La génomique, du laboratoire de recherche aux applications médicales.

philippe.glaser@pasteur.fr

Institut Pasteur

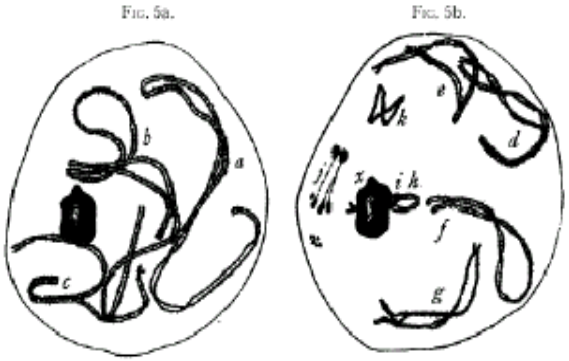
Plan de la présentation

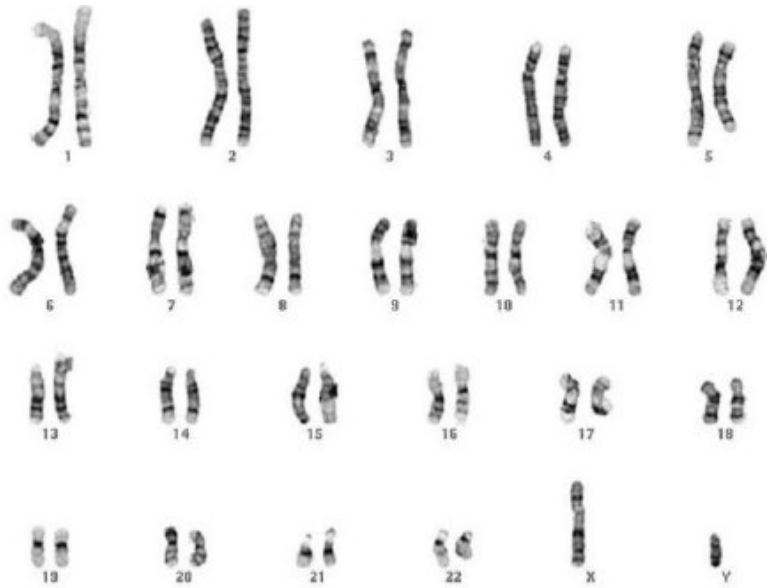
- De la génétique à la génomique : les bases
- L'évolution des méthodes de séquençage : comment décrypter les génomes toujours plus vite
- Les applications dans la caractérisation des génomes et de leur évolution
- Les applications du séquençage à la compréhension du fonctionnement des génomes

Les chromosomes le support de l'hérédité

Theodor Boveri - Walter Sutton (1902)

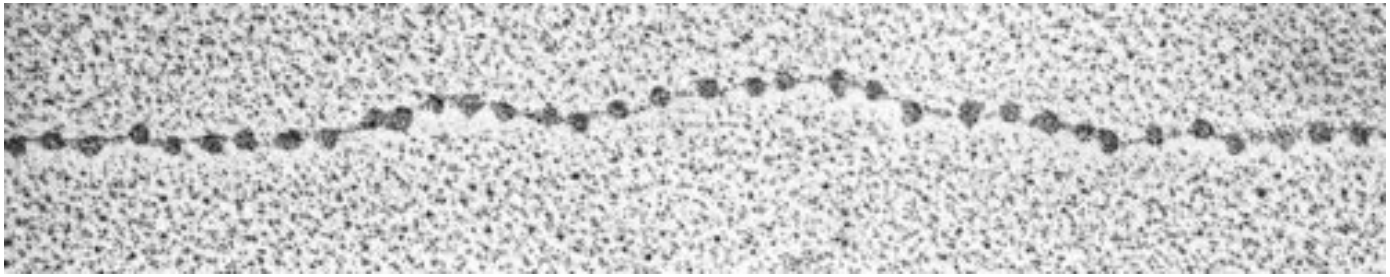
Fig. 12. Theodor Boveri.





Par l'analyse de leur distribution dans les cellules, leur forme, la comparaison entre les ovules et les spermatozoïdes, ces deux savants ont proposé que les chromosomes étaient le support de l'hérédité.

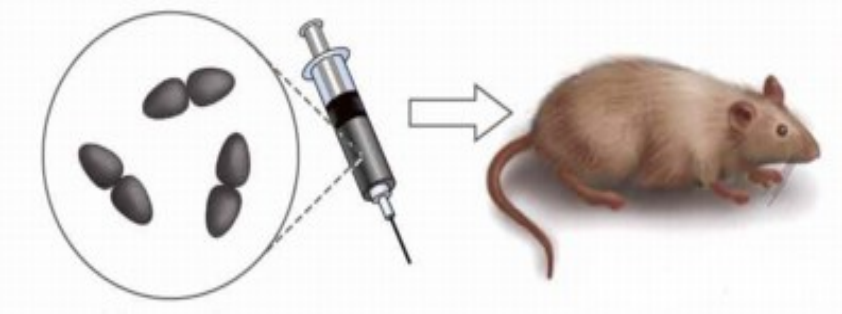
⇒ La cytogénétique



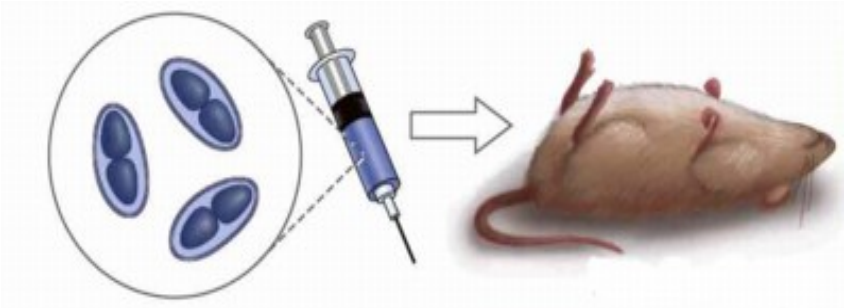
- Les chromosomes sont constitués d'ADN et de protéines - quel est le support de l'hérédité?
  - Réponse 1 : les protéines – des molécules très variées constituées de 20 acides aminés
  - Réponse 2 : l'ADN – semblait au départ un composant structurel bien moins complexe constitué de 4 bases

L'ADN est le support de l'hérédité

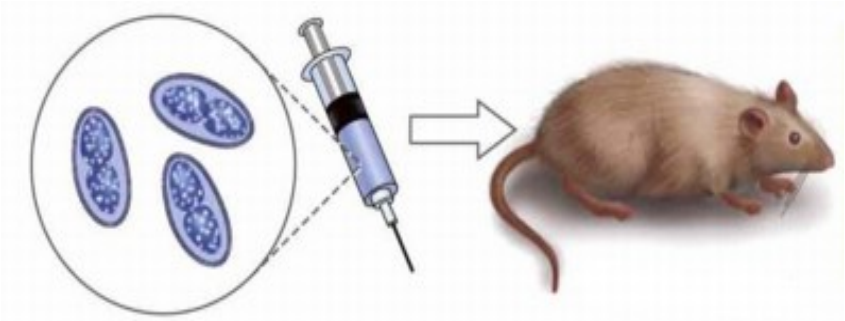
Transformation du pneumocoque - Oswald T. Avery



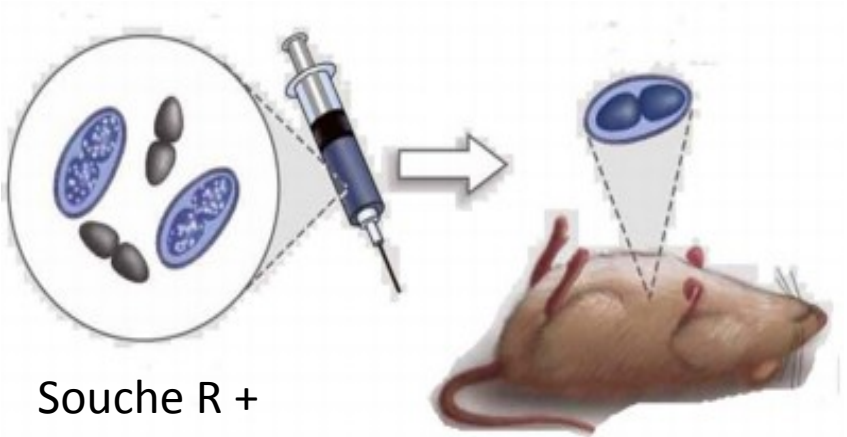
Souche R
(+ protéines souche S)



Souche S



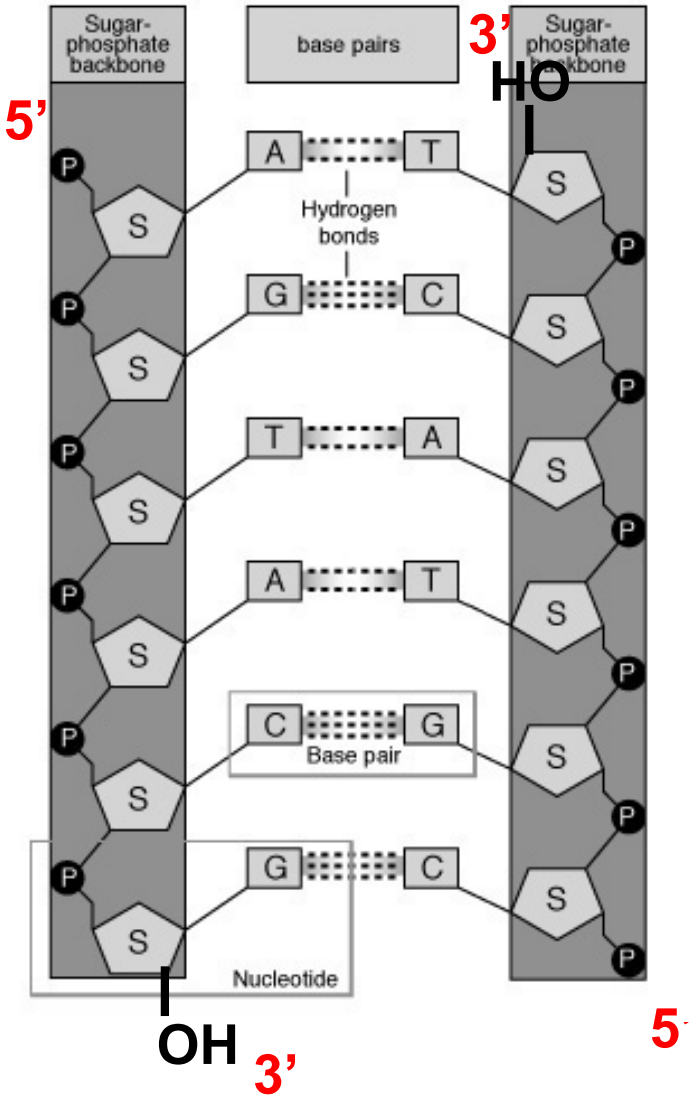
Souche S tuée

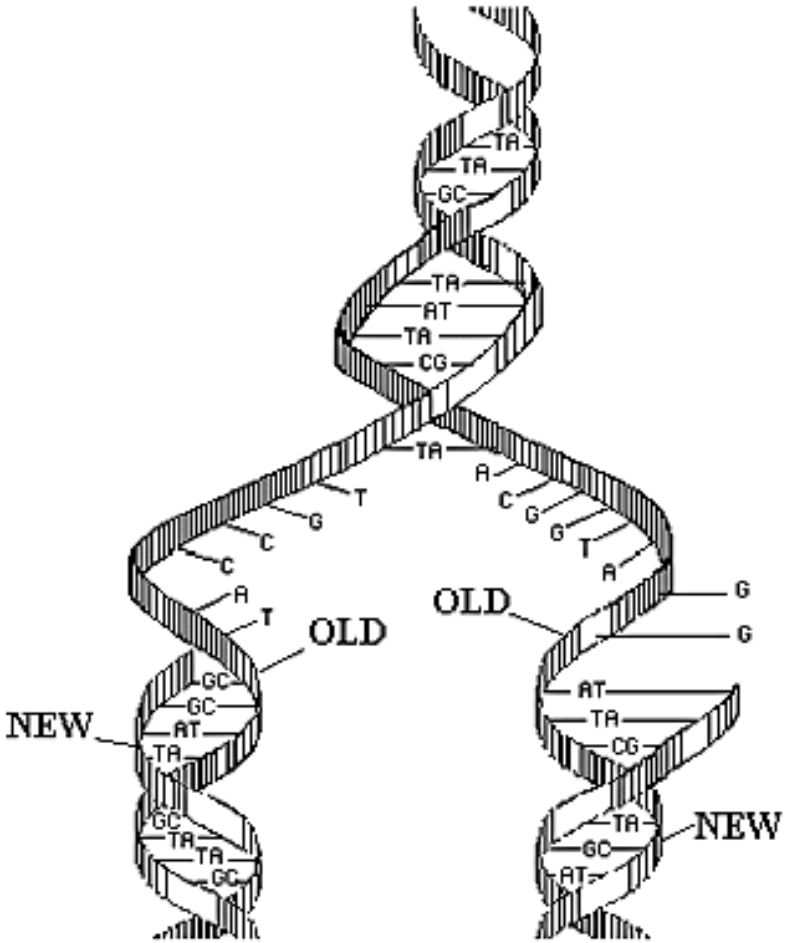


Souche R +
Souche S tuée ou ADN de souche S

⇒L'ADN est le support de l'hérédité!
COMMENT?

