DOI: 10.11844/cjcb.2013.06.0035

ChIP-Seq技术在研究转录因子调控干细胞分化中的应用

乌云毕力格[1] 顾婷玉[2] 何志颖[3] 吴 侠[1] 李光鹏[1] 李前忠[4] 丁小燕[2] 左永春[1]* 王 欣[1,2]*

([1]内蒙古大学哺乳动物生殖生物学及生物技术教育部重点实验室, 呼和浩特 010021;
[2]中国科学院上海生命科学研究院生物化学与细胞生物学研究所, 上海 200031;
[3]第二军医大学细胞生物学教研室, 上海 200433; [4]内蒙古大学物理科学与技术学院, 呼和浩特 010021)

摘要 研究胚胎发育进程中的转录因子调控及其相关网络, 是全面揭示胚胎分化机制需要阐明的首要问题。ChIP-Seq技术凭借其对DNA分子序列及丰度的双重解析能力, 已广泛应用于转录因子分子调控机制方面的研究。该文对ChIP-Seq技术和高通量测序技术的发展和更新及其在转录因子调控胚胎发育分化中的研究进展进行了综述, 重点论述了ChIP-Seq相关技术在胚胎干细胞和特定肝向分化领域内的最新研究进展。

关键词 染色质免疫共沉淀; 转录因子; 高通量DNA测序; 胚胎; 肝脏

Analysis of Transcription Factors Regulated Stem Cell Differentiation by ChIP-Seq

Uyunbilig Borjigin[1], Gu Tingyu[2], He Zhiying[3], Wu Xia[1], Li Guangpeng[1], Li Qianzhong[4], Ding Xiaoyan[2], Zuo Yongchun[1]*, Wang Xin[1,2]*

([1]*The Key Laboratory of Mammalian Reproductive Biology and Biotechnology, Ministry of Education, Inner Mongolia University, Hohhot 010021, China*; [2]*The Laboratory of Cellular and Molecular Biology, Shanghai Institute of Cell Biology and Biochemistry, Chinese Academy of Sciences, Shanghai 200031, China*; [3]*Department of Cell Biology, Second Military Medical University, Shanghai 200433, China*; [4]*School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China*)

Abstract Study of intricate transcription factor regulatory networks is main issue of illustrating mechanism of embryonic development and differentiation. ChIP-Seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. ChIP-Seq is widely used to analyze regulatory networks of transcription factors. This paper provides a brief overview of the ChIP-Seq method and application in transcription factors regulation during embyonic development and differentiation, mostly focuses on embryonic stem cell and committed hepatic differentiation.

Key words ChIP-Seq; transcription factor; massively parallel DNA sequencing; embryo; liver

引言

基因的转录调控是生物基因表达调控层次中最关键的一层, 转录因子(TFs)通过特异性结合调控区域的DNA序列来调控基因转录过程。其中, 基础转录因子通过与RNA聚合酶在转录起始位点附近的启动子区的相关作用在实现基因的准确转录中起着关键作用; 而调控性转录因子则通过与增强子结合在组织发育、细胞分化等基因表达水平调控中发挥

收稿日期: 2013-01-28 接受日期: 2013-03-25
国家自然科学基金(批准号: 31060168、31271469、31271474、31271042)资助的课题
*通讯作者。Tel: 0471-5227683, E-mail: yczuo@imu.edu.cn; E-mail: wangxxx6@yahoo.com
Received: January 28, 2013 Accepted: March 25, 2013
This work was supported by the National Natural Science Foundation of China (Grant No.31060168, 31271469, 31271474, 31271042)
*Corresponding author. Tel: +86-471-5227683, E-mail: yczuo@imu.edu.cn; E-mail: wangxxx6@yahoo.com
网络出版时间: 2013-05-22 13:11 URL: http://www.cnki.net/kcms/detail/31.2035.Q.20130522.1311.002.html

着极其重要的作用[1]。 所以, 理解转录因子与结合位点间的相互作用是准确揭示转录调控机制、构建转录调控网络的关键所在。转录因子及其DNA结合位点的鉴定, 以及它们构成的基因转录调控网络的构建已经成为目前生物信息学和系统生物学研究的重点领域, 也是生命科学研究的前沿热点。

1 转录因子调控机理研究的技术发展现状

转录调控及相关网络的搭建是目前分子细胞生物学研究中的一个热点领域, 而研究转录因子与靶基因间的相互作用及识别转录因子结合位点(transcription factor binding sites, TFBSs)是理解转录调控机制的关键。要弄清楚这些转录因子调控基因表达的分子机制, 就必须鉴定出这些转录因子全部的靶基因并构建其操纵的转录调控网络。

1.1 传统研究技术

基因转录过程的激活、抑制和调节主要通过转录因子蛋白与其在基因组序列中对应位置的结合位点之间的交互作用来实现。转录调控因子有序地结合在目标基因不同调控区域中的特殊位点, 启动基因的转录和控制基因的转录效率。这些位点被称为转录因子结合位点(TFBSs)。因此, 转录因子结合位点是转录因子结合到靶基因上的一段DNA基序, 长度范围从几个碱基到十几个碱基之间。由于同一转录因子往往同时调控若干个基因, 同时不同转录因子通过相互作用也可以协同调控同一靶基因, 转录因子与靶基因之间的相互作用也具有不同程度的特异性和亲和性, 所以不同基因上的结合位点保守性较弱。较短的DNA寡片段在规模较大基因组中重复出现的次数很多; 同一转录因子的结合位点又具有一定的可变性, 这给TFBS的识别研究带来了困难, 尽管科研工作者对TFBS研究已有多年, 但大多数转录因子调控靶基因的具体调控模式仍不明朗[2]。

