Genome Projects

- Identify genes and other functional elements (regulatory elements, etc.). Where are they?
- Predict the function of these genes. What do they do?

Identification and characterization of functional elements (genes, etc.)

- Experimental approach
  - Long and expensive

- Bioinformatics: provide predictions to guide the experiments
  - Rapid and cheap
  - Reliable ?

➧ **critical interpretation of the predictions of bioinformatic tools**

Genome Projects

- Identify genes and other functional elements (regulatory elements, etc.). Where are they?
- => gene prediction

- Predict the function of these genes. What do they do?
- => sequence similarity search

Plan du cours

- Introduction
- Projets Génome
- Banques de données (pour la biologie moléculaire)
- Algorithmes
  - Prédiction de gènes
  - Alignement de séquences
  - Recherche de similarité dans les banques de séquences

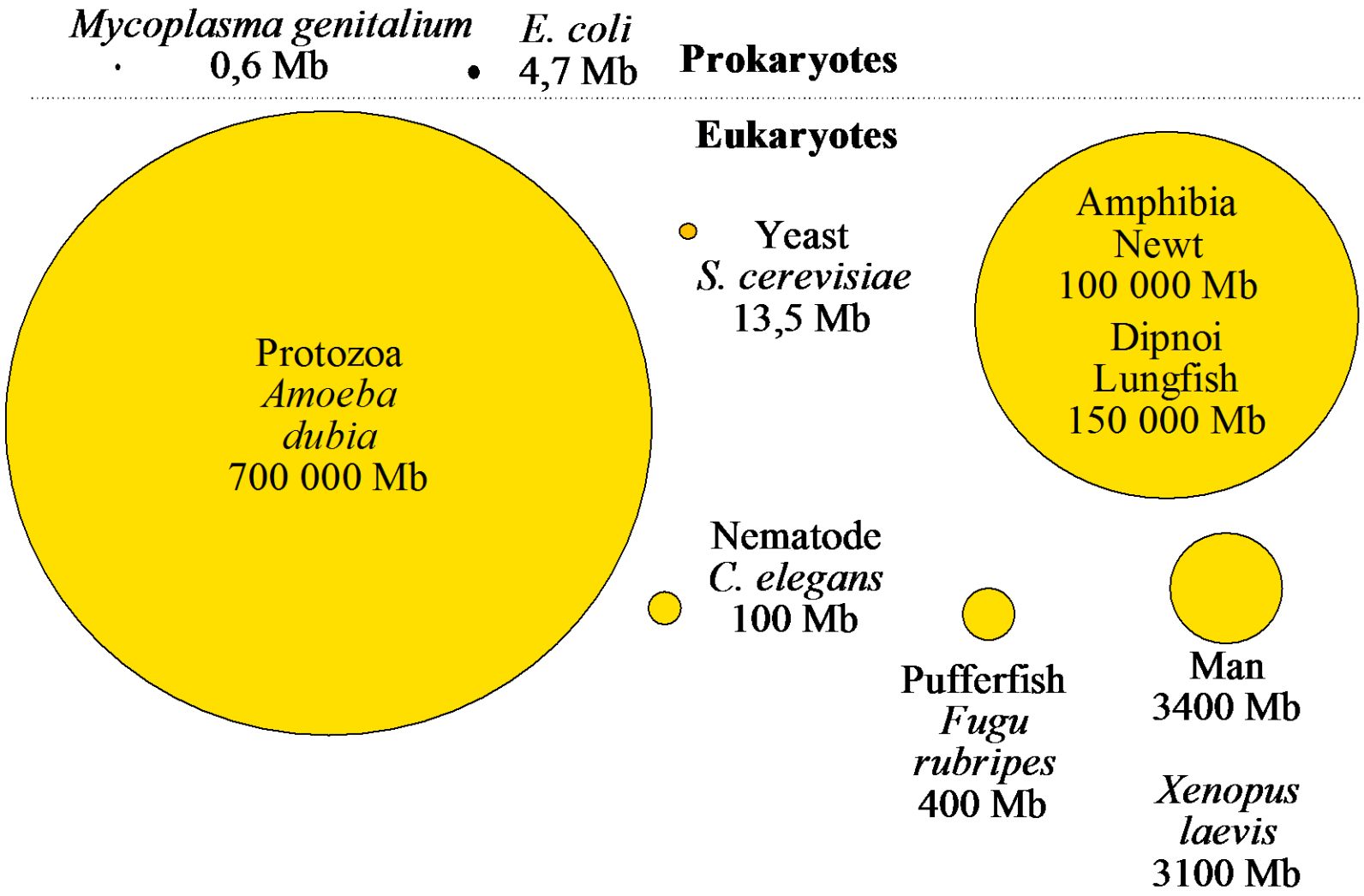
What is a genome ?

- 1911 - gene:
  - Elementary unit, responsible for the transmission of hereditary characters
- 1920 - genome:
  - Set of genes of an organism
- 1944 - Avery et al.
  - DNA is the molecule of heredity
- 1950-70 :
  - Double helix, Genetic code
  - Genome = set of DNA molecules present in a cell and transmitted to the offspring

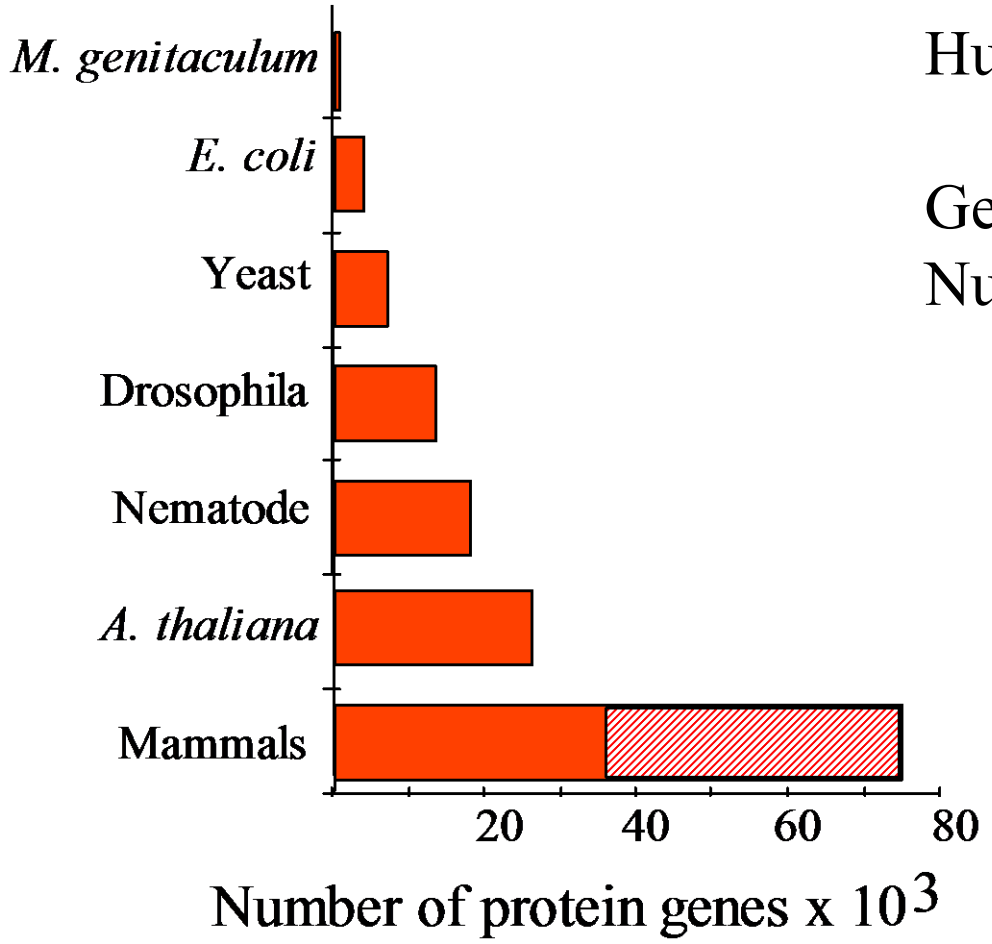
A genome is more than a set of genes

- 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

Genome size


Mycoplasma genitalium
0,6 Mb
E. coli
4,7 Mb
Prokaryotes
Eukaryotes
Protozoa
Amoeba
dubia
700 000 Mb
Yeast
S. cerevisiae
13,5 Mb
Amphibia
Newt
100 000 Mb
Dipnoi
Lungfish
150 000 Mb
Nematode
C. elegans
100 Mb
Pufferfish
Fugu
rubripes
400 Mb
Man
3400 Mb
Xenopus
laevis
3100 Mb

Number of protein genes



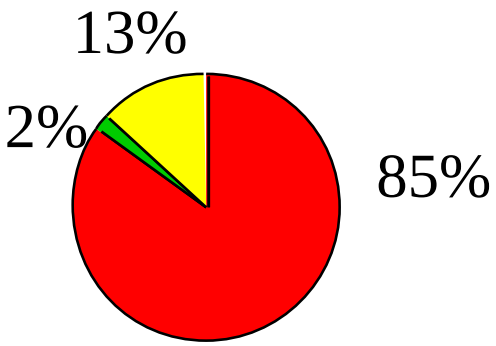
Human *vs E. coli*:

Genome size: x 1000
Number of genes: x 10

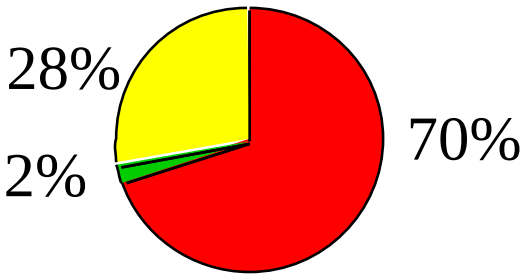
How many genes in the human genome ?

Technique	Gene estimate	Comments/assumptions
Estimation	100,000	if average size = 30 kb
Estimation	300,000	if average size = 10 kb
RNA reassociation	20,000 - 92,000	
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
Chromosome 21 (2000) + 22	40,000	
Exofish (2000)	28,000-34,000	Comparison human/fish
EST (2000)	35,000	Number of genes
EST (2000)	120,000	Number of transcripts
Complete genome (2001)	30,000-40,000	Known genes + predictions

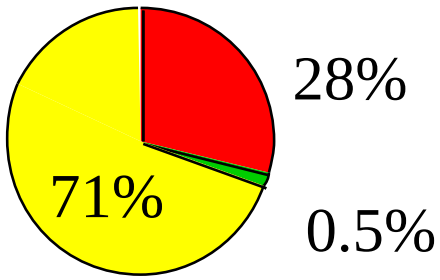
Proportion of functional elements within genomes



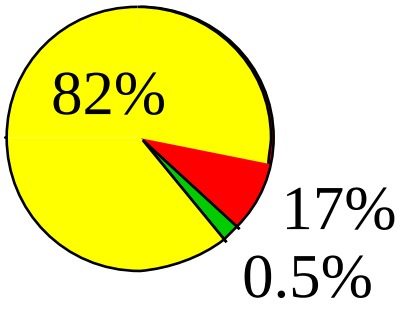
E. coli



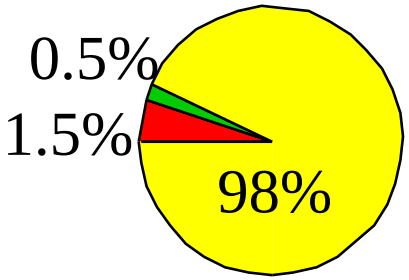
Yeast
S. cerevisiae



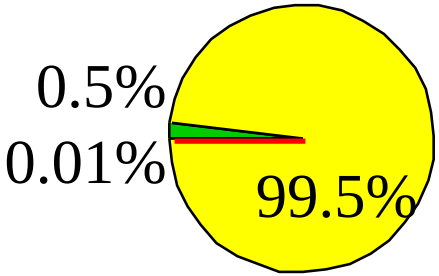
Nematode
C. elegans



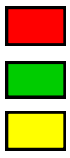
Drosophila



Human

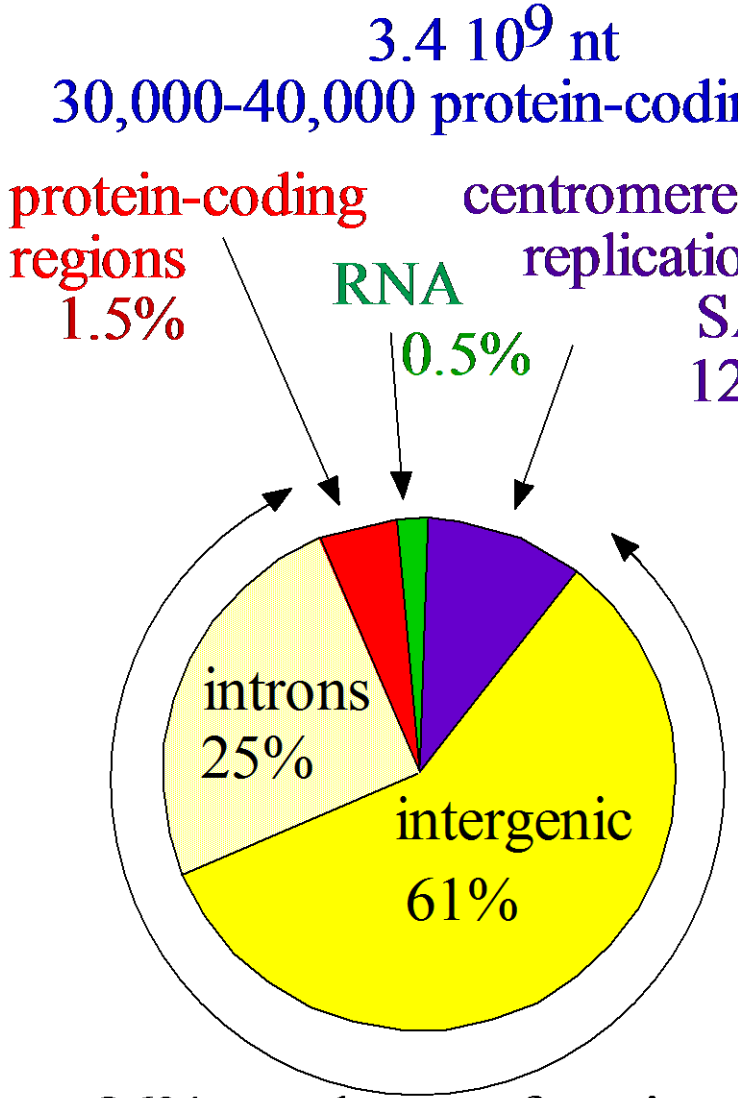


Lunfish
(dipnoi)



