

lncRNA的鉴定与功能推断

姚茜岚

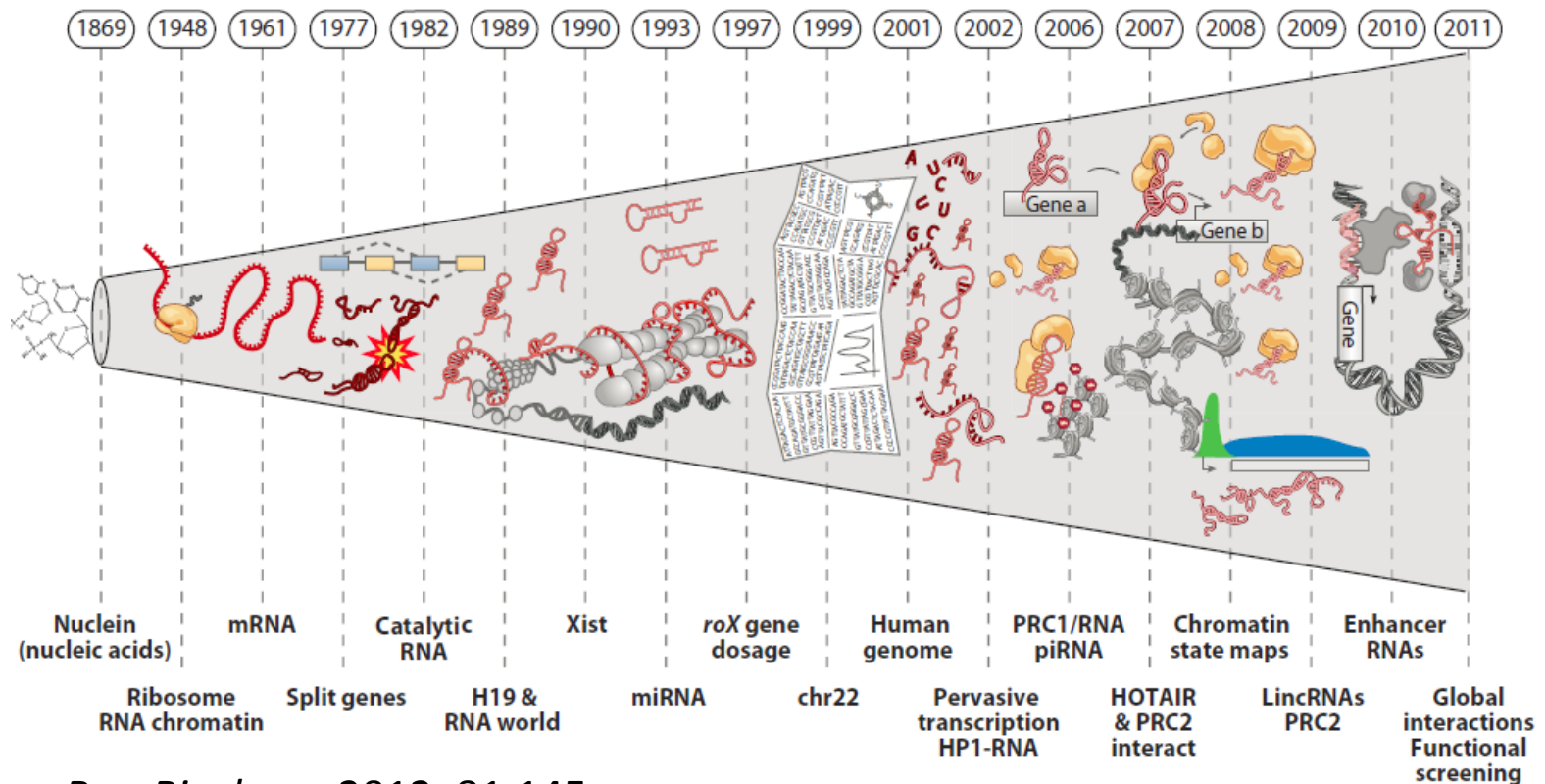
yaoqianlan@yahoo.com

提纲

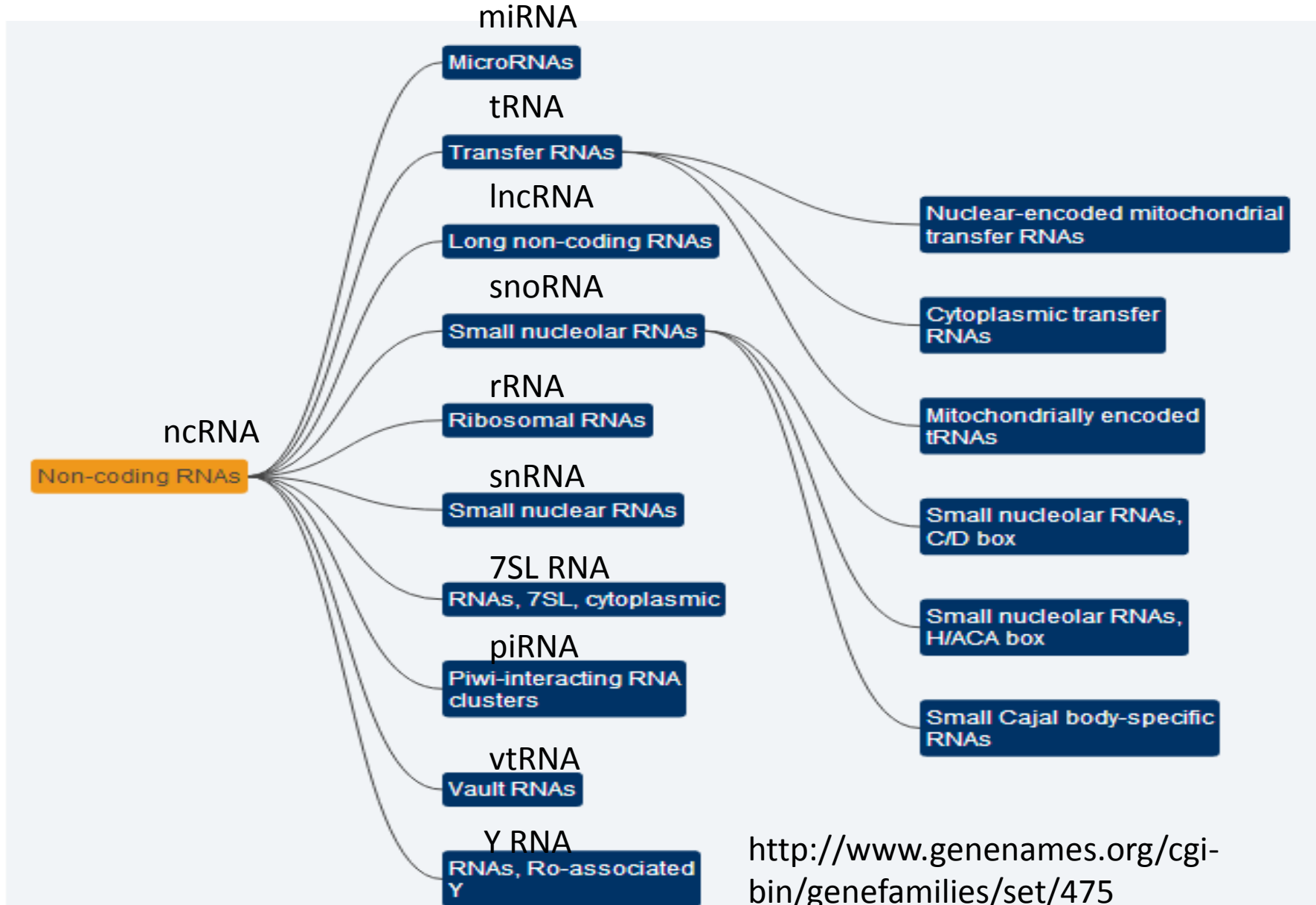
- lncRNA的生物学背景
- lncRNA的计算鉴定
- lncRNA的功能预测

非编码RNA (ncRNA)

- 经典的分子生物学中心法则
 - DNA→RNA →蛋白质
 - RNA作为遗传信息的载体编码蛋白质
- 随后50年来，各类起调控作用的非编码RNA被大量发现



人类基因组中ncRNA种类



长非编码RNA (lncRNA)

- 2002年在小鼠cDNA文库测序中首次发现
- 转录本长度大于200 核苷酸
- 缺乏编码蛋白质能力
- 位于细胞核或细胞质内
- 表达组织特异性较高，序列保守程度较低
- 可在多种层面（表观遗传、转录调控、转录后调控等）发挥调控作用

人类基因组lncRNA分布

Version 25 (March 2016 freeze, GRCh38) - Ensembl 85

General stats

Total No of Genes	58037	Total No of Transcripts	198093
Protein-coding genes	19950	Protein-coding transcripts	80087
Long non-coding RNA genes	15767	- full length protein-coding:	54755
Small non-coding RNA genes	7258	- partial length protein-coding:	25332
Pseudogenes	14650	Nonsense mediated decay transcripts	13769
- processed pseudogenes:	10725	Long non-coding RNA loci transcripts	27692
- unprocessed pseudogenes:	3400		
- unitary pseudogenes:	214		
- polymorphic pseudogenes:	51		
- pseudogenes:	21	Total No of distinct translations	60033
Immunoglobulin/T-cell receptor gene segments		Genes that have more than one distinct translations	13536
- protein coding segments:	411		
- pseudogenes:	239		

