the left vector of SSVD can be positive, negative or zero [36], the left vector for each block of HetRCNA is restricted to non-negative. When a layer is decomposed, the samples included in the subgroup of the current biclustering block are indicated by the positive coefficients of the vector, while the samples

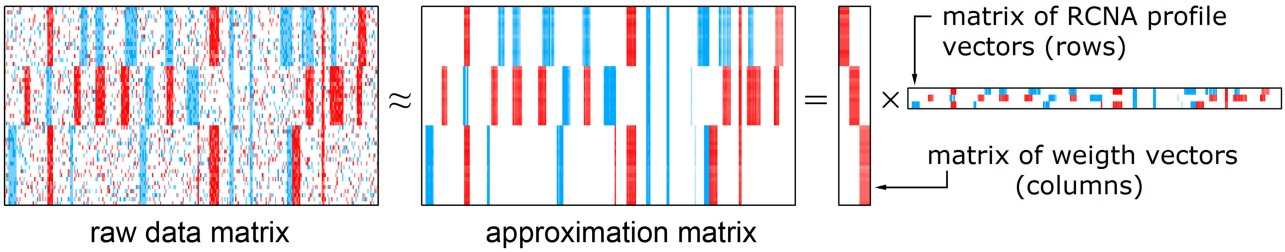



Fig. 1. An illustration of CNA matrix decomposition via HetRCNA, where red represents amplification, blue denotes deletion and white refers to copy number neutral. Left: Data matrix of aligned CNAs; middle: Approximation matrix of aligned CNAs yield from HetRCNA, with spontaneously random mutations removed; right: Subgroup sample matrix (consisting of sample weights vectors) and subgroup-specific RCNA matrix (consisting of RCNA regions vectors), of which the matrix multiplication equals to the approximation matrix (also the summation of the outer products of each pairs of column vectors and row vectors). The biclustering procedure of HetRCNA is based on matrix decomposition framework, and HetRCNA decomposes the data matrix through a layer-by-layer decomposing strategy [36]. There are usually several layers decomposed from the data matrix, and the approximation matrix is then obtained by the summation of these rank-one layers. Consequently, the rank of the approximation matrix is also usually larger than one, and equals to 3 in the case of the illustration.

not included in the subgroup are denoted by the zero coefficients of the vector. For the $k$th subgroup, the $i$th sample in set $\{i|\hat{u}_i^{(k)} = 0\}$ is not included in the current subgroup. Otherwise, the $i$th sample in set $\{i|\hat{u}_i^{(k)} > 0\}$ belongs to the subgroup and the absolute value $\hat{u}_i^{(k)}$ refers the contribution weight to the weighted average CNA profile. The normality of the left vector is also required to ensure that the vector is a sample weights vector. The weights are determined on factors as amplitude resolution or purity of samples. The assignment of all samples and all detected subgroups are indicated by the sample weights vectors of every layers obtained by HetRCNA.

The coefficient values have no constraint on the weighted average CNA profile vector $\hat{\boldsymbol{v}}^{(k)} = (\hat{v}_1^{(k)}, \ldots, \hat{v}_n^{(k)})$. Coefficients equaling to zero in $\hat{v}^{(k)}$ denote that the indicating CNA regions are copy-neutral, and positive and negative coefficients denote amplification and deletion in these CNA regions respectively at weighted average level.

### 3.2 Sample-RCNA Biclustering

The HetRCNA framework consists of two parts: 1) biclustering of sample weights and CNA profiles and 2) significance evaluation of RCNAs. Under the constraint that coefficients of $\hat{\boldsymbol{u}}$ satisfy the requirements of the weights vectors of sample-CNA matrix $\boldsymbol{X}$, we propose a biclustering method as the first part of HetRCNA, which efficiently measures subgroup samples in $\hat{\boldsymbol{u}}$ as series of weights and corresponding CNA profiles in $\hat{\boldsymbol{v}}$ as the weighted averages amplitudes. Inspired by SSVD [36], the biclustering framework can be regard the summation of different rank-one layers, and the $k$th layer is the block of its related samples (indicated by vector $\hat{\boldsymbol{u}}^{(k)}$) and related RCNAs (indicated by vector $\hat{\boldsymbol{v}}^{(k)}$). The summation of the different layers can be equivalent to matrix decomposition procedure, i.e.,

$$\boldsymbol{X} \simeq \sum_{k=1}^{K} \hat{\boldsymbol{u}}^{(k)} (\hat{s}^{(k)} \hat{\boldsymbol{v}}^{(k)})^{\mathrm{T}} = \boldsymbol{U}\hat{\boldsymbol{V}}^{\mathrm{T}}, \tag{1}$$

where the matrix $\boldsymbol{U} = [\hat{\boldsymbol{u}}^{(1)}, \ldots, \hat{\boldsymbol{u}}^{(k)}, \ldots, \hat{\boldsymbol{u}}^{(K)}]$ is a $n \times K$ matrix comprised of sample weights vectors and the matrix $\hat{\boldsymbol{V}} = [\hat{s}^{(1)}\hat{\boldsymbol{v}}^{(1)}, \ldots, \hat{s}^{(k)}\hat{\boldsymbol{v}}^{(k)}, \ldots, \hat{s}^{(K)}\hat{\boldsymbol{v}}^{(K)}]$ is a $p \times K$ matrix comprised of average CNA profile vectors, and the number $K$ is the rank of the two matrices. The positive scalars $\hat{s^{(k)}}$ for $k = 1, \ldots, K$ are the scales of the normalized vector $\hat{\boldsymbol{v}^{(k)}}$ to fit the approximation in (1), and these scalars are ordered from greatest to least. The rational of matrix decomposition based biclustering procedure of HetRCNA is demonstrated in Fig. 1. In HetRCNA, we decompose the matrix through a layer-by-layer decomposing strategy as proposed in [36]. Therefore, our method does not decompose only one rank-one layer to approximate the data matrix. Instead, our method decomposes several rank-one layers from the data, and use the summation of the layers to approximate the data matrix, and the layer number $K$ is obtained when no new layer can be decomposed from the data matrix. Note that the first rank-one layer of HetRCNA is the best rank-one matrix approximation of the samples-CNA matrix [36] under the weight constraints, i.e.,

$$\begin{aligned} (\hat{s}^{(1)}, \hat{\boldsymbol{u}}^{(1)}, \hat{\boldsymbol{v}}^{(1)}) &= \underset{\hat{s},\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}}{\operatorname{argmin}} \left|\left|\boldsymbol{X} - \hat{s}\hat{\boldsymbol{u}}\hat{\boldsymbol{v}}^{\mathrm{T}}\right|\right|_F^2 \\ \text{s.t. } \hat{\boldsymbol{u}} &\geq 0, \sum_{i=1}^{n} \hat{u}_i = 1, \end{aligned} \tag{2}$$

where $||\cdot||_F^2$ indicates the squared Frobenius norm. Factor $\hat{s}$ is a positive scalar, ensuring that each element of rank-one layer $\hat{s}\hat{\boldsymbol{u}}\hat{\boldsymbol{v}}^{\mathrm{T}}$ is the approximate estimation of the CNA amplitudes at the corresponding regions and samples. In each estimated rank-one layer, the corresponding samples share the same positive elements and negative elements, reflecting the same amplifications and deletions of subgroup-specific RCNAs respectively. If we used a layer of which the rank is larger than one, we cannot guarantee that the corresponding samples share same RCNA profiles. Also, the mutual-exclusivity of samples in the optimization ensures that the CNA amplitudes are approximated by the elements of each estimated layer, rather than by the weight summations of the elements of these layers. Consequently, the element of estimated rank one layer can well approximate the CNA amplitudes at the corresponding regions and samples.

Considering the sparsity of the RCNA component, we add L1-norm sparsity-inducing penalties [36] on $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ in the loss function of optimization problem of (2)

$$\begin{aligned} &\min_{\hat{s},\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}} \left|\left|\boldsymbol{X} - \hat{s}\hat{\boldsymbol{u}}\hat{\boldsymbol{v}}^{\mathrm{T}}\right|\right|_F^2 + \lambda_u ||\hat{\boldsymbol{u}}||_1 + \lambda_v ||\hat{\boldsymbol{v}}||_1 \\ &\text{s.t. } \hat{\boldsymbol{u}} \geq 0, \sum_{i=1}^{n} \hat{u}_i = 1, \end{aligned} \tag{3}$$

where $\lambda_u$ and $\lambda_v$ are penalty parameters. To solve $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ by lasso regression [36], [49], we replace triplet $(\hat{s}, \hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$ with $(s, \boldsymbol{u}, \boldsymbol{v})$ for calculating facility, restrained by $\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}} = 1$, $\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}} = 1$, and scaling equation $\hat{s}\hat{\boldsymbol{u}}\hat{\boldsymbol{v}}^{\mathrm{T}} = s\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$. The $(\boldsymbol{u}, \boldsymbol{v})$ in triplet $(s, \boldsymbol{u}, \boldsymbol{v})$ is the L2-normalized vectors of $(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$, i.e., $\boldsymbol{u} = \hat{\boldsymbol{u}}/||\hat{\boldsymbol{u}}||^2$ and $\boldsymbol{v} = \hat{\boldsymbol{v}}/||\hat{\boldsymbol{v}}||^2$, and $s = (||\hat{\boldsymbol{u}}||^2||\hat{\boldsymbol{v}}||^2)\hat{s}$ for balancing the scaling of two triplets. The inverse transformation of the triplets will be described at the end of this section.

