

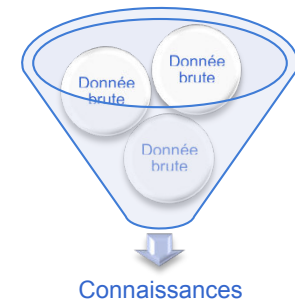
Bioinformatique et données biologiques

Cours d'introduction à la bioinformatique et de présentation des banques de séquences.

1^{ère} partie

Equipe Bonsai (2014)

QUELQUES MOTS SUR LA BIOINFO



Définition de la bioinformatique

Un domaine de recherche qui analyse et interprète des données biologiques, au moyen de méthodes informatiques, afin de créer de nouvelles connaissances en biologie.

Source : article présentant la bioinformatique, sur le site d'*Interstices*

Auteur(s) :

Isabelle Quinkal (Journaliste)

François Rechenmann (Chercheur)

Définition de la bioinformatique

en anglais : distinction entre
« *Bioinformatics* » et « *Computational Biology* »

- « *Bioinformatics* »

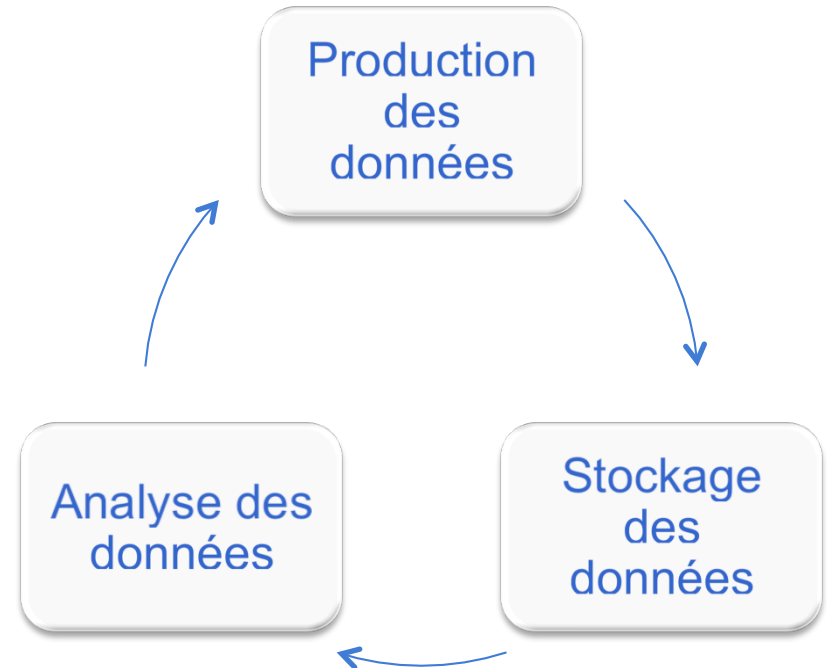
applique des algorithmes, modèles statistiques dans l'objectif d'interpréter, classer et comprendre des données biologiques.

- « *Computational Biology* »

développer des modèles mathématiques et outils associés pour résoudre des problèmes biologiques.

Qu'est-ce que la bioinformatique ?

- L'approche *in silico* de la biologie
- Trois activités principales :
 - Acquisition et organisation des données biologiques
 - Conception de logiciels pour l'analyse, la comparaison et la modélisation des données
 - Analyse des résultats produits par les logiciels



Quelques conseils

- Méfiez-vous des résultats donnés par les logiciels :
 - La qualité des résultats est parfois diminuée au profit de la rapidité
 - Certains problèmes admettent un ensemble infini de possibilités
 - Ce n'est pas toujours la solution la meilleure qui est trouvée
 - Beaucoup de logiciels ne font que de la prédiction
 - Prédiction : dire ce qu'on prévoit, par raisonnement, devoir arriver. (wiktionnaire)
- Méfiez-vous des banques de données :
 - Les données se sont pas toujours fiables
 - La mise à jour n'est pas toujours récente



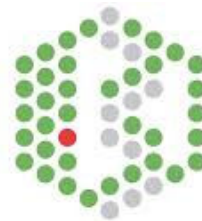
La réalité mathématique n'est pas la réalité biologique :

Les ordinateurs ne font pas de biologie, ils calculent ... vite !

En Europe : EBI

- European Bioinformatics Institute
<http://www.ebi.ac.uk/>
- Organisation académique à but non lucratif fondée en 92
- Centre de recherche et services en bioinformatique qui gère des banques de données biologiques (ADN-ARN, protéines, structures 3D)
- Met dans le domaine public et rend accessible gratuitement les informations issues de la recherche en biologie moléculaire et génomique afin de promouvoir le progrès scientifique

EMBL-EBI



Aux États-Unis d'Amérique : NCBI

- National Center for Biotechnology Information
<http://www.ncbi.nlm.nih.gov/>
- Ressource nationale pour l'information en biologie moléculaire fondée en 1988
- Création de banques publiques et recherche en bioinformatique
- Développe des outils informatiques pour analyser les données de génome et diffuser l'information médicale pour mieux comprendre les processus moléculaires touchant la santé humaine et la maladie

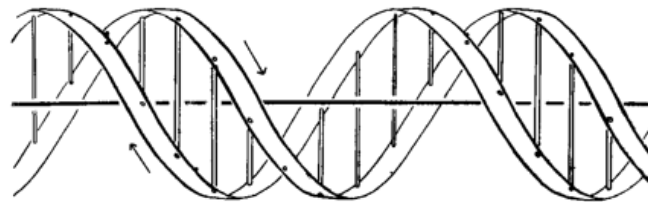


Comment s'assurer de la qualité de l'information ?

- Autorité :
 - Source de l'information, auteurs, statut, ...
- Péremption :
 - Date de création, de mise à jour, ...
 - Attention, ce qui est validé un jour peut être démenti par la suite !
- Transparence :
 - Documentation disponible
- Règles valables aussi bien pour une banque de données, que pour un logiciel, un site web, ...



GÉNOMIQUE ET BIOINFORMATIQUE



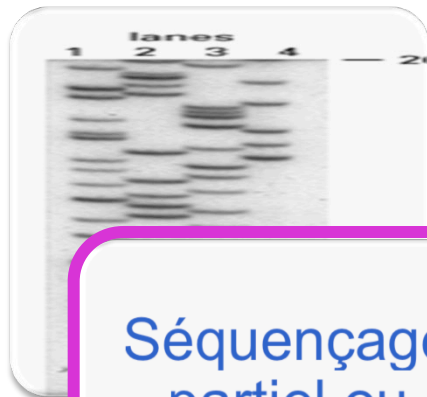
La génomique

- Etude des génomes et de l'ensemble de leurs gènes

- La structure
- Le fonctionnement
- L'évolution
- Le polymorphisme, ...

Nécessite des outils bioinformatiques

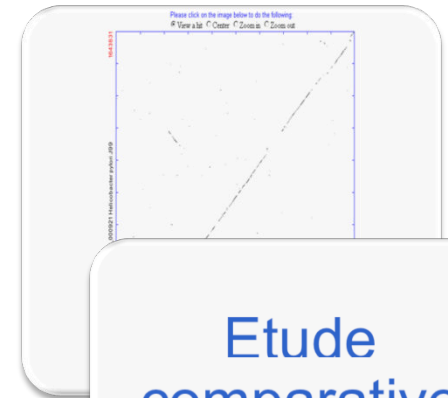
- Plusieurs étapes :



Séquençage
partiel ou
total d'un
génom



Etude et
annotation de
la séquence
du génome



Etude
comparative
de plusieurs
génom

Chronologie sur le séquençage de l'ADN

1^{er} gène ARN
par W. Fiers
et al.

1972

Technique de
Maxam-Gilbert
pour l'ADN

1975

Technique de
F. Sanger *et al.*
pour l'ADN

1977

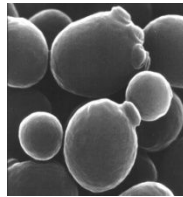
1^{er} virus phi
X174 par
Sanger *et al.*

1977

1^{er}
séquenceur
Applied
Biosystems

1987

1^{ère} bactérie
H. influenzae
1,83 Mb

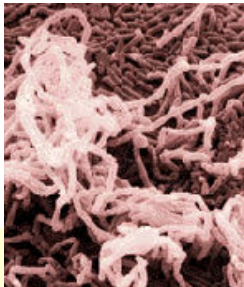


1^{er} pluricellulaire
C. elegans
100 Mb



Séquençage
massif et
parallèle

1995



1^{er} eucaryote
S. cerevisiae
12 Mb

1996

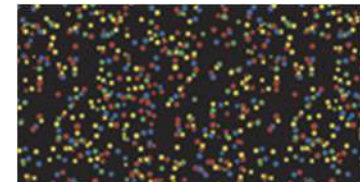
1998



2001

*Homo
sapiens*

2008



Bilan des projets « génomes » en 2014

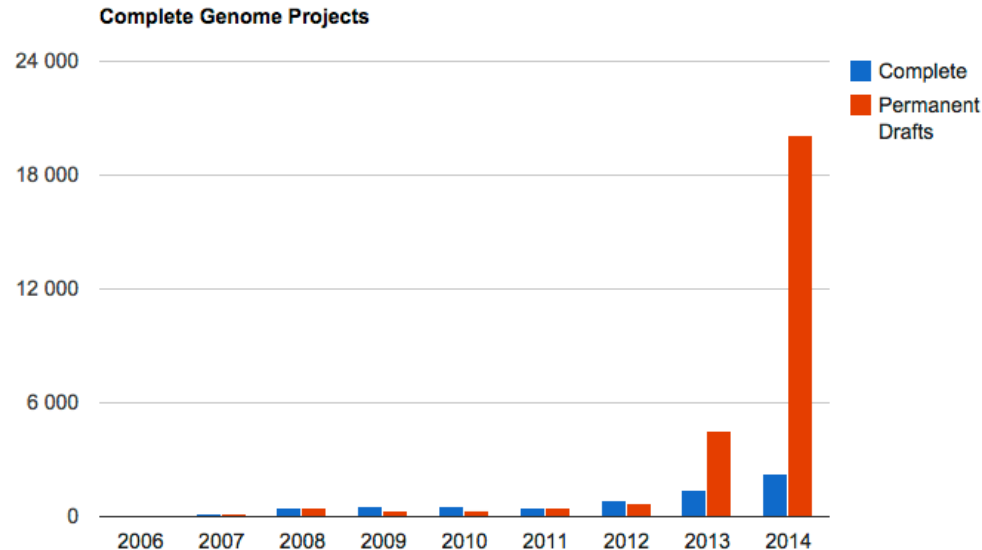
■ Genome Online Database

<http://www.genomesonline.org/>

■ Composition (projets au 1^{er} septembre)

- **6576** génomes complets +
- **22576** drafts dit « *permanents* » +
- **21020** drafts + **1007** targeted

GENOMES

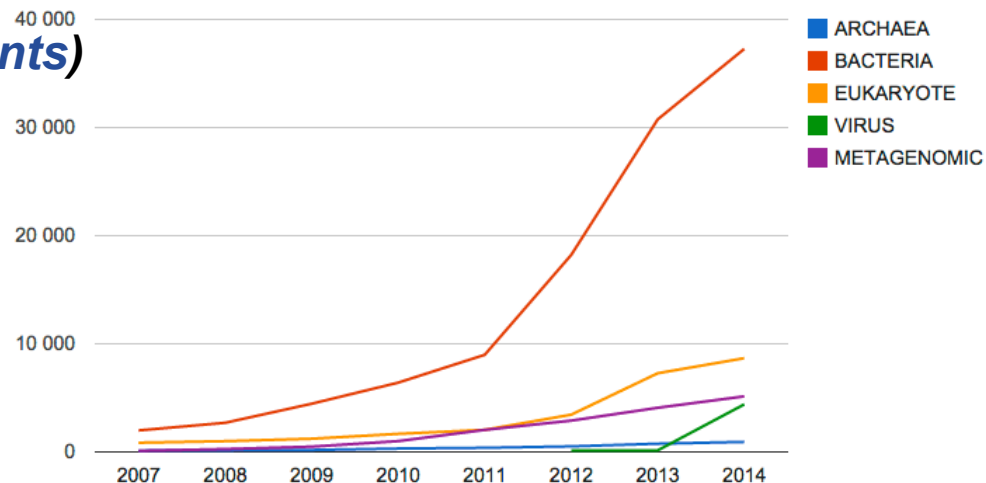


■ Distribution (organismes au 1^{er} septembre)

globale (complet + draft permanents)

- **37272 (24214)** eubactéries
- **926 (621)** archaebactéries
- **8667 (4382)** eucaryotes
- +
- **5373 (4841)** métagénomes

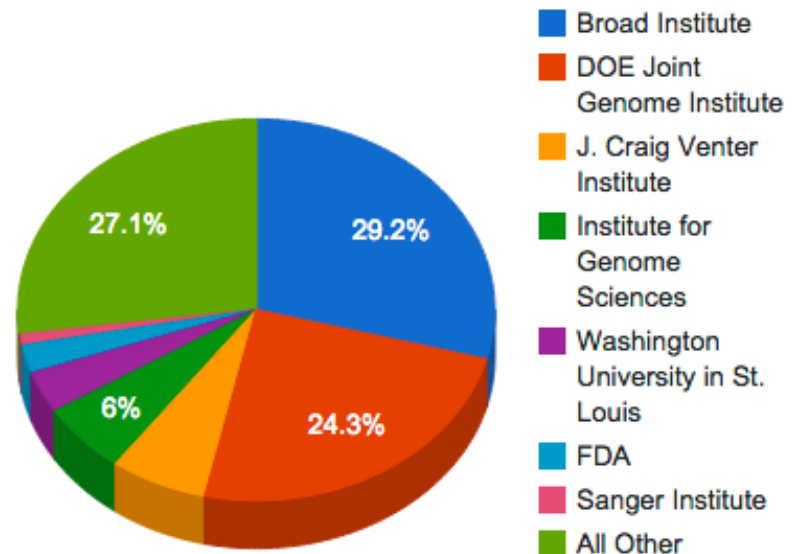
Projects by Domain



Les différents contextes de séquençage

- Séquences produites par des laboratoires pour étudier un gène, un groupe de gènes, une séquence intergénique, ...
 - Régions d'intérêts dont le génome complet n'est (n'était) pas connu
 - Etude des variations alléliques, ...
- Séquences produites par des centres de séquençage
 - Génomes complets (HTG, WGS) ou partiels (GSS)
 - STS
 - EST
 - Métagénomes