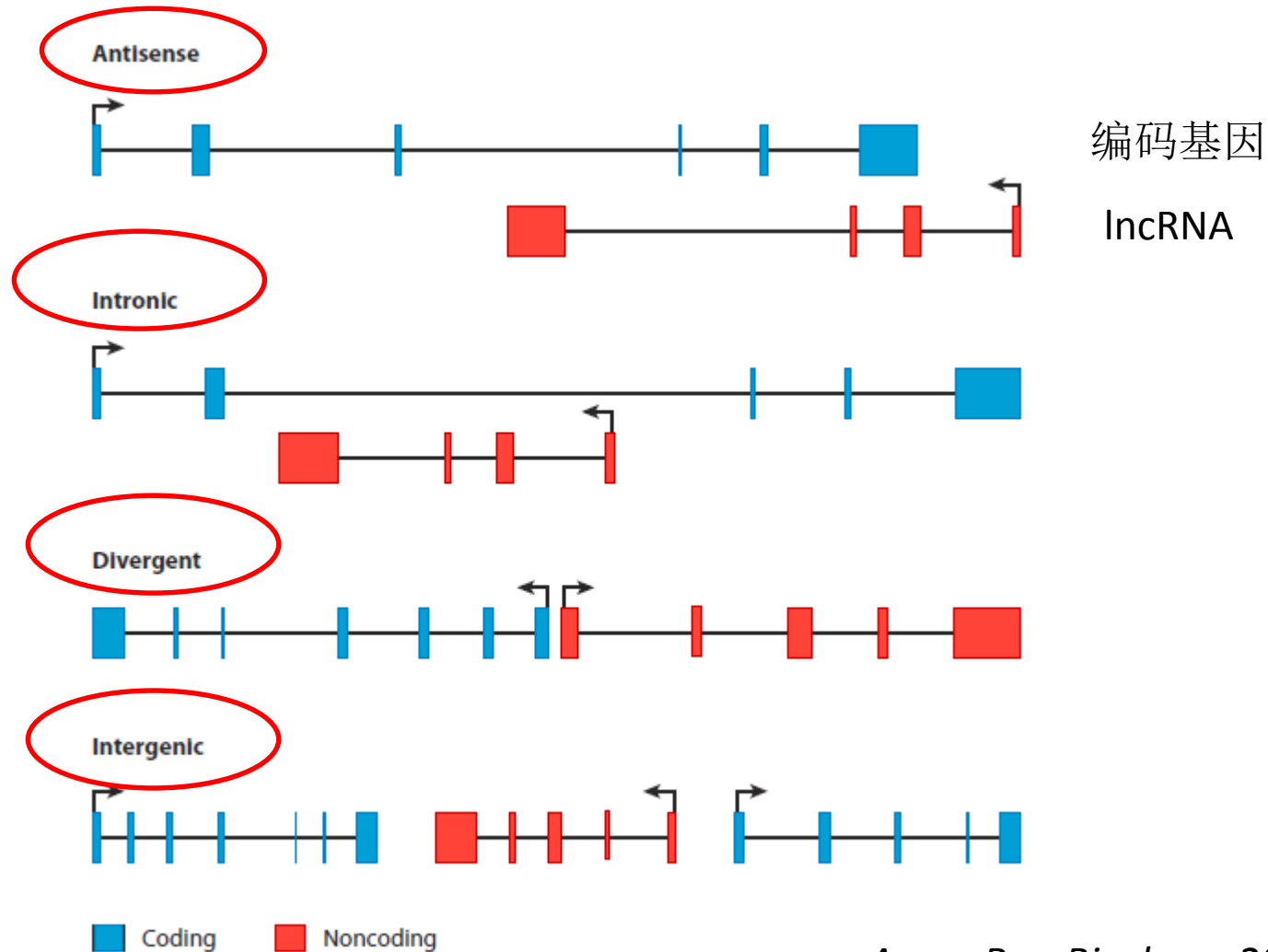
<http://www.encodegenes.org/stats/current.html>

- 据ENCODE (Encyclopedia of DNA Elements)项目统计，人类基因组中约75%区域都可以转录。其中1%为编码蛋白外显子，40%为蛋白编码基因（外显子+内含子）转录区域，剩下绝大多数转录区域产生lncRNA。

Nature 2012. 489:101; *Nature* 2012. 489:57

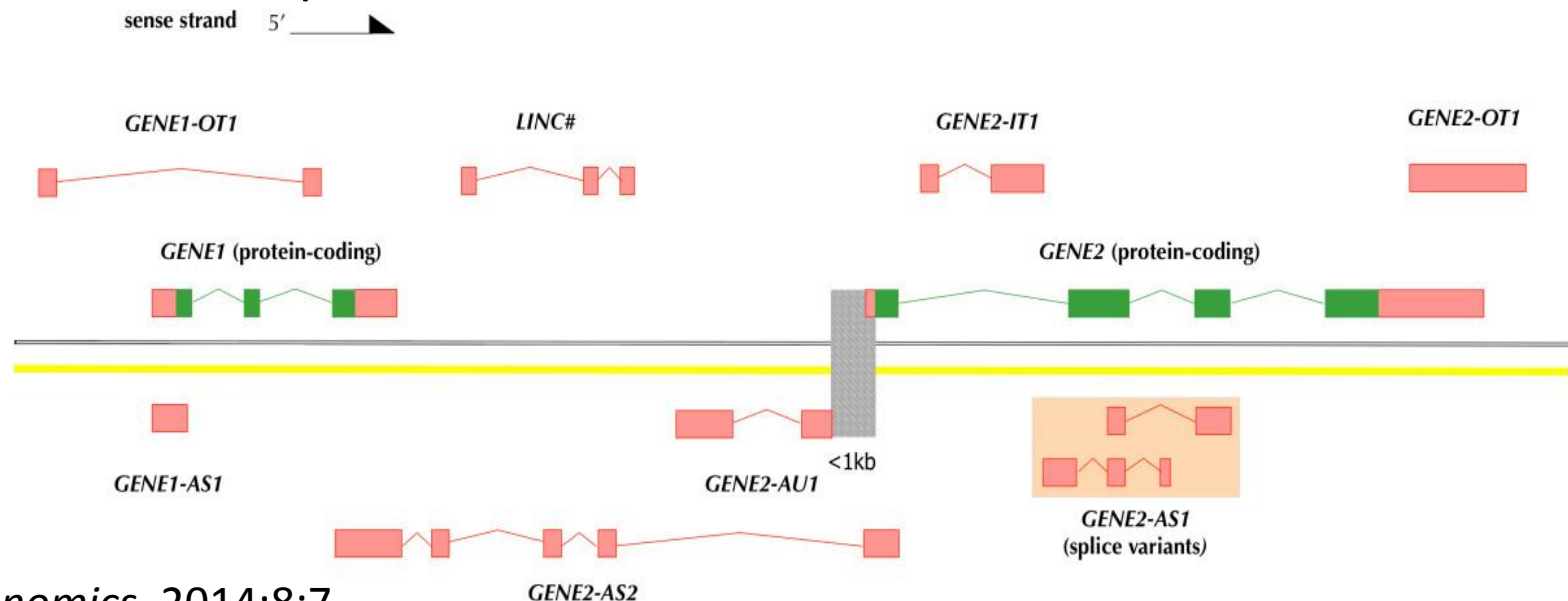
lncRNA分类

- lncRNA通常可以根据与邻近编码基因的位置关系来分类



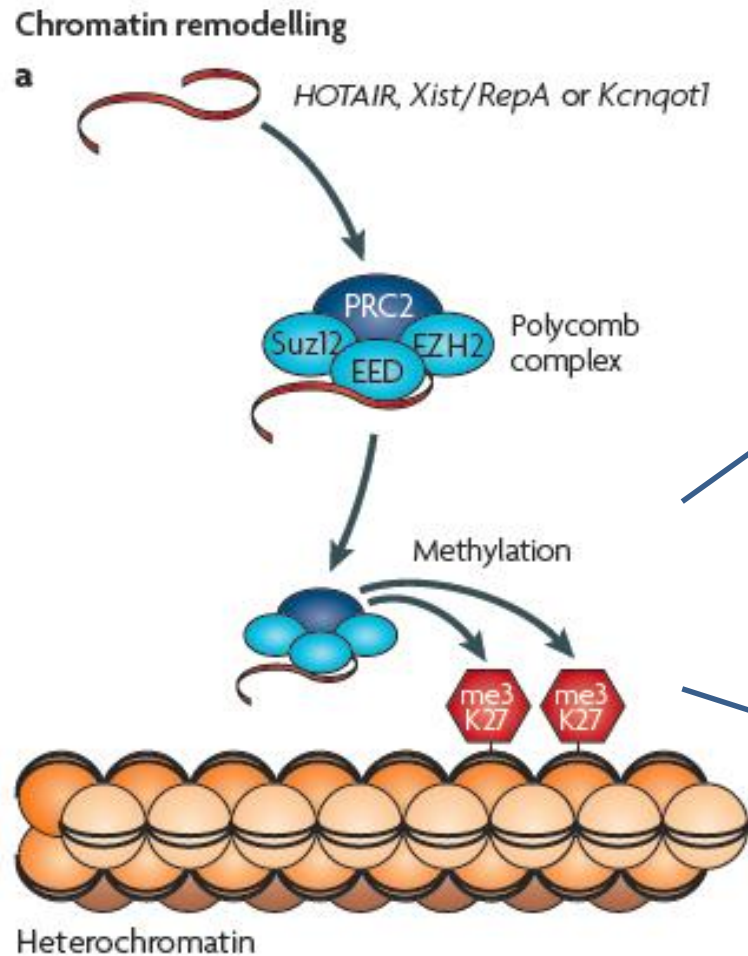
lncRNA命名

- 功能已知的lncRNA，根据功能命名
 - (e.g.) *Xist*: X inactive-specific transcript
- 功能未知的lncRNA，根据邻近编码基因命名
 - Antisense: '-AS' suffix, (e.g. *BOK-AS1*)
 - Intronic: '-IT' suffix, (e.g. *MAGI2-IT1*)
 - Sense overlapping: '-OT' suffix (e.g. *SOX2-OT*)
 - Intergenic: *LINC#*, (e.g. *LINC00152*)
 - Antisense upstream: '-AU' suffix



