Ch 4 – L1.2

ENCODE transcriptome

ARTICLE

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

This is the leading article that describes all the ENCODE project and gives a overall resumé of results obtained in the 2nd phase.

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

ENCODE official website: https://www.encodeproject.org/
ENCODE at the NHGRI: http://www.genome.gov/encode/
Nature ENCODE: http://www.nature.com/encode/#/threads

ARTICLE

doi:10.1038/nature11233

Landscape of transcription in human cells

Sarah Djebali[1]*, Carrie A. Davis[2]*, Angelika Merkel[1], Alex Dobin[2], Timo Lassmann[3], Ali Mortazavi[4,5], Andrea Tanzer[1], Julien Lagarde[1], Wei Lin[2], Felix Schlesinger[2], Chenghai Xue[2], Georgi K. Marinov[4], Jainab Khatun[6], Brian A. Williams[4], Chris Zaleski[2], Joel Rozowsky[7,8], Maik Röder[1], Felix Kokocinski[9], Rehab F. Abdelhamid[3], Tyler Alioto[1,10], Igor Antoshechkin[4], Michael T. Baer[2], Nadav S. Bar[11], Philippe Batut[2], Kimberly Bell[2], Ian Bell[12], Sudipto Chakrabortty[2], Xian Chen[13], Jacqueline Chrast[14], Joao Curado[1], Thomas Derrien[1], Jorg Drenkow[2], Erica Dumais[12], Jacqueline Dumais[12], Radha Duttagupta[12], Emilie Falconnet[15], Meagan Fastuca[2], Kata Fejes-Toth[2], Pedro Ferreira[1], Sylvain Foissac[12], Melissa J. Fullwood[16], Hui Gao[12], David Gonzalez[1], Assaf Gordon[2], Harsha Gunawardena[13], Cedric Howald[14], Sonali Jha[2], Rory Johnson[1], Philipp Kapranov[12,17], Brandon King[4], Colin Kingswood[1,10], Oscar J. Luo[16], Eddie Park[5], Kimberly Persaud[2], Jonathan B. Preall[2], Paolo Ribeca[1,10], Brian Risk[6], Daniel Robyr[15], Michael Sammeth[1,10], Lorian Schaffer[4], Lei-Hoon See[2], Atif Shahab[16], Jorgen Skancke[1,11], Ana Maria Suzuki[3], Hazuki Takahashi[3], Hagen Tilgner[1]†, Diane Trout[4], Nathalie Walters[14], Huaien Wang[2], John Wrobel[6], Yanbao Yu[13], Xiaoan Ruan[16], Yoshihide Hayashizaki[3], Jennifer Harrow[9], Mark Gerstein[7,8,18], Tim Hubbard[9], Alexandre Reymond[14], Stylianos E. Antonarakis[15], Gregory Hannon[2], Morgan C. Giddings[6,13], Yijun Ruan[16], Barbara Wold[4], Piero Carninci[3], Roderic Guigó[1,19] & Thomas R. Gingeras[2,12]

Eukaryotic cells make many types of primary and processed RNAs that are found either in specific subcellular compartments or throughout the cells. A complete catalogue of these RNAs is not yet available and their characteristic subcellular localizations are also poorly understood. Because RNA represents the direct output of the genetic information encoded by genomes and a significant proportion of a cell's regulatory capabilities are focused on its synthesis, processing, transport, modification and translation, the generation of such a catalogue is crucial for understanding genome function. Here we report evidence that three-quarters of the human genome is capable of being transcribed, as well as observations about the range and levels of expression, localization, processing fates, regulatory regions and modifications of almost all currently annotated and thousands of previously unannotated RNAs. These observations, taken together, prompt a redefinition of the concept of a gene.

from Djebali et al., 2012

……

Here we report identification and characterization of annotated and novel RNAs that are enriched in either of the two major cellular subcompartments (nucleus and cytosol) for all **15 cell lines studied**, and in three additional subnuclear compartments in one cell line.

In addition, we have sought to determine whether identified transcripts are modified at their 5’ and 3’ termini by the presence of a 7-methyl guanosine cap or polyadenylation, respectively.

These results considerably extend the current genome-wide annotated catalogue of long polyadenylated and small RNAs collected by the **GENCODE** annotation group.

ENCODE - Transcriptome

Djebali et al., 2012

RNA-Seq: identification of annotated and novel RNAs from either of the two major cellular subcompartments (nucleus and cytosol) for 15 cell lines.

To see the EXPERIMENTAL GRID :
http://genome.ucsc.edu/ENCODE/dataMatrix/encodeDataMatrixHuman.html

