Insights from within the ENCODE project

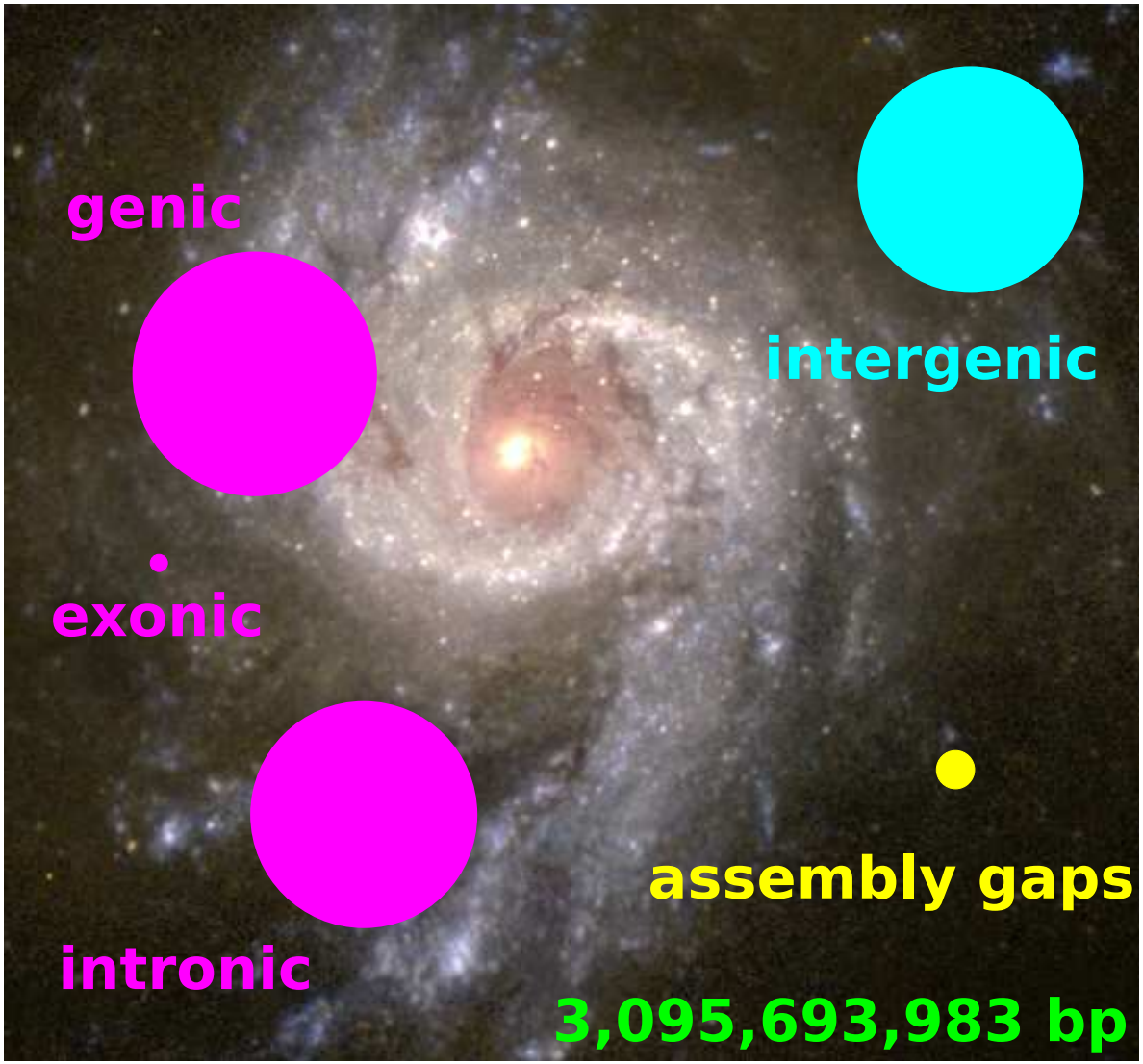
The interplay of small and long RNAs

Andrea Tanzer

Centre de Regulacio Genomica (CRG)
Dr. Aiguader, 88, E-08003 Barcelona
andrea.tanzer@crg.eu

The human genome (GRCh37/hg19)

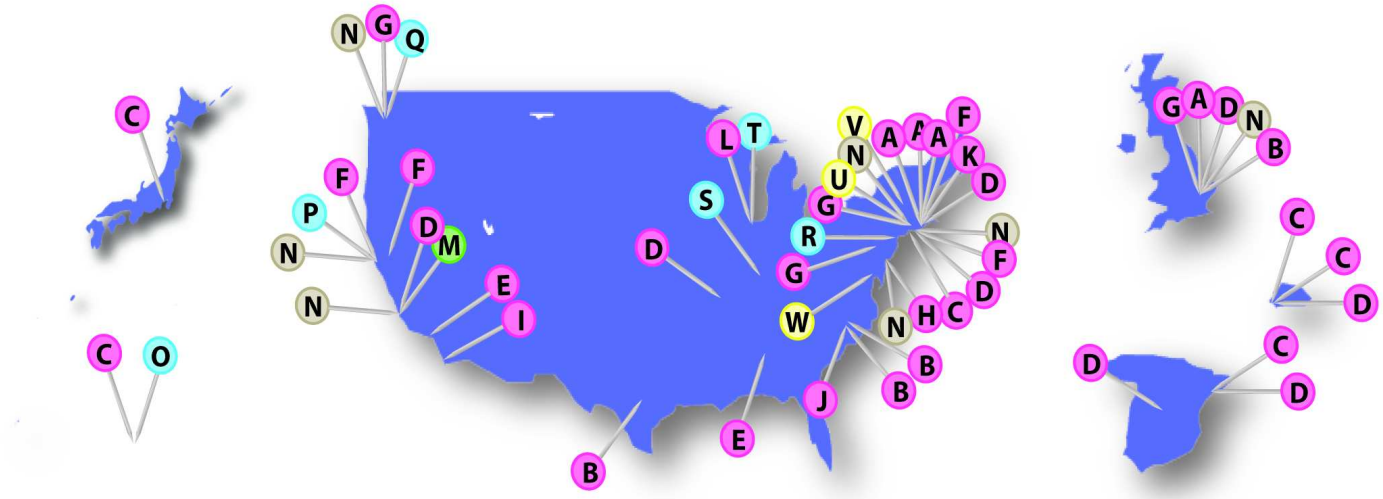
H. sapiens
(Symbolfoto)



ENCODE - What's this?

- **ENC**yclopedia **O**f **D**NA **E**lements
- mission: "enable the scientific and medical communities to interpret the human genome sequence"
- discover and define functional elements (genes, transcripts, transcriptional regulatory regions, chromatin states, DNA methylation patterns)
- introduce standards to ensure high-quality data
- develop novel algorithms
- data are public (with some restrictions)
- collective effort

ENCODE groups around the globe



Production Groups

- A **Broad Institute**, Harvard University, MIT, Cambridge University
- B **Duke University**, University of Texas, Austin, UNC-Chapel Hill, EBI
- C **CSHL**, University of Geneva, Centre for Genomic Regulation, RIKEN, University of Lausanne, Genome Institute of Singapore
- D **Sanger Institute** Washington University, Yale University, Centre for Genomic Regulation, UCSC, MIT, University of Lausanne, CNIO
- E **HudsonAlpha**, Caltech
- F **Stanford University**, Yale University, UC Davis, Harvard University
- G **University of Washington**, U Mass Med School, EBI, Princeton University
- H **Penn State**
- I **UC - San Diego**
- J **UNC - Chapel Hill**
- K **Boston University**
- L **University of Chicago**

Data Coordination Center

- M **UCSC**

Data Analysis Center

- N **EBI**, UC Berkeley, Yale University Penn State, UCSC, University of Washington, U Mass Med School