转录因子结合实验(transcription factor assay)、电泳迁移率变动分析(electrophoretic mobility shift assay)、DNase I足印法(DNase I footprinting)、酵母单杂交系统等实验技术是识别转录因子结合位点的传统方法[3]。虽然, 这些技术可以逐一鉴别出与特定转录因子结合的DNA序列片段, 但由于它们很难充分反映生理情况下DNA与蛋白相互作用的真实情况, 也很难捕捉到在染色质水平上基因表达调控的动态瞬时事件, 加上昂贵的费用及较长的时间花费说明已显然不适合开展后基因组时代DNA转录因子结合位点的鉴定[4]。近年, 染色质免疫沉淀技术(ChIP)被广泛用于研究体内转录调控因子与靶基因启动子上特异性核苷酸序列的结合, 并已成为研究染色质水平基因表达调控的最标准有效的方法[5]。

1.2 染色质免疫共沉淀技术

染色质免疫共沉淀技术(chromatin-immunoprecipitation, ChIP), 也称结合位点分析法, 因其能真实、完整地反映结合在DNA序列上的靶蛋白的调控信息, 是目前基于全基因组水平研究DNA-蛋白质相互作用的标准实验技术, ChIP技术由Orlando等[6]于1997年创立。其基本原理与过程是: 通过在特定时间点上用甲醛交联等方式“固定”细胞内所有DNA结合蛋白的活动, 相当于这一时间点上细胞内蛋白和DNA相互作用的关系被瞬时“快照(snapshot)”下来。再通过后续的裂解细胞、断裂DNA, 将蛋白质-DNA复合物与特定DNA结合蛋白的抗体孵育, 然后将与抗体特异结合的蛋白-DNA复合物洗脱下来, 最后将洗脱得到的特异DNA与蛋白解离、纯化DNA后, 进行下游分析[7](图1A)。

破碎DNA及具有较高的特异性和亲和力的抗体是ChIP实验成功与否的关键因素。抗体又有单抗与多抗之分, 选择起来也需要仔细考虑。单抗特异性强, 背景低。但致命的弱点是识别位点单一, 而在ChIP甲醛交联的过程中, 很有可能因该位点被其他蛋白或核酸结合而被封闭, 导致单抗不能识别靶蛋白; 而多抗特异性较差, 背景可能会偏高。常见的两种ChIP实验技术有N-ChIP和X-ChIP技术。N-ChIP采用核酸酶消化染色质, 适用于研究DNA与高结合力蛋白的相互作用, 比如组蛋白修饰等方面的研究[8]; X-ChIP则采用甲醛或紫外线进行DNA和蛋白交联, 通过超声波片段化染色质, 适合用来研究DNA与低结合力蛋白的相互作用问题, 例如大多数非组蛋白方面的蛋白研究[9]。

ChIP是相对成熟的技术, 但目前还存在一些技术难点。例如, ChIP实验涉及的步骤多, 结果的重复性较低, 需要大量的起始材料; 染色质免疫沉淀获得的DNA数量往往很多, 包含大量的非特异结合的假阳性结合基序; 又如对于像神经细胞和干细胞具有培养起来较难等局限性, 并且难以区分个别细胞与总体细胞的表型。在此背景下, 配合使用芯片或者第二代高通量测序技术检测这些DNA片段, 就形成

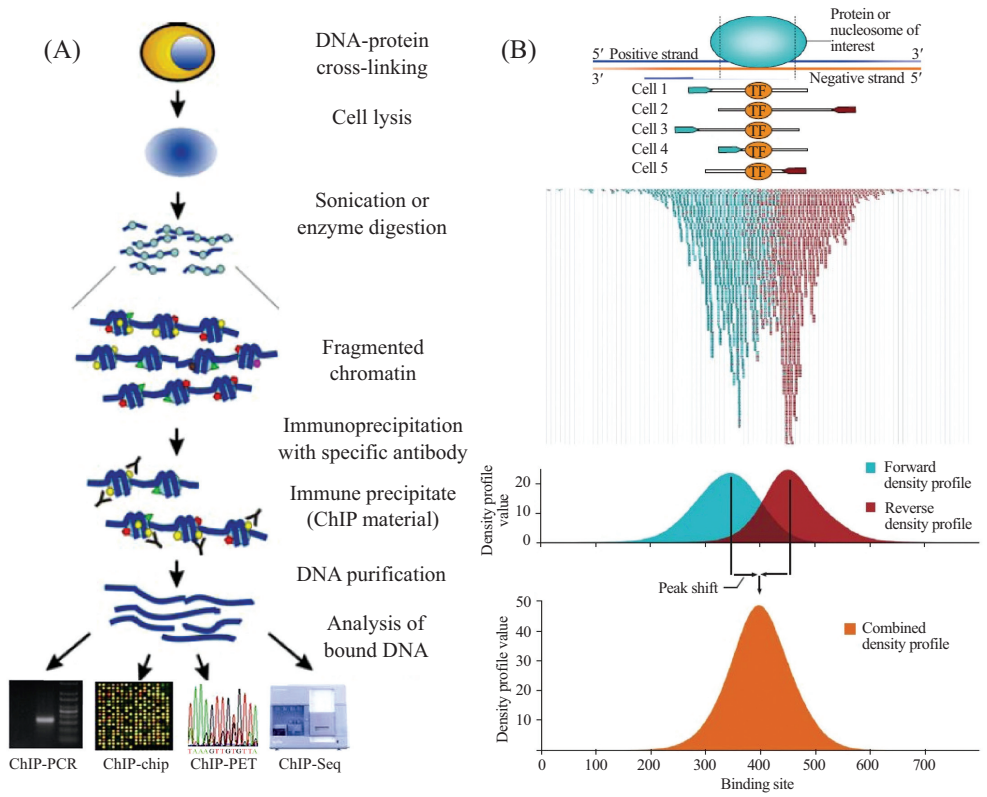

图1 ChIP实验流程(A, 根据参考文献[5]修改)及ChIP-Seq转录因子结合位点分析(B, 根据参考文献[36]修改)

Fig.1 The workflow of ChIP assay(A, modified from reference [5]) and the TFBSa analysis of ChIP-Seq methods (B, modified from reference [36])

