PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2008, 35(8): 964~968
www.pibb.ac.cn

丙型肝炎病毒基因分型及 core 蛋白区的分子演化分析

万祥辉　曾照芳*　杨细媚
(重庆医科大学检验系，临床检验诊断学省部共建教育部重点实验室，重庆市重点实验室，重庆 400016)

摘要　目前有多种方法对丙型肝炎病毒进行基因分型，但尚无一个“金标准”. 为确定丙型肝炎病毒基因分型的最佳区域，从 GenBank 中筛选出 15 条对各成熟肽区域有注释，来源于不同国家地区的丙型肝炎病毒全基因组序列，分别对 5′ UTR 区、core 区、E1 区、E2 区及 NS5B 区建系统进化树. 结果发现，以 5′ UTR 区建树基因分型不完全正确，而以 core 区、E1 区、E2 区及 NS5B 区建树，基因分型均完全正确，但同一基因型间的核苷酸演化距离存在差异. 计算 5 条 1a 型序列的 core 区、E1 区、E2 区、NS5B 区的核苷酸演化距离并和全基因组序列核苷酸演化距离比较，结果发现，NS5B 蛋白区基因分型最能反映病毒株间的演化关系. 同时分析各序列的 core 区的分子演化，为发明针对 core 区的新的 PCR-RFLP 基因分型方法提供新思路.

关键词　丙型肝炎病毒，基因型，分子演化，核心蛋白
学科分类号　Q811

丙型肝炎病毒(HCV)于 1989 年首次被 Choo 等克隆[1]，是慢性肝病的主要病原体，属黄病毒科，基因组(HCV RNA)为单正链 RNA(ssRNA)，目前全球约有 1.7 亿 HCV 感染患者[2]. HCV 基因组全长约 9.5 kb，分为 5′非编码区(5′UTR)、编码区和 3′非编码区(3′UTR). 编码区编码 3 个结构蛋白，即 core、E1 和 E2 蛋白，7 个非编码蛋白，即 p7、NS2、NS3、NS4A、NS4B、NS5A 和 NS5B 蛋白，基因组中 5′UTR 区 、core 区最保守，E1 区、E2 区最易变异[3, 4]. 不同国家，不同地区 HCV 基因型分布不同[5]，基因分型有利于丙型肝炎流行病学的研究. HCV 基因型还与干扰素治疗丙型肝炎的疗效及肝移植后再感染有关[6, 7]. 目前 HCV 基因分型的方法有多种，尚没有建立“金标准”，最准确的分型方法是全基因组测序，但这在临床难以实现. 本研究从 GenBank 中筛选出 15 条对各成熟肽区域有注释，来源于不同国家地区的丙型肝炎病毒全基因组序列，对序列的 5′UTR 区、core 蛋白区、E1 蛋白区、E2 蛋白区及 NS5B 蛋白区建系统进化树，找出最能替代全基因组分型的区域，并分析 core 区的分子演化情况.

1　材料和方法

从 GenBank (http://www.ncbi.nlm.nih.gov/) 中筛选出 15 条对各成熟肽区域有注释，来源于不同国家地区的 HCV RNA 全基因组序列，序列的接受号、基因型和国家见表 1. 根据注释分别选取全基因组序列的 5′UTR 区、core 区、E1 区、E2 区及 NS5B 区的基因序列，用 PHYLIP 的 seqboot.exe，dnaml.exe，consense.exe 程序(500 次 Bootstrap 检验)对上述区域分别建系统进化树，用 TreeView 观察进化树. 根据结果比较分析，用 MEGA4[8] (Kimura-2-parameter 模型)计算 5 条 1a 型序列的 core 区、E1 区、E2 区、NS5B 区的核苷酸演化距离并和全基因组序列核苷酸演化距离比较. 用 Clustalx 1.84[9]，MEGA4 对 15 条序列的 core 蛋白区进行变异分析.

