

DOI: 10.3724/SP.J.1005.2013.01253

## 非编码 DNA 序列的功能及其鉴定

秦丹<sup>1,2</sup>, 徐存拴<sup>1,2</sup>

1. 河南师范大学生命科学学院, 新乡 453007;
2. 河南省-科技部共建细胞分化调控国家重点实验室培育基地, 新乡 453007

**摘要:** 非编码 DNA 序列是指基因组中不编码蛋白质的 DNA 序列。这些序列可以结合调节因子、转录为功能性 RNA、单独或协同地调节生理活动和病理过程。文章围绕基因表达调控作用, 总结了近几年非编码 DNA 序列的研究成果, 对其结构、功能和可能的作用机制进行了初步阐述, 介绍了目前鉴定非编码 DNA 序列中功能元件的计算方法和实验技术, 并对非编码 DNA 未来的研究进行了展望。

**关键词:** 非编码 DNA 序列; 基因表达调控; 功能元件; 转录因子结合位点; 非编码 RNA; 表观遗传学

## Characterization and identification of functional elements in non-coding DNA sequences

QIN Dan<sup>1,2</sup>, XU Cun-Shuan<sup>1,2</sup>

1. College of Life Science, Henan Normal University, Xinxiang 453007, China;
2. State Key Laboratory Breeding Base Co-sponsored by Henan & Ministry of Science and Technology for Cell Differentiation Regulation, Xinxiang 453007, China

**Abstract:** The non-coding DNA sequences refer to non-protein-coding DNA sequences in genome. These sequences can bind with transcription factors or be transcribed as functional RNAs, thus participating in the regulation of many physiological activities and pathological processes. Aiming at gene expression regulation, this review focuses on the recent progress of non-coding DNA and illustrates their structures, functions and potential acting mechanisms. Meanwhile, some computational and experimental methods of identifying functional elements in the non-coding DNAs are introduced. Finally, further studies in this field are proposed.

**Keywords:** non-coding DNA sequences; gene expression regulation; functional elements; transcription factor binding sites; non-coding RNAs; epigenetics

在庞大的人类基因组中, 蛋白质编码序列所占比例不到 2%, 其余约 98% 的 DNA 序列都是不编码蛋白质的<sup>[1]</sup>。对于数量上占绝对优势的非编码 DNA

序列, 即所谓的“垃圾”DNA, 其存在意义如何? DNA 元件百科全书(Encyclopedia of DNA Elements, ENCODE)研究计划对此做了初步解读, 发现基因组

收稿日期: 2013-06-06; 修回日期: 2013-07-04

基金项目: 国家重点基础研究发展计划项目(973 计划)(编号: 2012CB722304)和河南省基础与前沿技术研究计划项目(编号: 102300413213)资助

作者简介: 秦丹, 硕士研究生, 专业方向: 细胞分化调控。Tel: 18336063263; E-mail: qindan2008@126.com

通讯作者: 徐存拴, 博士, 教授, 研究方向: 细胞分化调控。E-mail: xucs@x263.net

网络出版时间: 2013-9-12 2:44:29

URL: <http://www.cnki.net/kcms/detail/11.1913.R.20130912.0244.005.html>

中至少 80.4% 的序列都具有一定的生化活性, 它们可作为转录因子结合位点、DNA 修饰靶点, 也可转录为非编码 RNA<sup>[2]</sup>。这预示着人类基因组中至少 80% 的序列都有某些特定的功能。其中, 至少存在 400 万个遗传开关, 通过这些遗传开关的开启或关闭来调节蛋白质编码基因的表达、生物多样性的形成以及疾病的发生<sup>[3]</sup>。此外, 基因组中的非编码序列在 DNA 修复、免疫作用、新基因的形成和基因组的进化过程中起着重要作用<sup>[4]</sup>。目前, 非编码 DNA 序列的生物学作用, 尤其是对基因表达的调节作用日益受到人们的重视, 成为功能基因组学的研究热点。

## 1 非编码 DNA 序列的结构

综合各种序列比较分析结果(同一个体、同一物种不同个体间、不同物种间的基因组序列比较), 根据长度、数量以及在基因组中的覆盖率等特征, 可将非编码 DNA 序列分为以下几类(表 1)<sup>[5,6]</sup>。由表 1 可知, 非编码 DNA 多为重复序列。其中, 所占比例最大的是可转座元件, 以自我复制到新位点的方式增加基因组的规模。其次是变异结构, 其长度比单核苷酸多态性位点更长, 可达数十万碱基对, 通过染色体重排, 例如复制、删除、新序列插入或倒位等产生。重复片段是指重复单元较长的序列, 每单元长度约 1 000 到数百万碱基对, 同一条染色体上不同重复单元间通常相距 1 Mb 以内, 彼此有着 90% 以上的相似度。假基因在基因组中也大量存在, 包括复制型假基因、加工型假基因等, 由蛋白编码基因发生变异而来, 人体内多达 2 万个, 几乎与真基因

的数量相当。此外, 还有相当一部分未被分类的非编码 DNA 序列。

## 2 非编码 DNA 序列的功能

基因组中大量的非编码 DNA 序列都有某种功能。例如, 微卫星重复序列可结合转录因子, 作为应答元件调节某些肿瘤相关基因的表达<sup>[7]</sup>; 反转座子中存在 23 000 个候选调节区以及 2 000 多个双向转录区<sup>[8]</sup>, 短散在元件和长散在元件都对启动子具有调节作用<sup>[9]</sup>; 至少 20% 的人类假基因发生了转录<sup>[10]</sup>, 产生的反义转录本可以调节原基因的表达<sup>[11,12]</sup>; 一些重复序列具有拷贝数依赖的调节活性, 富含转录因子结合位点, 并且可进行非编码转录, 协同地参与人类某些遗传疾病的发生<sup>[13]</sup>; 基因间隔区的转录可通过影响某些基因的表达, 引发疾病等<sup>[14]</sup>。同时, 在非编码 DNA 序列中, 研究者发现了 399 124 个具有类似于增强子特征的区域以及 70 292 个具有类似于启动子特征的区域<sup>[2]</sup>, 使已知的顺式调节元件数量大大增加。这些研究结果表明, 非编码 DNA 序列中分布着众多功能元件, 它们可作为顺式调节元件和/或产生非编码 RNA(Non-coding RNAs, ncRNAs) 发挥作用<sup>[5,15]</sup>。

### 2.1 顺式调节元件的功能

研究表明, DNA 非编码区通常包含一些重要的顺式调节元件, 如启动子、增强子、绝缘子和沉默子等。这些调节元件在基因组中广泛分布, 可以近距离起作用, 也可以出现在距离靶基因较远的 5' 端

