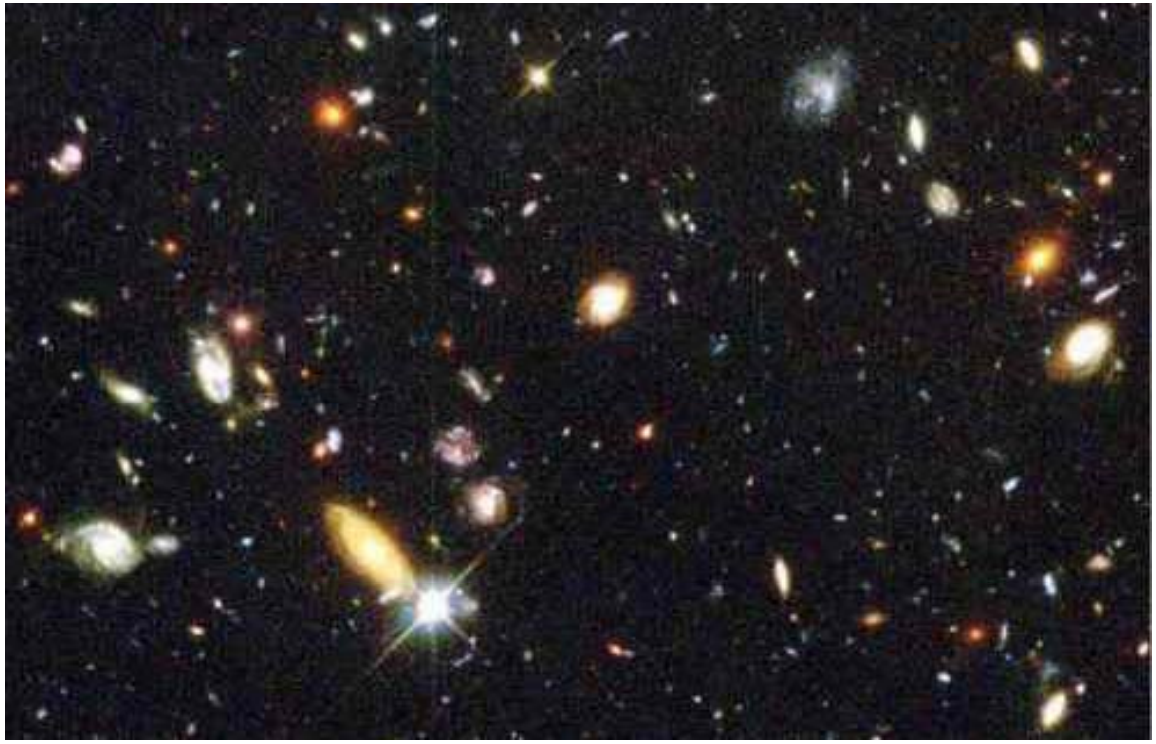
Non-Coding RNAs: Dark Matter of the Genome



Bioinformatics and Genomics Group (Professor Roderic Guigó)
Centre for Genomic Regulation, Barcelona
Rory.johnson@crg.eu
http://big.crg.cat/bioinformatics_and_genomics
www.crg.eu

1. **Please ask questions during the lecture**

2. Citations are usually given as (First Author et al, Pubmed ID)

3. For more questions or to get the slides from this presentation, please email me (address to bottom of slide).

4. Abbreviations:
    I. ncRNA = noncoding RNA
    II. lncRNA = long noncoding RNA (OR lincRNA)
    III. miRNA = microRNA

Today´s Lecture:

1. Introduction to Non-Coding RNAs (ncRNAs)

1.1 Discovery and classification
1.2 MicroRNAs
1.3 Methods for discovery and measurement
1.4 ncRNAs in human disease
1.5 Technological applications
1.6 Human evolution

2. Non-Coding RNA research at the CRG in Barcelona

Computational and functional analysis of **long non-coding RNAs**

1.1 Discovery and Classification

Complexity ~ Gene number?

Exactly 959 cells

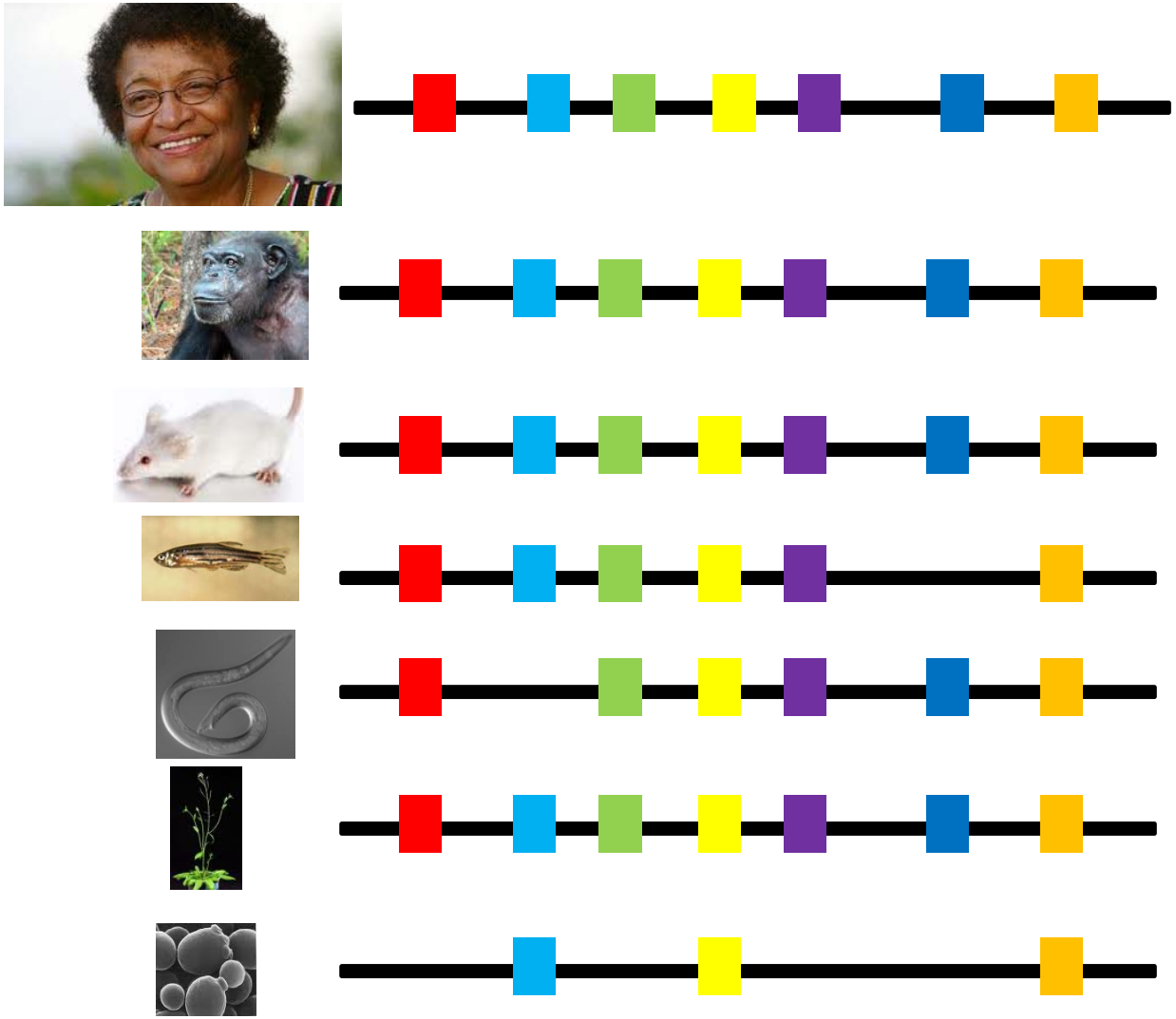
50 trillion (5×10^{12}) cells

What is the genomic basis of complexity?

20,000 protein genes
(ENSEMBL.org)

20,500 protein genes
(Clamp et PNAS 2007
104(49):19428-33)

The problem with genomics…



Protein coding gene number
(source: Ensembl.org)

21,000

20,000

22,000

22,000

20,000

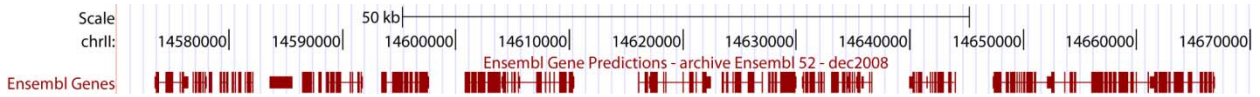
27,000

7,000

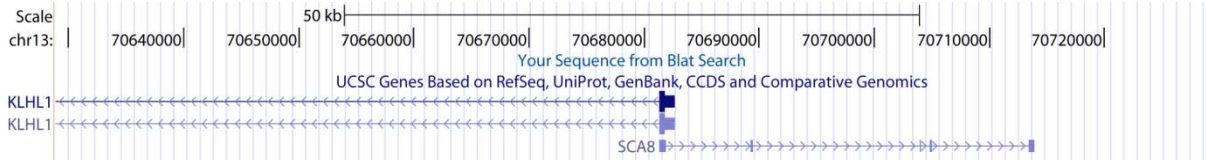
Non-coding DNA is very different between species

Does this hold the key to understanding organismal complexity?

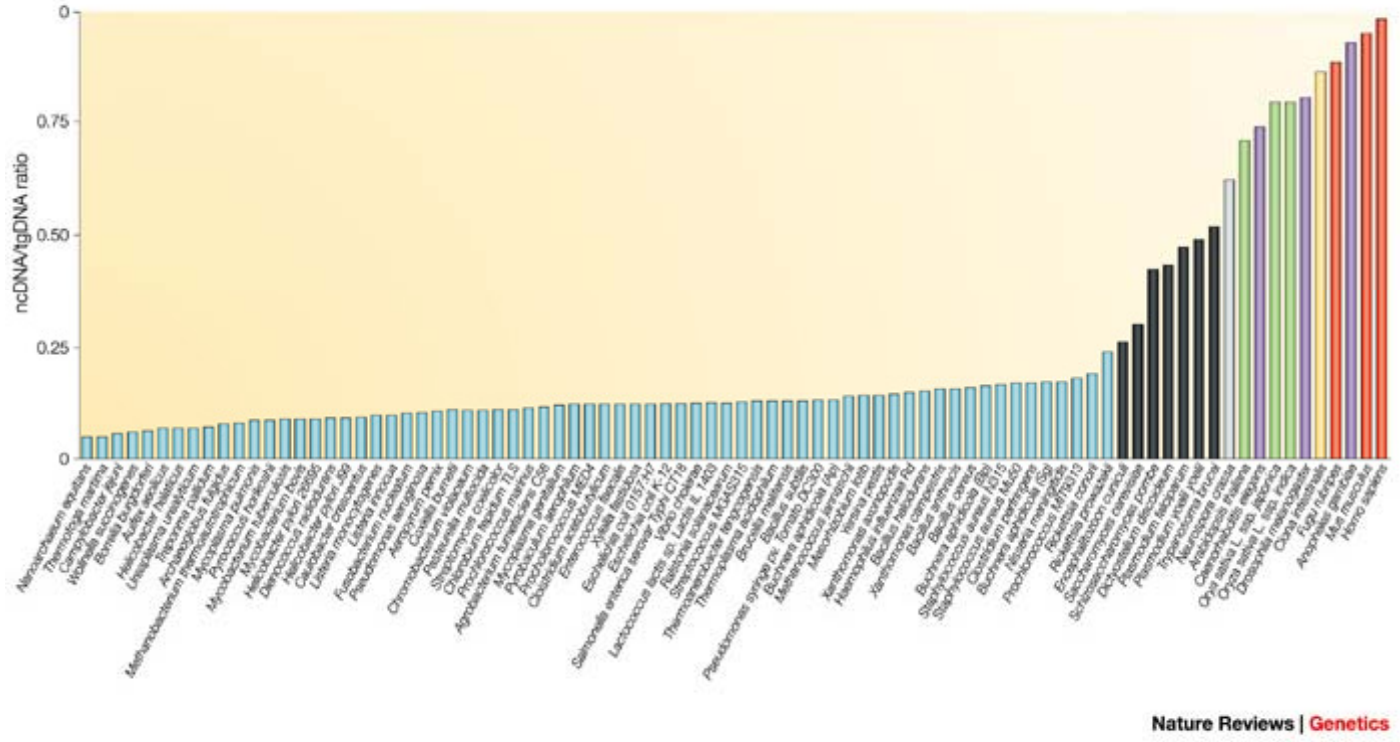
1x10^8 bp



3x10^9 bp



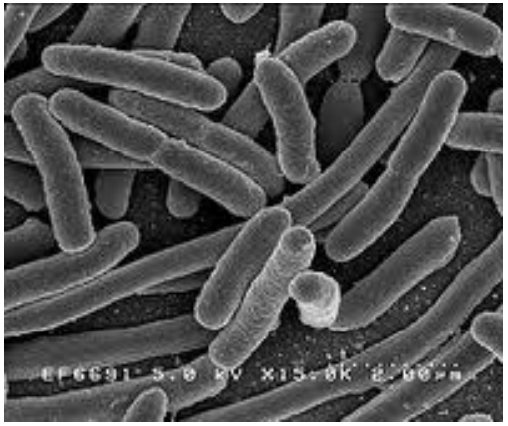
Noncoding DNA and animal complexity



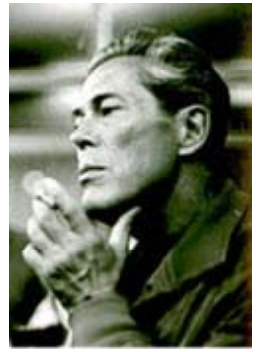
Mattick 2004 PMID15131654

The importance of noncoding DNA was not grasped in the early years of molecular biology

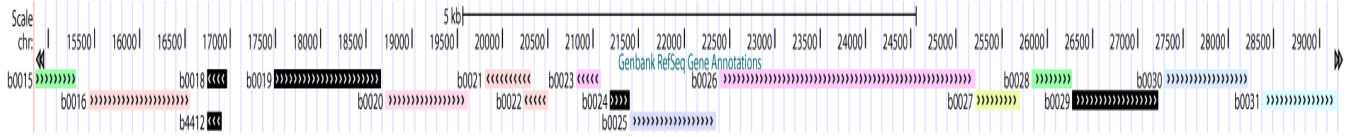
RNA as the "messenger"



"*What is true for E. coli is true for an elephant*"
-Monod



"*DNA makes RNA makes Protein*"
-Crick & Watson



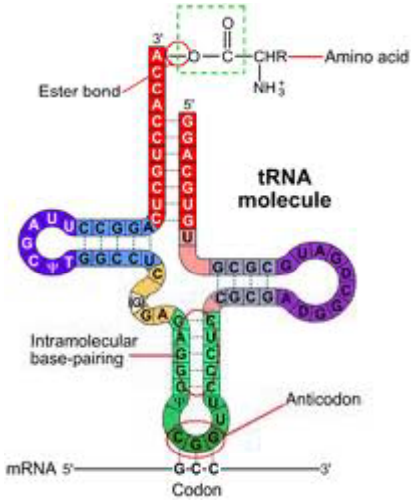
E. Coli have almost no noncoding DNA

Housekeeping ncRNAs – known since the 1960s (alanine tRNA discovered 1965, Robert Holley)

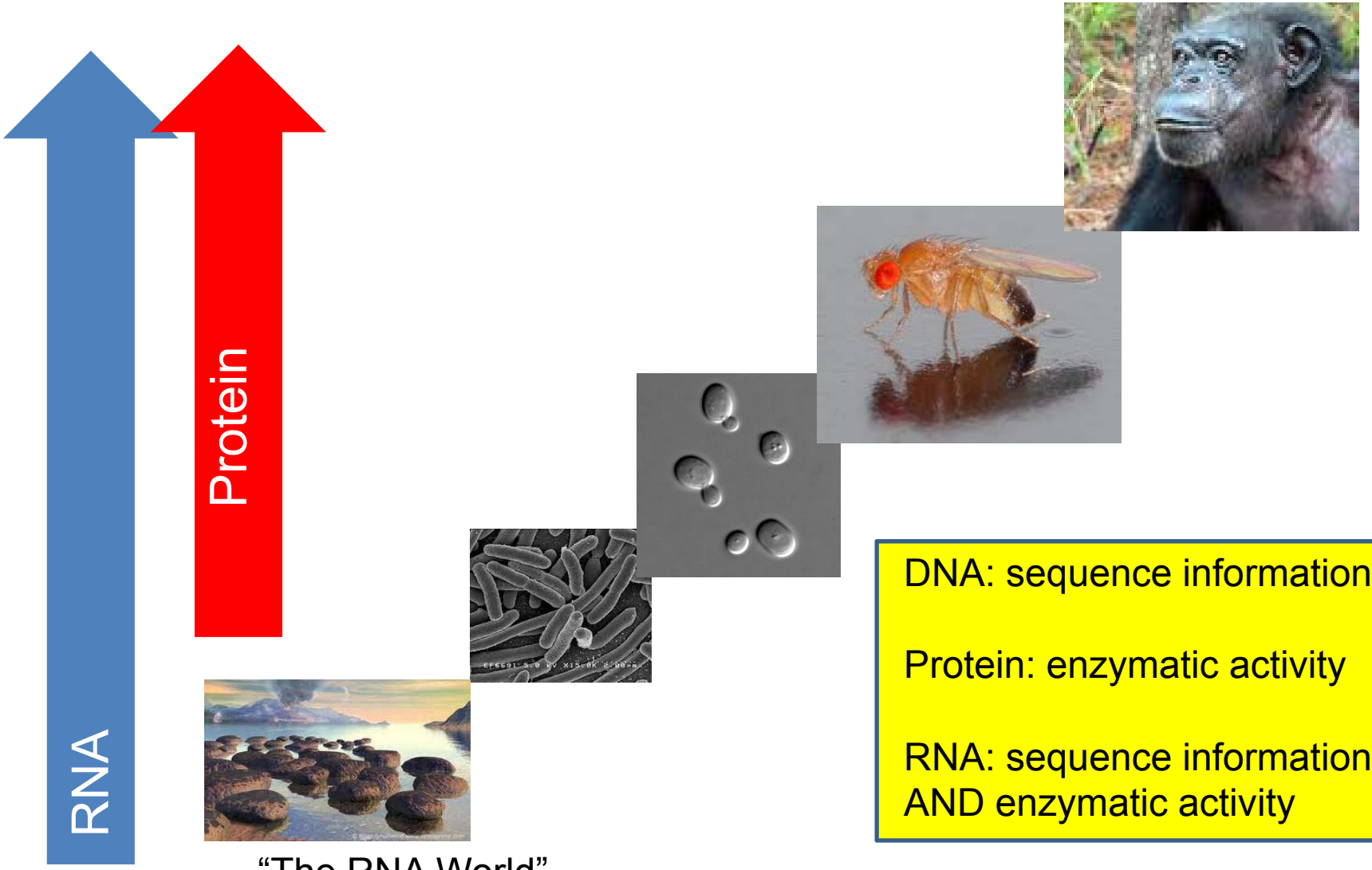
1. Transfer RNAs (tRNAs)
2. Ribosomal RNAs (rRNAs)
3. Spliceosomal RNAs (U1, U2, U4 etc)
4. RNaseP (tRNA processing)

Deeply conserved
Involved in fundamental cellular processes

- Taken to support the RNA World Hypothesis



Life probably began with non-coding RNA



"The RNA World"

DNA: sequence information

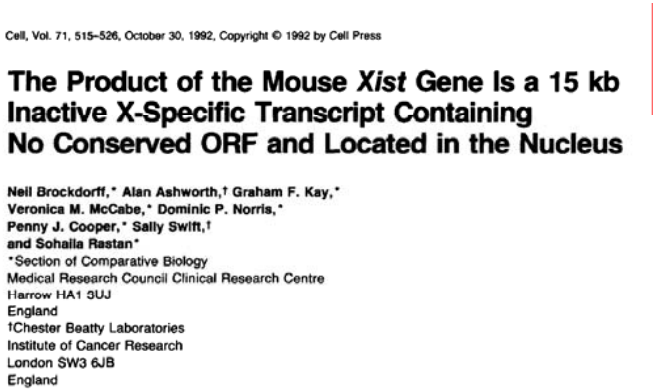
Protein: enzymatic activity

RNA: sequence information AND enzymatic activity

Early clues to the existence of regulatory non-coding RNAs….