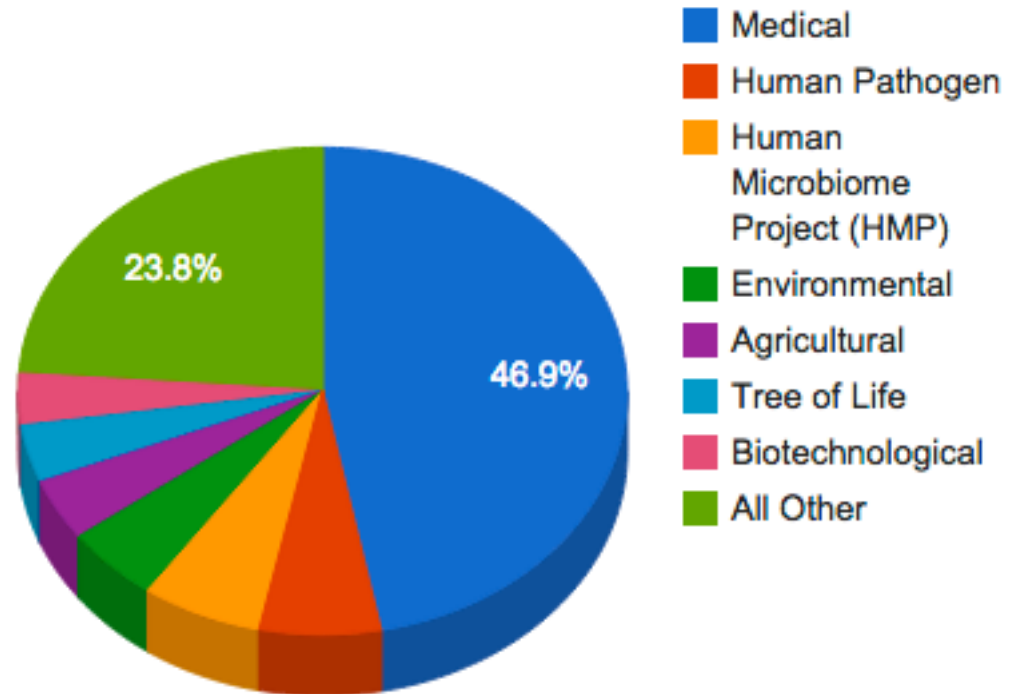
Projects by Sequencing Center



Pourquoi séquencer les génomes ?

- Intérêt économique
 - Médecine
 - Biotechnologies
 - Environnement
- Intérêt scientifique
 - Evolution des espèces
 - Fonctionnement des cellules
 - Etude des êtres vivants
- Utilité publique
 - Nutrition
 - Propagation des maladies
 - Environnement

Project Relevance of Bacterial Projects



Les méthodes de séquençage

- Méthode Sanger (1975)
- Méthode Maxam–Gilbert (1977)
- Automatisation de Sanger (de ~1980 à 2005)
 - Commercialisée en 1987 : premier séquenceur *Applied Biosystems 370A*
- Nouvelles Générations de Séquenceurs (depuis 2005)
 - NGS : *Next Generation Sequencing* (*désormais largement utilisés*) ou plutôt
 - HTS : *High-Throughput Sequencing*
- NNGS : *Next-Next Generation Sequencing* (en cours):
 - en particulier technologie SMS (*Single Molecule Sequencing*)

Séquençage : méthode Sanger (1975)

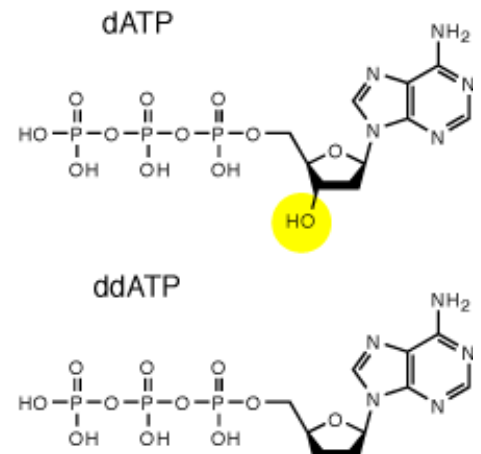
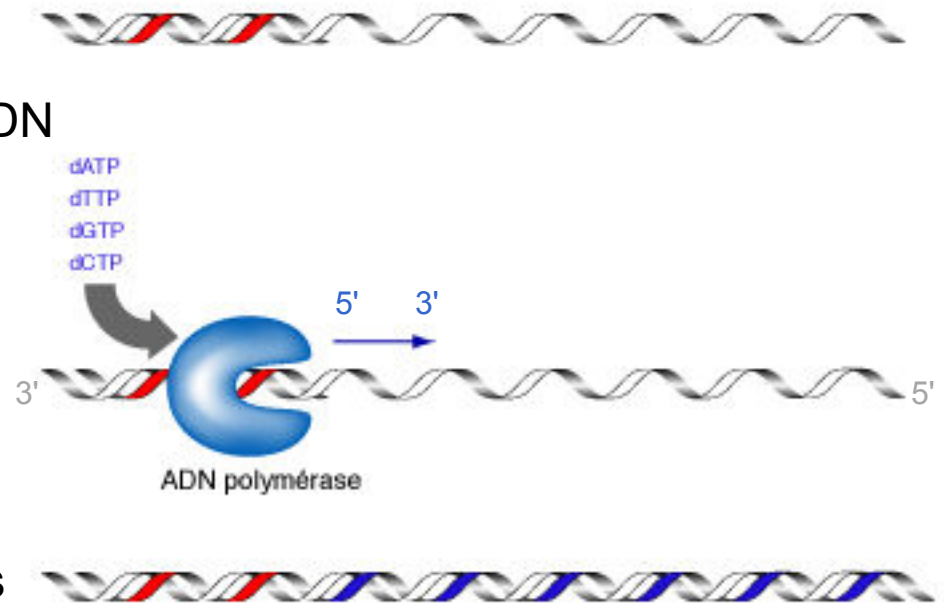
■ Idée

Amorcer une polymérisation de l'ADN

■ Elongation

- faite à l'aide de 4 désoxyribonucléotides (dATP, dCTP, dGTP, dTTP) majoritaires
- + faible concentration de **l'un des quatre** didésoxyribonucléotides (ddATP, ddCTP, ddGTP ou ddTTP) qui arrêtent l'élongation.

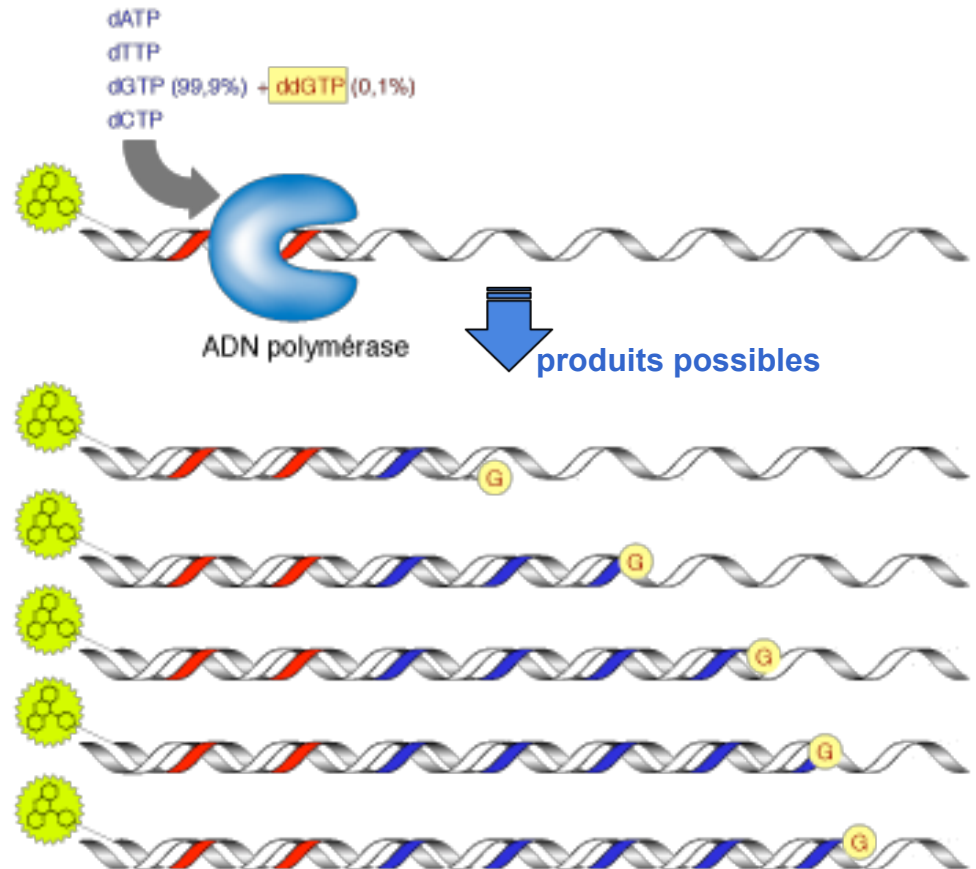
Note : il y a **4** expériences



Séquençage : méthode Sanger (1975)

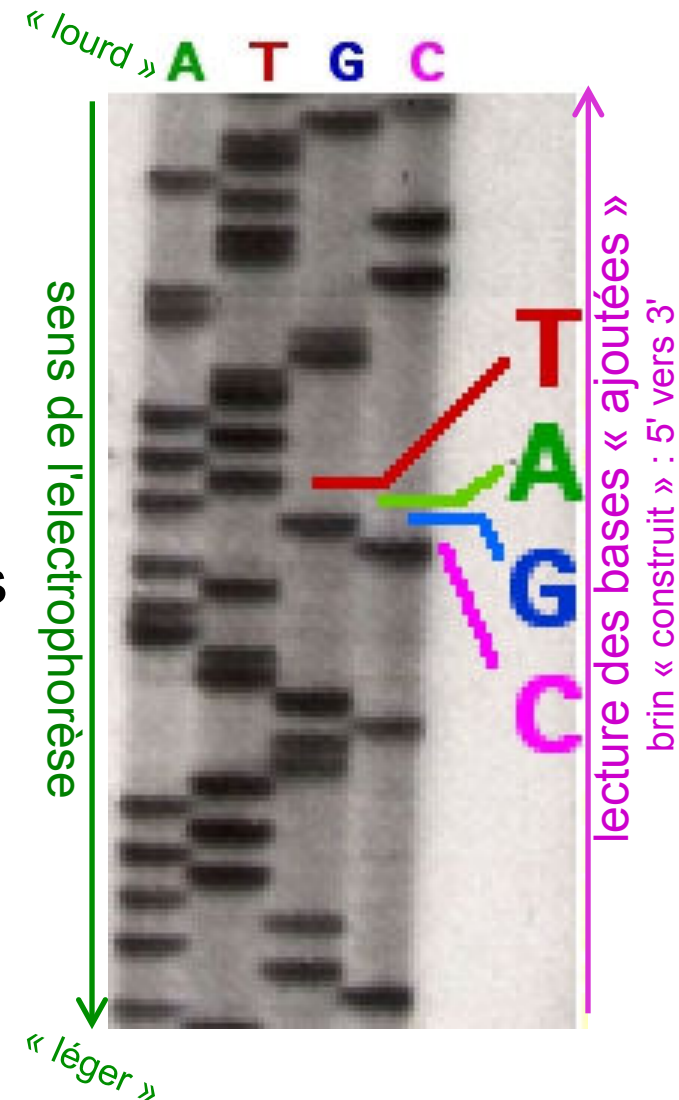
Exemple: expérience *ddGTP*

- Elongation statistique
 - Continue tant que des *d«N»TP* sont incorporés ...
« N »={A,C,G,T}
 - Arrêt si incorporation (par hasard ...) d'un *ddGTP*
 - Le hasard « dépend » ici de la concentration respective des *d«N»TP* et de *ddGTP*
- Tous les produits « normaux » terminent forcément par un **G**