Coding (protein)
RNA
Non-coding

Functional elements in the human genome



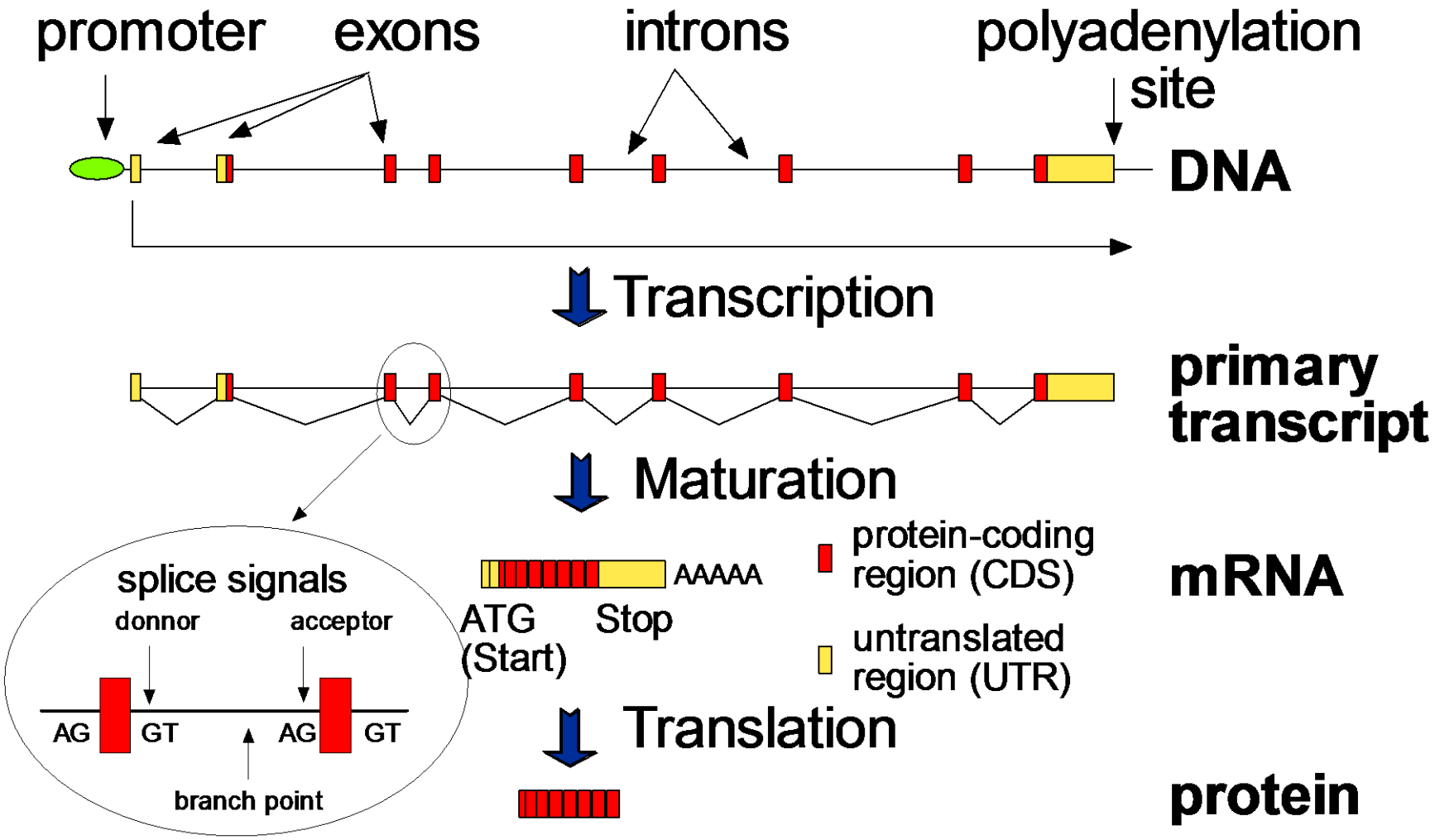
86% no (known) function

Untranslated RNAs: Xist, H19, His-1, bic, etc.

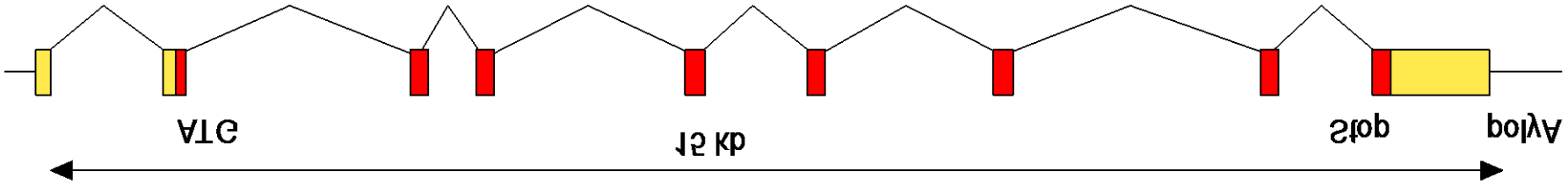
Regulatory elements: promoters, enhancers, etc.

Repeated sequences (SINES, LINES, HERV, etc.) : 40% of the human genome

Typical eukaryotic protein-coding gene



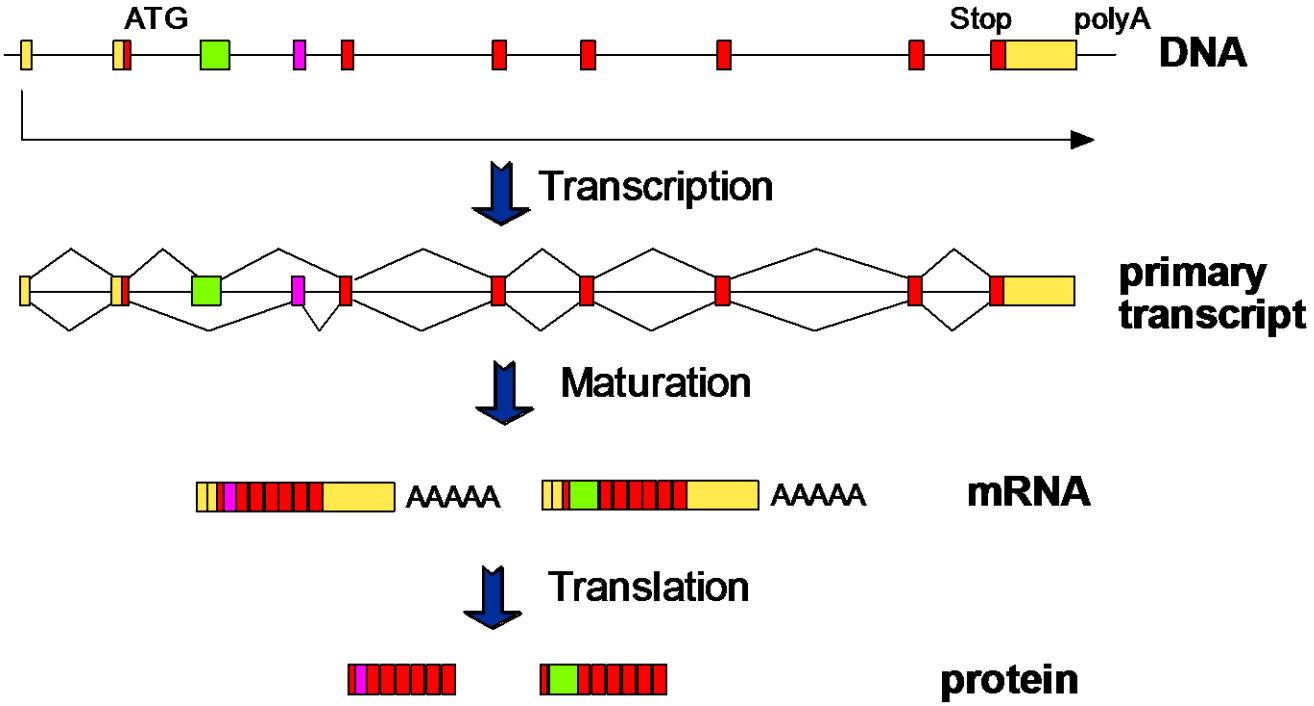
Structure of human protein genes



- 1396 complete human genes (exons + introns) from GenBank (1999)
- Average size (25%, 75%)
  - Gene 15 kb ± 23 kb (4, 16) (10% > 35 kb)
  - CDS 1300 nt ± 1200 (600, 1500)
  - Exon (coding) 200 nt ± 180 (110, 200)
  - Intron 1800 nt ± 3000 (500, 2000)
  - 5'UTR 210 nt (Pesole et al. 1999)
  - 3'UTR 740 nt (Pesole et al. 1999)
- Intron/exon
  - Number of introns: 6 ±3 introns / kb CDS
  - Introns / (introns + CDS): 80%
  - 5' introns in 15% of genes (more ?), 3 ’introns very rare

One gene, several products

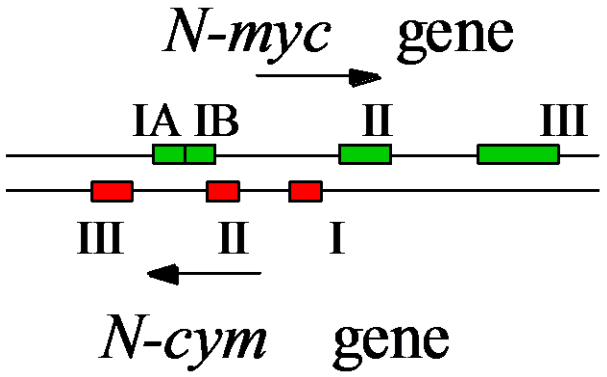
- Alternative splicing in more than 30% of human genes (Hanke et al. 1999)



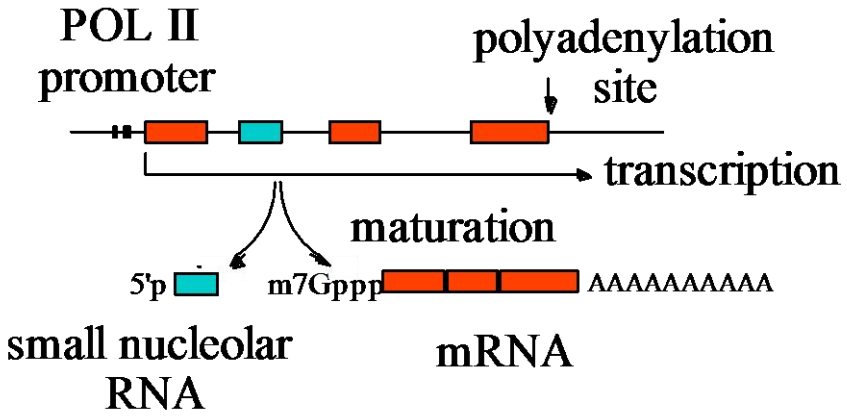
- Alternative promoter
- Alternative polyadenylation sites

Overlapping genes

Overlapping protein genes

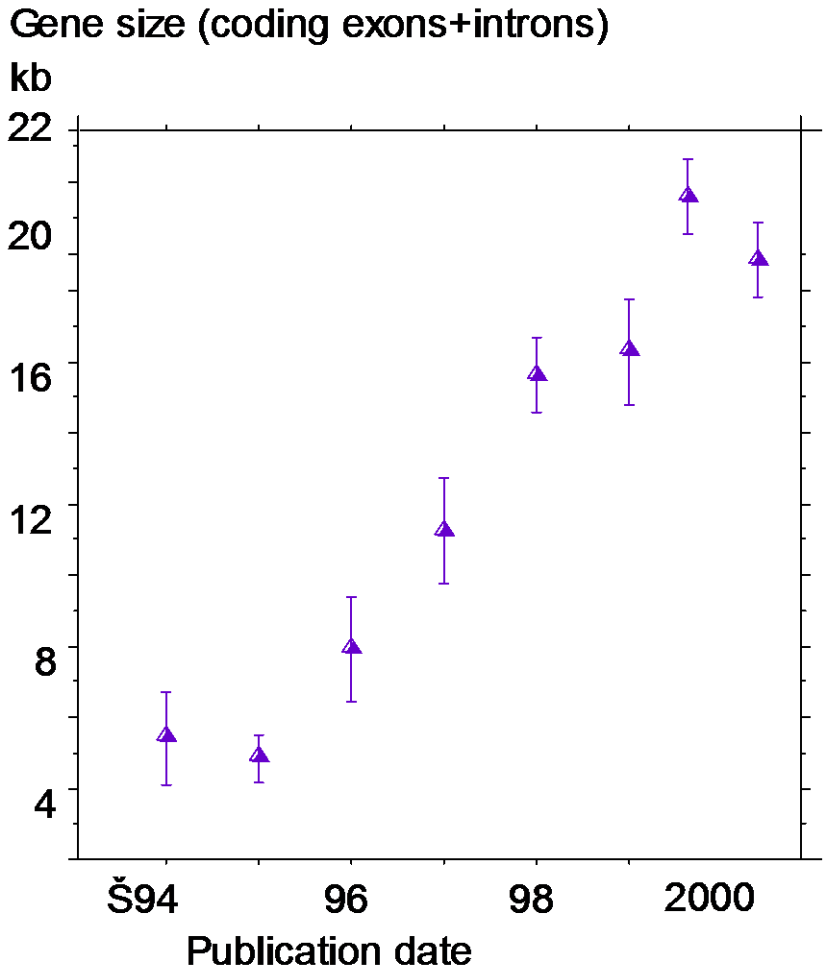


Small nucleolar RNA genes within introns of protein genes



Structure of human protein genes

- GenBank: bias towards short genes
- 2408 complete human genes (exons + introns)



