29 Novembre 2005

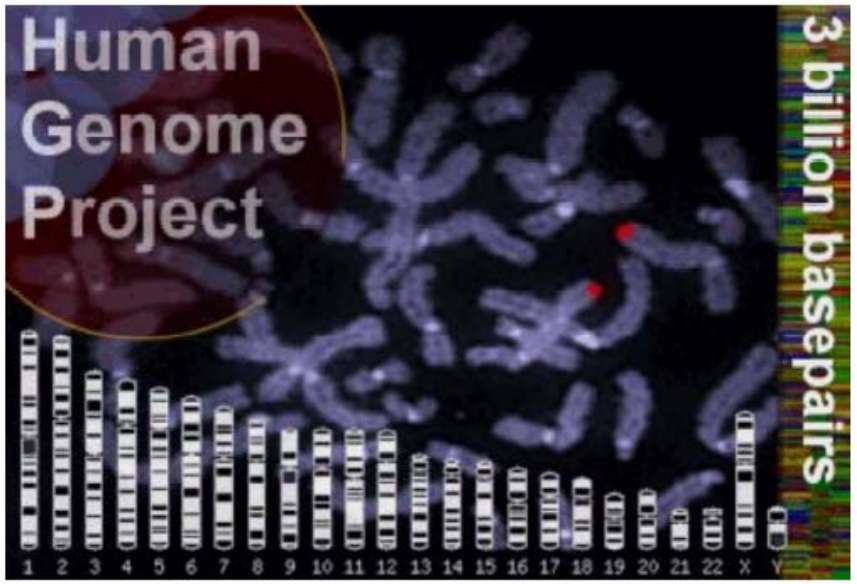
“Il trascrittoma dei mammiferi”

dott. Manuela Gariboldi

Gruppo di ricerca IFOM: Genetica molecolare dei tumori (responsabile dott. Paolo Radice)

The past 5 years have seen the completion of several mammalian genome sequences

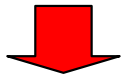
Organism	Number of genes in the genome
Myscoplasma genitalium	517
Saccharomyces cerevisiae	6,275
Arabidopsis thaliana	~ 20,000
Caenorhabditis elegans	19,099
Haemophilus influenzae	1,743
Drosophila melanogaster	13,601
Neisseria meningitidis	2,158
Homo sapiens	~ 30,000



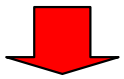
Only around 2% of the genome is translated into proteins

NEXT STEP

Decode the way genes are translated into functions required to create and mantain a mature organism



which 2% is translated and how is it controlled?



Transcriptome analysis

Transcriptome analysis

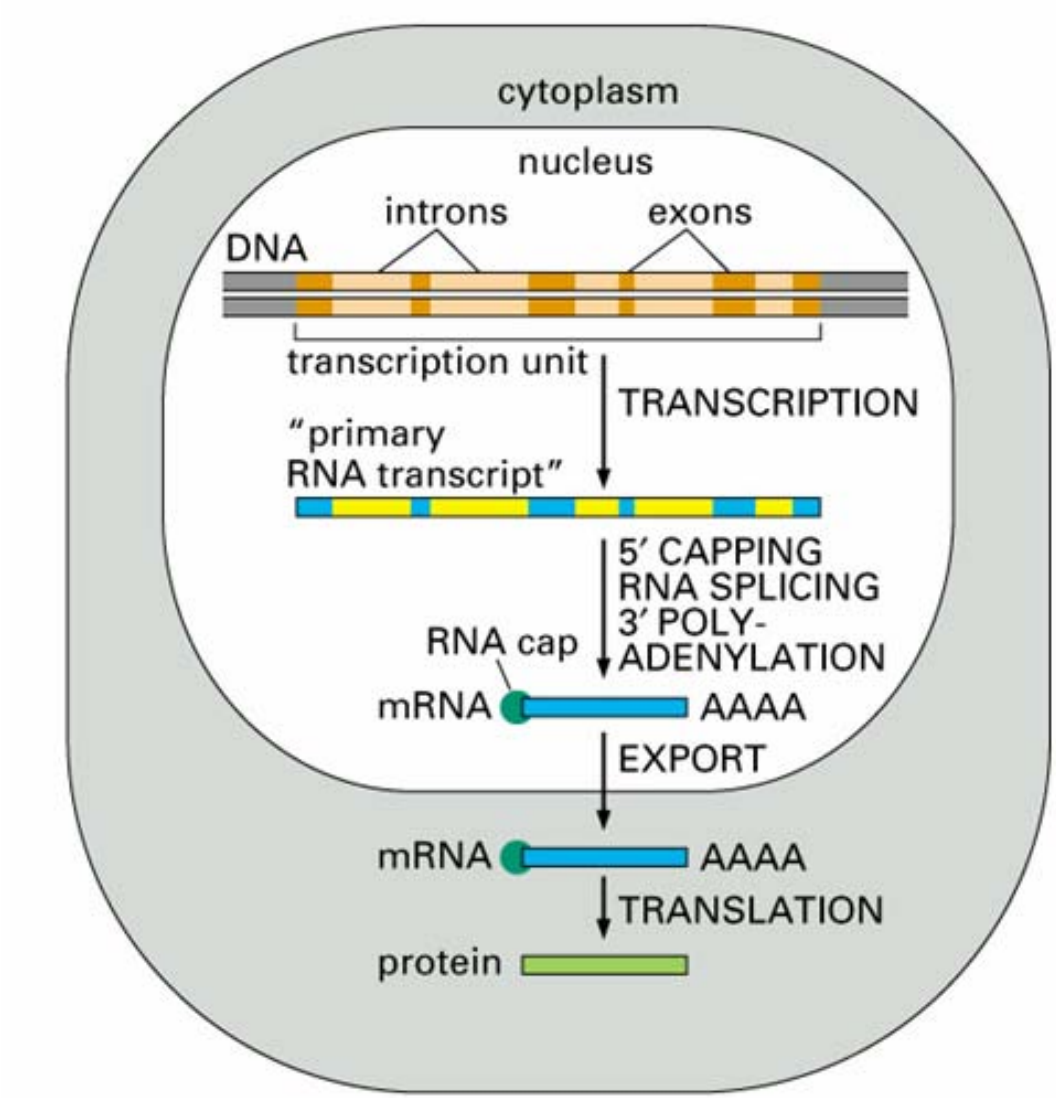


Figure 6–21 part 1 of 2. Molecular Biology of the Cell, 4th Edition.

RNA molecules must be individually isolated and sequenced

Different cell types express different genes

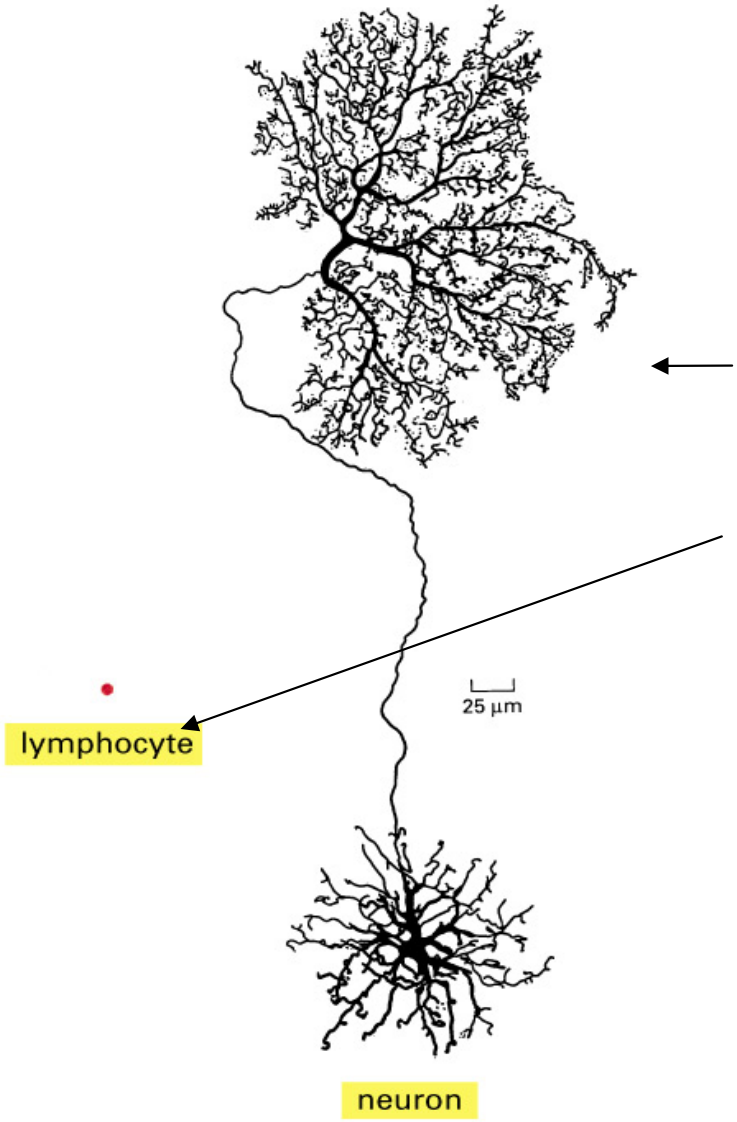


Figure 7–1. Molecular Biology of the Cell, 4th Edition.

A neuron has a greater size and more complex morphology (shape) than a lymphocyte

Although all cells always have all genes not all genes are activated in the same way in all cells

Different gene expression profiles (sets of genes) when regulated in concert (in a coordinated manner) lead to dramatic differentiation of cells

RNA molecules must be individually isolated and sequenced from all the different tissues present in an organism

Analysis of mammalian transcriptome

The FANTOM Consortium

(Functional Annotation of Mouse)

Established in 2000 in Japan by Riken Genomic Sciences Center and 45 research institutes from 11 countries (Australia, Germany, Greece, Italy, Japan, Singapore, Sweden, Switzerland, South Africa, the United Kingdom and the United States)

AIM

comprehensively annotate the transcripts from mammalian (mouse) genome

Analysis of transcripts expressed in almost all mouse organs, tissues and developmental stages

Why annotation of mouse?

