Chapter 4 – Transcriptomes and post-transcriptional regulation

L4.1 - Transcriptomes

Transcriptomics

Post-genome

The key aims of **transcriptomics** are:

1) to catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs;
2) to determine the transcriptional structure of genes, in terms of their start sites, 5′ and 3′ ends, splicing patterns and other post-transcriptional modifications;
3) to quantify the changing expression levels of each transcript during development and under different experimental or pathological conditions

Medicine: gene expression profiles in disease; diagnosis, prognosis, guide to treatment

Pharma: drug effects evaluation, drug targets evaluation

RNA analysis, BASICS

Accessing to RNA:

1. Hybridization – based methods
2. Sequencing - based methods

1. *The RNA sequence is not observed directly, but it is inferred since it hybridizes with probes or primers.*
2. *The RNA sequence is converted to DNA (cDNA) and the DNA sequenced (*)*

Pre-genomic

In pre-genomic years, transcriptome was accessed only using single-transcript measurement (or few in parallel)

Quantitative:

Northern blotting

RNase Protection Assay (RPA)

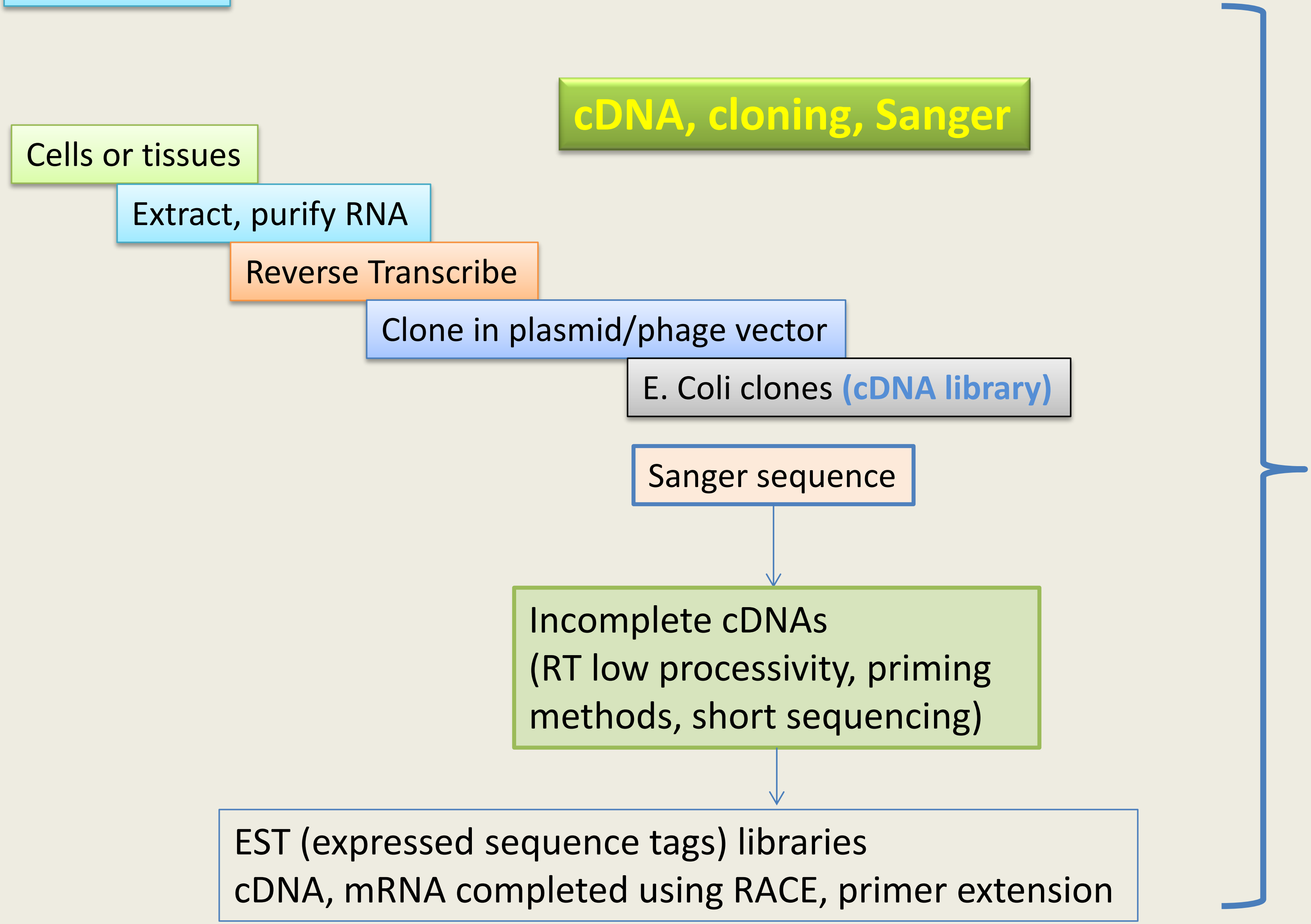
RT-PCR

qRT-PCR

Are they sequencing- or hybridization-based methods ?

Gene-by-gene methods to measure gene expression (mRNA)

Pre-genomic



cDNA, cloning, Sanger

Cells or tissues

Extract, purify RNA

Reverse Transcribe

Clone in plasmid/phage vector

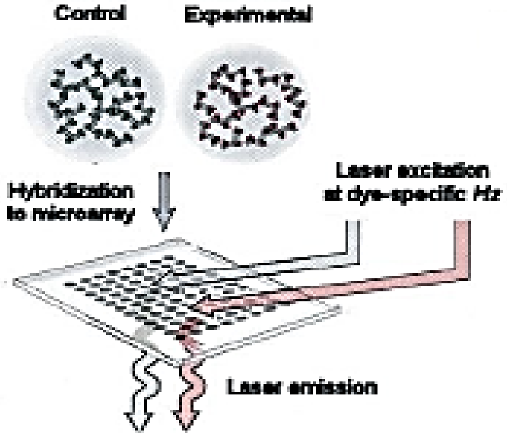
E. Coli clones **(cDNA library)**

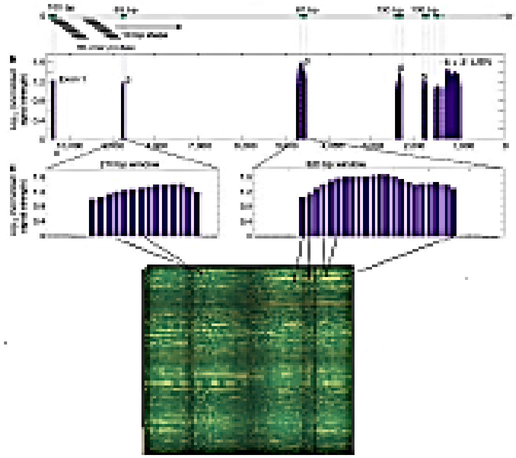
Sanger sequence

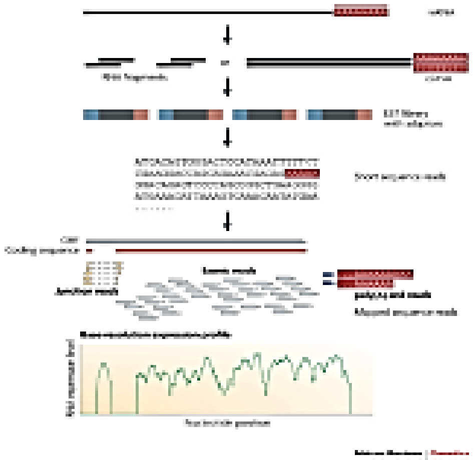
Incomplete cDNAs
(RT low processivity, priming methods, short sequencing)

EST (expressed sequence tags) libraries
cDNA, mRNA completed using RACE, primer extension

The evolution of transcriptomics


Control
Experimental
Hybridization to microarray
Laser excitation at dye-specific *Hz*
Laser emission





1995 P. Brown, et. al. Gene expression profiling using spotted cDNA microarray: expression levels of known genes

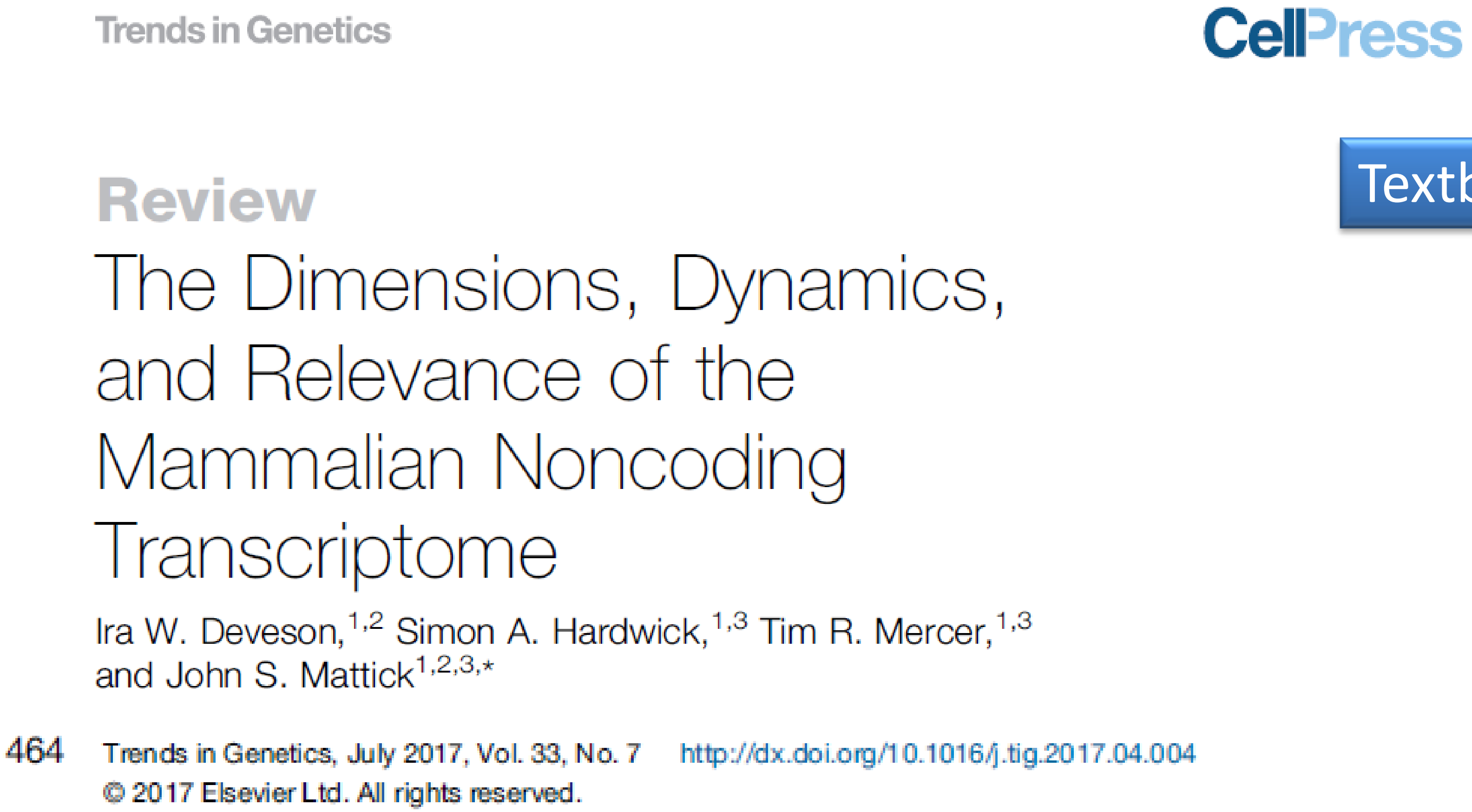
2002 Affymetrix, whole genome expression profiling using tiling array: identifying and profiling novel genes and splicing variants

2008 many groups, mRNA-seq: direct sequencing of mRNAs using next generation sequencing techniques (NGS)

Obviously microarray analysis is biased to previous knowledge of the transcripts.

Tiling microarrays have been used for transcript discovery, but with limited resolution, applicability and sensitivity

EST and SAGE methods have been used for unbiased analysis of the transcriptome. Major limitation: they describe only «parts» of the transcripts.



Trends in Genetics

CellPress

Review

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

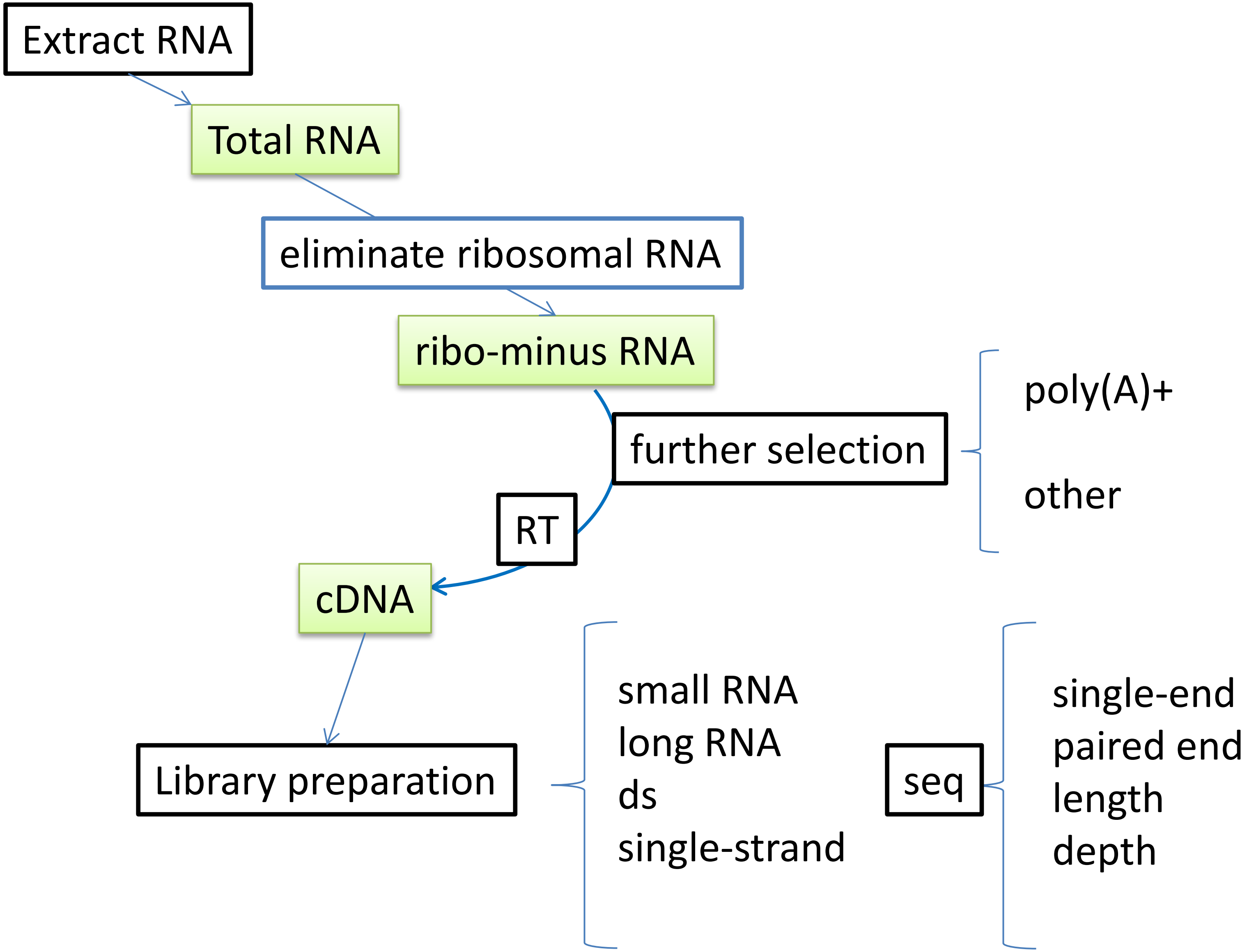
Ira W. Deveson,[1,2] Simon A. Hardwick,[1,3] Tim R. Mercer,[1,3] and John S. Mattick[1,2,3,*]

464 Trends in Genetics, July 2017, Vol. 33, No. 7 http://dx.doi.org/10.1016/j.tig.2017.04.004
© 2017 Elsevier Ltd. All rights reserved.

