基因结构变异检测方法综述*

连帅彬 郭东亮 戴宪华

(中山大学信息科学与技术学院 广东 广州 510006)

摘要:在人类基因组中结构变异(SVs),拷贝数变化(CNVs),单核苷酸多态性(SNP)是非常普遍的,而且和人类健康与疾病密切相关,因此检测这些结构变异对于人类生命健康非常重要。基于第二代基因测序平台,目前已经有很多结构变异检测算法,这些算法主要分为五大类:微阵列方法、读对方法、读深方法、分裂读取方法、序列组装方法。本文系统地阐述了这五类方法的基本原理、优缺点以及使用范围,并简要介绍了每一种方法的经典检测算法及应用范围、检测性能等,并对未来检测算法的研究提出了展望。

关键词:基因测序(GS);结构变异(SV);拷贝数变化(CNV);单核苷酸多态性(SNP)

中图分类号:Q-31 **文献标识码**:A **文章编号**:1673-6273(2012)18-3577-04

Review On Algorithms Of Detecting Genome Structural Variants*

LIAN Shuai-bin, GUO Dong-liang, DAI Xian-hua

(School of Information Science and Technology of Sun Yat-sen University, Guangzhou 510006, China)

ABSTRACT: Structural variants, copy number variants and single-nucleotide polymorphism are extensive present in the human genome and closely related to the disease and health of human, so it is very important to discover these variances for the human's health. Currently, there are many structural variants detecting methods based on second generation genome sequencing platform, including micro arrays, read-pair technology, read-depth methods, split-read methods and sequence assembly. In this paper, we systematically elaborate their basic principles, advantages, disadvantages and applications respectively, meanwhile, we briefly introduce the detecting performances and applications of classical detecting algorithms based on each method and put forward some prospects of detecting algorithms in future.

Key words: Genome sequencing; Structural variants; Copy number variants; Single-nucleotide polymorphism

Chinese Library Classification(CLC): Q-31 Document code: A

Article ID: 1673-6273(2012)18-3577-04

1 引言

大量研究表明[1-5]基因组中结构变异(Structure Variants (SVs))、拷贝数变化(Copy Number Variants(CNVs))、单核苷酸多态性(Single-Nucleotide Polymorphism(SNP))是非常普遍的。基因结构变异从单个碱基变异到大段染色体的结构变化[6,7],这些结构变异最初是指长度大于 1kb 的基因片段的插入、删除或者倒置等,基于第二代测序的短读取的大量出现[8],使得基因结构变异也包含小的结构变化(比如 <1kb)。精确地检测 SVs 是非常困难的,原因在于:第一 SVs 往往存在于 DNA 的重复或复制区域;第二 SVs 的种类很多,包含:插入,删除,倒置,移位,重复等,这些都增加了检测的难度。目前结构变异检测算法所面临的最大挑战就是如何精确而快速的检测出包含小结构变化在内的所有基因组结构的变化,从而加速人类对疾病和生命进化过程的理解。而在过去的几年里,由于第二代基因测序平台的问世,使得大量的检测算法纷纷出现,而每一种算法大多都只是针对于一类特定的变异,对其他种类的变异检测效果不理想。基于此本文系统的综述了目前检测基因变异的主要方法:基于微阵列技术方法和基于测序技术。基于微阵列技术的方法主要包括两类:CGH 阵列技术和 SNP 阵列技术;基因测序技术的方法主要包括:读对技术,读深技术,分裂读取以及序列组装方法。阐述了它们的主要原理、各自的优缺点以及适用范围,并分析现有算法的局限性,对未来检测算法的研究进行了探讨和展望。

2 微阵列测序方法

微阵列方法[9](Micro arrays)已经被广泛应用于 CNV 的发现[10]和基因分型[11,12](gene typing)。微阵列方法主要有两类:基因组杂交阵列比较(Array Comparative Genomic hybridization (array CGH))和 SNP 微阵列(SNP micro arrays)。

2.1 基因组杂交阵列比较(array CGH)

