基于特征挖掘的基因组缺失变异集成检测方法

张晓东 凌 诚 高敬阳
(北京化工大学信息科学与技术学院 北京 100029)

摘 要 随着高通量测序技术的应用与发展,基于测序的缺失变异检测方法大量涌现。然而,单一检测方法仍存在适用的局限性以及检测精度与敏感度不足的问题。为此,提出一种基于多检测理论融合的特征挖掘与机器学习算法集成的基因组缺失变异综合检测方法。该方法将多种工具应用于个体缺失变异检测,得到变异检测初始集;再根据多种检测理论对初始集中的缺失变异进行序列特征挖掘与特征提取;最后,将检测工具与机器学习算法相融合以获得集成的检测方法,剔除初始集中的假阳性变异,获得最终的结果集。基于千人基因组计划数据的实验表明,相较于单个工具的检测结果,该方法在检测精度和敏感度上均占优势;相较于多个工具检测结果的直接组合,该方法在损失少许检测敏感度的前提下显著地提高了检测精度。

关键词 缺失变异,特征挖掘,集成检测
中图法分类号 TP391,Q523 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.01.015

Integrated Feature Mining Based Approach for Calling Genomic Deletions

ZHANG Xiao-dong LING Cheng GAO Jing-yang
(College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract With the application and development of next generation sequencing technology, methods of calling genomic deletions based on sequencing have proliferated. However, using a single method to call deletions has limitation in application and insufficiency of precision and sensitivity. To solve these problems, an integrated approach for calling deletions was proposed based on feature mining according to combining multiple theory and machine learning algorithm. First, different callers are used for calling deletions. These results are merged as aninitial result set of deletions. Then, according to variety of detection strategies, features of the initial result set of deletions are extracted based on next generation sequencing data. Finally, to obtain the final result set of calling deletions, a machine learning model is trained to distinguish false positive deletions from initial call set. The experimental results show that compared with a single caller such as Pindel and SVseq2, the proposed approach has higher precision and sensitivity simultaneously. Compared with directly merging multiple deletion call sets, the proposed approach can significantly improve the precision with slight loss of sensitivity.

Keywords Deletion, Feature mining, Integrated detection

1 引言

基因组变异包括单核苷酸多态性(SNP)和结构变异(Structural Variation,SV),结构变异为非单碱基替换的基因组变化,包括缺失、插入、倒置、倍增、易位和拷贝数变异(CNV)等[1],其中较短(<50bp)的插入和缺失变异又简称为indel。而SV在人类基因组中覆盖的核苷酸总数远远超过SNP位点的总数,有研究发现SV对基因组的影响比SNP更大[2]。基因组结构变异与个体表现型、疾病的易感性和产生、生物进化等研究息息相关,如基因NRXN1的杂合变异与孤独症和精神分裂症相关联[3];儿童的先天性心脏缺陷与22q11.2区的缺失紧密相关[4]。因此,敏感且精准的全基因组范围内的SV检测具有重大的意义。

在过去的几年里,随着测序技术的不断发展和测序成本的不断降低,基于测序的SV检测方法也快速发展,其主要分为以下4类[5]:双末端测序序列映射分析(Read pair, Paired-end mapping)、测序序列分裂比对(Split read)、测序序列映射深度分析(Read depth)、序列拼接(Assembly)。基于以上4类检测理论的测序工具不断涌现,检测精度和敏感度不断提升,较优秀的SV检测工具包括:Pindel[6],SVseq2[7],DELLY[8],BreakDancer[9],CNVnator[10]和VariationHunter[11]等。然而,每种检测工具都有各自的使用条件和适宜性,且各有侧重,都不可能大而全地检测出所有变异。一般而言,针对同一个体,不同工具的检测结果有交集也有差异,既有多个工具共同支持的检测结果,也有部分工具支持的检测结果,甚至单一工具支持的检测结果也大量存在。图1为不同检测工具结果交集的示意图。

到稿日期:2015-10-03 返修日期:2016-01-22 本文受国家自然科学基金(61472026),广州市科技计划项目(2014J4100081)资助。
张晓东(1991—),男,硕士生,主要研究方向为生物信息学、机器学习;凌 诚(1987—),男,博士,讲师,主要研究方向为高性能计算、深度学习;高敬阳(1966—),女,博士,教授,主要研究方向为生物信息学、机器学习,E-mail:gaojy@mail.buct.edu.cn(通信作者)。

