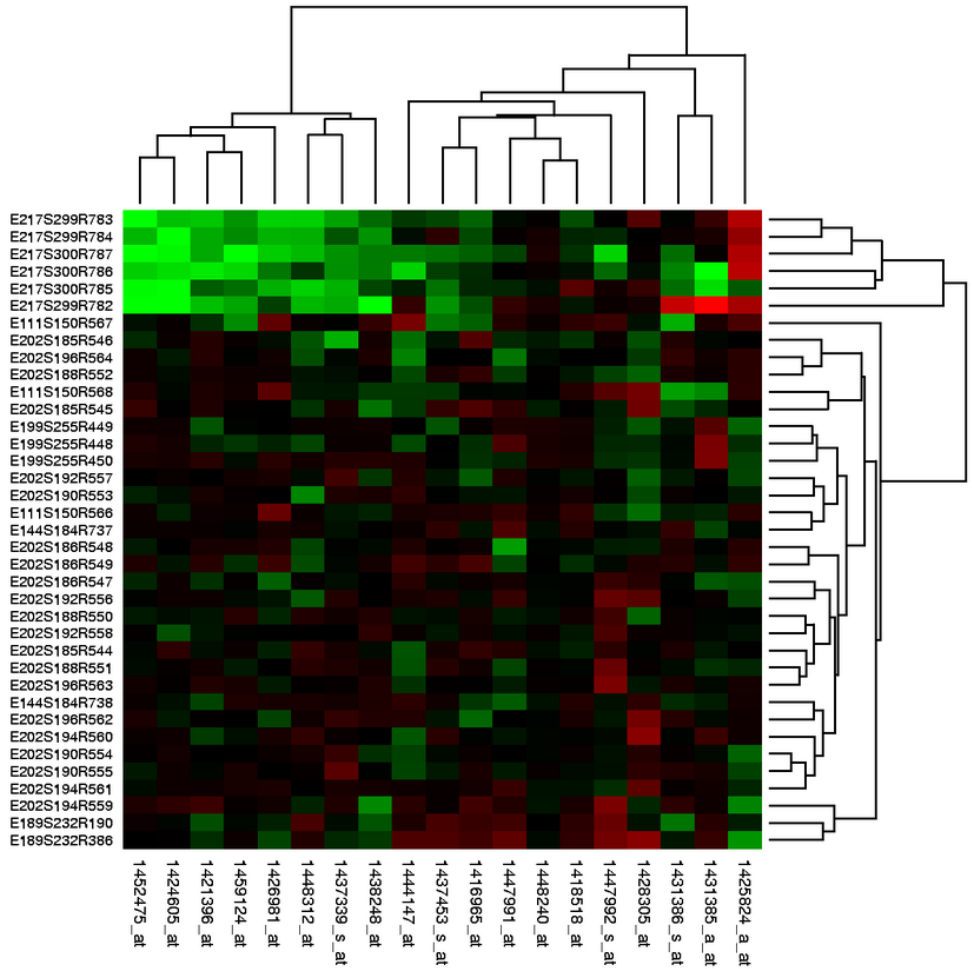
Leibniz Institute on Aging –
Fritz Lipmann Institute

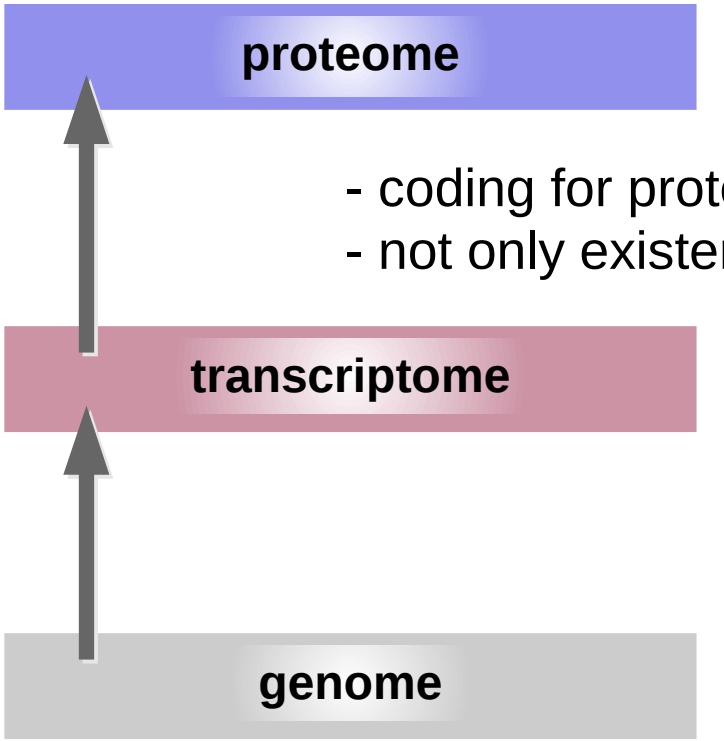
# Analytical Biochemistry




# Transcriptome Analysis

Karol Szafranski,
Genome Analysis Group (M. Platzer)

# Transcriptome – definition

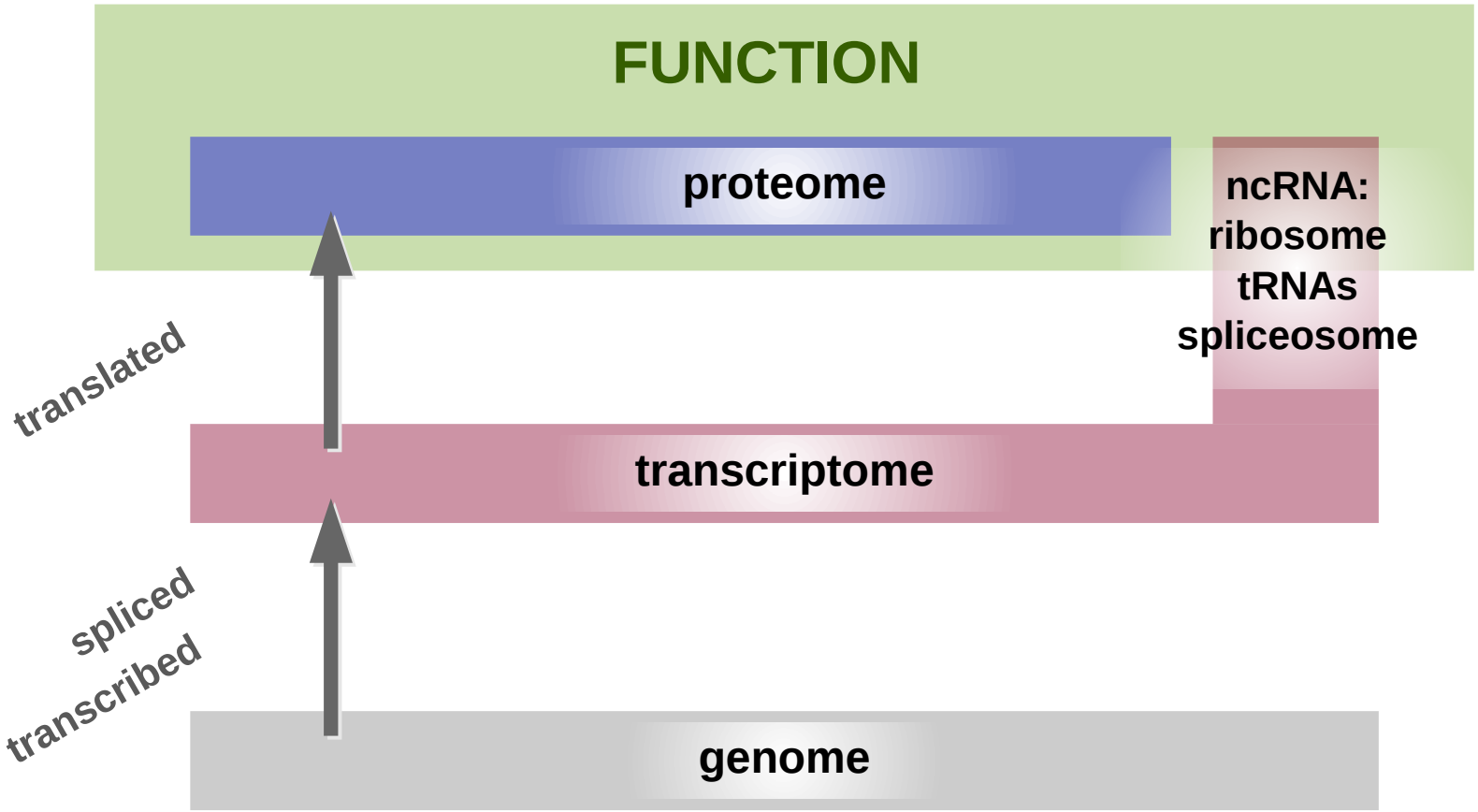



- coding for proteins (or functioning as such)
- not only existence, but **copy number matters!**

- RNA content of a cell
  (or a cell fraction, or a tissue)

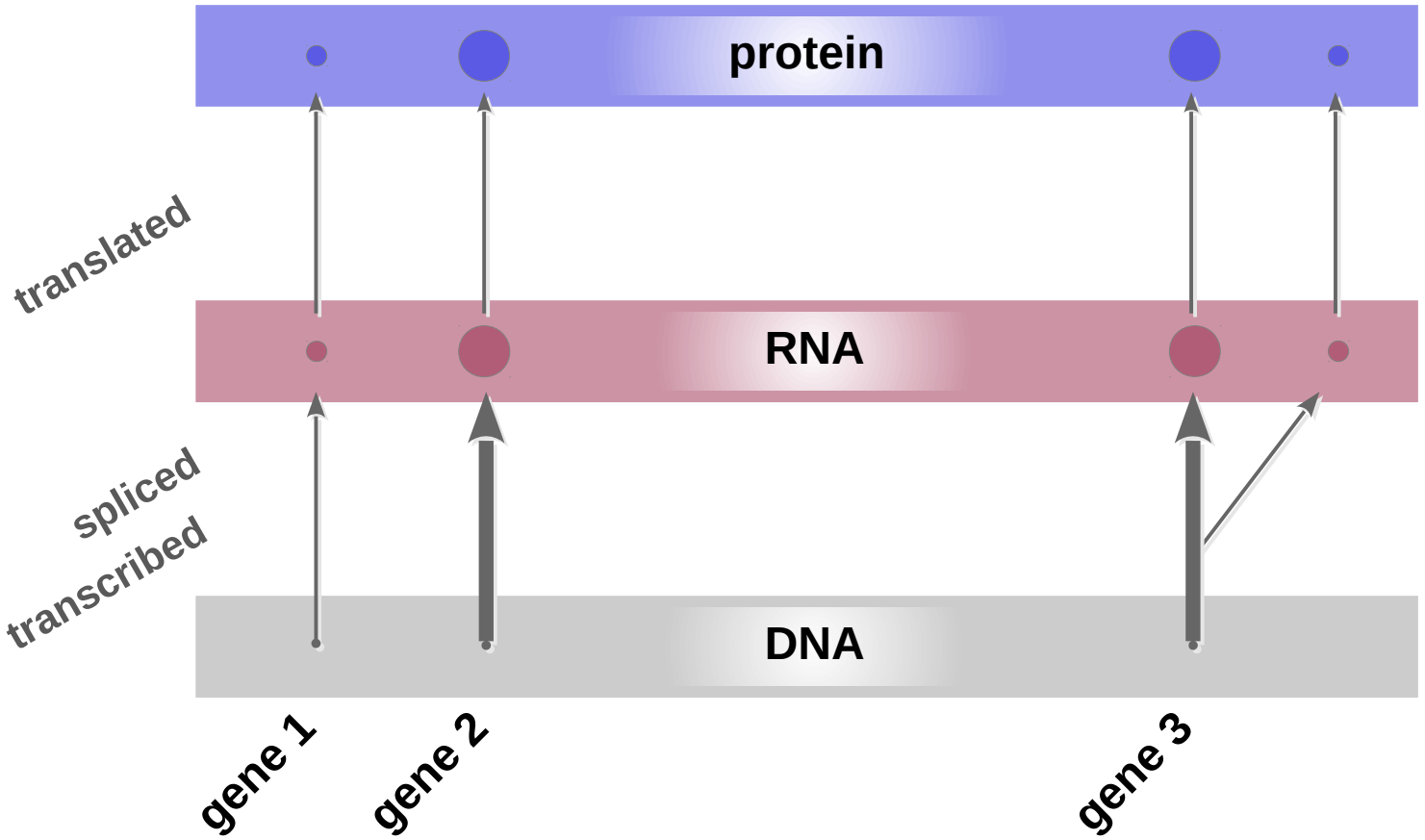
- DNA content of a cell
- genetic information storage

# Molecular levels of life



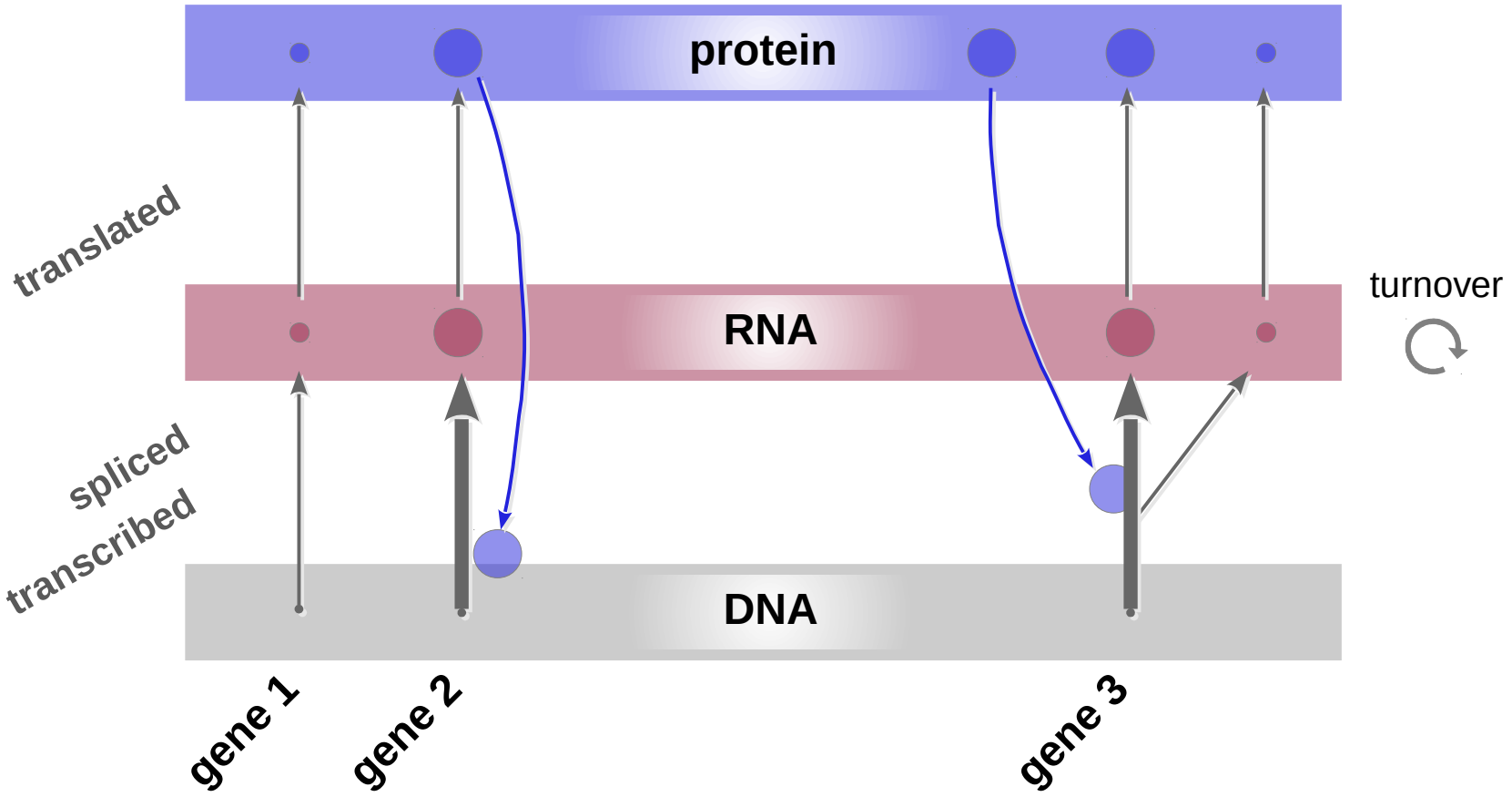
**mRNA levels** are regulated in

- time (development, environment, age)
- space (tissue-specific)