Array CGH 平台是基于参考基因和测试基因杂交比较的理念[9],它的原理是:将测试基因和参考基因之间的信号比率 ratio 标准化然后转换为以 2 为底的对数函数 $\log_2$ratio,将此对数函数作为基因拷贝数变化的度量依据,对数函数 $\log_2$ratio 的增加表示拷贝数的增加,相反它的减少表示拷贝数的减少。目前全基因组 CGH 阵列平台的主要供应者是 Roche NimbleGen 和 Agilent 科技,它们的每一个微阵列可以分别产生多达 2.1M

* 基金项目:国家自然科学基金(61174163)

作者简介:连帅彬(1982-),男,中山大学博士,研究方向:生物信息处理,E-mail:shuai_lian@qq.com

郭东亮,中山大学教师,研究方向:生物信息处理,E-mail: guodl@mail.sysu.edu.cn

戴宪华,中山大学教授,博导,研究方向:生物信息处理,E-mail: issdxh@mail.sysu.edu.cn

(收稿日期:2011-10-24 接受日期:2011-11-18)

和 1M 的低核苷酸(oligonucleotides)阵列,每一个 CNV 的检测要求至少 3-10 个连续探头的信号。因此 SNP 和 CGH 微阵列可以检测出每一个基因组中的几十个到几百个结构变化。文献[13,14] 中的研究就使用了基于 CGH 阵列的 SV 检测的高分辨阵列技术,能够检测出 500bp 的 CNVs,然而代价却相对较高。CGH 阵列的另外一个重要优点就是由两个核心制造商提供的高密度阵列探头,这也使得了 CGH 阵列取代了染色体组型分析(karyotype analysis)而被广泛应用于临床诊断[4];然而 CGH 阵列技术对于 CNV 的变化敏感度较差,检测参考样本的拷贝数变化效果不理想。比如,当只有一个样本被检测出时,就很难区分这个样本究竟是参考样本中的一个丢失(loss)还是测试样本中的一个获得(gain)。

2.2 SNP 阵列

SNP 阵列也是基于杂交技术的理念,它通过比较测试样本信号密度产生一个和 CGH 阵列相似的度量函数 $\log_2$ratio。但是 SNP 阵列的度量函数 $\log_2$ratio 表明:虽然 SNP 阵列比 Array CGH 有更低的信噪比,但是 SNP 阵列技术通过采用等位特别探头而增加了 CNV 的敏感性,能提供出更多的拷贝数的综合信息[9]。SNP 阵列也广泛的应用于 CNV 的检测,早期的 SNP 阵列显示在 CNV 区域的低覆盖,但是最近的阵列(Affymetrix6.0SNP, IIIumina 1M 平台)融合了更好的 SNP 选择标准[11-15],在文献 16 中通过结合 CGH 阵列和 SNP 平台来提供 CNV 检测的更高的可信度。

阵列技术已经被广泛应用于结构变异的检测,它在通量和费用上都有很多优点。在整个人类基因组中大 CNV 是很少的,研究表明在人类基因组中 CNVs>500kbp 的仅仅占 8%。由于 CGH 阵列和 SNP 平台的低费用以及大量的公共 SNP 数据的可用性,微阵列数据已经成为为分析大数据集的 CNV 的有效方法。另外微阵列技术自身也有一些缺点:(1)微阵列技术只局限于检测序列拷贝数的不同,不能提供拷贝数变化的具体位置,比如断点信息等;(2)阵列技术对于单拷贝数的增加不敏感[11,17];(3)在多重复和复制区域使用杂交试验,CGH 阵列和 SNP 平台都是假定参考基因的每一个位置都是二倍体的,但是在重复序列里这个假设是不成立的,因为 CNVs 和重复片段有很强烈的正相关而且很多断点都坐落在重复区域[9,11,13,18]。

3 基于测序的检测方法

下一代基因测序技术[19-21](Next Neneration Sequencing Technologies(NGS))的出现为基因结构变化的研究带来了革命性的变化,已经取代微阵列技术作为 SVs 发现和分型的一个平台。这类方法主要包括读对、读深、分裂读取、序列组装等,这些方法都是将读取映射到参考基因然后进行比对从而发现不同的结构变异及其类型。