For fixed $\hat{\boldsymbol{u}}$, the minimization of (3) in regard to $(s, \boldsymbol{v})$ is equivalent to minimization with respect to $\hat{\boldsymbol{v}} = (s\boldsymbol{v})$ of the formulation below:

$$\left|\left|\boldsymbol{X} - \boldsymbol{u}\hat{\boldsymbol{v}}^{\mathrm{T}}\right|\right|_F^2 + \lambda_v||\hat{\boldsymbol{v}}||_1 = \left|\left|\vec{\boldsymbol{X}} - (\mathbf{I}_p \otimes \boldsymbol{u})\hat{\boldsymbol{v}}\right|\right|^2 + \lambda_v||\hat{\boldsymbol{v}}||_1, \tag{4}$$

where $\vec{\boldsymbol{X}} \in \mathscr{R}^{np\times 1}$ is the vectorization of matrix $\boldsymbol{X}$ which results in a column vector $(\boldsymbol{x}_1^{\mathrm{T}}, \ldots, \boldsymbol{x}_p^{\mathrm{T}})^{\mathrm{T}}$ (here $\boldsymbol{x}_i$ is the $i$th column vector in matrix $\boldsymbol{X}$). The symbol $\otimes$ is the Kronecker product. On the other hand, for fixed $\boldsymbol{v}$ in (3), the optimization in regard to $\hat{\boldsymbol{u}} = (s\boldsymbol{u})$ is tantamount to the following equation:

$$\begin{aligned}\left|\left|\boldsymbol{X} - \hat{\boldsymbol{u}}\boldsymbol{v}^{\mathrm{T}}\right|\right|_F^2 + \lambda_u||\hat{\boldsymbol{u}}||_1 &= \left|\left|\vec{\boldsymbol{X}^{\mathrm{T}}} - (\mathbf{I}_n \otimes \boldsymbol{v})\hat{\boldsymbol{u}}\right|\right|^2 + \lambda_u \mathbf{1}^{\mathrm{T}}\hat{\boldsymbol{u}} \\ &\text{s.t. } \hat{\boldsymbol{u}} \geq 0.\end{aligned} \tag{5}$$

where $||\hat{\boldsymbol{u}}||_1$ in (3) is simplified as sum of coefficients in $\hat{\boldsymbol{u}}$ as a result of $\hat{\boldsymbol{u}} \geq 0$.

According to previous work [36], the right part of (4) is the lasso regression problem. The loss function is

$$\begin{aligned}&\left|\left|\vec{\boldsymbol{X}} - (\mathbf{I}_p \otimes \boldsymbol{u})\hat{\boldsymbol{v}}\right|\right|^2 + \lambda_v \sum_{j=1}^{p}|\hat{v}_j| \\ &\quad = \left|\left|\vec{\boldsymbol{X}}\right|\right|_F^2 + \sum_{j=1}^{p}(\hat{v}_j^2 - 2\hat{v}_j(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u})_j + \lambda_v|\hat{v}_j|),\end{aligned} \tag{6}$$

where $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u})_j$ is the $j$th coefficient of vector $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u}$, namely the ordinary least square (OLS) estimation of $\hat{\boldsymbol{v}}$ with no constraint in [36], which is $\{(\mathbf{I}_p \otimes \boldsymbol{u})^{\mathrm{T}}(\mathbf{I}_p \otimes \boldsymbol{u})\}^{-1}(\mathbf{I}_p \otimes \boldsymbol{u})^{\mathrm{T}}\vec{\boldsymbol{X}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{u}$ (given that fact that fixed $\boldsymbol{u}$ is L2-normalized as $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} = 1$). The coefficients in sparse vector $\hat{\boldsymbol{v}}$ of lasso regression in (6) are estimated by soft threshold estimator [36], [49]. The component-wise thresholding of (6) is

$$\tilde{v}_j^{(1)} = \mathrm{sign}((\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u})_j)\left(\left|(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u})_j\right| - \frac{\lambda_v}{2}\right)_+, \tag{7}$$

where $\mathrm{sign}(\cdot)$ represents the sign function that extracts the sign of the input number, and the operator $(\cdot)_+$ denotes $\max(\cdot, 0)$. The singular value is updated as $s_v^{(1)} = \left|\left|\hat{\boldsymbol{v}}^{(1)}\right|\right|^2$ and the L2-normalized vector $\boldsymbol{v}$ is calculated as $\boldsymbol{v}^{(1)} = \hat{\boldsymbol{v}}^{(1)}/s_v^{(1)}$.

For fixed $\hat{\boldsymbol{v}}$, the solution of the optimization of (5) is solved through a similar strategy [36]. The minimization of (5) is expanded to the formation above

$$\begin{aligned}&\left|\left|\vec{\boldsymbol{X}^{\mathrm{T}}} - (\mathbf{I}_n \otimes \boldsymbol{v})\hat{\boldsymbol{u}}\right|\right|^2 + \lambda_u \sum_{i=1}^{n}\hat{u}_i \\ &\quad = \left|\left|\vec{\boldsymbol{X}}\right|\right|_F^2 + \sum_{i=1}^{n}(\hat{u}_i^2 - \hat{u}_i\left[2(\boldsymbol{X}\boldsymbol{v})_i - \lambda_u\right]),\end{aligned} \tag{8}$$

where $(\boldsymbol{X}\boldsymbol{v})_i$ is the $i$th coefficient of OLS estimation for vector $\boldsymbol{u}$, which is $\{(\mathbf{I}_n \otimes \boldsymbol{v})^{\mathrm{T}}(\mathbf{I}_n \otimes \boldsymbol{v})\}^{-1}(\mathbf{I}_n \otimes \boldsymbol{v})^{\mathrm{T}}\vec{\boldsymbol{X}}^{\mathrm{T}} = \boldsymbol{X}\boldsymbol{v}$. The non-negative lasso is rather closely related to the non-negative least square (NNLS) problem [50]. To solve the NNLS and estimate the non-negative vector $\hat{\boldsymbol{u}}$, we utilize the component-wise thresholding rule proposed by [36] and NNLS by [50]. The solution is estimated by reshaped component-wise thresholding rule by adding the non-negative constraints of sparse vector $\hat{\boldsymbol{u}}$ [36], [50]

$$\tilde{u}_i^{(1)} = \left((\boldsymbol{X}\boldsymbol{v})_i - \frac{\lambda_u}{2}\right)_+, \tag{9}$$

which is the closed form expression of NNLS estimator for orthonormal design [50]. Similar with vector $\boldsymbol{v}$, $s_u^{(1)} = \left|\left|\hat{\boldsymbol{u}}^{(1)}\right|\right|^2$ is updated, with $\boldsymbol{u}^{(1)} = \hat{\boldsymbol{u}}^{(1)}/s_u^{(1)}$. Note that the calculation of $\boldsymbol{u}^{(1)}$ is depended on matrix decomposition framework and there is no constraint on the order of the samples. Therefore, the results of HetRCNA are not influenced by the order of the samples.

For objective function of (6) and (8), the optimal penalty parameters $\lambda_v$ and $\lambda_u$ are chosen by calculating the entire penalization path which can minimize the Bayesian information criterion [36]. Then the optimization procedures in (3) of two variables $\boldsymbol{u}$ and $\boldsymbol{v}$ are alternated iteratively until convergence. The convergence condition of the alternative procedure is that the euclidean distance between the latest estimated vector and the previous estimated vector is less than $10^{-4}$, for both left vector and right vector. Since the iteratively alternative procedure is deterministic, the outputs of the results only depend on the input matrix and the initial values of $\boldsymbol{u}$ and $\boldsymbol{v}$. In this study, we propose a procedure to make the initial values of the two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are determined by only the input matrix. We first decompose the input matrix by SVD and obtain the first left vector and the first right vector of SVD, of which the outer product is the first SVD layer of the data matrix. For the first left vector, if the number of negative coefficients of the vector is larger than the number of the positive coefficients of the vector, we then multiple both the first left vector and the first right vector by $-1$. This ensures that the two vectors are deterministic and their outer product is constant, and most coefficients of the left vector are positive. To satisfy the non-negativity of the sample weights vector, we replace the negative coefficient of the left vector with zeros. The two vectors obtained from the procedure above are then used as the initial values of $\boldsymbol{u}$ and $\boldsymbol{v}$. Since the initial values of the two vector $\boldsymbol{u}$ and $\boldsymbol{v}$ are deterministic when an input data matrix is given, the solution of HetRCNA is also unique.

After convergence, the estimation of the singular value of the current layer is $s = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{v}$. In this layer, we use the coefficients of $\boldsymbol{u}$ that are larger than half of the maximum value to obtain weight vector $\boldsymbol{u}^w$, and the samples of the remaining coefficients with insufficient contributions are reserved for the estimation of the next rank-one layer [46]. The inverse transformation of triplet $(s, \boldsymbol{u}, \boldsymbol{v})$ to $(\hat{s}, \hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$ is then described: we first normalize the weight coefficients to satisfy the normality of weight constraints, i.e., the summation $\hat{\boldsymbol{u}} = \boldsymbol{u}^w/\sum_{i=1}^{n}u_i^w$; as a result of $s = \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v}$ and $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = 1$, the weighted average vector is $\hat{\boldsymbol{v}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{u} = s\hat{\boldsymbol{v}}/\sum_{i=1}^{n}u_i^w$; finally, the positive scalar to balance the scaling of two triplets is calculated as $\hat{s} = \left(\sum_{i=1}^{n}u_i^w\right)^2$.

For each layer, we use a two-step method to satisfy the two constraints, i.e., nonnegative constraint and normality constraint. The first step in Formula (5) is used to obtain the relative values of vector $\hat{\boldsymbol{u}}$ through the alternative procedure, which ensure that the coefficients satisfy the nonnegative constraint when convergence. After removing coefficients with insufficient contributions and obtaining vector $\boldsymbol{u}^w$, we apply the second step to satisfy summation constraint via dividing the coefficients of the vector by their summation $\hat{\boldsymbol{u}} = \boldsymbol{u}^w / \sum_{i=1}^{n} u_i^w$, which make the summation of the coefficients of the vector a constant in this step. To assure the mutual-exclusivity of weight constraints on weights vector $\hat{\boldsymbol{u}}$ for each layer, the next sparse rank-one layer are obtained from the remaining rows (samples) of data matrix. We then applying the HetRCNA method to the matrix of remain samples iteratively until the number of rest samples is less than 10 percent of the total sample amount.

