遗传 HEREDITAS (Beijing) 20(6): 11～14 1998

应用大规模测序技术和生物信息学研究造血干/祖细胞的基因表达及新基因的识别和克隆

吴济生 茅 矛 付 刚 周 隽 张庆华 顾 健
黄秋花 沈 宇 俞亚萍 徐淑华 王亚新 陈 竺
(上海第二医科大学附属瑞金医院，卫生部和上海市人类基因组研究重点实验室，上海 200025)

摘 要 从新生儿脐血和成人骨髓中分选出造血干/祖细胞(HSC/HPC)，构建成 cDNA 文库，对其进行大规模表达序列标签(EST)测序，通过生物信息学等手段分析基因表达谱，并进行新基因的全长 cDNA 克隆。在所测的 10 512 条可分析 EST 序列中，有 9 866 条来自脐血 $CD34^+$细胞，其中 4 697 条(47.6%)为已知基因，2 603 条(26.4%)为已知 EST，1 415 条(14.3%)代表未知 EST。在已知基因中，8.2%基因与造血相关，22.7%涉及细胞代谢、结构和迁移，13.0%与细胞分裂和防御相关，26.2%与 RNA、蛋白质的合成相关，10.6%和细胞信号传递有关。对一些已知和未知的 EST，综合测序、生物信息学等方法，进行全长克隆，已获得 23 个新基因的全长 cDNA。

关键词 造血干/祖细胞，大规模测序，表达序列标签，基因表达，生物信息学，全长 cDNA 克隆
中图分类号 Q343

Study on Gene Expression of Hematopoietic Stem/ Progenitor Cells and Identification and Cloning of Novel Genes with Large-scale Sequencing and Bioinformatics

WU Jisheng MAO Mao FU Gang ZHOU Jun ZHANG Qinghua GU JianI
HUANG Qiuhua SHEN Yu YU Yaping XU Shuhua WANG Yaxin CHEN Zhu
(Key Laboratory for Human Genome Research, Ruijin Hospital, Shanghai Second Medical University, Shanghai 200025)

Abstract Hematopoietic stem/progenitor cells were isolated from umbilical cord blood and adult bone marrow, and subject to cDNA library construction. The gene expression pattern in $CD34^+$cells and the identification and cloning of novel genes were performed by sequencing ESTs and analyzing them with the tools of bioinformatics. Among the obtained 10 512 ESTs which could be further analyzed, 9,866 were from umbilical cord blood where 4 697(67.6%) were known genes, 2 603(26.4%) were known ESTs and 1 415(14.3%) represented novel ESTs. Within the identified genes, 8.2% was involved in hematopoiesis, 22.7% was associated with cell metabolism, structure and mobility, 13.0% was linked to cell division and defence, 26.2% was related to RNA protein synthesis and 10.6% was related with cell signal transduction. In parallel, we developed an efficient working system combining sequencing, bioinformatics, etc. and obtained 23 full-length cDNAs from both known and novel ESTs identified in this work.

Key words Hematopoietic stem/progenitor cell, Large-scale sequencing, Expressed sequence tag, Gene expression, Bioinformatics, Full-length cDNA cloning

大规模测序技术是刚兴起的一项技术，一般年测序达 100 万碱基对以上[1]。就基因识别而言，方法有基因组

DNA 测序和源于 mRNA 的 cDNA 测序。为获知整个人类基因的状况，基因组测序是一种非常有效的策略，但必须辅以有效的计算机和软件及昂贵的测序费用；后者，因长度一般为 500～8 000 bp，故易于克隆和测定，但因某些基因表达的特异性或低表达，难免造成遗漏。目前，要多、快、好、省地识别和克隆新基因，cDNA 测序可能是一条捷径。此外，生物信息学在大规模测序中起着极为重要的作用，它可将未知基因和已知基因进行同源性比较，从而推测未知基因的可能功能，并可利用电脑对 DNA 序列进行查询、组装、翻译甚至研究高级结构。

