计算分子生物(网络版), 2012 年, 第1 卷, 第 60-70 页
Jisuan Fenzi Shengwu (Online), 2012, Vol.1, 60−70
http://cmb.bioPublisher.com

评述与展望
Review and Progress

基于新一代测序数据分析组蛋白修饰的高通量计算方法

Lv Jie [1], Liu Hongbo [1], Wu Qiong [1], Zhang Yan [2]

1 哈尔滨工业大学生命科学与技术学院, 城市水资源与水环境国家重点实验室, 哈尔滨, 150001, 中国
2 哈尔滨医科大学生物信息科学与技术学院, 哈尔滨, 150081, 中国
通讯作者: xmin@ysu.edu; 作者
计算分子生物, 2012 年, 第 1 卷, 第 9 篇 doi: 10.5376/cmb.cn.2012.01.0009
收稿日期：2012 年 11 月 01 日
接受日期：2012 年 11 月 01 日
发表日期：2012 年 11 月 01 日

引用格式：
Lv et al., 2012, High-Throughput Computational Approaches to Analyzing Histone Modification Next-Generation Sequencing Data, Computational Molecular Biology, Vol.2, No.2 1-13 (doi: 10.5376/cmb.2012.02.0002)

摘 要 染色质免疫沉淀测序(ChIP–seq)促进染色体组蛋白尾的化学修饰的系统分析。随着下一代测序的成本的不断下降，全基因组的组蛋白修饰测序在各种表观遗传区研究中成为一种常见的测序方法。然而，现在高效的 ChIP–seq 数据分析方法面临的挑战正是解释组蛋白修饰 ChIP-Seq 数据的主要障碍，这也要求计算方法需要不断改进。我们提供了一个关于研究组蛋白修饰 ChIP-seq 数据可用的计算方法的实际概要。我们展现了关于系统地检测和功能化地定性不同类型的组蛋白修饰 ChIP-Seq 数据计算方法的最新进展，讨论了目前可用于执行短阅读定位，peak calling 下游基因鉴定和基因组可视化任务的软件。我们还展现了可以通过开发具体的组蛋白修饰 ChIP-Seq 数据的算法和方法推断组蛋白修饰对基因表达的调控作用。这种方法将有利于表观遗传调控网络的建设，并提供明确的基于进一步实验测试的生物假说。我们还描述了一些挑战和未来的基于 ChIP-seq 数据对组蛋白修饰的分析的重要方向。我们设想，计算方法的进步将为大规模的组蛋白修饰研究的带来一个更光明的未来。

关键词 下一代测序; 组蛋白修饰; 计算方法; 峰值识别; 染色质免疫沉淀测序

High-Throughput Computational Approaches to Analyzing Histone Modification Next-Generation Sequencing Data

Jie Lv[1], Hongbo Liu[1], Qiong Wu[1], Yan Zhang[2]
1 School of Life Science and Technology, State Key Laboratory of Urban Water Resource and Environment, Harbin Institute of Technology, Harbin, 150001, China
2 College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China

Abstract Chromatin immunoprecipitation followed by sequencing (ChIP–seq) facilitates systematic analysis of chemical modifications of histone tails. As the cost of next-generation sequencing continues to drop, genome-wide histone modification sequencing becomes a common approach in a variety of researches in the epigenetic area. However, challenges of efficient ChIP–seq data analysis are now the main hurdle to interpret the histone modification ChIP-seq data, calling for continued enhancements of computational approaches. Here we provide a pragmatic overview of available computational approaches for the study of histone modification ChIP-seq data. We present the latest advances of computational methods for systematically detecting and functionally characterizing various types of histone modification ChIP-seq data, discuss the software packages currently available for performing tasks from short read mapping, peak calling to downstream genomic characterization and genome-wide visualization. We also present that the regulatory roles of histone modifications upon gene expression can be inferred by developing algorithms and methods specifically for histone modification ChIP-seq data. Such approaches will facilitate the epigenetic regulatory network construction and provide explicit biological hypothesis for further experiment testing. We also describe some challenges and

important directions for histone modification analysis based on ChIP-seq data in future. We envision that the advances of computational approaches will bring more about a bright future for large-scale histone modification studies.

Keywords Next-generation sequencing; Histone modification; Computational approaches; Peak calling; ChIP sequencing

背景

核小体是染色质的基本单位，每个都包括两份核心组蛋白（H3、H4，H2A 和 H2B）并由 147 bp 的 DNA 覆盖。组蛋白是进化上高度保守的蛋白，拥有可接近的高度动态的氨基末端尾巴，也承担了介导组蛋白间相互作用的组蛋白折叠结构域。组蛋白尾巴的 N-末端由过百的不同翻译后修饰，包括甲基化乙酰化和磷酸化来进行广泛修饰(Kouzarides, 2007)。虽然近年来显著的进展表明，甲基化和乙酰化在转录调控中发挥重要的作用，但是到目前为止，大多数这些共价修饰的生物学意义还是不被人们理解。要系统地研究各种组蛋白修饰的全基因组模式，通常使用染色质免疫沉淀(ChIP)来收集 DNA 片段，这些 DNA 片段是通过使用抗体感兴趣的组蛋白修饰从染色质中分离(Collas, 2010)。DNA 片段分离，然后通过杂交到 DNA 微阵列或测序(Gilchrist et al., 2009)。

了解组蛋白修饰的发展和弊端的结构是很有趣的(Kurdistani, 2011; Ikegami et al., 2009; Aoki and Akiyama, 2007)。参考基因组序列和下一代测序平台的可用性要求有效地解释高通量基因组的组蛋白修饰数据的方法(Pepke et al., 2009; Mardis, 2007)。在这篇综述中，我们将描述可以分析根据原则和优势的下一代测序平台产生的组蛋白修饰的数据的计算方法。我们也将说明一些研究 组蛋白修饰在组蛋白修饰 ChIP-Seq 数据具体的的算法和方法的基础上对基因表达的计算推断调控作用。在我们讨论的计算方法之前，我们会首先说明基于 ChIP 的技术的历史