1. Discovery of a weird small RNA in nematodes by Victor Ambros and Gary Ruvkun (1993) – the first MicroRNA

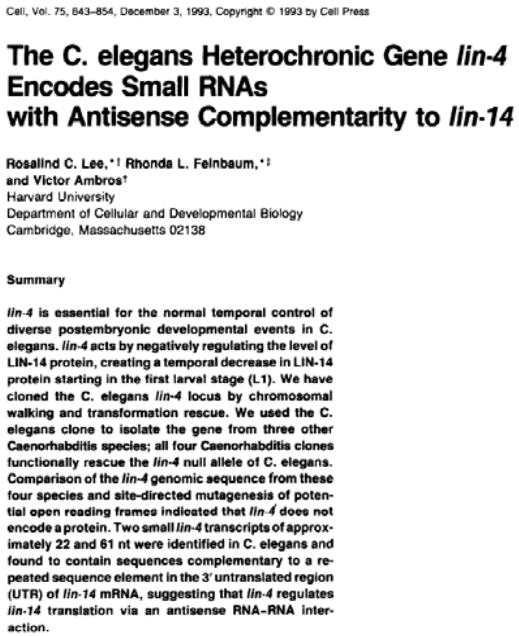
2. Discovery of XIST (1991) – the first Long Noncoding RNA



Cell, Vol. 71, 515–526, October 30, 1992, Copyright © 1992 by Cell Press

The Product of the Mouse *Xist* Gene Is a 15 kb Inactive X-Specific Transcript Containing No Conserved ORF and Located in the Nucleus

Neil Brockdorff,* Alan Ashworth,† Graham F. Kay,* Veronica M. McCabe,* Dominic P. Norris,* Penny J. Cooper,* Sally Swift,† and Sohaila Rastan*
*Section of Comparative Biology
Medical Research Council Clinical Research Centre
Harrow HA1 3UJ
England
†Chester Beatty Laboratories
Institute of Cancer Research
London SW3 6JB
England



Cell, Vol. 75, 843–854, December 3, 1993, Copyright © 1993 by Cell Press

The C. elegans Heterochronic Gene *lin-4* Encodes Small RNAs with Antisense Complementarity to *lin-14*

Rosalind C. Lee,*† Rhonda L. Feinbaum,*† and Victor Ambros†
Harvard University
Department of Cellular and Developmental Biology
Cambridge, Massachusetts 02138

Summary

***lin-4* is essential for the normal temporal control of diverse postembryonic developmental events in C. elegans. *lin-4* acts by negatively regulating the level of LIN-14 protein, creating a temporal decrease in LIN-14 protein starting in the first larval stage (L1). We have cloned the C. elegans *lin-4* locus by chromosomal walking and transformation rescue. We used the C. elegans clone to isolate the gene from three other Caenorhabditis species; all four Caenorhabditis clones functionally rescue the *lin-4* null allele of C. elegans. Comparison of the *lin-4* genomic sequence from these four species and site-directed mutagenesis of potential open reading frames indicated that *lin-4* does not encode a protein. Two small *lin-4* transcripts of approximately 22 and 61 nt were identified in C. elegans and found to contain sequences complementary to a repeated sequence element in the 3′ untranslated region (UTR) of *lin-14* mRNA, suggesting that *lin-4* regulates *lin-14* translation via an antisense RNA–RNA interaction.**

Annotation of human regulatory ncRNAs – work in progress

Decreasing Understanding!

RNA Class	Approx. Size (nt)	Approx. Number (min)	Evolutionary conservation	Function
microRNA	21	1424 (www.mirbase.org)	Metazoans, plants, algae	mRNA regulation
snoRNA	80-1000	456 (www-snorna.biotoul.fr/)	Eukaryotes, Archaea	RNA modification
piRNA	26-31	23,000 (pirnabank.ibab.ac .in)	Metazoans	Transposon suppression
Antisense RNA	>200	6,000 (Grinchuk et al 19906709)	Eukaryotes	mRNA regulation?
lncRNA	>200	>13,000 (www.gencodegen es.org)	Eukaryotes	Epigenetic regulation?
Circular RNAs	>200	???	???	???

Small RNAs Long RNAs

Discovery of lncRNA >> Characterisation

Annotated human lncRNA	13,000

Number of **characterised** human lncRNAs: **126** (lncrnadb.org)

1.2 MicroRNAs – Small but beautiful regulatory ncRNAs

- First discovered

- There are many (1000+ in human)

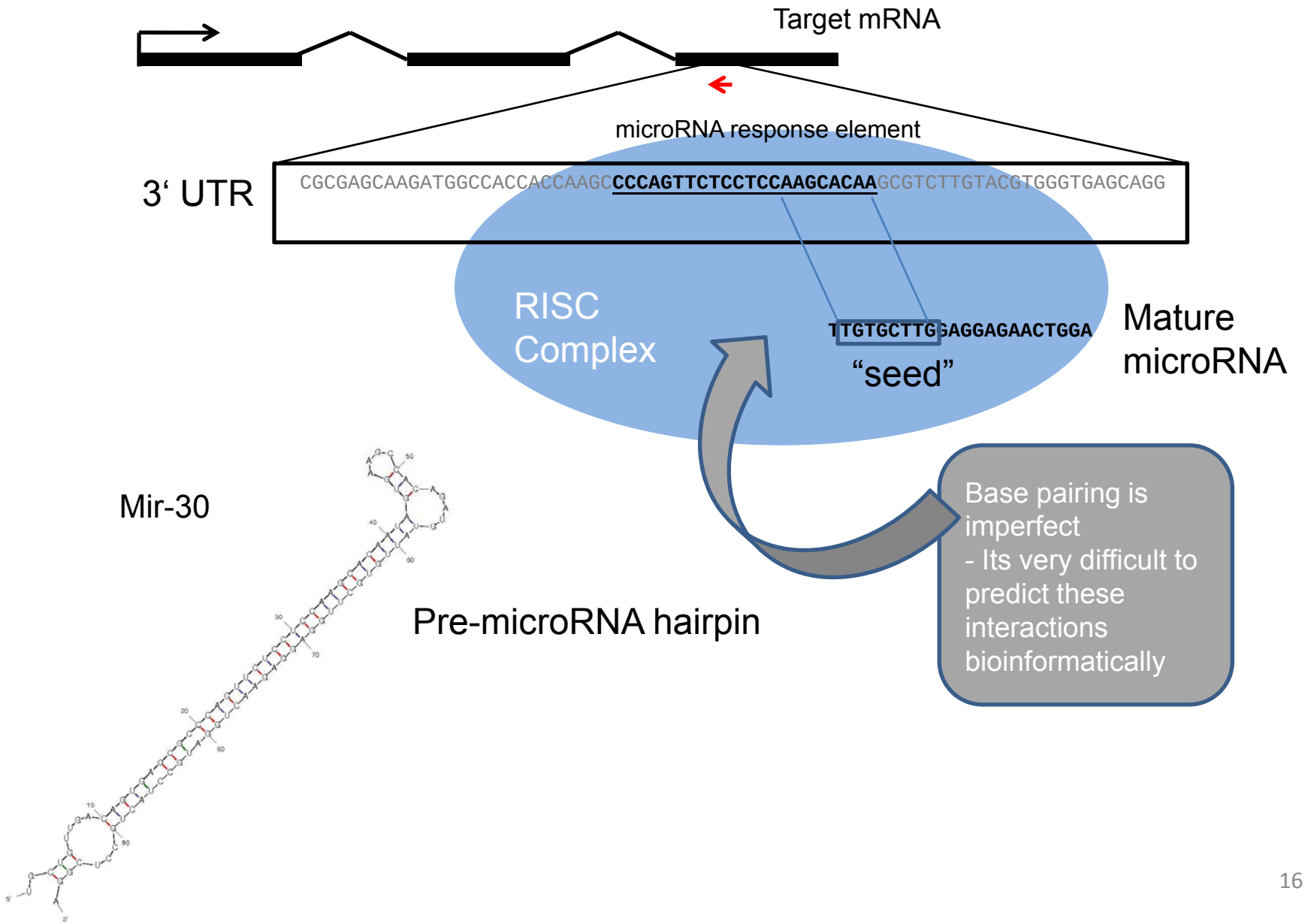
- They are found in diverse species

- They all have the same structure and size

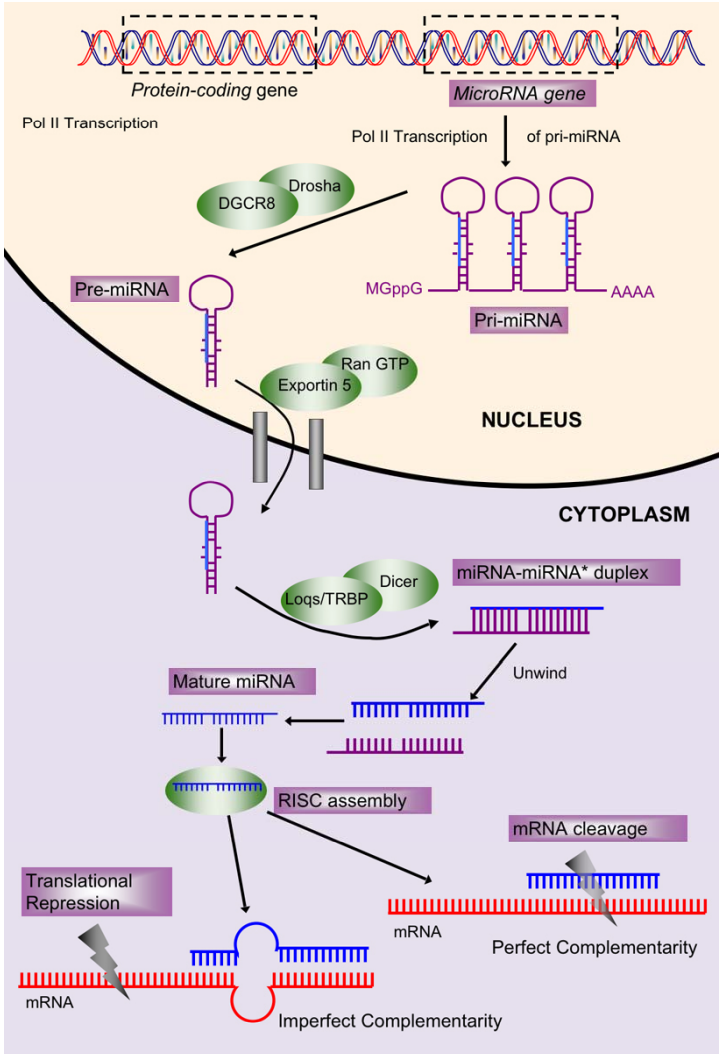
- They are all processed by the same pathway

- They all function in the same way:
  Post-transcriptional regulation of gene expression

MicroRNA – post-transcriptional gene repression by small RNAs



Biogenesis and function of microRNAs (miRNA)

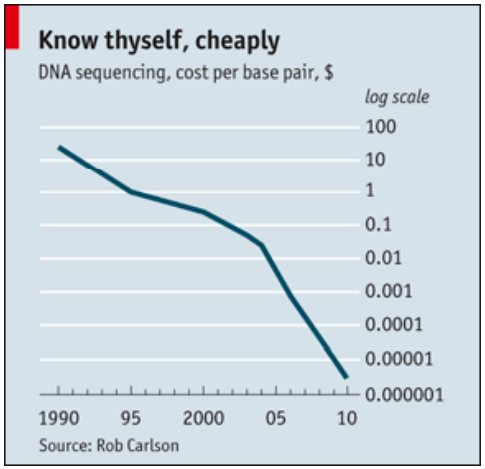


Kadri et al 19208136

- Function: negative post-transcriptional regulation of mRNAs
- Stereotypical biogenesis and functional pathways
- Sequence specific targeting of complementary RNA sequences
- Many 1000s of mRNAs are regulated by numerous miRNAs in any given cell
- Implicated in human diseases

1.3 Methods for Discovery and Measurement

Next Generation Sequencing (NGS) / High Throughput Sequencing
– a revolutionary technology for genome analysis



Various manufacturers, technologies, experimental applications.

Extremely rapid change in (a) cost, (b) throughput, (c) speed.

- The application of NGS to transcriptomics is called "**RNAseq**"

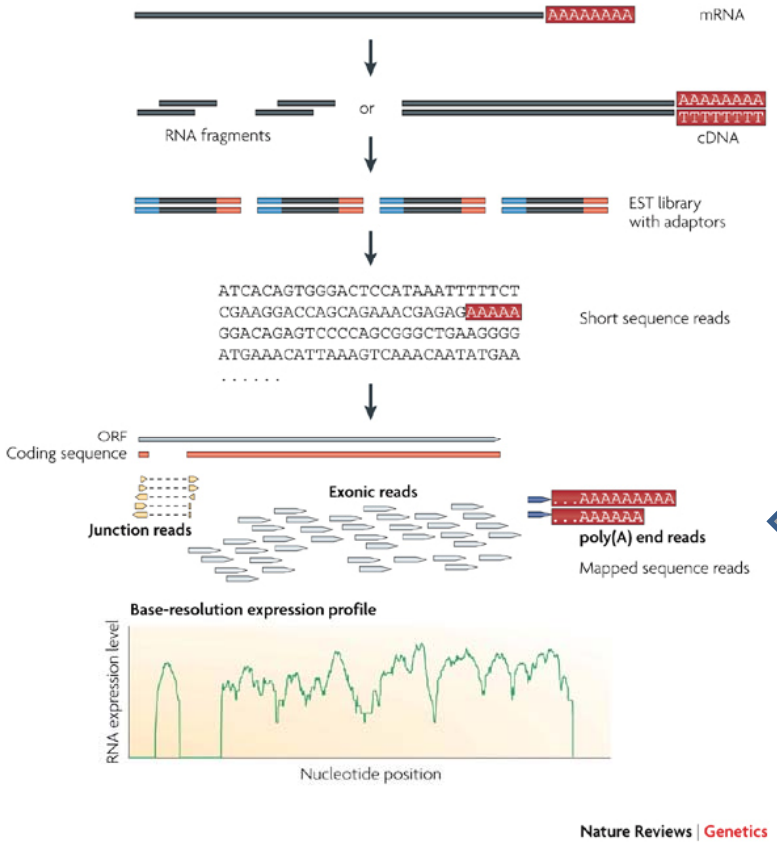
- We can use these data to:
    I. Identify new genes (including ncRNAs, short and long)
    II. Quantify the expression and splicing of any gene
    III. Examine the epigenetic signature of genes

For a good starting point to understand RNAseq see:

Mortazavi et al PMID 18516045
"Mapping and quantifying mammalian transcriptomes by RNA-Seq"

RNA sequencing (RNA seq) as a tool for detecting ncRNAs

-Advantage: (almost) unbiased measurement of (almost) all RNA in the cell
- approx 150 million reads, of approx 100 nt each
- can be adapted to short or long RNAs
- **can be used to DISCOVER new ncRNAs**



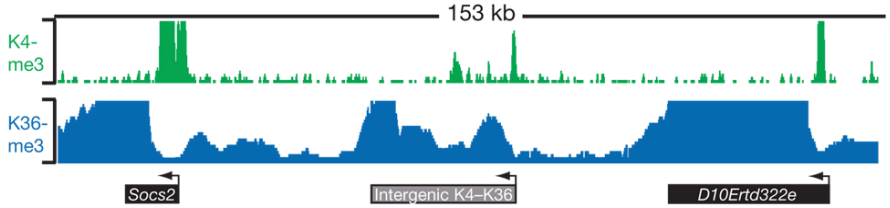
Solexa

Mapping and quantification of these reads is a contemporary bioinformatic challenge: Pioneered at CRG

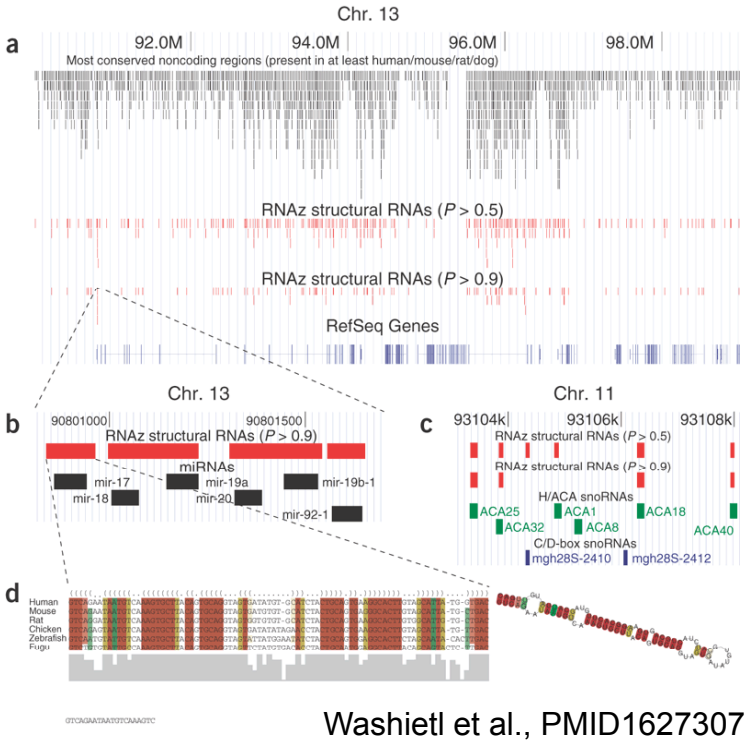
Wang et al PMID19015660

How we discover ncRNAs

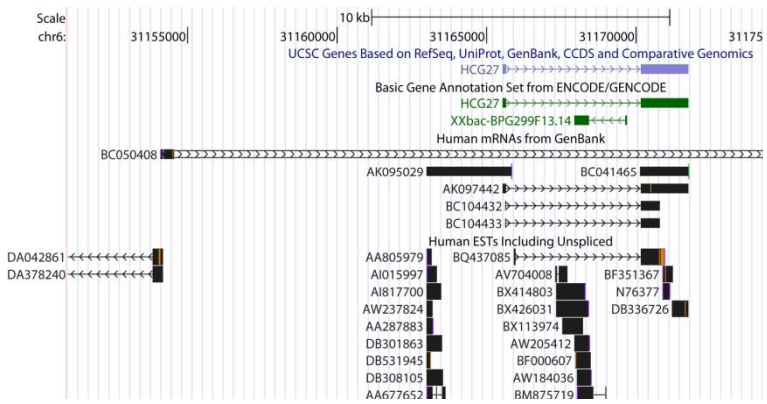
1. Genome sequence analysis
2. Manual annotation
3. Next generation sequencing
4. Epigenetic signatures



Guttman et al., PMID19182780



Washietl et al., PMID16273071

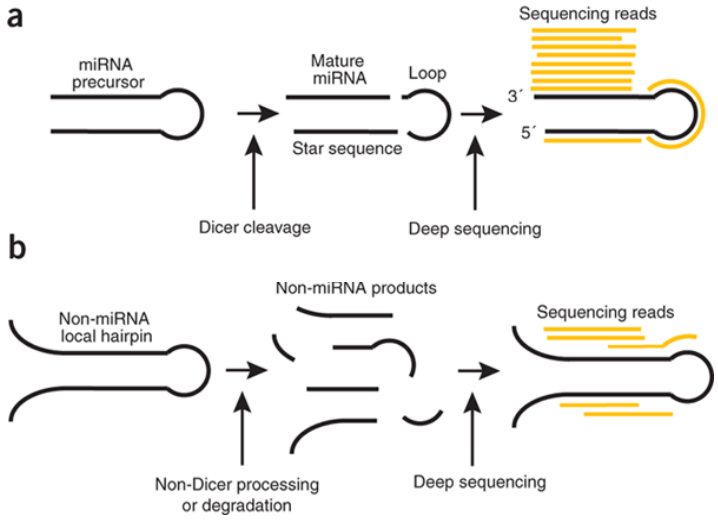


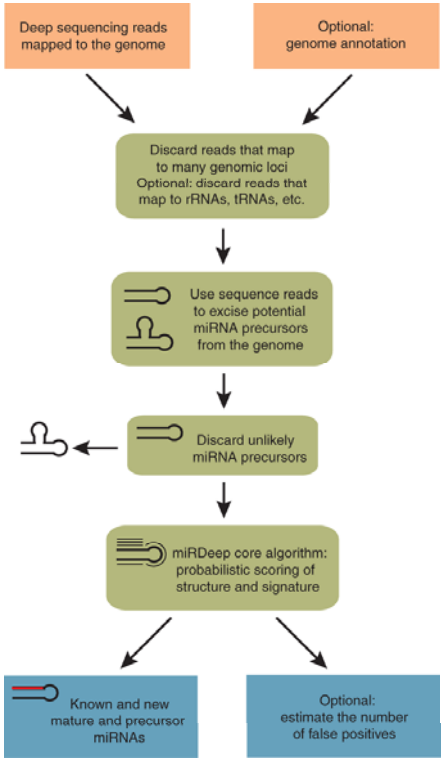
http://www.gencodegenes.org/

Example: Discovery of microRNAs using **MirDeep** program

Created by Marc Friedlaender – now a researcher at CRG

Aim: use experimental deep sequencing of small RNA sequences to discover new microRNAs.

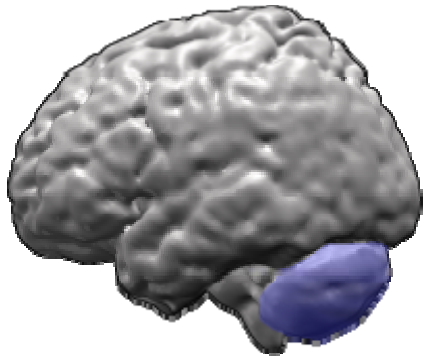




Friedlaender et al PMID 18392026

1.4 ncRNAs in human disease

- Problem of the “missing heritibility”
- Trinucleotide repeat disorders – “toxic RNAs” (eg **SCA8**)
- Neurodegenerative disease (eg **BACE1AS**)
- Blindness (Alu ncRNA expression in age-related macular degeneration)
- Heart disease (eg MIAT)
- Cancer (eg mir-34)
- Presently no drugs that can target ncRNAs, however oligonucleotide technology is promising.



