中国科学: 生命科学 2011 年 第 41 卷 第 9 期: 714 ~ 721
SCIENTIA SINICA Vitae www.scichina.com life.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

论 文

一种结合单张芯片序列捕获和高通量测序技术测序外显子组的方法

蒋涛[①†], 杨蕾[②③†], 蒋慧[①], 田埂[①②③], 张秀清[①②③*]

① 深圳华大基因研究院, 深圳 518083;
② 中国科学院北京基因组研究所, 北京 101300;
③ 中国科学院研究生院, 北京 100049
† 同等贡献
* 联系人, E-mail: zhangxq@genomics.org.cn

收稿日期: 2011-07-13; 接受日期: 2011-08-16
中国科学院对外合作重点项目(批准号: GJHZ07016)、中国科学院知识创新工程(批准号: KSCX-YW-N-023)、国家自然科学基金(批准号: 30725008, 90403130, 90608010, 30221004, 90612019 和 30392130)、国家重点基础研究发展计划(批准号: 2007CB815701, 2007CB815703 和 2007CB815705)、整合生物学丹麦平台和个性化疾病预测预防治疗应用医学基因组学 Lundbeck 基金会(批准号: LUCAMP)资助项目

摘要 随着高通量测序技术的发展, 全外显子测序已经成为一种研究人类疾病的重要方法. 本文展示了一种通过Nimblegen 2.1M芯片进行外显子DNA序列捕获和高通量测序的方法, 包括两步法文库制备. 测序的平均覆盖深度达33倍时, 95.6%的34M目标区域得到均衡覆盖, 特异性达到80%. 对比全基因组鸟枪法测序的结果, 此方法在检测SNP时的假阳性率为0.97%, 假阴性率为6.27%. 本方法对于全基因组扩增的DNA也适用. 结果显示, 全外显子测序技术可以在大规模的群体研究和医学研究中起到重要作用.

关键词
外显子组测序
外显子组捕获
高通量测序

近年来出现的大量DNA平行测序技术可产生大量短测序片段, 使得在较低的成本下测序的通量获得大幅提升[1]. 过去的两年间, 在这些新一代测序平台上已经完成多个个人基因组, 找到数以百万的遗传变异[2~4]. 然而对于许多研究而言, 尤其是那些需要对上亿人批量测序来确认DNA序列与各种疾病相关的遗传变异时, 一个完整的人类基因组重测序依然价格昂贵. 需要说明的是与疾病相关的变异大都发生于蛋白编码区, 这仅占人类基因组的 2%. 因此, 以成本效益至上的目标区域重测序方法有所发展, 依托这些方法研究高价值区域可大大降低成本, 使针对成千上万样本的测序研究得以进行.

近年来发展的外显子捕获技术通过杂交选择的方法有效地富集了目标区域. 但该技术的应用仍有局限, 如使用分子倒置探针(molecular inversion probes)会丢失约 80%的目标外显子[5]. 生物素标记的RNA捕获探针可很好地应用于1900个人类基因, 但是其对整个人类外显子的作用还未被证实[6]. 虽然已尝试过几种基于芯片的外显子捕获技术, 但是每张芯片上都只有一小部分人类外显子[1,7,8].

本文使用 NimbleGen 高密度 2.1M 芯片来捕获整个人类外显子组, 并用 Illumina 基因组分析仪进行了精确的测序. 使用一个已完成的鸟枪法重测序的亚洲人基因组来评价本实验方法[3]. 外显子组的

英文版见: Jiang T, Yang L, Jiang H, et al. High-performance single-chip exon capture allows accurate whole exome sequencing using the Illumina Genome Analyzer. Sci China Life Sci, 2011, 54, in press

完整性、平均覆盖度以及高精度变异位点的发现充分证明了此策略的有效性, 具备大规模多样本研究的潜力.

1　材料与方法

1.1　文库构建和测序

为使全外显子组捕获适应 Illumina 测序平台的要求, 使用两步法构建文库(图 1).

