PIBB 生物化学与生物物理进展
*Progress in Biochemistry and Biophysics*
2013, 40(3): 216~227
www.pibb.ac.cn

# 染色质免疫沉淀-测序：全基因组范围研究蛋白质-DNA 相互作用的新技术 *

梁 芳[1)] 徐 柯[1)] 龚朝建[1)] 李 俏[1)] 马 健[1, 2)]
熊 炜[1, 2)] 曾朝阳[1)**] 李桂源[1, 2)**]

([1)] 中南大学肿瘤研究所，卫生部癌变原理重点实验室及教育部癌变与侵袭原理重点实验室，长沙 410078；
[2)] 中南大学疾病基因组研究中心湖南省非可控性炎症与肿瘤重点实验室，长沙 410013)

**摘要** 染色质免疫沉淀 - 测序(ChIP-seq)是近年来新兴的将染色质免疫沉淀(ChIP)与深度测序技术相结合，在全基因组范围内分析 DNA 结合蛋白结合位点、组蛋白修饰、核小体定位和 DNA 甲基化的高通量技术．在新一代测序(NGS)技术的大力推动下，ChIP-seq 提供了一种相对于 ChIP-chip 高分辨率、低噪音、高覆盖率的研究方法．随着测序成本的降低，ChIP-seq 逐步成为研究基因调控和表观遗传机制的一种常用手段．本文就该技术的最近研究进展进行综述，并着重介绍 ChIP-seq 数据分析过程及该技术的实际应用情况.

**关键词** ChIP-seq，新一代测序技术，基因调控，表观遗传学，数据分析
**学科分类号** Q78，Q81 **DOI**: 10.3724/SP.J.1206.2012.00305

人类基因组序列测定工作的完成标志着生物学研究已经进入后基因组研究时代，功能基因组学的研究逐渐成为研究的热点，而基因表达的调控又是功能基因组学的一个重要研究领域．从全基因组层面考察核转录因子 -DNA 相互作用和表观修饰是全面认识基因转录调控过程必不可少的环节.

新一代测序(next-generation sequencing，NGS)技术的迅猛发展将基因组学水平的研究带入了一个新的阶段，使得许多基于全基因组的研究成为可能，如全基因组测序(whole-genome sequencing)[1–3]、RNA-Seq(RNA sequencing)[4–5]．染色质免疫沉淀测序技术(chromatin immunoprecipitation followed by sequencing，ChIP-Seq)是近年来新兴的将染色质免疫沉淀技术(chromatin immunoprecipitation，ChIP)与新一代测序技术相结合，在全基因组范围内分析转录因子结合位点(transcription factor binding sites，TFBS)、组蛋白修饰(histone modification)、核小体定位(nucleosome positioning)和 DNA 甲基化(DNA methylation)的高通量方法[6–10]．相对于传统的基于芯片的 ChIP-chip (chromatin immunoprecipitation combined with DNA tiling arrays)，ChIP-seq 提供了一种高分辨率、低噪音、高覆盖率的研究蛋白质 - DNA 相互作用的手段[11]，可以应用到任何基因组序列已知的物种，可以研究任何一种 DNA 相关蛋白与其靶定 DNA 之间的相互作用，并能确切得到每一个片段的序列信息．随着测序成本的降低，ChIP-seq 逐步成为研究基因调控和表观遗传机制的一种常用手段．本文主要就这一新兴技术的基本原理和实验设计、优点、数据分析、实际应用等方面进行讨论.

* 国家自然科学基金资助项目(81272298, 30871282, 30871365, 81172189, 81171930)，湖南省自然科学基金(10JJ7003)，霍英东高校青年教师基金(121036)，中央高校基本科研业务费专项资金(2011JQ020)，中南大学米塔尔创新创业项目(11MX27)和中南大学贵重仪器设备开放共享基金资助项目.
** 通讯联系人. Tel: 0731-84805383
曾朝阳. E-mail: zengzhaoyang@xysm.net
李桂源. E-mail: ligy@xysm.net
收稿日期：2012-06-21，接受日期：2012-11-28

## 1　ChIP-seq 基本原理及实验设计

ChIP-seq 实验包括染色质免疫沉淀实验和高通量测序两个部分，其基本原理是：首先通过染色质免疫共沉淀技术特异性地富集目的蛋白结合的 DNA 片段；然后构建测序文库，采用新一代测序技术对富集得到的 DNA 片段进行高通量测序；最后通过将获得的数百万条序列标签精确定位到基因组上，从而获得全基因组范围内与组蛋白或转录因子等 DNA 结合蛋白相互作用的 DNA 区段信息(ChIP-seq 具体实验流程见图 1).

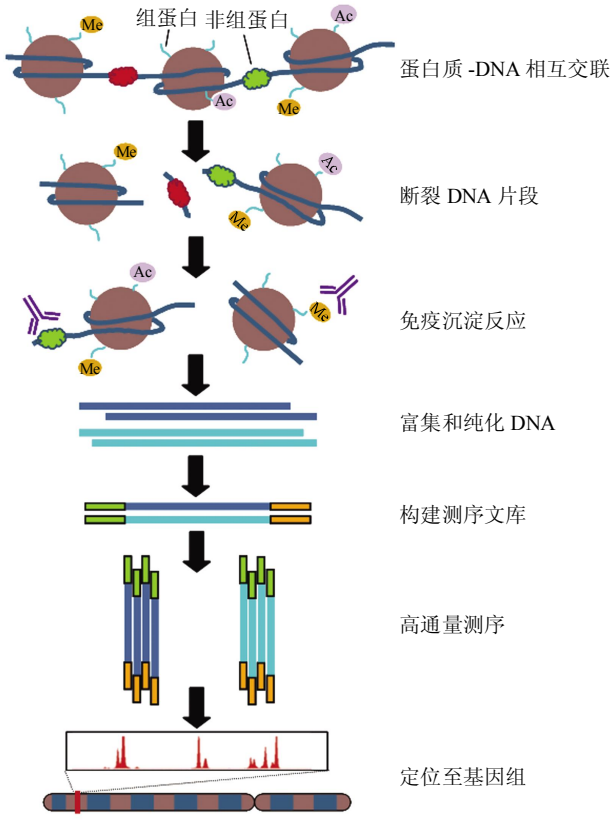


**Fig. 1　Workflow of a standard ChIP-Seq experiment**

**图 1　ChIP-seq 实验流程简图**

染色质免疫沉淀是该技术的基础，根据在 ChIP 过程中是否需要对蛋白质 -DNA 进行交联处理，ChIP 可分为 X-ChIP[12]和 N-ChIP[13]两类，其中 X 表示“交联”(crosslinking)，N 表示“自然状态”(native). X-ChIP 主要适用于结合力较弱的 DNA- 蛋白质相互作用的研究：首先，通过福尔马林处理处于生长状态下的活细胞，使 DNA 结合蛋白与 DNA 紧密交联；然后，采用超声仪超声处理染色质，将染色质随机打断为 200～1 000 bp 的小片段；接着，采用目的蛋白特异性的抗体免疫沉淀蛋白质 -DNA 复合体，富集目的蛋白靶定的 DNA 片段；最后，解交联蛋白质 -DNA 复合体，获取纯净的 DNA 片段. 而 N-ChIP 主要用于组蛋白修饰、核小体定位的研究. 由于组蛋白与 DNA 之间的作用力较强，因此在 N-ChIP 实验中不需要采用福尔马林处理活细胞，而是让 DNA 结合蛋白与 DNA 之间保持一种自然状态，然后，采用一种微球菌核酸酶(micrococcal nuclease，MNase)对染色质进行消化. 在染色质免疫沉淀过程中，抗体的质量和对照实验的设计是至关重要的. 选择特异且有效的抗体一方面能够尽可能全面地捕获相关信息，一方面减少了其他噪音的干扰，从而保障了后续实验的进行；而合理的对照试验设计则进一步保证了结果的客观性与真实性.

