# Bioinformatique: analyse des génomes

Céline Brochier-Armanet
(Laurent Duret)
Université Claude Bernard, Lyon 1
Laboratoire de Biométrie et Biologie évolutive (UMR 5558)
Celine.brochier-armanet@univ-lyon1.fr

## Présentation

- Equipe "Bioinformatique et Génomique Evolutive"
  - Labo. de Biométrie et Biologie Evolutive (CNRS, Univ. Lyon 1)
  - Pôle Bioinformatique Lyonnais (avec l'équipe de G. Deléage, IBCP): http://pbil.univ-lyon1.fr
- Développement d'outils informatiques pour l'analyse des génomes
  - Bases de données (Hogenom, Hovergen, Homolens, etc)
  - Algorithmes & outils
- Etude de l'organisation et de l'évolution des génomes: étudier l'évolution des génomes pour comprendre leur fonctionnement (et vice versa)
  - Évolution moléculaire
  - Analyse comparative de séquences
  - Phylogénie

## Introduction

- La biologie à l'heure du séquençage des génomes
- Séquençage de génomes:
  - Pourquoi?
  - Comment?
- L'annotation des génomes
  - Comment?
  - Problèmes techniques?
- Exploiter les données issues du séquençage
  - Questions scientifiques?
  - Approches?
  - Problèmes techniques?

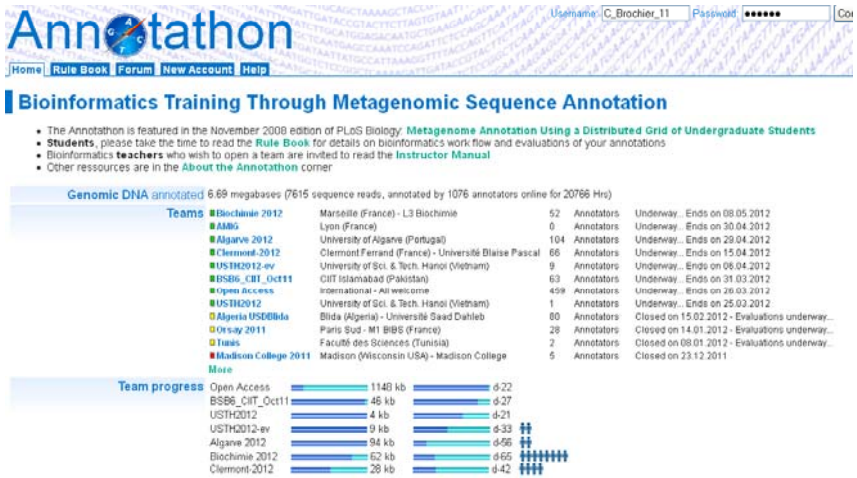
## Annotation des génomes

- Annotation structurale:
  - ⇒ Identifier tous les éléments fonctionnels présents dans les séquences génomiques
  - Gènes
  - mais aussi… les promoteurs, les sites de fixation des facteurs de transcription, les régions répétées, les origines de réplication, etc.
- Annotation fonctionnelle:
  - ⇒ Déterminer la fonction des produits des gènes
  - Comment définir la fonction d'un gène?

# Plan du cours

- Cours
  - Projets génomes
  - Structure des génomes procaryotes & eucaryotes
  - Annotation des gènes procaryotes
  - Annotation des gènes eucaryotes
  - Prédiction de fonction des gènes
  - Comparaison de séquences (alignement, recherche de similarités)
- TP: Annotathon (http://annotathon.org/)
  - Annotation de fragments d'ADN procaryotes
- Lecture d'articles

# Annotathon



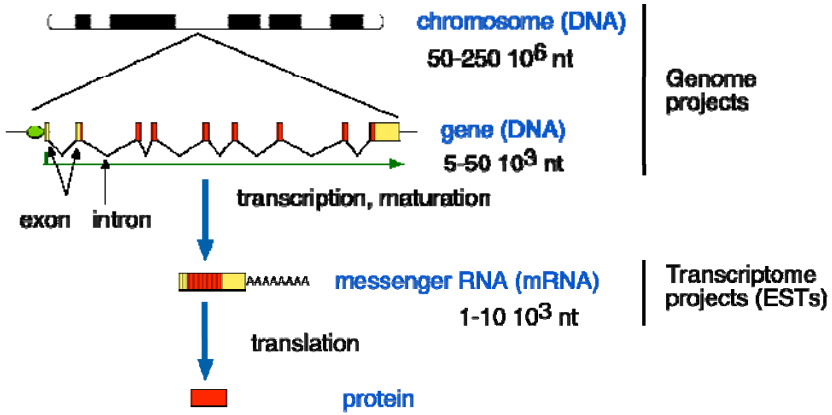

# Séquençage de l'ADN: historique

- 1943-1953: ADN support de l'information génétique
- 1977: Techniques modernes de séquençage de l'ADN (Maxam & Gilbert, Sanger et . al)
- 1981: Séquençage du génome mitochondrial humain
- 1982: Premières banques de données de séquences (GenBank, EMBL)
- 1990: Début du projet génome humain (cartographie)
- 1995: Premier génome complet d'un organisme cellulaire (*H. influenzae*)
- 1999: Chromosome 22 humain
- 2001: Première ébauche du génome humain
- 2003-??: Séquençage du génome humain achevé
- 2007: Première séquence complète du génome d'un individu
- 2011: >1900 génomes d'espèces différentes