1 分析组蛋白修饰的下一代测序技术

芯片是一种已经存在了一段时间的基本方法(Collas, 2009; 2010)。简单地说，DNA 共价连接到结合蛋白。然后，交联的 DNA 被分解成短片段。然后特定的组蛋白修饰的抗用于分离结合的 DNA。这个方法最突出的问题是，每一次只能研究个别的感兴趣的网站。ChIP 芯片中的部分问题已经解决，用一个早期的方法在全基因组水平上来研究 DNA 结合蛋白。ChIP-seq (Buck and Lieb, 2004; Horak and Snyder, 2002)是一种涉及 DNA 免疫沉淀反应的技术。DNA 免疫沉淀反应是用组蛋白修饰的特异性抗体，随后是 DNA 杂交阵列（芯片）。虽然 ChIP-seq 类似于染色质免疫沉淀的高通量测序 ChIP（ChIP-seq）的名称，但是其定位精度小于 ChIP-seq，且其量化表达的动态范围明显小于 ChIP-seq (Liu et al., 2010)。此外，所有的杂交方法，包括 ChIP 芯片的掩模重复序列。虽然对基因组组蛋白修饰的研究并不有效，但是对特定的基因或位点的有 ChIP 芯片自定义阵列在许多实验条件下的研究仍然是有用的。

后来，基于 ChIP 的高通量方法称为 ChIP-SAGE (Schones et al., 2011; Schones and Zhao, 2008)。总之，ChIP 之后进行了 SAGE（基因表达序列分析）。ChIP-SAGE 更接近于 ChIP-seq，21 bp 的短序列标签从测序文库中提取，并且定位到参考基因组。定位在一个基因组区域的标签的数量反映了该区域的组蛋白修饰水平。虽然没有直接比较这两种技术，但是由于 ChIP-SAGE 这项技术没有涉及到探针杂交的问题，其结果往往比 ChIP-seq 更定量。然而，因为技术的明显局限性，以及更划算更敏感的替代技术 ChIP seq 的到来，很少有研究使用 ChIP-SAGE (Park, 2009)。

今天，ChIP-seq 通常是研究全基因组的组蛋白修饰模式的首选方法，使数千万 DNA 在一个可接受的时间范围内被组蛋白修饰定向从而测序。ChIP-seq 具有低错误率、较高的灵敏度和特异性，同时每个库对于研究者来说保持可接受的成本(Johnson et al., 2007)。与 ChIP-chip 不同的是，ChIP-seq 完全消除了交叉杂交的潜在错误。ChIP-seq 主要的服务提供商是 Illumina 公司，使用一个高通量的大规模平行识别测序技术-像 Solexa 开发的技术(Cuddapah et al., 2009; Whiteford et al., 2009)。简而言之，ChIP 的 DNA 连接到适配器随后通过有限的扩增产生~ 200 ng 的 DNA，然后通过在固体表面上的杂交结合。30 ~ 6000 万 DNA 模板的短序列（25 ~ 50 bp）利用合成测序技术从序列末端进行测序，合成测序技术是改良的 Sanger 测序法。

在 CD4 $^+$ T 细胞中，ChIP-seq 首次运用来研究基因组的组蛋白修饰(Barski et al., 2007)。从概念上讲，定位到一个基因组位点的测序的阅读片段的数量是与其组蛋白修饰水平成正比的。ChIP-seq 的两个重要优点包括 PCR 扩增需要较少和探针杂交的独立性，使得它可能在不同的基因组区域更定量且更具有比较性(Johnson et al., 2007)。基于组蛋白修饰分析的下一代测序的一个格外关注点是如何深入到样本的每个库(Liu et al., 2010)。虽然超过饱和度的大规模测序提供了一个全覆盖及提高兴趣组蛋白修饰的信心，饱和度意味着进一步的测序将无法发现以上的背景的其他区域，但是低于或达到饱和的测序可能足以使测序成本可以接受而不明显降低覆盖。ChIP-chip，ChIP-SAGE 和 ChIP-Seq 技术参考表 1。

2 ChIP-seq 的组蛋白修饰的大数据资源

DNA 元素百科全书（ENCODE）组合进行了数百个 ChIP–seq 实验。ENCODE 是一个有价值的数据源并且提供有效的测序协议(Birney et al., 2007)。考虑到在 ENCODE 中正在检测和待检测的细胞类型的多样性，ChIP-seq 在从 ENCODE 的数据中获得的组织特异性和/或细胞类型特异性组蛋白修饰模式的各种基因组元素中是有用的。然而，应该指出的是，一个 ChIP 实验的成功很大程度上依赖于高度特异性的抗体结合的组蛋白修饰(Liu et al., 2010)。即使在相同的抗体独立制备的阶段，抗体的质量也不同，这由抗体在 ENCODE 的最近的评估和模式生物的 ENCODE (mod-ENCODE)项目中可以体现(Egelhofer et al., 2011)。在这项研究中，25%在特异性测试中失败和 20%在免疫沉淀实验中失败。因此，需要谨慎解释组蛋白修饰 ChIP-Seq 数据，特别是在比较不同的组蛋白修饰模式时。

modENCODE 项目启动为模式生物提供了一个全面的基因组功能元件百科全书，如线虫(*C. elegans*)和果蝇(*D. melanogaster*) (Washington et al., 2011; Muers, 2011)。数据内容的范围从基因结构、mRNA 和 ncRNA 基因表达谱到转录因子结合位点，组蛋白修饰等等。所有的数据都是公开的，可供下载和发布使用。