* 通讯联系人.
Tel: 023-68485095，E-mail: zeng000@126.com
收稿日期：2007-12-11，接受日期：2008-02-11

Table 1　The accession number and genotype of 15 complete genome sequences

Genotype	Accession number
1a	EU260396 EU260395 EU155380 EU234065 EU155355
1b	EU155377 EU155317 EU155381
1c	D14853
3a	NC_009824
3b	D49374
6a	DQ480524 DQ480523 DQ480522 DQ480521

2　结　　果

2.1　5′UTR、core、E1、E2 及 NS5B 区的系统进化树

5′UTR 区的系统进化树见图 1，core 区的系统进化树见图 2，E1 区的系统进化树见图 3，E2 区的系统进化树见图 4，NS5B 区的系统进化树见图 5．从图 1 中可以看出 5′UTR 区分型未能将 D14853 株和 EU155355 株分开，存在分型错误．图 2、图 3、图 4、图 5 显示，core 区、E1 区、E2 区、NS5B 区均能正确分型，但同一基因型不同病毒株的核苷酸演化距离不同．

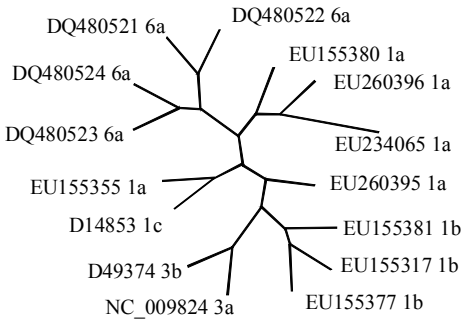

Fig. 1　Phylogenetic tree construction of the clear sequence on 5′ UTR region

It can not be able to distinguish between D14853 strain and EU155355 strain existence of the wrong type.

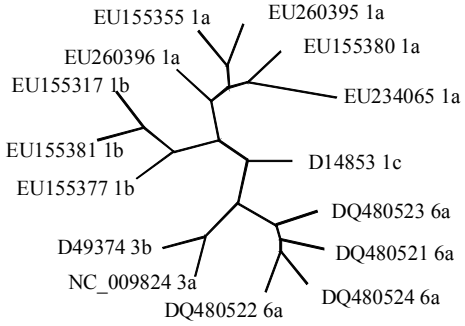

Fig. 2　Phylogenetic tree construction of the clear sequence on core region

It shows the genotypes are wholly correct on core region, but the phylogenetic distances of different virus strains which have the same genotypes are different with figures 3, 4, 5.

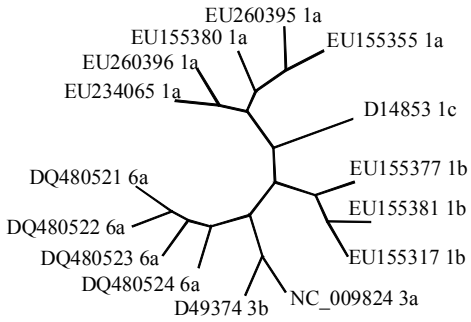

Fig. 3　Phylogenetic tree construction of the clear sequence on E1 region

It shows the genotypes are wholly correct on E1 region, but the phylogenetic distances of different virus strains which have the same genotypes are different with figures 2, 4, 5.

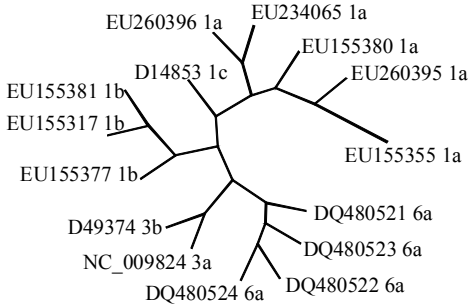

Fig. 4　Phylogenetic tree construction of the clear sequence on E2 region

It shows the genotypes are wholly correct on E2 region, but the phylogenetic distances of different virus strains which have the same genotypes are different with figures 2, 3, 5.