3.1 读对技术(Read-pair Technologies)

读对技术就是通过评价配对跨度以及读取方向,并且将配对跨度和读取方向与参考基因比对,将不一致的读对聚类。这种方法可以检测出大多数的变异,因此已经得到了广泛的应用,基于读对的算法有很多,比较经典的有:PEM[22],PEMer[23],HYDRA[24,25]等。

配对末端映射 PEM(Paired-end Mapping):主要涉及到长度等于或大于 3kb 的末端配对的隔离、准备以及基于 454 平台的大规模测序数据等。首先将大量的配对读取映射到人类基因组上,然后确定出由末端配对读取与对应的参考基因不一致的区域,从而预测基因结构变异。预测过程需要 5 个标签:删除、简单插入、配对插入、倒置、无配对插入。为了消除由于连接反应过程中形成的畸形结构所导致的误判,因此对于每一个结构变异类型都要求至少两个配对读取来支持。这种方法能够检测到大于 3 kb 的删除、倒置、配对插入、无配对插入以及 2 kb-3 kb 的简单插入。

PEMer:通过比较映射读对和基因组中平均插入大小的距离来检测插入和删除。主要算法步骤为:(1)构建预处理(construct pre-processing):将 PEM 数据设计成合理的结构;(2)读取排列(read-alignment):排列读取的两个末端到参考基因上;(3)最优配对位置(optimal paired-end placement):将映射配对末端与配对末端相结合;(4)孤立子确定(outlier-identification):确定孤立子配对末端,孤立子就是指读取末端映射到参考基因时由于有一定的距离而导致了和期望读对末端的偏离或者读取末端匹配到不同的染色体上。(5)孤立子聚类(outlier clustering):如果一个含有 N 个独立配对末端的集类与一个单一 SV 一致的话,那么孤立子被分类为 SVs。PEMer 评价了在一个集类中是否所有的配对末端都表示同一个变异类型。(6)聚类整合(cluster merging):有相同结构变异的集类被整合到一个单一集类中,当 SVs 在不同的的截断值以及不同的集类大小的情况下被并行搜索时,集类整合是非常必要的。PEMer 算法能敏感的检测出小于 1kb 的删除并且确定小片段的断点,但是不能确定多重复复杂区域的 SVs,同时当插入长度大于平均插入时算法也会失效。

HYDRA 也是一种基于读对的 SV 检测算法。它的原理就是在一个或者多个映射中使用启发式方法(heuristic approach)确定 SVs。HYDRA 通过比较相互之间不一致的映射并且确定配对集类,这些确定变异类型的映射集类都是来自于一个独立的 DNA 片段。当两个映射跨越相同的基因区域并且有相同的读取方向时,就认为彼此相互 " 支持 "。因此对于每一个预测的变异类型,HYDRA 都检查支持的映射数目并且选择被大多数其他映射支持的单映射。为了减少和单映射的重叠,其余映射都被整合到变异呼叫中。这个过程能使得定义变异类型的相互支持的映射数目最大化。每一个配对仅支持一个单变异类型,当多个变异存在时,HYDRA 就选择最多映射支持的那个作为变异类型。HYDRA 减少了对结构变异的假设,增加了算法敏感度。因此不仅能检测出删除、重复、倒置、任意长度的插入以及移位等,而且还能检测一个或者多个重复序列的断点。

3.2 读深方法(Read-depth Methods)

读深的方法通过假定映射深度是一个随机分布(泊松分布或者矫正泊松分布),然后利用映射深度的随机分布来检测序列样本的重复和删除。基本思想是利用映射深度和参考基因进行比对,重复区域显示高的读取深度,而删除区域则显示出低的读取深度。基于读深的算法也有很多,比较经典的有 EWT[26] 算法,CNVnator[27,28]算法。

