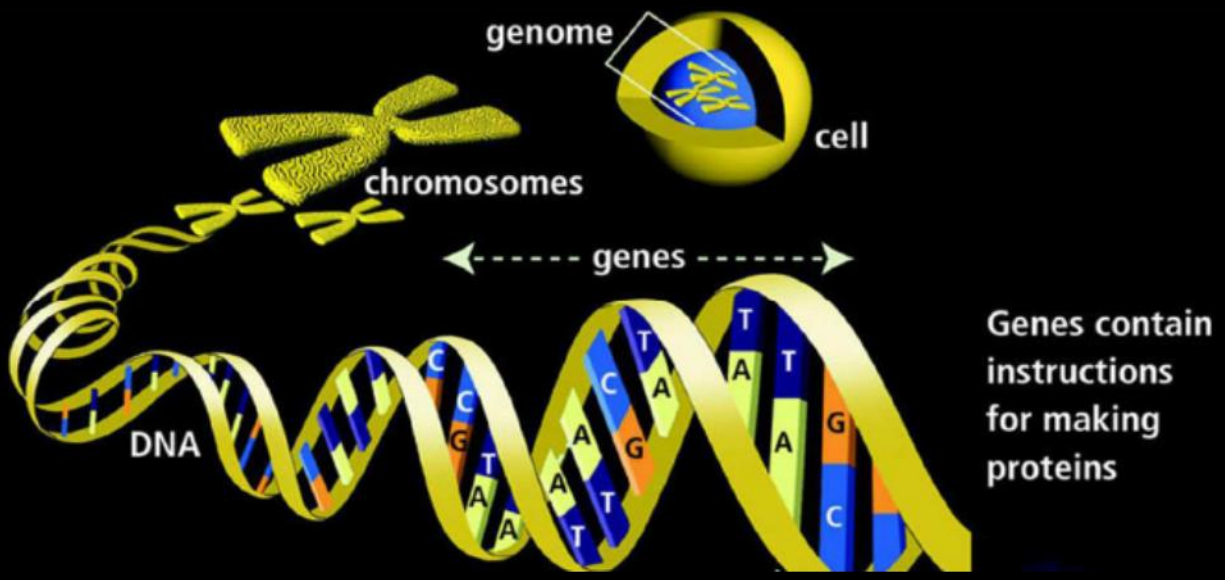
M1 Data Science

Introduction to Genomics

Sébastien Aubourg

sebastien.aubourg@inra.fr

Genome



Each cell contains a complete copy of the genome, distributed along chromosomes
(3x10^9 base pairs –3 Gb- in human DNA: 6 meters in each cell)

Encodes blueprint for all cellular structures and activities and which cells go where…

sebastien.aubourg@inra.fr

Genome

1920 : All genes of an organism

Now, the genome is all the DNA in a cell :
All the DNA on all the chromosomes
Includes genes, intergenic sequences, repeats

Eukaryotes can have 2-3 genomes

- Nuclear genome
- Mitochondrial genome
- Plastid genome

If not specified, "genome" usually refers to the nuclear genome.

sebastien.aubourg@inra.fr

Genomics

Genomics is the study of genome(s), including large chromosomal segments containing many genes.

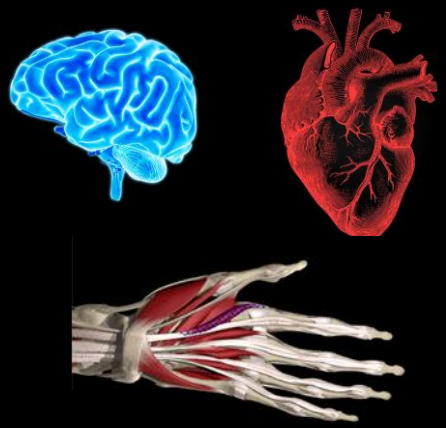
The initial phase of genomics aims sequence an initial set of entire genome(s).

Functional genomics aims to deduce information about the function of DNA sequences (structural and functional annotations from bioinformatics and high-throughput experimental approaches). Should continue long after the initial genome sequences have been completed…

sebastien.aubourg@inra.fr

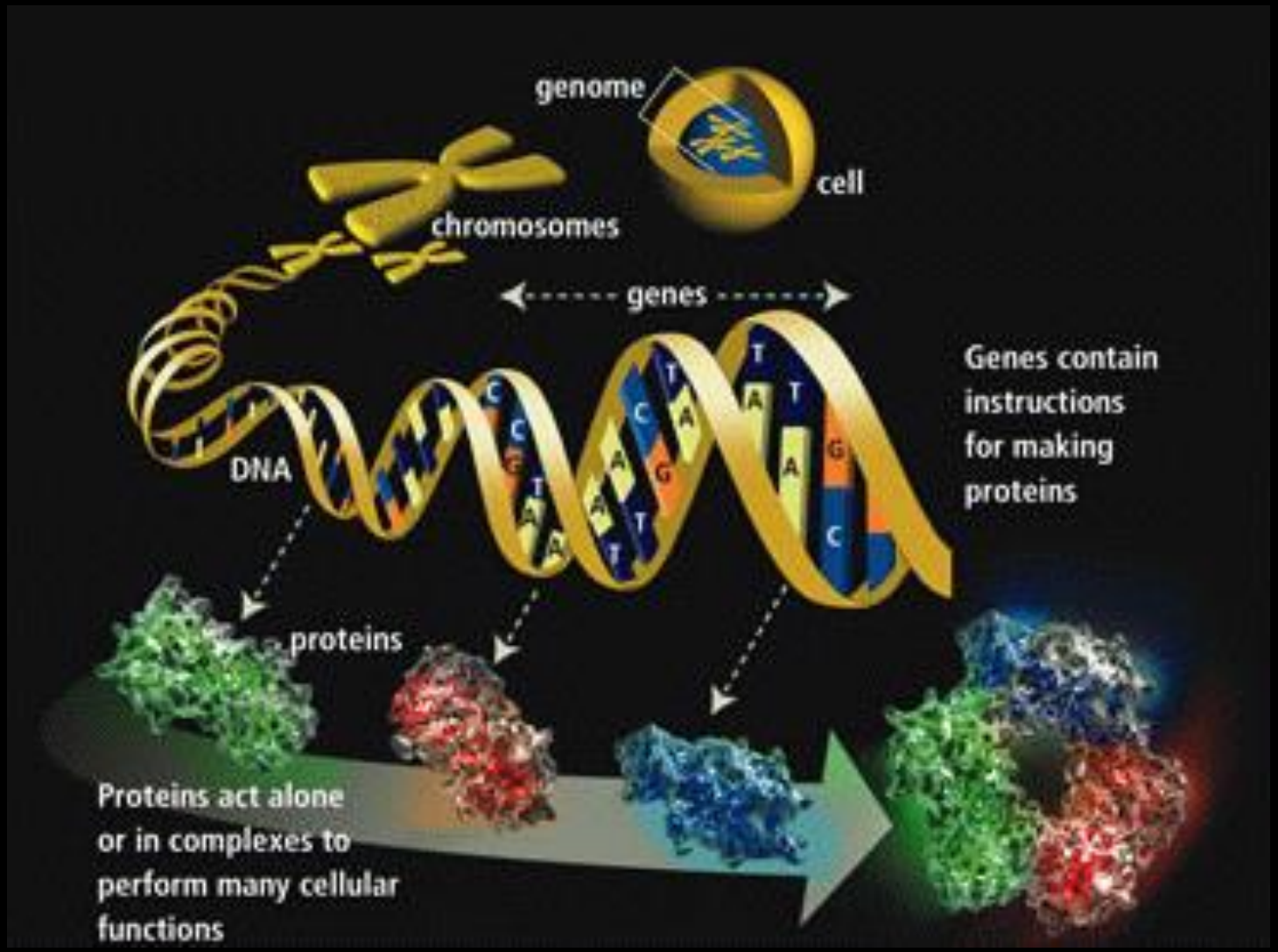
Different genomes

One genome -> different functions



sebastien.aubourg@inra.fr

Gene expression : from genes to proteins



sebastien.aubourg@inra.fr

Plan

- Gene expression
- Sequencing and sequences
- Genome annotation
  - Structure
  - Function
  - Evolution

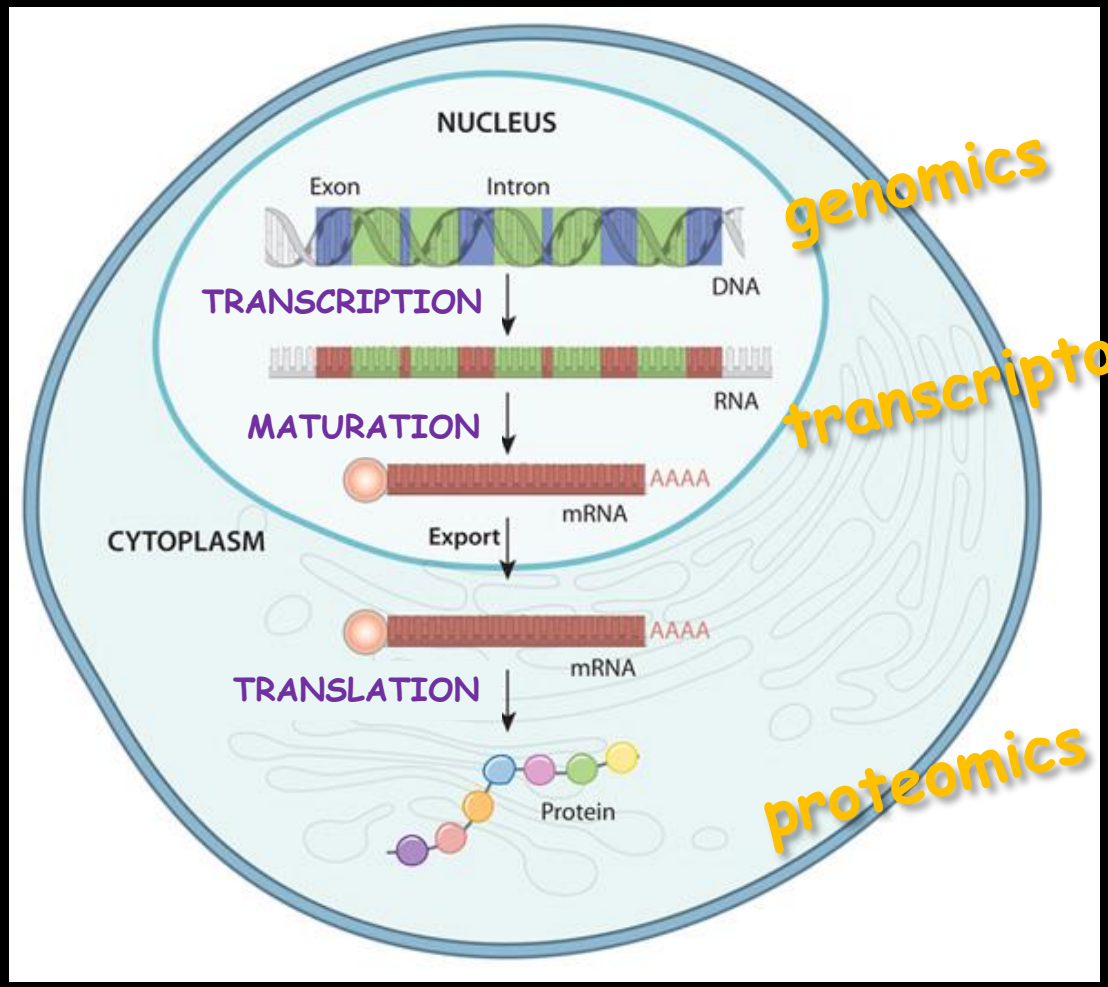
sebastien.aubourg@inra.fr

Gene Expression

Basis of molecular genetic

sebastien.aubourg@inra.fr

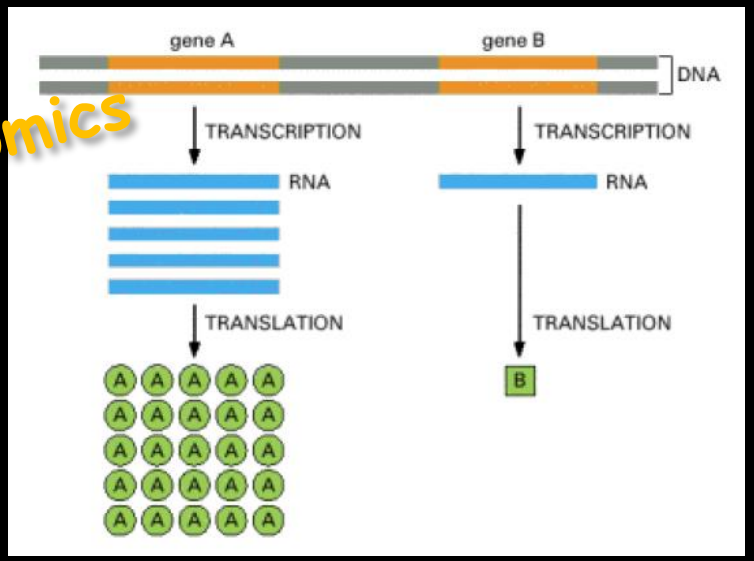
Gene expression : from genes to proteins



genomics

transcriptomics

proteomics



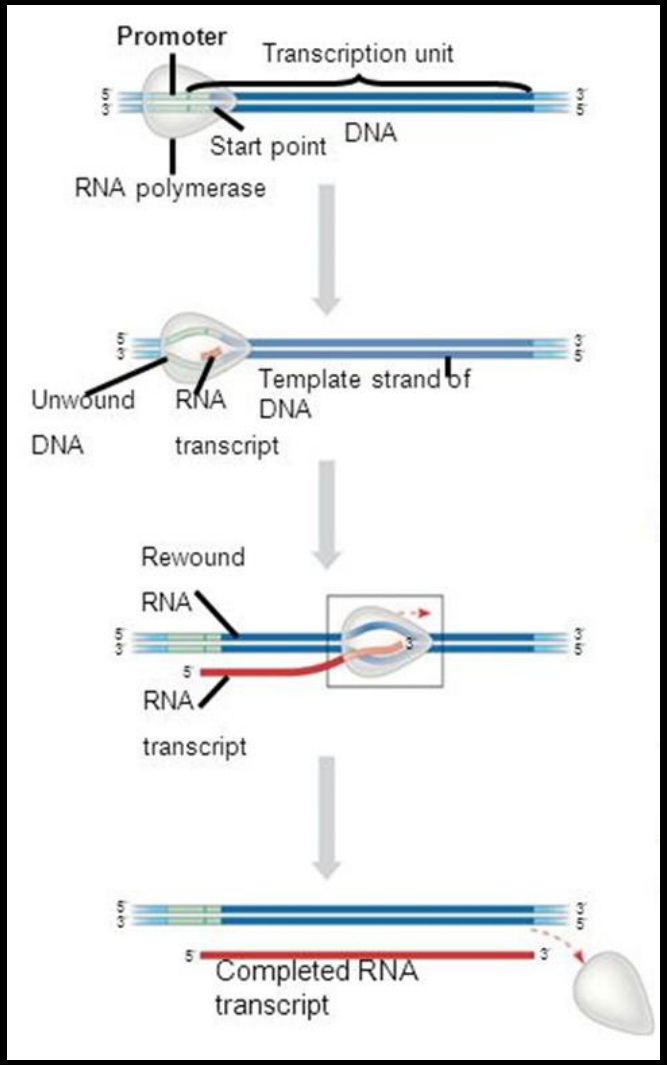
sebastien.aubourg@inra.fr

Transcription, 3 steps

1. Initiation

2. Elongation

3. Termination



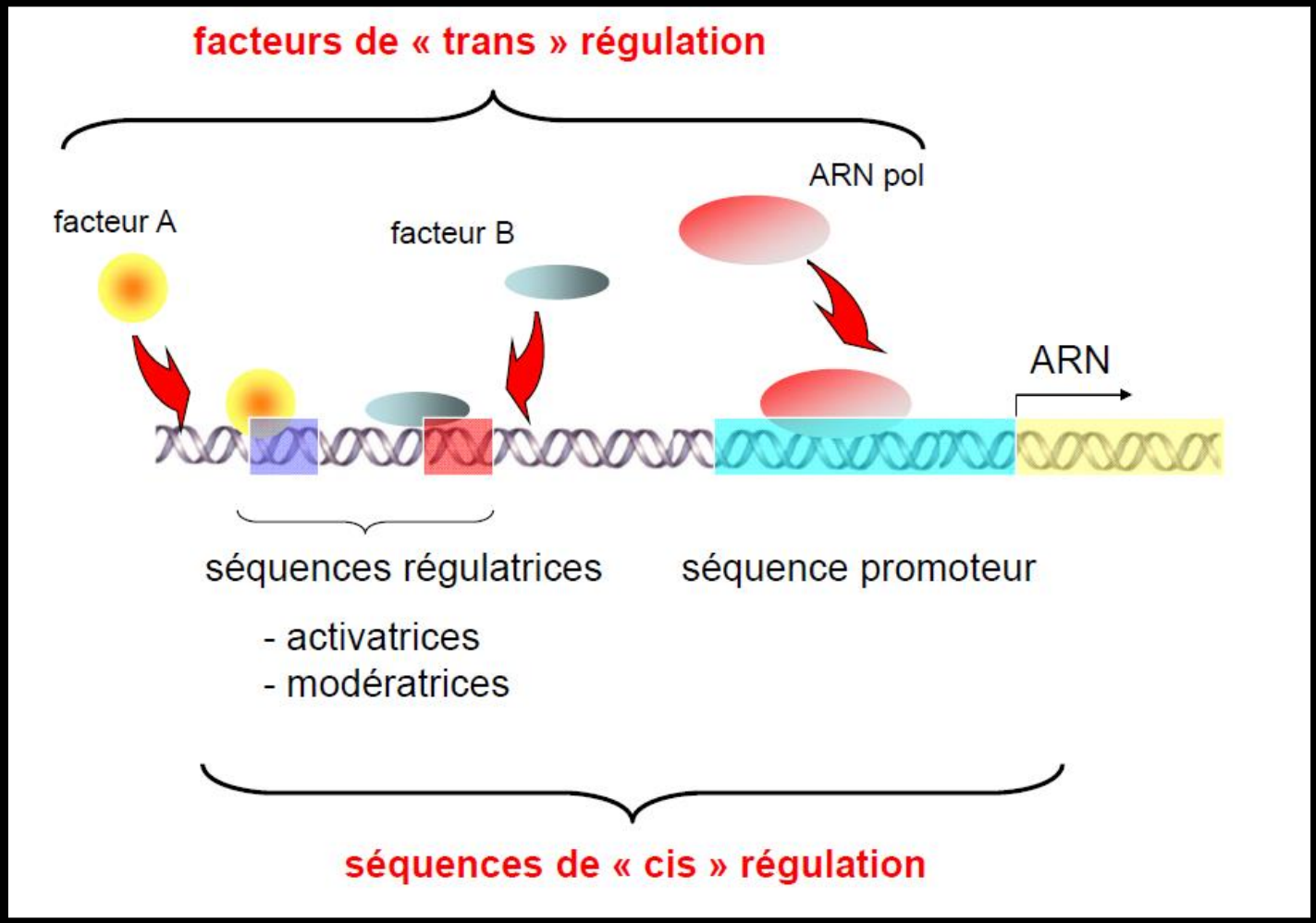
RNA polymerase binds to a promoter sequence

mRNA copy of gene (transcriptional unit) is synthesized 5' to 3'

Termination sequence causes transcription to stop

sebastien.aubourg@inra.fr

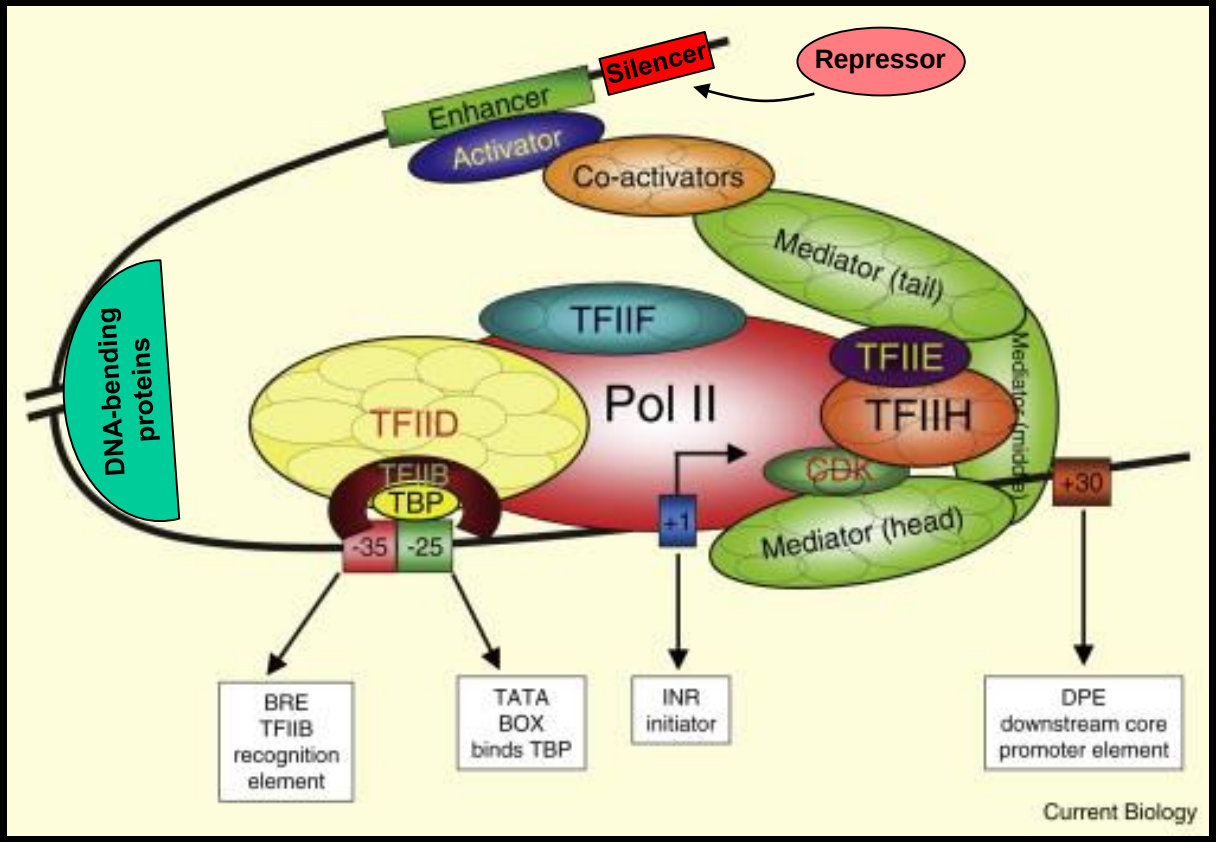
Transcription initiation



sebastien.aubourg@inra.fr

Transcription initiation

Chromatin remodeling allows DNA-protein interactions.
Transcription factors (TF), activators, co-activators, mediators recruit RNA polymerase II to constitute the transcription preinitiation complex.

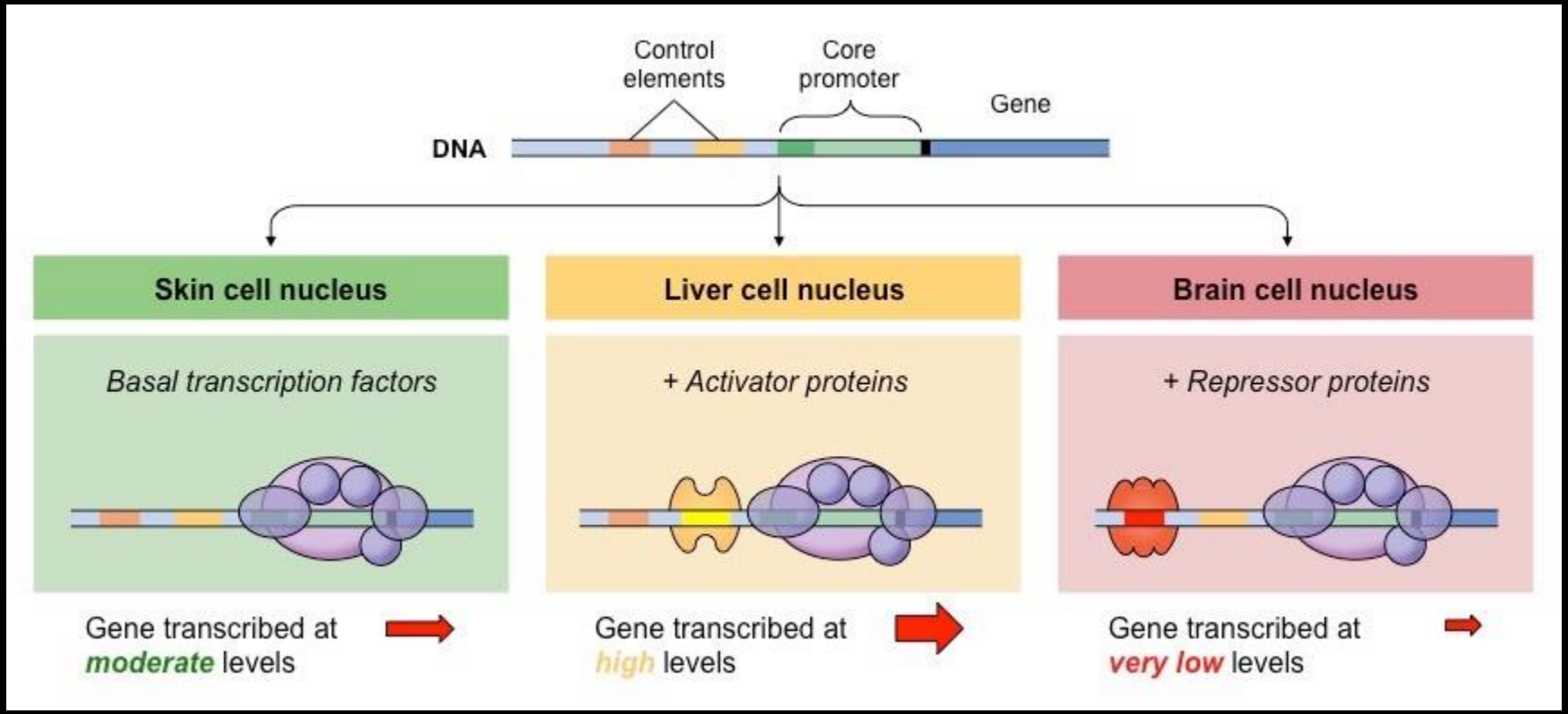


~ 100 proteins

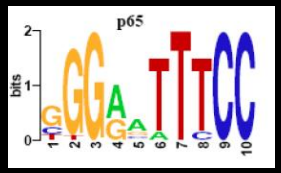
sebastien.aubourg@inra.fr

Transcription regulation

distal promoter proximal promoter introns, 3'



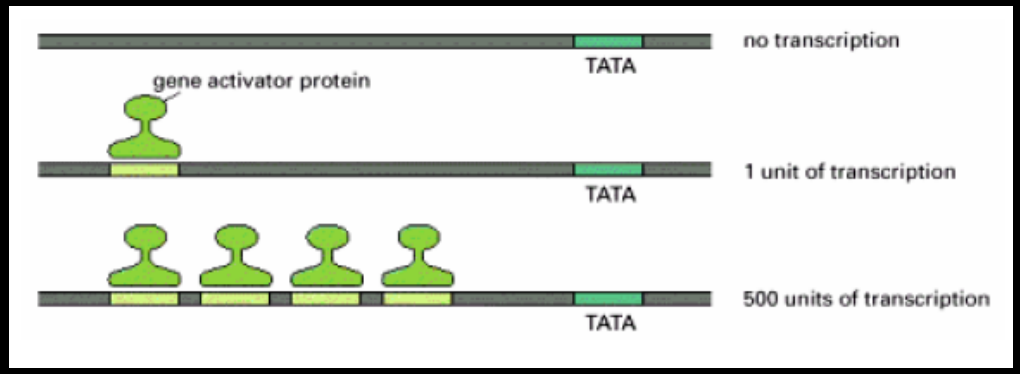
Each TF recognizes a target sequence, TFBS, such as :

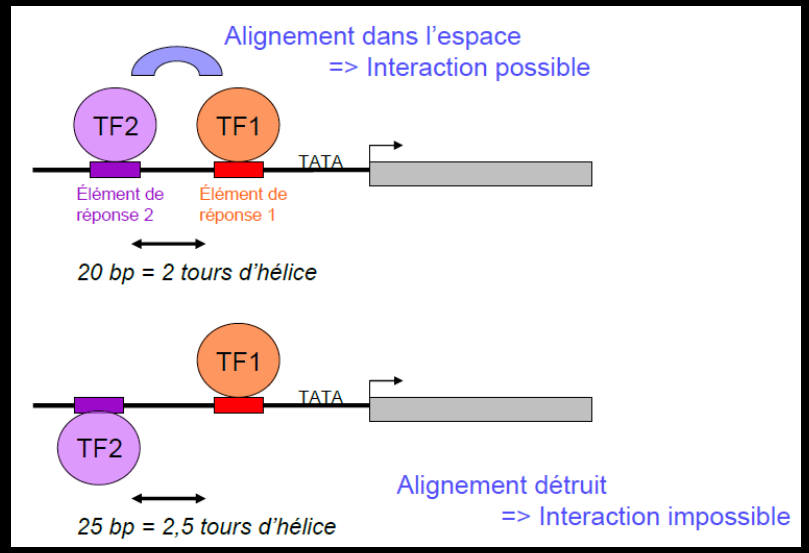


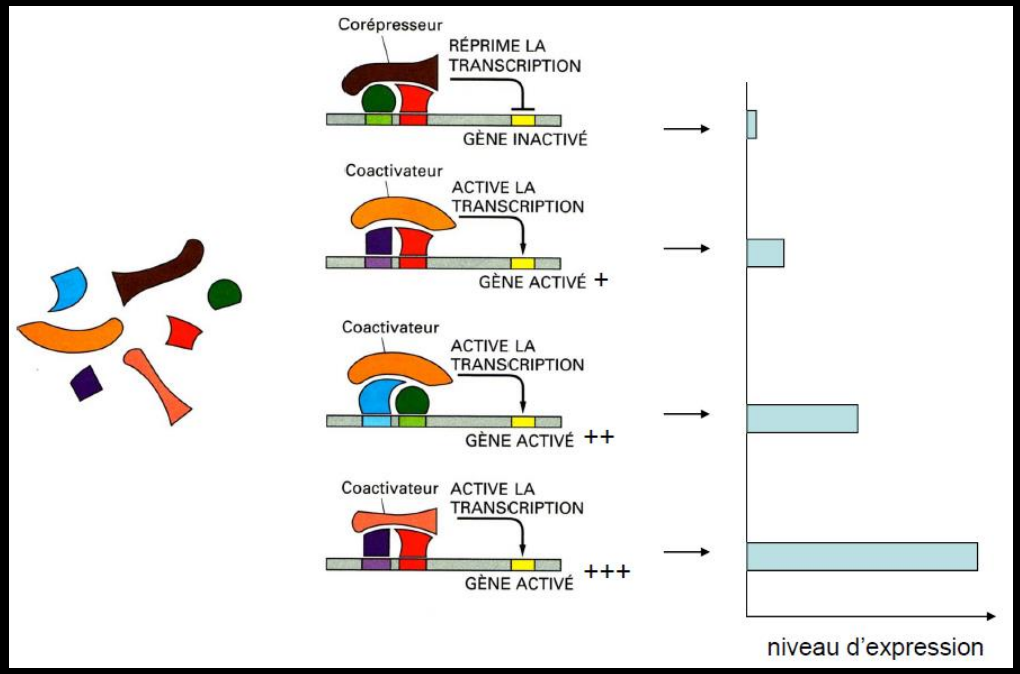
sebastien.aubourg@inra.fr

Transcription regulation

Cooperation/competition between TFs for regulation





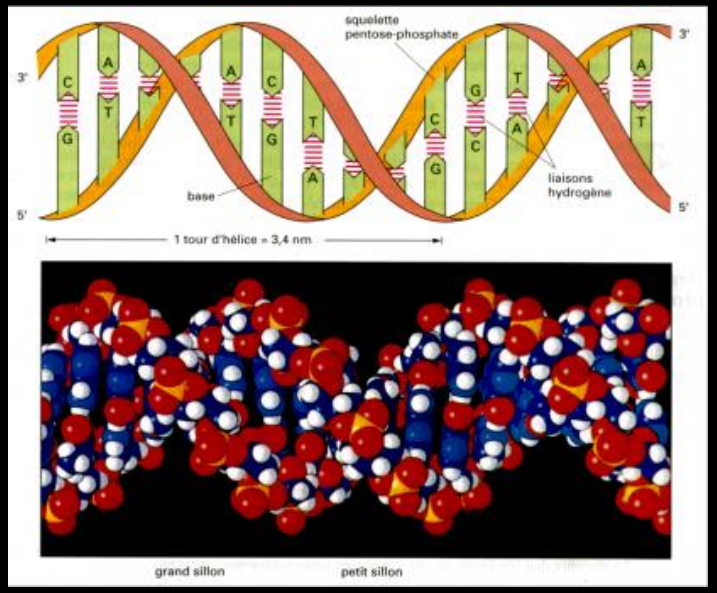


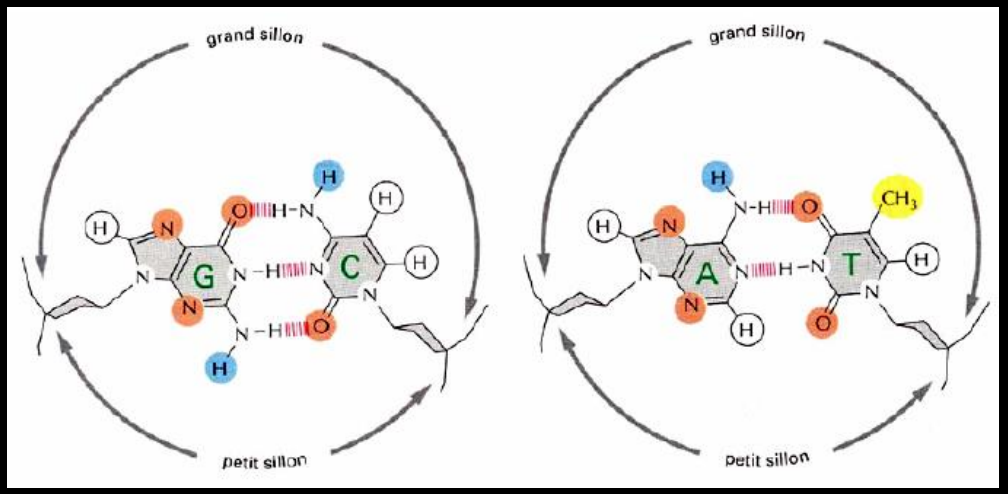
- Applications in biotechnologies for modulate/control gene expression

sebastien.aubourg@inra.fr

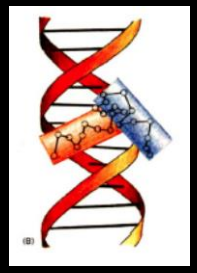
DNA-protein interactions

DNA-binding proteins mainly interact with the major groove of DNA

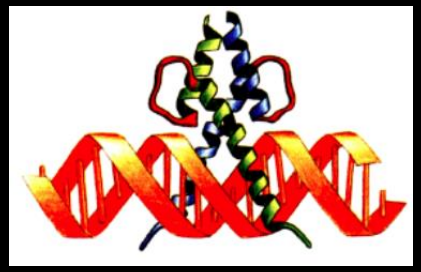




TFs are classified according their secondary structure which defined the interaction with double helix : basic-helix-loop-helix (leucine zipper…), zinc-fingers, helix-turn-helix, beta scaffold with minor groove contact…



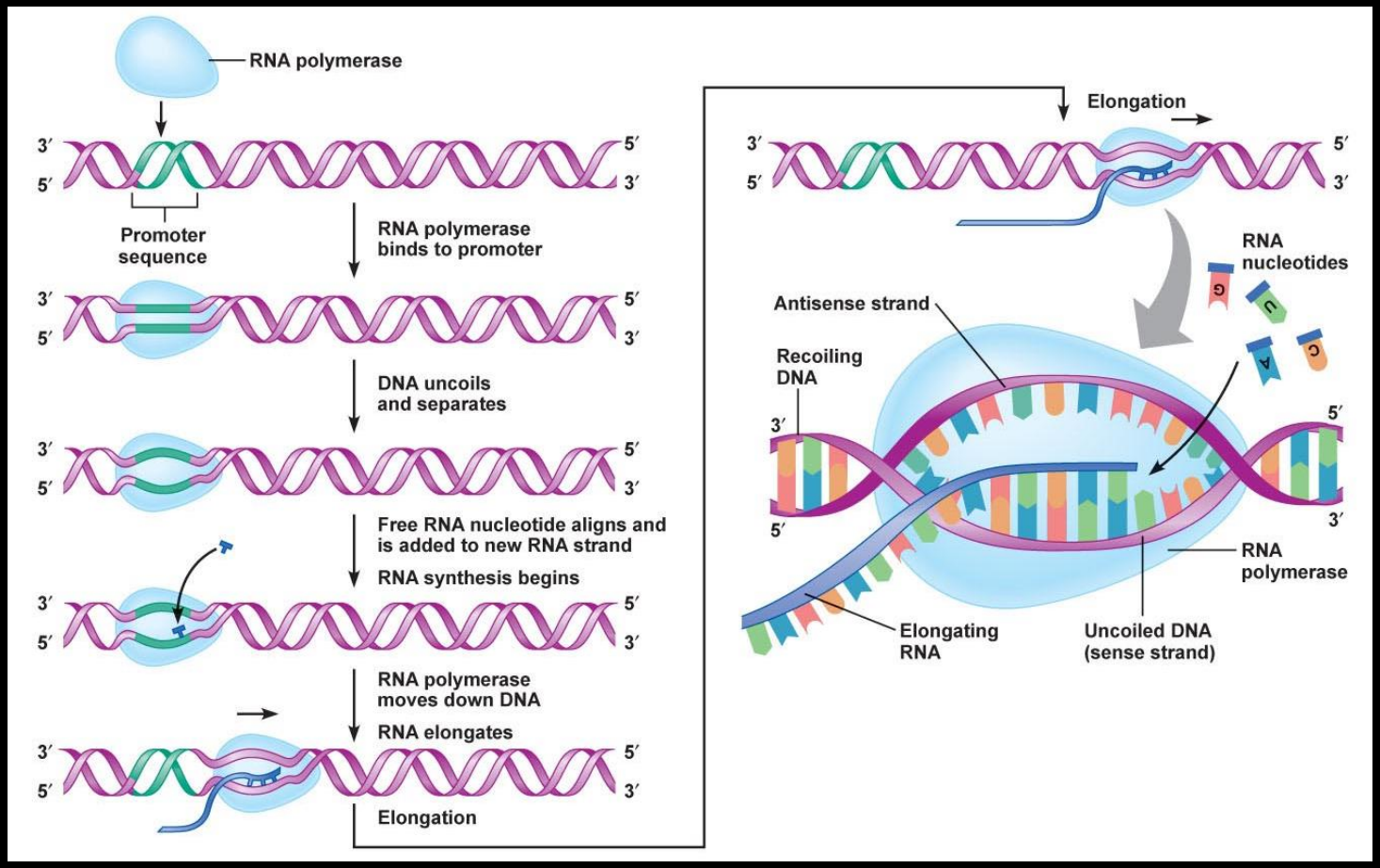
helix-turn-helix

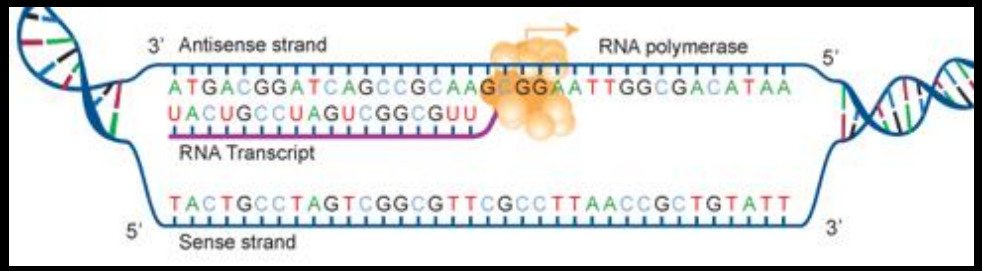


dimeric
helix-loop-helix

sebastien.aubourg@inra.fr

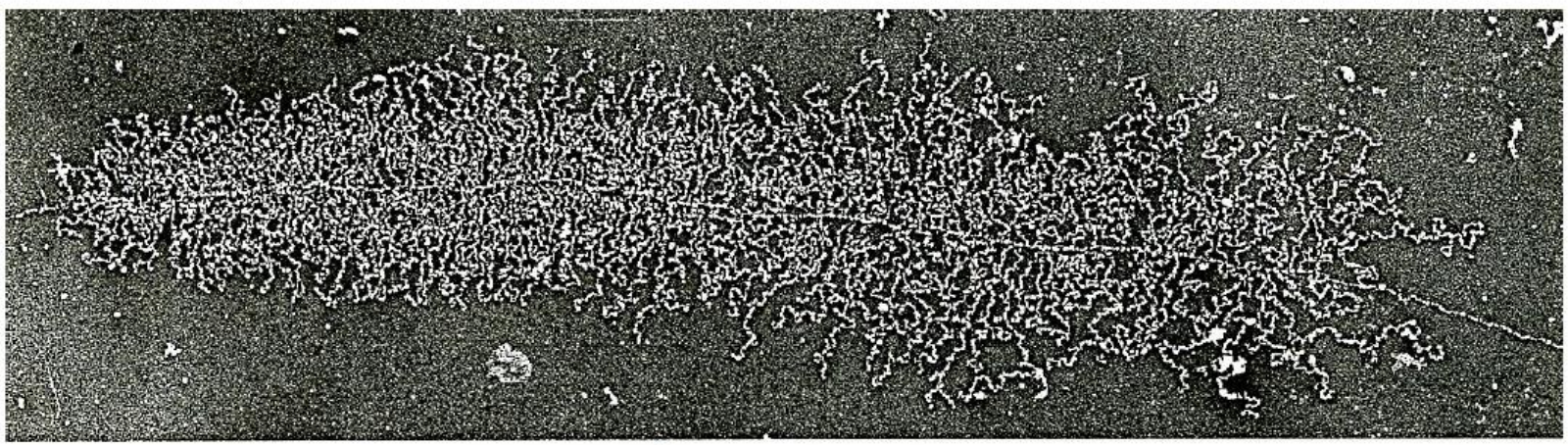
Transcription elongation





sebastien.aubourg@inra.fr

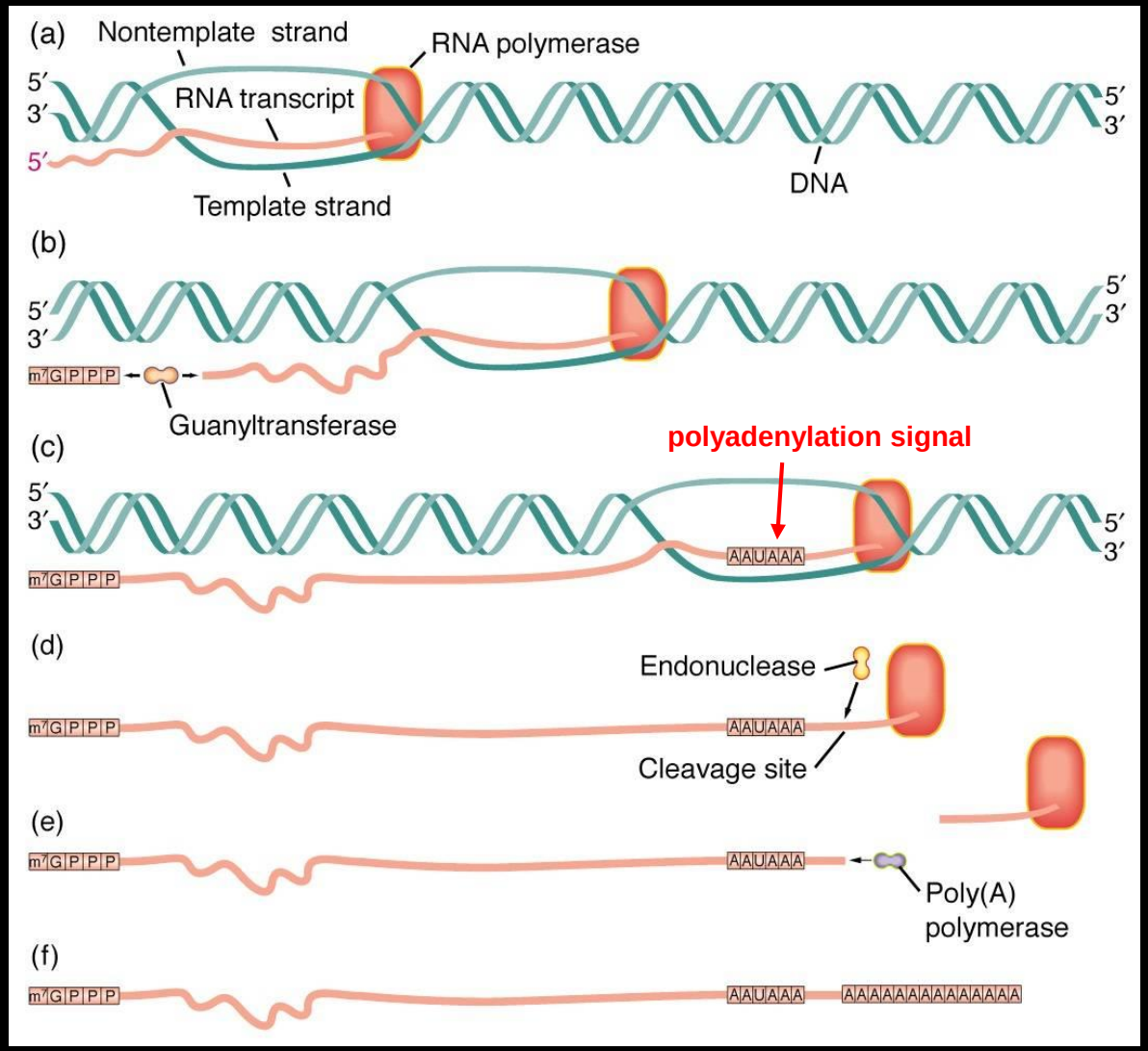
Transcription elongation



~ 40-80 nucleotides / sec

sebastien.aubourg@inra.fr

Transcription termination and maturation



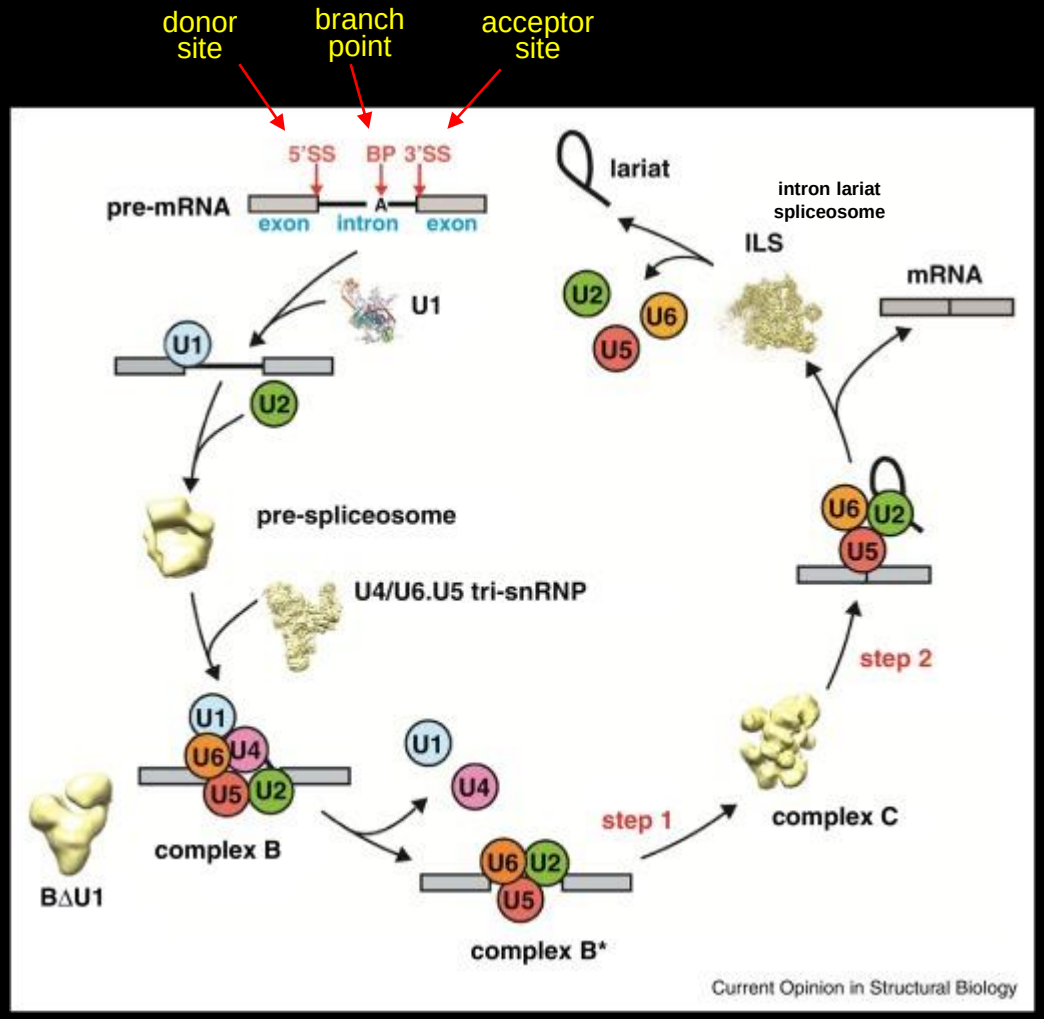
cap in 5′

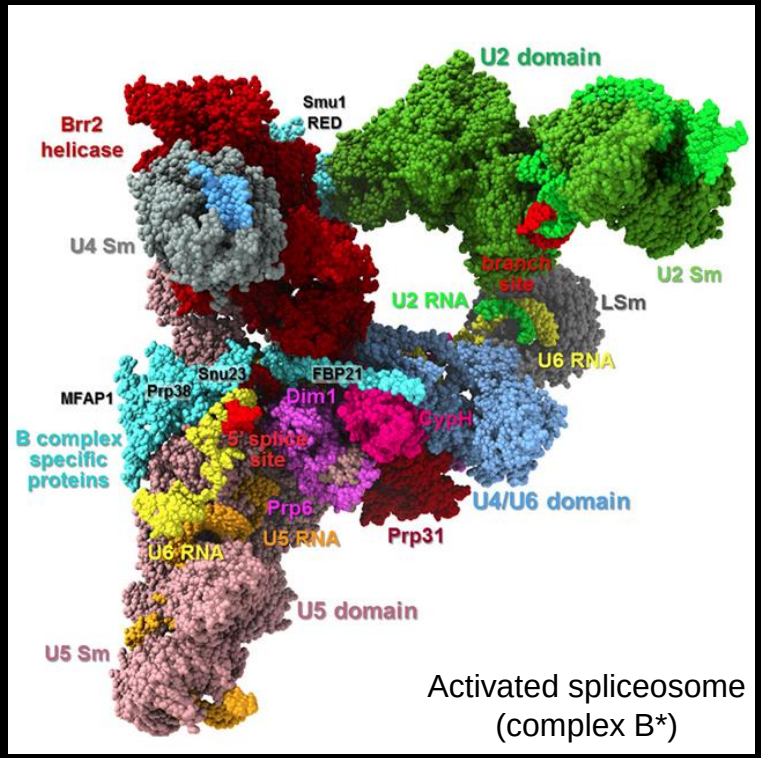
Poly(A) tail in 3′

sebastien.aubourg@inra.fr

Pre-mRNA splicing : intron removing

Small nuclear ribonucleic acids (snRNA, ~150 b) and proteins constitute small nuclear ribonucleoproteins (snRNP), components of spliceosome.