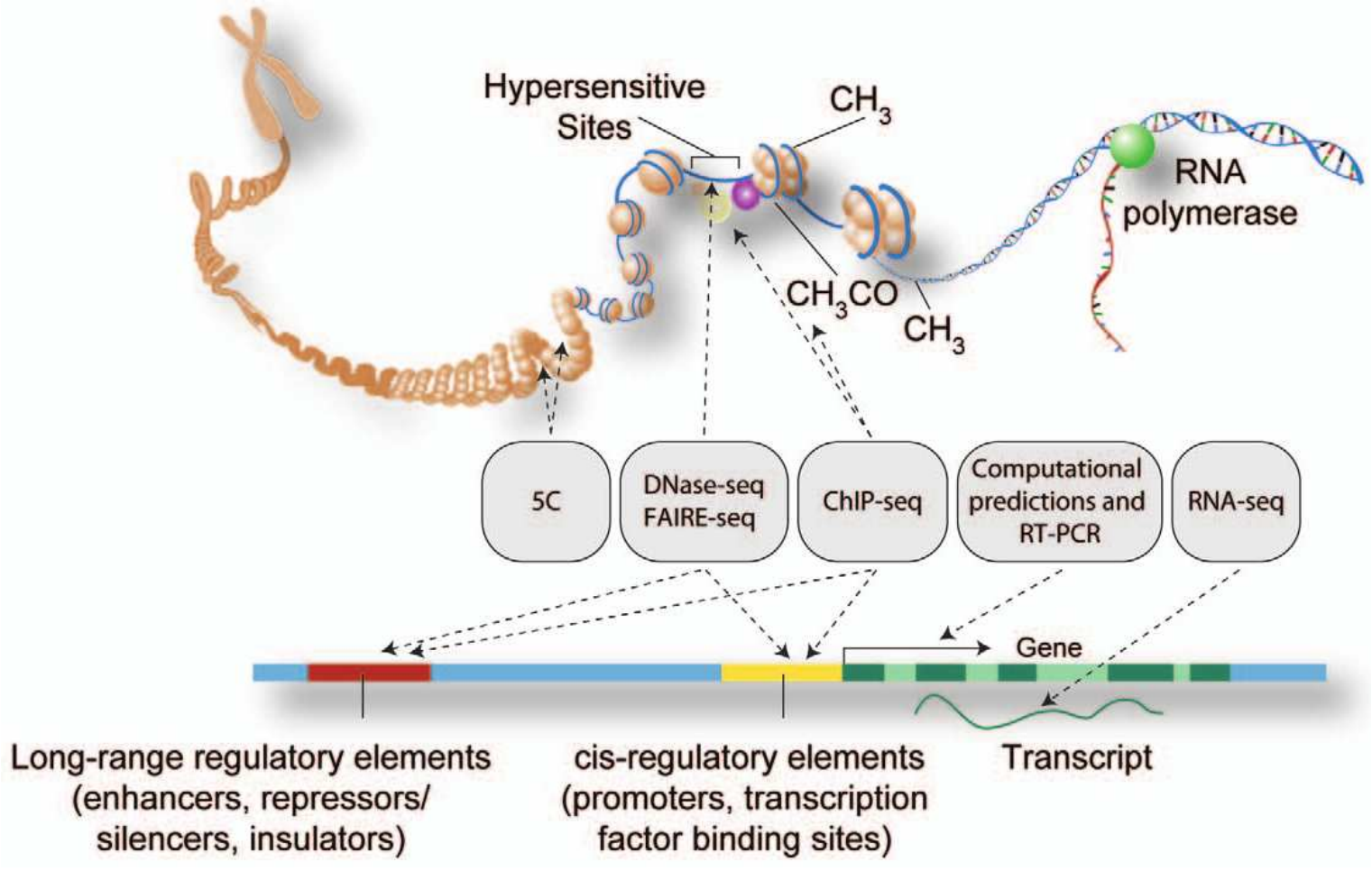
Technology Groups

- O **Genome Institute of Singapore**
- P **Stanford University**
- Q **University of Washington**
- R **Albert Einstein College of Med**
- S **Indiana University**
- T **University of Chicago**

