doi:10.3969/j.issn.1672-5565.2014.02.12

基于 ChIP-seq 的差异组蛋白修饰区域的筛选

祝让飞,刘洪波,苏建忠,王 芳,崔 颖,张 岩*
(哈尔滨医科大学生物信息科学与技术学院,哈尔滨 150081)

摘 要:组蛋白修饰是在基因组水平上起到重要调控作用的表观遗传修饰,随着 ChIP-Seq 的广泛使用,高通量数据的积累,为从全基因组研究组蛋白修饰模式奠定了基础。但目前缺乏在多样本间筛选疾病相关的调控区域的方法,因而本文开发了一种多细胞系的差异筛选算法来识别差异组蛋白修饰区域。本文通过窗口移动法来估计组蛋白修饰水平,并根据信息熵理论定量各个细胞系之间的差异。基于随机背景来确定差异显著性阈值。利用此算法来筛选人类全基因组 9 个细胞系间 H3K4me3 差异的区域,结果显示这些区域显著富集在基因启动子上和其他重要的染色质状态上,且与先前人们发现的活性启动子染色质状态显著重叠。通过文献挖掘进一步证实了与白血病相关的基因组标记。这些结果表明基于熵的策略可有效地挖掘多细胞系间以及与疾病相关的差异组蛋白修饰。

关键词:组蛋白修饰差异;高通量数据处理;表观遗传修饰
中图分类号:Q81 **文献标志码**:A **文章编号**:1672-5565(2014)-02-151-06

Identification of regions differentially modified by histone modification based on the ChIP-seq data

ZHU Rangfei, LIU Hongbo, SU Jianzhong, WANG Fang, CUI Ying, ZHANG Yan*
(*School of biological information and technology*, *Harbin medical university*, *Harbin* 150081, *China*)

Abstract: Histone modification plays an important regulating role in genome. With the wide use of ChIP-Seq, high-throughput data has been accumulating from whole genome researches related with epigenetic regulation of genes. However, the lack of effective methods for processing and analyzing of these data hinders the screening of disease-related regulatory regions. In this study, we developed a novel algorithm for identification of the regions differentially modified by histone modification across multiple cell lines. By this method, we estimated the level of histone modification in each cell line, and quantified the histone modification difference among multiple cell lines according to the information entropy theory. The statistically significant threshold based on the random background can be used to identify the regions differentially modified by histone modification across multiple cell lines. We applied the algorithm to genome-wide screening of the regions differentially modified by H3K4me3 across nine cell lines. We found a significant enrichment of these regions on gene promoters and other important chromatin states. It is also revealed that these regions significantly overlapped with chromatin status related with the active promoter. Further literature mining confirmed the specific high H3K4me3 of gene RHCE in K265 cell line. These results show that our proposed strategy based entropy is effective in identification of histone modification difference among multiple cell lines and mining epigenetic abnormalities in diseases.
Keywords: Histone modification, High-throughput data processing, epigenetic modification.

表观遗传学是研究表观遗传变异的遗传学分支学科,表观遗传变异是指在基因的 DNA 序列没有发生改变的情况下,基因功能发生了可遗传的变化,并最终导致了表型的变化,这在一定程度上并不符合

收稿日期:2013-11-01;**修回日期**:2014-01-06.
基助项目:黑龙江省大学生创新创业训练计划(201210226010)资助。
作者简介:祝让飞,男,本科,研究方向:生物信息学;E-mail:zhurangfei@gmail.com.
*通信作者:张岩,女,博士,教授,博士生导师;研究方向:生物信息学;E-mail:yanyou1225@gmail.com.

孟德尔遗传规律。表观遗传性状是一种来自于DNA序列变异的染色体改变所带来的稳定的可遗传的表型[1]。目前,最为广泛研究的表观遗传修饰主要包括DNA甲基化和组蛋白修饰。

在真核生物中,染色质被连续的组蛋白八聚体所包装,环绕其上的是147 bp的DNA序列。组蛋白是细胞核内的基本蛋白质,是一类小分子碱性蛋白质,有5种类型:H1、H2A、H2B、H3、H4。它们富含带正电荷的碱性氨基酸,能够同DNA中带负电荷的磷酸基团相互作用。其中H1不参加核小体的组建,其作用是连接核小体构成染色质,并带给染色质以极性,有一定的组织和种属特异性,在进化上不太保守。而H2A、H2B、H3、H4是核小体组蛋白,相对分子量较低,它们的作用是将DNA分子盘绕成核小体,没有组织和种属的特异性,在进化上也比较保守,特别是H3、H4是已知所有蛋白质中最为保守的。组蛋白修饰可以经共价修饰而发生乙酰化、甲基化和磷酸化,由此构成多种多样的组蛋白变体。

