文章编号　1672-6634(2020)05-0016-11　　**DOI**　10.19728/j.issn1672-6634.2020.05.003

基于分层矩阵能量谱的个体拷贝数变异检测算法

陈念华　袁细国

(西安电子科技大学 计算机科学与技术学院,陕西 西安 710071)

摘　要　拷贝数变异是基因组变异的一种重要形式,在癌症基因组中普遍存在,其包括复发性和个体性拷贝数变异模式,前者指多样本中共同发生的拷贝数变异区域,后者指个体特异性拷贝数变异区域.本文针对个体性拷贝数变异的检测问题,提出一种基于分层矩阵能量谱的检测算法 IndivCNV,其核心思想在于:通过全变分将观察到的信号进行平滑处理,利用潜变量模型将其重建为特征与权重的乘积,以检测拷贝数变异;然后对信号进行分层,依据分层矩阵能量谱在每层的占比,将个体性拷贝数变异进行鉴别.本文通过模拟数据测试所提方法的性能,并与三种同行方法进行比较,其结果表明所提方法的优势;同时,将方法应用于真实乳腺癌样本数据中,检测到一定数量的癌症关联基因.

关键词　拷贝数变异;分层矩阵能量谱;全变分

中图分类号　N32　　**文献标识码**　A

0　引言

癌症对人类的健康和生命威胁极大,从基因分子水平上研究癌症的预防和治疗策略是当代医学急需解决的问题.近年来国际生物医学界广泛关注的一种新的基因组变异形式:拷贝数变异(copy number variation,CNV),为此提供了新的线索和思路.CNV 是一种基因组结构性变异,主要表现为长度从几 Kb 至 Mb 的染色体片段的扩增或缺失[1,2],是促使人类个体间基因差异的重要因素之一,也是引发癌细胞产生和发展的重要现象.CNV 扩增是指基因组区域的拷贝数从正常细胞二倍体到多倍体的变化,CNV 缺失是基因组区域中拷贝数减少的变异.虽然 CNV 发生的频率较低,但累积的碱基数量却大大超过了单核苷酸多态.在癌细胞中,CNV 变异通常会引起相应区域中包含的基因的剂量变化,这会影响基因的正常功能[3,4].因此,在癌症基因组中 CNV 的准确检测对于癌细胞发展机理研究及癌症诊断具有重要的现实意义[5,6].

以多样本数据为背景的 CNV 检测与分析,其过程不仅涉及到癌症样本与正常样本信号的比较,而且涉及到癌症样本本身之间的比较,那么依据 CNV 在样本中出现的频率,可将其分为复发性和个体性 CNV 模式.复发 CNV 指在多数样本中共同发生的 CNV 区域,即 CNV 在多样本中表现的频率较高,目前相关检测方法的研究非常丰富[7,8].个体 CNV 指在少部分样本中共同发生或个体特异性的 CNV,即 CNV 在多样本中表现的频率较低[9].而目前为止,针对个体性 CNV 检测的研究方法较少,但这种 CNV 模式同样非常重要.通过研究个体 CNV 与癌症的关系,不仅可以发现更多与癌症发生发展密切相关的变异,还对在医学上进行个体化的有针对性的药物开发和治疗有极大的帮助.

因此,本文提出一种名为 IndivCNV(An individual copy number variation detection algorithm based on hierarchical matrix energy spectrum)的算法,与现有方法相比,该算法主要具有 3 个特点:(1) 可以从原始数据中实现个体性 CNV 模式的检测;(2) 通过全变分将观察到的信号进行平滑处理,利用潜变量模型将其

收稿日期:2020-05-09

基金项目:国家自然科学基金面上项目(61571341) 资助

通讯作者:袁细国,男,汉族,博士,副教授,研究方向:生物信息计算,xiguoyuan@mail.xidian.edu.cn.

重建为特征与权重的乘积，以应对噪声较高情况下 CNV 的检测；(3) 对信号进行分层，根据分层矩阵能量谱在每层的占比，将能量高的复发 CNV 信号层剔除，以更准确鉴别个体性 CNV.

1　相关工作

基于阵列的比较基因组杂交技术(array-based comparative genomic hybridization，aCGH)是一种高通量、高分辨率的方法，可以用于测量数千个 DNA 区域中拷贝数的变化. 要从 aCGH 数据中检测 CNV，就必须定位信号数据中 CNV 区域与非 CNV 区域间的变化点，这些变化点会将染色体分成多个离散的片段，进一步便可以检测出 CNV. 多样本 CNV 的检测涉及多个样本，以期发现那些单样本检测无法发现的模式. 目前有许多相关方法可以对 aCGH 数据进行多样本 CNV 检测，例如 PLA(Piecewise-constant and low-rank approximation for identification of recurrent copy number variations)[10]、fastRPCA(A fused lasso latent feature model for analyzing multi-sample aCGH data)[11]、FLLat(A variational approach to stable principal component pursuit)[12]等.

PLA 将多样本 CNV 检测问题转化为矩阵分解问题，其中原始数据矩阵被分解为低秩分量、稀疏分量和噪声分量. 这三个成分分别对应于复发 CNV、个体 CNV 和随机噪声. 通过主成分分析，也就是计算出输入矩阵的奇异值分解，并使用前几个奇异向量形成一个新的低秩矩阵，可以很容易地从低秩分量中识别出复发性 CNV，从稀疏分量中识别出个体 CNV.

类似地，fastRPCA 采用线性叠加的模型，为稳定主成分跟踪(stable principal component pursuit，SPCP)引入了新的凸公式，将原始信号分解为低秩分量和稀疏分量. fastRPCA 首先建立了一个凸变分框架，然后用准牛顿法对其进行加速，并使用此创新设计了通过变分框架的快速方法. 用 aCGH 数据作为原始输入，经过以上处理，便可以从低秩分量中识别出复发性 CNV，从稀疏分量中识别出个体 CNV.

FLLat 使用潜在特征模型对 aCGH 数据进行建模，其中每个样本均通过固定数量的特征的加权组合来建模. 这些特征代表了样本组 CNV 的关键区域，并与权重相结合，描述了每个单独样本中的 CNV 区域. FLLat 在特征的估计中使用了融合最小绝对值收敛和选择算子，这在估计中既保证了数据的平滑度，也保证了数据的稀疏性.

以上这些方法虽然能较好的从多样本数据中检测出 CNV，但是都不能对个体 CNV 进行针对性的检测，因此本文提出了可以对个体 CNV 进行针对性检测的算法 IndivCNV.

2　方法

IndivCNV 算法的基本框架如图 1 所示，其输入数据格式为大小为 $L\times S$ 的矩阵 $\boldsymbol{X}$，其中 L 代表探针数，S 代表一组数据中包含的样本个数. 该算法通过以下 5 个主要步骤实现对个体 CNV 的检测：(1) 基于全变分正则化的信号层次化分解，(2) 应用融合最小绝对值收敛和选择算子，(3) 计算约束权重与特征数量 J，(4) 模型参数估计，(5) 用分层矩阵能量谱识别个体 CNV，下面将会针对每一个步骤的相关理论和实现过程进行详细阐述.