Repeated sequences

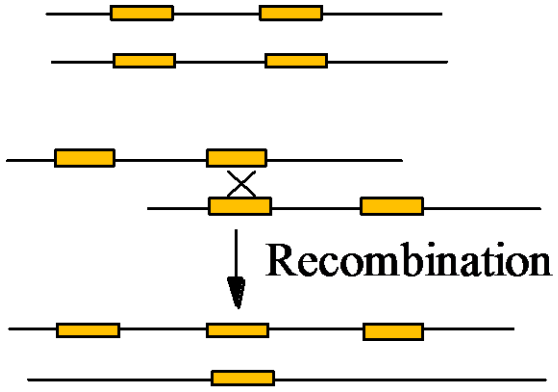
- Tandem repeats
  - Satellite
  - Minisatellite
  - Microsatellite
- Interspersed repeats
  - DNA transposons
  - Retroelements

Tandem repeats

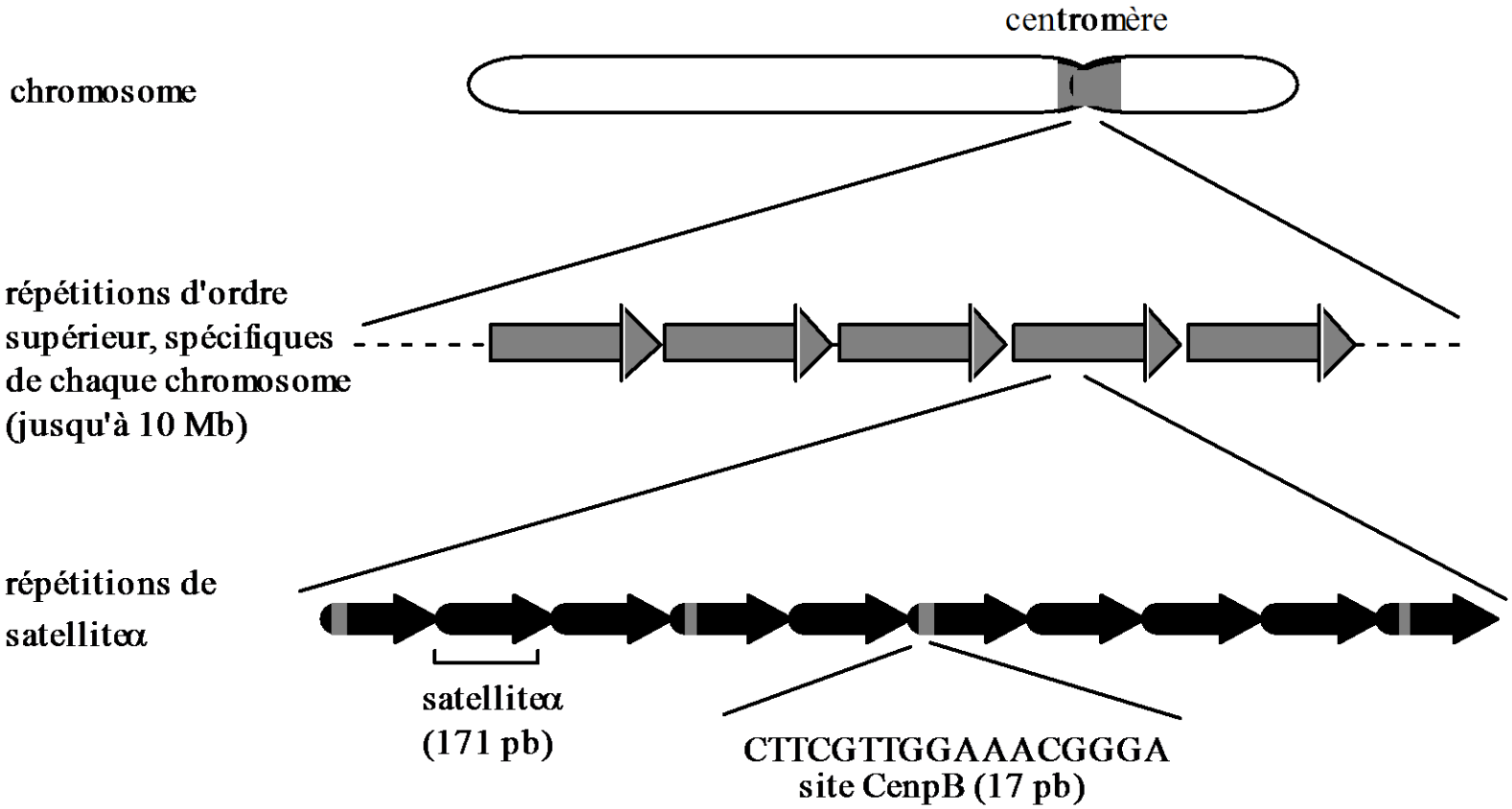
	motif	bloc size	% human genome
satellite:	2-2000 nt	up to 10 Mb	10%
minisatellite:	2-64 nt	100-20,000 bp	?
microsatellite:	1-6 nt	10-100 bp	2%

Slippage of the DNA polymerase: CACACACACACA

Unequal crossing-over:



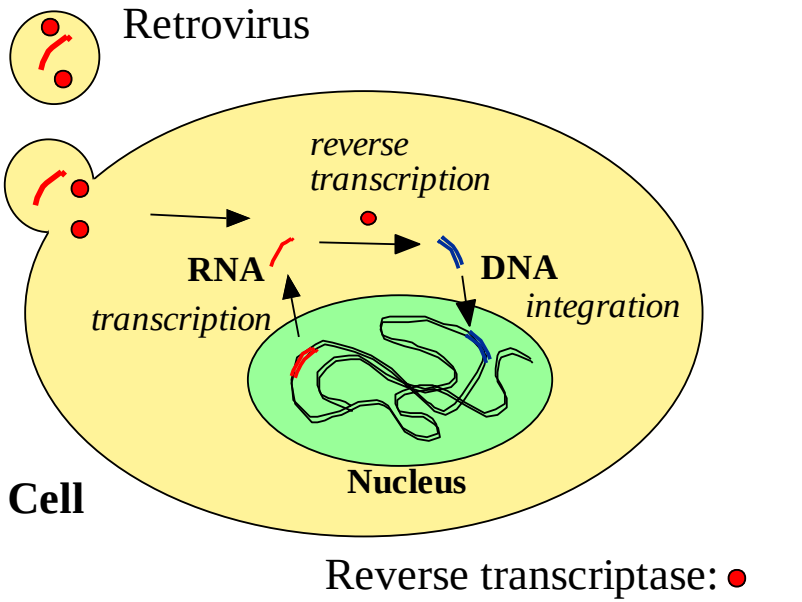
Centromeres, telomeres: Satellite DNA

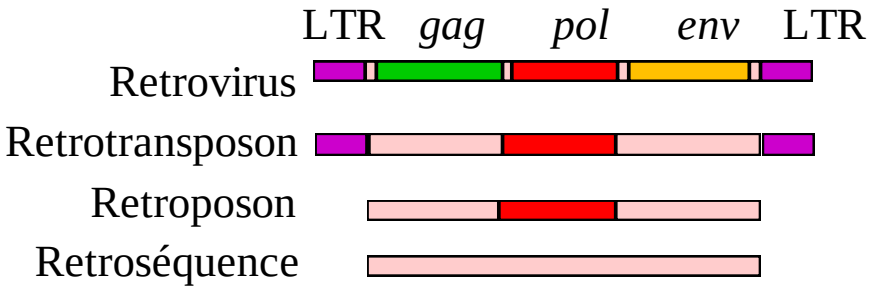


Interspersed repeats

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

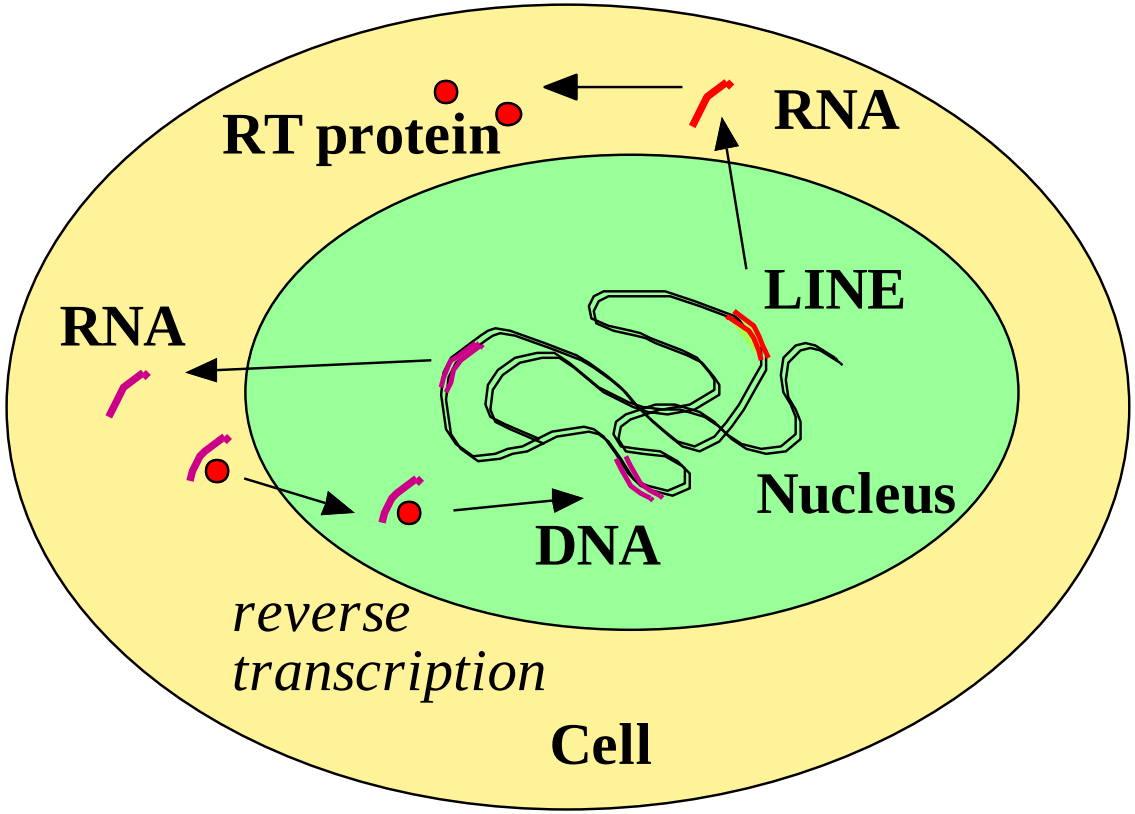
Retroelements





- **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

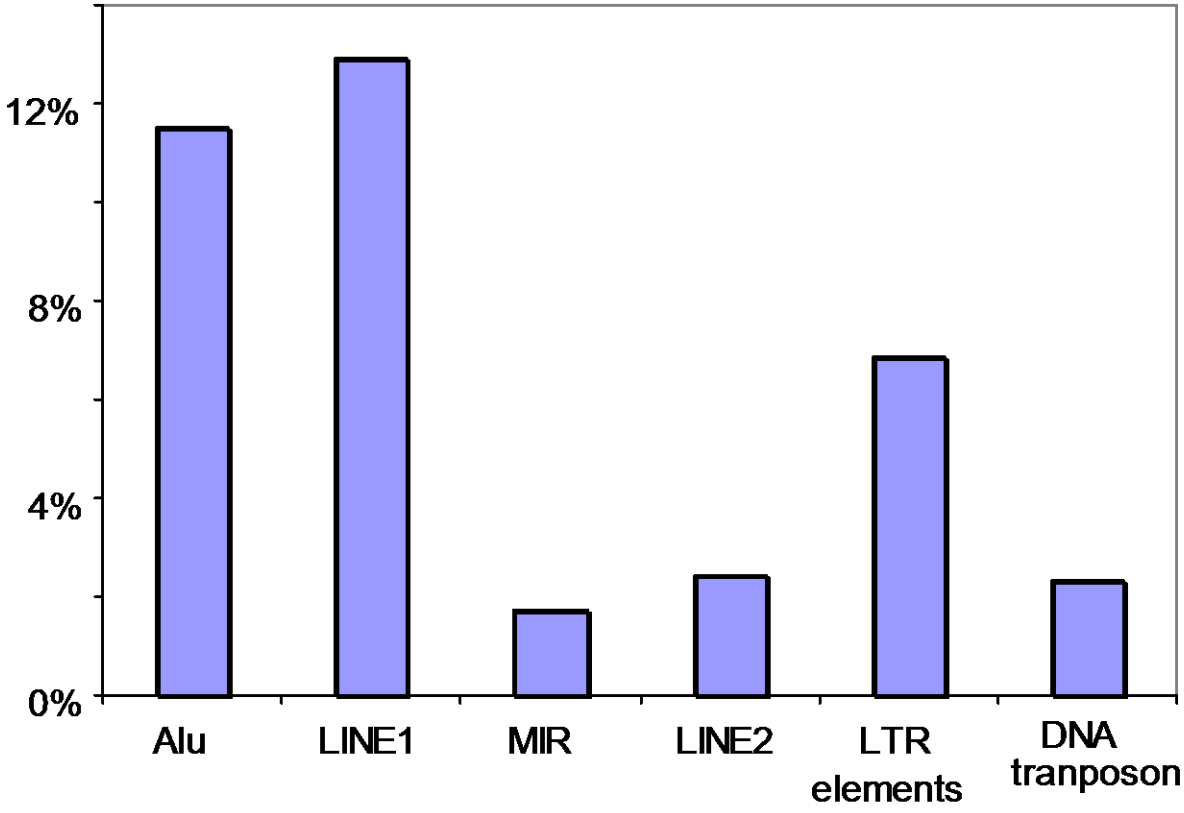
Retrosequences: opportunist retroelements



