

# 下一代测序技术在干细胞转录调控研究中的应用

刘亚军, 张峰, 刘宏德, 孙啸

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

**摘要:** 基因转录调控及其机制分析是后基因组时代生物学研究的重点之一。随着高通量测序技术的发展, 人们可以从不同层面研究基因的转录调控行为, 从转录组、转录因子结合, 到染色质局部结构和整体空间构象, 可系统分析转录调控的分子机制。干细胞分化过程的转录调控分析对研究再生医学和理解细胞癌变机制等具有重要意义。本文综述了下一代测序技术在干细胞转录调控研究中的应用, 包括: (1) 基于基因芯片或 RNA 测序的转录组分析; (2) 基于染色体免疫共沉淀(chromatin immunoprecipitation, ChIP)测序的表观基因组和转录因子结合信息的分析; (3) 基于 DNase 酶切测序(DNase-Seq)的染色质开放性分析; (4) 基于高通量染色质构象捕获(high-throughput chromosome conformation capture, Hi-C)技术的染色体远程相互作用分析。从基因表达谱、转录因子结合和基因组三维结构等层面展开介绍, 重点关注了一些多能性转录因子(Oct4、Sox2 和 Nanog 等)在维持干细胞干性和分化中的调控作用, 以期为干细胞转录调控的研究提供借鉴和参考。

**关键词:** 胚胎干细胞; 转录组; ChIP; DNase-seq; Hi-C

## The application of next-generation sequencing techniques in studying transcriptional regulation in embryonic stem cells

Yajun Liu, Feng Zhang, Hongde Liu, Xiao Sun

State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China

**Abstract:** The mechanism of transcriptional regulation has been the focus of many studies in the post-genomic era. The development of sequencing-based technologies for chromatin profiling enables current researchers to experimentally measure chromatin properties. Moreover, many studies aim at annotating the state of the chromatin into broad categories based on observed chromatin features and/or DNA sequences, then associating the resultant distal regulatory regions with the correct target genes based on DNA sequences, and predicting the dependence of epige-

收稿日期: 2017-02-27; 修回日期: 2017-05-08

基金项目: 江苏省重点研发计划项目(编号: BE2016002-3), 国家自然科学基金项目(编号: 31371339)和国家国家重点基础研究发展计划项目(973 计划)(编号: 2012CB316500)资助[Supported by the Key Research and Development Program of Jiangsu province (No. BE2016002-3), the National Natural Science Foundation of China (No. 31371339) and the National Basic Research Program of China (973 Program) (No. 2012CB316500)]

作者简介: 刘亚军, 博士, 专业方向: 生物信息学。E-mail: liuyajun\_biology@126.com

张峰, 硕士研究生, 专业方向: 生物信息学。E-mail: zf564js@163.com

刘亚军和张峰为并列第一作者。

通讯作者: 孙啸, 博士, 教授, 专业方向: 生物信息学。E-mail: xsun@seu.edu.cn

DOI: 10.16288/j.yczs.16-390

网络出版时间: 2017/7/4 16:20:57

URI: <http://kns.cnki.net/kcms/detail/11.1913.R.20170704.1620.008.html>

netic features on genetic variation. Stem cell biology has many applications in the area of regenerative medicine and tumorigenesis. In this review, we summarize recent research progresses on the application of next-generation sequencing techniques in studying transcriptional regulation in embryonic stem cells. This review mainly focuses on four areas: (1) microarray or RNA-seq; (2) chromatin immunoprecipitation (ChIP); (3) Dnase I hypersensitive sites (DHSs); (4) high-throughput chromosome conformation capture (Hi-C). These technologies have been utilized in studying chromatin on three levels, i.e., gene expression, transcription factor binding and genome three-dimensional structure. We especially emphasize three master transcription factors of pluripotency: Oct4, Sox2 and Nanog. We aim to track the frontier of stem cell transcriptional regulation research and share important progresses in this field.

**Keywords:** embryonic stem cell; transcriptome; ChIP; DNase-seq; Hi-C

胚胎干细胞(embryonic stem cell, ESC)是一类具有无限增殖能力和分化多能性的细胞<sup>[1]</sup>。在果蝇和哺乳动物中,过表达某一基因能够激活其他成体细胞的特征性基因表达,从而导致细胞的功能发生改变<sup>[2,3]</sup>,但是这种单因子的诱导方式,仅能实现细胞向相似细胞类型细胞的转变<sup>[2]</sup>。2006年,Yamanaka实验室证明在过表达4种转录因子 *OCT4*、*SOX2*、*c-Myc* 和 *KLF4* 后,分化细胞可转变为诱导多能性干细胞(induced pluripotent stem cells, iPSCs)<sup>[3,4]</sup>,表明转录因子调控对于细胞多能性的建立至关重要。

