ChIP-seq 数据综合分析实用指南

对 ChIP-seq 数据的有效分析需要足够的序列 read 覆盖（序列深度），所需的深度主要取决于基因组的大小和蛋白质结合位点的数量和大小。对于乳腺转录因子（tfs）和染色质修饰，如增强子相关组蛋白标记，这些标记通常局限于特定的狭窄部位，具有数千个结合部位，2000 万个 read 可能足够（蠕虫和苍蝇 tfs 为 400 万个 read）[7].具有更多结合位点（如 RNA pol II）或更广泛因子（包括大多数组蛋白标记）的蛋白质将需要更多的 read，哺乳动物 ChIP-seq 的 read 量高达 6000 万[8]。重要的是，在 TF 实验和涉及扩散广域染色质（diffused broad-domain chromatin）数据的实验中，对照样品的测序应明显比 ChIP 样品深。这是为了确保足够的基因组和非重复的常染色体 DNA 区域的覆盖。为确保所选的序列深度足够，建议进行饱和分析，在接下来的两个步骤（read 映射和 peak calling）对从实际 read 中随机选择的 read 数量增加时，调用的 peak 应一致。饱和分析内置于一些 peak callers 中（例如，SPP[9]）。如果这表明 read 次数不够，则可以组合从技术复制实验中 read。为了避免过多的排序和估计，将从额外的排序中得到一个最佳的多个[10]。同样，ENCODE 工具提供了一个称为 pcr 瓶颈系数（pbc）的质量度量，定义为只有一个唯一 read 的基因组位置与至少一个唯一 read 覆盖的基因组位置的分数。

1.ChIP-seq 技术简介

染色质免疫沉淀后测序（ChIP-seq），首次在 2007 年描述[1-4]，允许在体内确定蛋白质与基因组的结合位置，基因组可以是转录因子、DNA 结合酶、组蛋白、伴侣蛋白或核小体。ChIP-seq 首先将结合蛋白与染色质交联，对染色质进行碎片化，使用特定抗体捕获结合到一种蛋白的 DNA 片段，并使用下一代测序（NGS）对捕获片段的末端进行测序。测序 DNA 的计算图谱确定了结合 DNA 结合酶、修饰组蛋白、伴侣蛋白、核小体和转录因子（tfs）的基因组位置，从而阐明了这些蛋白质-DNA 相互作用在基因表达和其他细胞过程中的作用。与 ChIP-chip 分析（ChIP，然后是微阵列杂交）相比，NGS 的使用提供了相对高的分辨率、低噪声和高基因组覆盖率。ChIP-seq 现在是全基因组分析蛋白质-DNA 相互作用最广泛使用的程序[5]，其在组蛋白修饰图谱中的应用在表观遗传学研究中具有重要意义[6]。

2.ChIP-seq 数据测序深度分析

ChIP 序列数据的有效分析需要序列 read 的足够覆盖（测序深度）。所需的深度主要取决于基因组的大小以及蛋白质结合位点的数量和大小。对于乳腺转录因子（tfs）和染色质修饰，如增强子相关组蛋白标记，这些标记通常局限于特定的狭窄部位，具有数千个结合部位，2000 万个 read 可能足够（蠕虫和苍蝇 tfs 为 400 万个 read）[7].具有更多结合位点（如 RNA pol II）或更广泛因子（包括大多数组蛋白标记）的蛋白质将需要更多的 read，哺乳动物 ChIP-seq 的 read 量高达 6000 万[8]。重要的是，在 TF 实验和涉及扩散广域染色质（diffused broad-domain chromatin）数据的实验中，对照样品的测序应明显比 ChIP 样品深。这是为了确保足够的基因组和非重复的常染色体 DNA 区域的覆盖。为确保所选的测序深度足够，建议进行饱和分析，在接下来的两个步骤（read 映射和 peak calling）对从实际 read 中随机选择的 read 数量增加时，调用的 peak 应一致。饱和分析内置于一些 peak callers 中（例如，SPP[9]）。如果饱和分析表明 read 个数不够，则可以加入从技术性复制实验中的 read。为了避免过多的排序，也许将从额外的排序中估计一个最优排序[10]。同样，ENCODE 工具提供了一个称为 pcr 瓶颈系数（pbc）的质量度量，定义为只有一个唯一 read 的基因组位置与至少一个唯一 read

