·论　　著·

DNA-pool 全基因组高通量重测序分析原发性高血压单核苷酸多态性变异*

崔文博[1,2]，刘银河[2]，周义文[3△]

（1. 南华大学，湖南衡阳 421000；2. 深圳市孙逸仙心血管医院检验科，广东深圳 518001；
3. 南方医科大学深圳医院检验科，广东深圳 518001）

摘　要：目的　利用 DNA-pool 技术对原发性高血压患者进行测序，探索中国原发性高血压患者单核苷酸多态性变异（SNP）情况。**方法**　连续收集 2014 年 3 月至 2014 年 6 月深圳市孙逸仙心血管医院的高血压门诊患者 100 例。将基因组 DNA 片段化处理至 400～800 bp 进行建库测序，将测序结果与 NCBI（National Center of Biotechnology Information）人类基因库 hg19 进行比对。**结果**　共生成 120.8 Gb 原始序列数据，测序深度约为 36.13 倍，覆盖率达到了 99.88%。通过生物信息学分析共检测到 4 305 668个 SNP 点，其中 C:G→T:A 的变异类型最多，达到 12 314 个变异位点。**结论**　研究验证了使用 DNA-pool 的全基因组重排序的方法。研究所得数据也对中国原发性高血压的基因型数据库进行了补充，对未来原发性高血压基因研究提供了一定的帮助。

关键词：原发性高血压；　全基因组重排序；　单核苷酸多态性变异；　DNA-pool

DOI：10.3969/j.issn.1673-4130.2017.09.007　　**文献标识码：**A　　**文章编号：**1673-4130(2017)09-1172-04

DNA-pool high-throughput whole genome resequencing for exploring essential hypertension single nucleotide polymorphism mutation*

CUI Wenbo[1,2], *LIU Yinhe*[2], *ZHOU Yiwen*[3△]

(1. *University of South China*, *Hengyang*, *Hunan* 421000, *China*; 2. *Department of Clinical Laboratory*, *Shenzhen Municipal Sun Yat-sen Cardiovascular Hospital*, *Shenzhen*, *Guangdong* 518001, *China*;
3. *Department of Clinical Laboratory*, *Shenzhen Hospital*, *Southern Medical University*, *Shenzhen Hospital*, 518000, *China*)

Abstract: Objective　To use the DNA-pool technology to sequence patients with essential hypertension(EH) for exploring the single nucleotide polymorphism(SNP) mutation situation in Chinese patients with EH. **Methods**　One hundred EH outpatients in the Shenzhen Sun Yat-sen Cardiovascular Hospital from March to June 2014 were continuously collected. The genomic DNA was performed the fragmentation process to 400－800 bp for conducting the database creation and sequencing. The sequencing results were compared with hg19 in the human gene bank(National Center of Biotechnology Information). **Results**　A total of 120.8 Gb original sequence data were generated. The sequencing depth was 36.13 times, the coverage rate reached 99.88%. A total of 4 305 668 SNP loci were detected by the bioinformatic analysis, in which the C:G→T:A motation types were miximal, reaching 12 314 variation sites. **Conclusion**　This study verifies that the data obtained by using the DNA-pool whole genome resequencing method replenishes the Chinese gene database of EH and provides some help for EH gene reasearch in the future.

Key words: essential hypertension;　whole-genome re-sequencing;　SNP;　DNA-pool

原发性高血压是一类由遗传易感性和环境因素相互作用引起的，以血压升高为主要临床表现，伴或不伴有多种心血管危险因素的综合征，是最常见的慢性病，也是心脑血管病最主要的危险因素；世界卫生组织 2012 年 5 月 16 日发布的《2012 年世界卫生统计》报告全球 1/3 成年人患有高血压，这种病症的死亡人数约达中风和心脏病所导致的总死亡人数的一半[1]。此病不仅致残、致死率高，而且严重消耗医疗和社会资源，给家庭和国家造成沉重负担。探索原发性高血压患者基因有助于预防疾病并给予有效的指导治疗。

本研究应用 DNA-pool 技术（使用 Illumina 公司的 Illumina HiSeq 2000 平台）对 100 例中国原发性高血压患者基因进行了高通量重测序，以期在全基因组水平上“捕获”与原发性高血压患者性状相关的突变位点。

