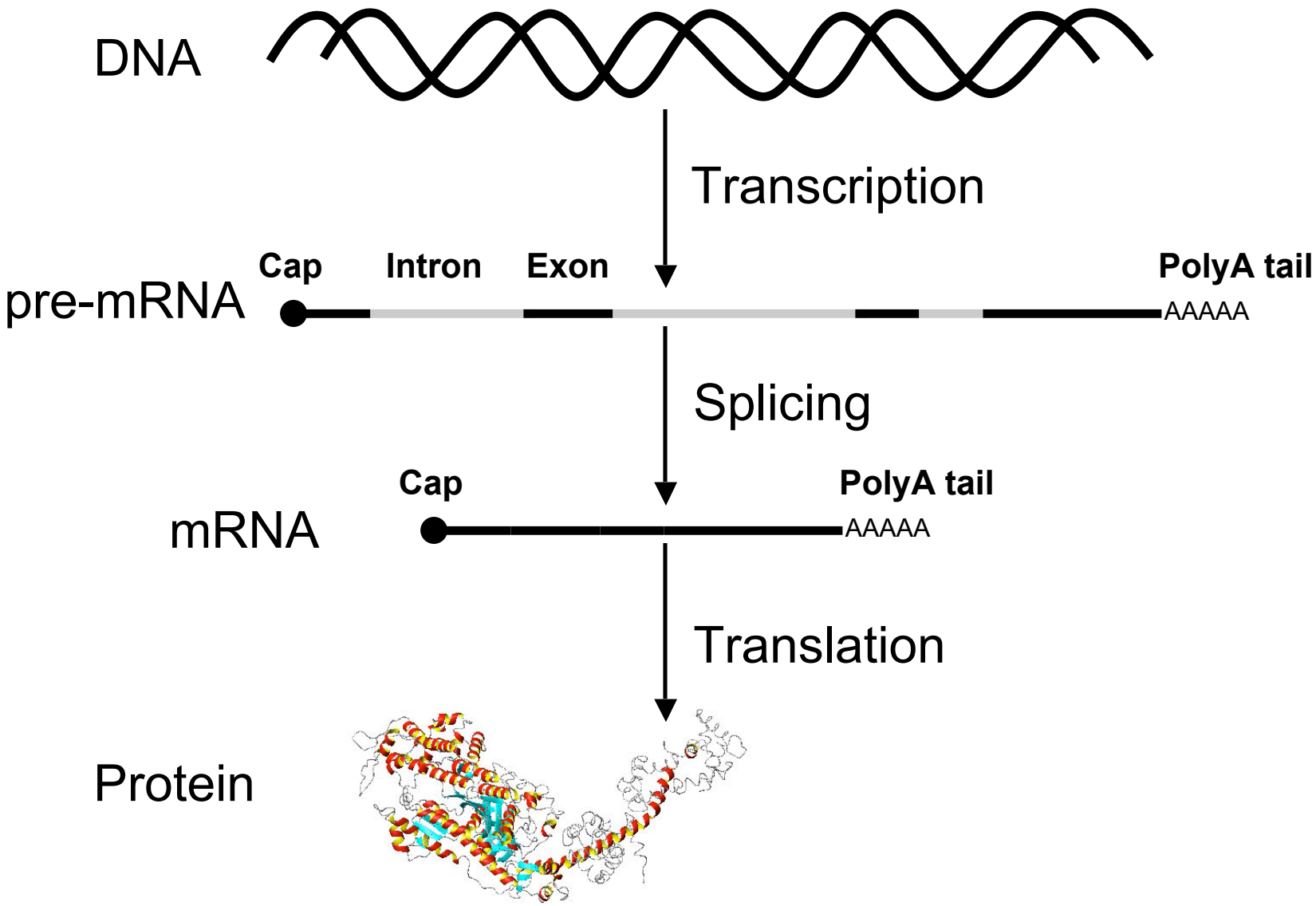
Computational approaches to RNA biology

Mihaela Zavolan
Biozentrum, Basel
Swiss Institute of Bioinformatics

BIOZENTRUM
Universität Basel

SIB

From genes to proteins



Alternative splicing

- Identification of alternative splice forms
- Regulation vs. noise in alternative splicing

Small regulatory RNAs

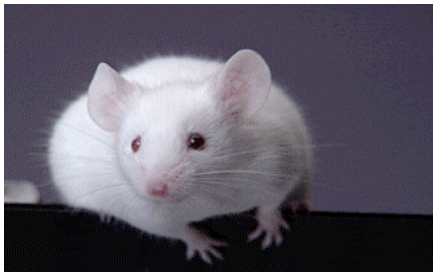
- miRNA gene identification
- miRNA expression profiling