核小体中每个核蛋白都存在不同的变体,部分是有翻译后修饰导致的,如甲基化。特异的组蛋白变体沿着染色质分布,影响基因表达,且能被染色质免疫共沉淀技术检测。运用此技术,染色质先被切割成核小体大小的片段,然后借助抗体将核小体特异的亚型和蛋白变体分离,用此法得到150~200 bp的DNA片段能通过探针点样杂交到微阵列(ChIP-chip)或是直接检测(ChIP-Seq),并被映射到基因组上。对照ChIP-chip数据,ChIP-Seq数据能提供更高的分辨率以及更广泛的蛋白质-DNA互作检测的基因组覆盖度。

与ChIP-chip相比,ChIP-Seq有所需原材料少,成本低及分辨率高等特点。并且在数据分析方面也有优势。首先,ChIP-Seq读取片段不仅仅能够测定已知的感兴趣的基因组区域,还可以用于检测未知的有潜在功能的基因组区域,如测定蛋白质-DNA结合位点等。此外,基于测序技术的ChIP-Seq促进了人们开发更多的信息学方法挖掘基因组中重要的调控元件。

随着ChIP-chip和ChIP-Seq技术的出现,人们可以在全基因组水平上分析染色质修饰结合位点,不同的细胞系各类修饰的日益增多,为人们解码染色质修饰模式提供了丰富的参考资料,过去已有一些方法分析表观遗传数据,确定功能描述组合模式和系统地定义DNA调控元件。例如组蛋白H3上的第4个和第27个赖氨酸的甲基化是胚胎干细胞的一个重要的表观遗传修饰,对基因的激活和沉默有着重要的作用[2]。还有基于有监督的回归分析框架,分析解释启动子区域强的组蛋白修饰及预测基因的表达,科研人员在最近的研究中利用隐式马尔科夫模型和贝叶斯网络挖掘组蛋白标记组合模式并发现了周期性的染色质状态[3]。然而,用于分析多细胞系间差异组蛋白修饰的方法的困乏直接限制了人们识别特定生命状态(如癌症)特异的表观遗传标记的进程。

本课题利用高通量的ChIP-Seq数据识别不同生命现象和过程的组蛋白修饰差异和功能模式,挖掘癌症特定基因或基因集合的表观遗传修饰的发生频率、稳定性和差异,用于疾病的早期诊断和预后;整合高通量实验技术测定的癌症中高通量的组蛋白修饰图谱数据,充分利用生物信息学优势,构建了合理的数学模型,开发新方法来挖掘癌症表观遗传特征。为癌症等复杂疾病的发病机制,分子诊断和靶点治疗提供新的思路。

1 材料与方法

1.1 材料

本文利用了Ernst等人的实验数据进行研究[3],该数据包括九个细胞系的H3K4me3修饰的全基因组染色质修饰谱数据。这9个细胞系包括:胚胎干细胞(H1 ES or H1),白血病细胞(K562),B淋巴细胞(GM12878),肝癌细胞(HepG2),脐带血管内皮细胞(HUVEC),骨骼肌成肌细胞(HSMM),正常人肺成纤维细胞(NHLF),正常表皮角质细胞(NHEK)和乳腺上皮细胞(HMEC)。

1.2 方法

1.2.1 组蛋白修饰定量

本文采用的是窗口移动法来定量组蛋白修饰强度。将全基因组划分为不重复的定长的窗口,为了捕获实际的位点,将所有读数片段向3'端方向扩展到300 bp,再将这些读数片段往基因组上覆盖,并计算各窗口中被读数片段覆盖的指标,约定被读取片段前200 bp覆盖到的权重为1,被其后的100 bp片段覆盖到的权重为0.25。这样,每个样本在基因组上的组蛋白修饰水平即为各窗口的覆盖的Reads数目与权重的乘积和,根据此方法本研究对所有样本在全基因组的组蛋白修饰进行定量。