Textbook

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.

RNA-Seq



Extract RNA

Total RNA

eliminate ribosomal RNA

ribo-minus RNA

further selection

poly(A)+

other

RT

cDNA

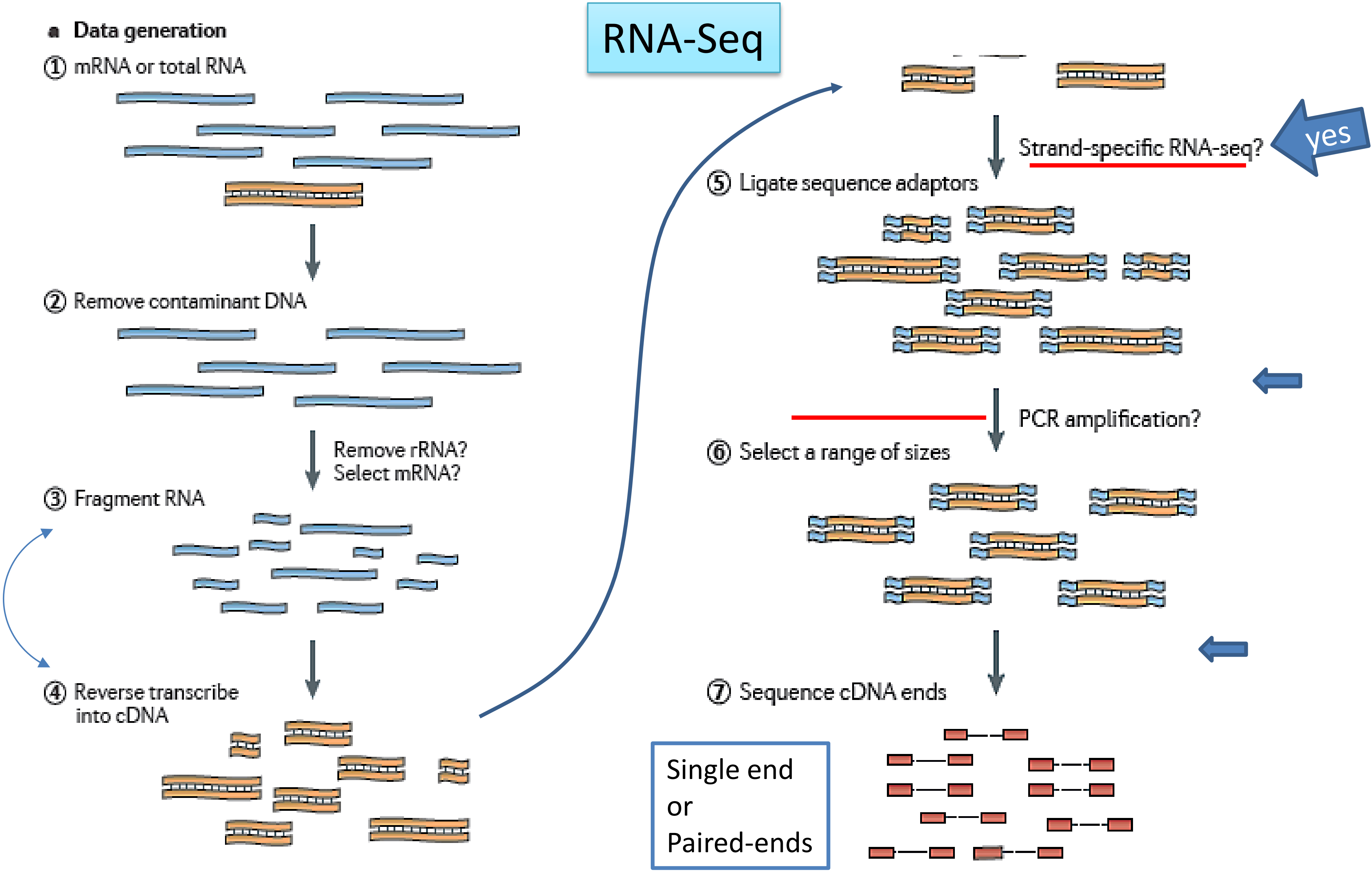
Library preparation

small RNA
long RNA
ds
single-strand

seq

single-end
paired end
length
depth

RNA-Seq



a Data generation

① mRNA or total RNA

② Remove contaminant DNA

Remove rRNA?
Select mRNA?

③ Fragment RNA

④ Reverse transcribe
into cDNA

Strand-specific RNA-seq?

yes

⑤ Ligate sequence adaptors

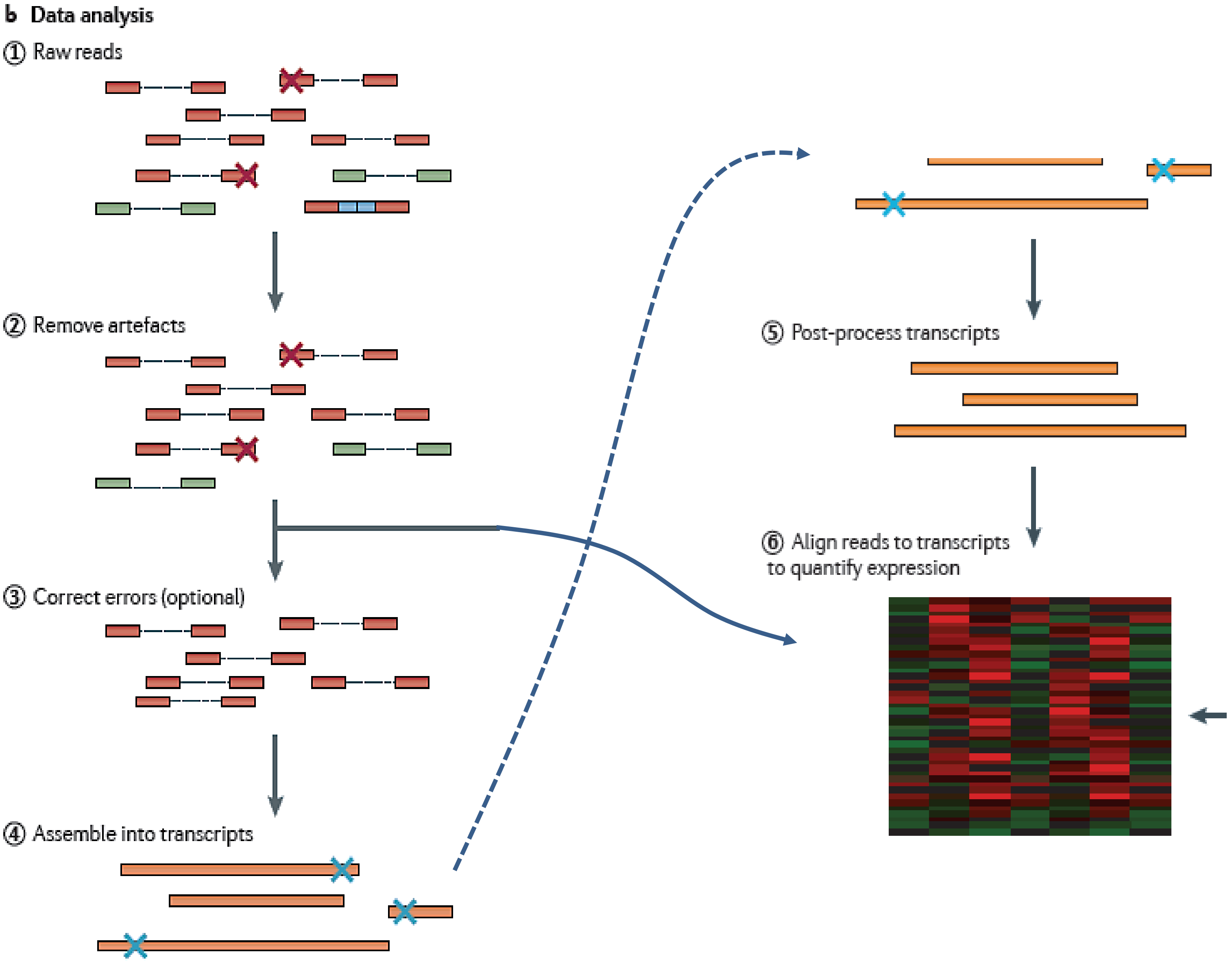
PCR amplification?

⑥ Select a range of sizes

⑦ Sequence cDNA ends

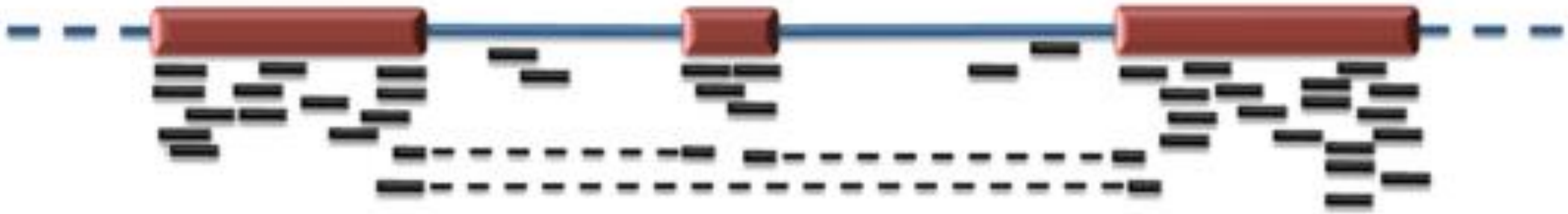
Single end
or
Paired-ends

Paired-end


b Data analysis
① Raw reads
② Remove artefacts
③ Correct errors (optional)
④ Assemble into transcripts
⑤ Post-process transcripts
⑥ Align reads to transcripts
to quantify expression

Heatmap

Mapping



Reads alignment to the genome

– Easy(ish) for genomic sequence

– Difficult for transcripts with splice junctions

Use of specific alignment tools
(i.e. Bowtie, Tophat, MapSplice…)

Quantitative (density over a region *or* transcription unit)

rpkm (reads per kilobase per million reads)

Double normalization for sequencing depth and gene length:

1- Divide the read counts by the “per million” scaling factor. This normalizes for sequencing depth, giving you reads per million (RPM)

2- Divide the RPM values by the length of the gene, in kilobases. This gives you RPKM.

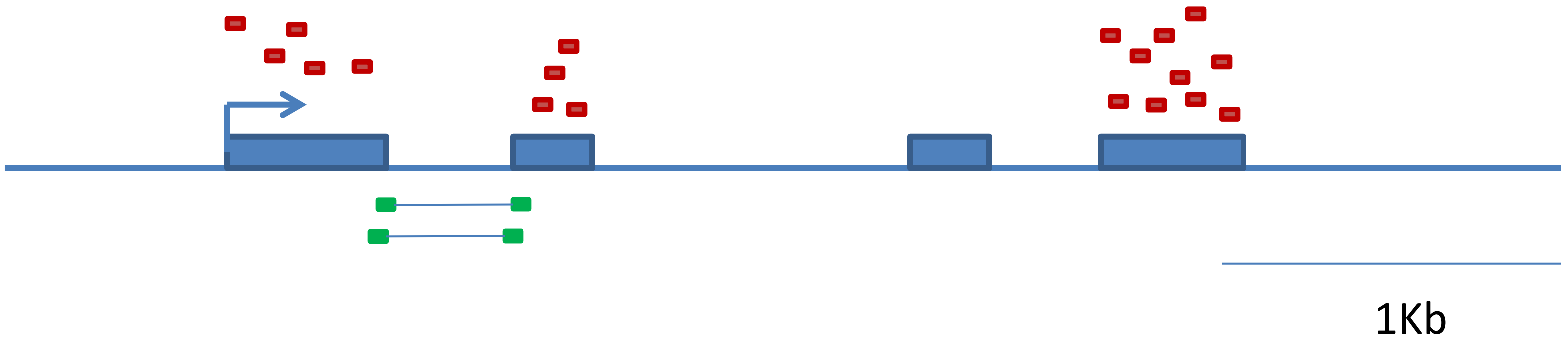
fpkm = fragments per kilobase per million
similar concept adapted for paired-end sequencing where two reads can map to one fragment

Quantitative

usually one reference set of «genes» (i.e. transcripts) is chosen and reads mapped to this.

then counts are taken by integrating all the reads falling in these models.

Caution: in the example below, one exon is not expressed. Nonetheless the gene is called «expressed»: algorithms should distinguish this and map to transcript isoforms instead of «genes».



One of the major variable in RNA-seq experiments (aside kind of RNA) is the sequencing depth = the number of reads («clusters» in Illumina sequencing) that you have decided to obtain for you sample.

Sequencing depth *versus* sensitivity

Always remember that the molecules you have sequenced are a «Sample» of the total possible reads from your biological sample.

How representative this sample is will depend on the number of molecules you have sequenced (i.e. the sequencing depth).

