The Abdus Salam
International Centre
for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

**2415-14**

**Winter School on Quantitative Systems Biology**

***26 November - 7 December, 2011***

# The Transcriptome: sequencing the effectors of the genome

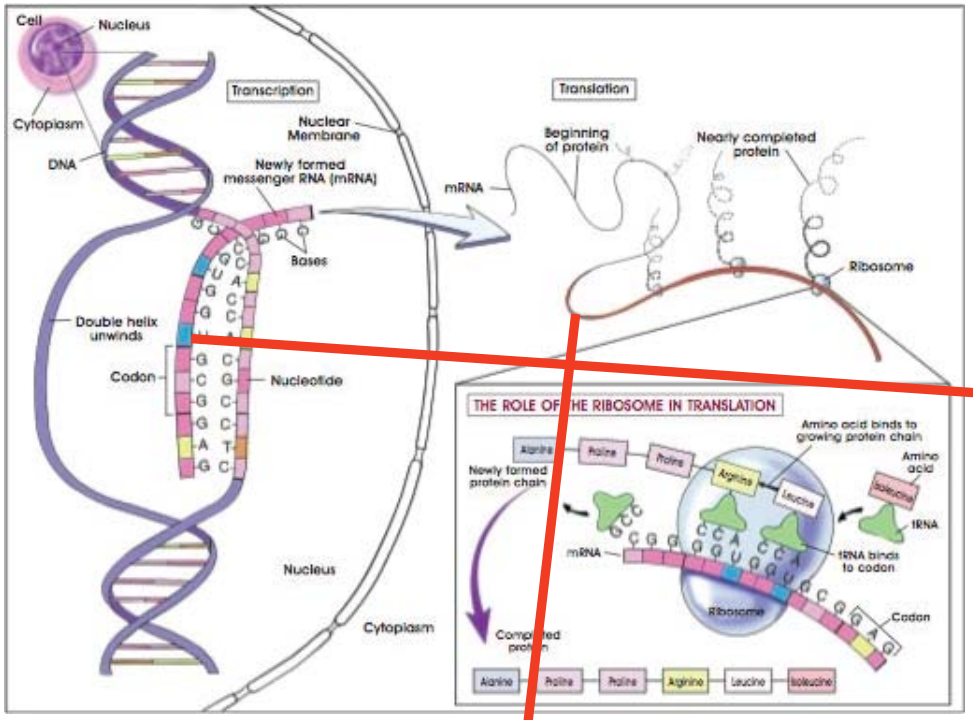
R. Sachidanandam
*Mount Sinai Sch. of Medicine*
*New York*
*USA*

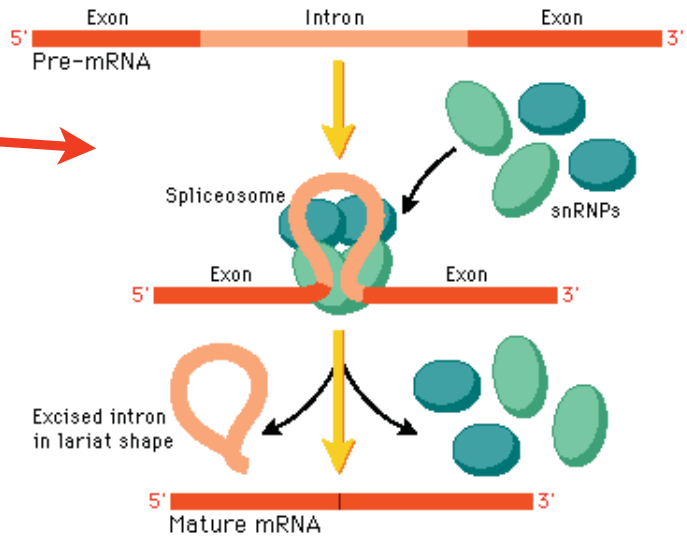
Strada Costiera, 11 - 34151 - Trieste - Italy • Tel. +39 0402240111 • Fax. +39 040224163 • sci_info@ictp.it • www.ictp.it
ICTP is governed by UNESCO, IAEA, and Italy, and it is a UNESCO Category 1 Institute

# Transcripts

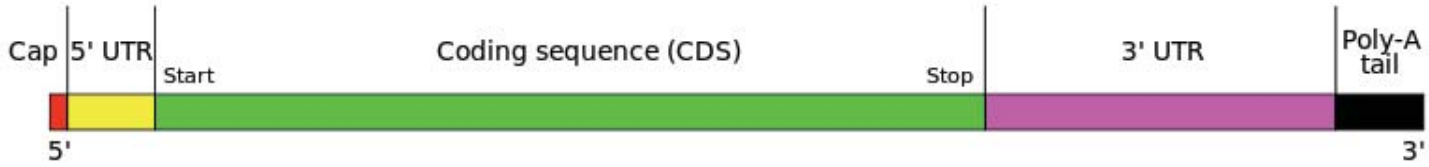
Ravi Sachidanandam
Mount Sinai School of Medicine

# mRNA transcription

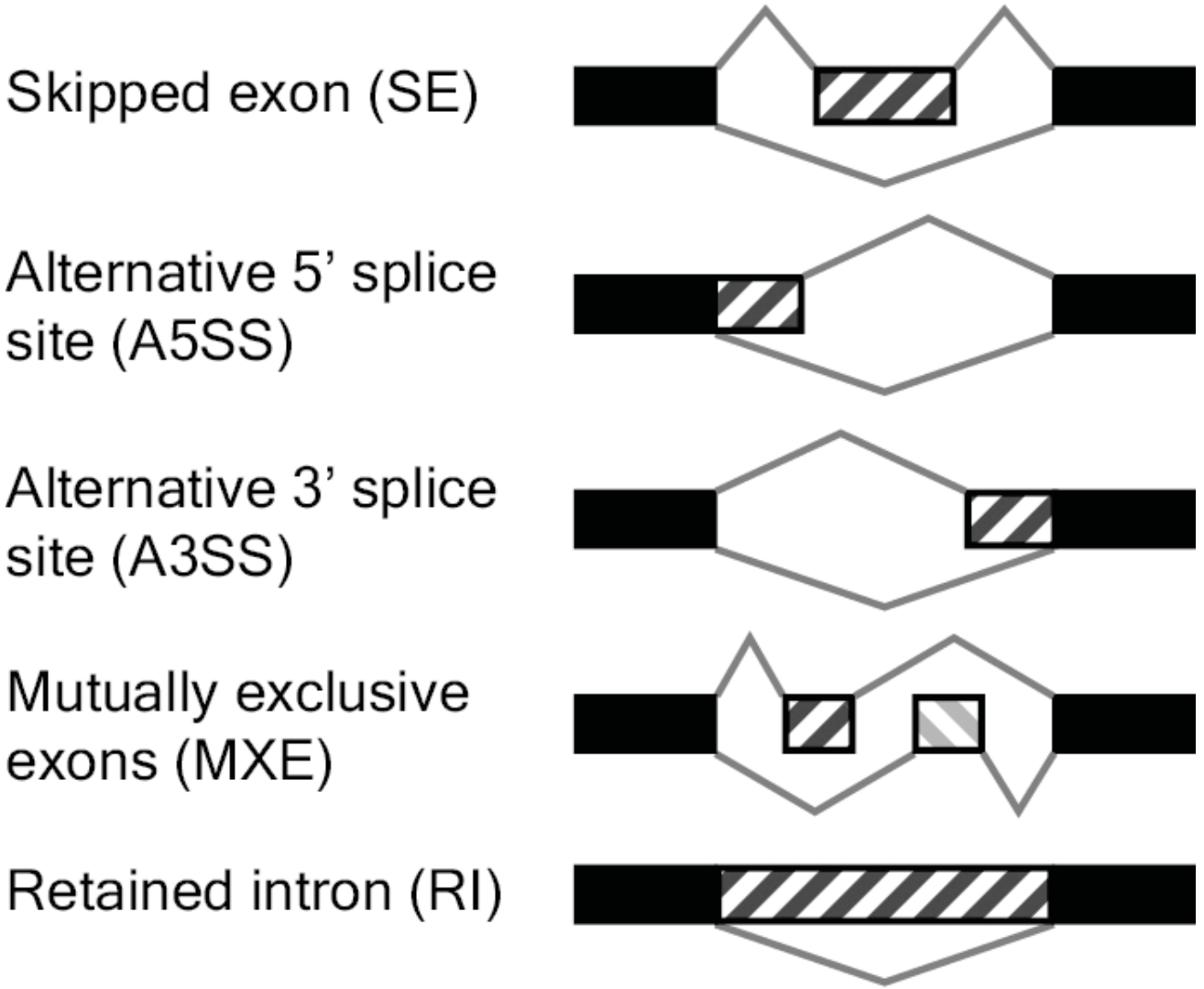



**The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)**

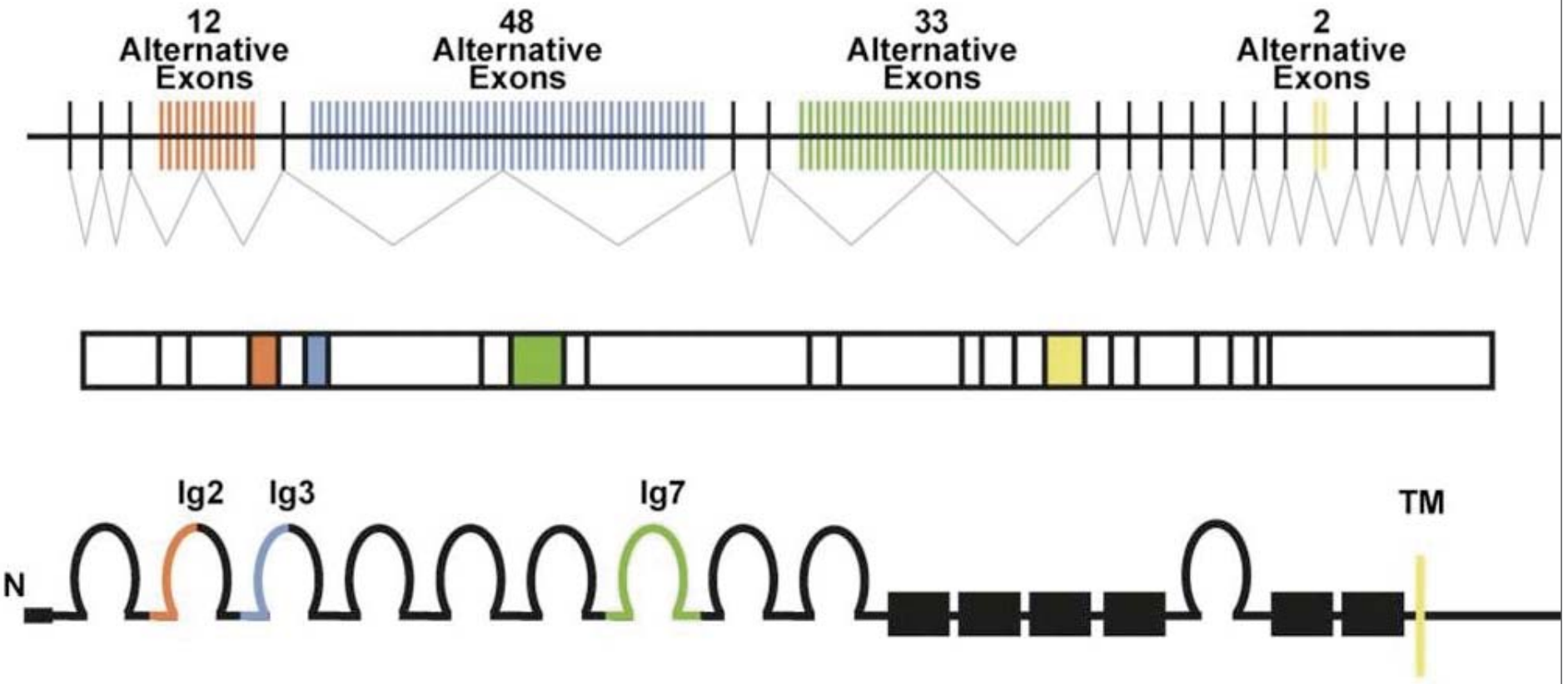


# Alternative Splicing Events



# Alternative Splicing of Dscam Potentially Generates 38,016 Isoforms

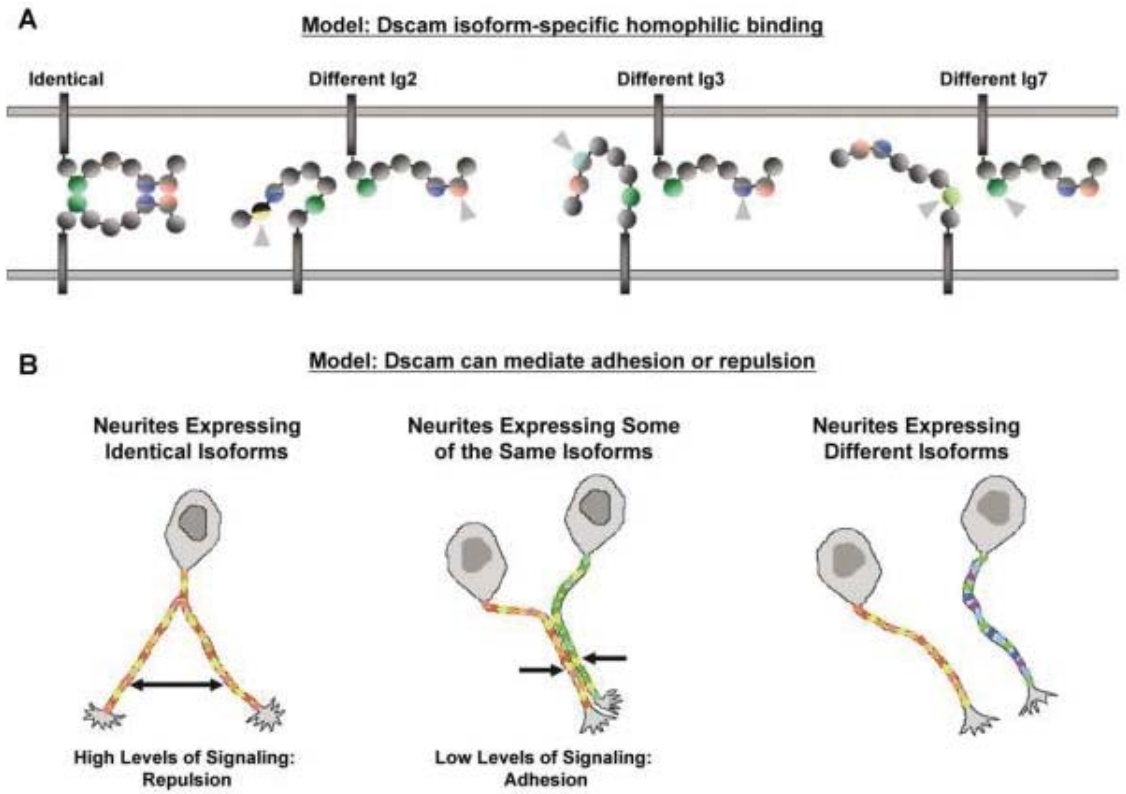
**DSCAM: Down syndrome cell adhesion molecule**




(A) Schematic representation of *Dscam* gene, mRNA, and protein. The Dscam protein contains both constant and variable domains. The variable domai are encoded by alternative exons. Each block of alternative exons is indicated by a different color. A transcript contains only one alternative exon from each block. The *Dscam* gene encodes 12 alternative exons for the N-terminal half of Ig2 (red), 48 alternative exons for the N-terminal half of Ig3 (blue), and 33 alternative exons for Ig7 (green). There are two alternative transmembrane domains (yellow).

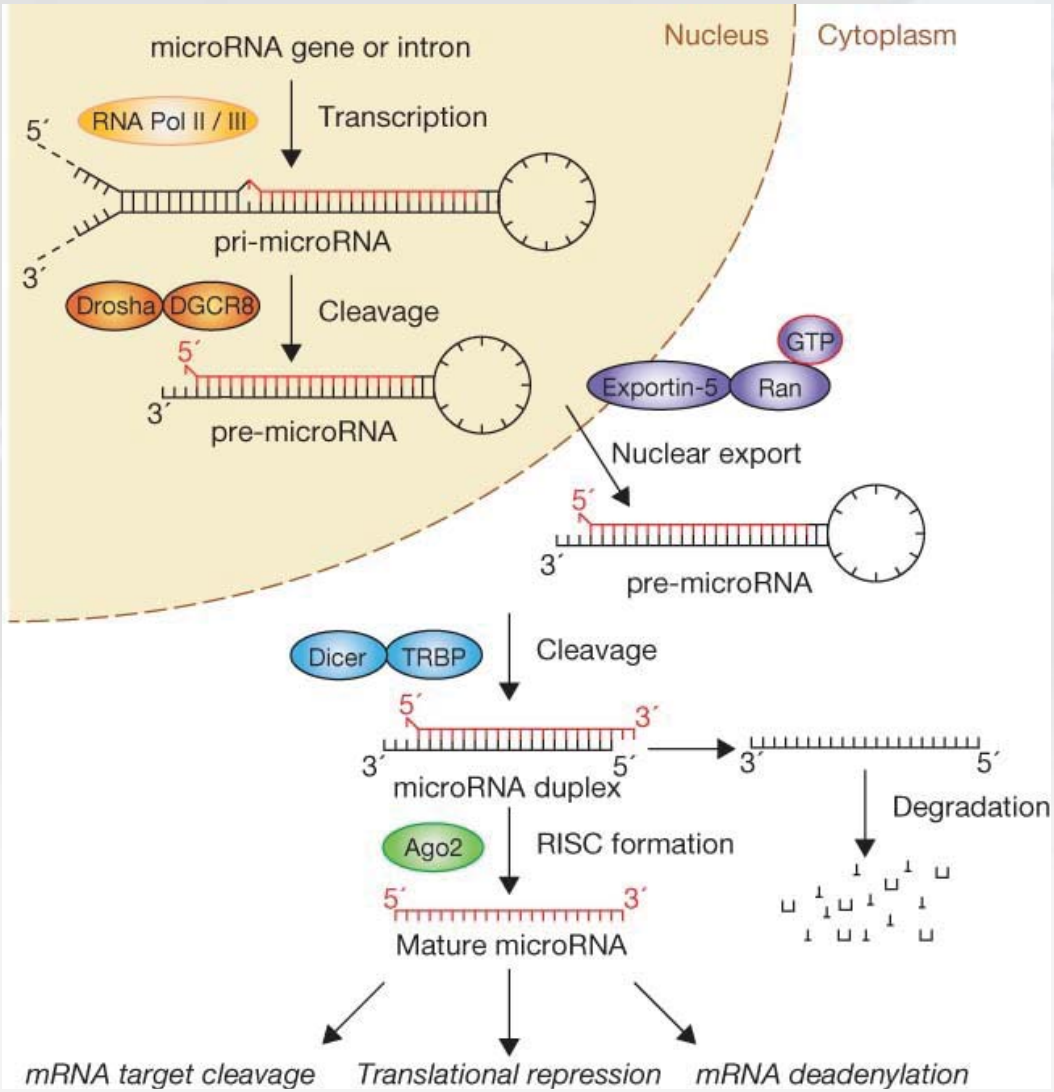


PNAS



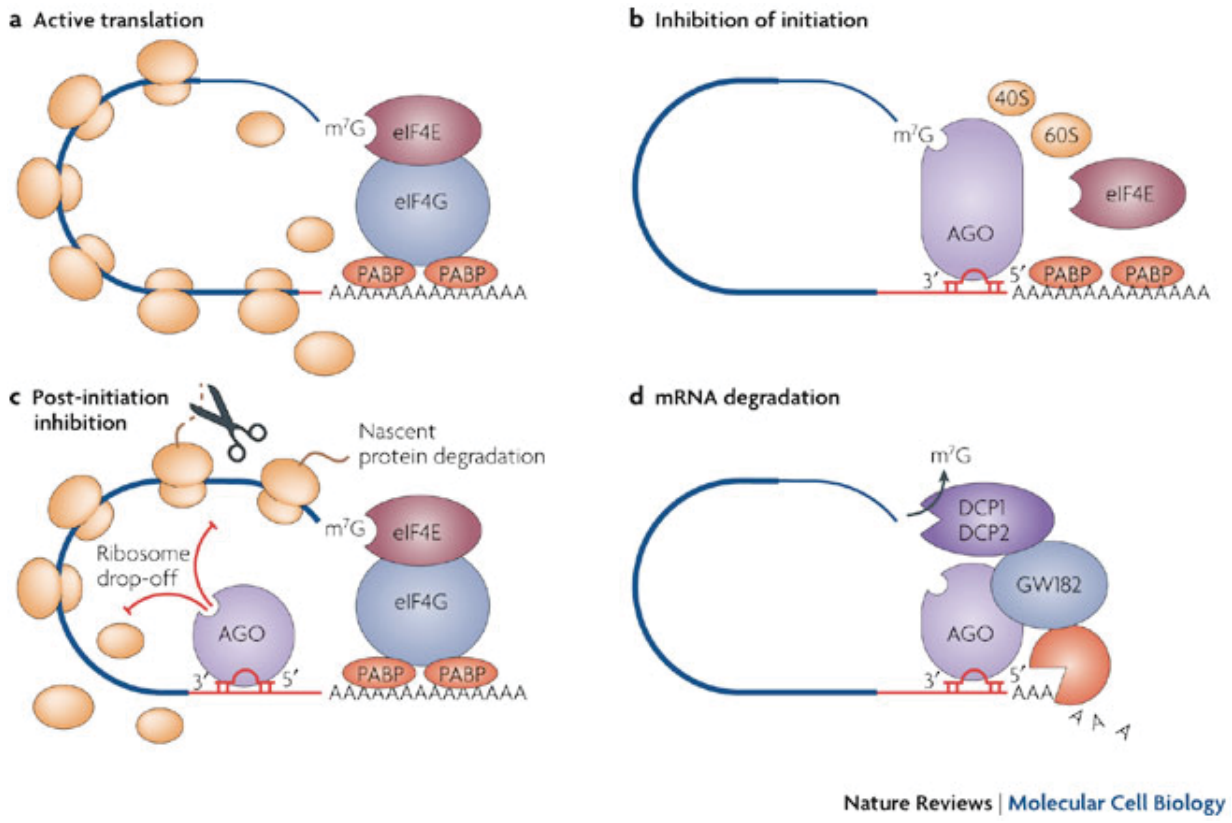

**A Model for Dscam-Mediated Interactions**

(A) Schematic representation of homophilic binding mediated by the three variable Ig domains. The eight N-terminal Ig domains (circles) are shown. Constant Ig domains are gray, and variable Ig domains are in color. The remainder of the protein is represented by a gray rectangle. Isoforms sharing identical variable Ig domains (represented by matching colors) bind to each other, while isoforms differing in only one variable Ig domain (gray arrowheads) do not. We propose that each variable Ig domain binds to the same variable Ig domain in an opposing molecule. As isoforms sharing any two identical variable Ig domains do not bind or exhibit reduced binding relative to binding between identical isoforms, it is likely that the binding of all three variable Ig domains is required to stabilize otherwise weak interactions between individual variable Ig domains. (B) Schematic representation of Dscam-mediated interactions between neurites. We propose that differences in levels of Dscam signaling influence the nature of the interactions between neurites. High levels of signaling between neurites expressing the same array of Dscam isoforms result in contact-dependent repulsion (left panel), while low levels of signaling between neurites

