Theoretical Physics Methods for Computational Biology.

Second lecture

M. Caselle
Dip di Fisica Teorica, Univ. di Torino

Berlin, 06/04/2006

Second lecture: Survey of most recent results in genome biology: the RNA revolution

- miRNA and post-transcriptional regulation.

- The Fantom project: non coding RNA's.

Question: how much do we *really* understand of our genome?

INTRODUCTION

In the Forests of RNA Dark Matter

For a long time, RNA has lived in the shadow of its more famous chemical cousin DNA and of the proteins that supposedly took over RNA's functions in the transition from the "RNA world" to the modern one. The shadow cast has been so deep that a whole universe (or so it seems) of RNA—predominantly of the noncoding variety—has remained hidden from view, until recently.

Nor is RNA quite so inert or structurally constrained as its cousin; its conformational versatility and catalytic abilities have been implicated at the very core of protein synthesis and possibly of RNA splicing. Noller (p. 1508) discusses how the basic building block of RNA—the double helix—has been fashioned into the intricate "protein-like" three-dimensional surfaces of the ribosome. A further parallel between RNA and protein is revealed in the structure of an RNA group I self-splicing intron, which uses an arrangement of two metal ions for phosphoryl transfer much like that seen in many protein enzymes (p. 1587). Another group I–like intron catalyzes the formation of a tiny RNA lariat, a reaction strikingly similar to one seen in group II introns and spliceosomal introns (pp. 1584 and 1530). This unusual lariat, at the very 5′ end of the resultant mRNA, is suggested to help protect the mRNA from degradation. The dynamics of the RNA messages passed between nucleus and cytoplasm provide a complex and sophisticated layer of regulation to gene expression, covered by Moore (p. 1514), who describes the teams of proteins that escort and regulate mRNA throughout the various stages of its life (and death). Death for many mRNAs occurs in cytoplasmic foci called P-bodies, which can also act as temporary storage depots for nontranslating mRNAs (see the *Science* Express Report by M. Brengues *et al.*).

Small noncoding microRNAs (miRNAs) have been found in such abundance that they have been christened the "dark matter" of the cell, a view reinforced by an analysis of the small RNAs found in *Arabidopsis* (pp. 1567 and 1525). The role of miRNAs and of their close cousins small interfering RNAs (siRNAs) in RNA silencing is discussed by Zamore

CONTENTS

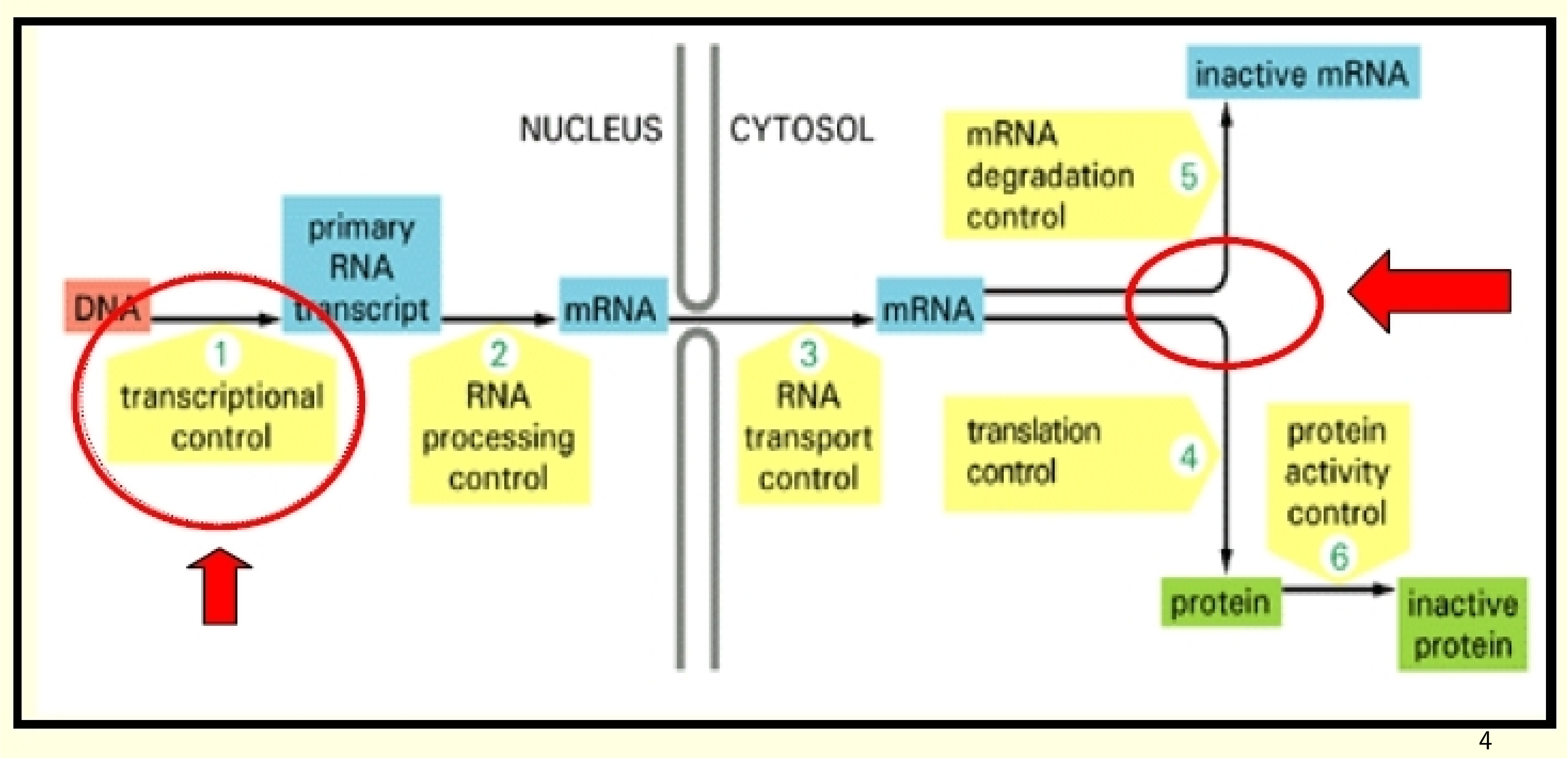
REVIEWS

1. miRNA.

Gene expression can be regulated at many of the steps in the pathway from DNA to RNA and protein.

MicroRNAs(miRNAs) are a family of 21 - 25 nucleotide small RNAs that negatively regulate gene expression at the post-transcriptional level.

Members of the miRNAfamily were initially discovered as small temporal RNAs(stRNAs) that regulate developmental transitions in Caenorhabditis Elegans(lin-4). (Chalfieet al. 1981; Lee et al. 1993)


inactive mRNA
NUCLEUS
CYTOSOL
mRNA degradation control
5
primary RNA transcript
DNA
mRNA
mRNA
1
transcriptional control
2
RNA processing control
3
RNA transport control
translation control
4
protein activity control
6
protein
inactive protein

1960s

July 1969
Britten and Davidson propose that RNA regulates eukaryotic gene expression

1970s

October 1972
Human cells are shown to contain nuclear double-stranded RNA

1990s

April 1990
Cosuppression discovered in plants

December 1993
The first microRNA, *lin-4*, is discovered

February 1994
RNA found to direct DNA methylation of plant viroids

May 1994
Calgene's "antisense" Flavr Savr tomato approved for sale by the FDA

May 1995
Both sense and antisense RNA found to inhibit gene expression in *C. elegans*

June 1997
An Argonaute protein, Piwi, is linked to stem cell maintenance

February 1998
Double-stranded RNA is discovered to be the trigger of RNA interference (RNAi)

October 1998
Plant viruses shown to encode RNA silencing suppressors

October 1999
Argonaute proteins found to be required for RNAi

October 1999–March 2000
Small interfering RNAs (siRNAs) discovered as guides for RNA silencing

2000s

October 2000
Double-stranded RNA shown to direct DNA methylation

January 2001
Dicer shown to make siRNAs

May 2001
RNAi discovered in human cells

July 2001
Dicer found to make microRNAs (miRNAs)

October 2001
miRNAs are established as a large class of gene regulators

July 2002
Plant miRNAs are discovered

July 2002
siRNAs are revealed as triggers of RNAi in mice

September 2002
Small RNAs guide the production of heterochromatin at centromeres

November 2002
miRNAs implicated in cancer

September 2003
It is clear that miRNA maturation begins in the nucleus

November 2003
Dicer shown to be required for mouse embryogenesis, and perhaps for stem cell production

March 2004
Human genome-wide RNAi libraries become available

April 2004
Animal viruses found to encode miRNAs

August 2004
First "investigational new drug" application filed for a therapeutic siRNA

September 2004
Argonaute is revealed as the RNAi endonuclease, "Slicer"