第一步先将 20 μg DNA 雾化后打断成 500 bp 左右的片段. 按照 NimbleGen 的实验流程将 5 μg 单链片段杂交到 2.1M 外显子组捕获芯片, 作为 PCR 扩增时起连接作用的接头. 洗掉未杂交上的片段后将杂交上的片段从芯片上洗脱. 这些片段再用连接介导的 PCR 方法(LM-PCR)进行后续扩增. 由此可得到一组 500 bp 左右片段的初级文库, 它们可直接应用到 Roche/454 平台的直接末端测序; 第二步, 将这些片段用 DNA 连接酶接在一起, 再剪切为 200 bp 左右的片段, 之后可与 Illumina 测序接头相连接. 这样就将原有的初级文库转变为一个适合 Illumina GA-Ⅱ平台插入片段大小的二级鸟枪文库, 之后会进入标准的 Illumina DNA 测序流程. 最终, 外显子富集后的鸟枪文库会使用标准测序引物按照制造商提供的实验流程在 Illumina GA-Ⅱ平台上测序. 在默认参数条件下, 使用 Genome Analyser Pipeline(1.3 版)进行图像分析和碱基判定.

1.2　使用的公共数据

使用 NimbleGen 2.1M 人类外显子组芯片捕获的目标 DNA 片段(http: //www.nimblegen.com/downloads/annotation/seqcap_exome/index.html).

CCDS 外显子信息下载于 CCDS 数据库(20080902 版, ftp: //ftp.ncbi.nlm.nih.gov/pub/CCDS/).

YH 一致序列基因型来自 http://yh.genomics.org.cn/.

1.3　read 比对

使用 SOAP aligner(2.01 版)在最多允许 2 个不匹

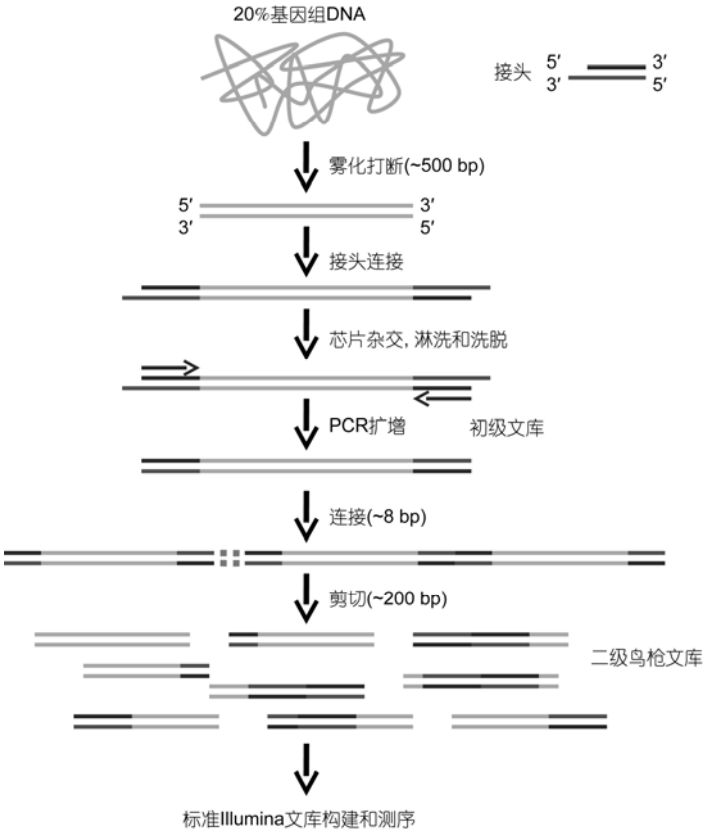

图 1　使用 NimbleGen 2.1M 探针序列捕获芯片进行全外显子组捕获和 Illumina GA 测序两步法文库构建的实验流程

配的原则下对原始测序 read 与参考序列比对. 参数设置为-a -D -o -r 1 -t -c -f 4. 本文所用的参考人类基因组是 NCBI build 36.1. 染色体序列下载至 UCSC 数据库 (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/). 只选取比对上的 read 进行后续分析和 SNP 判读.

1.4 SNP 判读和精度评估

基于 SOAP 比对结果, 使用 SOAPsnp 软件组合一致性序列、判读目标基因型. 参数设置如下: -i -d -o -r 0.00005 -e 0.0001 -M -t -u -L -s -2 –T. 首个亚洲人种的基因组序列和 SNP 被用来评估外显子组捕获序列基因型判读的精确度. YH 外显子组未比对的原始测序数据请见 http: //yh.genomics.org.cn.