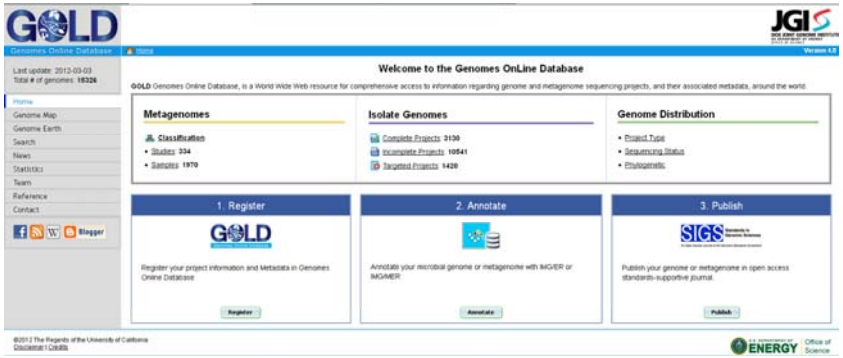
# De l'artisanat à l'industrie

- Séquencer pour répondre à une question donnée: de la biologie à la séquence (1980-1995)
  - ⇒ Phénotype => Gène
  - Equipement progressif en séquenceurs des laboratoires de biologie moléculaire
  - Séquençage de gènes ou d'ARNm (< 10 kb)
  - Informations biologiques associées aux séquences: riches
- Séquençage systématique à grande échelle: de la séquence à la biologie (>1995)
  - ⇒ Gènes => Phénotype
  - Apparition des grands centres de séquençage
  - Séquençage de grands fragments génomiques, chromosomes, génomes, etc. …
  - Informations biologiques associées aux séquences: pauvres