Increasing sequencing depth (higher coverage) helps identifying new transcripts

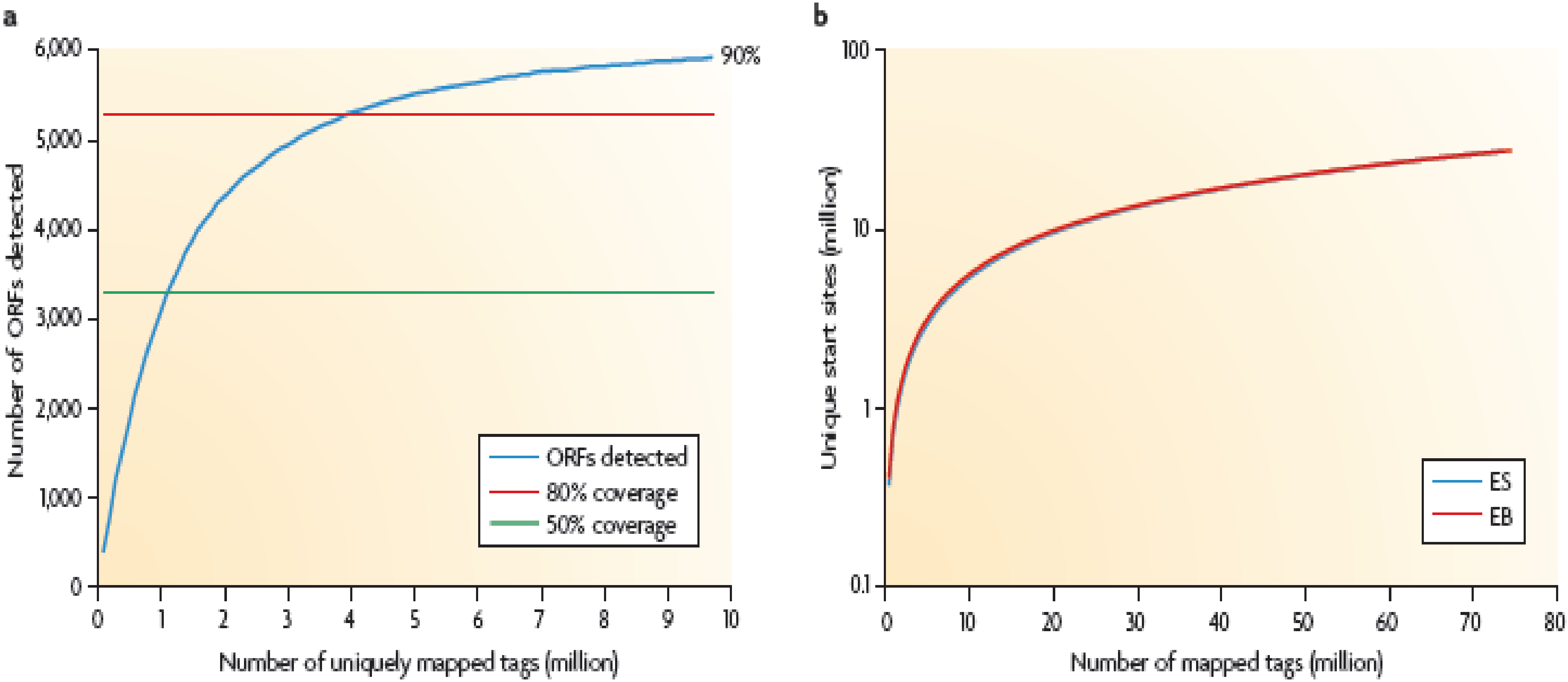


Figure 5 | **Coverage versus depth.** a | 80% of yeast genes were detected at 4 million uniquely mapped RNA-Seq reads, and coverage reaches a plateau afterwards despite the increasing sequencing depth. Expressed genes are defined as having at least four independent reads from a 50-bp window at the 3′ end. Data is taken from REF. 18. b | The number of unique start sites detected starts to reach a plateau when the depth of sequencing reaches 80 million in two mouse transcriptomes. ES, embryonic stem cells; EB, embryonic body. Figure is modified, with permission, from REF. 22 © (2008) Macmillan Publishers Ltd. All rights reserved.

Qualitative

Mapping

Reads are aligned to the reference genome, or to more limited reference of your choice:

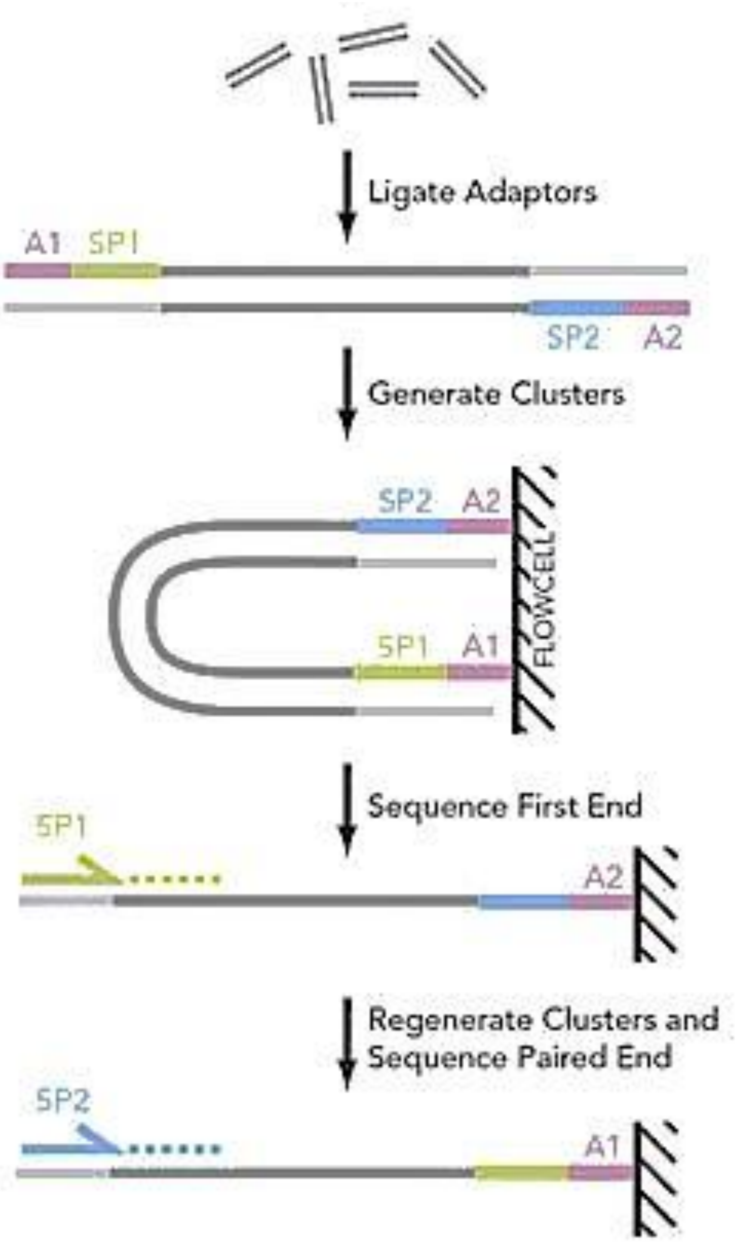
- known exons of protein-coding genes (exome)
- Spliced reads (*pay attention to this*!)
- Genes (sense and antisense)

New transcript definition

Common problems are difficulty in mapping reads can be solved by technical improvements:

- Longer reads
- Paired-end sequencing
- Strand-specific RNA-seq

Fragmented cDNA



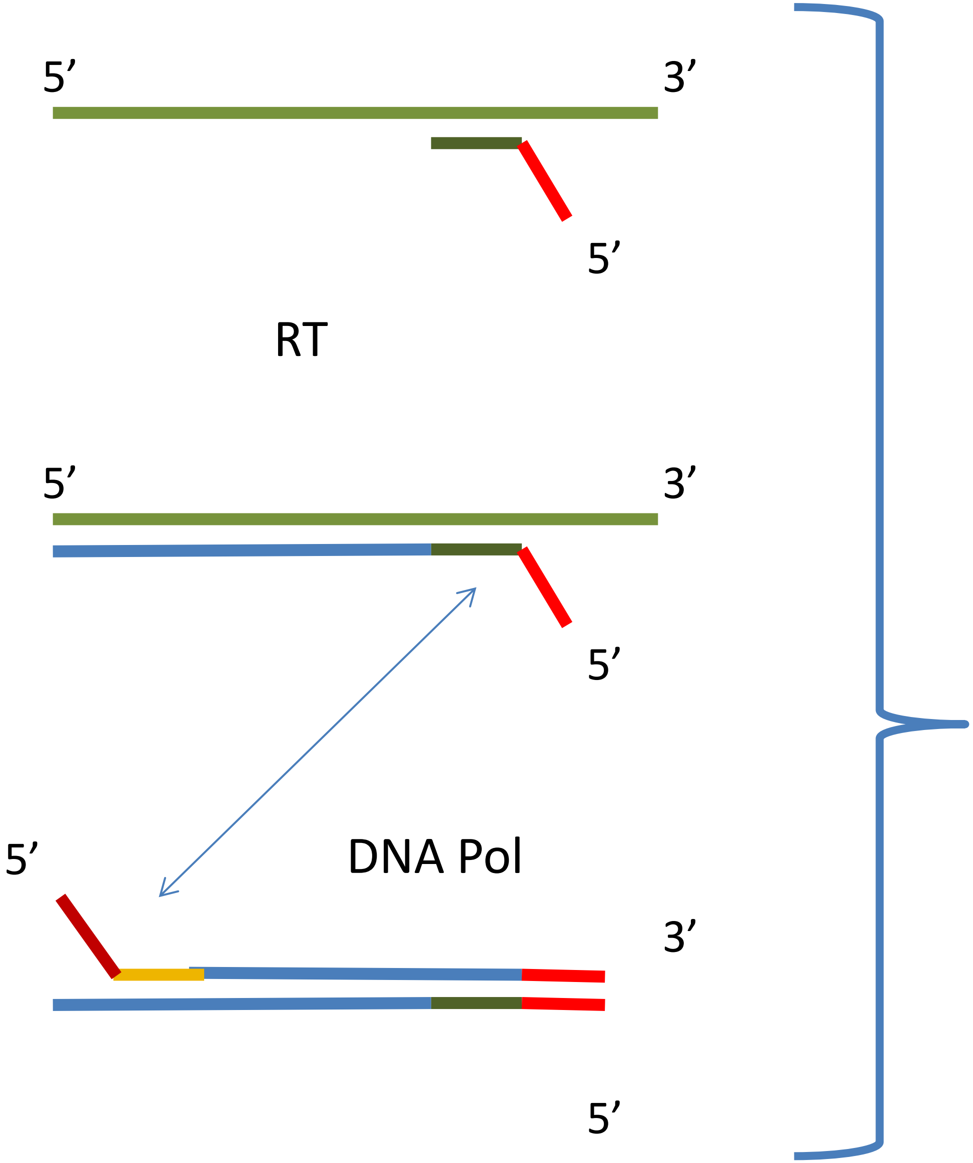
1. Longer reads

2. Paired-end sequencing (from Illumina)

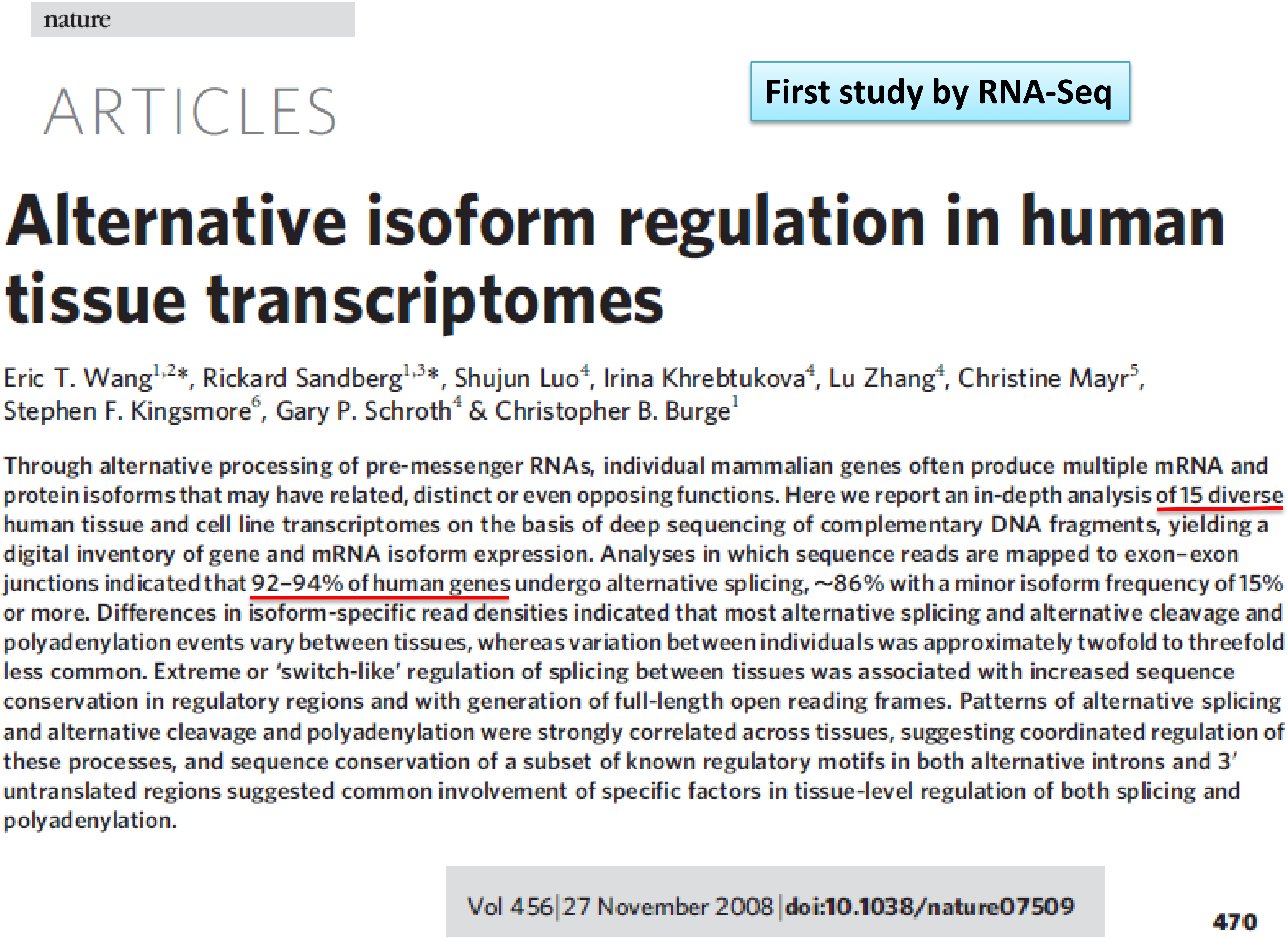
(Alternative methods exist)

Paired-end sequencing

Strand-specific library



First study by RNA-Seq



nature

ARTICLES

Alternative isoform regulation in human tissue transcriptomes

Eric T. Wang[1,2]*, Rickard Sandberg[1,3]*, Shujun Luo[4], Irina Khrebtukova[4], Lu Zhang[4], Christine Mayr[5], Stephen F. Kingsmore[6], Gary P. Schroth[4] & Christopher B. Burge[1]

Through alternative processing of pre-messenger RNAs, individual mammalian genes often produce multiple mRNA and protein isoforms that may have related, distinct or even opposing functions. Here we report an in-depth analysis of 15 diverse human tissue and cell line transcriptomes on the basis of deep sequencing of complementary DNA fragments, yielding a digital inventory of gene and mRNA isoform expression. Analyses in which sequence reads are mapped to exon–exon junctions indicated that 92–94% of human genes undergo alternative splicing, ~86% with a minor isoform frequency of 15% or more. Differences in isoform-specific read densities indicated that most alternative splicing and alternative cleavage and polyadenylation events vary between tissues, whereas variation between individuals was approximately twofold to threefold less common. Extreme or 'switch-like' regulation of splicing between tissues was associated with increased sequence conservation in regulatory regions and with generation of full-length open reading frames. Patterns of alternative splicing and alternative cleavage and polyadenylation were strongly correlated across tissues, suggesting coordinated regulation of these processes, and sequence conservation of a subset of known regulatory motifs in both alternative introns and 3′ untranslated regions suggested common involvement of specific factors in tissue-level regulation of both splicing and polyadenylation.