L'ADN est un acide nucléique double brin

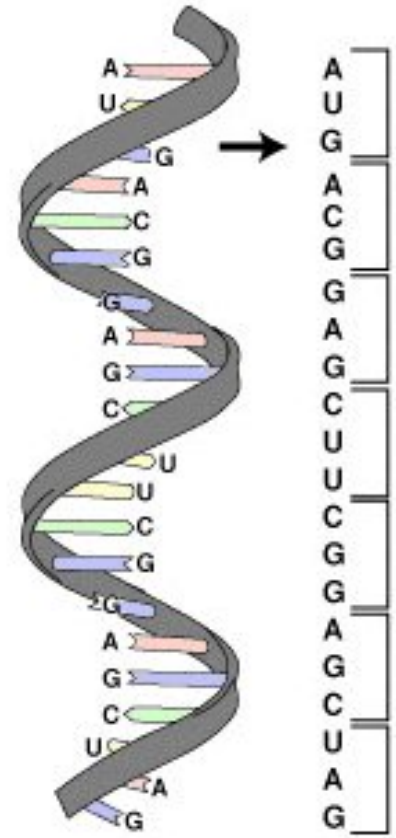




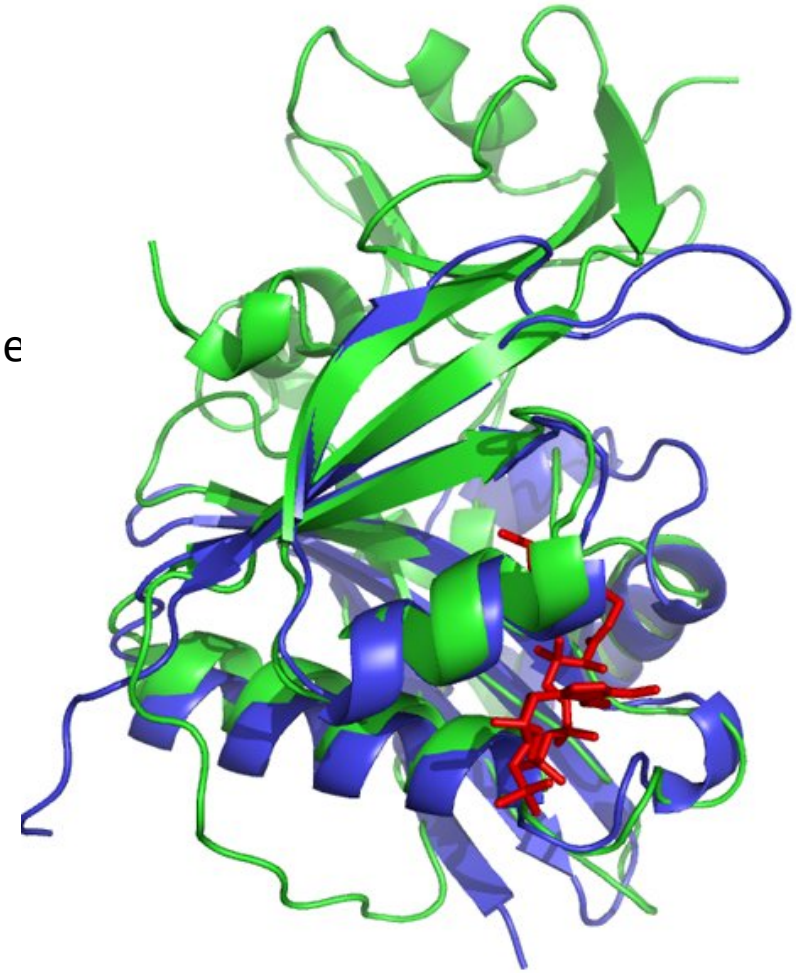
Réplication de l'ADN

Le code génétique

La structure des protéines est écrite dans le génome



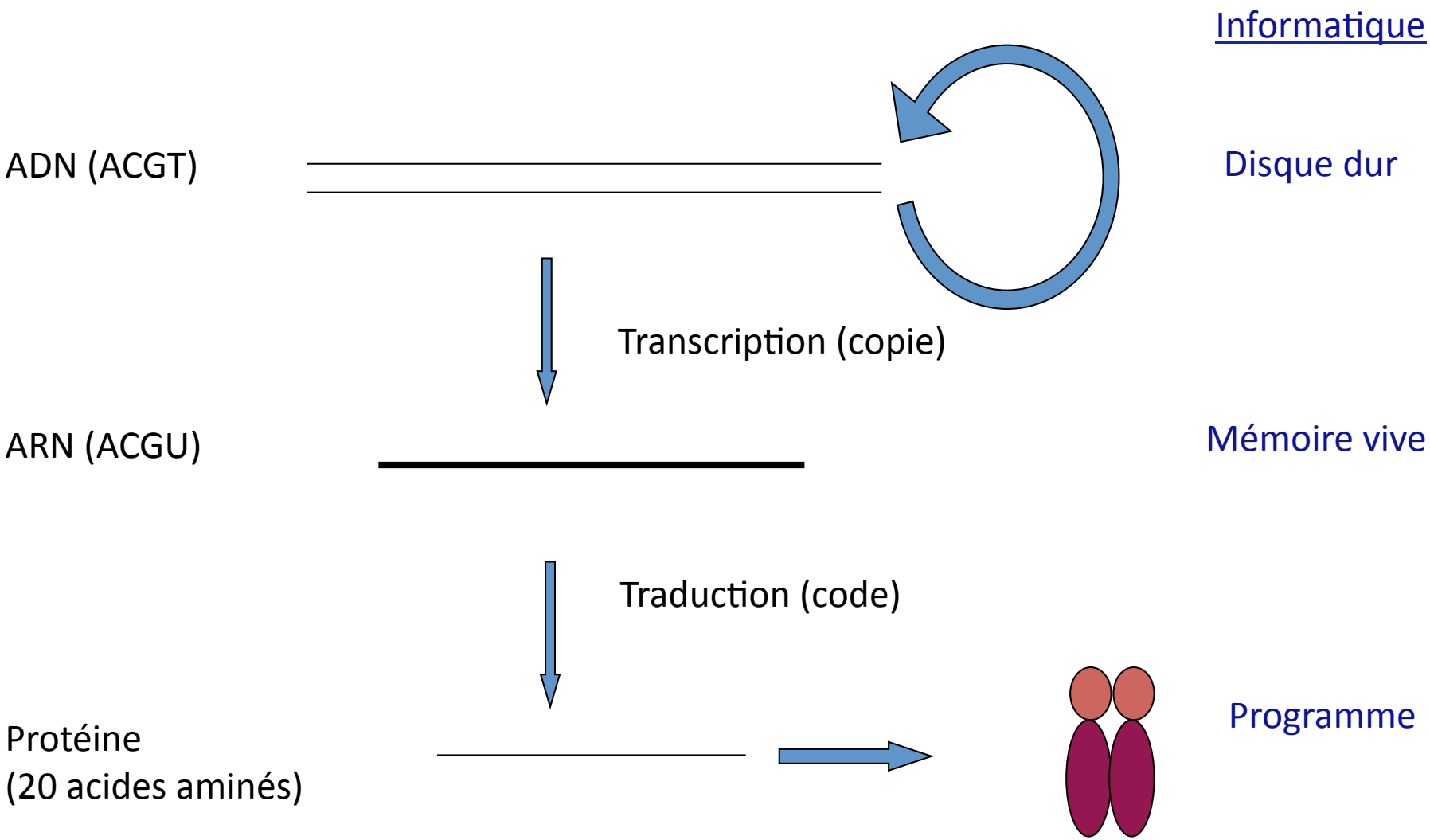
AUG	Codon 1	Méthionine
ACG	Codon 2	Thréonine
GAG	Codon 3	Acide glutamique
CUU	Codon 4	Leucine
CGG	Codon 5	asparagine
AGC	Codon 6	Sérine
UAG	Codon 7	Stop



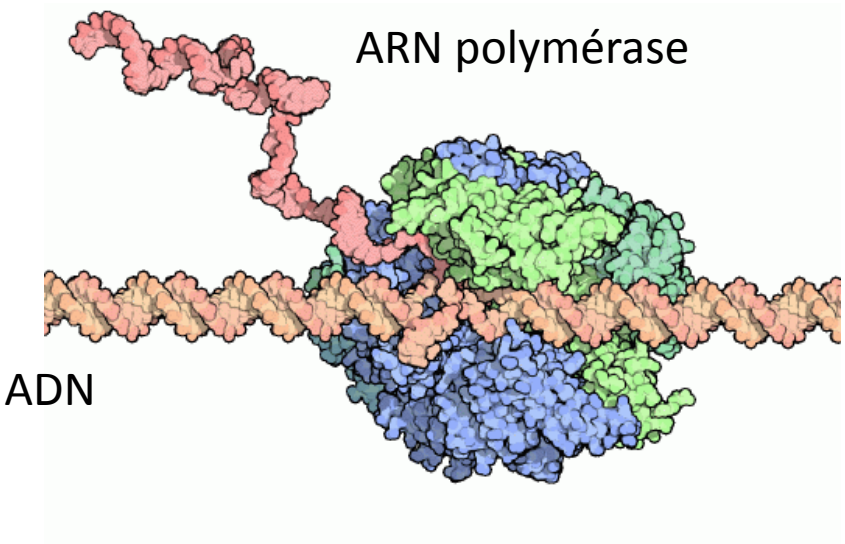
Message 1:

- L'ADN est le support de l'hérédité
- Sa structure lui permet d'être recopié et transmis fidèlement à la descendance
- L'agencement des 4 bases constitue un texte qui est le programme génétique transmis de génération en génération
- Le code génétique associe à des triplets de trois bases un des 20 acides aminés des protéines

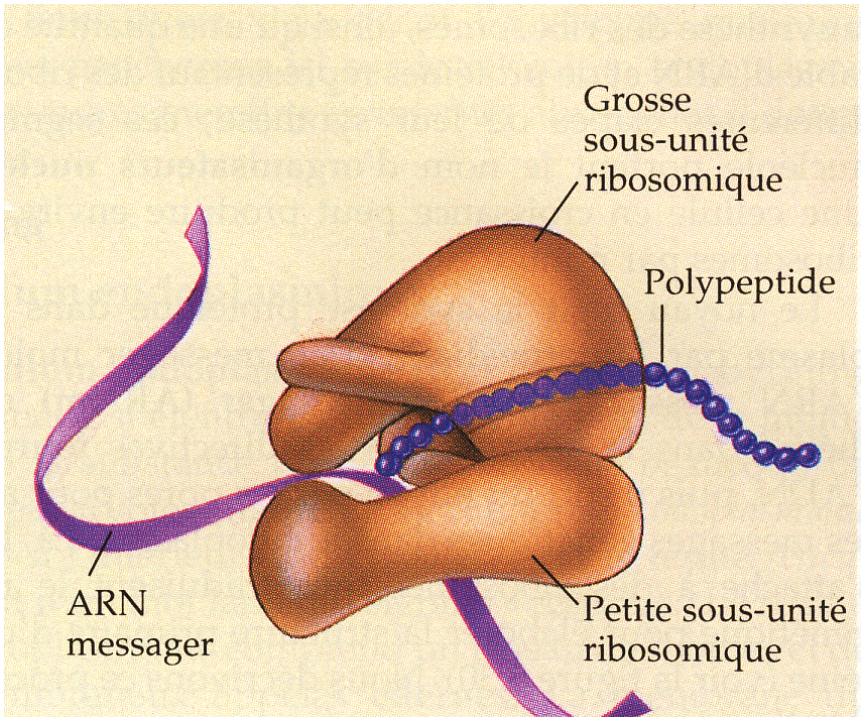
L'expression génétique et les technologies de l'information