造血干细胞(HSC)是造血组织中一类既能自我更新、又能分化为各类终末血细胞的细胞[2]。造血干／祖细胞的增殖、分化所形成的血细胞动态平衡过程，即为造血的过程。CD34 抗原在干细胞为强阳性，在早期祖细胞仍为阳性，可持续到晚期祖细胞[3]。约 1.5%的骨髓单核细胞(MNC)表达 $CD34^+$，而 $CD34^+$细胞群中 90%以上为祖细胞[4]。通常采用 $CD34^+CD38^-in^-$、$CD34^+CD33^-HLADR^-$参数抗原精选干细胞和祖细胞[5,6]。造血系统不仅发挥着重要的生理功能，而且在不同发育阶段基因表达亦有差异。它对细胞的分化、增殖和凋亡的精细调控使之成为研究发育调控和细胞分化的良好材料，同时造血生长因子在整个生物技术和生物制药工程中又处于中心位置，因此研究造血系统对整个生物医学工程来说是相当重要的。鉴于 HSCs 在整个造血系统中的重要地位，我们以不同发育阶段(胎儿和成人)的造血干／祖细胞($CD34^+$细胞)和造血干细胞($CD34^+$／$CD38^-$(细胞)研究其基因表达。

1 材 料 和 方 法

取新鲜脐血(共 643 例，约 15L，来自上海市五所医院产房)与成人骨髓(共 16 例，约 2.6 升，来自正常献髓者)，抗凝后 Ficoll 分离出单个核细胞，以抗 CD34 抗体标记后经 MACS 和 FACS 分选出 $CD34^+$／$CD38^-$(与 $CD34^+$／$CD38^+$两细胞群，用 $TRI_{ZOL}LS$ 试剂(GIBCO)抽提总 RNA。因 RNA 总量较少，选用 CLONTECH 的试剂作 RT／PCR 扩增得到 cDNA，加装衔接子，再与 ZAPII 载体(Strantagene 公司)相连，分别构建成脐血和成人骨髓 $CD34^+$细胞的 cDNA 文库。在辅助噬菌体帮助下进一步环化、包装、裂解，再感染 SOLR 细胞(Stratagene)，作 *in vivo* excision，用含 IPTG 和 x-gal 的氨苄平板鉴定，挑选白色菌落，扩增后提取质粒。测序反应用 PE 公司的 Dye-Primer 试剂，在 9600 PCR (Perkin Elmer)上进行延伸反应，产物乙醇沉淀后，由 ABI PRISM 377 DNA 自动测序仪(Perkin Elmer)进行测序，通过 Power Macintosh 7200／90(Apple)计算机控制，并用 Data Collection、DNA Sequencing Analysis 和 Factura (Perkin Elmer)等软件先对原始数据进行记录、收集、形成数据库、分析并读出其碱基组成，再经 Factura 去除载体和模糊序列。清晰序列输到 SUN 工作站(Enterprise 150)进行同源性分析。工作站数据库为 GenBank Release 100.0 和 EMBL Release 50.0 等，含 30 万条基因和 76 万条 EST 序列，所用的 GCG 软件包可对序列进行编辑、比较、装配和搜索。所测序列先用 FastA 软件和自己的数据库进行比较，筛掉冗余或有部分重叠的序列，再用 Blast 软件与 GenEMBL 进行比较，一般与已知基因有 95%以上同源性的序列即为已知基因。较低或无同源性的片段再与 EST 库进行比较，同上标准筛出已知 EST，最后所剩即是可能的未知基因，再根据与其相关的基因的同源性和功能及其全长大小((3.0kb)，挑选序列进行全长实验并明确其功能。