Vol 456|27 November 2008|**doi:10.1038/nature07509**

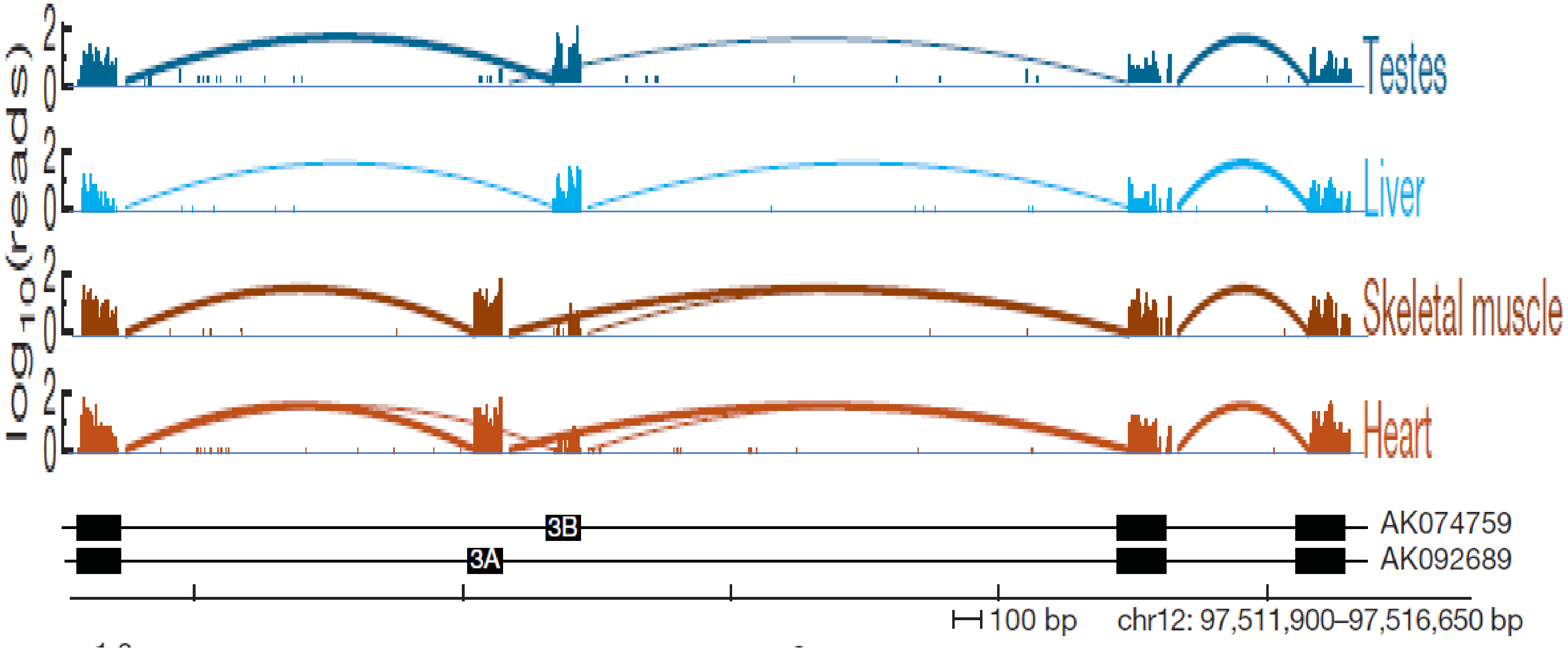
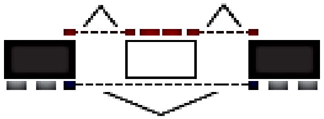
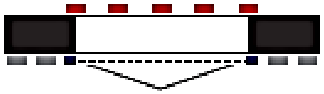
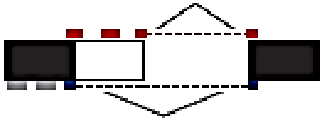
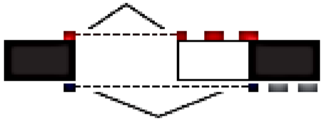
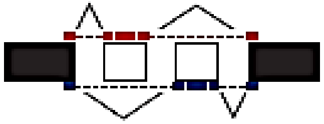
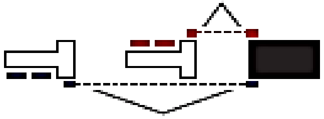
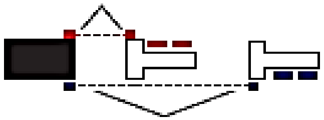
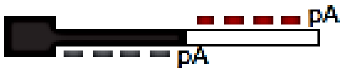


Figure 1 | Frequency and relative abundance of alternative splicing isoforms in human genes.

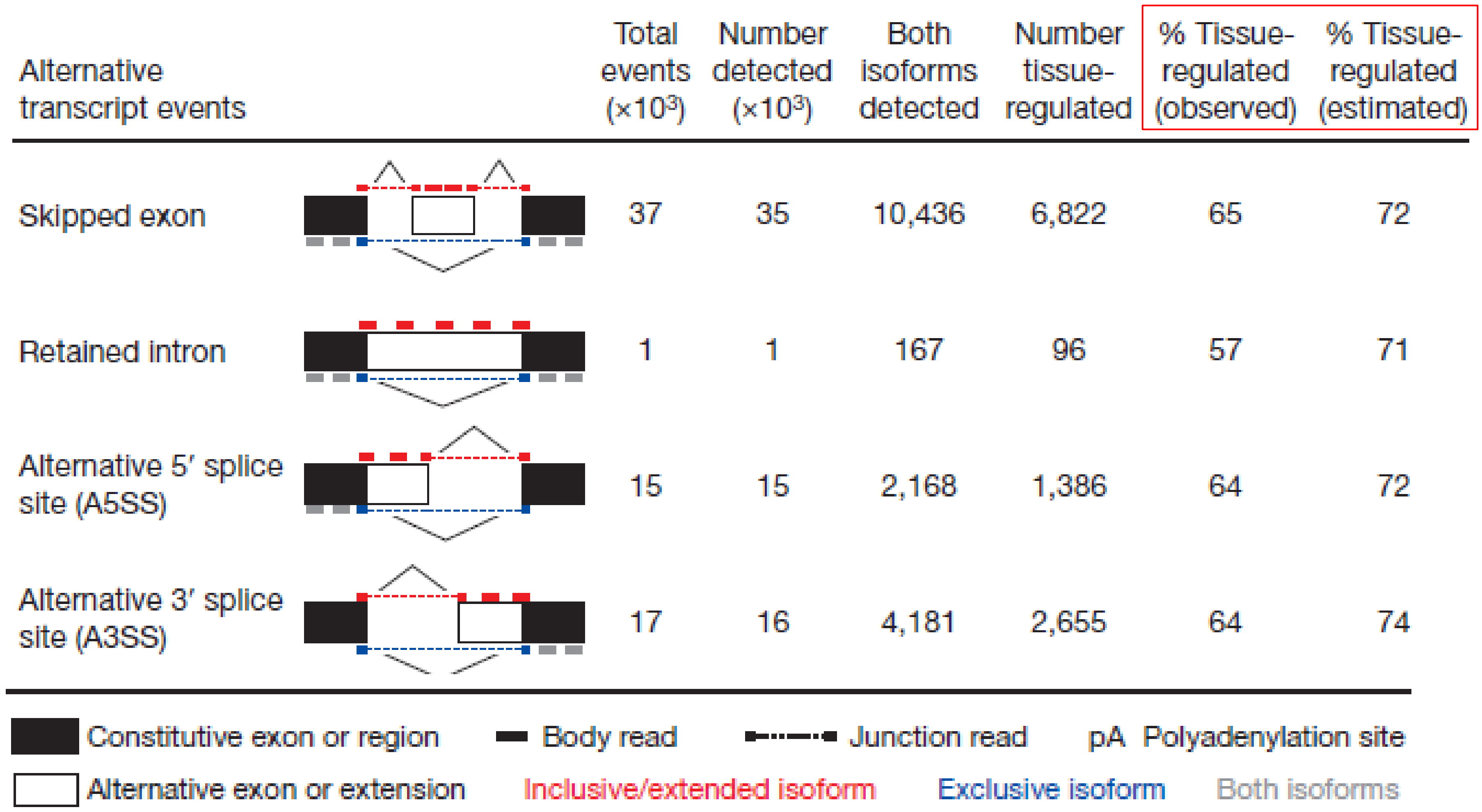
a, mRNA-Seq reads mapping to a portion of the SLC25A3 gene locus. The number of mapped reads starting at each nucleotide position is displayed (log10) for the tissues listed at the right. Arcs represent junctions detected by splice junction reads.

Bottom: exon/intron structures of representative transcripts containing mutually exclusive exons 3A and 3B (GenBank accession numbers shown at the right).

Alternative transcript events		Total events (×10³)	Number detected (×10³)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon		37	35	10,436	6,822	65	72
Retained intron		1	1	167	96	57	71
Alternative 5′ splice site (A5SS)		15	15	2,168	1,386	64	72
Alternative 3′ splice site (A3SS)		17	16	4,181	2,655	64	74
Mutually exclusive exon (MXE)		4	4	167	95	57	66
Alternative first exon (AFE)		14	13	10,281	5,311	52	63
Alternative last exon (ALE)		9	8	5,246	2,491	47	52
Tandem 3′ UTRs		7	7	5,136	3,801	74	80
Total		105	100	37,782	22,657	60	68

■ Constitutive exon or region ▬ Body read ■-------■ Junction read pA Polyadenylation site

□ Alternative exon or extension Inclusive/extended isoform Exclusive isoform Both isoforms

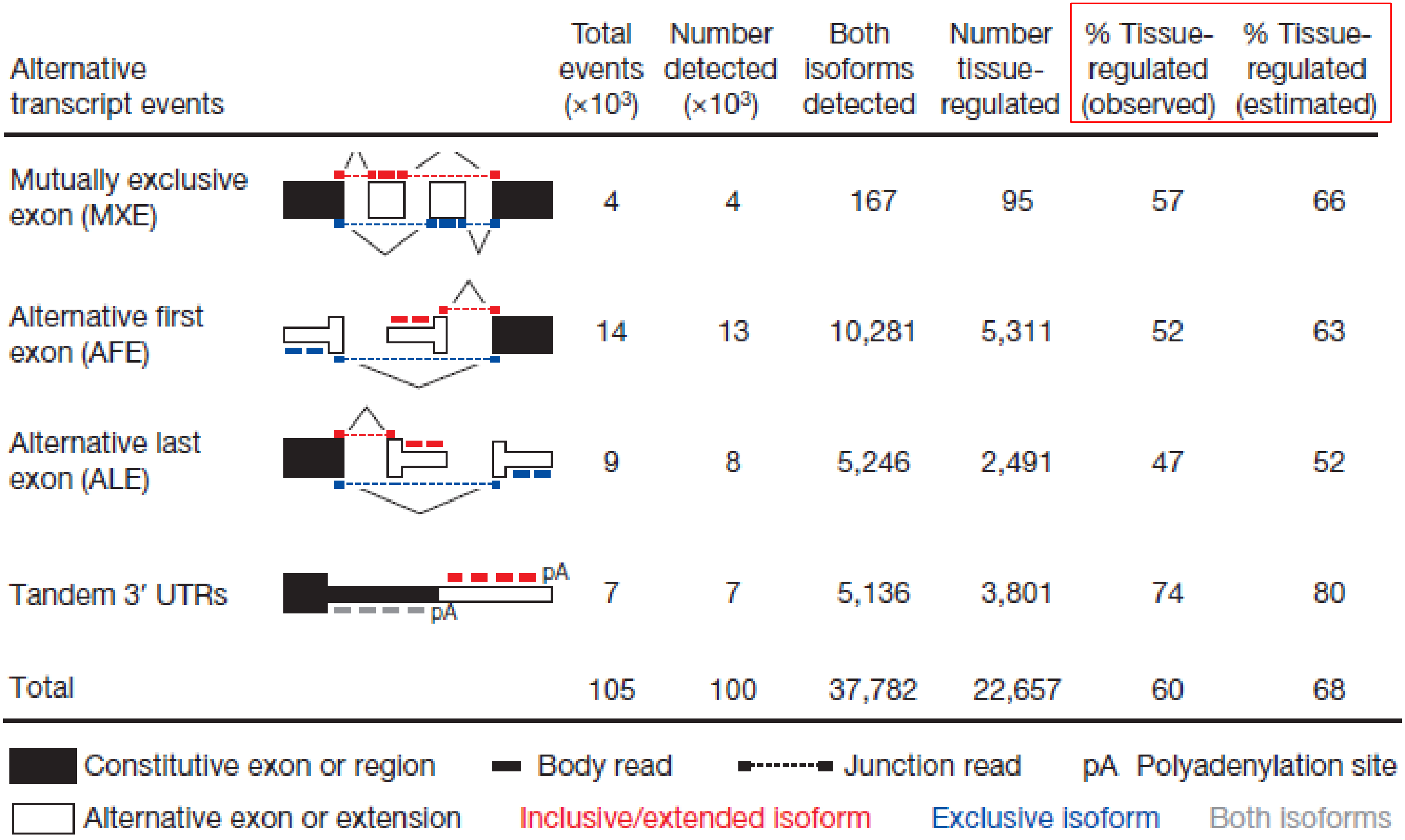


Alternative transcript events	Total events ($\times 10^3$)	Number detected ($\times 10^3$)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon	37	35	10,436	6,822	65	72
Retained intron	1	1	167	96	57	71
Alternative 5′ splice site (A5SS)	15	15	2,168	1,386	64	72
Alternative 3′ splice site (A3SS)	17	16	4,181	2,655	64	74

■ Constitutive exon or region ▬ Body read ■┅▬┅■ Junction read pA Polyadenylation site

□ Alternative exon or extension Inclusive/extended isoform Exclusive isoform Both isoforms

Figure 2 | Pervasive tissue-specific regulation of alternative mRNA isoforms. Rows represent the eight different alternative transcript event types diagrammed. Mapped reads supporting expression of upper isoform, lower isoform or both isoforms are shown in blue, red and grey, respectively. Columns 1–4 show the numbers of events of each type: (1) supported by cDNA and/or EST data; (2) with ≥ 1 isoform supported by mRNA-Seq reads; (3) with both isoforms supported by reads; and (4) events detected as tissue regulated (Fisher's exact test) at an FDR of 5% (assuming negligible technical variation).