转录因子(transcriptional factors, TFs)能识别特定的 DNA 序列并激活或抑制基因的转录,Jacob 和 Monod 对大肠杆菌 *Lac* 操纵子的研究为理解转录调控奠定了基础<sup>[5]</sup>。在基因组中,约 10%的编码基因转录翻译成为转录因子<sup>[6]</sup>,确定转录因子调控的靶基因是重建活细胞基因调控网络的前提。先前的研究多关注于启动子区近端的调控,如 DNA 序列特征、转录因子结合的基序(motif)、转录复合物的装配、转录的方向及协同作用等方面。传统方法往往通过构建单个基因启动子报告载体进行研究,但这种方法工作量大且效率低下。伴随着高通量测序技术的迅速发展,逐步实现了在全基因组水平上对各调控元件进行研究。本文综述了下一代测序技术在干细胞转录调控研究方面的应用。

## 1 基于基因芯片或 RNA 测序的干细胞转录组分析

高通量测序技术使得对一个物种的基因组和转录组进行全面细致的分析成为可能<sup>[7]</sup>。其中,微阵列(Microarray)和 RNA 测序(RNA-seq)技术是研究基

因组和转录组的重要手段。小鼠功能注释(functional annotation of the mouse, FANTOM)计划、人类基因组百科全书(encyclopedia of DNA elements, ENCODE)<sup>[8]</sup>计划等利用高通量测序技术,产生了大量的转录组数据和信息,这些计划的推行极大地促进了对真核转录调控规律的研究。

基因组上的某一段编码序列在转录之后平均能得到 6 种 mRNA,如果考虑到非编码 RNA 以及广泛存在的正反义链等因素,基因组序列和转录组信息之间的关系将会更加复杂。基因芯片和 RNA-seq 被广泛用于检测干细胞转录组,通过 RNA 干扰技术(RNA interference, RNAi)和转录组学手段的结合利用,可以寻找并识别与胚胎干细胞自我更新相关联的关键调控因子<sup>[9]</sup>。通过干扰关键的转录因子,对胚胎干细胞基因表达谱变化进行分析能深入理解多能性转录调控网络<sup>[10]</sup>。例如:下调 Oct4、Nanog 或 Sox2 可以引起转录组发生相似的动态变化模式,这表明这 3 个转录因子之间存在调控下游基因转录的协同作用。由于 RNA-seq 相对廉价,所以被广泛应用于各类物种的干细胞调控研究中。猪是优良的人类疾病模型动物,在药物评测中有重要价值,对猪干细胞进行转录组分析和基因表达调控研究的结果<sup>[11-13]</sup>表明转录组技术手段能揭示猪诱导多能干细胞的转录调控信息。但是为了深入和精确地检测转录因子之间相互作用的关系,仍需要对转录因子结合位点及其靶基因的调控做进一步的研究。

## 2 基于免疫共沉淀测序(ChIP-seq)的胚胎干细胞转录调控分析

转录因子可结合在启动子区,进而能调控下游

基因的表达<sup>[14]</sup>。X 射线晶体衍射、原子力显微镜、核磁共振波谱、扫描力显微镜等技术曾被用来进行转录因子和 DNA 的结合分析,但这些技术很难在全基因组尺度上分析转录因子的作用。基于蛋白质序列信息,以机器学习的模型可以预测蛋白质与核酸的相互作用<sup>[15,16]</sup>,但这些结果往往需要通过进一步的体内实验来确认。

高通量测序技术被广泛应用于转录因子结合位点的测定,将染色质免疫共沉淀 ChIP(chromatin immunoprecipitation)与第二代测序相结合的 ChIP-seq 技术,能在全基因组范围内测定那些与转录因子相结合的 DNA 序列信息<sup>[17,18]</sup>。通过特定算法分析得到转录因子结合域位点<sup>[19]</sup>,进而预测临近基因的表达<sup>[20]</sup>。转录因子结合与基因表达之间的相关性在细菌中已得到确证,但在真核细胞中,转录因子的结合与下游基因表达间的相关性并不显著,Gao 等<sup>[20]</sup>发现酵母中 67%的转录因子结合方式同基因表达谱间没有显著相关性。这表明基因的表达水平受到多个层次、多种方式的调控。

在多能性细胞中共表达的基因极可能受共同的转录因子调控,Oct4、Sox2 和 Nanog 能同时调控多个基因的表达,因此这些转录因子被认为是干细胞维持干性的核心因子。POU 家族转录因子主要在细胞核发育、胚胎及体外胚胎干细胞和胚胎生殖细胞(embryonic germ cells, EG)中表达<sup>[21]</sup>。Sox2 与 Oct4 协同作用,激活 Oct4、Sox2 和 Nanog 的增强子<sup>[22~24]</sup>。Oct4、Sox2 和 Nanog 中的任意一个基因都能靶定到另两个基因上,它们可以形成自组织反馈环来保证维持多能性的基因表达水平,进而形成互相调控。敲除 Nanog 后胚胎干细胞具有分化倾向,但仍能通过表达 Oct4 和 Sox2 来保持其多能性,这说明 Oct4 和 Sox2 是多能性构建的必需因子,而 Nanog 可以增强多能性水平<sup>[25]</sup>。

Walker 等<sup>[26]</sup>描绘 Oct4、Sox2 和 Nanog 之间的调控关系,发现多梳基因(Polycomb group genes)编码蛋白 Eed 和 Phc1 都受到 Oct4 和 Nanog 的调控,随着 Oct4 和 Nanog 的下调而下调,且 Eed 和 Phc1 的下游靶基因也受到了 Oct4、Sox2 和 Nanog 的靶定。Kim 等<sup>[27]</sup>检测了 6 个转录因子在小鼠胚胎干细胞中的功能,结果表明被较少的转录因子所靶定的基因

