doi:10.3969/j.issn.1672-5565.2014.01.02

基于高通量测序技术的基因组结构变异检测算法

高敬阳,齐 飞,管 瑞
(北京化工大学信息科学与技术学院,北京 100029)

摘 要:基因组结构变异的检测是生物信息学的重要方向之一。本文分别对基于高通量测序技术的双末端映射方法、映射分布方法、分裂片段方法和序列拼接方法等检测技术的四种算法进行详细的解读和说明,阐述了以上四种方法两两结合的检测算法,并分析了各种检测方法的性能和适用的条件,说明混合结合的方法将会成为未来发展的方向。

关键词:生物信息学;高通量测序技术;结构变异(SV)
中图分类号:R318.04 **文献标志码**:A **文章编号**:1672-5565(2014)-01-005-05

High-throughput based algorithm of detecting genome structural variation

GAO Jingyang, QI Fei, GUAN Rui
(*School of Information Science&Technology, Beijing University of Chemical Technology, Beijing* 100029, *China*)

Abstract: Structural variation detection is one of the most important directions of bioinformatics research. In this paper, we firstly illustrated four sequencing-based approaches in detail, read-pair, read-depth, split-read and assembly.Then we introduced algorithms based on pair wise combination of those four approaches, and analyzed their performance and conditions. Finally, we argued that the combined approaches will be the direction of the future.

Keywords: Bioinformatics; High-throughput sequencing(HTS); Structural variation(SV)

DNA测序技术即基因测序技术是研究基因组结构变异的重要方法。基因组结构变异是指一定长度范围的DNA序列上的差异,包括缺失、插入、重复、倒置等[1]。

如何利用测序技术检测基因组结构变异中的插入、缺失、重复、倒置等情形是问题的关键。目前基于测序技术的算法主要有四种,分别是双末端映射、映射分布、分裂片段和序列拼接技术。前三种技术均通过将短序列片段(Reads)映射到参考基因组上,通过对比个体基因组和参考基因组之间的差异信息来确定基因组结构变异,而序列拼接方法是通过拼接算法将短序列片段组装还原个体的整个基因,然后将这个拼接好的基因同参考基因组做对比来检测结构变异。

相比之前的sanger测序技术[2-3]而言,高通量测序技术有着测序速度快、耗费低等优点[1,4],人类全基因组测序使用高通量测序仪只需数千美元不到一个月即可完成[5-6],但高通量测序产生的测序片段并没有sanger测序结果准确,所以使得基于高通量测序的四种算法在未来的研究中有很大的改进和提升空间[7]。

1 结构变异检测方法

1.1 双末端映射方法

双末端映射方法(Paird-end mapping)也称为片段对方法(Read-pair),其首先通过高通量测序技术获得大量个体基因片段对,得到这些片段对中的如碱基的组成、片段的长度以及片段对之间的距离等信息,然后利用基于比对算法BWT、BWA、MAQ等的映射工具,将这些片段对映射到参考基因组上,获得如片段对在参考基因组的映射距离以及映射方向,最后将这些映射的信息与个体基因片段对信息进行对比,从而检测出基因

收稿日期:2013-10-25;**修回日期**:2013-11-27.
基金项目:国家自然基金资助项目(51275030)。
作者简介:高敬阳,女,副教授,博士。研究方向:人工智能,生物信息学;E-mail:gaojy@mail.buct.edu.cn.

组的结构变异。

双末端映射方法就是聚类至少两种不一致的片段对,包括片段对之间的距离和映射的方向。正常情况下,映射到参考基因上片段对之间的距离和映射的方向都是固定的,但是当存在结构变异时,例如片段中的缺失(见图 1),就会使得映射到参考基因上片段对之间的距离和库中的距离不一致,这时片段对之间的距离等于库中的距离加上缺失片段的长度。同理,如果存在片段插入(见图 2),映射到参考基因上的片段对之间的距离等于库中的距离减去插入片段的长度。当变异为片段倒置的时候(见图 3),片段对中的一个片段映射到该区域的方向和库中的一致,而另一个片段的映射方向相反,则说明在该区域内存在着基因片段的倒置。原则上,双末端映射方法可以检测到大多数的基因组结构变异。