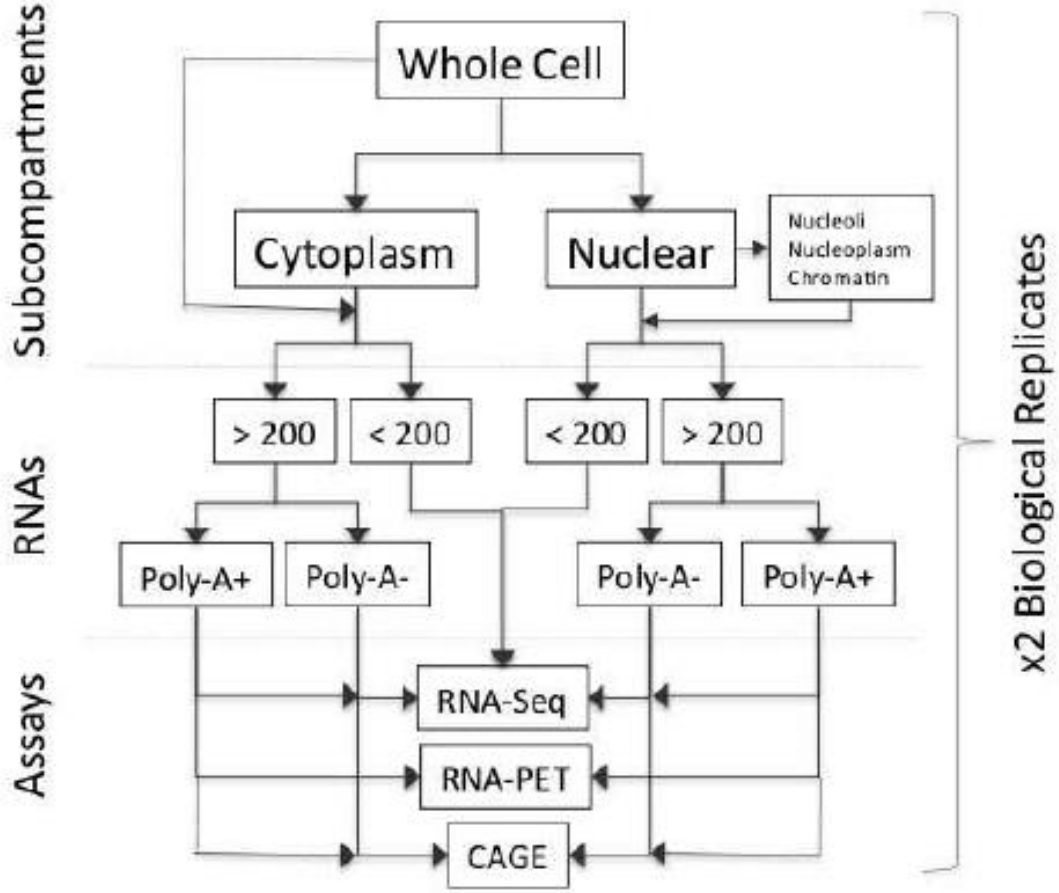
- 62.1% of genome covered by processed transcripts; 74.7% by unprocessed transcripts.
- Novel elements cover 78% of intronic nucleotides and 34% of intergenic sequences.
- Multiple isoforms per gene expressed simultaneously, with a plateau at 10-12 isoforms per gene per cell line.
- eRNA – transcripts starting from enhancers
- 6% of coding and noncoding overlap with small RNA (probably precursors)

Question: is this feature «conclusive» ?

RNA data set generation

We performed subcellular compartment fractionation (whole cell, nucleus and cytosol) before RNA isolation in 15 cell lines (Supplementary Table 1) to interrogate deeply the human transcriptome. For the K562 cell line, we also performed additional nuclear subfractionation into chromatin, nucleoplasm and nucleoli. The RNAs from each of these subcompartments were prepared in replica and were separated based on length into >200 nucleotides (long) and <200 nucleotides (short). Long RNAs were further fractionated into polyadenylated and non-polyadenylated transcripts. A number of complementary technologies were used to characterize these RNA fractions as to their sequence (RNA-seq), sites of initiation of transcription (cap-analysis of gene expression (CAGE)[9]) and sites of 5′ and 3′ transcript termini (paired end tags (PET)[10]; Supplementary Fig. 1). Sequence reads were

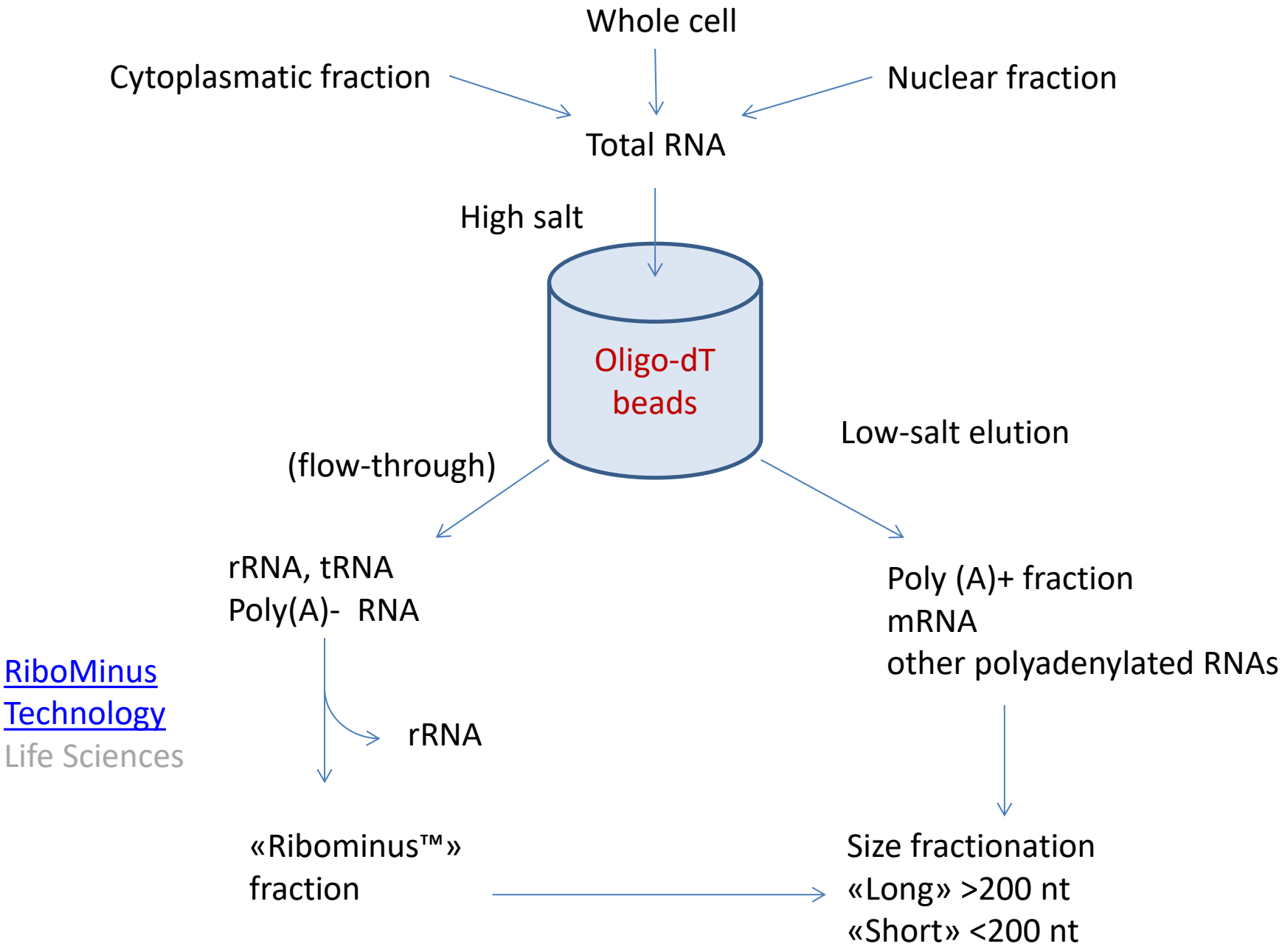
RNA-PET is a paired-end tag (PET) sequencing method for full-length mRNA analysis. RNA-PET captures and sequences the 5'- and 3'-end tags of full-length cDNA fragments of all expressed genes in a biological sample