通过染色质免疫沉淀过程中抗原 - 抗体特异反应，有效地富集大量目的蛋白结合的 DNA 片段，接下来以这些 DNA 片段为原料，构建 DNA 测序文库，然后进行高通量测序. 时下用于高通量测序的工具主要有罗氏公司(Roche)的 454 测序仪(Roche GS FLX sequencer)、Illumina 公司推出的基因组分析平台(包括 HiSeq2000、HiSeq1000、Solexa Genome Analyzer IIx)、ABI(Applied Biosystem)公司研发的 SOLiD 测序仪(包括 SOLiD3.0、SOLiD4.0)和 Helicos Biosciences 公司的单分子测序仪(HeliScope)，其中采用 Illumina Solexa 基因组分析平台进行 ChIP-Seq 已有较多文献报道[14].

## 2　ChIP-seq 与 ChIP-chip 技术特点对比

ChIP-seq 和 ChIP -chip 都是高通量研究蛋白质 -DNA 相互作用的方法，但 ChIP-seq 具有许多 ChIP-chip 技术无法媲美的优势[11]. a. ChIP-seq 最大的优势就是可以精确到单个核苷酸的分辨率. 尽管可以通过增加微阵列上排布的探针数目来提高 ChIP-chip 的分辨率，但是对于大型基因组而言，这就意味着需要增加大量的探针和费用[15]. 除此之外，探针杂交过程中的不确定性也限制了 ChIP-chip 的分辨率. b. ChIP-seq 减少了 ChIP-chip 探针杂交过程中产生的噪音. 核酸杂交过程是一个相对复杂的过程，受诸多因素影响，如靶序列和探针序列的 GC 含量、长度、拷贝数、二级结构等，因此，不完全匹配的序列间相互杂交事件经常发生，不可避免地引入大量噪音. c. 由于芯片信号识别

过程中弱信号常常会被丢弃而强信号会饱和，因此通过 ChIP-chip 获得的信号强度往往不是线性的，其信号强度的量程限定在一定范围内，而 ChIP-seq 有效地避免了这种情况．Alekseyenko 等[16]在相同的实验条件下比较了 ChIP-seq 和 ChIP-chip 数据识别能力，发现 ChIP-seq 中出现的一些明显的且具有生物学意义的峰(peaks)常常在 ChIP-chip 中被掩盖．d．ChIP-seq 的显著优势就是基因组覆盖范围不受固定在微阵列上探针的限制．这对分析基因组上的重复区域(如微卫星序列)尤其重要，ChIP-seq 可以通过分析重复序列的侧翼序列将这些重复序列精确定位到基因组，而这些信息往往在微阵列方法中被掩盖．在覆盖范围方面，ChIP-seq 可以覆盖整个基因组而 ChIP-chip 通常是选择性地扫描一些特定区域，因此可以在一个更广阔的视野总览基因组全貌．比如，为了考察某个转录因子在基因组上的结合情况，传统的 ChIP-chip 通常是选取已知基因的 GpC 岛或转录起始位点(transcriptional start site，TSS)附近的片段作为考察对象，这种方法通常可以获取大部分的转录因子结合位点．然而最近人们发现一些转录因子(如 p53、cMyc 等)其结合位点在基因组上的分布远非我们之前认为的那样局限而是广泛分布于整个基因组．对于这类转录因子，ChIP-chip 往往是不合适的．e．随着测序技术的发展，每一次测序运行(run)容纳的 reads 数目越来越多，通过在标本准备阶段对不同的样本进行不同的接头(adaptor)标记处理，可以实现在一次运行中同时检测多个样本[17]，如此大大提高 ChIP-seq 的性价比．f．在测序所需样本数量方面，ChIP-chip 需要多达 4～5 μg 的起始样本，通常在杂交之前需要进行连接介导聚合酶链反应(ligation-mediated polymerase chain reaction，LM-PCR)，可能导致背景增高或竞争性扩增导致假阳性[18]．而 ChIP-seq 仅需要 ng 级起始材料，可以只需要很少的扩增循环或完全不需要扩增，并且随着单分子测序平台(single-molecule sequencing platform)[19]以及即将面世的第三代测序技术(the third-generation sequencing technology)[20]的出现，测序所需的 DNA 总量更是大大减少，这对于那些难以大量培养的细胞(如神经细胞和干细胞)或难以区分个别细胞与总体细胞行为的实验是很有裨益的．

## 3　ChIP-seq 数据分析

ChIP-seq 测序完成后会产生海量的数据，正确而有效地处理这些数据是体现 ChIP-seq 技术应用价值的关键步骤．ChIP-seq 数据分析的具体流程见图 2．

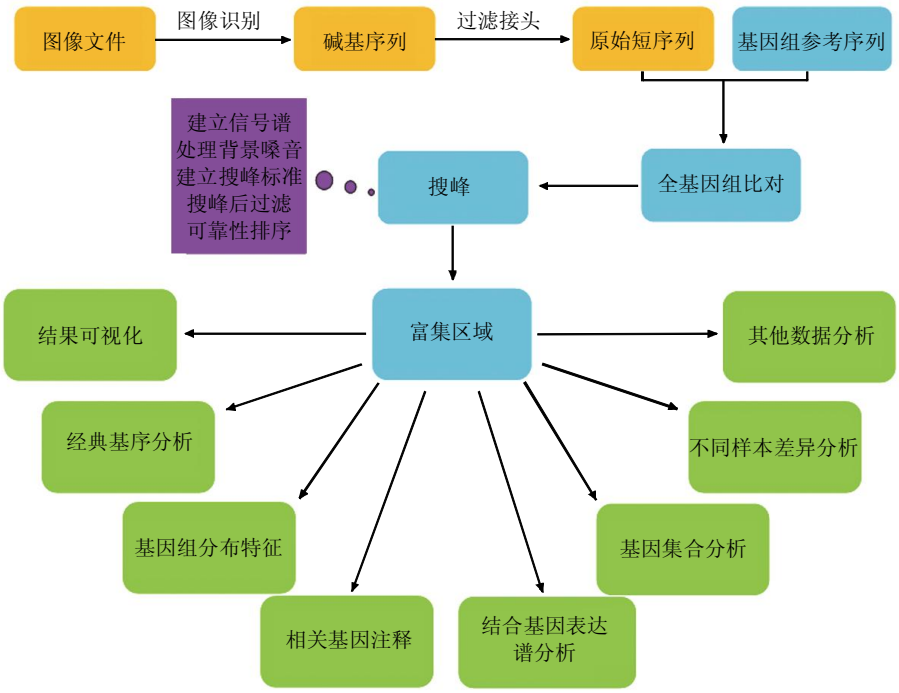


**Fig. 2　Overview of ChIP-seq data analysis**

**图 2　ChIP-seq 数据分析过程**

### 3.1　ChIP-seq 数据的初步处理及基本分析

新一代测序结果的原始数据形式是图像文件. 因此，对测序结果进行图像识别(base calling)是数据分析的第一步. 在每一次测序结束之后，这些图像通过一定的计算机算法处理可转换为碱基序列，然后通过对这些碱基序列进行接头序列过滤处理获得大量的具有一定长度的原始序列，又称 reads 或 tags，在此环节可完成 reads 长度和 reads 数量的统计以及数据产量的估算.