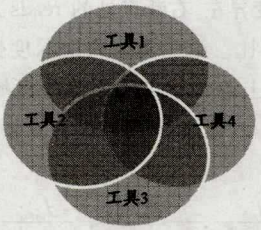


图 1　不同工具检测结果交集的示意图

通常，有研究人员将多个检测工具的检测结果进行直接合并，以换取最大化的检测敏感度，但同时也将引入大量的假阳性结果。本文针对长度大于 50bp 的缺失变异，提出一种基于多检测理论特征挖掘的集成多检测工具和机器学习算法的综合检测方法，该方法以 Read pair，Split read 和 Read depth 3 类检测方法为理论基础，以充分挖掘并提取与缺失变异相关的序列特征为途径，以检测工具与机器学习算法的集成为核心，最终实现对缺失变异的有效检测。

2　集成检测方法

本文提出一种新的缺失变异集成检测方法。其中集成的含义包括两个方面，一方面指不同数据的集成，另一方面指不同方法的集成。首先，利用多种检测工具对同一个体进行缺失变异检测，将不同工具的检测结果进行集成以构成初始集，以此来最大化检测敏感度；然后，根据多种检测理论对初始集中的缺失变异进行序列特征的挖掘和提取，并作为二次检测的基础；最后，将检测工具与机器学习相融合以获得集成的检测方法，即机器学习模型根据所挖掘的特征对初始集中的缺失变异进行判别，进而剔除由于极大化敏感度带来的大量假阳性变异，达到大幅提高检测精度的目的。缺失变异集成检测的总体检测流程如图 2 所示。全方位地挖掘并提取出有效的缺失变异序列特征是该方法成功的关键。

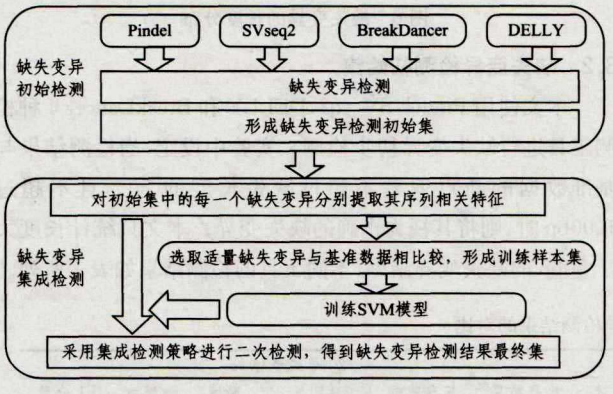


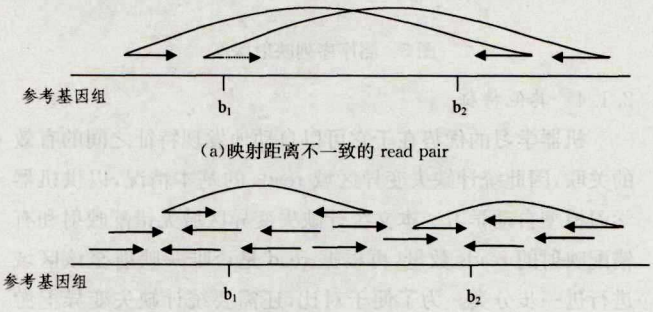
图 2　缺失变异集成检测总体流程图

2.1　特征挖掘

检测个体基因组缺失变异时，首先将测序得到的序列数据(reads)通过比对工具(如 BWA[12])映射到参考基因组上，从而得到包含 reads 比对信息的 SAM/BAM[13] 格式文件。缺失变异检测工具通常以 BAM 文件作为处理和分析对象，提取并分析 reads 的比对信息，对异常的 reads 进一步处理，最终找到缺失变异的准确位点。其中，支持或反对某区域存在缺失变异的信息称为缺失变异的序列特征。本文挖掘到的缺失变异特征主要包括以下 4 类。

2.1.1　双末端测序序列映射距离

对于初始集中的缺失变异$[b_1, b_2]$，首先对个体测序数据中缺失变异区域的 read pair 的映射距离进行分析。如图 3 所示，当 read pair 的映射距离大于测序文库制备时插入尺寸(insert size)的上限时，称其为不一致的序列，则该 read pair 支持此区域存在缺失变异(见图 3(a))；相反，若其映射距离在正常范围之内，则该 read pair 反对此区域存在缺失变异(见图 3(b))。本文设定阈值，当映射距离大于 $m+3*v$ 时为不一致的双末端测序序列，其中 m 为插入尺寸的均值，v 为标准差。