的表达会受到抑制,而当基因被超过 4 个转录因子靶定时则会被激活。Chen 等<sup>[28]</sup>用 ChIP-seq 确定了 13 个转录因子的转录调控情况,结果表明在 ES 细胞中上调的大多数基因受 Nanog、Oct4、Sox2、Smad1 及 STAT3,或 c-Myc 和 n-Myc 调控。Smad1(BMP 通路相关)和 STAT3(LIF 通路相关)同 Nanog、Oct4 和 Sox2 有许多共同结合位点。最近,有研究利用 ChIP-seq 技术分析了在诱导重编程各个阶段,Oct4、Sox2、Klf4 和 cMyc 的顺式调控机制,发现多能性调控因子在这一过程中多采用共调控策略<sup>[29]</sup>。

除了核心多能性转录因子,进化上保守的 Kruppel 样锌指转录因子家族也被证明同 Oct3/4 和 Sox2 共同作用激活下游一系列靶基因,但 ES 自我更新和多能性维持并不是 Klf4 依赖的。用维生素 A 诱导 ES 细胞分化可以检查到 901 个基因迅速下调,其中包括了 Klf 家族中的 Klf2、Klf4 和 Klf5<sup>[9]</sup>,这 3 个基因的作用具有互补性,ES 细胞在同时干扰掉这 3 个基因后分化。ChIP-chip 试验表明 Klf2, Klf4 和 Klf5 能调控多能性转录因子的表达<sup>[30]</sup>。

### 3 基于 DNase-seq 的胚胎干细胞转录调控分析

核小体是真核生物表观调控的一个重要环节,核小体的分布对于胚胎干细胞的转录调控至关重要,不同的核小体占据模式与基因的表达水平有关<sup>[31]</sup>。核小体是决定 DNA 序列对转录因子开放程度(accessibility)的重要因素<sup>[32]</sup>,转录因子与 DNA 的结合是通过与核小体竞争,导致核小体移出或滑动,进而使得转录因子结合到 DNA 上来。在转录因子结合位点处的核小体水平较低<sup>[33]</sup>,疏松状态的染色质区易被 DNase I 切割,形成 DNase 高敏位点(Dnase I hypersensitive sites, DHSs);而转录因子与基因组 DNA 相结合可以保护该结合区的 DNA 序列不被 DNase I 消化,这会使得转录因子结合处留下一个特别的“足迹”<sup>[34]</sup>,通过 DNase-seq 分析不同细胞的 DHSs,从而发现基因组上 DNA 调控元件,并通过 Motif 富集分析找出潜在的转录因子结合位点,此外,还可以利用这项技术获得转录因子-核小体互作信息以及调控元件周围核小体分布结构信息<sup>[35]</sup>。

相对 ChIP-seq 实验, DNase-seq 不需要制备特异性抗体,只需一次实验就可以鉴定出全基因组潜

在的调控元件。Stergachis 等<sup>[36]</sup>利用胚胎干细胞 DNase-seq 数据结合转录因子序列模体 Motif 特征,预测了 ES 细胞 DHSs 处富集的转录因子,成功识别出了 Oct4、Sox2 和 Nanog 等重要的转录因子,表明了这种策略在鉴定基因组调控元件方面的可行性。

干细胞转录调控分析呈现数据驱动的特点<sup>[37]</sup>,传统的转录因子结合预测依赖位置权重矩阵,其缺陷是需要大量的数据优化和确定相应的评分值。k-mer-SVM 方法预测调控 DNA 序列,这种方法使用短的 k-mer(6~8 bp)预测基因组较长功能序列元素的活性(500~2000 bp)<sup>[38]</sup>。基于核函数分类的方法需要对所有 N 个训练序列进行两两比较,计算和存储复杂性大,而卷积神经网络(convolutional neural networks, CNN)由于其深度学习特性,已成功用于预测转录因子结合位点<sup>[39]</sup>,深度学习模型<sup>[39]</sup>对高通量测序数据的分析具有广泛的应用前景。

#### 4 基于 Hi-C 技术研究胚胎干细胞转录调控

ENCODE 项目研究表明<sup>[40]</sup>,人类基因组大部分序列被潜在的调控元件占据着,这些元件中有许多与其调控目标有相当大的距离(数百 kb 或兆)。ChIP-seq 和 DNase-seq 技术能获得这些位点的线性分布信息,但无法获取调控元件的空间位置信息。

染色体构象捕获(3C)及其衍生技术的应用,可以通过研究远端调控元件与其靶基因在空间上的物理接触来分析转录调控。3C 技术最早是为了研究少数几个基因位点与其调控元件间的相互作用而产生的,其衍生技术—Hi-C 技术,试图捕获全基因组范围的染色质位点相互作用,通过简单的酶切、标记生物素、连接等步骤将染色质三维空间上所有位点间的相互作用转变成能被 PCR 检测的线性结构。Hi-C 技术捕获的染色质位点中包含了基因位点以及各种调控元件,分析这些相互作用能从中获得多层面的基因组组织信息<sup>[41,42]</sup>。主要的两个层面就是染色质组织结构和基因转录调控。其中基因组转录调控是通过调控元件与靶基因间的空间作用来实现的,并且多种调控元件与靶基因的综合作用产生的紧密的组织结构也可间接调控基因的转录表达。