### 3.3 RCNA Significance Assessing

For the CNA region vectors decomposed by HetRCNA of the heterogeneous tumor samples, coefficient $\hat{v}_j$ in $\hat{\boldsymbol{v}}$ of each layer is the weighted estimation of the pinpointed CNA amplitudes across the subgroup samples. Since the scores of the coefficients in $\hat{\boldsymbol{v}}$ are contributed by either subgroup specific RCNAs shared by samples of the subgroup or randomly passenger CNAs, we need a strategy to draw a fine line between the two counter-acting signals after the decomposition procedure. Here we use significance test to test whether a CNA region indicated by a coefficient of the vector is significantly recurrent among the samples of the related subgroup or not. The CNA regions that significantly recurrent in the subgroup samples are then regarded as subgroup specific RCNAs, while the insignificant ones are treated as regions containing only passenger CNAs.

To identify the RCNA in CNA regions of each subgroup, we then assess the significance of each CNA region indicated by $\hat{v}_j$ under the null hypothesis that all aberrations in the subgroup sample are spontaneous passenger mutations. Note that the empirical distribution of CNAs under the null hypothesis can be estimated by using the convolution of the amplitude distributions of every subgroup samples [16]. In this study, we use the procedure of weighted convolution [46] of the amplitude distributions $h_i(x)$ for the $i$th sample to estimate the hull distribution $\hat{v}_j$ for every CNA region $j = 1, \ldots, p$, i.e.,

$$H(x) = \frac{1}{\hat{u}_1} h_1\left(\frac{x}{\hat{u}_1}\right) * \frac{1}{\hat{u}_2} h_2\left(\frac{x}{\hat{u}_2}\right) * \ldots \frac{1}{\hat{u}_M} h_M\left(\frac{x}{\hat{u}_M}\right), \qquad (10)$$

where the weight $\hat{u}_i$ is the $i$th coefficient of $\boldsymbol{u}$, and $M$ is the number of non-zero coefficients of the related weight vector, indicating the amount number of samples in the current subgroup. Subsequently, the p-values for statistical significance of the CNA regions can be yielded by the empirical distribution $H(x)$.

The statistical significance of coefficients $\hat{v}_j$ denoting for amplifications or deletions are tested separately. By measuring FDR procedure on p-values and obtaining FDR q-values of each result, we assess the significant RCNAs by thresholding their related FDR q-values at an empirical value 0.25 proposed by previous study [16]. We then select all significant coefficients from CNA region vector $\hat{\boldsymbol{v}}$ to generate the RCNA vector $\hat{\boldsymbol{v}}^{sig}$. Finally, the columns of subgroup-specific significant RCNAs of the heterogeneous tumor samples and the rows of corresponding subgroups are indicated by non-zero coefficients of vector $\hat{\boldsymbol{v}}^{sig}$ and $\hat{\boldsymbol{u}}$ respectively. The flowchart as an overview of HetRCNA is illustrated in Supplementary Fig. S2, available online. The pseudocode of HetRCNA is also demonstrated in Algorithm 1.

**Algorithm 1.** HetRCNA

**Input**: $\boldsymbol{X}_{n\times p}$ ($n = \#$(samples); $p = \#$(regions)): CNA data.

**Output**: $K$: estimated layer number; $\boldsymbol{S}_{K\times 1}$: scaling scalars; $\boldsymbol{V}_{p\times K}$: subgroup-specific CNA profile matrix; $\boldsymbol{U}_{n\times K}$: sample weights matrix; $\boldsymbol{Q}_{p\times K}$: q-value matrix; $\boldsymbol{V}^{sig}_{p\times K}$: significant subgroup-specific RCNA matrix.

% initialize $\boldsymbol{S}, \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{V}^{sig}$ and $\boldsymbol{Q}$ with empty matrices
$\boldsymbol{S} \leftarrow [\,]$; $\boldsymbol{U} \leftarrow [\,]$; $\boldsymbol{V} \leftarrow [\,]$; $\boldsymbol{V}^{sig} \leftarrow [\,]$; $\boldsymbol{Q} \leftarrow [\,]$;
ind_assign $\leftarrow \boldsymbol{0}_{n\times 1}$; % initialize assigned sample index
$k \leftarrow 0$; % initialize the number of layers $k$
**repeat** % start: decompose new layer
  $\boldsymbol{u} \leftarrow$ first_left_vector_SVD($\boldsymbol{X}$); % initialize left vector
  $\boldsymbol{v} \leftarrow$ first_right_vector_SVD($\boldsymbol{X}$); % initialize right vector
  **if** $\{\mathrm{sum}(\boldsymbol{I}[\boldsymbol{u} < 0]) > \mathrm{sum}(\boldsymbol{I}[\boldsymbol{u} > 0])\}$
    $\boldsymbol{u} \leftarrow (-\boldsymbol{u})$; $\boldsymbol{v} \leftarrow (-\boldsymbol{v})$;
    **end_if** % ensure most elements of left vector nonnegative
  $\boldsymbol{u}(\boldsymbol{I}[\boldsymbol{u} < 0]) \leftarrow 0$; % remove negative values in left vector
  **repeat**
    $\boldsymbol{u}_{\mathrm{prev}} \leftarrow \boldsymbol{u}$; $\boldsymbol{v}_{\mathrm{prev}} \leftarrow \boldsymbol{v}$;
    $\lambda_v \leftarrow \mathrm{argmin}\{\mathrm{BIC}(\lambda_v, \boldsymbol{u}_{\mathrm{prev}})\}$;
    % Symbol $\circ$ is the element-wise product of two vectors
    $\boldsymbol{v} \leftarrow \mathrm{sgn}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u}_{\mathrm{prev}}) \circ (\left|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u}_{\mathrm{prev}}\right| - \frac{\lambda_v}{2})_+$;
    $\boldsymbol{v} \leftarrow \boldsymbol{v}/\|\boldsymbol{v}\|_2^2$;
    $\lambda_u \leftarrow \mathrm{argmin}\{\mathrm{BIC}(\lambda_u, \boldsymbol{v})\}$;
    $\boldsymbol{u} \leftarrow (|\boldsymbol{X}\boldsymbol{v}| - \frac{\lambda_u}{2})_+$;
    $\boldsymbol{u} \leftarrow \boldsymbol{u}/\|\boldsymbol{u}\|_2^2$;
  **until** $\{\left\|\boldsymbol{u} - \boldsymbol{u}_{\mathrm{prev}}\right\|_2 < 10^{-4}$ && $\left\|\boldsymbol{v} - \boldsymbol{v}_{\mathrm{prev}}\right\|_2 < 10^{-4}\}$
  % convergence condition of the alternation procedure
  $s \leftarrow \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{v}$; % scaling scalar % end of new layer
  $\boldsymbol{u}^w \leftarrow \eta_H\{\boldsymbol{u}, \frac{1}{2}\max(\boldsymbol{u})\}$;
  $\hat{\boldsymbol{u}} \leftarrow \boldsymbol{u}^w/\|\boldsymbol{u}^w\|_1$; $\hat{\boldsymbol{v}} \leftarrow s\boldsymbol{v}/\|\boldsymbol{u}^w\|_1$; $\hat{s} \leftarrow (\|\boldsymbol{u}^w\|_1)^2$;
  ind_assign($\boldsymbol{I}[\hat{\boldsymbol{u}} > 0]) \leftarrow 1$;
  $\#(\mathrm{assign}) \leftarrow \mathrm{sum}(\mathrm{ind\_assign})$; % #(assigned sample)
  $\boldsymbol{q} \leftarrow \mathrm{SignifTestByWeightedConv}(\boldsymbol{X}, \hat{s}, \hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$;
  $\hat{\boldsymbol{v}}^{sig} \leftarrow \hat{\boldsymbol{v}}(\boldsymbol{I}[\boldsymbol{q} \le 0.25])$;
  $\boldsymbol{S} \leftarrow [\boldsymbol{S}, \hat{s}]$; $\boldsymbol{U} \leftarrow [\boldsymbol{U}, \hat{\boldsymbol{u}}]$; $\boldsymbol{V} \leftarrow [\boldsymbol{V}, \hat{\boldsymbol{v}}]$; %update new layer
  $\boldsymbol{V}^{sig} \leftarrow [\boldsymbol{V}^{sig}, \hat{\boldsymbol{v}}^{sig}]$; $\boldsymbol{Q} \leftarrow [\boldsymbol{Q}, \boldsymbol{q}]$; $k \leftarrow k + 1$;
  $\boldsymbol{X}(\boldsymbol{I}[\hat{\boldsymbol{u}} = 0], :) \leftarrow 0$; % mutual-exclusivity for next layer
**until** $\{|n - \#(\mathrm{assign})|/n \le 0.1\}$
% stop condition: #(rest samples)/#(total samples) $\le 10\%$
$K \leftarrow k$; % obtain the number of layers $K$
**return** $K$, $\boldsymbol{S}$, $\boldsymbol{V}$, $\boldsymbol{U}$, $\boldsymbol{Q}$ and $\boldsymbol{V}^{sig}$

Note: $\#(\cdot)$ is the number of the input set; $\boldsymbol{I}[\cdot]$ is the indicator function.

$\eta_H\{\boldsymbol{w}, \lambda\} := \boldsymbol{w} \circ \boldsymbol{I}[|\boldsymbol{w}| > \lambda]$ is the hard-thresholding.

The pseudocode of $\mathrm{SignifTestByWeightedConv}(\boldsymbol{X}, \hat{s}, \hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$ is provided in Supplementary Algorithm S1, available online.