了ChIP-chip技术和ChIP-Seq技术。

1.3 ChIP-chip技术

ChIP-chip技术结合了基因芯片技术和ChIP技术, 具体来说是将生物芯片平台与ChIP实验相结合, 在全基因组或基因组较大区域上高通量分析DNA结合位点或组蛋白修饰位点的方法。其原理是: 全基因组已知的全部基因的调控区DNA片段, 汇合起来做成DNA芯片, ChIP后得到的DNA与此DNA芯片进行杂交后, 结合到的DNA片段就被识别出来。因此, ChIP-chip能大规模分析筛选ChIP后与转录因子相结合的DNA序列, 从而推断出转录因子所调控的下游基因。ChIP-chip获得的信息量主要取决于芯片表面固定的探针数量。探针的密度、分辨率和覆盖度是生物芯片的三大特征[10]。然而, 转录因子结合位点的长度往往在几个到几十个碱基, 而ChIP实验中通过对染色质随机打断得到的DNA片段长度大小不等, 有的甚至达到1 Kb, 使得ChIP-chip的分辨长度远远大于TFBS实际作用的位点区域, 这种方法并不能精确地标出转录因子结合位点。所以仍然需要计算方法的介入来更加精细地标注TFBS的位置。另外, 目前所能获得的芯片上固定的探针只能代表全基因组部分序列, 所获得的杂交信息具有偏向性。尽可能剔除假阳性并揭示出数据背后的意义是需要分子生物学与计算生物学工作者协同努力的问题。

2 高通量DNA测序技术

高通量DNA测序技术是在当今DNA测序技术研究领域产生革命性突破的前提下所实现的。随着DNA测序技术的飞速发展, 该领域技术已经对分子生物学的发展起到越来越大的作用, 也使得对基因组学的认识和思考上升到一个新的水平, 特别是其对DNA分子的序列和丰度的双重解析能力已经使研究范围从单一、局部的基因或基因片段转变为全基因组领域。美国《Times》杂志在2008年年末评出了2008年"50项最重要的发明", 其中"个人DNA测试服务"位居榜首当选该年度最佳发明。在过去的三年中, 高通量DNA测序技术对转录因子结合位点的研究亦产生深刻影响, 基于高通量DNA测序技术对转录因子结合位点的研究已成为目前研究DNA-蛋白质相互作用的标准方法。高通量DNA测序技术的

发展经过了从第一代到第三代的阶段。

2.1 第一代测序技术

DNA测序技术自发明至今在推动分子生物学发展方面一直起着关键作用, 诞生于20世纪70年代的Sanger等[11]发明的双脱氧核苷酸末端终止法和Maxam等[12]发明的化学降解法是最早被广泛应用的DNA测序技术, 主要用于解码DNA片段的碱基组成信息。例如第一代测序技术是人类基因组计划完成的基础。

1977年, Sanger等[11]发明的双脱氧链末端终止法就是根据核苷酸在某一固定的点开始, 随机在某一个特定的碱基处终止, 产生A、T、C、G四组不同长度的一系列核苷酸, 然后在尿素变性的PAGE胶上电泳进行检测, 从而获得DNA序列。后来, 研究人员陆续发明了荧光标记技术[13]、毛细管电泳以及杂交测序技术等方法。毛细管阵列电泳(Capilaries Arrays Electrophoresis)技术的应用也可视为Sanger测序方法的自动化时代[14]。

全基因组鸟枪法结合第一代测序技术已成功地组装了多种全基因组的高精度基因图谱, 例如人类[15]和果蝇[16]基因组的测序工作, 证明了它在测定大基因组上的可行性和有效性。然而由于“鸟枪法”是随机测序, 且需要对基因组进行高冗余测序, 这势必增加了测序周期和费用。虽然通过几十年的逐步改进, 尽管利用计算机技术和新的算法, 已经大大加快了拼接的速度, 但第一代测序技术在测序通量、速度及成本方面都已达到了极限。因其对电泳分离技术的依赖, 使其难以进一步提升分析速度、并行化程度和微量化, 降低测序成本, 因此, 急需开发新的测序技术来突破这些局限[17]。

2.2 第二代测序技术

随着对生命科学研究的逐步深入, DNA测序的目的从最初简单解码DNA片段的碱基组成, 过渡到如何通过解码序列去识别一个实验样本中目的DNA分子丰度及可变剪接信息。第二代测序技术是对毛细管电泳测序技术一次革命性的改变, 因其属于一类循环阵列合成测序的技术群, 可实现同时对几十万到几百万条DNA分子进行同时测定, 单次运行(run)数据量为Mbp, 故而被通称为高通量测序技术[18], 也是因以Sanger测序法为代表的第一代测序技术而得名。第二代测序中三种主流测序技术按先后顺序依次分别为: 美国的454 Life Sciences公司在2005年10月推出的GS20[19], 又称Roche/454焦磷酸测序; 随后美国Illumina公司在2006年推出Illumina/Solexa聚合酶合成测序[20], 以及2007年美国应用生物系统公司(ABI)推出ABI/SOLiD连接酶测序技术[21]。

三种二代测序技术的原理各不相同, 其数据量产出、数据质量和单Run运行成本也不一样: Roche/454焦磷酸测序方法运行速度快, 序列读段最长, 但其通量最低和准确率也不好, 对于基因组从头拼接和转录组测序, 它仍是最理想的选择[22]; ABI/SOLiD连接酶测序技术优势在于其无以伦比的通量和超精确检测模块, 测序通量大于20 GB/次, 但读长最短(50 bp), 但其创新之处在于双碱基校正技术的应用, 在测序过程中对每个碱基判读两遍, 从而减少原始数据错误, 提供内在的校对功能[23]。双重检测评价使15X覆盖率时的准确度可以达到99.999%, 超过98%的可定位碱基的质量值高于45, 是新一代基因分析技术中准确度最高的测序技术。

