ISSN 1007-7626
CN 11-3870/Q
中国生物化学与分子生物学报 http://cjbmb.bjmu.edu.cn
Chinese Journal of Biochemistry and Molecular Biology
2017 年 11 月
33(11):1168~1174

DOI: 10.13865/j.cnki.cjbmb.2017.11.12

基于混合泊松分布的新生突变识别算法

高迎心[1)], 温佳威[2)], 徐 尔[1)], 艾冬梅[1)*]

([1)]北京科技大学 数理学院 信息与计算科学系,北京 100083;[2)]河南偃师高级中学,河南 洛阳 471900)

摘要 对个体而言,不经父母遗传而后天获得的突变称为新生突变,绝大多数癌症都起自新生突变。构建快速精确的变异识别算法将有助于对癌症的研究。然而,针对前期新生突变识别算法准确率不高,且耗时多等问题,本文引入了基于变异位点的先验概率分布模型,运用基于混合泊松分布的期望最大化(EM)算法对新生突变识别算法进行改进与优化,研究了有亲缘关系的新生突变的识别,并在识别精度与运算速度方面与已有算法进行对比。结果表明,基于混合泊松分布的期望最大化算法在提高运算速度的同时降低了假阳性比率,具有良好的识别效果。

关键词 人类基因组; 新生突变; 混合泊松分布; 遗传疾病

中图分类号 TP301.6

Recognition of *de Novo* Mutations Based on Hybrid Poisson Distribution

GAO Ying-Xin[1)], WEN Jia-Wei[2)], XU Er[1)], AI Dong-Mei[1)*]

([1)] *Department of Information and Computing Science, School of Mathematics and Physics, University of Science and Technology Beijing, Beijing* 100083, *China*; [2)] *Yanshi Senior High School of Henan Province, Luoyang* 471900, *Henan, China*)

Abstract For the individual, gene mutations that are acquired without parental inheritance are the origins of vast majority of cancers. Application of fast and accurate recognition algorithms will be a great help to the study of cancer. Aiming at the problem of poor accuracy and time consumption, a prior probability model of mutation sites was introduced. To modify and optimize the recognition algorithm, the Expectation Maximum (EM) algorithm based on mixed Poisson distribution was used to identify the *de novo* mutation involving kinship data and compare with the existing algorithms in recognition accuracy and computing speed. The results show that the EM algorithm based on mixed Poisson distribution can improve the speed of operation and reduce the false positive ratio, which is of great significance for the recognition of cancer.

Key words human genome; *de novo* mutation; hybrid Poisson distribution; genetic disease

在全基因组水平上,与人类疾病相关的单核苷酸变异(single nucleotide variants, SNVs)、插入缺失(insertion-deletion, InDel)和结构变异(structural variation, SV)等多种突变信息,已经得到大量的检测[1-3]。其中,单核苷酸变异出现频率高并且能较稳定遗传,是人类可遗传变异中最常见的一种。人体的表现型、疾病的易感性以及抗药性等的差异都可能与其有关[4, 5]。大部分的单核苷酸变异,不会导致生物体性状发生明显的改变。若某一核苷酸位点的变异频率大于1%,则称在该位点发生了突变,会引起一定的表型变化[6, 7]。未经父母遗传而后天获得的基因突变,称为新生突变(*de novo* mutation)。新生突变只在后代个体中出现,通常会对表型产生更大的影响,并且容易引发许多复杂疾病[8-10]。研究表明,大多数罕见疾病都是由新生突变引起[11]。如果体细胞的某些特定基因发生新生突变,则该体细胞的后代就有更大的可能性发生癌变[12, 13]。因此,新生突变成为研究癌症等复杂疾病发病机制的有效切入点[14-16]。目前,通过检测三体家系(trio-

收稿日期: 2017-08-06; 修回日期: 2017-09-15; 接受日期: 2017-09-25
国家自然科学基金(No.61370131)资助

* 通讯作者 Tel: 010-62332349; E-mail: aidongmei@ustb.edu.cn

Received: August 6, 2017; Revised: September 15, 2017; Accepted: September 26, 2017
Supported by National Natural Science Foundation of China (No. 61370131)