2 结 果 与 讨 论

经 MACS 分选后富集了 $CD34^+$细胞，纯度达 95%～99%，再经 FACS 分选，纯度可达 99%。表 1 示分选前后克隆形成率的变化，证明细胞功能仍然完好。RNA 抽提量为：脐血 $CD34^+$细胞：238.7μg；$CD34^+$／$CD38^-$细胞：0.3μg；骨髓 $CD34^+$细胞：158.8μg；$CD34^+$／$CD38^-$细胞：0.3μg。共构建了脐血 $CD34^+$、$CD34^+$／$CD38^-$细胞和成人骨髓 $CD34^+$细胞三个 cDNA 文库。滴度测定：一般在 10^5～10^6pfu／μg；重组效率：蓝白噬菌斑比例在 1:1 左右，即重组子占 50%；插入长度见表 2 所示。截止到 1997 年 12 月 10 日，已测了 13 669 条 EST。有 10 512 条 EST 可作进一步分析，其中，9 866 条来自脐血 $CD34^+$细胞，646 条来自骨髓 $CD34^+$细胞。在来自脐

血的 EST 中，4 697 条(47.6%)序列代表已知基因，2 603 条(26.4%)代表已知 EST，1 415 条(14.3%)代表了未知基因。已知基因按其细胞生物学功能分为: (1) 造血; (2) 细胞分裂; (3) 细胞信号传导／联系; (4) 细胞结构／迁移; (5) 细胞／机体防御; (6) 基因／蛋白表达; (7) 代谢; (8) 功能不明。从其分布来看，造血相关基因占总数的 8.2%、35.7%与基础能量代谢、细胞结构、细胞分裂相关，26.2%与 RNA、蛋白的合成相关，另有 10.6%与细胞信号传递相关。并已获得 23 个全长 cDNA 克隆，其中部分全新的人类新基因的功能也已明确，如人类 ABC(ATP Binding Cassette)转运体基因。

表 1　MACS 分选前后集落形成率比较

	CFU-GM (unit／10^4)	CFU-E (unit／10^4)	CFU-MK (unit／10^4)
单核细胞	12.2	9.6	6.8
$CD34^+$细胞	512	252	208
$CD34^+$细胞富集倍数	42	26	31

表 2　随机克隆的插入片段长度分布

插入片段长度(kb)	脐血 $CD34^+$细胞文库(%)	骨髓 $CD34^+$细胞文库(%)
< 0.5	47	28
0.5～1.0	18	44
1.0～2.0	24	28
2.0～3.0	11	

大规模测序各步骤是紧密相连的，任何一环出现问题都会影响结果，必须对整个流程进行严密质量监控。因此，材料必须来源于健康人体，在离体后 4 小时内即进行分选。检测滴度、重组效率和插入长度来保证文库质量。滴度低表明插入序列和载体连接时出现问题，重组率较低可能是由于插入片段较短引起。随机挑选克隆，PCR 扩增后，通过琼脂糖凝胶电泳检测插入片段长度。所抽质粒，以超螺旋多者为佳，RNA 和基因组 DNA 的污染应少或无。所测序列，峰形应尖且高度适中，重叠区少，若模板纯度不高，会致背景信号复杂，影响分析。测序结果中 ACTG 比例应维持在一定水平，如某一碱基含量明显减少，可能在测序反应中有加样错误。数据经 Factura 处理后，再人为去除长度小于 100bp 或未读出比例大于 3%的序列。

PCR 方法建库对 RNA 需要量少，一般纳克量即可，文库含较高比例的全长 cDNA，且消除了基因组和 poly(A)RNA 的污染，可省去 mRNA 纯化的步骤，但在一定程度上会影响基因表达丰度的代表性，扩增时，循环数越多，越易导致一些非特异性的扩增。文库的污染有: 基因组 DNA、线粒体 DNA 和核糖体 RNA。基因组的污染多为 Alu，且所得 EST 和已知 Alu 的同源性均达 70%～90%。在体外，Alu 可由 RNA polⅢ转录，经自身引物反转录成 cDNA，再插入到基因组 DNA 中，这可能是造成其冗余度(Redundency)较高的原因[7]。冗余序列是指插入外源 cDNA 片段完全一致的序列，不包括来源于同一基因而有部分重叠的序列。冗余度较高原因有: 属高表达基因，如一些管家基因、PCR 人为导致和 *invivo* excision 所致。冗余度随 *invivo* excision 的时间延长而提高，因为 *invivo* excision 即有减切的过程，又有复制的过程，时间过长势必造成复制过度而冗余度增加。冗余度以每次 *in vivo* excision 后测定的前 150 条序列中每条序列出现的平均次数来计算。*in vivo* excision 的时间为 2.5、2.0 及 1.5 小时的时候，冗余度分别为 5.5、1.8 和 1.2。