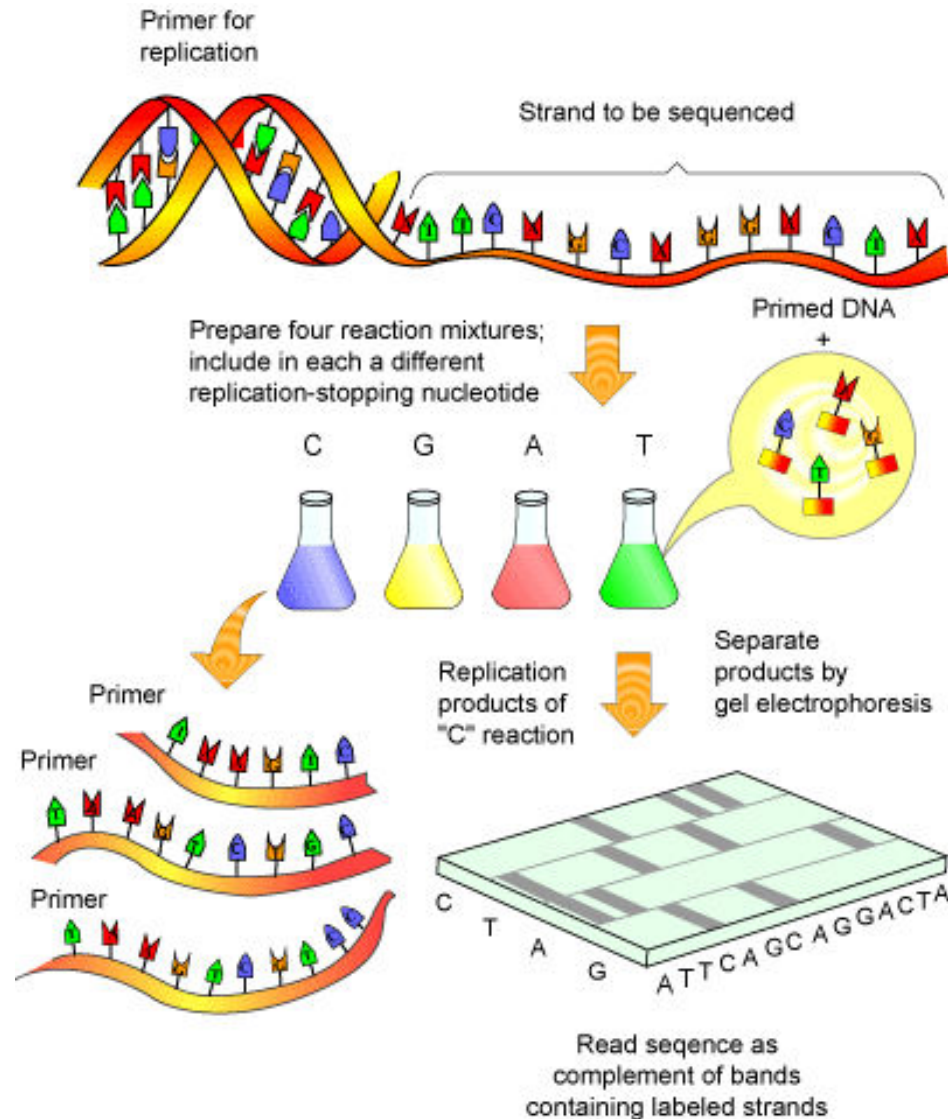


Séquençage : méthode Sanger (1975)

- Electrophorèse sur gel
 - Réalisé sur les quatre expériences en même temps.
 - Migration en fonction du poids des produits des 4 expériences...
- et lecture (manuelle) des bases ajoutées

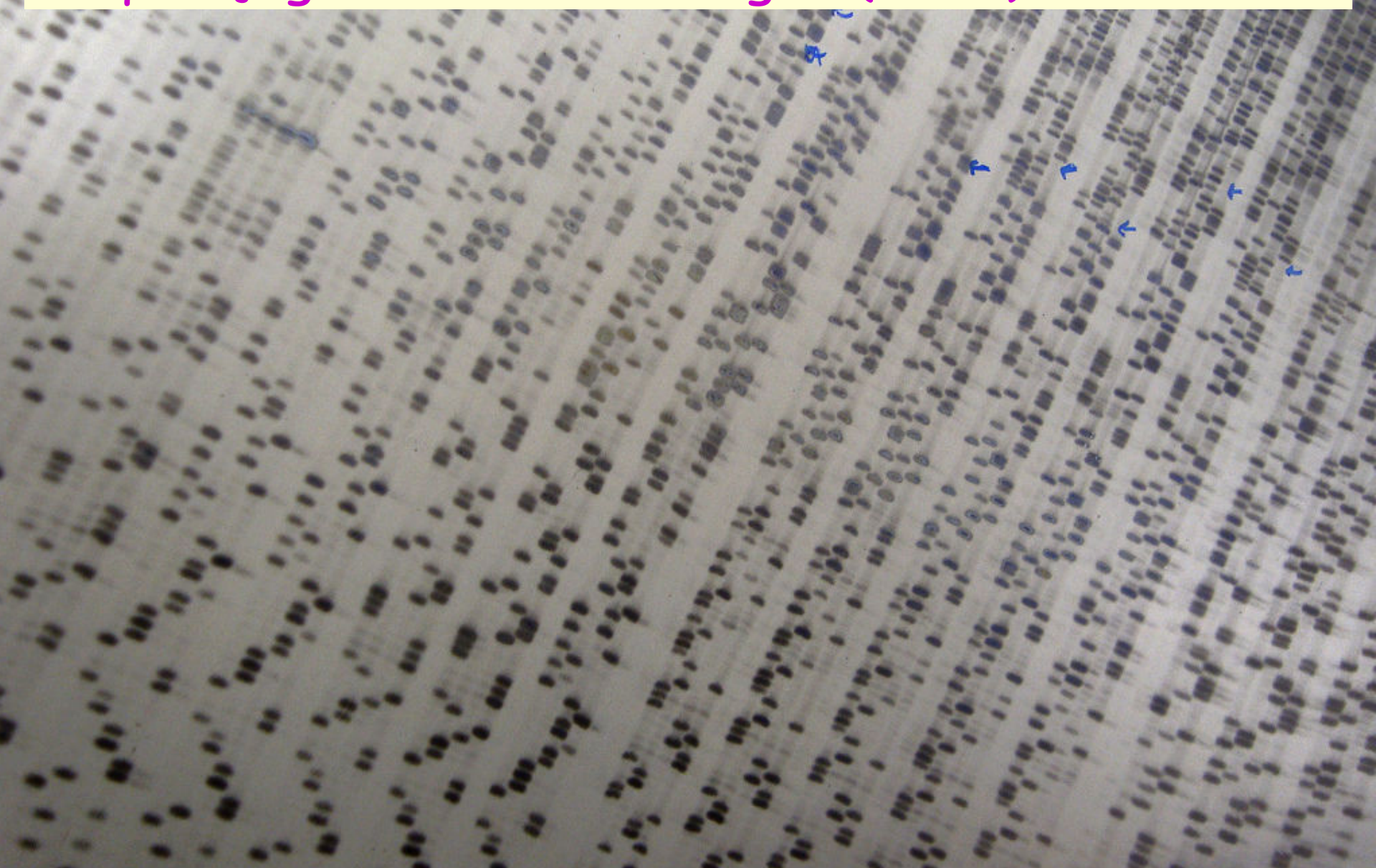


Séquençage : méthode Sanger (1975)



En résumé :

Séquençage : méthode Sanger (1975)



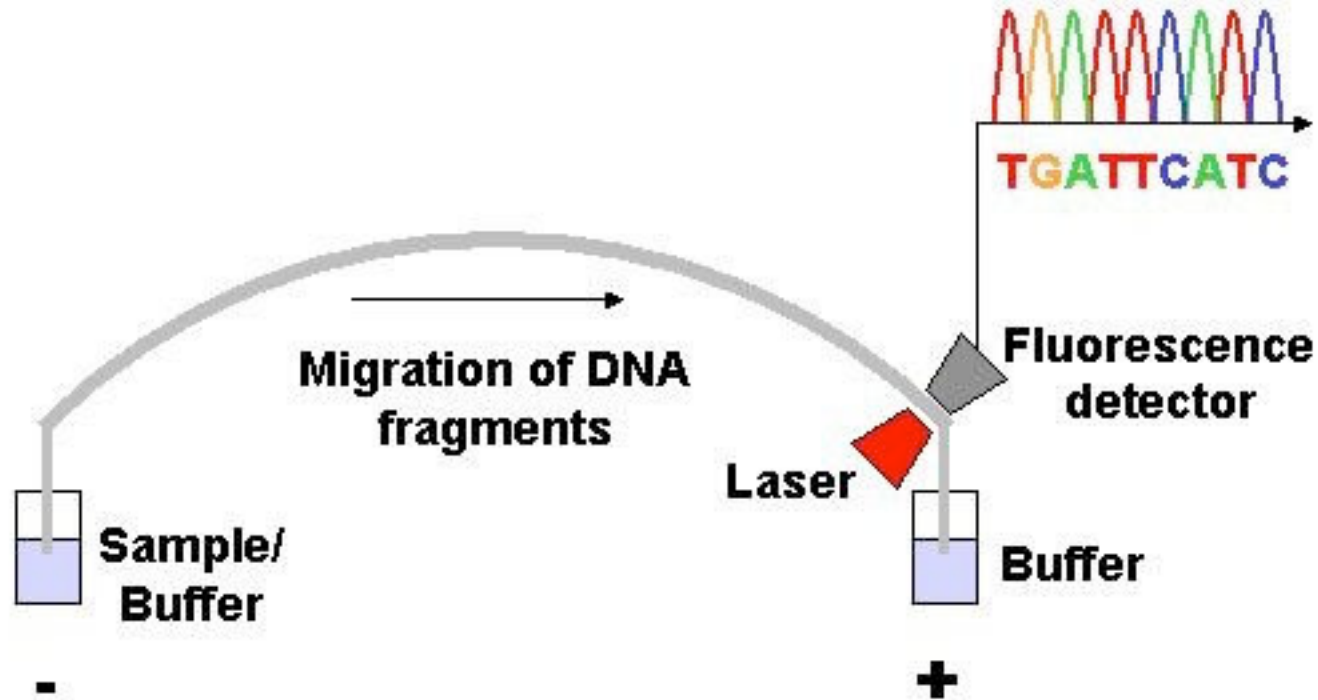
Séquençage : extension de la méthode Sanger ...

- Méthode Sanger avec « Dye terminator sequencing »

*An alternative to the labelling of the primer is to label the terminators instead, commonly called 'dye terminator sequencing'. The major advantage of this approach is the complete sequencing set can be performed **in a single reaction**, rather than the four needed with the labeled-primer approach. This is accomplished by **labelling each of the dideoxynucleotide** chain-terminators with a **separate fluorescent dye**, which fluoresces at **a different wavelength**.*

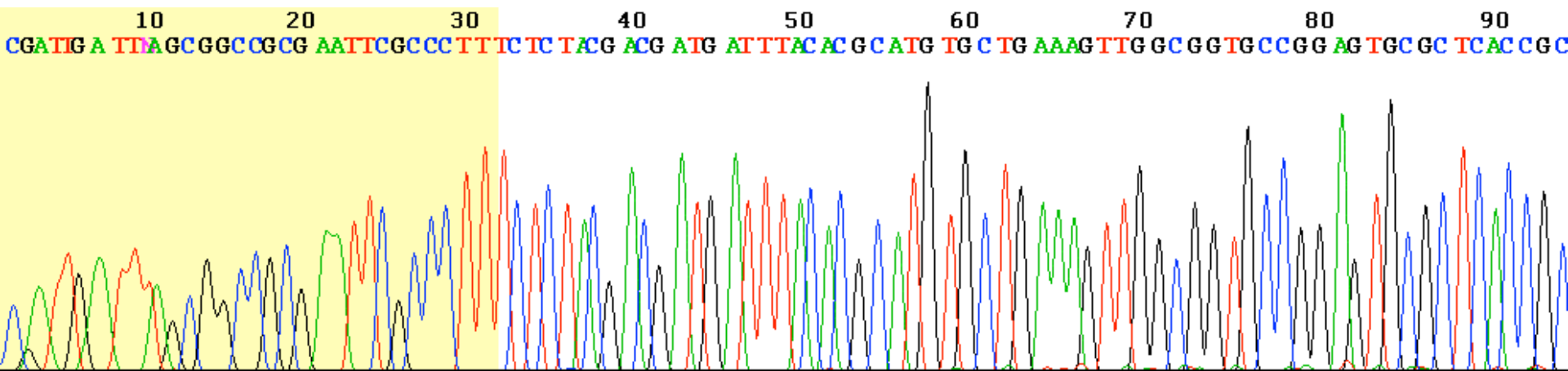
Séquençage : ... et automatisation ...

- Electrophorèse Capillaire
- Excitation à l'aide d'un laser, et lecture automatique des 4 longueurs d'onde possibles (associés au 4 ddNTP)



Séquençage : ... et automatisation ...

- Exemple de lecture Sanger « Automatisée »
(début de lecture) ...



Séquençage : 1er séquenceur automatique (1987)



APPLIED BIOSYSTEMS 370A DNA SEQUENCER

Item condition: --

Time left: 11d 16h (Mar 09, 2010 19:24:14 PST)

Price: **US \$400.00**

Buy It Now

or

Best Offer:

Make Offer

Watch this item

Shipping: Freight - See shipping details | [See all details](#)

Estimated delivery time varies for freight shipping.