June 2005
miRNAs shown to act as oncogenes

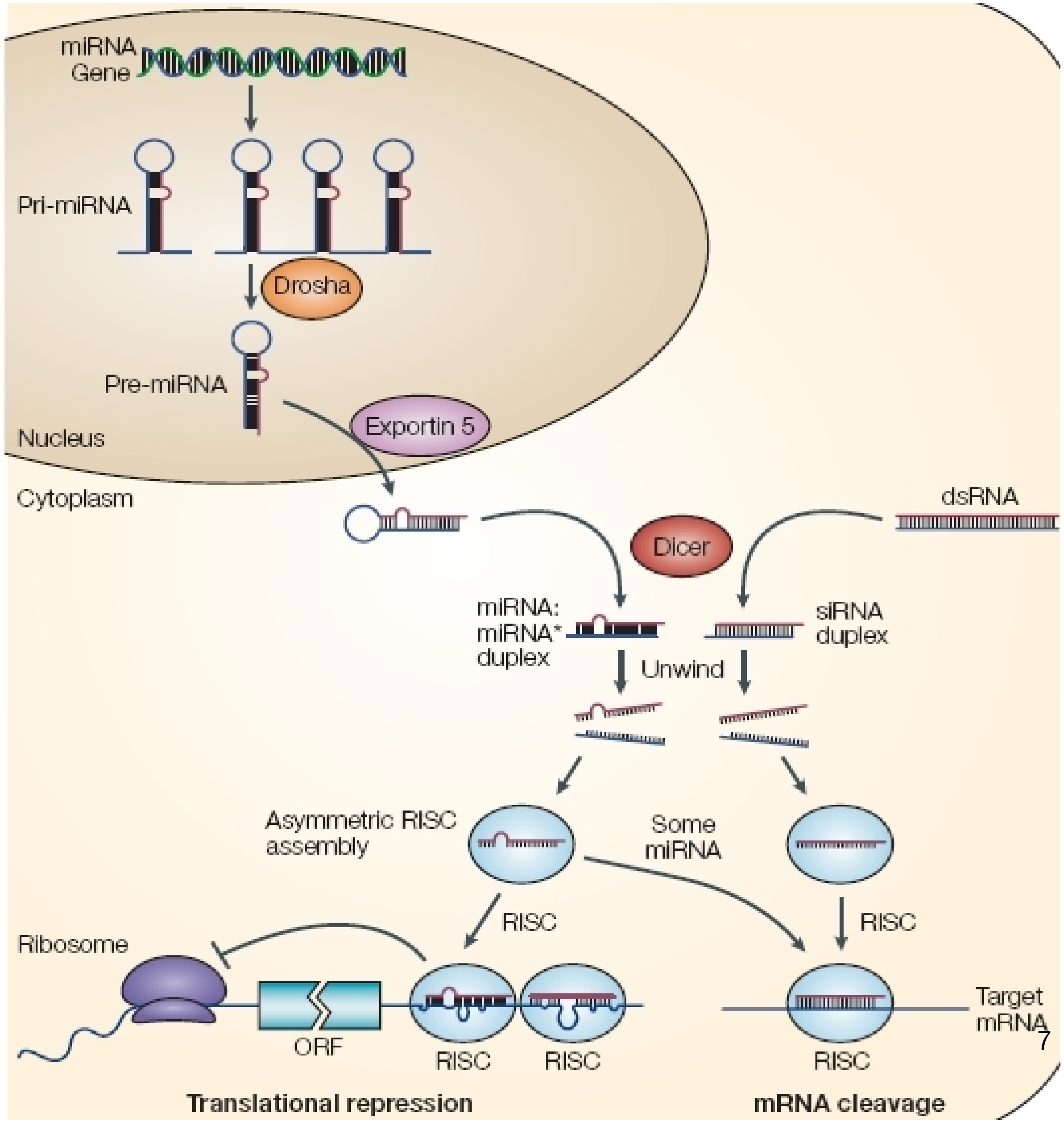
July 2005
Primate-specific miRNAs identified

miRNA biology

miRNAare derived from larger precursors that form imperfect stem-loop structures.

- the nascent miRNAtranscripts (pri-miRNA) are processed into 70 nucleotide precursors (pre-miRNA).

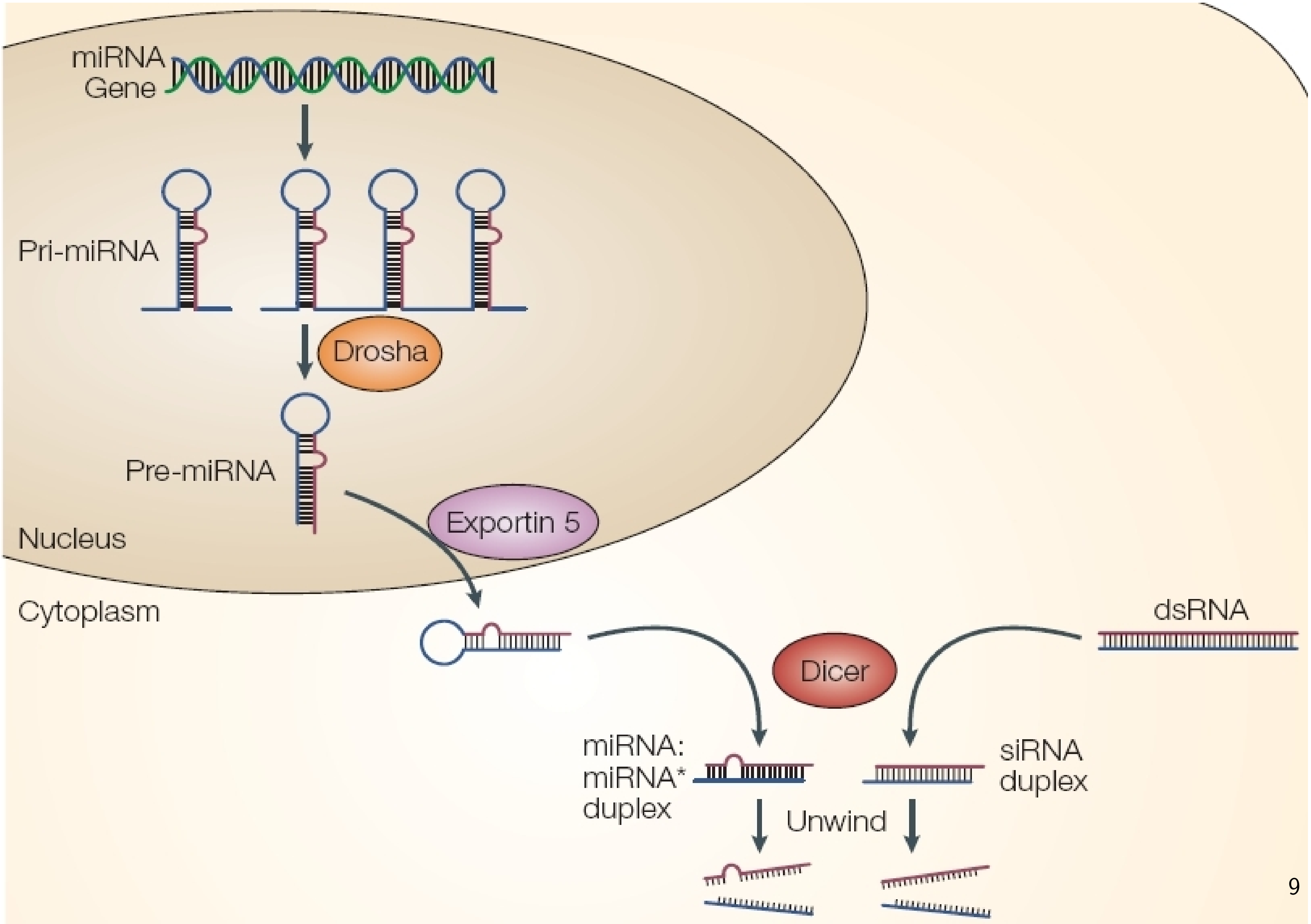
- The precursor is cleaved to generate 21 - 25 nucleotide mature miRNAs in cytoplasm.


miRNA
Gene
Pri-miRNA
Drosha
Pre-miRNA
Exportin 5
Nucleus
Cytoplasm
dsRNA
Dicer
miRNA:
miRNA*
duplex
siRNA
duplex
Unwind
Asymmetric RISC
assembly
Some
miRNA
RISC
RISC
Ribosome
ORF
RISC
RISC
RISC
Target
mRNA
Translational repression
mRNA cleavage

The miRNA gene regulation mechanism require the couple with a special protein complex called RNA-Induced Silencing Complex (RISC).

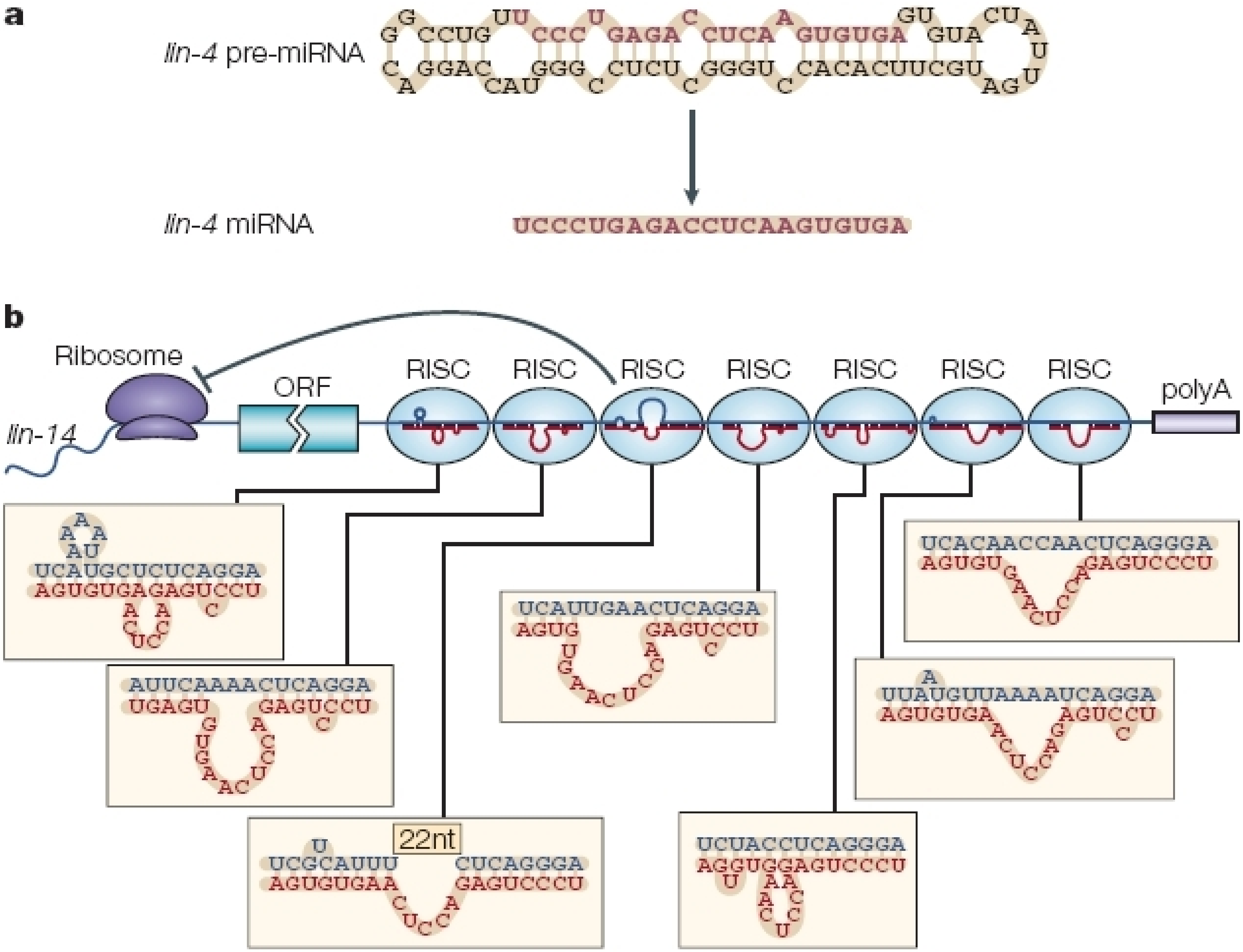
- Even though the precise mechanism of action of the miRNA / RISC complex is not very well understood, the current paradigm is that miRNAs are able to negatively affect the expression of a "target" gene via mRNA cleavage or translational repression, after antisense complementary base-pair matching to specific target sequences in the 3'-utr of the regulated genes.

