文章编号:1000-1638(2025)05-0518-14 DOI:10.13484/j.nmgdxxbzk.20250507

小鼠体细胞重编程过程关键转录因子在lncRNA区域的结合模式研究*

陈 星,左永春
(内蒙古大学生命科学学院,呼和浩特 010021)

摘要:诱导多能性干细胞(iPSC)在再生医学、疾病模型构建和药物筛选中具有广阔前景。Oct4、Sox2、Klf4和cMyc(简称OSKM)等转录因子在体细胞重编程中发挥先锋调控作用,驱动细胞命运重设。近年来,越来越多的研究发现lncRNA在重编程过程中亦发挥重要功能,然而关键转录因子在lncRNA区域的结合模式及其调控特征仍缺乏系统研究。首先基于公开ChIP-seq数据,构建了重编程关键阶段6种转录因子(Oct4、Sox2、Klf4、cMyc、Nanog和Esrrb)在小鼠基因组中的结合图谱,发现这些多能性相关因子主要富集于染色体9、10、13和17,且倾向于结合启动子区域含有ERVK元件的lincRNA。染色质可及性数据表明,OSKM在重编程过程中呈时间依赖性地分批结合至基因组靶点。基序和组蛋白修饰分析提示,OSKM可能通过与表观遗传因子协同调控lncRNA的转录活性。进一步分析发现,OSKM更易结合正义重叠、反义lncRNA,以及具ERVK启动子、长转录本、长第一内含子并邻近多能性基因的lncRNA。共表达网络分析鉴定出210个iPSC多能性网络相关lncRNA,筛选出多个可能参与iPSC多能性维持的候选lncRNA。研究结果可能对深入阐明体细胞重编程过程关键转录因子调控lncRNA的分子机制具有重要意义,为进一步筛选有效的非编码RNA重编程分子标记提供一定理论帮助。

关键词:关键转录因子;体细胞重编程;染色体结合偏好;表观遗传修饰;lncRNA类型;共表达网络

中图分类号:Q61;Q753 **文献标志码:**A

诱导多能性干细胞(Induced pluripotent stem cell,iPSC)因其具有类似胚胎干细胞的自我更新能力和多向分化潜能,在再生医学、体外疾病模拟和药物筛选等领域都有很好的应用前景[1-2]。Oct4(Octamer-binding transcription factor 4)、Sox2(SRY-box transcription factor 2)、Klf4(Kruppel-like factor 4)和cMyc(Cellular myelocytomatosis oncogene)(简称OSKM)等体细胞重编程过程关键转录因子(Transcription factor,TF),在体细胞身份转变中发挥着十分关键的先锋调控作用[3]。前期关于OSKM靶向调控体细胞基因组重编程的研究主要集中在蛋白编码(Protein-coding,PC)基因的启动子区域[4-5],筛选出了一批在维持或诱导多能性状态中具有功能作用的蛋白编码基因,部分甚至可以在特定条件下替代原始的OSKM组合,例如用Nanog(Nanog homeobox)或Esrrb(Estrogen related receptor beta)替代cMyc[6],或用Glis1(GLIS family zinc finger 1)等因子增强重编程效率[7]。

长链非编码RNA(Long non-coding RNA,lncRNA)指长度大于200 bp的无蛋白编码功能的转录

* **收稿日期:**2025-06-12;**修回日期:**2025-07-01
基金项目:国家自然科学基金项目(62171241)
作者简介:陈 星(1993—),女,河北石家庄人,博士。E-mail:chenxing.imu@outlook.com
通信作者:左永春(1982—),男,河北邢台人,教授,博士。主要从事生物信息学和细胞重编程中多组学的整合分析等方面研究。E-mail:yczuo@imu.edu.cn

本[8]。尽管 lncRNA 与蛋白编码 mRNA 在转录和加工机制上有诸多相似之处，但 lncRNA 通常缺乏有效的开放阅读框，因而不参与蛋白质的翻译过程。多项研究表明，lncRNA 的表达比蛋白编码基因具有更高的细胞类型特异性[9-10]，这使 lncRNA 成为研究特定生物学过程中关键调控事件的重要切入点。近期，越来越多的证据表明 lncRNA 在体细胞重编程过程中具有同样重要的调控作用[11-14]。然而，与蛋白编码基因相比，lncRNA 的功能研究仍相对滞后，尤其是在重编程过程中，关键转录因子在 lncRNA 基因组区域的靶向结合动态、结合位点（Binding site，BS）的选择偏好，以及结合是否与染色质状态和调控元件相关等方面，目前尚缺乏系统性的研究。这一空白限制了我们对非编码 RNA 在重编程调控网络中作用机制的深入理解。因此，亟需整合多组学数据，从转录因子结合图谱、染色质开放状态、调控元件特征等角度系统解析 OSKM 在 lncRNA 区域的调控模式。

本文围绕体细胞重编程过程中关键转录因子在 lncRNA 区域的结合特征与调控模式开展系统研究。作者基于公开的数据集，绘制了 6 种多能性相关转录因子的基因组结合图谱，并整合染色质可及性、组蛋白修饰及所结合基序的信息，揭示 OSKM 在重编程过程中呈时间依赖性地协同结合于特定 lncRNA 启动子区域，尤其偏好含内源性逆转录病毒 K（Endogenous retrovirus K，ERVK）元件、较长转录本、与多能性基因邻近的长链基因间非编码 RNA（Long intergenic non-coding RNA，lincRNA）。进一步构建 PC-lncRNA 共表达网络，筛选出可能参与 iPSC 多能性维持的候选 lncRNA。

1　材料与方法

1.1　数据来源

本研究使用 Chronis 等[15]上传至 NCBI 的小鼠体细胞重编程数据集（GSE90895）。包含小鼠胚胎成纤维细胞（MEF）、重编程 48 h、pre-iPSC 和 iPSC 四个阶段的 RNA-seq 和 ATAC-seq 数据，以及转录因子 Oct4、Sox2、Klf4、cMyc、Nanog、Esrrb 与组蛋白修饰（H3K4me3、H3K27me3、H3K79me2）及组蛋白变体 H3.3 的 ChIP-seq 数据。pre-iPSC 和 iPSC 的划分依据 Chronis 等[15]使用的 Oct4-绿色荧光蛋白（Green fluorescent protein，GFP）报告系统。pre-iPSC 指未激活内源性 Oct4、GFP 阴性的中间状态细胞；iPSC 指重编程第 10 天 GFP 阳性、具有多能性特征的细胞。

1.2　转录因子 ChIP-seq 数据分析

ChIP-seq 数据经 FastQC、Trim Galore、Trimmomatic[16]、Bowtie2[17]和 Samtools 依次进行质控、剪切、比对及格式转换，使用 MACS2（带宽 150 bp，$q<0.005$，富集倍数≥4）进行峰值探测，并借助 RIdeogram[18]可视化其基因组分布。

OSKM 结合强度以 $\log_2$（RPKM＋1）表示，采用 K-means（$k=7$）对 Oct4 结合信号聚类。通过 DeepTools2[19]展示 Oct4 结合区域±2 kb 内 SKM 信号分布，ChIPseeker[20]分析其与转录起始位点（Transcription start site，TSS）的距离，启动子定义为 TSS 上游 1.5 kb 至下游 0.5 kb。

1.3　染色质开放 ATAC-seq 数据分析

ATAC-seq 数据与 ChIP-seq 数据一致，转录因子结合基序通过 HOMER[21]识别。

1.4　基因组 lncRNA 注释

小鼠 lncRNA 注释文件（GENCODE[22] M21）包含位置、链别、ID 与类型等信息。转座元件（Transposable element，TE）注释来自 UCSC 的 rmsk 文件，提取 ERVK、ERV1、ERVL、L1 等 TE 亚类，计算 lincRNA 启动子区域内的 TE 数量作为其 TE 含量指标。

