文章编号：1004-0374(2012)01-0095-05

· 技术与应用 ·

外显子组测序技术在人类疾病研究中的应用

林正伟，赵晓航*

(中国医学科学院/北京协和医学院肿瘤研究所，分子肿瘤学国家重点实验室，北京 100021)

摘　要：外显子组测序是针对基因组中的蛋白质编码区，靶向富集外显子区域测序，以发现疾病相关遗传变异的技术。该技术近年越来越多地应用于发现人类基因组低频变异、鉴定单基因遗传病致病基因和肿瘤等复杂疾病易感基因研究，成为人类疾病相关变异研究的重要工具。综述了外显子组测序技术的基本原理及其在人类疾病相关基因研究中的应用。

关键词：外显子组测序技术；遗传疾病；易感基因

中图分类号：R394; R446.9　　**文献标志码：**A

Application of exome sequencing in research of human diseases

LIN Zheng-Wei , ZHAO Xiao-Hang*

(State Key Laboratory of Molecular Oncology, Cancer Institute, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100021, China)

Abstract: Whole exome sequencing (WES), which focuses on the protein coding region of the whole genome, aims to identify diseases related coding variants through sequencing the enriched exons of the genome. In recent years, WES is increasingly applied in several research fields, such as to identify novel low frequency variants, find causal genes of single gene inheritance diseases, in research of complex diseases like cancer. It has become one of the most important strategies in identifying diseases associated genes. Here we reviewed the basic technical principal of WES and its application in research of human diseases related genes.

Key words: exome sequencing; genetic disease; predisposing genes

目前，运用高通量二代测序技术已完成多种动植物和微生物的全基因组序列测定。但由于全基因组测序技术成本较高[1]，近年主要应用外显子组测序 (whole exome sequencing, WES) 技术开展单基因遗传病致病基因和复杂疾病易感基因的鉴定研究。

1　外显子组测序

外显子 (exon) 是真核生物基因的一部分，可翻译成相应的氨基酸，人类基因组大约含 180 000 个外显子 (约 30 Mb)，占人类基因组 1%，全部外显子称为外显子组 (exome)[2-3]。外显子组测序一般先应用靶向富集技术捕获基因组的外显子区域，再通过高通量测序获得编码区序列，进而发现疾病遗传特征[3]。人类大约 85% 的致病突变位于基因组的蛋白编码区[4]。

1.1　靶向富集

针对基因组中整个蛋白质编码区域的靶向富集称为外显子组捕获[5]。常用靶向富集方法有聚合酶链式反应 (polymerase chain reaction, PCR)、分子倒位探针 (molecular inversion probes, MIP)、芯片杂交

收稿日期：2011-07-29；修回日期：2011-09-11

基金项目：国家自然科学基金项目(91029725，81021061)；国家重点基础研究发展计划(“973”项目)(2011CB910703)；国际科技合作与交流专项(2008DFA31130)

*通信作者：E-mail: zhaoxh@cicams.ac.cn; Tel: 010-67709015

捕获 (on-array hybrid capture) 和液相捕获 (in-solution hybrid capture) 等 [6-7](表 1)。

1.2 高通量测序

将靶区域富集后再进行高通量测序，目前的高通量测序平台原理主要包括焦磷酸测序、可逆荧光终止和双碱基编码几种，分为边合成边测序 (sequencing by synthesis, SBS) 和边连接边测序 (sequencing by ligation, SBL) 模式 [9]。不同平台的读长、单位时间读取数据量、技术原理和仪器成本等方面具有不同特点 (表 2)。