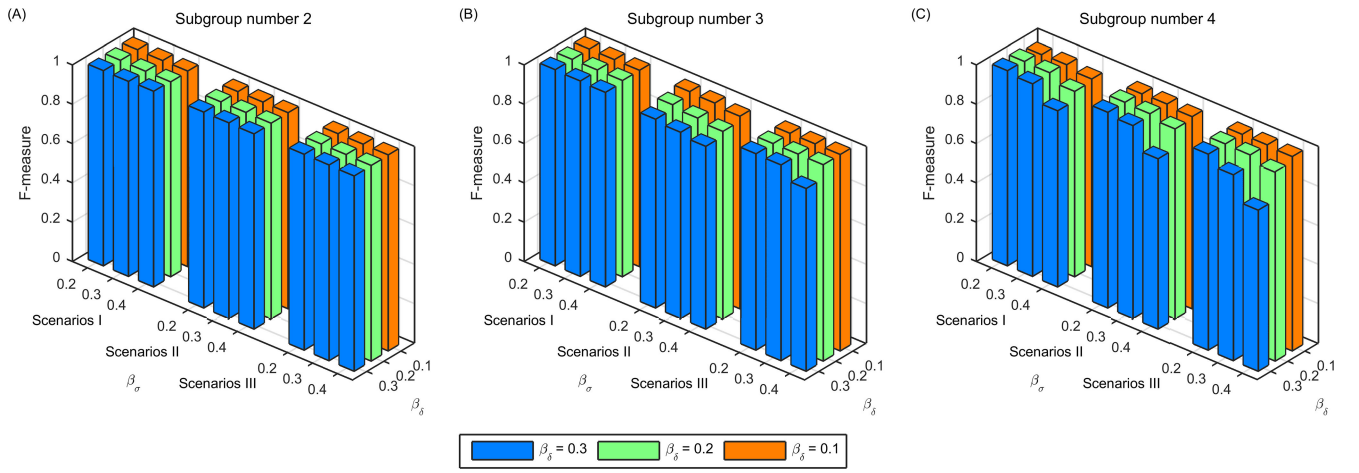



Fig. 2. F-measures of HetRCNA detection of three main scenarios under different subgroup numbers, with various background mutation rates and RCNA impurity rates. F-measures under (A) two subgroups, (B) three subgroups and (C) four subgroups are demonstrated to reflect the detection power, showing that the average level of detection powers decreases with the subgroup number increasing. In each diagram, the F-measures of Scenario II and III are more sensitive to noise than Scenario I due to complexity. The detection powers reduce by the noise increasing from either background mutation rate or impurity of RCNA region.

## 4 Experimental Result

### 4.1 Simulated Dataset Evaluation

To assess RCNA-finding algorithms quantitatively, Morganella et al. utilizes the three scenarios (Supplementary Fig. S1, available online) as synthetic data [18]. These three scenarios are considered as main fundamental scenarios because they are observed in real datasets and many other scenarios can be regarded as their combinations [18]. Scenario I consists of only communal RCNAs, sharing the same position among all samples. For Scenario II and III, the RCNA patterns show heterogeneity, which consist of subgroup-specific RCNAs from two subgroups (Supplementary Fig. S1, available online). In Scenario II, subgroup-specific RCNAs are non-overlapping at genome loci among different subgroup samples. Scenario III consists of different types of subgroup-specific RCNAs overlapping at genome loci, where amplifications and deletions from different subgroups occur in same regions. Since the subgroup numbers is usually more than two in real heterogeneous cancers [30], [31], [32], we configure the subgroup number as two, three and four when generating synthetic heterogeneous tumor data. For each case, we generated 100 samples with different subgroup numbers as heterogeneous tumor data. The RCNA patterns of each case are comprised by not only communal RCNAs of Scenario I but also subgroup-specific RCNAs of Scenario II and III, which are more similar to the real data of heterogeneous tumor samples.

Furthermore, we introduced two noise factors $\beta_\sigma$ and $\beta_\delta$ in each dataset by following previous work [13]. In non-RCNA regions, spontaneous passenger CNAs are added at a background mutation rate of $\beta_\sigma$, a parameter to quantify the level of spontaneous CNAs [10], [17]. In RCNA region, parameter $\beta_\delta$ measures the copy number neutral rate at RCNA regions. By varying the two parameters of the simulated data, we can use the detection results of HetRCNA to measure their influences on the detection performance. The amplitudes of every CNA regions are then calculated by adding the expectation amplitudes of the CNA regions and an a white Gaussian distribution [18].

To measure the statistical performance of HetRCNA, we involve two evaluation strategies: F-measure and Receiver Operation Characteristics (ROC) curve. F-measure is the harmonic mean of precision (the fraction of predictions that are underlying correct) and recall (the fraction of underlying true RCNAs that are predicted). Through F-measure, the detection performance can be measured by one numeric value, which help us comparing the detection performances under various configurations. ROC curve is a wildly-used comparison measurement for different computational methods, plotted by true-positive rate (TPR, also known as sensitivity) versus false-positive rate (FPR, also known as 1-specificity) with thresholding varying. In ROC curve, a better performance is indicated by the curve closer to the top and left borders. Since the power at small type I error levels is of interest, we investigate FPRs of the competing methods from 0 to 0.05, which is in consistent with previous studies [13], [14], [22].

#### 4.1.1 *Simulated Data Analysis with Various Configurations*

For synthetic heterogeneous tumor datasets with different subgroup numbers, the two noise factors $\beta_\sigma$ and $\beta_\delta$ are configured at various levels, increasing the difficulty of RCNA calling. Figs. 2A, 2B, and 2C shows the F-measures of HetRCNA of subgroup number 2 to 4 under different levels. Either $\beta_\sigma$ or $\beta_\delta$ levels increasing leads to F-measures decreasing, since more noise are contained in the datasets (Fig. 2 and Supplementary Table S1, available online). Scenario II and III are more sensitive to the two factors than Scenario I (the case of only communal RCNAs), which may be due to the heterogeneity of subgroup-specific RCNAs. For different subgroup numbers in heterogeneity datasets, the more subgroup number in the heterogeneous tumor data, the more sensitivity the detection results of HetRCNA to the two factors are observed. To observe the changes in F-measure more clearly, we also demonstrate the line and circle plots of the F-measures in Supplementary Fig. S3, available online. Generally, the performance of HetRCNA is influenced by factors of background mutation

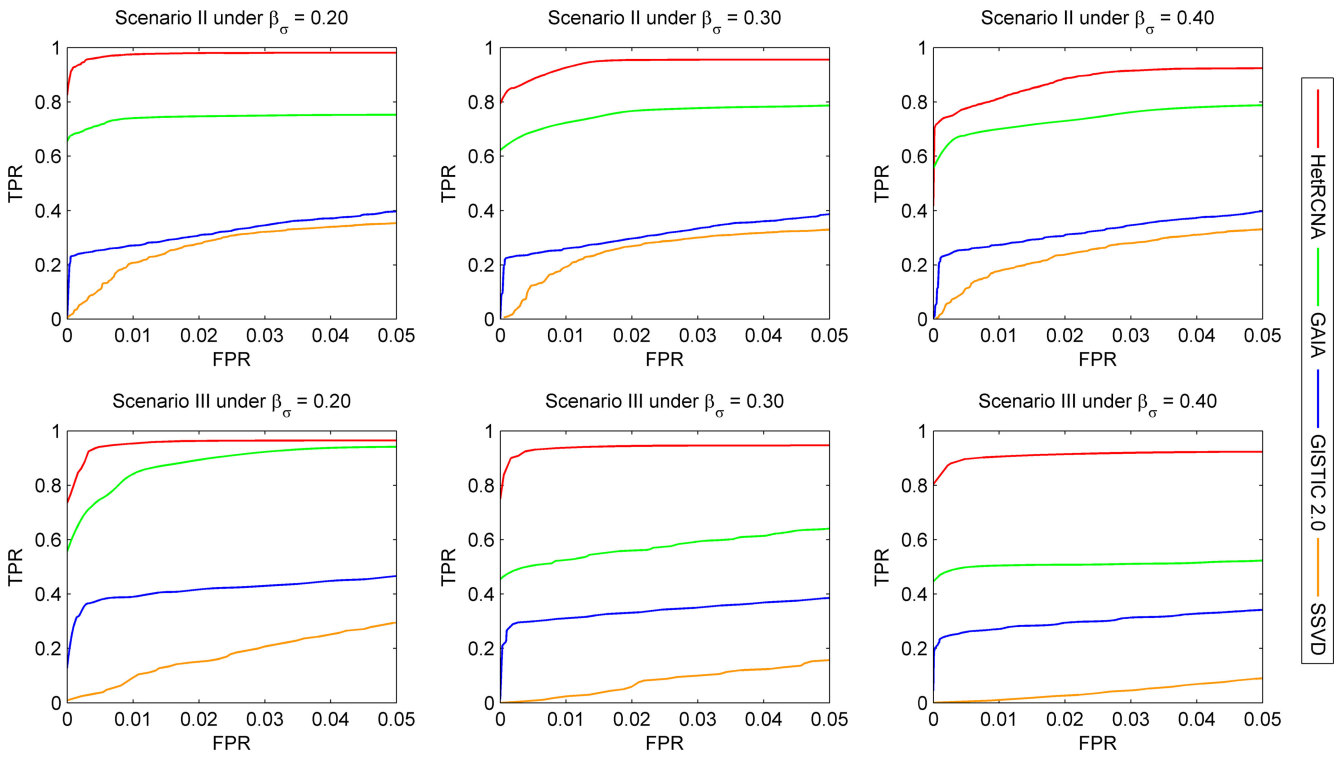



Fig. 3. ROC curves based on 100 replications of simulation. Since the power at small type I error levels is of interest [13], [14], [22], FPR is presented from 0 to 0.05. Configuration parameters are described as follows: background mutation rates vary at 0.2, 0.3 and 0.4 of two left panels, two middle panels and two right panels, respectively; Scenario II and III are showed at three top panels and three bottom panels; for all cases the copy number neutral rate at RCNA regions are 0.2. Red, green, blue and orange lines represent HetRCNA, GAIA, GISTIC 2.0 and SSVD, respectively.

rate, RCNA region neutral rate and heterogeneity of samples (subgroup number).