Targeting of alternative transcripts by miRNAs

Evolution of the number of genes

~13,500

~18,000

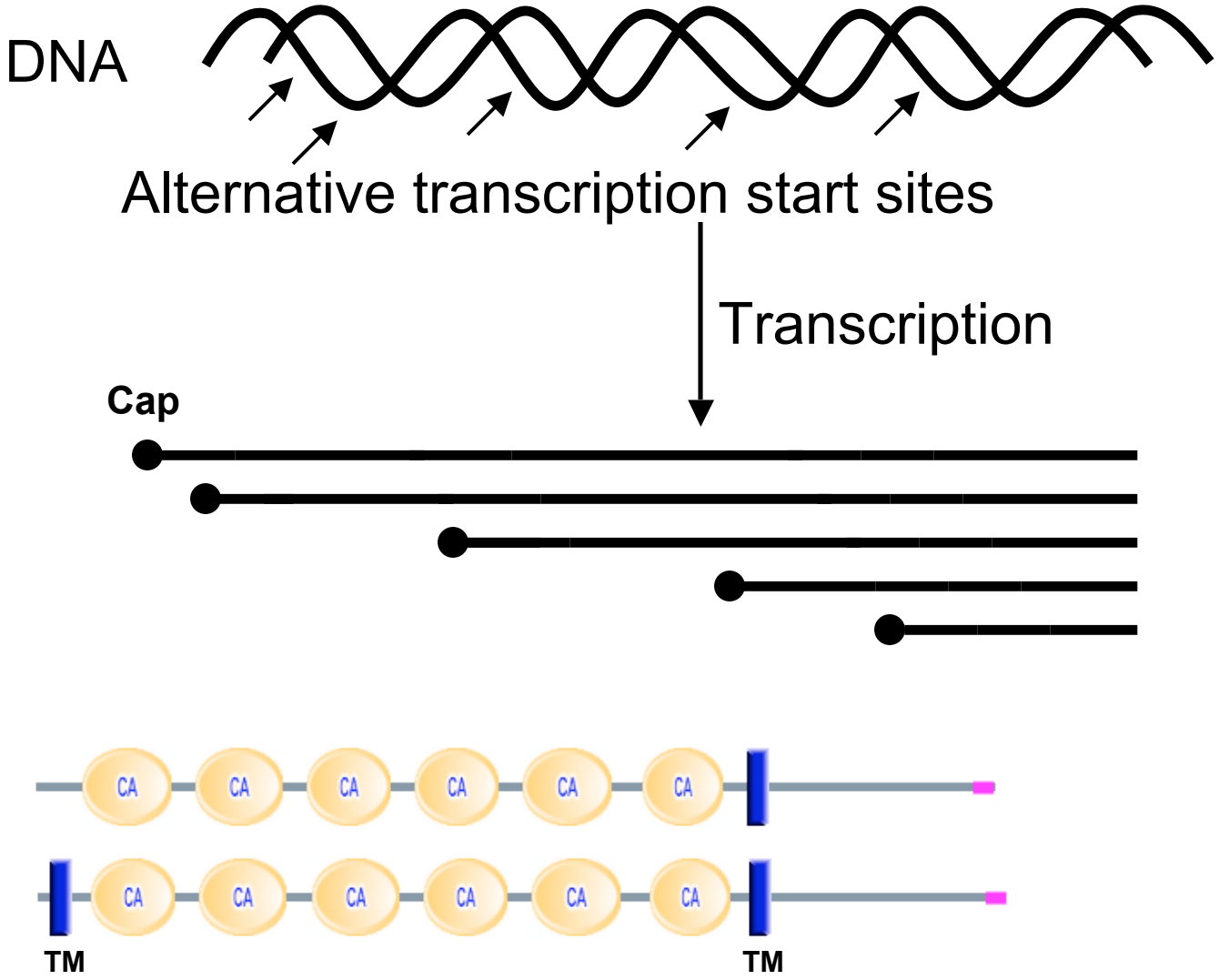


~28,000

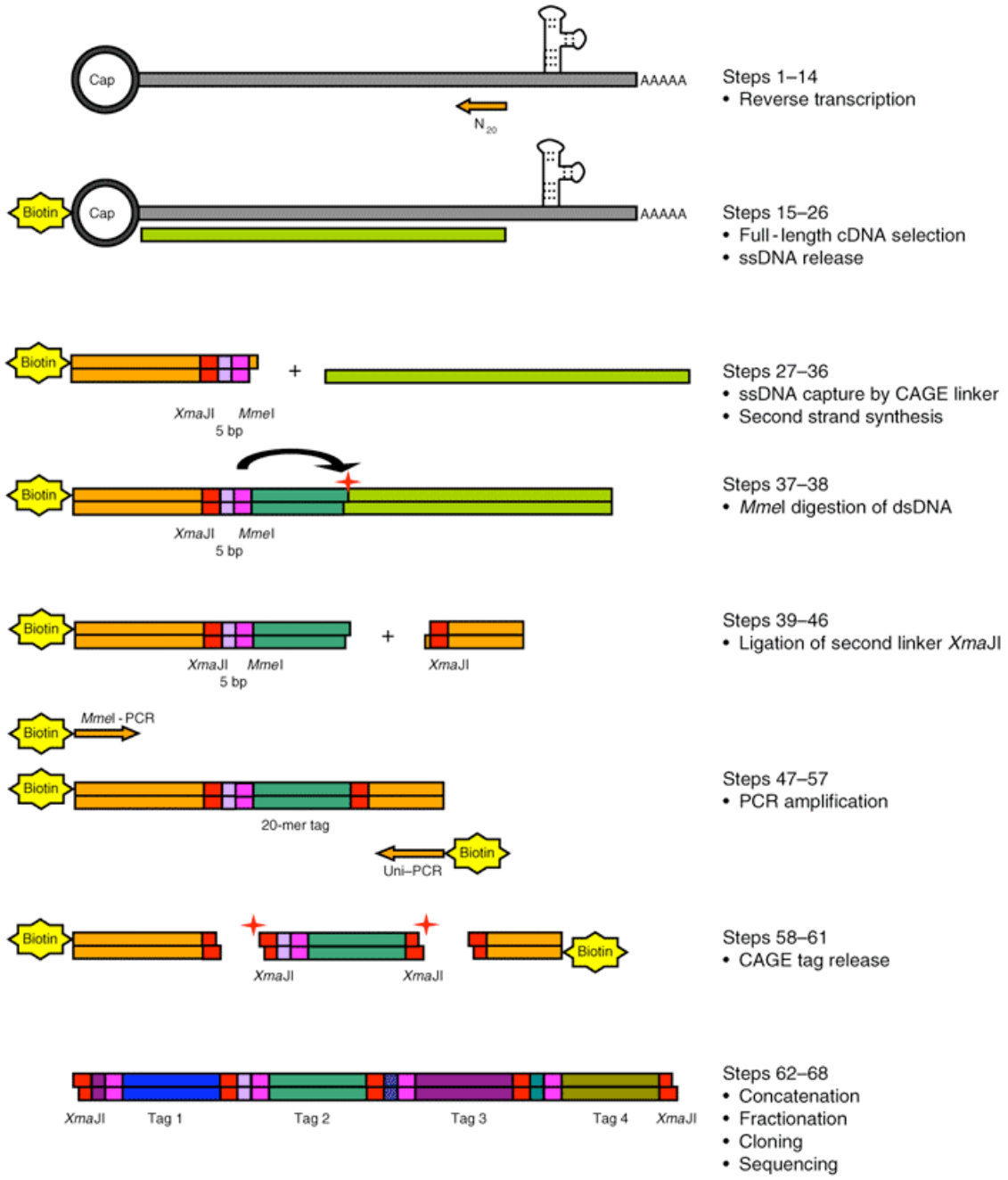
~35,000

Transcriptome complexity ⇒ Phenotypic complexity

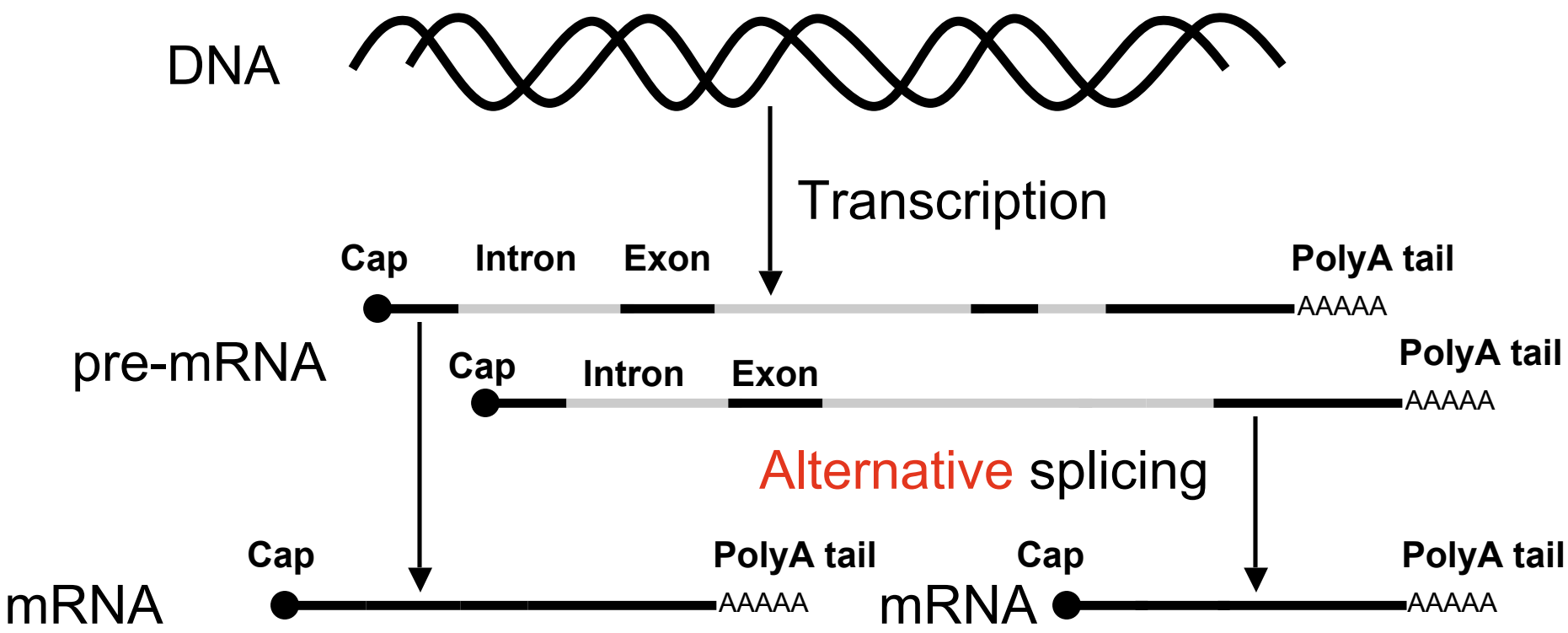
From genes to proteins



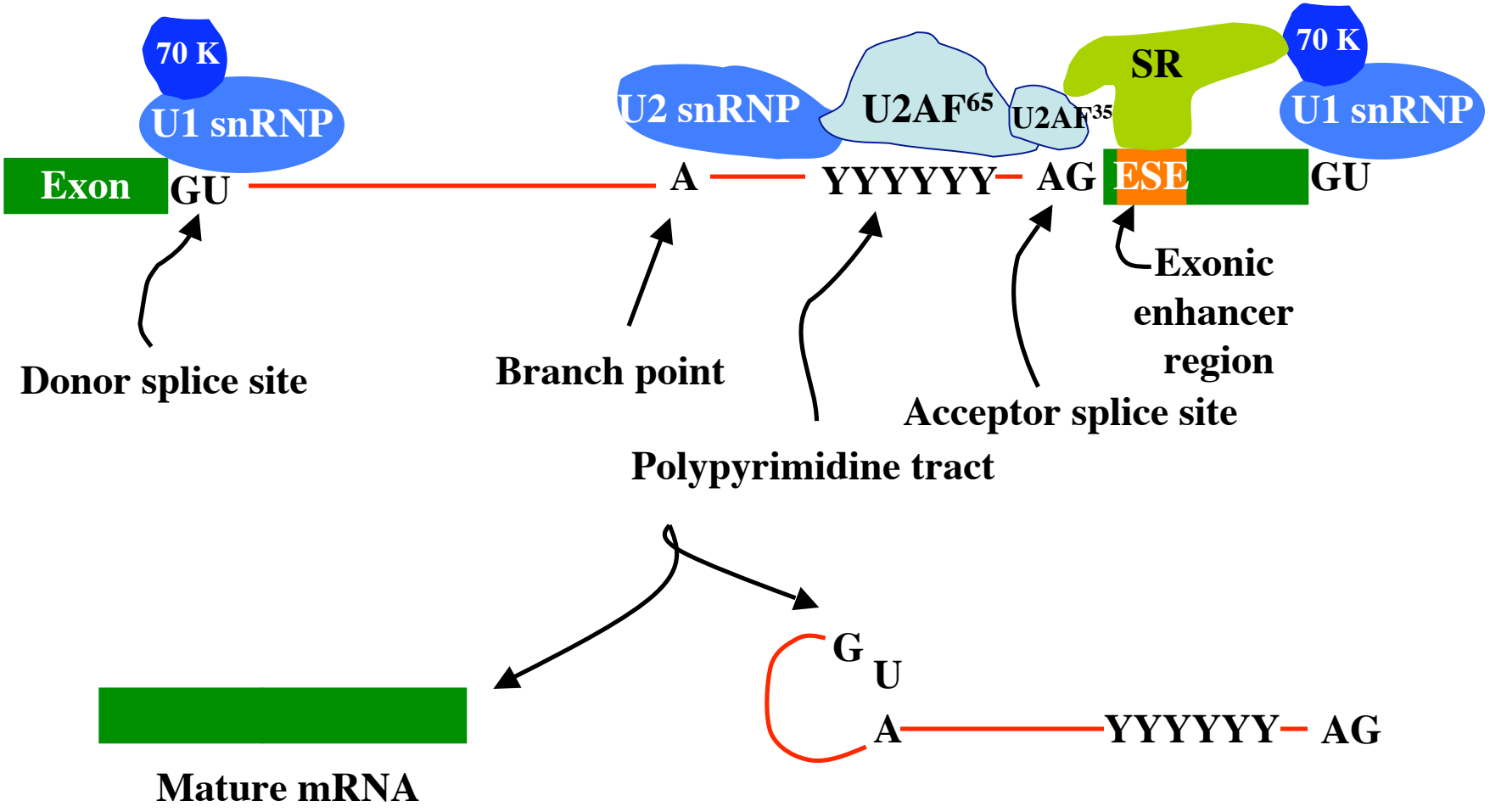
CAGE technology for determination of TSS



From genes to proteins



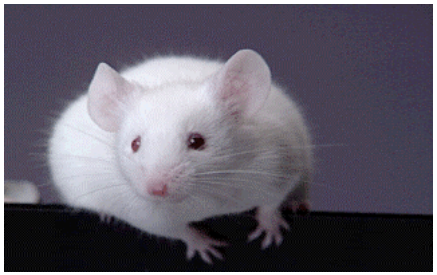
Splicing and splice signals



Evolution of the number of genes

~13,500

~18,000



~28,000

~35,000

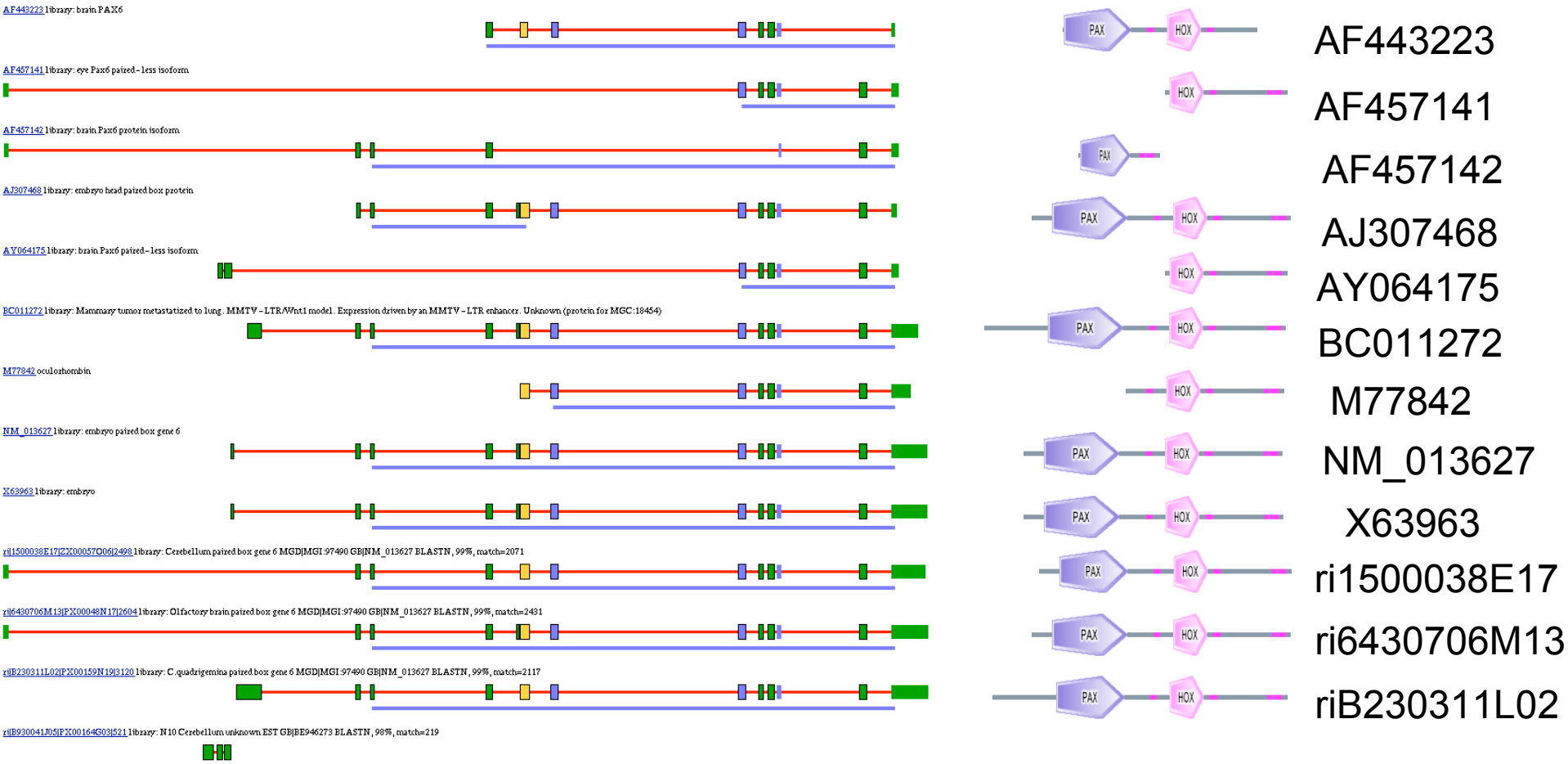
Transcriptome complexity ⇒ Phenotypic complexity

