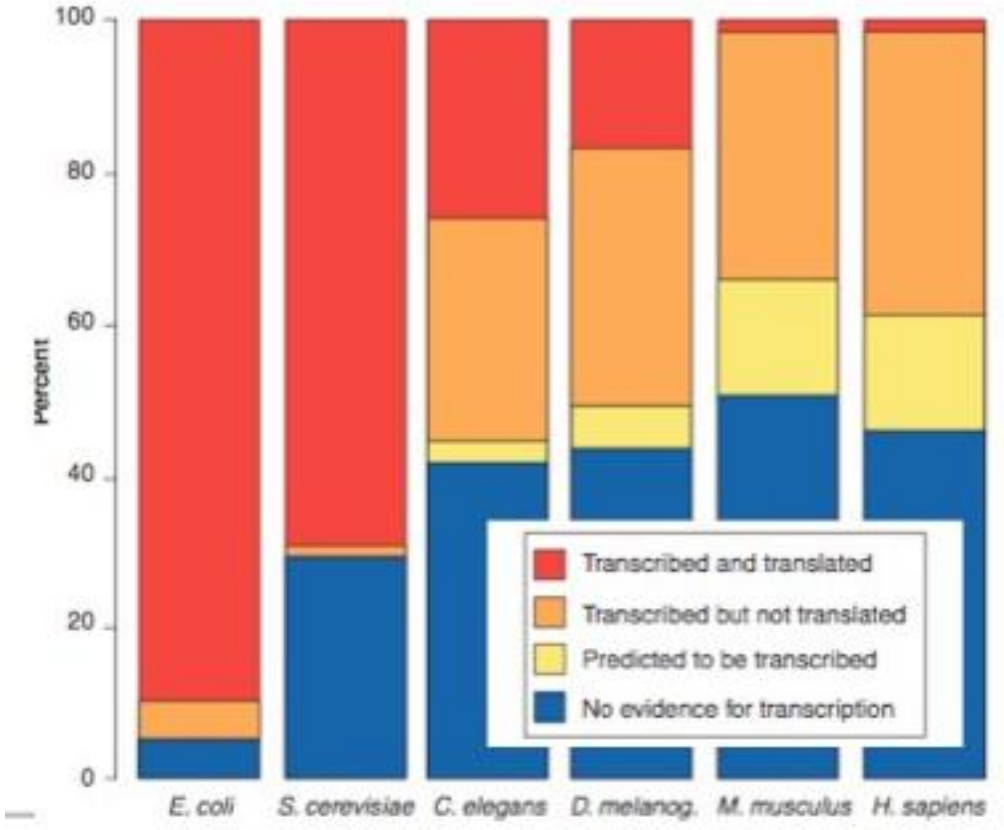
Long Noncoding RNAs

Functional genomics Non-coding genes

- The importance of noncoding regulatory is discovered relatively recently
- Very active area of biomedical research
- A **non-coding RNA** (ncRNA) is a functional **RNA** molecule that is transcribed from DNA but not translated into proteins
- Infrastructure/house-keeping
  - Ribosome , spliceosome, transfer RNAs
  - snoRNAs (small nucleolar RNAs) guide chemical modifications to above
- Regulatory
  - **microRNAs-**pair with complementary sequence in the UTR of coding genes to induce gene down-regulation or silencing. Processed from longer genes into small final products **Small interfering RNAs (siRNAs)**
  - **Long non-coding RNAs (lncRNAs)** lncRNAs are considered as non-protein coding transcripts >200 nt in length. The majority of non-coding RNAs belong to this group. Many RNAs in the group are treated by the cell as coding genes, they have exons, and are spliced.