在国家生物技术信息中心（NCBI）的表观基因组资源（www.ncbi.nlm.nih.gov/表观基因组）是全基因组的组蛋白修饰和其他表观遗传修饰的数据集的综合公共资源(Fingerman et al., 2011)。该数据是基于基因表达文库(Gene Expression Omnibus, GEO)的表观遗传修饰数据(Barrett and Edgar, 2006)。该资源是用户友好型的，并在持续更新中。表观基因组资源与其他 NCBI 数据库高度集成（Baxevanis, 2008）从而便于使用，数据库包括基因数据库(Maglott et al., 2011)和 PubMed (McEntyre and Lipman, 2001)。2011 年有超过 1100 个数据轨道涵盖了五个研究的物种。

表 1 ChIP-chip, ChIP-SAGE and ChIP-Seq 的比较

Table 1 Comparison of ChIP-chip, ChIP-SAGE and ChIP-Seq

	ChIP-chip	ChIP-SAGE	Chip-seq
定量 Quantification	定量有限，依赖于杂交效率 Limited quantitative and depends on the hybridization efficiency	定量的 Quantitative	定量的 Quantitative
分辨率 Resolution	取决于 ChIP 染色质片段的大小 Depends on size of the chromatin fragments for ChIP	取决于限制性内切酶位点 Depends on restriction enzyme sites	取决于染色质片段的大小和测序深度 Depends on the size of the chromatin fragments and sequencing depth
费用 Cost	全基因组 tiling 阵列费用高 High for whole-genome tiling arrays	比 ChIP–Seq 贵 More expensive than ChIP–Seq	便宜 Low
限制 Limitation	仅在微阵列上的预先选择的基因组区域 Only pre-selected genomic regions on a microarray	限制性内切酶的识别位点 Recognition sites for the restriction enzyme	只有非重复区域 Only non-repetitive regions

NIH 路标表观基因定位共同体（http://www.roadmapepigenomics.org/）是针对催化基础生物学和疾病研究的另一个人类表观基因组数据的公共资源(Bernstein et al., 2010)。在不同的细胞类型中的共同体定位组蛋白修饰和其他的染色质修饰，可能代表了人类疾病的组织和器官系统的正常。ChIP-seq 分析组蛋白修饰，随后是严格的特异性测试以保证抗体的特异性。此外，全面地进行描述和比较共同的细胞来源，从而确保不同的数据收集中心之间的一致性。

3 下一代组蛋白修饰数据分析的工具

新一代测序平台产生的组蛋白修饰 ChIP-seq 数据的分析仍然面临着挑战，部分是因为许多新一代测序平台的快速发展。下一代测序产生的组蛋白修饰数据分析可分为两个部分。

3.1 下一代组蛋白修饰数据的比对工具

从下一代测序平台产生的数据是碱基序列（Illumina 基因组分析仪，454 FLX）或颜色空间碱基过渡(SOLiD)以及相关的质量分数。

组蛋白修饰的 ChIP-seq 数据分析的第一步是使从公共资源上下载的或从服务提供商上获得的 ChIP-Seq 数据的 read 比对到参考基因组组装。分析的结果将是一个由参考基因组对齐序列和链上的基因组坐标组成的数据集。许多新一代测序比对工具已被用于定位参考基因组测序的 read (Pepke et al., 2009; Kim et al., 2011; Schones et al., 2011; Schones and Zhao, 2008; Hirst and Marra, 2010)。比对工具大多使用的“种子和扩展”的算法，包括 read 在内的子字符串比对到一个哈希表或最近一个参考基因组的 Burrows–Wheeler 转换生成的后缀数组。在找到匹配之前，read 在基因组上会“延长”至最大 read 长度。SAM/BAM 文件格式是比对工具可以输出的一种标准的文件格式。

虽然这些比对工具在速度和精度的差异很微小，但是这些差异在全面定位率和精度上有显著影响(Wilbanks and Facciotti, 2010)。最终用户可以根据从其他研究者或参考相关论文上的建议选择其中的一个比对工具。对齐的文件可以直接在基因组浏览器上看到，也可以通过调用 peak calling 来进行进一步处理。我们列举了许多常见的处理 ChIP-seq 数据的短 read 比对工具，如表 2 所示。

3.2 下一代组蛋白修饰数据的 Peak calling 工具

Peak calling 将原始的对齐 read 转化为的显著标签富集的峰-区域。峰被认为是与组蛋白修饰的入住率有关，这个入住率可由 Peak calling 工具建模（最近的一次评述(Pepke et al., 2009)）。一些算法简单地合并定位标签，而其他则使用链特定的信息来更精确地找到峰。一些 peak calling 工具需要控制测序 ChIP-seq 库而其他在不控制时仍然可以工作。假设有几个已知的 ChIP-seq 的测序偏好源，没有控制库的 peak calling 结果是不可靠的。基于不同的芯片的类库和控件库，用 P 值或错误发现率（FDR）量化定位峰的信心，即使不同的 peak-calling 算法在细节上相差很多。一般情况下，这样的工具可以分为两个部分，即使用的两个主要策略。第一个策略主要搜索组蛋白修饰标记，试图找到他们的基因组分布，如 H3K4me3 或 h3k27ac，以及试图建立 峰”标签分布的模型。虽然有大量的 peak calling 软件包，但不是所有的软件包都可以满足调用丰富的组蛋白修饰结构域的第一个策略。因此第二个策略更适合于有广泛的分布模式的组蛋白修饰，如 H3K36me3，这可以通过由组蛋白修饰设计的 peak calling 软件检测。表 2 为适用于组蛋白修饰 ChIP-seq 数据的公开 peak-calling 算法和几个适用于其他地方的相关评述(Pepke et al., 2009; Wilbanks and Facciotti, 2010; Szalkowski and Schmid, 2011)。未在表中列出的其他软件包可能涉及包含 peak-calling 功能的商业软件包。