2 外显子组测序与疾病相关基因研究

继人类基因组计划后，研究者开始运用高通量技术开展遗传变异与人类疾病的关联研究，目前主要涉及全基因组关联分析 (genome-wide association study, GWAS)、外显子组测序和全基因组测序等研究策略。基于全基因组单核苷酸多态性 (single nucleotide polymorphism, SNP) 芯片的 GWAS 技术流程和数据分析相对成熟，但芯片上的杂交探针是基于已知 SNP 位点设计，不能发现新的多态性位点 [6]。理论上，全基因组测序可以鉴定人类基因组序列的所有变异 [10]，但因目前成本较高，暂未得到广泛应用。外显子组测序在一定程度上解决了杂交分型芯片和传统测序技术的局限性，相对成本低，可以详细分析编码区域的遗传变异，成为目前疾病相关基因研究的重要工具 [11-12]。

表1 常用靶向富集方法比较[8]

	聚合酶链式反应	分子倒位探针	芯片杂交捕获	液相捕获
成本	高	高 (<10个样本) 低 (>100个样本)	中	中 (<10 个样本) 低 (>10 个样本)
DNA用量 (μg)	~8/Mb	0.2	10~15/30 Mb	3/30 Mb
敏感度 (%)	>99.5	>98	>98.6	>99.5
特异度 (%)	>70	>98	>70	>80
覆盖度 (%)/重数(×)	80/2	88/100	60/0.5~1.5	61/ 0.5~1.5

表2 二代测序平台比较

	Roche/454 GS FLX	Illumina/Solexa Genome Analyzer Ⅱ	Applied Biosystem/SOLiD 4
仪器成本	低	中	高
读长 (bp)	450	35~300	35~100
通量 (Gb)/天	1	6.5 (2×100 bp)	6
准确度(%)	>99	>90 (2×50 bp)	>99
技术原理	焦磷酸测序	可逆荧光终止法	双碱基编码
测序模式	SBS	SBS	SBL
PCR模式	油包水	桥式	油包水

目前发现与人类疾病相关的基因突变位点已超过10万个，但仍为人类基因组遗传变异的一小部分，尚有很多重要变异有待鉴定。外显子组测序是高效鉴定编码区变异的方法之一，其技术流程主要涉及病例和对照样本准备、靶向富集外显子组、高通量测序、数据分析并获得变异基因列表、与 dbSNP 数据库比较排除常见变异等 [13-14](图 1)。2008 年以来，PubMed 数据库收录的外显子组测序研究工作呈逐年递增趋势 (图 2)。

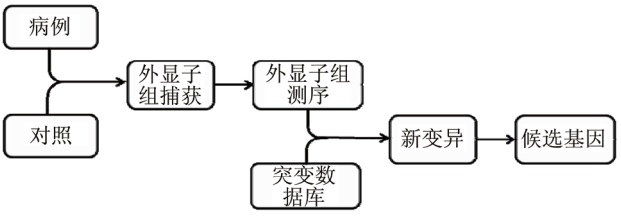
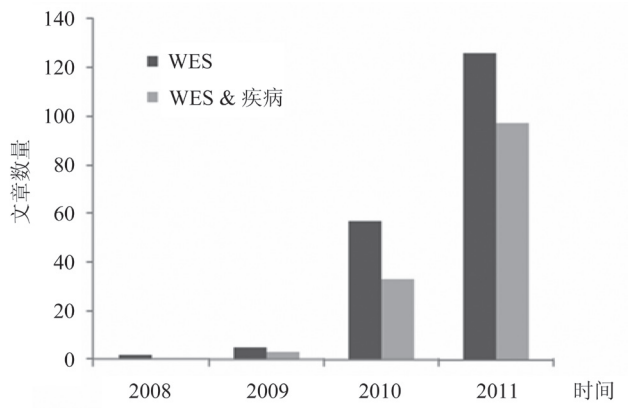


图1 基于外显子组测序的易感基因筛选策略

2.1 人类基因组中低频变异

通常把人类基因组中最小等位基因频率 (minor allele frequency, MAF) 低于 5% 的 SNP 称为低频变异。位于蛋白质编码区的 SNP 称为编码单核苷酸多态 (coding SNP, cSNP)，包含仅涉及基因编码序列改变而不引起氨基酸序列改变的同义编码单核苷酸多态 (synonymous cSNP) 和同时涉及基因编码序列与氨基酸序列改变的非同义编码单核苷酸多态 (non-synonymous SNP, nsSNP)。GWAS 基于“常见疾病，常见变异”的原则，主要研究 MAF>5% 的



