遗传 HEREDITAS (Beijing) 2012 年 6 月, 34(6): 773－783
ISSN 0253-9772 www.chinagene.cn

DOI: 10.3724/SP.J.1005.2012.00773

下一代测序中 ChIP-seq 数据的处理与分析

高山[1], 张宁[1], 李勃[2], 徐硕[3], 叶彦波[4], 阮吉寿[1]

1. 南开大学数学学院, 天津 300071;
2. 重庆大学生物工程学院, 重庆 400044;
3. 中国科学技术信息研究所, 北京 100038;
4. 中国科学院武汉病毒研究所, 武汉 430071;

摘要: 将染色质免疫共沉淀技术(ChIP)与下一代高通量测序技术相结合的染色质免疫共沉淀测序(ChIP-seq), 已成为功能基因组学、特别是基因表达调控领域研究的关键技术。ChIP-seq 实验带来的海量数据向生物信息学研究人员提出了新的挑战。由于此领域数据处理技术的发展大大滞后于实验技术进步, 有必要系统地介绍和回顾 ChIP-seq 数据处理的各个方面, 以便更多研究人员进入此领域设计或改进相应的算法。文章结合实例详细介绍了 ChIP-seq 数据整个流程, 并重点讨论了其中的主要问题和关键环节, 为这一研究领域的科研人员提供一个快速而深入的认识。

关键词: 下一代测序; ChIP-seq; 数据处理; 转录调控; 表观遗传学

Processing and analysis of ChIP-seq data

GAO Shan[1], ZHANG Ning[1], LI Bo[2], XU Shuo[3], YE Yan-Bo[4], RUAN Ji-Shou[1]

1. *College of Mathematics*, *Nankai University*, *Tianjin* 300071, *China*;
2. *College of Bioengineering*, *Chongqing University*, *Chongqing* 400044, *China*;
3. *Institute of Scientific and Technical Information of China*, *Beijing* 100038, *China*;
4. *Wuhan Institute of Virology*, *Chinese Academy of Sciences*, *Wuhan* 430071, *China*;

Abstract: The next-generation sequencing coupled with chromatin immunoprecipitation (ChIP-seq) is becoming a key technology for the study of transcriptional regulation in the context of functional genomics. Due to the overwhelming amount of data generated from ChIP-seq experiments, the ChIP-seq data processing brings many new challenges in the field of bioinformatics. Considering the development of data processing skills largely behind that of the ChIP-seq experiment techniques, it is urgent to give a review on the ChIP-seq data processing for more and more oncoming researchers to build or improve algorithms. This paper provides a brief overview of the ChIP-seq data processing, highlighting the main problems and methods in detail, to allow scientists to understand rapidly and deeply.

Keywords: next generation sequencing; ChIP-seq; data process; transcription regulation; epigenetics

收稿日期: 2011−10−10; 修回日期: 2011−12−04
基金项目: 国家自然科学基金项目(编号：31050110432, 68075049)资助
作者简介: 高山, 博士后, 研究方向：生物信息学。Tel: 022-23500736; E-mail: gao_shan@mail.nankai.edu.cn
通讯作者: 阮吉寿, 博士, 教授, 研究方向：生物信息学。Tel: 022-23501449; E-mail: jsruan@nankai.edu.cn
网络出版时间: 2012-4-5 11:00:30
URL: http://www.cnki.net/kcms/detail/11.1913.R.20120405.1100.001.html

下一代测序(Next generation sequencing, NGS)技术[1~3]不仅带来了测序领域(如基因组从头测序和重测序)的革命性变化, 而且与现有手段结合开拓了很多新的应用领域[4], 如染色质免疫共沉淀测序(ChIP-seq)[5]、RNA测序(RNA-seq)[6]和甲基化测序(Methyl-seq)[7,8]等。其中, ChIP-seq作为基因组学、特别是功能基因组学研究领域的热点技术之一, 其当前面临的主要问题就是数据分析手段的发展大大滞后于实验手段的进步。ChIP-seq数据的高通量、高噪声和高复杂度等特点催生了多种分析方法和软件工具, 然而这些方法和工具还仅在一定范围适用, 并且缺乏统一的标准和规范, 急需有力整合。ChIP-seq实验的标准化和自动化与数据处理的发展缓慢之间的矛盾严重阻碍了相应领域的数据挖掘和知识发现。本文在系统介绍ChIP-seq数据处理的技术框架的基础上, 结合经典数据集和已有方法, 对其中亟待解决的主要问题做了详细介绍和深入讨论, 旨在为该领域研究人员提供一个全面而简洁的介绍, 并为促进已有方法和工具的广泛应用和改进奠定基础。

1 ChIP-seq 简介

染色质免疫共沉淀技术(Chromatin immunoprecipitation, ChIP)也称结合位点分析法[9,10], 是全基因组水平研究DNA与蛋白质相互作用的有力工具和标准方法[11]。将ChIP与下一代高通量测序技术相结合的ChIP-seq技术, 由于具有成本低、效率高、检测的灵敏度和覆盖度高等优势[5], 已经成为这一领域的首选技术。

ChIP的实验原理如图 1 所示 :在生理状态下, 把细胞内的DNA与蛋白质交联(Crosslink)后裂解细胞, 分离染色体, 通过超声或酶处理将染色质随机切割, 利用抗原抗体的特异性识别反应, 将与目的蛋白相结合的DNA片段沉淀下来, 再通过反交联(Reverse crosslink)释放结合蛋白的DNA片段, 最后通过多种技术(定量PCR、芯片、测序等)获得DNA片段的序列。最常见的两种ChIP实验技术是N-ChIP[11,12]和X-ChIP[10]。前者用来研究DNA与高结合力蛋白的互作, 采用核酸酶消化染色质, 适用于组蛋白及其异构体方面的研究; 后者用来研究DNA与低结合力蛋白的互作, 采用甲醛或紫外线进行DNA和蛋白交联, 超声波片段化染色质, 适用于多数非组蛋白方面的蛋白的研究。

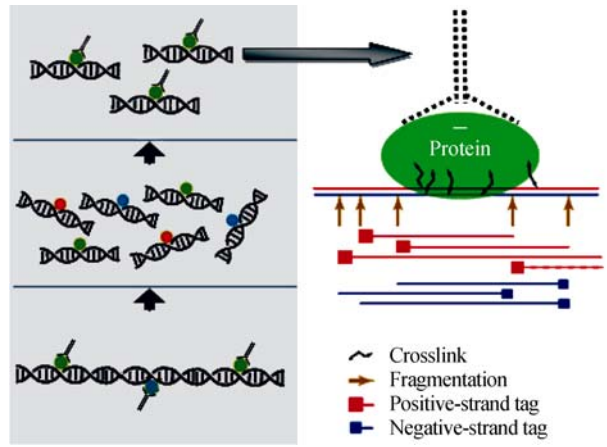

图 1 **ChIP-seq**实验原理示意图[13]
图中右侧的 DNA 长度对应测序得到的一个读长(Read), 红色表示 DNA 正义链, 蓝色表示反义链。