* Corresponding author Tel: 010-62332349;
E-mail: aidongmei@ustb.edu.cn

family)数据来发现新生突变,并探索新生突变与复杂疾病之间的关系,已成为人类基因组学研究中的热点问题。

新生突变识别方法有以下几种:GATK[17]和 Samtools[18],通过比较先验者与三体家系中亲本基因型来推测子代是否有新生突变;更有效的方法比如 DNMFilter[19],通过使损失函数在梯度方向上下降,从而不断优化序列特征分类模型来识别新生突变;mirTrios[20],经预先设定的标准测序质量值和测序深度值等参数过滤,得到新生突变;Triodenovo[21],引入贝叶斯模型,克服对预先设定值的过度依赖。但下一代测序数据维度高且噪声大,这些算法也面临精度不够、耗时过多的问题。

为了解决上述问题,本文结合变异位点的先验概率分布模型,提出了基于混合泊松分布的期望最大化算法(expectation maximum algorithm, EM)。首先利用先验信息对变异位点进行确定,从而根据变异特征缩小搜索范围,降低时间消耗。在此基础上,运用基于混合泊松分布的 EM 算法得到最优迭代过程,将变异属性带入此过程中,并通过设定阈值识别新生突变。

1　材料与方法

1.1　新生突变识别模型的建立

遗传变异过程中的新生突变与人类癌症、神经发育系统疾病密切相关。研究这些突变的发生机制、变异位点与突变率将有利于对复杂疾病的探索。如何将遗传因素与环境因素等与变异相关的先验信息融入先验概率统计模型,评估先验因素对变异发生的影响程度,很大程度上决定了此变异识别算法的准确度。目前,影响变异检测的先验因素复杂繁多且具有不确定性,先前的算法没有考虑这些因素。因此,本文将利用统计检验的方法,对先验信息进行选择归类,以确定最优化的先验信息特征集合。

1.2　基因变异先验因素变量的选择与聚类

随着大量生物基因数据的不断涌现,实际应用过程中所涉及到数据的特征维数逐渐增高,运用特征选择对数据进行降维的算法进一步发展,使得已选特征包含的类别信息尽可能多,同时使得特征子集内部的冗余程度尽可能小。

遗传或环境因素都会引起单核苷酸变异,造成表型的差异。化学诱变、温度和湿度等环境因素和基因组成等遗传因素都是造成基因组变异的先验因素。这些先验因素众多繁杂,因此必须充分利用这些先验信息,构造先验统计概率推断模型,按照对变异产生的贡献大小对这些因素进行选择聚类。

设 $X_i, Y_i, (i = 1, \cdots, n)$ 为独立的一组值,(X_i 表示引起基因突变第 i 种因素的观测值,Y_i 为 2 值变量 0 或 1,发生了基因突变为 1 否则为 0。指数族分布规范型表示如下:

$$f(Y_i \mid \theta_i) \propto \exp[Y_i\theta_i - b(\theta_i)] \tag{1}$$

其中 $b(\theta_i)$ 为对数配分函数,与 X_i 的分布有关。当 X_i 服从泊松分布时,$b(\theta_i) = \lambda$,$\theta_i = Z_i^T\beta$,$Z_i = (1, X_i^T)^T$,$\beta = (\beta_0, \beta_1, \cdots, \beta_p)^T$,$\beta$ 为权重系数向量,即先验因素对变异产生的贡献大小,β 过小将被视为先验因素影响作用微弱,需要去除。变量选择标准函数为

$$L_\lambda(\beta; X_i, Y_i) = L(\beta; X, Y) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

设 S^p 为 SNPs 先验信息的有限集,将 S^p 作为感兴趣的变量集,剩余的记为噪声参数集,为去除噪声干扰,拟对噪声参数集进行惩罚,并构建如下噪声惩罚回归(nuisance penalized regression, NPR)模型得到 $\hat{\beta}$:

$$L_{\kappa, S^p}(\beta; X_i, Y_i) = \underbrace{-\frac{1}{n}\sum_{i=1}^{n}\overbrace{[Y_i Z_i^T\beta - b(Z_i^\beta\beta)]}^{L(\beta)}}_{L_n(\beta)} + \kappa\sum_{j=1}^{p}|\beta_j| I(X_j \notin S^p) \tag{3}$$

