Discoveries in Genomes and Transcriptomes

Challenges in High Throughput Sequencing Data Analysis

Peter F. Stadler

Bioinformatics Group, Dept. of Computer Science &
Interdisciplinary Center for Bioinformatics,
University of Leipzig
Max Planck Institute for Mathematics in the Sciences
RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology
Institute for Theoretical Chemistry, Univ. of Vienna (external faculty)
Center for non-coding RNA in Technology and Health, U. Copenhagen
The Santa Fe Institute (external faculty)

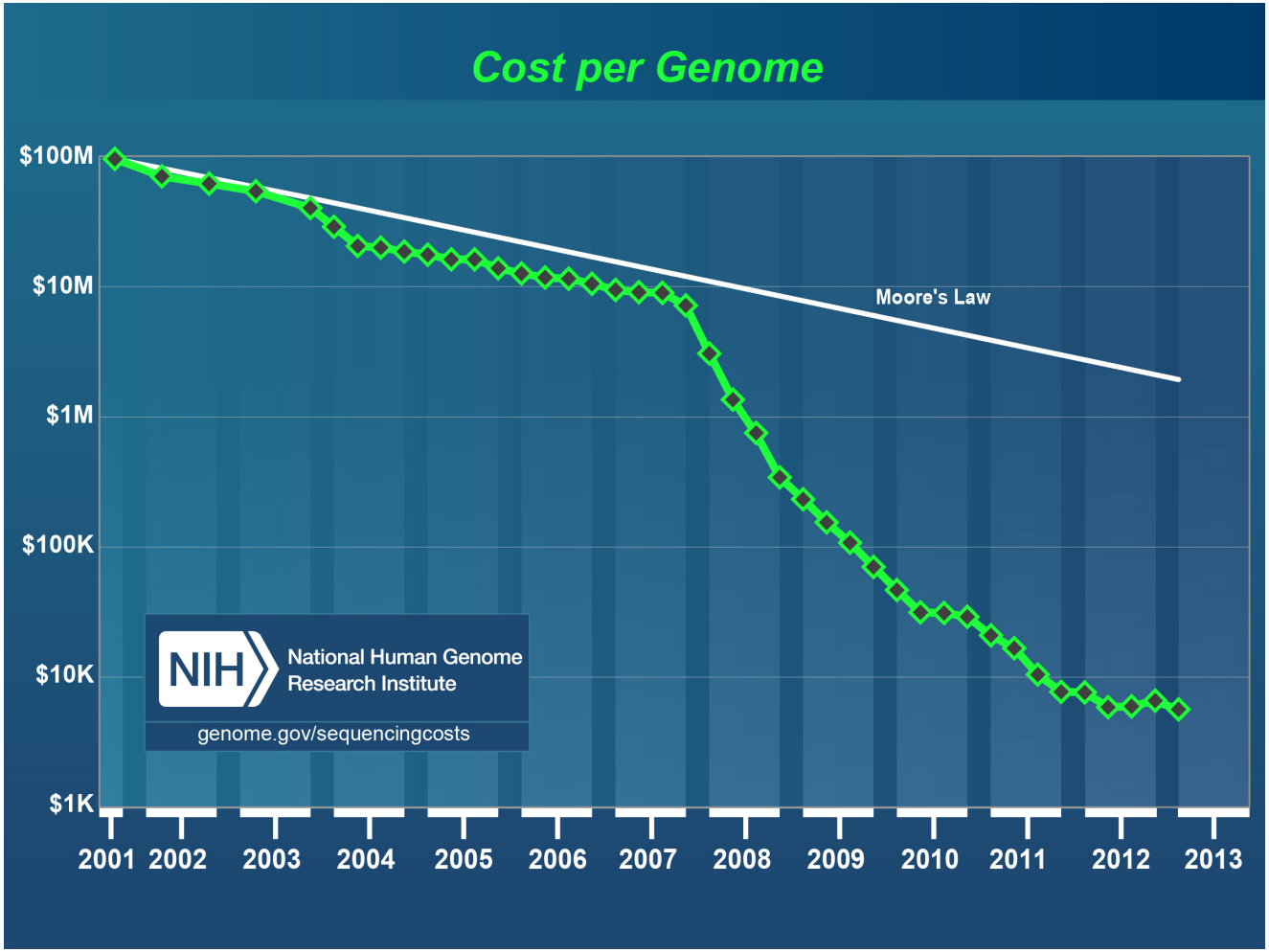
WU Wien, 26 Apr 2013

there be dragons ...

PART I:

Discoveries in Sequence Space

Progress in Sequencing Technologies

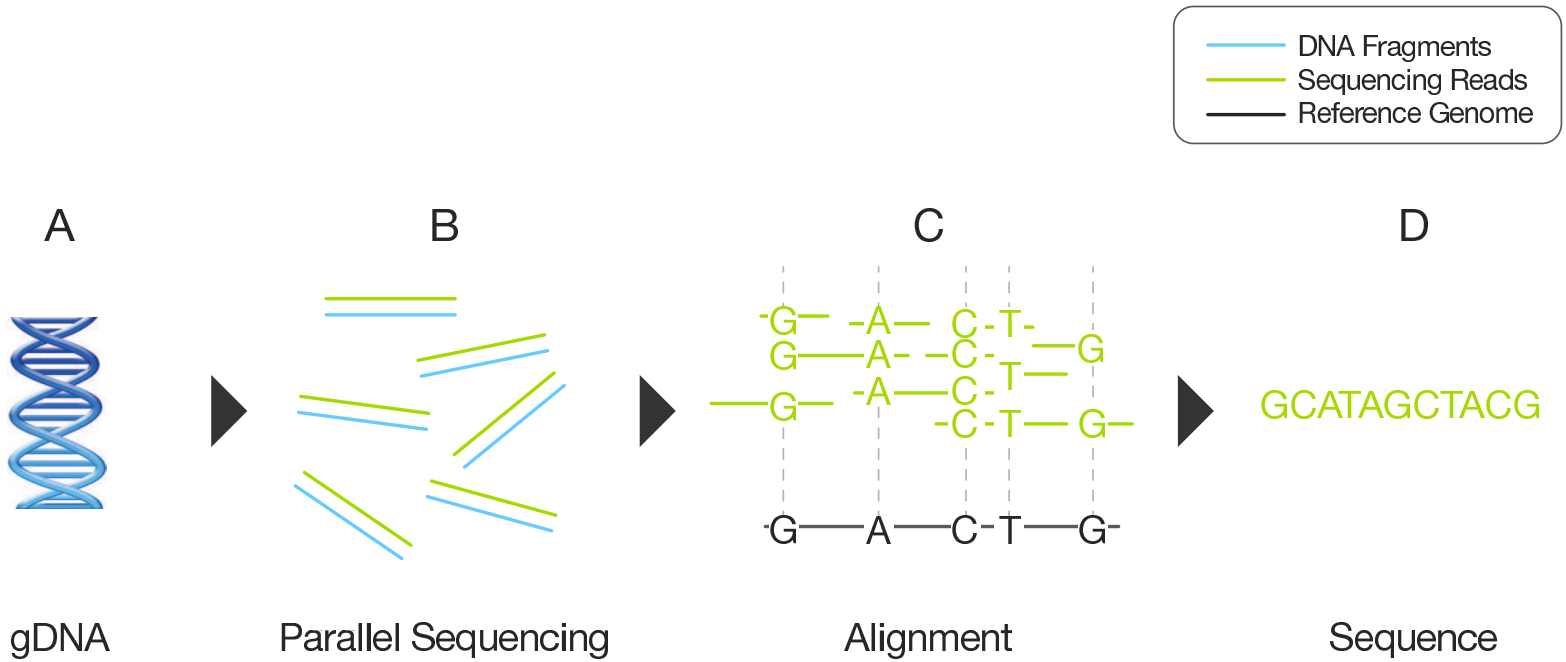


The Data Deluge

- Genome size (human) 3 Gb
- Transcripts:
  $\sim 20,000$ genes
  $10^6 ... 10^7$ RNA products (crude estimate)
- Sequencing run (Illumina HiSeq 2500) 600 Gb in 6×10^9 reads

Basic Workflow

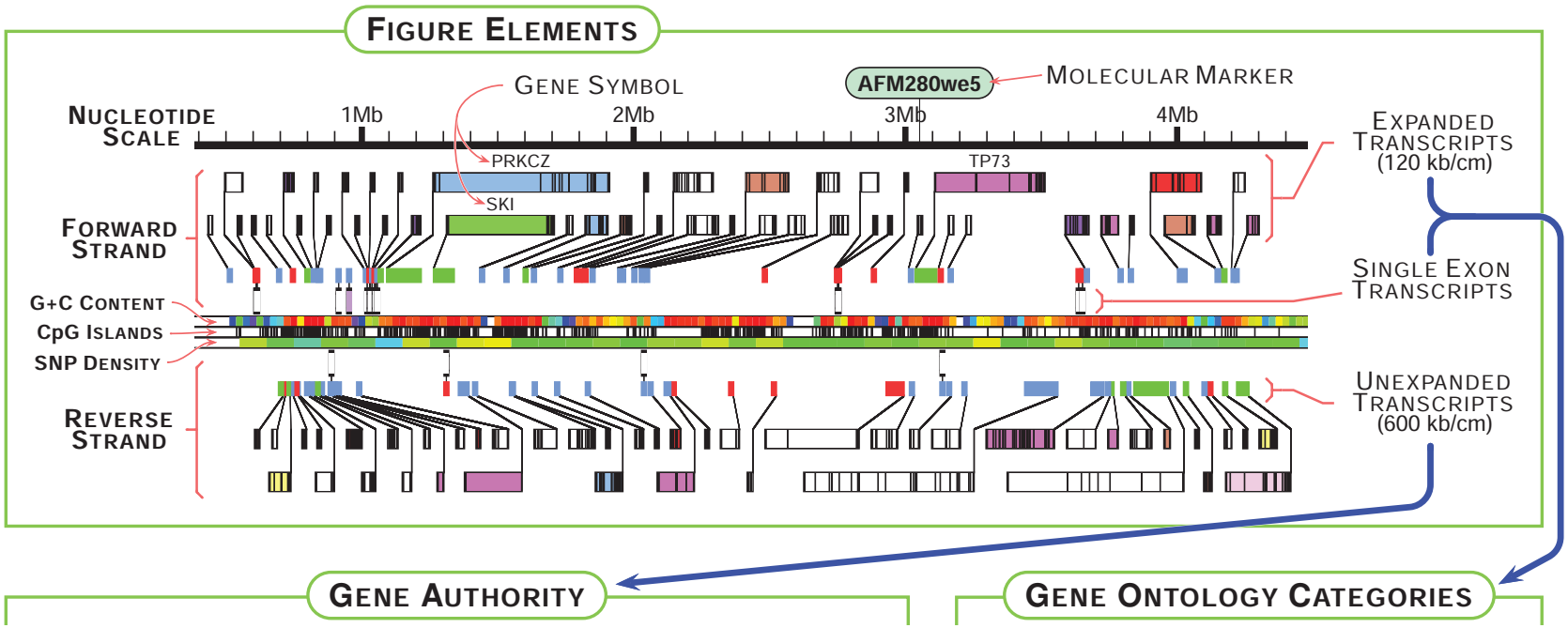
Figure 1: Conceptual Overview of Whole-Genome Resequencing



A. Extracted gDNA.
B. gDNA is fragmented into a library of small segments that are each sequenced in parallel.
C. Individual sequence reads are reassembled by aligning to a reference genome.
D. The whole-genome sequence is derived from the consensus of aligned reads.

Just a Decade ago ...

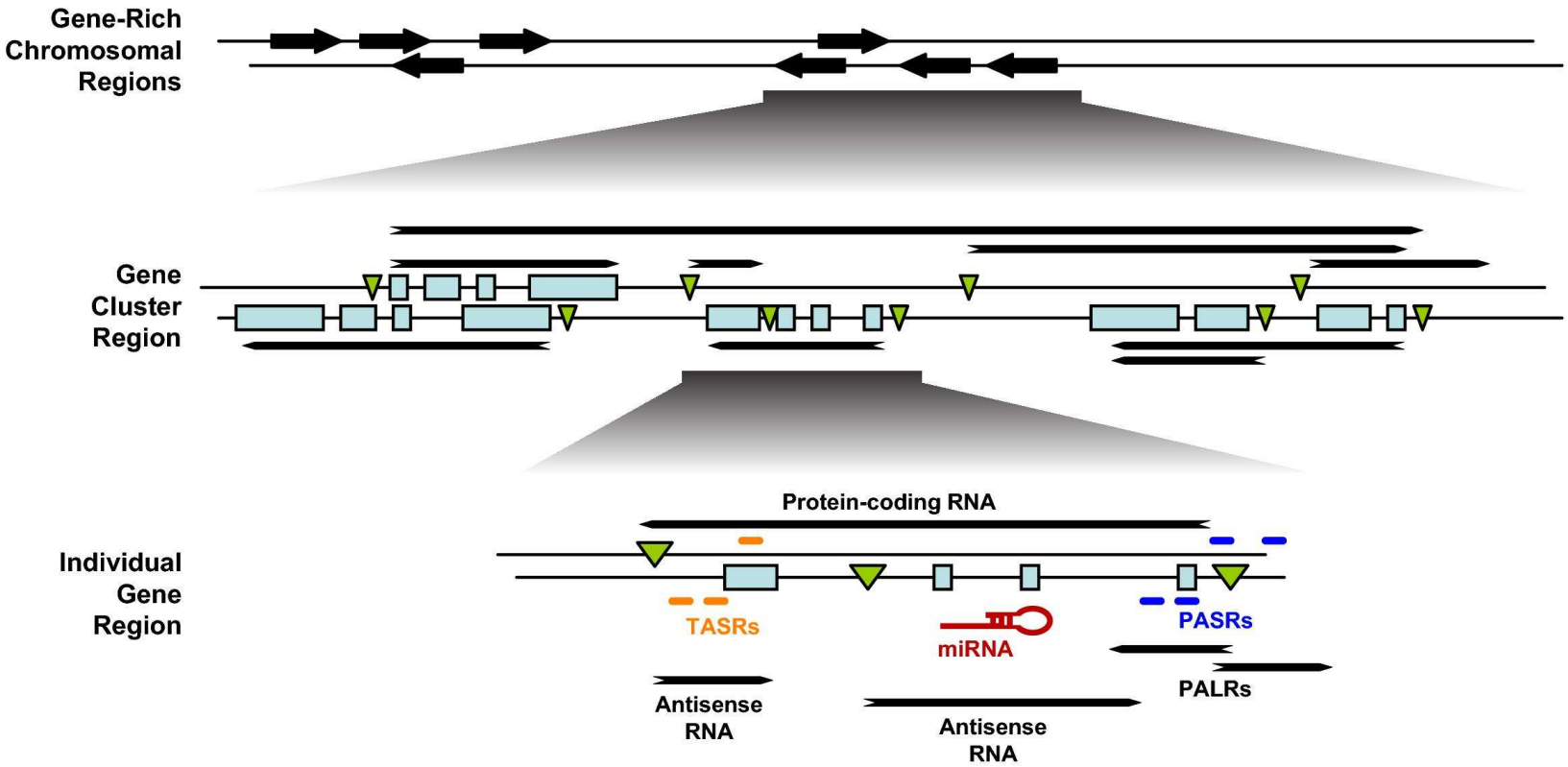
... we firmly believed that individual, separable genes are arranged like beads on a string ...