RIKEN Genomic Sciences Center

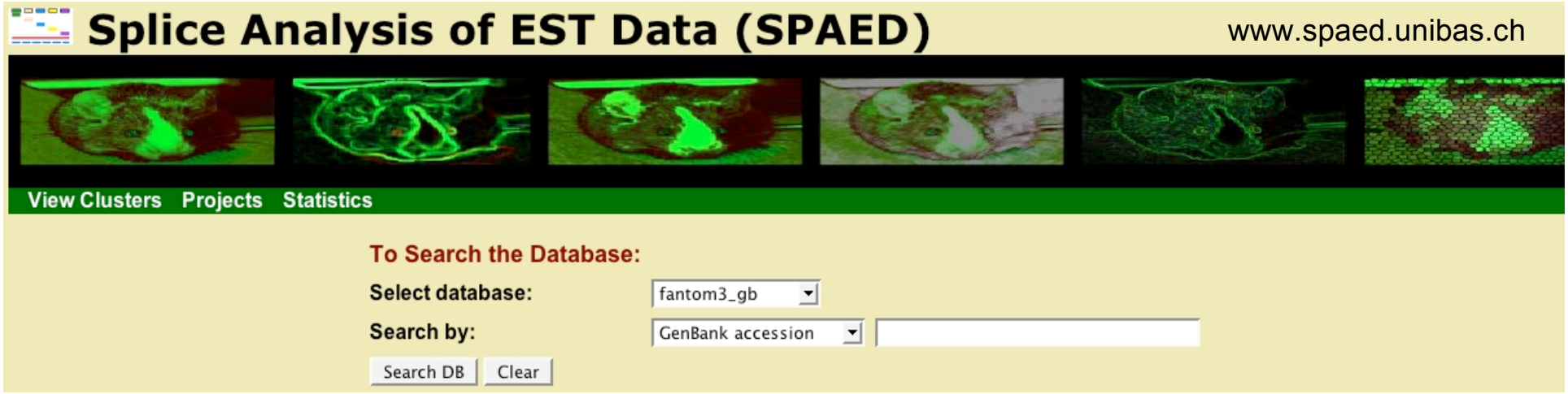
FANTOM
Functional Annotation of Mouse

Fantom3 (Science 2005): 102,801 cDNAs (>250 libraries)

Full-length, high-quality, extensively annotated sequences.


AF443223 library: brain PAX6
AF457141 library: eye Pax6 paired-less isoform
AF457142 library: brain Pax6 protein isoform
AJ307468 library: embryo head paired box protein
AY064175 library: brain Pax6 paired-less isoform
BC011272 library: Mammary tumor metastatized to lung. MMTV-LTR/Wnt1 model. Expression driven by an MMTV-LTR enhancer. Unknown (protein for MGC:18454)
M77842 oculorhombin
NM_013627 library: embryo paired box gene 6
X63963 library: embryo
ri|1500038E17|Z|X00057O06|2498 library: Cerebellum paired box gene 6 MGD|MGI:97490 GB|NM_013627 BLASTN, 99%, match=2071
ri|6430706M13|FX00048N17|2604 library: Olfactory brain paired box gene 6 MGD|MGI:97490 GB|NM_013627 BLASTN, 99%, match=2431
ri|B230311L02|FX00159N19|3120 library: C.quadrigemina paired box gene 6 MGD|MGI:97490 GB|NM_013627 BLASTN, 99%, match=2117
ri|B930041J05|FX00164G03|521 library: N10 Cerebellum unknown EST GB|BE946273 BLASTN, 98%, match=219
PAX
HOX
AF443223
AF457141
AF457142
AJ307468
AY064175
BC011272
M77842
NM_013627
X63963
ri1500038E17
ri6430706M13
riB230311L02

Identification of alternative splice forms



From raw sequences to web-accessible database of splice forms:

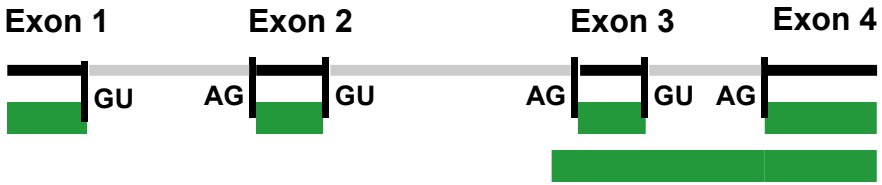
- Accurate and efficient mapping of transcripts to the genome

- Splice variation inferred through comparison of splice site usage in transcripts from a given genomic locus

- Classification of alternative splice forms in terms of relatively simple, known forms of variation.

Mapping transcripts to the genome

Spa: a probabilistic algorithm for spliced alignment



g: the genome from which the cDNA originates
t: an intron-exon structure indicated on the genome
c: the cDNA that one wants to map

P(t) = prior probability of the intron-exon structure (intron-exon lengths)

P(g|t) = probability of the genome nucleotides given the transcript (nucleotides at the splice boundaries)

P(c|t) = probability of the cDNA given the transcript (sequencing errors)

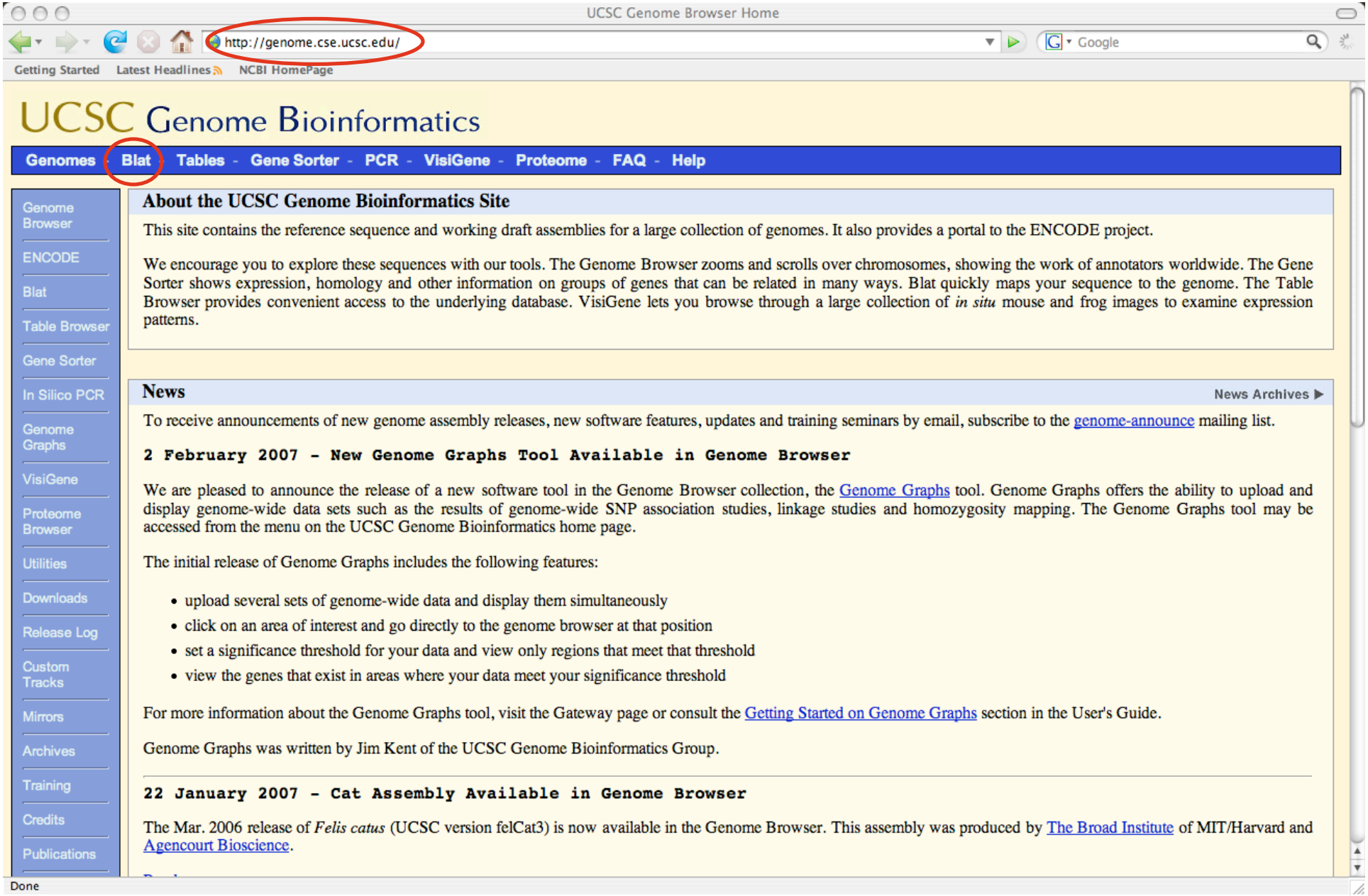
We want to calculate the probability of the transcript *t* given both the cDNA *c* and the genome *g:*

$$P(t \mid c,g) = \frac{P(c,g \mid t)P(t)}{P(c,g)} = \frac{P(c \mid t)P(g \mid t)P(t)}{P(c,g)}$$

