

# Initiation à la Génomique



**Daniel Gautheret**  
**ESIL, Université de la Méditerranée**

# Définition

La génomique est l'étude des génomes, de leur organisation et de leur évolution, ainsi que de l'expression et de la fonction des gènes

# En quoi consiste la génomique?

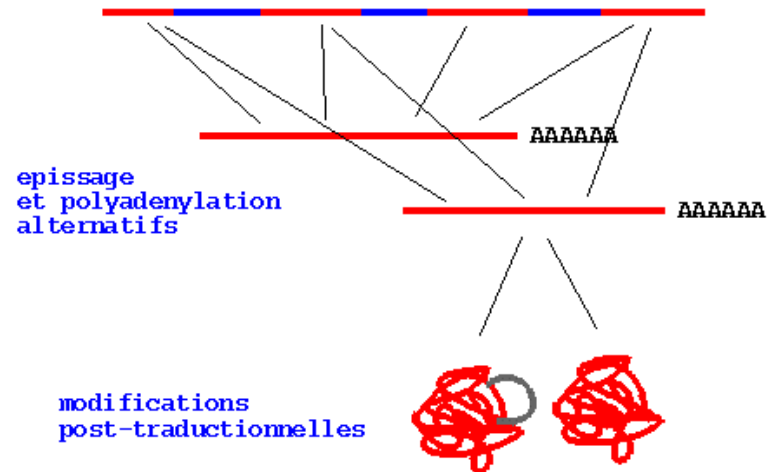
- ✦ Cartographier les génomes
  - ✦ Séquencer les génomes
  - ✦ Déterminer la structure des gènes (promoteur, ARNm, introns/exons, ORF, produit)
  - ✦ Déterminer les fonctions des gènes (activité, domaines, partenaires)
  - ✦ Etablir les profils d'expression des gènes (tissus, pathologies)
- (Dans les deux derniers cas, on parle de génomique fonctionnelle)

# Les omes et les omiques

« XXXome » = ensemble des XXX dans le génome

En raison des modifications posttranscriptionnelles, le transcriptome et le protéome ne découlent pas du génome de façon évidente.

- ★ Génome
- ★ Transcriptome
- ★ Protéome
  
- ★ Puis Interactome, etc.



Disciplines classiques abordées à l'échelle génomique

- ★ Génomique structurale
- ★ Génomique fonctionnelle
- ★ ARNomique

# Les applications de la génomique

## L'industrie utilise la génomique pour identifier

- ★ Les gènes impliqués dans les pathologies: cibles pharmaceutiques ou marqueurs diagnostiques
- ★ De nouveaux gènes permettant de synthétiser des molécules d'intérêt
- ★ Des gènes responsables de résistances (microbiologie ou agronomie)
- ★ Des gènes permettant de mieux comprendre des mécanismes-clés: cancer, vieillissement, etc.

## La Génomique, ce n'est pas que pour les maladies génétiques

### Toute pathologie implique à un moment ou à un autre les gènes et leur expression

- Prédisposition transmise (maladies héréditaires comme la dystrophie musculaire, certains cancers), MAIS AUSSI:
- Problèmes de réparation / mutation somatique (les modifications permettant le développement et le maintien des cancers ont leur origine dans l'ADN)
- Problèmes d'expression (tous les dysfonctionnements se traduisent tôt ou tard par une modification des profils d'expression des gènes.)

### Même dans le cas des maladies infectieuses:

- Gènes microbiens impliqués dans la pathogénicité (par ex. gouvernant la spécificité d'infection)
- Gènes impliqués dans les résistances aux antibiotiques

# Génomomes modèles

## Observation:

- ★ Plus de 70 gènes humains complètent des mutations chez la levure (1995)
- ★ La levure peut donc servir de modèle pour étudier la fonction de ces gènes chez l'homme. Par exemple: les gènes de cycle cellulaire.
- ★ Il existe des modèles pour toutes les grandes questions

On ne séquence pas des génomes-modèles parce qu'ils sont plus courts, mais:

- ★ Parce que les fonctions d'intérêt sont présentes dans ces espèces
- ★ Parce que les espèces-modèles sont beaucoup faciles à étudier expérimentalement (interférence ARN, KO, KI, etc.)

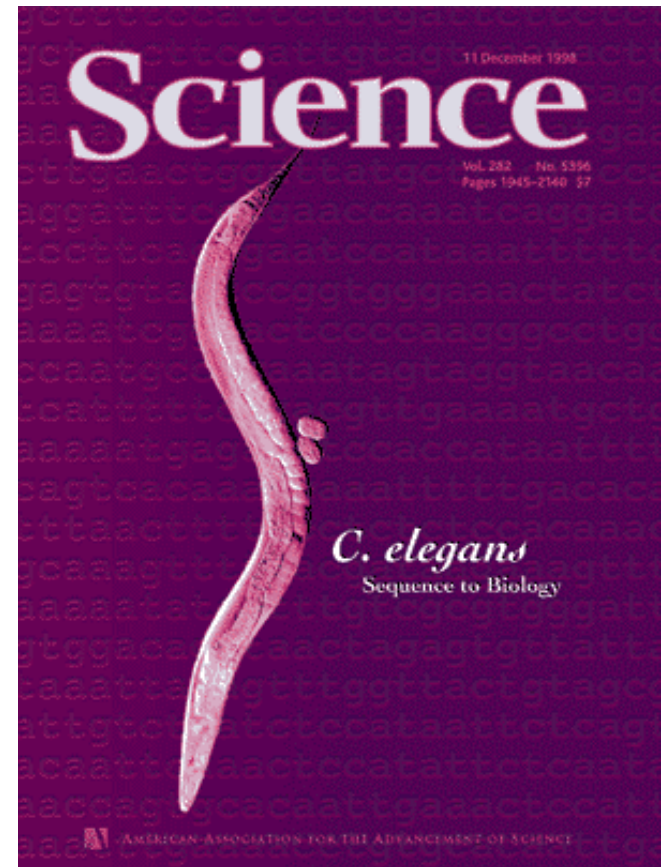
## Exemple

### Humain-drosophile (Banfi et al. Nature Genetics, 1996)

1. Collection des gènes causant des phénotypes mutants chez la drosophile
  2. Recherche de similitude dans tous les gènes (et cDNA) humains connus
  3. 66 cDNA humains montrent une similarité significative.
- Donc 66 cDNA candidats pour des maladies génétiques
  - Toutes les connaissances structurales et fonctionnelles acquises chez la drosophile sur ces gènes deviennent applicables.

# Un génome modèle: *C. elegans*

- ★ Un ver de 1mm de long.
- ★ pharynx, uterus, oocytes, oviducte, intestin, ovaire et même un comportement social en *959 cellules*.
- ★ Descendance: environ 300
- ★ 100.000.000 bp, 18.000 gènes
- ★ Génome disponible depuis 1999: le premier génome animal complet.
- ★ Pour chaque gène humain d'intérêt, l'homologue dans *C. elegans* peut être identifié en quelques minutes. Trouver sa fonction dans le ver est ensuite relativement facile par K.O. (Knock Out).



# Les programmes « Génome »



# Les programmes Génome

## Préhistoire

- ★ Séquençage de Sanger
  - ★ Avant le séquençage d'organismes: séquençage de bactériophages.
  - ★ Plus grande séquence obtenue avant projets « Génome »: bactériophage 1: 50.000pb
  - ★ Comment passer à plusieurs millions de bp?
- 
- ★ Bactérie: 0,5 – 8 Mb
  - ★ Eucaryote unicellulaire: à partir de 3Mb
  - ★ Mammifère: 3000 Mb

# Taille de quelques génomes

| Organisme                       | Nb. chrom. | Nbre gènes | Taille Mb |
|---------------------------------|------------|------------|-----------|
| <i>Amoeba dubia</i>             | 23         |            | 670 000   |
| Fougère                         | 23         |            | 160 000   |
| <i>Homo sapiens</i>             | 23         |            | 3000      |
| <i>Mus musculus</i>             | 21         |            | 3000      |
| Riz                             | 5          |            | 400       |
| <i>D. melanogaster</i>          | 4          |            | 165       |
| <i>Arabidopsis thaliana</i>     | 5          |            | 120       |
| <i>C. elegans</i>               | 6          |            | 100       |
| <i>Saccharomyces cerevisiae</i> | 16         |            | 13        |
| <i>Escherichia coli</i>         | 1          |            | 4,6       |
| <i>Encephalitozoon cuniculi</i> | 1          |            | 2,9       |
| <i>Mycoplasma genitalium</i>    | 1          |            | 0,6       |

# Programme Génome

## Le programme Génome tel que défini par le DOE en 1990

- ★ Carte génétique, résolution 2 à 5 cM (4 à 10 Mb chez l'homme)
- ★ Carte physique, résolution 100 kb
- ★ A long terme, séquençage du génome humain (après amélioration des techniques)
- ★ Séquençage du génome d'espèces modèles
- ★ Ajouté en 93: identification du plus grand nombre de gènes possible (par séquençage de cDNA)

# Cartographie génétique (cartes de liaison)

- ★ La fréquence de crossing over entre 2 marqueurs indique la distance qui les sépare. Les marqueurs doivent être polymorphes pour être suivis.
- ★ Les marqueurs moléculaires (phénotype visible sur gel) ont fait nettement progresser cette technique.

## RFLP (restriction fragment length polymorphism)

- ★ Des fragments d'ADN génomiques qui, hybridés à un ADN génomique digéré par une enzyme de restriction, donnent des bandes de taille différente selon les individus.
- ★ Les marqueurs sont assignés aux chromosomes par FISH (Fluorescent in situ hybridization). Résolution: sous-bande chromosomique (env. 10 Mb)

## Microsatellites

- ★ Des séquences répétées avec fort taux de polymorphisme: le nombre de répétitions varie fortement d'un individu à l'autre.
- ★ Une PCR suffit pour identifier quel allèle est présent chez un individu: deux amorces autour du microsatellite, amplification puis résolution sur gel. On observe soit deux bandes pour les allèles maternel et paternel, soit une seule bande si même allèle.

## STS (Sequence tagged Sites)

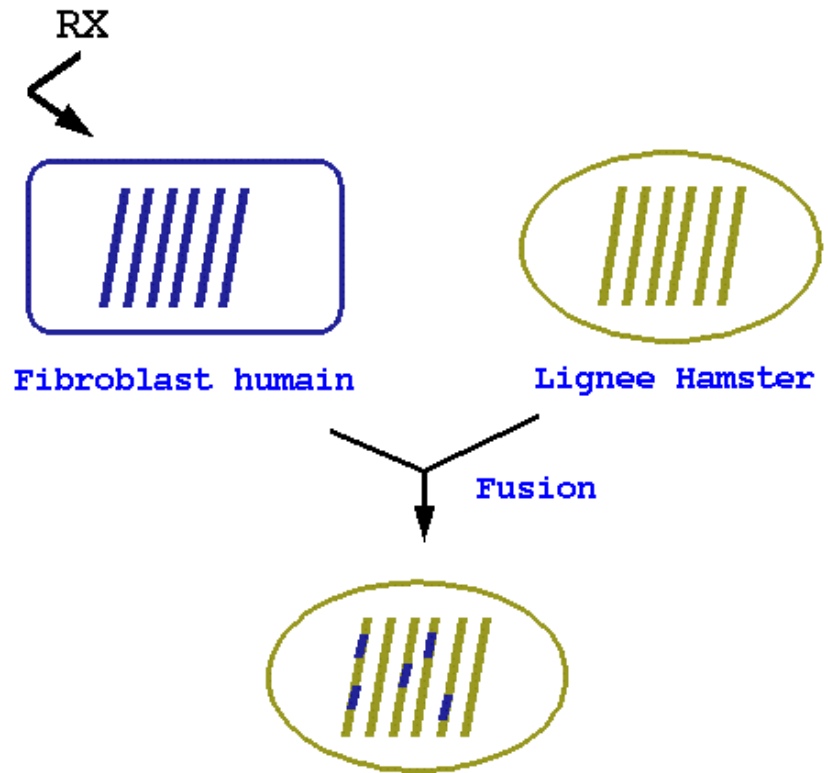
- ★ Séquence génomique unique de 200-500bp pouvant être amplifiée à partir d'amorces PCR connues.

# Cartographie génétique

## Hybrides d'irradiation

★ On recherche ce que deviennent des marqueurs: retrouver 2 marqueurs dans le même hybride indique leur proximité. Par exemple, utilisation des marqueurs STS: le marqueur STS s1 est retrouvé dans les hybrides hi, hj, hk, le marqueur s2 est retrouvé dans les hybrides hl, hm, etc. De ces données, on déduit une proximité ou non des marqueurs s1 et s2.

★ Résolution: 1-2Mb



## Carte Généthon 1996

★ 5264 microsatellites, résolution 1,6 cM (~3 Mb)

# Cartographie Physique

- ★ BUT: rendre disponible toute région voulue d'un génome sous forme d'un fragment d'ADN cloné.
  1. Création d'une banque d'ADN du génome étudié
  2. Identifier chaque clone
  3. Ordonner/positionner clones sur le chromosome

## Banques Génomiques

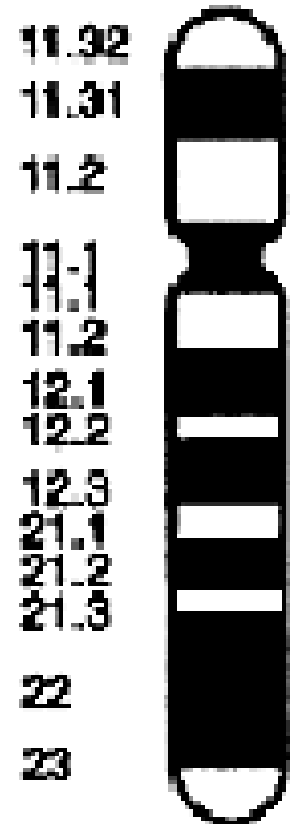
- ★ Phages, Cosmides: inserts de 15-20kb: pour petits génomes
- ★ YAC (Yeast Artificial Chromosome): inserts de 500kb-1Mb. Grands génomes.
- ★ Une banque génomique humaine se présente sous la forme d'une collection de plusieurs centaines de microplaques à 96 puits, chacun contenant un clone.

## Ordonnement des clones

- ★ Si l'on repère un marqueur de carte génétique dans un clone (par exemple par hybridation), ce clone se trouve directement positionné.
- ★ On peut également utiliser le recouvrement entre clones pour les positionner les uns par rapport aux autres.
- ★ On peut aussi utiliser les STS (Sequence Tagged Sites): des séquences quelconques uniques dans le génome. Le fait de retrouver le même STS dans 2 clones indique qu'ils se chevauchent.

