TOPIC 3. RNA-seq

Applications of RNA-seq: gene expression quantification and splicing variant annotation.

Omics Techniques

Bachelor’s Degree in Bioinformatics

Sònia Casillas, UAB

The transcriptome

What is the transcriptome?

All the transcripts of a cell, and their quantities, in a specific stage of development and a given physiological condition

The aims of transcriptomics

- To catalog all types of transcripts, including mRNAs, ncRNAs and sRNAs
- To determine the transcriptional structure of the genes, including the transcription start sites, 5’ and 3’ UTRs, the splicing patterns and other post-transcriptional modifications
- To quantify changes in the levels of expression of each transcript during development and under different physiological conditions

Why study of RNA is so important?

RNA profiling provides clues about:

- Genes and other expressed sequences of a genome
- Gene regulation and regulatory sequences
- Function of the genes and their interaction
- Functional differences between tissues and cell types
- Identification of candidate genes for any given process or disease

Experimental methods to analyze the transcriptome

ONE SINGLE GENE

- Northern Blot
- RT-PCR
- 5’ i 3’ RACE
- Quantitative RT-PCR (Real-Time PCR)

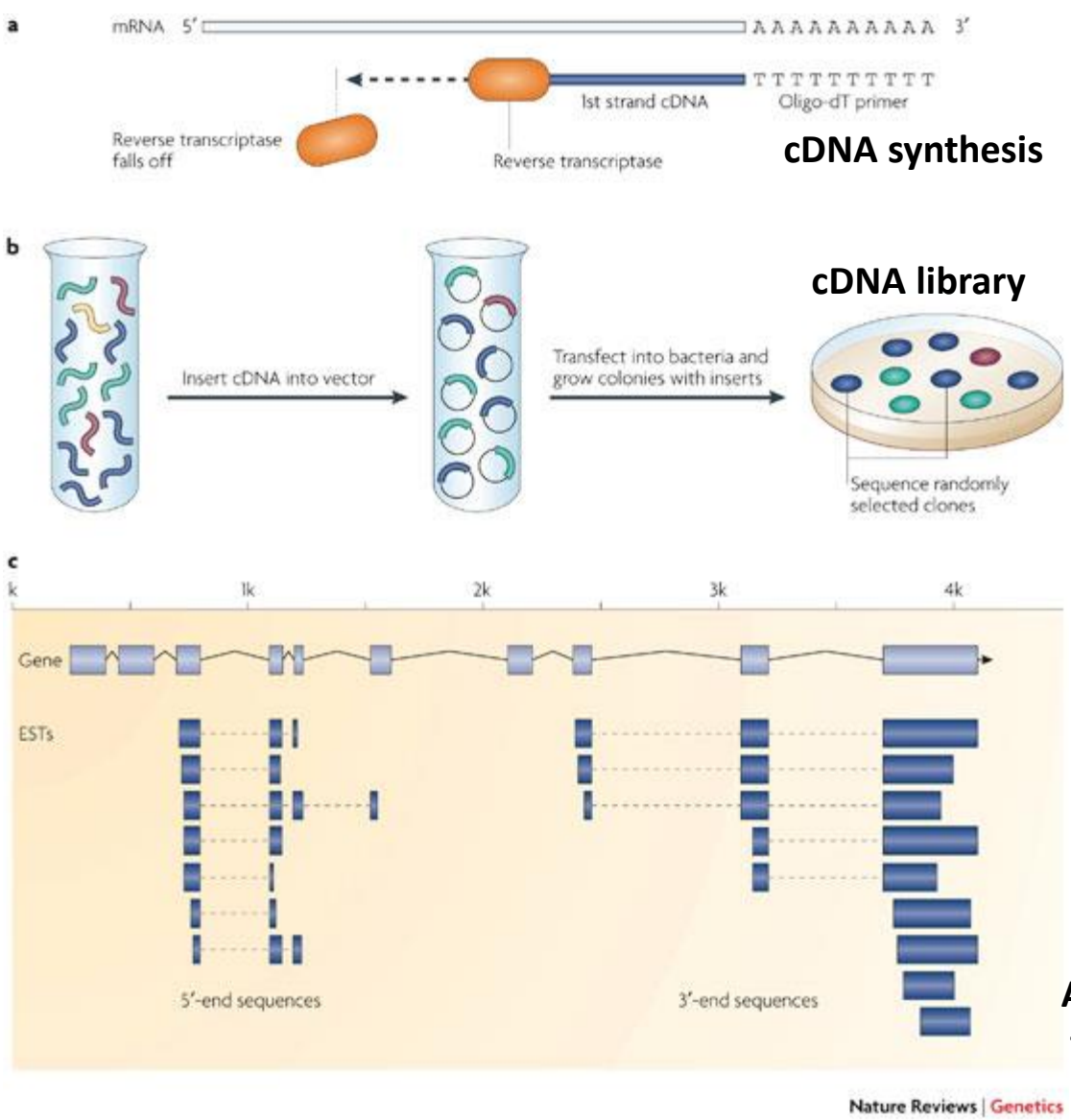
WHOLE TRANSCRIPTOME

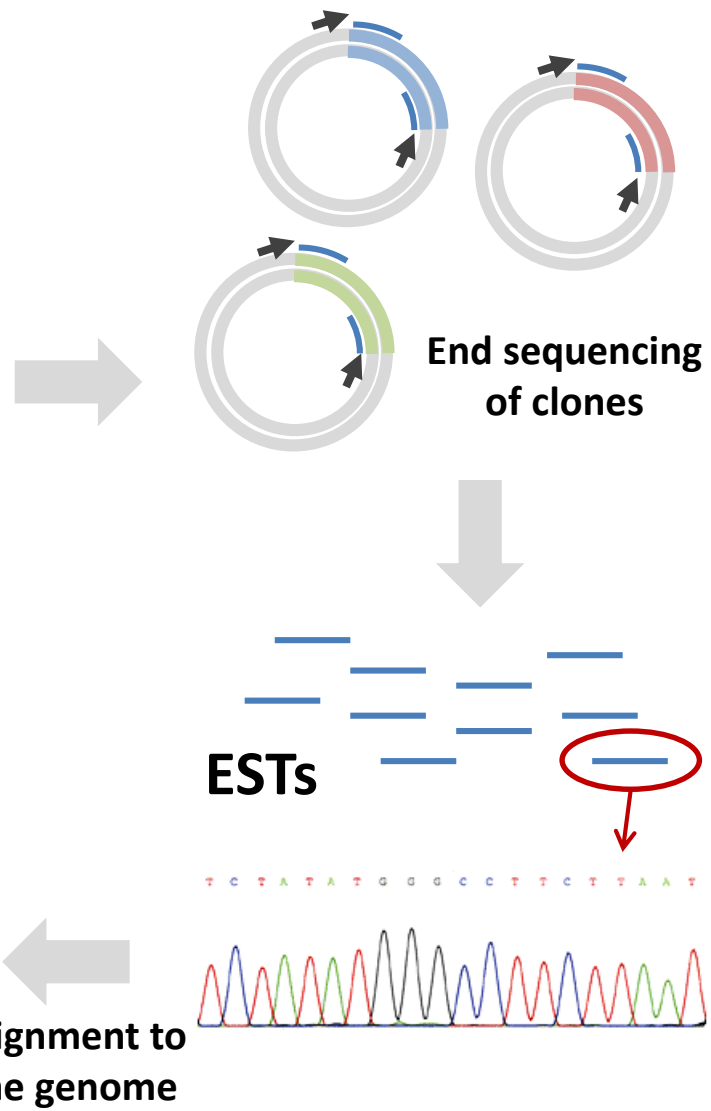
- EST sequencing
- Microarrays
- RNA-Seq

Experimental methods to analyze the transcriptome

- ESTs
  - Sequencing cDNA libraries
- Microarrays
  - Multiple DNA probes fixed on a glass slide
  - Hybridization of fluorescently-labelled RNA
- RNA-seq
  - Massive sequencing of cDNA libraries

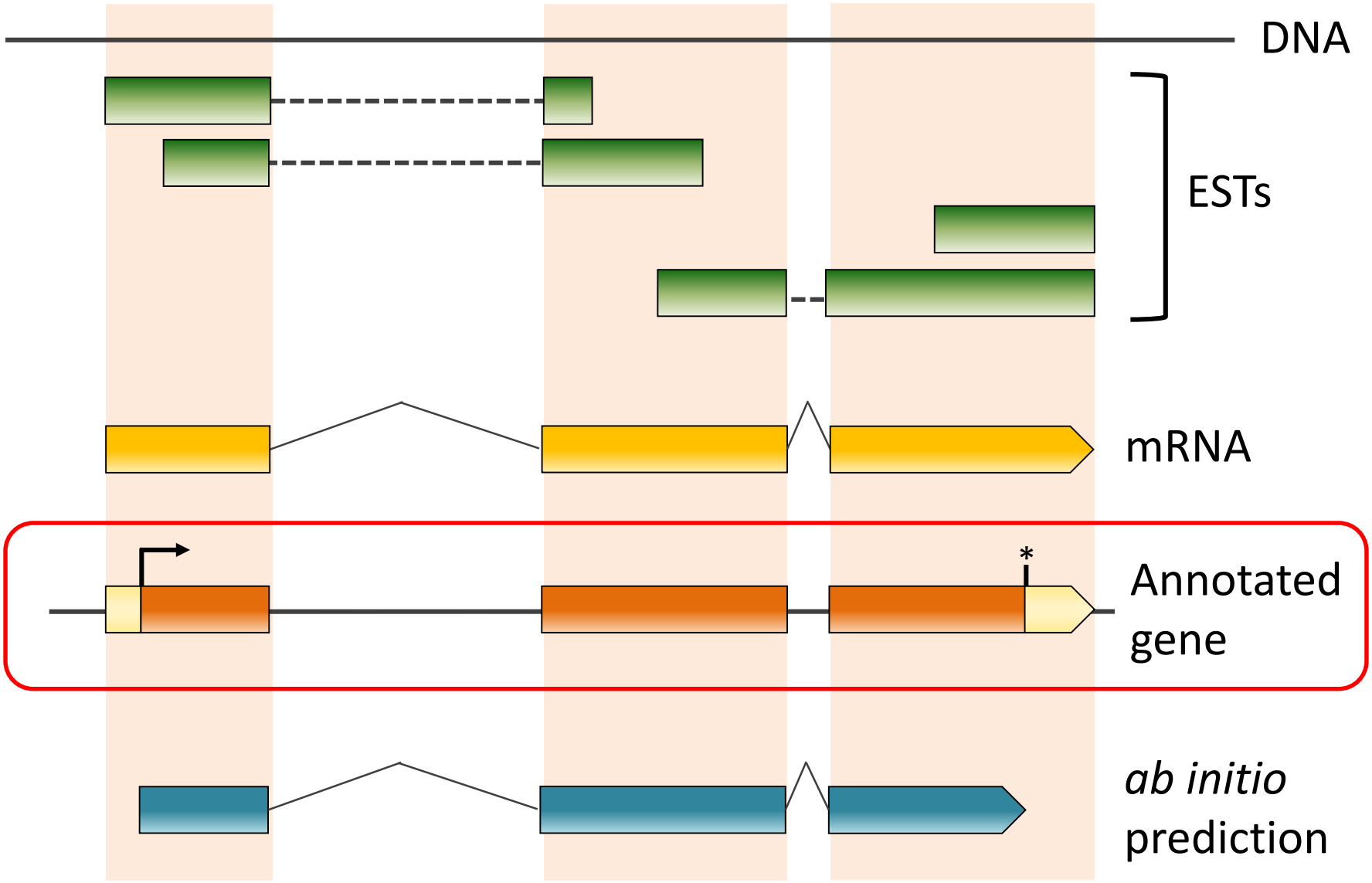
Expressed Sequence Tags (ESTs)