双末端映射方法是当前最广泛应用的方法,许多算法都是基于此方法,例如 PEMer[8], VaristionHunter[9-11], BreakDancer[12], MoDIL[13], MoGUL[14], HYDRA[15], Corona[16] 等算法。

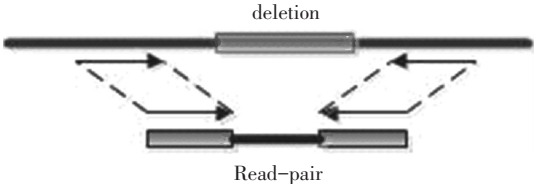

图 1 双末端映射方法检测到的缺失

Fig.1 Deletion detected by read-pair

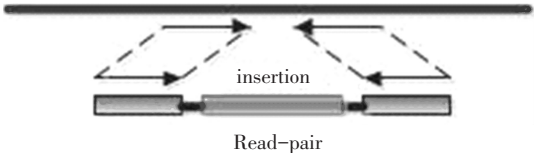

图 2 双末端映射方法检测到的插入

Fig.2 Insertion detected by read-pair

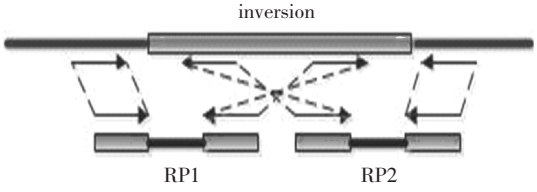

图 3 双末端映射方法检测到的倒置

Fig.3 Inversion detected by read-pair

1.2 映射分布方法

映射分布方法(Read-depth)也称为映射深度分析法,它是通过分析映射深度来判断该区域是否存在结构变异。所谓的映射深度是衡量该 reads 在参考基因组中的映射程度。

通过高通量测序可以从个体上得到大量的 reads,这些 reads 的长度,碱基组成是已知的。正常情况下,这些 reads 映射到参考基因组中各个区域的数量应该是相等的,但是当个体基因存在结构变异时,某个区域映射的 reads 会比其他区域映射的数量或者多或者少,这就说明在该区域存在着片段重复或者缺失(见图 4、图 5)的情况。EWT(The event-wise-testing)[17] 和 CNVnator 算法[18-20] 都是基于此方法。