From Supplementary 2

Supplementary Figure S1

Sample Flowchart. The ENCODE transcriptome data are obtained from several cell lines which have been cultured in replicates. They were either left intact (whole cell) and/or fractionated into cytoplasm and nucleus prior to RNA isolation. Total RNA was then isolated and partitioned into RNA ¿ 200bp (long) and ¡ 200bp (short). The long RNA was further partitioned over an oligo-dT column into polyA+ and polyA- fractions. The K562 cell line also underwent additional fractionation into nucleoli, nucleoplasm and chromatin, but no further partition into polyA+ and polyA- was done. RNA-seq was conducted on polyA+, polyA- and total (K562) RNA samples. CAGE was conducted primarily on polyA+ and total RNA but also on some polyA- samples. RNA-PET was conducted on PolyA+ samples only (not shown here are RNA-seq experiments performed at CalTech on polyA+ whole cell RNA extracts).

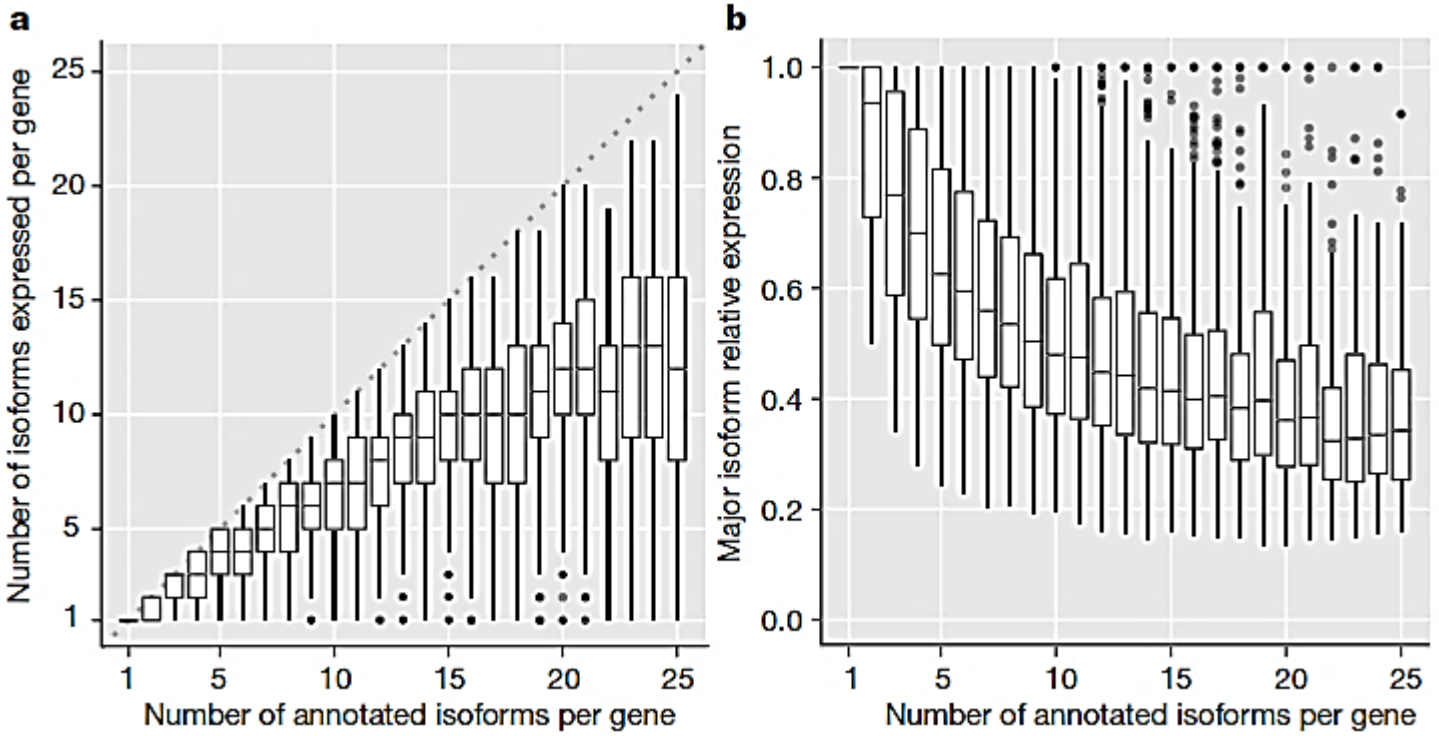

Whole cell
Cytoplasmatic fraction
Nuclear fraction
Total RNA
High salt
Oligo-dT beads
Low-salt elution
(flow-through)
rRNA, tRNA
Poly(A)- RNA
Poly (A)+ fraction
mRNA
other polyadenylated RNAs
RiboMinus Technology
Life Sciences
rRNA
«Ribominus™» fraction
Size fractionation
«Long» >200 nt
«Short» <200 nt

Such a deep, complete and detailed study in many cell lines allowed the discovery of:

