DOI：10.7652/xjtuxb201706015

一种碱基精度的肿瘤基因组单体型异质性识别算法

耿彧[1,2,4]，赵仲孟[2,4]，刘建业[2,4]，许静[2,4]，崔代兵[2,4]，萧笑[4,5]，王嘉寅[3,4]
（1.锦州医科大学公共基础学院，121001，辽宁锦州；2.西安交通大学计算机科学与技术系，710049，西安；
3.西安交通大学管理学院，710049，西安；4.西安交通大学陕西省医疗健康大数据工程研究中心，710049，西安；5.第四军医大学肿瘤生物学国家重点实验室，710032，西安）

摘要：针对肿瘤组织的异质性的子克隆解析，提出了一种通过多级子克隆的体细胞突变模式来识别单体型异质性的算法。该算法基于肿瘤组织的多文库测序数据提取文库特征和双末端读段约束，通过对体细胞突变位点的等位基因变异频率进行聚类估算出子克隆数目的一个先验；同时设计了一种拼接识别算法，通过遍历位点对应的读段来拼接单体型序列，拼接出的单体型序列的精度为碱基水平；采用后验概率的最大似然估计解出子克隆的个数、配比及演化关系。仿真实验表明，当基础文库满足一定测序覆盖度时，该算法对单体型异质性的识别精度可达到99%以上，能够取代目前数据分析中常用的两步法，且获得高精确的识别结果。

关键词：肿瘤异质性；子克隆解析；单体型异质性；多文库测序数据；拼接识别算法
中图分类号：TP399 **文献标志码**：A **文章编号**：0253-987X(2017)06-0092-05

An Algorithm with Base-Pair Resolution for Identifying Cancer Heterogeneity by Estimating Multiple Clonal Haplotypes

GENG Yu[1,2,4], ZHAO Zhongmeng[2,4], LIU Jianye[2,4], XU Jing[2,4], CUI Daibing[2,4], XIAO Xiao[4,5], WANG Jiayin[3,4]
(1. School of General Education, Jinzhou Medical University, Jinzhou, Liaoning 121001, China; 2. Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China; 3. School of Management, Xi'an Jiaotong University, Xi'an 710049, China; 4. Shaanxi Engineering Research Center of Medical and Health Big Data, Xi'an Jiaotong University, Xi'an 710049, China; 5. State Key Laboratory of Cancer Biology, The Fourth Military Medical University, Xi'an 710032, China)

Abstract: An algorithm for identifying haplotype heterogeneity in cancer genomes is proposed to consider somatic mutational events carried by multiple sub-clones. The algorithm is based on the genomic sequencing data with multiple libraries of tumor tissue and extracts the features from both the multi-library and the constraints of paired-end reads. A priori number of sub-clones is roughly estimated by clustering the allelic variant frequency of each somatic loci. A contig-and-extension algorithm is designed, and the haplotype sequences are assembled by traversing the reads mapping to the loci. Thus, the contigs present an identification resolution on base-pair level. The number and proportion of sub-clones and the evolution relationships among them are further estimated by maximizing the likelihood of the posterior probabilities. Simulation results

收稿日期：2016-12-07。 作者简介：耿彧(1976—)，女，讲师；王嘉寅(通信作者)，男，教授，博士生导师。 基金项目：国家自然科学基金资助项目(81400632)；陕西省自然科学基金资助项目(2014JM8350)；中央高校基本科研业务费专项资金资助项目(GLIJ002)。
网络出版时间：2017-03-28 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20170328.1133.006.html

show that the algorithm reaches 99% in accuracy when the sequencing based library satisfies some coverage. The proposed algorithm outperforms the existing two-stage pipeline, which is widely used in data analysis now.