# Séquençage : Génome / Transcriptome



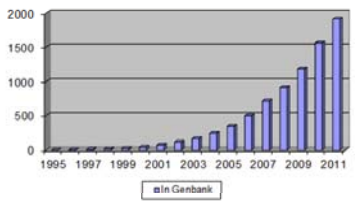

# Les projets de séquençage en quelques chiffres



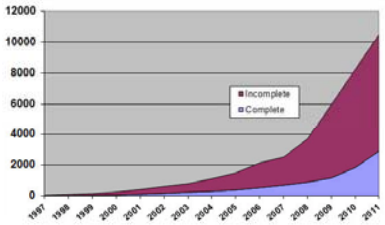

http://www.genomesonline.org/cgi-bin/GOLD/index.cgi

# Génomes complets & projets génomes

**Génomes complets**

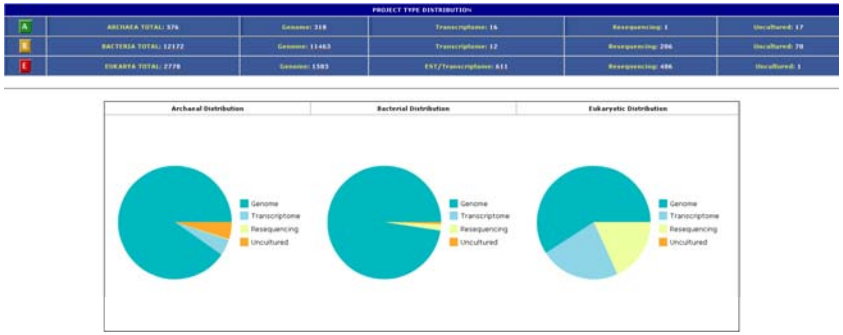



**Projets génomes**

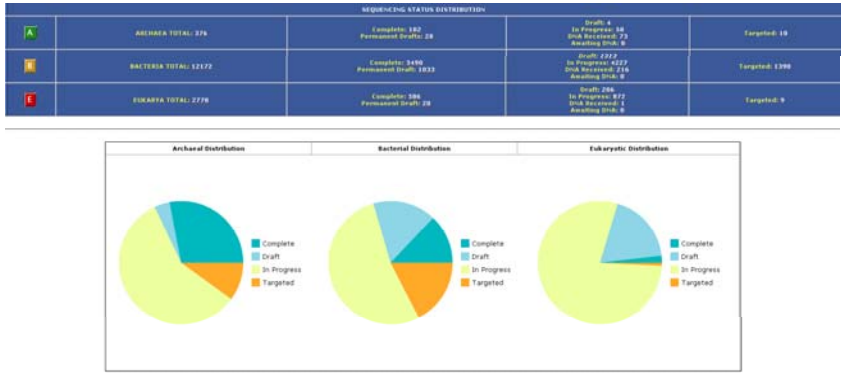





# Types de projets génomes

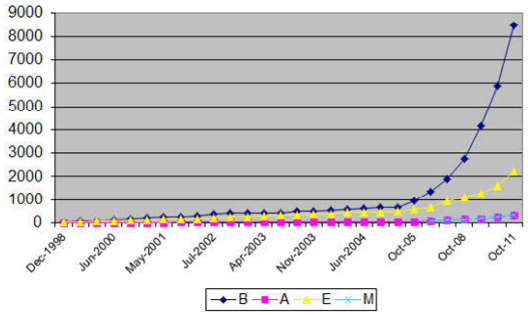


# Etats des projets génomes



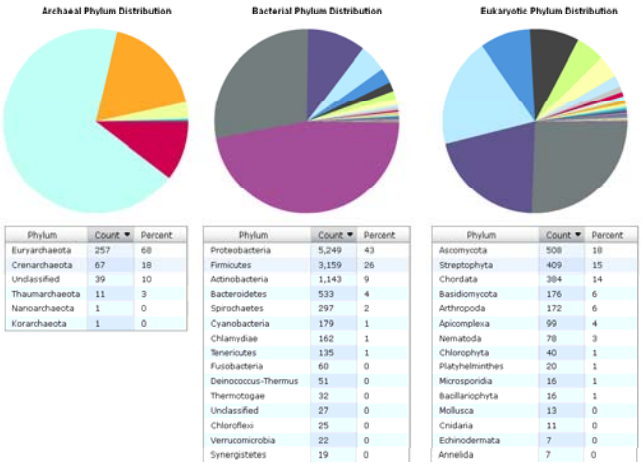

Octobre 2011

# Des inégalités taxonomiques entre les domaines




Octobre 2011

# Des inégalités taxonomiques au sein des domaines




| Phylum | Count | Percent |
|---|---|---|
| Euryarchaeota | 257 | 68 |
| Crenarchaeota | 67 | 18 |
| Undassified | 39 | 10 |
| Thaumarchaeota | 11 | 3 |
| Nanoarchaeota | 1 | 0 |
| Korarchaeota | 1 | 0 |

| Phylum | Count | Percent |
|---|---|---|
| Proteobacteria | 5,249 | 43 |
| Firmicutes | 3,159 | 26 |
| Actinobacteria | 1,143 | 9 |
| Bacteroidetes | 533 | 4 |
| Spirochaetes | 297 | 2 |
| Cyanobacteria | 179 | 1 |
| Chlamydiae | 162 | 1 |
| Tenericutes | 135 | 1 |
| Fusobacteria | 60 | 0 |
| Deinococcus-Thermus | 51 | 0 |
| Thermotogae | 32 | 0 |
| Unclassified | 27 | 0 |
| Chloroflexi | 25 | 0 |
| Verrucomicrobia | 22 | 0 |
| Synergistetes | 19 | 0 |

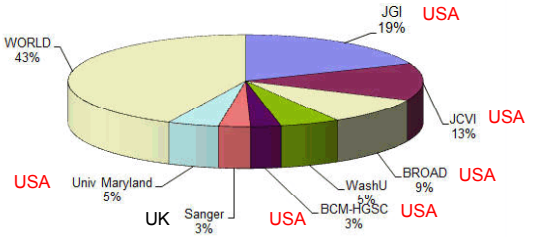
| Phylum | Count | Percent |
|---|---|---|
| Ascomycota | 508 | 18 |
| Streptophyta | 409 | 15 |
| Chordata | 384 | 14 |
| Basidiomycota | 176 | 6 |
| Arthropoda | 172 | 6 |
| Apicomplexa | 99 | 4 |
| Nematoda | 78 | 3 |
| Chlorophyta | 40 | 1 |
| Platyhelminthes | 20 | 1 |
| Microsporidia | 16 | 1 |
| Bacillariophyta | 16 | 1 |
| Mollusca | 13 | 0 |
| Cnidaria | 11 | 0 |
| Echinodermata | 7 | 0 |
| Annelida | 7 | 0 |

Octobre 2011

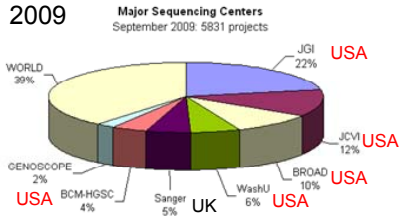
# Redistribution des sites de séquençage