接下来就是将这些测序序列比对到一个参考基因组上. 然而，将几十甚至几百兆的序列最优地比对到一个庞大的基因组(特别是哺乳动物基因组)上，并非一件容易的事. 因此，序列比对仍然是整个数据分析过程中运算量最大的步骤，在这一步中选择合适的比对策略是至关重要的. 时下比较盛行的比对软件主要有 ELAND、BWA[21]、MAQ[22]、Bowtie [23]、SOAP[24]等. 目前的比对软件虽然都综合平衡了准确度、速度、内存、灵活性等各方面，但每款软件都有各自优缺点，尚没有哪款软件可以适用于所有情况. 值得指出的是，由于测序过程中或多或少存在误差，或样本基因组与参考基因组之间存在一些细小差别，如单核苷酸多态性(single nucleotide polymorphism， SNPs)或插入与缺失(indels)等，故运用软件进行比对时应当允许适当的不匹配情况. 目前已发表的文献中一般允许不超过 2 个的错配[10]. 此外，大量软件在高通量处理数据的过程中往往自动丢弃一些非唯一的序列(non-unique tags)[25]，故当涉及基因组重复区域研究时，要慎重考虑这一特点，并小心地处理这些非唯一的序列.

在原始短序列成功地比对至参考基因组后，下一步工作就是确定 ChIP-seq 峰富集区域(ChIP-seq peak enrichment regions)，即搜峰(peak calling). 确定 ChIP-seq 峰富集区域的依据是样本数据相对对照数据具有显著差异，具体指标通常有两个：“峰”的分值与统计显著性指标. 其中“峰”的分值用以表示这个峰的信号大小(排除随机因素)，常用读段密度(tag density)和富集倍数(fold enrichment, 即样本数据与对照数据之比)衡量；而统计显著性指标则用以表示该峰的可信度，通常用错误发现率(false discovery rate，FDR)衡量. 大量现存的搜峰软件(peak callers)如 MACS[26]、Peakseq[27]、ZINBA[28]等都能沿着基因组进行全面扫描最终实现 ChIP-seq 峰富集区域的鉴定. 然而，研究人员通过比较不同 DNA 结合蛋白的峰形图，发现不同 DNA 结合蛋白在基因组上的分布模式是不同的，具体体现于 ChIP-seq 峰形的不同，如转录因子 TCF 的峰形为尖锐状(sharp peak，即信号高度集中)，组蛋白标记 H3K27me3 的峰形为连绵状(broad peak，信号跨越一定范围)，而 RNA 聚合酶 Pol Ⅱ 的峰形则两者兼有[29]. 不同峰型对算法的要求不同，研究人员应根据 DNA 结合蛋白的分布模式选择合适的搜峰算法，目前适用于分析 sharp peak 的代表性软件主要有 MACS[26, 30]，分析 broad peak 的代表性搜峰软件有 SICER[30-31]和 CCAT[30, 32]，而分析混合型峰的代表搜峰软件有 ZINBA[28, 30]. 此外，对搜峰软件返回结果的可靠性进行检验也是必要的，这需要借助于一些后续实验或数据分析，如随机选取一系列位点通过 qPCR 实验进行验证，或计算搜寻到的峰与其附近已知蛋白结合基序之间距离的分布情况进行验证，抑或检验一些已经被报道的靶定位点是否存在峰，如果转录因子的结合基序(motif)是已知的，则可以计算峰序列中包含基序序列的百分比，间接估计实验结果的可靠性. 至此，ChIP-seq 数据的初步分析基本完成.

### 3.2　ChIP-seq 数据的后续分析

通过对海量数据的初步分析确定了 ChIP-seq 的富集区域，然而，ChIP-seq 技术发挥真正价值的地方不仅在于获得 DNA 相关蛋白在基因组上的分布信息，更在于挖掘这些信息背后隐藏的更深奥的“秘密”. 许多优秀的相关文献其亮点就在于充分挖掘隐藏在数据背后的信息.

对于转录因子，最常见的后续研究就是分析其经典结合基序[33]. 将得分高的峰序列输入相关软件，如 MEME[34]、Weeder[35]等，软件将返回一些潜在的基序并附带其统计学显著性. 在一些情况下，返回的潜在基序中存在一个基序其相对其他基序具有显著的统计学意义，且这种显著性受输入序列数目的影响不大. 而在另一些情况下，返回的潜在基序其统计学显著性呈逐步下降趋势. 在这种情况下，进一步分析这些基序共同出现的概率可能有利于发现该转录因子的相互协助因子或发现基序间一些更为复杂的交互作用. 此外，还可以通过 Jaspar[36]获取一些已知的转录因子 DNA 结合基序，考察这些已知的基序在 ChIP-seq 峰附近的分布情况，有利于发现不同转录因子之间的联系[37].

我们还可以通过软件(如 GPAT[38])分析 ChIP-seq 峰在基因组不同区域分布的偏好性，如基因的转录起始位点、外显子 - 内含子拼接处、3′端等，考察目的蛋白在基因转录调控过程中可能发挥的作用. 例如，Schwartz 等[39]使用 CHIP-seq 技术测定了人类与小鼠核小体定位，发现核小体趋向定位于 3′端外显子上，H3K36me3 修饰与 RNA 聚合酶 Pol Ⅱ也都倾向于发生在外显子上的核小体中，在一定程度暗示外显子上核小体定位和组蛋白修饰情况可能与基因剪接有很大的关系. 又如，全基因组核小体定位图谱显示，一个活跃的基因其转录起始位点通常被认为含有丰富的 H3K4me3 而其增强子富含 H3K4me1，据此我们可以猜想 H3K4me3、H3K4me1 修饰可能具有促进基因转录的功能，同时通过考察基因调控区组蛋白修饰情况也可以帮助我们了解基因的活跃度[6, 40].

“峰”关联基因的 GO 注释及 Pathway 分析也是后续分析的重要分支. 倘若一群基因受同一个转录因子调控或具有相同的表观修饰，我们可以通过 GO(gene ontology)[41]或 KEGG(kyoto encyclopedia of genes and genomes)[42−43]数据库对它们进行注释，从而考察这些基因的共同属性(如共同参与的生物信号通路或生物学过程)，抑或通过 GSEA(gene set enrichment analysis)[44]软件分析这个基因集合与其他已知的具有一定功能注释的基因集合之间的联系，进而阐述调控它们的转录因子或表观修饰的生物学功能[45].

当然，结合其他数据来分析 ChIP-seq 数据或综合多个 ChIP-seq 数据将有利于获得更多有效信息. 如可以结合基因表达谱来分析 ChIP-seq 数据. 通过考察基因表达的差异性来初步确定某个转录因子直接靶定的基因群体或明确一个转录因子对某靶基因的调控作用(正调控还是负调控)，若一种染色质修饰标记常出现在高表达基因的启动子区，我们可以推断这种修饰可能具有转录促进作用. 又如，通过结合 SNP 数据，ChIP-seq 可用来发现等位基因特异性的转录因子结合或表观修饰[7]. 再如，甲基化标志 H3K4me3 被发现经常与基因活跃关联，而 H3K27me3 则经常与基因沉寂关联，结合分析这两个图谱更能明确基因活跃情况[7].

