高通量测序在遗传病临床基因检测中的效果评价

张彦[1,2] 吴柏林[2,3*]

(1. 广东省妇幼保健院医学遗传中心,广东广州 511442; 2. 哈佛大学医学院波士顿儿童医院实验医学部; 3. 复旦大学儿科医院、复旦大学生物医学研究院,上海 200433)

【摘要】 基于国内外迅猛发展的高通量测序技术(第二代测序技术)应用以及精准医学的需求,从技术和成本角度分析高通量测序技术用于同时检测单核苷酸变异(SNV) 和拷贝数变异(CNV) 的临床应用价值。

【关键词】 高通量测序;拷贝数变异;单核苷酸变异;基因诊断

【中图分类号】 R714.55 【文献标识码】 A

高通量测序技术(high throughput sequencing)也叫做第二代测序或下一代测序(next generation sequencing,NGS),由于其较高的检测效率而得到越来越多的推广和应用。但由于技术发展的局限和生物信息学的限制,临床应用尚存在着较多的结果可重复性问题和数据解读困难。本文就高通量测序技术在临床应用方面的实用性做一综述,拟对其同时检测拷贝数变异(copy number variation,CNV)和单核苷酸变异(single nucleotide variation, SNV)的效能进行客观评价。

1 NGS 简介

NGS 技术的发展目前大致可归结为 3 个方向。第一个方向可追溯到 1990 年的大规模平行测序技术 (massively parallel signature sequencing, MPSS)[1],由于其技术极其复杂而使其应用受限。2004 年与 Solexa 公司合并,而后者是基于 1998 年出现的另一技术-可逆染料终止技术(reversible dye-terminator technology)[2],后来被 Illumina 公司兼并,最终发展为现在的边合成边测序技术;第二个方向起始于 2005 年的聚合酶克隆测序技术(polony sequencing)[3],后来整合入 Applied Biosystems 公司的 SOLiD 平台,并最终相继被 Life Technologies 和 Thermo Fisher Scientific 公司并购,发展为现在的半导体测序技术;第三种为焦磷酸测序技术,出现于 2005 年,后被 Roche Diagnostics 公司并购。

还有一些其他高通量测序技术得到过推广,比如纳米球测序技术(DNA nanoball sequencing)[4]被 Complete Genomics 公司利用开发,后被华大基因并购;单分子荧光测序技术由 Helicos Biosciences 公司开发,但该公司很快破产;单分子实时测序(single molecule real time sequencing)[5]由 Pacific Biosciences 公司开发;另外,有些高通量测序技术仍处于研发阶段或初步使用,如纳米孔测序技术(Nanopore DNA sequencing)、微流控 Sanger 测序等[6,7]。

就目前已经在临床应用的几种高通量测序技术,按其技术原理和优缺点简单归纳如表 1。

2 NGS 同时检测 SNV 和 CNV 的临床应用尝试

自从高通量测序技术诞生开始,科学界对于各种遗传变异的检测要求就在不断提高,包括大到染色体数目和结构,基因组微缺失微重复等的变异(如 21-三体综合征、猫叫综合征)、不同长度片段拷贝数/CNV 的变异(如 α 地中海贫血、DMD),小到单个

doi: 10.13470/j.cnki.cjpd.2016.04.002

*通讯作者:吴柏林,E-mail:bai-lin.wu@childrens.harvard.edu

表1 4种高通量测序技术比较

方法	原理	读长	准确度(%)	应用范围	公司	优点	缺点
半导体测序	核苷酸合成过程中氢离子变化	<400bp	98.0	人基因组及其他	Life Technology	仪器相对便宜	检测in/del准确度较差,特定序列错读率高
边合成边测序	检测寡核苷酸片段合成过程中每种单核苷酸上的荧光标记	50～500bp	99.9	人基因组及其他	Illumina	针对各种检测需求的检测平台	设备和试剂昂贵
单分子实时测序	利用纳米孔连续监测单分子核苷酸链合成的荧光信号	10～40Kb	87.0	宏基因组及其他	Pacific Bioscience	长片段测序,减少拼接错误	设备昂贵,通量相对较低
焦磷酸测序	通过寡核苷酸链合成过程中释放的焦磷酸结合荧光信号进行序列读取	700bp	99.9	宏基因组	Roche	读长较长	运行成本高,特定序列错读率高