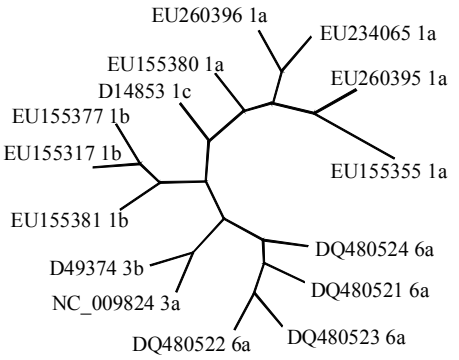

Fig. 5　Phylogenetic tree construction of the clear sequence on NS5B region

It shows the genotypes are wholly correct on NS5B region, but the phylogenetic distances of different virus strains which have the same genotypes are different with figures 2, 3, 4.

2.2　计算 5 条 1a 型序列的全基因组序列及 core、E1、E2、NS5B 区的核苷酸演化距离

core 区和 E1 区的核苷酸演化距离见表 2，E2 区和 NS5B 区的核苷酸演化距离见表 3，全基因组核苷酸演化距离见表 4．全基因组演化距离计算显示，EU260396 株与 EU155355 株的演化距离最

远，而与 EU155380 株的演化距离比 EU260395 株接近．core 区和 E2 区演化距离计算均显示 EU260396 株与 EU260395 株的演化距离最远，E1 区演化距离计算未能将 EU155380 株和 EU260395 株区分开，而 NS5B 区演化距离计算显示的各病毒株的演化关系与全基因组序列演化距离计算完全一致，所以以 NS5B 区最能代替全基因组序列进行基因分型，最能反映病毒株间的演化距离．

Table 2 Pairwise distance of HCV genotype 1a sequence between core region (in lower left) and E1 region (in upper right)

No.	1	2	3	4	5
1.**EU260396**	0.000	0.041	0.104	**0.084**	**0.084**
2.EU234065	0.014	0.000	0.087	0.072	0.078
3.EU155355	0.023	0.023	0.000	0.100	0.081
4.EU155380	0.029	0.025	0.042	0.000	0.092
5.**EU260395**	0.038	0.034	0.036	0.042	0.000

Bold indicates distinguished part. They show that EU260396 strain with the evolution relation of EU260395 strain is the farthest on core region and the phylogenetic distances are not be able to distinguish between EU155380 strain and EU260395 strain from E1 area.

Table 3 Pairwise distance of HCV genotype 1a sequence between E2 region (in lower left) and NS5B region (in upper right)

No.	1	2	3	4	5
1.**EU260396**	0.000	0.028	**0.066**	**0.048**	**0.057**
2.EU234065	0.069	0.000	0.064	0.048	0.060
3.EU155355	0.134	0.129	0.000	0.069	0.048
4.EU155380	0.136	0.120	0.150	0.000	0.059
5.**EU260395**	0.153	0.161	0.102	0.170	0.000

Bold indicates distinguished part. They show that EU260396 strain with the evolution relation of EU260395 strain is the farthest on E2 region and the evolution relation is exactly coincidence with the whole genome sequence from NS5B area.

Table 4 Pairwise distance of HCV genotype 1a complete genome sequence

No.	1	2	3	4	5
1.**EU260396**	0.000				
2.EU234065	0.043				
3.EU155380	**0.074**	0.070			
4.EU260395	**0.082**	0.077	0.092		
5.**EU155355**	0.092	0.088	0.103	0.067	0.000

Bold indicates distinguished part. They show that EU260396 strain with the evolution relation of EU155355 strain is the farthest, and with the evolution relation of EU155380 strain is closer than EU260395 strain.