Keywords: cancer heterogeneity; sub-clone deconvolution; haplotype heterogeneity; multi-library sequencing; contig-and-extension algorithm

肿瘤异质性是恶性肿瘤的重要特征，主要指肿瘤组织中存在多个子克隆，每个子克隆携带使之具备独特选择优势的体细胞突变等分子遗传学特征[1]。肿瘤异质性是肿瘤组织中子克隆演化的结果，普遍认为其与肿瘤生长、转移、复发、耐药性等密切相关，是非常重要的临床指标。准确估计患者肿瘤组织的异质性，包括子克隆个数、体细胞突变模式等特征，有助于推断子克隆之间、肿瘤-免疫微环境等的进化关系，进而指导肿瘤的精准诊疗[2]。

随着高通量基因组测序技术快速发展和普及，基于肿瘤基因组测序数据研究肿瘤异质性逐步成为主流，一批计算模型和算法陆续提出。根据算法的设计思想，大体可以分为两类。一类是通过构建系统发育树来推断肿瘤组织的子克隆结构，如：Schwartz 等使用基于几何模型的分离方法解析肿瘤组织的子克隆，并推断可能适合的系统发育树[3]；TrAp 方法利用贪婪算法生成所有满足特定的简约稀疏条件的系统发育树[4]；PhyloSub 和 PyClone 两种方法均采用马尔可夫-蒙特卡罗过程实现对所有可能的系统发育树的寻优[5-6]。另一类是通过构建概率模型，寻找具有最大似然的等位基因变异频率的子克隆结构，如 SciClone 和 AncessTree 方法分别采用变分贝叶斯混合模型和整数线性规划思想，重构子克隆结构[7-8]。

然而，上述方法的推断和重构结果都是在基因型层面上的，目前尚无在单体型层面解析异质性的方法。单体型相比基因型包含了更多的信息，具有更强的临床指导意义。数据分析中一般会在获得基因型结果之后接续一个或多个单体型定相步骤，这样操作的弱点在于：其一，高精度的单体型定相一般需要一个参考数据集，而体细胞突变呈现高度特异性，除了个别热点之外，不存在参考数据集；其二，两个步骤接续常常导致误差累积甚至放大。因此，本文提出一种利用肿瘤组织的多文库测序数据直接推断单体型的异质性识别算法，通过提取多文库特征和双末端读段约束，求解体细胞突变位点的等位基因变异频率的逆卷积，算法中采用一种分治策略来提高组装和拼接的效率，同时分析多个子克隆的单体型，并最终给出子克隆的个数、配比及进化关系的后验似然估计。该方法在各种贴近真实数据特征的实验中取得了良好的效果。

1　问题的数学模型

与现有方法类似，本文考虑满足完美系统发育树模型的肿瘤组织微进化过程[2]，其中体细胞突变模式遵从二次突变假设，具备选择优势，不考虑修复机制可以作用的位点[5]。

对于任意变异位点 p，其等位基因变异频率(variant allelic frequency，VAF)为支持突变的读段个数占该位点测序深度的比例，用 V_p 表示，相应数值可从读段比对数据中统计得出。令子克隆 S_i 的体细胞突变位点的集合为 M_i，$i\in\{0,1,\cdots,I\}$，测序样本的体细胞突变集合为 $M=\bigcup_i M_i$，其中克隆 S_0 表示由测序数据估计出的具有最大公共溯祖特征的子克隆，又称为基础克隆，I 表示子克隆总数。令 G_i^p 表示 S_i 携带的基因型，令子克隆 S_j 是 S_i 的子代克隆，则有原则：

(1)若 $V_p=1$，则 $p\in M_0$ 且 p 为纯合突变；

(2)若 $V_p=0.5$，则 $p\in M_0$ 且 p 为杂合突变；

(3)$p\in M_i\rightarrow p\in M_j$；

(4)$\forall\, p_i\in M_i$，若 $V_{pi}>V_{pj}$，则 $p_j\in M_j$。

其中，V_{pi}、V_{pj} 分别表示 M_i 和 M_j 上的等位基因变异频率。此外，单体型异质性遵循的继承原则包括克隆中不会出现与参考序列不同的纯合型变异位点 $\sum_k V_{i,p,k}\leqslant 1$ 以及进化的继承关系：