1.5 插入/缺失检测

为了判读插入/缺失, 对 YH 外显子组首先使用 GATK 流程, 再将 read 使用 BWA 与 hg18 参考基因组重比对[9]、应用 GATK 碱基质量分数重校正、插入/缺失重比对、重复删除、插入/缺失发现和基因分型, 同时使用标准的硬过滤参数或变异质量得分重新校准. 然后使用 Dindel 流程, 首先从 BWA 比对结果中选出候选插入/缺失, 再定义重比对窗口. 对每个重比对窗口, 生成代表参考序列替代序列的候选单倍型, 然后将测序序列与这些候选单倍型重比对, 最终用贝叶斯方法估计候选插入/缺失的统计学可能性. 两种插入/缺失判读方法同时认定, 外加至少 3 个 read 支持, 最终得到的插入/缺失可信度高.

2 结果

2.1 用于 Illumina 测序的外显子组捕获

使用 NimbleGen 2.1M 探针芯片富集炎黄一号的外显子组, 此款芯片具有定制合成的 60~90 寡聚核苷酸多聚体. 设计的 2.1M 杂交探针可覆盖 34 Mb 目的区域, 包括已被注释的 18654 个 *CCDS* 基因(http://www.ncbi.nlm.nih.gov/CCDS)中的 180640 编码区(表 1)、覆盖区域上下游 200 个碱基区域的序列和 698 个 microRNA. 虽然一些具有重复序列较难杂交的基因无法列入其中, 但是这款高密度芯片已涵盖了 92.8% 的注释基因, 接近一个完整的外显子组.

炎黄一号基因组的研究完成后[3], 对这个样本进行了外显子捕获(图 1). 简要地说, 首先将一份 20 μg 的基因组 DNA 样本超声随机打断, 得到的片段大小平均为 500~600 bp; 再将这些片段杂交到捕获芯片上, 严格的洗脱条件去除非特异吸附; 随后洗脱杂交上的 DNA, 按照标准文库构建步骤(详见方法部分)构建成测序文库, 并在 Illumina 测序平台上测序. 由此得到 67000000 reads(平均大小为 75 bp), 相当于 5.1G 测序碱基. 因为杂交时应用的接头序列和测序文库构建的接头序列都被引入最终的测序序列中, 利用 Smith-Waterman 算法找出可能的接头序列(图 2).

表 1 NimbleGen 2.1M 探针外显子组捕获芯片的目标区域构成情况 [a)]

项目	元件	长度(bp)	百分比(%)
启动子区	3062	50063	0.15
5′-UTR	14604	436388	1.28
编码外显子	180640	23939542	70.19
内含子	149008	8771941	25.72
3′-UTR	14505	477440	1.40
microRNA	545	46776	0.14
CCDS	18654	33720506	98.86
基因间隔区	–	388304	1.14

a) 目标区域全长为 34108810 bp, 因为一些元件的重叠造成此数值比各项总和值略小

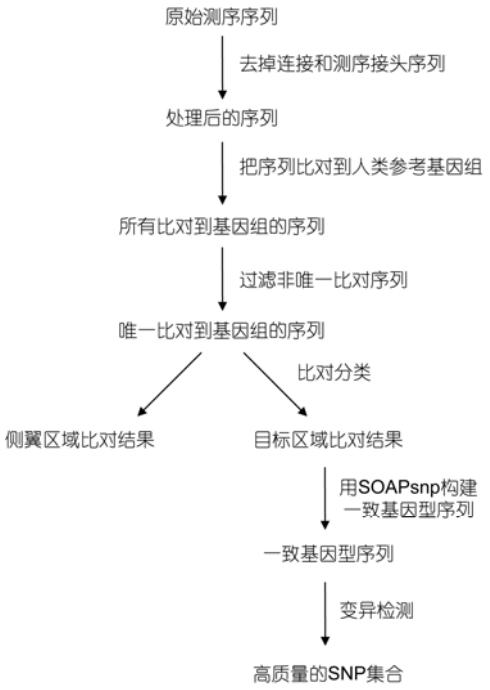

图 2 生物信息分析流程

经比对与接头连续有 12 bp 相同(最多有 2 个不匹配)的 reads 被认定为接头序列. 这些 reads 被截短后进行进一步的分析. 这个过滤步骤极大地提高了目标区域的数据质量和有效测序深度(表 2). 应用 SOAP 软件[10]与完整基因组比对, 发现与接头连接的 4.82 Gb 序列中的 3.38 Gb 都是唯一序列. 在这些序列(表 3)中, 1.14 Gb 的序列(34%)完全落在目标外显子区域, 1.57 Gb 的序列(46%)落在目标区域两端 500 bp 的侧翼区域. 目标测序区域测序深度为 33 倍, 而外展侧翼区域测序为 13×(表 3). 对那些比对到目标区域的 reads 随后用 SOAPsnp[11](参见方法部分)进行 SNP 鉴定.

