

人类基因组核苷酸多态性位点核小体定位分析*

刘宏德 孙 啸**

(东南大学生物电子学国家重点实验室, 南京 210096)

摘要 研究人类基因组核苷酸多态性位点周围核小体的定位, 对于分析核苷酸的变异机制有重要意义. 分析了人类基因组单核苷酸多态性(SNP)位点、简单插入位点、插入删除位点和删除位点的分布规律, 以及这些位点周围的核小体定位特征. 结果表明: 转录起始位点下游的核苷酸多态性位点分布呈现约 211 bp 的周期特征, 单核苷多态位点另有一个 146 bp 的周期; 约 211 bp 的周期与转录起始位点下游核小体的分布周期 204 bp 非常接近, 146 bp 的周期恰是核小体核心 DNA 的长度. 这些结果说明核小体与多态性位点的分布关系密切. 进一步研究证实, 单核苷酸多态性位点多分布于核心 DNA 上, 且多位于核心 DNA 的两端, 这使得单核苷酸多态性位点具有 146 bp 周期, 而插入、插入删除、删除多态性位点多分布于核小体排开区域, 间隔约为 204 bp. 转录起始位点下游核小体等间隔的规则分布使得多态性位点的分布也具有周期性. 研究表明, 相对于核小体, 不同类型变异发生的位置不同, 核小体定位在基因组多态性位点的形成过程中具有重要作用.

关键词 核苷酸多态性, 核小体定位, 周期

学科分类号 Q6

DOI: 10.3724/SP.J.1206.2010.00204

遗传变异影响个体性状特征、疾病发生风险及个体对环境的反应等, 是遗传学研究的主要内容^[1-2]. 在 DNA 层次上, 变异表现为核苷酸的多态性, 根据变异数目, 可将变异分为单核苷酸多态性(SNP)和结构变异(structure variation)(如拷贝数的变化, CNV)^[3-4]. SNP 是指 DNA 序列上发生的单个核苷酸碱基的变异, 已经发现很多基因的 SNP 位点与疾病有关联^[5]. 目前, 在 dbSNP 数据库中收集的简单变异位点超过 2 300 万个^[6]; 其中简单变异包括 SNP、小片段插入(insertion)、删除(deletion)和插入删除(in-del)等^[7]. 研究变异发生的机制对于分析多态性与疾病之间的关系及进化机制等具有重要意义.

核小体是真核生物基因组的基本构建单元, 75%~90%的真核基因组 DNA 包裹在组蛋白八联体上形成核小体^[8]. 核小体定位是指 DNA 双螺旋相对于组蛋白核的位置. 这种定位作用封闭了位于核小体上的蛋白质结合位点^[7-8], 进而阻止蛋白质与 DNA 的结合, 以此达到调节基因转录的作用.

近期, Higasa 等^[9]发现, 在转录起始位点(TSS)下游(3'方向), SNP 位点的分布呈现周期性, 周期为 146 bp. Sasaki 等^[10]通过分析青鳉(*Oryzias latipes*)遗

传变异位点的分布发现, 在 TSS 下游遗传变异存在一个约 200 bp 的周期, 该周期与染色体的结构和核小体的定位有密切的联系.

本文分析了人类基因组 TSS 下游多态性位点分布的周期特征, 并研究了多态性位点周围核小体定位的规律.

1 数据和方法

1.1 数据

人类全基因组多态性位点和 TSS 数据来自 UCSC (<http://genome.ucsc.edu/cgi-bin/hgTables>). 本文涉及的多态性位点是经过提交者验证的单核苷多态位点(single)、插入(insertion)、删除(deletion)和插入删除(in-del)位点. 利用 UCSC, 从人类单核苷酸多态性数据库 dbSNP (SNP130)中获取多态性位点数据. 转录起始位点数据取自 switchDbTss, 该数

* 国家自然科学基金(30800209, 61073141)和江苏省自然科学基金(BK2010406)资助项目.