**mRNA levels** are regulated in

- time (development, environment, age)
- space (tissue-specific)



# Transcriptome – what to learn from it?

## 1. Eukaryote gene structure

- exon/intron structure
- 5' end => promoter
- 3' end

- isoforms : alternative gene structures

## 2. mRNA expression levels

- RNA-seq, tag counting

## 3. mRNA processing steps / regulatory mechanisms

- identify/quantify binding factors:
  splicing regulators, surveillance factors

- A-to-I editing

## 4. protein translation

## Classical genetics

- identify functional mutant in a model system, e.g. in yeast

- identify/clone the gene through …
  - rescue experiments => rescue clone
  - tracking the insertion mutation site => tracked locus

## Identify human homolog : Era I (80ies to early 90ies)

- screen human phage library via hybridization
  and sequence positives

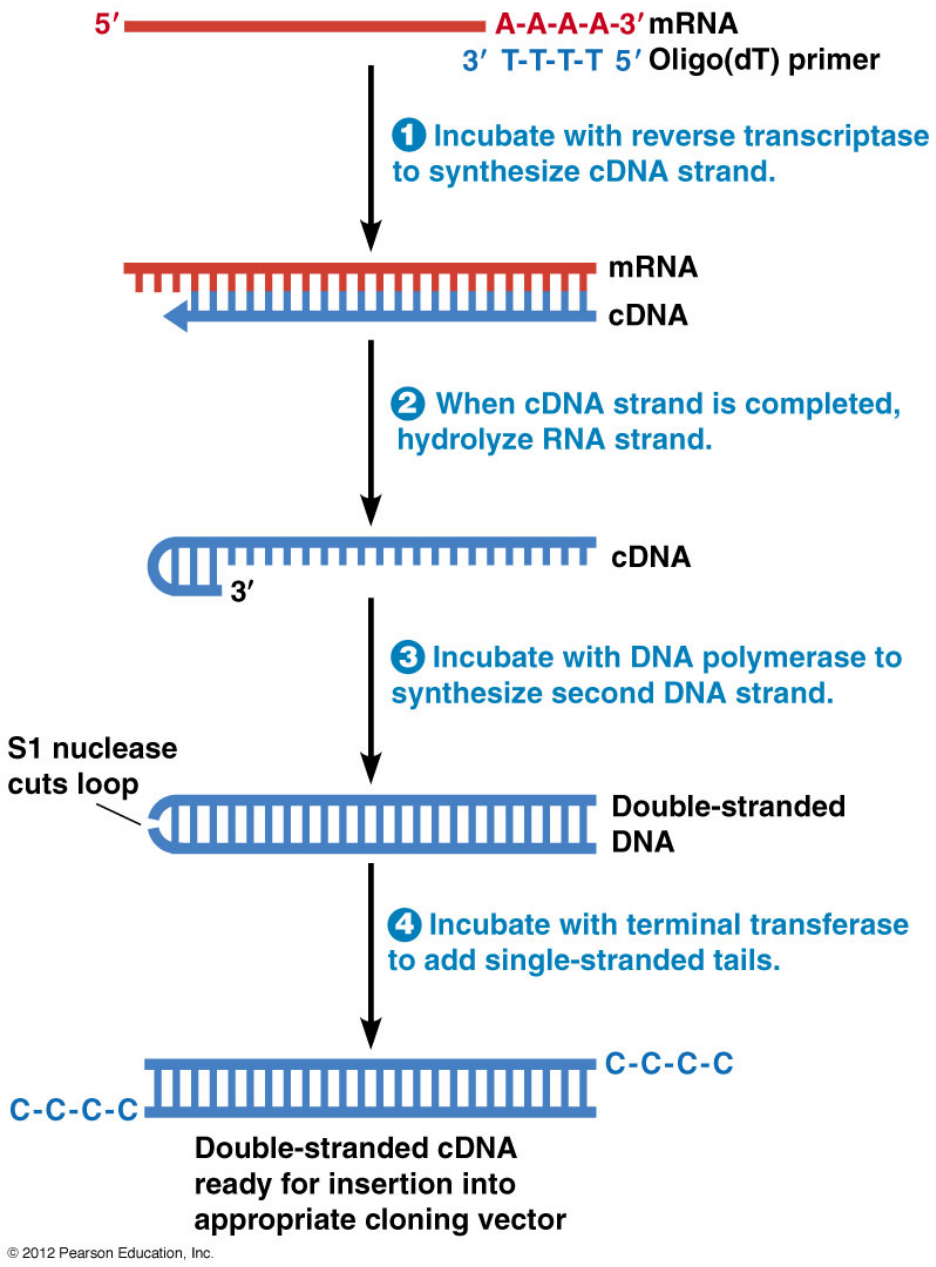
**=> a human gene (CDS) with attributed function**

**...**

**...**

**=> a human gene resolved in the genome**

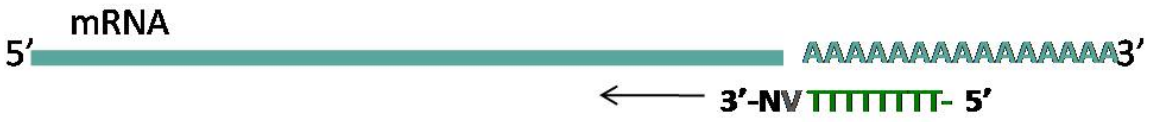
# Reverse transcriptase - cDNA



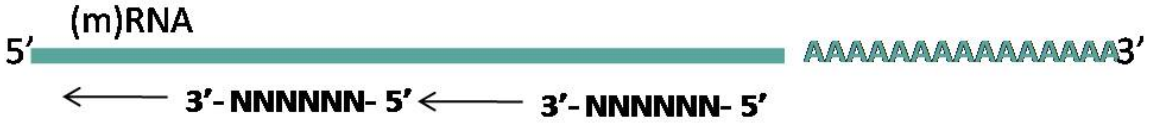
# Reverse transcriptase priming

**A) Oligo (dT) primer**


**B) Anchored oligo (dT) primer**



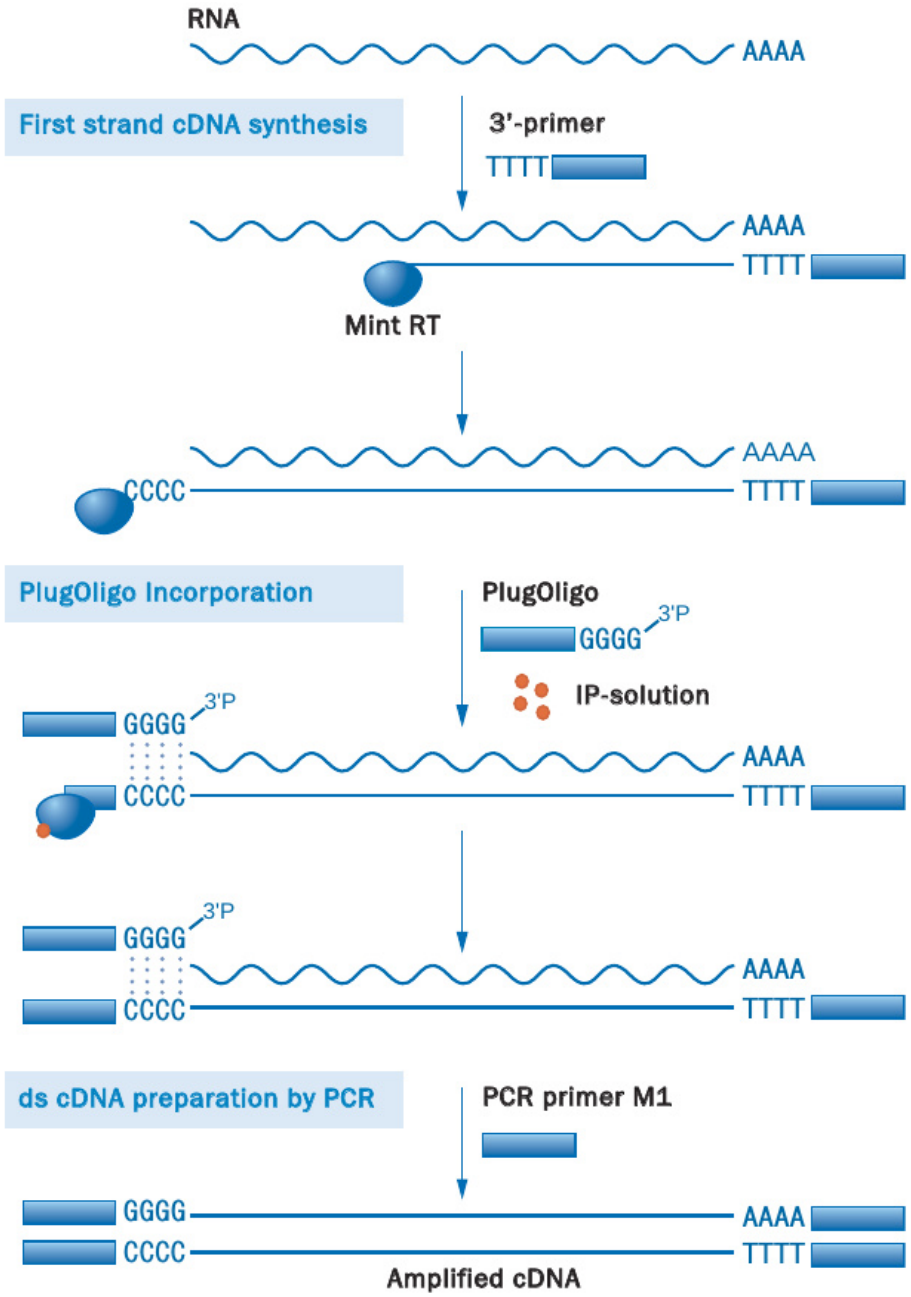

**C) Random hexamer primers**




**D) Gene-specific primer**


**N = any base**
**V = A, C, or G**

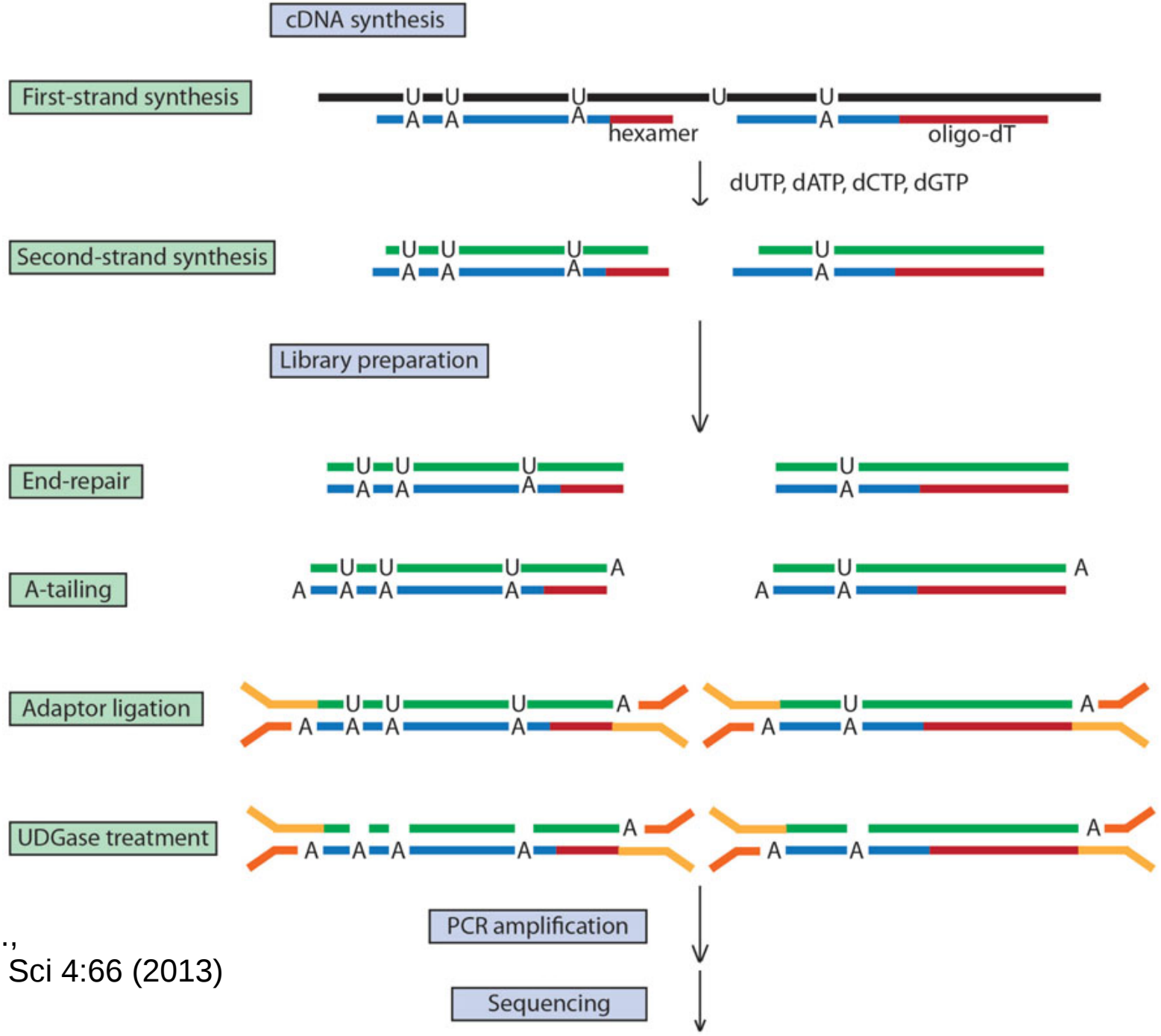
# cDNA strand-oriented I – SMART protocol




Evrogen Inc.: Mint

Takara Inc.: SMART

# cDNA strand-oriented II – dUTP protocol




Martin et al.,
Front Plant Sci 4:66 (2013)