2.3 core 蛋白区的分子演化分析

ClustalX1.84 多序列联配，MEGA4 分析发现，15 条序列的 core 蛋白区均为 573 个核苷酸，编码 191 个氨基酸，转换率为 7.0%，颠换率为 6.6%，转换率与颠换率之比为 1.1，Parsimony-information sites(包含 2 种以上碱基，而其中至少有 2 种碱基经常出现的位点) 有 114 个，Singleton site(包含 2 种以上碱基，其中有一种为主要类型的位点) 有 54 个，序列相似程度为 86%，可见 core 蛋白区虽是 HCV RNA 编码区中最保守的区域但异质性也较大且突变以转换为主．分析序列的 Singleton site，发现 1a 型和 6a 型突变位点较少．1a 型包括 EU260395 序列在第 18 位碱基的变异(A→C)，第 26 位碱基的变异(G→A)，第 528 位碱基的变异(C→T)．EU155355 序列在第 72 位碱基的变异(C→T)，第 451 位碱基的变异(C→T)．EU155380 序列在第 44 位碱基的变异(C→T)，第 243 位碱基变异(C→T)，第 518 位碱基变异(C→T)．6a 型包括 DQ480524 序列在第 34 位碱基的变异(A→C)，第 111 位碱基的变异(G→A)，第 280 位碱基的变异(A→C)，第 312 位碱基的变异(G→A)，第 498 位碱基的变异(A→G)，第 531 位的碱基变异(C→T)．DQ480522 在第 151 位碱基的变异(A→G)．3b 型 D49374 突变位点较多，包括第 22、33、45、46、61、77、98、105、109、117、128、143、194、231、276、295、354、390 位碱基的变异共 20 个突变位点．

3 讨 论

目前丙型肝炎病毒基因分型主要是以核苷酸序列为基础，主要有基因特异探针杂交法(LIPA)，PCR-RFLP 基因分型法，型特异性引物扩增 HCV 基因组特异区段法(T-S PCR)，特异引物错配延伸法(PSMEA)，异源分子迁移率法(HMA)和直接测序法．直接测序法中目前常用的区段是 5′UTR 区、core 区、NS5B 区，这 3 个区段相对保守，本文将高变区 E1 区、E2 区与 3 个相对保守区段一起分析，结果发现 E1 区、E2 区对 15 株不同国家地区来源的序列分型是正确的．国际上一些著名的病毒学家推荐以 HCV 基因组的 C 区、E1 区或 NS5B 区基因来反映全基因组序列的变异情况，对多株 HCV NS5B 区的序列分段发现，此区的变异足以区分不同系、不同型、亚型及株[10]．本文对多区段分别建立系统进化树并分析不同区段的核苷酸演化距

离与全基因组序列的核苷酸演化距离的接近程度，结果证实了 NS5B 最能反映全基因组序列的变异情况.

5′UTR 区是 HCV 基因组中最保守区域，邱国华等[11]采用 5′UTR ABC 程序酶切的方法进行基因分型，但建立在 5′UTR 基础上的分型方法已有研究表明有 5% ～10%的 1a，1b 型不能鉴别，也不能区分 2a 和 2c 型[12, 13]. 图 1 也显示以 5′UTR 区基因分型不完全正确，没有将 1a 型的 EU155355 株和 1c 型的 D14853 株区分开. Core 区异质性比 5′UTR 区大，但该区是 HCV 编码区最保守区段，分子演化分析显示，core 区不同基因型的相似性程度为 86%，系统树表明该区对 15 株不同国家地区的病毒株的分型是正确的. 分析 core 区的碱基突变位点是否存在酶切位点，如果突变位点存在酶切位点或根据突变位点用引物错配法设计酶切位点[14]则可用限制性内切酶酶切，根据片段长度多态性区分开不同的基因型，再根据酶切位点的个数限制，限制性内切酶的活性，PCR 反应体系筛选最佳的限制性内切酶组合，可以探索在 core 蛋白区建立新的 PCR-RFLP 基因分型方法.

本研究从生物信息学的角度进行序列分析，比较丙型肝炎病毒 5′UTR 区、core 区、E1 区、E2 区、NS5B 区 5 个区域的基因分型结果及演化距离，发现，5′UTR 区不能区分 1a 型和 1c 型，高变区(E1 区和 E2 区)对 15 株不同国家地区来源的序列分型是正确的，并最终确定最能反映丙型肝炎病毒株演化关系的区域是 NS5B 区，为丙型肝炎病毒基因分型确定“金标准”奠定理论基础. 同时分析 core 区的分子演化，发现 1a 型和 6a 型突变位点较少而 3b 型突变位点较多，进而分析 Singleton site，为发明针对 core 区的新的 PCR-RFLP 基因分型方法提供新的思路.