或数个碱基对/SNV的变异(如结节硬化症、成骨不全症)都可以导致遗传病。因此在同一平台同时检测各种致病基因突变(检测SNV)或基因组失衡(检测CNV)就成为遗传病基因诊断的必然要求,尤其对于罕见遗传病。

目前的高通量测序技术对于检测单个或少数几个核苷酸变异(SNV)一般不存在太多问题,各大主流测序平台(如Proton和Nextseq 550等)均可以检测外显子组或人基因组,检测准确性一般均可达到99%以上。对于拷贝数变异(CNV)和染色体微小结构变异(structural variation, SV)等的检测却不尽如人意,主要问题出现在成本效益比太低、特定染色体区域的捕获效率较差,以及短读长拼接错误。针对不同检测方案的检测效率大致可以分为如下几类:

2.1 基于外显子组测序(whole exome sequencing, WES)的尝试 针对目前已知的约2万个基因的所有编码外显子及两端侧翼序列进行捕获后测序。该方案优点是成本相对较低,而且将绝大多数已知致病基因的突变位点囊括在内,可以一次性将大部分已知疾病和未知疾病进行检测。缺点之一是对某些同源性较高的基因序列(如假基因和基因簇等)、高GC含量的基因序列以及基因动态突变(如亨廷顿舞蹈病的致病基因*HTT*等)等情况应对困难;另一缺点就是几乎无法准确检测到CNV以及SV,因为不连续的捕获将本身短读长的缺点加剧了,对于微小片段缺失或重复几乎不具备检测能力。虽然有许多研究报道用外显子组测序结合特定的算法可以有效地检测到某些较大的CNV[8-10],但总体而言,检测所有已知的和未知的不同大小的CNV,WES目前并不具备有效检测能力[11]。

2.2 基于特定基因群(gene panel)的尝试 不少的研究针对特定的染色体区域或者某些特定的基因组合设计测序方案。由于测序针对性强,测序范围一般较小,因此成本相对较低,可以兼顾编码区和非编码区[12]。但此种模式仅能针对性地检测某些疾病,无法大面积覆盖已知或未知遗传病,而且仍然需要解决特定区域的捕获效率问题,同样需要面对WES在特殊基因序列检测的困难[13]。

2.3 基于全基因组(whole genome sequencing, WGS)的尝试 针对全基因组的测序是理论上解决同时检测CNV和SNV的最有效方案,但是目前各种研究报道并不太多,原因主要包括:①成本太高,目前数据认为,大于20X的数据为可信数据,而30X以上比较均一的全基因组测序费用目前仍然需要数千美金,成本高昂[14];②特定区域数据有效性差,如着丝粒区、端粒区等异染色质区域和某些高GC含量区属于测序盲区,数据可信度较差。据目前研究显示,约有16%的基因组序列属于难测区域[15]。基于此,虽然有部分机构用WGS的平台检测CNV,但大多为低深度测序(低于10X),可信度相对较差。因此,国家食品药品检定研究院近期发布了《第二代测序技术检测试剂质量评价通用技术指导原则》,明确规定了"原则上全基因组测序一般不用于临床检测,除非有充分的技术适用性验证报告"[16]。

3 展望

尽管高通量测序作为目前最有可能实现同时检测SNV和CNV的技术手段,其对临床遗传病基因

检测的大规模广泛应用尚待时日。整合 SNV 和 CNV 乃至 SV 于同一平台检测是技术发展的必然趋势,从成本和技术原理角度,大致有两方面的发展趋势可以预测。

3.1 成本下降后的 WGS 目前全基因组测序是技术上最接近单一平台同时检测 SNV 和 CNV 的技术,但成本和读长是其受限因素。随着测序成本的继续降低,读长短和拼接错误多将成为限制目前几种常用的第二代测序(Life、Illumina 等)技术实现有效临床检测的主要瓶颈。