Brent (2008) *Nature Reviews Genetics* 9: 62-73

Annotation of genes using ESTs



Limitation of ESTs

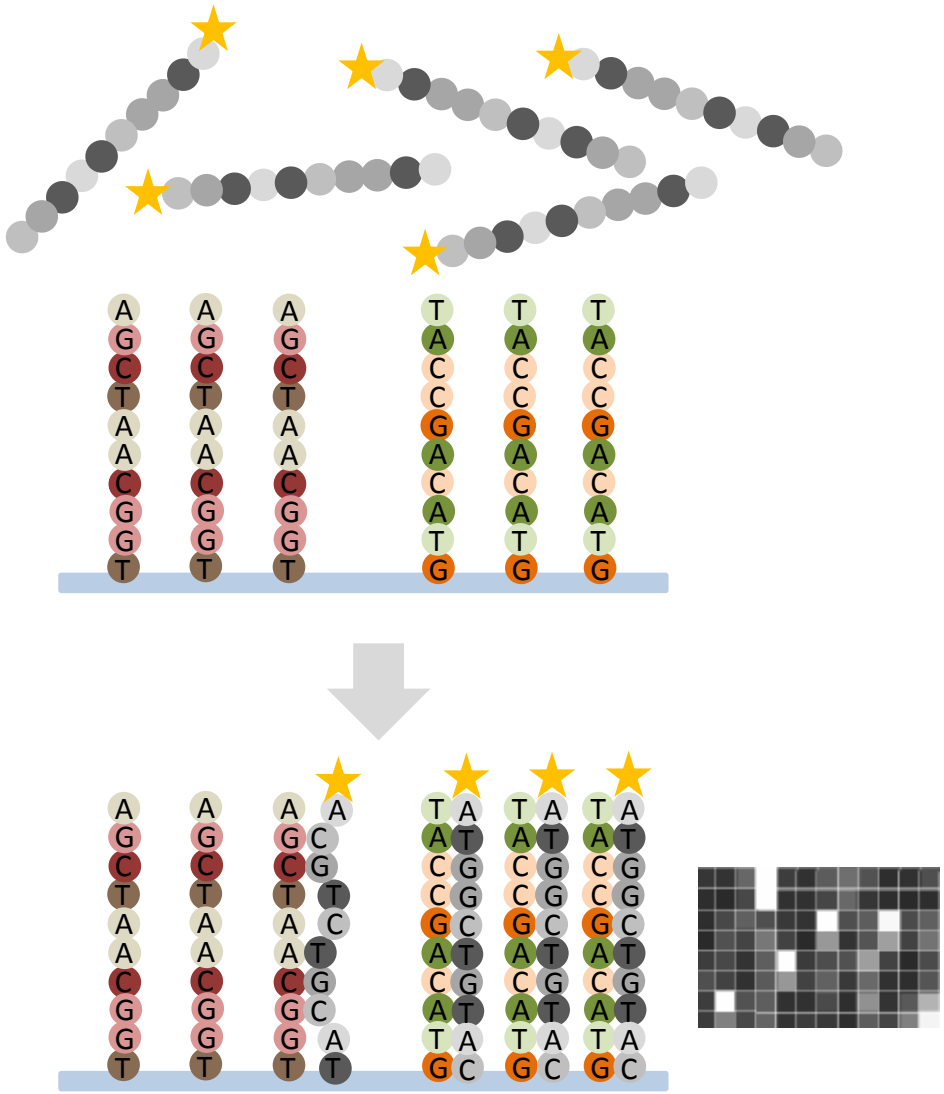
- Low-throughput
- Elevated cost
- Quantification is not accurate
- Different isoforms are generally indistinguishable

Experimental methods to analyze the transcriptome

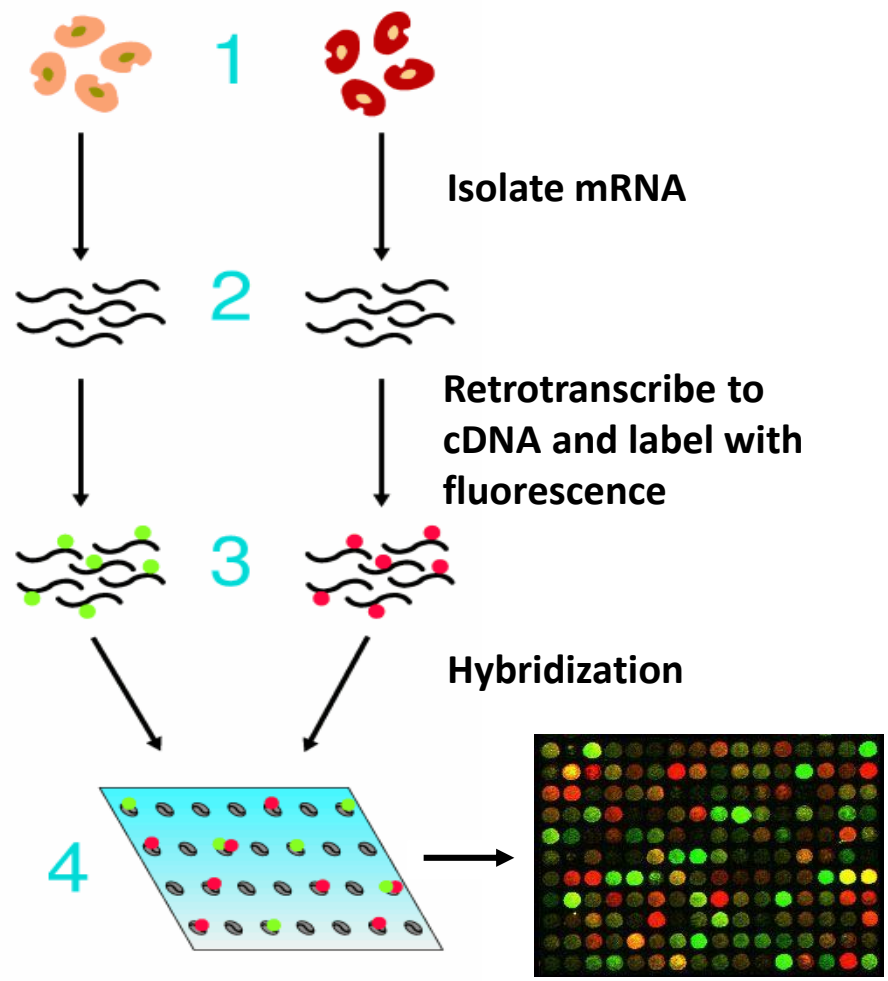
- ESTs
  - Sequencing cDNA libraries
- Microarrays
  - Multiple DNA probes fixed on a glass slide
  - Hybridization of fluorescently-labelled RNA
- RNA-seq
  - Massive sequencing of cDNA libraries

How do microarrays work?

Hybridization of **one sample**



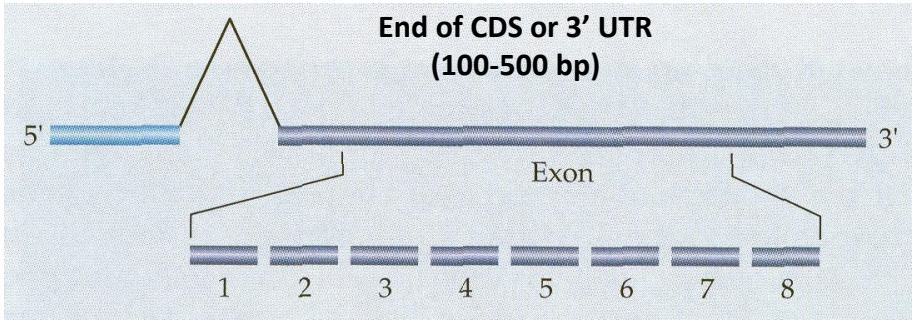
Competitive hybridization of **two samples**

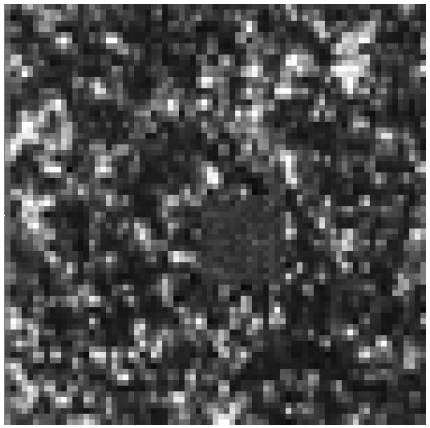


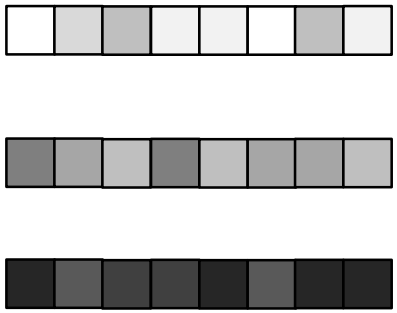
Gene expression arrays

- Gene expression arrays
  - Short oligonucleotides of 25 nucleotides
  - Multiple probes at the 3’-end exons of the gene
  - They allow quantifying the abundance of transcripts

Affymetrix GeneChip microarray







High level of expression

Medium level of expression

No expression

Genome tiling arrays

- Genome tiling arrays
  - Long oligonucleotides of 60 nucleotides
  - Overlapping probes that cover the region of interest
  - They allow identifying novel transcribed sequences