ChIP-seq当前的应用主要包括两个方面：一方面是DNA序列上转录因子结合位点(Binding sites)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[14]; 另一方面主要应用在表观遗传学领域[15], 包括研究基因组DNA甲基化、组蛋白修饰和核小体定位等问题[16]。下面用 2 个项目具体介绍这两类应用。

项目 1 是在全基因组范围内研究神经元限制性沉默因子NRSF(Neuron-restrictive silencer factor)的结合位点[17]。NRSF是一种锌指结构的负调控型转录因子, 广泛表达于胚胎干细胞、神经干细胞和非神经细胞中, 在神经分化等多种生命过程中发挥着重要作用[18,19]。项目 1 包括 2 组实验, 每组实验使用一个实验组和一个对照组, 实验 2 与 1 的唯一区别在于样本处理时是否进行了PCR预扩增, 两个实验结果的对比可以更好地确认识别的结合位点。本文只结合实验 1 的相关内容来介绍第一类应用, 不涉及实验 2 的内容。

NRSF数据集是一个非常经典的数据集, 它和另外 4 个转录因子(CTCF[20]、STAT1[21]、GABP[22]和FoxA1[22,23])数据集一起, 经常作为标准数据集来评价和比较各类结合位点识别算法和软件(下文详细介绍)。NRSF、CTCF和STAT1 的BED文件(原始数据(读长)定位后的文件格式)可查阅http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/sissrs/。

表 1　NRSF 原始数据集(测序后)

	实验 1		实验 2	
	实验组	对照组	实验组	对照组
所有读长	4 756 090	5 108 543	2 126 823	3 100 468
定义唯一的读长	3 661 543	3 834 288	1 697 893	2 319 582
用于分析的读长	3 661 543	3 661 543	1 697 893	1 697 893

注：表中数据来自文献[17]。

表 2　项目 2 中的原始数据集(测序后)

细胞类型	抗原表位	重复次数	定位唯一的标签
ES cells	pan-H3	1	4490474
	H3K4me3	2	8398790
	H3K9me3	2	4411447
	H3K27me3	2	7211279
	H3K36me3	2	7217118
	H4K20me3	2	5139339
	RNAP II	1	2736500
NPCs	H3K4me3	2	6995068
	H3K9me3	2	4614191
	H3K27me3	2	8166774
	H3K36me3	2	7899115
MEFs	H3K4me3	2	11371374
	H3K9me3	2	4468908
	H3K27me3	2	12208145
	H3K36me3	2	10315848

注：表中数据来自文献[24]。

项目 2 是在全基因组范围内, 用 7 种特征标志来描述 3 种小鼠细胞的染色体状态。每一种细胞可以得到一个用 7 种特征描述的染色体状态谱(Chromatin state profile), 也称染色体状态图(Chromatin state map)[24]。7 种特征包括 6 种组蛋白修饰和RNA聚合酶II的结合位点。其中 6 种组蛋白修饰分别是：三甲基化组蛋白H3 赖氨酸4(Trimethylated histone H3 lysine 4, H3K4me3)、H3K9me3、H3K27me3、H3K36me3、H4K20me3 与 pan-H3。3 种细胞分别是：胚胎干细胞 (Embryonic stem cells, ES cells)、神经前体细胞(Neural progenitor cells, NPCs)和胚胎成纤维细胞(Embryonic fibroblasts, MEFs)。

2　ChIP-seq 数据处理的流程

测序ChIP-seq数据的基本处理包括：读长定位和富集区域(“峰”)的识别。不同的应用在后续处理方面有不同的需求。项目 1[17]代表的一类应用中, 后续处理相对简单, 基本集中在转录因子结合位点基序分析等; 项目 2[24]代表的一类应用中, 后续处理包括更为广泛的内容：如“峰”在基因组不同区域的偏好性统计、“峰”关联基因的GO注释[25]及Pathway分析[26,27]、“峰”关联的各种突变分析等[28](图 2)。

3　ChIP-seq 数据的基本处理

3.1　读长定位

测序直接得到的核酸序列片段叫做读长(Read), 它只是被测序 DNA 片段 5′端开始的一个片段, 一般长度是 25～50nt, 对应着测序输出文件中的一个记录。读长定位也叫读长对齐(Alignment), 是把所有读长定位到参考基因组序列上, 能够定位的读长叫做标签, 标签和读长经常代表同一个意思。参考基因组往往是通过第一代测序技术得到的高准确度序列, 如人类基因组 hg18。

3.1.1　影响读长定位的主要因素

影响读长定位的主要因素是读长的长度和数据质量。读长过短会影响其定位到基因组唯一位点(Site)的可能性(图 3A)。读长过长会带有大量的低质量数据(如Illumina/Solexa平台的测序数据,大部分误差积累在 3′端[29]), 并由此带来定位时更多的错配; 而且会大大增加计算时间。图 3B是前面提到的项目 1 中统计的错配频率与其在读长中位置的关系。