- Many novel transcripts from known protein-coding genes (*)
- some undescribed potentially protein coding genes
- a large number of long noncoding RNAs from unsuspected genomic regions (intergenic)
- a laerge number of intragenic sense and antisense RNAs
- a large mumber of short and middle-short noncoding RNAs

(*) several transcript isoforms derived from altenative exon splicing, alternative TSS usage, alternative poly(A) site usage

Djebali 2012, Figure 4 – Transcript Isoforms



a) Number of expressed isoforms per gene per cell line. A plateau is evident between 10 and 12

b) Relative expression of the most abundant isoform per gene per cell line.

See an example at: https://www.ncbi.nlm.nih.gov/gene/7157

Djebali et al., 2012

Alternative transcription initiation and termination.

a total of 128,021 TSSs were detected across all cell lines (97,778 previously annotated ; 30,243 were novel intergenic/antisense TSSs).

CAGE tags…. identified a total of 82,783 nonredundant TSSs

48% of the CAGE-identified TSSs located within 500 base pairs (bp) of an annotated RNA-seq-detected GENCODE TSS,
additional 3% within 500 bp of a novel TSS

A large number of novel transcripts were classified as lncRNA

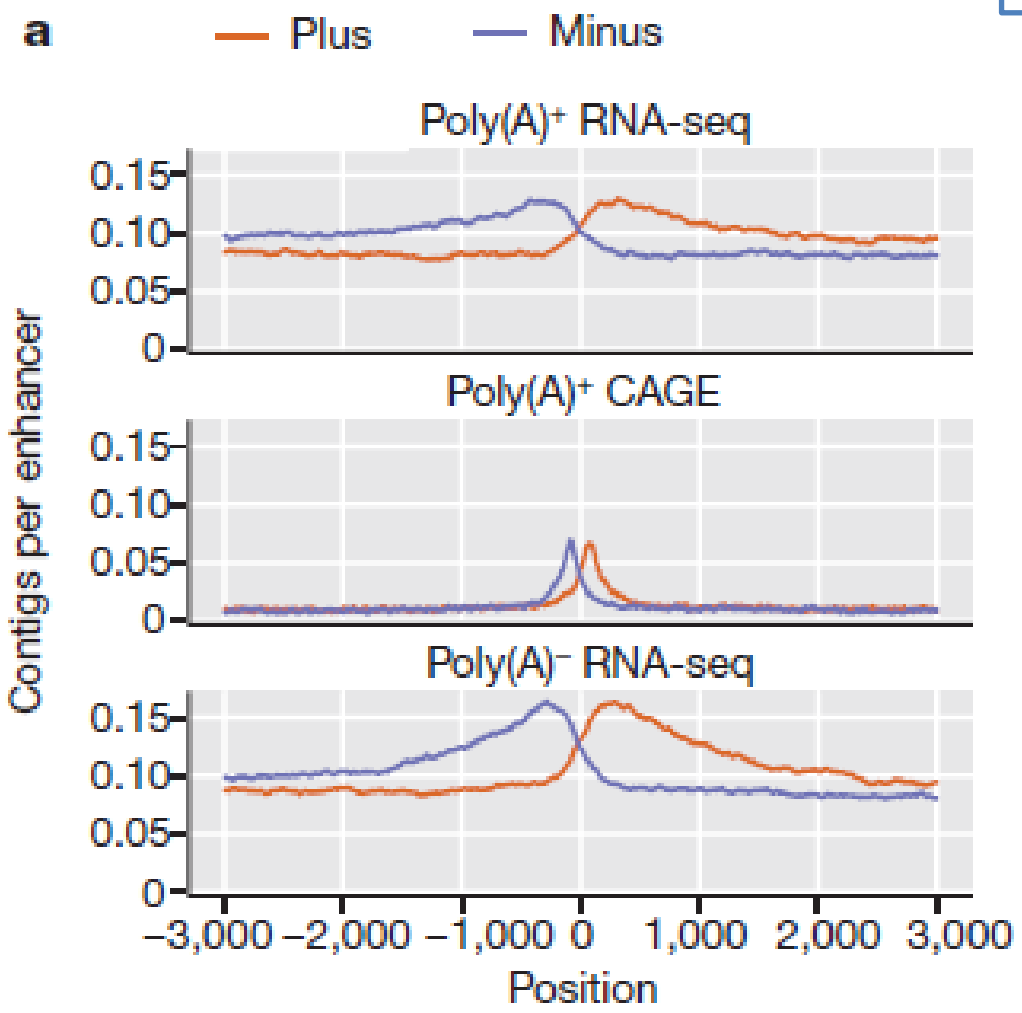
long noncoding RNA

How is the classification «*noncoding*» attributed ?