注：数据更新截至2011年9月10日。

图2 PubMed数据库收录外显子组测序文章发表趋势

SNPs 与疾病的关联性，不涵盖人类基因组中重要的低频变异，会引起部分遗传度缺失。外显子组测序可以发现基因编码区的低频变异及对疾病具有不同效应的稀有变异，在人类疾病相关基因研究中具有重要作用。

对单个个体的外显子组序列分析发现，10 400 个 nsSNPs 位点中 15%~20% 属于低频 SNP (MAF < 5%)，在所鉴定的 700 个插入 / 缺失突变中，50% 影响到编码蛋白质的氨基酸序列 [15]。除 SNP 外，基因组中还存在频率较低的涉及两个或两个以上核苷酸的变异，即多核苷酸多态性。通过联合分析外显子组和全基因组测序数据，发现数个与恶性胶质瘤显著相关的多核苷酸多态性 [16]。针对 200 个丹麦人个体的外显子组分析发现，低频变异对疾病的发生发展具有重要作用。在所鉴定的 121 870 个 SNPs 中，包括 53 081 个 cSNPs。通过对 SNP 响应值和基于人群的等位基因频率分析发现，MAF 在 2%~5% 范围内的 cSNP 有害突变和非同义突变是同义突变的 1.8 倍，多数低频变异多态性会引起编码蛋白改变，有可能引发疾病 [17]。外显子组测序对等位基因频率相关研究具有重要意义，尤其是那些芯片平台中不涵盖的稀有和低频变异。

2.2　孟德尔单基因遗传病致病基因

多数孟德尔单基因遗传病是由基因组中蛋白编码区的碱基突变导致蛋白功能和表型改变所致。对这些疾病致病基因的研究会受到样本数、外显率和位点异质性等因素影响 [3]，通过连锁分析通常很难找到致病基因。理论上，外显子组测序可发现同一基因座上外显子区域的所有突变，因而能快速直接地鉴定致病基因。

对 4 位相互独立的 Freeman-Sheldon 综合征个体的外显子组测序数据及 8 位 HapMap 正常个体的数据进行分析，同时结合变异数据库进行过滤，发现在所有病例样本中都可鉴定到已知致病基因的变异，该项研究是高通量测序研究单基因病的早期尝试，证明了外显子组测序技术能有效发现单基因病致病基因 [2]。另外一项在 3 个独立家系中寻找 Miller 综合征遗传变异的研究发现，*DHODH* 基因变异在 3 个家系中均存在，可能为新致病基因，提示对少量样本进行外显子组测序分析是鉴定单基因病致病基因的有力手段 [3]。对 10 名独立的 Kabuki 综合征先证者进行外显子组测序，通过已有的 SNP 数据库进行最严格条件过滤后，并没有在患病个体中发现包含未知突变的候选基因。次严格条件过滤会出现中度遗传异质性，但同时也鉴定到了多个候选基因。通过基因型和表型的分层分析发现，编码组蛋白甲基转移酶的 *MLL2* 基因在 7 位先证者中存在突变。通过 Sanger 测序法进一步在另外 43 个病例的 26 个个体中检测到该基因突变，通过变异个体父母 DNA 分析，确认该基因突变是新发现的遗传变异，遗传传递与表型一致。因此，对于遗传异质性强的疾病表型，采用散发病例进行研究时，需要适当处理可能由异质性带来的假阳性问题 [18]。皮质发育畸形致病基因定位受多因素影响，如位点异质性明显、亲缘关系弱和诊断分类边界不清楚等。针对此遗传特点，首先对两名患病个体进行全基因组基因分型鉴定出两个体的同源区段，再通过基于捕获芯片的外显子组测序，发现 *WDR62* 隐性突变是一系列重型大脑皮层畸形的致病原因，包括小脑畸形、巨脑回伴皮质肥厚和胼胝体发育不全等。这一结果提示在定位策略受到位点异质性、诊断边界不清等问题阻碍时，外显子组测序技术对于基因鉴定意义显著 [19]。