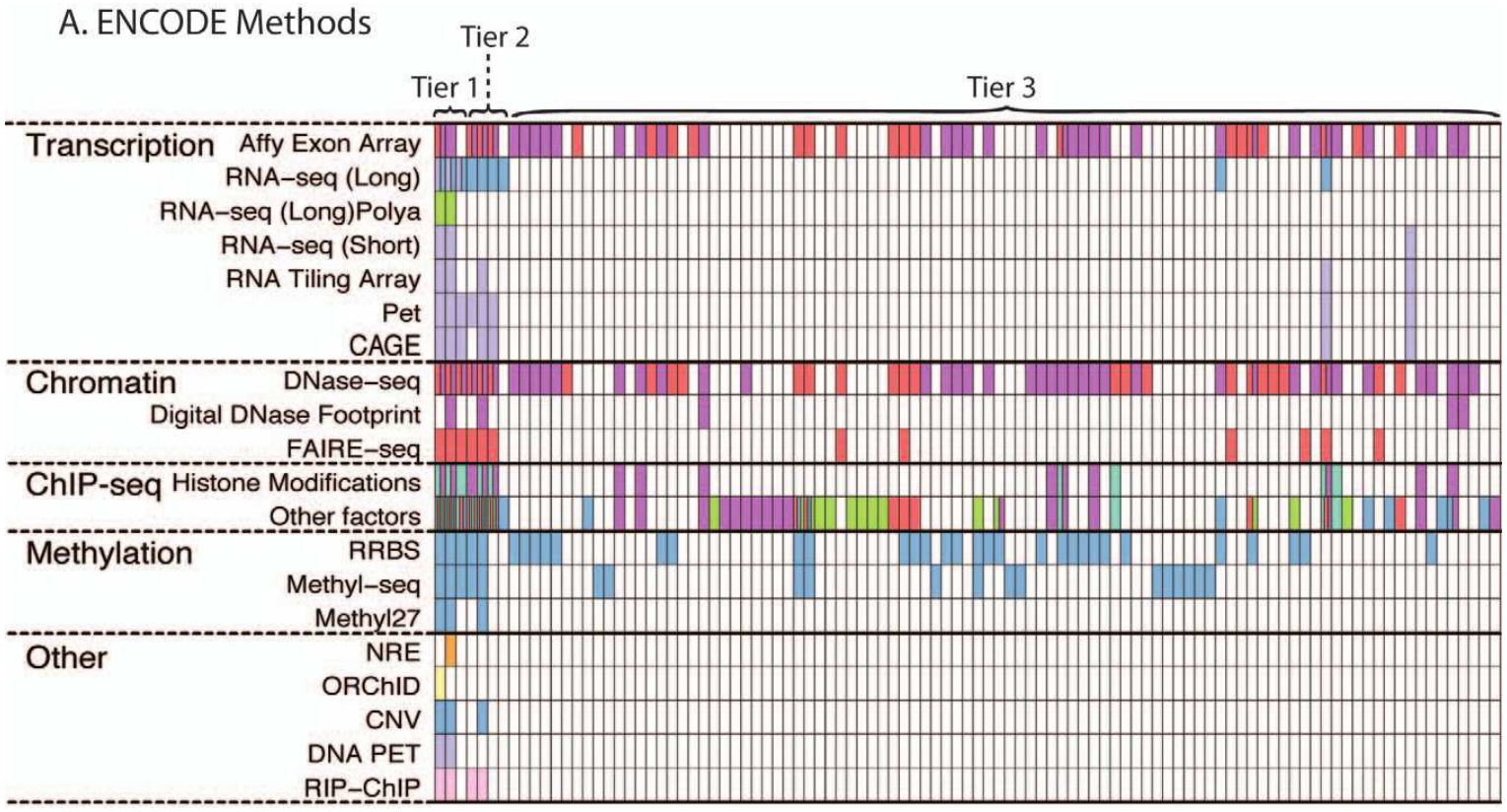
Other Groups

- U **U Mass Med School**
- V **SUNY Albany**
- W **NHGRI**

Organization

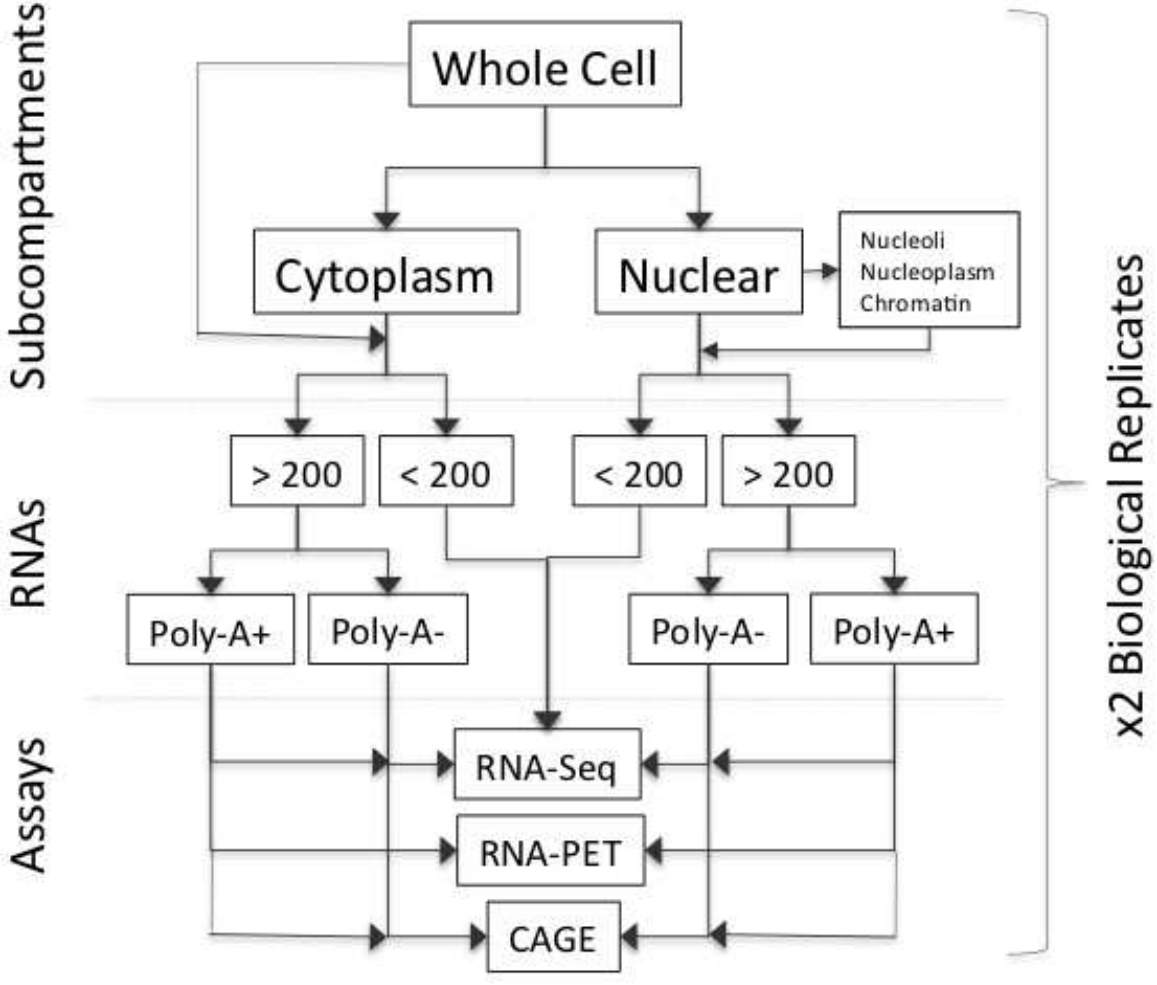


Cell lines used for RNAseq

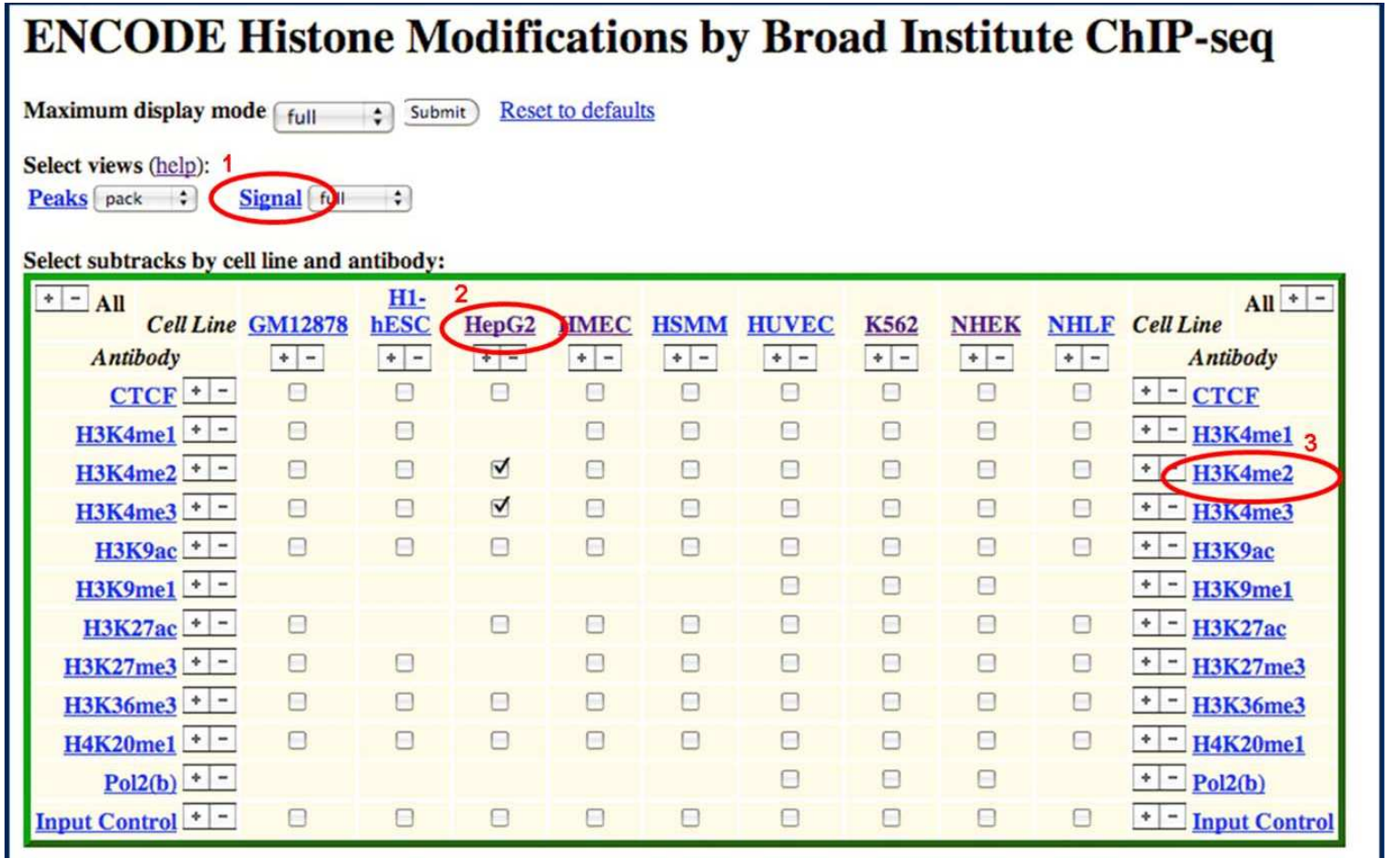


http://genome.ucsc.edu/ENCODE/cellTypes.html

RNA fractionation



Data Repository



ENCODE Histone Modifications by Broad Institute ChIP-seq

Maximum display mode: full — Submit — Reset to defaults

Select views (help): 1

Peaks: pack — Signal: full

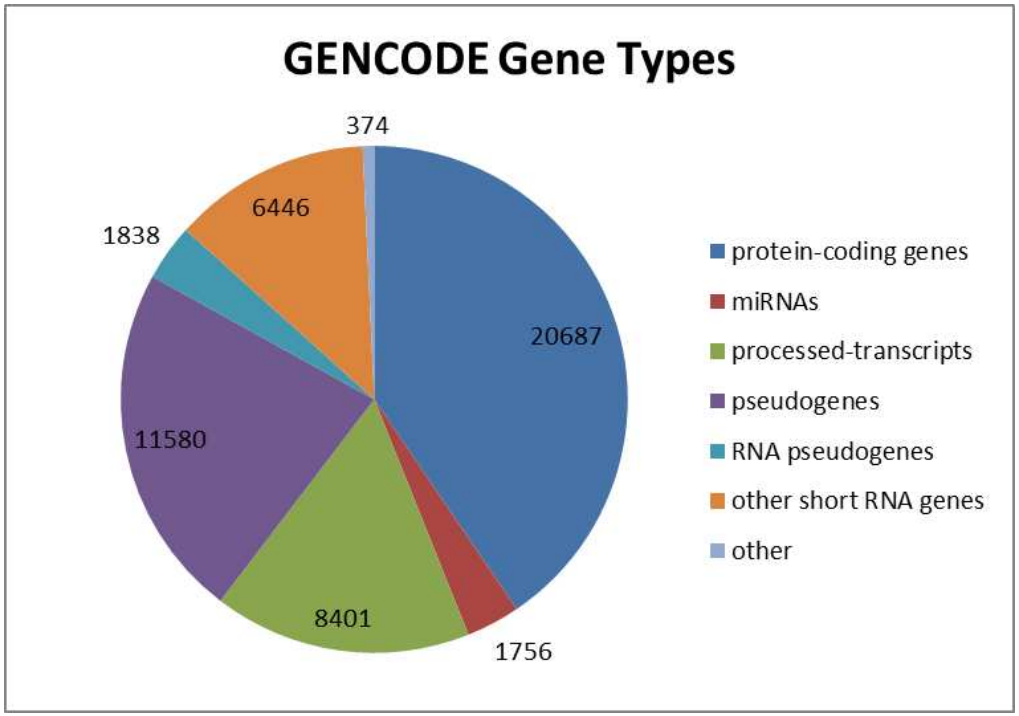
Select subtracks by cell line and antibody:

All	Cell Line / Antibody	GM12878	H1-hESC	HepG2	HMEC	HSMM	HUVEC	K562	NHEK	NHLF	Cell Line / Antibody	All
	CTCF	☐	☐	☐	☐	☐	☐	☐	☐	☐	CTCF	
	H3K4me1	☐	☐		☐	☐	☐	☐	☐	☐	H3K4me1	
	H3K4me2	☐	☐	☑	☐	☐	☐	☐	☐	☐	H3K4me2	
	H3K4me3	☐	☐	☑	☐	☐	☐	☐	☐	☐	H3K4me3	
	H3K9ac	☐	☐	☐	☐	☐	☐	☐	☐	☐	H3K9ac	
	H3K9me1						☐	☐	☐		H3K9me1	
	H3K27ac	☐		☐	☐	☐	☐	☐	☐	☐	H3K27ac	
	H3K27me3	☐	☐		☐	☐	☐	☐	☐	☐	H3K27me3	
	H3K36me3	☐	☐	☐	☐	☐	☐	☐	☐	☐	H3K36me3	
	H4K20me1	☐	☐	☐	☐	☐	☐	☐	☐	☐	H4K20me1	
	Pol2(b)						☐	☐	☐		Pol2(b)	