①若 $V_{i,p,k}=1$，则对所有的 $i'>i$ 有 $V_{i,p,k}=1$；

②若 $V_{i,p,k}=0$，则对所有的 $i'<i$ 有 $V_{i,p,k}=0$。

其中，$V_{i,p,k}$ 表示位点 p 在子克隆 S_i 的第 k 条单体型所呈现的等位基因变异频率。

2　分治拼接算法

本文所提算法的设计思想是：根据测序技术原理，一对双末端读段应来自同一条单体型，据此拼接单体型可分阶段进行。首先，对等位基因变异频率进行聚类，由此估计出子克隆数目并初始化每个子克隆的单体型；然后，根据等位基因变异频率将双末

端读段拼接成短链；接着，基于前述原则合并短链，链接短链组并根据比对信息验证其唯一性；最后，构建每个子克隆的单体型，再估算子克隆的配比。

2.1 预估子克隆单体型的先验信息

为了高效、准确地构建单体型，首先对等位基因变异频率进行聚类，解出子克隆个数的一个先验，作为构建单体型的先验信息。聚类方法采用 SciClone 方法，也是目前十分常用的一种子克隆个数估算方法，其设计思想是：假定等位基因变异频率服从贝塔分布、二项分布和高斯分布的混合分布，构建变分贝叶斯混合模型时要运用 Dirichlet 过程推断出较优的子克隆个数的解。需要说明的是，尽管聚类结果存在错误，但是作为先验信息已经具备提高拼接效率的功用。通过后续步骤，聚类错误大多被校正。

基于聚类结果，实施以下两个初始化步骤，分别初始化每个子克隆的单体型和每个子克隆的位点状态标志。令肿瘤样本包含 $I+1$ 个子克隆（包括基础克隆），M 的基数为 $|M|=N$。仅考虑二倍体，不考虑拷贝数变异，则子克隆的单体型共有 $2I+2$ 条。初始化每个子克隆的每条单体型，即对子克隆 S_i 设置两个数组 $c[2i]$ 和 $c[2i+1]$，并令 c 数组中 $2i$ 和 $2i+1$ 分别表示子克隆 S_i 的父链和母链单体型。在位点 p 处，数组有

$$\left.\begin{aligned} c[2i][p]&=A\\ c[2i+1][p]&=B \end{aligned}\right\}\quad (i\leqslant I, p\in M) \tag{1}$$

式中：A 表示位点的单体型与参考基因序列相同；B 表示位点的单体型与参考基因序列不同，即单体型在位点 p 处携带突变。

令 $f[2i][p]$ 和 $f[2i+1][p]$ 分别表示子克隆 S_i 的父链和母链单体型在位点 p 处的识别状态，若 $f[2i][p]$ 和 $f[2i+1][p]$ 均为 1，则子克隆 S_i 在位点 p 处已经完成单体型识别；否则，子克隆 S_i 在位点 p 处仍未获得收敛的单体型估计结果，其中 $i\leqslant I, p\in M$。

2.2 分治处理双末端读段的突变信息

当一对从双末端读段至少携带了 2 个突变时，可以从中提取出这些位点的单体型关系信息，用 R_E 表示。令这些关系的集合为 R，根据位点在参考基因组中的坐标对 R_E 排序后转化为 R 中的相对位置，删去那些相对位置重复的 R_E。然后，按照如下方法更新一遍数据 c 和 f：①将初始位点相同的 R_E 组成一组短链，具体来说，是一个整数规划问题，即规划的目标是使得链接得到的一组短链的基数（组中短链的个数）最小，而规划的约束是短链组必须能够支持所覆盖的所有的 R_E，这需考虑使用贪婪策略求解；②对于一条新的 R_E，若短链组不包含该 R_E，则将其纳入短链组中，否则暂时忽略该 R_E，保持短链组不变。这样设计的原因是，单体型拼接问题难以确定最优子结构，倘若盲目将一条新的 R_E 纳入短链组，可能会诱发假阳性错误，此类错误难以识别和消除，影响单体型构建精度。