- In plants, usually miRNA have perfect or near perfect complementarity to their mRNA target, whereas in animals the complementarity is restricted to the 5' regions of the miRNA, in particular requiring a "seed" of 6 nucleotides, usually from nucleotides 2 to 7.


miRNA Gene
Pri-miRNA
Drosha
Pre-miRNA
Nucleus
Exportin 5
Cytoplasm
dsRNA
Dicer
miRNA:
miRNA*
duplex
siRNA
duplex
Unwind

The functions in which miRNAs are involved are extremely wide and, in animals, they include: developmental timing, pattern formation and embryogenesis, differentiation and organogenesis, growth control and cell death, with putative involvement in human diseases in the case of *H. Sapiens*.

To date, hundreds of miRNAs and their relative targets are annotated in genomes of different metazoan organism, In addition, it is known that the miRNAs control is a one-to-many process, meaning that each miRNA is tough to control from one to hundreds of targets. Moreover, each specific miRNA binding site is also often overrepresented in a given 3'-utr sequence. And it is also a combinatorial mechanism, meaning that a certain mRNA can be under control of many different miRNAs


a
lin-4 pre-miRNA
G CCUG UU CCC U GAGA C CUCA A GUGUGA GU GUA CU A U
C GGAC CAU GGG C CUCU C GGGU C CACACUUCGU AG U
lin-4 miRNA
UCCCUGAGACCUCAAGUGUGA
b
Ribosome
lin-14
ORF
RISC
RISC
RISC
RISC
RISC
RISC
RISC
polyA
UCAUGCUCUCAGGA
AGUGUGAGAGUCCU
AUUCAAAACUCAGGA
UGAGU GAGUCCU
UCGCAUUU CUCAGGGA
AGUGUGAA GAGUCCCU
22nt
UCAUUGAACUCAGGA
AGUG GAGUCCU
UCUACCUCAGGGA
AGGUGGAGUCCCU
UUAUGUUAAAAUCAGGA
AGUGUGA AGUCCU
UCACAACCAACUCAGGGA
AGUGU GAGUCCCU

MiRNAs also show interesting evolutionary properties between different species. Up to one third of the miRNAs discovered in *C. elegans* have and orthologous in human. On the other hand, specie-specific miRNAs exist and, in particular, it is established that primates have an own class of miRNA genes.

However the rules followed by miRNAs in higher eukaryotes are slightly different from those of C.elegans or plants:

Rules in animals:

- Starting from 3-UTRs.

- Perfect complementarity or G-U pairing between the target 3-UTR and the first nucleotides 1-7 or 2-8 of miRNA.

- Favourable structural and thermodynamic heteroduplex formation between miRNAand its putative targets.

- Evolutionary conservation of miRNAtarget sites.

miRNA targets

So far, we have a catalogue of about 300 human miRNAgenes.

The functional characterization of miRNAs heavily relies on the identification of miRNA target genes.

num. human genes	~ 25000
num. human miRNA genes	~ 300
num. human genes regulated by miRNA	$\gg 5000$

miRNA References

- He and Hannon “MicroRNA: small RNAs with a big role in gene regulation.”
  Nat Rev. Genet. 2004 Jul;5(7):522-31.
- Bartel, D “MicroRNAs: Genomics, Biogenesis, Mechanism, and Function.”
  Cell. 2004 Jan 23;116(2):281-97.
- John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS “Human MicroRNA targets.”
  PLoS Biol. 2004 Nov;2(11):e363. Epub 2004 Oct 05.
- Griffiths-Jones S. “The microRNA registry.”
  Nucleic Acids Res. 2004 Jan 1;32 Database issue:D109-11.
- `http://www.microrna.org/`

2. The FANTOM project

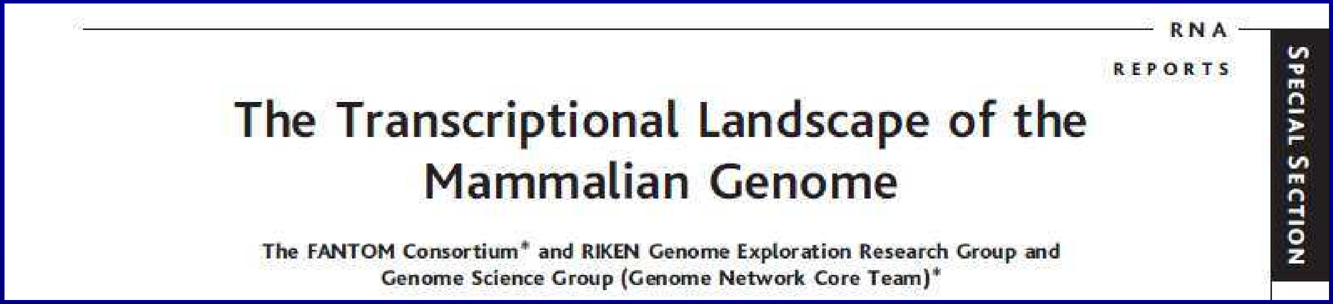
Functional annotation of the mouse

- Fantom 1 (2000), 21,000 cDNAs

- Fantom 2 (2002), 60,770 cDNAs

- Fantom 3 (2004-2005) functional libraries and 103,000 cDNAs:

Science 2005, September 2nd 「RNA Special Issue」



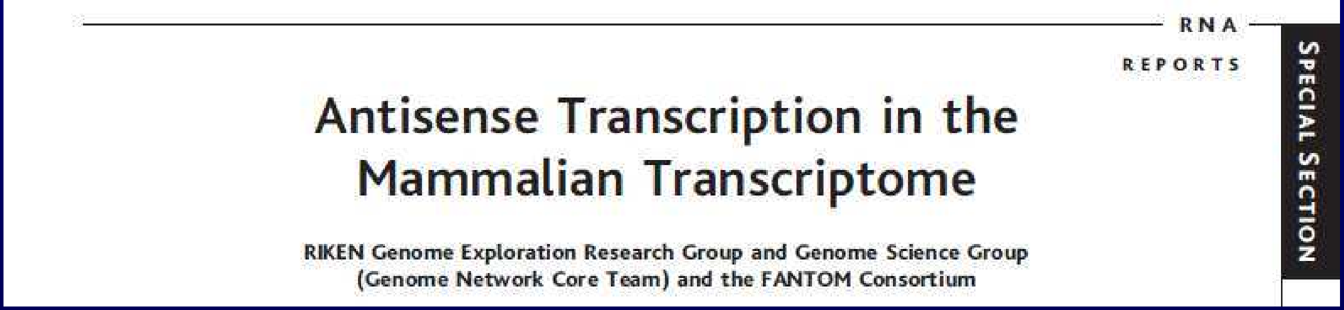
RNA

REPORTS

SPECIAL SECTION

The Transcriptional Landscape of the Mammalian Genome

The FANTOM Consortium* and RIKEN Genome Exploration Research Group and
Genome Science Group (Genome Network Core Team)*



RNA

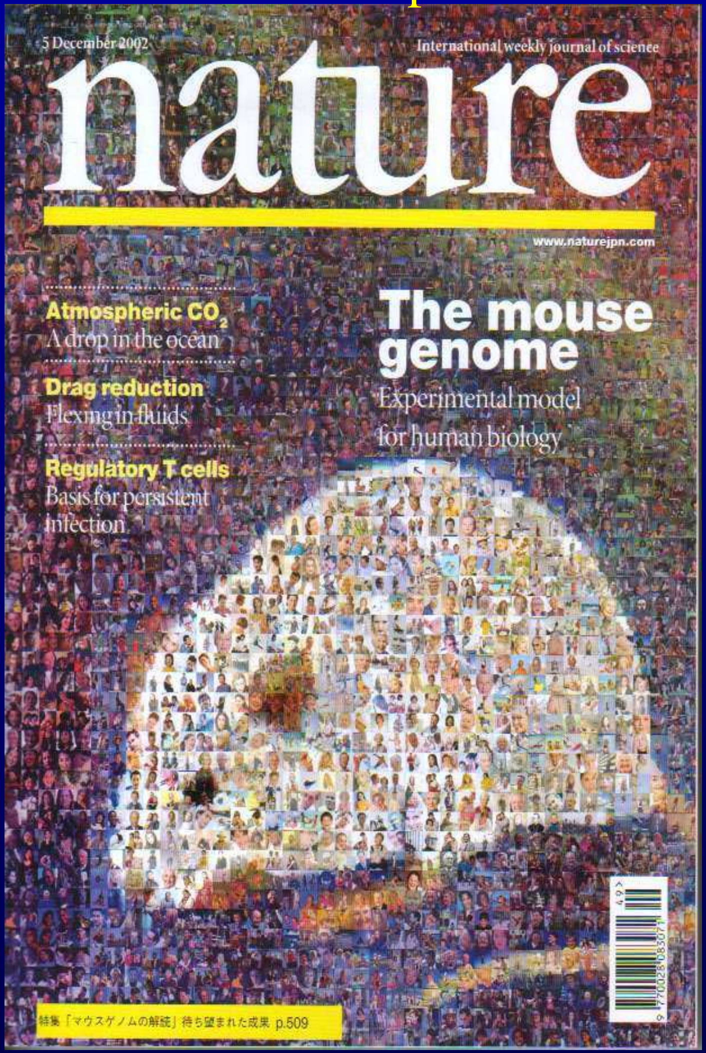
REPORTS

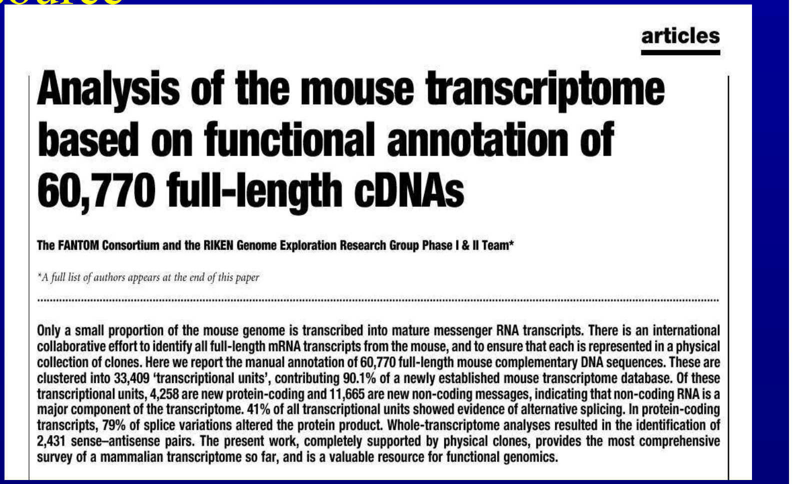
SPECIAL SECTION

Antisense Transcription in the Mammalian Transcriptome

RIKEN Genome Exploration Research Group and Genome Science Group
(Genome Network Core Team) and the FANTOM Consortium