Characterization of a novel testis transcript using tiling arrays

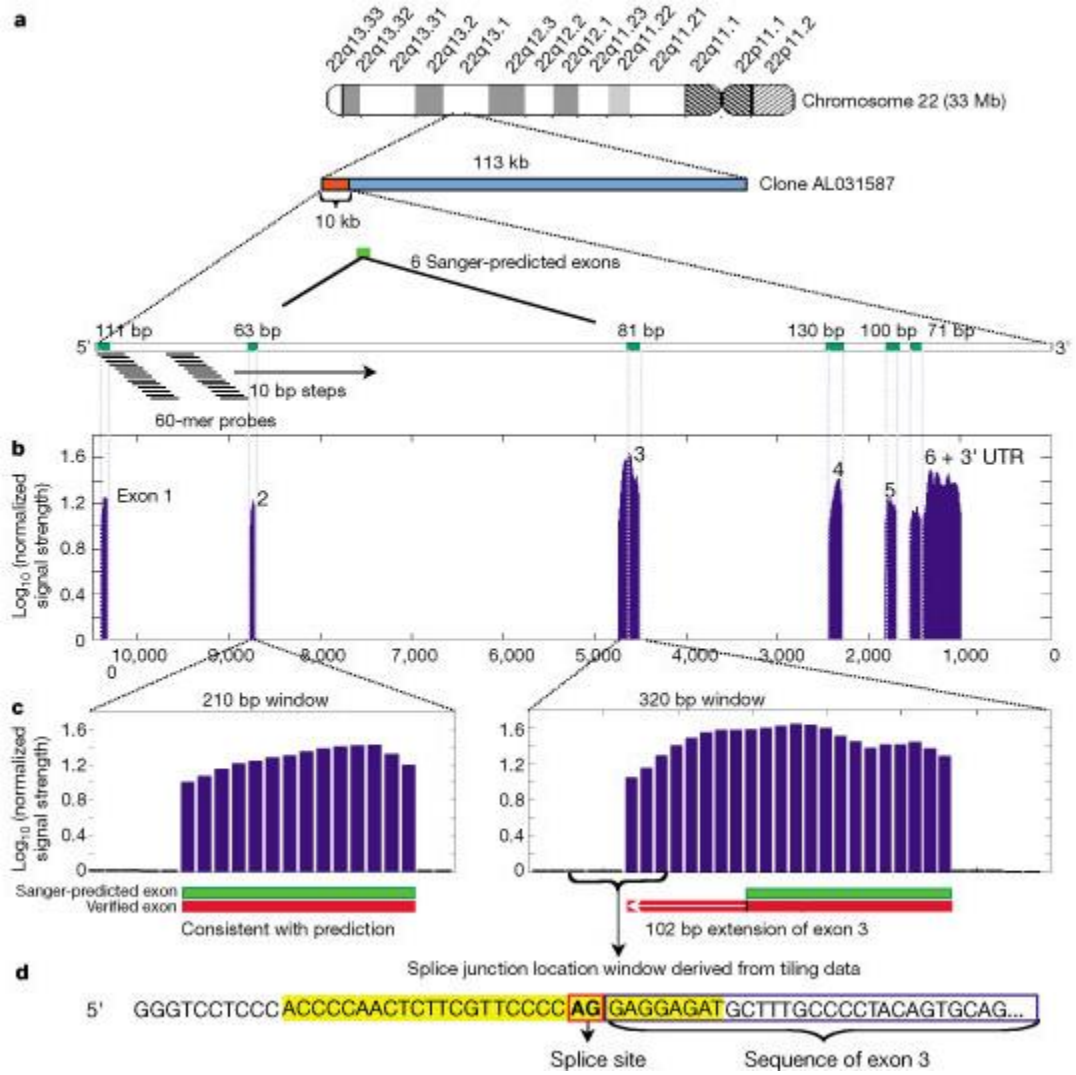


Figure 3. Shoemaker *et al.* (2001) *Nature* **409**, 922-927

Limitations of microarrays

- Lack of reproducibility & standardization
  - Cross-hybridization
  - Low affinity of probes
  - Batch effects
- Small dynamic range & saturation (~1000-fold range)
- Relative measures (always probe-dependent)
- Little information on actual gene
- Inability to detect unknown sequences (transcripts or SNPs)

In the case of ***genome tiling arrays***, also:

- Elevated cost

Experimental methods to analyze the transcriptome

- ESTs
  - Sequencing cDNA libraries
- Microarrays
  - Multiple DNA probes fixed on a glass slide
  - Hybridization of fluorescently-labelled RNA
- RNA-seq
  - Massive sequencing of cDNA libraries

RNA-Seq

RNA-Seq is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies. Studies using this method have already altered our view of the extent and complexity of eukaryotic transcriptomes. RNA-Seq also provides a far more precise measurement of levels of transcripts and their isoforms than other methods.

“RNA-Seq […] is expected to revolutionize the manner in which eukaryotic transcriptomes are analysed.”

Wang *et al*. (2009) *Nature Reviews Genetics* 10: 57-63

RNA-seq

Whole RNA sequencing using NGS technologies:

- Identification of transcribed sequences
- Quantification of transcript abundance
- Multiple reads along each RNA
- Analysis of alternative transcript isoforms

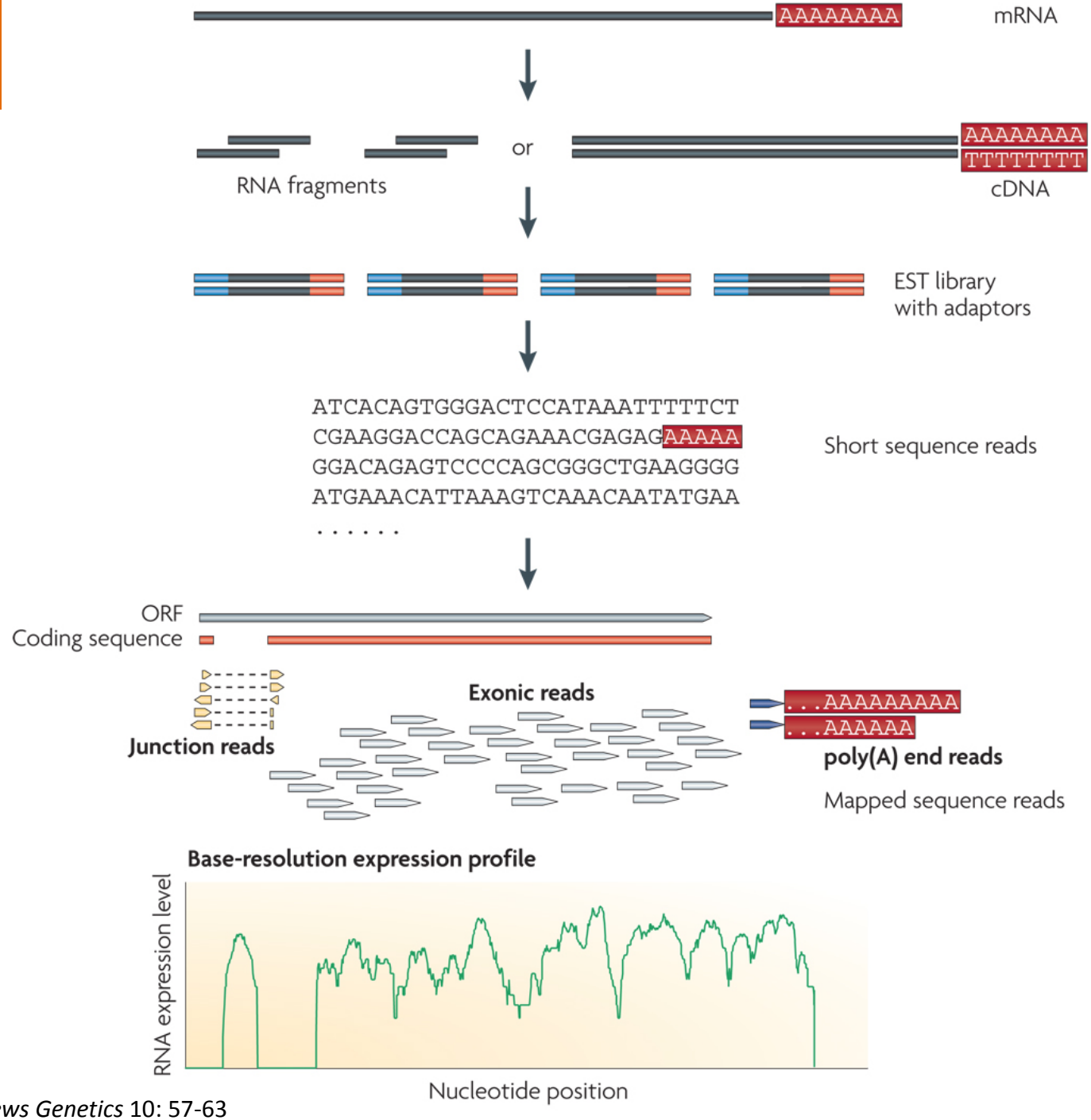
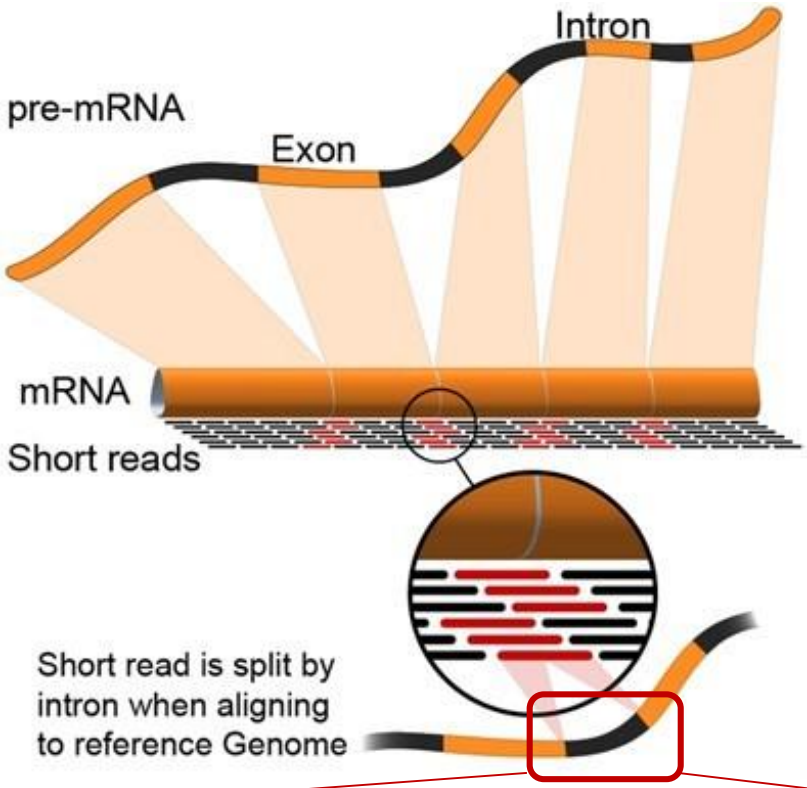


Figure 1. Wang *et al*. (2009) *Nature Reviews Genetics* 10: 57-63

RNA-seq mapping of short reads in exon-exon junctions



Junction reads
= Split reads

CCGAAAATCAAGTCATCCCTAAAGACTAAGTAAGTAACCATATTACATTAAGGAAGGCACTTTAAAAGTTTATAATCATTTGTAGACTCCCACCAAAGCCACTGACTCGCAAGG

Exon **Intron** **Exon**

Advantages of RNA-seq

- It is not necessary to know the genomic sequence of transcribed regions in advance
- It allows detecting novel transcripts
- High precision in the detection of the limits of the transcripts
- It allows detecting all *splicing* variants and alternative start and end sites of distinct transcripts
- It allows detecting SNPs in transcribed regions
- It allows detecting the specific transcription profiles of each allele
- Quantification of the levels of expression of each transcript is accurate (long dynamic range)
- Very reproducible
- Requires small amount of initial RNA