覆盖的基因组位置的分数。

3.read 映射和质量指标

在将 read 数据映射到参考基因组之前，应通过应用 quality cutoff（框 1）对其进行过滤。然后，应使用可用的映射器之一（如 Bowtie[11]、BWA[12]、SOAP[13]或 MAQ[14]）映射剩余的 read。最新版本支持间隙对齐（例如，bowtie2），但大多数 ChIP-seq 实验不需要检测 indel（基因组上小片段(>50bp)的插入或缺失）。重要的是要考虑映射器报告的唯一映射 read 的百分比。生物之间的百分比不同，对于人类、小鼠或拟南芥 ChIP 序列数据，超过 70%的唯一映射 read 是正常的，而少于 50%可能引起关注。唯一映射 read 低百分比通常是由于 PCR 步骤的过度扩增、read 长度不足或测质量序平台出现问题，但对于一些 ChIPed 蛋白，它也许不可避免（例如，如果蛋白质在重复 DNA 中频繁结合）。read 映射器设计为允许（用户可设置）read 中的不匹配数量，选择此参数以适合所使用的 NGS 平台是很重要的（请咨询制造商）。大量“多重映射”的最后一个潜在原因是蛋白质在重复的 DNA 区域频繁结合。在最后一种情况下，使用双末端排序来减少映射模糊性可能会有所帮助。应记住，大多数 peak calling 算法将忽略（过滤掉）多映射 read（参见“peak calling”一节），尽管它们可以驱使发现新的绑定位点[15]。映射后，应评估 ChIP-seq 实验的信噪比（SNR），例如，通过质量指标，如链互相关（strand cross-correlation）[7]或软件包 CHANCE[16]的 IP 富集估计（框注 2）。这些措施将检测 ChIP-seq 的几种可能失效模式：免疫沉淀步骤富集不足，碎片大小选择不当（fragment-size），或测序深度不足。链互相关分析内置于一些 peak callers 中（例如 SPP 或 MACS[17][2 版]）。

4.Peak Calling

ChIP-seq 的一个关键分析是通过寻找具有大量映射 read（peak）的区域来预测基因组中与 ChIPed 蛋白结合的区域。敏感性和特异性之间的好的平衡取决于选择合适的 peak calling 算法和归一化方法（框 3-6，表 S1 和[18,19]）根据蛋白 ChIPed 类型：点源因子，如大多数 TFS（框 3）；广泛富集因子，如组蛋白标记物。（框 4），以及同时具以上两种特征的，如 RNA pol II（框 5）[20]。强烈建议使用对照样本的映射 read（例如，从输入 DNA），尽管一些 peak callers 可以使用 GC 含量或可映射性作为必要信息去评估非特异性水平或背景结合水平。重复 read（相同的 5’端）可以在 peak calling 前删除，以提高特异性（框 7）。尽管一些 peak callers 同时支持单末端和双末端 reads（例如，MACS），但其他 peak callers 专门设计用于提高双末端 read 测序的敏感性和特异性（例如，SIPeS[21]）。现有的 peak callers 有许多用户可设置的参数，这些参数会极大地影响所调用 peak 的数量和质量。例如，大多数 peak callers 的富集指标，如 p 值或 FDR，可能会受到所使用的统计模型、测序深度或基因组中结合位点的实际数量的极大影响。因此，使用相同的 p 值或 FDR 阈值不能确保调用的 peak 数量在库（libraries）和不同的 peak calling 之间是可比较的[22]。一种更好的方法是不可恢复发现率（IDR）阈值[23]，它与 Motif 分析一起，也有助于选择最佳 peak calling 算法和参数集（见“可重复性评估”和“Motif 分析”章节）。