1.5　基因功能富集分析

使用 DAVID[23-24]进行基因本体论（Gene Ontology，GO）分析，以气泡图展示结果，count 表示富集基因数，P.adjust 为多重检验校正后的显著性水平，$P<0.05$ 视为差异显著。

1.6　RNA-seq 数据分析及共表达分析

RNA-seq 数据处理同 ChIP-seq 数据。用 Subread[25]计数比对到基因组的短读片段。利用 DE-

seq2[26]分析各阶段间差异表达的lncRNA和PC基因,筛选倍数变化(Fold change,FC)≥2、错误发现率(False discovery rate,FDR)<0.05的基因。对iPSC多能性PC基因与差异lncRNA进行皮尔森相关性分析(R语言cor函数),筛选相关系数$|r|\geq 0.95$、$FDR<0.01$的基因对构建共表达网络。并用Cytoscape[27]进行可视化,借助CytoHubba[28]和MCODE[29]分别识别hub基因和关键模块。

2 结果与讨论

2.1 关键多能性转录因子的靶向结合染色体分布模式偏好

首先分析转录因子在染色体上的结合偏好,初步宏观探究其在基因组空间结构中的分布,以及重编程中染色体水平的调控差异,并进一步研究转录因子在启动子、增强子等功能区域的结合特征。为此,选取了Oct4、Sox2、Klf4和cMyc四种代表性因子,同时纳入Nanog和Esrrb(作为关键调控因子与OSKM协同作用[15]),系统分析其结合分布及染色体偏好。图1显示,这些转录因子主要富集于常染色体,特别是第9、10、13和17号染色体。相比之下,性染色体结合频率明显较低。值得注意的是,从图中可以观察到,转录因子ChIP-seq信号峰密度较高的位置,往往与启动子区域含有内源性逆转录病毒家族成员ERVK的lincRNA分布存在一定程度的重合。这一观察结果支持了Kelley和Rinn关于ERVK在小鼠胚胎干细胞维持中发挥重要作用的报道[30]。

2.2 关键转录因子动态结合模式与染色质开放性关联分析

Yamanaka四因子中的Oct4是唯一不可被同家族成员替代的因子,在体细胞重编程过程中发挥核心作用[31]。以核心因子Oct4为例,分析其全基因组ChIP-seq结合信号,发现结合位点可分为7类(图2)。大部分位点(clusters 1~6)在重编程早期(48 h)即被占据,随后信号呈现增强或维持高水平结合(clusters 1、2、5)或减弱(cluster 6)趋势,少数位点(cluster 7)仅在iPSC阶段首次结合,显示Oct4按阶段性调控转录程序。值得注意的是,一旦Oct4的结合信号减少,几乎不会再有增强的趋势。其他SKM因子结合模式与其高度一致,表明四因子协同分批结合靶点。结合ATAC-seq数据显示,Oct4结合位点染色质开放程度与Oct4结合信号高度相关(图2)。

2.3 lncRNA启动子区域关键转录因子的动态结合模式

启动子区域是基因表达调控的关键位点。我们对开放染色质区域的转录因子结合基序进行了富集分析,比较了lncRNA与PC基因启动子区域的差异。表1显示,在iPSC阶段,lncRNA启动子中除常见CTCF和Sp家族外,多能性因子(如Oct4、Sox2、Klf家族、Nanog)基序显著富集,其排序基于富集的统计学显著性(P值)。重编程过程中,体细胞特异性转录因子如Fra1(Fos-related antigen 1)、Runx1(Runt-related transcription factor 1)、Tead4(TEA domain transcription factor 4))基序富集减少,而多能性因子的富集持续增强(图3),反映调控网络向多能性转变。

启动子区域,lncRNA与PC基因的TF富集模式存在差异。lncRNA启动子中Oct4和Sox2基序在48 h富集后下降;PC基因中,Oct4基序从48 h起逐渐富集,Sox2则始终较低(图3)。

依据Oct4在启动子区域的结合强度变化,将PC基因及lncRNA分为四类:1)持续结合(Persistent):48 h与iPSC的结合强度均大于等于0.66;2)结合减少(Lost):48 h的结合强度大于等于0.66,iPSC的结合强度小于0.66;3)结合增强(Gain):48 h的结合强度小于等于0.66,iPSC的结合强度大于0.66;4)无结合(None):48 h与iPSC的结合强度均小于0.66。发现SKM结合趋势与Oct4一致,且OSKM结合与H3K4me3、H3K79me2、H3.3等激活性标记呈正相关,与抑制性标记H3K27me3呈负相关(图4),提示OSKM可能通过调控表观修饰促进基因激活。

GO分析显示,PC基因中Persistent组富集于细胞周期与RNA加工等核心生物过程,表明这些基因作为管家基因,维持着细胞的基本功能;Lost组则与分化相关,主要富集在肌肉细胞生长发育、细胞迁移等生物学过程中;Gain组主要富集在阳离子跨膜运输、减数分裂等生物学过程中;None组主要富集在应激反应和药物代谢等生物学过程中(图5)。