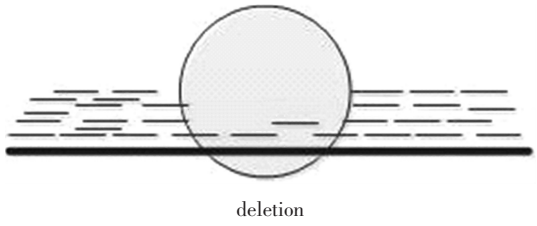

图 4 映射分布方法检测到的缺失

Fig.4 Deletion detected by read-depth

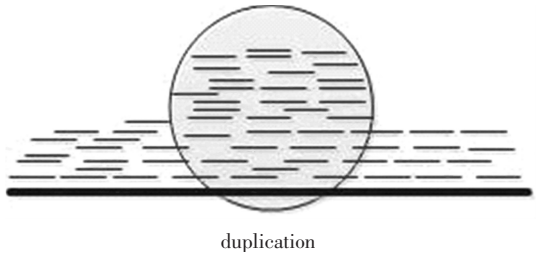

图 5 映射分布方法检测到的重复

Fig.5 Duplication detected by read-depth

1.3 分裂片段方法

分裂片段方法(Split-read)最初是为 sanger 测序开发的[21],片段的长度越长,检测基因结构变异的效果越好。

顾名思义,分裂片段方法就是将 reads 分裂成两部分分别映射的方法,一个完整的 reads 通过与参考基因组的映射,可能出现映射不成功的情况。该方法将一个没有映射的片段从不同的碱基位置依次分裂成两段,再将这两个小的片段分别映射到参考基因组中,如果这两个片段在某个区域能够分别映射,说明在该区域基因存在着片段缺失(见图 6),并且缺失长度为两个片段映射之后在参考基因中的距离。同理,当基因中存在片段插入时(见图 7),分裂片段的两部分只能映射上一个,而在另一种分裂情况下也只能映射一个,而这两次映射恰好在参考基因组中相邻。当分裂片段的两部分均能够映射到参考基因组上,但映射方向不同,这说明存在基因结构倒置(见图 8)。Kai Ye 等人提出的 Pindel[22] 算法、Alexej Abyzov 等人提出的 AGE[23] 算法和 Zhang Jing 等人提出的 SVseq[24] 算法就是基于该方法的典型算法。

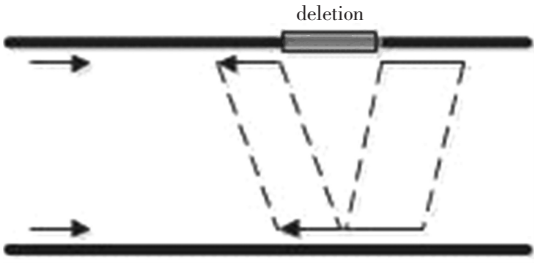

图 6　分裂片段方法检测到的缺失

Fig. 6　Deletion detected by split-read

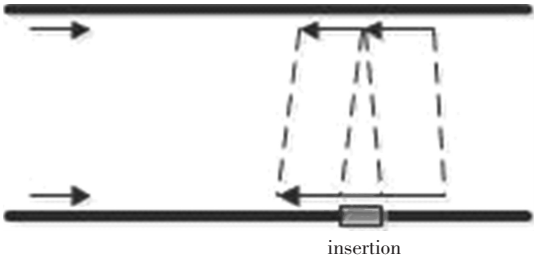

图 7　分裂片段方法检测到的插入

Fig.7　Insertion detected by split-read

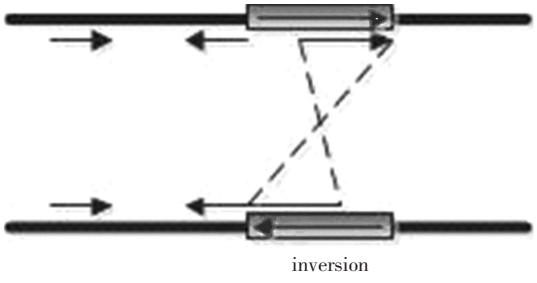

图 8　分裂片段方法检测到的倒置

Fig.8　Inversion detected by split-read

1.4　序列拼接方法

序列拼接方法(Assembly)是通过将测序得到的诸多基因片段重新组装,并与参考基因组进行对比,通过比较与参考基因组之间的差异,找到基因的结构变异,如图 9、10、11。

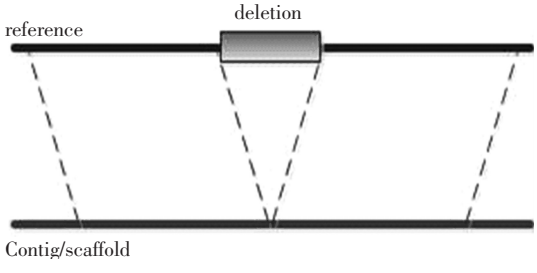

图 9　序列拼接方法检测到的缺失

Fig.9　Deletion detected by assembly

理论上,如果测序得到的片段足够准确以使得最原始的片段组装算法有效,那么所有的结构变异是可以被检测到的。但在实际中,序列组装还仅仅是在研究的初期,目前还只能应用原始和局部的结合算法恢复原始基因组。理想的情况下,高质量的原始片段组装法可以找到上千个结构变异。

