Bioinformatique: Annotation des génomes (eucaryotes)

M2 MIV - Octobre 2006

Laurent Duret
BBE – UMR CNRS n° 5558
Université Claude Bernard - Lyon 1

Activités de recherche

- Equipe “Bioinformatique et Génomique Evolutive”
  - Labo. de Biométrie et Biologie Evolutive (CNRS, Univ. Lyon 1)
  - Pôle Bioinformatique Lyonnais (avec l’équipe de G. Deléage, IBCP): http://pbil.univ-lyon1.fr
  - Groupe HELIX (INRIA)
- Développement d’outils informatiques pour l’analyse des génomes (bases de données, algorithmes)
- Etude de l’organisation et de l’évolution des génomes
  - Évolution moléculaire
  - Analyse comparative de séquences
  - Phylogénie

Projets génomes

- Identifier les gènes et autres éléments fonctionnels dans les séquences génomiques (où sont les gènes ?)

  => prédiction de gènes

- Déterminer la fonction des gènes (qu'est qu'ils font ?)

  => recherche de similarités entre séquences

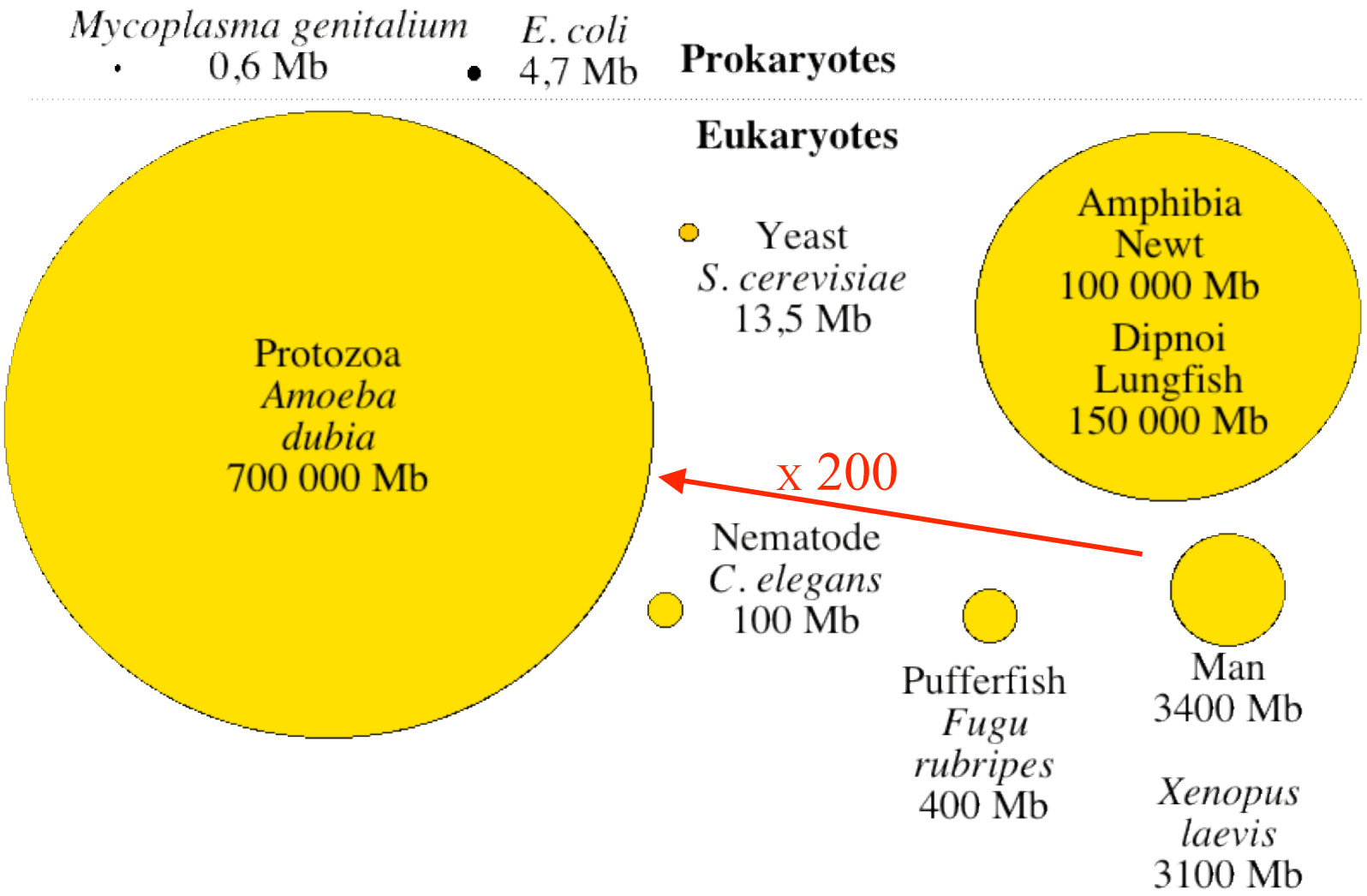
Plan du cours

- Structure des génomes eucaryotes
- Projets génomes (eucaryotes)
- Identification de gènes (eucaryotes)
- Prédire la fonction des gènes

Qu'est-ce qu'un génome ?

- Définition historique (1920) = ensemble des gènes d'un organisme
- Définition actuelle = ensemble des molécules d'acides nucléiques transmises de génération en génération
  - Génome nucléaire
  - Génome mitochondrial
  - Génome chloroplastique
  - ….

Genome size and the c-value paradox



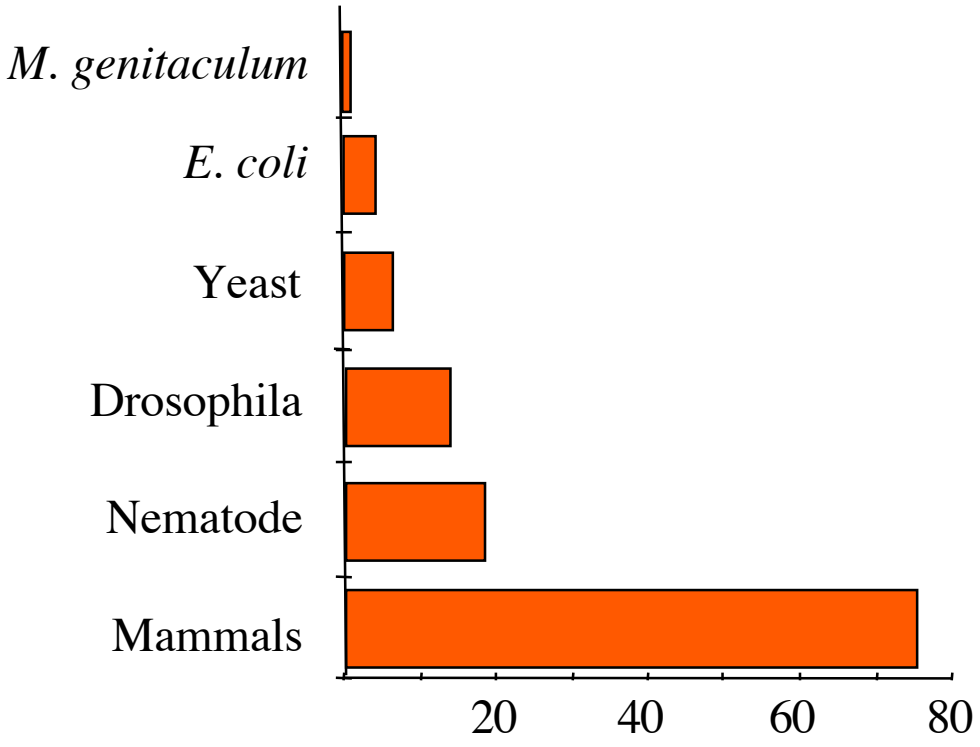
A genome is more than a set of genes

- Genes (transcription unit):
  - Protein-coding genes
  - RNA genes:
    - rRNAs, tRNAs, snRNAs, etc.
    - Untranslated RNA genes (e.g. Xist, H19)
- Regulatory elements (promoters, enhancers, etc.)
- Elements required for chromosome replication (replication origins, telomeres, centromeres, etc.)
- Non-functional sequences
  - Non-coding sequences
  - Repeated sequences
  - Pseudogenes

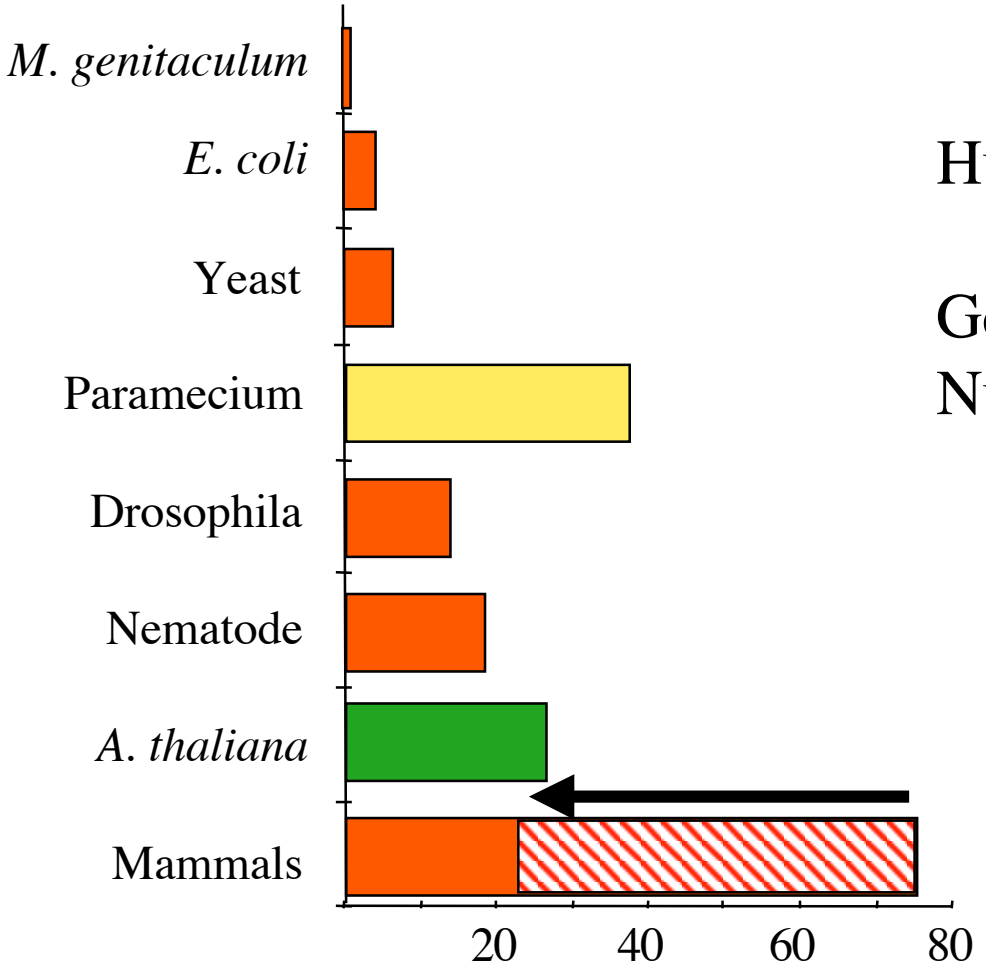
How much of the genome is functional *vs* non-functional ?

- The big surprise of mammalian genome projects …

Number of protein genes (1999)



Number of protein genes (2006)



Human *vs E. coli*:

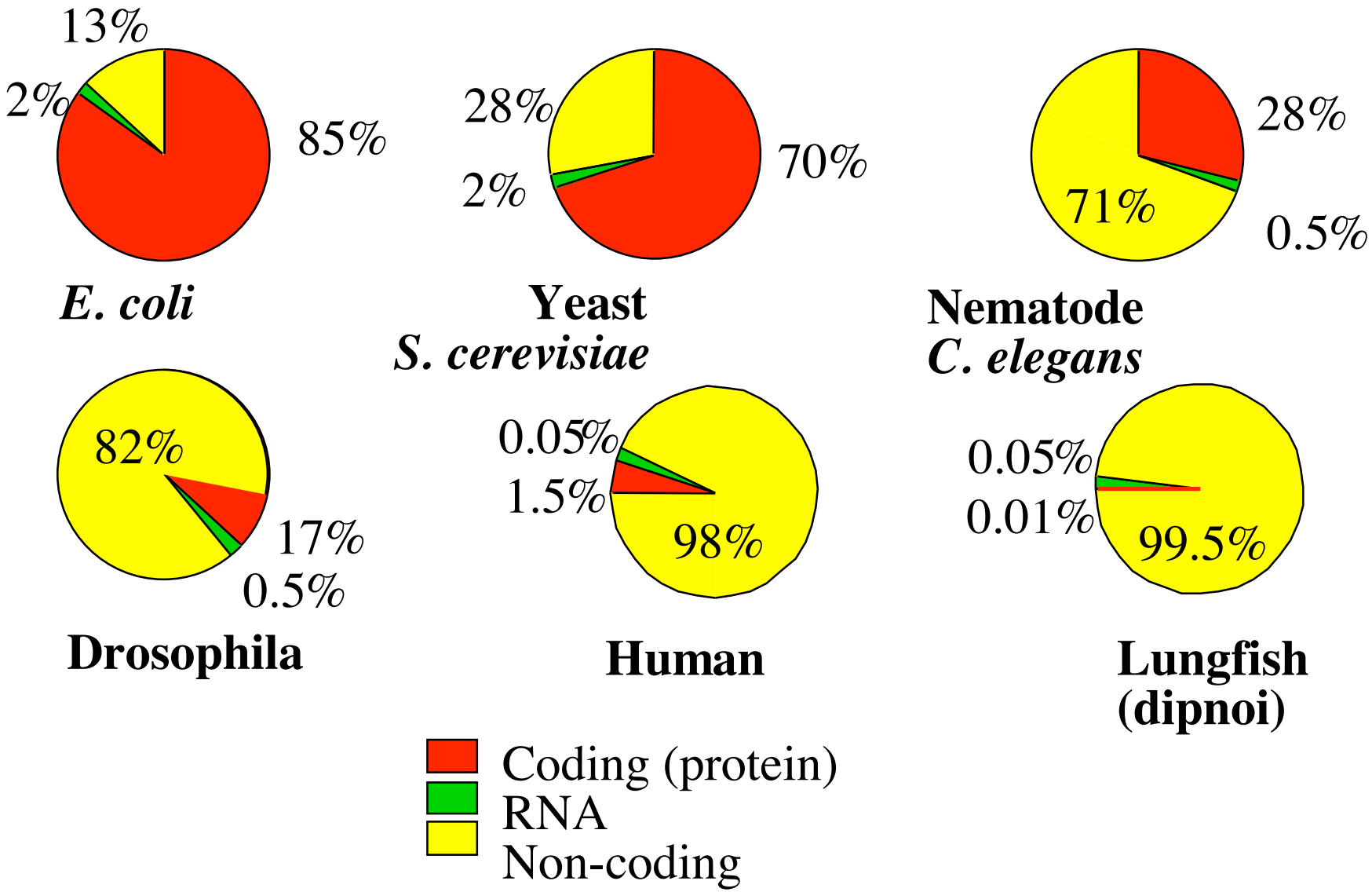
Genome size: x 600

Number of genes: x 5

How many genes in the human genome ?

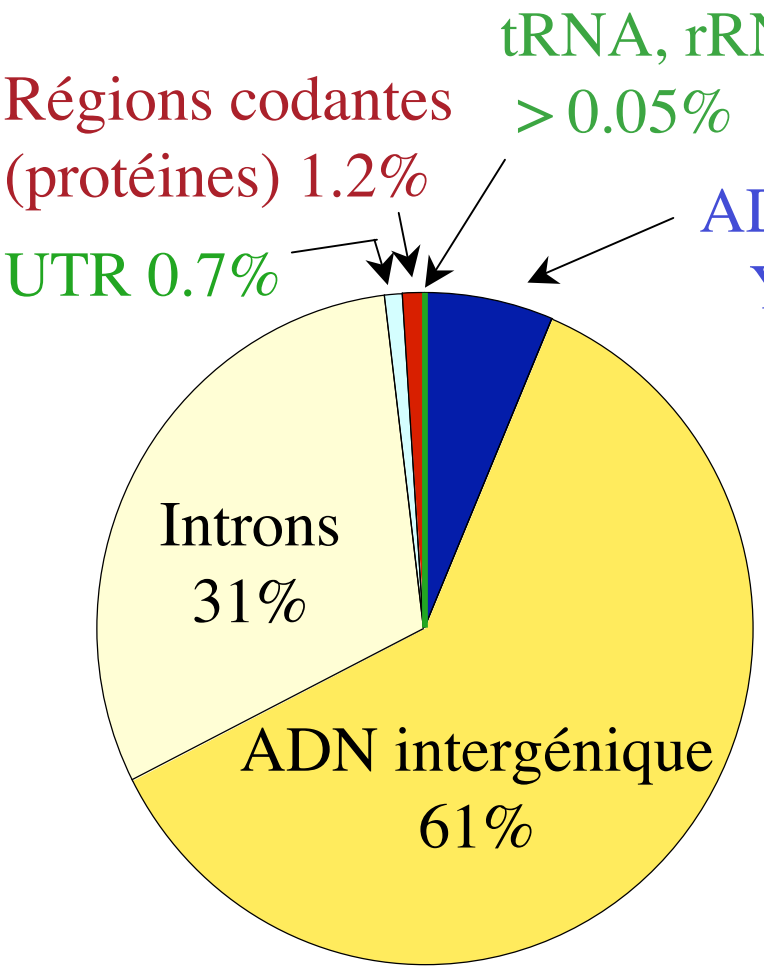
Technique	Gene estimate	Comments/assumptions
Text book (1990)	100,000	if average size = 30 kb
Genomic sequencing (1994)	71,000	biased toward gene-rich region?
CpG islands	67,000	assumes 66% human genes have CpG islands
EST analysis (1994)	64,000	matching with GenBank; 50% EST redundancy
Chromosome 22 (1999)	45,000	correction for high gene density on chrom. 22
Exofish (2000)	28,000-34,000	Comparison human/fish
EST (2000)	35,000	Number of genes
EST (2000)	120,000	Number of transcripts
First genome draft (2001)	30,000-40,000	Known genes + predictions
Comparison / mouse (2002)	30,000	Known genes + predictions
Finished genome (2004)	20,000-25,000	Known genes + predictions

Proportion of functional elements within genomes



Eléments fonctionnels dans le génome humain

3.1 10^9 nt 20 000 - 25 000 gènes protéiques



Elements fonctionnels non-quantifiés:

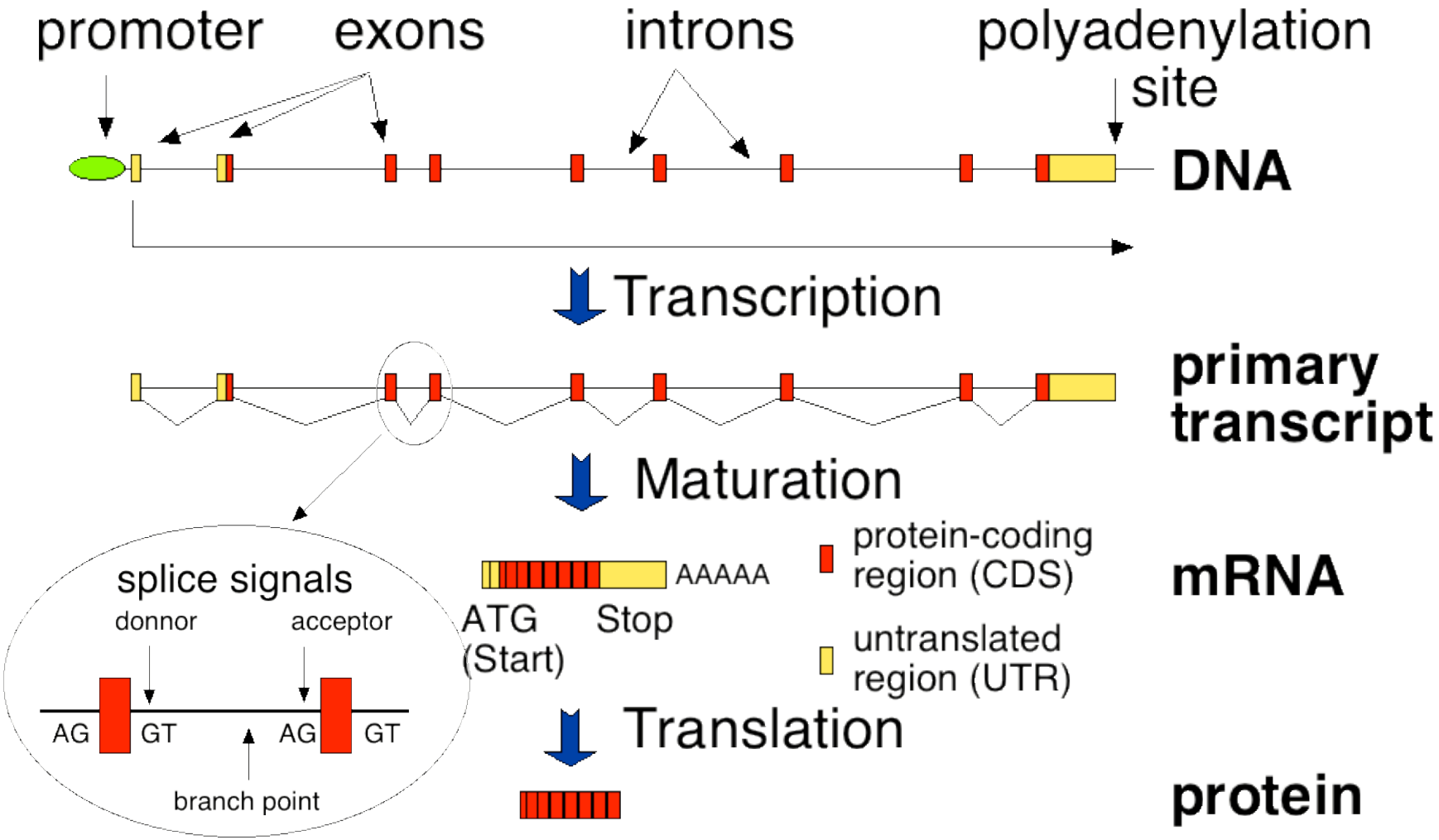
- ARN non-traduits: Xist, H19, etc.
- Eléments régulateurs: promoteurs, enhancers, etc.
- Origines de réplications, MAR, télomères

Eléments non-fonctionnels:

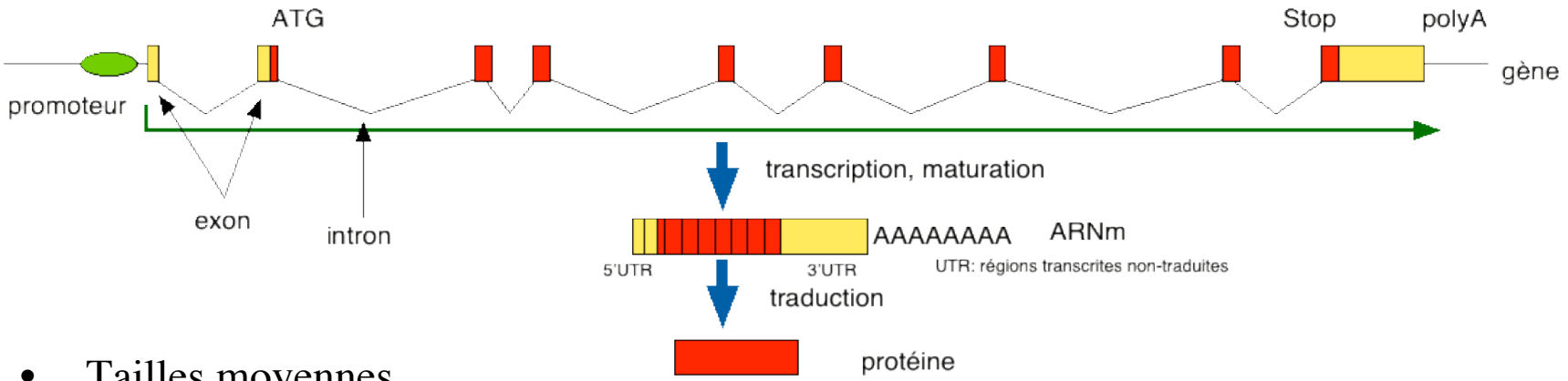
- Pseudogènes : 1.2%
- Elements transposables (SINES, LINES, HERV, etc.) : 42%

>90% sans fonction connue

Typical eukaryotic protein-coding gene



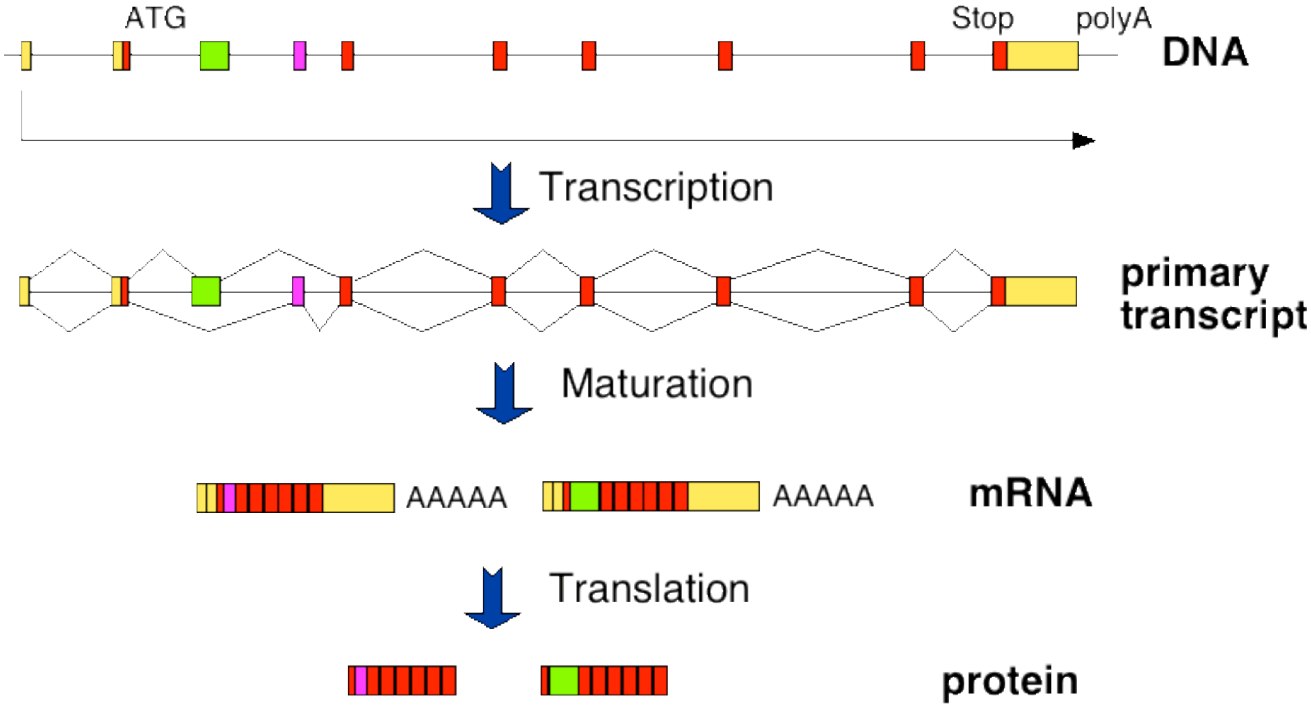
Structure des gènes humains



- Tailles moyennes
  - Gene 45 kb
  - CDS 1500 nt
  - Exon (interne) 145 nt
  - Intron 5200 nt
  - 5'UTR 210 nt
  - 3'UTR 740 nt
- Intron/exon
  - Nombres d'introns: 6 ±3 introns / kb CDS
  - Introns / (introns + CDS): 92%
- Epissage alternatif dans plus de 30% des gènes

One gene, several products

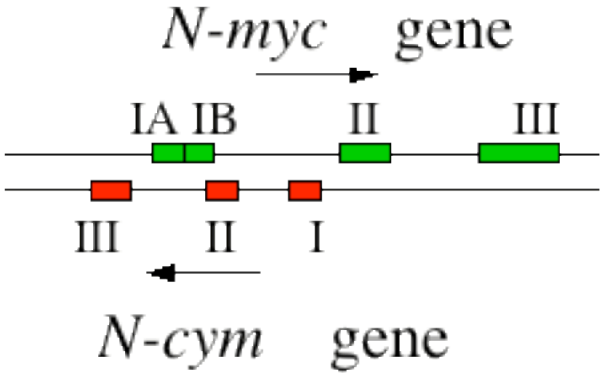
- Alternative splicing in more than 30% of human genes (Hanke et al. 1999)



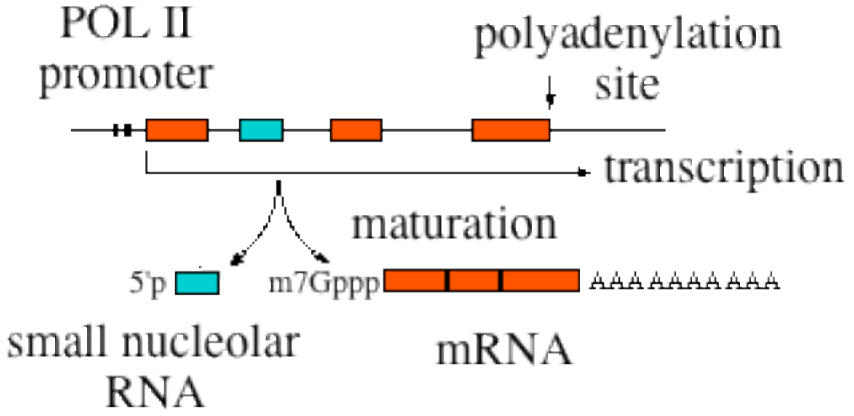
- Alternative promoter
- Alternative polyadenylation sites

Overlapping genes

Overlapping protein genes



Small nucleolar RNA genes within introns of protein genes



Gènes non-traduits dans le génome humain

- tRNA (70-100 nt) : ≈ 350 gènes
- rRNA :
  - 18S (1800 nt), 5.8S (160 nt), 28S (5000 nt) : 150-200 gènes
  - 5S (120 nt): 200-300 gènes
- snRNA (small nuclear RNA) (70-200 nt): fonctions diverses (notamment épissage: U1, U2, U4, ...): > 100 gènes
- snoRNA (small nucleolar RNA) (70-200 nt): maturation des rRNA dans le nucléole: > 100 gènes
- miRNA (micro RNA) (régulation traduction, stabilité mRNA, transcription): 250 gènes identifiés (total ??)
- autres: Xist (17 kb), H19 (2 kb), UHG (U22 snoRNA host gene, 1-2 kb), ... : ?? gènes

Repeated sequences

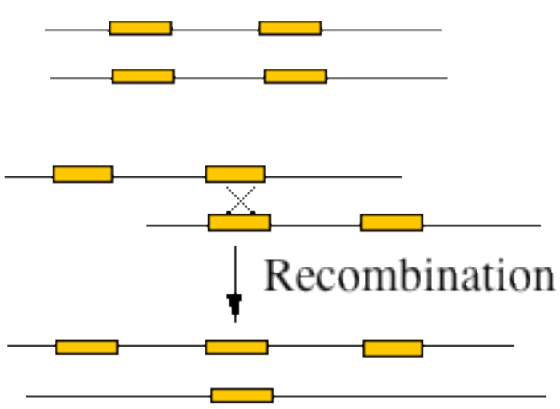
- Tandem repeats
  - Satellite
  - Minisatellite
  - Microsatellite
- Interspersed repeats
  - DNA transposons
  - Retroelements

Tandem repeats

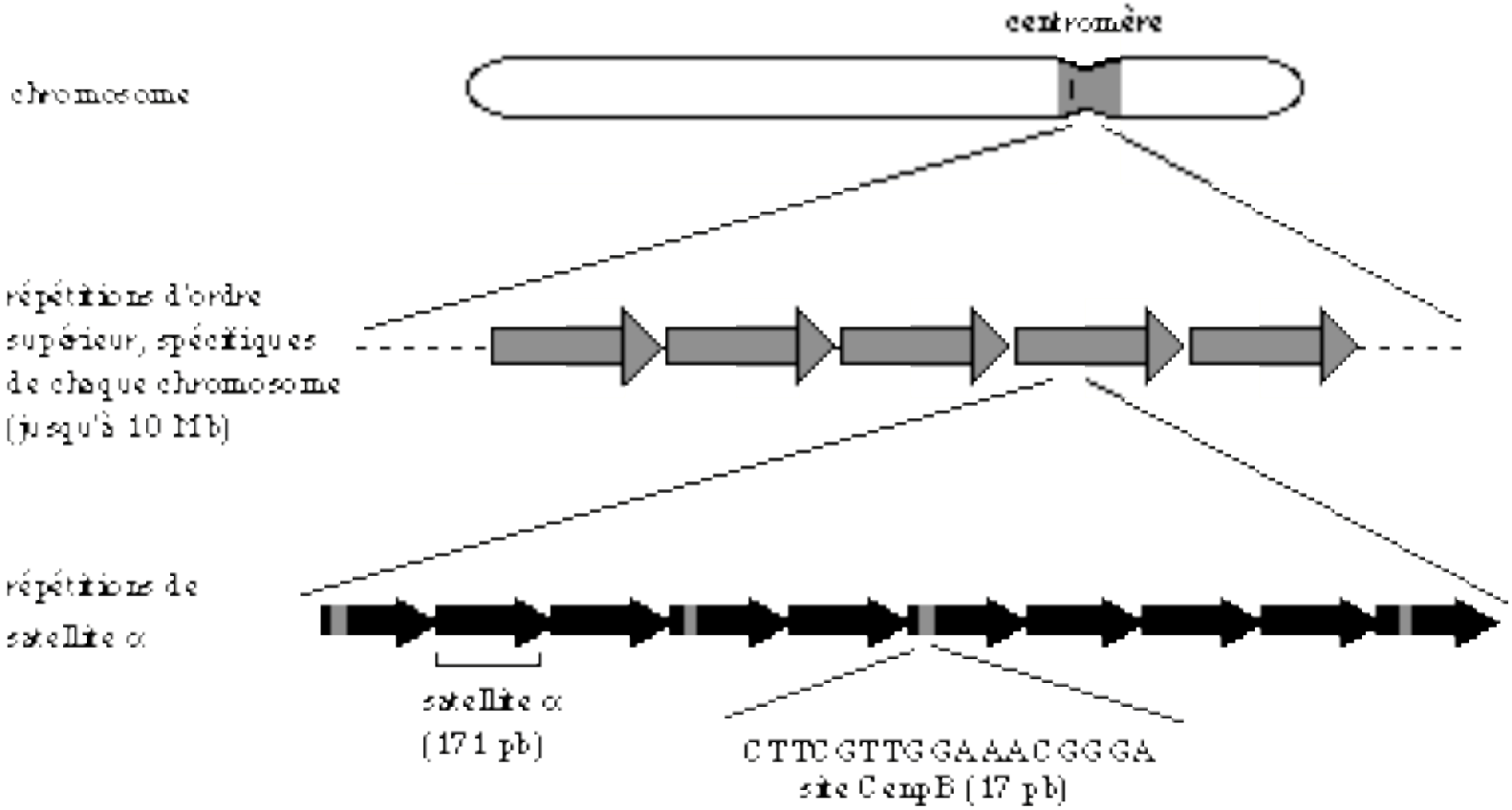
	motif	bloc size	% human genome
satellite:	2-2000 nt	up to 10 Mb	6.5%
minisatellite:	2-64 nt	100-20,000 bp	0.3%
microsatellite:	1-6 nt	10-100 bp	2%

Slippage of the DNA polymerase: CACACACACACA

Unequal crossing-over:



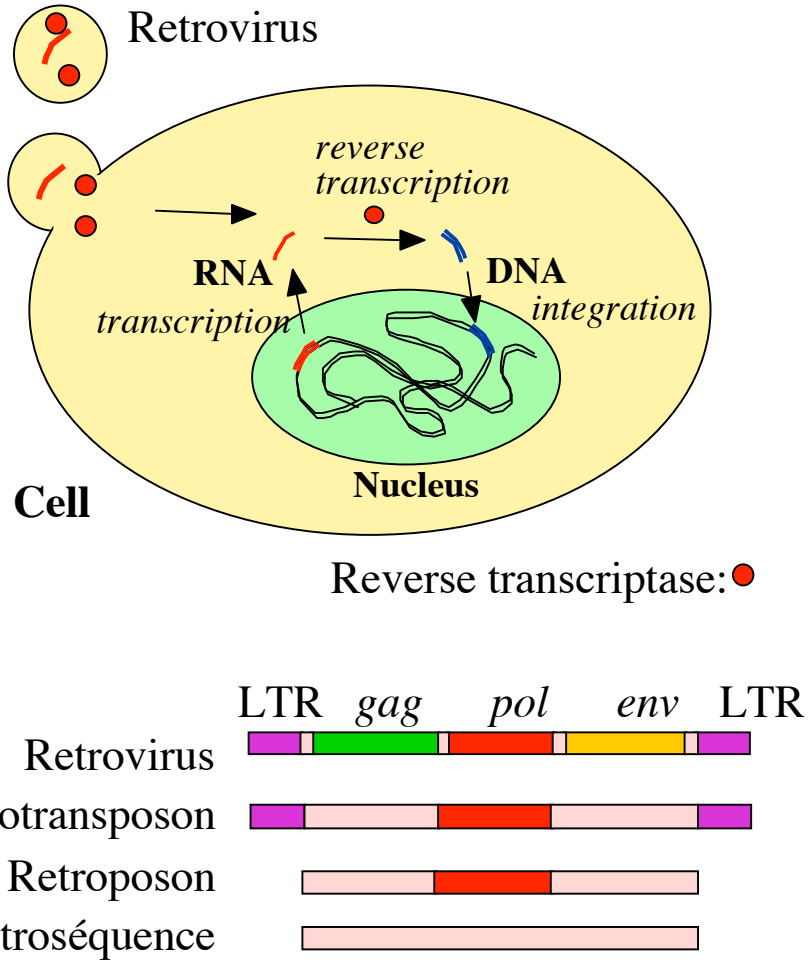
Centromeres, telomeres: Satellite DNA



Interspersed repeats

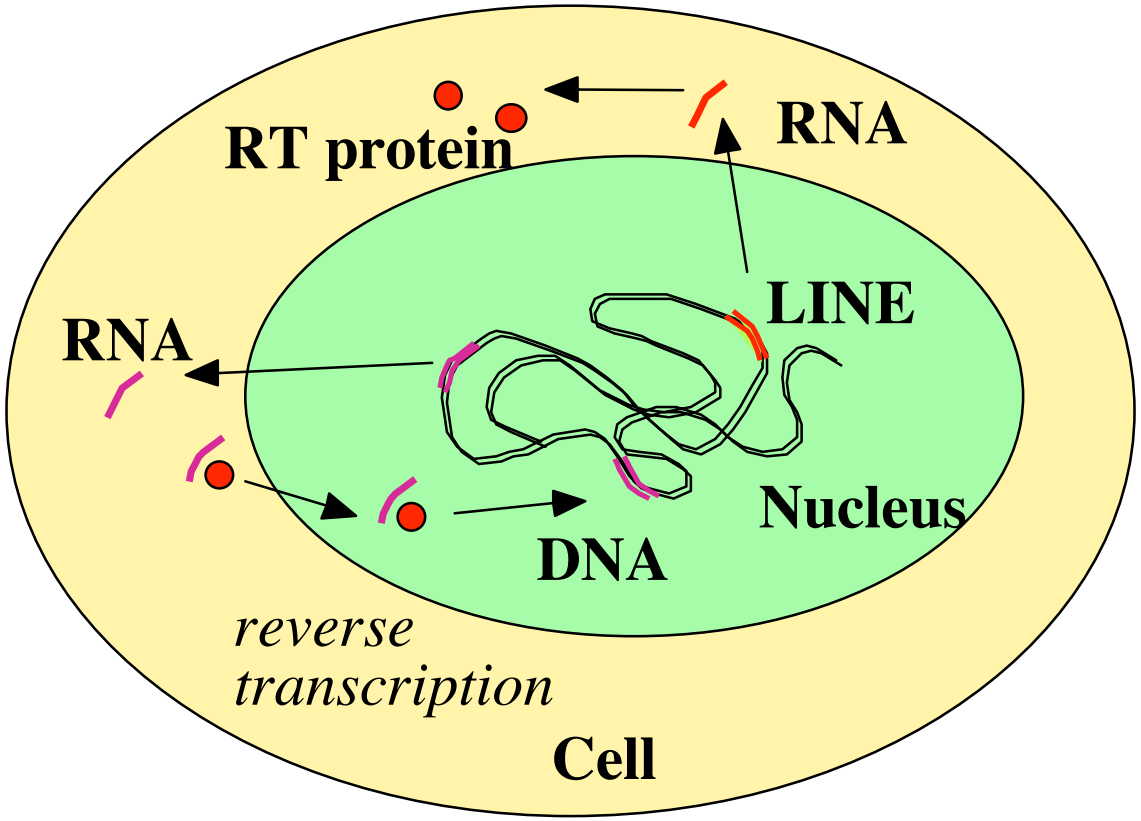
- Transposable elements (autonomous or non-autonomous) :
  - DNA transposons (rare in mammals)
  - Retroelements

Retroelements



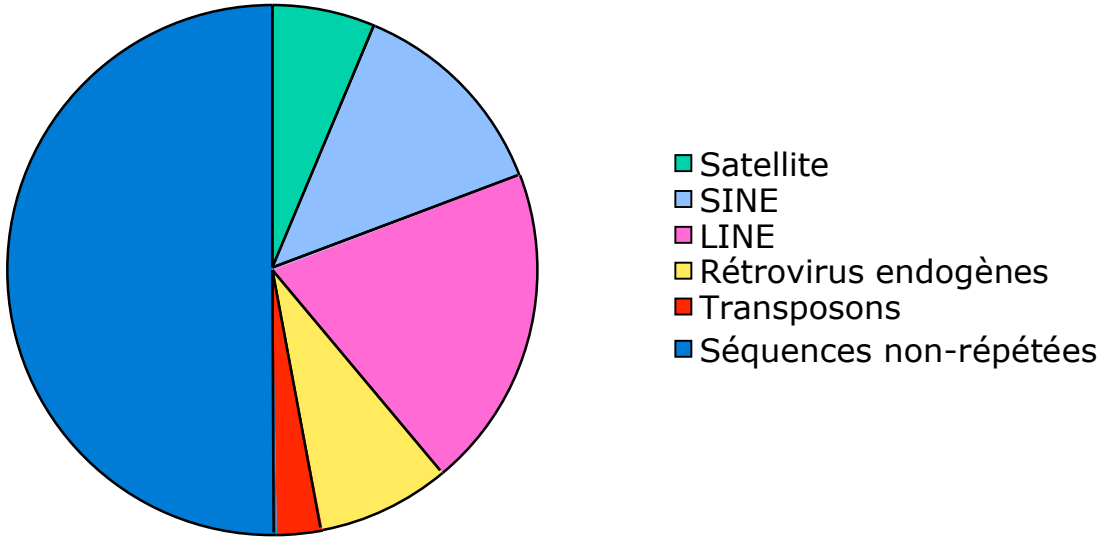
- **LINEs** (long interspersed elements): 6-8 kb retroposons
- **SINEs** (short interspersed elements):80-300 bp small-RNA-derived retrosequences (tRNA), pol III
- **Endogenous Retroviruses**: 1.5-10 kb

Retrosequences: opportunist retroelements

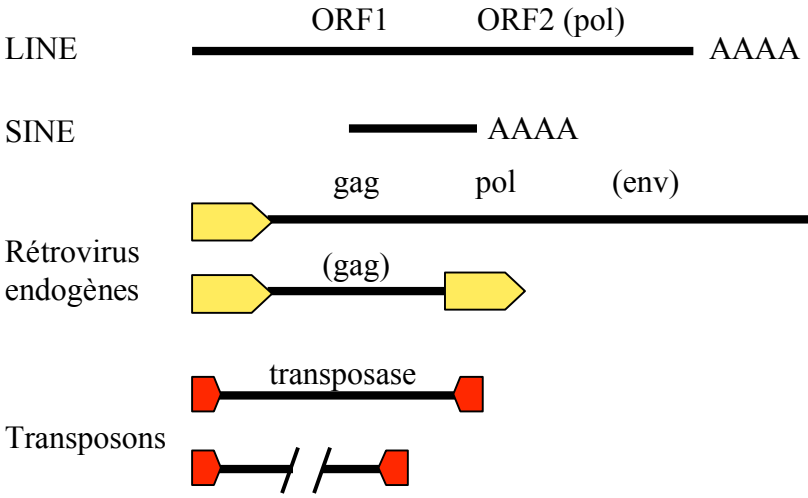


- LINE reverse transcriptase

Contenu en séquences répétées du génome humain



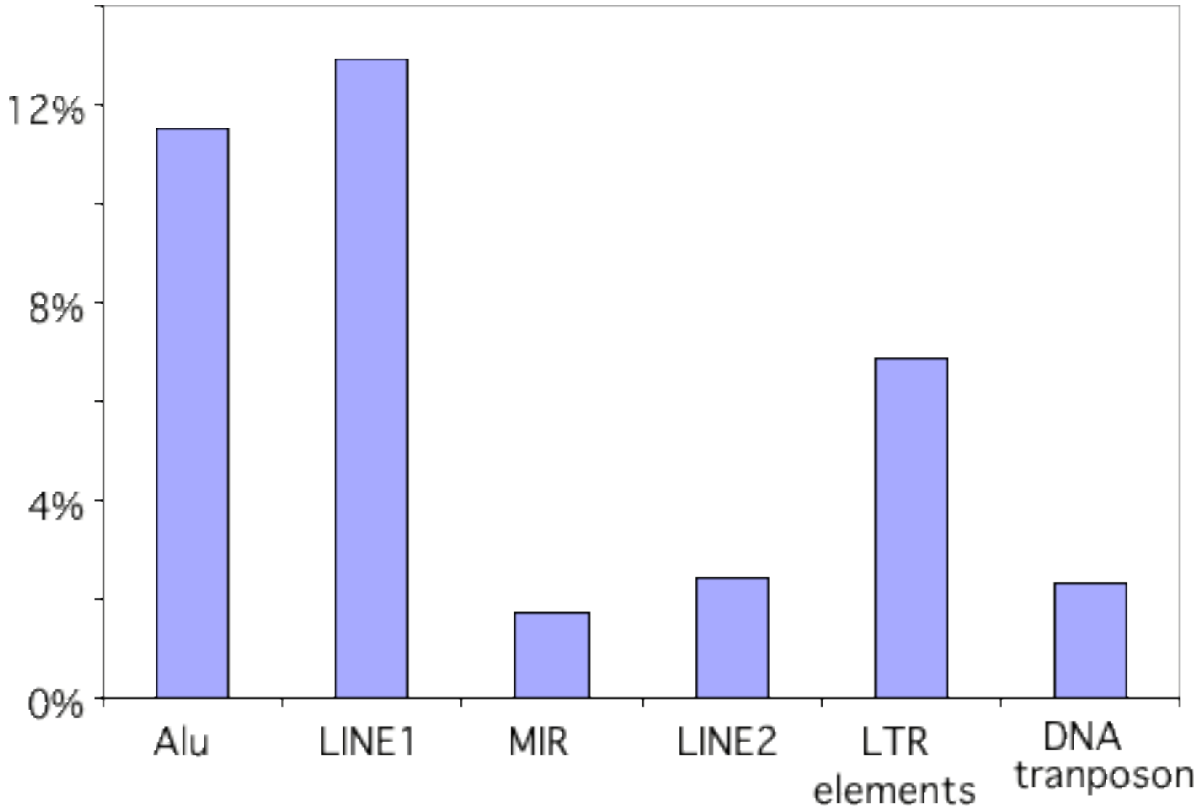
Principales classes d'éléments transposables dans le génome humain