Zang 等（2009）分析了在基因组背景模型下的随机 read 的分数分布，并采用他们的理论来确定空间集群，这是作为一个软件实现的，不可能偶然出现。Rashid 等人（2011）发展了 ZINBA（零膨胀负二项算法）来识别丰富的 ChIP-seq 的基因组区域，这些区域建模和说明了共有背景或实验信号的因素，如 G / C 含量。Xu 等人（2010)提出了一种线性信号噪声模型，其中引入了噪声率。他们开发了使用控件库估计噪声率迭代算法，衍生了图书馆交换策略估计 FDR。该算法的软件实现，命名为 CCAT（基于控制的 ChIP-seq 分析工具）。H3K4me3 和 H3K36me3 应用数据表明，CCAT 比以前的方法预测明显更多的 ChIP 丰富的位点。张等（2008)提出一种根据 Perl 的模型依据的分析 ChIP-seq 数据分析

软件，MACS 来分析 ChIP-seq 数据。MACS 有一个无模型参数来为组蛋白修饰如H3K36me3广泛分布模式提供支持。Boyle 等（2008）提出了 F-seq 检测开放的染色质区域，也可用于检测组蛋白修饰 ChIP-seq 数据。这些算法的重要参数如表 3 所示。

这种 peak calling 工具的多样性是由于测序技术的快速进步和多样性。从事基于 ChIP-seq 研究组蛋白修饰的研究人员需要判断哪个将是最适合自己数据的工具。然而在不久的将来，预计这样的工具在表观遗传学研究中会变的规范化。

4 用于下一代组蛋白修饰数据的差分组蛋白修饰区鉴定工具

差异组蛋白修饰位点（DHMSs）在研究不同的细胞类型，阶段或环境反应的组蛋白修饰调节的动态性质上是很重要的。虽然 ChIP-seq 因为比较长的 read 长度很少出错，但是一些程序，如样品制备、标签扩增和序列比对，在比较不同的 ChIP-seq 数据来提取真正的生物相关信号上存在着一些挑战（Taslim et al., 2009）。虽然我们希望所有样本的差异反映生物条件，但是依旧存在更多不能被模拟的因素，从而可能使结果产生偏差。因此，十分需要有效的计算和统计的方法来可靠地检测不同区域的不同 ChIP-seq 数据。

此前，Xu 等（2008）提出了一种 ChIPDiff 的方法。这种方法用于从 ChIP-seq 识别的组蛋白修饰丰富区域的基因组比较。他们采用了一个隐藏的马尔可夫模型（HMM）来推断每个基因组位置的组蛋白修饰的变化。黄等（2011）开发了一个有效的框架，来确定全基因组差异的组蛋白修饰区。他们开发了一个软件工具 EpiCenter，可以有效地执行相关的数据处理。此外，Taslim 等人（2009）运用一种基于局部加权回归两步非线性归一化（LOESS）的方法，利用正常指数的混合模型在多个样本和模型比较 ChIP-seq 数据的差异。

5 下一代组蛋白修饰数据的可视化工具

许多网络和独立的工具可用于对齐的表观基因组数据包括组蛋白修饰 ChIP-seq 数据集在内的可视化。使用最广泛的工具是由加利福尼亚大学圣克鲁斯拥有的基因组浏览器（UCSC）。本地安装 UCSC 基因组浏览器来进行未发表的 ChIP-seq 数据可视化，受到许多研究者的欢迎。作为基因组注释背景下的直线轨道，UCSC 是用于全基因组数据可视化的早期工具，对之后的相关工具都有所影响。即使它在人工基因组图上非常强大，但是很难同时可视化很多 ChIP-Seq 数据。为支持 BAM 格式下很多大的 ChIP-seq 数据，之后开发了几个基因组浏览器，如 GBrowse，GenomeView (Abeel et al., 2012)，JBrowse (Skinner et al., 2009)和 ABrowse。此外，单机工具如 IGV 和 IGB 在查看非常大的对齐的 ChIP-seq 数据也是受欢迎的工具。任何不能建立基于浏览器的工具的人也可以使用这样的单机工具。有用的可视化工具如表 4 所示。

表 2 可用于组蛋白修饰 ChIP-Seq 比对的短 read 比对工具的集合

Table 2 A subset of short read aligners available for histone modification ChIP-seq alignment

软件工具 Software tool	网址 Web address
种子和延伸战略 Seed and extend strategy	
MAQ	http://maq.sourceforge.net/
SOAP	http://soap.genomics.org.cn/index.html
SHRiMP	http://compbio.cs.toronto.edu/shrimp/
ZOOM	http://www.bioinfor.com/zoom
BFAST	http://sourceforge.net/projects/bfast/
使用哈希表或 Burrows–Wheeler 变换生成的最近一个后缀数组 Using a hash table or more recently a suffix array generated from Burrows–Wheeler transform	
BOWTIE	http://bowtie-bio.sourceforge.net
BWA	http://bio-bwa.sourceforge.net
SOAP2	http://soap.genomics.org.cn/index.html