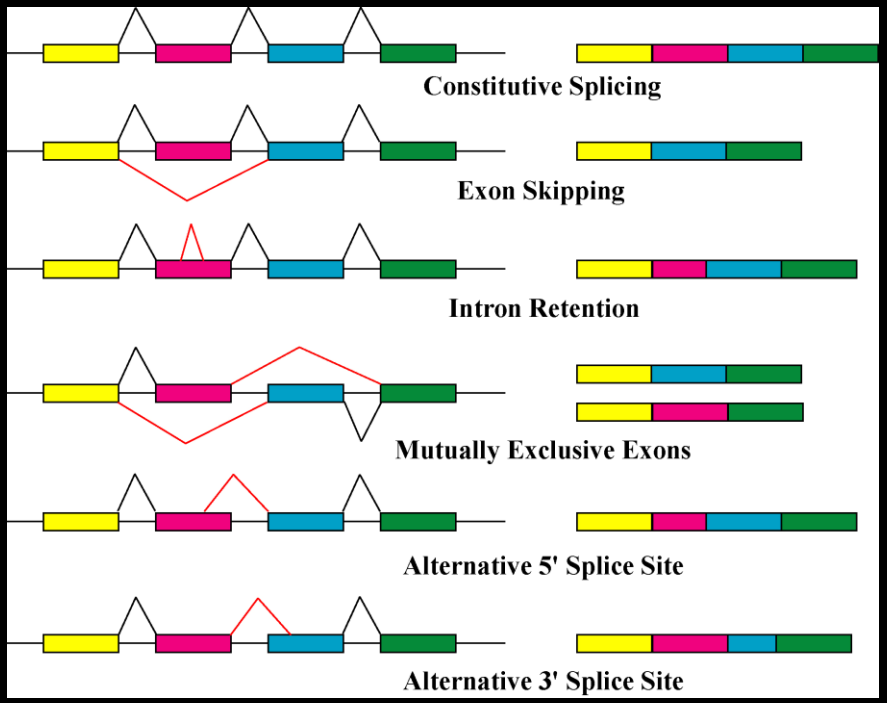




Activated spliceosome (complex B*)

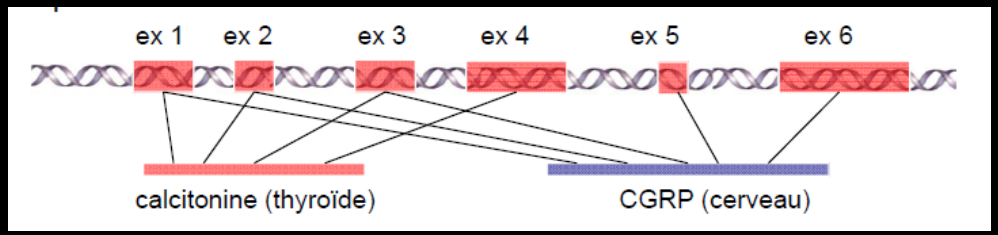
sebastien.aubourg@inra.fr

Post-transcriptional regulation: alternative splicing



One pre-mRNA can produce a large diversity of mature RNA

Exemple 1 :



Calcitonin Gene-Related Peptide

2 peptides
2 functions

Hormone
reduces Ca^{2+} in blood

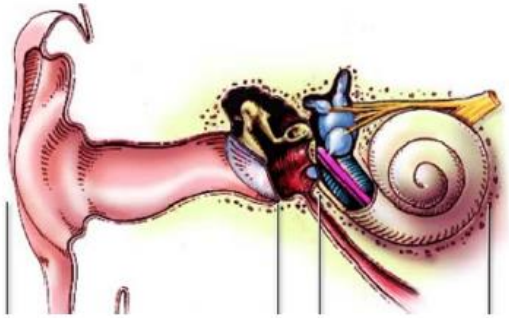
Vasodilator
nociception

sebastien.aubourg@inra.fr

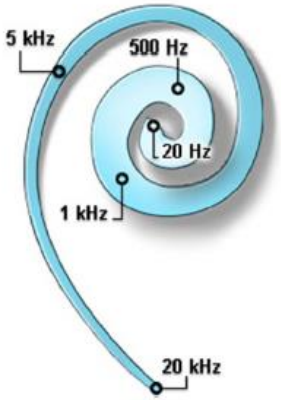
Post-transcriptional regulation: alternative splicing

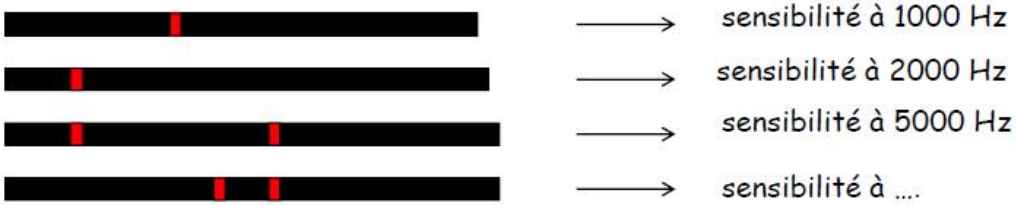
Exemple 2 :

Gene slo : 35 exons including 8 optional ones
500 different transcripts ► 500 proteins SLO



cochlée





Le gène *SLO* code pour une protéine membranaire régulant l'entrée et la sortie de potassium dans les cellules pileuses de la cochlée qui répondent aux différentes fréquences (Hz) en fonction de la nature de la nature de la protéine SLO. La diversité des protéines SLO obtenues grâce à l'épissage alternatif détermine l'étendue de l'audition chez l'homme

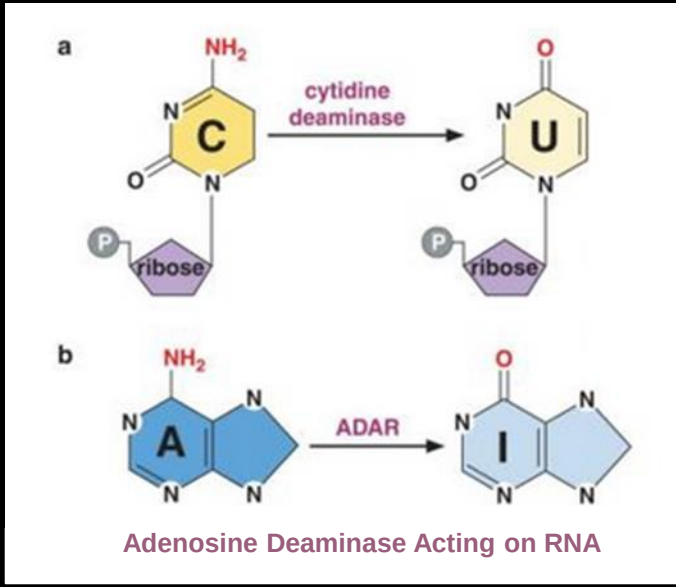
sebastien.aubourg@inra.fr

Post-transcriptional regulation: RNA editing

RNA editing is a dynamic mechanism that generates site-specific deamination. The most common type of RNA editing involves nucleotide substitution that consists of :

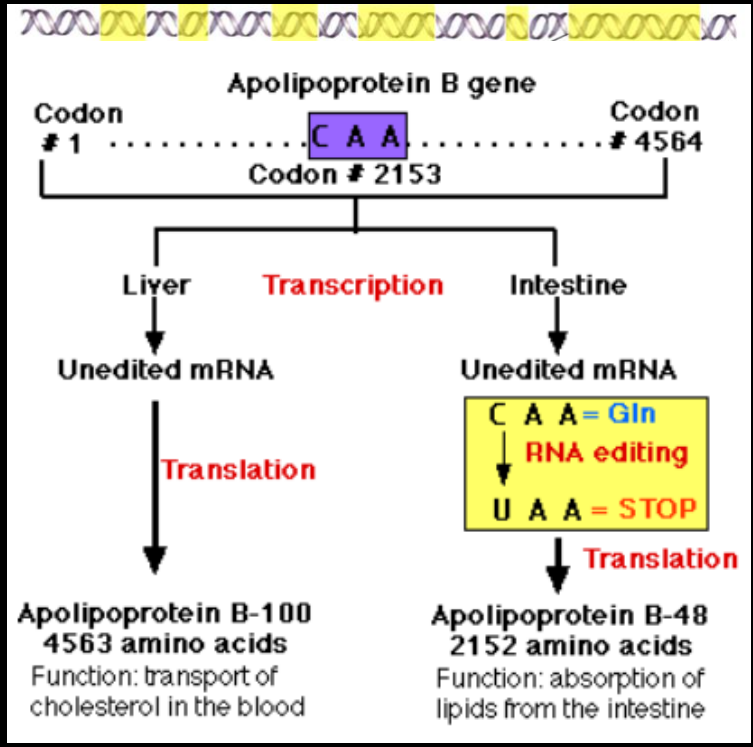
Adenosine A ⟶ inosine I (mainly in human), I is read as G by the translation machinery of the cell.

cytidine C ⟶ uridine U (mainly in plants).



Adenosine Deaminase Acting on RNA

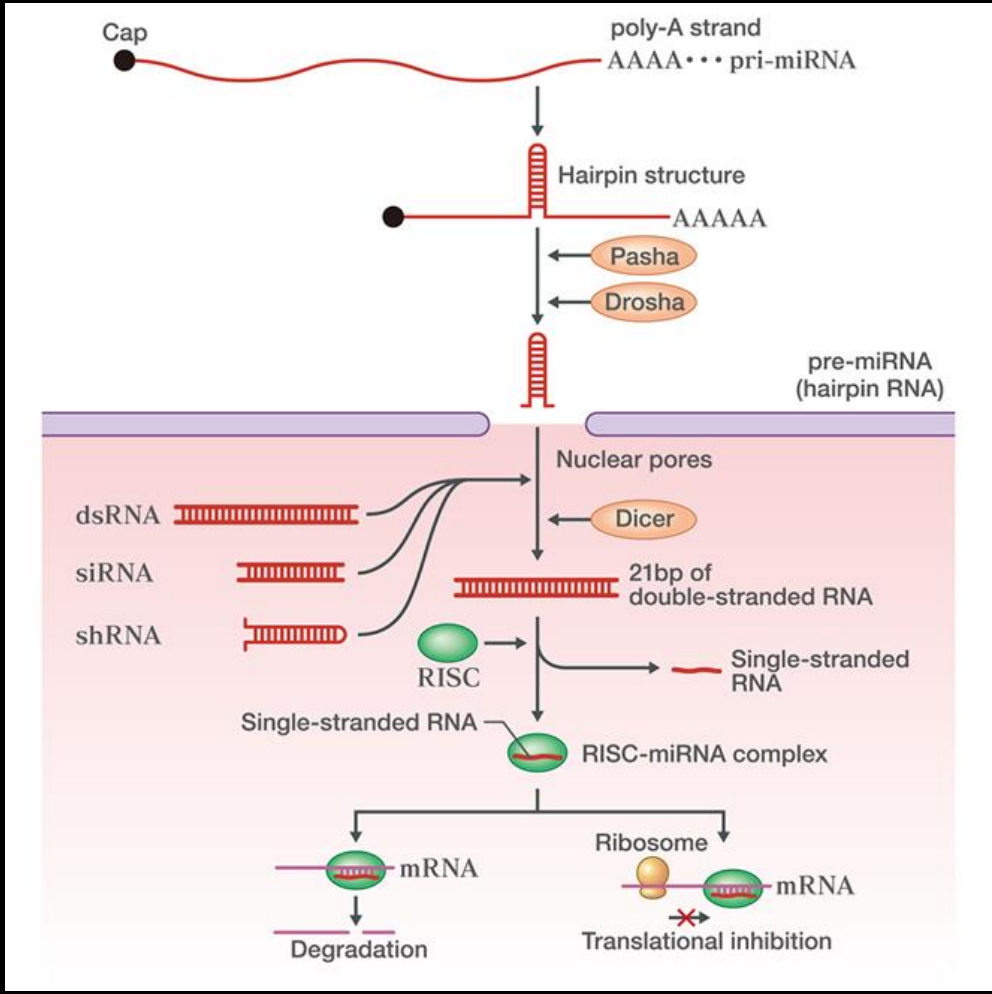
Exemple



sebastien.aubourg@inra.fr

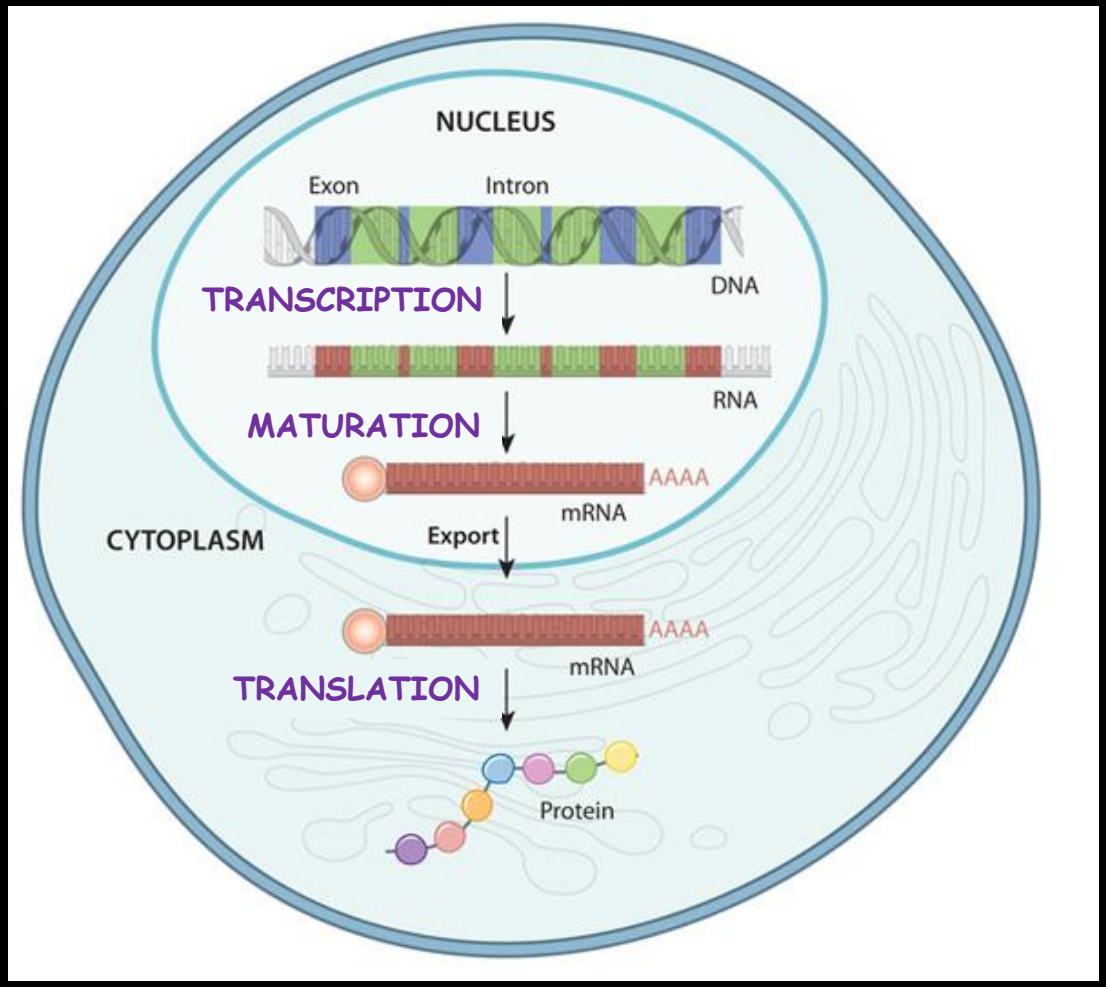
Post-transcriptional regulation: miRNA

MicroRNAs (miRNAs) are a class of small non-coding RNAs that post-transcriptionally control gene expression via either translational repression or mRNA degradation.



sebastien.aubourg@inra.fr

Translation



Actors
mature mRNA
Ribosomes
tRNAs

Place
Cytoplasm
/
ER associated

sebastien.aubourg@inra.fr

Ribosome : translational machinery

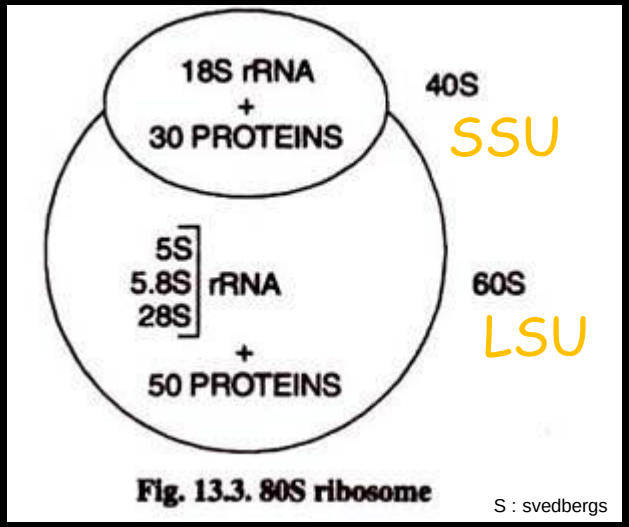
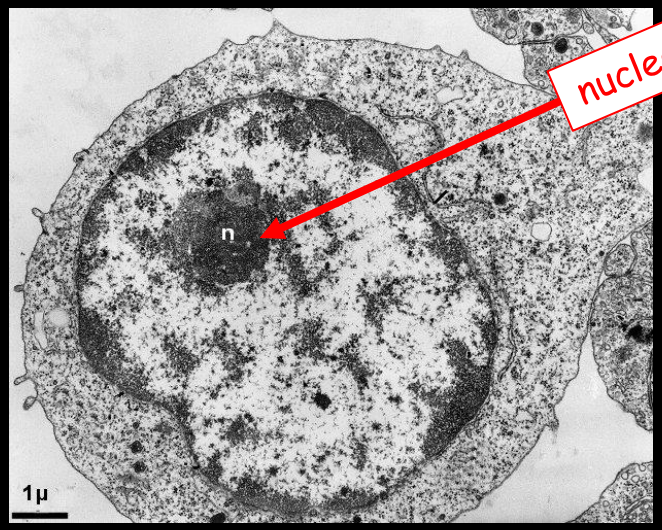
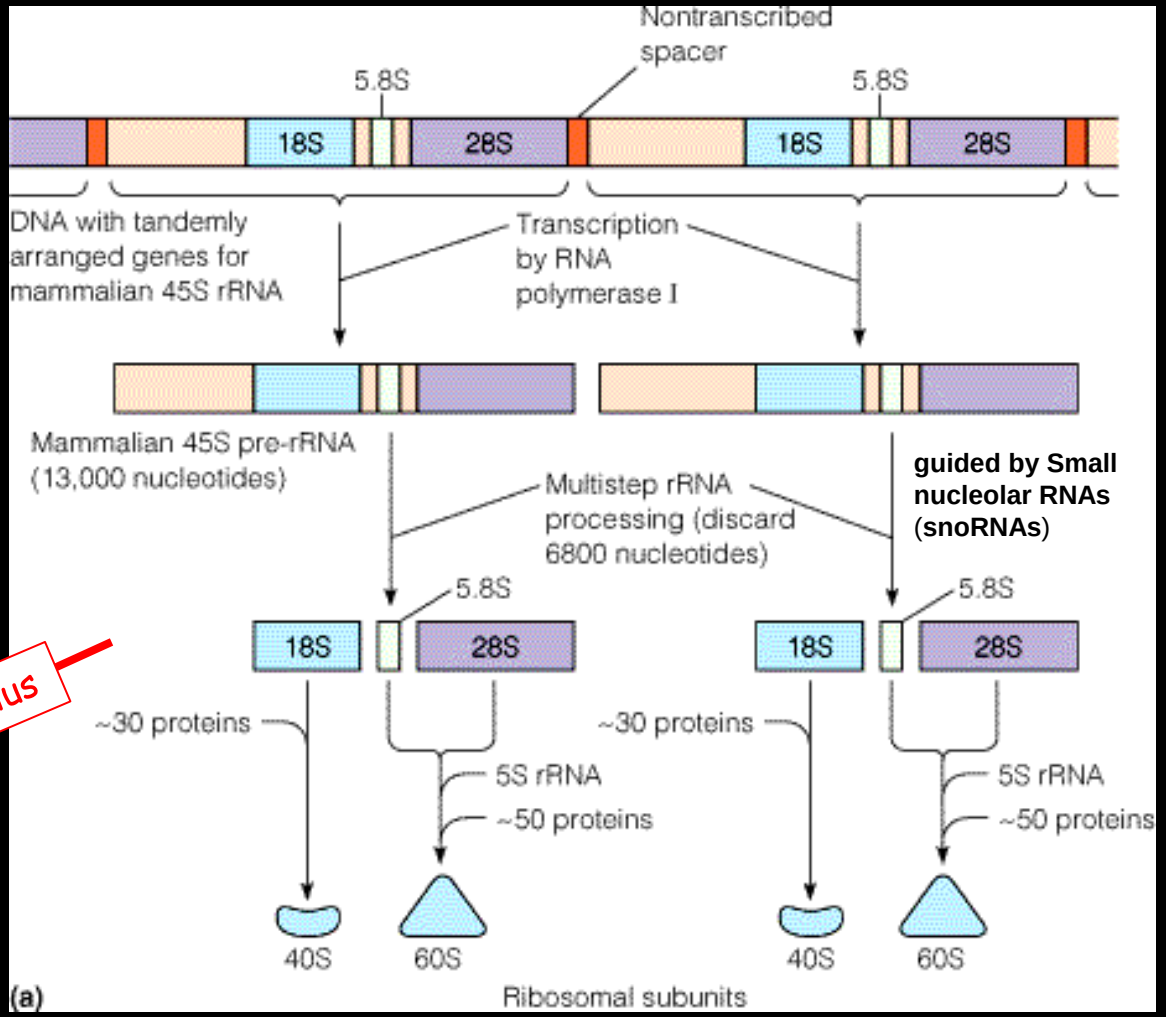


Fig. 13.3. 80S ribosome

S : svedbergs



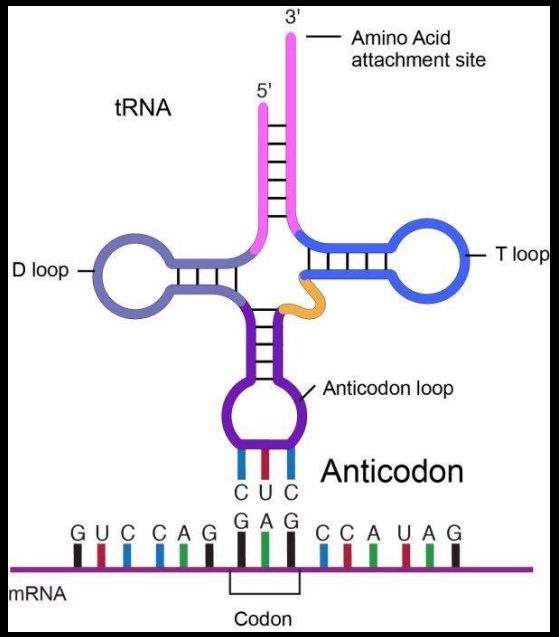
nucleolus

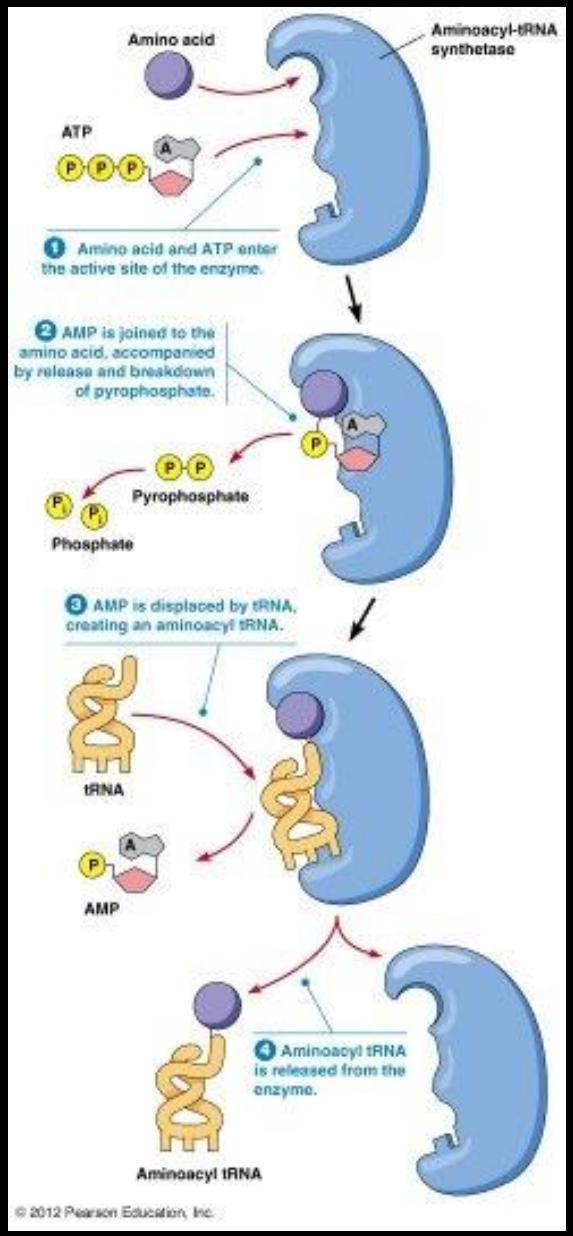


rRNA : ~90% of the RNA in a cell

sebastien.aubourg@inra.fr

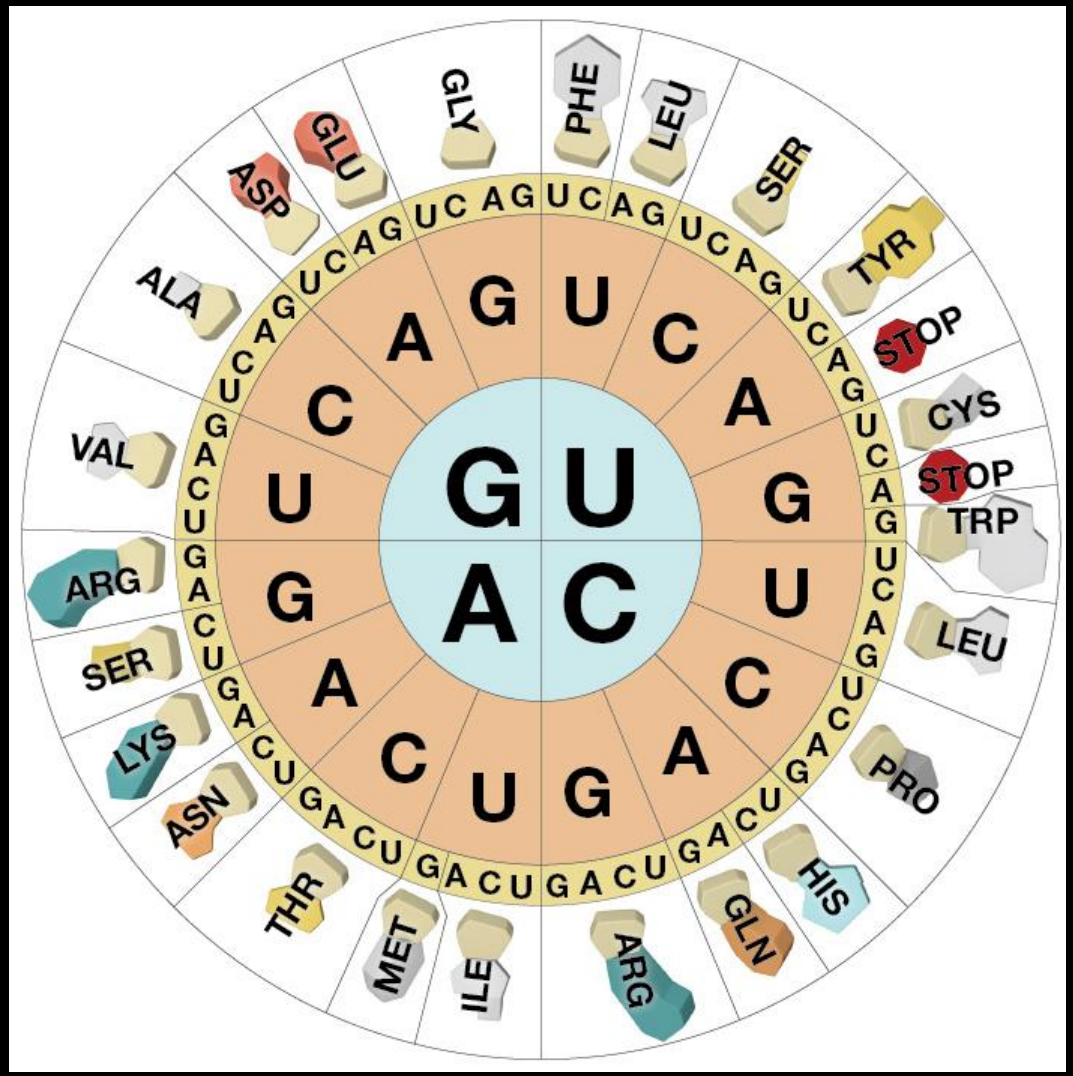
tRNA : links between nucleotides and amino-acids





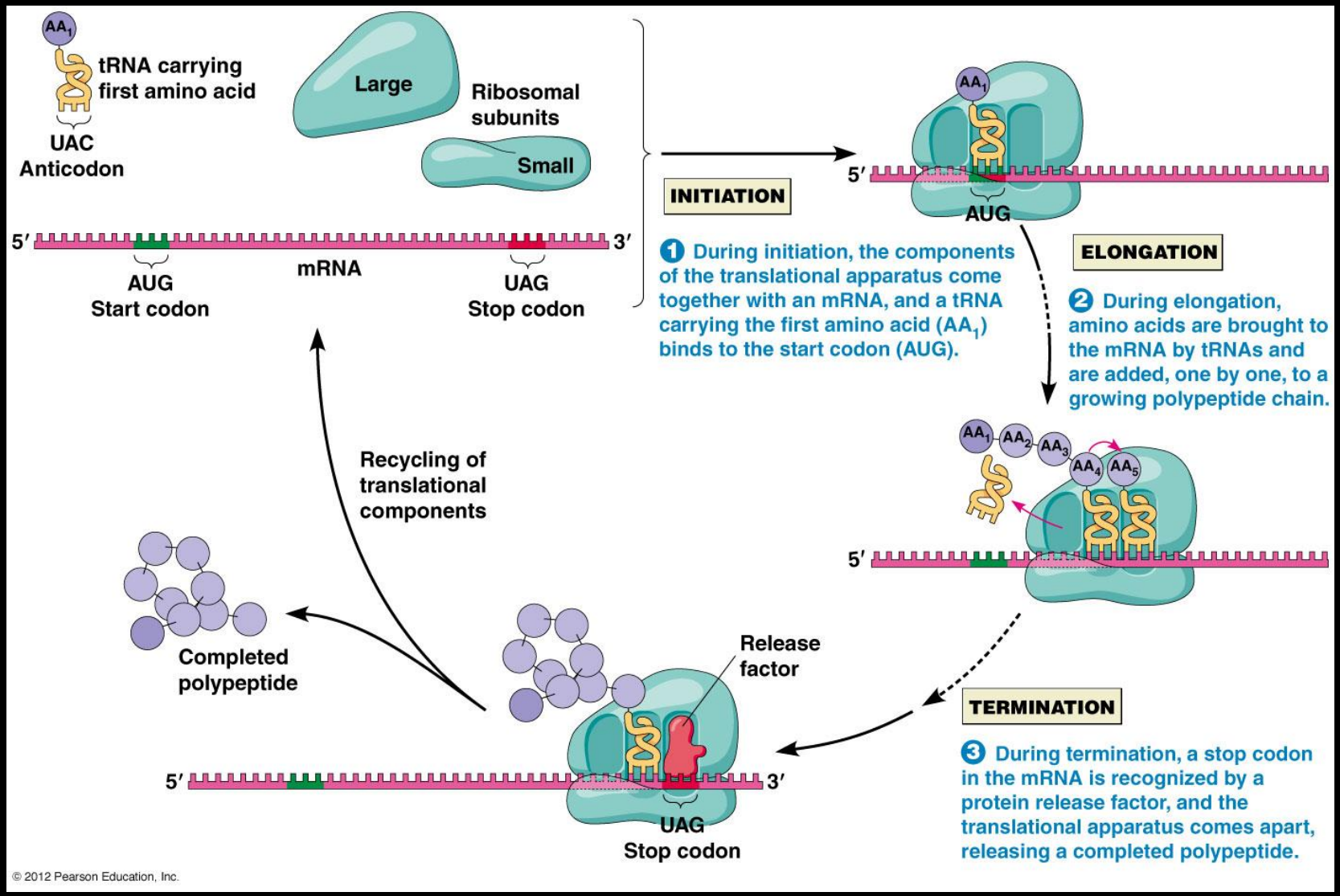
Stem loop structure
One tRNA per codon

The genetic code



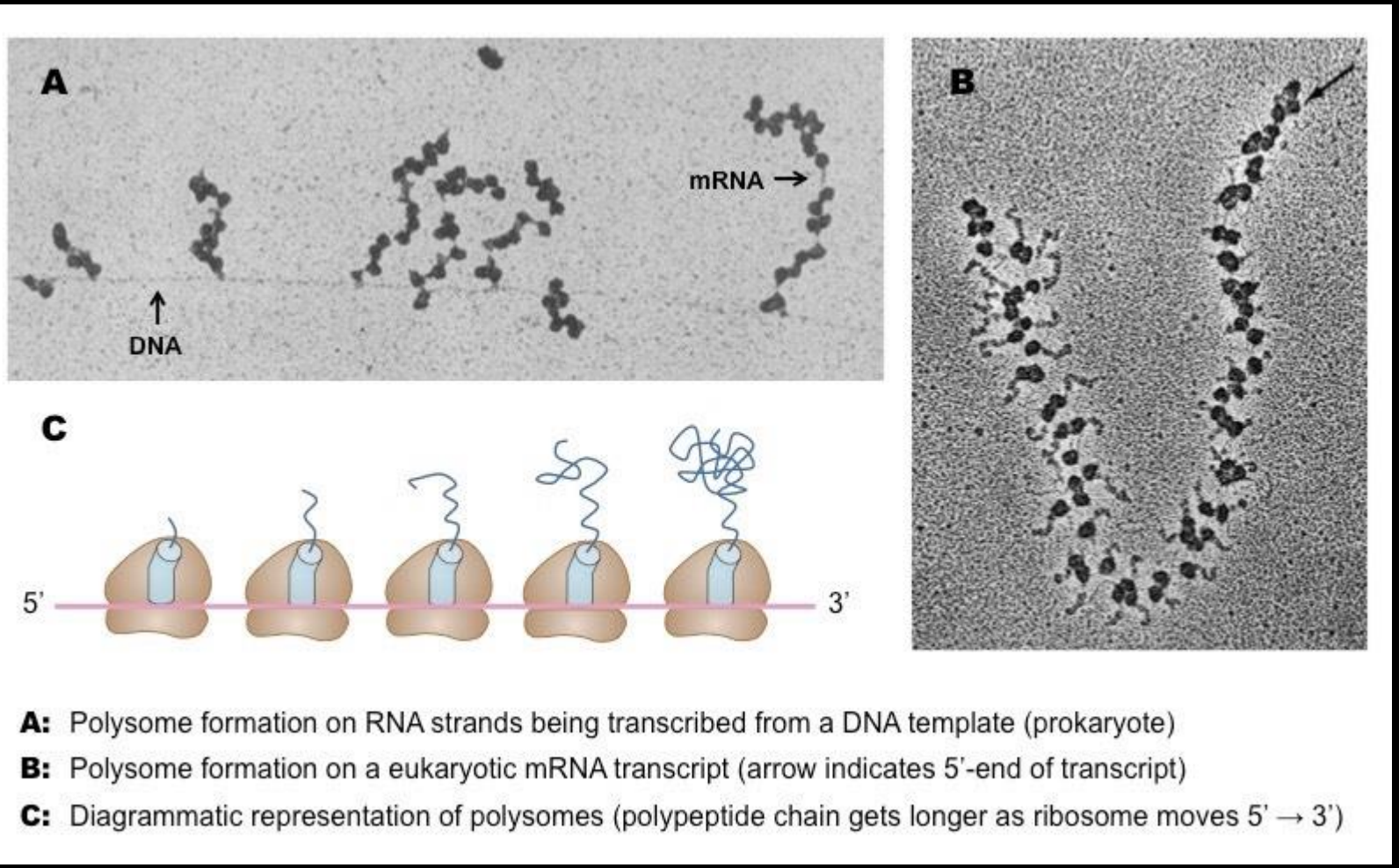
sebastien.aubourg@inra.fr

Translation



sebastien.aubourg@inra.fr

Polysomes



A: Polysome formation on RNA strands being transcribed from a DNA template (prokaryote)

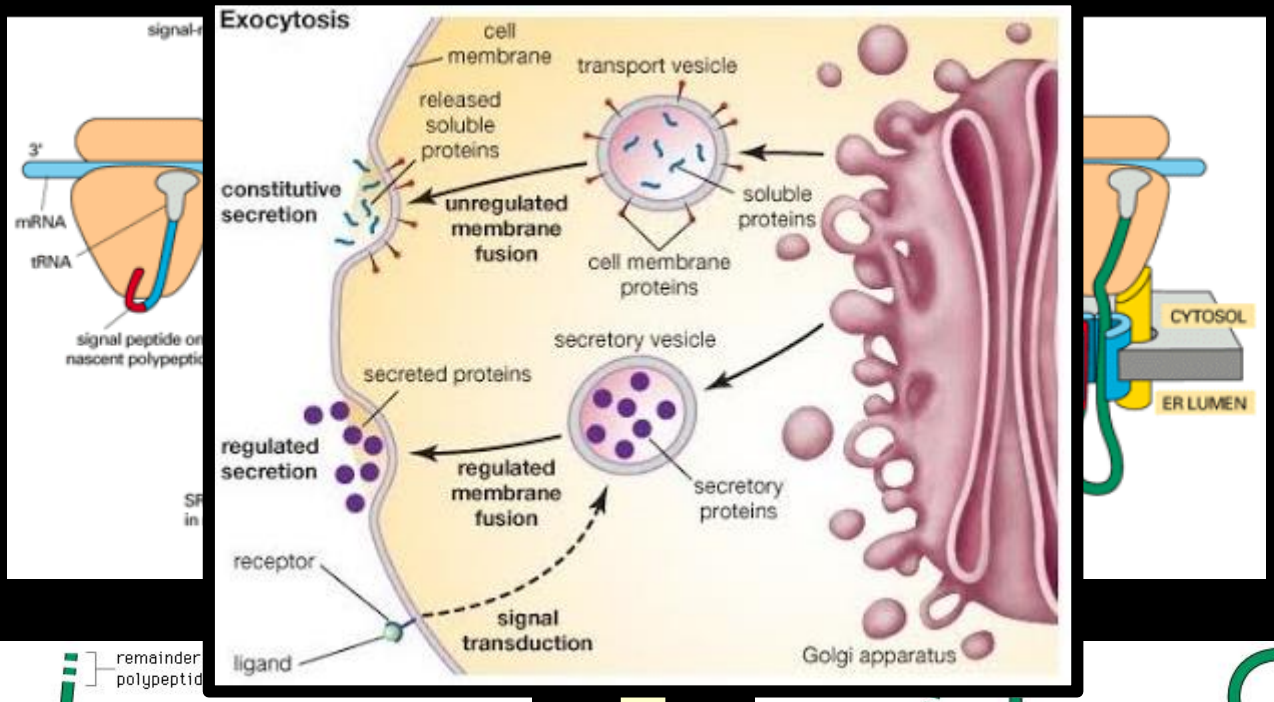
B: Polysome formation on a eukaryotic mRNA transcript (arrow indicates 5'-end of transcript)

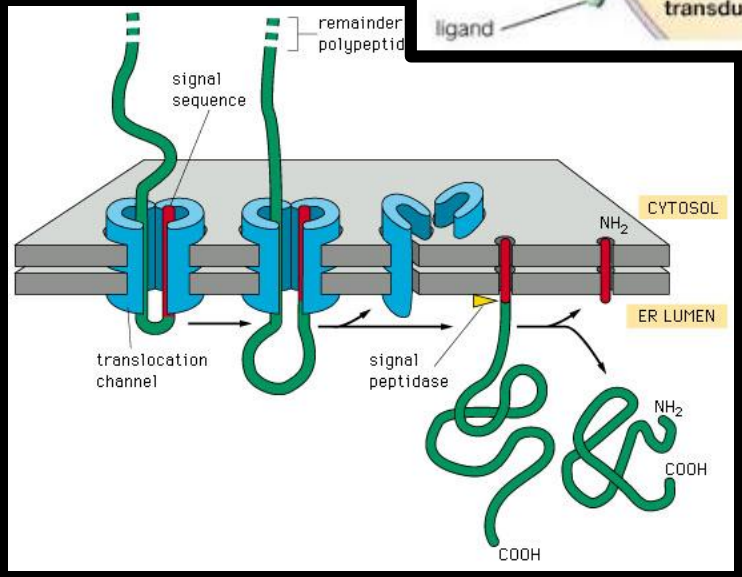
C: Diagrammatic representation of polysomes (polypeptide chain gets longer as ribosome moves 5' → 3')

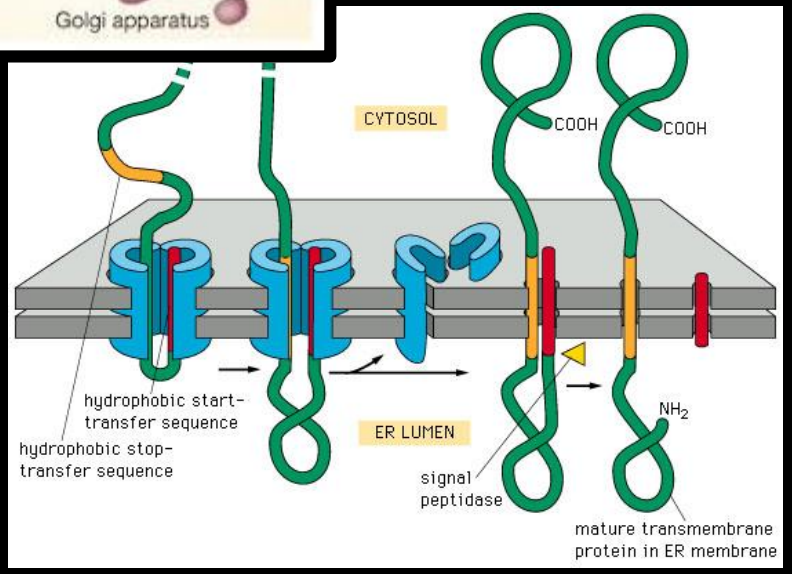
~ 5 aa (eukaryote) to 50 aa (prokaryote) / sec

sebastien.aubourg@inra.fr

Proteins targeted to the endoplasmic reticulum

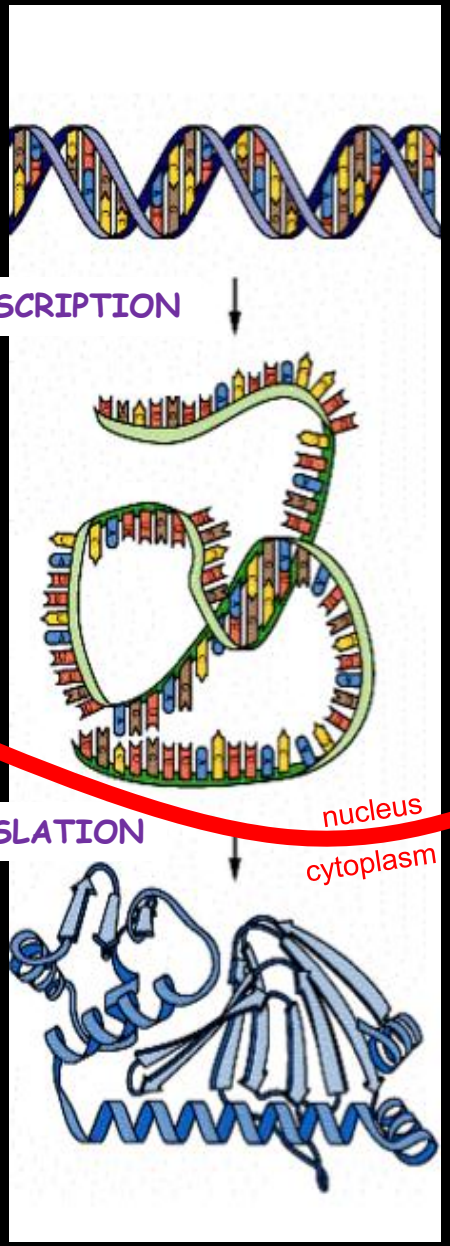






sebastien.aubourg@inra.fr

Levels of regulation of gene expression



Transcriptional control
DNA access
chromatin structure
DNA methylation
Transcription factors binding
Initiation, elongation, termination

Post-transcriptional control
Maturation
Splicing
Editing
Stabilization (miRNA control, polyA)
Export