4.再现性评估

为了确保实验结果是可重复的，建议对每个 ChIP-seq 实验进行至少两次生物复制，并检查 read 和识别峰的重复性[7,24]。read 的重复性可以通过计算每个基因组位置（映射的）read 数的皮尔逊相关系数（PCC）来测量[25]。PCC 的范围通常为 0.3–0.4（对于未相关样品）到 0.9（对于高质量实验中的复制样品）。低值通常意味着一个或两个复制品的质量可能较低。然而，这一数量可由少数高度富集区域控制，因此它可能无法反映较低富集区域的再现性

[26]。因此，在计算 PCC 之前，重要的是去除具有高 ChIP 信号的人工区域，例如中心体附近的区域、端粒、卫星重复区域以及 ENCODE 和 1000 个基因组黑名单区域。为了在 peak calling 水平上测量再现性，IDR 分析（框注 8）[23]可应用于从一对回复中识别的两组 peak。此分析评估复制之间已识别 peak 的等级一致性，并输出通过用户指定的可复制性阈值（例如，IDR=0.05）的 peak 数量。据报道，使用基于再现性的度量（如 IDR），而不是基于浓缩度的度量（如 FDR 或 P 值），使得声明的峰的数量在实验中更具可比性[7]。此外，IDR 分析还可用于比较和选择 peak callers[8,23]以及识别低质量的经验[7]。

5.差异绑定分析

随着 NGS 项目的稳步增加，对不同条件或组织中越来越多的蛋白结合区域进行比较性 ChIP-seq 分析。例如，ChIP-seq 实验的时间或发展设计可以为同一 tf 提供不同的结合信号快照，揭示基因调控的阶段特定模式[27,28]。考虑到这一点，我们应该注意到两组峰的简单二元重叠（例如 bedtools[29]）并不代表比较峰时的最佳方法[25]。

提出了两种备选方案。第一个定性方法对多个重叠的峰集[30]进行假设测试，因此扩展了上述两个重叠集方法。第二个定量分析提出了基于 peak 区域 read 总数或 read 密度的条件之间的差异绑定分析，即单个基因组位置的 read 重叠计数（表 S3 和[31,32]）。不建议直接计算没有控制的处理样品之间的差异结合区域（即使用其中一个作为对照），因为高富集区域可能由于人工制品或不同染色质结构而被识别，而不是由于真实结合事件。

通常，这两种方法都假定在每种情况下都预先独立地发现了显著的（见“peak calling”一节）和可再现的（见“再现性评估”一节）peak。为了提高检测差分结合区域的灵敏度（以增加假阳性数量为代价），可以使用更宽松的阈值来查找每个条件下的 peak。然后，根据生物学问题的不同，在任何一种情况下所调用的峰集都可以单独考虑，或者折叠成一个或多个有意义的共识峰区域列表。我们可以用定性的方法得到微分结合的初始视图。然而，在所有条件下确定的峰决不会仅仅基于峰的位置，通过这种方法被宣布为有差异的结合点[33]。定量方法适用于 read 计数（如 dbchip[33]）或 read 密度（如 manorm[34]）超过 peak 区域，并且具有较高的计算成本，但由于它提供了不同条件下（如 p 值或 q）差异结合的精确统计评估，因此建议使用这种方法。-链接到 read 丰富内容折叠更改的值）。强烈建议验证数据是否满足分析所选软件的要求。例如，dime[35]假设有很大一部分 peak 与比较项下的条件相同，manorm 假设在两种条件下相同的 peak 不会发生显著变化，而其他方法可能期望在条件[25]下出现恒定数量的 peak。重要的是，使用一些工具，只有两种条件可以同时提交进行比较（例如，manorm），并且一些条件可能会更好地取决于 ChIP 化的蛋白质（例如，组蛋白标记的 chipdiff[36]和 RNA pol II 的 polyphemus[37]）。