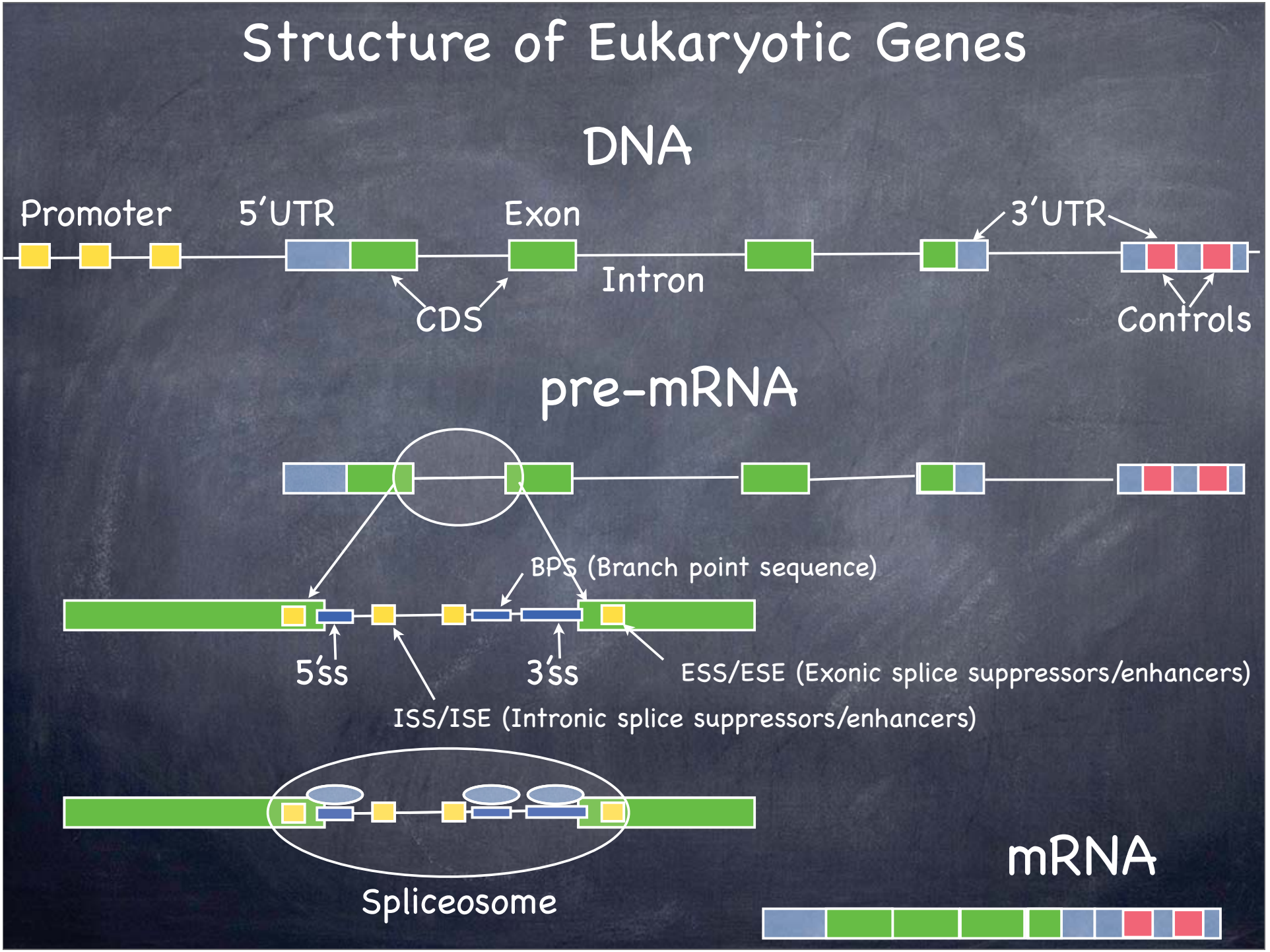



The miRNA processing pathway has long been viewed as linear and universal to all mammalian miRNAs. This canonical maturation includes the production of the primary miRNA transcript (pri-miRNA) by RNA polymerase II or III and cleavage of the pri-miRNA by the microprocessor complex Drosha–DGCR8 (Pasha) in the nucleus. The resulting precursor hairpin, the pre-miRNA, is exported from the nucleus by Exportin-5–Ran-GTP. In the cytoplasm, the RNase Dicer in complex with the double-stranded RNA-binding protein TRBP cleaves the pre-miRNA hairpin to its mature length. The functional strand of the matur miRNA is loaded together with Argonaute (Ago2) proteins into the RNA-induced silencing complex (RISC), where it guides RISC to silence target mRNAs through mRNA cleavage, translational repression or deadenylation, whereas the passenger strand (black) is degraded. In this review we discuss the many branches, crossroads and detours in miRNA processing that lead to the conclusion that many different ways exist to generate a mature miRNA.

# mRNA post-translational regulation by miRNA



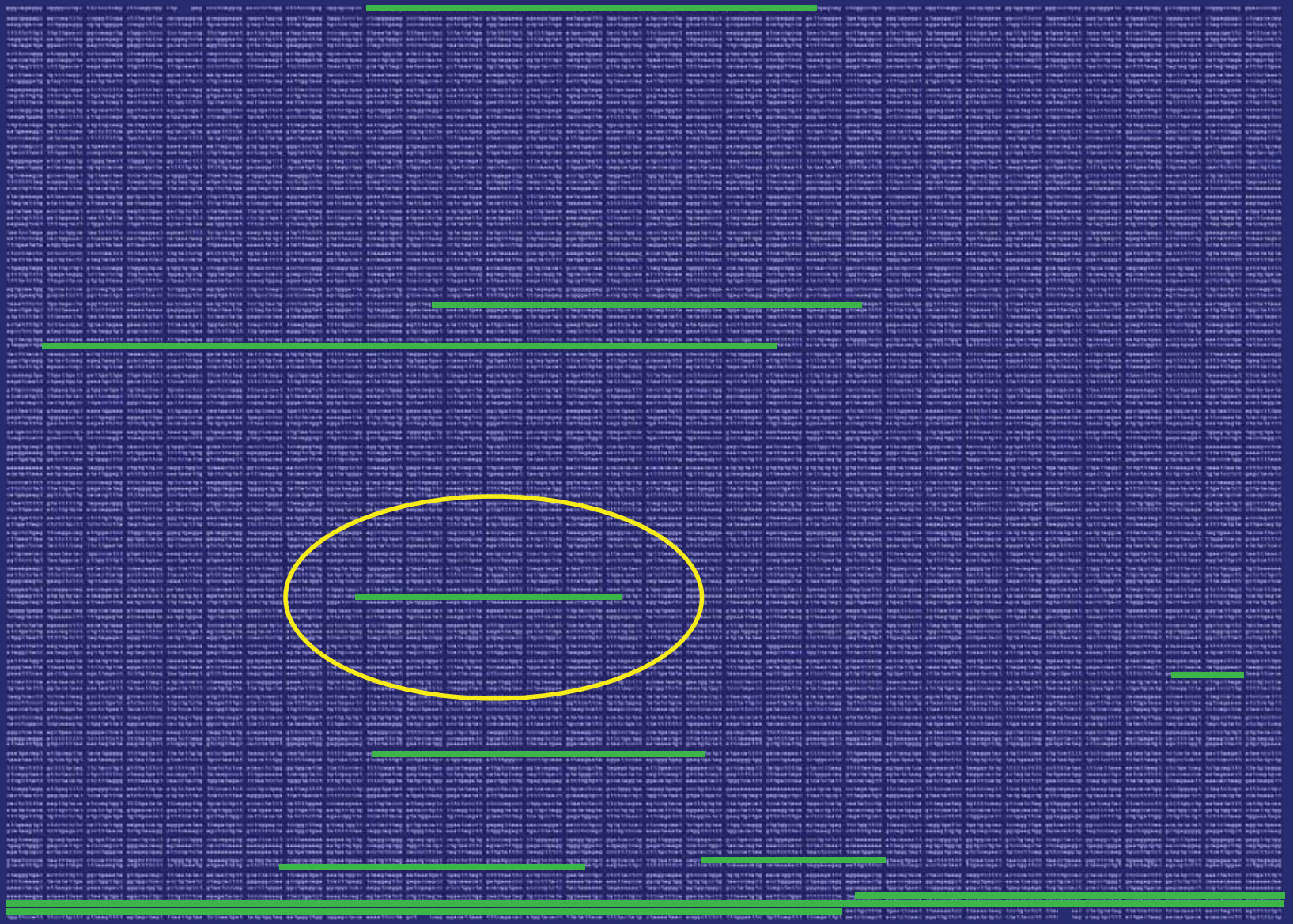

Nature Reviews | Molecular Cell Biology

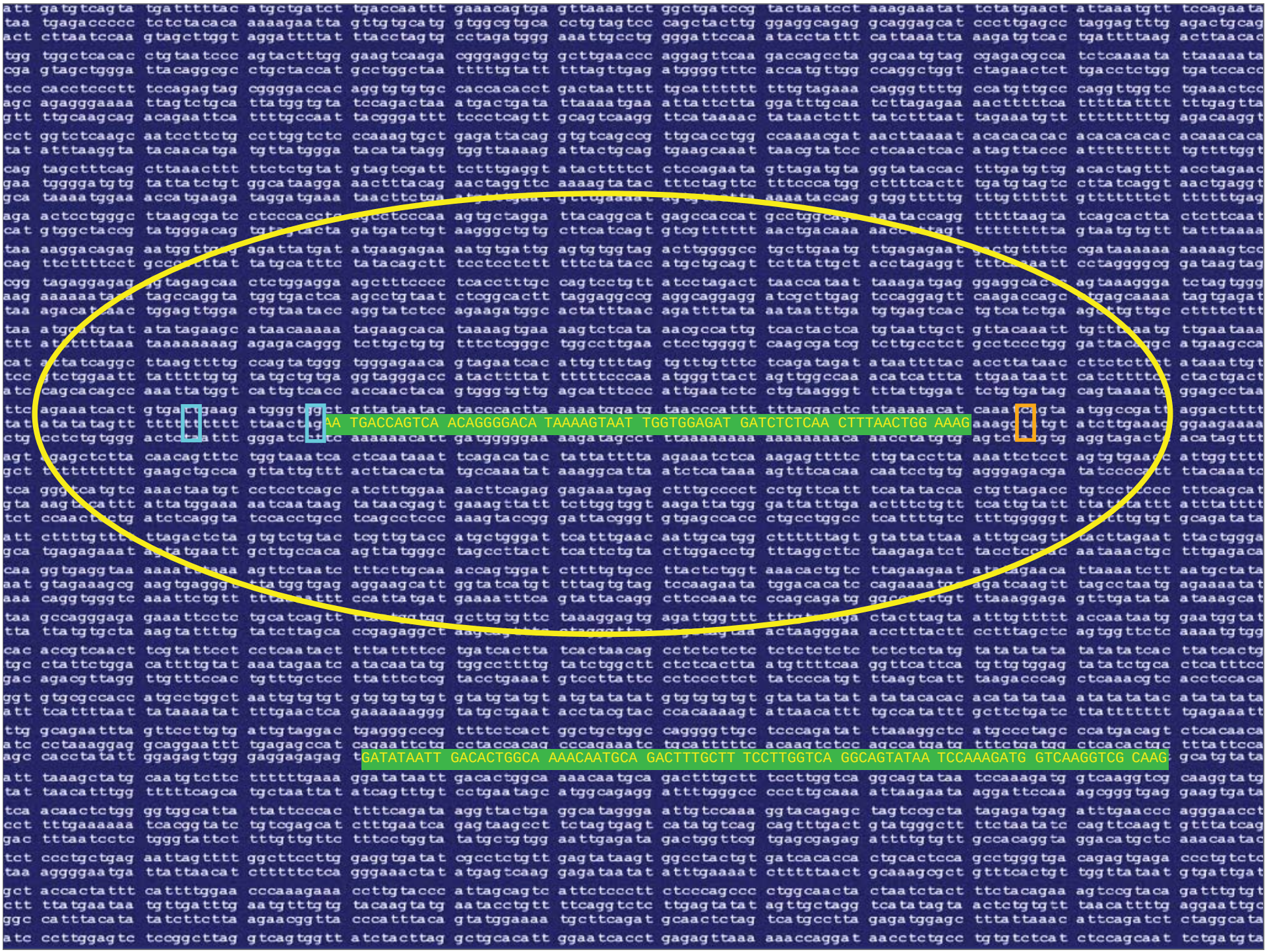

Structure of Eukaryotic Genes
DNA
Promoter
5'UTR
Exon
Intron
3'UTR
CDS
Controls
pre-mRNA
BPS (Branch point sequence)
5'ss
3'ss
ESS/ESE (Exonic splice suppressors/enhancers)
ISS/ISE (Intronic splice suppressors/enhancers)
Spliceosome
mRNA

# Splice-site prediction

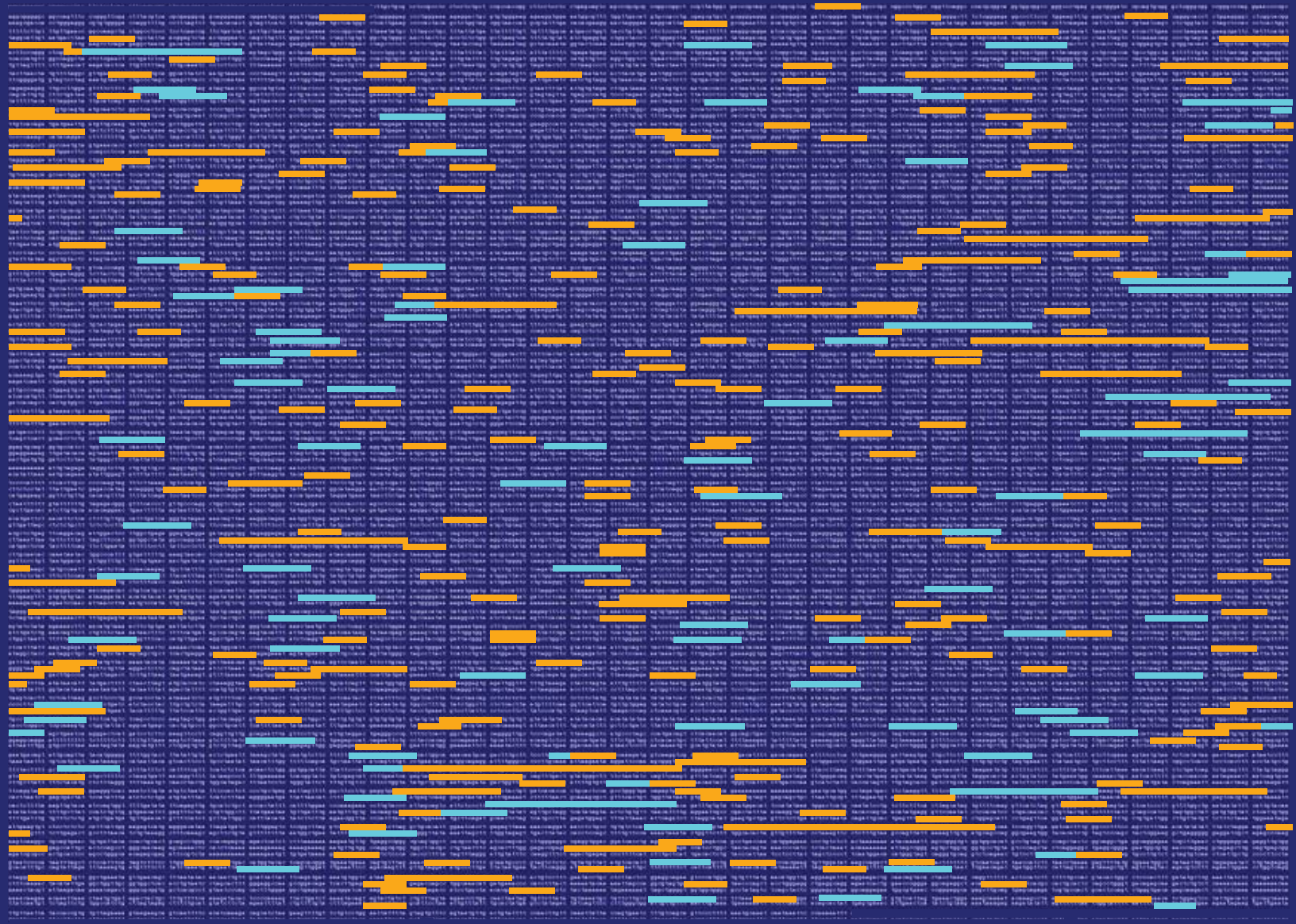
- Good splice-site prediction implies understanding of its recognition by the spliceosome.
- Splice-site predictor/classifier: The Mouse-trap of bioinformatics ?
- Predictors generally improve with more data.
- SpliceRack is the result.

# *HPRT1* Exon-Intron Structure

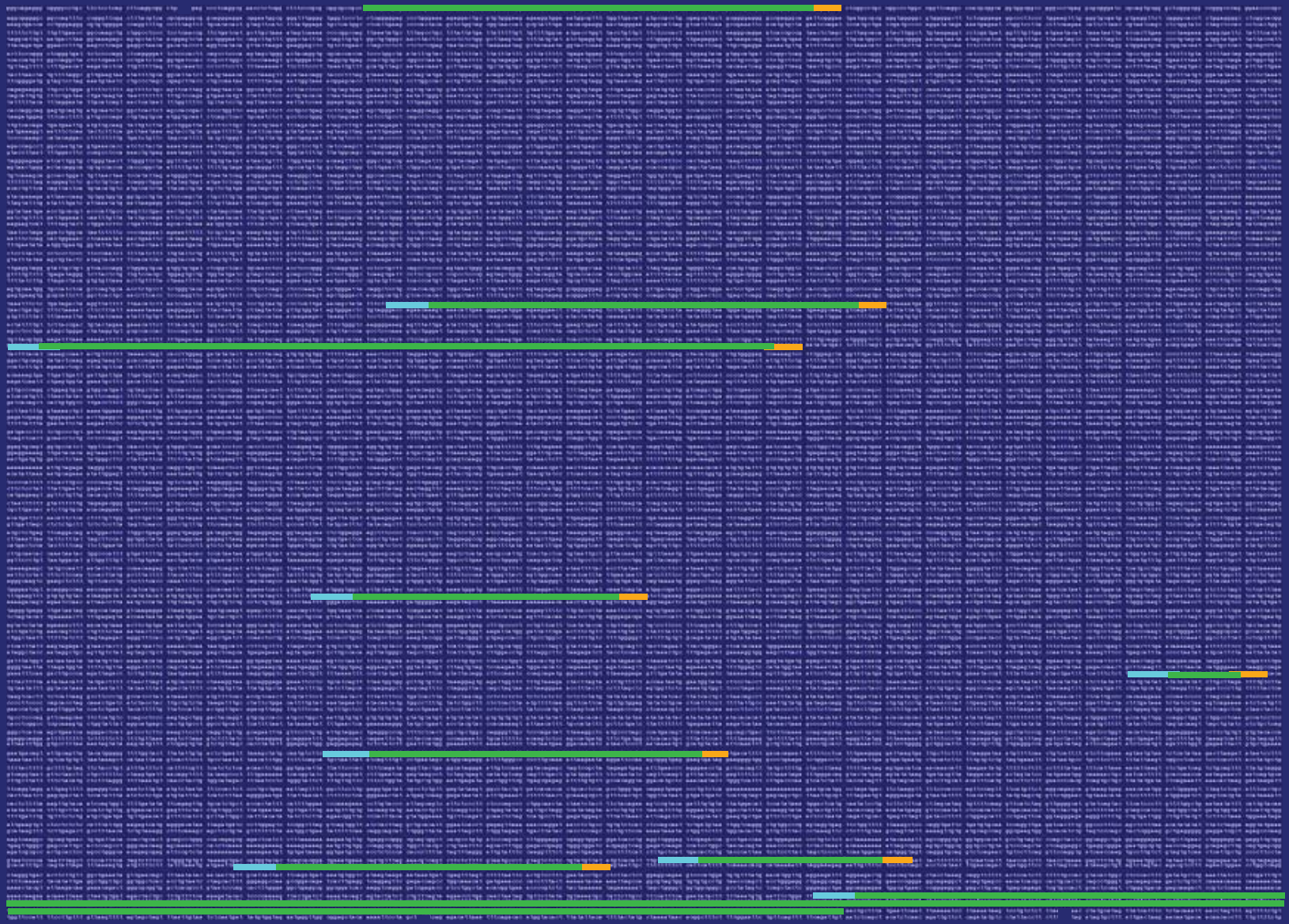



AA TGACCAGTCA ACAGGGGACA TAAAAGTAAT TGGTGGAGAT GATCTCTCAA CTTTAACTGG AAAG
GATATAATT GACACTGGCA AAACAATGCA GACTTTGCTT TCCTTGGTCA GGCAGTATAA TCCAAAGATG GTCAAGGTCG CAAG

# Pseudo 5' Splice Sites and 3' Splice Sites



# Authentic 5' Splice Sites and 3' Splice Sites



# SpliceRack

- Largest and more thoroughly curated collection of splice sites.
- Allows exploration of splice sites in 5 genomes (H. sapiens, M. musculus, D. melanogaster, C. elegans, A. thaliana).
- Uses new methods of classifying splice sites into the four categories.
- Identified rare non-canonical sites, conserved in several species.
- Offers a platform for further genomic exploration.