Translational control
Initiation, elongation, termination
Codon usage

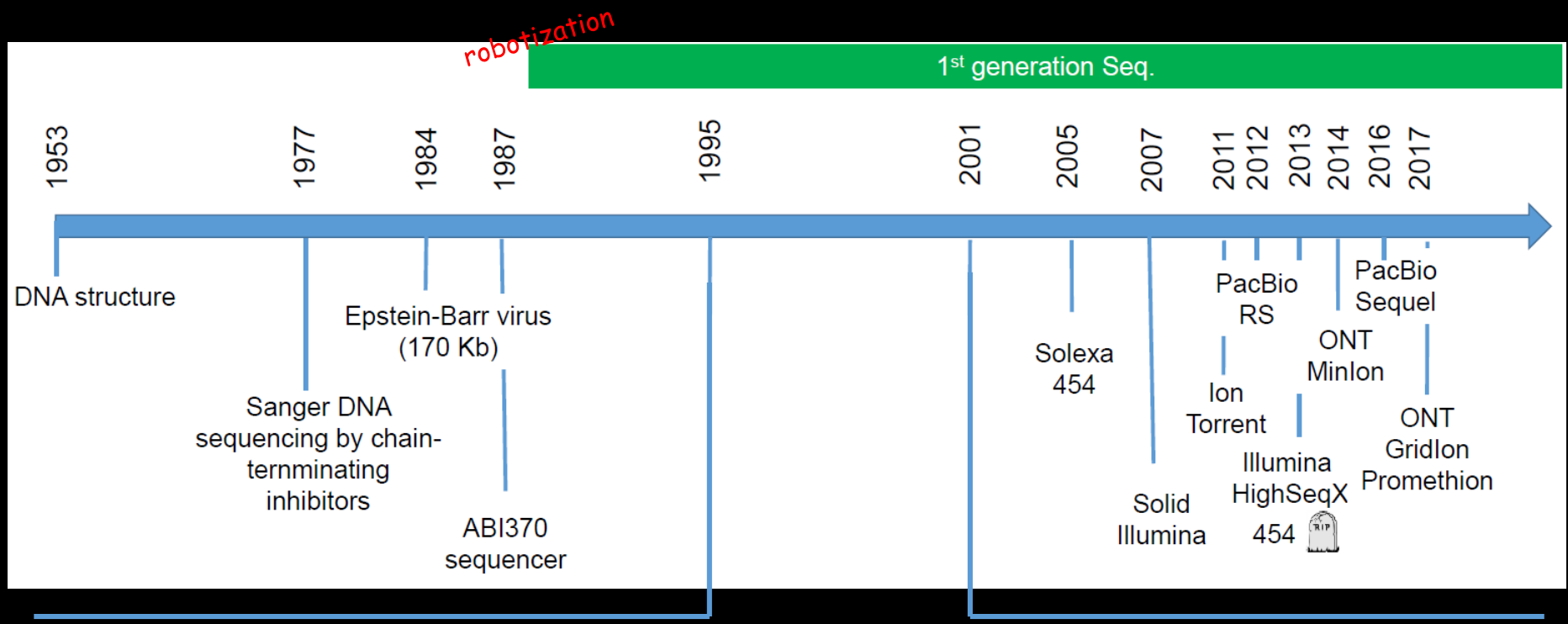
Post-translational control
Targeting
Cleavage
Modifications (glycosylation, acetylation…)

sebastien.aubourg@inra.fr

Sequencing and Sequences

sebastien.aubourg@inra.fr

History



H. influenzae
1,8 Mb
(1995)

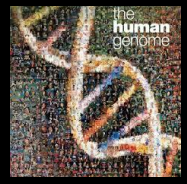
E. coli
4,6 Mb
(1997)

S. cerevisiae
6 Mb
(1997)

C. elegans
98 Mb
(1998)

A. thaliana
130 Mb
(2000)

D. melanogaster
165 Mb
(2000)

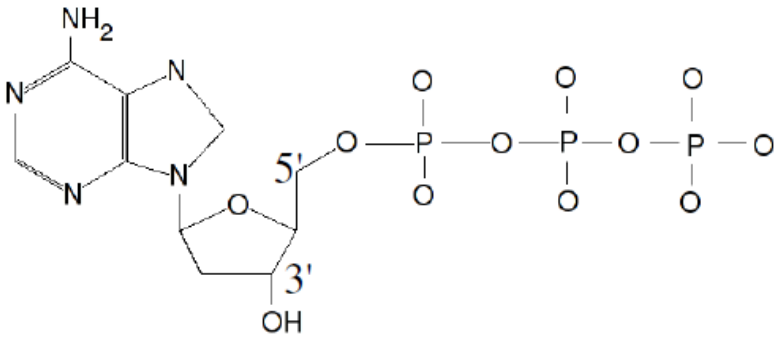


H. sapiens
3 Gb
(2001-2004)

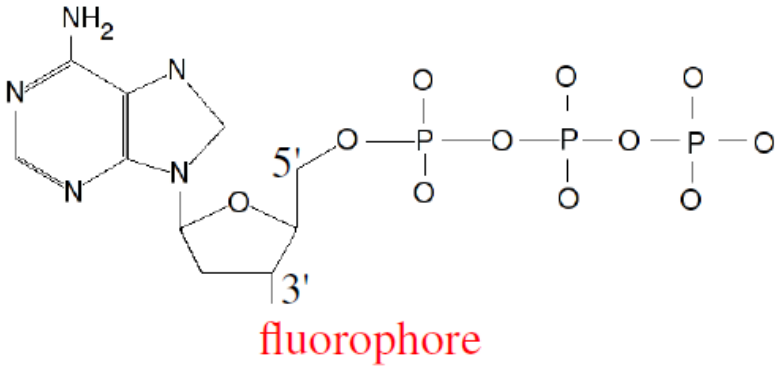
sebastien.aubourg@inra.fr

Sanger DNA sequencing

Sanger sequencing is based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication



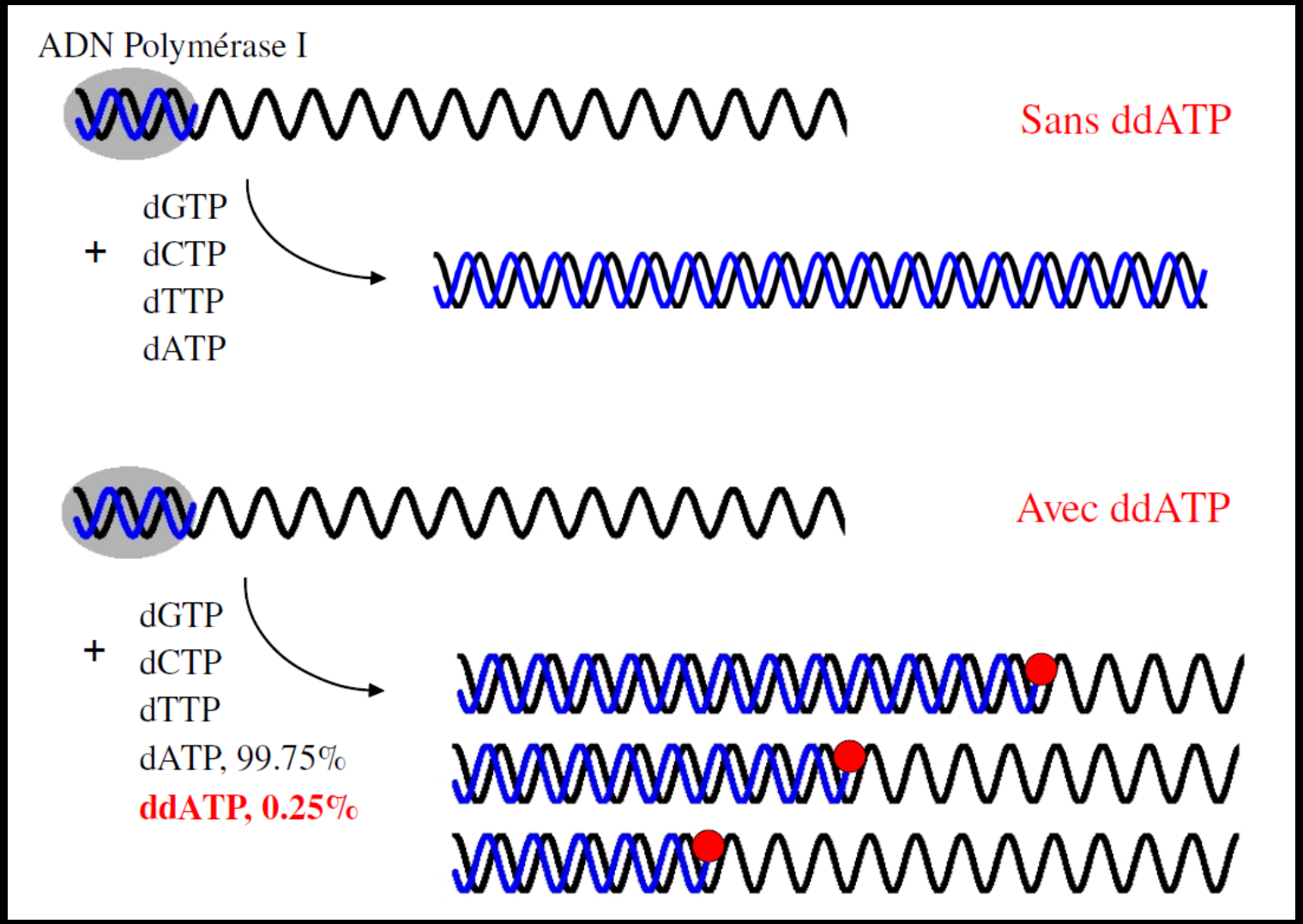
dATP



ddATP

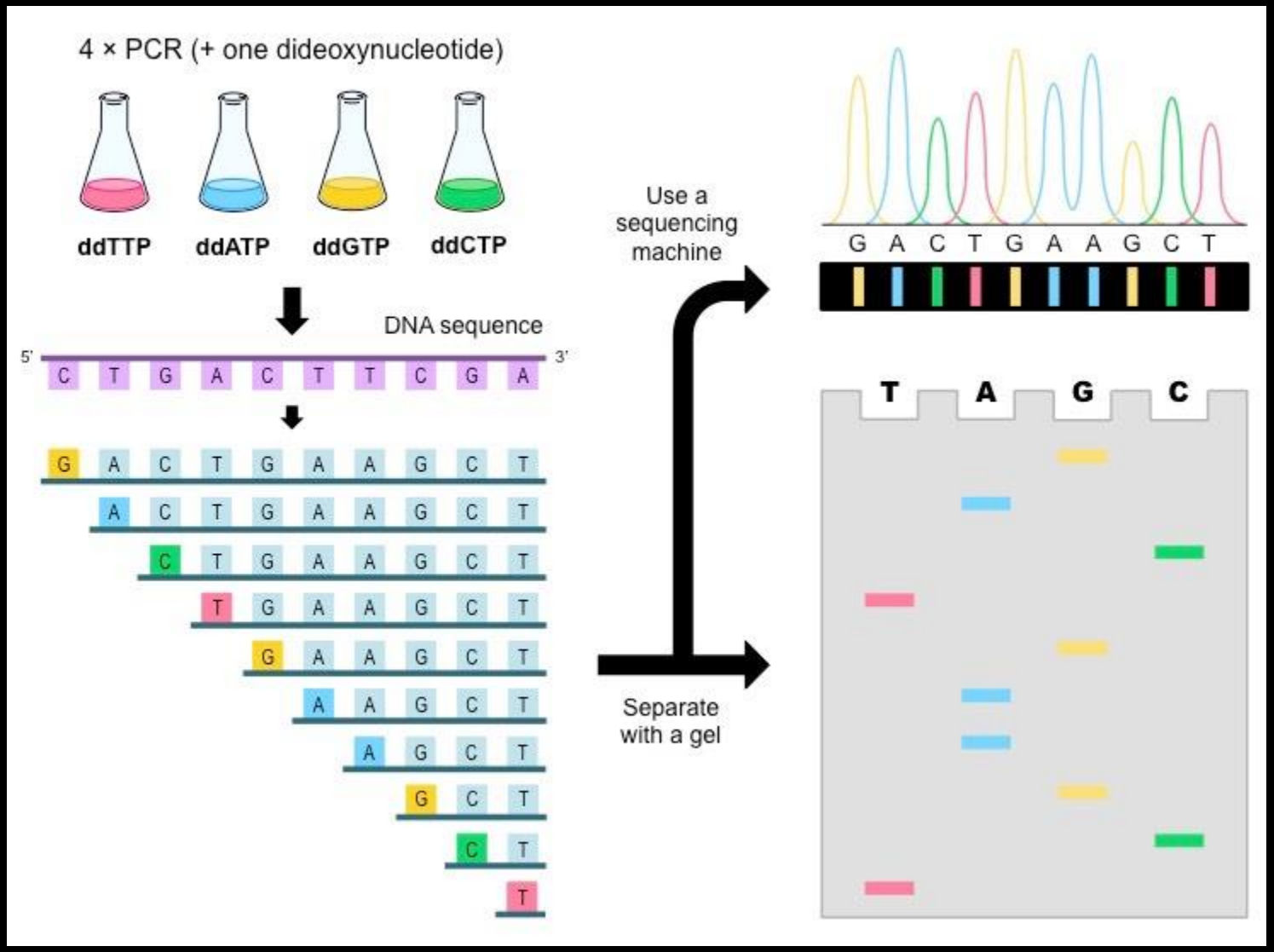
sebastien.aubourg@inra.fr

Sanger DNA sequencing



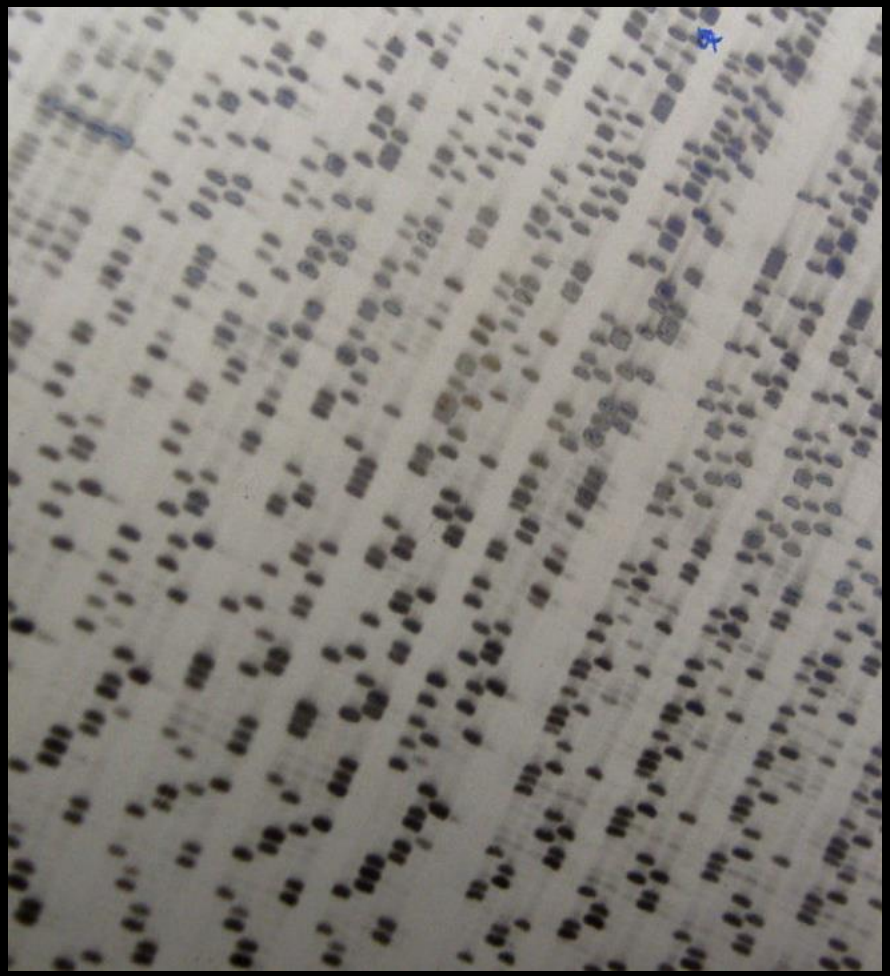
sebastien.aubourg@inra.fr

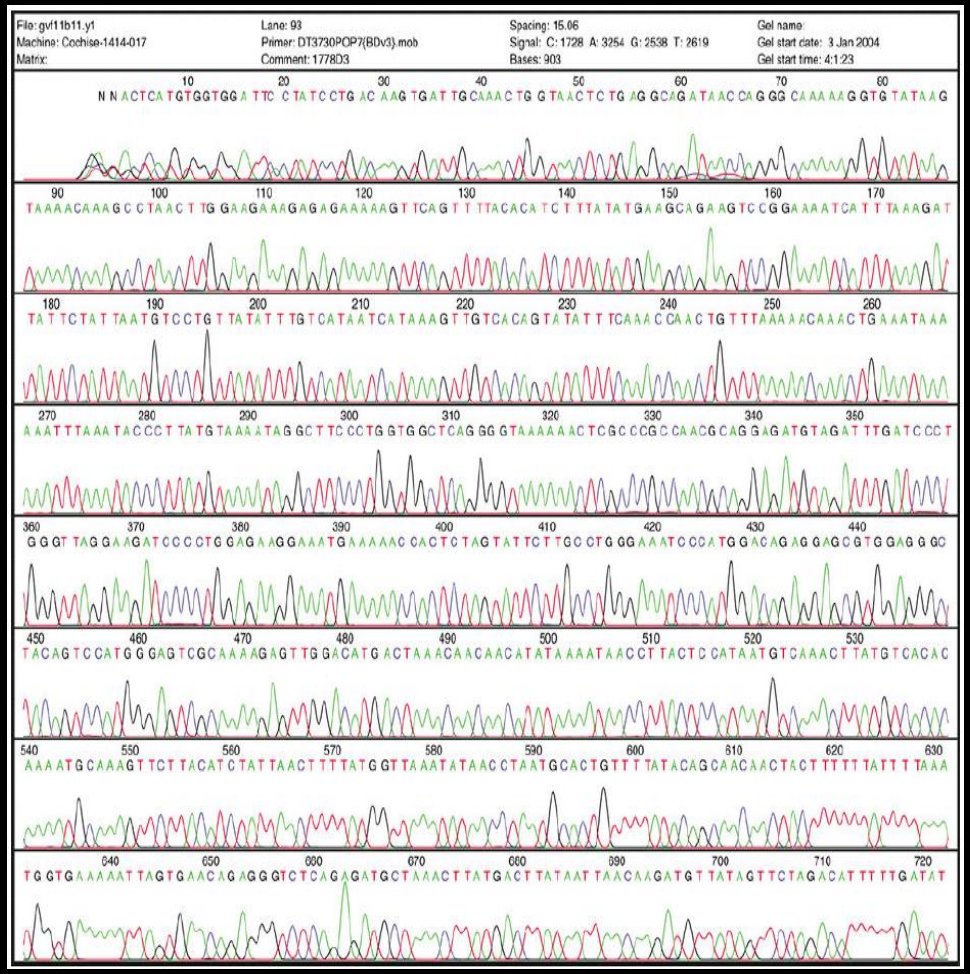
Sanger DNA sequencing



sebastien.aubourg@inra.fr

Sanger DNA sequencing

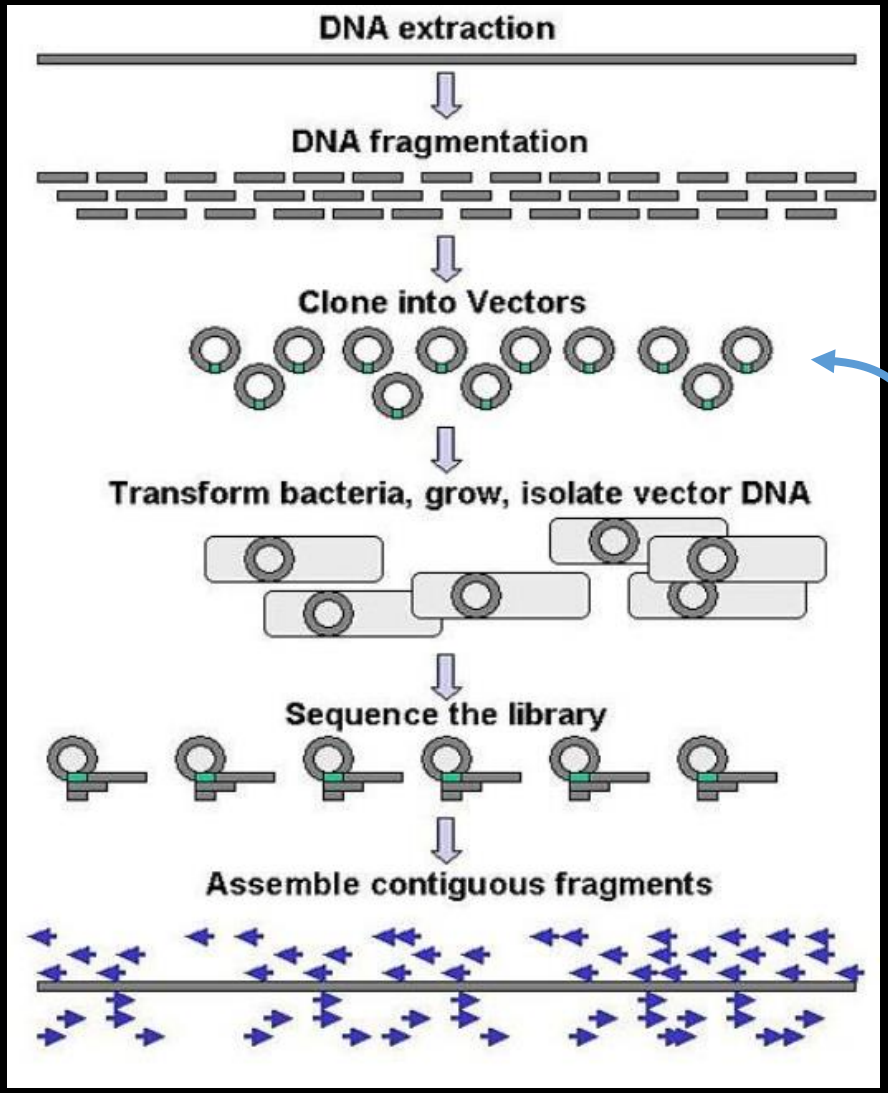




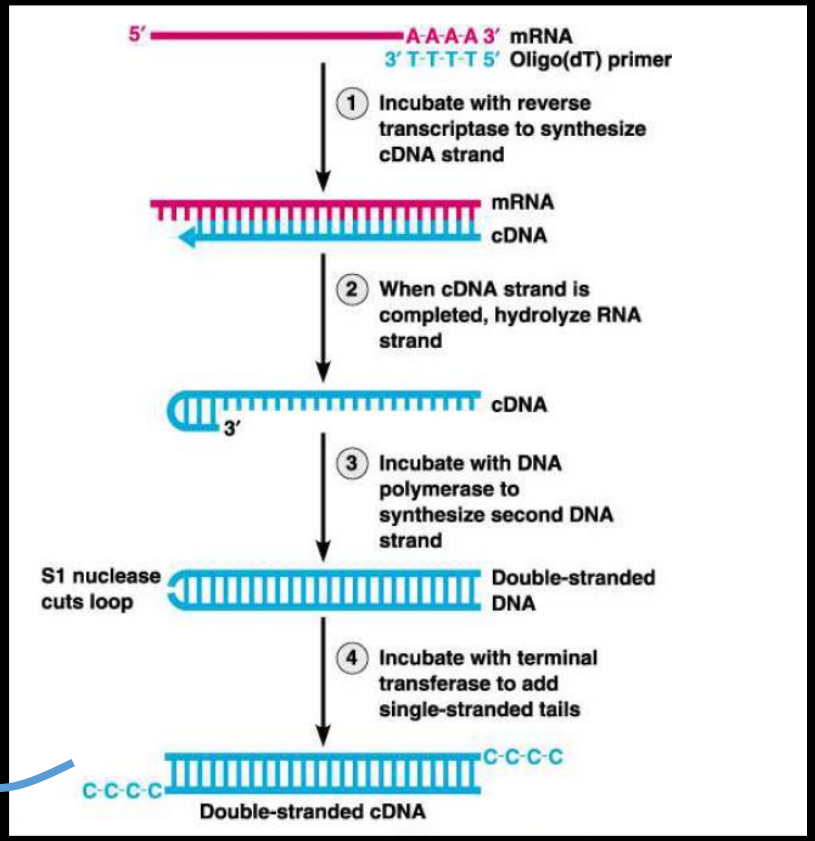
sebastien.aubourg@inra.fr

A limiting cloning step

DNA



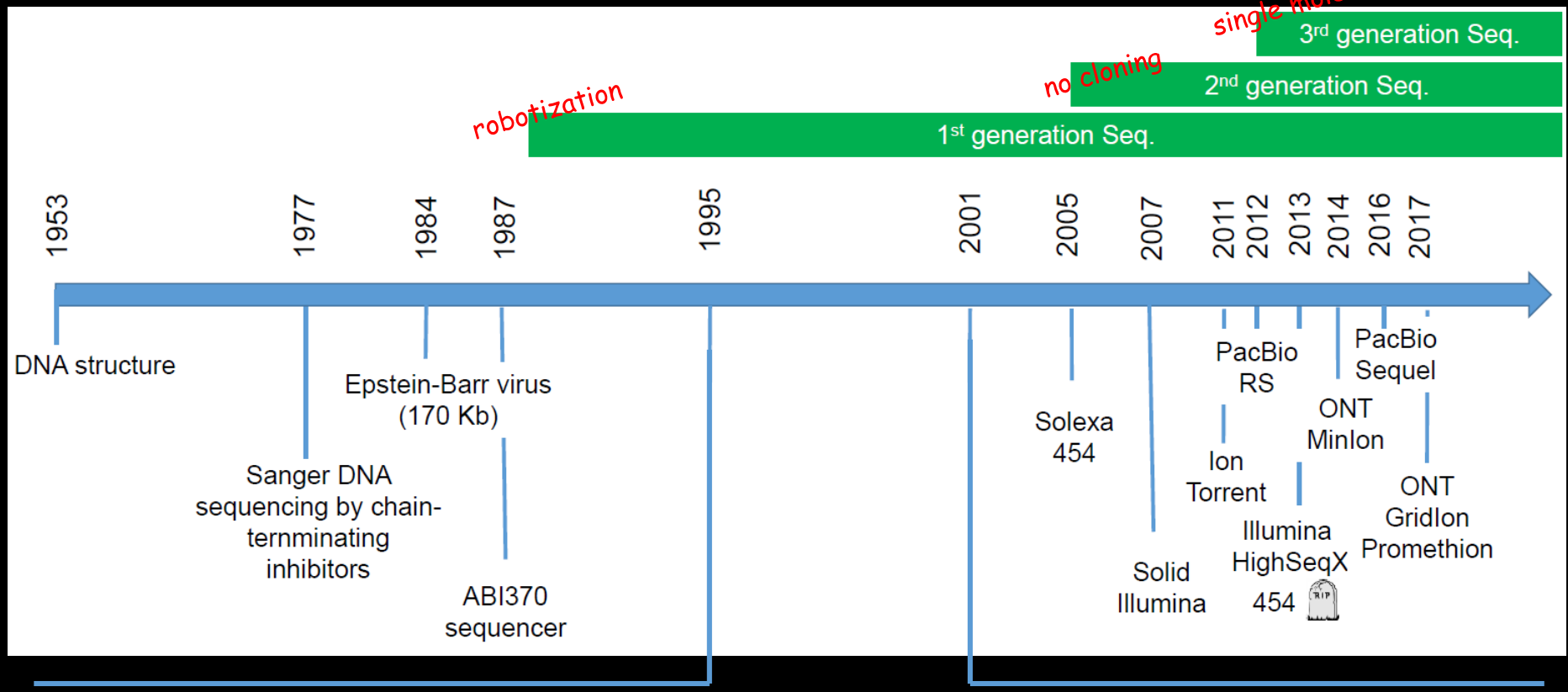
RNA

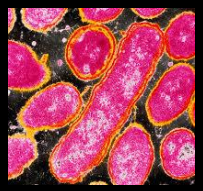


Reverse transcription

sebastien.aubourg@inra.fr

History





H. influenzae
1,8 Mb
(1995)

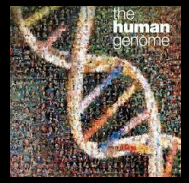
E. coli
4,6 Mb
(1997)

S. cerevisiae
6 Mb
(1997)

C. elegans
98 Mb
(1998)

A. thaliana
130 Mb
(2000)

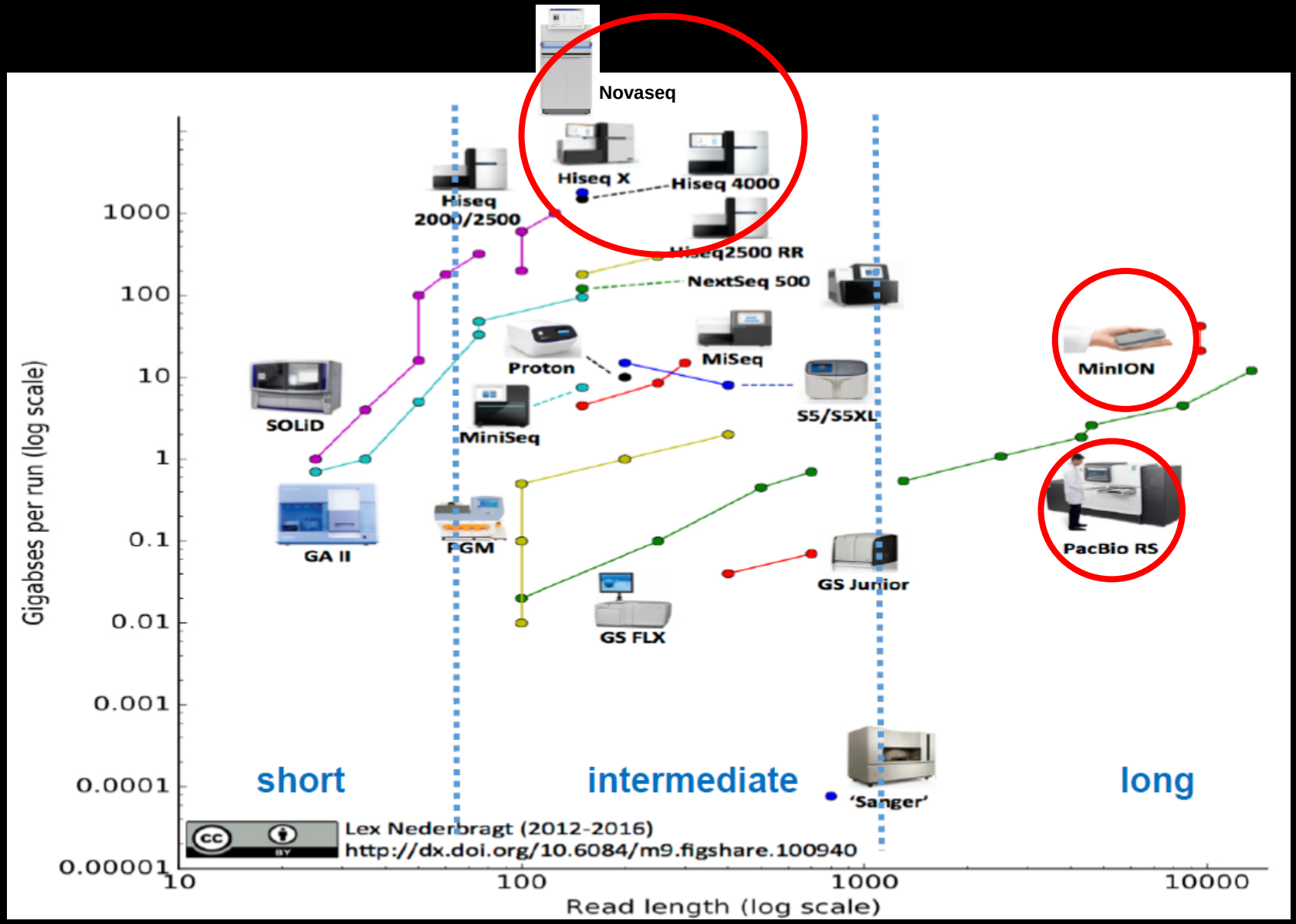
D. melanogaster
165 Mb
(2000)



H. sapiens
3 Gb
(2001-2004)

sebastien.aubourg@inra.fr

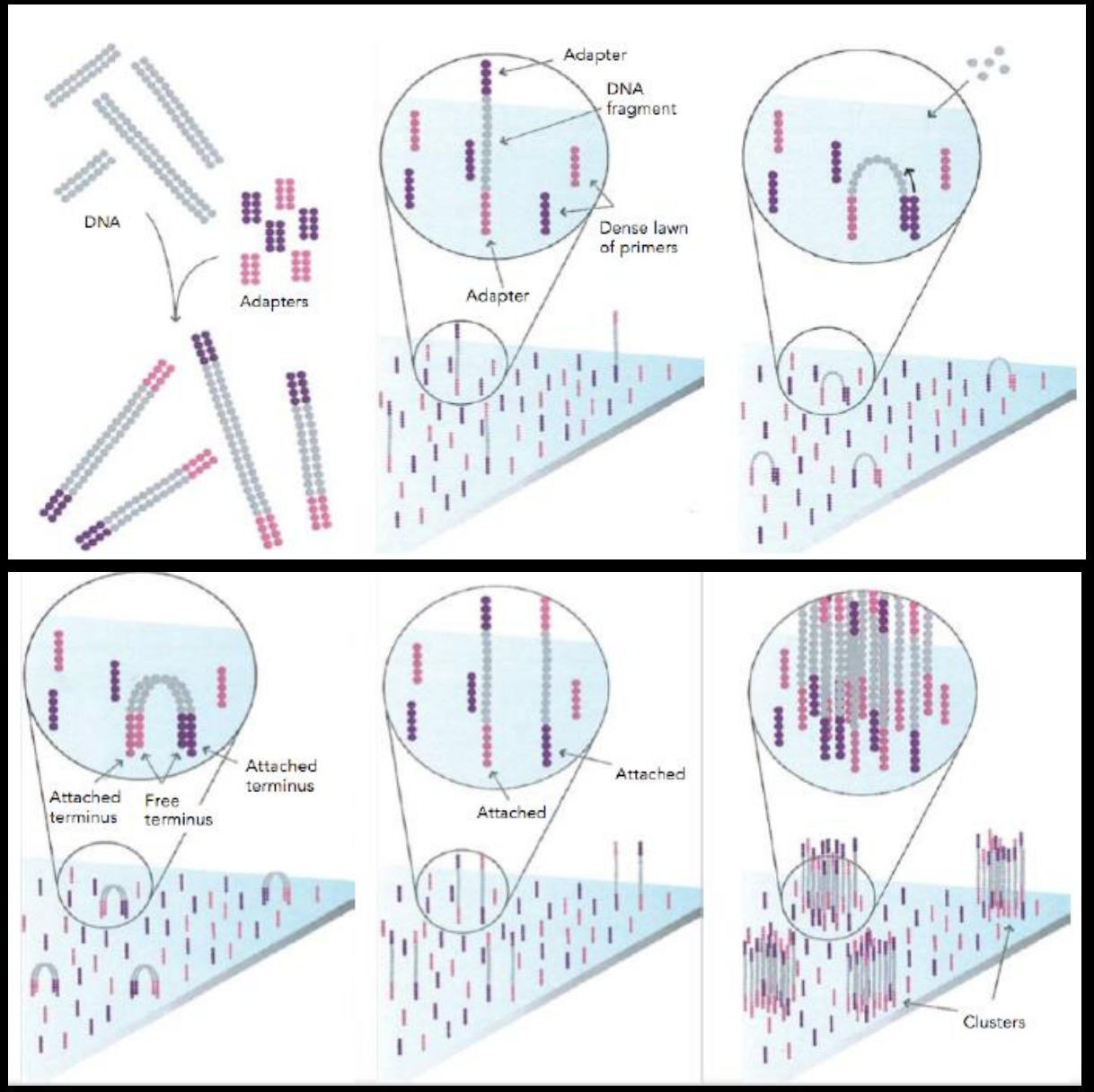
Next generation sequencing



sebastien.aubourg@inra.fr

Illumina technology (preparation)

- DNA library (fractioning, sizing)
- Adding specific adaptors
- DNA denaturation
- Random fixation on the flow cell
- PCR bridge
- Creation of clusters of amplified fragments

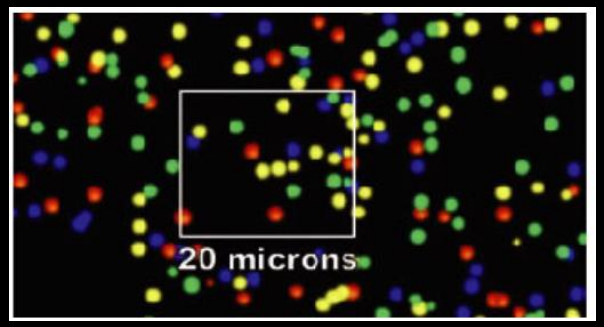


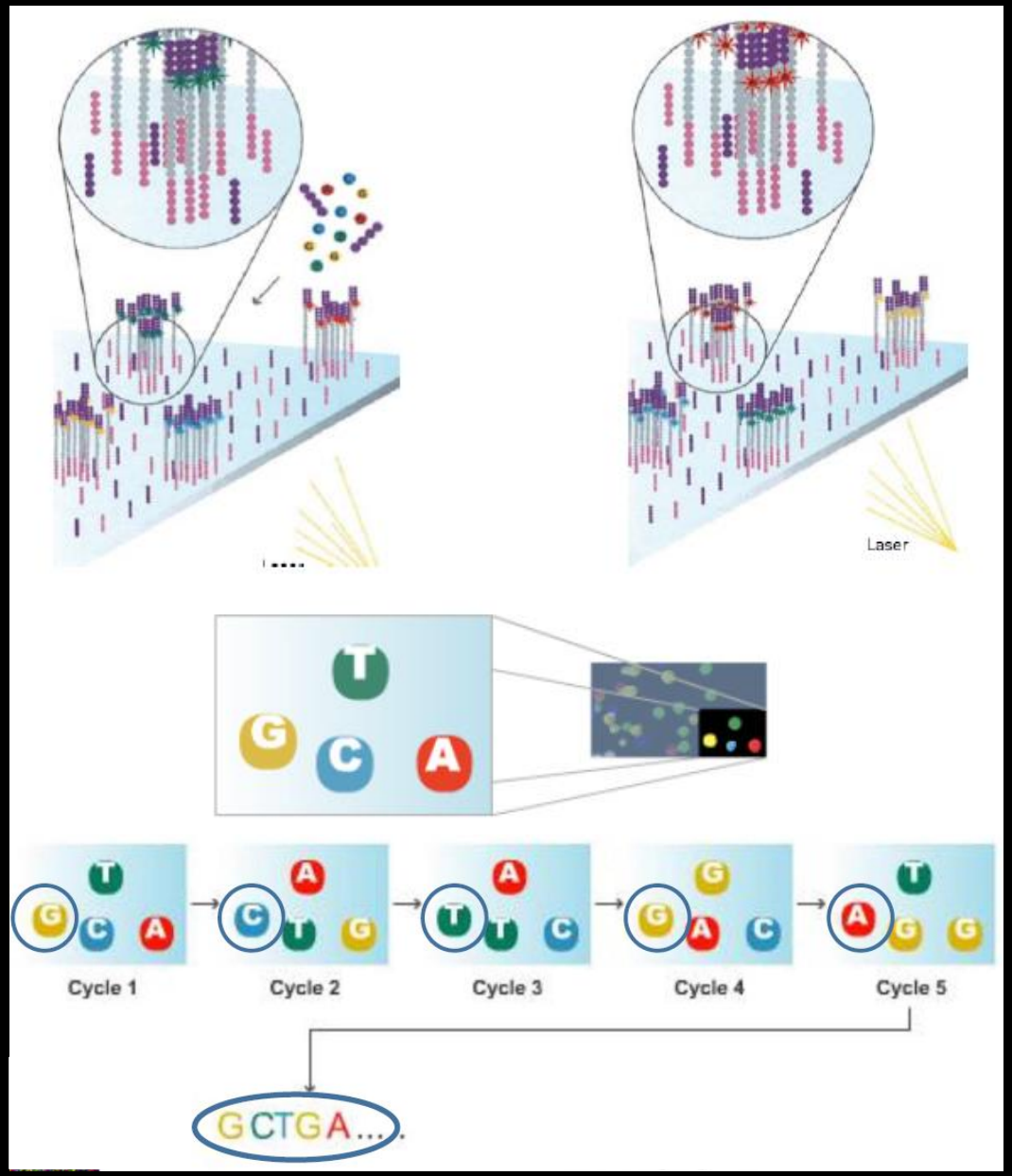
illumina®

sebastien.aubourg@inra.fr

Illumina technology (sequencing)

- Sequencing cycle with
  - 4 reversible terminators labeled with fluorophore
  - primers
  - DNA-polymerase
- Laser excitation
- Image capture to identify the incorporated base for each cluster
- Next cycle after removing the blocked 3' terminus and the fluorophore from each added base

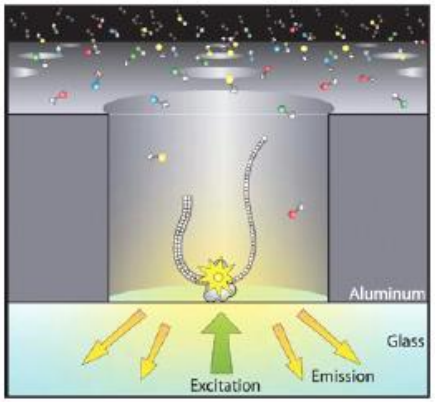


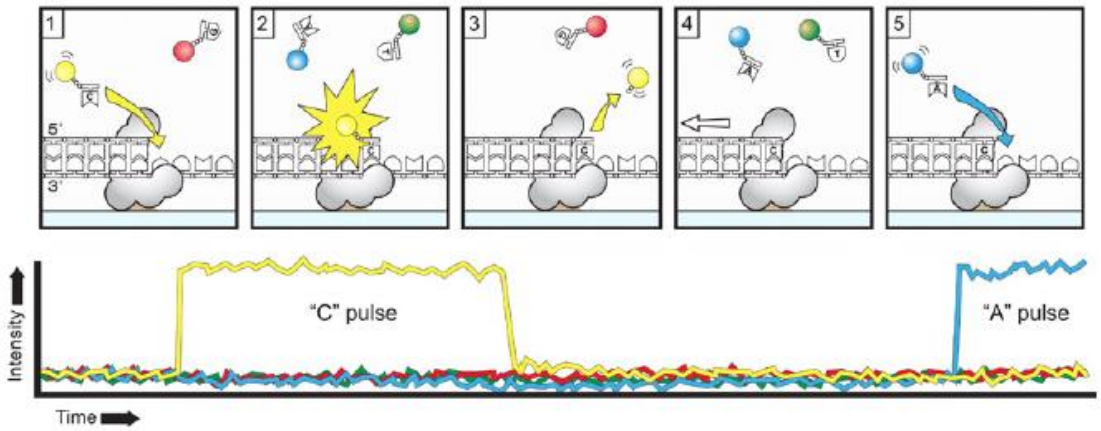


sebastien.aubourg@inra.fr

PacBio technology

Single Molecule, Real-Time (SMRT) sequencing / Long-read sequencing
DNA-polymerase is fixed in a micro-well (10^{-21} liters)
Labeled nucleotides incorporation is detected in real-time

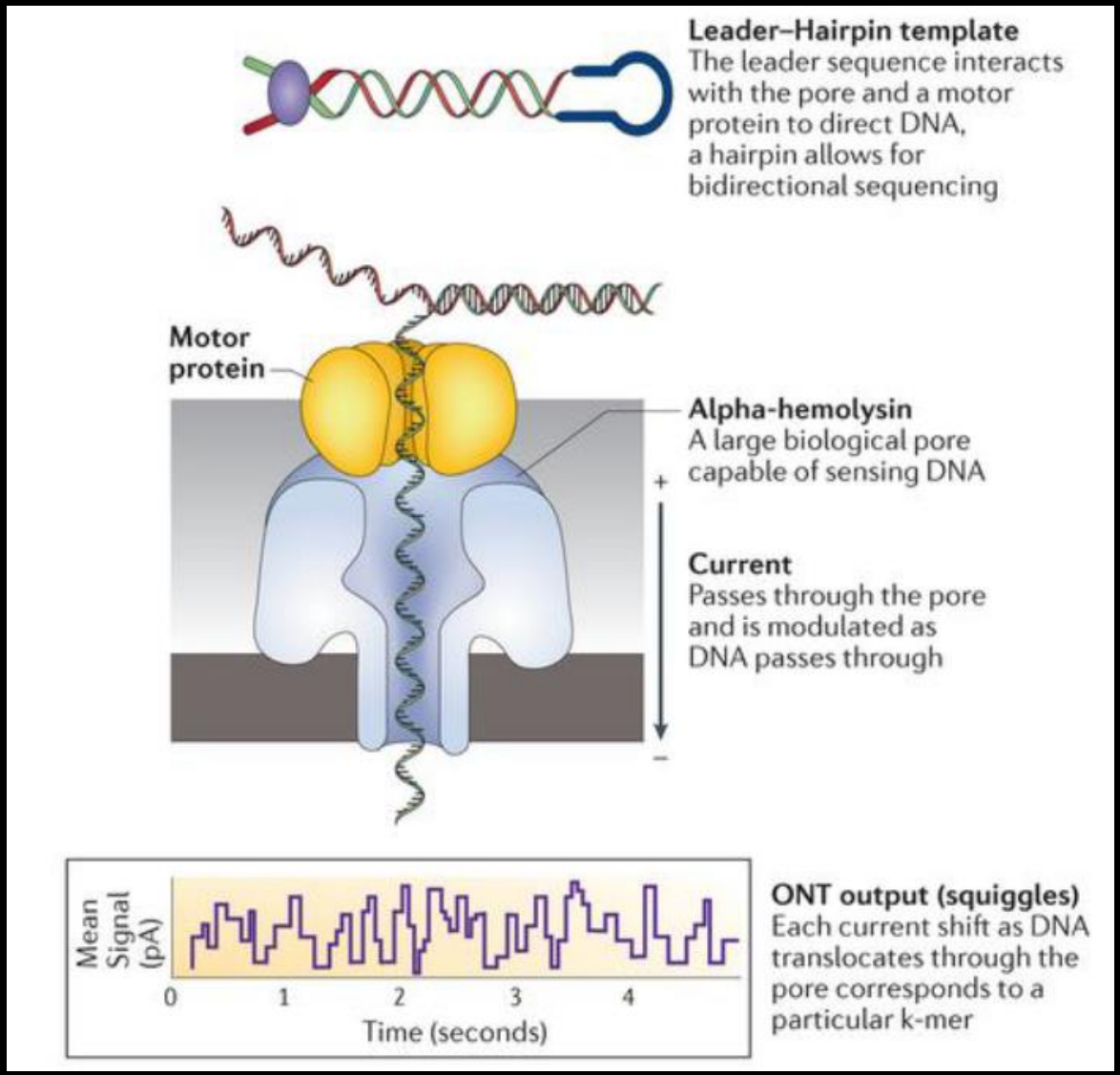




PACIFIC BIOSCIENCES®

sebastien.aubourg@inra.fr

Oxford Nanopore technology



Oxford NANOPORE Technologies®

sebastien.aubourg@inra.fr

Some numbers…

Technology	Read length (b)	Reads per run	Bases per run (Gb)	Error rate
ABI Sanger	800	96	0,0000768	< 1%
Illumina	150	600.10^6 to 6.10^9	7 to 2000	< 1%
PacBio	13 500	660 000	12 000	10-15%
Nanopore	9 545	$4{,}4.10^6$	42	5-15%

