科学通报 2010 年 第 55 卷 第 14 期：1347 ~ 1357
专题：生物信息学 论 文 www.scichina.com csb.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

基于二次打断 IPed DNA 片段 ChIP-Seq 的模拟分析

王薇，施小龙，陆祖宏*

东南大学生物科学与医学工程学院，生物电子学国家重点实验室，南京 210096
* 联系人, E-mail: zhlu@seu.edu.cn

2009-04-30 收稿, 2009-11-25 接受
国家自然科学基金(30871393)和国家高技术研究发展计划(2006AA020702)资助项目

摘要 ChIP-Seq 是在全基因组水平上研究活体细胞中蛋白质和 DNA 相互作用谱的有效手段. 近年来, 随着高通量短序列 DNA 测序技术的快速发展, 研究基于新一代 DNA 测序方法的 ChIP-Seq 分析算法已经成为热点之一. 然而, 目前报道的分析方法主要是基于对免疫共沉淀获得的 DNA 片段进行片段大小选择后的 ChIP-Seq 数据, 也就是主要针对 Solexa 系统获得的数据进行分析的算法. SOLiD 系统是目前测序通量最高的新一代 DNA 测序系统. 在 SOLiD 系统的 DNA 测序文库制备过程中, 采用对免疫共沉淀获得的 DNA 片段进行二次超声打断可以满足 ePCR 对序列长度的要求, 因此 SOLiD 测序文库中的 DNA 测序片段较短. 到目前为止, 基于 SOLiD 系统测序特点的 ChIP-Seq 研究很少报道. 本文旨在研究测序文库中 DNA 片段的长度对 ChIP-Seq 分析的影响. 通过真实的 ChIP-seq 数据和模拟产生的 ChIP-Seq 数据, 对目前 3 种主要的 ChIP-Seq 分析方法(CisGenome, SISSRs 以及 MACS)的特点进行研究. 有报道表明来自 Solexa 系统的 ChIP-Seq 数据局部有明显的正负链双峰特征, 而通过对真实的来自 SOLiD 系统的 ChIP-Seq 数据特征的挖掘, 我们发现单个峰局部无明显的正负链双峰特征, 并且峰的局部的序列分布大部分符合正态分布. 基于这些特征, 我们模拟了两个不同测序平台的 ChIP-Seq 实验. 在控制了模拟实验的可比性后, 我们发现当前基于 Solexa 文库制备方案的 ChIP-Seq 数据发展的算法, 并不能有效地捕获来自 SOLiD 系统的 ChIP-Seq 数据特征. 我们的研究还表明, 误用 ChIP-seq 软件可能是导致部分 SOLiD 的 ChIP-seq 实验失败的原因. 因此, 需要开发一种新的基于二次打断 IPed DNA 片段的 ChIP-Seq 分析策略.

关键词
蛋白质与 DNA 相互作用
下一代测序技术
序列方向性
乳液 PCR
ChIP-Seq
SOLiD

染色质免疫共沉淀(ChIP)是一种研究活体内蛋白质和 DNA 直接相互作用的有效手段. 在过去几年中, 人们建立了一些分析染色质免疫共沉淀后(IPed)的 DNA 片段的测序技术平台. 如 ChIP-SAGE[1], ChIP-SACO[2], 这些技术主要通过对串联 IPed DNA 片段进行测序来寻找转录结合位点. 然而, 这些技术都依赖于传统测序技术, 对于大多数实验室和研究人员来说, 通过上述技术去获得某个感兴趣的蛋白质在全基因组范围的转录活性都过于昂贵和费力. 因此, 尽管 ChIP 技术和微阵列芯片技术(ChIP-chip)的结合存在精度低和检测偏性等问题, 但由于其高通量和可接受的价格, 它仍然得到了较多的应用.

近年来新一代高通量测序技术发展迅速, ChIP 实验与这种高通量并行的测序技术的结合(ChIP-Seq)正开始转变蛋白质-DNA 相互作用的研究现状. ChIP-Seq 可在全基因组范围定量分析蛋白质-DNA

英文版见: Wang W, Shi X L, Lu Z H. Simulation of ChIP-Seq based on extra-sonication of IPed DNA fragments. Chinese Sci Bull, 2010, 55, doi: 10.1007/s11434-010-3013-y

的结合以及染色质的修饰. 与 ChIP-chip 相比, ChIP-Seq 技术的主要优势如下: 单个碱基精度的直接测序, 较少的背景噪声, 无需预先确定结合区域, 较少的起始 DNA 量的要求[3,4]. 已有报道使用 ChIP-Seq 技术进行组蛋白修饰在人类 T 细胞[5]和鼠的胚胎干细胞[6]的定位, 研究转录因子 STAT1[7]和 NRSF[8]的结合位点.

目前, 有 3 种商业化的下一代高通量并行测序平台, 它们分别是 454(Roche), Solexa(Illumina) 和 SOLiD(ABI Life technologies). 在这 3 个测序平台中, 454 的测序读长为 200~400 bp, 而 Solexa 和 SOLiD 的测序读长只有 30~50 bp. SOLiD 测序通量最大而 454 的测序通量相对较低. 通过基因组的可比对性分析, 30~50 bp 长的序列足以唯一的匹配到~79.6%~86.7% 的人类基因组[9,10](NCBI build 36.1/UCSC hg18). 到目前为止, 唯有 Solexa 在 ChIP-Seq 领域被广泛报道. 尽管 SOLiD 系统可产生与 Solexa 系统的相同读长并拥有更高的通量, 它在 ChIP-Seq 领域的应用潜力却没有被完全挖掘. 最近, ABI 公司自己发表了一篇在 SOLiD 系统上进行 ChIP-Seq 研究的应用小短文[11]. 在该文中, IPed DNA 的起始量高达 0.5 μg, 而这个数量在很多 ChIP-Seq 实验中较难达到. 除此之外, 该实验选择了 3 个不同范围的 DNA 片段长度 (150~200 bp, 200~250 bp, 250~300 bp), 这个步骤同 Solexa 的片段大小选择一致, 但这种处理显然并不适用于 SOLiD v2 系统的文库制备指南. 因此, 在 SOLiD 系统上进行 ChIP-Seq 实验的优势并没有展现出来.