** 通讯联系人.

Tel: 025-83795174, E-mail: xsun@seu.edu.cn

收稿日期: 2010-09-22, 接受日期: 2011-03-15

数据库中的转录起始位点是利用实验数据结合经验得分函数确定的^[11]。实验检测激活的 CD4⁺ T 细胞核小体位置数据取自文献[12]。

1.2 方法

1.2.1 转录起始位点下游多态性位点及核小体分布的周期特征分析。

已有的研究表明, TSS 下游变异位点的分布呈现 146 bp 和 200 bp 的周期^[9-10], 为了进一步在人类基因组中验证这一结论, 本文利用快速傅立叶变换 (FFT) 分析多态性位点分布的周期特征。对每种多态性类型(single, insertion, deletion 和 in-del), 首先将多态性位点以 TSS 为对齐点, 将 TSS 上下游各 3 kb 内的多态性位点数进行加和, 加和结果进行标准化处理(每个数据减去平均值, 然后除以标准方差), 得到多态性位点在 TSS 周围的分布曲线。然后对 TSS 下游(0~3k bp)多态性位点的分布信号进行快速傅立叶变换。

为了考察多态性位点分布与核小体分布的关系, 利用实验检测的核小体定位数据, 计算了 TSS 上游 1 kb 和下游 3 kb 范围内核小体的分布, 并对

分布信号进行了频谱分析。TSS 周围核小体分布的计算为: 以 TSS 为对齐点, 加和核小体定位的信号, 最后数据进行标准化处理(每个数据减去平均值, 然后除以标准方差)。实验检测数据为 CD4⁺ T 细胞在激活状态下核小体定位信号([http:// dir. nhlbi. nih.gov/papers/lmi/epige-nomes/Hgtcellnucleosomes.aspx](http://dir.nhlbi.nih.gov/papers/lmi/epige-nomes/Hgtcellnucleosomes.aspx))^[12]。

1.2.2 多态性位点周围核小体定位分析。

多态性位点周围核小体定位信号的计算过程为: 以多态性位点为参考点, 取其上下游各 1 kb 的序列(每条序列的长度为 2 kb), 计算每条序列的核小体定位信号(实验检测的核小体定位信号^[12]); 加和所有序列的核小体定位信号, 加和所得的信号经过标准化后即多态性位点周围核小体的分布信号。