2011

USA=54%; France=??%

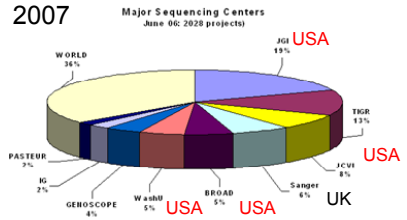



2009




USA=56%; France=2%

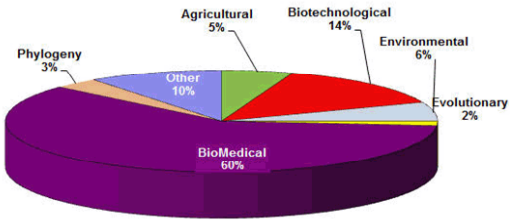
2007



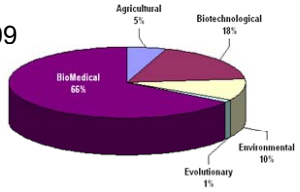

USA=50%; France=6%

# Motivations scientifiques pour le séquençage des génomes

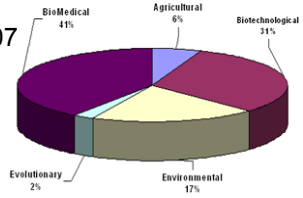
2011




2009




2007




# Pourquoi séquencer des génomes complets?

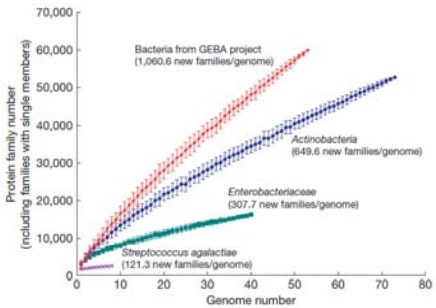
- Inventorier l'information génétique nécessaire au développement et à la reproduction d'un organisme
- Comprendre les bases génétiques de la variabilité phénotypique (e.g. pathologie; projet 1000 génomes)
- Etablir et comprendre l'histoire du vivant
- Analyser la biodiversité: Classification & identification automatisée des taxa (BarCoding, MLSA), Diagnostique, Epidémiologie
- Ingénierie & génie génétique: Applications médicales, agronomiques, industrielles

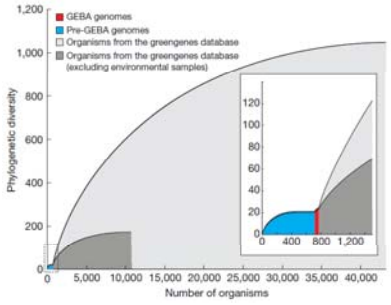
# Une encyclopédie génomique

LETTERS 2009

**A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea**

Dongying Wu[1,2], Philip Hugenholtz[3], Konstantinos Mavromatis[1], Rüdiger Pukall[3], Eileen Dalin[1], Natalia N. Ivanova[1], Victor Kunin[1], Lynne Goodwin[1], Martin Wu[1], Brian J. Tindall[3], Sean D. Hooper[1], Amrita Pati[1], Athanasios Lykidis[1], Stefan Spring[3], Iain J. Anderson[1], Patrik D'haeseleer[1,4], Adam Zemla[5], Mitchell Singer[1], Alla Lapidus[1], Matt Nolan[1], Alex Copeland[1], Cliff Han[1], Feng Chen[1], Jan-Fang Cheng[1], Susan Lucas[1], Cheryl Kerfeld[1], Elke Lang[3], Sabine Gronow[3], Patrick Chain[1,4], David Bruce[4], Edward M. Rubin[1], Nikos C. Kyrpides[1], Hans-Peter Klenk[3] & Jonathan A. Eisen[1,2]




# Pourquoi séquencer des transcriptomes?

- Identifier les gènes présents dans les génomes
- Identifier les variants d'épissage (eucaryotes)
- Etudier l'expression des gènes (en quelle quantité, dans quels tissus, à quels stades du développement, en réponse à des variations de l'environnement, …)

# Pourquoi séquencer des métagénomes?

- Ecologie microbienne: inventorier la diversité des écosystèmes microbiens naturels (soils, océans, intestin, etc.) ou non (industrie alimentaire, vinification, etc.)
- Comprendre les interactions des microorganismes dans les écosystèmes
- Décrypter le fonctionnement des écosystèmes (cycles géochimiques, flux d'énergie, etc.)
- Suivre l'évolution spatiale et temporelle de la diversité dans les écosystèmes (saisons, perturbations, etc.)