Illumina/Solexa测序仪因可扩展的超高通量、需要较少样品量、简单、快速、自动化等优势使其迅速成为目前使用最广泛的测序技术。Illumina/Solexa公司目前拥有三种测序平台, 分别为HiSeq 2000、HiSeq 1000、Genome Analyzer, 例如, Genome Analyzer系统目前每次运行后可获得超过20 GB的高品质过滤数据。经优化后通量还有望上升到95 GB, 相当于人类基因组的30倍覆盖度。该系统需要的样品量低至100 ng, 能应用在很多样品有限的实验中[24]。另外, 系统提供了最简单和简洁的工作流程。系统在很短时间内便可完成样品文库制备和得到高精确度的数据。因此, RNA-Seq和ChIP-Seq研究目前基本上都采用该技术。

2.3 第三代测序技术

高通量测序技术的迅猛发展, 将分子生物学水平的研究提到了一个新的高度, 也大大加速了人类对生命遗传机制和疾病分子生物学发生过程的认识。新一代测序技术的发展均以提高测序通量与读序长度、降低测序成本、简化测序步骤为目标, 在所有上述第二代测序技术中, 序列都是在荧光或者化学发光物质的协助下, 通过聚合酶将碱基连接到DNA链上过程中释放出的光学信号来确定具体碱基类型的。昂贵的光学仪器、频繁光学图像处理以及大量的试剂和耗材使测序复杂性使得成本大大增加[17]。

第三代单分子测序技术正在努力通过直接读

取序列信息, 不使用化学试剂。例如, 采用具有足够高的分辨率的非光学显微镜成像技术直接区分4种碱基; 单分子测序技术能够直接观察和测定核苷酸序列, 它将是非光学影像技术[25]、纳米技术[26]、石墨烯和碳纳米管[27]等多项技术的结晶, 例如, 非光学显微镜成像或原子力显微镜分别采用高分辨率和原子力分辨检测技术直接读取空间线性排列的具有不同物理结构特征的四类碱基[28-29]; 纳米孔单分子测序技术则根据四种碱基通过直径为1纳米的生物孔时产生的电流特异性判断碱基类型[30]、石墨烯和碳纳米管则利用碳原子排列和碳纳米管的电物理特性及其导电性的读取DNA碱基[31-32]。

目前, 第三代单分子测序技术最具代表性的包括Heliscope单分子测序仪, 单分子实时合成测序法, Oxford纳米孔测序技术等。其中, Oxford纳米孔测序技术的样本处理极其简单, 并且无需DNA聚合酶或者连接酶等, 该技术根本无需纯化的荧光素试剂, 也无需进行DNA扩增, 因此其测序成本十分低廉, 比其他测序技术更有可能实现1 000美元基因组目标[26,33]。相比较第一代和第二代测序技术, 第三代测序技术正在从生物化学或化学手段向基于物理方法的直接测序手段过渡。

3 基于高通量测序技术研究DNA-蛋白质的相互作用

高通量实验技术的出现加速了人们对转录调控机制的理解过程, 而从高通量数据中提取转录调控和转录因子结合位点相关信息是生物信息学面临的一个主要挑战。

3.1 ChIP-Seq技术的发展

ChIP-Seq是将深度测序技术与ChIP实验相结合分析全基因组范围内DNA结合蛋白结合位点、组蛋白修饰、核小体定位或DNA甲基化的高通量方法, 可以应用到任何基因组序列已知的物种, 并能确切得到每一个片段的序列信息。相对于ChIP-chip技术, ChIP-Seq是一种对ChIP富集DNA的无偏检测技术, 能够完整显示ChIP富集DNA包含的信息。ChIP-chip技术的缺点在于它是个“封闭系统”, 只能检测有限的已知的序列信息, 而ChIP-Seq优势在于其强大“开放性”, 强大的发现和寻找未知信息的能力, 所以ChIP-Seq与传统的ChIP-chip技术相比具有明显的优势[34]:

(1)灵敏度很高。传统的ChIP-chip实验要求起始DNA的量在4 μg以上, 而一般ChIP-Seq实验对起始DNA量的要求是10 ng。这直接反映在起始细胞数目的减少, 对于像早期胚胎发育相关的研究中更占优势。

(2)灵活性很强。ChIP-chip实验以研究对象的特定物种相符的全基因组DNA芯片平台为基础, 所以不适合应用于很多基因组序列信息不丰富或缺少相关芯片平台开发的物种。ChIP-Seq技术则不存在这方面的限制, 可以应用到任何基因组序列已知的物种, 并能确切得到每一个片段的序列信息。

(3)分辨率极高。传统的微阵列芯片技术受制于当前芯片的容量, 事实上不能涵盖真正的全基因组DNA序列信息, 这导致ChIP-chip的实验结果分辨率不高, 精确定位蛋白与DNA的结合位点存在一定的困难。而ChIP-Seq技术辅之以强大的生物信息计算能力, 可以高效地将测序得到的序列定位到特定基因组的精确碱基位置上, 分辨率大大提升。

(4)不具备其它一些与芯片相关的负效应, 如由核酸非特异杂交带来的噪音信号。

因此, 随着目前测序成本不断降低及通量迅速增高等优势, ChIP-Seq已经基本上取代ChIP-chip成为研究转录因子、RNA聚合酶、核小体等DNA结合蛋白体内结合靶点的主打技术。

3.2 测序数据格式及ChIP-Seq分析步骤

目前, Sanger测序和Illumina公司测序仪器产出的测序数据基本都是FASTQ格式, 即一种含有测序质量的FASTA文件[35]。FASTQ格式以测序读段为单位存储, 每条读段占四行, 第一行开头为“@”后接读段标识, 第二行为测序出的碱基序列, 第三行开头为“+”后接读段ID, 因读段ID一般与第一行相同, 所以有时可以省略以节省空间。第四行为测序质量, 一般用字符表示, 长度与第二行相同, 对应于相应位置碱基的测序质量。目前, FASTQ格式主要已有多个版本, 如Sanger格式、Illumina 1.8+格式、Solexa格式、Illumina 1.3+格式等; 其中, Illumina 1.8+格式采用Phred+33编码方式, 即用ASCII 33至126对应的字符编码Phred质量得分的0至93。