Returns: No Returns Accepted

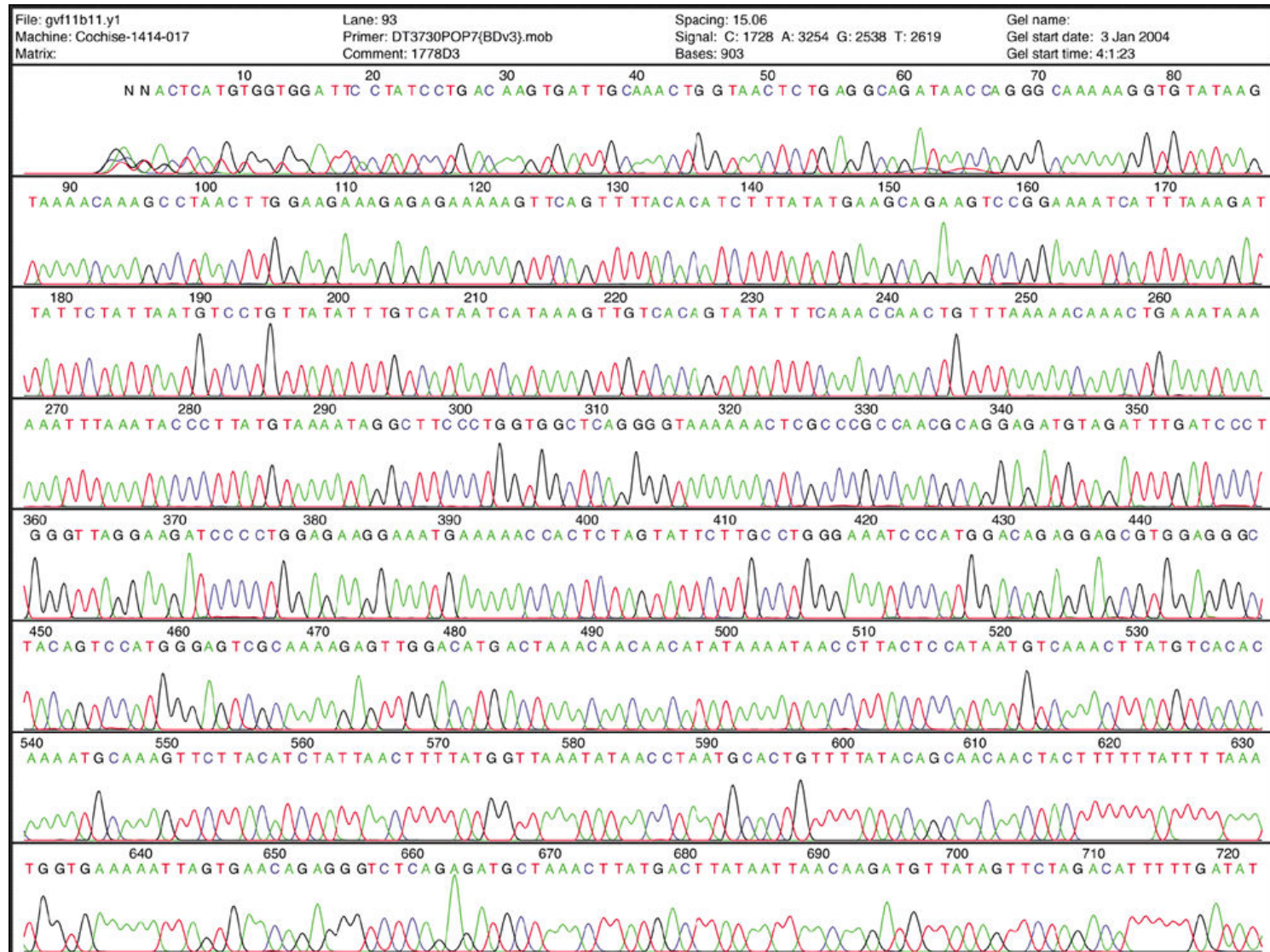
... désormais collector (1987) ...

Séquençage : évolution des modèles (1990-2000)

Voir : <http://www.biology.iupui.edu/biocourses/biol540/14genome2k6.html>



Séquençage : exemple de lecture "actuelle" (2004)



NGS : Next Generation Sequencing (>2005)



NGS : Next Generation Sequencing

ou « high-throughput sequencing »

Nouvelles technologies de séquençage à **Haut Débit**

- **Récentes:**
 - 1ere commercialisé en 2005 (actuellement **Roche 454**),
 - Depuis x autres ont suivi (**Illumina Solexa**, **Applied Biosystems SOLiD** [moribond], **Ion torrent**, **Pacbio**, ...)
- **Rapides:**
 - ~ 3 jours au lieu de 3 mois
 - Coût initial + production en baisse régulière
ex: 1000 génomes humains à « 1000\$ »
- **Reads (Lectures) plus courts (pour le moment) :**
 - taux d'erreur *actuellement* plus élevé => reads plus courts

NGS : Next Generation Sequencing

■ **Haut Débit :**

séquençage de ~~milliers~~ millions de « reads » en parallèle

- **Read** = « lecture » de l'ordre de ~100 à ~400 bases.
- **Reads** = comment sont-ils obtenus ?? principe général simplifié :
 - chaque lecture d'une lettre génère un point de couleur à une position donnée sur une « image »
 - une suite d'images lue donne une suite de couleurs, et (selon un code) une suite de nucléotides ...

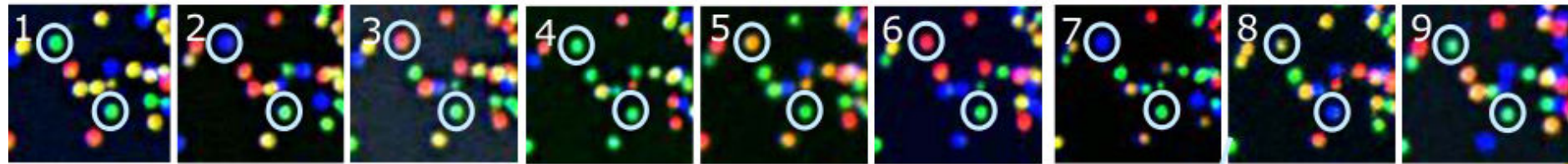
[voir exemple sur slide suivant]

■ **Avantage :**

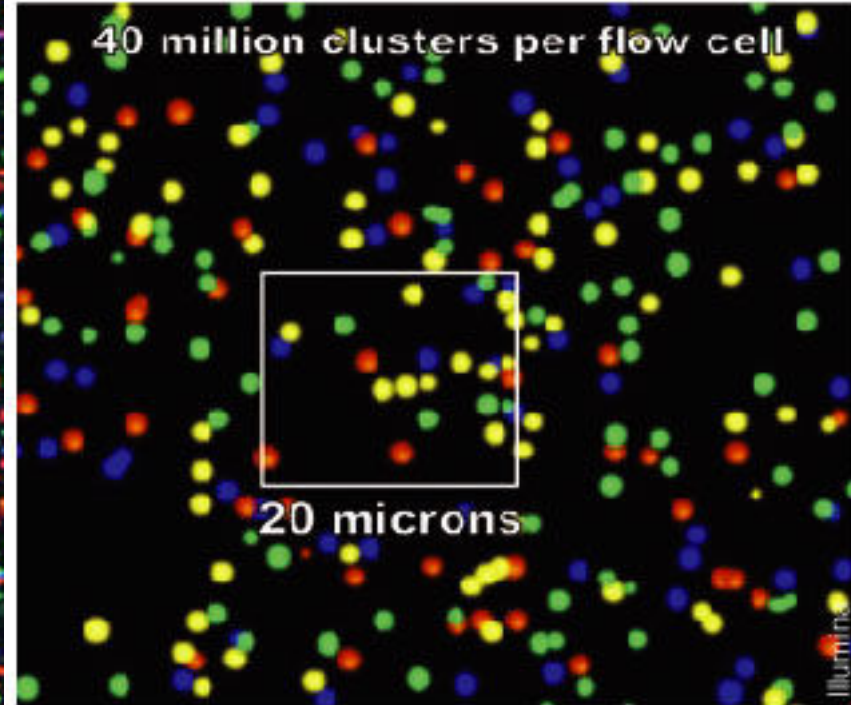
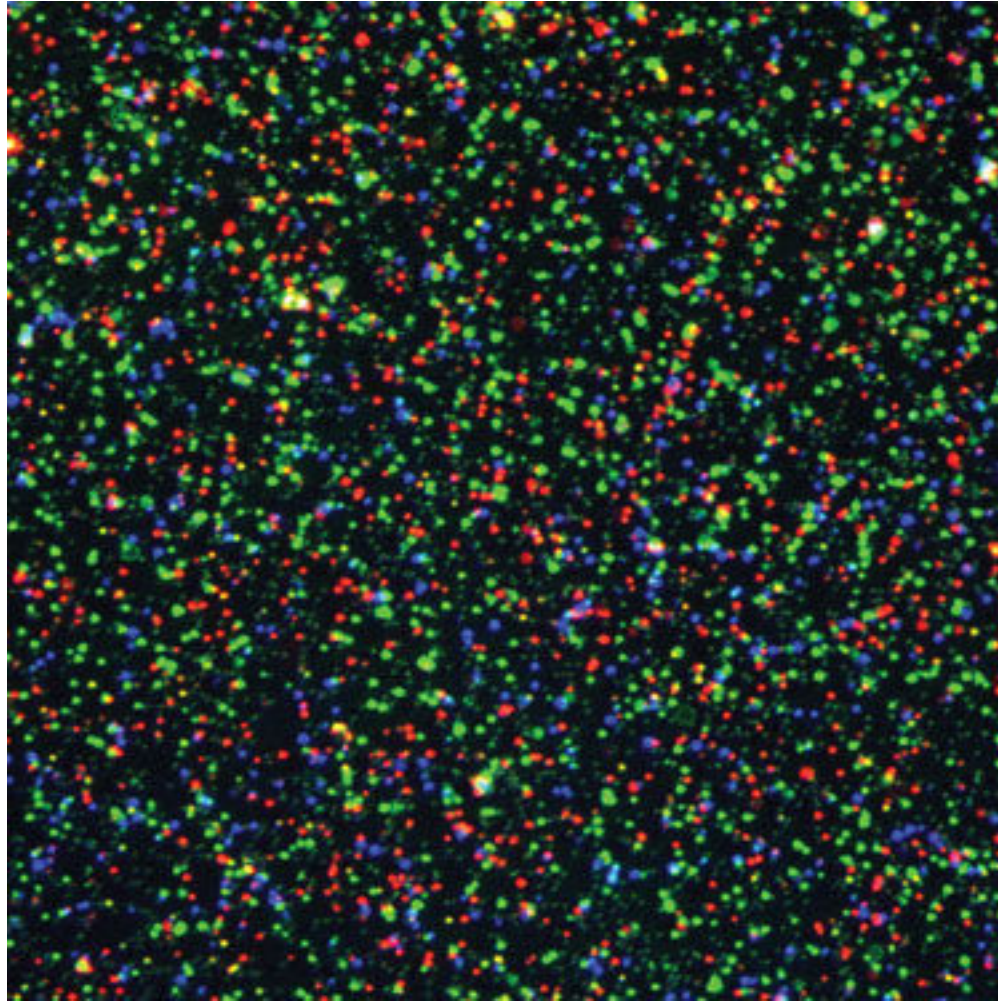
- Génère des centaines de ~~milliers~~ millions de lectures en parallèle (dépend de la densité en points colorés)

NGS : principe (exemple sur Illumina-Solexa)

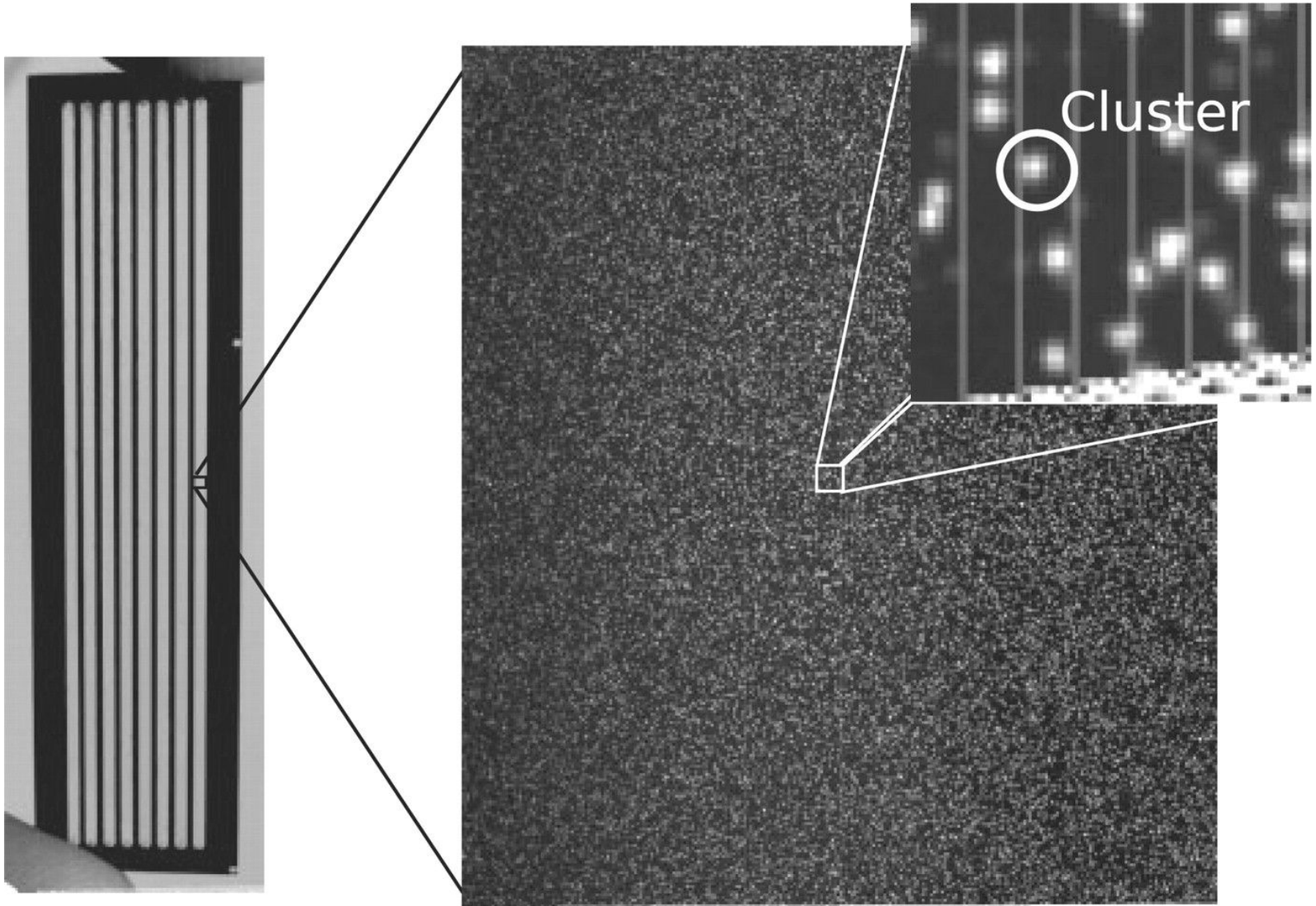
T G C T A C G A T ...