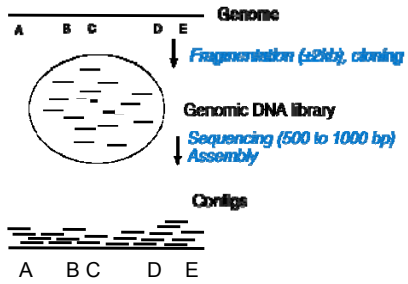
# Séquençage de génomes

- Trois étapes:

1- Purification et découpage de l'ADN

2- Séquençage de fragments d'ADN

3- Assemblage des fragments



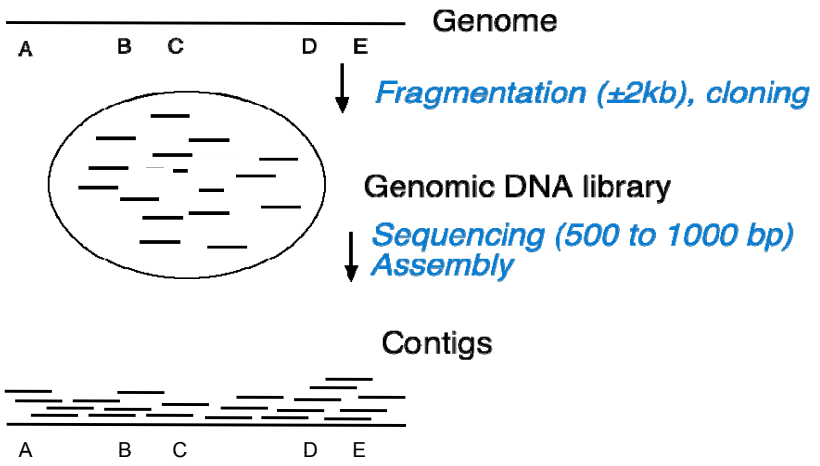

# Technologies de séquençage à haut débit

| Platform | Library/ template preparation | NGS chemistry | Read length (bases) | Run time (days) | Gb per run | Machine cost (US$) | Pros | Cons | Biological applications | Refs |
|---|---|---|---|---|---|---|---|---|---|---|
| Roche/454's GS FLX Titanium | Frag, MP/ emPCR | PS | 330* | 0.35 | 0.45 | 500,000 | Longer reads improve mapping in repetitive regions; fast run times | High reagent cost; high error rates in homo-polymer repeats | Bacterial and insect genome de novo assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics | D. Muzny, pers. comm. |
| Illumina/ Solexa's $GA_x$ | Frag, MP/ solid-phase | RTs | 75 or 100 | 4[9], 9[8] | 10[9], 35[5] | 540,000 | Currently the most widely used platform in the field | Low multiplexing capability of samples | Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics | D. Muzny, pers. comm. |
| Life/APG's SOLID 3 | Frag, MP/ emPCR | Cleavable probe SBL | 50 | 7[9], 14[5] | 30[9], 50[5] | 595,000 | Two-base encoding provides inherent error correction | Long run times | Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics | D. Muzny, pers. comm. |
| Polonator G.007 | MP only/ emPCR | Non-cleavable probe SBL | 26 | 5[5] | 12[5] | 170,000 | Least expensive platform; open source to adapt alternative NGS chemistries | Users are required to maintain and quality control reagents; shortest NGS read lengths | Bacterial genome resequencing for variant discovery | J. Edwards, pers. comm. |
| Helicos BioSciences HeliScope | Frag, MP/ single molecule | RTs | 32* | 8[5] | 37[5] | 999,000 | Non-bias representation of templates for genome and seq-based applications | High error rates compared with other reversible terminator chemistries | Seq-based methods | 91 |
| Pacific Biosciences (target release: 2010) | Frag only/ single molecule | Real-time | 964* | N/A | N/A | N/A | Has the greatest potential for reads exceeding 1 kb | Highest error rates compared with other NGS chemistries | Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks | S. Turner, pers. comm. |

*Average read-lengths. [5]Fragment run. [9]Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLID, support oligonucleotide ligation detection.

*(Michael Metker Nat rev Genetics 2010)*

# Séquençage par shotgun



## Modélisation du processus de séquençage en shotgun

- Lander & Waterman (1988): processus aléatoire d'échantillonnage de:
  - *N* lectures de taille *L*
  - Génome de taille *G*
- Couverture: $a = N\,L\,/\,G$
- Quelle "couverture" est nécessaire pour séquencer un génome complet?