6.peak 注释

注释的目的是将 ChIP-seq 峰与功能相关的基因组区域相关联，例如基因启动子、转录起始位点、基因间区域等。在第一步中，上传峰并 read（以适当的格式，例如，BED 或 GFF 用于峰、WIG 或 BED 正常大小的 read 覆盖图；见文本 S1 和[38-41]）到基因组浏览器，在那里可以手动检查区域，以寻找与注释基因组特征的关联。如果有可比较的数据（例如，chip-qpcr），可以将其与 ChIP-seqpeak 进行比较，并在浏览器中手动 read。也可以使用 bedtools 等软件包中的工具进行系统分析，以计算从每个 peak 到最近地标的距离（例如，tss），或者识别在给定 peak 距离内的基因。例如，使用 CEAS[42]或生物导电剂包 Chippeakano[43]获得的此类“位置分析”的输出可进一步与表达数据相关（例如，确定基因与 peak 的接近程度是否与其表达相关），或与基因的表达相关。Tology 分析（例如，确定 ChIP 蛋白是否参与特定的生物过程）。基因本体分析可以使用 david[44]、great[45]或 gsea[46]来完成。有时，会绘

制与特定注释特征相关的 read 密度图，并在不同样本之间进行比较，从而揭示它们之间的蛋白质结合模式差异[47]。

7.Motif 分析

motif 分析不仅可以识别 tf-ChIP-seq 峰中的因果 DNA 结合基序，而且还非常有用。当 ChIP 蛋白的基序已知时，基序分析为实验的成功提供了验证。即使事先不知道基序，通过基序分析确定大部分峰的中心位置基序，也表明试验成功。基序分析还可以识别其他蛋白质的 DNA 结合基序，这些蛋白质结合复合物或与 ChIP 蛋白结合，从而阐明转录调控的机制。motif 分析对组蛋白修饰 ChIP-seq 也很有用，因为它可以发现与这些标记相关的未预料到的序列信号。表 S4 和[48,49]列出了用于 motif 分析的公共工具的小样本。

Motif 分析应用于由 peak calling 算法识别的基因组区域。因此，motif 分析的第一步是以 fasta 格式组装一组基因组序列，对应于所有重要的 ChIP-seq 峰[50–54]。Motif 分析的第二步是 Motif 发现，建议将 peak 序列输入能够发现未对齐 DNA 序列中序列基序的两个或多个算法[55–58]，因为这些算法具有互补的优势和弱点。一些 motif 发现算法构成了执行多个 motif 分析步骤（例如，meme chip[57]和 peak motif[58]）的管道的一部分，包括基于单词的 motif 发现算法和 motif 浓缩算法，这些算法可以识别只存在于一小部分 peak 中的 motif。在 motif 发现后，使用 motif 比较软件[59,60]将发现的基序与已知的 dna 基序进行比较，如果知道 ChIPtf 基序（或其 tf 族）的结合基序，则可用于确认其存在。结果还将提供有关绑定到 ChIP 化 tf 附近的其他 tfs 的提示。接下来，中心基序富集分析将确定其他已知的 DNA 基序是否在 ChIP-seq 峰的中心（或顶点）附近富集[61]。对以基因组标志物为中心的区域（如转录起始位点与 ChIP 序列峰重叠）进行局部基序富集分析也很有用[61]。此外，motif 间隔分析检测首选距离和成对的 motif 的排列，这可以指示 tfs 之间的物理交互[62]。最后，motif 预测映射并可视化每个 ChIP-seq 区域的 motif 基因组位置[63,64]。在这一步中，发现或丰富的基序被用来扫描 ChIP 序列的 peak 区域，匹配的坐标被上传到基因组浏览器中进行可视化。

8.展望

ChIP-seq 的挑战需要新的实验、统计和计算解决方案。目前的进展将使 ChIP-seq 能够分析含有更少细胞的样本，大大扩展其在胚胎学和发育等领域的适用性，在这些领域中，大样本的成本过高或难以获得。Nano ChIP-seq 可以分析多达 10000 个细胞的样本[65]。同样重要的是修剪今天比实际转录因子结合位点宽得多的峰。这对于区分人工制品和真正的联合结合事件是必要的：大多数转录因子竞争、合作或与其他转录因子、转录机制或辅因子共结合。上下文相关的调节机制的作用可以从根本上不同于个体结合事件的作用[66]。为了解决这个问题，基因组定位系统（GPS）使用分段期望最大化算法来解决近距离 peak 问题[67]。定位窄峰的一种有希望的实验方法是 ChIP 外显子，该外显子使用噬菌体 L 外显子核酸酶来消化不与蛋白质结合的 DNA 片段末端[68]。