2008 年, Valouev 等人报道了基于对 ChIP-Seq 数据的序列方向性的处理算法, 对早期的算法进行了较大的改进. 这些改进包括算法中参数的自估计以及结合位点在基因组上定位的精度[12~15]. 方向性来自 IPed DNA 片段长度(150~300 bp)和测序读长(26~35 bp)的差异. 150~300 bpDNA 片段的两端可以等概率地测出, 但是测序读长无法跨越整个 DNA 片段长度, 因此一个真实的结合位点附近的序列密度理想状况下可呈现两个对称的峰. 正链的序列富集峰在上游, 而负链峰在下游. 因此, 基于对上述方向性处理的不同, 算法可大致分为两类: 搜峰前估计类和搜峰后优化类. 如 QuEST[12], SISSRs[13]和 MACS[14]是典型的搜峰前估计类算法, 使用序列的方向性信息来平移序列的位置或估计片段长度从而精确定位转录结合位点. 与此不同的是, CisGenome[15]提出了二个优化的功能, 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, 实现了结合位点的精确定位. 然而, 序列的方向性信息仅来自用桥式 PCR 进行较长 DNA 片段文库制备的 Solexa 系统. 根据 SOLiD 的操作手册, 进行乳液 PCR 的最优扩增的 DNA 片段长度范围为 60~110 bp, 因此当前基于序列方向性的分析 ChIP-Seq 的算法可能不再适用于分析来自 SOLiD 系统的数据.

本文考察了 Solexa 和 SOLiD 系统在文库制备方面的 DNA 片段长度差异. 我们挖掘了来自 SOLiD v2.0 系统的 ChIP-Seq 数据的局部序列分布特征. 基于这两种系统的不同的文库制备方案所带来的真实结合位点的链特异性的序列分布情况, 模拟产生了两批 ChIP-Seq 数据. 用 3 个算法 SISSRs, MACS 和 CisGenome 对模拟数据进行了分析并比较了其结果. 我们讨论在 SOLiD 系统上进行 ChIP-Seq 实验的可能性及其潜力.

1 材料和方法

虽然测序成本已大大下降, 但是进行一次 ChIP-Seq 实验的成本仍然不低. Zhang 等人[16]在计算机上模拟的 ChIP-Seq 实验信号与真实的 ChIP-Seq 的信号的对比证明了 ChIP-Seq 在全基因组范围的信号分布非均匀分布, 这对处理 ChIP-Seq 的算法设计提供了一个重要的信息, 同时也提供了模拟实验的方法依据. 本文主要考察的是算法的通用性问题, 通过在计算机模拟实验数据的方法较为经济可行. Solexa 的 ChIP-Seq 富集区域的序列分布的局部特征已由前人研究, 本文挖掘了 SOLiD 的 ChIP-Seq 富集区域的序列分布的局部特征, 作为模拟实验的依据. 在模拟实验的过程中, 严格控制了一些参数, 预期能够准备地判断适用于 Solexa ChIP-Seq 数据的分析策略是否能用到 SOLiD 的 ChIP-Seq 数据中.

1.1 SOLiD 和 Solexa 平台下 ChIP-Seq 示意图

两个测序平台下的 ChIP 步骤是相同的. 首先, 用甲醛将活体细胞交联, 随后将细胞核内的染色质用超声打断成目标长度(通常 0.2~1 kb)的短片段形式(图 1). 用所研究的蛋白特异性抗体将蛋白质结合的 DNA 片段免疫共沉淀, 接下来将蛋白质-DNA 复合物解交联. 如果是在 SOLiD 系统上进行 ChIP-Seq 实验中测序部分的实验, 分离纯化后的 DNA 片段需进

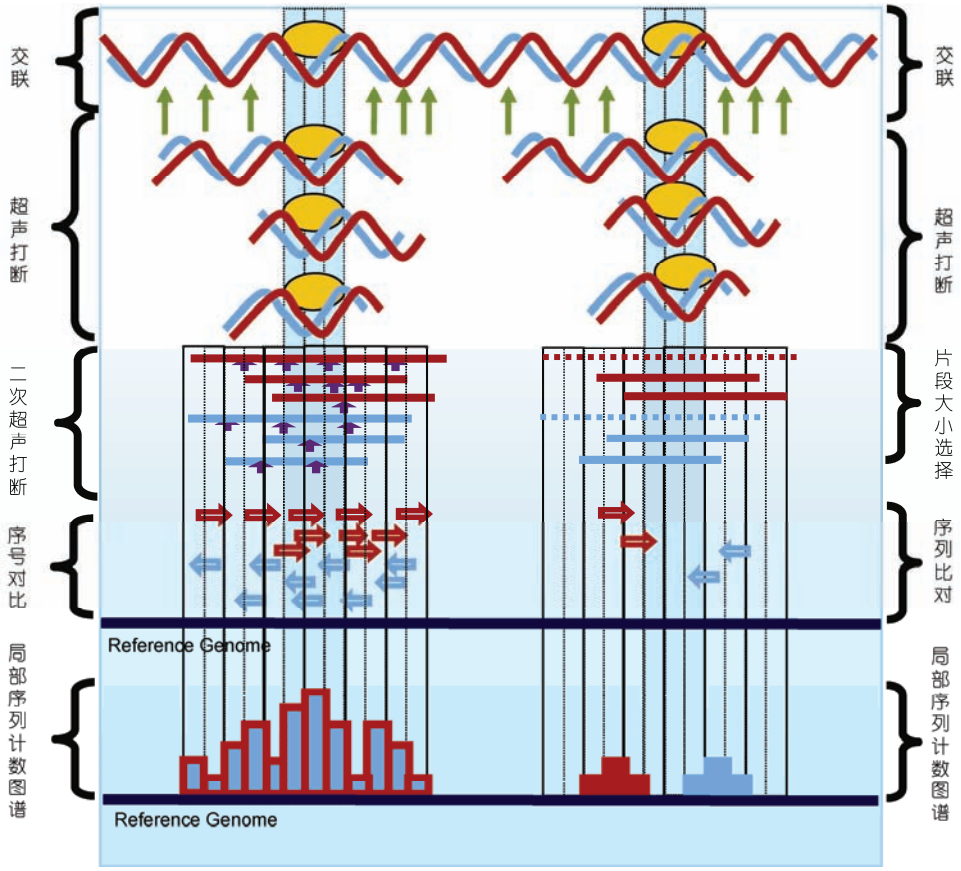

图 1　SOLiD(左)和 Solexa(右)平台下 ChIP-Seq 示意图

将交联的染色质用超声随机打断至所需的短片段长度(通常为 0.2~1 kb), 蛋白质的结合位点因蛋白质的保护作用而得以保留在短片段中. 将蛋白质-DNA 复合物解交联并纯化 DNA 片段后, 对于 SOLiD 测序系统, 为满足其文库制备要求需将 DNA 进一步用超声打断到 60~110 bp 左右的小片段, 而对于 SOlexa 系统, 一般则是对 DNAs 进行 150~300 bp 范围左右的片段大小选择, 超出此范围的 DNA 片段则被舍弃(虚线). 所测序列首先进行基因组比对然后用不同的算法进行结合位点的统计. 对 SOLiD 系统来说, 正链序列和负链序列应随机分布在富集的区域的每一个位置, 在蛋白质结合的位点的密度最高. 而对 Solexa 系统来说, 富集区域将呈现出典型的双峰模式, 正链序列集中在上游而负链序列集中在下游