Mouse is the most used and known mammalian experimental model

The use of mouse allows to sample starting tissues appropriately for constructing cDNA libraries

Analysis of samples including early embryonic stages and preimplantation embryos

RIKEN prepared and characterized full-length cDNA libraries from about 240 mouse tissues and cell types

cDNA library: collection of clones which includes all the RNA transcripts (copied in cDNA) expressed in a tissue, including the less represented (rare mRNAs)

Full-length: complete conding sequence of the transcript (from starting to termination site)

➡ predictions of human full-length cDNA sequences in-silico are done by homology searches with the mouse full-length cDNAs

Result: three papers on mouse transcriptome

NATURE | VOL 409 | 8 FEBRUARY 2001 | www.nature.com

Functional annotation of a full-length mouse cDNA collection

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

Annotated 21.076 transcripts

NATURE | VOL 420 | 5 DECEMBER 2002 | www.nature.com/nature

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*

Annotated 60.770 transcripts

The Transcriptional Landscape of the Mammalian Genome

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

www.sciencemag.org SCIENCE VOL 309 2 SEPTEMBER 2005

Annotated 181.047 transcripts

FANTOM3: Functional Annotation Of Mouse - 3

The Transcriptional Landscape of the Mammalian Genome

Analyzed over 2 million sequences of RNAs produced from the mouse genome

obtained more than 180,000 full length copies of these RNAs

Annotated and mapped on the mouse genome

Used several technologies:

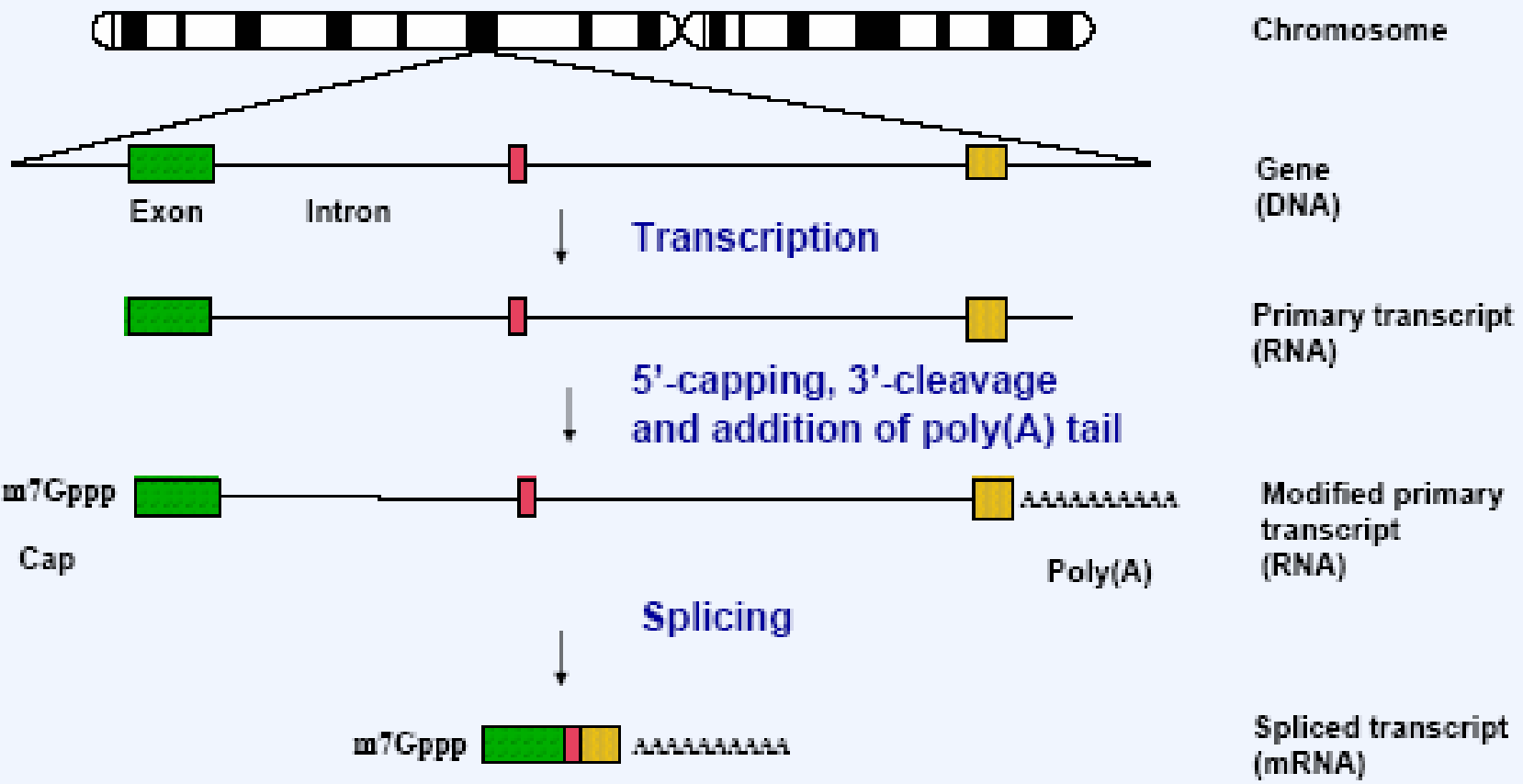
A) full length cDNA isolation and 5'- and 3'-end sequencing of cloned cDNAs

B) CAGE (cap analysis gene expression)

C) GIS (gene identification signature)

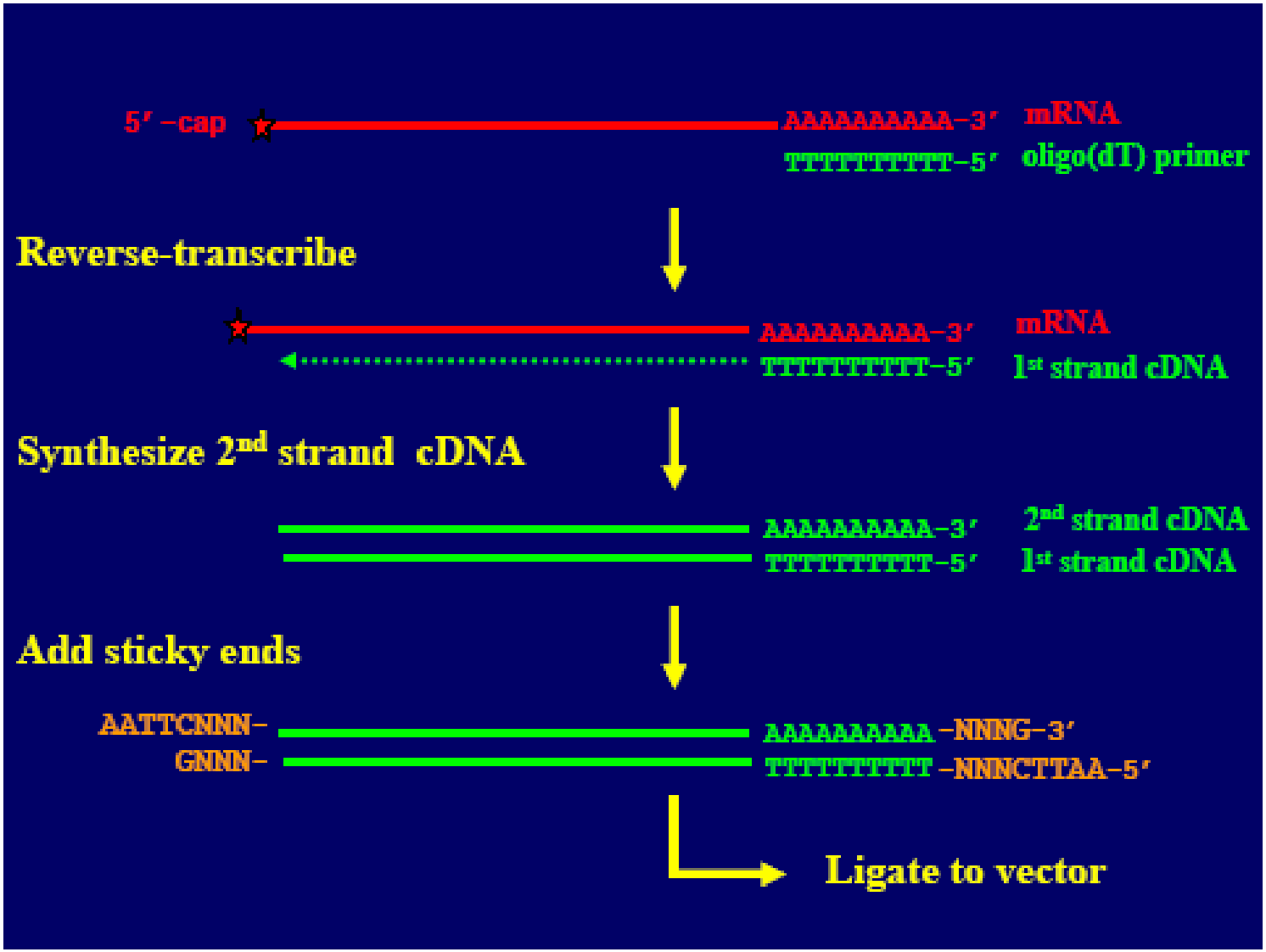
D) GSC (gene signature cloning)