# Notation et Carte Cytogénétique

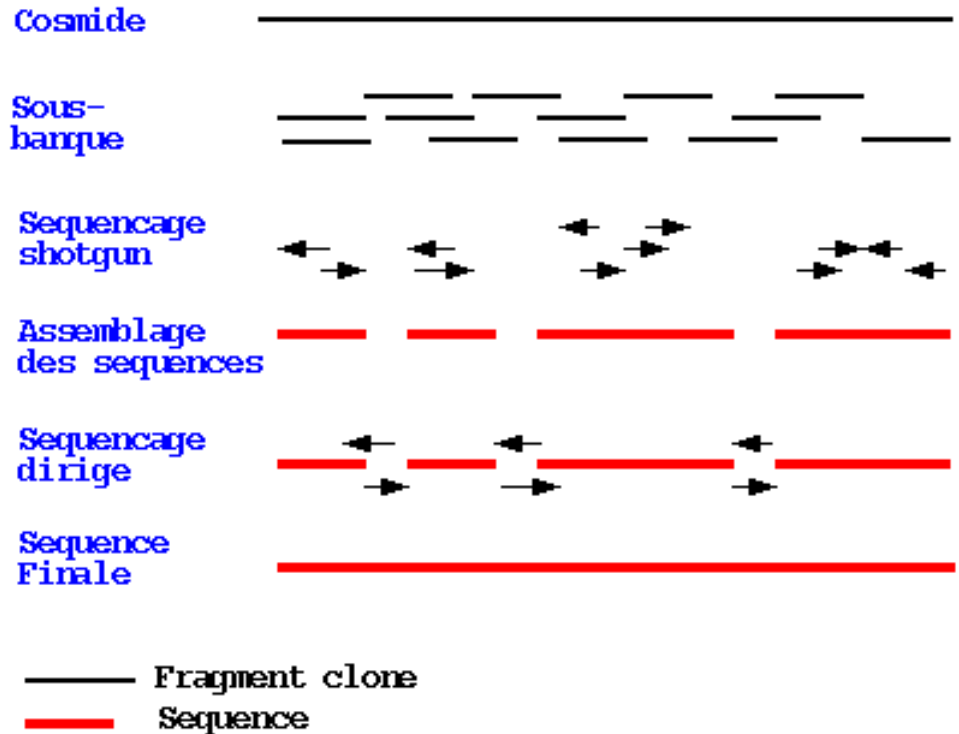
- ★ Chaque chromosome humain possède un **bras court** ("p" pour "petit") et un **bras long** ("q" pour "queue"), séparés par un **centromère**. Les extrémités du chromosome sont appelées **télomères**.
- ★ Chaque bras chromosomique est divisé en régions ou **bandes cytogénétiques** visibles au microscope à l'aide de certaines colorations. Les bandes sont nommées p1, p2, p3, q1, q2, q3, etc., en allant du centromère vers le télomère. A plus haute résolution, des sous-bandes apparaissent à l'intérieur des bandes. Les sous-bandes sont numérotés dans le même ordre.
- ★ Par exemple, la localisation du gène CFTR sur la carte cytogénétique est 7q31.2, c'est à dire:
- ★ Chromosome 7, bras q, bande 3, sous-bande 1, et sous-sous-bande 2. Les extrémités du chromosome sont nommées p<sub>tel</sub> et q<sub>tel</sub>. La notation 7q<sub>tel</sub> indique la fin du bras long du chromosome 7.



# Séquençage Shotgun

## Principe général

- ★ Cosmide ou BAC sous-cloné en fragments de petite taille
- ★ Des clones prélevés aléatoirement sont séquencés
- ★ Des séquences contigües sont reconstruites par recouvrement
- ★ Les séquences restantes sont réalisées de façon dirigée.
- ★ On voit tout de suite que cette approche est problématique en présence de longues régions répétées.



D'après A. Bernot: L'Analyse des Genomes, Nathan Université (1996)

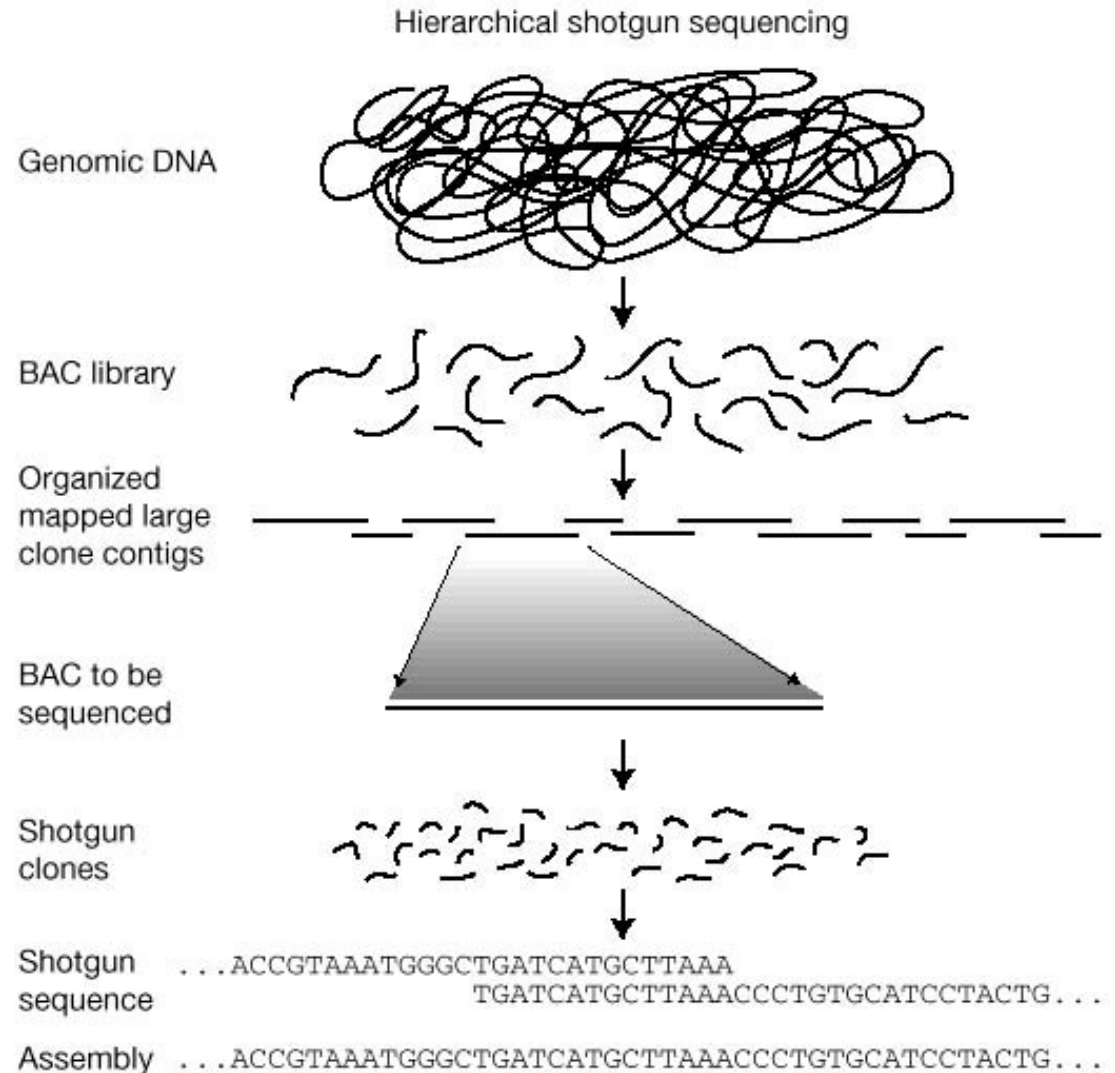


# Shotgun « génome complet »

- ★ le shotgun "genome complet" (whole-genome shotgun) s'applique normalement aux génomes relativement simples, en exploitant au maximum les informations de cartographie et la bioinformatique pour éviter les misassemblages.
- ★ Employé par Celera pour le séquençage du génome de la Drosophile, puis du génome humain

# Shotgun hierarchique

- ★ Le shotgun hierarchique ou "clone par clone" implique de générer un jeu de clones de grande taille (100-200kb) couvrant le génome puis de soumettre au shotgun uniquement des clones bien choisis. Ceci élimine les risques d'erreur "longue distance".



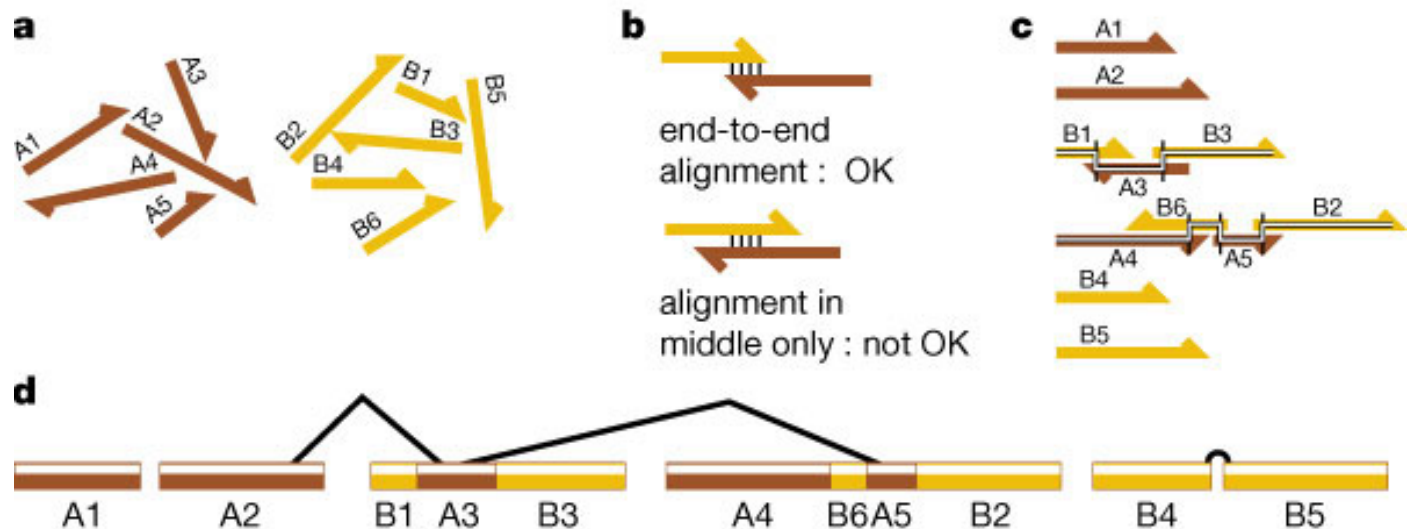
# Shotgun hierarchique

- ★ Dans le projet de séquençage du génome humain par le consortium international, les BAC à séquencer ont été choisis en analysant les profils de restriction (fingerprint) de l'équivalent d'une couverture 20X en BAC, combiné à différents marqueurs physiques.
- ★ Extrait de l'article de Nature sur le Génome humain:

A version of the draft genome sequence was prepared on the basis of the map and sequence data available on 7 October 2000. For this version, the mapping effort had assembled the fingerprinted BACs into 1,246 fingerprint clone contigs. The sequencing effort had sequenced and assembled 29,298 overlapping BACs and other large-insert clones, comprising a total length of 4.26 gigabases (Gb). This resulted from around 23 Gb of underlying raw shotgun sequence data, or about 7.5-fold coverage averaged across the genome (including both draft and finished sequence).

# Assemblage des clones en contigs

- ★ On doit assembler les clones pour reconstituer le chromosome.
- ★ Ceci commence par un assignement des clones au bon chromosome, à l'aide des cartes physiques, STS etc. (présence de marqueurs dans le clone et dans le chromosome).
- ★ Puis on recherche des recouvrements dans les séquences des clones. Le recouvrement peut comprendre plusieurs séquences initiales avec gaps, donc pas forcément facile (voir figure ci-dessous).
- ★ L'assemblage final est appelé un "squelette de contigs" (contig scaffold).



# Etat des programmes (2004)

- ★ 168 génomes bactériens (18 archae et 150 bactéries)
  - 1er génome: haemophilus influenza (1995)
- ★ Génomes eucaryotes:
  - Levure
  - Cenorhabditis (ver plat)
  - Drosophile
  - Arabidopsis Thaliana, riz
  - Humain, Souris
  - Fugu, Tetraodon
  - Cione (cordé ancestral)
- ★ millions de transcrits (pleine longueur et EST)

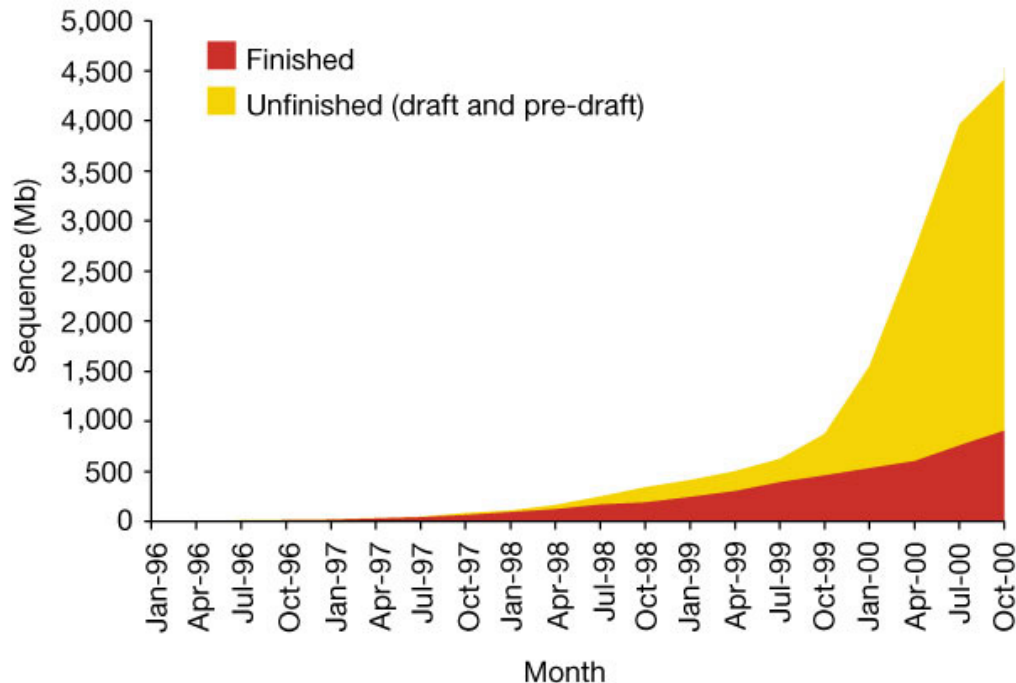
# Génomes bactériens... (liste de [www.tigr.org](http://www.tigr.org))

## Published complete microbial genomes (listed alphabetically)

| Link | Genome                                 | Strain  | Domain            | Size (Mb) | Institution   | Funding                            | Publication  |
|------|--|---------|-------------------|-----------|---|------------------------------------|--|
|      | <i>Aeropyrum pernix</i>                | K1      | <a href="#">A</a> | 1.67      | <a href="#">Biotechnology Center</a>                            | <a href="#">NITE</a>               | <a href="#">Kawarabayasi et al., DNA Research 6: 83-101 (1999)</a>   |
|      | <i>Agrobacterium tumefaciens</i>       | C58     | <a href="#">B</a> | 5.3       | <a href="#">University of Washington Genome Center / Cereon</a> | <a href="#">NSF / Cereon</a>       | <a href="#">Wood et al., Science 294:2317-2323 (2001) / Goodner et al., Science 294:2323-2328 (2001)</a>       |
|      | <i>Aquifex aeolicus</i>                | VF5     | <a href="#">B</a> | 1.50      | <a href="#">Diversa</a>   | <a href="#">DOE, Diversa</a>       | <a href="#">Deckert et al., Nature 392:353 (1998)</a>  |
|      | <a href="#">Archaeoglobus fulgidus</a> | DSM4304 | <a href="#">A</a> | 2.18      | <a href="#">TIGR</a>  | <a href="#">DOE</a>                | <a href="#">Klenk et al., Nature 390:364-370 (1997)</a>  |
|      | <i>Bacillus halodurans</i>             | C-125   | <a href="#">B</a> | 4.2       | <a href="#">Japan Marine Science and Technology Center</a>      |                                    | <a href="#">Takami et al., Nuc. Acid Res. 28: 4317-4331 (2000)</a>   |
|      | <i>Bacillus subtilis</i>               | 168     | <a href="#">B</a> | 4.20      | <a href="#">International Consortium</a>                        | <a href="#">EC</a>                 | <a href="#">Kunst et al., Nature 390: 249-256 (1997)</a>   |
|      | <a href="#">Borrelia burgdorferi</a>   | B31     | <a href="#">B</a> | 1.44      | <a href="#">TIGR</a>  | <a href="#">Mathers Foundation</a> | <a href="#">Fraser et al., Nature, 390: 580-586 (1997) / Casjens et al., Mol Microbiol, 35: 490-516 (2000)</a> |
|      | <i>Buchnera sp.</i>                    | APS     | <a href="#">B</a> | 0.64      | Univ. Tokyo / RIKEN   |                                    | <a href="#">Shigenobu et al., Nature 407: 81-86 (2000)</a>   |

# Articles Génome humain 2001

- ★ Versions publique et Celera terminée à 90% début 2001.
- ★ Séquences Celera disponible avec restriction
- ★ Séquence publique disponible dans Genbank
- ★ Le bond après juillet 99 correspond à l'arrivée d'une nouvelle génération de séquenceurs (à capillaire)



[Voir l'article complet](#)  
(accès Internet requis)

# Articles Génome humain 2001

- ★ Draft (brouillon): Pour atteindre le statut de brouillon, un clone doit avoir été séquencé au moins avec une couverture de 3. Ceci revient généralement à une couverture effective de 96%, avec des gaps de 500bp en moyenne. Au dessous d'une couverture 3, on parle de pré-draft.