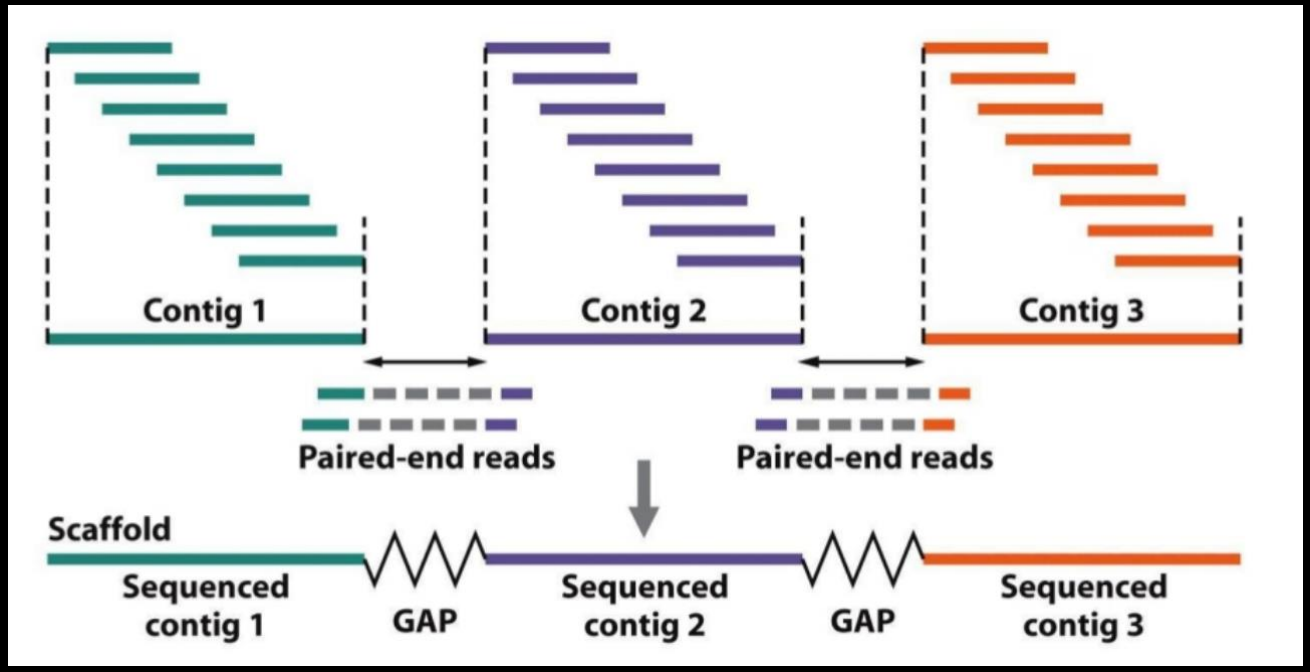
Human Genome Project

13 years (1990-2003)
$ 3,000,000,000

Illumina XTen technology (2014)
1 week
$ 1,000

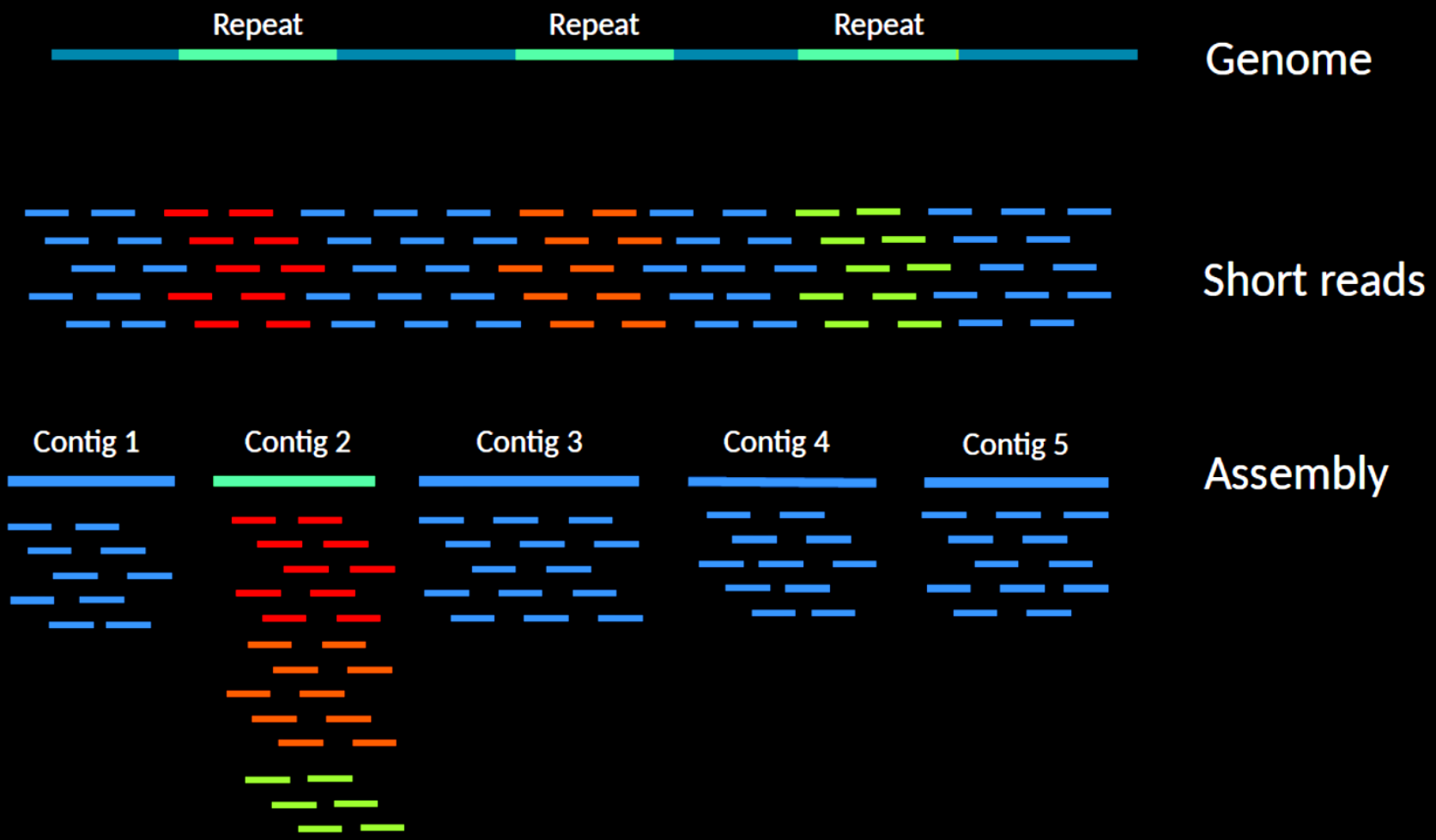
sebastien.aubourg@inra.fr

Genome assembly : the challenge



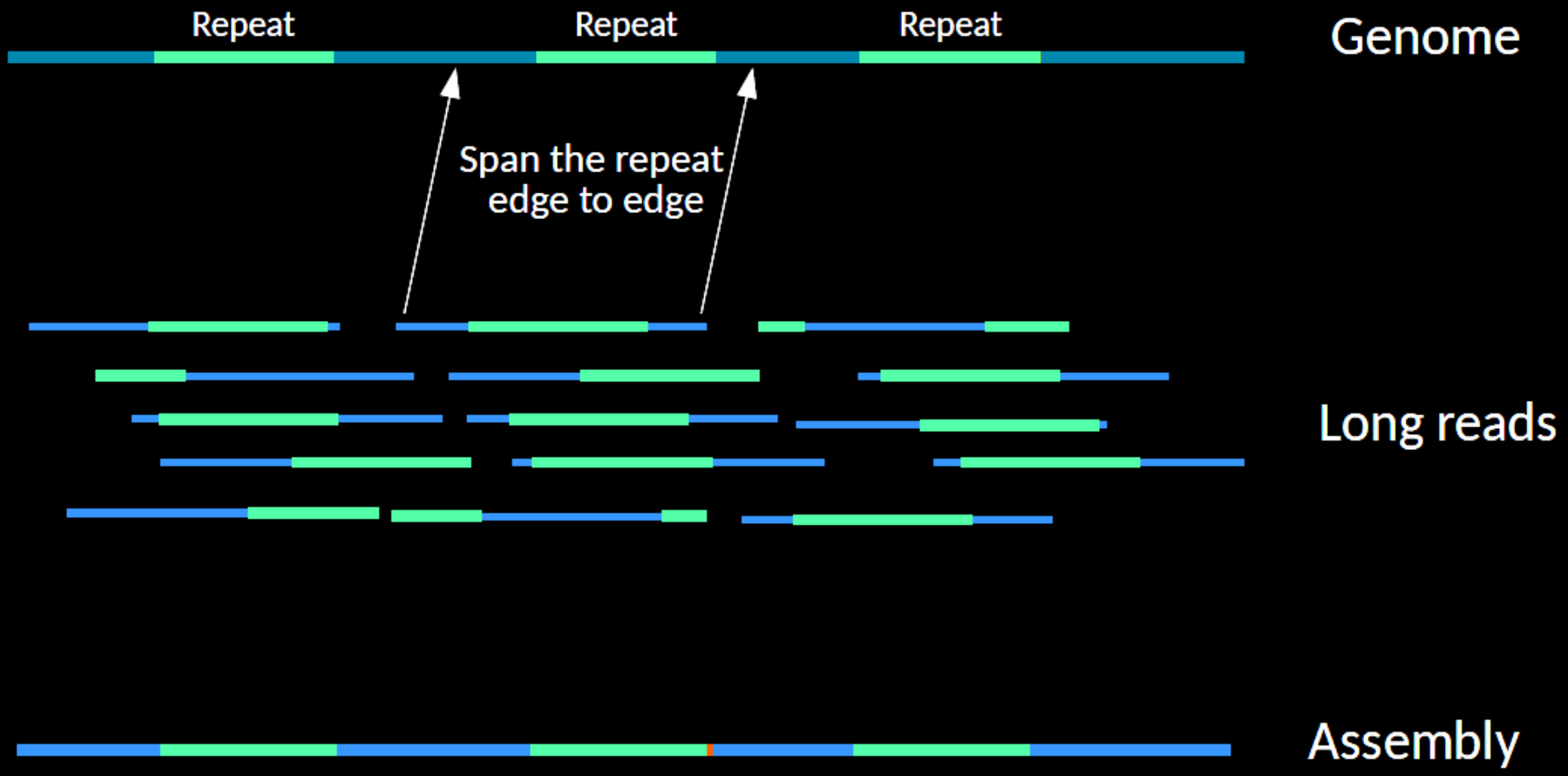
sebastien.aubourg@inra.fr

Genome assembly : advantage of long reads



sebastien.aubourg@inra.fr

Genome assembly : advantage of long reads



sebastien.aubourg@inra.fr

Genome assembly : optical mapping

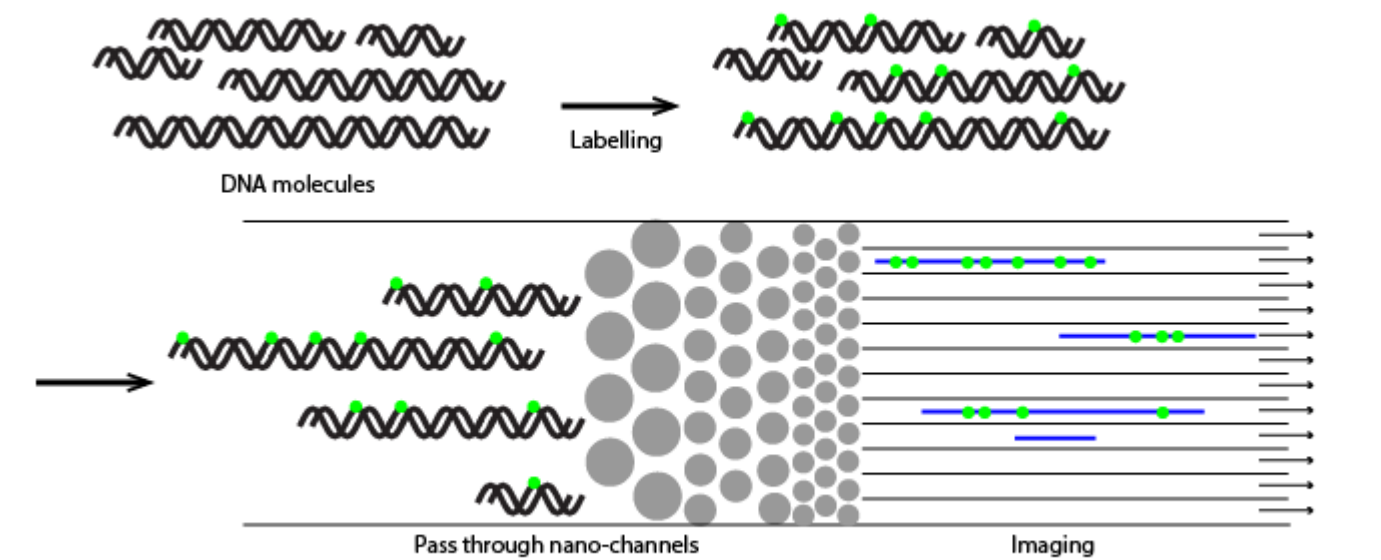
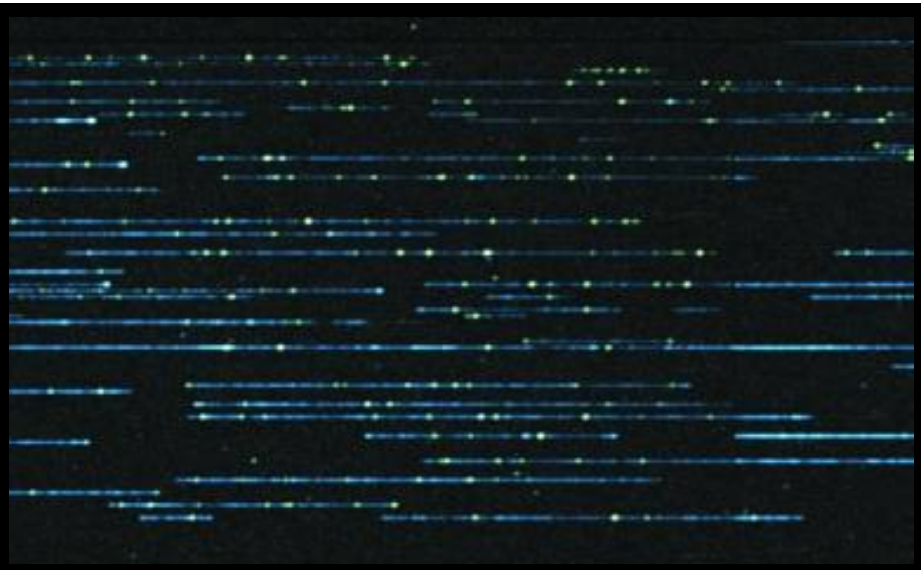
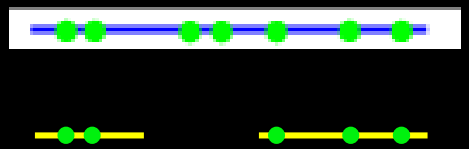


Fig.3 An overview of generation of optical mapping data using Bionano Genomics Inc. technologies.

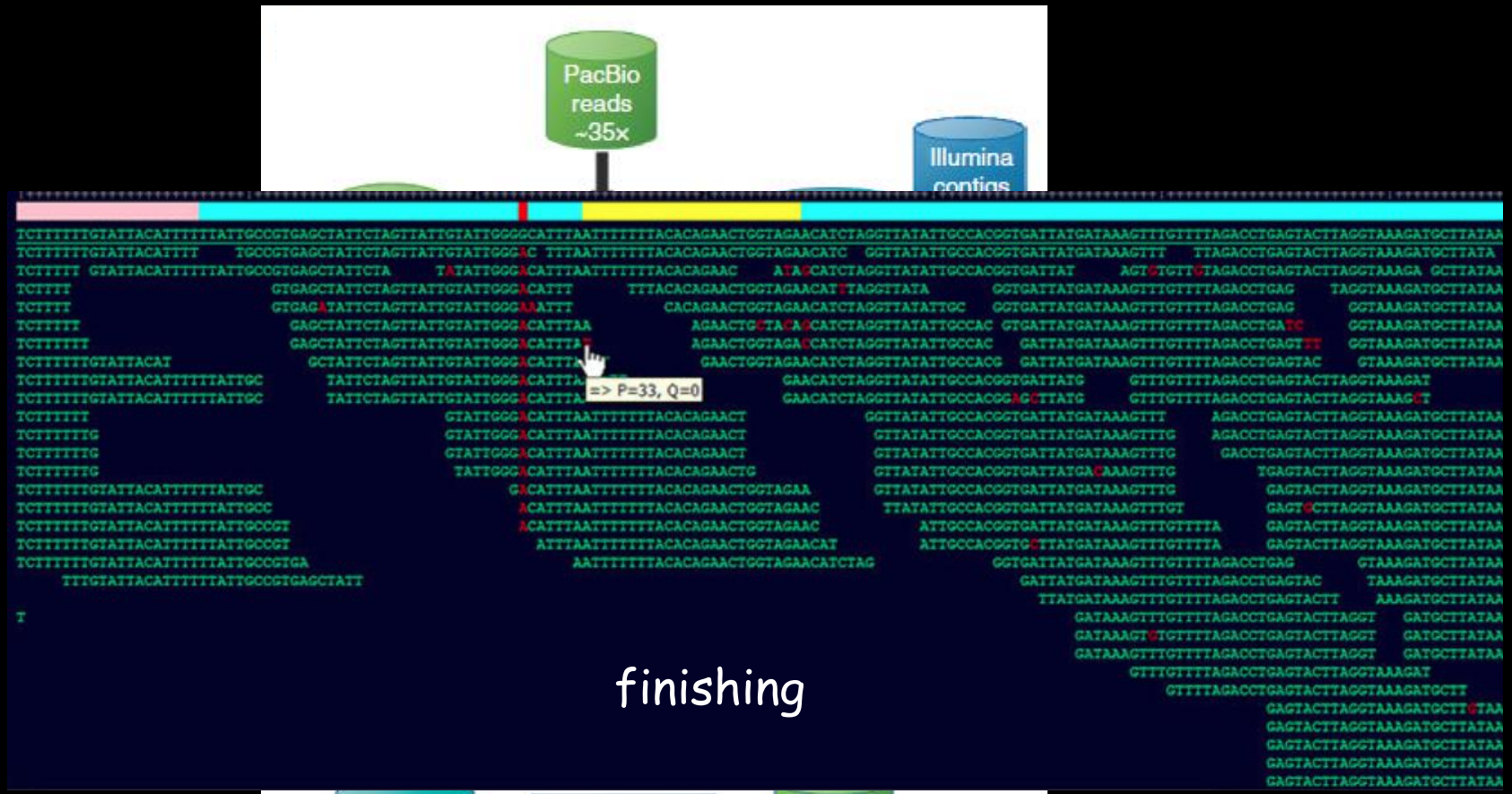




scaffolding

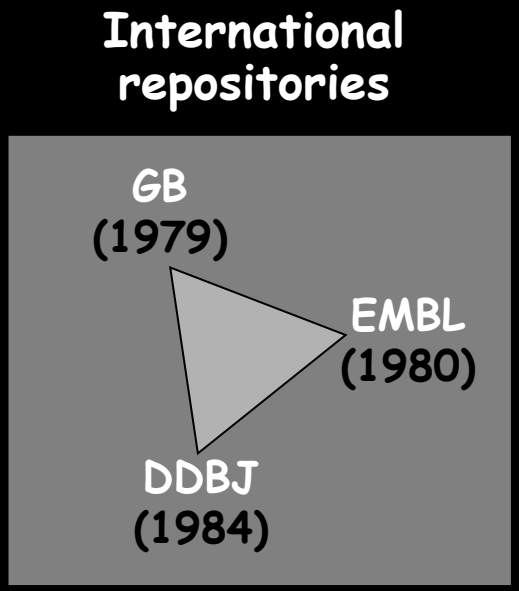
sebastien.aubourg@inra.fr

Genome assembly : Hybrid assembly strategy



sebastien.aubourg@inra.fr

Bank of sequences



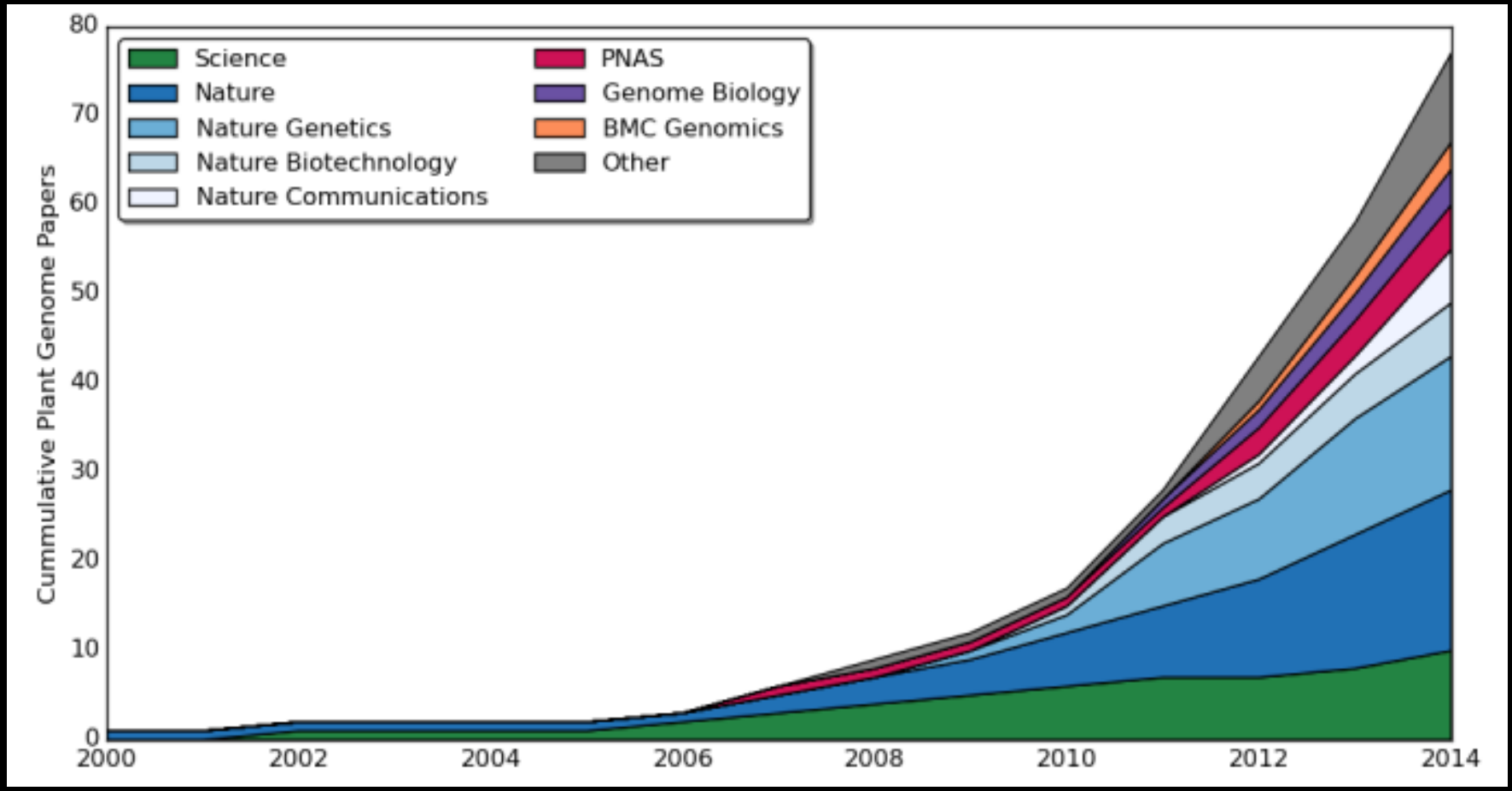
- GENBANK @ NCBI (National Center for Biotechnology Information)
- EMBL (European Molecular Biology Laboratory) @ EBI (European Bioinformatics Institute)
- DDBJ (DNA DataBase of Japan) @ DDBJ center

sebastien.aubourg@inra.fr

High throughput sequencing

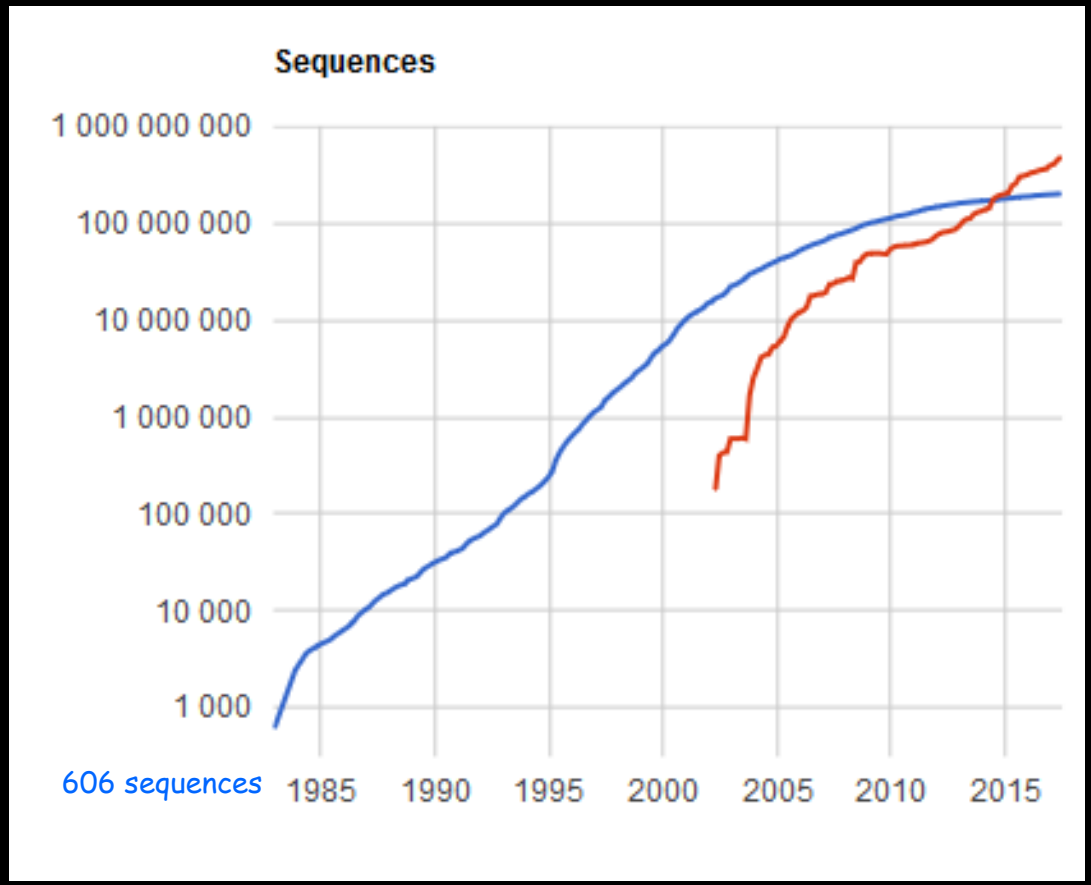
Genomics sizes up

NATURE|Vol 451|17 January 2008



sebastien.aubourg@inra.fr

High throughput sequencing



GenBank r229
Dec 2018

WGS
3656.10^9 nt

GENBANK
285.10^9 nt

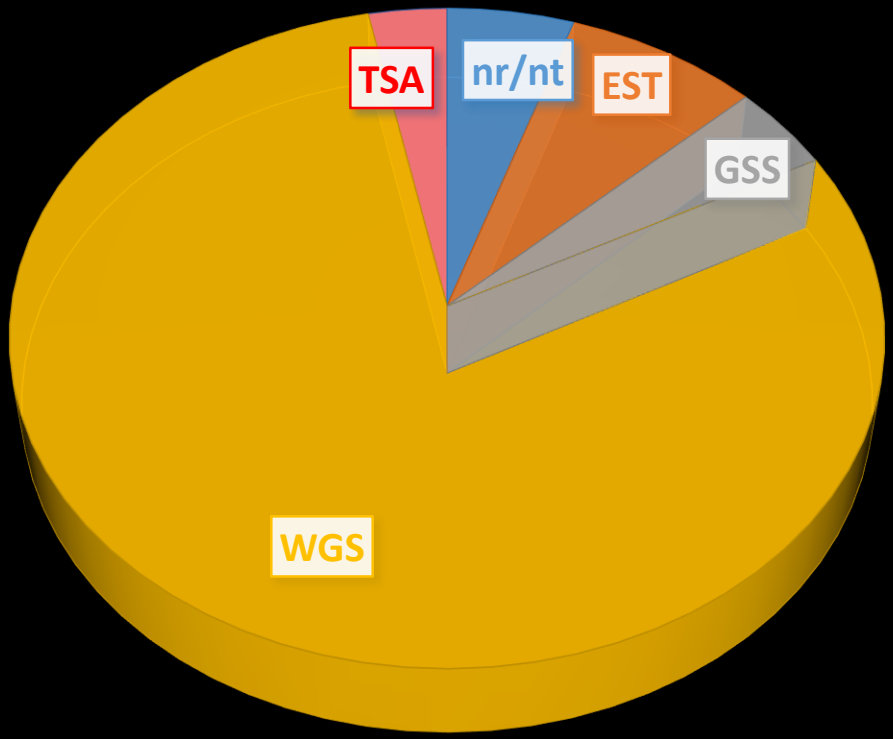
GenBank r7
1987

release occurs every 2 months
doubling every 15 months

sebastien.aubourg@inra.fr

GenBank divisions

Huge diversity of sequences with different origins (species, mutants), lengths (few bases to whole chromosomes), qualities (clones, assembly, single-pass, gold finishing), types (transcripts, genomic sequences, metagenomes, synthetics)...



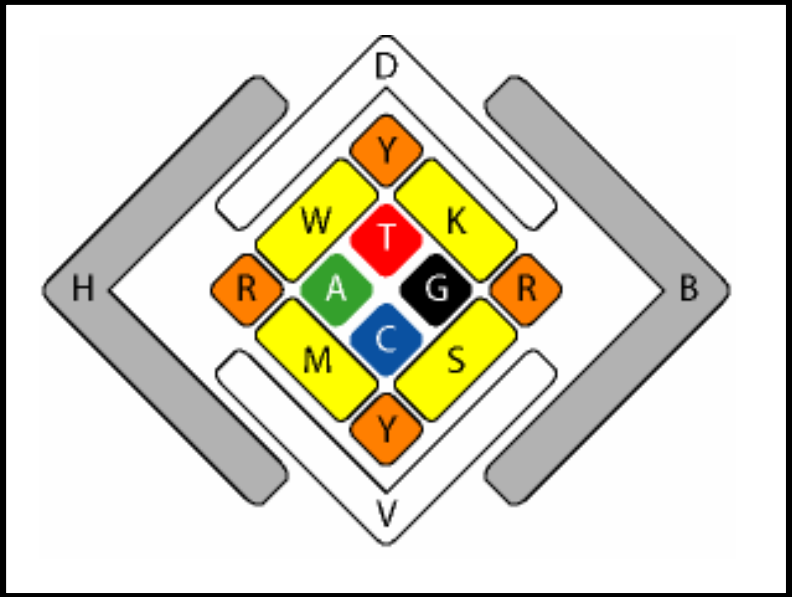
millions of sequences

nr/nt	50
EST	77
GSS	41
TSA	31
WGS	773

sebastien.aubourg@inra.fr

GenBank notes

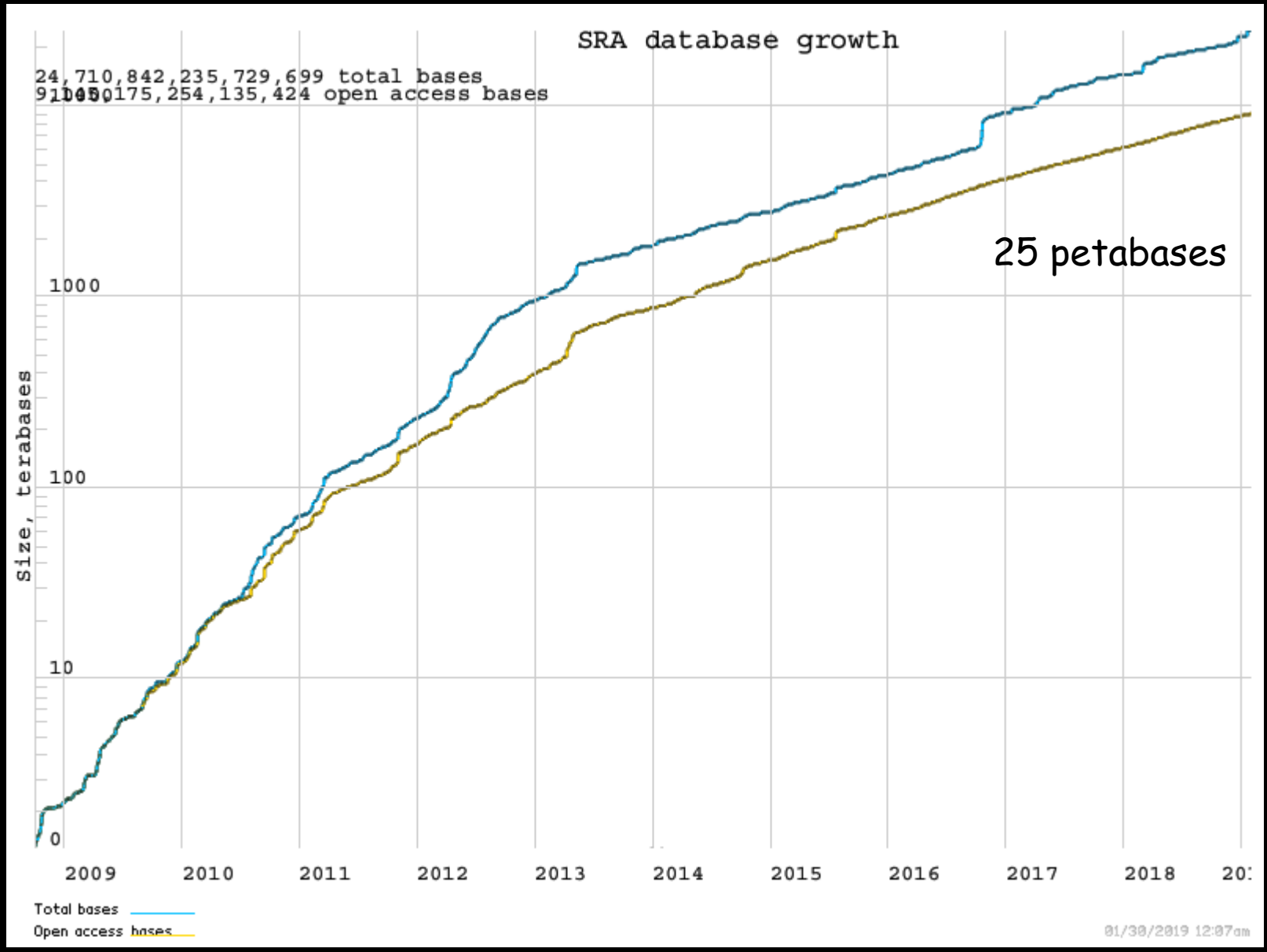
All sequences (RNA or DNA) contains A, T, G, C and degenerated IUB code but never U.



The strand available in GenBank is named '+' or 'direct' without consideration regarding the coding strand.

sebastien.aubourg@inra.fr

Sequence Read Archive (SRA division)



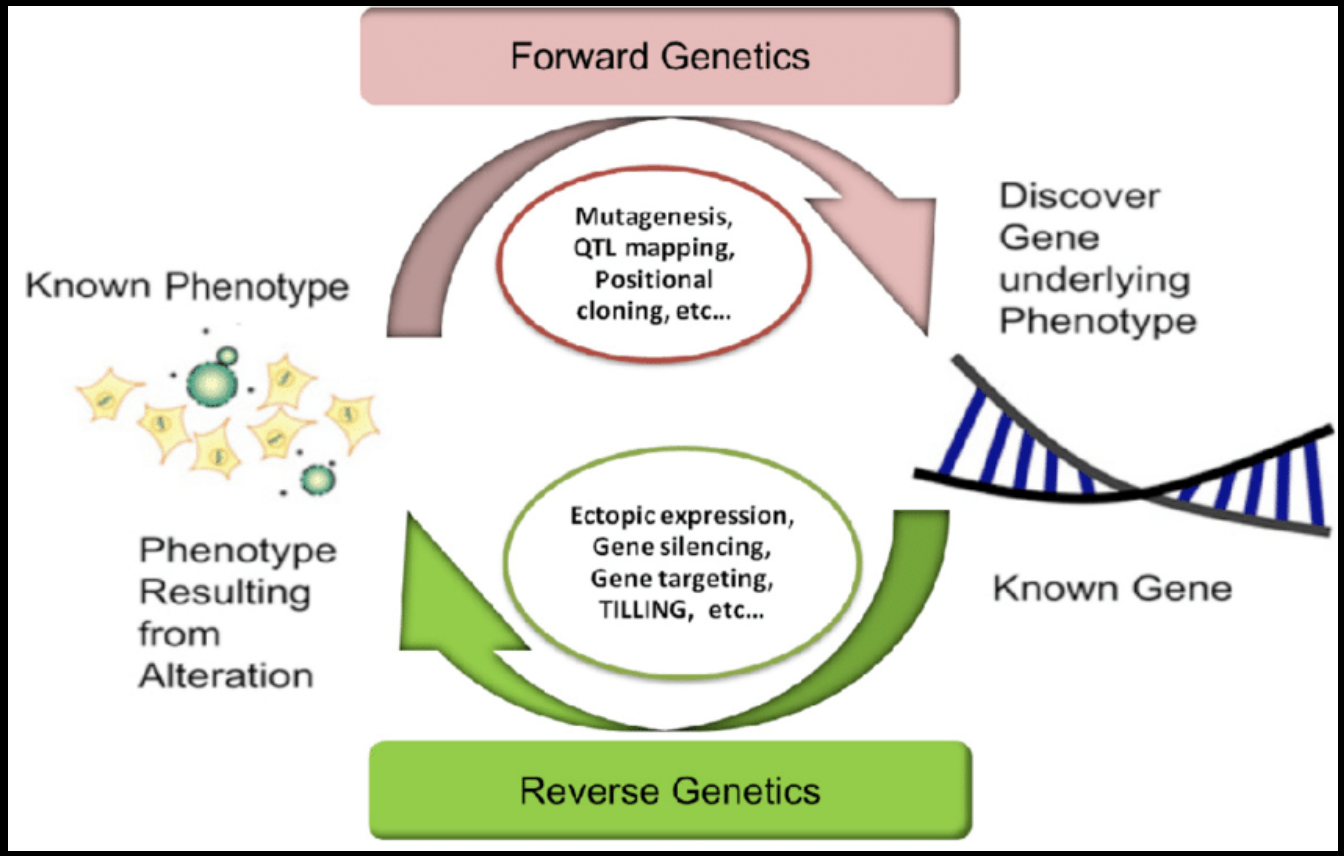
sebastien.aubourg@inra.fr

From genome to genes… from genes to functions

sebastien.aubourg@inra.fr

Genomics : from forward to reverse genetics

Forward 'classical' genetics seeks to find the genetic base of a phenotype or trait.
Reverse genetics seeks to find what phenotype arise a result of particular genetic sequence.



sebastien.aubourg@inra.fr

Genome Annotation

sebastien.aubourg@inra.fr

Annotation

Sequences ⟷ Biological information

Structural Annotation

- gene localization ?
- gene structure ?
- coding regions ?
- regulatory sites ?

Functional Annotation

- biochemical function ?
- biological functions ?
- regulation and interactions ?
- expression ?

Experimental work

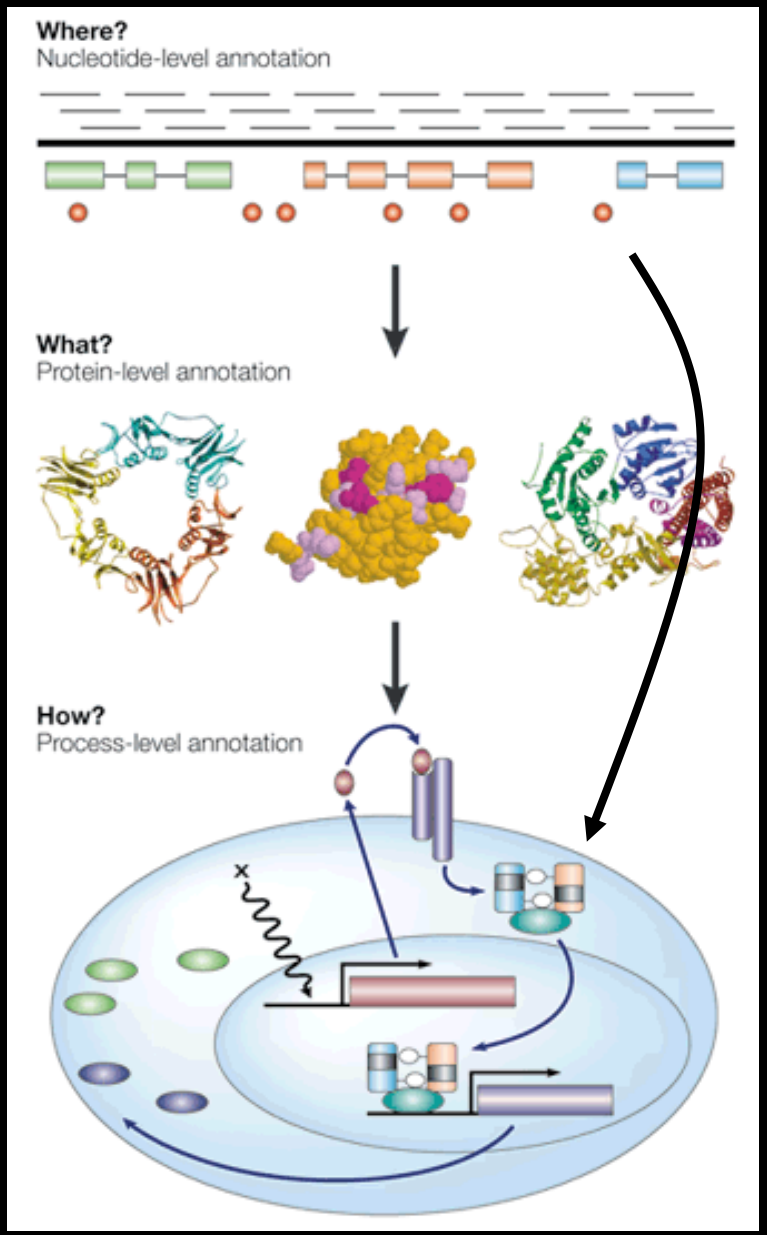
in silico predictions

sebastien.aubourg@inra.fr

Annotation

Structural Annotation

Functional Annotation



biochemical function

biological function

Relational Annotation

Systems Biology

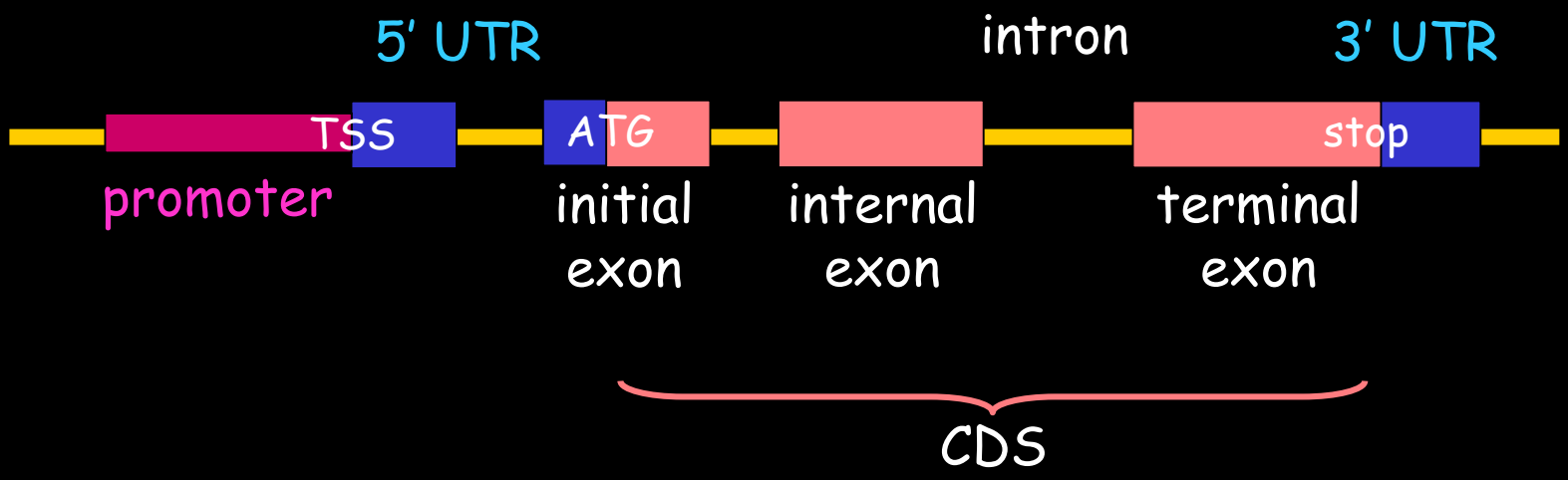
sebastien.aubourg@inra.fr

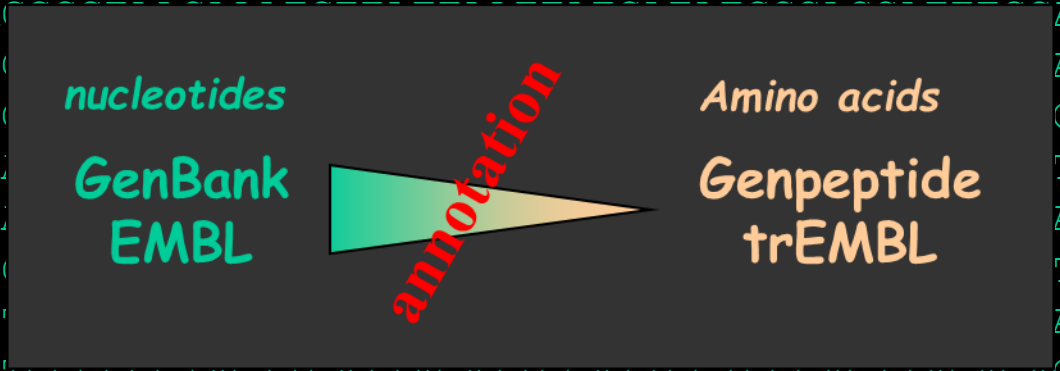
Genome Annotation

Structure

sebastien.aubourg@inra.fr

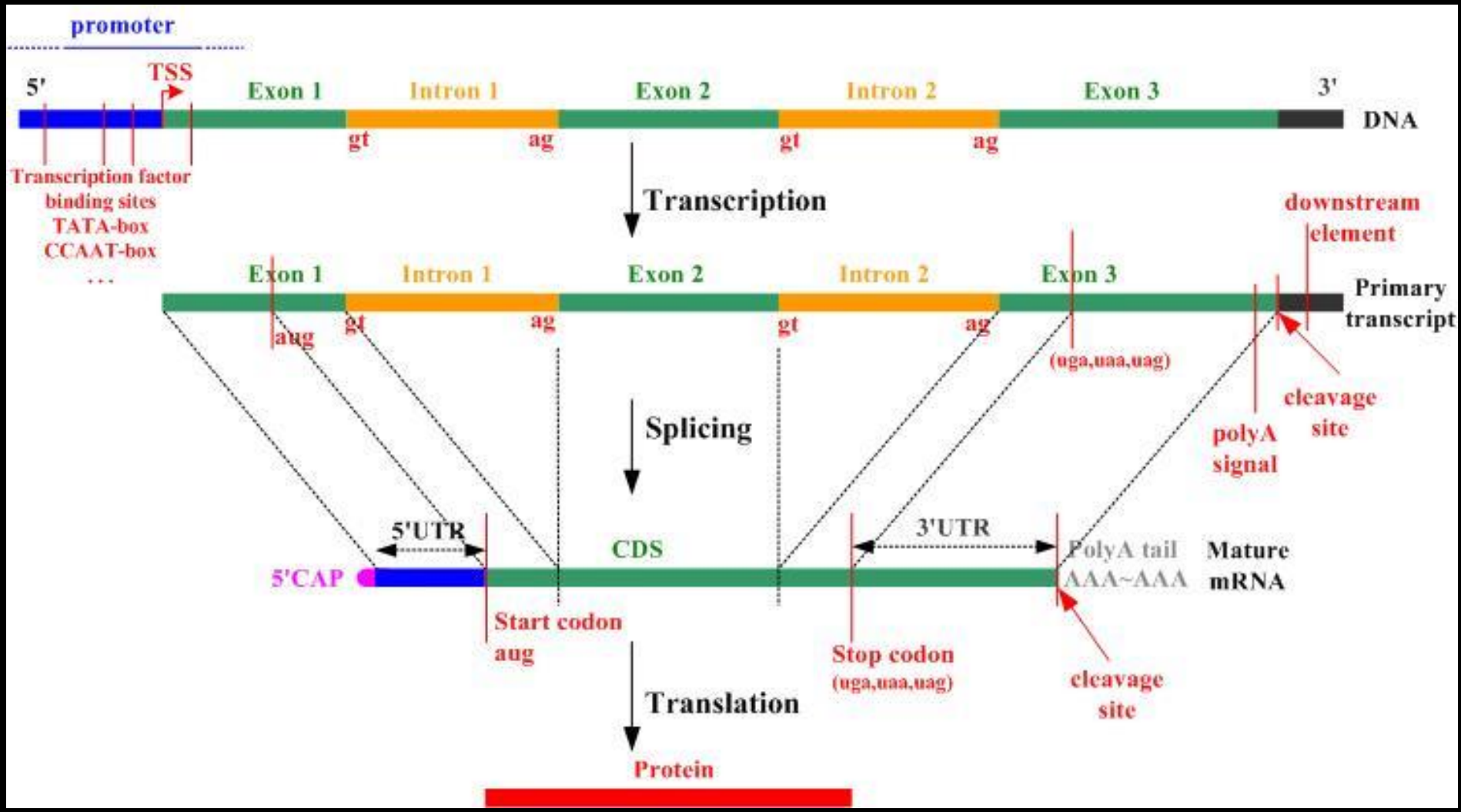
Structural annotation: goals





sebastien.aubourg@inra.fr

As a reminder…



sebastien.aubourg@inra.fr

Prediction methods

by similarity

ab initio

Sequence comparisons with databank content

BLAST

+

Definition of a 'coding potential'

composition

+

Looking for splicing sites

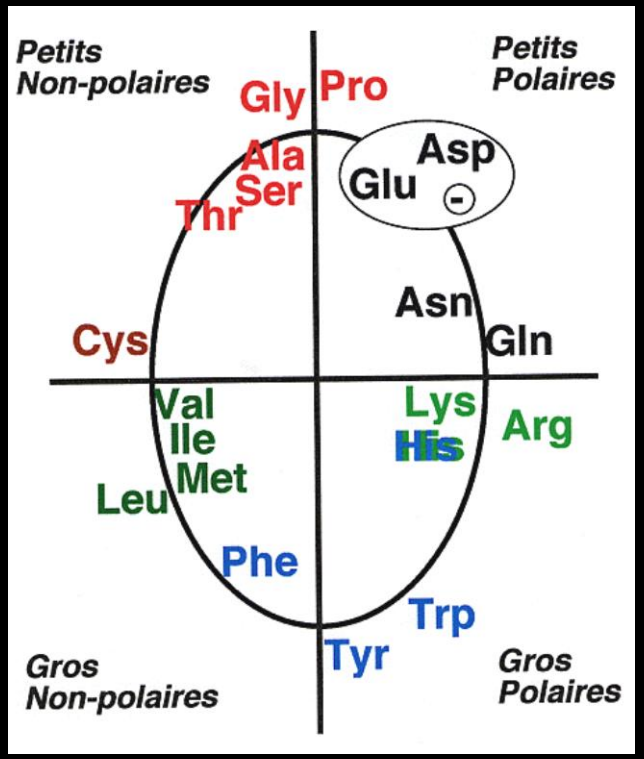
...GT:AG...
...GC:AG...