表 1 非编码 DNA 序列的结构分类

类型	亚类	平均长度(bp)	数量	覆盖率(%)
短串联重复序列	简单重复	63	415 917	0.84
	卫星 DNA	1 444	8 997	0.42
	低复杂度序列	46	370 102	0.55
DNA 转座子	自主型、非自主型	215	459 524	3.17
反转座子	长散在元件	426	1 490 241	20.4
	短散在元件	约 200	1 600 000	13
假基因	复制型	6 607	2 413	0.51
	加工型	723	8 303	0.19
重复片段	无	5 740	26 469	4.89
变异结构	无	8 761	96 874	27.3
未分类的间隔 DNA	无	不确定	不确定	25

或 3'端区域, 它们可以独自或者协同作用调节基因表达<sup>[16]</sup> (图 1)。RNA pol 启动子通常位于它所调控的基因 5'端上游附近, 分为核心启动子与近端启动子, 是基因表达的“开关”。核心启动子位于转录起始位点附近 100 bp 范围内, 通过结合通用转录因子促进 RNA pol 的成分在转录起始位点聚集, 从而启动基因的转录。而近端启动子距转录起始位点稍远一些, 但通常锁定在几百 bp 范围内, 可以为一些激活蛋白提供结合位点, 启动基因的组织特异性表达<sup>[17]</sup> (图 1a)。增强子多为重复序列, 长约 50 bp, 适于结合蛋白因子, 其内部通常含有一个核心序列(G)TGGA/TA/TA/T(G), 是产生增强效应所必需的。增强子通过结合激活蛋白, 远距离地作用于启动子, 从而提高基因的转录效率(图 1b)。例如, 音猬因子 *SHH* 的增强子位于距离音猬因子基因 1 Mb 的 *LMBRI* 基因的一个内含子中<sup>[18]</sup>。沉默子是对靶基因具有负调节作用的元件(图 1c), 可以出现在增强子内, 也可以

作为一种独立的元件起作用, 也有一些沉默子只在位于启动子和非翻译区(UTR)之间时起作用<sup>[19]</sup>。绝缘子通过形成不连续的调节结构域, 维持毗邻基因转录的保真性。这类调节元件具有增强子屏蔽活性(图 1d)和/或抑制活性(图 1e)。脊椎动物的结合转录因子 CTCF 是绝缘子发挥增强子抑制活性的普遍条件<sup>[17,20]</sup>。同时, 上述增强子、沉默子、绝缘子还可聚簇形成基因座控制区, 整合多种调节元件所下达的指令, 调控区域内基因表达。非编码 DNA 序列还可以形成富含 AT 的核基质结合区, 结合核基质或者染色质重塑蛋白, 调节染色质结构, 形成有利于转录因子接近和结合的位点<sup>[21,22]</sup>。

## 2.2 非编码转录本的功能

基因组中可转录的区域有 70%~90%, 产生的非编码转录本约占整个转录组的 2/3<sup>[18,23]</sup>。这些非编码 RNA 可根据长度划分为两类(图 2): 少于 200 个

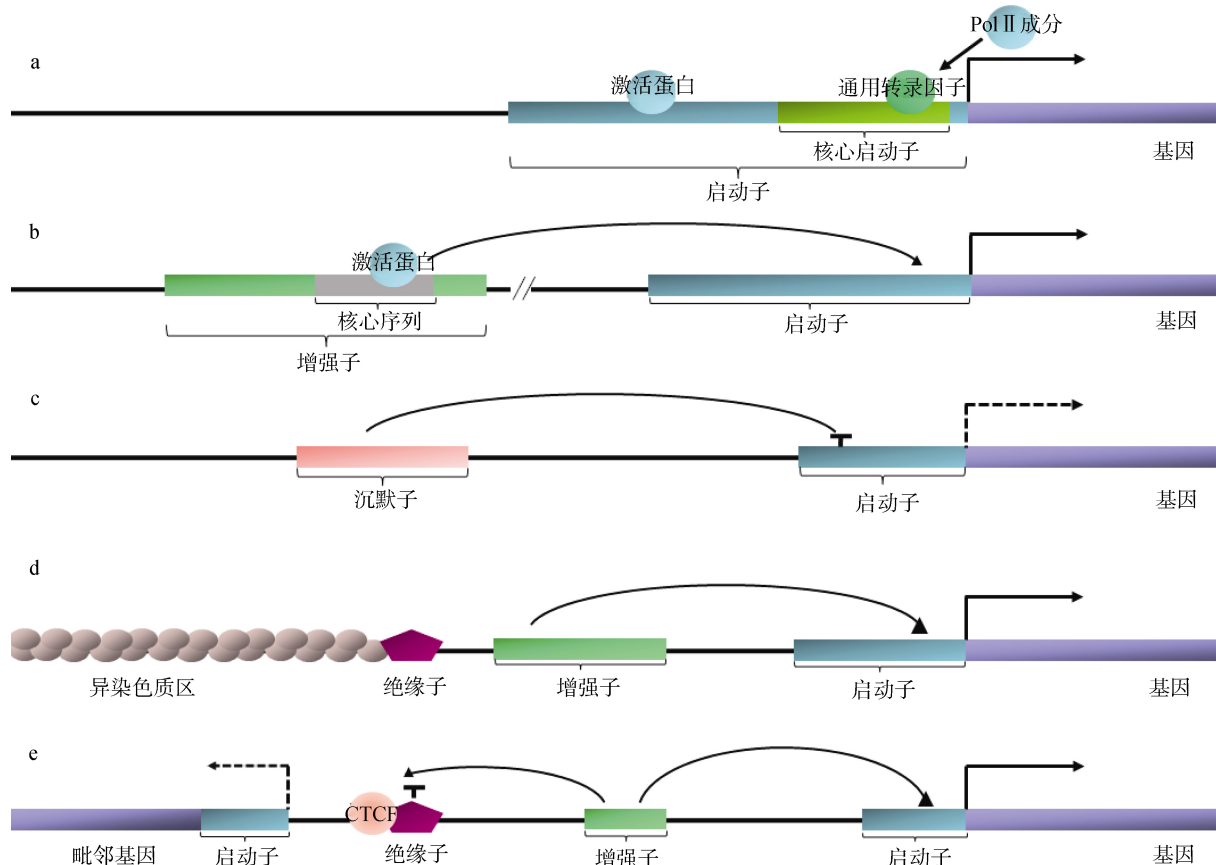


图 1 非编码 DNA 序列中的顺式调节元件

a: 启动子结构及作用模式; b: 增强子结构及作用模式; c: 沉默子及作用模式; d: 屏蔽增强子活性的绝缘子及其作用模式; e: 抑制增强子活性的绝缘子及其作用模式。