在生成初始短链组后，给出短链组的全排列，再根据第 1 节原则分支限界。对于任意的位点 p，将起始于位点 p 处的 R_E 分为一组，逐组处理，直至所有的 R_E 都被遍历。

2.3 链接短链组得出子克隆单体型

构建子克隆单体型需要进一步链接短链组。设短链组的总数为 G，按照 R 中的相对位置对短链组排序，即 $1\leqslant g-1\leqslant g\leqslant g+1\leqslant G$，然后依次链接。当链接短链组 g 时，根据相对位置对短链组 g 包含的排列逐个检查，依据第 1 节原则，判断其是否可以与前 $g-1$ 个短链组的最优子链进行链接。

当可行解不唯一时，采用如下算法处理：①若存在至少两种链接方式，且单体型存在冲突，则令冲突位点的 $f[2i]$ 和 $f[2i+1]$ 置 0，表明需要更多的 R_E 参与校正；②令 $H_k^{g-1,g}$ 表示短链组 g 与短链组 1 至 $g-1$ 既成链的第 k 种链接组合，为了降低算法的时间复杂度，可以截取短链组 1 至 $g-1$ 既成链的一个子链，用于后续分析；③针对 f 值为 0 的位点，从现有的文库中选择一个新的文库，根据新文库的 R_E 判断 $H_k^{g-1,g}$ 的正确性，对既有子链进行微调；④在对既有子链进行微调时，需要对调整结果进行验证，即如果存在一个 R_E 支持 $H_k^{g-1,g}$，且存在其他的 R_E 否定 $H_{k'\neq k}^{g-1,g}$，那么更新链至短链组 R_E，更新对应的数组 $f[2i]$ 和 $f[2i+1]$。

2.4 估算各子克隆的混合比例

在构建出各子克隆的单体型之后，根据链和各变异位点的变异频率迭代计算出各子克隆在肿瘤样本中的混合比例，即

$$O_i=\sum_{p=1}^{N}V_pT(i,p)/\sum_{p=1}^{N}T(i,p) \tag{2}$$

式中：O_i 表示子克隆 S_i 的初始 V_p 值。若 $p\in M_i$ 且 $p\notin M_{i-1}$，则 $T(i,p)=1$，否则，$T(i,p)=0$。令

$$T'(i,p)=\theta T(i,p) \tag{3}$$

若 $|V_p-O_i|>\varepsilon$，则 $\theta=0$；否则，$\theta=1$。继而更新 O_i，当 O_i' 与 O_i 的差小于 ε 时，默认迭代收敛，否则，继续迭代更新 O_i。

利用

$$O_i = O'_i / \sum_{i=0}^{I-1} O'_i \tag{4}$$

可对 O_i 做归一化处理。用 α_i 表示子克隆 S_i 的配比

$$\alpha_i = \begin{cases} 2(O_i - O_{i+1}), & 0 \leqslant i \leqslant I-1 \\ 2O_i, & i = I \end{cases} \tag{5}$$

3　仿真和对比实验

将本文算法称为 MulPE 方法，下面在不同参数配置下通过仿真实验检验 MulPE 方法的性能。从人类参考基因组中随机采样连续的 10^6 个碱基作为参考序列，仿真单点变异，令血系变异发生率为 0.1%，体细胞突变率是 1%。考虑到肿瘤样本的纯度问题，令正常细胞占比为 30%，同时令肿瘤仿真数据包含 2 个子克隆，子克隆 S_1 和子克隆 S_2 的占比比例为 5∶2。

3.1　基础文库覆盖度的影响分析

在选择文库和覆盖度（测序得到的碱基数量与待测基因组中碱基数量的比值）时，一般先确定基础文库和其覆盖度，从中提取 V_E 信息。在此，将用于计算等位基因变异频率所用的文库称作基础文库，图 1 为 MulPE 方法中基础文库覆盖度对子克隆单体型识别准确率的影响分析，其充分表明了 MulPE 方法在覆盖度大于 50 时具有较好的鲁棒性。