L'expression génétique



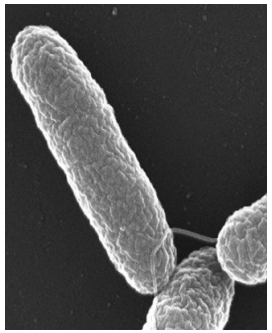
L'ADN est recopié en ARN dans le noyau



L'ARN est "traduit" en protéines
dans le cytoplasme

Deux autres niveaux de complexité de l'expression génétique

Les gènes morcelés



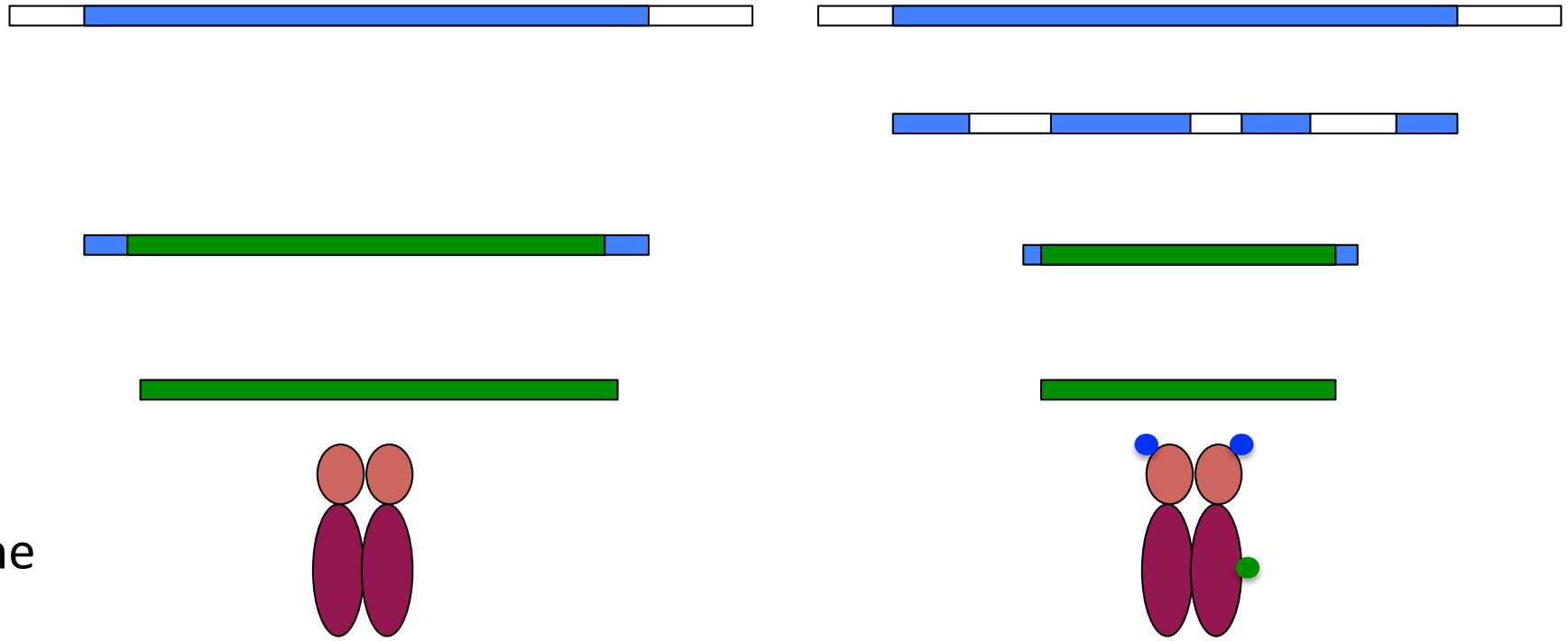
Escherichia coli : 4000 gènes

Homo sapiens : 25 000 gènes

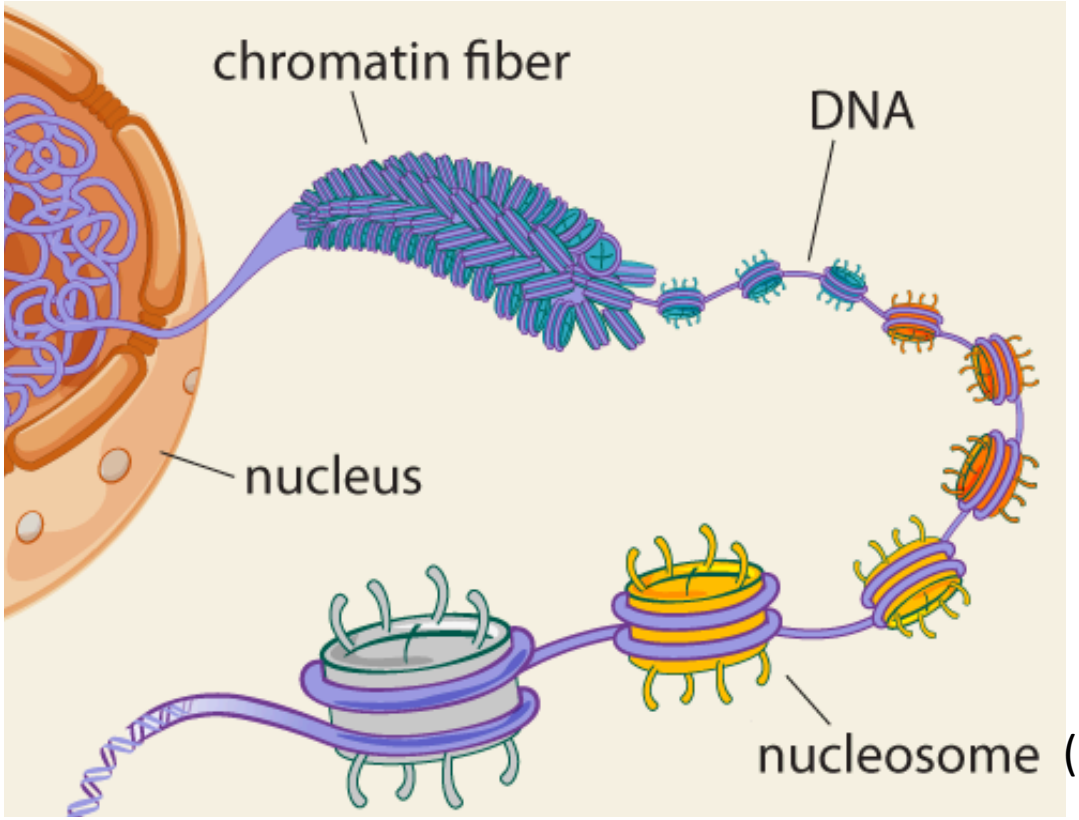
ADN

ARN

Protéine



La chromatine et l'épigénétique



(histones)

- L'ADN est associé à des protéines dans le noyau
- La modification des histones affecte la transcription des gènes
- La méthylation de l'ADN modifie également la transcription

Message 2:

- Un gène code pour une ou plusieurs protéines

- Le fonctionnement du génome – le programme génétique assure la synthèse de chaque protéine à la bonne concentration – on parle d'expression génétique

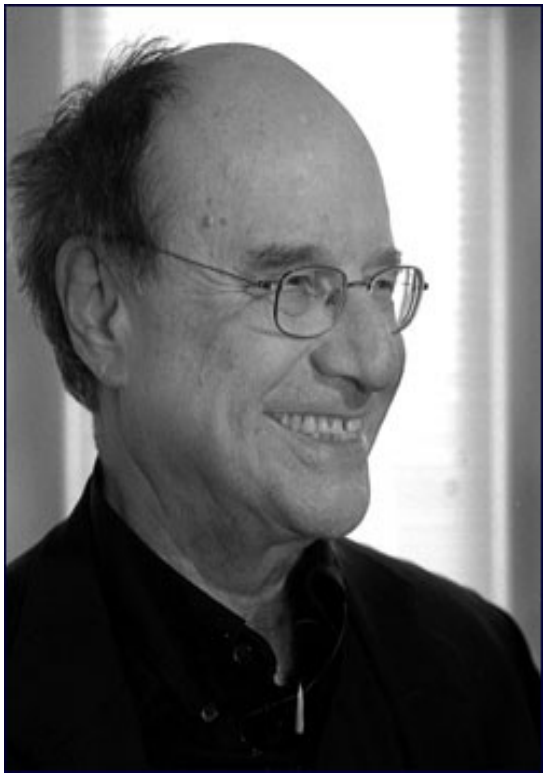
- Le phénotype résulte à la fois des caractéristiques des protéines et de leur quantité.

Le séquençage des acides nucléique permet de caractériser:

I. L'ADN
II. L'ARN
III. La traduction de l'ARN
IV. La modification de la chromatine
V. La fixation de régulateurs sur leur cible

- Décrypter le génome: le séquençage de l’ADN

1977 : Le séquençage de l'ADN



Walter Gilber - Université d' Harvard
Méthode chimique

Fred Sanger - MRC Cambridge
Méthode enzymatique

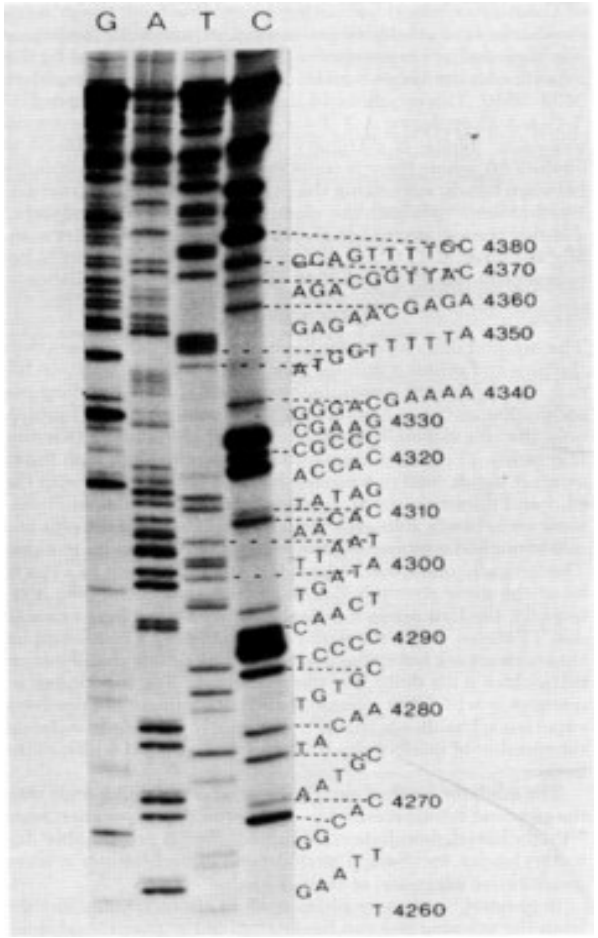
La méthode de Sanger

```
ACGTGGGCTAAGTGCGTATGCATGCGTGCT
|||||||||||||||||||||||||||||
TGCACCCGATTCACGCATACGT
TGCACCCGATTCACGCATACG
TGCACCCGATTCACGCATAC
TGCACCCGATTCACGCATA
TGCACCCGATTCACGCAT
TGCACCCGATTCACGCA
TGCACCCGATTCACGC
TGCACCCGATTCACG
```

L'ADN est recopié par l'ADN polymérase

La synthèse se **termine** à une base connue: A C G T

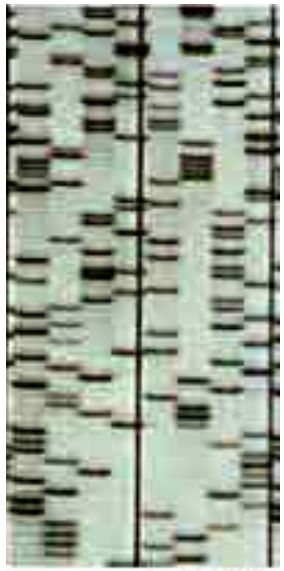
Toutes les molécules ont le même **début**



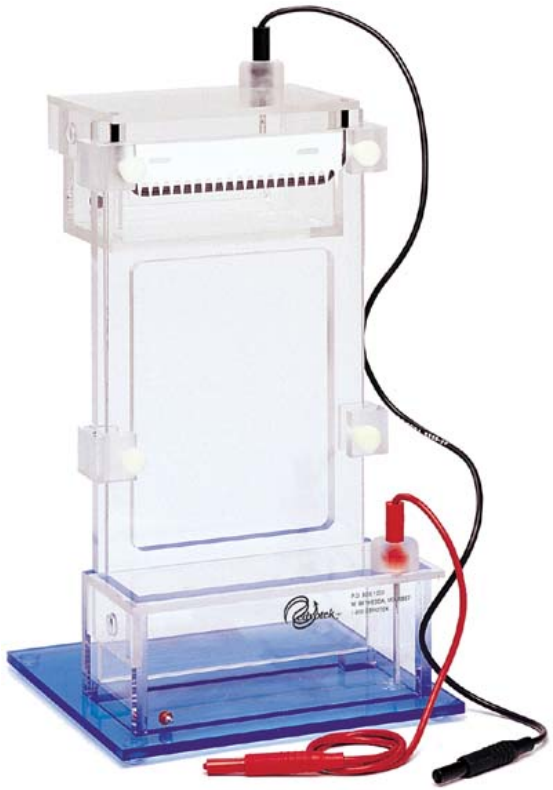
Séparation des molécules par électrophorèse

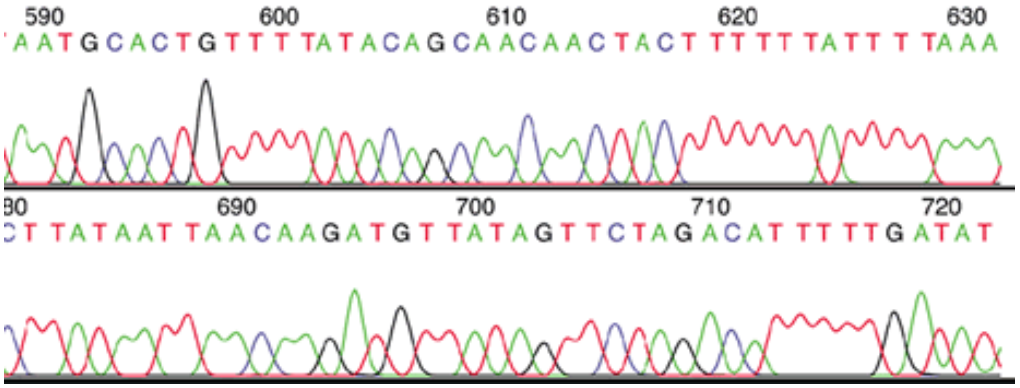
1977 - 2003

- Automatisation et optimisation de la méthode de Sanger
- Séquençage capillaire et fluorescence