3.1.2　常用的读长定位算法与工具

读长定位算法属于短序列比对算法范畴, 王曦等[30]将这方面主要的算法及工具分为 3 大类：

(1)空位种子片段索引法, 如Maq[31]、ELAND等, 首先将读长切分, 并选取其中一段或几段作为种子建立搜索索引, 再通过查找索引、延展匹配来

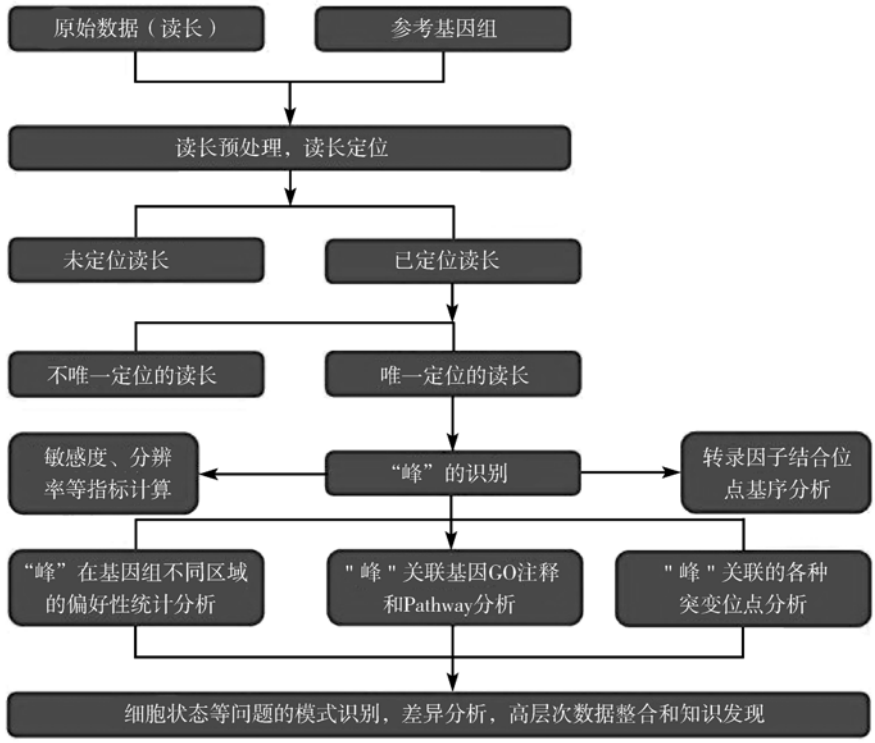

图 2　ChIP-seq 数据处理流程

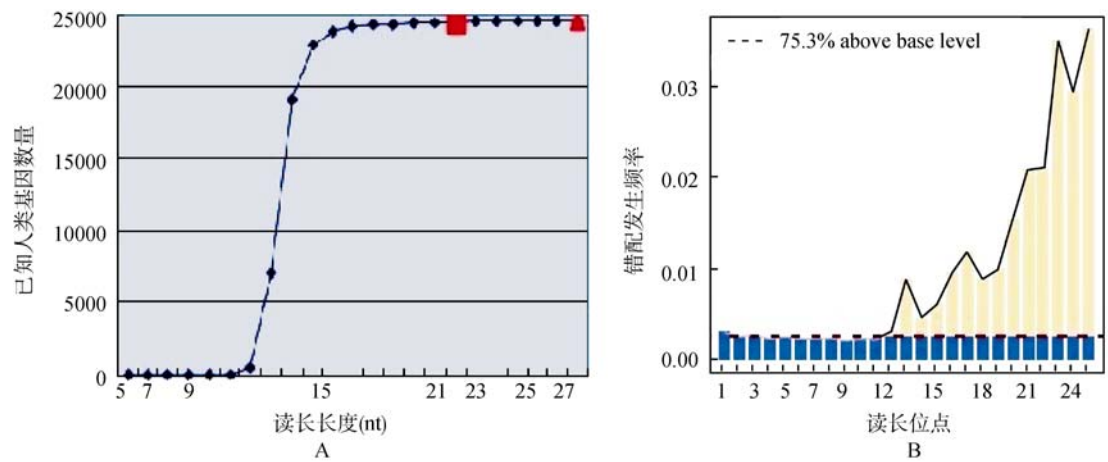

图 3　影响读长定位的主要因素[13]

A：已知人类基因数量与读长长度关系，长度增加到 21 时，已经很好的保证读长定位的唯一性；B：读长位点与其错配发生概率的关系，图中错配基准值(Base level)的估计基于位点 2~5(用红色虚线标出)，高于基准值的部分用浅黄色画出，低于的部分用蓝色，读长第 1 个位置的误差值往往较高，不符合整体的变化规律，在计算时往往不予考虑。

实现读长定位;

(2)Burrows-Wheeler 转换法，如 Bowtie[32]、BWA、SOAP2 等，通过B-W转换将基因组序列按一定规则压缩并建立索引，再通过查找和回溯来定位读长;

(3)Smith-Waterman 动态规划算法，如 BFAST、SHRiMP 等，由于效率太低，在此领域几乎不使用。

当前最为常用的 Bowtie 和 BWA 中，前者在时间效率上要大大优于后者，因此对于人类等较大基因组，Bowtie 的应用更为普遍；BWA 更适于具有较

多插入删除突变的情况。

3.1.3　读长定位前的预处理

最简单的读长定位方法就是把测得的整条读长, 不经过预处理就直接在参考基因组上定位, 然后根据读长与基因组对齐的结果进行筛选, 显然这样是不考虑对齐的质量和计算效率的。常用的读长定位方法是选取固定长度的读长, 如项目 2 就设定为 36nt, 这样既保证了映射唯一性又可以尽量删除掉 3′端的低质量数据。

最好的方法是在定位前对读长数据进行预处理, 根据质量注释去除低质量数据, 以得到高保真区(从 5′端开始的一段连续区域, 它包括的所有位点的质量分数要大于一个阈值, 如 Q20)。除了去除低质量数据, 预处理工作还包括去除各种污染, 这些污染来自接头、PCR 引物、非目标物种的 DNA 或 RNA 等序列。读长数据的预处理又叫做数据清理(Data clean), 通常需要编制较复杂的程序。

3.1.4　定位的质量控制和输出选择

经过预处理的数据, 还是会包括很多测序误差和污染, 再加上基因组个体差异(如 SNP)等原因, 需要在读长对齐过程中考虑一定的错配, 同时可假定没有插入/删除错误(仅对 Illumina 测序数据适用)。Bowtie 提供两种错配控制方式, 一种(v 模式)是控制总的错配数量, 适用于已经清理过的数据, 另外一种(n 模式)将读长序列分为高保真(种子)区和低保真区, 可以通过参数设置同时控制高保真区的长度, 高保真区内错配总数, 以及总错配的质量分数和, 这种方式参数设定更为精细。

按照一定的对齐质量要求, 读长在基因组上的定位可以得到 1 个或多个位点。通常有多个位点的读长会被去除, 只保留有唯一定位位点的读长。另外, 为了去除 PCR 不均匀扩增等带来的影响, 经常把定位到同一个位点上的多个读长去除到 1 个。

3.2　ChIP-seq 数据的可视化及注释

基因组数据的高度复杂性要求相应的可视化工具能够在各种尺度和不同视角显示 DNA 序列及其注释数据, 这是后期统计分析和作图的先决条件。典型的基因组可视化工具为基因组浏览器, 其中最有影响力的就是 UCSC Genome Browser。与其他基因组浏览器相比, 其主要特点包括：支持 SAM、BAM、BED 等多种数据格式; 可以通过 Web 访问, 也可以通过建立本地镜像使用; 与多种统计分析工具如 Galaxy 无缝连接; 最主要特点就是后端有多种注释数据库支持。