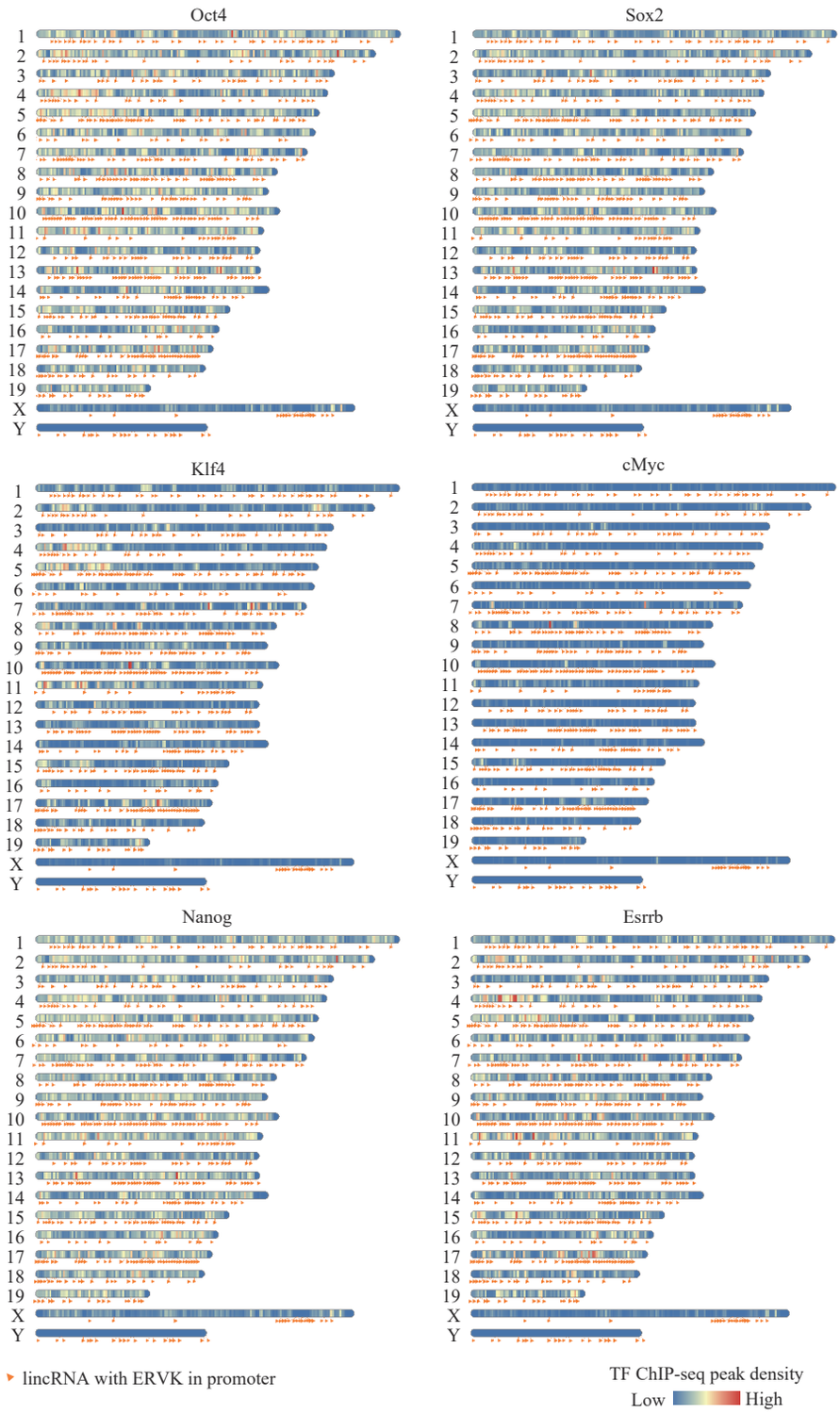


图 1　多能性转录因子在 iPSC 阶段的染色体结合分布

Fig. 1　Chromosomal distribution of pluripotency transcription factors binding at the iPSC stage

2.4　体细胞重编程过程中 OSKM 结合启动子偏好与 lncRNA 特征的关系

为解析 OSKM 在 lncRNA 启动子区域的结合偏好，从几个 lncRNA 的特征进行了系统分析。

1）lncRNA 类型　不同类型 lncRNA 启动子的 OSKM 结合频率存在显著差异（图 6）。与蛋白编码序列关系密切的 lncRNA（如 sense_overlapping、antisense、双向启动子型）被 OSKM 更频繁地结合；

而如 retained_intron 和 sense_intronic 类型,它们位于内含子区域,远离外显子,其启动子对 OSKM 的结合频率显著较低。

2) lncRNA 转录本长度　在 iPSC 阶段,较长的 lncRNA 转录本在启动子区域更倾向于显示出更多的 OSKM 结合位点,表现出一定的正相关趋势(图 7)。

3) 与多能性基因的距离　以 lincRNA(数量最多,也最具功能潜力的一类 lncRNA)为代表,分析 iPSC 阶段 lincRNA 启动子与多能性基因 Oct4、Sox2、Klf4 和 Nanog(OSKN)TSS 的距离发现:距离多能性基因越近的 lincRNA,其启动子区域被 Oct4 结合的频次越高;特别是多能性基因 TSS 10～20 kb 内的 lincRNA,其启动子结合 Oct4 水平显著高于平均水平(图 8)。这提示这些邻近的 lincRNA 可能通过顺式调控参与多能性建立与维持。

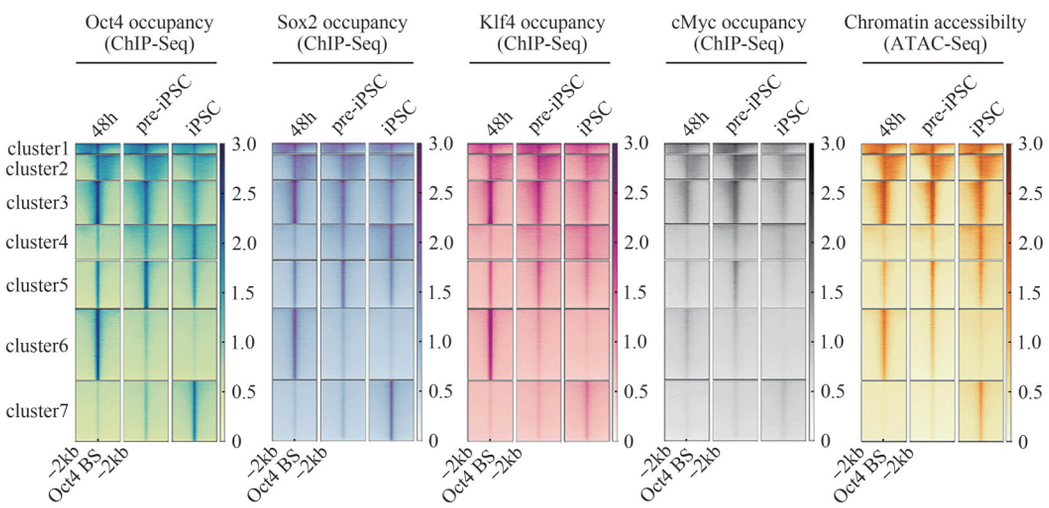


图 2　重编程过程中 OSKM 的动态结合模式与染色质开放性的变化

Fig. 2　Dynamic binding patterns of OSKM binding patterns and chromatin accessibility during reprogramming

表 1　iPSC 阶段 lncRNA 启动子 ATAC-seq 峰中富集的前 10 个转录因子结合基序

Table 1　Top 10 TF-binding motifs enriched in lncRNA promoter ATAC-seq peaks at the iPSC stage

排序	转录因子	转录因子结合基序	*P* 值
1	CTCF(CCCTC-binding factor)		1×10^{-120}
2	Ctcfl(CCCTC-binding factor like)		1×10^{-92}
3	Sp5(Sp5 transcription factor)		1×10^{-57}
4	Sp2(Sp2 transcription factor)		1×10^{-54}
5	OCT4-SOX2-TCF-NANOG		1×10^{-46}
6	KLF14		1×10^{-44}
7	KLF5		1×10^{-44}
8	Pou3f3(POU domain,class 3,transcription factor 3)		1×10^{-38}
9	KLF6		1×10^{-35}
10	Oct4		1×10^{-34}

4）启动子转座元件含量　转座元件（TE）是一类能改变自身在基因组中位置的重复序列，已被证实对于维持干细胞多能性具有重要作用[32]。lincRNA 中富含 TE，其中 ERVK 对胚胎干细胞有极强的影响[30]。据此我们假设，OSKM 在重编程过程中对 lincRNA 启动子的结合偏好可能与其 TE 含量相关。为验证该假设，将小鼠 lincRNA 分为 3 类：ERVK-lincRNA，启动子区域含 ERVK 元件；OtherTEs-lincRNA，启动子区域含除 ERVK 外的其他转座元件；dTE-lincRNA，启动子区域不含任何转座元件。如图 9（a）所示，iPSC 阶段 ERVK-lincRNA 启动子上的 OSKM 结合事件数量高于另外两组，表明 OSK 的结合偏好与 TE，尤其是 ERVK 的存在密切相关。比较发现：ERVK-lincRNA 转录本最长且拥有更长的第一内含子，见图 9（b）；启动子 TE 数量与转录本长度呈正相关，见图 9（c）。这提示较长的转录本为 TE 提供了更多嵌入空间，也表明 OSKM 结合偏好可能更多受第一内含子结构影响。