	Input Control	☐	☐	☐	☐	☐	☐	☐	☐	☐	Input Control	

2 3

data release policy:
`http://www.encodeproject.org/ENCODE/terms.html`

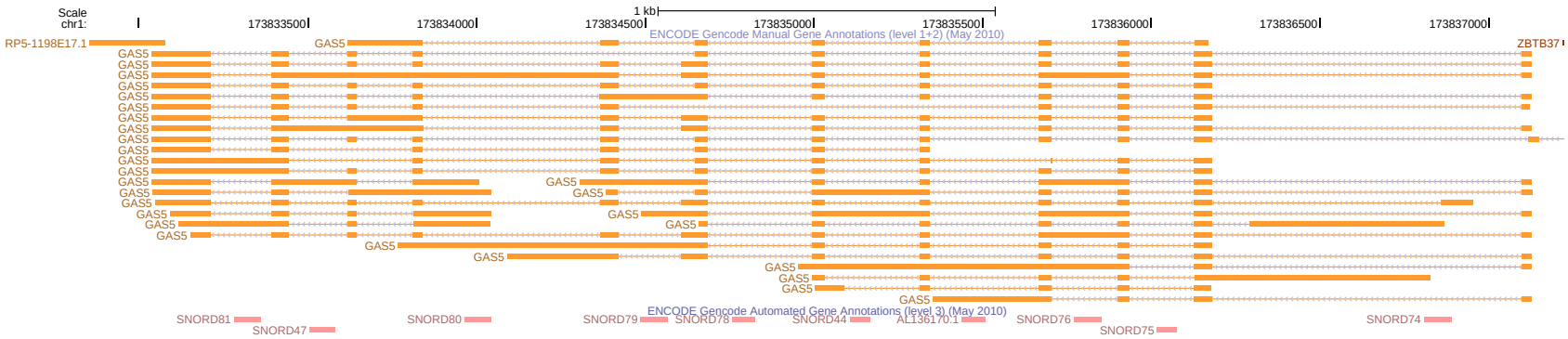
GENCODE Annotation



- provided by Sanger center
- integration of ENSEMBL and HAVANA
- manually curated annotation
- 51,082 genes
- 161,375 transcripts
- majority of tx is noncoding

Host Genes

- several small noncoding RNAs (ncRNAs) not transcribed as individual genes
- processed from introns of longer, multi-exonic, and often non protein coding polII transcripts
- $\rightarrow$ 'polycistronic transcripts' driven by a single promoter, several genes (i.e. small RNAs) are transcribed and subsequently processed into the functional products by nucleases.
- examples: snoRNA host genes GAS5 and RPL7a



Processing of small ncRNAs from host genes

snoRNAs:

- processed from introns (lariat debranching)
- or exonucleolytic cleavage

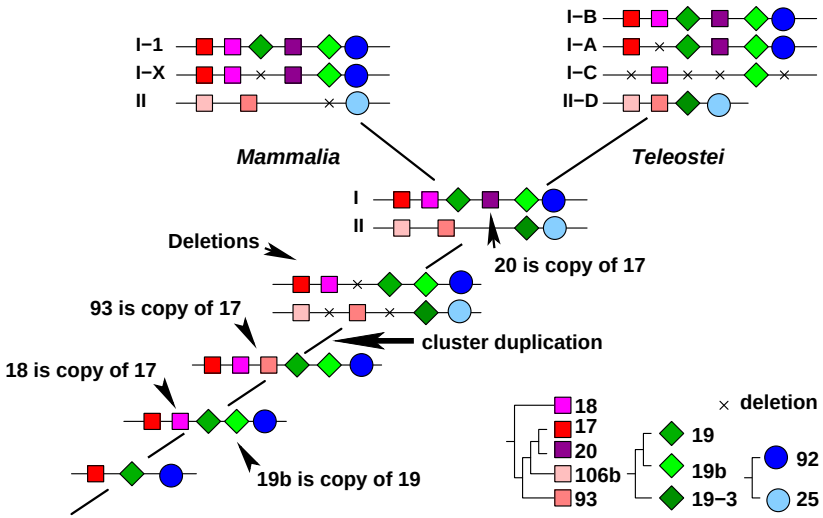
miRNAs

- processed post- or co-transcriptional
- processing does not interfere with splicing