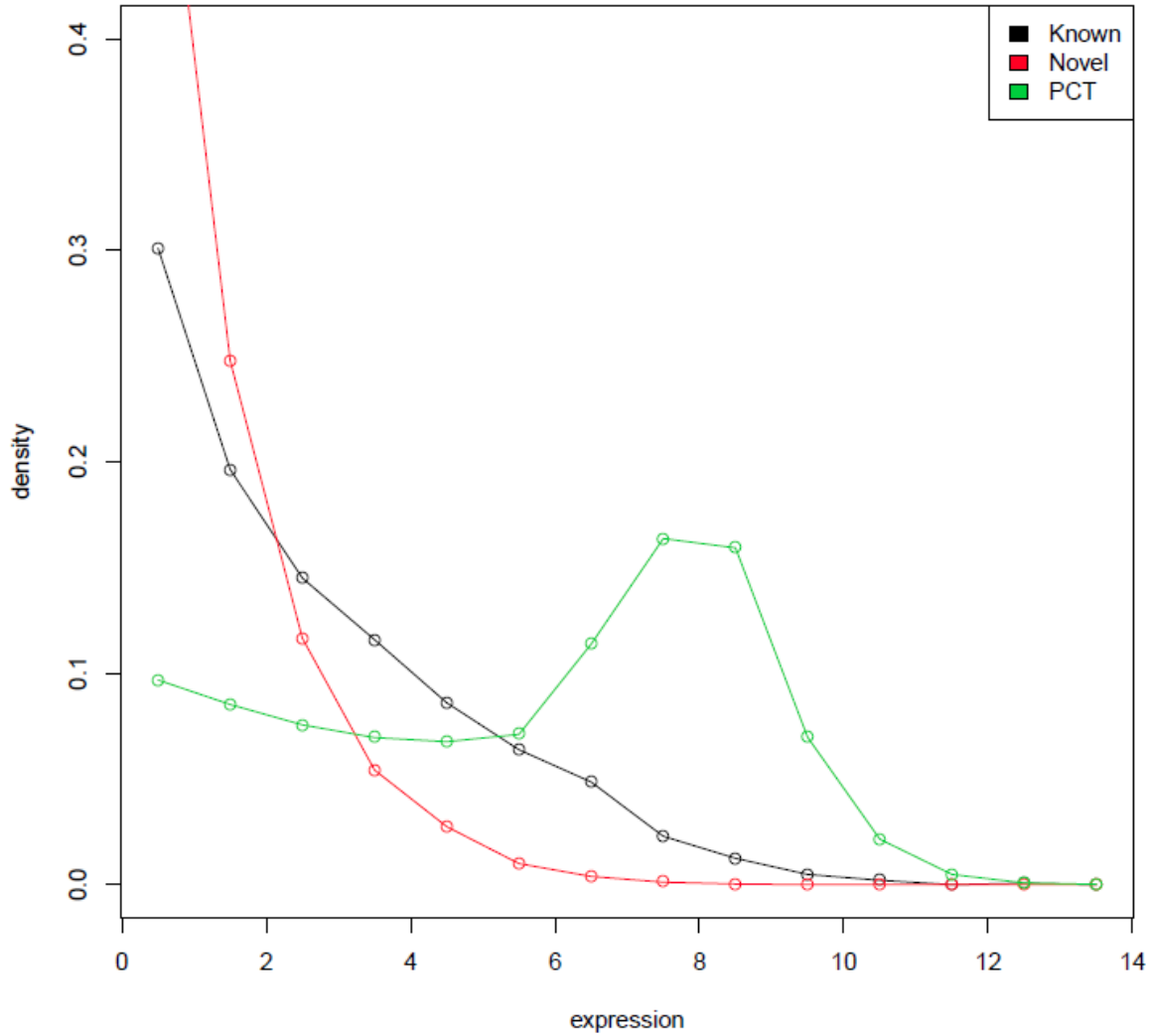
Long non-coding RNA

- 80% of the transcription in mammalian genomes is exclusively associated with long non-coding RNAs (lncRNAs)
- >2 (some >100) kb in length, spliced and could contain polyA signals
- No obvious open reading frame—can't get a long protein sequence without hitting a stop codon
- Mouse transcriptome (~180,000)
  - ~20,000 protein coding genes
  - ~160,000 lncRNAs



lncRNAs—lots of transcripts but typically small amounts

- PCT is protein coding transcript
- Known lncRNAs have been previously found in other datasets
- Novel- detected from de-novo assembly in this dataset
- Expression in log counts



Pervasive transcription

- 2% of the mammalian genome codes for amino acids in proteins.
- evidence over the past decade has suggested that the vast majority of the genome is transcribed, well beyond the boundaries of known genes -- **pervasive transcription**
- Functionality has to be demonstrated via a phenotype

OPEN ACCESS Freely available online PLoS BIOLOGY

Most "Dark Matter" Transcripts Are Associated With Known Genes

Harm van Bakel[1], Corey Nislow[1,2], Benjamin J. Blencowe[1,2], Timothy R. Hughes[1,2]*

1 Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada, 2 Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

OPEN ACCESS Freely available online PLoS BIOLOGY

Perspective

The Reality of Pervasive Transcription

Michael B. Clark[1], Paulo P. Amaral[1,3], Felix J. Schlesinger[2,3], Marcel E. Dinger[1], Ryan J. Taft[1], John L. Rinn[3], Chris P. Ponting[4], Peter F. Stadler[5], Kevin V. Morris[6], Antonin Morillon[7], Joel S. Rozowsky[8], Mark B. Gerstein[8], Claes Wahlestedt[9], Yoshihide Hayashizaki[10], Piero Carninci[10], Thomas R. Gingeras[2]*, John S. Mattick[1]*

Human Molecular Genetics, 2010, Vol. 19, Review Issue 2 **R162–R168**
doi:10.1093/hmg/ddq362
Advance Access published on August 25, 2010

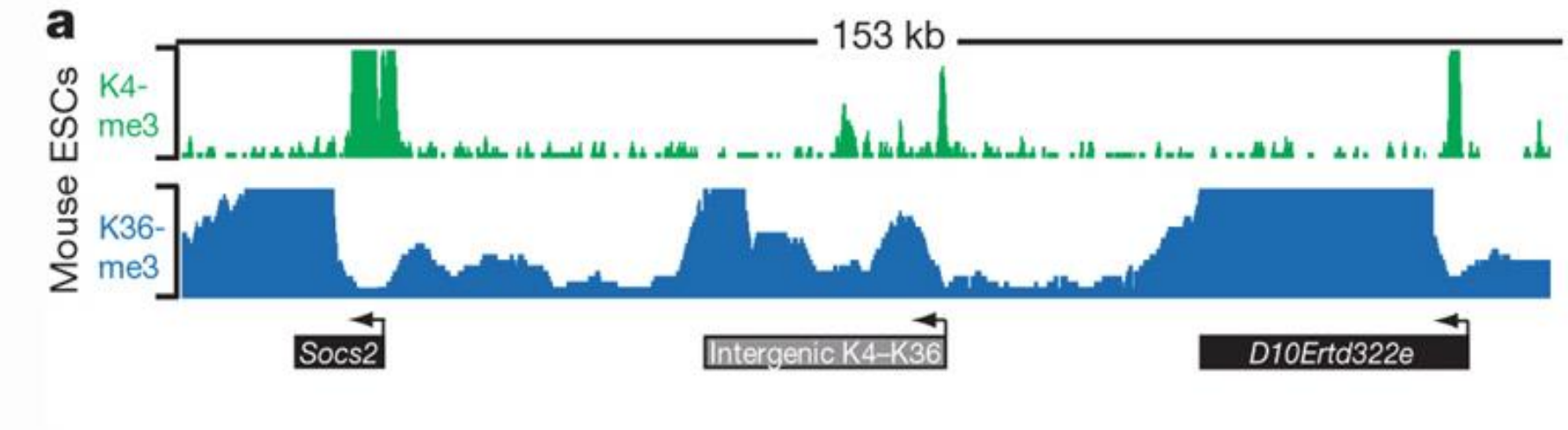
Transcribed dark matter: meaning or myth?