Sheth et. al., NAR, 2006, http://katahdin.cshl.org:9331/SpliceRack

# Characterizing the Splice sites

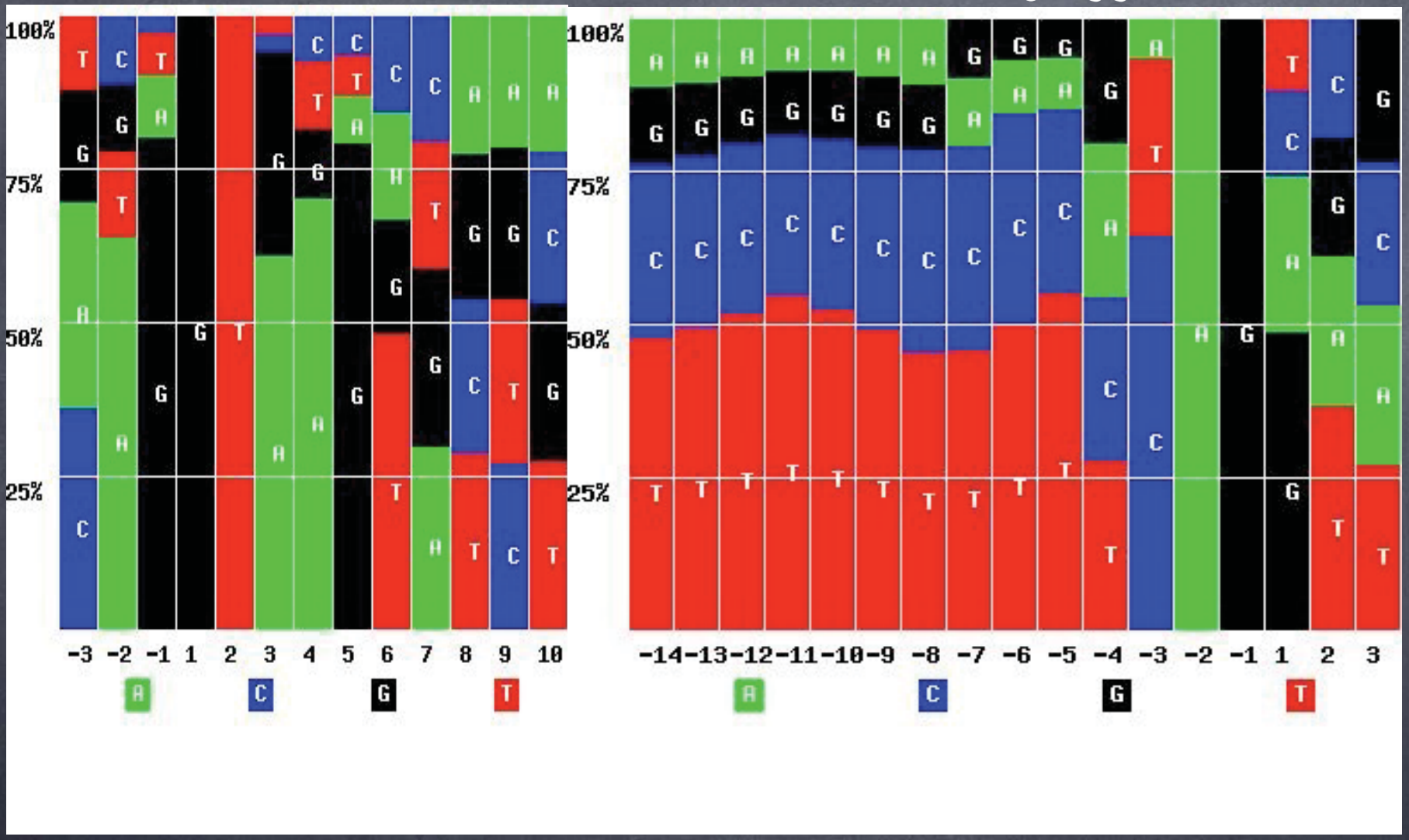
## Position Weight Matrix

| | A | C | G | T |
|---|---|---|---|---|
| −3 | 0.334 | 0.362 | 0.185 | 0.119 |
| −2 | 0.637 | 0.107 | 0.115 | 0.141 |
| −1 | 0.099 | 0.027 | 0.805 | 0.069 |
| 1 | 0.000 | 0.000 | 1.000 | 0.000 |
| 2 | 0.000 | 0.000 | 0.000 | 1.000 |
| 3 | 0.597 | 0.027 | 0.350 | 0.026 |
| 4 | 0.699 | 0.071 | 0.119 | 0.111 |
| 5 | 0.089 | 0.054 | 0.782 | 0.075 |
| 6 | 0.181 | 0.149 | 0.193 | 0.477 |
| 7 | 0.296 | 0.194 | 0.295 | 0.215 |
| 8 | 0.226 | 0.250 | 0.237 | 0.287 |
| 9 | 0.224 | 0.262 | 0.242 | 0.272 |
| 10 | 0.227 | 0.238 | 0.255 | 0.280 |

Occurrence of G at -1

U2: GT_AG 5'ss: Consensus Motif → CAG GTAAGT

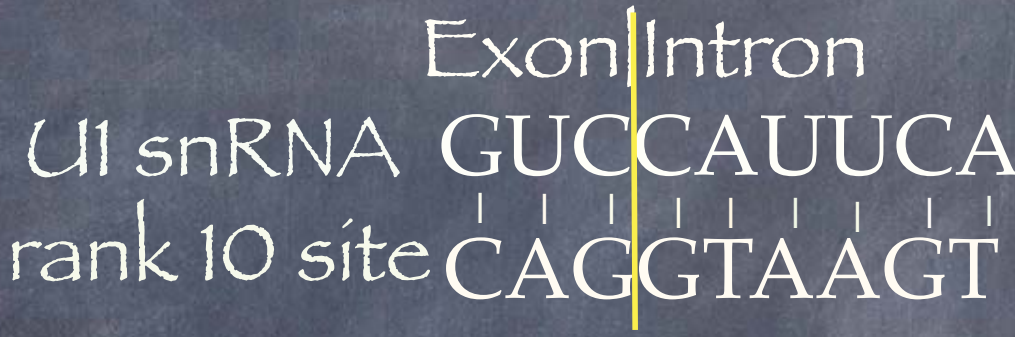
# Position Weight Matrices

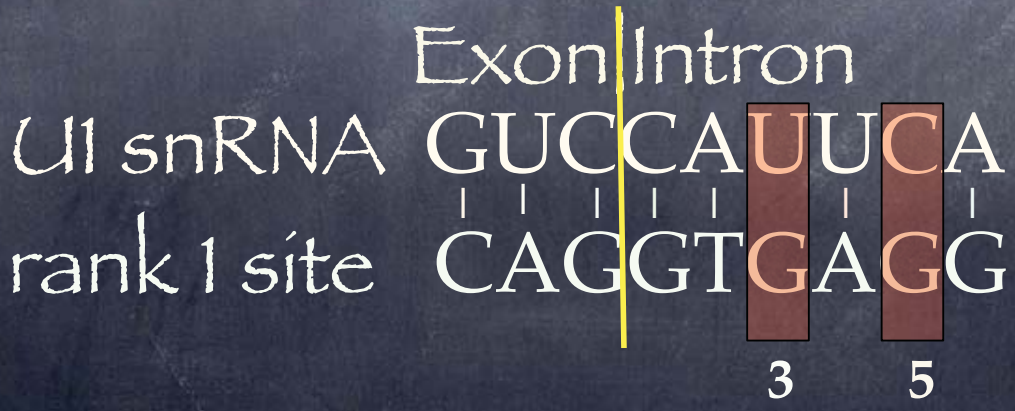



5′ is snRNA-mRNA binding, while 3′ is protein-mRNA binding, less sequence-specific.

# One reason for prediction failure

The weight matrix reflects U1-snRNP 5' splice-site binding



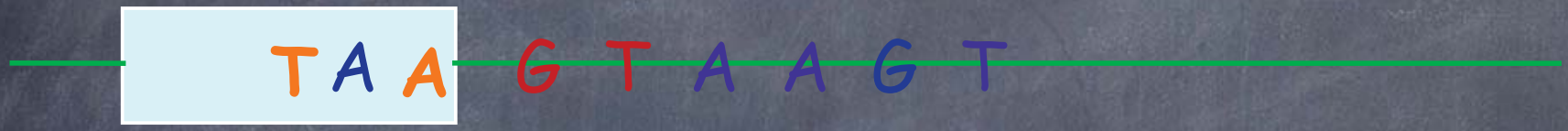



# Summary I

- Collections of splice sites give useful information about the splicing machinery.
- 5′ donor sites show ancient origins, while 3′ acceptor sites show shuffling between U2 and U12 types.
- Weight matrices can explain many features of splice sites, but not all.

# Next Level of splice site characterization

- 2-pt correlation is a natural extension.
- Only feasible at the 5'ss GT-AG-U2 type




Observed frequency = O(-3T,-1A) = 2537

Expected frequency = E(-3T,-1A)

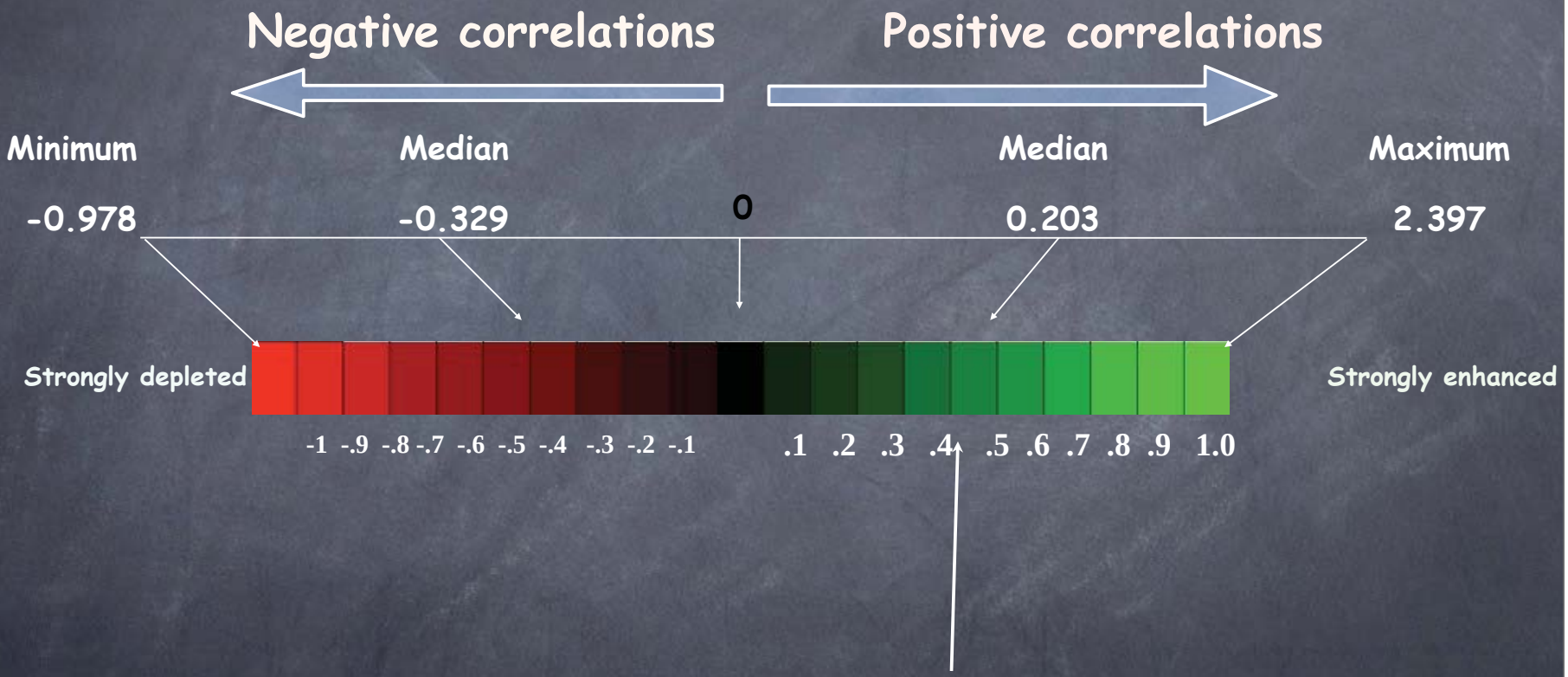
= p(-3T) × p(-1A) × N(5'ss)

= 0.1192 × 0.0995 × 183682

= 2178.11

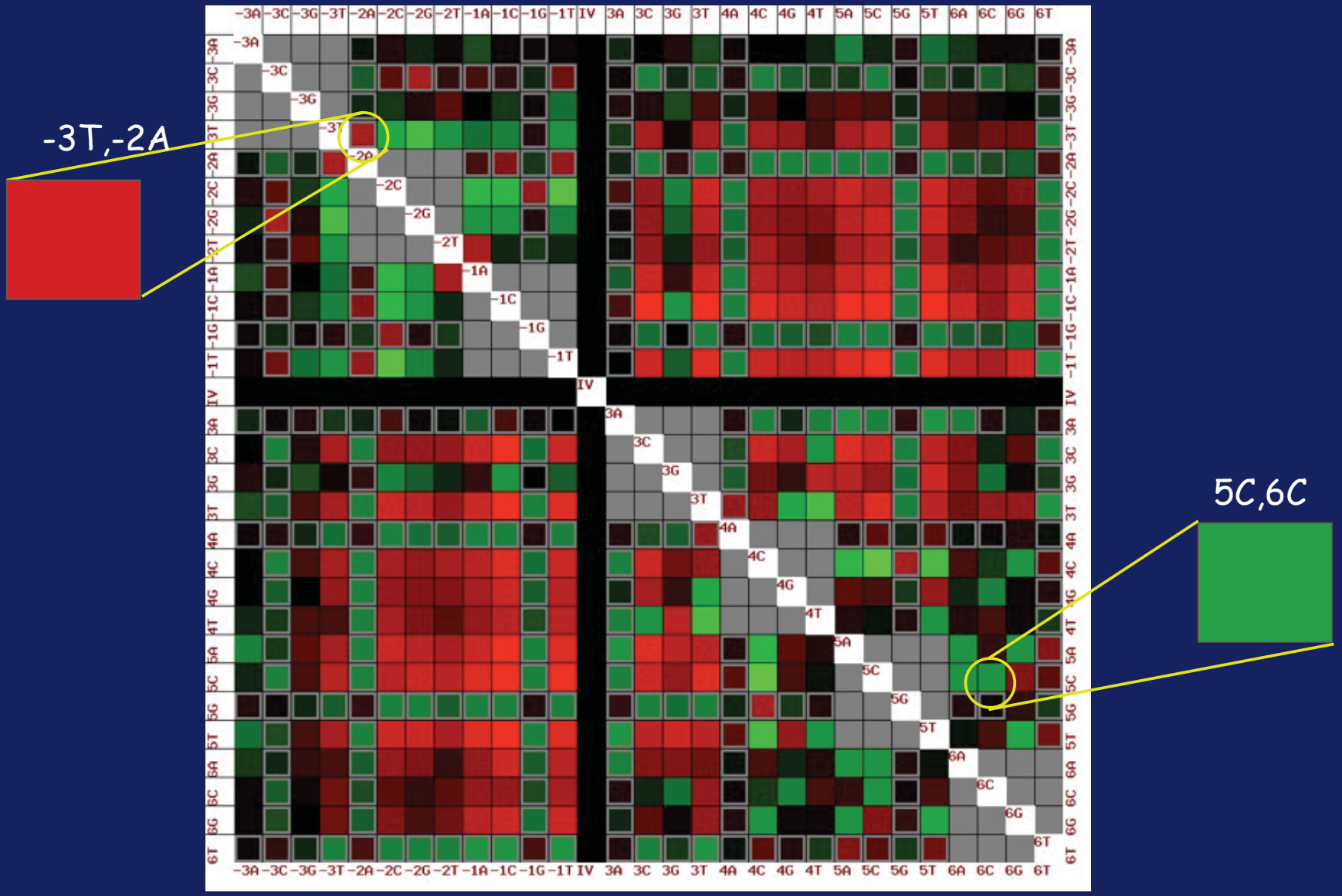
Correlation = (Obs. freq. – Exp. freq)/Exp. Freq = 0.164

# Color-Scheme for Correlations



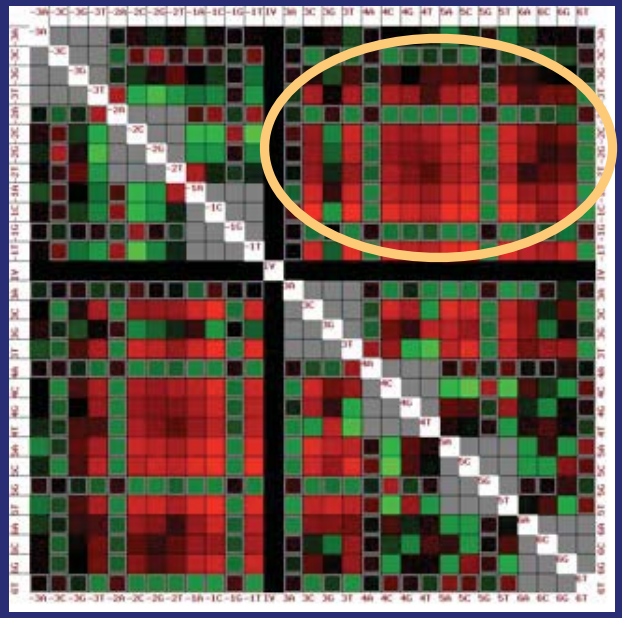

Original Correlation (-3T,-1A) = 0.164 ≈ 0.403 on transformed scale

# Two nucleotide correlation in H. sapiens

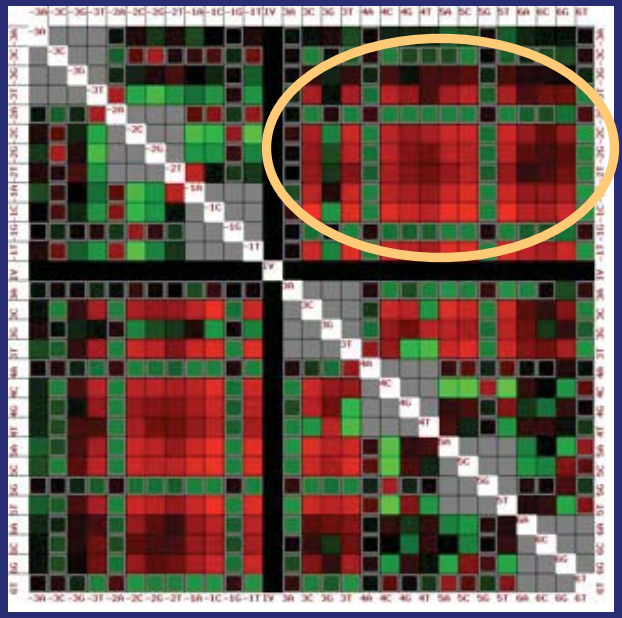



Positions 1 and 2 are Invariant (IV)