		Taille	Nombre de copies	Fraction du génome
LINE	autonome	6-8 kb	850 000	20%
SINE	non-autonome	100-300 pb	1 500 000	13%
Rétrovirus endogènes	autonome	6-11 kb	450 000	8%
	non-autonome	1,5-3 kb		
Transposons	autonome	2-3 kb	300 000	3%
	non-autonome	80-3000 pb		

Frequency of transposable elements in the human genome

- Total = 46% (Smit 1999, MGSC 2002)
- Probably underestimated



Vagues d'insertions dans les génomes de mammifères

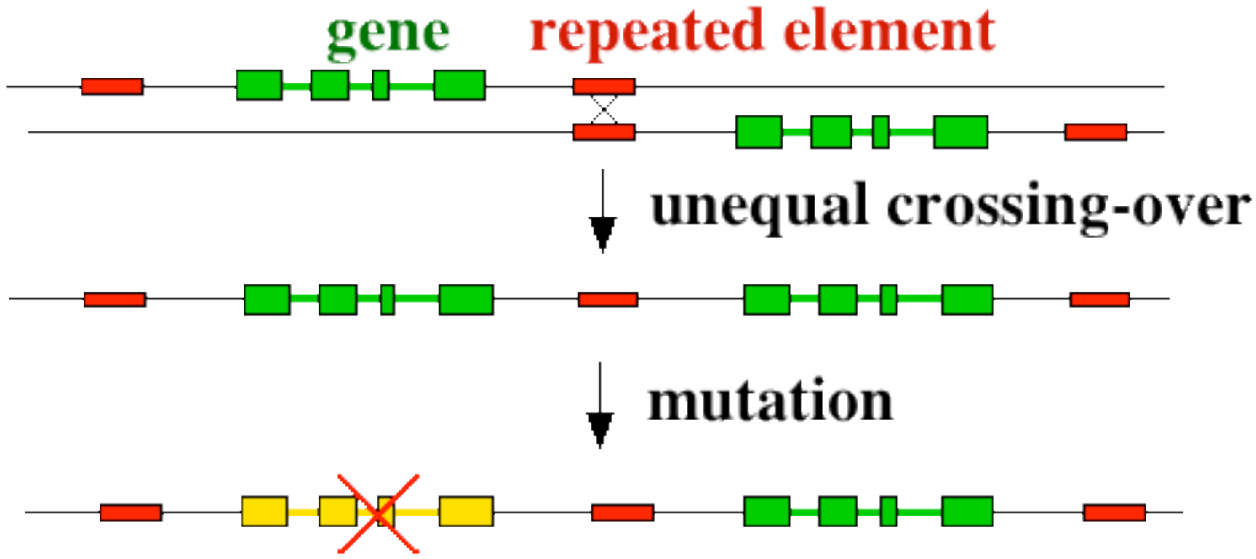
- La plupart des copies présentes dans le génome de mammifères sont tronquées et ont perdu la capacité de transposer (délétions, mutations non-sens)
- Par exemple, LINE1: 500,000 copies dans le génome humain, dont seulement 80-100 encore actives
- Vagues successives d'insertions

Comment expliquer l'omniprésence des éléments transposables dans les génomes ?

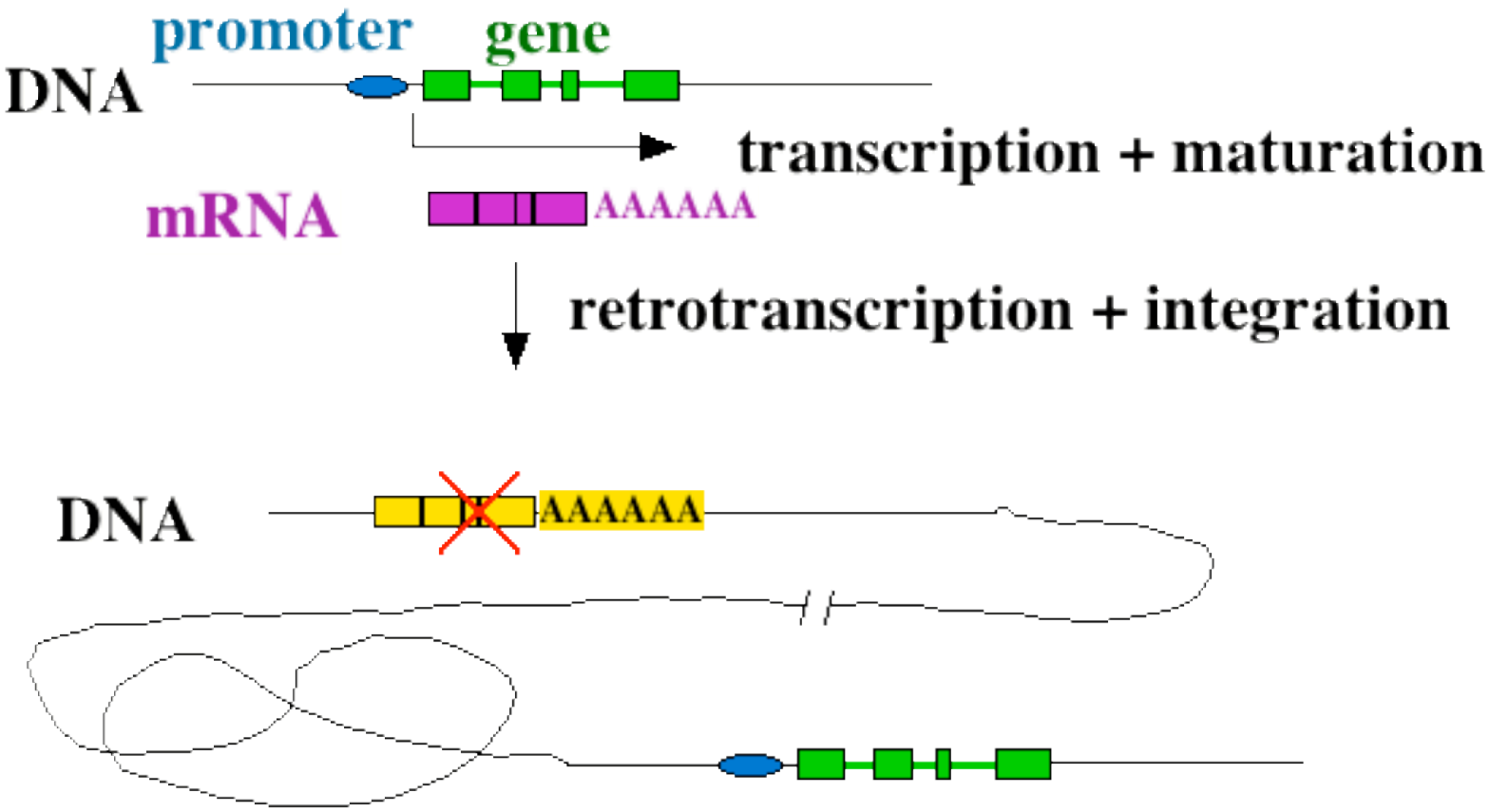
- Hypothèse 1: fonction structurale pour l'organisation du noyau
- Hypothèse 2: source d'innovations bénéfiques pour l'adaptation des espèces
  - Exemples d'éléments transposables domestiqués (régions codantes, régulatrices)
  - Sélection de second ordre
- Hypothèse 3: ADN « égoïste »
  - Conflits entre éléments transposables et leur génome hôte

Pseudogenes

- After a gene duplication:
  - evolution of new function (sub-functionalization or neo -functionalization)
  - or gene inactivation



Retropseudogenes



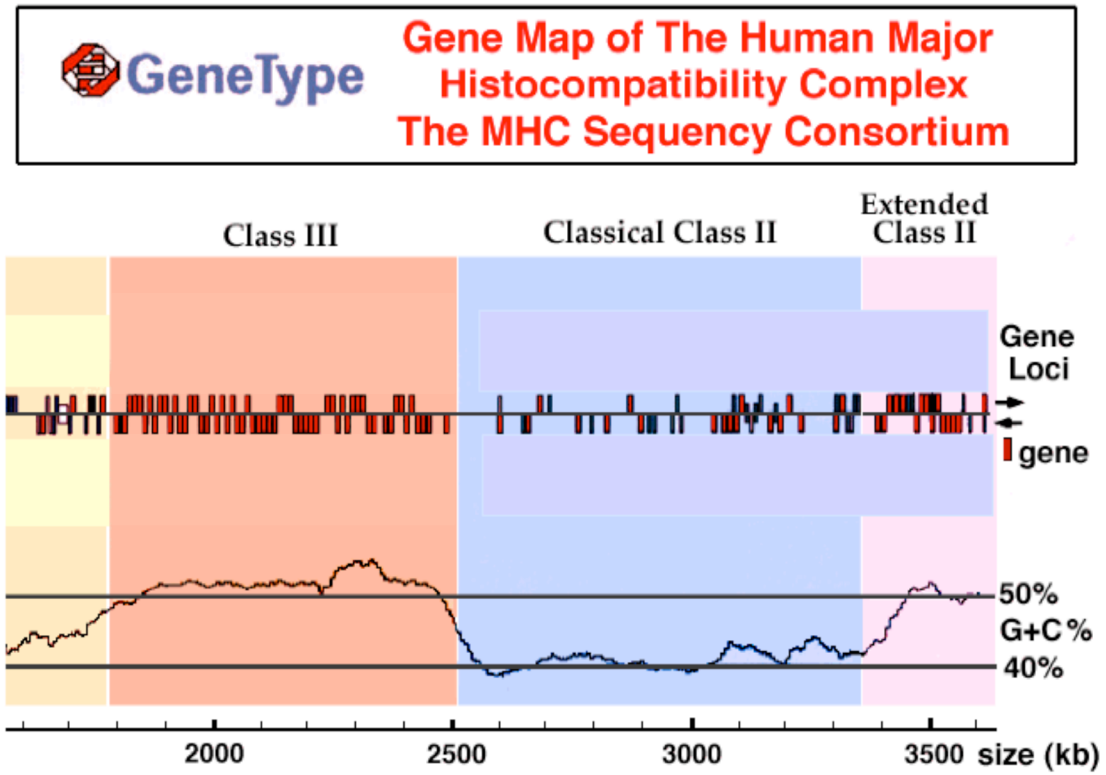
Pseudogenes

- About 20,000 retropseudogenes in the human genome
- Often derive from housekeeping genes

- About 2,000 unprocessed pseudogenes

Vertebrate genome organization: variations of base composition along chromosomes

Sequence of human MHC

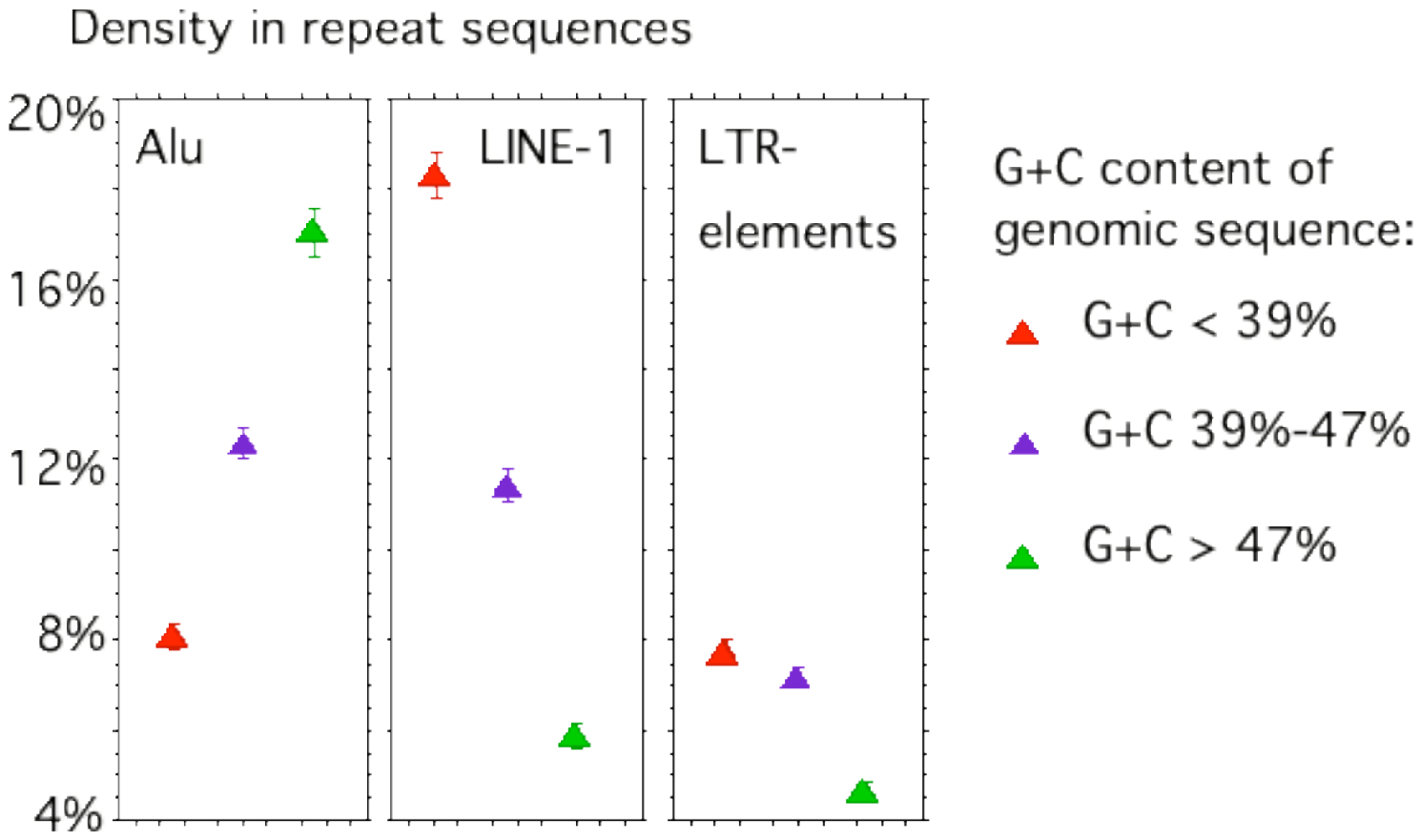


Isochore organization of vertebrate genomes

- Large scale variations of GC-content along chromosomes
  - Affect both coding and non-coding regions (introns, intergenic regions)
  - Large regions (> 300 kb) of relatively homogenous base composition = **isochores**
  - Bernardi et al. (1985, 2000)

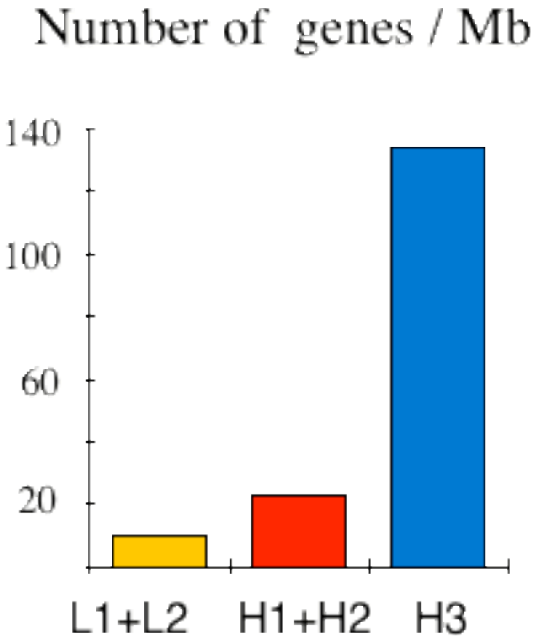
- Correlation wih other genomic features:
  - Insertion of repeated sequences (A. Smit 1996)
  - Recombination frequency (Eyre-Walker 1993)
  - Chromosome banding (Saccone, 1993)
  - Replication timing (Bernardi, 1998)
  - Gene density (Mouchiroud, 1991)
  - Gene structure (Duret, 1995)
  - Gene expression ?? -> No

Isochores and insertion of repeat sequences (Smit 1999)