1　资料与方法

1.1　一般资料　连续收集 2015 年 3 月至 2015 年 6 月深圳市心血管医院门诊的高血压患者 100 例。患者年龄范围控制在 60 岁以内，无血缘关系，并且所有患者均居住在深圳市。本实验所有患者知晓实验内容及实验方法并签署知情同意书。本研究相关病历资料按《医疗机构病历管理规定》保管。依据中国高血压防治指南（第三版）对高血压的诊断标准对原发性高血压患者进行筛选[2]。纳入标准：（1）收缩压≥140 mm Hg 和/或舒张压≥90 mm Hg 并排除继发性高血压可能的患者；（2）有家族高血压病史（高血压家族史系指家族中父母至少 1 例、直系亲属中有 1 例或 1 例以上患有原发性高血压）；（3）发病年龄≤60 岁。在收集标本时均除外有感染、肿瘤等伴发疾病。并尽量选择遗传背景一致的病例和年龄、性别、地区来源相一致的对照组标本，以避免人群分层造成的假阳性关联。并且向入选本课题研究的患者收集以下资料：一般状况、病史、家族史、个人史、生活饮食习惯及体格检查结果。血压分级以血压达到的最高水平进行评估。连续吸烟 1 年以上为有吸烟史，

*　**基金项目：**广东省深圳市科技研发资金资助项目（JCYJ20140415151845365）。

作者简介：崔文博，女，在读硕士研究生，主要从事心血管疾病基因方向研究。　△　**通信作者**，E-mail：yiwenzhou21@aliyun.com。

戒烟 10 年以上视为不吸烟。连续饮酒超过 1 年以上为有饮酒史。体育锻炼评定标准按照国家相关健康指南设定。

1.2 DNA 的提取、浓度测定及 DNA-pool 制备　空腹抽取静脉血 2 mL，EDTA 抗凝，采用血液基因 DNA 提取试剂盒提取 DNA(武汉华大医学检验所有限公司进行)。以聚合酶链反应(PCR)进行扩增。用紫外分光光度计测量每个 DNA 样品浓度 3 次，取平均值，再稀释为 100 ng/μL，将原发性高血压病例组的 100 个 DNA 样品各取 5 μL 混合构建原发性高血压病例 DNA-pool。

1.3 Illumina HiSeq 2000 进行全基因组高通量重测序　将基因组 DNA 片段化处理至 400～800 bp，凝胶电泳切下所需长度的 DNA 片段。使用 2 种酶分别把 DNA 片段的两端补平。在 3′端挂上一个 A 碱基。接头的 3′端有一个突出的 T，它们通过类似 T-A 克隆的方式连接起来。留下两端连接完整的产物，去掉残缺不全的片段。然后再进行 10～12 循环的 PCR，使产物富集。PCR 完成后可以进行一次凝胶电泳，检测文库的质量和浓度。将扩增的 DNA 文库在 cBot 上进行簇生成。之后在武汉华大医学检验所有限公司进行边合成边测序(SBS)[3-5]。

1.4 生物信息学分析　测序完成之后生成大量的原始数据需要进行生物信息学分析(武汉华大医学检验所有限公司进行)。首先，从原始数据中去除接头序列和含有大量 Ns 的低质量碱基序列。通过这一步骤将产生“纯净数据”。其次，应用 Burrows-Wheeler Aligner(BWA)序列取匹配参考序列[6]。匹配信息应用 BAM 格式文件存储，对所得数据进行以下步骤的处理，包括修复 mate-pair 信息，增加序列组信息，标记由 PCR 引起的重复序列取。通过以上步骤得到最终的 BAM 数据用来进行变体识别。应用 SOAPsnp 检测单核苷酸多态性(SNPs)[7]。应用 Segseq 方法进行自我识别[8]。应用自我识别的方法检测序列上未标明的病毒整合序列。管道质控还包括纯度估计。经过滤之后得到更有意义的变异结果。使用 ANNOVAR 注释变异的结果[8]，用来做进一步的高级分析。为了得到纯净数据在生物信息学分析的每个阶段都要进行包括管道清洁，序列对齐和变体识别等质控方法。质控的步骤用于确定原始序列是否合格，图 1 中 A 曲线与 T 曲线，G 曲线与 C 曲线没有重合，它显示一个不平衡的组合。图 2 中 A 曲线与 T 曲线重叠，G 曲线与 C 曲线重叠，它显示了一个平衡的组合。图 3 在横轴表示序列，纵轴是碱基在相应序列上的分布。图像中的每个点代表碱基在序列上相对应的位置。如果低质量的碱基(＜20)所占百分比非常高，那么这个轨道的测序质量不符合要求。图 4 中的每个点代表碱基在序列上相对应的位置。如果低质量碱基(＜20)所占的百分比很低，表示这个轨道的测序质量。对于不符合要求的测序结果，需要重测。