## Modélisation du processus de séquençage en shotgun

- Lander & Waterman (1988): processus aléatoire d'échantillonnage de:
  - *N* lectures de taille *L*
  - Génome de taille *G*
- Couverture: $a = N\,x\,L\,/\,G$
- Nombre de contig obtenu ($N_c$) en fonction de la couverture:

  $N_C = (a \times G / L)\, e^{-a}$
- Taille moyenne des contigs:

  $L_C = (e^a - 1)\, L / a$

## Application

- Génome de la levure
  - G = 12 Mb
  - L = 500 (séquençage Sanger)
- Nombre de contig obtenu ($N_C$): $N_C = (a \times G / L)\, e^{-a}$
- Taille moyenne des contigs: $L_C = (e^a - 1)\, L / a$

| Couverture (*a*) | $N_c$ | $L_c$ |
|---|---|---|
| 1 x | 8829,1 | 860 pb |
| 10 x | 10,8 | 1 101 273 pb |

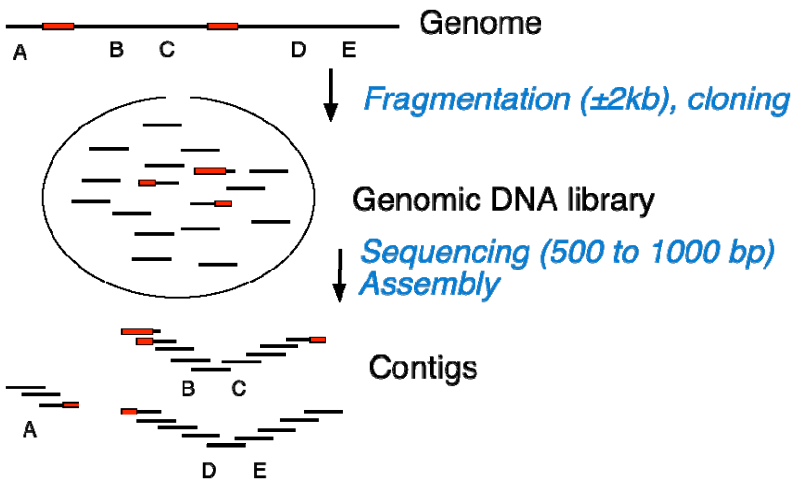
## Prise en compte du chevauchement entre les fragments lors de l'assemblage

- Chevauchement minimal entre 2 lectures ($L_O$):

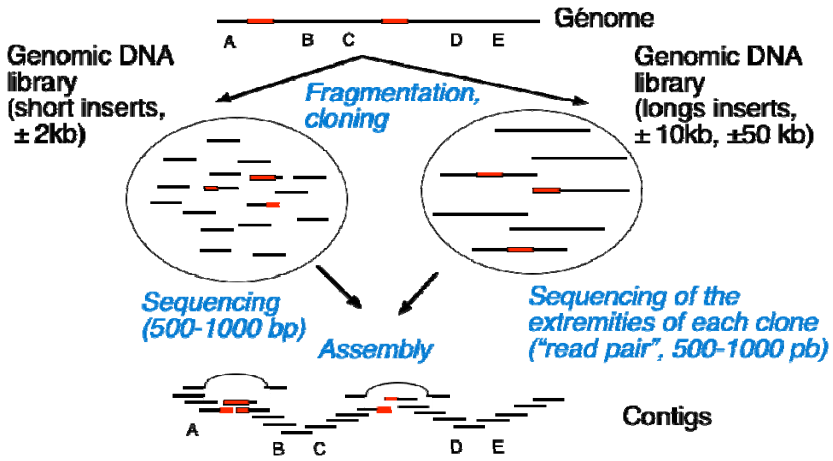
  $O = L_O/L$
- Nombre de contig obtenu ($N_c$) en fonction de la couverture:

  $N_c = (a \times G / L)\, e^{-(1-O)a}$
- En réalité, situation plus complexe
  - Couverture variable le long des génomes (biais techniques liés à la composition des séquences)
  - Séquences répétées

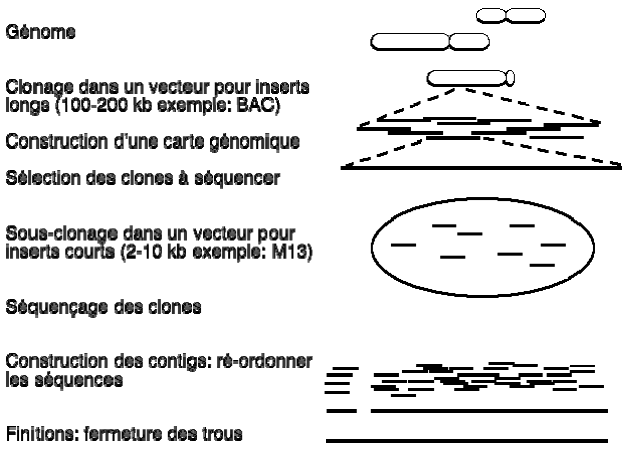
# Séquençage par Shotgun: le problème des répétitions