2 结果与讨论

2.1 多态性位点在 TSS 周围的分布

图 1 显示了人类基因组 24 条染色体上 TSS 周围多态性位点的分布情况及 TSS 下游多态性位点

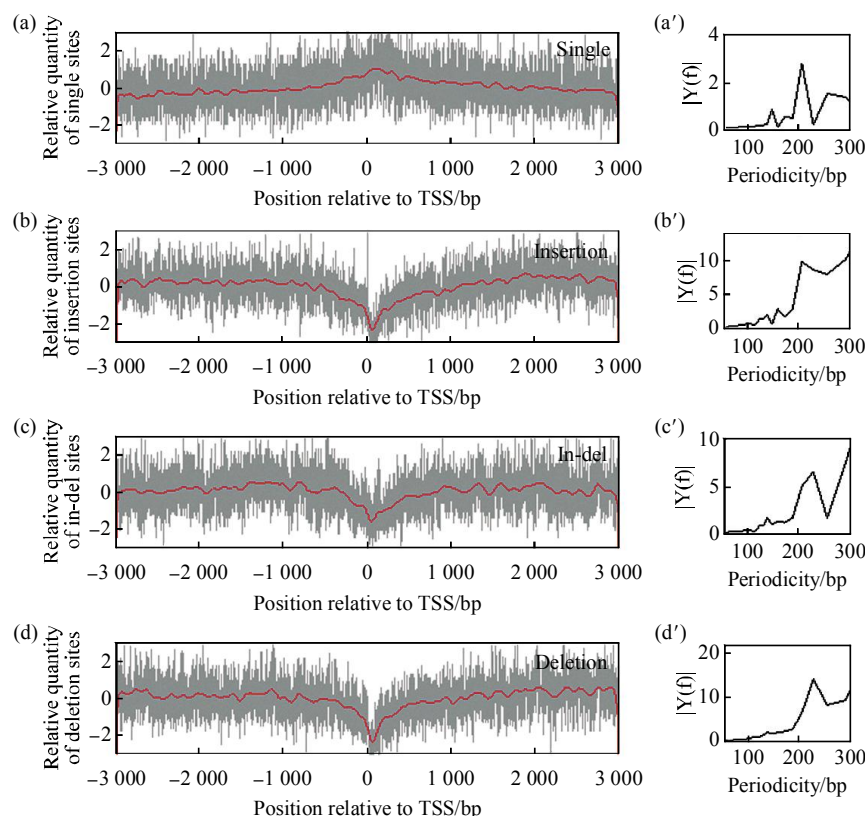


Fig. 1 The distributions of human nucleotide polymorphism sites near transcription start site (a, b, c and d) and the frequency spectrum of the distributions downstream of TSS (a', b', c' and d')

(a) Single nucleotide polymorphism (single). (b) Insertion. (c) Insertion-deletion(in-del). (d) Deletion. (a'), (b'), (c') and (d') Indicating the frequency spectrum of (a), (b), (c) and (d) in range of 0~3k bp, respectively.

分布的频谱分析. 图 1a、1b、1c 和 1d 分别为 TSS 周围单核苷酸多态位点、插入位点、插入删除位点和删除位点的分布. 从图 1 中可见, 多态性位点的分布在 TSS 周围(尤其在下游)呈现有规律的震荡, 即具有周期特征, 特别是从 500 到 2 500 bp 的区域中周期性更加明显. 图 1a~1d 显示的另外一个有趣的结果是, 在 TSS 下游 200 bp 范围内, 单核苷酸多态位点分布较多, 而插入、插入删除和删除位点分布较少. TSS 下游 200 bp 左右, 有一个稳定的核小体^[7], 因此上述结果说明不同类型的多态性位点对核小体有不同偏好性(见 2.3).

为分析多态性位点分布的周期特性, 对图 1a、b、c 和 d 的信号进行了频谱分析, 结果分别示于图 1a'、1b'、1c'和 1d'. 单核苷酸多态位点和插入位点的分布周期为约 204 bp(见图 1a'和 1b'); 插入删除位点和删除位点的分布周期分别约为 210 bp 和 227 bp(见图 1c'和 1d'); 四者的平均周期约为 211 bp. 此外, 对于单核苷酸多态位点(图 1a'), 另有一个 146 bp 的峰可见, 其他类型的多态位点, 此峰不明显. 这些结果表明: 多态性位点在 TSS 下游呈周期分布, 周期约为 211 bp. 这些结果与文献报道的结论基本一致^[9-10], Higasa 等^[9]的研究表明, 有 CpG 岛的人类启动子, 单核苷酸多态性位点在 TSS 周围的分布呈 146 bp, 且单核苷和二核苷在此区域没有 146 bp 的周期, 推测该周期是核小体在 TSS 周围定相的结果. Shin 等发现青鳉 (*Oryzias latipes*)遗传变异位点的分布有 200 bp 的周

期, 该周期与核小体分布的周期有关^[10].

2.2 转录起始位点下游核小体的分布

为进一步确认核小体和多态性位点在 TSS 下游的分布关系, 计算了 TSS 下游核小体的分布曲线及其周期特征(图 2). 在图 2a 中, 核小体在 TSS 下游的分布呈现振荡现象, 尤其在下游 400 bp 到 2 500 bp 的范围内. 在 TSS 上游 200 bp 范围内, 可以观察到一个明显的凹陷信号, 指示此处为核小体缺失区域(NFR), 且 NFR 两侧稳定的核小体也明显可见. 同时也注意到此 NFR 信号不是特别强烈(谷不是很深), 我们推测有两个方面的原因, 一个原因是图 2a 显示的结果为所有 TSS 周围的核小体分布特征, 即既包括活动的 TSS, 也包括沉默的 TSS. 因此, 有一部分 TSS 被核小体占据, 导致 TSS 上游 200 bp 范围内整体信号偏高. 另一个原因是本文所使用的 TSS 数据可能有小部分存在偏差, 即并非真实的 TSS^[11], 这可能导致 NFR 的特征被掩盖.

