Hans Journal of Biomedicine 生物医学, 2011, 1, 1-5
http://dx.doi.org/10.12677/hjbm.2011.11001 Published Online July 2011 (http://www.hanspub.org/journal/hjbm/)

High Throughput Sequencing Methods and Applications of Read Mapping Algorithm

Huili Li[1], Feng He[1], Hang Yang[1], Yan Zheng[2], Xiaoming Wu[1*]
[1]The Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology Xi'an Jiaotong University, Xi'an
[2]Department of Dermatology, Second Affiliated Hospital, Xi'an Jiaotong University, Xi'an
Email: wxm@mail.xjtu.edu.cn; lihuili8892@163.com
Received: Apr. 18th, 2011; revised: May 20th, 2011; accepted: May 21st, 2011

Abstract: Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) has become a tool for studying DNA-binding proteins profiles and histone modifications. At the same time, it also arouse requirements for effective computational methods to map short DNA reads and to compare mapping profiles, which are crucial for uncovering biological mechanisms. In this article, we introduced the principle of new generation sequencing and corresponding data format, then gave an overview of current method of reads mapping and peak identification. Finally, we demonstrated the application of this method in histone modifications analysis and transcription factor binding sites identification.

Keywords: Protein-DNA Interaction; Chip-Seq; Genome; Sequence Read Mapping

高通量测序及读序映射算法的应用

李慧丽[1]，何　凤[1]，杨　航[1]，郑　焱[2]，吴晓明[1*]
[1]西安交通大学生命科学与技术学院，生物医学信息工程教育部重点实验室，西安
[2]西安交通大学，第二附属医院皮肤科，西安
Email: wxm@mail.xjtu.edu.cn; lihuili8892@163.com
收稿日期：2011 年 4 月 18 日；修回日期：2011 年 5 月 20 日；录用日期：2011 年 5 月 21 日

摘　要：免疫共沉淀－DNA 高通量测序二者的结合是研究蛋白质与基因组 DNA 相互作用及组蛋白修饰的新实验工具，它同时也对短 DNA 读序在基因组上的映射、映射结果比较提出了新的算法需求。本文介绍新一代测序原理及数据的特点、相关的读序映射算法的基本原理及对应软件，并说明了该方法在组蛋白修饰、转录因子结合位点分析中的应用。

关键字：蛋白质 DNA 相互作用；ChIP-Seq；基因组；读序映射

1. 引言

生物体内基因调控的一个重要环节是细胞中的功能分子如启动子、聚合酶、分子伴侣等，同遗传物质 DNA 之间进行相互作用，从而影响基因的转录调控。基因组 DNA 的甲基化、组蛋白甲基化、乙酰化、磷酸化，染色体结构变化都影响基因的调控，进而对生命活动的各个环节产生影响。此时，DNA 的结构不再是规则的双螺旋结构，而是由于蛋白－DNA 之间的作用，染色体的局部结构发生空间变化。这些变化不仅出现在基因的转录起始区域，而是会出现在和基因相距较远的区域。多种不同的染色体结构变化以远程作用的方式，影响着不同基因、不同条件下的调控。分子层次引起 DNA 结构变化的途径有很多，仅组蛋白修饰就有 100 多种。在这些结构变化中，蛋白质与 DNA 的相互作用占据绝对的优势。染色质免疫共沉淀(Chromatin immunoprecipitaion, ChIP)技术是全基因组范围内识别 DNA 与蛋白质体内相互作用的标准方法，可用于组蛋白修饰研究、转录因子研究等多方面，且其应用领域还在不断扩展。

高通量的 DNA 测序是一种并行测序技术，按照边扩增、边检测的方式来完成。目前典型的、常被实验采用的高通量测序方法有 Iilumina 公司的 Solexa 技

术，Applied Biosystem 公司的 SOLiD 技术以及 Roche 公司的 454-pyrosequencing 技术。每种测序方法在测序通量和成本上都在迅速地发展[1]。