- ORF search in all the possible frames
- Short ORFs evaluated on «codon usage»
- Proteomic database interrogated
- Association with ribosomes (*poly-ribosome purification and RNA-seq*)

eRNA

Enhancer attribution by means of PTMs+TF ChIP-Seq data



Djebali et al., 2012

Expression level - quantity

Transcripts range in a 6-order magnitude (poly A+)(10^{-2} to 10^{4} rpkm) or 5 orders of magnitude (poly A-) (10^{-2} to 10^{3} rpkm)

Assuming that 1–4 r.p.k.m. approximates to 1 copy per cell (*Montazavi et al., 2008*):

- one quarter of protein-coding RNAs *and*
- 80% of long noncoding RNAs (lncRNA)

Are expressed at <u>1 or <1 molecules per cell</u>

i.e. the majority of lncRNAs are expressed at a very low level

Novel lncRNAs discovered here contains also a class showing *rpkm* from 10^{-4} to 10^{-1}: << extremely low expression >>

Question: what does it mean «less than one molecule per cell ?

Expression level by class

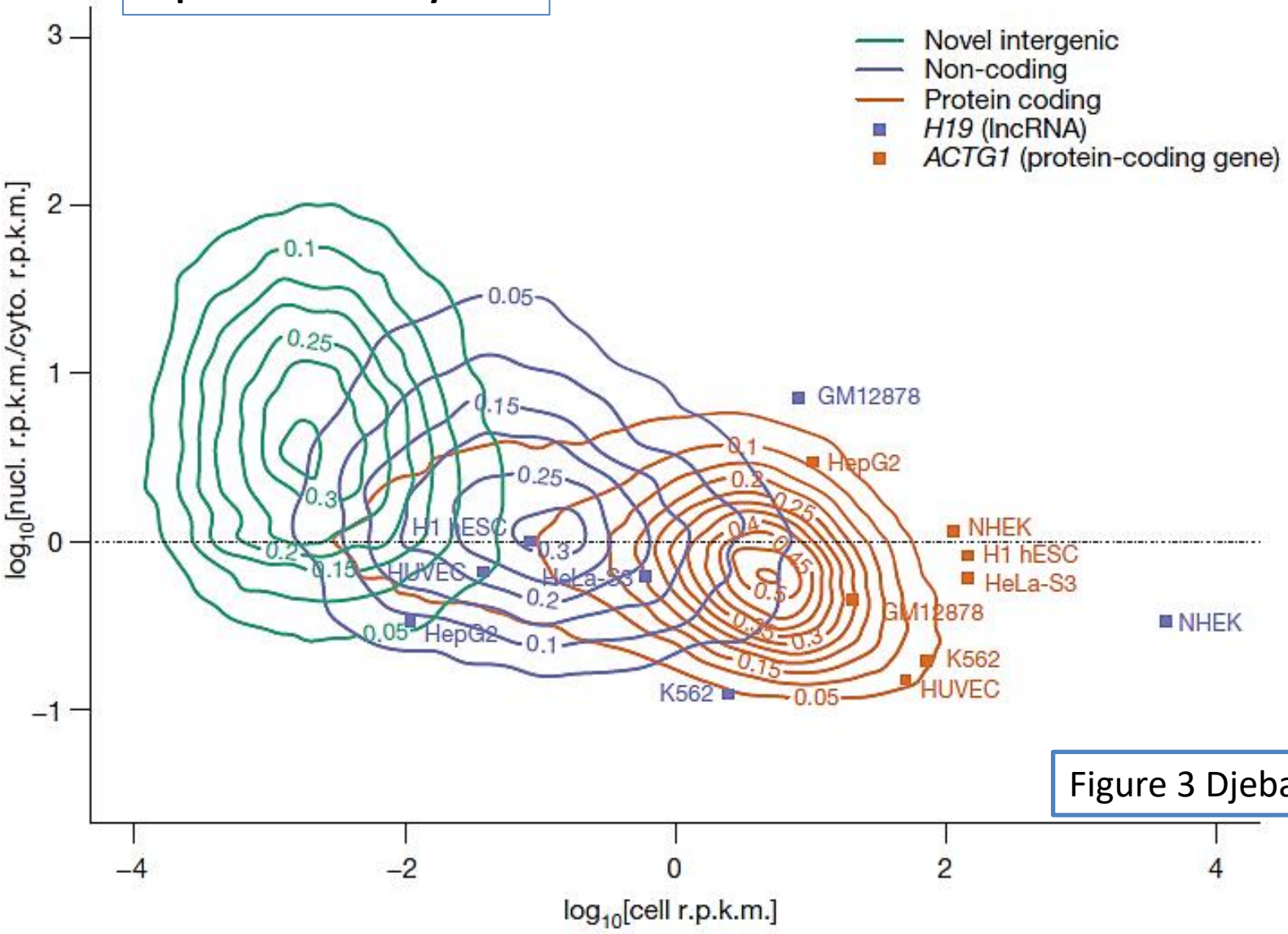


Figure 3 Djebali et al. 2012

Protein coding transcripts are the only class that is enriched in the cytoplasm

Trends in Genetics

CellPress

Textbook

Review

The Dimensions, Dynamics, and Relevance of the Mammalian Noncoding Transcriptome

Ira W. Deveson,[1,2] Simon A. Hardwick,[1,3] Tim R. Mercer,[1,3] and John S. Mattick[1,2,3,*]