Eukaryotic Gene Transcription



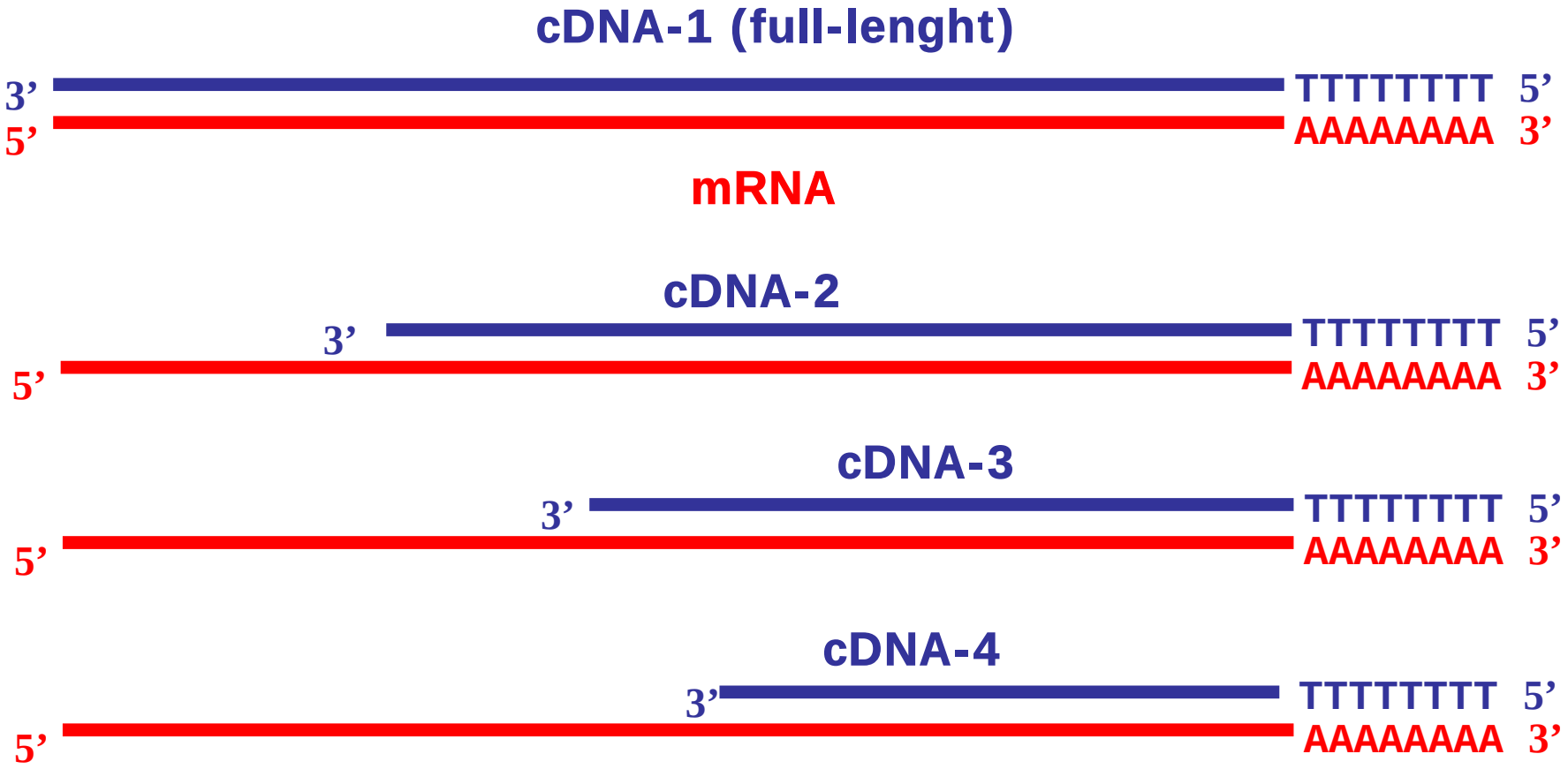
A) Full-length cDNA isolation and 5'- and 3'-end sequencing of cloned cDNAs

cDNA library construction



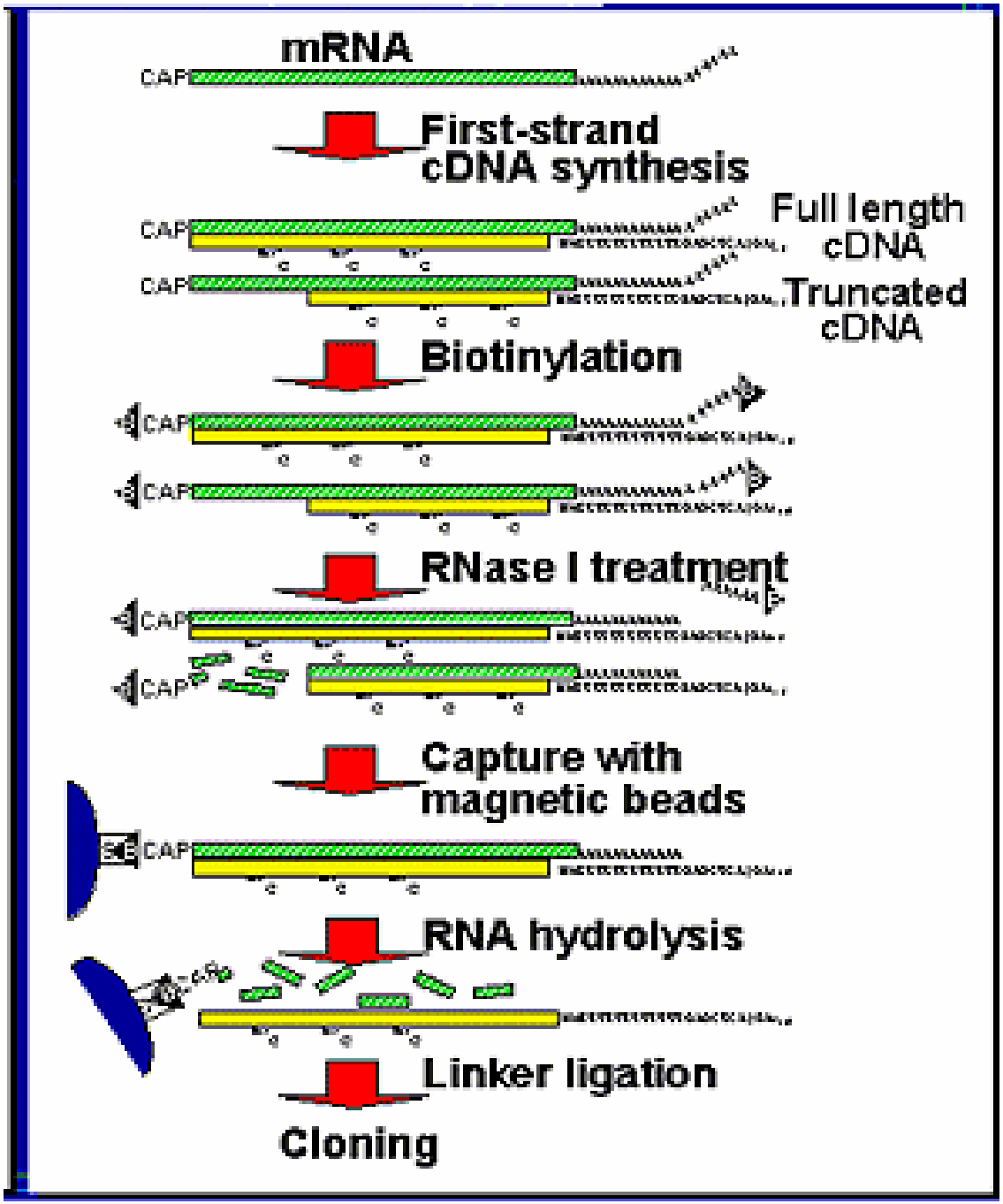
Full-length cDNA isolation and 5’- and 3’-end sequencing of cloned cDNAs

Problem: cDNA synthesis starting form poly-A makes molecules with different length



Full-length cDNA isolation allows cloning protein coding regions, which is necessary for genomics functional studies

Full-length RIKEN cDNA libraries construction

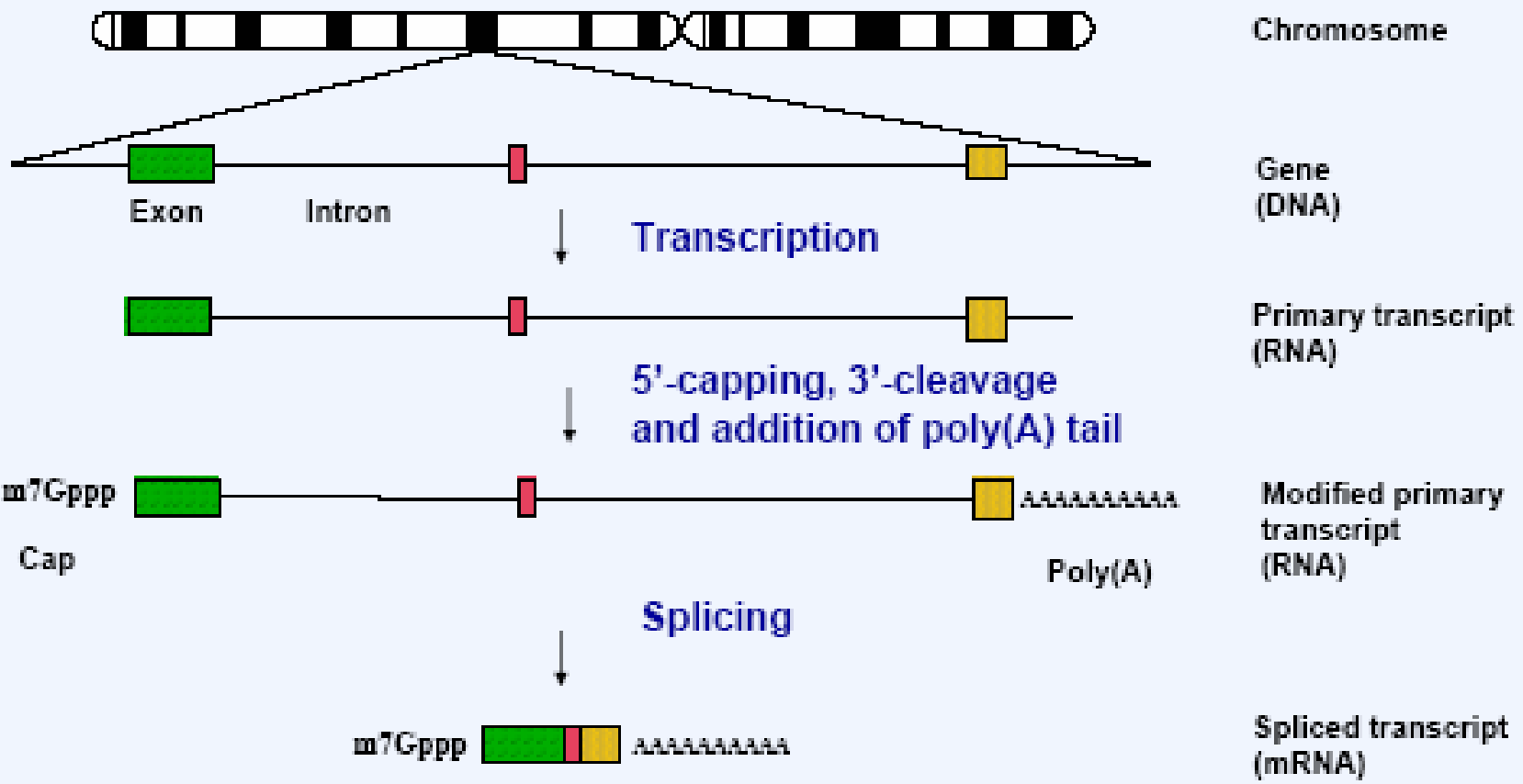


Library enrichment in rare cDNAs :