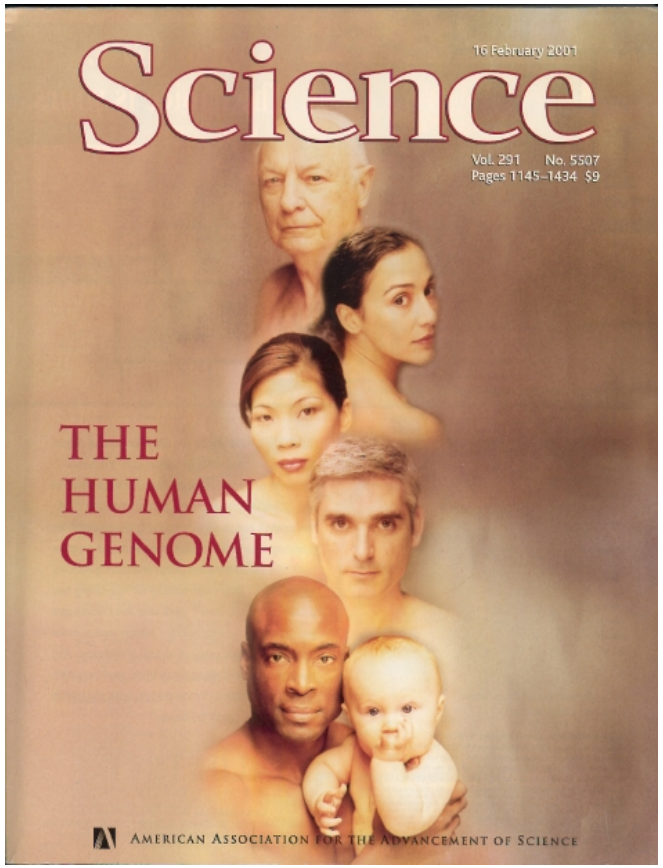


X 300





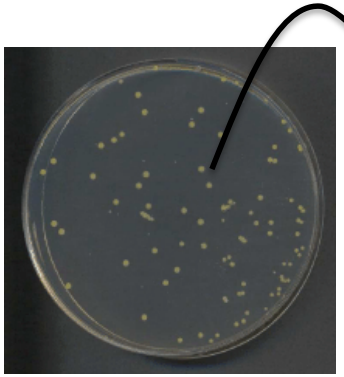
Le génome humain :



- Première séquence du génome humain en 2001
- Coût total de l'ordre de 1 milliard de $

Les nouvelles technologies de séquençage

Séquençage Sanger

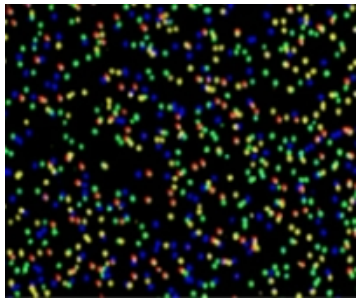


1 ADN

1 réaction de séquence

96 séquences 800 bases

Séquençage Nouvelle Génération



N ADN amplifiés

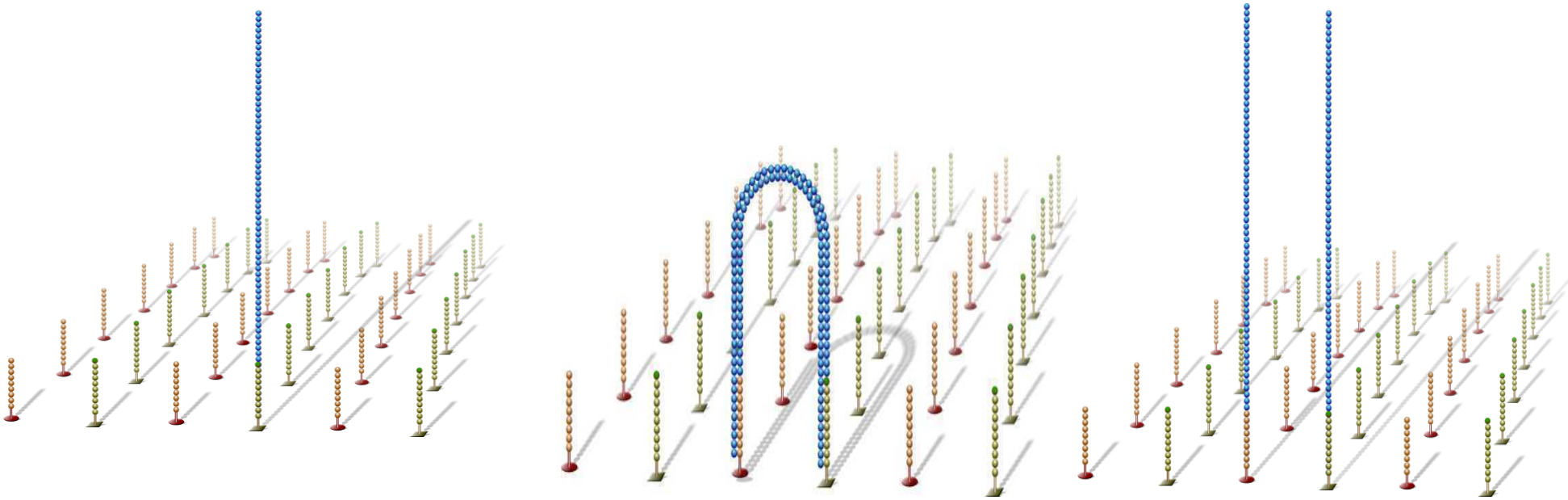
N réactions de séquence et séquençage de 150 bases

HiSeq2500 - Illumina

Clonage in vitro

Séquençage

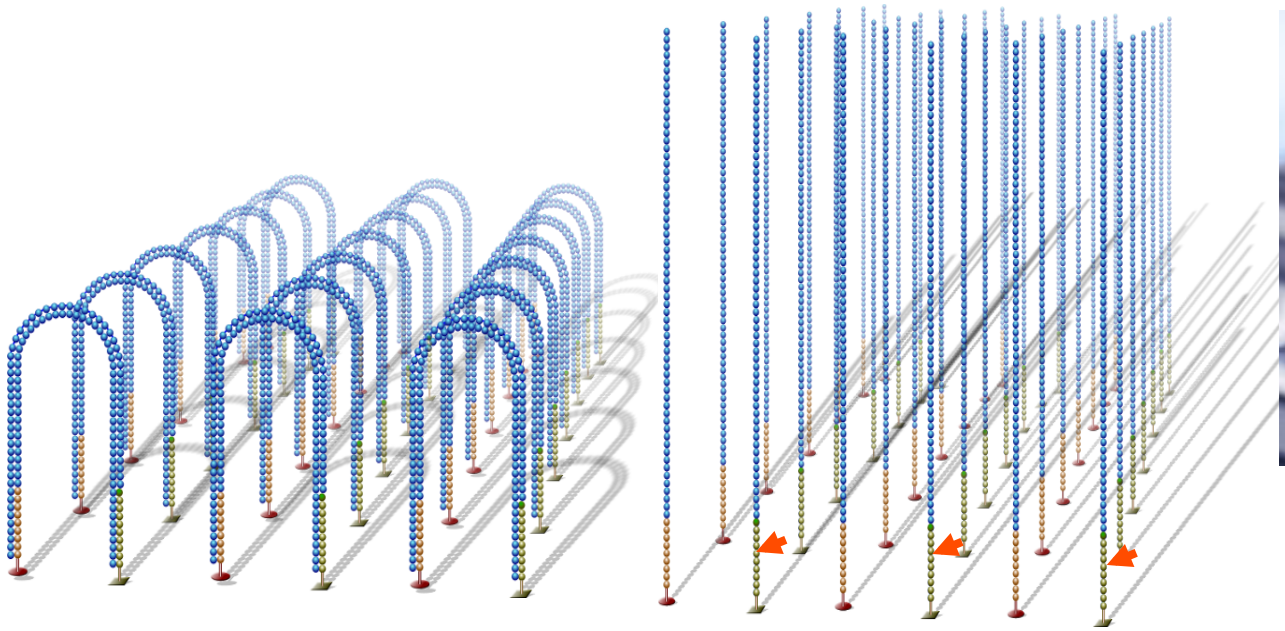
Le clonage *in vitro*: le recopiage de molécules d'ADN

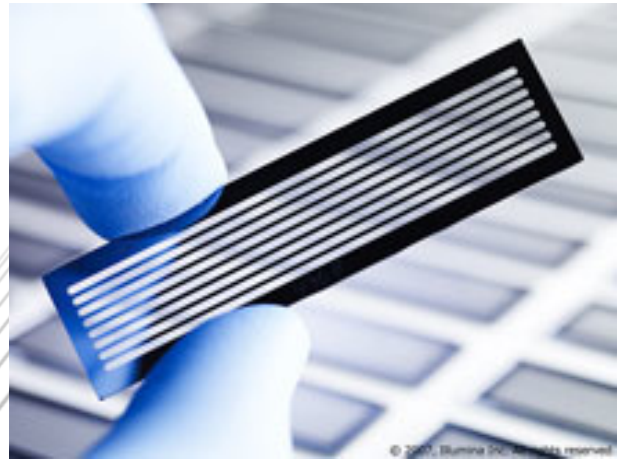


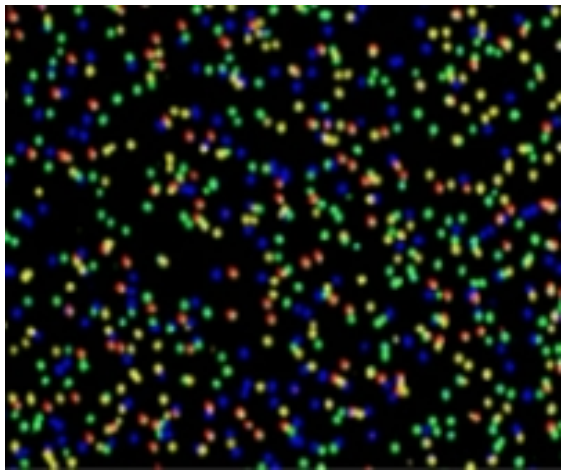
Fixation de molecules
sur une lame de verre

Chaque molécule d'ADN est recopiée localement
par l'ADN polymérase

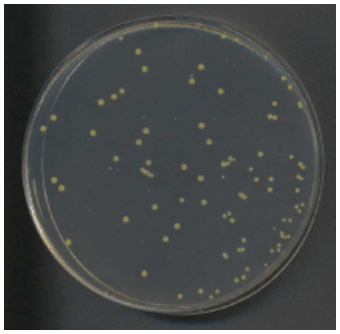
Le clonage *in vitro*: le recopiage de molécules d'ADN







Le processus réitéré permet d'obtenir
1000 molécules identiques



1 milliard de "clones"
sur une lame

Séquençage par détection de la base incorporée

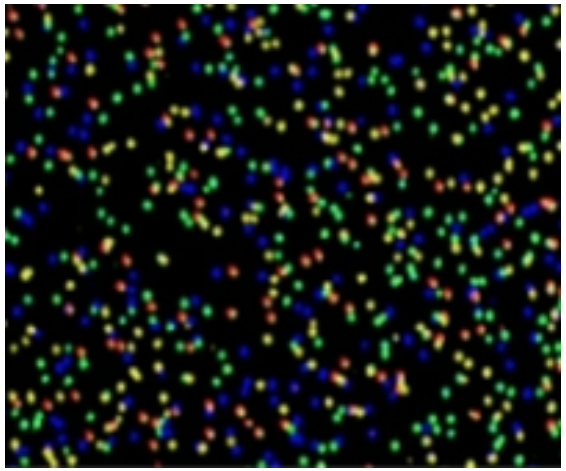
ACGTGTGTCAGTGCTA
TGCACACAGT

+ **A**ppp **C**ppp **G**ppp **T**ppp

Recopier une base bloquée

ACGTGTGTCAGTGCTA
TGCACACAGT**C**

+ **A**ppp **C**ppp **G**ppp **T**ppp



Prise d'image

Séquençage par détection de la base incorporée

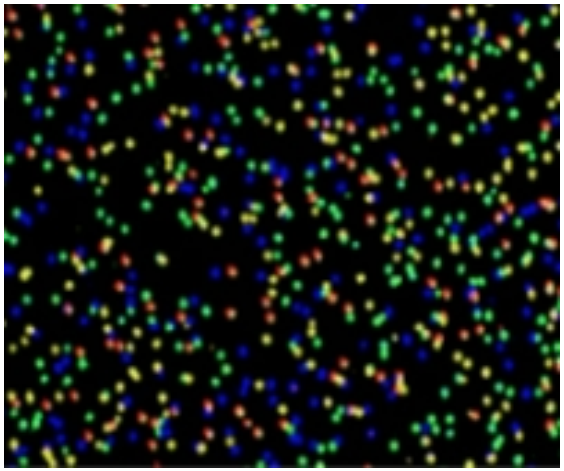
ACGTGTGTCAGTGCTA
TGCACACAGT**C**

Déblocage et élimination de l'étiquette fluorescente

ACGTGTGTCAGTGCTA
TGCACACAGT**C**

Nouveau cycle de synthèse

\+ **A**ppp **C**ppp **G**ppp **T**ppp



Prise d'image

ACGTGTGTCAGTGCTA
TGCACACAGT**C****A**

Grands instruments ou usines

- The Beijing Genomics Institute (BGI)
- 140 séquenceurs
- 300 génomes humains par semaine