参 考 文 献

1 Choo Q L, Kuo G, Weiner A J, *et al*. Isolation of cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. Science, 1989, **244**(4902): 359～362

2 Cohen J. The science challenge of hepatitis C. Science, 1999, **285** (5424): 133～159

3 Otsuka M, Kato N, Omata M. Recent progress and prospective view of chronic hepatitis C research. Nippon Rinsho, 2001, **59** (7): 1243～1247

4 Smith D B, McAllister J, Casino C, *et al*. Molecular epidemiology of hepatitis C virus. J Gastroenterol Hepatol, 1997, **12**(7): 522～527

5 Mellor J, Holmes E C, Jarvis L M, *et al*. Investigation of the pattern of hepatitis C virus sequence diversity in different geographical regions: implications for virus classification. J Gen Virol, 1995, **76**: 2493～2507

6 谢　尧, 徐道振, 陆志檬, 等. HCV 基因型对慢性丙型肝炎干扰素疗效的影响. 中华肝脏病杂志, 2004, **12**(2): 72～75
Xie X, Xu D Z, Lu Z M, *et al*. Chin J Hepatology, 2004, **12**(2): 72～75

7 Cyrille F, Michelle G, Didier S, *et al*. Influence of the genotypes of hepatitis C virus on the severity of recurrent liver disease after liver transplantation. Gastroenterology, 1995, **108**(4): 1088～1096

8 Tamura K, Dudley J, Nei M, *et al*. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution, 2007, **24**(8): 1596～1599

9 Thompson J, Gibson T, Plewniak F, *et al*. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research, 1997, **25** (24): 4876～4882

10 Pawlotsky. Molecular diagnosis of viral hepatitis. Gastroenterology, 2002, **122**(6): 1554～1568

11 邱国华, 张　瑞, 杜绍财, 等.中国丙型肝炎病毒基因分型的进化树分析. 临床检验杂志, 2005, **23**(3): 165～167
Qiu G H, Zhang R, Du S C, *et al*. Chin J Clin Laboratory Science, 2005, **23**(3): 165～167

12 Donald B S, Janet M, Lisa M J, *et al*. Variation of the hepatitis C virus 5′ non-coding region: implications for secondary structure, virus detection and typing. J Gen Virol, 1995, **76**(7): 1749～1761

13 Donald G, Bernard W, Marc D, *et al*. Use of sequence analysis of the NS5B region for routine genotyping of hepatitis C virus with reference to C/E1 and 5′ untranslated region sequences. J Clin Microbiology, 2007, **45**(4): 1102～1112

14 赵春江, 李　宁, 邓学梅. 应用创造酶切位点法检测单碱基突变. 遗传, 2003, **25**(3): 327～329
Zhao C J, Li N, Deng X M. Hereditas, 2003, **25**(3): 327～329

Genotyping of Hepatitis C Virus and Analysis of The Molecular Evolution Based on Core Region Sequence

WAN Xiang-Hui, ZENG Zhao-Fang*, YANG Xi-Mei

(*Key Laboratory of Laboratory Medical Diagnostics, Ministry of Education, Department of Laboratory Medicine, Chongqing Medical University, Chongqing* 400016, China)

Abstract At present, there are many methods for genotype of hepatitis C virus , but not a gold standard. In order to establish the rationale for genotypic determination of optimal region sequence, fifteen complete genome sequences of hepatitis C which had been given the annotation about every region and derived from different country were downloaded from GenBank. Phylogenetic trees on 5′ UTR, core, E1, E2 and NS5B region were established. The results demonstrated that genotyping group was not all correct on 5′ UTR region while genotyping groups were wholly correct on core, E1, E2 and NS5B region. Comparing phylogenetic distances on core, E1, E2 and NS5B region with that on complete genome sequence demonstrated that the NS5B area was the best genotyping region instead of the complete genome sequence. In addition, analysis of the molecular evolution on each core region could supply some clues for creating novel genotyping method based on PCR-RFLP.

Key words hepatitis C virus，genotype，molecular evolution，core protein

*Corresponding author.

Tel: 86-23-68485095, E-mail: zeng000@126.com

Received: December 11, 2007 Accepted: February 11, 2008