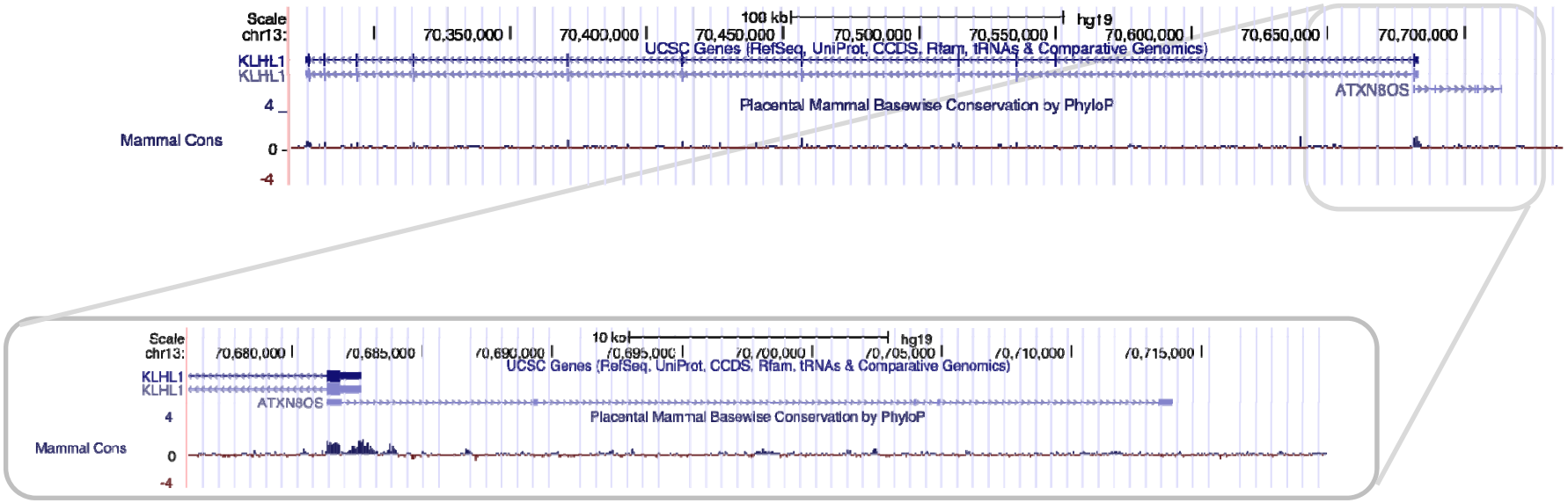
Example: Spinocerebellar ataxia, Type 8

• A member of the trinucleotide repeat expansion class of neurodegenerative disorders

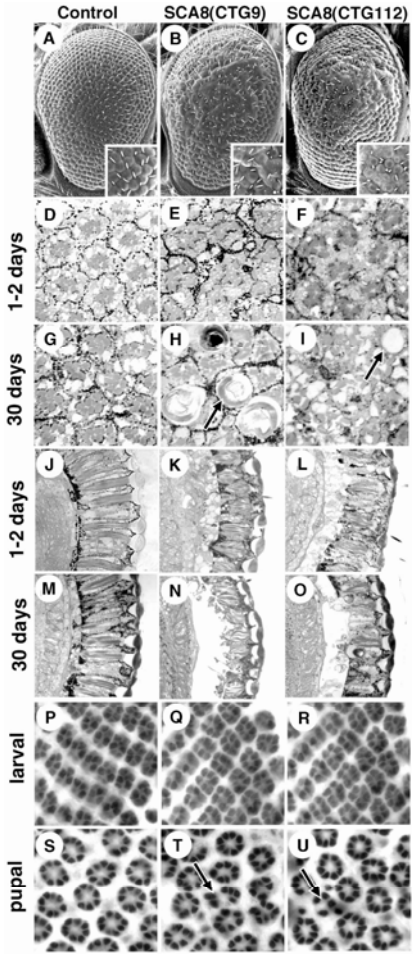
•Symptoms: Progressive loss of coordination in walking, speech, hands, eyes, due to degeneration of neurons in the cerebellum

• Cause: trinucleotide CUG repeat expansion in exon 5 of **SCA8** noncoding RNA

CUG repeat in spinocerebellar ataxia 8



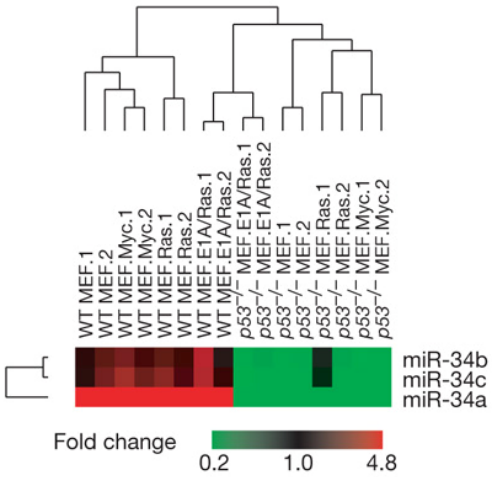
• SCA8 expression leads to neurodegeneration in flies

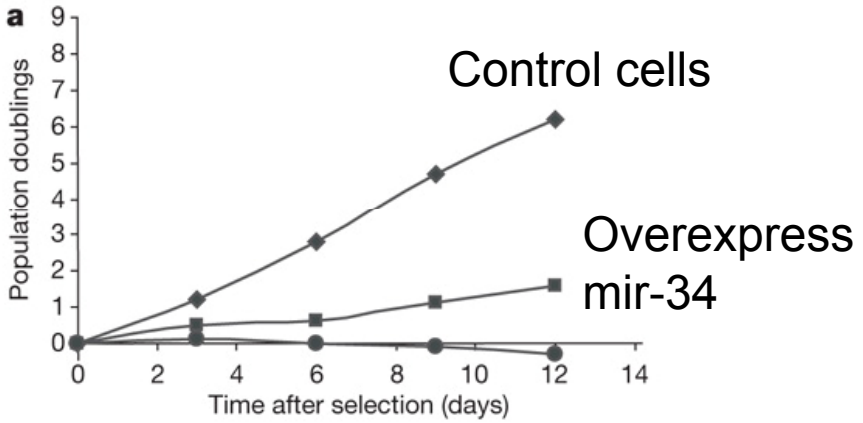


Mutsuddi et al
Current Biology
Volume 14, Issue 4, 17 February 2004,
Pages 302-308

Example: microRNA mir-34 tumour-suppressor microRNA

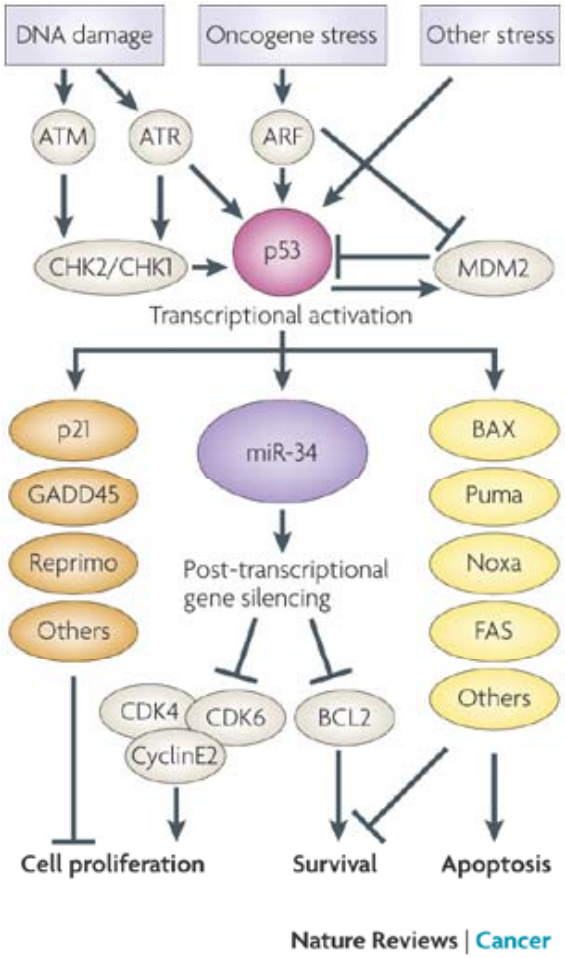
- Many changes in microRNA expression observed in cancers
- Genetic amplifications and deletions of microRNA observed in tumours
- Both tumour suppressor and oncogenic microRNAs identified





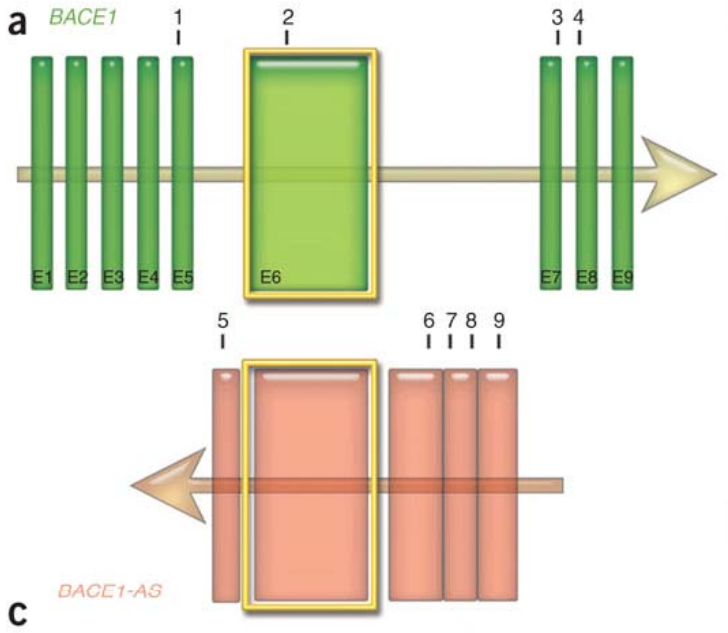
He et al., PMID 17554337

Mir-34 functions within the p53 tumour suppressor gene network



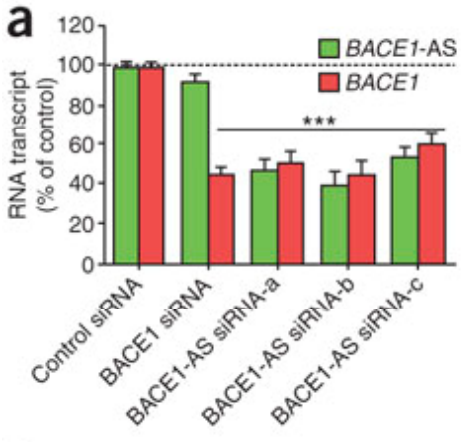
Example: an antisense ncRNA in Alzheimer´s disease

BACE1 gene encodes β-secretase, a key contributor to beta amyloid production in Alzheimer´s brain.

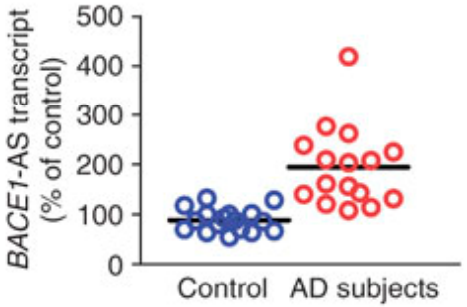


Faghihi et al., PMID 18587408

Reduction of BACE1AS leads to reduction of BACE1



BACE1AS is elevated in Alzheimers patients brains



BACE1AS is thus a good drug target for Alzheimers….

1.5 Technological Applications

“**RNA interference**” (RNAi)

Notably **siRNA (small interfering RNA)** technology, based on the microRNA pathway.

Can be used to “**knock down**” (reduce) almost any gene.

Extremely useful in

1. Research – investigate gene function
2. Medicine – can potential knock down disease genes

Nobel Prize, Physiology and Medicine, 2006: Fire and Mello

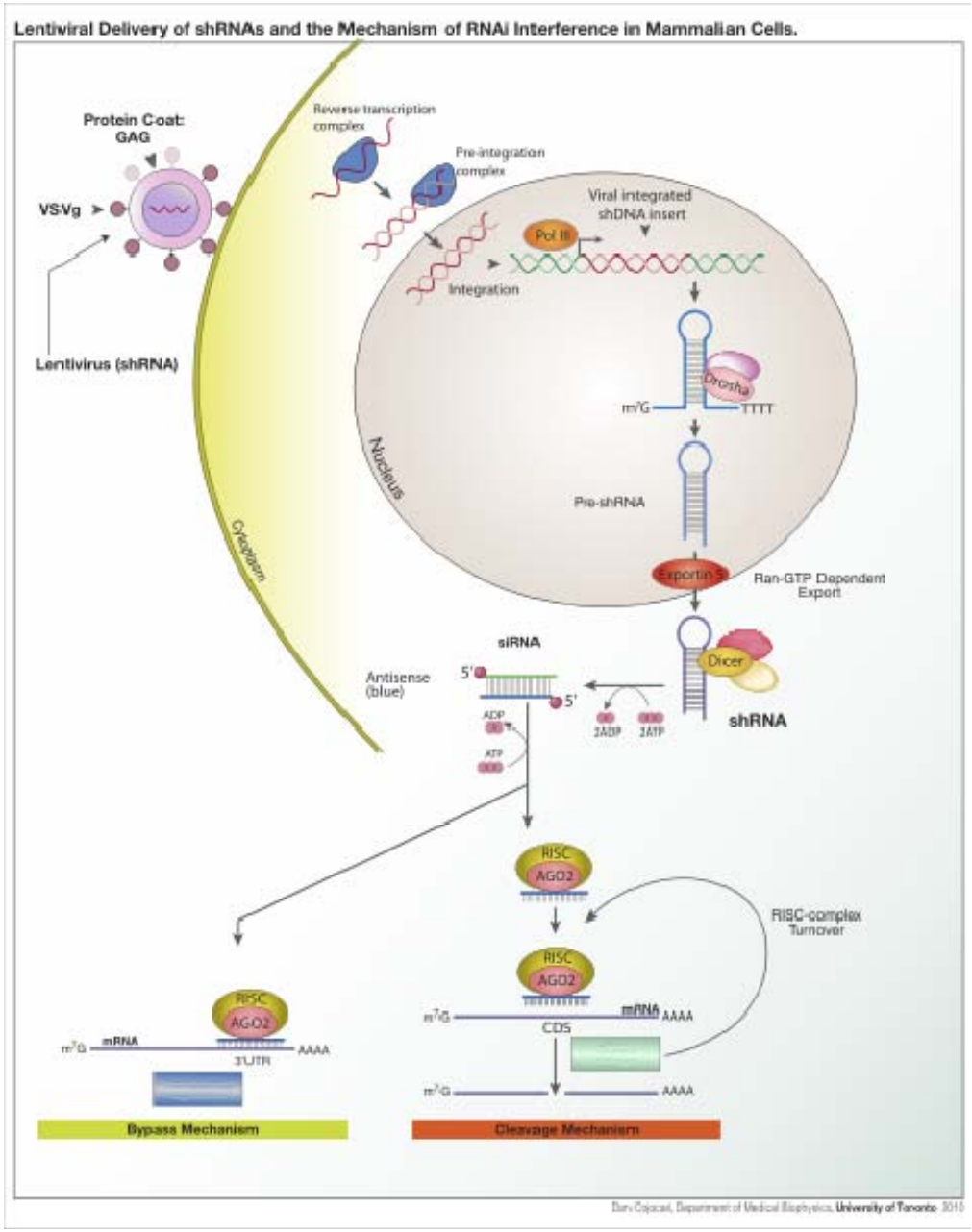
Must design effective siRNA sequence for target gene

Various methods of delivery:
• transfection, lentivirus, retrovirus

Can be delivered in vivo

Chemically-modified nucleic acids have longer half-life

Essentially an “artificial microRNA”



1.6 ncRNAs and Human Evolution

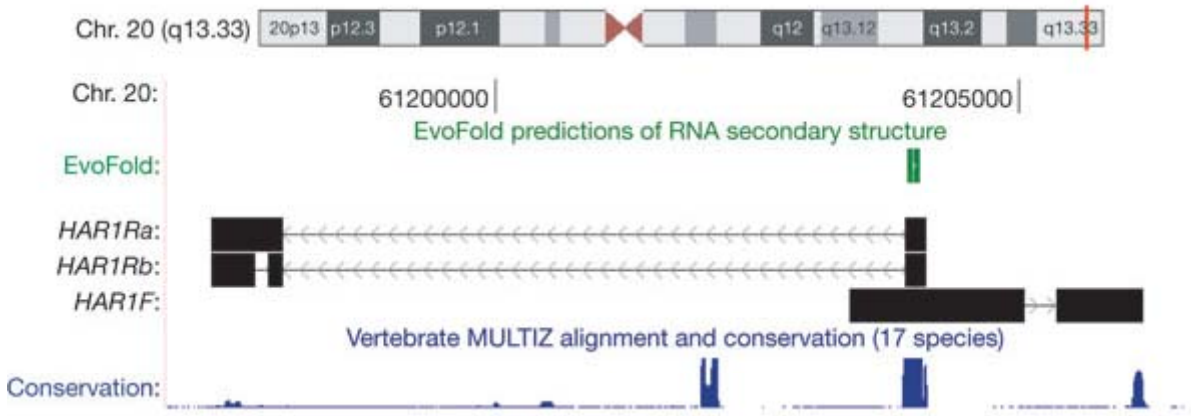
1. ncRNAs evolve rapidly
2. Many ncRNAs are expressed in the human brain
3. ncRNAs are involved in epigenetics and gene regulation

=> candidates for the evolution of human-specific traits, like intelligence

See Review articles by John Mattick:
Eg Mattick JS PMID21557942

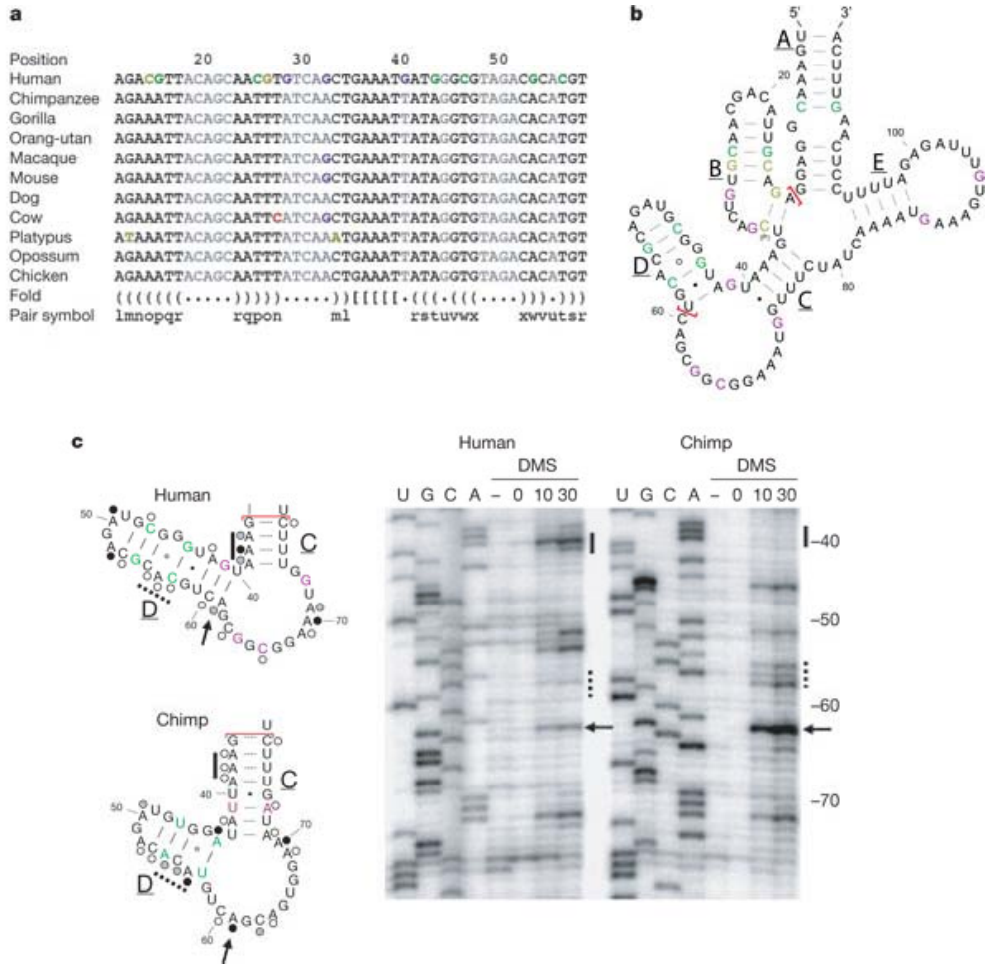
Example: HAR1 (Human accelerated region 1) noncoding RNAs

Discovered in a genome-wide search for regions of the human genome where evolutionary change has accelerated recently (since divergence from other apes)