行再次超声打断到约 60~110 bp, 以满足文库制备中乳液 PCR 的要求, 这一步区别于 Solexa 系统下的直接片段大小选择步骤(图 1). 在文库制备后, 在 SOLiD 测序仪的流动池中进行测序, 所测序列然后在参考基因组上进行比对. 如图所示, 相比于 Solexa 系统上的 ChIP-Seq, SOLiD 系统在文库制备过程中增加的短片段可产生更多测序序列, 原因是那些较长片段在再次超声过程中被打断而不至于被舍弃. 此外, SOLiD 系统上的一个真实结合位点的局部序列分布可表现出正链峰的模式与负链峰的模式相同. 不同的高级分析算法用来富集结合区域, 这些区域即可被认为是所研究蛋白在基因组上的候选结合区域.

1.2　数据集

来自 Solexa 系统的转录因子 NRSF 的 ChIP-Seq 数据从 NCBI, GEO(Gene Expression Omnibus)数据库(GSE13047)下载[8]. 考虑到大多数的 ChIP-Seq 实验无法达到 70 次染色质免疫共沉淀的起始 DNA 量, 因此本研究仅使用了实验二(4 次染色质免疫共沉淀加 PCR 扩增)的 ChIP-Seq 数据.

来自 SOLiD 系统的转录因子 PU.1 的 ChIP-Seq 数据由本实验产生. 从 K562 细胞系大约获得 50 ng IPed DNA 进行文库制备. 文库制备的过程如同 1.1 节所描述的那样再加上一个 PCR 预扩增的过程. 并且, 对 DNA 片段做了二次超声的处理, 将其打断到 60~110 bp 的长度, 接下来这些更短片段的 DNA 的两端被接上连接子(adaptors)在 SOLiD v2 系统上进行测序. 经过图像处理过程总共获得 21253931 条序列(35 nt). 将序列与 hg18 人类染色体组进行比对, 最多允许颜色编码空间的 3 个错配. 获得 7189990

(33.83%)条可比对到多个位置的序列, 经过滤, 留下 5071807 条唯一匹配的序列进行后续分析.

虽然转录因子 NRSF 的 ChIP-Seq 数据来自 Solexa 系统, 转录因子 PU.1 的 ChIP-Seq 数据来自 SOLiD 系统, 但因为二者分别来自不同的细胞系和不同的转录因子, 通常不同的转录因子在某个特定细胞形态的全基因组范围的靶点数并不一致, 因此所产生的数据无直接的可比性, 所以本文考虑利用不同系统的 ChIP-Seq 数据的特征来进行模拟实验, 从而考察针对 Solexa 系统的 ChIP-Seq 实验的分析方法 CisGenome, SISSRs 和 MACS 是否适用来自 SOLiD 的 ChIP-Seq 数据.

1.3 真实的 SOLiD ChIP-Seq 数据特征提取

真实的 SOLiD ChIP-Seq 数据主要用来查看富集区域的正链的峰和负链的峰之间是否有明显的位移差, 局部的单链序列分布是否服从正态分布.

根据文库制备过程以及短序列测序仪的特点我们假设来自 SOLiD 的 ChIP-Seq 数据其富集区域的正链峰和负链峰应重叠. 由于 CisGenome 有支持单样本分析的优势和无事先处理序列方向性的步骤, 为检验这一假设, 我们选择 CisGenome 分析来自二级比对产生的 5071807 条唯一匹配的序列, 参数设置如下: 滑动窗口大小设为 100 bp, 步长 25 bp, 富集一个峰的序列数的阈值首先采用默认值 10, 在这些参数设置下总共获得 1371 个富集区域. 我们统计了所有富集区域的正链峰和负链峰的位移差的频数. 随后, 考虑到 CisGenome 提出的用负二项分布模型计算的 10%的 FDR 水平[15], 我们采用了一个更加严格的阈值, 即每个窗口范围内至少满足 24 条序列, 这样有 281 个区域富集出来.

根据实验过程我们假设来自 SOLiD 的 ChIP-Seq 数据其富集区域的局部序列分布服从正态分布. 为检验这一假设, 我们依照链的方向将富集区域的测序序列分开, 然后用 100 bp 的窗宽和 25 bp 的步长统计富集区域的单链序列分布情况. 将所统计的序列分布进行正态分布验证. 此验证使用 Matlab 7.0.4 平台上的 lillietest 算法, lillietest 可检验在无定均值和定方差的情形下随机变量 X 是否服从正态分布, 支持小样本分析.

1.4 模拟

本次实验进行的模拟实验的序列数量关系参考的是转录因子 NRSF 的 ChIP-Seq 数据, 在所有的 1697893 条序列中, 有 343234 条序列富集了 2171 个可能的结合位点. 为了具有可比性, 模拟两个不同平台的 ChIP-Seq 过程中将上述数量设为相同.

ChIP-Seq 数据的全局特征曾被仔细分析过[16], 并且提出了一个在计算机上进行 ChIP-Seq 实验的方案. 除了结合位点内部的序列分布模拟, 我们采用了其中的一些步骤, 如去除基因组的盲区和重复区, 随机选择结合位点, 添加服从 gamma 分布的背景噪声和服从幂函数分布的真实信号.

基于我们的假设, 这两个系统的 ChIP-Seq 数据的最主要的差别在于局部序列的方向性分布上. Solexa 系统的 ChIP-Seq 数据的该特征已有前人研究[12~14], 而 SOLiD 系统的该特征则通过本次实验真实的 ChIP-Seq 数据来挖掘. 为模拟 Solexa 测序仪在某个结合位点局部双峰现象, 我们选择用对称的两个 gamma 分布来分别模拟正链序列和负链序列的分布情况(图 2). 不同长度的结合位点的单链序列的密度最高处不同, 从而可模拟不同片段长度的正链峰与负链峰之间的不同的位移差. 由于进一步超声, 我们假设 SOLiD 系统正负链序列的密度最高处应出现在同一个碱基位置上, 且真正的结合位点为正负链序列重叠的最高点, 因此我们用相同的均值在结合区域正中间的正态分布来模拟正负链的序列分布.