(a)映射距离不一致的 read pair

(b)映射距离一致的 read pair

图 3　双末端测序序列映射距离

分别统计映射距离一致与不一致的 read pairs 数量。reads 中包括有错配映射和无错配映射两种情况(见图 3(a))，不同的情况对缺失变异的支持程度不尽相同，故分开统计。此外，唯一映射至该区域与非唯一映射至该区域的 reads 也不能一概而论，前者更能反映该区域的序列特征，而后者可能在其他区域也存在最优映射，两种情况分开统计有助于识别 CNV 型的缺失变异。统计上述共 8 种情况下 read pair 的数量作为一类特征。

2.1.2　双末端测序序列分裂比对

通常，跨越断点的 reads 往往最能反映出缺失变异的精确位点信息，如图 4 所示，跨越断点的 reads 不能作为一个整体映射到参考基因组，只能以分裂比对的形式部分映射，这种 reads 更能支持该缺失变异的存在。这一类特征反映跨越缺失变异断点的 reads 情况，包括左、右断点两部分区域，对于每个区域，分别统计无错配映射、部分映射以及其他情况映射的 reads 数量。此外，根据其配对 read 的相对位置(上游还是下游)以及 read 是否唯一映射至该区域进行进一步分类，最终挖掘得到 24 个该类特征。

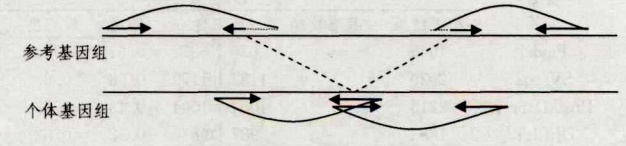


图 4　双末端测序序列分裂比对

2.1.3　测序序列映射深度

第二代测序技术中，测序覆盖深度的概率分布服从正态分布或泊松分布，即测序所得 reads 基本是均匀分布的。个体基因组相较于参考基因组存在缺失时，测序数据映射到参考基因组上后，如图 5 所示，缺失区域的平均覆盖深度要远远小于其他正常区域，故覆盖深度也可作为一组特征。

根据公式$\frac{\sum_{i=1}^{l_d} d_i}{l_d}$计算平均覆盖深度。其中，$l_d$ 为统计区域的长度；d_i 为该区域位置 i 处的覆盖深度，本文使用 SAMtools[7]进行计算。令 l_d 为缺失变异的长度，统计缺失变异区域的平均覆盖深度；此外，为了便于比较，还需要计算缺失变异上游 l_d 和下游 l_d 区域的平均覆盖深度，这一类共包括 3 个特征。

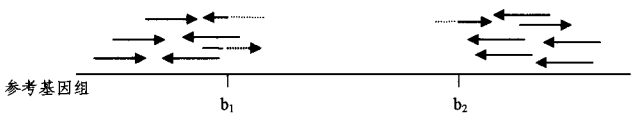


图 5 测序序列映射深度

2.1.4 其他特征

机器学习的优势在于它可以自动地发现特征之间的有效的关联，因此统计缺失变异区域 reads 的基本情况，以供机器学习模型自动学习。本文统计缺失变异区域无错配映射和有错配映射的 reads 数量，再根据 read 是否唯一映射至该区域进行进一步分类。为了便于对比，还需要统计缺失变异上游 l_d 和下游 l_d 区域的上述特征，合计 12 个。此外，将缺失变异长度也作为缺失变异的一个特征。

2.2 特征的标准化

对初始集合中的缺失变异进行特征挖掘和提取之后，还不能直接将其应用于机器学习模型的判别。因为不同覆盖深度的数据提取的特征绝对值有差距，如果训练数据和测试数据的覆盖深度不同，会导致训练的模型不能对测试数据进行有效分类。即使采取同样覆盖深度的训练和测试数据，不同缺失变异周围的覆盖深度也可能有所差异。如果一个缺失变异区域的测序覆盖深度很低，那么即使是很少数的不一致 read pairs，仍能强烈地支持该区域存在缺失变异；反之，如果覆盖深度很高，那么少量的不一致 read pairs 就不能提供强烈支持该区域存在缺失变异的信号。所以，要对每一个缺失变异的特征进行标准化的处理。

第一类为 read pair 的映射距离提供的特征，首先对所统计的 read pairs 数量进行求和，即 8 个统计量之和，它代表了跨越缺失变异区域的 read pairs 总数，然后将每一个特征除以该值，处理过后的特征代表该种类的 read pairs 数量相对于总数的占比。第二类为序列的分裂比对提供的特征，同样地，分别统计跨越缺失变异左、右断点处的 reads 总数，求得相应特征 reads 数量的占比。第三类为映射深度提供的特征，这里把缺失变异区域、上游 l_d 区域和下游 l_d 区域的平均覆盖深度记作 d_{in}，d_{up} 和 d_{down}，通过如下公式计算得到 4 个新的特征值：

$$\frac{d_{in}}{d_{in}+d_{up}},\frac{d_{up}}{d_{in}+d_{up}},\frac{d_{in}}{d_{in}+d_{down}},\frac{d_{down}}{d_{in}+d_{down}}$$