small RNAs often functionally related to host genes

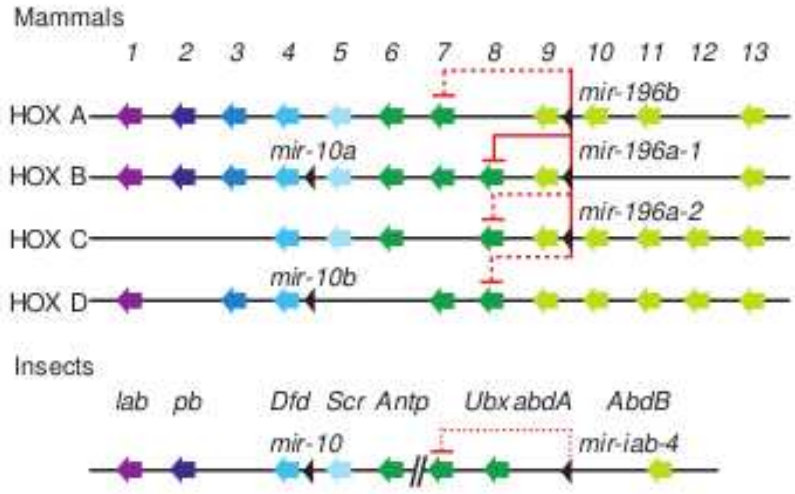
Evolution of miRNAs

miRNA-17 cluster

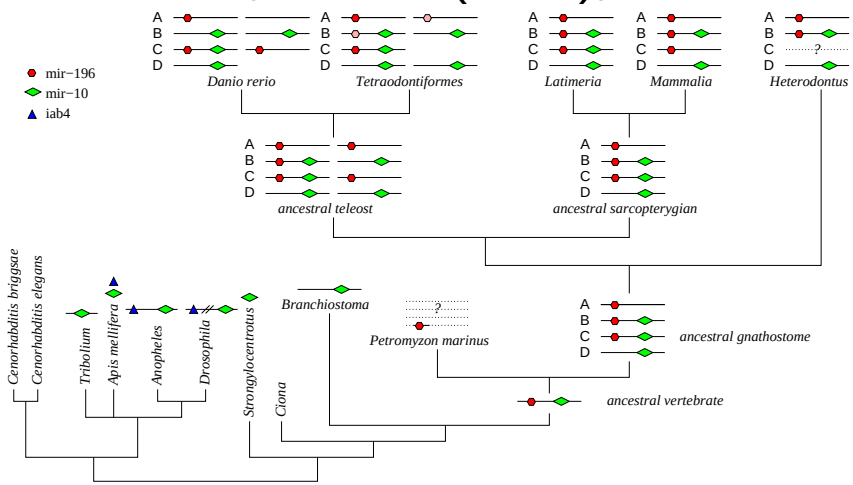


Tanzer and Stadler. J Mol Biol. 2004; 339:327-35

miRNA 10 & 196



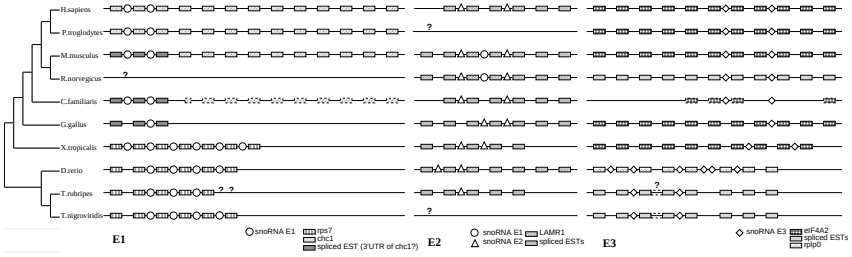
Yekta *et al.*, Science(2004), 304:594-6



Tanzer *et al.* J Exp Zoolog B Mol Dev Evol. 2005; 304:75-85

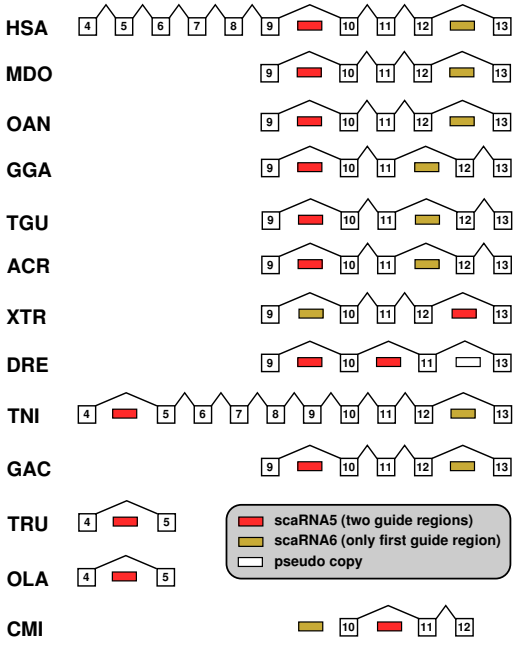
snoRNAs and their host genes

snoRNAs E1, E2, E3



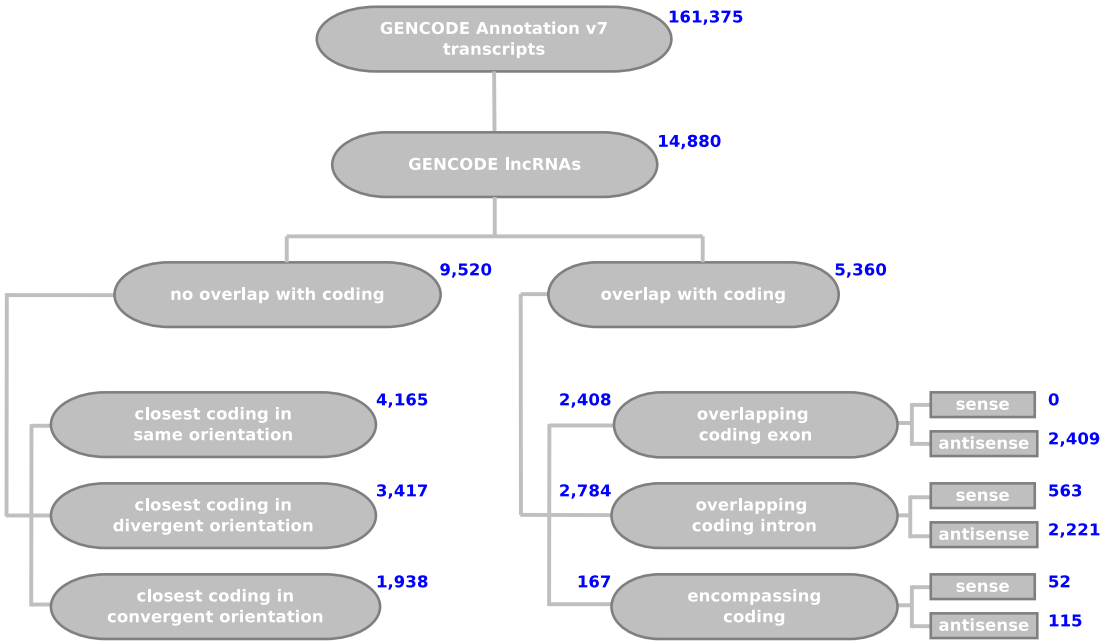
Bompfünewerer *et al.* Th. Biosci. 2005; 123:301-369