假阳性峰的数目可以通过实验和计算来减少。提高抗体特异性是一项长期的努力，尽管取得了显著进展，但仍有四分之一的组蛋白修饰抗体未能通过特异性测试[69]。消除大量假阳性峰的另一种方法是将调控结合位点限制在核小体缺失区域，这些区域可用于调控结合。这些区域通过 dnase-i 超敏性测序（dnase-seq）和类似技术绘制：Thurman 等人发现 94%的人转录因子结合位点属于 dnase 超敏区，只有少数例外，如转录因子 znf274、kap1 和 setdb1，它们也与封闭染色质结合[70]。假阳性 peak 也是由于大多数方法中使用的不现实统计模型产生的不现实 p 值（以及因此产生的 fdr）所致[71]。peak calling 的计算分析仍处于初级阶段，扩展了方法的多样性和特定于条件的性能[72,73]，因此我们建议使用几种方法进行 peak calling。

也许最重要的新进展与远端调控区的检测和分析有关，这些调控区序列较远，但通过DNA 弯曲在三维空间中靠近。为了揭示这种 3-D 的转录调控机制，出现了两种主要的技术：成对末端标记（chia-pet）的染色质相互作用分析（74）和染色体构象捕获分析（如圆形染色体构象捕获（4c）[75]或染色质构象捕获）。捕获副本（5C）[76]。

结合位点的生物学功能不一定由峰的可重复性或 fdr/idr 值（框注 8，[7,23,77,78]）来表示。这一问题在编码项目期间再次出现，该项目根据严格的质量标准[7]提供了前所未有的监管信息[66,79]。DNA 蛋白质结合是动态的，结合事件的测量强度取决于（除其他外）发生的样本中的细胞比例（通常不均匀），以及它在给定细胞中所占的时间比例。因此，“弱”结合位点，无论使用什么标志性阈值，都可能具有强大的生物学功能[80-82]。ChIP-seq 还检测到蛋白质（通过另一种蛋白质或复合物）的间接 DNA 结合，因此不包含基序的预测位点也可能具有功能。最后，绑定并不一定意味着函数，因此仍然需要使用附加信息（例如表达式或 chro-matin 构象数据）来可靠地推断单个绑定事件的函数[83]。

本文所讨论的各种实验和计算方法，通过调节转录、影响翻译和几乎所有生物过程，使我们对复杂网络的理解发生了逆转。

9.支持信息

文本 S1：基因组浏览器可视化的标准图形跟踪数据格式。(DOCX)

图 S1：使用链互相关法评估 read 质量。链的相互关系用不同链位移距离 k 下的正链和负链分布之间的皮尔逊关系计算。相互关系（图 A）通常在两个位移距离处达到 peak，一个对应于 read 长度，另一个对应于平均 fra。G 库的长度。两个峰的绝对高度和相对高度可用于评估 IP 富集程度。改编自 Landt 等人[7]。(TIF)

图 S2：用于评估 ChIP-seq 数据集可重复性的不可恢复率（IDR）框架。图 A 显示了两个重复的 ChIP-seq 实验中确定的峰的显著性分数的散点图。IDR 方法将峰分为可复制（黑色）和不可生成（红色）组，并计算每个峰属于不可生成组的概率。它根据这个概率对峰进行排序和选择，并计算出 IDR，即所选峰中不可恢复发现的预期速率。图 B 显示了按原始显著性评分对 peak 进行排序时，不同等级阈值下的估计 IDR。(TIF)

表 S1：ChIP-seq 中使用的 peak calling 器示例。该列表包括允许处理和后处理不同类型窄读富集区（峰）、宽富集区（域）和混合信号（如 RNA pol II ChIP-seq）的工具。(DOCX)

表 S2：ChIP 序列数据集比较分析的归一化方法。(DOCX)