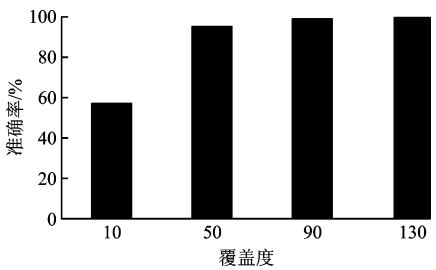

图 1　MulPE 方法中基础文库覆盖度对单体型识别准确率的影响

如图 1 所示，单体型识别的准确率随覆盖度加大而提高。当覆盖度低于 50 时，由于各子克隆采样不均衡，使得变异频率的统计信息出现偏差而导致聚类不准确，影响单体型的识别精度。当覆盖度达到 100 时，随着聚类准确性的提高，准确率可以达到 99%以上。目前，肿瘤基因组全外显子组测序的标准深度为 100，美国肿瘤基因组路线图计划的全外显子组测序数据的平均覆盖度约为 167.5，这样的实验设计已经被证明能够将低频变异位点很好地检测出来[9]。因此，MulPE 方法需要 50 倍覆盖度的要求易被满足。

3.2　新增文库覆盖度的影响分析

当基础文库确定后，新增多个文库，直至所有的变异位点都被包含。在新增文库时，文库长度默认以 500 个碱基对递增。新增文库的覆盖度对子克隆单体型识别准确率的影响分析如图 2 所示。当新增文库的覆盖度从 5 到 50 变化时，子克隆单体型识别准确率的变化甚微（在 1%以内）。由此可见，子克隆单体型识别的准确率主要取决于基础文库的覆盖度，而受新增文库覆盖度的影响较小。

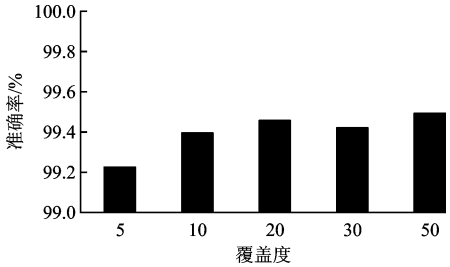

图 2　MulPE 方法中新增文库的覆盖度对单体型识别准确率的影响

3.3　与其他工具对比

CITUP 和 MulPE 两种方法都可实现子克隆结构与比例识别，但目标略有差异。CITUP 侧重于子克隆基因型识别，比较时要将 CITUP 得出的子克隆基因型随机拆分为单体型[10]。在此，设置基础文库的长度为 1 000 个碱基对，其余 2 个文库长度的均值分别为 1 500 个碱基对和 2 000 个碱基对，标准差均为 30 个碱基对。如表 1 所示，MulPE 在估算子克隆混合比例上的精度远高于 CITUP。

表 1　变异率为 1%时 MulPE 与 CITUP 方法比对结果

方法	覆盖度	单体型正确率/%	正常细胞占比/%	S_1 占比/%	S_2 占比/%
MulPE	60	98.41	30.07	48.85	21.09
CITUP		69.26	35.41	26.49	38.10
MulPE	70	98.20	29.40	50.40	20.21
CITUP		68.30	33.55	35.32	31.14
MulPE	80	99.30	30.55	50.44	19.02
CITUP		69.11	31.68	39.74	35.00
MulPE	90	99.44	30.77	50.09	19.15
CITUP		69.88	32.53	32.47	35.00
MulPE	100	99.45	30.09	50.28	19.63
CITUP		68.86	31.15	40.423	28.42
预设值			30	50	20

3.4 不同聚类方法在低覆盖度下的结果分析

子克隆单体型的先验依赖于 SciClone 的结果，而在低覆盖度时聚类结果偏差较大，在此分别采用不同聚类方法对低覆盖度下的效果进行分析，如图 3 所示。其中，聚类方法选取了基于密度的聚类方法 DBSCAN、基于层次的聚类方法 AGNES、混合高斯聚类方法 GMM 和 SciClone 共 4 种。由分析可见，各种聚类方法的精度主要依赖于等位基因变异频率的文库覆盖度及测序偏差，在低覆盖度下的聚类结果均无法达到理想值。MulPE 方法提出基础文库最低覆盖度需满足 50 的要求方可使每个变异位点都能够被有效采样，为子克隆单体型的精确构建提供了必要条件。