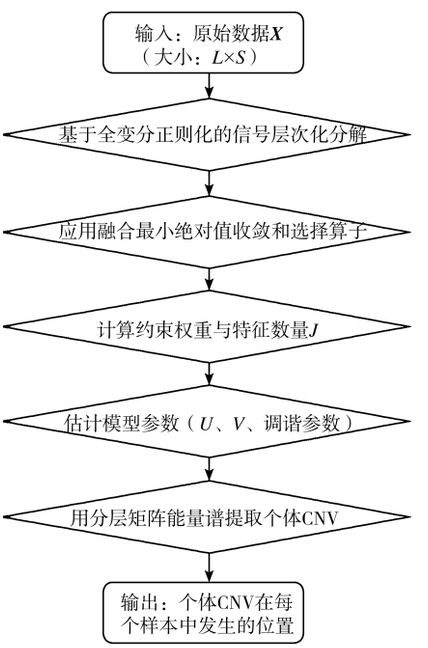


图 1　IndivCNV 的主要步骤

2.1　基于全变分正则化的信号层次化分解

本文使用潜在特征模型来模拟多样本数据，并且提出逐层分解信号的策略，通过将 CNV 的原始数据重建为不同特征模式的组合来发现原始数据中的 CNV 模式. 将两个秩为 j 的矩阵相乘的形式用 j 个秩为 1 的列向量与行向量相乘的加和来等价表示，以此来表示原始矩阵的分层分解，即

$$\boldsymbol{X}=\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}=\sum_{j=1}^{J}\hat{u}^{j}\,(\hat{v}^{j})^{\mathrm{T}}+\boldsymbol{Z}\,,\tag{1}$$

其中 $\boldsymbol{X}$ 是大小为 $L\times S$ 的矩阵，$\boldsymbol{U}=[\hat{u}^1,\cdots,\hat{u}^j,\cdots,\hat{u}^J]$ 是大小为 $L\times J$ 的矩阵，代表潜在的特征，$\hat{u}^j$ 是其列向量；$\boldsymbol{V}^{\mathrm{T}}=[(\hat{v}^1)^{\mathrm{T}},\cdots,(\hat{v}^j)^{\mathrm{T}},\cdots,(\hat{v}^J)^{\mathrm{T}}]$ 是大小为 $J\times S$ 的矩阵，代表权重，其中 $(\hat{v}^j)^{\mathrm{T}}$ 是其行向量；$\boldsymbol{Z}$ 是大小为 $L\times S$ 的矩阵，代表噪声. 这种信号分层的形式有助于后期将复发 CNV 和个体 CNV 有效区分开来.

该模型说明了样本组的 CNV 的重要特征是由 J 个特征共同总结的. 具体来说，每个特征代表 CNV 的特定模式. 然后，给定样本的权重确定每个特征对该样本的贡献程度. 换句话说，通过这些特征的权重可以知道不同特征的发生频率，以此来推断复发 CNV 和个体 CNV.

2.2 应用融合最小绝对值收敛和选择算子

CNV 区域倾向于在整个染色体的连续区域中发生，区域具有相同的拷贝数. 对于未显示 CNV 的染色体的其余部分，预期的信号强度应为零. 因此，如果我们将生物芯片数据视为沿着染色体的 1 维信号，则信号的大部分都为零，非零区域出现在平滑区域中. 通过这种 1 维信号的稀缺性和平滑性的组合可以自然地想到融合最小绝对值收敛和选择算子信号近似器(fused lasso signal approximator，FLSA[13]). FLSA 可以解决优化问题

$$\min_u \sum_{i=1}^{n}(x_i-u_i)^2+\lambda_1\sum_{i=1}^{n}|u_i|+\lambda_2\sum_{i=2}^{n}|u_i-u_{i-1}|\ , \tag{2}$$

其中 $\boldsymbol{u}=(u_1,\cdots,u_p)^{\mathrm{T}}$ 是估计所述有序结果的参数的向量. 第一个惩罚项负责惩罚每个参数大小，这可以促使解决方案稀疏，第二个惩罚项负责惩罚相邻参数之间的绝对差异，这可以促使解决方案平滑. 有 2 个相应的调谐参数，λ_1 和 λ_2 ，分别控制稀疏性和平滑性.

在多样本的情况下，需要估计在模型(1)中的 J 个参数向量 $\hat{u}^1,\cdots,\hat{u}^J$. 因为每个特征都分别描述了 CNV 的特定模式，所以应当对每个特征都应用 FLSA，以使每个估计特征中的平滑性和稀疏性更加明显. 具体而言，为了适应模型(1)，可以最小化以下式子估计 $\boldsymbol{U}$ 和 $\boldsymbol{V}$

$$F(\boldsymbol{U},\boldsymbol{V})=\left\|\boldsymbol{X}-\sum_{j=1}^{J}\hat{u}^j(\hat{v}^j)^{\mathrm{T}}\right\|_F^2+\sum_{j=1}^{J}P_{\lambda_1,\lambda_2}\hat{u}^j\ , \tag{3}$$

其中 $P_{\lambda_1,\lambda_2}\hat{u}^j=\lambda_1\sum_{l=1}^{L}|u_{lj}|+\lambda_2\sum_{l=2}^{L}|u_{lj}-u_{l-1,j}|$. 式(3)中的第一项是平方误差的平方和，第二项是应用于每个特征的 FLSA. 使用交替最小二乘算法来解决这个最小化问题，也就是通过固定 $\boldsymbol{U}$ 来更新 $\boldsymbol{V}$，然后固定 $\boldsymbol{V}$ 来更新 $\boldsymbol{U}$，直到解决方案收敛. 关于收敛问题，这是一个双凸优化问题，所以可能存在许多局部最小值. 因此，无法保证最终结果是收敛到全局最优的. 但是，如果在算法的每个步骤中都使式(3)降低，结果将收敛到局部最小值. 增加获得全局最小值可能性的一种可能方法是使用一系列不同的初始值进行交替运算，并选择达到最小标准的解决方案. 可以通过将 $\boldsymbol{U}$ 设置为 $\boldsymbol{X}$ 的 J 列的随机选择来选择初始值. 或者，可以通过将 $\boldsymbol{U}$ 设置为 $\boldsymbol{X}$ 的第一个主成分来对算法进行初始化.

2.3 约束权重与特征数量 J

对于有用的特征的估计，有必要适当地约束权重. 不受约束的权重可能导致融合最小绝对值收敛和选择算子惩罚以及模型可识别性的问题. 例如，将特定特征 $\hat{u}^{(j)}$ 乘以常数 $0<c<1$，并且将相应的权重除以相同的常数使得拟合不变，但是减少了惩罚. 考虑到这些问题，本方法在权重上放置了以下 L 2 约束

$$\sum_{s=1}^{S}v_{js}^2\leqslant 1\text{for each } j\ , \tag{4}$$

约束(4)对每行 $\boldsymbol{V}$ 的大小设置了限制，即对应于给定特征的权重. 在此认为这是限制权重大小的最合适方式. 首先，它使估计的特征之间的直接比较更有意义；其次，它可以防止大部分权重仅分布在少数几个特征上.