表 3 每个 peak calling 算法的重要参数

Table 3 Important parameters for each peak calling algorithm

算法 Algorithm	重要参数 Important parameters
CCAT	最小得分: 归一化差的最小值 Minimum score: minimum score of normalized difference 最小计数: 峰值读取计数的最小值 Minimum count: minimum number of read counts at the peak 移动步: 窗口滑动的一步 Moving Step: step of window sliding SlidingWinSize: 滑动窗口的大小 SlidingWinSize: size of sliding window 自助通：在引导过程中传球次数 Bootstrap pass: number of passes in the bootstrapping process
MACS	NoLambda: 如果属实，MACS 将使用固定的背景 λ 作为每个峰区本地 λ NoLambda: if True, MACS will use fixed background lambda as local lambda for every peak region NoModel: 是否建立移动模型 NoModel: whether or not to build the shifting model MFold: 在可信度高的富集率 MFOLD 范围内的区域而不是背景区域进行建模 MFold: regions within MFOLD range of high-confidence enrichment ratio against background to build model PValue: P 值的截止峰值检测 PValue: p-value cutoff for peak detection
SICER	WindowSize: 扫描基因组宽度的窗口大小 WindowSize: size of the windows to scan the genome width GapSize: islands 间的碱基对的允许间隙 GapSize: allowed gap in base pairs between islands FDR: 错误发现率控制意义 FDR: false discovery rate controlling significance
ZINBA	Select model: 指定选择模型 = 假跳过模型选择过程完全可以节省大量时间 Specifying select model = FALSE skips the model selection process altogether and may save a significant amount of time 扩展：平均片段库长度（选择大小） Extension: average fragment library length (size selected) Win Size: 选择一个更大的窗口大小增加的分析速度，但降低分辨率和灵敏度来检测丰度 Win Size: Selecting a larger window size increases speed of analysis but decreases resolution and sensitivity to detect enrichment offset: 较小的非零偏移距离的增加的敏感性也增加了计算 builder offset: Smaller non-zero offset distances increase sensitivity but also increase computational builden FDR: FDR =真正的指定模型使用 FDR 阈值而不是后验概率。这通常会导致更自由的 peak calls。如果为假，然后利用后验概率通过 1-阈值接近峰。 FDR: FDR = TRUE specifies the model to use the FDR threshold rather than posterior probabilities. This typically results in more liberal peak calls. If false, then uses posterior probability to threshold peaks using 1-threshold.
F-seq	特征：特征长度的长度 Feature Length: feature length 阈值：标准偏差 Threshold: standard deviations

表 4 使组蛋白修饰 ChIP-seq 数据更加可视化的工具列表

Table 4 The list of more visualization tools for histone modification ChIP-seq data

网络服务器 / 软件 Web server / Software	网址/下载链接 Website / Download Link
UCSC Genome Browser	http://genome.ucsc.edu/
GBrowse	http://www.gbrowse.org/index.html
Ensembl	http://asia.ensembl.org/index.html
GenomeView	http://genomeview.org/
JBrowse	http://jbrowse.org/
ABrowse	http://www.abrowse.org/
Artemis	http://www.sanger.ac.uk/resources/software/artemis/
Avadis Genome Browser	http://www.avadis-ngs.com/features/genome_browser
IGV	http://www.broadinstitute.org/igv/
IGB	http://bioviz.org/igb/

6 下一代组蛋白修饰数据的下游分析工具

Peak calling 通常是通过特定的组蛋白修饰由下游分析来注释和定性丰富的区域。通常情况下，还需要基因组注释来发现与丰富的组蛋白修饰区的潜在的有趣的关联。注释可以在许多公共库，例如 UCSC，Ensembl (Flicek et al., 2012)和许多零散的网站上进行。注释丰富的组蛋白修饰区，有利于在各种常见的基因组注释，如染色体和基因的情况下显示这些区域的基因组景观，来寻找有趣的生物协会。例如，通过比较有增强子的 H3K27ac 峰位置甚至通过更先进的生物信息学分析发现与增强子的可能关系。许多工具可以做到这一点，如 Galaxy（Goecks et al., 2010）和 Bedtools (Quinlan and Hall, 2010)。一些可视化工具如 CEAS（Shin et al., 2009）和 ChIPseeqer (Giannopoulou and Elemento, 2011)可以支持显示平均组蛋白修饰富集信号内部或附近的基因组元件如增强子和基因起始而事先不需要 peak calling，从而帮助生物学家更好地理解组蛋白修饰。

此外，它在已知的基因组元素也有助于探索丰富的组蛋白修饰峰，以获得没有任何特定先验知识的特定组蛋白修饰的潜在监管功能和本地化偏好的全局视图。特别是，研究基因结构特定的组蛋白修饰的优先目标是很有趣的，如外显子，启动子和远端上游区域。这两种互补的方法在染色质生物学研究中都是常见的。研究人员可以选择基于他们自己的生物假说的一个或两个方法。

7 调控组蛋白修饰基因表达

积累的组蛋白修饰 ChIP-Seq 数据使研究人员开展全球染色质的知识挖掘，这在表观遗传学领域是很有趣的。从计算的角度来看，它可以通过提出不同算法利用这些数据来分析组蛋白修饰的之间的相互作用。Yu 等（2011）进行了一项开创性的研究，在通过已提出的贝叶斯网络与基因表达的关联基础上，推断组蛋白修饰和其他转录调控因子之间的组合关系。他们建立了染色质调控网络并且根据一套在人类 CD_4^+ T 细胞上的 23 个 ChIP-seq 数据推断出许多染色质相互作用关系，这套数据在那时是最全面的组蛋白修饰的数据。许多进一步的研究表明组蛋白修饰、基因组元素和基因表达之间的相关性更为复杂。例如，Karlić 等人（2010）使用线性回归模型，以进一步探讨类似的问题，发现有高的 GC 含量和低的 GC 含量的启动子的基因受不同的组蛋白修饰。Costa 等（2012）在 H3K4me3 和 H3K27me3 上进一步应用线性回归模型的混合，发现与转录因子结合相比，他们对基因表达的预测性更强。do Rego 等人（2012）采用稀疏线性回归混合模型对基因表达进行建模，并进行转录因子的有效特征选择。通过使用该模型，作者确定了与血液发展相关的组蛋白修饰和转录因子(do Rego et al., 2012)。有趣的是，通过不同的计算模型与输入的标签数量转换，这些研究模拟了基因表达。然而，其他的方法使用其他派生的功能，如峰的形状和位置，以及信号频率模拟了基因表达。Beck 等人（2012）提出了一种新的策略，利用模式和信号位置量化了 ChIP-seq 的轮廓。UCAR 等人（2011）介绍了子空