表 S3 ：ChIP-seq 中差分绑定分析的软件包。该表显示了使用 ChIP-seq 数据进行差分绑定分析的可用算法示例。(DOCX)

表 S4：ChIP 序列 peak 的 Motif 分析软件工具及其应用。该表给出了公开的软件工具示例，用于对 ChIP-seqpeak 或附近基因进行每一个形成基序分析。工具按主要任务（“类别”）分组，复选标记指示每个工具执行的特定步骤。基于 web 的 motif discovery input size limits chipmunk:未知；completemotifs:500000 个碱基对；meme chip:50000000 个碱基对；peak motifs:无限制；citrome:5000 个峰。(DOCX)

方框 1。序列 read 的质量度量

通常，对 ChIP-seq 数据的预处理将类似于任何其他排序数据的预处理，并将评估原始 read 的质量，以确定可能的排序错误或偏差（fastqc 可用于数据质量概述）。phred-quality 分数用于描述每个序列标记中每个基调用的置信度，与错误概率呈对数关系，并可用于过滤低质量的 read。在这个过滤步骤之后，可能还需要修剪低质量的 read 结束（参见镰刀，https://github.com/najoshi/scille）。此外，库复杂度是 ChIP-seq 库的常见质量度量（来自编码工具的 preseq package[10]或 pcr 瓶颈系数[pbc]，

https://code.google.com/p/phantompeakqualtools/），库复杂度与许多因素有关，如抗体质量、交叉线 King，材料数量，超声波，或 PCR 过度扩增。后者可以通过系统地识别和删除冗余的读来纠正，这在许多 peak callers 中实现，因为它可以提高他们的特异性。读者可能会对 Galaxy 工具箱感兴趣，该工具箱提供了访问此处介绍的许多工具的权限[50]。

方框 2。read 计数的质量指标

在 ChIP-seq 实验中，链交叉相关分析[7]通过测量免疫沉淀（IP）片段聚类的程度来评估数据质量。这是基于以下观察而发展的：（1）一个高质量的 ChIP-seq 实验通常显示了在受感兴趣的蛋白质结合的位置上富集的 dna 序列标记的显著聚集；（2）在正向和反向链上富集的序列标记位于与结合点中心的距离取决于片段大小分布[9]。该方法通过计算两股线之间的相互关系来量化聚类程度，即，作为应用于两股线之一的位移（k）函数的股线特定 read 密度分布之间的皮尔逊相关性（图 s1）。互相关通常在对应于片段长度的移位和对应于 read 长度的移位处达到 peak。片段长度的互相关与背景互相关之间的比率，称为归一化链互相关系数（nsc），片段长度的互相关与 read 长度的互相关之间的比率，称为重新关联。相对链互相关系数（RSC）共同反映了 ChIP 序列数据中的信噪比。非常成功的 ChIP 实验通常有 NSC.1.05 和 RSC.0.8[7]，尽管 ChIP 序列数据中仍然存在不符合这些标准的重要生物学信息。读者可以参考[7]了解编码数据上所示的交叉相关的原型配置文件。

软件 chance[16]通过估计和比较抗体和背景下拉的 IPread 来评估 IP 强度，方法称为信号提取缩放[77]。对于每个样本，它首先将基因组放入 IP 和输入端的非重叠箱中，然后通过比较 IP 和输入端箱中标记计数的累积分布，将箱分为信号区和背景区。接下来，它根据每种类型区域中 read 的百分比分配，计算一个 p 值作为扩展的重要性。根据一组编码 IP 输入和输入实验对人体数据计算的经验 P 值分布，分别将这两类实验视为真阳性和假阳性，从而估计出 Q 值。因此，Q 值被解释为与编码数据比较的分数，编码数据在用户数据水平上显示差异富集，但结果是输入的技术复制。该软件根据 Q 值确定实验的成功率，并报告了一些描述性质量统计数据，如 IP 中平均标记密度相对于输入的增加百分比和被分类为信号区的基因组百分比。因为 Q 值是根据人类数据计算的，所以用户应该意识到，如果 Q 值是由其他有机体产生的，那么 Q 值可能不相关。