1.5 转录因子 NRSF 和模拟的 ChIP-Seq 数据分析

因为 CisGenome, SISSRs 和 MACS 在其发表时都有测试 NRSF 数据集, 因此, 在分析 NRSF 数据时, 参数都设置为默认值. 然而, 对于模拟 Solexa 系统的 ChIP-Seq 数据, 当用 MACS 分析时, mfold 这一参数必须设为 20 以满足其足够的配对峰的数量的要求, 而其他所有的参数也取的是默认值. 序列簇集计数用的是一个 perl 脚本[16], 其中片段长度参数设置为 100 bp.

2 结果和讨论

2.1 ChIP-Seq 进一步超声的优势

ChIP-Seq 实验所适用的片段文库制备的方案是低输入量 DNA 方案. Solexa 和 SOLiD 系统在片段文库制备方面最主要的差异是扩增步骤. Solexa 采用的是桥式 PCR 而 SOLiD 选择用乳液 PCR 增加 DNA 克隆群的拷贝数. 尽管桥式 PCR 和乳液 PCR 都不能扩增

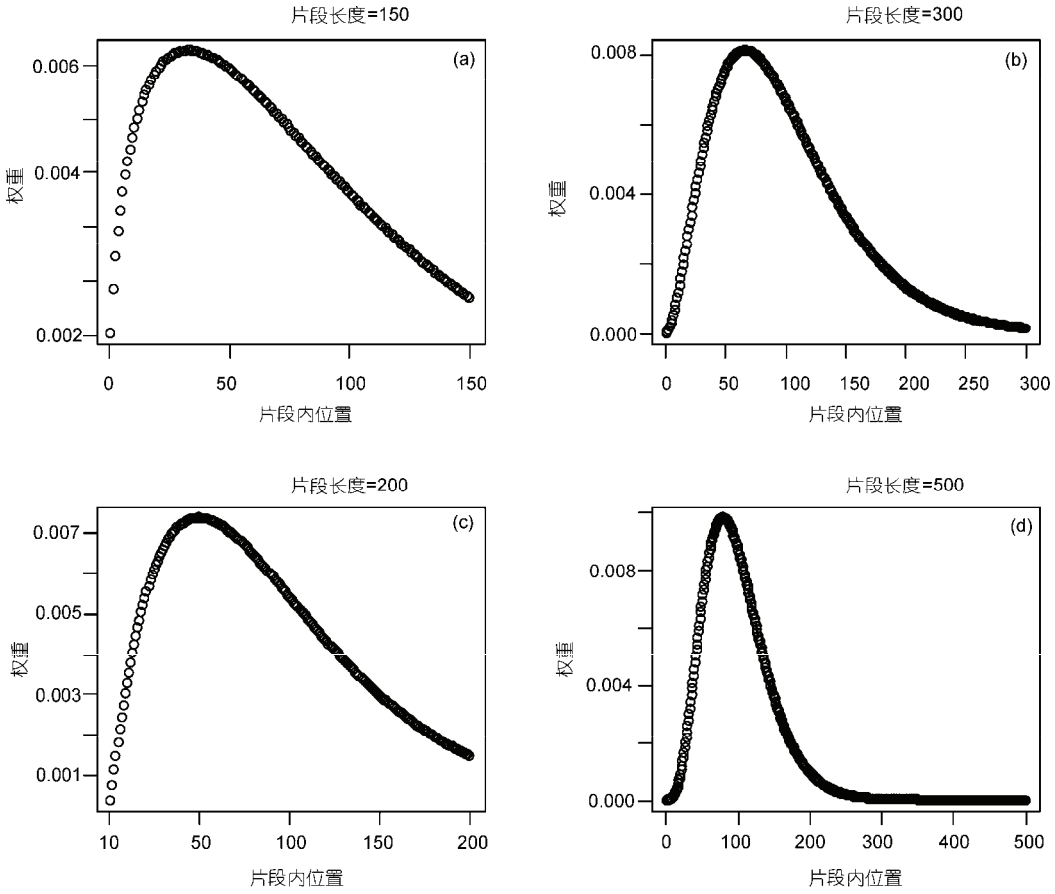

图 2　模拟正链序列分布函数

不同的片段长度可在不同点达到序列密度最大

太长的片段，但一个值得注意的差别是，乳液 PCR 相较于桥式 PCR 对片段长度的要求更为严格. 乳液 PCR 的扩增效率与扩增子长度的数量关系已有研究[17]. 那些较短片段的产物的荧光强度明显比那些长扩增子的强度要高. 不同的是，桥式扩增的有效片段长度范围相对较广. 由于 IPed DNA 的长度在 200~2 kb 不等，因此 Solexa 可能更适于 ChIP-Seq 的研究.

由于分离和纯化后的 IPed DNAs 的片段长度不一，文库制备前的额外的 150~300 bp 的片段大小选择可增加空间位置的精度，也可为 Solexa 测序平台提供最优的测序底物. 窄一些的片段大小选择可增加分子克隆群的大小均一性，从而增加 Solexa 平台的测序序列的得率. 另外，Johnson 等人[8]发现 DNA 片段长度可影响最终序列的输出结果，其中短序列的测序效率更高. 这种现象也有其他报道[14]. 一个模型被用来计算 DNA 片段的平均长度，发现所估计出来的长度都小于 150 bp. 也就是说，在一个给定的 DNA 输入样本中，较短的 DNA 片段在测序结果的出现概率更高，由此看来短序列测序仪如 Solexa 可能更倾向测 IPed DNA 池中的短片段. 这对于在 SOLiD 系统上进行 ChIP-Seq 来说是一个有希望的预测.

已有的 ChIP-Seq 实验都跳过文库制备中的 DNA 片段化的步骤，直接进行 DNA 末端修复. 然而，如果在 SOLiD 系统上进行 ChIP-Seq 实验，根据 SOLiD v2 的片段文库制备指南，将 DNA 进一步进行打断将有助于后续实验. 在这个二次打断过程中，原来 200~2000 bp 长的 DNA 的可被打断到更短，其中 60~110 bp 的长度对乳液 PCR 来说最为合适. 由于乳液 PCR 的限制所驱使的二次超声似乎给整个 ChIP-Seq 实验及接下来的数据分析带来了麻烦. 但是，事实上，这可能有助于研究抑制性转录因子和增加序列的最终得率，而这两个问题是在 ChIP-Seq 研究领域中亟

需解决的难题. 一方面, 如上所言, 短序列测序仪, SOLiD 也归为这一类, 倾向于对短的 DNA 片段测序. 如果原始 DNA 样本进一步被打断成小片段, 那么小片段可在乳液 PCR 扩增中易被扩增也易被测序, 可增加序列的得率. 另一方面, 抑制性转录因子结合结构相对较紧密的染色质, 这一部分 DNA 不容易被超声打断到易溶解的提取物. 因此, 来自抑制性因子的 IPed DNA 比来自开放染色质区域的 DNA 要长, 在片段大小选择和测序中被淘汰的概率更高. 例如, H3K27me3 和 H3K9me3, 这两个抑制性的组蛋白修饰, 原应在细胞分化时同激活性修饰拥有一样的 ChIP-Seq 效率, 反常地表现出调控较少的位点[14]. 如果在解交联后(图 1)它们被进一步超声打断, DNA 的长度可能不再是一个障碍.