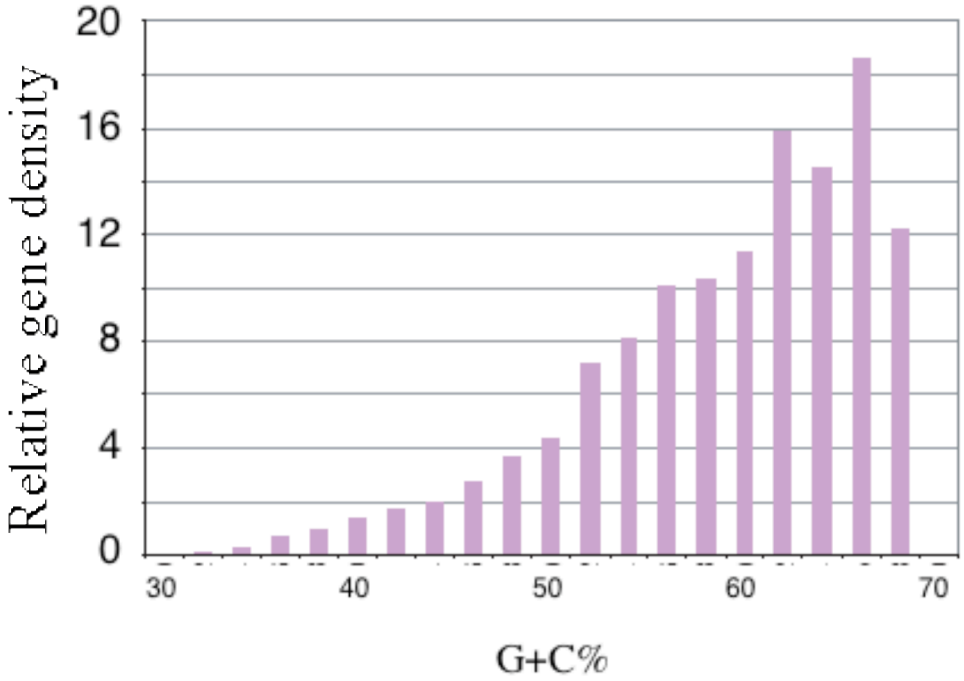


4419 human genomic sequences > 50 kb

Isochores and gene density

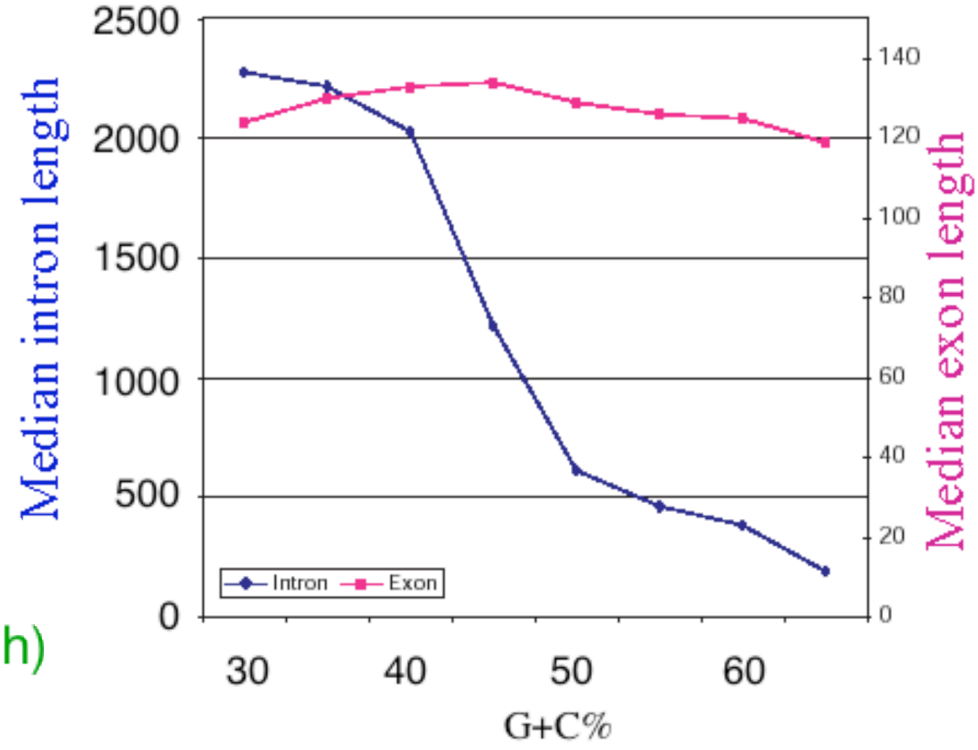


Mouchiroud et al 1991



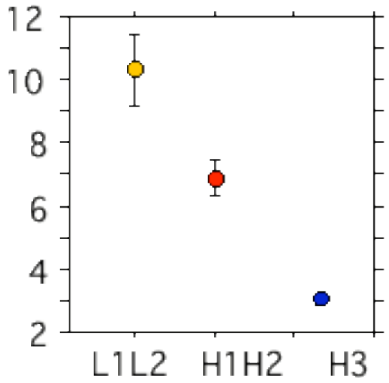
IHGSC 2001

Isochores and introns length



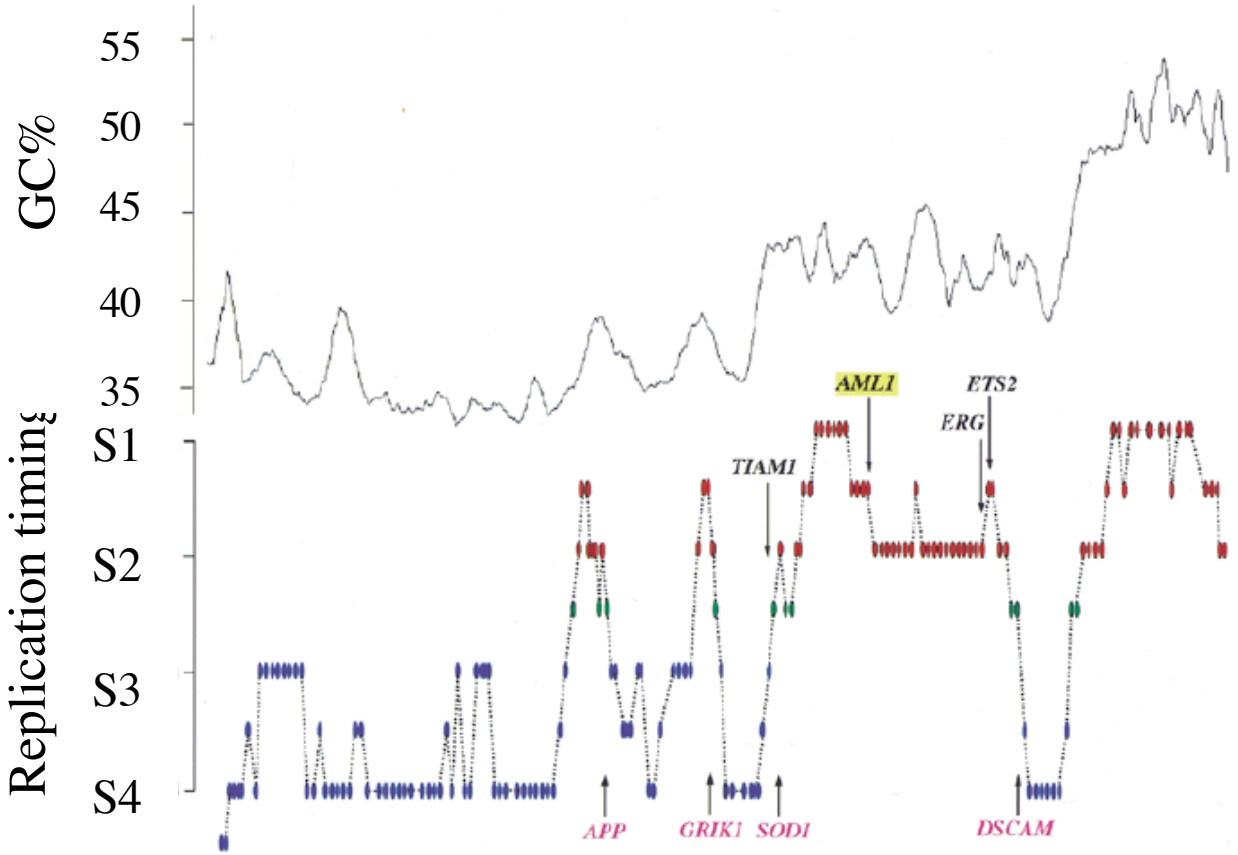
IHGSC 2001

Gene compaction
(intron length/coding region length)



Isochores and replication timing

(Watanabe et al. 2002)

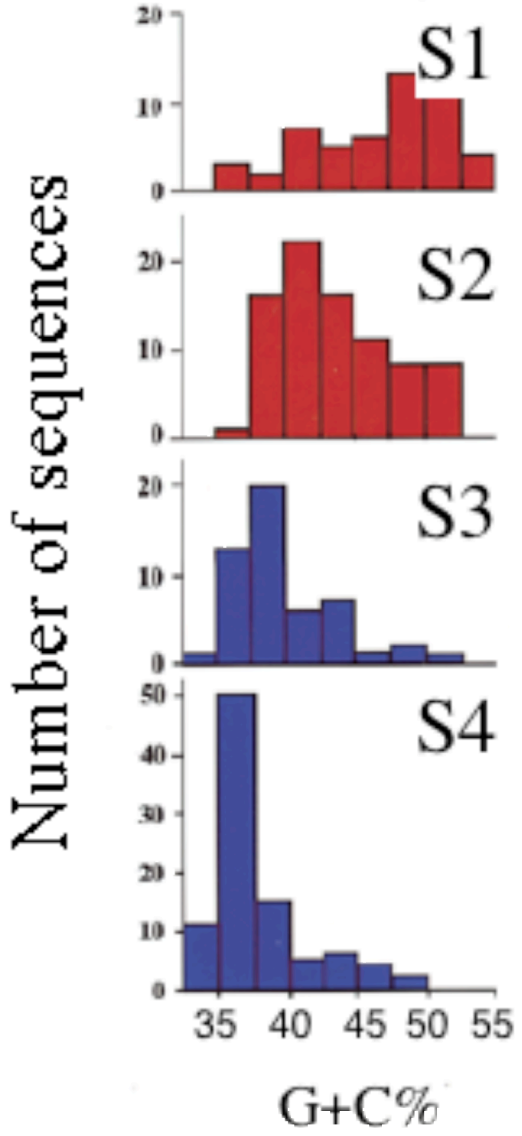


Chromosome 21q (~35 Mb)

Isochores and replication timing

(Watanabe et al. 2002)

- Human chromosomes 21q and 11q (total: 180 Mb)



Human genome: summary

- Genes, regulatory elements: ~ 2-5%
- Non-coding sequences: ~ 95-98%
  - Satellite DNA (centromeres) ~ 6-7%
  - Microsatellites ~ 2%
  - Transposable elements ~ 46%
  - Pseudogenes ~ 1%
  - Other (ancient transposable elements?) ~ 42%
- Variations in gene and repeat density along chromosomes

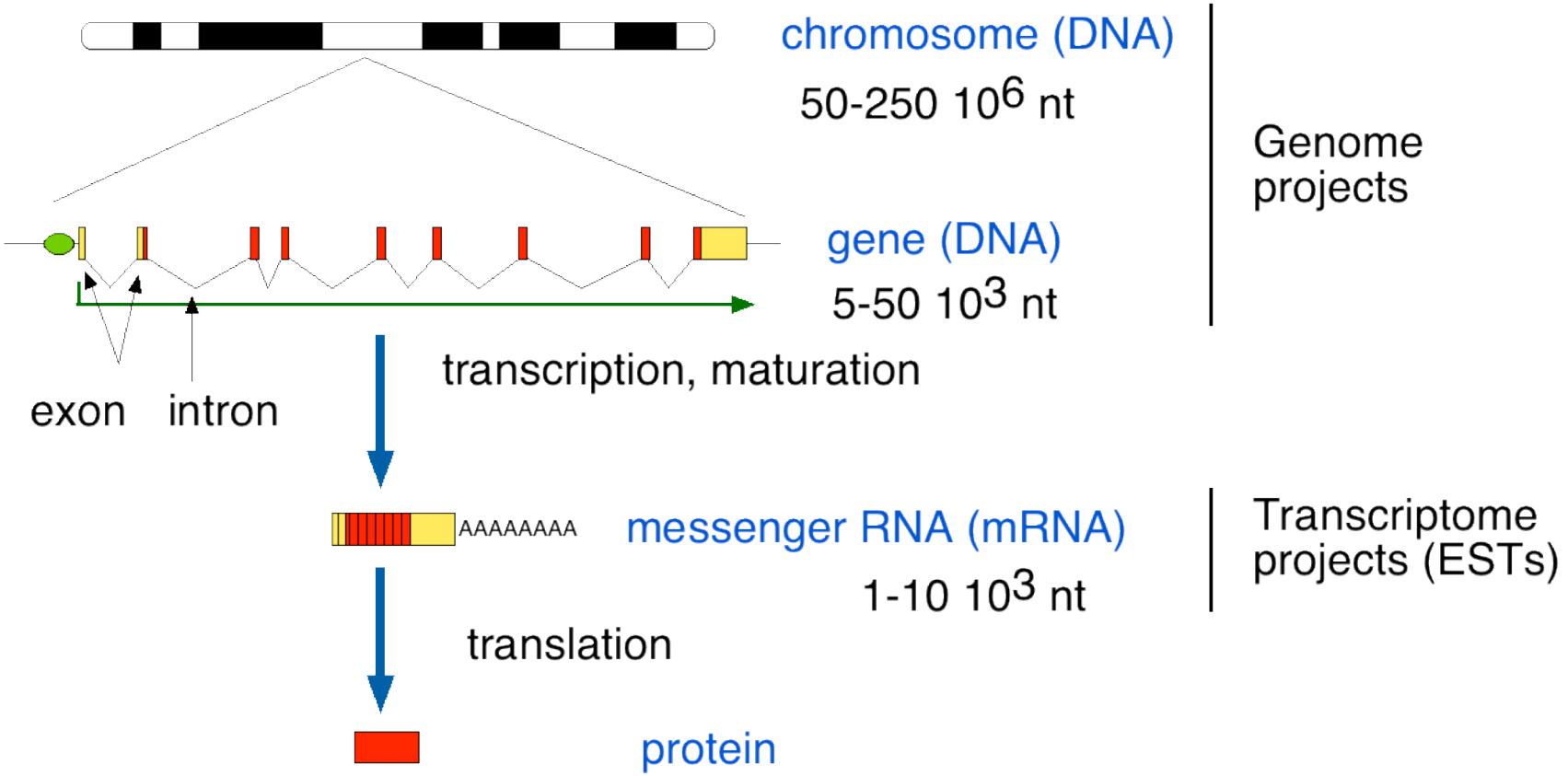
Séquençage de l'ADN: historique

- 1943-1953: ADN support de l'information génétique
- 1977: techniques modernes de séquençage de l'ADN (Maxam & Gilbert, Sanger *et . al*)
- 1982: création des premières banques de données de séquence (GenBank, EMBL)
- 1990: début du projet génome humain (cartographie)
- 1995: premier génome complet d'un organisme cellulaire (*H. influenzae*)
- 2001: première ébauche du génome humain
- 2004: environ 200 génomes complets
- 2005: nouvelle technique de séquençage (x 30)

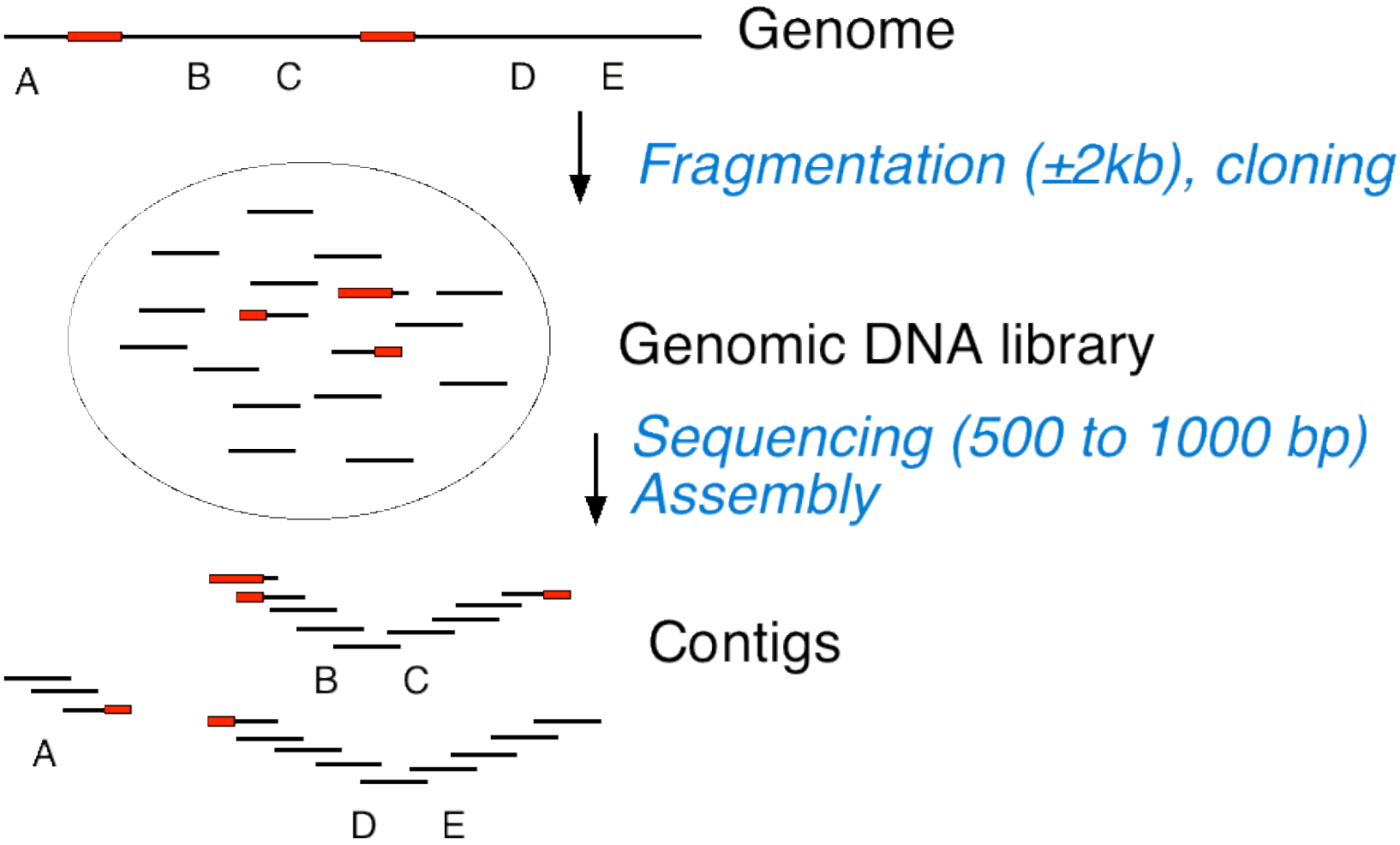
Les projets génomes: objectifs

- faire l'inventaire de l'information génétique nécessaire au développement et à la reproduction des organismes
- comprendre l'organisation du génome (système d'information intégré ?)
- comprendre l'évolution des génomes
- applications médicales, agronomiques, industrielles

Projets de séquençage : Génome / Transcriptome

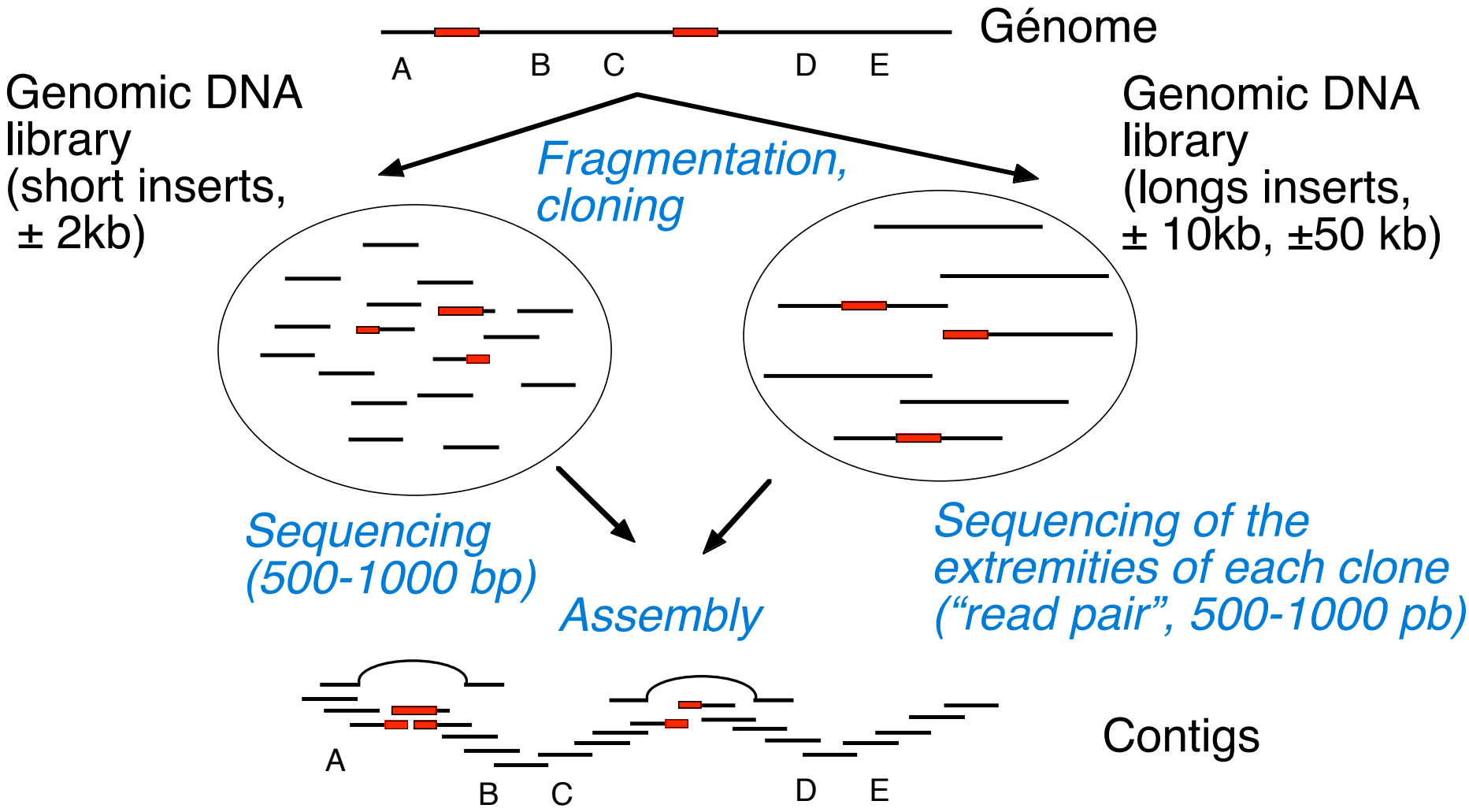


Shotgun sequencing



Shotgun sequencing: improvement

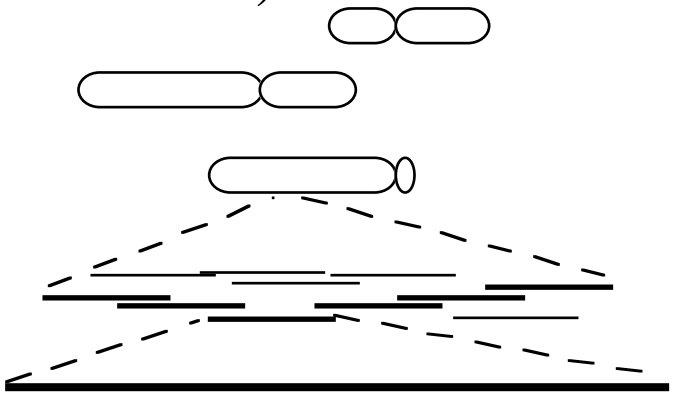
(E. Myers)



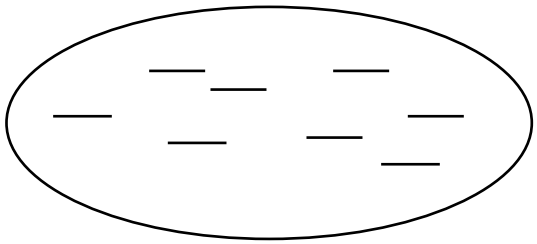
Strategy for sequencing the human genome

(Academic international consortium)

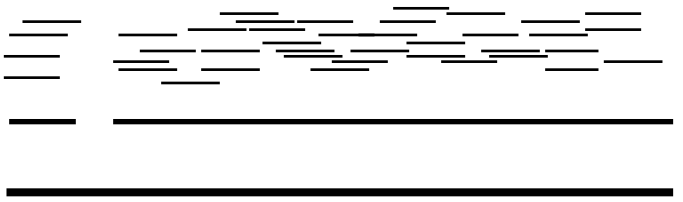
- Genome
- Cloning of long inserts (e.g. BAC DNA library : 100-200 kb)
- Genomic mapping
- Selection of clones to sequence



- Sub-cloning of short inserts (e.g. M13 DNA library : 1-20 kb)

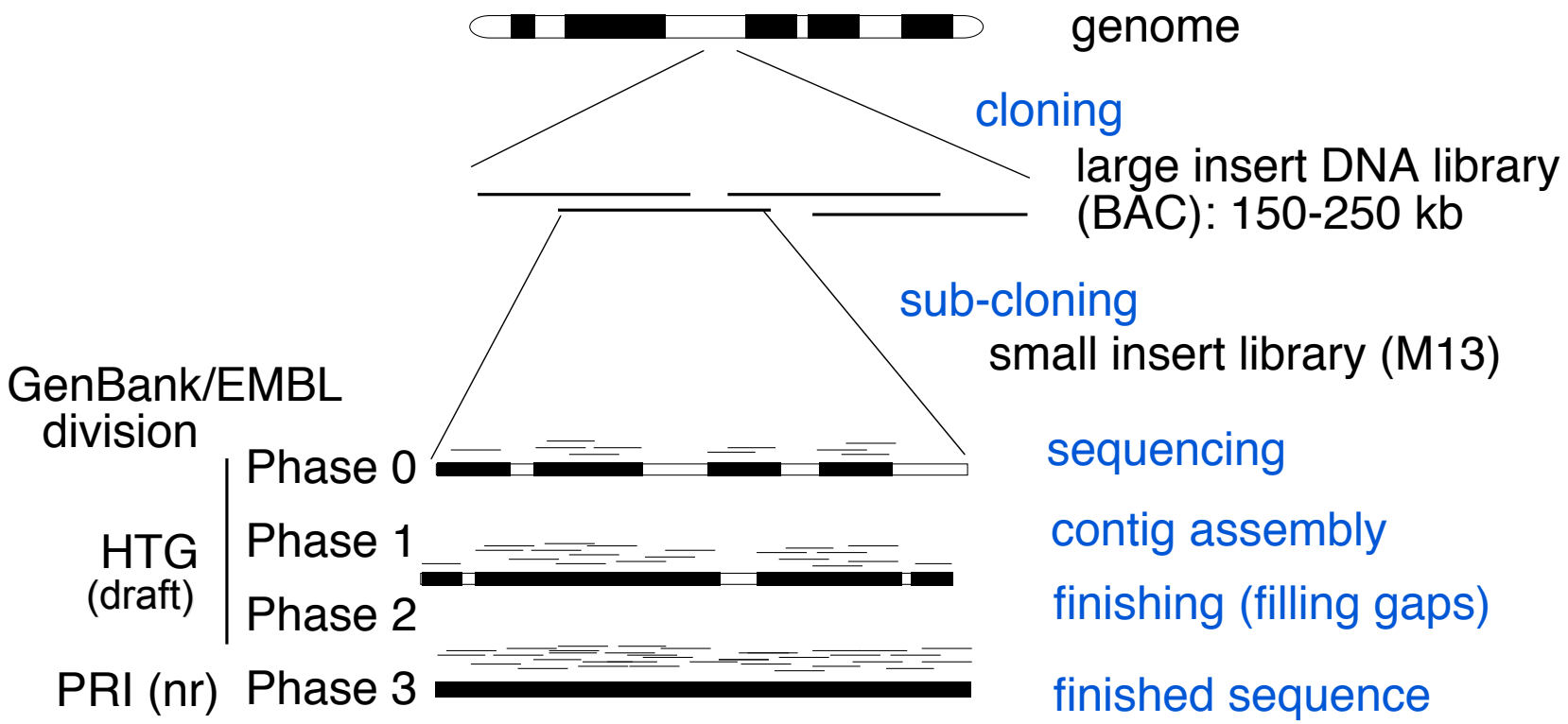


- Sequencing M13 clones
- Assembly: contigs
- Finishing: gap closure



Genomic Sequences

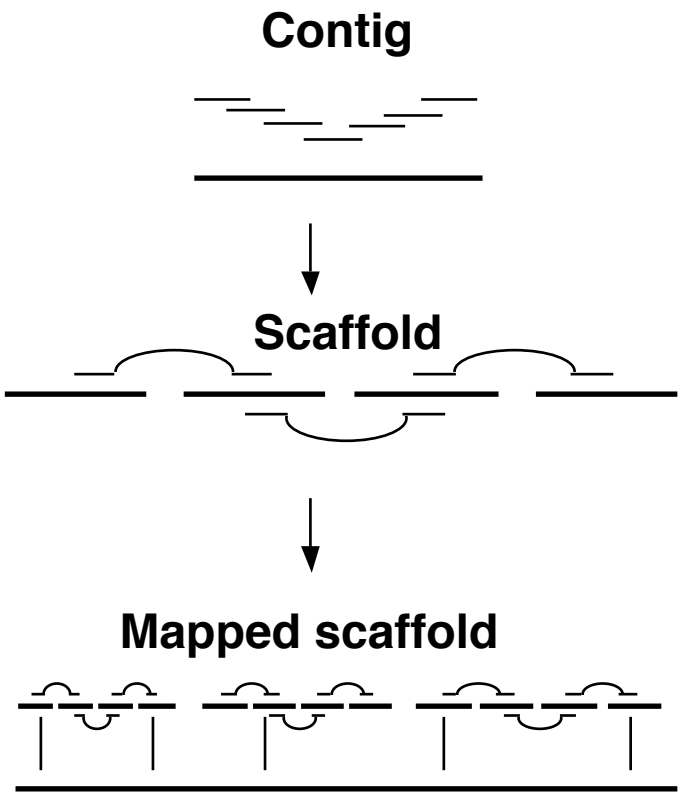
GenBank/EMBL HTG division : High Troughput Genome sequences



Phase 0	single-few pass reads of a single clone (not contigs).
Phase 1	Unfinished, may be unordered, unoriented contigs, with gaps.
Phase 2	Unfinished, ordered, oriented contigs, with or without gaps.
Phase 3	Finished, no gaps (with or without annotations)

Complete genome sequence ?