基于高通量测序技术的原始组装算法主要有 EULER-USR[25], ABySS[26], SOPdenovo[27] 和 ALLPATHS-LG[28]等。

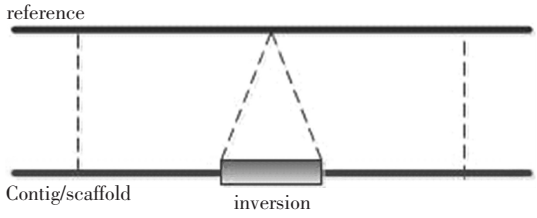

图 10　序列拼接方法检测到的插入

Fig.10　Insertion detected by assembly

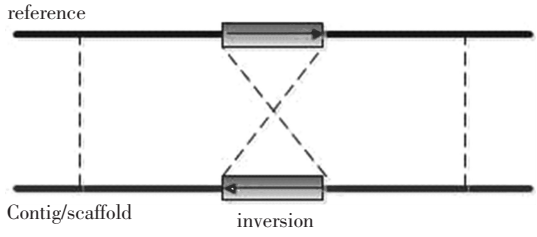

图 11　序列拼接方法检测到的倒置

Fig.11　Inversion detected by assembly

2　检测方法的性能

根据基因变异类型的不同,以上四种方法有着各自的优缺点和应用范围。当前,90%以上的高通量测序片段长度小于 1 kb,而且大部分结构变异都为缺失而非插入[18-19]。双末端映射方法虽然很具优势,能够检测到几乎所有的结构变异,但是在检测结构重复时并不是很准确,而且如果需要检测真正的断点就必须建立紧密的片段分布,这会使得库的建立非常的困难和消耗巨大[29]。映射分布方法可以真正的找到重复的区域,但是却很难确定断点的准确位置,所以该算法主要被用来检测重复的数量。分裂片段方法不仅可以像双末端映射方法一样检测到缺失、插入和倒置,而且还可以确定移动元素插入的位置,但检测移动元素插入时,片段的长度必须大于这个移动元素的长度。虽然分裂片段方法可以找到许多基因结构变异的断点,但由于高通量测序产生的片段都相对较短,所以,制约着分裂片段方法的效果。序列拼接方法是最通用的算法,但是当该区域发生片段重复时,就可能使得该方法在该区域产生崩溃性的错误[30-31]。上述提到的方法只能找到相

对较小的基因组结构变异而且尚存在较多不足之处。

高通量测序技术特点之一是产生的片段长度较之前的 sanger 测序的长度短。由于人类基因组非常的复杂,所以需要通过片段的模糊映射来提高映射的专一性和敏感性。一项评估表明:即使长度超过 1 kb 的片段也会有超过 1.5%的人类基因组很难被唯一的映射[32]。测序的覆盖度也是影响结构变异检测敏感性和精确性的一个重要因素。正因为如此,促使一些新的算法的涌现来提高检测的敏感性和精确性。

3 混合检测方法

以上四种检测方法每一种在独立应用方面虽然有许多优点,但是缺陷和限制也非常明显。因此,有研究者在实际应用中尝试将其中两种算法相结合来检测基因组结构变异。算法的结合主要是为了克服使用一种算法时的限制,从而得到更好的检测效果。

3.1 双末端映射方法和映射分布方法结合

CNVer[33]是将双末端映射方法和映射分布方法相结合的算法,其主要被应用在检测基因组结构的重复,也称为拷贝数的检测,该算法克服了独立应用一种方法时的不足,例如利用双末端映射方法检测的插入片段时,该长度必须小于片段对中间的距离,否则无法检测出该插入片段,但是高通量测序技术产生的片段对之间的距离往往小于 1 kb,所以很可能漏掉此片段。并且两种方法结合还可以提高检测的鲁棒性。

3.2 双末端映射方法和分裂片段方法结合

Pindel 方法是双末端映射和分裂片段方法结合的算法,它是第一个能够检测到缺失的长度达到 10 kb而片段对的长度只为 36 bp 的算法。而且该算法也提出了一个新的检测断点的方法:增长模式法。该方法可以相对快速的检测结构变异中的断点。该算法的出现利用双末端映射方法来减小潜在的结构变异的搜索空间,因此,减少了短片段映射到参考基因组时局部间隙的计算量,提高了检测效率。