在项目 1 中, 利用基因组浏览器可以方便地观察几种转录因子在全基因组范围内的分布; 在项目 2 中, 数据的可视化尤为重要。项目 2 的主要工作之一就是用两种标志(H3K4me3 和 H3K27me3)来刻画 3 种细胞的染色体状态差异, 其结论之一就是 ES 细胞中的很多启动子区域同时携带这 2 种标志, 因此细胞处于平衡状态。

作为一种甲基化标志, H3K4me3 被发现经常与基因活跃关联, 而 H3K27me3 则经常与基因沉默关联, 两者结合可以作为双价标志将细胞划分为不同的状态。双价标志(H3K4me3/H3K27me3)存在时, 细胞处于平衡状态; 只有标志 H3K4me3 存在时, 细胞处于活跃(Active)状态; 只有标志 H3K27me3 存在时, 细胞处于抑制状态。通过基因组浏览器, 就可以直观的观察到每种细胞所处的状态(图 4), 为后续的统计分析提供思路。

3.3　结合位点的识别

ChIP-seq 实验首先富集与蛋白质结合的 DNA 片段, 然后通过测序来得到它们的序列。由于测序前染色质被随机打断为一些大片段, 而目前的测序技术只能测得这些片段(长度范围 100～500nt, 集中在 200～300nt)的 5′端的一小段, 因此不能直接得到对应实际结合位点的富集区域(Enrichment region)或者叫做“峰”(Peak), 而只能根据读长定位后出现在保护蛋白两侧的标签簇(Tag cluster)来计算得到真正的结合位点(图 5)。结合位点的识别算法, 也叫做“峰”寻找算法(Peak finding algorithms), 就是依据这些标签簇来确定结合位点的位置。结合位点的识别算法是当前 ChIP-seq 数据处理领域的一项重要工作。

3.3.1　结合位点识别的原理

富集区域、结合位点和“峰”这些概念经常混用。“峰”本质上共有两个概念：一个是指读长定位到基因组后得到的富集区域和“峰”, 它们由标签簇, 也称重叠标签(Overlapped tags)组成, 在真正结合位点的周围形成双峰模式(Bimodal pattern); 另一个是指

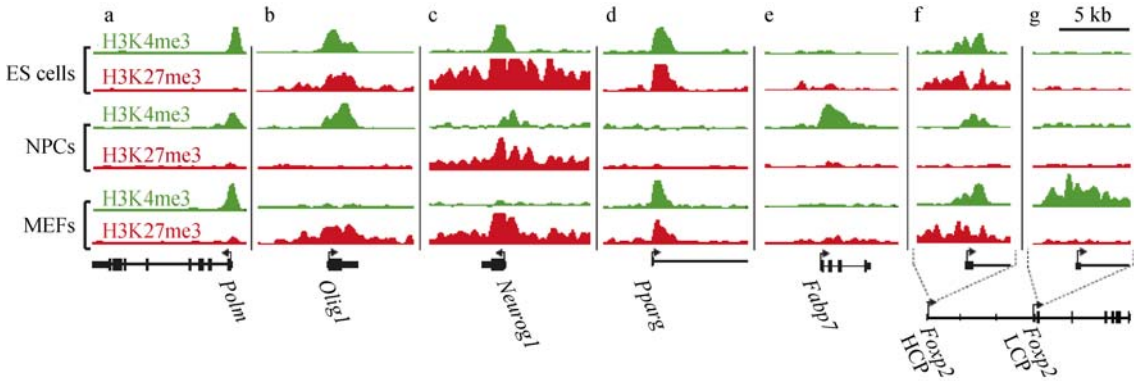

图 4　双价标志在 3 种启动子区域的分布[24]

a：H3K4me3 与 *Polm* 基因启动子区在 3 种细胞中都关联；b：H3K4me3/H3K27me3 与 *Olig1* 基因启动子区在 ES 细胞中关联，但是在 NPCs 细胞中只有 H3K4me3, MEFs 细胞中只有 H3K27me3; c：基因 *Neurog1* 启动子区情况; d：基因 *Pparg* 启动子区情况; e：基因 *Fabp7* 启动子区情况; f :基因 *Foxp2* 高 CpG 启动子区(High CpG promoters, HCP)情况; g :基因 *Foxp2* 低 CpG 启动子区(Low CpG promoters, LCP)情况。

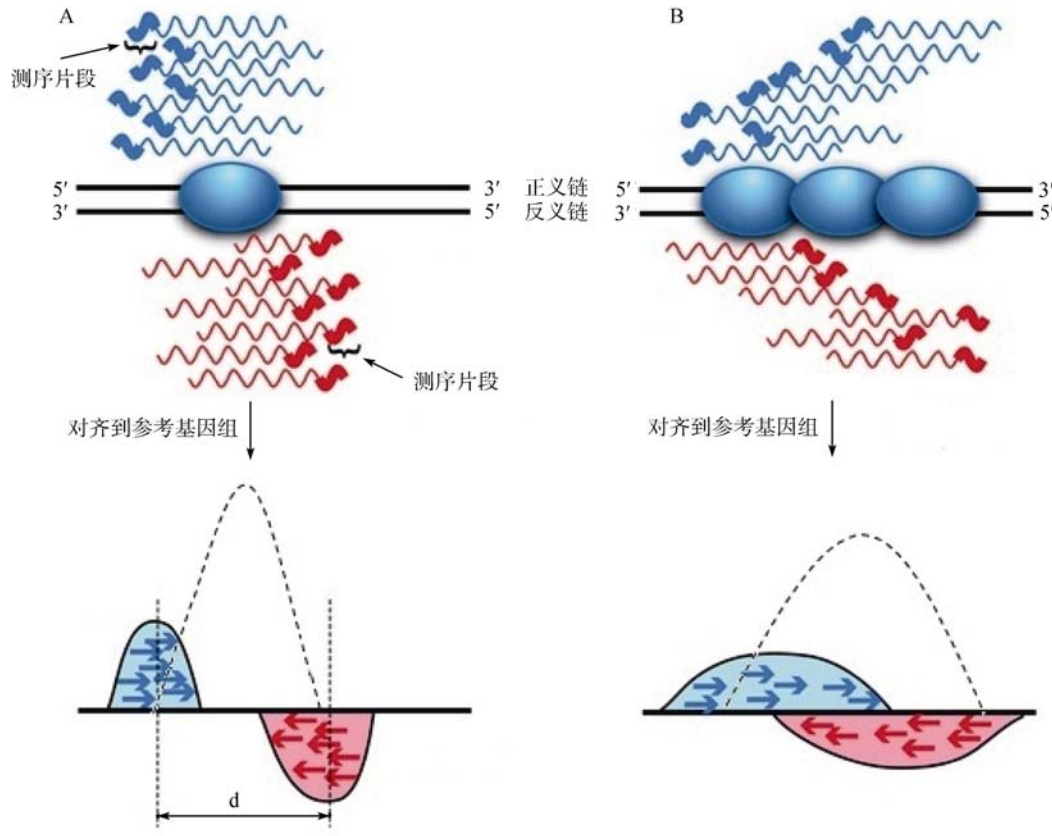

图 5　从双峰模式中寻找结合位点(“峰”)[33]