中国科学家近年在该领域也取得了一定的成果。研究者通过外显子组测序和连锁分析相结合的策略对一个 4 代家系中的 4 名常染色体显性小脑共济失调患者进行了分析。鉴定到 *TGM6* 基因第 10 外显子的 1 个错义突变 (L517W)，连锁分析所定位区域与该结果吻合，而且在另外一个家系中鉴定到 *TGM6* 基因第 7 外显子的另一个错义突变 (D327G)，但是在 500 例健康对照个体中均未检测到这两个错义突变，进一步验证了该基因的致病性。该研究提示与连锁分析相结合可以有效减少研究所需样本量 [20]。逆向性痤疮 (化脓性汗腺炎) 是一种常染色体显性遗传病，通过 1 个典型家系的全基因组连

锁分析发现，其致病区域位于 1p21.1~1q25.3 区段 [21]。对同一家系中的 2 个病例和 1 个正常个体的外显子组测序，发现 2 个病例共有 85 个基因突变，而 1p21.1~1q25.3 区域包含了其中 8 个突变基因 [22-23]。可见，外显子组测序技术往往能够更加具体地重复连锁分析的结果，同时鉴定到之前未知的相关区域变异。

研究提示，利用外显子组测序技术在小样本中可以鉴定已知和新的致病基因，尤其对于罕见孟德尔疾病的研究，可以对一个家系中的患病个体做外显子测序，是一种相对高效低成本的策略。并且通过与连锁分析等策略相结合的方式在提高结果可信度的同时也减少了研究所需样本量。

2.3 癌症等复杂疾病相关基因变异

通常，来源于父母的肿瘤遗传易感基因具有以突变形式遗传、经常突变产生终止密码、在肿瘤发生过程中另一个未突变的等位基因在体细胞中发生突变失活等特征。对 1 名有家族史的胰腺癌患者外显子组测序分析，发现 15 461 个种系基因组 DNA 突变，含 7 318 个同义突变、7 721 个错义突变、64 个无义突变和 250 个插入 / 缺失突变 [24]。其中，*SERPINB12*、*RAGE* 和 *PALB2* 基因为源于父母的胰腺癌候选遗传易感基因。有报道称，*PALB2* 与乳腺癌易感性和贫血相关，进一步对 96 个有家族史的高加索人胰腺癌病例和 1 084 个正常对照的 *PALB2* 基因测序发现，该基因在 3 例胰腺癌中发生截短突变，而对照均未检测到突变。遗传性胰腺癌中 *BRCA2* 是突变频率最高的基因，而 *PALB2* 的突变频率屈居第二。结合外显子组和全基因组测序对 38 例多发性骨髓瘤组织样本及其外周血样本的遗传变异分析，发现了几个具有原癌基因致病作用的体细胞突变，涉及蛋白质翻译、组蛋白甲基化和凝血功能等。除发现了大量不同形式的变异之外，该研究的一个重要发现是鉴定到与 RNA 处理、蛋白翻译及蛋白去折叠响应相关基因的频繁突变，这类突变在近一半的患者中都能鉴定到。结果提示其中突变率最高的 *DIS3* 突变可能导致蛋白质翻译失调，是多发性骨髓瘤的一种致癌机制。通常认为肿瘤是一种分子网络疾病，因此，研究者也进行了基因集合突变情况分析，试图检测到有变异富集的通路，其中涉及突变最多的基因集合编码的蛋白参与凝血级联系统中纤维蛋白凝块的形成过程。这一研究中不同的变异筛选策略对于肿瘤等复杂疾病提供了良好的研究思路 [25]。