EWT 算法是基于读取深度加窗的显著性测试算法。基本

思想是：确定连续窗口内读取深度的增加和减少，以此作为基因拷贝数的定量测量，首先在 100bp 窗口内使用 GC 矫正读取深度，多个连续窗口的覆盖的增加和减少就表明拷贝数的增多和减少。EWT 算法能够检测出 99.9%的删除（>1000bp），并且整体误测率被显著性测试控制并随多测试而调整的。和 PEM 算法相比，EWT 算法检测的变异类型在重复片段是非常丰富的，而且在重复片段的复杂区域很难使用 PEM 算法确定 CNV，因为这些区域的很多读取不是唯一的映射到参考基因中。EWT 算法和 PEM 算法是互补的，因此两种算法结合将会增强基于下一代基因测序数据的 SVs 检测精度。

CNVnator 也是基于读取深度的 CNV 检测算法。为了计算读深信号，它将整个基因组分成大小不等且不重叠的区域并且使用每一个小区域内映射数目作为读取深度信号。然后利用移均值技术[29](mean-shift technique)把这些信号分区成包含潜在 CNVs 的小片段，最后在每个小片段上应用统计显著性测试来预测 CNVs。这种方法具有很高的敏感度（86%-96%），低误测率(false-discovery rate)（3 %-20 %），高精确分型率(high genotyping accuracy)（93 %-95 %）以及高断点分辨率(high resolution in breakpoint discovery)（90 %）的特点。而且 CNVnator 是对分裂读深取方法和读对方法的一个直接补充，它能检测出半数以上分裂读取和读对方法不能检测到的 CNVs[28]。

3.3 分裂读取方法（Split-read Methods）

分裂读取方法能够检测删除与小的插入而被首次应用到长 Sanger 序列读取中[27]。分裂读取的目标就是要定义出结构变异的断点，而且只要读取足够长(>400bp)的话这种方法还可以检测出移动插入[27](Mobile-Element Insertions(MEIs))。但是基于下一代基因测序（NGS）的大量短读取，大大增加了排列的难度从而限制在分裂读取方法的应用。而 Pindel[30]算法运用末端配对读取技术减少了分裂读取的搜索空间，从而减少了排列短序列到参考基因的复杂度。

Pindel：就是利用模式增长算法搜索最大 - 最小子串的方法来确定大删除（1 bp-10 kb）与中等长度的插入（1-20 bp），并计算出精确的断点。模式增长算法举例：比如定义一个基因序列数据为 S=ATCAAGTATGCTTACGC 以及模式 P=ATGCA，利用模式增长算法搜索出的最大最小子串分别为：ATGC 和 ATG。Pindel 的主要思想就是首先应用 SSAHA2[31]方法把所有读取映射到参考基因，然后检查映射结果确保这些配对读取只有一端能被映射。对于每一个读对，映射端必须唯一精确的映射到基因中，而另一端在给定阈值 s 的条件下不能被映射到基因中的其他位置。Pindel 使用映射端来确定参考基因上的锚点以及无映射读取的的搜索方向。锚点和无映射读取搜索方向确定之后，用户就可以定义最大删除参数（Max_D_Size），从而参考基因中的一个子区域就可以被确定下来，Pindel 将无映射读取打断为 2 个删除或者 3 个插入，然后分别映射这些片段的两个终端。该算法能够精确有效的检测出大删除和中等插入的断点信息。

3.4 序列组装方法（Sequence Assembly）

理论上如果可以 de novo 组装的话，所有形式的结构变异都能够精确的被检测出来。然而序列组装的方法刚刚起步，并且下一代测序技术产生了大量的短读取序列(30bp)，这对于整个基因组的 de novo 组装是个严峻挑战。应用传统组装方法需要找出所有重叠增加了组装难度。另外，理论上配对读取更容易组装，但实际上配对读取组装比无配对读取组装更加复杂。目前有很多结合 de novo 组装和局部组装的算法。比如：ABYSS[32]，ALLPATHS[33,34]等。

ABYSS：是一种并行的序列组装算法。它最大的创新就是利用 de Bruijn 图，从而允许组装算法在多台计算机上并行计算。ABYSS 算法过程的两个阶段：(1)初始化重叠群(contigs)：从序列读取中产生所有可能的长度为 k 的子串，消除 k-mer 数据集的读取误差从而构建初始化重叠群(contigs)。(2)扩展重叠群：应用配对信息通过消除重叠群区域的重叠模糊性来扩展重叠群。这个算法能够精确快速的组装大量的短读取序列，文献[32]通过精确组装由 NA18507 的整个基因序列产生的 35 亿个短读取序列证实了该算法的组装能力。