Alternative transcript events	Total events (×10^3)	Number detected (×10^3)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Mutually exclusive exon (MXE)	4	4	167	95	57	66
Alternative first exon (AFE)	14	13	10,281	5,311	52	63
Alternative last exon (ALE)	9	8	5,246	2,491	47	52
Tandem 3′ UTRs	7	7	5,136	3,801	74	80
Total	105	100	37,782	22,657	60	68

■ Constitutive exon or region ▬ Body read ■-------■ Junction read pA Polyadenylation site

□ Alternative exon or extension Inclusive/extended isoform Exclusive isoform Both isoforms

Columns 5 and 6 show: (5) the observed percentage of events with both isoforms detected that were observed to be tissue-regulated; and (6) the estimated true percentage of tissue-regulated isoforms after correction for power to detect tissue bias (Supplementary Fig. 6) and for the FDR. For some event types, 'common reads' (grey bars) were used in lieu of (for tandem 39UTR events) or in addition to 'exclusion' reads for detection of changes in isoform levels between tissues.
Note that Aa use the following definition for "tissue-specific": at least 10% variation in isoforms.

This paper describes a number of «known» features of genes

1) the usage of Alternative Promoters

2) Alternative splicing of internal exons

3) the usage of alternative polyadenylation sites

The real news is the frequency and extension of these phenomena

What do we know of Alternative Splicing ?

Alternative Splicing

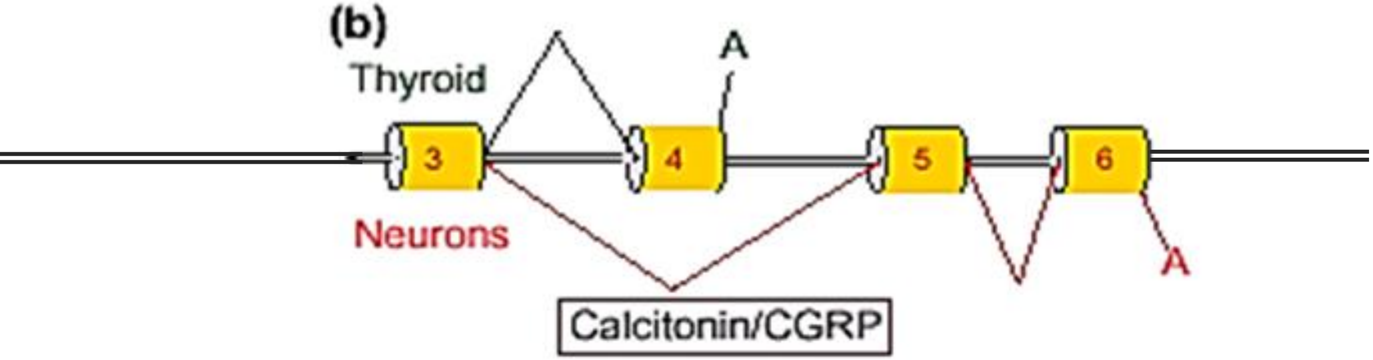
It was discovered in Viruses more than 40 years ago

In 1983 a Vertebrate gene discovered to make AS, producing two different mRNAs encoding for Calcitonin and CGRP

Several other gene transcripts shown to undergo AS

in 2000, the insect Dscam gene transcript was shown to undergo a complex AS combination among a number of mutally exclusive exons, leading potentially to 38,000 different mRNAs.

In 2008-2012 most Human genes show to undergo AS among the first discovered …

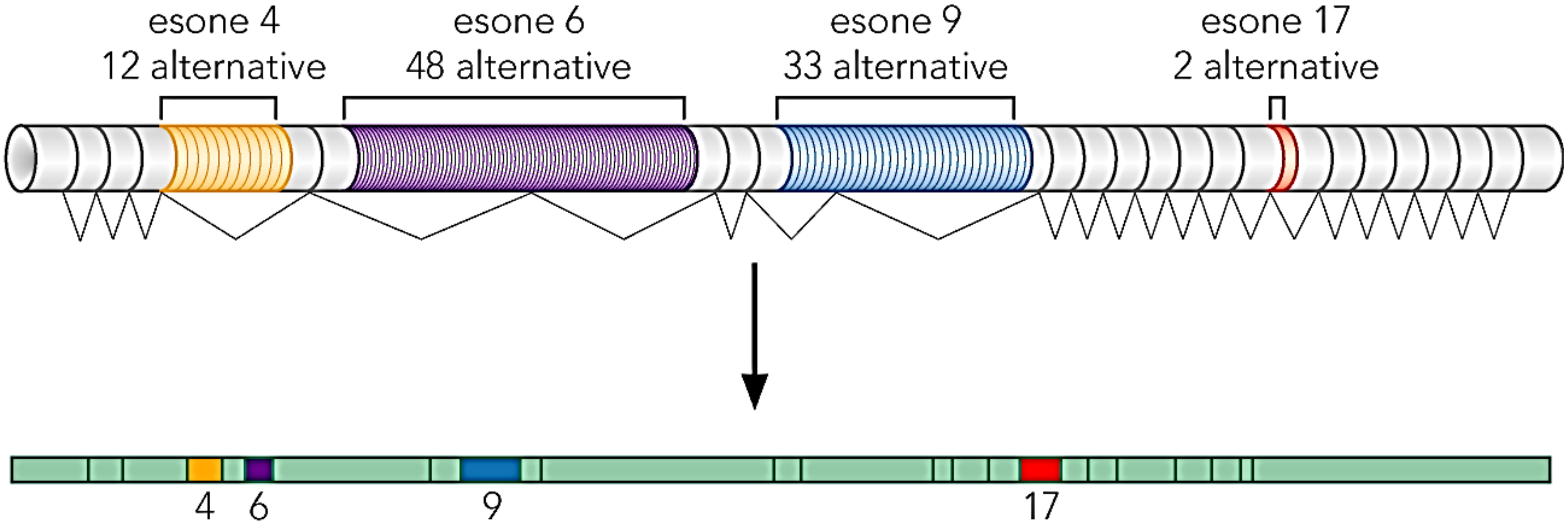


Calcitonin is a 32-aminoacid peptide hormone that is produced by the parafollicular tyroid cells in Humans. The first function of calcitonin is homeostatic: it lowers calcium concentration in blood.

CGRP is produced in both peripheral and central neurons. It is a potent peptide vasodilator and can function in the transmission of pain. In the spinal cord, the function and expression of CGRP may differ depending on the location of synthesis.

The Dscam gene in D. melanogaster

primary transcript



Potentially 38,000 splicing variants

Vertebrate Dscam1 does **not** behave in a similar way.

Not all exons in a gene can undergo «alternative» splicing

Alternative exons = weak exons = average one-two per gene

«Pure» alternative splicing

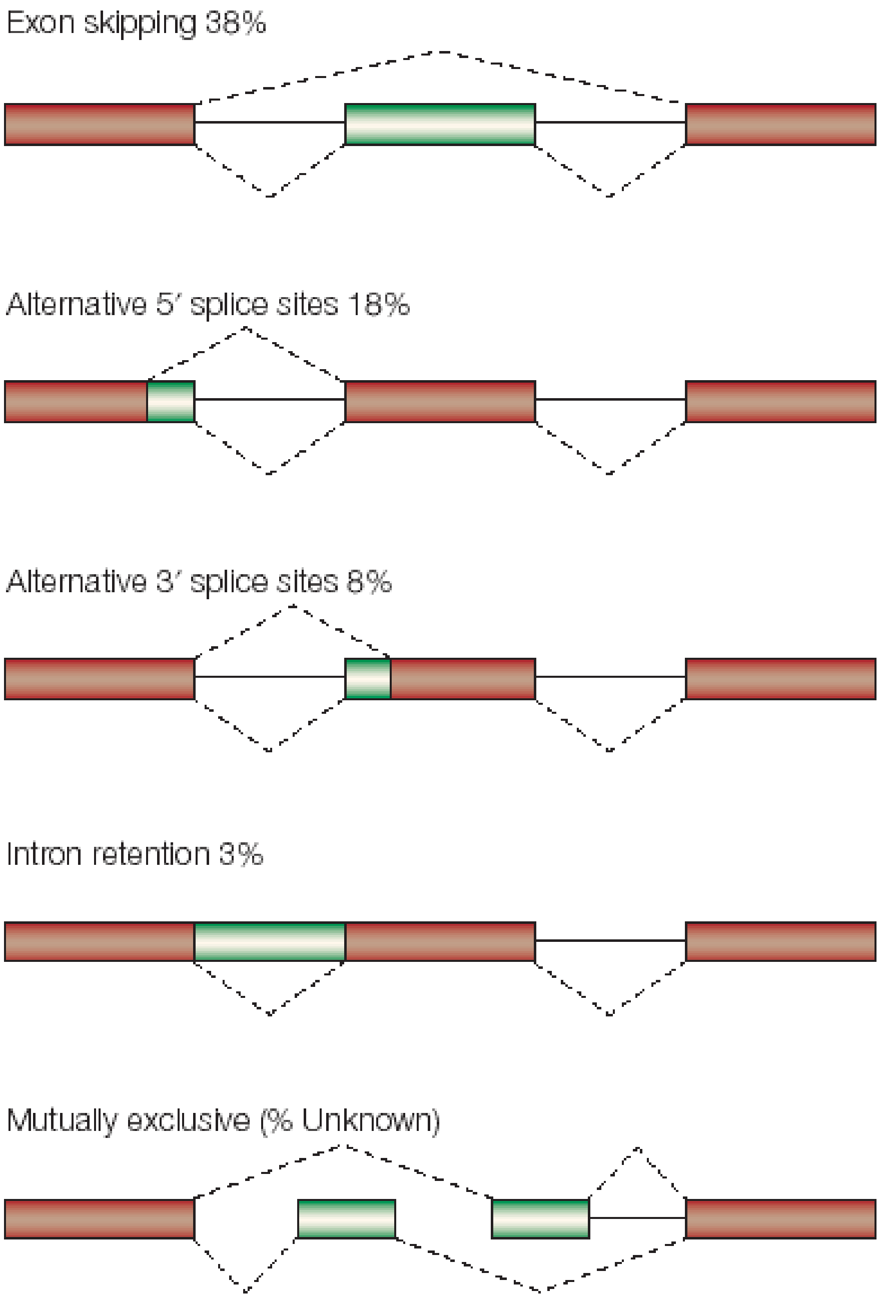

Exon skipping 38%
Alternative 5′ splice sites 18%
Alternative 3′ splice sites 8%
Intron retention 3%
Mutually exclusive (% Unknown)

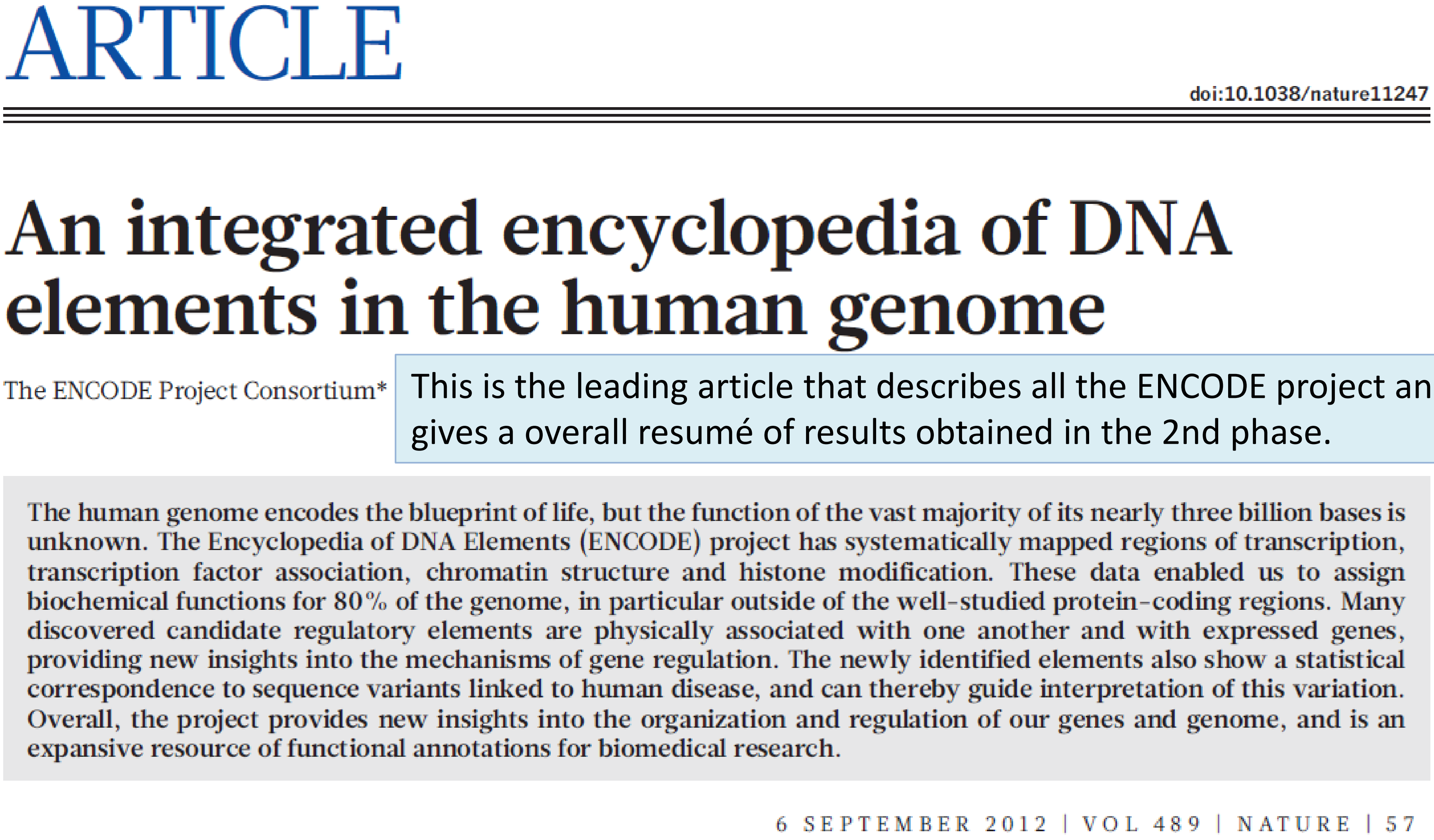
Figure 3

Types of alternative splicing.

In all five examples of alternative splicing, constitutive exons are shown in red and alternatively spliced regions in green, introns are represented by solid lines, and dashed lines indicate splicing activities. Relative abundance of alternative splicing events that are conserved between human and mouse transcriptomes are shown above each example (in % of total alternative splicing events).

From: Ast G. (**2004**)
Nature Rev Genetics 5: 773.