1.2.2 组蛋白修饰差异的定量

在信息论中,熵被用来衡量一个随机变量出现的期望值。这一期望值代表了信号在被接收之前的传输过程中损失的信息量,又被称为信息熵。基于香农信息熵[4]开发了一种新的有效的多样本组蛋白修饰差异筛选的算法。在过去的研究中,人们已

经利用香农信息熵开发了一些用于衡量基因表达组织特异性、甲基化特异性的方法，香农信息熵在对筛选具有组织特异性的效能已经得到证实[5-6]。差异熵的计算方式是将基因在 n 个样本中的表达量视为一个 n 维的向量 $X_i=(x_{i1},x_{i2},\ldots\ldots,x_{i(n-1)},x_{in})$，用 p_t 表示基因 i 在组织 t 中的相对表达量，则计算公式为 $p_{it}=x_{it}/\sum_{j=1}^{n}x_{ij}$，对应基因 i 的差异熵的计算公式是 $H_i=E(\log_2{}^{pi})$，H 的单位是 bits，值域为 $(0\sim\log_2{}^{N})$，代表组织差异性的高低，0 表示该基因只在一个组织中特异表达，$\log_2{}^{N}$ 则表示基因在所有组织中具有完全相同的表达量。这种方法是建立在固定的某些区域内的，对于 ChIP-Seq 数据，将每个窗口作为相应的区域，每个窗口上的修饰指标作为修饰值，通过上述方法来定量每个窗口的差异。

1.2.3　组蛋白修饰差异的显著性阈值

基于量化的组蛋白修饰差异，本研究进一步确定了一个阈值来筛具有选显著差异的组蛋白修饰。利用随机的思想模拟测序生成随机数据，构建随机测序模型，该模型中各样本的修饰是均匀分布的，依此差异定量的公式计算出的信息熵服从正态分布。我们的零假设是不同样本间在窗口中修饰是相同的，那么在一定的显著性水平下，可以确定窗口的显著性阈值。

1.2.4　组蛋白修饰差异的筛选

使用此算法筛选以上 9 个细胞系在一号染色体上的 H3K4me3 修饰的差异。显著性水平为0.001，选择的窗口长度为 25 bp，通过筛选得到了213 506 个具有显著性差异的窗口。考虑到核小体的长度约为 147 bp，所以将差异的窗口间距离低于为 147 bp 的区域进行合并成一个区域，这样得到了15 033个区域。通常情况下低于 147 bp 的区域是难以单独行使功能的，所以去除了低于 147 bp 的区域以减少噪音，这样总共得到了 9 828 个差异的区域。

1.2.5　方法评估

将差异区域映射到已知的基因组特征上，探讨它们潜在的功能角色。将人类基因组所有碱基分类为七类：up2kb、5’ UTR、exon、intron、3’ UTR、down2kb 和 intergenic。具体来说，在 refSeq 基因的转录起始位点 2kb 以内的碱基标记为 up2kb，处于距离转录终止位点 2kb 以内的下游碱基标记为 down2kb，2kb 以外的碱基标记为 intergenic。差异区域依次被分类到这七种特殊的类别中。如果这些差异区域中有属于多个类别的，将按 up2kb>5’ UTR>3’ UTR>exon>intron>down2kb>intergenic 的级别进行处理。对于每个类型的基因组区域，差异区域相应的比例除以全基因组比例来确定富集比例。

将这些差异区域映射到各细胞系 Ernst 等人研究的 15 种染色质状态中[3]，检查这些差异区域的潜在的功能。凡是与状态区域有交集的即把差异区域分类到该状态类别下，差异区域依次被分类到这 15 个状态中，如果这些差异区域属于多个类别的，则这些状态类别分别计数加一。对于每个细胞系的每个染色质状态区域，差异区域相应的比例减去该状态的比例再除以该状态的比例来确定富集比例。通过差异区域找到了白血病特异的组蛋白修饰区域，并对这些区域与白血病的关系进行了文献挖掘证实。

2　结果与讨论

2.1　结果

2.1.1　组蛋白修饰定量结果

为确定组蛋白修饰定量的可靠性，将定量后的修饰上传到 UCSC 与先前的研究作对比（见图 1）。由图中两种方法定量的结果对比可以看出无明显的差异，由此确定定量方式是有效的。

2.1.2　组蛋白修饰差异的显著性阈值

通过 10 次随机模拟和 10 个不同测序深度的随机模拟及 20 个不同读数片段的扩展长度的随机模拟的结果表明随机的背景是稳定。最终计算得到了 9 个样本在 0.001 的显著性水平下的阈值。