在经典 Hi-C 分析中,染色质被描述为 A(活跃)和 B(抑制)区域,A/B 区域具有高度可塑性并且受细

胞类型限制<sup>[43]</sup>,ES 细胞分化过程中 A/B 区域会发生广泛的相互转换,影响基因组区域对核心多能性转录因子(如 Oct4、Sox2 和 Klf4)的开放程度。每个区域可进一步划分为拓扑关联结构域(Topology associate domain, TAD),它能够限定基因与其调控元件相互作用的空间,是具有独立生物学功能的基本单位。TAD 及其内部的亚结构能够通过控制增强子-启动子相互作用调控微环境,进而建立适当的基因转录与表达模式。TAD 边界上富集了许多调控因子(如 CTCF),能够调控基因的转录与表达。

目前利用 Hi-C 分析基因组转录调控主要关注点在于调控元件与基因的启动子、增强子间的相互作用。启动子在染色质三维结构上主要参与 TAD 之间的接触,大范围调控染色质构型。Hi-C 数据中存在大量相互作用的启动子,通过构建启动子-启动子互作网络发现,ES 细胞中 Oct4、Sox2 和 Nanog 等基因的启动子彼此具有强连接关系<sup>[44]</sup>。在相应细胞类型中,与关键转录调控因子结合的启动子之间相关联的频率高于预期,表明特定因子的占据有利于特定基因群之间的联系。总的来说,那些参与相关途径、由共同转录因子调控的、具有相似转录输出的基因,在空间上彼此优先接触,有助于细胞类型特异性基因表达程序的协调。也就是说,ES 细胞中与多能性因子结合的基因也优先以组织特异性方式彼此连接,共同协调基因的转录与表达。此外,与 ES 细胞特异性转录因子结合的非启动子区优先与活性基因接触,从而加强这些元件的潜在调控作用。

相比于启动子,增强子多参与细胞类型特异<sup>[45]</sup>的染色质成环,调控局部相互作用。这些环通常将启动子与增强子连接起来,组织成增强子-启动子相互作用网络,共同协调基因集转录的失活或激活。增强子-启动子网络中富集有特定的生物过程和顺式调控元件<sup>[46]</sup>,具有细胞系类型的特异性<sup>[47,48]</sup>,对细胞分化有重要影响。但是 CTCF 能够阻断增强子与启动子之间的通信<sup>[49]</sup>,将增强子活性限制在一定功能区域内,从而抑制基因转录。通过对增强子-启动子相互作用的标定,发现增强子的作用范围在分化过程中会发生明显变化,这种现象即使对于多数组织中普遍表达的基因的增强子也是存在的。

总之,染色质三维结构将基因和调控序列封装



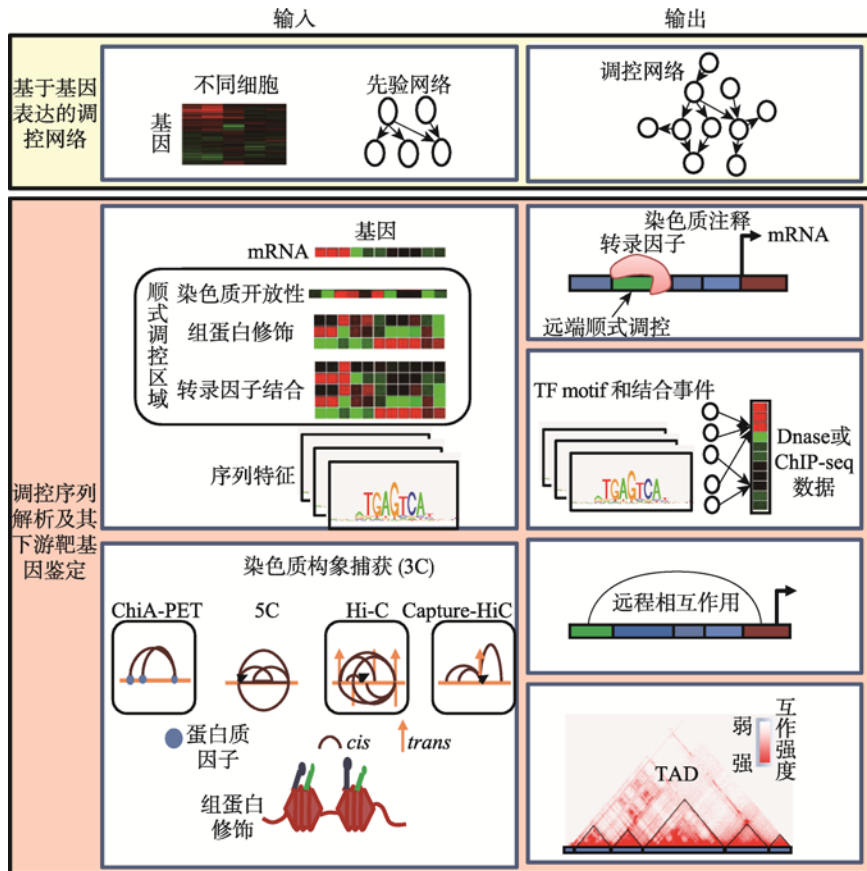


图 1 基于下一代测序方法数据推断细胞特异性调控网络的输入和输出

Fig. 1 Inputs and outputs for inference of cell type specific regulatory networks based on next-generation sequencing technologies