lncRNA调控机制

- 表观遗传调控—招募染色质修饰复合物至特定位置

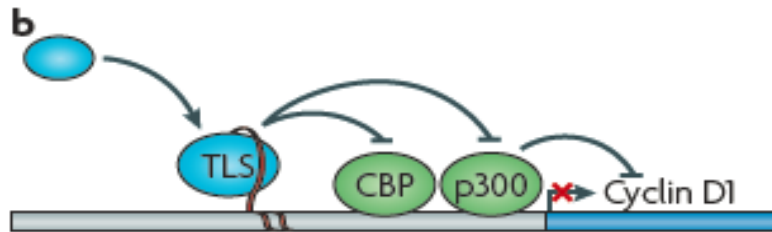


在剂量补偿效应中，*Xist*编码的lncRNA招募PRC2复合物导致X染色体失活

*HOTAIR*通过招募 PRC2复合物沉默 *HOXD*基因表达

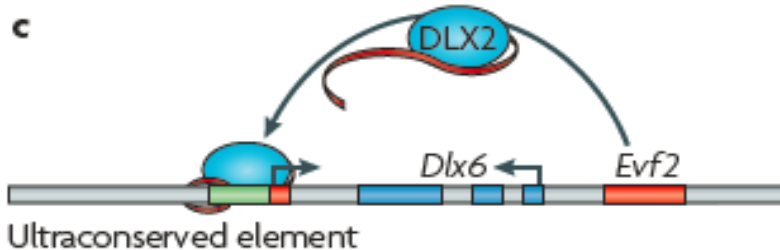
lncRNA调控机制

Transcriptional control

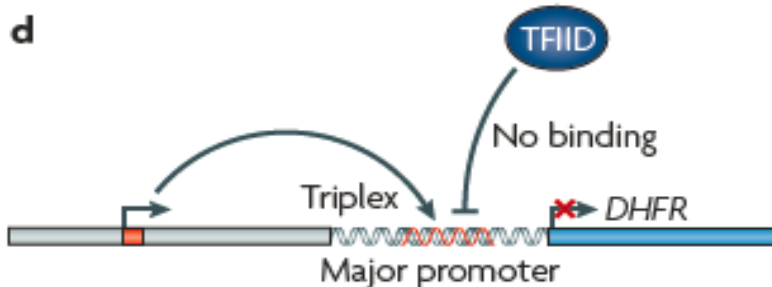


- 转录调控—通过结合调控蛋白、调控元件等多种机制影响下游基因表达

DNA损伤时，Cyclin D1相关lncRNA结合TLS蛋白抑制下游Cyclin D1表达



lncRNA *Evf2*结合转录因子DLX2诱导附近基因表达

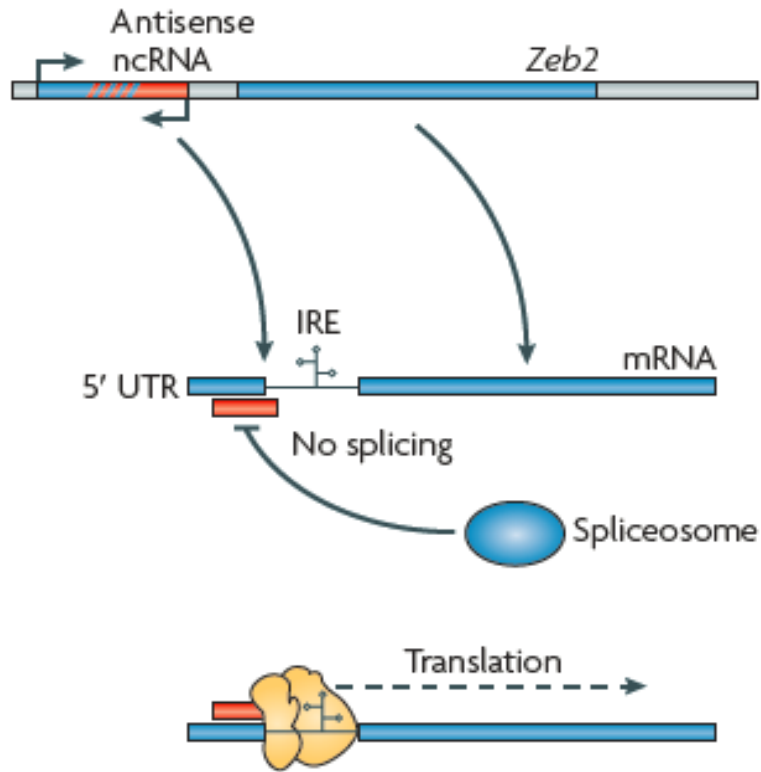


上游lncRNA通过结合 DHFR启动子来阻止转录因子TFIID结合

lncRNA调控机制

Post-transcriptional processing

e



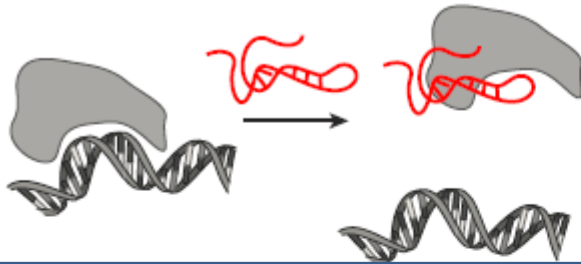
- 转录后调控—lncRNA与mRNA结合影响mRNA剪切、翻译和降解

Zeb2-antisense lncRNA通过与Zeb2 mRNA 结合，抑制其内含子剪切

作用机制模型

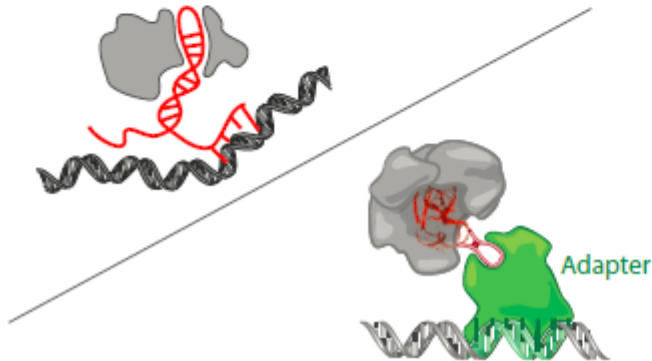
诱饵 — 抑制调控蛋白与DNA结合

a Decoy



向导 — 通过RNA-DNA或RNA-蛋白质相互作用招募调控蛋白

c Guide