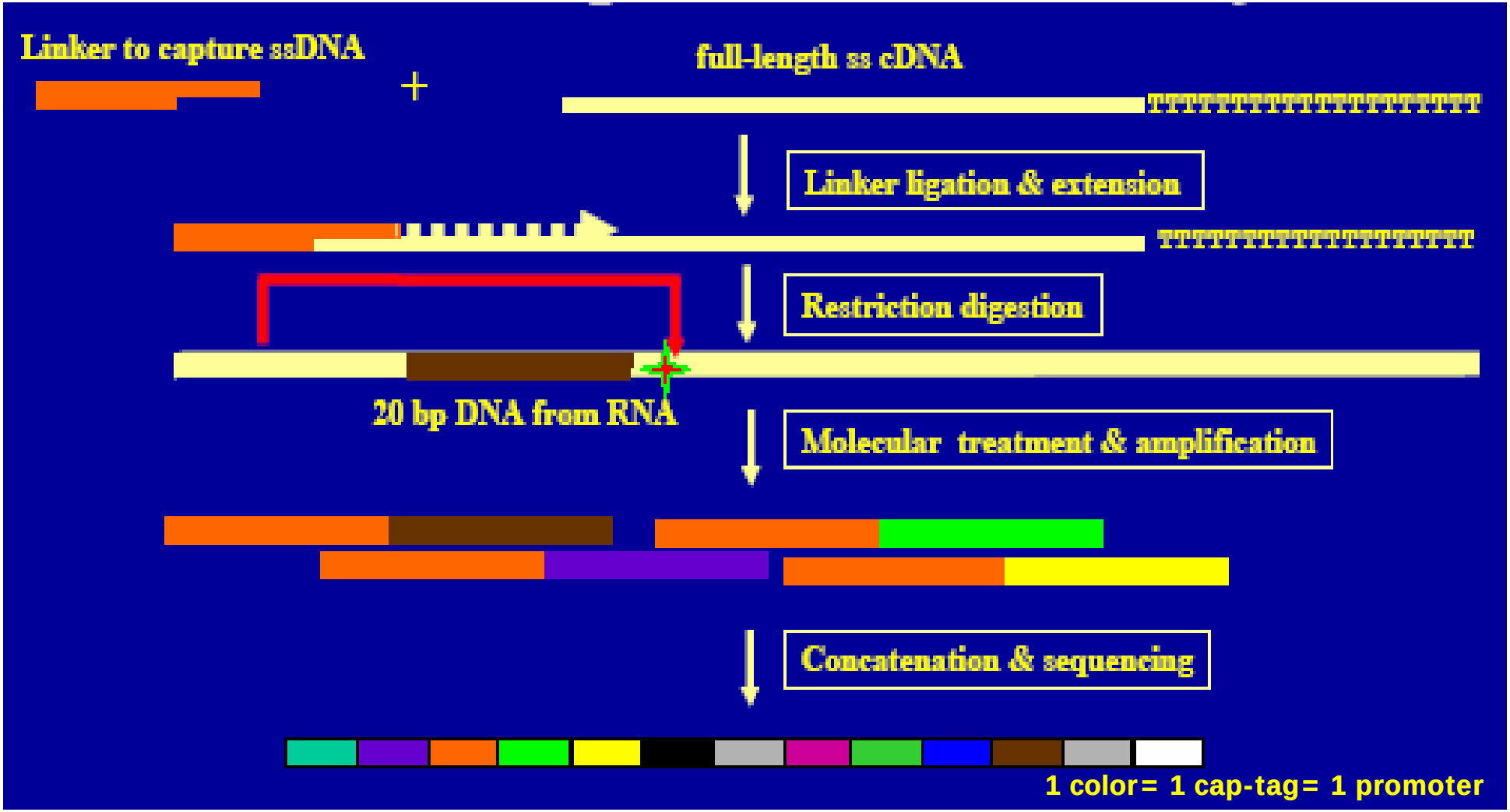
Libraries are normalized (abundant RNAs are reduced)
and subtracted (already identified RNAs are removed)

All clones identified: sequenced at 5’ and 3’ ends

Eukaryotic Gene Transcription

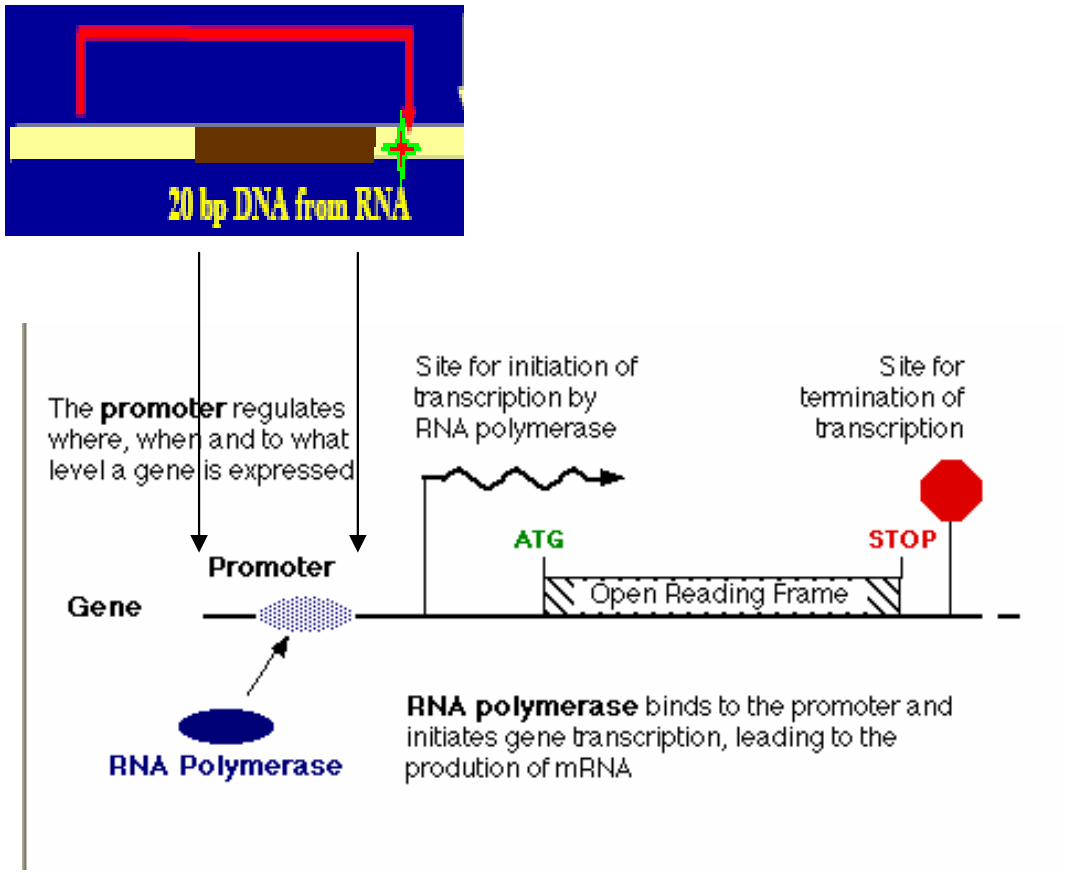


B) CAGE
Cap Analysis of Gene Expression



High-throughput technique (1 clone = 20-30 cap-tags) which identifies transcriptional start points and promoter regions

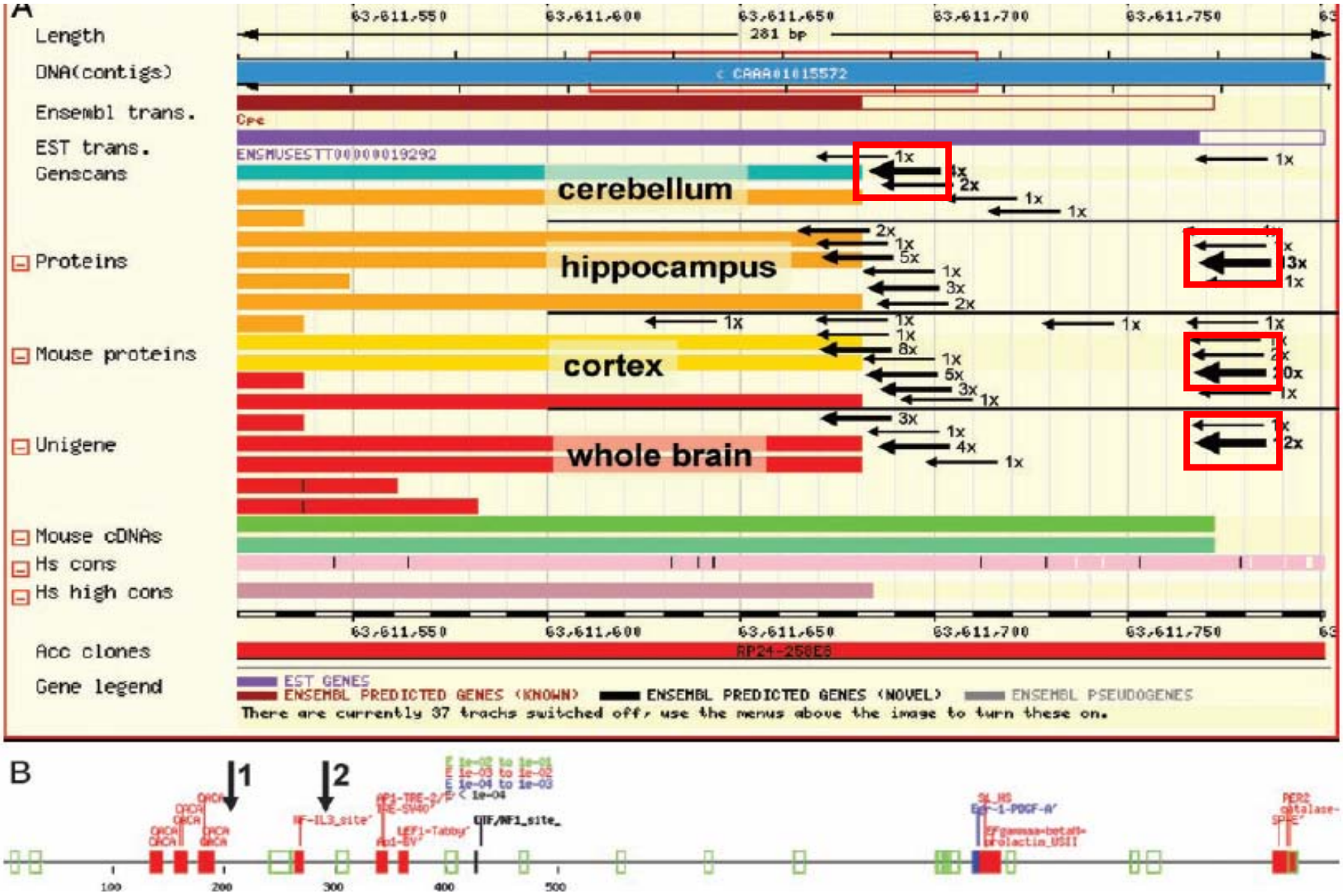
1 cap-tag contains the promoter region of a gene



Promoter sequencing allows to identify the transcription starting site of the transcript