至于为何 ABI 公司可成功地在 SOLiD 系统上进行起始 DNA 片段长度超过建议的 60~90 bp 范围甚至到达 300 bp 的 ChIP-Seq 实验, 其中的一个原因是 SOLiD v3 系统支持更大的长度范围, 但是 ABI 如何升级其 SOLiD 系统并非本文的议题. 如果一个转录因子结合启动子区域的连续两个区域, 较宽的长度(200~300 bp)可囊括这些复杂的结合现象. 同样的, 如果算法设置得当的话二次超声也不会使这类信息丢失.

尽管在 SOLiD 系统上进行二次超声打断 IPed DNA 的测序有如此潜力, 但本实验所提到的 ChIP-Seq 实验并没有对照实验, 即将 IPed DNA 只进行一次超声打断, 如同那些在 Solexa 上进行的实验一样. 然而, 这可由模拟实现. 由于 SOLiD 和 Solexa 的富集区域的局部序列分布特征已被挖掘, 基于此我们仔细模拟了两个数据集, 一个模拟来自 Solexa 的, 另一个模拟来自 SOLiD 的, 用来测试基于 Solexa ChIP-Seq 数据的分析方法是否同样适用于来自 SOLiD 的 ChIP-Seq 数据.

2.2 SOLiD ChIP-Seq 数据的局部正链和负链的峰的特点

用 CisGenome 对 5071807 条唯一匹配的序列在较松的阈值条件下进行搜峰后得到 1371 个富集区域, 该软件同样计算了 1371 个负链峰和正链峰的位移差. 如图 3 所示, 位移差大多集中在±10 bp 左右, 位移差为 1 bp 的占 39%, 证实了在二次随机超声后双模式峰之间的明显位移差的消失.

使用 CisGenome 的浏览器的功能, 可观察富集区域内的单链峰上的序列分布情况. 我们假设其服从正态分布, 这个假设由正态分布检验证实(表 1). 除去那些特别窄的富集区域以至无法形成足够的样本量(Null)进行正态检验, 830 个富集区域中的~70%的单链序列分布都被正态分布假设检验接受. 当富集标准设置得更为严格时, 该比率达到了 80%. 因此, 在模拟 SOLiD 系统的 ChIP-Seq 实验时我们采用局部序列服从正态分布的假设.

2.3 模拟的序列的全局和局部分布图谱

ChIP-Seq 实验的全局序列分布可由幂函数成功模拟真实的结合信号, gamma 分布模拟复杂的背景噪声. 由于在处理来自实验的 NRSF 数据时, 其基因组范围和单个染色体范围的序列频数图谱(图 4)与转录因子 STAT1 的 ChIP-Seq 数据十分类似, 因此本文也采用了这一模拟策略. 该图谱的特征是, 中小序列频数簇集占较大比例而大和极大序列频数的簇集仅占很小一部分[16]. 此外, 除了由总的富集区域数引起的小差异外, 模拟的 NRSF ChIP-Seq 数据与真实的数据的序列簇集规律类似.

来自 Solexa 系统的局部的正链和负链的序列频

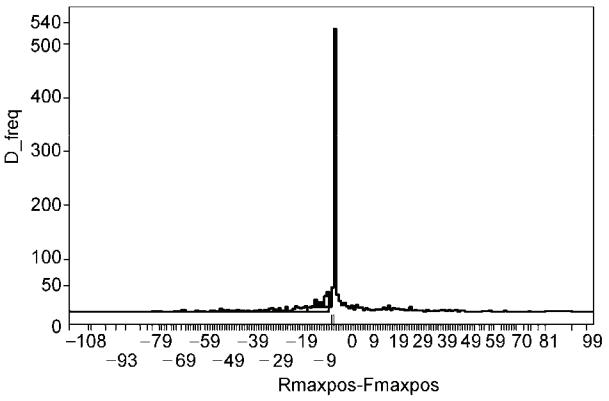

图 3 富集区域负链峰和正链峰距离的频率图谱
两个小竖条分别表示 0 和 1

表 1 富集区域单链和双链的序列正态分布检验结果

W100/S25	阈值	Alpha	接受(比例)	拒绝	Null
正链	10	0.01	560(67.5%)	270	541
正链	24	0.01	208(82.5%)	44	29
负链	10	0.01	578(70.1%)	247	546
负链	24	0.01	204(81.3%)	47	30
正负链	10	0.01	637(71.7%)	251	483

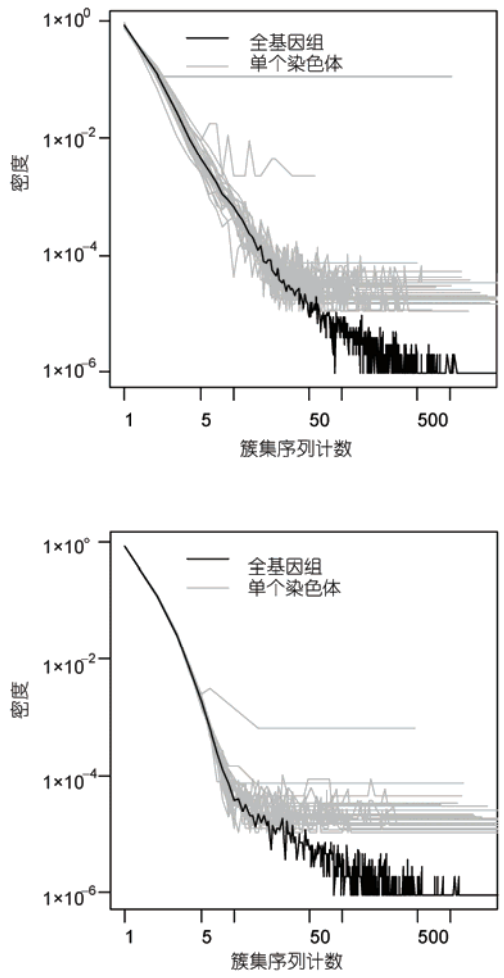

图 4　真实和模拟的 NRSF ChIP-Seq 数据的簇集序列计数图谱
(a) 真实 NRSF ChIP-Seq 数据全基因组和单个染色体上的簇集序列计数及其百分比; (b) 模拟的 NRSF ChIP-Seq 数据全基因组和单个染色体上的簇集序列计数及其百分比

数的带位移差的分布由 gamma 分布施加的方向性权重实现(图 2). 值得一提的是, 由于各种偏差和噪声局部的真实的 ChIP-Seq 信号可更为复杂(图 5(a), (c),(e)黑色柱状图). 因此, 它们通常不会显示出一个平滑的数学意义上的正态分布信号图谱(图 5(b),(d),(f)黑色柱状图). 然而, 我们真正关注的是其正义链和反义链上的序列图谱及其相对位置. 没有一个算法特别地要求一个完美的正态分布信号图谱来富集一个区域.