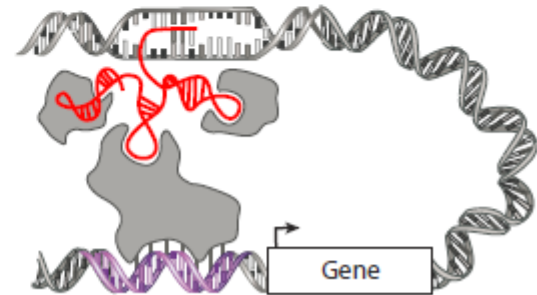
支架 — 将多个蛋白质结合为复合物

b Scaffold

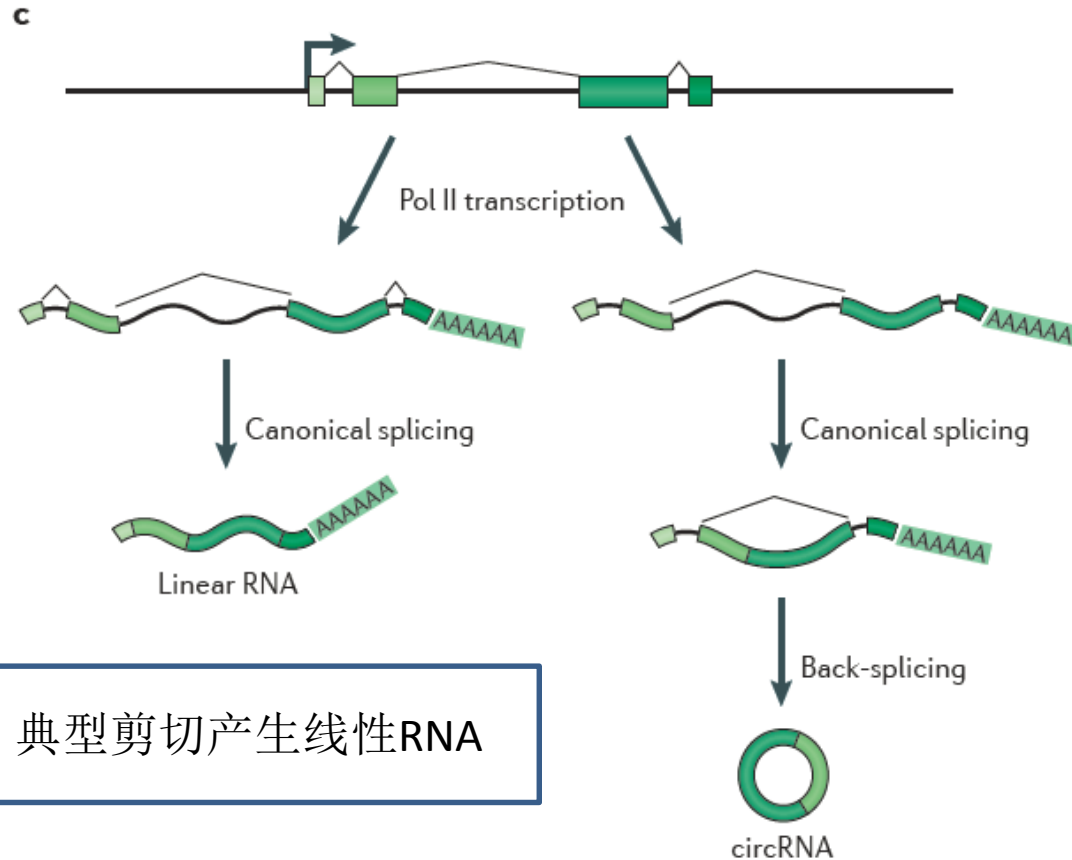


增强子 — 介导染色质成环，形成转录增强子结构

d Enhancer

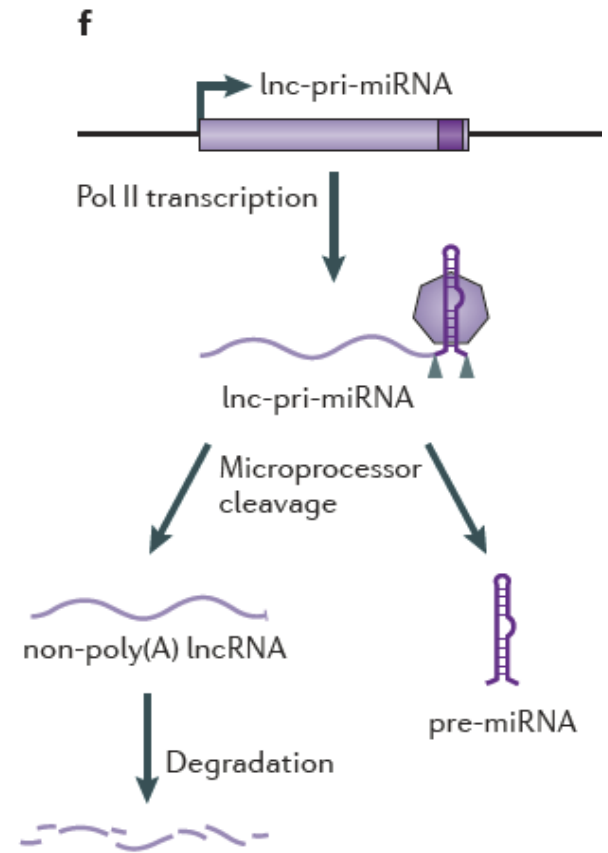


lncRNA作为其他ncRNA前体



典型剪切产生线性RNA

反向剪切方式产生环状RNA



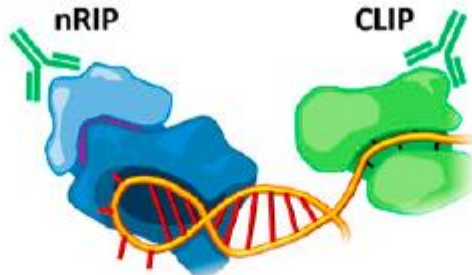
作为miRNA 前体

lncRNA作用鉴定方法

蛋白质-lncRNA相互作用:

- RNA免疫共沉淀 (RIP)
- 交联免疫共沉淀 (CLIP)

A PROTEIN INTERACTIONS

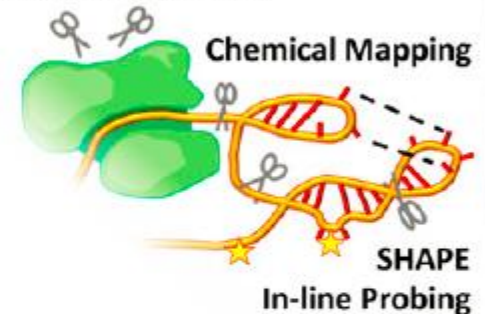


lncRNA高级结构捕获:

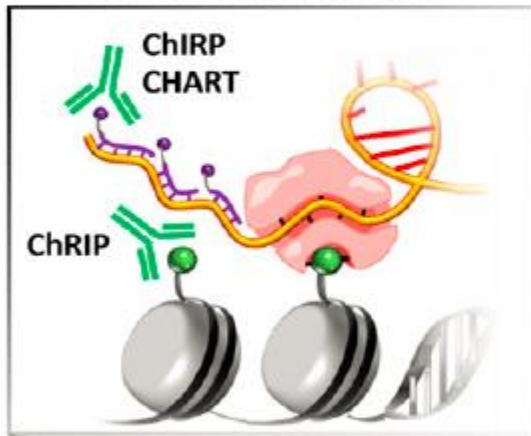
- 核酸酶剪切印迹(RNase footprint)
- 局部可变核苷酸捕获 (SHAPE)

C STRUCTURAL FEATURES

RNase Footprinting

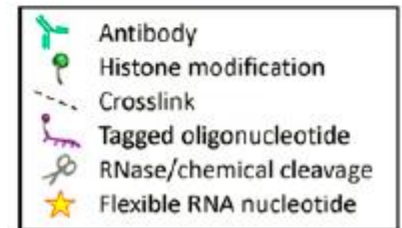


B DNA INTERACTIONS

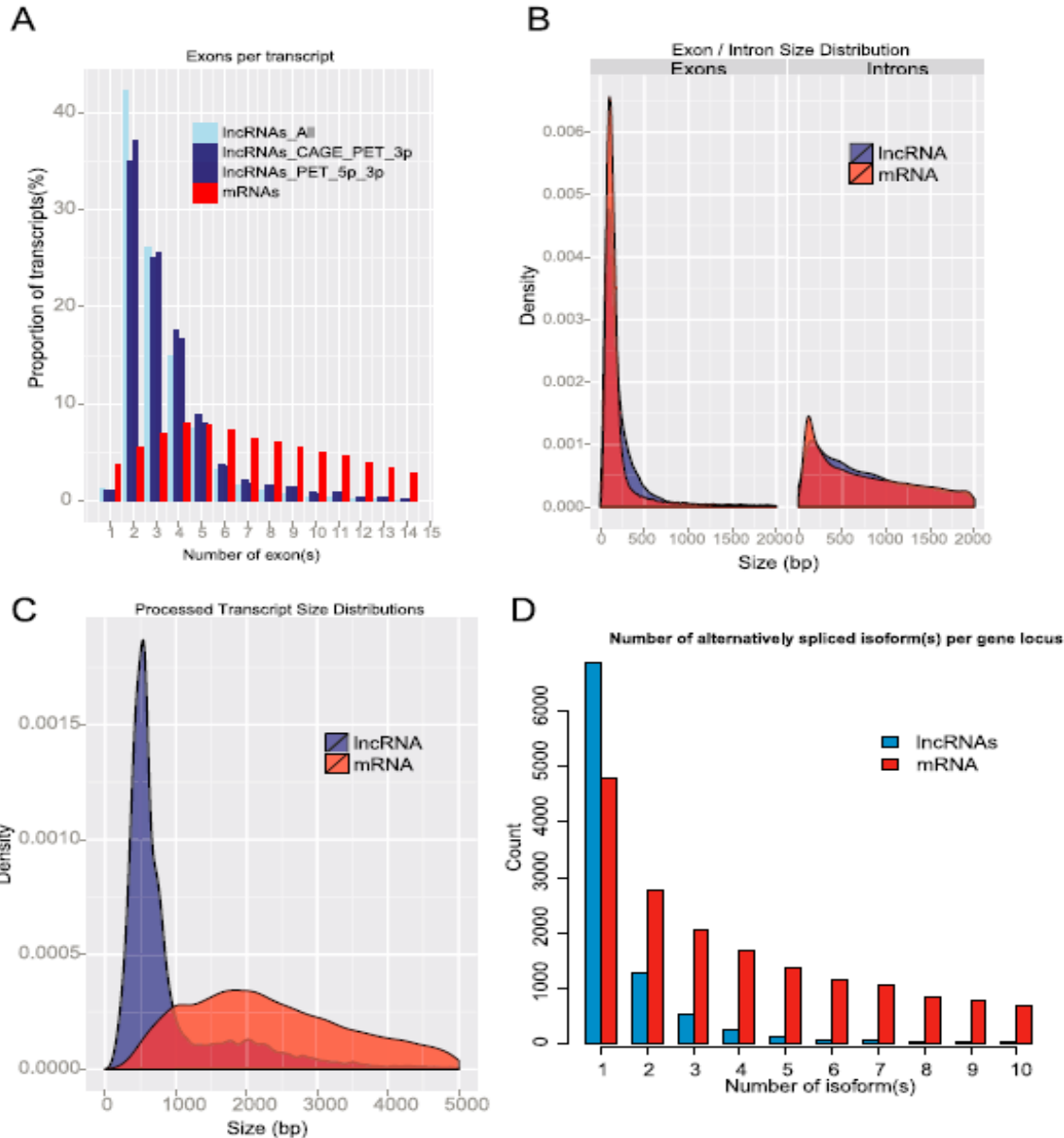


DNA-lncRNA相互作用:

- 染色质RNA免疫共沉淀 (ChRIP)
- 通过RNA纯化的染色质分离(ChIRP)
- RNA靶标捕获杂交 (CHART)