间聚类算法，详尽地确定组合修饰模式也确定不同类别的功能的DNA分子的组合的组蛋白修饰特征。总之，利用不同的细胞和发育阶段的组蛋白修饰ChIP-Seq 谱的算法和模型有助于理解染色质修饰网络是如何调节基因的表达。

8 结论和观点

组蛋白修饰的重要性促使 ChIP-seq 数据的持续累积，从而识别和定性组蛋白修饰和组合调控在基因表达时的作用。ChIP-seq 实际上是用于识别基因组蛋白修饰景观的标准。然而，从下一代的ChIP-seq 数据可靠地获得生物学知识中，技术和计算的局限性还是障碍。在这里，我们重点关注从原read 到下游分析，组蛋白修饰 ChIP-Seq 数据处理的计算方面。

在过去的五年中，下一代测序将表观遗传学研究带到了一个快速发展的时代。组蛋白修饰的研究从方法论到生物学解释发生了进化。大量的组蛋白修饰 ChIP-Seq 数据可以从公共数据库中下载，如NCBI 的 GEO。因为第三代测序平台将很快被商业化，所以这一发展预计将继续。然而，第二代测序平台仍将盛行很长一段时间。因此，掌握第二代测序数据处理的基本概念和方法是十分必要的。

由于不成熟的下一代测序技术的持续发展，因此很明显有许多不同的工具来执行相同的任务。在不久的将来，预计处理方法或指标将大大标准化。实现这一目标的关键是与在基因组研究中的基础质量得分类似，即为富集的组蛋白修饰数据研究可用的计算指标或开发一个新的指标。如果开发后，一个共同的指标将在不同的 ChIP-seq 实验中启用有意义的比较，在未来，考虑到这些丰富的数据集是关键。

虽然组蛋白修饰的数据正在以前所未有的速度积累，来处理和整合大量的数据的更有效的计算工具的发展已经有一点落后。差异组蛋白修饰识别方法不仅可以比较不同的生物样品，而且也有利于揭示疾病相关的组蛋白修饰。事实上，最近的研究已经证明了组蛋白修饰标记物在诊断和治疗中的能力(Zhao and Zhang, 2011)。为了能够识别在各种疾病中的改变的组蛋白修饰模式，研究人员需要继续开发和比较基于下一代测序数据相关的组蛋白修饰差异鉴定相关的更强大的工具。那些基于网络和独立，具有更好的显示效果以及更支持大量的数据轨道的工具将更受欢迎。这将有利于在发展中的阶段，疾病类型等，不同类型的组蛋白修饰数据比较。我们设想，计算方法的进步将为大规模的组蛋白修饰研究的带来一个更光明的未来。

致谢

感谢中国国家自然科学基金资助。本研究由中国国家自然科学基金（31171383，31371334，31371478）和中央高校基本科研基金（HIT.NSRIF.2010027）和黑龙江省自然科学基金（C201217）共同资助。

参考文献

Abeel T., Van Parys T., Saeys Y., Galagan J., and Van De Peer Y., 2012, GenomeView: a next-generation genome browser, Nucleic Acids Res, 40: e12

Aoki F., and Akiyama T., 2007, [Involvement of histone modification and histone variants replacement in genome reprogramming during oogenesis and preimplantation development], Tanpakushitsu Kakusan Koso, 52: 2170-2176

Barrett T., and Edgar R., 2006, Gene expression omnibus: microarray data storage, submission, retrieval, and analysis, Methods Enzymol, 411: 352-369

Barski A., Cuddapah S., Cui K., Roh T.Y., Schones D.E., Wang Z., Wei G., Chepelev I., and Zhao K., 2007, High-resolution profiling of histone methylations in the human genome, Cell, 129: 823-837

Baxevanis A.D., 2008, Searching NCBI databases using Entrez, Curr Protoc Bioinformatics, Chapter 1: Unit 1 3

Beck D., Brandl M.B., Boelen L., Unnikrishnan A., Pimanda J.E., and Wong J.W., 2012, Signal analysis for genome-wide maps of histone modifications measured by ChIP-seq, Bioinformatics, 28: 1062-1069

Bernstein B.E., Stamatoyannopoulos J.A., Costello J.F., Ren B., Milosavljevic A., Meissner A., Kellis M., Marra M.A., Beaudet A.L., Ecker J.R., Farnham P.J., Hirst M., Lander E.S., Mikkelsen T.S., and Thomson J.A., 2010, The NIH Roadmap Epigenomics Mapping Consortium, Nat Biotechnol, 28: 1045-1048

Birney E., Stamatoyannopoulos J.A., Dutta A., Guigo R., Gingeras T.R., Margulies E.H., Weng Z., Snyder M.,