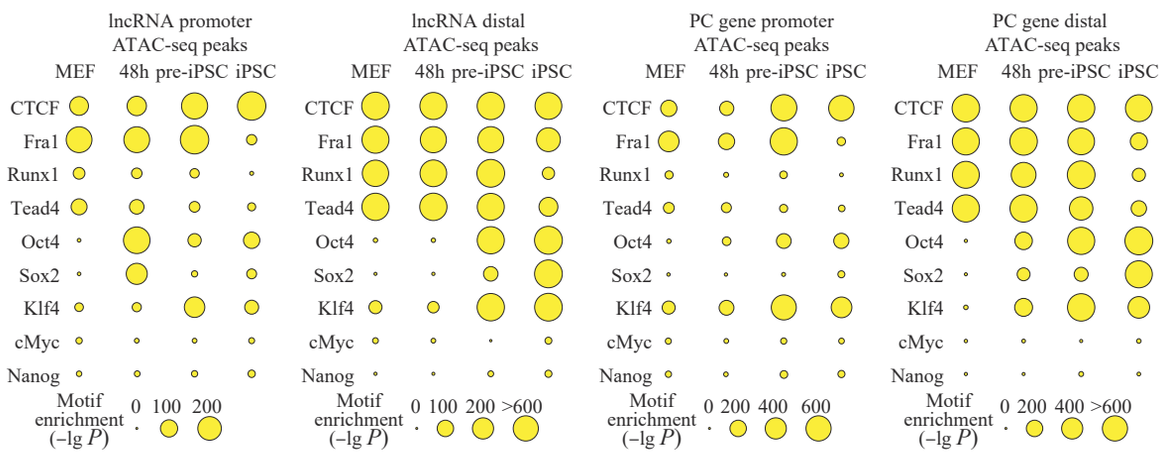


图 3　重编程过程中富集于 lncRNA 与 PC 基因启动子/远端 ATAC-seq 峰的 TF 结合基序及其 *P* 值

Fig. 3　Enriched TF motifs and *P* values at lncRNA and PC gene promoter/distal ATAC-seq peaks during reprogramming

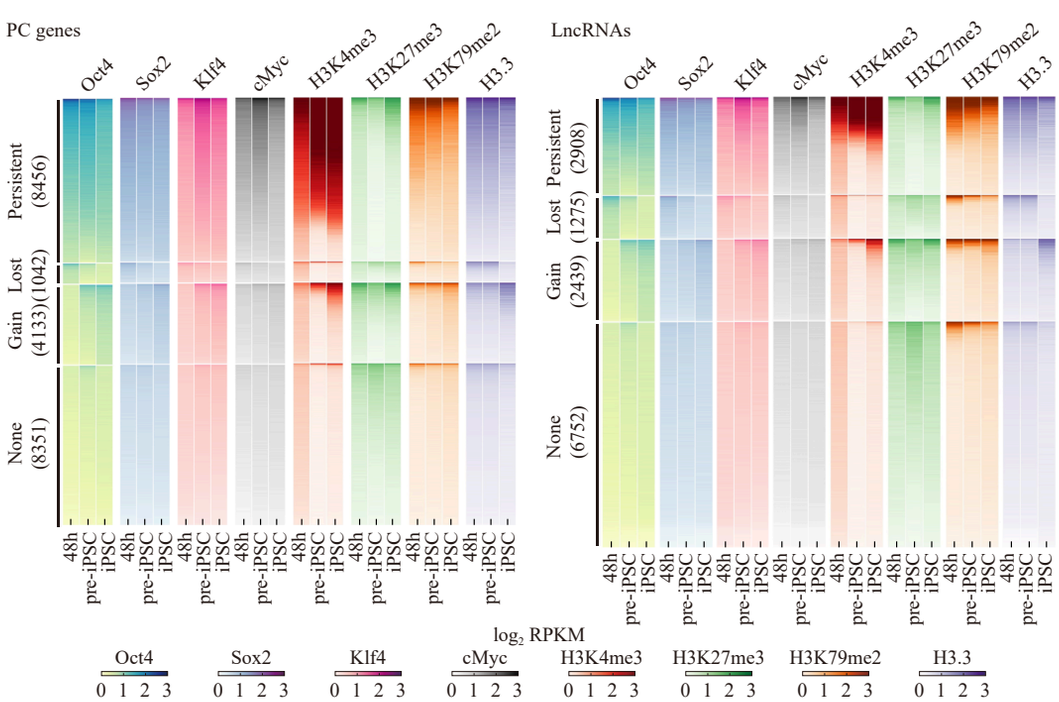


图 4　重编程过程中 PC 基因与 lncRNA 启动子的 OSKM 结合及表观修饰动态变化

Fig. 4　Dynamics of OSKM binding and epigenetic marks at PC gene and lncRNA promoters during reprogramming