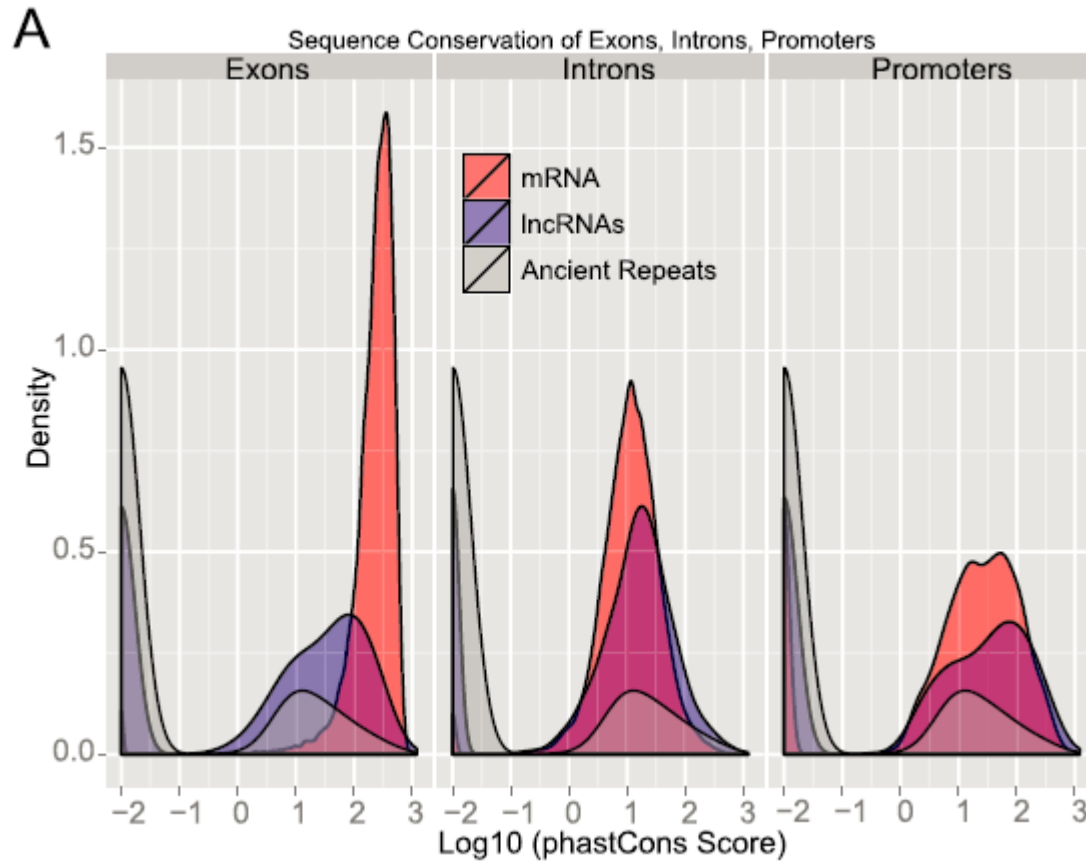


lncRNA基因组特征



- 绝大多数lncRNA含2-4个外显子，显著少于mRNA。
- lncRNA 外显子/内含子长度大于mRNA
- lncRNA 转录本长度低于mRNA
- lncRNA isoform类型显著少于mRNA。

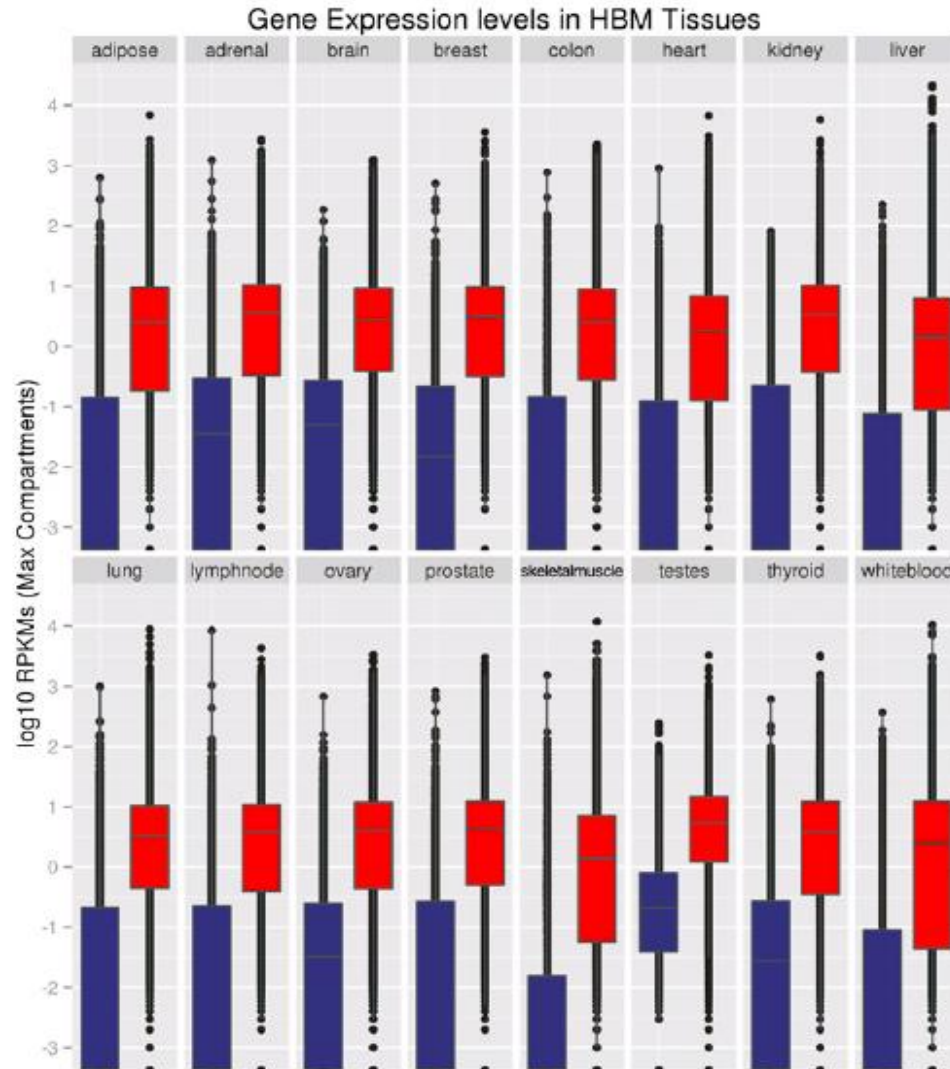
lncRNA基因组特征



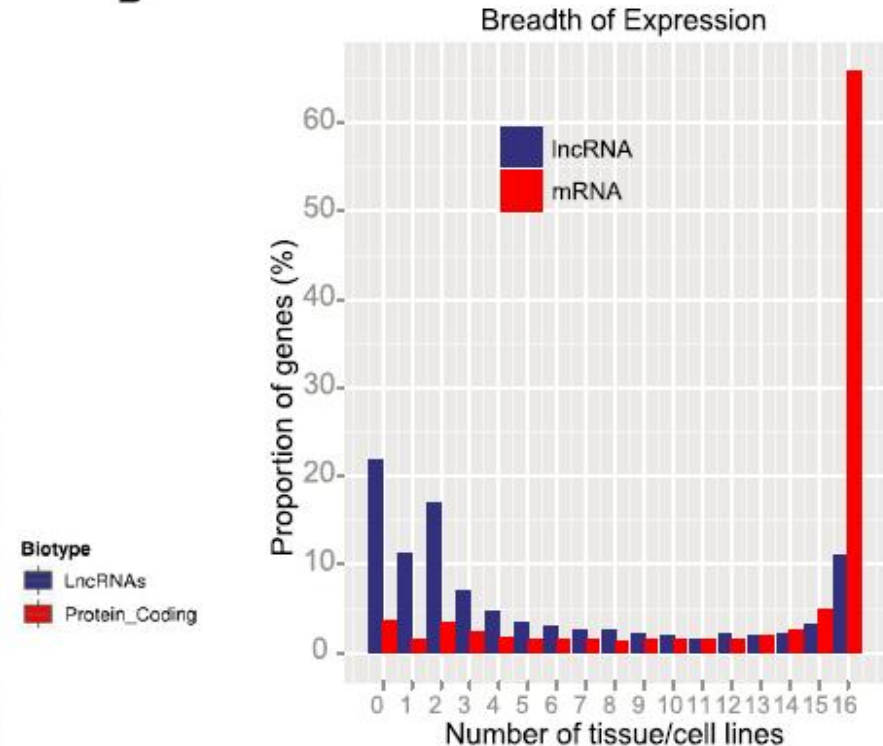
- lncRNA序列保守性显著低于mRNA（特别是外显子区域），但高于基因组重复序列。