Dermitzakis E.T., Thurman R.E., Kuehn M.S., Taylor C.M., Neph S., Koch C.M., Asthana S., Malhotra A., Adzhubei I., Greenbaum J.A., Andrews R.M., Flicek P., Boyle P.J., Cao H., Carter N.P., Clelland G.K., Davis S., Day N., Dhami P., Dillon S.C., Dorschner M.O., Fiegler H., Giresi P.G., Goldy J., Hawrylycz M., Haydock A., Humbert R., James K.D., Johnson B.E., Johnson E.M., Frum T.T., Rosenzweig E.R., Karnani N., Lee K., Lefebvre G.C., Navas P.A., Neri F., Parker S.C., Sabo P.J., Sandstrom R., Shafer A., Vetrie D., Weaver M., Wilcox S., Yu M., Collins F.S., Dekker J., Lieb J.D., Tullius T.D., Crawford G.E., Sunyaev S., Noble W.S., Dunham I., Denoeud F., Reymond A., Kapranov P., Rozowsky J., Zheng D., Castelo R., Frankish A., Harrow J., Ghosh S., Sandelin A., Hofacker I.L., Baertsch R., Keefe D., Dike S., Cheng J., Hirsch H.A., Sekinger E.A., Lagarde J., Abril J.F., Shahab A., Flamm C., Fried C., Hackermuller J., Hertel J., Lindemeyer M., Missal K., Tanzer A., Washietl S., Korbel J., Emanuelsson O., Pedersen J.S., Holroyd N., Taylor R., Swarbreck D., Matthews N., Dickson M.C., Thomas D.J., Weirauch M.T., Gilbert J., et al., 2007, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, Nature, 447: 799-816

Boyle A.P., Guinney J., Crawford G.E., and Furey T.S., 2008, F-Seq: a feature density estimator for high-throughput sequence tags, Bioinformatics, 24: 2537-2538

Buck M.J., and Lieb J.D., 2004, ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments, Genomics, 83: 349-360

Collas P., 2009, The state-of-the-art of chromatin immunoprecipitation, Methods Mol Biol, 567: 1-25

Collas P., 2010, The current state of chromatin immunoprecipitation, Mol Biotechnol, 45: 87-100

Cuddapah S., Barski A., Cui K., Schones D.E., Wang Z., Wei G., and Zhao K., 2009, Native chromatin preparation and Illumina/Solexa library construction, Cold Spring Harb Protoc, 2009: pdb prot5237

Do Rego T.G., Roider H.G., De Carvalho F.A., and Costa I.G., 2012, Inferring epigenetic and transcriptional regulation during blood cell development with a mixture of sparse linear models, Bioinformatics, 28: 2297-2303

Dreszer T.R., Karolchik D., Zweig A.S., Hinrichs A.S., Raney B.J., Kuhn R.M., Meyer L.R., Wong M., Sloan C.A., Rosenbloom K.R., Roe G., Rhead B., Pohl A., Malladi V.S., Li C.H., Learned K., Kirkup V., Hsu F., Harte R.A., Guruvadoo L., Goldman M., Giardine B.M., Fujita P.A., Diekhans M., Cline M.S., Clawson H., Barber G.P., Haussler D., and James Kent W., 2012, The UCSC Genome Browser database: extensions and updates 2011, Nucleic Acids Res, 40: D918-923

Egelhofer T.A., Minoda A., Klugman S., Lee K., Kolasinska-Zwierz P., Alekseyenko A.A., Cheung M.S., Day D.S., Gadel S., Gorchakov A.A., Gu T., Kharchenko P.V., Kuan S., Latorre I., Linder-Basso D., Luu Y., Ngo Q., Perry M., Rechtsteiner A., Riddle N.C., Schwartz Y.B., Shanower G.A., Vielle A., Ahringer J., Elgin S.C., Kuroda M.I., Pirrotta V., Ren B., Strome S., Park P.J., Karpen G.H., Hawkins R.D., and Lieb J.D., 2011, An assessment of histone-modification antibody quality, Nat Struct Mol Biol, 18: 91-93

Fingerman I.M., Mcdaniel L., Zhang X., Ratzat W., Hassan T., Jiang Z., Cohen R.F., and Schuler G.D., 2011, NCBI Epigenomics: a new public resource for exploring epigenomic data sets, Nucleic Acids Res, 39: D908-912

Flicek P., Amode M.R., Barrell D., Beal K., Brent S., Carvalho-Silva D., Clapham P., Coates G., Fairley S., Fitzgerald S., Gil L., Gordon L., Hendrix M., Hourlier T., Johnson N., Kahari A.K., Keefe D., Keenan S., Kinsella R., Komorowska M., Koscielny G., Kulesha E., Larsson P., Longden I., Mclaren W., Muffato M., Overduin B., Pignatelli M., Pritchard B., Riat H.S., Ritchie G.R., Ruffier M., Schuster M., Sobral D., Tang Y.A., Taylor K., Trevanion S., Vandrovcova J., White S., Wilson M., Wilder S.P., Aken B.L., Birney E., Cunningham F., Dunham I., Durbin R., Fernandez-Suarez X.M., Harrow J., Herrero J., Hubbard T.J., Parker A., Proctor G., Spudich G., Vogel J., Yates A., Zadissa A., and Searle S.M., 2012, Ensembl 2012, Nucleic Acids Res, 40: D84-90

Giannopoulou E.G., and Elemento O., 2011, An integrated ChIP-seq analysis platform with customizable workflows, BMC Bioinformatics, 12: 277

Gilchrist D.A., Fargo D.C., and Adelman K., 2009, Using ChIP-chip and ChIP-seq to study the regulation of gene expression: genome-wide localization studies reveal widespread regulation of transcription elongation,

Methods, 48: 398-408

Goecks J., Nekrutenko A., and Taylor J., 2010, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, Genome Biol, 11: R86

Hirst M., and Marra M.A., 2010, Next generation sequencing based approaches to epigenomics, Brief Funct Genomics, 9: 455-465

Horak C.E., and Snyder M., 2002, ChIP-chip: a genomic approach for identifying transcription factor binding sites, Methods Enzymol, 350: 469-483