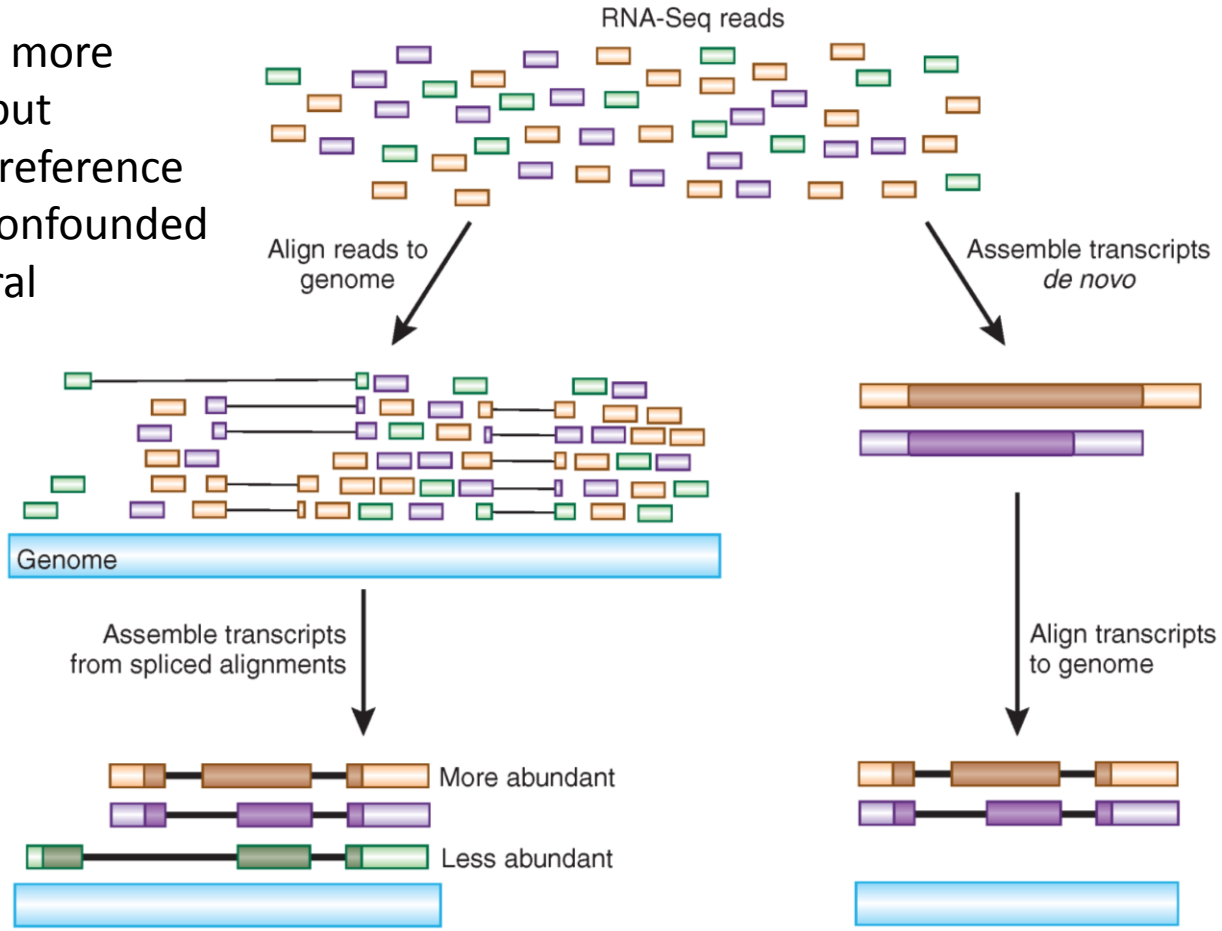
Transcript reconstruction by RNA-seq

Align-then-assemble

Potentially more sensitive, but requires a reference genome, confounded by structural variation

de-novo assembly

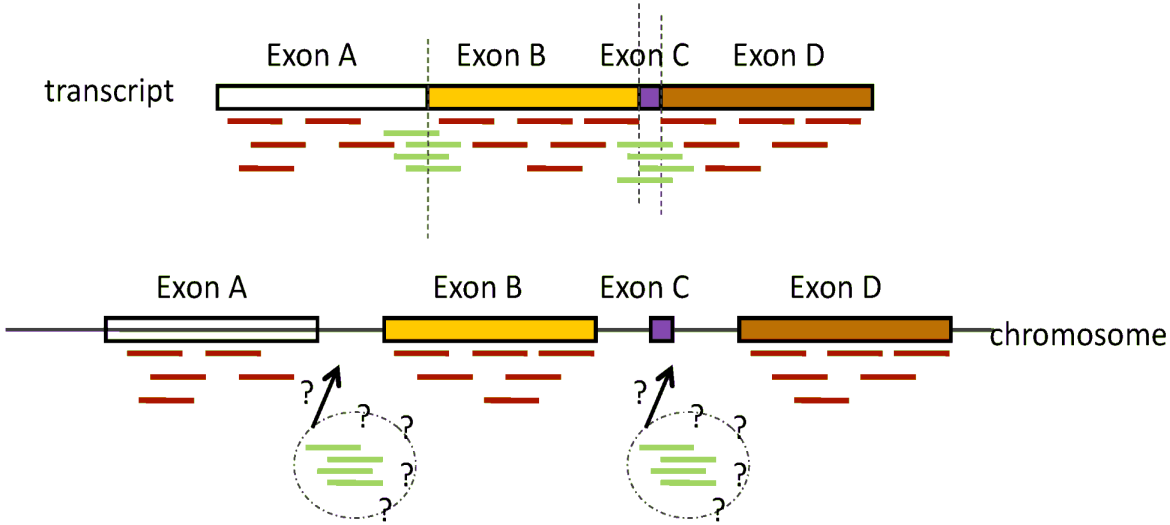
Likely to only capture highly expressed transcripts, but does not require a reference genome, robust to variation



Unspliced vs. spliced mapping

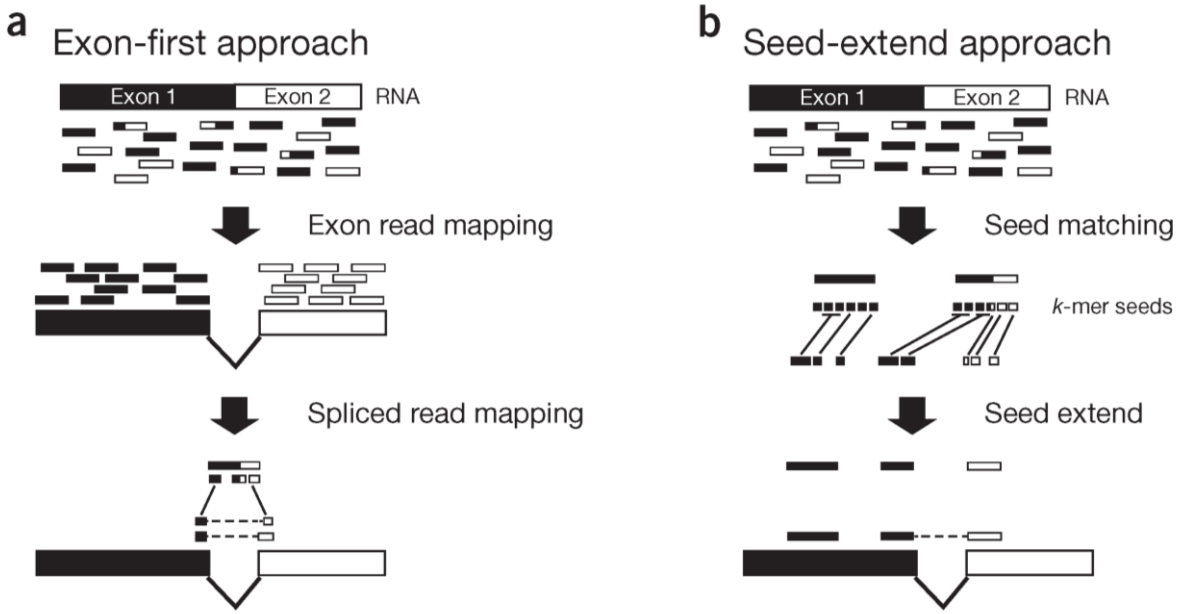
Unspliced mapping

e.g. BWA, Bowtie



Spliced mapping

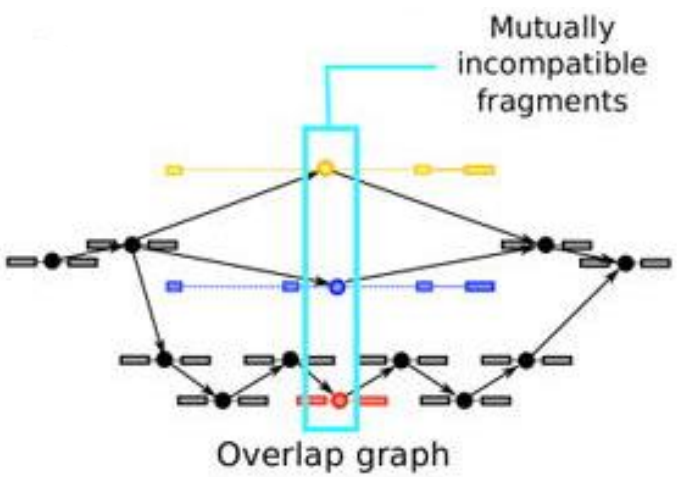
e.g. Tophat

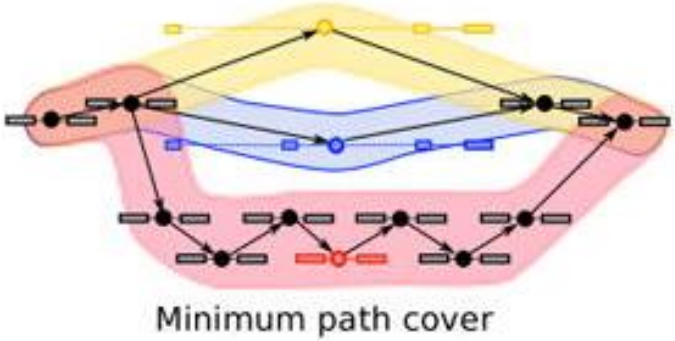


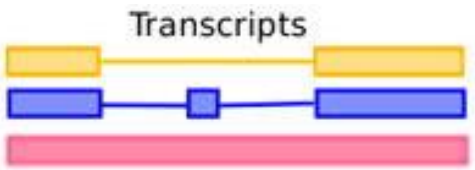
Assembly of transcripts from spliced alignments

Assembly of transcripts

e.g. Cufflinks





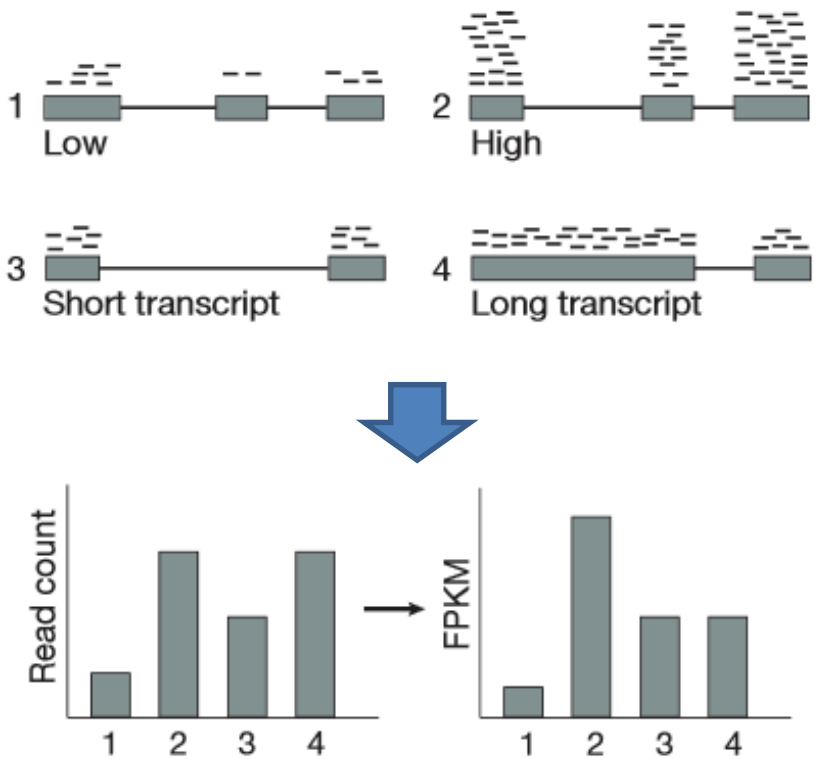


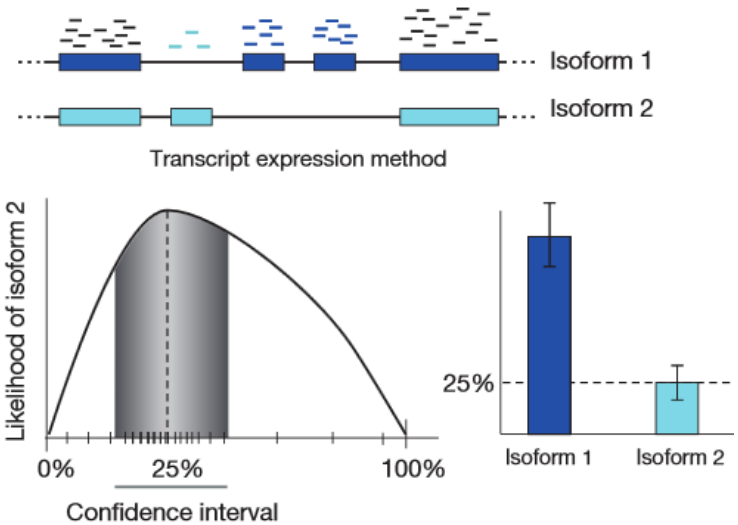
What genes are expressed in the sample?