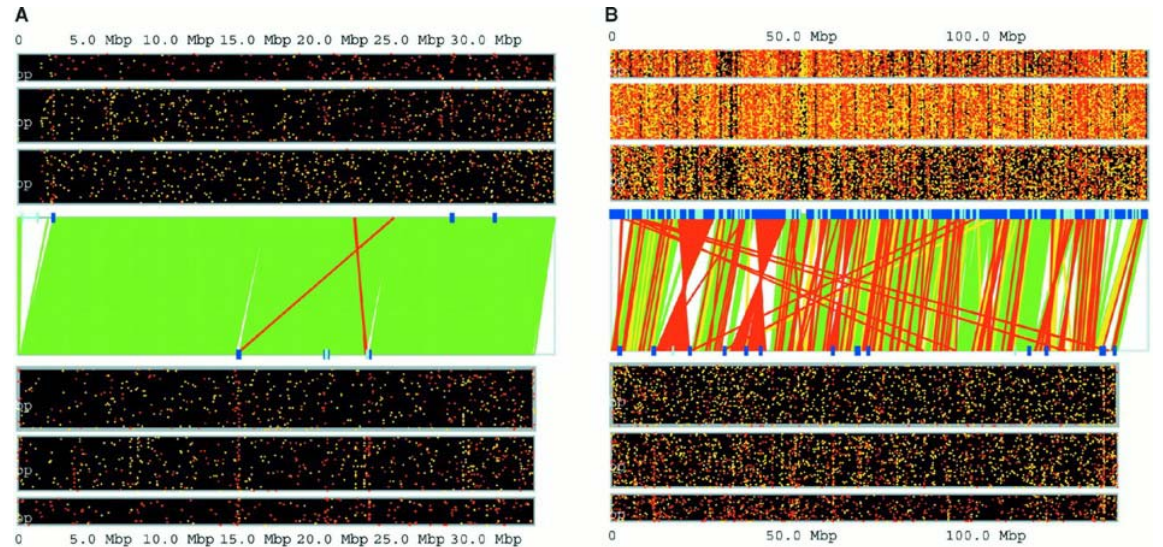
**Table 7 Sequence level contiguity of the draft genome sequence**

| Chromosome | Initial sequence contigs |                 | Sequence contigs |                 | Sequence-contig scaffolds |                 |
|------------|--------------------------|-----------------|------------------|-----------------|---------------------------|-----------------|
|            | Number                   | N50 length (kb) | Number           | N50 length (kb) | Number                    | N50 length (kb) |
| All        | 396,913                  | 21.7            | 149,821          | 81.9            | 87,757                    | 274.3           |
| 1          | 37,656                   | 16.5            | 12,256           | 59.1            | 5,457                     | 278.4           |
| 2          | 32,280                   | 19.9            | 13,228           | 57.3            | 6,959                     | 248.5           |
| 3          | 38,848                   | 15.6            | 15,098           | 37.7            | 8,964                     | 167.4           |
| 4          | 28,600                   | 16.0            | 13,152           | 33.0            | 7,402                     | 158.9           |
| 5          | 30,096                   | 20.4            | 10,689           | 72.9            | 6,378                     | 241.2           |
| 6          | 17,472                   | 43.6            | 5,547            | 180.3           | 2,554                     | 485.0           |
| 7          | 12,733                   | 86.4            | 4,562            | 335.7           | 2,726                     | 591.3           |
| 8          | 19,042                   | 18.1            | 6,984            | 38.2            | 4,631                     | 198.9           |
| 9          | 15,955                   | 20.1            | 6,226            | 55.6            | 3,766                     | 216.2           |
| 10         | 21,762                   | 18.7            | 9,126            | 47.9            | 6,886                     | 133.0           |
| 11         | 29,723                   | 14.3            | 8,503            | 40.0            | 4,684                     | 193.2           |
| 12         | 22,050                   | 19.1            | 8,422            | 63.4            | 5,526                     | 217.0           |
| 13         | 13,737                   | 21.7            | 5,193            | 70.5            | 2,659                     | 300.1           |
| 14         | 4,470                    | 161.4           | 829              | 1,371.0         | 541                       | 2,009.5         |
| 15         | 13,134                   | 15.3            | 5,840            | 30.3            | 3,229                     | 149.7           |
| 16         | 10,297                   | 34.4            | 4,916            | 119.5           | 3,337                     | 356.3           |
| 17         | 10,369                   | 22.9            | 4,339            | 90.6            | 2,616                     | 248.9           |
| 18         | 16,266                   | 15.3            | 4,461            | 51.4            | 2,540                     | 216.1           |
| 19         | 6,009                    | 38.4            | 2,503            | 134.4           | 1,551                     | 375.5           |
| 20         | 2,884                    | 108.6           | 511              | 1,346.7         | 312                       | 813.8           |
| 21         | 103                      | 340.0           | 5                | 28,515.3        | 5                         | 28,515.3        |
| 22         | 526                      | 113.9           | 11               | 23,046.1        | 11                        | 23,046.1        |
| X          | 11,062                   | 58.8            | 4,607            | 218.6           | 2,610                     | 450.7           |
| Y          | 557                      | 154.3           | 140              | 1,388.6         | 106                       | 1,439.7         |
| UL         | 1,262                    | 21.4            | 613              | 46.0            | 297                       | 166.4           |

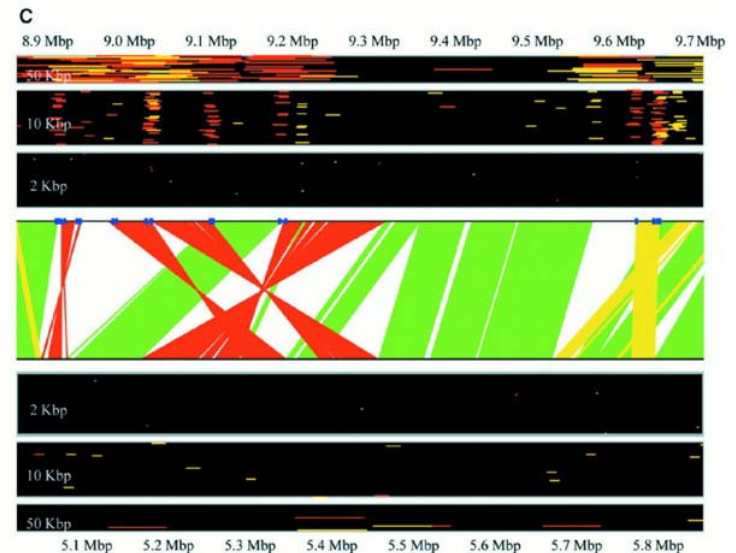
This Table is similar to Table 6 but shows the number and N50 length for various types of sequence contig (see Box 1). See legend to Table 6 concerning treatment of gaps. For sequence contigs in the draft genome sequence, the N50 length ranges from 1.7 to 5.5 times the arithmetic mean for initial sequence contigs, 2.5 to 8.2 times for merged sequence contigs, and 6.1 to 10 times for sequence-contig scaffolds.



# Comparison Celera / Consortium



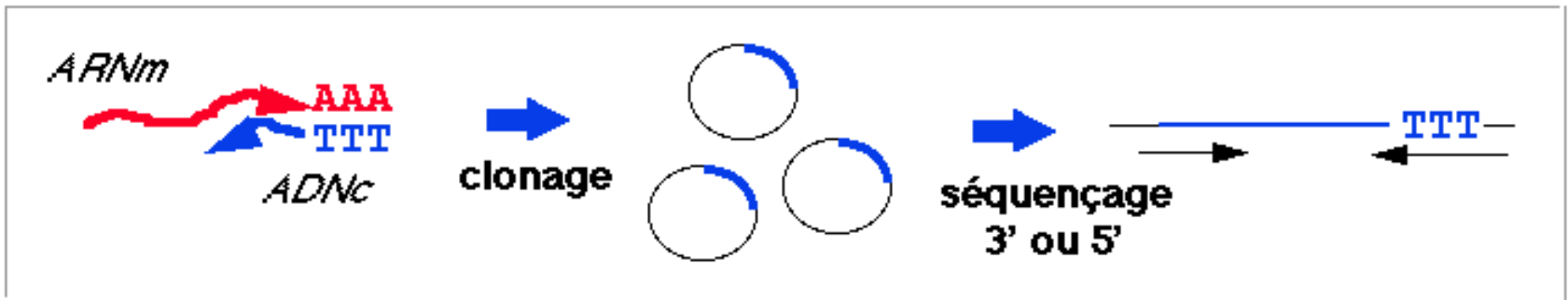
Comparison of the CSA and the PFP assembly. (A) All of chromosome 21, (B) all of chromosome 8, and (C) a 1-Mb region of chromosome 8 representing a single Celera scaffold. The PFP assembly is indicated in the upper third of each panel; the Celera assembly is indicated in the lower third. In the center of the panel, green lines show Celera sequences that are in the same order and orientation in both assemblies and form the longest consistently ordered run of sequences. Yellow lines indicate sequence blocks that are in the same orientation, but out of order. Red lines indicate sequence blocks that are not in the same orientation. For clarity, in the latter two cases, lines are only drawn between segments of matching sequence that are at least 50kb long. The top and bottom thirds of each panel show the extent of Celera mate-pair violations (red, misoriented; yellow, incorrect distance between the mates) for each assembly grouped by library size. (Mate pairs that are within the correct distance, as expected from the mean library insert size, are omitted from the figure for clarity.) Predicted breakpoints, corresponding to stacks of violated mate pairs of the same type, are shown as blue ticks on each assembly axis. Runs of more than 10,000 Ns are shown as cyan bars.



# Transcriptome

# Expressed Sequence Tags (ESTs)

- ★ Idée originale: pourquoi vouloir tout séquencer (95% de junk DNA) si ce sont les gènes qui nous intéressent?
- ★ EST = Séquences partielles d'ADNc clonés et prélevés aléatoirement.

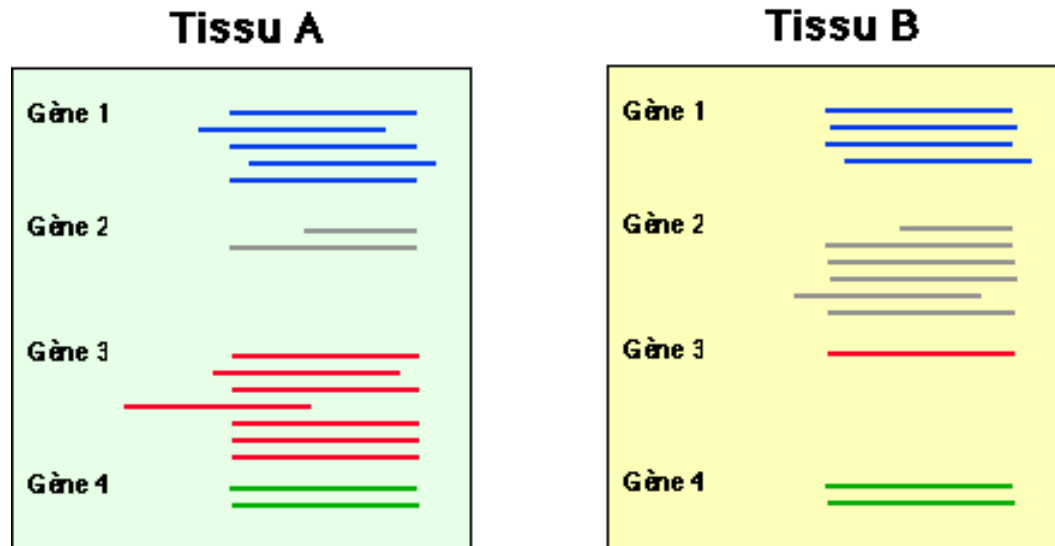


- ★ Grandes applications:
  - Cataloguage des gènes
  - Profils d'expression/Northern virtuels

# Utilisations des EST

## Profils d'expression

- Un test statistique (Test de Fisher) permet de déterminer si les différences de niveaux d'expression sont significatives.
- Le mode d'obtention des EST peut introduire des biais dans ces profils (par ex. banques normalisées: les EST les plus fréquents sont réduits).
- Vu le coût associé à la création de banques d'EST utilisables en analyse quantitative, on se penche sur d'autres techniques de mesure d'expression comme les puces à ADN.