## **Identify human homolog : Era II** (mid-90ies to now)

- human genome **resource**

- cDNA library **resources**,
  sequenced at random

- identified gene loci by autom. mapping of cDNA sequence to genome

- refined gene loci using CDS prediction programs

**=> ~all human genes at high resolution**
**http://genome-euro.ucsc.edu**

research task to find target protein:
(eventually, via sequence tag)

- explore sequence similarity between genomes

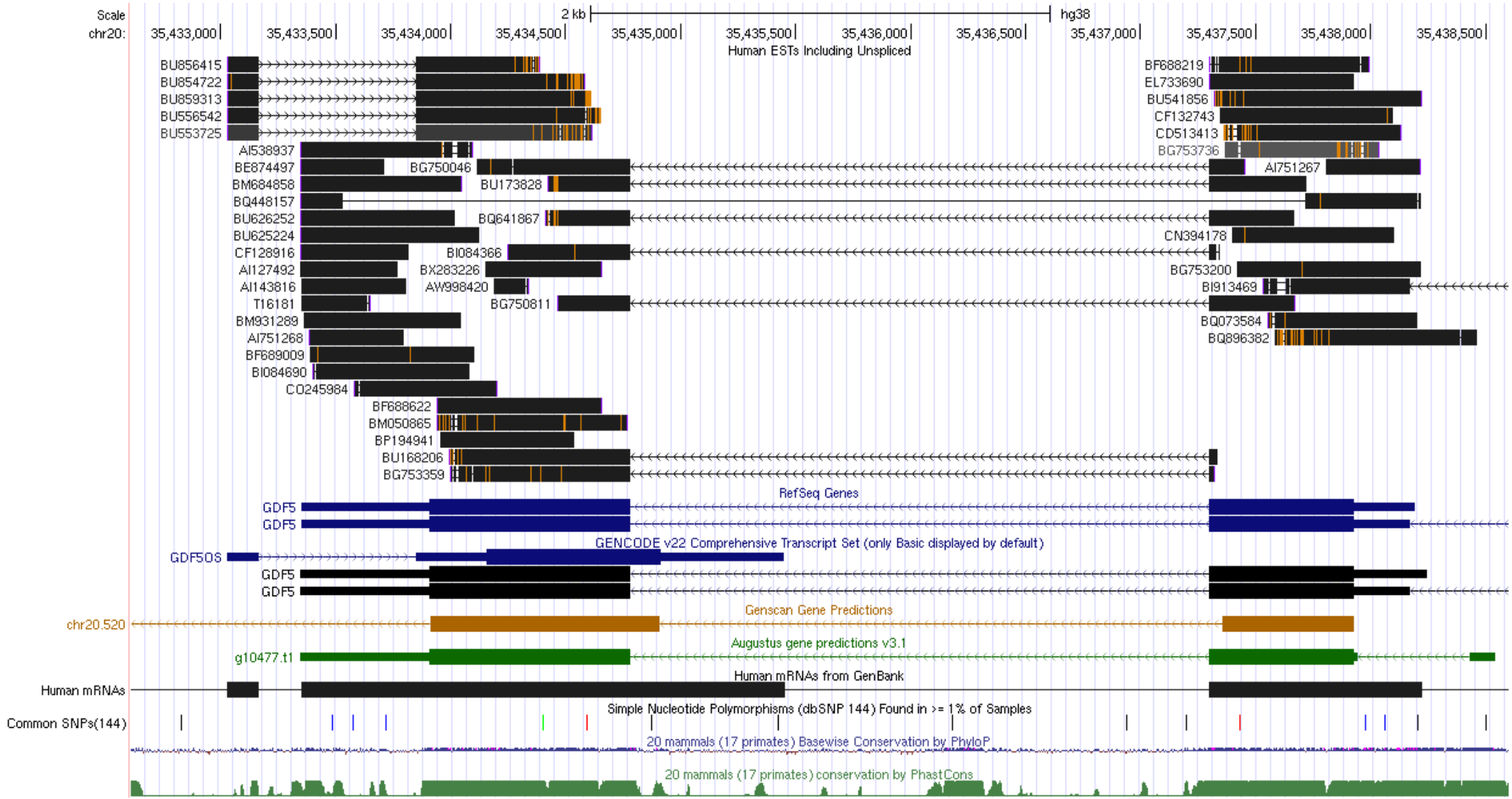
## Mapping/alignment of cDNA sequence to genome

program EXALIN, Zhang & Gish, Bioinformatics 22:13 (2006)

```
T:  GCGCCCGCTGGCGGCGTTGCTTCCGCTAACACAGACGAAGCTGCAGACGCCGAGAGCGGG   540
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
G:  GCGCCCGCTGGCGGCGTTGCTTCCGCTAACACAGACGAAGCTGCAGACGCCGAGAGCGGG   1587

T:  ATCCGCAGTCGGCAGTTGCAGCAGCTCATCTCTTTTTCCATGGTTTGGGCTAACCATGTC   600
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
G:  ATCCGCAGTCGGCAGTTGCAGCAGCTCATCTCTTTTTCCATGGTTTGGGCTAACCATGTC   1647

T:  TTCCTTGGGTGCCG..............GTACTCTCAAAAAGTTATGGATAAAATTCTT   645
    |||||||||||||>>>>>>...>>>>>>|||||||||||||||||||||||||||||||
G:  TTCCTTGGGTGCCGGTGAGT...TTTTAGGTACTCTCAAAAAGTTATGGATAAAATTCTT   2170

T:  AGTATGGCTGAAGGCATCAAAGTGACAGATGCTCCAATCCATACAACAAGAGACGAACTG   705
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
G:  AGTATGGCTGAAGGCATCAAAGTGACAGATGCTCCAATCCATACAACAAGAGACGAACTG   2230

T:  GTTGCCAAGGTGAAGAAAAGAGGGATATCGAGTAGCAATG..............AAGGG   750
    ||||||||||||||||||||||||||||||||||||>>>>>>...>>>>>>|||||
G:  GTTGCCAAGGTGAAGAAAAGAGGGATATCGAGTAGCAATGGTTAGC...TTTTAGAAGGG   2670

T:  GTAGAAGAGCCATCCAAAAAACGAATTGTAGAAGGAAAAAACAATTCTGCAGTTGAGCGA   810
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
G:  GTAGAAGAGCCATCCAAAAAACGAATTGTAGAAGGAAAAAACAATTCTGCAGTTGAGCGA   2730

G: genome
T: transcript
```

# Elucidation of eukaryote gene structure

## Example: human growth differentiation factor 5 (*GDF5*) locus



## Identify human homolog : Era II (mid-90ies to now)

- human genome **resource**

- cDNA library **resources**,
  sequenced at random

- identified gene loci by autom. mapping of cDNA sequence to genome

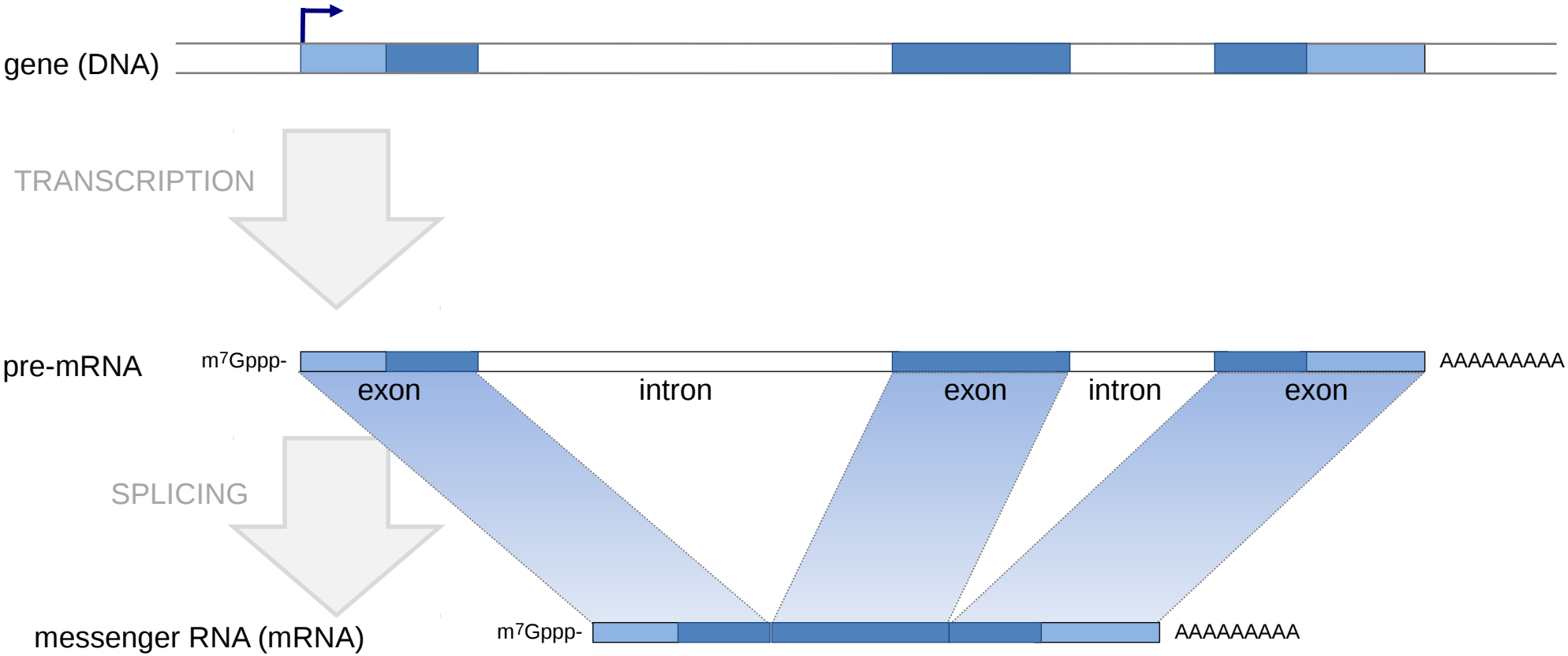
- refined gene loci using CDS prediction programs

**=> ~all human genes at high resolution**
http://genome-euro.ucsc.edu

result for target protein:

**=> human gene at high resolution:**
**- exon/intron structure (possibly alternative)**
**- promoter (possibly alternative)**

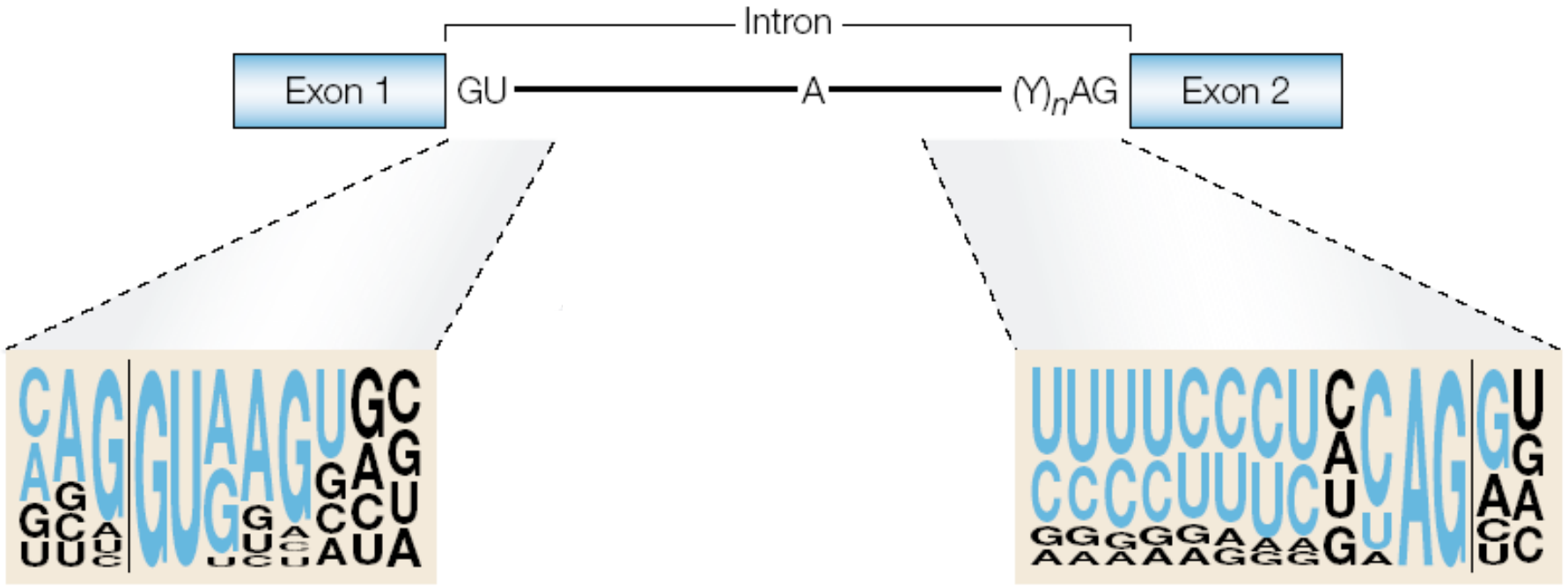
# Elucidation of eukaryote gene structure