What transcripts are expressed for each gene?

Quantification of transcripts' expression levels

What are the genes' and transcripts' expression levels?

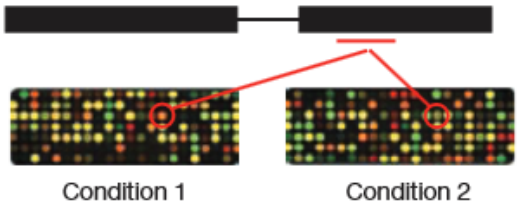


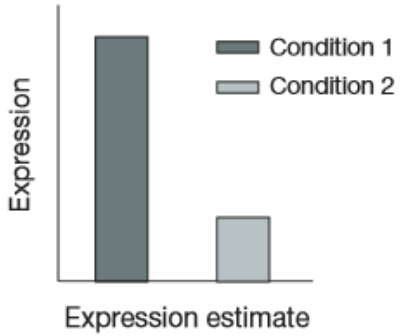


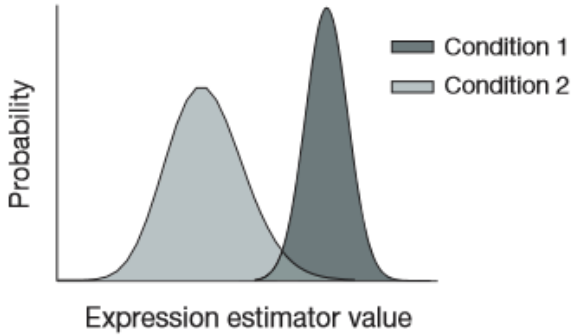
Gene expression **QUANTIFICATION**, typically reported in **RPKM/FPKM**: # of reads (fragments) per kb of <u>exonic bases</u> per million reads in the <u>library</u>

Testing for differentially-expressed genes

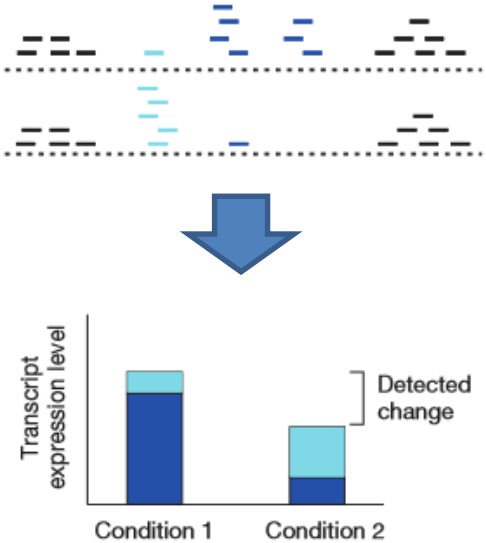
How do expression levels and/or splicing patterns differ between two conditions?







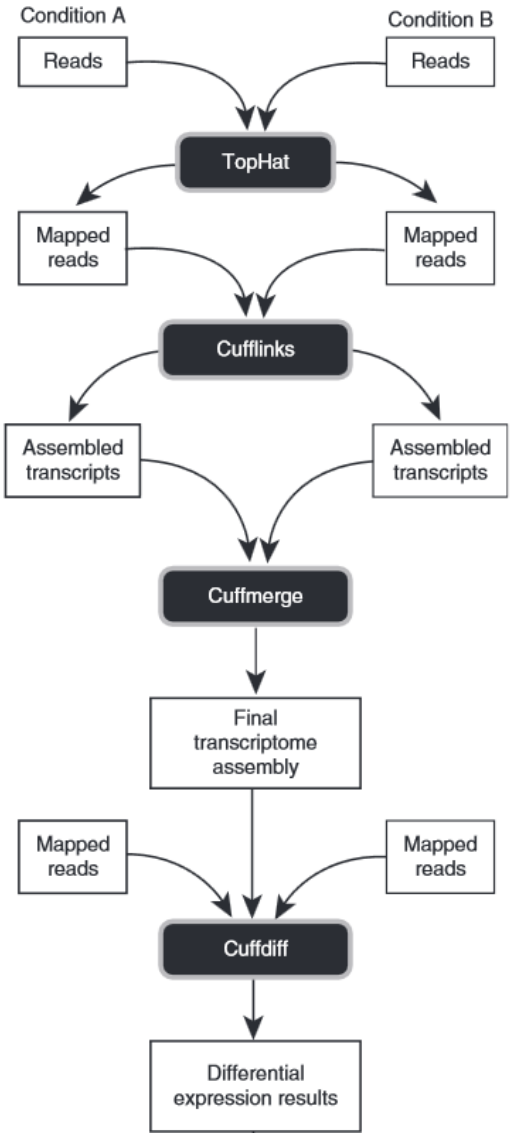
Gene **DIFFERENTIAL** expression for individual isoforms between conditions



A typical workflow for RNA-seq

Tuxedo Workflow (steps 1-4)

1. Map reads to a reference genome: **Tophat**
2. Assemble reads into transcripts: **Cufflinks**
3. Reconcile transcripts across multiple samples: **Cuffmerge**
4. Quantify isoform expression and compare among samples: **Cuffdiff**
5. Visualization: **Integrated Genomes Viewer (IGV)**



Summary

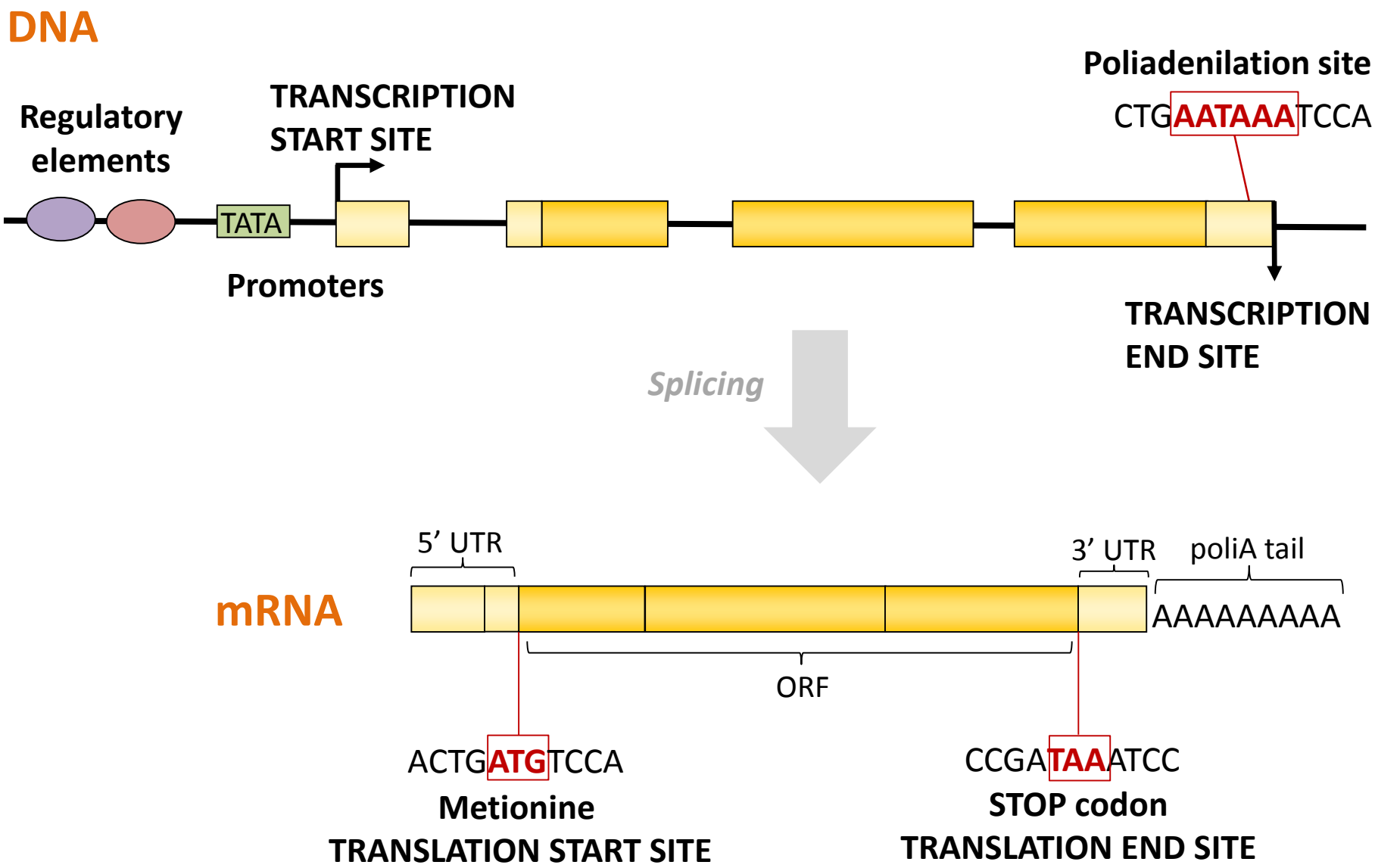
Table 1 | **Advantages of RNA-Seq compared with other transcriptomics methods**

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

Table 1. Wang *et al*. (2009) *Nature Reviews Genetics* 10: 57-63

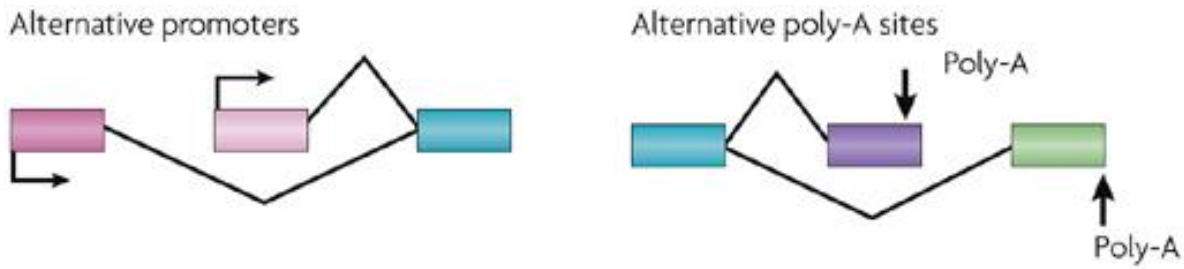
The transcriptional landscape of a genome