# EST

## Normalisation

- ✦ Pour éviter de réamplifier sans cesse les transcrits les plus fréquents: réhybridation (normalisation) ou hybridation contre bibliothèque de référence (soustraction).
- ✦ **Incompatible avec utilisation pour profils d'expression**

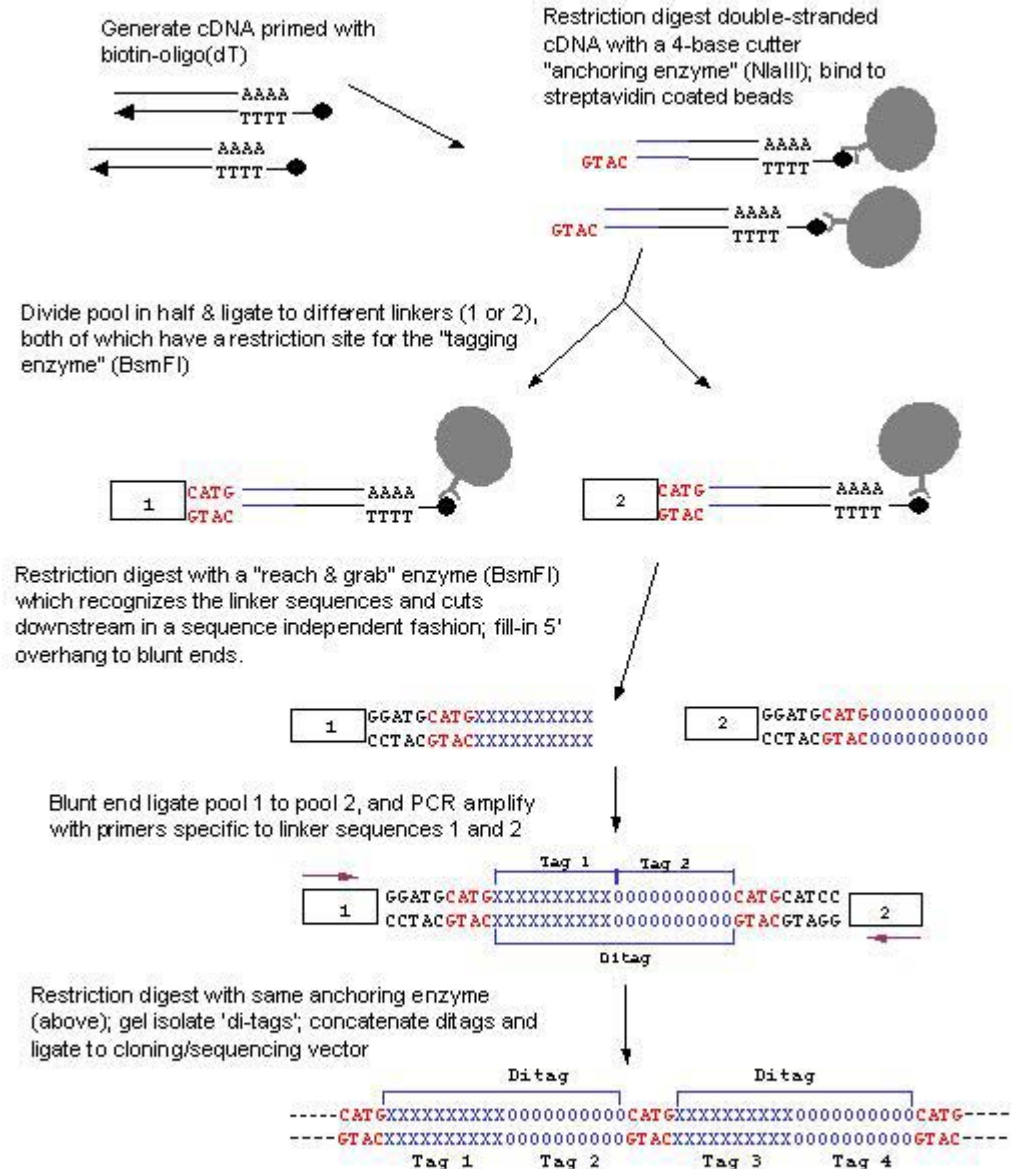
## Limitations de l'approche EST

- ✦ Chimérisme
- ✦ Clônes inversés
- ✦ Priming interne
- ✦ Introns
- ✦ Epissage alternatif
- ✦ Coût en ce qui concerne le cataloguage: 85%-95% des transcrits sont de faible abondance.

# SAGE (Serial Analysis of Gene Expression)

- ★ Génération d'ADNc double-brin via amorce poly(T) biotinylée.
  - ★ Digestion des ADNc par enzyme coupant en moyenne tous les 256 bases
  - ★ Partie 5' des ADNc récupérée par billes magnétiques couplées à une molécule de streptavidine
  - ★ Séparation en 2 lots et fixation d'un linker à l'extrémité des cDNA (linkers différents pour les 2 lots)
  - ★ Clivage par enzyme d'étiquetage, coupant 20 bases après l'extrémité (la taille du linker est ajustée de façon à conserver un bout d'ADNc de 10 bases)
  - ★ Ligation et amplification des fragments deux par deux (ditags) en utilisant les deux linkers comme amorce (assure que tous les fragments sont amplifiés dans la même façon).
  - ★ Les ditags sont séparés du linker puis concaténés.
  - ★ Les concaténats sont clonés et séquencés.
- 
- ★ Découverte de nouveaux gènes par SAGE: une étude SAGE du transcriptome de levure pendant 3 phases de croissance a permis de mettre en évidence 160 nouveaux gènes (Velculescu et al. 1997)

# SAGE



# Structure des génomes



# La composition en séquence et ses variations

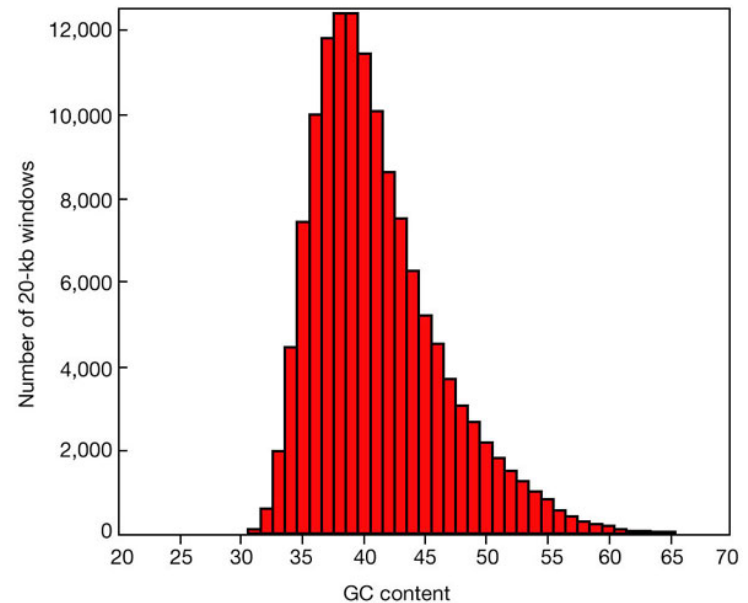
## La règle de Chargaff

- ★ Les règles d'appariement expliquent que, quelque soit la quantité d'adénine (A) dans l'ADN d'un organisme, la quantité de Thymine (T) est la même. De la même façon, G=C.
- ★ Mais les génomes ne sont pas uniformément constitués de 25%A, 25%T, 25%G, 25%C.
- ★ Il existe de grandes variations d'un génome à l'autre. Par exemple, les génomes d'archaebactéries sont généralement très riches en AT.

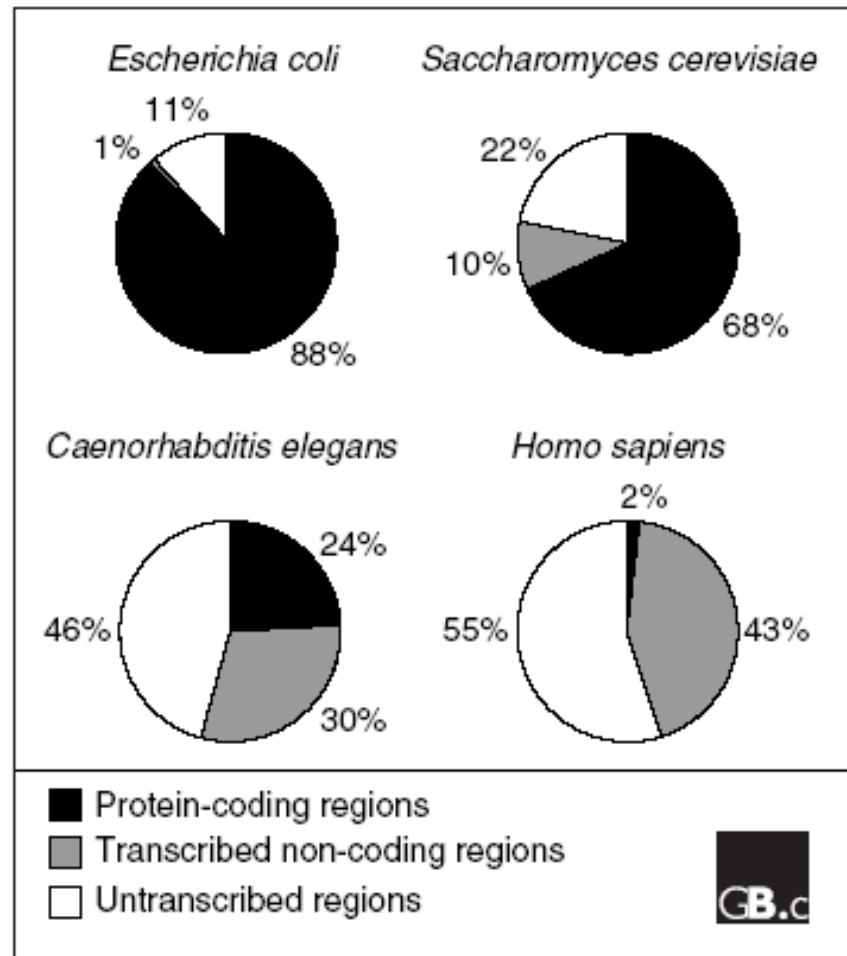
| <b>Organisme</b> | <b>A</b> | <b>T</b> | <b>G</b> | <b>C</b> |
|------------------|----------|----------|----------|----------|
| Humain           | 30.9     | 29.4     | 19.9     | 19.8     |
| Poulet           | 28.8     | 29.2     | 20.5     | 21.5     |
| Sauterelle       | 29.3     | 29.3     | 20.5     | 20.7     |
| Oursin           | 32.8     | 32.1     | 17.7     | 17.3     |
| Blé              | 27.3     | 27.1     | 22.7     | 22.8     |
| Levure           | 31.3     | 32.9     | 18.7     | 17.1     |
| E. coli          | 24.7     | 23.6     | 26.0     | 25.7     |
| Methanococcus    | 34.5     | 34.2     | 15.9     | 15.5     |

# Contenu en GC et isochores

- ★ Il existe également de grandes variations à l'intérieur d'un génome. Par exemple, les régions transcrites sont souvent plus riches en GC que les régions non-transcrites.
  - ★ Le contenu en GC moyen du génome humain est 41%
  - ★ On trouve pourtant des régions de plusieurs centaines de kb avec des contenus en GC de 33% ou 59%, ce qui représente une variation beaucoup plus grande que si la distribution était uniforme.
  - ★ La variation est en fait 15 fois supérieure à la variation attendue, avec une importante "queue" de régions riches en GC.
- 
- ★ On a proposé que le génome soit constitué d'une mosaïque de régions de composition homogène appelées **isochores**.
  - ★ En fait, on ne trouve pas de régions vraiment homogènes en composition, mais plutôt de régions plus ou moins riches en GC.
  - ★ Il existe une forte corrélation entre la richesse en GC et certaines propriétés: densité en gènes, contenu en répétitions, etc.
  - ★ Un tiers du génome contient 50% des gènes (gene-rich third)



# Codant et non codant...



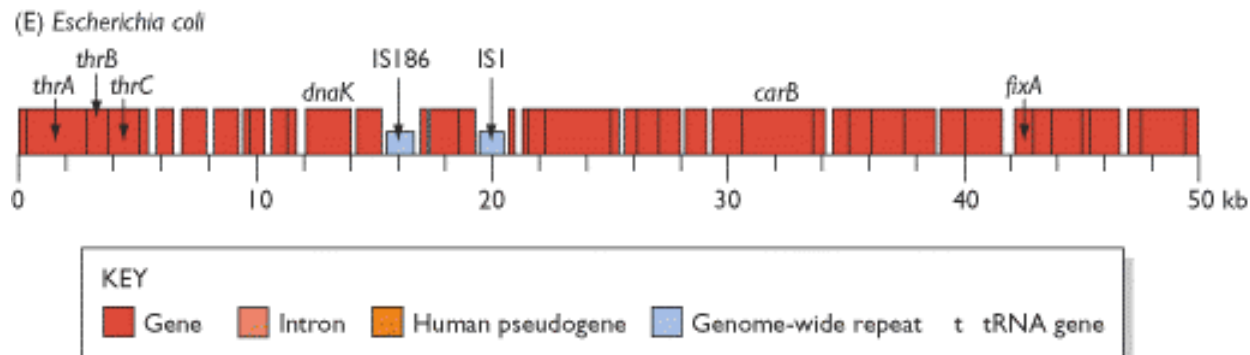
**Figure 1**

Ratios of the protein-coding, non-coding, and untranscribed sequences in bacterial, yeast, nematode and mammalian genomes. Estimations of the transcribed and protein-coding parts of genomes are based on the sequence length of annotated genes [3, 12, 13, 73]. Estimation of the transcribed portion of the human genome is based on the sequence length occupied by the annotated genes on chromosomes 6, 7, 14, 20, and 22 [5].

# Statistiques sur les gènes

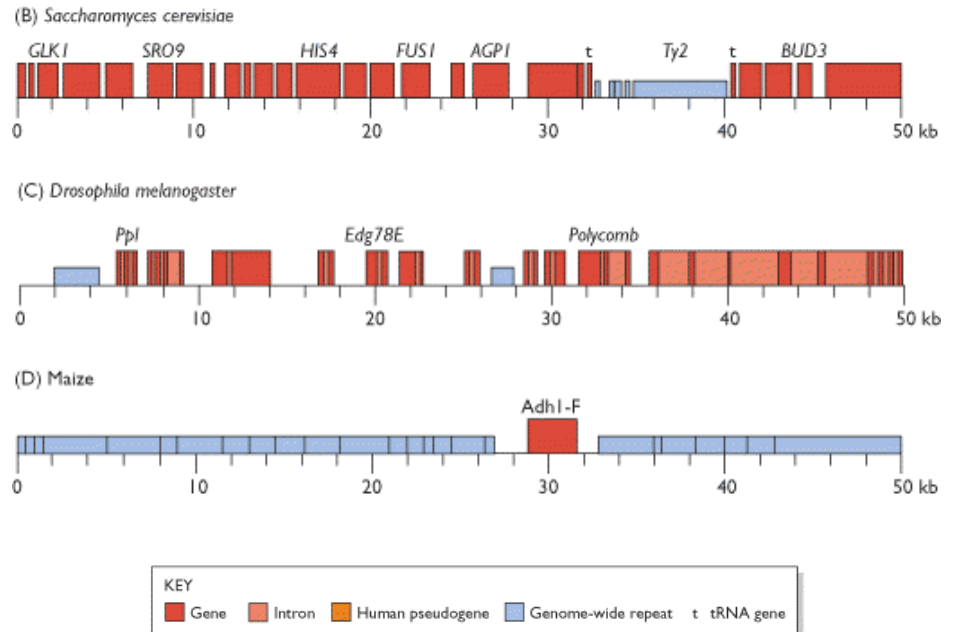
## Gènes procaryotes

- ★ longueur gène 950 nt. en moyenne (coli)
- ★ Densité en gènes. 95% du génome est transcrit chez E. coli.
- ★ Gènes organisés en opérons. 600 opérons dans le génome de Coli.



# Gènes eucaryotes

- ★ Gène humain moyen: 8,8 introns, 27 kb, 3'UTR:770bp, 5'UTR:300bp, CDS:1340bp, exon moyen: 145 bp (218 bp pour *C. elegans*), intron moyen:3365 bp (mais pic à 87 bp).
- ★ Gènes "monstres": dystrophine: 2,4 Mb; Facteur de coagulation VIII: 186 kb, 26 exons; Tinine: CDS de 80780 bp, 178 exons
- ★ Densité: humain: 1 gène tous les 100kb en moyenne; *C.elegans*: 1gène/5-6kb (25%); *S. Cerevisiae*: 1 gène/2kb.