Celera genome paper, Science **291**: 1304-1351 (2001)

Transcriptome Complexity

after few years of high throughout transcriptomics we see a complex network of interleaved transcriptional activity



Science 316: 1484-1488 (2007)

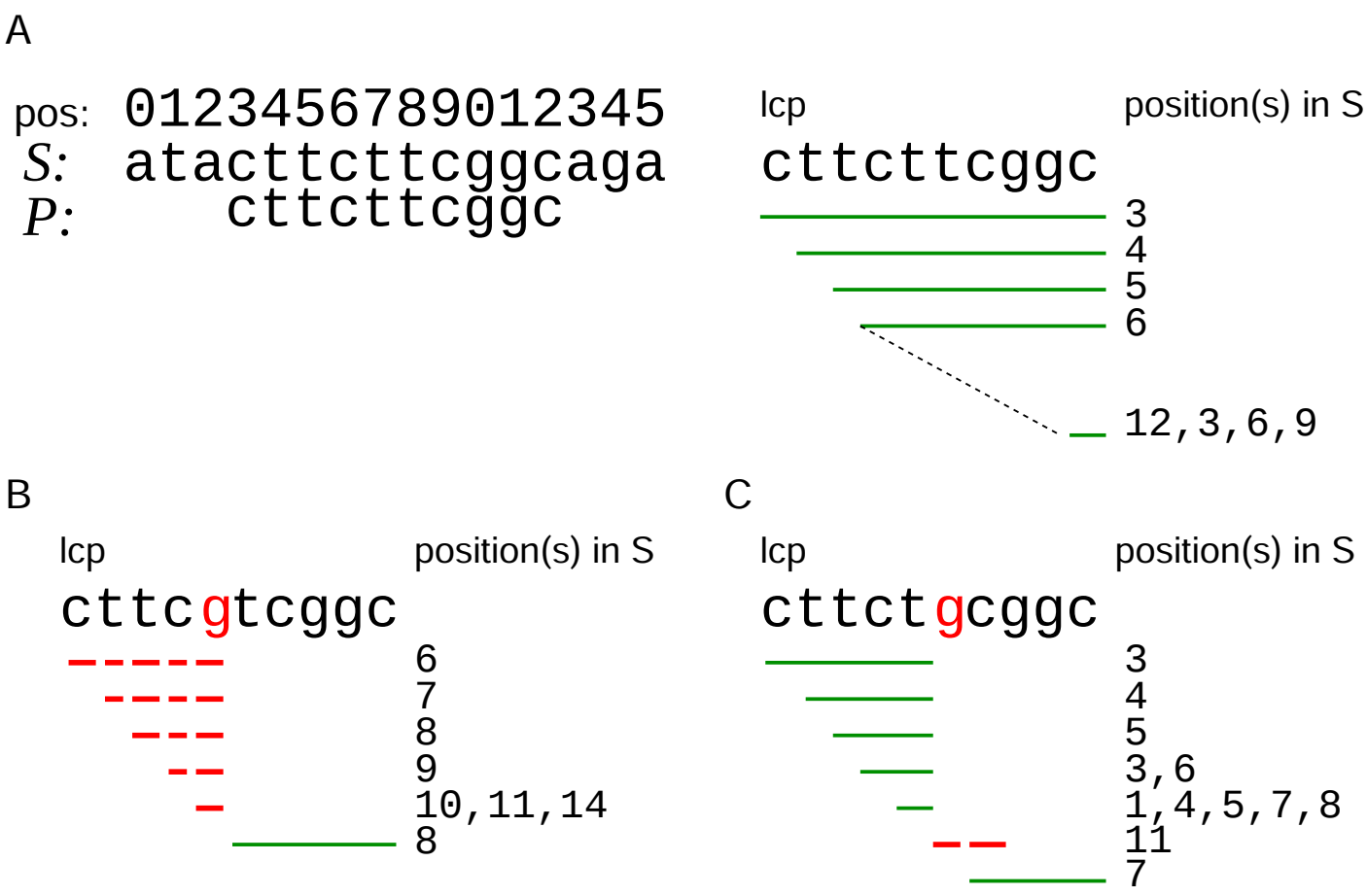
Transcriptomics

Analysis of expressed RNAs.
Simplest case: A reference genome is known
Mapping Problem Align reads to the reference genome
Efficient Read-Mapping with In/Dels: `segemehl`

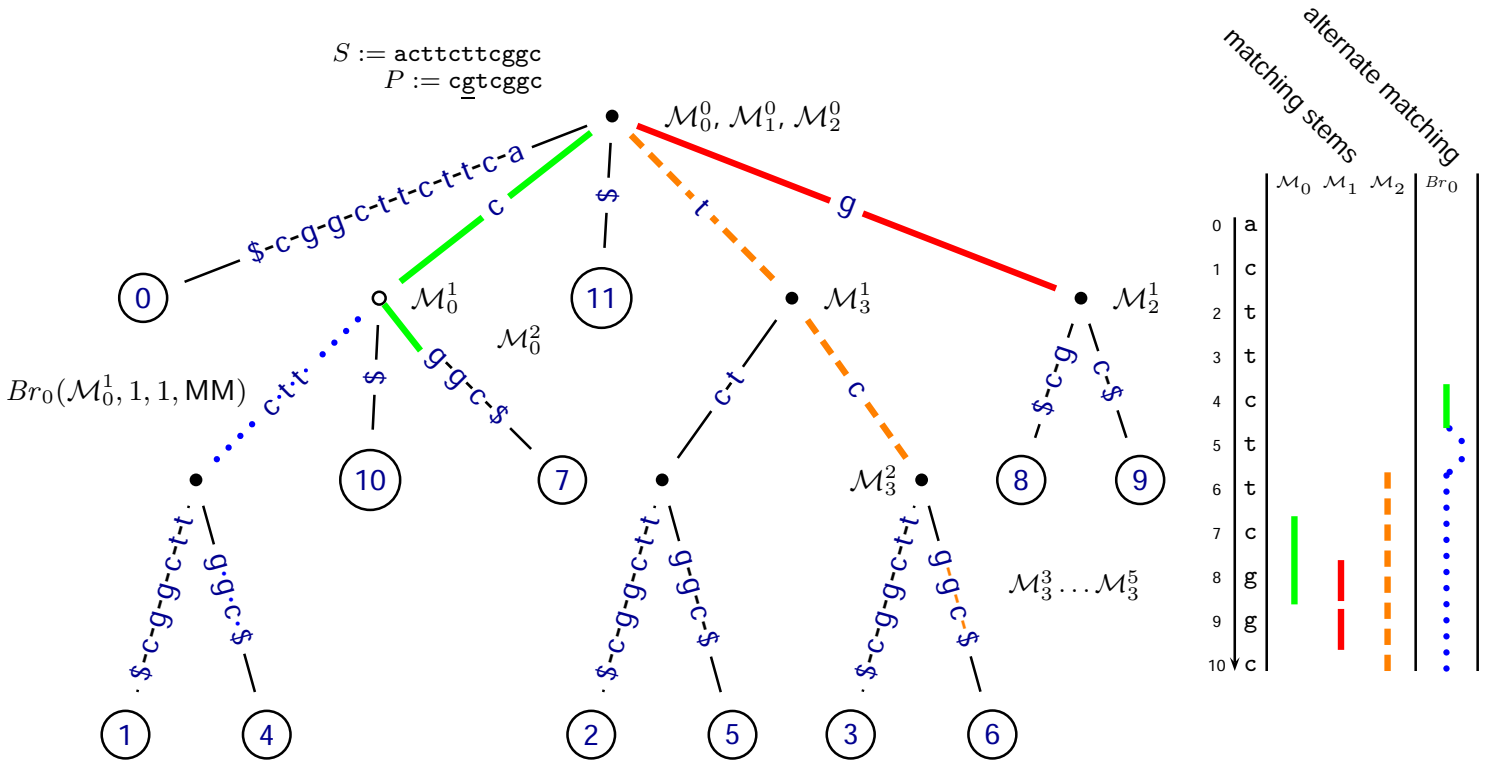
- **In principle**, mapping reads to the genome is a simple local alignment problem.
- **In practise**, there are several problems:
  - huge volume of data $\rightarrow$ classical methods to slow
  - index-based methods (suffix trees, suffix arrays) have problems with in/dels
  - short reads: problems with significance

Suffix trees or the more efficient suffix arrays solve the problem if there are few mismatches and in/dels.

The problem: Longest prefix matches may fail to deliver the position of the optimally scoring local alignment

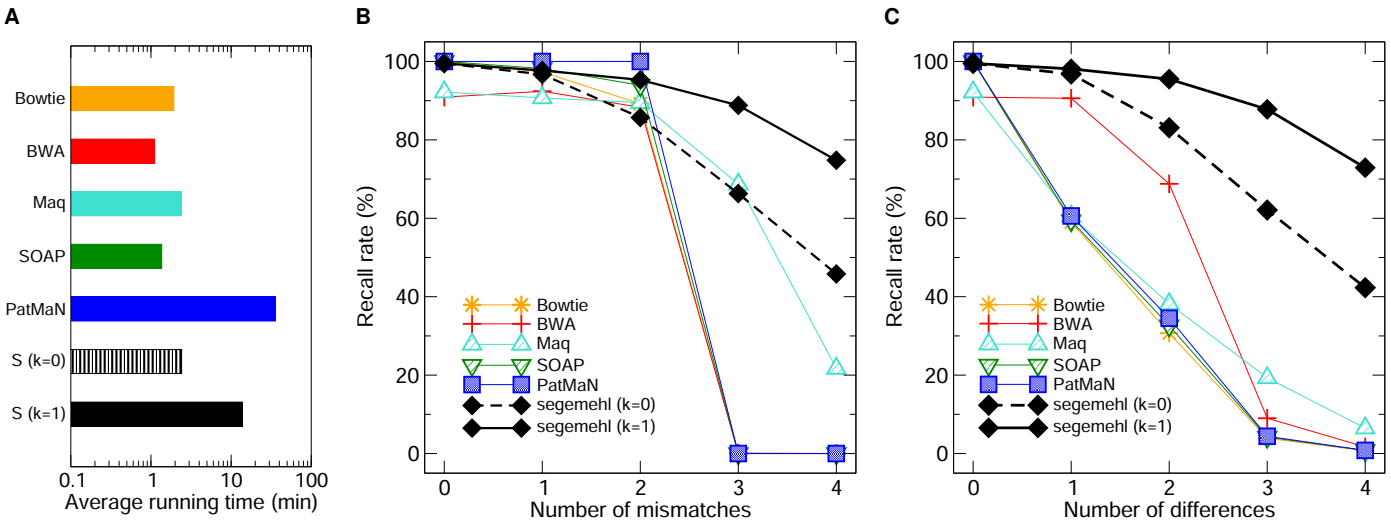


The solution: “matching stems” allow to “jump over” individual mismatches and in/dels.



PLoS Comp. Biol. **5** e1000502 (2009)

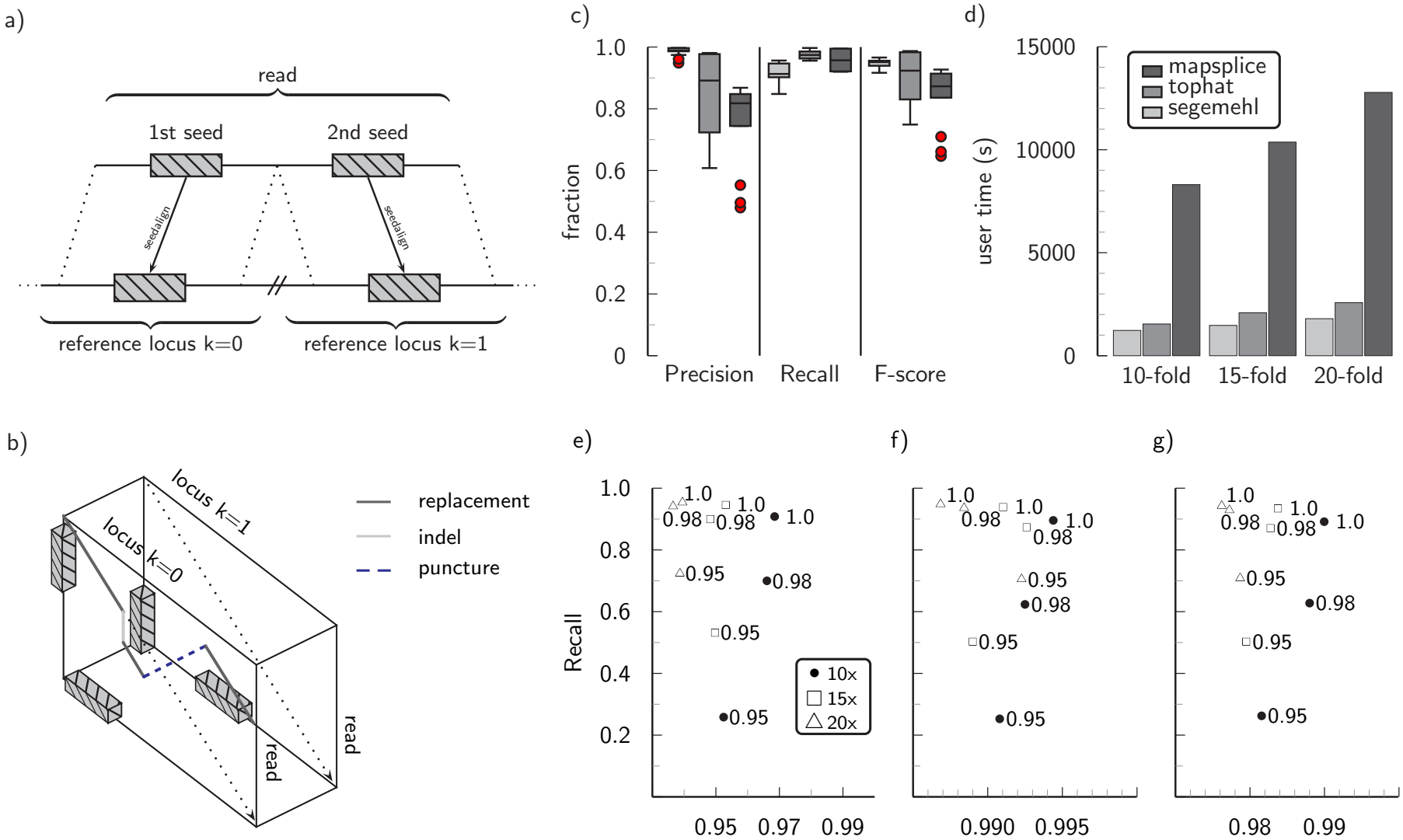
Performance: yes it works [almost as good as full enumeration of all mismatch/indel combinations], it is (reasonably) fast, and it can deal very well with poor-quality reads.