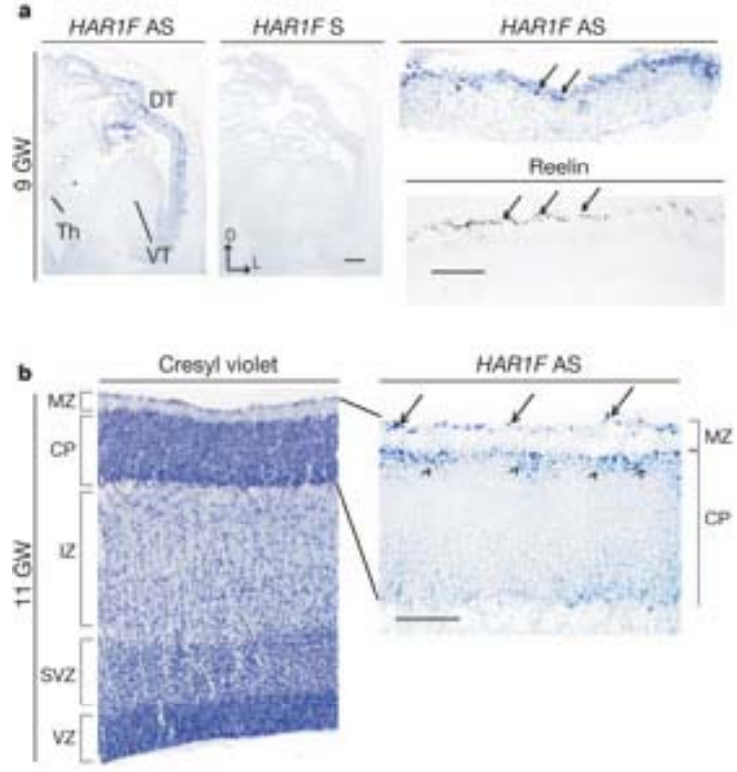


Pollard et al., PMID16915236

Changes to HAR1 sequence in very recent human evolution may have altered the structure and function of the RNA



HAR1 RNA is expressed in developing cortical neurons of human embryos



DT- dorsal telencephalon

Cerebral cortex

Part 2: Studying human long noncoding RNAs (lncRNA)

Roderic Guigó group

Bioinformatics and Genomics

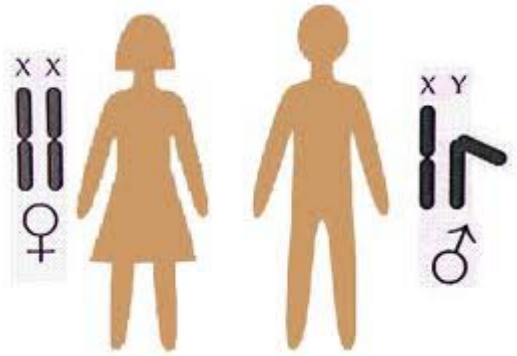
Centre for Genomic Regulation

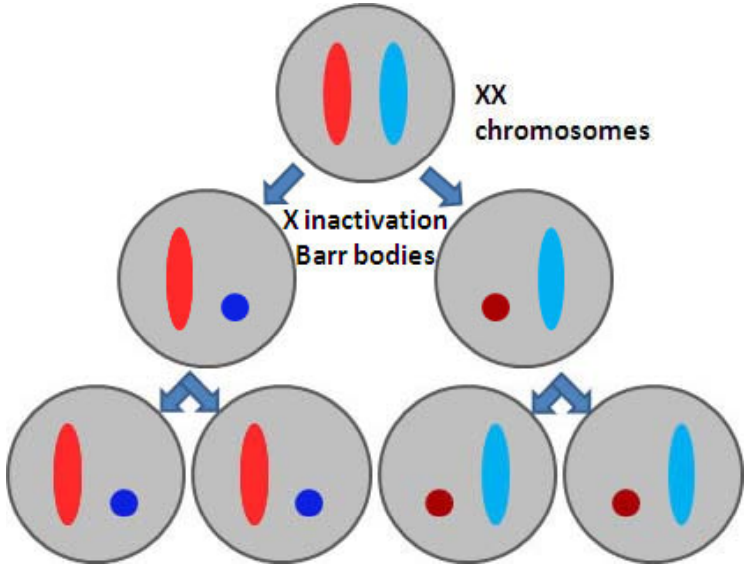
Website: http://big.crg.cat/bioinformatics_and_genomics

For the BEST (in my opinion) review of lncRNA see:
Ulitsky and Bartel. lincRNAs: Genomics, Evolution,
and Mechanisms. PMID 23827673

Maceo Parker

X chromosome inactivation
“Dosage compensation”





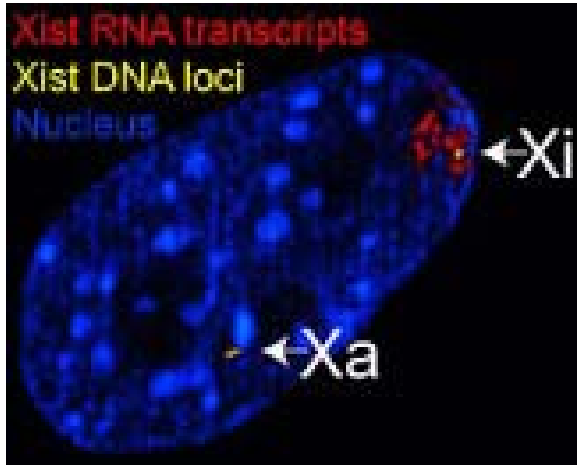
A long noncoding RNA “XIST” controls mammalian dosage compensation

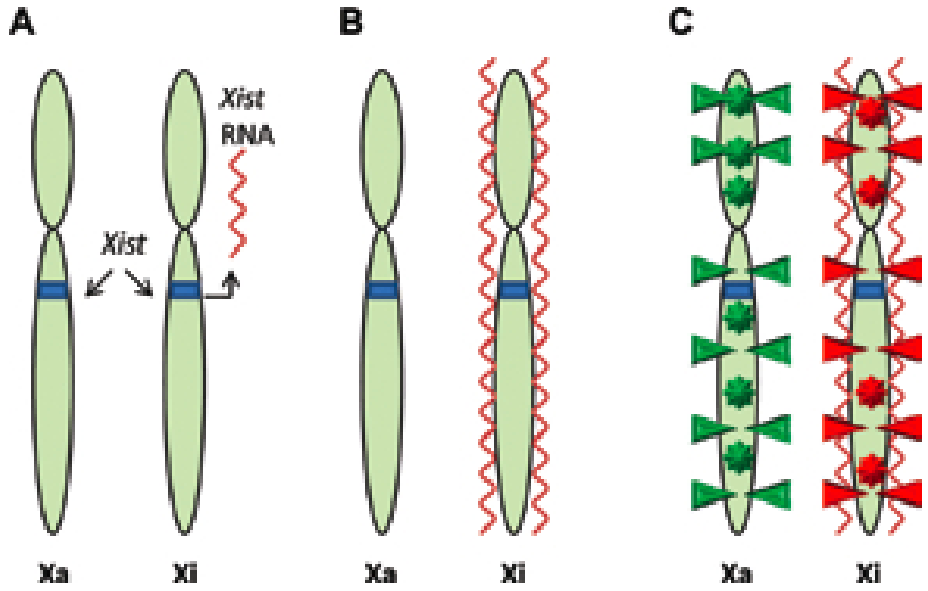
Cell, Vol. 71, 515–526, October 30, 1992, Copyright © 1992 by Cell Press

The Product of the Mouse *Xist* Gene Is a 15 kb Inactive X-Specific Transcript Containing No Conserved ORF and Located in the Nucleus

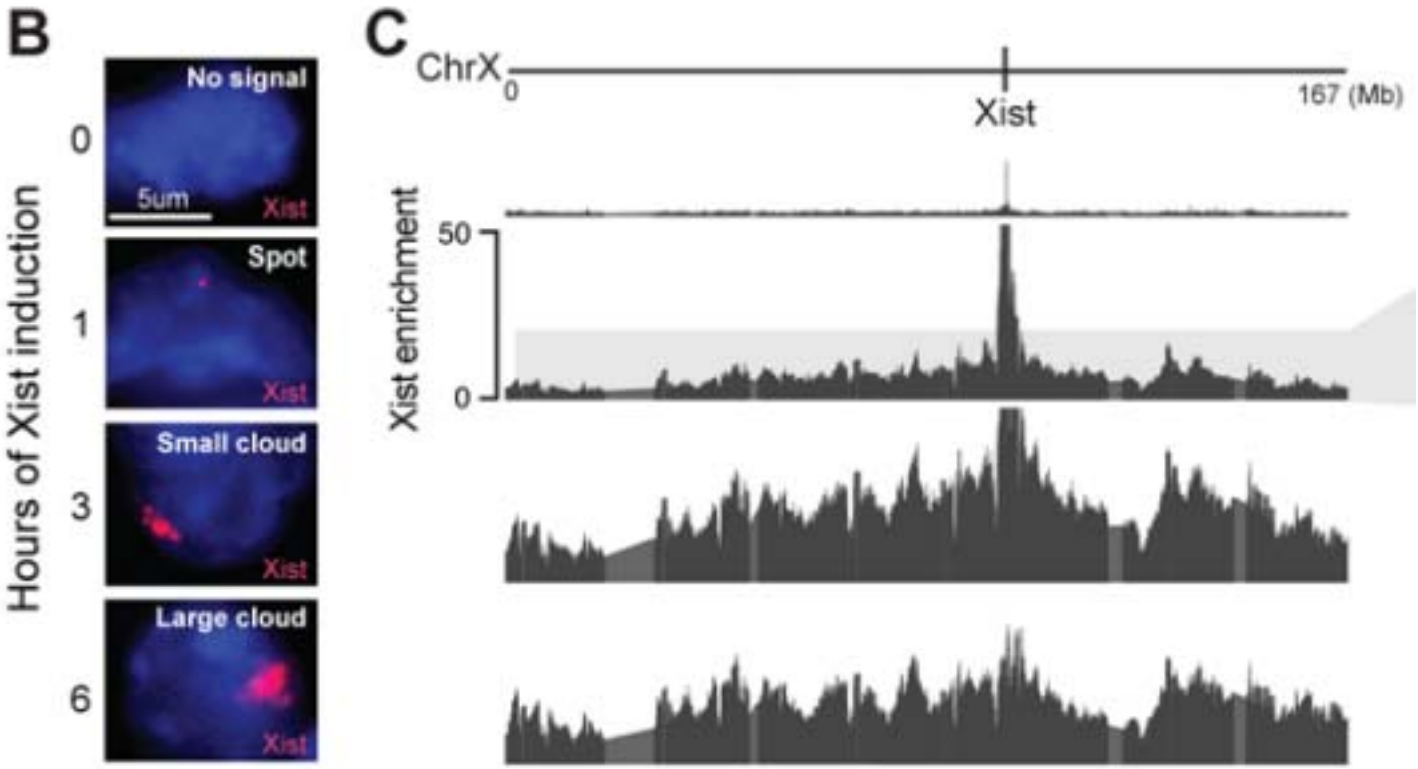
Neil Brockdorff,* Alan Ashworth,† Graham F. Kay,*
Veronica M. McCabe,* Dominic P. Norris,*
Penny J. Cooper,* Sally Swift,†
and Sohaila Rastan*
*Section of Comparative Biology
Medical Research Council Clinical Research Centre
Harrow HA1 3UJ
England
†Chester Beatty Laboratories
Institute of Cancer Research
London SW3 6JB
England

XIST “the Mother of All Noncoding RNAs”

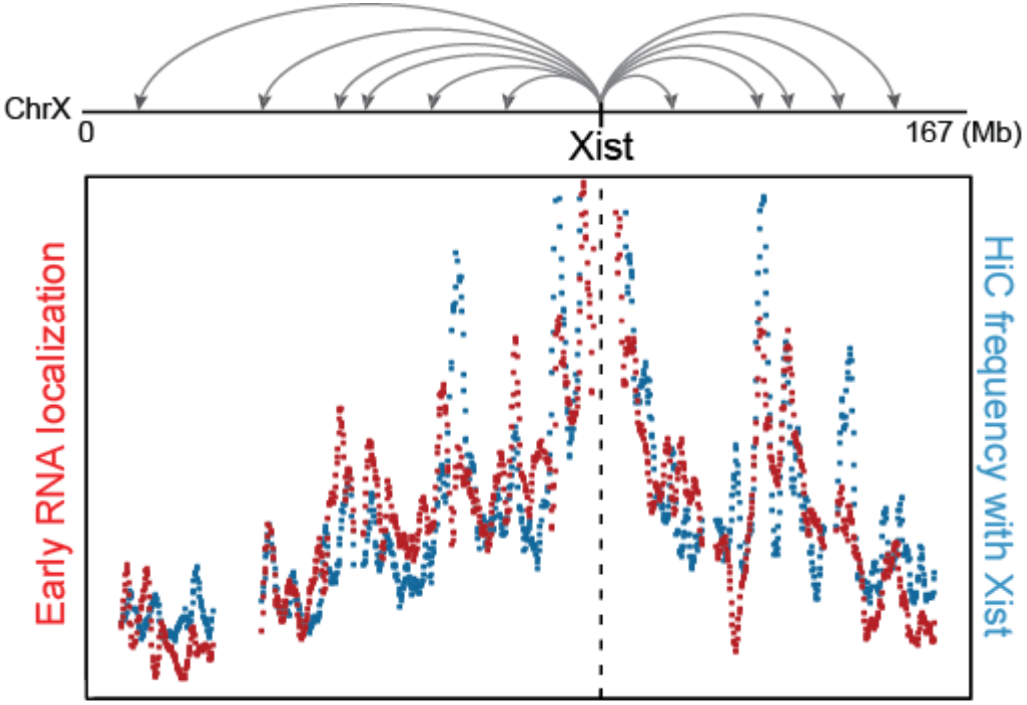


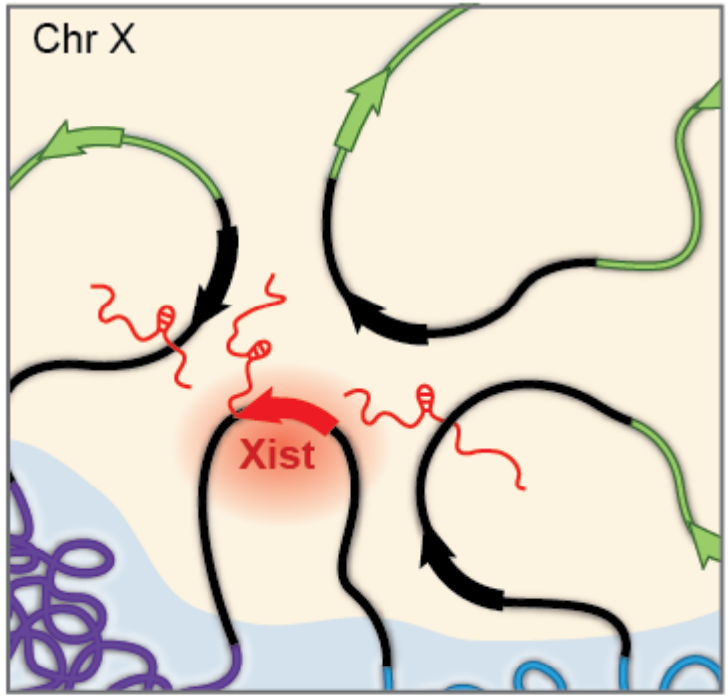


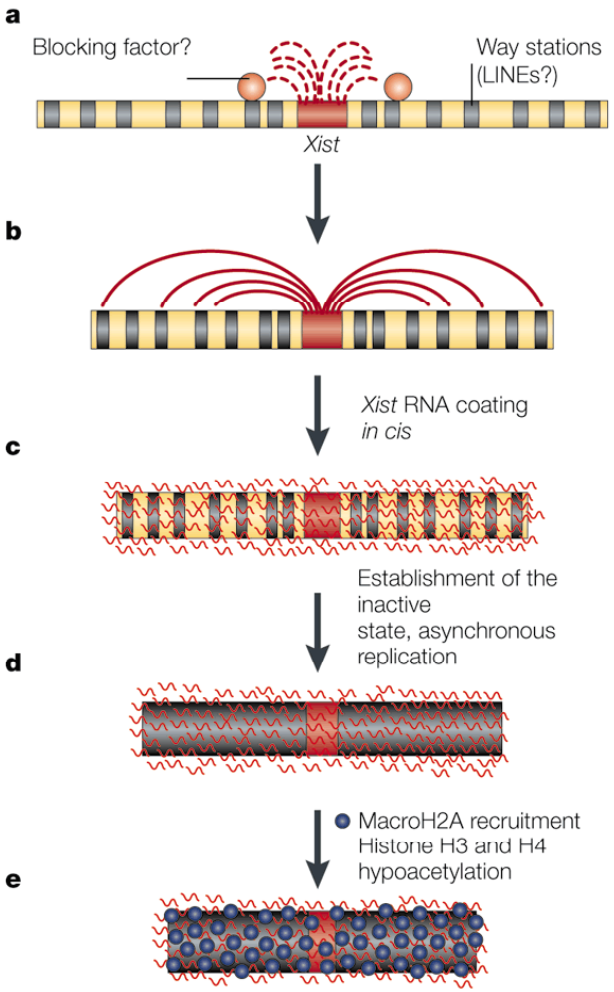
XIST functions by a poorly understood “spreading” mechanism from “Chromosome Entry Sites”



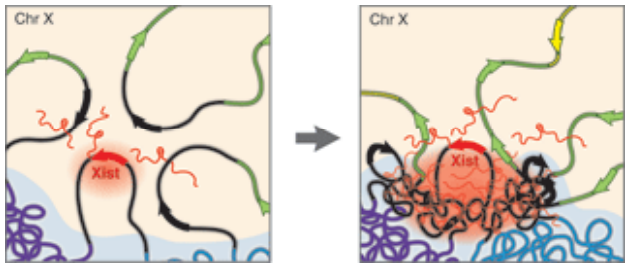
Entry sites may be determined by the 3 dimensional architecture of the genome

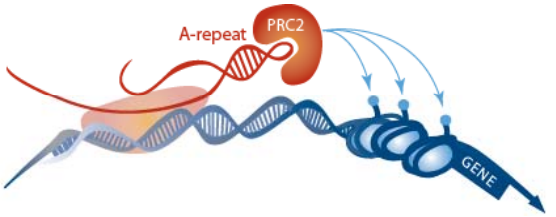




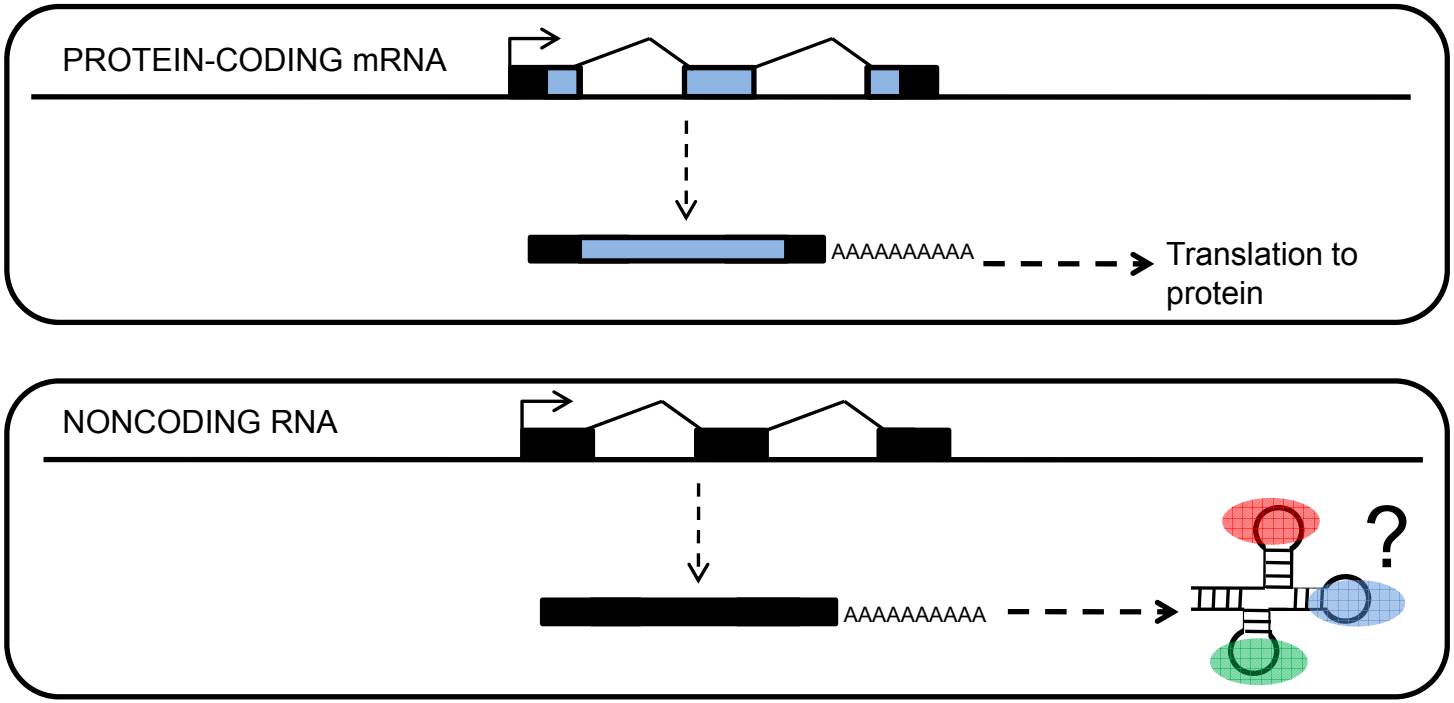


Nature Reviews | Genetics





What does a lncRNA look like?



Overview of lncRNA

- **LncRNA are mRNA-like transcripts that do NOT encode a protein product**

- Our genome contains >10,000 lncRNA
- We know almost nothing about 99% of them
- They are under evolutionary conservation, but weaker
- They are found in all multicellular species
- A few very convincing examples have been studied in detail – we do not know how much we can infer from these.

WARNING!