**Genes are fragmented in eukaryote genomes**

# GU-AG rule

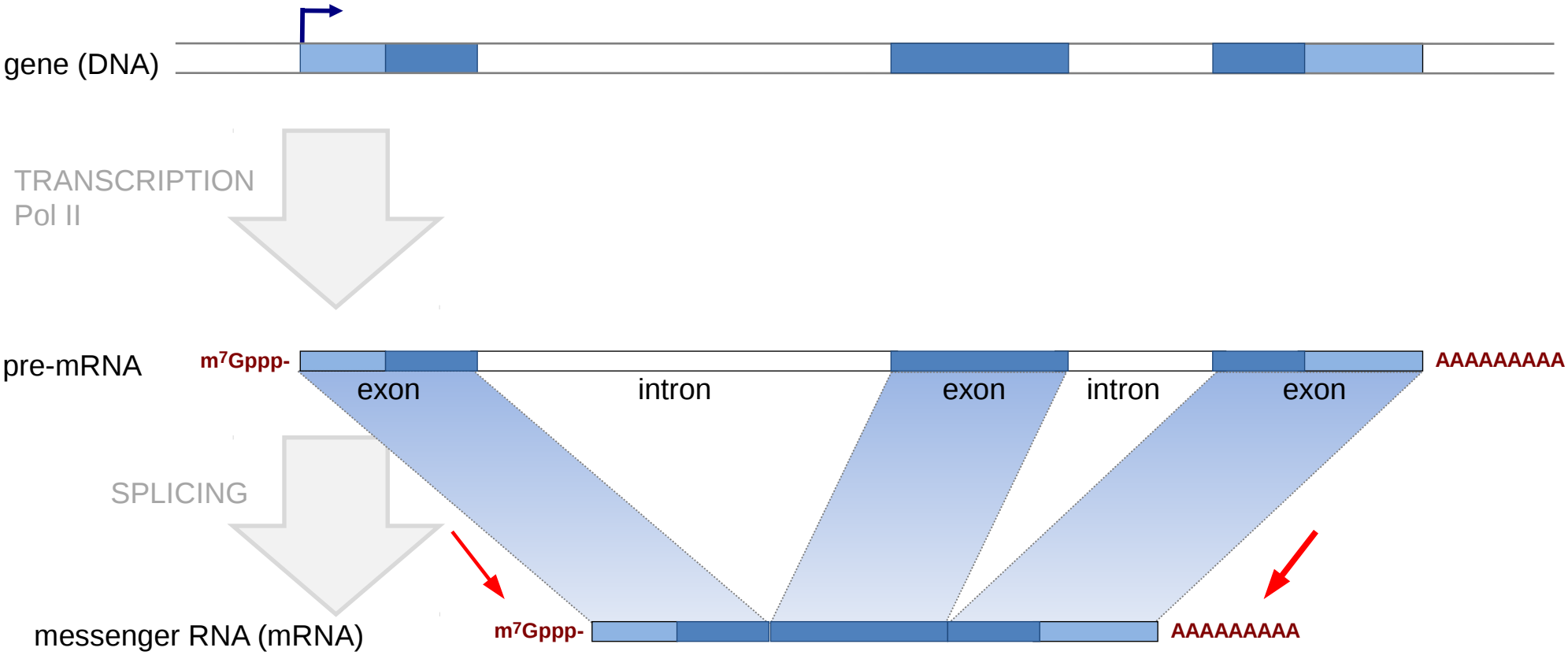
Breathnach & Chambon, PNAS (1978); Crick, Science (1979)