lncRNA基因组特征

A



B



- HBM (Human Body Map)表达谱数据显示，lncRNA表达量显著低于mRNA，表达的组织的组织特异性显著高于mRNA。

提纲

- lncRNA的生物学背景
- lncRNA的计算鉴定
- lncRNA的功能预测

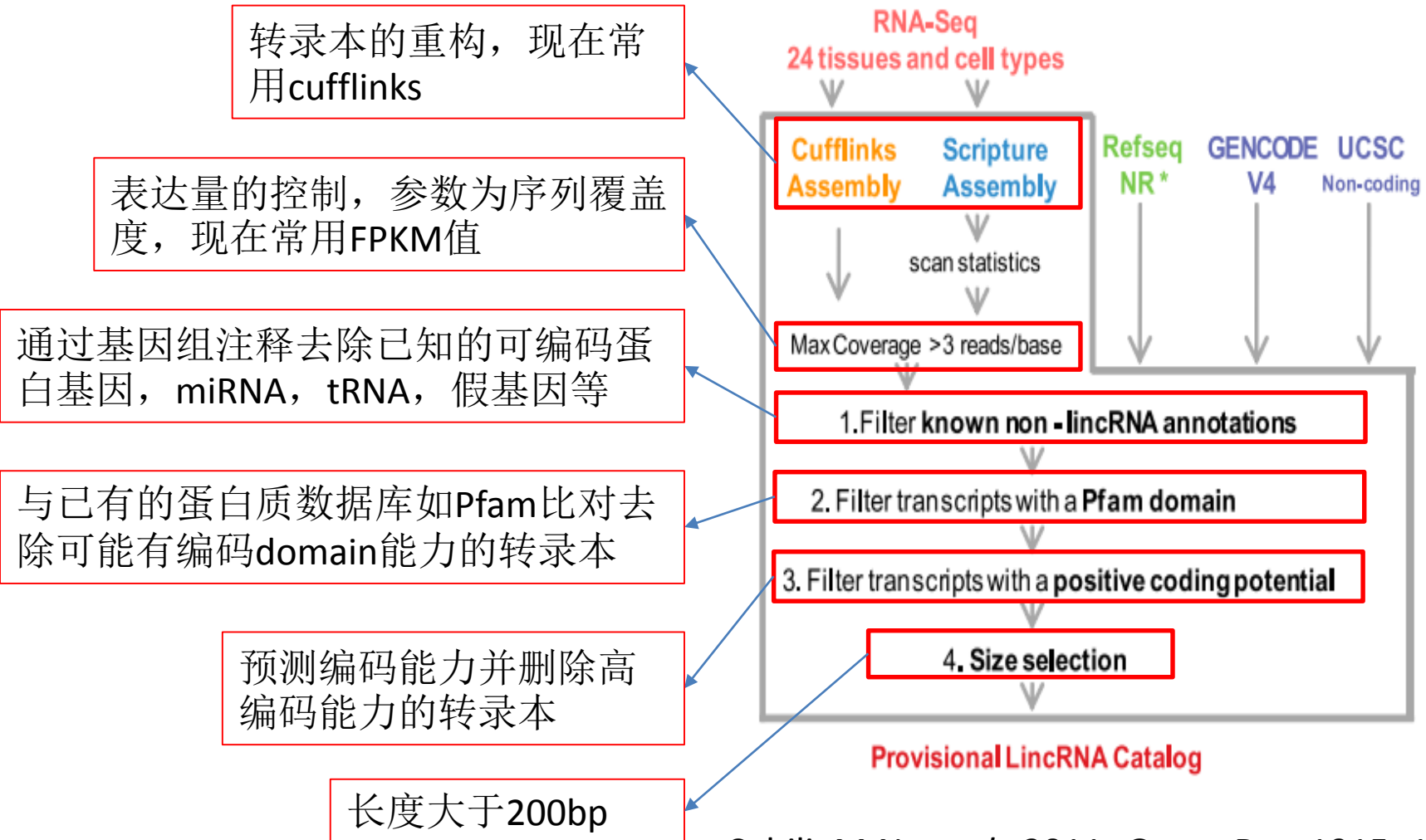
已知lncRNA鉴定流程

- 下载物种已知的参考ncRNA序列
 - Ensemble:
<http://uswest.ensembl.org/info/data/ftp/index.html>
- 选用下载的ncRNA序列为参考基因组
 - 序列比对（如Tophat）
 - 转录本重构（如Cufflinks）
- 成功比对上的转录本鉴定为已知的lncRNA

lncRNA常用数据库

- NONCODE: <http://www.noncode.org/>
 - 中科院计算所构建的ncRNA综合性数据库
 - 包含16种常见模式生物
 - 数据来源为文献挖掘(PubMed)和序列数据库(Ensembl, RefSeq, lncRNAdb, GENCODE)
 - 提供基因组, 表达谱, 进化保守性, 预测的功能和疾病关联信息
- lncRNAdb: <http://lncrnadb.org/>
 - 已知功能的lncRNA参考数据库
 - 提供人工校对的结构, 表达, 细胞定位和功能证据
- GENCODE: <http://www.gencodegenes.org/>
 - ENCODE项目发布的高质量参考基因注释
 - 包括protein-coding, lncRNAs, pseudogenes, small RNAs
 - 目前只针对人类和小鼠

RNA-Seq来源的新lncRNA鉴定步骤

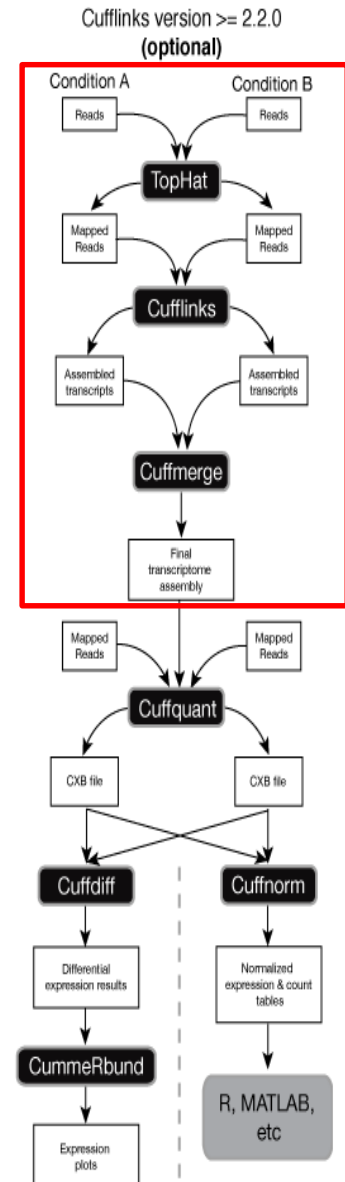


基因组比对

- 可在此之前添加质量控制
- 目的：原始数据与参考基因组比对，得到每一条序列在基因组上的相对位置
- 软件
 - Tophat: <https://ccb.jhu.edu/software/tophat/index.shtml>
 - Bowtie: <http://bowtie-bio.sourceforge.net/index.shtml>
- 基因组数据下载
 - Ensemble:
<http://uswest.ensembl.org/info/data/ftp/index.html>
 - UCSC: <http://hgdownload.soe.ucsc.edu/downloads.html>

转录本重构

- 目的：得到所有的转录本并检测可能的新转录本
- 软件
 - cufflinks: <http://cole-trapnell-lab.github.io/cufflinks/>
- 步骤
 - Cufflinks: 获取每个样本的所有拼接转录本
 - Cuffmerge: 对样本的结果进行整合, 得到文件 **merged.gtf**



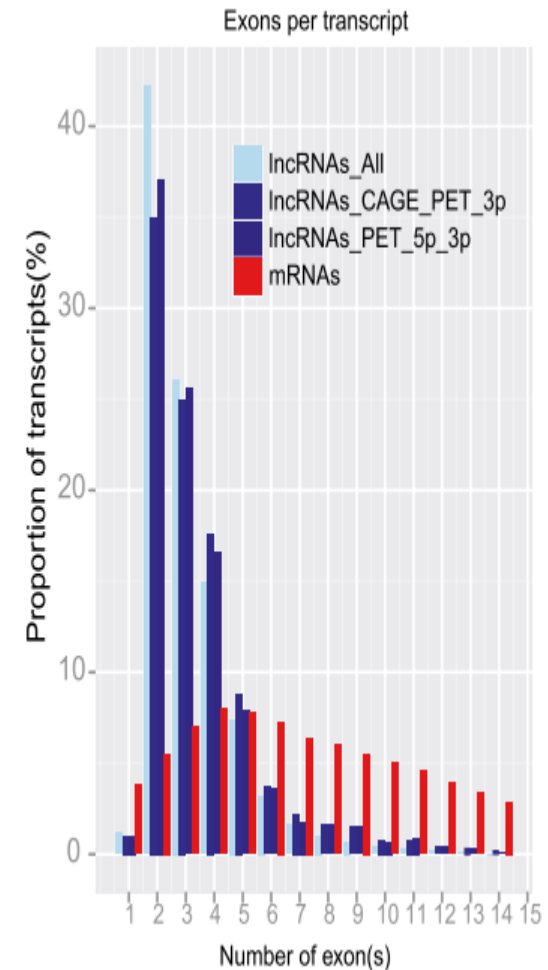
Exon个数的控制

- 目的：控制假阳性率
- 步骤

统计merged.gtf文件中每一条转录本的exon个数，根据实际的分析需求考虑去掉单exon的转录本

- 示例

右图为GENCODE数据库对人类LncRNA的exon个数分析比较



去除已知的非lncRNA

- 方法一：根据基因组的注释文件进行筛选
 - 标准：与注释mRNA, miRNA, tRNA, 假基因等没有外显子序列上的重合。
- 方法二：利用merged.gtf文件自带的classcode特征进行筛选
 - 去除显著被标记为mRNA的序列。
 - 如：classcode “=” 表示与mRNA序列完全匹配；
classcode “c” 表示与mRNA序列部分匹配，均不可能是lncRNA。
 - 仅仅挑选可能被标记为非编码RNA的序列。
 - 如：classcode “i” 表示完全落在内含子区域的转录本；
classcode “u” 表示落在基因间区的转录本，有可能是lncRNA。

与已知的蛋白质数据库如Pfam比较

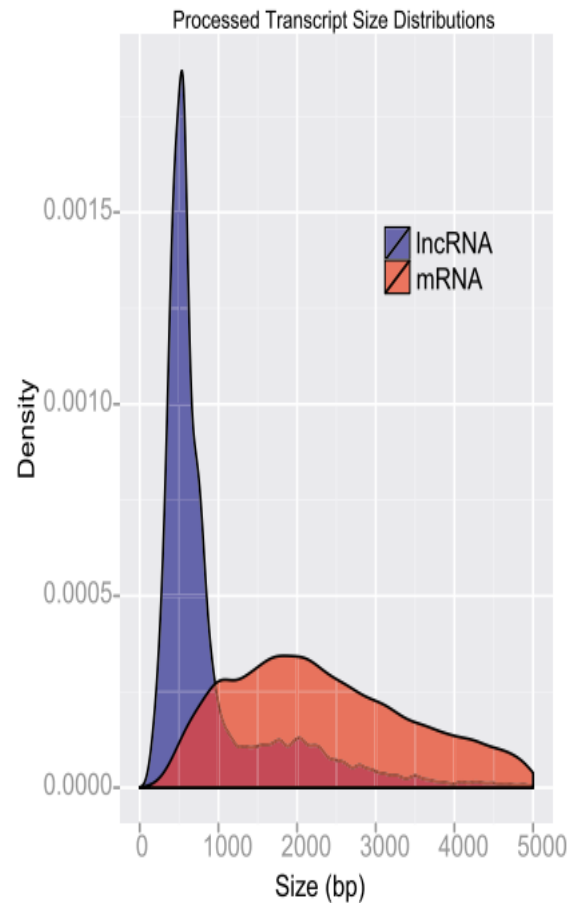
- 目的：去除可能有编码蛋白质domain能力的转录本
- 方法：
 - 翻译出每条转录本所有可能对应的氨基酸序列：ORF Finder, TransDecoder
 - 将每个转录本所有可能的氨基酸序列与Pfam数据库进行比对，有任意一条氨基酸序列比对上数据库则删除该转录本：HMMER