3.2 第三代长片段测序的应用 目前已有不少研究机构尝试单分子测序,即俗称的第三代测序技术。相较于第二代测序技术,其最大的优势在于读长的明显加长。第二代测序技术的单片段读长一般不超过 500bp,但目前可查到的第三代测序技术平台,如 Pacific Bioscience 一般均为 kb 级,另有其他机构宣称研发的测序方法平均读长已可达 200kb,如 Nanopore。

显著延长的读长可以极大地减少第二代测序技术的拼接错误,而且可以避开基因组复杂结构对捕获和测序的影响,从而大幅度地提高检测准确性。虽然各种第三代测序技术还有各种各样的问题,但假以时日仍有较大的发展空间。

在精准医学和出生缺陷防控的政策支持和不断提高的优生优育的需求推动下,整合 SNV、CNV 甚至 SV 检测技术至同一平台是大势所趋。在儿科领域,诸多遗传病,尤其是罕见遗传病存在诊断困难,目前的遗传学检测往往需要结合多种技术进行(如高通量测序和芯片等);在产科和产前诊断领域,针对孕期超声异常的胎儿的遗传学原因排查也常需要多种方法结合进行(如核型分析、芯片等)。如能将 SNV、CNV 和 SV 检测在同一平台进行,将明显提高检测效率、减少患儿/胎儿父母等待时间、减少沟通成本、节省检测成本以及缓减实验室压力,极大地有利于出生缺陷的防控。

参 考 文 献

[1] Brenner S, Johnson M, Bridgham J, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays[J]. Nat Biotechnol, 2000, 6: 630-634.

[2] Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry[J]. Nature, 2008, 7218: 53-59.

[3] Shendure J, Porreca GJ, Reppas NB, et al. Accurate multiplex polony sequencing of an evolved bacterial genome[J]. Science, 2005, 5741: 1728-1732.

[4] Drmanac R, Sparks AB, Callow MJ, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays[J]. Science, 2010, 5961: 78-81.

[5] Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules[J]. Science, 2009, 5910: 133-138.

[6] Ammar R, Paton TA, Torti D, et al. Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes[J]. F1000 Res, 2015, 4:17.

[7] Kan CW, Fredlake CP, Doherty EA, et al. DNA sequencing and genotyping in miniaturized electrophoresis systems[J]. Electrophoresis, 2004, 25(21-22): 3564-3588.

[8] D'Aurizio R, Pippucci T, Tattini L, et al. Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2[J]. Nucleic Acids Res, 2016. [Epub ahead of print]

[9] Fromer M, Purcell SM. Using XHMM software to detect copy number variation in whole-exome sequencing data[J]. Curr Protoc Hum Genet, 2014, 81:7.23.1-21.

[10] Jiang Y, Oldridge DA, Diskin SJ, et al. CODEX: a normalization and copy number variation detection method for whole exome sequencing[J]. Nucleic Acids Res, 2015, 6: e39.

[11] Hong CS, Singh LN, Mullikin JC, et al. Assessing the reproducibility of exome copy number variations predictions[J]. Genome Med, 2016, 1: 82.

[12] Zhang X, Ma D, Zou W, et al. A rapid NGS strategy for comprehensive molecular diagnosis of Birt-Hogg-Dube syndrome in patients with primary spontaneous pneumothorax [J]. Respir Res, 2016, 1: 64.

[13] Wang Y, Yang Y, Liu J, et al. Whole dystrophin gene analysis by next-generation sequencing: a comprehensive genetic diagnosis of Duchenne and Becker muscular dystrophy[J]. Mol Genet Genomics, 2014, 5: 1013-1021.

[14] Plothner M, Frank M, von der Schulenburg JG. Cost analysis of whole genome sequencing in German clinical practice [J]. Eur J Health Econ, 2016. [Epub ahead of print]

[15] Telenti A, Pierce LC, Biggs WH, et al. Deep sequencing of 10,000 human genomes[J]. Proc Natl Acud Sci U S A, 2016,113(42):11901-11906.

[16] 中国食品药品检定研究院. 第二代测序技术检测试剂质量评价通用技术指导原则[EB/OL]. 2016. http://www.nicpbp.org.cn/CL0149/8495.html

(收稿日期:2016-09-21)

编辑:熊诗诣