splice donor

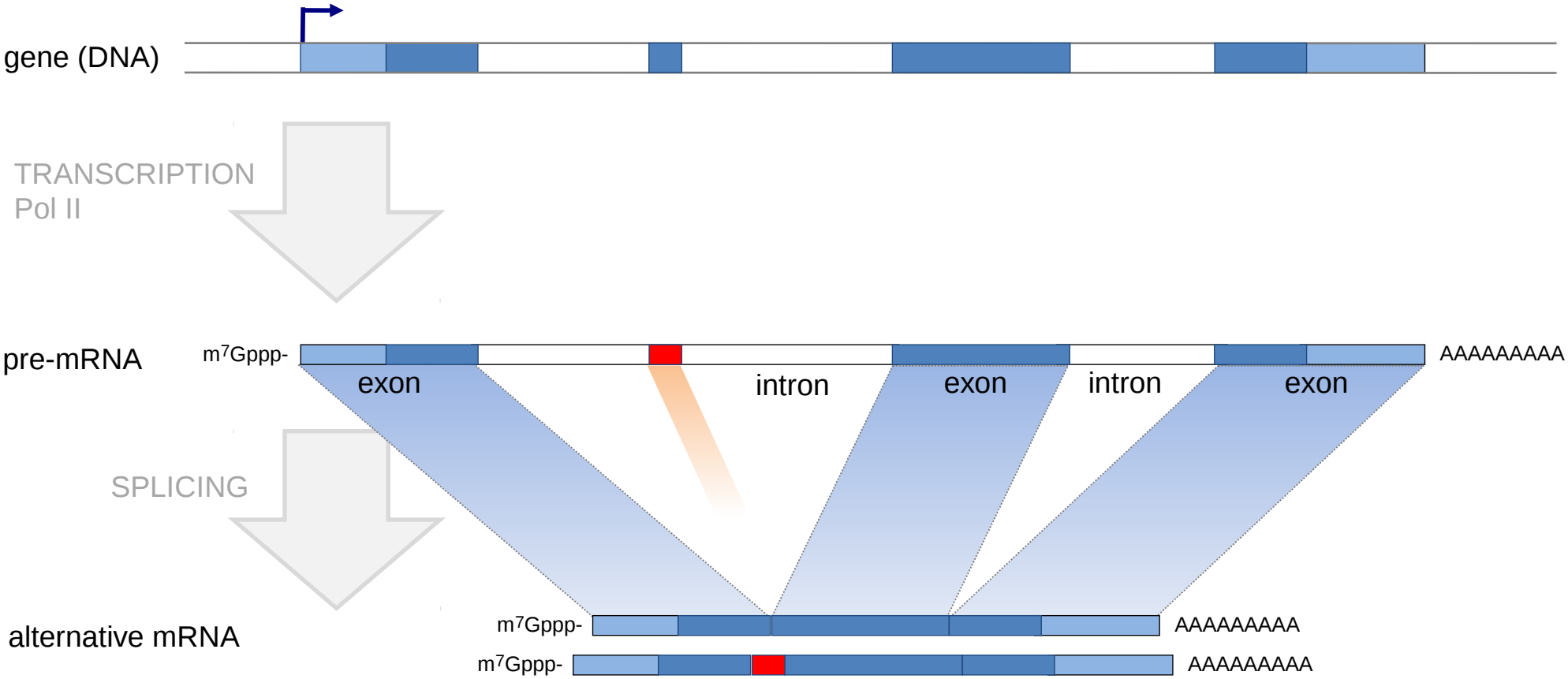
splice acceptor

# Elucidation of eukaryote gene structure

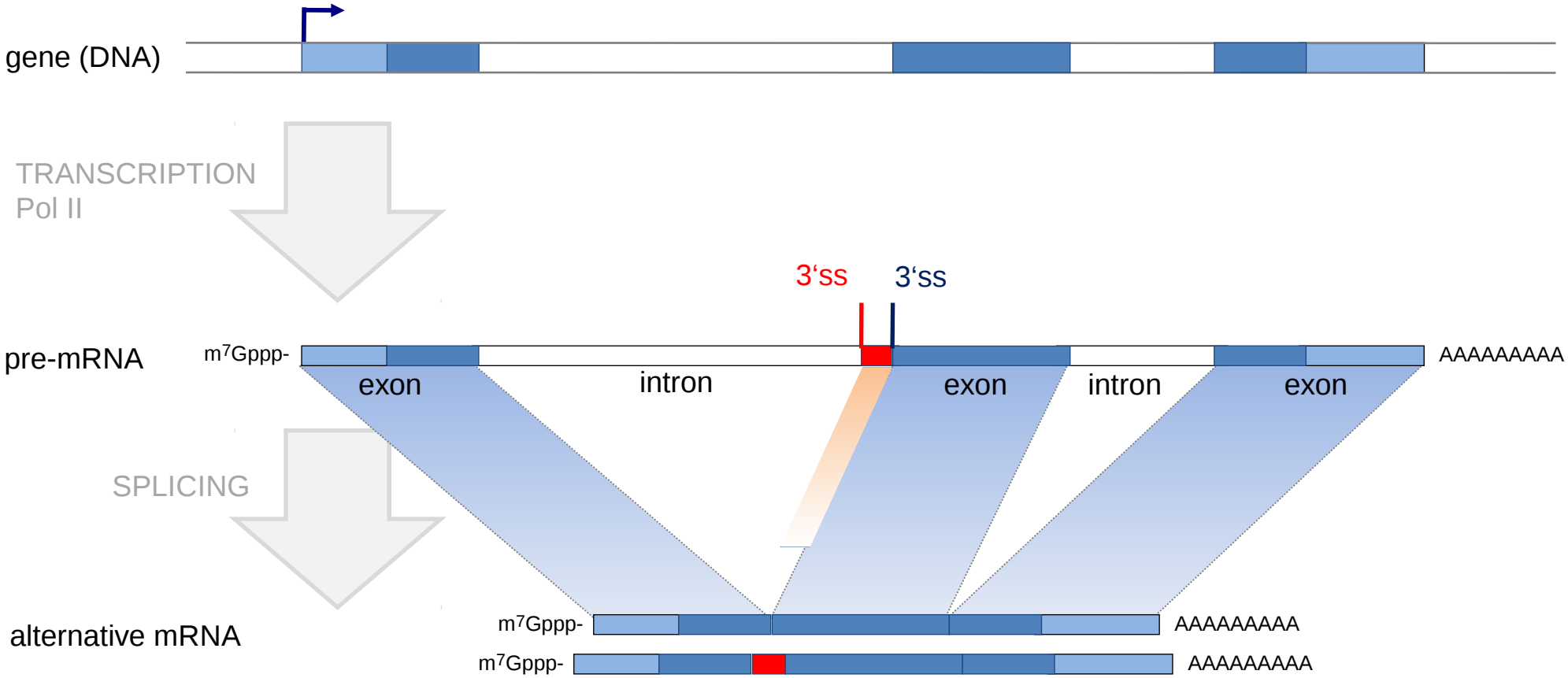



**Characteristic modifications of mRNA molecules**

# Splicing – alternative splicing



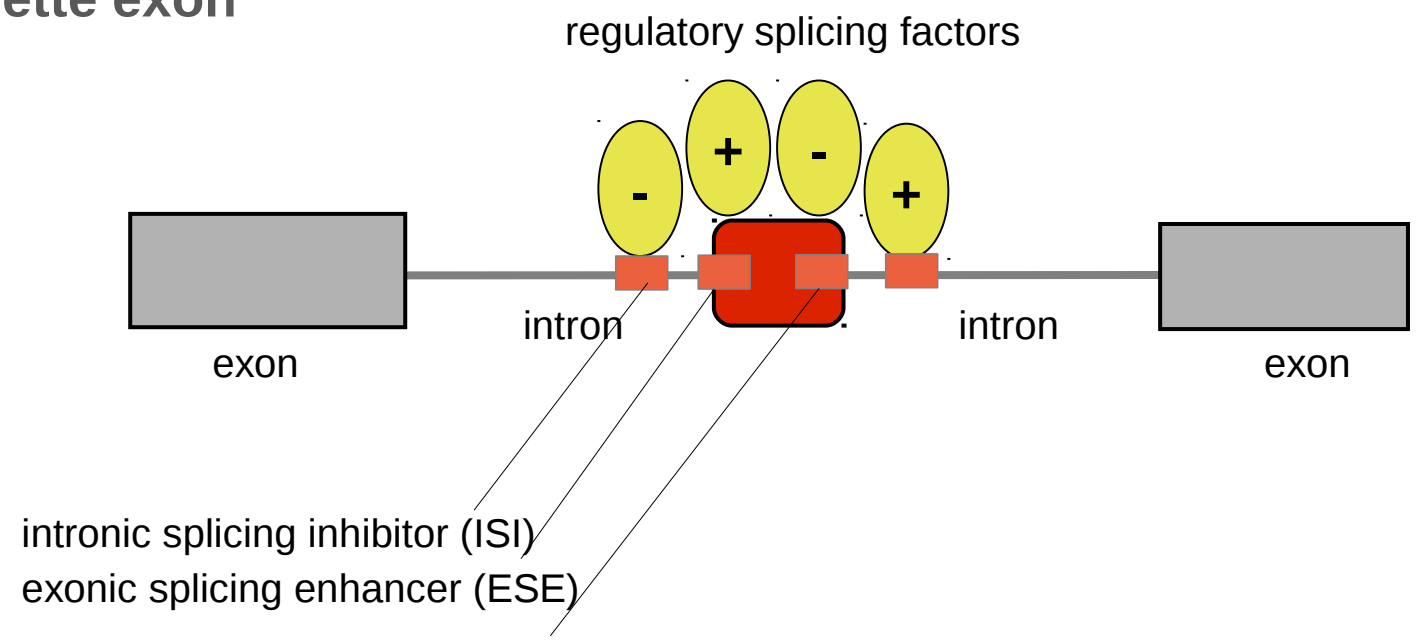
# Splicing – alternative splicing




Hiller et al. 2004, Nat Genet 36:1255

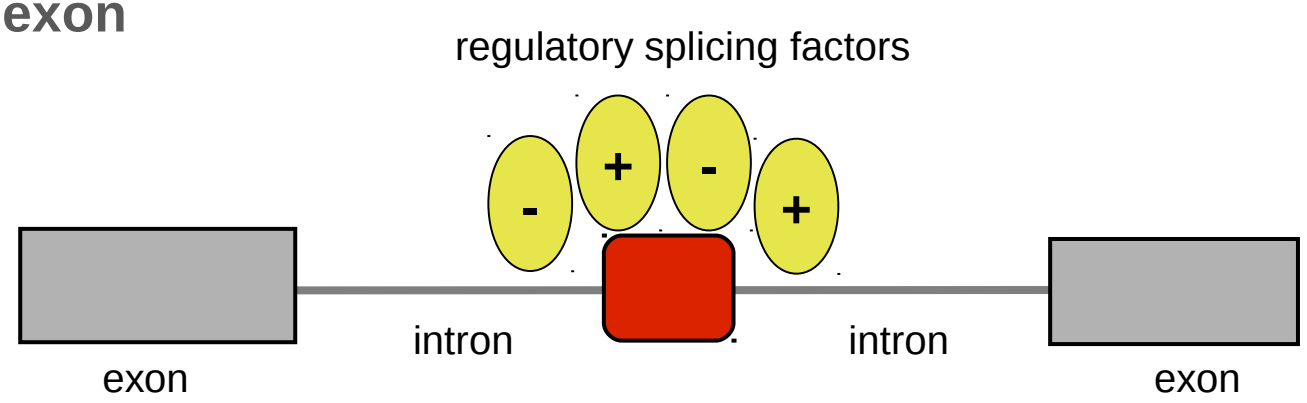
# Splicing – alternative splicing

## Cassette exon

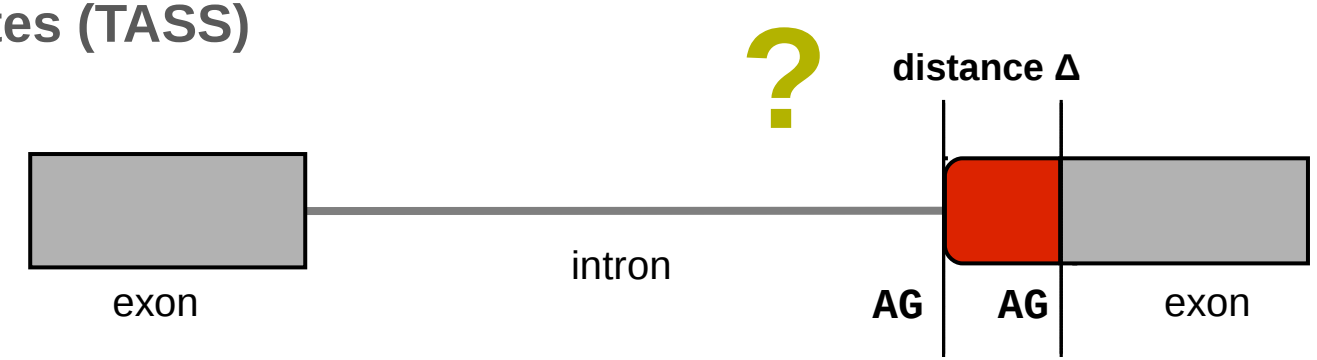


# Splicing – alternative splicing

## Cassette exon



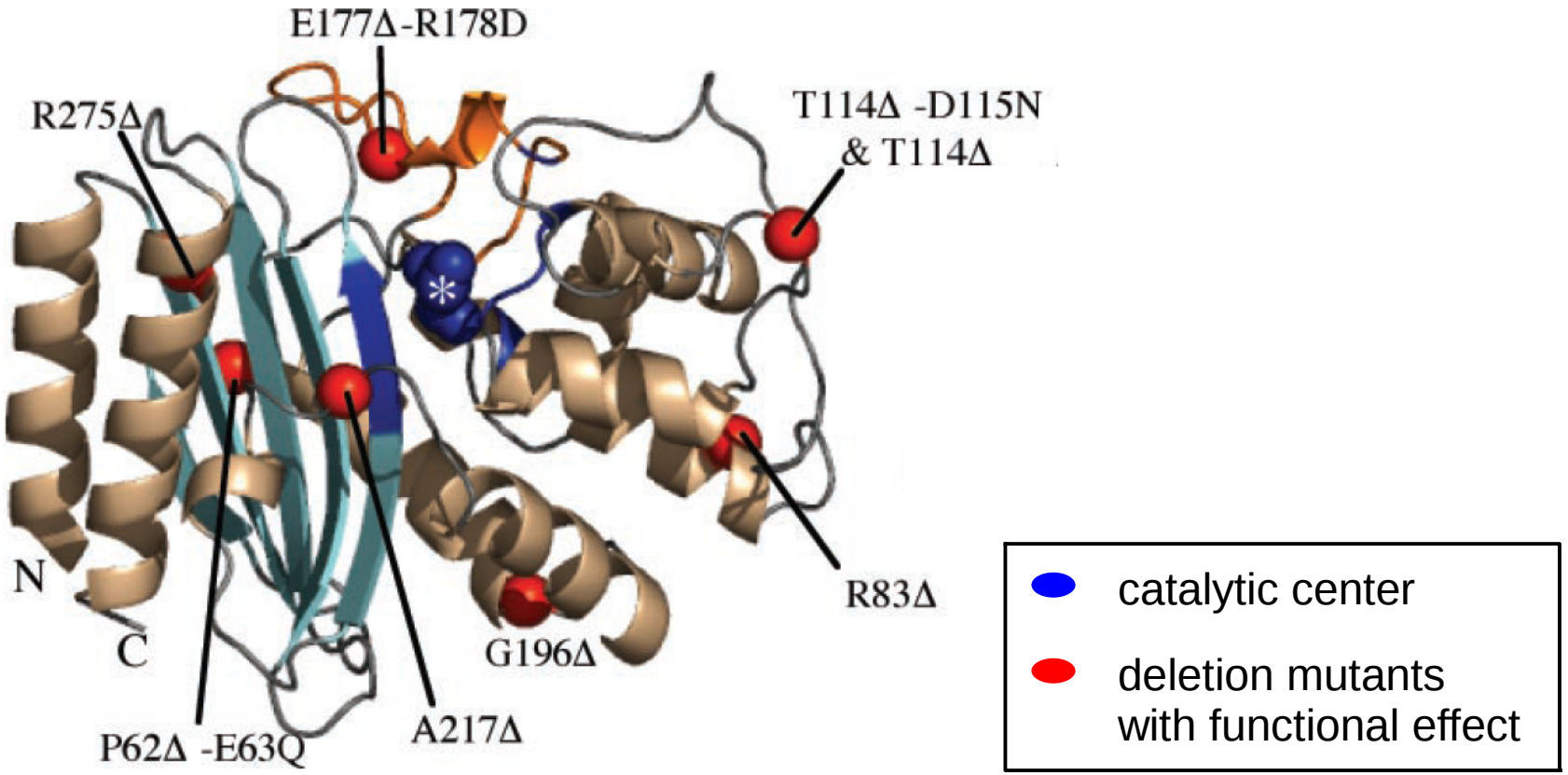

## Tandem alternative splice sites (TASS)



## Do small differences actually have an effect?

Random codon deletion in β-lactamase (TEM-1)

Jones, NAR 33:e80 (2005)




**=> Effect on catalytic activity in 14 of 22 mutants (64%)**

# Transcriptome – what to learn from it?

## 1. Eukaryote gene structure

- exon/intron structure
- 5' end => promoter
- 3' end

- isoforms : alternative gene structures

## 2. mRNA expression levels

- RNA-seq, tag counting

## 3. mRNA processing steps / regulatory mechanisms

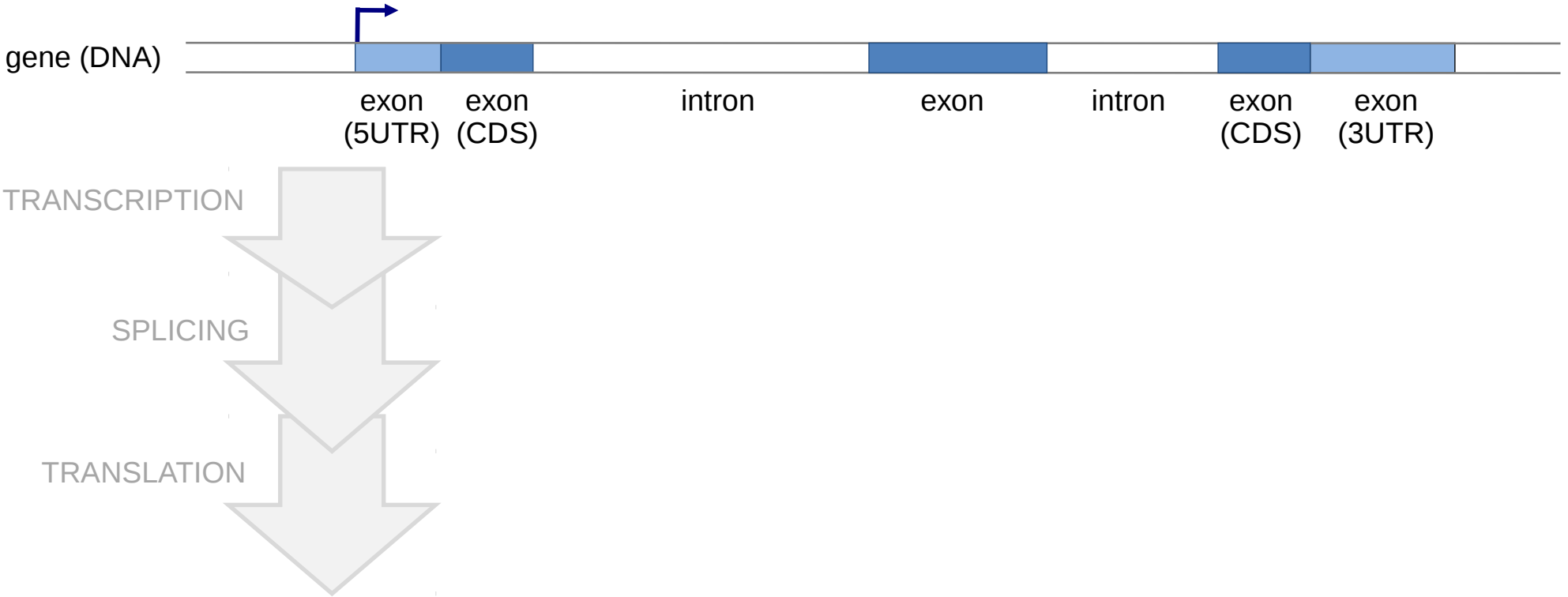
- identify/quantify binding factors:
  splicing regulators, surveillance factors

- A-to-I editing

## 4. protein translation

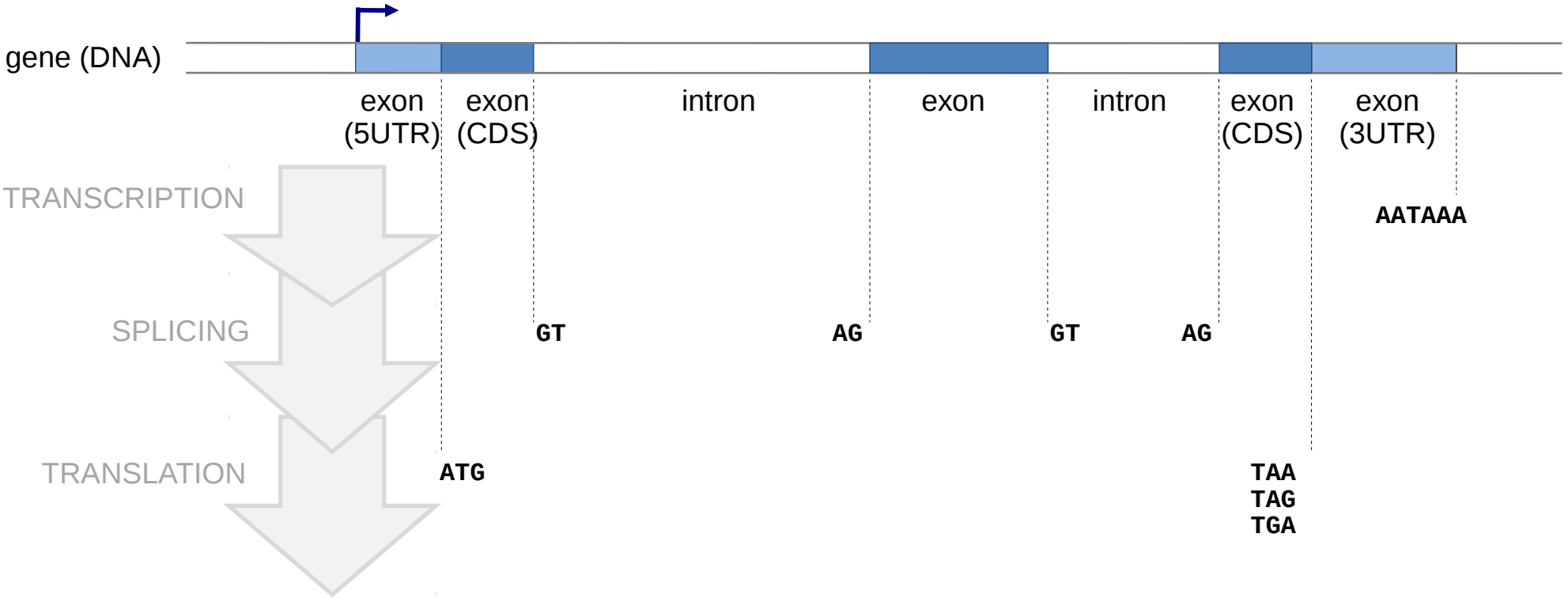
# *Ab initio* gene prediction

search genome for signs; do **not** use homology information



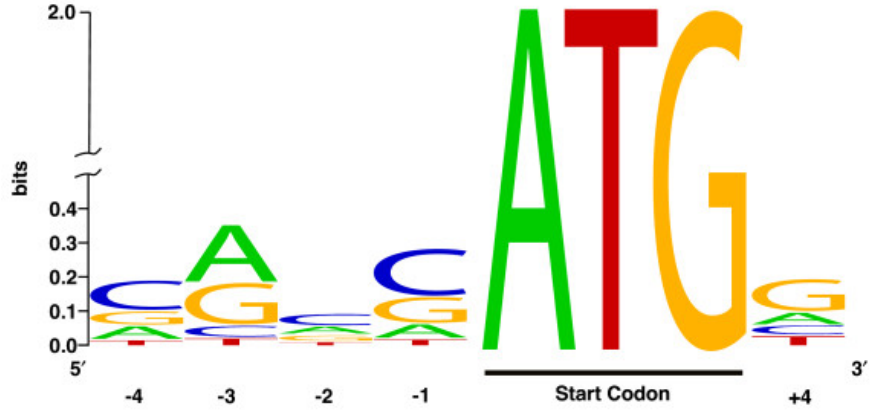
# *Ab initio* gene prediction

search genome for signs; do **not** use homology information



# *Ab initio* gene prediction

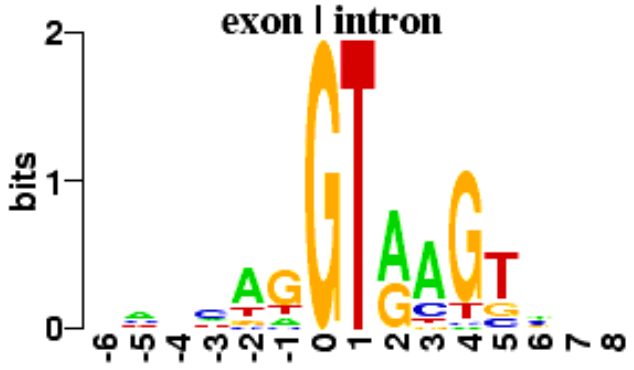
**Kozak motif**
(translation initiation; bovine)



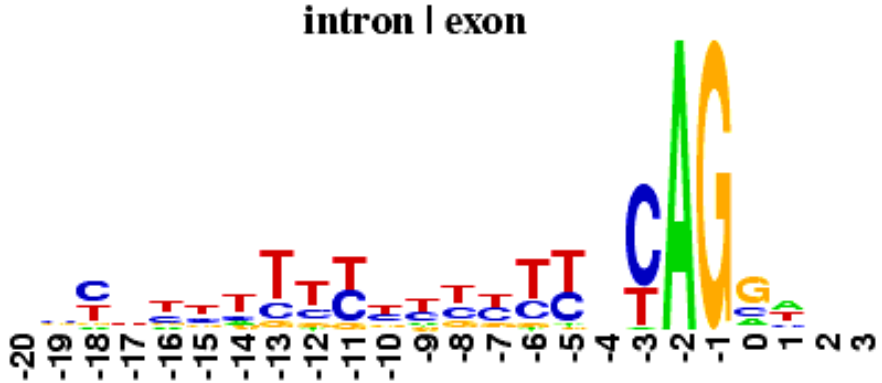

Harhay et al.,
BMC Genomics 6:166 (2005)

weblogo
http://weblogo.berkeley.edu/
examples.html

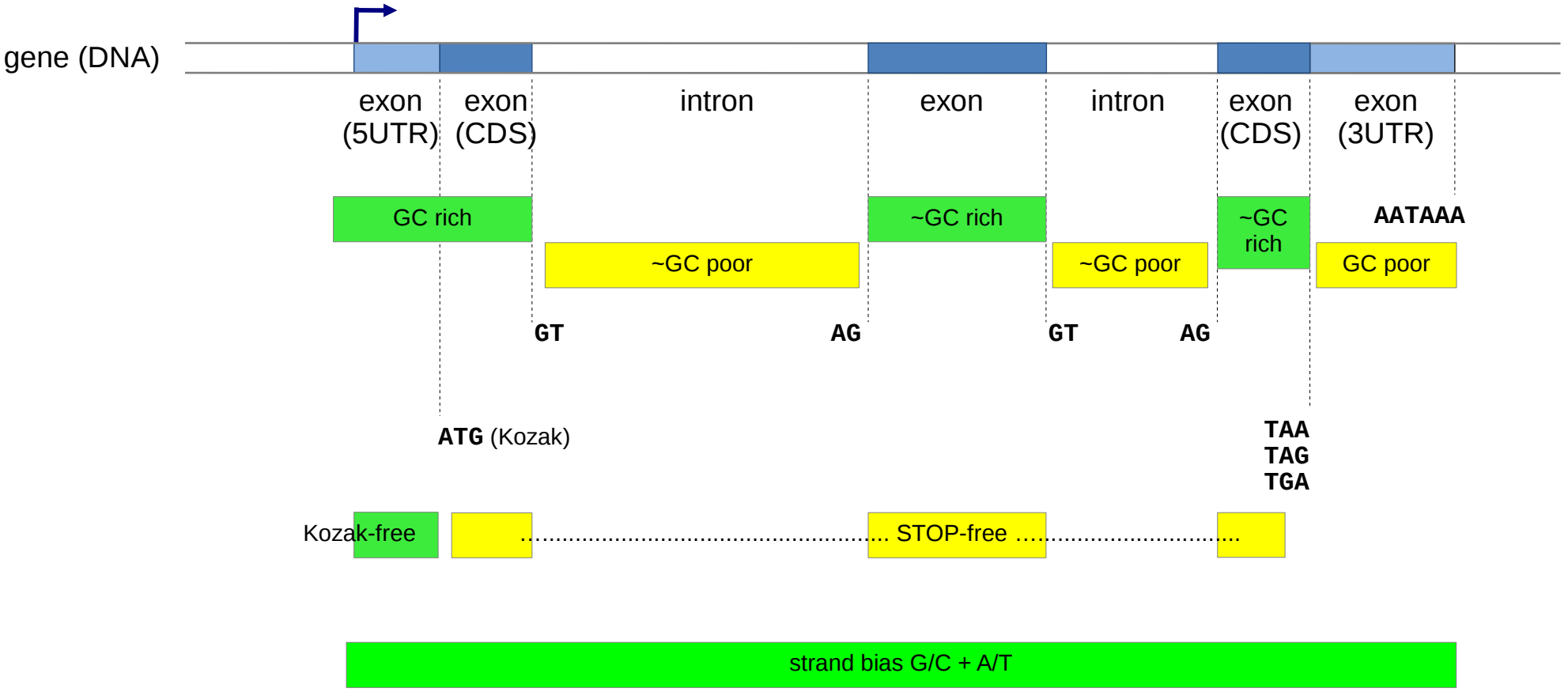
**5' splice site**
(splice donor)