设 $H = E[-\nabla_{\beta\beta} L(\beta)]$,且假设 $\beta_{S^p} \perp \beta_{(S^p)^c}$,即 $E[-\nabla_{\beta_{S^p}\beta_{(S^p)^c}} L(\beta)] = 0$。记 $\| n^{-1}\nabla_{\beta\beta} L_n(\hat{\beta}) - H \|_{\max} = O_P(\eta_n)$。

利用 KKT(Karush-Kuhn-Tucker)条件和一阶泰勒展开,对上述模型进行修正,得到校正变量选择标准函数如下:

$$L_{\lambda,\eta}(\beta; X_i, Y_i, \hat{Y}_i^p) = L(\beta; X_i, Y_i) + \lambda\sum_{j=1}^{p}|\beta_j| + \eta L(\beta; X_i, \hat{Y}_i^p) \tag{4}$$

其中 $\hat{Y}_i^p = (\hat{y}_1^p, \cdots, \hat{y}_n^p)$ 用来平衡实际数据与先验信息,是事先臆测的向量。

$\eta = 0$:校正变量选择标准函数退化为依据先验信息的假设检验;

$\eta \to \infty$:校正变量选择标准函数完全依赖事先臆测的信息。

得到系数向量估计值 $\hat{\beta}_{S^p}$ 后,即完成了变量选择的过程。随后,按照各个先验因素变量之间的相似度关系进行变量聚类。

实际操作中为避免遗漏重要因素,人们会选用尽量多的相关因素对某一指标进行衡量,按照变量

之间的相关关系,将他们聚合成为不同的类别,经常采用相似性来衡量变量之间的亲疏关系,使得同一类中的数据具有相同或相似的主要特征。本文选用的变量相似系数计算如下:

设变量 u_i 和 u_j 的取值分别为 $x_1, x_2, \cdots, x_p$ 和 $y_1, y_2, \cdots, y_q$。n_{pq} 表示 u_i 取 x_p 和 u_j 取 y_q 的样本数。

则相关系数表示为:$r_{ij} = \sqrt{\frac{\chi^2}{n}}$,式中

$$\chi^2 = n\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}^2 - n_{i.}n_{.j}}{n_{i.}n_{.j}} \quad (5)$$

$|r_{ij}|$ 越接近 1,表示两变量相关程度越高,将相关程度高的变量聚为一类。以此来减少变量类数,提高运算速度。

1.3 基于混合泊松分布的 EM 算法模型

由于泊松分布是描述单位时间内随机事件出现的次数,符合基因组变异随机发生的情境,该方法实现了由静态模拟向动态模拟的转变,抽样分布随着模拟进程而随时调整改变,有利于提升基因型估计的准确性。在测序、比对和变异检测中,会产生较多误差。而仅仅通过测序深度、基因型和质量值等基本参数,不足以降低假阳性比率。因此,将 EM 算法应用到变异属性筛选的过程中,通过简单的迭代算法来计算密度函数,转化为参数估计问题。为进一步减少数据文件中的假阳性变异,采用基于泊松分布的最大期望算法确定最优迭代过程,将上一步筛选的参数带入最优迭代过程,通过设定阈值,识别出新生突变及罕见序列变异。

假设随机变量 $x_j = \{x_1, x_2 \cdots x_n\}$ 是来自由 m 个泊松分布总体 $G_1, G_2 \cdots G_m$ 且分别以 $\pi_1, \pi_2 \cdots \pi_m$ 为权重混合而成的分布 G。其和为 1,则 m 阶混合泊松分布的概率密度函数 $f(x|\lambda)$ 就可以表示为:

$$f(x_j|\lambda) = \pi_1 f(x_j|\lambda_1) + \pi_2 f(x_j|\lambda_2) + \cdots + \pi_m f(x_j|\lambda_m) \quad (6)$$

其中,$f(x_j|\lambda_i) = \frac{\lambda_i^{x_j}}{x_j}e^{-\lambda_i} i = 1, 2 \cdots m$ 为相应总体 G_i 的密度函数,λ_i 为未知参数,整个总体参数 θ 由 λ_i 和 π_i 组成,令 $\psi = (\pi_1, \pi_2 \cdots \pi_{m-1}, \lambda_1, \lambda_2 \cdots \lambda_m)^T$。

在 EM 框架下,每个 x_j 被认为来自混合模型(式(6))的 m 个分量中的 1 个。用 $z = \{z_1, z_2 \cdots z_n\}$ 表示不可观测的分量的指示向量。其中

$$z_{ij} = \begin{cases} 1 & x_j \text{ 是来自第 i 个分量} \\ 0 & x_j \text{ 不是来自第 } i \text{ 个分量} \end{cases} \quad (7)$$