此外，通过 ChIP-seq 数据分析还可以获得许多新的发现. 有人通过分析 H3K4me3 和 H3K36me3 的 ChIP-seq 数据，发现这些通常分别位于基因启动子或基因转录区域的组蛋白修饰可以用来鉴定多数非编码 RNAs[46]. 2012 年，Birnbaum 等[47]通过 ChIP-seq 分析启动子在基因组上的位置分布情况，发现一些基因的启动子坐落在其邻近基因的编码区. 该研究提示我们，某一个基因编码区的变异除了可能改变该基因编码的蛋白质之外，也有可能影响另外一个基因的转录过程.

### 3.3　ChIP-seq 数据分析综合性软件 CisGenome

在 ChIP-seq 数据分析过程中，借助各种生物信息软件和统计软件是必不可少的. 表 1 列举了 ChIP-seq 数据分析过程一些关键环节中常用到的软件，各软件之间的具体比较此处不赘述，详情可参见文献[48−51]. 此处主要介绍其中的一款综合性分析软件 CisGenome (http://www.biostat.jhsph.edu/~hji/cisgenome/)[52−54].

**Table 1　A subset of software tools available for the key steps in ChIP-seq data analysis**

**表 1　ChIP-seq 数据分析关键环节中的一些常用软件**

| 数据分析流程中的应用阶段 | 软件 | 主要特点 |
|---|---|---|
| 序列比对 | ELAND | Illumina 默认软件；比对过程中不允许碱基的空缺，且比对序列长度受限. |
| | BWA | 基于 BWT(Burrows-Wheeler transform)算法；运算快速高效，比对过程中允许适度插入与缺失. |
| | Bowtie | 基于 BWT(Burrows-Wheeler transform)算法；速度超快，且具备高存储效率. |
| | MAQ | 比对过程中不允许碱基的空缺，但能考虑到每个碱基的质量指数. |
| | SOAP | 比对过程中允许少量碱基的空缺和错配. |
| 搜峰 | MACS | 能自动将数据调整成动态泊松分布；且搜峰过程可以不依赖对照组数据，自动进行数据拟合. |
| | PeakSeq | 搜峰过程中能兼顾基因组区域结构特点；通过计算 FDR 来确定峰富集区域. |
| | ZINBA | 搜峰过程中能兼顾基因组区域结构特点；可以分析尖锐状峰型和连绵状峰型两类 ChIP-seq 数据. |

CisGenome 目前主要用于分析 ChIP-chip 和 ChIP-seq 数据，其独特的模块化设计开创了可视化用户界面和数据自定义批量处理功能，支持数据间的交互式分析．首先，嵌入式的浏览器可实现多种数据的可视化，如微阵列图像、各序列信号强度、序列相关基因的结构以及序列在基因上的具体位置、峰序列的保守性分析、DNA 序列及基序信息等[52, 54]．同时，能够实现数据的标准化、搜峰并计算各 ChIP-seq peaks 富集区的 FDR 值、考察各 ChIP-seq peaks 富集区在物种间的保守性、明确各峰分布在哪些基因中以及其在基因上的位置、计算各 reads 的丰度或信号强度、发现新的结合基序或将已知的基序定位至特定的 DNA 序列中等等[52-53]．在搜峰算法方面，CisGenome 提供了两个优化功能，即在搜峰后过滤只有单链富集的峰和将峰平移片段长度的一半以搜峰后优化的方式来处理方向性的信息，进一步实现了结合位点的精确定位．此外，CisGenome 浏览器是典型的本地版基因组浏览器，所有 reads 数据、注释信息都存于本地文件，因此不需要网络连接，方便内部考查数据用．借助 CisGenome 来处理 ChIP-seq 海量数据将会事半功倍．

除 CisGenome 经典软件外，目前可用于 ChIP-seq 数据分析的综合软件还有 Pyicos[55]、CASSys[56]等．在不久的将来，大量用户界面友好的分析软件还将大量涌现，然而各软件的偏重可能不尽相同，研究人员必须根据自己的实验设计和研究目的选用适合的软件．

## 4　ChIP-seq 的实际应用

关于 ChIP-seq 技术的公开报道最早开始于 2007 年[6-10]，且最先应用于表观遗传学研究．目前，ChIP-seq 主要应用于两个方面：一方面是 DNA 序列上转录因子结合位点的识别，如启动子、增强子等各种顺式作用元件(*Cis*-acting element)的识别；另一方面主要应用在表观遗传学领域，包括研究全基因组组蛋白修饰、核小体定位和 DNA 甲基化等问题．

### 4.1　转录因子结合位点研究

转录因子(transcription factor，TF)是一类很重要的蛋白质分子，其可以通过靶定调控一些下游效应分子，引发一系列级联反应，从而发挥强大的生物学作用．全基因组范围内明确这些转录因子的结合位点是揭示这些转录因子生物学功能和机制的基础，同时也是绘制基因调控网络不可缺少的部分．

2007 年，Robertson 等[9]最先使用 ChIP-seq 技术在全基因组范围内考察了 IFN-γ 诱导和非诱导两种状态下转录因子 STAT1 (signal transducer and activator of transcription 1) 的结合情况．同年，Johnson 等[10]也使用 ChIP-seq 技术研究了神经元限制性沉默因子 NRSF (neuron-restrictive silencer factor)在全基因组上的结合情况．这些研究验证了 ChIP-seq 技术应用于该领域的可行性和有效性，开辟了使用高通量测序技术在全基因组范围内研究转录因子调控作用的新纪元．

目前，ChIP-seq 已广泛应用于研究一些经典转录因子和一些新的转录因子在全基因组上的结合情况，实现全基因组层面分析一些经典的转录因子的调控网络，或为新转录因子的功能研究提供一些线索．研究人员可通过分析这些转录因子结合序列的特征发现它们的经典作用基序(binding motif)或协同作用因子．如 Copeland 等通过分析转录因子 EVI1 靶定序列的特征，发现在 EVI1 的 ChIP-seq 峰附近常常伴随有 FOS 结合位点，并通过实验证实 FOS 是 EVI1 的相互作用分子[37]．研究人员也可通过对目的蛋白靶定的下游效应分子的注释与分析来揭示它们发挥的生物学作用或具体机制．Durant 等[57]在 $CD4^+$T 细胞中分析了 STAT3 的靶基因，发现 STAT3 的大量靶基因涉及 Th17 细胞的分化、增殖、生存等过程，明确了 STAT3 在机体炎症发生过程中的作用．研究人员亦可通过分析这些转录因子在基因组上的位置分布情况来拓展其基因调控作用．之前认为转录因子通常分布在基因启动子区和增强子区，然而通过 ChIP-chip 和 ChIP-seq 考察一些转录因子在基因组上的分布情况，发现有些转录因子结合位点分布在远离已知基因 TSS 的位置或广泛分布在整个基因组上[58-59]，这可能暗示新的靶基因或新的调控机制的存在．此外，通过比较某转录因子在不同阶段或不同状态的组织或细胞中靶定位点的差异，研究人员还可分析该转录因子在细胞不同阶段或不同状态下的不同作用．Ross-Innes 等[60]通过全基因组考察乳腺癌患者雌激素受体的结合情况，发现雌激素受体在不同的患者中结合位点存在差异，而且这种差异与患者的预后密切相关．

### 4.2　组蛋白修饰研究

组蛋白是存在于真核生物染色质中的一组进化上非常保守的碱性蛋白质，分为 H1、H2A、H2B、H3、H4 5 种类型，其中 H2A、H2B、H3、H4 被

称作核心组蛋白(core histone)，四者各 2 个形成组蛋白八聚体，构成核小体的核心，是 DNA 压缩组装时所依赖的线轴；而 H1 被称做连接组蛋白(linker histones)，与核小体间的 DNA 结合．通常情况下，这些结合在 DNA 上的组蛋白常常被修饰，修饰方式如甲基化、乙酰化、磷酸化等．各组蛋白可在不同氨基酸残基位点发生不同方式、不同数目的修饰，如此便大大丰富了组蛋白修饰的形式，更复杂的是，不同形式的组蛋白修饰之间又存在相互作用，因此，早先就有科学家提出“组蛋白密码子(histone code)”的假说[61-63].