Definition of lncRNA: any transcript >200nt with no protein coding capacity

This definition is imperfect, and not based on function

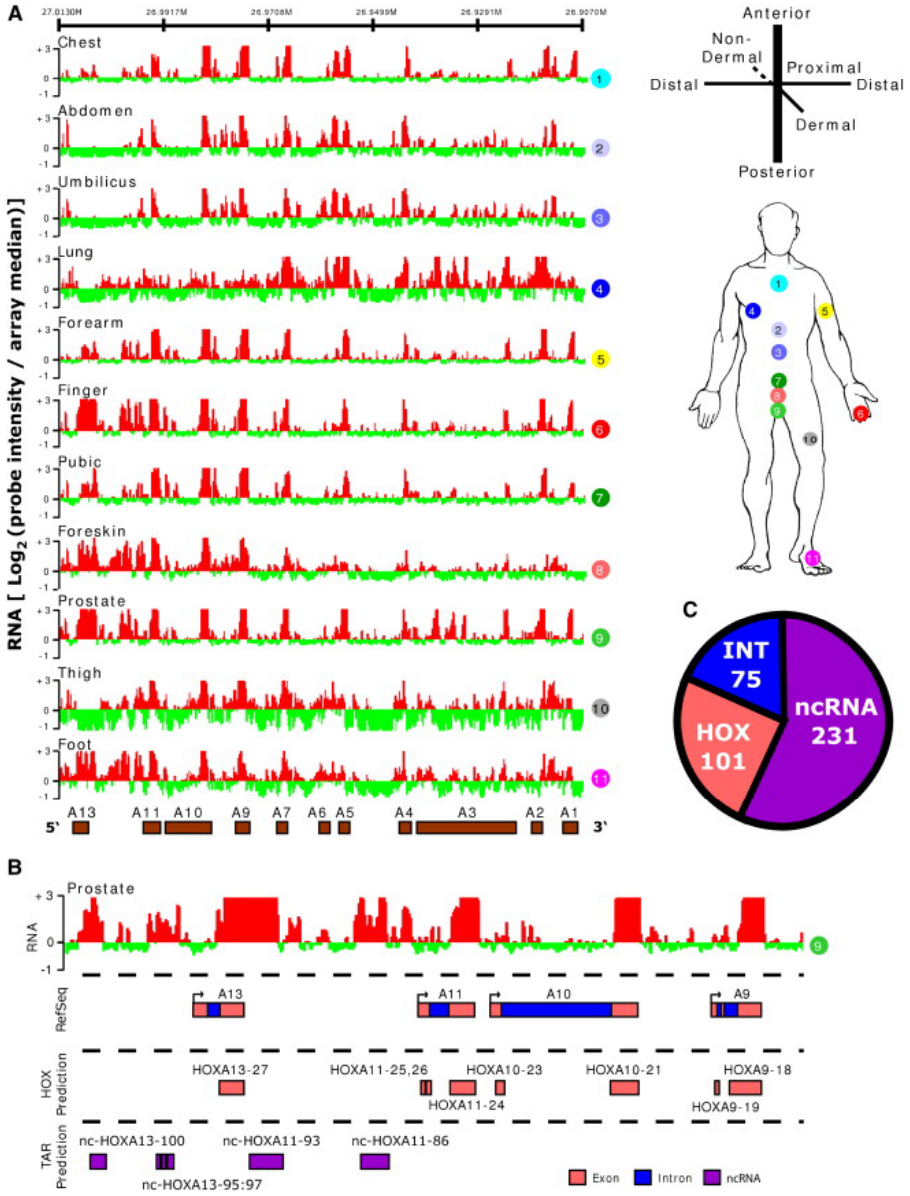
It has been argued that some / many / all lncRNAs are “transcriptional noise”

We lack most of the fundamental tools that are available to proteins: evolutionary signature, secondary structure prediction, tertiary structure prediction.

Key Topics in LncRNA

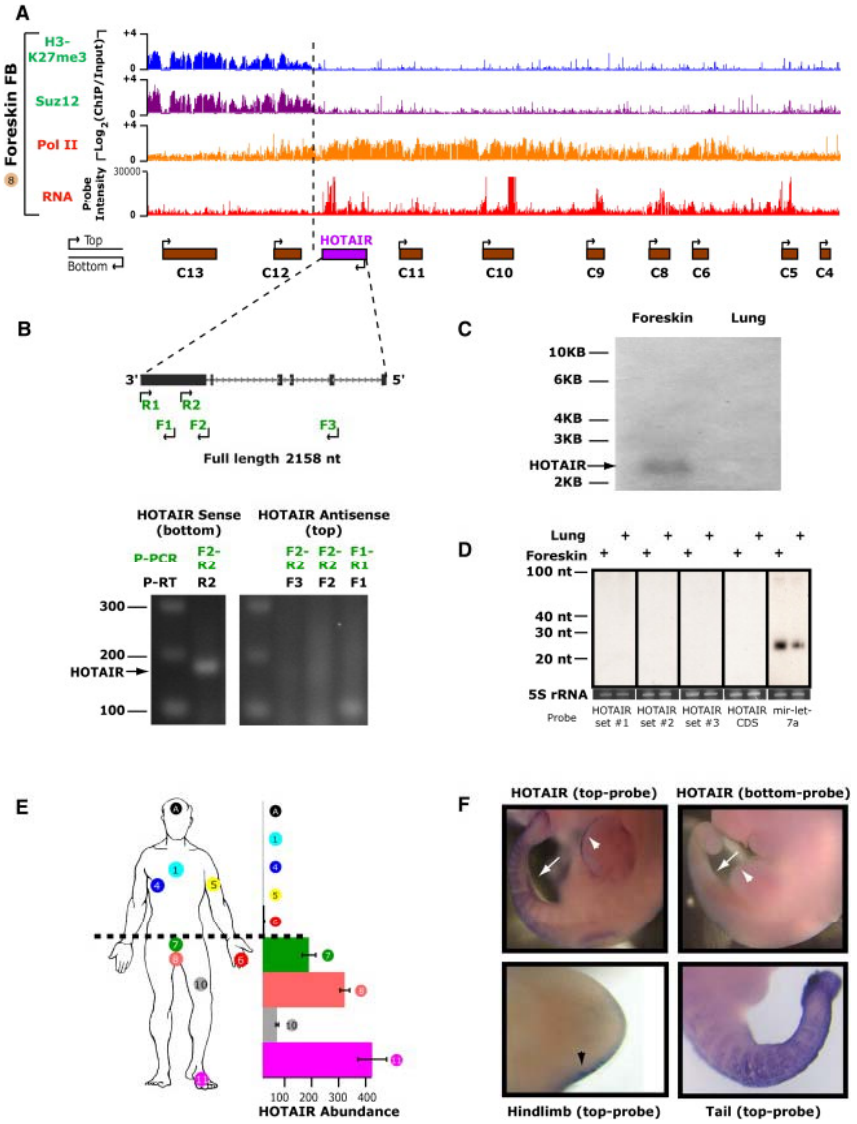
- Creating catalogues
- Effective discrimination of protein-coding/NONcoding
- Evolutionary conservation
- Cellular localisation
- Biological function
- Molecular mechanisms
- Roles in disease

HOTAIR
(*HO*X *A*ntisense *I*nter genic *R*NA): a paradigm for epigenetic regulatory lncRNAs



Rinn et al., PMID17604720

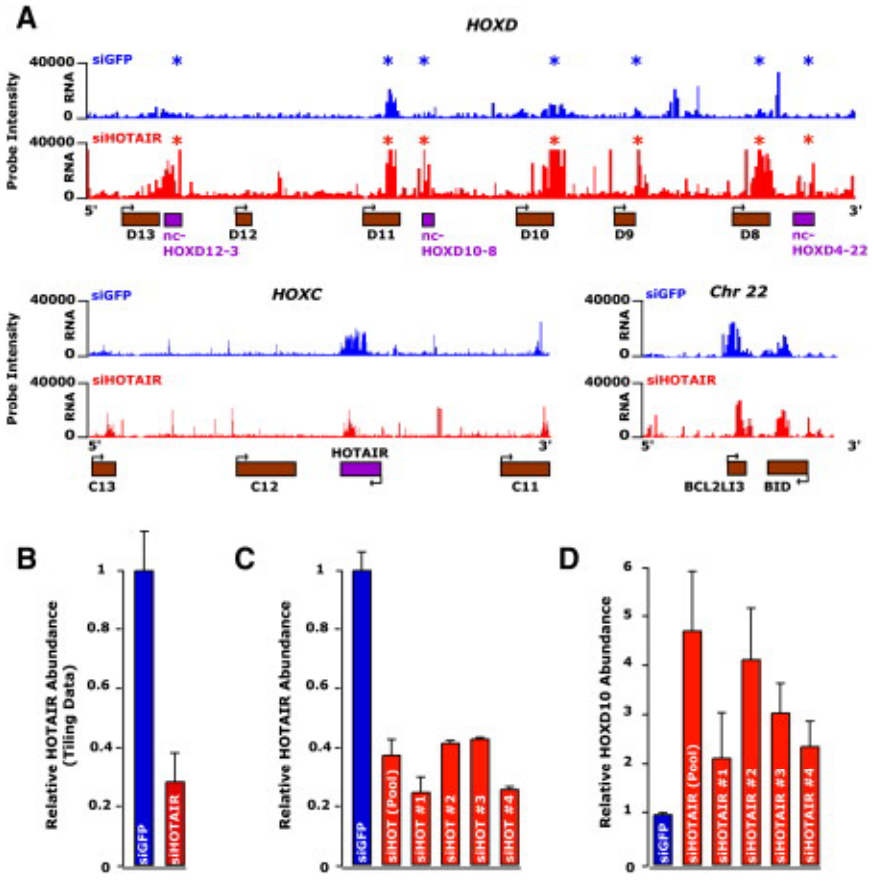
HOTAIR is highly expressed in the lower half of the body



HOTAIR represses other genes in *trans*

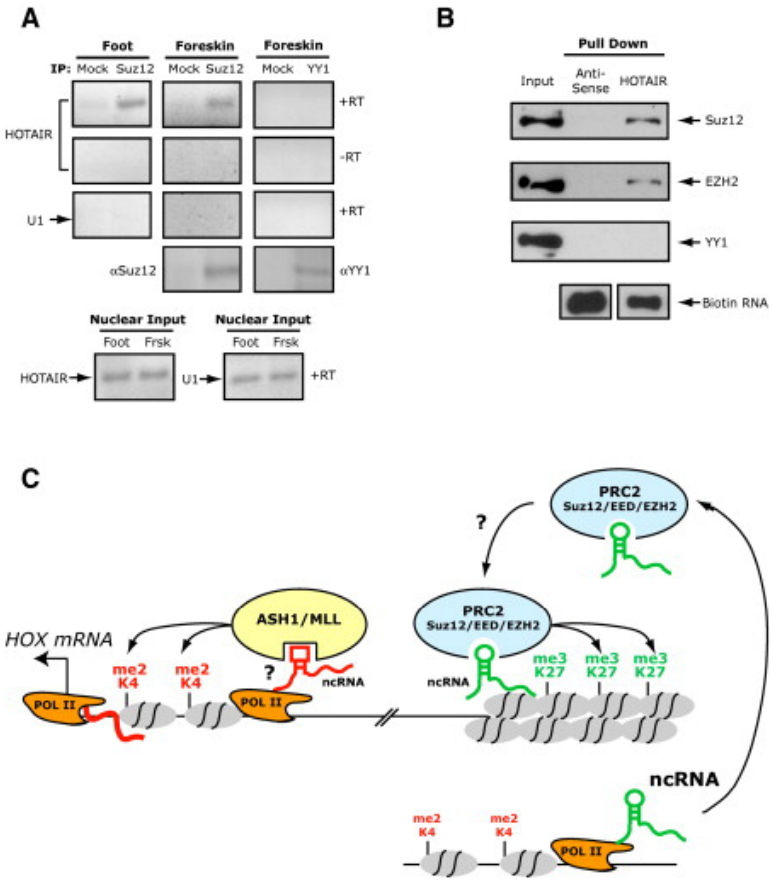
Ch2

Ch12

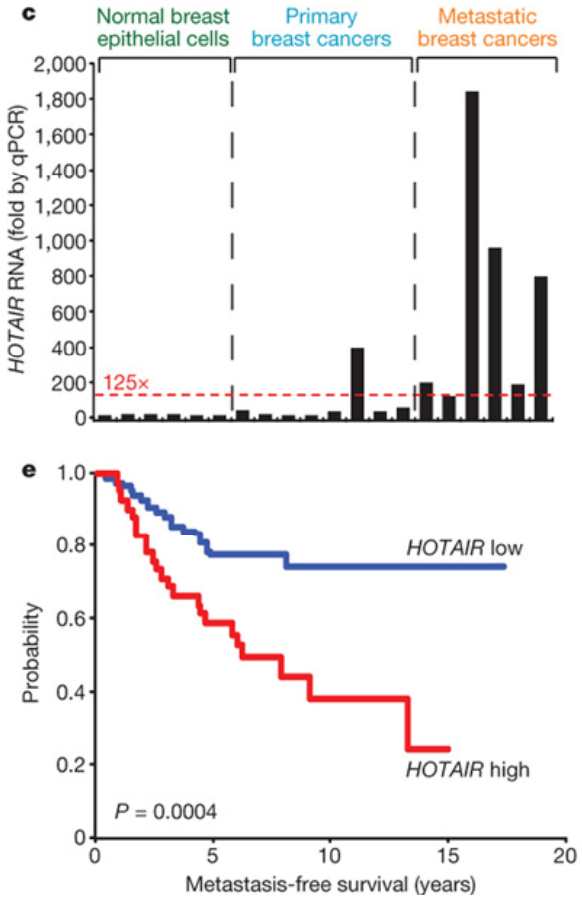


HOTAIR recruits epigenetic regulatory proteins

RNA
Immunoprecipitation



HOTAIR is an oncogene in breast cancer



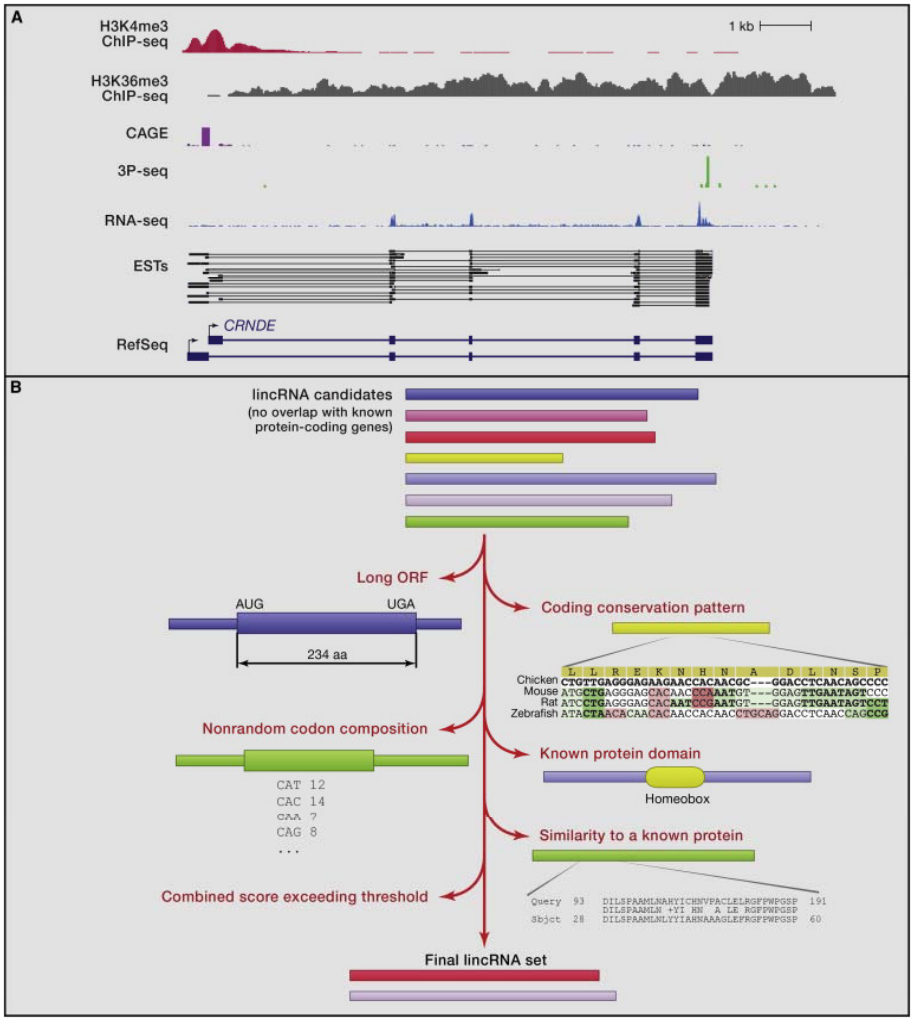
Gupta et al., PMID 20393566

Our aim: to discover how many more genes like this exist

Approach:

1. Develop a good catalogue of lncRNAs
2. Use functional genomics approaches to discover functional lncRNAs

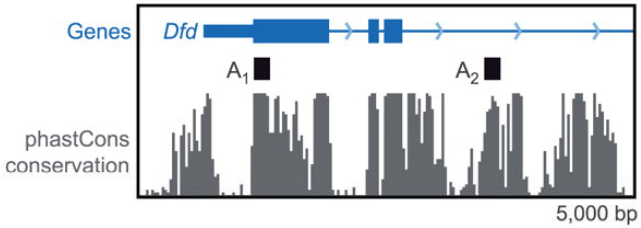
How to discover lncRNA – Annotations / Catalogues etc

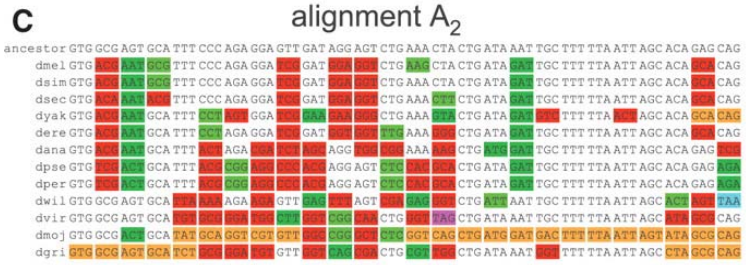


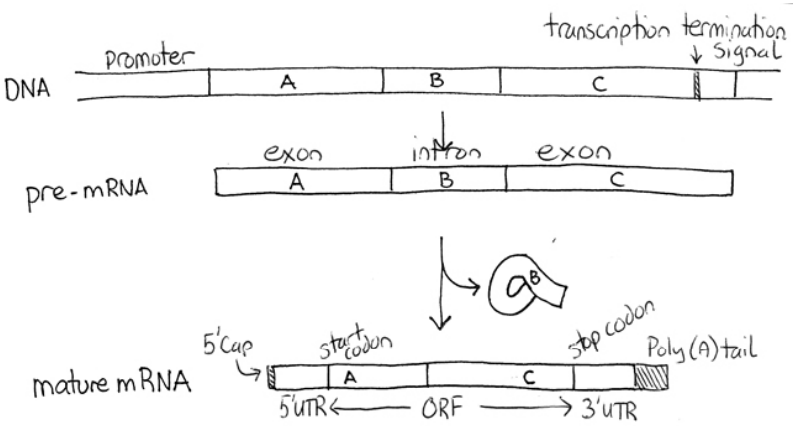
How to work out if something encodes protein?

