Master of Molecular Medicine
Module IV Functional Genomics

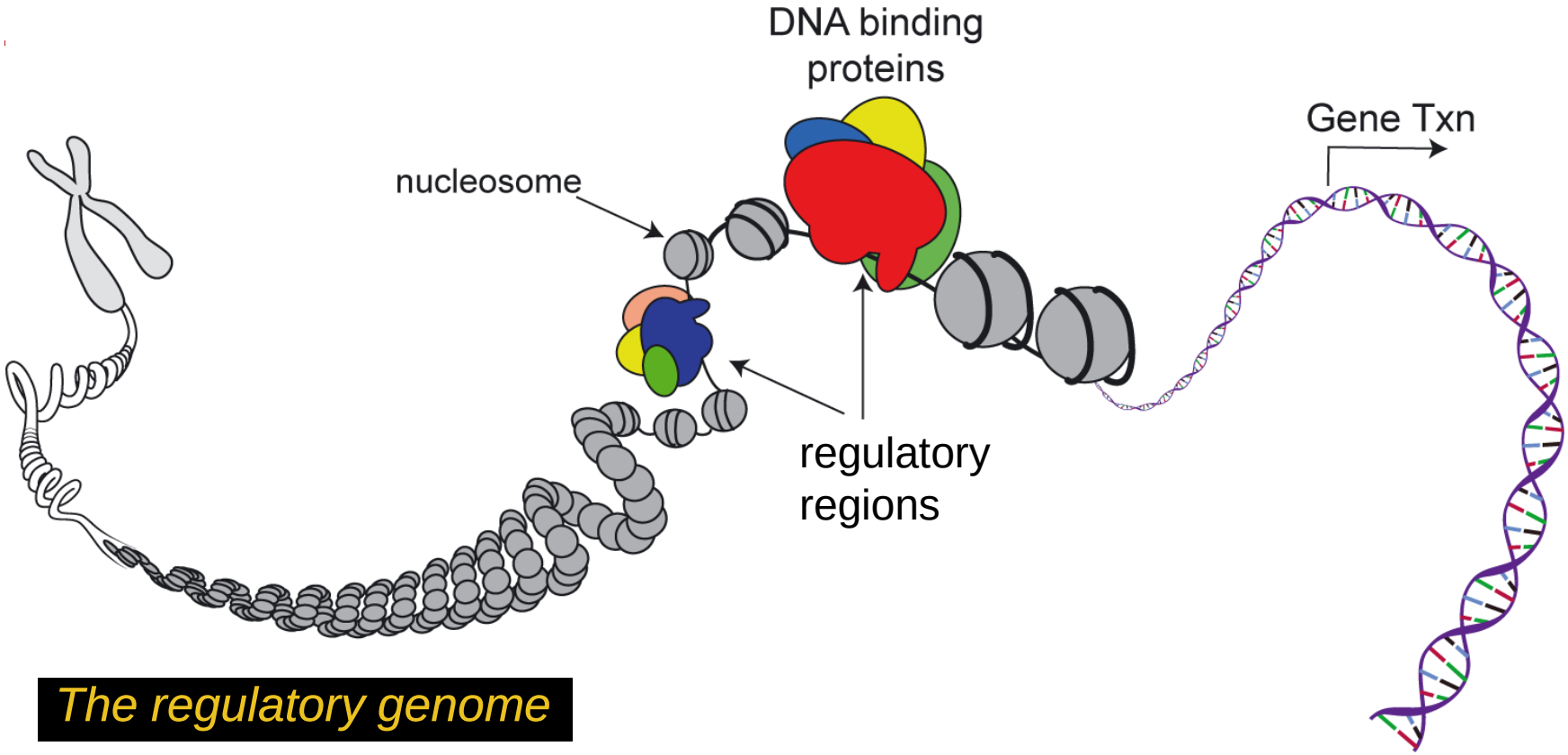
RNAseq & alternative splicing

Uwe Ohler

Max Delbrueck Center

uwe.ohler@mdc-berlin.de

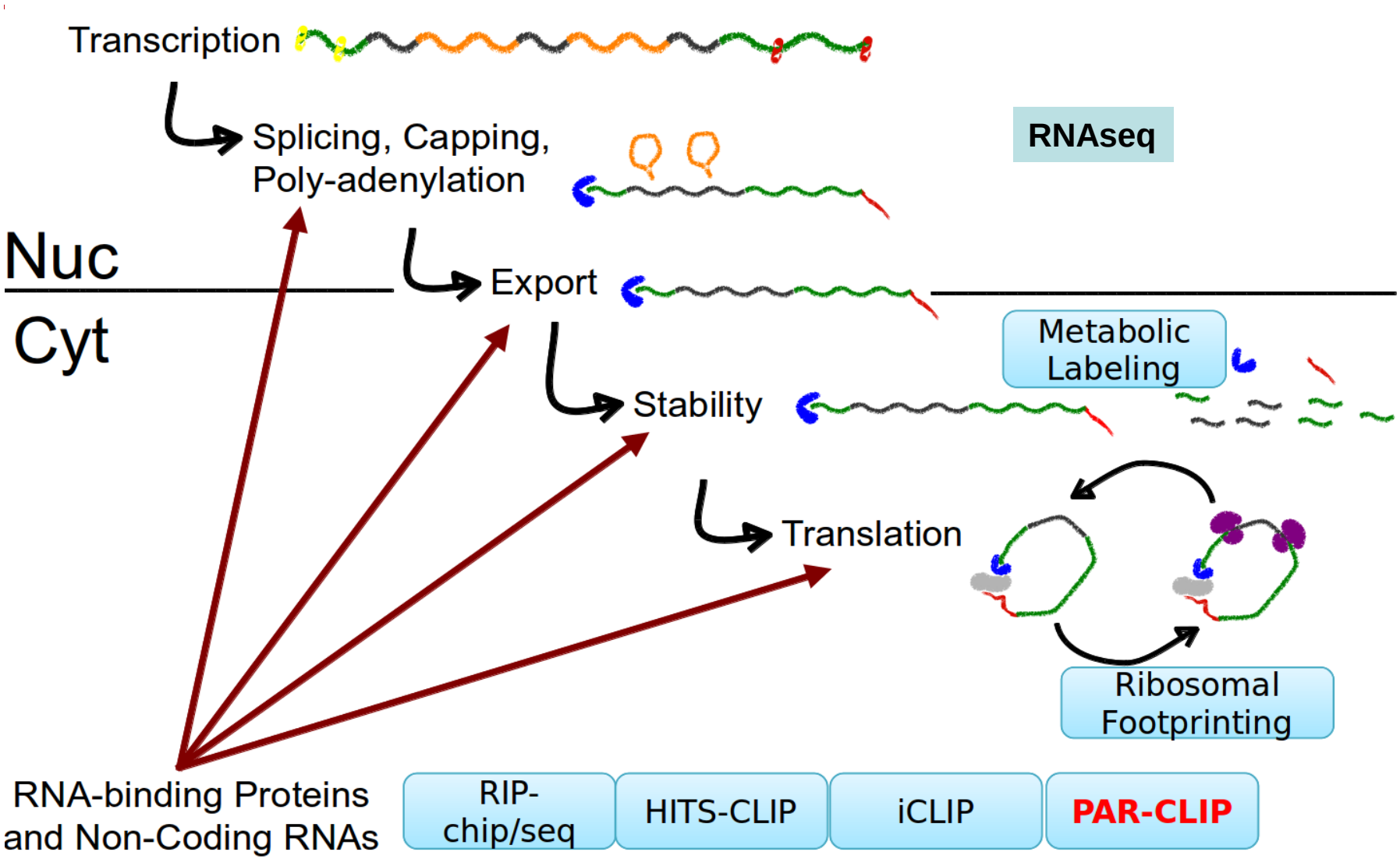
The regulatory genome: transcription



The regulatory genome

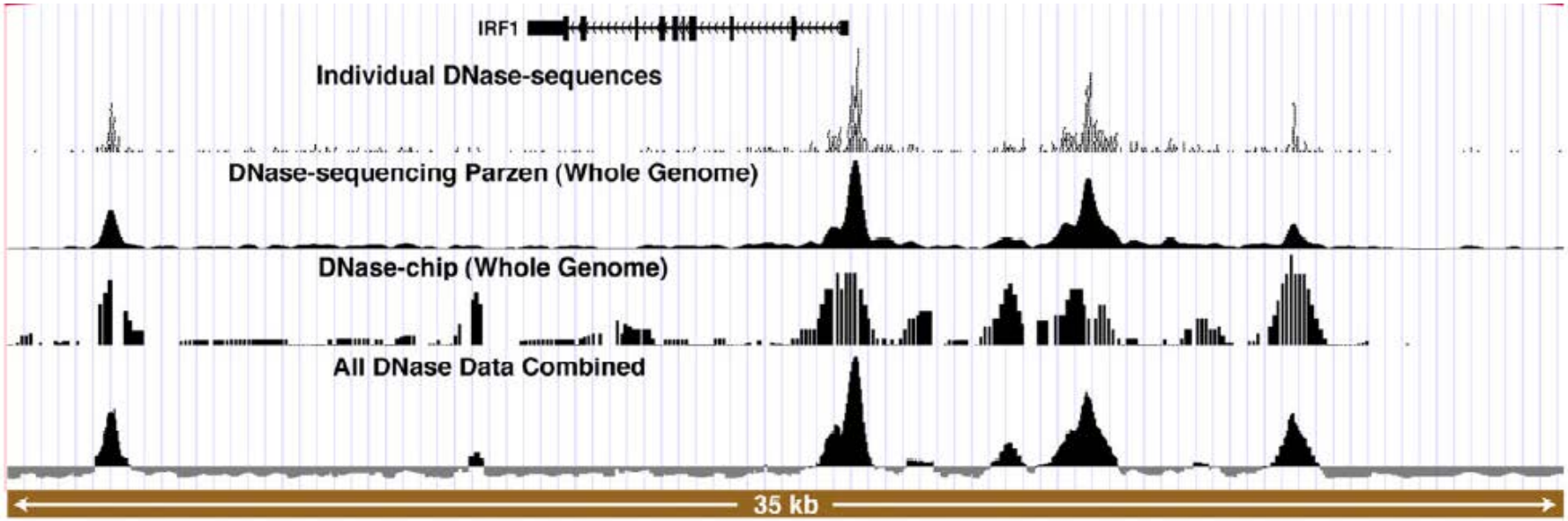
Variation & evolution (whole-genome sequencing)
Chromatin structure (histone modifications; “epigenetic code”)
Promoter & enhancers (mapping target sites of regulatory factors)
Transcript variants + expression (RNA sequencing)

Post-transcriptional regulation



The impact of deep sequencing

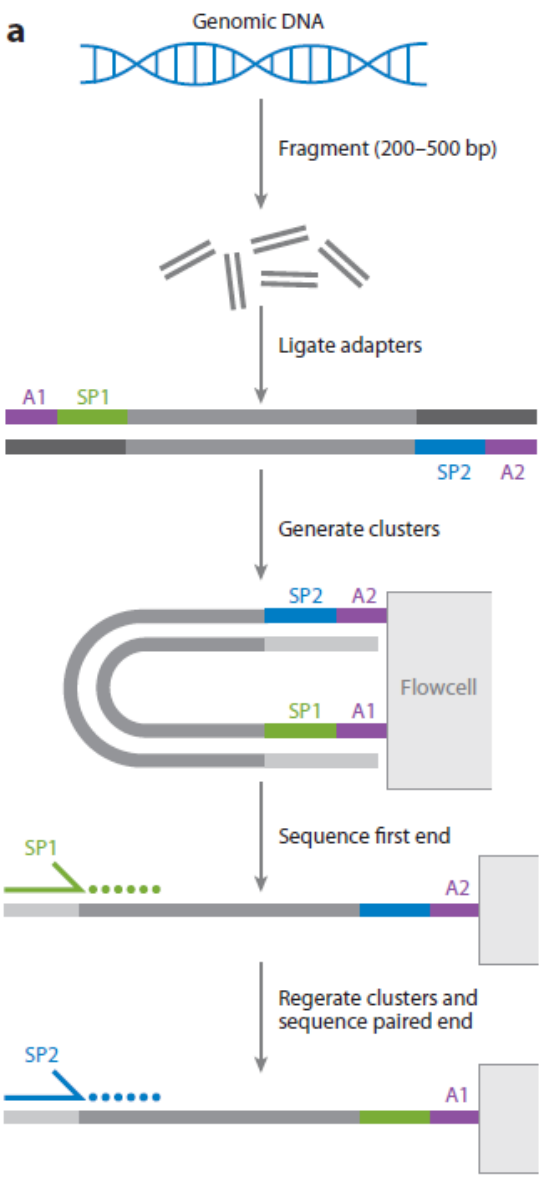
- In the past years, new sequencing technologies have made the anticipated quantum leap
  - Illumina HiSeq: typically ~100-200 mio reads of ~100 bp
  - Unbiased exploration of genomes (DNA) & transcriptomes (RNA)
  - We can now study the 98% of the genome that is non-coding (and not only the 2% that codes for proteins)