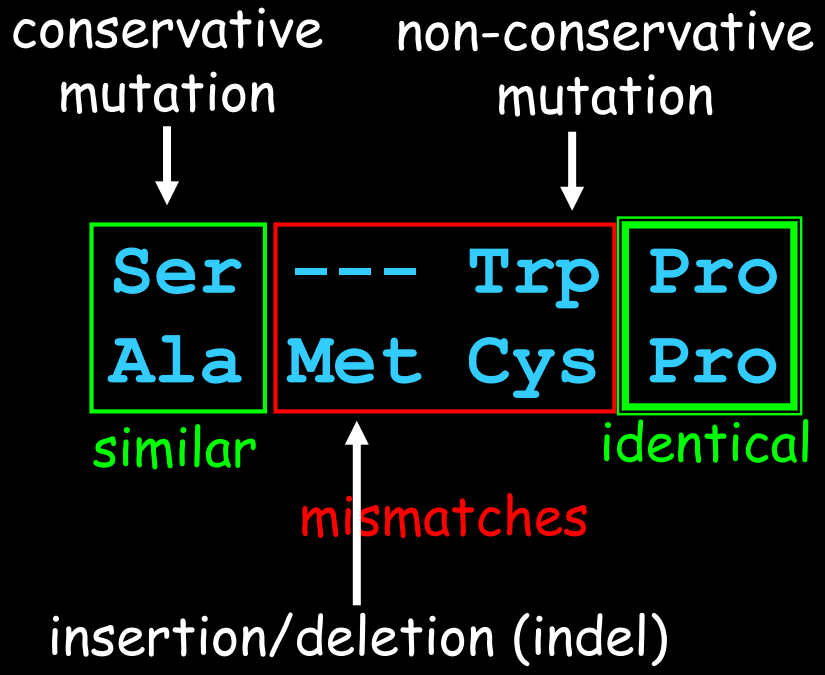
Learning step is necessary

sebastien.aubourg@inra.fr

Quantification of similarities

Representation of a DAYHOFF matrix of mutations (substitutions) by a plan projection.





sebastien.aubourg@inra.fr

Quantification of similarities

PAM (1978)

BLOSUM (1992)

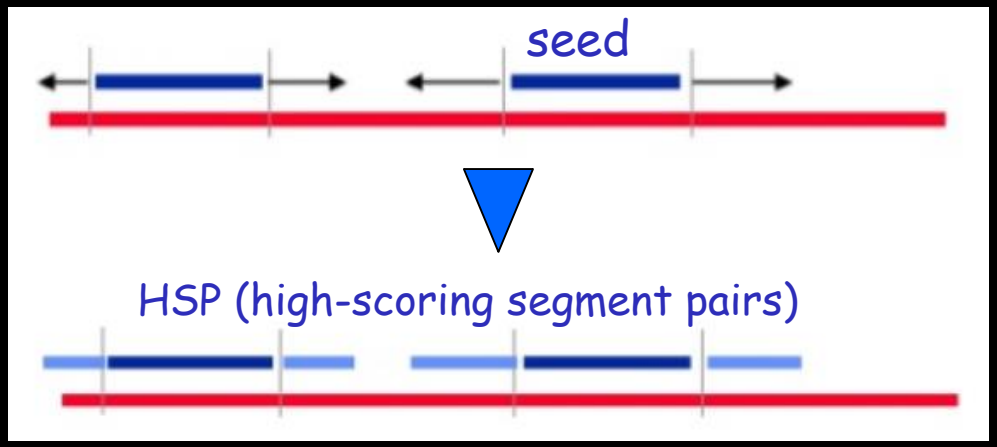
BLOSUM62

	C	G	A	T	S	N	D	E	Q	K	R	V	I	L	M	W	F	Y	H	P
C	9	-3	0	-1	-1	-3	-3	-4	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2	-3	-3
G	-3	6	0	-2	0	0	-1	-2	-2	-2	-2	-3	-4	-4	-3	-2	-3	-3	-2	-2
A	0	0	4	0	1	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	-3	-2	-2	-2	-1
T	-1	-2	0	5	1	0	-1	-1	-1	-1	-1	0	-1	-1	-1	-2	-2	-2	-2	-1
S	-1	0	1	1	4	1	0	0	0	0	-1	-2	-2	-2	-1	-3	-2	-2	-1	-1
N	-3	0	-2	0	1	6	1	0	0	0	0	-3	-3	-3	-2	-4	-3	-2	1	-2
D	-3	-1	-2	-1	0	1	6	2	0	-1	-2	-3	-3	-4	-3	-4	-3	-3	-1	-1
E	-4	-2	-1	-1	0	0	2	5	2	1	0	-2	-3	-3	-2	-3	-3	-2	0	-1
Q	-3	-2	-1	-1	0	0	0	2	5	1	1	-2	-3	-2	0	-2	-3	-1	0	-1
K	-3	-2	-1	-1	0	0	-1	1	1	5	2	-2	-3	-2	-1	-3	-3	-2	-1	-1
R	-3	-2	-1	-1	-1	0	-2	0	1	2	5	-3	-3	-2	-1	-3	-3	-2	0	-2
V	-1	-3	0	0	-2	-3	-3	-2	-2	-2	-3	4	3	1	1	-3	-1	-1	-3	-2
I	-1	-4	-1	-1	-2	-3	-3	-3	-3	-3	-3	3	4	2	1	-3	0	-1	-3	-3
L	-1	-4	-1	-1	-2	-3	-4	-3	-2	-2	-2	1	2	4	2	-2	0	-1	-3	-3
M	-1	-3	-1	-1	-1	-2	-3	-2	0	-1	-1	1	1	2	5	-1	0	-1	-2	-2
W	-2	-2	-3	-2	-3	-4	-4	-3	-2	-3	-3	-3	-3	-2	-1	11	1	2	-2	-4
F	-2	-3	-2	-2	-2	-3	-3	-3	-3	-3	-3	-1	0	0	0	1	6	3	-1	-4
Y	-2	-3	-2	-2	-2	-2	-3	-2	-1	-2	-2	-1	-1	-1	-1	2	3	7	2	-3
H	-3	-2	-2	-2	-1	1	-1	0	0	-1	0	-3	-3	-3	-2	-2	-1	2	8	-2
P	-3	-2	-1	-1	-1	-2	-1	-1	-1	-1	-2	-2	-3	-3	-2	-4	-4	-3	-2	7

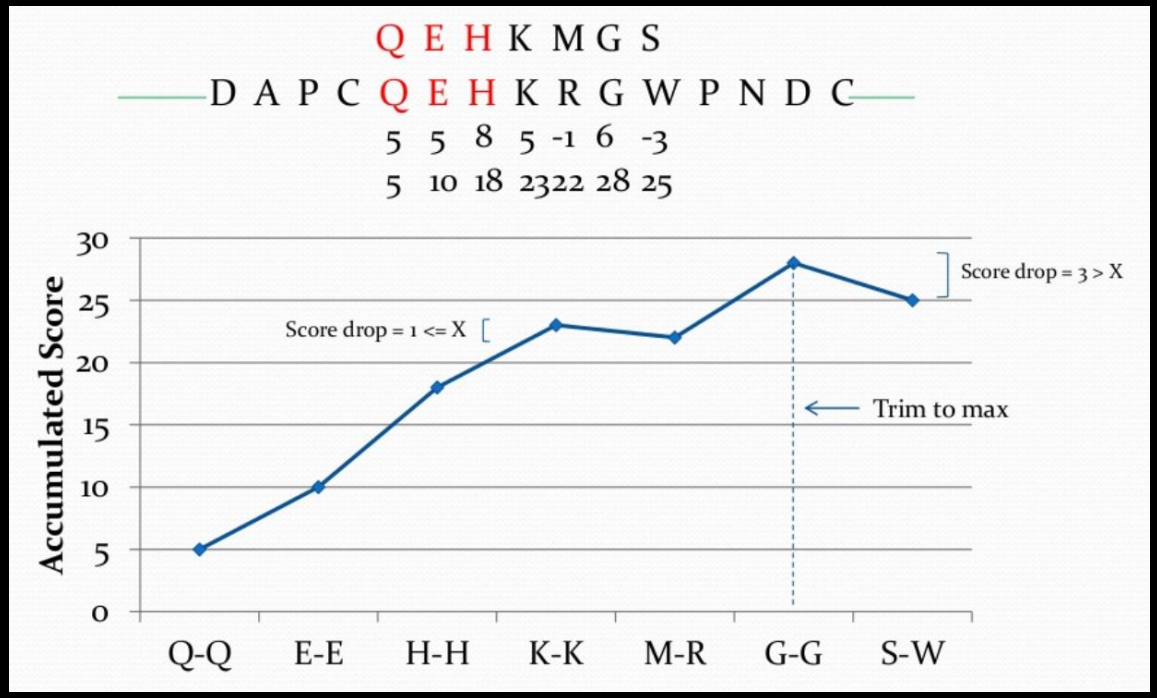
sebastien.aubourg@inra.fr

Search for sequence similarities: BLAST

Basic
Local
Alignment
Search
Tool

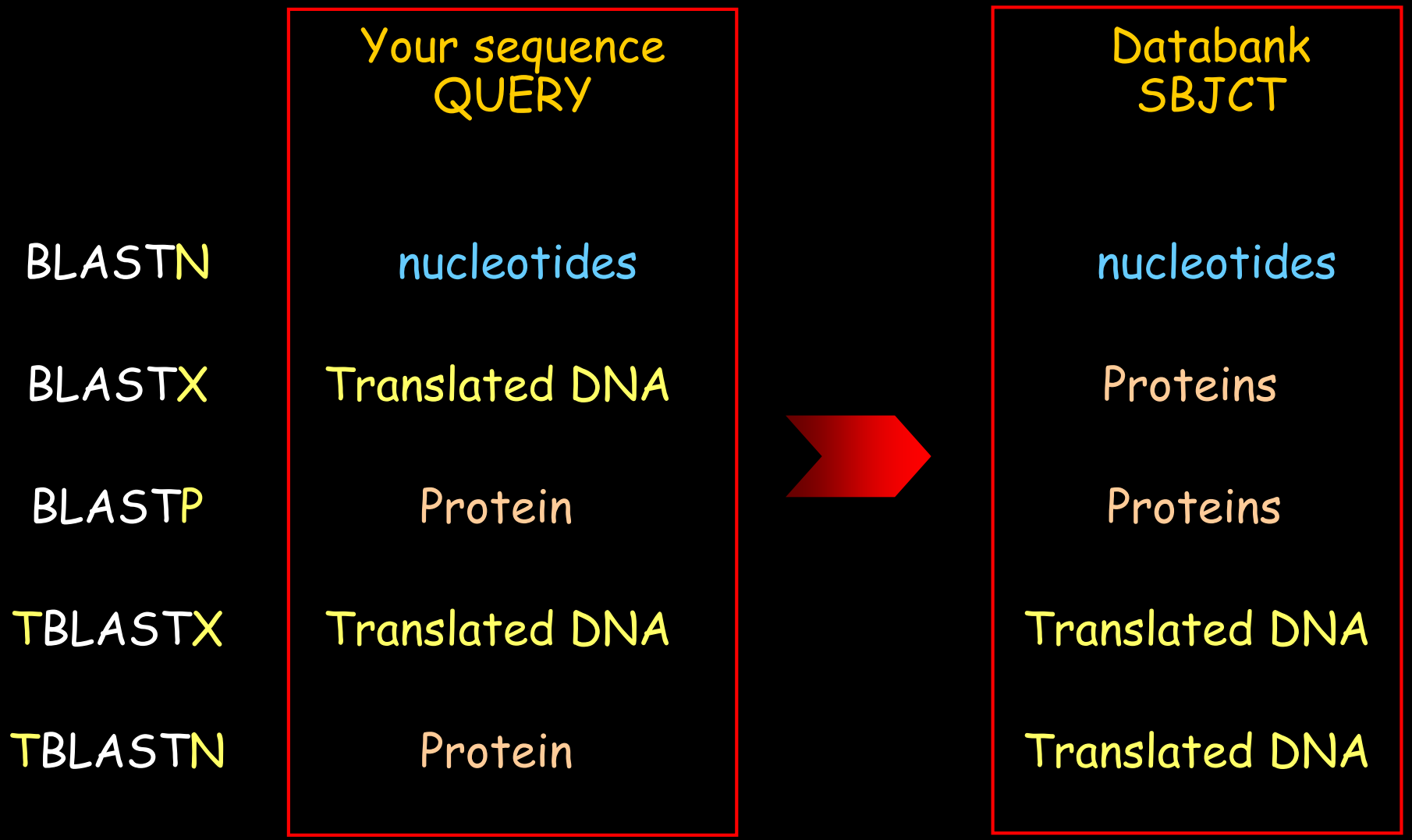


Algorithm for comparing a given sequence against sequences in a database



sebastien.aubourg@inra.fr

BLAST algorithms



sebastien.aubourg@inra.fr

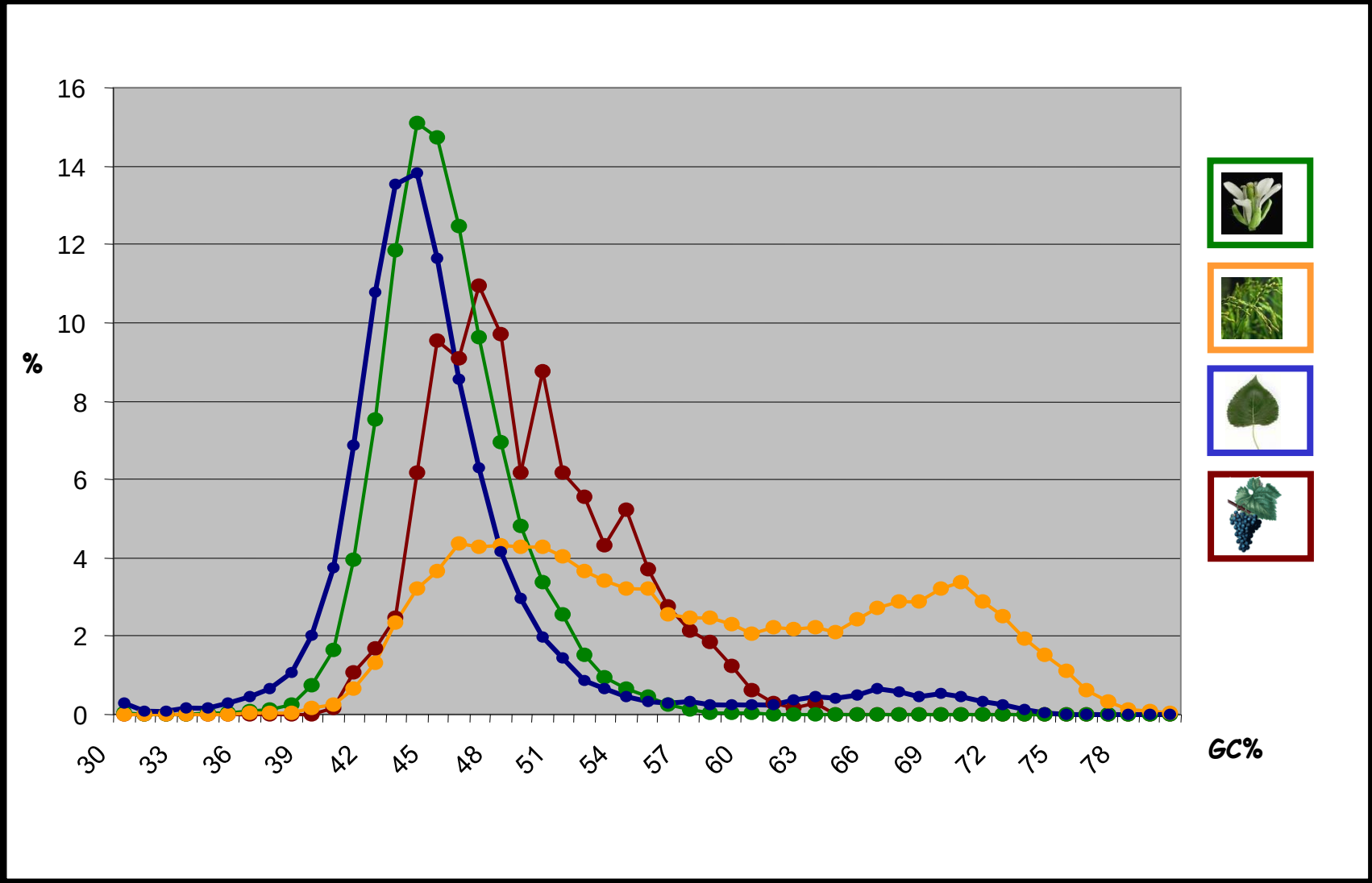
The 'coding potential'

- Each organism use differently the genetic code redundancy
- Amino acids and synonym codons have not the same frequency in a genome (W rare, bias in codon usage)
- There are relations between successive codons.

The probability of the presence of a nucleotide depends to the previous K nucleotides (K: order of the Markov chain). This probability is different between coding and non-coding regions.

sebastien.aubourg@inra.fr

Coding potential: learning step is necessary



GC content in CDS

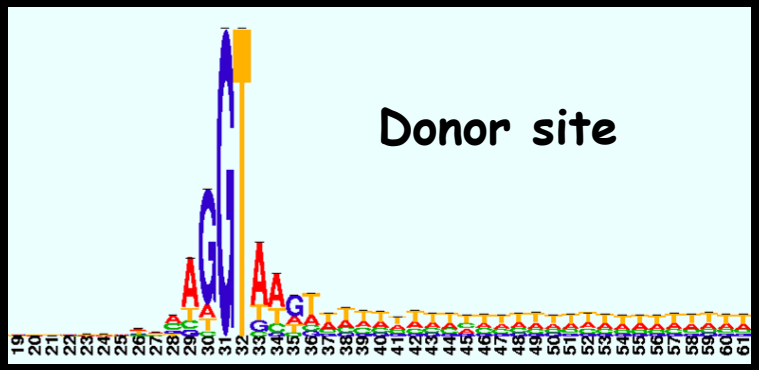
sebastien.aubourg@inra.fr

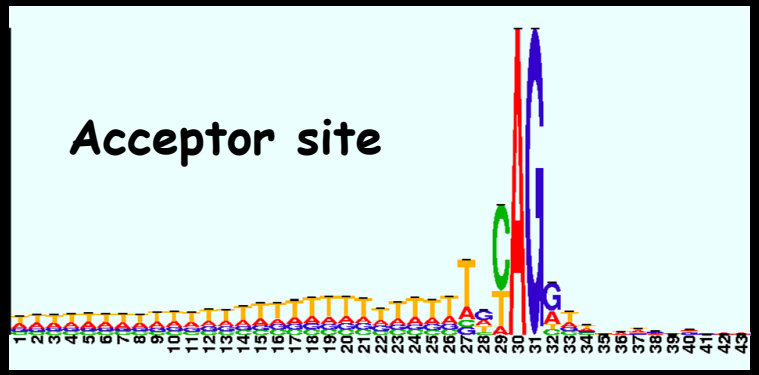
Looking for splicing sites

Only in eukaryotic genomes…

Positional Weight Matrices - WAM

LOGO



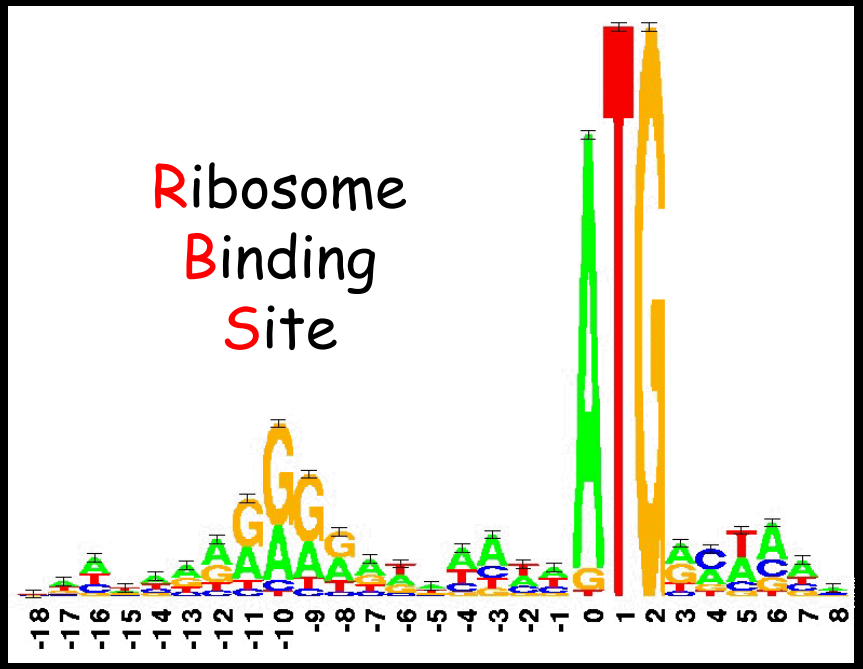


U2	GT::AG	99%
	GC::AG	1%
U12	AT::AT AT::AC	1‰

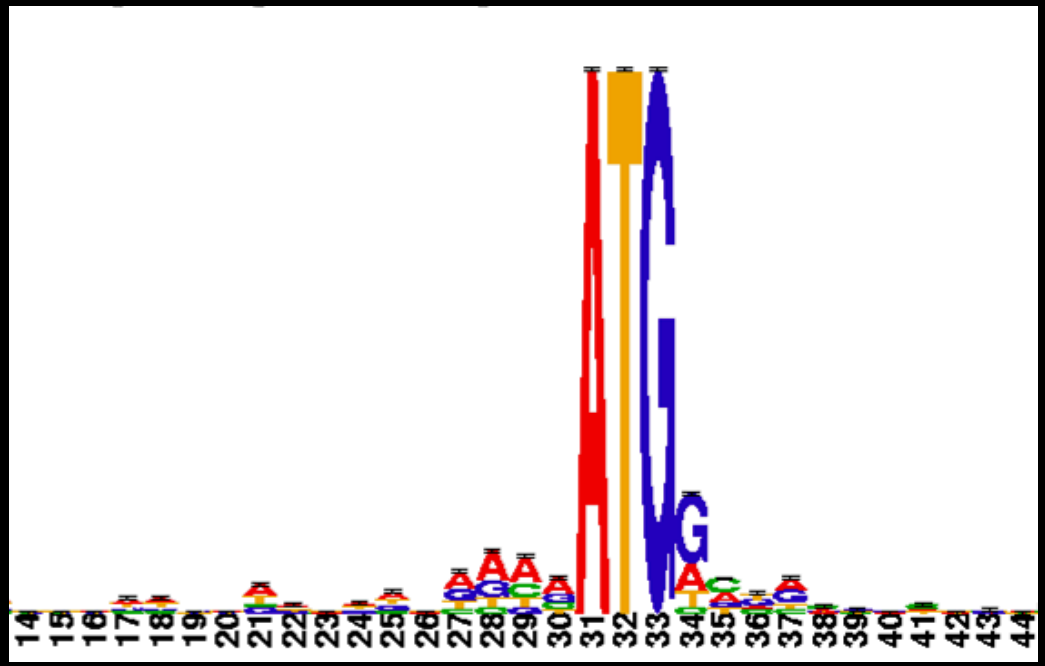
sebastien.aubourg@inra.fr

Other signal: Translation initiation sites

Prokaryotes (*E. coli*)



Eukaryotes (*A. thaliana*)

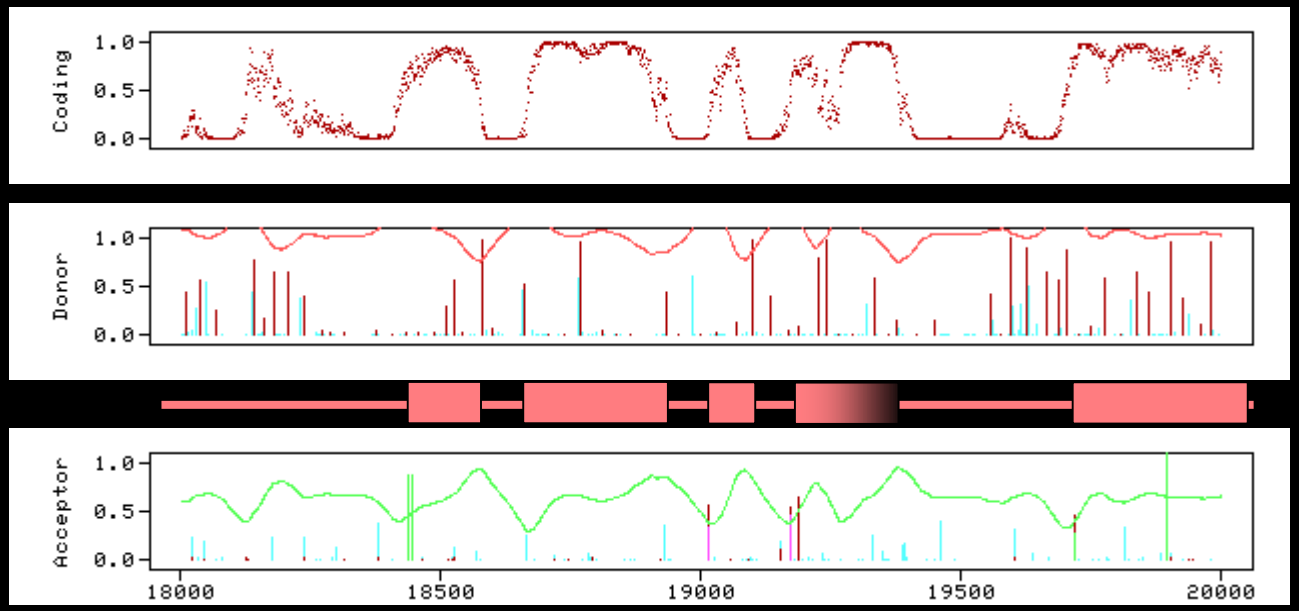


sebastien.aubourg@inra.fr

Exon and intron predictions

Restrictions
- minimum length of introns
- respect of reading frame
- initial exon: ATG codon
- terminal exon: STOP codon

NetGene2

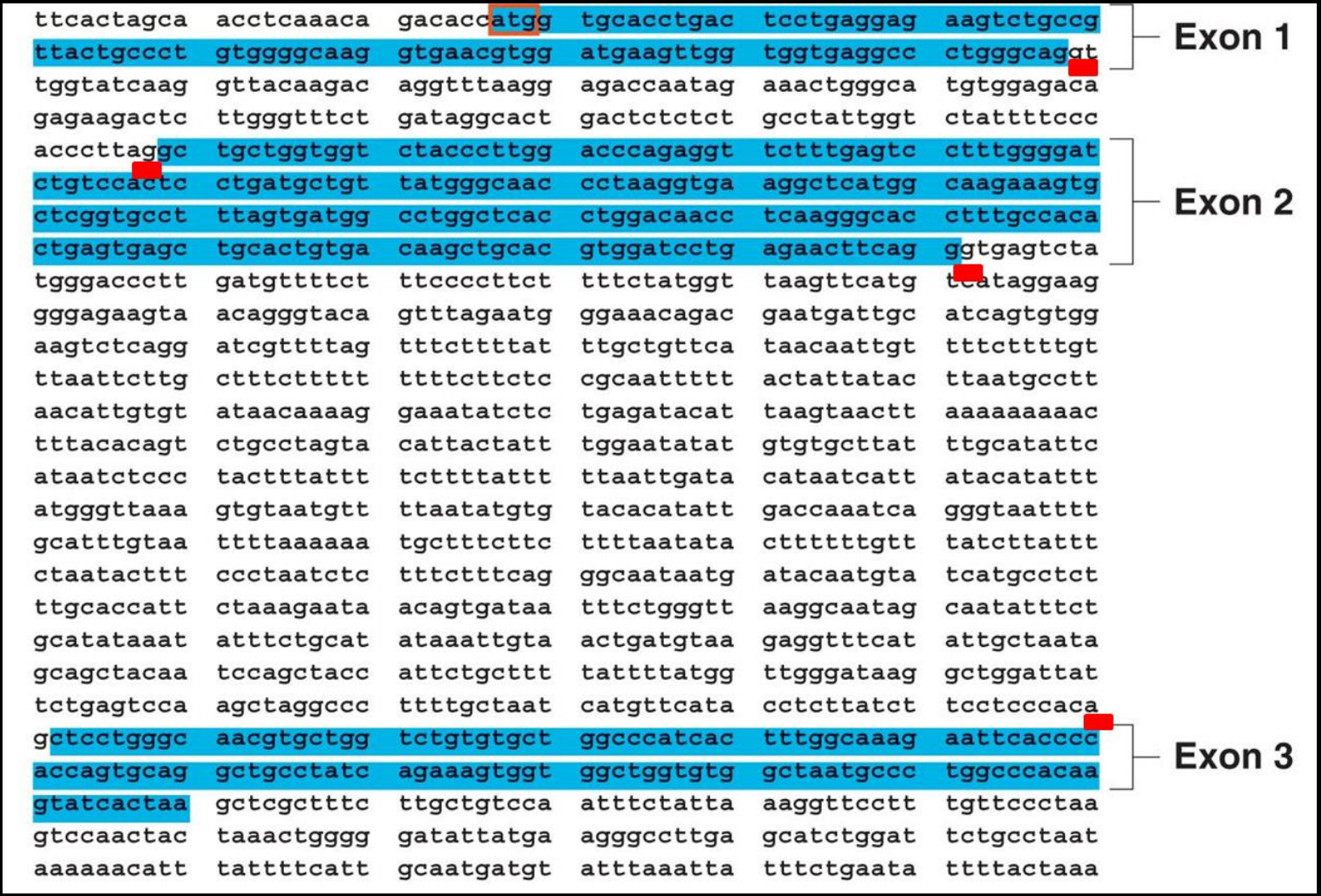


The prediction only regards the coding exons
The UTRs are difficult to predict

▶ CDS

sebastien.aubourg@inra.fr

Exon and intron predictions

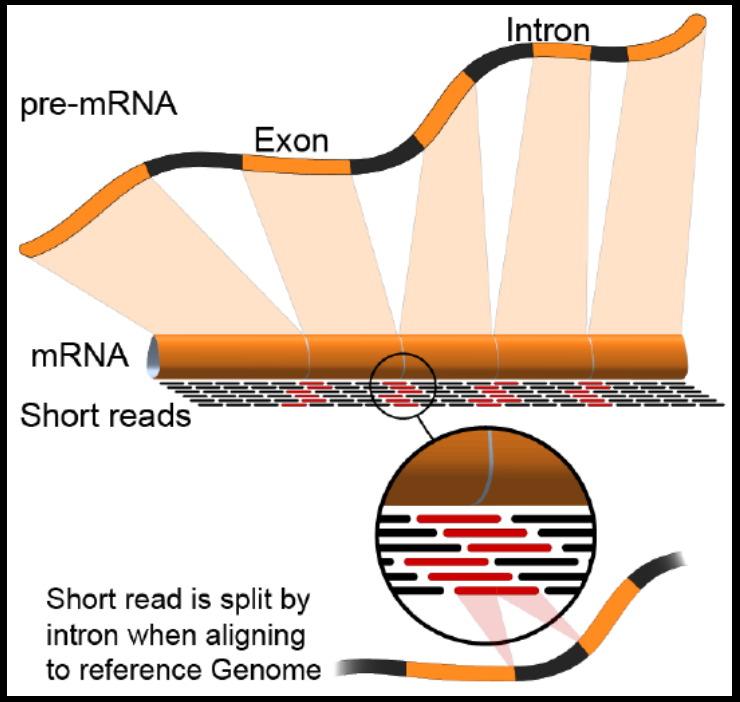


sebastien.aubourg@inra.fr

Transcripts and genomic sequences

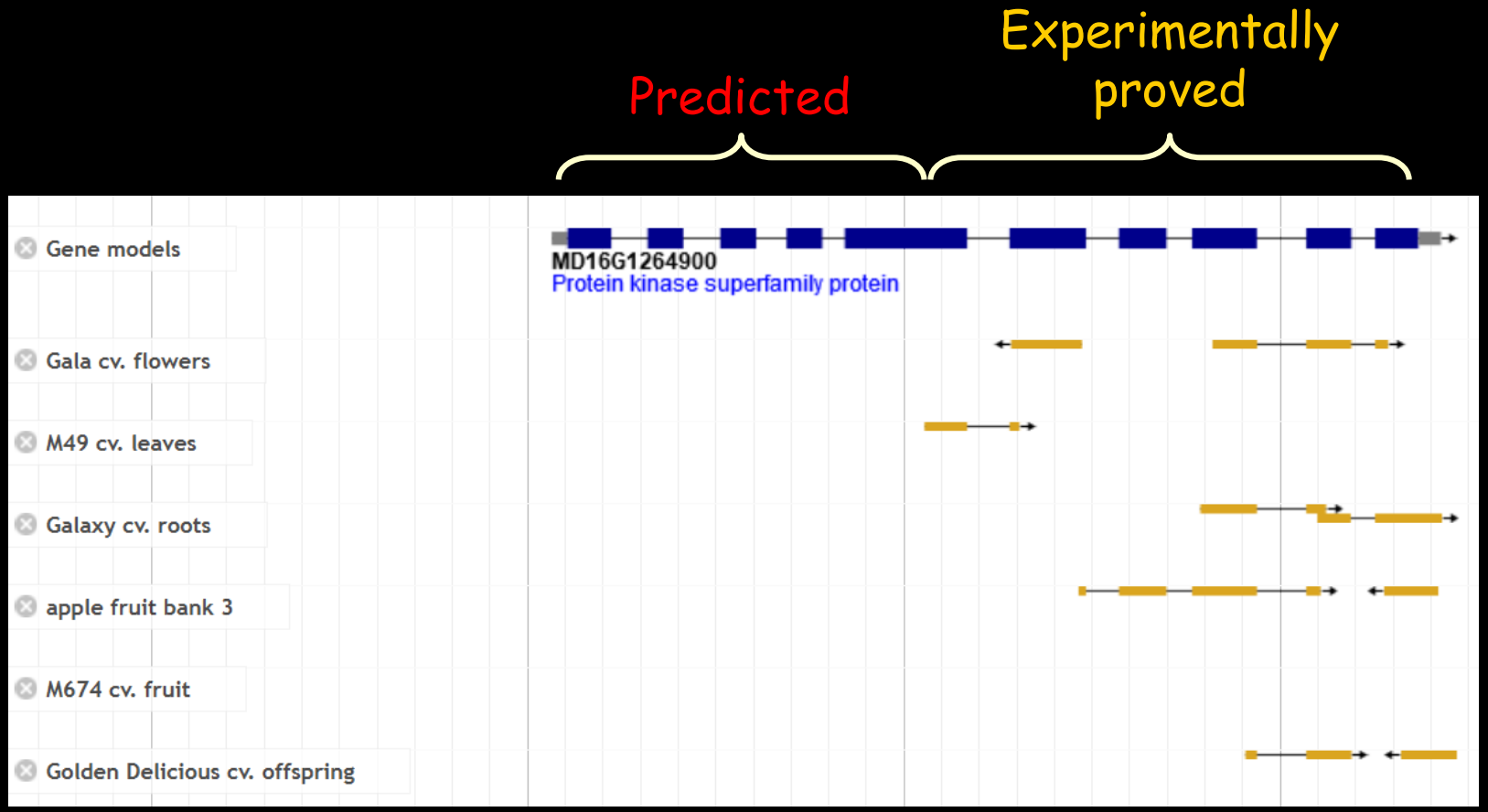
Alignment of cDNA, EST or consensus of short read contigs with the cognate genomic sequence (gmap, star, sim4, spidey…) :

- Spliced full length alignment
- Deal with low quality sequences
- Take into account known splicing sites



sebastien.aubourg@inra.fr

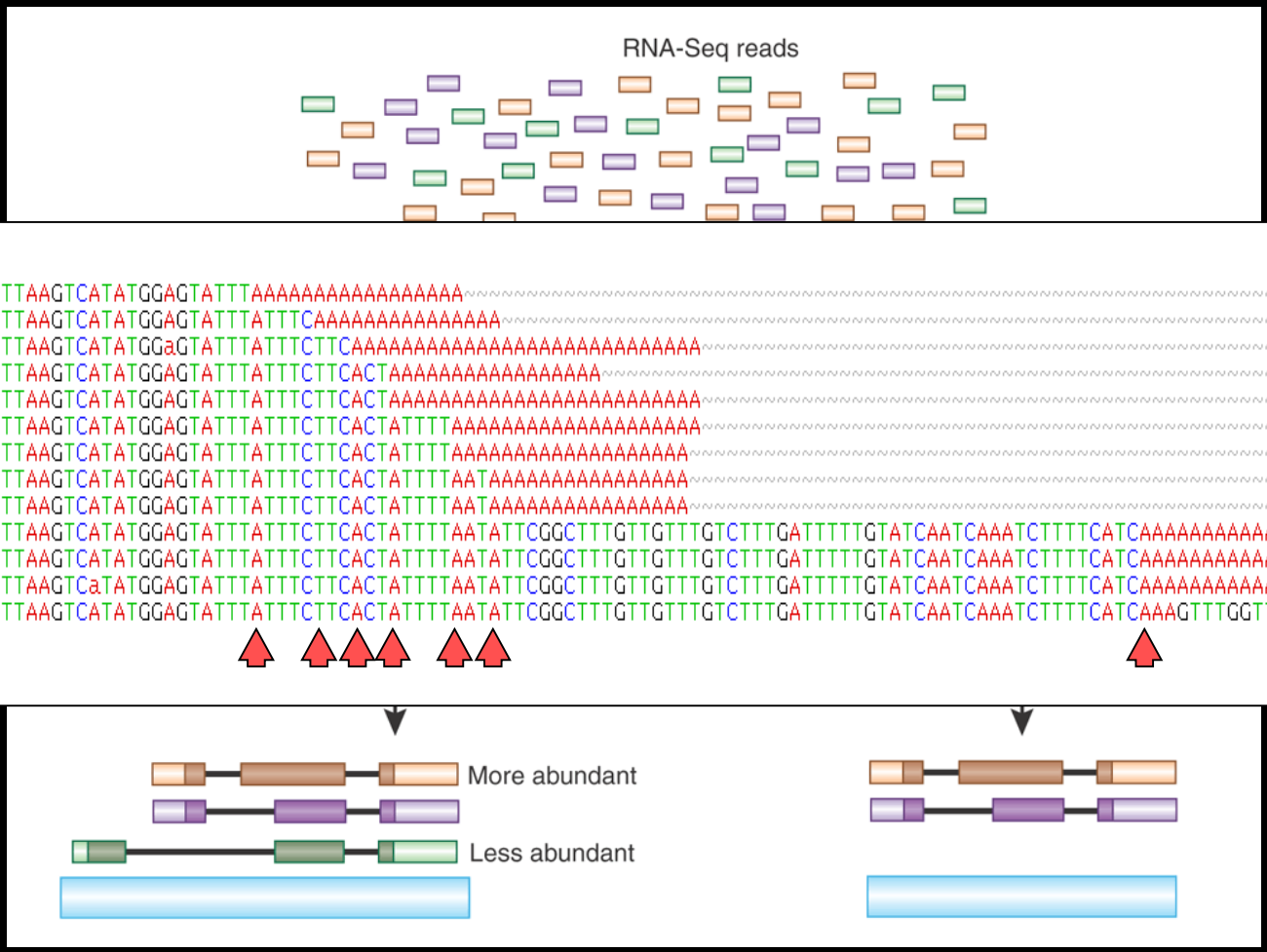
Transcripts and genomic sequences



Heterogeneous quality of structural annotation inside a gene !

sebastien.aubourg@inra.fr

Transcript mapping / de novo assembly



Assembly goals

- more informative consensus sequence
- reduction of complexity
- remove sequencing errors
- highlight gene structure and alternative events

sebastien.aubourg@inra.fr

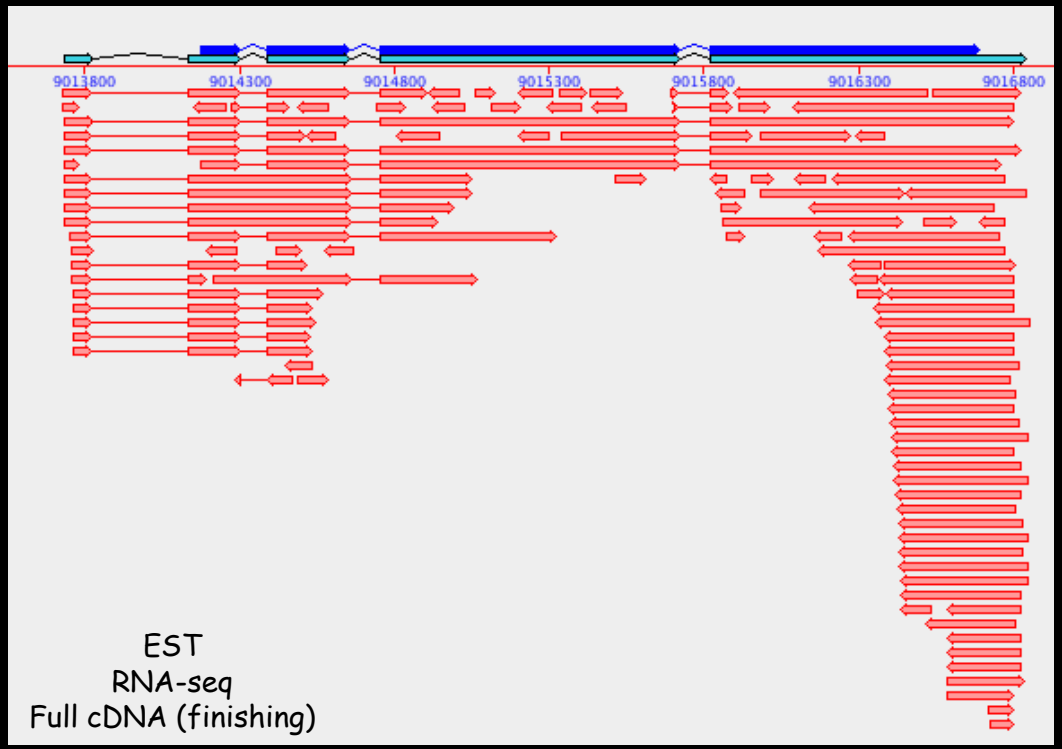
Transcript exploitation

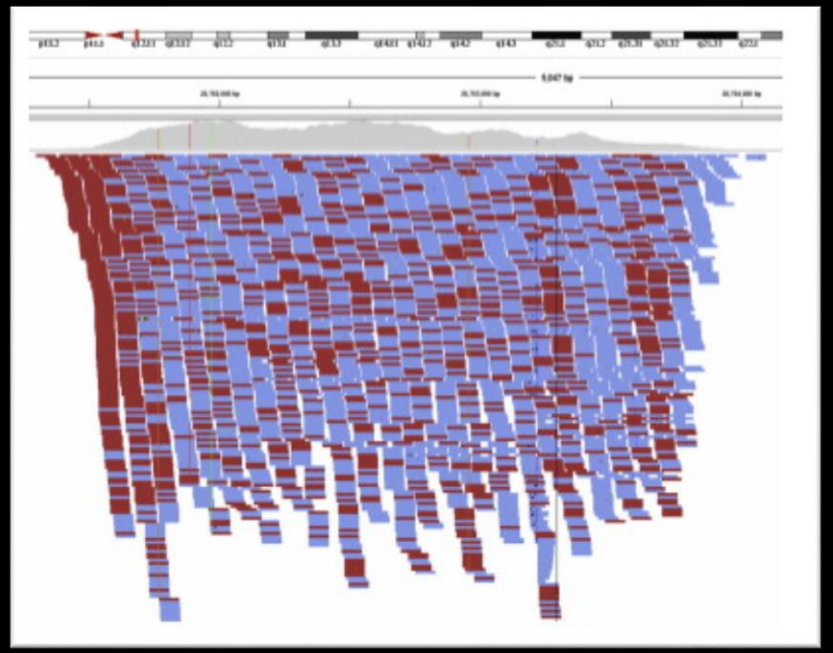
Qualitative

Proof of gene transcription in a specific condition / intron-exon structure / UTRs

Quantitative

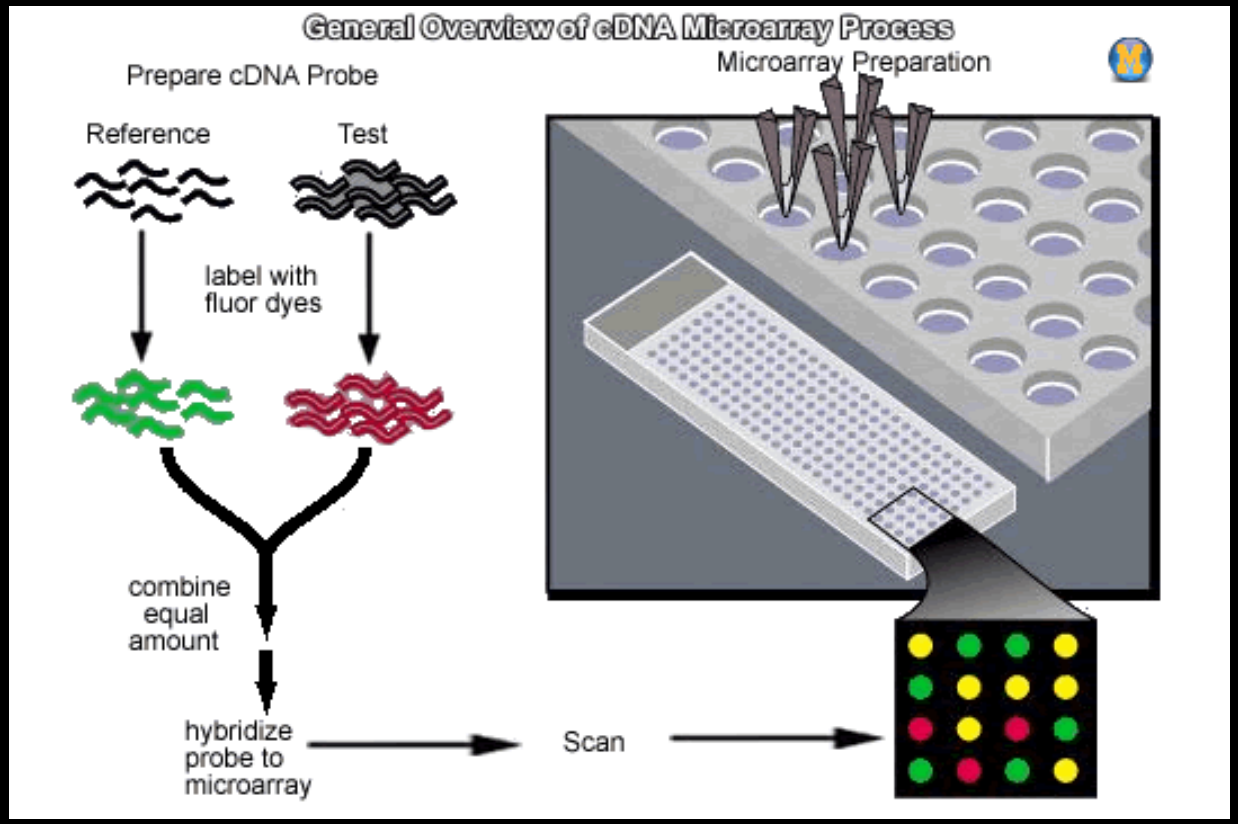
The number of cognate reads reflects the relative proportion of transcripts in the biological sample used