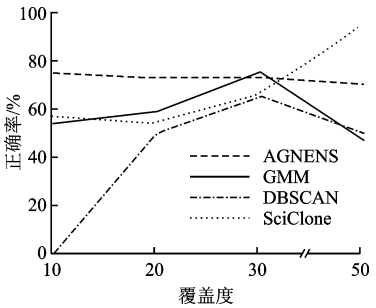

图 3 低覆盖度下不同聚类方法的结果分析

4 结 论

本文所提出的 MulPE 方法在一定覆盖度下，可以较为精确地构建子克隆单体型，其实现结果具有一定的实用价值，可以更为清晰地刻画癌克隆进展情况，为个性化的癌症诊疗方案、追踪较强的子克隆并清除等相关癌症研究提供理论依据。

由于肿瘤异质性的存在，使得子克隆单体型识别难度较大。如果直接使用组装方法构建子克隆，由于双末端变异信息容纳的变异点数目和构成的短链长度相对于子克隆单体型少而短，使得组装过程中的不确定性太大，无法得出最正确而合适的组装结果。从时间复杂度上分析，若肿瘤组织中包含 n 个变异位点，每个变异位点有两种等位基因，则有 2^n 条单体型链，故完全采用贪婪算法所需时间复杂度为 $O(2^n)$。MulPE 方法采用分治法处理双末端变异信息，恰当定义分治长度和连接方法，使得分治后得到的短链可以准确反映子克隆单体型的片段。在分治法中，首先将双末端变异信息根据其起始位置分成 G 组进行处理，假设已知子克隆数目为 I，则单体型链的数目是确定的，即为 $2I$，故全排列算法的时间复杂度为 $O((2I)!)$，所以使用分治法所用的时间复杂度是 $O(G(2I)!)$，远小于贪婪法的指数级时间复杂度。

虽然 MulPE 方法在子克隆单体型的识别较为准确，但在很多方面仍需进一步研究。如：解决基础文库低覆盖度时聚类冗余的问题；当子克隆微进化模式较为复杂时对应的单体型识别问题；综合考虑多种变异事件如何更加精准地构建子克隆单体型等。

参考文献：

[1] 涂超峰，綦鹏，李夏雨. 肿瘤异质性：精准医学需破解的难题 [J]. 生物化学与生物物理进展，2015(10)：881-890.

[2] GREAVES M, MALEY C C. Clonal evolution in cancer [J]. Nature, 2012, 481(7381): 306-313.

[3] SCHWARTZ R, SHACKNEY S E. Applying unmixing to gene expression data for tumor phylogeny inference [J]. BMC Bioinformatics, 2010, 11(1): 1-20.

[4] FRANCESCO S, FARISI P, MARIANN M. TrAp: a tree approach for fingerprinting subclonal tumor composition [J]. Nucleic Acids Research, 2013, 41(17): e165.

[5] WEI Jiao, VEMBU S, DESHWAR A G. Inferring clonal evolution of tumors from single nucleotide somatic mutations [J]. BMC Bioinformatics, 2014, 15(1): 1-16.

[6] ROTH A, KHATTRA J, YAP D. PyClone: statistical inference of clonal population structure in cancer [J]. Nature Methods, 2014, 11(4): 396-398.

[7] MILLER C A, WHITE B S, DEES N D. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution [J]. Plos Computational Biology, 2014, 10(8): e1003665.

[8] EL-KEBIR M, OESPER L, ACHESON-FIELD H. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data [J]. Bioinformatics, 2015, 31(12): 62-70.

[9] NETWORK C G A. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia [J]. New England Journal of Medicine, 2013, 368(22): 2059-2074.

[10] MALIKIC S, MCPHERSON A W, DONMEZ N. Clonality inference in multiple tumor samples using phylogeny [J]. Bioinformatics, 2015, 31(9): 1349-1356.

（编辑 武红江）