Svseq 算法的出现使得准确率进一步提升,该方法相对于 Pindel 方法不同,分为两步,一是利用加强的分裂片段映射来找到多个候选的缺失,第二步利用已经映射的片段对过滤掉候选中的假缺失,保留下真缺失。

3.3 分裂片段方法和机器学习方法结合

众所周知,如何利用高通量测序技术精确地检测基因组结构的变异是一项重大的挑战。而现存的方法通常通过检测某个区域映射的信息,例如映射的分裂片段的数量,然后人为的设定一个映射片段数量的阈值,大于这个阈值的被检测到的变异为真正的基因结构变异。但这个阈值往往很难确定。

Dominik Grimm 等人提出了一种关于机器学习的基因组结构变异检测方法[34],该方法主要是将支持向量机和分裂片段方法相结合,所谓的支持向量机就是一个分类模型,它通过一个超平面将样本分为不同的两类。该检测方法根据参考基因映射的特征来训练一个支持向量机的模型,这个训练的数据是通过 sanger 测序得到的。首先利用分裂片段方法来检测基因中的插入和缺失,将这部分检测到的插入和缺失作为候选,然后再利用训练好的支持向量机从候选的插入和缺失中筛选出正确的基因结构变异。

利用机器学习最大的优势是可以和任何一种检测方法相结合,该检测过程可以从机器学习过程中自动的获得权值参数,而不需要人为的设定,所以避免了人为的错误,提高了检测的精度。

4 结 论

本文介绍了双末端映射、映射分布、分裂片段和序列拼接四种基因组结构变异检测方法,详细阐述了各种检测方法以及其优势与适用的条件,并总结和归纳了几种检测方法相结合的混合检测算法。混合检测的目的是为了克服各种独立检测方法在检测基因组结构变异时的缺点和不足,其中介绍了一种机器学习与分裂片段检测方法相结合的算法,该算法的出现大大提高了检测速度和检测精度,并且实现了检测的半自动化。

总之,利用现存的四种检测方法中的两种或者与类似于机器学习方法相结合来检测基因组结构变异有种种的优势,它不仅不用人为的设定阈值,而且还可以集两种方法的优点于一身提高检测精度,因此,机器学习方法,例如贝叶斯分类器、决策树、神经网络等算法在基因组结构变异检测中有很广阔的应用前景。

参考文献(References)

[1] SCHUSTER S C. Next-generation sequencing transforms today's biology [J]. Nature Methods, 2008, 5(1): 16-18.

[2] SANGER F, NICKLEN S, COULSON A R. DNA sequencing with chain-terminating inhibitors [C]. Proc. Natl. Acad. Sci. USA, 1977, 74(12): 5463-5467.

[3] BENTLEY D R. Whole-genome re-sequencing [J]. Current Opinion Genetics&Development, 2006, 16(6): 545-552.

[4] SHENDURE J, HANLEE J I. Next-generation DNA sequencing [J]. Nature Biotechnology, 2008, 26(10):