2.4　比较 CisGenome, SISSRs 和 MACS 对真实的 NRSF 数据的分析结果

使用 3 种不同的算法处理同一数据集, 我们可以发现 3 个算法所计算出的片段长度估计和总峰数完全不一致. 对真实的 NRSF 数据来说, CisGenome 估计出来的平均片段长度是 60 bp, 短于 MACS 估计的 68 bp, 远短于 SISSRs 估计的 124 bp. 这些差异可由这三个算法的原理来解释. 对 CisGenome 来说, 长度估计并非其直接的和有意义的结果. 事实上, 在某种程度来说, 片段长度估计由扫描全基因组的滑动窗口大小决定[15]. 这个参数在 CisGenome 里设置得偏小, 如 CisGenome 的用户指南上所言, 起始的测试窗口大小可设为 100 bp, 这是因为在设这个参数时没有嵌入任何的实验方面信息. 与之不同的是, MACS 在设置这一参数时将超声打断的平均长度融入其模型构建当中[14]. 这个参数大小的默认值是 300 bp, 可随不同的 ChIP 的结果而由用户自行设定, 比 100 bp 大几倍. 我们在分析中使用的全是默认参数, 由于 MACS 假设的结合位点在建模 ChIP-Seq 序列的平移距离时要宽, 因此 MACS 所得的长度估计比 CisGenome 的要长是合理的. 而 SISSRs 所估计的片段长度, 不是由窗口扫描而来的. 取而代之的是, 它选择了一种点到点的方式, 避免了由窗口扫描引入的位置偏差[13]. 简单来说, 当考虑到大多数交联的染色质不能被超声打断到小于一个核小体的长度时, 124 bp 的片段长度是更容易让人接受的, 核小体的典型结构是由 146 bp DNA 缠绕 4 组核心蛋白而构成的一个八聚体[18].

使用默认参数并且没有经过对照 ChIP 的假阳性过滤, CisGenome, SISSRs 和 MACS 分别找到 2534, 4937, 7195 个峰(表 2). 没有经过进一步验证, 如富集区域的 motif 分析, 将 ChIP-Seq 的结果与前人的 ChIP-chip 的结果比较从而查看重叠率, 或者是 qPCR 检验, 这些峰的绝对数量是毫无意义的, 因为更多的峰可能表示更多的区域可能是假阳性结合位点. 将这些数据列于此是为了给处理模拟数据提供一个参考, 以期经过这 3 个算法搜峰处理后得到同样的一个趋势.