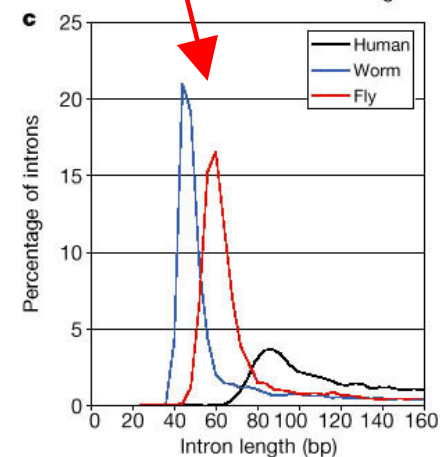
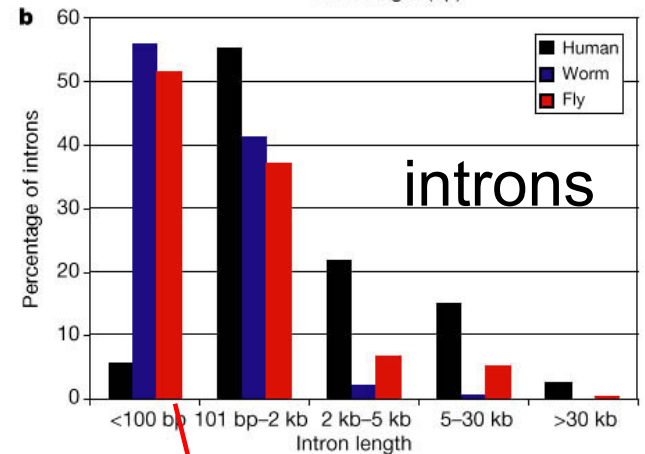
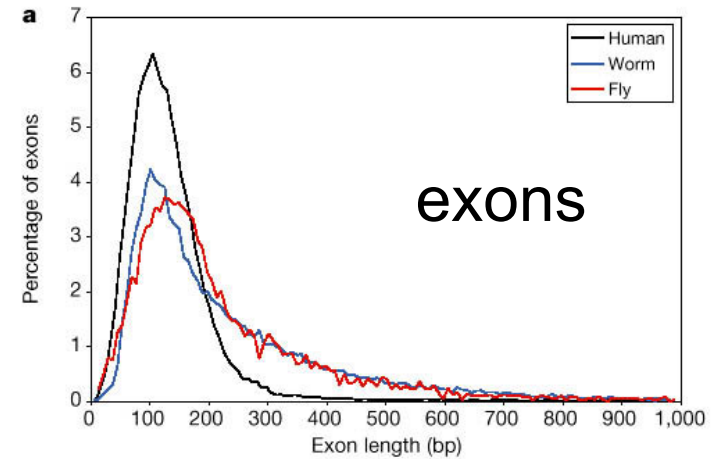


From « Genomes 2 », T.A. Brown

Remarque: 35000 gènes de 30kb en moyenne font 1Gb: donc au moins 1/3 du génome humain est transcrit. Par contre, seul 1,5% est codant!

# Gènes eucaryotes: les introns

- ★ intron humain moyen:3365 bp
- ★ Intron de ver moyen:267bp
- ★ Intron de mouche moyen:487bp
- ★ *S. cerevisiae*: 4% des gènes ont des introns, seulement 10 gènes sur 6200 en ont plus d'un.



Introns  
Courts  
(<100)

# Relation taille / nb gènes

| Organisme                       | Nb. chrom. | Nbre gènes | Taille Mb |
|---------------------------------|------------|------------|-----------|
| <i>Amoeba dubia</i>             | 23         | ??         | 670 000   |
| Fougère                         | 23         | ??         | 160 000   |
| <i>Homo sapiens</i>             | 23         | 30-40.000  | 3000      |
| <i>Mus musculus</i>             | 21         | 30-40.000  | 3000      |
| Riz                             | 5          | ??         | 400       |
| <i>D. melanogaster</i>          | 4          | 13600      | 165       |
| <i>Arabidopsis thaliana</i>     | 5          | 26000      | 120       |
| <i>C. elegans</i>               | 6          | 18.000     | 100       |
| <i>Saccharomyces cerevisiae</i> | 16         | 6000       | 13        |
| <i>Escherichia coli</i>         | 1          | 4000       | 4,6       |
| <i>Encephalitozoon cuniculi</i> | 1          | 2000       | 2,9       |
| <i>Mycoplasma genitalium</i>    | 1          | 400        | 0,6       |

Duplication  
de fragments/  
gènes

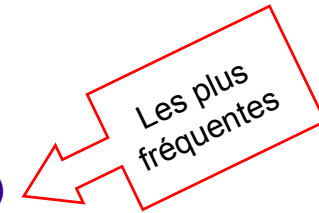
Séquences  
répétées

Expansion  
des introns

# Junk DNA: les séquences répétées dans le génome humain

## 5 classes de séquences répétées

- ★ Répétition de type transposon (ou interspersed repeats)
- ★ copie rétro-transposées inactives de gènes (protéines ou ARN) (processed pseudogenes)
- ★ Répétition simples de k-mères courts, p. ex. (A)<sub>n</sub>, (CA)<sub>n</sub> ou (CGG)<sub>n</sub>
- ★ Segments dupliqués: blocs de 10–300 kb copiés d'une région à l'autre
- ★ Blocs de séquences répétées en tandems (centromères, télomères, clusters de gènes ribosomiques)



Un ADN pas si "poubelle" que ça qui joue un grand rôle dans la transformation des gènes et l'apparition de nouveaux gènes.

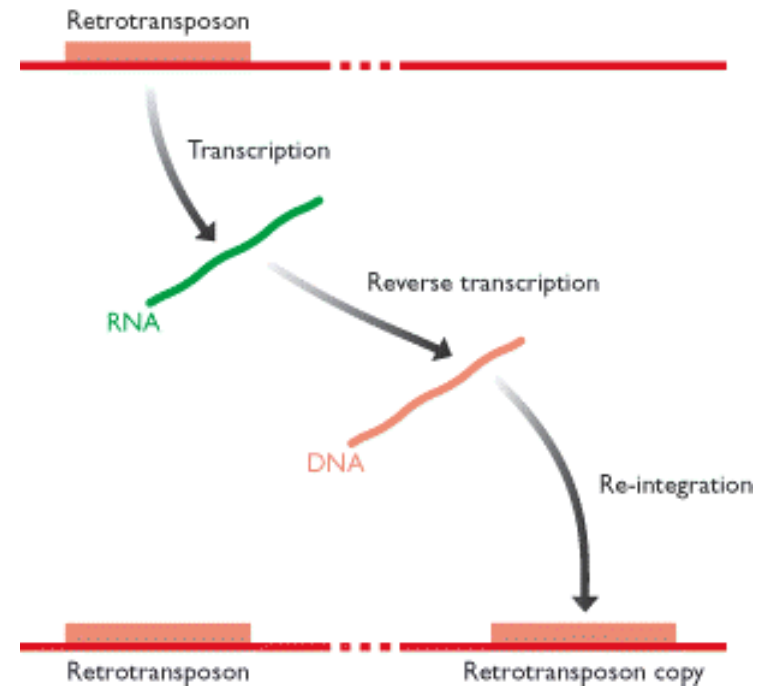
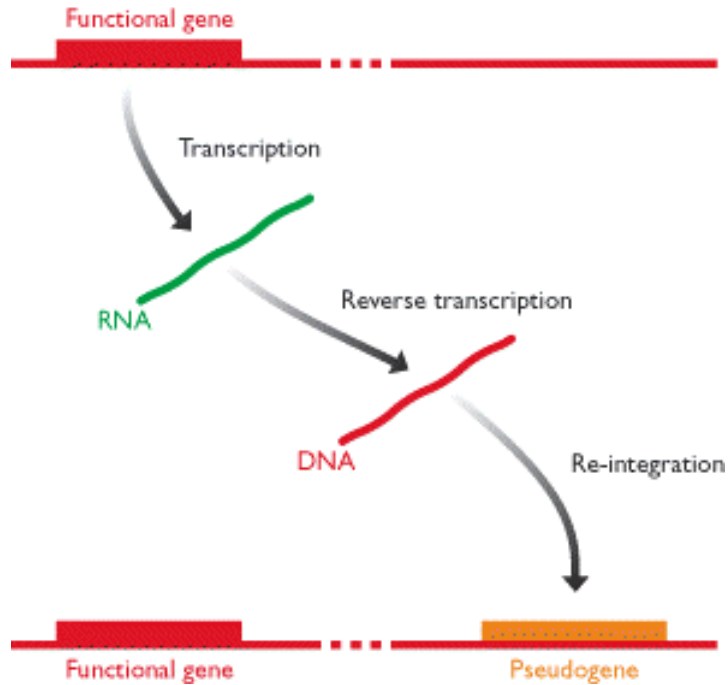
Il y a des régions pauvres en répétitions (p. ex. région des gènes HOX) et des régions riches (région de 500kb du chr. 11 contenant 89% d'éléments transposables)

Les répétitions humaines sont relativement anciennes comparées à celles qu'on trouve dans le génome de drosophile. Notre génome a des difficultés pour se débarrasser des répétitions.



# Retroposons et pseudogènes







## Deux mécanismes proches



# Séquences répétées de type transposon

Les séquences répétées de type transposon représentent plus de 1/3 du génome des vertébrés  
 Génome humain: 45% !!

Classes of interspersed repeat in the human genome

|                          |                |  | Length      | Copy number | Fraction of genome |
|--------------------------|----------------|--|-------------|-------------|--------------------|
| LINEs                    | Autonomous     |    | 6–8 kb      | 850,000     | 21%                |
|                          | Non-autonomous |     | 100–300 bp  |             |                    |
| Retrovirus-like elements | Autonomous     |    | 6–11 kb     | 450,000     | 8%                 |
|                          | Non-autonomous |  | 1.5–3 kb    |             |                    |
| DNA transposon fossils   | Autonomous     |  | 2–3 kb      | 300,000     | 3%                 |
|                          | Non-autonomous |  | 80–3,000 bp |             |                    |

# LINES (Long interspersed repeated sequences)



- ★ 21% du génome humain (850.000 copies). Le plus commun de ces éléments est L1: 6kb.
- ★ Contiennent un promoteur Pol-II et 2 ORFs.
- ★ Après traduction, l'ARN LINE s'assemble avec ses propres protéines et se déplace vers le noyau où l'ARN est reverse transcrit et s'insère dans le génome au niveau d'une coupure simple brin. La transcription inverse s'interrompt souvent avant terme, créant de nombreux inserts tronqués (la plupart en fait)
- ★ La machinerie LINE est responsable également de la retrotranscription des éléments SINE.
- ★ Il y a en fait trois familles de LINE dans le génome humain (LINE1, LINE2, LINE3), mais seule L1 est active.

```
>L1 element| Human L1 interspersed repetitive sequence-full length copy
ggcgggtggagccaagatgaccgaataggaacagctccagtctatagctcccatcgtgagt
gagcagaagacgggtgatttctgcatttccaactgaggtaccaggttcattctcacaggg
aagtgccaggcagtgagggtgcaggacagtagtgcagtgactgtgcatgagccgaagcagg
gagagcatcacctcacccgggaagcacaaggggtcaggaattccctttcctagtcaaa
gaaaaggtgacagatggcacctggaaaatcgggtcactcccgcctaatactgcgctct
tccaacaagcttaacaaatggcacaccaggagattatatcccatgcctggctcagagggt
ctacgcccattgagcctcgctcattgctagcacagcagtcaggtctgaggtcaaactgcaaggt ...
```

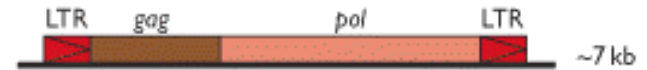
# SINEs (Short interspersed repeated sequences)

- ★ Contiennent un promoteur pol-III mais pas d'ORF
- ★ Vivent sur le dos des LINE
- ★ 13% du génome humain. 3 types: Alu, MIR et Ther2/MIR3.
- ★ (1 500 000 SINEs par génome humain haploïde, dont 1 000 000 Alu)
- ★ La plus connue est ALU, la seule SINE active: 290 bp constitué de 2 répétitions en tandem de 130 bp.
- ★ On ne trouve pas ces séquences dans les régions codantes, mais souvent dans l'unité de transcription, soit dans les introns, soit dans les parties non traduites des ARNm.



```
>ALU | Human ALU interspersed repetitive sequence - consensus A
ggccggggcgcggtggctcacgcctgtaatcccagcactttgggaggccgagggcgggcgga
tcacctgaggtcaggagttcgagaccagcctggccaacatggtgaaaccccgctctctact
aaaaatacaaaaattagccggggcgtggtggcgcgcgctgtaatcccagctactcgggag
gctgaggcaggagaatcgcttgaacccgggaggcgaggttgcagtgagccgagatcgcg
ccactgcactccagcctgggcgacagagcgagactccgtctcaaaaaaaaa
```

# Autres transposons



## Retrotransposons à LTR

- ★ Long terminal repeats avec tous les éléments de régulation de transcription
- ★ A l'intérieur: protease, reverse transcriptase, RNase H et integrase
- ★ L'acquisition du gène env suffit pour en faire un véritable rétrovirus exogène.

## Transposons ADN

- ★ Ressemble aux transposons bactériens
- ★ Flanqué par répétitions inversées
- ★ Code pour une transposase
- ★ Au moins 7 classes de transposons ADN dans le génome humain

# Autres séquences répétées

## Les répétitions simples (Simple Sequence Repeats - SSR)

- ★ Répétition parfaite ou non d'un fragment de longueur k (k-mère).
- ★ Fragment court (1 à 13b): microsatellite
- ★ Fragment long (14 à 500b): minisatellite
- ★ Représentent 3% du génome humain, 0,5% provenant des répétitions de dinucléotides (85% AC ou AT).
- ★ Un SSR tous les 2kb en moyenne

## Les duplications de segments

- ★ Blocs de 1 à 200kb transférés entre différentes positions du génome (inter ou intra-chromosomique)
- ★ Exemple sur chromosome 17: 3 copies d'un segment de 200-kb séparées par environ 5 Mb + 2 copies d'un segment de 24-kb séparées par 1.5 Mb.
- ★ Génome humain: représentent 3,6% du génome (à un niveau > 90% identité).
- ★ Arabidopsis: 58% du génome est constitué de segments dupliqués

## Péricentromères et subtelomères

- ★ Chaque centromère est constitué de zones de duplication interchromosomique.
- ★ P. ex. le Chromosome 22 contient une région de 1.5 Mb adjacente au centromère constituée à 90% de duplications interchromosomiques (en provenance d'autres chromosomes).
- ★ Ces duplications sont complexes, avec de nombreux événements, souvent récents, séparés par des répétitions de type minisatellite riches en AT ou en GC.

# Evolution et variation des génomes

# Réarrangements chromosomiques

- ★ Translocations
- ★ Inversions
- ★ Délétions



# Apparition des nouveaux gènes

## Par duplication

- ★ Duplication du génome entier ou polyploïdisation (plusieurs cas chez les eucaryotes)
- ★ Duplication d'un gène ou d'un groupe de gènes (fréquent)
- ★ Duplication d'un chromosome ou d'une partie (rare car délétère)
- ★ La duplication est suivie le plus fréquemment de la perte de gènes: 90% des gènes dupliqués à l'origine des vertébrés auraient été perdus depuis.