2.1.3　9 个细胞系 H3K4me3 的差异结果

通过筛选最终得到了 9 828 个差异的区域在人类一号染色体上，显著性水平为 0.001。通过映射已知的基因组特征上，共有 6 296 个区域富集到 refGene（一号染色体共 4 412 个基因）上，有7 080 个区域富集到了 refGene 基因的 2kb 范围内如图 2（A）所示。从图中可以看出，大部分差异的区域落在了基因相关的区域上，这说明筛出的差异区域是有功能的，有 64.06%以上是与基因相关的。

差异区域映射到已知基因组特征的结果见表 1。通过表 1，我们可以看出差异区域大部分富集在 up2kb，5’ UTR，intron 和 intergenic 上，但是 intron 和 intergenic 是全基因组占比例最大的区域，而 up2kb 和 5’ UTR 是在启动子区域范围内的，在全基因组所占比例较小，而富集的比例却很大，这说明这些差异区域显著富集在基因启动子区域。并已有很多研究表明组蛋白修饰和基因表达密切相关[7-9]。

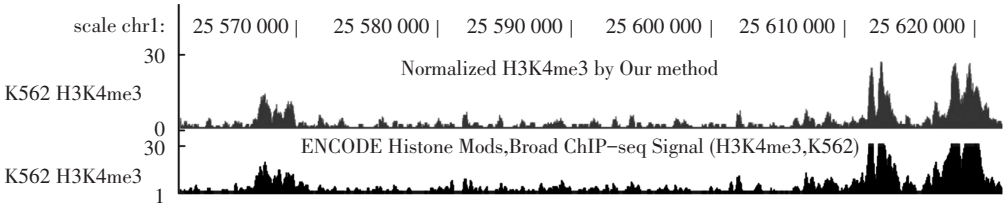

图 1 校正后的组蛋白修饰水平

Fig.1 The normalized histone modification level

表 1 差异 H3K4me3 修饰区域的基因组分布

Table 1 The distribution of regions differentially modified by H3K4me3

基因组特征	up2kb	5'UTR	3'UTR	exon	intron	down2k	intergenic
个数	1 502	1 508	373	449	3 010	238	2 748
比例	15.28%	15.34%	3.80%	4.57%	30.63%	2.42%	27.96%

2.1.4 映射到各细胞系染色质状态的差异区域的特征

将这些差异区域映射到各细胞系 Ernst 等人研究的 15 种染色质状态中[3],检查这些差异区域的潜在的功能。最终得到了差异区域在各个细胞系中 15 种状态的富集情况(见表 2)。

表 2 中体现了 9 个细胞系差异区域在先前研究的 15 种染色质状态下的富集个数,从中可以看出 Promoter 状态下的富集程度较高,这与 H3K4me3 作用域相一致。

表 2 差异 H3K4me3 修饰区域的与染色质状态

Table 2 The regions differentially modified by H3K4me3 and chromatin states modified by H3K4me3

染色质状态 / 细胞系	Gm12878	H1hesc	Hepg2	Hmec	Hsmm	Huvec	K562	Nhek	Nhlf
1 Active Promoter	1 891	1 230	1 973	1 526	1 496	1 302	1 685	1 686	1 585
2 Weak Promoter	2 176	1 953	2 316	1 560	1 723	1 233	1 363	1 269	1 585
3 Poised Promoter	494	902	248	290	368	490	299	471	392
4 Strong Enhancer	1 979	235	914	1 158	880	1 306	2 717	1 311	674
5 Strong Enhancer	524	229	229	840	639	654	705	662	795
6 Weak Enhancer	1 348	1 707	1 490	1 460	1 486	1 233	871	1 363	1 439
7 Weak Enhancer	620	1 178	535	1 247	887	1 032	1 016	987	1 015
8 Insulator	167	257	135	159	203	225	312	263	188
9 Txn Transition	216	510	495	288	613	443	765	435	339
10 Txn Elongation	278	303	414	394	692	319	447	503	454
11 Weak Txn	1 071	2 611	1 600	2 028	2 062	1 881	1 112	1 967	1 910
12 Repressed	891	413	454	463	749	1 170	1 277	840	1 190
13 Heterochromatin; low signal	2 250	2 857	3 007	2 889	2 508	2 823	1 458	2 596	2 547
14 Repetitive/CNV	513	177	203	187	230	204	263	168	205
15 Repetitive/CNV	174	177	110	121	166	192	192	87	375