464 Trends in Genetics, July 2017, Vol. 33, No. 7 http://dx.doi.org/10.1016/j.tig.2017.04.004

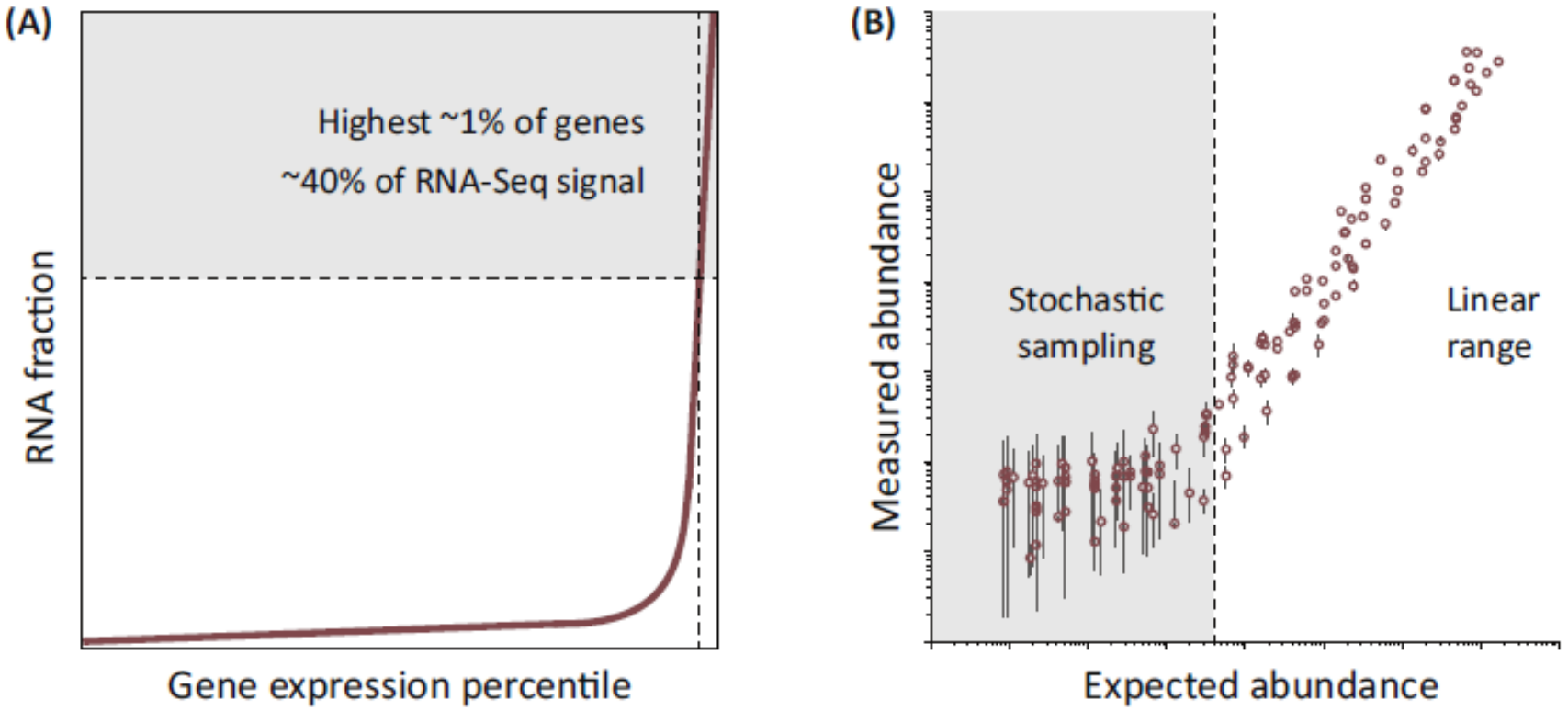
The proliferation and evolution of RNA-Seq, including the advent of methods for targeted, single-molecule, and single-cell sequencing, continues to enlarge our understanding of **transcriptional diversity**.

Textbook Deveson et al.

LncRNA are on average expressed at a lower range than protein-coding genes

To explore and understand less expressed lncRNAs, the Targeted RNA-seq method was developed. In practice, rare transcripts are selected using appropriate primers so that the sequencing library is enriched.

LncRNA databases vary greatly in number, and this is due to the criteria assumed to accept a lncRNA.
GENECODE is the most conservative,
MiTranscriptome lists 58,648 lncRNAs compared to 21,313 protein-coding


(A)
Highest ~1% of genes
~40% of RNA-Seq signal
RNA fraction
Gene expression percentile
(B)
Measured abundance
Stochastic sampling
Linear range
Expected abundance

Expression of lncRNAs is highly tissue-specific

ENCODE: 50% of the features were seen only in one cell line.