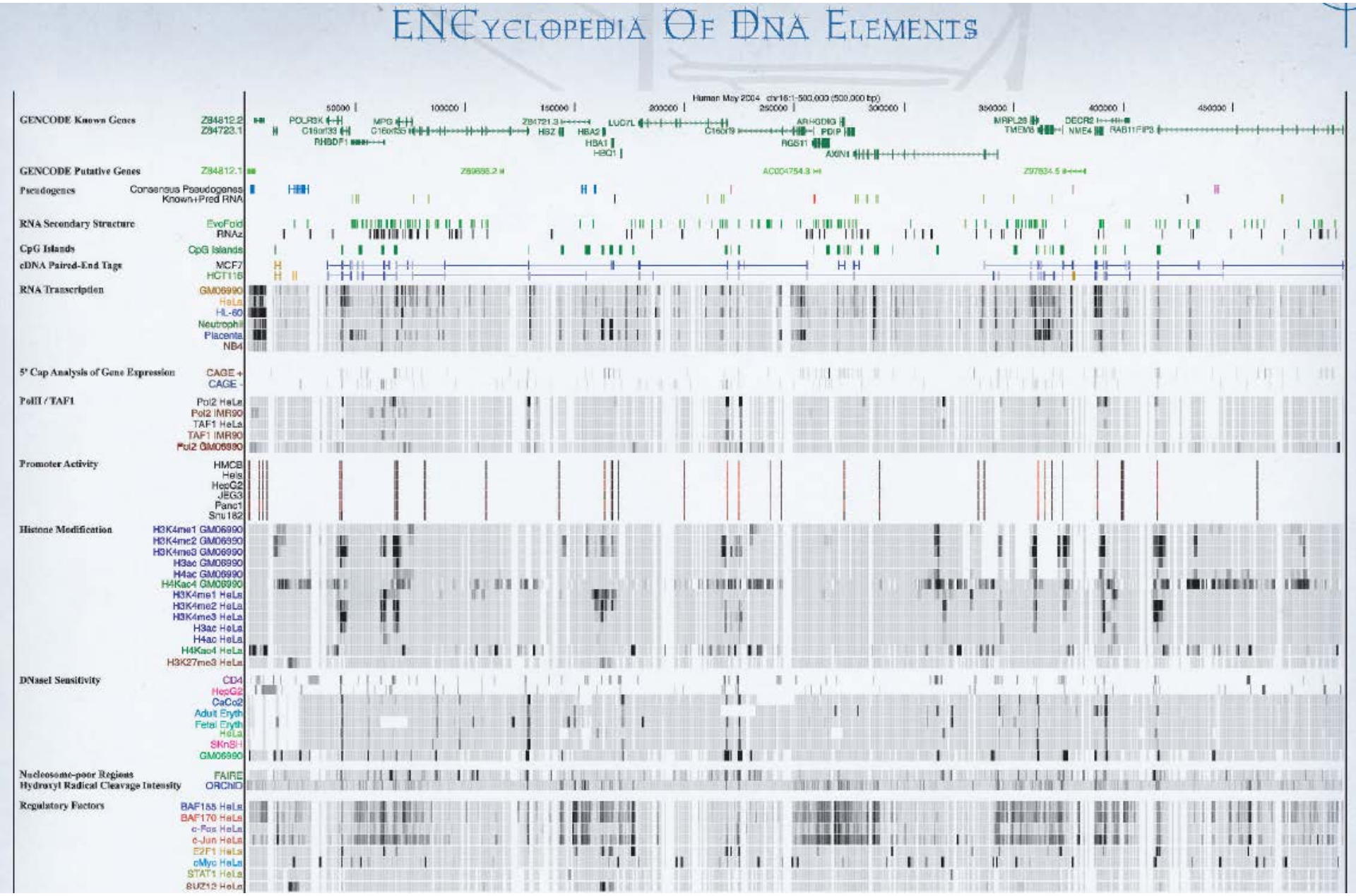


Boyle et al 2008

Platforms

- Most popular approach: amplification & sequencing by synthesis
  - “short” reads, 100+ bp
  - Reversible dyes (Illumina, since 2008)
  - Microfluidics, pH based (Ion Torrent, since 2010)
- Single molecule sequencing
  - Longer reads, 1000+bp; each molecule is sequenced multiple times
  - Example: PacBio




ENCyclopedia Of DNA Elements
Human May 2004 chr16:1-500,000 (500,000 bp)
GENCODE Known Genes
GENCODE Putative Genes
Pseudogenes
Consensus Pseudogenes
Known+Pred RNA
RNA Secondary Structure
EvoFold
RNAz
CpG Islands
cDNA Paired-End Tags
MCF7
HCT116
RNA Transcription
GM06990
HeLa
HL-60
Neutrophil
Placenta
NB4
5' Cap Analysis of Gene Expression
CAGE +
CAGE -
PolII / TAF1
Pol2 HeLa
Pol2 IMR90
TAF1 HeLa
TAF1 IMR90
Pol2 GM06990
Promoter Activity
HMCB
Hela
HepG2
JEG3
Panc1
Snu182
Histone Modification
H3K4me1 GM06990
H3K4me2 GM06990
H3K4me3 GM06990
H3ac GM06990
H4ac GM06990
H4Kac4 GM06990
H3K4me1 HeLa
H3K4me2 HeLa
H3K4me3 HeLa
H3ac HeLa
H4ac HeLa
H4Kac4 HeLa
H3K27me3 HeLa
DNaseI Sensitivity
CD4
HepG2
CaCo2
Adult Eryth
Fetal Eryth
HeLa
SKnSH
GM06990
Nucleosome-poor Regions
Hydroxyl Radical Cleavage Intensity
FAIRE
ORCHID
Regulatory Factors
BAF155 HeLa
BAF170 HeLa
c-Fos HeLa
c-Jun HeLa
E2F1 HeLa
cMyc HeLa
STAT1 HeLa
SUZ12 HeLa

ENCODE 2012

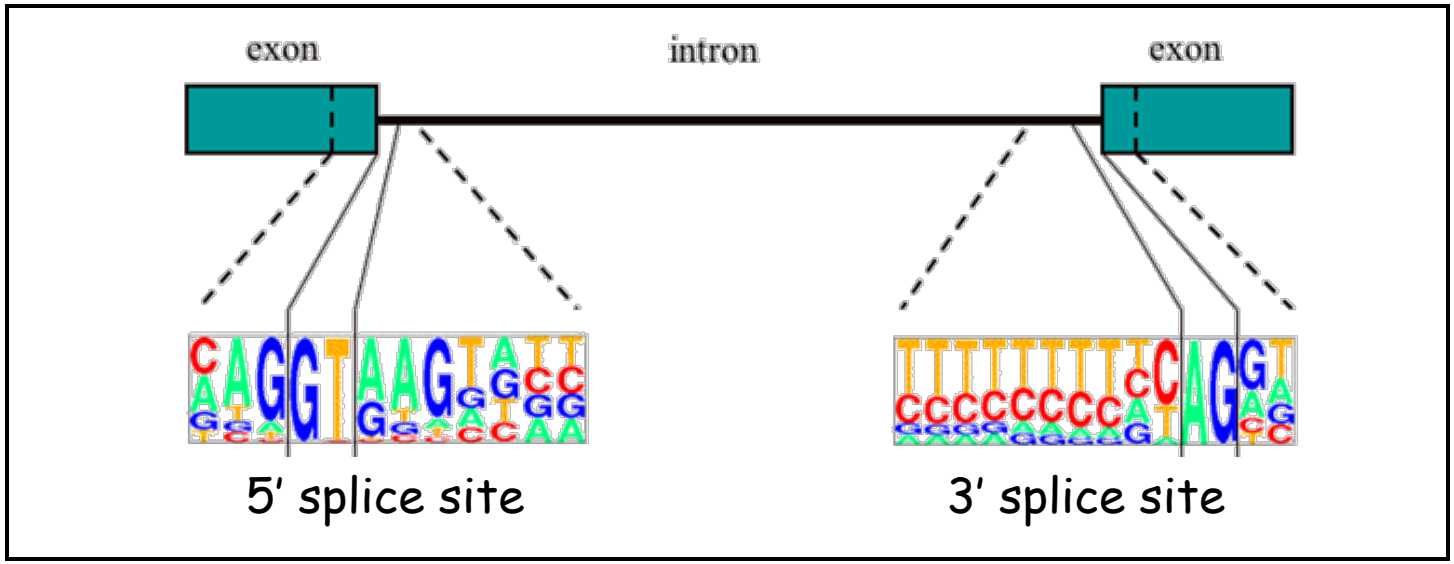
Computational challenges

"Deep sequencing" is a catch-phrase for 100s of different biochemical assays,
with some common and some application specific tasks

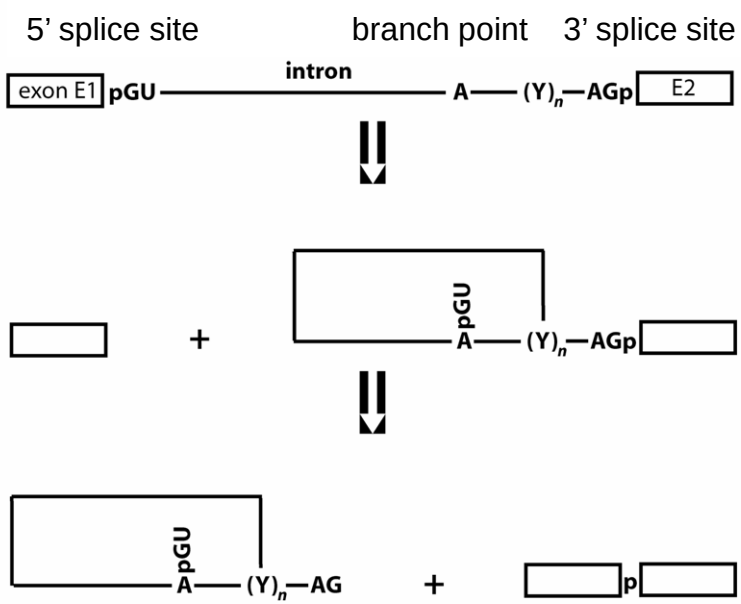
1. Mapping reads to the genome or transcriptome
   - Spliced or contiguous
   - Example read aligners: Bowtie, BWA, STAR (Cufflinks; Scripture, etc)
2. Quantifying read density
   - Identification of sequence variation (exome seq)
   - Gene expression changes
   - Protein/DNA (or RNA) interactions

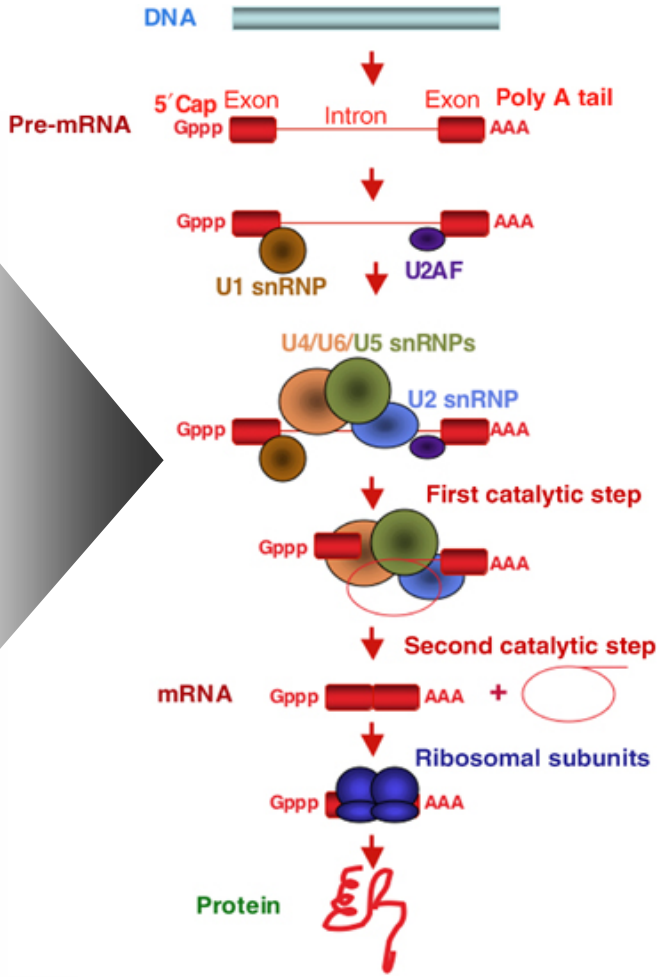
Protein coding genes and splice sites

- Most eukaryotic protein-coding genes have a **split gene structure** of exons and introns
- Conserved **sequence motifs** mark beginning and end of exons