在隔绝的基因组邻域中, 控制基因和调控序列之间的接触。染色质构象变化可以使启动子、增强子等调控元件与靶基因相互靠近或形成空间位阻结构阻碍调控元件作用于靶基因, 从而促进或抑制细胞中分化基因及其调控因子的特异性结合, 以调控基因表达。在干细胞分化过程中, 染色质会折叠形成不同的构象, 通过调控多能性调控因子与目标基因的结合, 以适应基因表达的要求。

## 5 其他研究方法

最新的基因组编辑工具 CRISPR/Cas9 (clustered regularly interspaced short palindromic repeats/CRISPR-associated protein 9) 利用一个短的 RNA (guide RNA, gRNA) 序列引导细菌来源的 Cas9 蛋白和基因组互补。通过简单地改变 gRNA 的 5' 末端, Cas9 蛋白可以靶定基因组中几乎任何位点作用, 大大简化

了瞄准新的基因组位点的程序<sup>[50]</sup>。激活的调控元件显示低水平的 DNA 甲基化和高水平的组蛋白乙酰化, 因此, 采用表观遗传调控开关抑制增强子, 可以增加 DNA 甲基化水平或降低调控元件的组蛋白的乙酰化水平为目标。Jaenisch 等<sup>[51]</sup>通过融合 DNA 羟甲基化转移酶 TET1 或 DNA 甲基转移酶 DNMT3A 与具有催化活性的 Cas9 (dcas9), 实现表观基因组编辑, 从而改变了所调控元件附近基因表达水平。这些进展表明 CRISPR 技术结合诱导多能干细胞技术将对理解基因组转录调控机制提供帮助。

G-四联体 (G-quadruplex) 是一种特殊的核苷酸二级结构, 由四条链上的鸟嘌呤组成一个氢键连接的方形环状结构即 G-四分体 (G-tetrads) 通过堆叠形成 G-四联体。研究发现 G-四联体倾向于在基因的启动子区形成, 因此可以推测基因调控区域的 G-四联体可能具有对转录过程的调控功能。MYC 能极大提

高诱导重编程的效率, 研究发现 MYC 的调控与 G-四联体密切相关<sup>[52]</sup>, 未来需要关注 DNA 结构在转录调控中的作用。

## 6 结语和展望

细胞是生物体结构与生命活动的基本单位, 分子生物学在近半个世纪取得了极为显著的成就, 但是它并没有完全揭示出生命的奥秘, 部分原因在于它研究的生命复杂性层次仅仅处于“线性”思维和还原论水平上, 这对于许多生命现象的复杂性是难以解释的。随着高通量测序技术的发展, 干细胞生物学快速进入大数据时代, 多种层次数据大量涌现, 这对于深入理解生命起源, 进化和发育提供了契机。

本文主要对下一代测序技术在干细胞转录调控研究中的应用进行综述。基于 RNA-seq 数据构建基因表达网络来推断调控元件和目标基因之间的关系, 从而确定一个静态的调控网络或随着时间的推移动态变化网络; 基于 DNase-seq 和 Chip-seq 数据检测不同的转录因子结合位点的基因组片段; 基于染色体构象捕获技术检测两个基因组区域的三维接近的长程基因调控, 通过长程相互作用识别调控元件及其目标预测, 从而分析发现 TAD 拓扑单元(图 1)。总之, 下一代测序技术及其衍生的转录组分析、染色质免疫共沉淀分析、染色质高敏感位点分析、染色质构象分析等相互融合, 并结合生物信息分析技术, 可以从不同侧面研究干细胞的基因转录调控。这些技术对于人和小鼠以外其它物种以及人类胎盘发育的转录调控<sup>[53,54]</sup>相关研究也具有参考意义。同时, 结合单细胞技术<sup>[55,56]</sup>、转基因实验<sup>[57]</sup>、高通量报告载体<sup>[58]</sup>、基因敲除实验<sup>[59]</sup>以及 CRISPR 等实验方法, 可以对鉴定出的转录调控关系进行验证。

RNA-seq 鉴定了干细胞中编码基因的的表达谱, 然而基因组非编码序列对于生物表型同样也具有至关重要的作用。ChIP-seq 研究蛋白质与 DNA 的相互作用, 能够得到高分辨率的单个转录因子结合数据和组蛋白修饰等表观遗传学数据。DNase-seq 不需要特异性抗体就可以鉴定出全基因组潜在的调控元件, 然而, 由于人类基因组只有大约一半的转录因子被较准确鉴定<sup>[60]</sup>, 并且目前的算法很难区别出与靶定序列十分接近的转录因子<sup>[61]</sup>, 因此这项技