NGS : principe (exemple: Illumina-Solexa)



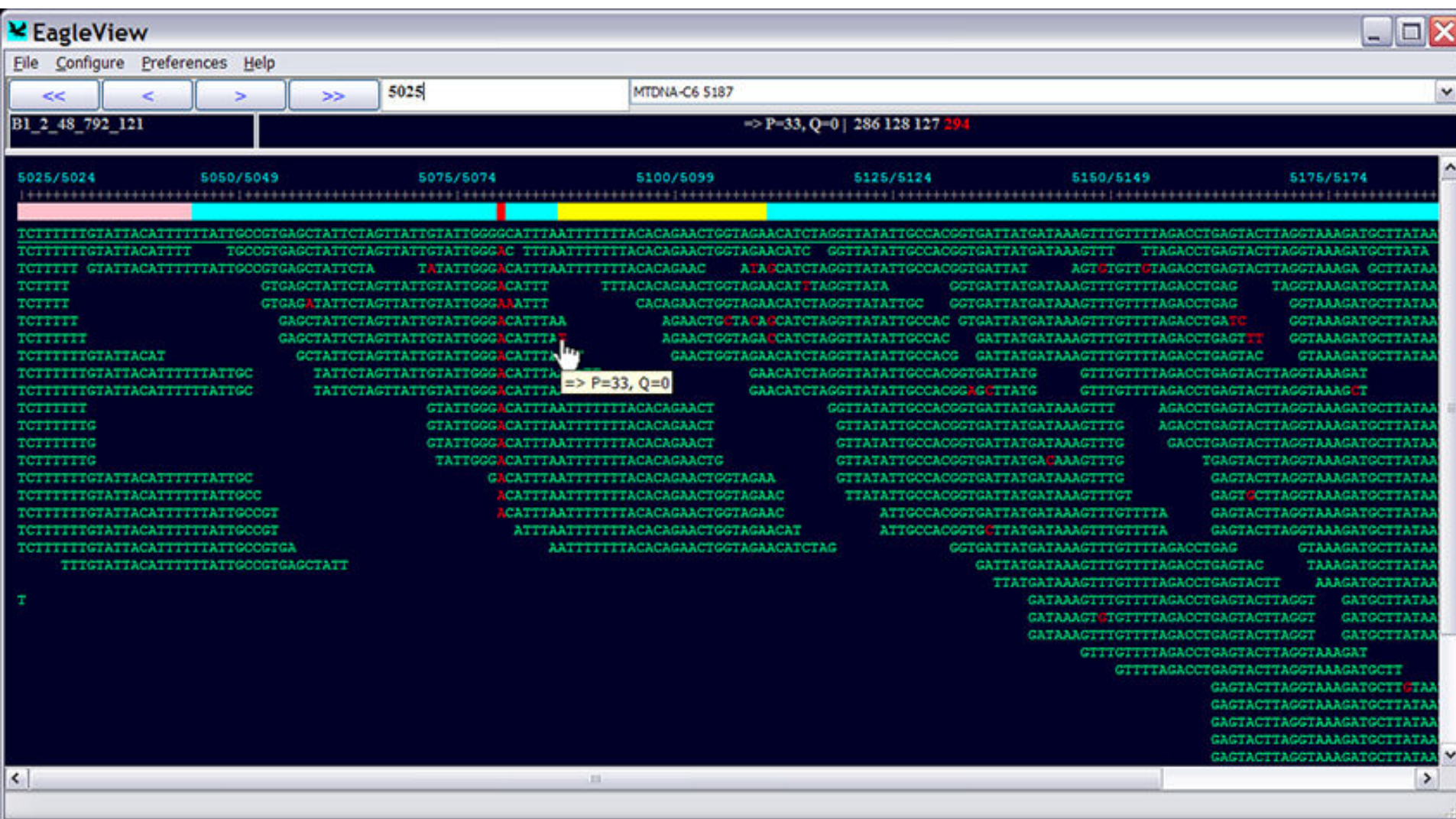
NGS : principe (exemple: Illumina-Solexa GA2)



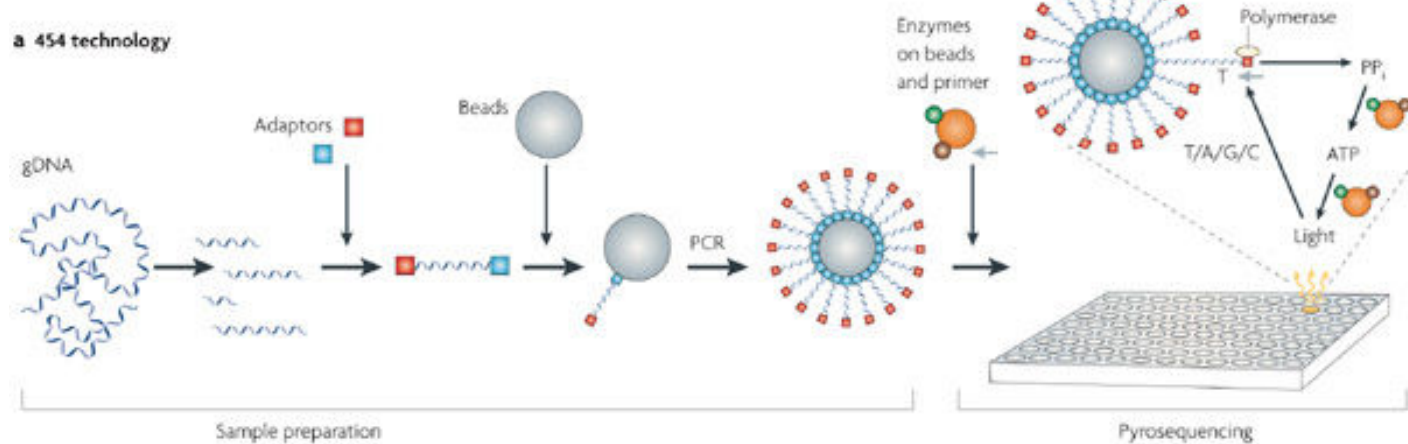
GACCAGATAGAAACCCAGATAGACCCATAGTACTGTACATCCAGATAGAAATGGCTAGGTA

NGS : reads (remappés sur génome connu)

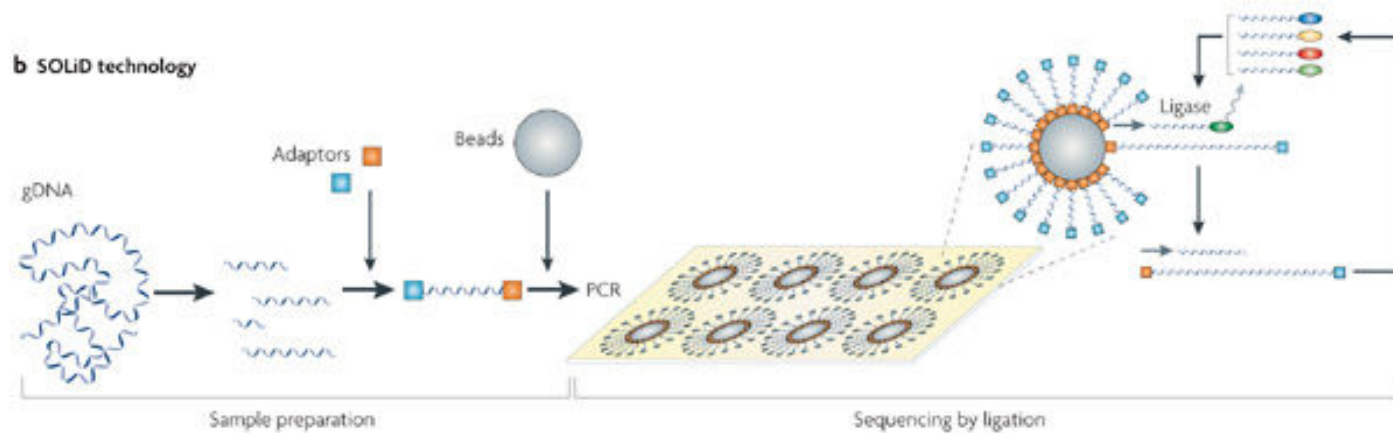
(des erreurs de lecture + 1SNP)



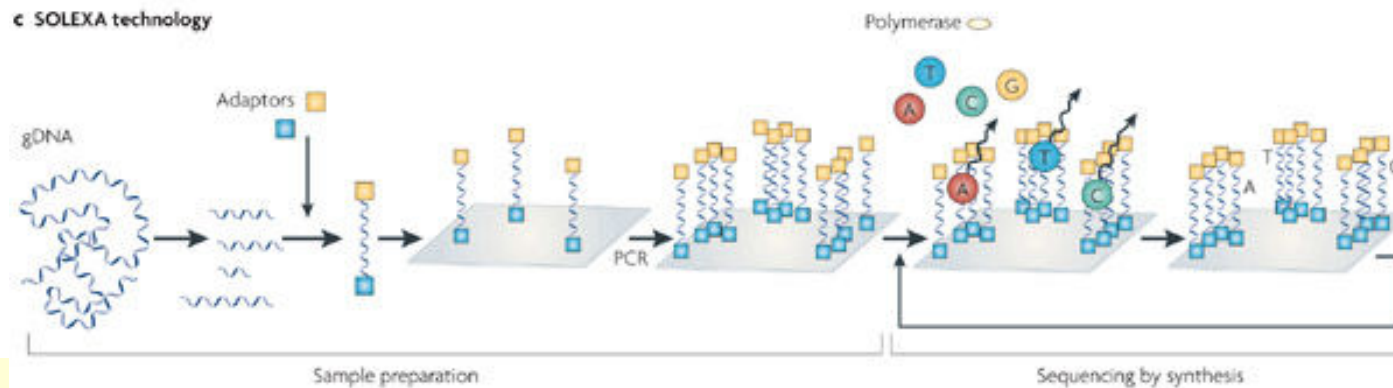
a 454 technology



b SOLiD technology



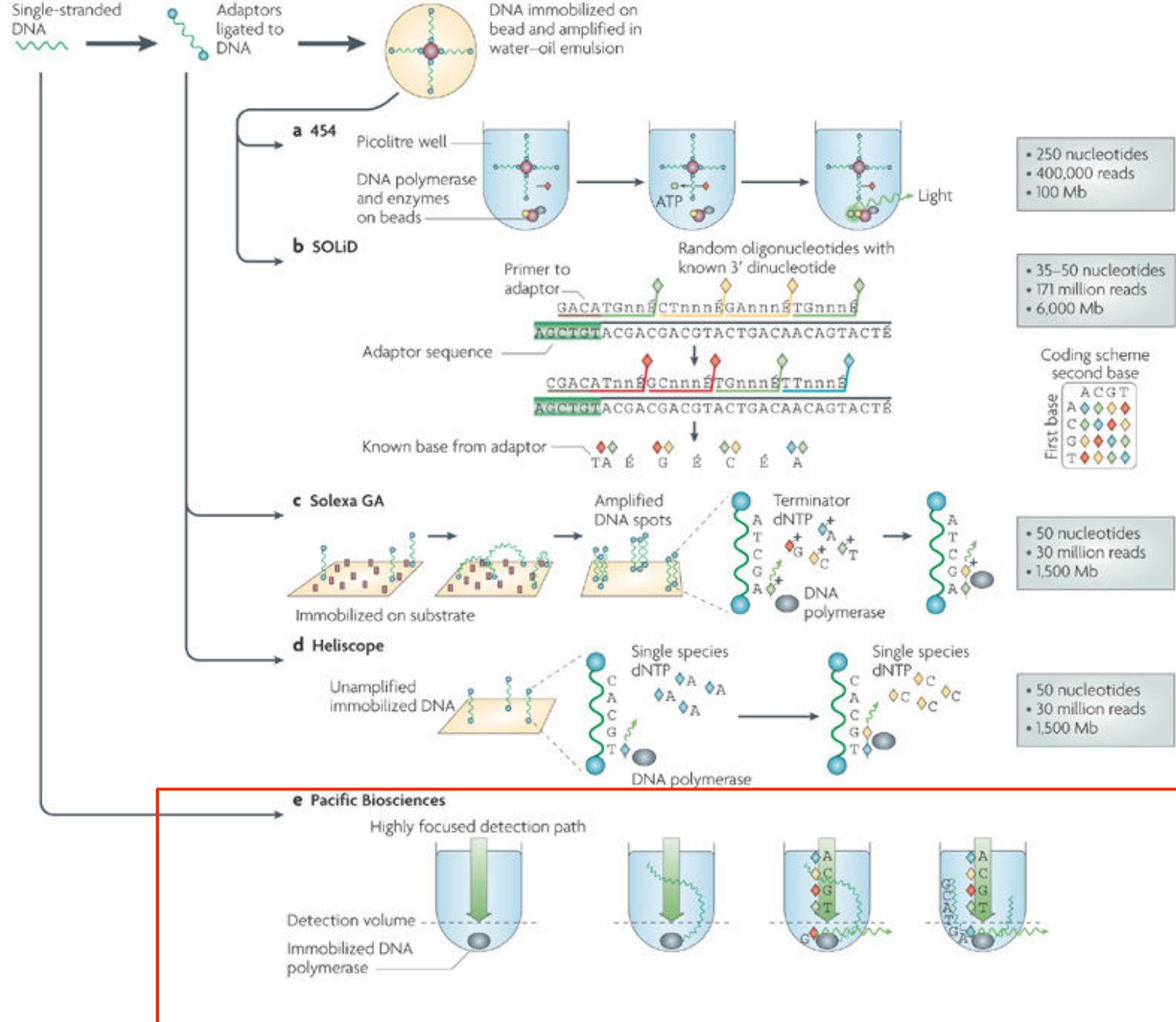
c SOLEXA technology



NNGS : Next Next Generation Sequencing

- *SMS : Single Molecule Sequencing*
 - *Éviter l'étape d'amplification (séquençage d'une seule molécule, ou d'un seul fragment non amplifié de la molécule)*
- Modèles commercialisés (Pacific Biosciences, Nanopore)





Au final ... les « non NGS » ...