ChIP-Seq 是免疫共沉淀实验和高通量 DNA 测序结合的实验技术。该实验的对象是分布在整个基因组范围的、能够和特定蛋白质结合的 DNA 序列，实验结果是大量短的 DNA 序列片段，该序列片段在基因组上的位置、分布特征、分布变化情况等需要利用生物信息学的技术进行分析，并在数据分析阶段排除实验过程中的各种假阳性数据。最终结果将结合具体的实验设计，达到对实验数据的解析和基因组功能分析工作。

ChIP-Seq 实验目前最主要的研究体现在两个方面，分别是：1. 特定的转录因子研究，包括识别转录因子的位置、长度、DNA 结合强度、结合特异性等方面。2. 组蛋白修饰研究，包括特异性修饰的在基因组的位置、核小体的位置等信息。此外，它可以用于对 RNA、SNP 基因多态性进行分析。

2. 实验技术-高通量测序

高通量测序技术使用并行的方法进行，但三种主要方案还是有所区别。在 454 测序技术中，先使特异性测序引物和单链 DNA 模板结合后，在多种酶及底物的共同参与下，将每一个 dNTP 的聚合与荧光信号的释放偶联起来。通过 CCD 光学系统即可获得一个特异的检测峰，峰值的大小表示对应的碱基数。然后在反应体系中，加入另一种 dNTP，使以上反应重复进行，根据获得的峰值图即可读取准确的 DNA 序列信息[2]。

Solexa 测序技术在进行 DNA 聚合酶的链延伸反应中，使用经过了特殊修饰的 dNTP，四种不同的 dNTP 分别被标记上了不同的荧光基团和 3’末端保护基团，以使每一步反应只能延伸一个碱基。每步反应所收集到的荧光信号对应了所要检测的碱基[2]。

SOLiD 测序技术在待测的短的 DNA 片段的两侧，连上称为 P1 接头和 P2 接头的 SOLiD 接头，然后利用能够和双碱基配对的分子，从接头处进行扩增和延伸，不同的双碱基对应不同的荧光信号，从这些荧光信号中最终分析出序列信息。

更新的实验技术是基于单分子的测序技术，能够以更高的速度、更低的成本、更少的样本完成 DNA、RNA 的测序工作，它们在进行基因组功能研究方面也将会有很大的作为。

3. 实验数据的格式

由于采用荧光信号的检测方式，每个检测获得的最原始的数据为荧光图像。根据图像中荧光点的位置、颜色、强度等信息，这些原始数据会被转换为带有一定数据质量的序列信息，用不同的文件格式来保存。通常见到的是 FASTQ 格式。该格式是包含有序列和序列中每个碱基数据 Phred 质量的文件，文件格式简洁紧促，最早用于表示毛细管电泳 DNA 测序的结果。新一代测序技术 Solexa/Illumina 的读序文件也是 FASTQ 格式，但其表示碱基质量的方法和传统的 FASTQ 格式有一些不同。该文件利用 ASCII 代码大于 90 的字符表示对应碱基的数据质量。例如序列的描述：

```
@SRR002004.12 Nanog: 1: 1: 681: 134 length = 36
GTTATGGGCGGTGGGTTTATTTGTAGTATATTTATT
+ SRR002004.12 Nanog: 1: 1: 681: 134 length = 36
III = IIII"I$' < + I = 8. /&.('%%("/&)%!!()
```

第一行和第三行表示序列的名称，第二行表示序列，第四行是用 ascII 码字符表示的一个正整数，代表序列中每个碱基的数据质量。每个测序得到的读序被映射到基因组之后，就能得到读序在基因组中的分布。BED 格式和 WIG 格式为包含读序在染色体中起始位置的文本文件，其中 WIG 格式是能够显示连续数值的格式[3]。GFF (General Feature Format) 最早被开发出来用于表示映射过的 SOLiD 数据，每行描述一个读序的信息，包括在基因组中的位置，正链还是负链，颜色代码表示的序列等（四个颜色代码表示 16 个双碱基）[4]。

4. 分析方法

4.1. 读序映射

高通量测序获得的是众多的短的 DNA 读序片段，它们是在实验中分离得到、和实验相关的 DNA 序列，需要找到它们在基因组中的位置，这就需要进行读序映射工作。该工作实际上是一个序列比对，但由于基