- **Contig**: overlapping sequences without any gap
- **Scaffold**: set of ordered and orientated contigs; gaps of known length
- **Mapped scaffold**: set of scaffold localized along chromosomes (but not always ordered and orientated, gaps of unknown length)
- 2002: 97% of the human genome is sequenced, but only 85% in scaffolds assembly = draft sequence



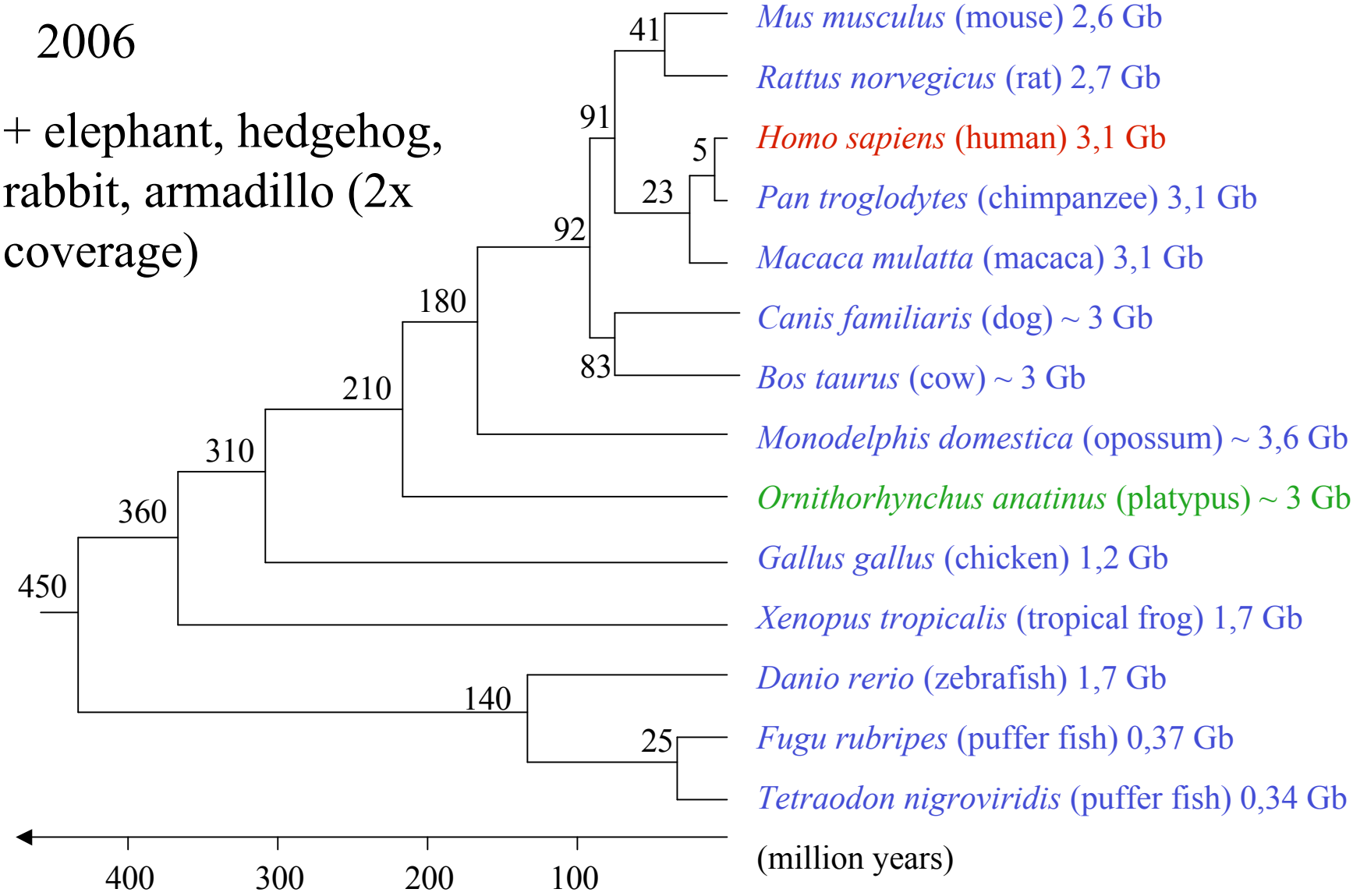
http://genome.ucsc.edu/

Genome Projects

- 263 complete genomes publically available (Oct. 2005) (215 bacterias, 22 archeaes, 26 eukaryotes)
- The 10 largest sequenced genomes
  - *Tetraodon nigroviridis* 365 Mb (2004) (draft)
  - chicken 1200 Mb (2004) (draft)
  - zebrafish 1600 Mb (2004) (draft)
  - mous 2600 Mb (2002) (draft)
  - rat 2600 Mb (2003) (draft)
  - chimpanzee 3000 Mb (2004) (draft)
  - human 3000 Mb (2001-2003)
  - opossum 3000 Mb (2005) (draft)
  - dog 3000 Mb (2005) (draft)
  - cow 3000 Mb (2005) (draft)

2006

+ elephant, hedgehog, rabbit, armadillo (2x coverage)



Complete sequence, finished assembly

Nearly complete sequence; preliminary assembly (draft)

Sequencing in progress

Projets EST (transcriptome)

- Expressed Sequence Tags (EST)
- Inventaire des ARNm exprimés par un organisme, dans différents tissus, stades de développement, pathologies, …
- Extraction et clonage des ARNm (banques d ’ADNc)
- Séquençage systématique des clones
  - Séquences partielles d ’ARNm (300-500 nt) - mauvaise couverture des extrémités 5’
  - Erreurs de séquence (1-3%)
  - Redondance (gènes fortement exprimés)
  - Qualité suffisante pour identifier un gène
- Automatisation
- Possibilité d’obtenir les clones cDNA (consortium IMAGE) (http://image.llnl.gov/)

Large scale EST projects

	Number of ESTs (Feb. 05)
• *Homo sapiens*	6,029,601
• *Mus musculus* (mouse)	4,329,768
• *Rattus sp*. (rat)	691,985
• *Ciona intestinalis*	684,319
• *Xenopus tropicalis* (tropical frog)	634,367
• *Danio rerio* (zebrafish)	592,837
• *Triticum aestivum* (wheat)	587,650
• *Gallus gallus* (chicken)	531,351
• *Bos taurus* (cow)	513,065
• *Xenopus laevis* (xenope)	449,492
• *Zea mays* (maize)	417,803

Projets GSS

- Genome Survey Sequence (GSS)
- Echantillonage aléatoire de séquence génomiques: donner un premier aperçu du contenu d'un génome
- Banques d ’ADN génomique
- Séquençage systématique de clones
  - Séquences courtes (< 1kb)
  - Erreurs de séquence (1-3%)
  - Qualité suffisante pour identifier un gène
- Automatisation

Projets GSS à grande échelle

	Nombre de GSS (Sept. 2003)
• *Mus musculus* (souris)	952,000
• *Homo sapiens*	876,000
• *Zea mais* (maïs)	642,000
• *Brassica* (choux fleur)	567,000
• *Rattus norvegicus* (rat)	307,000
• *Arabidopsis thaliana* (arabette)	240,000
• *Tetraodon nigroviridis*	189,000
• *Danio rerio* (poisson zèbre)	159,000
• *Pan troglodytes* (chimpanzé)	158,000
• *Gallus gallus* (poulet)	137,000

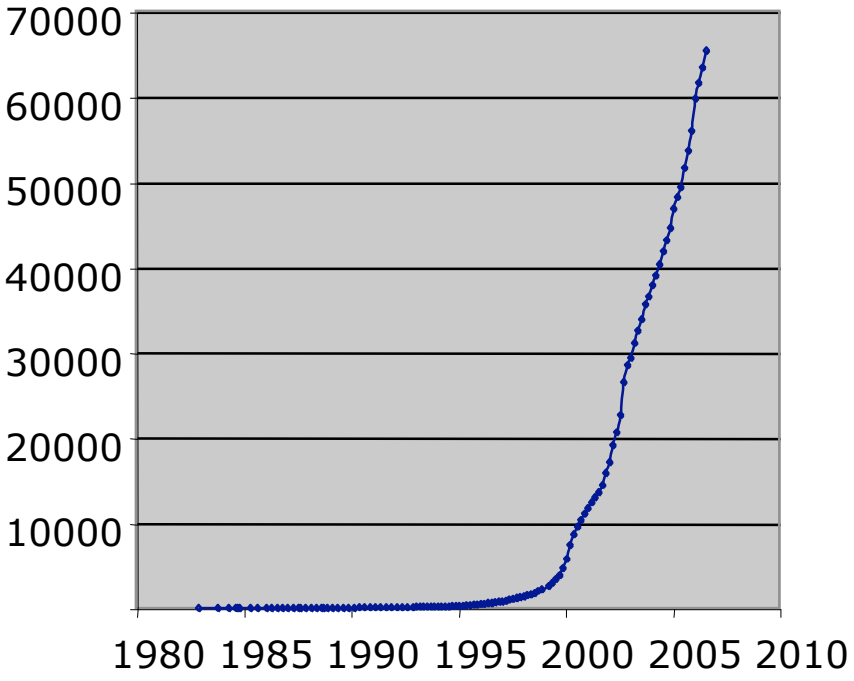
Different types of nucleotide sequences in current databases

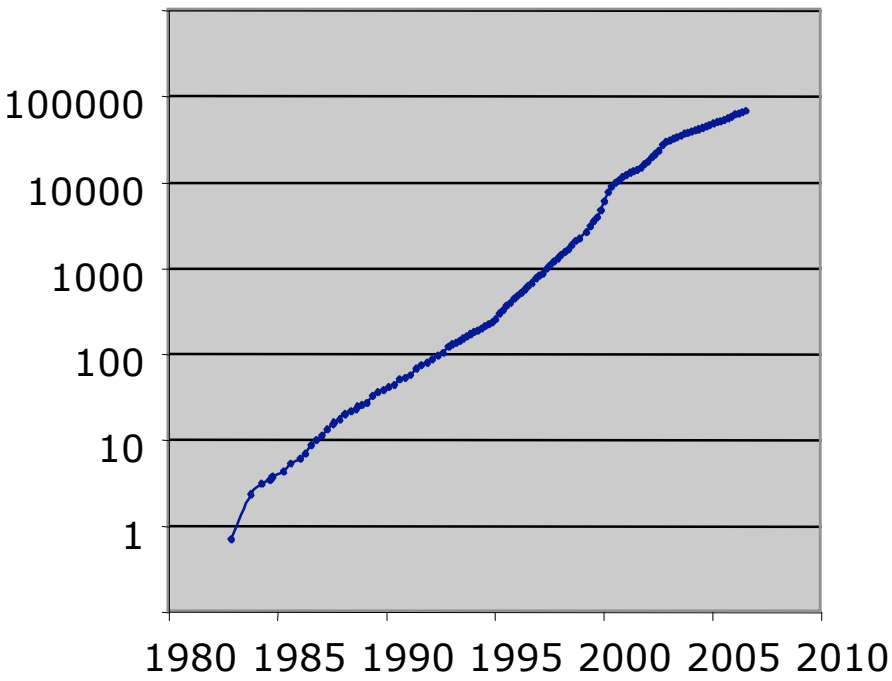
	Standard	High throughput genome (HTG)	Genome survey sequence (GSS)	Expressed sequence tags (EST)
Contents	biologically characterized genes and RNAs, finished clones from genome projects	unfinished clones from genome projects	single pass sequences from random genomic clones	single pass sequences from random cDNA clones
Length	variable	>20,000 bp	<1,000 bp	<1,000 bp
Accuracy	medium-high	high	low	low
Annotation	medium to high, rich biological annotation	technically use- ful, biologically poor	technically use- ful, biologically poor	technically use- ful, biologically poor

Augmentation exponentielle des données du séquençage

- Doublement tous les 18 mois

Quantité de séquences publiées (Mb)





Pas seulement des séquences ...

- Expression des gènes (puces à ADN, filtres haute densité, SAGE, EST…)
- Polymorphisme (SNP, microsatellites)
- Structure 3D des protéines
- Interactions moléculaires
- ...

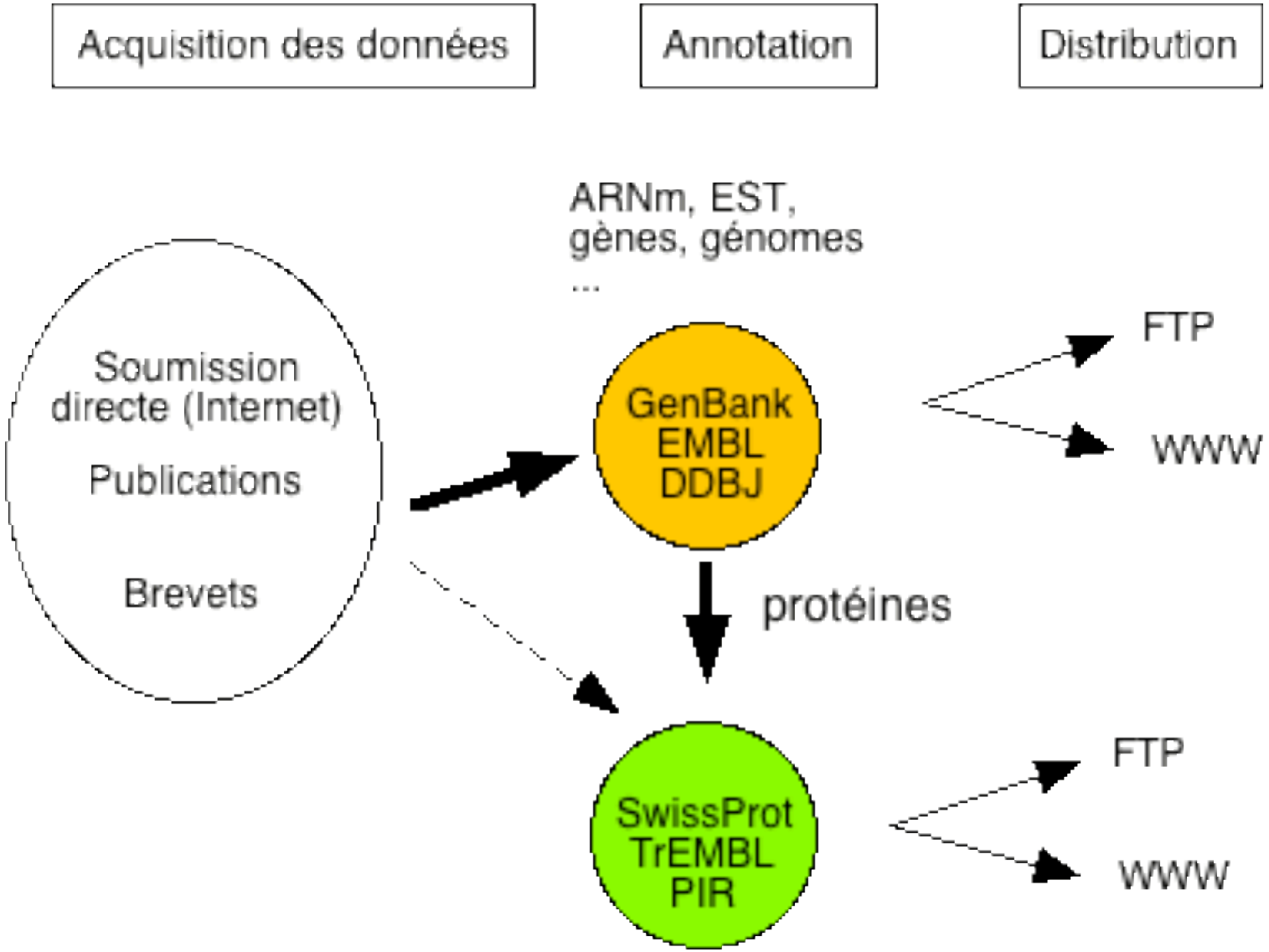
Banques de données en biologie moléculaire

- Séquences
  - Banques généralistes (nucléiques, protéiques)
  - Banques spécialisées
- Structure des protéines
- Cartographie génomique
- Maladies génétiques, phénotypes
- Bibliographie
- …

Les banques de séquences généralistes

- Banques de séquences nucléotidiques :
  - EMBL (Europe) (1980)
  - GenBank (USA) (1979)
  - DDBJ (Japon) (1984)
  - Ces 3 centres échangent leurs données quotidiennement
    => contenu identique

- Banques de séquences protéiques :
  - SwissProt-TrEMBL (Suisse, Europe) (1986 et 1996) => UniProt


Acquisition des données
Annotation
Distribution
ARNm, EST,
gènes, génomes
...
Soumission
directe (Internet)
Publications
Brevets
GenBank
EMBL
DDBJ
FTP
www
protéines
SwissProt
TrEMBL
PIR
FTP
www

Taille de GenBank/EMBL

(Septembre 2006)

- 67 10^9 nucléotides.
- 43 10^6 séquences.
- 3 364 000 gènes (protéines et ARN).
- 439 000 références bibliographiques.
- 236 giga-octets sur disque.
- Augmentation de 200% en 24 mois.

GenBank release 155 (August, 2006)

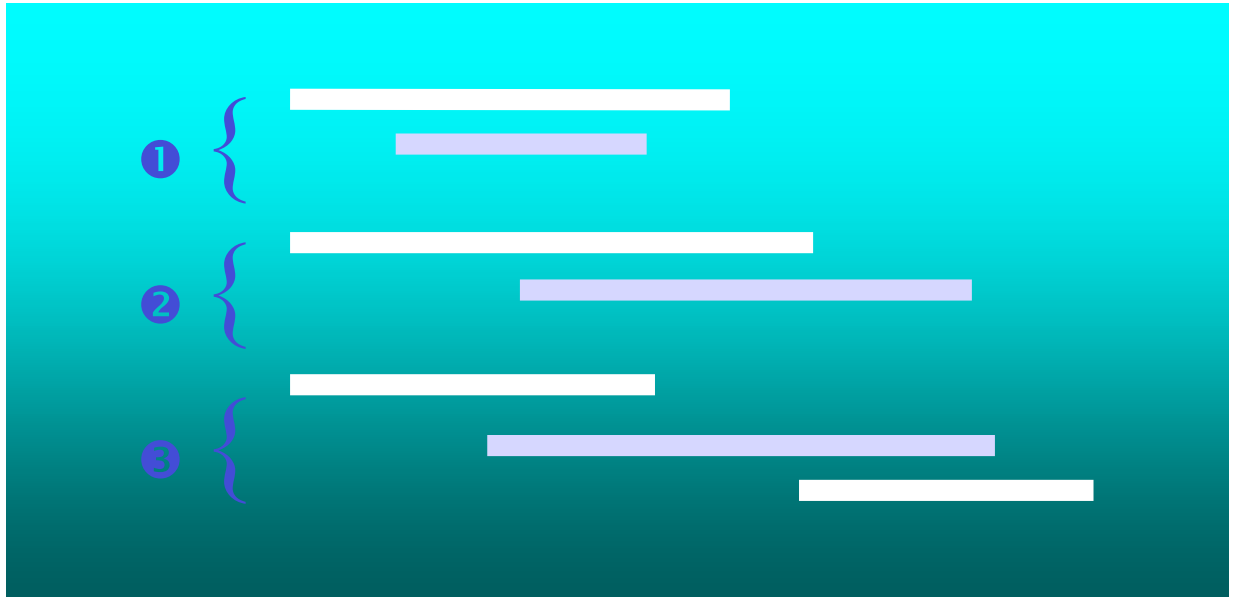
Division	Nucleotides	% nt
EST	21.4 Gb	32%
HTG	16.2 Gb	24%
GSS	9.9 Gb	15%
Other	19.5 Gb	29%
Total	67.0 Gb	100%
Human	12.4 Gb	19%

Contenu des banques de séquences nucléiques: échantillonage taxonomique

- 268,000 espèces pour lesquelles on a au moins une séquence
- 10 espèces (0.004%) représentent à elles seules 60% des séquences
  - Homo sapiens 19%
  - Mus musculus (souris) 12%
  - Rattus norvegicus (rat) 9%
  - Bos taurus (vache) 5%
  - Danio rerio (poisson zèbre) 4%
  - Zea mays (maïs) 3%
  - Oryza sativa (riz) 2%
  - Strongylocentrotus purpuratus (oursin) 2%
  - Sus scrofa (porc) 2%
  - Xenopus tropicalis (grenouille tropicale) 1%

La redondance

- Un problème majeur des banques est celui de la redondance.