Mouse over image to zoom

ABI 3100 Genetic Analyzer, DNA Sequencer, Licence included, ABI Refurbished HUGE DISCOUNT!! This Week ONLY! MAKE OFFER!

Item condition: **Manufacturer refurbished**

Time left: 22d 10h (Sep 21, 2012 20:22:01 PDT)

Price: **US \$35,000.00**

Buy It Now

Add to cart

Best Offer:

Make Offer

[Add to Watch list](#)

Shipping: Freight - Read the item description or contact the seller for details | [See all details](#)

Delivery: Varies for freight shipping

Payments: [See details](#)

Returns: 60 days money back, buyer pays return shipping | [Read details](#)

Share: [Email](#) [Facebook](#) [Twitter](#) [Pinterest](#) | [Add to Watch list](#)

Seller information

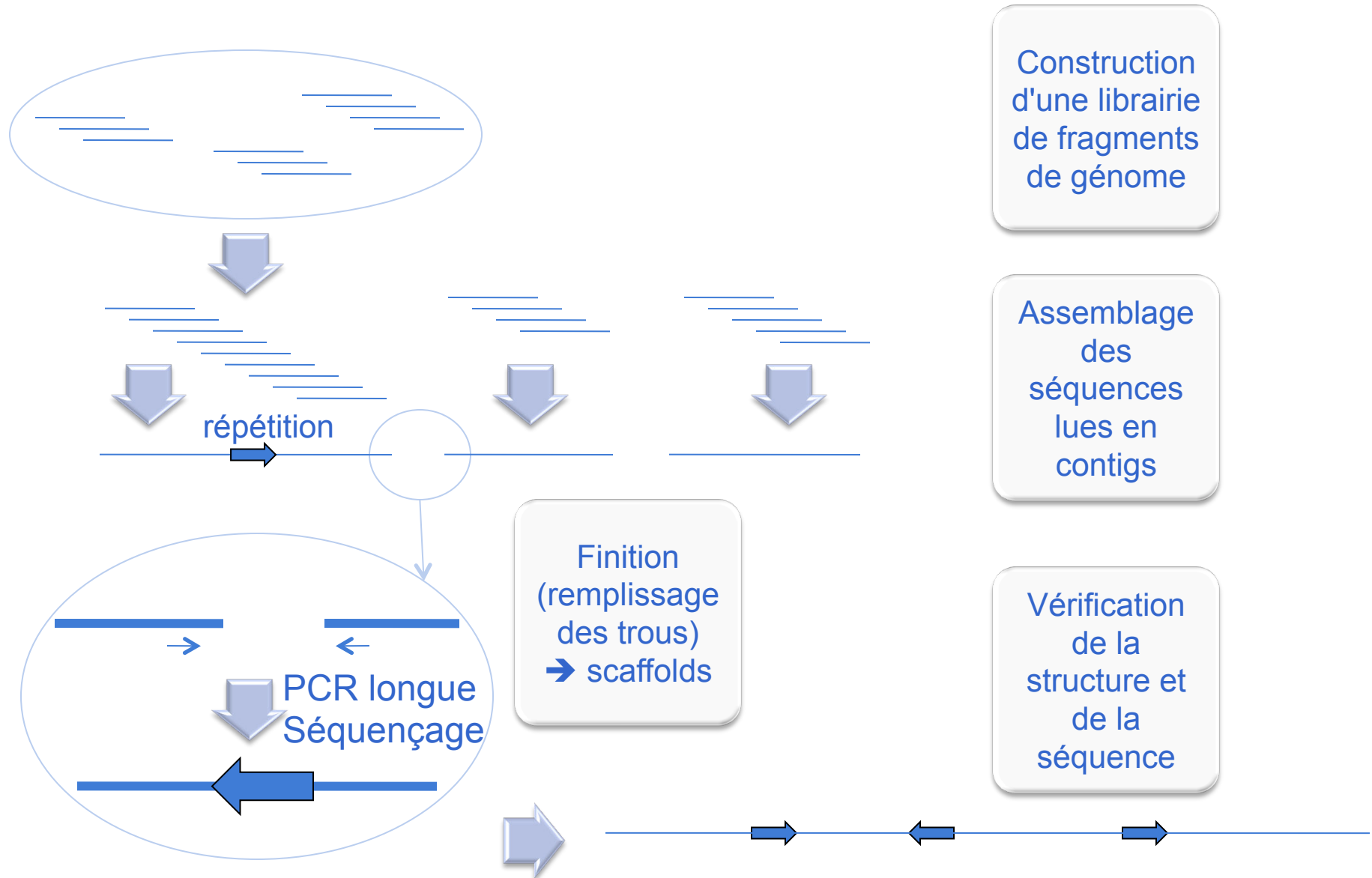
[eBay Store](#)
100% Positive Feedback

[Save this seller](#)
[See other items](#)



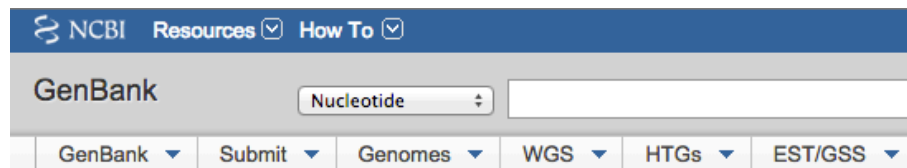
Et pourtant ...

WGS : Whole Genome Shotgun



GSS : Genome Survey Sequence

- Séquences génomiques courtes et contenant des erreurs
 - Lecture aléatoire de fragments, une seule lecture (single-pass)
- La banque dbGSS en contient différents types :
 - Séquences de génomes (lecture aléatoire)
 - Séquences des extrémités de cosmides, BAC ou YAC
 - Capture d'exons sur le génome
 - Séquences d'ALU (séquences répétées présente chez l'Homme)
 - Séquences de transposons



dbGSS release 130101

Summary by Organism - 01 January 2013

Number of public entries: 35,422,228

Mus musculus + domesticus (mouse)	4,107,950
marine metagenome	2,642,942
Zea mays + subsp. mays	2,092,607
Homo sapiens (human)	1,727,870
Nicotiana tabacum	1,420,595
Sus scrofa	1,161,435
Rattus norvegicus	867,131
Canis lupus familiaris (dog)	853,938

STS : Sequence Tagged Site

- Court (200 à 500 nt) fragment d'ADN dont
 - La séquence est unique sur un génome
 - La localisation sur le génome est connue
- Utilisé comme
 - Marqueur génétique
 - Point de repère pour construire les cartes physiques ou pour assembler les séquences génomiques

The screenshot displays the EMBL-EBI STS search interface. At the top, there's a search bar with 'All Databases' selected and a search input field containing 'emb1-Class:sts'. Below the search bar, a navigation menu includes 'Databases', 'Tools', 'EBI Groups', 'Training', 'Industry', 'About Us', and 'Help'. A secondary menu shows 'Quick Search', 'Library Page', 'Query Form', 'Tools', 'Results', 'Projects', 'Views', 'Databanks', and a 'HELP' button. The search results section shows 'Query found 945908 entries'. On the left, there are 'Apply Options to:' (selected results only, unselected results only) and 'Result Options' (Launch analysis tool: NCBI BLASTN, Show tools relevant to these results: Tools, Link to related information:). The main table lists search results with columns: EMBL, Primary Accession (Links to SVA), Accession List, Description, and Sequence Length.

EMBL	Primary Accession (Links to SVA)	Accession List	Description	Sequence Length
<input type="checkbox"/> EMBL:AJ250940	AJ250940	AJ250940	Homo sapiens STS 31 K9	394
<input type="checkbox"/> EMBL:AJ250941	AJ250941	AJ250941	Homo sapiens STS 976 I12	347
<input type="checkbox"/> EMBL:AJ250942	AJ250942	AJ250942	Homo sapiens STS 976 I12	280
<input type="checkbox"/> EMBL:AJ250943	AJ250943	AJ250943	Homo sapiens STS 924 I5	342
<input type="checkbox"/> EMBL:AJ250944	AJ250944	AJ250944	Homo sapiens STS 924 I5	331
<input type="checkbox"/> EMBL:AJ250945	AJ250945	AJ250945	Homo sapiens STS 191 L2	272
<input type="checkbox"/> EMBL:AJ250946	AJ250946	AJ250946	Homo sapiens STS 191 L2	480
<input type="checkbox"/> EMBL:AJ250947	AJ250947	AJ250947	Homo sapiens STS 264 C4	431
<input type="checkbox"/> EMBL:AJ250948	AJ250948	AJ250948	Homo sapiens STS 175 I10	390

EST : Expressed Sequence Tag

- Court fragment de séquence transcrite et épissée
 - Une seule lecture (single-pass) des ADNc d'un tissu, ...
 - Contient beaucoup d'erreurs, taille comprise entre 200 et 800 nt
- Localisation des séquences transcrites sur les génomes
- Assemblage des EST pour reconstruire les ARN complets
- Information sur les conditions d'expression des transcrits

NCBI Resources How To

GenBank Nucleotide

GenBank Submit Genomes WGS HTGs EST/GSS Metagenome

dbEST release 130101

Summary by Organism - 01 January 2013

Number of public entries: 74,186,692

Homo sapiens (human)	8,704,790
Mus musculus + domesticus (mouse)	4,853,570
Zea mays (maize)	2,019,137
Sus scrofa (pig)	1,669,337
Bos taurus (cattle)	1,559,495
Arabidopsis thaliana (thale cress)	1,529,700
Danio rerio (zebrafish)	1,488,275
Glycine max (soybean)	1,461,722
Triticum aestivum (wheat)	1,286,372
Xenopus (Silurana) tropicalis (western clawed frog)	1,271,480
Oryza sativa (rice)	1,253,557
Ciona intestinalis	1,205,674
Rattus norvegicus + sp. (rat)	1,162,136
Drosophila melanogaster (fruit fly)	821,005

La métagénomique

- Etude du matériel génétique provenant de communautés entières de micro-organismes
 - Extraites de différents environnements (océan, terre, flore intestinale, ...)
 - Accès à des organismes non cultivables et non connus
- Métagénome
 - Ensemble des fragments d'ADN issus d'un échantillon

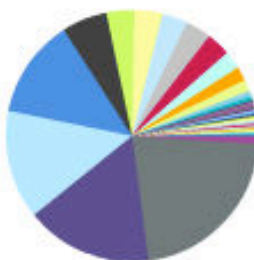
Ecosystems



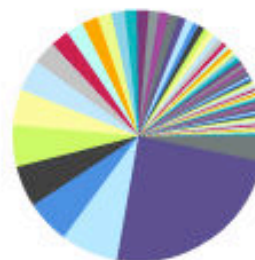
Ecosystem Category



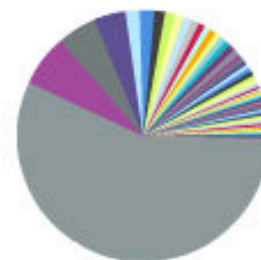
Ecosystem Types



Ecosystem Subtypes



Specific Ecosystems



Ecosystem	Cou
Environmental	231
Host-associated	128
Engineered	34
Unclassified	0

Ecosystem C	Cou
Aquatic	172
Terrestrial	51
Human	37
Arthropoda	34
Mammals	24

Ecosystem T	Cou
Marine	86
Digestive system	64
Freshwater	55
Soil	50
Unclassified	23

Ecosystem S	Cou
Unclassified	96
Large intestine	29
Lentic	22
Intertidal zone	21

Specific Eco	Cou
Unclassified	212
Fecal	25
Sediment	19
Bioreactor	15
Grasslands	7

Et la bioinformatique ?

Programmes de bioinformatique :

- Utilisés à différentes étapes du séquençage des génomes
 - Lecture des séquences à la sortie des séquenceurs
 - Assemblage des génomes à partir des fragments séquencés
 - Recherche des répétitions pour corriger les mauvais assemblages
- Utilisés pour l'exploitation des séquences d'EST
 - Regroupement des séquences appartenant à un même gène
 - Localisation des EST sur les génomes
- Utilisés pour comparer les séquences obtenues
 - Comparaison 2 à 2, multiple, une séquence contre une banque

Banques de données :

- Collecte puis stockage des séquences et bien plus ...