编码能力的评估

- 目的：对每条转录本的编码能力进行评估，去除编码能力过强的转录本
- 原理：各种评估编码能力的软件侧重点均有所不同
 - CPC和CPAT是以现有的开放阅读框（ORF）知识和蛋白数据库为依据
 - PhyloCSF以进化模型为依据
 - CNCI则不依靠现有注释，以相邻核苷酸三联体（ANT）为依据

长度控制

- 标准：Fasta格式的碱基序列长度一般要大于200bp

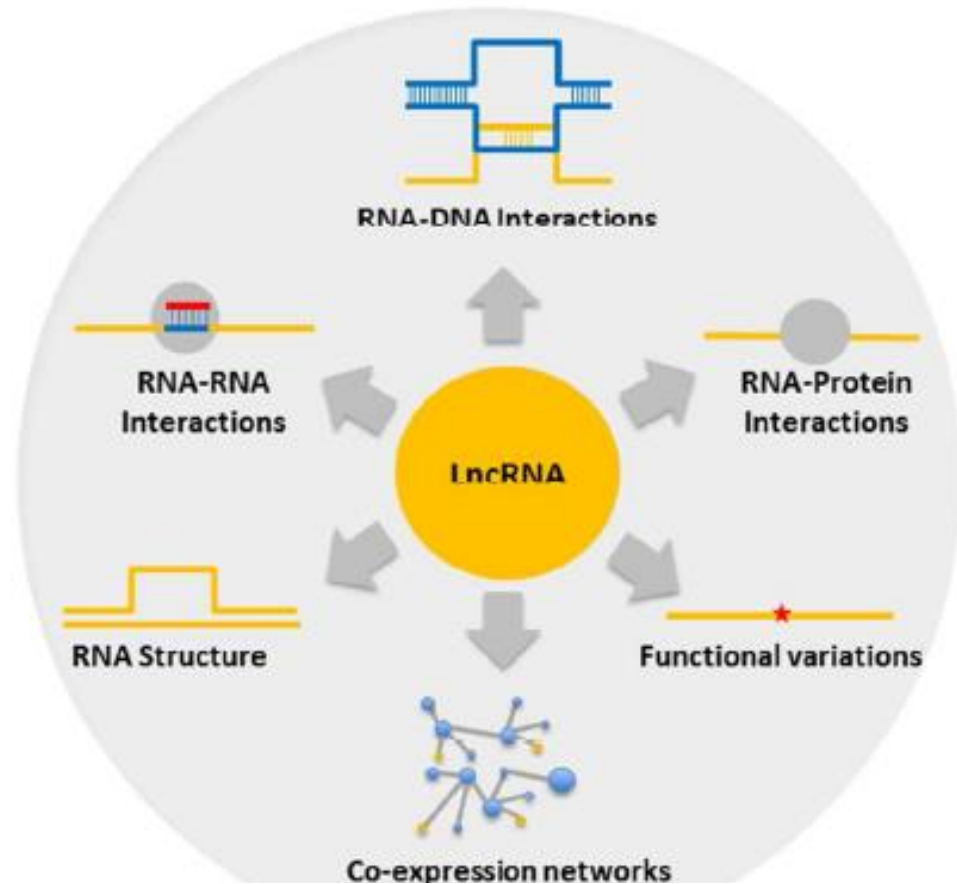


提纲

- lncRNA的生物学背景
- lncRNA的计算鉴定
- lncRNA的功能预测

功能预测的主要方法

- 共表达方法
- Interaction方法
 - RNA-DNA
 - RNA-RNA
 - RNA-protein
- lncRNA结构
- lncRNA功能变异



Review

Computational approaches towards understanding human long noncoding RNA biology

Saakshi Jalali^{1,2,\$}, Shruti Kapoor^{1,2,\$}, Ambily Sivadas^{1,\$}, Deeksha Bhartiya^{1,2} and Vinod Scaria^{1,2,*}

¹GN Ramachandran Knowledge Center for Genome Informatics, CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Mathura Road, Delhi 110020, India

²Academy of Scientific & Innovative Research (AcSIR), 2 Rafi Marg, Anusandhan Bhawan, New Delhi 110001, India

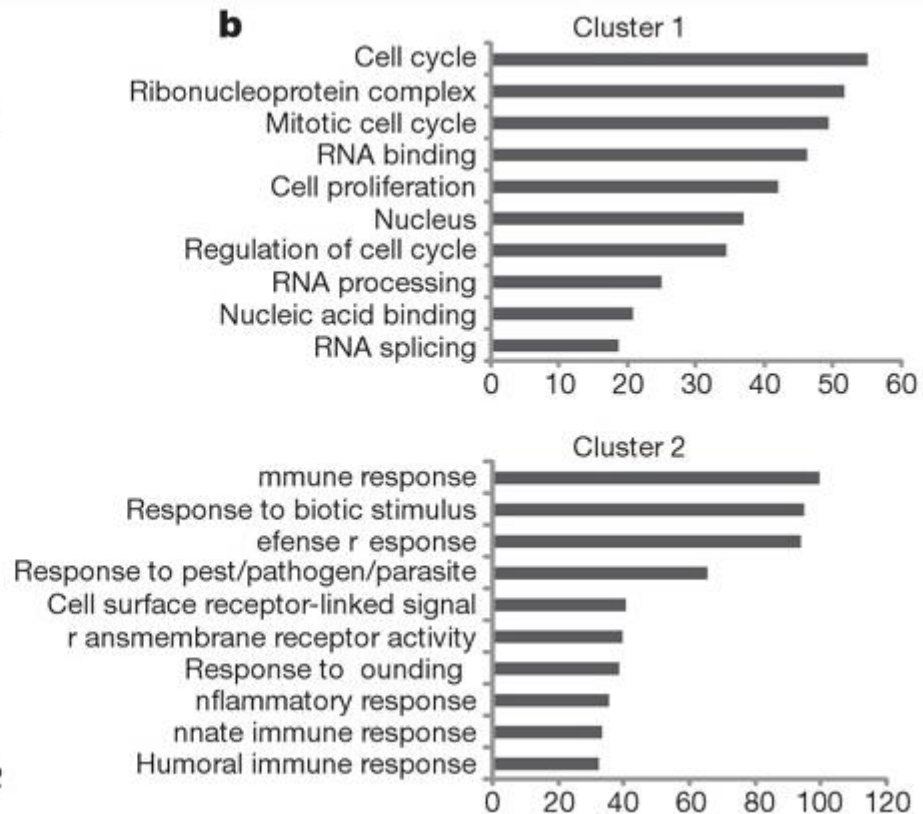
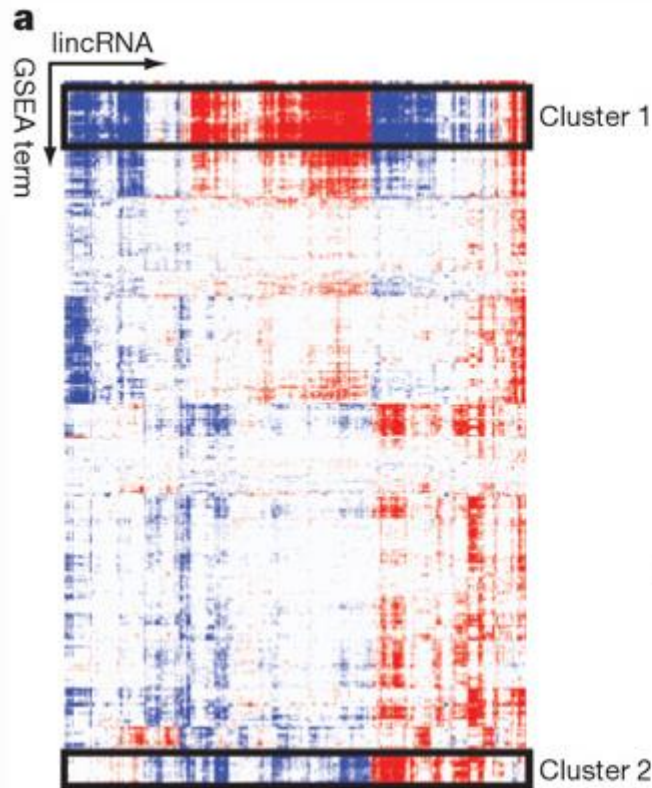
Associate Editor: Dr. Jonathan Wren

基于基因组位置的功能预测

- 基本假设：lncRNA对其附近的编码基因有cis的调控作用，可能行使共同的生物学功能
 1. 查找lncRNA附近（2kb）的基因
 2. 对这些基因进行功能分析，如GO，KEGG富集分析
- 该方法曾在lncRNA功能分析初试阶段使用，目前较少单独使用