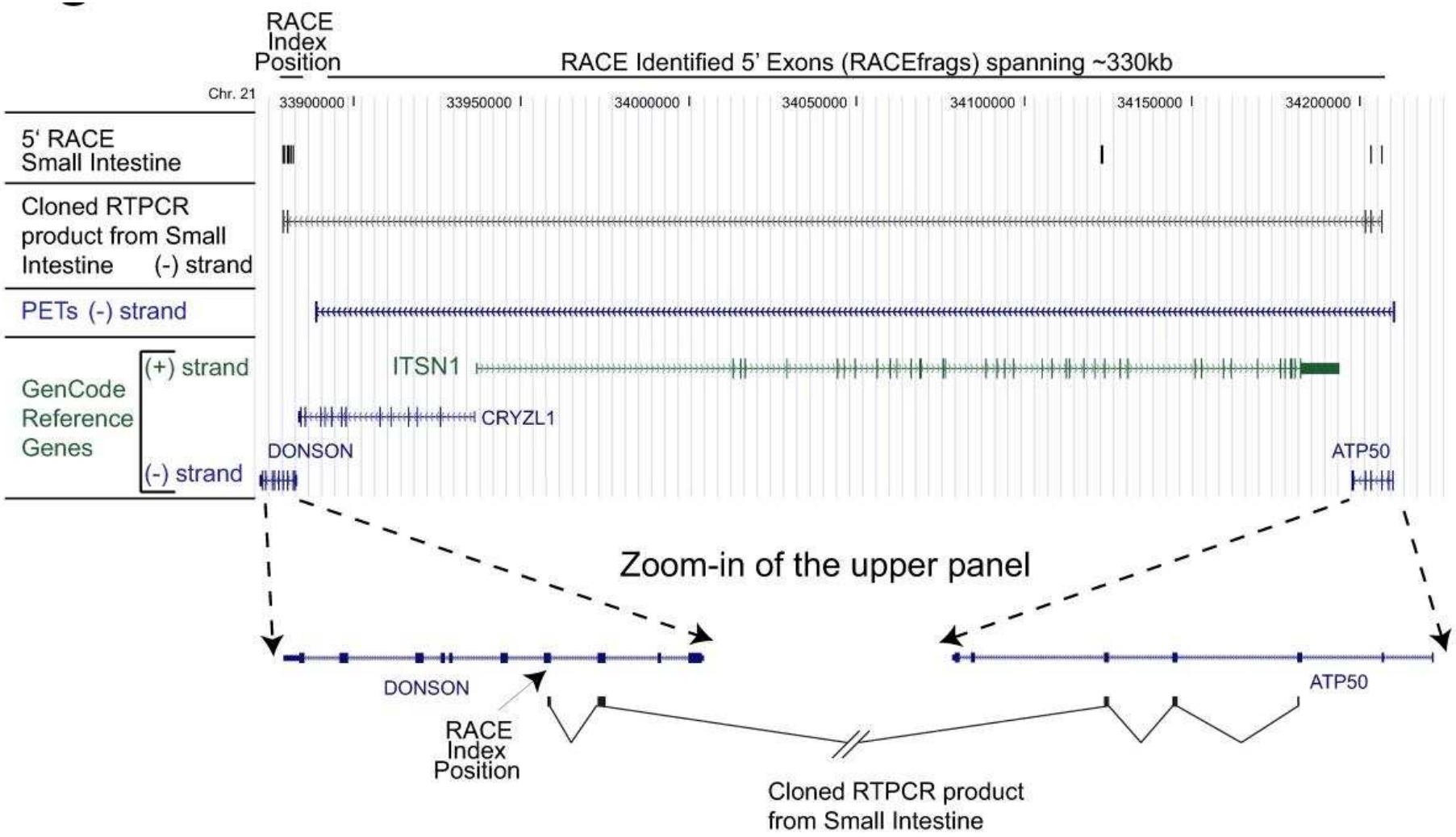
. . . as customary, your own methods always works best :-)

PLoS Comp. Biol. **5**: e1000502 (2009)

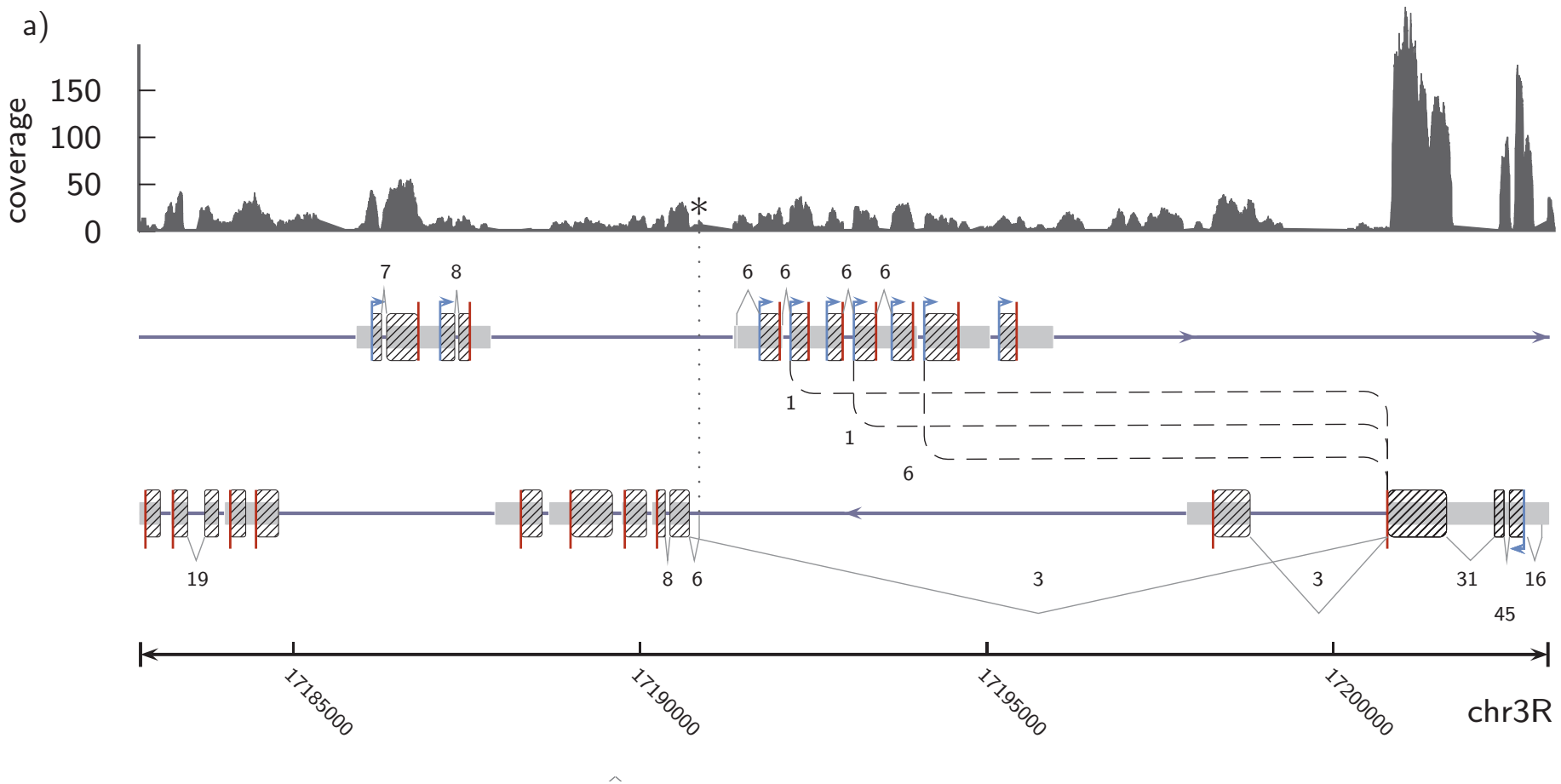
Mapping Split Reads: segemehl



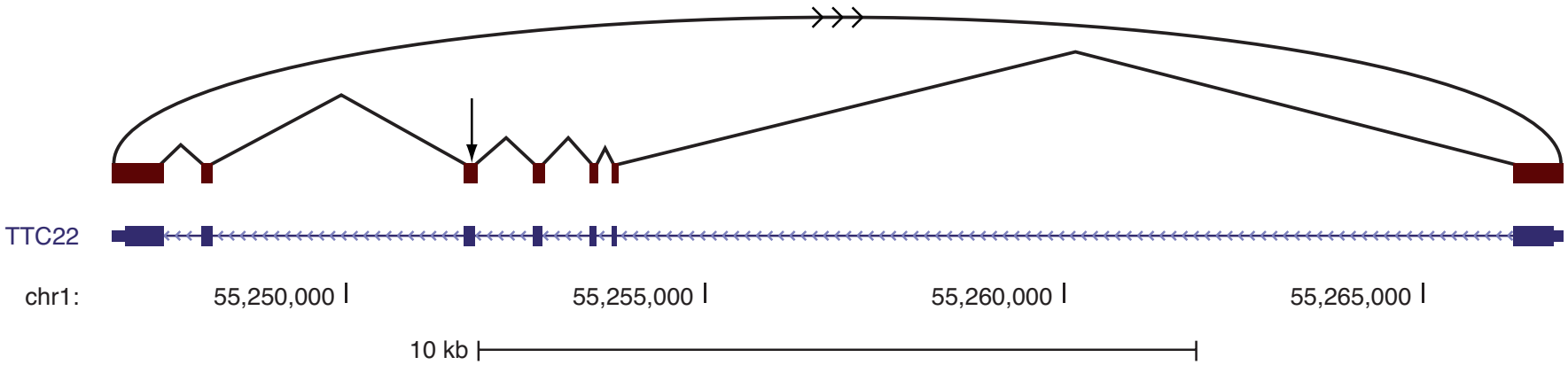
Mosaic Transcripts



not uncommon in ENCODE data ...

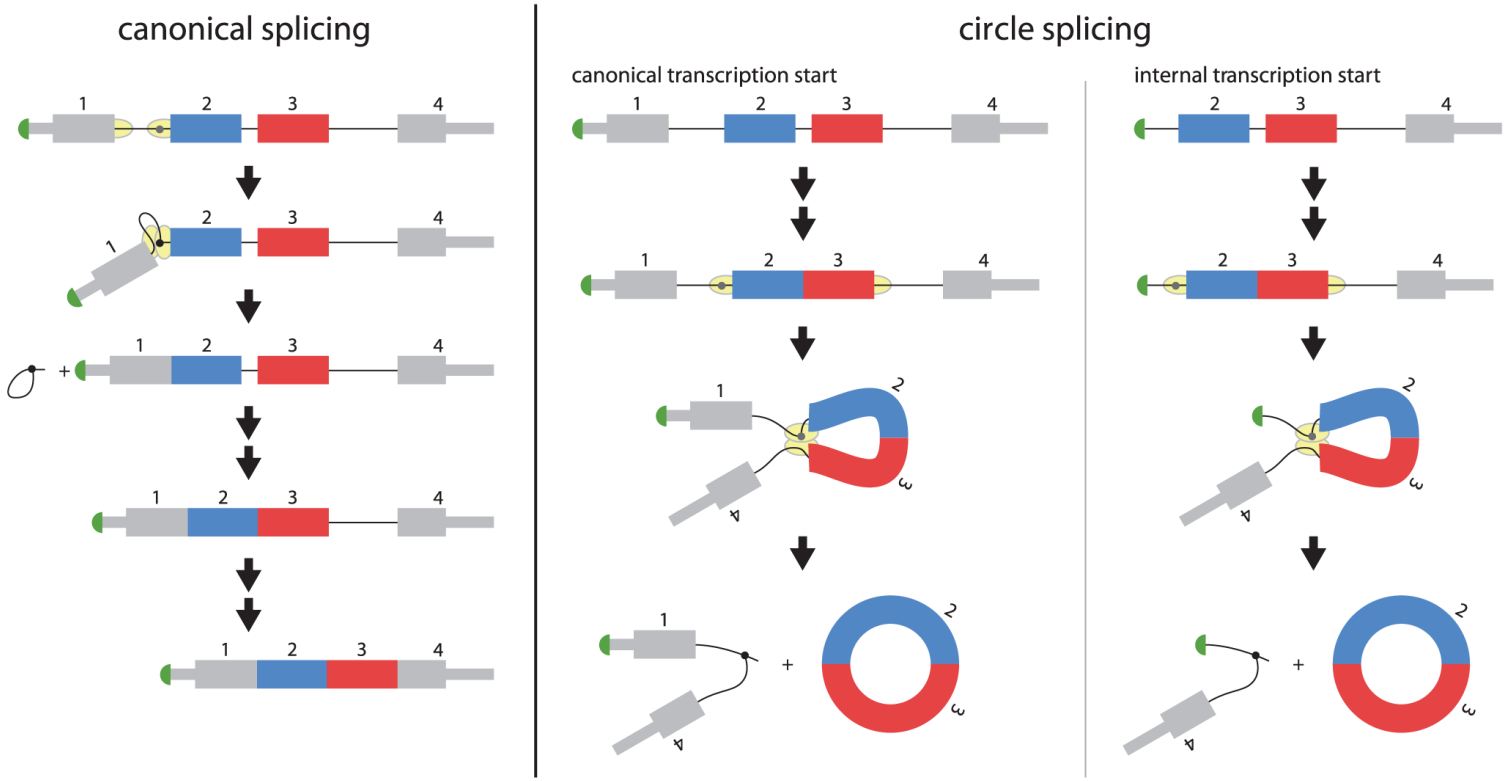

a)
coverage
150
100
50
0
7
8
6
6
6
6
1
1
6
19
8
6
3
3
31
16
45
17185000
17190000
17195000
17200000
chr3R