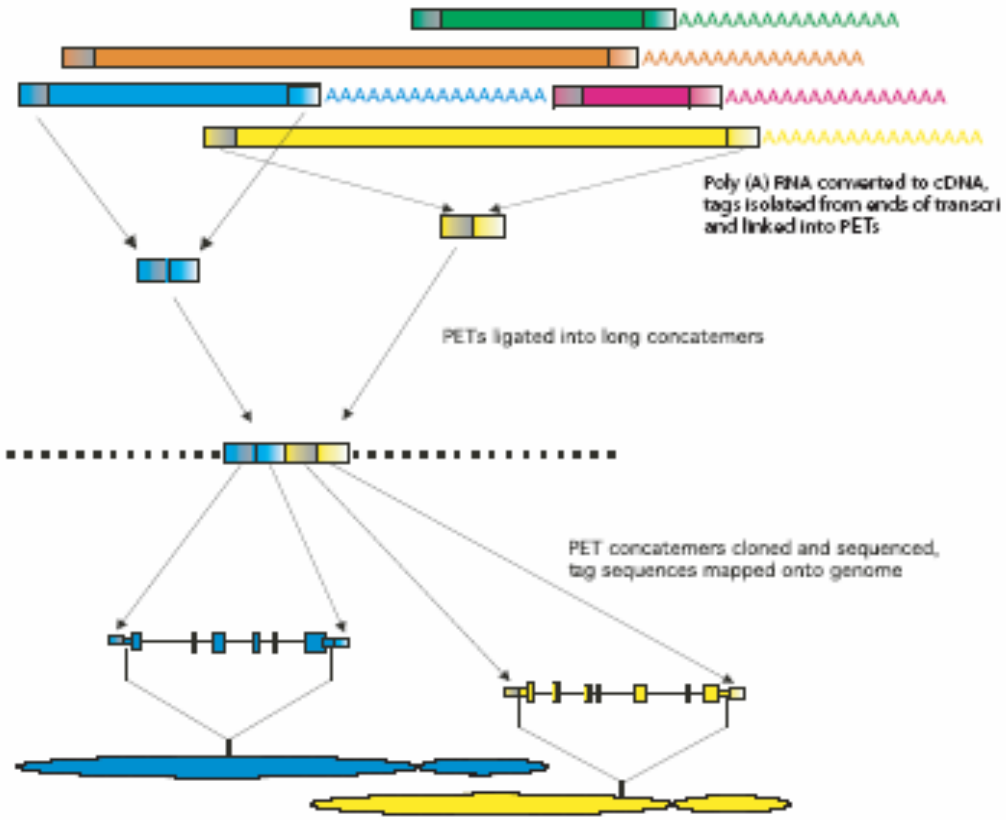
CAGE analysis of carboxypeptidase E

Different tissues have specific trascription starting points



Gene identification signature (GIS) analysis

Figure 1 | Identifying the 5′ and 3′ ends of transcripts using GIS analysis. mRNA is isolated and converted to cDNA, and short tags are obtained from each end of the transcript molecule. The tags are linked together to form a PET. PETs are ligated together to form large concatemers, which can be cloned and sequenced. Sequencing each concatemer can identify >15 PETs per reaction and tens to hundreds of thousands of PETs can be identified from each sample. The sequence of each PET can be mapped to the genome to determine the boundaries of the corresponding transcript.



GSC uses subtracted libraries, which allow detection of rare transcripts

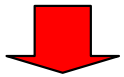
Fantom annotation of cDNA clones
(done by >100 curators in a teleconference)

- **Flowchart of annotation:**

– Define the CDS (coding sequences), if any

– Provide a controlled vocabulary and nomenclature (gene and protein name)

– Provide GO terms (give specific functions)



Identified 31.166 proteins, 16.274 of which are newly described and 5,154 transcripts encode for proteins that are very considerably different than known proteins or completely new

Results 1: transcripts

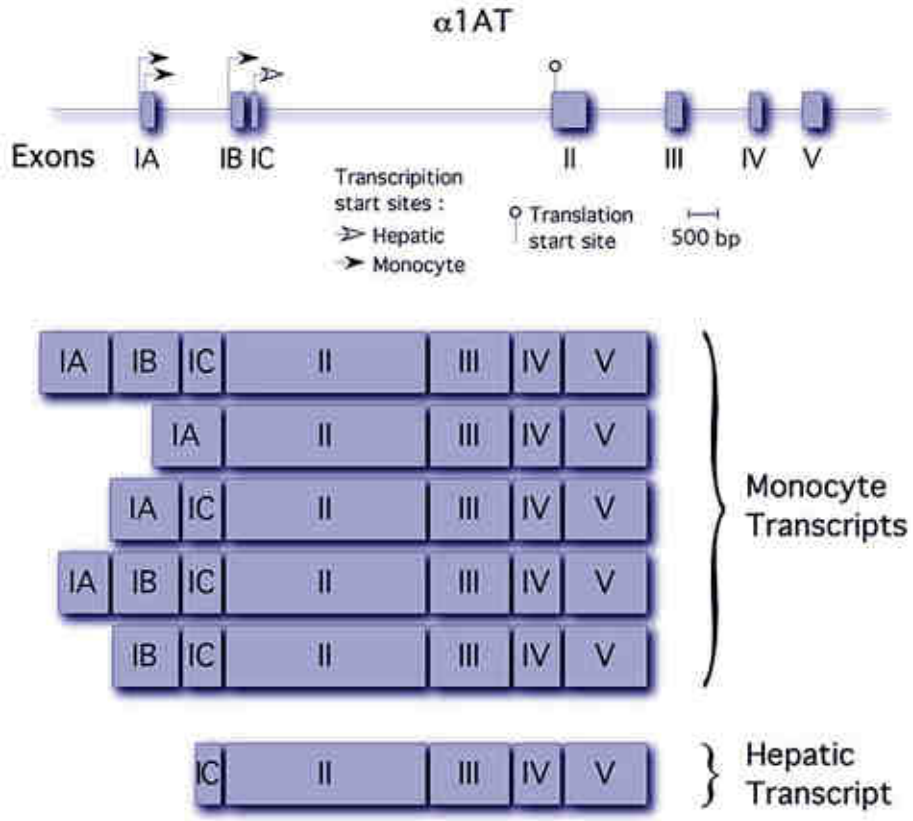
Extensive experimental evidence of a great variability of transcripts

A) Identified 181,047 transcripts with defined transcriptional boundaries (paired initiation and termination sites)

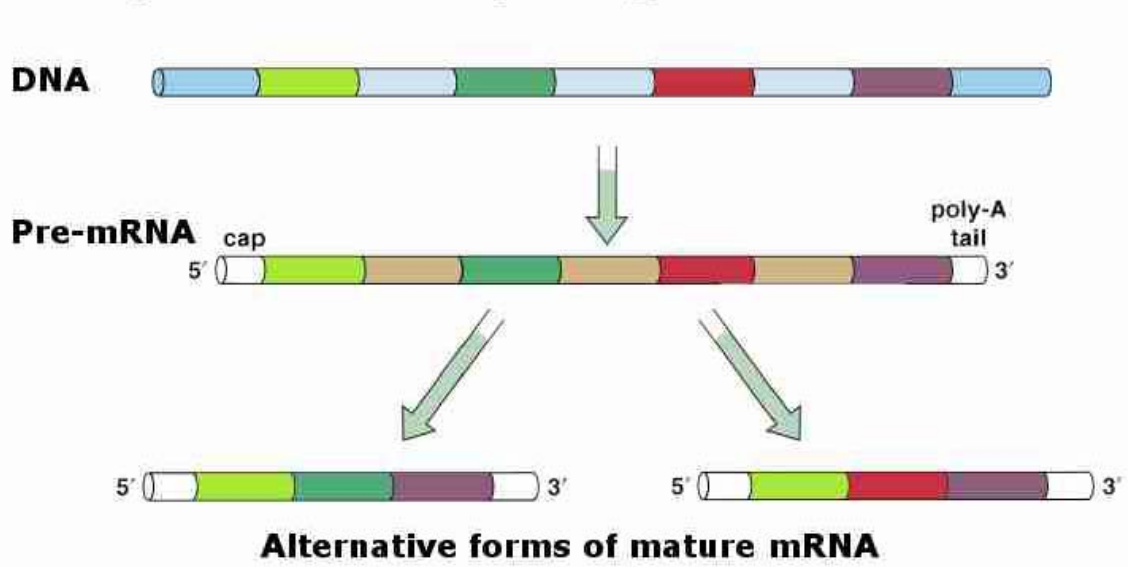
the number of transcripts is at least one order of magnitude larger than the estimated 22,000 genes in the mouse genome

B) Noticed an extensive variation in transcripts arising from alternative promoter usage, splicing and polyadenylation