1. ORF (Open Reading Frame) features: length / similarity to other proteins eg CPC

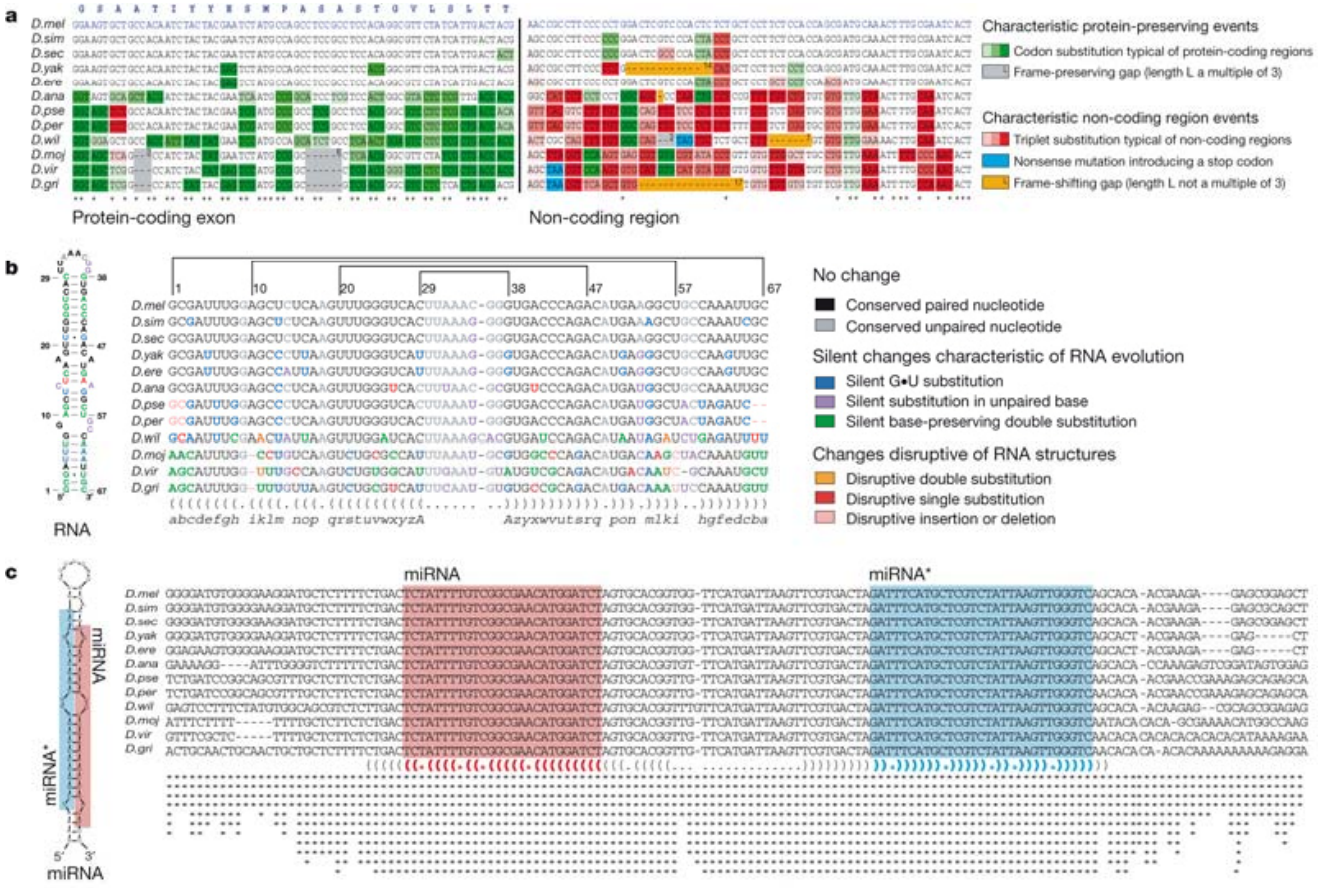
2. Evolutionary signatures eg PhyloCSF

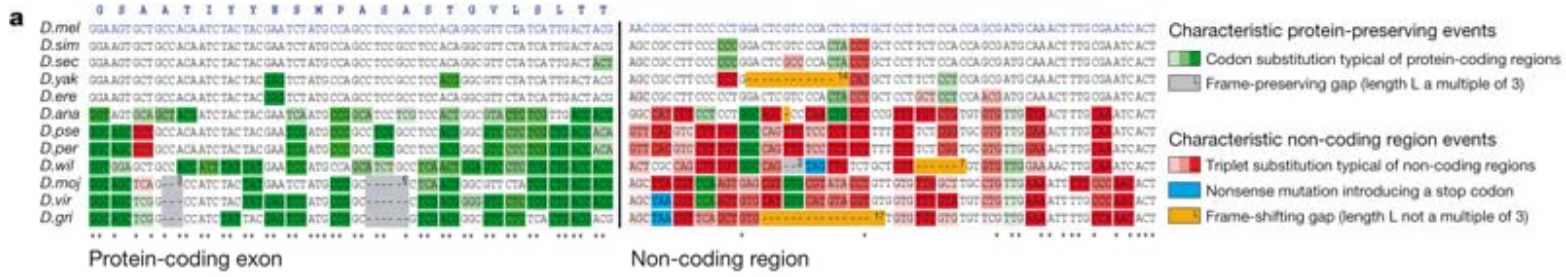


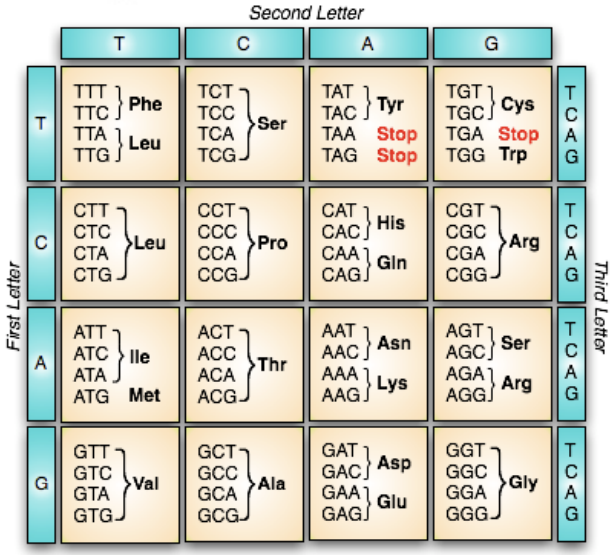




Different types of DNA evolve in different ways



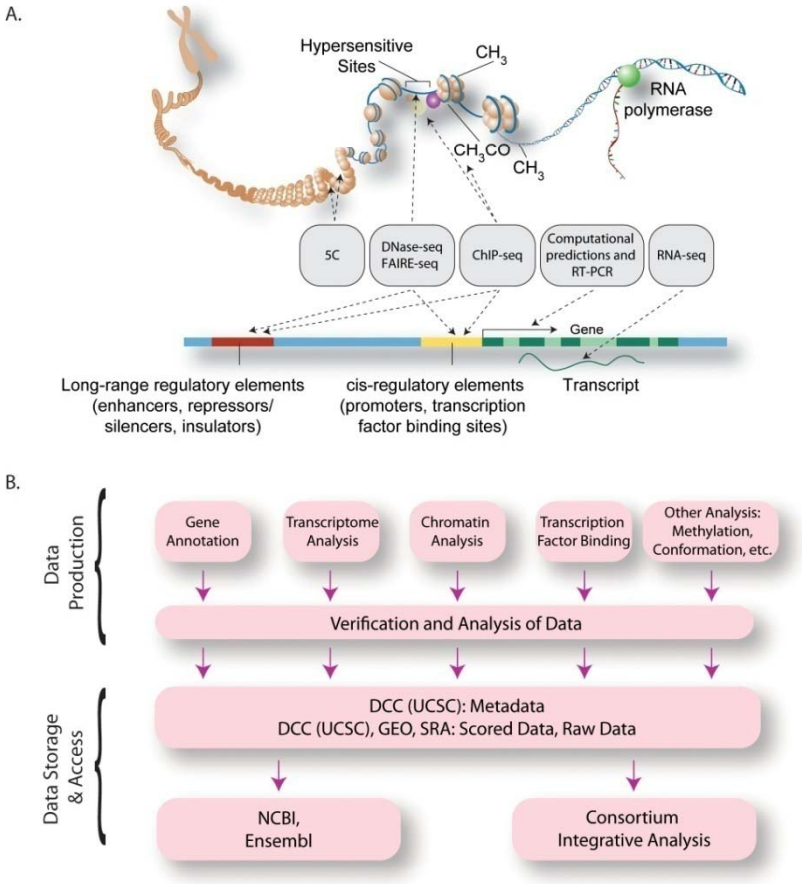

a
D.mel
D.sim
D.sec
D.yak
D.ere
D.ana
D.pse
D.per
D.wil
D.moj
D.vir
D.gri
Protein-coding exon
Non-coding region
Characteristic protein-preserving events
Codon substitution typical of protein-coding regions
Frame-preserving gap (length L a multiple of 3)
Characteristic non-coding region events
Triplet substitution typical of non-coding regions
Nonsense mutation introducing a stop codon
Frame-shifting gap (length L not a multiple of 3)


Second Letter
First Letter
Third Letter
T
C
A
G
TTT Phe
TTC Phe
TTA Leu
TTG Leu
TCT Ser
TCC Ser
TCA Ser
TCG Ser
TAT Tyr
TAC Tyr
TAA Stop
TAG Stop
TGT Cys
TGC Cys
TGA Stop
TGG Trp
CTT Leu
CTC Leu
CTA Leu
CTG Leu
CCT Pro
CCC Pro
CCA Pro
CCG Pro
CAT His
CAC His
CAA Gln
CAG Gln
CGT Arg
CGC Arg
CGA Arg
CGG Arg
ATT Ile
ATC Ile
ATA Ile
ATG Met
ACT Thr
ACC Thr
ACA Thr
ACG Thr
AAT Asn
AAC Asn
AAA Lys
AAG Lys
AGT Ser
AGC Ser
AGA Arg
AGG Arg
GTT Val
GTC Val
GTA Val
GTG Val
GCT Ala
GCC Ala
GCA Ala
GCG Ala
GAT Asp
GAC Asp
GAA Glu
GAG Glu
GGT Gly
GGC Gly
GGA Gly
GGG Gly

The ENCODE Project: ENCyclopedia Of DNA Elements

Funded by NIH (National Institutes of Health, USA) www.genome.gov

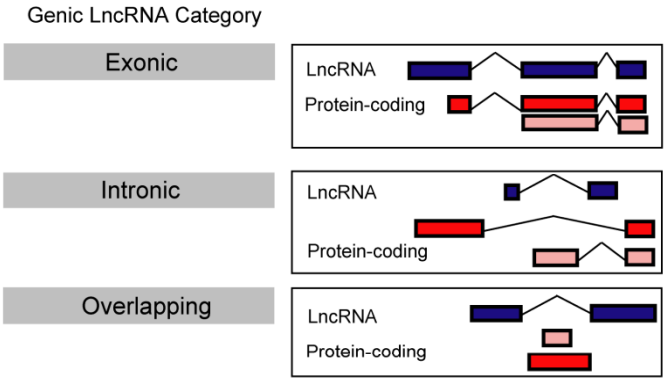
Aim: “to identify all functional elements in the human genome sequence”



The GENCODE manually-curated human lncRNA catalogue:

9,277 gene loci producing **14,880** transcripts (Gencode version 7)
www.gencodegenes.org

B



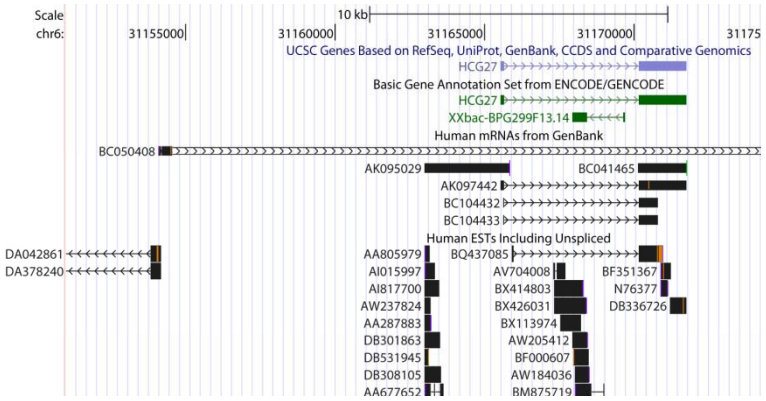
C

Gencode LncRNAs transcripts (14,880)								
Intergenic (9,520)			Genic (5,360)					
Same Strand	Convergent	Divergent	Exonic (2,409)		Intronic (2,784)		Overlapping (167)	
			S	AS	S	AS	S	AS
4,165	1,938	3,417	NA	2,409	563	2,221	52	115

Latest GENCODE version 14:

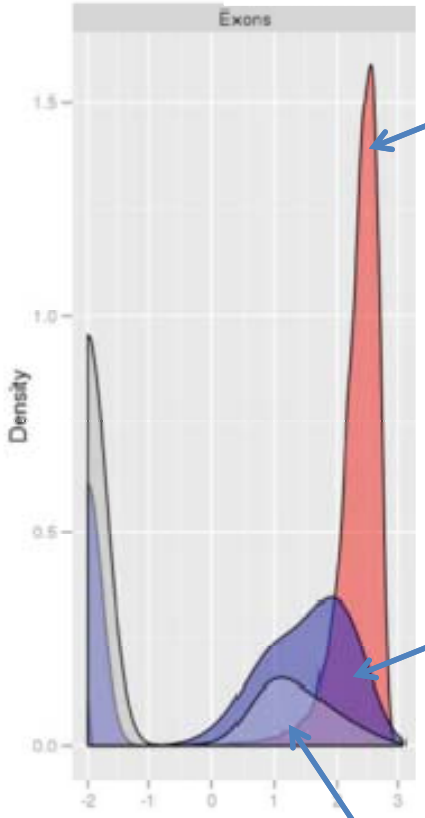
12,933 lncRNA gene loci
21,271 lncRNA transcripts

Based on manual annotation

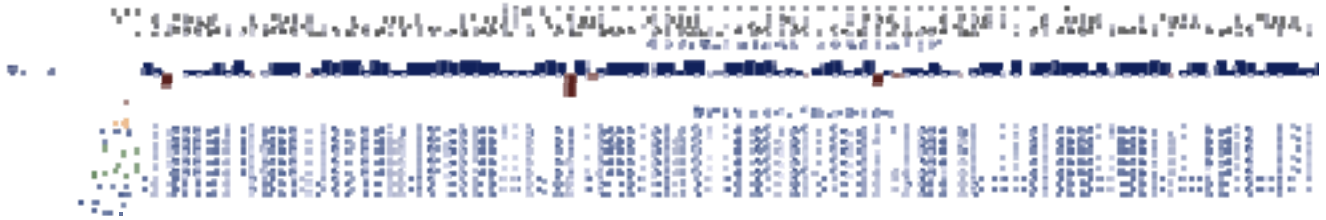


Derrien et al PMID 22955988

Using evolutionary conservation as evidence of function for LncRNAs



Protein coding genes – expected to be functional



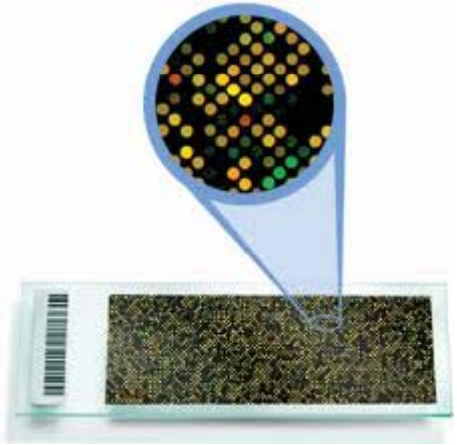
LncRNAs

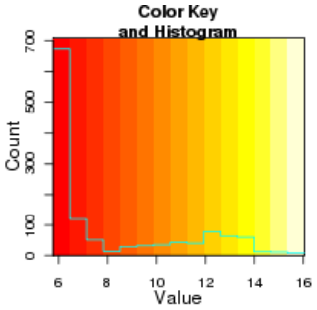
Ancestral repeats – expected to be non-functional

Derrien et al
PMID 22955988

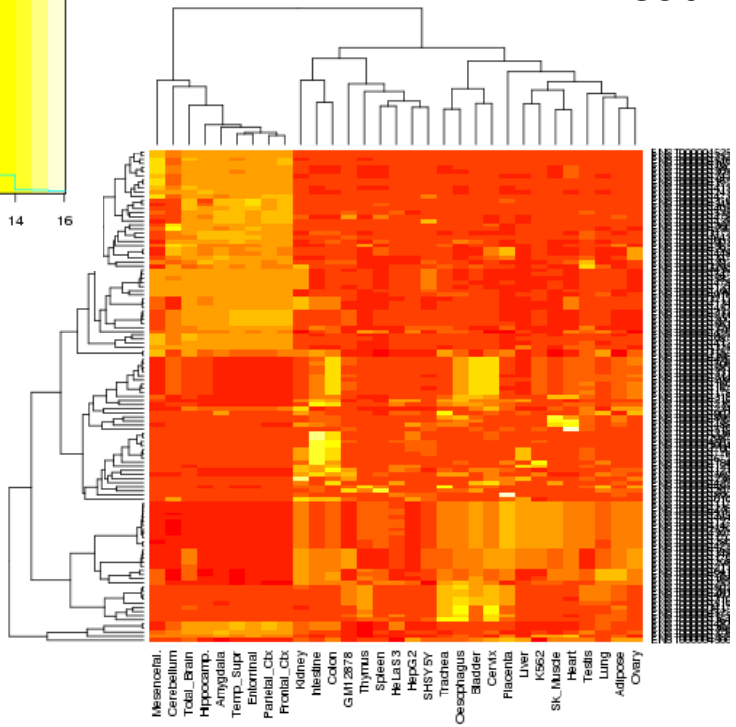
Microarray as a tool for detecting ncRNA expression

- Can monitor up to ~1 million genes in a single experiment
- Requires a known annotation of ncRNA – NOT unbiased like RNAseq
- Sensitive, reproducible, cheap, quick
- Commercial designs available for short ncRNAs
- Possible to create **custom RNA designs** to quantify new ncRNA gene sets

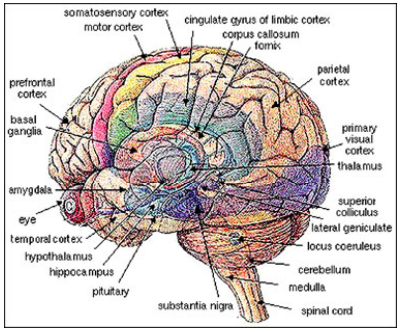


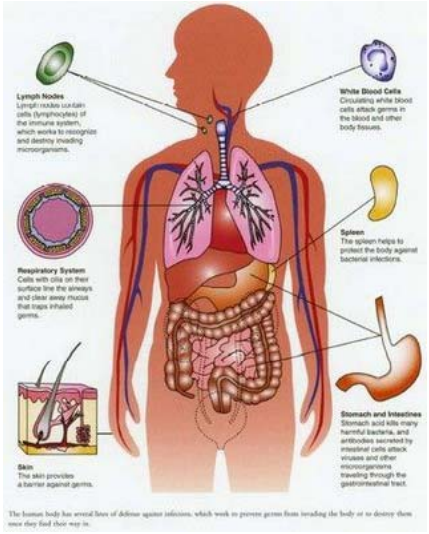


Heatmap

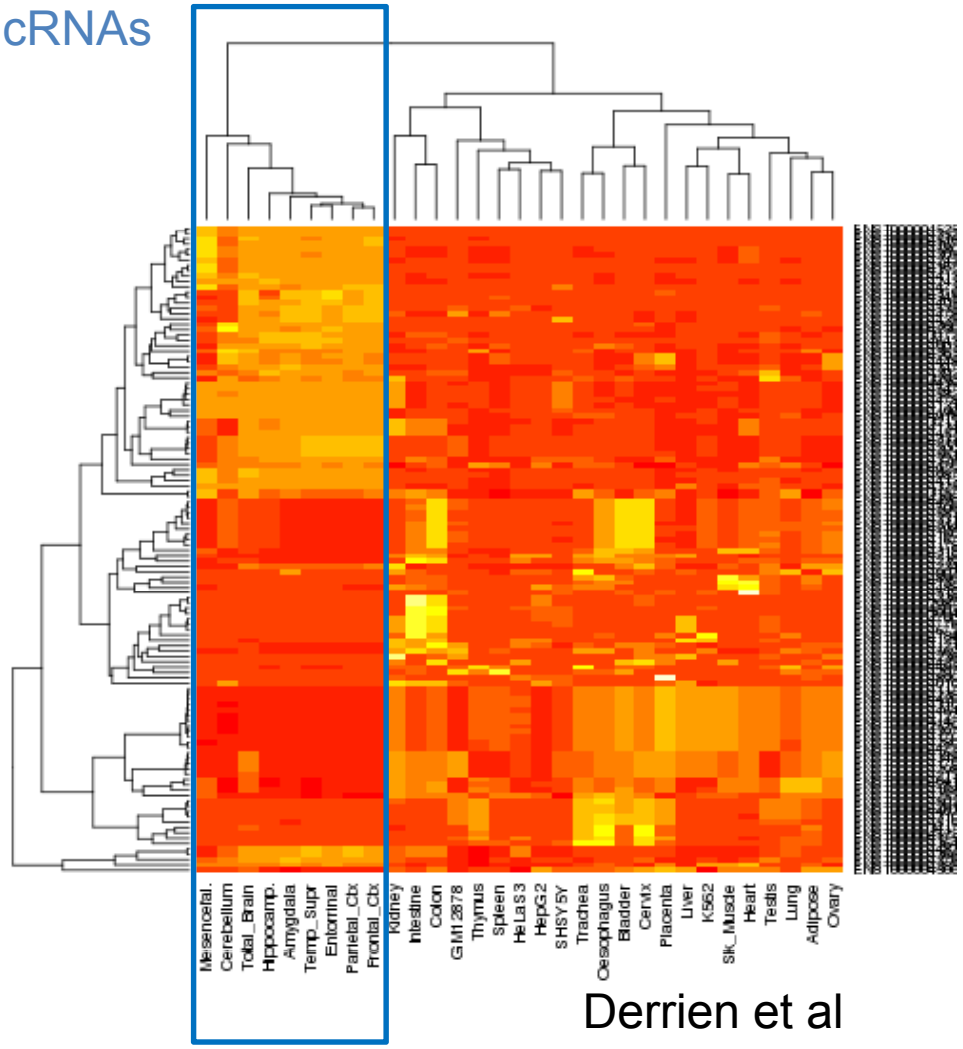


Expression maps of lncRNAs in the 31 tissues / cell lines (10 of neural origin)