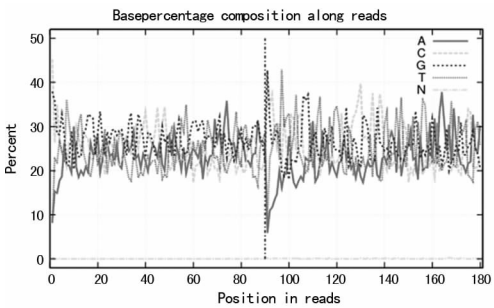

图 1　不平衡原始序列

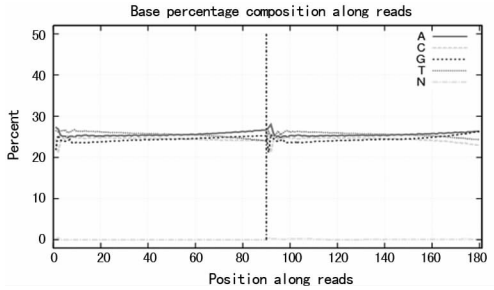

图 2　平衡原始序列

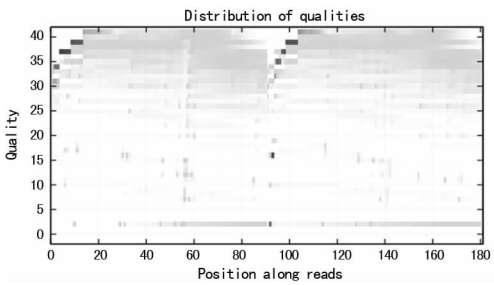

图 3　低质量碱基在序列上的分布

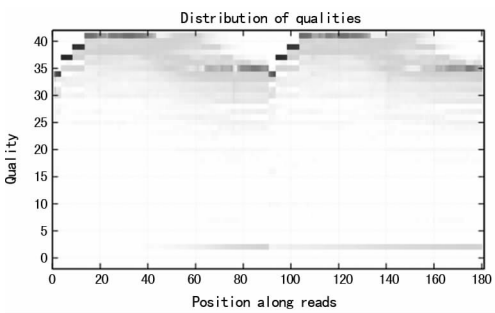

图 4　高质量碱基在序列上的分布

2　结　果

2.1 数据质控　DNA-pool 测序在 HiSeq 2000 平台(美国，圣地亚哥，Illumina 公司)上进行。测序深度约 30 倍共有 120.8 Gb 原始序列数据生成(表 1)。所有数据均经过质控，质控内容包括去除含有接头的 reads、去除未知碱基 N 超过 10％的 reads 和去除低质量的 reads(指质量分数 Q≤5 的碱基数占整个序列的 50％以上)，之后得到纯净数据。在进行任何进一步的分析之前，均需要对原始数据进行过滤质量控制以便检测数据是否合格。去除含有未知碱基 N≥10％的 reads 为 1 653 434，去除低质量 reads 为 52 622 044，去除接头 reads 为 4 318 602，纯净数据/原始数据为 95.63％。