1135-1145.
[5] WHEELER D A,SRINIVASAN M,EGHOLM M,et al. The complete genome of an individual by massively parallel DNA sequencing [J].Nature,2008,452(7189):872-876.
[6] DRMANAC R, SPARKS A B, CALLOW M J, et al. Human genome sequencing using unchained Based reads on self-assembling DNA nanoarrays [J]. Science, 2010, 327(5961):78-81.
[7] 林勇.面向下一代测序技术的 de novo 序列拼接工具综述[J].小型微型计算机系统,2013,34(3):627-631.
LIN Yong. Survey of de novo Assembly Tools for Next-generation Sequencing Technology [J].Journal of Chinese computer systems,2013,34(3):627-631.
[8] KORBEL J O, ABYZOV A, MU X J, et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data [J]. Genome Biology,2009,10(2):R23.
[9] HORMOZDIARI F, ALKAN C, EICHLER E E, et al. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes [J].Genome Res,2009, 19:1270-1278.
[10] HORMOZDIARI F, HAJIRASOULIHA I, DAO P, et al. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery [J]. Bioinformatics, 2010,26:i350-i357.
[11] HORMOZDIARI F, HAJIRASOULIHA I, McPherson A, et al. Simultaneous structural variation discovery in multiple paired-end sequenced genomes [J]. Genome Research,2011,21:2203-2212.
[12] CHEN R, WALLIS J W, MCLELLAN M D, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation [J]. Nature Methods, 2009, 6: 677-681.
[13] LEE S,HORMOZDIARI F,ALKAN C, et al. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions [J].Nature Methods,2009,6:473-474.
[14] LEE S, XING E, BRUDNO M. MoGUL: detecting common insertions and deletions in a population [M]. Research in Computational Molecular Biology, 2010, 6044:357-368.
[15] QUINLAN A R, CLARK R A, SOKOLOVA S, et al. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome [J].Genome Research, 2010,20:623-635.
[16] STUART J R, MALEK J A, MANNING J M, et al. Blanchard A P. Sequence and structural variation in a human genome uncovered by short-read massively parallel ligation sequencing using two-base encoding [J]. Genome Research,2009,19:1527-1541.
[17] YOON S, XUAN Z Y,MAKAROV V,et al. Sensitive and accurate detection of copy number variants using read depth of coverage [J].Genome Research,2009,19:1586-1592.
[18] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing [J]. Nature,2010,467(7319):1061-1073.
[19] MILLS R E, WALTER K, STEWART C, et al. Mapping copy number variation at fine scale by population scale genome sequencing [J].Nature,2011,470:59-65.
[20] 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, 2011,21:974-984.
[21] MILLS R E,LUTTING C T, LARKINS C E, et al. An initial map of insertion and deletion (INDEL) variation in the human genome [J].Genome Research,2006,16:1182-1190.
[22] YE R,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,2009,25:2865-2871.
[23] ABYZOV A, GERSTEIN M. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision [J]. Bioinformatics,2011,27:595-603.
[24] ZHANG J, WU Y F. SVseq: an approach for detecting exact breakpoints of deletions with low-coverage sequence data [J].Bioinformatics,2011,27(23):3228-3234.
[25] CHAISSON M J, BRINZA D, PEVZNER P A.De novo fragment assembly with short mate-paired reads: does the read length matter? [J].Genome Research,2009,19:336-346.
[26] SIMPSON J T, WONG K, JACKMAN S D, et al. ABySS: a parallel assembler for short read sequence data [J].Genome Research,2009,19:1117-1123.
[27] LI R Q, ZHU H M, RUAN J,et al. De novo assembly of human genomes with massively parallel short read sequencing [J].Genome Research,2009,20:265-272.
[28] GNERREA S, MACCALLUMA I, PRZYBYLSKI D,et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data [J].Proc.Natl.Acad.Sci. USA,2011,108(4):1513-1518.
[29] MEDVEDEV P, STANCIU M, BRUDNO M. Computational methods for discovering structural variation with next-generation sequencing [J].Nature Methods,2009,6: S13-S20.
[30] SHE X W, JIANG Z S, CLARK R A, et al. Shotgun sequence assembly and recent segmental duplications within the human genome [J].Nature,2004,431(21):927-930.
[31] ALKAN C, SAJJADIAN S, EICHLER E E. Limitations of next-generation genome sequence assembly [J]. Nature Methods,2011,8:61-65.
[32] SCHATZ M C, DELCHER A L, SALZBERG S L. Assembly of large genomes using second-generation sequencing [J].Genome Research,2010,20:1165-1173.
[33] MEDVEDEV P, FIUME M, DZAMBA M, et al. Detecting copy number variation with mated short reads [J].Genome Res,2010,20:1613-1622.
[34] GRIMM D, HAGMANN J, KOENIG D, et al. Accurate indel prediction using paired-end short reads [J]. BMC Genomics,2013,14:132.