Chris P. Ponting* and T. Grant Belgard

MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, UK

Non-coding RNAs have an epigenetic profile similar to coding genes

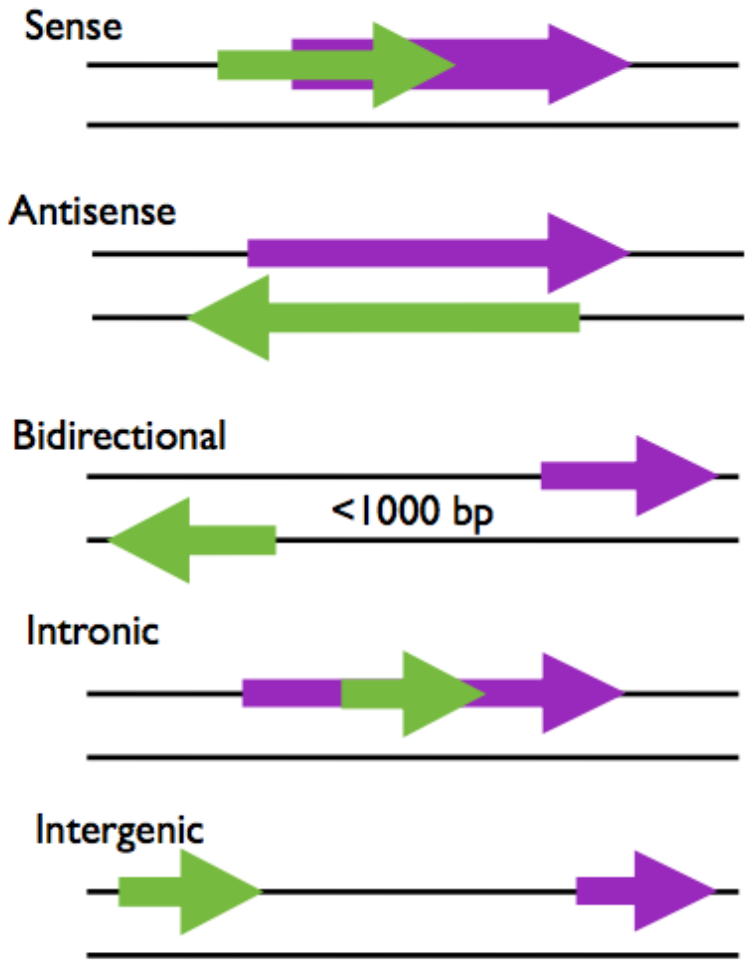
- Guttman M, et al. (2009) Chromatin Structure Reveals Over a Thousand Highly Conserved, Large Non-coding RNAs in Mammals. Nature
- Genes have stereotypic chromatin modifications that mark the promoter and gene body
- We can look for new genes just based on the epigenetic profile



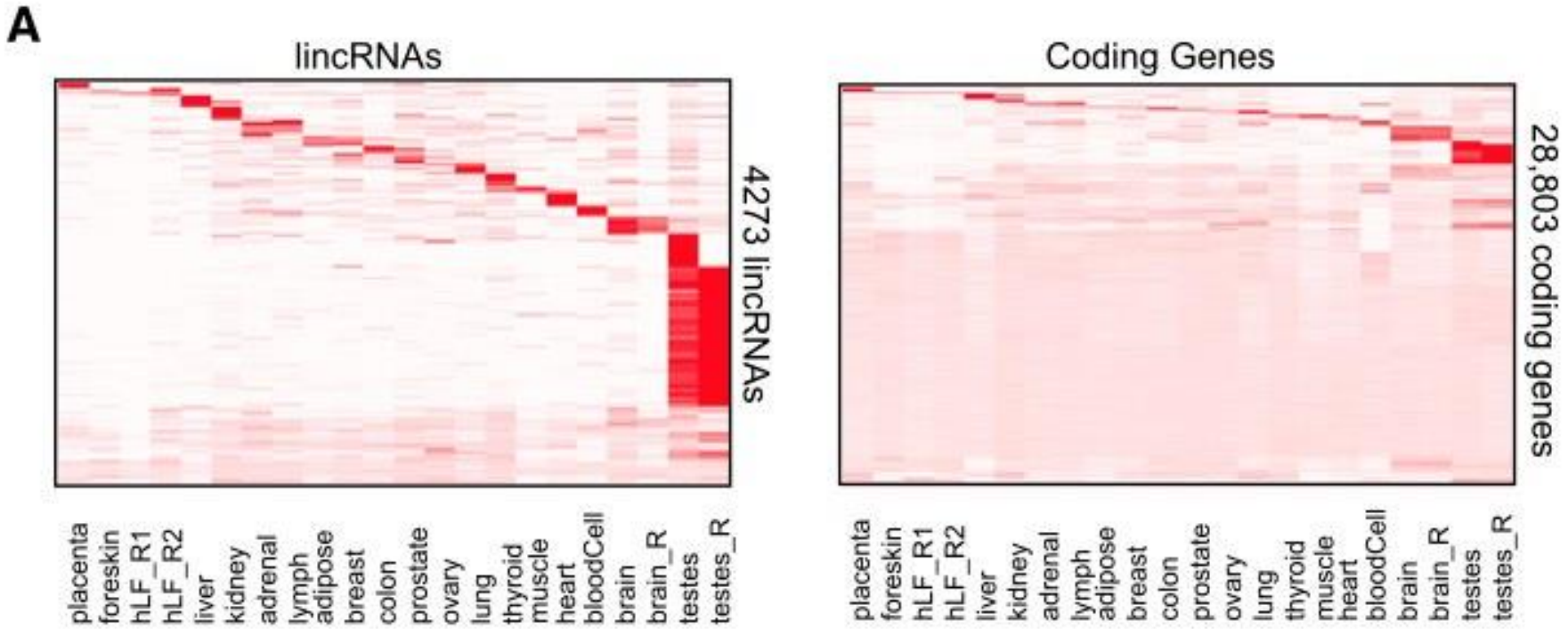
Coding gene proximity categorization

LncRNA classes

- miRNA host genes,
- snoRNA host genes
- **Divergent --lncRNAs that are transcribed in the opposite orientation of a coding gene with which they share a promoter**
- Intron
- Same strand (sense)
- **Intergenic**
- looking for lncRNAs from RNAseq data
  - May not have a polyA—need total RNA protocol
  - Can be transcribed opposite of another gene—stranded protocol very helpful here