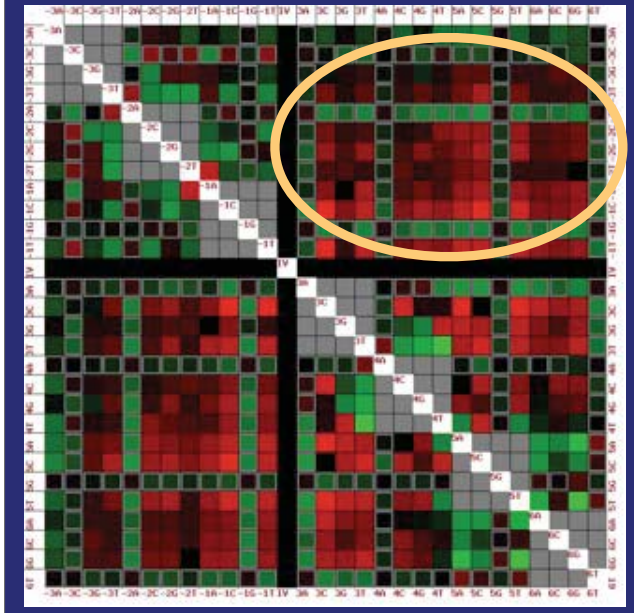
*H.sapiens*



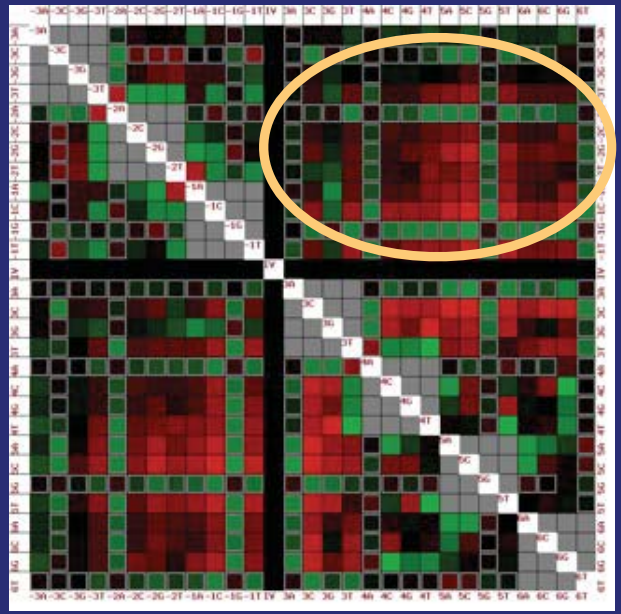
*M.musculus*



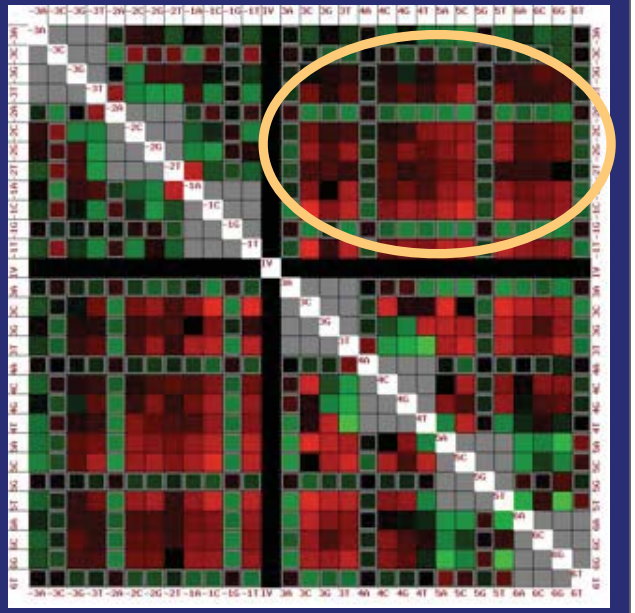
*D.melanogaster*



*C.elegans*

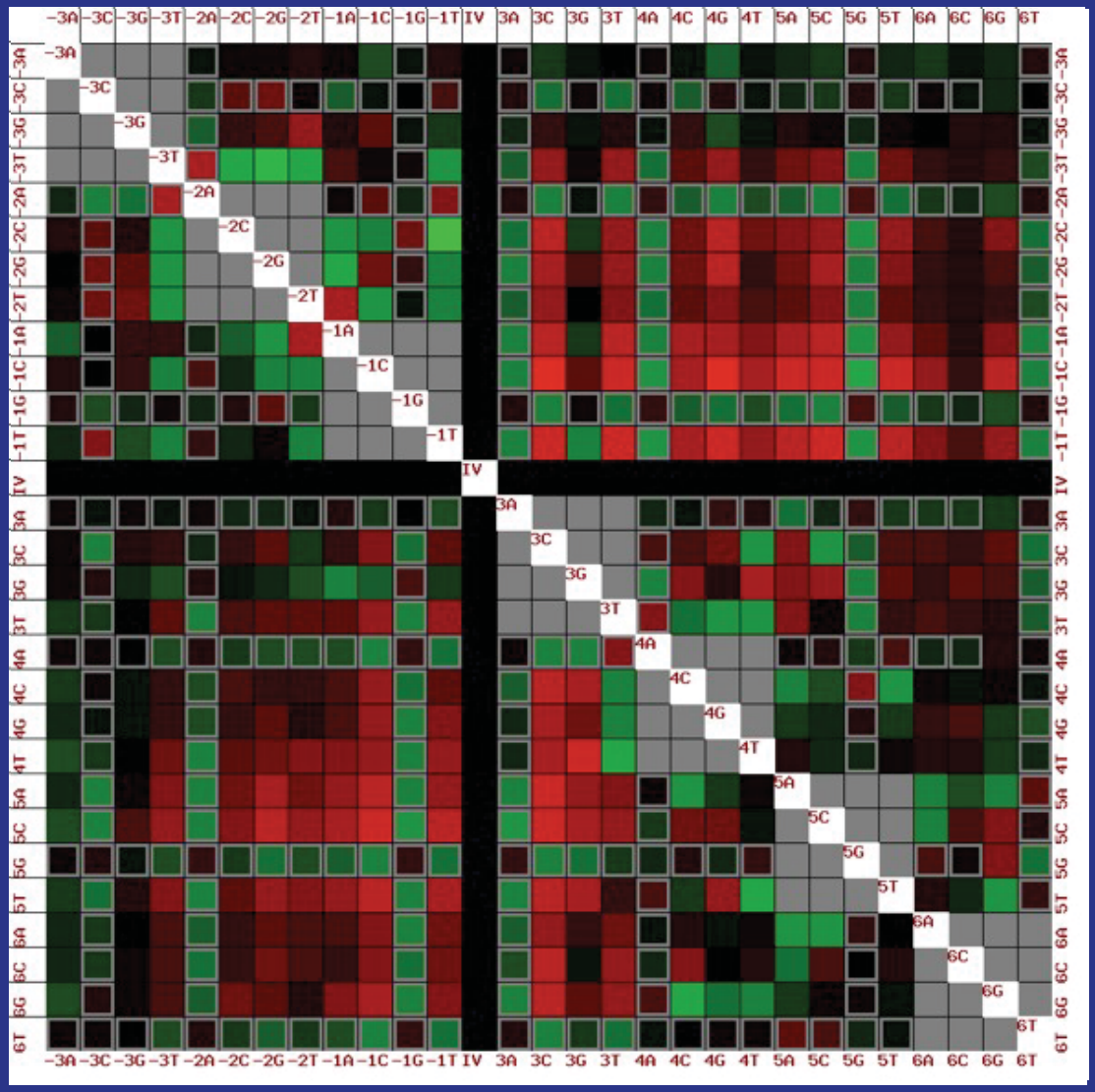


*A.thaliana*



# Cross-species dinucleotide correlations

*A.thaliana*




*C.elegans*

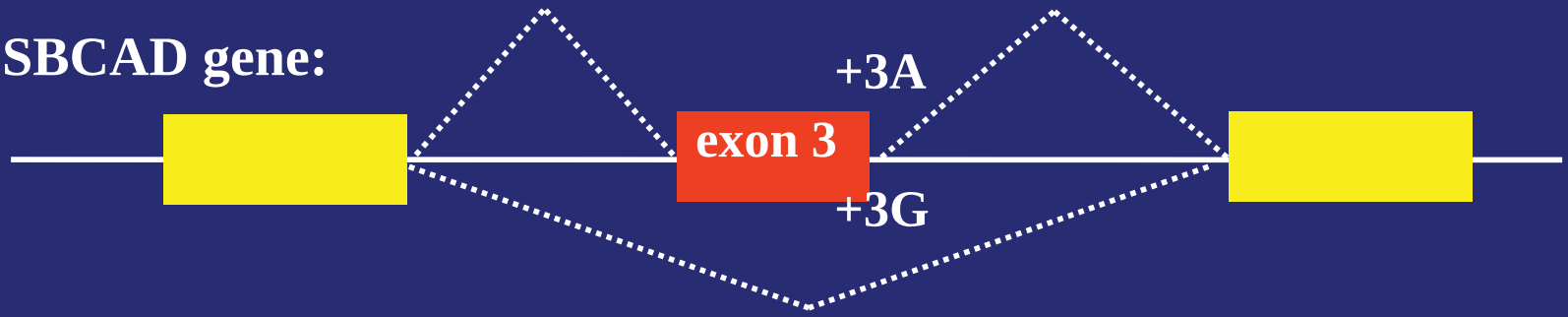
# Are the correlations relevant ?

- Second-order effect, probably visible only in weak splice sites.
- Context is important in splicing, experiments difficult to perform. But Nature provides us with two examples
- Mutations - sporadic changes that cause disorders.
- SNPs - Single Nucleotide Polymorphisms, natural differences at single positions in the DNA between people.

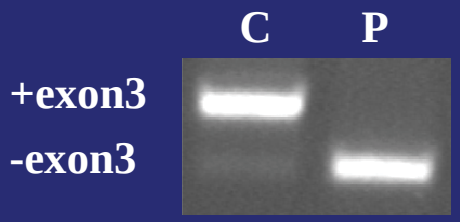
# +3 A-->G mutation in disease

**In collaboration with Brage S. Andresen's lab**

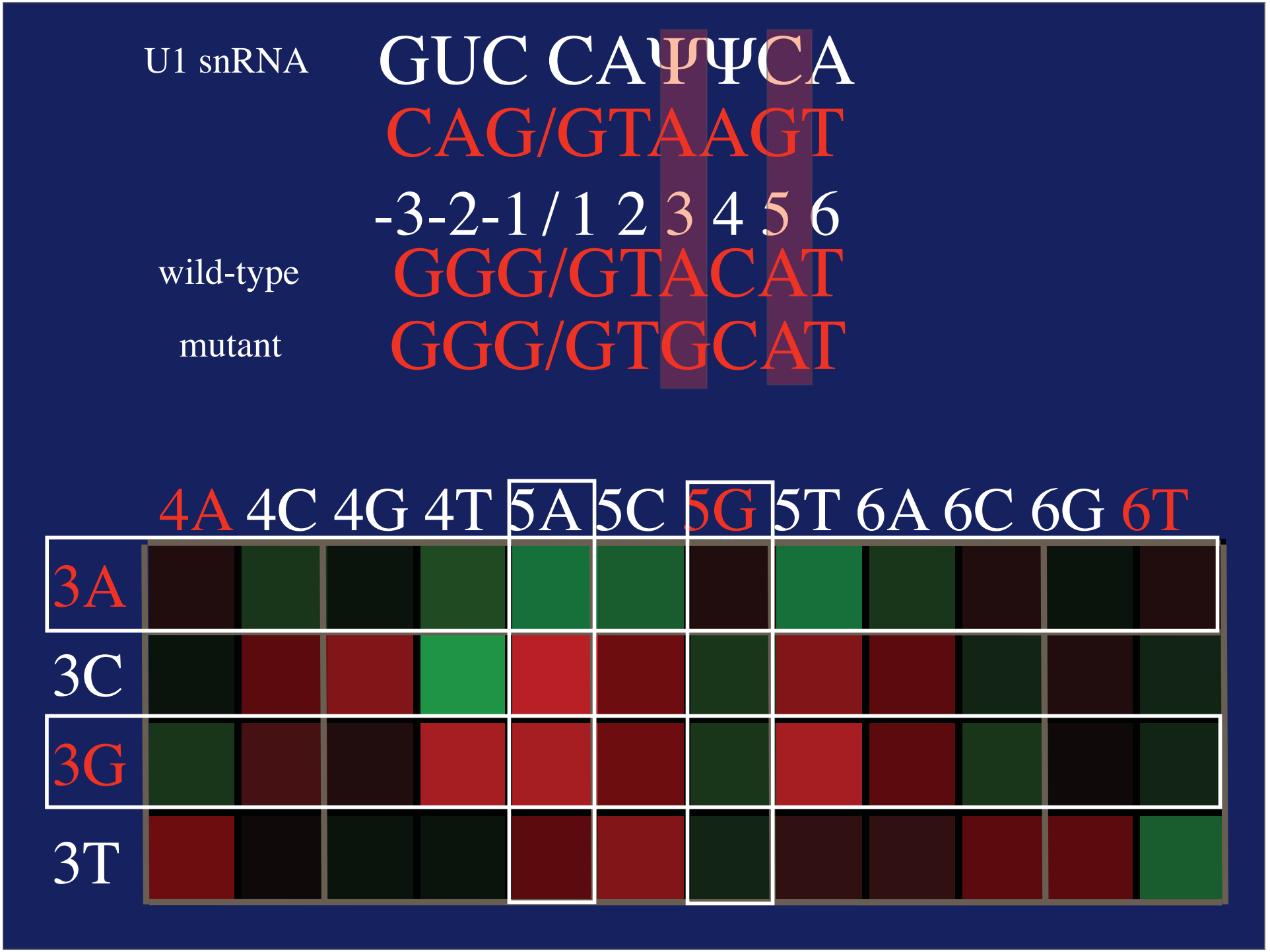
**Research Unit for Molecular Medicine and Institute of Human Genetics, Aarhus University, Aarhus, Denmark**




**SBCADD: Autosomal recessive disorder of L-Isoleucine catabolism**




Madsen et al., HG, 2006

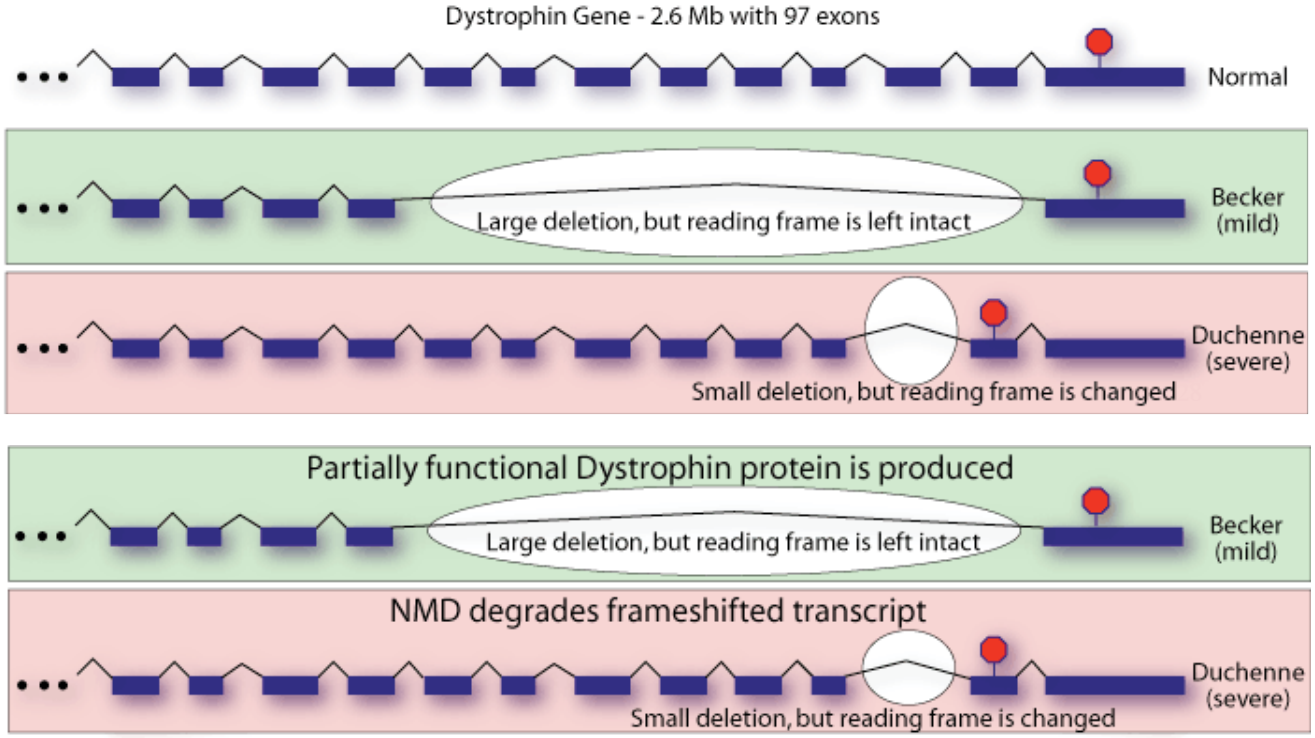

U1 snRNA
GUC CAΨΨCA
CAG/GTAAGT
-3-2-1 / 1 2 3 4 5 6
wild-type
GGG/GTACAT
mutant
GGG/GTGCAT
4A 4C 4G 4T 5A 5C 5G 5T 6A 6C 6G 6T
3A
3C
3G
3T

# Use of splice-site database and tools

- We do better than many machine-learning based approaches.
- We cannot ab-initio predict good splice sites
- BUT, we can tell when a splice-site change will be deleterious, and have explained several rare-genetic disorders using our splice-site database.

# Muscular Dystrophy

- Genetic Disorder leading to muscle wasting
- Defects in Dystrophin, helps connect cytoskeleton of muscle fibre to the surrounding ECM.
- on X, 2.6MB gene, 97 exons, 16 hours to transcribe, mRNA is 14kb, protein 3500

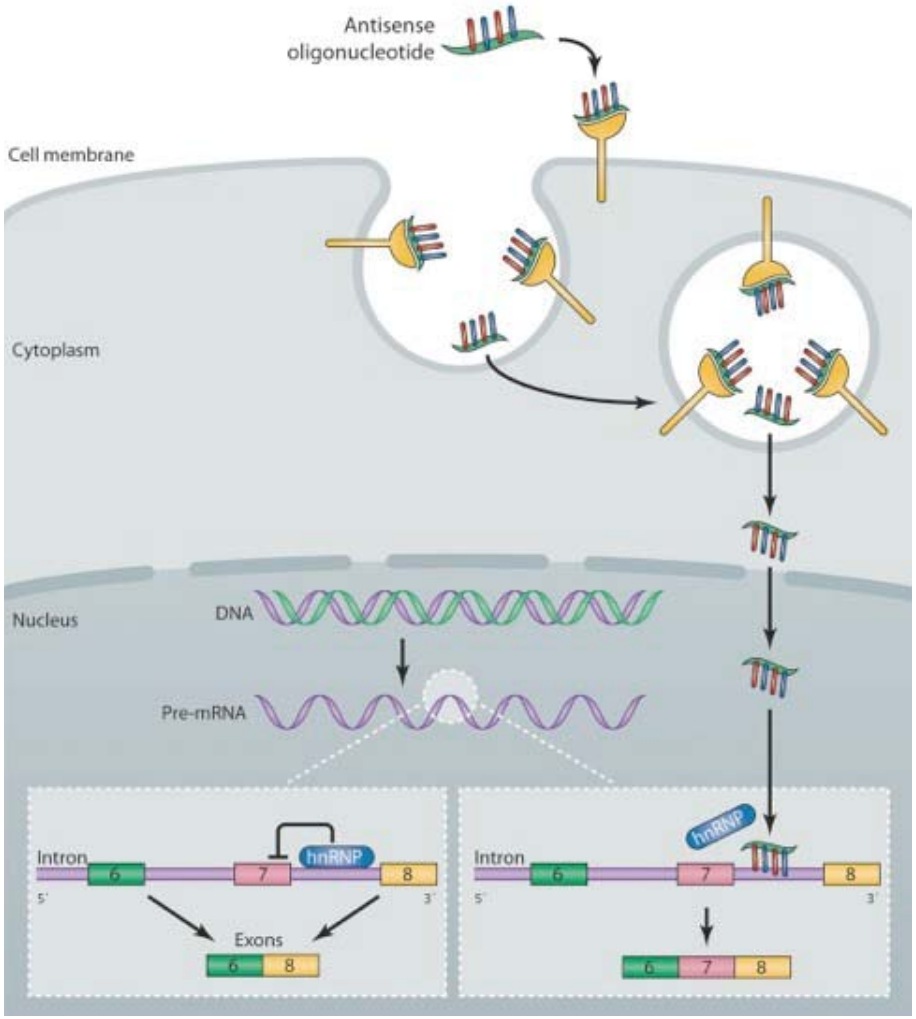


# Spinal Muscular Atrophy (SMA)

SMA is caused by the homozygous loss of the survival of motor neuron 1, telomeric (SMN1) gene; either by deletion or rarely by mutation. Leads to death of affected patient, in childhood.

Humans have a paralogue gene called SMN2, also on chromosome 5, which differs from SMN1 by 11 nucleotides but has an identical coding sequence. One of the nucleotide changes between SMN1 and SMN2 genes is a C-to-T transition within exon 7, and although it is a synonymous change, it weakens the 3′ splice site, resulting in skipping of exon 7. Some good transcripts are made. Can you make SMN2 take over the role of SMN1 ?

## Mechanism of action of an antisense drug that modulates SMN2 splicing.




**Rigo F et al. J Cell Biol 2012;199:21-25**

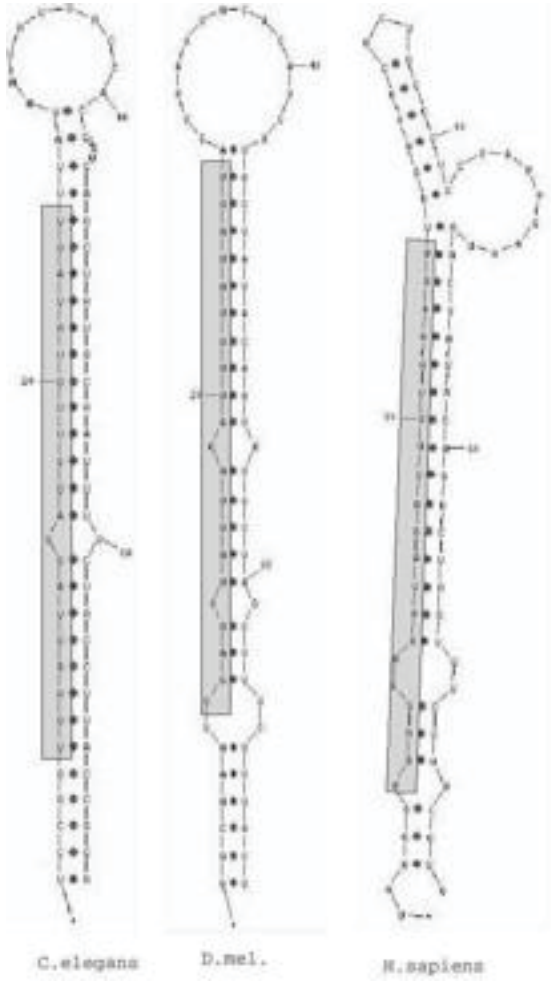


JCB

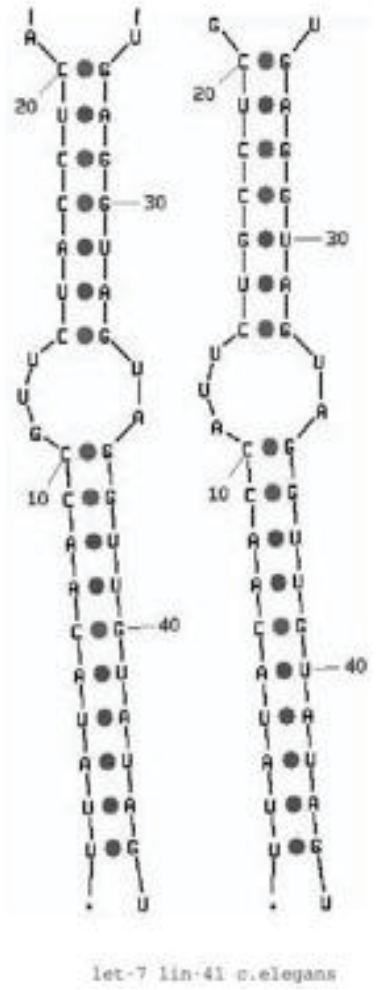
# What is RNAi ?