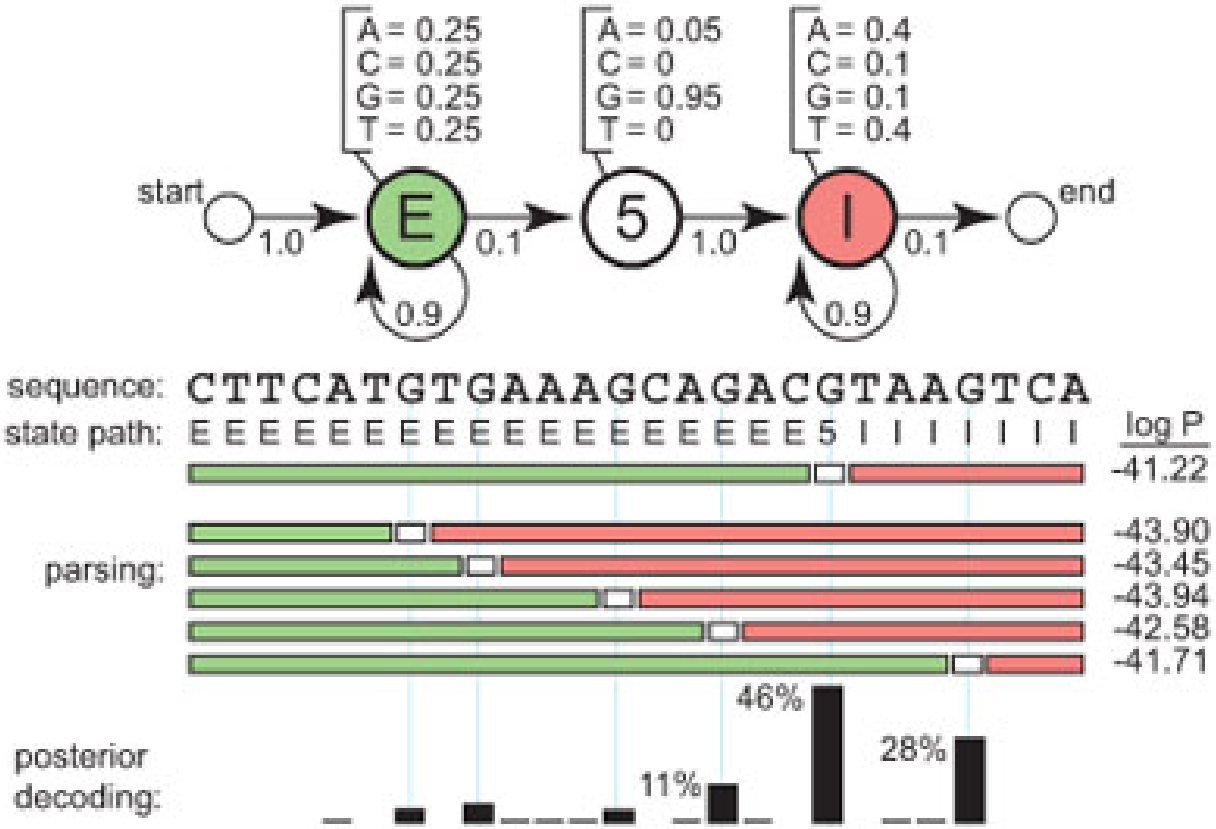

**3' splice site**
(splice acceptor)



# *Ab initio* gene prediction



# *Ab initio* gene prediction




Heyman, The Scientist 19:26 (2005)
Eddy, Nat Biotechnol 22:1315 (2004)

# Transcriptome – what to learn from it?

## 1. Eukaryote gene structure

- exon/intron structure
- 5' end => promoter
- 3' end

- isoforms : alternative gene structures

## 2. mRNA expression levels

- RNA-seq, tag counting

## 3. mRNA processing steps / regulatory mechanisms

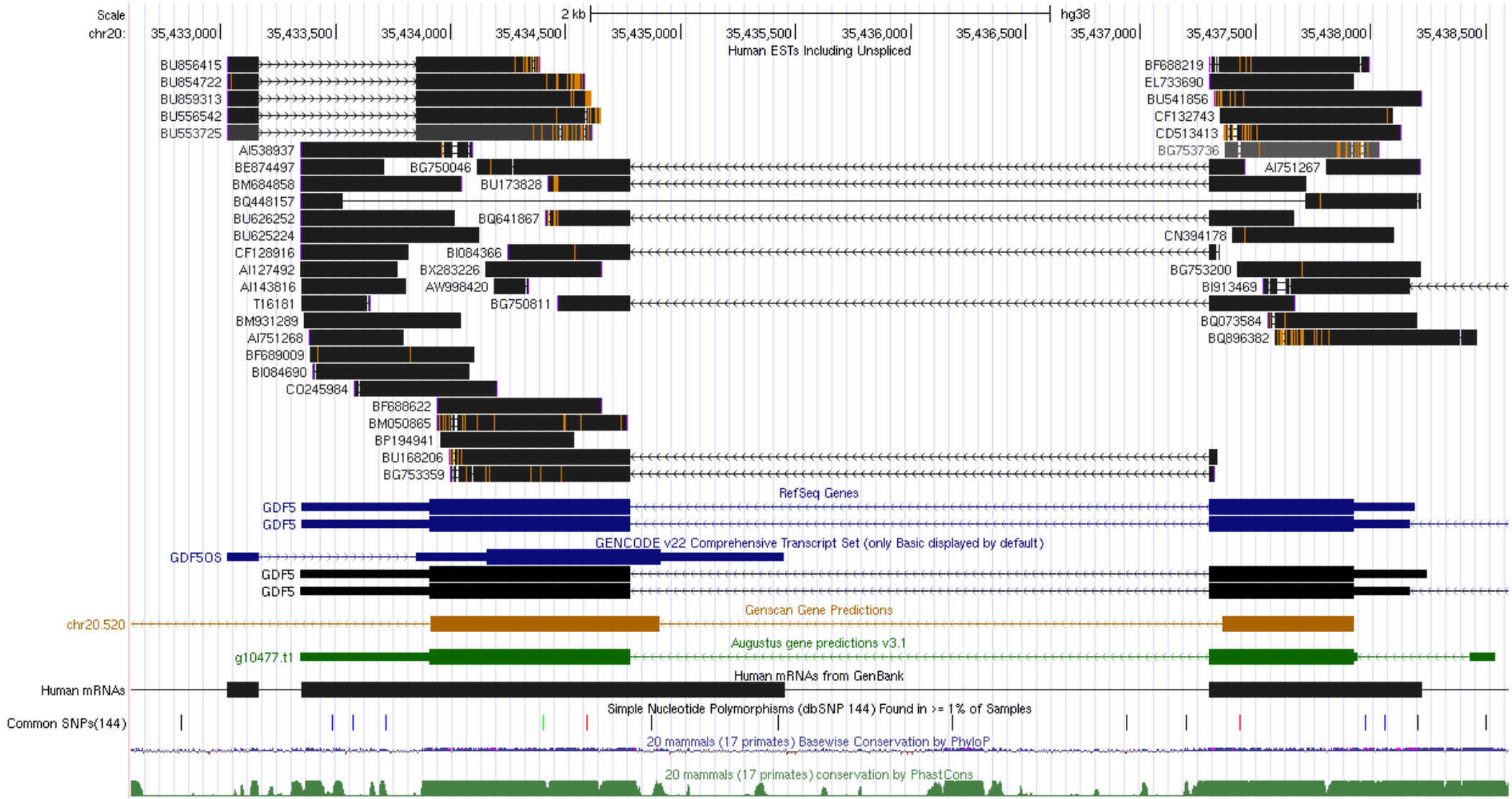
- identify/quantify binding factors:
  splicing regulators, surveillance factors