对于每个细胞系的每个染色质状态区域,差异区域相应的比例减去该状态的比例再除以该状态的比例得到差异区域在个细胞系各状态下的富集比例如图 2(B)。

由图 2(B)中可以看出差异区域在各个细胞系中染色质状态 1 Active Promoter 的富集比例最为显著,这与先前人们发现的活性启动子染色质状态显著重叠,说明本研究的方法找出的差异区域是有生物学功能的。

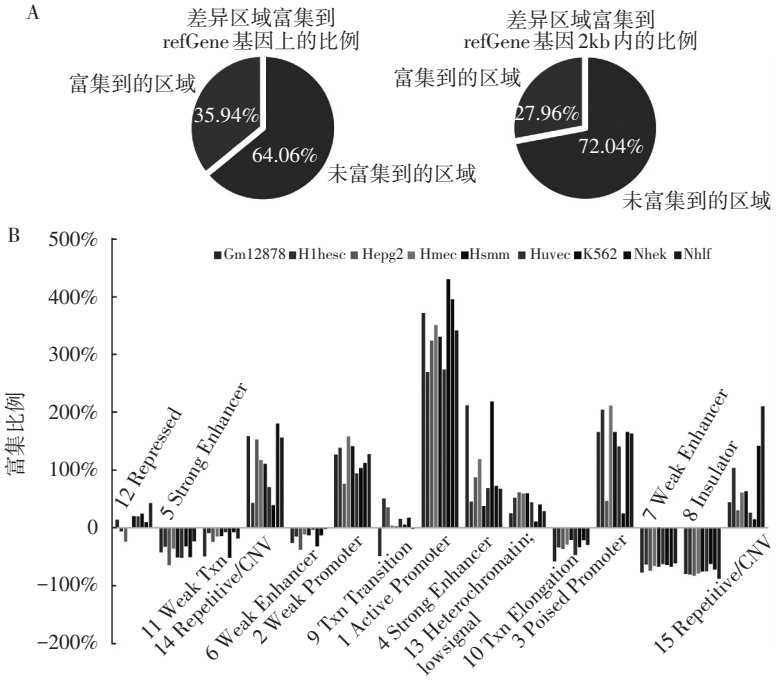

图 2　差异组蛋白修饰区域在基因组的分布

Fig.2　The distribution of regions differentially modified by H3K4me3

注：(A)为差异 H3K4me3 修饰区域的在基因附近的分布；(B)为差异 H3K4me3 修饰区域在染色质状态中的分布。

Notes：(A) The distribution of regions differentially modified by H3K4me3 nearby genes, (B) The distribution of the regions differentially modified by H3K4me3 on the chromatin states Indifferent cell lines.

2.1.4　文献挖掘验证与白血病和肝癌相关基因组标记

通过文献挖掘找到了被差异区域富集的基因 RHCE，从图 3 中可以看出 K562 上的组蛋白修饰水平明显高于其他各细胞系，在筛出的差异区域主要位于 RHCE 启动子位置附近，说明该差异区域是与 RHCE 相关的，且在 K562 细胞系中 H3K4me3 修饰特异高，在过去的研究中已被证实 H3K4me3 修饰影响基因的表达。而白血病的发生与 RHCE 的表达有关[10-12]，这与本研究的发现相一致。