Circular Transcripts



Abundant circular transcripts

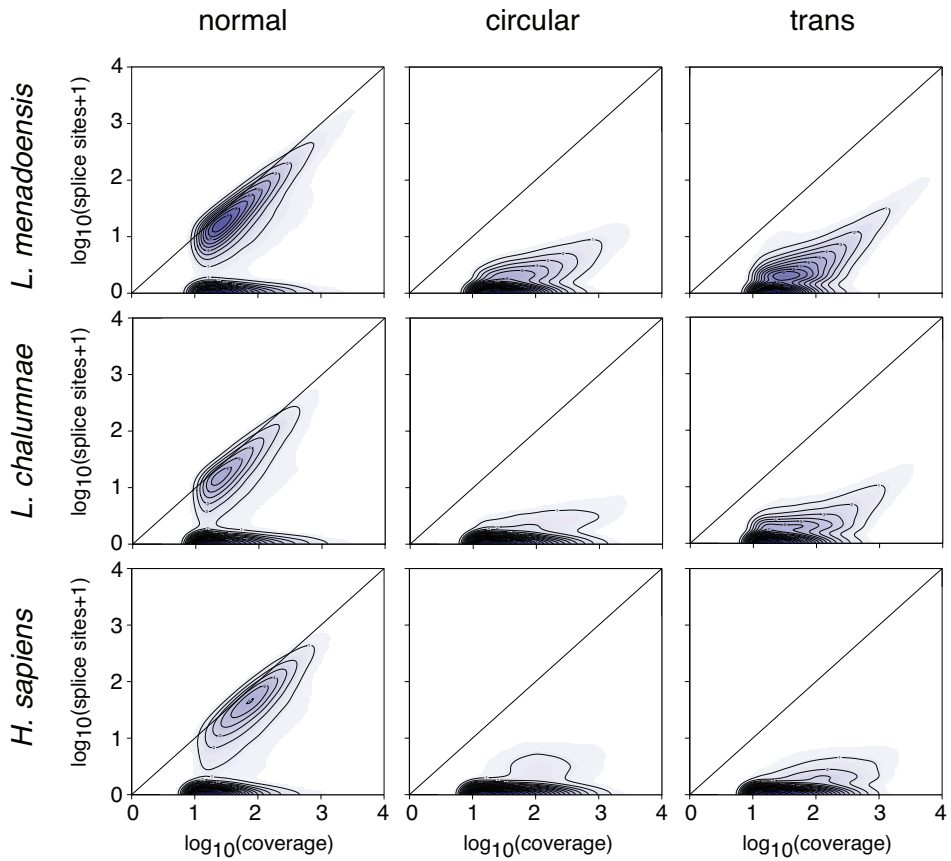
Generation of Circular Transcripts



Salzman et al PloS ONE 2011

Circular transcripts are functional e.g. in the ANRIL ncRNA

Abundant circular and chimeric transcripts



Latimeria menadoensis

(genome just published)

seems to be a generic feature of (at least) vertebrate genomes

PART II:

What's The Function

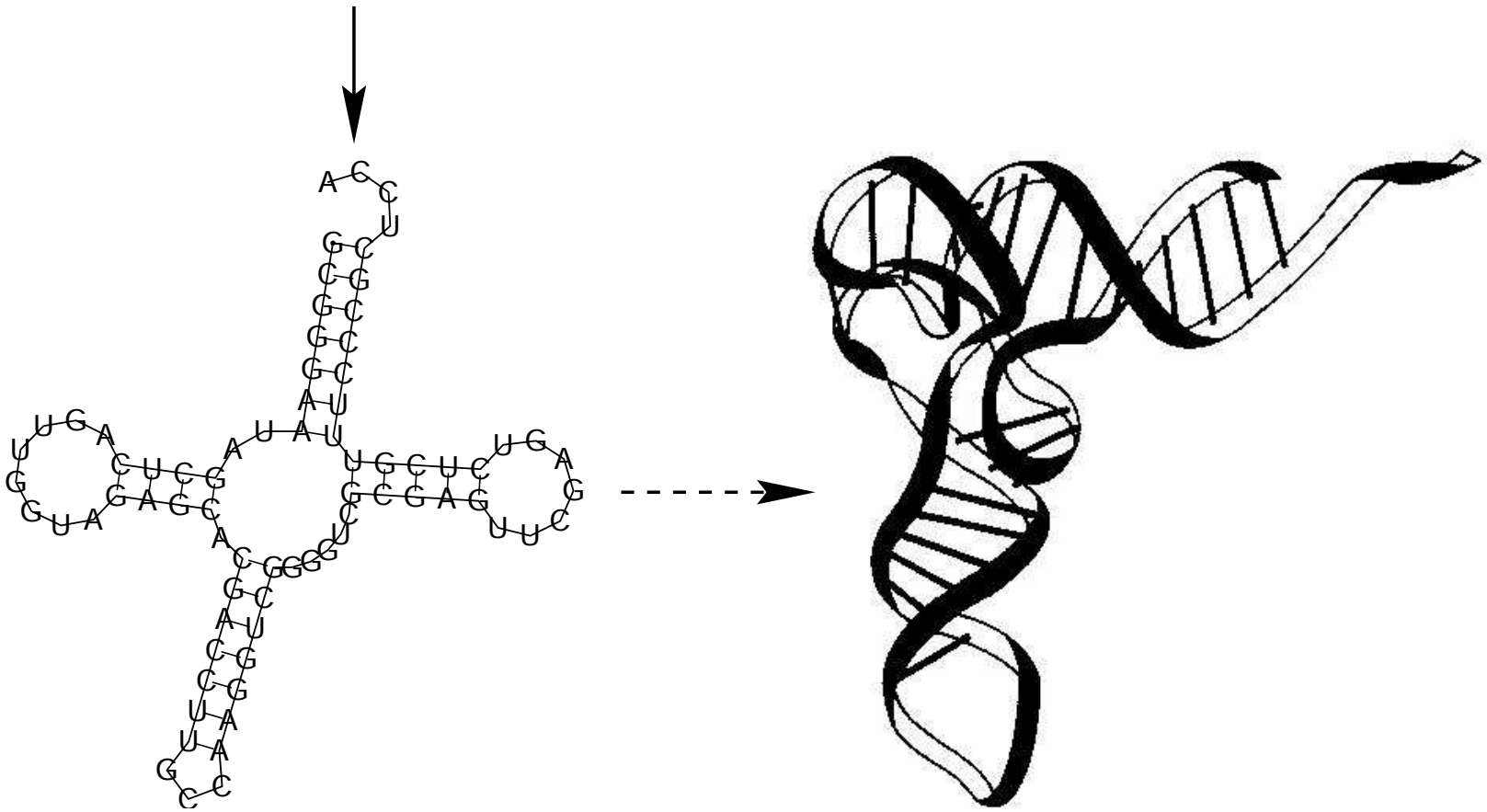
of all these RNAs

The RNA Zoo

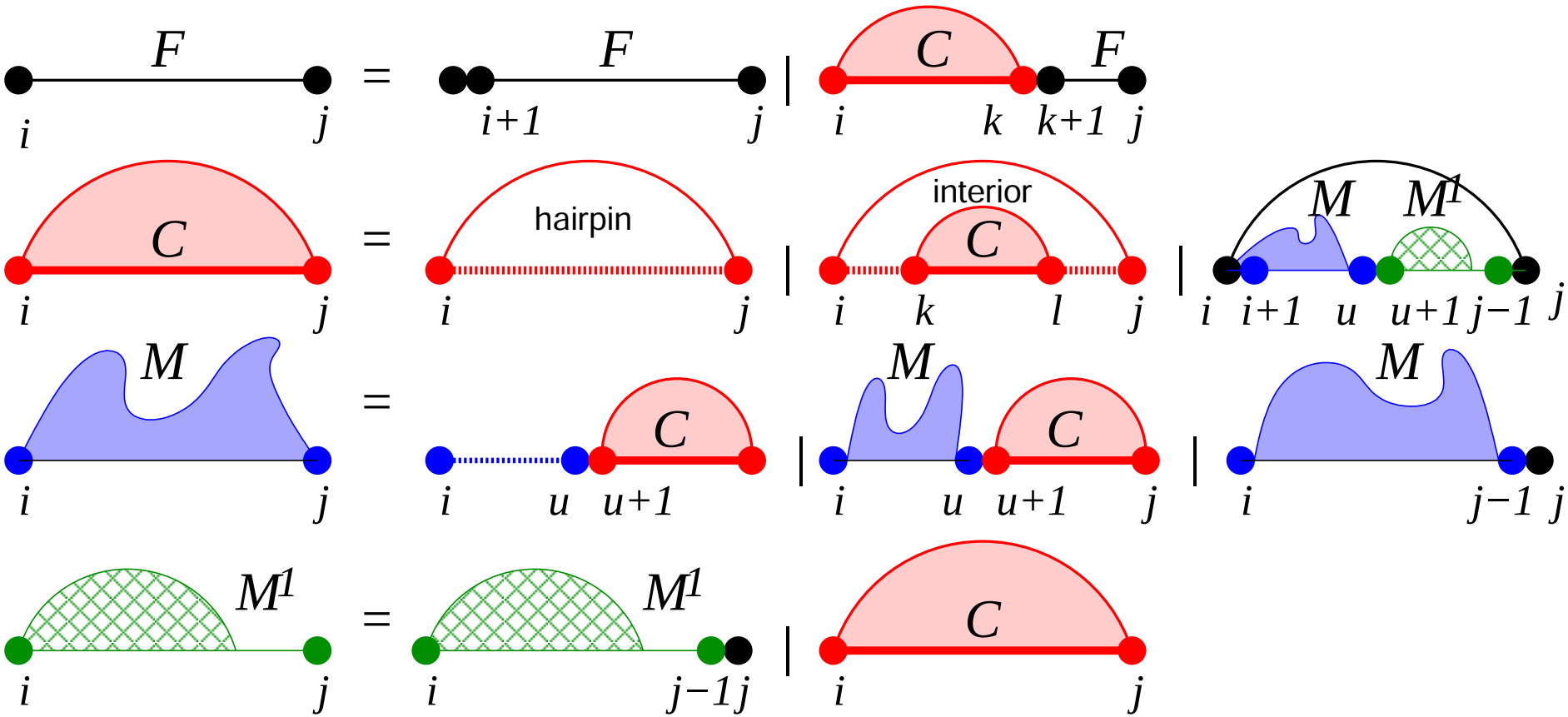
1. small RNAs
   microRNAs, piRNAs, siRNAs, xiRNAs, ...
2. medium-size housekeeping RNAs
   tRNAs, snoRNAs, snRNAs, etc
   typically very well-conserved sequence and secondary structure
3. long RNAs
   usually not well conserved, only small structural elements under stabilizing selection