Typical structure of a protein-coding gene



Alternative splicing

Initial/final exons



Internal exons

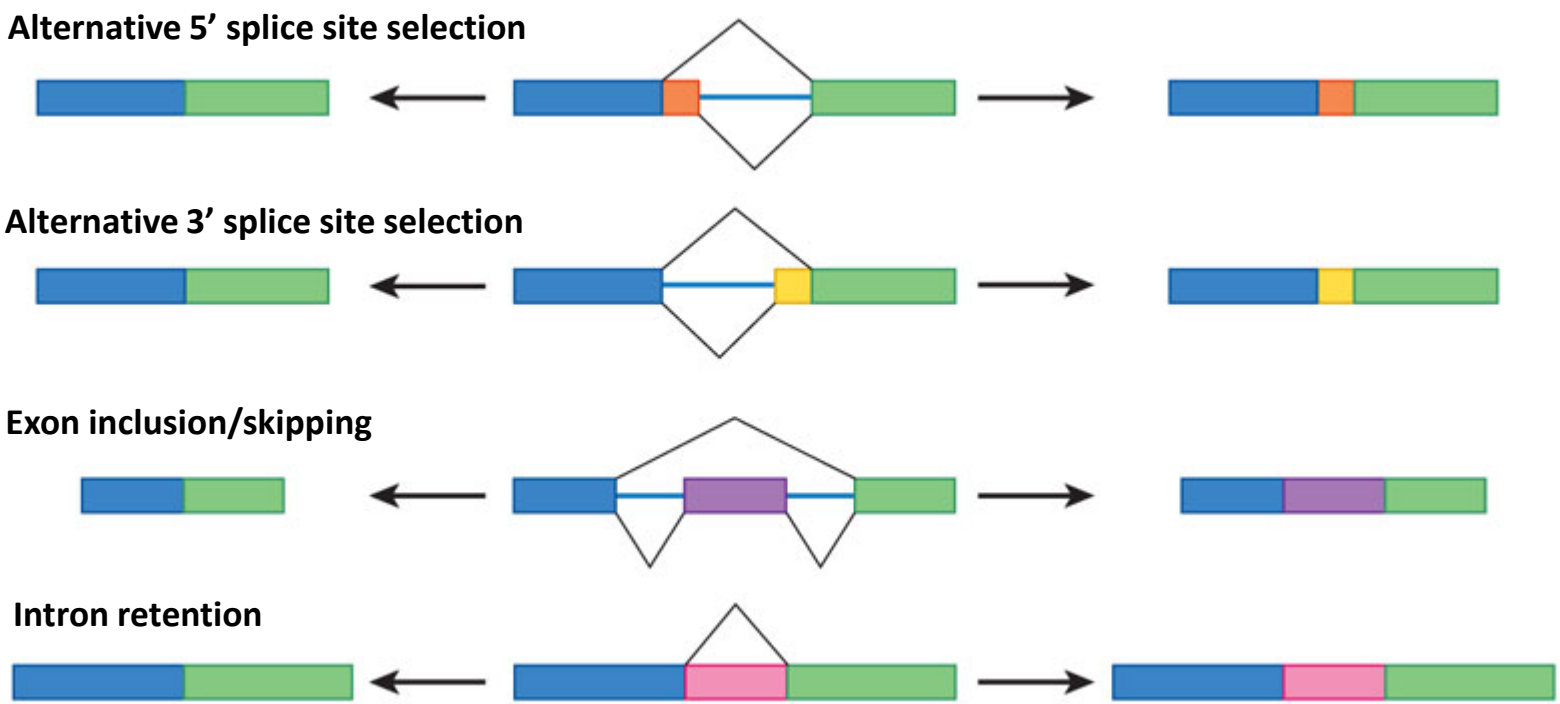


Figure 1. Nielsen and Graveley (2010) *Nature* 463: 457-463

Figure 1. Li *et al*. (2007) *Nature Reviews Neuroscience* 8: 819-831.

Example of alternative splicing: α-tropomiosine

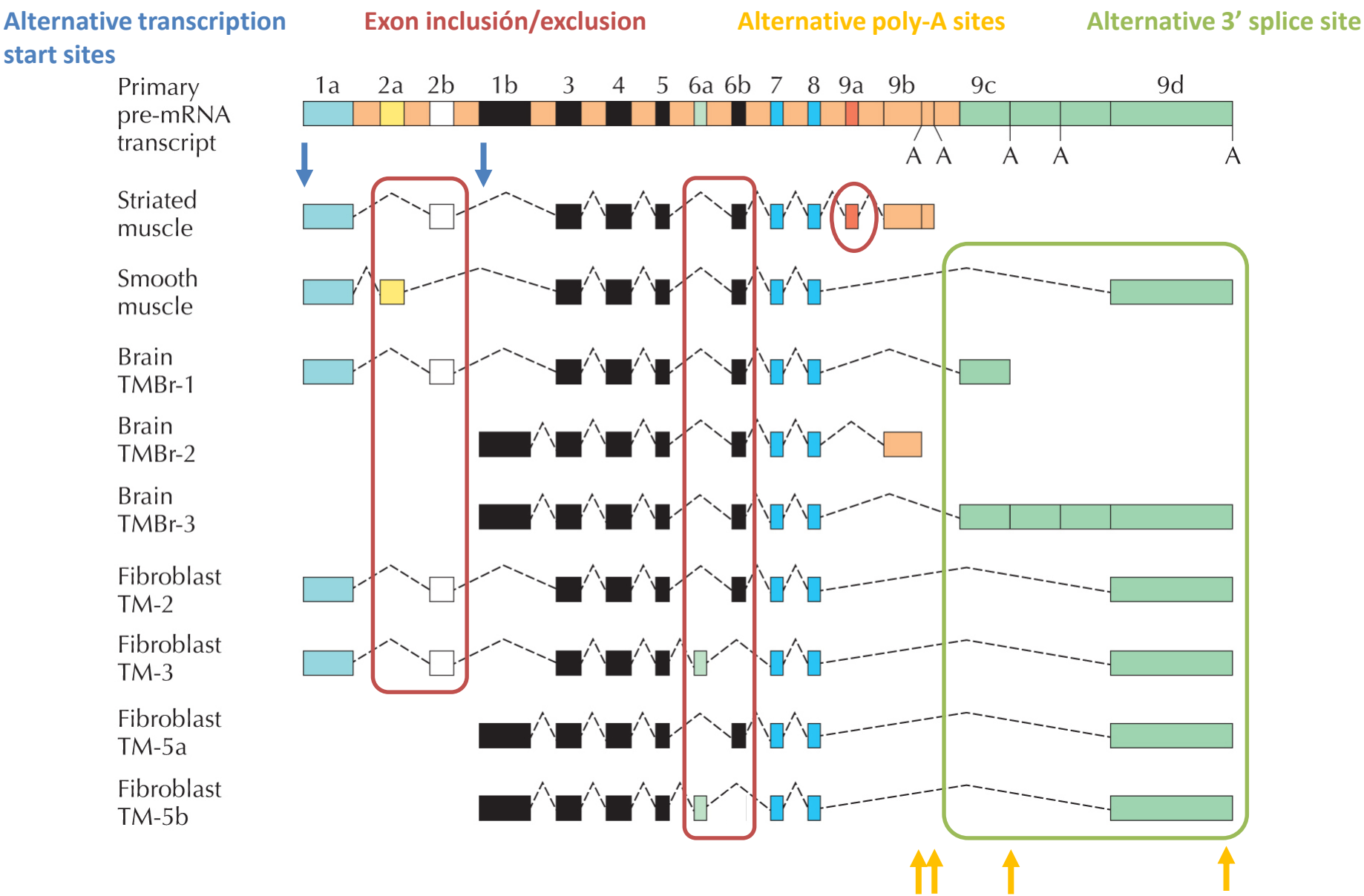


Figure 8.22. *Evolution.* Barton *et al*. (2007) Cold Spring Harbor Laboratory Press

Other examples of alternative splicing

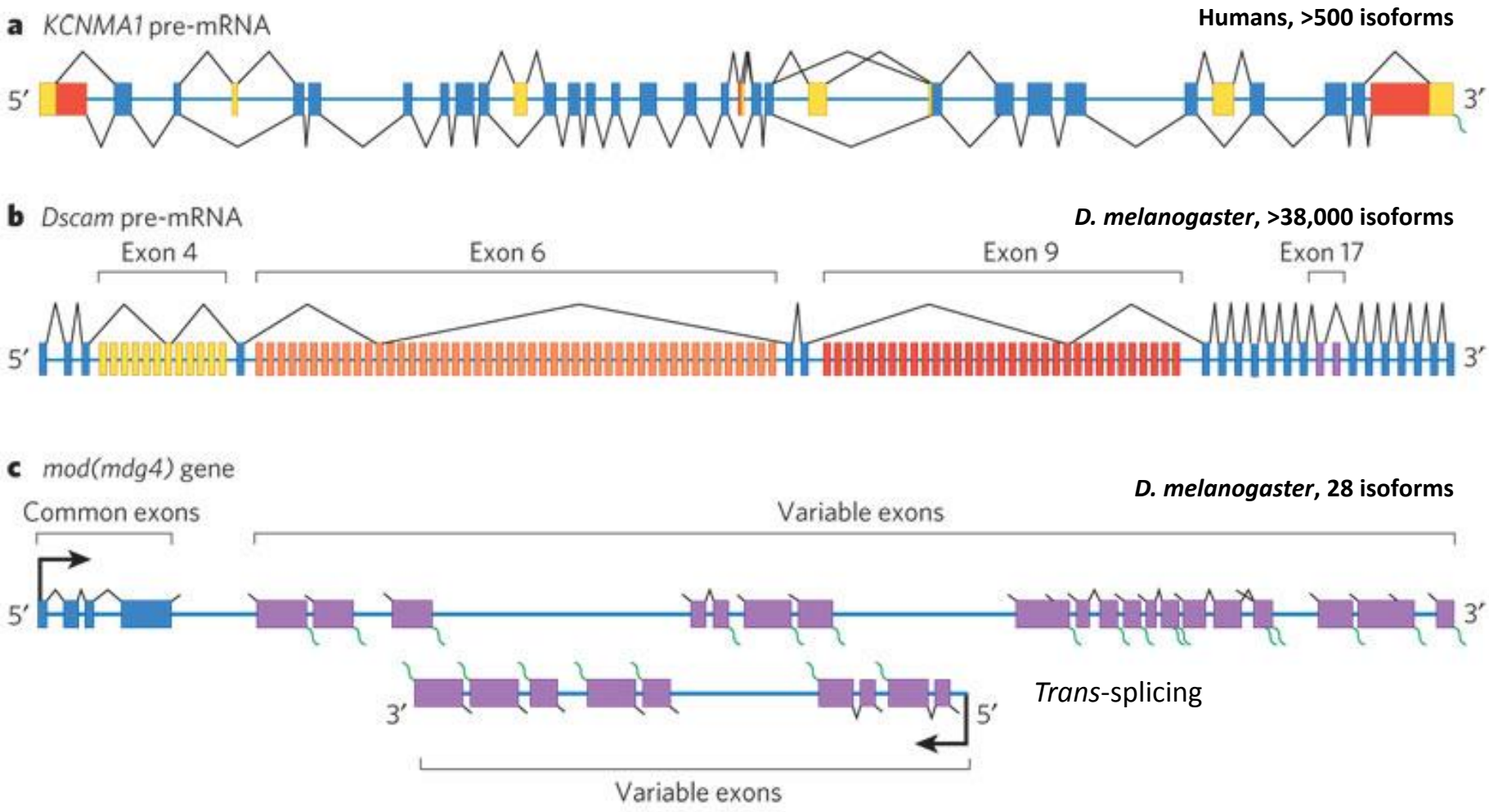
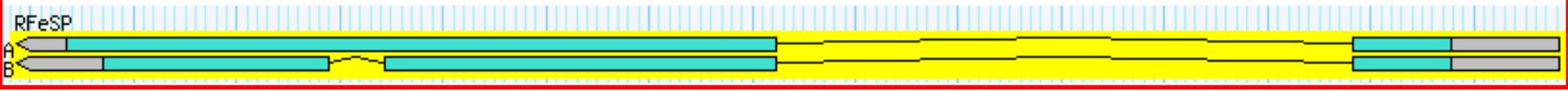