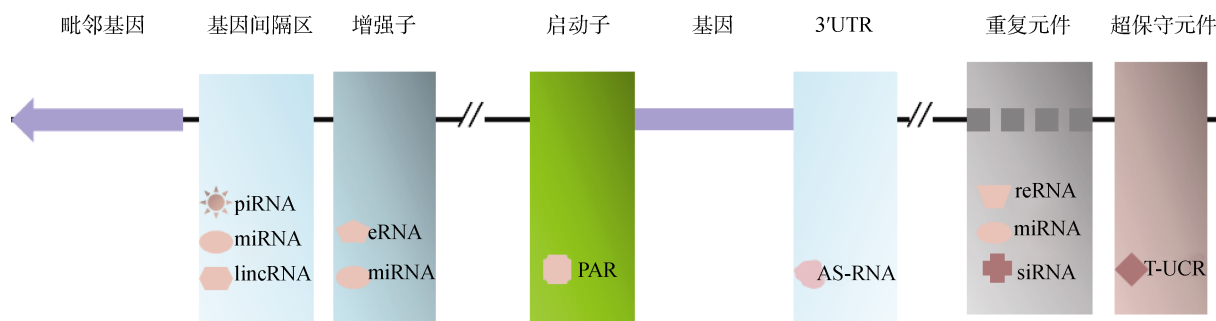


图 2 非编码转录本的代表类型及其来源

核苷酸的短RNA或者小RNA, 包括piRNA、siRNA和miRNA等<sup>[24,25]</sup>; 长度大于 200 个核苷酸的长非编码RNA(Long non-coding RNA, lincRNA)<sup>[26,27]</sup>, 包括反义RNA(AS-RNA)、启动子相关转录本(PAR)、增强子相关RNA(eRNA)、基因间隔区RNA(Long intergenic non-coding RNA, lincRNA)、超保守元件转录本(T-UCR)和假基因转录本等。ncRNAs可通过自身转录、DNA甲基化、基因印记、组蛋白修饰、染色质重塑等表观遗传学途径调控基因表达, 影响多种生物学过程。

在基因转录水平上, ncRNAs可通过自身转录行为, 包括转录起始复合体的组装以及Pol I 沿着染色质移动, 改变染色质构象或者转录因子的结合, 提供持续、低水平的转录活性以维持一个开放的染色质状态, 从而影响周围的染色质结构以及基因的转录<sup>[28]</sup>; ncRNAs还可以通过自身结构特性结合一些调节因子, 引起染色质修饰与/或DNA甲基化的改变, 从而影响基因的转录<sup>[29]</sup>。在体细胞向诱导型多能干细胞(Induced pluripotent stem cell, iPSC)转化的重编程过程中, lincRNA参与了表观基因组和基因表达的整体重塑<sup>[30]</sup>, 预示着它可能参与了细胞分化命运的调节。*H19* 基因的母系等位基因特异性地编码一种lincRNA, 在基因组印记中起着重要作用<sup>[31]</sup>。Xist的反义RNA Tsix可诱导Xist/Tsix区域组蛋白甲基化, 对该区域染色质重塑具有重要作用, 有助于启动X染色体失活<sup>[32]</sup>。在裂殖酵母中, 着丝粒处染色质修饰活动以及异染色质的形成依赖于一些小RNA<sup>[33]</sup>。采用突变技术破坏piRNA的产生, 能够改变异染色质化所引起的重复元件沉默, 预示着piRNA可以驱动相应位点染色质状态的改变<sup>[34]</sup>。

ncRNAs还可以在基因转录后水平发挥表观遗传

传学效应, 参与基因表达调控。例如, 小RNA可与AGO蛋白结合成复合体, 降解与其互补的mRNA或抑制其翻译<sup>[23]</sup>。人类*FAS*基因内含子 1 对应的互补链区域, 可转录产生AS-RNA, 选择性剪接蛋白质编码基因的初始转录本, 使其产生不同类型的蛋白质, 保护T细胞免受与*FAS*基因相关的细胞凋亡<sup>[35]</sup>。反义序列*Wrap53* RNA可通过靶定*p53* mRNA的5'端非翻译区, 与其相互作用, 调节*p53* mRNA及其蛋白水平<sup>[36]</sup>。

ncRNAs引起的表观遗传异常改变与某些疾病相关。例如, *HBA2* 特异性的AS-RNA可诱导CpG岛发生甲基化从而引起*HBA2* 基因表达沉默, 引发 $\alpha$ -地中海贫血症<sup>[37]</sup>。*p15* 反义RNA的表达可引发异染色质化、DNA甲基化, 抑制*p15* 正义基因表达, 构成白血病发生的部分分子机制<sup>[38]</sup>。LincRNA可以招募染色质修饰复合体或者RNA结合蛋白来改变基因表达程序<sup>[39]</sup>, 其作为癌症潜在的诊断和预后指标以及治疗靶点已引起关注<sup>[14]</sup>。

此外, ncRNAs还与表观遗传的记忆有关。起初, 人们认为组蛋白修饰是一种永久的标记, 可以在细胞世代间遗传。随着研究的深入, 现在推测组蛋白修饰是表观遗传的执行者, 而RNA则是表观遗传记忆的传递者<sup>[40]</sup>。在配子形成及受精过程中, 染色质重塑对随后的染色质状态有着深远的影响, 先后在植物和小鼠中发现, 精子中出现的小RNA可以被传递到后代, 在后代中引发表观遗传表型<sup>[41]</sup>。以小RNA这种核酸分子传递某种特定染色质状态, 这样的模型可能是完美的。

### 3 非编码 DNA 序列的功能元件鉴别

细胞核中蛋白质与DNA特异性结合, 进而招募



染色质修饰酶, 改变染色质结构, 是调控真核细胞基因组功能, 介导调节元件发挥效应的重要机制<sup>[42]</sup>。未被解析的非编码DNA序列中分布着许多调节元件, 提供了大量的转录因子结合位点以及染色质修饰靶点, 同时, 基因组还存在着广泛转录, 产生大量不翻译蛋白质的功能性转录本, 以ncRNAs的形式发挥作用。因此, 鉴别非编码序列中的功能元件可以从识别转录因子结合位点、染色质修饰标记以及非编码转录区着手。

### 3.1 识别转录因子结合位点

#### 3.1.1 根据已知信息预测

迄今, 通过实验所鉴定出的对应不同转录因子的调节元件有 4 000 多种, 这些信息已被收录到 TRANSFAC<sup>[43]</sup>和 JASPAR<sup>[44]</sup>等数据库中。研究者可以提交序列到相应数据库, 通过 Expectation- Maximization algorithms<sup>[45]</sup>、Gibbs-sampling methods<sup>[46]</sup>等算法进行比对, 找出潜在的结合位点及其对应的转录因子。目前, 在算法方面的改善主要集中于提高特征序列的分辨率, 以及同时识别多种特征序列。