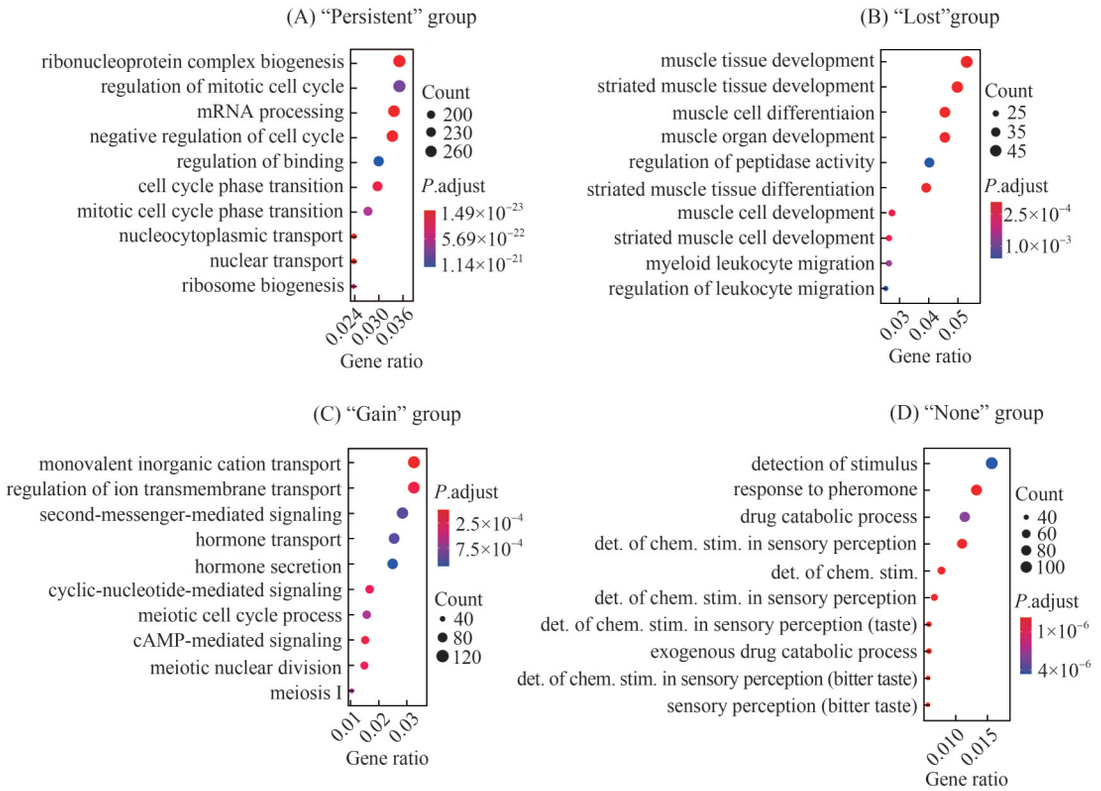


图 5 各组蛋白编码基因的 GO 富集分析

Fig. 5 GO enrichment analysis of each group of protein-coding genes

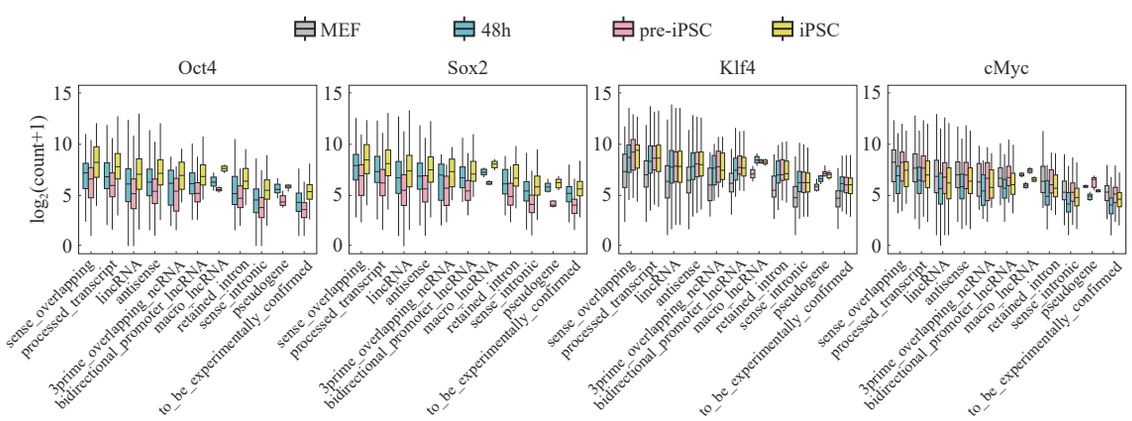


图 6 重编程过程中各类 lncRNA 启动子上 OSKM 结合数量变化

Fig. 6 Dynamics of OSKM binding at promoters of different lncRNA types during reprogramming

2.5 小鼠 iPSC 多能性相关 PC-lncRNA 基因共表达网络分析

为挖掘在 iPSC 多能性维持中发挥关键作用的 lncRNA，首先筛选出差异表达（$|\log_2 FC| \geqslant 2$，$FDR < 0.05$）的 lncRNA 和蛋白编码基因（Protein-coding gene，PCG）。结果显示在重编程 48 h、pre-iPSC 及 iPSC 阶段，上调的 lncRNA 数量普遍多于下调，见表 2 及图 10(a)。随后将 288 个已知多能性相关蛋白编码基因（Pluripotency PCGs）[33]与筛选出的 4548 个 iPSC 差异表达 PCGs（iPSC-DE-PCGs）对比，发现有 101 个基因重合，见图 10(b)。这不仅在一定程度上验证了差异表达分析结果的可靠性，更重要的是，这些基因可能在 iPSC 多能性维持中发挥关键作用，我们将其定义为 iPSC 多能性相关蛋白编码基因（iPSC pluripotency PCGs）。

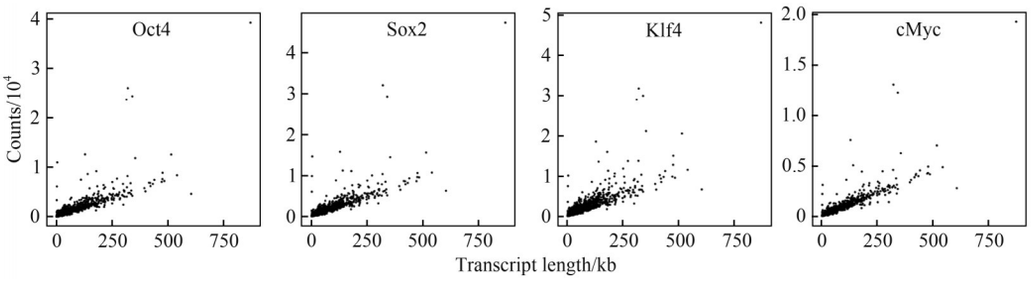


图 7　iPSC 阶段 lncRNA 启动子 OSKM 结合数量与转录本长度的关系

Fig. 7　Correlation between OSKM binding and lncRNA transcript length at the iPSC stage

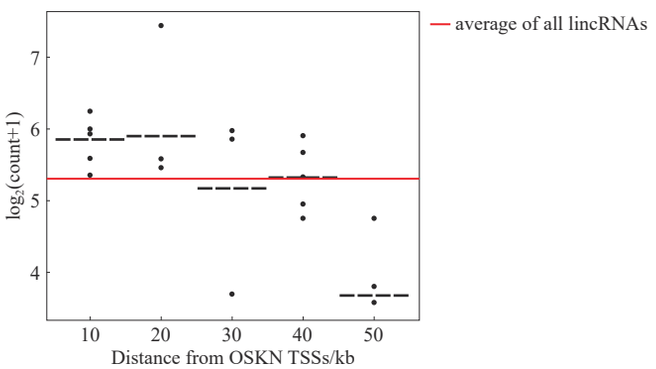


图 8　iPSC 阶段多能性基因(OSKN)TSS 不同距离范围内 lincRNA 启动子上 Oct4 结合数量

Fig. 8　Oct4 binding at lincRNA promoters located at varying distances from OSKN TSS

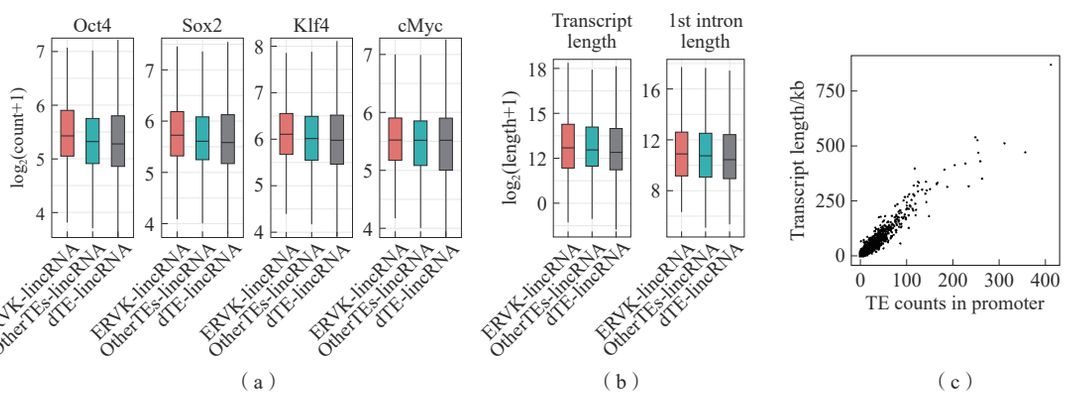


图 9　lincRNA 启动子中转座元件含量与 OSKM 结合偏好相关

Fig. 9　Association between transposable element content in lincRNA promoters and OSKM binding preference

我们对 iPSC 多能性蛋白编码基因与 iPSC 中差异表达的 lncRNA(iPSC-DE-lncRNAs)进行了共表达分析。发现 91 个 iPSC 多能性蛋白编码基因与 210 个 iPSC-DE-lncRNAs 存在显著相关($|r|\geqslant 0.95$, $FDR<0.01$)。据此构建了由蛋白编码基因与非编码基因组成的共表达网络,共形成 797 条连接(图 11)。