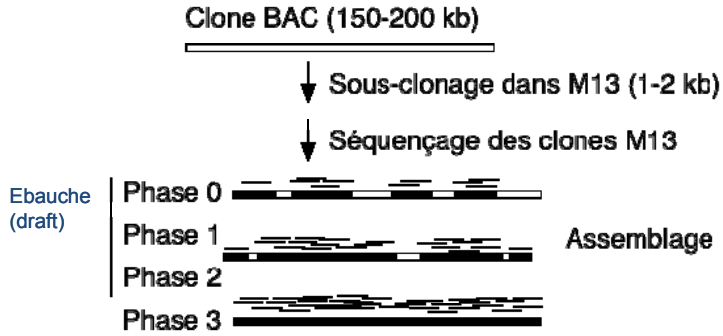

# "Paired-end sequencing"




# Stratégie de séquençage du génome humain: séquençage hiérarchisé




# Etapes du séquençage génomique




Phase 0-1: séquence non-terminée; contigs non-ordonnés, non-orientés; gaps
Phase 2: séquence non-terminée; contigs ordonnés, orientés; gaps
Phase 3: séquence terminée

Phase 0-2: séquences mentionnées HTG (High Throughput Genome sequences) dans les banques de données

# Assemblage: des séquences ... à **LA** séquence

- Séquençage hiérarchisé:
  1- Séquençage de BAC (shotgun)
  2- Assemblage des BAC

- Problèmes:
  - Certaines régions du génome ne sont pas clonables
  - Certaines régions sont très riches en séquences répétées (e.g. régions centromériques)
  - Il existe de grandes régions dupliquées dans le génome

⇒ impossible d'assembler le génome humain uniquement à partir de séquences fragmentées

⇒ utilisation de cartes génomiques

# Cartographie des génomes

- Cartographie génétique

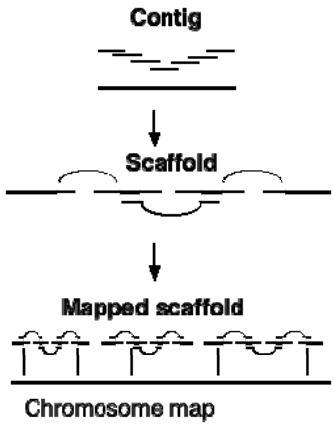

- Cartographie physique
  - Cartographie d'enzymes de restriction
  - Hybrides de radiation

- 1992: carte génétique du génome humain (30,000 marqueurs microsatellite – Genethon)

- 1993: carte physique haute densité

# Assemblage: des séquences ... à LA séquence

- **Contig**: ensemble de séquences chevauchantes, sans gap
- **Scaffold**: ensemble de contigs orientés et ordonnés; gaps de longueur connue
- **Mapped scaffold**: ensemble de scaffolds localisés le long des chromosomes (mais pas toujours ordonnés ou orientés; gaps de longueur inconnue)
- Scaffolds ordonnés et orientés
- 2002: 97% du génome humain séquencé, mais seulement 85% assemblé dans des scaffolds




# Assemblage: des séquences ... à LA séquence

- 2004: séquençage et assemblage "terminés" (*IHGSC, 2004, Nature 431:931-45*)
  - 99% du génome humain séquencé (99% de la fraction euchromatine uniquement)
  - 341 gaps (souvent à proximité de duplications segmentaires)
  - Il manque l'hétérochromatine (environ 6% du génome): 33 gaps
    - 24 centromères (50 Mb) (ADN satellite)
    - 3 constrictions secondaires (proches de centromères)
    - 5 bras courts chromosomes acrocentriques (rDNA, ~50 copies 43 kb sur chaque bras, + autres répétitions)
    - Grande région chromosome Y

- !! Génome entièrement séquencé ≠ génome séquencé et assemblé

# "Le" génome humain?

- Séquence de référence publiée en 2004 = mosaïque de 5 individus (anonymes, différentes origines éthniques, des 2 sexes)
- Il existe du polymorphisme au sein des populations:
  - SNP: single nucleotide polymorphism
  - Petits indels, microsatellites
  - Eléments transposables
  - CNV: copy number variations
  - Réarrangements chromosomiques
- NB: le polymorphisme contribue aux difficultés d'assemblage (il n'existe pas un génome, mais des génomes humains)

# Les génomes humains: "personal genome projects"

- Oct. 2007: 1er séquence du génome d'un individu (Craig Venter)
  - NB: organisme diploïde => polymorphisme
  - Sanger sequencing
- Avril 2008: James Watson (454-FLX)
- Nov. 2008:
  - Un individu asiatique
  - Un individu nigérien (yorubé) - Solexa
  - Un génome d'un individu atteint de cancer
- 2008-2011: projet 1000 génomes