然而, 真核细胞中, 往往多种特定转录因子紧密地结合在一起, 形成顺式调节模块(Cis-regulatory modules, CRMs), 以此调控基因转录<sup>[47]</sup>。而且, 单个转录因子结合位点的长度通常在 5~20 bp 范围内<sup>[48]</sup>, 在较大规模的基因组中, 序列如此短的DNA片段重复出现的频率很高, 许多计算方法都能把这些位点收录在内, 导致识别单个结合位点往往存在假阳性偏高问题。考虑到这些特点, 只鉴定出单个或少数结合位点不足以揭示其对基因转录的调控作用。因此, 有必要识别出多个结合位点聚簇而成的复合顺式调节模块, 这也是调控序列计算分析的活跃研究领域。这种方法已被广泛地应用于启动子的识别, 例如 PromoterScan、PromFind、PromoterInspector、McPromoter<sup>[17]</sup>等。融合了基因表达加帽分析(CAGE)的计算方法, 大大提高了识别核心启动子以及转录起始位点的分辨率与精确度。类似地, 统筹考虑序列保守性、基因表达数据、染色质状态, 以及多个转录因子结合位点之间内在联系的CRMs算法也已经被用于预测远端调节元件。

预测顺式调节模块既要考虑那些高密度结合位点, 还需考虑不同结合位点的组合方式。可对同一

类转录因子结合位点进行同型聚类, 以发现具有某种特定功能的CRM。也可聚类某一特定生理活动或细胞类型中的功能元件位点, 以特征序列的出现频率以及相对分布来构建模型。其中, 定量高分辨率成像技术已被用于检测细胞核中能与DNA序列结合的转录因子密度, 以转录因子密度及其位置权重矩阵来模拟一段DNA序列调节基因表达的可能性<sup>[49]</sup>。结合DNase HS 信息和/或组蛋白修饰情况所预示的染色体结构, 会使CRMs的预测精确度得到更大提高。

#### 3.1.2 *de novo* 预测

许多转录因子的DNA结合模式是未知的, 顺式调节模体内转录因子的共结合信息也非常匮乏。为解决该问题, 以共调控基因有着相同或相似的调节机制这一假设为基础, 科研人员开发出了一系列从头发现转录因子结合位点的方法。这种方法以一组共调控基因作为输入信息, 用吉布斯取样、Monte Carlo 策略、多重位置权重矩阵<sup>[50~54]</sup>、后缀树结构<sup>[55,56]</sup>等算法搜索在这些基因上游调控序列中富集的特征序列。

#### 3.1.3 系统发育足迹分析

生物进化过程中由于选择压力, 具有序列特异性的功能区域突变积累较为缓慢, 因而导致不同物种间的这些区域具有较高的序列相似性, 物种间同源基因经受着相同的调节机制。基于这一假设, 已建立了系统发育足迹分析法(Phylogenetic footprinting)来预测转录因子结合位点, 相应的软件有FootPrinter<sup>[57]</sup>和Gibbs Motif Sampler<sup>[58]</sup>等。预测过程可分为 3 个步骤: 选择进化分歧合适的物种, 确定待比较的同源基因; 序列比对; 筛选出显著保守的非编码序列。研究者已用这种方法鉴定出了灵长类 *COX5B*、脊椎动物 *HOXB-1* 和 *HOXB-4*、哺乳动物 *IL-2Ralpha* 等基因的启动子、增强子之类的调节元件<sup>[59]</sup>。然而, 保守和功能并非一定对应, 转录因子结合位点也并非都有跨物种保守性。为了进一步加强预测的准确性, 可比较两种以上物种的同源序列, 也可考虑以比对特征序列代替单个核苷酸的比对, 把CRMs模型与系统发育足迹法叠加应用等。

当然, 以上这些识别转录因子结合位点的计算方法都着重于分析 DNA 序列, 预测 DNA 序列与转

录因子结合的可能性。而对于二者的结合强度以及结合能的高低对靶基因表达的影响,则缺乏定量的评价。此外,功能元件对基因转录的调控离不开染色质结构的影响,因此,对非编码调控序列的识别,可以考虑开发一些计算方法,描述其与转录因子结合形成的构象所带来的染色质效应。随着科技的发展,这些问题都可以得到解决,预测非编码 DNA 调控序列的精确性也将不断提高。

### 3.2 识别染色质修饰标记

组蛋白密码假说<sup>[60]</sup>认为,某种程度上,遗传信息的转录受控于组蛋白的化学修饰。活性启动子区和增强子区富含H3K4me3 和H3K4me1<sup>[61]</sup>,活性增强子元件带有H3K27ac,非活性或平衡状态的增强子元件只携带H3K4me1 标记<sup>[62]</sup>,大多数基因的启动子区含有H3K4me3 和H3K9ac标记<sup>[63]</sup>,常染色质和异染色质区域的边界元件绝缘子可与转录因子CTCF结合,并招募组蛋白修饰酶,以此维持常染色质或异染色质状态。因此,借助于组蛋白上带有的特殊修饰标记,能够识别出不同的转录调节元件。

Wang 等<sup>[64]</sup>采用 8 种不同的组蛋白标记(H3K4me1、H3K4me2、H3K4me3、H3K9ac、H3K27ac、H3K36me3、H3K20me1 和CTCF),并分别使用线性条件随机域模型(CRF)及 hidden Markov model (HMM)来识别调节元件,不同的标记组合预示着不同类型的调节元件。其分析过程包括:(1)输入信息;(2)分别建立多元HMM模型以及CRF模型以发现染色质状态;(3)用训练过的HMM和CRF模型预测染色质状态;(4)借助外部数据,如参考序列的转录起始位点以及DNase I 高敏感信息等,分析不同组合标记的频率及丰度,并以此来解析染色质状态。在预测正调控相关区域方面,此方法显示出强大的潜力。若要进一步提高预测结果的精确性以及解析的可靠性,则需参考更多其他实验数据的附加信息。

### 3.3 识别非编码转录区

常用的方法有Tilling arrays、RNA-Seq、RNA CaptureSeq等。Tilling arrays<sup>[65]</sup>技术是无任何偏倚地把基因组的双链序列,或按着一定的间隔规律、或者以序列交叠的方法、或者以序列首尾相接的方式制成探针,对全基因组水平的转录信息进行高通量探测。其探针的筛选和芯片制备不依赖于已有的基