染色质免疫沉淀技术是相对比较成熟的技术, ChIP-Seq的难点是测序后的生物信息学分析。DNA打碎方法、染色质开放程度的不均一性、PCR扩增偏向性、基因组的重复程度以及测序和序列比对过

程中的错误都会引入系统误差造成假阳性, 尽可能剔除假阳性并揭示出数据背后的机制是需要分子生物学与计算生物学工作者协同努力的问题。通常基于ChIP-Seq技术研究转录因子结合位点的生物信息学分析主要有以下步骤: (1)数据预处理及读段定位; (2)读段富集区域识别; (3)结合位点识别; (4)结合位点功能注释及靶基因Pathway分析[36](图1B)。

3.3 双末端标记测序系统(PET)技术的产生

新一代测序平台是低花费和高通量的, 但其阅读长度较短。根据Chip-Seq技术流程, 测序前染色质往往被随机打断为长度不等的片段序列, 片段长度普遍较长, 而目前的高通量测序技术只能测得片段5′-端的一小部分, 且不能直接测出实际结合位点的富集区域。最近, Fullwood等[37]开发的基因组学研究方法体系——全基因组双末端标记测序系统(paired-end-tag, PET)改进了上述缺陷, PET是一种革新性的基因测序方法, 这种方法能帮助解决新一代测序平台阅读长度较短的缺点问题, 利用多种应用的配对末端标签PET测序[38], 从超高通量测序的常DNA片段末端引出短的配对标签, 从而就能精确地绘制基因组, 区别PET所在DNA片段的基因组边界和鉴定靶DNA片段。PET测序法已经发展为转录组, 转录因子结合位点和染色体结构分析。PET测序技术的独特优点在于能够暴露DNA片段两末端的连接处。由于该优点, PET测序可以揭示非传统的融合转录物, 染色体结构变化, 甚至分子相互作用。

PET测序技术通过将前期实验得到较长的DNA片段, 取出其前后两个端的小段序列, 称作paired-end-tags, 将这些tags进行测序然后map到基因组上, 从这两个tags就可以知道得到了基因组上哪段区域。PET可以有效降低冗余的数据量, 从而降低测序费用。PET测序可以发现DNA片段两端的链接关系, 由这一特征可以用于进一步发现融合的转录子、基因组上的结构变异以及染色体上的远距离相互作用[39]。随后, Wei等[40]集合PET和ChIP两者的优势, 开发出了称为ChIP-PET的新实验系统, 用于基因组层面蛋白结合位点的鉴定研究, 他们对包含约50万个PET序列的饱和样品进行了分析, 获得了65 572个单独的ChIP实验DNA片段, 建立了用特殊标记标定p53结合位点的重叠PET簇。在这一基础上, 研究人员最终确定了最少542个p53精确结合位点, 包括98个之未发现的p53目标基因。

ChIA-PET(chromatin interaction analysis using paired-end-tag sequencing)分析方法是一种新型的测序方法[41], ChIA-PET技术是利用PET测序技术研究免疫沉淀后邻近式连接的DNA片段, 以得到染色质相互作用的技术。ChIA-PET技术的出现, 对于识别具有单一功能性转录因子的结合位点是一个重要突破, 相比较ChIP-Seq技术, ChIA-PET技术在确定结合位点的同时, 能够确定结合位点间的相互作用[42]。同时, 这也是其与ChIP-PET技术的唯一区别。另外, ChIA-PET技术通过超声打断DNA-蛋白质复合体, 有效减少因利用限制性内切酶引入的一系列相互作用的噪音[43]。

4 ChIP-Seq用于研究干细胞分化中的转录因子调控过程

胚胎发育中的器官发生是一个十分复杂的生物学过程。然而研究其内在本质, 往往是由前体组织接收诱导信号后启动一些转录因子的活性, 再由它们下游调控基因的各种纷繁复杂的调控来完成器官发生所需的一系列细胞分化和形态发生的过程。在器官发生过程中一些关键转录因子起着至关重要的作用, 在全基因组水平将转录因子定位于靶基因DNA是认识转录调控网络的有效方法之一。

2007年, Johnson等[44]开创了一种基于直接超高通量DNA测序的大规模染色质免疫沉淀分析ChIP-Seq方法, 将这种测序方法用于体内神经限制性沉默因子(neuron-restrictive silencer factor, NRSF)与人类基因组中1 946个位点的结合作图, 得到了高结合位点分辨率[±50碱基对]的体内蛋白-DNA相互作用图谱。同年, Robertson等[45]也宣布发明了ChIP-Seq技术, 并用其定位了转录因子STAT1在正常的人类Hela S3细胞和受IFN-gamma刺激后的人类HelaS3细胞中的全基因组上的结合情况。他们比较ChIP-Seq与ChIP-PCR、ChIP-chip并用后两者的结果进一步证明了ChIP-Seq的特异性和敏感性。

4.1 ChIP-Seq在胚胎干细胞多能性调控中的研究

胚胎干细胞(embryonic stem cells, ESCs)是一种高度未分化细胞。它具有发育的多能性, 能分化出成体动物的所有组织和器官, 包括生殖细胞。Oct4、Nanog、Sox2等多能性相关转录因子通过自调节或反馈调节等方式调控胚胎干细胞的自我更新和分化。2008年, Chen等[46]用ChIP-Seq检测了Nanog、