A：转录因子特异性结合 DNA 的情况，产生了理想的双峰模式，这时两波峰之间的距离 d 相当于测序片段的平均长度；B：分布式结合 DNA 的情况(如组蛋白或 RNA 聚合酶)，产生了另外一种双峰模式。蓝色椭圆代表结合到 DNA(黑色加粗的线条表示)的蛋白(如转录因子)；波浪线代表从 ChIP 实验中富集的 DNA 片段(也就是下一步要测序的片段)，蓝色来自正义链，红色来自反义链；带箭头的蓝色、红色粗线条表示测序得到的读长，来自于上半区的波浪线。

真实存在的结合位点，或通过识别算法得到的目的富集区域，也就是假定的结合位点。本文中用到的“峰”和结合位点指的都是后面这个概念。双峰模式由正义标签簇(Sense tag cluster)和反义标签簇(Antisense tag cluster)组成(图 5)，也叫做 Watson tags 和 Crick tags。

寻找结合位点的基本思路就是从实验中得到的双峰模式中计算出真正的“峰”所在的位置。每种算法都会得到一组候选的“峰”，并且用两个指标来描述这些“峰”。第一个就是“峰”的得分(Score)用以表示这个“峰”的信号大小(排除随机因素)，常用的有读长密度(Tag density)和富集倍数(Fold enrichment)；第二个是统计显著性指标，用以表示这个“峰”的可信度。一般是首先根据不同的模型得到的 *p*-value、*q*-value、*t*-value 或最大后验概率；然后，将这些“峰”根据其得分降序排列，也可以用得分与可信度二级排序；最后，根据设定的阈值来选取前 n 个“峰”作为输出。

“峰”选取(阈值设定)的方法有很多种，最简单的是百分比方法，如Top 10%；或可信度方法，如 *p*-value<0.05；或某个经验值，如富集倍数>7.5。另外一个常用的阈值设定方法是控制经验错误发现率(False discovery rate, FDR)，如低于 0.01，FDR的估计需要有对照组的数据(见公式 1)。公式 1 中，$N_E(s)$表示实验组中分数大于等于s的“峰”的数量，$N_C(s)$表示对照组中所有分数大于等于s的“峰”的数量，b是偏置因子，主要用于防止分母为 0，在文献[13]中 b=0.5。

$$\frac{N_E(s)+b}{N_C(s)+b} \tag{1}$$

在计算得分和可信度这两个指标时，必须考虑背景噪声的影响。因此，大部分的算法都考虑了如何利用对照组的数据。例如，得分指标中的富集倍数就是 ChIP 数据与对照数据之比(信噪比)，还有的方法计算 ChIP 数据与正规化之后的对照数据之差。FDR 等其他可信度指标的计算也要用到对照数据。对照数据主要有两类，分别是无抗体(No antibody)富集取样测量得到的和非特异抗体(Non-specific antibody)富集取样测量得到的对照。没有对照数据的时候，需要对噪声假定一定的参数模型来计算相应的指标，常用的有泊松分布、局部泊松分布和条件二项分布。

3.3.2　算法的评价

为了公平进行比较，各种算法的参数都应设定为默认值。考虑到算法输出“峰”的数量依赖于阈值，无法设定一个统一的阈值来比较各类算法。因此，可以通过绘制各类指标随阈值变化的曲线来比较各种算法的差异。

$$\begin{aligned} Sn &= \left[TP_n/\left(TP_n+FN_n\right)\right] = TP_n/P_n \\ Sp &= \left[TP_n/\left(TP_n+FP_n\right)\right] = TP_n/\#Called\ peaks \end{aligned} \tag{2}$$

算法的主要评价指标包括敏感度(Sensitivity, *Sn*)，也叫做覆盖度(Coverage)；特异度(Specificity, *Sp*)。TP(True positive)表示真实存在而且算法也找到了的真阳性样本，FP(False positive)表示真实不存在而算法找到了的假阳性样本，TN(True negative)表示真实不存在而算法也没找到的真阴性样本，FN(False negative)表示真实存在而算法没找到的假阴性样本，P 表示所有真实存在的“峰”。公式 2 中所有下标均表示这种样本的个数，“Called peaks”表示某种算法找到的“峰”，一般根据得分或者可信度降序排列，“#”表示其个数。

求敏感度 *Sn*(见公式(2))过程中的关键问题就是如何获得所有已知结合位点(数据 P)。目前主要有两种方法：一种是用实验方法验证，如实时荧光定量 PCR(Real-time quantitative PCR, qPCR)检测；另外一种就是用已知的转录因子结合位点的经典基序通过全基因组扫描得到所有的“峰”，最常见是用 MEME/MAST 软件扫描，然后选择超过一定可信度阈值(如 $P<1\times10^{-7}$)的结合位点组成 P 的集合。qPCR 作为获得 P 的金标准，其优点是可信度高，但成本高，因而可用数据量小，不足以覆盖全基因组；后一种方法的主要问题是符合未知的非经典基序(Non-canonical motif)的样本不能被识别为 TP。由于两种方法都不可能得到全部的真实结合位点，最好的办法就是用敏感度(TP/P)随阈值变化的曲线(简称敏感度曲线)来显示不同算法在“峰”寻找能力上的差异。已知 P 后，TP 的确定相对简单：一个预测的“峰”的顶点的坐标如果在一个已知结合位点中心的上下游若干个核苷酸之内(如 250nt)，就认定为一个 TP。

图 6 是在 NRSF 数据集上，应用上述两种方法获得数据 P 后分别画出的 11 种算法的敏感度曲线，它们所用的阈值都是前 n 个“峰”(已按照打分或可信度排序)。方法比较时，由于算法 Sole-Search 和 GisGenome 只识别了 1 800 个“峰”，因此要在 x 轴选取一个公共的截断点(x=1800)来计算各条曲线下的面积值，做最终的比较。