因组注释信息。它的这些特点和优势非常适用于研究非编码DNA序列中的转录区。其识别表达信号的算法有滑窗(Sliding window, SW)法<sup>[66]</sup>、亮度分布(Signal distribution, SD)法<sup>[65]</sup>、HMM信号识别法<sup>[67]</sup>等。RNA-Seq<sup>[68]</sup>,即RNA测序,又称转录组测序,是最近发展起来的利用深度测序来分析转录组的技术。该技术能够在单核苷酸水平上检测基因组的整体转录活动,提供全面而精确的转录组信息。相对于传统的芯片杂交技术,它无需设计探针,具有更高的检测通量以及更广泛的检测范围,可用于研究转录本的结构、非编码区的功能以及发现低丰度全新转录本。然而,对RNA-Seq结果的处理和分析也面临着挑战,包括:如何把测序读段准确定位到参考转录组或基因组上,根据测序结果重建转录组,转录本表达水平、不同情况下表达差异的量化分析等。为克服RNA-Seq的某些缺陷,研究人员又开发出了一种靶向RNA测序新技术RNA CaptureSeq<sup>[69]</sup>。该技术结合了序列捕获以及饱和测序方法,使用定制型基因芯片,用于探测目标区域的转录信息。与常规RNA测序相比,能够得到更高的覆盖深度,适用于研究特定区域的详细转录信息,可鉴定出转录水平很低,或者具有转录差异的细胞亚群中(各发育阶段的细胞、不同组织的细胞、病变细胞等)出现的转录子。通过这种方法,只对特定区域序列感兴趣的研究人员可以低成本、高效益地获取相关信息,例如,可根据全基因组关联研究(Genome wide association studies, GWAS)所提供的信息,集中研究疾病相关的非编码区转录。

## 4 非编码 DNA 的功能验证

目前,用计算方法预测顺式调节元件、RNA-Seq等技术识别非编码转录区,普遍存在假阳性偏高问题。因此,需要通过实验验证分析结果,现将常用的验证方法总结如下。

### 4.1 检测报告基因(Reporter assays)

根据调节元件的种类和作用,设计质粒内部各调节元件和报告基因的排列。将构建好的质粒导入受体细胞,通过检测报告基因的表达情况确定调节元件的活性。如果导入的受体细胞是受精卵,则可以反映调节元件的组织特异性作用。

## 4.2 检测电泳迁移率变化(Electrophoretic mobility shift assay, EMSA)

DNA 片段与转录因子发生特异性结合后, 其电泳时的迁移率会发生改变, 以此确认二者的互作。此方法也可用于分析非编码 DNA 的转录本 ncRNA 与特定蛋白的结合。然而, 该方法依赖特定蛋白与靶序列的高亲和力。因此, 许多情况下, 只能获得少数靶序列。

## 4.3 紫外交联免疫沉淀分析(Cross-linking and immunoprecipitation, CLIP)

ncRNAs 可结合蛋白质(RNA-binding proteins, RBPs), 形成核糖核蛋白(RNPs)复合物, 发挥调节作用。目前, 用于研究非编码RNA功能的技术有: 紫外交联免疫沉淀结合高通量测序(Cross-linking immunoprecipitation and high-throughput sequencing, CLIP-Seq)、RNA纯化的染色质分离(Chromatin isolation by RNA purification, ChIRP)以及非编码RNA沉默与定位分析(Combined knockdown and localization analysis of non-coding RNAs, c-KLAN)等。CLIP-Seq<sup>[70]</sup>, 也称HITS-CLIP, 是一项在转录组水平揭示RNA分子与RBPs互作的技术。其原理是: 在紫外照射下, RNA分子可与RBPs发生共价结合, 形成的复合物又可被RBPs的特异性抗体沉淀, 回收沉淀, 提取其中的RNA片段进行高通量测序和分析。在此基础上发展的两种改良方法有光活性增强的核糖核苷紫外交联免疫共沉淀(Photoactivatable ribonucleoside enhanced CLIP, PAR-CLIP)<sup>[71]</sup>以及单核苷酸分辨率紫外交联免疫共沉淀(Individual-nucleotide resolution CLIP, iCLIP)<sup>[72]</sup>, 与传统的RNA免疫沉淀相比, 它们能够特异性识别与RNAs直接结合的RBPs, 并能鉴定出其在RNAs分子上特定的识别或结合基序。此外, ChIRP用于研究RNA和染色质的互作, 与高通量测序技术结合(ChIRP-Seq)可用来在全基因组范围内定位ncRNAs的染色体结合位点<sup>[73,74]</sup>。c-KLAN技术主要用于对ncRNAs进行功能缺失研究和细胞定位, 该技术结合了核糖核酸内切酶制备的小干扰RNA(Endoribonuclease-prepared siRNA, esiRNA)和荧光原位杂交技术(Fluorescence in situ hybridization, FISH), 提供了一种可靠、快速检测ncRNAs调控作用的方法<sup>[73,75]</sup>。

## 4.4 染色质结构分析

通过分析染色质结构鉴定DNA调节元件的方法有Dnase 足迹法、染色质免疫共沉淀(Chromatin immunoprecipitation, ChIP)等。Dnase 足迹法<sup>[76,77]</sup>是根据Dnase 超敏感位点(Dnase hypersensitive sites), 即染色体上核小体结合比较松散或者无核小体结合的部位, 对Dnase 的消化较为敏感, 这些部位通常是结合了转录因子的调节元件。收集这些片段进行芯片分析(DNase-chip)或高通量测序(DNase-Seq), 获得其信息。ChIP方法<sup>[78]</sup>是在活细胞状态下, 用甲醛作为交联剂, 固定DNA-蛋白质复合物, 获得二者的互作信息, 很大程度上改善了对结合位点的评估。将该方法结合芯片技术(ChIP-chip)或者高通量测序(ChIP-Seq), 可以鉴定出某一给定转录因子在全基因组范围内的结合位点。目前, 研究者已鉴定出了SP1 的 12 000 个结合位点, p53 蛋白的 1 600 个结合位点<sup>[79]</sup>。此外, ChIP方法还可用来研究组蛋白的共价修饰与基因表达的相关性, 以及ncRNAs(RNA-ChIP)在基因表达调控中的作用。然而, ChIP方法也具有一定的局限性, 它需要提供大量的细胞样品( $\sim 10^7$ 个细胞)、对DNA-蛋白质直接或间接互作的分辨率不高、需要复杂程序来分析实验数据等, 需要进一步改进和提高。

## 4.5 染色质构象分析

上述Dnase 足迹法、ChIP方法尚不能分析染色质高级结构的构象。用于分析染色质空间构象的方法有染色质构象捕获(Chromatin Conformation Capture, 3C)及其衍生方法<sup>[80]</sup>、配对末端标签测序分析染色质互作(Chromatin interaction analysis by paired-end tag sequencing, ChIA-PET)技术<sup>[81]</sup>等。3C技术用于揭示转录因子介导的DNA调控元件之间的相互作用。如上所述, 增强子结合与目标启动子部位转录机器有关的转录因子, 引起增强子元件和启动子元件的间接接触, 发挥其转录调节功能。因此, 可以通过检测这种互作关系来鉴定增强子元件。在3C基础上开发的新的检测方法有: 4C (Chromosome conformation capture on-chip/circular chromosome conformation), 5C(Chromosome conformation capture carbon copy)和Hi-C。其中, 4C、5C的优势是采用了芯片技术或高通量测序技术, 可在一次