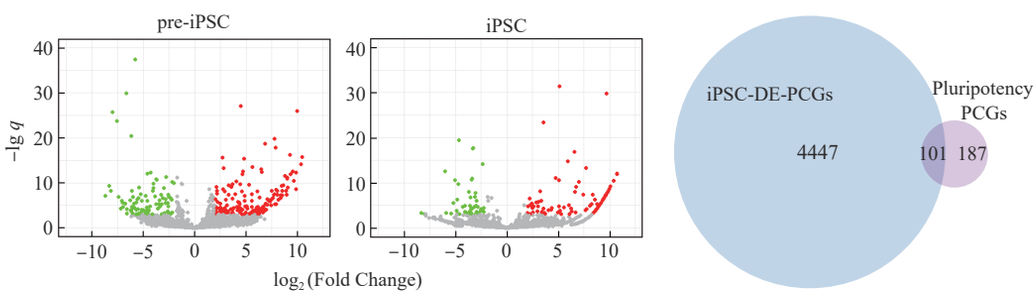
我们进一步从共表达网络中筛选出关键枢纽(hub)基因,确定得分最高的 10 个 hub 基因,包括 Eppk1(Epiplakin 1)、Tdgf1(Teratocarcinoma-derived growth factor 1)、Dppa2(Developmental pluripotency associated 2)、Fb1(Fumonisin B1)、Phc1(Polyhomeotic homolog 1)、Pbx1(Pre B cell leukemia homeobox 1)、Nanog、Platr15(Pluripotency associated transcript 15)、AC132307.5 和 Stxbp2(Syntaxin binding protein 2)(图 12)。其中多个蛋白编码基因与多能性维持及重编程密切相关。如 Tdgf1 是 Nodal 信号通路的新调节分子,维持胚胎干细胞的多能性;Dppa2,这是近年来发现的在多能性细胞系中特异性表达的基因之一,可增强体细胞重编程效率并激活合子基因组激活相关基因[34-36],并可能

在小鼠胚胎干细胞状态的维持与增殖中发挥重要作用[37];经典多能性因子 Nanog 也是我们鉴定出的 hub 基因之一,其在维持胚胎干细胞全能性中的关键功能已明确[38-40];Phc1 属于 Polycomb 抑制复合体,通过染色质重塑参与基因沉默,在调控多能状态及分化中扮演重要角色[41]。

表 2 三种细胞类型中差异表达的 lncRNA 和蛋白编码基因数量

Table 2 Number of differentially expressed lncRNAs and PCGs across three cell types

Type	48 h	pre-iPSC	iPSC
PCG_Down	47	2335	2521
PCG_Up	199	1778	2027
PCG_All	246	4113	4548
Lnc_Down	2	227	171
Lnc_Up	35	449	244
Lnc_All	37	676	415



(a) 火山图展示了pre-iPSC和iPSC阶段上调和下调的lncRNAs (b) 多能性蛋白编码基因与iPSC中差异表达蛋白编码基因的韦恩图

图 10 基因差异表达分析

Fig. 10 Gene differential expression analysis

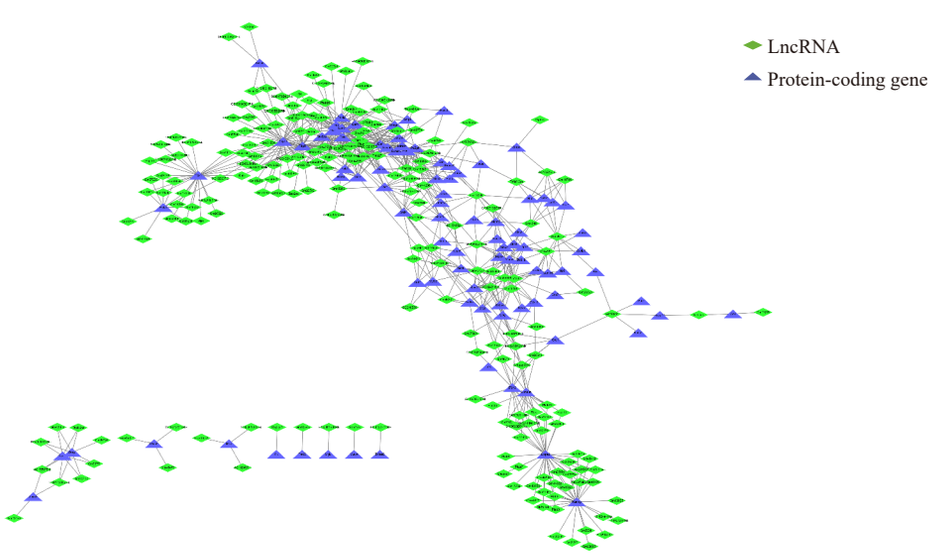


图 11 蛋白编码基因-lncRNA 的共表达网络

Fig. 11 PCG-lncRNA co-expression network

另一方面,我们发现两个高评分的 lncRNA 基因——Platr15 和 AC132307.5,也表现出典型的 hub 特征。其中,Platr15 属于 Platr 家族,该家族多个成员已被证实参与胚胎干细胞多能性的维

持[42-44]。值得注意的是，Platr家族的另一个成员 Platr27 亦在网络中排名第 27，提示该家族成员可能共同参与多能状态的调控。

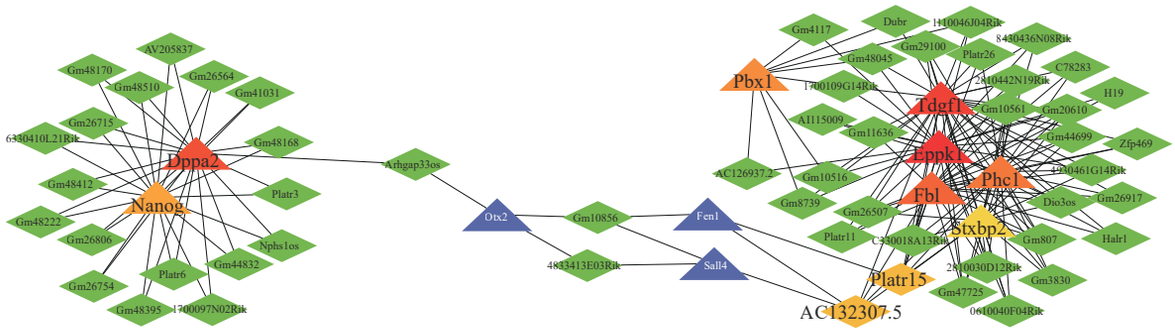


图 12　hub基因筛选

Fig. 12　Selection of hub genes

我们进一步识别出共表达网络中的关键模块，这些模块是共表达网络中基因联系最紧密的模块，其网络得分介于 2.7 至 4.0 之间（图 13）。模块中含多个多能性相关的蛋白编码基因，如 Chaf1a（Chromatin assembly factor 1 subunit A），在维持多能性状态下的胚胎植入前稳定性中发挥关键作用，敲低会导致胚胎发育阻滞[45]；Mcm6（Minichromosome maintenance complex component 6）在 DNA 复制受阻时参与复制恢复，以维持基因组稳定性，对胚胎干细胞至关重要[46-48]。