- LINE reverse transcriptase

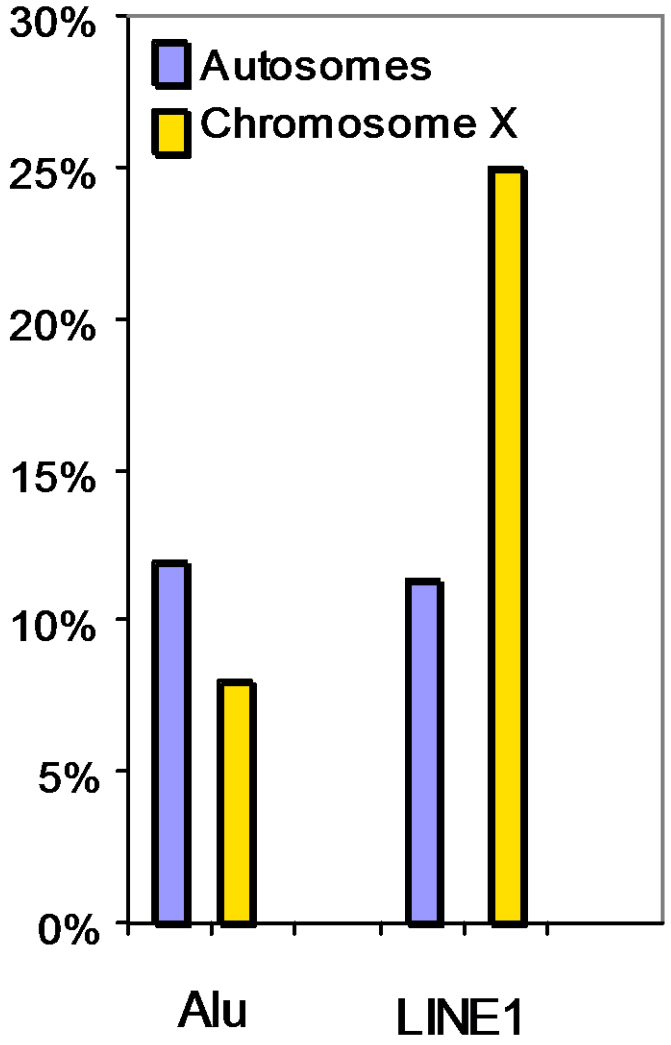
Frequency of transposable elements in the human genome

- Total = 42% (Smit 1999)
- Probably underestimated



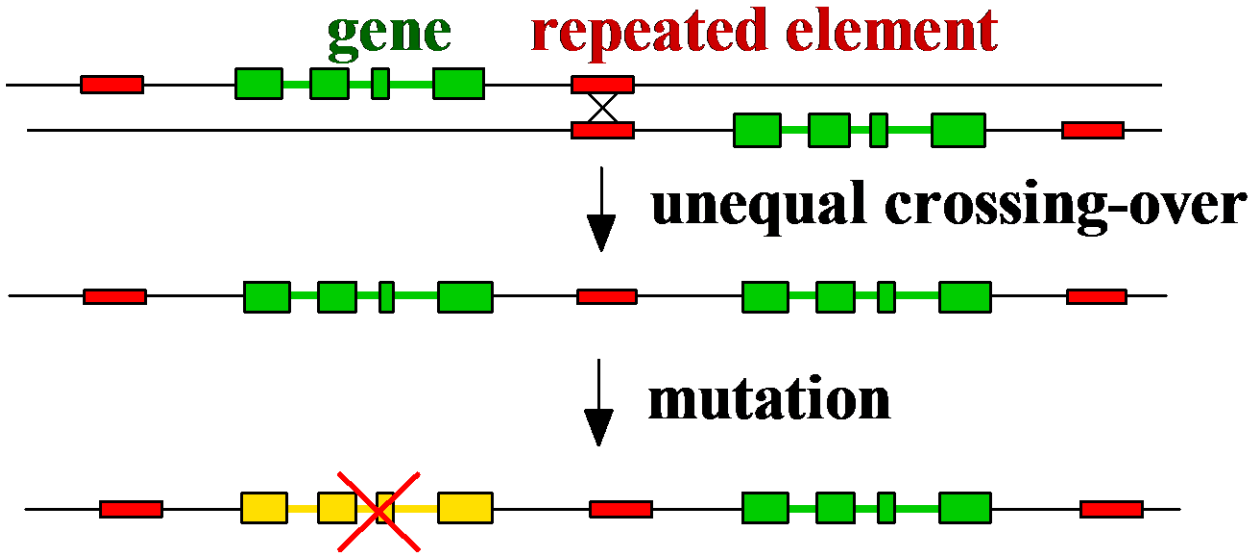
The frequency of transposable elements is not uniform along the human genome:

e.g. inter-chromosomic variations
(Smit 1999)

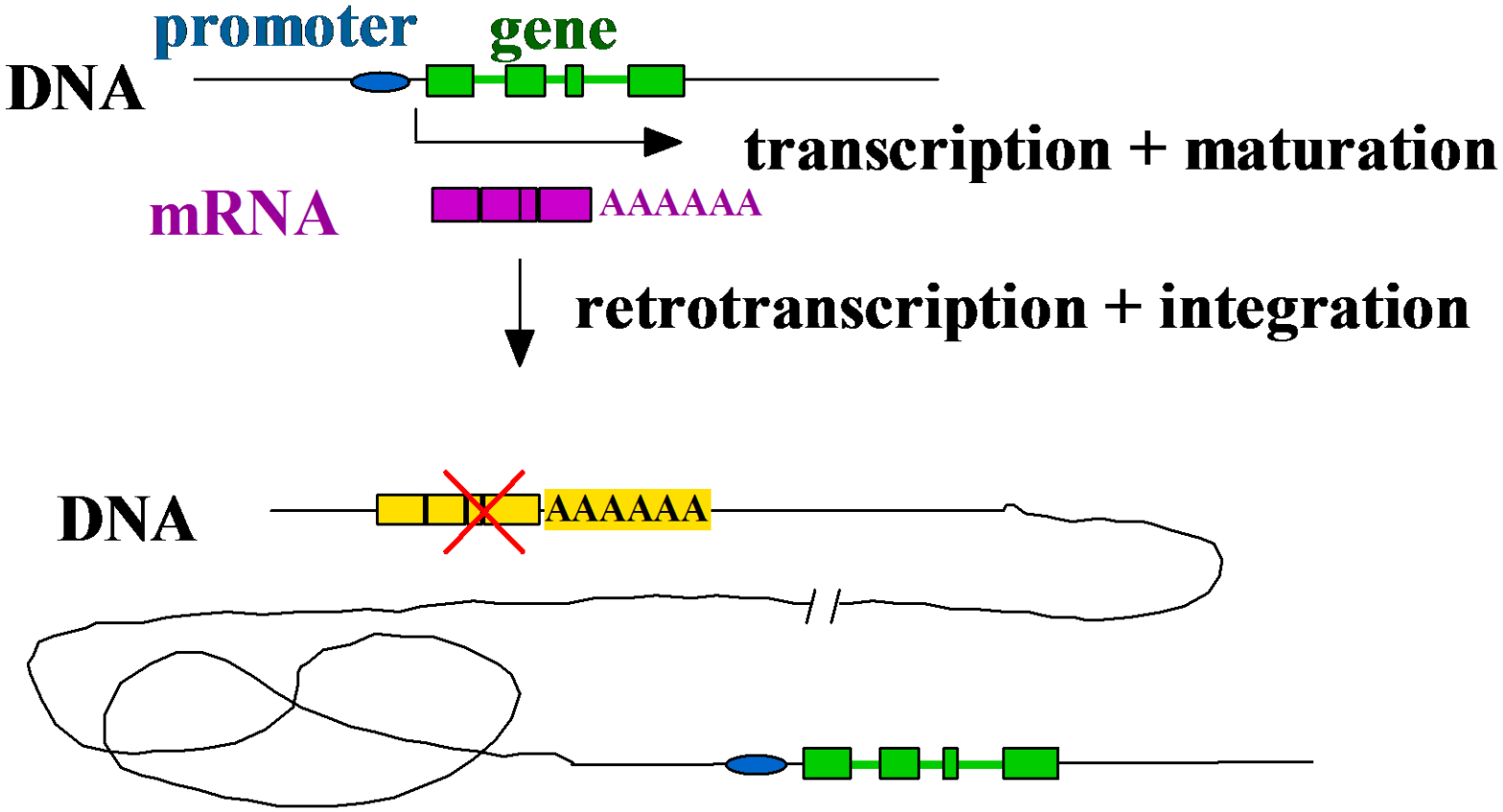


Pseudogenes

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



Retropseudogenes

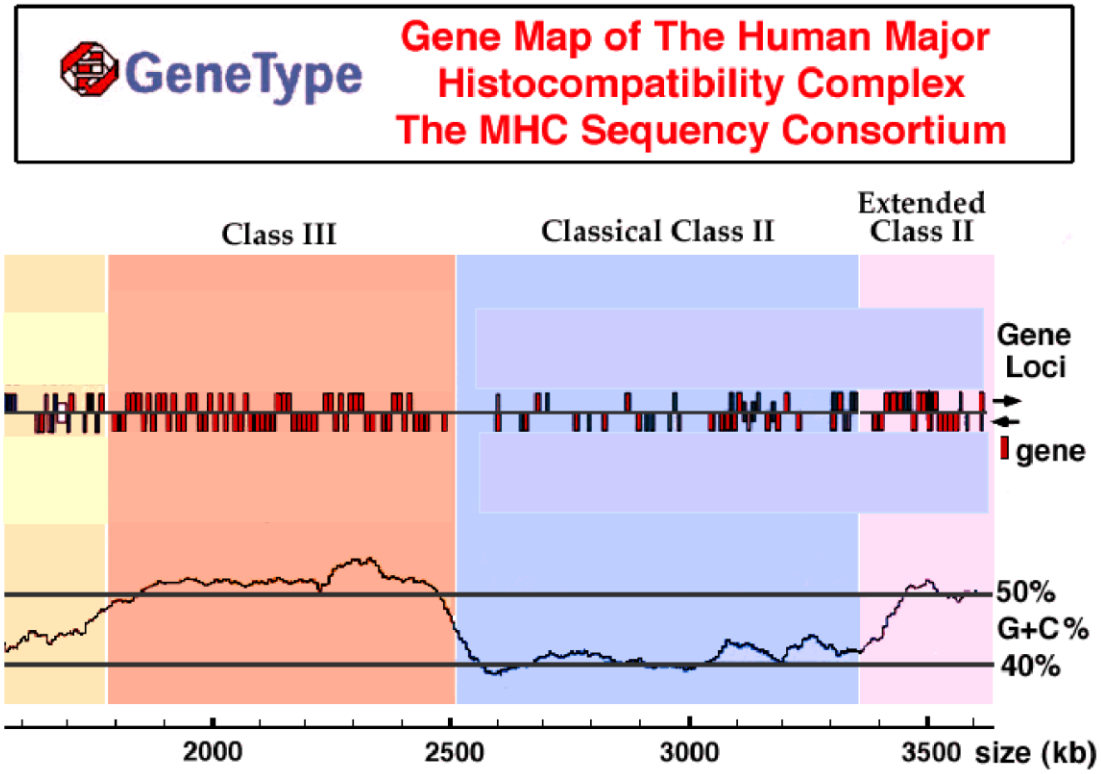


Retropseudogenes

- 23,000 to 33,000 retropseudogenes in the human genome
- Often derive from housekeeping genes

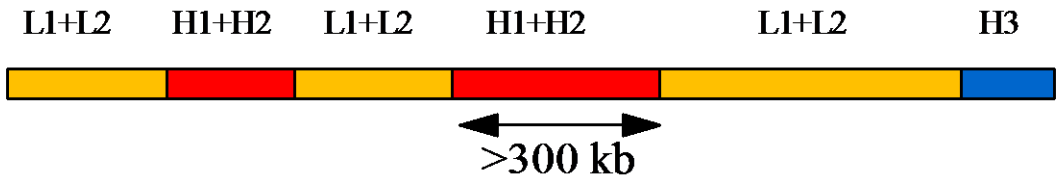
Vertebrate genome organization: variations of base composition along chromosomes

Sequence of human MHC



Isochore organization of vertebrate genomes

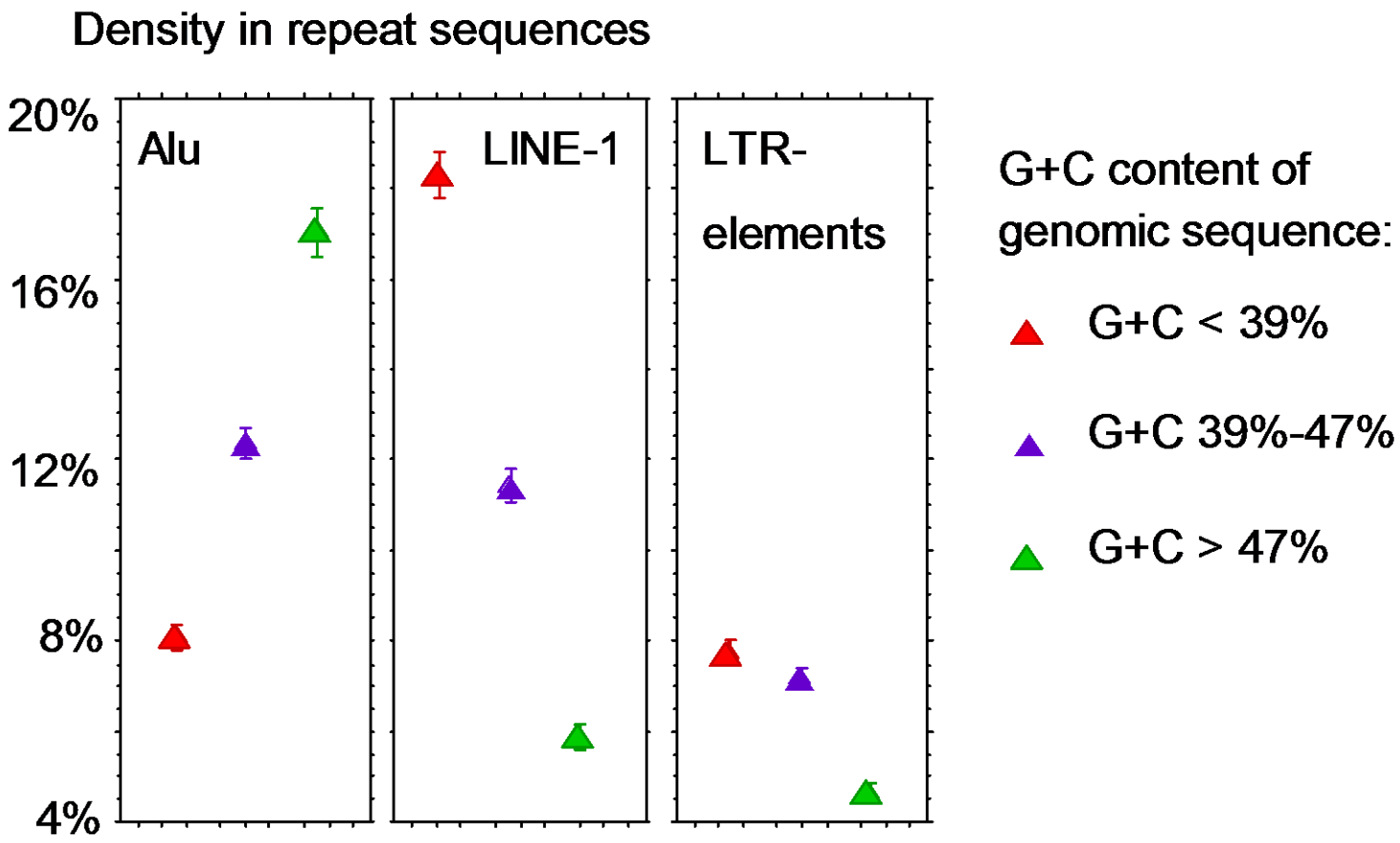
Bernardi et al. 1985



isochore	%C+G	% total genomic DNA
L1+L2 :	33%-44%	62 %
H1+H2 :	44%-51%	31%
H3 :	51%-60%	3-5%