图 2b 显示了图 2a 的频谱分析结果, 在 204 bp 处, 有一明显的峰, 该峰表明核小体在 TSS 下游的分布具有 204 bp 的周期. 此前的研究表明, 核小体在 TSS 下游呈规则地等间隔的排布^[7], TSS 下游第一个核小体定位信号强, 随后定位强度(核小体的稳定性)依次降低, 这种核小体在 TSS 下游的规则分布是该区域核小体分布曲线具有周期特征的原因, 204 bp 对应核小体 DNA 序列的长度与连接 DNA 序列长度之和.

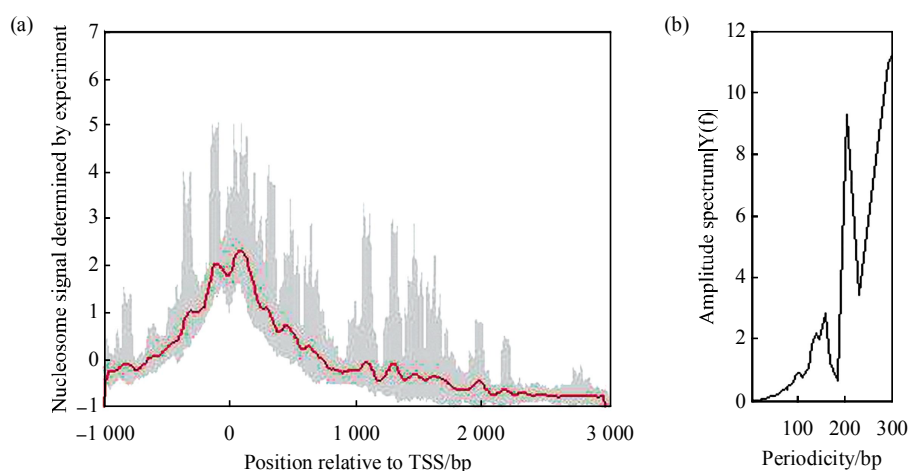


Fig. 2 The distribution of nucleosome in the vicinity of human transcription start site (TSS) (a) and the frequency spectrum of the distribution downstream of TSS (b)

In subplot (a), nucleosome signal is a standardized one; Gray line represents the experimental data; Red line is a 27-points smoothed result for the gray line. — : Raw signal; — : Smoothed signal.

2.3 多态性位点周围核小体的定位特征

从图 1 和图 2 的结果可见: TSS 下游, 多态性位点的分布周期(约 211 bp)和核小体分布的周期(204 bp)非常接近; 多态性位点分布的 146 bp 周期(图 1a')恰好是核小体核心 DNA 序列的长度. 这些结果和文献报道的结论强烈暗示多态性位点的分布与核小体的分布关系密切^[9-10], 因此有必要进一步分析多态性位点周围核小体的分布特征.

图 3 显示了人类基因组多态性位点周围核小体的分布, 核小体定位信号为实验检测的激活 CD4⁺ T 细胞的核小体定位数据^[12]. 结果非常有趣: 在单核苷酸多态性位点(single)附近, 核小体定位信号明显高(图 3a), 说明 single 类型的多态性位点更

多地分布于核小体 DNA 上. 另一个重要的细节是, 恰在 single 多态位点处, 核小体定位信号有一个小凹陷(见图 3a 的右上角放大图), 而定位信号的最高值却位于两侧距离单核苷酸多态位点 50 bp 的位置. 由此结果推测, 一些单核苷酸多态位点倾向于分布于核小体核心 DNA 序列的两端, 因此, 单核苷酸多态位点的分布就会有一个 146 bp 的周期(图 1a 和图 1a'), 恰和核心 DNA 序列的长度一致. 由于核小体在 TSS 下游的分布具有等间隔(约 204 bp)的特征, 因此倾向于分布于核小体上的单核苷多态位点也有相似的周期特征(约 204 bp, 图 1a').