Cell and tissue specific expression



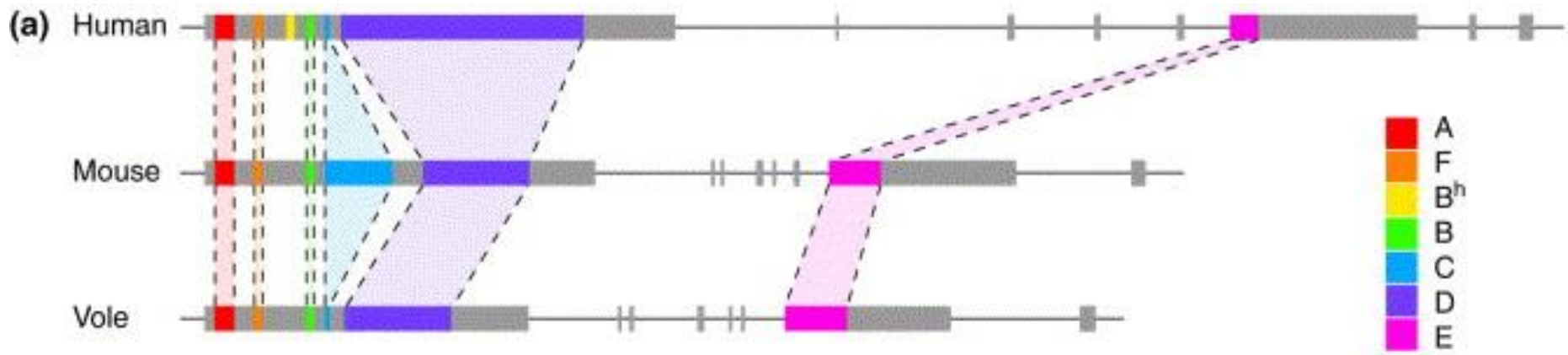
Cabili MN. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011 Sep 15;25(18):1915-27

Functionality controversy

- **Transcriptional Noise**
  - Low affinity binding of RNA polymerase to randomly generated promoter sequences.
  - More efficient to allow random transcripts than to downregulate nonspecific transcription.
  - LncRNAs are generally expressed at low levels
  - LncRNA sequences are not well conserved between species.
  - Sequencing with splicing/polyadenylation signals can occur by chance-regional chromatin state would direct tissue specific transcription
- **LncRNAs are Functional**
  - LncRNAs do not have the strict sequence conservation constraint that protein-coding genes do.
  - LncRNAs may be more plastic then protein coding genes and thus can evolve rapidly.
  - LncRNA promoter sequences are very well conserved.
- General consensus: some are functional and some are not but disagreement over relative frequency

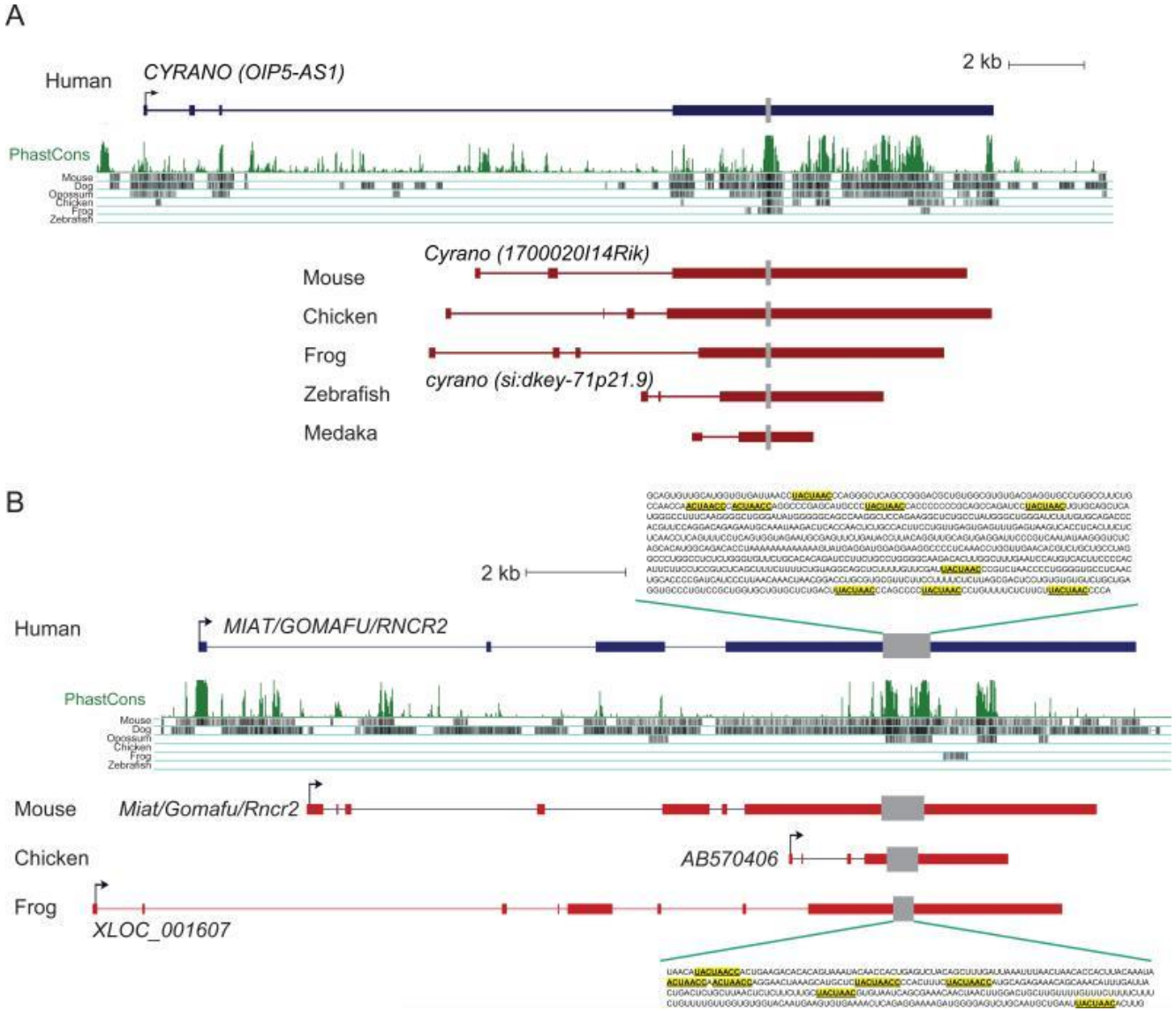
Xist – well characterized lncRNA

- single X chromosome is transcriptionally inactivated during development in XX female mammals
- Inactivation is random but once it has occurred, X inactivation is extremely stable and is maintained through subsequent cell divisions
- *XIST* (human) and *Xist* (mouse) RNA is a large non-protein-coding transcript that coats the inactive X chromosome
- Poor sequence conservation
- Exons are composed of variable length repeats but relative exon order is conserved



What about conservation?