Van Nimwegen *et al.* PLoS Genetics 2006

Spa: a probabilistic algorithm for spliced alignment

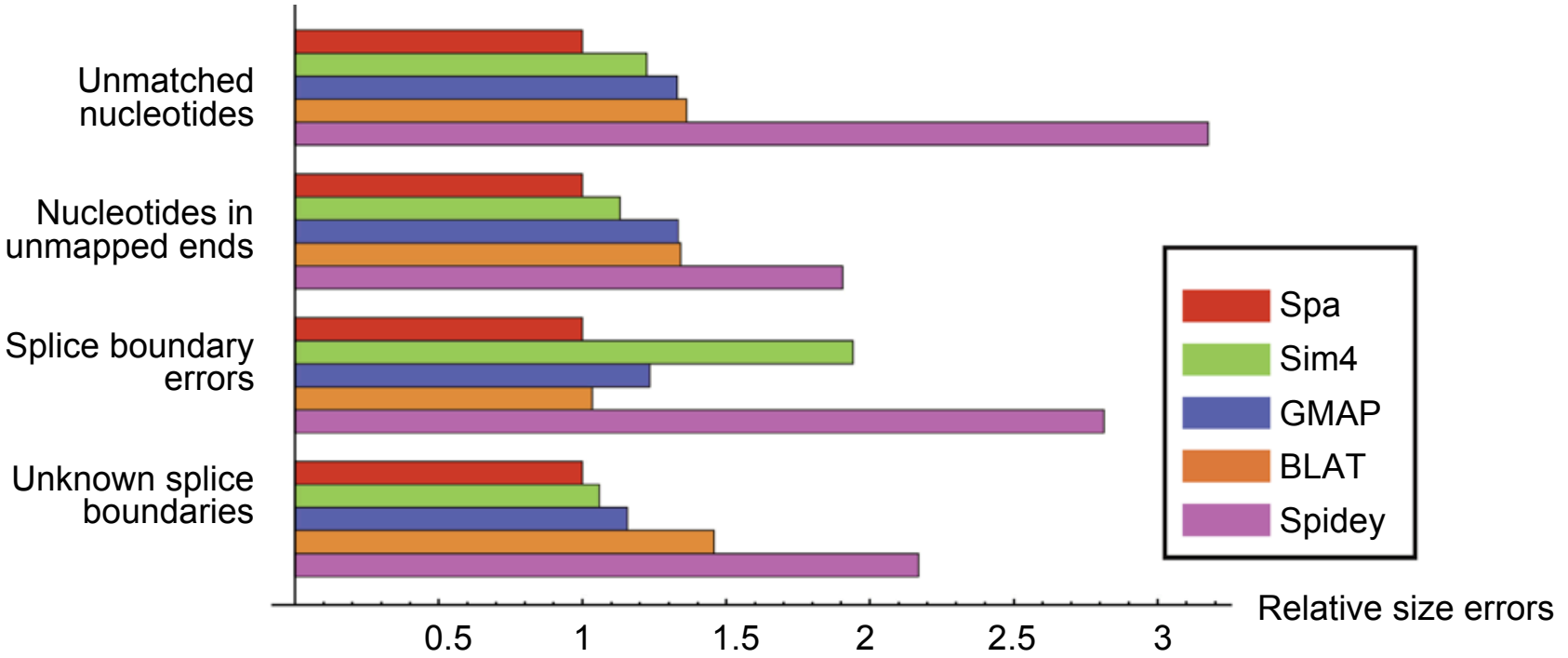
- Use the BLAT gfServer to identify genomic loci to which the cDNA might map.


UCSC Genome Browser Home
http://genome.cse.ucsc.edu/
Google
Getting Started
Latest Headlines
NCBI HomePage
UCSC Genome Bioinformatics
Genomes
Blat
Tables
Gene Sorter
PCR
VisiGene
Proteome
FAQ
Help
Genome Browser
ENCODE
Blat
Table Browser
Gene Sorter
In Silico PCR
Genome Graphs
VisiGene
Proteome Browser
Utilities
Downloads
Release Log
Custom Tracks
Mirrors
Archives
Training
Credits
Publications
About the UCSC Genome Bioinformatics Site
This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides a portal to the ENCODE project.
We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database. VisiGene lets you browse through a large collection of in situ mouse and frog images to examine expression patterns.
News
News Archives
To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list.
2 February 2007 - New Genome Graphs Tool Available in Genome Browser
We are pleased to announce the release of a new software tool in the Genome Browser collection, the Genome Graphs tool. Genome Graphs offers the ability to upload and display genome-wide data sets such as the results of genome-wide SNP association studies, linkage studies and homozygosity mapping. The Genome Graphs tool may be accessed from the menu on the UCSC Genome Bioinformatics home page.
The initial release of Genome Graphs includes the following features:
upload several sets of genome-wide data and display them simultaneously
click on an area of interest and go directly to the genome browser at that position
set a significance threshold for your data and view only regions that meet that threshold
view the genes that exist in areas where your data meet your significance threshold
For more information about the Genome Graphs tool, visit the Gateway page or consult the Getting Started on Genome Graphs section in the User's Guide.
Genome Graphs was written by Jim Kent of the UCSC Genome Bioinformatics Group.
22 January 2007 - Cat Assembly Available in Genome Browser
The Mar. 2006 release of Felis catus (UCSC version felCat3) is now available in the Genome Browser. This assembly was produced by The Broad Institute of MIT/Harvard and Agencourt Bioscience.
Done

Spa: a probabilistic algorithm for spliced alignment

- Use the BLAT gfServer to identify genomic loci to which the cDNA might map.
- For each locus assemble a set of cDNA/genome coordinate pairs (genome/cDNA word matches extended to the first mismatch and expanded at the ends) that we call *defined positions*.
- Compute optimal alignment using dynamic programming.
- If parts of the cDNA fail to map retile the gap regions with smaller word size and repeat dynamic programming step.
- Report the optimal alignment from all the loci and orientations.

Spa: comparison with other programs



SPA		Sim4		GMAP		BLAT		Spidey	
Splice Site	**Percent**	**Splice Site**	**Percent**	**Splice Site**	**Percent**	**Splice Site**	**Percent**	**Splice Site**	**Percent**
GT-AG	96.0%	GT-AG	94.1%	GT-AG	96.6%	GT-AG	94.9%	GT-AG	92.5%
GC-AG	1.06%	GC-AG	1.36%	GC-AG	1.06%	GC-AG	1.01%	GC-AG	1.05%
AG-CC	0.27%	CT-AC	0.55%	AT-AC	0.17%	CT-AC	0.20%	GC-CA	0.20%
TG-CC	0.18%	GT-AA	0.38%	CT-GG	0.04%	AT-AC	0.14%	GT-TG	0.17%
AT-AC	0.17%	GT-TG	0.22%	CA-CC	0.03%	GA-AG	0.13%	GT-CT	0.15%
AG-GC	0.15%	CT-GC	0.21%	CC-GG	0.03%	CT-AG	0.11%	GT-GG	0.15%
TG-GC	0.14%	GA-AG	0.20%	CA-GC	0.03%	GT-TG	0.10%	GC-CT	0.15%
AC-CC	0.13%	AT-AG	0.16%	GC-GT	0.03%	AT-AG	0.09%	GG-CA	0.14%
TC-CC	0.11%	CT-AA	0.14%	CC-TG	0.03%	GT-GG	0.09%	GC-GG	0.12%
AA-AA	0.11%	CT-AG	0.13%	CA-GG	0.03%	CC-AG	0.09%	GT-CA	0.11%

Van Nimwegen *et al.* PLoS Genetics 2006

Spa vs. BLAT: examples of alignment differences

SPA:

```
TACGCC---------ATCACT
||||||>>>...>>>||||||
TACGCCATC---CTCATCACT
```

BLAT:

```
TACGCCAT---------CACT
||||||||>>>...>>>||||
TACGCCATCTA---CATCACT
```

SPA:

```
GTGAGGCACTGTCTCAACTAAACTAAACTAAACTAAACTAAACTGAAACAAAACAAAATGCT
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
GTGAGGCACTGTCTCAACTAAACTAAACTAAACTAAACTAAACTGAAACAAAACAAAATGCT
```

BLAT:

```
.......ACTGTCTCAACTAAACTAAACTAAACTAAACTAAACTGAAACAAAACAAAATGCT
---------------|||||||||||||||||||||||||-----|||||||||||||||||
---------------AACTAAACTAAACTAAACTAAACT-----GAAACAAAACAAAATGCT
```

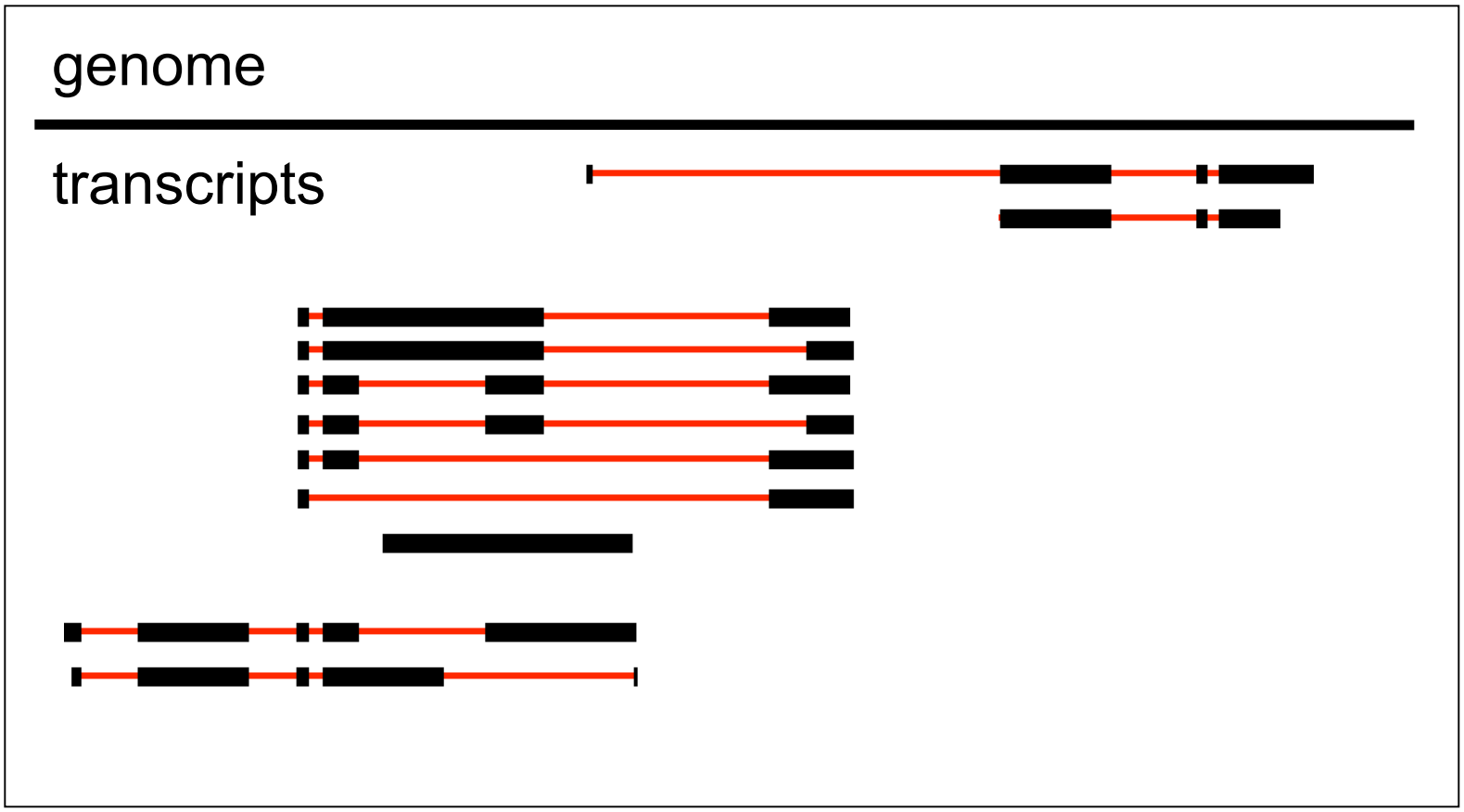
SPA:

```
GTATGGCCCTGGCTGTCCTGGAACTCACTTTGTAGATCAG---------GCTGG
||||  |||||||||||||||||||||||||||||||||<<<...<<<|||||
GTATAGCCCTGGCTGTCCTGGAACTCACTTTGTAGATCAGGCT---CAGGCTGG
```