Note that the indicated percentages derive from older studies and are slightly different from those demonstrated by recent, RNA-Seq based evaluations



ARTICLE

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

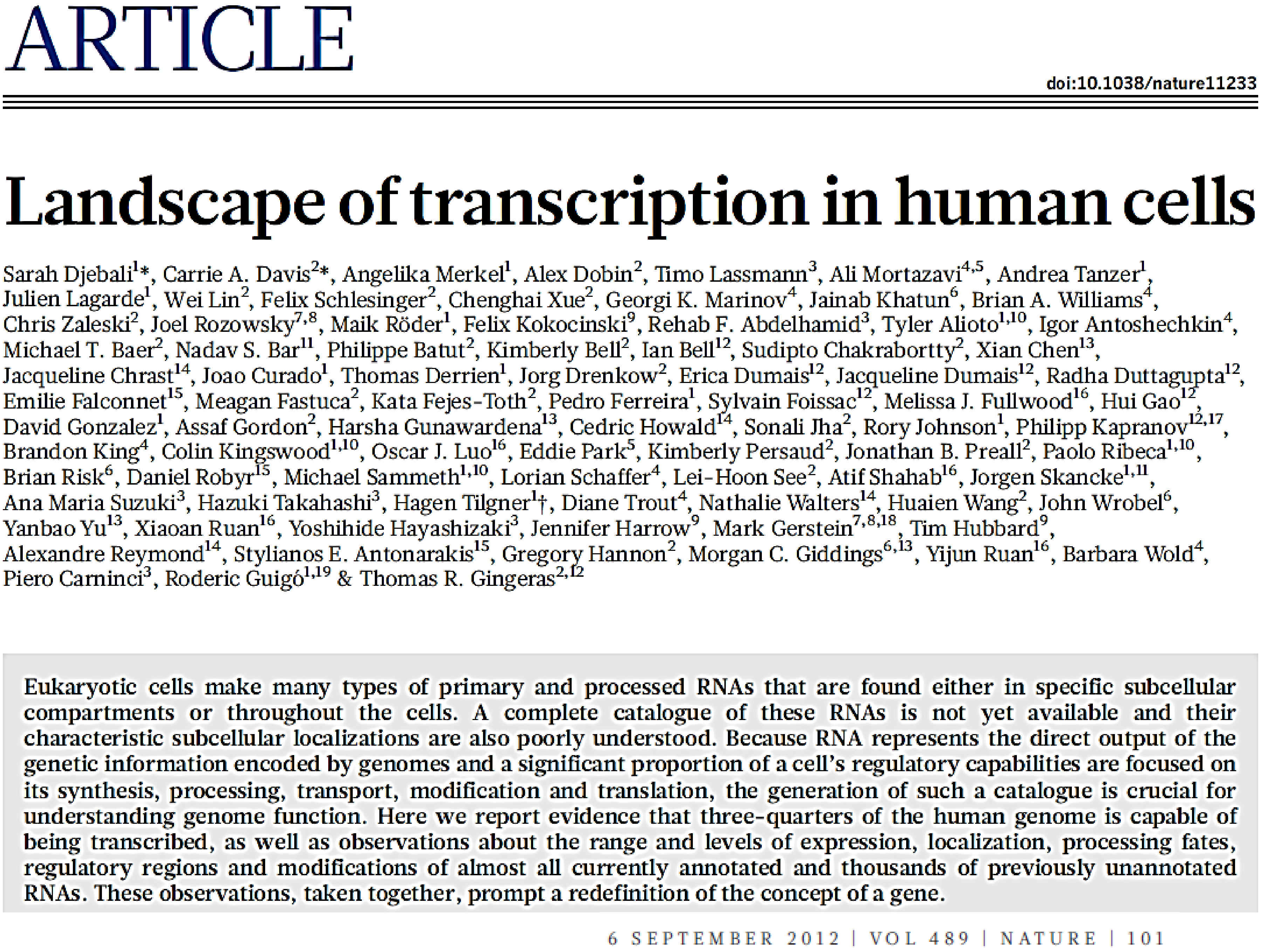
The ENCODE Project Consortium*

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.

6 SEPTEMBER 2012 | VOL 489 | NATURE | 57

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

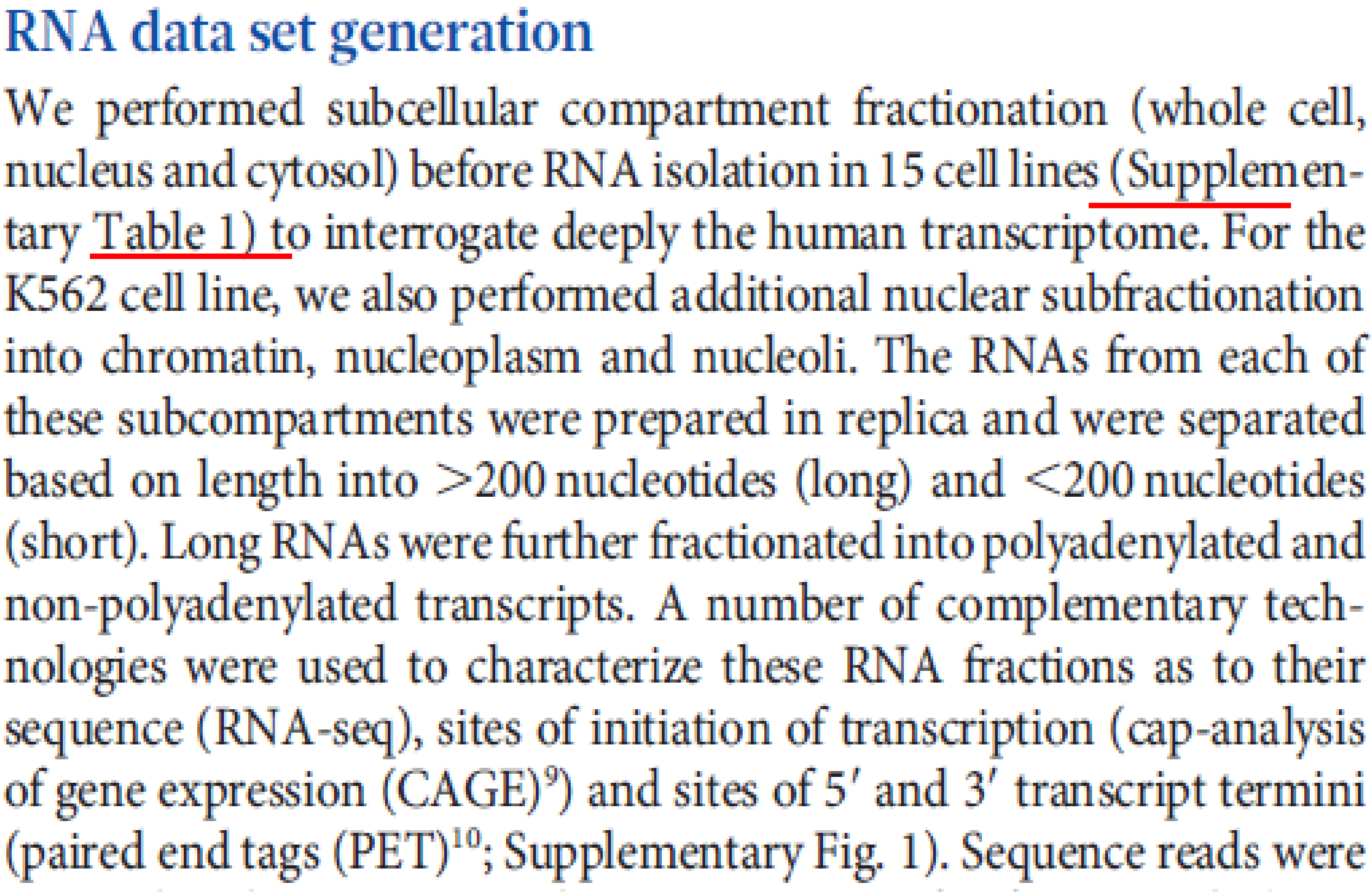
Djebali et al., 2012

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» ?



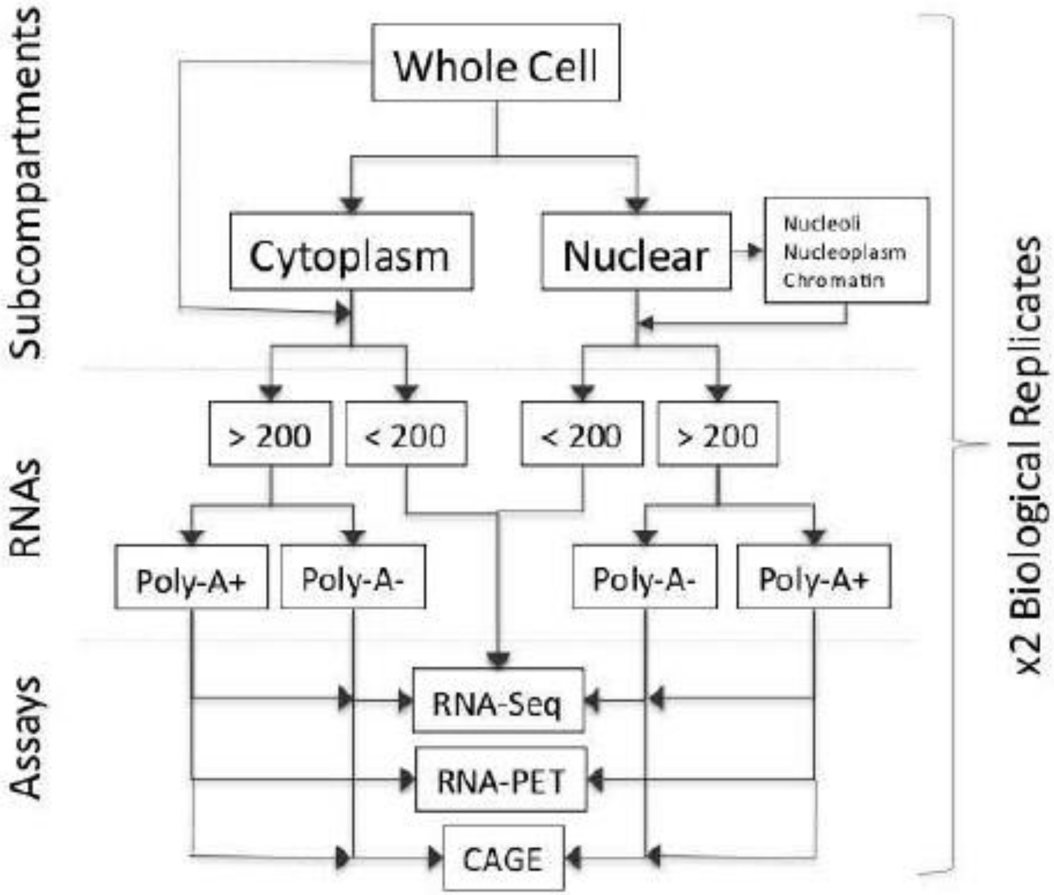
RNA data set generation

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

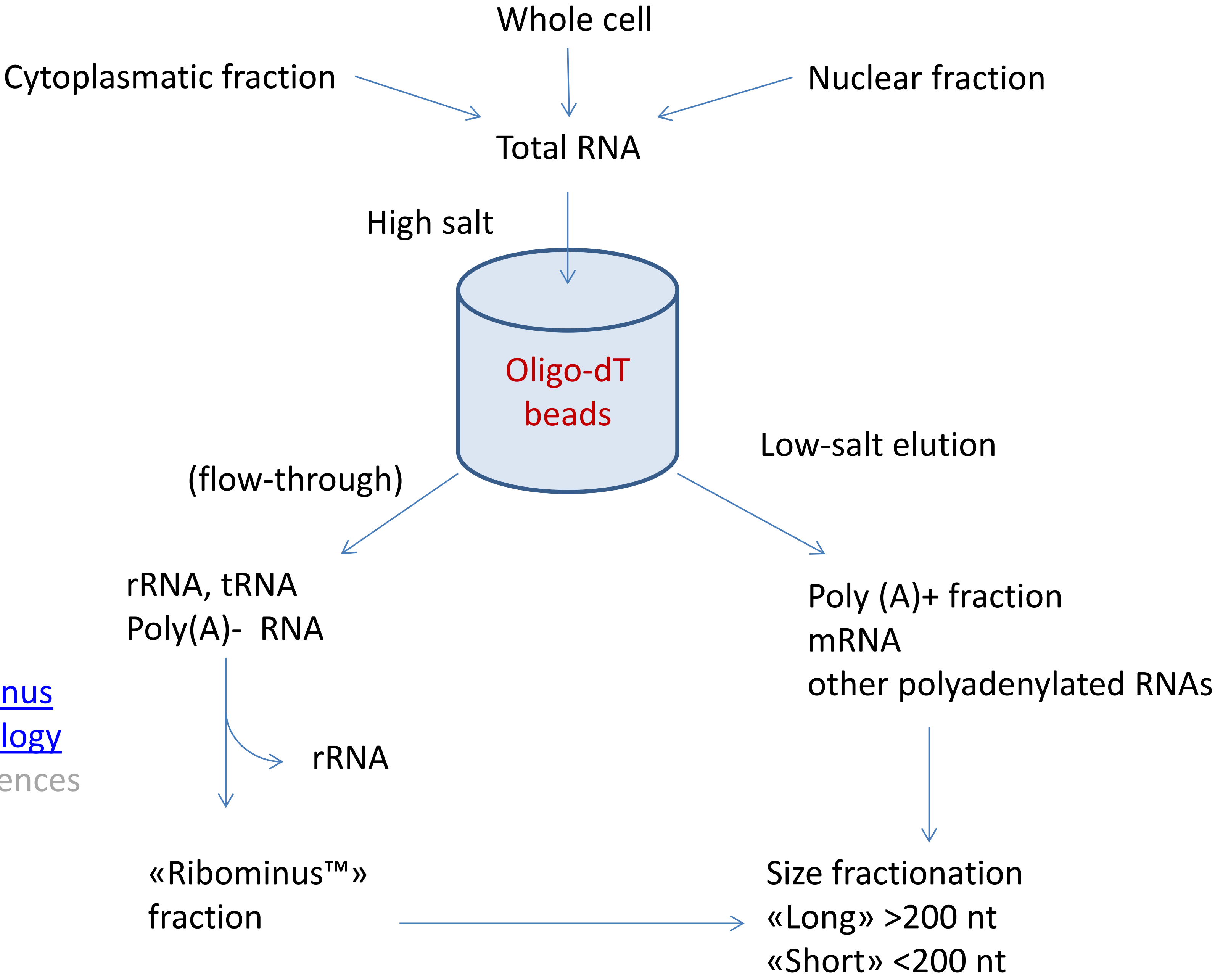
RNA-PET captures and sequences the 5′- and 3′-end tags of full-length cDNA fragments of all expressed genes in a biological sample



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

(flow-through)