2.2　完整外显子组捕获方法的评价

通过特异性、均一性和可重复性 3 个参数对这个完整外显子组捕获方法进行评价.

特异性是指序列中通过特异选择选出的来自于目标区域的部分, 而非来源于系统误差或随机效应. 在这项评价中, 只考虑特定测序序列所占的比例作为捕获特异性的指标, 即在所有唯一比对的 reads 中, 特异比对到目标外显子匹配或外显子附近区域的 reads 所占百分比. 如上所述(表 3), 3.38 Gb 唯一比对的 reads 中约有 2.7 Gb(80%)定位于目标外显子或其侧翼区域. 基于这些结果, 预计此方法的特异性可达 80%. 通过归一化后的碱基深度图发现, 超出目标区域边界的碱基深度有明显下降, 这进一步印证了此方法的高特异性(图 3). 结果指出对于整个外显子组进行了外显子捕获后的特异性接近于小规模的人类外显子捕获实验结果[6].

均一性是指对于整个外显子捕获和测序的随机性. 这点对于不偏移的发现基因组变异很关键. 本实验方法在目标区域覆盖度达到 95.6%, 而其侧翼区域覆盖度也可达 86.2%(表 3). 通过随机抽取比对到目标区域上的测序数据发现, 在测序深度只有 6 倍的情况下, 目标区域被唯一比对 reads 覆盖的比例已超过 90%, 而随着测序深度的增加改善不大(图 4). 这些数据证实了本实验捕获测序结果的随机性, 同时结果也几乎达到了先前测试阶段部分外显子捕获覆盖度的上限.

表 2　接头屏蔽与否的序列比对结果比较 [a)]

	未屏蔽连接接头/测序接头[b)]	已屏蔽连接接头/测序接头
原始数据量(Mb)	5096	4818
比对上的 reads(Mb)	3178	3707
唯一比对(Mb)	2897	3379
目标区域唯一比对(Mb)	1001	1136
目标区域平均深度	29.3×	33.3×

a) 屏蔽掉接头后虽然输入的原始数据较少, 但与基因组比对会增加 14%的数据; b) 连接接头: linker; 测序接头: adaptor

表 3　GA-Ⅱ测序比对汇总 [a)]

	目标区域	侧翼区域	基因组背景	总量
基因组长度(Mb)	34	117	2707	2858
唯一比对(Mb)	1136	1566	677	3379
平均深度	33.3	13.4	0.25	–
覆盖度(%)	95.6	86.2	9.5	–

a) 目标区域平均测序深度为 33 倍, 比侧翼区域高 2.5 倍. 这符合实验原理, 即与探针杂交碱基多的片段比少的片段更易被捕获. 对非特异杂交的 0.25 倍基因组覆盖度非特异的背景 DNA 也进行了测序