用 x_j 表示观测数据向量,z 表示缺失数据向量,$x = (x_j^T, z^T)^T$ 表示完整数据向量。在有限混合泊松分布模型中,基于参数 ψ 的完整数据对数似然函数为:

$$\log L(\Psi) = \sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij}\{\log\pi_i + \log f(x_j, \lambda_i)\} \quad (8)$$

在 EM 算法的第 $k+1$ 次迭代中,E-step 计算函数 $Q(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}}\{\log L(\Psi)|x\}$ (9)

$$E_{\Psi^{(k)}}(z_{ij}|x) = \tau(x_j; \Psi^{(k)}) = \frac{\pi_i^{(k)} f(x_j, \lambda_i^{(k)})}{\sum_{h=1}^{m}\pi_h^{(k)} f(x_j, \lambda_h^{(k)})} \quad (10)$$

由式(8)和(10)可以得到

$$Q(\Psi; \Psi^{(k)}) = \sum_{i=1}^{m}\sum_{j=1}^{n}\tau(x_j; \Psi^{(k)})\{\log\pi_i + \log f(x_j, \lambda_i)\} \quad (11)$$

M-step 中参数估计的更新公式为:

$$\pi_i^{(k+1)} = \frac{\sum_{j=1}^{n}\tau_i(x_j; \Psi^{(k)})}{n}(i = 1, 2 \cdots m) \quad (12)$$

$$\theta_i^{(k+1)} = \frac{\sum_{j=1}^{n}\tau_i(x_j; \Psi^{(k)})x_j}{n\pi_i^{(k+1)}}(i = 1, 2 \cdots m) \quad (13)$$

如此循环执行 E-step 与 M-step,直到 $L(\Psi^{(k+1)})$ 与 $L(\Psi^{(k)})$ 的差值小于给定的阈值时停止迭代。如此得到 $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2 \cdots \hat{\pi}_m), \hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2 \cdots \hat{\lambda}_m)$。

利用 32 个子代患有自闭症谱系障碍(autism spectrum disorder, ASD)三体家系数据生成 EM 算法中,与每一个新生突变相关属性的初始值,如 n, π_j, λ_i 等。VCF 文件中,与新生突变相关的性质有 QUAL——比对质量值、Depth——测序深度、QD——变异置信度、MQ0——映射质量值为 0 的读片数量、PL——父母或子代基因型的最大 Phred-scaled 值、PRT——子代读片在参考基因组上的最大覆盖率、PART——亲代读片在参考基因组上的最大覆盖率等。将变异属性代入由上述 EM 算法产生的最优迭代过程,并利用获得的阈值进一步筛选出潜在的新生突变。在混合泊松分布下,EM 算法的模型用如下流程表示。

初始化,确定模型分量个数 m,设置参数初始值 ψ。

1. 计算混合泊松分布中 X 的期望值
2. 重新估计分布参数,使得 X 的似然函数最大,给出 X 的期望估计

重复上述迭代过程直至期望值收敛,本次的参

数估计值即为最终估计值。

(**Procedure** Expectation maximization algorithm model based on Mixture Poisson distribution

Require the parameter λ for iterative process, the total number of Poisson distributions m, the weight of Poisson distribution π, the sample set.

Ensure the iterative process

Repeat

1. Calculate the likelihood function of variant X in mixture Poisson Z

2. Calculate the mathematical expectation of conditional distribution

3. Minimize the likelihood function to obtain new parameters

Until the parameter λ tends to be stable

Return the iterative process λ)

将每 1 个三体家系数据带入上述算法流程,识别新生突变,输出结果包括变异所在的染色体及外显子或内含子区域、起始位点、基因名称、变异类型及是同义或非同义替换等信息。