因组往往很大，且高通量实验得到的读序数量非常多，长度却很短，因此需要快速、有效、能容纳一些错误匹配的算法来实现。

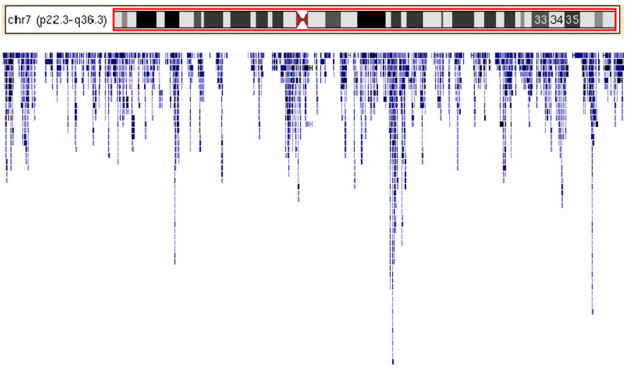
基于动态规划算法的比对是最为准确的，但同时也是最耗费计算机时间的算法。基于索引的算法能够对序列进行快速的定位，但当序列较长时，所需要的索引文件就非常大。例如，当索引所有长度为 10 的 DNA 序列时，需要的索引的个数为 4^{10} 个，即 2 百万个。此时当前的计算机还能够应付，但读序的长度往往超过 30bp，此时需要的索引的数量为 4^{30}，这是一个非常大的数字，远远超过当前计算机的存贮和检索的能力。较为有效的方法是把读序切分成多个更小的片段，然后在利用索引的方法进行序列位置的确定。基于此思路的读序映射方法包括有 SeqMap[5]、ELAND、SOAP、RMAP、MAQ[6]、Bowtie[7]、SSAHA、SSHRiMP[8]、SOCS[9]等这几个算法虽然功能相似，但具体实现方案有所差异。例如，SeqMap 获得的序列比对能够包含多达 5 个的字符替换、插入或剔除，使用非常灵活。能够处理包括 FASTA 格式的序列输入形式和多种序列输出格式(如 ELAND 格式)。同时它提供了命令行的执行方式，在进行读序映射时更加灵活。其思路是对读序建立索引，利用索引的哈希值进行匹配和比较，因而对计算机的内存容量要求较低，典型的映射工作可在普通的计算机、数小时的计算时间完成，同时，也非常容易在计算机集群上并行运行。而 BLAT 和 SOAP 程序是对基因组构建索引和哈希表。需要更多的内存来容纳索引[5]。

ELAND 是和 Illumina 的测序仪器配套的分析软件，能够完成读序映射工作，并输出所有的读序在基因组中的位置信息。这些信息可被转换成为 BED 格式或 WIG 格式，它们均可以在基因组浏览工具中以图形化的形式显示读序的位置和数量。

在映射时需要考虑的参数还包括：最大允许错配碱基数目、读序在基因组的映射位置数目等。最终的目的是获得读序基因组分布模式，并通过构建模型获得分布模式有显著性差异的基因组区域。

4.2. 结果显示

由于基因组是一维的序列，在显示时候，既要能够显示局部区域的细节，又需要显示整个基因组范围



(http://genome.ucsc.edu/cgi-bin/hgTracks)

Figure 1. UCSC Genomic feature visualization tool
图 1. USCS 基因组特征显示工具

的宏观分布，因此，需要能够在不同尺度采用不同的显示手段。常见的基因组显示工具，利用缩放技术能够很好的完成显示。UCSC 提供的基因组浏览器就能够很好的实现映射结果的显示，但它需要把映射好的数据上传的服务器来实现。图 1 为采用 UCSC 基因组浏览器进行 7 号染色体上基因分布显示结果。网页为 http://genome.ucsc.edu/cgi-bin/hgTracks，选择其中的 Genes and Gene Prediction Tracks，参数中 Display mode 设定为 squish，track by codons 设定为 OFF。

CisGenome 是一个可以在本地计算机运行的软件，并实现各种注释数据的显示，用它可以进行读序分布的显示，并能够完成不同读序分布谱的比较，找出差异的读序分布位点[10]。它同时也是一个峰识别工具，利用滑动窗口识别基因组上读序富集的区域。在识别时使用了负二项分布来描述背景读序的出现，该方法优于泊松分布的分析结果。当有对照样本时，利用二项模型判断实验中获得的读序富集的区域和对照数据中相比是否有统计学的差异。