RNA Secondary Structures

GCGGGAAUAGCUCAGUUGGUAGAGCACGACCUUGCCAAGGUCGGGGUCGCGAGUUCGAGUCUCGUUUCCCGCUCCA



RNA Folding

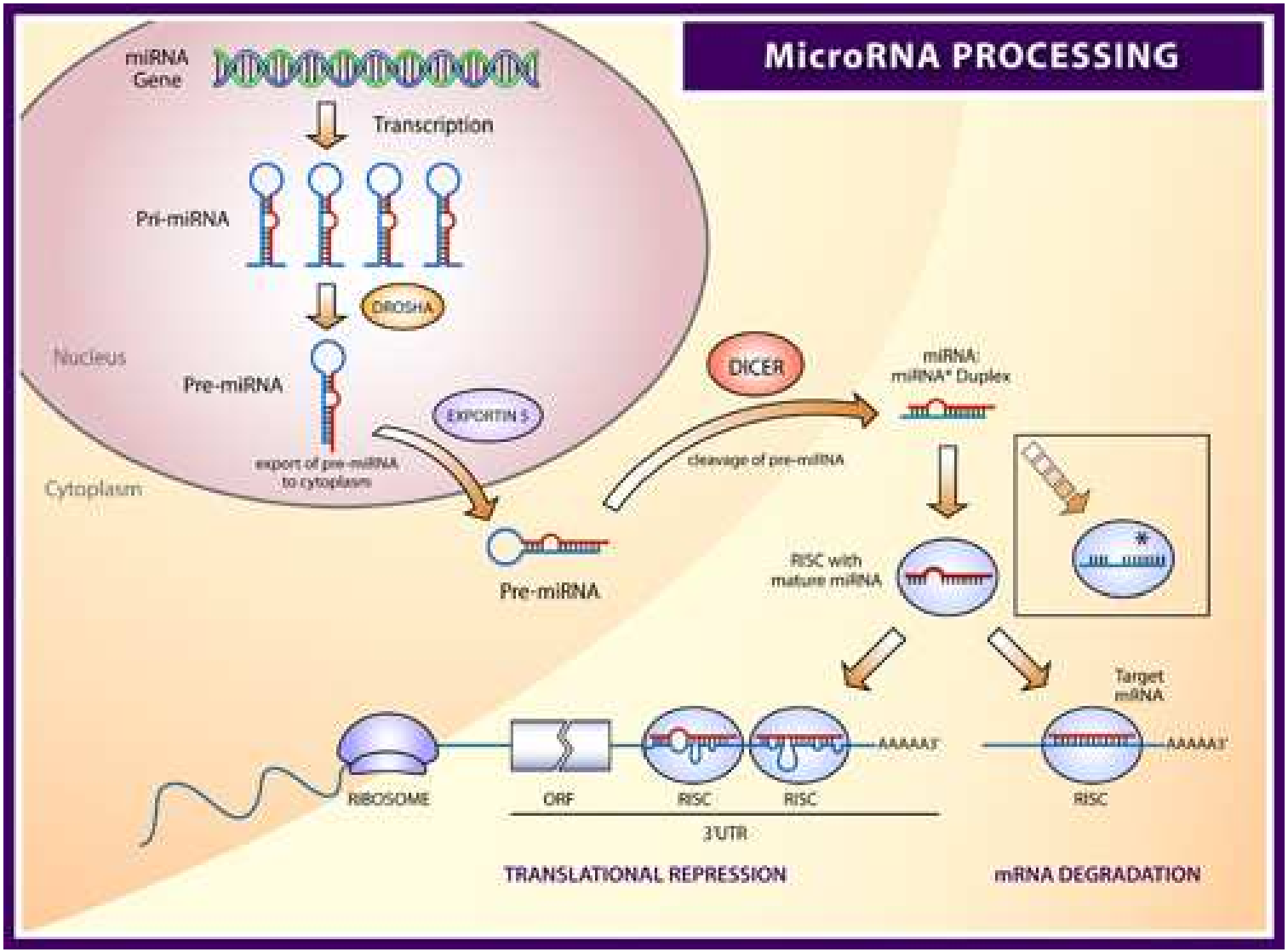


efficient solution by Dynamic Programming

Vienna RNA Package

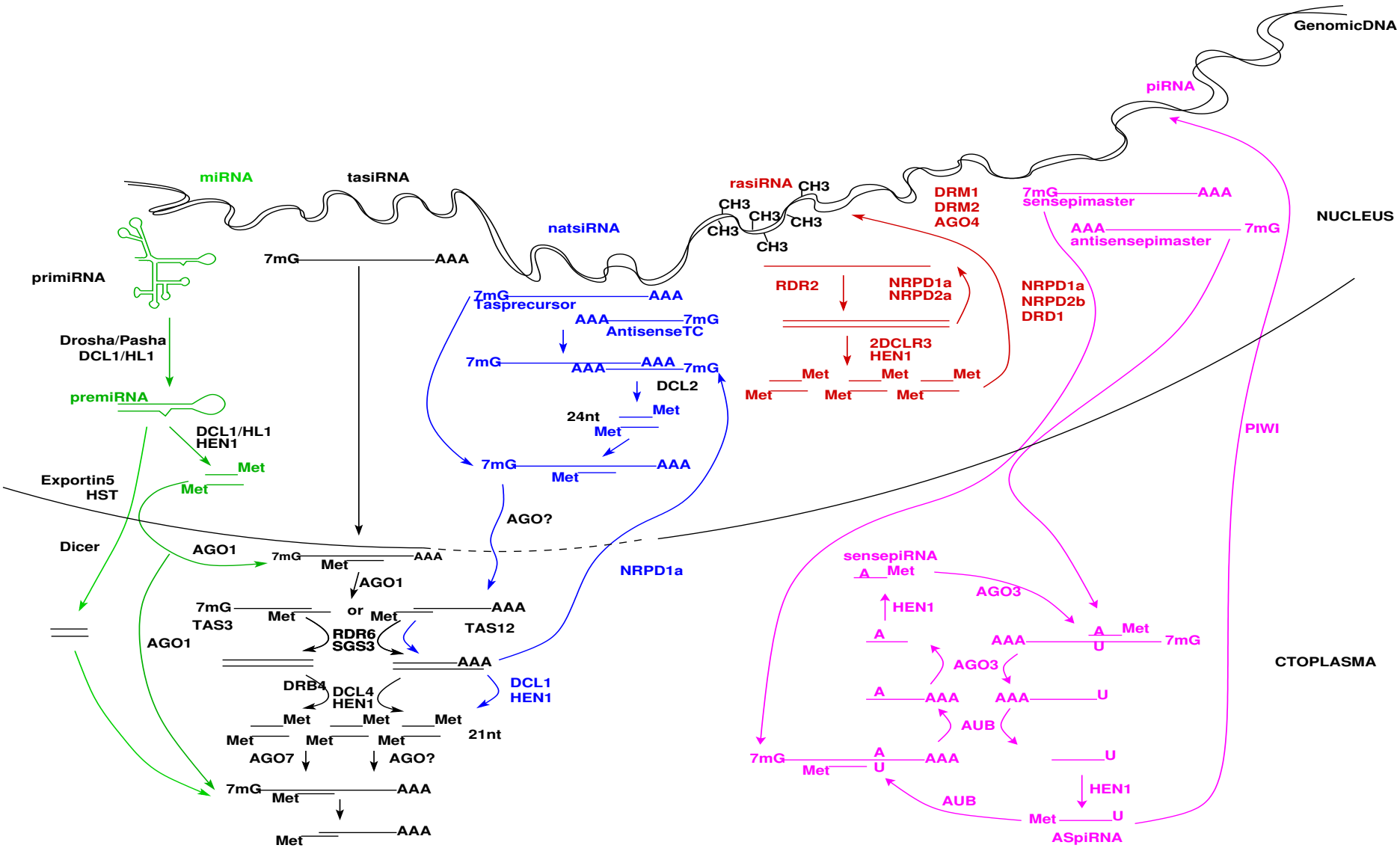
Monatsh.Chem. **124**: 167-188 (1994), Alg.Mol.Biol. **6**: 26 (2011)

MicroRNAs



http://biologiaecologia1globalwarming.files.wordpress.com

Many Pathways to Small RNAs



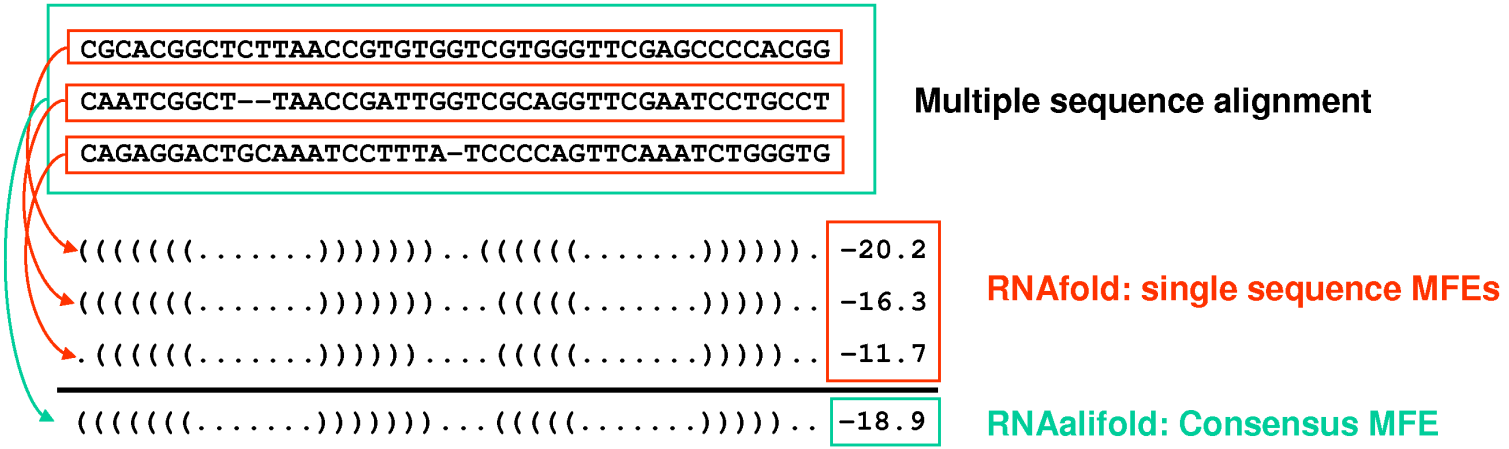
Consensus folding using `RNAalifold`

- `RNAalifold` uses the same algorithms and energy parameters as `RNAfold`
- Energy contributions of the single sequences are averaged
- Covariance information (e.g. compensatory mutations) is incorporated in the energy model.
- It calculates a consensus MFE consisting of an energy term and a covariance term:

```
((((((..(((.........))).(((((.......)))))).....(((((.......)))))))))).
GTTTCCGTAGTGTAGCGGTTATCACATTCGCCTCACACGCGAAAGGTCCCCGGTTCGATCCCGGGCGGAAACA
GTTTCCGTAGTGTAGTGGTTATCACGTTCGCCTAACACGCGAAAGGTCCCCGGTTCGAAACCGGGCGGAAACA
GTTTTCGTAGTGTAGTGGTTATCACGTGTGCTTCACACGCACAAGGTCCCCGGTTCGAACCCGGGCGAAAACA
**** ********** ********* *  ** * ******  ****************  ******* *****
(-24.76 = -23.43 +  -1.33)
```

J.Mol.Biol. **319**:1059-1066 (2002)

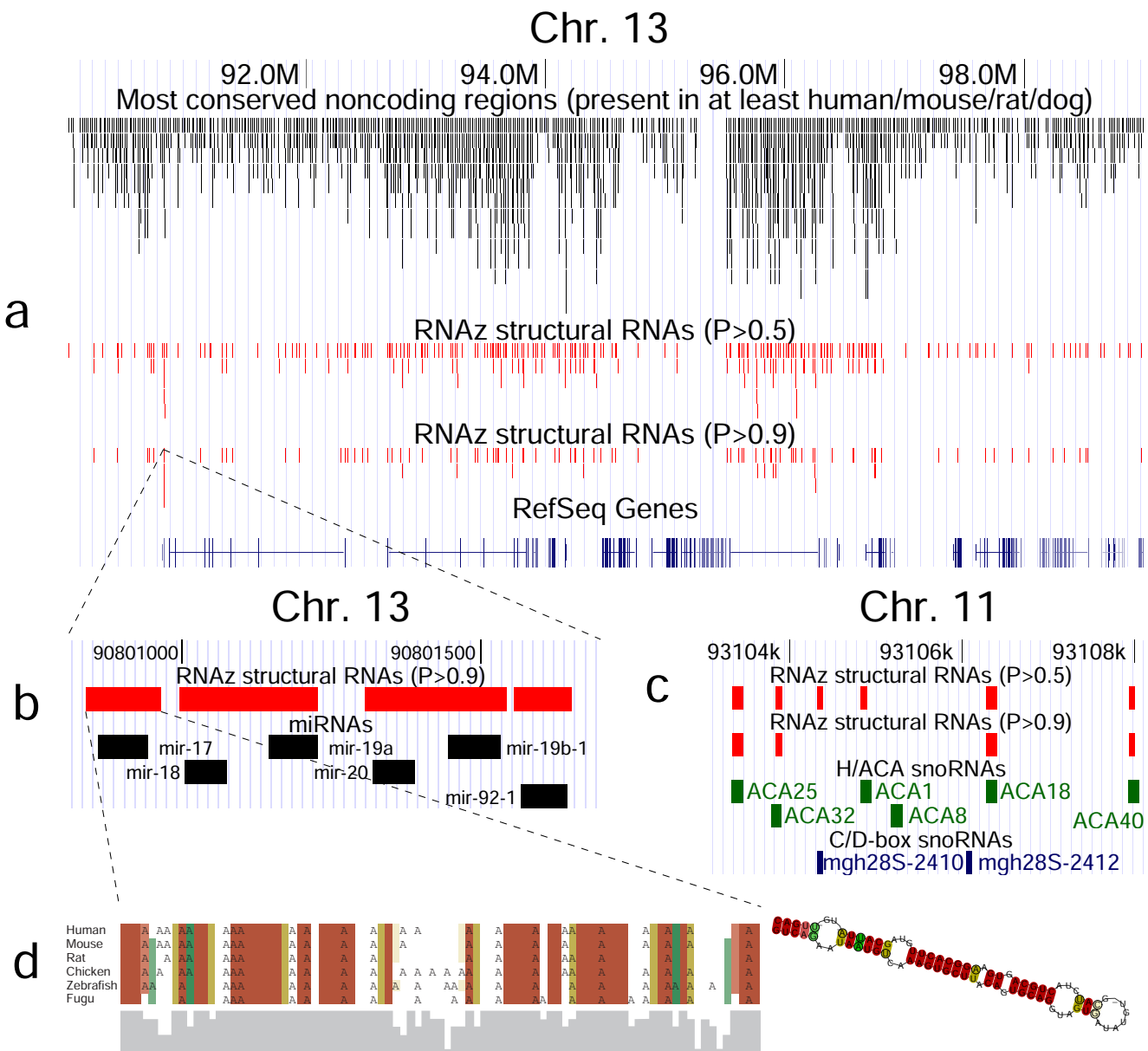
The Structure Conservation Index



$$\text{SCI} = \frac{\text{Consensus MFE}}{\text{Mean single MFEs}}$$

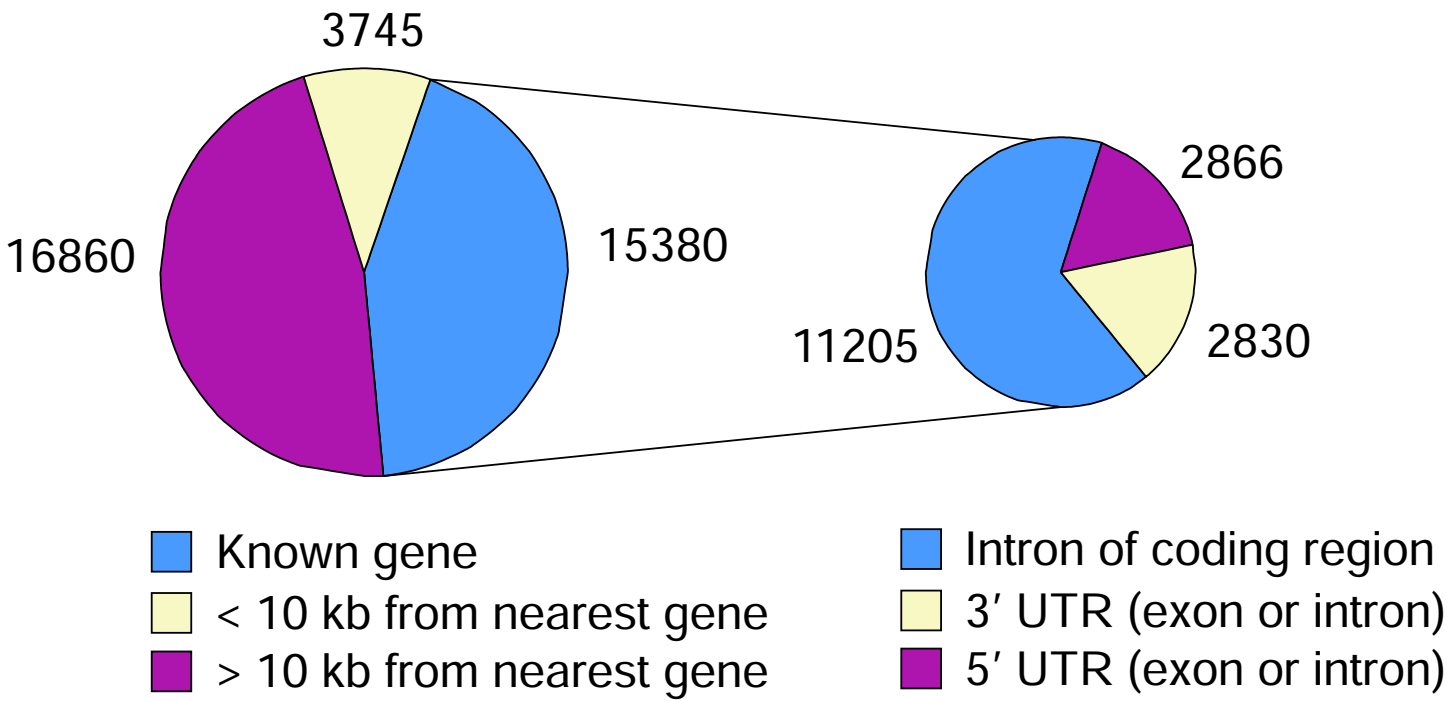
- The SCI is an efficient and convenient measure for secondary structure conservation.