Variations dans les séquences

- Les doublons présentent fréquemment des variations dans leurs séquences et leurs annotations.
- Il est impossible de décider si ces différences sont issues :
  - D'un polymorphisme.
  - D'erreurs de séquençage.
  - De duplications de gènes.
- GenBank: 20% de redondance parmi les séquences de gènes protéiques de vertébrés; 40% de redondance parmi les séquences génomiques humaines

UniProt: SWISS-PROT et son complément TrEMBL

- Collaboration entre le *Swiss Institute of Bioinformatics* (SIB) et l'EBI (EMBL).
- SwissProt:
  - Expertise manuelle poussée: annotations riches (fonction des protéines, localisation sub-cellulaire, strucutre, modifications post-traductionnelles, ...)
  - Redondance minimale
  - Incomplète
- TrEMBL: traduction des séquences codantes d 'EMBL qui ne sont pas déjà dans SwissProt
  - Annotation automatique: annotations moins riches
- UniProt: SwissProt+TrEMBL: ensemble complet, redondance minimale

Banques de séquences spécialisées

- *PROSITE, PFAM, PRODOM, PRINTS, INTERPRO* : banques de motifs protéiques
- *Protein Data Bank* (PDB): structures 3D (protéines, DNA, RNA)
- *Ribosomal Database Project* (RDP) : rRNA
- Banques de données taxon-spécifiques:
  - Homme: OMIM: phénotypes, maladies génétiques, mutations
  - Génomes animaux: Ensembl
  - Levure (LISTA, SGD, YPD).
  - Nématode (ACeDB).
  - Drosophile (FlyBase).
  - …

Banques de données bibliographiques

- MEDLINE/PUBMED: médical, génétique, biologie moléculaire, biochimie, physiologie, …
  - http://www.ncbi.nlm.nih.gov/entrez/
- ISI Science Citation Index: général
- EMBASE: médical
- PASCAL: général

Interrogation des bases de données

- Sélection des entrées en fonction de :
  - Noms et numéros d'accession des séquences.
  - Références bibliographiques (auteurs, articles, …).
  - Mots-clés.
  - Taxonomie (espèce, genre, ordre, …).
  - Date de publication
  - Organelles, hôte
  - ...
- Accès aux aux régions fonctionnelles décrites dans les *features*.
  - régions codantes (CDS), tRNA, rRNA, ...

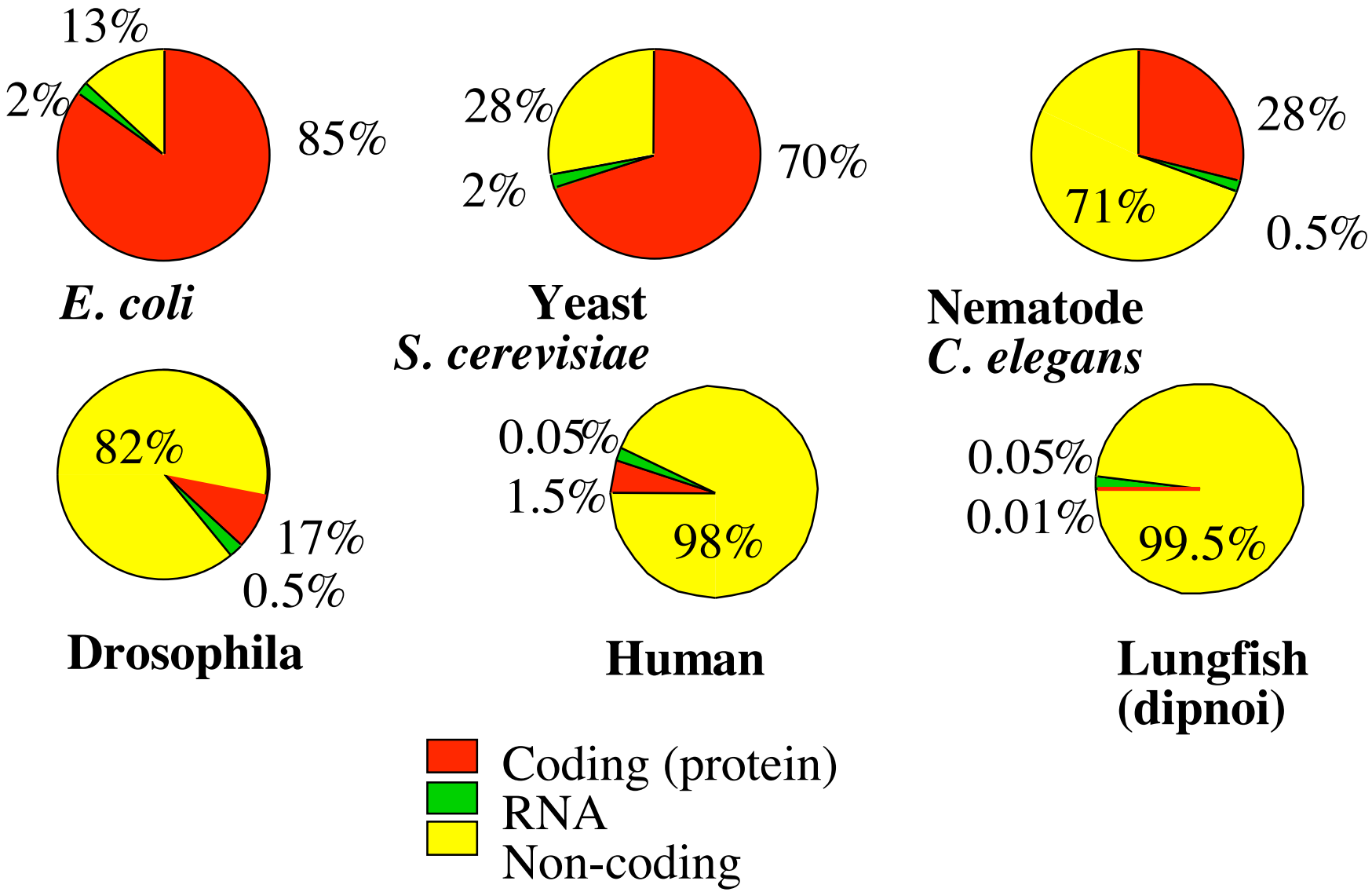
Logiciels d ’interrogation

- ACNUC/Query : http://pbil.univ-lyon1.fr/
  - Accès à GenBank, EMBL, SWISS-PROT, TrEMBL.
  - Requêtes complexes
  - Sélection et extraction de sous-séquences (e.g. CDS, tRNAs, rRNAs, …)
- SRS (sequence retrieval system) http://srs.ebi.ac.uk/
  - À ce jour, près de 90 banques consultables sous SRS.
  - Permet des interrogations multi-banques.
- Entrez http://ncbi.nlm.nih.gov/
  - Accès aux banques du NCBI : GenBank, GenPept, NRL_3D, MEDLINE.
  - Recherches par voisinage:
    - séquences : recherche de similarités.
    - références bibliographiques : mots-clés communs dans les titres ou les résumés.

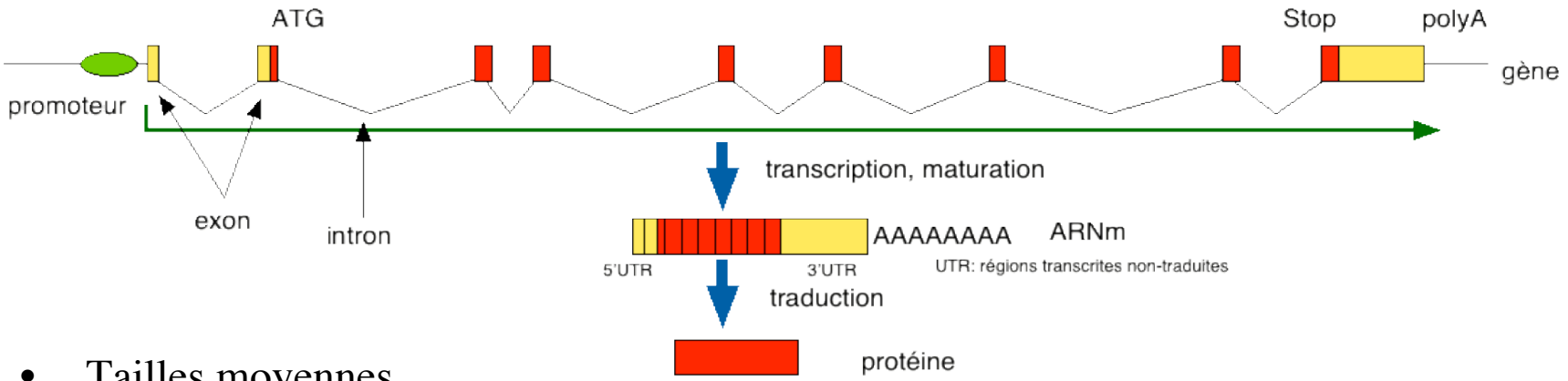
Genome annotation

- Identification of repeats (RepeatMasker, Reputer, …)
- Prediction of protein-coding genes
  - Intrinsic methods (GenScan, Genmark, Glimmer, ...)
  - Genomic/mRNA (EST) comparison (blastn, sim4, …)
  - Genomic/protein comparison (blastx, GeneWise, …)
- Prediction of RNA genes
  - Intrinsic methods (tRNA: tRNAScanSE, snoRNA …)
  - Genomic/RNA (EST) comparison (blastn, sim4, …)
- And more …
  - Replication origins (bacteria) (oriloc)
  - Pseudogenes (by similarity) (blastn, blastx)
  - Regulatory elements (CpG islands, promoters ??)

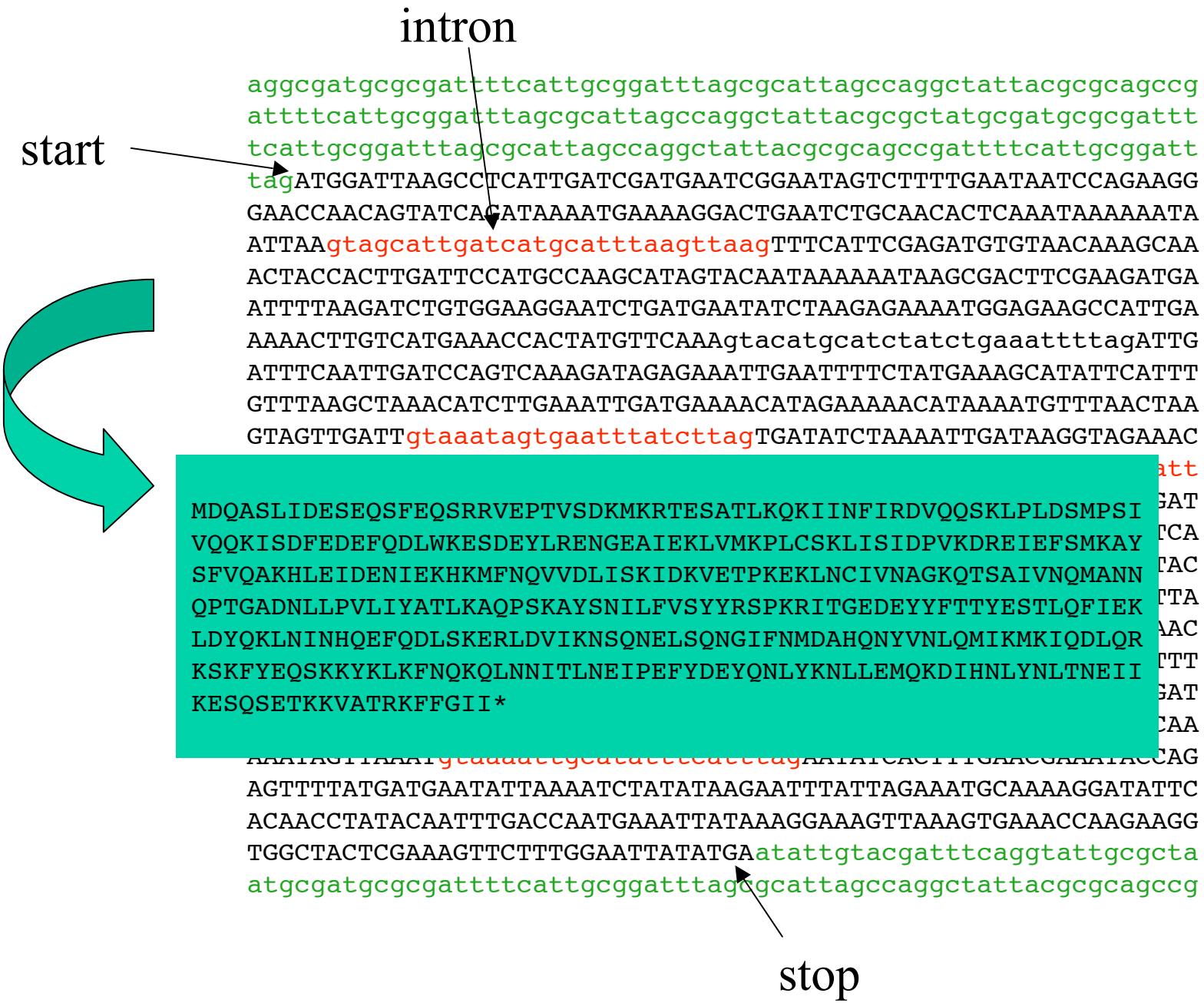
Proportion of functional elements within genomes



Structure des gènes humains



- Tailles moyennes
  - Gene 45 kb
  - CDS 1500 nt
  - Exon (interne) 145 nt
  - Intron 5200 nt
  - 5'UTR 210 nt
  - 3'UTR 740 nt
- Intron/exon
  - Nombres d'introns: 6 ±3 introns / kb CDS
  - Introns / (introns + CDS): 92%
- Epissage alternatif dans plus de 30% des gènes


intron
start
aggcgatgcgcgattttcattgcggatttagcgcattagccaggctattacgcgcagccg
attttcattgcggatttagcgcattagccaggctattacgcgctatgcgatgcgcgattt
tcattgcggatttagcgcattagccaggctattacgcgcagccgattttcattgcggatt
tagATGGATTAAGCCTCATTGATCGATGAATCGGAATAGTCTTTTGAATAATCCAGAAGG
GAACCAACAGTATCAGATAAAATGAAAAGGACTGAATCTGCAACACTCAAATAAAAAATA
ATTAAgtagcattgatcatgcatttaagttaagTTTCATTCGAGATGTGTAACAAAGCAA
ACTACCACTTGATTCCATGCCAAGCATAGTACAATAAAAAATAAGCGACTTCGAAGATGA
ATTTTAAGATCTGTGGAAGGAATCTGATGAATATCTAAGAGAAAATGGAGAAGCCATTGA
AAAACTTGTCATGAAACCACTATGTTCAAAgtacatgcatctatctgaaattttagATTG
ATTTCAATTGATCCAGTCAAAGATAGAGAAATTGAATTTTCTATGAAAGCATATTCATTT
GTTTAAGCTAAACATCTTGAAATTGATGAAAACATAGAAAAACATAAAATGTTTAACTAA
GTAGTTGATTgtaaatagtgaatttatcttagTGATATCTAAAATTGATAAGGTAGAAAC
MDQASLIDESEQSFEQSRRVEPTVSDKMKRTESATLKQKIINFIRDVQQSKLPLDSMPSI
VQQKISDFEDEFQDLWKESDEYLRENGEAIEKLVMKPLCSKLISIDPVKDREIEFSMKAY
SFVQAKHLEIDENIEKHKMFNQVVDLISKIDKVETPKEKLNCIVNAGKQTSAIVNQMANN
QPTGADNLLPVLIYATLKAQPSKAYSNILFVSYYRSPKRITGEDEYYFTTYESTLQFIEK
LDYQKLNINHQEFQDLSKERLDVIKNSQNELSQNGIFNMDAHQNYVNLQMIKMKIQDLQR
KSKFYEQSKKYKLKFNQKQLNNITLNEIPEFYDEYQNLYKNLLEMQKDIHNLYNLTNEII
KESQSETKKVATRKFFGII*
AGTTTTATGATGAATATTAAAATCTATATAAGAATTTATTAGAAATGCAAAAGGATATTC
ACAACCTATACAATTTGACCAATGAAATTATAAAGGAAAGTTAAAGTGAAACCAAGAAGG
TGGCTACTCGAAAGTTCTTTGGAATTATATGAatattgtacgatttcaggtattgcgcta
atgcgatgcgcgattttcattgcggatttagcgcattagccaggctattacgcgcagccg
stop

Prédiction de gènes: informations utilisées

- 1- caractérisation de la taille et du contenu des régions (codantes/non-codantes)
- 2- caractérisation des signaux au niveau de sites fonctionnels (e.g. signaux d'épissage, début et fin de traduction, ...)
- 3- utilisation de similarité ADN/protéines, ADN/ARNm, ADN/ADN

- méthodes intrinsèques (ab initio): utilisent 1 et 2
- méthodes extrinsèques (approche comparative): utilisent 3, et éventuellement 2

Prédiction de gènes : méthodes intrinsèques

- Prédiction des régions codantes uniquement !
- Recherche de phases ouvertes de lecture (ORF: open reading frame) = série de codon sans STOP

```
Phase +0
Phase +1
Phase +2
                ATGTACCGTCGATCGTAGCTTGATCGATCG
                TACATGGCAGCTAGCATCGAACTAGCTAGC
Phase -0
Phase -1
Phase -2
```

  - Taille moyenne des ORF: ± 150 nt

- Distinction codant/non-codant : contenu et taille des séquences
  - usage des codons: utilisation non aléatoire des codons synonymes
  - fréquence des amino-acides (e.g. tryptophane est rare)
  - corrélations entre amino-acides (codons) successifs
  - taille des exons et introns

  - Apprentissage sur un ensemble de gènes connus
  - Fréquence d'oligomères (e.g. hexamères)
  - chaines de Markov

Prédiction de gènes : méthodes intrinsèques (suite)

- Recherche de signaux: sites fonctionnels conservés
  - signaux d'épissage: site donneur, accepteur d'épissage, point de branchement
  - codon d'initiation de la traduction
  - codon stop

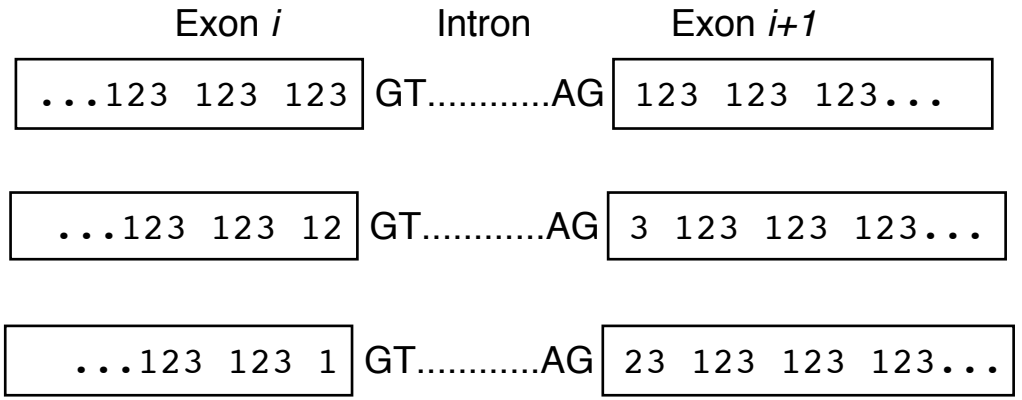
  - Utilisation de consensus (historique): e.g.

donneur	accepteur
A/CAG GT RAGT	YYYYYYYYY*C AG G

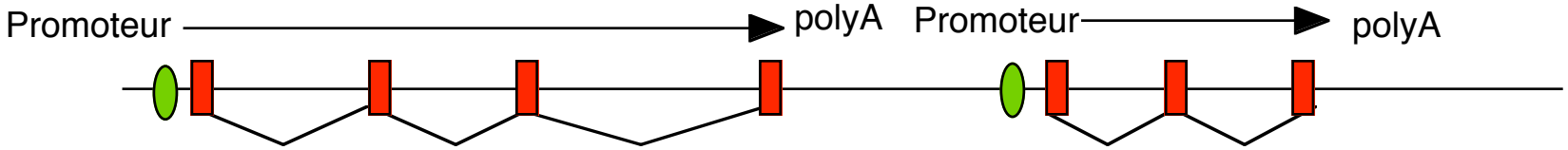
  - Utilisation de matrices de pondération position-dépendantes (profils)

Prédiction de gènes : méthodes intrinsèques (suite)

- Construction d ’un modèle de gène protéique
  - Combinaison d ’exons de phases compatibles (pondération en fonction des scores de chaque exon potentiel) - pas de codons stop en phase!