1.4　数据来源

本文选取来自于千人基因组计划(1 000 Genomes Project)第三阶段的测序数据。数据库中的 26 个群体分布在 5 个区域,从中随机选取居住在美国犹他州的北欧人和西欧人 CEU(Utah residents (CEPH) with Northern and Western European ancestry)作为欧洲代表,尼日利亚依巴丹区的约鲁巴人 YRI(Yoruba in Ibadan, Nigeria)作为非洲人代表,居住在北京的中国人 CHB(Han Chinese in Beijing, China)作为亚洲代表,居住在洛杉矶的墨西哥人 MXL(Mexican Ancestry in Los Angeles, California)和居住在波多黎各的波多黎各人 PUR(Puerto Rican in Puerto Rico)作为美洲代表,居住在胡志明市的越南人 KHV(Kinh in Ho Chi Minh City, Vietnam)作为东南亚代表,每个人群选取 1 至 2 组三体家系(trio-family)数据,将 UCSC 数据库中的参考基因组序列 hg19 通过 BWA 进行双端序列比对,处理后得到用于描述 SNVs 与 Indels 的 VCF(variant calling file)文件。本文选取了 CEU、YRI、CHB、PUR 与 MXL 的三体家系数据对基因组变异进行分析。将个体的读片数据分别与参考基因组 hg19 进行比对处理,得到专门用于描述 SNVs 与 Indels 的 VCF(ariant calling file)。文件所包含变异属性及其含义如下 Table 1 所示:

Table 1　VCF file format for genomic variation of trio-families

Variation characteristic	Meaning
CHROM	Reference sequence, refers to the chromosome number in the human genome
POS	Variation position, corresponding to the position on the reference genome
ID	The mutation's ID number in dbSNP database, equaling to the rs number in dbSNP
REF and ALT	The base of reference genome and genome alignment in the mutation sites
QUAL	The possibility of variation at this site
FILTER	To further filter the variation
GT	The genotype of the sample, for diploid organisms, GT represents the sample carries two alleles at that site
AD	The coverage of allele in samples
DP	The total number of reads covering a particular site, referring to the depth of the site
GQ	The quality value of the most likely genotype
PL	PL corresponds to three values, showing the likelihood value of three different genotype, that is 0/0, 0/1 and 1/1.

2　结果

2.1　计算结果

本文选取的样本数据号如 Table 2 所示。

传统的变异检测,依赖于研究对象序列与参考序列的比对。而基于三体家系数据的变异检测,则充分考虑样本之间的相互关系,以包括父、母、后代在内的核心家庭为单位,利用高通量测序技术在全基因组范围内进行变异识别。将上表所示的 8 组核心家庭的 8 个子代(NA12878、NA12882、NA19238、NA18485、HG00512、NA19675、HG02024、HG00733)在不考虑家庭结构关系的情况下,进行单独的变异检测。然后,与基于三体家系整体结构进行变异识别的结果进行对比(Fig. 1)。结果显示,对于每 1 组

三体家系数据,对子代进行变异检测,不考虑家庭结构因素影响的条件下,能检测出更多的变异,即发现了新生突变的存在。

Table 2 ID number of trio-families sample data

Race	ID number of trio-family
CEU	NA12878, NA12891, NA12892
CEU	NA12882, NA12877, NA12878
YRI	NA19238, NA19239, NA19240
YRI	NA18485, NA18487, NA18489
CHB	HG00512, HG00513, HG00514
MXL	NA19675, NA19678, NA19679
KHV	HG02024, HG02025, HG02026
PUR	HG00733, HG00731, HG00732

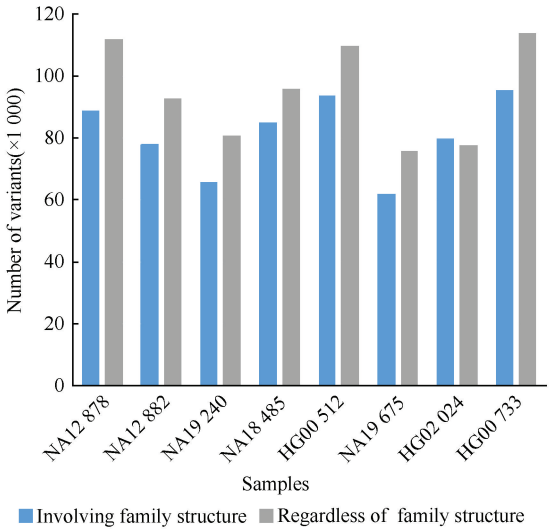

Fig. 1 Comparison of sample variation test results
Blue represented the variation numbers involving family structure. Gray represented the number of variations regardless of family structure