## Par réarrangements de gènes

- ★ Domain shuffling
- ★ Domain duplication

## Par transfert horizontal

- ★ Très important entre génomes procaryotes
- ★ Arrive de procaryote à eucaryote

# Role des éléments transposables

- ★ Par leur insertion, en perturbant des gènes existants
- ★ Par leur capacité à initier des recombinaisons entre différentes parties du génome (dû à leurs séquences presque identiques)

# L'apparition des introns

- ★ (Concerne les introns de type GT-AG, pas les introns catalytiques issus du RNA world)

## 2 hypothèses

- ★ « Intron early ». Les introns étaient présents dans l'organisme ancestral (y compris les bactéries). Ils disparaissent petit à petit
- ★ « Intron late »: apparition chez les premiers eucaryotes, puis prolifération
- ★ Aucune hypothèse n'est clairement réfutée à ce jour

# La ressemblance entre génomes

## Homme/chimpanzé

- ★ Codant: <1,5% de différence
- ★ Non codant ~3% de différence
- ★ Quelques duplications/délétions importantes de région de quelques dizaines de kb

## Homme/souris

- ★ Codant: 90% identique
- ★ Non codant: la plus grande partie des régions non codantes est sans identité apparente, mais il y aurait ~2000nt conservés dans chaque région intergénique chez l'ensemble des mammifères
- ★ Mutations neutres: 0,6/site

## Homme/poulet

- ★ Mutations neutres: 1,5/site

# Génomique comparative

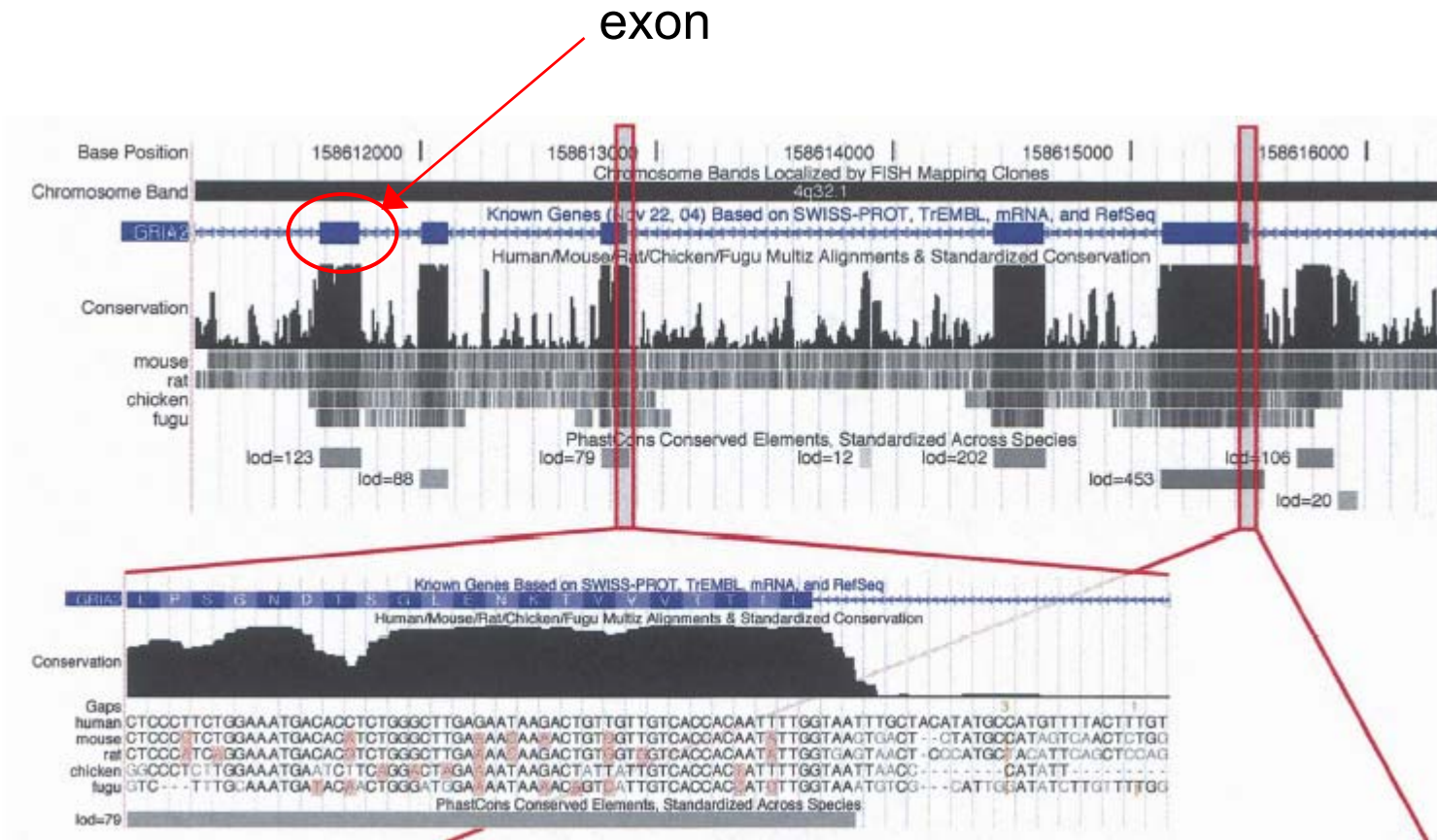
★ Génomique comparative = comparaison de génomes entièrement séquencés

★ **Les applications:**

- Aide l'annotation en identifiant les régions fonctionnelles (les régions non fonctionnelles sont non conservées)
- Identifier le jeu de gènes de chaque organisme
- Comprendre les solutions trouvées par des organismes différents pour une même fonction
- Etudier des gènes/fonctions particuliers par comparaison de séquence (voir cours de bioinformatique)
- Autres questions spécifiques: adaptation, résistance, pathogénicité, etc.

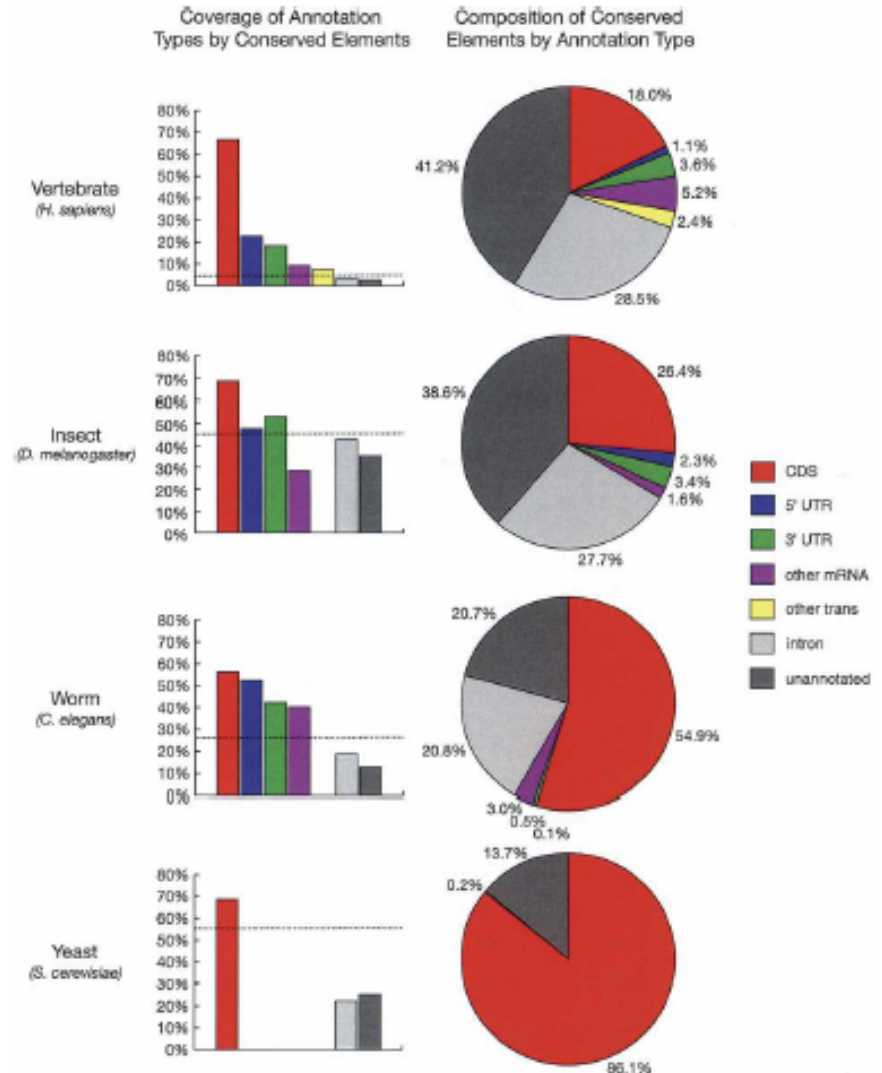
# Génomique comparative

Exemple: gène humain GRIA2 (Siepel et al. Genome Res. 2005)



# L'importance des ARN non codants dans les régions conservées

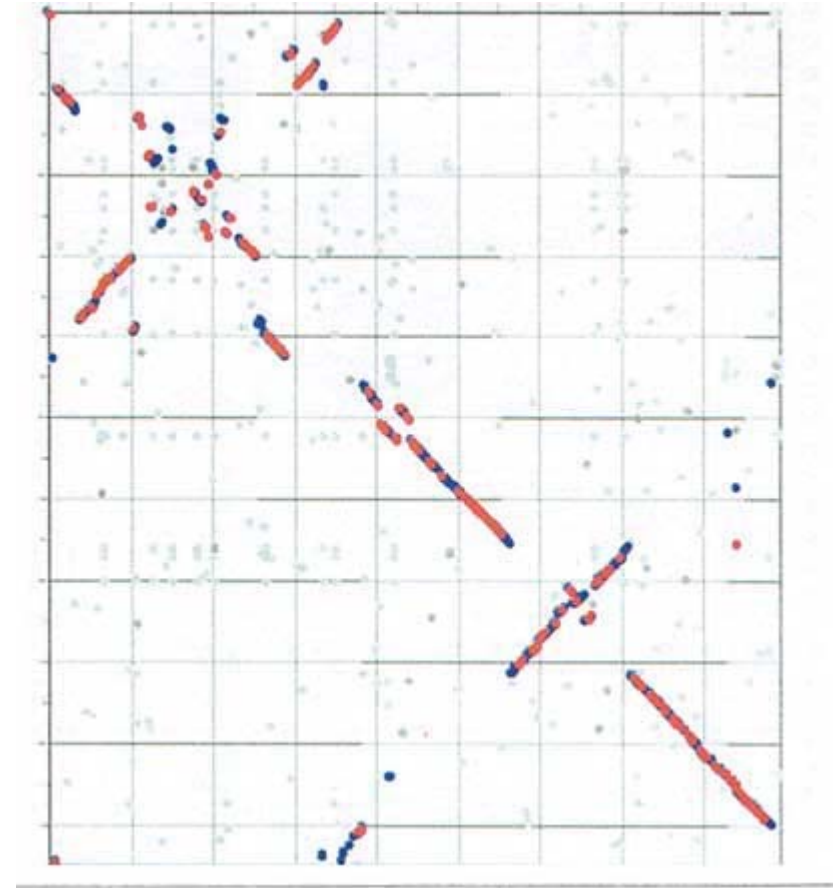
(Siepel et al. Genome Res. 2005)



**Figure 3.** Fractions of bases of various annotation types covered by predicted conserved elements (*left*) and fractions of bases in conserved elements belonging to various annotation types (*right*). Annotation types include CDS, 5' UTR, 3' UTR, other mRNA, other trans, intron, and unannotated.

# La sythénie

- ✦ Le séquençage des premiers génomes a révélé que l'ordre des gènes était beaucoup moins conservé que les séquences.
- ✦ Une région observée chez deux organismes est dite sythénique lorsqu'elle n'a pas subi de réarrangement depuis l'ancêtre commun de ces deux organismes.
- ✦ Humain/souris: environ 150 réarrangements. Régions sythéniques de 8,8cM en moyenne.



*Chlamidia trachomatis* / *chlamydomonada reinhardtii*



# Les variations du génome dans une population

## Très importantes médicalement

- ★ Pharmacogénomique: comment chaque patient répond aux drogues
- ★ Marqueurs de susceptibilité aux maladies

## Polymorphismes dans le génome humain

- ★ Insertions, délétions, duplications, réarrangements
  - rares et peu étudiés
- ★ Microsatellites etc..
- ★ Single Nucleotide (SNP)

# Single Nucleotide Polymorphism (SNP)

- ✦ Le polymorphisme le plus commun chez l'homme.
- ✦ Stable
- ✦ Beaucoup n'ont pas d'implications fonctionnelles
- ✦ 1 SNP tous les 100 à 300 nt. 3 200 000 dans le génome.
- ✦ En comparant 2 individus, on trouve un SNP toutes les 1000/2000 bases.

## Applications

- ✦ Les SNP constituent une trace historique pour l'étude de la phylogénie humaine: ils mutent lentement et ont peu de chance de réapparaître de façon récurrente.
- ✦ Les SNP sont à l'origine de susceptibilité ou de résistance à de nombreuses maladies.
- ✦ Cartographie de maladies à caractères complexes (cancers, diabète, maladies mentales)
- ✦ Prédiction des réponses aux drogues.

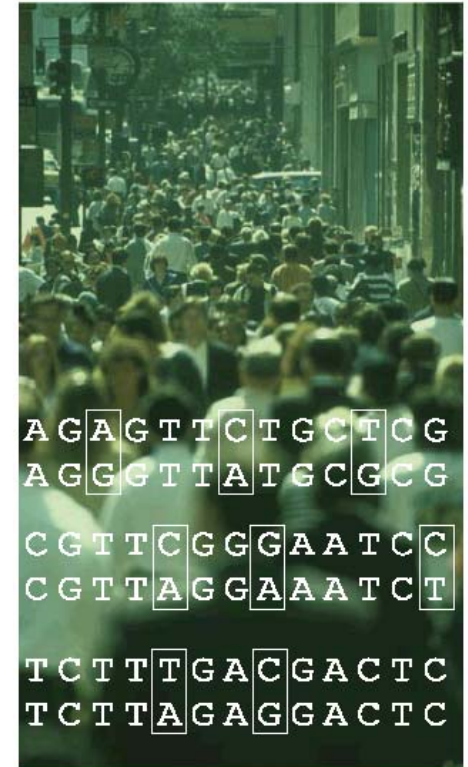
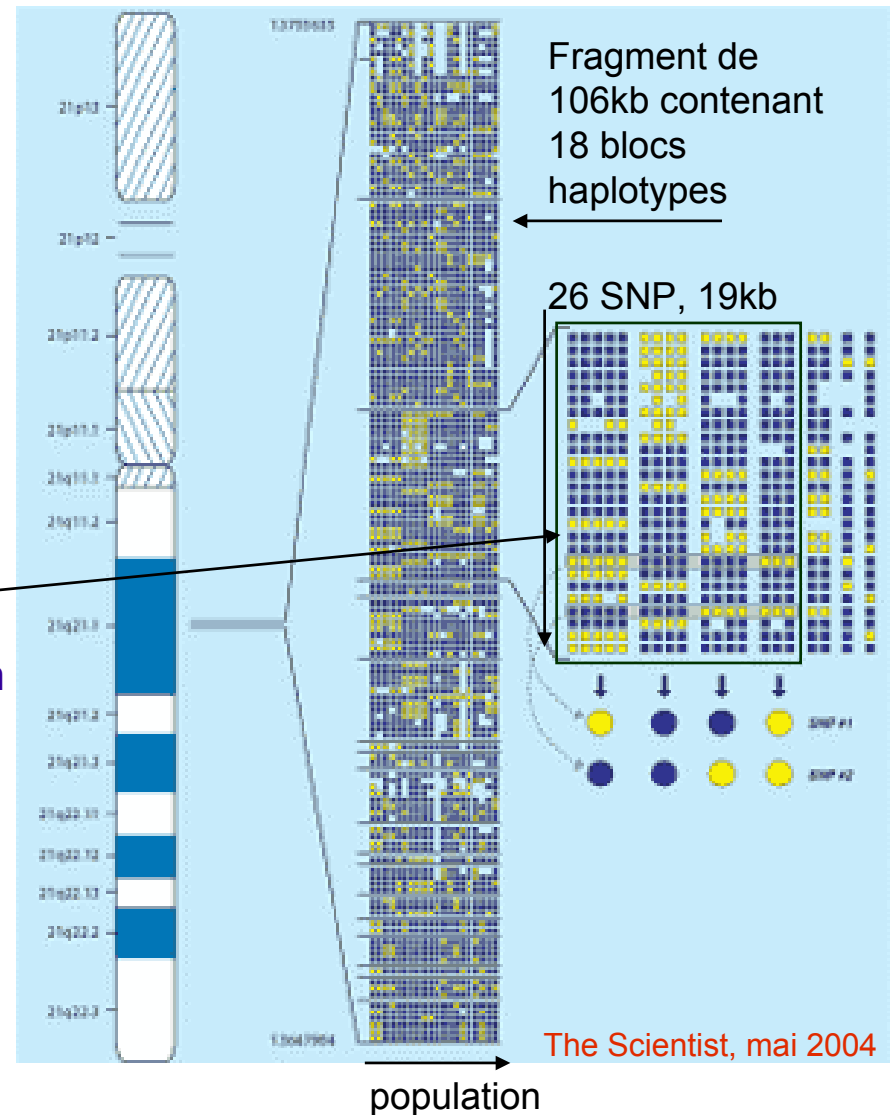


Figure 1 The most common sources of variation between humans are single nucleotide polymorphisms (SNPs) — single base differences between genome sequences. Fragments of two sequences, with eight SNPs, are shown.