  - Recherche de limites de gènes
    - Exons terminaux (5 ’, 3 ’)
    - Promoteur
    - Signal de polyadénylation



Qualité de la prédiction par exon

- Évaluation de la fiabilité de la prédiction
  - essai des logiciels de prédiction sur un ensemble de séquences caractérisées expérimentalement (différentes de celles utilisées pour entrainer les logiciels)

- Sensibilité : fraction des exons présents dans la séquence qui sont retrouvés par le logiciel

$$sensibilité = \frac{VraisPositifs}{VraisPositifs + Faux\,Négatifs}$$

- Spécificité : fraction des vrais exons parmi tous ceux prédits

$$spécificité = \frac{Vrais\,Positifs}{VraisPositifs + Faux\,Positifs}$$

Prédiction de gènes eucaryotes: qualité de la prédiction

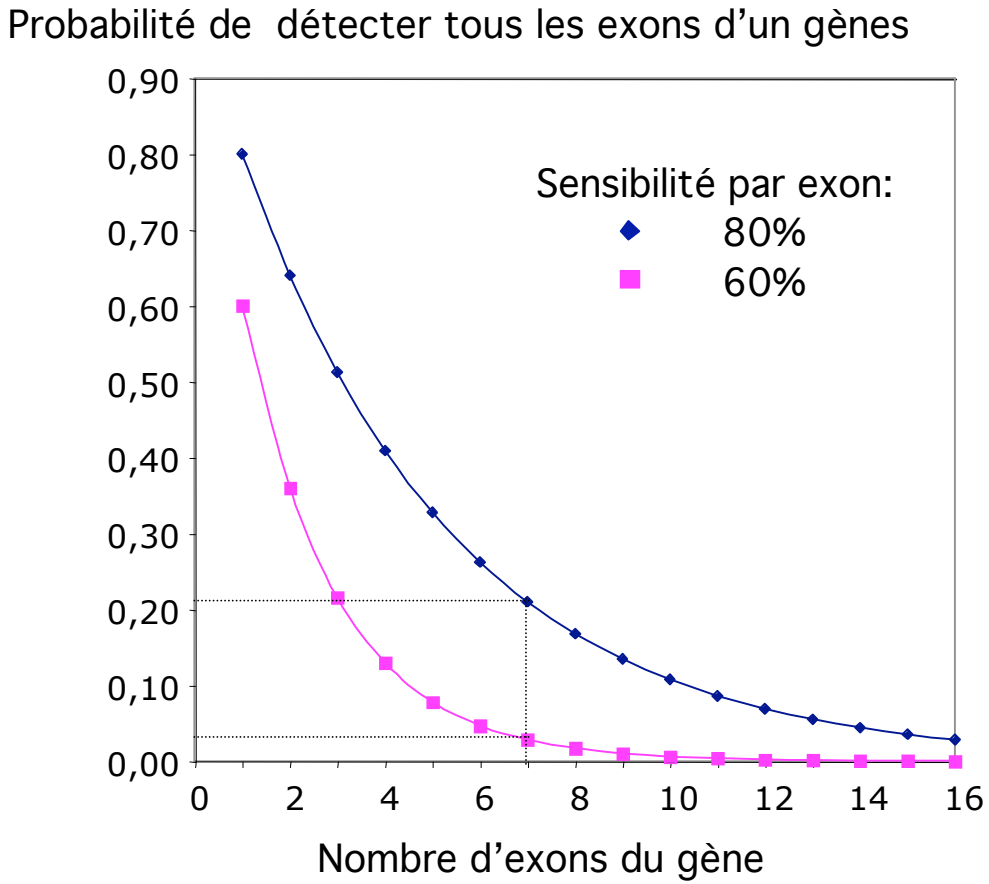
- Comparaison des différents logiciels: sensibilité/spécificité
  - Sn: sensibilité Sp: spécificité par exon (sn_e, sp_e) ou par nucéotide (sn_e, sp_e)

  - Locus BRCA2 (1.4 Mb, chrom. 13q) (Sanger Centre 1999): région "difficile" pour les logiciels de prédiction. 159 exons

	Sn_e	Sp_e	Sn_n	Sp_n
GenScan	0.66	0.36	0.81	0.44
FGENES 1.6	0.69	0.57	0.79	0.66
FGENES 1.6 masked	0.69	0.65	0.79	0.74
GenScan+FGENES	0.61	0.82	0.67	0.90

Prédiction de gènes protéiques complets

- Prédiction de gènes complets: sensibilité ?



  - + les faux positifs ! + épissage alternatif ! + exons non-codants !

Un peu d ’optimisme

- Fraction de la longueur des gènes correctement prédits:

70-80%

- Probabilité que deux exons potentiels consécutifs soient réels (et donc positifs en RT-PCR)

0.5

Prédiction de gènes : méthodes intrinsèques (bilan)

- Procaryotes (pas d'intron):
  - sensibilité et spécificité > 95% (dépend du taux de G+C du génome)
- Eucaryotes: efficacité variable (dépend du taux de G+C du génome et du nombre et de la taille des introns)
  - prédiction d'exons: sensibilité et spécificité 60-80%
  - prédiction de gènes complets:
    - levure: >90% des gènes correctement prédits
    - nématode: 50% des gènes correctement prédits
    - homme: 20% (?) des gènes correctement prédits
- très utile pour guider les expérimentations

Prediction of protein genes : mRNA / DNA comparison

- Large scale transcriptome data
  - ESTs
  - full-length cDNA sequencing projects

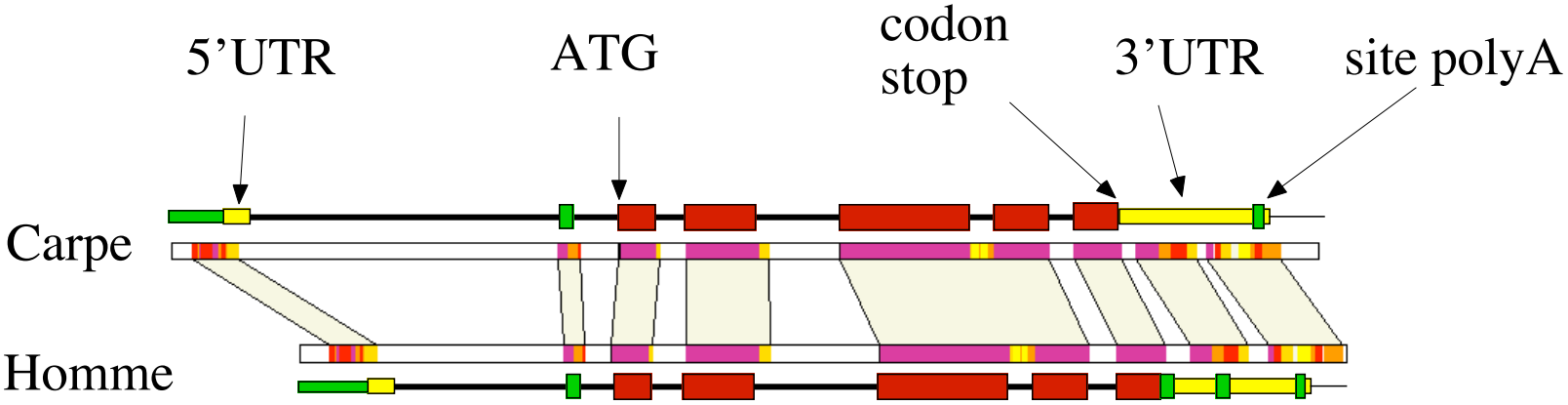
- Alignment genomic DNA / mRNA : identification of exons (blastn, sim4, est2genome)
  - information on alternative splicing, gene expression pattern
  - not restricted to protein-coding regions (UTRs, non-coding RNAs)
- Problems:
  - weakly expressed genes; genes with a restricted tissue-distribution
  - artefacts in EST sequences (contamination with nuclear RNA, DNA)

Prediction of protein genes: comparative approach

- Comparison of a genomic sequence with genes that have been already characterized (e.g. in other species)
  - DNA/protein alignments: blastx, genewise

- Comparaison of homologous genomic sequences
  - DNA/DNA alignments

Analyse comparative des gènes de β-actine de l'homme et de la carpe



introns:

régions codantes:

éléments régulateurs:

échelle de similarité:

pas de similarité significative

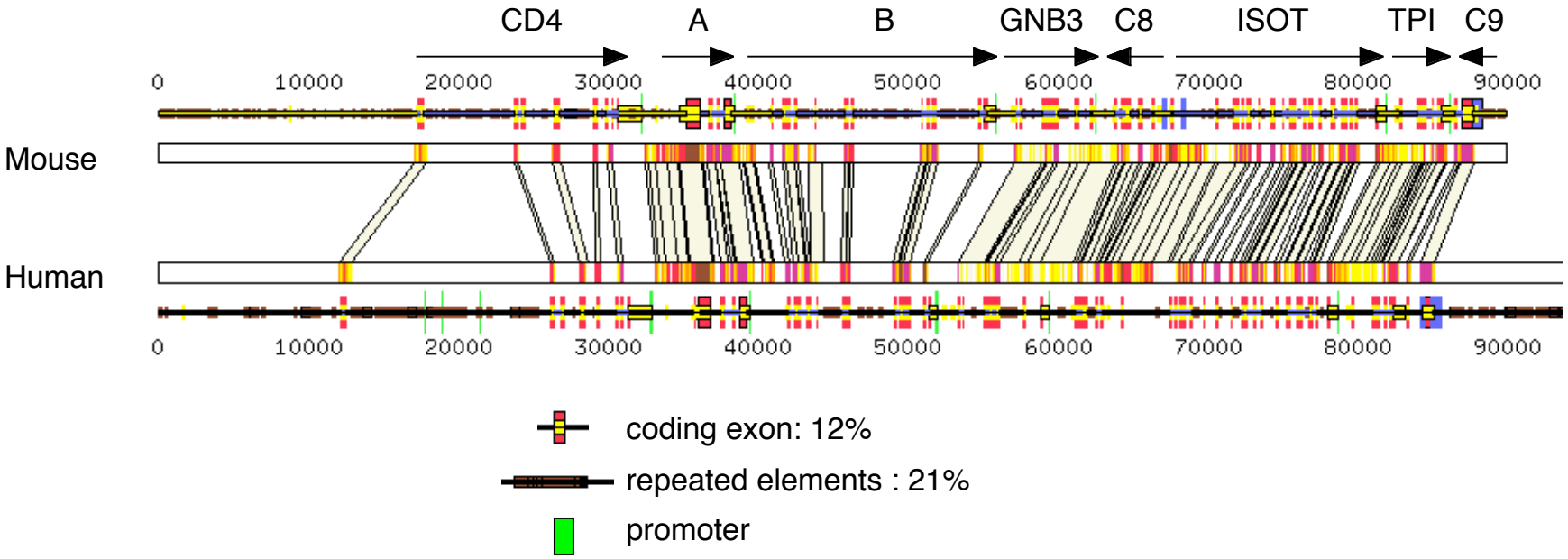
80 - 90% identité

70 - 80% identité

Comparison of human and mouse CD4-C9 locus: gene-rich, repeated-element poor, G+C-rich region (50.5%)

Human chromosome 12p13
Mouse chromosome 6

8 genes: CD4, A, B, GNB3, C8, ISOT, TPI, C9

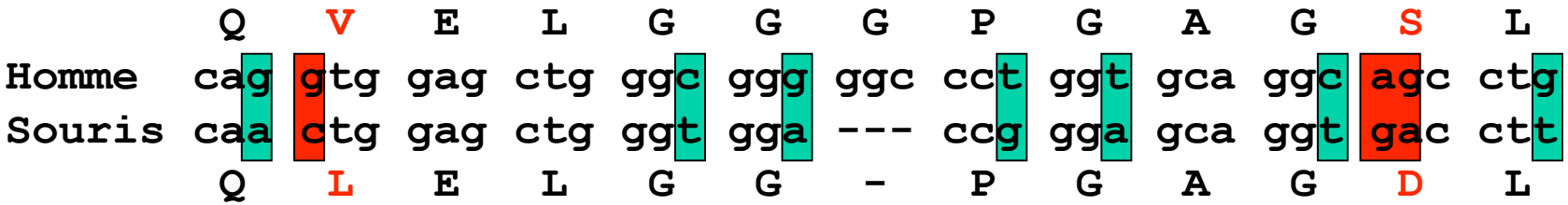


Prediction of protein genes: comparative approach

- Problems: sensitivity depends on the evolutionary distance
  - rapidly evolving genes
  - lineage-specific genes (orphans)

- How to distinguish protein-coding regions from conserved non-coding sequences ?

Distinction des régions conservées codantes *vs.* non-codantes



- Substitutions synonymes (Ks)
- Substitutions non-synonymes (Ka)
- \- Insertion ou délétion

Ratio Ka/Ks << 1 => région codante

Transposable elements: noise for gene prediction

- Transposable elements: ubiquitous in eukaryotic genomes (50% of mammalian genomes)
- TEs contain coding-regions (transposases, reverse-transcriptase) => recognized as “genes” by gene prediction software
- Domesticated (recruited) TEs are very rare
- Mask TEs before running gene prediction software (RepeatMasker)

Prédiction de gènes : démarche

- 1- recherche de séquences répétées (RepeatMasker)
- 2- méthodes intrinséques (consensus de différentes méthodes)
- 3- recherche de similarité ADN/protéines (blastx/genewise)
- 4- recherche de similarité ADN/mRNA (blastn/sim4)
- 5- recherche de similarité ADN/ADN (blastn)
- COMBINER LES RESULTATS

- 6- prédiction de gènes RNA
  - tRNA: tRNAScanSE
  - rRNA: par similarité
  - snRNA ...

Annotation systématique du génome humain

- ENSEMBL project
  - http://www.ensembl.org/
- Human Genome Project Working Draft at UCSC
  - http://genome.ucsc.edu/

Prédiction de régions régulatrices

- Méthodes intrinsèques (*ab initio*)
  - Prédiction de promoteurs
  - Îlots CpG
- Approche comparative

Prédiction de promoteurs eucaryotes

- Combinaison de sites de fixation de facteur de transcription (ordre, orientation, distance)
- Motifs courts, dégénérés
  - Difficile de distinguer les vrais sites des faux positifs:
  - Motif à 4 bases: ≈1/256 pb (1/128 pb sur les deux brins)
- Boîtes TATA, CAAT , GC: absents dans beaucoup de promoteurs

- Banques de données de sites de fixation de facteurs de transcription (TRANSFAC), de promoteurs caractérisés expérimentalement (EPD)

- PromoterScan (Prestridge 1995): Mesure de la densité en sites potentiels de fixation de facteurs de transcription de long de la séquence (pondération en fonction de la fréquence des sites dans ou en dehors des vrais promoteurs)

Prédiction de promoteurs: sensibilité, spécificité

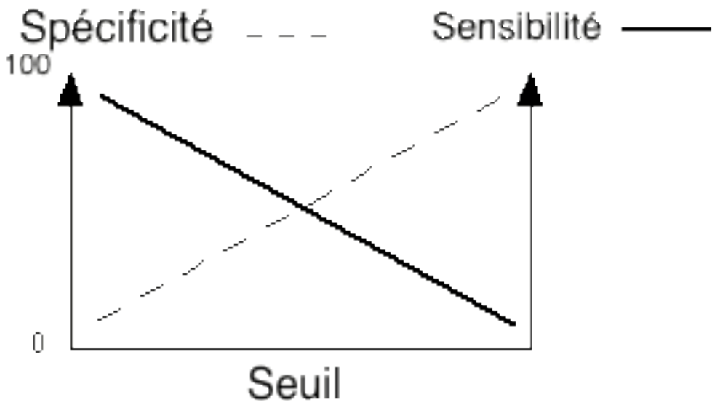
- Sensibilité: fraction des promoteurs qui sont trouvés par le logiciel

$$sensibilité = \frac{vrais\_positifs}{vrais\_positifs + faux\_négatifs}$$

  - PromoterScan: sensibilité = 70% (promoteurs à boîte TATA)
- Spécificité: fraction des vrais promoteurs parmi ceux qui ont été prédits

$$spécificité = \frac{vrais\_positifs}{vrais\_positifs + faux\_positifs}$$

  - PromoterScan: spécificité = 20%
  - Un faux positif / 10 kb



- Génome humain:
- ≈25 000 gènes, ≈1 promoteur/100 kb

Îlots CpG

- Génome de vertébrés :
  - méthylation des C dans les dinucléotides 5 ’-CG-3 ’(CpG)
- Me-C fortement mutable -> T

```
5 ’-CG- 3 ’      =>    5 ’-TG-3 ’     ou     5 ’-CA-3 ’
3 ’-GC- 5 ’            3 ’-AC-5 ’            3 ’-GT-5 ’
```

- Génome des vertébrés: globalement dépourvu en CpG (excès de TG, CA)

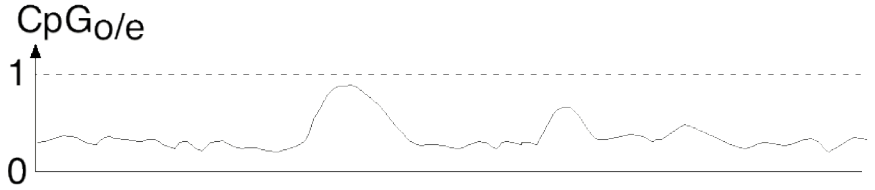
$$CpG_{o/e} = \frac{Nombre\_de\_CpG\_observé}{Nombre\_de\_CpG\_attendu} = 0.25$$

- Certaines régions (200 nt à plusieurs kb) échappent à la méthylation
  - Pas de déplétion en CpG: $CpG_{o/e}$ proche de 1
  - Riche en G+C
  - Îlot CpG:

    Longueur > 500 nt

    CpGo/e > 0.6

    G+C > 50%



Îlots CpG

- Bird (1986), Gardiner-Garden (1987) Larsen (1992) ref
  - 40% des gènes tissu-spécifiques possèdent un îlot CpG en 5 ’
  - 100% des gènes ‘ housekeeping ’ possèdent un îlot CpG en 5 ’

- Rechercher des îlots CpG pour prédire des régions promotrices ?
  - Sensibilité: 40-100%
  - Spécificité ?? (Quelle fraction des îlots CpG correspond effectivement à des régions promotrices ?)

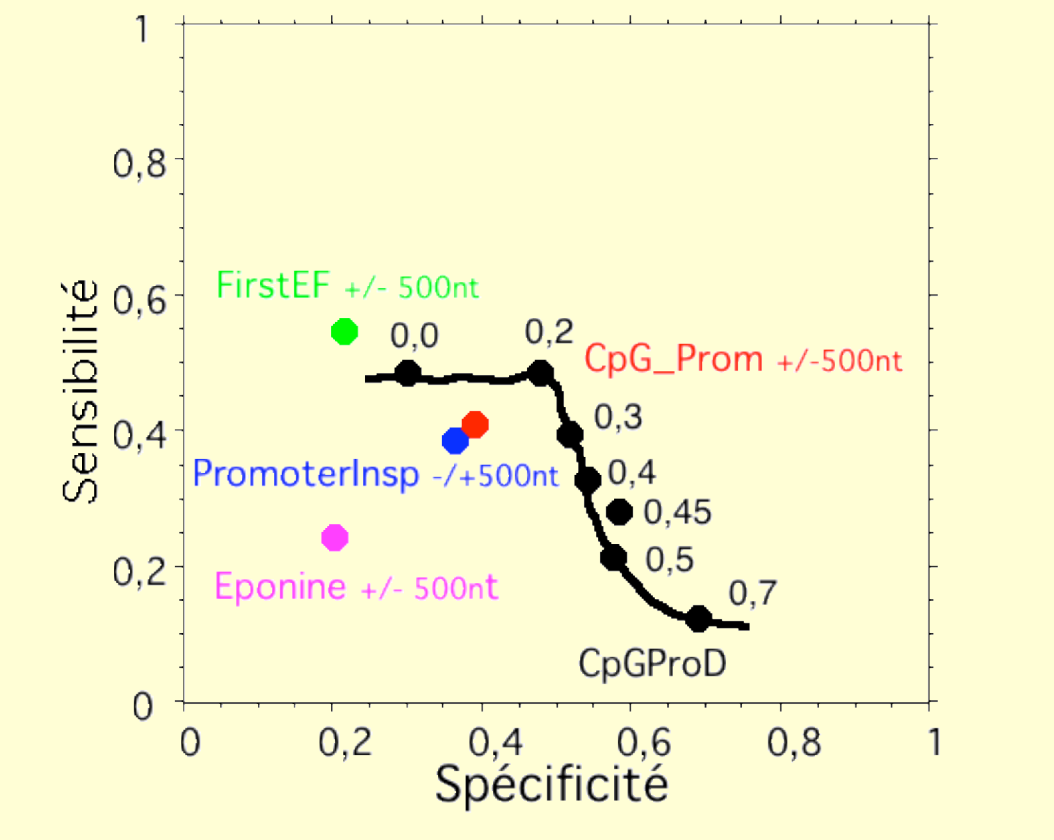
- Ponger (2001): comparaison des îlot CpG qui recouvre ou non le site d ’initiation de la transcription

Prédiction de promoteurs

La prédiction des régions promotrices:

sn=nb prom de gènes connus préd / nb prom de gènes connus

sp=nb prom préd / nb de prédictions

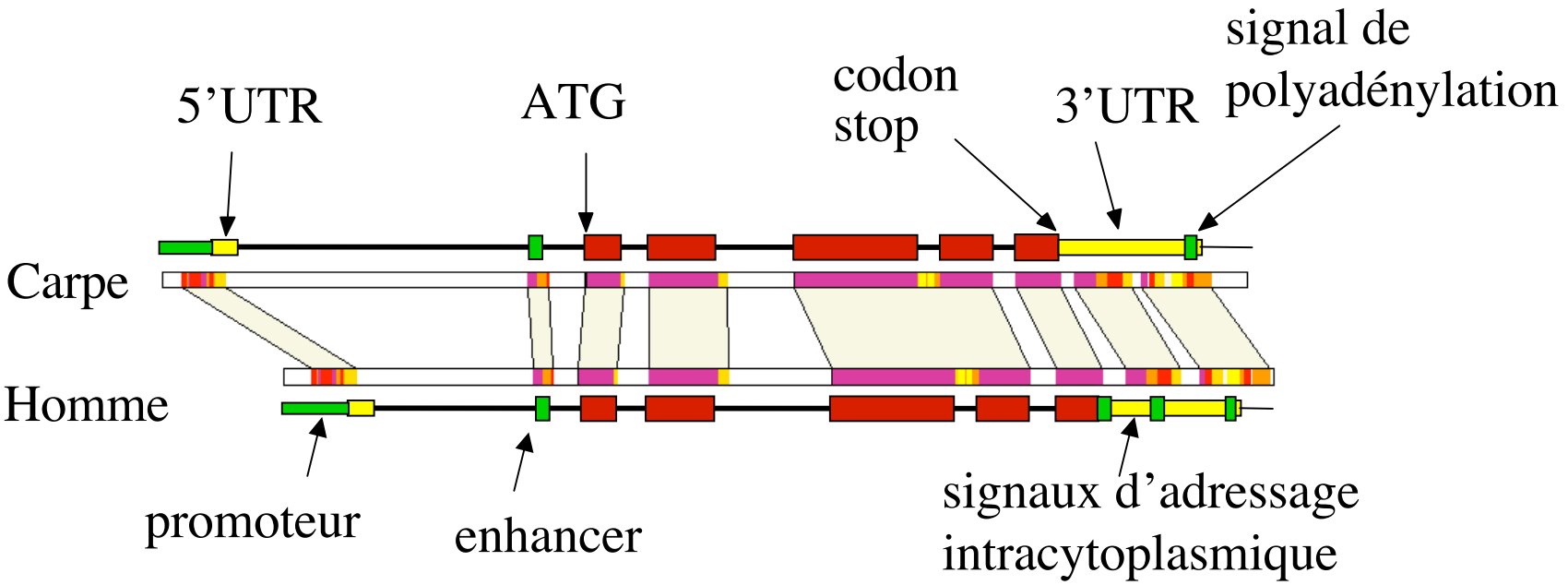


Recherche de régions régulatrices par analyse comparative (empreintes phylogénétiques)

- Goodman et al. 1988: régulation de l'expression des gènes du cluster β-globine au cours du développement
  - Alignement de séquences orthologues de 6 mammifères (> 270 Ma d'évolution)
  - 13 empreintes phylogénétiques: ≥ 6 nt, conservation 100%
  - Analyse par retard de bande sur gel:
  - 12/13 (92%) correspondent à des sites de fixation de protéines

- 1996: 35 empreintes phylogénétiques avec protéines fixatrices identifiées

Analyse comparative des gènes de β-actine de l'homme et de la carpe



introns:

régions codantes:

éléments régulateurs:

échelle de similarité

pas de similarité significative

80 - 90% identité

70 - 80% identité

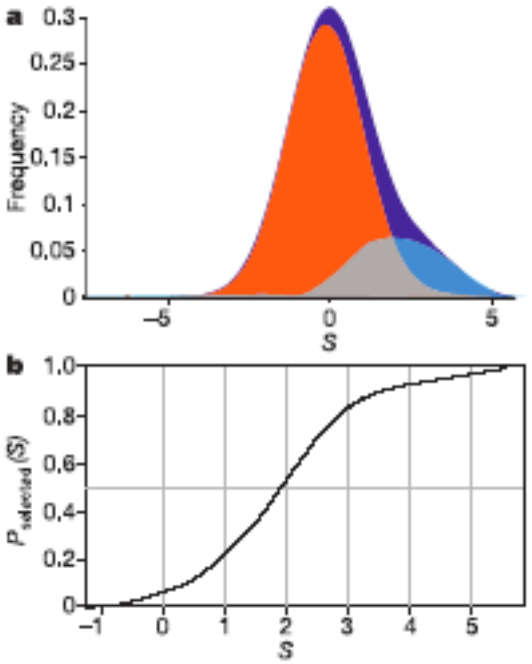
Recherche d'empreintes phylogénétiques à l'échelle du génome

- Empreintes phylogénétique = séquences qui évoluent plus lentement que des séquences non soumises à pression de sélection
- En absence de sélection, le taux d'évolution est égal au taux de mutation (= évolution neutre)
- Mammifères: taux de mutation environ 3.10^{-9} substitution/site/an
- Variation des taux de mutation le long du génomes (et entre espèces)
- Utilisation de marqueurs neutres pour mesurer le taux de mutation local : pseudogènes, éléments transposables défectifs

Comparaison des génomes de l'homme et de la souris

- Alignement des génomes de l'homme et de la souris: 40% du génome humain est alignable avec le génome de la souris
- Utilisation de marqueurs neutres pour mesurer le taux de mutation local : éléments transposables défectifs orthologues homme/souris (i.e. insérés dans le génome avant la divergence primates/rongeurs)
- Comparaison des taux de substitution dans les séquences non-répétées et dans les marqueurs neutres

Comparaison des génomes de l'homme et de la souris



Distribution des taux de substitution

- Marqueurs neutres
- Séquences non-répétées

Probabilité d'être sous pression de sélection négative

- Plus de 5% du génome des mammifères est sous pression de sélection négative
- NB: seulement 1.2% du génome est codant !! 4 fois plus des regions non-codantes fonctionnelles que de régions codantes !!
- Problème: faible discrimination des séquences sous pression de sélection !!