单核苷酸碱基的变异有两类,分别是转换(transition)和颠换(transversion)。前者指的是嘌呤被嘌呤取代或嘧啶被嘧啶取代,后者指的是嘌呤与嘧啶之间进行交换。碱基的插入或缺失也是导致单核苷酸变异的原因之一。但在一般情况下,SNP 是指碱基的转换与颠换,且通常是二等位多态性的。因此,本文针对转换(AT-GC,GC-AT)和颠换(AT-CG,AT-TA,GC-TA,GC-CG)这 6 种变异的发生比率进行比较(Fig. 2)。Fig. 2 的结果表明,转换的发生率明显高于颠换,且以 C 转换为 T 为主,转换型变异的 SNP 约占全部变异的 2/3,其

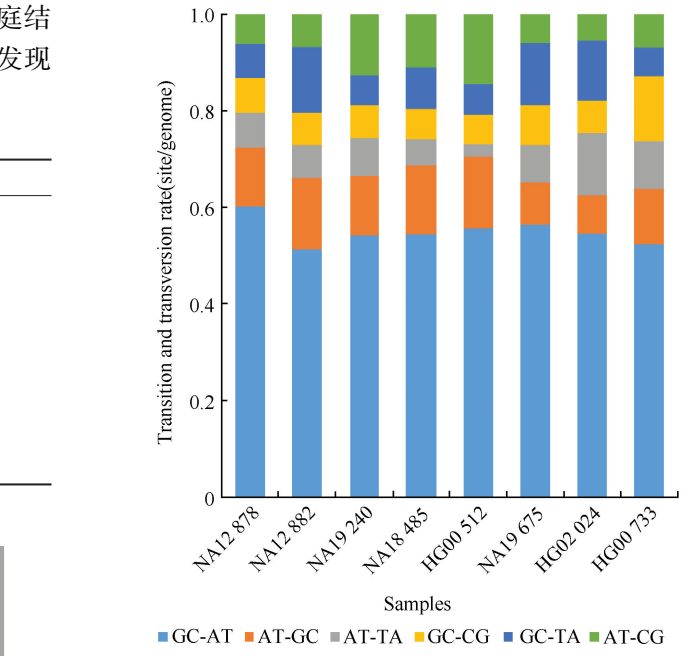

Fig. 2 Comparison of transition and transversion rate
Light blue represented AT in exchange of GC. Orange represented GC in exchange of AT. Gray represented AT in exchange of CG. Yellow represented AT in exchange of TA. Dark blue represented GC in exchange of TA. Green represented GC in exchange of CG

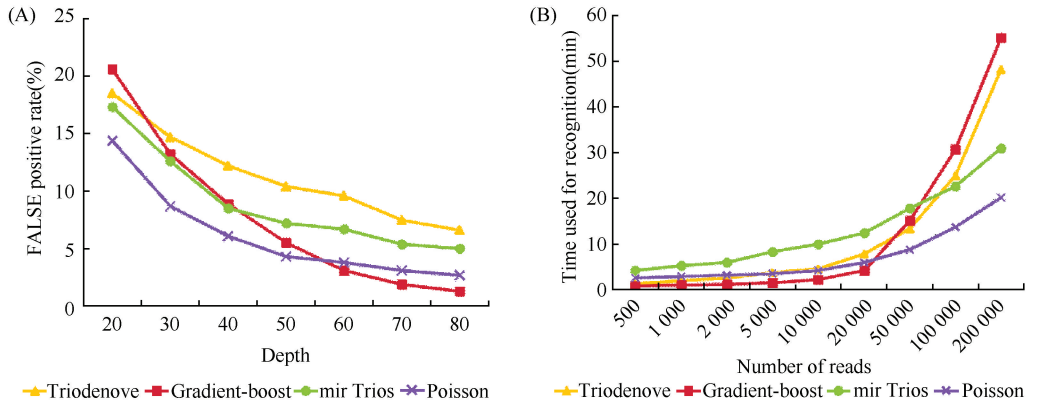

Fig. 3 Comparison of different recognition algorithms (A) Comparison of false positive rate of different recognition algorithms. (B) Comparison of operation speeds of different recognition algorithms. Purple represented mixed Poisson. Green represented mirTrios. Yellow represented Triodenovo. Red represented Gradient-boost