表 1　数据质量控制

项目	纯净数据	原始数据
测序所得 reads	1 283 669 642	1 342 263 722
数据大小	115 530 267 780	120 803 734 980
fq1 文件中 N 碱基的数量	2 338 506	25 057 177
fq2 文件中 N 碱基的数量	3 227 327	59 833 054
fq1 文件中 GC 含量(％)	41.00～41.11	41.09～41.31
fq2 文件中 GC 含量(％)	41.12～41.18	41.2～41.32
fq1 中质量≥20 的碱基(％)	97.59～98.71	96.81～97.87
fq2 中质量≥20 的碱基(％)	94.93～96.35	92.29～94.15
fq1 中质量≥30 的碱基(％)	92.06～95.51	90.98～94.21
fq2 中质量≥30 的碱基(％)	87.62～91.09	85.09～88.12

2.2 质控校准　本研究将人类基因组 build37(hg19)作为本

课题的参考基因组，通过 BWA 与测到的序列进行对比。去除杂质数据之后的纯净数据与参考序列比对，36.52 倍的测序深度得到了 99.89%的极高覆盖率，与参考序列的匹配率达到了 95.77%，错误匹配仅有 0.36%，证明本研究的测序结果质量非常高，为后续的生物信息学分析奠定了坚实基础，对齐质控结果如表 2 所示。本研究通过 SAMtools 软件分析每一个碱基测序深度分布、累计测序深度分布，如图 5 所示 X 轴指测序深度，Y 轴指大于或等于给定的测序深度的碱基所占的百分比。图 6 所示 X 轴指测序深度，Y 轴指大于或等于给定的测序深度的碱基所占的百分比。它大致遵循泊松分布，这表示 exome-capturing 目标区域是均匀采样。

表 2　对齐质量控制数据

项目	数量	项目	数量
纯净数据	1 323 532 708	重复序列所占比率(%)	6.78
纯净碱基(bp)	119 117 943 720	错配碱基	405 962 278
匹配 reads	1 280 793 667	错配率(%)	0.36
匹配碱基(bp)	113 873 432 418	平均测序深度	36.52
匹配率(%)	96.77	覆盖率(%)	99.89
特别 reads	1 225 525 181	至少 4 条 reads 覆盖率(%)	99.30
特别碱基(bp)	108 969 083 786	至少 10 条 reads 覆盖率(%)	97.63
特别 reads 比率(%)	95.68	至少 20 条 reads 覆盖率(%)	93.51
重复 reads	86 789 137		

2.3　SNP 的识别与注释分析结果　本研究使用短寡核苷酸分析软件包 SOAPsnp 识别 SNPs，在每个确定轨迹的单独测序样品概率最高的基因型和标本的一致性序列组装和保存为 CNS 格式。使用一致性序列所确定的基因型和参考序列之间的多态位点可以被过滤，SNP 确定之后，使用进行 ANNOVAR 注释和分类。在全基因组范围内检测到 4 305 668 个 SNP，其中新发现 28 355 个 SNP 位点，在外显子区有 24 900 个 SNP 位点，内含子区有 1 478 121 个 SNP 位点，根据 SIFT(sorting intolerant from tolerant)评分，低于 0.05 就被认为是有可能存在损害，本研究发现有 3 106 个 SNP 位点 SIFT 评分低于 0.05，详见表 3。