Chance 还提供了一个具有基因组覆盖的 IP 强度的图形化可视化，通过绘制容器覆盖的标签的经验累积百分比，这些标签按照 IP 和输入的 read 密度的增加顺序排序。通过检查和比较 IP 和输入曲线，可以确定质量问题，例如序列深度不足、放大偏差和 IP 浓缩不足。

方框 3。peak calling：目前点状源转录因子

由于点状源因子的 ChIP-seq 数据是最丰富的类型，因此大多数 peak calling 都是针对这些因子设计和微调的。现有的 peak callers 在信号平滑和背景建模方面存在差异。由于相互作用位点周围的 DNA 更容易被剪切，因此，在文库准备过程中，ChIPDNA 片段的末端会在 DNA 上形成足迹，其大小与蛋白质-DNA 相互作用的关系比与尺寸选择的关系更大。那些能够捕获这个实验特定信息的 peak callers 可以大大提高预测的准确性。例如，peak callersspp[9]和 macs[17]（版本 2）使用互相关来发现映射到正负链的 read 之间的延迟，即实际蛋白质-DNA 相互作用区域的大小。平滑之后，背景模型被用来直接从控制样本或基因组序列的特征（如 GC 含量或可映射性）中去除噪声（Beads[84]）。peak 最终被称为高于用户定义的信噪比水平。用于丰富区域（peak）统计评估的模型从泊松（csar[85]）、局部泊松（macs）、负二项式（cisgenome[56]）到零膨胀负二项式（zinba[86]），甚至扩展到更复杂的机器学习建模技术，如隐藏的 M。Arkov 模型（Hpeak[87]和 Bayespeak[88]）。

大多数 peak calling 算法都采用基于窗口的方法来检测 peak，因此可能会错误地合并附

近的绑定事件。为了提高绑定事件预测的空间分辨率，一些 peak callers 使用 peak 形状作为线索。peak 分割器（PeakSplitter）[89]可以在包含多个子峰的更广区域中查找局部最大值。GPS[67]建立了给定候选峰区域 ChIP 序列 read 分布的概率模型，以对附近的同种类型事件进行去卷积。R 封装的多峰和窄峰可以分析峰的形状，从而分别重新排列和缩小最终的峰列表。强烈建议将这些方法作为一般 peak calling 点源因素后的后处理步骤。

框 4.调峰：由于对表观遗传调控的兴趣增加

由于组蛋白标记的广泛富集区域，组蛋白修饰、DNA 甲基化和染色质重塑因子等表观遗传标记正在通过 ChIP-seq 进行探索。其中一些标记在狭窄的基因组区域（例如基因启动子处的 h3k4me3）中高度富集，并且可以使用适合点源因子的 peak callers（在框注 3 中讨论）。然而，大多数组蛋白标记往往具有更广泛的扩散和较弱的模式（例如，H3K27ME3）。一些 peak callers 专门设计用于从 ChIP-seq 数据预测广泛区域，包括 sicer[90]、ccat[91]、zinba 和 rseg[92]。其他 peak callers，包括 SPP、MACS（版本 2）和 PeakRanger[93]也可以通过使用它们的选项来增加“带宽”或放宽“peak cutoff”来与这种类型的 ChIP-seq 数据一起使用。

对于广义标记，富集模式应描述为“域”。“”而不是“peak”，因为没有明确定义的 peak。映射 read 模式的另一种表示是分层的：组合多个级别的扩展。例如，macs（第 2 版）和经文[94]（最初是为 rna-seq 设计的）可以在更广的范围内进行强富集的狭窄调用，以实现与域边界相关的弱富集。

方框 5。peak calling

混合信号也有一些因子（如 RNA pol II）在变异较大的区域与 DNA 结合。众所周知，一些 RNA-pol-II 复合物停滞，而另一些则随着活性转录而移动[95]。在第一种情况下，理想情况下应将数据视为点源因子，而在第二种情况下，应将数据视为具有宽标记的因子。理想的算法应该同时适应这两种模式，这意味着 peak calling 应该更通用。有些工具可以选择窄 peak 和宽 peak calling，例如 SPP、Mac、Zinba 和 PeakRanger。但是，如果仔细调整参数，任何适合于宽 peak 检测的算法都可以用于这类数据。