Structured RNAs in the Human Genome



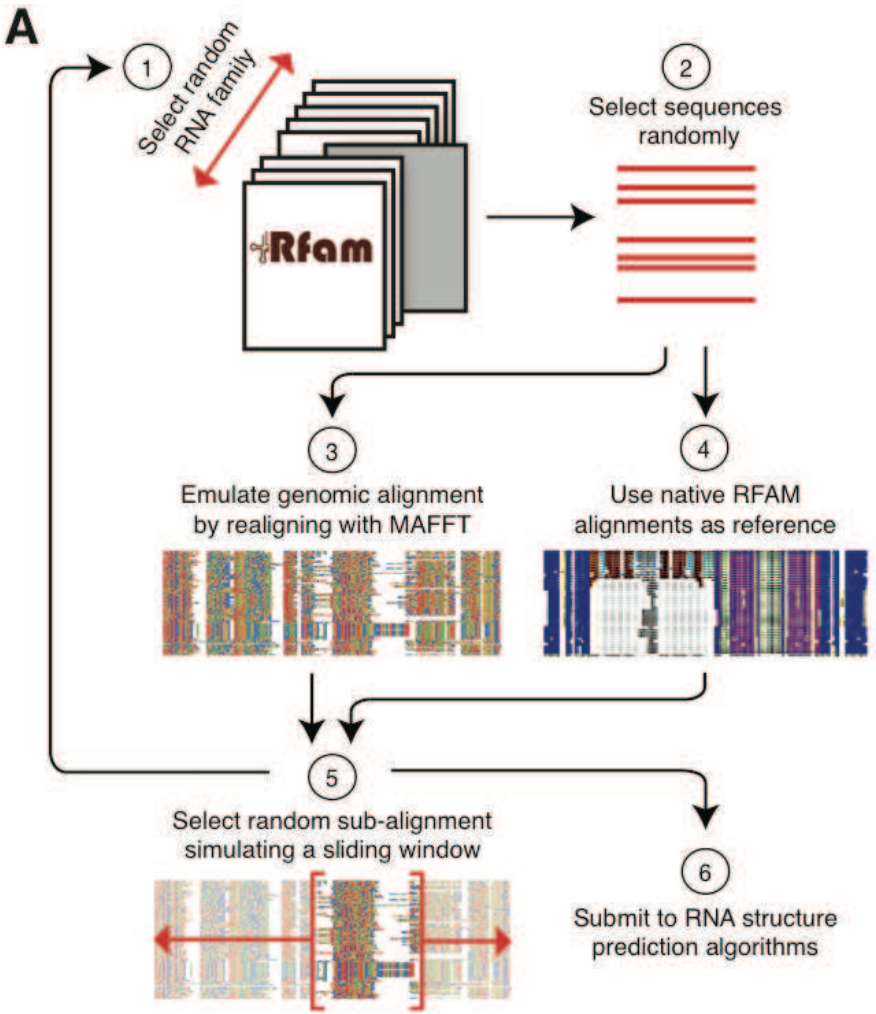
Structured RNAs in the Human Genome

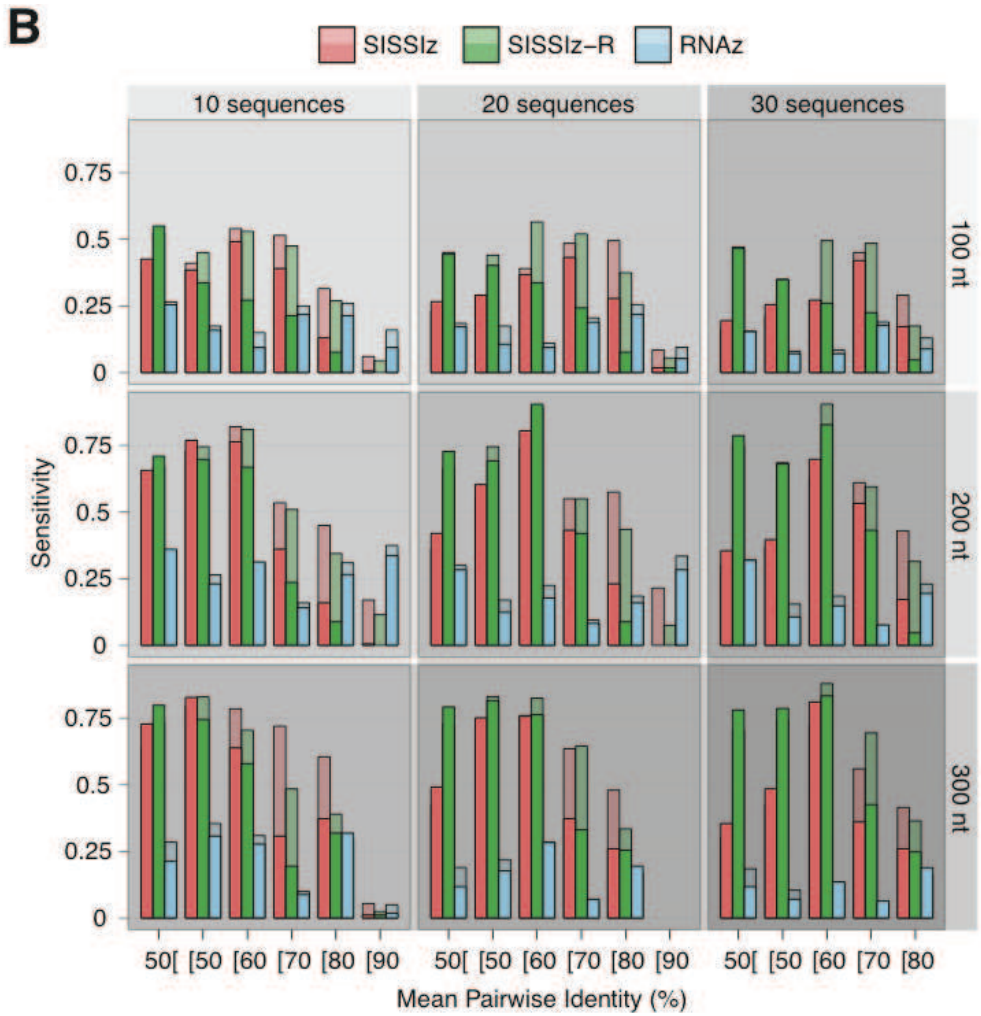
Mammalian genomes contain $\sim 10^5$ structured RNA motifs
Statistics of the highest-confidence fraction ($\sim$ 36000):



Nature Biotech. **23** 1383-1390 (2005)

A new screen

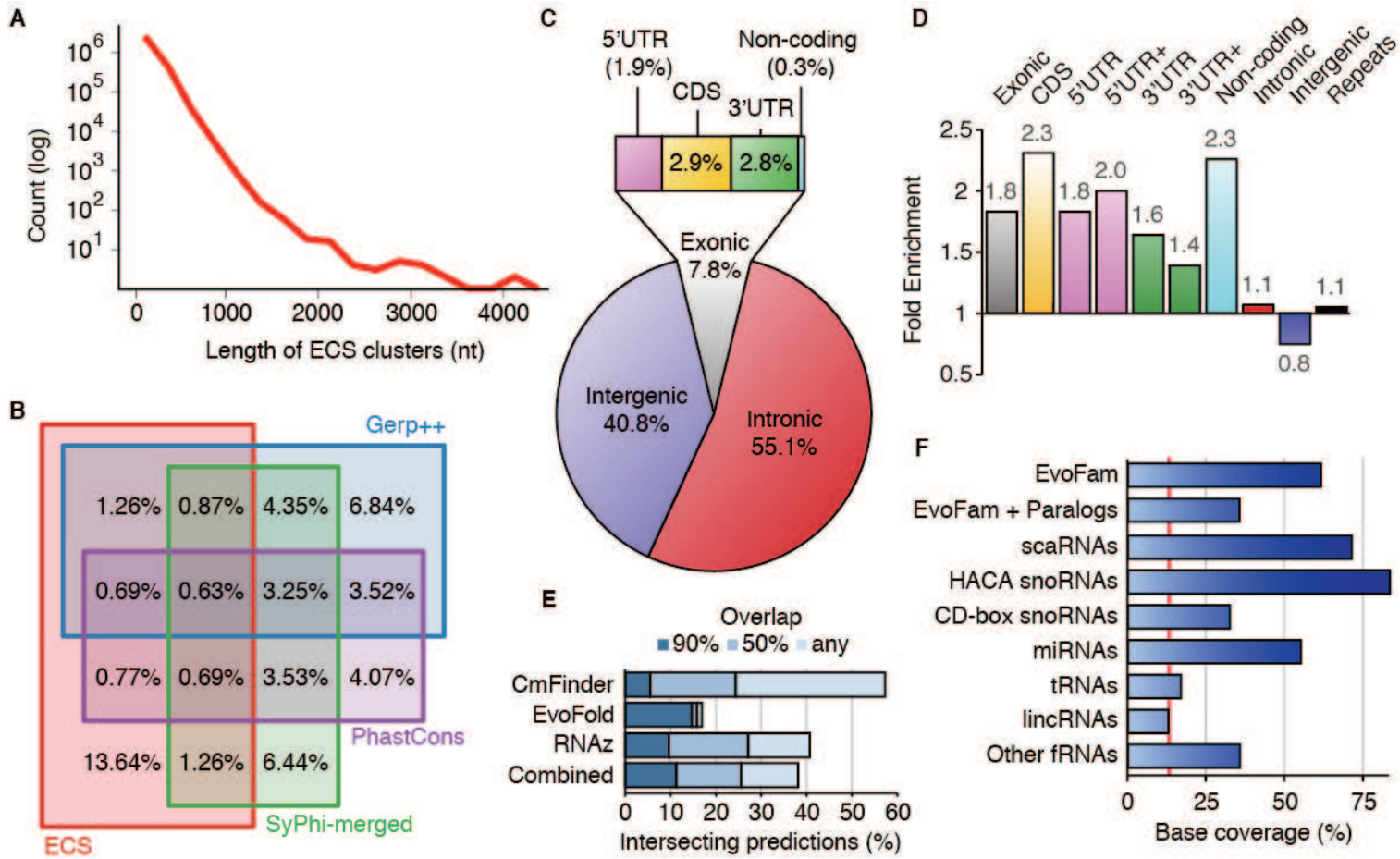




combination of `RNAz` and `sissiz`

with Martin Smith, Tanja Gesell, and John Mattick (under review 2012)

A new screen



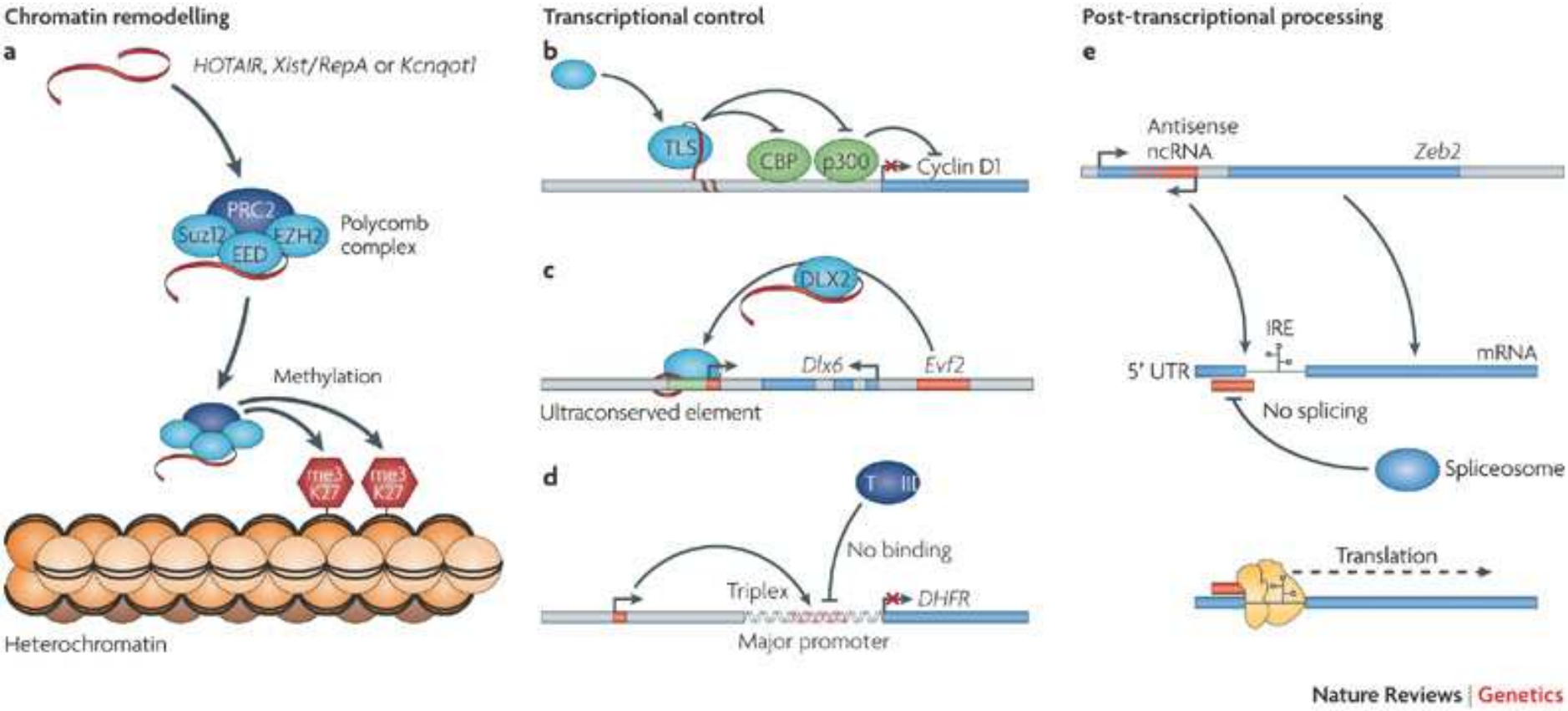
13.6 % of the genome is under selection for RNA secondary structure about 88% of these are not constrained at sequence level

with Martin Smith, Tanja Gesell, and John Mattick (under review 2012)