它几种变异的发生几率相近。SNP 中 CG 碱基的转换出现频率最高,这与此前的研究结果是一致的。胞嘧啶在人类基因组中最易发生突变是因为 CG 中的胞嘧啶 C 大多为甲基化的,可自发地脱去氨基而形成胸腺嘧啶。

2.2　分析结果

测序深度指的是测序获得的总碱基数量与基因组大小的比值,它与基因组覆盖度呈正相关关系。一般而言,测序引发的误差或假阳率会随着测序深度的增加而降低。本文将混合泊松算法与 Triodenovo、Gradient-boost 以及 mirTrios 变异识别算法的假阳性率进行了比较,假阳性率即错误识别的读片数占总读片数的百分比,随测序深度的变化不同算法假阳性率的变化情况见 Fig. 3 A。其中横坐标表示测序深度,纵坐标表示假阳性率。由 Fig. 3 A 可知,4 种算法的变异识别假阳性率均随测序深度的增加而下降。当测序深度大于 50 倍以后,假阳性率的变化趋于平缓。通过比较表明,本文中基于泊松分布的变异识别算法假阳性率最低,虽然基于梯度算法(Gradient-boost)的假阳性率在测序深度大于 50 倍时低于本文算法,但综合测序成本与运行时间,基于泊松分布的变异识别算法仍然具有较大优势。

当读片数量不断增多,运算量也迅速增大,导致变异检测所消耗的时间也会有所增加。算法运行时间随读片数量变化的情况如 Fig. 3B 所示。在 Fig. 3B 中,用横坐标表示读片数量,纵坐标表示变异识别耗时。不同变异识别算法的运算速度随着读片数量增加而加大,在读片数量小于 50 000 时,运行时间变化缓慢,读片数量超过 100 000 时运行时间急剧增加。本文算法由于加入了先验概率估计模型,在读片数量较少时耗时也较其他算法多,然而随着读片数量增多,运行时间增速缓慢,适合处理大量样本的变异识别。

单核苷酸多态性是由变异频率大于 1% 的单核苷酸变异引起。在人类基因组中,大约每 1 000 个碱基就会出现 1 个 SNP,尤其是对于那些既非来源于父亲,又非来源于母亲的新生突变具有特殊的研究意义。将 8 组三体家系数据与独立样本数据的变异检测结果进行比较,发现结合家庭结构的变异识别算法,能够检测出更多的突变。基因组上的单个核苷酸的变异,主要包括转换和颠换,且转换与颠换的比值约为 2∶1,其中以胞嘧啶转换为胸腺嘧啶为主。对于 20 倍、30 倍直至 60 倍等不同的测序深度,变异检测的假阳率不断下降。为平衡测序成本与精确度,在大多数的研究中,以 30 到 40 倍测序深度为宜。变异检测耗时随读片数量的增加而增加,当读片数量超过 50 000 时,耗时激增。

3　讨论

对于个体而言,不经父母遗传而后天获得的新生突变是许多单基因遗传病的主要病因,并且新生突变也参与了某些复杂疾病的发病过程。实际上,绝大多数癌症都起自新生突变。因此,相较于以往只对单个个体进行变异分析而言,利用基于父、母、子三代的三体家系对新生突变进行识别,对罕见疾病的诊疗具有重要意义,也是功能基因组学与医药领域的重要研究内容。研究表明,基于核心家庭进行人类全基因组数据分析,不仅将研究对象的基因组序列与参考序列进行比对,还将其与具有亲缘关系的个体序列进行比对,从而能够深层次发掘稀有罕见变异及新生突变,提高了基因型估计的准确性,有效降低了假阳性并降低了对于高测序深度产生的大量读片的处理与运算时间。

本文的统计算法可以被整合并拓展为人类全基因组数据生物分析的标准化工具。该流程将实现从结构变异识别到形成机制分析到疾病诊断分析为一体的流水线,极大地方便一般生物学家开展全基因组数据的分析工作。

参考文献(References)

[1] 邵谦之,姜毅,吴金雨. 全基因组测序及其在遗传性疾病研究及诊断中的应用[J]. 遗传(Shao QZ, Jiang Y, Wu JY. Whole-genome sequencing and its application in the research and diagnoses of genetic diseases[J]. Hereditas(Beijing), 2014,**36**(11):1087-1098

[2] 1000 Genomes Project Consortium, Auton A, Brooks L D, *et al*. A global reference for human genetic variation[J]. Nature, 2015, **526**(7571):68-74