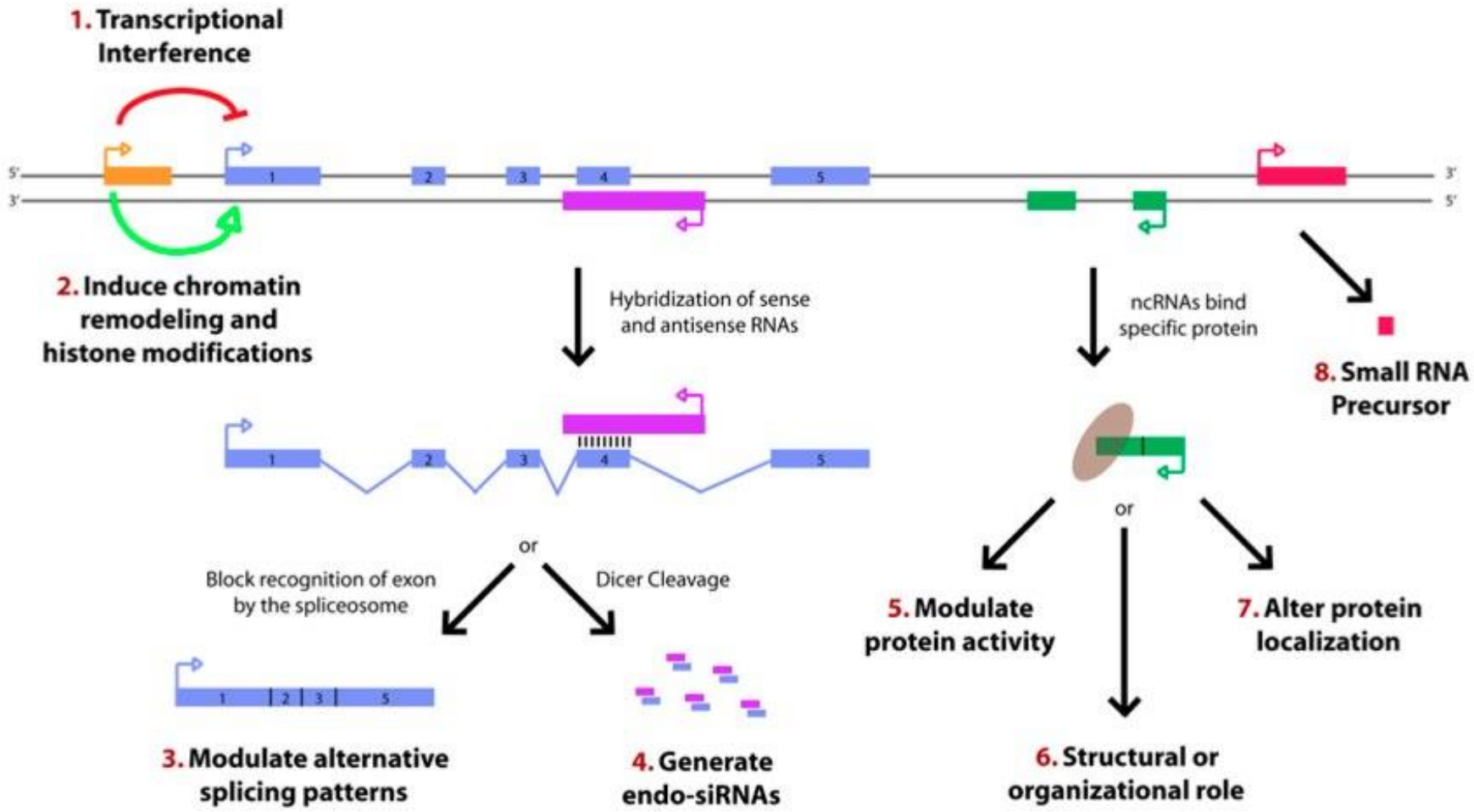
- Xist is a functional lincRNA with poor sequence conservation but significant exon structure conservation
- The *Cyrano* lincRNA is
  - conserved in vertebrates
  - required for proper morphogenesis and neurogenesis in zebrafish
- *Megamind*
  - conserved in vertebrates
  - required for proper brain development in zebrafish
  - No sequence level conservation
  - Cannot be identified via blastn alignment –but can be identified with HMM



lincRNA functional mechanisms

- Cis-acts on nearby gene only, depends on the site of transcription
  - Two possible modes of action
    - Requires transcription only
    - Requires a processed transcript
- Trans-acts elsewhere in the genome
  - Does not depend on the site of transcription