Steps in RNA splicing





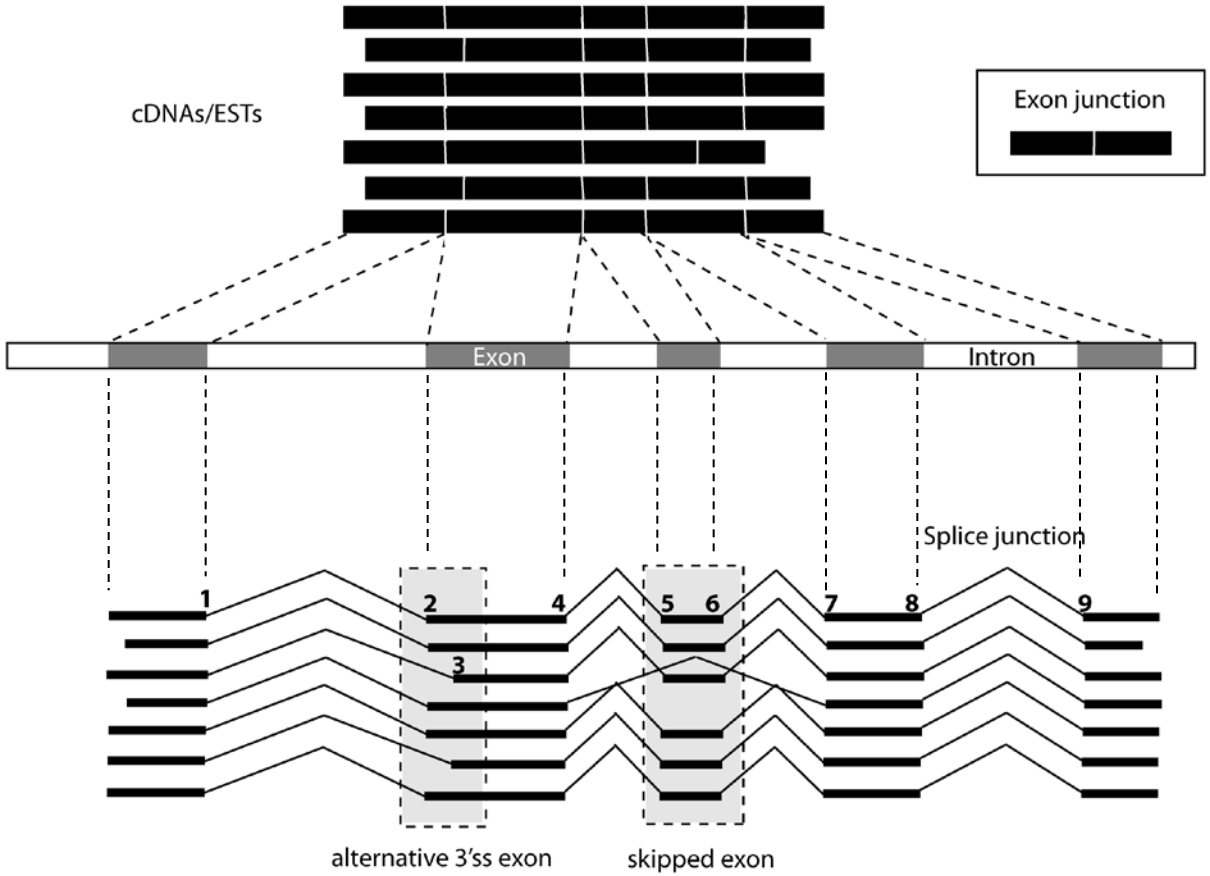
Soares & Valcarcel (2006)

Classical (long read) transcript structure assembly

Given an incomplete and complementary transcript sequence (**cDNA/mRNA**), find the complete genomic sequence (**primary transcript structure**) and determine the gaps (**introns**)

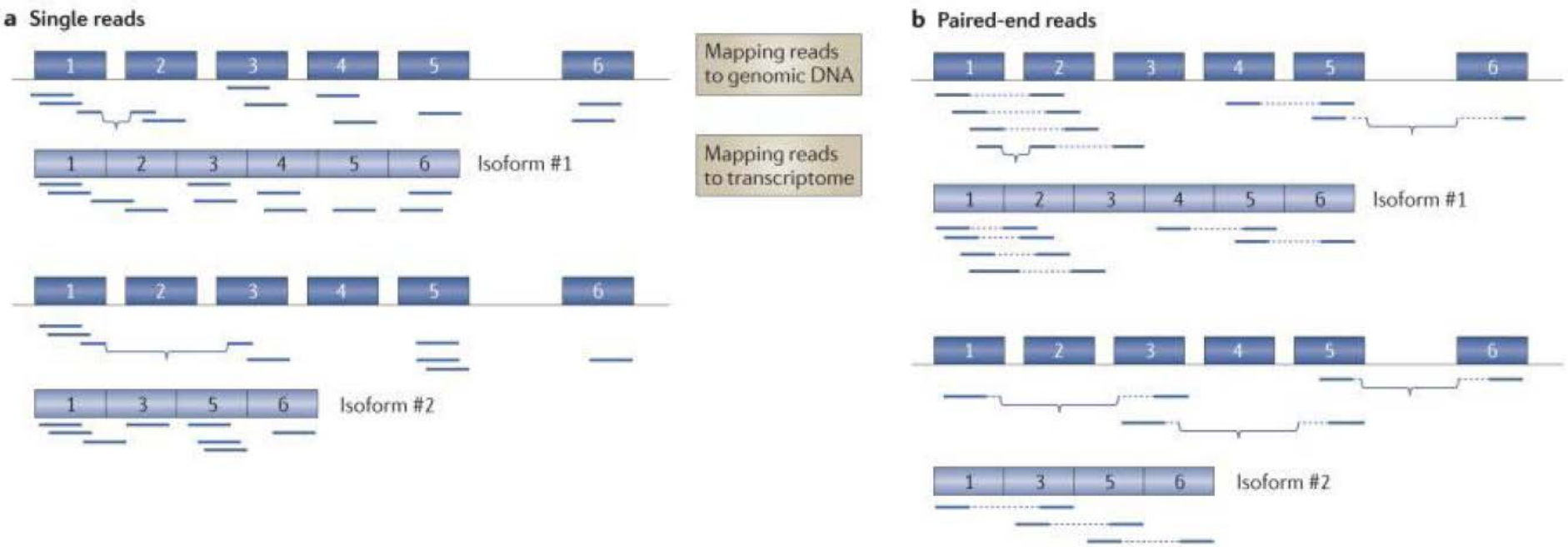
- Align sequences with gaps and determine 'best' match

- Gaps start/end with 5'ss and 3'ss signals

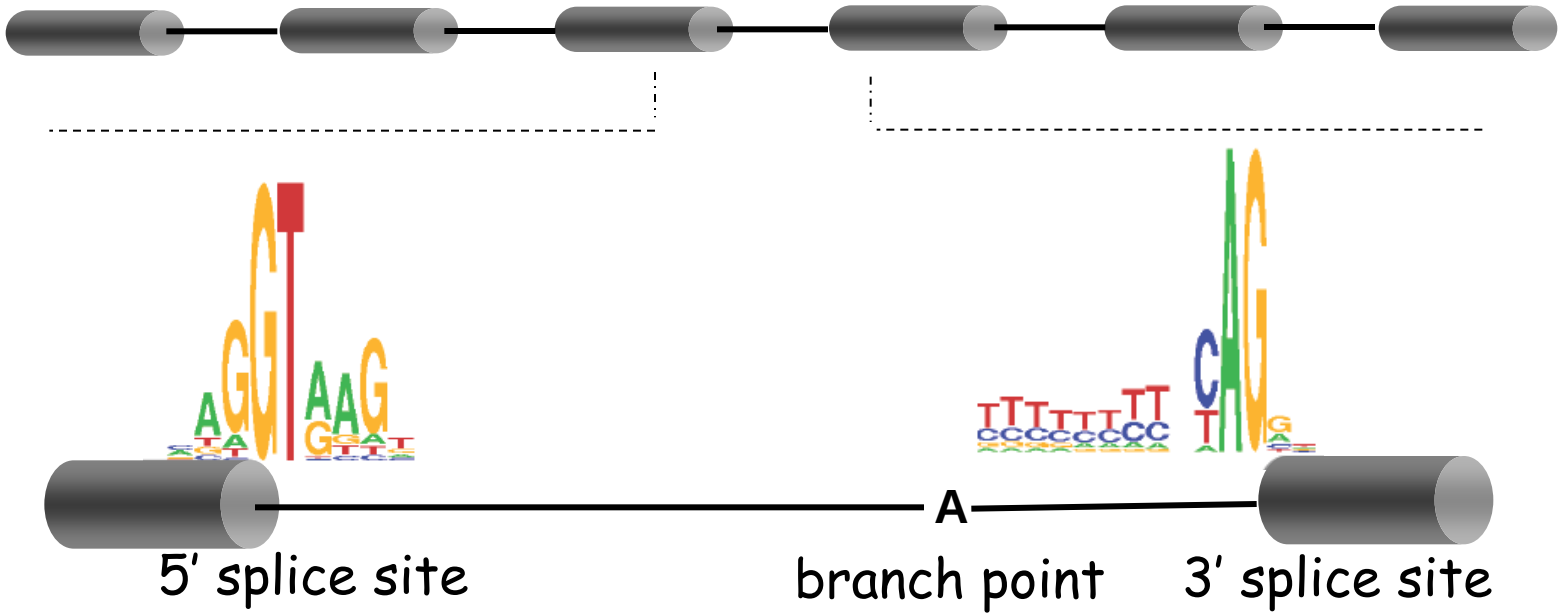


Short-read spliced alignment

- From expression quantification of whole genes, to exons, and isoforms
  - Alignment to known isoforms
  - Simultaneous reconstruction and quantification



Signals: RNA binding sites recognized by the spliceosome

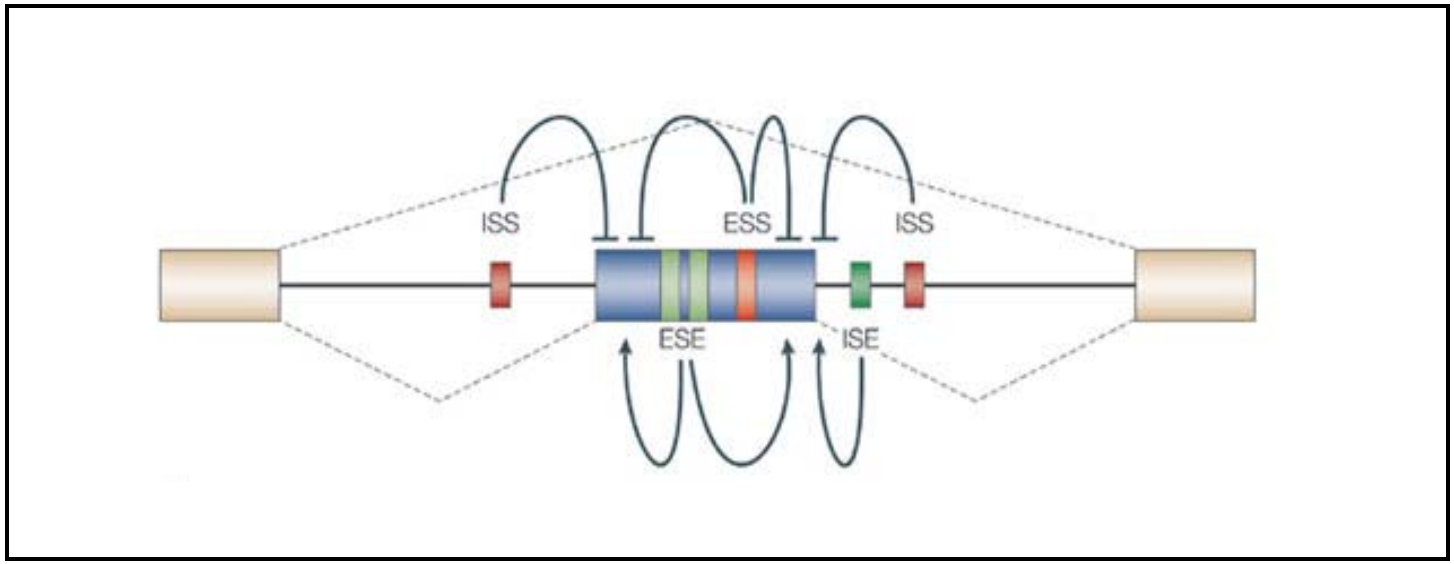


Signal/Species	5'ss	BP	3'ss	
S.cerevisiae	11	12	7	30
C.elegans	8	5	11	24
D.melanogaster	9	5	10	24
H.sapiens	8	5	8	21