We needed to complement this resource


5 December 2002
International weekly journal of science
nature
www.naturejpn.com
Atmospheric CO_2
A drop in the ocean
Drag reduction
Flexing in fluids
Regulatory T cells
Basis for persistent infection
The mouse genome
Experimental model for human biology
特集「マウスゲノムの解読」待ち望まれた成果 p.509



articles

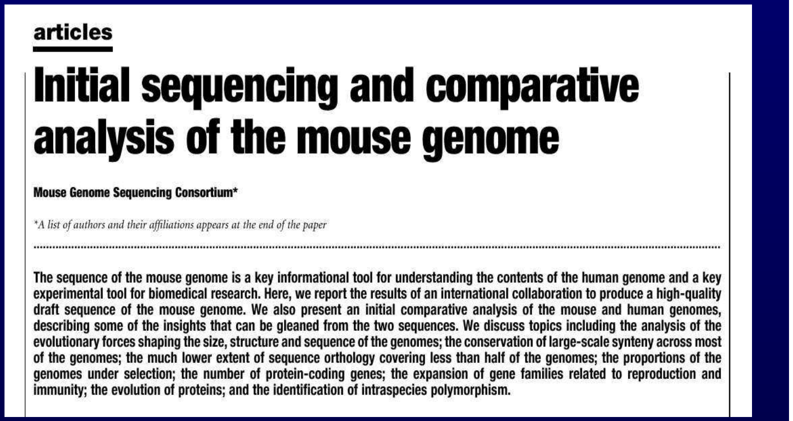
Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs

The FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team*

*A full list of authors appears at the end of this paper

Only a small proportion of the mouse genome is transcribed into mature messenger RNA transcripts. There is an international collaborative effort to identify all full-length mRNA transcripts from the mouse, and to ensure that each is represented in a physical collection of clones. Here we report the manual annotation of 60,770 full-length mouse complementary DNA sequences. These are clustered into 33,409 'transcriptional units', contributing 90.1% of a newly established mouse transcriptome database. Of these transcriptional units, 4,258 are new protein-coding and 11,665 are new non-coding messages, indicating that non-coding RNA is a major component of the transcriptome. 41% of all transcriptional units showed evidence of alternative splicing. In protein-coding transcripts, 79% of splice variations altered the protein product. Whole-transcriptome analyses resulted in the identification of 2,431 sense–antisense pairs. The present work, completely supported by physical clones, provides the most comprehensive survey of a mammalian transcriptome so far, and is a valuable resource for functional genomics.

↕ Riken · MGSC collabora



articles

Initial sequencing and comparative analysis of the mouse genome

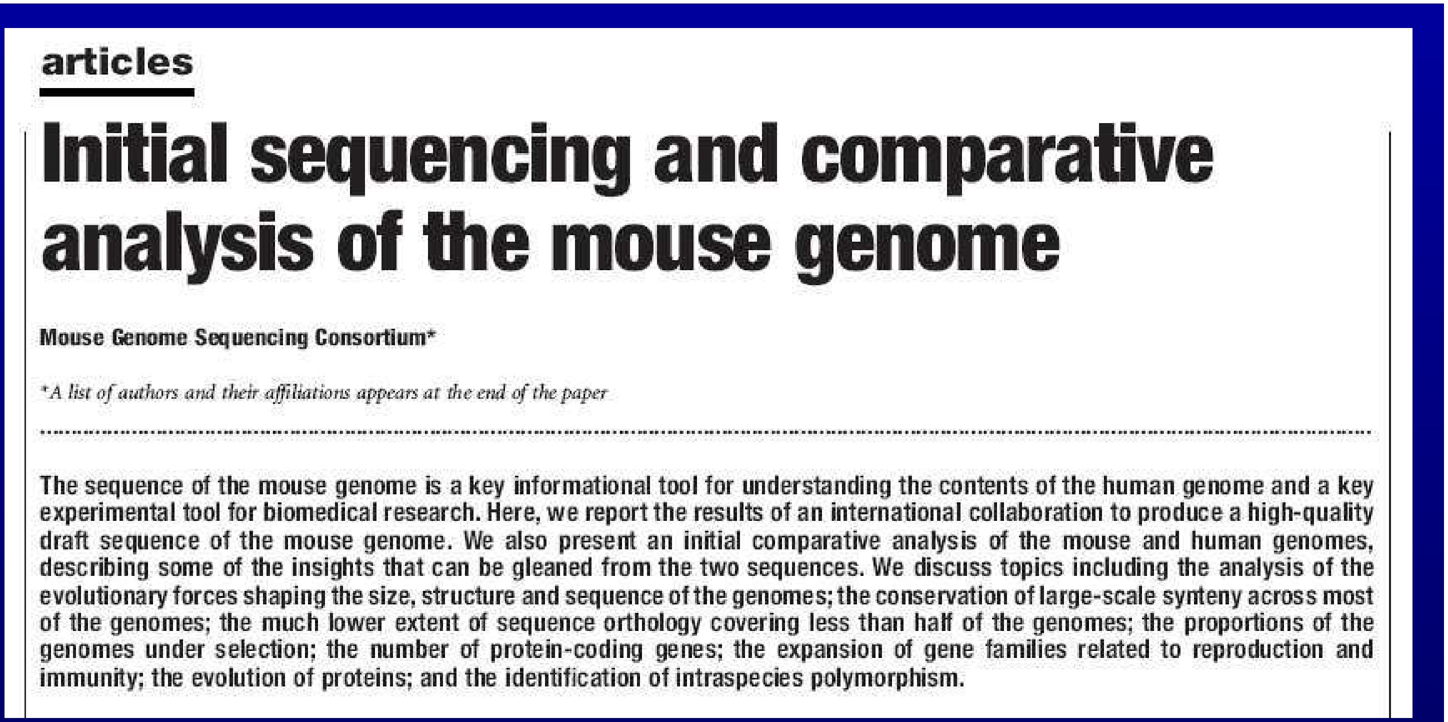
Mouse Genome Sequencing Consortium*

*A list of authors and their affiliations appears at the end of the paper

The sequence of the mouse genome is a key informational tool for understanding the contents of the human genome and a key experimental tool for biomedical research. Here, we report the results of an international collaboration to produce a high-quality draft sequence of the mouse genome. We also present an initial comparative analysis of the mouse and human genomes, describing some of the insights that can be gleaned from the two sequences. We discuss topics including the analysis of the evolutionary forces shaping the size, structure and sequence of the genomes; the conservation of large-scale synteny across most of the genomes; the much lower extent of sequence orthology covering less than half of the genomes; the proportions of the genomes under selection; the number of protein-coding genes; the expansion of gene families related to reproduction and immunity; the evolution of proteins; and the identification of intraspecies polymorphism.

Nature, volume 420, (December 5, 2002)

Nature, 420, 520-562, 2002



articles

Initial sequencing and comparative analysis of the mouse genome

Mouse Genome Sequencing Consortium*

*A list of authors and their affiliations appears at the end of the paper

The sequence of the mouse genome is a key informational tool for understanding the contents of the human genome and a key experimental tool for biomedical research. Here, we report the results of an international collaboration to produce a high-quality draft sequence of the mouse genome. We also present an initial comparative analysis of the mouse and human genomes, describing some of the insights that can be gleaned from the two sequences. We discuss topics including the analysis of the evolutionary forces shaping the size, structure and sequence of the genomes; the conservation of large-scale synteny across most of the genomes; the much lower extent of sequence orthology covering less than half of the genomes; the proportions of the genomes under selection; the number of protein-coding genes; the expansion of gene families related to reproduction and immunity; the evolution of proteins; and the identification of intraspecies polymorphism.

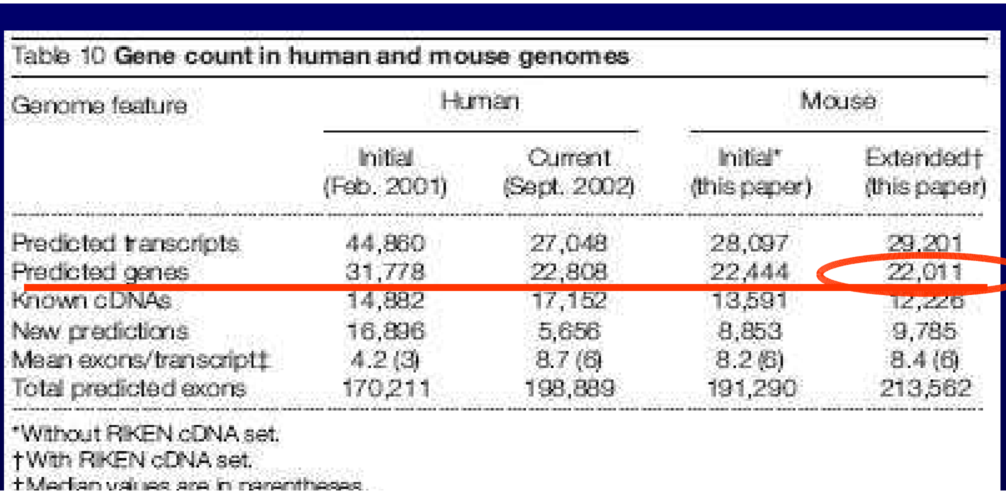


Table 10 **Gene count in human and mouse genomes**

Genome feature	Human		Mouse	
	Initial (Feb. 2001)	Current (Sept. 2002)	Initial* (this paper)	Extended† (this paper)
Predicted transcripts	44,860	27,048	28,097	29,201
Predicted genes	31,778	22,808	22,444	22,011
Known cDNAs	14,882	17,152	13,591	12,226
New predictions	16,896	5,656	8,853	9,785
Mean exons/transcript‡	4.2 (3)	8.7 (6)	8.2 (6)	8.4 (6)
Total predicted exons	170,211	198,889	191,290	213,562

*Without RIKEN cDNA set.
†With RIKEN cDNA set.
‡Median values are in parentheses.