#### 4.1.2 Performance Comparison

Under a configuration of three subgroups with $\beta_\delta = 0.2$ and $\beta_\sigma \in \{0.2, 0.3, 0.4\}$, we compared the performance of HetRCNA against origin SSVD and two previous approaches, GAIA [18] and GISTIC 2.0 [17], through 100 replications of synthetic data. Since negative coefficients in left vector of SSVD can confound amplifications and deletions, we only consider the absolute amplitudes of alterations for ROC curves of SSVD. The ROC curves of the four methods under different configurations are illustrated in Fig. 3. For communal RCNAs in heterogeneous tumor data, all methods except SSVD achieve nearly perfect performances in the case of Scenario I. As for the cases of Scenario II that containing subgroup-specific RCNAs, HetRCNA and GAIA achieve the top two TPRs among the four methods, since HetRCNA considers tumor sample heterogeneity and GAIA takes into account within-sample homogeneity. The detection results of GISTIC 2.0 show low FPRs, which may be due to its powerful peel-off procedure. For Scenario III, HetRCNA maintains its performance best among the four methods. The results of all method show robustness when background mutation level $\beta_\sigma$ increases. For the data with larger noise factors, we can find that HetRCNA also achieves comparable or better performance (Supplementary Fig. S4, available online). Overall, the ROC curve evaluation indicates that HetRCNA has a clear advantage among the four competing methods.

When the scenario is complex, we introduce simulated data of Scenario IV in [24] (Supplementary Fig. S5A, available online). In Scenraio IV, there are three groups of samples with different patterns of RCNAs. The first group of samples contains recurrent amplification regions, where some of the regions are group-specific RCNAs of the the first group and the other regions are shared with the second group. The second group of samples contains regions of both recurrent amplifications and deletions, and the deletion regions are shared with the third group. The third group of samples includes only recurrent deletion regions, which are the same deletion regions of the second group. When HetRCNA is implemented on the data of Scenario IV, HetRCNA yields three subgroups as expected (Supplementary Fig. S5B, available online). When we compare the performances of HetRCNA with those of other existing methods through ROC curves, HetRCNA outperforms the other existing methods (Supplementary Fig. S5D, available online). Notably, the recurrent amplification regions in the first group of samples are highly scored by HetRCNA as subgroup-specific RCNAs, while these regions are not assigned with high scores by the other competing methods (Supplementary Fig. S5C, available online), indicating the group-specific advantage of HetRCNA.

### 4.2 Application on Breast Cancer

Breast cancer is one of the most common cancers threatening women's health worldwide [3]. Here we applied GISTIC 2.0, GAIA, and HetRCNA on 112 breast cancer samples using Illumina 109K SNP arrays from a previous study [51]. GPHMM [39] is used as CNA calling preprocessing, with CG coefficients (bias noise on SNP array platform) and impurity rectified. Comparison analysis and RCNA calling in heterogeneous tumor samples of HetRCNA are measured in details as below.

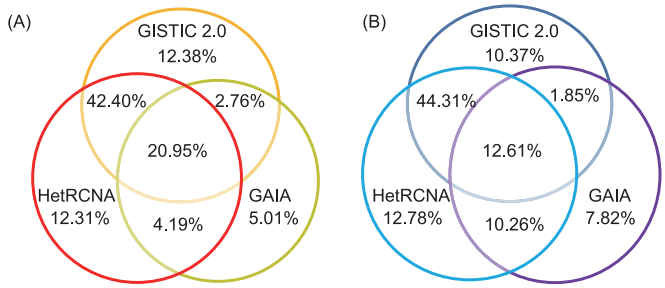



Fig. 4. Venn diagrams of intersections between (A) amplification and (B) deletion RCNA regions predicted by HetRCNA, GAIA and GISTIC 2.0 in breast cancer dataset. Percentages are according to the union of genomic regions predicted by the three methods.

#### 4.2.1 Comparison Analysis

The breast cancer data are implemented by HetRCNA, GISTIC 2.0 and GAIA with their default settings. Among the whole genome-wide regions, HetRCNA discovers 71 RCNA amplification regions and 70 RCNA deletion regions across all subgroups. Meanwhile, 68 amplification and 68 deletions are detected by GISTIC 2.0, while 137 amplification and 274 deletions, which are more fractional in size, are found by GAIA finds. Through comparison analysis, a fair amount of predicted regions are overlapping among the three approaches, while many other regions are unique to a single method, or pairs of methods. The proportions of prediction regions of the three methods across genome are showed as Venn diagrams (Fig. 4).

In our study, the proportions of common regions is 20.95 and 12.61 percent for amplifications and deletions respectively, showing the consistency among these RCNA finding approaches (Fig. 4). The values of the proportions are much larger than the proportions of common regions (less than 1 percent) in previous study [21], which are also based on three methods and also include GISTIC 2.0 and GAIA. Among the Venn diagrams, the detected RCNAs shared by HetRCNA and GISTIC 2.0 but not GAIA show the largest proportions for both amplifications (42.40 percent) and deletions (44.31 percent). GAIA shares more RCNA regions with HetRCNA with proportions of 4.19 and 10.26 percent comparing to GISTIC 2.0 of 2.76 and 1.85 percent, for amplifications and deletions respectively. The proportions of regions unique to GAIA are relative smaller than HetRCNA and GISTIC (5.01 percent in amplifications and 7.82 percent in deletions), which could be explained by the relatively conservative feature of GAIA [14].

Since the driver aberrations in these samples are unknown, we then align the detected RCNA regions of the three approaches with overlapping genes listed in [52]. For amplifications, the RCNA regions identified by at least two methods harbor a bunch of genes that are highly consistent with previous studies [3], [27], [28], [29], [53], [54], [55], [56], including ERBB2, CCND1, GRB7, PIK3CA, ARNT, C17orf37, FADD, FGF3, PPFIA1, RAD51C, PPM1D, TBX, NUPR1, AKT3, BIRC5, CRABP2, CSNK1D, GATA3 and JARID2 [3], [27], [29], [53], [55]. Further, some reported oncogenes are unique to HetRCNA: BIRC7, CCNE1, EEF1A2, FLRT3, JUN, LYN, and PRDM14 [3], [27], [29], [53], [55]. For example, CCNE1 gene is only detected by HetRCNA, which is capable of inducing chromosome instability by centrosome duplication and inappropriate initiation of DNA replication [57], [58]. Previous studies have demonstrated that CCNE1 is clinically associated with poor prognosis in patients with breast cancer [59] and the treatment of breast cancers [60], [61]. For another gene that unique to the detection results of HetRCNA, JUN (also known as AP-1) has been reported to play an important role in regulating breast cancer cell proliferation [62], [63], [64], which have also been reported as an oncogene for other types of cancers [65], [66].

For deletion, some genes highly associated to breast cancer [53], [54] are detected by at least two methods, e.g., NCOA7, ATM, CBFB, CDH1, CDH13, CDKN2A, CDKN2B, CTCF, DLC1, GATA4, HDAC2, MAP2K4, MYB, TP53 and TUSC3 [3], [27], [28], [29], [53]. Tumor suppressor genes unique to HetRCNA as AKT1, APOBEC3B, BRCA1, BRCA2, CHEK2, MAPT, NF1, NF2, PTEN, RB1, SMARCB1 and TIMP3 are detected in deletion RCNAs [3], [27], [29], [53], [55]. Taking PTEN as an example, it has been identified as a tumor suppressor gene in breast cancers [67]. To trastuzumab-based therapy, patients with PTEN-deficient breast cancers have been reported to show significantly poorer responses than those with normal PTEN [68], [69], indicating that PTEN activation in breast cancer contributes to trastuzumab's antitumor activity [70]. Genes APOBEC3B, BRCA1, BRCA2 and RB1 have been found in chromosome deletions in tumor samples of breast cancer and have been curated by Cancer Gene Census [71] as known cancer driver genes. For gene SMARCB1, it has been previously curated as a driver gene of malignant rhabdoid tumors [71]. In a recently study, aberrations on SMARCB1 have been found in Her2-positive breast cancers although it has not been previously observed in the results of TCGA on Her2-positive breast cancers [72]. Finally, a full list of reported genes identified by HetRCNA with more detailed information is provided in Supplementary Table S2, available online.

#### 4.2.2 Subgroups of Heterogeneous Tumor Samples

By HetRCNA, five distinct subgroups with corresponding RCNA regions are demarcated, with a remained sample subset including no significant RCNA (Figs. 5A and 5B). We denote the six subgroups as "Sub 1" to "Sub 5", and "Remain" respectively. The sample numbers for the subgroups are 19, 24, 32, 15 and 6, and the "Remain" subset contains 17 samples (Figs. 5A and 5B). According to the previous work of the investigated dataset [51], 17 samples among all 112 tumor samples are reported to be normal-like and have no acceptable solution (21 cases in primary version). Notably, compared with the other subsets, the copy numbers of tumor samples in "Remain" subset are closest to diploid (copy number=2) (Supplementary Fig. S6A, available online), implying that the 17 samples are highly correlated with normal-like state.