模型(1)中需要对特征J 的数量做出选择. 从理论上讲，J 可以取{1，2，…，S }中的任何值，其中 S 是样本数. J 的最好的选择对于任何给定的数据集都是难以确定的，并可能取决于许多因素，例如，噪声的水平，调谐参数λ_1和λ_2的值，以及 S 的值. 因此，J 的值通常留给用户指定，默认设置为 min {15，S/2}. 本方法也提供选择J 的半自动过程，这是基于解释的变化百分比(PVE). 对于给定的J 值，PVE 被定义为

$$PVE_J=1-\frac{\sum_{s=2}^{S}\sum_{l=1}^{L}\left(x_{ls}-\sum_{j=1}^{J}\bar{u}_{lj}\bar{v}_{js}\right)^2}{\sum_{s=2}^{S}\sum_{l=1}^{L}(x_{ls}-\bar{x}_s)^2}\ , \tag{5}$$

其中 $\bar{u}_{lj}$ 和 $\bar{v}_{js}$ 是由上一步产生的估计值，并且 $\bar{x}_s = \sum_{l=1}^{L} x_{ls}/L$. 随着更多特征被添加到模型中，估计的拟合度会得到改善并且使 PVE 增加. 然而，在某一点之后，继续增加特征数量将不会显著改善估计的拟合度，并且基本上是多余的. 因此，PVE 将倾向于超越这一点. 因此，通过将 PVE 与特征的数量相关联，用户可以选择使得 PVE 值开始平稳的 J 值.

2.4　模型参数估计

2.4.1　*估计* $\boldsymbol{U}$ *和* $\boldsymbol{V}$. 当 $\boldsymbol{V}^{\mathrm{T}}$ 被保持固定，就使用一个块坐标下降[14]的方法来估计 $\boldsymbol{U}$. 也就是说，要估计每个特征 $\hat{u}^j$ ，就需要通过循环 j=1,2,…,J，直到估计值收敛. 具体地，对于固定 $\boldsymbol{V}^{\mathrm{T}}$ 并且 $\{\hat{u}^k\}_{k\neq j}$ ，$\hat{u}^j$ 的解由下式给出

$$\hat{u}^j = \underset{u^j}{\operatorname{argmin}} \sum_{l=1}^{L} (\dot{x}_{lj} - u_{lj})^2 + P_{\lambda_1,\lambda_2}\hat{u}^j \ , \tag{6}$$

其中 $\dot{x}_{lj} = \sum_{s=1}^{S} \breve{x}_{lsj}\, v_{js} / \sum_{s=1}^{S} v_{js}^2$ ，$\breve{x}_{lsj} = x_{ls} - \sum_{k=j} u_{lk}\, v_{ks}$ ，$\dot{\lambda} = \lambda / \sum_{s=1}^{S} v_{js}^2$. 在此可以通过将 *FLSA* 应用于 $(\dot{x}_{1j},\cdots,\dot{x}_{Lj})^{\mathrm{T}}$ 来解决(6). 由于该解决方案取决于 $\{\hat{u}^k\}_{k\neq j}$ ，因此需要循环每个 j 直到解收敛.

当 $\boldsymbol{U}$ 保持固定时，便再次使用块坐标下降方法来估计 $\boldsymbol{V}^{\mathrm{T}}$. 通过循环 $j=1,2,\cdots,J$ 估计权重 $(\hat{v}^j)^{\mathrm{T}} = (v_{1j},\cdots,v_{Sj})$ ，直到估计值收敛. 对于固定 $\boldsymbol{U}$ 并且 $\{(\hat{v}^k)^{\mathrm{T}}\}_{k\neq j}$ ，$(\hat{v}^{(j)})^{\mathrm{T}}$ 的解由(7)给出

$$(\hat{v}^j)^{\mathrm{T}} = (\tilde{x}_{j1},\cdots,\tilde{x}_{jS}) / \max\Big\{ \sum_{l=1}^{L} u_{lj}^2, (\sum_{s=1}^{S} \tilde{x}_{js}^2)^{\frac{1}{2}} \Big\} \ , \tag{7}$$

其中 $\tilde{x}_{js} = \sum_{l=1}^{L} \breve{x}_{lsj}\, u_{lj}$ ，$\breve{x}_{lsj} = x_{ls} - \sum_{k=j} u_{lk}\, v^{\mathrm{T}}{}_{ks}$ ，$s=1,2,\cdots,S$. 由于该解决方案取决于 $\{(\hat{v}^k)^{\mathrm{T}}\}_{k\neq j}$ ，因此需要循环每个 j 直到解收敛.

2.4.2　*选择融合最小绝对值收敛和选择算子调谐参数* $\boldsymbol{\lambda_1}$ *和* $\boldsymbol{\lambda_2}$. 通常，给定模型的最佳调谐参数的选择都是一个困难的任务，并且随着调谐参数数量的增加会更加复杂. 为了简化对最佳调谐参数的搜索，本方法通过引入 λ_0 和 $\alpha \in (0,1)$ 来重新定义参数 λ_1 和 λ_2 ，使得 $\lambda_1 = \alpha\lambda_0$ ，$\lambda_2 = (1-\alpha)\lambda_0$. 在此可以认为 λ_0 是整体调谐参数，它和 α 一起确定对稀疏度与平滑度的重视程度. 通过固定 α 可能采取的值，可以有效地将对两个参数 λ_1 和 λ_2 的搜索简化为仅对一个参数 λ_0 的搜索.

具体来说，本方法最初修正了 α 的可能值(例如{0.1,0.3,0.5,0.7,0.9}). 对于每个 α 值确定的值 λ_0 ，这样每个估计的特征都是恒定的，由 $\lambda_{0,\alpha}^{max}$ 表示该值. 然后对 λ_0 从区间 $(0,\lambda_{0,\alpha}^{max})$ 选择的候选值的固定数目(例如 5). λ_0 和 α 的最优值通过在这 2D 网格中最小化以下基准值进行搜索选择

$$(SL)\cdot\log\Big(\frac{\Big\|X - \sum_{j=1}^{J}\hat{u}^j\,(\hat{v}^j)^{\mathrm{T}}\Big\|_F^2}{SL}\Big) + k_{\alpha,\lambda_0}\log(SL) \ . \tag{8}$$

在此，定义 $k_{\alpha,\lambda_0} = \sum_{j=1}^{J} k_{\alpha,\lambda_0}(j)$ ，其中 $k_{\alpha,\lambda_0}(j)$ 是第 j 个特征 $\hat{u}^j$ 里非零元素的数目. k_{α,λ_0} 代表了模型的复杂度，其值越大则说明模型的复杂度越高，S 和 L 分别是样本数和探针数. 当假设(1)中的模型误差是正态分布时，式(8)就类似于贝叶斯信息准则. 这个式子背后的基本原理是，通过最小化(8)，可以尽量找到一个合适的模型而不会导致过度拟合数据，因为对于复杂模型，第一项往往会更小，而第二项则趋向于在简单模型中的值更小. 出于计算的考虑原因，相对于基于交叉验证的方法，用这种方法来选择最佳调谐参数会更合适.