- RNA interference.
- Mechanism for gene regulation.
- PTGS (non-coding genes, miRNAs).
- TGS (epigenetic modifications).
- Silencing of transposons, stem-cell differentiation etc. Other components of the pathway (e.g. piRNAs)

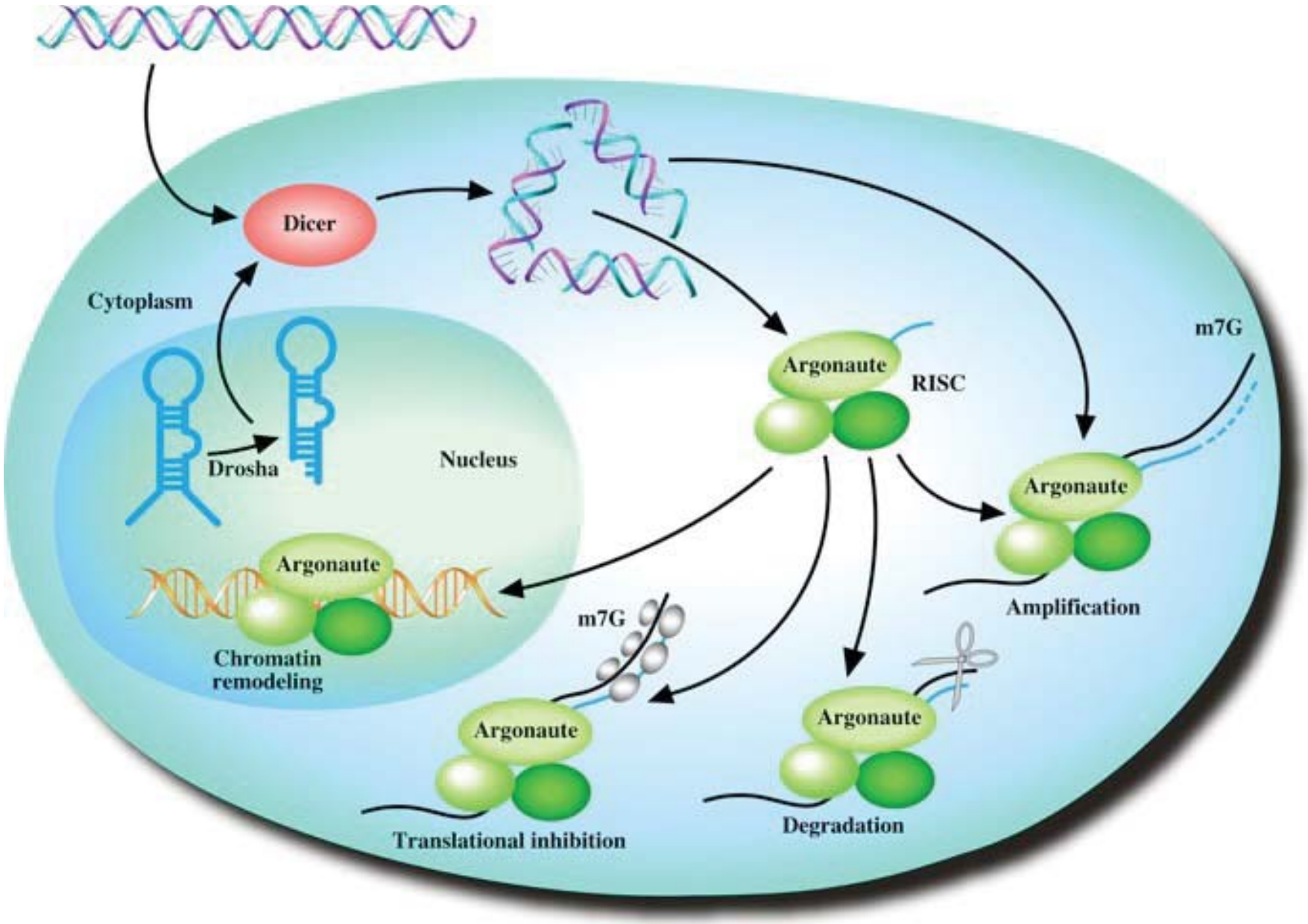
# Non-coding silencing genes

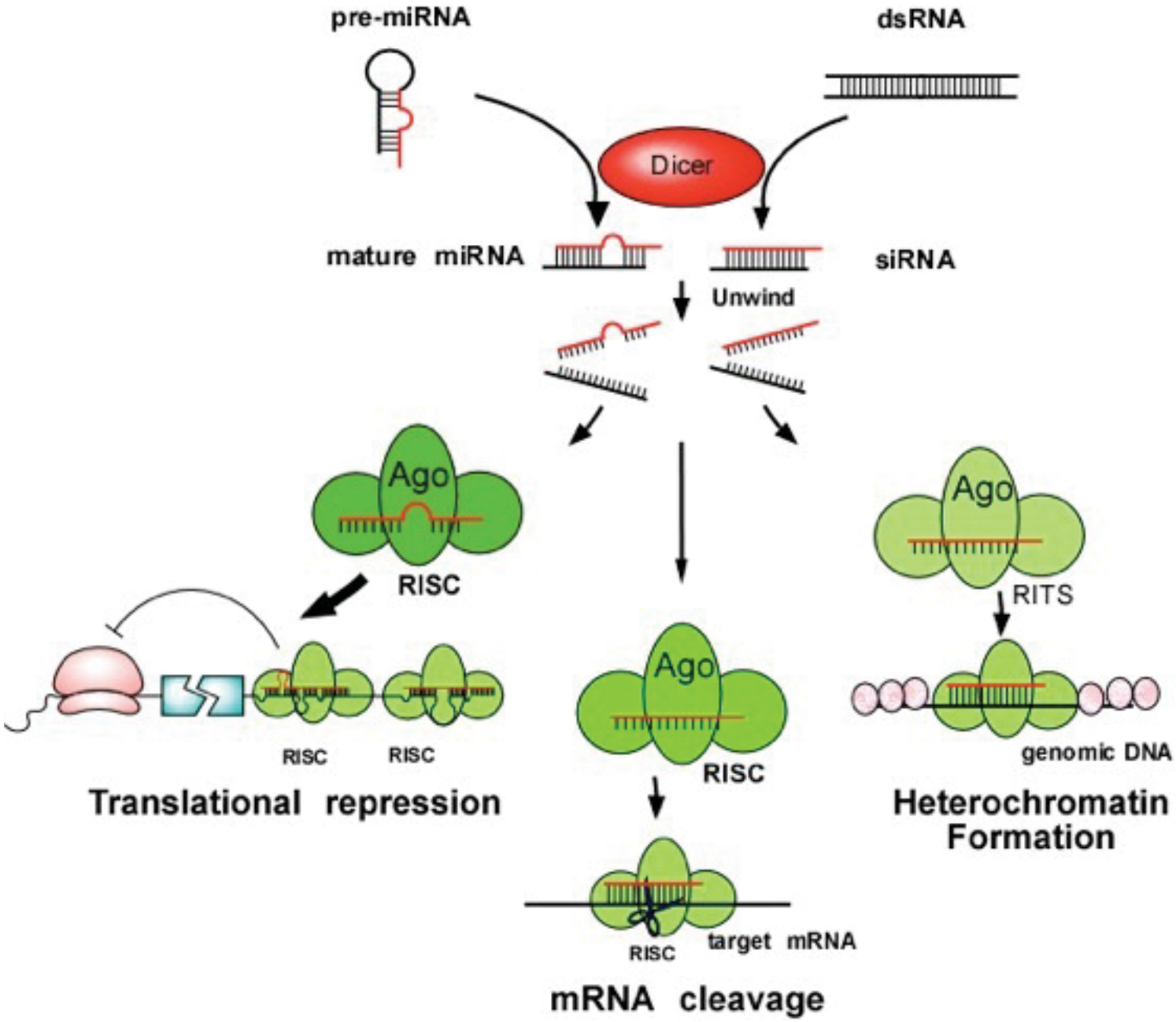



let-7 genes in several species

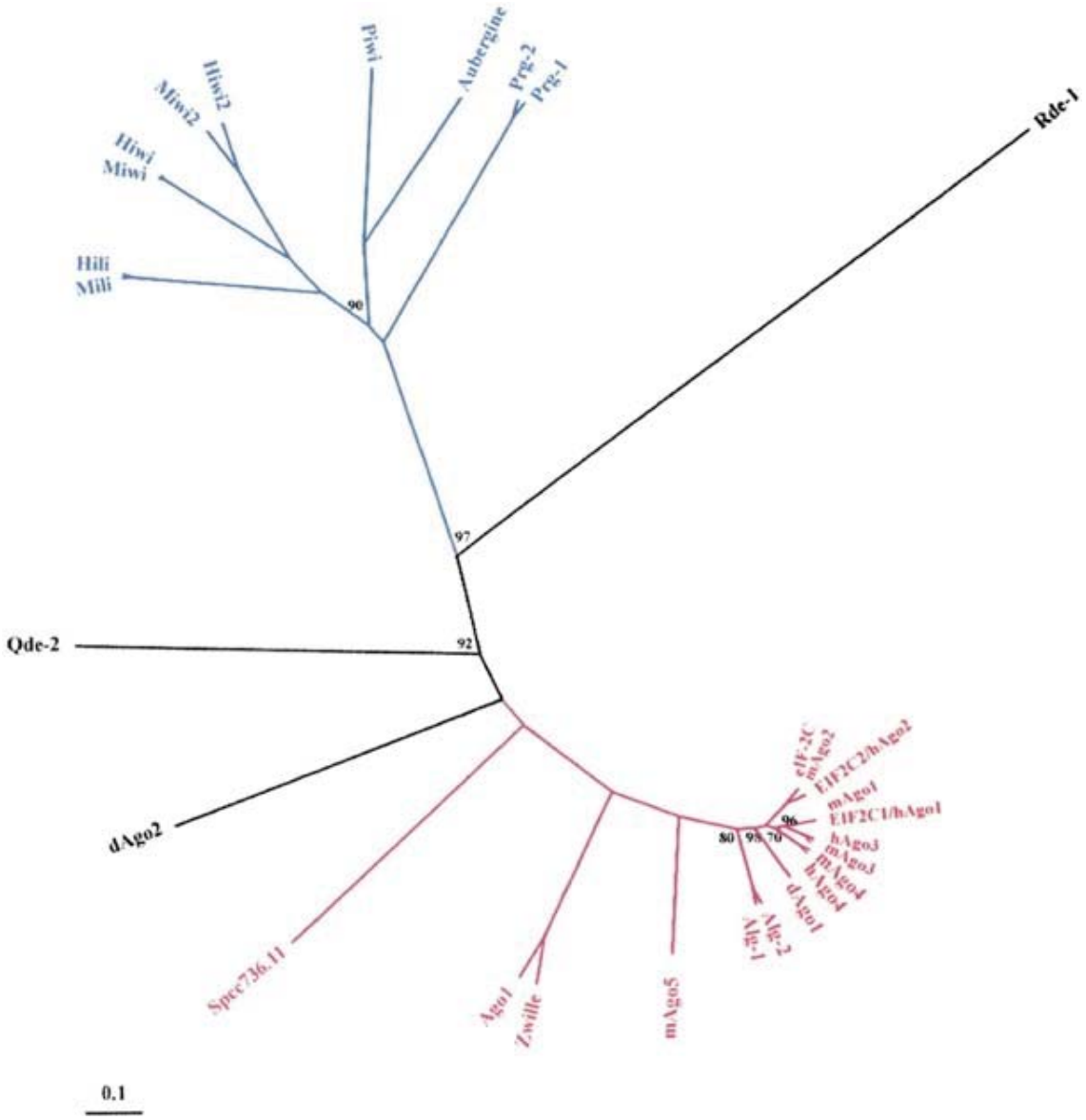


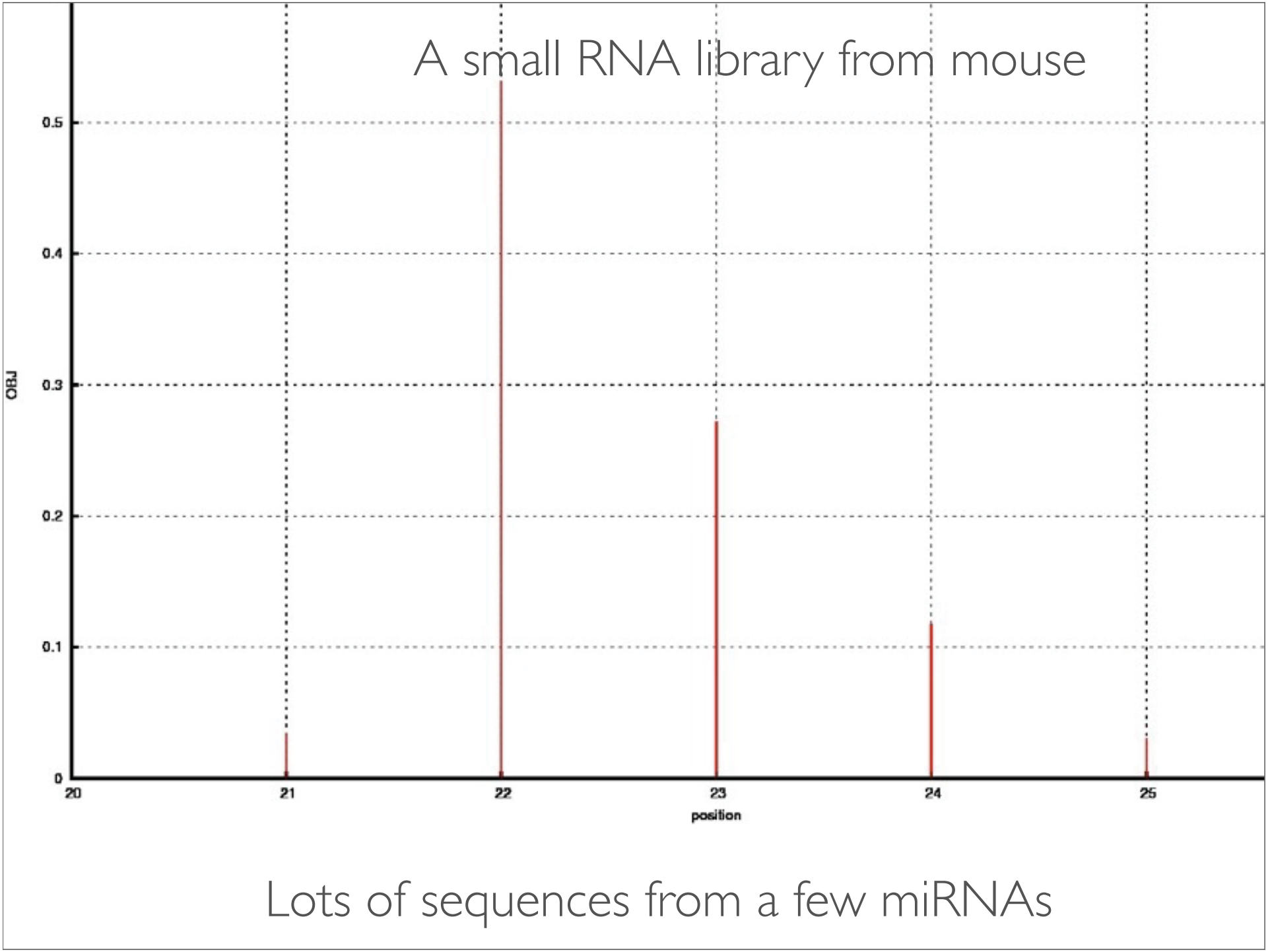

let-7 binding sites in 3' UTR
of C.elegans lin-41


Dicer
Cytoplasm
Drosha
Nucleus
Argonaute
RISC
m7G
Argonaute
Amplification
Argonaute
Chromatin
remodeling
m7G
Argonaute
Translational inhibition
Argonaute
Degradation

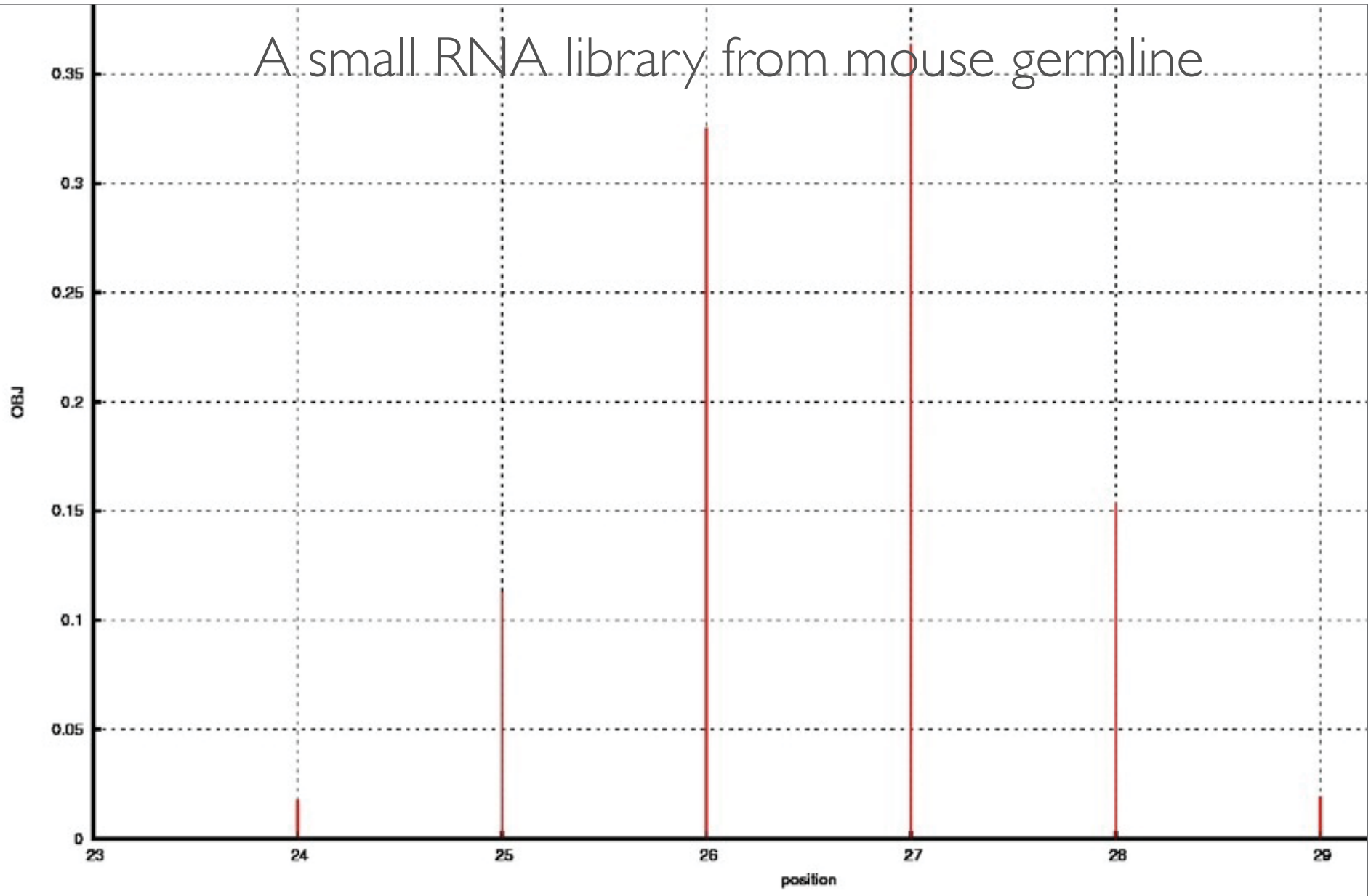

pre-miRNA
dsRNA
Dicer
mature miRNA
siRNA
Unwind
Ago
RISC
Ago
RITS
Ago
RISC
RISC
RISC
genomic DNA
Translational repression
RISC
target mRNA
Heterochromatin Formation
mRNA cleavage