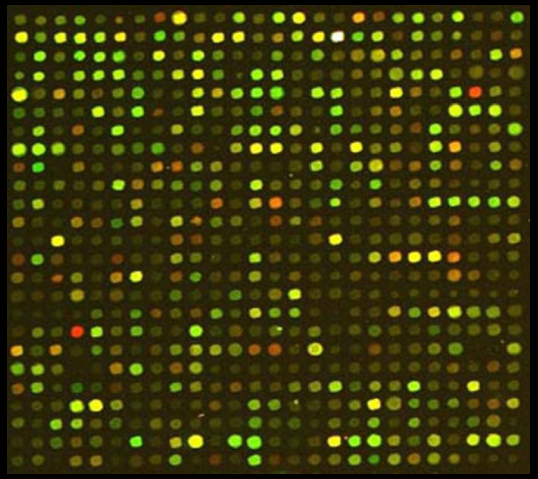




sebastien.aubourg@inra.fr

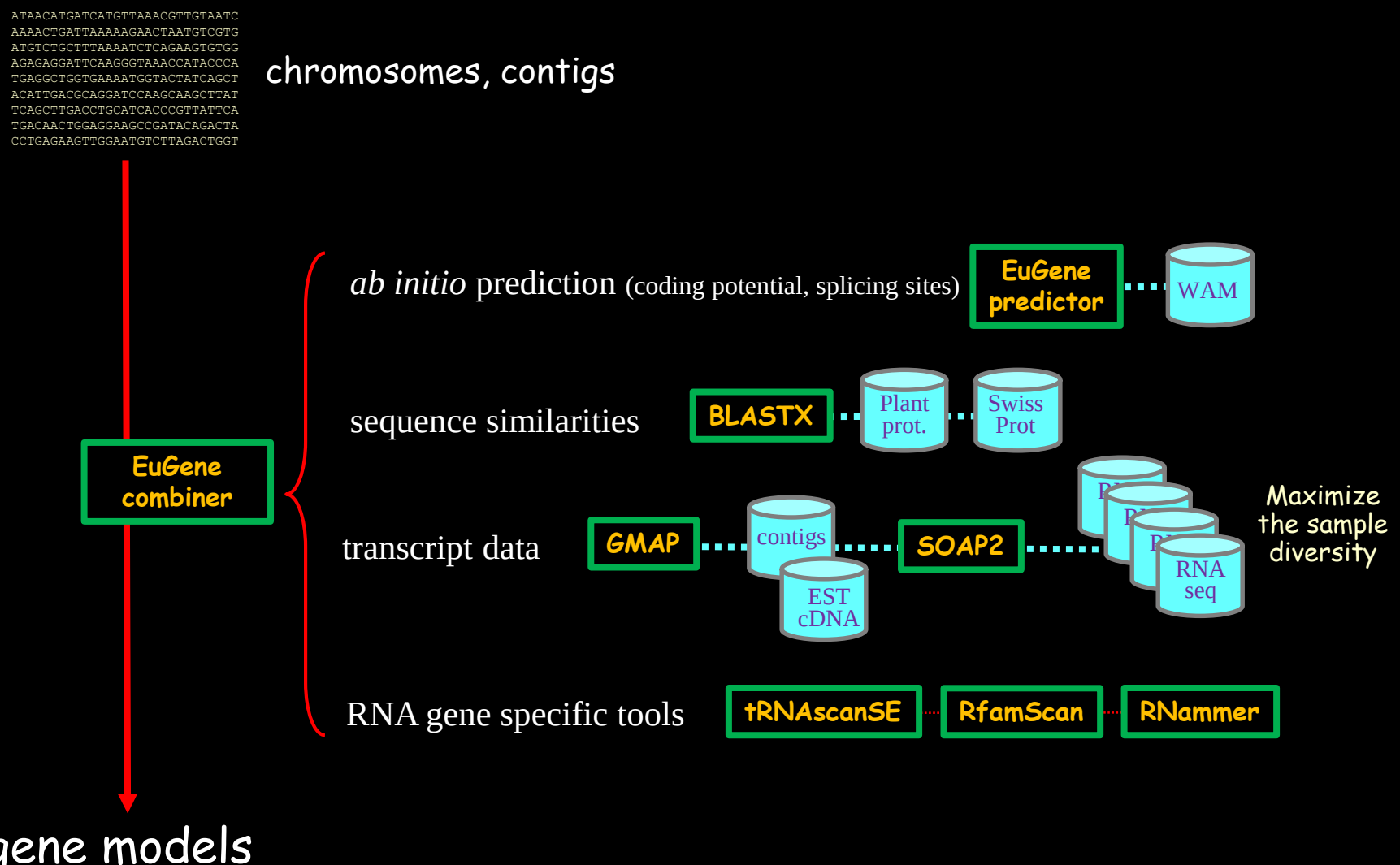
Expression analysis with micro-array





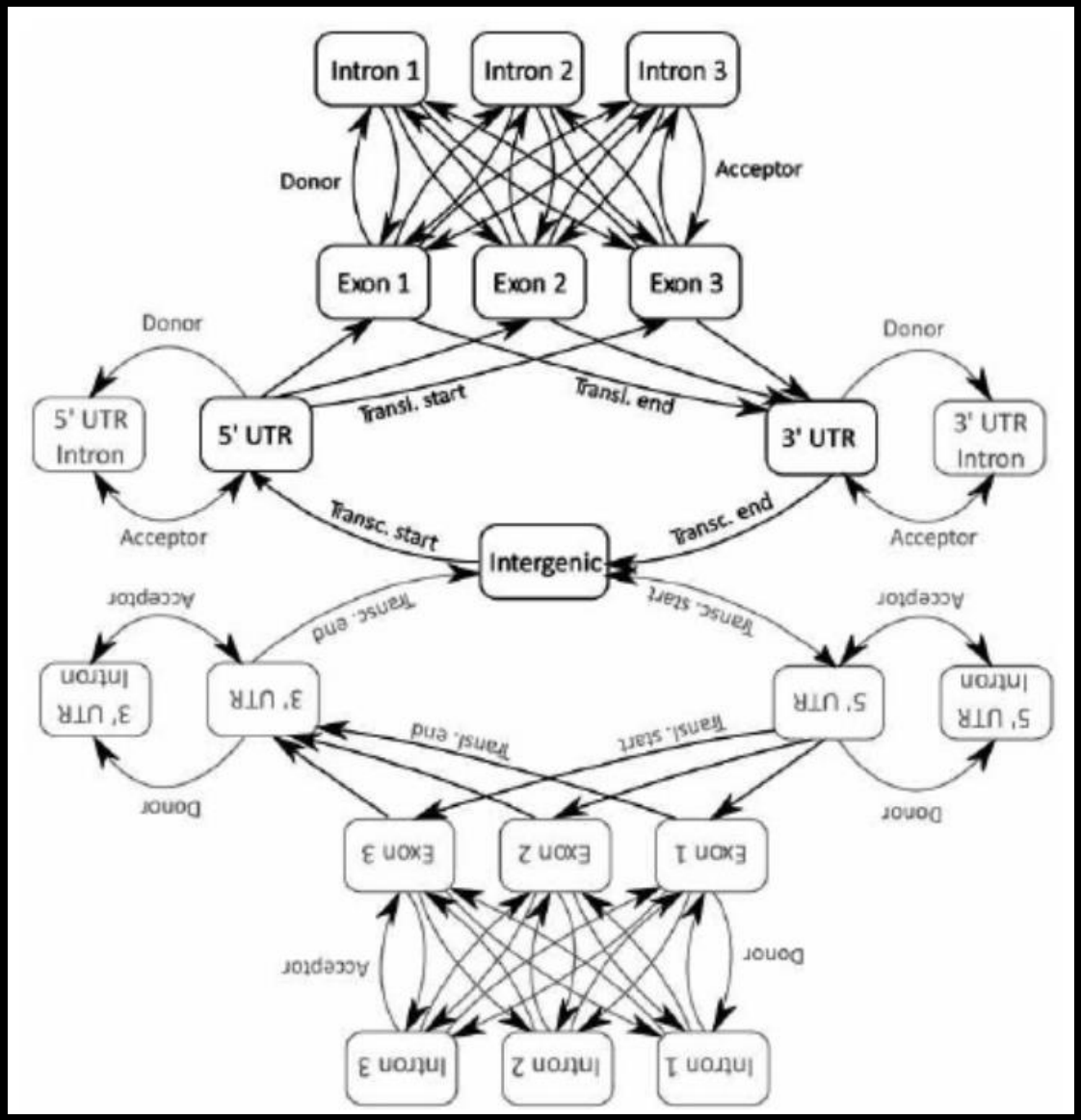
sebastien.aubourg@inra.fr

EuGene: a complete pipeline for gene modeling



sebastien.aubourg@inra.fr

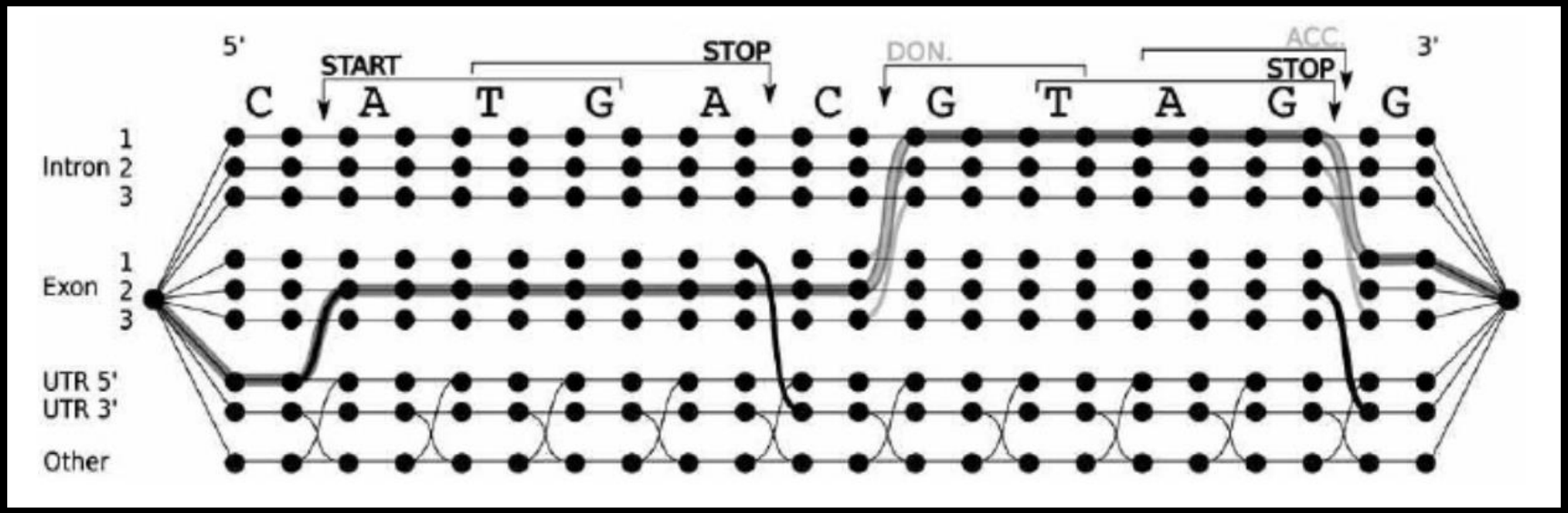
Gene modeling : the states of a genomic sequence



EuGene (T. Schiex)

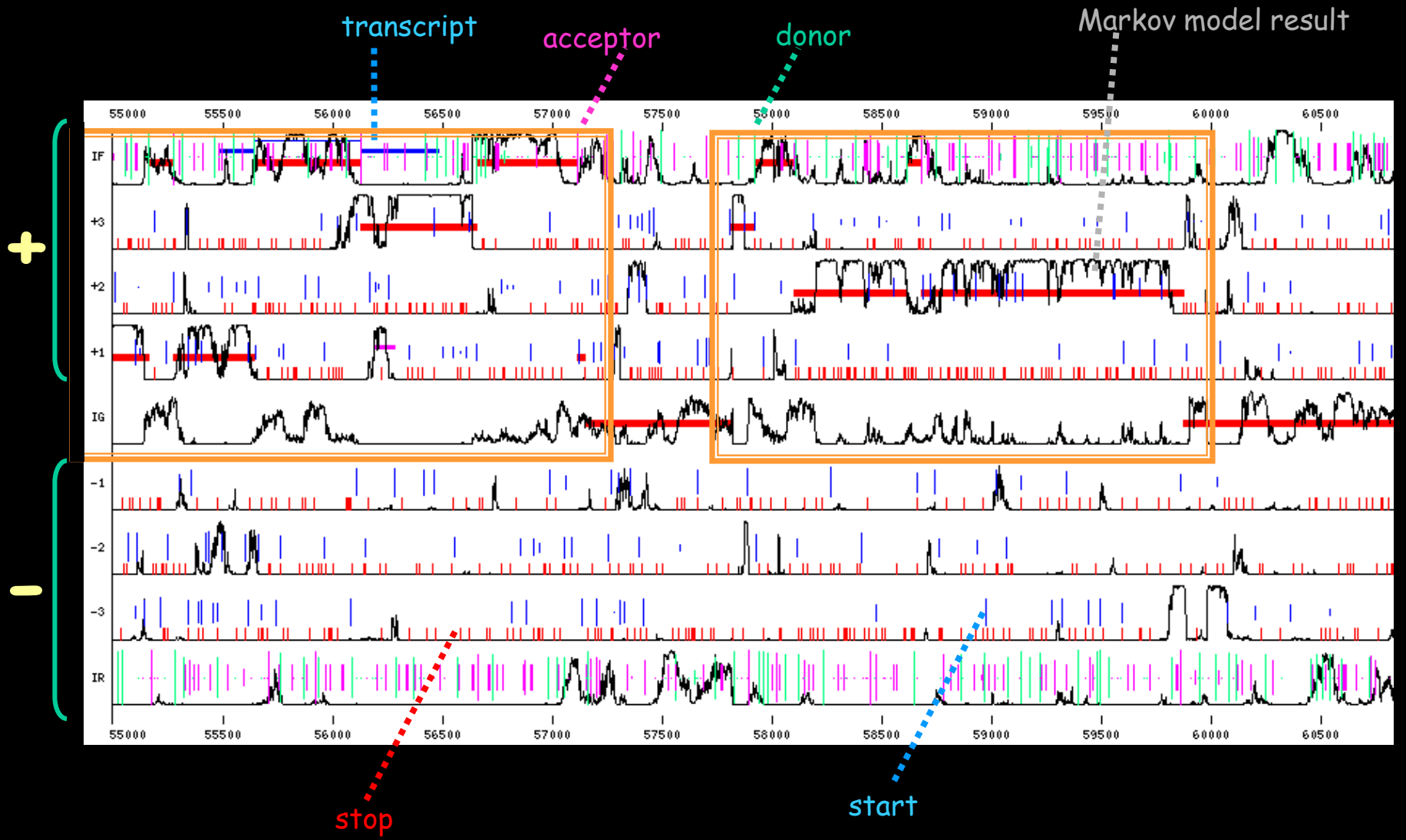
sebastien.aubourg@inra.fr

Gene modeling : the states of a genomic sequence



Eugene (T. Schiex)

Integration for gene prediction



sebastien.aubourg@inra.fr

Examples of gene prediction tools

GeneMark.HMM
Glimmer
GeneFinder
EuGene
FgeneSH
Augustus
Genie
HMMgene
MagPie
GeneId
Grail
GeneScan
GeneWise
…

~ 70 predictors

sebastien.aubourg@inra.fr

Problems and errors: where and why ?

Gene extremities are difficult to predict

- ✓ No or low differences between intergenic and UTR regions
- ✓ in 3': polyadenylation signals are rare or degenerated (AAUAAA)
- ✓ in 5': no systematic CpG islands and/or TATA box
- ✓ intergenic regions can be short and introns can be very long
- ✓ N- and C-terminal regions are often lowly conserved

► Gene breaking and gene merging

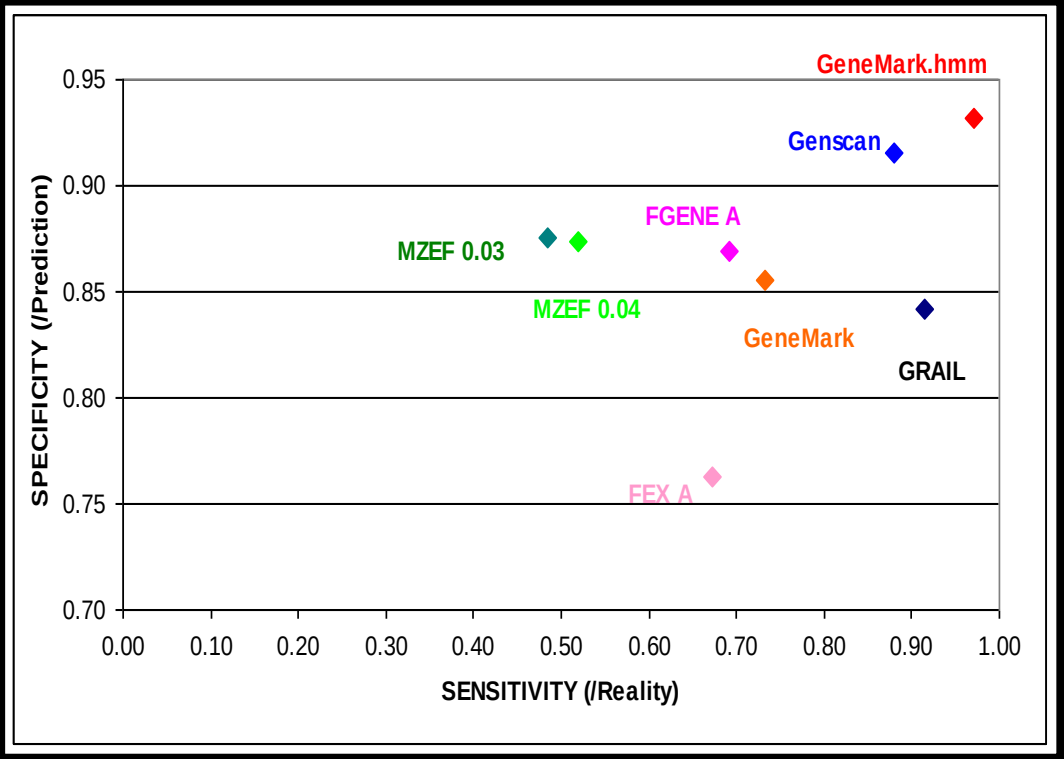
Bottlenecks of spliced alignments

- ✓ Assembly step can produce chimeric transcripts
- ✓ Very short exons can fail the alignment process (<10 bp)
- ✓ Mainly GT::AG sites are looked for to open gaps
- ✓ Sequencing of partially spliced transcripts

sebastien.aubourg@inra.fr

Evaluation and efficiency of gene predictors

true positive (TP): reality well predicted
true negative (TN): false predict to be false
false positive (FP): false predict to be true
false negative (FN): reality not predicted



$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

Fraction of reality that is well predicted

$$\text{Specificity} = \frac{TP}{TP+FP}$$

Fraction of the prediction that is true

…at different levels: splicing sites, exons or genes

sebastien.aubourg@inra.fr

Evaluation of genome annotation

BUSCO v3

Benchmarking Universal Single-Copy Orthologs
Plant dataset : 1440 genes

Sunflower : 92%
Medicago : 95,4%
Rosa : 96,5%
Apple : 96,8%

complete	**96,8 %**
fragmented	1,3 %
missing	1,9 %
transcript support	89,8 %
PFAM signature	74,8 %
TAIR or SwissProt homolog	89,6 %

U : 93,2 %

expertise of 1194 genes (50 families)

CDS ok	site	ATG	exons	split	merge	underpred	pseudo
753	**18**	**55**	**54**	**23**	**20**	**32**	**239**
78,8 %	**1,9 %**	**5,8 %**	**5,7 %**	**2,4 %**	**2,1 %**	**3,4 %**	**20 %**

955 functional genes

sebastien.aubourg@inra.fr

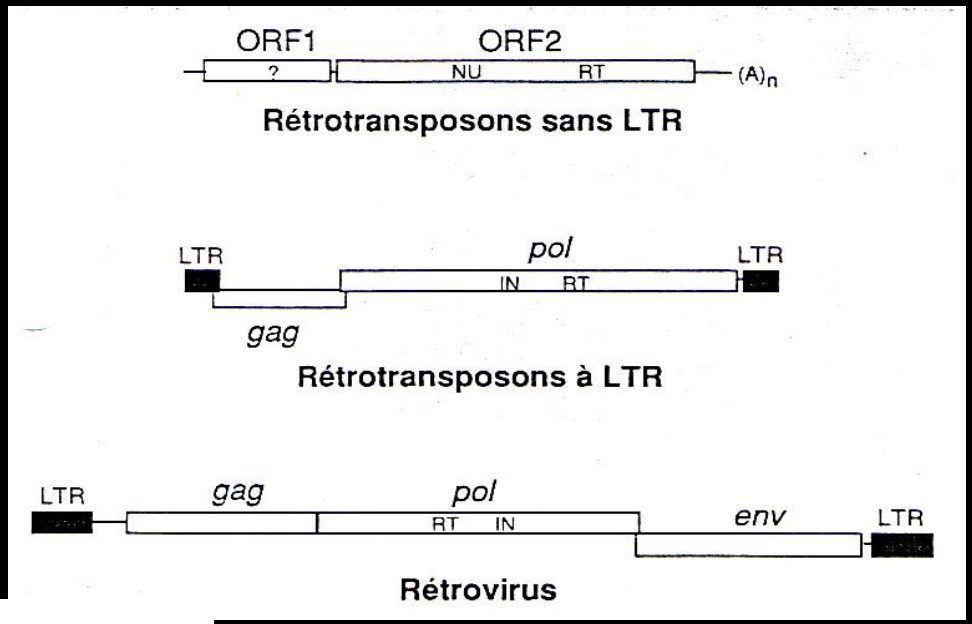
Annotation of transposable elements

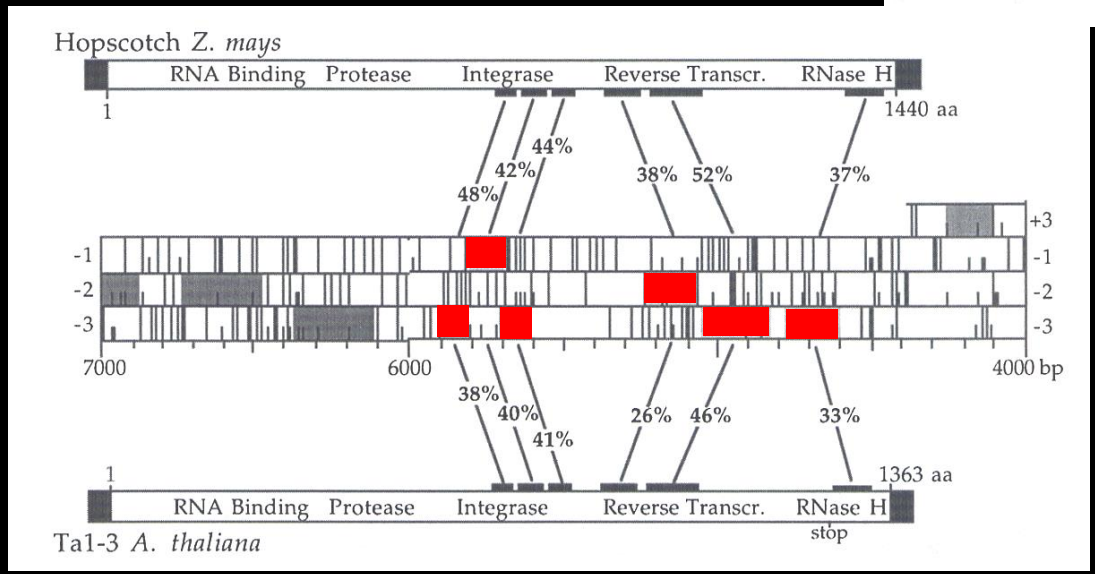
RepeatMasker

- repeated sequences
- Transposables elements

(RepBase)

Repet pipeline

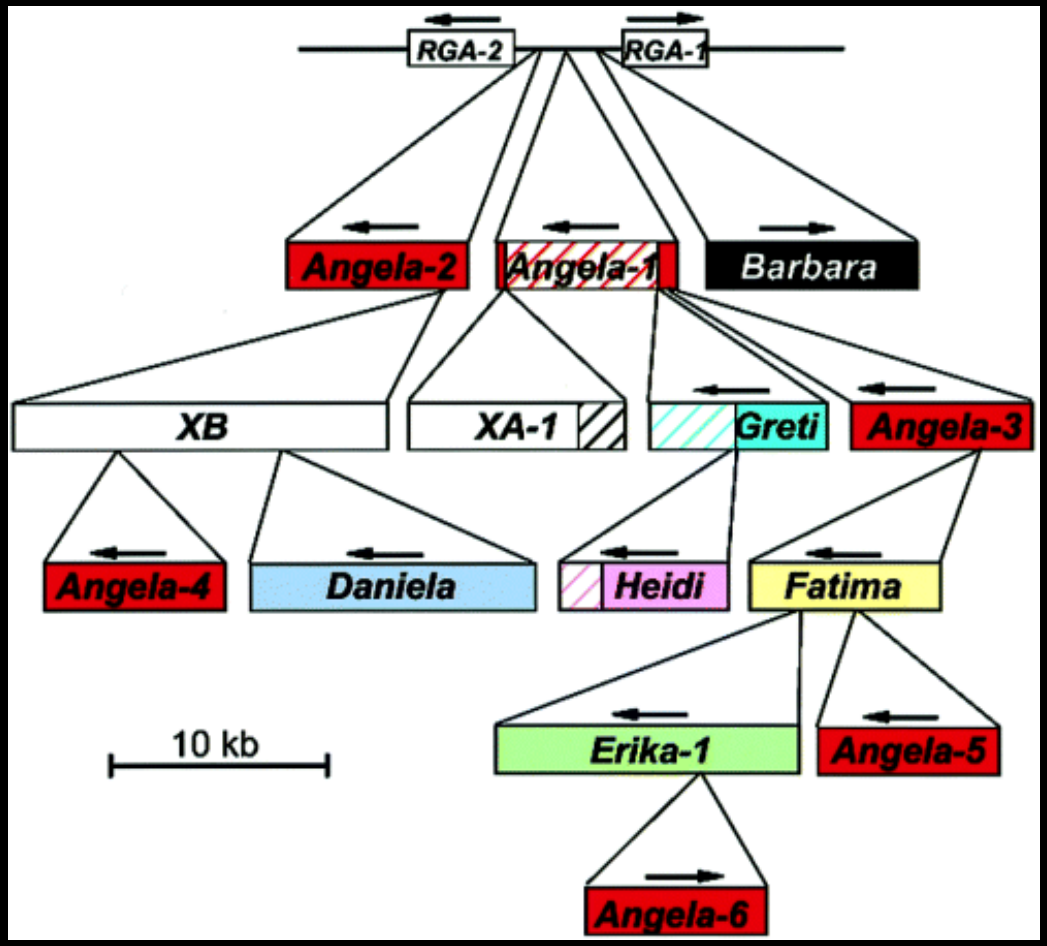




Quesneville et al.
Combined evidence annotation of transposable elements in genome sequences.
PLoS Comput Biol. 2005 Jul;1(2):e22
PMID: 16110336

sebastien.aubourg@inra.fr

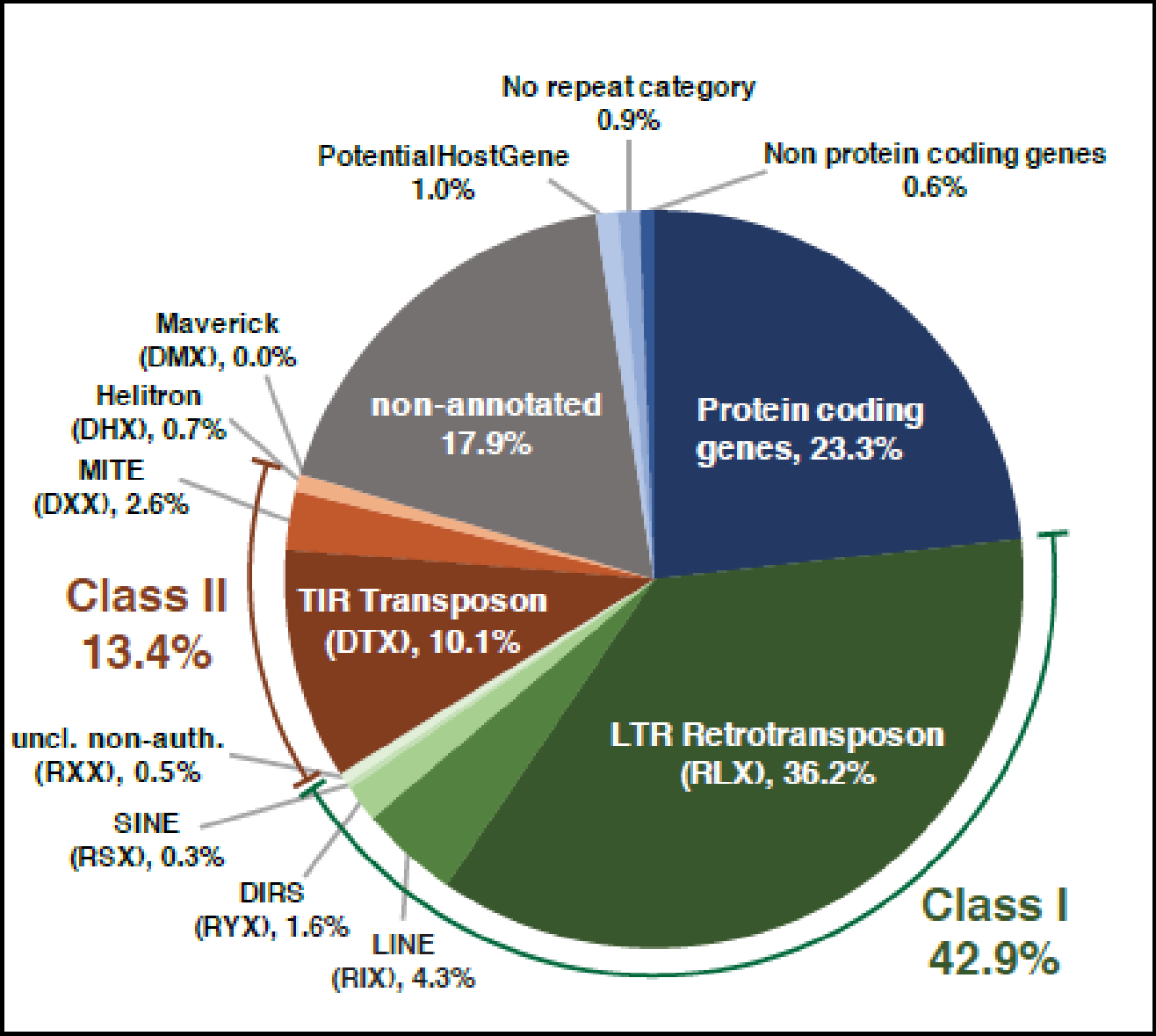
Annotation of transposable elements: evolution



Wicker et al. (wheat, 2001)

sebastien.aubourg@inra.fr

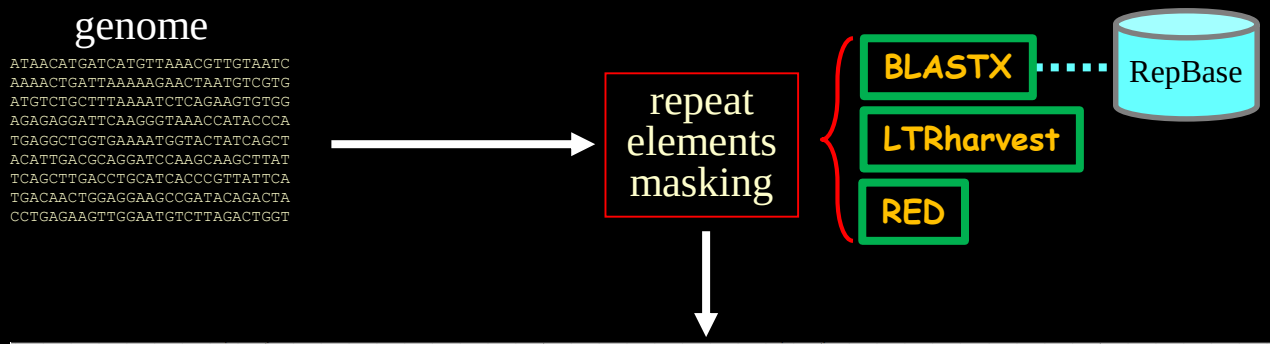
TE overview in the apple genome

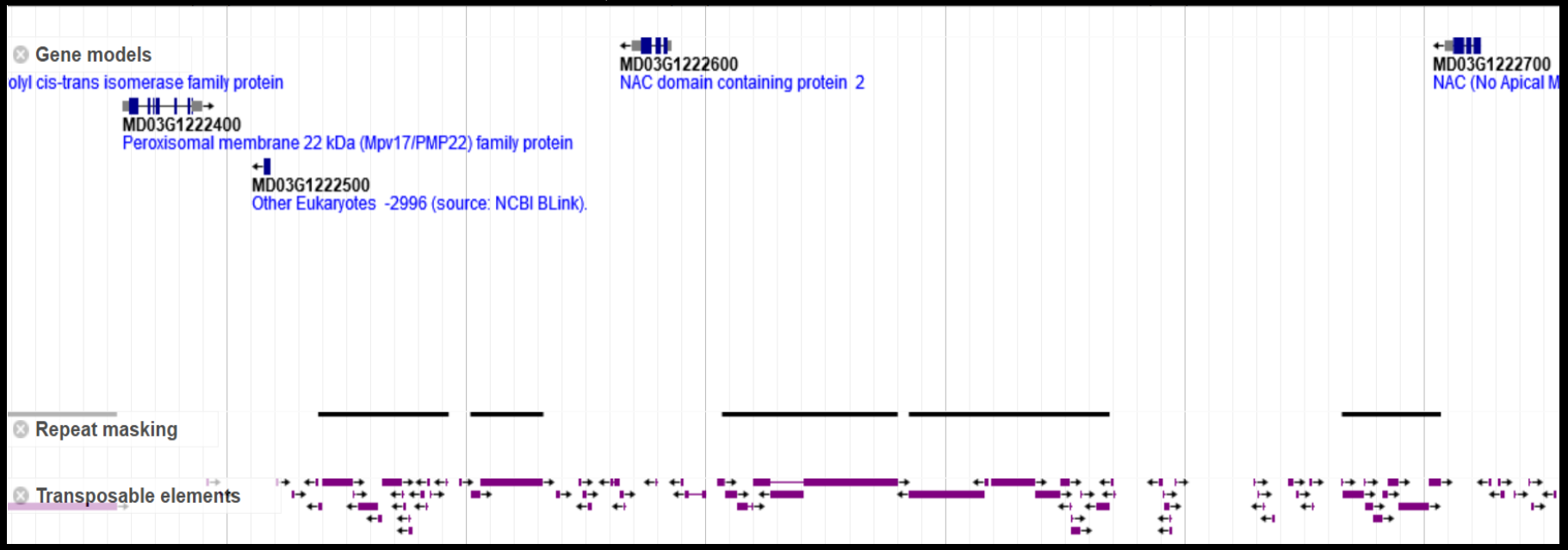


sebastien.aubourg@inra.fr

Repeat masking

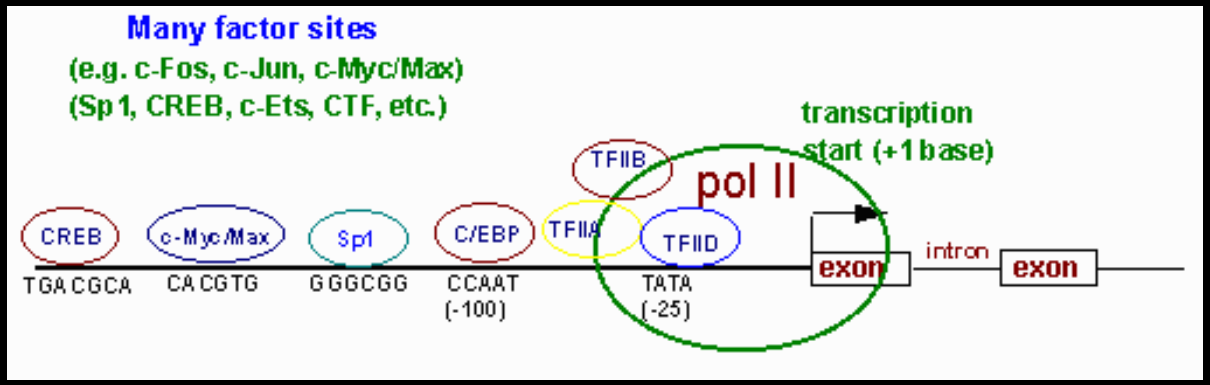
Goals : to avoid over-prediction of genes in
- disrupted coding regions of TE (complete or not)
- low complexity regions





sebastien.aubourg@inra.fr

Annotation of promoters



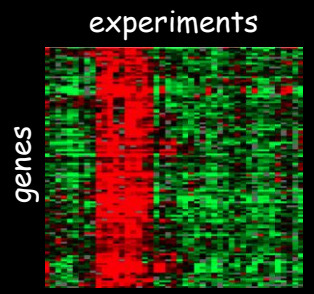
- ✓ short and degenerated motifs ► numerous false positives
- ✓ Transcription Start Sites (TSS) identification are based on transcripts
- ✓ No systematic CpG islands and/or TATA box

► Never done at the genome scale

sebastien.aubourg@inra.fr

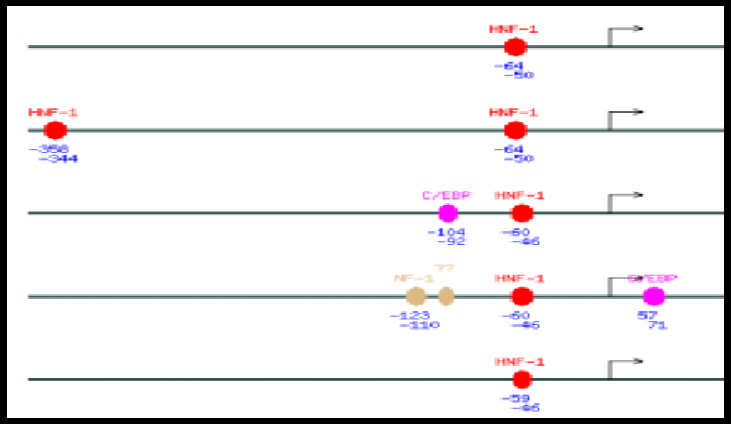
Annotation of promoters in co-expression clusters

Transcriptomics



Clustering

Gibbs Sampling



- TRANSFAC
- MOTIF SAMPLER (TOUCAN package)
- RSAT
- MATINSPECTOR
- VISTA
- PROMO
- ALIGNACE
- MEME
- SCPD (yeast)
- PLACE (plant)
- PLANTCARE (plant)

sebastien.aubourg@inra.fr

Example of plant genomes

	ab initio	similarities	transcripts	combiner	
 arabidopsis	Genscan+ GeneMark.hmm GlimmerA GeneSplicer	BLAST	PASA BLAT SIM4	AAT COMBINER	**TIGR** **TAIR**
 oryza	Genscan+ GeneMark.hmm GlimmerM FGENESH	BLAST	PASA2	AAT COMBINER	**TIGR**
 populus	EUGENE GrailExp6 GeneWise FGENESH	BLAST S&W	SIM4	JGI R.Syst.	**JGI** **PSB** **ORNL**
 vitis	GeneId GlimmerHMM SNAP EUGENE	EXOFISH BLAST	EST2Genome GeneWise	GAZE	**GENOSCOPE** **URGV**
 medicago	FGENESH EUGENE	BLAST	PASA	EUGENE	**IMGAG**
 malus	EUGENE	BLAST	SOAP2 GMAP	EUGENE	**IRHS**

sebastien.aubourg@inra.fr

Conclusions about structural annotation

- Automatic annotation is quick and provides (preliminary) data wanted by the scientific community

- Numerous errors but not too much missed full coding genes

- The different prediction pipelines have to be known, understood and took into account to read and exploit the annotations

- Other approaches are necessary (by gene family, by experts..)

- Take precautions:

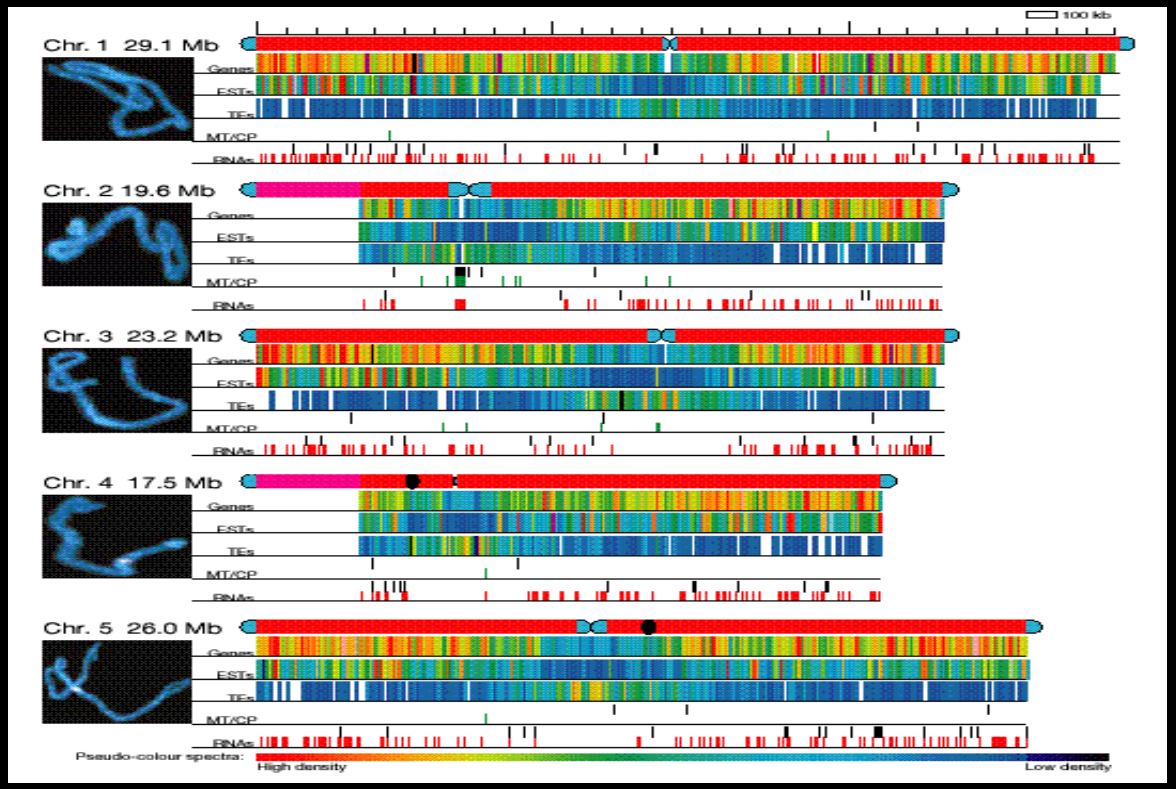
  - **BLASTX, P or PSI (bank of proteins + translations of CDS)**
  - **Prodom, Pfam...**

  Mix of proved and only supposed and predicted data

- Rare and novel situations can not be predicted !

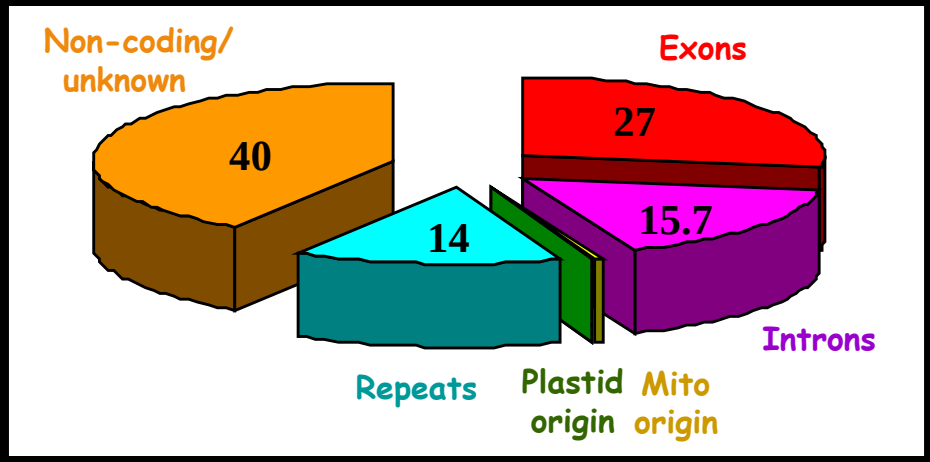
sebastien.aubourg@inra.fr

The Arabidopsis genome



118 997 677 bp
27 665 genes
4 853 pseudogenes + TE

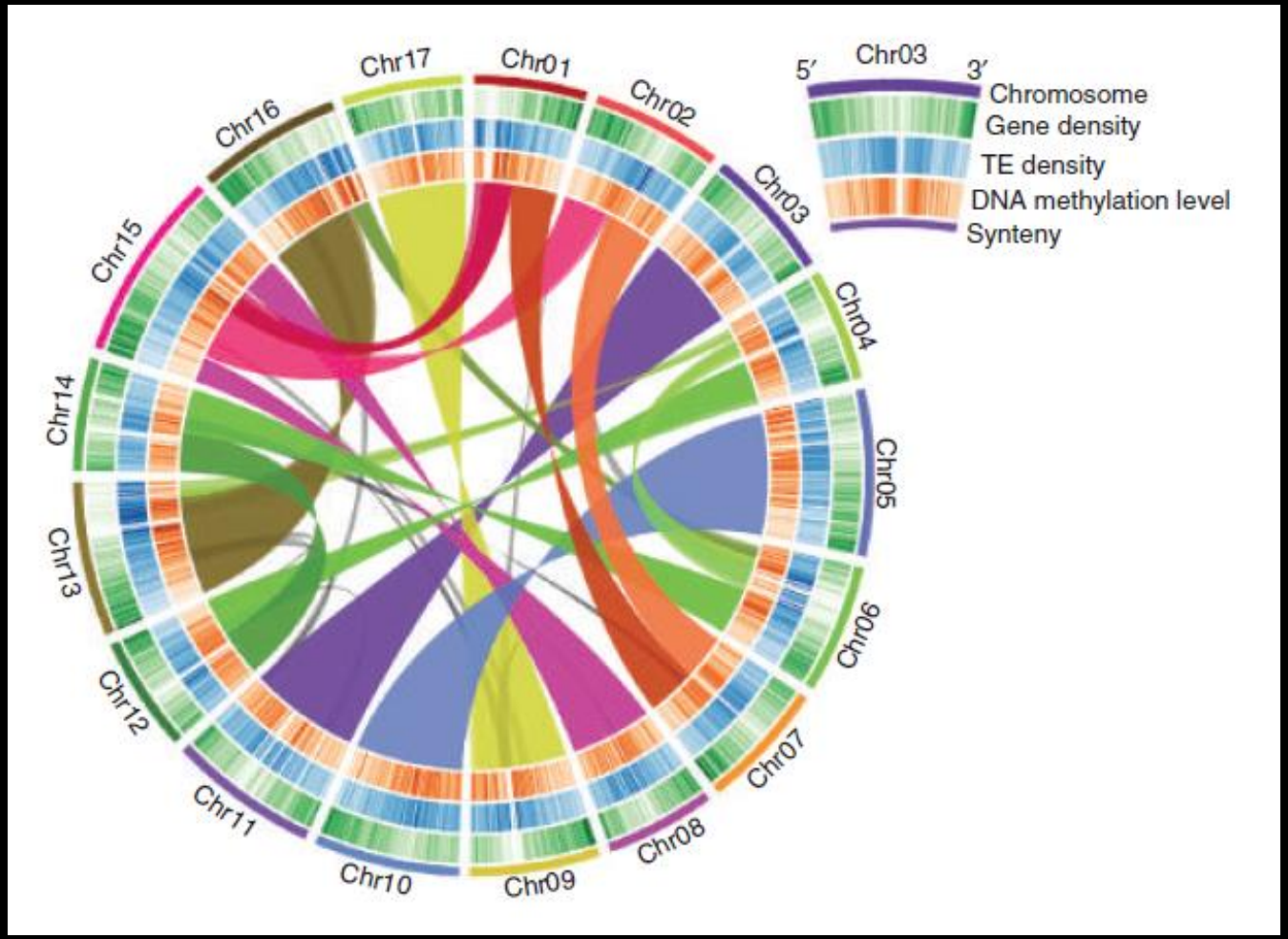
10 695 alternatively spliced genes



sebastien.aubourg@inra.fr

The *Malus domestica* genome

coding genes	**45116**
snoRNA	410
rRNA	4369
tRNA	654
miRNA	141
snRNA	181
other ncRNA	54
undefined ncRNA	1816
total genes	**52741**



sebastien.aubourg@inra.fr

Format of annotations @ GB/EMBL

```
repeat_region   11095..11216
                /rpt_family="(TAAAA)n"
     gene       12166..13245
                /gene="T27C4.4"
                /note="similar to ribosomal protein L17 GB:AAA34113.1 from
                [Nicotiana tabacum]"
     mRNA       join(12166..12188,12416..12700,12785..12826,13063..13245)
                /gene="T27C4.4"
     CDS        join(12176..12188,12416..12700,12785..12826,13063..13145)
                /gene="T27C4.4"
                /codon_start=1
                /product="ribosomal protein L17, putative"
                /protein_id="AAF63771.1"
                /db_xref="GI:7547099"
                /translation="MSKRGRGGTSGNKFRMSLGLPVAATVNCADNTGAKNLYIISVKG
                IKGRLNRLPSACVGDMVMATVKKGKPDLRKKVLPAVIVRQRKPWRRKDGVFMYFEDNA
                GVIVNPKGEMKGSAITGPIGKECADLWPRIASAANAIV"
     gene       <13674..>14042
                /gene="T27C4.5"
                /note="predicted by genscan+"
     mRNA       join(<13674..13775,13914..>14042)
                /gene="T27C4.5"
     CDS        join(13674..13775,13914..14042)
                /gene="T27C4.5"
                /note="hypothetical protein"
                /codon_start=1
                /protein_id="AAF63772.1"
                /db_xref="GI:7547100"
                /translation="MVNPVGFRFRPTKEEIVDHYLRPTNFDGDTSHVDRNIMFMQDNR
                NDYRPPNSLTGVFSDCSSDDNDSDLLSPKTVS"
     repeat_region   complement(15431..15636)
                /note="Limpet1 transposon GB U76697"
```

sebastien.aubourg@inra.fr

Format of annotations @ GB/EMBL

```
repeat_region   11095..11216
                /rpt_family="(TAAAA)n"
gene            12166..13245
                /gene="T27C4.4"
                /note="similar to ribosomal protein L17 GB:AAA34113.1 from
                [Nicotiana tabacum]"
mRNA            join(12166..12188,12416..12700,12785..12826,13063..13245)
                /gene="T27C4.4"
CDS             join(12176..12188,12416..12700,12785..12826,13063..13145)
                /gene="T27C4.4"
                /codon_start=1
                /product="ribosomal protein L17, putative"
                /protein_id="AAF63771.1"
                /db_xref="GI:7547099"
                /translation="MSKRGRGGTSGNKFRMSLGLPVAATVNCADNTGAKNLYIISVKG
                IKGRLNRLPSACVGDMVMATVKKGKPDLRKKVLPAVIVRQRKPWRRKDGVFMYFEDNA
                GVIVNPKGEMKGSAITGPIGKECADLWPRIASAANAIV"
gene            <13674..>14042
                /gene="T27C4.5"
                /note="predicted by genscan+"
mRNA            join(<13674..13775,13914..>14042)
                /gene="T27C4.5"
CDS             join(13674..13775,13914..14042)
                /gene="T27C4.5"
                /note="hypothetical protein"
                /codon_start=1
                /protein_id="AAF63772.1"
                /db_xref="GI:7547100"
                /translation="MVNPVGFRFRPTKEEIVDHYLRPTNFDGDTSHVDRNIMFMQDNR
                NDYRPPNSLTGVFSDCSSDDNDSDLLSPKTVS"
repeat_region   complement(15431..15636)
                /note="Limpet1 transposon GB U76697"
```