2.5　分层矩阵能量谱

对于 $\boldsymbol{X} = \boldsymbol{U}\,\boldsymbol{V}^{\mathrm{T}} = \sum_{j=1}^{J}\hat{u}^j\,(\hat{v}^j)^{\mathrm{T}}$ ，每层特征被提取出来，此时面临的任务就是如何区分出复发 CNV 与个体 CNV. 本文提出分层矩阵能量谱的方法，根据每层特征的能量占比，来将二者区分. 通常来讲，复发 CNV 会在所有或者大量样本中同时出现，这就导致复发 CNV 的变异模式出现频率很高，代表其特征的特定层矩阵所占有的能量也必然相较于其他特征更高；个体 CNV 则与之相反，由于个体 CNV 是随机发生在不同样本中，个体性更高，因此出现的频率相对较低，代表其特征的特定层矩阵所占有的能量就相对较低. 基于以上

原理,可由下式算出每一层的能量

$$\sum_{l=1}^{L}\sum_{s=1}^{S}(\hat{u}^j(\hat{v}^j)^{\mathrm{T}})_{ls}^2\ , \tag{9}$$

其中 $(\hat{u}^j(\hat{v}^j)^{\mathrm{T}})_{ls}^2$ 代表第 j 层矩阵第 l 行 s 列元素的平方,整个式子代表该层矩阵的能量. 根据不同层之间的能量大小排序,可认为能量占比大于某阈值的层数代表复发 CNV,剩下的则代表个体 CNV,将代表复发 CNV 的矩阵层剥离,就得到个体 CNV 的矩阵信息,即

$$\text{if}\ \frac{\sum_{j=1}^{M}\sum_{l=1}^{L}\sum_{s=1}^{S}(\hat{u}^j(\hat{v}^j)^{\mathrm{T}})_{ls}^2}{\sum_{j=1}^{J}\sum_{l=1}^{L}\sum_{s=1}^{S}(\hat{u}^j(\hat{v}^j)^{\mathrm{T}})_{ls}^2}>T\ ,\text{Then}\ I=X-\sum_{j=1}^{M}\sum_{l=1}^{L}\sum_{s=1}^{S}(\hat{u}^j(\hat{v}^j)^{\mathrm{T}})_{ls}^2\ , \tag{10}$$

其中 T 代表设定的占比阈值,$\boldsymbol{I}$ 是大小为 $L\times S$ 的矩阵,代表个体 CNV. 得到最终的个体 CNV 矩阵 $\boldsymbol{I}$ 以后,需要按照样本将数据区分为 S 个大小为 $L\times 1$ 的矩阵,每个矩阵代表每个样本的结果. 此时,需要再选定一个阈值 H,若某探针处的绝对值大于 H,则认为该处有个体 CNV,反之则认为是正常. 因为个体 CNV 在样本间有很大的差异,所以需要按上述对每个样本的结果数据都分别判断.

3 实验结果

3.1 模拟数据

3.1.1 *模拟数据介绍.* 为了评估 IndivCNV 算法对个体 CNV 的检测性能,本节将采用模拟数据进行实验,并与三种现有方法(PLA、FLLat、fastRPCA)进行比较. 在文献[15]里,详细地定义了六种不同的复发 CNV 场景. 在本文的研究里,将采用这六种场景来生成模拟数据. 在每一种场景下生成 50 组数据,每组数据是 50×5000 的矩阵,其中 50 代表 50 个样本,5000 代表每个样本上的 5000 个探针. 在生成每组数据时,无 CNV 区域的信号值设为 0;复发 CNV 区域位于探针 1876 到 3125 之间,其模式参考图 2,将缺失变异区域的信号值设为 −1,扩增变异区域设为 1. 每个样本还需要在不与复发 CNV 区域重合的部分,随机选取一个位置,添加一个长度为 500 探针的个体 CNV,个体 CNV 的信号值从{−2,−1,1,2}中随机选取,最后再向整个数据加入高斯噪声.

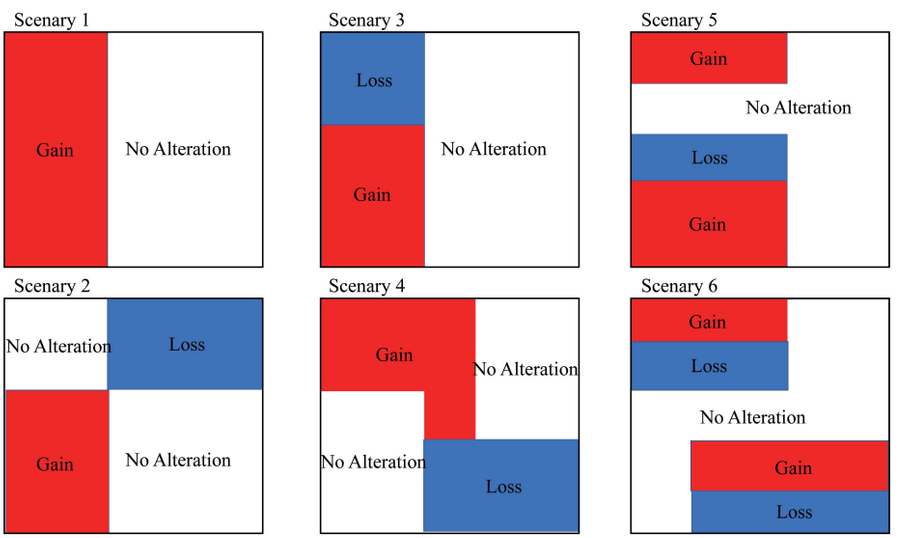


图 2 在 Rueda and Diaz-Uriarte (2010)里定义的六种常见复发 CNV 的模式. 每个场景的纵轴代表样本,横轴代表探针

6 种不同场景生成模拟数据的过程展示在图 3,图中黄色区域代表扩增,蓝色区域代表缺失. 其中第一行是根据文献[15]中对不同场景的描述生成的只有复发 CNV 的数据,第二行是在复发 CNV 的基础上随机添加个体 CNV 的数据,第三行是添加了噪声水平为 1 的高斯噪声的最终模拟数据. 每组数据的纵向代表样本,横向代表探针. 从图上可以看出,这六种场景可以分为两类,场景 1、3、5 为一类,它们只有一个复发 CNV 区域;场景 2、4、6 为一类,它们含有多个复发 CNV 区域. 本文的研究任务是从这些最终的模拟数据里准确恢复出个体 CNV.