实验中识别出某一给定启动子与多个调控元件的互动。最新出现的 Hi-C 技术用生物素对互作的序列贴上标签,以便后续大规模平行测序的富集和鉴定,它已被用于构建低分辨率的染色质互作图谱。随着测序覆盖度和分辨率的提高,该技术有望应用于全基因组范围内鉴定远距离的调节互作。ChIA-PET 技术是把染色质免疫沉淀技术、染色质邻近式连接技术、配对末端标签技术和新一代测序技术融为一体,在基因组三维折叠和套环状态下分析基因表达和调控。叠加 Hi-C 图谱和 ChIA-PET 图谱,以及现有的注释,将有助于了解三维空间的基因调控。

## 5 展望

基因表达的调控是在基因组水平上,由一系列调节元件、转录因子、辅助因子等相互作用,共同构成复杂的调控网络来完成的。一直以来,人们主要关注决定生物体结构和功能多样性的蛋白质,以及编码这些蛋白质的基因,认为那些非编码序列都是无用的垃圾。实际上,这些非编码 DNA 才是遗传的中心,正是由它们决定基因何时何地何量表达,以及如何高效生产出各种蛋白质。随着科技的发展、新方法的开发,研究者能够从基因组中获得大量的信息。然而,解读基因组中非编码调控序列,完善基因组注释,进而提高人们对基因表达调控的理解,尚面临着巨大的挑战。

在顺式调节元件研究方面,如何鉴定出调控基因表达的功能元件,尤其是增强子、沉默子之类的远距离作用元件是相当困难的。目前,许多研究者正致力于用计算生物学和功能基因组学方法,在全基因组范围内鉴定调节元件与转录因子互作的结合模式及互作时染色质结构状态,以挖掘出更多远距离顺式调节元件。今后的努力方向包括:系统分析各种方法所得数据,使其变为更有意义的信息;开发灵敏度、精确度更高的预测方法,发现更多调节元件,同时,开发不依赖预测等附加信息的方法,直接鉴定基因组范围内对应某一类顺式调节元件的全部转录因子;开发出合适的高通量方法,系统地研究对基因表达起负调控作用的调节元件;探索顺式调节元件在时间(生理或病理过程各阶段)和空间上的(物种、个体、组织)特异性作用;评估可能引发疾病的非编码调节序列的变异。

在非编码 RNA 研究方面,尽管越来越多的 ncRNAs 被发现,极大地丰富了人们对 RNA 世界及基因表达调控复杂性的认识,但对它们的研究,尤其是长非编码 RNA 的研究还处于起始阶段。目前,对于 ncRNAs,主要集中于研究其来源、表达模式、调控基因表达的分子机制、表观遗传学效应,以及对某些生理或病理过程的影响。解密 ncRNAs 所传达的信息仍是一个巨大的挑战,许多问题尚待解决:在分子水平上,ncRNAs 功能基序的变异对其功能结构域有何影响,及其如何实现与蛋白质特异性互作,靶定染色质修饰复合体到特定染色质区域,并以何种信号途径参与细胞功能的调节;其细胞定位对其发挥功能的贡献;其自身的生理行为、活性及功能又是受何种机制所调节的;不同 ncRNAs 之间的联系及如何形成复杂的基因表达调控网络;ncRNAs 在时间、空间上的表达特异性;建立更多预测及验证 ncRNAs 功能的计算方法和实验方法。

在研究手段方面,高通量测序及分子交联技术为识别和分析顺式调节元件及非编码转录区的功能奠定了基础,但仍需改进,今后的努力方向包括:发展第三代测序技术,以获得更长的读长,高效拼接重复序列,进一步提高测序结果的可靠性;发展高效处理和分析高通量测序所产生海量数据的生物信息学方法,这也是高通量测序技术能够在这一科学领域中发挥重大作用的关键;以紫外光代替甲醛作为交联剂提高了交联特异性,然而,常规紫外交联的低效率会导致高的噪声比,难以分辨交联与非交联的靶 RNA,可发展使用改良方法,以获得真实交联的序列。

最后,非编码 DNA 发挥调节作用的方式不是单一的,综合考虑各要素的作用,有可能发现新的调控机制。构建包含非编码 DNA 序列、非编码转录本、调控蛋白在内的基因表达调控网络,从基因组、转录组、蛋白质组水平全面认识非编码 DNA 序列的作用,对于揭示生物发生、发育、进化、遗传变异、疾病等生命现象的本质和机理具有重要的意义。

## 参考文献(References):

- [1] Ponting CP, Belgard TG. Transcribed dark matter: meaning or myth? *Hum Mol Gene*, 2010, 19(R2): R162-R168.  
[\[DOI\]](#)



- [2] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 2012, 489(7414): 57–74. [\[DOI\]](#)
- [3] Vickers KC, Palmisano BT, Remaley AT. Remaley. The role of noncoding "junk DNA" in cardiovascular disease. *Clin Chem*, 2010, 56(10): 1518–1520. [\[DOI\]](#)
- [4] Xie C, Zhang YE, Chen JY, Liu CJ, Zhou WZ, Li Y, Zhang M, Zhang R, Wei L, Li CY. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genet*, 2012, 8(9): e1002942. [\[DOI\]](#)
- [5] Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. Annotating non-coding regions of the genome. *Nat Rev Genet*, 2010, 11(8): 559–571. [\[DOI\]](#)
- [6] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrum J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissole SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang HM, Yu J, Wang J, Huang GH, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kasprzyk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrino A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 2001, 409(6822): 860–921. [\[DOI\]](#)
- [7] Gangwal K, Lessnick SL. Microsatellites are EWS/FLI response elements: genomic "junk" is EWS/FLI's treasure. *Cell Cycle*, 2008, 7(20): 3127–3132. [\[DOI\]](#)
- [8] Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, Waki K, Hornig N, Arakawa T, Takahashi H, Kawai J, Forrest AR, Suzuki H, Hayashizaki Y, Hume DA, Orlando V, Grimmond SM, Carninci P. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet*, 2009, 41(5): 563–571. [\[DOI\]](#)
- [9] Zuckerkandl E, Cavalli G. Combinatorial epigenetics, "junk DNA", and the evolution of complex organisms. *Gene*, 2007, 390(1–2): 232–242. [\[DOI\]](#)
- [10] Zheng DY, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, Lu YT, Denoeud F, Antonarakis SE, Snyder M, Ruan YJ, Wei CL, Gingeras TR, Guigó R, Harrow J, Gerstein MB. Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res*, 2007, 17(6): 839–851. [\[DOI\]](#)
- [11] Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, Hannon GJ. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, 2008, 453(7194): 534–538. [\[DOI\]](#)
- [12] Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, Surani MA, Sakaki Y, Sasaki H. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, 2008, 453(7194): 539–543. [\[DOI\]](#)
- [13] Cabianca DS, Casa V, Bodega B, Xynos A, Ginelli E, Tanaka Y, Gabellini D. A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. *Cell*, 2012, 149(4): 819–831. [\[DOI\]](#)