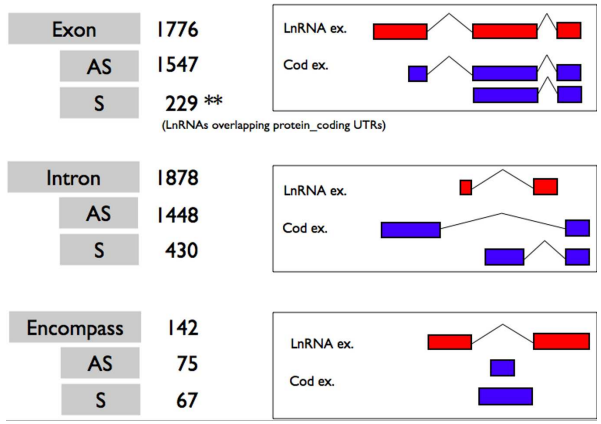
scaRNA5&6



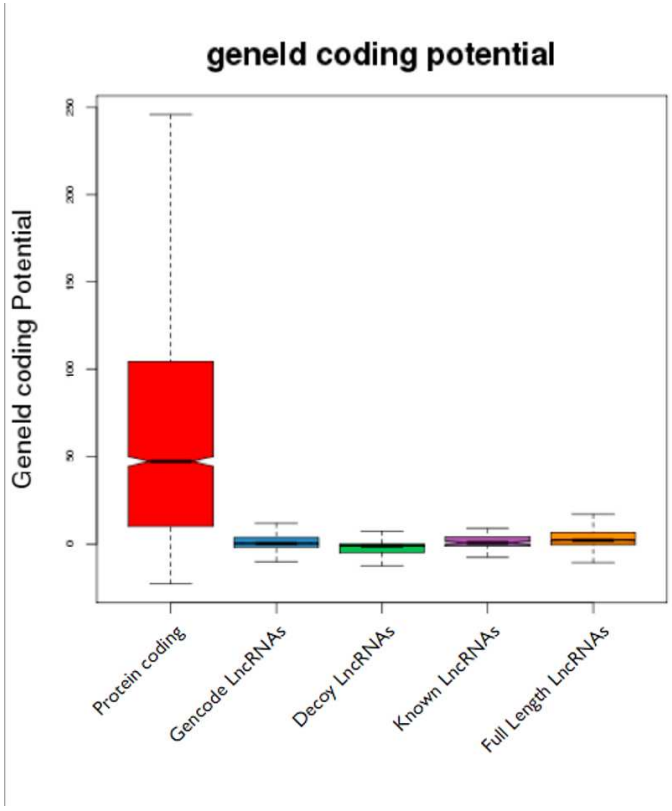
Marz et al., RNA Biol.(2011), 8:938-46

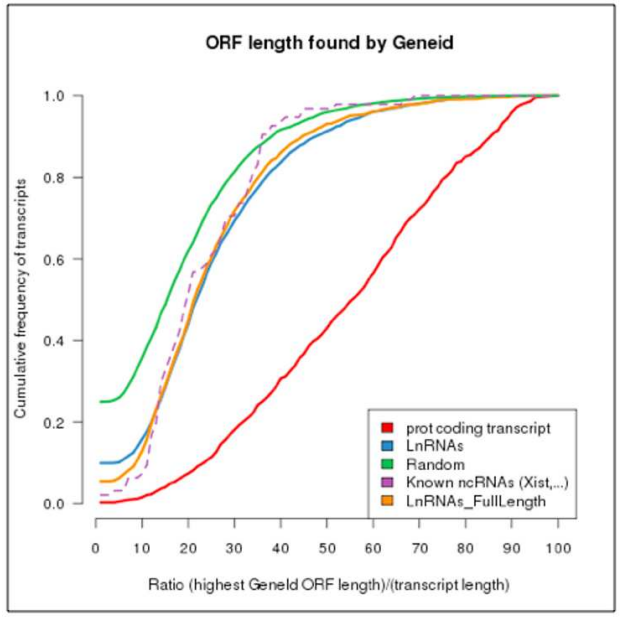
Long noncoding RNAs - genomic context





Are long noncoding RNAs non coding?

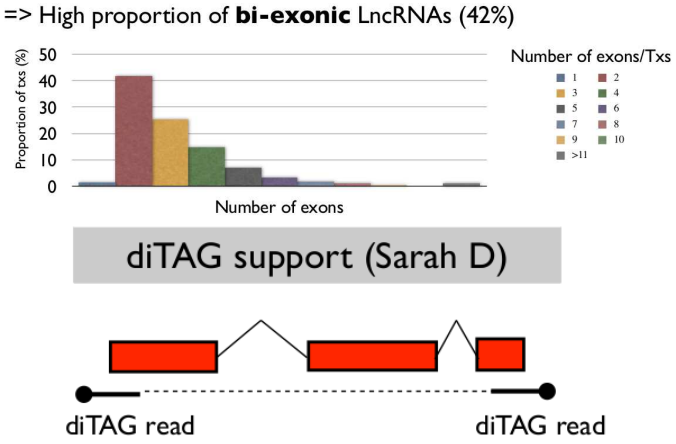




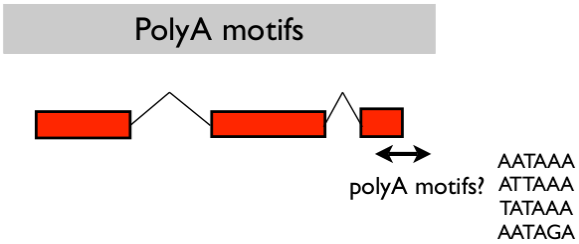
- coding potential in the range of other known ncRNAs
- ORF length: 20% of sequence (protein coding: 60%)

lncRNAs share mRNA like features

exon structure and number

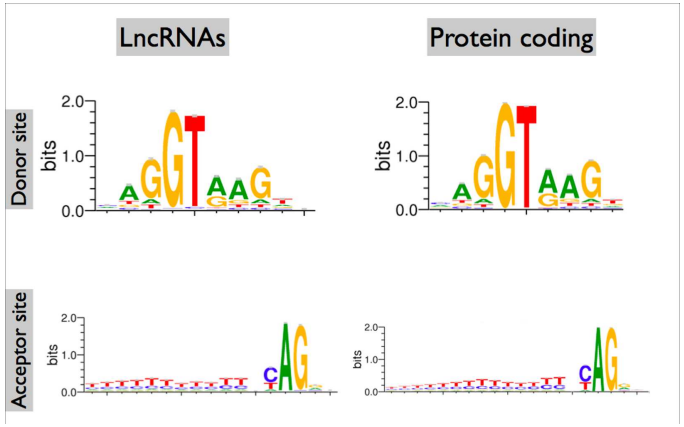


polyA sites



Category	% transcripts 100 bp TTS	200 bp TTS
Coding	54.7	80.3
LncRNAs	53.5	81.7

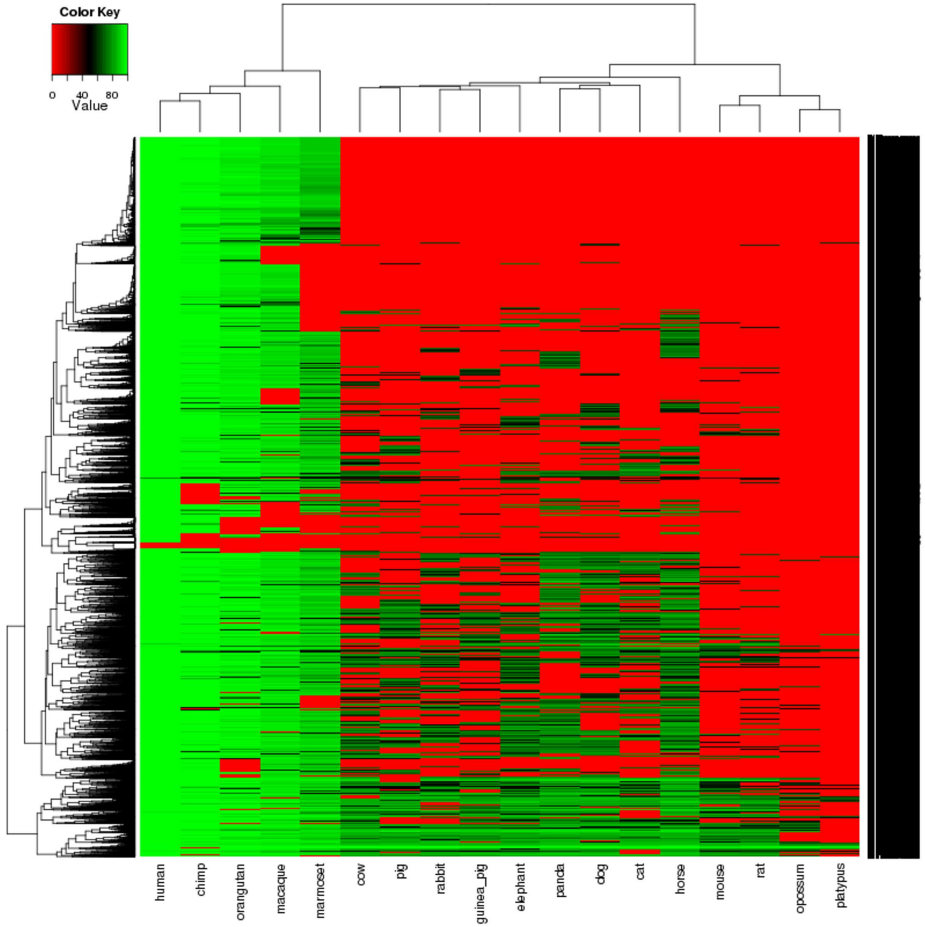
splice sites



	% introns lncRNA	prot. cod.
Donor site		
GT	95.046	99.027
GC	2.271	0.797
AT	0.993	0.016
Acceptor site		
AG	97.087	99.796
AC	1.088	0.070
GG	0.308	0.015