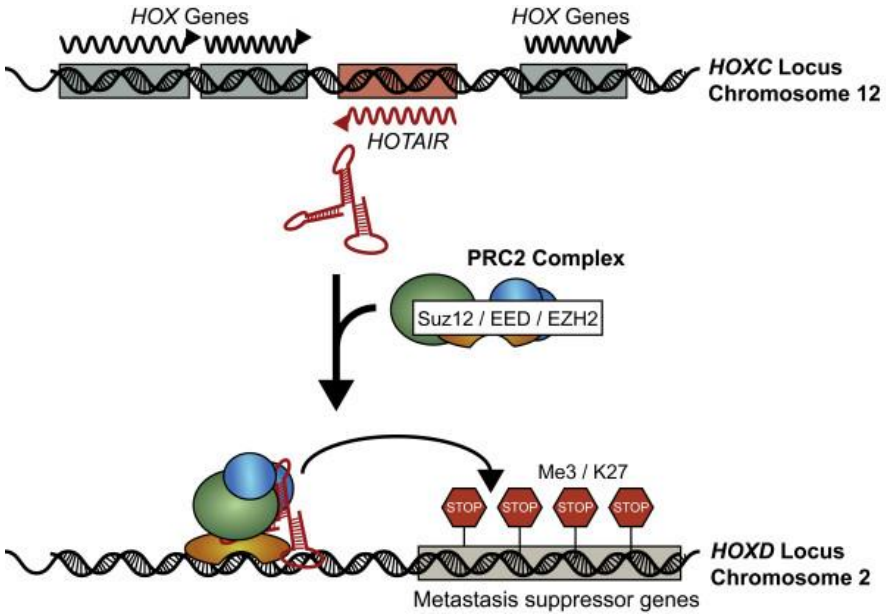
Potential functions of lncRNA



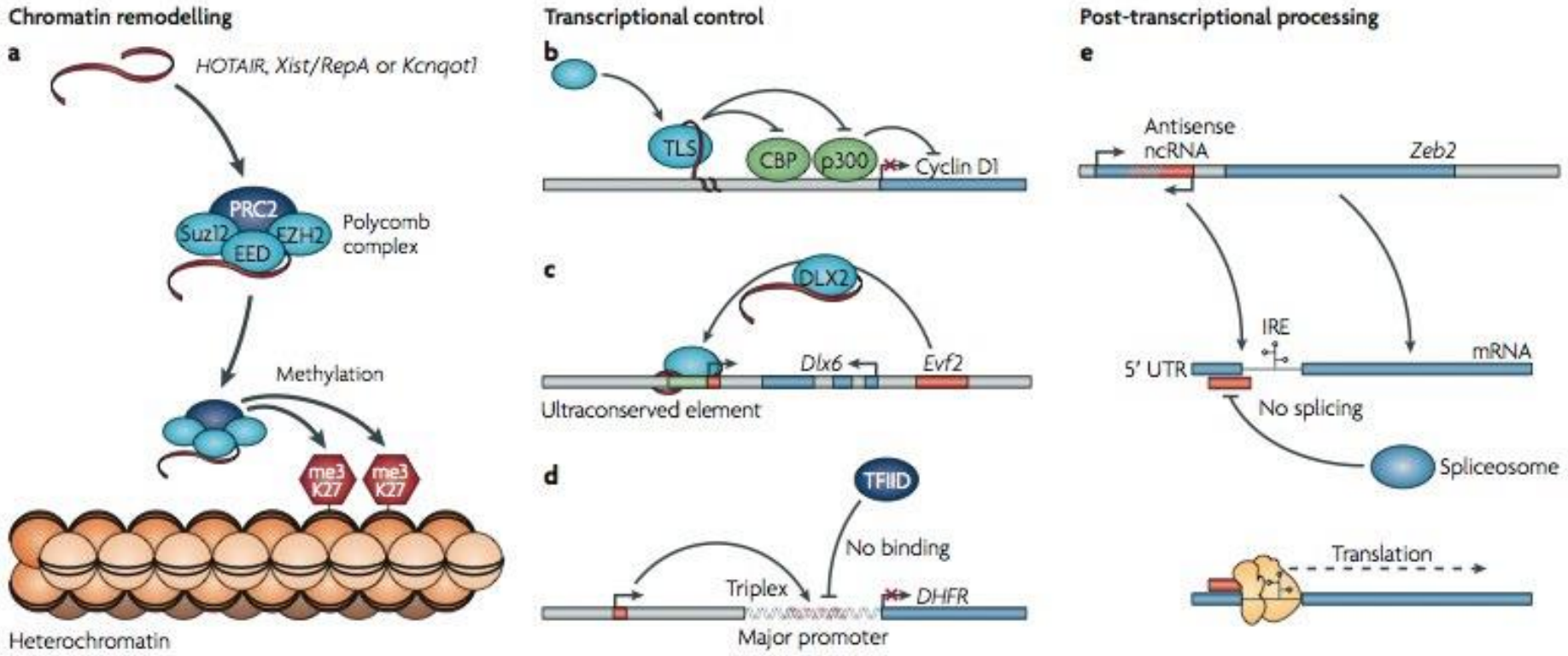
Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. 2009 23(13):1494-504.

Example: HOTAIR

- HOTAIR (for HOX transcript antisense RNA)is first example of an RNA expressed on one chromosome that has been found to influence transcription on another chromosome
- It is required for gene-silencing of the HOXD locus
- It is hypothesized to be important for epigenetic differentiation of skin over the surface of the body.
- HOTAIR was shown to contain distinct protein interaction domains that can associate with polycomb repressive complex (PRC2) and the CoREST–LSD1 complex, which together are required for correct function