To show the difference between the subgroups, we involve contrast ploidy boxplots, copy number distributions and genome-wide frequencies of subgroup samples. The sample ploidy of HetRCNA subgroups are demonstrated in boxplots (Supplementary Fig. S6A, available online), illustrating the distinction of the subgroup ploidies. The plots of copy number distributions in HetRCNA subgroups fitted by Gaussian mixture model (Supplementary Fig. S6B, available online) indicate different copy number proportions

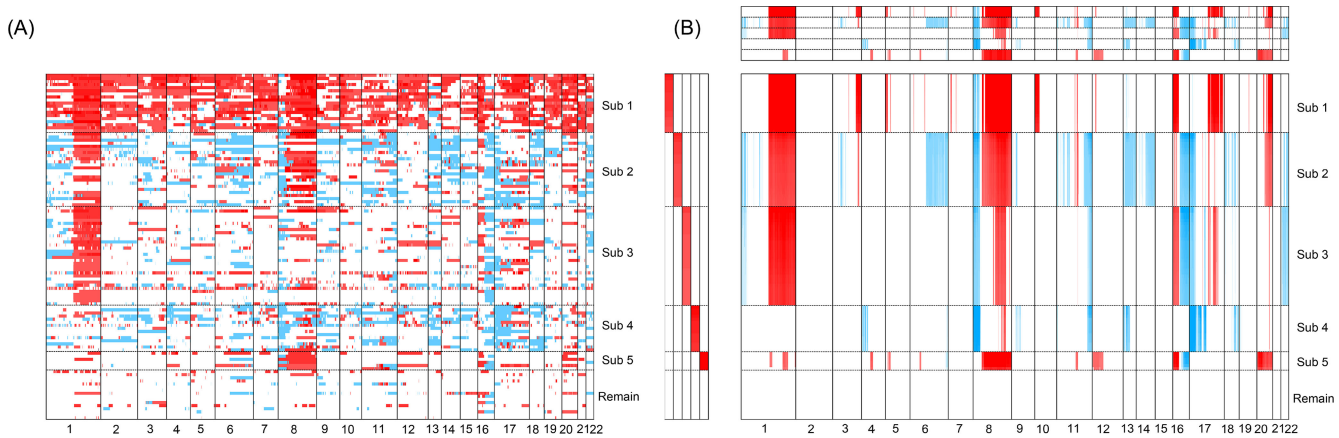



Fig. 5. (A) Sample-CNA data matrix of breast cancer dataset, of which rows represent samples (realigned by HetRCNA subgroups for illustration) and columns refer to CNA regions among genome. The red and blue colors indicate amplification and deletion regions. (B) Approximate sample-CNA data matrix (main part) obtained by matrix multiplication of the matrix comprised of sample weights vectors (left) $U$ and the transposed of matrix comprised of average CNA profile vectors (top) $\hat{V}^{\mathrm{T}}$.

among the subgroups. The distinction of genome-wide frequencies of amplifications and deletions among subgroup populations (similar to average CNA profile but only considering aberration frequency rather than weight and amplitude) are plotted in Supplementary Fig. S7A-S7E, available online.

As shown in Fig. 6 and Supplementary Fig. S8A-S8F, available online, the FDR q-values of CNA regions calculated by significant evaluation of HetRCNA among each subgroup demonstrate the significance of numerous RCNA regions, which harbor many breast cancer associated oncogenes and tumor suppressor genes. ERBB2 gene amplified is labeled in both "Sub 2" and "Sub 3"; CCND1 amplified in "Sub 1", "Sub 4" and "Sub 5"; BRCA2 deleted in "Sub 2" and "Sub 4"; TP53 deleted in "Sub 2", "Sub 3" and "Sub 4"; GATA4 deleted in "Sub 2", "Sub 3" and "Sub 4" (Figs. 6A and 6B; Supplementary Fig. S8A-S8F, available online; Supplementary Table S3, available online).

## 5 Discussion and Conclusion

The detection of RCNAs strongly associated with cancer tumors has thrown light on the research of oncology. When no prior information of cancer subgroup is available, detecting subgroup-specific RCNAs is still a challenging task for the existing RCNA detecting approaches. In this article, we introduce a method called HetRCNA to address subgroups and subgroup-specific RCNA finding problem for heterogeneous tumor samples, which is based on biclustering framework with weight constraints and RCNA significance assessment. In simulated dataset studies, HetRCNA shows high performance in separating both communal and subgroup-specific RCNAs from spontaneous passenger abberations. When compared with two existing methods, HetRCNA show better detection power when these methods are evaluated by ROC curves. When we apply HetRCNA on a real breast cancer data, a brunch of genes is unique to the results of HetRCNA rather than the results of the other two methods, including

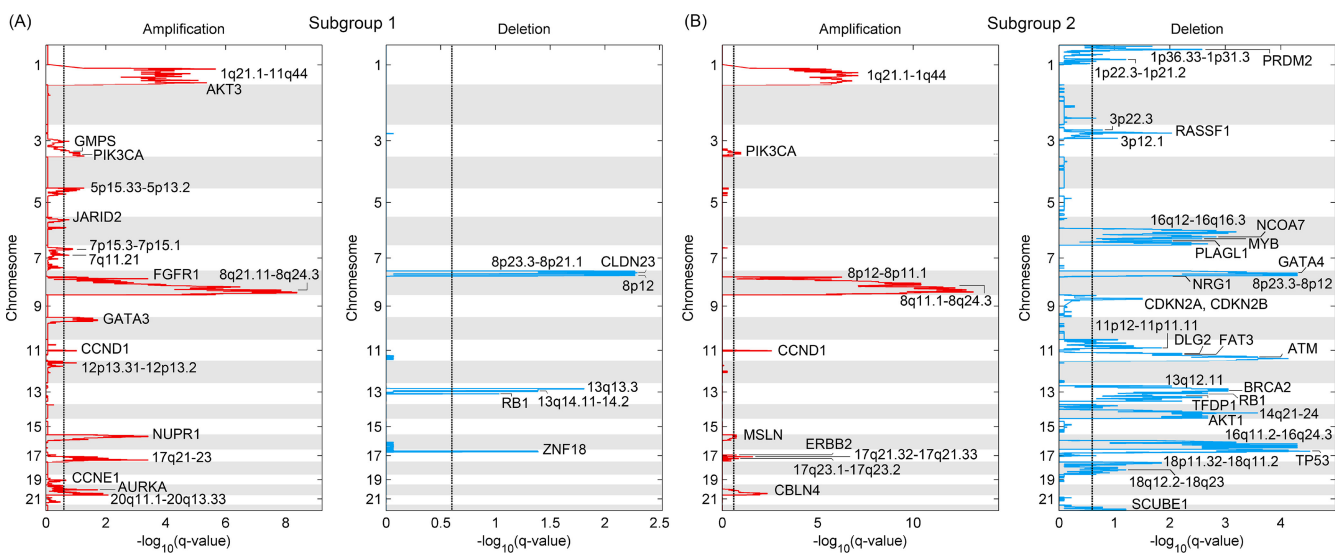



Fig. 6. Genome-wide statistical significance of subgroup-specific RCNAs detected by HetRCNA for breast cancer dataset. The subgroups are demonstrated: (A) "Sub 1"; (B) "Sub 2". The "Sub 3", "Sub 4" and "Sub 5" are shown in Supplementary Fig. S8A-S8B, S8C-S8D, S8E-S8F respectively, available online. In each plot, statistical significance for amplification (red) and deletion (blue) are evaluated via FDR corrected q-values. Dash lines refer to the significant threshold of 0.25.

some breast cancer associated genes. Finally, we assessed the subgroups and their related RCNA regions and genes harbored, demonstrating that subgroups and related RCNAs detected by HetRCNA show high consistency with previous studies.

There are mainly three differences between HetRCNA and SSVD [36]. The first difference is that the non-negative constraint is applied on the left vectors of HetRCNA to represent sample weights, while the constraint is not in SSVD. The second difference is that the left vector are mutual-exclusive to each other, while there can be overlaps among the left vectors in SSVD. The third difference is that the right vectors of HetRCNA of the subgroups are further processed by statistical significance test based on weighed convolutions to detect significantly subgroup-specific RCNAs, while this procedure is not included in SSVD.

In HetRCNA, the two tuning parameters $\lambda_u$ and $\lambda_v$ are chosen by Bayesian information criterion (BIC) in each iteration. To study the robustness of the two parameters, we disturb the two chosen parameters for each iteration. When the parameters are increased/decreased by 20 percent, the values of the output vectors only changes by 0.5 percent averagely. Therefore, the results of HetRCNA is robust to the two tuning parameters.

Despite the advantages above, a limitation of HetRCNA is that it might detect some spurious peaks at RCNA closely located regions, due to correlations between the recurrence scores. This challenge that both identifying independent peaks within a region and discarding spurious peaks, has been argued by previous studies [16], [17], [18], [19], [20], [21] with different peel-off algorithms. However, in HetRCNA, the RCNA region vectors are sparsified that the correlation of closely located RCNAs are broken by adding zeros into continuous alternation regions. Future investigation on designing peel-off algorithm to sparsified subgroup-specific RCNAs is warrant.

In conclusion, we present HetRCNA, a biclustering based algorithm to simultaneously detect cancer subgroups and their related RCNAs. Not only used for cancer subgroup-RCNA finding, HetRCNA can be generalized for many other group-feature labeling problems of any matrix-type data. We expect it will be utile for heterogeneous cancer subgroup labeling and subgroup-specific genome interpretation, and show its potential assistance on developing individualized therapy of heterogeneous cancers in clinical diagnosis and treatment.

## Acknowledgments

The authors would like to thank Yuanning Liu for helpful suggestions. This work is supported by the National Natural Science Foundation of China (Grant No. 61571414, No. 61471331 and No. 31100955).

## References

[1] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nature Rev. Genetics*, vol. 7, no. 2, pp. 85–97, 2006.

[2] D. G. Albertson, C. Collins, F. McCormick, and J. W. Gray, "Chromosome aberrations in solid tumors," *Nature Genetics*, vol. 34, no. 4, 2003, Art. no. 369.

[3] Cancer Genome Atlas Network and others, "Comprehensive molecular portraits of human breast tumors," *Nature*, vol. 490, no. 7418, 2012, Art. no. 61.

[4] A. Shlien and D. Malkin, "Copy number variations and cancer," *Genome Med.*, vol. 1, no. 6, 2009, Art. no. 62.

[5] B. S. Taylor, J. Barretina, N. D. Socci, P. DeCarolis, M. Ladanyi, M. Meyerson, S. Singer, and C. Sander, "Functional copy-number alterations in cancer," *PloS One*, vol. 3, no. 9, 2008, Art. no. e3179.