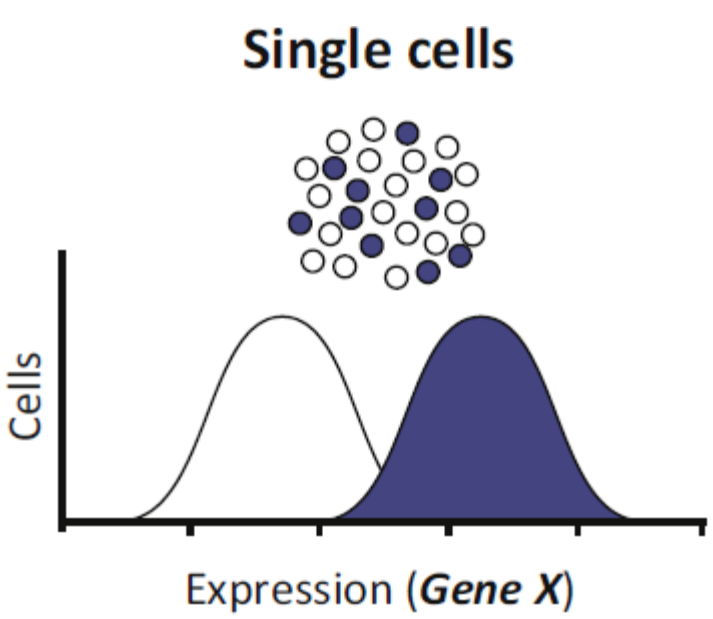
By FISH analysis: expression highly limited to cell types (in brain)

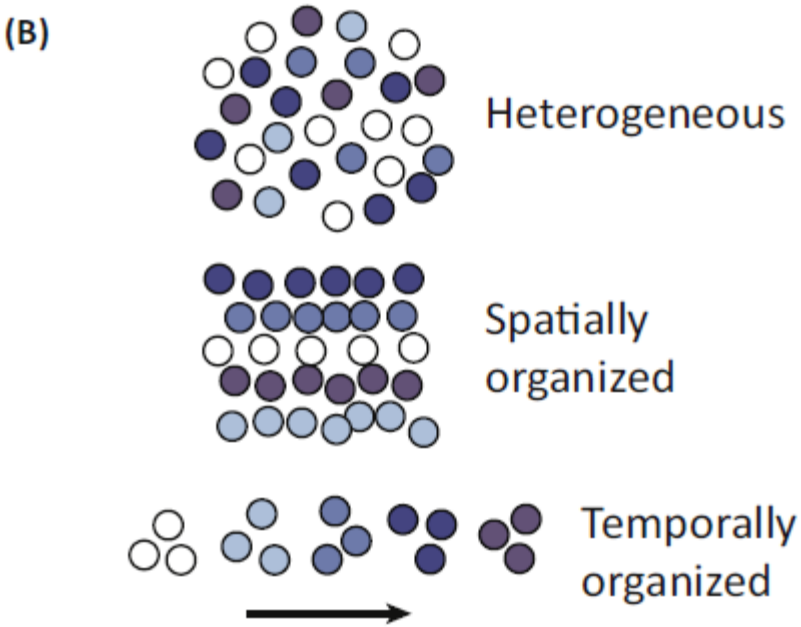
Single-cell RNA-Seq : lncRNAs expressed only in some cells.

Thus, average expression is low, but single-cell expression can be high

Cell-subtype determinants?

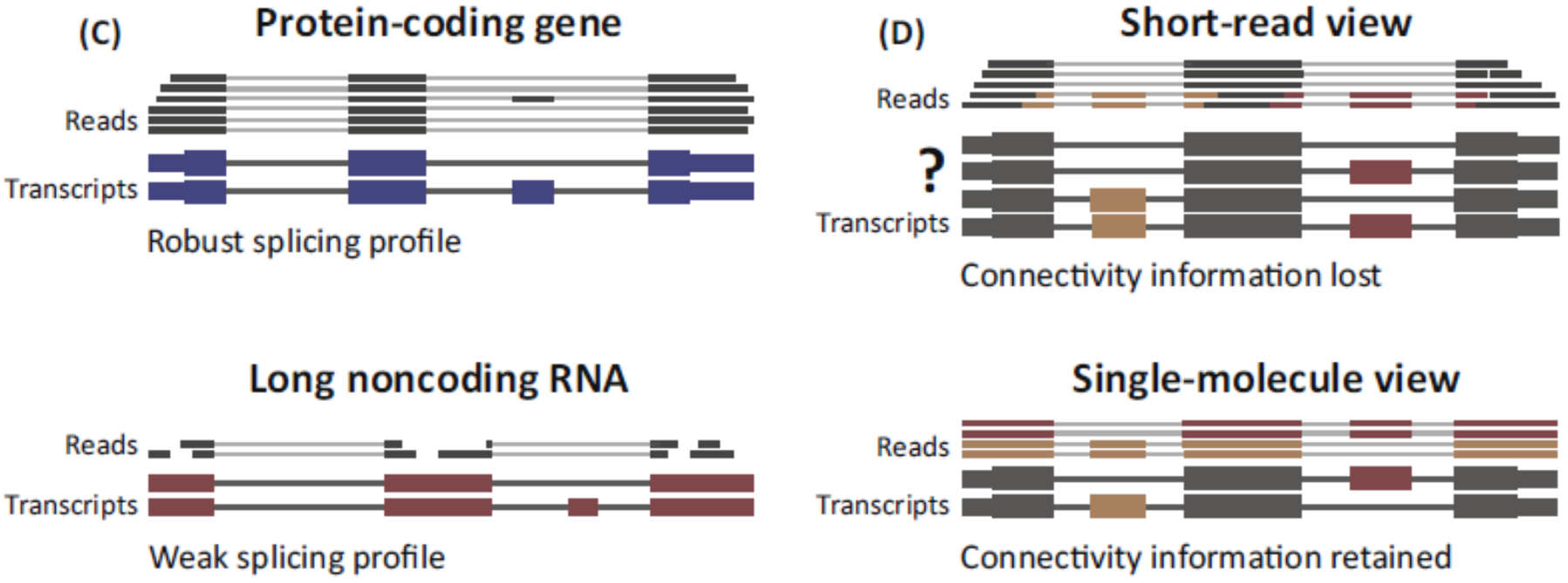
Diversity not seen in cell cultures, possibly since much more homogeneous.