术的开发需要实验生物学家和计算生物学家的共同努力。染色质的三维空间结构处于变化之中, 是一个动态平衡的过程, Hi-C 技术能揭示转录调控在三维空间发生情况, 然而 Hi-C 数据模式依赖于分辨率, 分辨率太高, 模式可能看不出来或被掩盖在噪音中; 分辨率太低, 模式可能无法表现, 结合 Hi-C 和 CHIP 发展起来的 HiChIP<sup>[62]</sup>, 能够获得特异蛋白的染色质三维调控情况, 这项技术避免了 Hi-C 技术冗余信息过多的缺陷, 提高信噪比。

综上所述, 本文综述了下一代测序技术对于细胞的转录调控进行多方位解析的研究进展, 然而, 这些数据能提供的是系统的部分属性, 细胞作为一个系统, 受到强烈的“反馈”作用, 因此, 未来基于多个数据源的相互提升和加强, 为挖掘新的调控机制提供可能; 另一方面, 数据清理和数据集成的困难, 以及对数据源之间复杂相互作用的挖掘也是数据挖掘中存在的挑战。

已有包括核移植、细胞融合和转录因子介导在内的方法可以实现将分化细胞转变为多能干细胞, 这表明终末分化的成体细胞仍然具有重编程获得多能性的潜力。胚胎干细胞转录调控网络的研究为了解诱导干细胞的产生机制提供了线索, 然而对于克隆胚胎中的体细胞如何实现重编程还知之甚少, 未来可以综合使用多种组学技术, 为揭示这一生命发生的起始机制提供更多的线索。

## 致谢

感谢张星星对本文关于 G-四联体部分提供的参考; 感谢李馥雨、杨金晶和明文龙对文章提出的意见。

## 参考文献(References):

- [1] Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, Marshall VS, Jones JM. Embryonic stem cell lines derived from human blastocysts. *Science*, 1998, 282(5391): 1145–1147. [DOI]
- [2] Farah MH, Olson JM, Sucic HB, Hume RI, Tapscott SJ, Turner DL. Generation of neurons by transient expression of neural bHLH proteins in mammalian cells. *Development*, 2000, 127(4): 693–702. [DOI]
- [3] Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cul-

- tures by defined factors. *Cell*, 2006, 126(4): 663–676. [DOI]
- [4] Yamanaka S. Strategies and new developments in the generation of patient-specific pluripotent stem cells. *Cell Stem Cell*, 2007, 1(1): 39–49. [DOI]
- [5] Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 1961, 3(3): 318–356. [DOI]
- [6] Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, 2009, 10(4): 252–263. [DOI]
- [7] Shen S, Qu YC, Zhang J. The application of next generation sequencing on epigenetic study. *Hereditas (Beijing)*, 2014, 36(3): 256–275.  
沈圣, 屈彦纯, 张军. 下一代测序技术在表观遗传学研究中的重要应用及进展. *遗传*, 2014, 36(3): 256–275. [DOI]
- [8] Ding N, Qu HZ, Fang XD. The ENCODE project and functional genomics studies. *Hereditas (Beijing)*, 2014, 36(3): 237–247.  
丁楠, 渠鸿竹, 方向东. ENCODE 计划和功能基因组研究. *遗传*, 2014, 36(3): 237–247. [DOI]
- [9] Ivanova N, Dobrin R, Lu R, Kotenko I, Levorse J, DeCoste C, Schafer X, Lun Y, Lemischka IR. Dissecting self-renewal in stem cells with RNA interference. *Nature*, 2006, 442(7102): 533–538. [DOI]
- [10] Matoba R, Niwa H, Masui S, Ohtsuka S, Carter MG, Sharov AA, Ko MSH. Dissecting Oct3/4-regulated gene networks in embryonic stem cells by expression profiling. *PLoS One*, 2006, 1(1): e26. [DOI]
- [11] Liu YJ, Ma YY, Yang JY, Cheng D, Liu XP, Ma XL, West FD, Wang HY. Comparative gene expression signature of pig, human and mouse induced pluripotent stem cell lines reveals insight into pig pluripotency gene networks. *Stem Cell Rev Rep*, 2014, 10(2): 162–176. [DOI]
- [12] Zhang W, Zhong L, Wang J, Han JY. Distinct MicroRNA expression signatures of porcine induced pluripotent stem cells under mouse and human ESC culture conditions. *PLoS One*, 2016, 11(7): e0158655. [DOI]
- [13] Ezashi T, Telugu BPVL, Alexenko AP, Sachdev S, Sinha S, Roberts RM. Derivation of induced pluripotent stem cells from pig somatic cells. *Proc Natl Acad Sci USA*, 2009, 106(27): 10993–10998. [DOI]
- [14] Arnone MI, Davidson EH. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 1997, 124(10): 1851–1864. [DOI]
- [15] Ma X, Sun X. Sequence-based predictor of ATP-binding residues using random forest and mRMR-IFS feature selection. *J Theor Biol*, 2014, 360: 59–66. [DOI]
- [16] Ma X, Guo J, Liu HD, Xie JM, Sun X. Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information. *IEEE/ACM Trans Comput Biol Bioinform*, 2012, 9(6): 1766–1775. [DOI]
- [17] Maerkl SJ, Quake SR. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 2007, 315(5809): 233–237. [DOI]
- [18] Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen XY, Kuznetsov H, Wang CF, Coburn D, Newburger DE, Morris Q, Hughes TR, Bulyk ML. Diversity and complexity in DNA recognition by transcription factors. *Science*, 2009, 324(5935): 1720–1723. [DOI]
- [19] Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 2004, 431(7004): 99–104. [DOI]
- [20] Gao F, Foat BC, Bussemaker HJ. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinform*, 2004, 5: 31. [DOI]
- [21] Pesce M, Schöler HR. *Oct-4*: gatekeeper in the beginnings of mammalian development. *Stem Cells*, 2001, 19(4): 271–278. [DOI]
- [22] Chew JL, Loh YH, Zhang WS, Chen X, Tam WL, Yeap LS, Li P, Ang YS, Lim B, Robson P, Ng HH. Reciprocal transcriptional regulation of *Pou5f1* and *Sox2* via the Oct4/Sox2 complex in embryonic stem cells. *Mol Cell Biol*, 2005, 25(14): 6031–6046. [DOI]
- [23] Kuroda T, Tada M, Kubota H, Kimura H, Hatano SY, Sumerai H, Nakatsuji N, Tada T. Octamer and Sox elements are required for transcriptional *cis* regulation of *Nanog* gene expression. *Mol Cell Biol*, 2005, 25(6): 2475–2485. [DOI]
- [24] Rodda DJ, Chew JL, Lim LH, Loh YH, Wang B, Ng HH, Robson P. Transcriptional regulation of *Nanog* by Oct4 and Sox2. *J Biol Chem*, 2005, 280(26): 24731–24737. [DOI]
- [25] Chambers I, Silva J, Colby D, Nichols J, Nijmeijer B, Robertson M, Vrana J, Jones K, Grotewold L, Smith A. Nanog safeguards pluripotency and mediates germline development. *Nature*, 2007, 450(7173): 1230–1234. [DOI]