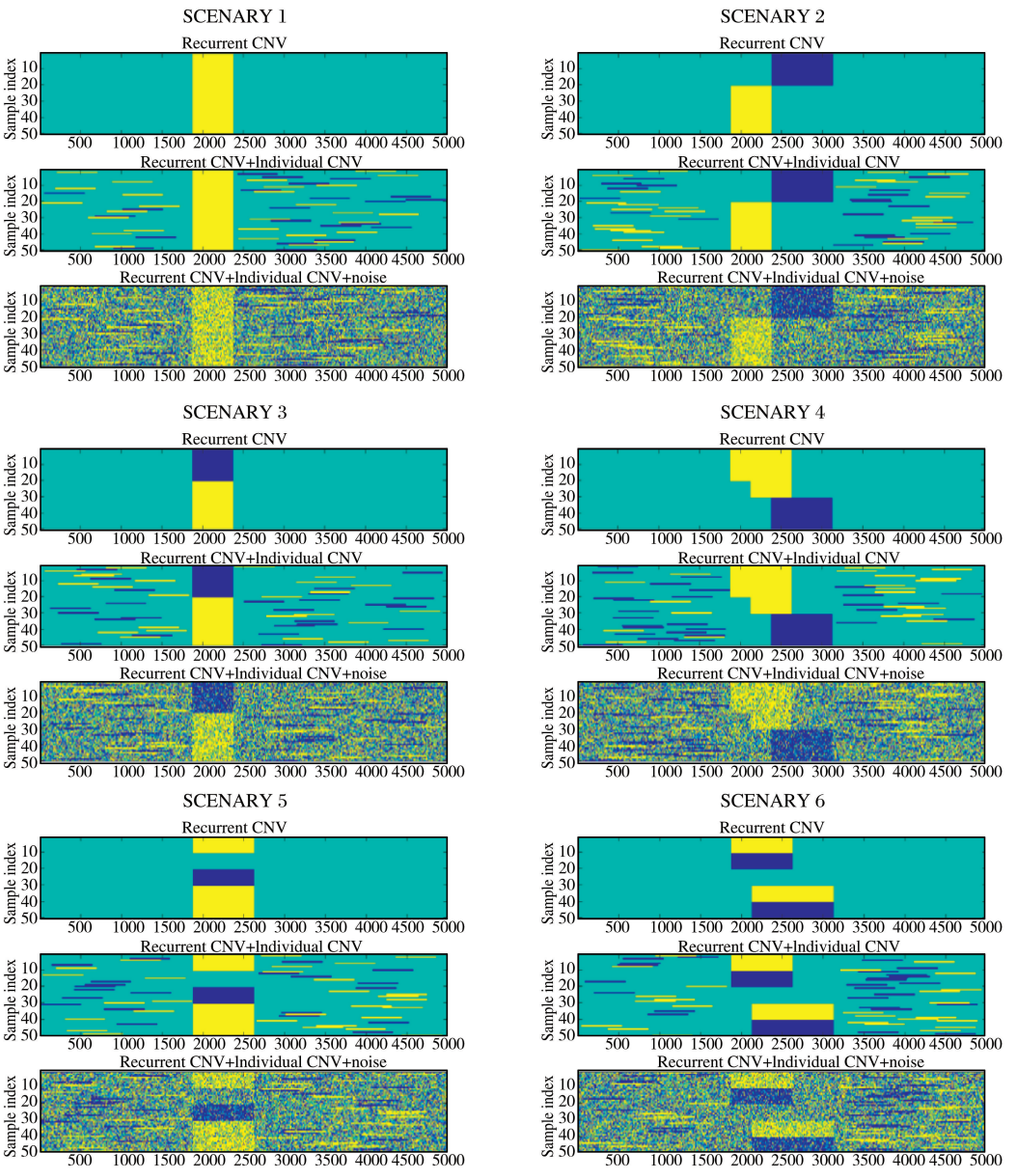
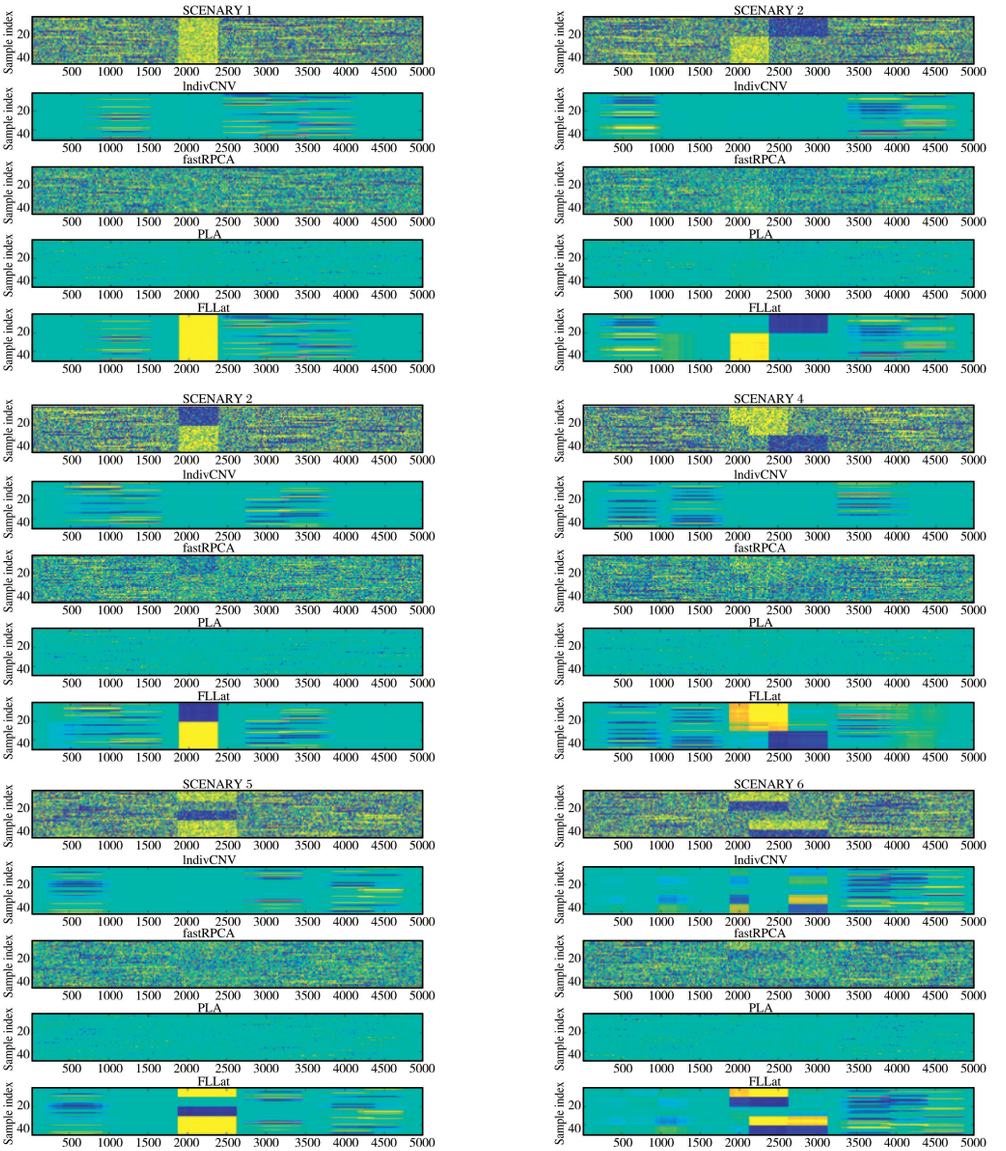