# The Piwi family



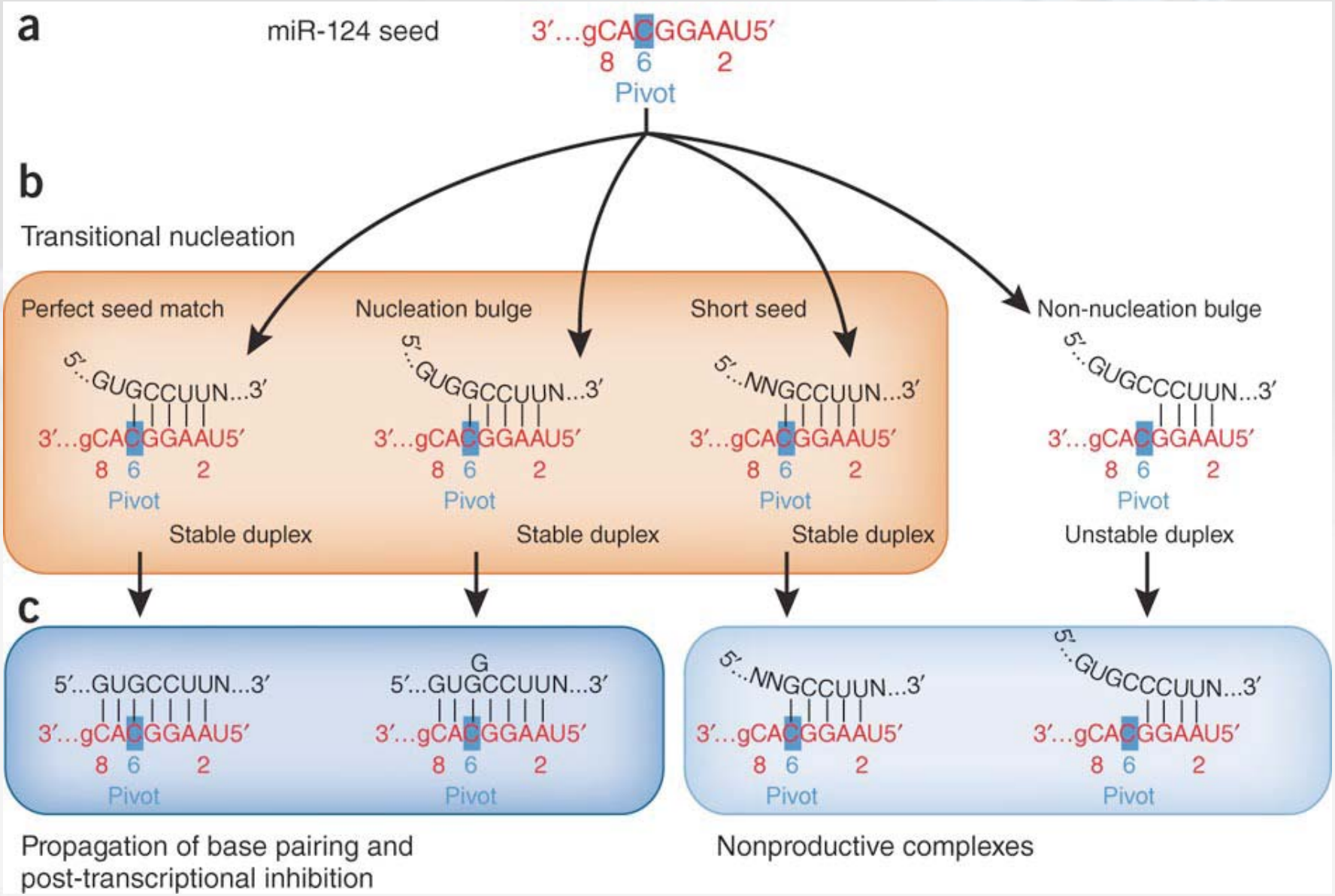

A small RNA library from mouse
0.5
0.4
0.3
0.2
0.1
0
OBJ
20
21
22
23
24
25
position
Lots of sequences from a few miRNAs

# A small RNA library from mouse germline



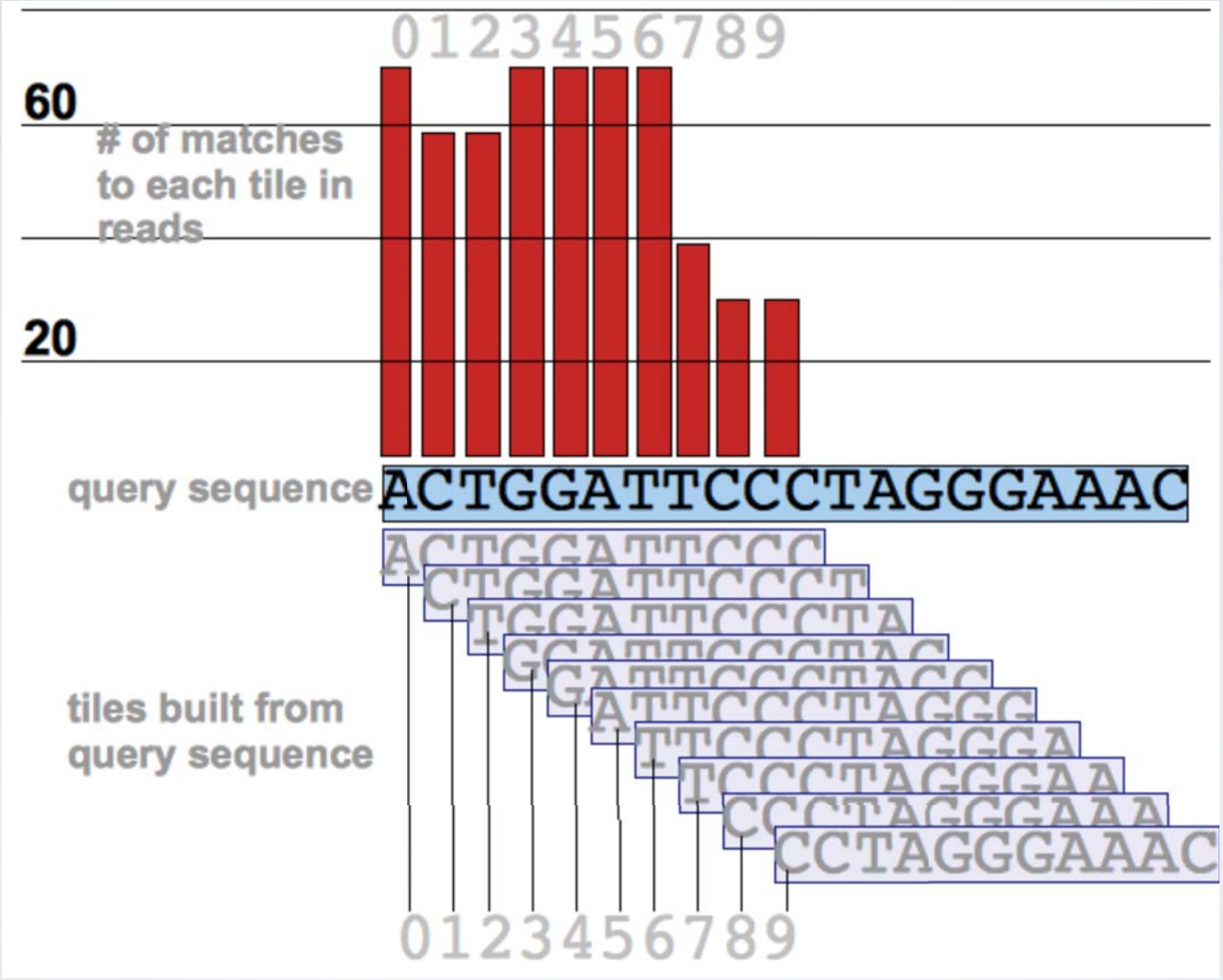

Lots of sequences from a millions of piRNAs

# miRNA Targeting rules



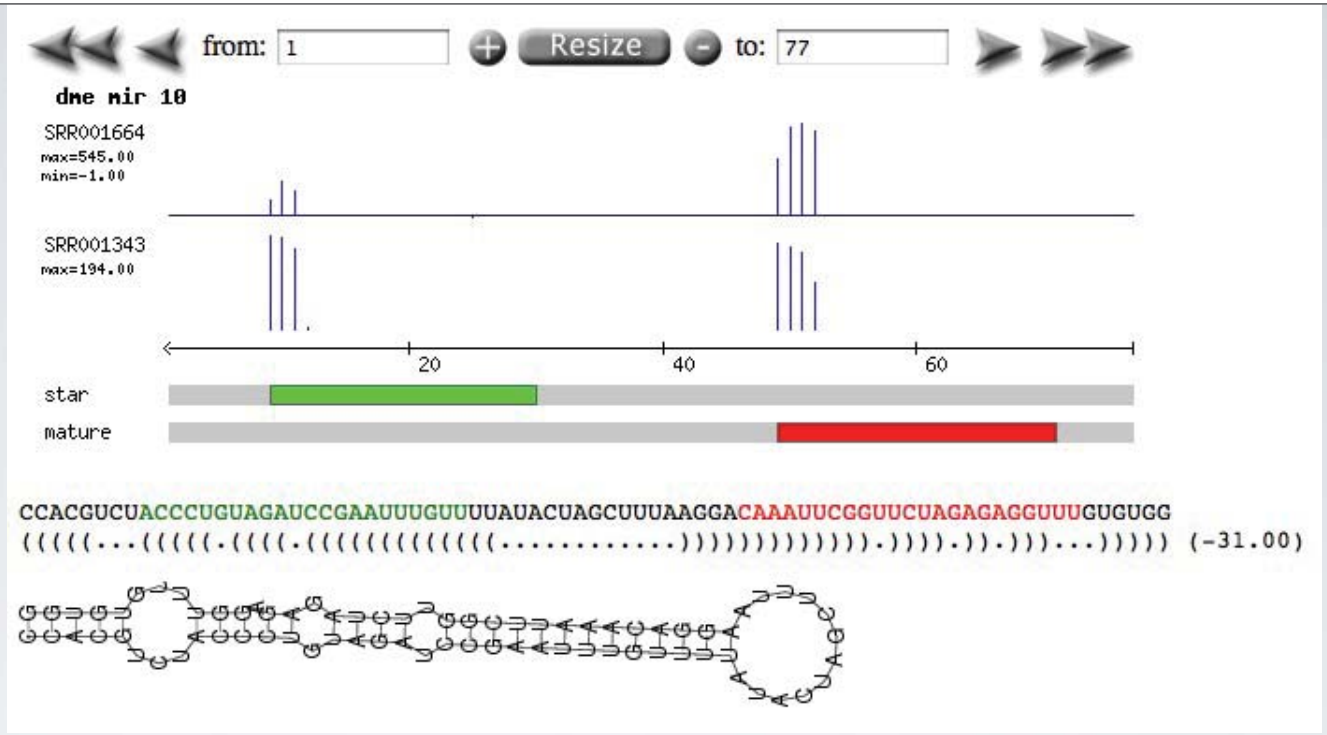
# GeoSeq-Algorithm

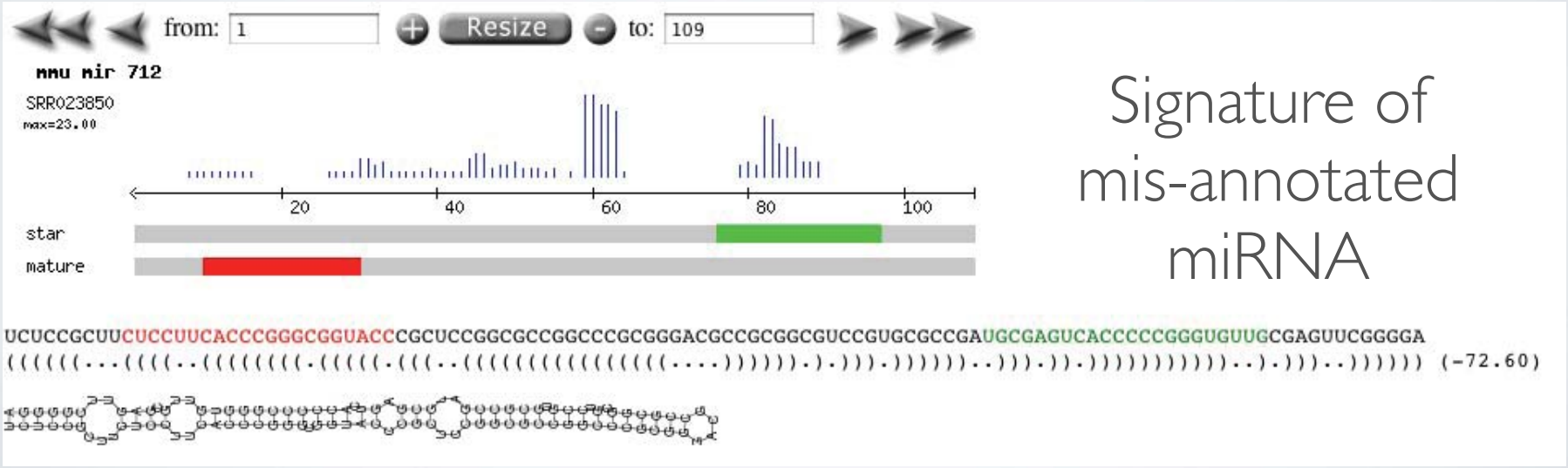
## used to detect spurious miRNA in miRBASE




Gurtowski J, Cancio A, Shah H, Levovitz C, George A, Homann R, Sachidanandam R.
*Geoseq: a tool for dissecting deep-sequencing datasets.*
BMC Bioinformatics.2010 Oct 12;11:506. PubMed PMID: 20939882;

Signature of correctly annotated miRNA

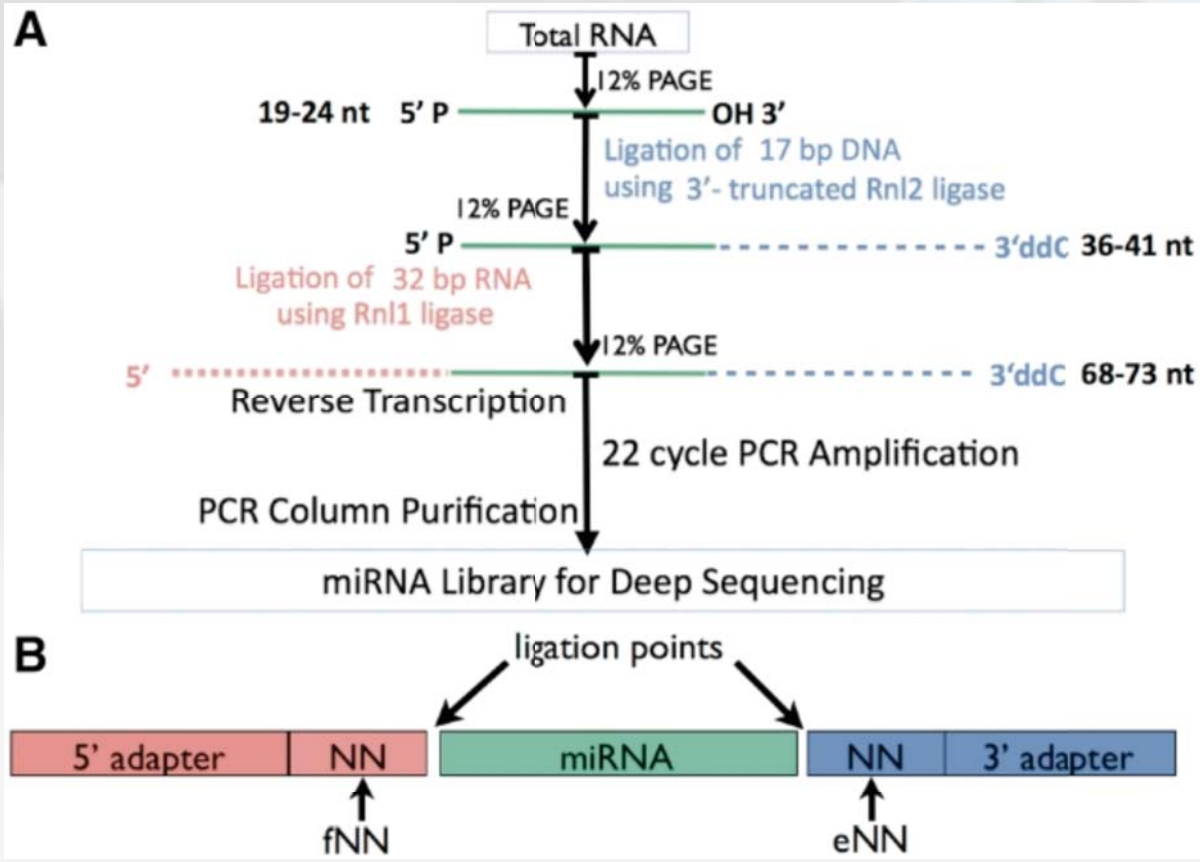



Signature of mis-annotated miRNA

Gurtowski J, Cancio A, Shah H, Levovitz C, George A, Homann R, Sachidanandam R.
*Geoseq: a tool for dissecting deep-sequencing datasets.*
BMC Bioinformatics.2010 Oct 12;11:506. PubMed PMID: 20939882;

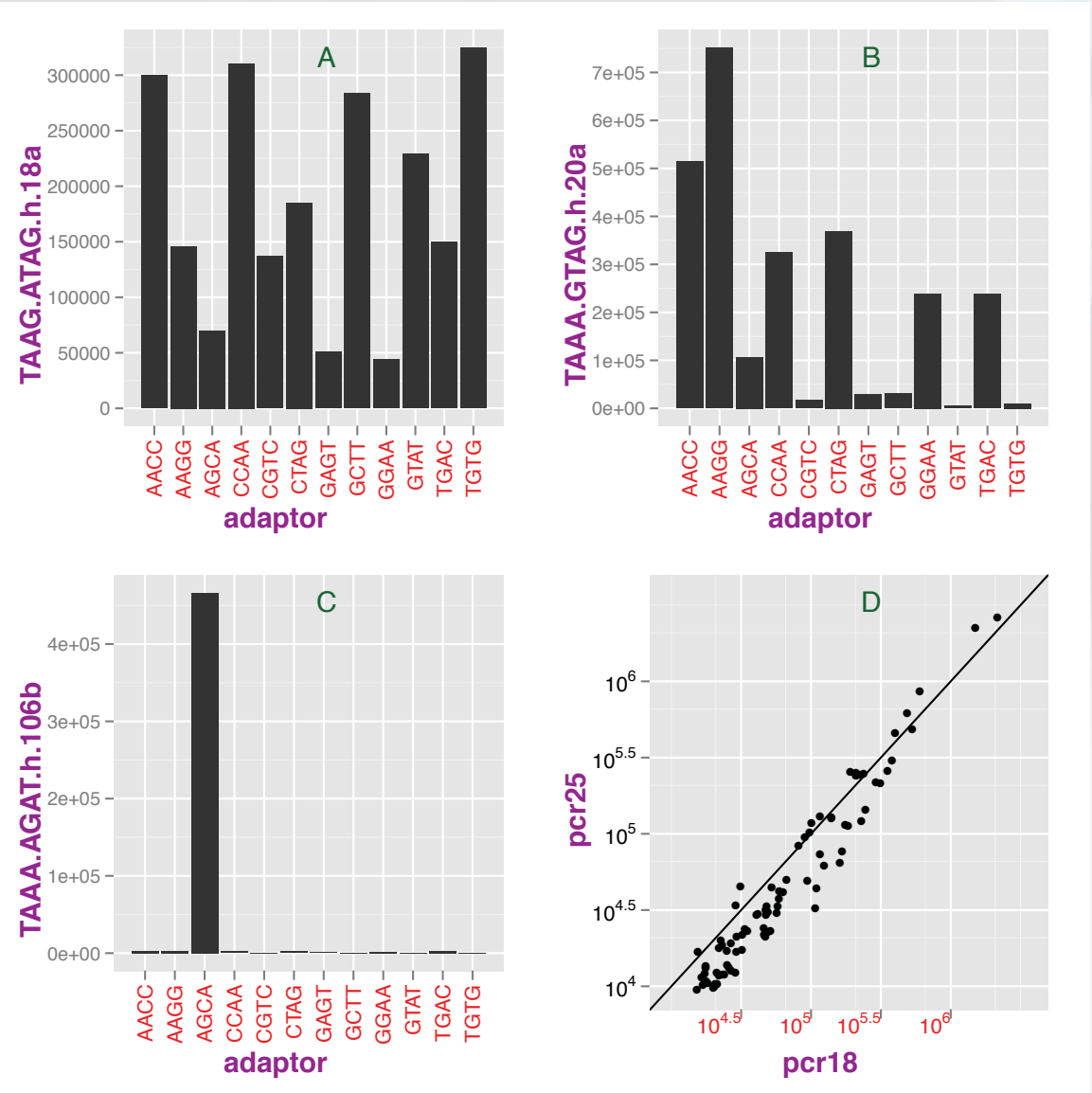
# Protocol to prepare small RNAs for deep-sequencing





Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R.
*Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing.*
Nucleic Acids Res. 2011 Nov;39(21):e141. Epub 2011 Sep 2.

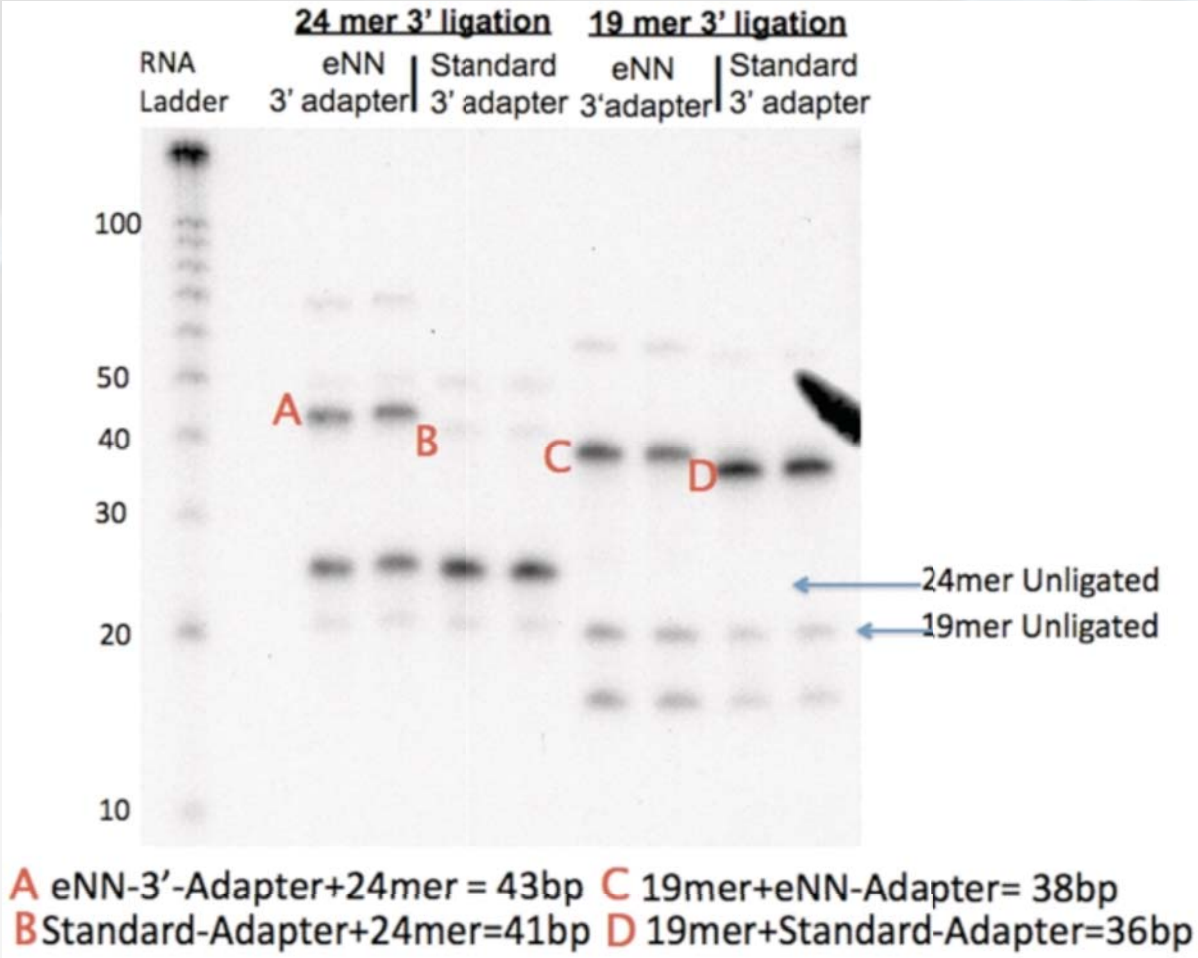
# Different adaptor sequences give different results from the same sample





Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R.
*Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing.*
Nucleic Acids Res. 2011 Nov;39(21):e141. Epub 2011 Sep 2.