Oct4、STAT3、Smad1和Sox2等13个序列特异性的转录因子与基因组DNA的结合情况。这些转录因子都是LIF和BMP途径的重要调控分子, 发现有些位点是几个转录因子结合共同起调节作用, 如Nanog、Sox2、Oct4、Smad1和STAT3共同组成胚胎干细胞增强体(ES-cell-specific enhanceosomes)。Ouyang等[47]利用ChIP-Seq技术研究小鼠胚胎干细胞, 发现大约有65%的基因表达是由12个转录因子调控的。他们鉴定了两组转录因子。其中第一组(E2f1、Myc、Mycn和Zfx)通常作为激活剂起作用, 第二组(Oct4、Nanog、Sox2、Smad1、Stat3、Tcfcp2l1、Esrrb)可能依赖于靶点不同或作为激活剂, 或作为抑制剂。这两组转录因子紧密协作, 激活胚胎干细胞中差异化上调的基因。在缺乏第一组转录因子结合时, 第二组转录因子结合胚胎干细胞中被抑制的基因及早期分化中去抑制的基因。最近, Li等[48]证实组蛋白乙酰转移酶Mof在维持胚胎干细胞自我更新和多能性中发挥了至关重要的作用。Mof缺失导致胚胎干细胞失去特征性形态、碱性磷酸酶(AP)染色和分化潜能。这些细胞的核心转录因子Nanog、Oct4和Sox2表达异常, 当Nanog过表达时可部分抑制无Mof的胚胎干细胞的表型, 说明胚胎干细胞中Mof是Nanog的上游调控因子。他们利用ChIP-Seq进一步证实了Mof是ESC核心转录网络的一个必需元件, 主导了不同发育程序的基因。此外, Mof也在关键调控位点中招募Wdr5和H3K4甲基化中起重要作用, 从而突显了胚胎干细胞中各种染色质调控因子的复杂的网络调控。

4.2 ChIP-Seq在肝脏转录因子调控中的研究

肝脏是一个以代谢功能为主的器官, 并扮演着去氧化、储存肝糖、合成分泌性蛋白质等角色。同时, 肝脏也制造消化系统中的胆汁。Foxa2等转录因子在肝脏发育和代谢过程中起着重要作用。Wederell等[49]研究成年小鼠肝脏组织中转录因子Foxa2结合位点, 一共识别了11 000个位点, 其中43.5%的肝脏表达基因含有相关的Foxa2结合位点。Schmidt等[50]利用ChIP-Seq技术研究多种脊椎动物肝脏中表达的两个转录因子CEBPA和HNF4A, 虽然两个转录因子都有高度保守的DNA结合结构域, 绝大多数情况下表现出种属特异性。 Bochkis等[51]通过ChIP-Seq全基因组定位分析高度同源的转录因子Foxa1和Foxa2, 发现虽然Foxa1和Foxa2肝脏中的结合位点大部分重叠, 在体内它们各自还有跟其他调控元件结合。Foxa1和p53共同结合位点附近的基因经常在细胞周期调控中起到重要作用。Foxa2独有结合位点附近的基因与类固醇和脂质代谢相关。因此, Foxa1和Foxa2在成体的肝脏中展示出不同的角色, 确保这两个基因在进化过程中的重要性。肝脏的发育经历了一系列内胚层和中胚层之间复杂的相互作用, 在我们先期的研究中应用ChIP-Seq技术发现了Foxa2在DE细胞全基因组中的结合位点[52], 并在全基因组的水平上证明Foxa2具有先锋因子的作用, 且Foxa2对特异靶基因的作用是通过修饰靶基因promoter/enhancer片段中的H3K4me2组蛋白位点实现。在肝向分化研究中我们发现个别新的肝向基因的promoter/enhancer片段上找到GATA4与Foxa2相互作用的新证据。

核受体是与类固醇激素受体同源的一类配体依赖性转录因子超家族, 主要有法呢醇受体(farnesoid X receptor, FXR)、肝脏受体类似物-1(liver receptor-homolog-1, LRH-1)、肝脏X受体(liver X receptors, LXR)、孕烷X受体(pregnane X receptor, PXR)等组成复杂的转录调控网络调节各种代谢活动。2010年, Hansook等[53]利用ChIP-Seq分析转录因子FXR在肝细胞染色质中的结合, 发现一个额外的核受体半位点(half-site)和FXR结合。他们根据FXR转录因子的DNA结合谱, 推测这个FXR结合的核受体是LRH-1转录因子。为了进一步证实在整个基因组范围内LRH-1是否与FXR具有相互作用, 2012年, Hansook等[54]分析LRH-1在全基因组中的结合位点。他们检测超过10 600的LRH-1在肝染色质中结合域, 其中超过20%的结合域位于已知小鼠基因的5′-端的2 Kb区域内。此外, 分析结果还表明, LRH-1结合的基因位点和FXR的结合位点靠近。LRH-1/FXR共同结合的基因都与脂质代谢相关。这些结果表明, LRH-1招募FXR激活脂质代谢相关基因的表达。研究还发现, 部分FXR跟LRH-1没有共同结合域, 表明FXR可能跟RORs和NR4As家族成员相互作用, 调节其他代谢途径, 比如FXR已被证明在葡萄糖代谢中发挥重要的作用。

4.3 ChIP-Seq在其它转录因子调控中的研究

细胞核接头蛋白LDB1是多蛋白转录复合体中的一个核心成分。Li等[55]研究发现, 在小鼠胚胎和成体造血干细胞中LDB1起到关键作用。在造血干

细胞和前体细胞中敲除该基因会造成一些维持多能性相关基因的转录下调, ChIP-Seq结果显示, LDB1形成的复合体结合在这些基因的启动子区域, 暗示LDB1维持造血干细胞中起核心作用。

多能造血前体细胞是一种干细胞样细胞, 表达各种基因, 且有分化成包括免疫系统细胞在内的大量不同类型血细胞的能力。Zhang等[56]利用ChIP-Seq和RNA-seq技术在小鼠全基因组内找出造血前体细胞转化成定型T细胞中起作用的所有基因, 并确定了每个基因在发育过程中的转录时间点。同时, 还追踪了指引前体细胞到各种替代途径中的基因。结果不仅揭示了T细胞发育过程的时空调控网络, 也表明了T细胞发育过程如何关闭启动替代命运的基因。