Weaker splice sites, yet more signals enhancing (ESE) or silencing (ESS) exons in higher eukaryotes

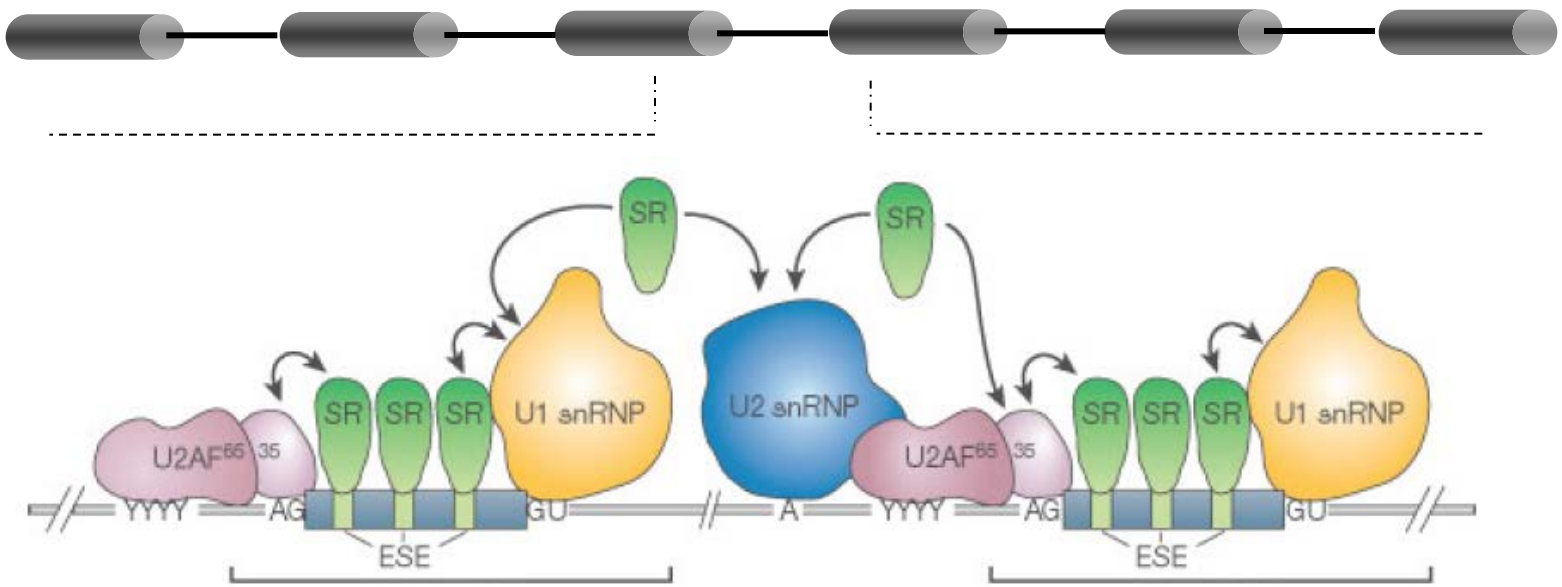
The reality is of course more complicated

- In higher eukaryotic genomes (beyond yeast), splice sites alone often do not contain sufficient information for accurate splicing, compensated for by **splicing regulatory elements** in both exons and introns



Smith *et al* (2005)

Signals and *trans*-acting factors

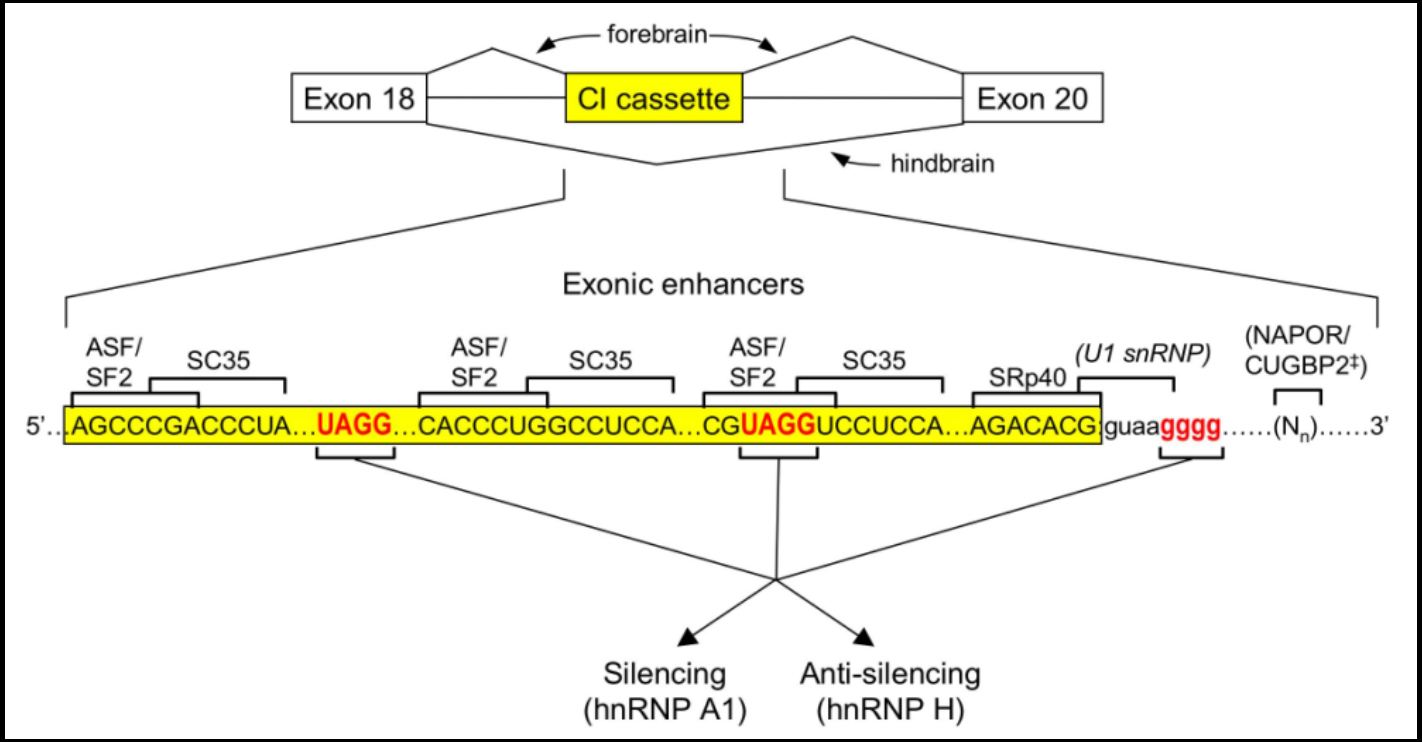


Signal/Species	5'ss	BP	3'ss	
S.cerevisiae	11	12	7	30
C.elegans	8	5	11	24
D.melanogaster	9	5	10	24
H.sapiens	8	5	8	21

Weaker splice sites, yet more signals enhancing (ESE) or silencing (ESS) exons in higher eukaryotes

Maniatis (2002)

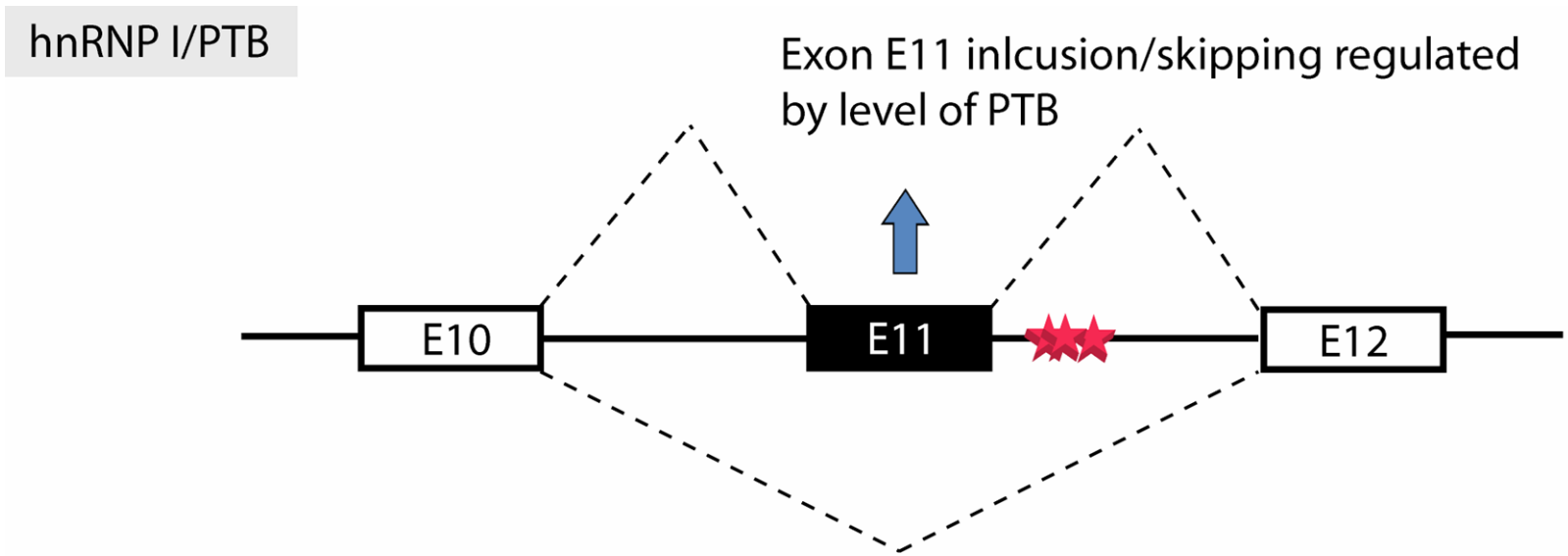
Illustrative example: the *NMDA*-type glutamate receptor



RNA-binding proteins attracted to multiple sequence elements: SR proteins, hnRNPs many (but not all) are conserved

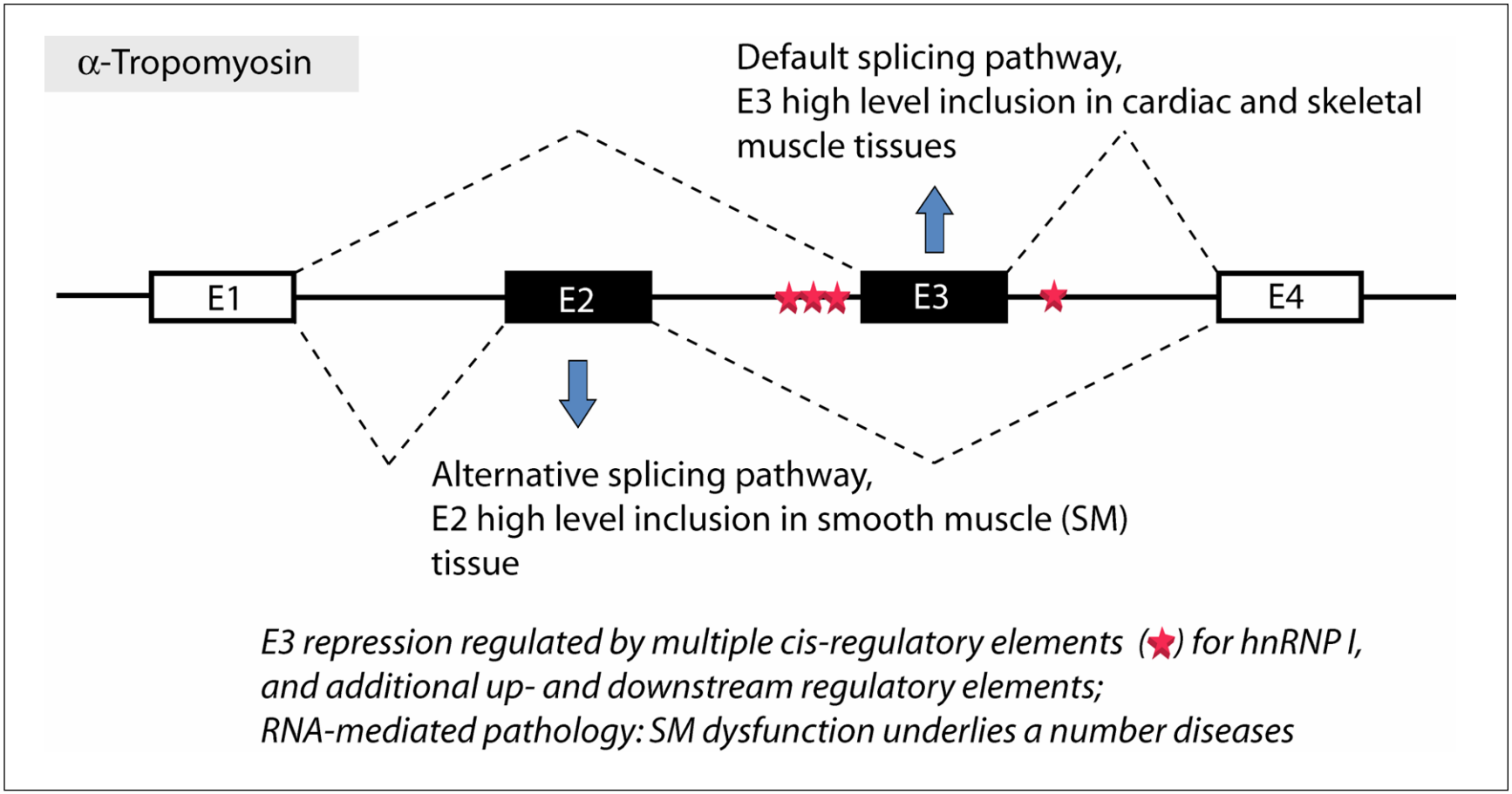
Grabowsky (2005)