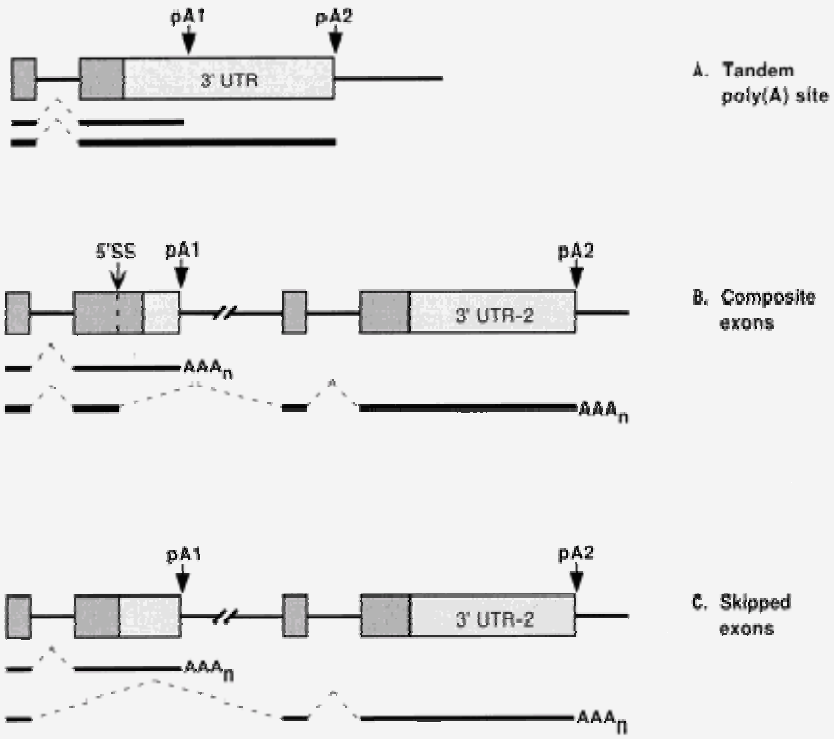
Alternative promoter



Alternative splicing

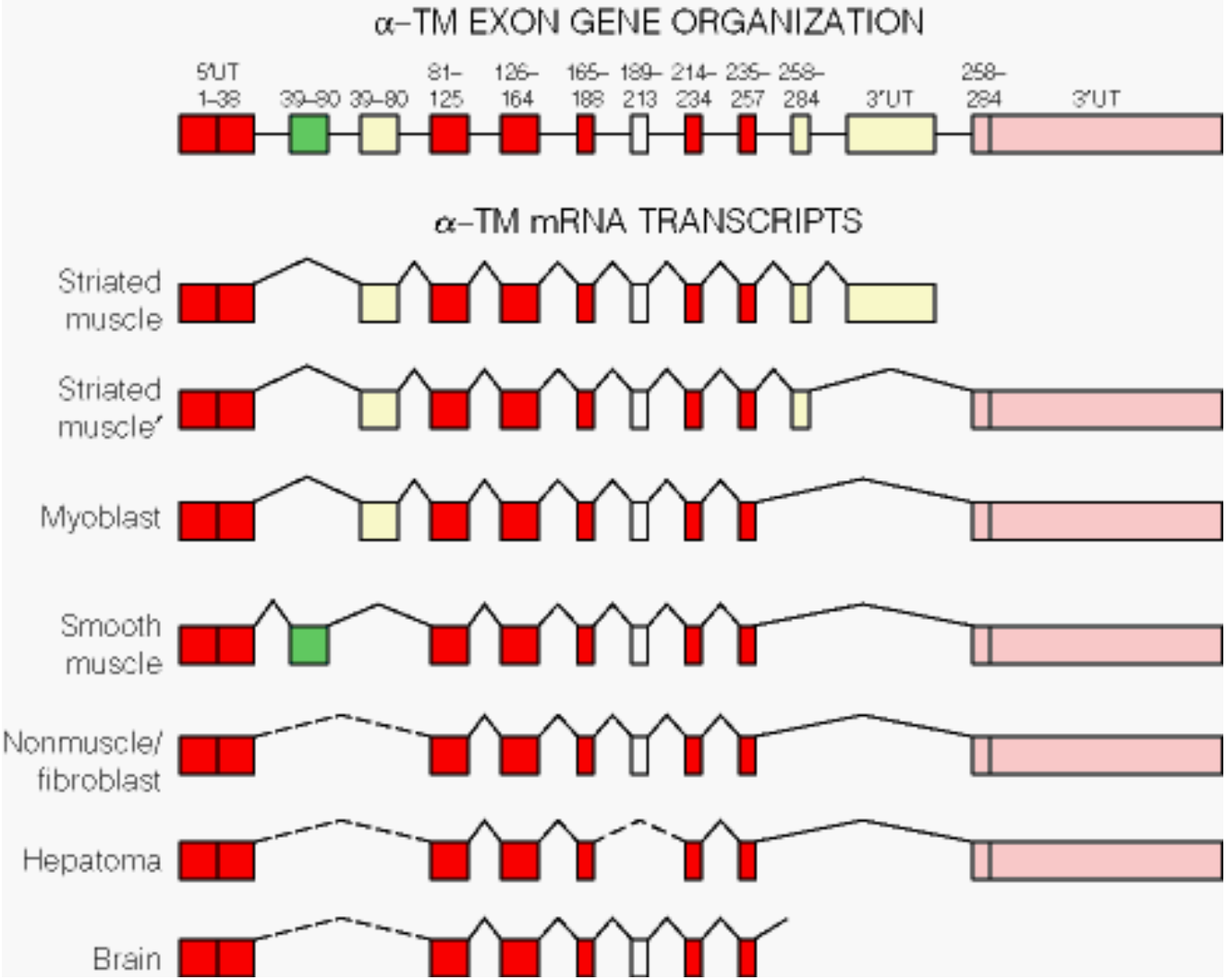


Alternative polyadenilation



Edwalds Gilbert et al. , NAR, 25, 2547, 1997

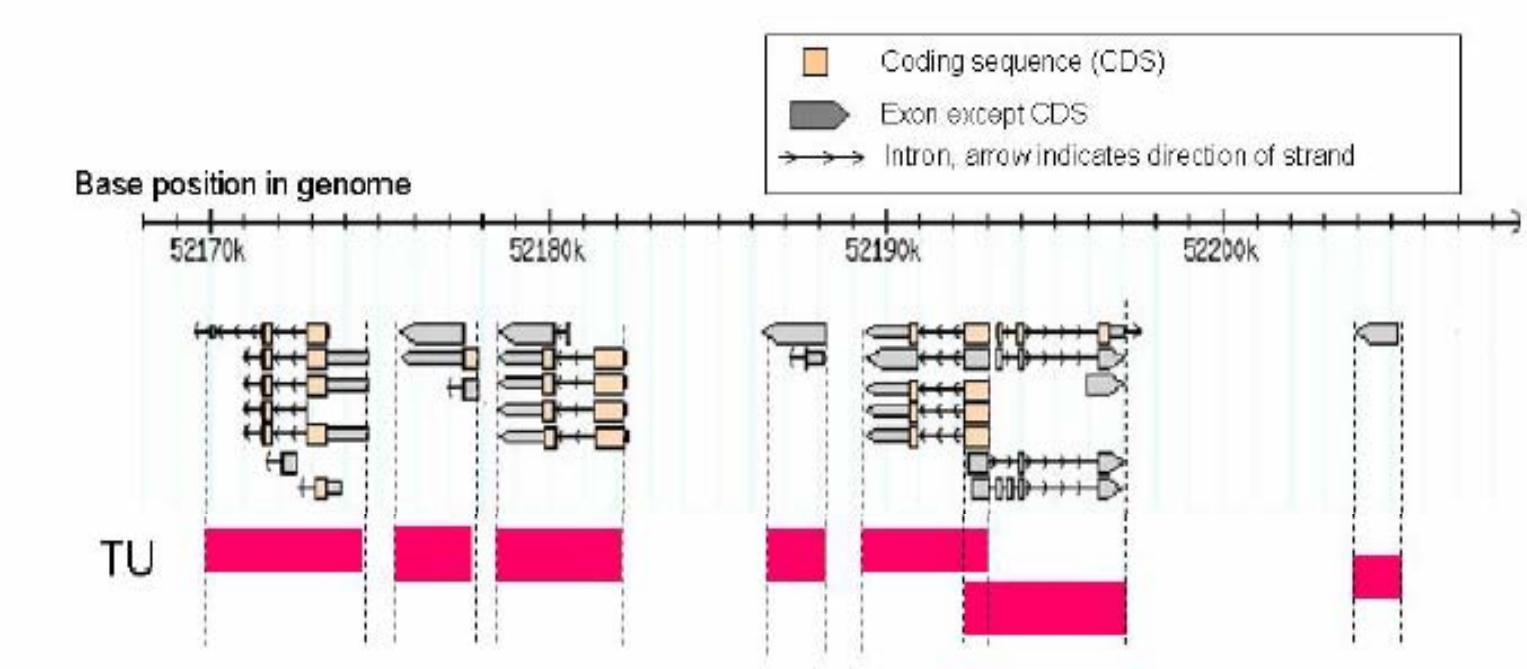
Splicing and termination variants are often cell type specific



Results 2: genome mapping of transcripts

Genes are organized in transcriptional units (TUs)

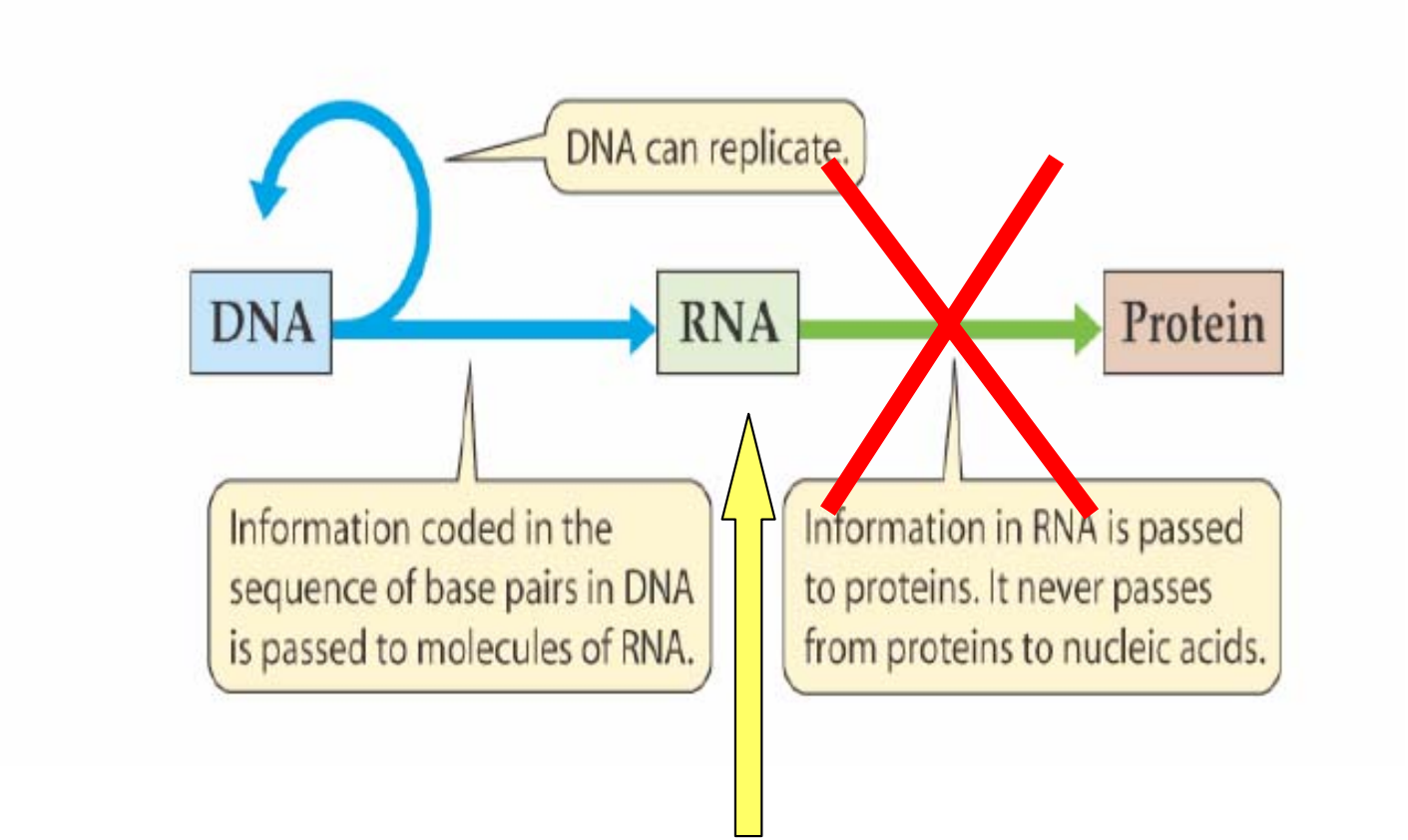
Genomic segment that groups transcripts with exon overlaps on the same strand (sharing direction and partly location)