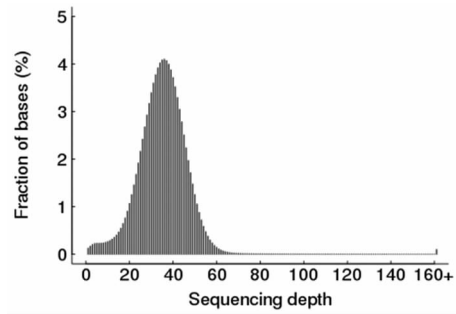

图 5　测序深度分布

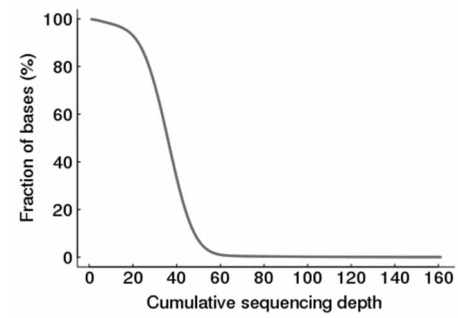

图 6　累计测序深度分布

表 3　SNPs 数据

类别	数量	类别	数量
总计 reads	4 305 668	剪切位点	174
千人基因组和 dbSNP135 数据库	4 142 411	非编码 RNA	108 685
千人特异性基因组数据库	58	5′端非翻译区	5 243
dbSNP135 特异性基因数据库	134 844	5′-3′非翻译区	11
dbSNP 出现率(%)	99.34	内含子区	1 478 121
新发现基因位点	28 355	转录起始位置上游	22 869
纯合子	553 072	转录起始位置上游和转录终止位置下游	698
杂合子	3 752 596	转录终止位置下游	24 545
同义突变	13 528	基因间区	2 612 499
错义突变	11 594	SIFT 评分<0.05	3 106
密码子变为终止密码子突变	82	碱基转换与颠换比	2.122 9
密码子变为非终止密码子	36	dbSNP 区域碱基转换与颠换比	2.129 6
外显子区域	24 900	新发现位点碱基转换与颠换比	1.362 5
外显子和剪切位点区域	340		

3　讨　　论

目前对于原发性高血压易感基因的研究国内外主要采取候选基因研究、基于遗传标志的连锁分析及全基因组关联研究等方法，通过这些方法，在高血压易感基因的探索上取得了一些成果。在候选基因研究中，目前研究主要集中在肾素-血管紧张素-醛固酮系统、G-蛋白信号传导系统、儿茶酚胺肾上腺素能系统、离子通道、炎症、内皮相关因子等[9-15]。但是针对高血压相关基因位点的研究多集中于常见变异，对于罕见变异在高血压发病机制中的作用还知之甚少。研究发现的与血压相关的基因位点只在特定的人群中存在，由于遗传背景、环境等因素的差异，大多不能进一步在不同人种之间进行重复验证；研究大多需要巨大的样本量才能得到有显著关联性的结果，而且发现的单个基因位点的变异对血压的影响都极其微弱。新近的研究利用多个已经证实的对血压有影响的基因位点而设计的危险评分对不同人群的高血压的风险进行评估，发现该危险评分与血压升高和高血压发病独立相关，说明利用多个基因位点变异的累积效应来评价高血压的发病风险可能获益会更大[16-17]。全基因组关联研究(GWAS)在时间和成本上限制了许多研究的可行性和有效性[18-20]。

本研究应用 DNA-pool 为中国原发性高血压进行了全基因

组重测序。共生成 120.8 Gb 原始序列数据。其中4 305 668个 SNP 位点，36.13 倍的测序深度得到了 99.88%的极高覆盖率，与参考序列的匹配率达到了 95.68%，错误匹配仅有0.36%，体现了 DNA-pool 的高效性与准确性，为探索常见的家族性疾病的遗传基础提供了一个高效的方法。DNA-pool 已被证实是一种进行基因个体化诊断和治疗的有效方法[20-21]。这意味着对上千例的样本进行 GWA 研究，包括对遗传变异进行大规模检测，在技术上和经济上是可行的。此外，本研究还发现了基因拷贝数变异、插入缺失、单核苷酸变异等等。

通过本研究，本组总结出使用 DNA-pool 进行全基因组重测序相关研究的经验，并提出质控的重要性。在这项研究中，测序轨道产生了含有大量 Ns 或低质量的碱基的 120.8 Gb Illumian 原始数据。小的系统误差很容易对真正的关联产生影响[21-22]。本研究使用包括管道清洁、序列对齐和变体识别等质控方法最小化系统误差从而得到纯净数据。这种排序方法也导致序列冗余达到平均 36.13 倍。因此，一致性序列精度比较高，特别适用于对杂合的等位基因的识别[22]。应用 DNA-pool 进行全基因组重排序技术拥有非常高的吞吐量，1 亿 DNA 片段可以并行测序芯片。用于本研究的 Illumina 公司 HiSeq 2000 平台可以提供每天 55 Gb 的高质量的数据，大大提高了 GWAS 研究在技术、时间和经济上的可行性。

参考文献

[1] World Health Organization. World health statistics 2012 [M]. Geneva, Switzerland: World Health Organization, 2012.

[2] 中国高血压防治指南修订委员会. 中国高血压防治指南 2010[J]. 中华高血压杂志，2011，19(8)：701-743.