- **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 expression ??** -> No
- **Gene structure** (Duret, 1995)

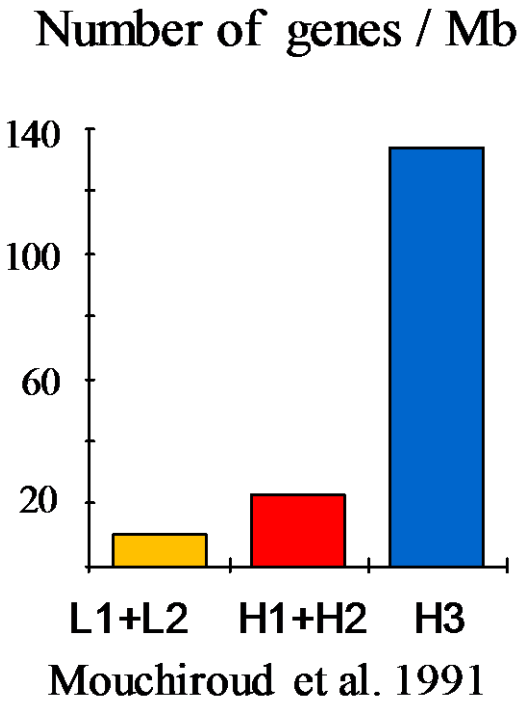
Isochores and insertion of repeat sequences (Smit 1999)



4419 human genomic sequences > 50 kb

Isochores and gene density

isochore	% total genomic DNA	%total genes
L1+L2 :	62 %	31%
H1+H2 :	31%	39%
H3 :	3-5%	30%



MHC locus (3.6 Mb) (The MHC sequencing consortium 1999)

Class I, class II (H1-H2 isochores): 20 genes/Mb, many pseudogenes

Class III (H3 isochore): 84 genes/Mb, no pseudogene

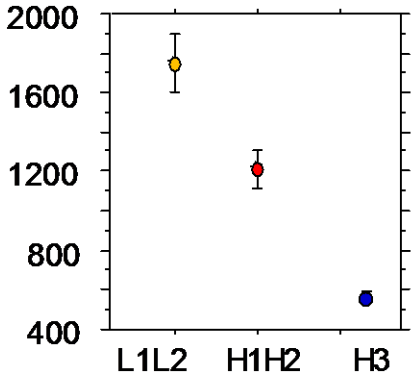
Class II boundaries correlate with switching of replication timing

Isochores and introns length

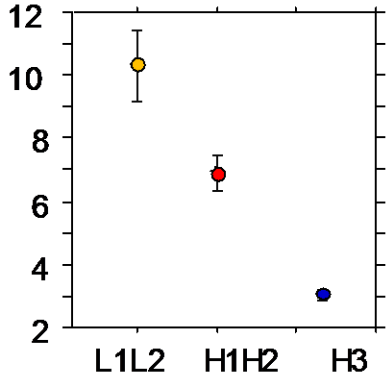
Duret, Mouchiroud and Gautier, 1995

- 760 complete human genes
- L1L2: intron G+C content < 46%
- H1H2: intron G+C content 46-54%
- H3: intron G+C content >54%

Average intron length (bp)



Gene compaction
(intron length/coding region length



Mammalian genomes: summary

- Genes, regulatory elements: ~ 2%
- Non-coding sequences: ~ 98%
  - Satellite DNA (centromeres) ~ 10%
  - Microsatellites ~ 2%
  - Transposable elements ~ 42%
  - Pseudogenes ~ 1%
  - Other (ancient transposable elements?) ~ 43%
- 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*)
- 2000: environ 40 génomes complets

Passage de l'artisanat à l'industrie

- 1980-1995: séquencer pour répondre à une question donnée: de la biologie à la séquence
  - séquenceurs: tous les laboratoires de biologie moléculaire
  - séquences: des gènes ou des ARNm (< 10 kb)
  - informations biologiques associées aux séquences: riches

phénotype ↛ gène

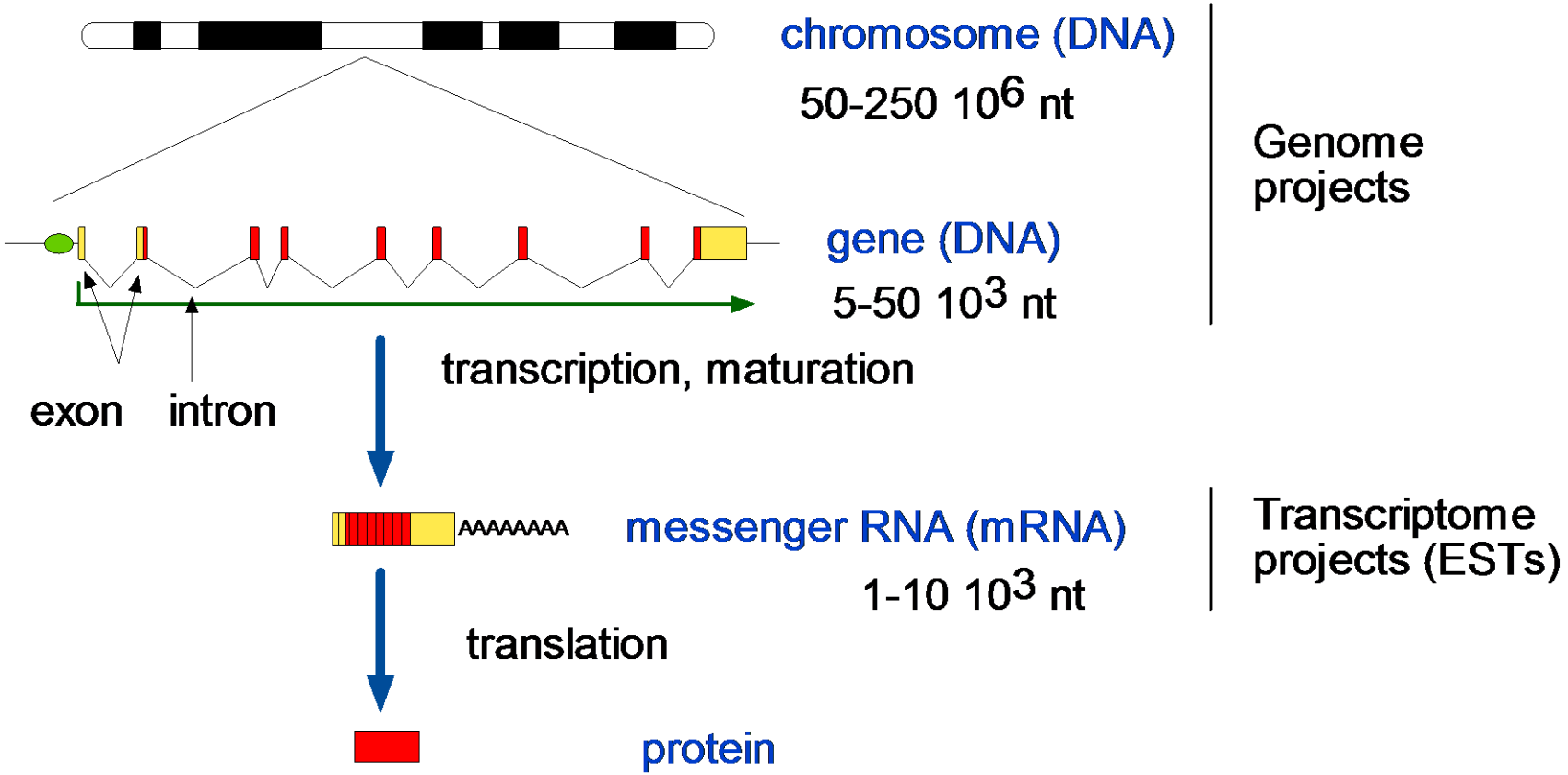
- >1995: séquençage systématique à grande échelle: de la séquence à la biologie
  - séquenceurs: quelques grands centres de séquençage
  - séquences: grands fragments génomiques, chromosomes, etc ...
  - informations biologiques associées aux séquences: pauvres

gène ↛ phénotype

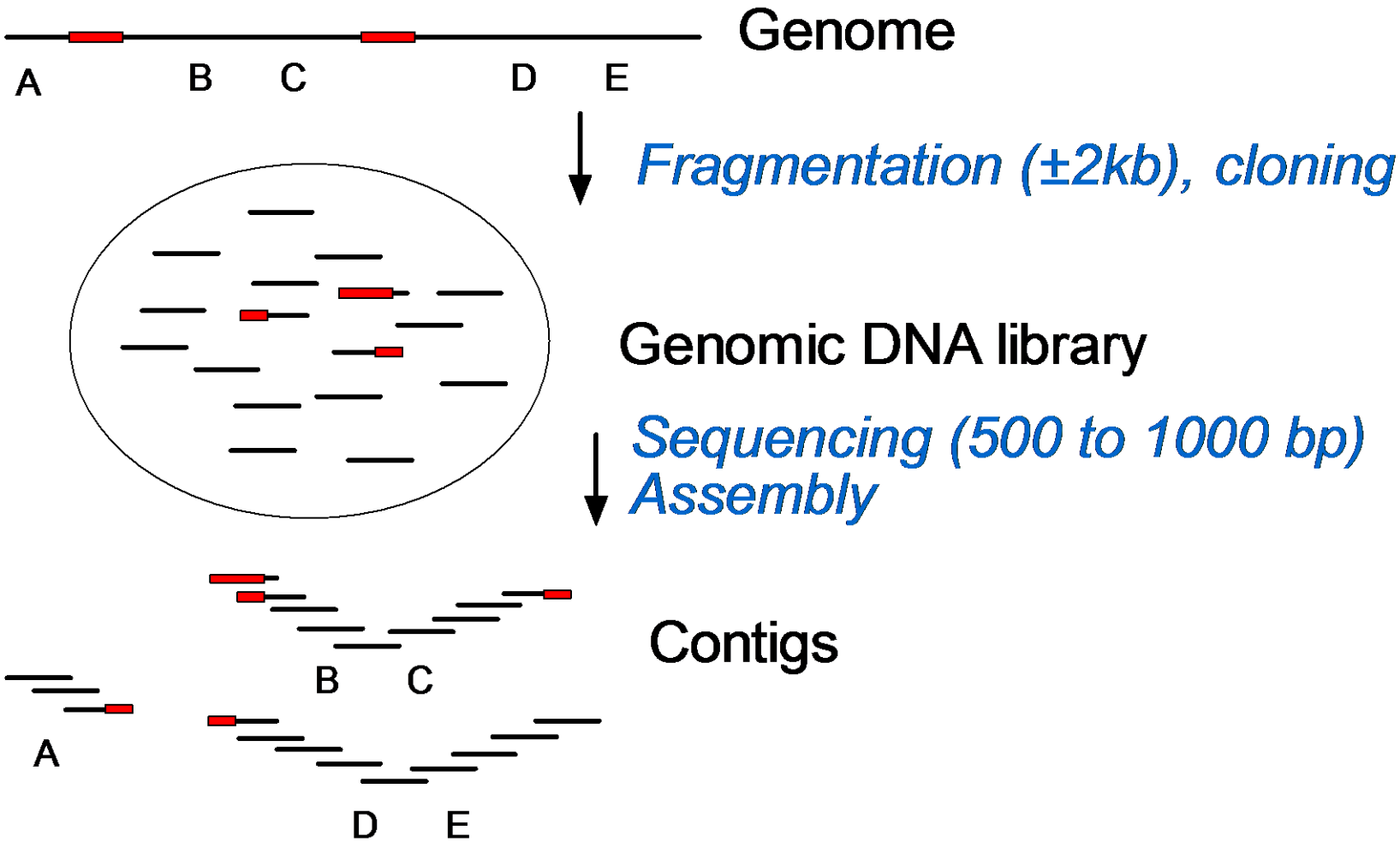
Genome projects

- Make the inventory of all the genetic information necessary for the development and reproduction of an organism
- Understand genome organization (bag of genes or integrated information system ?)
- Understand genome evolution
- Applications in medicine, agronomy, industry