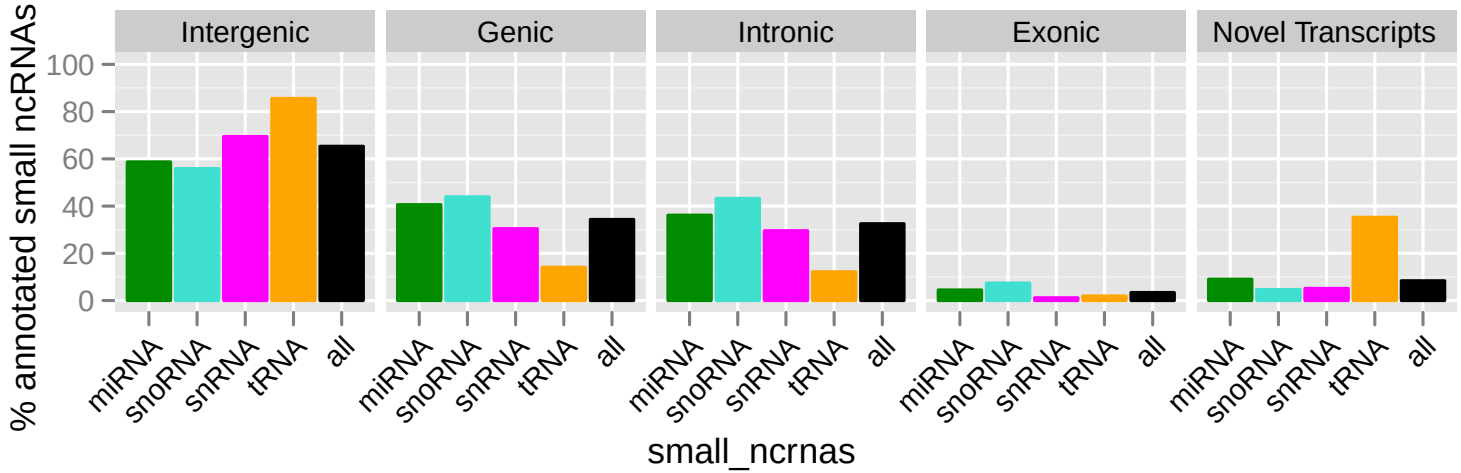
- mean transcript length: 938bp
- mRNA like poly A sites
- 42% bi-exonic transcripts
- full length: only 10% supported by ditags ← low expression?

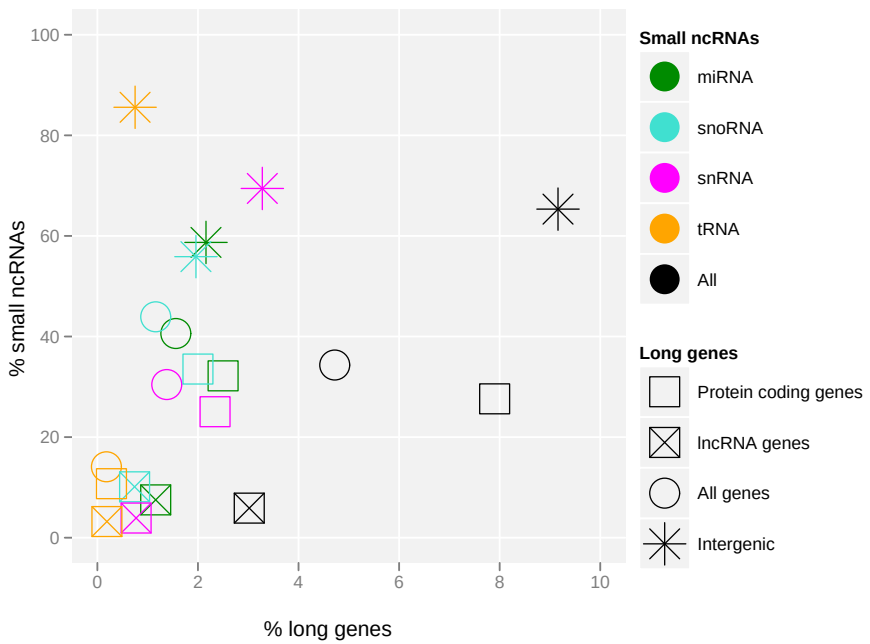
Conservation of lncRNAs



- blast/exonerate
- 31% primate specific
- 4% conserved acroos all species
- 194 families

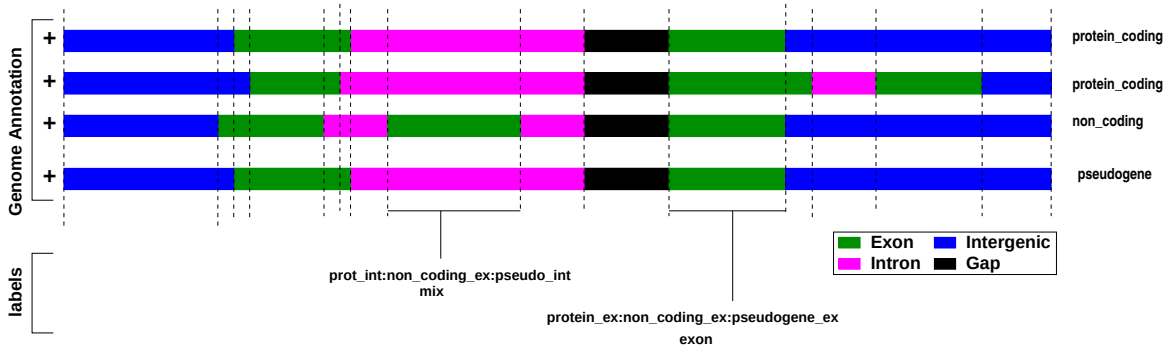
Small ncRNAs in GENCODE annotations

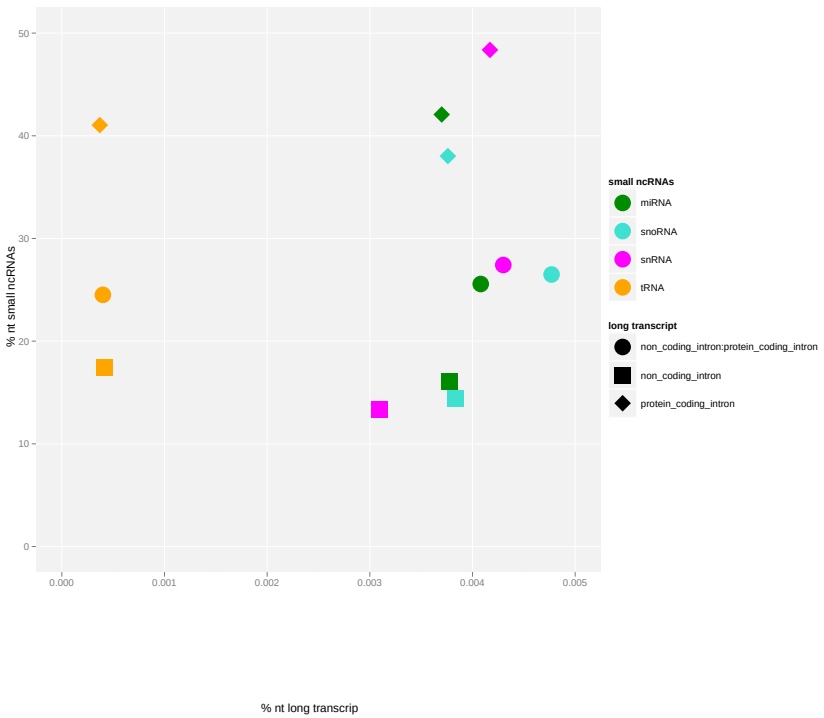




- ca. 60% small ncRNAsintergenic
- miRNAs and snoRNAs more genic
- if genic, then intronic $\rightarrow$ separation of information
- in both coding and noncoding genes