Comparaison à plus grande distance évolutive

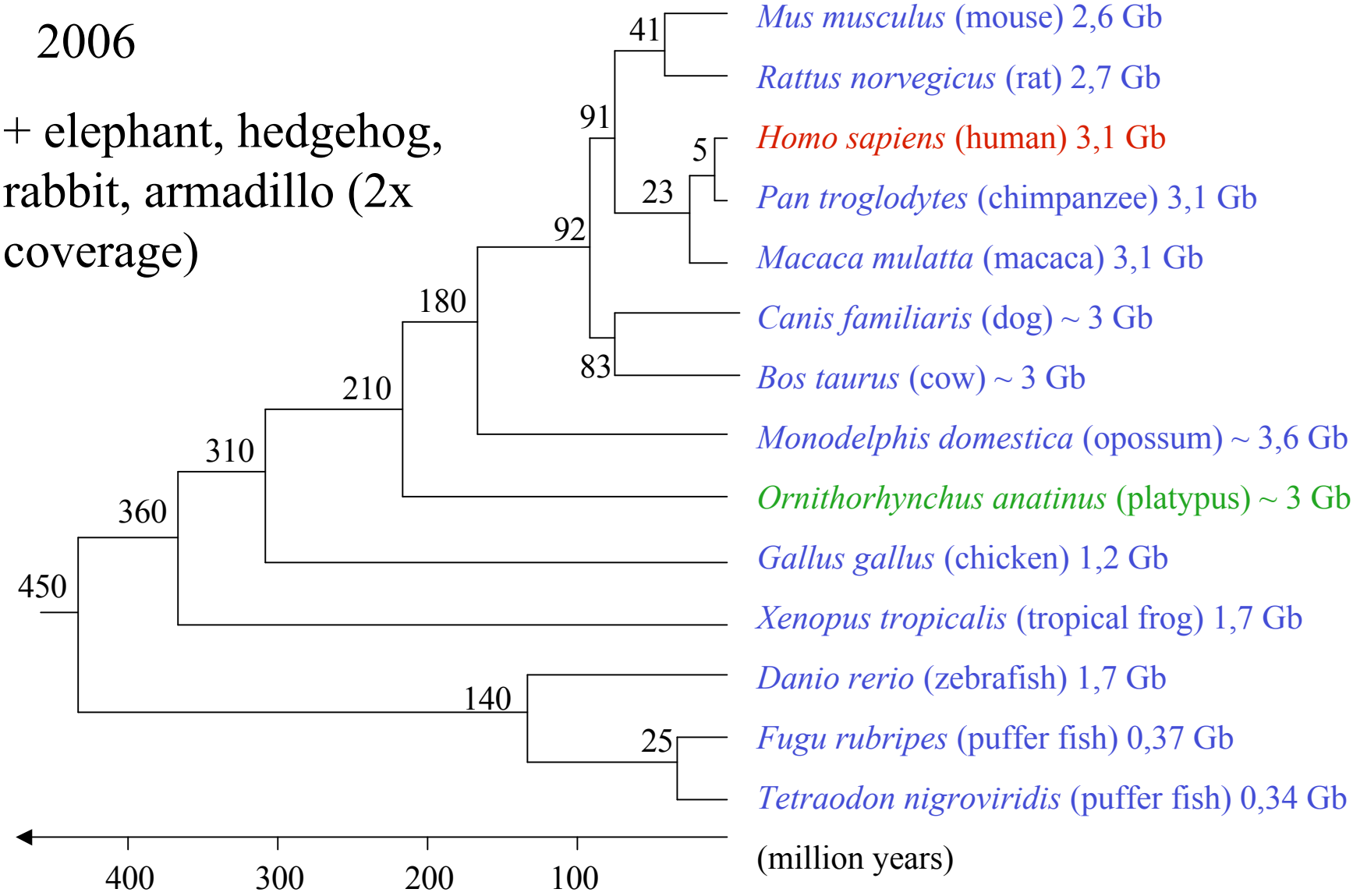
- Mammifères/Oiseaux
  - Empreinte phylogénétique couvrent 2,5% du génome humain
  - 56% des ces empreintes sont non-codantes
- Mammifères/Poissons:
  - Empreinte phylogénétique couvrent 1,8% du génome humain
  - 50% des ces empreintes sont non-codantes
  - Empreintes phylogénétiques non-codantes souvent à proximité des gènes impliqués dans le développement
  - 36 testées expérimentalement: 27 (75%) ont une activité enhancer (Nobrega 2003)

Empreintes phylogénétiques: compromis sensibilité/spécificité

- Grande distance évolutive: ne détecte pas les éléments fonctionnels spécifiques d'une lignée
  - Mammifères/poissons: 1.8% génome humain
  - Mammifères/oiseaux: 2.5% génome humain
  - Primates/rongeurs: 5% génome humain
  - Homme/grands singes: 10% génome humain ??
- Faible distance évolutive: ne discrimine pas les régions conservées contraintes vs. neutres
- Solution: augmenter le nombre d'espèces !

2006

+ elephant, hedgehog, rabbit, armadillo (2x coverage)



Complete sequence, finished assembly

Nearly complete sequence; preliminary assembly (draft)

Sequencing in progress

Prediction of gene function

- Analysis of expression pattern (ESTs, SAGE,…)
- Prediction of the subcellular location of the protein : nucleus, membrane, excreted, etc.
  - SignalPep : http://www.cbs.dtu.dk/services/SignalP/
  - Psort: http://psort.nibb.ac.jp/
  - etc. (see http://www.expasy.org/tools/)
- Search for functional motifs (e.g. DNA binding domains, catalytic sites, …)

  http://hits.isb-sib.ch/cgi-bin/PFSCAN
- Prediction by homology

Function prediction by homology ?

- Similarity between proteins => homology
- Homology => conserved structure
- Conserved structure => conserved function
- Yes, but …
  - Function: fuzzy concept
    - Identical biochemical activity ?
    - Identical expression pattern (tissu-specific isoforms) ?
    - Identical subcellular location (cytoplasm, mitochondria, etc.) ?
  - Homologous proteins with different function
    - e.g. homologous proteins binding a same receptor but opposite activity (activator/repressor)
    - homologous proteins with totally different functions: τ -cristalline / α-énolase
  - Orthology/paralogy
  - Modular evolution

Function prediction by homology ?

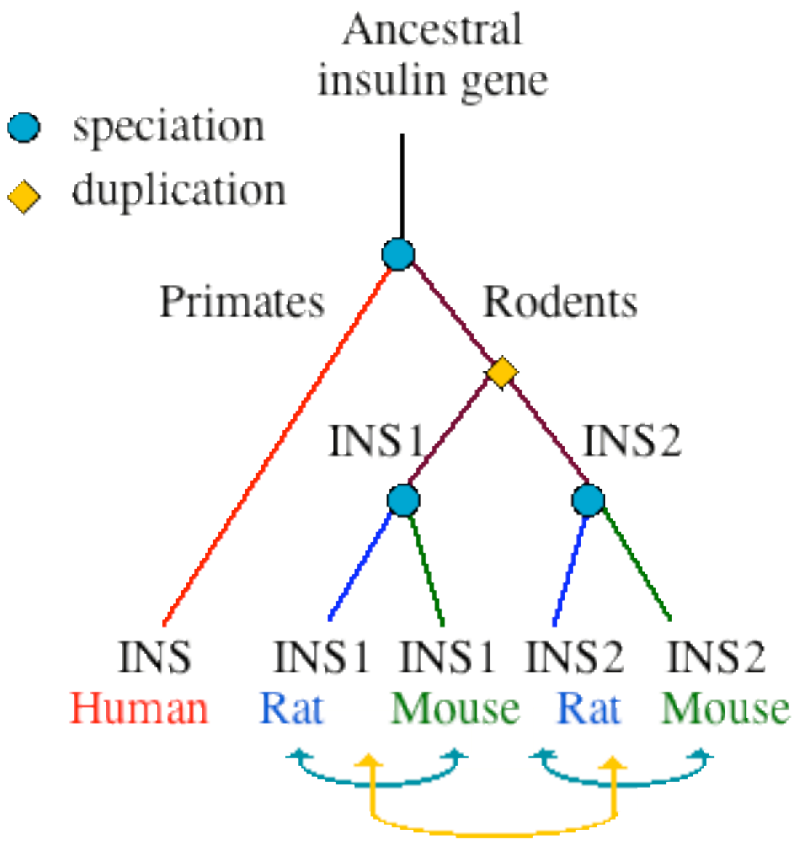
```
MZEORFG:  1   NSPDRACNLAKQAFDEAISELDSLGEESYKDSTLIMQLLDNLTLWTSDNE 59
              N+P++AC LAKQAFD+AI+ELD+L E+SYKDSTLIMQLLDNLTLWTSD E
BOV1433P: 186 NAPEQACLLAKQAFDDAIAELDTLNEDSYKDSTLIMQLLDNLTLWTSEGE 244

 Score = 87.4 bits (213), Expect = 1e-17
 Identities = 41/59 (69%), Positives = 50/59 (84%)

LOCUS       BOV1433P     1696 bp    mRNA            MAM   26-APR-1993
DEFINITION  Bovine brain-specific 14-3-3 protein eta chain mRNA
ACCESSION   J03868

LOCUS       MZEORFG       187 bp    mRNA    linear   PLN 01-FEB-2001
DEFINITION  Zea mays putative brain specific 14-3-3 protein,
            tau protein homolog  mRNA, partial cds.
ACCESSION   M95066
```

Orthology/paralogy



Homology: two genes are homologous if they share a common ancestor

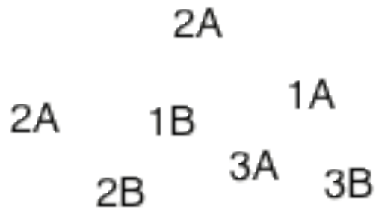
↔ Orthologues: homologous genes that have diverged after a speciation

↔ Paralogues: homologous genes that have diverged after a duplication

⚠ Orthology ≠ functional equivalence

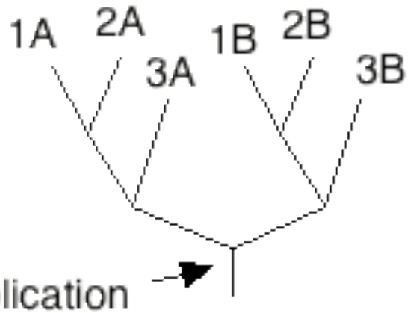
Phylogenetic approach for function prediction

1) Identify homologues

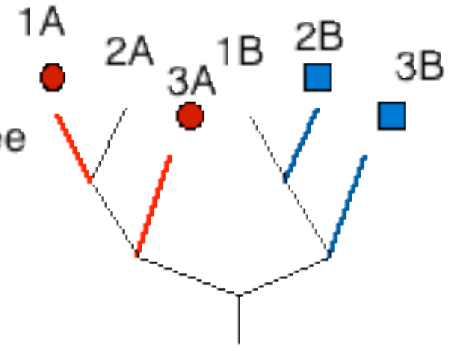


2) Align sequences

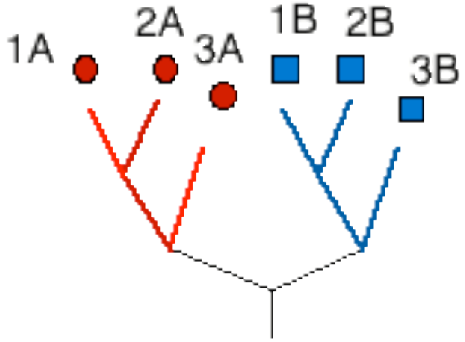
3) Compute the phylogenetic tree



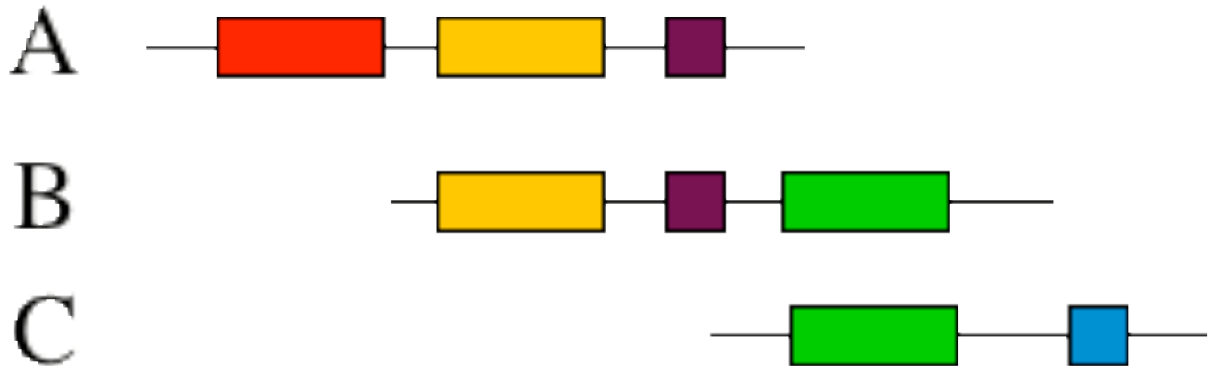
4) Place known functions on the tree



5) Infer the likely function of other genes



Modular evolution



VIEWPOINT

The ENCODE (ENCyclopedia Of DNA Elements) Project

The ENCODE Project Consortium*†

The ENCyclopedia Of DNA Elements (ENCODE) Project aims to identify all functional elements in the human genome sequence. The pilot phase of the Project is focused on a specified 30 megabases (~1%) of the human genome sequence and is organized as an international consortium of computational and laboratory-based scientists working to develop and apply high-throughput approaches for detecting all sequence elements that confer biological function. The results of this pilot phase will guide future efforts to analyze the entire human genome.

approaches, such as cDNA-cloning (*4*, *5*) and chip-based transcriptome a (*6*, *7*), have revealed the existence o transcribed sequences of unknown fu As a reflection of this complexity, ab of the human genome is evoluti conserved with respect to rodent g sequences, and therefore is inferred

Science. 2004 Oct 22;306(5696):636–40.

Encode Project: make the inventory of all functional elements in the human genome

Genes (transcription units, coding or not), promoters, enhancers, silencers, replication origin/termination, TFBS, methylation, chromatin modification, conserved regions of unknown function

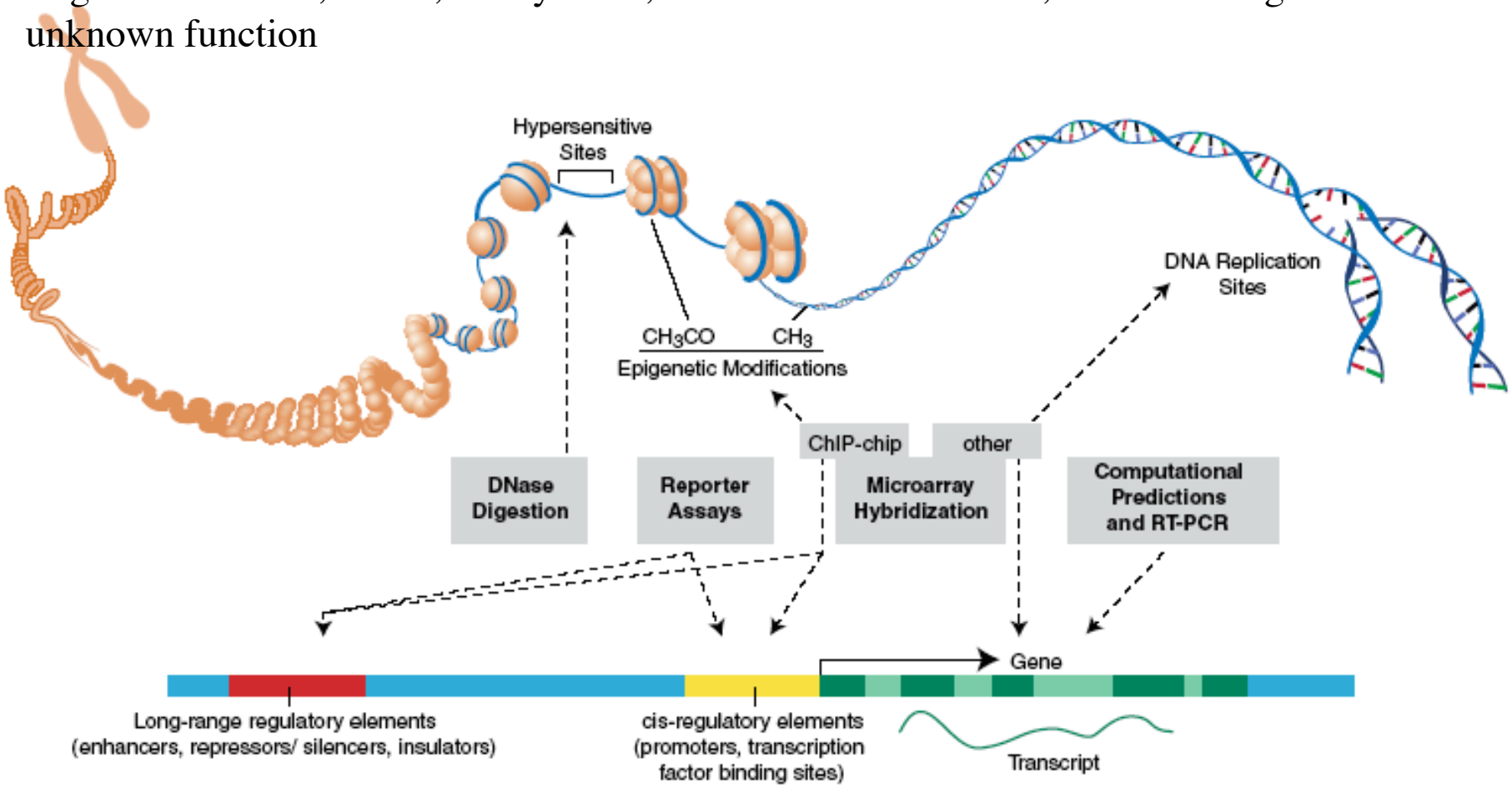
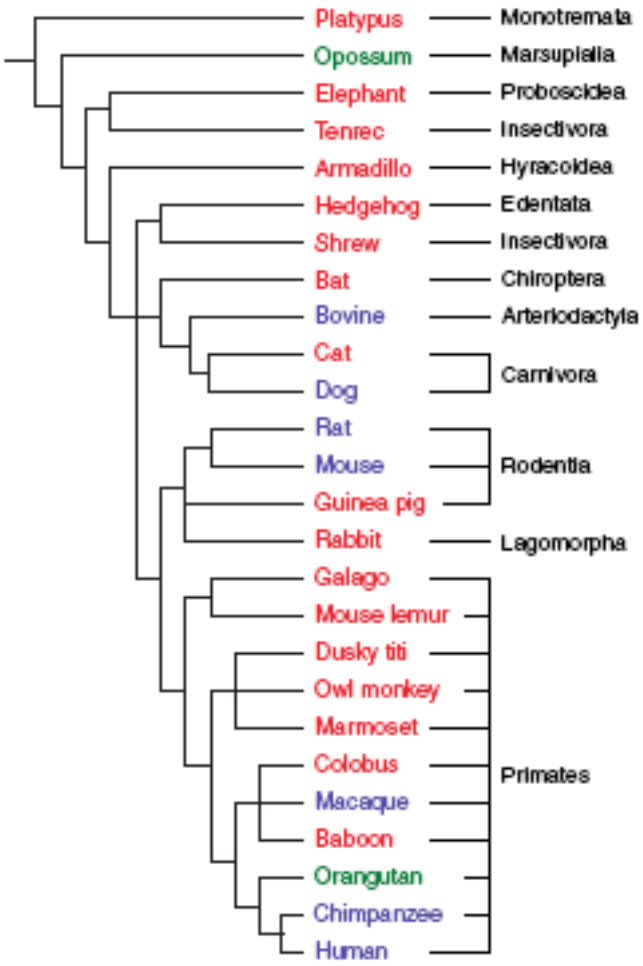


Fig. 1. Functional genomic elements being identified by the ENCODE pilot phase. The indicated methods are being used to identify different types of functional elements in the human genome.

Encode Project

- Combine all possible methods to identify functional elements

- Large scale experiments:
  - transcriptome analysis
  - Chip on chip experiments
  - ...
- Comparative genomics



Inter-species Comparaisons

Polymorphism Analysis (48 individuals)

Open Consortium

- Phase 1 pilot: 44 regions (30 Mb = 1% of the human genome)
  - Target regions (e.g. globin, CFTR)
  - Randomly sampled regions
  - Test various approaches
- Phase 2: technology development (large scale!)
- Phase 3: production