- [14] Tsai MC, Spitale RC, Chang HY. Long intergenic non-coding RNAs: new links in cancer progression. *Cancer Res*, 2011, 71(1): 3–7. [\[DOI\]](#)
- [15] Hemberg M, Gray JM, Cloonan N, Kuersten S, Grimmond S, Greenberg ME, Kreiman G. Integrated genome analysis suggests that most conserved non-coding sequences are regulatory factor binding sites. *Nucleic Acids Res*, 2012, 40(16): 7858–7869. [\[DOI\]](#)
- [16] Mullapudi N, Joseph SJ, Kissinger JC. Identification and functional characterization of *cis*-regulatory elements in the apicomplexan parasite *Toxoplasma gondii*. *Genome Biol*, 2009, 10(4): R34. [\[DOI\]](#)
- [17] Narlikar L, Ovcharenko I. Identifying regulatory elements in eukaryotic genomes. *Brief Funct Genomic Proteomic*, 2009, 8(4): 215–230. [\[DOI\]](#)
- [18] Lee JT. Epigenetic regulation by long noncoding RNAs. *Science*, 2012, 338(6113): 1435–1439. [\[DOI\]](#)
- [19] Ogbourne S, Antalis TM. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem J*, 1998, 331 (Pt 1): 1–14. [\[DOI\]](#)
- [20] Noonan JP, McCallion AS. Genomics of long-range regulatory elements. *Annu Rev Genomics Hum Genet*, 2010, 11(1): 1–23. [\[DOI\]](#)
- [21] Noordermeer D, de Laat W. Joining the loops:  $\beta$ -globin gene regulation. *IUBMB Life*, 2008, 60(12): 824–833. [\[DOI\]](#)
- [22] Hart CM, Laemmli UK. Facilitation of chromatin dynamics by SARs. *Curr Opin Genet Dev*, 1998, 8(5): 519–525. [\[DOI\]](#)
- [23] Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest ARR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impimbato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christofels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasaki Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pe-sole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schönbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusica V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y; FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). The transcriptional landscape of the mammalian genome. *Science*, 2005, 309(5740): 1559–1563. [\[DOI\]](#)
- [24] Moazed D. Small RNAs in transcriptional gene silencing and genome defence. *Nature*, 2009, 457(7228): 413–420. [\[DOI\]](#)
- [25] Kim VN. Small RNAs just got bigger: Piwi-interacting RNAs (piRNAs) in mammalian testes. *Genes Dev*, 2006, 20(15): 1993–1997. [\[DOI\]](#)
- [26] Atkinson SR, Marguerat S, Bähler J. Exploring long non-coding RNAs through sequencing. *Semin Cell Dev Biol*, 2012, 23(2): 200–205. [\[DOI\]](#)
- [27] Gibb EA, Brown CJ, Lam WL. The functional role of long non-coding RNA in human carcinomas. *Mol Cancer*, 2011, 10(1): 1–17. [\[DOI\]](#)
- [28] Bickel KS, Morris DR. Silencing the transcriptome's dark matter: mechanisms for suppressing translation of intergenic transcripts. *Mol Cell*, 2006, 22(3): 309–316. [\[DOI\]](#)
- [29] Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, Wysocka J, Lei M, Dekker J, Helms JA, Chang HY. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature*, 2011, 472(7341): 120–124. [\[DOI\]](#)
- [30] Loewer S, Cabili MN, Guttman M, Loh YH, Thomas K, Park IH, Garber M, Curran M, Onder T, Agarwal S, Manos PD, Datta S, Lander ES, Schlaeger TM, Daley GQ, Rinn JL. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet*, 2010, 42(12): 1113–1117. [\[DOI\]](#)
- [31] Gabory A, Jammes H, Dandolo L. The *H19* locus: role of an imprinted non-coding RNA in growth and development. *BioEssays*, 2010, 32(6): 473–480. [\[DOI\]](#)

- [32] Navarro P, Pichard S, Ciaudo C, Avner P, Rougeulle C. Tsix transcription across the *Xist* gene alters chromatin conformation without affecting *Xist* transcription: implications for X-chromosome inactivation. *Genes Dev*, 2005, 19(12): 1474–1484. [\[DOI\]](#)
- [33] Volpe TA, Kidner C, Hall IM, Teng G, Grewal SIS, Martienssen RA. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science*, 2002, 297(5588): 1833–1837. [\[DOI\]](#)
- [34] Pal-Bhadra M, Leibovitch BA, Gandhi SG, Chikka MR, Bhadra U, Birchler JA, Elgin SCR. Heterochromatic silencing and HP1 localization in *Drosophila* are dependent on the RNAi machinery. *Science*, 2004, 303(5658): 669–672. [\[DOI\]](#)
- [35] Yan MD, Hong CC, Lai GM, Cheng AL, Lin YW, Chuang SE. Identification and characterization of a novel gene *Saf* transcribed from the opposite strand of *Fas*. *Hum Mol Genet*, 2005, 14(11): 1465–1474. [\[DOI\]](#)
- [36] Mahmoudi S, Henriksson S, Corcoran M, Méndez-Vidal C, Wiman KG, Farnebo M. Wrap53, a natural *p53* antisense transcript required for *p53* induction upon DNA damage. *Mol Cell*, 2009, 33(4): 462–471. [\[DOI\]](#)
- [37] Tufarelli C, Stanley JA, Garrick D, Sharpe JA, Ayyub H, Wood WG, Higgs DR. Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nat Genet*, 2003, 34(2): 157–165. [\[DOI\]](#)
- [38] Yu WQ, Gius D, Onyango P, Muldoon-Jacobs K, Karp J, Feinberg AP, Cui HM. Epigenetic silencing of tumour suppressor gene *p15* by its antisense RNA. *Nature*, 2008, 451(7175): 202–206. [\[DOI\]](#)
- [39] Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. *Cell*, 2009, 136(4): 629–641. [\[DOI\]](#)
- [40] Kouzarides T. Chromatin modifications and their function. *Cell*, 2007, 128(4): 693–705. [\[DOI\]](#)
- [41] Rassoulzadegan M, Grandjean V, Gounon P, Vincent S, Gillot I, Cuzin F. RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature*, 2006, 441(7092): 469–474. [\[DOI\]](#)
- [42] Kadonaga JT. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell*, 2004, 116(2): 247–257. [\[DOI\]](#)
- [43] Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 2006, 34(Database issue): D108–D110. [\[DOI\]](#)
- [44] Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 2004, 32(Database issue): D91–D94. [\[DOI\]](#)
- [45] Liu XS, Brutlag DL, Liu JS. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol*, 2002, 20(8): 835–839. [\[DOI\]](#)
- [46] Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouzé P, Moreau Y. A higher-order background model improves the detection of potential promoter regulatory elements by Gibbs sampling. *Bioinformatics*, 2001, 17(12): 1113–1122. [\[DOI\]](#)
- [47] GuhaThakurta D. Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Res*, 2006, 34(12): 3585–3598. [\[DOI\]](#)
- [48] Ureta-Vidal A, Ettwiller L, Birney E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nature Rev Genet*, 2003, 4(4): 251–262. [\[DOI\]](#)
- [49] Surkova S, Kosman D, Kozlov K, Manu, Myasnikova E, Samsonova AA, Spirov A, Vanario-Alonso CE, Samsonova M, Reinitz J. Characterization of the *Drosophila* segment determination morphome. *Dev Biol*, 2008, 313(2): 844–862. [\[DOI\]](#)
- [50] GuhaThakurta D, Stormo GD. Identifying target sites for cooperatively binding factors. *Bioinformatics*, 2001, 17(7): 608–621. [\[DOI\]](#)
- [51] Liu X, Brutlag DL, Liu JS. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, 2001, 6: 127–138. [\[DOI\]](#)
- [52] Zhou Q, Wong WH. CisModule: *de novo* discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci USA*, 2004, 101(33): 12114–12119. [\[DOI\]](#)
- [53] Gupta M, Liu JS. *De novo* cis-regulatory module elicitation for eukaryotic genomes. *Proc Natl Acad Sci USA*, 2005, 102(20): 7079–7084. [\[DOI\]](#)
- [54] Thompson W, Palumbo MJ, Wasserman WW, Liu JS, Lawrence CE. Decoding human regulatory circuits. *Genome Res*, 2004, 14(10A): 1967–1974. [\[DOI\]](#)
- [55] Eskin E, Pevzner PA. Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, 2002, 18(Suppl. 1): S354–S363. [\[DOI\]](#)
- [56] Marsan L, Sagot MF. Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J Comput Biol*, 2000, 7(3–4): 345–362. [\[DOI\]](#)
- [57] Blanchette M, Tompa M. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res*, 2002, 12(5): 739–748. [\[DOI\]](#)
- [58] Newberg LA, Thompson WA, Conlan S, Smith TM, McCue LA, Lawrence CE. A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site pre-