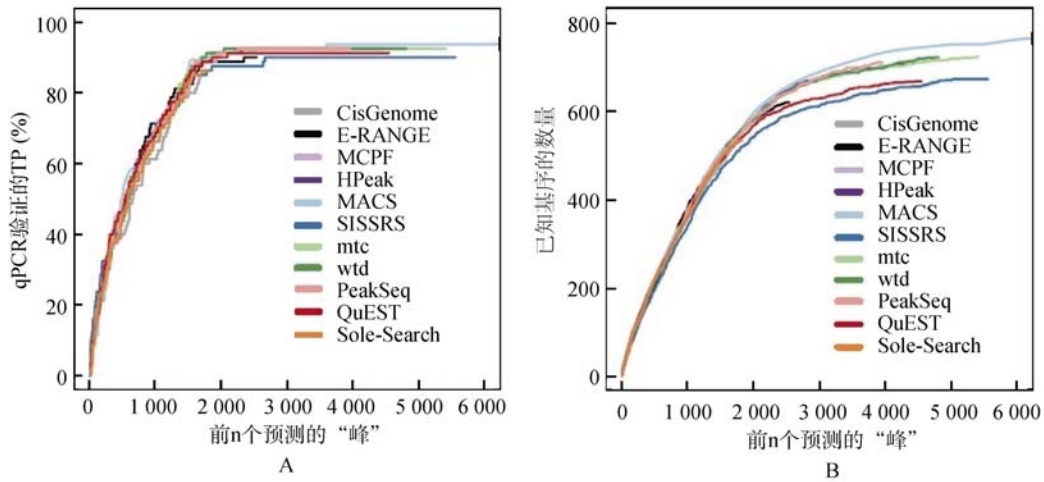

图 6 根据两种验证手段得到的敏感度曲线[33]

A：通过 qPCR 在 NRSF 数据集上共发现 83 个真实的结合位点，坐标 y 轴是算法找到的前 n 个“峰“中包括真实”峰“的个数与全部真实”峰“之比，也就是敏感度(TP/P)，坐标 x 轴是算法找到的前 n 个“峰”的个数; B：MEME/MAST 软件包扫描并得到了 NRSF 数据集的经典基序 NRSE2，坐标 y 轴是算法找到的前 n 个“峰“中包括真实”峰“的个数，坐标 x 轴是算法找到的前 n 个“峰”的个数。

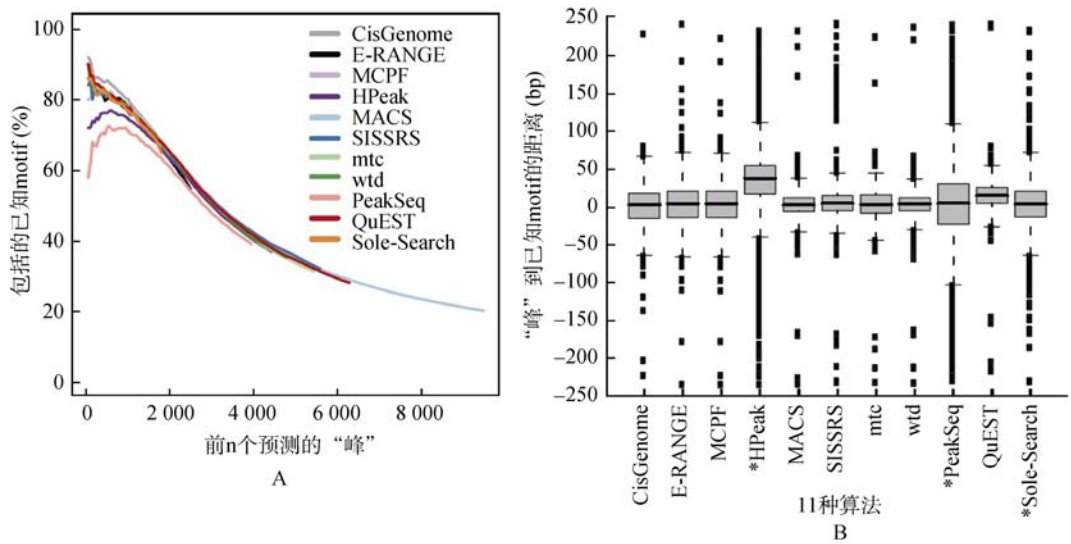

图 7 **NRSF**数据集上 **11** 种算法的特异度曲线与分辨率分布[33]

A：根据是否含有经典基序来判定TP(True positive)，y轴表示的就是特异度(TP/#Called peaks)[33]; B：图中*表示算法不提供顶点的基因组坐标，只能通过“峰”的起止位点估算出。所有算法都选取前 1 500 个“峰”进行统计。

求特异度 *Sp*(见公式(2))的难点在于如何鉴定 TP，用实验手段(如 qPCR)来鉴定所有算法找到的“峰”(即“Called peaks”)是否是 TP 是不现实的。一个常用的替代方法就是：如果一个找到的“峰”的上下游若干个核苷酸之内(如 250nt)含有一个已知的经典基序，就认定为一个 TP，然后构建特异度随阈值变化的曲线(简称特异度曲线)来评价不同算法之间的性能差异(图 7A)。

另外一种常用的算法评价指标—ROC曲线，可以同时包括敏感度和特异度的全部信息，在Kharchenko等[13]的工作中得到使用。但是正如前面提到的，敏感度和特异度只是一种估计值，因此由这些信息得到的ROC曲线不能够很好地反应一个算法的总体性能。

在结合位点寻找的应用中，不仅需要尽量把需要的“峰”都找到，还需要评估找到的“峰”的质量等其他指标，最主要就是“峰”的分辨率。一般分辨率的定义是指算法找到的“峰”的顶点到周围(一般是左右

250nt)最近的高可信度的已知基序中心的距离。有的算法预测“峰”后只输出其范围，而没有顶点在基因组上的坐标信息，这时可以用这些“峰”的起止位点的中点坐标来估算“峰”的顶点的坐标。一般来讲，一种算法在一个数据集上的分辨率围绕一个中心值上下波动，图 7B 就是 NRSF 数据集上 11 种算法的分辨率分布图。