图 3　场景 1-6 模拟数据的生成过程

3.1.2　检测结果热图展示. 在图 4 中展示了在 6 种场景下不同方法对个体 CNV 的检测结果. 从图中可以看出来 IndivCNV 检测出了绝大部分的个体 CNV，并且能很好地把个体 CNV 与复发 CNV 区分开来，没有将复发 CNV 误判为个体 CNV. FastRPCA 可以分辨出一部分个体 CNV，但是没有将噪声很好地剔除，因此难以识别检测出的个体 CNV 的模式；而 PLA 则倾向于将一个完整的个体 CNV 切割成多个小段，有明显的缺失；FLLat 的特点是它做检测时不对复发 CNV 与个体 CNV 进行区分，导致两种类型的 CNV 都存在于结果数据中. 由以上分析可知 IndivCNV 在检测个体 CNV 时确实更加有优势，但是从图中可以看出它还是存在一定的缺陷，因为它更趋向于检测出发生个体 CNV 频率较高的位置的变异，而对于发生频率小的个体 CNV，则很难检测出.

3.1.3　检测结果 ROC 曲线. 为了可量化地评估这些方法，本研究进一步通过 ROC 曲线评估各方法在六种场景下的个体 CNV 识别性能. ROC(receiver operating characteristic curve)是一种显示分类模型在所有分类阈值下的效果的图表，其横轴是假阳性率（False Negative Rate，FPR），纵轴是真阳性率（True Negative Rate，TPR). FPR 指的是所有非个体 CNV 区域中被误判为个体 CNV 的比率，该值越小越好，TPR 指的是在所有检测出来为个体 CNV 的区域里，确实是个体 CNV 的比率，该值越大越好. ROC 曲线的作用在于，在很多分类器分析中，得到的预测值通常不是 0 或 1，而是一个 0-1 之间的概率值，此时就需要人为设定一个阈

值,比如设定大于 0.6 则为 1,反之则为零.但是不同的阈值所带来的预测结果一定有差异,此时就可以用 ROC 曲线来刻画不同阈值给分类器带来的影响.通过上文对 FPR 和 TPR 含义的介绍可知,ROC 曲线越靠近左边沿和上边沿,说明模型越好,因为此时 TPR 足够大,FPR 足够小,说明分类的正确率很高.而 ROC 曲线上不同的点对应着模型对不同阈值的预测水平,简单来讲,阈值越大,点越靠近左下,反之越靠近右上.



注:第 1 行是模拟数据,第 2 行到第 5 行分别是 IndivCNV、fastRPCA、PLA、FLLat 的检测结果.

图 4 不同方法对场景 1-6 模拟数据的个体 CNV 的检测结果

图 5 展示了各方法在 6 种场景下的 ROC 曲线.这些 ROC 曲线是通过对各方法检测出来的结果数据设定不同的阈值生成的.从图上可以看出,IndivCNV 检测个体 CNV 的性能优于其他三种方法.例如在场景 1 的 ROC 曲线中,当 FPR=0.1 时,IndivCNV 的 TPR 就已达到 0.8,而 FLLat 的 TPR 只有 0.45,PLA 和 fastRPCA 的 TPR 仅有 0.3;在场景 2 中,虽然当 FPR 值大于 0.3 时,FLLat 和 IndivCNV 的曲线基本重合,但是 IndivCNV 在 FPR=0.05 时 TPR 就已经达到了 0.7,这说明 IndivCNV 在低 FPR 水平就可以实现较高水平的 TPR;在场景 3、5、6 中,呈现出同样的趋势:当 FPR 较高时,FLLat 与 IndivCNV 的曲线十分接近,但是始终都低于 IndivCNV,只有在场景 4 中曲线的后半段 FLLat 超过了 IndivCNV,尽管如此,其前半段依旧远低于 IndivCNV 的 ROC.

综上所述,与 fastRPCA 和 PLA 相比,IndivCNV 和 FLLat 算法对个体 CNV 的识别结果具有更高的 TPR.然而,FLLat 的性能与 IndivCNV 虽然较为接近,但仅表现在 FPR 较高的情况,当 FPR 较低时,其

ROC 曲线依旧远低于 IndivCNV. 因此，在对个体 CNV 的检测中，IndivCNV 算法具有更明显的优势.