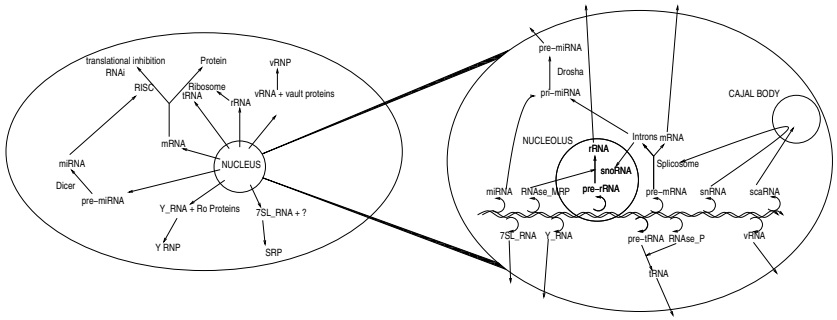
Detailed few of intronic small RNAs

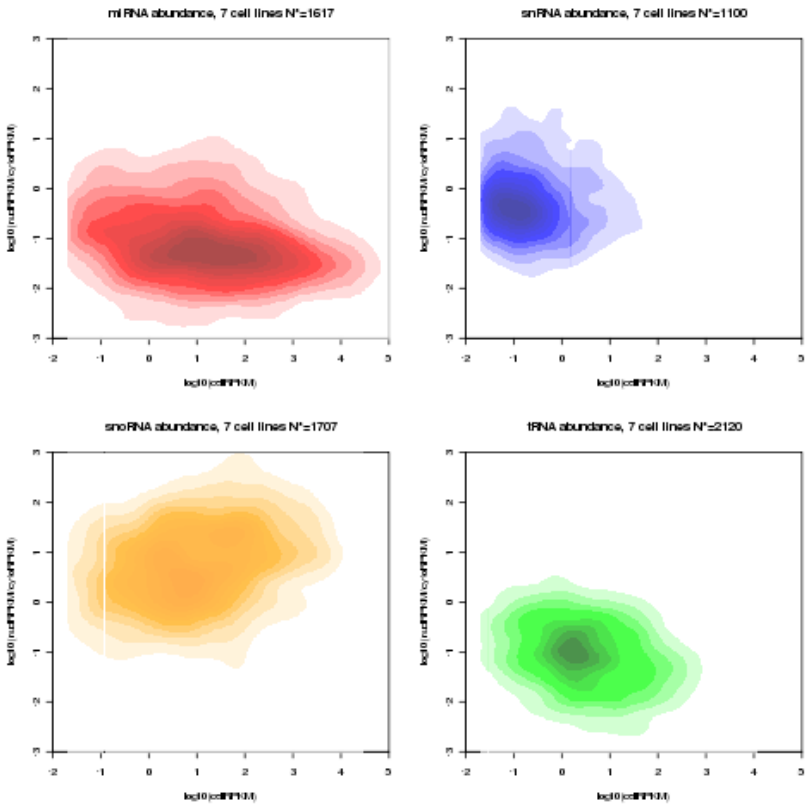


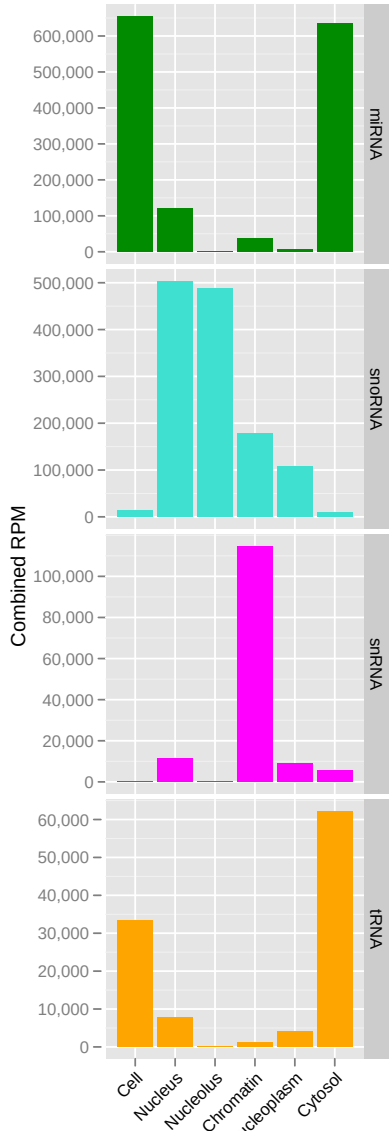


- tRNAs independent loci
- lncRNAs and protein coding gens contain miRNAs, snoRNAs, snRNAs at similar frequency
- lnc-protein overlapping regions: are slightly enriched $\rightarrow$ backup system?

Subcellular compartimentalization of small ncRNAs

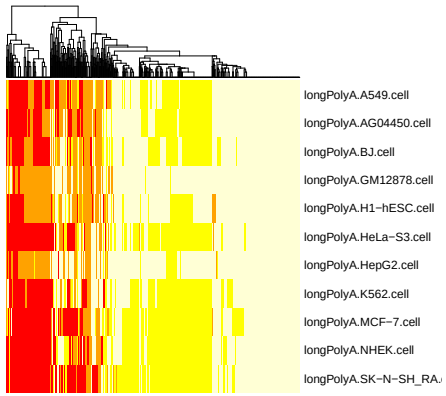




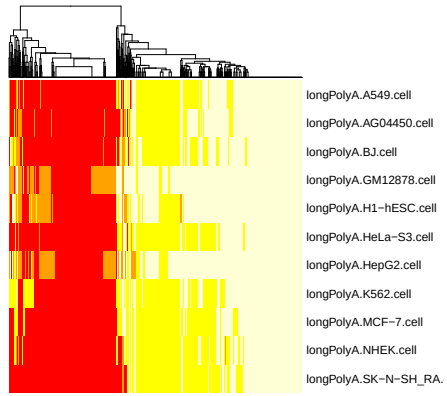


Coexpression of annotated small ncRNA & annotated long transcripts

miRNA



snoRNAs

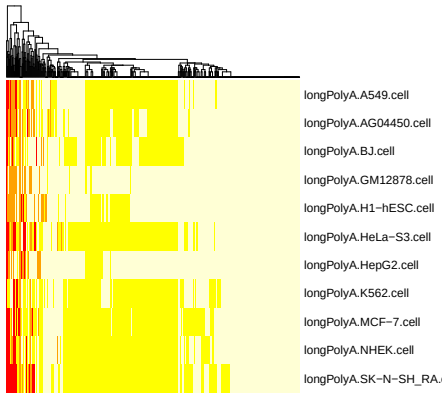


red: both, orange: only small, yellow: only host, white: none

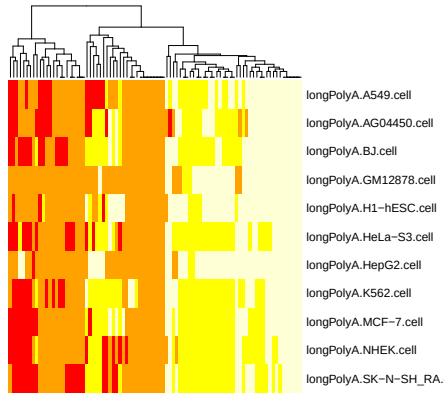
- we see coexpression
- about half of small-long pairs (based on annotation) are expressed
- snoRNAs: uniform expression
- miRNA: some uniformly expressed, but also tissue specific expression
- for both: hosts expressed, but not small $\rightarrow$ annotated long is not the host, or regulation of procesing of small

Coexpression of annotated small ncRNA & annotated long transcripts

snRNA



tRNA



red: both, orange: only small, yellow: only host, white: none

- snRNAs: hardly any genic snRNAs expressed
- tRNAs independent of long

Summary

- overview of ENCODE project
- and GENCODE gene annotation
- small ncRNAs in different types of host genes
- characterized long ncRNAs in gencode annotated regions
- lncRNAs are non protein coding, but share other mRNA features
- small ncRNA reside in introns often shared by coding-noncoding transcripts
- majority of small snoRNAs and miRNAs are intergenic
- 20% of intergenic small ncRNAs in transcribed regions
- coexpression of host-small ncRNA pairs

Acknowledgments

CRG

Roderic Giugo
Sarah Djebali
Angelika Merkel
David Gonzales
Julien Lagard

Rory Johnson
Thomas Derrien
Cedric Notredame
Giovanni Buscotti

Sanger Center

Tim Hubbard
Jennifer Harrow
Felix Kokocinski
Jose Gonzales

UCSC

Mark Diekhans
Rachel Harte

Univ. Lausanne:

Alexandre Reymond
Cedric Howald

CSHL:

Tom Gingeras
Carrie Davies
Alex Dobin

...and probably many more

CRG Centre for Genomic Regulation
Barcelona Biomedical Research Park