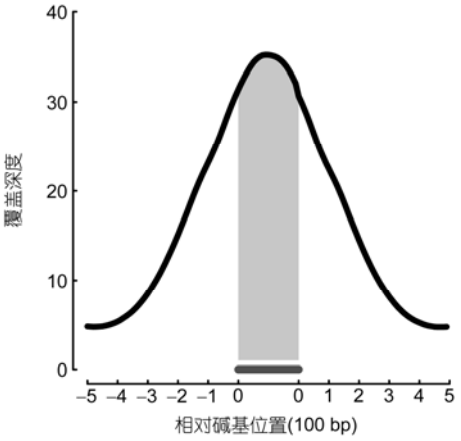

图 3　相对于目标区域的比对数据分布

目标区域(灰色区域)具有最高测序深度(约 33 倍). 比对数据显示 500 bp 侧翼序列同样被富集, 但在延展至非目标区域时测序深度降低

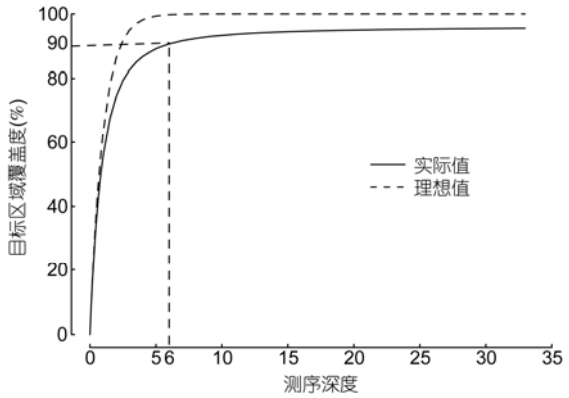

图 4　外显子捕获测序深度与目标区域覆盖度的相关性

在 4.5%未被覆盖到的目标区域中: 2%的区域可通过比对到多个位置的 reads(即重复比对序列)弥补; 而剩余的 2.5%则很可能是对杂交技术而言非常难捕获的区域. 后续分析表明, 这些区域通常有过高的 G+C 含量, G+C 含量大都超过 60%(图 5). 接下来检查了覆盖区的单碱基深度分布, 以此衡量覆盖区的均匀度(图 6). 大于 5 倍标准差时, 分布情况几乎符合预期中的理想模型(泊松分布). 分析表明, 覆盖度的差异是由于高 G+C 含量所带来的偏差造成的(图 7). 中等 G+C 含量(40%~60%)的目标区域会得到更高的覆盖度, 而过高或过低 G+C 含量的区域覆盖度较差. 与此类似的是在测序捕获和测序中所运用的杂交和 PCR 技术, 通常会优化实验条件到适合中等 G+C 含量片段, 高 G+C 含量片段对其而言同样是个难题. 目前的实验条件可覆盖外显子组的大部分, 认为这是一个合理的取舍. 目标区 76.8%的碱基测序深度达到 16 倍(平均深度的一半)以上, 此结果相比以前对于部分外显子捕获的研究数据更好[6].

可重复性是指实验流程能否重复并应用到其他样本, 这点对多种样本的大规模应用非常关键. 为测试可重复性, 做了另外一个样本(称为 EMC)完整外显子组的实验, 几乎得到同等数量的数据. 两个样本目标区域测序深度分别就各自整体的平均深度进行归一化. 归一化后, YH 和 EMC 两个样本测序深度的相关系数为 0.7, 而 log 值趋近于 0. 这就说明完整外显子组捕获方法可在另一个样本中被重复(图 8).