迄今为止，组蛋白修饰的具体作用仍不甚清楚，目前已有的大量研究表明组蛋白修饰主要参与基因转录调控，且不同的组蛋白修饰其转录调控作用不同[64]．如 H3K4Me3、H3K36Me3 常常与基因活化密切相关，而 H3K27Me3、H3K9Me2/3、H4K20Me3 常常与基因活性抑制密切相关．2007 年，Barski 等[6]率先利用新一代测序技术平台 Solexa IG 发明了 ChIP-seq 技术，并用该技术绘制了人 $CD4^+$T 细胞中 20 种组蛋白甲基化修饰、组蛋白变体 H2A.Z、RNA 聚合酶 PolⅡ和绝缘子结合蛋白 CTCF 的全基因组高分辨定位图谱．随后，他们又继续在人 $CD4^+$T 细胞中研究了 18 种组蛋白乙酰化修饰[65]．这些研究提示了各种组蛋白修饰的新功能以及多种组蛋白修饰协作的重要性．Mikkelsen 等[7]应用 ChIP-seq 技术在小鼠胚胎干细胞及其他两种已分化细胞中考察了多种组蛋白修饰(H3K4me3、H3K27me3、H3K36me3、H3K9me3 和 H3K20me3)的情况，探究了各种组蛋白修饰在细胞分化中的作用．

同时，组蛋白修饰也与 DNA 损伤修复相关．如，当细胞受到外界损伤导致 DNA 双链断裂后可诱导组蛋白变体 H2AX 第 139 位丝氨酸磷酸化形成 γ-H2AX，γ-H2AX 进一步招募损伤修复相关蛋白聚集于 DNA 损伤位点进行修复工作．γ-H2AX 已被广泛认为是 DNA 双链断裂的标记物[66-67]．2011 年底，Dmitrieva 等[68]采用 ChIP-seq 检测了高浓度 NaCl 条件下细胞 DNA 双链断裂发生的频率以及断裂发生的位点，发现在高浓度 NaCl 条件下 DNA 双链更容易发生断裂，但这些断裂位点主要分布于非基因区，该结果从另一个视角解释了为什么高浓度 NaCl 诱导的 DNA 双链断裂相对无害．还有研究表明，H3K56Ac 与基因组稳定性的维持密切相关[69-70]．这些研究都将为后续进一步揭示组蛋白修饰的生物学功能奠定了基础.

### 4.3 核小体定位研究

核小体(nucleosome)是构成真核生物染色质的基本结构单位，由 DNA 缠绕在组蛋白八聚体上构成．各核小体通过核小体之间的自由 DNA 双链(linker DNA)念珠样串联成染色质．念珠样串联的核小体在基因组 DNA 分子上的精确位置称为核小体定位，核小体定位能影响染色质的包装状态和 DNA 的开放程度，已被证实在基因转录调控、DNA 复制与修复、可变剪切等过程中扮演着重要的角色[71-74]．因此，了解核小体定位及其功能有助于揭示染色质结构对基因表达的影响，同时对阐述生物体多种生物学过程也具有十分重要的作用．

然而，核小体在 DNA 上的定位非常复杂，其一方面要完成 DNA 的高度压缩组装，另一方面还要对基因的表达进行精确的调控．那么核小体在 DNA 上具有怎样的分布规律，以及在基因转录过程中有哪些调控功能呢？目前的研究结果仍无法准确地回答这个问题．因此，核小体定位及其功能研究是一项复杂而意义重大的任务．许多研究者使用包括生化实验、生物信息学等多种方法来揭示体内核小体在基因组上的定位规律及其功能．特别是最近几年来，随着 ChIP-chip 与 ChIP-seq 等高通量技术的发展，多种生物(如酵母[8, 75]、线虫[76]、果蝇[77]、小鼠[78]、人[79]等)的高分辨全基因组核小体定位图谱已被测定并公开．这些研究一方面在全基因组层面验证了核小体的分布规律，如：活动基因的转录起始位置和转录终止位点存在核小体空缺区域，而一些重要的保守区域核小体结构出现频繁；功能性的转录因子结合位点周围核小体定位水平较低；核小体定位会因时空的改变而动态摆动，其动态性主要体现在不同细胞类型、细胞的不同状态、不同发育阶段、以及特定刺激等条件下，核小体定位状态不同．另一方面探索了核小体定位在基因表达调控中的功能，指出核小体定位与转录相关的蛋白质因子(包括 RNA 聚合酶、转录因子、染色质改构复合物等)之间是相互联系的，它们以一种协调的、系统的方式共同调控基因的转录水平，在基因转录起始阶段、转录延伸阶段、基因表达模式多样化以及可变剪接等过程中挥发着重要的调控作用．Chen 等[80]在酵母中考察 Cyc8 或 Tup1 蛋白缺失的情况下核小体在基因组上的定位情况，验证了酵母 Cyc8-Tup1 复合物能通过促进核小体在靶基因启动子区定位，进而抑制下游靶基因的转录调控.

### 4.4 DNA 甲基化研究

DNA 甲基化是表观遗传修饰的一个重要机制，在基因转录调控中起着重要作用．近年来，大量研究都表明 DNA 甲基化修饰在胚胎发育、基因组印记、肿瘤发生、基因调控以及转座子沉默等生物过程中发挥着重要作用．因而，寻找基因组上的甲基化区域，明确各种细胞、组织、甚至疾病样本的 DNA 甲基化修饰模式，并比较不同细胞、组织、甚至疾病样本间的 DNA 甲基化修饰模式的差异，有助于明确基因转录调控机制和疾病发生机制．

目前，最经典的全基因组研究 DNA 甲基化的方法是高通量的重亚硫酸盐测序(bisulfite conversion followed by pyrosequencing)[81–82]，其原理为先采用重亚硫酸盐处理 DNA，使 DNA 中未发生甲基化的胞嘧啶脱氨基转变成尿嘧啶，而甲基化的胞嘧啶保持不变，然后进行 PCR 扩增反应，则尿嘧啶全部转化成胸腺嘧啶，最后对 PCR 产物进行测序，并且与未经重亚硫酸盐处理的序列比较，判断哪些位点发生了甲基化．此方法具有很高的可靠性及精确度，但是费用较高．事实上，ChIP-seq 也常常应用于全基因组甲基化研究[83–84]，通过 5′- 甲基胞嘧啶抗体特异性富集基因组上发生甲基化的 DNA 片段，然后高通量测序定位 DNA 甲基化位点．用于甲基化研究的 ChIP-seq 又称为 MeDIP-seq (methylated DNA immunoprecipitation sequencing)．MeDIP-Seq 是基于抗体富集的实验方法而进行的 DNA 甲基化分析，因此可以通过测序数据在基因组上的富集性分布而反映出相应区域的甲基化状态，但是该方法尚不能确定富集下来的 DNA 片段中发生甲基化 C 碱基的确切位点，所以无法像亚硫酸氢盐测序方法那样精确得到单个碱基的甲基化情况．