Low-salt elution

rRNA, tRNA
Poly(A)- RNA

Poly (A)+ fraction
mRNA
other polyadenylated RNAs

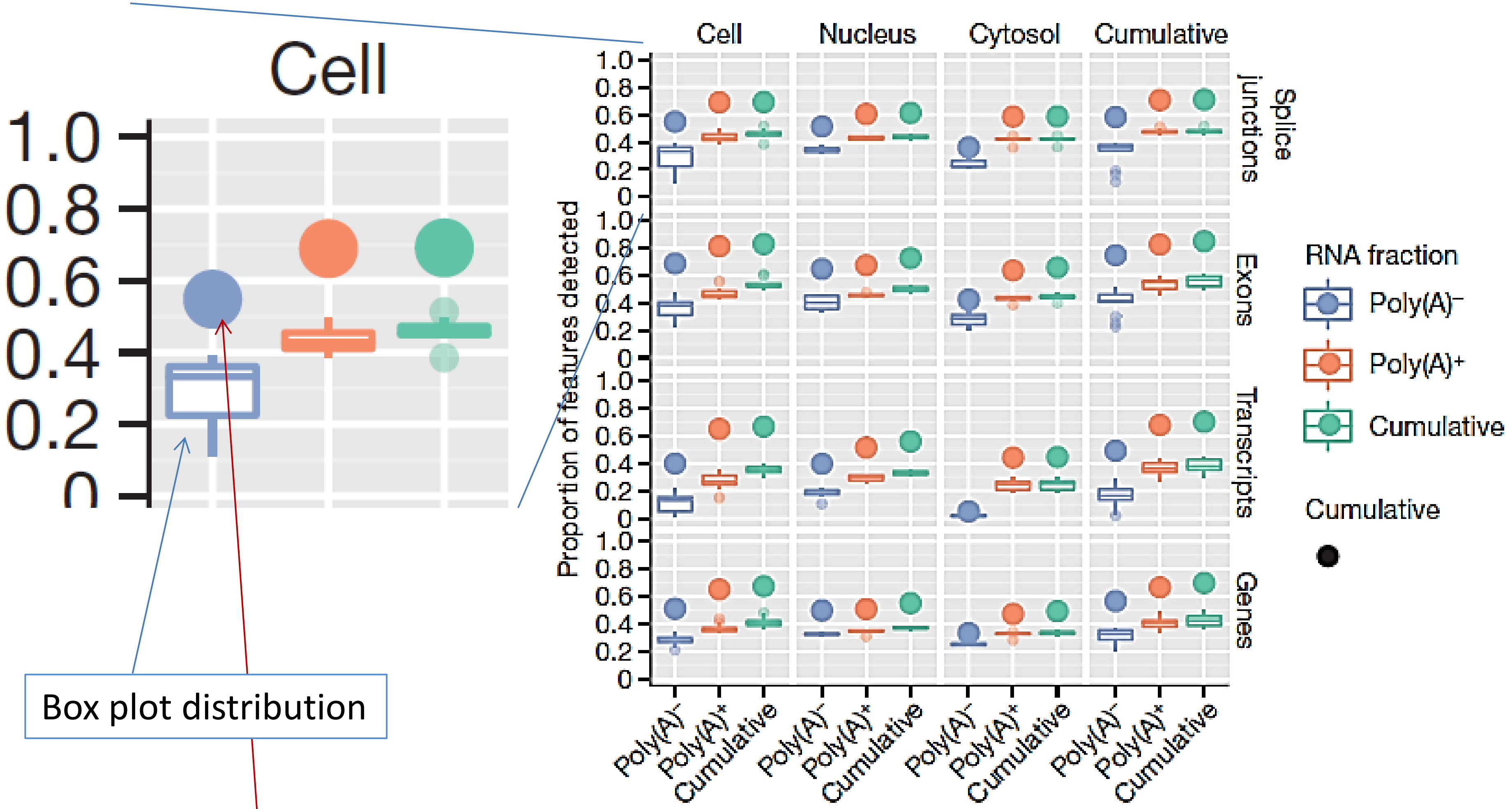
rRNA

«Ribominus™» fraction

Size fractionation
«Long» >200 nt
«Short» <200 nt

RiboMinus Technology
Life Sciences

novel elements covered 78% of the intronic nucleotides and 34% of the intergenic sequences (Supplementary Fig. 4). Overall, the unique contribution of each cell line to the coverage of the genome tends to be small and similar for each cell line (Supplementary Fig. 5). We used the Cufflinks algorithm (see Supplementary Information), and predicted over all long RNA-seq samples 94,800 exons, 69,052 splice junctions, 73,325 transcripts and 41,204 genes in intergenic and antisense regions (Table 1b). These novel elements increase the GENCODE collection of exons, splice sites, transcripts and genes by 19%, 22%, 45% and 80%, respectively.



Box plot distribution

cumulative

Figure 1 | A large majority of GENCODE elements are detected by RNA-seq data. Shown are GENCODE-detected elements in the polyadenylated and non-polyadenylated fractions of cellular compartments (cumulative counts for both RNA fractions and compartments refer to elements present in any of the fractions or compartments). Each box plot is generated from values across all cell lines, thus capturing the dispersion across cell lines. The largest point shows the cumulative value over all cell lines.

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)

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 1 or <1 molecules per cell

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

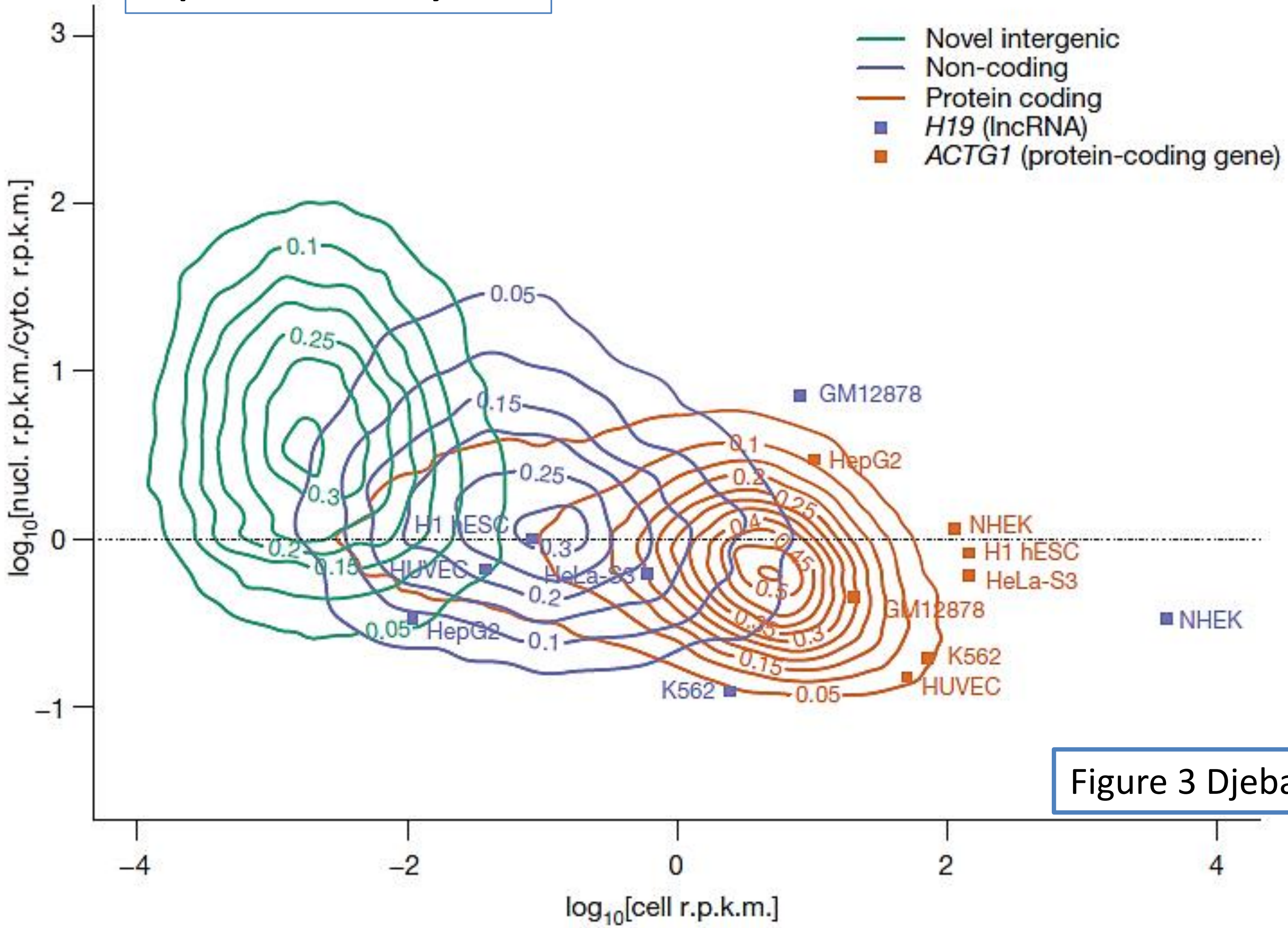
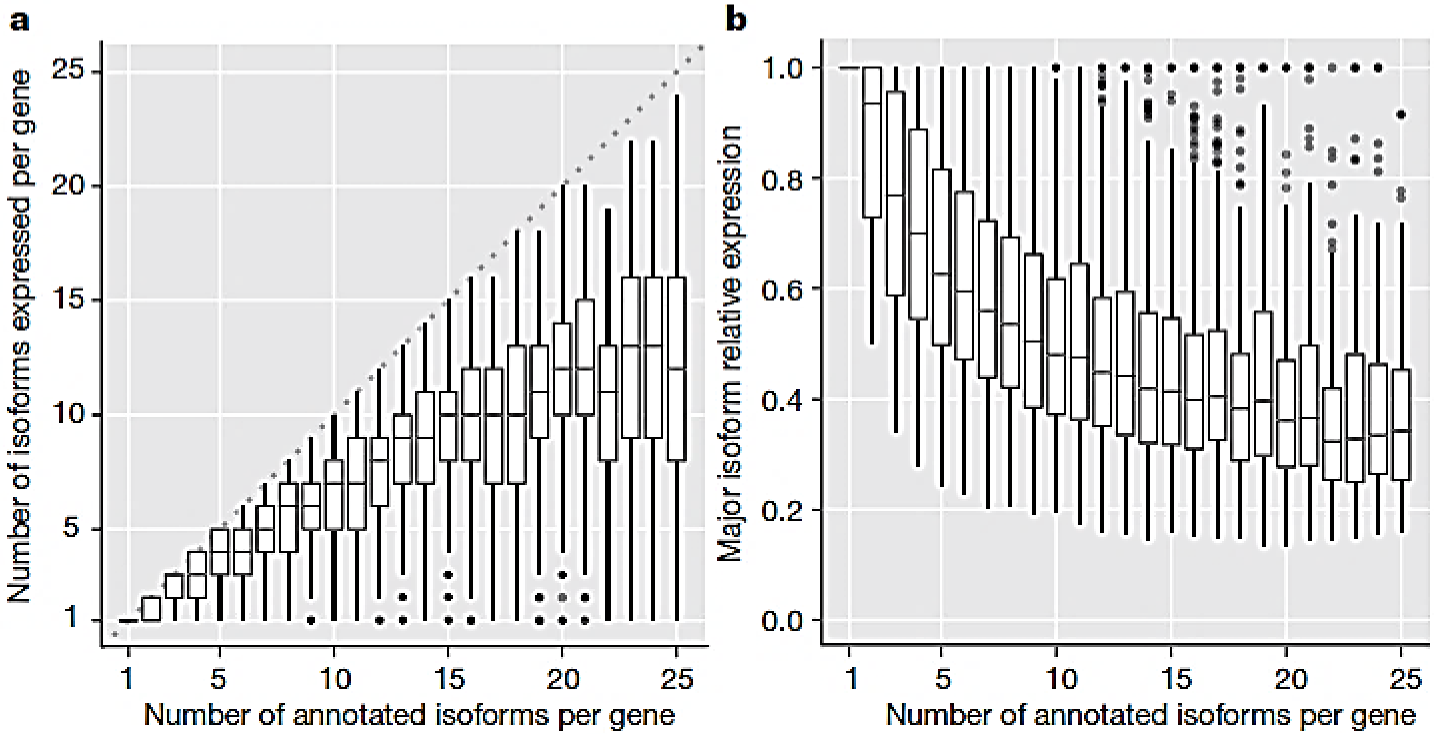


Figure 3 Djebali et al. 2012

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

Djebali 2012, Figure 4 - Isoforms (**alternative splicing**)



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.

Djebali et al., 2012

Alternative transcription initiation and termination.

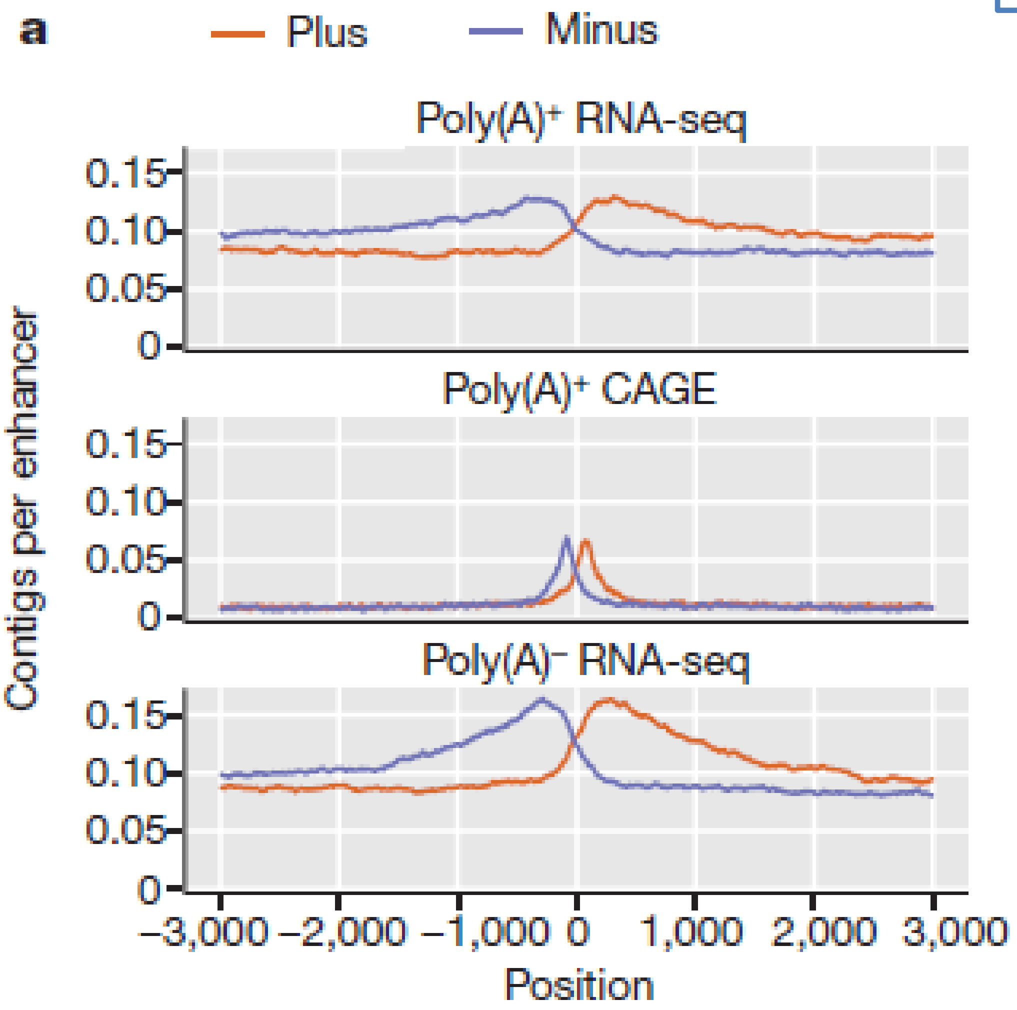
a total of 128,021 TSSs were detected across all cell lines (97,778 previously annotated ; 30,243 were novel intergenic/antisense TSSs).