Splicing regulation of PTB exon skipping by its own protein

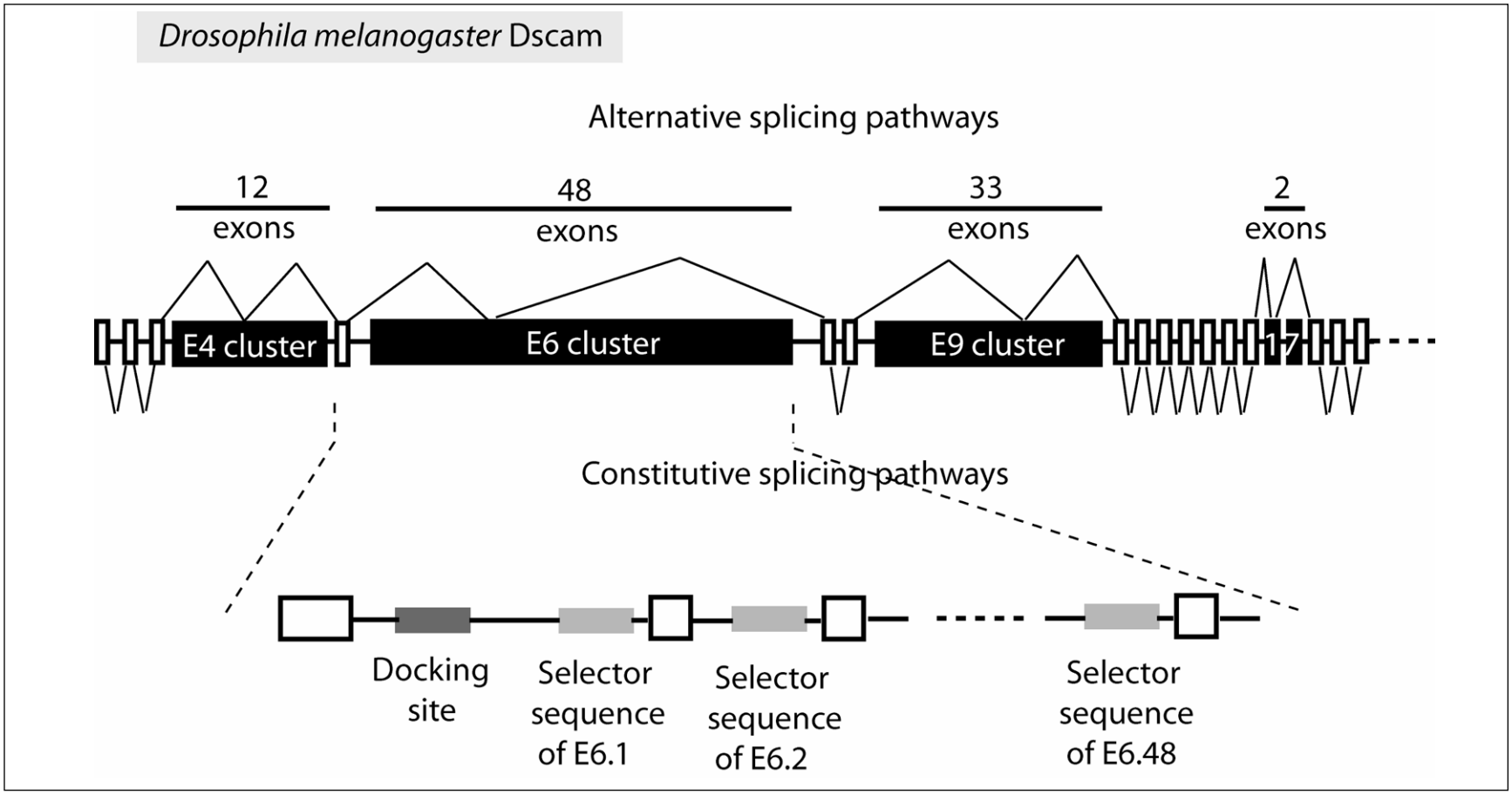


Autoregulatory functional loop: steady state level of E11 skipping low (<1%), skipping of E11 causes frame-shift and creation of premature termination codon (PTC) in E12 (>55 bp of authentic stop codon), substituting the substrate to NMD pathway

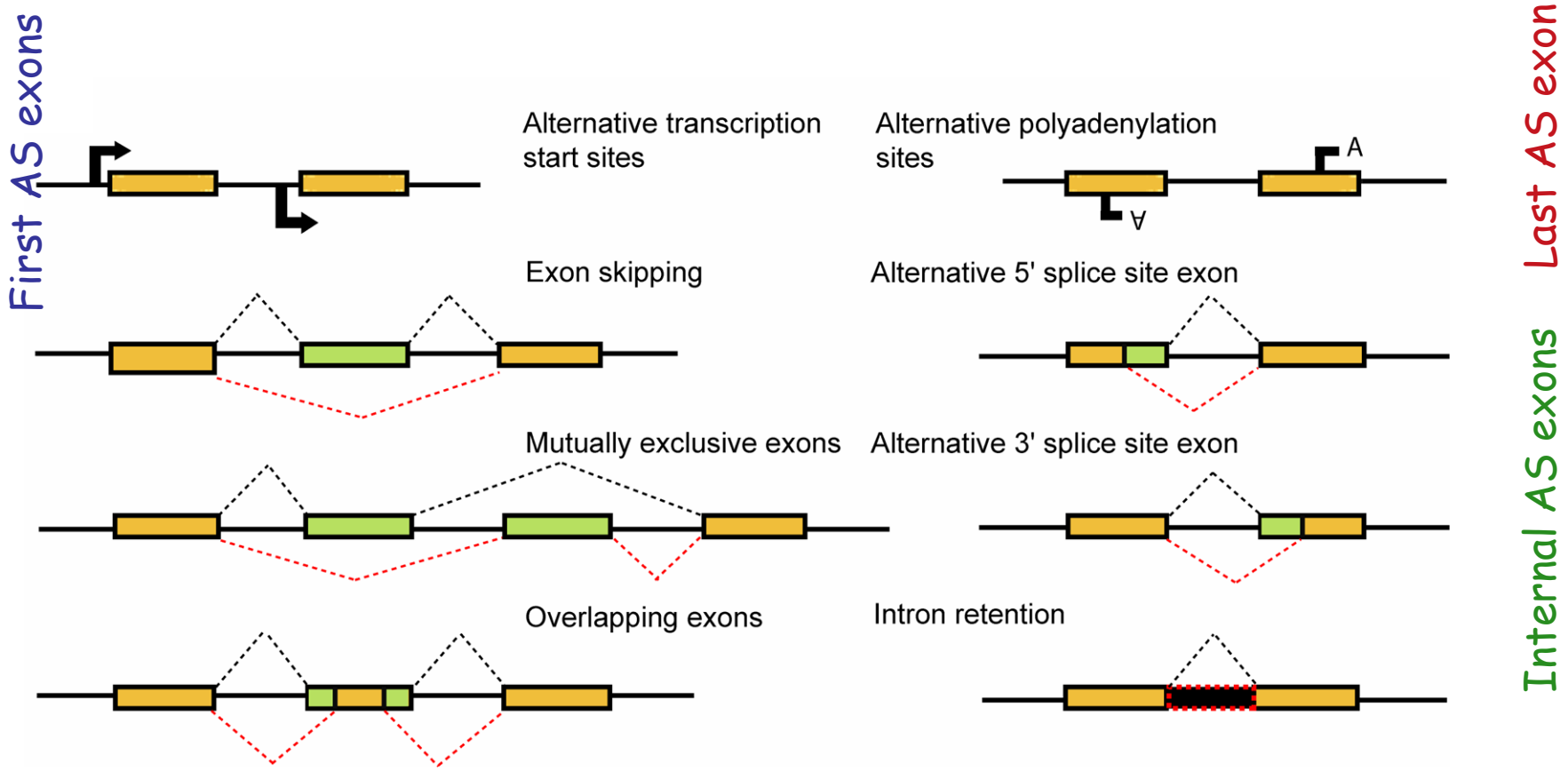
Tissue-specific exon switch using *trans*-regulatory factors



Clusters of alternative exons in *Drosophila's Dscam* gene



A repertoire of alternative splicing patterns



Functional roles and consequences of alternative splicing

RNA splicing and Plasticity of alternative splicing

Spliceosome function

- Precise recognition of splice sites among many pseudo-sites
- Removal of introns
- Production of correct message

RNA processing

- Coupling interaction of RNA splicing with gene expression:
  > Transcription
  > Capping and Polyadenylation
  > mRNA export
  > Surveillance and mRNA degradation

Protein functional diversity

- Cell and tissue-specific mRNA isoforms
- Developmental stage regulated isoforms
- Inducible control and expression of isoforms

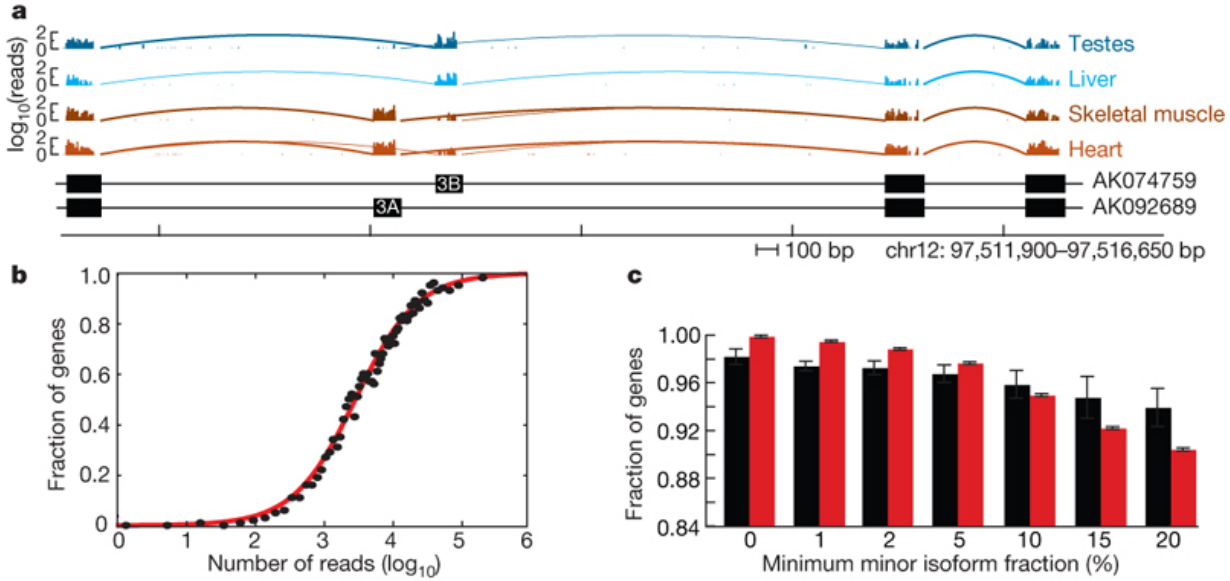
Splicing errors in human disease

- Inherited mutations and effects of genetic background affecting
  > Authentic splice sites
  > Alternative splice sites
  > Basal splicing machinery
  > Trans-acting regulators of AS
- Tumorigenic phenotypes and progression

adapted from **Grabowsky** (2004)

Deep sequencing surveys

- Frequency and relative abundance of alternative splicing isoforms in human genes [Wang et al, 2008]
  - 10 tissues, 5 breast cancer cell lines, > 400 mio reads
  - 94% of human genes are multi-exon;
    ~86% have >=2 isoforms at reasonable levels