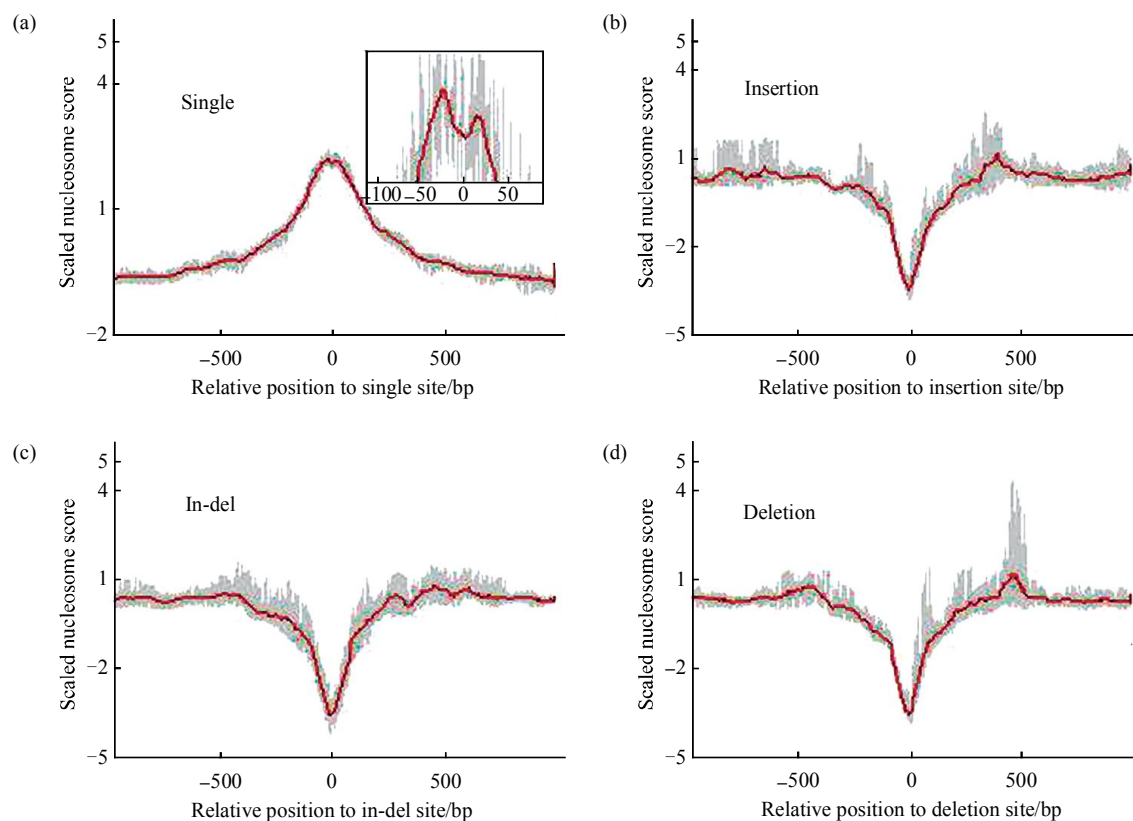


Fig. 3 The distributions of nucleosomes near nucleotide polymorphism sites in human genome

(a) Single nucleotide polymorphism (single). (b) Insertion. (c) Insertion-deletion(in-del). (d) Deletion. Gray line represents the experimental data in human activated CD4⁺ T cell; Red line is a 13-points smoothed result for the gray line. The nested plot in (a) shows a magnified profile for (a). —: Raw data; —: Smoothed data.