- A-to-I editing

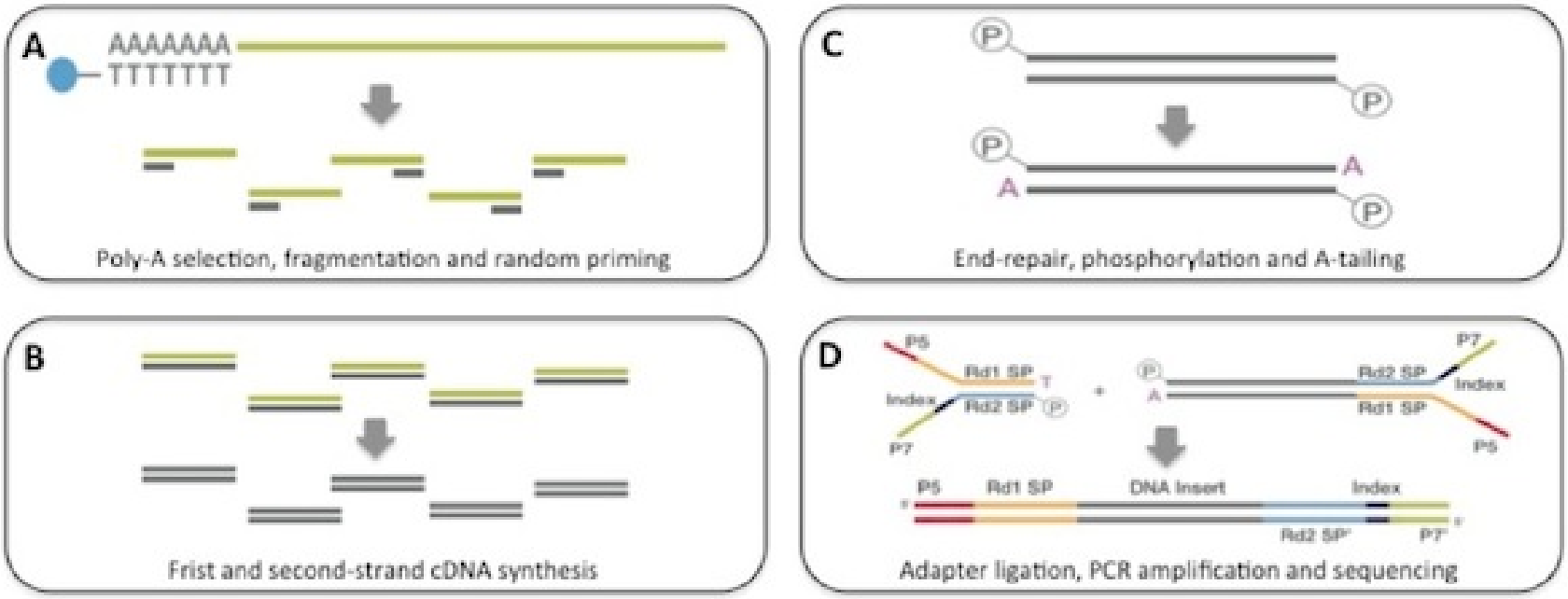
## 4. protein translation

# cDNA counts are proportional to mRNA levels

## human growth differentiation factor 5 (*GDF5*) locus

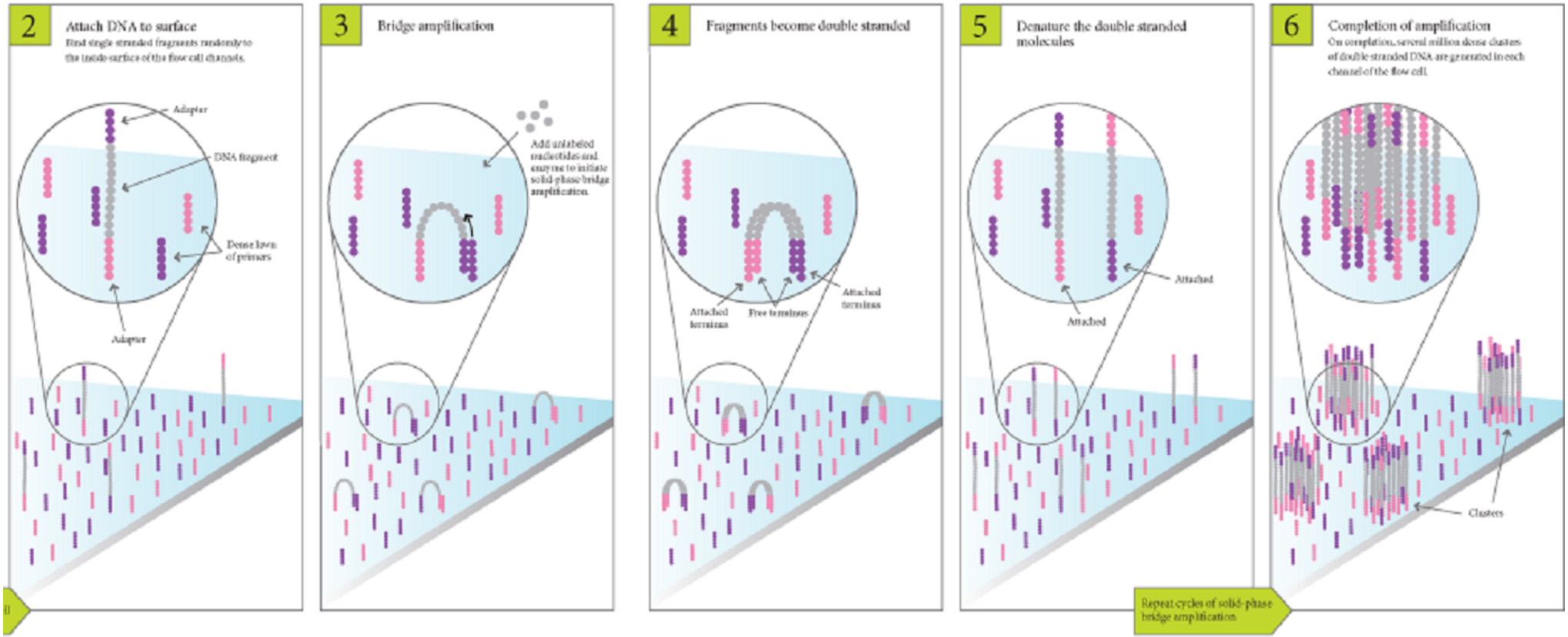


# Illumina Tru-Seq RNA-seq protocol



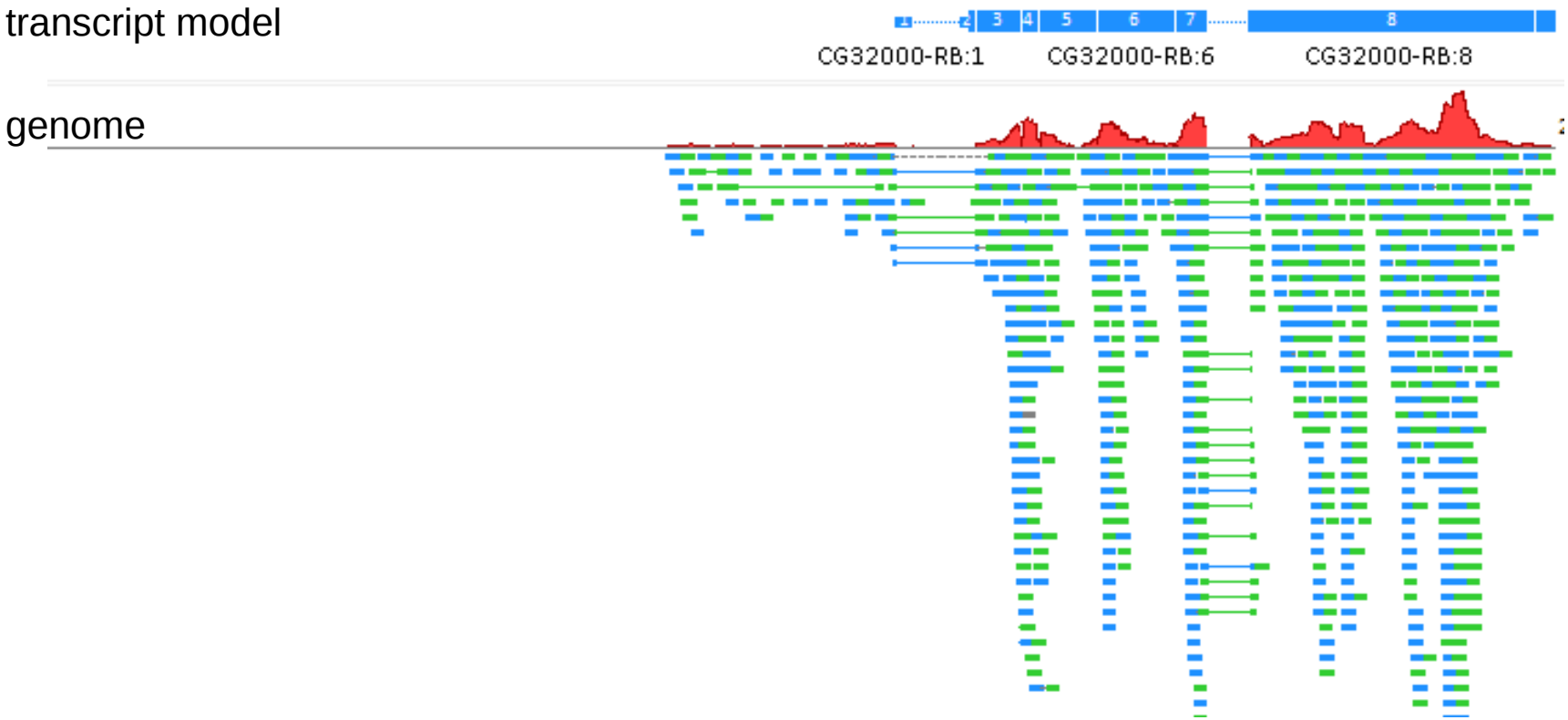

Library prep begins from 100ng-1ug of Total RNA which is poly-A selected (A) with magnetic beads. Double-stranded cDNA (B) is phosphorylated and A-tailed (C) ready for adapter ligation. The library is PCR amplified (D) ready for clustering and sequencing.

# Illumina technology : solid-phase clonal single molecule PCR "Bridge amplification"



## Read mapping,

- against reference transcripts, bowtie
- against annotated genome using “spliced alignment”, e.g. RedHat, or



## Counting of cDNA sequence tags

example MAQC-3 dataset (GSM475204)

| | |
|---|---|
| GAPDH | 2824.293 |
| ACTB | 2289.077 |
| GFAP | 2010.641 |
| RN7SL2 | 1992.531 |
| CKB | 1814.053 |
| MT3 | 1719.392 |
| MBP | 1629.907 |
| FTH1 | 1610.553 |
| TUBB4A | 1502.924 |
| PTGDS | 1474.918 |
| UBC | 1427.698 |
| CST3 | 1407.278 |
| UBB | 1391.971 |
| CLU | 1300.067 |
| MIF | 1242.009 |
| ALDOC | 1233.735 |
| MTRNR2L2 | 1217.551 |
| TMSB4X | 1190.765 |
| CALM3 | 1129.767 |
| MTRNR2L8 | 1090.615 |
| RN7SL1 | 1053.638 |
| RPL3 | 1021.194 |
| EEF2 | 1003.831 |

definition:

RPKM :=
reads per kb transcript per million readings

Mortazavi et al.,
Nat Methods 5:621 (2008)

# Transcriptome – what to learn from it?

## 1. Eukaryote gene structure

- exon/intron structure
- 5' end => promoter
- 3' end

- isoforms : alternative gene structures

## 2. mRNA expression levels

- RNA-seq, tag counting

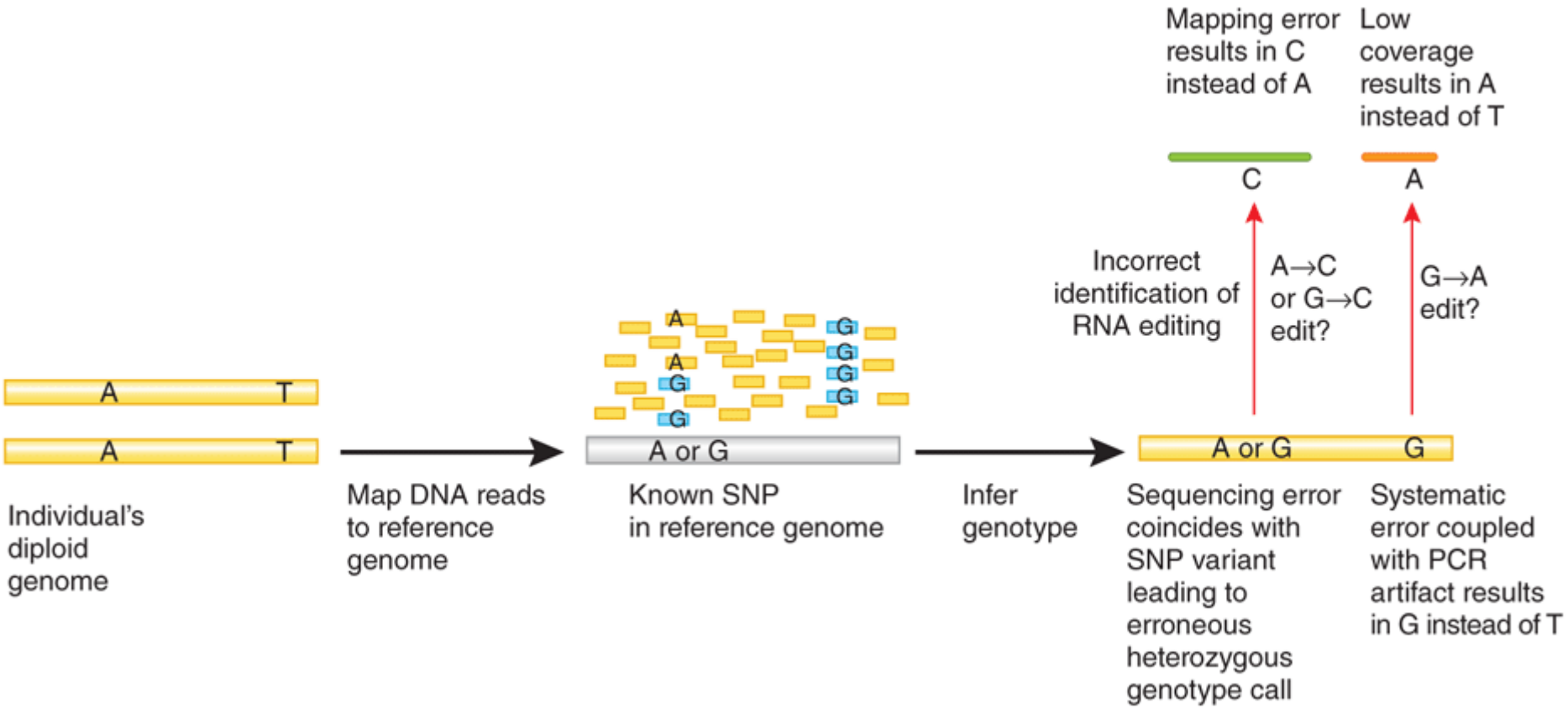
## 3. mRNA processing steps / regulatory mechanisms

- A-to-I editing

- identify/quantify binding factors:
  splicing regulators, surveillance factors

## 4. protein translation

# Analysis of A-to-I editing




Pachter, Nat Biotech 30:246 (2012)

# Factors binding to mRNA

## - splicing regulators

cause alternative gene structures
depending on tissue or environment

unproductive mRNA structures cause degradation (NMD)
=> influence on expression level

## - surveillance factors

regulate stability of mRNA (+/-)

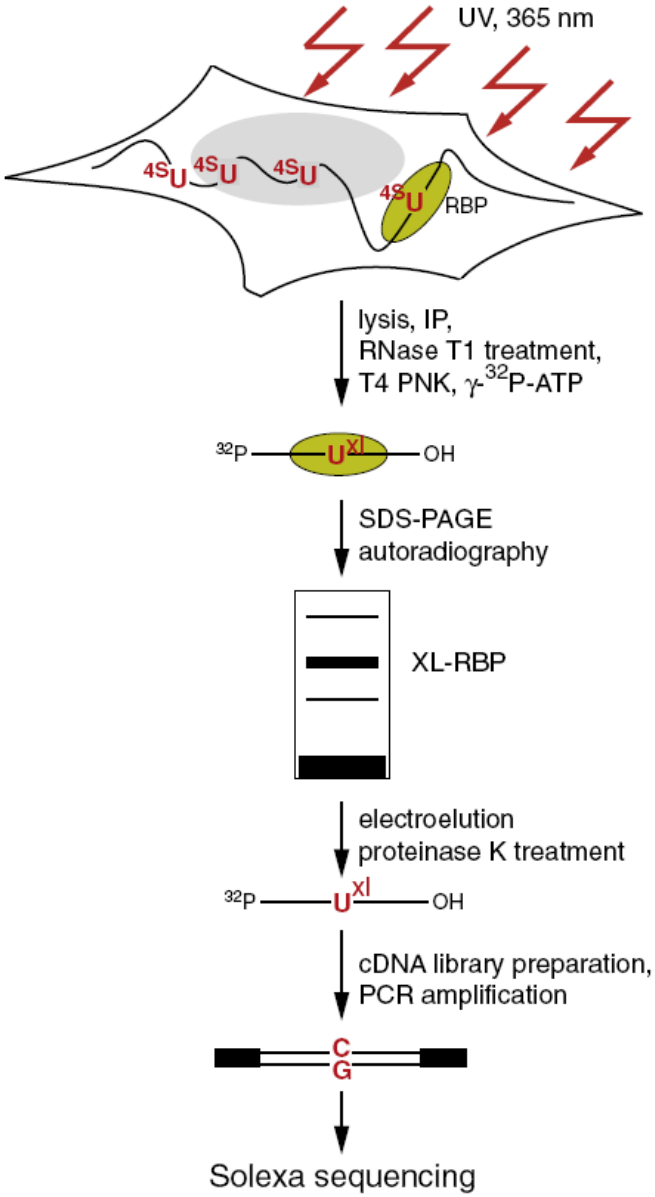
## - translation factors

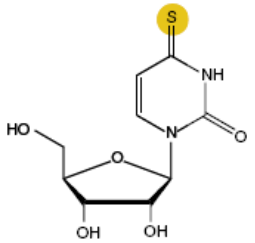
regulate affinity to the ribosome and
commitment to translation (elongation factors)

upstream ORFs influence the translation of mRNA into protein

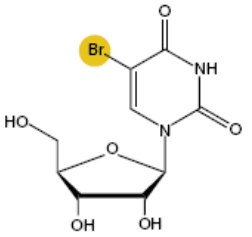
## CLIP method

Hafner et al.,
Cell 141:129 (2010)

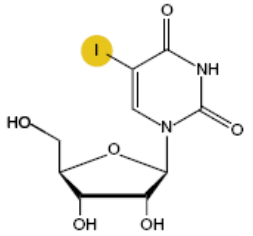


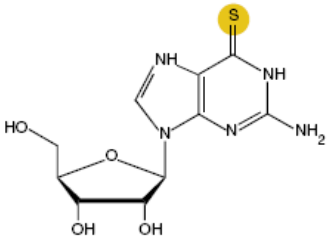

4-thiouridine (4SU)




5-bromouridine (5BrU)




5-iodouridine (5IU)




6-thioguanosine (6SG)

## CLIP method

Licatalosi et al., Nature 456:464 (2008)