### 4.5 其他方面

真核生物的基因组染色质并非是简单的线性 DNA 双链，而是通过蛋白质介导 DNA 长距离相互作用从而组建成高级染色质结构．ChIP-seq 可以通过捕获这些发挥介导作用的蛋白质进而揭示高级染色体结构．如，转录因子 CTCF(CCCTC-binding factor)是一类经典的染色质组装蛋白，可将两条远离的 DNA 双链聚集在一起形成染色质环(chromatin loop)，从而改变染色质的 3D 结构进而影响基因的表达调控．Wang 等[85]通过 ChIP-seq考察了人类 19 种细胞 CTCF 结合情况，发现 CTCF 结合位点与 DNA 甲基化区域联系紧密；同时，Lee 等[86]发现 CTCF 结合位点与黏连蛋白结合位点高度重叠，这些研究都暗示了高级染色质结构与基因转录调控机制之间的相互联系．

端粒是存在于真核细胞线状染色体末端的 DNA 重复序列，它与端粒结合蛋白一起构成了特殊的“帽子”结构，能够维持染色体的完整．Vaquero-Sedas 等[87]利用 ChIP-seq 数据分析端粒表观遗传特征，发现端粒处组蛋白 H3 分布较着丝粒处密集，其中异染色质标记(H3K9Me2 和 H3K27Me)分布较着丝粒处少，而一些常染色质标记(H3K4Me2 和 H3K9Ac)相对较多，且含有较多的 H3K27Me3 结合位点，该结果提示端粒也是基因调控的特殊位点．

## 5　问题和展望

ChIP 技术自 1997 年由 Orlando 等[88]创立以来，已经成为阐述转录级联调控过程和破译染色质编码信息的主要工具．在芯片技术和高通量测序技术的推进下，ChIP 技术先后经历了 ChIP 到 ChIP-chip 再到 ChIP-seq 的转型．ChIP-chip 和 ChIP-seq 都可以从全基因组范畴分析转录因子或其他染色质蛋白结合位点及表观遗传机制，但是 ChIP-seq 相对于 ChIP-chip 已经具有明显的优势，它能够以较低的成本获取高分辨率和低噪音的数据．目前，ChIP-seq 在一些领域已取得显著的成绩，其中最典型的成就包括在核小体水平描绘了全基因组染色质修饰的情况以及全基因组范畴精确鉴定出基因转录调控相关的 DNA 序列元件．随着空间分辨率的进一步提高，测序所需细胞数目的减少以及测序容量的继续大幅度增长，ChIP-seq 预期可以实现跨多个组织、细胞类型、生理条件和发展阶段的综合分析，有望在未来几年内成为该研究领域主导的方法．

人类基因组序列测定工作完成后，科学家们致力于研究基因组的功能．起始于 2003 年 9 月的 DNA 分子百科全书计划(the encyclopedia of DNA elements，ENCODE)[89–90]和 NIH 于 2008 年启动的表观基因组学路线图计划(roadmap epigenomics project)[91]是研究基因表达调控奥秘的国际合作大项目．其中，ENCODE 计划旨在鉴定人类 1%基因组序列中所有的结构和功能元件，从而构建一份完整的人类基因组的“元件目录”，而表观基因组学路线图计划旨在鉴定、分类和描述人类组织特异性的全基因组表观修饰概貌．在大量科研工作者的共同

努力下，目前已存在大量的在线公开数据库，如GEO(gene expression omnibus)[92-93]、Oncomine[94-95]，结合这些公共数据综合分析 ChIP-seq 数据结果以获取更多有效的信息也是至关重要的．例如，ChIP-seq 数据结合基因表达芯片数据或 RNA-Seq 数据，也许更能促进我们对基因调控网络以及表观基因组和转录组之间相互作用的理解．不可置疑，ChIP-seq 将在一定程度上推进上述计划的完成．

像其他任何技术一样，ChIP-seq 也不可避免地存在一些问题：首先包括染色质免疫沉淀实验过程中存在的问题，如染色质的打碎程度与开放程度、抗体的可获得性与有效性、实验设计是否合理等；其次还包括高通量测序过程中存在的问题，如高通量测序误差的存在、测序数据 GC 含量偏向、reads 长度有限、测序深度选择的难题等．这些问题随着实验操作的规范和技术的改进将逐步得以缓解．然而，许多实验室在 ChIP-seq 技术应用过程中面临的最大挑战是如何有效地管理和分析 ChIP-seq 获取的海量数据．因此，迫切需要实验研究人员和生物信息工作人员共同努力以开发用户界面友好且功能强大的数据管理和数据分析软件．最近，ENCODE 研究团队通过归纳和提炼目前已完成的 400 多个 ChIP-seq 实验，发展了一套具体的实验规范和数据分析指南[96]，将有效地指导我们正确运用 ChIP-seq 这一新技术．

## 参 考 文 献


[1] Hillier L W, Marth G T, Quinlan A R, *et al*. Whole-genome sequencing and variant discovery in *C. elegans*. Nat Methods, 2008, **5** (2): 183-188

[2] Wheeler D A, Srinivasan M, Egholm M, *et al*. The complete genome of an individual by massively parallel DNA sequencing. Nature, 2008, **452** (7189): 872-876

[3] Ley T J, Mardis E R, Ding L, *et al*. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature, 2008, **456** (7218): 66-72

[4] Nagalakshmi U, Wang Z, Waern K, *et al*. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science, 2008, **320** (5881): 1344-1349

[5] Wilhelm B T, Marguerat S, Watt S, *et al*. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature, 2008, **453** (7199): 1239-1243

[6] Barski A, Cuddapah S, Cui K, *et al*. High-resolution profiling of histone methylations in the human genome. Cell, 2007, **129** (4): 823-837

[7] Mikkelsen T S, Ku M, Jaffe D B, *et al*. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature, 2007, **448** (7153): 553-560

[8] Albert I, Mavrich T N, Tomsho L P, *et al*. Translational and rotational settings of H2A. Z nucleosomes across the *Saccharomyces cerevisiae* genome. Nature, 2007, **446** (7135): 572-576

[9] Robertson G, Hirst M, Bainbridge M, *et al*. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods, 2007, **4** (8): 651-657

[10] Johnson D S, Mortazavi A, Myers R M, *et al*. Genome-wide mapping of *in vivo* protein-DNA interactions. Science, 2007, **316** (5830): 1497-1502

[11] Schones D E, Zhao K. Genome-wide approaches to studying chromatin modifications. Nat Rev Genet, 2008, **9** (3): 179-191

[12] Orlando V. Mapping chromosomal proteins *in vivo* by formaldehyde-crosslinked-chromatin immunoprecipitation. Trends Biochem Sci, 2000, **25** (3): 99-104

[13] O'Neill L P, Turner B M. Immunoprecipitation of native chromatin: NChIP. Methods, 2003, **31** (1): 76-82