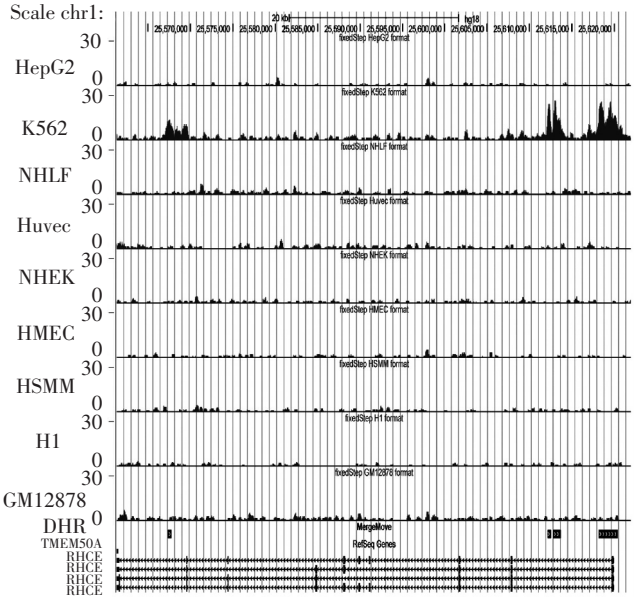

图 3　RCHE 基因区域的 K562 细胞特异的 H3K4me3 修饰模式

Fig.3　K562 cell line specific H3K4me3 modification in RCHE gene

注：图中每一行表示一个细胞系中 RCHE 基因上的 H3K4me3 修饰强度，其中 DHR 表示筛出的差异 H3K4me3 修饰区域。

Notes：For each line, the H3K4me3 pattern of gene RCHE in a cell line is shown. DHR represents the regions differentially modified by H3K4me3.

2.2 讨论

本文开发了一个有效的、快速的筛选多样本的全基因组的组蛋白修饰差异的算法。基于 ChIP-Seq 技术是目前最流行的测定组蛋白修饰的方法。本文提出的基于信息熵的方法对多样本的差异的识别更有准确,效率更高。将该算法应用到人类 9 个细胞系的 H3K4me3 修饰中,分别识别出了这 9 个细胞系中的大量功能组蛋白修饰差异模式,通过分析发现,这些差异的组蛋白修饰区域倾向于定位在基因相关的功能区域,且可能参与癌症等复杂疾病的发生。虽然本文的方法能有效的识别出差异的组蛋白修饰区域,但是没有一个成型的工具,使用并不方便,后续我们将改善算法的效率,开发出方便实用的工具以便其更好的应用于疾病相关的差异组蛋白修饰区域识别,为揭示疾病发生中的表观遗传调控机制及挖掘疾病表观遗传标记提供便利。

参考文献(References)

[1] BIRD A. DNA methylation patterns and epigenetic memory [J]. Genes Dev, 2002, 16(1): 6-21.

[2] BARSKI A, CUDDAPAH S, CUI K, et al. High-resolution profiling of histone methylations in the human genome [J]. Cell, 2007, 129(4): 823-837.

[3] ERNST J, KHERADPOUR P, MIKKELSEN T S, et al. Mapping and analysis of chromatin state dynamics in nine human cell types [J]. Nature, 2011, 473(7345): 43-49.

[4] SHANNON C E. The mathematical theory of communication. 1963 [J]. MD computing : Computers in Medical Practice, 1997, 14(4): 306-317.

[5] SCHUG J, SCHULLER W P, KAPPEN C, et al. Promoter features related to tissue specificity as measured by Shannon entropy [J]. Genome Biology, 2005, 6(4): R33.

[6] ZHANG Yan, LIU Hongbo, LÜ Jie, et al. QDMR: a quantitative method for identification of differentially methylated regions by entropy [J]. Nucleic Acids Res, 2011, 39(9): e58.

[7] BELZIL V V, BAUER P O, PRUDENCIO M, et al. Reduced C9orf72 gene expression in c9FTD/ALS is caused by histone trimethylation, an epigenetic event detectable in blood [J]. Acta Neuropathologica, 2013, 126(6): 895-905.

[8] RADUWAN H, ISOLA A L, BELDEN W J. Methylation of histone H3 on lysine 4 by the lysine methyltransferase SET1 protein is needed for normal clock gene expression [J]. The Journal of Biological Chemistry, 2013, 288(12): 8380-8390.

[9] TAN Dunyong, TAN Si, ZHANG Jie, et al. Histone trimethylation of the p53 gene by expression of a constitutively active prolactin receptor in prostate cancer cells [J]. The Chinese Journal of Physiology, 2013, 56(5):1-9.

[10] SMYTHE J S, AVENT N D, JUDSON P A, et al. Expression of RHD and RHCE gene products using retroviral transduction of K562 cells establishes the molecular basis of Rh blood group antigens [J]. Blood, 1996, 87(7): 2968-2973.

[11] SMYTHE J S, ANSTEE D J. Expression of C antigen in transduced K562 cells [J]. Transfusion, 2001, 41(1): 24-30.

[12] 严力行,许先国,朱发明,等.RHD 基因的克隆及其在 k562 细胞中的表达[J].中国实验血液学杂志,2005,13(3):492-495.
YAN Lixing, XU Xianguo, ZHU Faming, et al. Cloning of human RHD gene and its expression in K562 cells [J]. Journal of Experimental Hematology, 2005, 13(3): 492-495.