Long non-coding RNAs

1. mRNA-like: spliced and often polyadenylated
   - microRNA precursors (not all are spliced)
   - snoRNA precursors
   - piRNA precursors
   - “lincRNAs” associating with protein complexes that read, write, or erase chromatin marks
   - ceRNAs, i.e., microRNA sponges and possibly other decoys
   - enhancer-like ncRNAs
   - ...
2. other types of lncRNAs
   - totally and partially intronic transcripts (TINs, PINs)
   - independent UTRs (uaRNAs)
   - long unspliced RNAs such as MALAT-1 and MENβ
   - macroRNAs (hundreds of kilobases long transcribed regions)

Functions of long non-coding RNAs



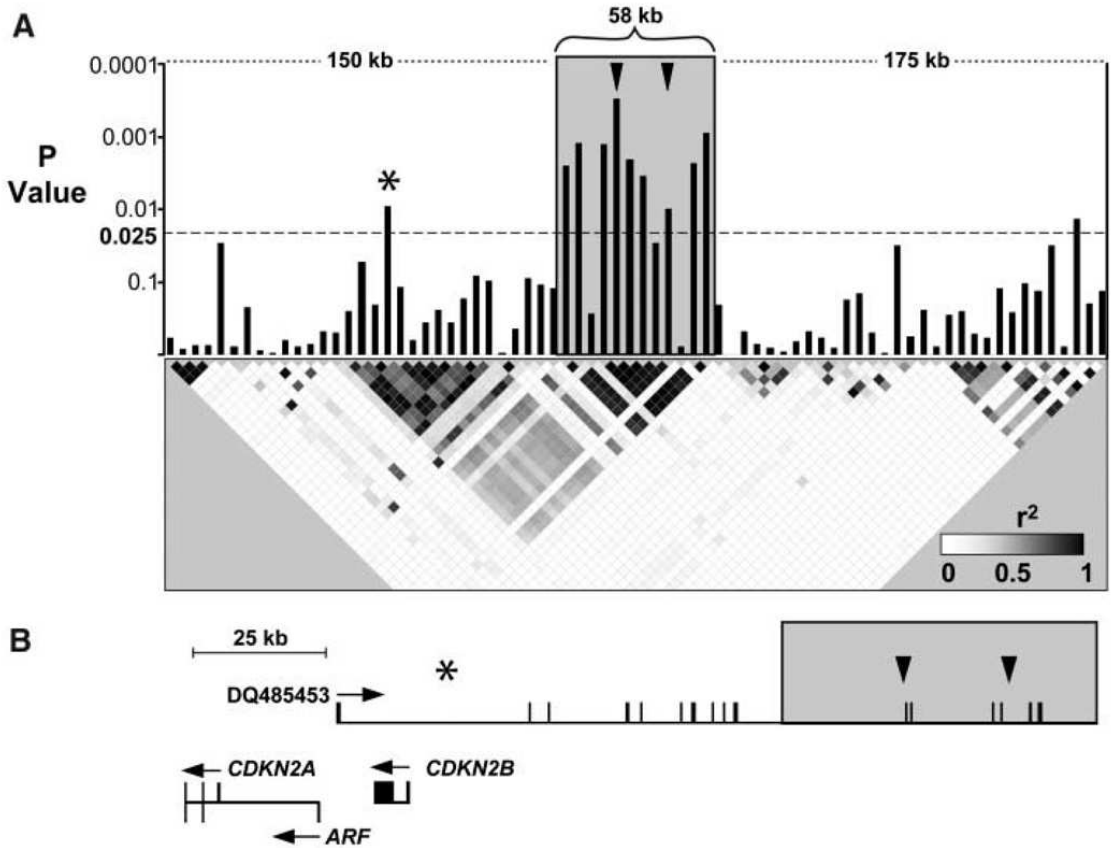
Mercer *et al* 2009

... and many more

ANRIL

Most QTLs for complex multi-genic diseases hit noncoding regions

Association of coronary heart disease (CHD) with a 58kb region on chr. 9p21

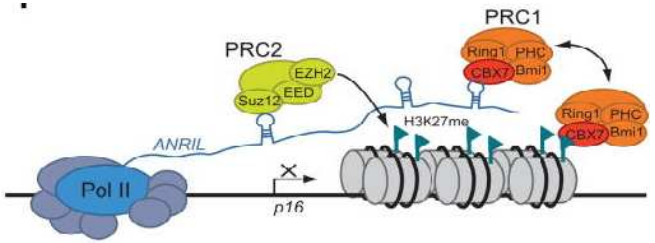


McPherson *et al.*, Science (2007)

ANRIL transcript(s) in many isoforms
associated with the atherosclerosis risk

Holdt *et al.* (2010)

and it appears in many other GWAS studies

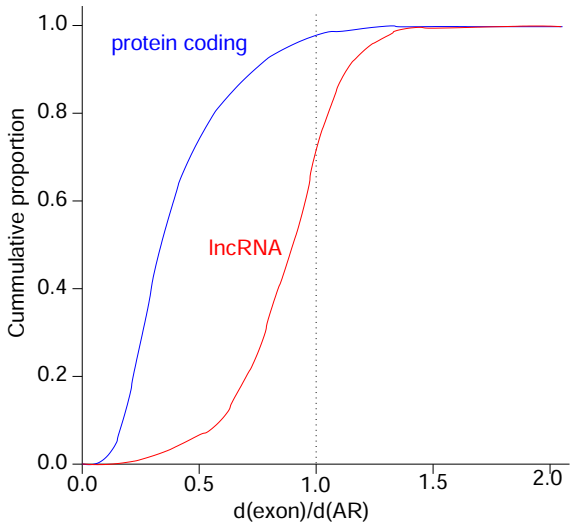


Yap *et al* (2010)

mRNA-like ncRNAs

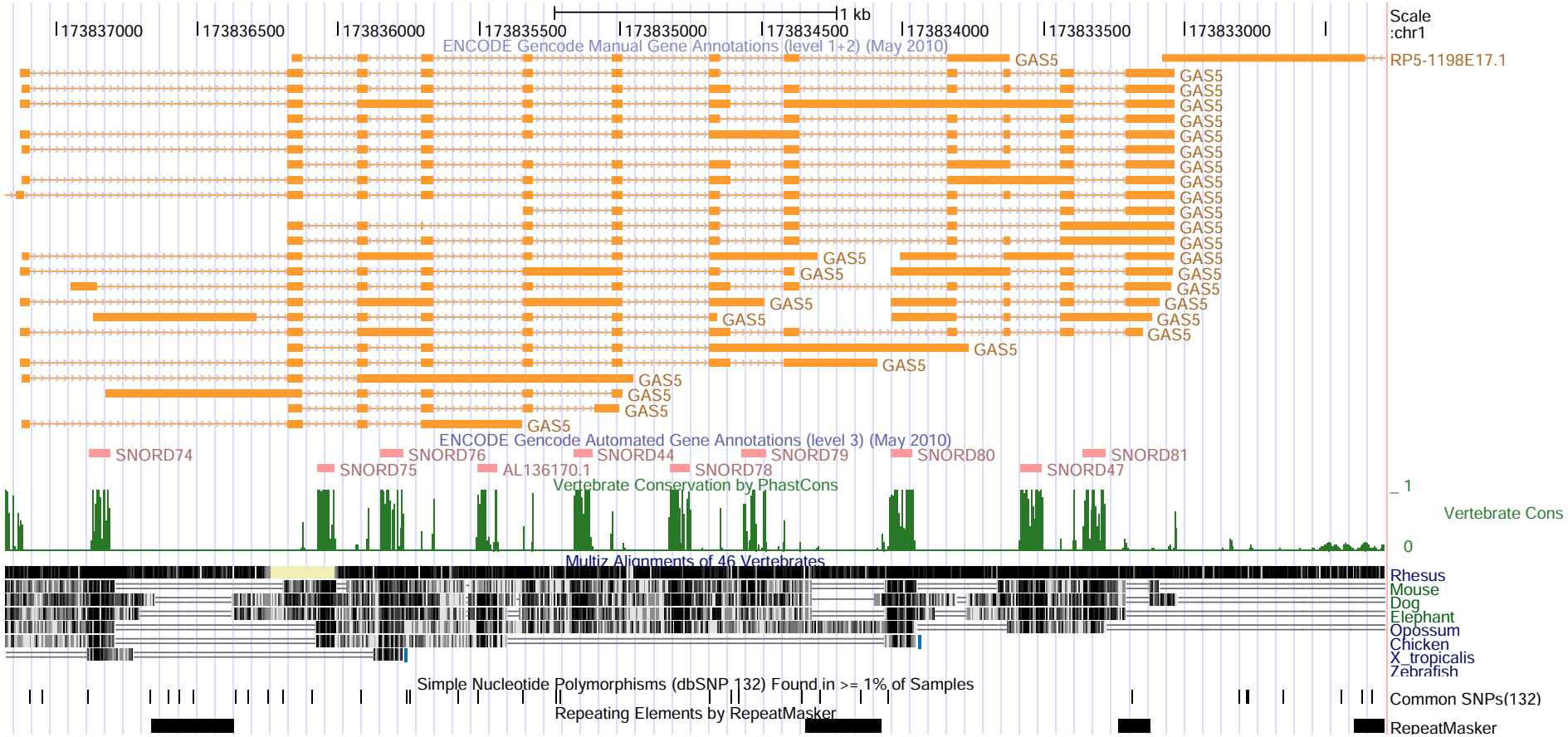
over the last few years ncRNAs that otherwise look quite similar to mRNAs have become a major research topic
(using, as usual, a variety of acronyms) mlncRNA, lincRNAs,

- How well conserved are lncRNAs?
  Two answers:
  1. "relatively low degree of sequence constraint"
     (Marques & Ponting 2009)
  2. but ... some very well-conserved examples
     (Chodroff *et al.* 2010, ...)



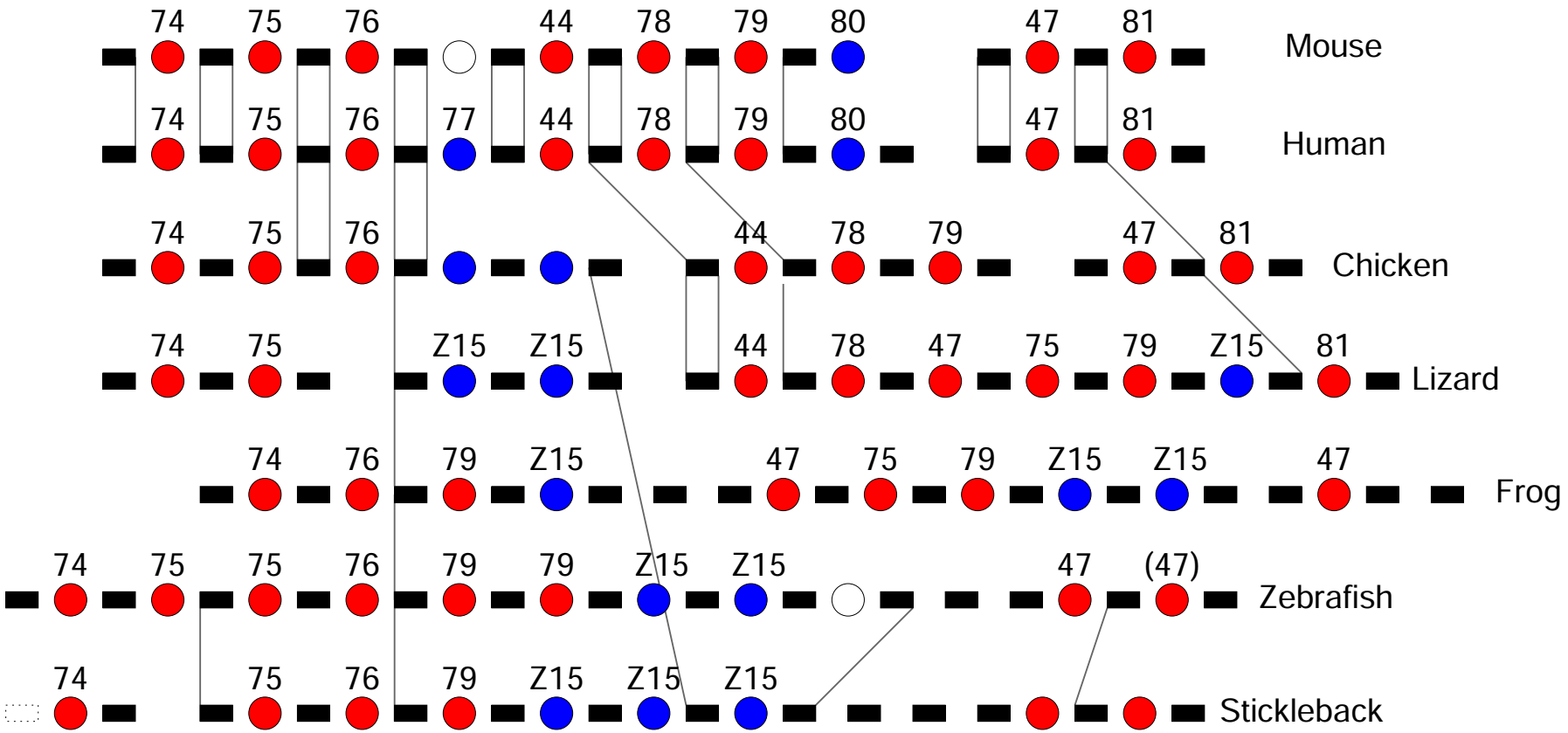
- One additional problem:
  sequence conservation does not necessarily imply conservation of the ncRNA!