4.3. 峰值确定和比较

读序的位置被确定后，就可根据不同基因组区域读序的数量确定那些被读序高度覆盖的区域。在基因组中，数量较多的序列片段都映射到的共同区域最有可能是蛋白质－DNA 结合的区域，当按照读序对应的位置和读序的数量进行可视化，这些区域会形成峰。这个工作称为读序峰识别，峰所在的位置可以看作是蛋白质的特异结合位置。多种软件和算法被开发出来完成这个工作，这些算法包括：Peak Finder[11]、

MACS[12]、Hpeak[13]、FindPeaks[14]、SISSRs[15]、ChIP-Peak 等。这些搜峰软件的目的是从基因组中得到这些区域，并最终识别出具有重要生物意义的峰值信号。

Peak Finder 是最早的峰识别软件，峰被定义为 *k* 个读序位于 *n* 个碱基之间(默认长度 *n* 为 75 bp)，并且至少有 5 个读序重叠。当有对照样本时，读序的数量为对照样本的 5 倍以上[11]。Hpeak 是一个基于隐马尔科夫模型的读序峰值搜索软件，用于识别基因组中的蛋白质结合区域[13]。它的识别过程分为 4 步：(1) 导入读序在基因组中的位置信息，这些信息可以是 ELAND 等软件的输出，或者 BED 格式；(2) 把基因组分割成为默认长度为 25 bp 的小区间，并统计每个区间读序的数目，形成图谱；(3) 应用两状态的隐马尔科夫模型，从图谱中识别连续的读序富集区域和背景区域，识别出峰。免疫沉淀产生序列多的区域识别为峰值，其余部分被识别为背景。(4) 输出 WIG 格式的分析结果以及序列信息，供可视化软件显示和序列软件分析和进行基因组注释。

FindPeaks：将读序进行延伸到实验中 DNA 片段的平均长度，重叠的部分叠加形成峰，从这些峰中识别结合位点，HDFs 的数量超过阈值时，被识别为峰。利用 Monte Carlo 模拟的方法计算每个结合位点的错误识别率，该模拟基于随机序列。

MACS：分离不同链上的读序，利用泊松模型描述背景分布时候读序的分布，并且该模型是一个动态模型，模型中的参数可以在一定范围内改变，以便识别局部的序列变化。当有对照样本时，错误识别率 FDR 的计算方法是对照样本中识别到的峰的数目除以在免疫沉淀实验得到的峰值数目。

5. 实际应用

5.1. 组蛋白修饰

甲基化是一个典型的组蛋白修饰，人们通过实验研究了位于启动子区、增强子区、转录区域等区域组蛋白甲基化的特征，发现 H3K27，H3K9，H4K20，H3K79，H2BK5 的单甲基化通常和基因活化相关，而 H3K27，H3K9，H3K79 的三甲基化同基因的抑制相关。H3K4 的甲基化也同染色体的断裂相关[16]。从鼠胚胎肝细胞等细胞中，研究者发现 lysine 4、lysine 27 三甲基化状态能够用于区分和表达、抑制的基因，而 Lysine 36 的三甲基化和转录体是否编码有关，该结果可用于基因组的标注[17]。

5.2. 转录因子结合位点识别

2007 年 Johnson 等人用 ChIP-Seq 对转录因子 NRSF 在 DNA 上的结合位点进行了全基因组的筛查，获得了 1946 个结合位点，为研究 DNA-蛋白相互作用提供了新的方法和数据，并可以用于识别非典型的结合位点形式[11]。Robertson 等人也用该方法检测干扰素(IFN-gamma)刺激和非刺激的 HeLa S3 细胞中 STAT1 结合位点分布的情况，分别识别到了 41,582 和 11,004 个潜在的结合位点，并验证了已知结合位点[18]。