Feature

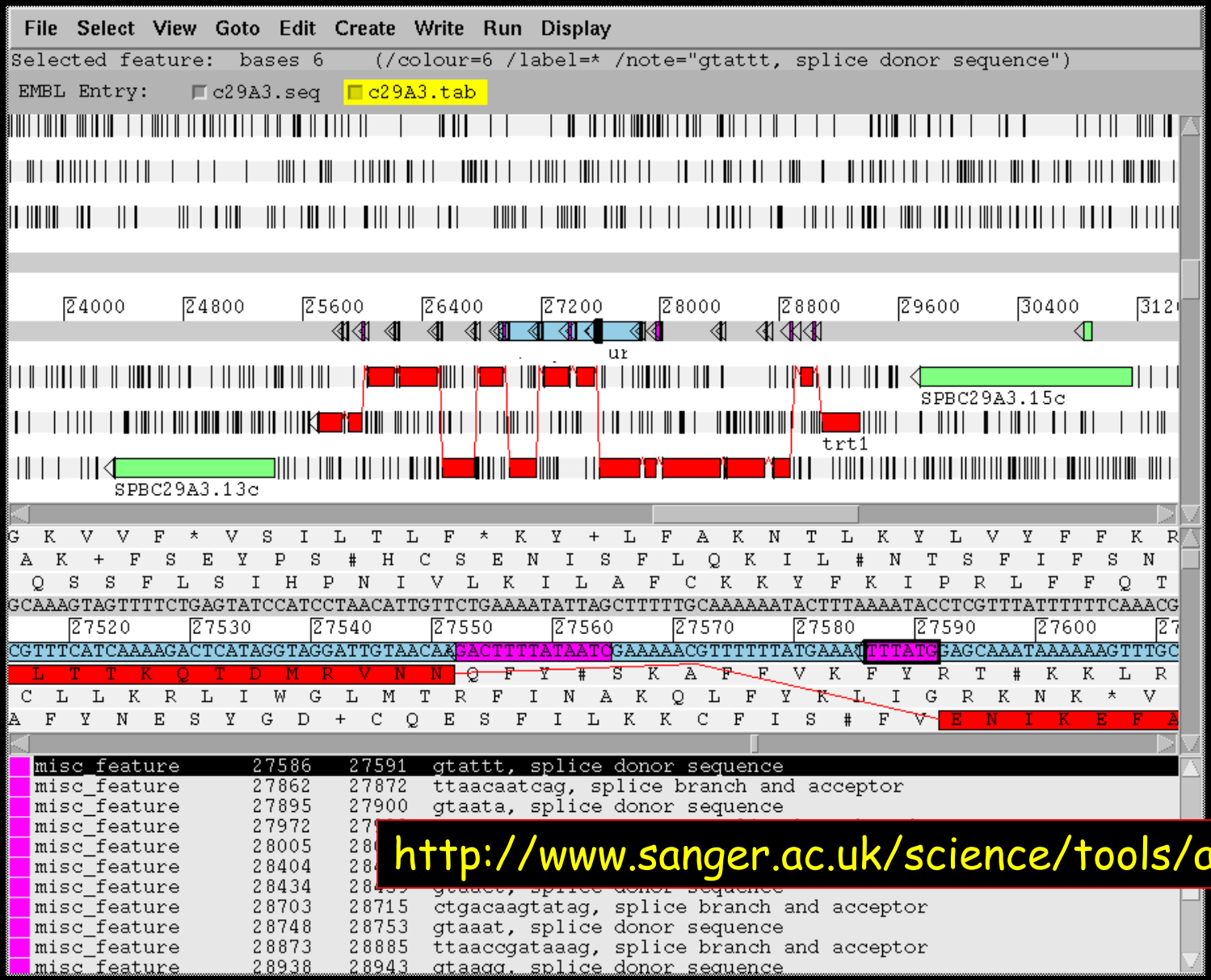
/Qualifier

sebastien.aubourg@inra.fr

Genome Browser

sebastien.aubourg@inra.fr

Annotation tool



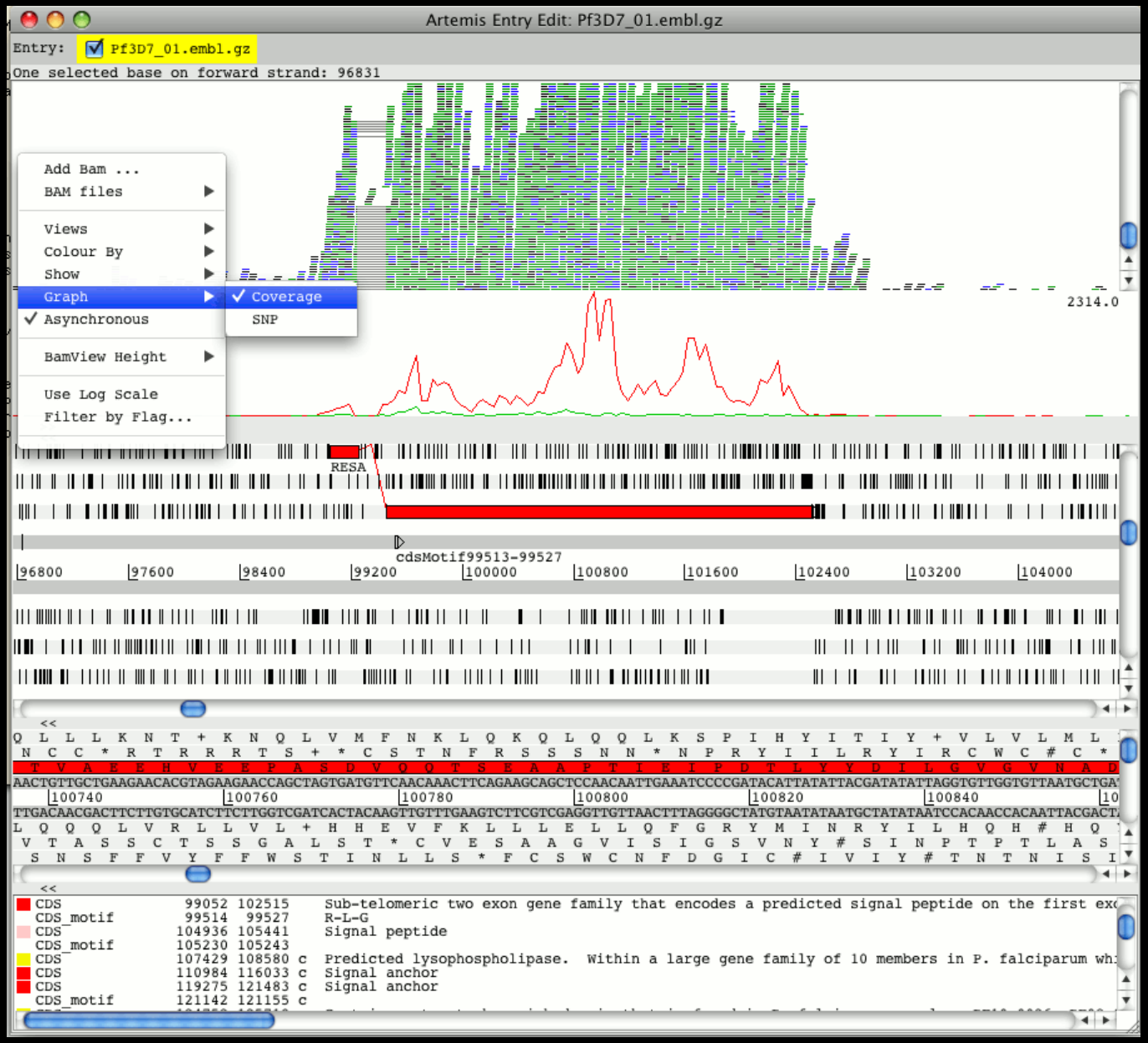
Artemis

Sanger

http://www.sanger.ac.uk/science/tools/artemis

sebastien.aubourg@inra.fr

Genome browser and annotation tool

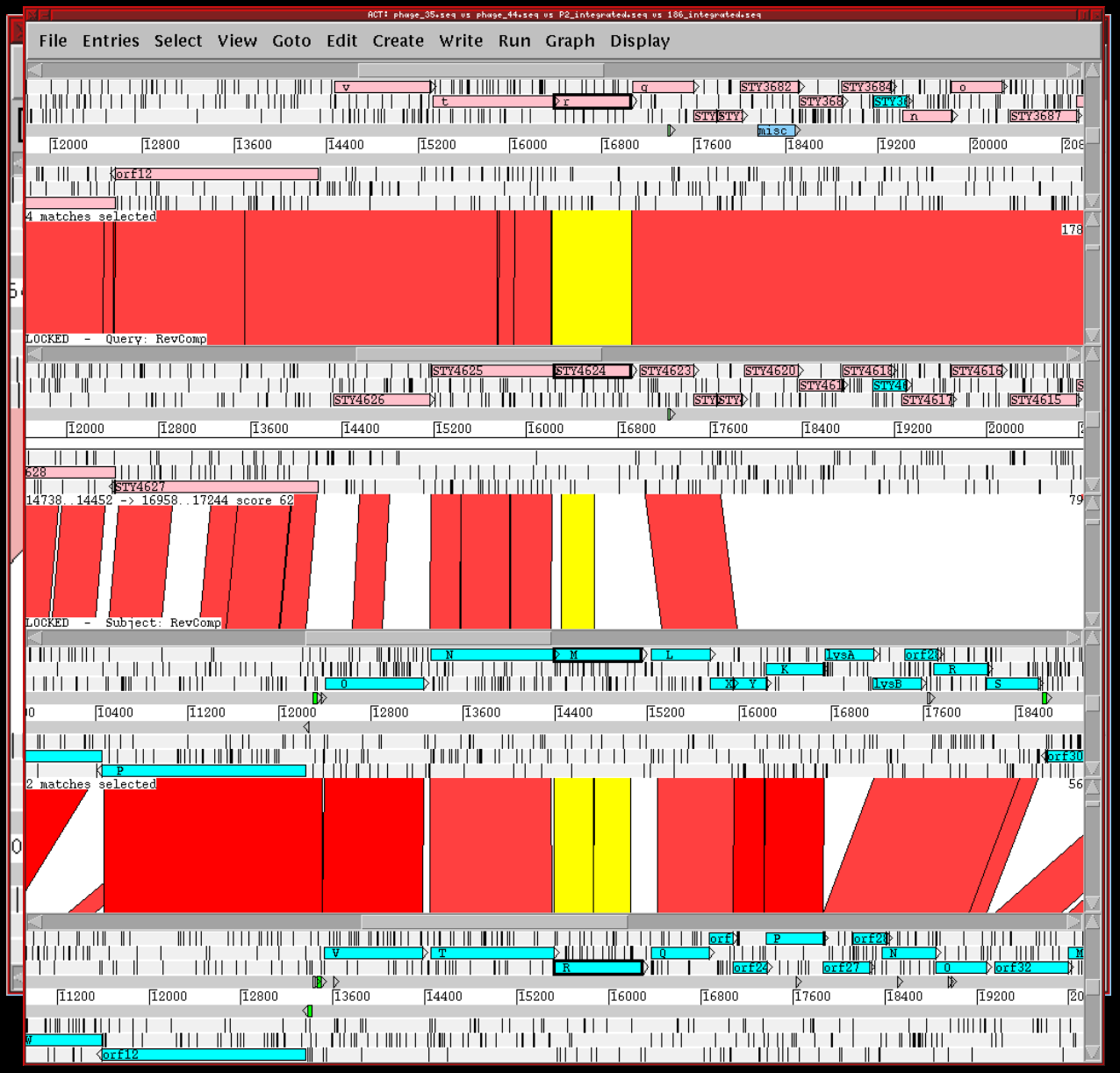


Plug-in BamView

RNA-seq integration

sebastien.aubourg@inra.fr

Comparative genomics for annotation



Plug-in
ACT

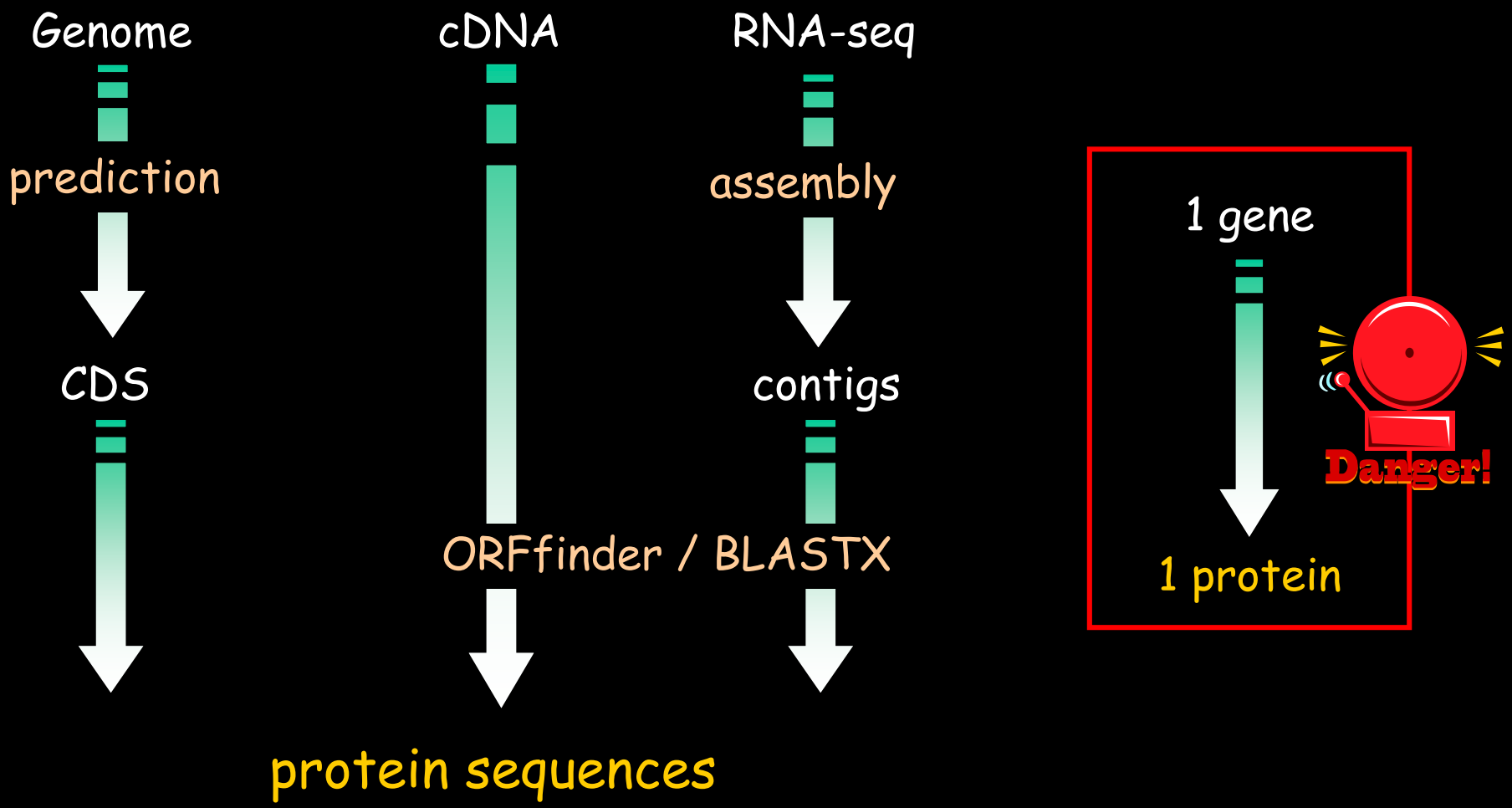
sebastien.aubourg@inra.fr

Genome Annotation

Function

sebastien.aubourg@inra.fr

Translation : from nucleotides to proteins



sebastien.aubourg@inra.fr

Automatic functional annotation

Prediction of protein function

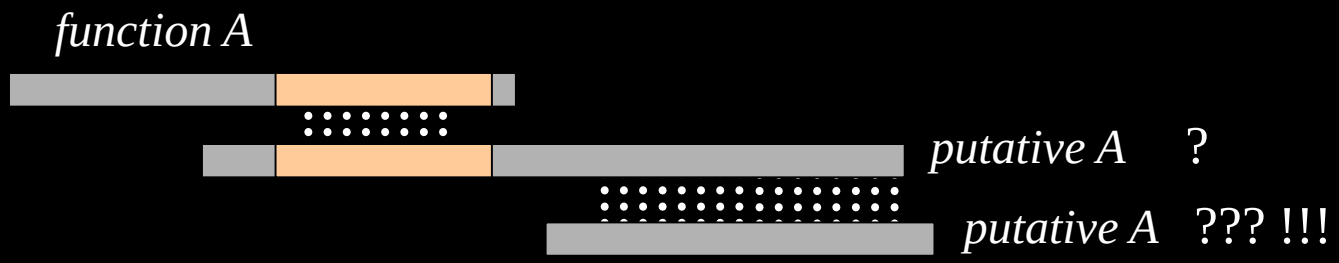
- inference by similarity (BLAST)
- motif and domain detection

“ homologue to putative protein similar to hypothetical orphan - like ”

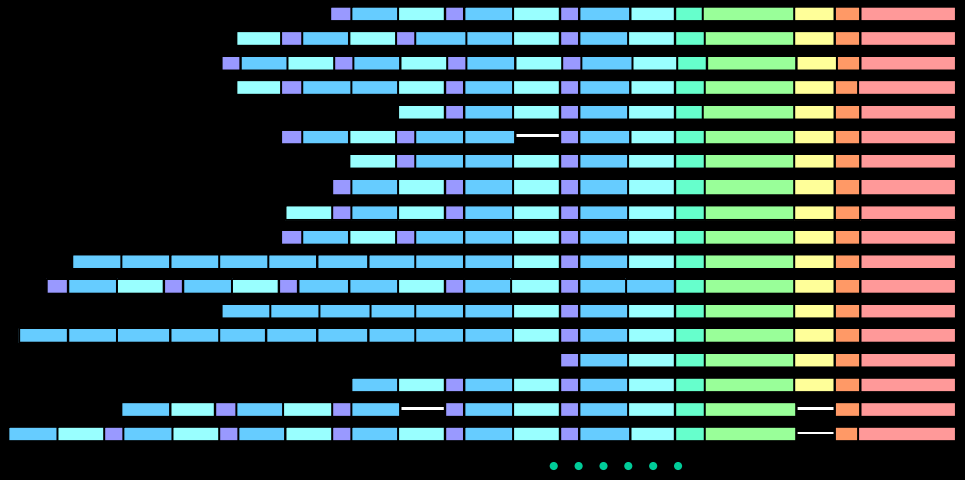
no hit	hypothetical
hit RNAseq	unknown/expressed
hit protein	putative/-like

◀ nomenclature

Errors and databank contamination

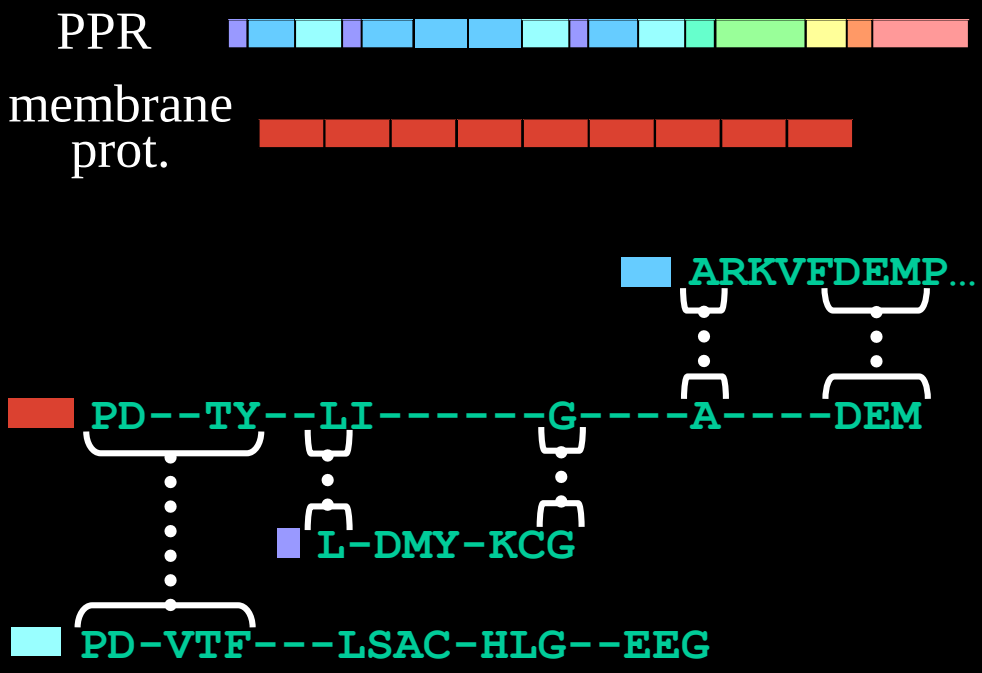


sebastien.aubourg@inra.fr



PPR family

400 genes
in Arabidopsis



Annotations

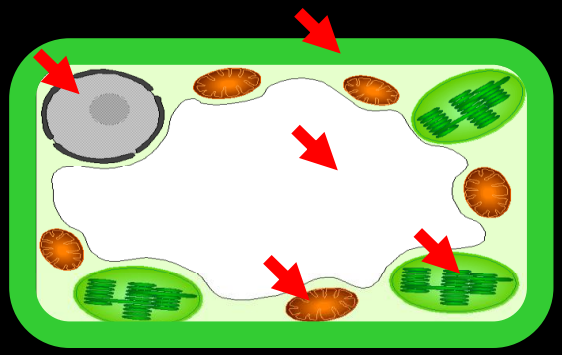
- PPR protein
- seleniu[illegible]binding
- membr[illegible]ne a[illegible]ociated

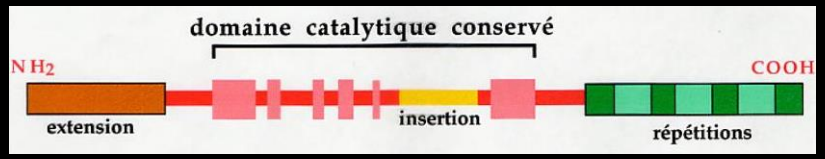
sebastien.aubourg@inra.fr

Expert functional annotation

« Swiss-Prot method »

Bioinformatics predictions → function





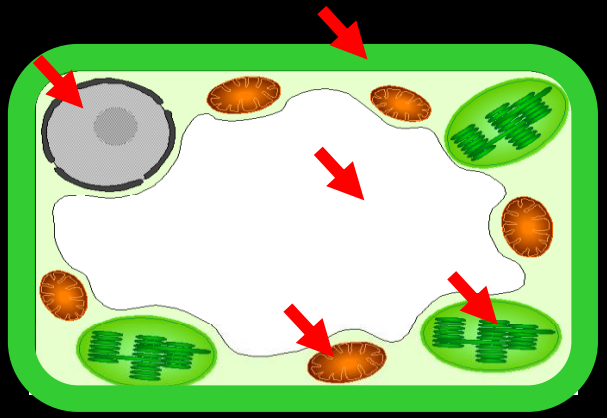
- ✓ targeting and signal peptides for subcellular localization
- ✓ transmembrane segments
- ✓ motifs, domains and signatures
- ✓ post-translational maturation sites
- ✓ secondary structures and 3D models (PBIL)

sebastien.aubourg@inra.fr

Motifs and domains

Signal searching

TmPRED TmHMM
PSORT
PREDOTAR
TargetP MitoP ChloroP…



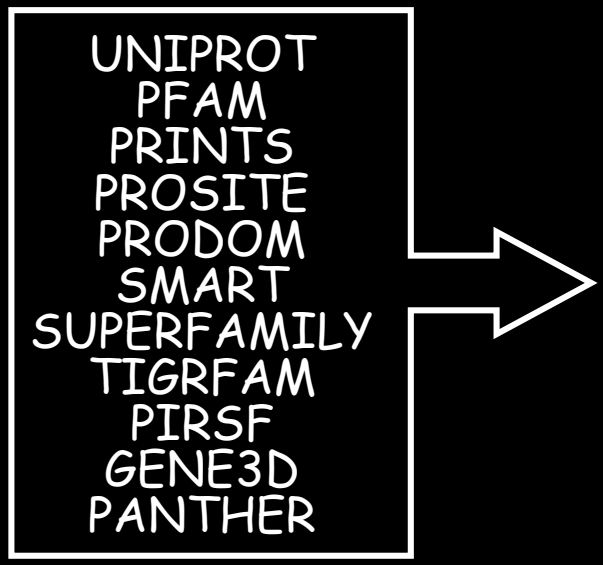
Known motif searching

Prosite (SIB-expertised)	1 500 entries
Pfam-AB (profiles hmm-GeneWise)	12 000 entries
Prints (footprints – signatures)	2 000 entries
Prodom (automatic clustering)	574 000 entries
Panther (evolution-expertised)	6 000 entries

sebastien.aubourg@inra.fr

Motifs and domains

Looking for known motifs

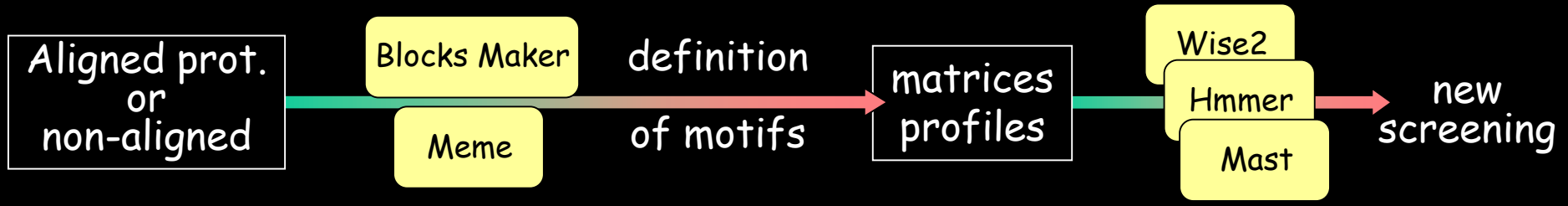


InterPro

Release 39
23 800 entries
24 10^6 proteins
(80% trEMBL)

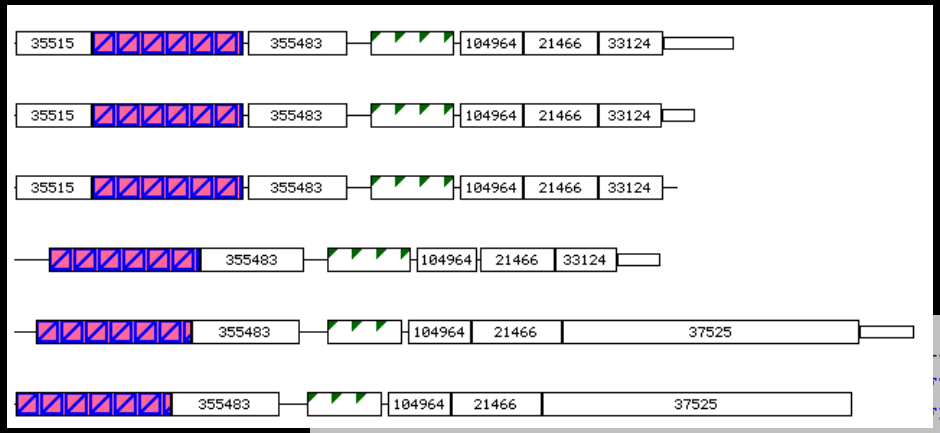
http://www.ebi.ac.uk/interpro/index.html
Nucleic Acids Research vol 29(1):37-40

Definition of new motifs

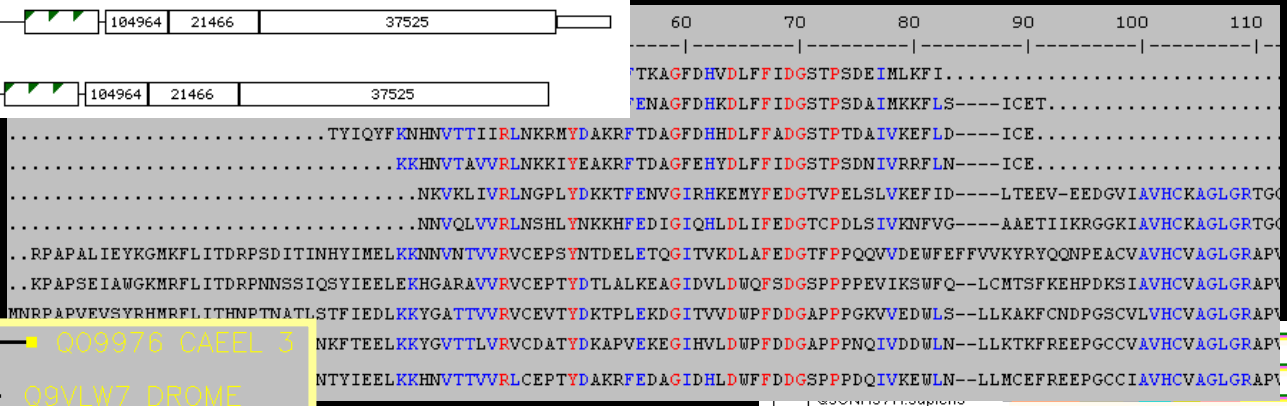


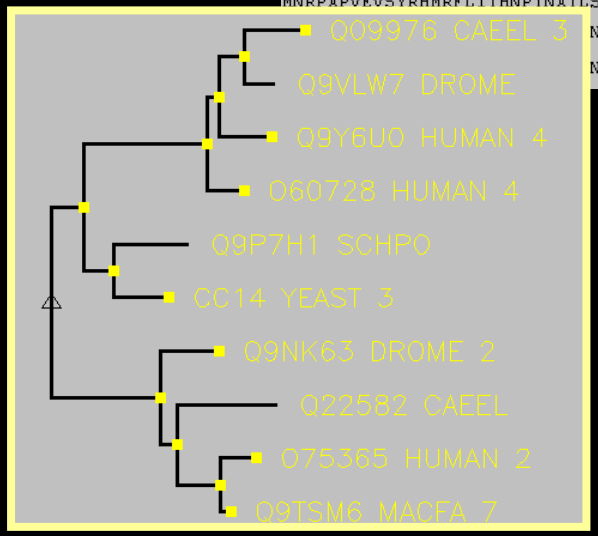
sebastien.aubourg@inra.fr

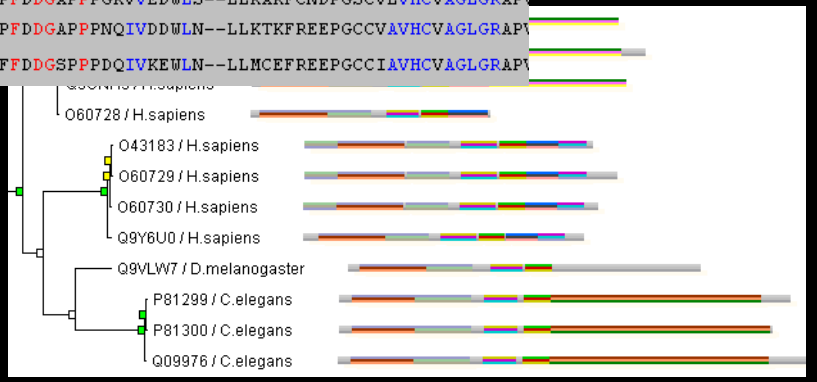
INTERPRO / INTERPROSCAN : the reference



- phylogeny
- cross-references
- biochemical function
- examples







Links:

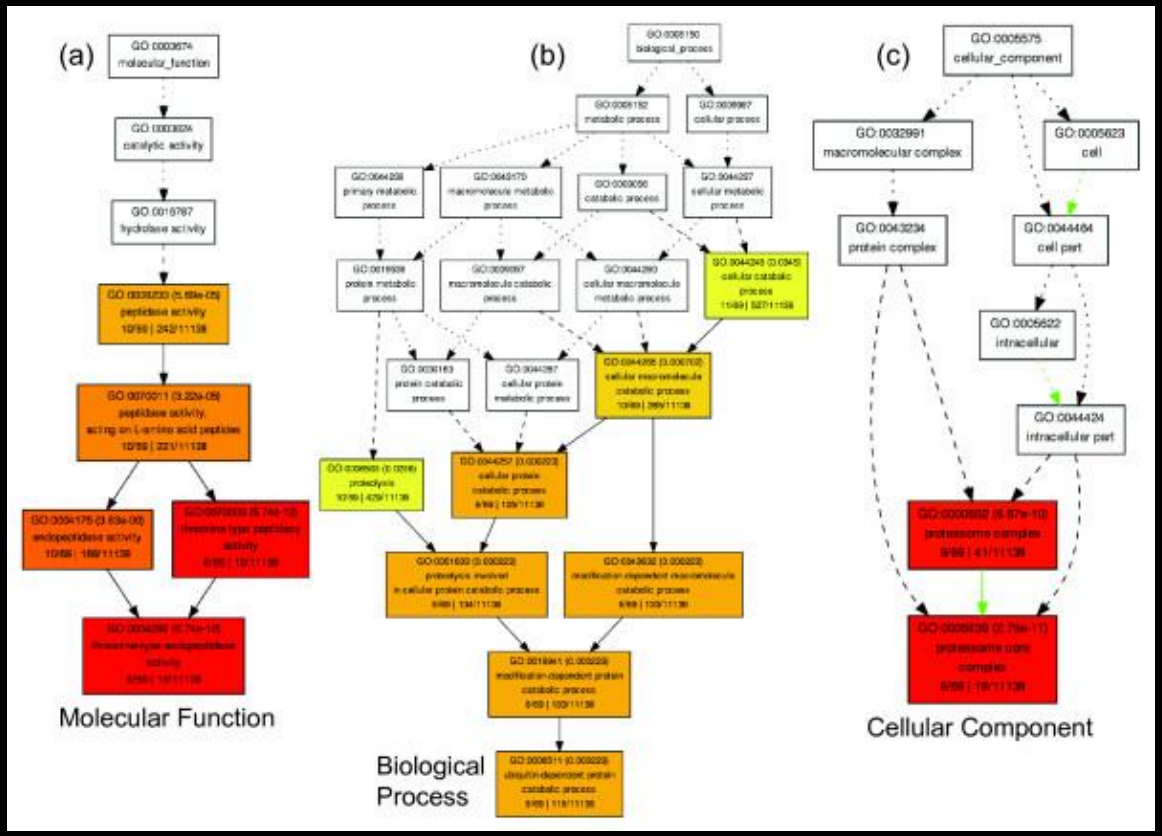
- Medline
- PDB
- generic databanks
- ...

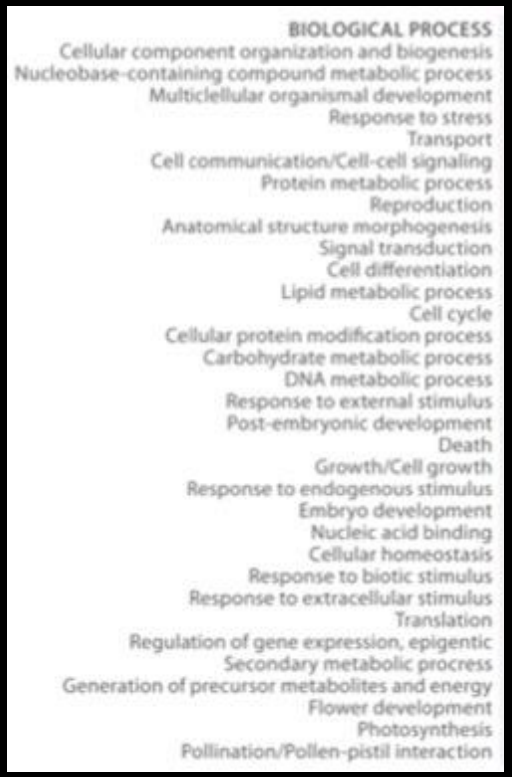
sebastien.aubourg@inra.fr

Functional classification of genes

Gene Ontology Consortium

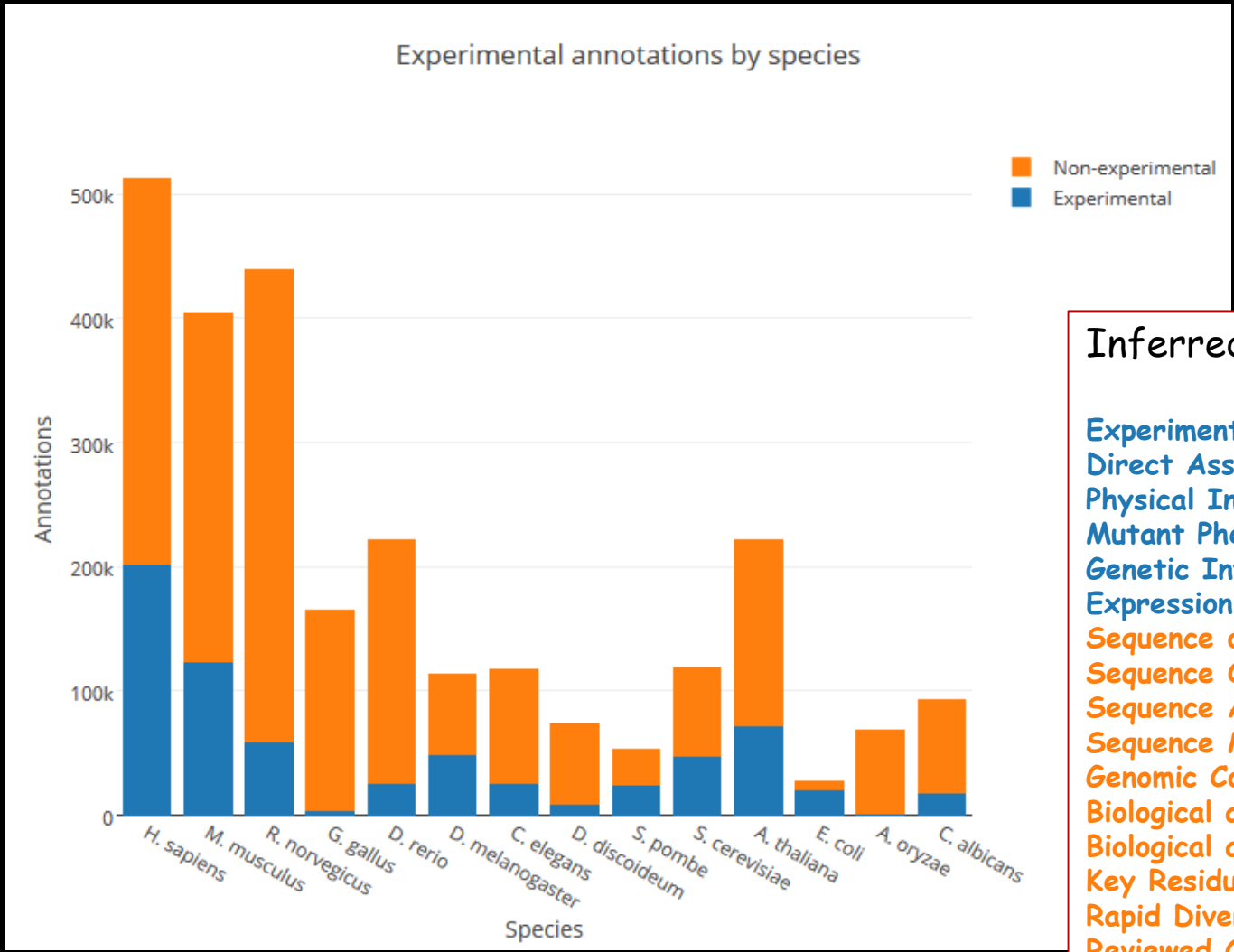
- molecular function (transcription factor, protease…)
- biological process (mitosis, sugar metabolism…)
- cellular component (nucleus, ribosome…)





sebastien.aubourg@inra.fr

Functional classification of genes

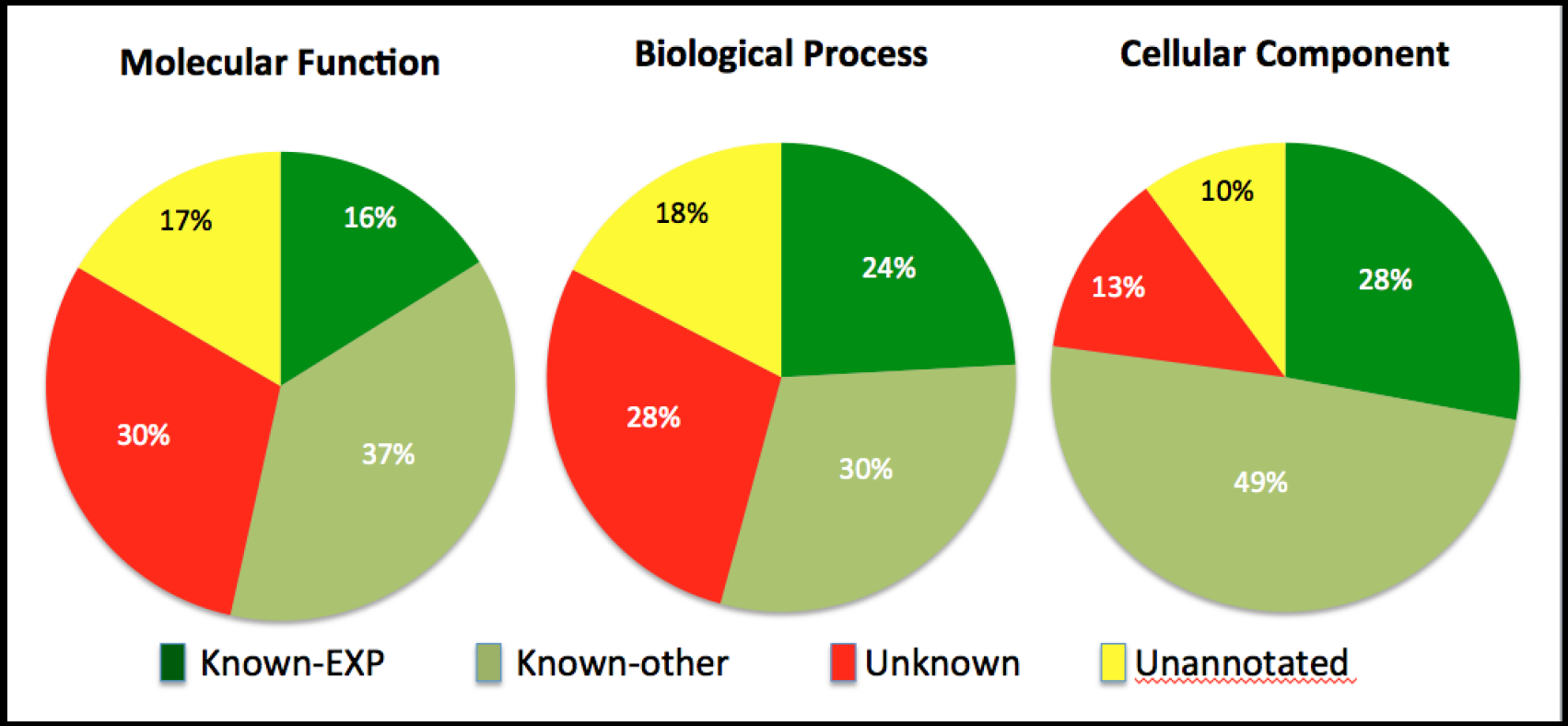


Inferred from:

Experiment (EXP)
Direct Assay (IDA)
Physical Interaction (IPI)
Mutant Phenotype (IMP)
Genetic Interaction (IGI)
Expression Pattern (IEP)
Sequence or structural Similarity (ISS)
Sequence Orthology (ISO)
Sequence Alignment (ISA)
Sequence Model (ISM)
Genomic Context (IGC)
Biological aspect of Ancestor (IBA)
Biological aspect of Descendant (IBD)
Key Residues (IKR)
Rapid Divergence (IRD)
Reviewed Computational Analysis (RCA)
Curator (IC)
Electronic Annotation (IEA)

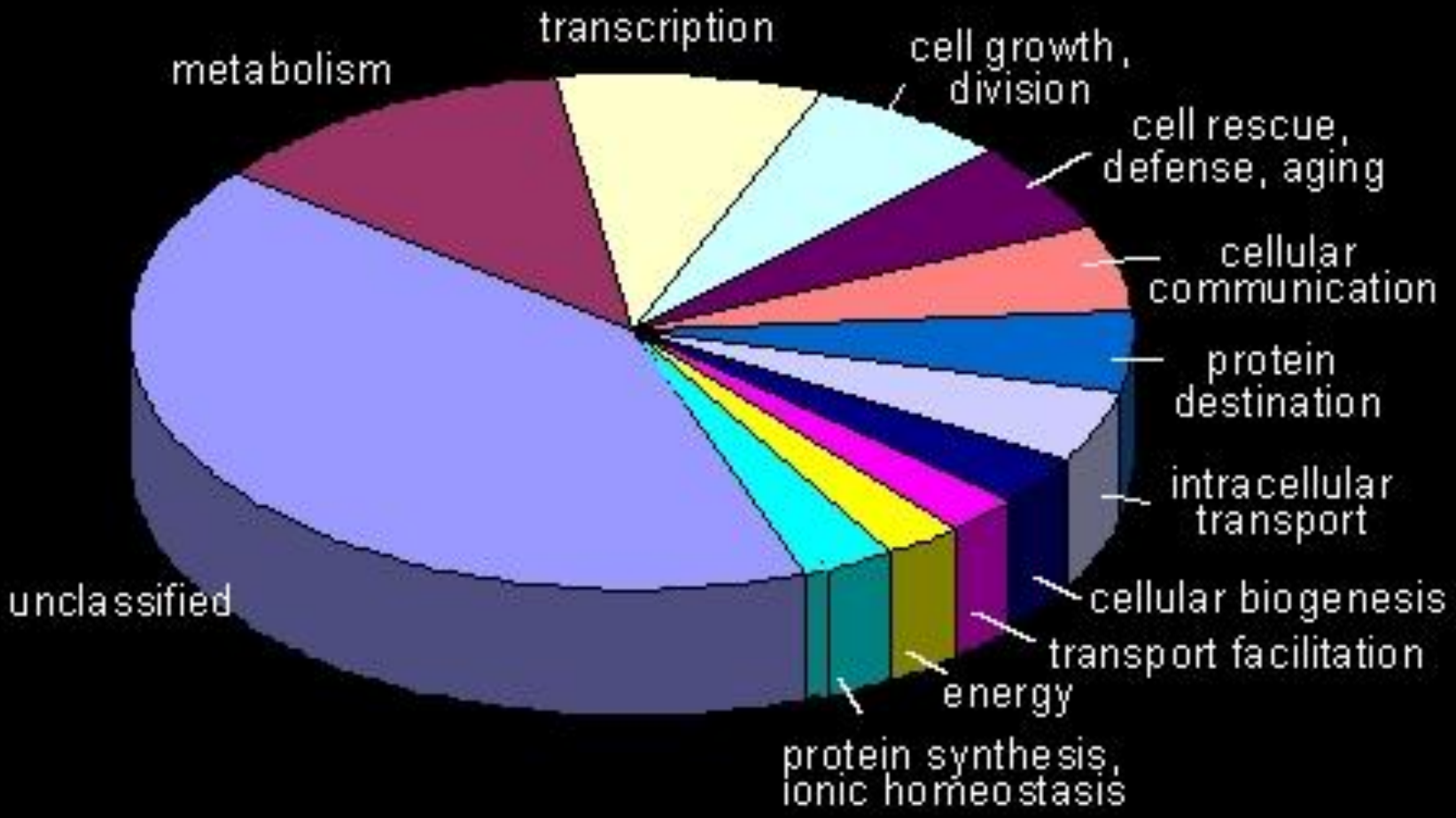
sebastien.aubourg@inra.fr

Functional annotation of the Arabidopsis genome



sebastien.aubourg@inra.fr

Functional annotation of a whole genome



sebastien.aubourg@inra.fr

UniProt & Swiss-Prot

UniProt KnowledgeBase

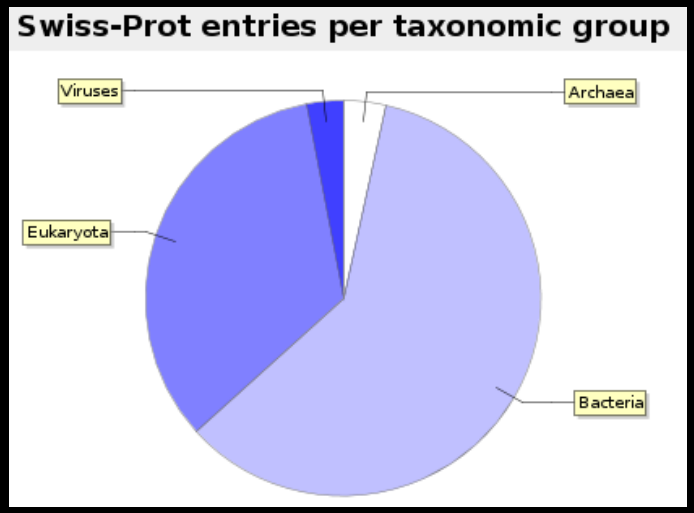
TrEMBL

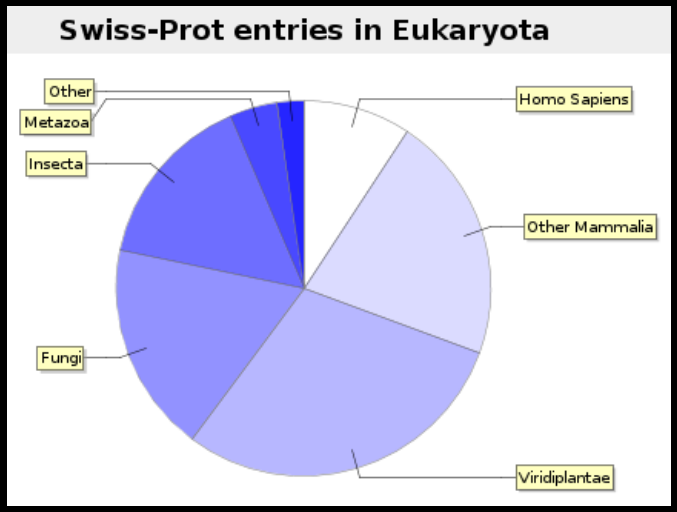
Release 2018_07
120 10^6 entries

Swiss-Prot

Release 2018_07
558 000 entries (2%)