- During the initial years of the human genome project, there were many debates on the number of coding genes. Now apparently settled, around 25,000

- Then many debates started on number of transcripts: latest FANTOM number: 181,047 !!

which is the origin of such a large discrepancy?

Main result: unexpected complexity of the Transcriptome

- Multiple starts/ends of mRNAs/ genes/transcriptional units (TU)
- Transcript density and overlap: Sense-antisense
- Alternative splicing
- Gene fusion
- Promoters identification: Alternative promoters

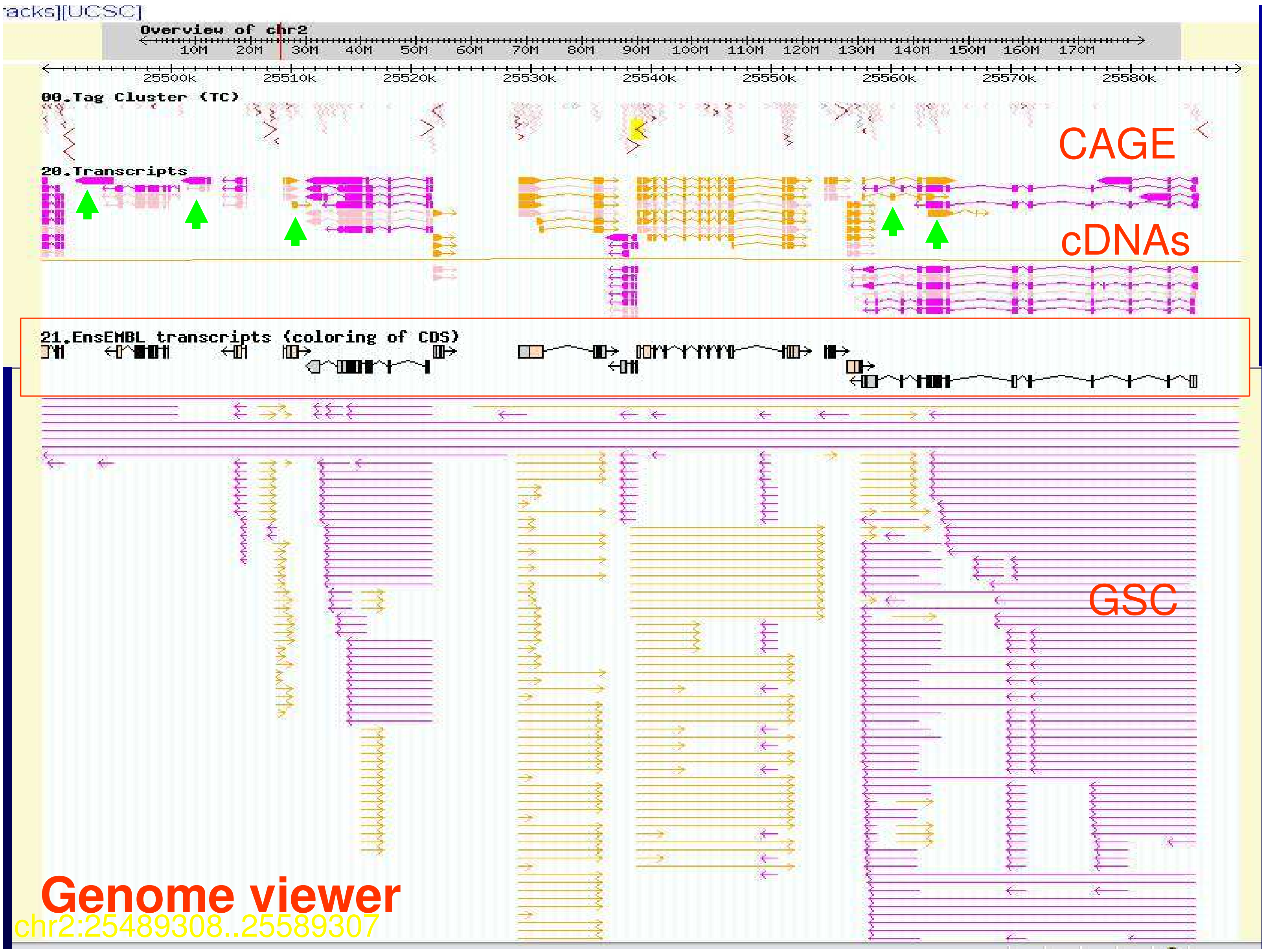
Transcriptional ”jungle” complexity

Extensive transcripts overlap:

Need for a new organization of the Transcriptome.

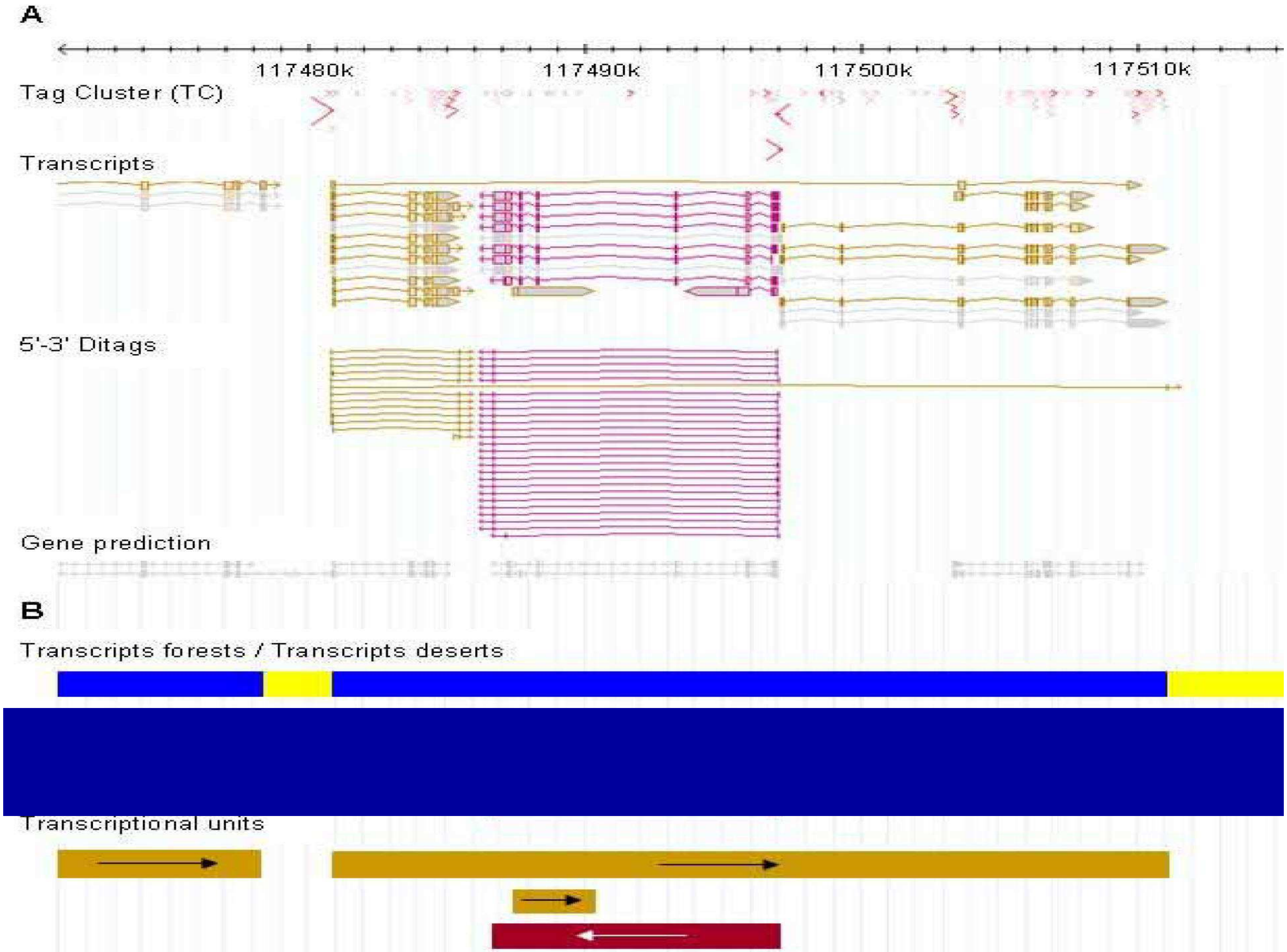
Fantom proposal: hierarchical scheme:

- Transcripts definition: Identification of true starting and termination sites pairs
- Transcriptional units (TU): Share a part of transcribed sequence in the same orientation
- Transcriptional forests


acks][UCSC]
Overview of chr2
10M 20M 30M 40M 50M 60M 70M 80M 90M 100M 110M 120M 130M 140M 150M 160M 170M
25500k 25510k 25520k 25530k 25540k 25550k 25560k 25570k 25580k
00.Tag Cluster (TC)
CAGE
20.Transcripts
cDNAs
21.EnsEMBL transcripts (coloring of CDS)
GSC

Genome viewer

chr2:25489308..25589307


A
117480k
117490k
117500k
117510k
Tag Cluster (TC)
Transcripts
5'-3' Ditags
Gene prediction
B
Transcripts forests / Transcripts deserts
Transcriptional units

Main results at the proteome level:

- No good overlap with ENSEMBL (predicted) genes

- More than 5000 novel proteins were identified.