ALLPATHS：算法利用图论中全通路理论进行组装。全通路方法保留了由于数据集限制以及二倍体基因多态性而产生的内在相关信息，从而实现了对短读取的良好组装。ALLPATHS 算法的两个核心概念：(1) 找出跨越一个给定读对的所有通路，即从一个读取到另一个读取的所有序列都被其他序列所覆盖；(2)采用配对将基因组分成小区域然后利用分别组装的方法来确定位点。因此，ALLPATHS 组装算法能够获取基因数据的精确信息，这种方法不仅对短读取有效，对其他类型的 DNA 序列也同样有效。随着第三代基因测序技术高通量、长读取特点的出现，将会减少重叠从而降低组装算法的复杂度，使得组装更加容易和准确。另外，利用高深度覆盖的杂交组装技术也能够很有效地提高组装的连续性，从而解决重复区域的问题。

3.5 基于测序算法的普遍局限性

基于第二代基因测序 SVs 检测算法的最大问题就是基因数据本身，因为第二代基因测序平台产生大量的短读取序列(30 bp 左右)，这会导致更多的重复和复制出现，因此需要长读取和长插入通过增加读取映射的特异性来矫正偏差。另外一个缺点是：对这些短读取数据的分析和存储需要大量的投资和资源，因此提高短读取数据的分析和存储的效率也迫在眉睫。这四种基于测序的方法没有一种方法是万能的，每一种方法都有不同的侧重点，它们都依赖于变异的类型或者序列的特性。尽管读深的方法是唯一能够精确预测拷贝数的方法[35]，但是不能提供详细的断点信息。读对方法虽然可以检测出多种结构变异但是很难解决重复区域的模糊映射分配问题，而且 SV 断点的精确预测依赖于片段大小分布。分裂读取方法能够检测多种结构变异并且有很高的分辨率，但是目前分裂读取方法仅仅在基因组特定区域内可靠。序列组装方法通过实施成对基因组比较技术被认为是最有效的方法，但是在重复和复制区域由于组装失败而导致算法对于重复和复制严重偏离[36]。基于这样一种现实，目前已经有一些算法通过整合多种方法来提高敏感性和特异性。比如：SVMERGE[37]，SPANNER[27]，CNVer[38]等它们是通过结合读对和读深方法来提高 CNV 检测可靠性的。

4 总结与展望

基因结构变异(SVs)，拷贝数变化(CNVs)以及单核苷酸多

态性(SNPs)在人类基因组中非常普遍,已经成为生物医学领域研究的热点,因此精确地发现和检测这些变异能够加速人类对于生命过程以及重大疾病的理解。第二代基因测序平台的问世和大规模应用为这些检测算法提供了极大的发展空间。目前检测基因变异的方法大致有五类:微阵列技术、读对技术、读深技术、分裂读取技术、序列组装等。这些方法每一种都又包含很多经典的算法,每一种算法都有各自的优缺点和应用范围,使用相同的 DNA 样本,不同的实验方法和数据分析方法检测出的结构却有很低的重叠,因此目前最严重的挑战就是缺少一个评价这些算法的黄金标准。这些问题都要寄希望于基因测序技术的不断发展。

我们预测第三代基因测序技术将会体现出高通量、长读取、低费用的特点,读取长度将会是大于 1kb,这些优点不仅使得小结构变异检测更加容易,而且能更好的确定结构变异断点的信息,同时也改善了断点分辨率、拷贝数精度等。但是检测基因结构变异不是一蹴而就的,目前基于基因测序技术的检测方法在重复区域显示低敏感度,然而在这些区域却显示出更多的变异类型[39,40]。随着测序技术的不断成熟,新的方法将会被不断的提出改进结构变异的预测效果。特别地,未来更有效的检测算法必将融合多种方法于一体,这些方法应该将会有更好的检测效果,能够提供更多的断点分辨率,因此这个领域依然有很大的发展空间。

参 考 文 献 (References)