Single cells
Cells
Expression (*Gene X*)


(B)
Heterogeneous
Spatially organized
Temporally organized

Most lncRNA have the same structure as protein-coding: Exons & Introns

Short-reads sequencing make it difficult to discriminate among transcripts



Definite improvements are expected for single-molecule, long-read sequencing technologies

Oxford Nanopore: https://vimeo.com/211385238

Pacific Bio:
https://www.pacb.com/smrt-science/smrt-sequencing/

Functional characterization: only few lncNAs

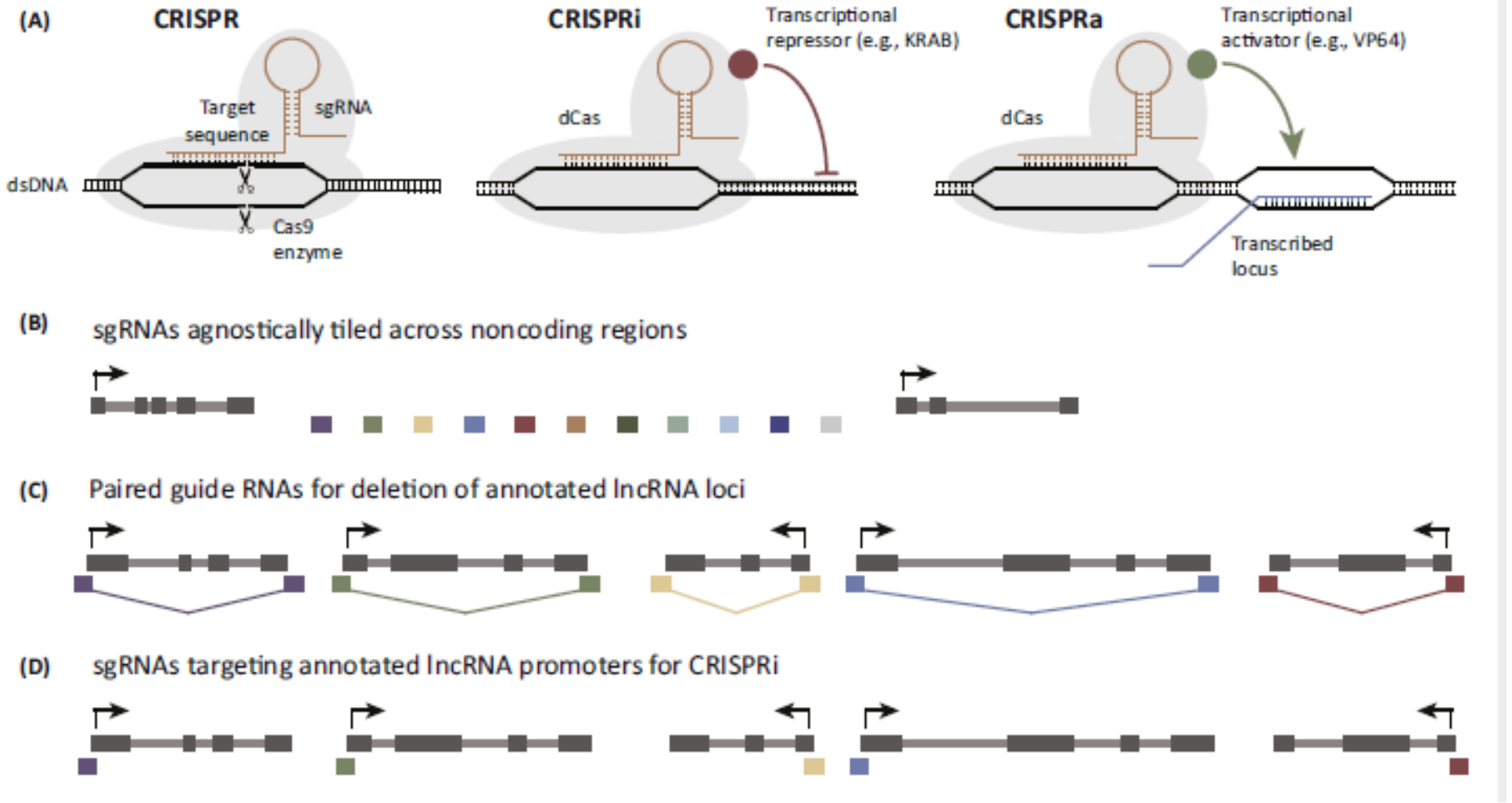
CRISPR screening are made today

Many lncRNA KO or KD → lethal phenotype

Many lncRNAs participate in <u>epigenomic regulation</u>

(examples from monoallelic expression lesson, interactiong with PRC2)

-HOTAIR binds both PRC2 and LSD1 (KDM) (repressor scaffolds)
-NEAT1 in paraspeckles
-Xsist in X-chr inactivation
-RNA-a at regulatory sequences (Enhancers?) – bind to Mediator for looping
-Other lncRNAs in imprinted regions (local repressor)



from Deveson et al. (TextBook)