Sequencing Projects : Genome / Transcriptome

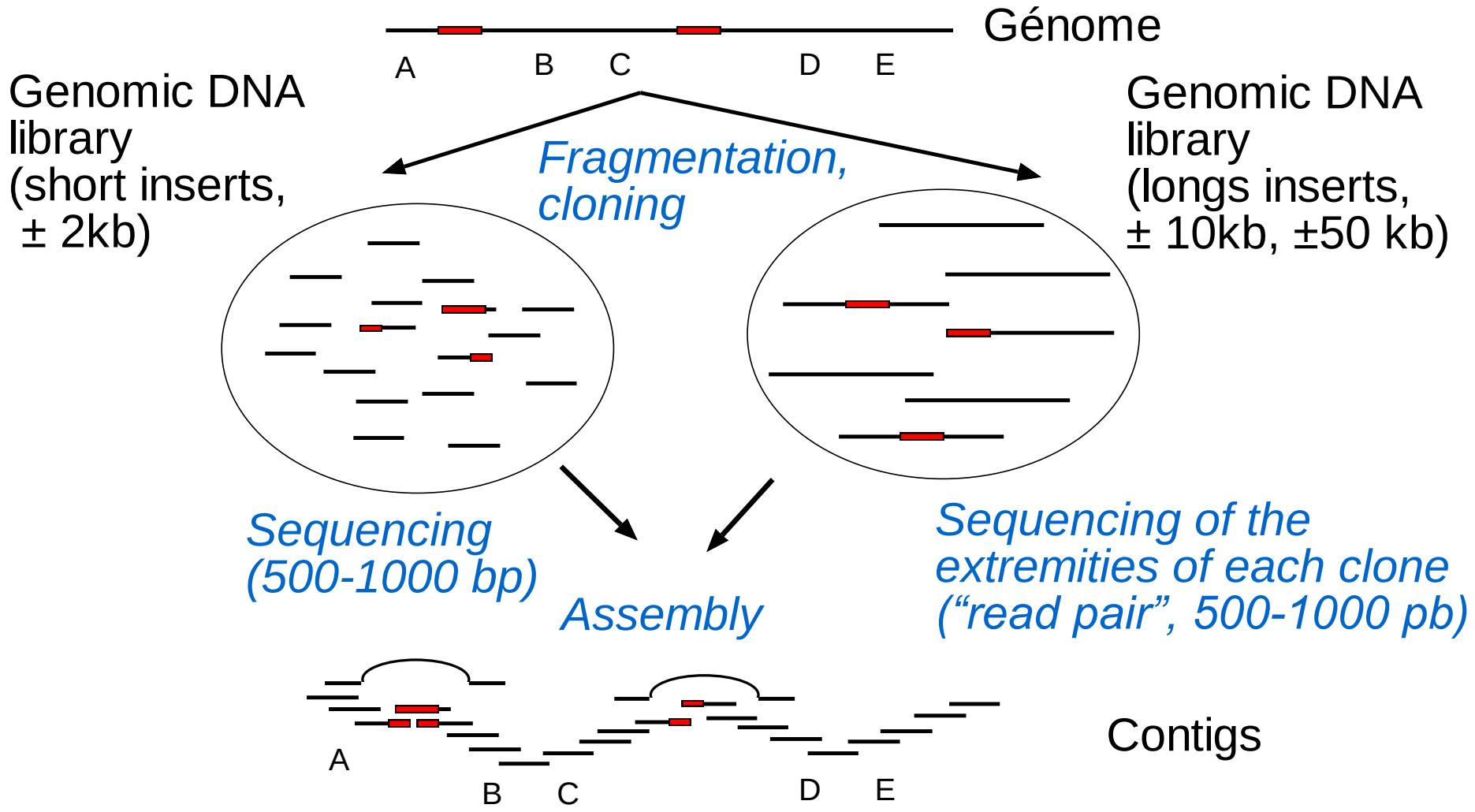


Shotgun sequencing



Shotgun sequencing: improvement

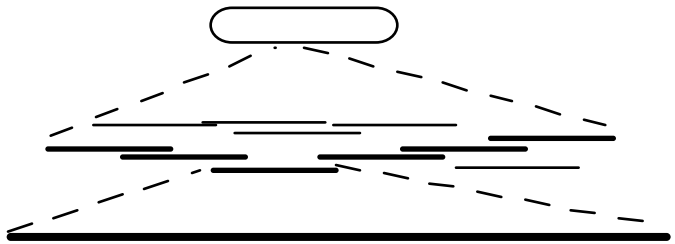
(E. Myers)



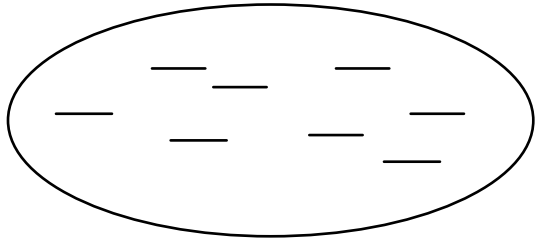
Strategy for sequencing the human genome

(Academic international consortium)

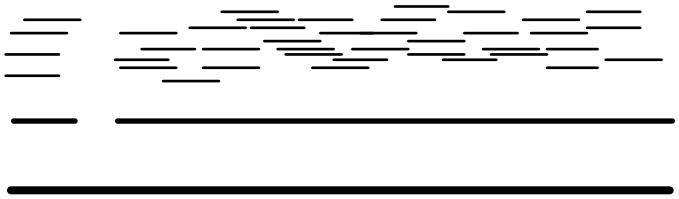
- Genome
- Cloning of long inserts (e.g. BAC DNA library : 100-200 kb)
- Genomic mapping
- Selection of clones to sequence



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

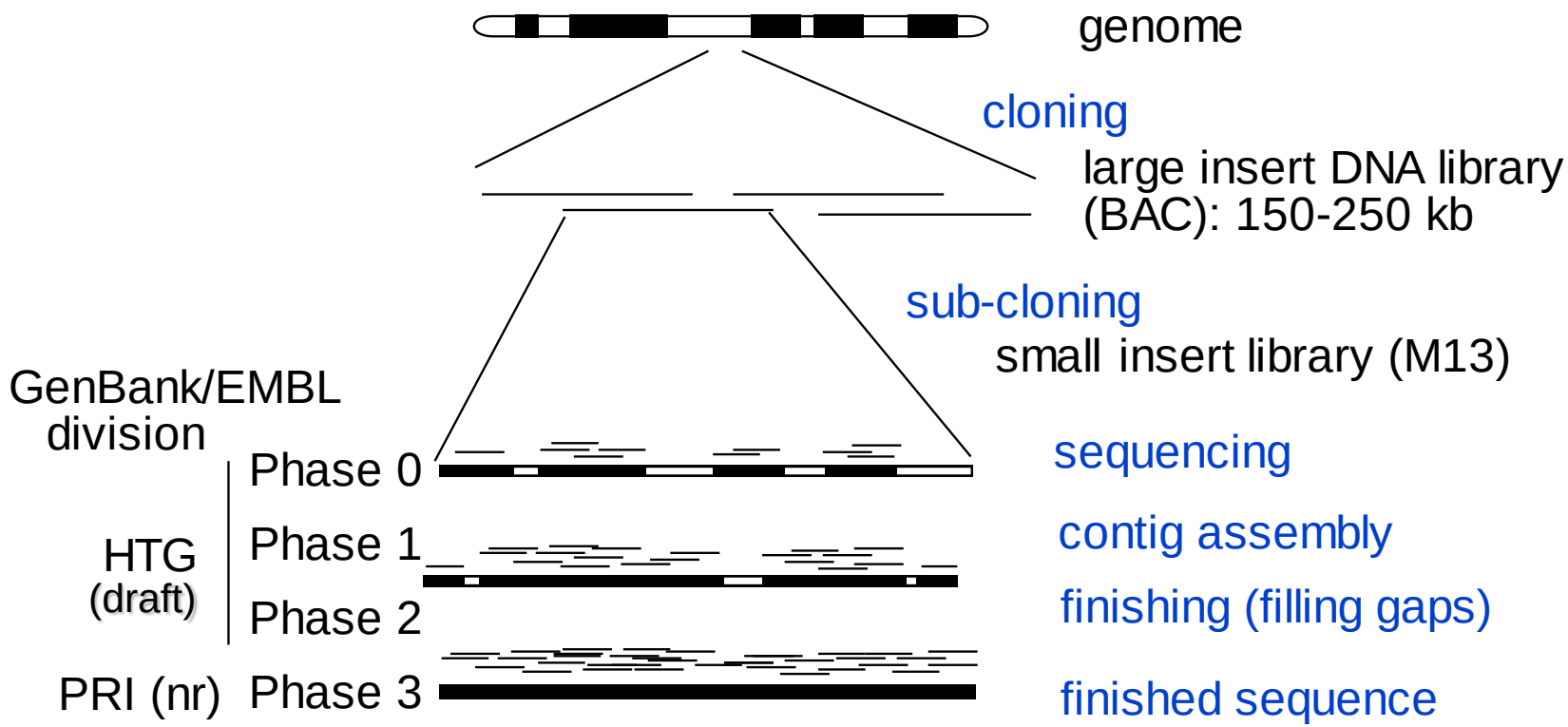


- Sequencing M13 clones
- Assembly: contigs
- Finishing: gap closure



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)

The human genome sequencing project
Where are we today (March 2001) ?

- According to Philipp Bucher (SIB, Lausanne) statistics and genome coverage estimates (see also EBI's statistics: http://www.ebi.ac.uk/~sterk/ genome-MOT)

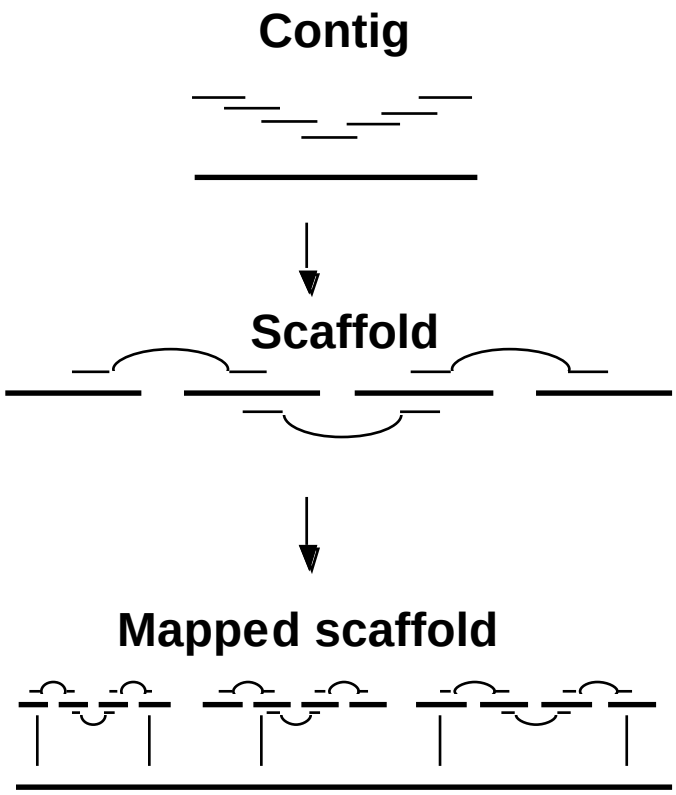
Estimated size of human genome	3400 MB	100.00%
EMBL sequences in HUM division:	1200 MB	35.30%
(10,073 entries, ave. Size: 120 kb)		
Human sequences in HTG division:	3813 MB	112.00%
(24953 entries, ave. Size: 153 kb)		
Total:	5013 MB	147.00%
Estimated redundancy (35%)	-1755 MB	-51.60%
Corrected total:	3258 MB	95.80%

Complete genome sequence ?

- **Contig**: sequence without any gap
- 170,000 contigs, 16 kb in average (cover 95% of the genome). Longest contig: 2 Mb

- **Scaffold**: set of ordered and orientated contigs; gaps of known length
- 1935 long scaffolds (>100 kb), 1.4 Mb in average (cover 86% of the genome), 100,000 gaps (2kb in average) + 51,000 short scaffolds (5% of the génome)

- **Mapped scaffold**: set of scaffold localized along chromosomes (but not always ordered and orientated, gaps of unknown length)
- Scaffolds ordered and orientated: 70% of the genome
- Scaffold ordered: 84% of the genome



http://genome.ucsc.edu/

- CELERA: similar results

Genome projects: complete sequencing

- Bacteria: 45 complete genomes (19 during the last 12 mounths !)
- Archea: 10 complete genomes
- Eukaryotes: 5 (6) complete genomes
  - *G. theta* (nucleomorph) 0.5 Mb 100%
  - yeast: 13 Mb 100%
  - *C. elegans* 100 Mb 95%
  - *A. thaliana* 120 Mb 95%
  - *Drosophila* 170 Mb 60% (100%)
  - human 3200 Mb 95%
    - 2/3 « draft » sequence, finished in 2003
  - mouse 3000 Mb 10%
    - 3 x « draft » sequence in 2001

Genome Survey Sequence (GSS) projects

- Random sampling of genomic sequences: give (at low cost) an overview of the content of a genome

- Genomic DNA library
- Sequencing of clones:
  - Short sequences (< 1kb)
  - Single read => high rate of sequencing errors (1-3%)
  - Accurate enough to identify genes (exons)
  - Largely automated => low cost

Large scale GSS projects

Species	Nb. of GSS
Mus musculus (mouse)	937 975
Homo sapiens	870 073
Tetraodon nigroviridis	188 963
Oryza sativa (rice)	93 164
Trypanosoma brucei	91 319
Strongylocentrotus purpuratus (sea urchin)	76 019
Arabidopsis thaliana (plant)	61 266
Takifugu rubripes (pufferfish)	47 111
Drosophila melanogaster	45 323

From GenBank (September 2001)

Transcriptome projects: Expressed Sequence Tags (ESTs)

- Inventory of all mRNAs expressed by an organism, in different tissues, development stages, pathologies, …
  - Single pass sequences: high error rate (>1%), partial mRNA sequences (300-500 bp)
  - Redundancy (highly expressed genes)
  - Accurate enough to identify genes (exons)
  - Largely automated
- Very useful to identify genes in genomic sequences, + information on expression pattern
  - Usually derived from poly-dT-primed cDNA -> bad coverage of 5' regions of long mRNAs
  - 60-80% of human genes represented in public EST database, but only 25-50% of the total coding part of the genome
- Possibility to get cDNA clones from the IMAGE consortium (http://image.llnl.gov/)