# A gel-shift to show 3'-adapter ligations to a synthetic construct depend on sequence of the adapter

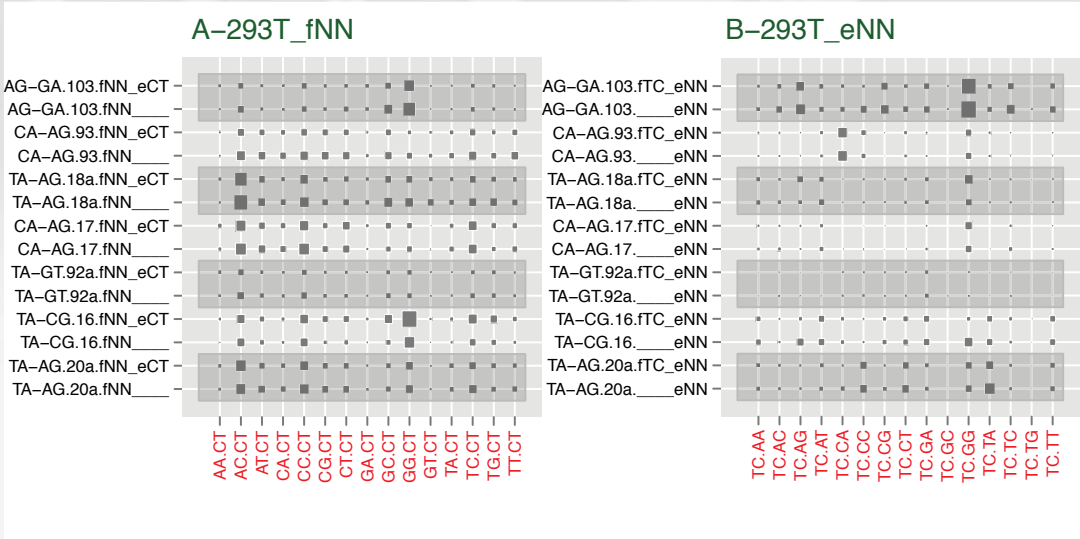




Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R.
*Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing.*
Nucleic Acids Res. 2011 Nov;39(21):e141. Epub 2011 Sep 2.

# Modeling the ligation efficiency and experimental verification of the model

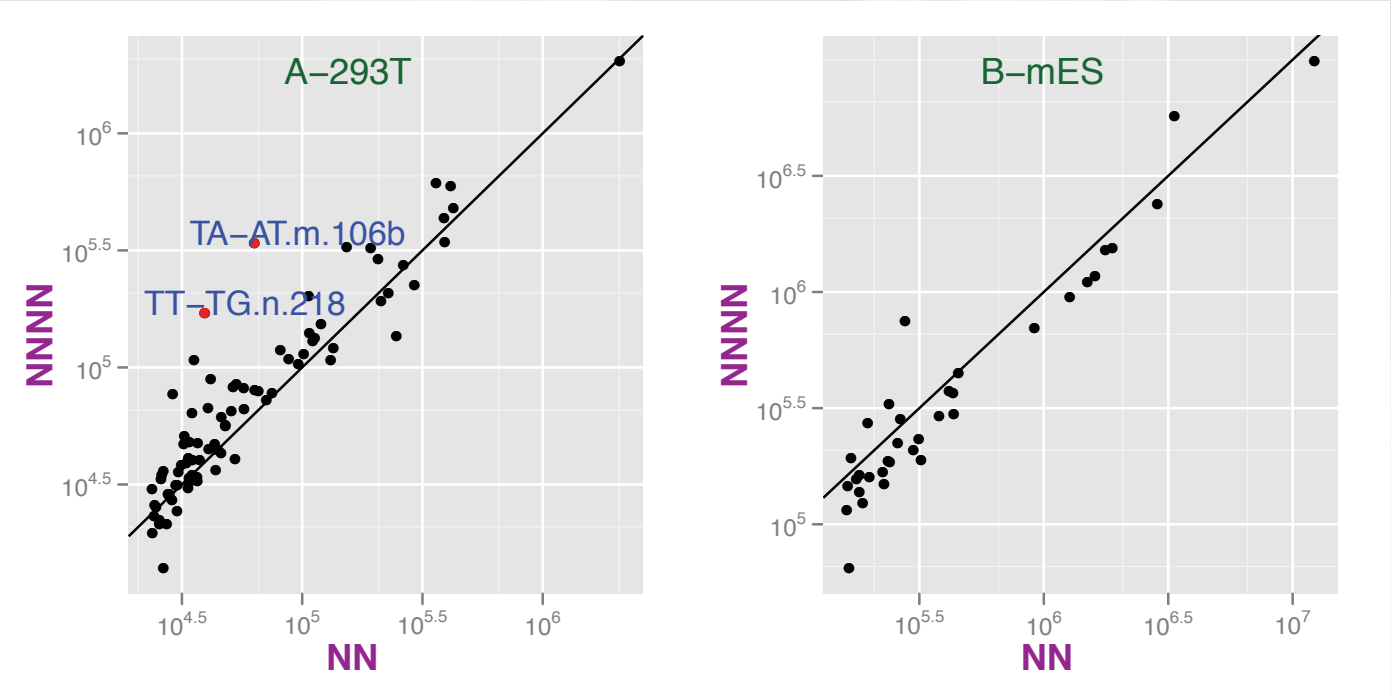
$$m_{in}^{k} = F_{ij} * M^{k} * E_{mn}$$





Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R.
*Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing.*
Nucleic Acids Res. 2011 Nov;39(21):e141. Epub 2011 Sep 2.

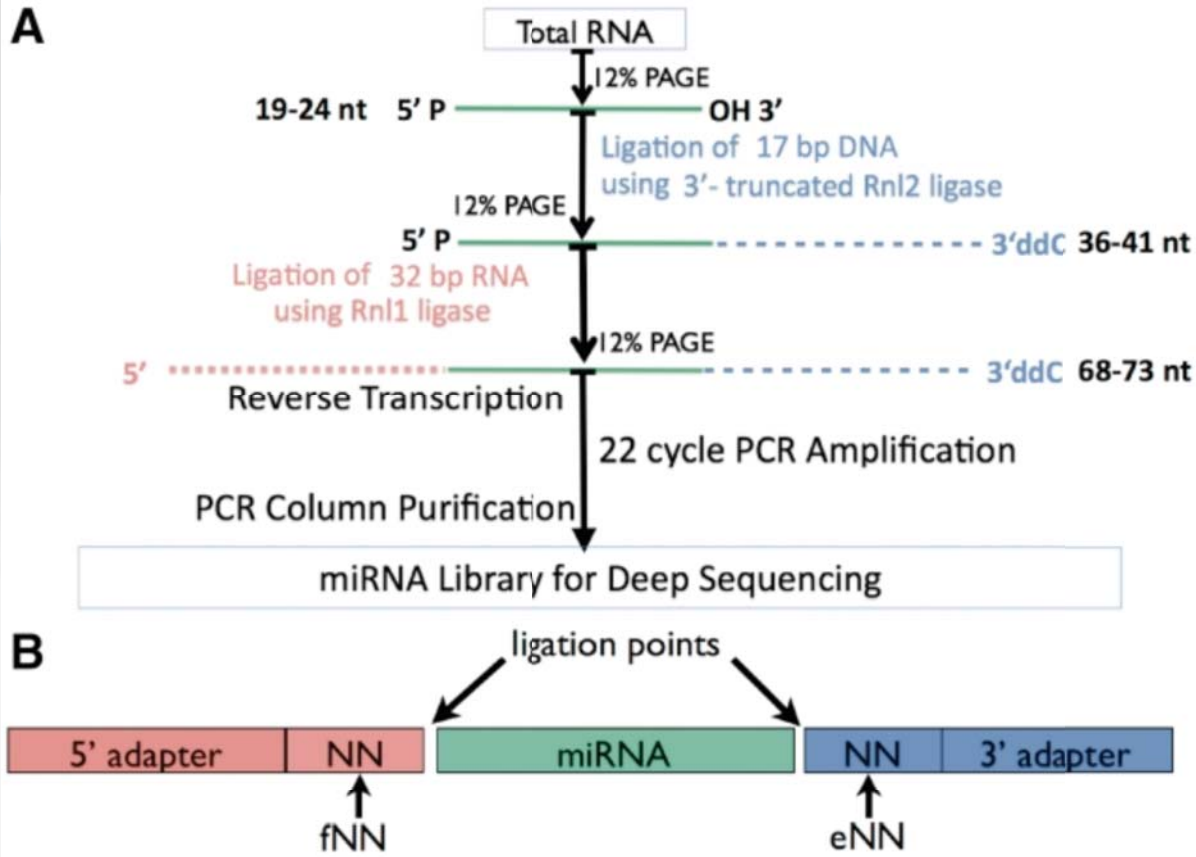
# Most variability comes from the 2 nucleotides at the adapter's ligating ends





Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R.
*Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing.*
Nucleic Acids Res. 2011 Nov;39(21):e141. Epub 2011 Sep 2.

# Final improved protocol for unbiased small RNA sequencing




Use of random-ends (N's) on the adapter averages biases


Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R.
*Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing.*
Nucleic Acids Res. 2011 Nov;39(21):e141. Epub 2011 Sep 2.

# Measurement of dme-mir-6 from early drosophila embryos in various experiments over the years.

| Publication | 3' adapter | 5' adapter | Platform | 3'-adapter ligation | 5'-adapter ligation | Result |
|---|---|---|---|---|---|---|
| Elbashir *et al.*, G&D, 2001[27] | UUUaaccg (RNA/dna hybrid) | tcactAAA (RNA/dna hybrid) | Sanger | T4 RNA-ligase | T4 RNA ligase | miR-6-3p **high**, miR-6-1,2,3-5p **low** |
| Aravin *et al.* Dev. Cell, 2003 [28] | UUUaaccg (RNA/dna) | tcactAAA (RNA/dna) | Sanger | T4 RNA-ligase | T4 RNA ligase | miR-6-3p **high**, miR-6-1,2,3-5p **low** |
| Ruby *et al.* Genome Res., 2007 [29] | CTGTAGG(DNA) | UGAAA (RNA) | 454 | T4 RNA-ligase | T4 RNA ligase | miR-6-3p **high**, miR-6-1,2,3-5p **low** |
| Wang and Liu ★ Frontiers in Gen., 2011[30] | CTGTAGG(DNA) adenylated 5' dideoxy | 3' CGCAUC (RNA) | GAII-(Solexa) | Truncated T4 RNA-ligase | T4 RNA ligase | miR-6-3p **low**, miR-6-1,2-5p **low**, miR-6-3-5p **high** |

★ mod-ENCODE data is inconsistent with other data

| NAME | SEQUENCE |
|---|---|
| dme-miR-6-1-5p (star-1) | agggaauaguugcugugcugua |
| dme-miR-6-2-5p (star-2) | agggaacuucugcugcugauaua |
| dme-miR-6-3-5p (star-3) | agggaacgguugcugaugaugua |
| dme-miR-6-3p (mature) | uaucacaguggcuguucuuuuu |

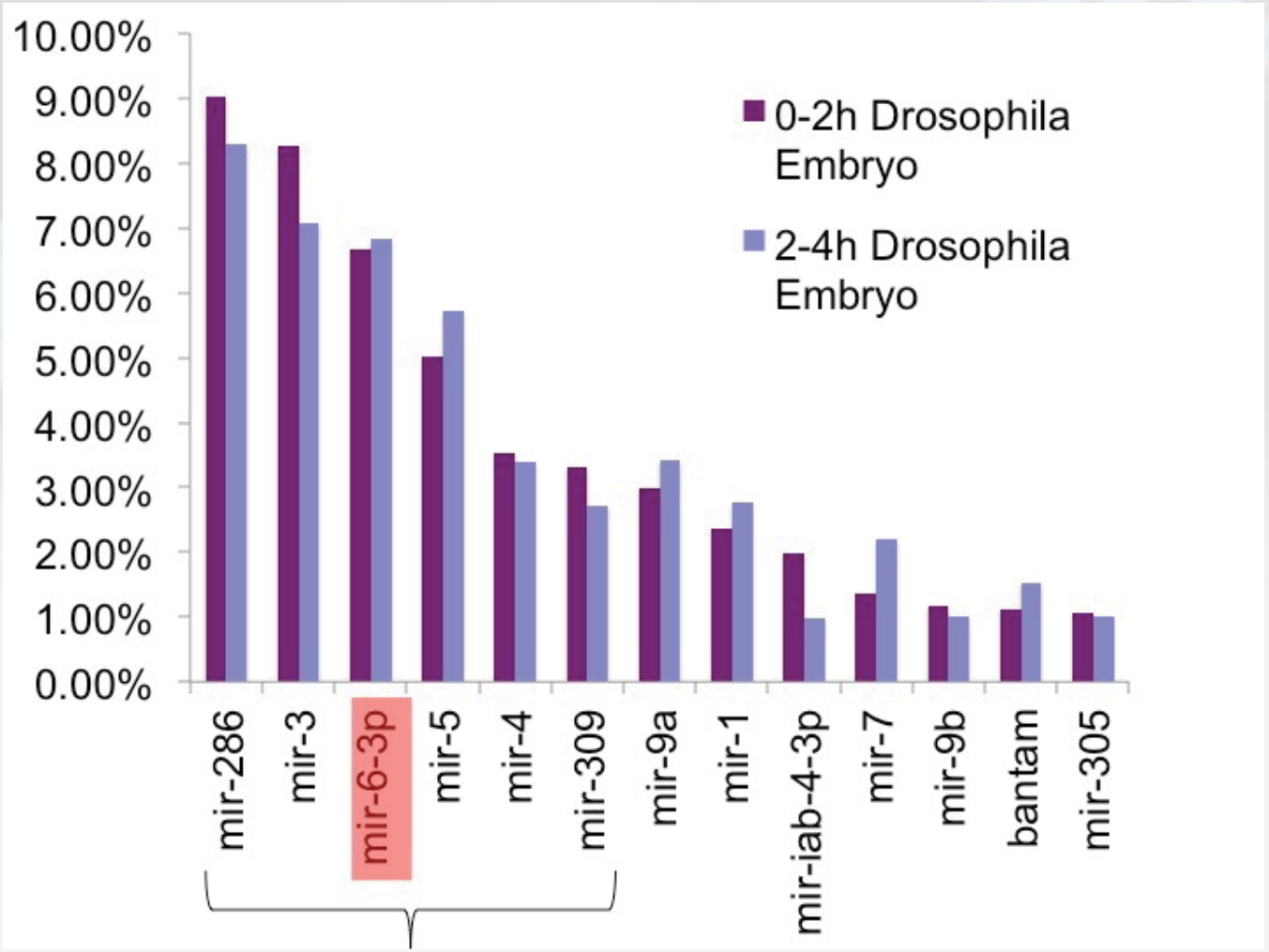
Three instances of dme-miR-6 in drosophila genome from the same cluster





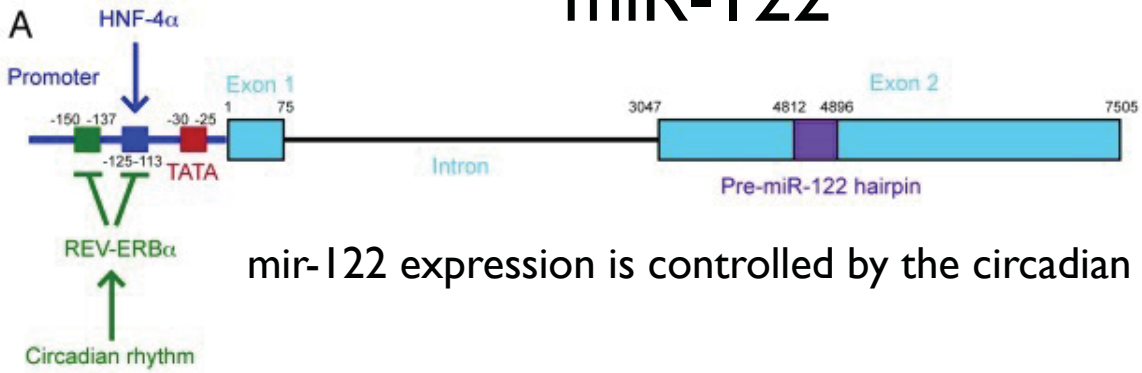
unpublished

# Correct distribution of miRNAs is restored using our new protocol with randomized adapter ends




unpublished

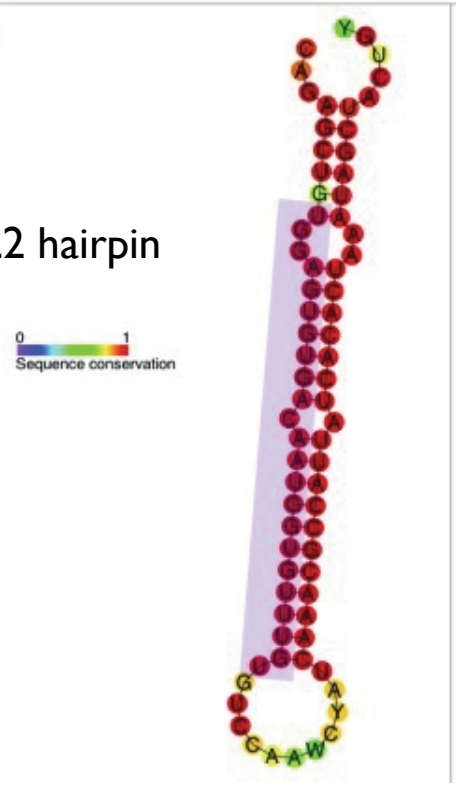
# Liver specific miRNA in humans, miR-122




mir-122 expression is controlled by the circadian rhythm

B

mir-122 hairpin




C

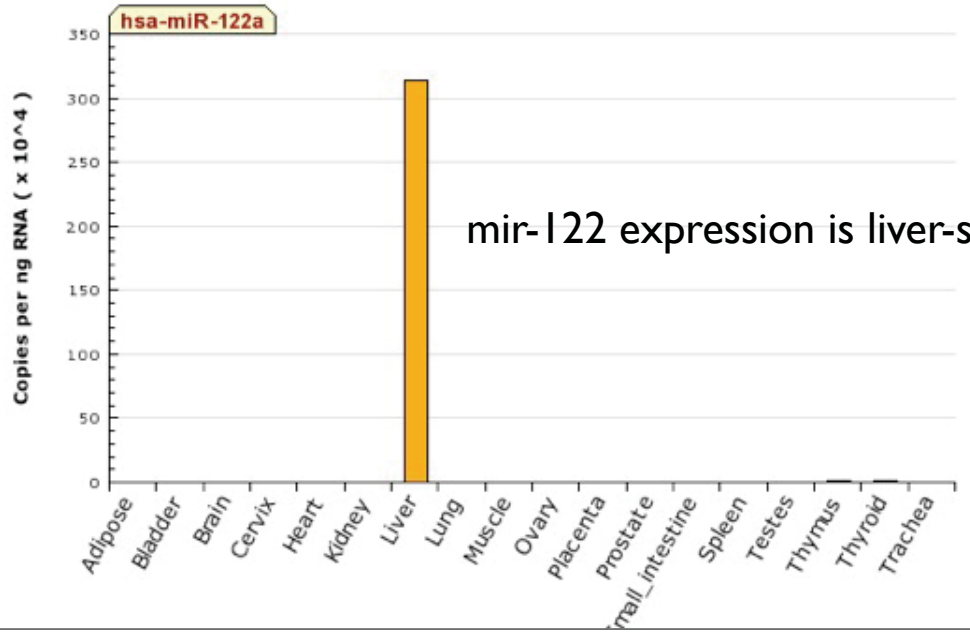
5'-UGGAGUGUGACAAUGGUGUUU-3' 11%

5'-UGGAGUGUGACAAUGGUGUUUG-3' 27%

5'-UGGAGUGUGACAAUGGUGUUUGU-3' 43%

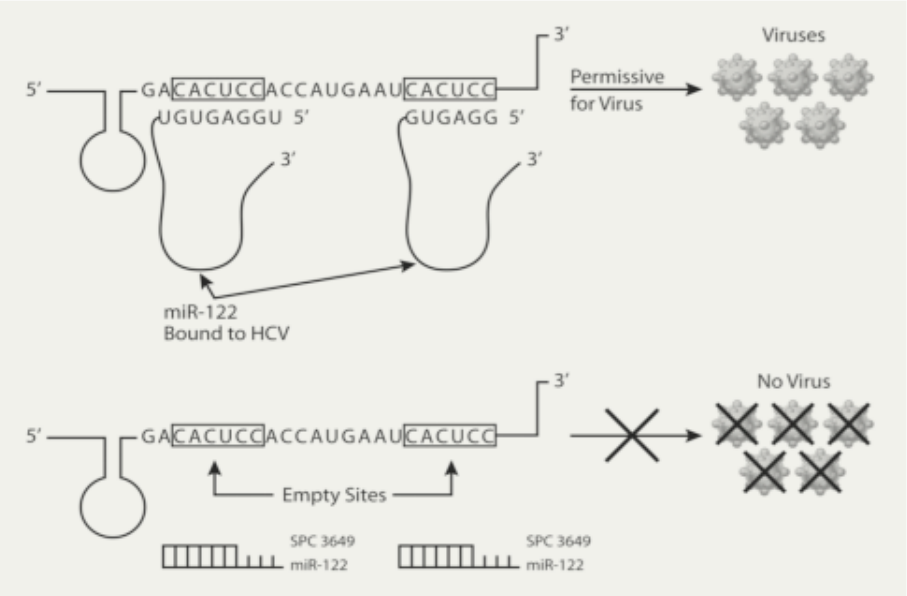
5'-UGGAGUGUGACAAUGGUGUUUGA-3' 19%

mir-122 mature sequences.