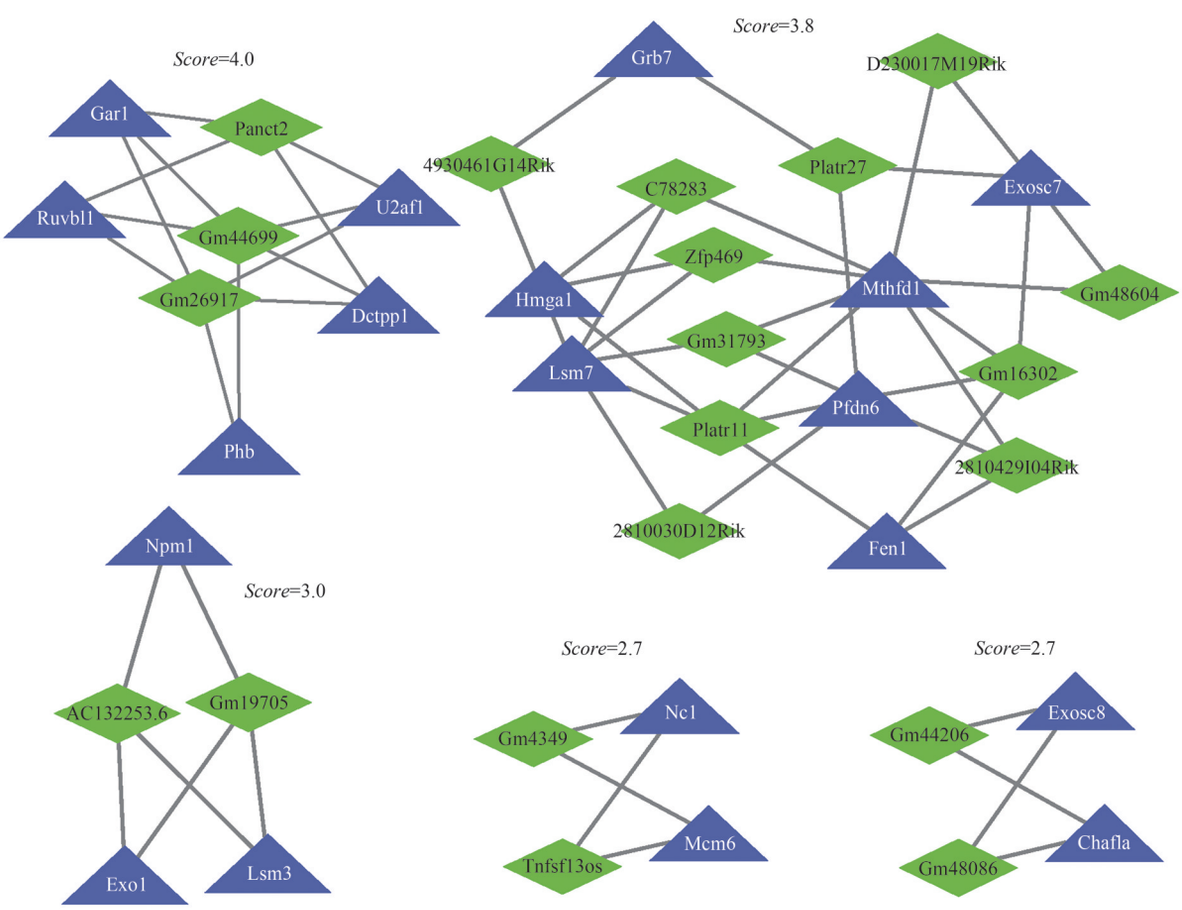


图 13　MCODE分析结果中网络得分最高的5个亚网络

Fig. 13　Five sub-networks with highest scores in results of MCODE analysis

关键模块还富含与多能性密切相关的lncRNA,如Platr家族的Platr11、Platr27以及Panct(Pluripotency-associated noncoding transcript)家族的Panct2。已有研究表明,Panct1、Panct2、Panct3的表达与小鼠胚胎干细胞的自我更新及多能性状态密切相关,其功能涉及调控基因表达、染色质结构及细胞命运决定[49-50]。具体而言,Panct2可能通过与核心多能性转录因子或表观遗传调控因子互作,参与维持干细胞的开放染色质状态和转录活性。

3 结论

通过整合关键转录因子的ChIP-seq与ATAC-seq数据,系统性地描绘了小鼠体细胞重编程过程中多个关键时间点上,Oct4、Sox2、Klf4、cMyc、Nanog和Esrrb等转录因子在全基因组范围内,特别是在lncRNA启动子区域的动态结合模式。研究发现,这些转录因子的结合呈现明显的时间依赖性,且与染色质开放状态高度相关,反映出它们在重编程进程中具有阶段性、协同性的调控特征。基于转录因子结合基序和组蛋白修饰数据的进一步分析显示,OSKM等转录因子可能通过与染色质修饰因子协同作用,共同调控lncRNA的转录活性。我们还发现,OSKM更倾向于结合具有特定特征的lncRNA启动子区域,如正义重叠和反义lncRNA,启动子含有ERVK、具有较长转录本或第一内含子,并且与多能性基因位于相邻区域。这些结果可能揭示了关键转录因子在lncRNA区域较为精细的调控图谱。更为重要的是,我们通过构建蛋白编码基因与差异表达lncRNA的共表达网络,鉴定出210个与iPSC多能性密切相关的lncRNA,初步勾勒出多能性调控中非编码RNA的网络架构。这些lncRNA可能通过多种机制——包括作为分子支架、调节染色质修饰酶的定位、影响转录因子活性等——充当多能性基因表达的关键调控节点。共表达网络中筛选出的候选lncRNA可能为未来深入功能研究提供线索。未来的功能验证研究,如基因敲除或过表达实验,将有助于揭示这些lncRNA在维持iPSC多能性中的具体机制。同时,这也可能为开发更高效的重编程策略和非编码RNA分子标记提供新的潜在靶点。综上所述,本研究可能为理解重编程过程中转录因子与非编码RNA协同调控机制提供理论参考,也为后续功能验证实验及重编程相关lncRNA分子标记的筛选奠定初步基础。

参考文献:

[1] AVIOR Y, SAGI I, BENVENISTY N. Pluripotent stem cells in disease modelling and drug discovery[J]. Nature Reviews Molecular Cell Biology, 2016, 17(3): 170-182.

[2] TROUNSON A, DEWITT N D. Pluripotent stem cells progressing to the clinic[J]. Nature Reviews Molecular Cell Biology, 2016, 17(3): 194-200.

[3] TAKAHASHI K, YAMANAKA S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors[J]. Cell, 2006, 126(4): 663-676.

[4] CHEN X, XU H, YUAN P, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells[J]. Cell, 2008, 133(6): 1106-1117.

[5] SOUFI A, GARCIA M F, JAROSZEWICZ A, et al. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming[J]. Cell, 2015, 161(3): 555-568.

[6] NAKAGAWA M, KOYANAGI M, TANABE K, et al. Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts[J]. Nature Biotechnology, 2008, 26(1): 101-106.

[7] MAEKAWA M, YAMAGUCHI K, NAKAMURA T, et al. Direct reprogramming of somatic cells is promoted by maternal transcription factor Glis1[J]. Nature, 2011, 474(7350): 225-229.

[8] RINN J L, CHANG H Y. Genome regulation by long noncoding RNAs[J]. Annual Review of Biochemistry, 2012, 81: 145-166.

[9] CABILI M N, TRAPNELL C, GOFF L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses[J]. Genes & Development, 2011, 25(18): 1915-1927.

[10] DJEBALI S, DAVIS C A, MERKEL A, et al. Landscape of transcription in human cells[J]. Nature, 2012, 489(7414): 101-108.

[11] LOEWER S, CABILI M N, GUTTMAN M, et al. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells[J]. Nature Genetics, 2010, 42(12): 1113-1117.

[12] NG S Y, JOHNSON R, STANTON L W. Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors[J]. The EMBO Journal, 2012, 31(3): 522-533.

[13] BAO X C, WU H T, ZHU X H, et al. The p53-induced *lincRNA-p21* derails somatic cell reprogramming by sustaining H3K9me3 and CpG methylation at pluripotency gene promoters[J]. Cell Research, 2015, 25(1): 80-92.

[14] LIU S J, HORLBECK M A, CHO S W, et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells[J]. Science, 2017, 355(6320): aah7111.

[15] CHRONIS C, FIZIEV P, PAPP B, et al. Cooperative binding of transcription factors orchestrates reprogramming [J]. Cell, 2017, 168(3): 442-459. e20.

[16] BOLGER A M, LOHSE M, USADEL B. Trimmomatic: A flexible trimmer for Illumina sequence data[J]. Bioinformatics, 2014, 30(15): 2114-2120.

[17] LANGMEAD B, SALZBERG S L. Fast gapped-read alignment with Bowtie 2[J]. Nature Methods, 2012, 9(4): 357-359.

[18] HAO Z D, LV D K, GE Y, et al. RIdeogram: Drawing SVG graphics to visualize and map genome-wide data on the idiograms[J]. PeerJ Computer Science, 2020, 6: e251.