方框 6。标准化

无论是将一个 ChIP 样本与输入 DNA（声波 DNA）、peak calling 中的“模拟”ChIP（非特异性抗体，如 IgG）进行比较，还是在差异分析中将一个 ChIP 样本与另一个 ChIP 样本进行比较，都有线性和非线性归一化方法可使两个样本“可比”（表 S2）。尽管许多方法都侧重于对控制样本的规范化，但没有一种方法能够区分所使用的控制样本的类型。一种直观且常用的线性归一化技术称为序列深度归一化。在这种方法中，read 次数乘以一个比例因子，使不同样本中的总 read 次数相同（详见[9,96]。PeakSeq[24]中使用了对该方法的轻微修改，其中使用线性回归在区域（，10 kb）中估计了比例因子。许多其他现有的方法也使用标准化因子来线性缩放样本，重点是对照样本的标准化（参见 cisGenome[56]、macs[17]和 useq[97]）。在[98]中提出的另一种称为 rpkm 的缩放标准化方法（每百万次映射 read 的序列范围的每千基 read 数）调整了由于更高的 read 概率落入更长区域的偏差。

非线性归一化调整具有非线性趋势的偏差。在[28]中描述的方法中，使用局部加权回归（黄土）对数据的平均值和方差进行归一化。这是基于这样一个假设：生物条件变化的影响不会引起全球结合变化。例如，当比较具有不同疾病进展阶段的样本时，或在特定治疗前后的样本时（参见“差异结合分析”一节），可以应用此假设。这种非线性归一化的一个改进版本被实现为 manorm[34]，假设两种情况下的常见 peak 不经历全局变化。还开发了名为 polyphemus[37]的 R 包，实现了两种归一化方法：（1）非线性方法，如[28]所述；（2）分位

数归一化，使不同样本中的分布相同。目前，标准化问题尚未得到充分利用，尽管它们可能对结果产生重大影响[28,37,99]。

方框 7。重复 read

重复（相同的）的 read 是一个挑战，因为它们可能来自独立的 DNA 片段或单个片段的 PCR 扩增。在前一种情况下，重复 read 是信号，在后一种情况下是噪声（实验性人工制品）。一个安全的解决方案是根据测序深度保持每个基因组位置的固定点击数（考虑不同的链作为不同的位置），这样就可以获得更好的特异性（较少的假阳性 peak）。然而，在估计蛋白质对一个特定基因组区域的亲和力方面，考虑所有的点击更为合理。这可以在设计良好的管道中完成，在调峰前后有一定的步骤。例如，可以删除一定数量的重复项以称为自信峰，然后将重复项放回以优化这些峰的特性，如峰高和边界。

方框 8。不可回收发现率（IDR）

如果一组 peak 需要一对复制数据集，则可以根据显著性标准（如 p 值、q 值或折叠富集）对 peak 进行排序。显著峰通常在重复序列中比显著性较低的峰排列更一致。这提供了从实际信号到噪声的转换指示。IDR[23]通过将峰分为可重复和不可重复组来量化这种转变，其中可重复组中的峰应比不可重复组更高、更一致地在重复组中排列（图 S2）。它为每个信号分配一个再现性指数，该指数估计其可再现的概率，并以类似于错误发现率（FDR）的方式报告选定 peak（称为 IDR）中不可再现发现的预期速率。计算 IDR 的 R 包在[23]中给出。使用编码数据说明的原型示例可在[7]中找到。使用 IDR 时，建议使用相对宽松的 peak calling 阈值，因为 IDR 算法需要对信号和噪声分布进行采样，以评估 peak 的再现性。

IDR 方法的一个主要优点是它独立于 peak calling 算法，并且可以跨实验室和平台应用于各种显著性标准。已经证明，它产生一个稳定的阈值，在实验室、抗体和分析方案（例如，peak calling 者）中比 FDR 测量值更一致[7]。