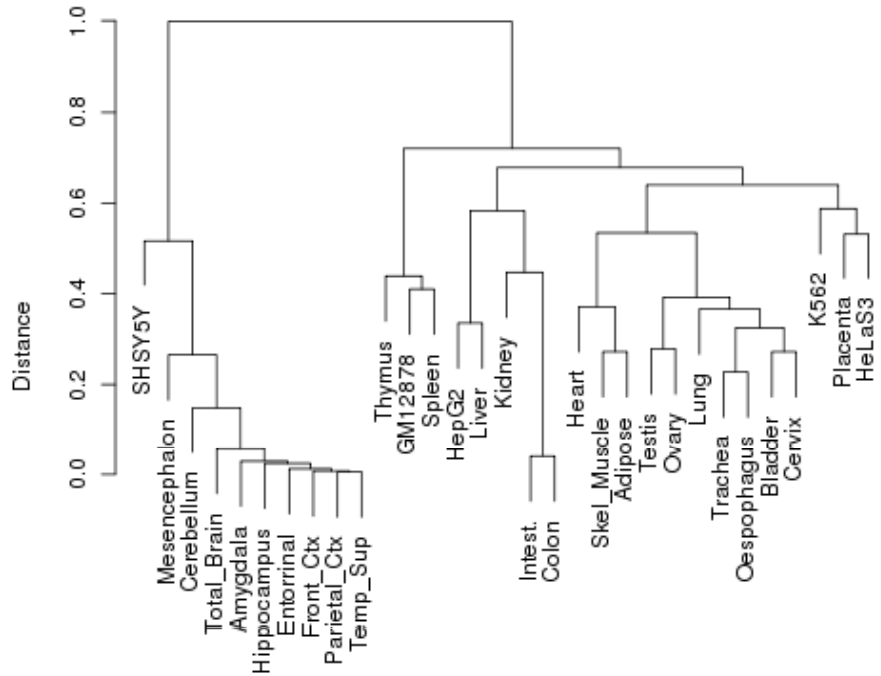


Brain-specific lncRNAs

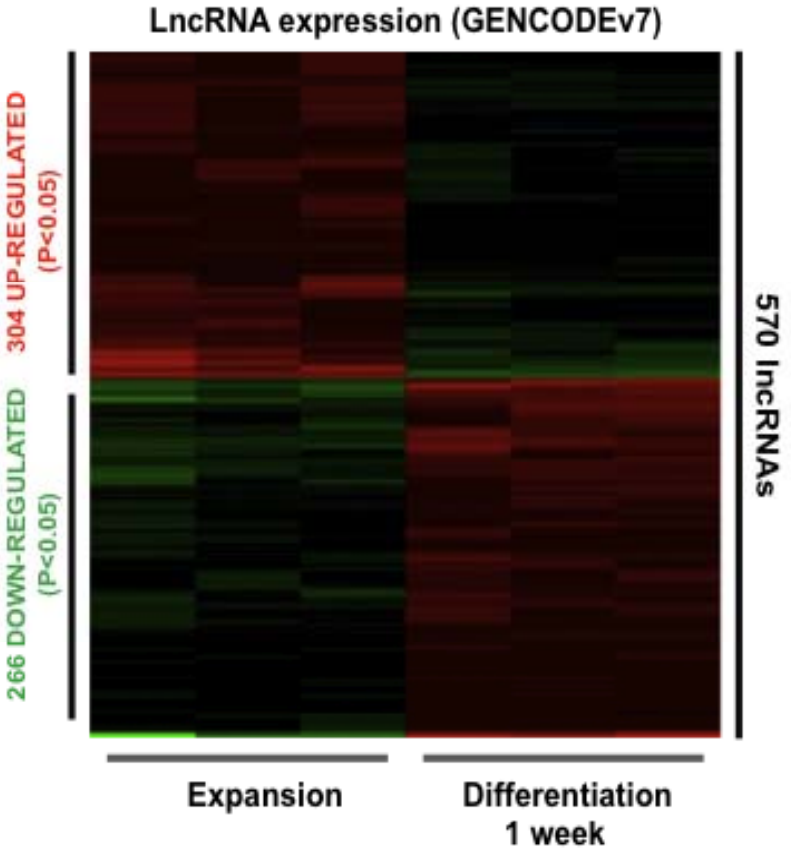


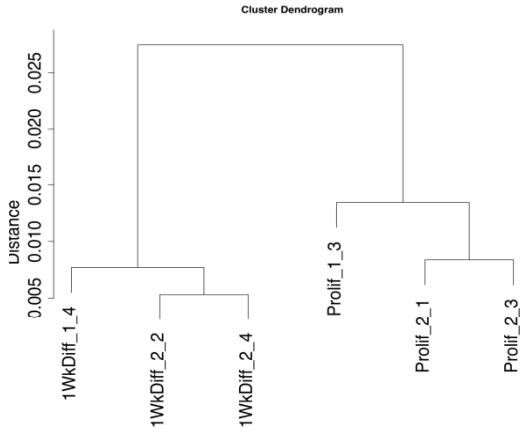
Derrien et al
PMID 22955988

Correct tissue clustering using lncRNAs

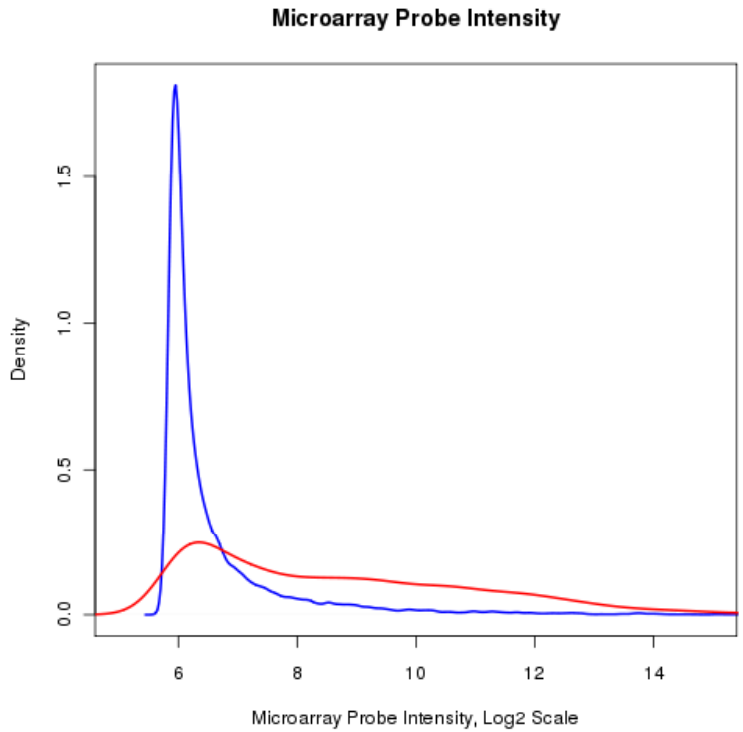


We can use microarrays to discover lncRNAs that change in disease or development

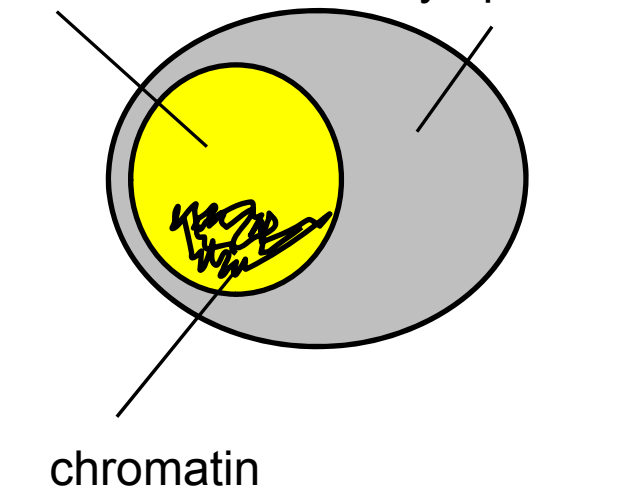




LncRNAs are expressed lower than protein-coding mRNAs

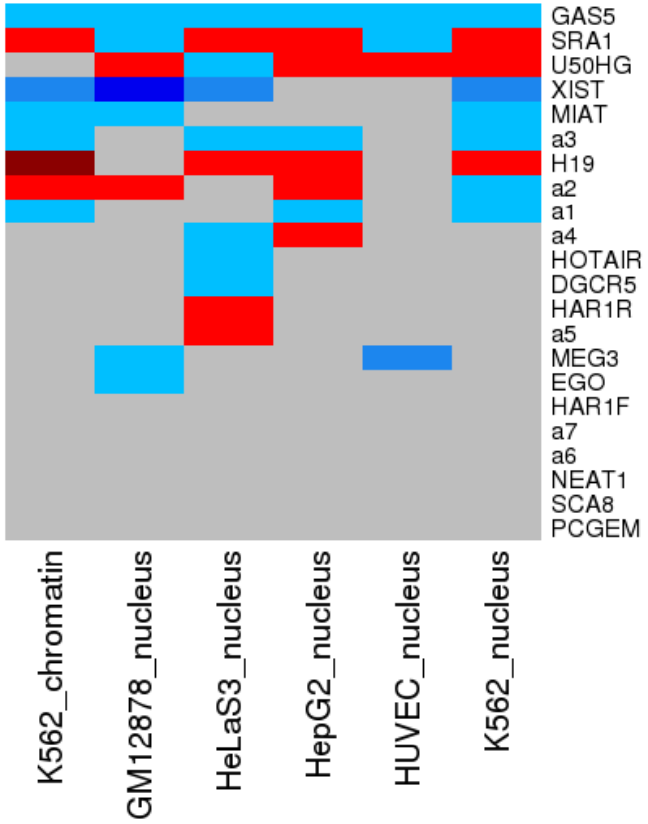


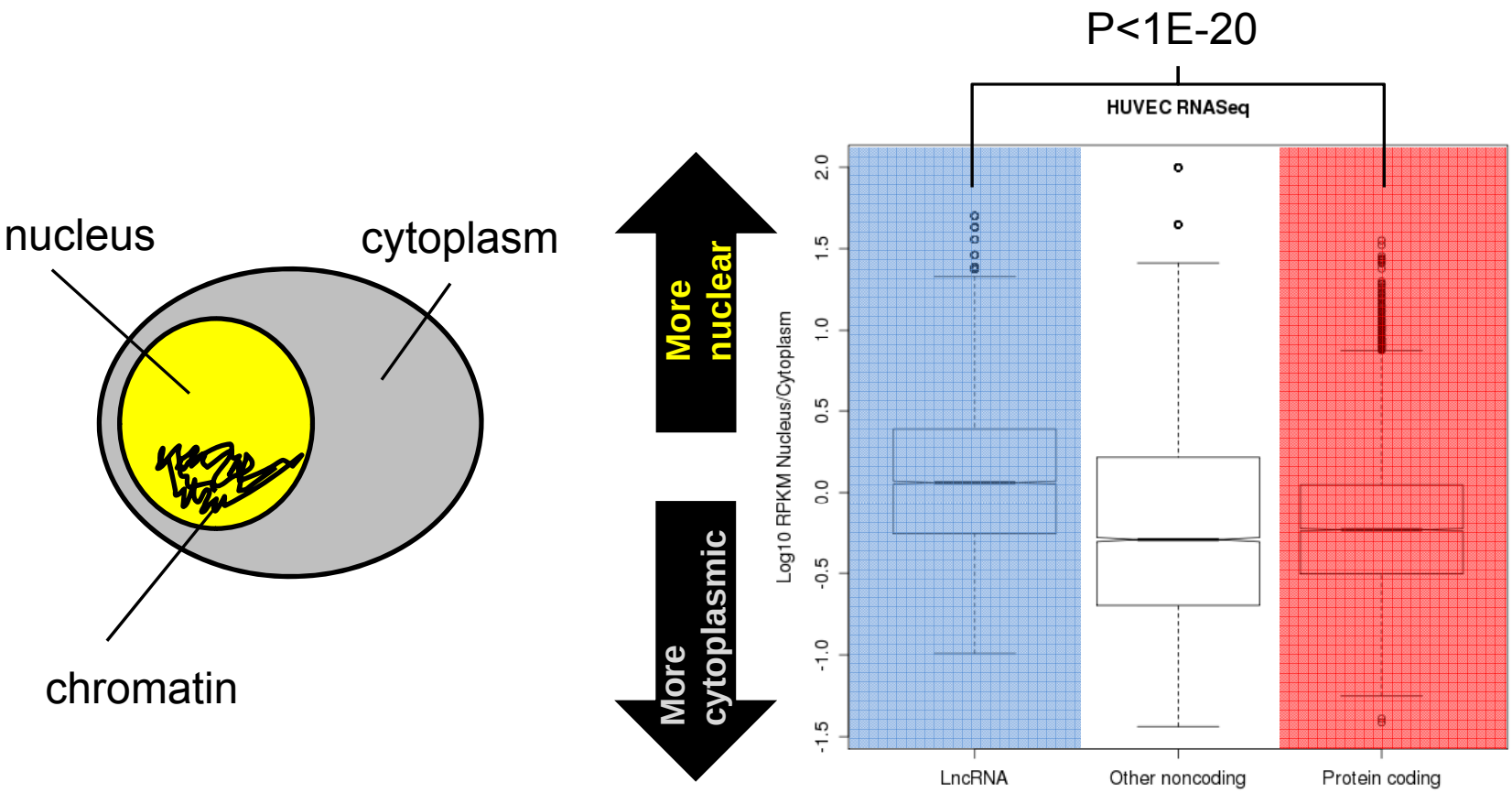
Where are lncRNAs in the cell?



More nuclear

More cytoplasmic



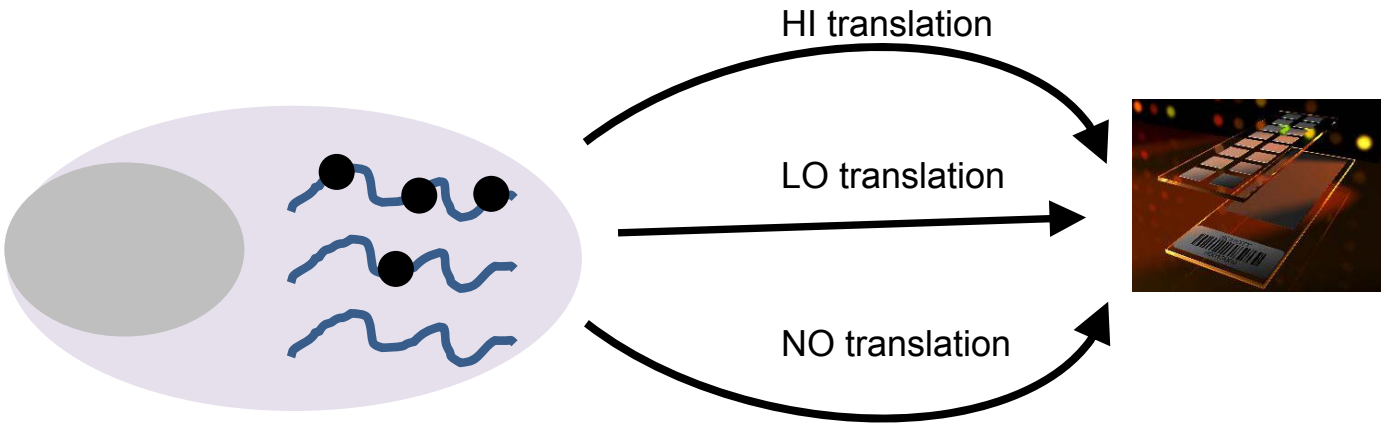


ENCODE consortium,
Tom Gingeras

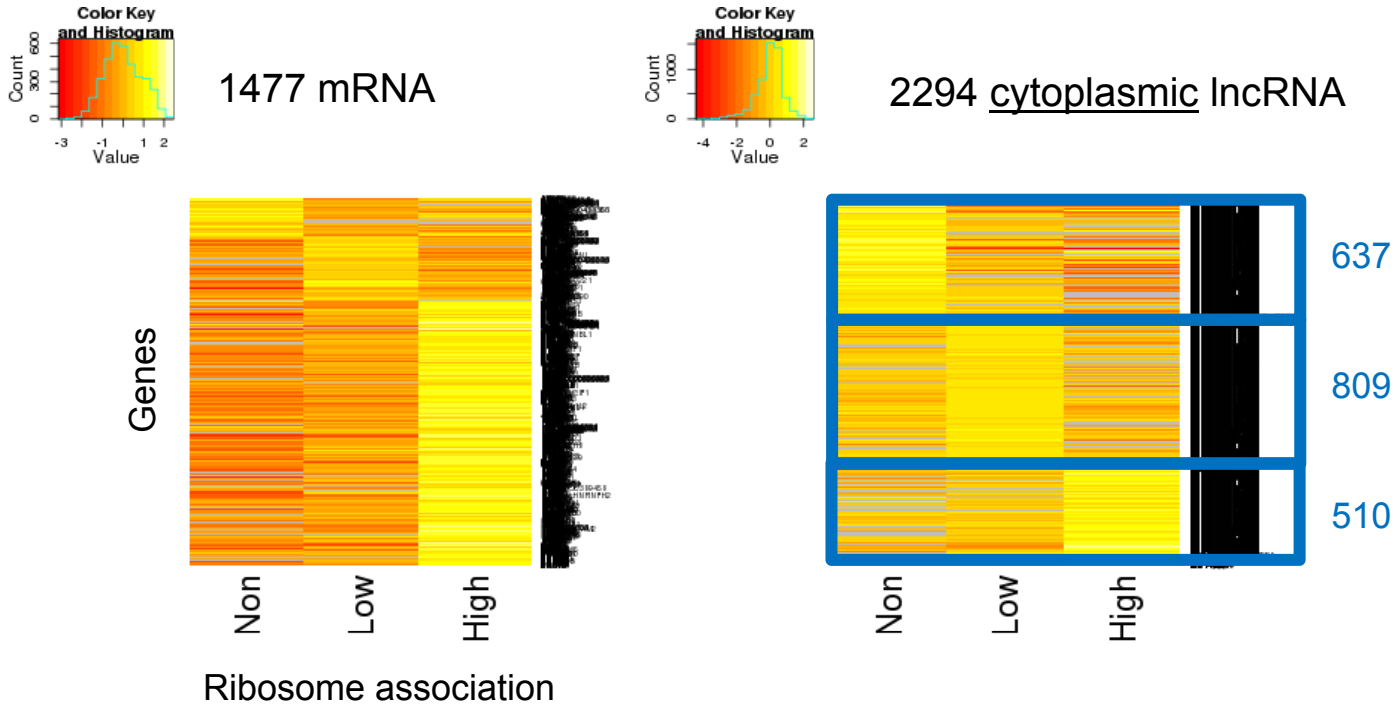
Do lncRNA have roles outside the nucleus?

Ribosome profiling of lncRNA

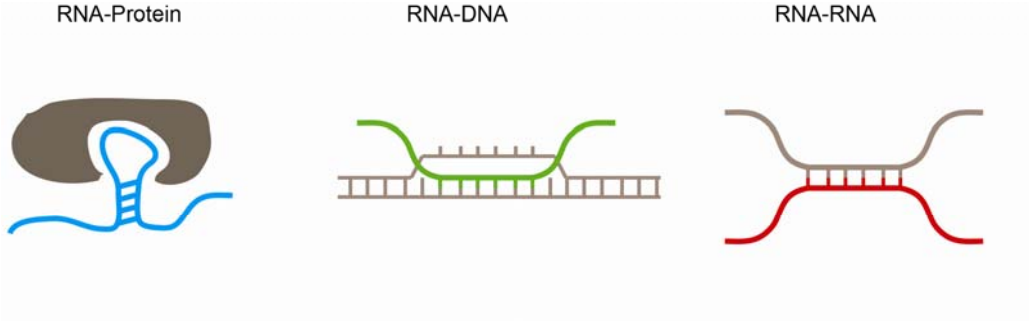
1. Purify ribosome-bound RNA by sucrose gradient ultracentrifugation
2. Hybridise to custom lncRNA arrays
3. Bioinformatic analysis to discover lncRNA associated with ribosomes
4. Search for clues as to their function