[1] Lee, S., Hormozdiari, F., Alkan, C., et al. Detecting small indels from clone-end sequencing with mixtures of distributions[J]. Nat.Methods, 2009, 6: 473-474

[2] McCarroll, S.A. Integrated detection and population-genetic analysisof SNPs and copy number variation[J]. Nat. Genet, 2008, 40: 1166-1174

[3] Kidd, J.M. Mapping and sequencing of structural variation from eighthuman genomes[J]. Nature, 2008, 453: 56-64

[4] Cooper, G.M., Zerr, T., Kidd, J.M., et al. Systematic assessment of copy number variant detection via genome-wide SNP genotyping[J]. Nat. Genet, 2008, 40: 1199-1203

[5] 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

[6] Iafrate, A. J. Detection of large-scale variation in the human genome [J]. Nature Genet, 2004, 36: 949-951

[7] Redon, R. Global variation in copy number in the human genome[J]. Nature, 2006, 444: 444-454

[8] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing [J]. Nature, 2010, 467: 1061-1073

[9] Can Alkan, Bradley P. Coe and Evan E. Eichler. Genome structural variation discovery and genotyping. Nature reviews [J]. Genetics, 2011, 12 (5): 363-376

[10] Locke, D. P. BAC microarray analysis of 15q11-q13 rearrange- ments and the impact of segmental duplications[J]. J. Med. Genet, 2004, 41: 175-182

[11] Itsara, A. Population analysis of large copy number variants and hotspots of human genetic disease [J]. Hum. Genet., 2009, 84: 148-161

[12] Sebat, J. Large-scale copy number polymorphism in the human genome[J]. Science, 2004, 305:525-528

[13] Conrad, D. F. Origins and functional impact of copy number variation in the human genome[J]. Nature, 2010, 464: 704-712

[14] Park, H. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing[J]. Nature Genet, 2010, 42: 400-405

[15] Winchester, L., Yau, C., Ragoussis, J. Comparing CNV detection methods for SNP arrays [J]. Brief. Funct. Genomic Proteomic, 2009, 8: 353-366

[16] Gusev, A. Whole population, genome-wide mapping of hidden relatedness[J]. Genome Res, 2009, 19: 318-326

[17] Coe, B. P. Resolving the resolution of array CGH [J]. Genomics, 2007, 89: 647-653

[18] Bailey, J. A. Recent segmental duplications in the human genome[J]. Science, 2002, 297: 1003-1007

[19] Wheeler, D. A. The complete genome of an individual by mas- sively parallel DNA sequencing[J]. Nature, 2008, 452:872-876

[20] Bentley, D. R. Accurate whole human genome sequencing using reversible terminator chemistry[J]. Nature, 2008, 456: 53-59

[21] McKernan, K. J. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding[J]. Genome Res, 2009, 19: 1527-1541

[22] Jan O. Korbel1, Alexander Eckehart Urban. Paired-End Mapping Reveals Extensive Structural Variation in theHuman Genome [J]. Science, 2007, 318(5849): 420-426

[23] Korbel, J. O. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data [J]. Genome Biol, 2009,10: R23: 1-14

[24] Quinlan, A. R. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome [J]. Genome Res, 2010,2 0: 623-635

[25] Chen, K. Break Dancer: an algorithm for high-resolution mapping of genomic structural variation[J]. Nature Methods, 2009, 6: 677-681

[26] Yoon, S., Xuan, Z., Makarov, V., et al. Sensitive and accurate detection of copy number variants using read depth of coverage [J]. Genome Res, 2009,19:1586-1592

[27] Mills, R. E. Mapping copy number variation at fine scale by population scale genome sequencing[J]. Nature, 2011, 470: 59-65

[28] 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 Res, 2011, 21: 974-984

[29] Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis [J]. IEEE Trans Pattern Anal Mach Intell, 2002, 24: 603-619

[30] Ye, K., Schulz, M. H., Long, Q. 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

[31] Ning,Z. SSAHA: a fast search method for large DNA databases[J]. Genome Res, 2001, 11: 1725-1729

(下转第 3588 页)