就库的整体情况，造血特异基因在已知基因中所占比例并不高，而一些管家基因，包括核糖体蛋白、组蛋白及与代谢相关的酶等有较高表达，它们在所有组织细胞中表达可能都较高，这无益于本研究。表 3 为所测 10 512 条序列的总体评价。

由于才刚起步，所获信息尤其是与造血相关的信息还不多，但通过已做的工作，一套稳定的大规模测序体系已经建立起来，造血干／祖细胞基因表达的轮廓也初步被勾画出来，近二十多个全长 cDNA 已被克隆，相信随着工作的进一步深入，越来越多有意义的基因会被我们所获取。

表 3 文库质量的评估(基于 1 492 条 cDNA 序列)

序 列	克隆数	含 量(%)	序 列	克隆数	含 量(%)
细菌 DNA	0	0	无插入片段克隆	16	1.1
Alu 顺序	36	2.4	短插入片段克隆	54	3.6
线粒体 DNA	53	3.6	有用序列	1198	80.3
核糖体 RNA	18	1.2	无用序列	294	19.7

参 考 文 献

1 Hunkapiller T, Kaiser R J, Koop B F *et al.* Large-scale and automated DNA Sequence determination. Science, 1991, 254: 59～67

2 Morrison S J, Uchida N, Weissman I L. The Biology of Hematopoietic Stem Cells. Annu. Rev. Cell Dev. Biol., 1995, 11: 35～71

3 Krause D S, Fackler M J, Civin C I *et al.* CD34: Structure, Biology, and Clinical Utility. Blood, 1996, 87: 1～13

4 Lu L, Xiao M, Shen R-N *et al.* Enrichment, characterization, and responsiveness of single primitive $CD34^{+}$++ human umbilical cord blood hematopoietic progenitors with high proliferative and replating potential. Blood, 1993, 81: 41～48

5 Traycoff C M, Abboud M R, Laver J *et al.* Evaluation of the in vitro behavior of phenotypically defined populations of umbilical cord blood hematopoietic progenitor cells. Exp. Hematol., 1994, 22(2): 215～222

6 Almici C, Carlo-Stella C, Wagner J E *et al.* Umbilical cord blood as a source of hematopoietic stem cells: from research to clinical application. Haematologica, 1995, 80: 473～479

7 Watson J D *et al.* Molecular Biology of Gene, 1987, 4th edn, pp. 668～670

1998-07-20 收稿, 1998-09-25 修回.

中国人群 5-羟色胺 2A 受体基因中 T102C 多态性与精神分裂症的联系①

李 胜[1)] 顾牛范[2)] 冯国鄞[2)] 刘万清[1)] 沈 韬[1)] 贺 林[1,3)]

(1 中国科学院上海生命科学研究中心, 上海 200031)

(2 上海市精神卫生中心, 上海 200030) (3 中国科学院上海脑研究所, 上海 200031)

摘 要 在研究 5-羟色胺 2A 受体基因多态性与精神分裂症的关联分析中, 调查了 202 例精神分裂症患者及 202 例正常对照。各相匹配组间比较未发现基因型和等位基因频率的显著性差异。结果提示, 在中国人群中 5-羟色胺 2A 受体的静态 T102C 突变与精神分裂症之间不存在关联。

关键词 5-羟色胺 2A 受体基因多态性, 精神分裂症

中图分类号 Q987

Analysis of Association Between T102C Polymorphism in the Serotonin 2A Receptor Gene and Schizophrenia in Chinese Population

LI Sheng[1)] GU Niufan[2)] FENG Guoyin[2)] LIU Wanqing[1)] SHEN Tao[1)] HE Lin[1,3)]

(1 Shanghai Research Center of Life Science, Chinese Academy of Sciences, Shanghai 200031)

①本课题研究得到了中国科学院及上海市科委的经费支助。