- 50 annotators (SIB) + experts
- Monthly updated
- Curated predictions /validation
- Data mining





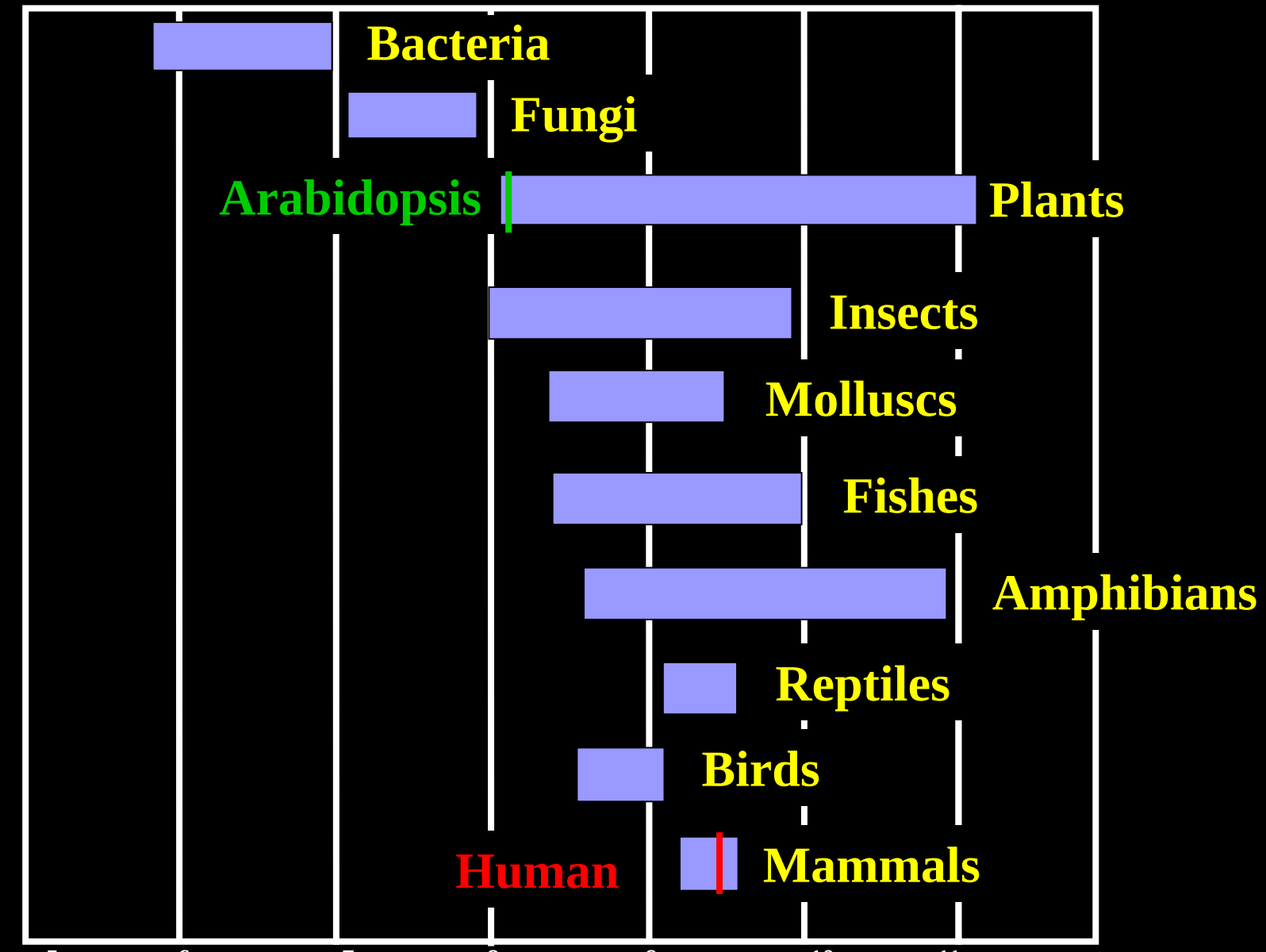
sebastien.aubourg@inra.fr

Genome Annotation

Evolution

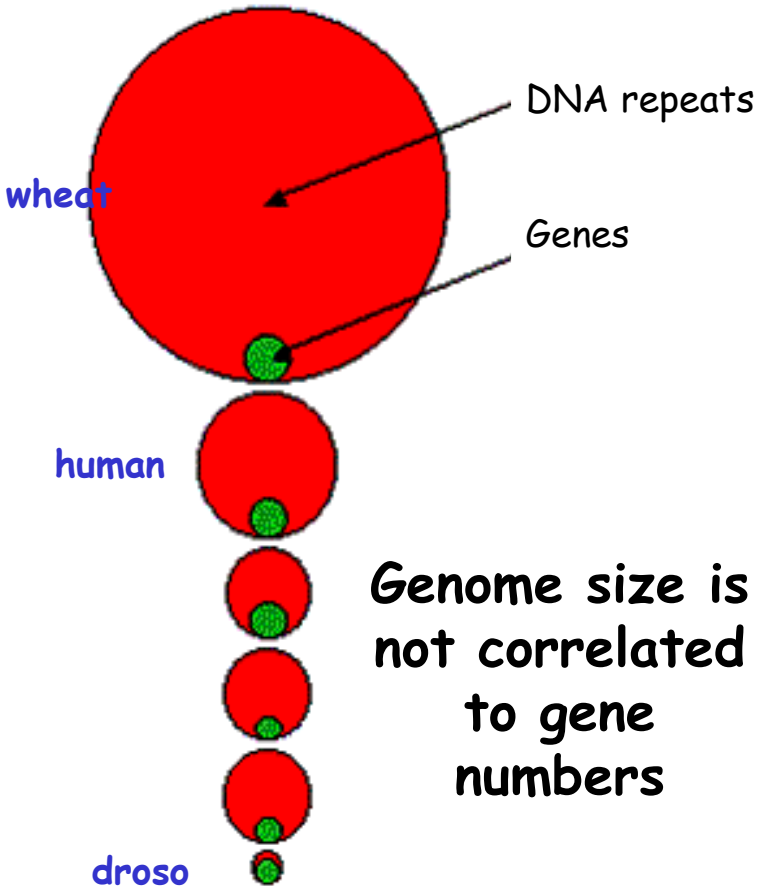
sebastien.aubourg@inra.fr

Genome size

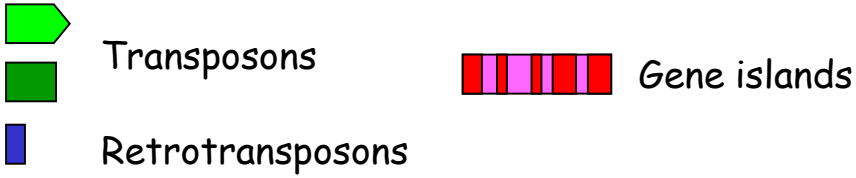


sebastien.aubourg@inra.fr

Gene Space



Genome size is not correlated to gene numbers



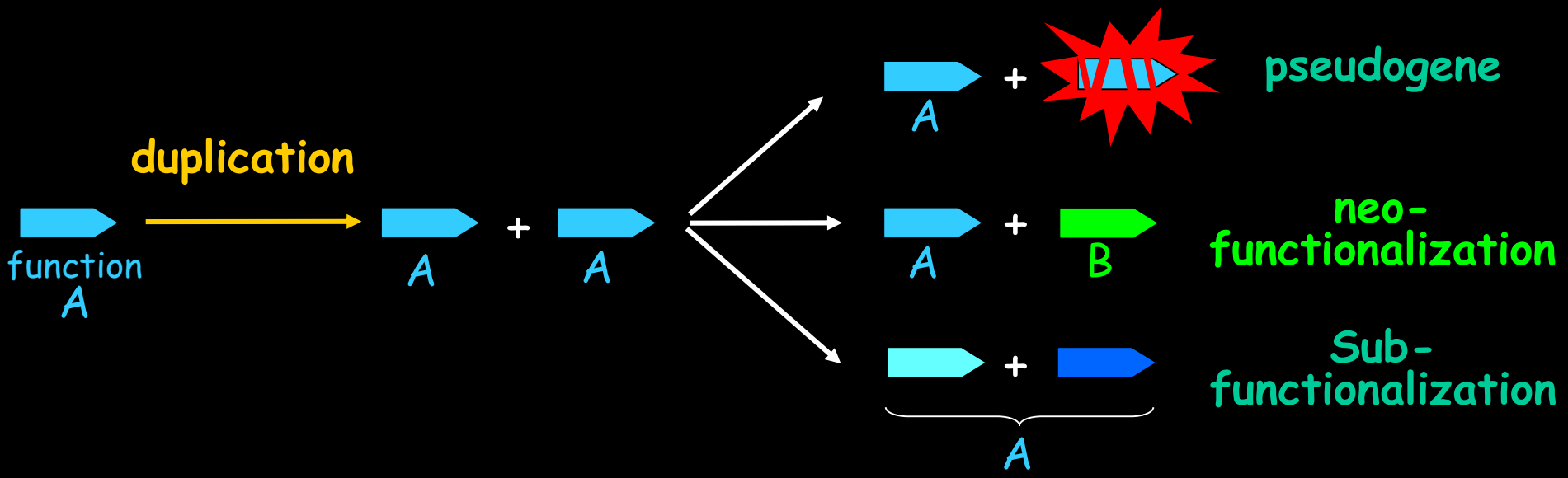
Genome size is the results of opposite processes :

- Retrotransposon invasion by multiple insertions
- Duplications (at different scales)
- Deletions by recombination

sebastien.aubourg@inra.fr

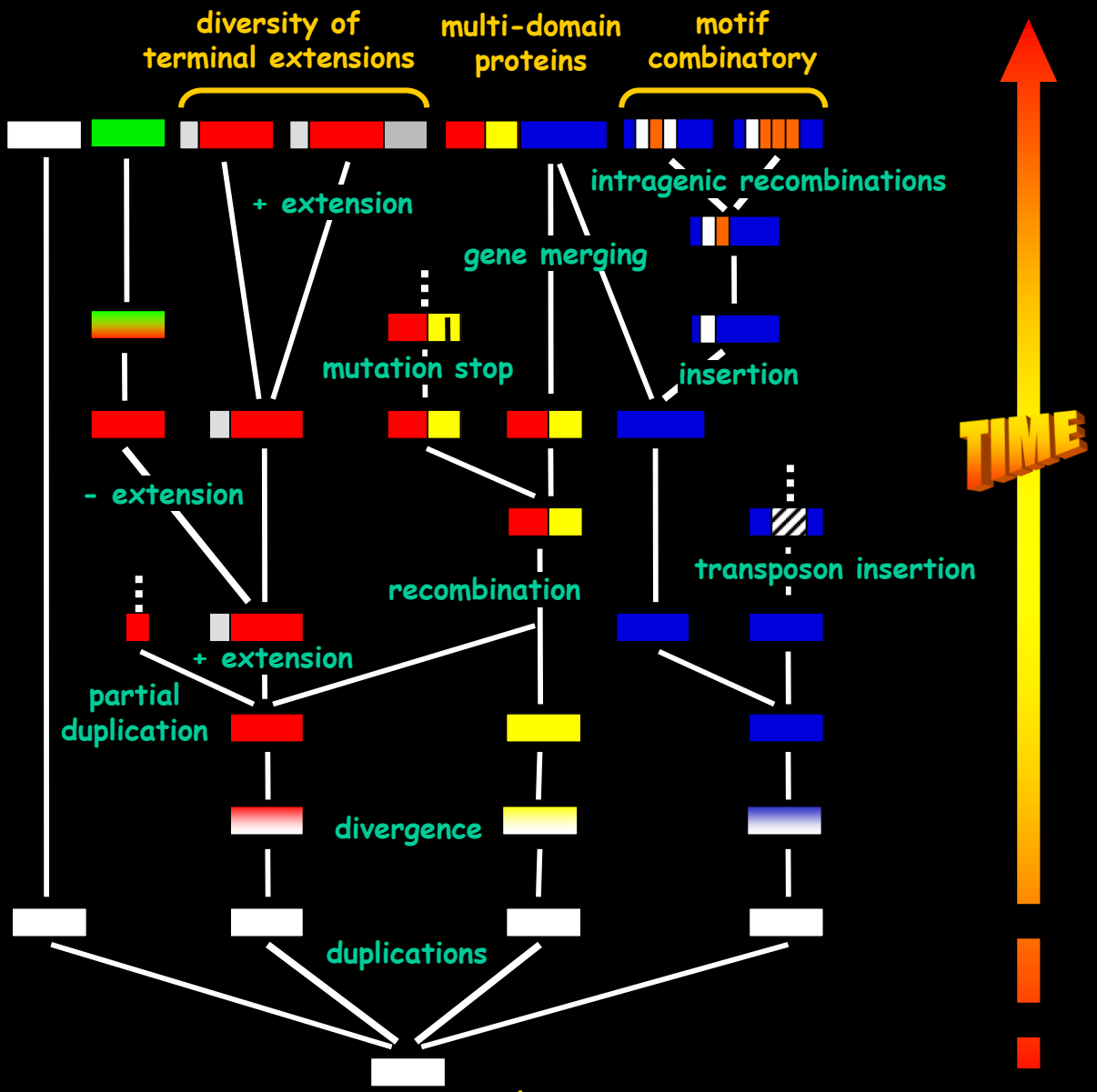
Duplication and divergence

'Evolution by gene duplication '
Ohno, 1970



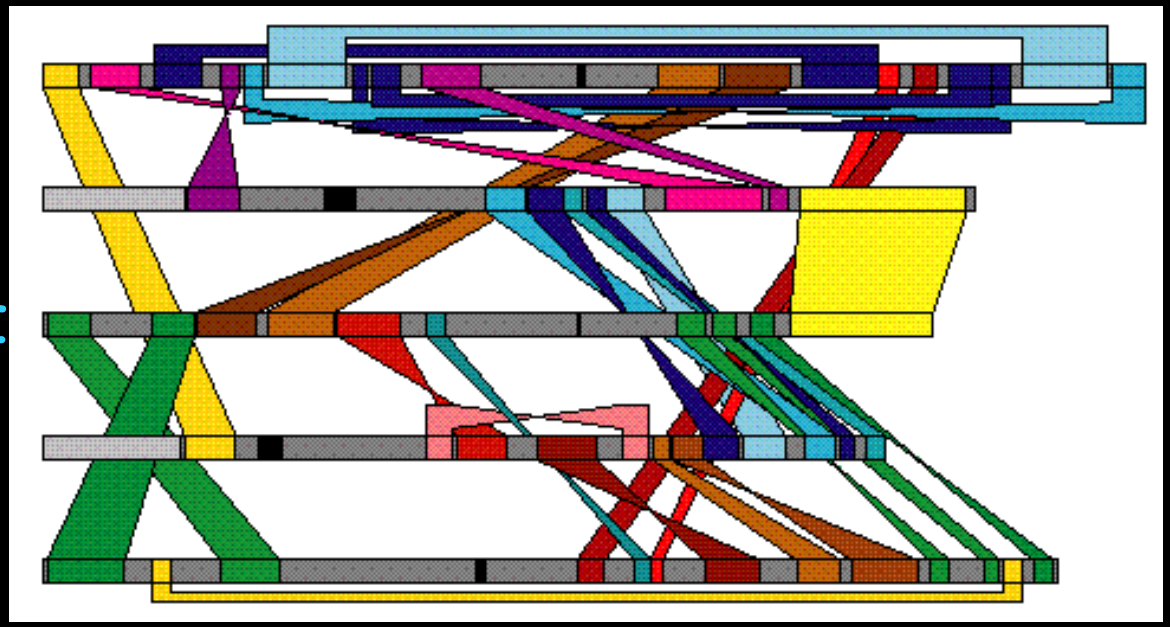
sebastien.aubourg@inra.fr

Functionalization

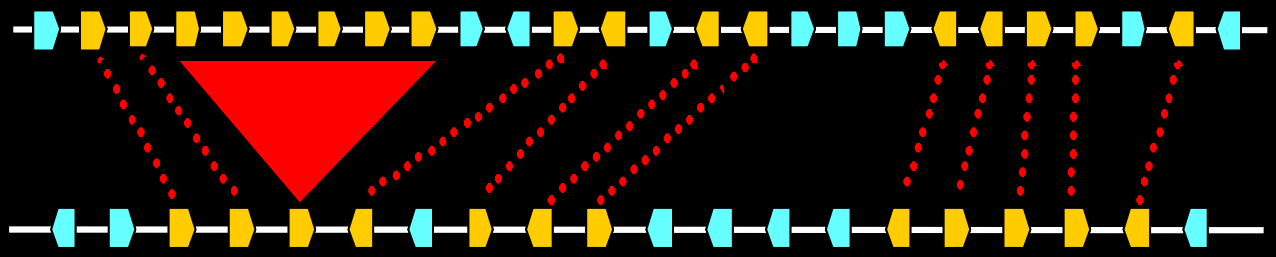


sebastien.aubourg@inra.fr

Dynamic of genomes

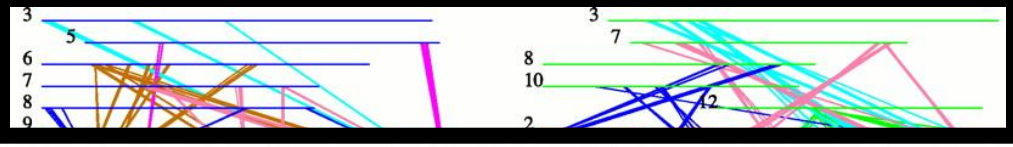


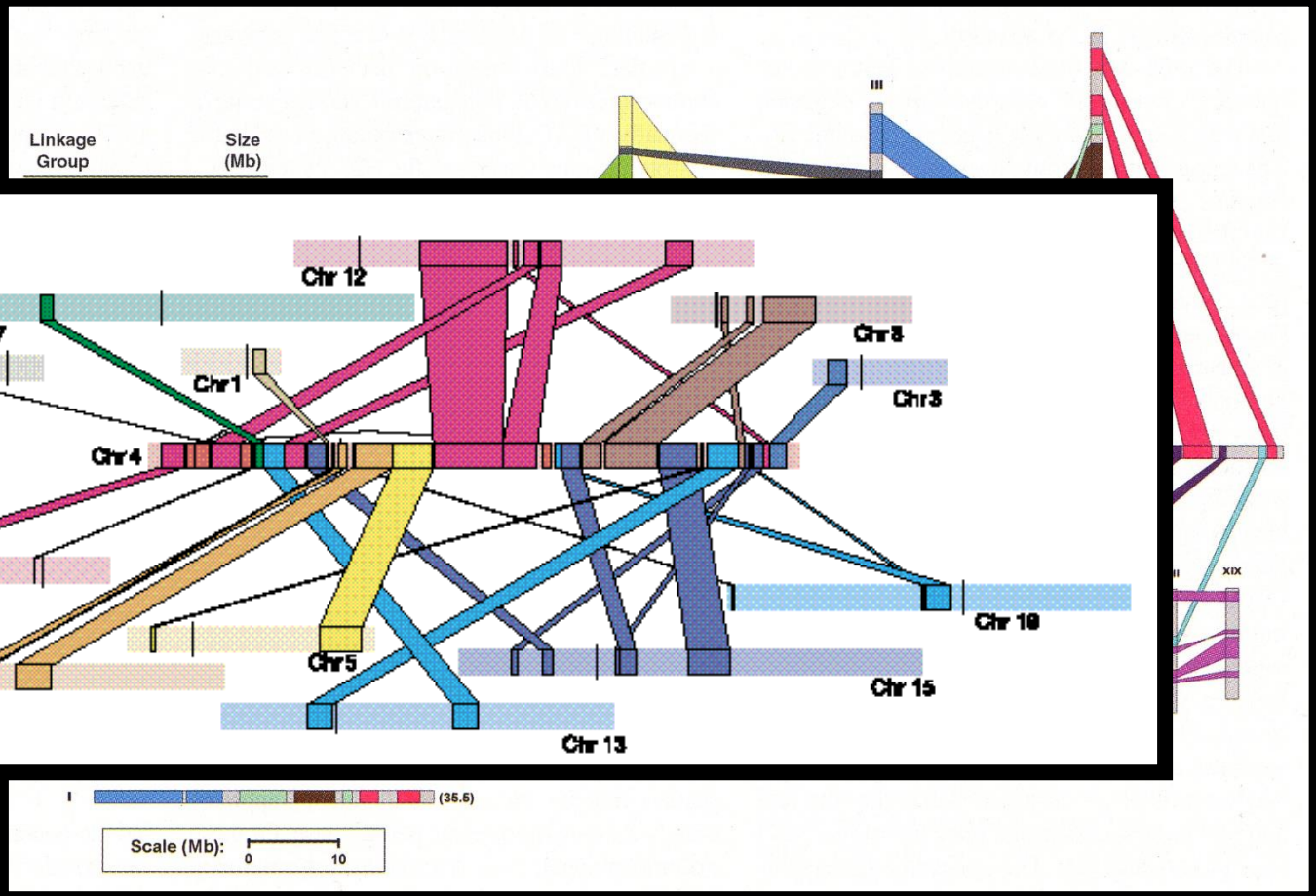
80% of the Arabidopsis genome is duplicated



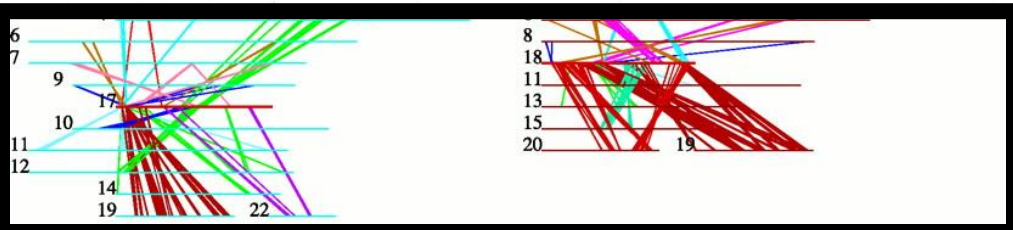
sebastien.aubourg@inra.fr

Whole Genome Duplication





yeast



sebastien.aubourg@inra.fr

Homology

Etymology : A homologous trait is any characteristic of organisms that is derived from a common ancestor, a common origin. The homology is a similarity acquired, inherited from a common ancestor.

The significant similarities between two sequences are considered as representative of their homology: They have a common origin, they derived from the same ancestral sequence.

Similarity (quantitative) ⟹ Homology (qualitative)

sebastien.aubourg@inra.fr

Definitions

HOMOLOG

PARALOG

Homologous genes are paralogs if they were separated by a gene duplication event.

ORTHOLOG

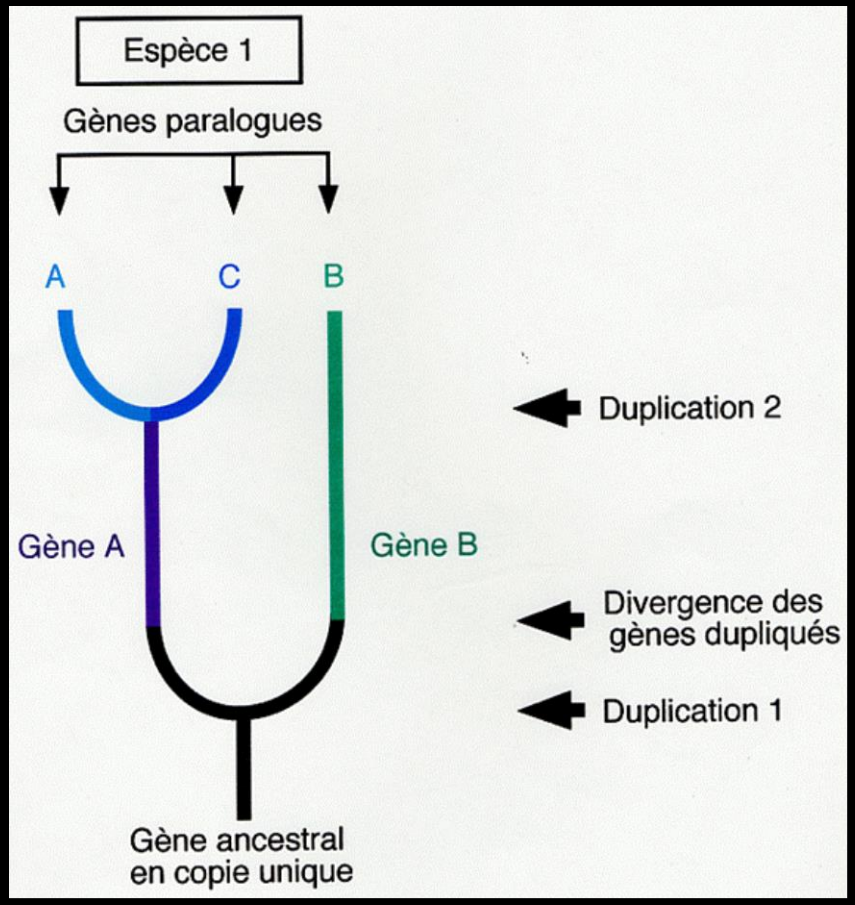
Homologous genes are orthologs if they were separated by a speciation event: when a species diverges into two separate species, the divergent copies in the resulting species are orthologs.

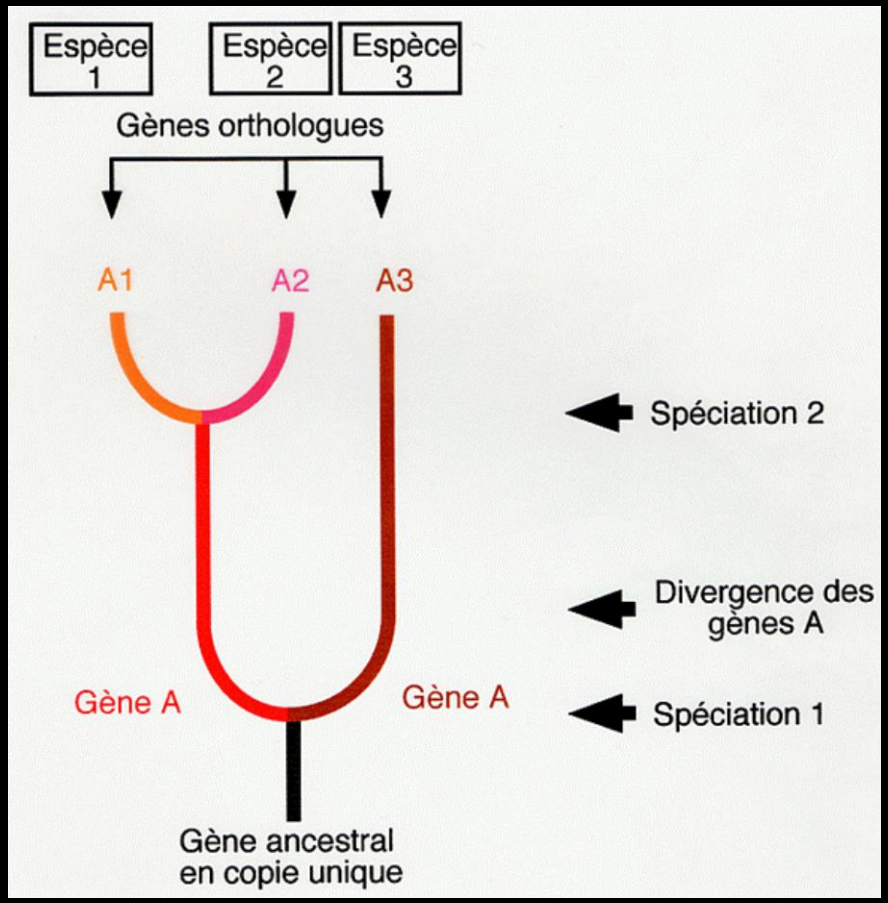
GENE FAMILY

Group of homologous genes. The genes belonging to the same family have therefore a common ancestor gene and share significant similarities. They encode proteins having a common biochemical function but can perform different biological functions.

sebastien.aubourg@inra.fr

Paralogs and orthologs



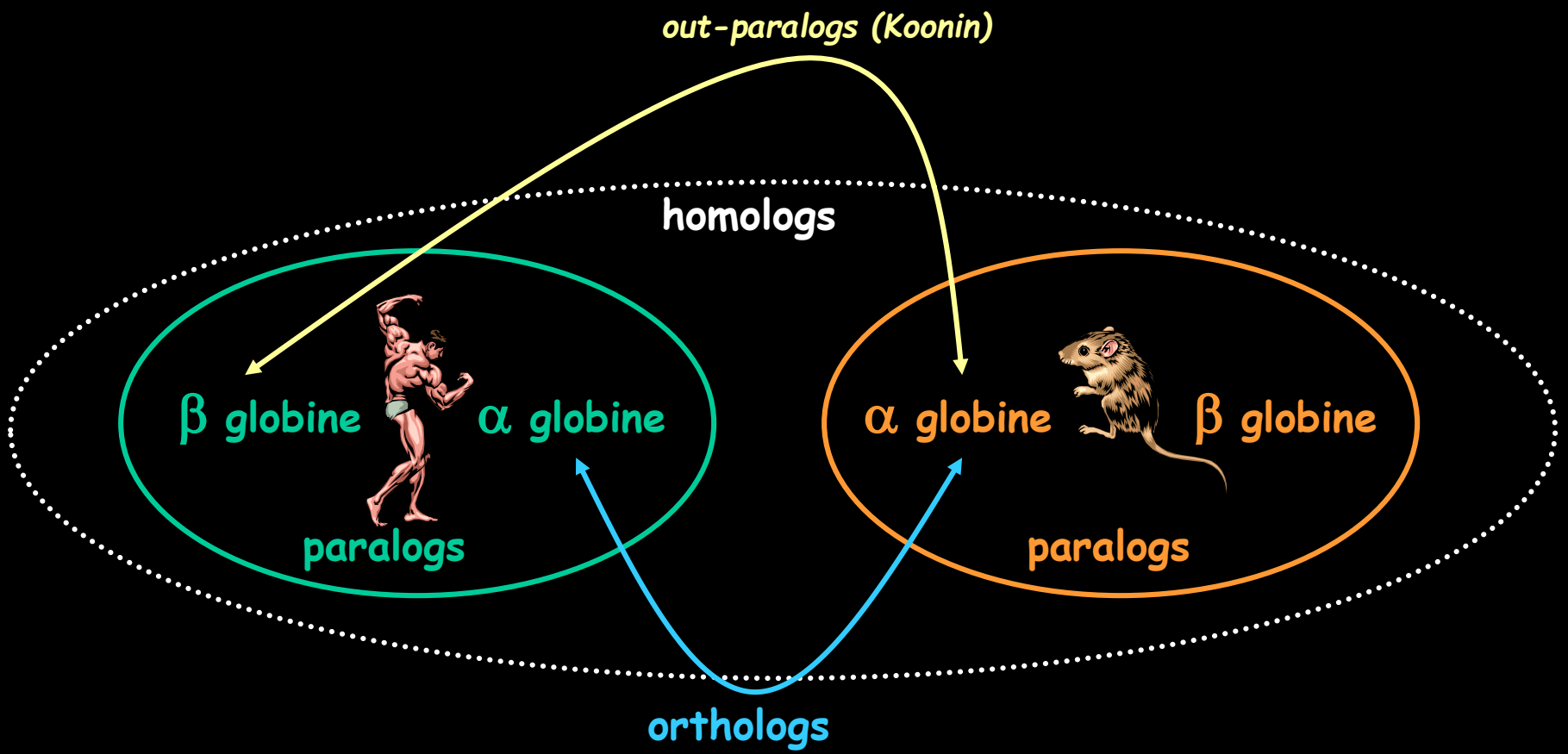


sebastien.aubourg@inra.fr

Concretely…

The genes AtMYB29 and OsMYB43 have 53% of in their coding region.

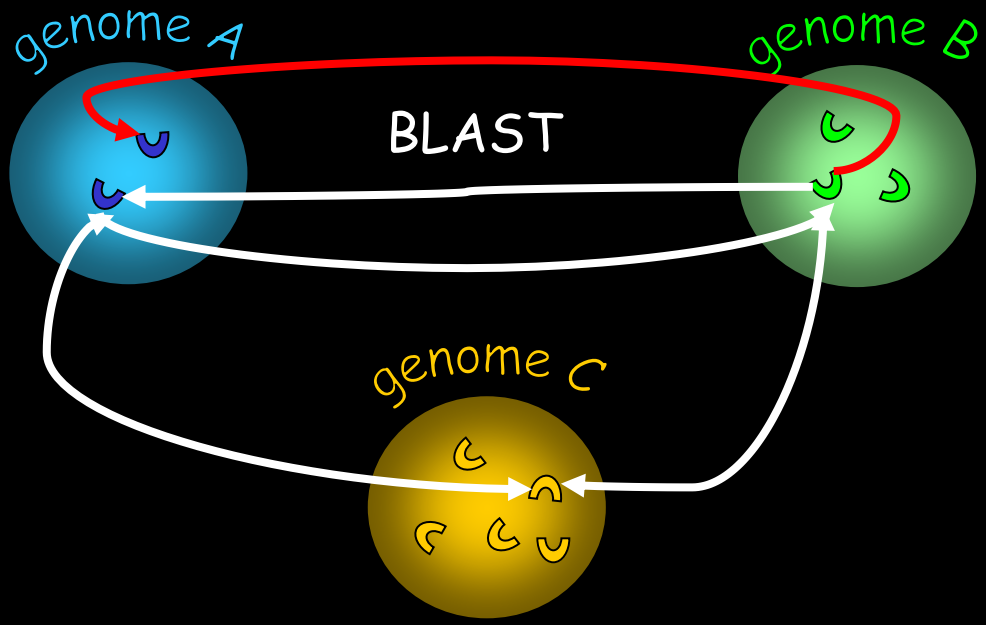
- ☐ similarity
- ☑ identity
- ☐ homology



sebastien.aubourg@inra.fr

The ortholog quest ('functional orthologs')

Diversify the approaches… collect the clues



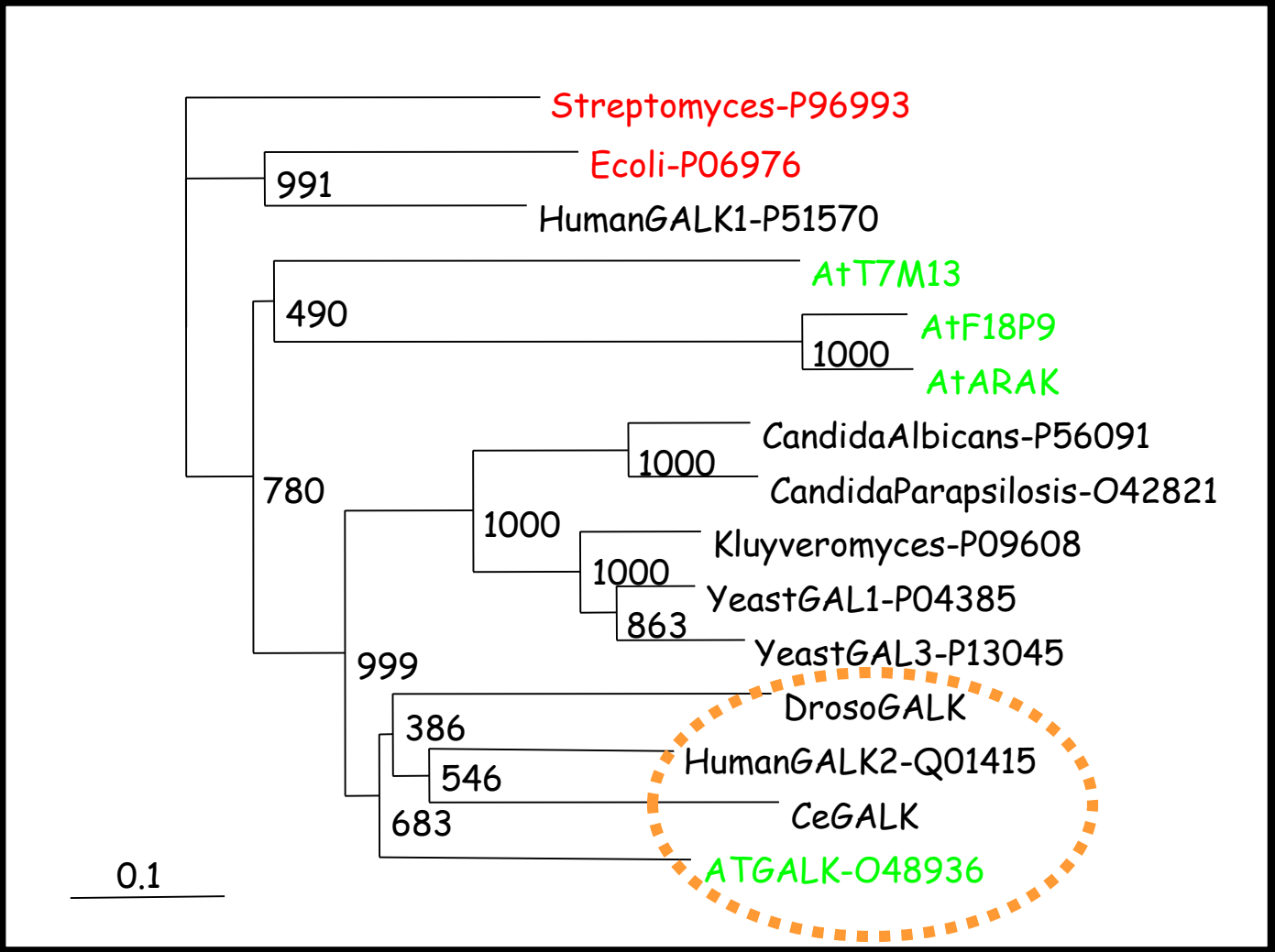
- Cross the genomes:
  'BDBH'

  bidirectional best hit
  multidirectional reciprocal best blast hit

- Look for the presence of common insertion or extension
- Compare intron-exon structure of genes
- Find protein partners (Rosetta Stone)
- Exploit the genomic environment (synteny)
- Explore promoter regions (promoter shadowing)
- Exploit transcriptomic or proteomic data (expression profiles)

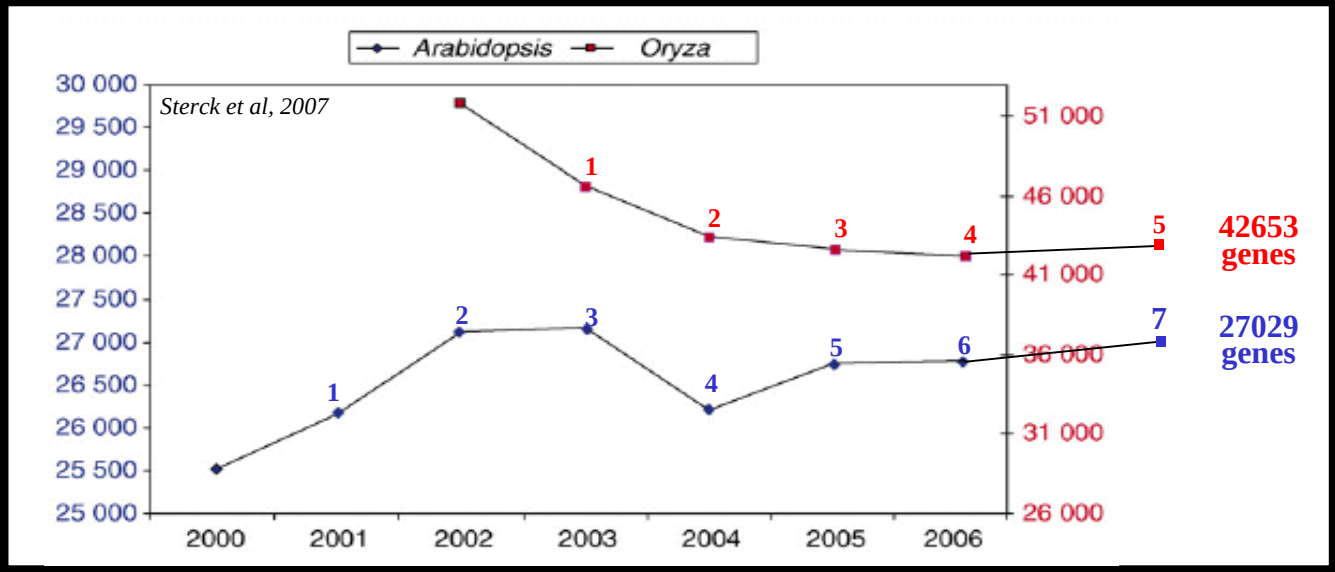
sebastien.aubourg@inra.fr

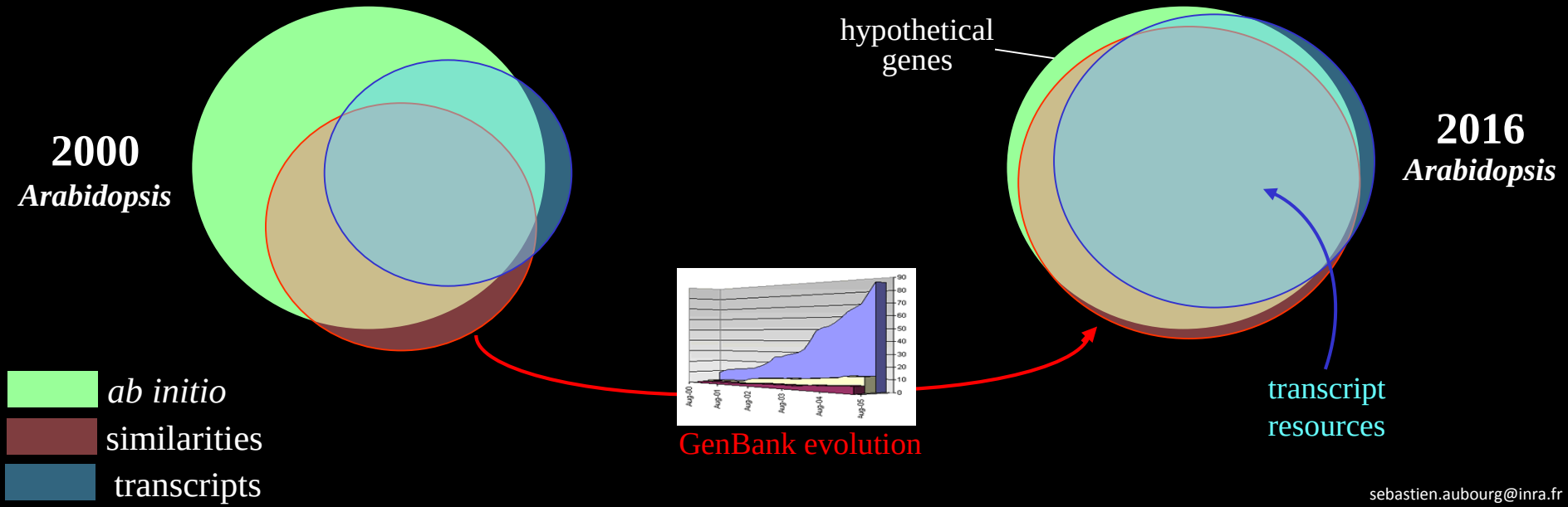
Phylogeny and orthology



sebastien.aubourg@inra.fr

Annotation is a dynamical process





sebastien.aubourg@inra.fr

Toward a relational annotation

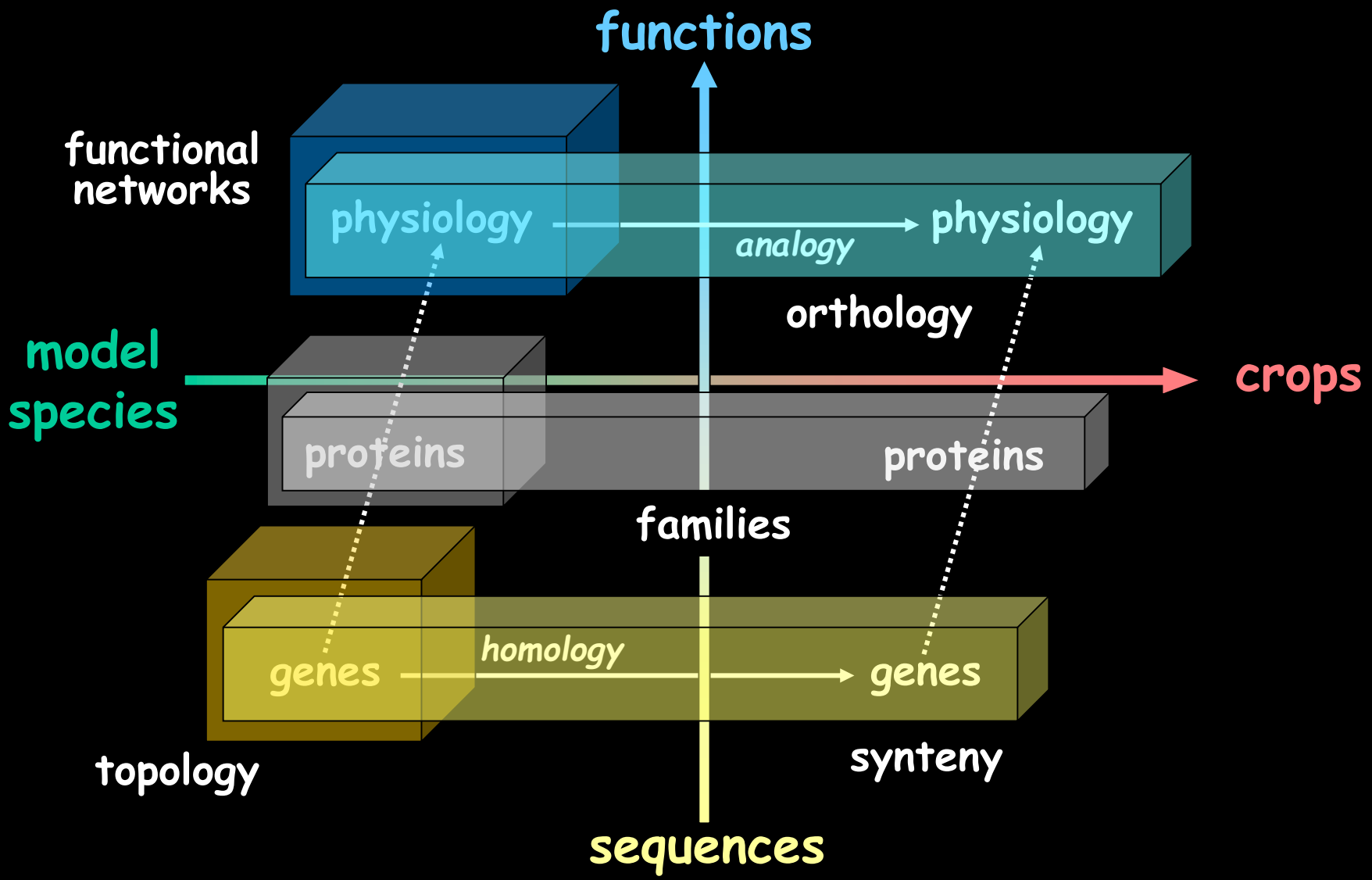
Comparative genomics (gene context) Phylogenetic profiles	►	functional collaboration
miRNA prediction and their target(s) Relationships between TFBS/TF (ChIP-)	►	regulation
Transcriptome Proteome	►	co-expression
Interactome (systematic Y2H) Definition of interologs	►	direct interaction
Text-mining (co-citation)	►	functional links

► gene/protein networks

Integrative biology

sebastien.aubourg@inra.fr

From sequences to functions



sebastien.aubourg@inra.fr