[3] Mardis ER. Next-generation DNA sequencing methods. [J]. Ann Rev Genomics Hum Genet, 2008, 9(1): 387-402.

[4] Wang J, Wang W, Li R, et al. The diploid genome sequence of an Asian individual [J]. Nature, 2008, 456 (7218): 60-65.

[5] Wheeler DA, 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] Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform [J]. Bioinformatics, 2010, 25(5): 1754-1760.

[7] Li R, Li Y, Fang X, et al. SNP detection for massively parallel whole-genome resequencing [J]. Genome Res, 2009, 19(6): 1124-1132.

[8] Chiang DY, Getz G, Jaffe DB, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing[J]. Nat Methods, 2009, 6(1): 99-103.

[9] Tchelougou D, Kologo JK, Karou SD, et al. Renin-angiotensin system genes polymorphisms and essential hypertension in Burkina Faso, West Africa[J]. Int J Hypertens, 2015, 2015: 979631.

[10] Santoro ML, Santos CM, Ota VK, et al. Expression profile of neurotransmitter receptor and regulatory genes in the prefrontal cortex of spontaneously hypertensive rats: relevance to neuropsychiatric disorders [J]. Psychiatry Res, 2014, 219(3): 674-679.

[11] Grandbois J, Khurana S, Graff K, et al. Phenylethanolamine N-methyltransferase gene expression in adrenergic neurons of spontaneously hypertensive rats[J]. Neurosci Lett, 2016, 635: 103-110.

[12] Mir SA, Zhang K, Milic M, et al. Analysis and validation of traits associated with a single nucleotide polymorphism Gly364Ser in catestatin using humanized chromogranin A mouse models[J]. J Hypertens, 2016, 34(1): 68-78.

[13] Abramova TO, Smolenskaya SE, Antonov EV, et al. Expression of catechol-O-methyltransferase (Comt), mineralocorticoid receptor(Mlr), and epithelial sodium channel (ENaC) genes in kidneys of hypertensive ISIAH rats at rest and duringresponse to stress[J]. Genetika, 2016, 52 (2): 206-214.

[14] Liang YF, Zhang DD, Yu XJ, et al. Hydrogen sulfide in paraventricular nucleus attenuates blood pressure by regulating oxidative stress and inflammatory cytokines in high salt-induced hypertension[J]. Toxicol Lett, 2017, 270: 62-71.

[15] Jiang X, Sheng H, Li J, et al. Association between renin-angiotensin system gene polymorphism and essential hypertension: a community-based study[J]. J Hum Hypertens, 2009, 23(3): 176-181.

[16] Chen X, Zhou B, Hou X, et al. Associations between CLCNKA_B tag SNPs with essential hypertension and interactions between genetic and environmental factors in an island population in China[J]. Clin Exp Hypertens, 2015, 37(7): 519-525.

[17] Damasceno A, Azevedo A, Silvamatos C, et al. Hypertension prevalence, awareness, treatment, and control in mozambique: urban/rural gap during epidemiological transition. [J]. Hypertension, 2009, 54(1): 77-83.

[18] Baum AE, Akula N, Cabanero M, et al. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder[J]. Mol Psychiatry, 2008, 13(2): 197-207.

[19] Galvan A, Falvella FS, Frullanti E, et al. Genome-wide association study in discordant sibships identifies multiple inherited susceptibility alleles linked to lung cancer[J]. Carcinogenesis, 2010, 31(3): 462-465.

[20] Forstbauer LM, Brockschmidt FF, Moskvina V, et al. Genome-wide pooling approach identifies SPATA5 as a new susceptibility locus for alopecia areata[J]. Eur J Hum Genet, 2011, 20(3): 326-332.

[21] Clayton DG, Walker NM, Smyth DJ, et al. Population structure, differential bias and genomic control in a large-scale, case-control association study[J]. Nat Genet, 2005, 37(37): 1243-1246.

[22] Zondervan KT, Cardon LR. The complex interplay among factors that influence allelic association[J]. Nat Rev Genet, 2004, 5(2): 89-100.

（收稿日期：2016-12-02　修回日期：2017-02-18）