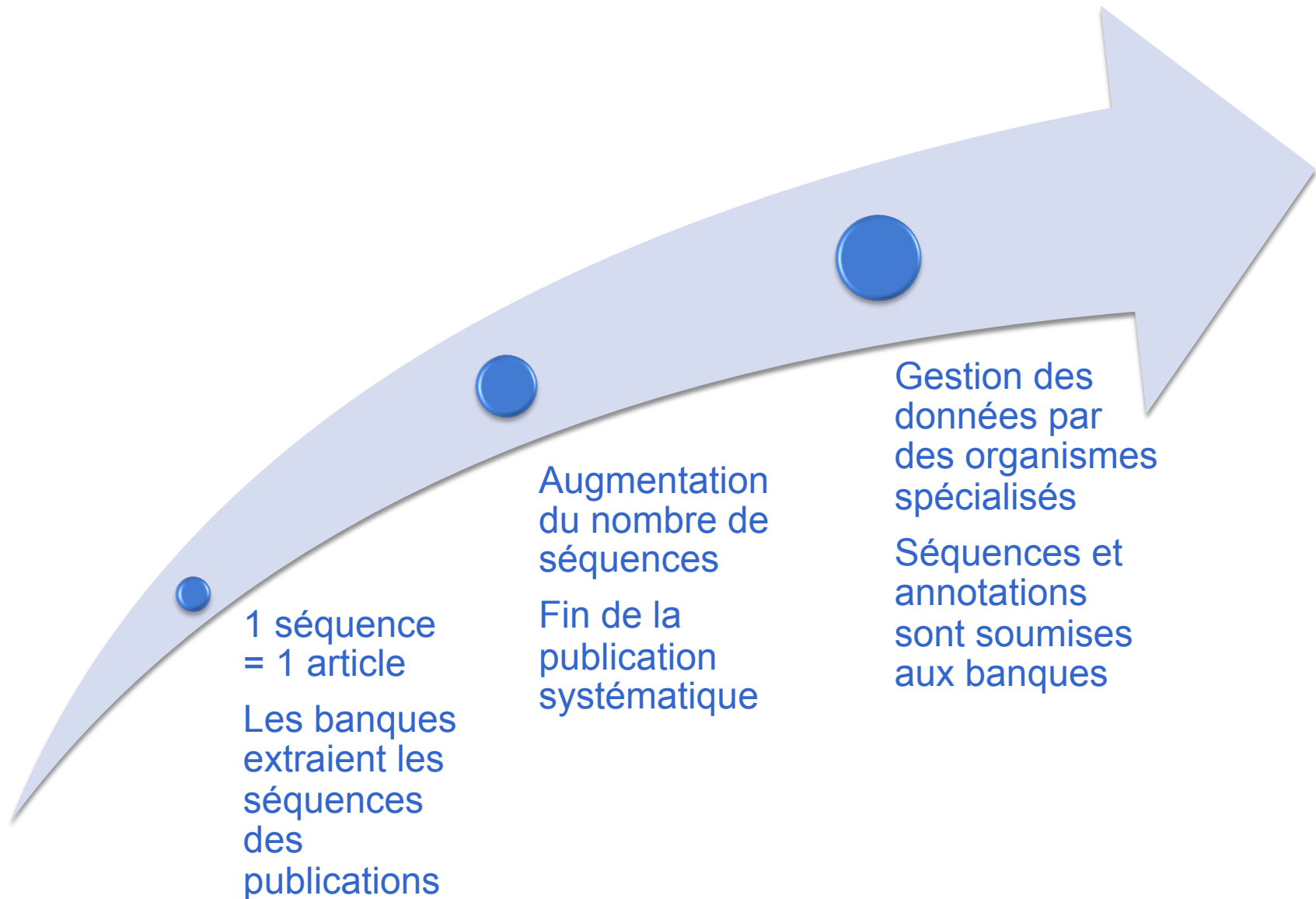
Qu'est-ce qu'une banque de données ?

- Ensemble de données relatives à un domaine, organisées par traitement informatique, accessibles en ligne et à distance
- Souvent, les données sont stockées sous la forme d'un fichier texte formaté (respectant une disposition particulière)
- Besoin de développer des logiciels spécifiques pour interroger les données contenues dans ces banques

Les banques de séquences nucléiques

- Origine des données
 - Séquençage de molécules d'ADN ou d'ARN
- Les données stockées :
 - 1 séquence + ses annotations = 1 entrée
 - Fragments de génomes
 - Un ou plusieurs gènes, un bout de gène, séquence intergénique, ...
 - Génomes complets
 - ARNm, ARNt, ARNr, ... (fragments ou entiers)
- Note 1 : toutes les séquences (ADN ou ARN) sont écrites avec des T
- Note 2 : le brin donné dans la banque est appelé brin + ou brin direct, pas de rapport avec le brin codant

Banques nucléiques, les débuts



Banques nucléiques, collaboration

International Nucleotide Sequence Database Collaboration

- Association des 3 banques nucléiques :
 - ENA (European Nucleotide Archive) – EMBL-EBI
<http://www.ebi.ac.uk/embl/>
 - GenBank (banque des Etats-Unis d'Amérique) – NCBI
<http://www.ncbi.nlm.nih.gov/Genbank/>
 - DDBJ (DNA DataBank of Japon) – CIB
<http://www.ddbj.nig.ac.jp/>
- Echange quotidien des données
- Répartition de la collecte des données
 - Chaque banque collecte les données de son continent

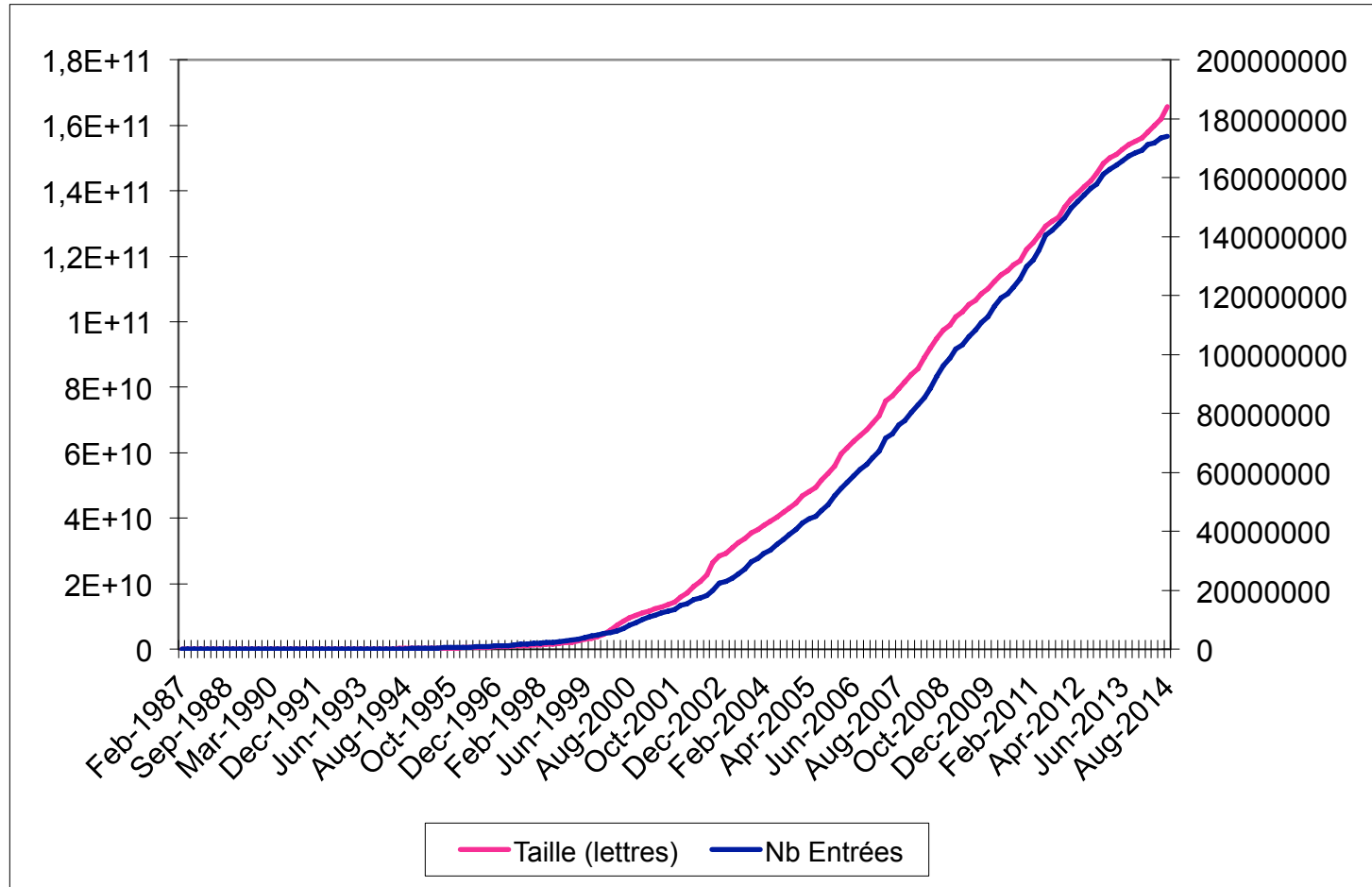


Banques nucléiques, mises à jour de la banque

- Une nouvelle version est disponible plusieurs fois par an
 - Date et numéro de version (release)
 - Données figées à une date fixée (toutes les séquences collectées jusque là)
- Mise à disposition des « Updates »
 - Mise à jour quotidienne des données
 - Toutes les nouvelles séquences depuis la dernière version
- Facilite le traitement des données
 - Pas besoin de télécharger la banque entière tous les jours
 - Possibilité de faire des calculs longs

Banques nucléiques : l'explosion relative

Taille de Genbank : genbank/statistics



Banques nucléiques, format d'une entrée

- 3 parties :

Description générale
de la séquence

« Features »

Description des objets
biologiques présents
sur la séquence

La séquence

```
ctccggcagc ccgaggtcat cctgctagac tcagacctgg atgaacccat agacttgcgc      60
tcggtcaaga gccgcagcga ggccggggag ccgcccagct ccctccaggt gaagcccagag    120
acaccggcgt cggcggcggt ggcggtggcg gcggcagcgg caccaccac gacggcggag      180
```

- Chaque ligne commence par
un mot-clé

- Deux lettres pour EMBL
- Maximum 12 lettres pour
Genbank et DDBJ

- Fin d'une entrée : //

EMBL, description générale de la séquence

- ID : toujours la 1ère ligne d'une entrée

Accession	Version	Topologie	Molécule	Classe	Taxonomie	Taille seq
M71283	SV 1	linear	genomic DNA	STD	BCT	1322 BP

- AC : numéros d'accension
 - Un n°acc principal pour chaque entrée, unique
 - Une liste de n°acc secondaires (historique de l'entrée)
- DT : dates de création et de dernière version
- DE : description du contenu de l'entrée
- KW : mots-clés ; peu renseigné
- OS, OC : organisme contenant la seq. et sa taxonomie
- RN, RC, RX, RP, RA, RT, RL : réf. bibliographiques
 - Uniquement les références données par les auteurs de l'entrée

GenBank et DDBJ, description générale

- LOCUS : toujours la première ligne d'une entrée

Locus name	Taille seq	Molécule	Topologie	Division	Date
BACCOMQP	1322 bp	DNA	linear	BCT	26-APR-1993

- DEFINITION = DE
- ACCESSION = AC
- VERSION ~ DT
- KEYWORDS = KW
- SOURCE, ORGANISM = OS, OC
- REFERENCE, AUTHORS, TITLE, JOURNAL, ... = R...

Banques nucléiques, lignes FT (Features)

Format (partagé par toutes les banques) :

- **Key** : un seul mot indiquant un groupe fonctionnel
 - Vocabulaire contrôlé, hiérarchique
 - gene : séquence complète du gène (y compris les introns)
 - CDS : séquence codante (sans les introns, entre ATG et Stop)
- **Location** : instructions pour trouver l'objet sur la séquence de l'entrée
 - Voir description du format plus loin
- **Qualifiers** : description précise du groupe fonctionnel
 - Format : /qualifier=' 'commentaires libres'
 - /gene="comQ": nom du gène concerné
 - /product="comQ": nom de la protéine produite
 - /protein_id="BAB13491.1": numéro accession protéine
 - /note="competence protein Q;...": information sur la fonction

Banques nucléiques, exemples de « Key » (1/2)

- Mot-clé le plus général : `misc_feature`
- Changements dans la séquence : `misc_difference, ...`
- Régions répétées : `repeat_region, ...`
- Régions des Ig : `immunoglobulin_related, ...`
- Structures secondaires : `misc_structure`
 - `stem_loop`
 - `D-loop`
- Régions impliquées dans la recombinaison :
`misc_recomb, ...`

Banques nucléiques, exemples de « Key » (2/2)

gene	misc_RNA	
misc_signal		prim_transcript
promoter		precursor_RNA
CAAT_signal		mRNA
TATA_signal		5'clip
-35_signal		3'clip
-10_signal		5'UTR
GC_signal		3'UTR
RBS		exon
polyA_signal		CDS
enhancer		intron
attenuator		
terminator	polyA_site	

Banques nucléiques, localisation des objets bio

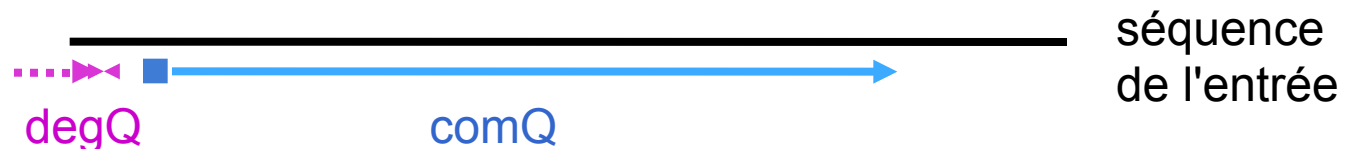
- **467** : l'annotation ne concerne qu'une seule base
- **109..1105** : entre les positions 109 et 1105 (incluse)
 - Toujours la position la plus petite en premier
- **<1..21 ou 1275..>1322** : « Keys » tronqués
 - Commence avant le premier nt de l'entrée
 - Se termine après le dernier nt de l'entrée (taille seq = 1322)
- **<234..888** : début réel inconnu, mais avant 234
- **234..>888** : fin réelle inconnue, mais après 888
- **complement(340..565)** : séquence complémentaire inversée à celle de l'entrée (brin -)
- **join(12..78,134..202)** : fragments indiqués mis bout à bout (concaténés) ; nombre de fragments illimité

Banques nucléiques, Qualifiers

- Vocabulaire contrôlé entre « / » et « = » puis texte libre
 - Le vocabulaire dépend du Key auquel le Qualifier se réfère
- Nom de gène
 - /gene= ou /name=
- Fonction de la protéine codée par le gène
 - /product=
- Traduction de la séquence codante
 - /translation=
- Origine de l'annotation
 - /evidence=
- Texte libre
 - /note=