  In some cases genes are hidden inside other genes!

- More than 78,000 different splicing patterns were detected

Novel genes are important!

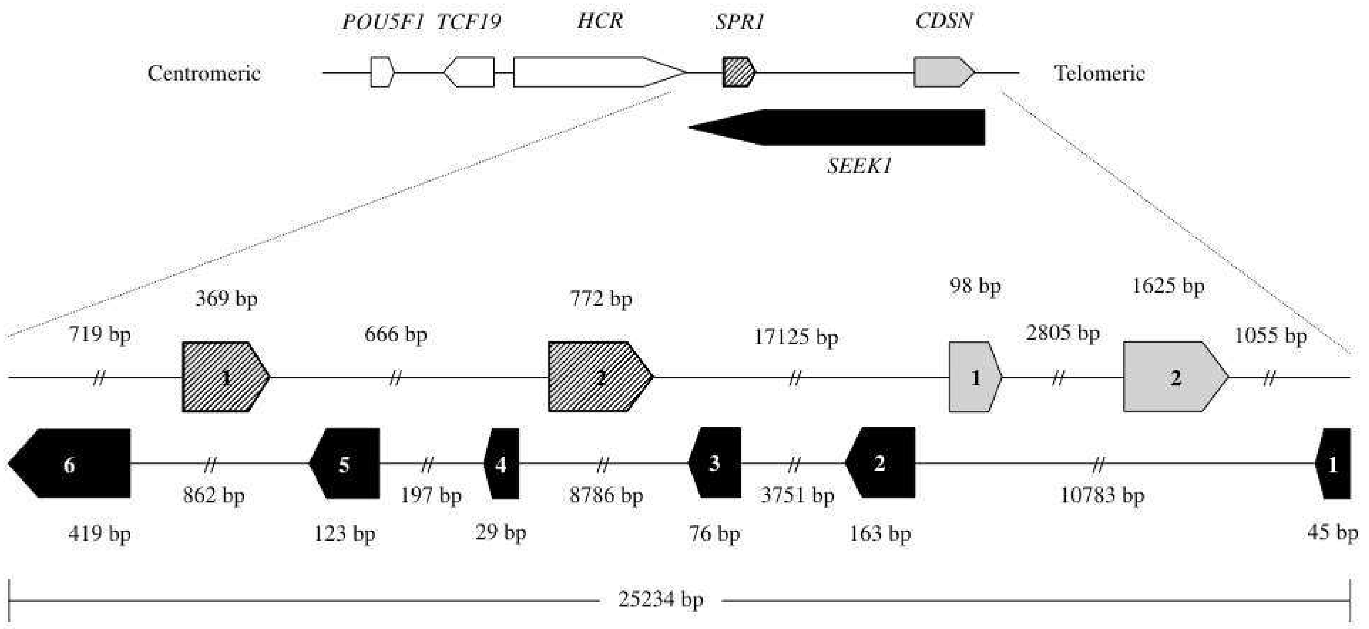


Figure 1. *SEEK1* gene overlaps with the *SPR1* and *CDSN* genes on 6p21.3. A gene map of a subsection of the PSORS1 region is shown, below which the 25234 bp region spanning the *SEEK1* gene is expanded in more detail. Gray boxes indicate the *CDSN* gene and exons, striped boxes indicate the *SPR1* gene and exons, which are both orientated in a centromeric to telomeric direction. Black boxes indicate the *SEEK1* gene and exons located on the opposite strand. Exon and intron sizes are given. Figure not to scale.

The challenge: noncoding RNA

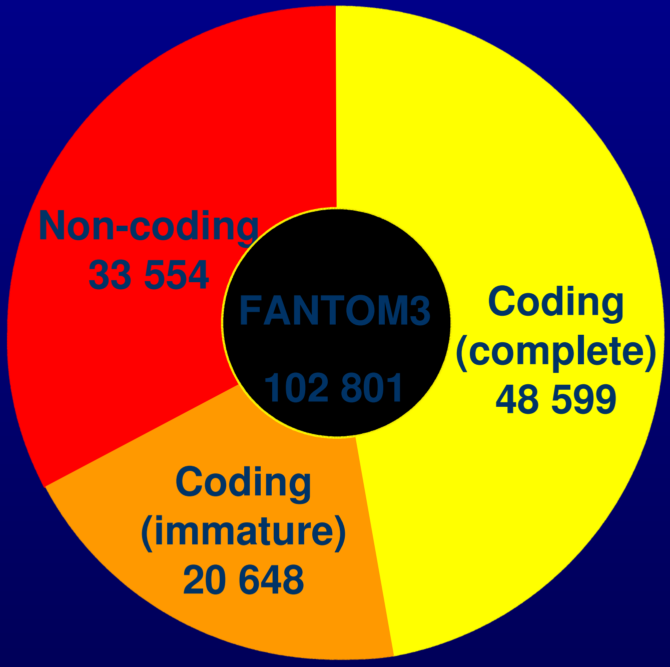
- Protein-coding clones: 64,672 most of them can be safely annotated

- Non-protein coding clones: 38,129 (even with a more conservative annotation pipeline more than 33,000 clones are certainly non-coding)

  Can we say anything about them??

  They are less conserved But their promoters are conserved!

Annotation result
Including noncoding cDNA pipeline



Coding (immature) includes “Truncated”, “intron retention”,
“UTR”, and “internally primed”.

The mistery: Sense-Antisense pairs

Huge amount of sense-antisense pairs detected.

Except RNAi, other potential functions are:

- Transcription interference
- CpG island overlap
- Triplex
- ??? ???

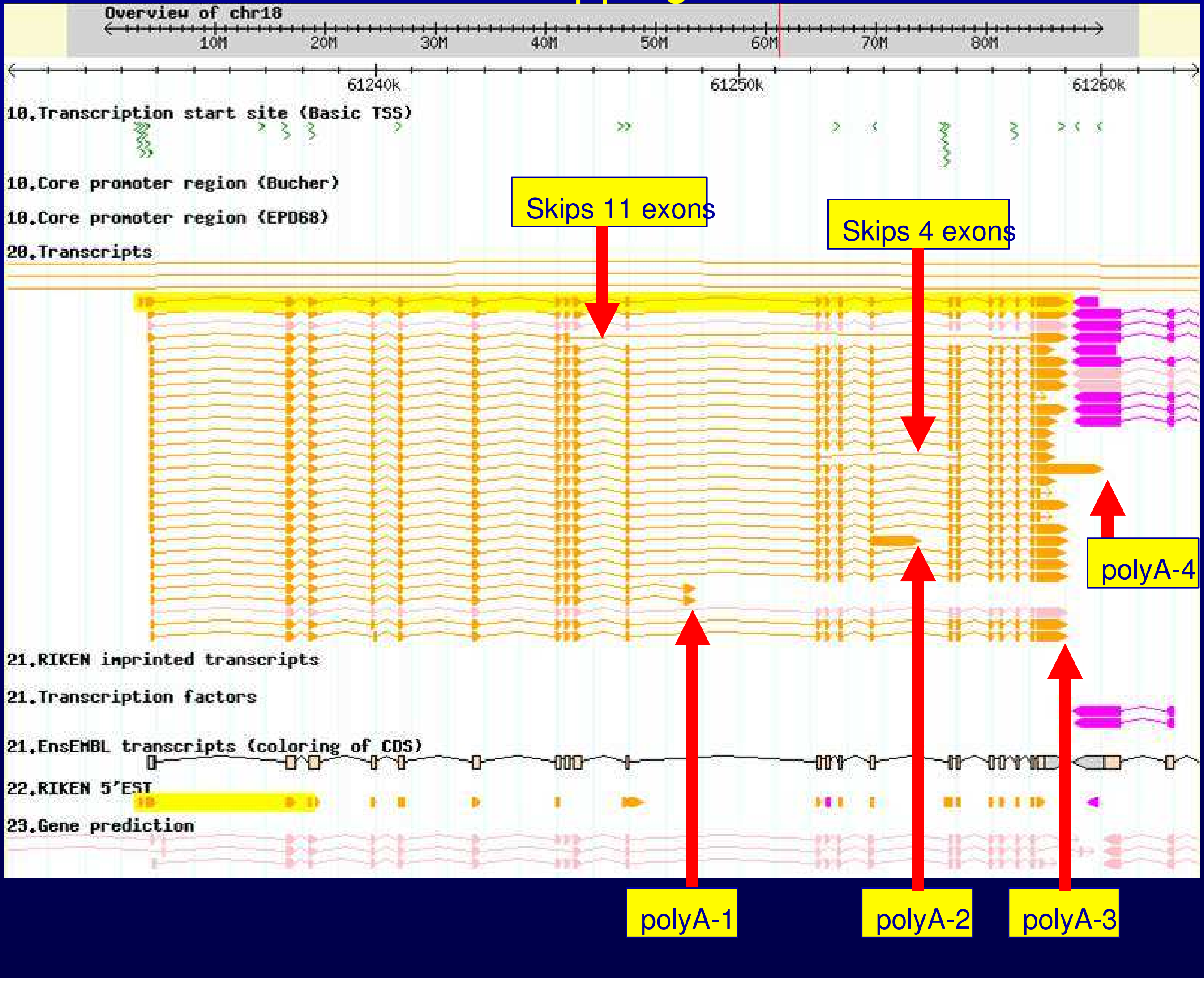
Alternative splicing and Gene fusion

These were both known events in transcription. However Fantom has shown that they are much more frequent than expected:

- nearly each gene is alternative spliced (more than 78,000 splicing patterns were observed)

- almost 1500 event of gene fusion were detected!!