Huang W., Umbach D.M., Vincent Jordan N., Abell A.N., Johnson G.L., and Li L., 2011, Efficiently identifying genome-wide changes with next-generation sequencing data, Nucleic Acids Res, 39: e130

Ikegami K., Ohgane J., Tanaka S., Yagi S., and Shiota K., 2009, Interplay between DNA methylation, histone modification and chromatin remodeling in stem cells and during development, Int J Dev Biol, 53: 203-214

Johnson D.S., Mortazavi A., Myers R.M., and Wold B., 2007, Genome-wide mapping of in vivo protein-DNA interactions, Science, 316: 1497-1502

Karlic R., Chung H.R., Lasserre J., Vlahovicek K., and Vingron M., 2010, Histone modification levels are predictive for gene expression, Proc Natl Acad Sci U S A, 107: 2926-2931

Kim H., Kim J., Selby H., Gao D., Tong T., Phang T.L., and Tan A.C., 2011, A short survey of computational analysis methods in analysing ChIP-seq data, Hum Genomics, 5: 117-123

Kouzarides T., 2007, Chromatin modifications and their function, Cell, 128: 693-705

Kurdistani S.K., 2011, Histone modifications in cancer biology and prognosis, Prog Drug Res, 67: 91-106

Liu E.T., Pott S., and Huss M., 2010, Q&A: ChIP-seq technologies and the study of gene regulation, BMC Biol, 8: 56

Maglott D., Ostell J., Pruitt K.D., and Tatusova T., 2011, Entrez Gene: gene-centered information at NCBI, Nucleic Acids Res, 39: D52-57

Mardis E.R., 2007, ChIP-seq: welcome to the new frontier, Nat Methods, 4: 613-614

Mcentyre J., and Lipman D., 2001, PubMed: bridging the information gap, CMAJ, 164: 1317-1319

Mcghee J.D., and Felsenfeld G., 1980, Nucleosome structure, Annu Rev Biochem, 49: 1115-1156

Muers M., 2011, Functional genomics: the modENCODE guide to the genome, Nat Rev Genet, 12: 80

Park P.J., 2009, ChIP-seq: advantages and challenges of a maturing technology, Nat Rev Genet, 10: 669-680

Pepke S., Wold B., and Mortazavi A., 2009, Computation for ChIP-seq and RNA-seq studies, Nat Methods, 6: S22-32

Quinlan A.R., and Hall I.M., 2010, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics, 26: 841-842

Rashid N.U., Giresi P.G., Ibrahim J.G., Sun W., and Lieb J.D., 2011, ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions, Genome Biol, 12: R67

Schones D.E., Cui K., and Cuddapah S., 2011, Genome-wide approaches to studying yeast chromatin modifications, Methods Mol Biol, 759: 61-71

Schones D.E., and Zhao K., 2008, Genome-wide approaches to studying chromatin modifications, Nat Rev Genet, 9: 179-191

Shin H., Liu T., Manrai A.K., and Liu X.S., 2009, CEAS: cis-regulatory element annotation system, Bioinformatics, 25: 2605-2606

Skinner M.E., Uzilov A.V., Stein L.D., Mungall C.J., and Holmes I.H., 2009, JBrowse: a next-generation genome browser, Genome Res, 19: 1630-1638

Szalkowski A.M., and Schmid C.D., 2011, Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts, Brief Bioinform, 12: 626-633

Taslim C., Wu J., Yan P., Singer G., Parvin J., Huang T., Lin S., and Huang K., 2009, Comparative study on ChIP-seq data: normalization and binding pattern characterization, Bioinformatics, 25: 2334-2340

Ucar D., Hu Q., and Tan K., 2011, Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering, Nucleic Acids Res, 39: 4063-4075

Washington N.L., Stinson E.O., Perry M.D., Ruzanov P., Contrino S., Smith R., Zha Z., Lyne R., Carr A., Lloyd P., Kephart E., Mckay S.J., Micklem G., Stein L.D., and Lewis S.E., 2011, The modENCODE Data Coordination Center: lessons in harvesting comprehensive experimental details, Database (Oxford), 2011: bar023

Whiteford N., Skelly T., Curtis C., Ritchie M.E., Lohr A., Zaranek A.W., Abnizova I., and Brown C., 2009, Swift: primary data analysis for the Illumina Solexa sequencing platform, Bioinformatics, 25: 2194-2199

Wilbanks E.G., and Facciotti M.T., 2010, Evaluation of algorithm performance in ChIP-seq peak detection, PLoS One, 5: e11471

Wood A.C., Rijsdijk F., Asherson P., and Kuntsi J., 2011, Inferring Causation from Cross-Sectional Data: Examination of the Causal Relationship between Hyperactivity-Impulsivity and Novelty Seeking, Front Genet, 2: 6

Xu H., Handoko L., Wei X., Ye C., Sheng J., Wei C.L., Lin F., and Sung W.K., 2010, A signal-noise model for significance analysis of ChIP-seq with negative control, Bioinformatics, 26: 1199-1204

Xu H., Wei C.L., Lin F., and Sung W.K., 2008, An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data, Bioinformatics, 24: 2344-2349

Zang C., Schones D.E., Zeng C., Cui K., Zhao K., and Peng W., 2009, A clustering approach for identification of enriched domains from histone modification ChIP-Seq data, Bioinformatics, 25: 1952-1958

Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., Bernstein B.E., Nusbaum C., Myers R.M., Brown M., Li W., and Liu X.S., 2008, Model-based analysis of ChIP-Seq (MACS), Genome Biol, 9: R137

Zhao Q., and Zhang Y., 2011, Epigenome sequencing comes of age in development, differentiation and disease mechanism research, Epigenomics, 3: 207-220