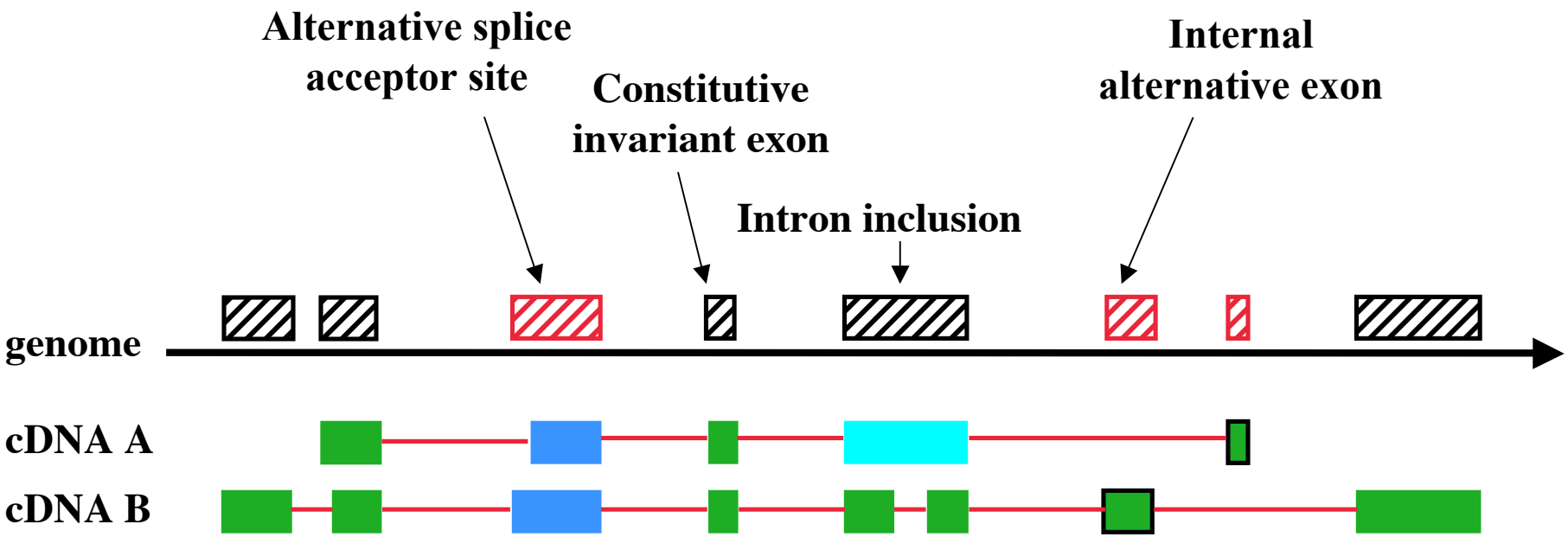
BLAT:

```
GTAT---------GGCCCTGGCTGTCCTGGAACTCACTTTGTAGATCAGGCTGG
||||<<<...<<< ||||||||| ||| ||| || ||  |||||| ||||||||
GTATAGC---TATAGCCCTGGCTTTCCAGGACCTTACCATGTAGACCAGGCTGG
```

Mapping transcripts to genome



Annotation of splice variation

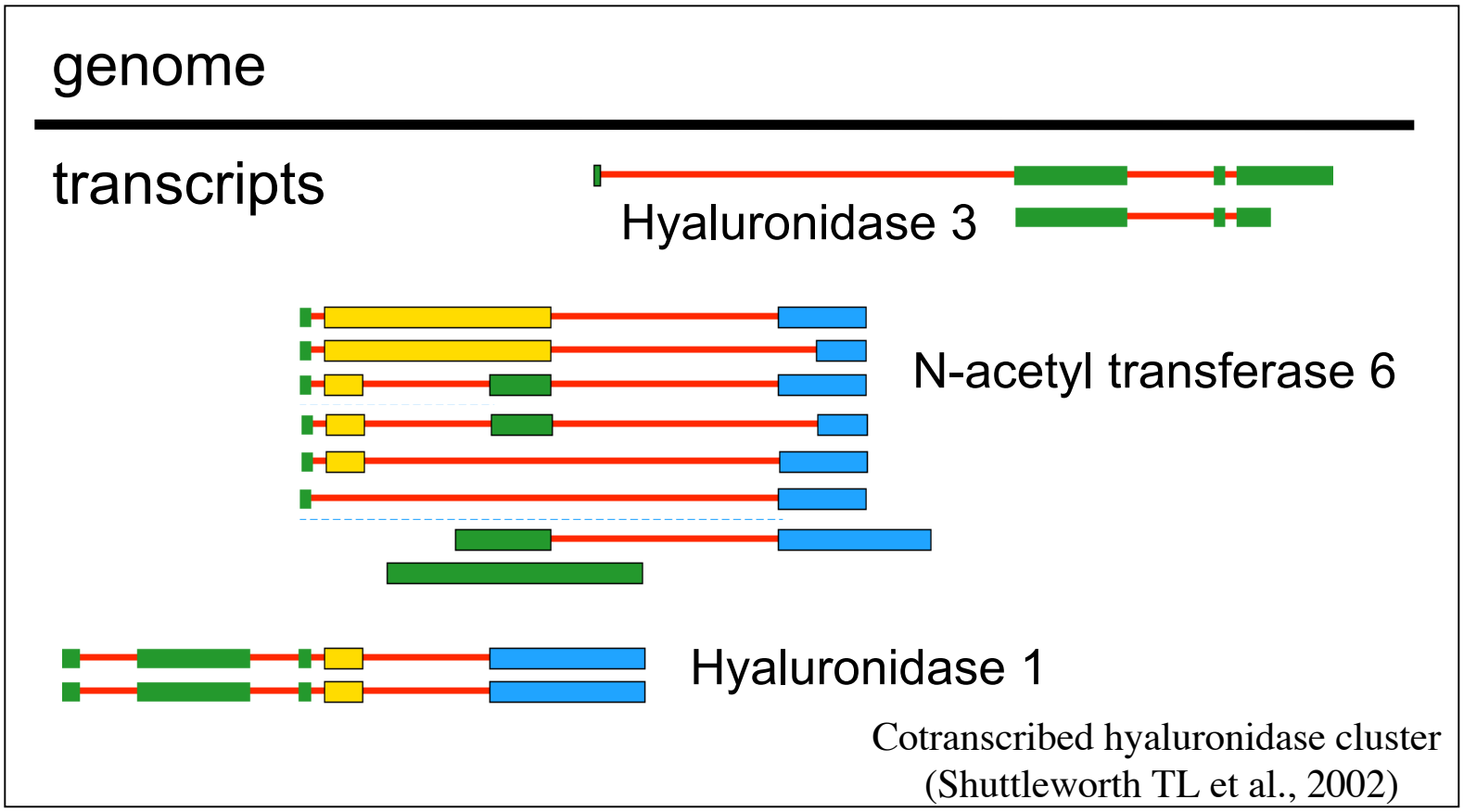


Variant exon

Invariant exon

Expression: constitutive/alternative
Splice site variation: 5'/3'/both
Intron inclusion: yes/no.

Complexity of eukaryotic gene structure



Up to 60% mouse genes have alternative splice forms (Zavolan *et al.* Genome Research 2003).

How much of this variation is functionally relevant?

Alternative splicing

- Identification of alternative splice forms
- Regulation vs. noise in alternative splicing

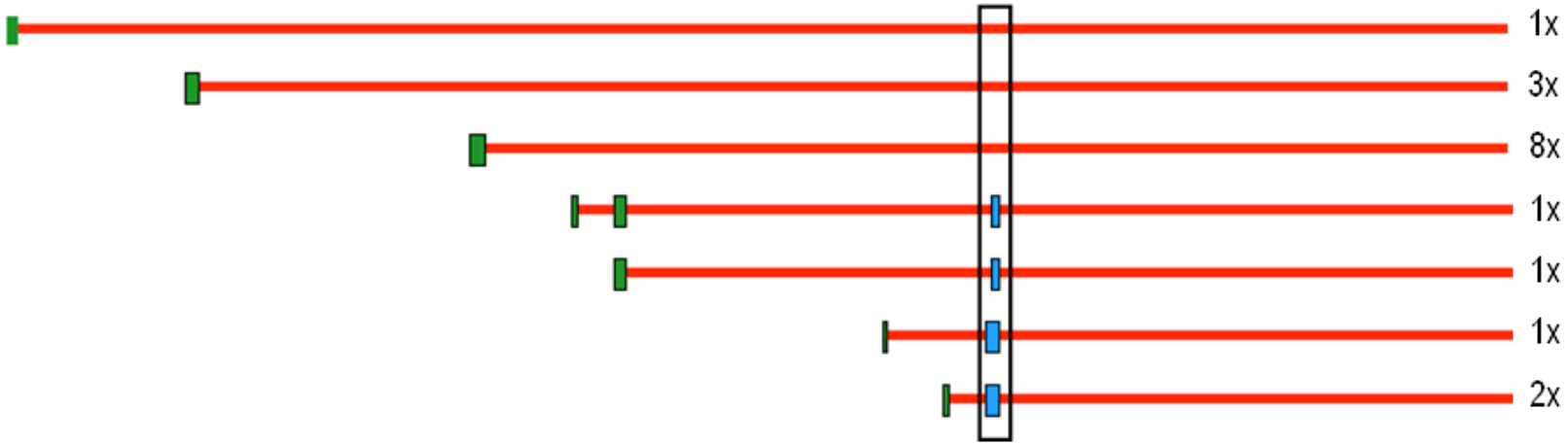
Small regulatory RNAs

- miRNA gene identification
- miRNA expression profiling

Targeting of alternative transcripts by miRNAs

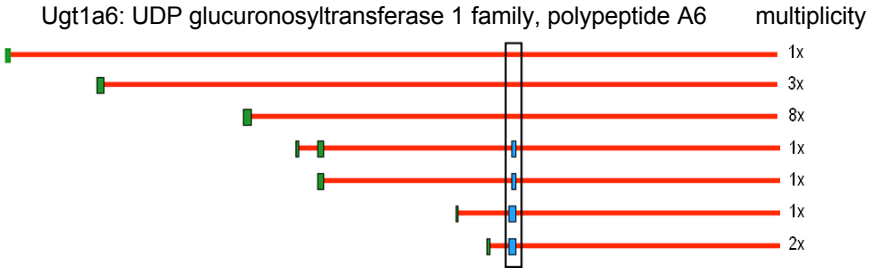
Correlation between promoter choice and internal splicing