- [26] Walker E, Chang WY, Hunkapiller J, Cagney G, Garcha K, Torchia J, Krogan NJ, Reiter JF, Stanford WL. Polycomb-like 2 associates with PRC2 and regulates transcriptional networks during mouse embryonic stem cell self-renewal and differentiation. *Cell Stem Cell*, 2010, 6(2): 153–166. [DOI]
- [27] Kim J, Chu JL, Shen XH, Wang JL, Orkin SH. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, 2008, 133(7): 1290. [DOI]
- [28] Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang WW, Jiang JM, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan YJ, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 2008, 133(6): 1106–1117. [DOI]
- [29] Chronis C, Fiziev P, Papp B, Butz S, Bonora G, Sabri S, Ernst J, Plath K. Cooperative binding of transcription factors orchestrates reprogramming. *Cell*, 2017, 168(3): 442–459.e20. [DOI]
- [30] Jiang JM, Chan YS, Loh YH, Cai J, Tong GQ, Lim CA, Robson P, Zhong S, Ng HH. A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat Cell Biol*, 2008, 10(3): 353–360. [DOI]
- [31] Huang H, Chen J, Liu HD, Sun X. The nucleosome regulates the usage of polyadenylation sites in the human genome. *BMC Genomics*, 2013, 14: 912. [DOI]
- [32] Huang H, Liu HD, Sun X. Nucleosome distribution near the 3' ends of genes in the human genome. *Biosci Biotechnol Biochem*, 2013, 77(10): 2051–2055. [DOI]
- [33] Nie YM, Liu HD, Sun X. The patterns of histone modifications in the vicinity of transcription factor binding sites in human lymphoblastoid cell lines. *PLoS One*, 2013, 8(3): e60002. [DOI]
- [34] Vierstra J, Stamatoyannopoulos JA. Genomic footprinting. *Nat Methods*, 2016, 13(3): 213–221. [DOI]
- [35] Vierstra J, Wang H, John S, Sandstrom R, Stamatoyannopoulos JA. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nat Methods*, 2014, 11(1): 66–72. [DOI]
- [36] Stergachis AB, Neph S, Reynolds A, Humbert R, Miller B, Paige SL, Vernot B, Cheng JB, Thurman RE, Sandstrom R, Haugen E, Heimfeld S, Murry CE, Akey JM, Stamatoyannopoulos JA. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell*, 2013, 154(4): 888–903. [DOI]
- [37] Del Sol A, Thiesen HJ, Imitola J, Salas REC. Big-data-driven stem cell science and tissue engineering: vision and unique opportunities. *Cell Stem Cell*, 2017, 20(2): 157–160. [DOI]
- [38] Fletez-Brant C, Lee D, McCallion AS, Beer MA. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Res*, 2013, 41(W1): W544–W556. [DOI]
- [39] Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol*, 2015, 33(8): 831–838. [DOI]
- [40] Maher B. ENCODE: the human encyclopaedia. *Nature*, 2012, 489(7414): 46–48. [DOI]
- [41] Hu WQ, Hou Y, Zhang F, Liu HD, Sun X. A chromatin conformation analysis technology Hi-C and extracting of chromatin conformation information. *Genomics Appl Biol*, 2015, 34(11): 2319–2327.  
胡文桥, 侯越, 张峰, 刘宏德, 孙啸. 染色质构象解析技术——Hi-C 及染色质构象信息提取. 基因组学与应用生物学, 2015, 34(11): 2319–2327. [DOI]
- [42] Gorkin DU, Leung D, Ren B. The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell*, 2014, 14(6): 762–775. [DOI]
- [43] Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, Diao YR, Liang J, Zhao HM, Lobanenko VV, Ecker JR, Thomson JA, Ren B. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 2015, 518(7539): 331–336. [DOI]
- [44] Schoenfelder S, Furlan-Magaril M, Mifsud B, Tavares-Cadete F, Sugar R, Javierre BM, Nagano T, Katsman Y, Sakthidevi M, Wingett SW, Dimitrova E, Dimond A, Edelman LB, Elderkin S, Tabbada K, Darbo E, Andrews S, Herman B, Higgs A, LeProust E, Osborne CS, Mitchell JA, Luscombe NM, Fraser P. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res*, 2015, 25(4): 582–597. [DOI]
- [45] Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature*, 2012, 489(7414): 109–113. [DOI]
- [46] Roy S, Siahpirani AF, Chasman D, Knaack S, Ay F, Stewart R, Wilson M, Sridharan R. A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Res*, 2015, 43(18): 8694–8712. [DOI]
- [47] He WH, Ren BJ, Mao FF, Jing ZY, Li YF, Liu Y, Peng B, Yan H, Qi YH, Sun YY, Guo JT, Sui JH, Wang FC, Li WH.