**Splicing factor Nova, driving exon inclusion or exclusion**

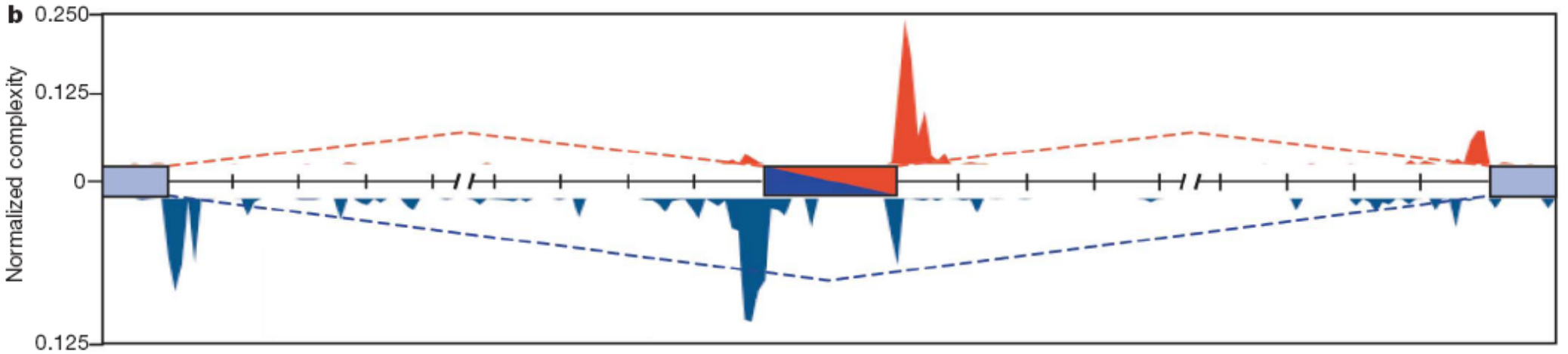
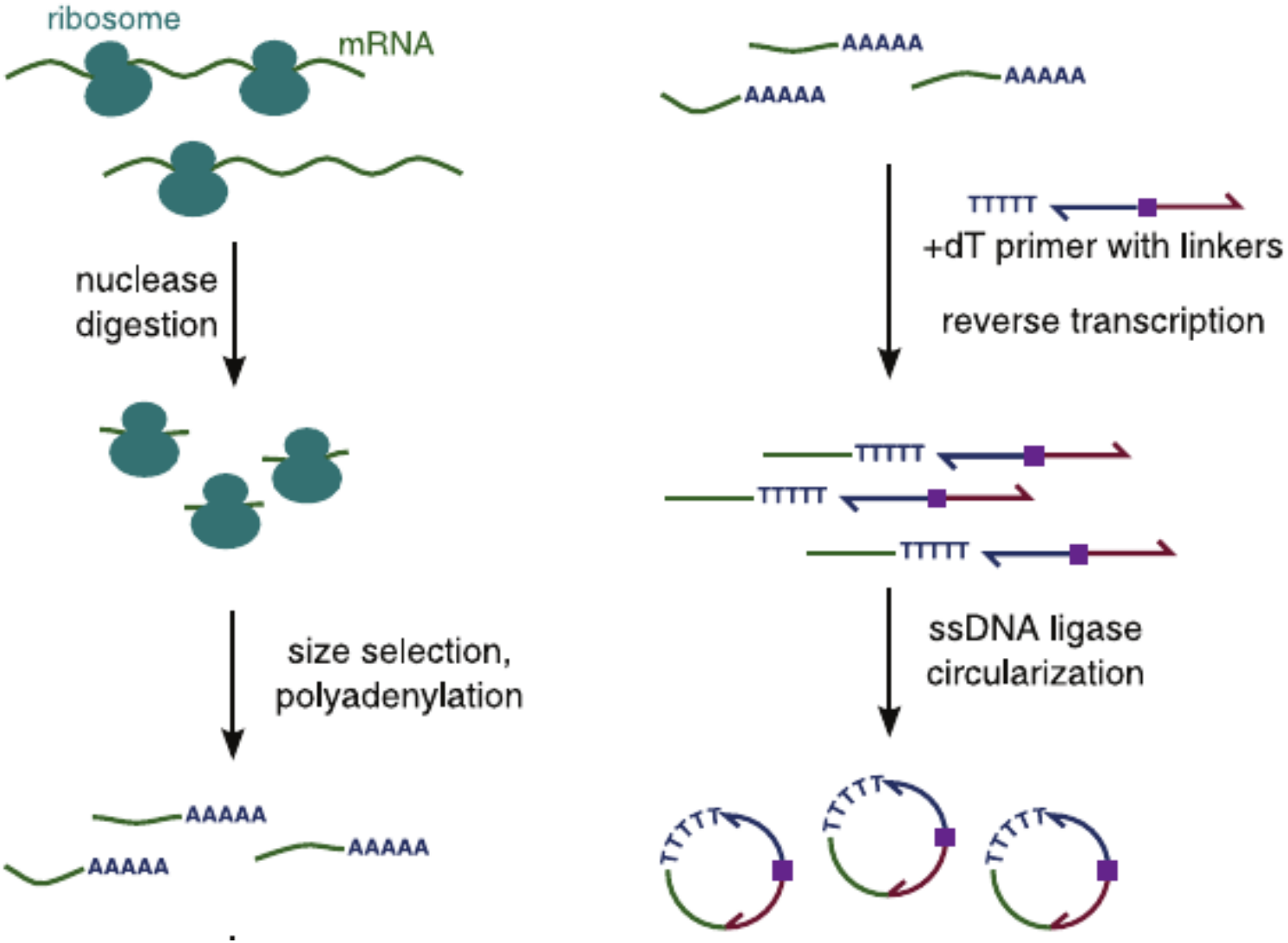



**Figure 2 | Nova–RNA interaction maps associated with Nova-dependent splicing regulation.** **a**, CLIP tags around all known Nova-regulated cassette exons, with one colour per transcript. Tags were mapped onto a composite transcript containing an alternative (dark blue/red box) and flanking constitutive (light blue box) exons. Tags are from transcripts showing Nova-dependent exon inclusion (top panel) or exclusion (bottom panel); representative examples of experimentally validated target RNAs (*Arpp21* and *Mcf2l*) are shown (insets). **b**, Normalized complexity map (see Methods) of Nova–RNA interactions recapitulate predicted maps[18] (insets) for Nova-dependent exon inclusion (red) or exclusion (blue).

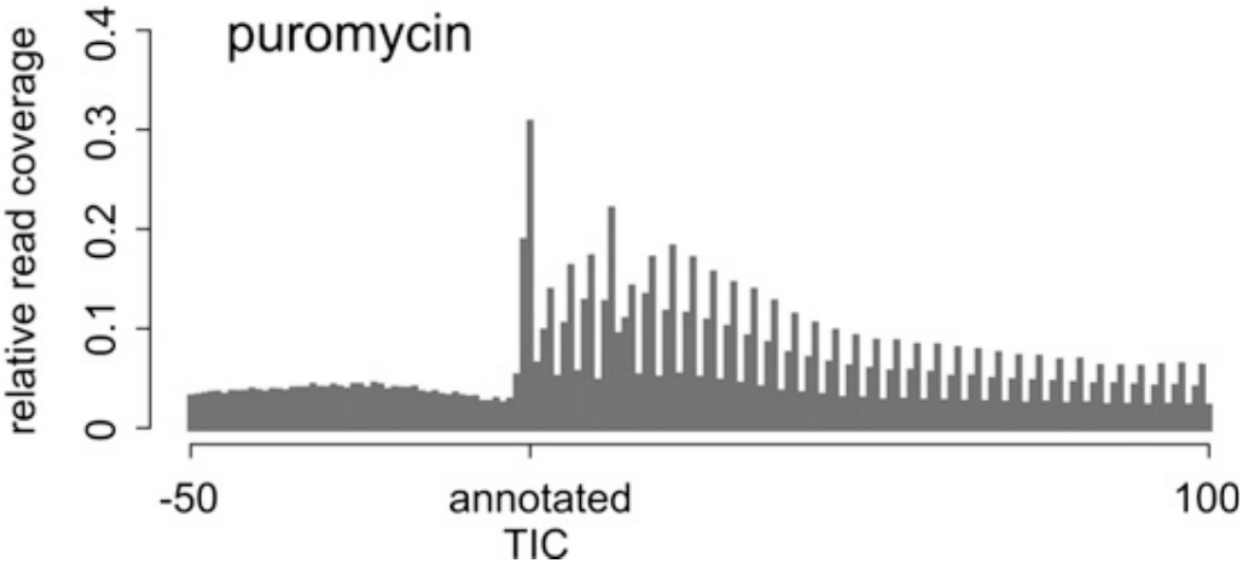
## Ribosomal footprinting

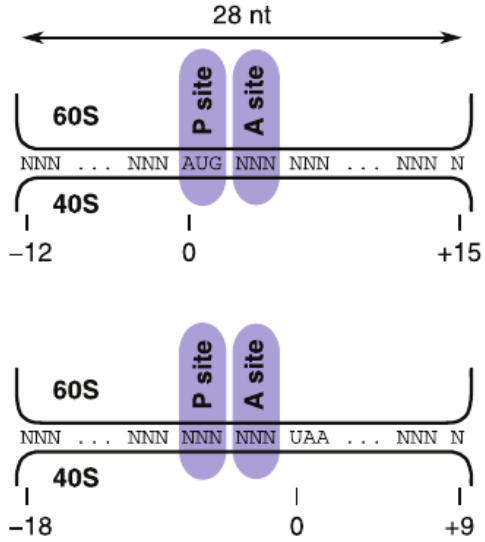
Ingolia et al., Science 324:218 (2009)

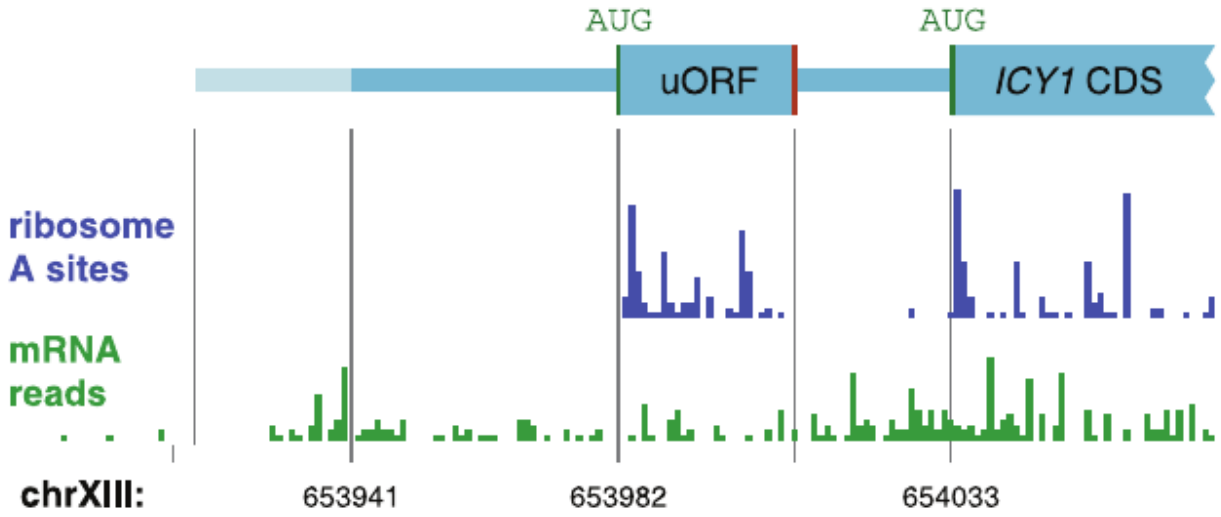


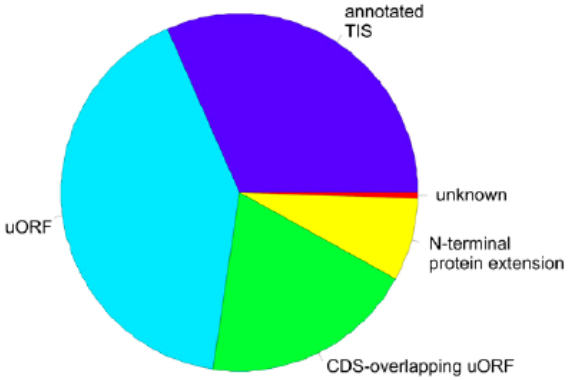
## Ribosomal footprinting

Ingolia et al., Science 324:218 (2009)

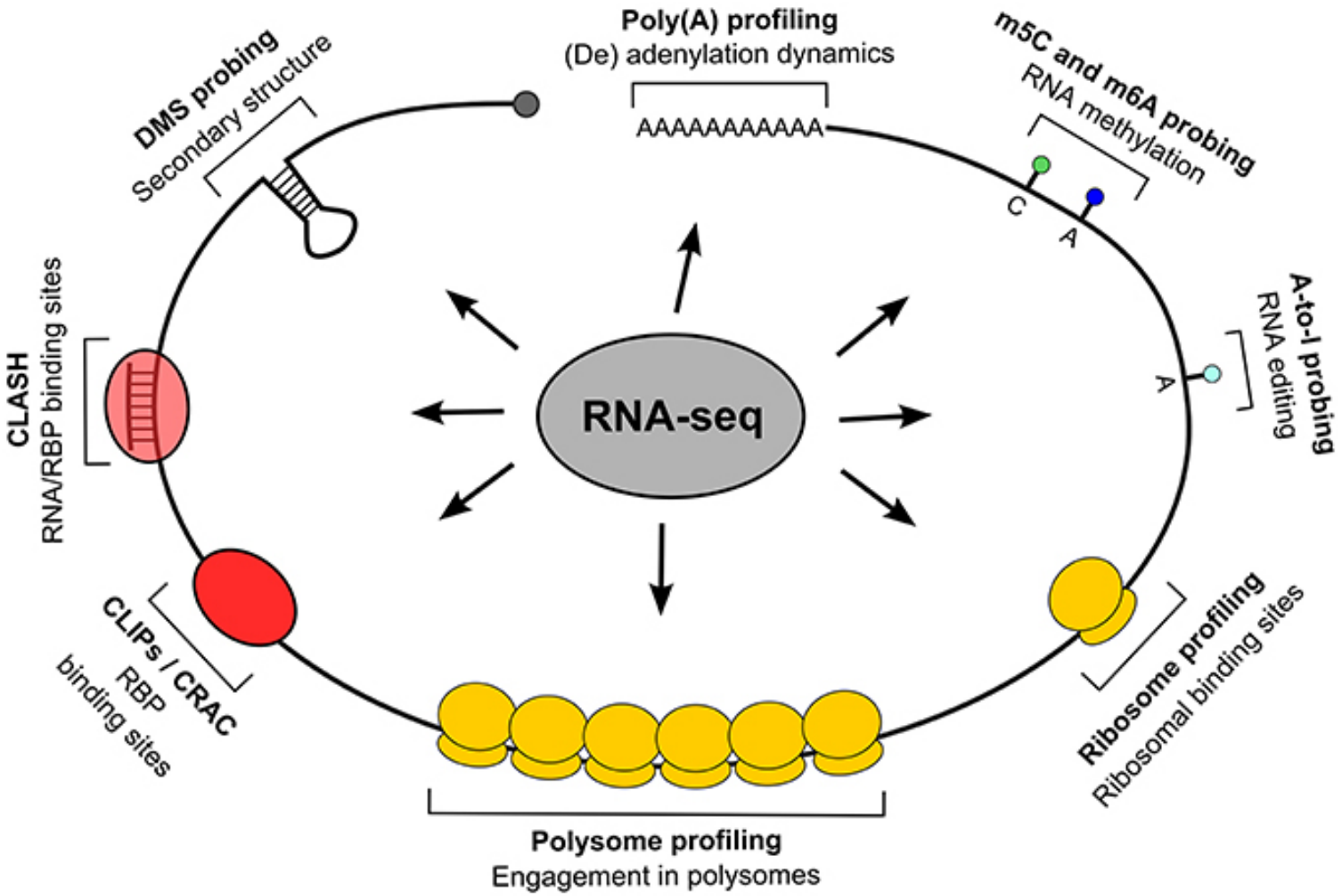


# Methods for specific aspects of the transcriptome



# Summary

## 1. Eukaryote gene structure

- exon/intron structure
- 5' end => promoter
- 3' end

- isoforms : alternative gene structures

## 2. mRNA expression levels

- RNA-seq, tag counting

## 3. mRNA processing steps / regulatory mechanisms

- A-to-I editing

- identify/quantify binding factors:
  splicing regulators, surveillance factors

## 4. protein translation

**genome.leibniz-fli.de**
**→ Lectures**

**karol.szafranski@...**