mir-122 expression is liver-specific

# Curing hepatitis-C

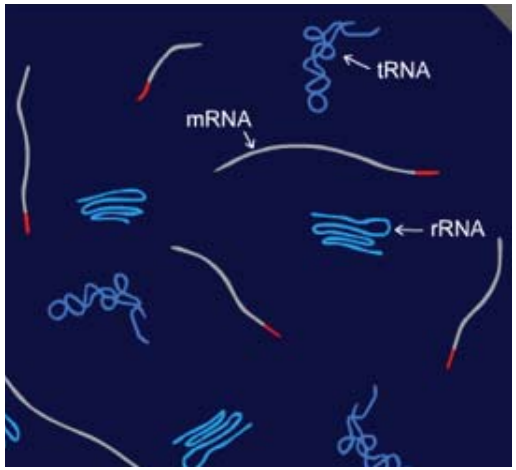



Hep-C virus binds to mature miR-122 and uses that to enhance its replication

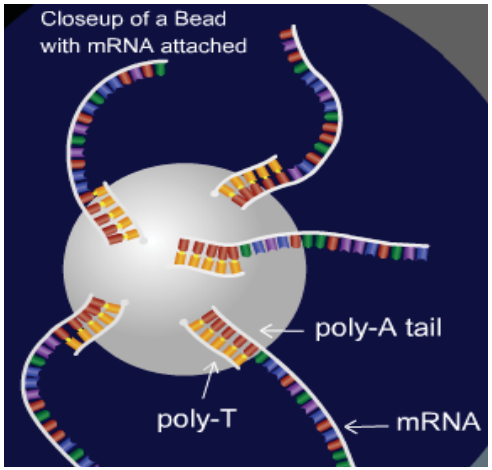
Blocking this binding is one way to halt infection and effect a cure

# mRNA-seq: sequencing the mRNA

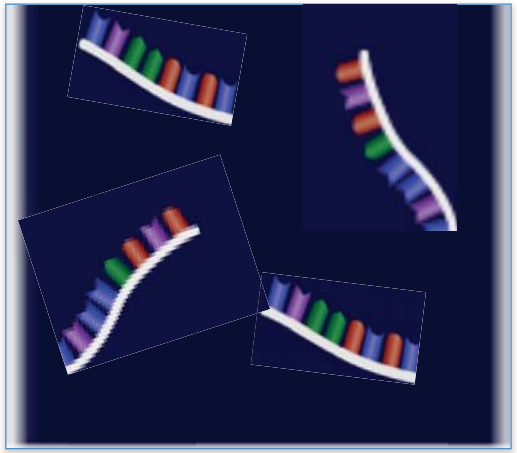
## Total RNA Isolation



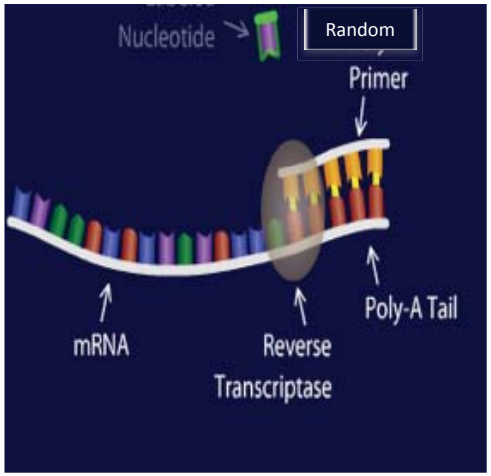

## mRNA isolation




## Random Fragmentation

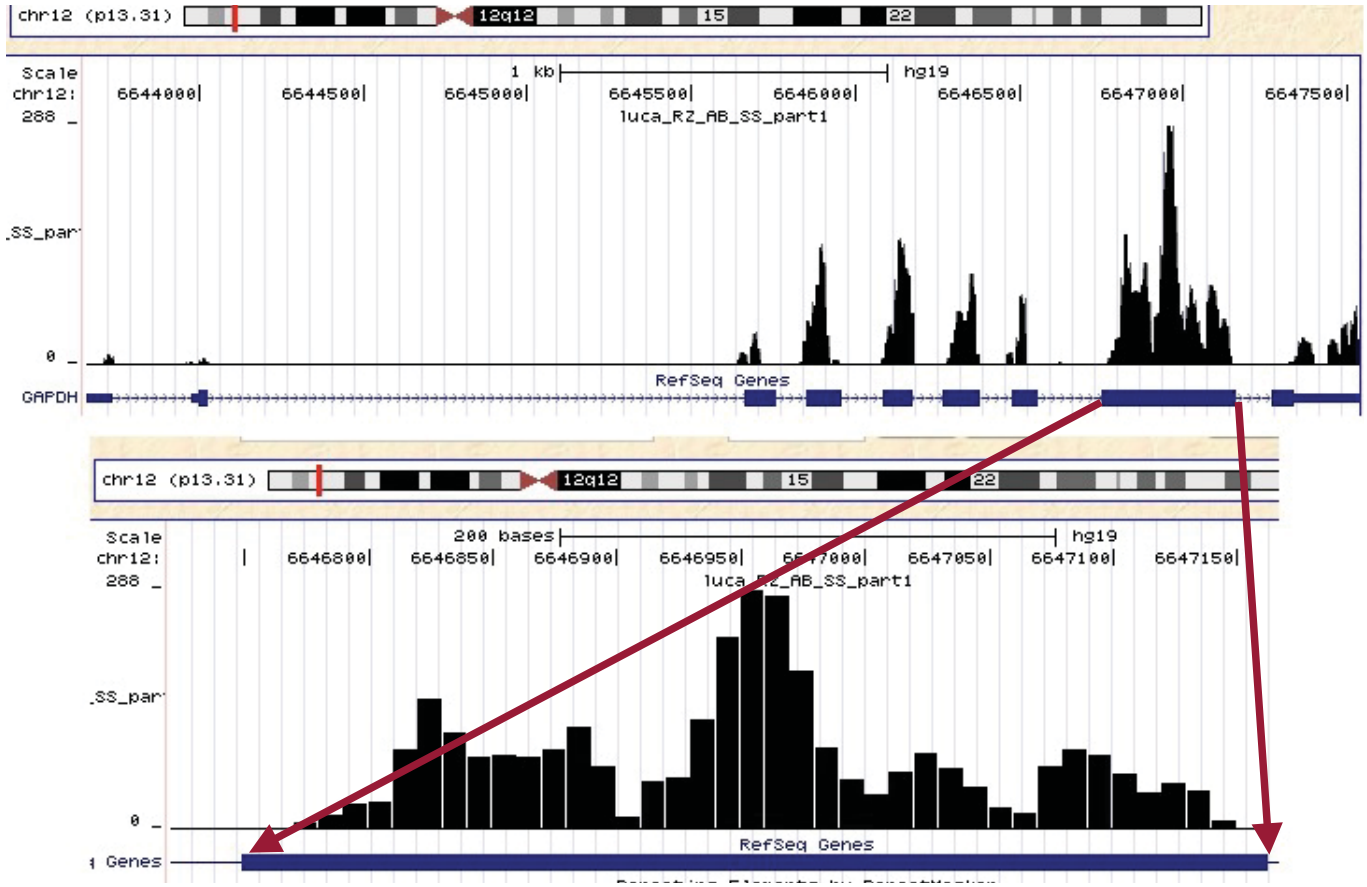


## c-DNA synthesis



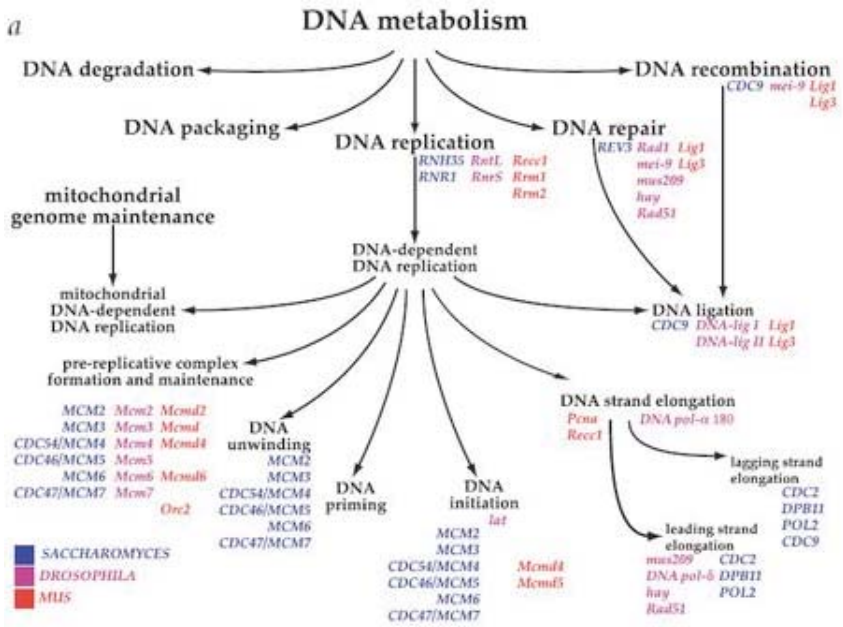

Anitha Jayaprakash

# Inhomogenous coverage of exons




How do you assign a number to the expression level of a gene from such inhomogenous data ?

# Gene Ontology (GO)

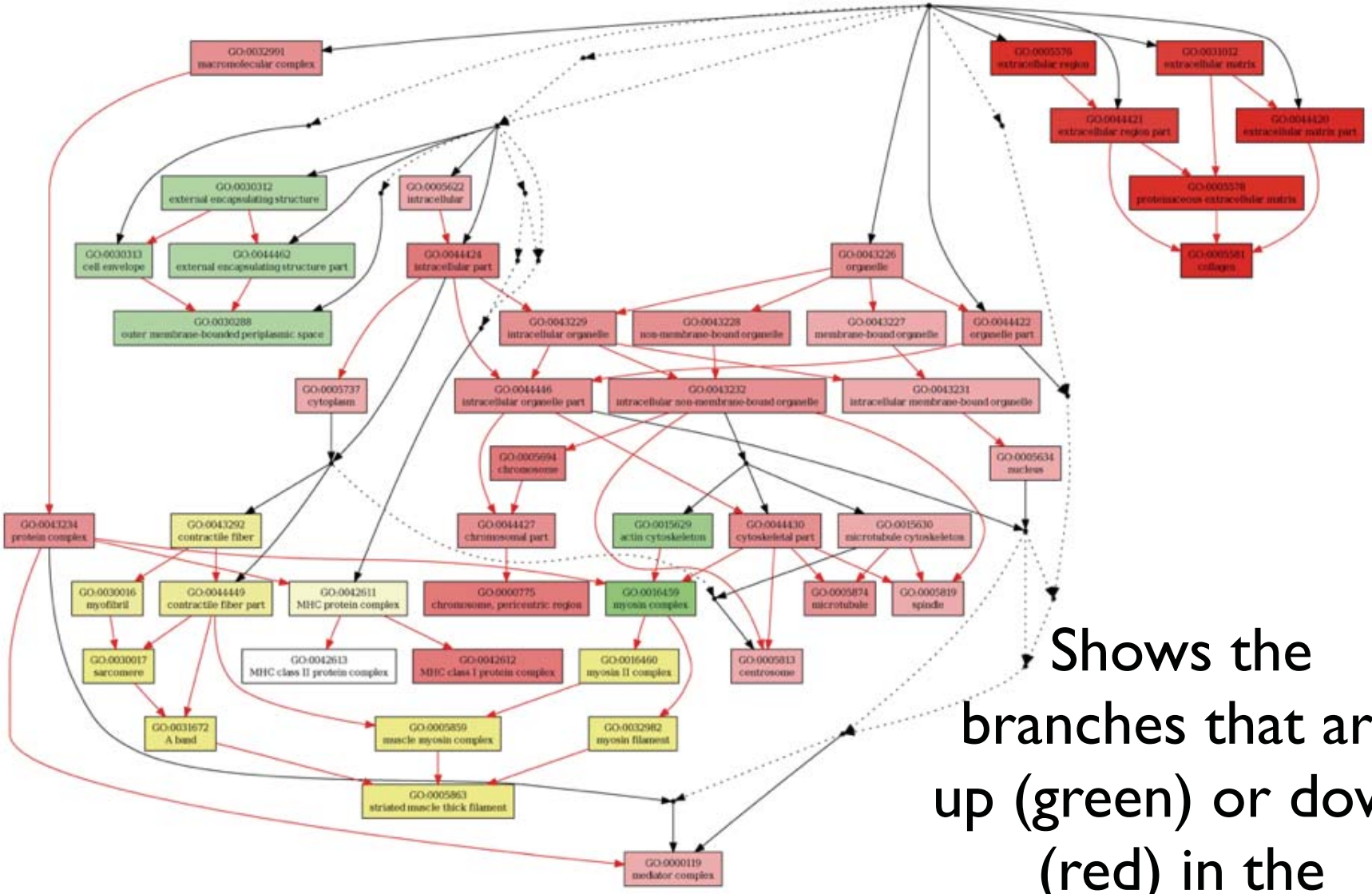



From Gene Ontology, a portion of the biological process ontology describing DNA metabolism. Note that a node may have more than one parent. For example, *DNA ligation* has three: *DNA-dependent DNA replication*, *DNA repair* and *DNA recombination*. *This is a directed acyclic graph (DAG).*

- GO organizes the genes in hierarchies, according to function.
- Changes in expression levels are used to identify branches or pathways in GO that are affected in the experiment

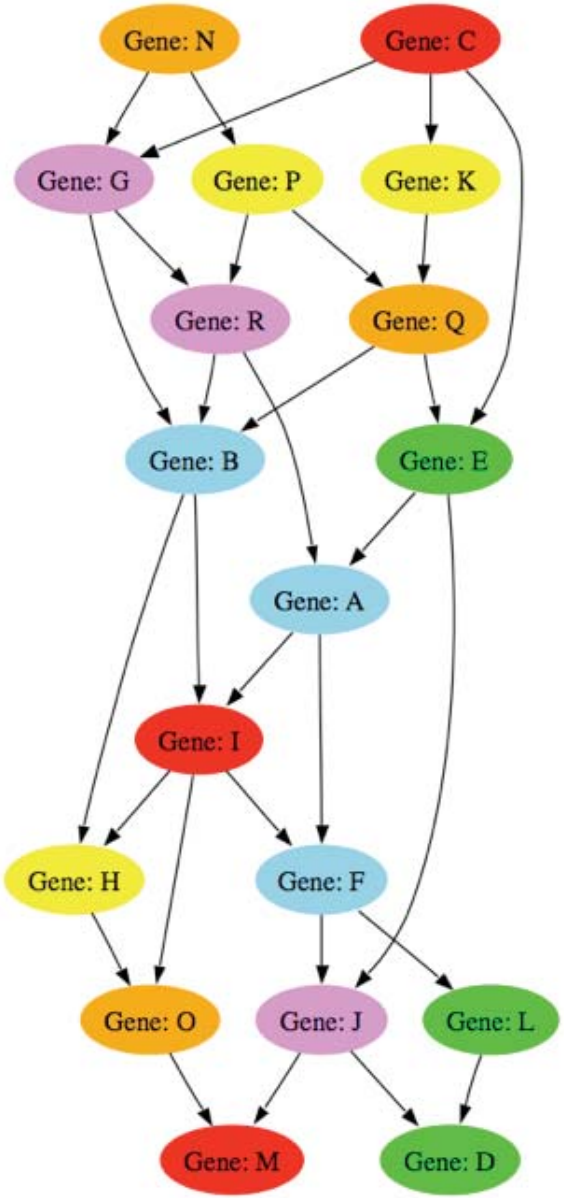
Gene Ontology : http://www.geneontology.org/

# Graphical display of affected terms




Shows the branches that are up (green) or down (red) in the experiment

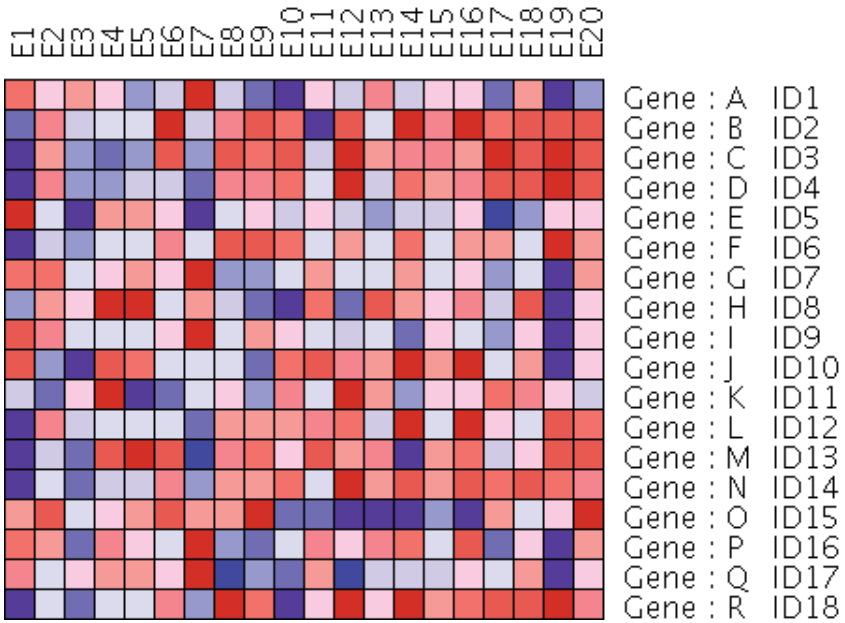
**GO Enrichment:**
Gene Ontology allows the genes to b mapped to processes via gene products. For a given list of genes (selected via any class comparison like SAM etc.), the GO processes tha are enriched in it can be identified. There are standard tools like DAVID that perform these functions. A hypergeometric test can be used to test for GO term enrichment.




**Bayesian Networks.**
Causality relations can be derived from data and modeled using bayesian networks.

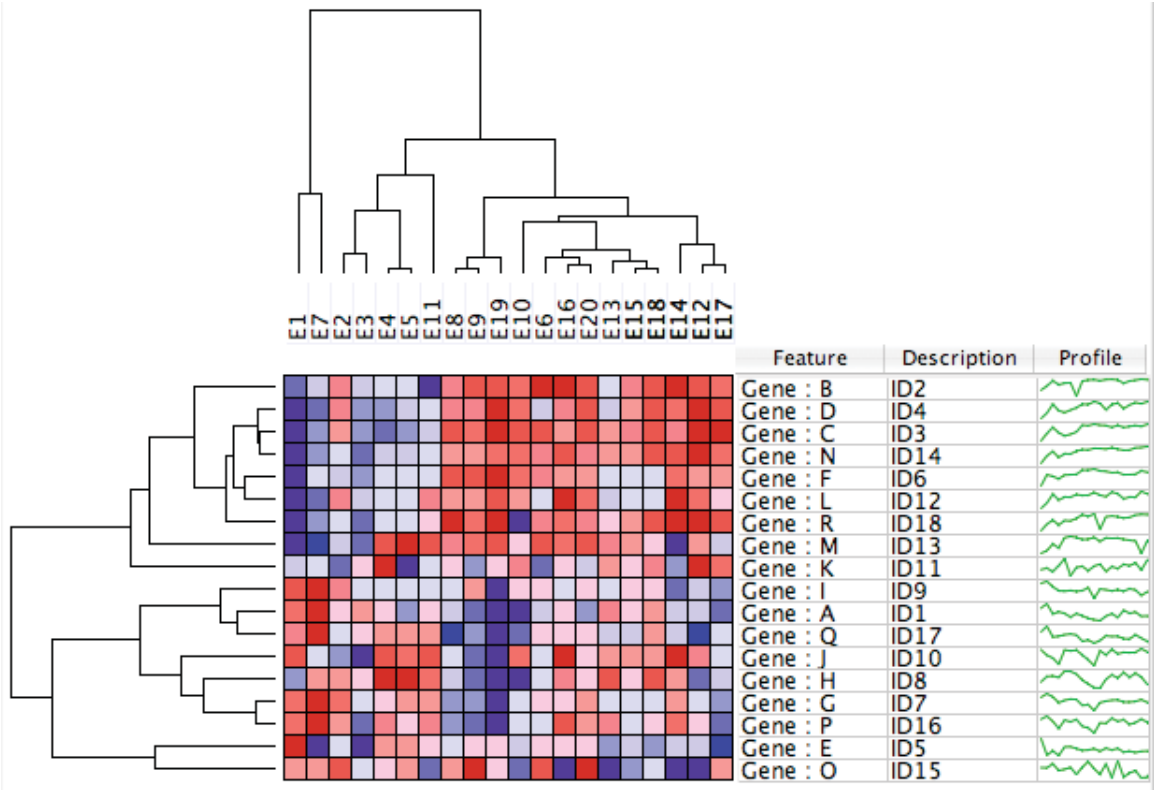
# Heat Maps

- A matrix with genes in the rows and experiments in the columns
- Colors represent values relative to mean
- Visually striking patterns can be detected, especially by organizing rows and columns, using clustering methods (shown in the next slide)



# Heat maps with clustering of rows and columns

Patterns in the colors show groups of genes/experiments that are co-regulated



# Questions ?

Contact me at ravi.mssm@gmail.com for criticisms, comments and questions.

I have tried to give a quick overview in lectures 1 and 2 of the broad ideas in genomics/ bioinformatics.

I have attached to this presentation, slides containing notes, for some material I worked out on the board, as well as some extra population genetics material that I did not have time to get into.