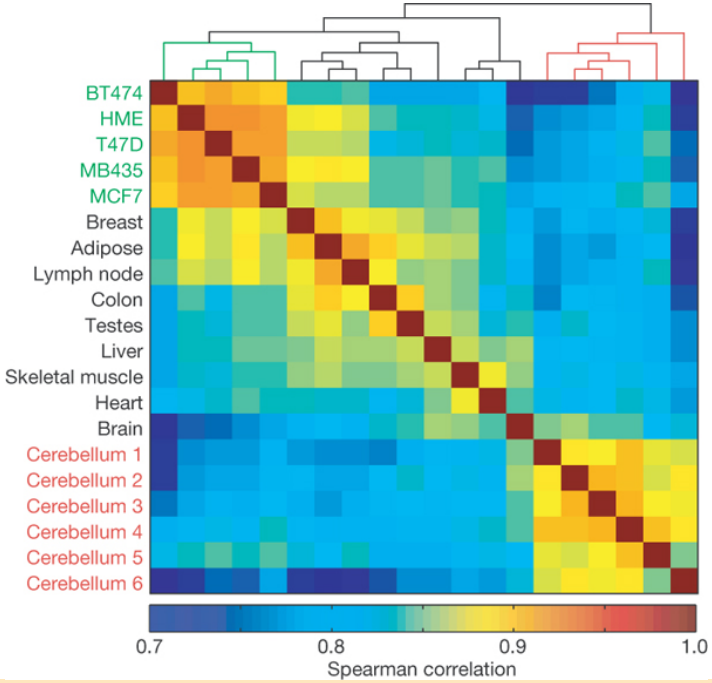


Types and abundances of mammalian AS patterns

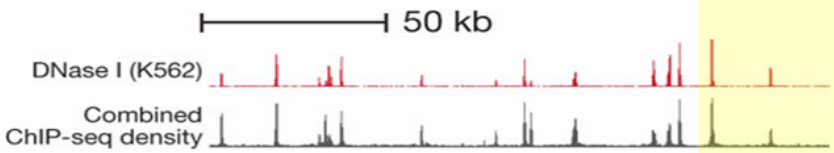
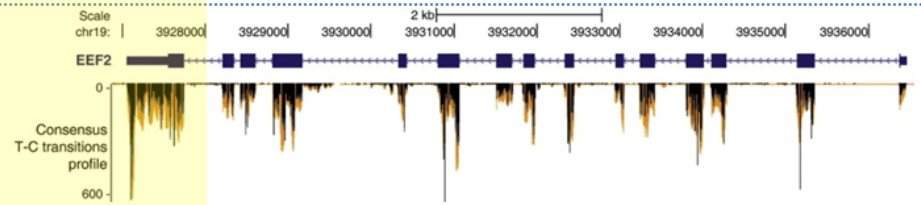
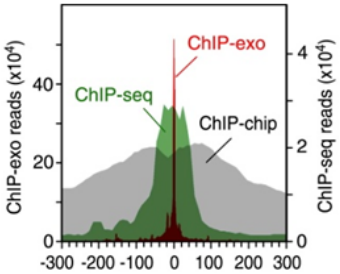
Alternative Transcript Events	No.	Det.	Both Iso.	Biased	Obs. biased (%)	Est. true biased (%)
Skipped exon (SE)	37K	35K	10,436	6822	65	72
Retained Intron (RI)	1K	1K	167	96	57	71
Alternative 5' splice site (A5SS)	15K	15K	2168	1386	64	72
Alternative 3' splice site (A3SS)	17K	16K	4181	2655	64	74
Mutually exclusive exons (MXE)	4K	4K	167	95	57	66

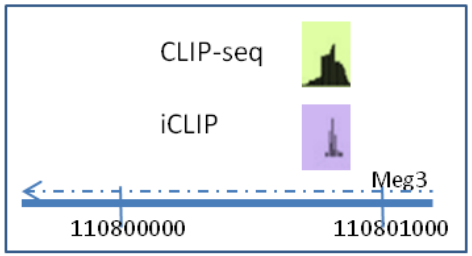
Tissue-specific vs individual-specific variation

- Tissue-specific variation exceeds individual-specific variation
  - 10-30% of AS events vary between individuals; 47-74% differ between tissues



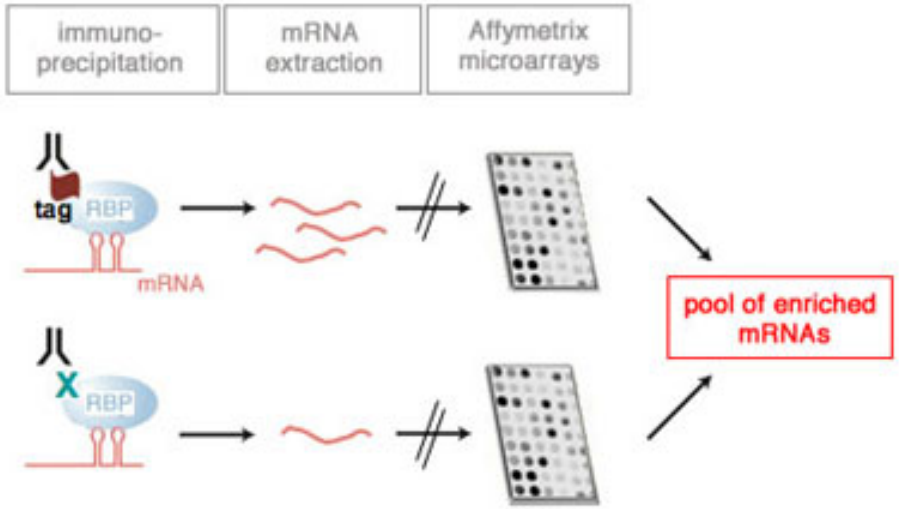
Quantification of regulatory interactions

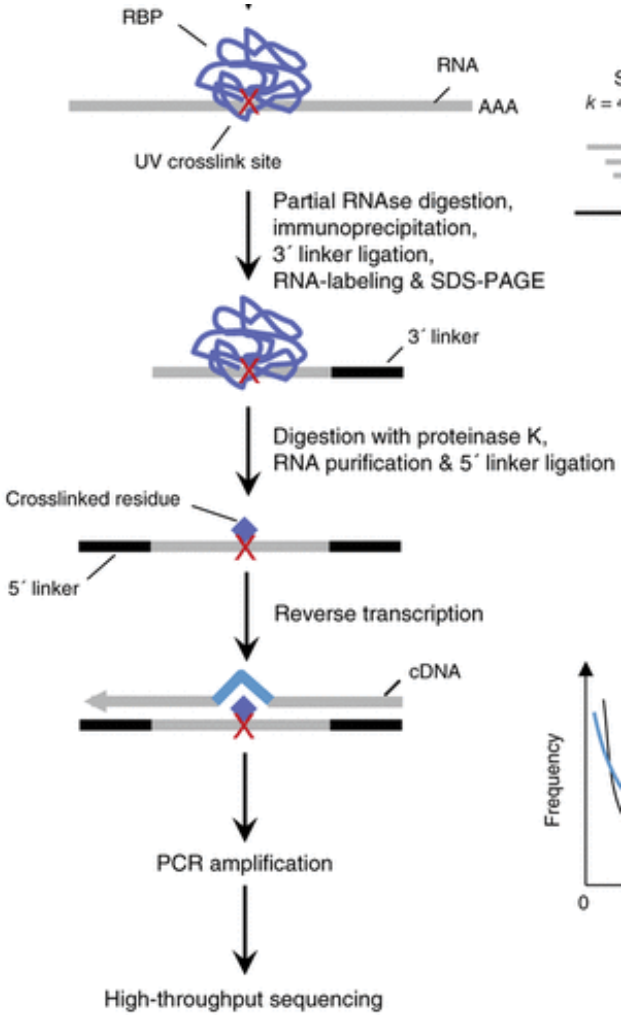
genome	Resolution(nt)	*transcriptome*
		
Dnase-seq	500-1,000	Protein Occupancy Profiling
		
ChIP-seq	100-250	CLIP-seq
ChIP-exo	20-50	i/PAR-CLIP



RNA binding proteins: From RIP-chip to CLIP-seq

- RIP: RNA immunoprecipitation (transcripts)
- CLIP: RNA cross-linking and IP (sites)

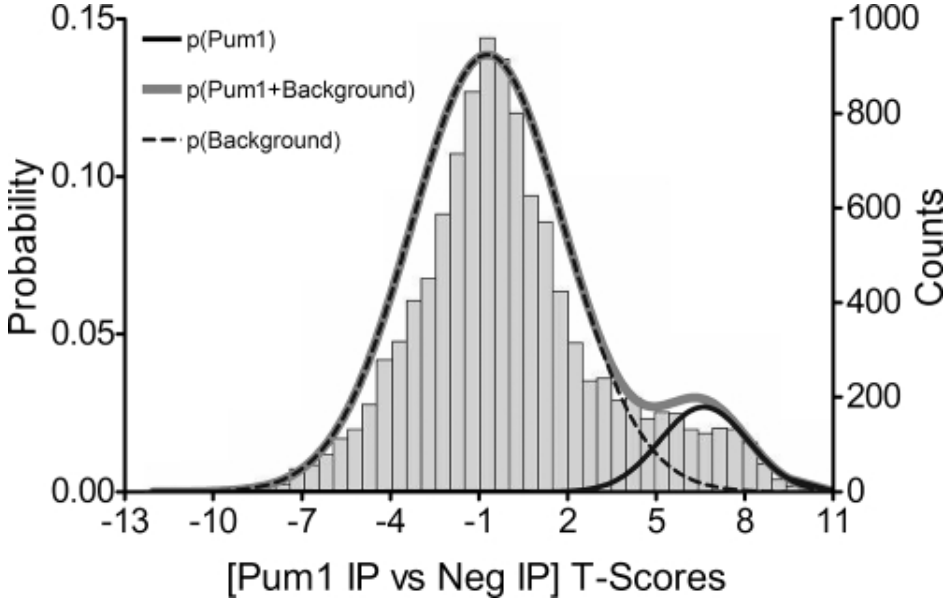


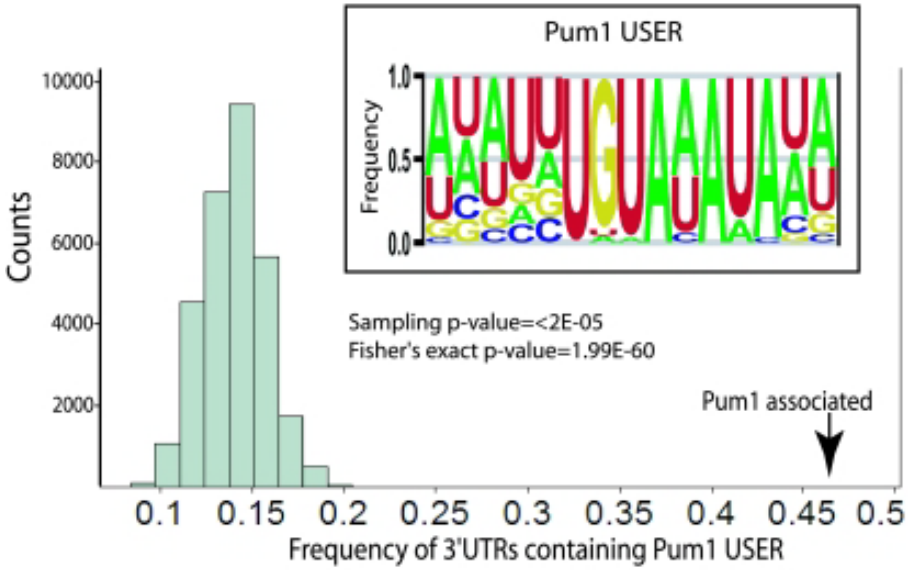


Keene, Darnell labs

Pum1 RIP-chip

- Pum1: Member of a very widely conserved RBP family (yeast to human)
  - Enhances decay, represses translation





Morris et al, MCB 20008

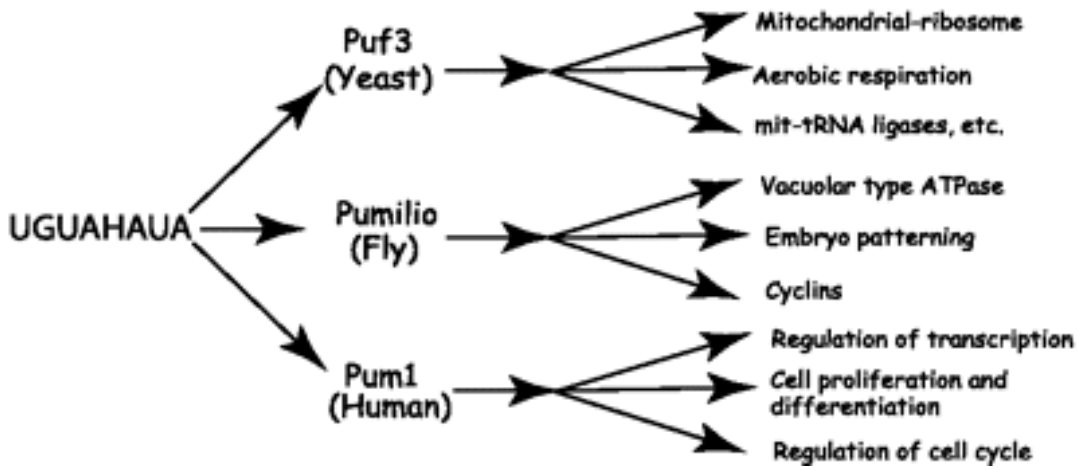
Changing targets across evolution

- Conserved protein, conserved motif, but different targets with different functions