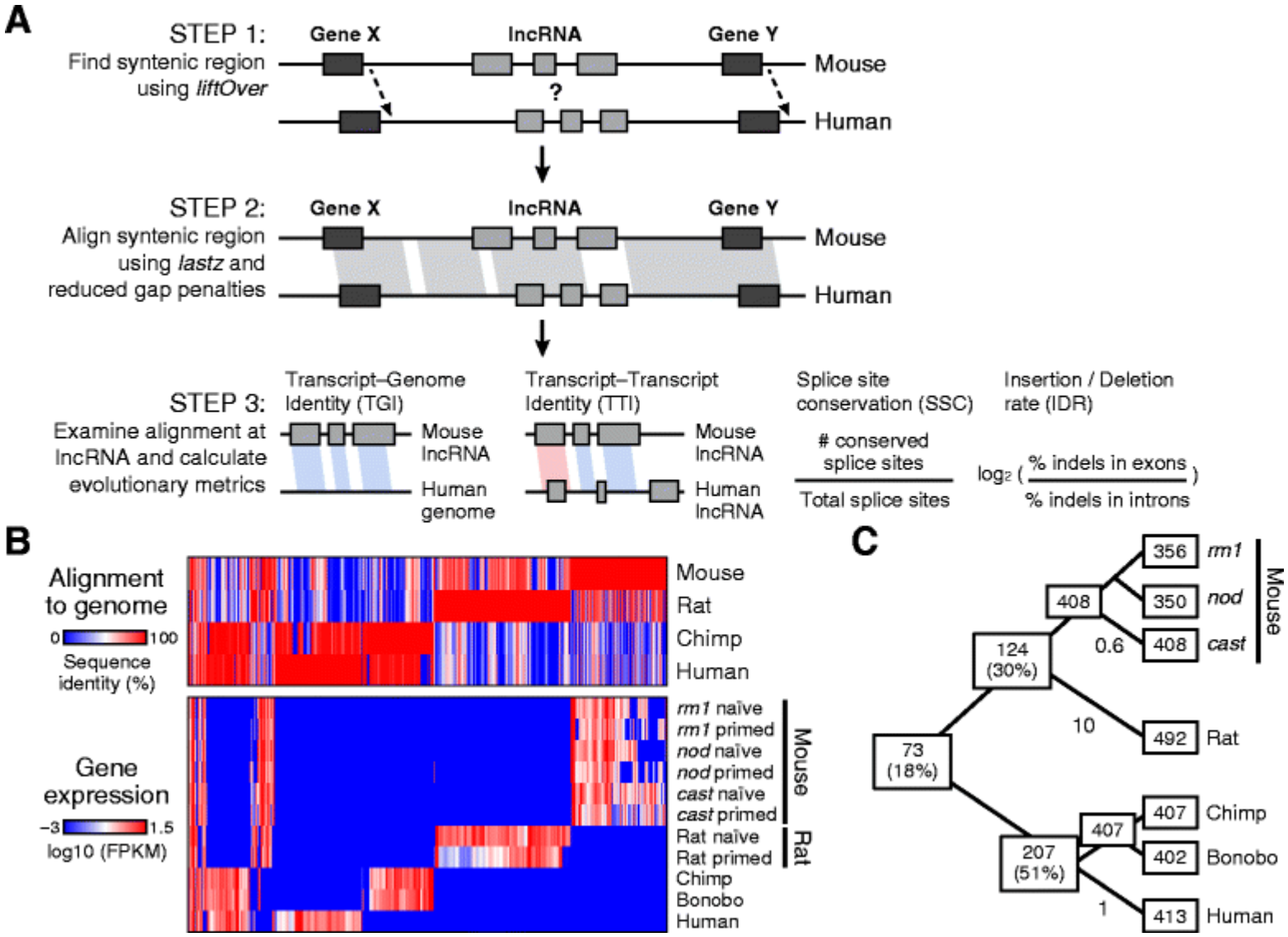


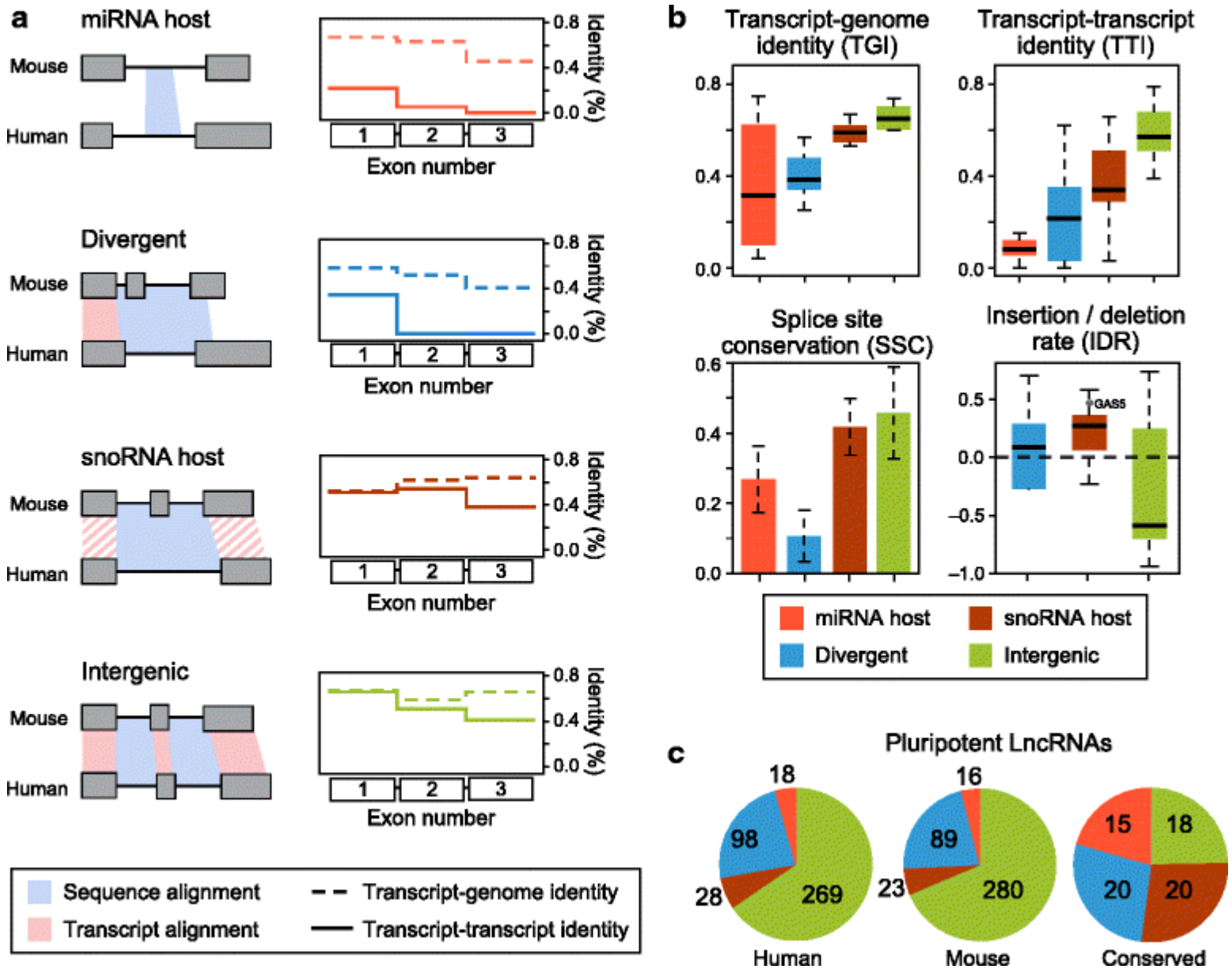
Known Examples



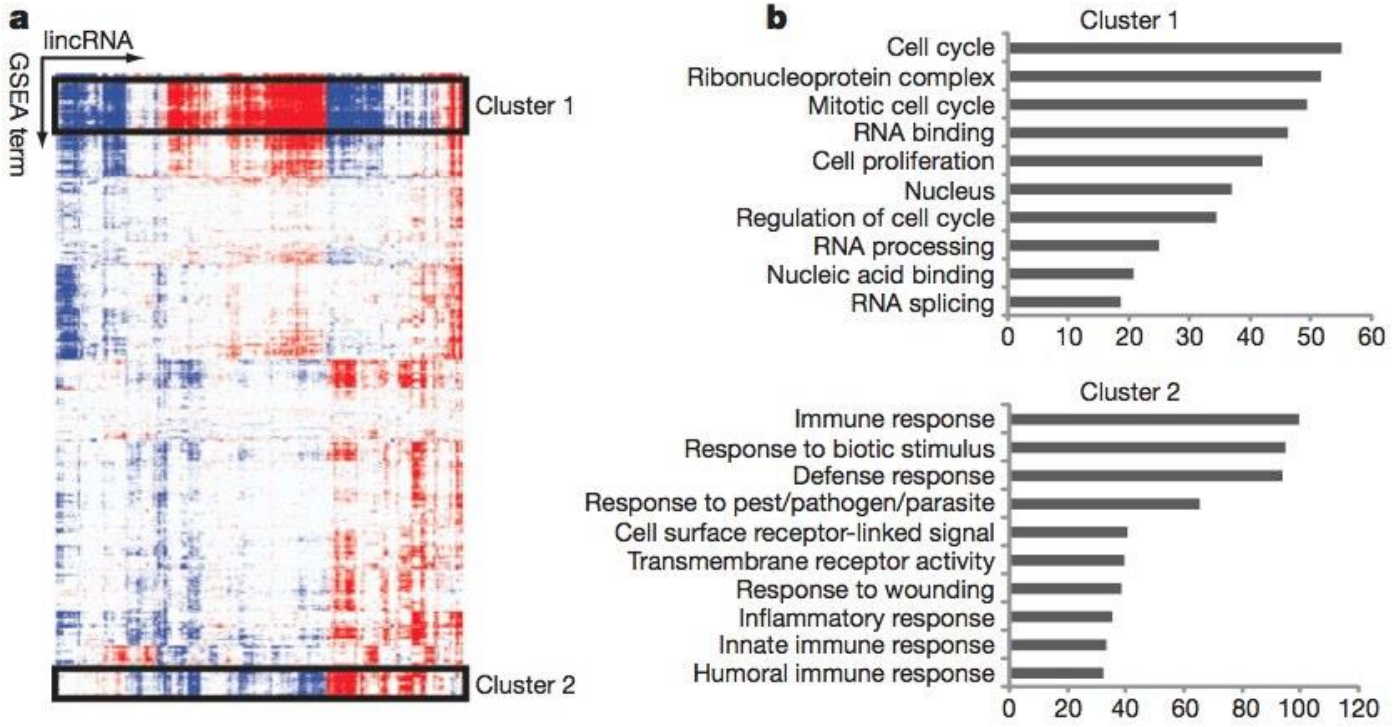
Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet. 2009 Mar;10(3):155-9

Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs- *Genome Biology* 2016




a
miRNA host
Mouse
Human
Identity (%)
0.8
0.4
0.0
1
2
3
Exon number
Divergent
snoRNA host
Intergenic
Sequence alignment
Transcript alignment
Transcript-genome identity
Transcript-transcript identity
b
Transcript-genome identity (TGI)
Transcript-transcript identity (TTI)
Splice site conservation (SSC)
Insertion / deletion rate (IDR)
GAS5
miRNA host
snoRNA host
Divergent
Intergenic
c
Pluripotent LncRNAs
18
98
28
269
Human
16
89
23
280
Mouse
15
18
20
20
Conserved

Functional assignment based on gene expression correlation



Guttman M. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009 458(7235):223-7

Knock-out studies

- Selected 18 lincRNAs beased on
  - Conservation
  - Chromatin features
  - Low protein-coding potential
    - Sequence based filter
    - Mass-spec based filter
- postnatal lethal phenotypes in three mutant strains (*Fendrr*, *Peril*, and *Mdgt*), the latter two exhibiting incomplete penetrance and growth defects in survivors