# Haplotypes

- ★ Haplotypes: combinaison d'allèles tendant à être transmis ensemble (qq dizaines de kb de long)
- ★ Les SNP sont parfaitement adaptés pour identifier les haplotypes
- ★ Chaque haplotype existe en quelques versions dans la population
- ★ Dans l'exemple ci-contre une région de 19kb est dominée par 4 haplotypes
- ★ 4 SNP suffisent pour repérer l'haplotype d'un patient
- ★ Evite des reséquençages

Chromosome 21:



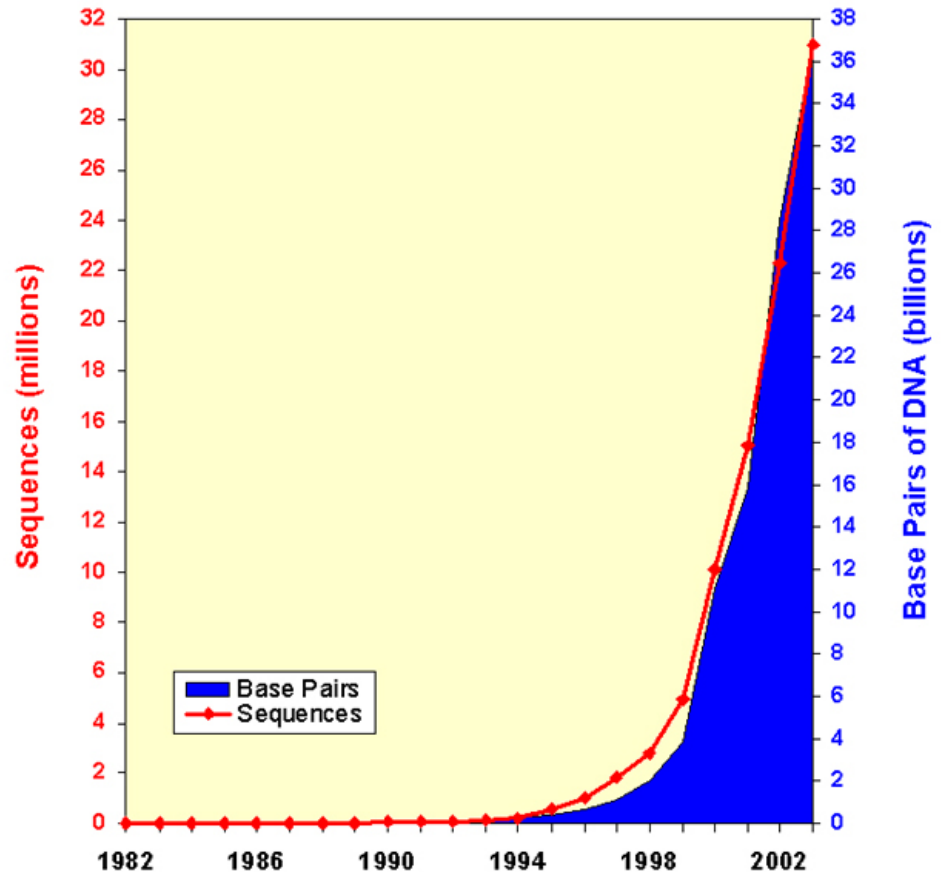
# Les banques de données génomiques

# Genbank: La banque d'ADN du NIH

## Etat au 2-2004

- ★  $38 \times 10^9$  bases
- ★  $32 \times 10^6$  séquences
- ★ Genbank double environ tous les 14 mois depuis ses débuts en 1982.
- ★ Nouvelle version tous les 2 mois

## Growth of GenBank



# Divisions de Genbank

- ★ **ESTs** (Expressed sequence tags):  
Principale division ed Genbank. 18  
10<sup>6</sup> sequences, 580 organismes  
différents
- ★ **GSS** (Genome Sequence Survey):  
résultats de séquençages aléatoire  
de BAC, dans le cadre de projets  
Génome
- ★ **HTGS** (High Throughput Genomic  
Sequences): séquences génomiques  
en cours d'assemblage. Une fois  
assemblées, les séquences passent  
dans les divisions « organisme ».
- ★ Bactéries (**BCT**), virus (**VRL**),  
primates (**PRI**), rongeurs (**ROD**) etc:  
divisions « organismes ».
- ★ 17 divisions en tout.

| Nb entrées | Nb. bases | Espèce                              |
|------------|-----------|-------------------------------------|
| 1355113    | 854232260 | Homo sapiens                        |
| 378892     | 179249409 | Mus musculus                        |
| 76471      | 139699685 | Caenorhabditis elegans              |
| 66177      | 69663817  | Arabidopsis thaliana                |
| 48963      | 53428355  | Drosophila melanogaster             |
| 10571      | 28658828  | Saccharomyces cerevisiae            |
| 39568      | 25816686  | Rattus norvegicus                   |
| 4923       | 17859484  | Escherichia coli                    |
| 32221      | 16490243  | Fugu rubripes                       |
| 31480      | 13072925  | Oryza sativa                        |
| 28406      | 11746328  | Rattus sp.                          |
| 9540       | 10912762  | Schizosaccharomyces pombe           |
| 24125      | 10712174  | Human immunodeficiency virus type 1 |
| 1086       | 9893044   | Bacillus subtilis                   |
| 15370      | 5794059   | Brugia malayi                       |
| 661        | 5701954   | Mycobacterium tuberculosis          |
| 4852       | 5585160   | Gallus gallus                       |
| 4680       | 5400457   | Plasmodium falciparum               |
| 5063       | 4559072   | Bos taurus                          |
| 10845      | 4409926   | Toxoplasma gondii                   |

Organismes dans Genbank (en 2002)

# Enregistrement Genbank

- ★ Chaque enregistrement se voit attribuer un numéro d'accession, stable et unique, et chaque séquence un numéro GI.
- ★ Quand un changement est effectué dans un enregistrement Genbank, le num. d'accession reste, le GI change.

# Enregistrement Genbank avec annotation

```
LOCUS       L10986               47233 bp    DNA    linear   INV 21-SEP-2004
DEFINITION  Caenorhabditis elegans cosmid F10E9, complete sequence.
ACCESSION   L10986
VERSION     L10986.2  GI:38638818
KEYWORDS    HTG.
SOURCE      Caenorhabditis elegans
  ORGANISM  Caenorhabditis elegans
            Eukaryota; Metazoa; Nematoda; Chromadorea; Rhabditida;
            Rhabditoidea; Rhabditidae; Peloderinae; Caenorhabditis.
REFERENCE   1  (bases 1 to 47233)
AUTHORS     .
CONSRTM     WormBase Consortium
TITLE       Genome sequence of the nematode C. elegans: a platform for
            investigating biology. The C. elegans Sequencing Consortium
JOURNAL     Science 282 (5396), 2012-2018 (1998)
MEDLINE     99069613
PUBMED     9851916
```

```
FEATURES             Location/Qualifiers
     source           1..47233
                     /organism="Caenorhabditis elegans"
                     /mol_type="genomic DNA"
                     /strain="Bristol N2"
                     /db_xref="taxon:6239"
                     /chromosome="III"
                     /clone="F10E9"
     gene             265..26728
                     /gene="mig-10"
                     /locus_tag="F10E9.6"
     CDS              join(265..338,3266..3515,15194..15317,21507..21
                     21727..21887,23171..23335,24302..24472,24524..24608,
                     25012..25827,26284..26430,26478..26728)
                     /gene="mig-10"
     /translation="MDSCEEECDLEVDSDEEDQLFGEKICISLLSSLLPLSSSTLLSNA
INLELDEVERPPPLLNVLEEQQFPPKVCANIEEENELEADTEEDIAETADDEESKDPVE
KTENFEPSVTMDTYDFDPYPVQIRARPVPPKPPIDTVRYSMNNIKESADWQLDELL
EELEALETQLNSSNGGDQLLLGVSGIPASSSRENVKSISTLPPPPALS YHQTPQQPQ
. . .
QVYTIGIWEEKYKSPWPWCISIKLTALQMKRSQFIKYICAEDEMTFKKWLVALRIAKN
GAELLENYERACQIRRETLGPASSMSAASSSTAISEVPHLSLHHQRTPSVASSIQLSS
HMMNNPHTPLSVNVRNQSPASFVSNVCQQSHPSRTSAKLEIQYDEQPTGTIKRAPLDV
LRRVSRASTSSPTIPQEESSDSEDFPAPPPVASVMRMPVPVTPPKPCTPLTSKKAPPP
PKRSDTTKLQSASPMAPAKNDLEAALARRREKMATMEC"
```

```
BASE COUNT      2598 a    2024 c    1888 g    2449 t
ORIGIN
1  ttctaaaagt cgaaaaacga gcaatTTTTg atgctagatt ttttgattg acgaatTTTT
61 tcagttTTTT ttcttTaaaa aaggTTTTtg acccctTaaa gttttccttt cccttccaat
121 tttttccttc ttctttatac gacttctcaa gtttcaactc taaaacaag ctacatgtac
181 atttccggta aactttgtgt ctcagaagat ccattttctt tttgttacct ttattcaaga
241 ttgaattcca aaatTtcagc caatatggac agttgcgaag aggaatgcga tctggaagtt
301 gacagtgcgc aagaagatca acctTTTTgt gaaaagtgtg gaggttctat tgtggtaacc
361 aaagaaatgt cagtggTccg taacacttg actcccaaat ggtttctcgt aattacctta
421 tgcacacttt tcaagtgttt gccgTttgat cttagccaat ttgaacgTt tagatgttaa
481 atggaaaatg ggtaaaagtt tttatTttat agaaaaaagg tttggaaaaa aatcgagtca
541 ctgaatagtt tgaagaacgg aaaaaTaaa ctttccaaa atcataaaac atttagtTt
601 tcgaaaatta tagtgTTTT tttgtTgTta tgttttgaca aaagctaaac catctttatt
661 gtagTttTgt aaaatgtTca caaagatgcg tttttTttc aaattTggca ggctatcttt
721 acattcacat ttggataatt caaatTTTT ttatcgtaa caaatTttcc tattttTcca
781 attattcgtt ttTataaagc tttgTtagta tgtTgtgtct atctTtagtG gTcatcagtt
. . .
//
```



# Les banques de gènes

| Nom                                    | Type                   | Organisme                                       | Description   | Nb. Enrgst.                     |
|--|------------------------|---|---|---------------------------------|
| <a href="#">Refseq</a>                 | gene+<br>mRNA+<br>prot | H, M, R   | Itération à partir de Blast seed mRNA/EST vs contigs de Genbank. +Annotation manuelle: publications, UTR prolongées, etc. | 11 405(H) 5 749 (M)             |
| <a href="#">HGI (Human Gene Index)</a> | mRNA                   | H, M, R, D, Arabid., Fugu, riz, etc.            | Genbank mRNA+EST contigués. Toutes les solutions alternatives sont conservées.  | 388 000                         |
| <a href="#">Ensembl</a>                | gènes +<br>transcrits  | H + eucaryotes?                                 | Genscan sur contig, puis Blast vs prot, mRNA, EST, PFAM   | 35 500 genes, 44 860 transcrits |
| <a href="#">Unigene</a>                | clusters               | H, M, R, bovin, zebrafish, blé, riz, mais, orge | Clustering itératif à partir de mRNA+CDS génomique+ESTs. Pas de contigage.  | 89 371 cl                       |

★ [Unigene](#): banque d'ESTs classifiés ("clusterisés"). Dans chaque cluster Unigene sont regroupés des EST ayant une similarité de séquence significative. On peut donc trouver des transcripts différents et des artefacts (chimères, etc.). Unigene ne propose pas de mRNA reconstruits (contigs) à partir des séquences d'un cluster.

★ TIGR Human Gene Index ([HGI](#)). Ici encore on a clusterisé les EST, mais HGI est une banque de "contigs", c.a.d. de séquences de mRNA reconstruites à partir des EST d'un même cluster. Les clusters étant souvent hétérogènes, ils produisent souvent plusieurs contigs. Ces contigs doivent théoriquement correspondre à des mRNA alternatifs.