2002,527:309-314

[8] Henrich S, Lindberg I, Bodew, et al. Proprotein convertase models based on the crystal structures of furin and kexin: explanation of their specificity [J]. J Mol Biol, 2005,345(2):211-227

[9] Thomas G. Furin at the cutting edge: from protein traffic to embryogenesis and disease [J]. Nat Rev Mol Cell Biol, 2002,3(10):753-766

[10] Anderson E D, Molloy S S, Jean F, et al. The ordered and compartment-specfific autoproteolytic removal of the furin intramolecular chaperone is required for enzyme activation [J]. Biol Chem, 2002, Apr 12;277(15):12879-12890

[11] Shapiro J, Sciaky N, Lee J, et al. Localization of endogenous furin in cultured cell lines [J]. Histochem Cytochem, 1997,45(1):3-12

[12] Molloy SS,Thomas L,VanSlyke JK,et al. Intracellular trafficking and activation of the furin proprotein convertase: localization to the TGN and recycling from the cell surface[J]. EMBO, 1994, 13(1):18-33

[13] Rockwell NC,Thorner JW.The kindest cuts of all: crystal structures of Kex2 and furin reveal secrets of precursor processing [J]. Trends-Biochem Sci, 2004,29:80-87

[14] Henrich S, Cameron A, Bourenkov GP, et al. The crystal structure of the proprotein processing proteinase furin explains its stringent specificity [J]. Nat Struct Biol, 2003,10(7):520-526

[15] Siezen RJ, Creemers JW, Van de Ven WJ. Homology modelling of the catalytic domain of human furin. A model for the eukaryotic subtilisin-like proprotein convertases [J]. Eur J Biochem, 1994,222(2): 255-266

[16] Wise RJ, Barr PJ, Wong PA, et al. Expression of a human proprotein processing enzyme: correct cleavage of the von Willebrand factor precursor at a paired basic amino acid site [J]. Proc Natl Acad Sci USA, 1990,87(23):9378-9382

[17] Van de Ven WJ, Voorberg J, Fontijn R, et al. Furin is a subtilisin-like proprotein processing enzyme in higher eukaryotes [J]. Mol Biol Rep, 1990,14(4):265-275

[18] Bresnahan PA, Leduc R, Thomas L, et al. Human fur gene encodes a yeast KEX2-like endoprotease that cleaves pro-beta-NGF in vivo [J].J Cell Biol,1990, 111(6 Pt2):2851-2859

[19] Lee R, Kerman IP, Teng KK, et al. Regulation of cell survival by secreted proneurotrophins [J]. Science, 2001,294(5548):1945-1948

[20] Vardar D, North CL, Sanchez-IR, et al. Nuclearmagnetic resonance structure of a prototype Lin12-Notch repeat module from human Notch1 [J]. Biochemistry, 2003,42(23):7061-7067

（上接第 3580 页）

[32] Simpson, J. T. ABySS: a parallel assembler for short read sequence data[J]. Genome Res, 2009, 19: 1117-1123

[33] Jonathan Butler, Iain MacCallum, Michael Kleber, et al. ALLPATHS: De novo assembly of whole-genome shotgun microreads [J]. Genome Res, 2008,18: 810-820

[34] Li, R. De novo assembly of human genomes with massively parallel short read sequencing[J]. Genome Res, 2009, 20: 265-272

[35] Sudmant, P. H. Diversity of human copy number variation and multicopy genes[J]. Science, 2010, 330: 641-646

[36] Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly[J]. Nature Methods, 2011, 8: 61-65

[37] Kim Wong, Thomas M Keane, James Stalker, et al. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly[J]. Genome Biology, 2010, 11, R128:1-9

[38] Medvedev, P., Fiume, M., Dzamba, M., et al. Detecting copy number variation with mated short reads [J]. Genome Res, 2010, 20: 1613-1622

[39] Kim, P.M. Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history [J]. Genome Res, 2008,18: 1865-1874

[40] Cooper, G.M., Nickerson, D.E. & Eichler, E.E. Mutational and selective effects on copy-number variants in the human genome [J]. Nat Genet, 2007, 39:S22-S29