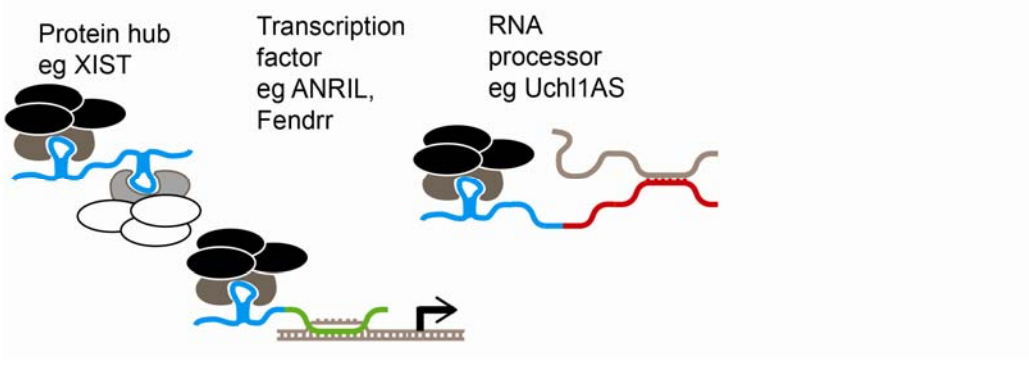


Discovery of ribosome-associated lncRNAs

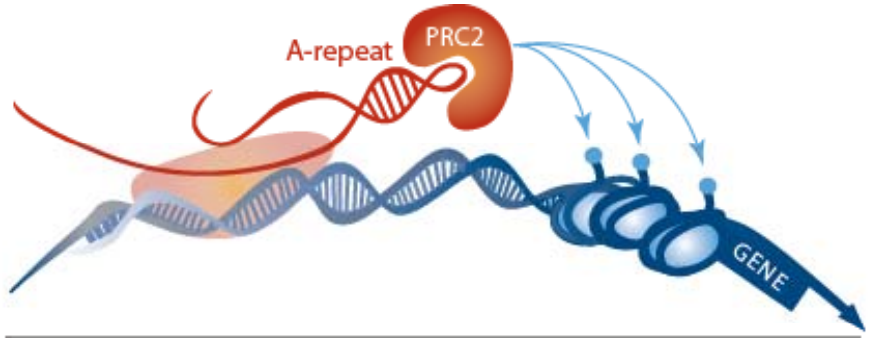


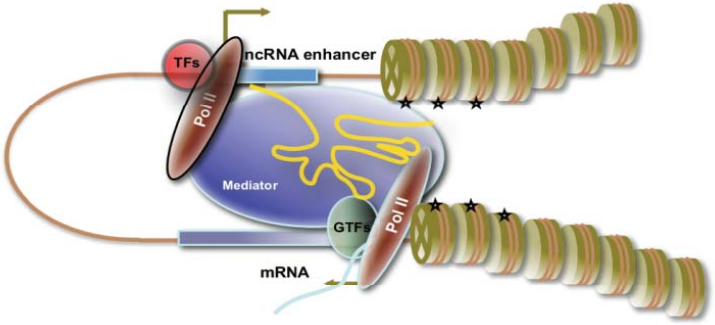
Molecular Mechanisms



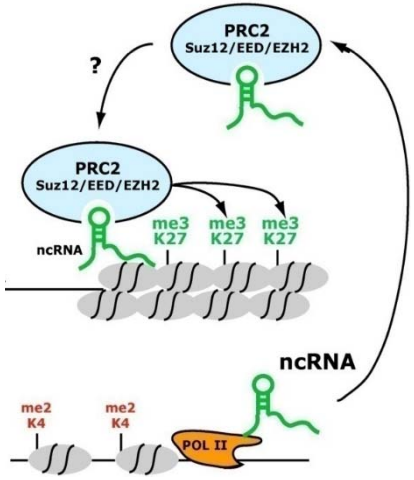


Example: epigenetic gene regulation



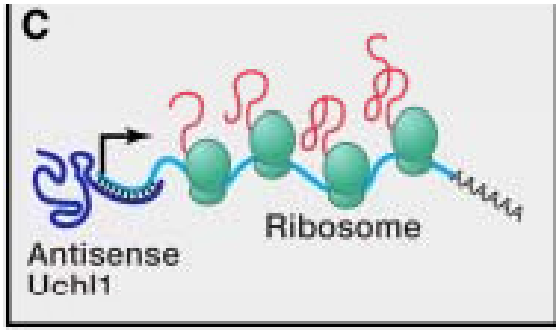


Fan et al, Nature 2013



Rinn et al, Cell 2007

Example: Regulation of mRNA translation by Uchl1-as



Batista and Chang, 2013

LETTER

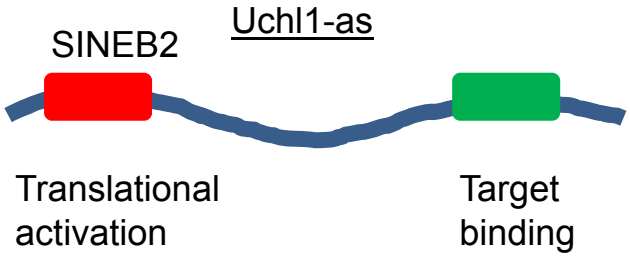
doi:10.1038/nature11508

Long non-coding antisense RNA controls *Uchl1* translation through an embedded SINEB2 repeat

Claudia Carrieri[1]*, Laura Cimatti[1]*, Marta Biagioli[1,2], Anne Beugnet[3], Silvia Zucchelli[1,2], Stefania Fedele[1], Elisa Pesce[3], Isidre Ferrer[4], Licio Collavin[5,6], Claudio Santoro[7], Alistair R. R. Forrest[8], Piero Carninci[8], Stefano Biffo[3,9], Elia Stupka[10] & Stefano Gustincich[1,2]

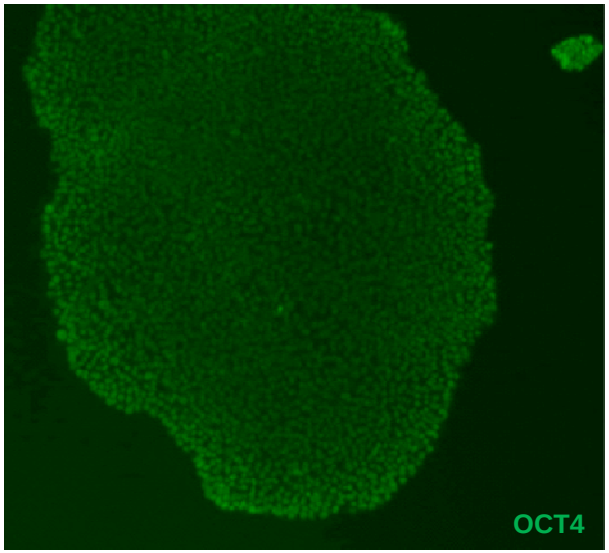
Capable of sequence specific regulation of:

RNA stability
RNA translation

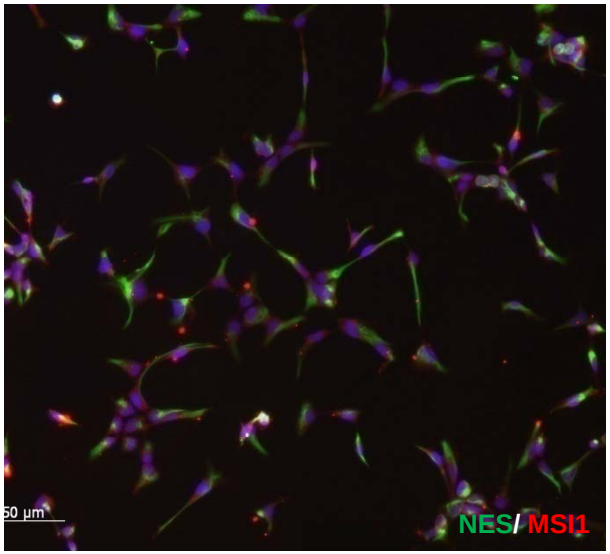


Functional Genomics Approach to LncRNAs: Identifying lncRNAs in human neuronal differentiation

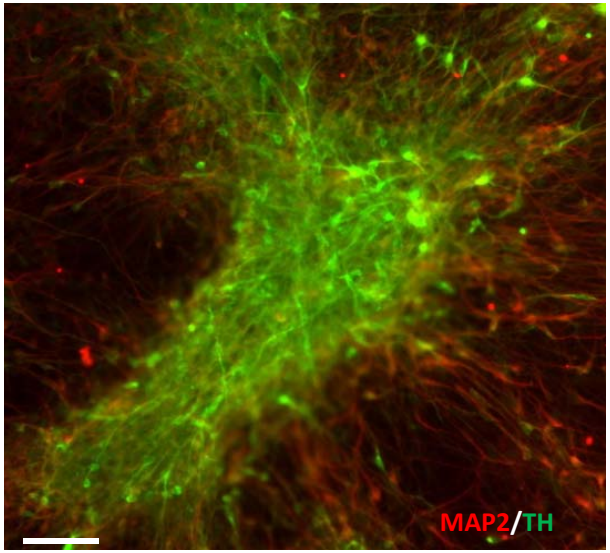
hESCs→ neural progenitors (NPCs) →dopaminergic neurons



97.5% OCT4^{+}



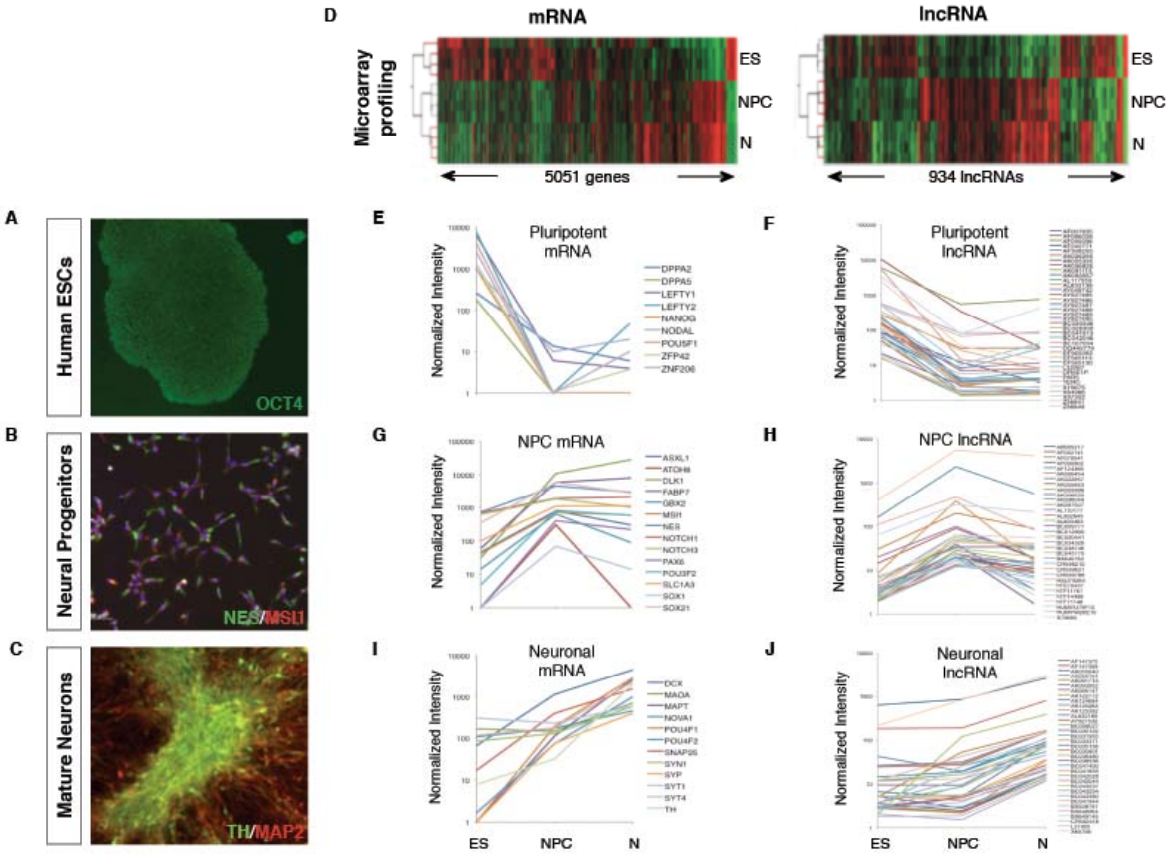
99.6% NESTIN^{+}
98.9% VIMENTIN^{+}
86.0% BLBP^{+}



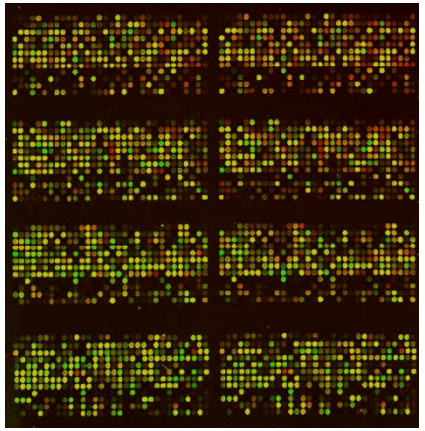
90% MAP2^{+}
85.3% TH^{+}/MAP2^{+}

Ng SY et al EMBO J
PMID 22193719

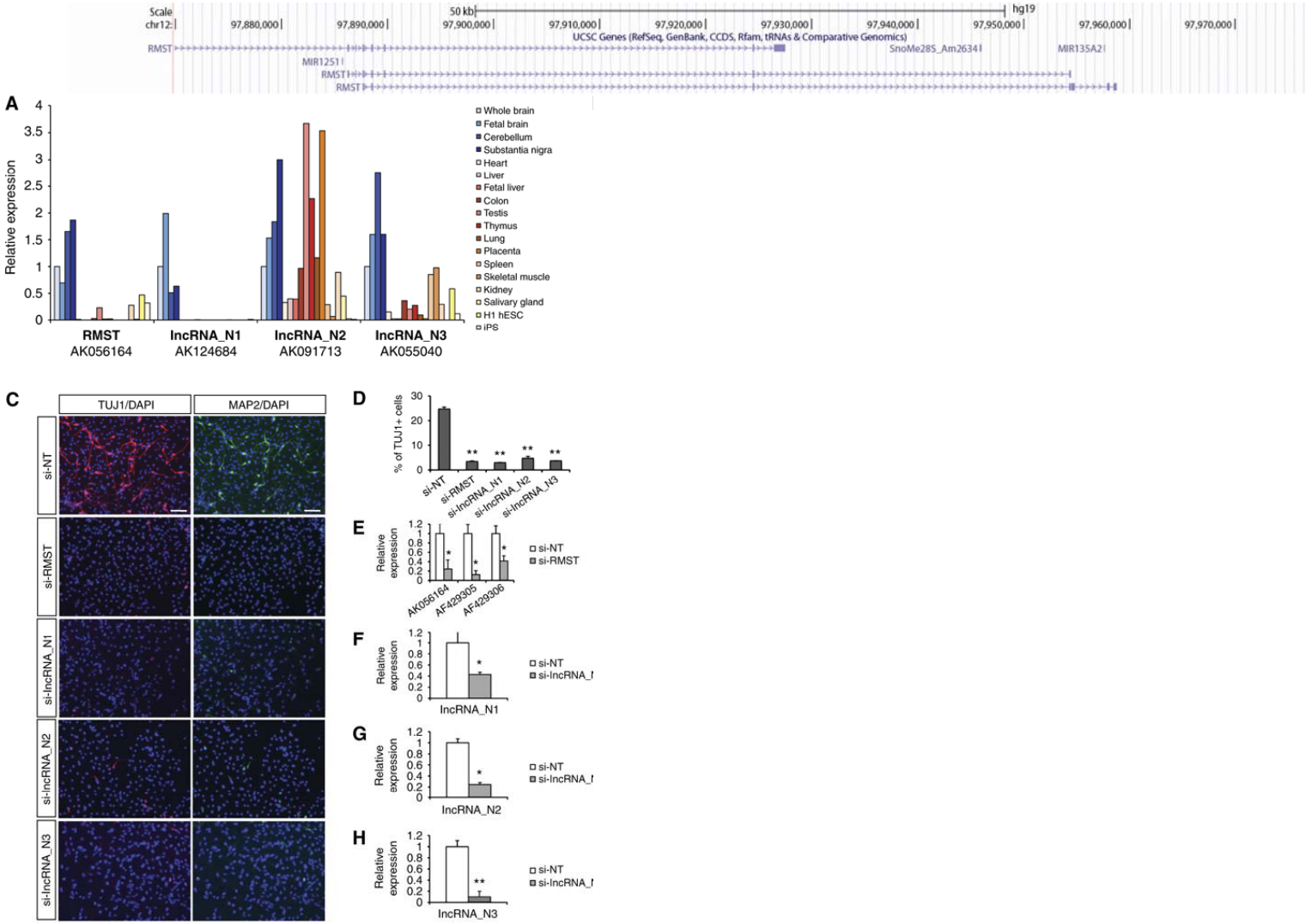
Genome-wide identification of neural lncRNAs



lncRNA microarray

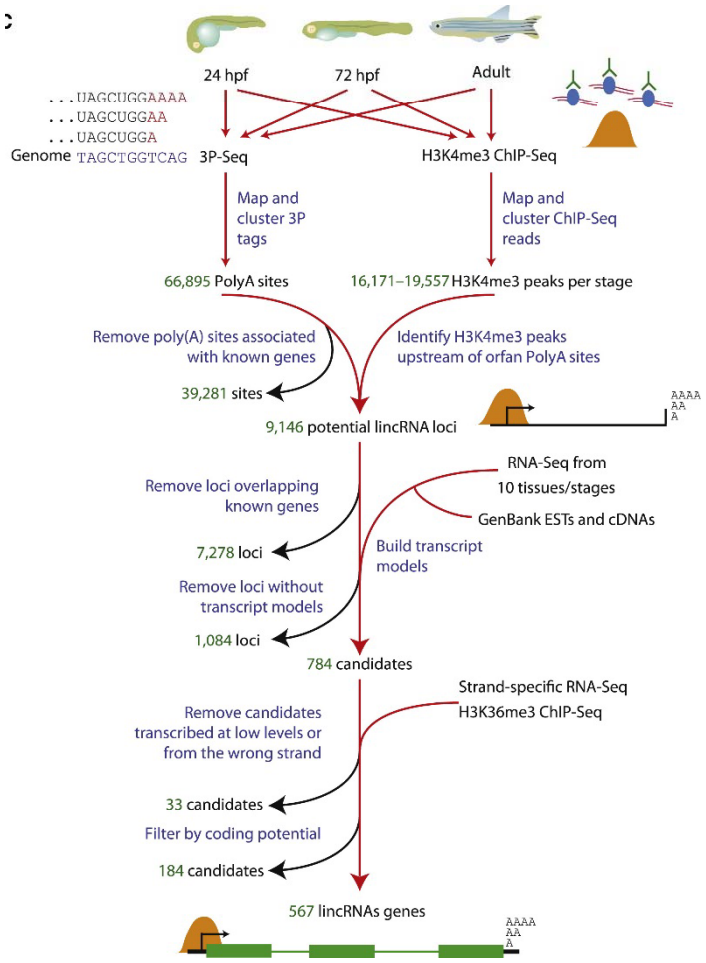


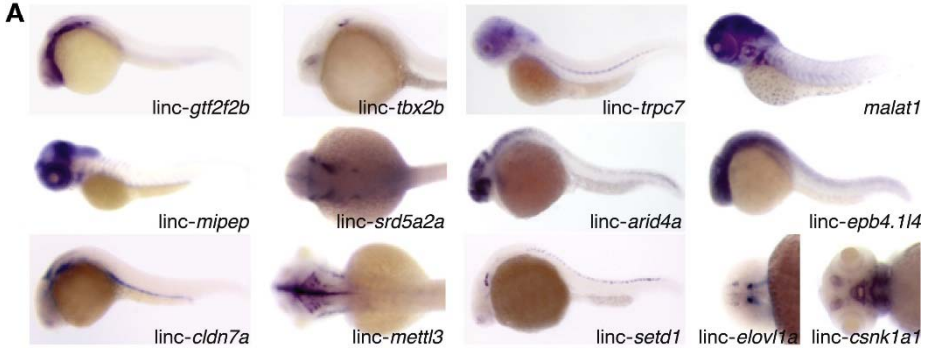
Discovery of lncRNAs necessary for dopaminergic neuron production



Ng SY et al EMBO J
PMID 22193719

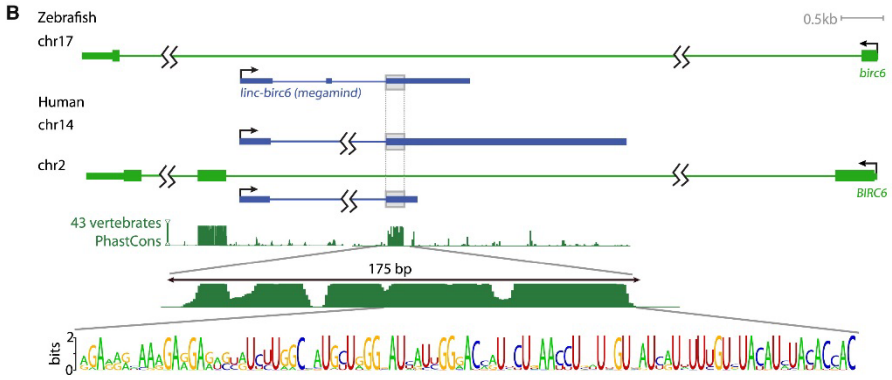
LncRNAs in non-human species

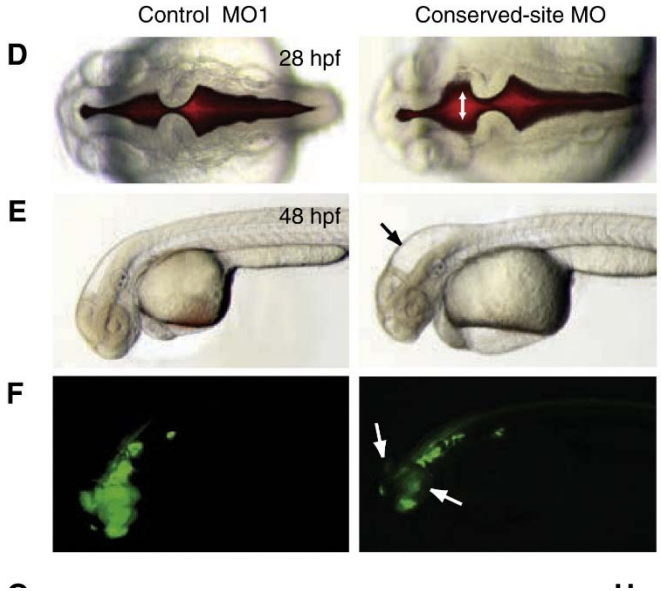




Ulitsky et al
PMID 22196729

Deep conservation of lncRNA between Fish and Human eg Megamind





3. Next generation functional genomics of lncRNA

Annotated human lncRNA	13,000

?

Number of **characterised** human lncRNAs: **126** (lncrnadb.org)

Summary

1. The human genome contains thousands of non-coding RNAs, most of which have unknown function.

2. Non-coding RNAs are diverse: size, biogenesis, evolutionary conservation, function.

3. Non-coding RNAs may help to explain fundamental biological questions: human evolution, human disease, genomic complexity.

Almost everything remains to be discovered!