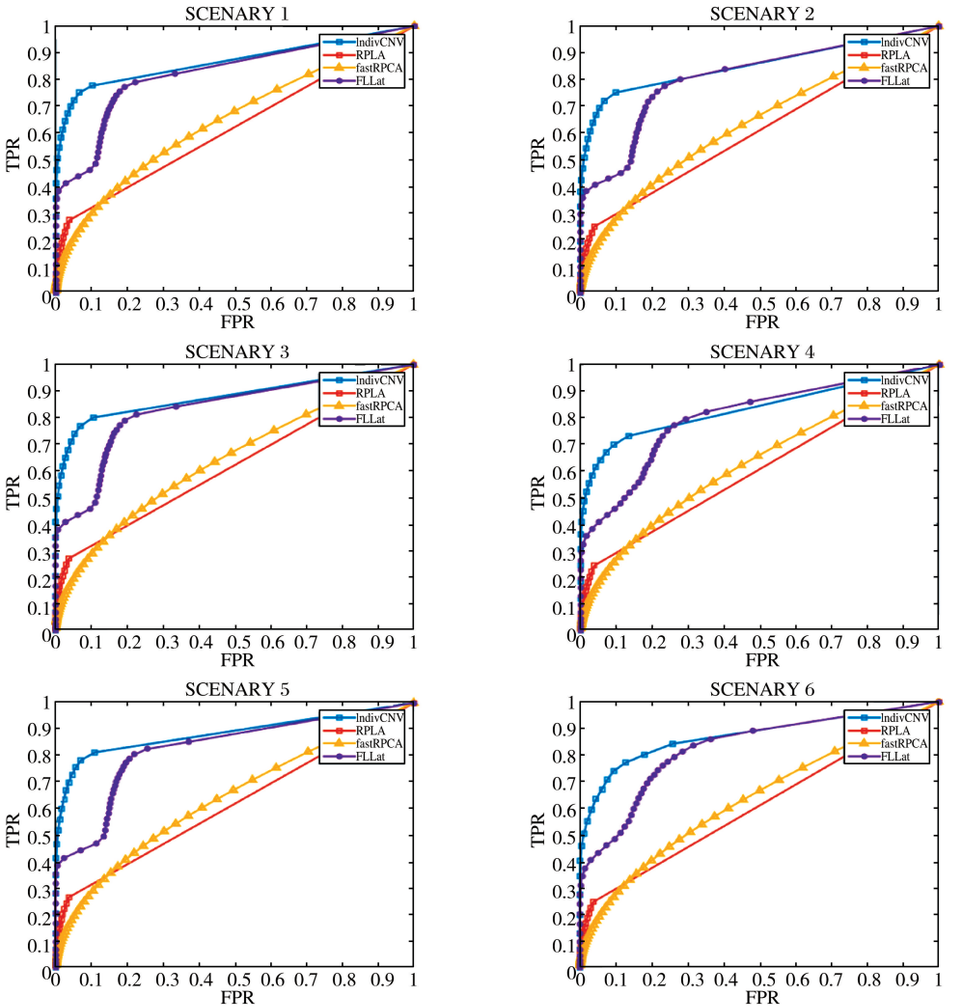


图 5　不同方法在场景 1-6 下的 ROC 曲线

3.2　真实数据

为了证明 IndivCNV 在真实数据上的可用性，本实验引入异质性乳腺癌 CNA 真实数据集对算法进行验证. 这个数据集中包含了 112 个乳腺癌样本的 SNP array 数据，每个样本都有 23 条染色体上的不同数据，每条染色体的探针各不相同，由 Illumina 109 K SNP array 平台采集. 在进行实验时，首先将每个样本不同染色体上的数据分割开来，然后将处理所得的 CNV 分段在基因组区域对齐，成为一个大小为 $112 \times p_i$ 的变异强度矩阵，其中 112 代表样本数，p_i 代表在第 i 条染色体上的探针数，即分割完成后有 22 个变异强度矩阵（因为乳腺癌是常染色体上的疾病，所以仅对前 22 条常染色体进行实验），并分别对这 22 个信号矩阵进行实验分析. 在实验过程中，使用 IndivCNV 对数据进行分析，阈值 T 设为 0.1. 为了消除每个样本中的波谱偏差，需通过局部中值减去信号数据，中值计算的窗口大小是染色体长度的四分之一.

对于 IndivCNV 算法在乳腺癌数据中所发现的个体 CNV 区域，本研究通过乳腺癌相关文献报道的 CNV 区域对算法结果进行验证. 对于 IndivCNV 算法所发现的个体 CNV 区域，其中许多区域被现有文献报道为乳腺癌 CAN 驱动区域. 例如，IndivCNV 算法成功识别出 17 号染色体上的 ERBB2 基因[16]，该基因曾被多项研究报道为乳腺癌 CAN 驱动变异. 同时，IndivCNV 在 14 号染色体发现 AKT1 基因[17]，而该基因则被报道与乳腺癌的发生发展密切相关. 表 1 汇总了 IndivCNV 所发现的个体 CNV 与现有文献报道发现与乳腺癌有密切关系的基因重合的结果. 上述结果表明，IndivCNV 算法所发现的个体 CNV 区域与已报道 CNV 驱动变异区域具有较高的一致性.

表 1 IndivCNV 检测出与现有文献报道发现与乳腺癌有密切关系的基因重合的结果

序号	染色体	基因名称	开始位置	结束位置
1	chr1	ARID1A	26895109	26981188
2	chr2	BARD1	215301520	215382673
3	chr2	CASP8	201806411	201860679
4	chr3	BAP1	52410065	52419049
5	chr3	FBLN2	13565625	13654923
6	chr3	PBRM1	52554408	52694906
7	chr6	ARID1B	157140756	157573603
8	chr6	ESR1	152053324	152466101
9	chr11	CCND1	69165054	69178423
10	chr12	CDKN1B	12761569	12766572
11	chr12	ETV6	11694055	11939592
12	chr13	BRCA2	31787617	31871809
13	chr13	RB1	47775884	47916841
14	chr14	AKT1	104306732	104333125
15	chr16	CDH1	67328696	67426945
16	chr16	CTCF	66153811	66230588
17	chr17	ERBB2	35097919	35138441
18	chr20	SALL4	49833990	49852455

表 2 IndivCNV 的复发 CNV 模式所匹配到的乳腺癌相关基因

序号	染色体	基因名称	开始位置	结束位置
1	chr3	MAP3K13	186563664	186683322
2	chr3	PIK3CA	180349005	180435191
3	chr5	MAP3K1	56146657	56227735
4	chr10	GATA3	8136673	8157170
5	chr 11	CCND1	69165054	69178423
6	chr 12	SMARCD1	48765250	48780761
7	chr 14	FOXA1	37128942	37134240
8	chr 15	NTRK3	86220992	86600665
9	chr 17	MAP2K4	11864860	11987776
10	chr 17	NCOR1	15874134	16059599
11	chr 17	PPM1D	56032336	56096818
12	chr 17	TP53	7512445	7531588
13	chr 19	KEAP1	10457796	10475054
14	chr 20	SALL4	49833990	49852455

正如第 3 节所说，IndivCNV 在做个体 CNV 模式检测的过程中，会将复发 CNV 的模式剔除. 在此，本实验在用该真实数据检测时，将剔除的复发 CNV 数据也另行保存，并对复发 CNV 模式进行驱动基因匹配. 表 2 中汇总了在 IndivCNV 的复发 CNV 模式中发现的乳腺癌驱动基因，表 3 汇总了 IndivCNV 在真实数据检测出的个体 CNV 模式在复发 CNV 模式之外发现的驱动基因. 由表 2、3 可以看出，个体 CNV 的检测可以很大程度上弥补复发 CNV 对驱动基因发现的不足，例如，在表 2 复发 CNV 的检测结果里，未发现 1 号染色体和 13 号染色体上有与乳腺癌相关的基因，而在个体 CNV 模式里则发现了 1 号染色体上的 ARID1A 基因，13 号染色体上的 BRCA2 基因和 RB1 基因，这几个基因都是乳腺癌相关基因，并被权威癌症基因数据库 Cancer Gene Census 所收录[18-20]. 上述结果表明，IndivCNV 算法的个体 CNV 发现结果可有效弥补现有方法发现结果的不足，同时也证明了个体 CNV 检测对于癌症研究的重要性.