Ugt1a6: UDP glucuronosyltransferase 1 family, polypeptide A6

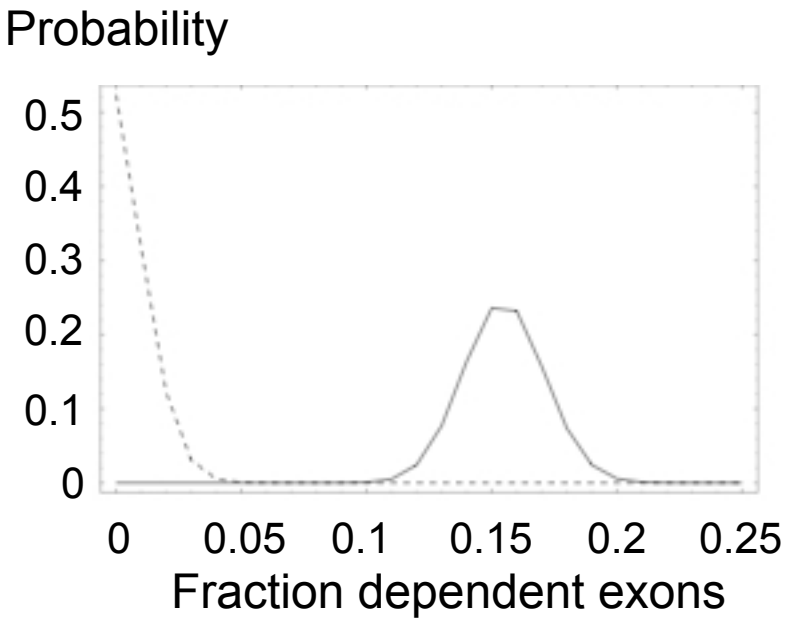


Klrb1c, Klrb1d: Natural Killer cell receptor, subfamily B

Correlation between promoter choice and internal splicing



Promoter	Exon inclusion (n_i)	Exon exclusion (m_i)
1	0	1
2	0	3
3	0	8
4	1	0
5	1	0
6	1	0
7	2	0



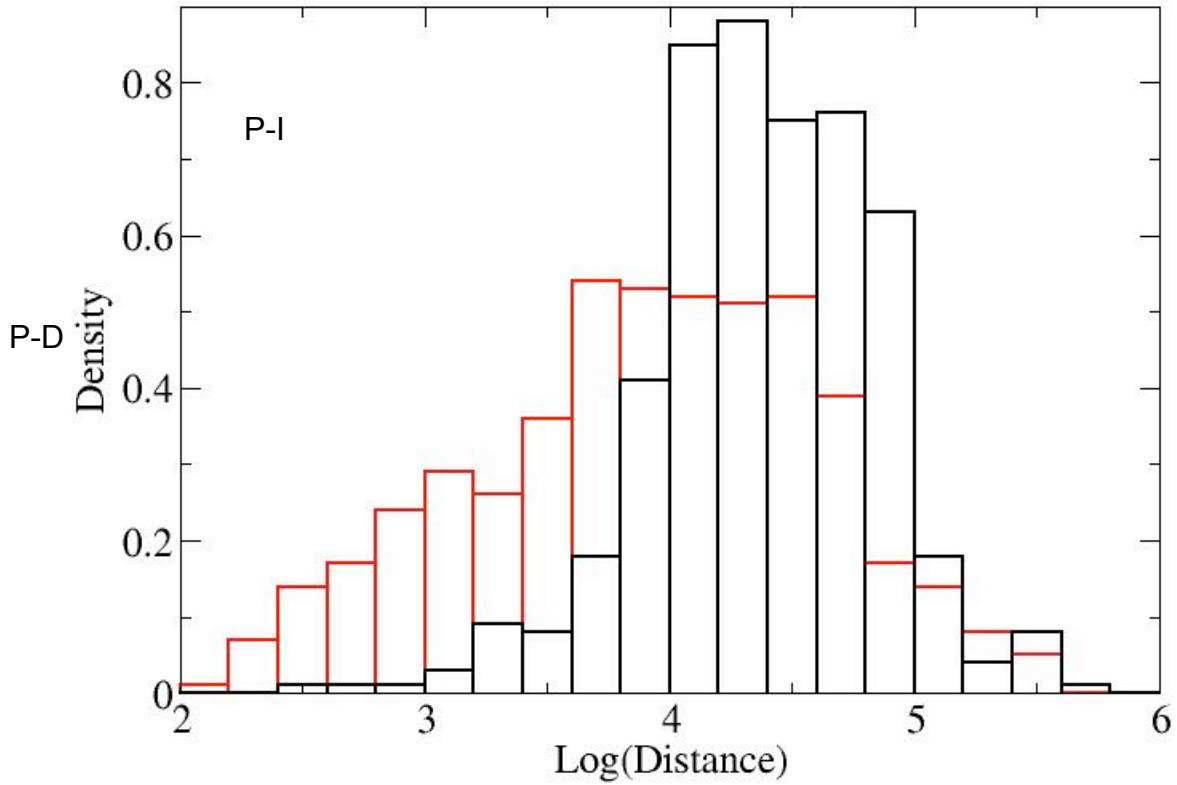
$$P(D \mid f) = \prod_{k \in \text{Exons}} \left[L_{indep}(k)(1-f) + L_{dep}(k) f \right]$$