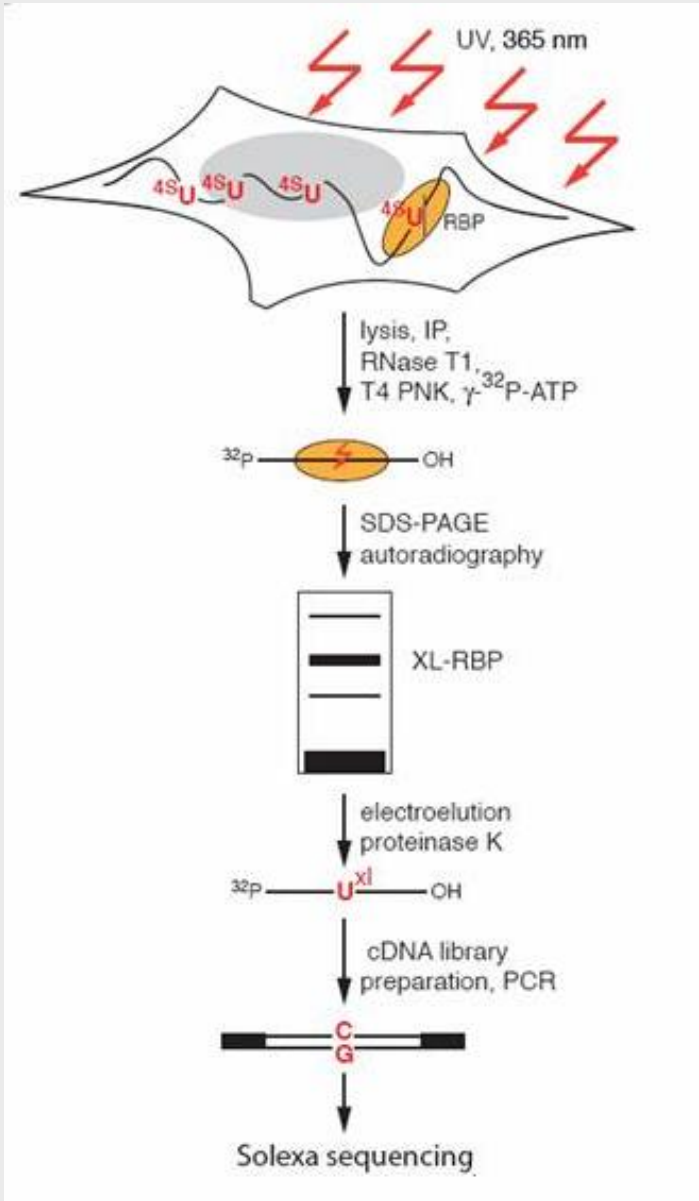
A.

	Genes with human orthologues	Pum1 target genes	Percent conserved target genes	P-value of enrichment (sampling)	P-value of enrichment (Fisher's exact)
Puf3 target genes	89	7	7.87%	0.83	0.79
Pumilio target genes	502	73	14.54%	0.018	0.036

B.



PAR-CLIP



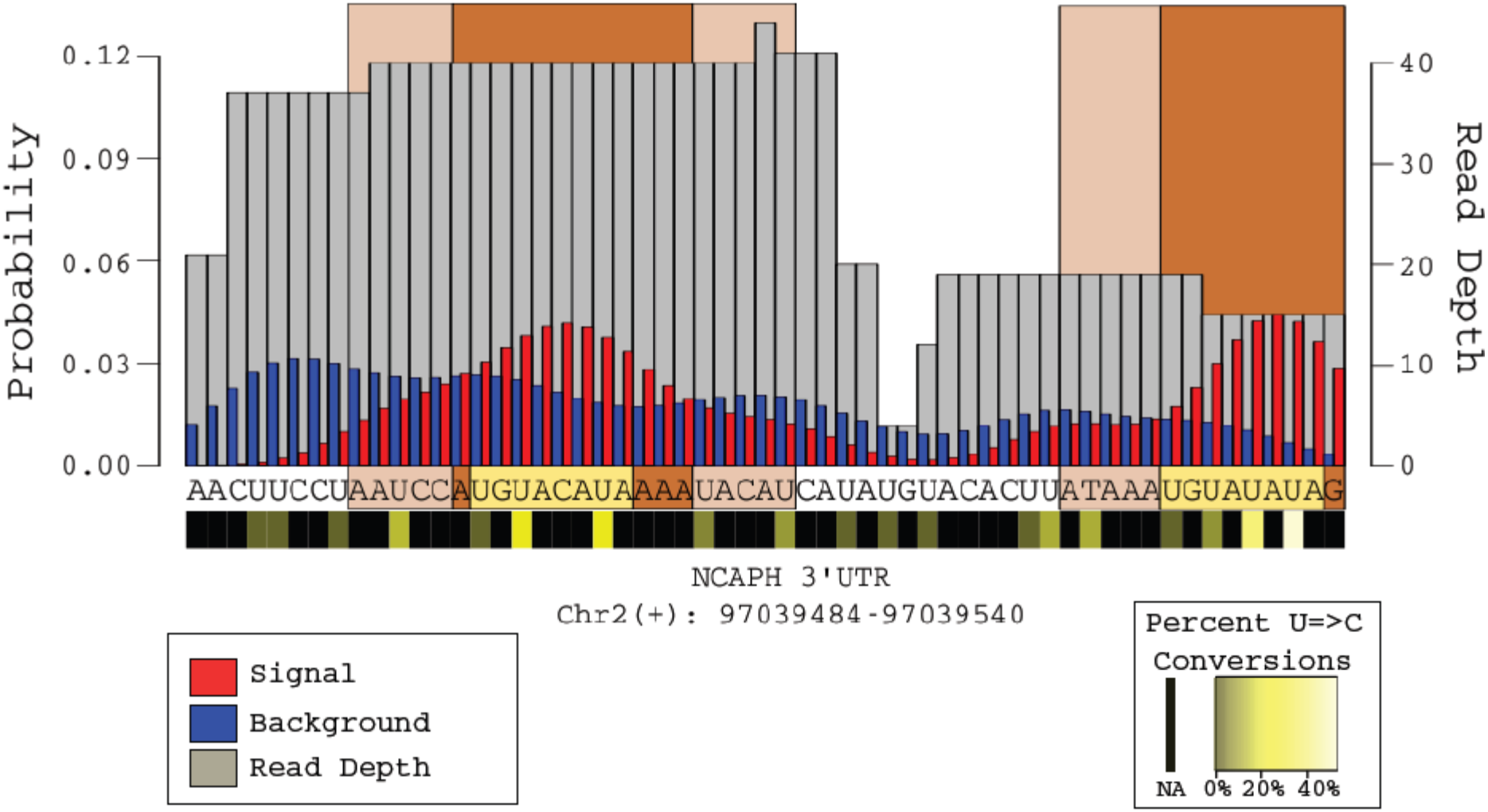
Identifies the locations of interaction between RNA-binding proteins (RBPs) and their mRNA targets in a high-throughput manner

- Think of it as a ChIP-Seq on an RNA level
- Presence of T->C conversions indicates crosslinked sites

Hafner *et al, Cell* 2010

PARalyzer: RBP site identification

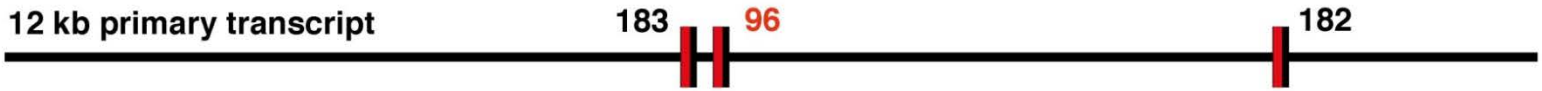
- Example: human 3' UTR bound by Pum2



Corcoran *et al, Genome Biol* 2011

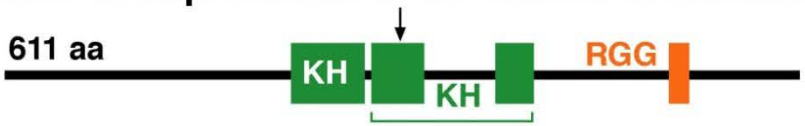
Genetic diseases caused by loss or mutation of miRNAs or RBPs

- Mutations in seed of miR-96 cause progressive hearing loss



- Monoallelic loss of *MIR17HG* causes Feingold syndrome

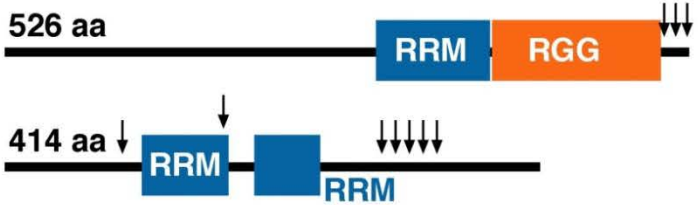
- Loss of expression or nonsense mutations in FMR1 cause Fragile X mental retardation



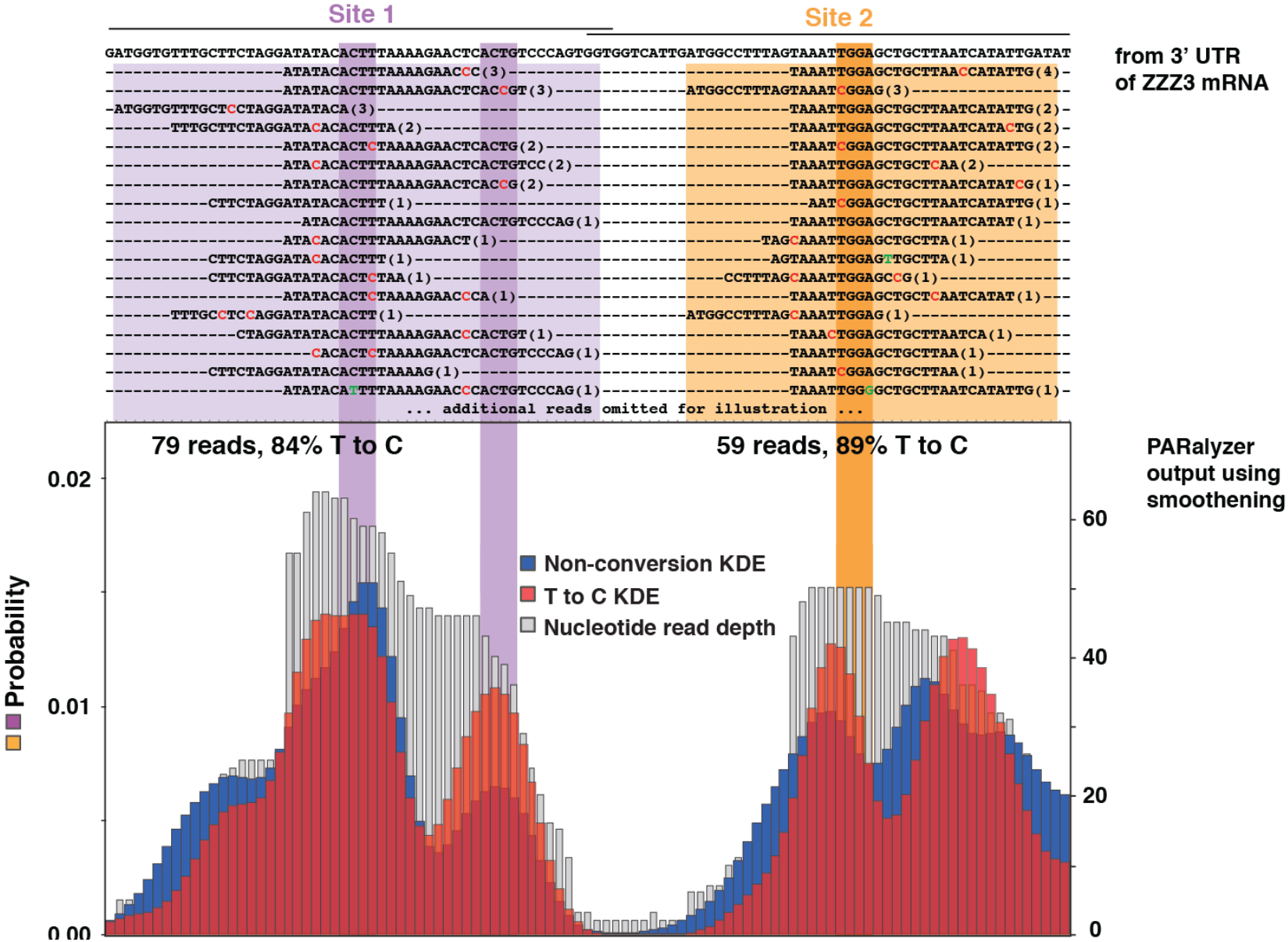
- Missense mutations in RBM20 underly dilated cardiomyopathy