表 3　IndivCNV 的个体 CNV 模式在其复发模式之外检测到的乳腺癌相关基因

序号	染色体	基因名称	开始位置	结束位置
1	chr1	ARID1A	26895109	26981188
2	chr2	BARD1	215301520	215382673
3	chr2	CASP8	201806411	201860679
4	chr3	BAP1	52410065	52419049
5	chr3	FBLN2	13565625	13654923
6	chr3	PBRM1	52554408	52694906
7	chr6	ARID1B	157140756	157573603
8	chr6	ESR1	152053324	152466101
10	chr12	CDKN1B	12761569	12766572
11	chr12	ETV6	11694055	11939592
12	chr13	BRCA2	31787617	31871809
13	chr13	RB1	47775884	47916841
14	chr14	AKT1	104306732	104333125
15	chr16	CDH1	67328696	67426945
16	chr16	CTCF	66153811	66230588
17	chr17	ERBB2	35097919	35138441

4　结论

CNV 是导致癌症发生发展的重要因素之一，由于现有研究更侧重于对复发 CNV 的研究，对个体 CNV 的关注程度不够，忽略了个体 CNV 的研究价值，因此本文通过分析个体 CNV 的模式，提出了一种新的适用于发现个体 CNV 的算法 IndivCNV. IndivCNV 首先需要使原始信号趋于平滑，因此采用了全变分正则化的方式达到此目的；接着将原始数据的每个样本建模为固定数量的特征的加权和，这一步使用了潜变量模型和融合最小绝对值收敛和选择算子惩罚；然后使用信号层次化分解，将不同模式的 CNV 用不同层的矩阵表示；最后利用分层矩阵能量谱，根据复发 CNV 模式能量占比大，个体 CNV 模式的能量占比小的原理，将复发 CNV 与个体 CNV 区分开来，最终达到检测个体 CNV 的目的.

在本文的实验中，首先将 IndivCNV 应用到六种不同场景的模拟数据上，同时将 fastRPCA、PLA、FL-Lat 这三种算法也应用到该模拟数据上，以 ROC 曲线为性能判断标准，根据检测结果选定不同阈值绘制 ROC，以此进行性能对比，实验结果表明，IndivCNV 检测个体 CNV 的性能显著高于已有的三种方法的性能. 然后又使用 IndivCNV 检测异质性乳腺癌 CNA 真实数据集中的个体 CNV，检测个体 CNV 结果中包含许多现有文献已报道过与乳腺癌相关的基因，还发现了复发 CNV 模式没有发现的与乳腺癌相关的基因，因此 IndivCNV 的性能在实际数据上也得到了验证. 综上所述，IndivCNV 在个体 CNV 方面的检测性能确实有了大幅提升.

参　考　文　献

[1] Beroukhim R, Mermel C H, Porter D, et al. The landscape of somatic copy-number alteration across human cancers[J]. Nature, 2010, 463(7283): 899-905.

[2] Zhang J, Feuk L, Duggan G E, et al. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome[J]. Cytogenetic & Genome Research, 2006, 115(3): 205-214.

[3] Stephens P J, McBride D J, Lin M-L, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes[J]. Nature, 2009, 462(7276): 1005-1010.

[4] Bignell G R, Greenman C D, Davies H, et al. Signatures of mutation and selection in the cancer genome[J]. Nature, 2010, 463(7283): 893-898.

[5] Conrad D F, Pinto D, Redon R, et al. Origins and functional impact of copy number variation in the human genome[J]. Nature, 2010, 464

(7289):704-712.

[6] Yuan Xiguo,Yu Jiaao,Xi Jianing,et al. CNV-IFTV: an isolation forest and total variation-based detection of CNVs from short-read sequencing data[J]. IEEE/ACM Trans Comput Biol Bioinform,2019,19:1-12.

[7] Yuan Xiguo,Yu Guoqiang,Hou Xuchu,et al. Genome-wide identification of significant aberrations in cancer genome[J]. BMC Genomics,2012,13(1):1-14.

[8] Yuan Xiguo,Zhang Junying,Yang Liying,et al. Detection of Significant Copy Number Variations From Multiple Samples in Next-Generation Sequencing Data[J]. IEEE Transactions on Nanobioscience,2017,17(1):12-20.

[9] Curtis C,Shah S P,Chin S F,et al. The genomic and transcriptomic architecture of 2, 000 breast tumors reveals novel subgroups[J]. Nature,2012,486(7403):346-352.

[10] Zhou Xiaowei,Liu Jiming,Wan Xiang,et al. Piecewise-constant and low-rank approximation for identification of recurrent copy number variations[J]. Bioinformatics,2014,30(14):1943-1949.

[11] Nowak G,Hastie T,Pollack J R,et al. A fused lasso latent feature model for analyzing multi-sample aCGH data[J]. Biostats,2011,12(4):776-791.

[12] Aravkin A,Becker S,Cevher V,et al. A variational approach to stable principal component pursuit[J]. Mathematics,2014,14:1-9.

[13] Tibshirani R,Saunders M,Rosset S,et al. Sparsity and smoothness via the fused lasso[J]. Journal of the Royal Statistical Society,2005,67(1):91-108.

[14] Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization[J]. Journal of Optimization Theory and Applications,2001,109(3):475-494.

[15] Rueda O M,Diaz-Uriarte R. Finding Recurrent Copy Number Alteration Regions: A Review of Methods[J]. Current Bioinformatics,2010,5(1):1-17.

[16] Haverty P M,Fridlyand J,Li L,et al. High-resolution genomic and expression analyses of copy number alterations in breast tumors[J]. Genes,Chromosomes and Cancer,2008,47(6):530-542.

[17] Hicks J,Krasnitz A,Lakshmi B,et al. Novel patterns of genome rearrangement and their association with survival in breast cancer[J]. Genome Research,2007,16(12):1465-1479.

[18] Futreal P A,Coin L,Marshall M,et al. A census of human cancer genes[J]. Nature Reviews Cancer,2004,4(3):177-183.

[19] Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer[J]. Nature,2012,487(7407): 330-337.

[20] Lancaster J M,Wooster R,Mangion J,et al. BRCA2 mutations in primary breast and ovarian cancers[J]. Nature Genetics,1996,13(2):238-40.

An Individual Copy Number Variation Detection Algorithm Based on Hierarchical Matrix Energy Spectrum

CHEN Nian-hua YUAN Xi-guo

Abstract Copy number variation is an important form of genomic variation that is ubiquitous in cancer genomes. It includes recurrent and individual copy number variation patterns. The former refers to areas of copy number variation that co-occur in multiple samples,and the latter refers to individual specificity copy number variation areas. In this paper, for the detection of individual copy number variation,a detection algorithm named IndivCNV which is based on hierarchical matrix energy spectrum is proposed. Its main ideas are as follows: Firstly,smooth the observed signal through total variation and use latent features to reconstructs it into the product of features and weights;then the signal is layered,and the individual copy number variations are identified according to the proportion of the energy spectrum of the layered matrix in each layer. To test the performance of the proposed method,this paper used simulated data and compared with three peer methods. The results showed the advantages of the proposed method;at the same time,the method was applied to real breast cancer sample data,and a certain number of cancer-related genes were detected.

Key words individual copy number variation;hierarchical matrix energy spectrum;total variation