2.5　比较 CisGenome, SISSRs 和 MACS 分析模拟的 Solexa ChIP-Seq 数据和 SOLiD ChIP-Seq 数据

模拟序列的方向性的方法已在前文中有所阐述, 局部序列的方向性的差异见图 6. 一个富集区域的配对峰由对称的两个 gamma 分布模拟序列方向实现

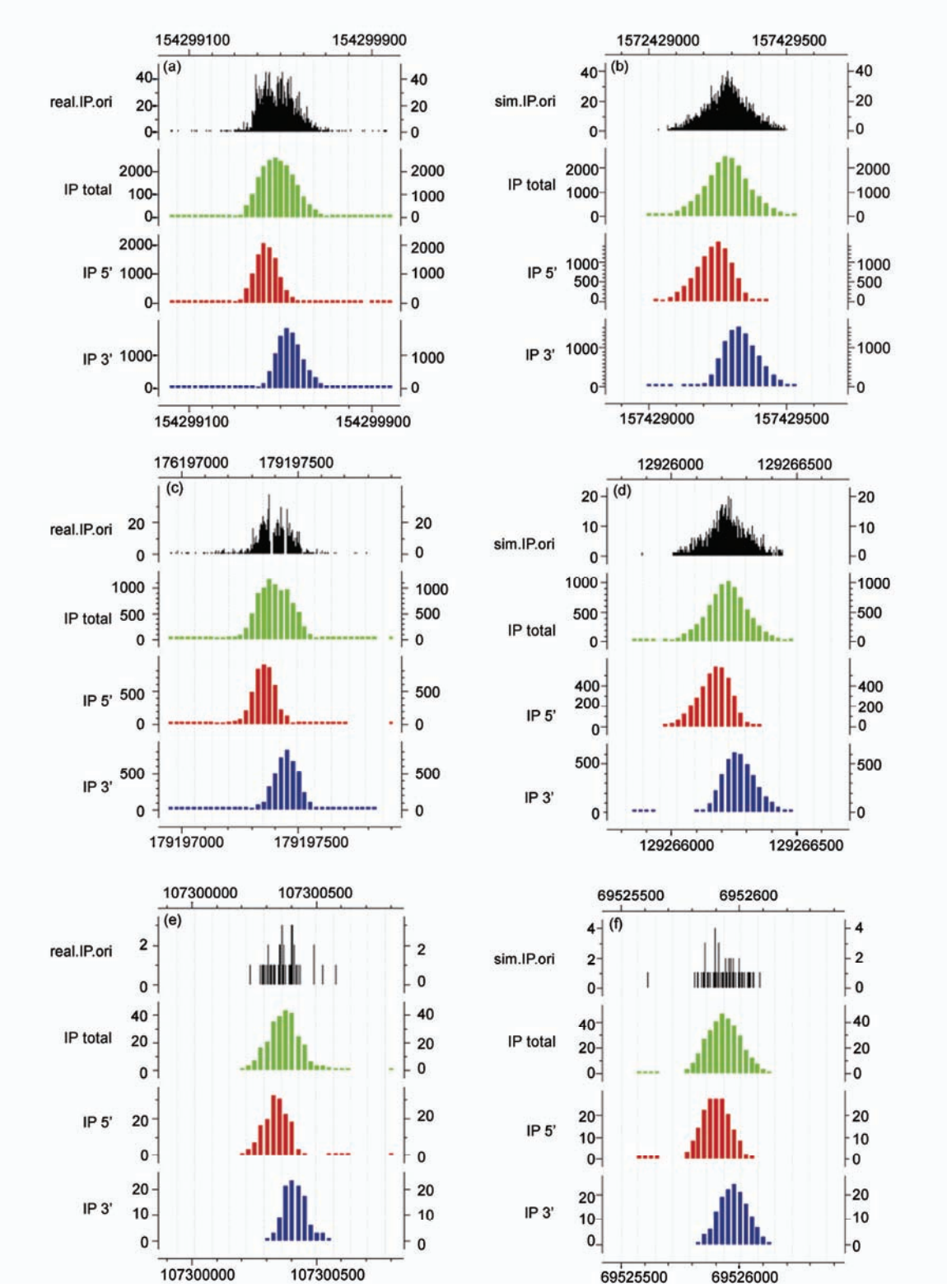

图 5 真实和模拟的 Solexa ChIP-Seq 序列方向性分布的比较

(a), (c), (e)是真实的 ChIP-Seq 富集区域在不同的总序列数下的局部序列方向性分布. (b), (d), (f)是模拟的 ChIP-Seq 富集区域在不同的总序列数下的局部序列方向性分布. 黑色柱状图是原始的序列频数信号, 绿色柱状图是富集区域内以滑动窗口为单位的序列频数信号;红色和蓝色分别为正链和负链的窗口序列频数信号

表 2　真实的 NRSF ChIP-Seq 数据的分析结果总结

	CisGenome	SISSRs	MACS
片段长度/bp	60	124	68
总峰数	2534	4973	7195

(图 6(a)). 相反, 我们控制 SOLiD 系统上的正链峰与负链峰重叠(图 6(b)).

在前文中的模拟部分我们提到, 为便于比较, 我们控制两批模拟数据的总峰数一致都为 2171, 由 343234 条序列富集. 如果全局结合信号服从均匀分布, 那么平均每个峰可富集到 158 条序列. 但是前人研究表明 ChIP-Seq 实验中由于实验固有的随机噪声, 结合位点不可能包含同等数量的序列, 事实上所获得的结合信号序列的全局分布是幂函数分布和 gamma 分布的叠加, 即中小序列频数的簇集占较大比例而大和极大序列频数的簇集占较小比例. 在这种情况下, 我们可能从分析中无法再复原真实的 2171 个峰, 但是由于服从 gamma 分布的噪声信号的贡献, 我们仍可能得到这个数量级的峰数, 只是对于真实的 ChIP-Seq 实验来说, 其中包含了许多假阳性. 在实验成本允许的情况下, 这种假阳性需要由 input DNA 或以 IgG 蛋白为抗体作为对照的样本来尽量剔除.

CisGenome, SISSRs 和 MACS 3 个算法对两批模拟数据的分析结果中, 在所搜到的富集峰的数量方面, CisGenome 在两批模拟数据中所搜到的峰数(1320 和 1279)差别不大, 说明 CisGenome 在其常规处理中对富集峰局部序列分布的处理不敏感. 而 MACS 在两批数据中所得到的配对峰的数量上的显著差异(1489 和 699), 证实了 MACS 对链特异性的序列分布的要求. MACS 能在局部双链序列分布重叠的情况下(即模拟 SOLiD ChIP-Seq 数据)仍能找到 699 个配对峰, 这可能是来自噪声信号的随机贡献或是由于模拟真实信号样本量过小, 不能满足精准的正负链序列分布重叠, 从而出现了正负链峰之间的位移差. 虽然 SISSRs 在两批模拟数据中所找到的总峰数有差异, 但与 MACS 相比, 其差异并不大, 这与 SISSRs 的算法对正负链峰的对称性的非严格要求有关.

尽管分析那些模拟在 SOLiD 系统上进行 ChIP-Seq 实验所产生的局部没有典型的双峰现象的数据集, SISSRs 和 MACS 仍然会有结果出现. 但不妙的是, 这些结果具有迷惑性. 对于模拟 SOLiD 系统上采用进一步超声的 ChIP-Seq 数据来说, 它的 SISSRs 的片段长度估计达到 116 bp(表 3), 这个数值是误导性的但又是可以理解的. SISSRs 对两个分离的峰的对称性要求较少, 它重点关注的是正链富集区与负链富

表 3　模拟 Solexa 和 SOLiD ChIP-Seq 数据的分析结果总结

		Solexa sim	SOLiD sim
CisGenome	片段长度/bp	48	–28
	总峰数	1320	1279
SISSRs	片段长度/bp	190	116
	总峰数	1389	1716
MACS	配对峰	1489	699
	片段长度/bp	73	31
	总峰数	5299	2416

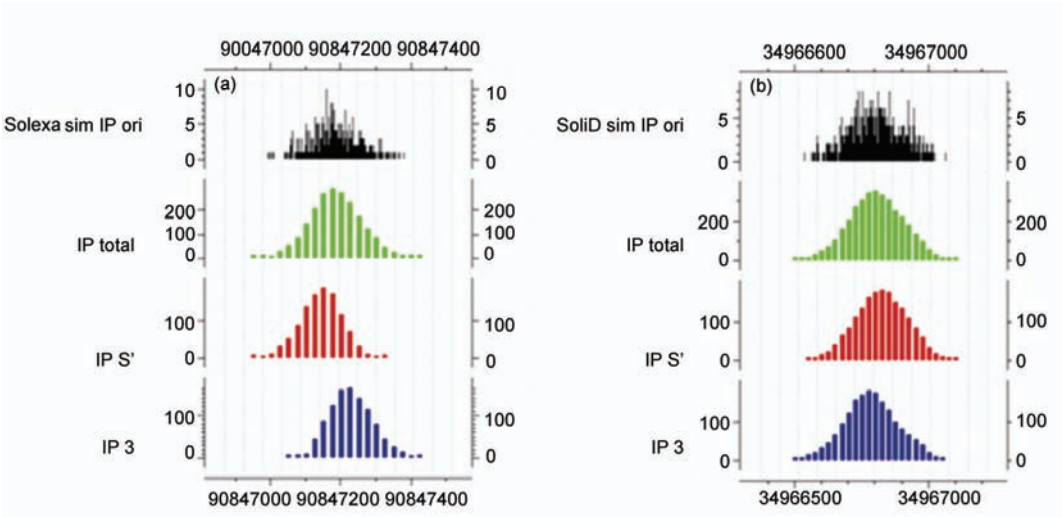

图 6　模拟 Solexa 和 SOLiD ChIP-Seq 数据局部序列方向性分布的比较

(a) Solexa 系统上的 ChIP-Seq 呈典型的分离双峰模式; (b) SOLiD 系统上的 ChIP-Seq 表现出正负链峰重叠

集区的转折点, 这种转折点也可出现在模拟 SOLiD 局部正链序列和负链序列服从同一分布和在同一点达到最高密度的 ChIP-Seq 中. 然而, 尽管 SISSRs 可通过其网络计分方程找到一定数量的零点(1716), 这些点可能不再是准确的结合位点.