$$P(f \mid D) = \frac{P(D \mid f)}{\int P(D \mid f) df}$$

$$L_{dep}(k) = \prod_i \int p_i^{m_i} (1-p_i)^{n_i} \, dp_i$$

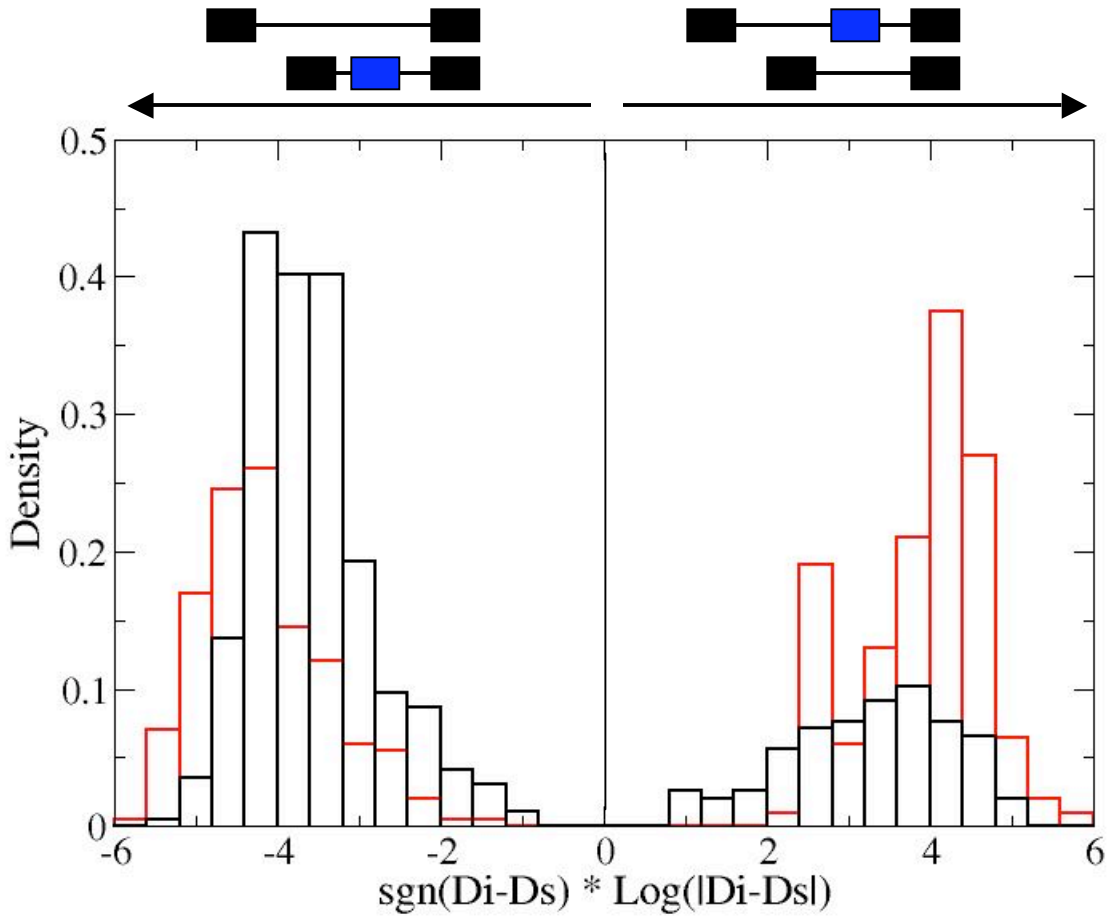
$$L_{indep}(k) = \int p^m (1-p)^n dp.$$

Promoter-dependent exons are closer to promoter than promoter-independent exons

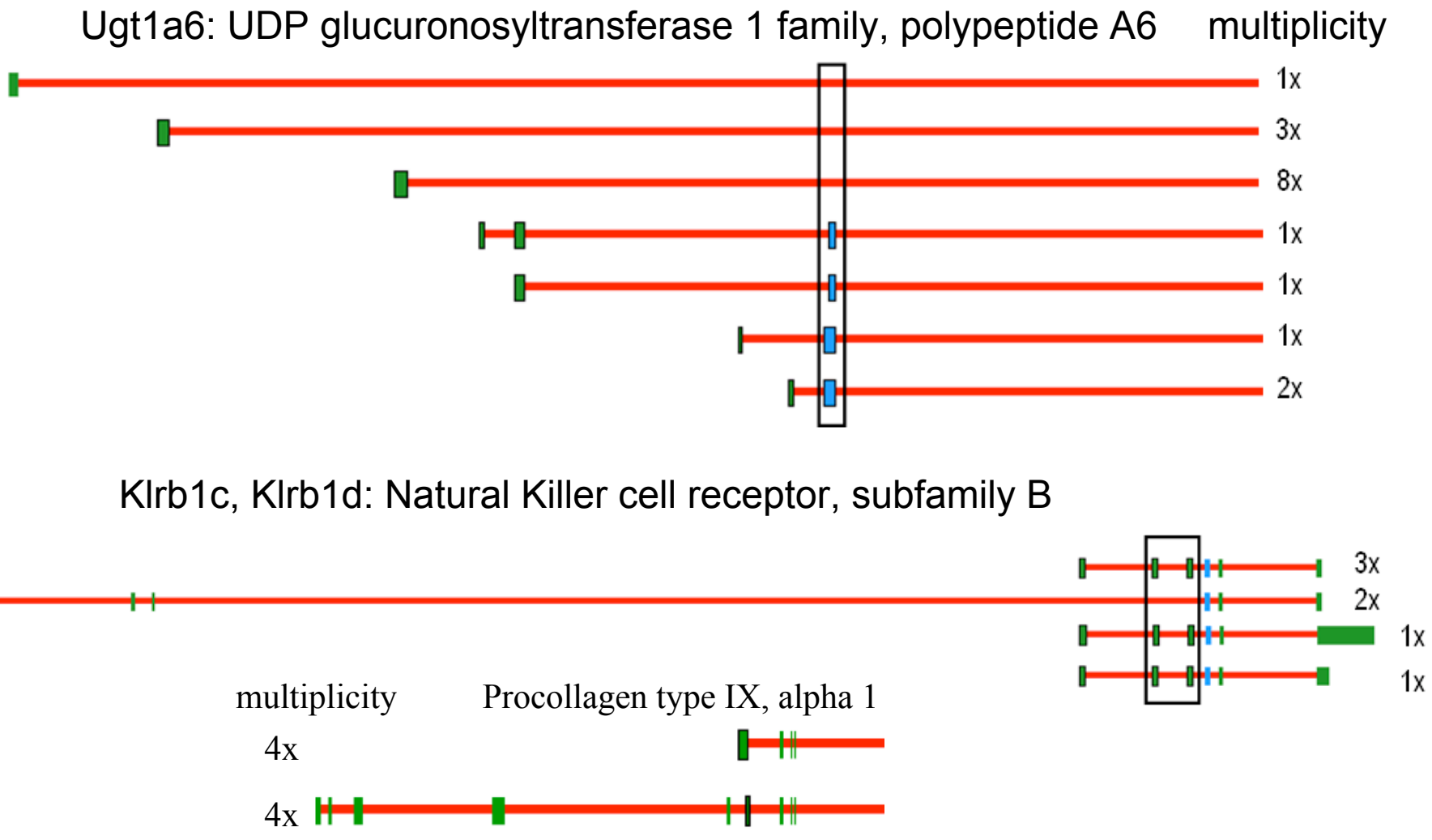


Min(Avg distance included,
Avg distance skipped)

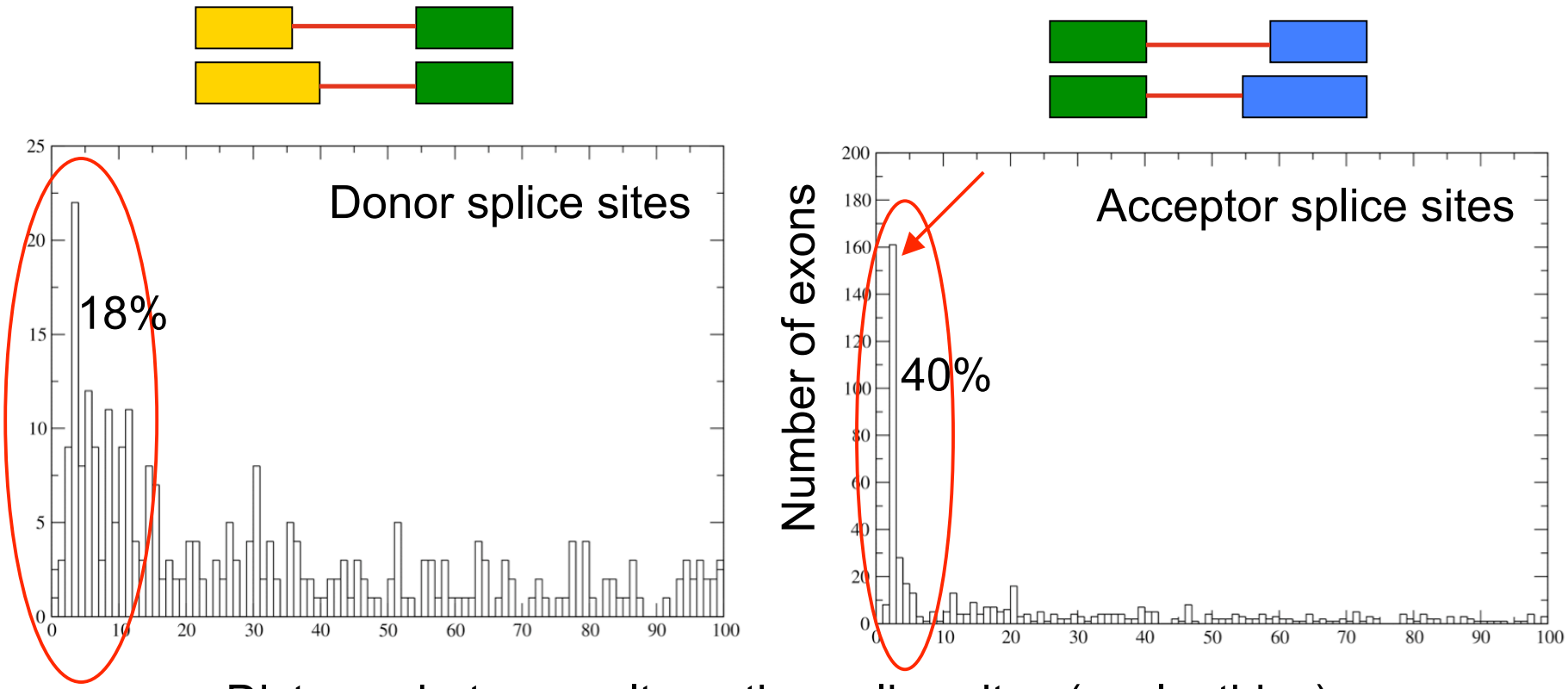
Distance from promoter does not predict inclusion/exclusion



Distance from promoter does not predict inclusion/exclusion



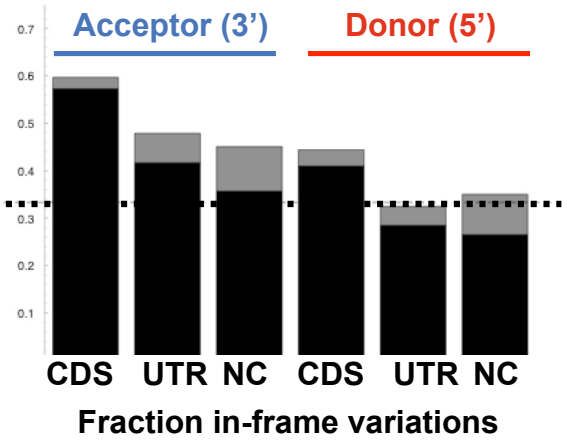
Small variations in exon length



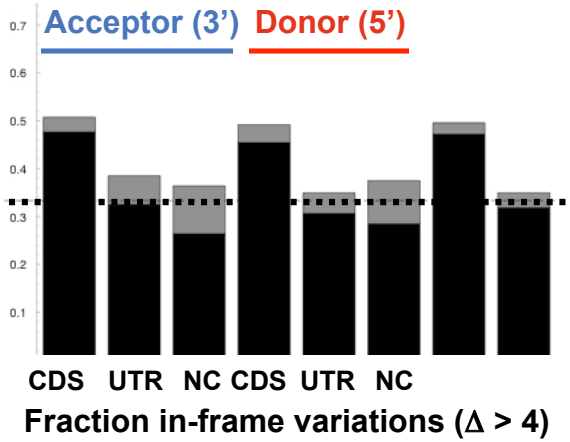
Zavolan *et al.* Genome Res. 2003

Are small variations in exon length regulated or are they mostly due to splicing “noise”?

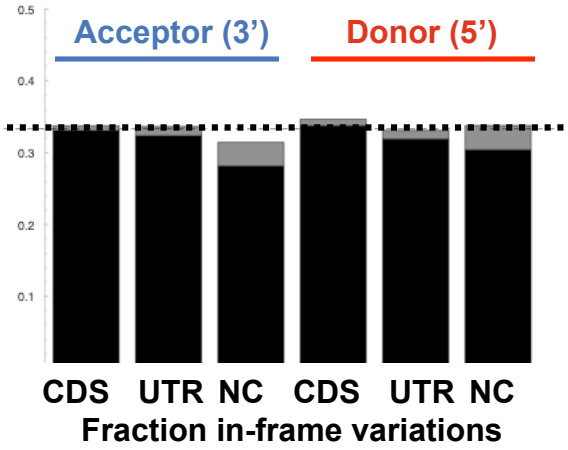
Preference for frame preservation and nonsense-mediated decay



Fraction in-frame variations



Fraction in-frame variations (Δ > 4)



Fraction in-frame variations (predicted based on sequence)

ρ_i = fraction of in-frame variations before NMD

ρ_o= fraction of observed in-frame variations

f = fraction of frame-shifting variations that survive NMD

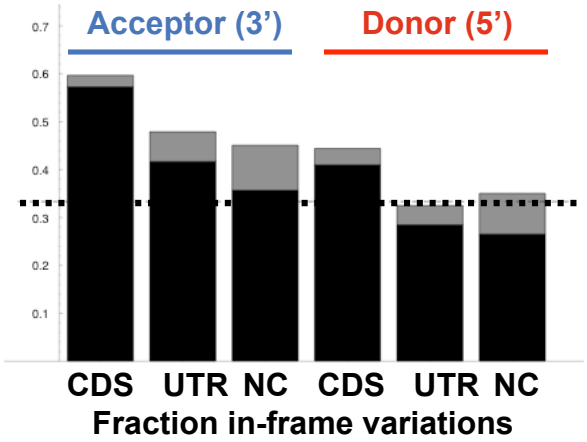
$$\rho_o = \frac{\rho_i}{\rho_i + f(1-\rho_i)}$$

ρ_i = 0.333

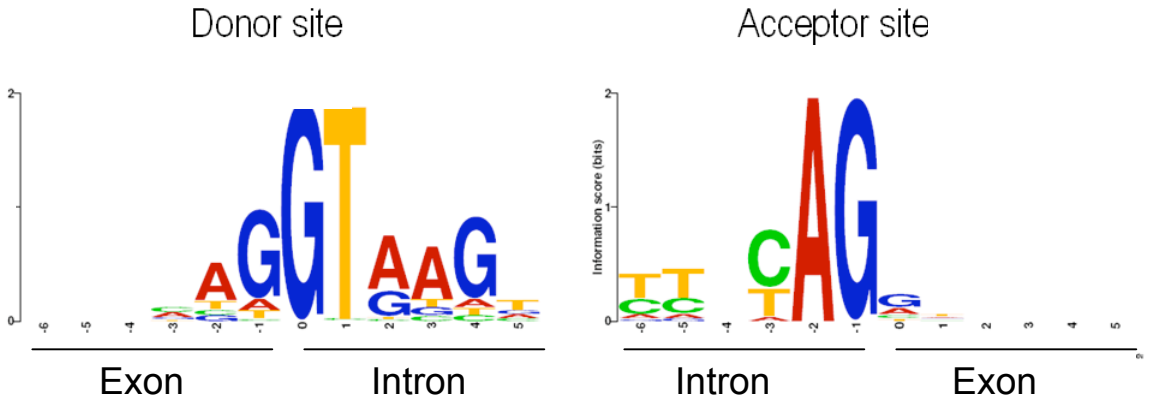
ρ_o= 0.484

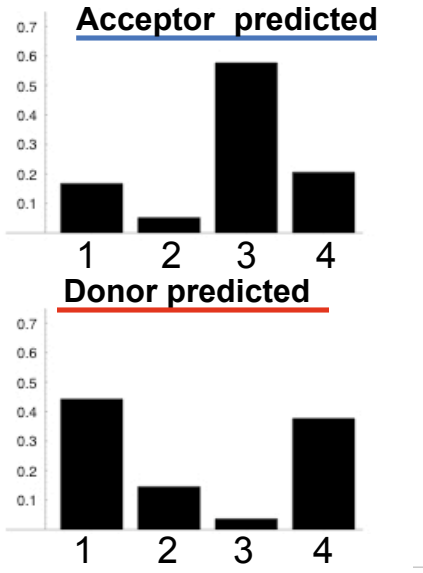
f = 0.53

Small exon variation are largely due to spliceosome “repositioning”