少突胶质细胞(oligodendrocyte, OL)是中枢神经系统的髓鞘形成细胞, 来源于其前体细胞。Olig2是一种少突胶质细胞转录因子, 与前体细胞增殖和向少突胶质细胞分化密切相关。Yu等[57]利用ChIP-Seq进行全基因组分阶段研究发现Olig2充当了一种预定位(prepatterning)因子, 引导染色质重塑酶Smarca4/Brg1到达少突胶质细胞活性靶点, 从而通过Brg1激活少突胶质细胞特定基因表达。Olig2和Brg1在染色质上的结合还指导重要的表观遗传修饰信息, 并预测重要的少突胶质细胞分化调控基因。

5 结语与展望

随着功能基因组学研究的深入开展, 通过染色质水平研究基因的表达调控成为全面阐明真核基因表达调控机制的必经之路。ChIP技术能够揭示转录因子的结合位点和确定直接的靶基因序列, 可在体内分析特定启动子的分子调控机制, 因此被广泛应用于转录调控机制的研究方面。由于传统的Sanger测序由于测序成本和速率限制, 分析大量的免疫沉淀DNA序列从时间和花费上都是不现实的。高通量测序技术对DNA分子的序列兼丰度的双重解析能力为ChIP技术的应用如虎添翼。这也是ChIP-Seq技术迅速成为目前转录因子研究的主导技术的优势。

然而, ChIP-Seq技术在研究转录因子结合位点上也存在的一定的缺陷, 例如, ChIP实验抗体品质较差往往容易造成非特异性DNA沉淀, 产生高背景和假阳性。另外, ChIP实验只能鉴定出转录因子在很短时间处于结合态的DNA序列。由于受众多因素的干扰, 使得部分蛋白质结合的DNA难以被ChIP捕获。另外, 结合位点与功能的不对等性、结合位点与靶基因的对应关系不确定性以及二代测序技术普遍的读段较短等问题都对转录因子结合位点的深入挖掘造成一定影响。而ChIA-PET技术也存在一定的局限性, 如只能检测依赖蛋白质因子的染色质相互作用, 不能确定蛋白质因子中每一种相互作用的蛋白质。随着人们对转录调控过程的了解的深入, 只有通过各种数据的融合和相互校正,以及其他新技术的介入, 才能挖掘出可靠的转录调控关系和TFBSs。

尽管基于Chip-Seq和ChIA-PET转录因子分析技术还存在上述缺点, 但这两种技术借鉴了ChIP技术的优点, 利用特异性的抗体捕获与蛋白质因子结合的染色质, 进而得到蛋白质因子结合位点和结合位点间的相互作用, 解决了全基因组分析的复杂性, 所以ChIP技术仍然是目前研究转录因子体内结合靶点的难以替代的核心技术, 因此, 该技术在转录因子体内结合谱的研究中仍将发挥重要作用。ChIP-Seq等技术通过系统整合DNA与蛋白质相互作用的数据, 在揭示基因表达调控的若干机制及构建更加详细的基因表达调控网络图谱中发挥无可替代的作用。

参考文献 (References)

1 Messina DN, Glasscock J, Gish W, Lovett M. An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. Genome Res 2004; 14(10B): 2041-7.

2 Muller F, Demeny MA, Tora L. New problems in RNA polymerase II transcription initiation: matching the diversity of core promoters with a variety of promoter recognition factors. J Biol Chem 2007; 282(20): 14685-9

3 Im H, Grass JA, Johnson KD, Boyer ME, Wu J, Bresnick EH. Measurement of protein-DNA interaction *in vivo* by chromatin immunoprecipitation. Methods Mol Biol 2004; 284: 129-46.

4 Elnitski L, Jin VX, Farnham PJ, Jones SJ. Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques. Genome Res 2006; 16(12): 1455-64.

5 Collas P. The current state of chromatin immunoprecipitation. Mol Biotechnol 2010; 45(1): 87-100.

6 Orlando Strutt H, Paro R. Analysis of chromatin strncture by *in vivo* form aldehyde cross−linking. Methods 1997; 1l(2): 205-14.

7 Lee TI, Johnstone SE, Young RA. Chromatin immunoprecipitation and microarray-based analysis of protein location. Nat Protoc 2006; 1(2): 729-48.