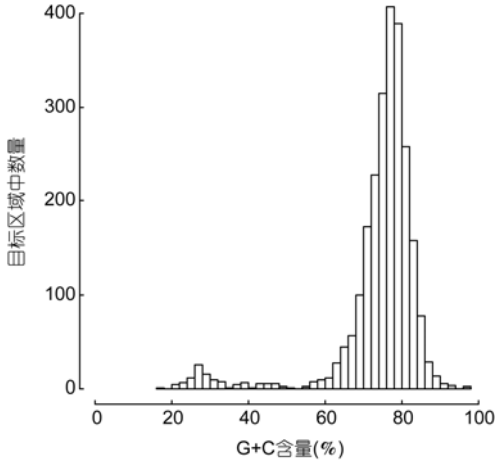

图 5　未被捕获的目标区域中 G+C 含量分布

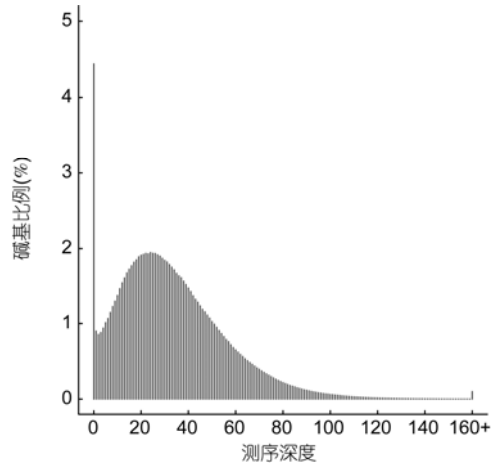

图 6　目标区域中单碱基深度分布

尽管有部分目标区域未被捕获, 但是仍大致符合理论泊松模型的单峰分布

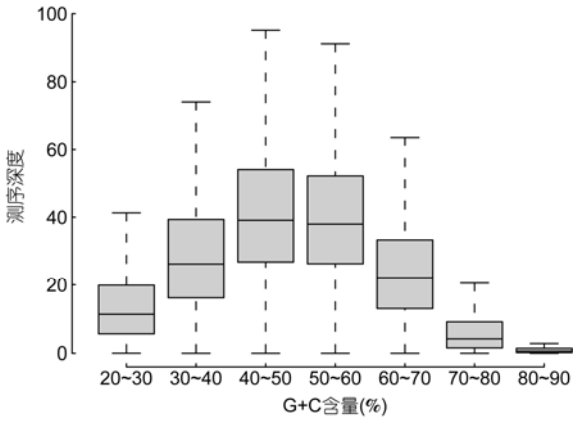

图 7　在不同 G+C 含量水平下目标区域平均测序深度

矩形框表明第一四分值和第三四分值之间的范围. G+C 含量在 40%~60%的目标区域具有最高的测序深度; 低 G+C 含量和高 G+C 含量目标区域测序数据较少

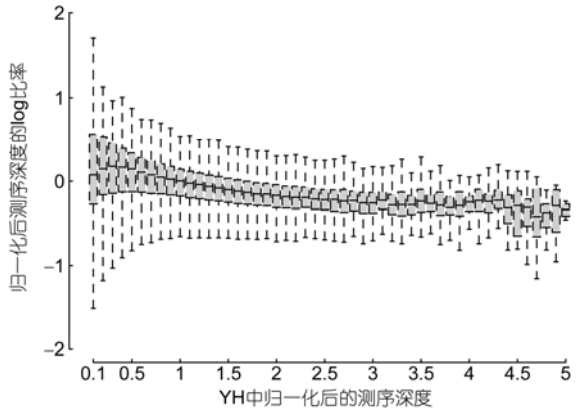

图 8　YH 和另一个独立样本外显子组捕获后归一化的测序深度分布的可重复性

2.3 运用外显子组测序进行变异检测

开创性地对整个人类基因组进行重测序, 在此实验中证实了使用 Illumina Genome Analyzer 进行 SNP 鉴定精度高、一致性好. 它与基于芯片基因分型得出的等位基因一致率超过 99.9%[3]. 这就提供了一个前所未有的机会, 在整个基因组水平上运用外显子捕获技术来评估 SNP 鉴定的准确性. 目前, 通过 SOAPsnp[11]从外显子捕获测序数据建立了 YH 外显子组的一致序列, 参数与 YH 基因组重测序项目相同. 总数为 32387638(95%)的外显子组一致性的基因型超过 Q20 基线, 其中的 32350665 同样与超过 Q20 基线的 YH 基因组一致性的基因型相重叠. 在 YH 基因组与 YH 外显子组一致性的序列中只有 1568(4.8×10^{-5})个碱基对不符(表 4). 两组数据的高一致性表明, YH 外显子组分析提供了精确的基因型读取结果, 本方法只引入很小的误差.

将 YH 外显子组的一致性序列与 NCBI36 人类参考序列相比较, 共发现 19972 个潜在的高质量 SNP. 以 YH 基因组一致性序列作为金标准, YH 外显子组 SNP 中有 195 个是过读取(Overcalls), YH 基因组 SNP 中有 1323 个是未读出(Undercalls), 表明在 Q20 下, 过读取 SNP 误差率是 0.97%的假阳性和 6.27%的假阴性. 这说明了此方法只引入了很小的假阳性. 但是外显子捕获技术所引入的假阴性有些高, 这点应进一步优化. 问题很可能是由于杂交探针是根据参考序列设计合成, 倾向于捕获与参考序列相同的等位基因. 与参考序列相比, 一些含有 SNP 数目较多的区域使用捕获探针无法杂交, 因而丢失了这部分序列.

为了确定测序深度对 SNP 覆盖度和精度的影响, 随机抽取不同测序深度的部分序列, 用同样的方法检测 SNP. 测序深度的增加理论上会增强 SNP 的检出率. 事实也是如此, 随着测序深度的增加, SNP 判读精度获得了稳步提升. 在大概 11 倍测序深度时, 可找到 65%的高质量 YH SNP, 它们的精确度可达 91%. SNP 检测的全部敏感度比 Choi 等人[12]检测的要高, 即 65%比 50%. 而低精度很大原因是因为高假阴性率. 同样, YH 外显子组和 YH 基因组具有相同的 SNP 位点, 一致率高达 99.65%, 所以说 SNP 检测敏感性和精确度只受限于测序深度.