其他类特征也做类似处理，即每一个特征值均除以该类 reads 的总数，从而得到相应特征 reads 数量的比率。

经上述处理之后，所有的特征值均变为[0,1]之间的实数，这样就消除了覆盖深度对缺失变异特征的影响，最终共得到 49 个特征值用于缺失变异的二次检测，即机器学习模型对初始集中的缺失变异结果进行判别。

3 实验与分析

3.1 数据来源

实验基于国际千人基因组计划[15]第三阶段的测序数据，包括 YRI 种群 17 个个体的 11 号和 20 号染色体，平均测序覆盖深度为 6.51×，平均插入尺寸均为 446bp，标准差在 57～78bp 范围内。实验采用的 BAM 文件是通过 BWA 将测序数据映射到 NCBI37 版本的参考基因组上而得到的。基准变异数据使用国际千人基因组计划发布的文件。根据基准数据，17 个个体的 11 及 20 号染色体中合计存在 1904 和 921 个缺失变异，其长度分布如图 6 所示。

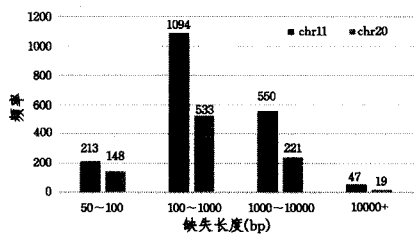


图 6 缺失变异的长度分布

3.2 缺失变异检测初始集

本文使用 Pindel，SVseq2，DELLY 和 BreakDancer 4 种检测工具进行缺失变异初步检测。实验中设定，当检测结果与基准数据的断点偏差不超过缺失长度的 30% 且不超过 1000bp 时，则将其视为正确的缺失变异。本文只统计长度大于 50bp 的缺失变异结果，不同工具的检测结果如表 1 所列。

表 1 不同工具缺失变异检测结果的对比

检测工具	11 号染色体						20 号染色体					
	检测结果	基准数据	真阳性	精度	敏感度	F1 分数	检测结果	基准数据	真阳性	精度	敏感度	F1 分数
Pindel	1043	1904	636 | 635	0.61	0.33	0.43	521	921	325|325	0.62	0.35	0.45
SVseq2	2030		1182 | 1179	0.58	0.62	0.60	963		539 | 537	0.56	0.59	0.57
BreakDancer	2215		1004 | 1004	0.45	0.53	0.49	1091		451 | 451	0.41	0.49	0.45
DELLY	1883		987 | 987	0.52	0.52	0.52	883		456 | 456	0.52	0.50	0.51

注："|"后面的数字为结果集中覆盖基准数据中缺失变异的个数。

从表 1 中可以看出，SVseq2 检测结果的 F1 分数最高，总体效果最好，其敏感度最高，精度比 Pindel 稍差；Pindel 的精度最高，但敏感度最低，在 F1 分数上相比于其他工具偏低；DELLY 和 BreakDancer 不论检测精度还是敏感度均处于中等水平。

将上述 4 个工具的检测结果进行组合，得到缺失变异检测结果初始集。由于不同工具的检测结果存在交集，需要将初始集中的缺失变异进行归并去重。最终，获得 8268 个缺失变异。

3.3 集成检测结果与分析

本文选取机器学习算法中的支持向量机(SVM)作为分类模型，用 LibSVM[14]工具来进行实验。对缺失变异检测初

始集中的每一个缺失变异，提取前述的相关序列特征。之后，从初始集中随机选取适量的缺失变异，通过与基准数据进行对比，为其加注真、假标签，以此作为机器学习算法 SVM 模型的训练数据，核函数采用默认的径向基函数，并使用网格遍历来寻找最优参数。期间，采用 10 折交叉验证来估计模型精度。得到训练好的模型之后，用其来判别初始集中的缺失变异，将分类为“真”的变异作为最终的检测结果，如表 2 所列。

表 2　集成检测结果与初始集的对比

结果集	检测结果	基准数据	真阳性	真阴性	假阳性	假阴性	精度	敏感度
初始集	8268	2825	4062 | 2267		4206	558	0.491	0.803
最终集	3978		3669 | 2005	3897	309	+393|262	0.922	0.710