近几年中国研究者对于癌症的外显子组测序研究也有了诸多新发现。对 9 例急性单核细胞白血病个体外周血和骨髓样本的外显子组测序，发现 266 个可能的体细胞变异，含 220 个单核苷酸变异和 46 个插入 / 缺失突变。其中，影响蛋白编码序列的非同义突变 59 个；影响 ORF 表达的插入 / 缺失突变 10 个，其中 66 个得到进一步验证。112 例样本中的 23 个个体存在 *DNMT3A* 基因的错义突变，均位于进化保守的氨基酸残基。该突变改变 DNA 甲基转移酶活性，从而导致 DNA 甲基化水平和基因表达异常。*DNMT3A* 突变型个体伴有 *HOXB* 基因簇 CpG 岛低甲基化 [26]。对 14 例黑色素瘤及其配对正常组织的外显子组测序分析，发现存在 1 个 *TRRAP* 基因频发突变、*GRIN2A* 基因突变率较高，此外，大部分突变基因聚集在谷氨酸代谢通路 [27]。移行细胞癌是膀胱癌的主要类型，对 9 例患病个体的组织样本及配对的外周血样本进行外显子组测序，通过比较 dbSNP 数据库及千人基因组先期亚洲人群数据，筛选到 465 个可能的体细胞突变，包括 329 个错义突变。共鉴定到 49 个新候选易感基因，其中以 *UTX*(突变率 21%) 为代表的 8 个基因涉及到染色质重塑过程，提示这些突变可能与肿瘤的早期发展相关 [28]。无独有偶，对 10 名 HCV 病毒相关的肝细胞肝癌个体的肿瘤组织及对应的正常组织的外显子组测序分析发现的新易感基因 *ARID2*，也是染色质重塑基因之一。该研究同样也鉴定到上百个潜在的体细胞突变，通过不同种族不同类型的肝癌样本验证后发现，*ARID2* 在 18.2% 的研究个体中存在失活突变，可能为新的抑癌基因 [29]。

上述研究发现的肿瘤相关基因突变大部分为非同义突变，也发现了许多插入 / 缺失突变，有的基因同时存在两种以上突变形式，提示肿瘤遗传变异形式复杂，频率较高。同时，染色质重塑基因在上述两种肿瘤中的高突变率也再一次有力地提示遗传改变与表观遗传的失调可能共同协作促进肿瘤的生成和进展，显示了表观基因组学 (epigenomics) 在肿瘤研究中的必要性。另外，可认识到在不考虑非编码区变异的情况下，外显子组测序还不能覆盖所有编码区已知致病变异，可能因为对某些变异区测序深度不够，或涉及重复区域以及存在数据拼接等问题。在小样本量研究中，为避免由于测序错误而引起假阳性结果，需保证足够的测序深度，30 重以上测序深度较为适宜。研究发现 5 到 20 重深度测序，可以大幅提升发现遗传变异的敏感度；而 20 至 50

重测序，敏感度缓慢升高；大于 50 重测序后进入平台期[4]。测序深度不仅会影响总测序量、发现新变异几率和成本，更影响测序数据的准确度。

总之，外显子组测序是介于全基因组关联分析与全基因组测序之间的基因分析策略[30]。该技术能较系统地发现基因组中蛋白编码区的主要遗传变异，与全基因组测序相比，工作量与分析成本相对较低。目前外显子组测序还不能覆盖所有编码区的致病变异，需要通过选择不同富集与测序平台、提高测序深度，以及改进数据分析与拼接方法等措施进一步完善。

[参 考 文 献]

[1] Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods, 2008, 5(1): 16-8

[2] Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature, 2009, 461(7261): 272-6

[3] Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet, 2010, 42(1): 30-5

[4] Choi M, Scholl UI, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA, 2009, 106(45): 19096-101