# Autres banques de séquences

## Nucléotidiques

- ★ gbEST / dbEST: Division EST dans Genbank
  - ★ EMBL: Equivalent européen de Genbank. Format différent, contenu presque identique.
  - ★ Banques spécialisées Certaines collections de séquences, bien que généralement présentes dans Genbank, sont beaucoup plus utiles lorsqu'elles sont rassemblées dans des banques spécialisées, par ex:
    - Récepteurs des lymphocytes T (Réarrangements de l'ADN)
    - Génomes HIV, etc.
- ★ NR nucléique (Non-redundant). Banque combinée: Genbank+refseq (20x10e9 nt / oct. 2002)

# Ensembl ([www.ensembl.org](http://www.ensembl.org))

## ★ Plusieurs banques en une:

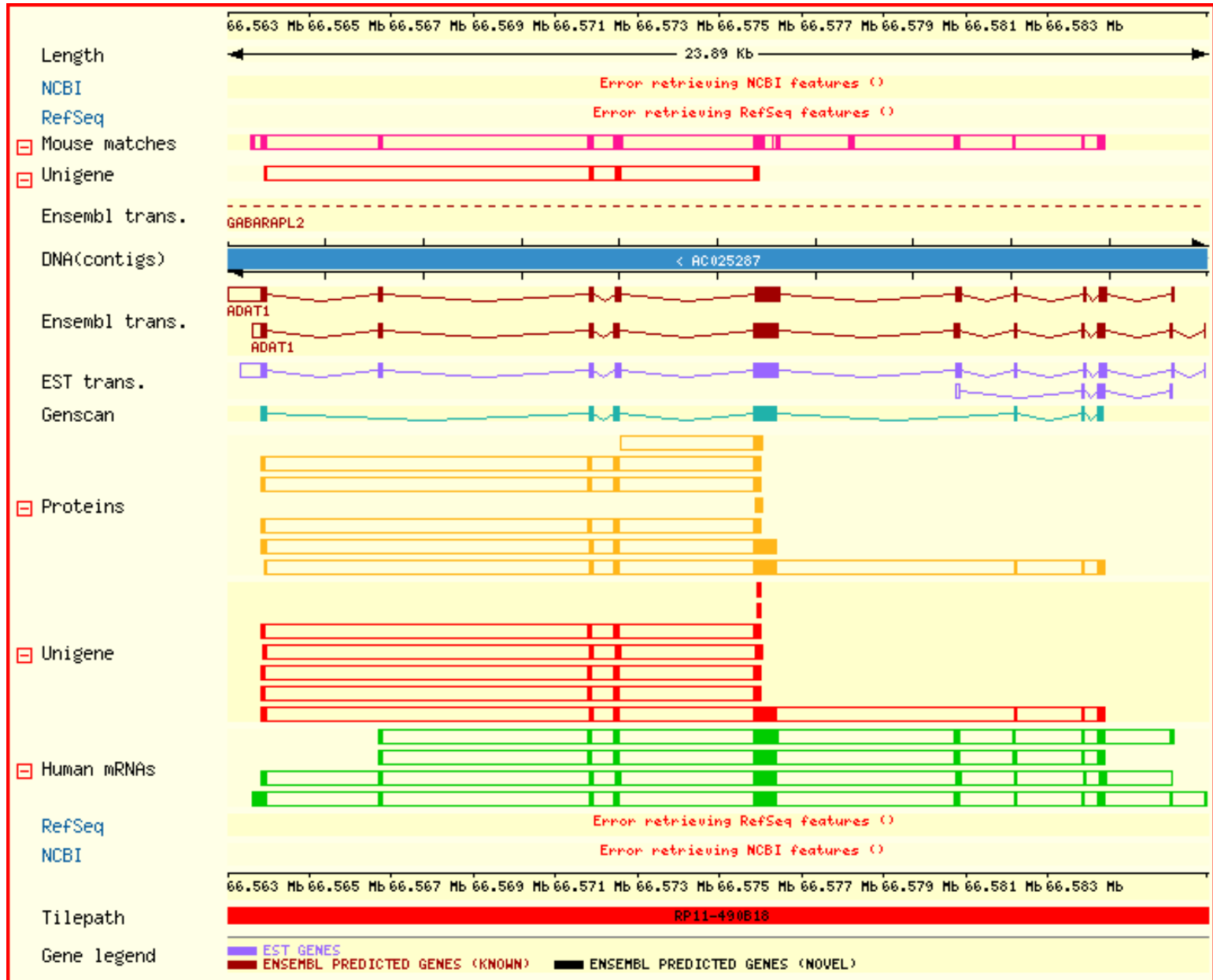
- Peptides confirmés
- Transcrits confirmés
- peptides prédits
- Transcrits prédits
- Génome assemblé (golden path)

## ★ Méthode de prédiction (système Genewise): Genscan sur contig, puis Blast contre: protéines, mRNA, EST

## ★ Version Juillet 2001, humain: Confirmed genes: 21921; Predicted genes: 24636; Confirmed exons: 143479; Predicted exons: 770562; Transcripts: 23931; Contigs: 329154; Sequences: 29080; base pairs: 4318661441.

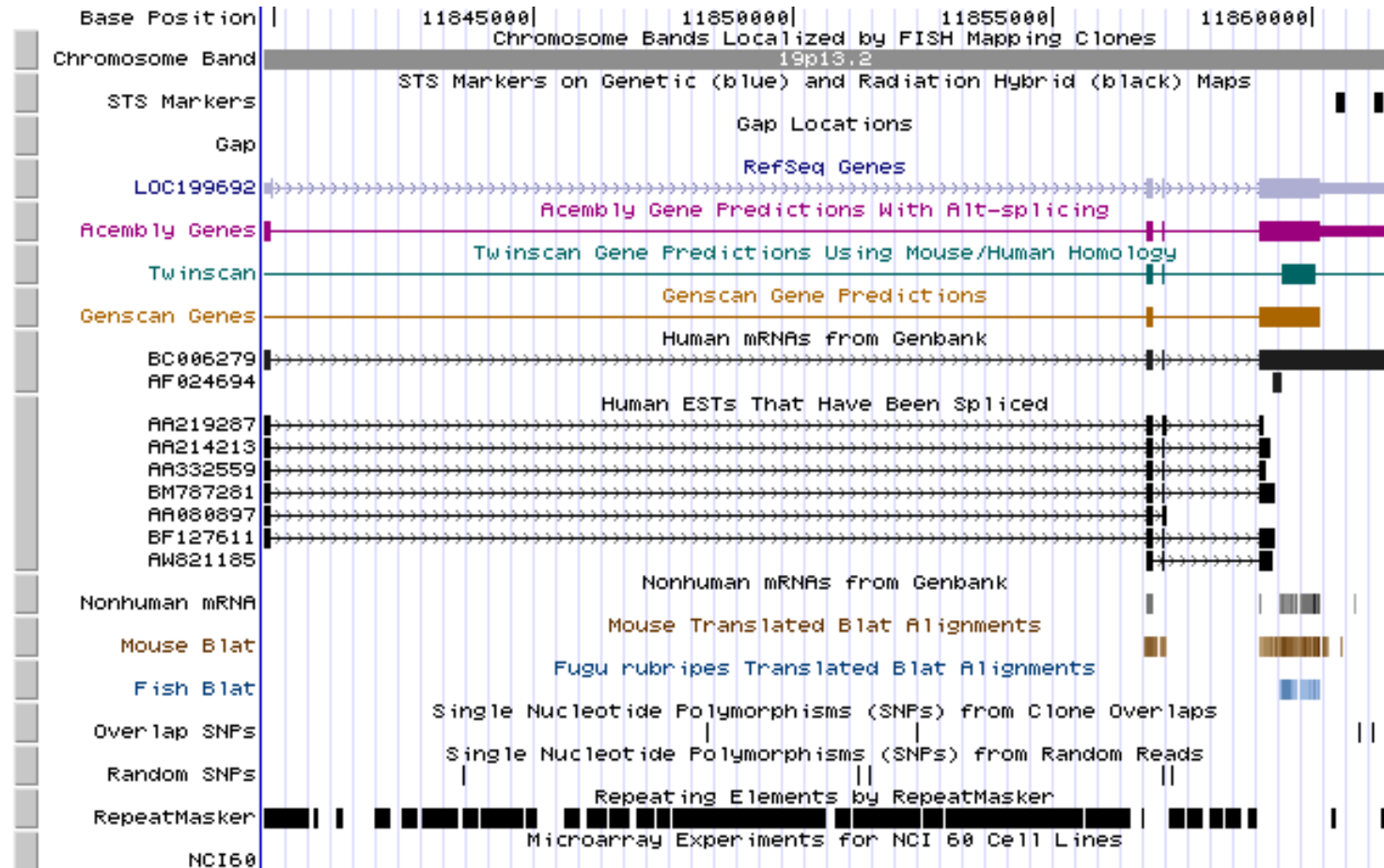
| Species - Ensembl v24 |             |                   |
|-----------------------|-------------|-------------------|
| Human                 | <i>pre!</i> | NCBI 34 Jul 04    |
| Mouse                 |             | NCBI m33 Jul 04   |
| Zebrafish             |             | WTSI Zv4 Sep 04   |
| Rat                   |             | RGSC 3.1 Jul 04   |
| Chicken               |             | WASHUC1 Jul 04    |
| Mosquito              |             | MOZ 2 Apr 04      |
| Fugu                  |             | Fugu v2.0 May 04  |
| Fruitfly              |             | BDGP 3.1 Jul 03   |
| Chimp                 |             | CHIMP1 May 04     |
| Honeybee              |             | Amel1.1 Sep 04    |
| Tetraodon             |             | TETRAODON7 Sep 04 |
| Dog                   | <i>pre!</i> | BROADD1           |
| <i>C. elegans</i>     |             | WS 116 Apr 04     |
| <i>C. briggsae</i>    |             | cb25.aqp8 Jul 03  |

# Ensembl: « contig view »



# Banque génomique UCSC

★ <http://www.genome.ucsc.edu/>



# Banques protéiques

- ★ Swissprot. La mieux annotée des banques protéiques.  
Release 39 (2001): 101247 entrées, 37 135 523 aa.  
Attention: toutes les protéines connues n'y sont pas! Visiter le serveur
- ★ PIR (Protein Identification Resource), EMBL.
- ★ NR Protéique (Non-redundent): Banque protéique du NCBI  
= Traduction de tous les CDS de GenBank + PDB + SwissProt + PIR + PRF - redondances.
- ★ Banques spécialisées
  - Cazy (Carbohydrate Active Enzymes)
  - Etc.

# SRS

## Sequence Retrieval System

## Database selection page

The screenshot shows the SRS website interface in a Netscape browser window. The browser's address bar displays the URL: <http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?page+top+id+ufno10a1eD>. The website header includes the SRS@EMBL-EBI logo and navigation tabs for Quick Search, Library Page, Query Form, Tools, Results, Projects, Views, and Databanks. A search bar with a "Quick Search" button and a "Reset" button is located below the header. The main content area is titled "Available Databanks" and features a "Show databanks tooltips:" checkbox which is checked. The page is organized into several sections, each with a "collapse all" and "expand all" control:

- Literature, Bibliography and Reference Databases**
  - MEDLINE  OMIM  TAXONOMY  GENETICCODE
  - Patent Abstracts  KarynsGenomes
  - Literature, Bibliography and Reference Databases - subsections*
    - MEDLINE (Main Release)  MEDLINE (Updates)  OLDMEDLINE  MED2PUB
- Nucleotide sequence databases**
  - EMBL  EMBL (Contig)  EMBL (Contigs expanded)  Genome Reviews
  - IMGTHLA  IMG/LIGM-DB  LiveLists  PATENT\_DNA
  - RefSeq Genome DB
  - Nucleotide sequence databases - subsections*
    - EMBL (Release)  EMBL (Updates)  EMBL (Third Party Annotation)  EMBL (Coding Sequences)
    - RefSeq Genome Release  RefSeq Genome Updates
- UniProt Universal Protein Resource**
  - UniProt  UniParc  UniRef100  UniRef90  UniRef50
  - UniProt/Swiss-Prot  UniProt/TrEMBL
- Other protein sequence databases**
  - RefSeq Proteome DB  IPI  EPO Proteins  JPO Proteins
  - USPTO Proteins  MHCBN  BCIPEP  SWISSCHANGE
  - Protein sequence databases - subsections*
    - Refseq Proteome Release  Refseq Proteome Updates
- Deprecated Protein Databases**
- Nucleotide related databases**
- Protein function databases**
- Protein structure databases**

On the left side of the page, there is a "Search Options" section with instructions: "1. Select the databanks you want to search" and "2. Enter your search terms in the Quick Search box, or choose a query form from below". It includes buttons for "Standard Query Form" and "Extended Query Form", and a "Browse Entries" button. Below this is a "Tips" section with links to bookmark the link, return to the project, and a document regarding linking to the SRS server. At the bottom left, there is a "BookMarkLets" section.

# SRS

Query:

droso: OK  
elegans: NON  
\*elegans: OK

Standard Query Form - Netscape

File Edit View Go Bookmarks Tools Window Help

http://srs.embl.ac.uk/srsbin/cgi-bin/wgetz

Home Local Institutions Journaux Mot/Annu Cours/Guides MolBio 1 RNA FP6-ATD trad

SRS@EMBL-EBI Quick Search Library Page Query Form Tools Results Projects Views Database

Reset search UniProt/Swiss-Prot

**Search Options**

Combine search terms with: & (AND)

Use wildcards

Get results of type: Entry

**Fields you can search**

Your search terms

In a single field, you can separate multiple values by &, |, !

|               |            |
|---------------|------------|
| AllText       | argonaute  |
| Organism Name | drosophila |
| AllText       |            |
| AllText       |            |

**Result Display Options**

View results using: UniprotView

or

Create a view

Show 30 results per page

**Create a view**

Select the fields you want displayed in your view and choose the format

Choose 1 or more fields:

- ID
- EntryName
- AccessionNumber
- Creation Date
- Seq Mod Date
- Annot Mod Date
- Description

Display As:  Table  List

Sequence Format: swiss



# Entrez

The screenshot shows a Netscape browser window titled "Entrez cross-database search - Netscape". The address bar contains "http://www.ncbi.nlm.nih.gov/Entrez/". The browser's menu bar includes "File", "Edit", "View", "Go", "Bookmarks", "Tools", "Window", and "Help". The browser's toolbar shows "Home", "Local", "Institutions", "Journaux", "Mot/Annu", "Cours/Guides", "MolBio 1", "fRNA", "FP6-ATD", and "trad".

The main content area features the NCBI logo on the left and the Entrez logo with the tagline "Entrez, The Life Sciences Search Engine" on the right. Below the logo is a navigation bar with links for "HOME", "SEARCH", "SITE MAP", "PubMed", "Entrez", "Human Genome", "GenBank", "Map Viewer", and "BLAST".

A search bar is located below the navigation bar, with the text "Search across databases" and a "GO" button. To the right of the search bar are "CLEAR" and "Help" buttons.

The main content area is titled "Welcome to the new Entrez cross-database search page". It displays a grid of database icons and descriptions:

- PubMed:** biomedical literature citations and abstracts
- PubMed Central:** free, full text journal articles
- Books:** online books
- OMIM:** online Mendelian Inheritance in Man
- Site Search:** NCBI web and FTP sites
- Nucleotide:** sequence database (GenBank)
- Protein:** sequence database
- UniGene:** gene-oriented clusters of transcript sequences
- CDD:** conserved protein domain database
- Genome:** whole genome sequences
- 3D Domains:** domains from Entrez Structure
- Structure:** three-dimensional macromolecular structures
- UniSTS:** markers and mapping data
- Taxonomy:** organisms in GenBank
- PopSet:** population study data sets
- SNP:** single nucleotide polymorphism
- GEO Profiles:** expression and molecular abundance profiles
- Gene:** gene-centered information
- GEO DataSets:** experimental sets of GEO data
- HomoloGene:** eukaryotic homology groups
- Cancer Chromosomes:** cytogenetic databases
- PubChem Compound:** small molecule chemical structures
- PubChem BioAssay:** bioactivity screens of chemical substances
- PubChem Substance:** chemical substances screened for bioactivity
- Journals:** detailed information about the journals indexed in PubMed and other Entrez databases
- MeSH:** detailed information about NLM's controlled vocabulary
- NLM Catalog:** catalog of books, journals, and audiovisuals in the NLM collections

The browser's status bar at the bottom shows "Done" and various system icons.

# Glossaire de génomique...

[http://www.sciencemag.org/cgi/content/full/291/  
5507/1197](http://www.sciencemag.org/cgi/content/full/291/5507/1197)

(*Science*, Vol 291, 1197)