注：“|”后面的数字为结果集中覆盖基准数据中缺失变异的个数。

表 2 显示，经过集成模型的判别后，初始集存在的 4206 个假阳性结果中只有 309 个未能正确排除，检测精度由 49.1%提升至 92.2%；初始集中未检测出的缺失变异即假阴性个数为 558，而最终集的假阴性个数增加了 393，其覆盖了基准数据中 262 个缺失变异，这使得敏感度有所下降，由 80.3%降低为 71%。初始集与最终集相比，在损失少量检测敏感度的前提下，检测精度得到显著提高，即集成检测方法排除了大量的假阳性结果，在大幅提高检测精度的同时只损失了少许检测敏感度。

将各种单一工具的检测结果、初始检测结果以及二次集成检测结果进行对比，如图 7 所示。多个检测工具组合的结果的敏感度最高，但是它同时带入了大量的假阳性缺失变异。而本文提出的集成检测方法首先合并多个结果集以最大化检测敏感度，然后利用机器学习模型在损失少许检测敏感度的情况下有效过滤假阳性结果；相比于其他工具，集成检测方法的检测精度和敏感度得到明显提升。

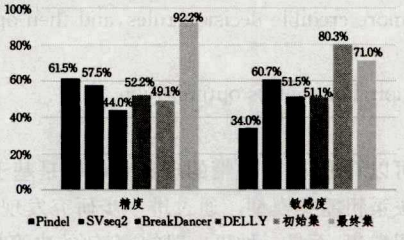


图 7　不同缺失变异检测结果集合的精度与敏感度对照图

结束语　本文提出了一种基于特征挖掘的基因组缺失变异集成检测方法，该方法将多种检测工具与机器学习算法在数据和方法两方面进行集成。首先将多个工具的缺失变异检测结果进行集成，从而最大化检测敏感度，以此作为缺失变异检测初始集；然后，在多种检测理论的基础上全方位挖掘缺失变异的序列特征，并对初始结果集中的缺失变异进行特征提取；最后，结合机器学习算法，根据所提取的特征对初始集中的缺失变异进行真假判别，将分类为“真”的变异作为最终的检测结果。基于千人基因组计划数据的实验表明，与其他方法相比，本文所提方法在检测精度和敏感度上均具有明显优势。

参 考 文 献

[1] EICHLER E E, NICKERSON D A, ALTSHULER D, et al. Completing the map of human genetic variation [J]. Nature, 2007, 447(7141): 161-165.

[2] 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.

[3] PAK C H, DANKO T, ZHANG Y, et al. Human neuropsychiatric disease modeling using conditional deletion reveals synaptic transmission defects caused by heterozygous mutations in NRXN1[J]. Cell Stem Cell, 2015, 17(3): 316-328.

[4] LEE M Y, WON H S, BAEK J W, et al. Variety of prenatally diagnosed congenital heart disease in 22q11. 2 deletion syndrome [J]. Obstetrics & Gynecology Science, 2014, 57(1): 11-16.

[5] ALKAN C, COE B P, EICHLER E E. Genome structural variation discovery and genotyping [J]. Nature Reviews Genetics, 2011, 12(5): 363-376.

[6] YE K, SCHULZ M H, LONG Q, et al. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads[J]. Bioinformatics, 2009, 25(21): 2865-2871.

[7] ZHANG J, WANG J, WU Y. An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data [J]. BMC Bioinformatics, 2012, 13(Suppl 6): 1-11.

[8] RAUSCH T, ZICHNER T, SCHLATTL A, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis[J]. Bioinformatics, 2012, 28(18): i333-i339.

[9] CHEN K, WALLIS J W, MCLELLAN M D, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation[J]. Nature Methods, 2009, 6(9): 677-681.

[10] ABYZOV A, URBAN A E, SNYDER M, et al. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing [J]. Genome Research, 2011, 21(6): 974-984.

[11] HORMOZDIARI F, HAJIRASOULIHA I, DAO P, et al. Next-generation Variation Hunter: combinatorial algorithms for transposon insertion discovery [J]. Bioinformatics, 2010, 26 (12): i350-i357.

[12] LI H, DURBIN R. Fast and accurate short read alignment with Burrows-Wheeler transform[J]. Bioinformatics, 2009, 25(14): 1754-1760.

[13] LI H, HANDSAKER B, WYSOKER A, et al. The sequence alignment/map format and SAMtools[J]. Bioinformatics, 2009, 25(16): 2078-2079.

[14] CHANG C C, LIN C J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 389-396.

[15] 1000 Genomes Project Consortium. An integrated map of genetic variation from 1092 human genomes [J]. Nature, 2012, 491 (7422): 56-65.