基于共表达的功能预测

- 目前最流行最常用方法
- 主要思想: guilt-by association
- 假设: 若lncRNA与编码基因具有相似表达模式, 其有相似的功能



nature

ERS

基于共表达的功能预测

3864–3878 *Nucleic Acids Research*, 2011, Vol. 39, No. 9
doi:10.1093/nar/gkq1348

Published online 18 January 2011

Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network

1 构建coding-non-coding 共表达网络

2 分析lncRNA功能

hub-based方法：只有网络中直接coding gene邻居数目大于10的lncRNA进行分析，对其coding gene邻居进行GO富集

基于相互作用的功能预测

lncRNA的生物学功能可以通过与其互作的其他生物分子来进行研究

- RNA-protein interactions
- RNA-DNA interactions
- RNA-RNA interactions

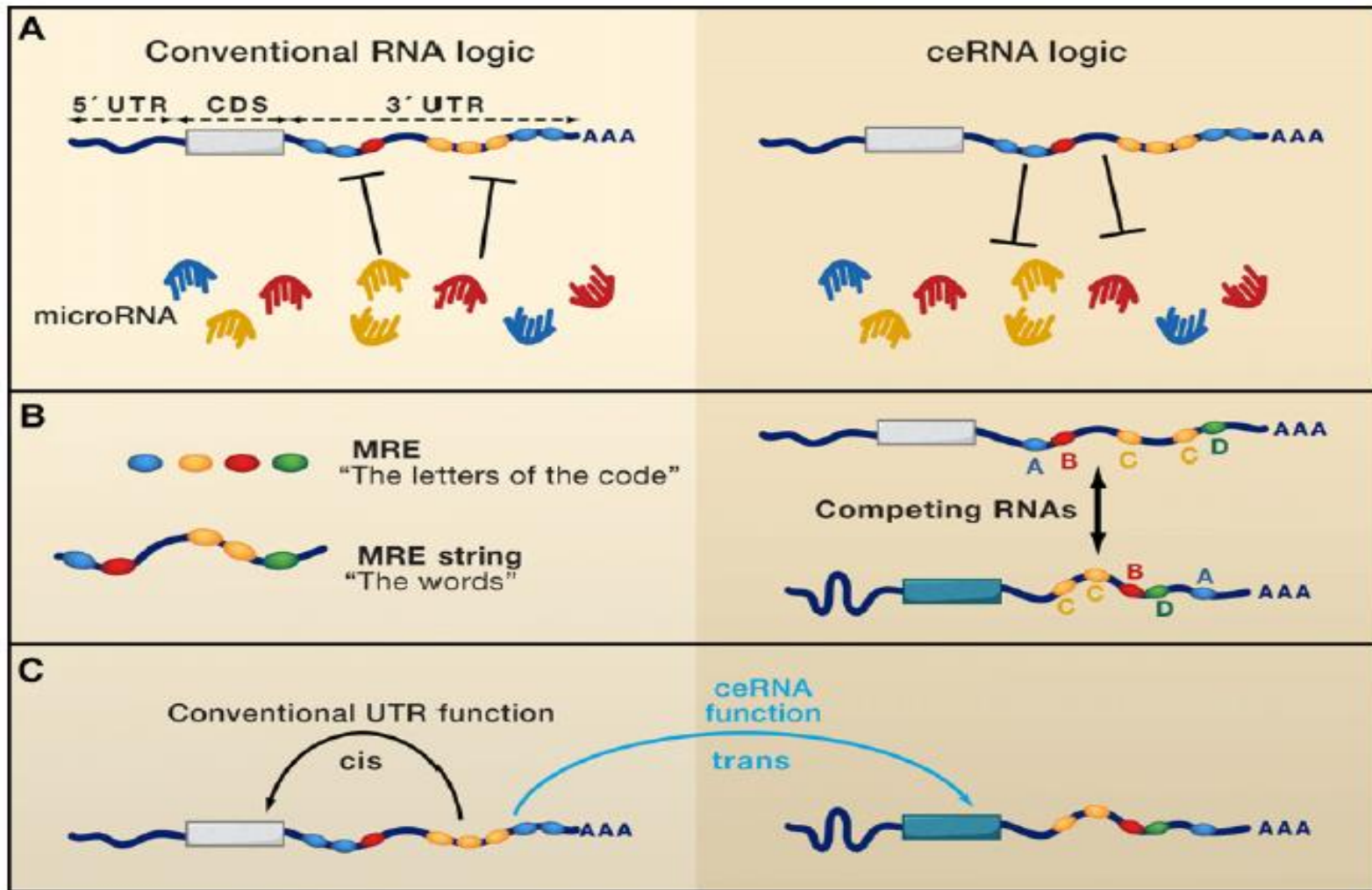
lncRNA-蛋白质相互作用预测

- 预测方法主要基于序列、结构、以及理化性质
 - RPIseq (Muppirala et al., 2011). 主要基于序列信息
<http://pridb.gdcb.iastate.edu/RPISeq/>.
 - CatRAPID (Agostini et al., 2013) 基于序列二级结构、氢键绑定及范德华力
<http://s.tartagliablab.com/catrapid/omics>
 - lncPro (Lu et al., 2013) 也是基于序列二级结构、氢键绑定及范德华力
<http://cmbi.bjmu.edu.cn/lncpro>

lncRNA-DNA相互作用预测

- lncRNA 能够形成与DNA双链形成DNA-RNA三链结构。
 - DHFR上游的lncRNA转录本能够通过通过与DHFR启动子区域结合形成DNA-RNA三链结构抑制下游编码基因(Blume et al., 2003; Martianov et al., 2007)。
- 计算方法
 - Triplexator (Buske et al., 2012)

lncRNA-miRNA互作及ceRNA 理论



Cell 146,
August 5,
2011

目前广泛认为lncRNA能够作为海绵（ceRNA）来调节miRNA的功能
lncRNA能够和mRNA同时竞争miRNA来调节mRNA的表达

一个预测lncRNA-miRNA互作以及ceRNA关系的网站



CLIP-Seq, Pan-Cancer. Portal, Visualize, Analyze, Discover.



Home miRNA-lncRNA miRNA-mRNA ceRNA Network Pan-Cancer Protein-RNA Function Prediction Download Contact Us

Welcome to starBase

starBase is designed for decoding **Pan-Cancer** and **Interaction Networks** of lncRNAs, miRNAs, competing endogenous RNAs(ceRNAs), RNA-binding proteins (RBPs) and mRNAs from large-scale CLIP-Seq (HITS-CLIP, PAR-CLIP, iCLIP, CLASH) data and **tumor samples (14 cancer types, >6000 samples)**.

starBase is also developed for deciphering **Protein-RNA** and **miRNA-target interactions**, such as protein-lncRNA, protein-sncRNA, protein-mRNA, protein-pseudogene, miRNA-lncRNA, miRNA-mRNA, miRNA-circRNA, miRNA-pseudogene, miRNA-sncRNA interactions and ceRNA networks from **108 CLIP-Seq (HITS-CLIP, PAR-CLIP, iCLIP, CLASH)** datasets.

starBase provides **miRFunction** and **ceRNAFunction** web tools to **predict the function** of ncRNAs (miRNAs, lncRNAs, pseduogenes) and protein-coding genes from the **miRNA-mediated (ceRNA) regulatory networks**.

How to Cite: **NEW** [Li et al. Nucleic Acids Res. 2014](#) & [Yang et al. Nucleic Acids Res. 2011](#).

What's New

starBase v2.0 release at Sept. 2013

- **NEW** [Pan-Cancer Analysis](#)
- [starBase v2.0 paper in Nucleic Acids Res.](#)
- 14 Cancer Types
- > 6000 tumor samples
- 111 CLIP-Seq data

NEW [starBase Pan-Cancer Tutorial](#)
[starBase v2.0 Tutorial](#)

Data Sets

starBase contains **108** CLIP-Seq data from **37** studies. [Details](#)

<http://starbase.sysu.edu.cn/>

lncRNA结构预测

- lncRNA次级结构决定了其分子功能
 - 例如：lncRNA MALAT1 3' 末端有类似t-RNA的结构。 lncRNA *HOTAIR*有一个复杂的RNA结构motif, 可以作为蛋白复合物的功能性支架(scanfold)
- 目前预测方法
 - prediction of secondary structure from a single sequence
 - prediction of the consensus secondary structure for a set of aligned sequences

<http://rna.tbi.univie.ac.at/>

lncRNA上的SNP功能预测

- 尽管目前大多数研究表明lncRNA序列进化上不保守，然而其功能变异对lncRNA还是会产生一定的影响

1 SNP对lncRNA结构有影响

2 SNP对RNA-lncRNA 互作的影响

3 GWAS研究表明疾病相关SNP落在lncRNA上

- 工具

LincSNP <http://bioinfo.hrbmu.edu.cn/LincSNP>

IncSNP <http://bioinfo.life.hust.edu.cn/IncRNASNP/>

基因组关联分析方法

Table 1. Some of the lincRNA *cis*-eQTLs are disease-associated SNPs.

<i>Cis</i> -eQTL SNP	eQTL <i>P</i> on lincRNA	Proxies ($R^2 > 0.8$) associated with disease/trait	Chr	Trait/Disease	eQTL affected lincRNA	eQTL tissue
rs13278062	4.31×10^{-32}	rs13278062	8	Exudative age-related macular degeneration	XLOC_006742	Blood
rs11066054	4.09×10^{-11}	rs6490294	12	Mean platelet volume	XLOC_010202	Blood
rs206942	3.63×10^{-5}	rs206936	6	Body mass index	XLOC_005690	Blood
rs11065766	6.67×10^{-5}	rs10849915	12	Alcohol consumption	XLOC_009878	Blood
	6.67×10^{-5}	rs10774610	2	Drinking behavior		
rs1465541	1.84×10^{-4}	rs11684202	2	Coronary heart disease	XLOC_002026	Blood
rs12125055	1.84×10^{-4}	rs7542900	1	Type 2 diabetes	XLOC_000922	Blood
rs199439	8.25×10^{-6}	rs199515	17	Parkinson's disease	XLOC_012496	SAT
		rs415430	17	Parkinson's disease		SAT
		rs199533	17	Parkinson's disease		SAT
rs17767419	1.05×10^{-8}	rs17767419	16	Thyroid volume	XLOC_011797	SAT, VAT
		rs3813582	16	Thyroid function		SAT, VAT

Chr chromosome, SAT Saturated adipose tissue, VAT Visceral adipose tissue.
doi:10.1371/journal.pgen.1003201.t001

GWAS研究表明很多SNP落在lincRNA区域，该研究也发现了一些lincRNA *cis*-eQTLs是疾病相关的SNP。

Kumar,V. et al. (2013)

基于基因组位置上下文

通过邻居基因及miRNA与疾病的关系预测lncRNA与疾病的关系。

- 如果lncRNA基因组位置2kb内的coding基因或者miRNA与某种疾病相关，那么这个lncRNA有可能是该疾病lncRNA。

缺点：比较粗糙，预测准确性比较差

Nucleic Acids Res. 2013 Jan;41(Database issue):D983-6. doi: 10.1093/nar/gks1099. Epub 2012 Nov 21.

LncRNADisease: a database for long-non-coding RNA-associated diseases.

谢谢