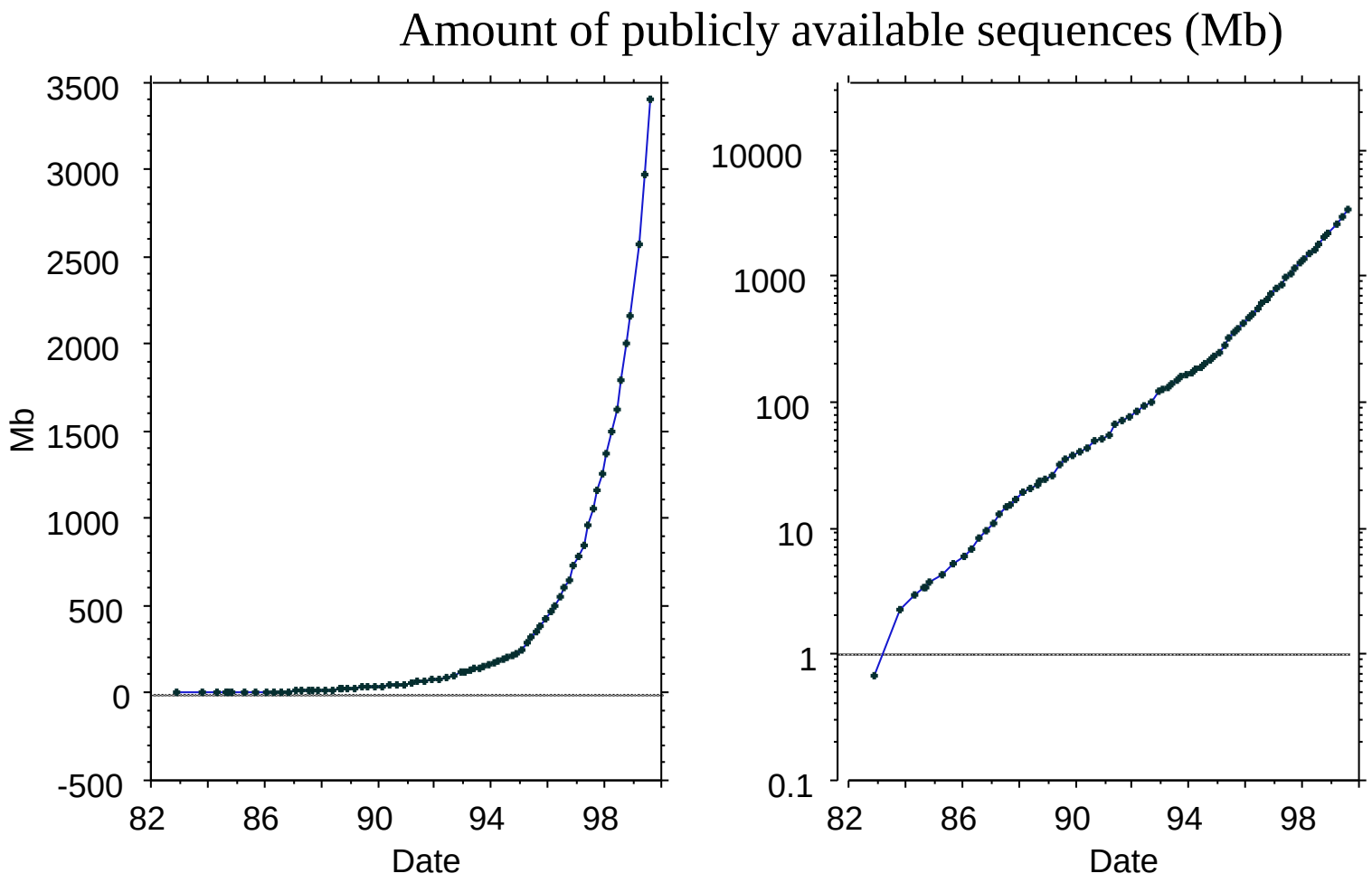
Large scale EST projects

Species	Nb. of ESTs
Homo sapiens	3 789 914
Mus musculus (mouse)	2 153 036
Rattus sp.	317 066
Drosophila melanogaster	255 456
Caenorhabditis elegans (nematode)	135 203
Lycopersicon esculentum (tomato)	126 736
Danio rerio (zebrafish)	117 276
Arabidopsis thaliana (plant)	113 331
Zea mays	106 595
Oryza sativa (rice)	80 365

From GenBank (September 2001)

Exponential increase of sequence data

- Doubling time: 13 mounths



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 ??)

Prediction of gene function

- Analysis of expression pattern (ESTs, …)
- 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 ?

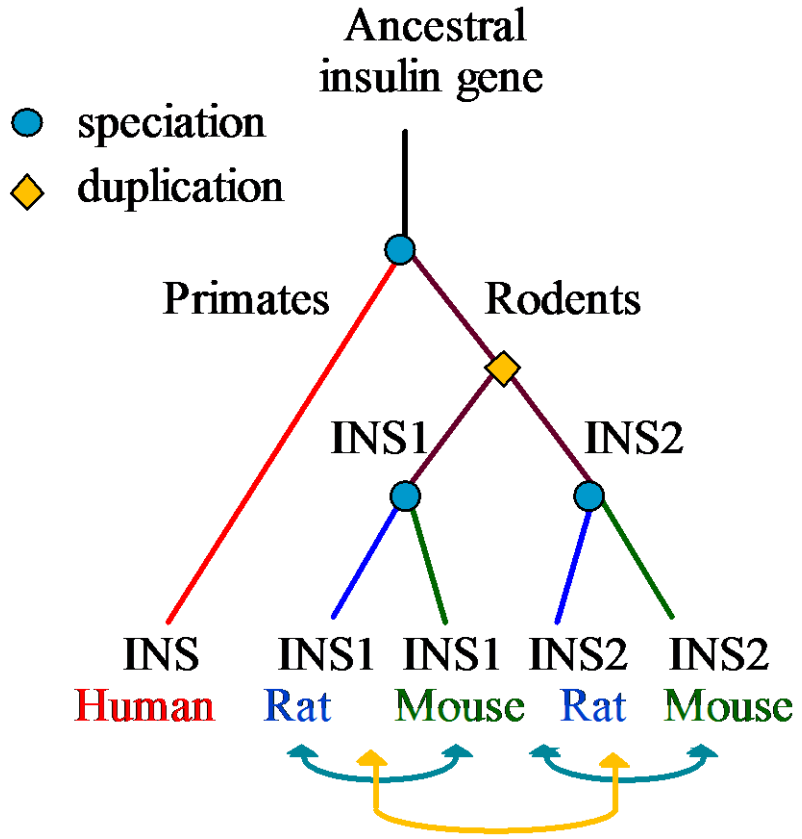
```
MZEORFG:  1   ILNSPDRACNLAKQAFDEAISELDSLGEESYKDSTLIMQLLXDNLTLWTSDTNEDGGDE 59
              I N+P++AC LAKQAFD+AI+ELD+L E+SYKDSTLIMQLL DNLTLWTSD  ++   E
BOV1433P: 186 IQNAPEQACLLAKQAFDDAIAELDTLNEDSYKDSTLIMQLLRDNLTLWTSDQQDEEAGE 244

 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, complete cds
ACCESSION   J03868

LOCUS       MZEORFG       187 bp    mRNA            PLN       31-MAY-1994
DEFINITION  Zea mays putative brain specific 14-3-3 protein, tau protein
            homolog  mRNA, partial cds.
```

Orthology/paralogy



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

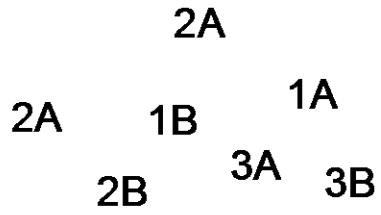
↔ Orthologues: homologous genes that have diverged after a speciation

↔ Paralogues: homologous genes that have diverged after a duplication

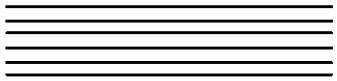
! Orthology ≠ functional equivalence

Phylogenetic approach for function prediction

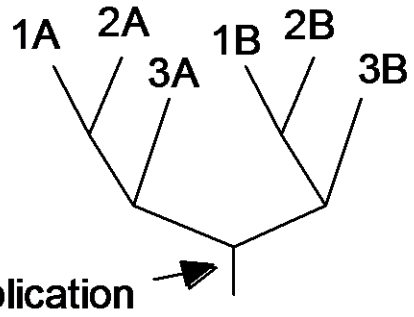
1) Identify homologues



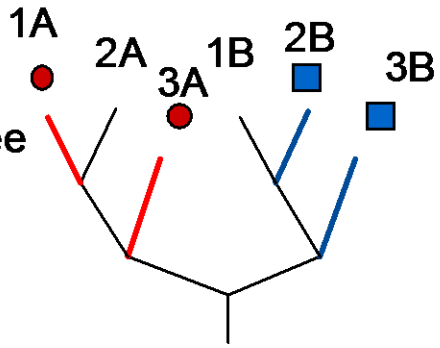
2) Align sequences



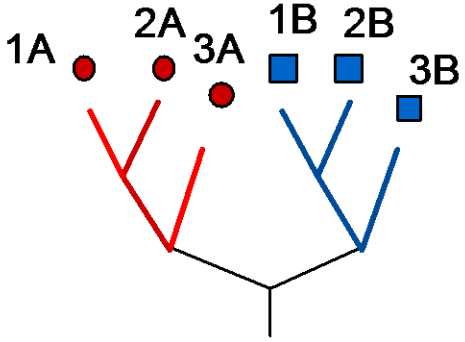
3) Compute the phylogenetic tree



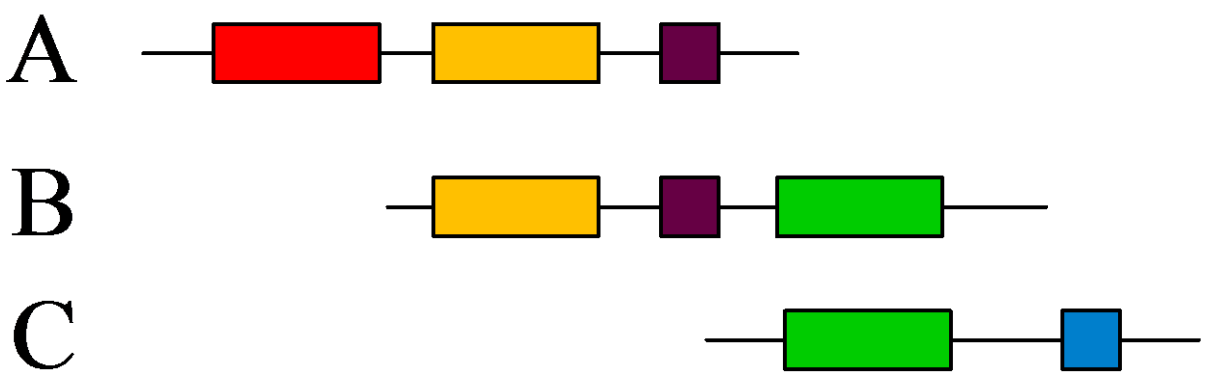
4) Place known functions on the tree



5) Infer the likely function of other genes



Modular evolution



Systematic annotation of the human genome

- ENSEMBL project
  - http://www.ensembl.org/
- Human Genome Project Working Draft at UCSC
  - http://genome.ucsc.edu/
- The genome channel
  - http://compbio.ornl.gov/channel/index.html

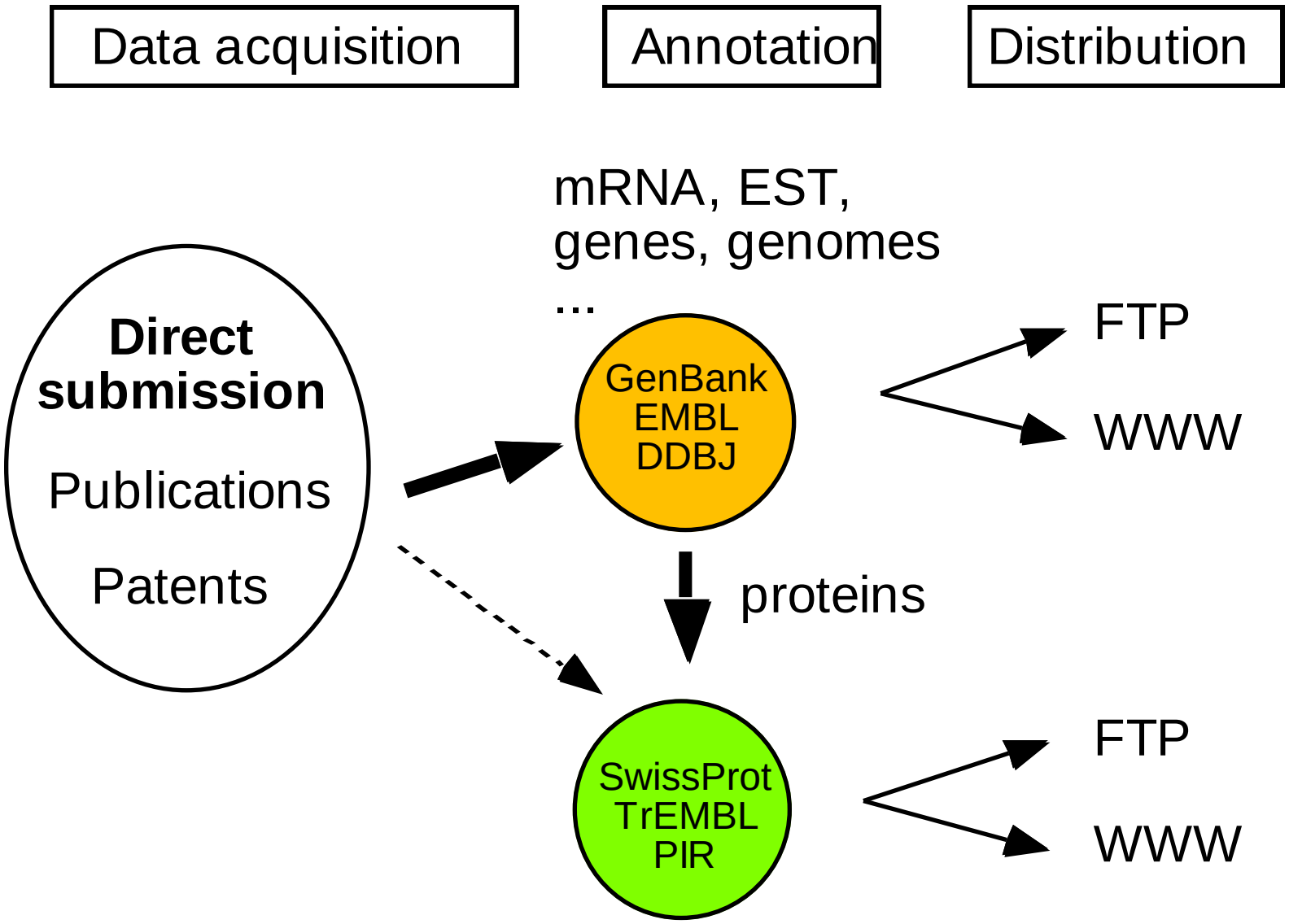
Databases for molecular biology

- Sequences
  - General databases (DNA, proteins)
  - Specialised databases
- Polymorphism
- Proteins structure
- Genomic mapping
- Gene expression
- Genetic diseases, phenotypes
- Bibliography
- …
- Databases of databases (dbCAT)

General sequence databases

- DNA databases :
  - EMBL (Europe) (1980)
  - GenBank (USA) (1979)
  - DDBJ (Japan) (1984)
  - These 3 centres exchange their data daily
    - ↘ identical content

- Protein databases :
  - SwissProt-TrEMBL (Switzerland, Europe) (1986 and 1996)
  - PIR (International)


Data acquisition
Annotation
Distribution
mRNA, EST,
genes, genomes
...
Direct submission
Publications
Patents
GenBank
EMBL
DDBJ
FTP
WWW
proteins
SwissProt
TrEMBL
PIR
FTP
WWW

Size of GenBank/EMBL
(October 2001)

- 14.2 10^9 nucleotides.
- 13.3 10^6 sequences.
- 764 000 genes (proteins and RNAs).
- 256 000 bibliographic references.
- 57 giga-bits on disk.