- diction. *Bioinformatics*, 2007, 23(14): 1718–1727. [\[DOI\]](#)
- [59] Duret L, Bucher P. Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol*, 1997, 7(3): 399–406. [\[DOI\]](#)
- [60] Jenuwein T, Allis CD. Translating the histone code. *Science*, 2001, 293(5532): 1074–1080. [\[DOI\]](#)
- [61] Heintzman ND, Stuart RK, Hon G, Fu YT, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu CX, Ching KA, Wang W, Weng ZP, Green RD, Crawford GE, Ren B. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, 2007, 39(3): 311–318. [\[DOI\]](#)
- [62] Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA*, 2010, 107(50): 21931–21936. [\[DOI\]](#)
- [63] Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 2007, 130(1): 77–88. [\[DOI\]](#)
- [64] Wang HY, Zhou X. Detection and characterization of regulatory elements using probabilistic conditional random field and hidden Markov models. *Chin J Cancer*, 2013, 32(4): 186–194. [\[DOI\]](#)
- [65] Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu XW, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M. Global identification of human transcribed sequences with genome tiling arrays. *Science*, 2004, 306(5705): 2242–2246. [\[DOI\]](#)
- [66] Kampa D, Cheng JL, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res*, 2004, 14(3): 331–342. [\[DOI\]](#)
- [67] Du J, Rozowsky JS, Korbel JO, Zhang ZD, Royce TE, Schultz MH, Snyder M, Gerstein M. A supervised hidden markov model framework for efficiently segmenting Tiling Array data in transcriptional and CHIP-chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics*, 2006, 22(24): 3016–3024. [\[DOI\]](#)
- [68] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 2009, 10(1): 57–63. [\[DOI\]](#)
- [69] Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddleloh JA, Mattick JS, Rinn JL. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol*, 2011, 30(1): 99–104. [\[DOI\]](#)
- [70] Murigneux V, Saulière J, Roest Crolius H, Le Hir H. Transcriptome-wide identification of RNA binding sites by CLIP-seq. *Methods*, 2013, 63(1): 32–40. [\[DOI\]](#)
- [71] Hafner M, Lianoglou S, Tuschl T, Betel D. Genome-wide identification of miRNA targets by PAR-CLIP. *Methods*, 2012, 58(2): 94–105. [\[DOI\]](#)
- [72] König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. iCLIP-transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. *J Vis Exp*, 2011, 30(50): e2638. [\[DOI\]](#)
- [73] 夏天, 肖丙秀, 郭俊明. 长链非编码RNA的作用机制及其研究方法. *遗传*, 2013, 35(3): 269–280. [\[DOI\]](#)
- [74] Chu C, Qu K, Zhong FL, Artandi SE, Chang HY. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol Cell*, 2011, 44(4): 667–678. [\[DOI\]](#)
- [75] Chakraborty D, Kappei D, Theis M, Nitzsche A, Ding L, Paszkowski-Rogacz M, Surendranath V, Berger N, Schulz H, Saar K, Hubner N, Buchholz F. Combined RNAi and localization for functionally dissecting long noncoding RNAs. *Nat Methods*, 2012, 9(4): 360–362. [\[DOI\]](#)
- [76] Carey MF, Peterson CL, Smale ST. DNase I footprinting. *Cold Spring Harb Protoc*, 2013, 2013(5): 469–478. [\[DOI\]](#)
- [77] Cockerill PN. Structure and function of active chromatin and DNase I hypersensitive sites. *FEBS J*, 2011, 278(13): 2182–2210. [\[DOI\]](#)
- [78] Hao H. Genome-wide occupancy analysis by ChIP-chip and ChIP-Seq. *Adv Exp Med Biol*, 2012, 723: 753–759. [\[DOI\]](#)
- [79] Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, 2004, 116(4): 499–509. [\[DOI\]](#)
- [80] Ethier SD, Miura H, Dostie J. Discovering genome regulation with 3C and 3C-related technologies. *Biochim Biophys Acta*, 2012, 1819(5): 401–410. [\[DOI\]](#)
- [81] Zhang JY, Poh HM, Peh SQ, Sia YY, Li GL, Mulawadi FH, Goh Y, Fullwood MJ, Sung WK, Ruan XA, Ruan YJ. ChIA-PET analysis of transcriptional chromatin interactions. *Methods*, 2012, 58(3): 289–299. [\[DOI\]](#)