Quelques chiffres

- 2 00 000 000 réactions en parallèle
- 250 bases lues
- 180 fois le génome humain
- 1 semaine
- 5000 $ pour un génome humain

Deux types de séquenceur

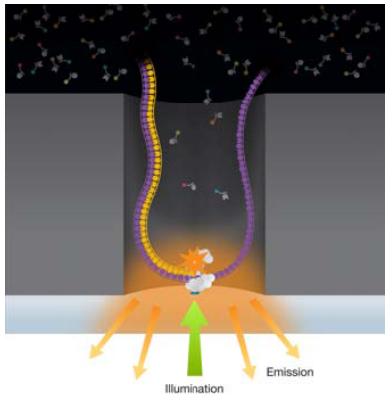
HiSeq (Ion Proton): très haut débit
2 fois 500 Milliards de bases en 6 jours
⇒ Séquence de 30 génomes humain
⇒ Séquence de 20 000 *Escherichia coli*

MiSeq (Ion Torrent): moyen débit - Flexible
Jusqu'à 15 Milliard de bases en 2 jours
⇒ Séquence de 30 génomes d'*E. coli*
⇒ Application en recherche et dans le diagnostic

Le séquençage de 3ème génération

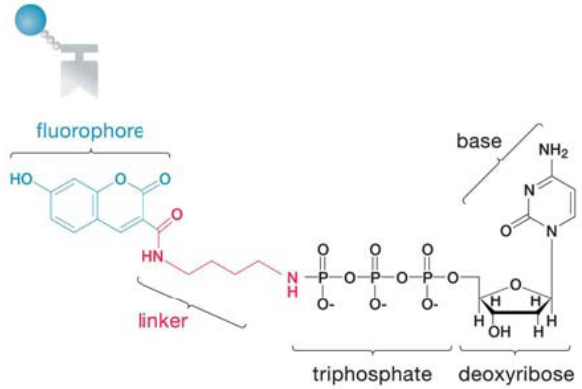
- Molécule unique sans amplification
- Lectures longues (1000 nucléotides)
- Identification des modification des bases
- …

Pacific bioscience – SMRT technology



Zeptochamber

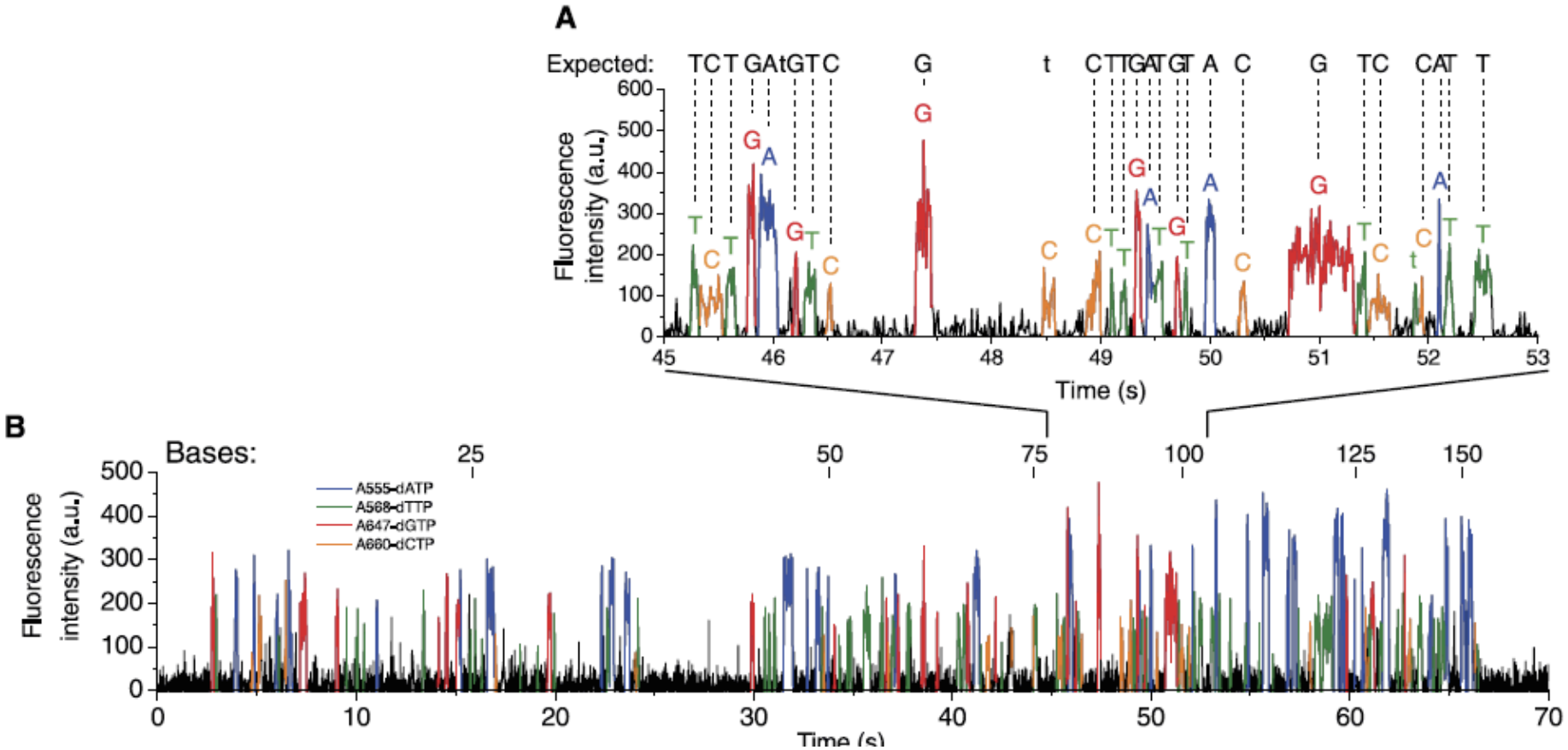
Technologie des semi-conducteurs.
Visualisation: chambre nanophotonique de 20 zeptolitres (10^{-21} l)



Marquage fluorescent des nucléotides tri-P

La polymérase est immobilisée

Observation d'une polymérase unique ajoutant des nucléotides (1-3 bases / sec)


A
Expected: TCT GAtGT C G t CTTGATGT A C G TC CAT T
Fluorescence intensity (a.u.)
Time (s)
B
Bases:
A555-dATP
A568-dTTP
A647-dGTP
A660-dCTP
Fluorescence intensity (a.u.)
Time (s)

Pacific bioscience

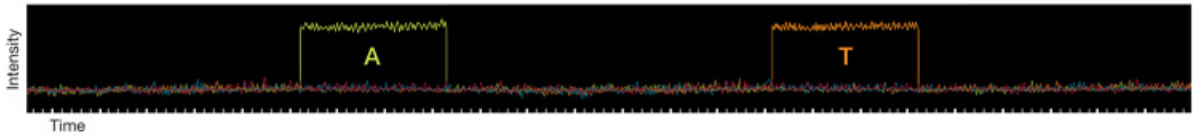
Spécifications

23 000 lectures

1 heure

> 5000 bases

Exactitude: 90%!



Message 3.

- Le séquençage de l'ADN a été inventé il y a 36 ans
- Le premier génome humain a été séquencé en 2001
- De très nombreuses innovations entrainent une augmentation de la vitesse et une diminution des couts
- Le séquençage du génome humain pour 1000$ devient une réalité
- Les étapes limitantes sont l'informatique et l'analyse des données

Evolution du séquençage

- Séquençage Sanger: production de données et de données de référence
  - Ces données sont stockées dans des banques de séquences (Genbank, EMBL)
- Séquençage nouvelle génération: méthode de laboratoire d'observation (sorte de microscope moléculaire) donc d'étude de mécanismes

La bioinformatique

```
@M01626:70:000000000-A6E12:1:1101:14394:1725 1:N:0:5
GCACGATTACTTGCTCTACGCTCTTTACGACGCTTTTCATCAACTTTCCATTGAATTTTTTTCTTGTAACCTGGTTTTATTTTTTTCTTTTTCTTTTTCACAAGTCCAATCATT
TCTGTATCTAATTTTTGTTAGGACTTTTCACGGTTATTACGACGATCACGGTCGTATGTATCTTGAAATTCACCGTTT
+
1>1>11>>1BDD1AFEF1A11AAAFG3B00A00AAABF2FF21AABBBA2AD111DFGDEAEADA1D12DFE00BB0F2BFFGHGECFFGFGFGGHHHF2>210B>>11BB1BF
BBF2B2BBF21FGHE1<0B21<1BBGHHF2@//??/@F22//</>F.0<.<.-<.</=0<<=00<0=DH00:=.:.:
@M01626:70:000000000-A6E12:1:1101:15300:1737 1:N:0:5
TTCAATACAAGAGCCCCTTACTTCCTTTTTCTATACCTTTTTCCGTAATTTCCATAATCGCCGTAAGTAGCAACAGATTCATTAACTTTATTAAGATTAATACCTAAGAACTTT
GAACCACTTTGTTCCATCTGTTCTTTTGCTTTTTCAACATAATTACGTTTTATTCTGCCTGCTTGGGTTACTAAAATAAAACCATCCCAACCATTAGCGCTTATTGCGGCATCA
ACAACTAAACCAATCGTTTGAGTATCCATGTTAATTTAAACATCGGACATCCCACCCCCTTCCCCCCCTCCC
+
>AA11D3BB1111AFEAAF1AAFEAFEGGHCF3D3B2DFGGHGF0B0/ADBFF2F21F2//A/A//0D22BA1B000AB@1DF2DFGGG2@D2@112BB1F2FCF11011FFGF
121B?0?BFG1FGFF2BBB1BFFGGFHFB>GGHFB22F1<D1BG2BA<GFF2GGFH2@<G1@CG1A/AF/1@1111?11F1D..>>0.1>.>.<DD0/<-<<.=<0::-;-:/0
0:0..;9C0CFA.//;...=0.;00;00000000///9/9///9-9--9//;9/-;--;---/--:9--;-;
@M01626:70:000000000-A6E12:1:1101:16123:1748 1:N:0:5
```

- Le séquenceur produit des fichiers FASTQ:
  - La séquence nucléotidique
  - Des valeurs de qualité pour chaque base

⇒ Fichier de plusieurs millions de lectures

⇒ Les logiciels utilisent ces fichiers en entrée

Séquençage génomique

Deux approches:

- Séquençage de novo: assemblage de toutes les séquences pour obtenir la séquence d'un organisme
- Re-séquençage: alignement des séquences individuelles sur une référence pour détecter les différences, les mutations, les recombinaisons, les duplications …

Application du séquençage d'ADN

Caractérisation d'une espèce vivante

nature COMMUNICATIONS

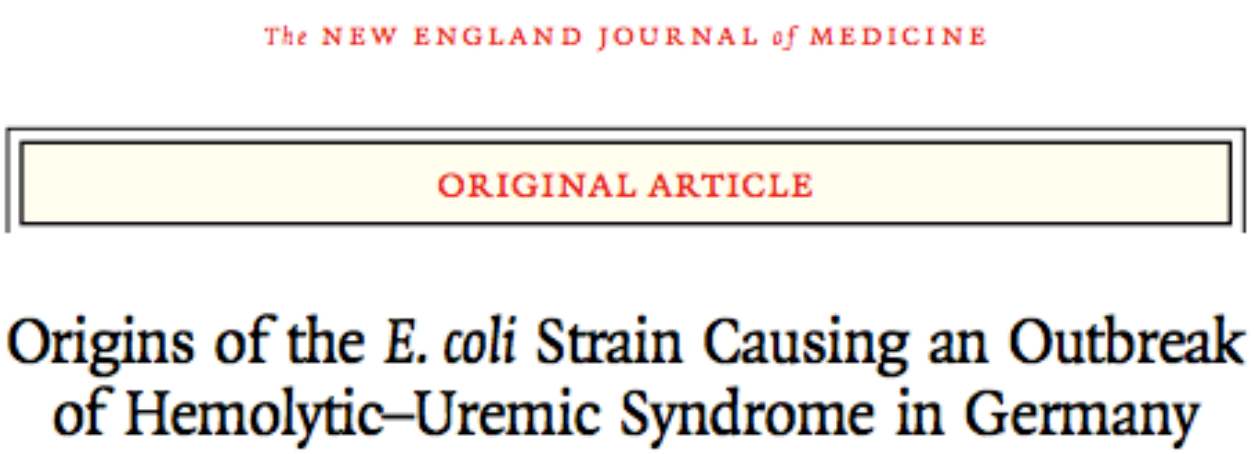
ARTICLE

Received 7 Aug 2013 | Accepted 31 Mar 2014 | Published 6 May 2014

DOI: 10.1038/ncomms4765 OPEN

Spider genomes provide insight into composition and evolution of venom and silk

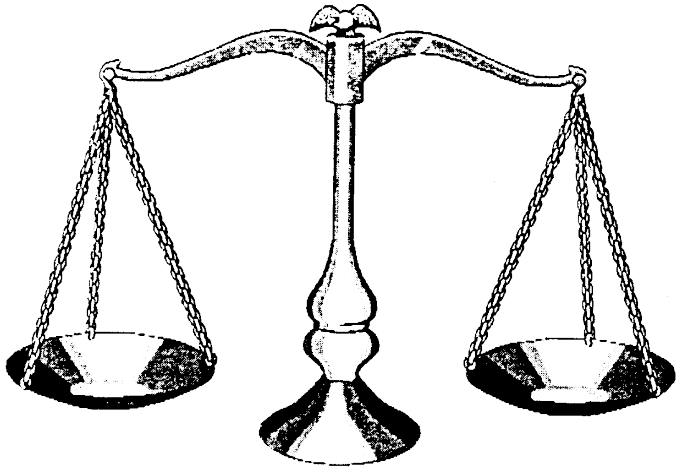
Caractérisation d'une espèce vivante



- Souche d'*Escherichia coli* responsable l'épidémie de syndromes hémolytiques urémiques de 2011 en Europe.
- Compréhension de la spécificité de cette souche, de sa virulence et de sa capacité épidémique.
- Mise au point d'un outil de diagnostic