MACS 对两批模拟数据的分析结果截然不同. 模拟的 SOLiD 系统 ChIP-Seq 产生的配对峰的数量, 估计的片段长度和总峰数都是 Solexa 系统的一半. 由于 MACS 处理数据的方式类似于链式反应, 这种结果的出现是正常的. 如果片段长度估计得较短, 那么由于扫描窗宽是片段长度估计的两倍, 那么扫描窗宽就会较小. 进一步, 在在较小的扫描窗宽条件下, 其窗口的序列数就越难超过给定 P 值下算出的序列阈值. 因此, 如果用 MACS 来分析 SOLiD 系统的 ChIP-Seq 数据, 所估计出来的短片段可能与来自二次超声的片段长度一致, 但是配对峰的数量可能不足以进行该参数的估计, 并且可获得的配对峰可能是一种随机现象. 此外, 较少的总峰数可能意味着 MACS 在处理这类 ChIP-Seq 时可能会丢失一些结合信息.

CisGenome 分析这两批模拟的 ChIP-Seq 数据, 一个来自 Solexa, 另一个来自 SOLiD, 得到的峰的数量无明显差异(表 3). 这是由于 CisGenome 在其进行搜峰时没有依赖任何的序列方向性方面的先验知识. 它的优化由后处理方式实现. 如果仅由一条链上的序列富集成一个结合区域, 那么这些区域可被过滤, 这个功能对 Solexa 和 SOLiD 来说都可行. 尽管 CisGenome 运用候选结合区域的 5′端和 3′端的序列计数去精确定位该区域的结合位点, 对 SOLiD 来说是不合适的, 因此从 CisGenome 富集的峰可能缺乏空间分辨率.

2.6 处理真实的 SOLiD ChIP-Seq 数据的建议

在处理丢失了序列方向性的 SOLiD ChIP-Seq 情况下, 如何达到一个可接受的空间分辨率成为一个重要的议题. 相较于滑动窗口和序列延伸方法(XSET)[7,9,19] 的统计序列的方法, 核密度估计(KDE), 又名 Parzen 窗口[20], 可有助于精确定位所研究蛋白的结合位点. 这个方法对一个区域内的每个碱基处都打分, 其原则是离所测序列位点较近的区域计分权重高, 离所测序列位点较远的区域计分权重低, 这个方法已被应用于 QuEST[12]和 F-Seq[21], 并取得了不错的结果. 如果将来有十分可靠的 SOLiD ChIP-Seq 数据的发表, 可首先尝试用这种方法来处理.

3 结论

通过真实和模拟的 ChIP-Seq 数据, 我们研究了分析 ChIP-Seq 数据 DNA 片段长度的影响, 并比较了三个主要的 ChIP-Seq 算法(CisGenome, SISSRs 和 MACS). 当前通过 Solexa 文库制备方案所产生的 ChIP-Seq 数据的处理办法, 不能有效地捕获来自 SOLiD 系统的 ChIP-Seq 数据的特征. 因此有必要开发一些新算法去处理这些即将出现的新数据集.

目前处理 ChIP-Seq 的算法仍处于其成长期. 本文仅讨论了 ChIP-Seq 数据处理中的部分话题. 当处理真实的来自 SOLiD 系统的 ChIP-Seq 数据时, 还有许多其他要考虑的问题, 如当没有 IgG 或基因组 DNA 对照 ChIP-Seq 的前提下如何控制单样本的偏向性, 基因组的重复区域的比率的问题, PCR 扩增的偏向性和测序误差以及测序深度的问题. 实验生物学家和计算生物学家应意识到每一个 ChIP-Seq 数据集可有其独特的特征, 最好的算法就是能捕获这种特征的算法.

参考文献

1 Roh T, Ngau W C, Cui K, et al. High-resolution genome-wide mapping of histone modifications. Nat Biotechnol, 2004, 22: 1013—1016

2 Impey S, McCorkle S R, Cha-Molstad H, et al. Defining the CREB regulon: A genome-wide analysis of transcription factor regulatory regions. Cell, 2004, 119: 1041—1054

3 Park P J. Epigenetics meets next-generation sequencing. Epigenetics, 2008, 3: 318—321

4 Hoffman B G, Jones S J M. Genome-wide identification of DNA-protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing. J Endocrinol, 2009, 201: 1—13

5 Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. Cell, 2007, 129: 823—837

6 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: 553—560

7 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: 651—657
8 Johnson D S, Mortazavi A, Myers R M, et al. Genome-wide mapping of *in vivo* protein-DNA interactions. Science, 2007, 316: 1497—1502
9 Rozowsky J, Euskirchen G, Auerbach R K, et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol, 2009, 27: 66—75
10 Whiteford N, Haslam N, Weber G, et al. An analysis of the feasibility of short read sequencing. Nucleic Acids Res, 2005, 33: e171
11 Shah A. Chromatin immunoprecipitation sequencing (ChIP-Seq) on the SOLiD™ system. Nat Methods, 2009, 6: ⅰ—ⅲ
12 Valouev A, Johnson D S, Sundquist A, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods, 2008, 5: 829—834
13 Jothi R, Cuddapah S, Barski A, et al. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res, 2008, 36: 5221—5231
14 Zhang Y, Liu T, Meyer C A, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol, 2008, 9: R137
15 Ji H, Jiang H, Ma W, et al. An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol, 2008, 26: 1293—1300
16 Zhang Z D, Rozowsky J, Snyder M, et al. Modeling ChIP wequencing in silico with applications. PLoS Comput Biol, 2008, 4: e1000158
17 Shendure J, Porreca G J, Reppas N B, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science, 2005, 309: 1728—1732
18 Schmid C D, Bucher P. ChIP-Seq data reveal nucleosome architecture of human promoters. Cell, 2007, 131: 831—832; author reply 832—833
19 Fejes A P, Robertson G, Bilenky M, et al. FindPeaks 3.1: A tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics, 2008, 24: 1729—1730
20 Parzen E. On estimation of a probability density function and mode. Ann Math Stat, 1962, 33: 1065—1076
21 Boyle A P, Guinney J, Crawford G E, et al. F-Seq: A feature density estimator for high-throughput sequence tags. Bioinformatics, 2008, 24: 2537—2538