[6] T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C.-Z. Zhang, J. Wala, C. H. Mermel, et al., "Pan-cancer patterns of somatic copy number alteration," *Nature Genetics*, vol. 45, no. 10, pp. 1134–1140, 2013.

[7] R. Beroukhim, C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. S. Boehm, J. Dobson, M. Urashima, et al., "The landscape of somatic copy-number alteration across human cancers," *Nature*, vol. 463, no. 7283, pp. 899–905, 2010.

[8] B. J. Raphael, J. R. Dobson, L. Oesper, and F. Vandin, "Identifying driver mutations in sequenced cancer genomes: Computational approaches to enable precision medicine," *Genome Med.*, vol. 6, no. 1, 2014, Art. no. 5.

[9] O. M. Rueda and R. Diaz-Uriarte, "Finding recurrent copy number alteration regions: A review of methods," *Current Bioinf.*, vol. 5, no. 1, pp. 1–17, 2010.

[10] L. Ding, M. C. Wendl, J. F. McMichael, and B. J. Raphael, "Expanding the computational toolbox for mining cancer genomes," *Nature Rev. Genetics*, vol. 15, no. 8, pp. 556–570, 2014.

[11] A. J. Aguirre, C. Brennan, G. Bailey, R. Sinha, B. Feng, C. Leo, Y. Zhang, J. Zhang, J. D. Gans, N. Bardeesy, et al., "High-resolution characterization of the pancreatic adenocarcinoma genome," *Proc. Nat. Academy Sci. United States America*, vol. 101, no. 24, pp. 9067–9072, 2004.

[12] C. Rouveirol, N. Stransky, P. Hupé, P. L. Rosa, E. Viara, E. Barillot, and F. Radvanyi, "Computation of recurrent minimal genomic alterations from array-CGH data," *Bioinf.*, vol. 22, no. 7, pp. 849–856, 2006.

[13] X. Yuan, J. Zhang, S. Zhang, G. Yu, and Y. Wang, "Comparative analysis of methods for identifying recurrent copy number alterations in cancer," *PloS One*, vol. 7, no. 12, 2012, Art. no. e52516.

[14] X. Yuan, G. Yu, X. Hou, I.-M. Shih, R. Clarke, J. Zhang, E. P. Hoffman, R. R. Wang, Z. Zhang, and Y. Wang, "Genome-wide identification of significant aberrations in cancer genome," *BMC Genomics*, vol. 13, no. 1, 2012, Art. no. 342.

[15] S. J. Diskin, T. Eck, J. Greshock, Y. P. Mosse, T. Naylor, C. J. Stoeckert, B. L. Weber, J. M. Maris, and G. R. Grant, "STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments," *Genome Res.*, vol. 16, no. 9, pp. 1149–1158, 2006.

[16] R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, I. Vivanco, J. C. Lee, J. H. Huang, S. Alexander, et al., "Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma," *Proc. Nat. Academy Sci. United States America*, vol. 104, no. 50, pp. 20 007–20 012, 2007.

[17] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz, "GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers," *Genome Biol.*, vol. 12, no. 4, 2011, Art. no. R41.

[18] S. Morganella, S. M. Pagnotta, and M. Ceccarelli, "Finding recurrent copy number alterations preserving within-sample homogeneity," *Bioinf.*, vol. 27, no. 21, pp. 2949–2956, 2011.

[19] F. Sanchez-Garcia , U. D. Akavia, E. Mozes, and D. Pe'er, "JISTIC: Identification of significant targets in cancer," *BMC Bioinf.*, vol. 11, no. 1, 2010, Art. no. 189.

[20] V. Walter, A. B. Nobel, and F. A. Wright, "DiNAMIC: A method to identify recurrent DNA copy number aberrations in tumors," *Bioinf.*, vol. 27, no. 5, pp. 678–685, 2010.

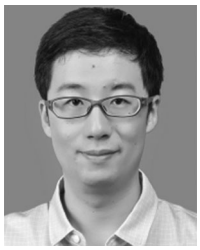
[21] H.-T. Wu, I. Hajirasouliha, and B. J. Raphael, "Detecting independent and recurrent copy number aberrations using interval graphs," *Bioinf.*, vol. 30, no. 12, pp. i195–i203, 2014.

[22] Q. Zhang, L. Ding, D. E. Larson, D. C. Koboldt, M. D. McLellan, K. Chen, X. Shi, A. Kraja, E. R. Mardis, R. K. Wilson, et al., "CMDS: A population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data," *Bioinf.*, vol. 26, no. 4, pp. 464–469, 2010.

[23] R. Pique-Regi, A. Ortega, and S. Asgharzadeh, "Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA," *Bioinf.*, vol. 25, no. 10, pp. 1223–1230, 2009.

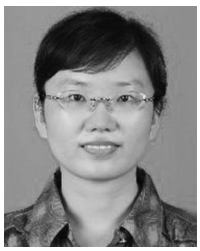
[24] X. Zhou, J. Liu, X. Wan, and W. Yu, "Piecewise-constant and low-rank approximation for identification of recurrent copy number variations," *Bioinf.*, vol. 30, no. 14, pp. 1943–1949, 2014.
[25] C. Klijn, H. Holstege, J. de Ridder, X. Liu, M. Reinders, J. Jonkers, and L. Wessels, "Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data," *Nucleic Acids Res.*, vol. 36, no. 2, pp. e13–e13, 2008.
[26] E. van Dyk, M. J. Reinders, and L. F. Wessels, "A scale-space method for detecting recurrent DNA copy number changes with analytical false discovery rate control," *Nucleic Acids Res.*, vol. 41, no. 9, pp. e100–e100, 2013.
[27] J. Hicks, A. Krasnitz, B. Lakshmi, N. E. Navin, M. Riggs, E. Leibu, D. Esposito, J. Alexander, J. Troge, V. Grubor, et al., "Novel patterns of genome rearrangement and their association with survival in breast cancer," *Genome Res.*, vol. 16, no. 12, pp. 1465–1479, 2006.
[28] J. Staaf, G. Jönsson, M. Ringnér, B. Baldetorp, and Å. Borg, "Landscape of somatic allelic imbalances and copy number alterations in HER2-amplified breast cancer," *Breast Cancer Res.*, vol. 13, no. 6, 2011, Art. no. R129.
[29] S.-H. Jung, A. Lee, S.-H. Yim, H.-J. Hu, C. Choe, and Y.-J. Chung, "Simultaneous copy number gains of NUPR1 and ERBB2 predicting poor prognosis in early-stage breast cancer," *BMC Cancer*, vol. 12, no. 1, 2012, Art. no. 382.
[30] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, et al., "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, pp. 346–352, 2012.
[31] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinf.*, vol. 25, no. 22, pp. 2906–2912, 2009.
[32] G. Wong, C. Leckie, and A. Kowalczyk, "FSR: Feature set reduction for scalable and accurate multi-class cancer subtype classification based on copy number," *Bioinf.*, vol. 28, no. 2, pp. 151–159, 2011.
[33] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes," *Sci.*, vol. 339, no. 6127, pp. 1546–1558, 2013.
[34] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, et al., Cancer Genome Atlas Research Network, "The cancer genome atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.
[35] M. D. Leiserson, F. Vandin, H.-T. Wu, J. R. Dobson, J. V. Eldridge, J. L. Thomas, A. Papoutsaki, Y. Kim, B. Niu, M. McLellan, et al., "Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes," *Nature Genetics*, vol. 47, no. 2, pp. 106–114, 2015.
[36] M. Lee, H. Shen, J. Z. Huang, and J. Marron, "Biclustering via sparse singular value decomposition," *Biometrics*, vol. 66, no. 4, pp. 1087–1095, 2010.
[37] D. Y. Chiang, G. Getz, D. B. Jaffe, M. J. O'Kelly, X. Zhao, S. L. Carter, C. Russ, C. Nusbaum, M. Meyerson, and E. S. Lander, "High-resolution mapping of copy-number alterations with massively parallel sequencing," *Nature Methods*, vol. 6, no. 1, pp. 99–103, 2009.
[38] P. Hupé, N. Stransky, J.-P. Thiery, F. Radvanyi, and E. Barillot, "Analysis of array CGH data: From signal ratio to gain and loss of DNA regions," *Bioinf.*, vol. 20, no. 18, pp. 3413–3422, 2004.
[39] A. Li, Z. Liu, K. Lezon-Geyda, S. Sarkar, D. Lannin, V. Schulz, I. Krop, E. Winer, L. Harris, and D. Tuck, "GPHMM: An integrated hidden Markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome SNP arrays," *Nucleic Acids Res.*, vol. 39, 2011, Art. no. gkr014.
[40] S. Morganella, L. Cerulo, G. Viglietto, and M. Ceccarelli, "VEGA: Variational segmentation for copy number detection," *Bioinf.*, vol. 26, no. 24, pp. 3020–3027, 2010.
[41] A. B. Olshen, E. Venkatraman, R. Lucito, and M. Wigler, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics*, vol. 5, no. 4, pp. 557–572, 2004.
[42] R. Xi, J. Luquette, A. Hadjipanayis, T.-M. Kim, and P. J. Park, "BIC-seq: A fast algorithm for detection of copy number alterations based on high-throughput sequencing data," *Genome Biol.*, vol. 11, no. S1, 2010, Art. no. O10.
[43] Z. Yu, Y. Liu, Y. Shen, M. Wang, and A. Li, "CLImAT: Accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data," *Bioinf.*, vol. 30, no. 18, pp. 2576–2583, 2014.
[44] Y. Cheng and G. M. Church, "Biclustering of expression data," in *Proc. 8th Int. Conf. Intell. Syst. Mol. Biol.*, 2000, pp. 93–103.
[45] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 1, no. 1, pp. 24–45, Jan.–Mar. 2004.
[46] J. Xi and A. Li, "Discovering recurrent copy number aberrations in complex patterns via non-negative sparse singular value decomposition," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 4, pp. 656–668, Jul./Aug. 2016.
[47] Q. Sheng, Y. Moreau, and B. De Moor, "Biclustering microarray data by Gibbs sampling," *Bioinf.*, vol. 19, no. suppl_2, pp. ii196–ii205, 2003.
[48] C. Tang, L. Zhang, A. Zhang, and M. Ramanathan, "Interrelated two-way clustering: An unsupervised approach for gene expression data analysis," in *Proc. IEEE 2nd Int. Symp. Bioinf. Bioeng. Conf.*, 2001, pp. 41–48.
[49] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statistical Soc. Series B (Methodological)*, vol. 58, pp. 267–288, 1996.
[50] M. Slawski and M. Hein, "Sparse recovery by thresholded non-negative least squares," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 1926–1934.
[51] P. Van Loo, S. H. Nordgard, O. C. Lingjærde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, et al., "Allele-specific copy number analysis of tumors," *Proc. Nat. Academy Sci. United States America*, vol. 107, no. 39, pp. 16 910–16 915, 2010.
[52] T. H. Hwang, G. Atluri, R. Kuang, V. Kumar, T. Starr, K. A. Silverstein, P. M. Haverty, Z. Zhang, and J. Liu, "Large-scale integrative network-based analysis identifies common pathways disrupted by copy number alterations across cancers," *BMC Genomics*, vol. 14, no. 1, 2013, Art. no. 440.
[53] P. M. Haverty, J. Fridlyand, L. Li, G. Getz, R. Beroukhim, S. Lohr, T. D. Wu, G. Cavet, Z. Zhang, and J. Chant, "High-resolution genomic and expression analyses of copy number alterations in breast tumors," *Genes Chromosomes Cancer*, vol. 47, no. 6, pp. 530–542, 2008.
[54] X. Hu, H. M. Stern, L. Ge, C. O'Brien, L. Haydu, C. D. Honchell, P. M. Haverty, B. A. Peters, T. D. Wu, L. C. Amler, et al., "Genetic alterations and oncogenic pathways associated with breast cancer subtypes," *Mol. Cancer Res.*, vol. 7, no. 4, pp. 511–522, 2009.
[55] R. Murria, S. Palanca, I. de Juan, C. Egoavil, C. Alenda, Z. García-Casado, M. J. Juan, A. B. Sánchez, A. Santaballa, I. Chirivella, et al., "Methylation of tumor suppressor genes is related with copy number aberrations in breast cancer," *Amer. J. Cancer Res.*, vol. 5, no. 1, 2015, Art. no. 375.
[56] J. Staaf, G. Jönsson, M. Ringnér, J. Vallon-Christersson, D. Grabau, A. Arason, H. Gunnarsson, B. A. Agnarsson, P.-O. Malmström, O. T. Johannsson, et al., "High-resolution genomic and expression analyses of copy number alterations in HER2-amplified breast cancer," *Breast Cancer Res.*, vol. 12, no. 3, 2010, Art. no. R25.
[57] S. Ekholm-Reed, J. Méndez, D. Tedesco, A. Zetterberg, B. Stillman, and S. I. Reed, "Deregulation of cyclin E in human cells interferes with prereplication complex assembly," *J. Cell Biol.*, vol. 165, no. 6, pp. 789–800, 2004.
[58] K. Kawamura, H. Izumi, Z. Ma, R. Ikeda, M. Moriyama, T. Tanaka, T. Nojima, L. S. Levin, K. Fujikawa-Yamamoto, K. Suzuki, et al., "Induction of centrosome amplification and chromosome instability in human bladder cancer cells by p53 mutation and cyclin E overexpression," *Cancer Res.*, vol. 64, no. 14, pp. 4800–4809, 2004.
[59] K. Keyomarsi, S. L. Tucker, T. A. Buchholz, M. Callister, Y. Ding, G. N. Hortobagyi, I. Bedrosian, C. Knickerbocker, W. Toyofuku, M. Lowe, et al., "Cyclin E and survival in patients with breast cancer," *New England J. Med.*, vol. 347, no. 20, pp. 1566–1575, 2002.
[60] D. Etemadmoghadam, B. A. Weir, G. Au-Yeung, K. Alsop, G. Mitchell, J. George, S. Davis, A. D. DAndrea, K. Simpson, W. C. Hahn, et al., "Synthetic lethality between CCNE1 amplification and loss of BRCA1," *Proc. Nat. Academy Sci. United States America*, vol. 110, no. 48, pp. 19 489–19 494, 2013.
[61] E. Kuhn, A. Bahadirli-Talbott, and I.-M. Shih, "Frequent CCNE1 amplification in endometrial intraepithelial carcinoma and uterine serous carcinoma," *Modern Pathology*, vol. 27, no. 7, 2014, Art. no. 1014.