The 181,047 transcripts identified groups in 44,147 TU
(65% of TUs contain multiple splice variants)

Results 3: non-coding RNAs

More than a third (34,030) of the cDNAs in the FANTOM3 data set lack any protein coding sequence (non-coding RNAs)



ncRNAs

- **implicated in different molecular and cellular events in eukaryotic cells**

- grouped in three classes of transcripts according to their number of nucleotides:

21–25 nucleotides: microRNAs

are the reverse complement of another gene's mRNA transcript and inhibit the expression of the target gene

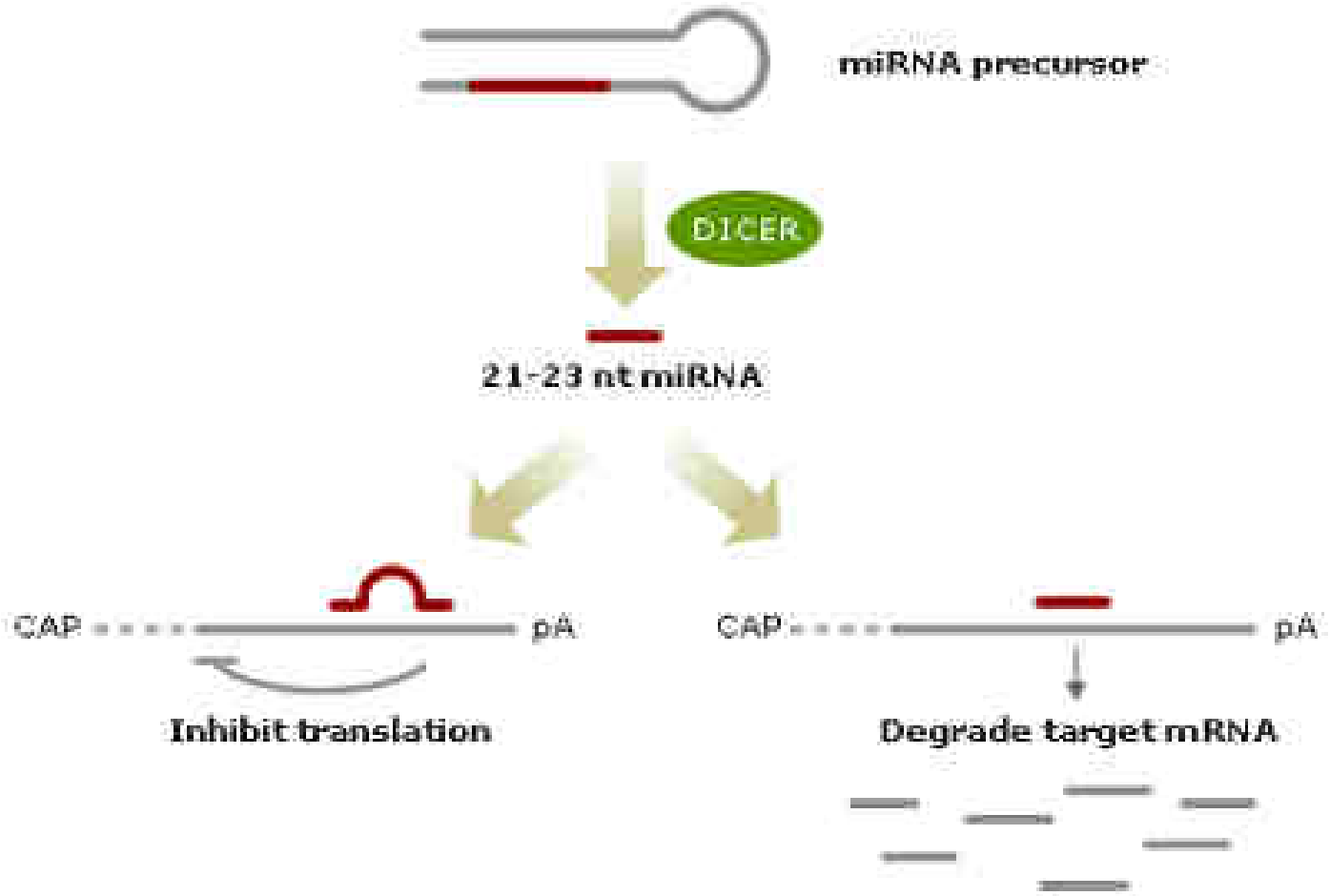
100–200 nucleotides: translational regulators

transfer RNA (tRNA) which are involved in the process of translation and gene expression
Small nuclear RNA (snRNA) found within the nucleus of eukaryotic cells. Involved in a variety of important processes such as RNA splicing and maintaining the telomeres
Small nucleolar RNA (snoRNA) involved in chemical modifications of ribosomal RNAs (rRNAs) and other RNA genes

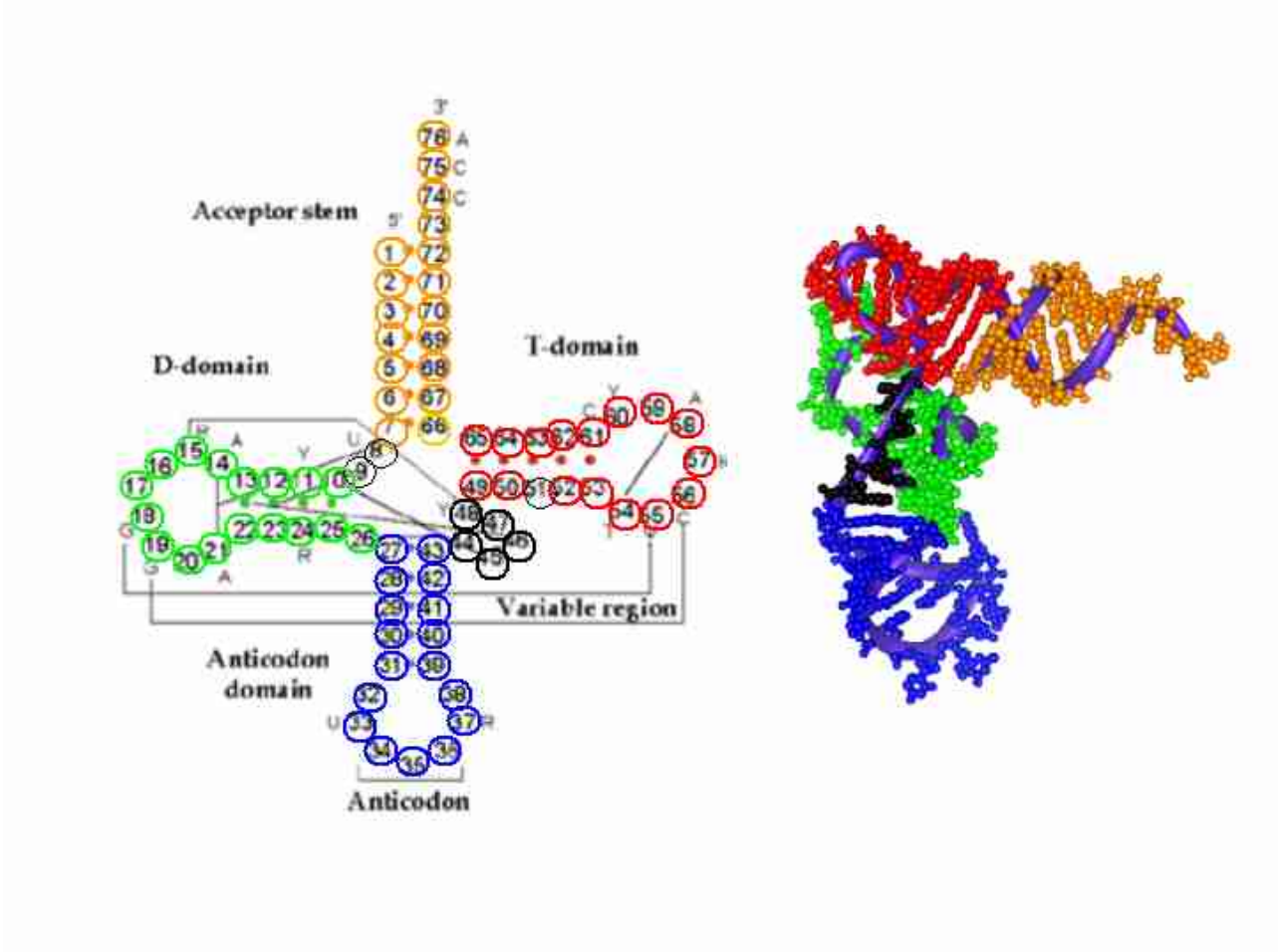
up to over 10,000: involved in gene silencing

(the XIST gene was cloned as a large ncRNA being expressed exclusively from the inactive X-chromosome)

microRNA


miRNA precursor
DICER
21-23 nt miRNA
CAP
pA
Inhibit translation
CAP
pA
Degrade target mRNA

transfer RNA (tRNA)



ncRNAs function

- role in protein synthesis (ribosomal and transfer RNAs)

- implicated in control processes such as genomic imprinting and perhaps more globally in control of genetic networks

- regulate expression of genes

- changes in expression levels of ncRNAs have been described in complex diseases such as cancer and neurological diseases