Application du re-séquençage: L'évolution bactérienne et la virulence

Le streptocoque du groupe A



- Bactérie commensal inoffensive
- Infections bénignes : angines

Infections invasives gravissimes:

- Fasciite nécrosante
- Choc sceptique

Analyse de la transition commensal => pathogène invasif

- Séquençage de groupes d'isolats d'infections invasives ou de portage provenant de l'entourage
- Génome de 2 000 000 bases – séquençage Illumina
- Deux cas avec une seule différence dans un gène même régulateur

=> Ces mutation entraine une plus forte expression des gènes de virulence dans l'isolat de portage

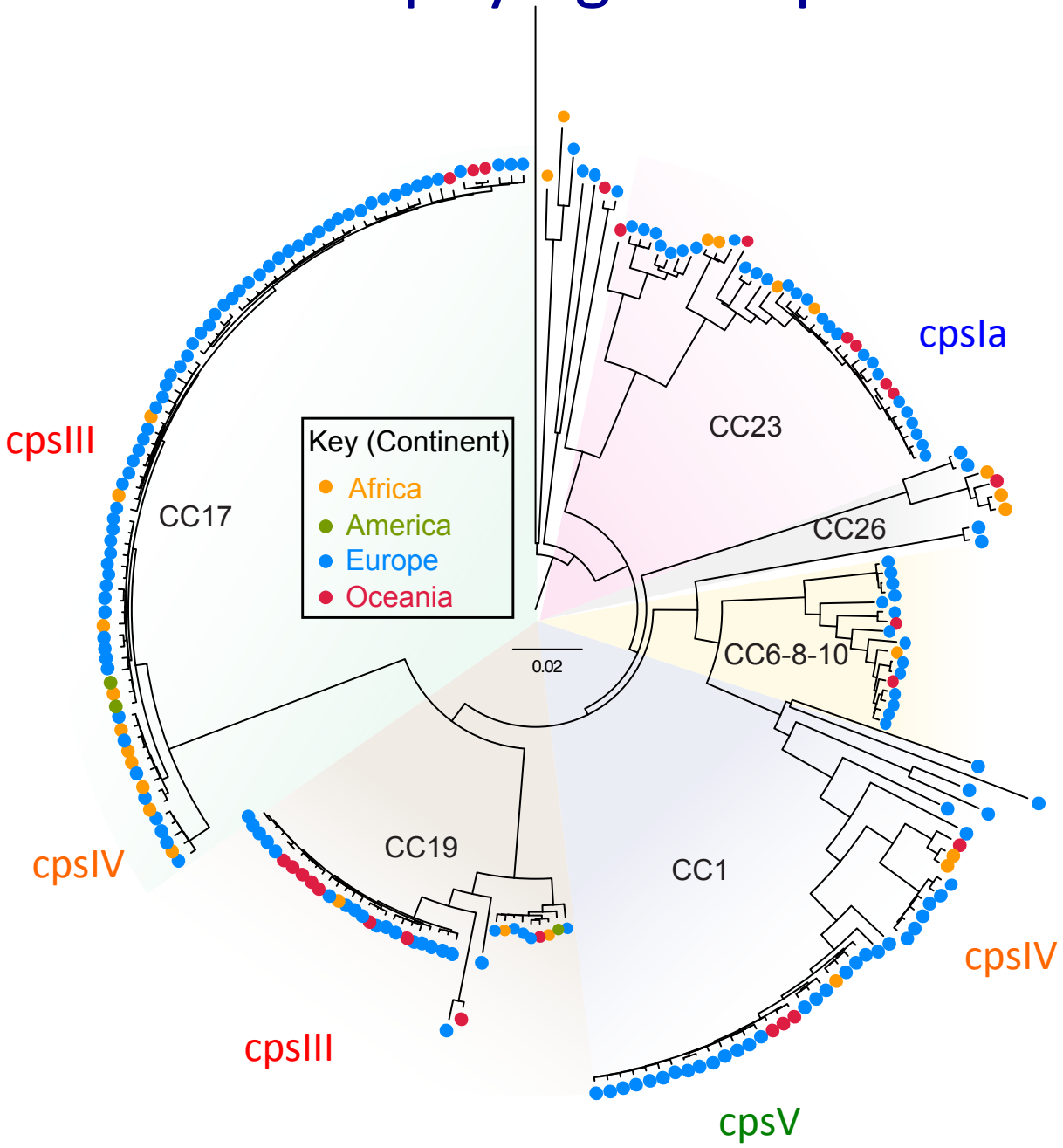
Coll Claire Poyart, CNR des streptocoques

Répondre à une question d'épidémiologie

- L'émergence des infections néonatales à *Streptococcus agalactiae* (streptocoque du groupe B) dans les années 60 en Europe et aux Etats-Unis

- Méthodologie:
  - séquençage complet de 230 souches isolées de 4 continents sur plus de 50 ans.
  - Assemblage *de novo* et détection de mutations (re-séquençage)
  - Phylogénie basée sur les polymorphismes entre ces souches

Relations phylogénétiques entre les 230 souches

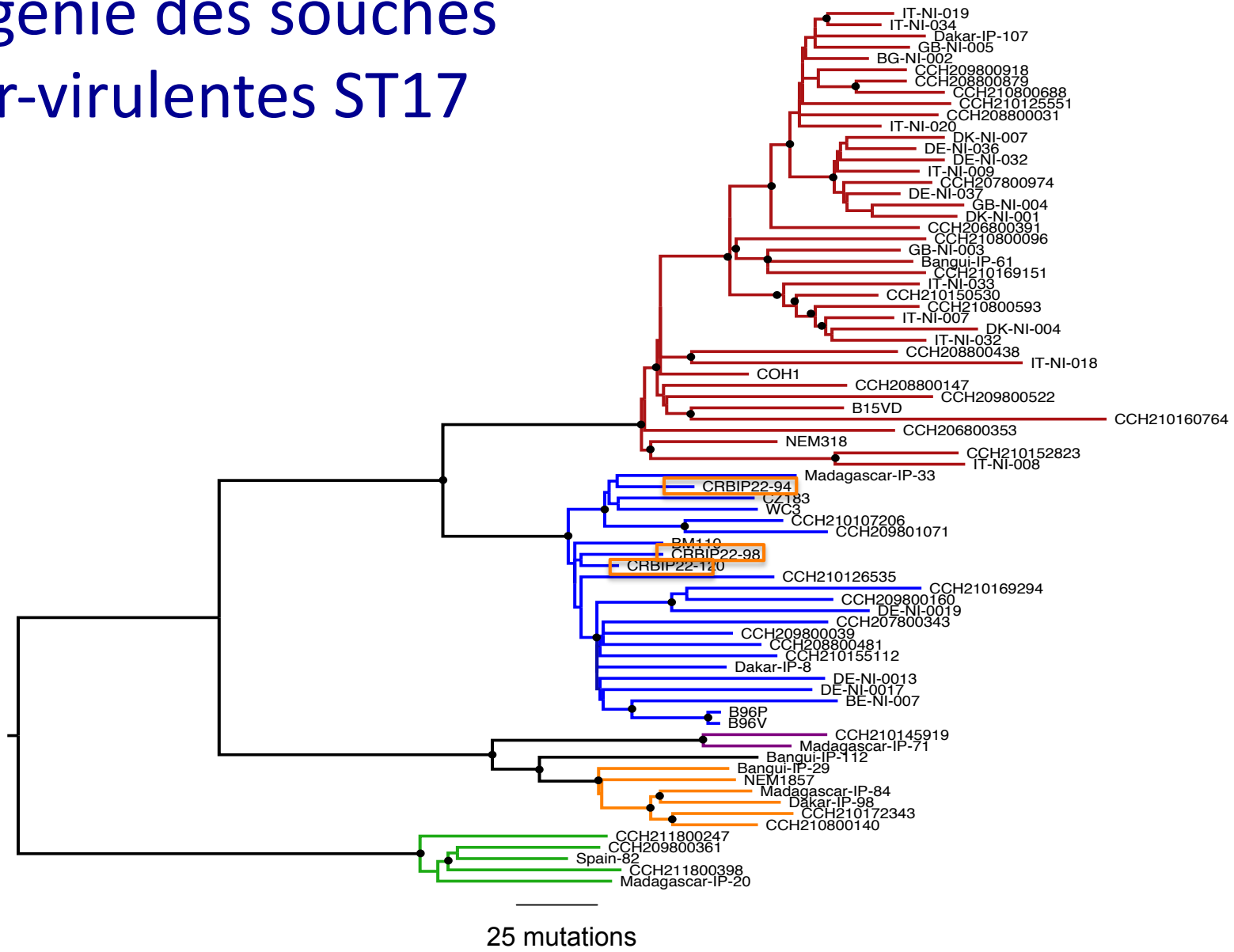


⇒Des lignages bien définis

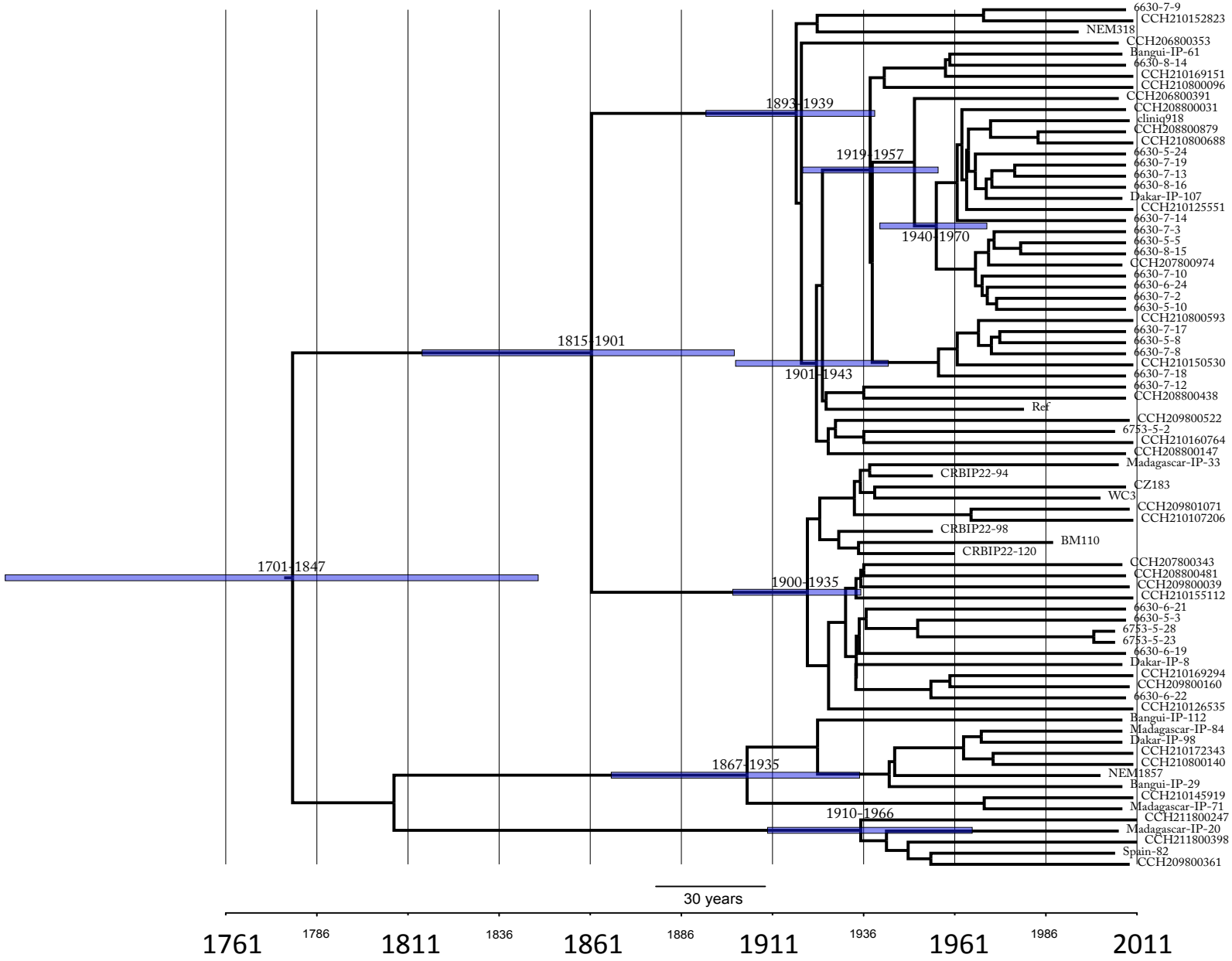
⇒Présence de clones dominants dans chaque lignage très peu divers