# Changements d'échelle: nouvelles méthodes de séquençage

QUICKER, SMALLER, CHEAPER

| Genome sequenced (publication year) | HGP (2003) | Venter (2007) | Watson (2008) |
|---|---|---|---|
| Time taken (start to finish) | 13 years | 4 years | 4.5 months |
| Number of scientists listed as authors | > 2,800 | 31 | 27 |
| Cost of sequencing (start to finish) | $2.7 billion | $100 million | < $1.5 million |
| Coverage | 8–10 × | 7.5 × | 7.4 × |
| Number of institutes involved | 16 | 5 | 2 |
| Number of countries involved | 6 | 3 | 1 |

2008: Yorubé: 0,25 millions $ (8 semaines) Solexa

# Changements d'échelle: nouvelles méthodes de séquençage

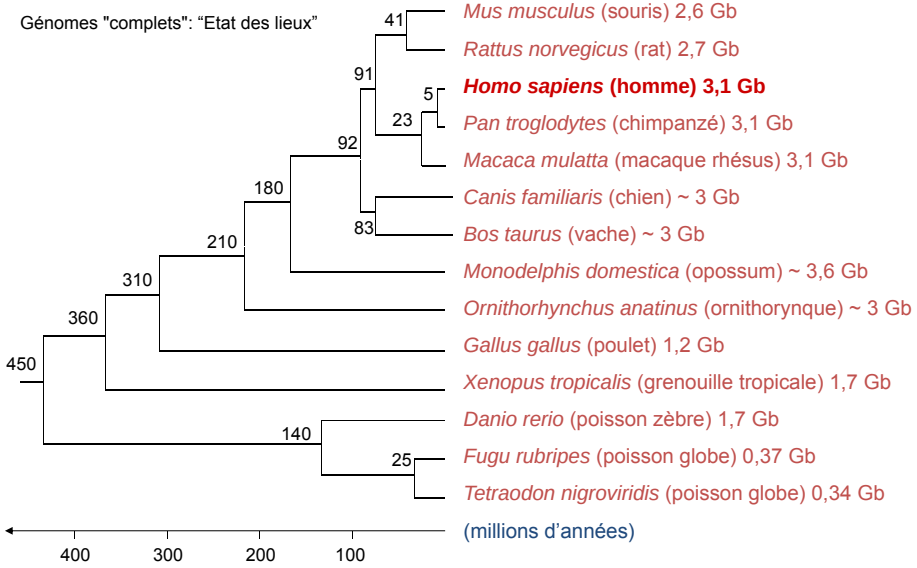
Table 1.
A comparison of new-generation DNA sequencing platform

| Features | Second-generation sequencers | | | Third-generation sequencers (single molecule-SBS) | | | |
|---|---|---|---|---|---|---|---|
| | 454-FLX | Solexa | SOLiD | Helicos tSMS | PacBio SMRT | Nanopore and modified forms | ZS Genetics TEM |
| Read-length (bp) | 240–400 | 35 | 35 | 30 | 100 000 | Potentially unlimited? | Potentially unlimited? |
| Cost/human genome (US$) | 1 000 000 | 60 000 | 60 000 | 70 000 | Low | Low | Low |
| Run time (h/Gb) | 75 | 56 | 42 | ~12 | <1 | >20 | ~14 |
| Ease of use | Difficult | Difficult | Difficult | Easy | Easy | Easy | Easy |

Gupta (2008) Trends Biotechnol 26:602-11.

- 2008: compromis coût/temps/longueur des lectures
- 2014: un génome humain pour 1000 $ ??

Génomes "complets": “Etat des lieux”




*Mus musculus* (souris) 2,6 Gb

*Rattus norvegicus* (rat) 2,7 Gb

***Homo sapiens* (homme) 3,1 Gb**

*Pan troglodytes* (chimpanzé) 3,1 Gb

*Macaca mulatta* (macaque rhésus) 3,1 Gb

*Canis familiaris* (chien) ~ 3 Gb

*Bos taurus* (vache) ~ 3 Gb

*Monodelphis domestica* (opossum) ~ 3,6 Gb

*Ornithorhynchus anatinus* (ornithorynque) ~ 3 Gb

*Gallus gallus* (poulet) 1,2 Gb

*Xenopus tropicalis* (grenouille tropicale) 1,7 Gb

*Danio rerio* (poisson zèbre) 1,7 Gb

*Fugu rubripes* (poisson globe) 0,37 Gb

*Tetraodon nigroviridis* (poisson globe) 0,34 Gb

**Séquençage complet, assemblage terminé**

Séquençage presque complet; version préliminaire de l’assemblage disponible

+ 15 génomes de mammifères en faible couverture (2x): non assemblés