对于插入、删除和插入删除类型的多态性位点(图 3b、3c 和 3d), 核小体定位信号在这些类型的多态性位点周围明显凹陷, 凹陷区域的宽度约为 150 bp, 恰接近一个核小体 DNA 序列的长度. 这些结果表明: 在插入、插入删除和删除类型的多态

性位点周围, 核小体缺失, 或者说这些类型的多态性位点多分布于核小体缺失或者连接 DNA 的区域. 图 3 和图 1 的结果也是相互印证的. 据目前的研究, 在紧邻 TSS 的下游(约 200 bp), 有一个稳定定位的核小体(在图 2a 中也可见)^[6, 12-13]. 图 1 显示,

单核苷酸多态位点(图 1a)的分布在此区域较多, 而其他 3 种类型多态位点在此区域分布较少(图 1b~d), 这说明单核苷酸多态位点更倾向于分布于核小体上, 而插入、插入删除和删除类型的多态位点更偏好于核小体缺失或者连接 DNA 区。而图 3 的结果直接说明了 4 种类型的多态位点周围核小体的分布模式, 即单核苷酸多态位点在核小体上较多, 而其他 3 种类型的多态位点则倾向于分布于核小体排开区域。

为了更清晰地说明本文的结果及推论, 绘制了多态性位点与核小体分布的关系图(图 4)。虽然单核苷酸多态性位点和其他 3 种多态性位点的分布相对于核小体的位置不同, 但是明显地, 这些多态性位点的分布均有染色质结构(核小体分布)的印迹, 即

分布上的密切关联性; 核小体的分布在 TSS 的下游呈现等间距的规则分布^[7, 13], 是多态性位点分布呈周期分布的原因。转录耦联的 DNA 修复过程对被转录的区域有一定保护作用, 这个作用过程与染色质的可接近性有关^[10], 而且在化学诱导突变过程中^[9], 也涉及致突变物质与核小体的相互作用, 因此, 核小体在多态性形成过程中有重要作用。不同类型的多态性形成的机制存在差异, 因此发生的位点相对于核小体便会不同, 本文揭示单核苷酸多态性位点与另外 3 种多态性位点周围核小体定位模式的不同, 可能正是这种差异的反映。研究表明, 插入和删除均与点突变关联, 且插入和删除的发生存在相似的基因组因素^[14], 这也与本文发现的结果一致。

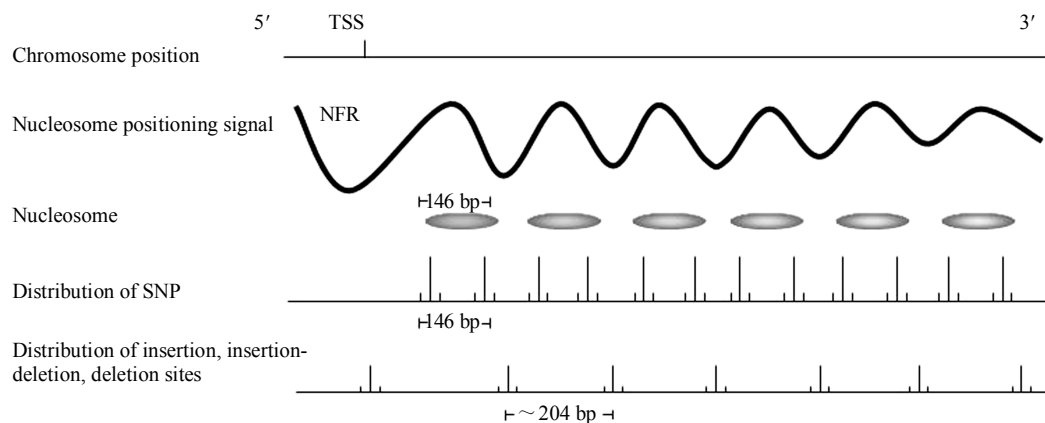


Fig. 4 Distribution mode of nucleosome and nucleotide polymorphism site

3 结 论

本文通过对人类基因组多态性位点分布规律的研究发现, 在转录起始位点下游的多态性位点分布呈现约 211 bp 的周期性特征, 单核苷酸多态位点还有一个 146 bp 的周期, 146 bp 是核小体核心 DNA 序列的长度。而且核小体的分布在转录起始位点下游也呈 204 bp 的周期, 这种周期上的吻合说明多态性位点的分布与核小体关系密切。进一步的分析发现单核苷酸多态性位点多分布于核心 DNA 的两侧, 而插入、删除、插入和删除多态性位点多分布于核小体排开区域。转录起始位点下游核小体等间隔的规则分布使得多态性位点的分布具有周期性。本文的研究表明核小体在基因组 DNA 的突变和修复过程中及多态性位点的形成过程中具有重要作用。