[3] 熊昕昕,贺苗,陈晓钎. 孤独症研究新进展:新发突变及 CHD8 功能[J]. 生理科学进展(Xiong XX, He M, Chen XQ. Progress in Autism: de novo mutation and CHD8 functions[J]. Prog Physiol Sci, 2014, **45**(3): 185- 189

[4] Sudmant P H, Rausch T, Gardner E J, *et al*. An integrated map of structural variation in 2 504 human genomes [J]. Nature, 2015,**526**(7571):75-81

[5] Schwender H, Taub M A, Beaty T H, *et al*. Rapid testing of SNPs and gene-environment interactions in case-parent trio data based on exact analytic parameter estimation[J]. Biometrics, 2012,**68**(3):766-773

[6] Samocha K E, Robinson E B, Sanders S J, *et al*. A framework for the interpretation of de novo mutation in human disease[J]. Nat Genet,2014,**46**(9):944-950

[7] Barrick J E, Lenski R E. Genome dynamics during experimental evolution[J]. Nat Rev Genet, 2013,**14**(12): 827- 839

[8] 兰风华,王志红,柯龙凤,等. 新生突变与人类疾病[C]. 中

国遗传学会代表大会暨学术讨论会. 中国重庆;2008. (Lan F H, Wang Z H, Ke L F, *et al*. De novo mutations and human diseases: the eighth congress of genetics society of China [C], Chongqing, China, 2008

[9] Neale B M, Kou Y, Liu L, *et al*. Patterns and rates of exonic de novo mutations in autism spectrum disorders[J]. Nature,2012, **485**(7397):242-245

[10] 廖娟,周春燕,郭小艳,等. 一个多发性骨软骨瘤家系的 EXT1 基因突变分析[J]. 分子诊断与治疗杂志(Liao J, Zhou CY, Guo XY, *et al*. Mutational analysis of the EXT1 gene from a pedigree with hereditary multiple osteochondromas [J]. J Mol Diagn Ther), 2011, **3**(4):227-231

[11] Ku C S, Tan E K, Cooper D N. From the periphery to centre stage: de novo single nucleotide variants play a key role in human genetic disease[J]. J Med Genet,2013,**50**(4):203-211.

[12] Veltman J A, Brunner H G. De novo mutations in human genetic disease[J]. Nat Rev Genet,2012,**13**(8): 565- 575

[13] Ware J S, Samocha K E, Homsy J, *et al*. Interpreting de novo variation in human disease using denovolyzeR [J]. Curr Protoc Hum Genet,2015,**87**:7. 25. 1-15

[14] Ku C S, Vasiliou V, Cooper D N. A new era in the discovery of de novo mutations underlying human genetic disease[J]. Hum Genomics,2012,6:27

[15] Al-Aama J Y, Al-Ghamdi S, Bdier A Y, *et al*. De novo mutation in the KCNQ1 gene causal to Jervell and Lange-Nielsen syndrome [J]. Clin Genet,2014,**86**(5):492-495

[16] Lee JR, Srour M, Kim D, *et al*. De novo mutations in the motor domain of KIF1A cause cognitive impairment, spastic paraparesis, axonal neuropathy, and cerebellar atrophy [J]. Hum Mutat,2015,**36**(1):69-78

[17] DePristo M A, Banks E, Poplin R, *et al*. A framework for variation discovery and genotyping using next- generation DNA sequencing data[J]. Nat Genet,2011,**43**(5):491-498

[18] Li H, Handsaker B, Wysoker A, *et al*. The Sequence Alignment/Map format and SAMtools[J]. Bioinformatics, 2009, **25**(16):2078-2079

[19] Liu Y, Li B, Tan R, *et al*. A gradient-boosting approach for filtering de novo mutations in parent-offspring trios [J]. Bioinformatics,2014,**30**(13):1830-1836

[20] Li J, Jiang Y, Wang T, *et al*. mirTrios: an integrated pipeline for detection of de novo and rare inherited mutations from trios-based next-generation sequencing[J]. J Med Genet, 2015, **52**(4):275-281

[21] Wei Q, Zhan X, Zhong X, *et al*. A Bayesian framework for de novo mutation calling in parents-offspring trios [J]. Bioinformatics,2015,**31**(9):1375-1381