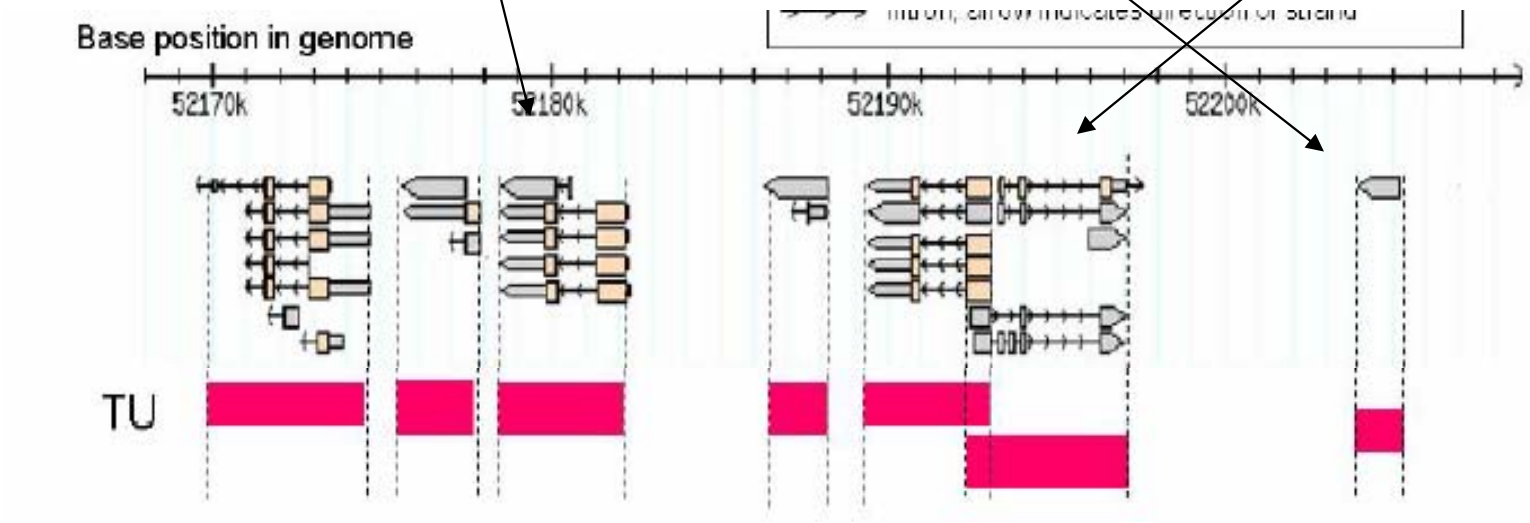
Results 4: genome mapping of ncRNA

ncRNA distribution in the genome

- Many are singletons

- Many appear to start from initiation sites in 3’ untranslated regions of protein-coding loci

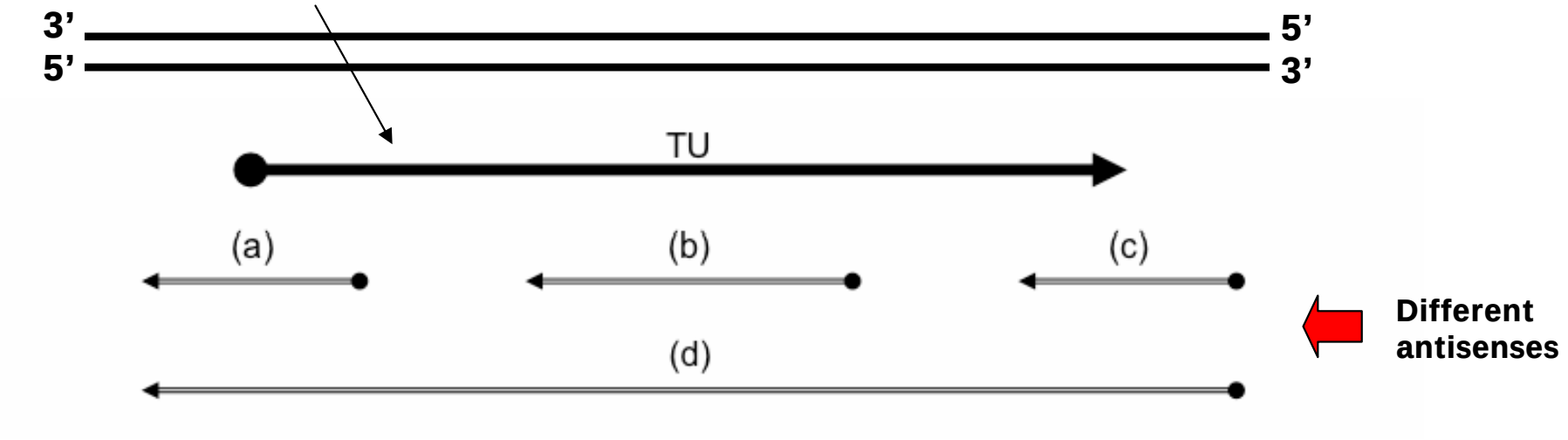
- More than half of TUs contain exclusively non-protein coding RNA



Results 5: sense/antisense pairs

RNA is transcribed in sense/antisense pairs

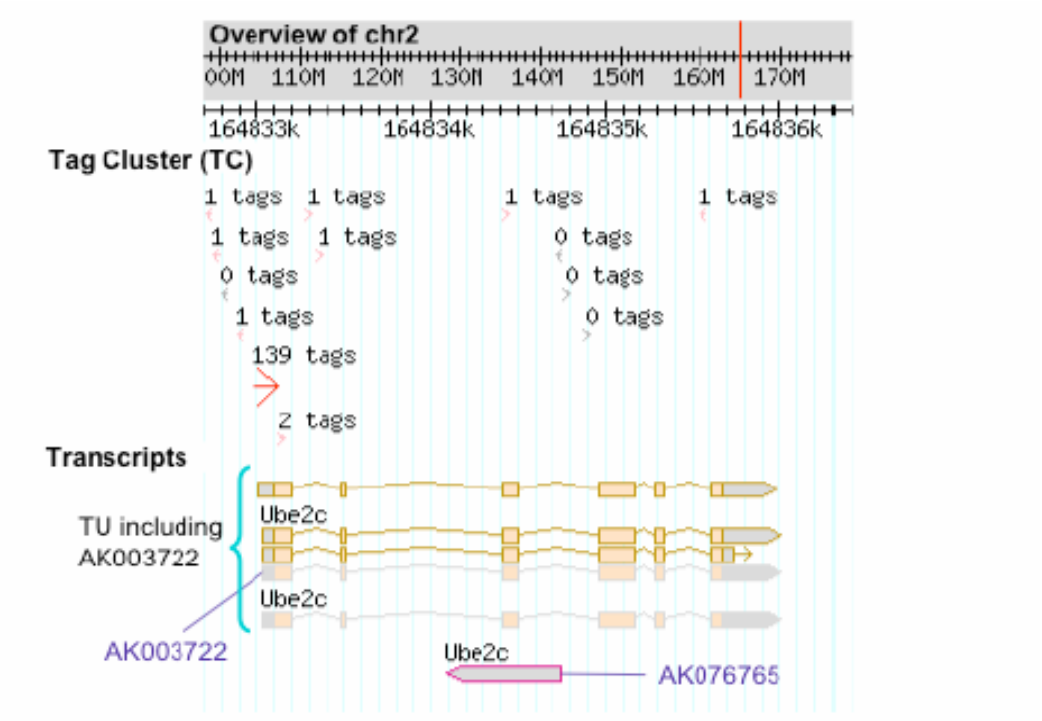
The sense strand of DNA generally provides the template for production of mRNA, which in turn encodes proteins



Transcription from the opposite (antisense) strand can produce transcripts that hybridize with the coding DNA strand to interfere with transcription or mRNA stability

Results 5: sense/antisense pairs

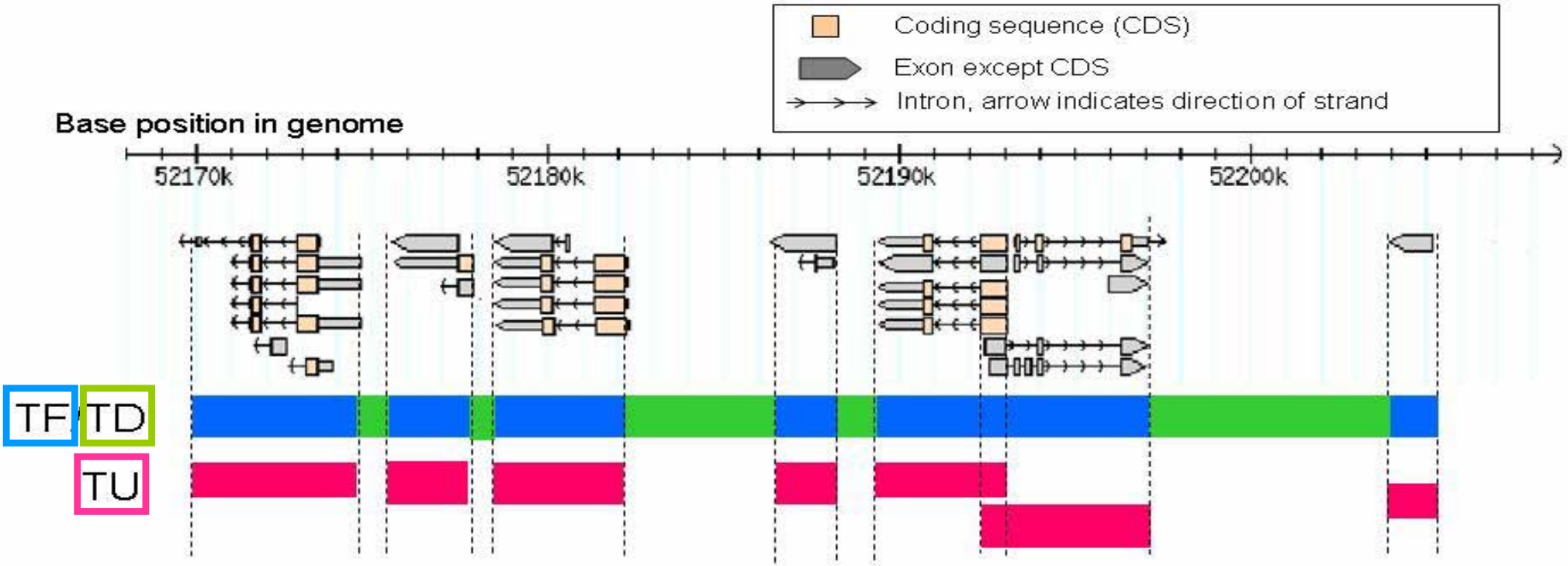
Sense/antisense distribution in the genome



- 4520 TU pairs contain full-length transcripts, which form S/AS pairs on exons
- antisense transcription is more widespread than previously thought in the mammalian genome
- Co-expressed S/AS pairs show tissue-specific regulation

Genome organization

TUs can be clustered together into transcript forests (TF) that are transcribed in either strands without gaps
TFs encompass 62.5% of the genome are separated by transcription deserts (TD), regions devoid of transcription



Conclusions

A gene is no longer only the section of DNA that is transcribed to RNA to be translated into a protein

RNA is not simply an intermediary between DNA and functional protein

RNA has a multitude of intrinsic functions and activities

Perspectives

The data provide tools to understand the control networks that are needed to create a complex organism such as a mammalian

The development of multicellular organisms like mammals is controlled by vast amounts of regulatory noncoding RNAs that until recently were not suspected to exist or be relevant to our biology

The fraction of protein-coding DNA in the genome decreases with increasing organism complexity

Many of the differences between species may be embedded in the differences in the RNA regulatory control systems, which are evolving much faster than the protein components