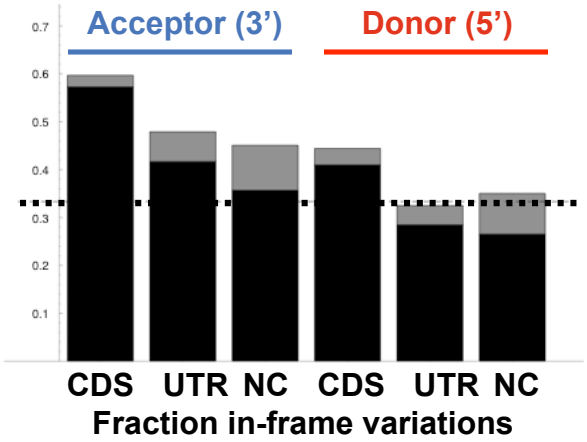
Invariant exons

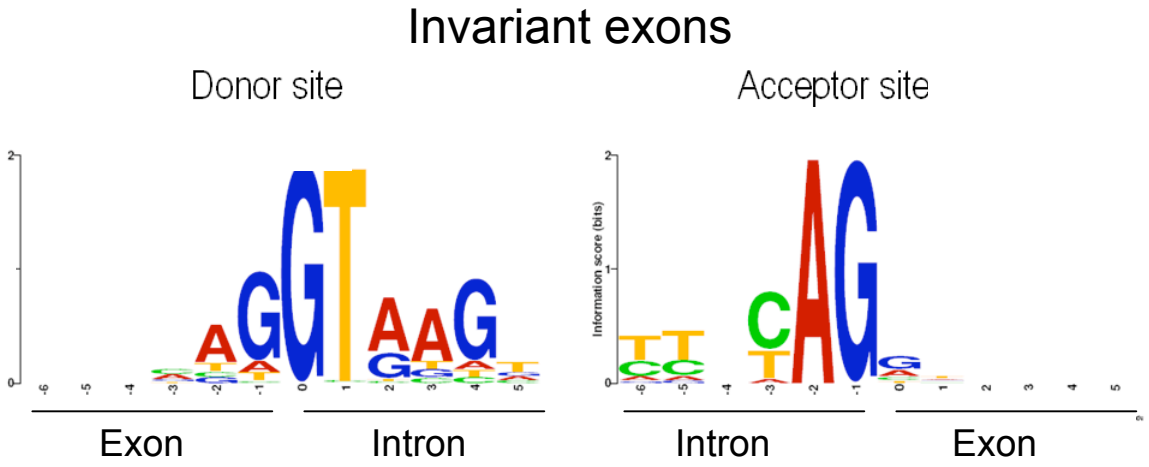




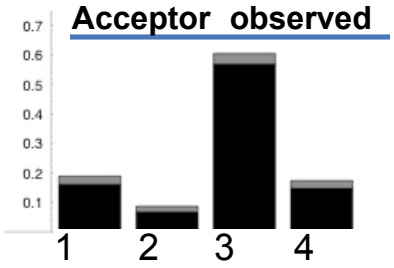
Variations of 1,2,3,4 nucleotides predicted on the basis of the sequence around observed splice sites

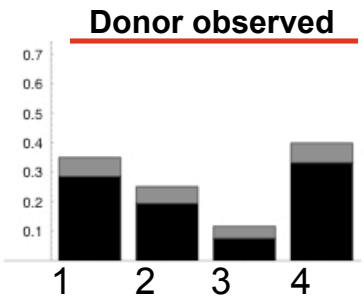
Small exon variation are largely due to spliceosome “repositioning”

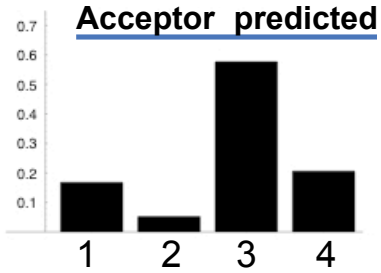


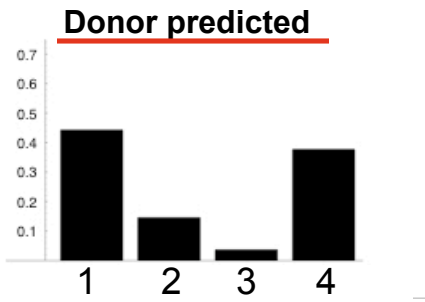


Observed variations of 1,2,3,4 nucleotides









Variations of 1,2,3,4 nucleotides predicted on the basis of the sequence around observed splice sites

Sequence-based prediction of variant tandem acceptor sites

XXXNAGNAGXXX

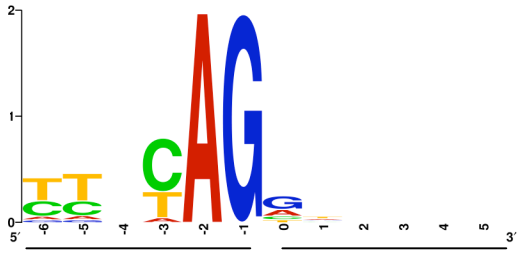
XXXNAGNAGXXX (2,557)

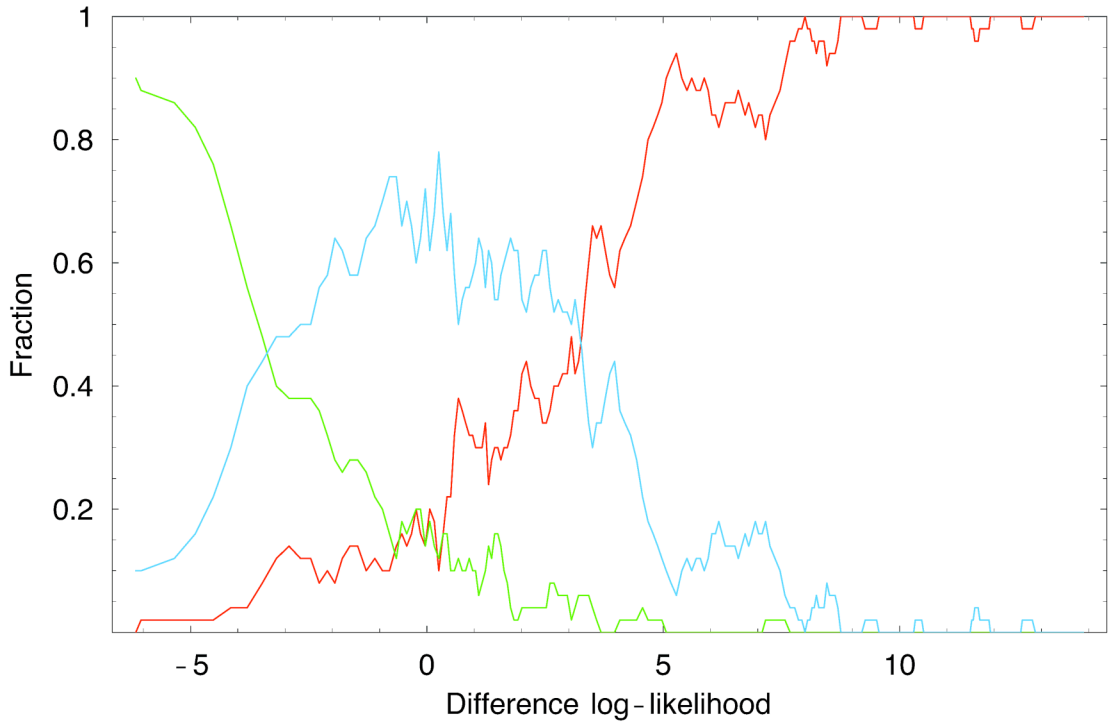
XXXNAGNAGXXX (230)

XXXNAGNAGXXX (414)

N = A, C, G, or T/U

Boundary of invariant exons





Alternative splicing

- Identification of alternative splice forms
- Regulation vs. noise in alternative splicing

Small regulatory RNAs

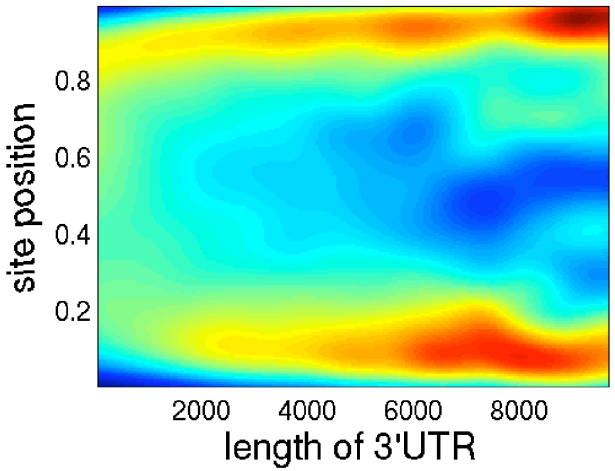
- miRNA gene identification
- miRNA expression profiling

Targeting of alternative transcripts by miRNAs

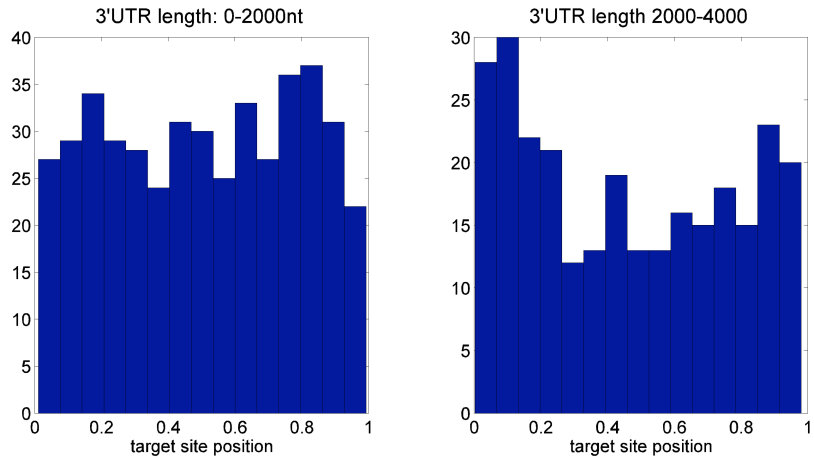
Targeting of alternative transcripts by miRNAs

Location bias of miRNA target sites

Predicted miRNA targets



siRNA off-targets



Targeting of alternative transcripts by miRNAs

Alternative 3’ UTRs of Ago2