[62] Y. Liu, C. Lu, Q. Shen, D. Munoz-Medellin, H. Kim, and P. H. Brown, "AP-1 blockade in breast cancer cells causes cell cycle arrest by suppressing G1 cyclin expression and reducing cyclin-dependent kinase activity," *Oncogene*, vol. 23, no. 50, 2004, Art. no. 8238.
[63] C. Lu, Q. Shen, E. DuPre, H. Kim, S. Hilsenbeck, and P. H. Brown, "cFos is critical for MCF-7 breast cancer cell growth," *Oncogene*, vol. 24, no. 43, 2005, Art. no. 6516.
[64] K. Dahlman-Wright, Y. Qiao, P. Jonsson, J.-Å. Gustafsson, C. Williams, and C. Zhao, "Interplay between AP-1 and estrogen receptor $\alpha$ in regulating gene expression and proliferation networks in breast cancer cells," *Carcinogenesis*, vol. 33, no. 9, pp. 1684–1691, 2012.
[65] O. Mariani, C. Brennetot, J.-M. Coindre, N. Gruel, C. Ganem, O. Delattre, M.-H. Stern, and A. Aurias, "JUN oncogene amplification and overexpression block adipocytic differentiation in highly aggressive sarcomas," *Cancer Cell*, vol. 11, no. 4, pp. 361–374, 2007.
[66] T. Taniguchi, S. Karnan, T. Fukui, T. Yokoyama, H. Tagawa, K. Yokoi, Y. Ueda, T. Mitsudomi, Y. Horio, T. Hida, et al., "Genomic profiling of malignant pleural mesothelioma with array-based comparative genomic hybridization shows frequent non-random chromosomal alteration regions including JUN amplification on 1p32," *Cancer Sci.*, vol. 98, no. 3, pp. 438–446, 2007.
[67] L. H. Saal, S. K. Gruvberger-Saal, C. Persson, K. Lövgren, M. Jumppanen, J. Staaf, G. Jönsson, M. M. Pires, M. Maurer, K. Holm, et al., "Recurrent gross mutations of the PTEN tumor suppressor gene in breast cancers with deficient DSB repair," *Nature Genetics*, vol. 40, no. 1, pp. 102–107, 2008.
[68] Y. Nagata, K.-H. Lan, X. Zhou, M. Tan, F. J. Esteva, A. A. Sahin, K. S. Klos, P. Li, B. P. Monia, N. T. Nguyen, et al., "PTEN activation contributes to tumor inhibition by trastuzumab, and loss of PTEN predicts trastuzumab resistance in patients," *Cancer Cell*, vol. 6, no. 2, pp. 117–127, 2004.
[69] F. J. Esteva, H. Guo, S. Zhang, C. Santa-Maria, S. Stone, J. S. Lanchbury, A. A. Sahin, G. N. Hortobagyi, and D. Yu, "PTEN, PIK3CA, p-AKT, and p-p70S6K status: Association with trastuzumab response and survival in patients with HER2-positive metastatic breast cancer," *Amer. J. Pathology*, vol. 177, no. 4, pp. 1647–1656, 2010.
[70] D. Egas-Bejar, P. M. Anderson, R. Agarwal, F. Corrales-Medina, E. Devarajan, W. W. Huh, R. E. Brown, and V. Subbiah, "Theranostic profiling for actionable aberrations in advanced high risk osteosarcoma with aggressive biology reveals high molecular diversity: The human fingerprint hypothesis," *Oncosci.*, vol. 1, no. 2, 2014, Art. no. 167.
[71] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, "A census of human cancer genes," *Nature Rev. Cancer*, vol. 4, no. 3, 2004, Art. no. 177.
[72] M. de Oliveira Taveira, S. Nabavi, Y. Wang, P. Tonellato, F. J. Esteva, L. C. Cantley, and G. M. Wulf, "Genomic characteristics of trastuzumab-resistant Her2-positive metastatic breast cancer," *J. Cancer Res. Clinical Oncology*, vol. 143, pp. 1255–1262, 2017.



**Jianing Xi** received the BS degree in electronic science and technology from the School of Information Science and Technology, University of Science and Technology of China (USTC), in 2013, where he is currently working toward the PhD degree. His current research interest include machine learning algorithms for computational biology and bioinformatics. He is a student member of the IEEE.

**Ao Li** received the BS degree in biophysics from the School of Life Science, University of Science and Technology of China (USTC), in 2000 and the PhD degree in biomedical engineering from the School of Information Science and Technology, USTC, in 2005. Currently, he is an associated professor with the School of Information Science and Technology and Centers for Biomedical Engineering, USTC. His research contributions are in computational cancer genomics and bioinformatics with a focus on issues concerning systematic identification and evaluation of genome-wide variants in cancer. He is a member of the IEEE.



**Minghui Wang** received the BS degree from the School of Gifted Youth, University of Science and Technology of China (USTC), and the PhD degree in biomedical engineering from the School of Information Science and Technology, USTC, in 2006. She is an associated professor with the School of Information Science and Technology and Centers for Biomedical Engineering, USTC. Her research interests include bioinformatics, biostatistics, and machine learning. She is a member of the IEEE.