8 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): 70-6.
9 Sun JM, Chen HY, Davie JR. Differential distribution of unmodified and phosphorylated histone deacetylase 2 in chromatin. J Biol Chem 2007; 282(45): 33227-36
10 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(Li Minli, Wang Wei, Lu Zuhong. Genomic analysis of DNA-protein interaction by chromatin immunoprecipitation. Hereditas) 2010; 32(3): 219-28.
11 Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977; 74(12): 5463-7.
12 Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci USA 1977; 74(2): 560-4.
13 Smith LM, Fung S, Hunkapiller MW, Hood LE. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: Synthesis of fluorescent DNA primers for use in DNA sequence analysis. Nucleic Acids Res 1985; 13(7): 2399-412
14 Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, *et al*. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics 2008; 24(24): 2818-24.
15 Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, *et al*. The Sequence of the Human Genome. Science 2001; 291(5507): 1304-51.
16 Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, Patel S, *et al*. Finishing a whole-genome shotgun: Release 3 of the *Drosophila* melanogaster euchromatic genome sequence. Genome Biol 2002; 3(12): research0079.
17 周晓光, 任鲁风, 李运涛, 张 猛, 俞育德, 于 军. 下一代测序技术: 技术回顾与展望. 中国科学: 生命科学(Zhou Xiaoguang, Ren Lufeng, Li Yuntao, Zhang Meng, Yu Yude, Yu Jun. The next-generation sequencing technology: A technology review and future perspective. Sci China Life Sci) 2010; 40(1): 23-37.
18 Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods 2008; 5(1) : 16-8.
19 Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, *et al*. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005; 437(7057): 376-80.
20 Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008; 456(7218): 53-9.
21 Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, *et al*. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Res 2008; 18(7): 1051-63.
22 Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet 2008; 24(3): 133-41.
23 Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, *et al*. Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res 2008; 18(10): 1638-42.
24 Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour 2011; 11(5): 759-69.
25 Tanaka H, Kawai T. Partial sequencing of a single DNA molecule with a scanning tunnelling microscope. Nat Nanotechnol 2009; 4(8): 518-22.
26 Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol 2009; 4(4): 265-70.
27 Albertorio F, Hughes ME, Golovchenko JA, Branton D. Base dependent DNA-carbon nanotube interactions: Activation enthalpies and assembly-disassembly control. Nanotechnology 2009; 20(39): 395101-9.
28 Driscoll RJ, Youngquist MG, Baldeschwieler JD. Atomic-scale imaging of DNA using scanning tunnelling microscopy. Nature 1990; 346(6281): 294-6.
29 Ikai A. STM and AFM of bio/organic molecules and structures. Sur Sci Rep 1996; 26: 261-332.
30 Butler TZ, Pavlenok M, Derrington IM, Niederweis M, Gundlach JH. Single-molecule DNA detection with an engineered MspA protein nanopore. Proc Natl Acad Sci USA 2008; 105(52): 20647-52.
31 Postma HW. Rapid sequencing of individual DNA molecules in graphene nanogaps. Nano Lett 2010; 10(2): 420-5.
32 Albertorio F, Hughes ME, Golovchenko JA, Branton D. Base dependent DNA-carbon nanotube interactions: Activation enthalpies and assembly-disassembly control. Nanotechnology 2009; 20(39): 395101.
33 Stoddart D, Heron AJ, Mikhailova E, Maglia G, Bayley H. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. Proc Natl Acad Sci USA 2009; 106(19): 7702-7.
34 Park PJ. ChIP-seq: Advantages and challenges of a maturing technology. Nat Rev Genet 2009; 10(10): 669-80.
35 Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 2010; 38(6): 1767-71.
36 Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, *et al*. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 2008; 5(9): 829-34.
37 Fullwood MJ, Wei CL, Liu ET, Ruan Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res 2009; 19(4): 521-32.
38 Chiu KP, Wong CH, Chen Q, Ariyaratne P, Ooi HS, Wei CL, *et al*. PET-Tool: A software suite for comprehensive processing and managing of Paired-End diTag (PET) sequence data. BMC Bioinformatics 2006; 7: 390.
39 Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, *et al*. CTCF-mediated functional chromatin interactome in pluripotent cells. Nat Genet 2011; 43(7): 630-8.
40 Wei CL,Wu Q, Vega VB, Chiu KP, Ng P, Zhang T, *et al*. A global map of p53 transcription-factor binding sites in the human genome. Cell 2006; 124(1): 207-19.
41 Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohanmed YB, *et al*. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 2009; 462(7269): 58-64.
42 Li G, Fullwood MJ, Xu H, Mulawadi FH, Velkov S, Vega V, *et al*. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol 2010; 11(2): R22.
43 Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, *et al*. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 2012; 148(1/2): 84-98.

44 Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of *in vivo* protein-DNA interactions. Science 2007; 316(5830): 1497-502.

45 Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, *et al*. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 2007; 4(8): 651-7.

46 Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, *et al*. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 2008; 133(6): 1106-17.

47 Ouyang Z, Zhou Q, Wong WH. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proc Natl Acad Sci USA 2009; 106(51): 21521-6.

48 Li X, Li L, Pandey R, Byun JS, Gardner K, Qin Z, *et al*. The histone acetyltransferase MOF is a key regulator of the embryonic stem cell core transcriptional network. Cell Stem Cell 2012; 11(2): 163-78.

49 Wederell ED, Bilenky M, Cullum R, Thiessen N, Dagpinar M, Delaney A, *et al*. Global analysis of *in vivo* Foxa2-binding sites in mouse adult liver using massively parallel sequencing. Nucleic Acids Res 2008; 36(14): 4549-64.

50 Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, *et al*. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 2010; 328(5981): 1036-40.

51 Bochkis IM, Schug J, Ye DZ, Kurinna S, Stratton SA, Barton MC, *et al*. Genome-wide location analysis reveals distinct transcriptional circuitry by paralogous regulators Foxa1 and Foxa2. PLoS Genet 2012; 8(6): e1002770.

52 Xu C, Lu X, Chen EZ, He Z, Uyunbilig B, Li G, *et al*. Genome-wide roles of Foxa2 in directing liver specification. J Mol Cell Biol 2012; 4(6): 420-2.

53 Chong HK, Infante AM, Seo YK, Jeon TI, Zhang Y, Edwards PA, *et al*. Genome-wide interrogation of hepatic FXR reveals an asymmetric IR-1 motif and synergy with LRH-1. Nucleic Acids Res 2010; 38(18): 6007-17.

54 Chong HK, Biesinger J, Seo YK, Xie X, Osborne TF. Genome-wide analysis of hepatic LRH-1 reveals a promoter binding preference and suggests a role in regulating genes of lipid metabolism in concert with FXR. BMC Genomics 2012; 13: 51.

55 Li L, Jothi R, Cui K, Lee JY, Cohen T, Gorivodsky M, *et al*. Nuclear adaptor Ldb1 regulates a transcriptional program essential for the maintenance of hematopoietic stem cell. Nat Immunol 2011; 12(2): 129-36.

56 Zhang JA, Mortazavi A, Williams BA, Wold BJ, Rothenberg EV. Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. Cell 2012; 149(2): 467-82.

57 Yu Y, Chen Y, Kim B, Wang H, Zhao C, He X, *et al*. Olig2 targets chromatin remodelers to enhancers to initiate oligodendrocyte differentiation. Cell 2013; 152(1/2): 248-61.