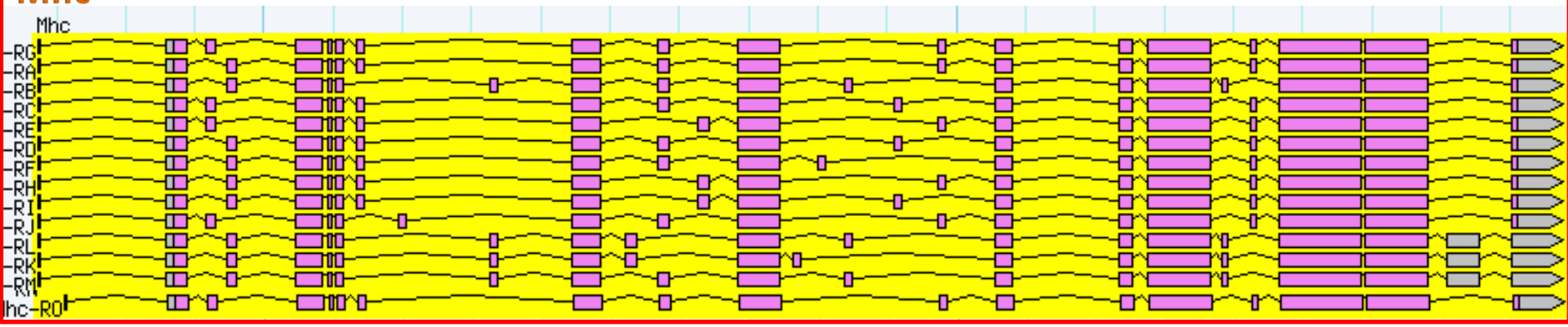
Figure 2. Nielsen and Graveley (2010) *Nature* 463: 457-463

QUIZ

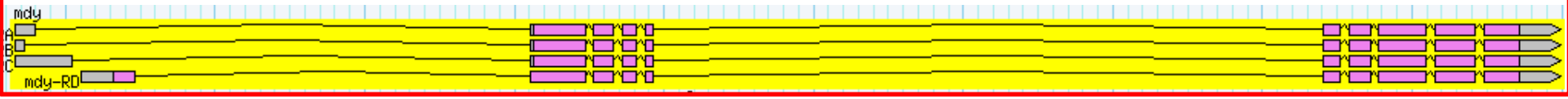
RFeSP



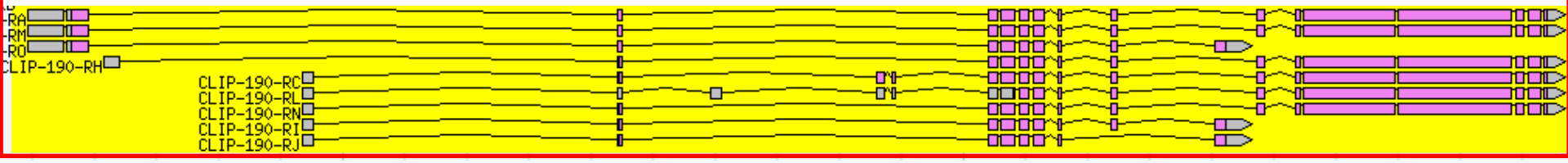
Mhc



mdy



CLIP-190



TRANSCRIPT START/END

- ● Inici transcripció alternatiu
- ● Final transcripció alternatiu

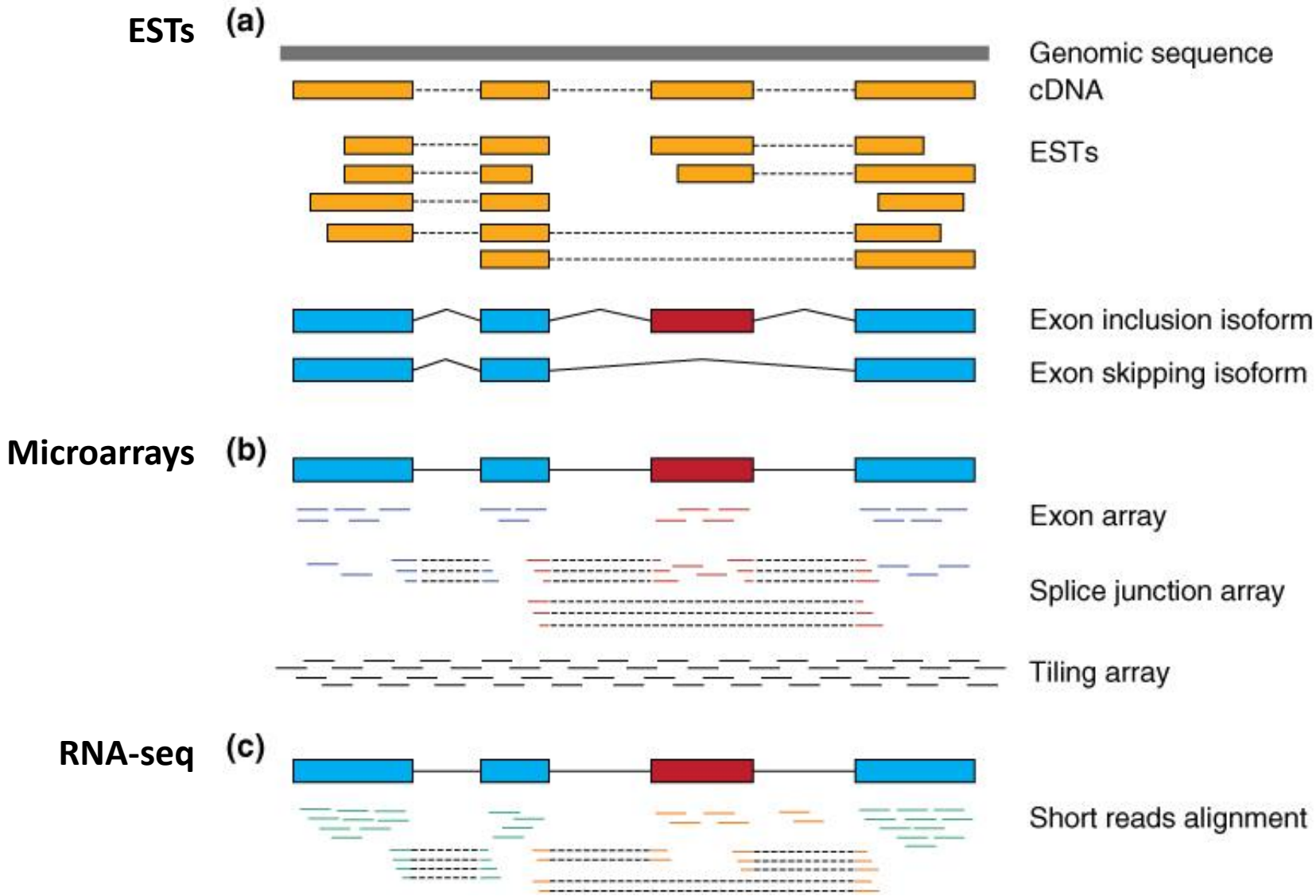
INTERNAL EXONS

- ● Inclusió/Exclusió d'exons
- ○ Exons mutualment excloents

INTRON SPLICING

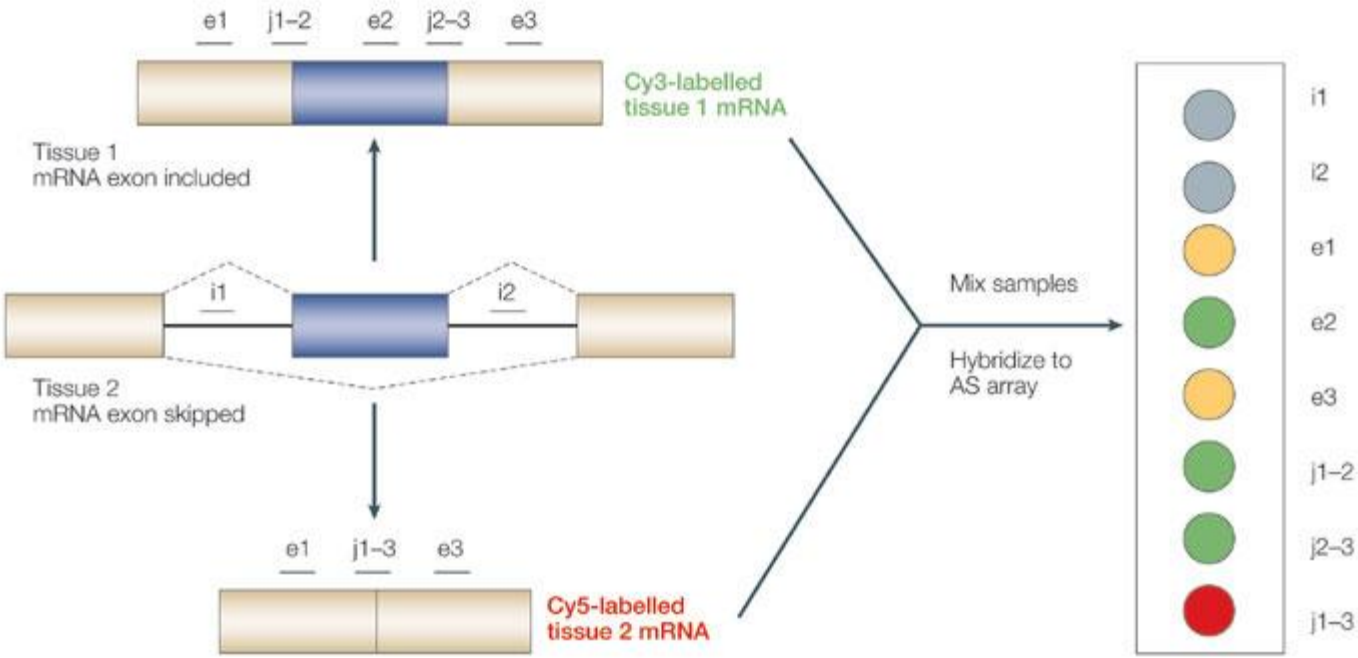
- ○ Punt splicing 5' alternatiu
- ○ Punt splicing 3' alternatiu
- ○ Retenció d'introns

Methods to specifically analyze alternative splicing



Xiao *et al.* (2010) *WIREs Syst Biol Med* DOI: 10.1002/wsbm.084

- Alternative splicing arrays
  - Probes at the exons and exon junctions
  - They allow quantifying the expression of different isoforms



Nature Reviews | Molecular Cell Biology

Box 2. Matlin *et al*. (2005) *Nature Reviews Molecular Cell Biology* 6: 386-398.

Alternative splicing in *Drosophila melanogaster*

Table 1 | Classification of alternative splicing events

Splicing event	Diagram	FlyBase r5.12	modENCODE	New events	Short poly(A)$^+$ RNA-Seq	Significantly changing
Cassette exons		793	2,717	2,014	2,369	1,539
Alternative 5′ splice sites		843	5,192	4,599	4,583	3,142
Alternative 3′ splice sites		879	6,253	5,505	5,579	3,242
Mutually exclusive exons		229	251	123	228	226
Coordinate cassette exons		301	1,227	979	992	467
Alternative first exons		1,767	4,936	3,442	4,473	3,996
Alternative last exons		227	604	432	553	471
Retained/unprocessed introns		1,434	2,679 (5,667)	1,275 (4,263)	2,439 (35,641)	868 (8,998)
Total		6,437	23,859 (26,847)	18,369 (21,478)	21,216 (54,418)	13,951 (22,081)

The number of retained/unprocessed introns in parentheses indicates the total number identified, whereas the number not in parentheses indicates the subset of identified events that have been validated by cDNA sequences or FlyBase 5.12 annotations.