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

eRNA

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



Expression level by class

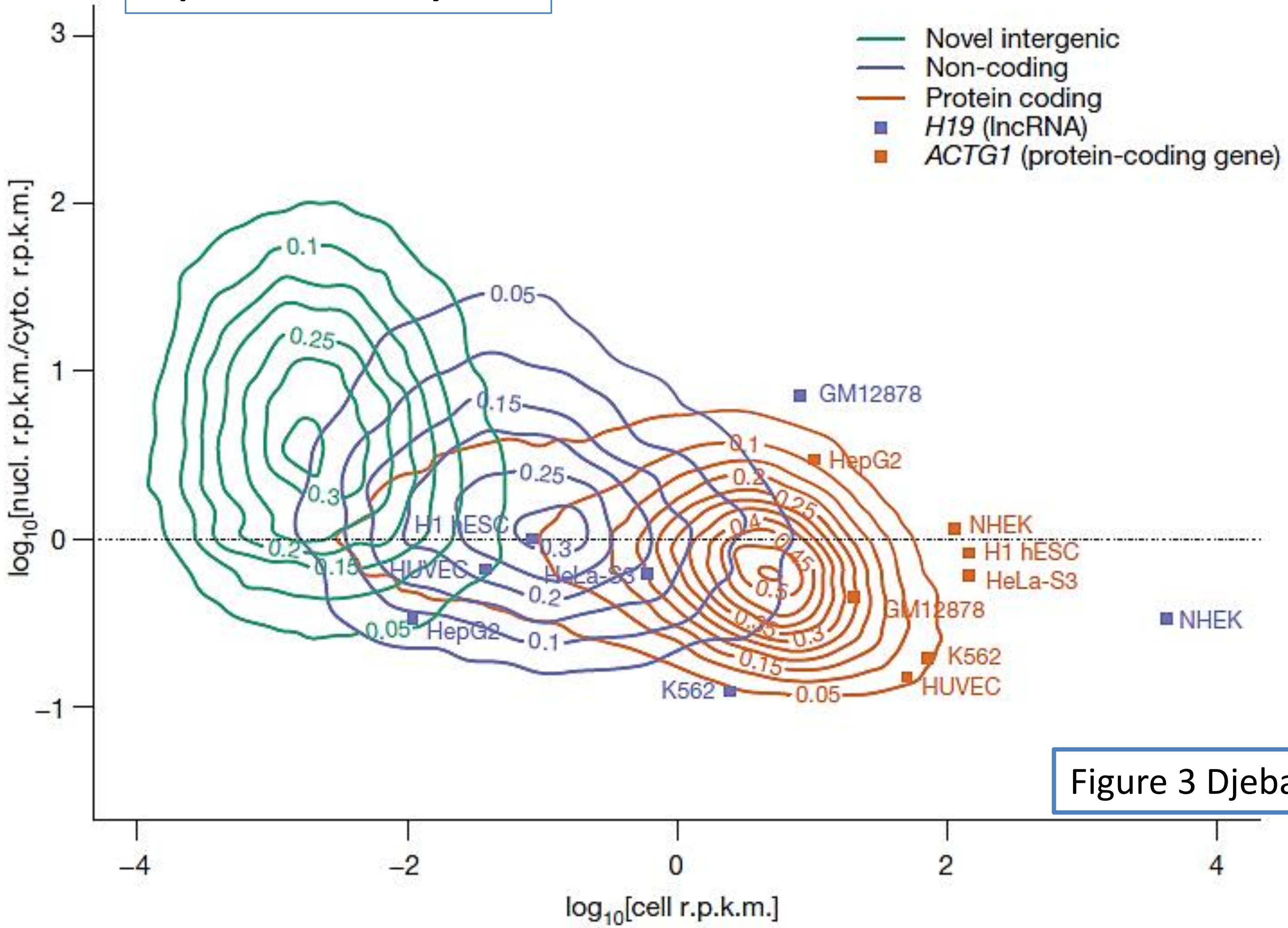


Figure 3 Djebali et al. 2012

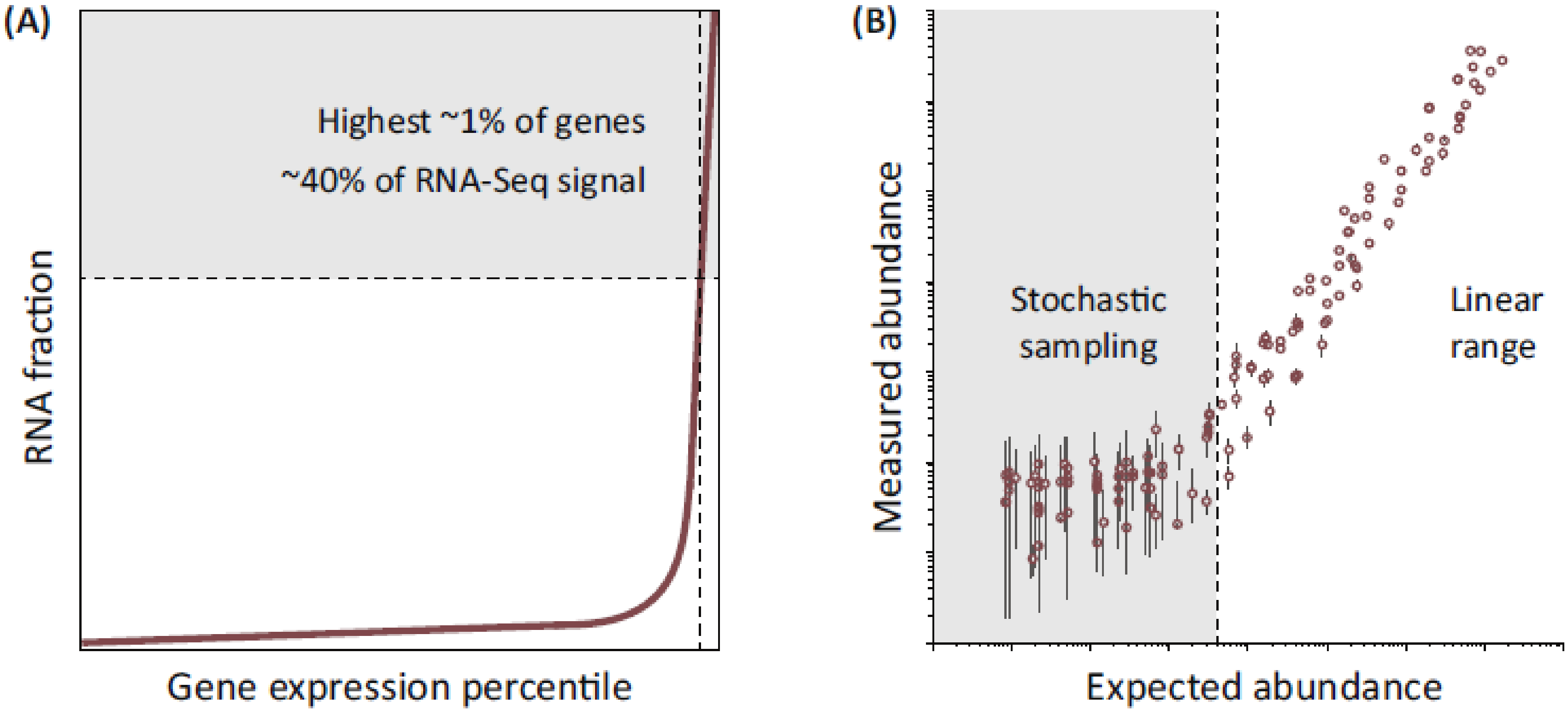
Protein coding transcripts are the only class that is enriched in the cytoplasm

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)
Stochastic sampling
Linear range
Measured abundance
Expected abundance

Expression of lncRNAs is highly tissue-specific

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

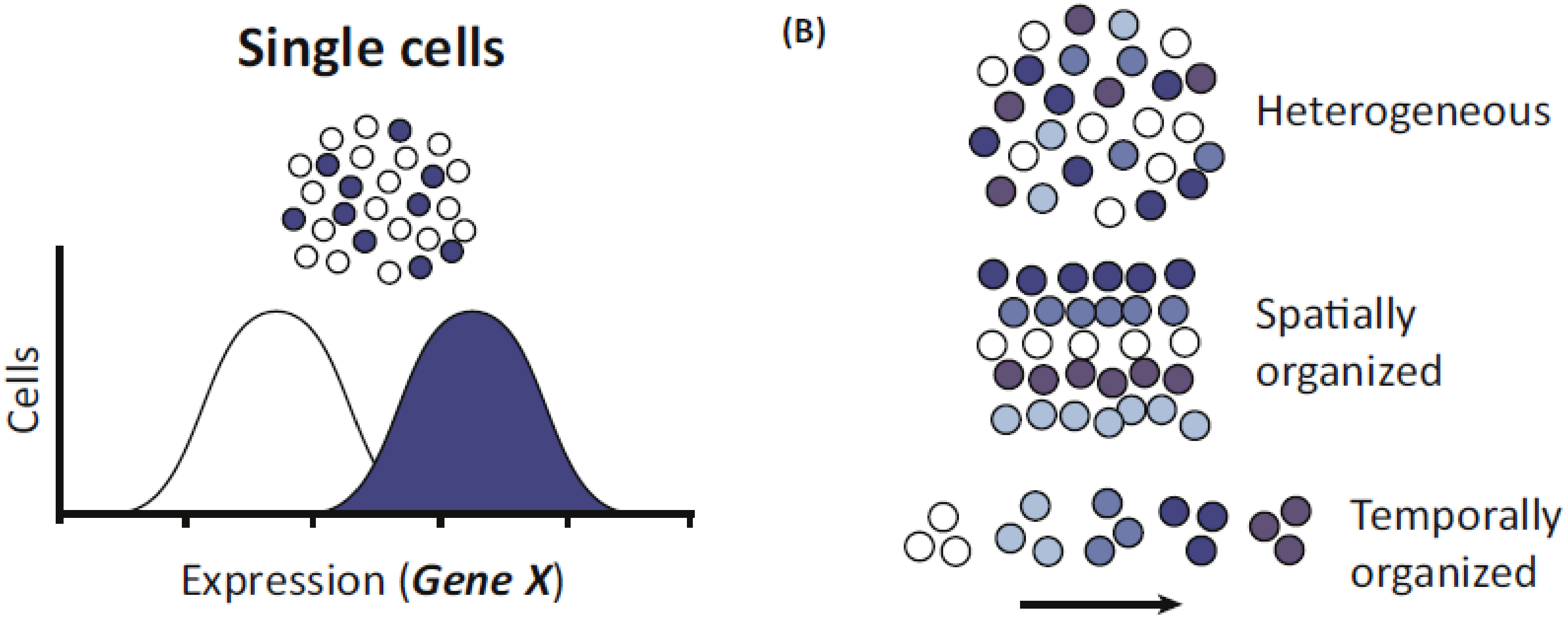
By FISH analysis: expression highly limited to cell types (in brain)

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

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

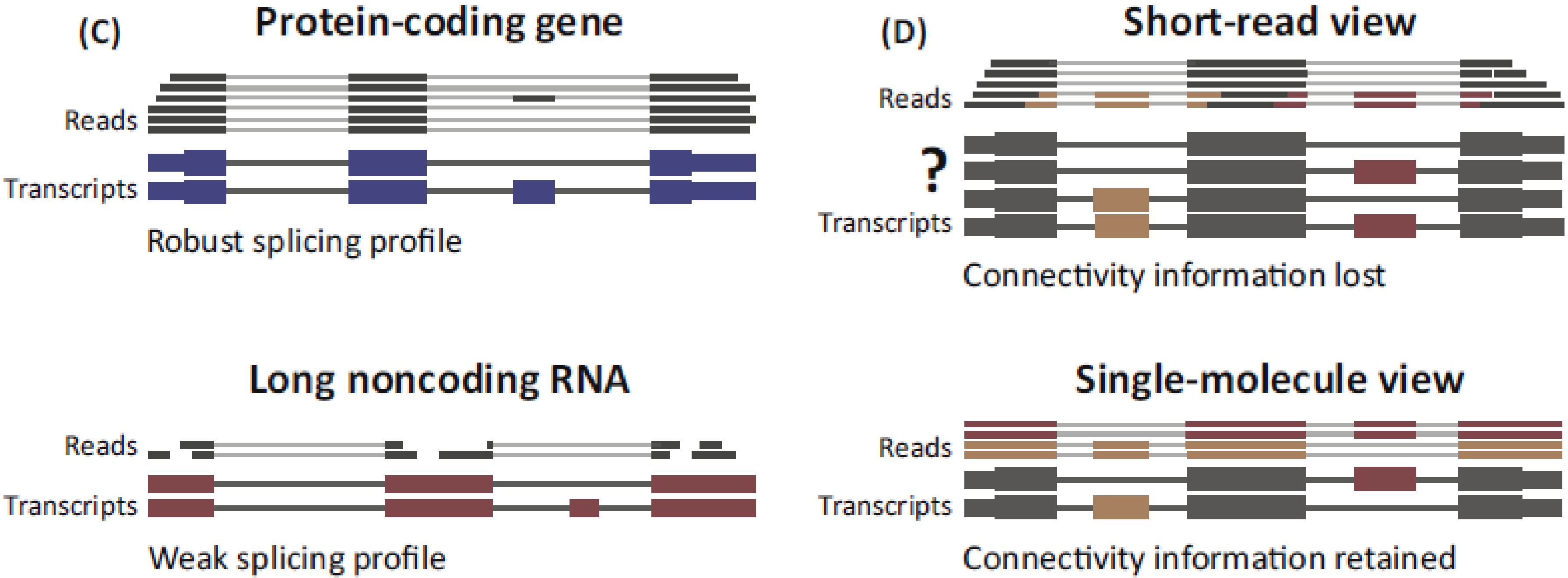
Cell-subtype determinants?

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 epigenomic regulation

(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)