[19] RAMÍREZ F, RYAN D P, GRÜNING B, et al. deepTools2: A next generation web server for deep-sequencing data analysis[J]. Nucleic Acids Research, 2016, 44(W1): W160-W165.

[20] YU G A, WANG L G, HE Q Y. ChIPseeker: An R/bioconductor package for ChIP peak annotation, comparison and visualization[J]. Bioinformatics, 2015, 31(14): 2382-2383.

[21] HEINZ S, BENNER C, SPANN N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities[J]. Molecular Cell, 2010, 38(4): 576-589.

[22] FRANKISH A, DIEKHANS M, FERREIRA A M, et al. GENCODE reference annotation for the human and mouse genomes[J]. Nucleic Acids Research, 2019, 47(D1): D766-D773.

[23] HUANG D W, SHERMAN B T, LEMPICKI R A. Systematic and integrative analysis of large gene lists using David bioinformatics resources[J]. Nature Protocols, 2009, 4(1): 44-57.

[24] HUANG D W, SHERMAN B T, LEMPICKI R A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists[J]. Nucleic Acids Research, 2009, 37(1): 1-13.

[25] LIAO Y, SMYTH G K, SHI W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads[J]. Nucleic Acids Research, 2019, 47(8): e47.

[26] LOVE M I, HUBER W, ANDERS S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2[J]. Genome Biology, 2014, 15(12): 550.

[27] SHANNON P, MARKIEL A, OZIER O, et al. Cytoscape: A software environment for integrated models of biomo-

lecular interaction networks[J]. Genome Research,2003,13(11):2498-2504.

[28] CHIN C H,CHEN S H,WU H H,et al. cytoHubba:Identifying hub objects and sub-networks from complex interactome[J]. BMC Systems Biology,2014,8(4):S11.

[29] BADER G D,HOGUE C W V. An automated method for finding molecular complexes in large protein interaction networks[J]. BMC Bioinformatics,2003,4:2.

[30] KELLEY D,RINN J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs[J]. Genome Biology,2012,13(11):R107.

[31] JERABEK S,MERINO F,SCHÖLER H R,et al. OCT4:Dynamic DNA binding pioneers stem cell pluripotency [J]. Biochimica et Biophysica Acta,2014,1839(3):138-154.

[32] ST LAURENT G,SHTOKALO D,DONG B,et al. VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer[J]. Genome Biology,2013,14(7):R73.

[33] MÜLLER F J,LAURENT L C,KOSTKA D,et al. Regulatory networks define phenotypic classes of human stem cell lines[J]. Nature,2008,455(7211):401-405.

[34] BORTVIN A,EGGAN K,SKALETSKY H,et al. Incomplete reactivation of *Oct4*-related genes in mouse embryos cloned from somatic nuclei[J]. Development,2003,130(8):1673-1680.

[35] MALDONADO-SALDIVIA J,VAN DEN BERGEN J,KROUSKOS M,et al. *Dppa2* and *Dppa4* are closely linked SAP motif genes restricted to pluripotent cells and the germ line[J]. Stem Cells,2007,25(1):19-28.

[36] ECKERSLEY-MASLIN M,ALDA-CATALINAS C,BLOTENBURG M,et al. Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program[J]. Genes & Development,2019,33(3/4):194-208.

[37] DU J,CHEN T J,ZOU X,et al. Dppa2 knockdown-induced differentiation and repressed proliferation of mouse embryonic stem cells[J]. Journal of Biochemistry,2010,147(2):265-271.

[38] CHAMBERS I,COLBY D,ROBERTSON M,et al. Functional expression cloning of Nanog,a pluripotency sustaining factor in embryonic stem cells[J]. Cell,2003,113(5):643-655.

[39] MITSUI K,TOKUZAWA Y,ITOH H,et al. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells[J]. Cell,2003,113(5):631-642.

[40] CHAMBERS I,SILVA J,COLBY D,et al. Nanog safeguards pluripotency and mediates germline development [J]. Nature,2007,450(7173):1230-1234.

[41] ALOIA L,DI STEFANO B,DI CROCE L. Polycomb complexes in stem cells and embryonic development[J]. Development,2013,140(12):2525-2534.

[42] BERGMANN J H,LI J J,ECKERSLEY-MASLIN M A,et al. Regulation of the ESC transcriptome by nuclear long noncoding RNAs[J]. Genome Research,2015,25(9):1336-1346.

[43] DU Z H,WEN X,WANG Y C,et al. Author correction:Chromatin lncRNA *Platr10* controls stem cell pluripotency by coordinating an intrachromosomal regulatory network[J]. Genome Biology,2021,22(1):272.

[44] YAN P X,LU J Y,NIU J,et al. LncRNA Platr22 promotes super-enhancer activity and stem cell pluripotency[J]. Journal of Molecular Cell Biology,2021,13(4):295-313.

[45] WANG C F,LIU X Y,GAO Y W,et al. Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development[J]. Nature Cell Biology,2018,20(5):620-631.

[46] BOCHMAN M L,SCHWACHA A. The Mcm complex:Unwinding the mechanism of a replicative helicase[J].

Microbiology and Molecular Biology Reviews:MMBR,2009,73(4):652-683.

[47] DOKSANI Y,BERMEJO R,FIORANI S,et al. Replicon dynamics,dormant origin firing,and terminal fork integrity after double-strand break formation[J]. Cell,2009,137(2):247-258.

[48] GE X Q,HAN J,CHENG E C,et al. Embryonic stem cells license a high level of dormant origins to protect the genome against replication stress[J]. Stem Cell Reports,2015,5(2):185-194.

[49] CHAKRABORTY D,KAPPEI D,THEIS M,et al. Combined RNAi and localization for functionally dissecting long noncoding RNAs[J]. Nature Methods,2012,9(4):360-362.

[50] CHAKRABORTY D,PASZKOWSKI-ROGACZ M,BERGER N,et al. lncRNA *Panct1* maintains mouse embryonic stem cell identity by regulating TOBF1 recruitment to Oct-Sox sequences in early G1[J]. Cell Reports,2017,21(11):3012-3021.

(责任编委　那顺布和)

Genome-wide Binding Patterns of Core Transcription Factors at lncRNA Loci during Mouse Somatic Cell Reprogramming

CHEN Xing,ZUO Yongchun

(*School of Life Sciences*, *Inner Mongolia University*, *Hohhot* 010021, *China*)

Abstract:Induced pluripotent stem cells (iPSCs) offer broad applications in regenerative medicine, disease modeling, and drug discovery. Transcription factors, Oct4, Sox2, Klf4, and cMyc (OSKM), are essential for somatic cell reprogramming, but how they regulate long non-coding RNAs (lncRNAs) remains unclear. In this study, we analyzed ChIP-seq data to map genome-wide binding of six pluripotency-related transcription factors (Oct4, Sox2, Klf4, cMyc, Nanog, Esrrb) during mouse reprogramming. These factors were enriched on chromosomes 9, 10, 13, and 17, showing preferential binding to lincRNA promoters containing ERVK elements. Integration with chromatin accessibility and histone modification analyses revealed temporally ordered OSKM binding, likely coordinated with epigenetic regulators. OSKM favored lncRNAs that were sense-overlapping, antisense-oriented, ERVK-associated, and characterized by longer transcripts, extended first introns, and proximity to pluripotency genes. A co-expression network identified 210 lncRNAs potentially involved in iPSC pluripotency regulation. These findings reveal how key transcription factors interact with lncRNAs during somatic reprogramming and offer a basis for identifying functional non-coding RNA markers.

Key words: key transcription factors; somatic cell reprogramming; chromatin binding preference; epigenetic modification; lncRNA types; co-expression network