同时应用 GATK[13]和 Dindel[14]软件包在 YH 外显子组数据集中寻找插入/缺失. 最终, 利用 GATK 方法找到 1128 个插入/缺失, 其中 98 个位于编码序列; 而 Dindel 发现 1130 个, 167 个位于编码区(表 5). 插入/缺失数目的不同可能因为两种软件算法的不同.

2.4 以整个基因组扩增样本评估外显子组捕获方法

对整个外显子组捕获和测序所需的 DNA 量为 20 μg, 这可能会限制本方法在那些只具有少量 DNA 医疗样本中的应用. 近来, 对整个基因组扩增 DNA 样本的序列捕获测试已取得成功[7]. 同样将全基因组扩增(WGA)技术应用到 YH DNA, 并用此 WGA DNA 测试本实验流程. 此 WGA DNA 目标区域的测序深度接近 12 倍, 覆盖度为 86.4%. 这比同样测序深度下、未经过 WGA 技术处理的 93.8%覆盖度要低. 单碱基深度分布也表明, 其具有较低的均一性, 此分布不符合理想模型曲线(图 9). 在 Q20 标准下, WGA 后一致性序列 SNP 误差率具有 1.73%的假阳性率和 9.9%的假阴性率, 表明 WGA 外显子组取样会有更大的偏差. 然而, 结果与先前结果类似, 仍可为只有少量基因组 DNA 的医学研究提供有意义的参考数据.

3 讨论

为了降低采用新一代测序技术对个人基因组和大样本全基因组关联(GWA)研究的成本, 最好的方法就是对通过探针杂交捕获后的整个外显子组进行重测序. 在这项研究中, 全外显子组来自于一个基因组已经被很好地重测序的个体, 对此全外显子组区域进行了深度测序和广泛分析. 由此系统地评估了

表 4 YH 外显子组与基因组对比 a)

	等位基因型	YH 外显子组			总量	差异率
		HOM ref.	HOM mut.	HET		
YH 基因组	HOM ref.	32329370	6	189	32329565	6.03×10^{-6}
	HOM mut.	29	8413	11	8453	4.70×10^{-3}
	HET	1294	39	11314	12647	1.05×10^{-1}

a) HOM ref.代表基因型与 NCBI36 人类参考基因组相同; HOM mut.代表 NCBI36 的纯合子; HET 代表杂合子

表 5　YH 外显子组鉴定到的插入/缺失

方法	GATK	Dindel
插入/缺失总量	1128	1130
插入编码	29	46
缺失编码	69	121
剪切位点	42	43
内含子	916	848
5′ UTRs	35	24
3′ UTRs	35	45
基因间隔区	2	3
所有插入	438	433
所有缺失	690	697
杂合插入/缺失	669	716
纯合插入/缺失	459	414

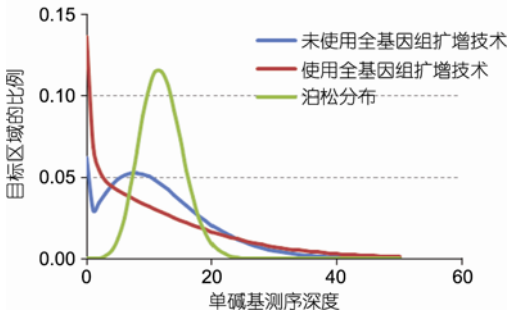

图 9　WGA 样本外显子组和测序的单碱基测序深度分布
平均 12 倍测序深度时, 13.4%目标区域仍未被任何 read 所覆盖. 而在非 WGA 样本中, 该区域只占 6.2%

外显子捕获技术的效率和在实验、测序过程中可能引入的潜在测序误差. 本实验数据表明, 所捕获到的序列是外显子区特异性地富集序列, 只带有很少量的侧翼区域序列. 虽然高 G+C 含量区域序列可能会在捕获过程中有所丢失, 但对大多数目标区域而言, 在高测序深度时对比参考序列是均匀分布的, 因此在单碱基水平达到高精度. 此外, 该方法还可以在另一个样本中重建全基因组数据, 提示此方法广泛应用的可能性.