- Hepatitis D virus infection of mice expressing human sodium taurocholate co-transporting polypeptide. *PLoS Pathog*, 2015, 11(4): e1004840. [DOI]
- [48] Heidari N, Phanstiel DH, He C, Grubert F, Jahanbani F, Kasowski M, Zhang MQ, Snyder MP. Genome-wide map of regulatory interactions in the human genome. *Genome Res*, 2014, 24(12): 1905–1917. [DOI]
- [49] Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 2014, 159(7): 1665–1680. [DOI]
- [50] Li S, Yang YY, Qiu Y, Chen YH, Xu LW, Ding QR. Applications of genome editing tools in precision medicine research. *Hereditas (Beijing)*, 2017, 39(3): 177–188. 李爽, 杨圆圆, 邱艳, 陈彦好, 徐璐薇, 丁秋蓉. 基因组编辑技术在精准医学中的应用. *遗传*, 2017, 39(3): 177–188. [DOI]
- [51] Liu XS, Wu H, Ji X, Stelzer Y, Wu XB, Czauderna S, Shu J, Dadon D, Young RA, Jaenisch R. Editing DNA methylation in the mammalian genome. *Cell*, 2016, 167(1): 233–247.e217. [DOI]
- [52] David AP, Margarit E, Domizi P, Banchio C, Armas P, Calcaterra NB. G-quadruplexes as novel cis-elements controlling transcription during embryonic development. *Nucleic Acids Res*, 2016, 44(9): 4163–4173. [DOI]
- [53] Liu FL, Zhou J, Zhang W, Wang H. Epigenetic regulation and related diseases during placental development. *Hereditas (Beijing)*, 2017, 39(4): 263–275. 刘福林, 周瑾, 张蔚, 汪晖. 胎盘发育过程中的表观遗传学改变及其相关疾病. *遗传*, 2017, 39(4): 263–275. [DOI]
- [54] Liu YJ, Ding DW, Liu HD, Sun X. The accessible chromatin landscape during conversion of human embryonic stem cells to trophoblast by BMP4. *Biol Reprod*, 2017, doi: 10.1093/biolre/iox028. [DOI]
- [55] Wen L, Tang FC. Recent progress in single-cell RNA-Seq analysis. *Hereditas (Beijing)*, 2014, 36(11): 1069–1076. 文路, 汤富酬. 单细胞转录组高通量测序分析新进展. *遗传*, 2014, 36(11): 1069–1076. [DOI]
- [56] Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 2015, 348(6237): 910–914. [DOI]
- [57] Visel A, Blow MJ, Li ZR, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 2009, 457(7231): 854–858. [DOI]
- [58] Melnikov A, Murugan A, Zhang XL, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG Jr, Kinney JB, Kellis M, Lander ES, Mikkelsen TS. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol*, 2012, 30(3): 271–277. [DOI]
- [59] Knott JG, Paul S. Transcriptional regulators of the trophoblast lineage in mammals with hemochorial placentation. *Reproduction*, 2014, 148(6): R121–R136. [DOI]
- [60] Weirauch MT, Hughes TR. A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. In: Hughes TR, ed. *A Handbook of Transcription Factors*. Netherlands: Springer, 2011: 25–73. [DOI]
- [61] Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M, Lewsey MG, Huang EY, Mukherjee T, Chen XT, Reece-Hoyes JS, Govindarajan S, Shaulsky G, Walhout AJ, Bouget FY, Ratsch G, Larrondo LF, Ecker JR, Hughes TR. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 2014, 158(6): 1431–1443. [DOI]
- [62] Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, Chang HY. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods*, 2016, 13(11): 919–922. [DOI]

(责任编辑: 张飞雄)