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

GenBank release 125 (October 2, 2001)

Division	Entries	Nucleotides	% nt
EST	9,014,899	4,104,167,129	29%
HTG	88,432	4,608,681,226	32%
GSS	2,706,132	1,480,201,675	10%
Other	1,459,835	4,036,209,322	28%
Total	13,269,298	14,229,259,352	100%
Human	5,006,832	7,942,037,394	56%

Content of DNA databases: taxonomic sampling

- 72,000 species for which there is at least one sequence
- 9 species (0.01%) totalize 85% of sequences
  - Homo sapiens 62.1%
  - Mus musculus 7.7%
  - Drosophila melanogaster 6.1%
  - Caenorhabditis elegans 3.3%
  - Arabidopsis thaliana 2.9%
  - Oryza sativa 1.3%
  - Rattus norvegicus 0.8%
  - Danio rerio 0.6%
  - Saccharomyces cerevisiae 0.6%

Structure of database entries

- The format of entries is different in EMBL and GenBank/DDBJ
- The content is the same
- Text with structured fields

Fields ID, AC, NI and DT

Identifiers (sequence name and accession number), date of creation and last modification of the entry.

```
ID   BSAMYL    standard; DNA; PRO; 2680 BP.
XX
AC   V00101; J01547
XX
NI   g39793
XX
DT   13-JUL-1983 (Rel. 03, Created)
DT   12-NOV-1996 (Rel. 49, Last updated, Version 11)
```

Fields DE, KW, OS and OC

General information on sequences (definition, keywords, taxonomy).

```
DE   Bacillus subtilis amylase gene.
XX
KW   amyE gene; alpha-amylase; amylase; amylase-alpha;
KW   regulatory region; signal peptide.
XX
OS   Bacillus subtilis
OC   Eubacteria; Firmicutes; Clostridium group
OS   firmicutes; Bacillaceae; Bacillus.
```

Fields RN, RX, RA and RT

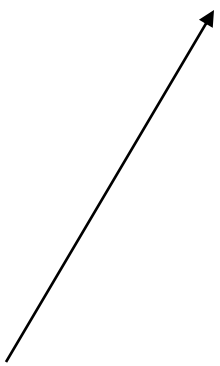
Bibliographic references.

```
RN   [1]
RP   1-2680
RX   MEDLINE; 83143299.
RA   Yang M., Galizzi, A., Henner, D.J.;
RT   "Nucleotide sequence of the amylase gene from
RT   Bacillus subtilis";
RL   Nucleic Acids Res. 11:237-249(1983).
…
```

Fiels FT: FEATURE TABLE

Description of functional regions.

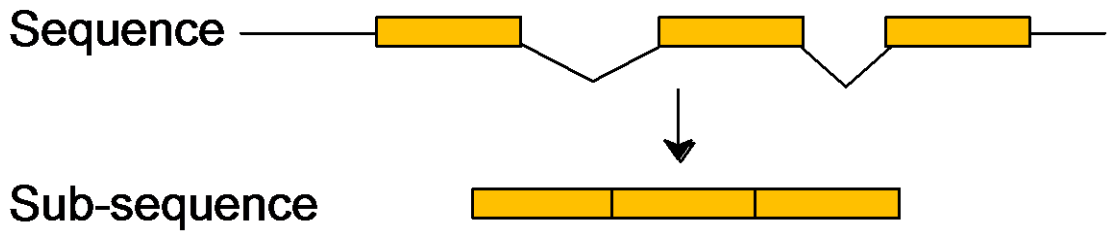
```
FT   promoter  369..374
FT             /note="promoter sequence P2 [3] (amyR1)"
FT   mutation  381..381
FT             /note="g is a gra-5 and gra-10 mutation [3]"
FT   RBS       414..419
FT             /note="rRNA-binding site rbs-1 [3]"
FT   CDS       498..2480
FT             /gene="amyE"
FT             /db_xref="SWISS-PROT:P00691"
FT             /product="alpha-amylase precursor"
FT             /EC_number="3.2.1.1"
FT             /translation="MFAKRFKTSLLPLFAGFLLLFHLVLAGPAA
FT             ASAETANKSNELTAPSIKSGTILHAWNWSFNTLKHNMKDIHDAG
...
```



Cross-references

Field FT

"join" operator




```
FT   CDS             join(242..610,3397..3542,5100..5351)
FT                   /codon_start=1
FT                   /db_xref="SWISS-PROT:P01308"
FT                   /note="precursor"
FT                   /gene="INS"
FT                   /product="insulin"
...
```

Field SQ

```
SQ   Sequence 2680 BP; 825 A; 520 C; 642 G; 693 T; 0 other;
     gctcatgccg agaatagaca ccaaagaaga actgtaaaaa cgggtgaagc agcagcgaat        60
     agaatcaatt gcttgcgcct ttgcggtagt ggtgcttacg atgtacgaca gggggattcc       120
     ccatacattc ttcgcttggc tgaaaatgat tcttcttttt atcgtctgcg gcggcgttct       180
     gtttctgctt cggtatgtga ttgtgaagct ggcttacaga agagcggtaa aagaagaaat       240
     (...)
     gatggtttct tttttgttca taaatcagac aaaacttttc tcttgcaaaa gtttgtgaag      2580
     tgttgcacaa tataaatgtg aaatacttca caaacaaaaa gacatcaaag agaaacatac      2640
     cctgcaagga tgctgatatt gtctgcattt gcgccggagc                            2680
//
```

Errors in sequence databases

- There are many errors in general sequence databases (notably for DNA databases) :
  - Annotations errors.
  - Sequence errors :
    - Sequencing errors (compression, etc.)
    - Contamination with cloning vector
    - Contamination with foreign DNA
    - Etc.

Redundance

- Major problem for DNA sequence databases.

Variations in sequences

- Redundant sequences are often not totally identical.
- It is impossible to determine whether the observed differences between two nearly-identical sequences are due to :
  - Polymorphism.
  - Sequencing errors.
  - Gene duplication
- GenBank: 20% of redundance among vertebrate protein-coding genes; 35-40% of redundance among human genomic sequences

SWISS-PROT and its complement TrEMBL

- Collaboration between the *Swiss Institute of Bioinformatics* (SIB) and the *European Bioinformatics Institute* (EBI).
- SwissProt:
  - Manual expertise of protein sequences: very rich annotations (protein function, subcellular localization, post-translational modification, structure, …)
  - Minimal redundance
  - Incomplete
- TrEMBL: translation of protein-coding sequences described in EMBL and not in SwissProt
  - Automatic annotation: annotations moins riches
- SwissProt+TrEMBL: complete data set, minimal redundance

Specialized sequence databases ...

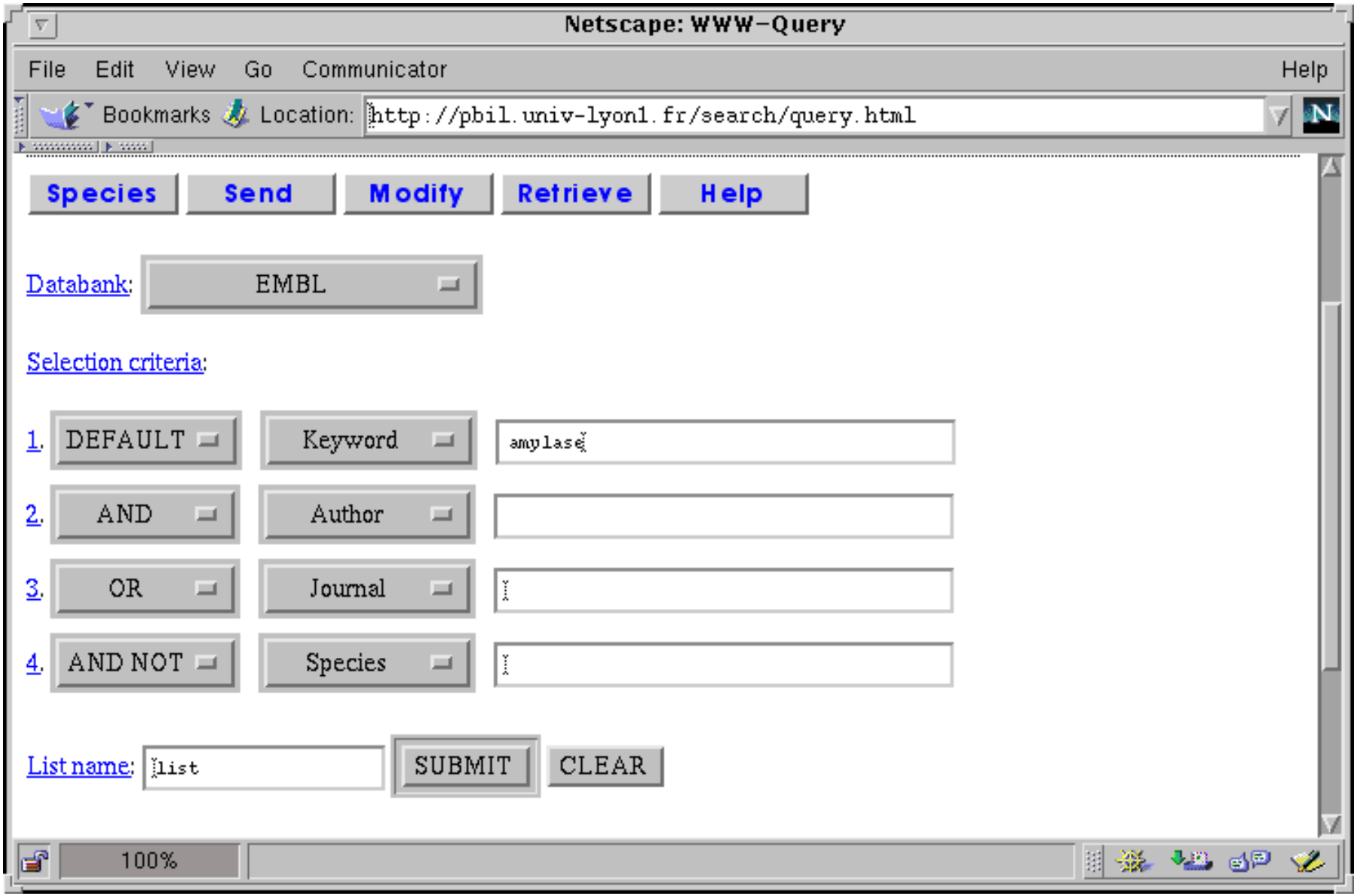
- *PROSITE, PFAM, PRODOM, PRINTS, INTERPRO* : databases of protein motifs
- *Protein Data Bank* (PDB) 3D structures of sequences (proteins, DNA, RNA)
- *Ribosomal Database Project* (RDP) : data on rRNAs
- Species-specific databases:
  - Human: OMIM: phenotypes, genetic diseases, mutations
  - Bacteria (ECD, NRSub, MycDB, EMGLib).
  - Yest (LISTA, SGD, YPD).
  - Nematode (ACeDB).
  - Drosophila (FlyBase).
  - …
- And many others … see dbCAT:
    - http://www.infobiogen.fr/services/dbcat/

Sequence retrieval in databases

- Selection of database entries according to :
  - Name or accession numbers of sequences.
  - Bibliographic references (author, article, …).
  - Keyword.
  - Taxonomy (species, gender, order, …).
  - Publication date
  - Organelle (mitochodria, chloroplaste, nucleus), host ...
  - …
- Access to functional regions described in the feature table:
  - Coding regions (CDS), tRNA, rRNA, ...

Database query software

- ACNUC/Query : http://pbil.univ-lyon1.fr/
  - Access to databases in GenBank, EMBL, SWISS-PROT or PIR formats.
  - Complex queries
  - Easy selection and extraction of subsequences (e.g. CDS, tRNAs, rRNAs, …)
- SRS (sequence retrieval system) http://srs.ebi.ac.uk/
  - 90 databases available through SRS.
  - multi-database queries.
- Entrez http://ncbi.nlm.nih.gov/
  - Access to NCBI databases: GenBank, GenPept, NRL_3D, MEDLINE.
  - Search by neighboring: sequences, bibliographic references


Netscape: WWW-Query
File Edit View Go Communicator Help
Bookmarks Location: http://pbil.univ-lyon1.fr/search/query.html
Species Send Modify Retrieve Help
Databank: EMBL
Selection criteria:
1. DEFAULT Keyword amylase
2. AND Author
3. OR Journal
4. AND NOT Species
List name: list SUBMIT CLEAR
100%