Human GAS5 – a complex locus



- most famous snoRNA host gene with 10 different snoRNAs
- The exonic part (“mRNA”) sequesters and inhibits the glucocorticoid receptor
- conserved at least in gnathostomes

Evolution of GAS5

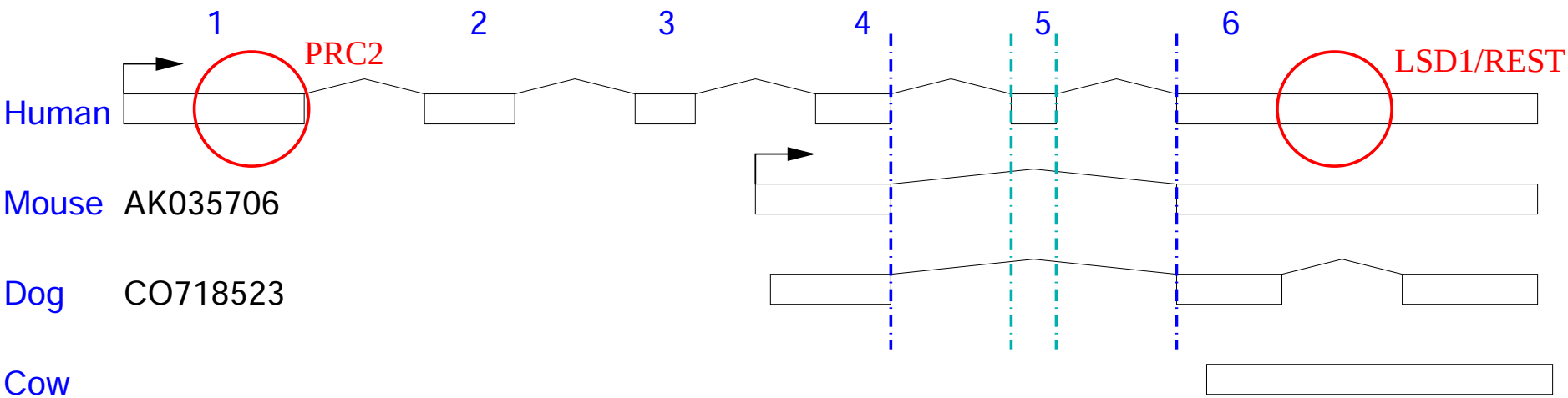


Two superimposed effects

- changes in the structure of the host gene itself
  gain & loss of splice sites
- snoRNAs can be behave like mobile elements

Evolution of mlncRNAs: HOTAIR

- transcribed from the HOXC cluster in antisense direction from the HoxC12-HoxC11 intergenic region
- directs PRC2 to the HOXD locus, silencing HoxD11-HoxD8. [Rinn et al 2007, Tsai et al 2010]
- however, the mouse homolog does not have this function [Schorderet & Duboule 2011]



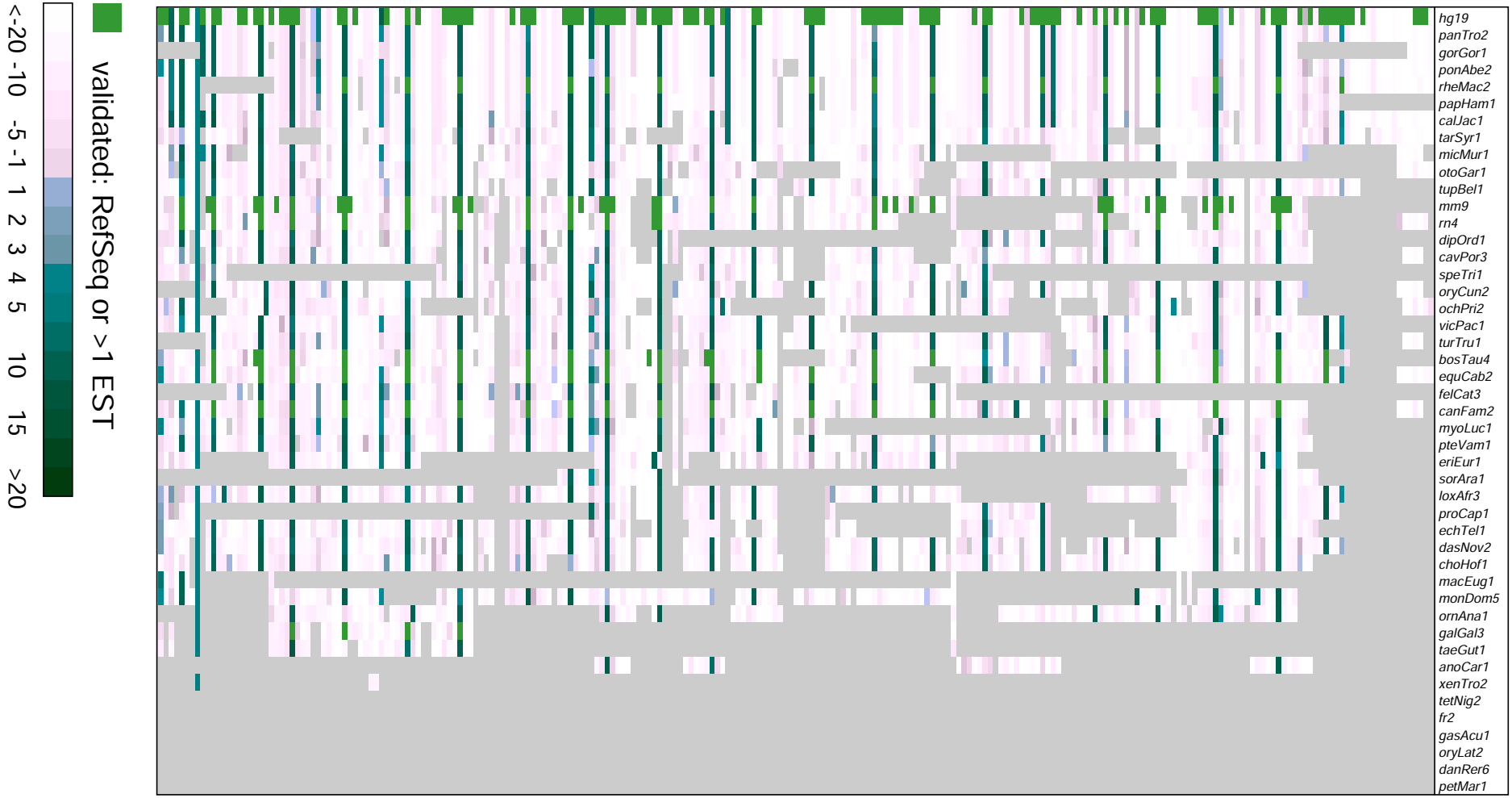
Schorderet P, Duboule D. (2011): Mouse HOTAIR has a different structure, presumably lacks PRC2 binding domain

Comparative Map of Splice Sites

Simple idea:

1. use a genome-wide multiple sequence alignment
   1. UCSC 46-way multiz alignment
   2. ENSEMBL 12-way EPO alignment
2. map all splice sites that are experimentally known to the alignment
   RefSeq plus all ESTs

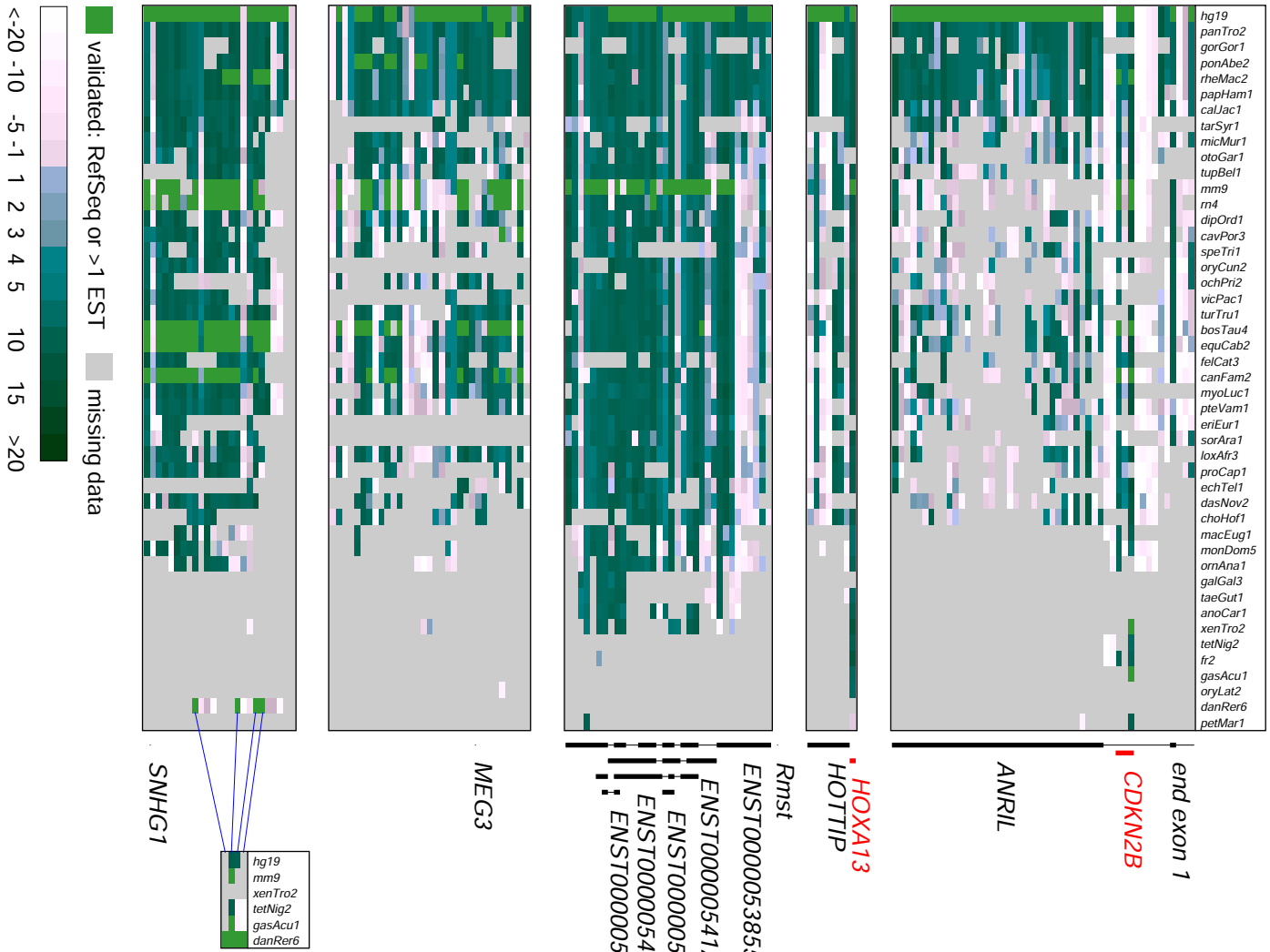
Splice Site Map for GAS5



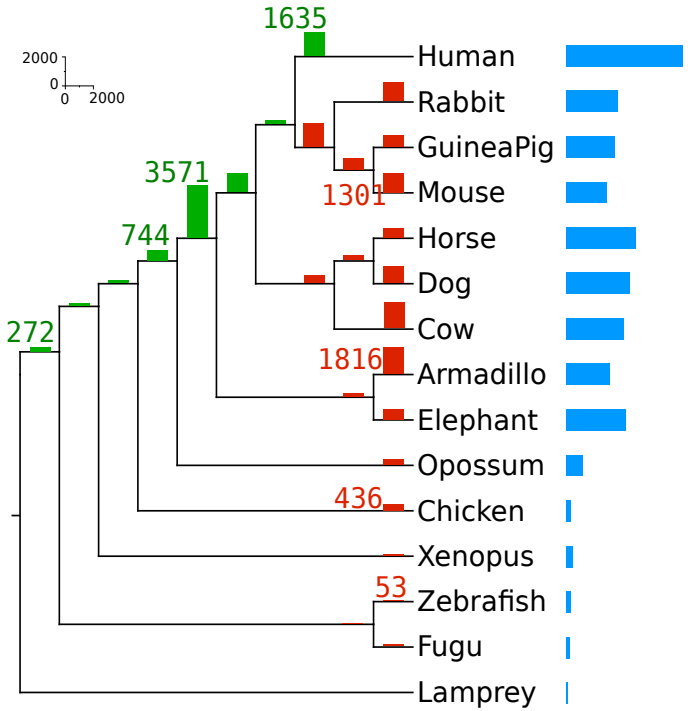
GAS5 is conserved throughout vertebrates. Very little aligned sequence outside amniotes.

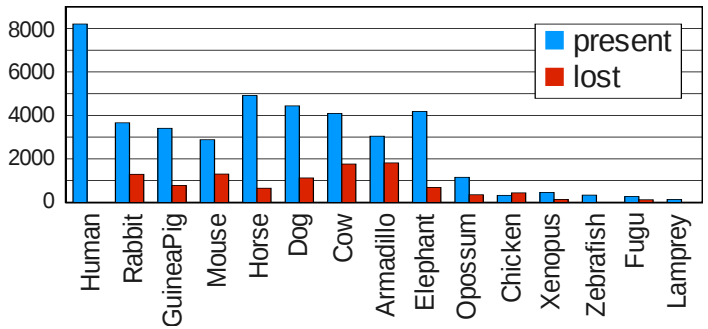
$\Longrightarrow$ sensitivity is limited by alignment quality

Some more examples



Conservation and Innovation of mlncRNAs





	aligned	cons.	known
	269 human microRNA host genes		
mouse	195	120	31
dog	237	191	13
5 eutheria	247	216	46
	118 snoRNA host genes		
mouse	95	73	57
dog	105	88	46
5 eutheria	111	96	63
	2,076 mouse lncRNAs [1]		
human	1,770	1,113	446
dog	1,628	944	185
4 eutheria	1,776	1,237	472
	1,508 zebrafish lncRNAs [2,3]		
teleosts	953	513	112
vertebrates	476	170	56

1. Guttmann *et al.* Nature 477: 295-300 (2011)
2. Pauli *et al.* Genome Res. 10.1101/gr.133009.111 (2011)
3. Ulitsky *et al.* Cell 147: 1537-1550(2011)

Many, many thanks . . .

- Leipzig: Jana Hertel, Hakim Tafer, Jan Engelhardt, Anne Nitsche, Sebastian Bartschat, Steffi Kehr, and many others
  Steve Hoffmann, Christian Otto, David Langenberger, Gero Doose, Stephan H. Bernhart
  Sonja J. Prohaska and her Comp. EvoDevo group
  FH RNomics group: Jörg Hackermüller, Kristin Reiche, ...
  FG ncRNAs: Friedemann Horn, Thomas Arendt, Kurt Engeland, Peter Ahnert, . . .
- Vienna: Ivo L. Hofacker, Christoph Flamm, Sven Findeiß, Andreas Gruber, and many others in Peter Schuster's Lab over the years
- Halle: Günter Reuter's Lab
- München: Daniel Teupser, Lesca Holdt
- Dresden: Michael Hiller
- Marburg: Manja Marz and her group
- Freiburg: Rolf Backofen, Dominic Rose
- Copenhagen: Jan Gorodkin, Stefan Seemann, Peter Menzel, and the RTH
- Barcelona: Roderic Guigó, Andrea Tanzer
- Strassbourg: Catherine Florentz, Joern Pütz, Frank Jühling
- MIT: Stefan Washietl, Sebastian Will
- Affymetrix: Tom Gingeras, Phil Kapranov, *et al.*
- PICB Shanghai: Axel Mosig and Phil Khaitovich and their students (PICB/SIBS)
- ASU Tempe: Julian L. Chen and his lab
- ENCODE: Ewan Birney and $10^{2.5}$ coauthors
- Funding by the DFG, DAAD, the EU 6th and 7th framework programme, the VW Foundation