- Nonsense and frameshift mutations in RBM10 cause a syndromic form of cleft palete

- Missense mutations in FUS and TDP43 cause amyotrophic lateral sclerosis



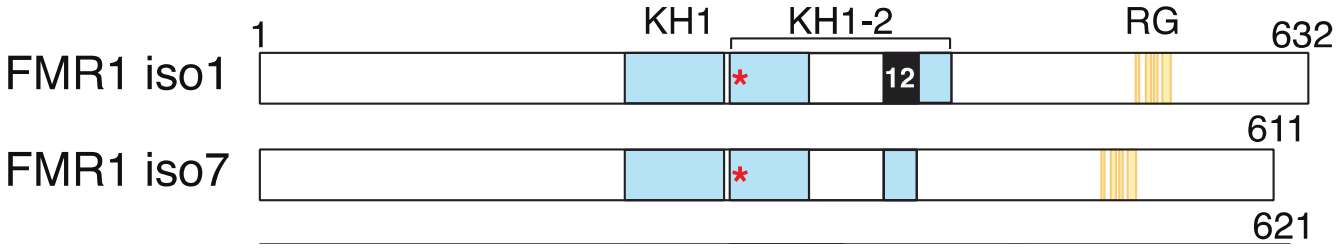
Example I: Deep sequencing delineates target sites and sequence preferences of FMR1



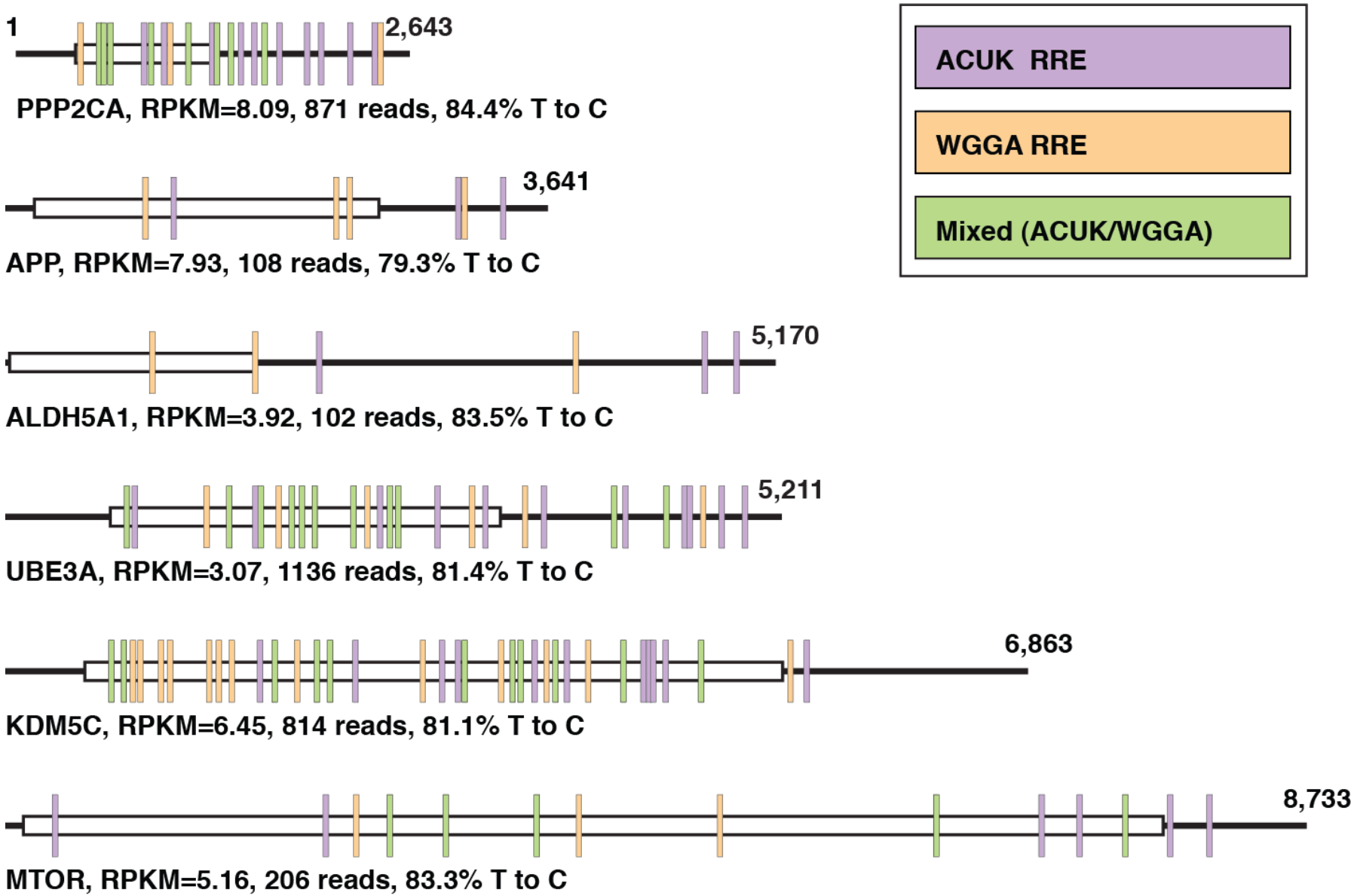
Ascano *et al, Nature* 2012

I304N mutation:
Lower Read Depth of ACUK Clusters

-CLIP
FMR1 iso7
wt vs. I304N

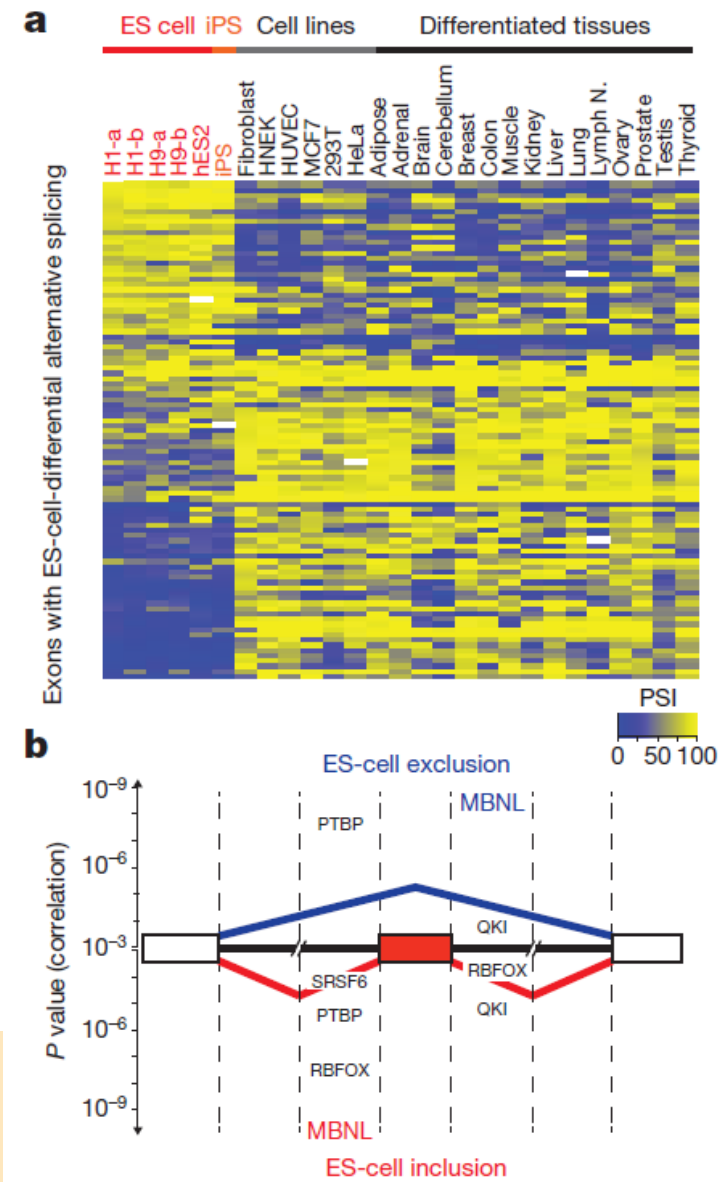


FMR1 binding sites on selected mRNAs

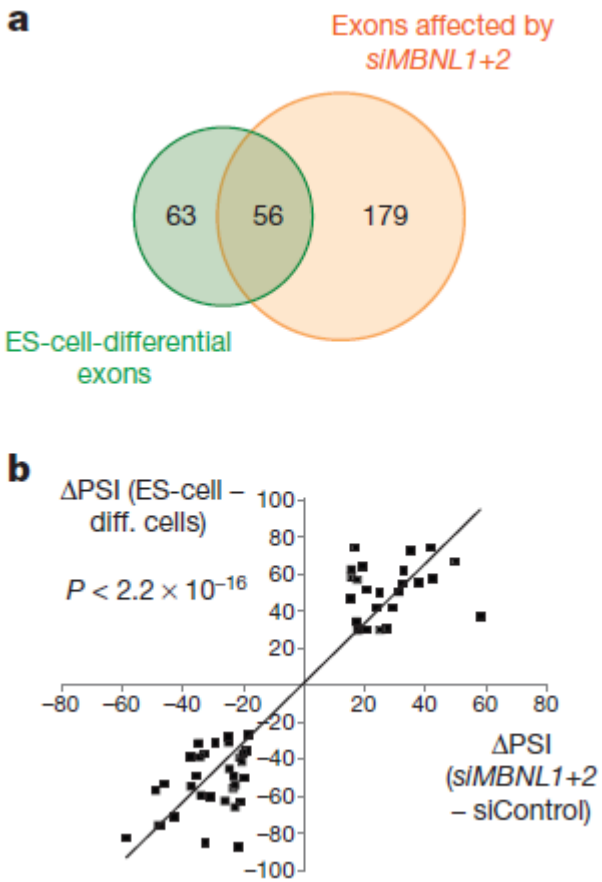


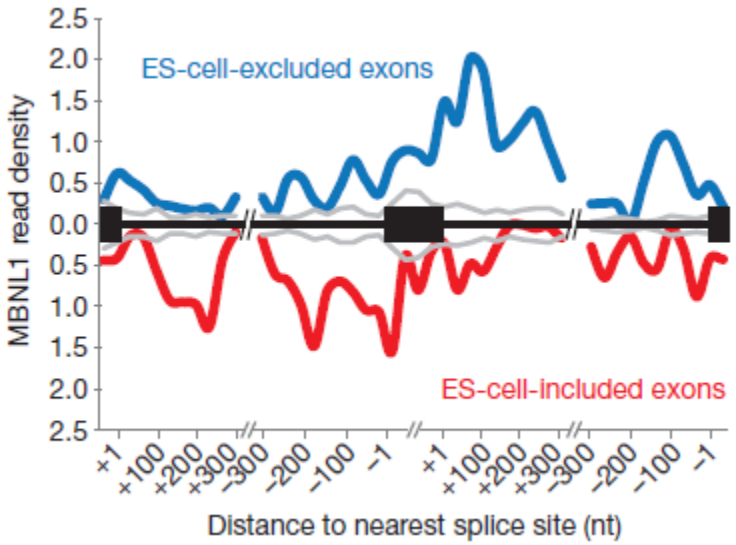
Embryonic stem cell differentiation and alternative splicing

- Muscle blind-like proteins (MBNL): conserved and direct regulators of exon skipping/inclusion
- Overexpression in ES cells promotes AS patterns of differentiated cells
- One of the targets: FOXP1
  - Switched exon spans TF binding domain, changes affinity in ES cells
  - ES cell splicing directs it to express pluripotency factors



Linking splicing changes and RBP binding





Han *et al, Nature* 2013

Take home points: next-gen sequencing is…

- Unbiased
  - Not limited to classical coding genes
  - Not limited to gene proximal regions
- High resolution
  - Expression on the level of isoforms
  - Interactions up to single nucleotide precision
- Cheap
  - Whole genome now < $1,000