[5] Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. Hum Mol Genet, 2010, 19(R2): R145-51

[6] Kahvejian A, Quackenbush J, Thompson JF. What would you do if you could sequence everything? Nat Biotechnol, 2008, 26(10): 1125-33

[7] Bainbridge MN, Wang M, Burgess DL, et al. Whole exome capture in solution with 3 Gbp of data. Genome Biol, 2010, 11(6): R62

[8] Mamanova L, Coffey AJ, Scott CE, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods, 2010, 7(2): 111-8

[9] Lander ES. Initial impact of the sequencing of the human genome. Nature, 2011, 470(7333): 187-97

[10] Lalonde E, Albrecht S, Ha KC, et al. Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. Hum Mut, 2010, 31(8): 918-23

[11] Zaghloul NA , Katsanis N. Functional modules, mutational load and human genetic disease. Trends Genet, 2010, 26(4): 168-76

[12] Cohen JC, Kiss RS, Pertsemlidis A, et al. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science, 2004, 305(5685): 869-72

[13] Summerer D, Schracke N, Wu H, et al. Targeted high throughput sequencing of a cancer-related exome subset by specific sequence capture with a fully automated microarray platform. Genomics, 2010, 95(4): 241-6

[14] McClellan JM, Susser E, King MC. Schizophrenia: a common disease caused by multiple rare alleles. Br J Psychiatry, 2007, 190: 194-9

[15] Ng PC, Levy S, Huang J, et al. Genetic variation in an individual human exome. PLoS Genet, 2008, 4(8): e1000 160

[16] Rosenfeld JA, Malhotra AK, Lencz T. Novel multi-nucleotide polymorphisms in the human genome characterized by whole genome and exome sequencing. Nucleic Acids Res, 2010, 38(18): 6102-11

[17] Li Y, Vinckenbosch N, Tian G, et al. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat Genet, 2010, 42: 969-72

[18] Ng SB, Bigham AW, Buckingham KJ, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet, 2010, 42(9): 790-3

[19] Bilguvar K, Ozturk AK, Louvi A, et al. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature, 2010, 467(7312): 207-10

[20] Wang JL, Yang X, Xia K, et al. TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Brain, 2010, 133(Pt 12): 3510-8

[21] Gao M, Wang PG, Cui Y, et al. Inversa acne (hidradenitis suppurativa): a case report and identification of the locus at chromosome 1p21.1-1q25.3. J Invest Dermatol, 2006, 126(6): 1302-6

[22] Liu Y, Gao M, Lv YM, et al. Confirmation by exome sequencing of the pathogenic role of NCSTN mutations in acne inversa (hidradenitis suppurativa). J Invest Dermatol, 2011, 131: 1570-2

[23] Wang B, Yang W, Wen W, et al. γ-secretase gene mutations in familial acne inversa. Science, 2010, 330(6007): 1065

[24] Jones S, Hruban RH, Kamiyama M, et al. Exomic sequencing identifies PALB2 as a pancreatic cancer susceptibility gene. Science, 2009, 324(5924): 217

[25] Chapman MA, Lawrence MS, Keats JJ, et al. Initial genome sequencing and analysis of multiple myeloma. Nature, 2011, 471(7339): 467-72

[26] Yan XJ, Xu J, Gu ZH, et al. Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. Nat Genet, 2011, 43: 309-15

[27] Wei X, Walia V, Lin JC, et al. Exome sequencing identifies GRIN2A as frequently mutated in melanoma. Nat Genet, 2011, 43(5): 442-6

[28] Gui Y, Guo G, Huang Y, et al. Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. Nat Genet, 2011, 43(9): 875-8

[29] Li M, Zhao H, Zhang X, et al. Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. Nat Genet, 2011, 43(9): 828-9

[30] Mardis ER. A decade's perspective on DNA sequencing technology. Nature, 2011, 470(7333): 198-203