参 考 文 献

- [1] Brookes A J. The essence of SNPs. *Gene*, 1999, **234**(2): 177-186
- [2] Sherry S T, Ward M H, Kholodov M, *et al.* dbSNP: the NCBI database of genetic variation. *Nucl acids Res*, 2001, **29**(1): 308-311
- [3] <http://www.ncbi.nlm.nih.gov/projects/SNP/>
- [4] Mills R E, Luttig C T, Larkins C E, *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res*, 2006, **16**(9): 1182-1190
- [5] Pennisi E. Breakthrough of the year. Human genetic variation. *Science*, 2007, **318**(5858): 1842-1843
- [6] Lewin B. *Gene VIII*. Pearson Prentice Hall, Person Education, Inc., 2004: Chap. 20
- [7] Segal E, Mittendorf Y F, Chen L, *et al.* A genomic code for nucleosome positioning. *Nature*, 2006, **442**(17): 772-778
- [8] Henikoff S. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nat Rev Genet*, 2008, **9**(1): 15-26
- [9] Higasa K, Hayashi K. Periodicity of SNP distribution around transcription start sites. *BMC Genomics*, 2006, **7**: 66

- [10] Sasaki S, Mello C C, Shimada A, *et al.* Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. *Science*, 2009, **323**(5912): 401–404
- [11] Trinklein N D, Aldred S J, Saldanha A J, *et al.* Identification and functional analysis of human transcriptional promoters. *Genome Res*, 2003, **13**(2): 308–312
- [12] Schones D E, Cui K, Cuddapah S, *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 2008, **132**(5): 887–898
- [13] Jiang C Z, Pugh B F. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet*, 2009, **10**(3): 161–172
- [14] Kvikstad E M, Francesca C, Makova K D. Ride the wavelet: A multiscale analysis of genomic contexts, flanking small insertions and deletions. *Genome Res*, 2009, **19**(7): 1153–1164

Analysis of Nucleosome Positioning in The Vicinity of Sites of Nucleotide Polymorphism in Human Genome*

LIU Hong-De, SUN Xiao**

(State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China)

Abstract Studies of nucleosome positioning around the sites of nucleotide polymorphism are important for analysis of mechanism of genome variation. The distributions of sites single nucleotide polymorphism (SNP), simple insertion, insertion-deletion, and deletion are analyzed for human genome. Characteristics of nucleosomes in the vicinity of the polymorphism sites are also studied. The results indicated polymorphism sites downstream of transcription start sites (TSSs) occurs with an ~ 211 bp periodicities. For single nucleotide polymorphism, there is also a periodicity with 146 bp. The periodicity with ~ 211 bp is very close to the periodicity (204 bp) of nucleosome distribution downstream of TSS. The 146 bp periodicity is just the length of linker DNA sequence of a nucleosome. The results indicate that the distribution of polymorphism sites has an intimate relationship with nucleosome positioning. Further studies suggest that most of single nucleotide polymorphism sites are at two ends of core DNA, while sites of insertion, deletion, and insertion and deletion (in-del) tend to be in nucleosome-depleted region. The equally-spaced configuration of nucleosomes downstream of TSS causes the periodic distribution of polymorphism sites. The studies suggest genome variations occur in different regions relatively to nucleosomes, and nucleosome positioning has a role in forming nucleotide polymorphism.

Key words nucleotide polymorphism, nucleosome positioning, periodicity

DOI: 10.3724/SP.J.1206.2010.00204

*This work was supported by grants from The National Natural Science Foundation of China (30800209, 61073141) and The Jiangsu Natural Science Foundation (SBK201022749).

**Corresponding author.

Tel: 86-25-83795174, E-mail: xsun@seu.edu.cn

Received: September 22, 2010 Accepted: March 15, 2011