6. 结论

新一代的测序方法能够高通量、低成本地对由分子生物学实验获得的大量 DNA 序列片段进行测序。通过读序映射的方法，这些序列片段在基因组中的位置得以确定；进而，基因组不同位置读序的覆盖度也可以获得。结合免疫沉淀等方法，可以有目的的检测不同实验条件下分离得到的 DNA 序列，然后利用读序映射的方法，确定这些序列片段在染色体中的位置信息，这些位置信息对应了蛋白质在基因组上的结合分布。在基因组功能分析方面、转录因子结合位点研究方面已经取得巨大的成功。该技术将会有力的推动在基因的调控、染色体的结构变异等方面的研究。值得注意的一点是，该技术所涉及的测序工作，较第一代的测序而言，成本已经大大降低。但当它用于进行基因组范围的分析时候，由于涉及到的测序的 DNA 片段的数量巨大，实验所需费用仍旧较高。第三代的基于单分子的测序技术采用序列边复制边检测的原理[19]，能够高通量、高速度地进行序列测序，会在不远的将来展现其强有力的价格优势和高的测序质量。

7. 致谢

本研究得到国家自然科学基金(No. 60601017)，陕西省卫生厅科研基金(No. 2010E06, 2010D21)，西安交通大学基本科研业务费基金项目资助。

参考文献 (References)

[1] M. L. Li, W. Wang, and Z. H. Lu. Genomic analysis of DNA-protein interaction by chromatin immunoprecipitation.

Hereditas, 2010, 32(3): 219-228.
[2] C. Chen, H. Wan, and Q. Zhou. The next generation sequencing technology and its application in cancer research. Chinese Journal of Lung Cancer, 2010, 13(2): 154-159.
[3] Browser UG. UCSC Genome Browser: Wiggle Track Format (WIG)[URL]. http://genome.ucsc.edu/goldenPath/help/wiggle.html, 2011-7-16 /2011-7-16.
[4] Welcome Trust Sanger Institute, Genome Research Limited. GFF (General Feature Format) Specifications Document—Welcome Trust Sanger Institute [URL]. http://www.sanger.ac.uk/resources/software/gff/spec.html, 2011 -4-19/2011-7-16.
[5] H. Jiang, W. H. Wong. SeqMap: Mapping massive amount of oligonucleotides to the genome. Bioinformatics, 2008, 24(20): 2395-2396.
[6] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research, 2008, 18(11): 1851-1858.
[7] B. Langmead, C. Trapnell, M. Pop, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol., 2009, 10(3): R25.
[8] S. M. Rumble, P. Lacroute, A. V. Dalca, et al. SHRiMP: Accurate mapping of short color-space reads. PLoS Comput Biol, 2009, 5(5): Article ID e1000386.
[9] B. D. Ondov, A. Varadarajan, K. D. Passalacqua, et al. Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. Bioinformatics, 2008, 24(23): 2776-2777.
[10] H. Ji, H. Jiang, W. Ma, et al. An integrated software system for analyzing ChIP-Chip and ChIP-Seq data. Nat Biotechnol, 2008, 26(11): 1293-1300.
[11] D. S. Johnson, A. Mortazavi, R. M. Myers, et al. Genome-wide mapping of in vivo protein-DNA interactions. Science, 2007, 316(5830): 1497-1502.
[12] Y. Zhang, T. Liu, C. A. Meyer, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biology, 2008, 9(9): R137.
[13] Z. S. Qin, J. Yu, J. Shen, et al. HPeak: An HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. BMC Bioinformatics, 2010, 11: 369.
[14] A. P. Fejes, G. Robertson, M. Bilenky, et al. FindPeaks 3.1: A tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics, 2008, 24(15): 1729-1730.
[15] R. Jothi, S. Cuddapah, A. Barski, et al. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res., 2008, 36(16): 5221-5231.
[16] A. Barski, S. Cuddapah, K. Cui, et al. High-resolution profiling of histone methylations in the human genome. Cell, 2007, 129(4): 823-837.
[17] T. S. Mikkelsen, M. Ku, D. B. Jaffe, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature, 2007, 448(7153): 553-560.
[18] G. Robertson, M. Hirst, M. Bainbridge, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature Methods, 2007, 4(8): 651-657.
[19] J. Eid, A. Fehr, J Gray, et al. Real-time DNA sequencing from single polymerase molecules. Science, 2009, 323(5910): 133-138.