3.3.3　主要的结合位点识别算法及其特点

Wilbanks等[33]将主要的结合位点识别算法分成 4 类：第一类方法最简单，直接将标签向 3′端移动一定长度，或者延长得到一定长度的扩展标签。这一类算法最典型的就是XSET(Extended sequence tags)[34]。XSET将所有标签向 3′端延长得到 200nt长度的扩展标签，记下基因组每个位点上重叠覆盖的扩展标签的个数，超过阈值(默认 11)的位点被识别为“峰”的区域，“峰”的高度就是该“峰”区域包括的所有扩展标签个数的最大值。第二类方法是滑动窗口(Sliding window)方法。典型的有MACS(Model- based analysis of ChIP-seq)[35]。MACS使用用户指定的窗口长度(默认 300)的 2 倍去扫描基因组，记录窗口内标签个数是背景标签个数n倍(默认是 32)的位点并连成一个“峰”，随机取 1 000 个这样的高质量的“峰”，找到每个“峰”内正义标签簇和反义标签簇的中心，并定义两个中心之间的距离为该“峰”的宽度，求这 1 000 个峰的平均宽度d(图 5)，将正义标签簇和反义标签簇都向 3′端移动d/2 的距离，最后用 2d宽度的窗口再次扫描基因组以得到最后的“峰”。滑动窗口方法容易产生依赖于窗口长度的边缘效应，因此又出现了其改进方法(第三类)，其基本思想就是引入高斯核密度函数(Gaussian kernel density function)，使窗口的取样结果更为平滑，这类方法的代表有QuEST(Quantitative enrichment of sequence tags)[36]。最后一类方法不同于前面提到的方法直接选出“峰”，而是先根据用户提供的参数推荐出一些候选的“峰”及其得分和可信度，而后用户可以进一步通过排序选择最终的“峰”的列表。最后一类方法也被称作方向性打分方法(Directional scoring methods)，其典型方法是SISSRS(Site identification from short sequence reads)[33]。

Kharchenko等[13]在NRSF、CTCF和STAT1 三个数据集上对CSP(ChIPSeq peak locator)、XSET、WTD(Window tag density)、MTC(Mirror tag correlation)和MSP(Matching strand peaks)等 5 种结合位点识别算法的性能进行了比较和评估；Wilbanks[33]等在NRSF、GABP和FoxA1 三个数据集上对MTC、WTD、SISSRS、PeakSeq、MACS、MCPF、Sole-Search、CisGenome 、 E-RANGE 、 QuEST 和 HPeak(HMM-based peak-finding algorithm)等 11 种算法性能进行了比较和评估。以上研究结果表明，各类算法在敏感度和特异度两项指标上差异不大，但是在分辨率方面差异较大。

4　ChIP-seq 数据的后续分析

在从实验数据中得到“峰”后，项目 1 代表的一类应用中的数据后续分析基本集中在转录因子结合位点基序分析，通常大部分“峰”都符合非经典基序(Non-canonical motif)，从不符合经典基序的“峰”中可以归纳出以前没有发现的非经典基序，图 8A 是 NRSF 的基序。项目 2 代表的一类应用中，数据的后续分析范围十分广范，找到的“峰”可以与基因组中不同种类的区域关联，并统计其偏好性(图 8B)。当这些“峰”与邻近基因关联后，可以对所有关联基因做 GO 注释、Pathway 分析、或进行基因表达的分析，这些内容是 ChIP-seq 实验进一步揭示生物体基因表达调控和细胞各种生命活动规律的关键。

5　结语与展望

本文在整体介绍ChIP-seq数据处理框架的基础上，以两个项目为例，详细介绍了两类不同应用中ChIP-seq数据处理需要解决的问题和已有方法。在找“峰”算法方面，由于DNA打碎方法、染色质压缩程度的不规则性、PCR扩增的偏好性、基因组中序列的冗余程度以及测序过程都会引入错误而造成假阳性[38]，因此，只有通过统计方法去除噪声，才能准确地得到“峰”的信号。除了算法层面，还可以从改进实验手段或策略方面入手，配合算法不断提高各项性能指标。此外，当前的算法最初都是基于单末端测序技术，双末端测序(Paired-end sequence)实验技术可以大大提高找“峰”算法的性能。ChIP-seq的后续分析是未来此领域数据处理的热点和难点。

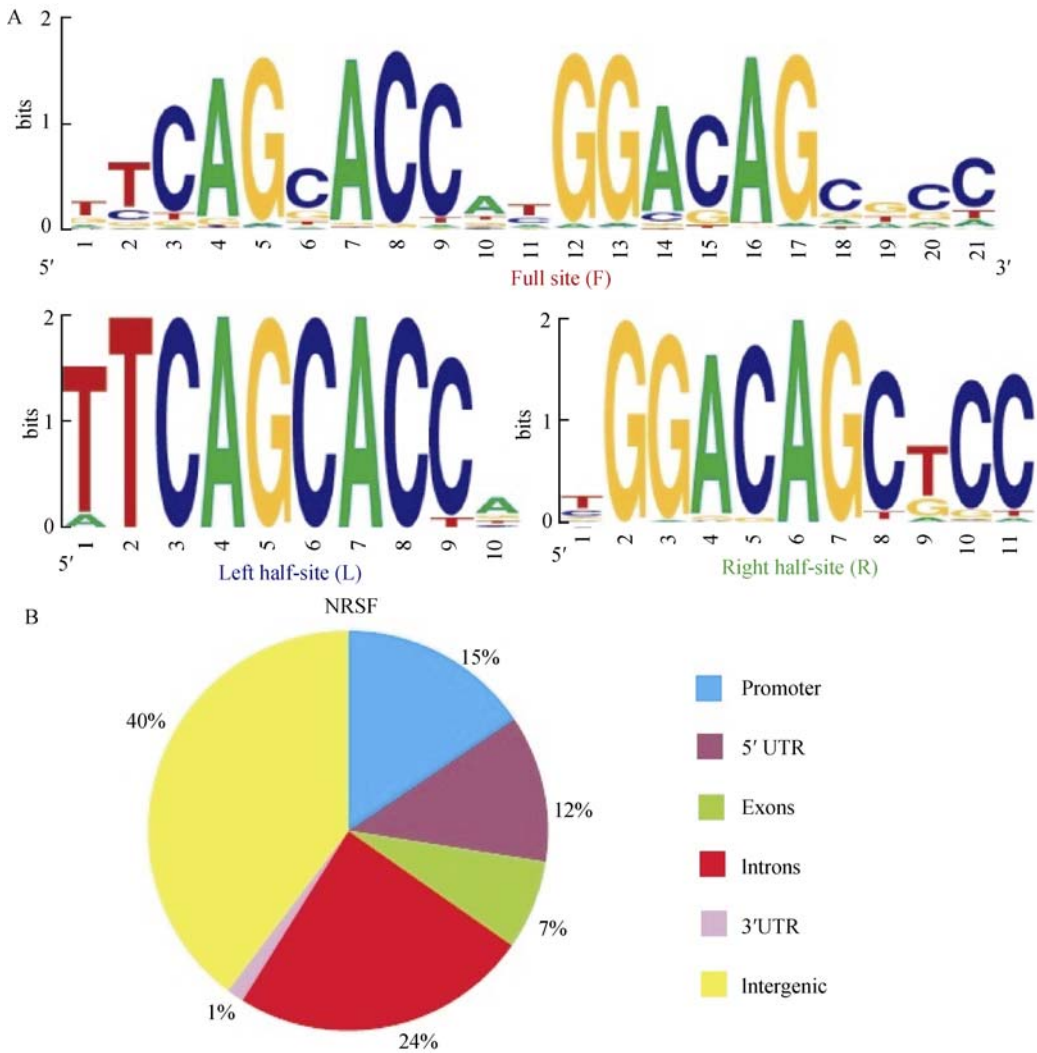

图 8　ChIP-seq数据的后续分析[37]