- growth defects for two additional mutant strains (*linc–Brn1b* and *linc–Pint*)

Multiple knockout mouse models reveal lincRNAs are required for life and brain development

Strain	+/+	+/-	-/-	Total	p-value
linc-Brn1a	12 (13)	32 (26)	7 (13)	51	0.1168
linc-Brn1b	16 (17)	39 (33)	11 (17)	66	0.1952
linc-Cox2	10 (10)	19 (20)	11 (10)	40	0.9277
Fabl	18 (23)	52 (45)	20 (23)	90	0.3220
linc-Enc1	16 (12)	17 (23)	13 (12)	46	0.2252
Manr	21 (20)	37 (40)	22 (20)	80	0.7886
Fendrr	36 (23)	57 (47)	0 (23)	93	8.9 E-8
Haunt	20 (19)	44 (39)	13 (19)	77	0.2741
Hottip	8 (8)	16 (17)	9 (8)	33	0.9122
Mdgt	25 (17)	37 (34)	6 (17)	68	0.0038
Celr	11 (19)	43 (38)	21 (19)	75	0.1202
Cmde	20 (19)	41 (39)	16 (19)	77	0.7302
Spasm†	13 (22)	29 (22)	47 (45)	89	0.0498
linc-Pint	14 (12)	23 (23)	9 (12)	46	0.5818
linc-p21	19 (20)	40 (39)	19 (20)	78	0.9391
linc-Ppara	13 (14)	35 (28)	8 (14)	56	0.1112
Peril	34 (32)	79 (63)	13 (32)	126	0.0005
Tug1	15 (11)	19 (21)	8 (11)	42	0.2574

lincRNA summary

- Many are not functional but some are
- Sequence conservation is poor but we can look for
  - Small conserved regions
  - Promoter conservation
  - Exon-intron structure
  - Synteny
  - Non-alignable conserved feature
- lincRNAs most likely come from different classes that differ
  - Functionality
  - Mechanism
  - Gene proximity
  - Conservation