Un exemple de « Feature » d'une séquence ADN

```
FT   CDS                <1..21
FT                               /codon_start=1
FT                               /db_xref="SWISS-PROT:Q99039"
FT                               /transl_table=11
FT                               /gene="degQ"
FT                               /protein_id="AAA22322.1"
FT                               /translation="YAMKIS"
FT   terminator          21..47
FT                               /gene="degQ"
FT   promoter            109..140
FT                               /gene="comQ"
FT   mRNA                146..1105
FT                               /partial
FT                               /gene="comQ"
```



Banques nucléiques, mise à jour des données

- Evolution possibles des entrées
 - Changements dans la séquence, dans les annotations
 - Ajout d'une séquence, d'une annotation, d'une publication
- Les entrées sont mises à jour par leurs auteurs
- Limites de ce processus
 - Seuls les auteurs d'une entrée peuvent la corriger
 - Seules les données issues de séquençage sont admises
- Création de TPA : Third Party Annotation
 - TPA experimental : la séquence et ses annotations doivent avoir été vérifiées par des expériences en laboratoire “humide”
 - TPA inferential : séquence et/ou annotations proviennent de prédictions basée sur des études de familles de gènes, par exemple

Banques nucléiques, inconvénients

- Difficulté de mise à jour des données
 - Version plus récente d'une séquence ou d'une annotation dans d'autres banques (ex : banques dédiées à un génome complet)
- Forte redondance
 - Un même fragment de séquence présent dans plusieurs entrées
- Annotations peu normalisées
 - Difficulté de recherche d'une information particulière
- Annotations peu précises
 - Peu de descriptions sur les gènes et leurs produits
- Erreurs dans les annotations

RefSeq (NCBI) = Reference Sequence collection

- « *The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major research organisms. »*
- « *Curated collections from a number of biologically significant organisms »*
- Avantages :
 - Non redondante
 - Liens explicites entre les séquences nucléiques et protéiques
 - Mise à jour régulière par le personnel du NCBI avec indication du statut de l'entrée
 - Validation des données et consistance des formats
 - Synthèse des informations issues de plusieurs entrées nucléiques ou protéiques

Différents niveaux de correction des données

Indiquées dans le champ « COMMENT »

- Reviewed

- Revu par un membre du NCBI qui a ajouté des informations provenant de publications scientifiques et de différentes entrées de séquences

- Validated

- Une première révision a été effectuée par un membre du NCBI, mais l'annotation est en cours

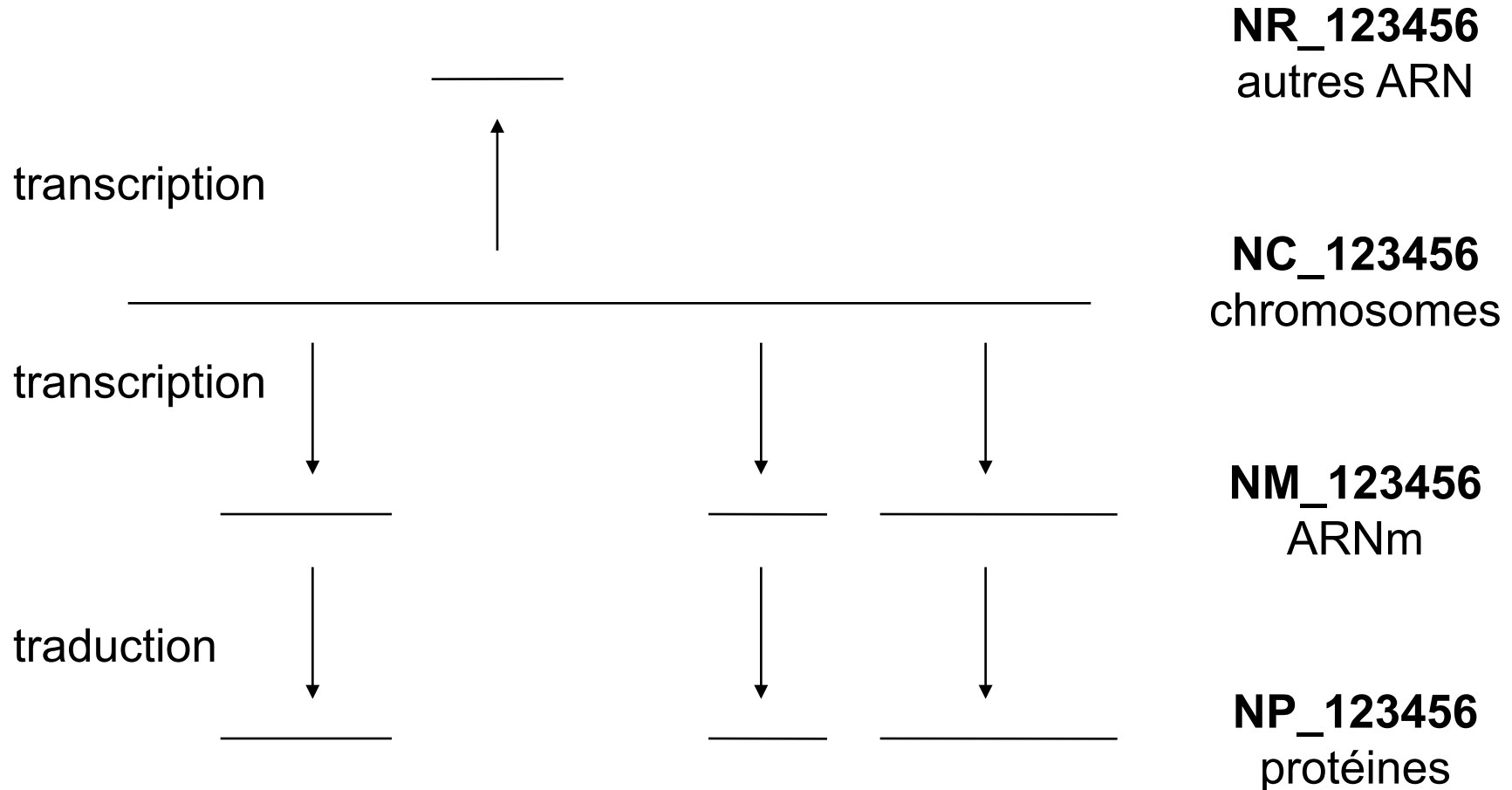
- Provisional

- Entrée non lue par un annotateur, mais qui contient sûrement un vrai transcrit ou une vraie protéine

- Predicted

- Transcrit ou protéine issu d'une prédiction à l'aide d'un programme informatique

Quelques numéros d'accèsion de RefSeq



Autres banques du NCBI

■ Gene :

- Banque centrée sur les gènes
- Source : RefSeq ou centres reconnus d'annotation des génomes
- Localisation sur le génome, variants d'épissage, protéines codées par le gène, bibliographie, gènes homologues, ...

■ UniGene : *transcriptome*

- Regroupement de séquences nucléiques dicté par les gènes
- Un groupe contient toutes les séquences qui représentent un gène unique (ARNm et EST)
- Données mises à jour régulièrement
- Problème : gestion des familles de gènes répétés

Banques généralistes de génomes

- 3 banques : Ensembl (EBI), UCSC Genome (USA), NCBI genome (USA)
- Les **même** séquences brutes
- 3 méthodes **différentes** pour annoter les séquences
 - Principe de base : localiser sur la séquence des informations provenant de différentes sources
 - Gènes connus (annotations provenant d'autres banques)
 - ARNm et EST localisés sur le génome (variants d'épissage)
 - Protéines localisées sur le génome (traduction du génome)
 - Prédictions statistiques
- Données de comparaison entre génomes

Quelques formats de données biologiques

- Format des banques, exemples :
 - Séquences ADN/ARN : EMBL ; GenBank et DDBJ
 - Séquences protéiques : SwissProt et TrEMBL ; PIR ; ...
- Formats lus par la plupart des outils en bioinformatique
 - FASTA
 - Séquence brute (« raw sequence »)
- Conversion de formats
 - Lors de la consultation des banques
 - Le programme ReadSeq (n'importe quel format en entrée, choix du format de sortie)

Le format FASTA

- Utilisé par les logiciels d'analyse de séquence
- Une ligne de commentaires précédée de « > »
- La séquence brute (pas d'espace, ni de nombre)

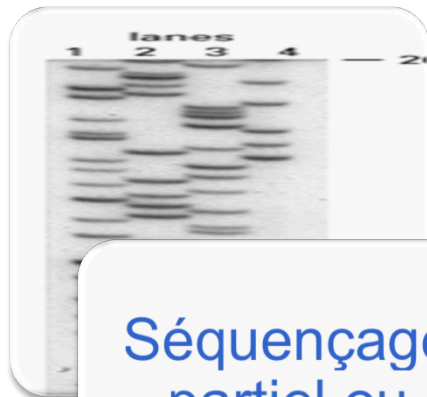
```
>Human Polycomb 2 homolog (hPc2) mRNA, partial cds
ctccggcagcccgagggtcatcctgctagactcagacctggatgaacccat
agacttgcgctcgggtcaagagccgcagcgaggccggggagccgcccagct
ccctccagggtgaagcccgagacaccggcgctcggcggcggtggcggtggcg
Gcggcagcggcacccaccacgacggcgggagaagcct
```

```
>hPc2 gene
```

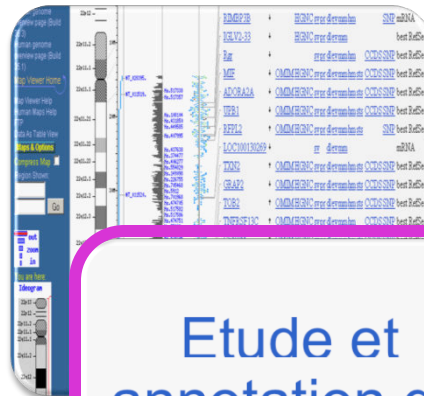
```
ggacgaacctgcagagtcgctgagcgagttcaagcccttctttgggaata
taattatcacccgacgtcacccggaactgcctcacccgttactttcaaggag
tacgtgacggtg
```

La génomique

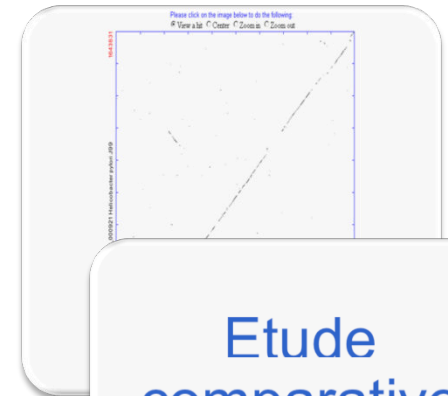
- Etude des génomes et de l'ensemble de leurs gènes
 - La structure
 - Le fonctionnement
 - L'évolution
 - Le polymorphisme, ...
- Plusieurs étapes :



Séquençage
partiel ou
total d'un
génom



Etude et
annotation de
la séquence
du génome



Etude
comparative
de plusieurs
génom

Ce que souhaiterait connaître chaque biologiste :

- Le jeu complet et précis des gènes ainsi que leur position sur le génome,
- L'ensemble des transcrits d'un génome,
- Le lieu et le moment de l'expression de chaque transcrit,
- La protéine produite par chaque transcrit,
- Le lieu et le moment de l'expression de chaque protéine,
- La structure complète de chaque protéine,
- La fonction de chaque protéine,
- Les mécanismes cellulaires auxquels participent les protéines.

Annotation des séquences nucléiques

- Petites séquences : annotation manuelle
 - Prédiction des gènes à « ARN » ou « à protéine » présents sur la séquence à l'aide de programmes
 - Localisation, fonction des produits, ...
 - Permet d'orienter les expérimentations
 - Les techniques seront présentées dans un prochain cours
- Génomes complets
 - Annotation réalisée entièrement (ou presque) par des programmes informatiques
 - Risque important d'erreurs



Ce ne sont que des prédictions, une vérification expérimentale est indispensable

La génomique

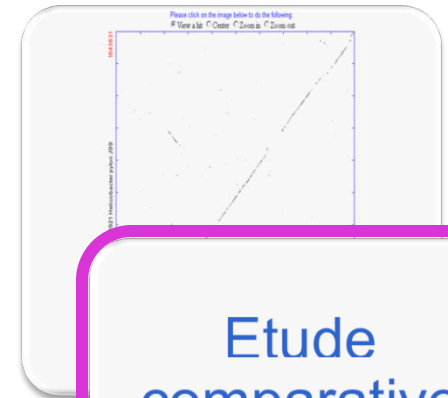
- Etude des génomes et de l'ensemble de leurs gènes
 - La structure
 - Le fonctionnement
 - L'évolution
 - Le polymorphisme, ...
- Plusieurs étapes :



Séquençage
partiel ou
total d'un
génom



Etude et
annotation de
la séquence
du génome



Etude
comparative
de plusieurs
génom

Génomique comparative

■ Objectifs :

- Etudier l'évolution entre espèces à l'échelle du génome
- Identifier des gènes spécifiques à une espèce (pathogénicité, ...)
- Retrouver des régions de synténie (conservation de l'ordre de gènes homologues dans le génome d'espèces différentes)
- Étude du polymorphisme au sein d'une même espèce

■ Méthodes

- Comparaison de cartes génétiques
- Alignement de génomes
- Alignement de toutes les protéines de plusieurs génomes
- Etude de l'ordre des gènes

Phylogénie

- Objectifs des études phylogénétiques :
 - Mieux comprendre les mécanismes de l'évolution et les mécanismes moléculaires associés.
 - Connaître l'arbre de la vie (taxonomie).
 - Etudier la biodiversité, l'origine géographique des espèces, ...
- Phylogénie moléculaire :
 - Détermination de l'arbre phylogénétique d'un ensemble de séquences
- Arbre phylogénétique :
 - Configuration **la plus probable** pour rendre compte du degré de parenté existant entre des séquences.