⇒Les branches plus longues correspondent à des recombinaisons

Phylogénie des souches hyper-virulentes ST17

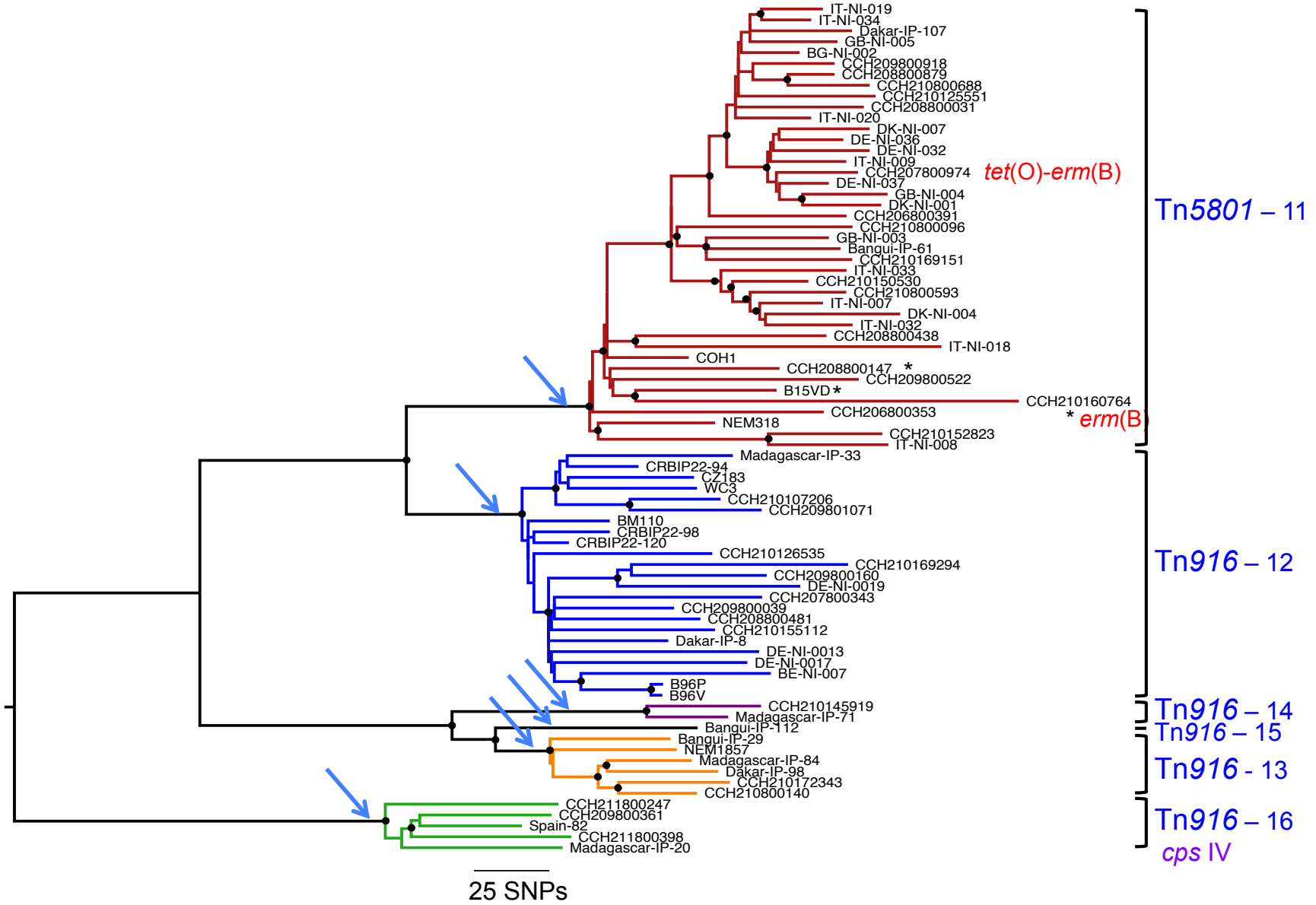


Très petit nombre de mutation => Origine très récente de ces lignages



L’analyse statistique permet de dater l’origine de ces clones

- L’émergence des infections néonatales à streptocoque B correspond à l’expansion d’un petit nombre de clones
- Les années 50 - 60 correspondent à la généralisation de l’utilisation des antibiotiques
- 90% des souches humaines de *S. agalactiae* sont résistantes à la tétracycline



- Chaque clone résulte de l'insertion d'un transposon codant pour la résistance à la tétracycline
- Ces clones provenant d'une seule bactérie se sont disséminés mondialement

Conclusions

- L'étude génomique montre que l'émergence des infections néonatales à *S. agalactia*e est réelle et non la conséquence d'une meilleure détection.

- L'utilisation massive de la tétracycline dans les années 50 a résulté dans le remplacement de la population par quelques clones résistants

- Malgré la diminution de l'utilisation de cet antibiotique ces clones sont stables

Application du séquençage en microbiologie clinique

- Compréhension et surveillances des maladies infectieuses et des épidémies
- Surveillance des maladies nosocomiales
- Analyse de la résistance aux antibiotiques et sa dissémination (antibiogramme moléculaire)
- Diagnostic en microbiologie et en virologie
- Alerte pour le bio terrorisme

…

Etudes des maladies rares

- Hypothèse: une mutation est responsable du syndrome
- Analyse familiale, étude de grandes familles, étude de familles consanguines
- Difficultés:
  - Maladies très rares
  - Mutations dans différents gènes responsables du même syndrome.

Exemple d'identification de mutation

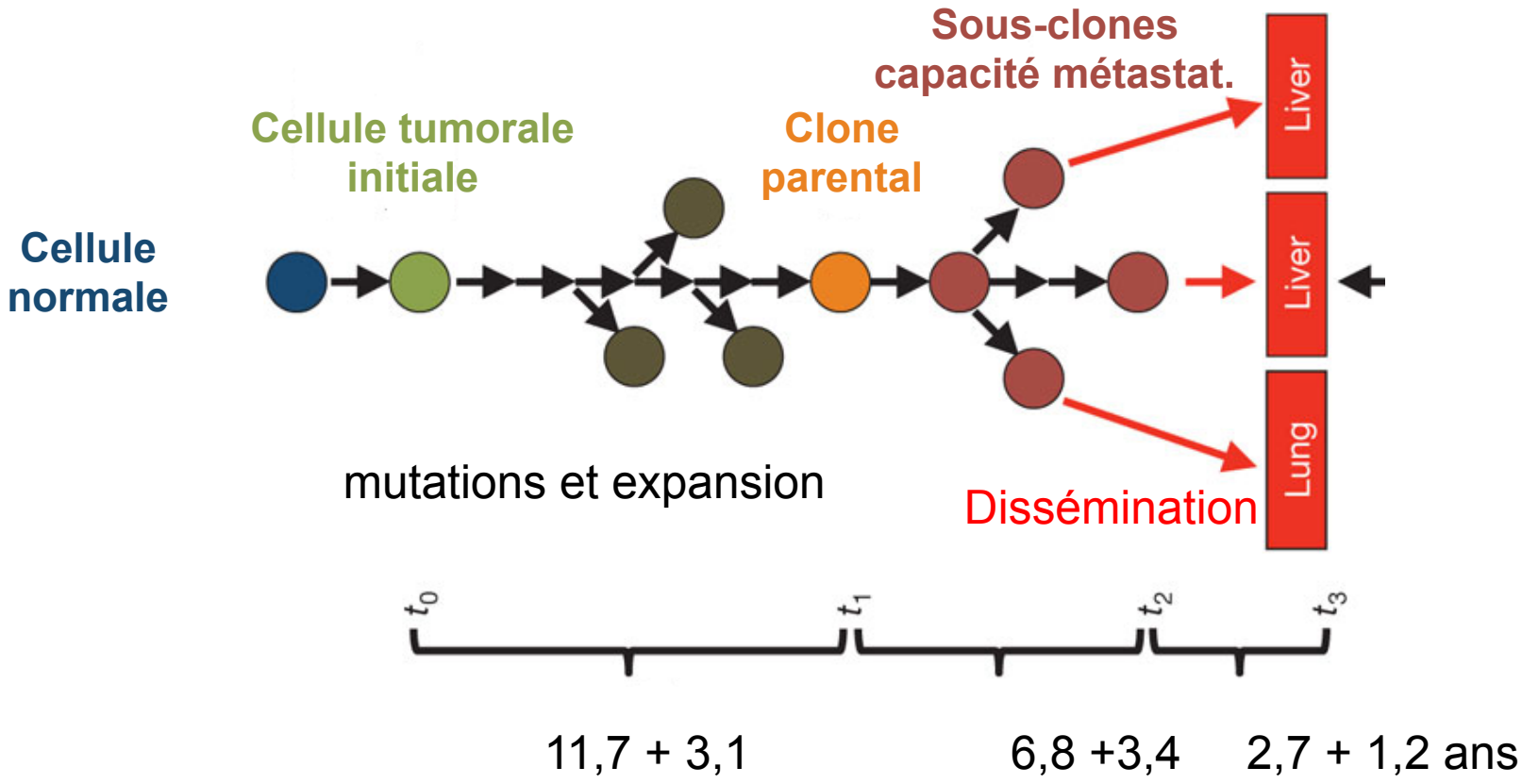
- Une famille de quatre enfants sains, mais trois d'entre eux perdent peu à peu la vision – syndrome de rétinite pigmentaire.
- Test négatif pour plus de 50 mutations connues associées à ce syndrome.
- Proposition du généticien de séquencer les génomes au sein de cette famille

$\Rightarrow$ Identification d'une mutation dans le gène DHDDS qui ajoute un sucre à la rhodopsin

$\Rightarrow$ L'inactivation de ce gène chez le poisson zèbre donne le même phénotype

Analyse du développement d'un cancer

⇒ Principe: séquençage de l'ADN de tumeurs et de métastases



Développement d'un cancer pancréatique

Application du séquençage pour la santé humaine

- Génétique: identification de mutations responsable de maladies mais aussi de la prédisposition à des pathologies
- Caractérisation des mutations au cours du processus cancéreux
- Potentiel considérable de la médecine personnalisée
- Identification de variants génétiques associés à des traits phénotypiques: taille, couleurs de la peau … psychologie
- Identification des personnes pour la police scientifique

Le risque éthique

- Séquençage de son génome à titre privé
- Comment utiliser les informations de « prédisposition »
- Risque d'erreur dans l'analyse des données
- Problème de confidentialité

Des applications très innovantes

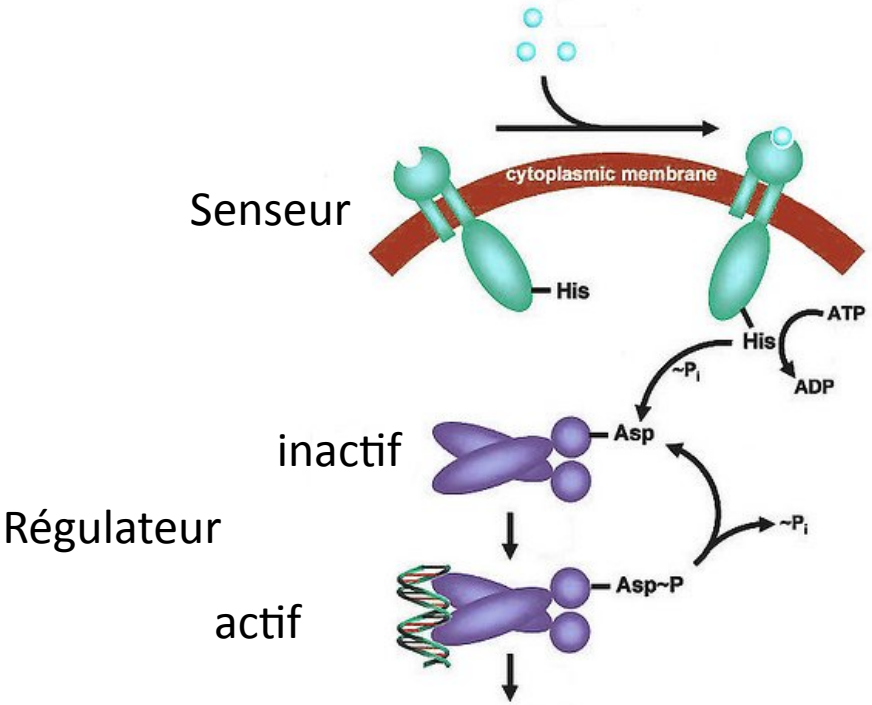
Diagnostic de la trisomie et de certains cancer par séquençage de l'ADN circulant

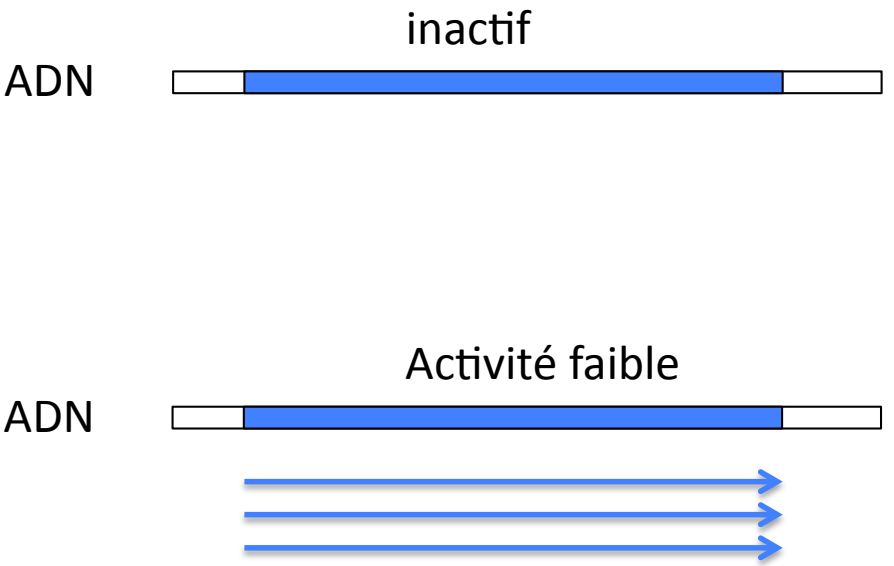
Message 5:

L’analyse de la diversité génétique des individus et au cours du processus cancéreux présente un potentiel considérable en recherche et en médecine mais doit être précisément encadré.