[14] Schuster S C. Next-generation sequencing transforms today's biology. Nat Methods, 2008, **5** (1): 16-18

[15] Kim T H, Barrera L O, Zheng M, *et al*. A high-resolution map of active promoters in the human genome. Nature, 2005, **436** (7052): 876-880

[16] Alekseyenko A A, Peng S, Larschan E, *et al*. A sequence motif within chromatin entry sites directs MSL establishment on the *Drosophila* X chromosome. Cell, 2008, **134** (4): 599-609

[17] Lefrancois P, Euskirchen G M, Auerbach R K, *et al*. Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. BMC Genomics, 2009, **10**: 37

[18] O'Geen H, Nicolet C M, Blahnik K, *et al*. Comparison of sample preparation methods for ChIP-chip assays. Biotechniques, 2006, **41** (5): 577-580

[19] Pushkarev D, Neff N F, Quake S R. Single-molecule sequencing of an individual human genome. Nat Biotechnol, 2009, **27** (9): 847-850

[20] Branton D, Deamer D W, Marziali A, *et al*. The potential and challenges of nanopore sequencing. Nat Biotechnol, 2008, **26** (10): 1146-1153

[21] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 2009, **25** (14): 1754-1760

[22] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res, 2008, **18** (11): 1851-1858

[23] Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 2009, **10** (3): R25

[24] Li R, Li Y, Kristiansen K, *et al*. SOAP: short oligonucleotide alignment program. Bioinformatics, 2008, **24** (5): 713-714

[25] Newkirk D, Biesinger J, Chon A, *et al*. AREM: aligning short reads from ChIP-sequencing by expectation maximization. J Comput Biol, 2011, **18** (11): 1495-1505

[26] Zhang Y, Liu T, Meyer C A, *et al*. Model-based analysis of ChIP-Seq (MACS). Genome Biol, 2008, **9** (9): R137

[27] Rozowsky J, Euskirchen G, Auerbach R K, *et al*. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol, 2009, **27** (1): 66−75

[28] Rashid N U, Giresi P G, Ibrahim J G, *et al*. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. Genome Biol, 2011, **12** (7): R67

[29] Park P J. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet, 2009, **10** (10): 669−680

[30] Kidder B L, Hu G, Zhao K. ChIP-Seq: technical considerations for obtaining high-quality data. Nat Immunol, 2011, **12** (10): 918−922

[31] Zang C, Schones D E, Zeng C, *et al*. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics, 2009, **25** (15): 1952−1958

[32] Xu H, Handoko L, Wei X, *et al*. A signal-noise model for significance analysis of ChIP-seq with negative control. Bioinformatics, 2010, **26** (9): 1199−1204

[33] Tompa M, Li N, Bailey T L, *et al*. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol, 2005, **23** (1): 137−144

[34] Bailey T L, Williams N, Misleh C, *et al*. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res, 2006, **34** (Web Server issue): W369−373

[35] Pavesi G, Mereghetti P, Mauri G, *et al*. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res, 2004, **32**(Web Server issue): W199−203

[36] Sandelin A, Alkema W, Engstrom P, *et al*. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res, 2004, **32** (Database issue): D91−94

[37] Bard-Chapeau E A, Jeyakani J, Kok C H, *et al*. Ecotopic viral integration site 1 (EVI1) regulates multiple cellular processes important for cancer and is a synergistic partner for FOS protein in invasive tumors. Proc Natl Acad Sci USA, 2012, **109** (6): 2168−2173

[38] Krebs A, Frontini M, Tora L. GPAT: retrieval of genomic annotation from large genomic position datasets. BMC Bioinformatics, 2008, **9**: 533

[39] Schwartz S, Meshorer E, Ast G. Chromatin organization marks exon-intron structure. Nat Struct Mol Biol, 2009, **16** (9): 990−995

[40] Heintzman N D, Stuart R K, Hon G, *et al*. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet, 2007, **39** (3): 311−318

[41] Ashburner M, Ball C A, Blake J A, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000, **25** (1): 25−29

[42] Kanehisa M, Goto S, Sato Y, *et al*. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res, 2011, **40** (Database issue): D109−114

[43] Ogata H, Goto S, Sato K, *et al*. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 1999, **27** (1): 29−34

[44] Subramanian A, Tamayo P, Mootha V K, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA, 2005, **102** (43): 15545−15550

[45] Orford K, Kharchenko P, Lai W, *et al*. Differential H3K4 methylation identifies developmentally poised hematopoietic genes. Dev Cell, 2008, **14** (5): 798−809

[46] Guttman M, Amit I, Garber M, *et al*. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature, 2009, **458** (7235): 223−227

[47] Birnbaum R Y, Clowney E J, Agamy O, *et al*. Coding exons function as tissue-specific enhancers of nearby genes. Genome Res, 2012, **22**(6): 1059−1068

[48] Kim H, Kim J, Selby H, *et al*. A short survey of computational analysis methods in analysing ChIP-seq data. Hum Genomics, 2012, **5** (2): 117−123

[49] Wilbanks E G, Facciotti M T. Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One, 2012, **5** (7): e11471

[50] Laajala T D, Raghav S, Tuomela S, *et al*. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. BMC Genomics, 2009, **10**: 618

[51] Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform, 2012, **11** (5): 473−483

[52] Ji H, Jiang H, Ma W, *et al*. An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol, 2008, **26** (11): 1293−1300

[53] Ji H, Jiang H, Ma W, *et al*. Using CisGenome to analyze ChIP-chip and ChIP-seq data. Curr Protoc Bioinformatics, 2011, Chapter 2: Unit2 13

[54] Jiang H, Wang F, Dyer N P, *et al*. CisGenome Browser: a flexible tool for genomic data visualization. Bioinformatics, 2010, **26** (14): 1781−1782

[55] Althammer S, Gonzalez-Vallinas J, Ballare C, *et al*. Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. Bioinformatics, 2011, **27** (24): 3333−3340

[56] Alawi M, Kurtz S, Beckstette M. CASSys: an integrated software-system for the interactive analysis of ChIP-seq data. J Integr Bioinform, 2011, **8** (2): 155

[57] Durant L, Watford W T, Ramos H L, *et al*. Diverse targets of the transcription factor STAT3 contribute to T cell pathogenicity and homeostasis. Immunity, 2010, **32** (5): 605−615

[58] Wederell E D, Bilenky M, Cullum R, *et al*. Global analysis of *in vivo* Foxa2-binding sites in mouse adult liver using massively parallel sequencing. Nucleic Acids Res, 2008, **36** (14): 4549−4564

[59] Yu M, Riva L, Xie H, *et al*. Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. Mol Cell, 2009, **36** (4): 682−695

[60] Ross-Innes C S, Stark R, Teschendorff A E, *et al*. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature, 2012, **481** (7381): 389−393

[61] Strahl B D, Allis C D. The language of covalent histone modifications. Nature, 2000, **403** (6765): 41-45

[62] Jenuwein T, Allis C D. Translating the histone code. Science, 2001, **293** (5532): 1074-1080

[63] 蒋智文, 刘新光, 周中军. 组蛋白修饰调节机制的研究进展. 生物化学与生物物理进展, 2009, **36**(10): 1252-1259
Jiang Z W, Liu X G, Zhou Z J. Prog Biochem Biophys, 2009, **36**(10): 1252-1259