先前, 大多数 GWAS 研究依赖芯片完成, 这只能基于现有的 SNP 数据库(如 dbSNPs)[15]. 然而, 对一个亚洲人种全基因组的高分辨率重测序已经鉴定到成千上万的新 SNP, 这些都未涵盖在 dbSNPs[3]. 这说明仍有许多在不同人种中未被发现的 SNP. 可见, 将芯片应用到 GWAS 研究一个很重要的局限就在于它可能忽略那些稀有 SNP, 而这些 SNP 却可能起到很重要的生物学作用. 采用目前的方法进行高精度 SNP 全基因组扫描, 结果支持了在 GWAS 研究中对全外显子区使用外显子捕获进行目标性重测序的可能性. 虽然在杂交过程中引入了很小的 GC 偏差要进行矫正, 但是高覆盖度的目标区域和高精度的 SNP 都表明该偏差不会影响后续分析. GWAS 研究中很重要的一点是样本收集的限制. 本实验数据显示, 进行全基因组扩增可以大幅度降低对样本 DNA 量的需求. 但由于全基因组扩增中可能引入的扩增偏向性, 以及扩增酶引入的扩增错误, 全基因组扩增会引入更高的假阳性率和假阴性率.

近来, 在全外显子组重测序中使用外显子捕获已被成功地应用到鉴定罕见的孟德尔疾病的致病基因[16~19]和基因诊断[12]上. 本实验室还在 50 例西藏人群中使用该捕获方法进行外显子组重测序, 以此揭示高海拔适应的原因[20]. 本实验室同时完成了 200 例人类外显子组的重测序, 鉴定到大量低频非同义编码变异, 其中大部分被预测为罕见的有害突变[21]. 下一步, 将采用外显子捕获和新一代测序技术相结合的策略来阐释代谢疾病中的基因风险. 在当前框架中, 已经提供了此策略可应用到大规模全外显子组重测序的直接证据. 特别是通过这种方法检测到一组相关的高精度 SNP. 本研究可能促使外显子组测序在疾病关联研究中的进一步应用.

致谢　　感谢深圳华大基因研究院 Lars Bolund 教授、李奇斌和胡宇洁对本工作提供的帮助.

参考文献

1　Albert T J, Molla M N, Muzny D M, et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods, 2007, 4: 903–905
2　Levy S, Sutton G, Ng P C, et al. The diploid genome sequence of an individual human. PLoS Biol, 2007, 5: e254
3　Wang J, Wang W, Li R, et al. The diploid genome sequence of an Asian individual. Nature, 2008, 456: 60–65

4 Wheeler D A, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature, 2008, 452: 872–876
5 Porreca G J, Zhang K, Li J B, et al. Multiplex amplification of large sets of human exons. Nat Methods, 2007, 4: 931–936
6 Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol, 2009, 27: 182–189
7 Hodges E, Xuan Z, Balija V, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet, 2007, 39: 1522–1527
8 Okou D T, Steinberg K M, Middle C, et al. Microarray-based genomic selection for high-throughput resequencing. Nat Methods, 2007, 4: 907–909
9 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 2009, 25: 1754–1760
10 Li R, Yu C, Li Y, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, 2009, 25: 1966–1967
11 Li R, Li Y, Fang X, et al. SNP detection for massively parallel whole-genome resequencing. Genome Res, 2009, 19: 1124–1132
12 Choi M, Scholl U I, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA, 2009, 106: 19096–19101
13 McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res, 2010, 20: 1297–1303
14 Albers C A, Lunter G, Macarthur D G, et al. Dindel: accurate indel calls from short-read data. Genome Res, 2010, 21: 961–973
15 Frazer K A, Murray S S, Schork N J, et al. Human genetic variation and its contribution to complex traits. Nat Rev Genet, 2009, 10: 241–251
16 Ng S B, Turner E H, Robertson P D, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature, 2009, 461: 272–276
17 Ng S B, Bigham A W, Buckingham K J, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet, 2010, 42: 790–793
18 Ng S B, Buckingham K J, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet, 2010, 42: 30–35
19 Bilguvar K, Ozturk A K, Louvi A, et al. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature, 2010, 467: 207–210
20 Yi X, Liang Y, Huerta-Sanchez E, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. Science, 2010, 329: 75–78
21 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: 969–972