Alternative splicing is present in:

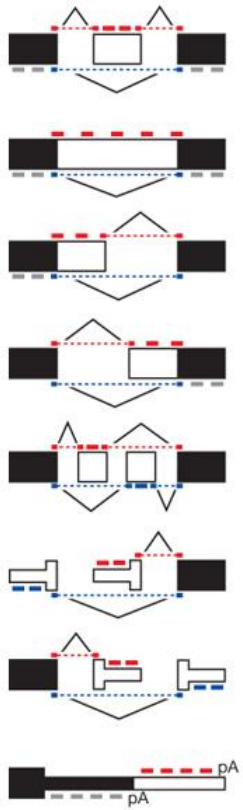
7473 genes

60.7% of the 12,295 genes expressed and with multiple exons

Table 1. Graveley *et al*. (2011) *Nature* 471: 473-479

Alternative splicing in humans

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
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

- 92-94% of the human genes show alternative splicing
- 86% of them in appreciable quantities (>15% frequency of the rarest transcript)
- Most alternative transcripts are expressed in different tissues as a result of a specific regulation

Figure 2. Wang *et al*. (2008) *Nature* 456: 470-476.

Regulation of alternative splicing

Developmentally regulated splicing variants in *D. melanogaster*

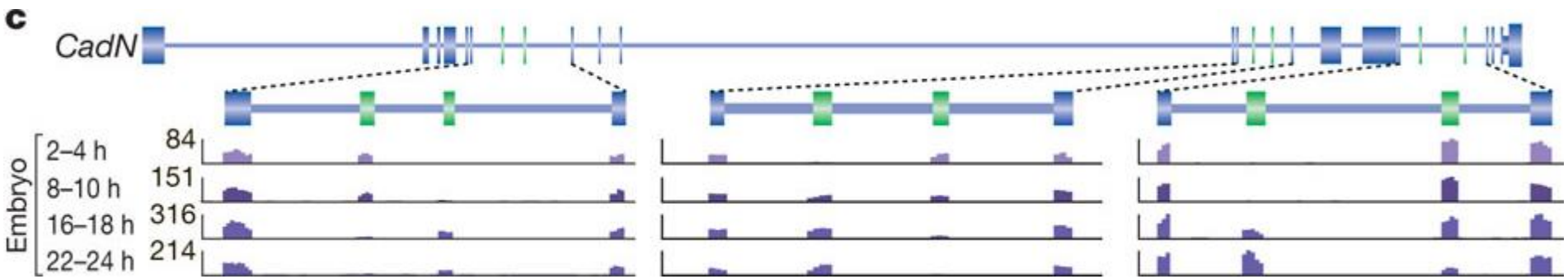


Figure 4. Graveley *et al*. (2011) *Nature* 471: 473-479

Tissue-regulated splicing variants in humans

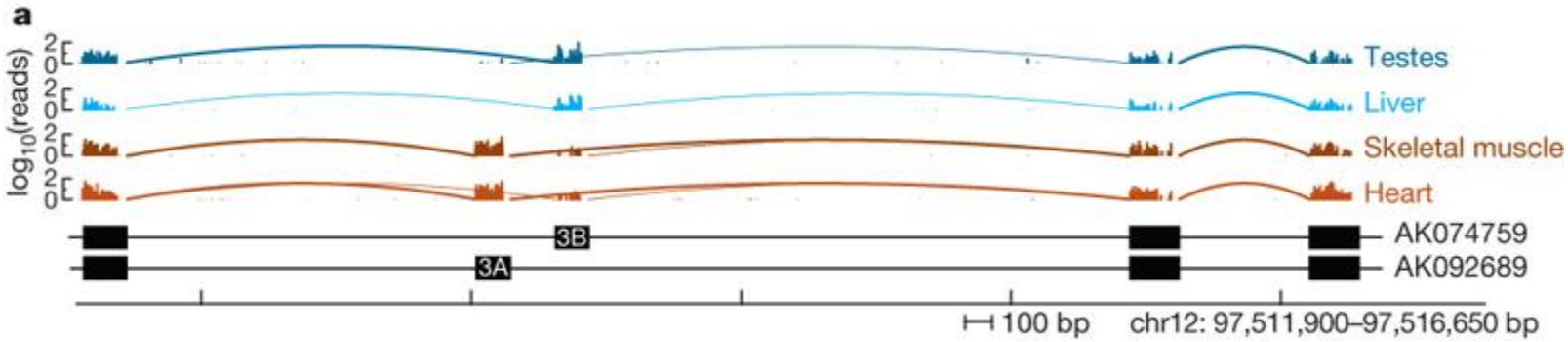


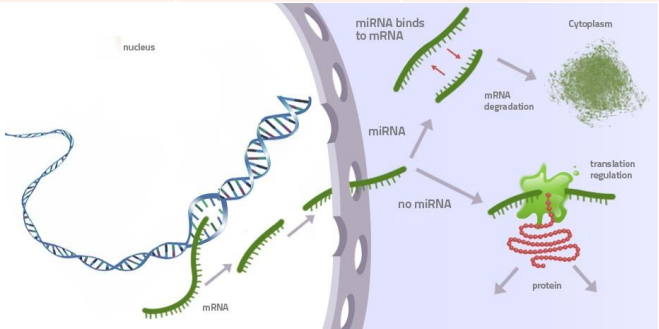
Figure 1. Wang *et al*. (2008) *Nature* 456: 470-476.

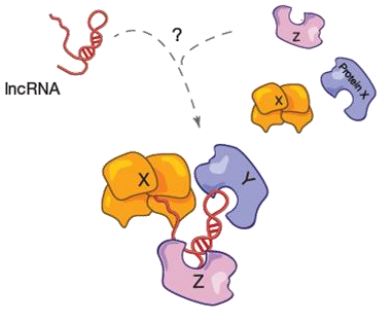
Non-coding RNAs

Type	Name	Size	Genes	Transcripts	Function
Small non-coding RNAs					
rRNAs	ribosomal RNAs	114-5000 nt	549 + 2	549 + 2	Component of ribosome
tRNAs	transfer RNAs	73-93 nt	624* + 22	624* + 22	Translation
miRNAs	micro RNAs	21-23 nt	3,837	3,837	Gene expression regulation
snRNAs	small nuclear RNAs	100-300 nt	1,912	1,912	Splicing
snoRNAs	small nucleolar RNAs	60-300 nt	978	978	RNA modification
Long non-coding RNAs					
lncRNAs	long non-coding RNAs	>200 nt	15,877	26,414	Regulation, imprinting…

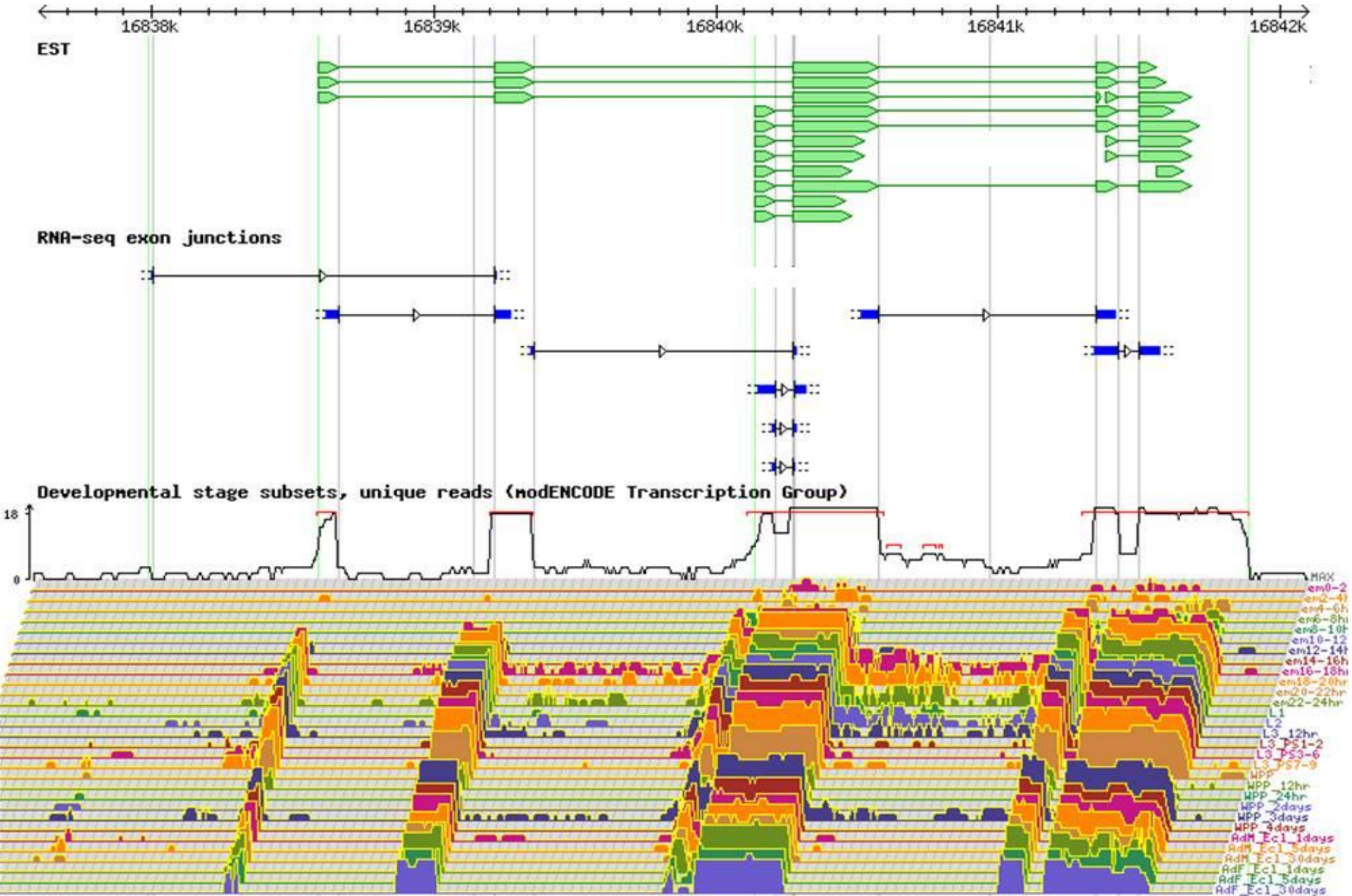
Number of transcripts from GENCODE v21 data, http://www.gencodegenes.org/stats.html

* Number of transcripts from GENCODE v7 data





QUIZ


16838k
16839k
16840k
16841k
16842k
EST
RNA-seq exon junctions
Developmental stage subsets, unique reads (modENCODE Transcription Group)
18
0
MAX
em0-2
em2-4h
em4-6h
em6-8hr
em8-10h
em10-12
em12-14h
em14-16h
em16-18hr
em18-20hr
em20-22hr
em22-24hr
L1
L2
L3_12hr
L3_PS1-2
L3_PS3-6
L3_PS7-9
WPP
WPP_12hr
WPP_24hr
WPP_2days
WPP_3days
WPP_4days
AdM_Ecl_1days
AdM_Ecl_5days
AdM_Ecl_30days
AdF_Ecl_1days
AdF_Ecl_5days
AdF_Ecl_30days