[64] Kouzarides T. Histone methylation in transcriptional control. Curr Opin Genet Dev, 2002, **12** (2): 198-209

[65] Wang Z, Zang C, Rosenfeld J A, *et al*. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet, 2008, **40** (7): 897-903

[66] Celeste A, Petersen S, Romanienko P J, *et al*. Genomic instability in mice lacking histone H2AX. Science, 2002, **296** (5569): 922-927

[67] Rogakou E P, Pilch D R, Orr A H, *et al*. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. J Biol Chem, 1998, **273** (10): 5858-5868

[68] Dmitrieva N I, Cui K, Kitchaev D A, *et al*. DNA double-strand breaks induced by high NaCl occur predominantly in gene deserts. Proc Natl Acad Sci USA, 2011, **108** (51): 20796-20801

[69] Ozdemir A, Spicuglia S, Lasonder E, *et al*. Characterization of lysine 56 of histone H3 as an acetylation site in *Saccharomyces cerevisiae*. J Biol Chem, 2005, **280** (28): 25949-25952

[70] Masumoto H, Hawke D, Kobayashi R, *et al*. A role for cell-cycle-regulated histone H3 lysine 56 acetylation in the DNA damage response. Nature, 2005, **436** (7048): 294-298

[71] Jiang C, Pugh B F. Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet, 2009, **10** (3): 161-172

[72] 刘 辉, 壮子恒, 关佶红,等.核小体定位的转录调控功能研究进展. 生物化学与生物物理进展, 2012, **39** (9): 843-852
Liu H, Zhuang Z H, Guan J H, *et al*. Prog Biochem Biophys, 2012, **39** (9): 843-852

[73] Hodges C, Bintu L, Lubkowska L, *et al*. Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase Ⅱ. Science, 2009, **325** (5940): 626-628

[74] Tilgner H, Nikolaou C, Althammer S, *et al*. Nucleosome positioning as a determinant of exon recognition. Nat Struct Mol Biol, 2009, **16** (9): 996-1001

[75] Lee W, Tillo D, Bray N, *et al*. A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet, 2007, **39** (10): 1235-1244

[76] Valouev A, Ichikawa J, Tonthat T, *et al*. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Res, 2008, **18** (7): 1051-1063

[77] Mavrich T N, Jiang C, Ioshikhes I P, *et al*. Nucleosome organization in the *Drosophila* genome. Nature, 2008, **453** (7193): 358-362

[78] Li Z, Schug J, Tuteja G, *et al*. The nucleosome map of the mammalian liver. Nat Struct Mol Biol, 2011, **18** (6): 742-746

[79] Schones D E, Cui K, Cuddapah S, *et al*. Dynamic regulation of nucleosome positioning in the human genome. Cell, 2008, **132** (5): 887-898

[80] Chen K, Wilson M A, Hirsch C, *et al*. Stabilization of the promoter nucleosomes in nucleosome free regions by the yeast Cyc8-Tup1 corepressor. Genome Res, 2013, **23** (2): 312-322

[81] Meissner A, Mikkelsen T S, Gu H, *et al*. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature, 2008, **454** (7205): 766-770

[82] Lister R, Pelizzola M, Dowen R H, *et al*. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature, 2009, **462** (7271): 315-322

[83] Down T A, Rakyan V K, Turner D J, *et al*. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol, 2008, **26** (7): 779-785

[84] Ruike Y, Imanaka Y, Sato F, *et al*. Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. BMC Genomics, 2010, **11**: 137

[85] Wang H, Maurano M T, Qu H, *et al*. Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res, 2012, **22** (9): 1680-1688

[86] Lee B K, Iyer V R. Genome-wide studies of CCCTC-binding factor (CTCF) and cohesin provide insight into chromatin structure and regulation. J Biol Chem, 2012, **287** (37): 30906-30913

[87] Vaquero-Sedas M I, Luo C, Vega-Palas M A. Analysis of the epigenetic status of telomeres by using ChIP-seq data. Nucleic Acids Res, 2012, **40** (21): e163

[88] Orlando V, Strutt H, Paro R. Analysis of chromatin structure by *in vivo* formaldehyde cross-linking. Methods, 1997, **11** (2): 205-214

[89] Birney E, Stamatoyannopoulos J A, Dutta A, *et al*. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 2007, **447** (7146): 799-816

[90] Dunham I, Kundaje A, Aldred S F, *et al*. An integrated encyclopedia of DNA elements in the human genome. Nature, 2012, **489** (7414): 57-74

[91] Bernstein B E, Stamatoyannopoulos J A, Costello J F, *et al*. The NIH roadmap epigenomics mapping consortium. Nat Biotechnol, 2010, **28** (10): 1045-1048

[92] Edgar R, Domrachev M, Lash A E. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res, 2002, **30** (1): 207-210

[93] Barrett T, Troup D B, Wilhite S E, *et al*. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res, 2009, **37** (Database issue): D885-890

[94] Rhodes D R, Kalyana-Sundaram S, Mahavisno V, *et al*. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia, 2007, **9** (2): 166-180

[95] Rhodes D R, Yu J, Shanker K, *et al*. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia, 2004, **6** (1): 1-6

[96] Landt S G, Marinov G K, Kundaje A, *et al*. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res, 2012, **22** (9): 1813-1831

# ChIP-seq: a New Technique for Genome-wide Profiling of Protein-DNA Interaction*

LIANG Fang[1)], XU Ke[1)], GONG Zhao-Jian[1)], LI Qiao[1)], MA Jian[1, 2)],
XIONG Wei[1, 2)], ZENG Zhao-Yang[1)**], LI Gui-Yuan[1, 2)**]

([1)] *Key Laboratory of Carcinogenesis of Ministry of Health, Key Laboratory of Carcinogenesis and Cancer Invasion of Ministry of Education, Cancer Research Institute, Central South University, Changsha* 410078, China;
[2)] *Hunan Key Laboratory of Nonresolving Inflammation and Cancer, Disease Genome Research Center, Central South University, Changsha* 410013, China)

**Abstract** Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a burgeoning technique which combines Chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to detect protein-DNA binding events, histone modifications, nucleosomes positioning and DNA methylation on a genome-wide scale. Motivated by the tremendous progress in next-generation sequencing (NGS) technology, ChIP-seq offers higher resolution, less noise, and broader coverage than conventional microarray based ChIP-chip. With the decreasing cost of sequencing, ChIP-seq has become an indispensable tool for studying gene regulation and epigenetic mechanisms. In this review, we describe its latest advances, with an emphasis on issues related to data analysis and its application.

**Key words** ChIP-seq, next generation sequencing, gene regulation, epigenetics, data analysis
**DOI**: 10.3724/SP.J.1206.2012.00305

* This work was supported by grants from The National Natural Science Foundation of China (81272298, 30871282, 30871365, 81172189 and 81171930), The Hunan Province Natural Sciences Foundation of China (10JJ7003), The Fok Ying Tong Education Foundation (121036), The Fundamental Research Funds for The Central Universities (2011JQ020), The Mittal Innovative Entrepreneurial Project of Central South University (11MX27) and The Open-End Fund for the Valuable and Precision Instruments of Central South University.
**Corresponding author. Tel: 86-731-84805383
ZENG Zhao-Yang. E-mail: zengzhaoyang@xysm.net
LI Gui-Yuan. E-mail: ligy@xysm.net
Received: June 21, 2012 Accepted: November 28, 2012