A：NRSF 的基序。Full site (F)：整体基序; Left half-site (L)：左半基序; Right half-site (R)：右半基序。B：“峰“在基因组不同区域的分布，转录起始点上游 5 kb 的区域定义为启动子, RefSeq 作为参考基因。

“峰”找到后，往往要统计这些“峰”在基因组上的分布特征，也可以用以描述不同细胞或染色体状态，以进一步揭示细胞发育、疾病等生物学现象。

参考文献(References):

[1] Schuster SC. Next-generation sequencing transforms today’s biology. *Nat Methods*, 2008, 5(1): 16–18. DOI

[2] Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*, 2008, 26(10): 1135–1145. DOI

[3] Metzker ML. Sequencing technologies-the next generation. *Nat Rev Genet*, 2009, 11(1): 31–46. DOI

[4] Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet*, 2008, 24(3): 133–141. DOI

[5] Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 2009, 10(10): 669–680. DOI

[6] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 2009, 10(1): 57–63. DOI

[7] Wold B, Myers RM. Sequence census methods for functional genomics. *Nat Methods*, 2008, 5(1): 19–21. DOI

[8] Fouse SD, Nagarajan RP, Costello JF. Genome-scale DNA methylation analysis. *Epigenomics*, 2010, 2(1): 105–117. DOI

[9] Carey MF, Peterson CL, Smale ST. Chromatin immunoprecipitation (ChIP). *Cold Spring Harbor Protocols*, 2009, 2009(9): pdb. prot5279. DOI

[10] Sun JM, Chen HY, Davie JR. Differential distribution of

unmodified and phosphorylated histone deacetylase 2 in chromatin. *J Biol Chem*, 2007, 282(45): 33227. DOI

[11] Massie CE, Mills IG. ChIPping away at gene regulation. *EMBO Reports*, 2008, 9(4): 337–343. DOI

[12] Cosseau C, Azzi A, Smith K, Freitag M, Mitta G, Grunau C. Native chromatin immunoprecipitation (N-ChIP) and ChIP-Seq of *Schistosoma mansoni*. Critical experimental parameters. *Mol Biochem Parasitol*, 2009, 166(1): 70–76. DOI

[13] Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol*, 2008, 26(12): 1351–1359. DOI

[14] Yamaguchi-Shinozaki K, Shinozaki K. Organization of *cis*-acting regulatory elements in osmotic-and cold-stress-responsive promoters. *Trends Plant Sci*, 2005, 10(2): 88–94. DOI

[15] Laird PW. Cancer epigenetics. *Hum Mol Genet*, 2005, 14(Suppl 1): R65–R76. DOI

[16] Barski A, Zhao KJ. Genomic location analysis by ChIP-Seq. *J Cell Biochem*, 2009, 107(1): 11–18. DOI

[17] Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, 2007, 316(5830): 1497–1502. DOI

[18] Schoenherr CJ, Anderson DJ. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science*, 1995, 267(5202): 1360–1363. DOI

[19] Kim CS, Hwang CK, Choi HS, Song KY, Law PY, Wei LN, Loh HH. Neuron-restrictive silencer factor (NRSF) functions as a repressor in neuronal cells to regulate the μ opioid receptor gene. *J Biol Chem*, 2004, 279(45): 46464–46473. DOI

[20] Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng ZP, Furey TS, Crawford GE. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 2008, 132(2): 311–322. DOI

[21] Qin ZS, Yu JJ, Shen JC, Maher CA, Hu M, Kalyana-Sundaram S, Yu JD, Chinnaiyan AM. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics*, 2010, 11(1): 369. DOI

[22] Wallerman O, Motallebipour M, Enroth S, Patra K, Bysani MSR, Komorowski J, Wadelius C. Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing. *Nucleic Acids Res*, 2009, 37(22): 7498–7508. DOI

[23] Laajala TD, Raghav S, Tuomela S, Lahesmaa R, Aittokallio T, Elo L. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics*, 2009, 10(1): 618. DOI

[24] Mikkelsen TS, Ku MC, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie XH, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 2007, 448(7153): 553–560. DOI

[25] Leonelli S, Diehl AD, Christie KR, Harris MA, Lomax J. How the gene ontology evolves. *BMC Bioinformatics*, 2011, 12: 325. DOI

[26] Torres NV, Voit EO. Pathway analysis and optimization in metabolic engineering. Cambridge: Cambridge University Press, 2002. DOI

[27] Bader GD, Cary MP, Sander C. Pathguide: a pathway resource list. *Nucleic Acids Res*, 2006, 34(Suppl. 1): D504–D506. DOI

[28] Wang ZB, Zang CZ, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui KR, Roh TY, Peng WQ, Zhang MQ, Zhao KJ. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet*, 2008, 40(7): 897–903. DOI

[29] Whiteford N, Skelly T, Curtis C, Ritchie ME, Löhr A, Zaranek AW, Abnizova I, Brown C. Swift: primary data analysis for the Illumina Solexa sequencing platform. *Bioinformatics*, 2009, 25(17): 2194–2199. DOI

[30] 王曦, 汪小我, 王立坤, 冯智星, 张学工. 新一代高通量RNA测序数据的处理与分析. 生物化学与生物物理进展, 2010, 37(8): 834–846. DOI

[31] 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. DOI

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

[33] Wilbanks EG, Facciotti MT, Veenstra GJC. Evaluation of algorithm performance in ChIP-seq peak detection. *Plos One*, 2010, 5(7): e11471. DOI

[34] Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao YJ, Zeng TJ, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, 2007, 4(8): 651–657. DOI

[35] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 2008, 9(9): R137. DOI

[36] Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods*, 2008, 5(9): 829–834. DOI

[37] Jothi R, Cuddapah S, Barski A, Cui KR, Zhao KJ. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res*, 2008, 36(16): 5221–5231. DOI

[38] 李敏俐, 王薇, 陆祖宏. ChIP技术及其在基因组水平上分析DNA与蛋白质相互作用. 遗传, 2010, 32(2): 219–228. DOI