Exon skipping in Csf1r



Overview of chr18

10. Transcription start site (Basic TSS)

10. Core promoter region (Bucher)

10. Core promoter region (EPD68)

20. Transcripts

21. RIKEN imprinted transcripts

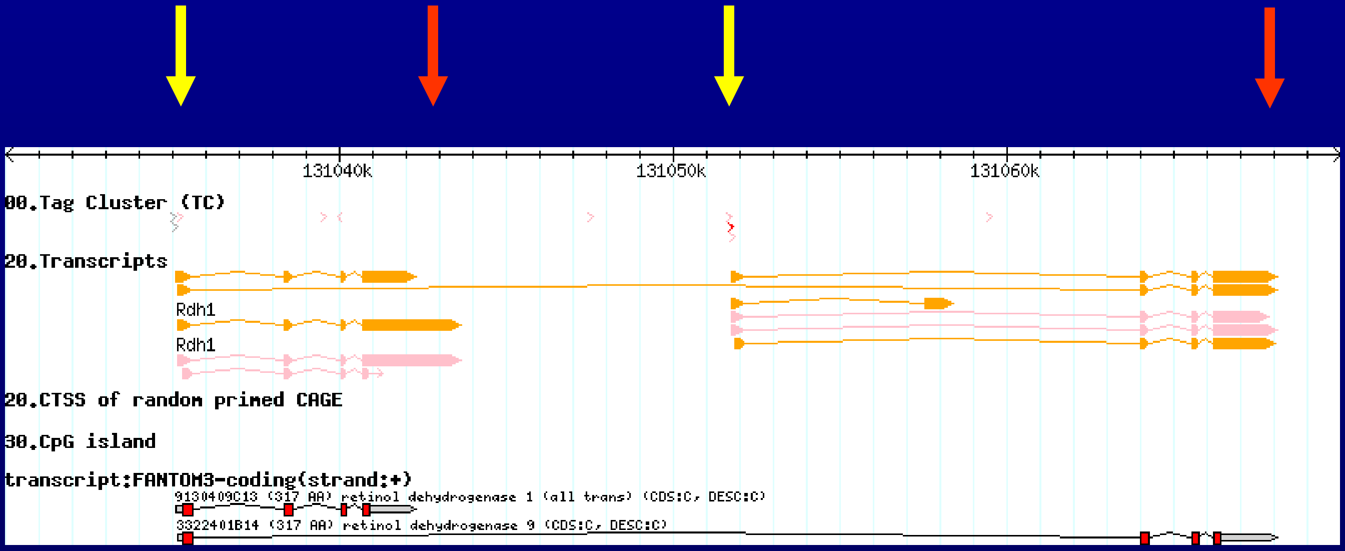
21. Transcription factors

21. EnsEMBL transcripts (coloring of CDS)

22. RIKEN 5'EST

23. Gene prediction

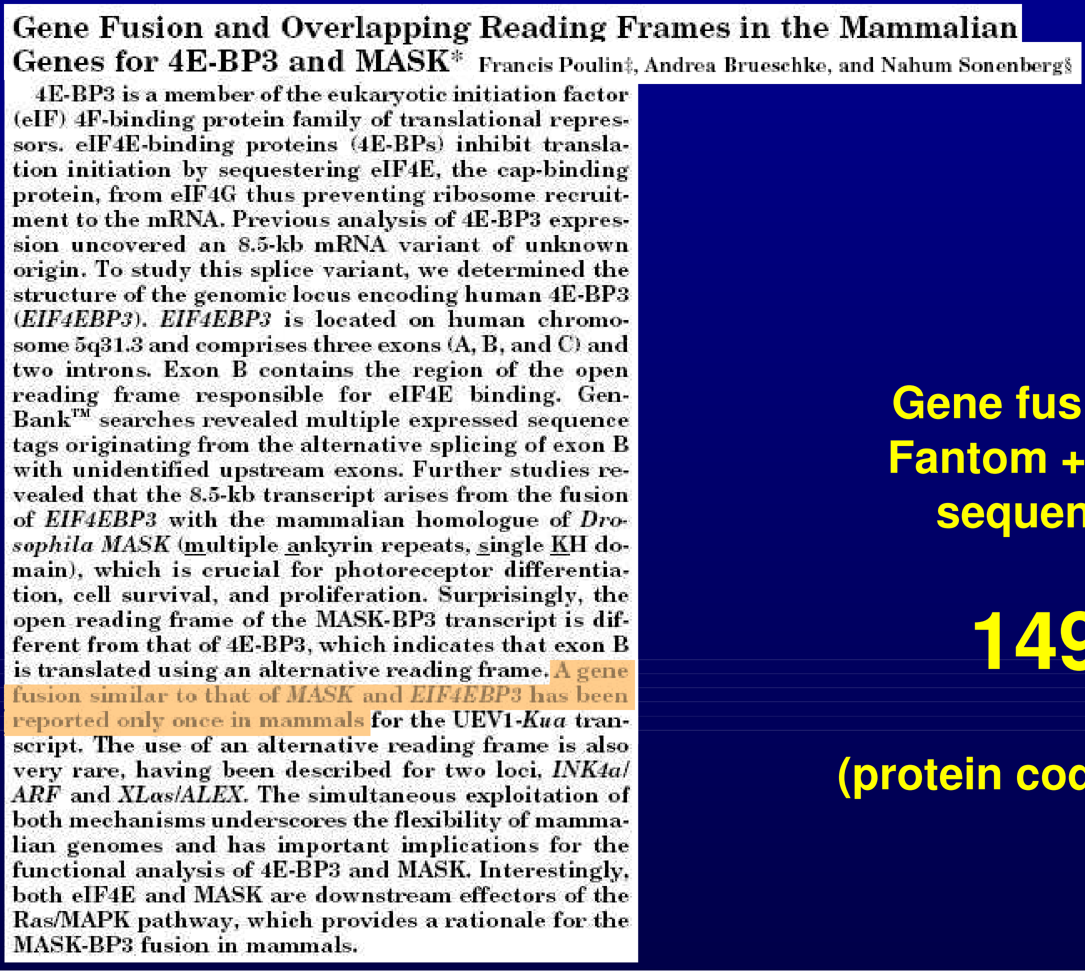
Gene Fusion


131040k
131050k
131060k
00.Tag Cluster (TC)
20.Transcripts
Rdh1
Rdh1
20.CTSS of random primed CAGE
30.CpG island
transcript:FANTOM3-coding(strand:+)
9130409C13 (317 AA) retinol dehydrogenase 1 (all trans) (CDS:C, DESC:C)
3322401B14 (317 AA) retinol dehydrogenase 9 (CDS:C, DESC:C)

Retinol dehydrogenase 1

Retinol dehydrogenase 9

Gene Fusions



Gene Fusion and Overlapping Reading Frames in the Mammalian Genes for 4E-BP3 and MASK*

Francis Poulin‡, Andrea Brueschke, and Nahum Sonenberg§

4E-BP3 is a member of the eukaryotic initiation factor (eIF) 4F-binding protein family of translational repressors. eIF4E-binding proteins (4E-BPs) inhibit translation initiation by sequestering eIF4E, the cap-binding protein, from eIF4G thus preventing ribosome recruitment to the mRNA. Previous analysis of 4E-BP3 expression uncovered an 8.5-kb mRNA variant of unknown origin. To study this splice variant, we determined the structure of the genomic locus encoding human 4E-BP3 (*EIF4EBP3*). *EIF4EBP3* is located on human chromosome 5q31.3 and comprises three exons (A, B, and C) and two introns. Exon B contains the region of the open reading frame responsible for eIF4E binding. GenBank™ searches revealed multiple expressed sequence tags originating from the alternative splicing of exon B with unidentified upstream exons. Further studies revealed that the 8.5-kb transcript arises from the fusion of *EIF4EBP3* with the mammalian homologue of *Drosophila MASK* (multiple ankyrin repeats, single KH domain), which is crucial for photoreceptor differentiation, cell survival, and proliferation. Surprisingly, the open reading frame of the MASK-BP3 transcript is different from that of 4E-BP3, which indicates that exon B is translated using an alternative reading frame. A gene fusion similar to that of *MASK* and *EIF4EBP3* has been reported only once in mammals for the UEV1-*Kua* transcript. The use of an alternative reading frame is also very rare, having been described for two loci, *INK4a/ARF* and *XLas/ALEX*. The simultaneous exploitation of both mechanisms underscores the flexibility of mammalian genomes and has important implications for the functional analysis of 4E-BP3 and MASK. Interestingly, both eIF4E and MASK are downstream effectors of the Ras/MAPK pathway, which provides a rationale for the MASK-BP3 fusion in mammals.

Gene fusions in Fantom + public sequences:

1499

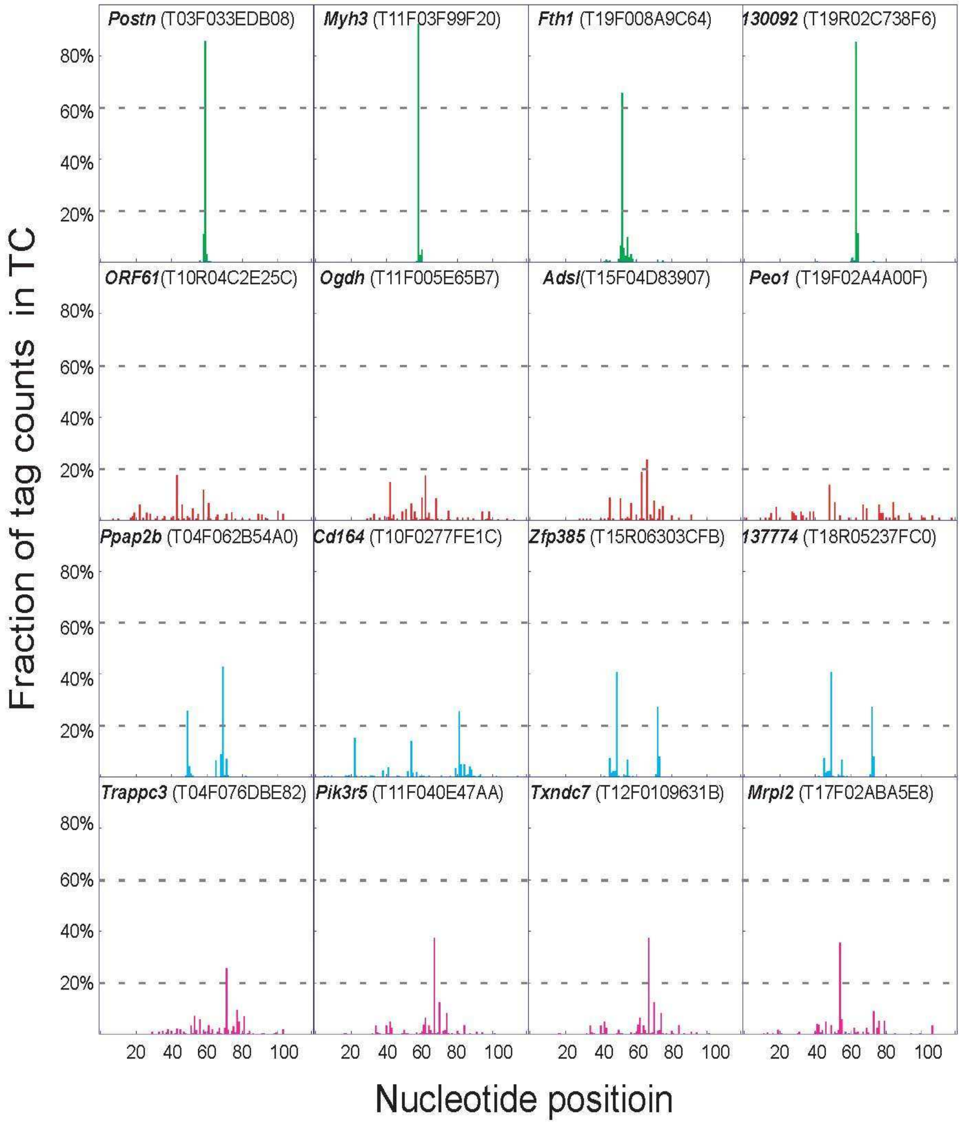
(protein coding only)

TSS

Transcriptional start sites are much less precise than expected. They can be organized in four classes

- ”traditional ones” with a precise isolated localization
- broad TSS
- bimodal (sometimes multimodal)
- a localized TSS within a broad background

Genomic mappings of TSS: modality



Single dominant peak (SP)

Broad (BR)

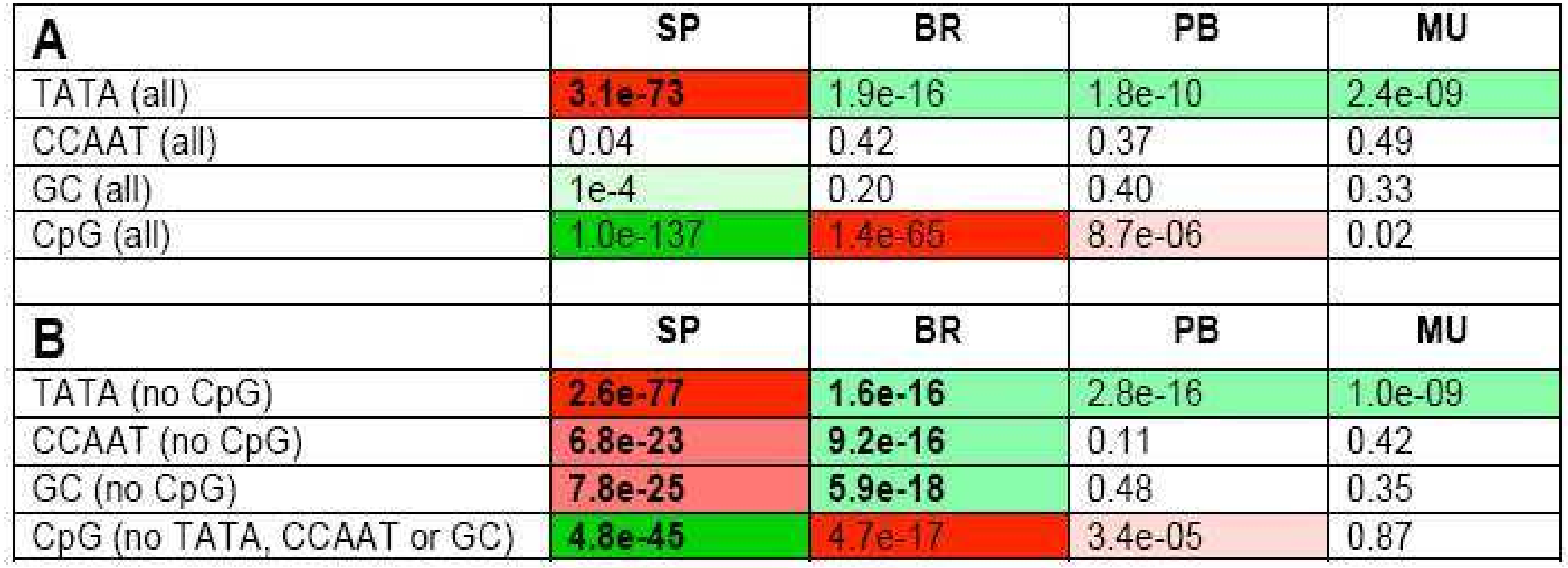
Bi- or Multimodal (MU)

Broad with dominant peak (PB)

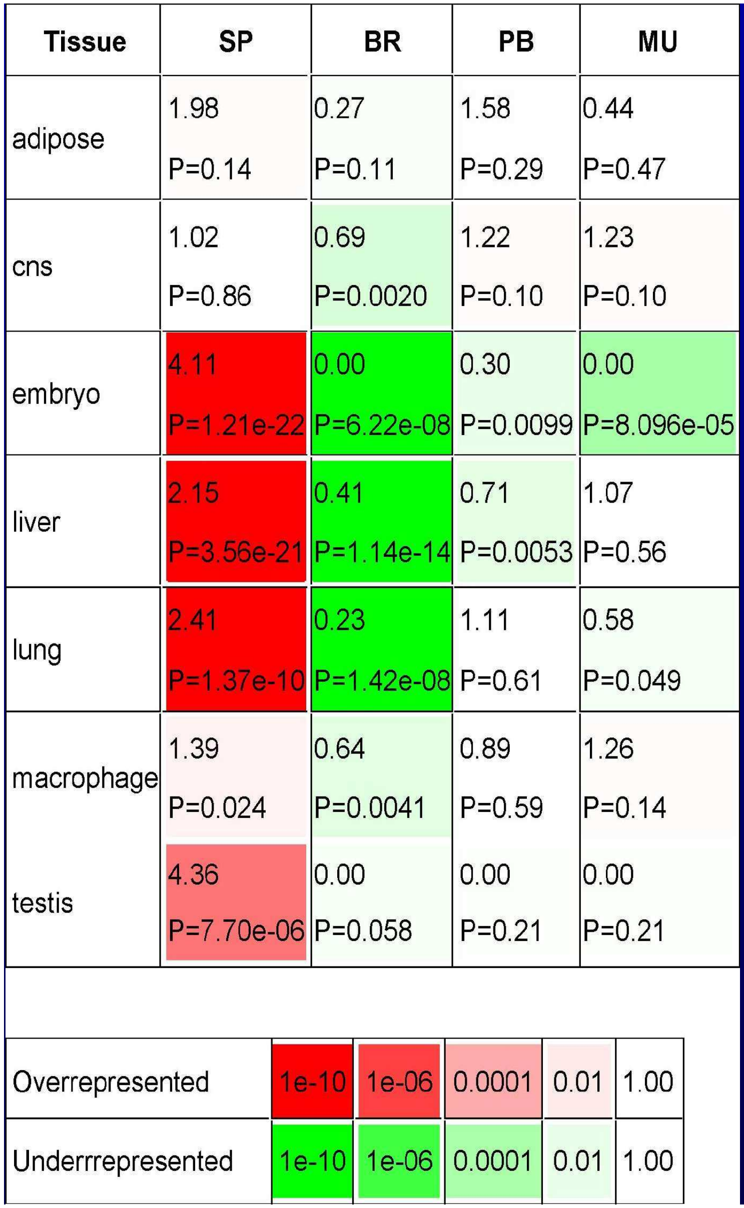
Association of TSS types with features of promoter sequences (TSS with >=100 tags)

🟥 RED: overrepresented

🟩 Green: underrepresented



A	SP	BR	PB	MU
TATA (all)	🟥 3.1e-73	🟩 1.9e-16	🟩 1.8e-10	🟩 2.4e-09
CCAAT (all)	0.04	0.42	0.37	0.49
GC (all)	🟩 1e-4	0.20	0.40	0.33
CpG (all)	🟩 1.0e-137	🟥 1.4e-65	🟥 8.7e-06	0.02
B	**SP**	**BR**	**PB**	**MU**
TATA (no CpG)	🟥 **2.6e-77**	🟩 **1.6e-16**	🟩 2.8e-16	🟩 1.0e-09
CCAAT (no CpG)	🟥 **6.8e-23**	🟩 **9.2e-16**	0.11	0.42
GC (no CpG)	🟥 **7.8e-25**	🟩 **5.9e-18**	0.48	0.35
CpG (no TATA, CCAAT or GC)	🟩 **4.8e-45**	🟥 4.7e-17	🟥 3.4e-05	0.87



Tissue	SP	BR	PB	MU
adipose	1.98 P=0.14	0.27 P=0.11	1.58 P=0.29	0.44 P=0.47
cns	1.02 P=0.86	0.69 P=0.0020	1.22 P=0.10	1.23 P=0.10
embryo	4.11 P=1.21e-22	0.00 P=6.22e-08	0.30 P=0.0099	0.00 P=8.096e-05
liver	2.15 P=3.56e-21	0.41 P=1.14e-14	0.71 P=0.0053	1.07 P=0.56
lung	2.41 P=1.37e-10	0.23 P=1.42e-08	1.11 P=0.61	0.58 P=0.049
macrophage	1.39 P=0.024	0.64 P=0.0041	0.89 P=0.59	1.26 P=0.14
testis	4.36 P=7.70e-06	0.00 P=0.058	0.00 P=0.21	0.00 P=0.21

Overrepresented	1e-10	1e-06	0.0001	0.01	1.00
Underrrepresented	1e-10	1e-06	0.0001	0.01	1.00

Tissue specific promoters are generally TATA box, SP type, while “housekeeping transcript” are controlled by CpG

http://fantom.gsc.riken.go.jp/

- **Databases**
- cDNA Annotation (Annotation strategy)
  - RIKEN cDNA Annotation Viewer
  - Public cDNA Viewer
- Sense/Antisense
  - SADB (Sense/Antisense Database)
- CAGE (Cap-Analysis Gene Expression)
  - CAGE Basic Viewer (CAGE primary Database) [mouse | human]
  - CAGE Analysis Viewer (Promoter Database) [mouse | human]
- Genomic Elements Viewer [mouse | human]