Analyse de l’activité des gènes:
Expression génétique et épigénétique

L'expression génétique et sa régulation les bactéries





- Le séquençage permet de compter le nombre de molécules d'ARN et donc l'activité des gènes.
- Il permet aussi d'identifier les différentes formes d'ARN

Analyse de l'expression génétique

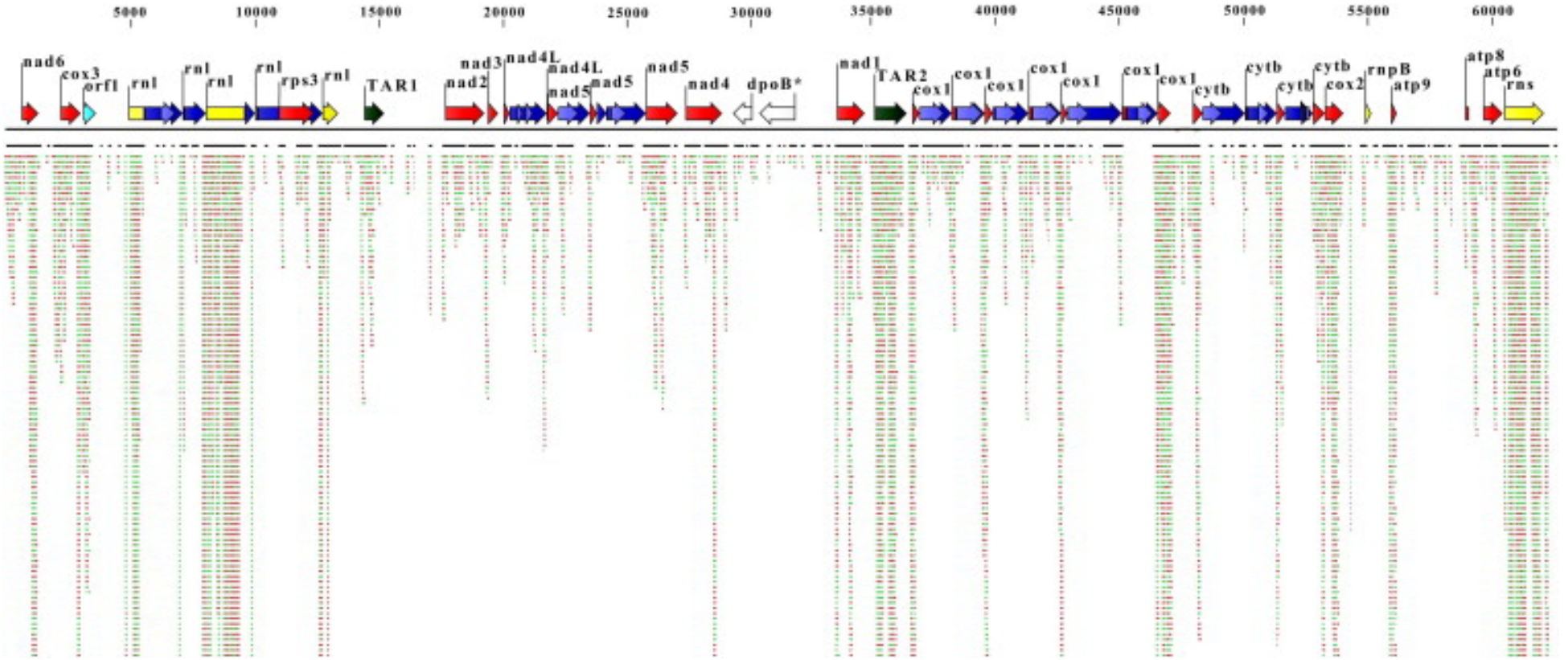
Très nombreux protocoles en fonction du type d'analyse:

- Conversion de l'ARN en ADNc
- Ligation d'adaptateurs en 3' et/ou en 5'
- Autres traitements enzymatiques
- Purification (déplétion des ARN ribosomique, purification des ARN polyA)

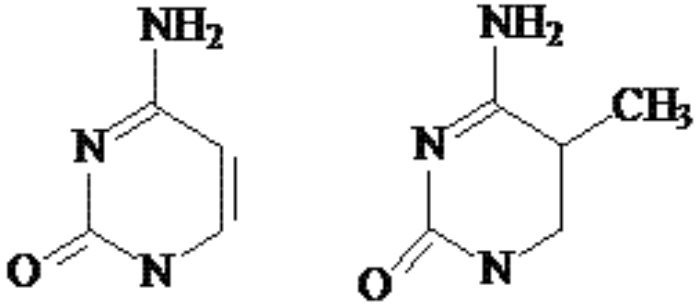
Analyse des ARNs

- Analyse quantitative de l'activité des gènes: comptage des ARNs pour chaque gène (remplace les puces à ADN)
- Identification du début des ARNs et donc des promoteurs
- Caractérisation des épissages alternatifs et quantification des iso-formes
- Découverte et analyse des ARN non codants et des petits ARN (miRNA, piRNA …)

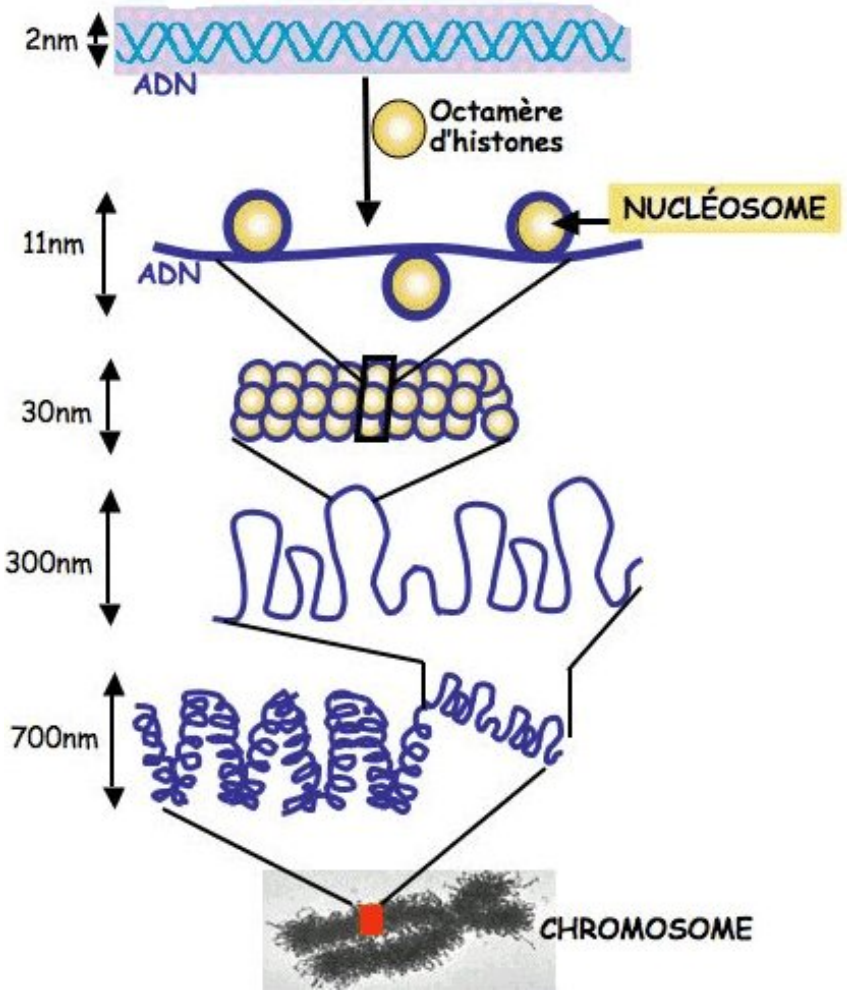
Analyse globale de l'expression génétique



Analyse de la chromatine

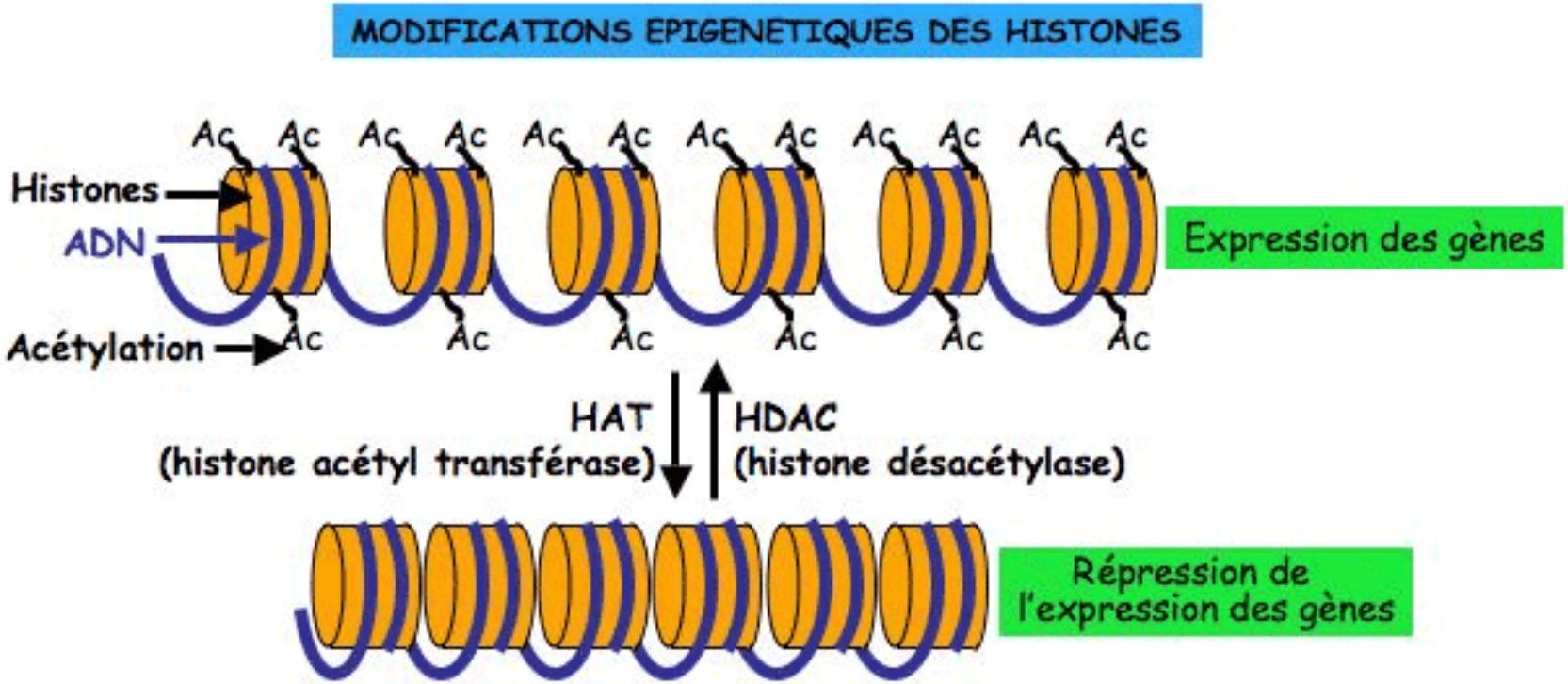


L'ADN peut être méthylé sur les cytosines



L'ADN est associé à des protéines: les histones

La modification des histones module l'activité des gènes



Détection de site de liaison des protéines – analyse de la chromatine

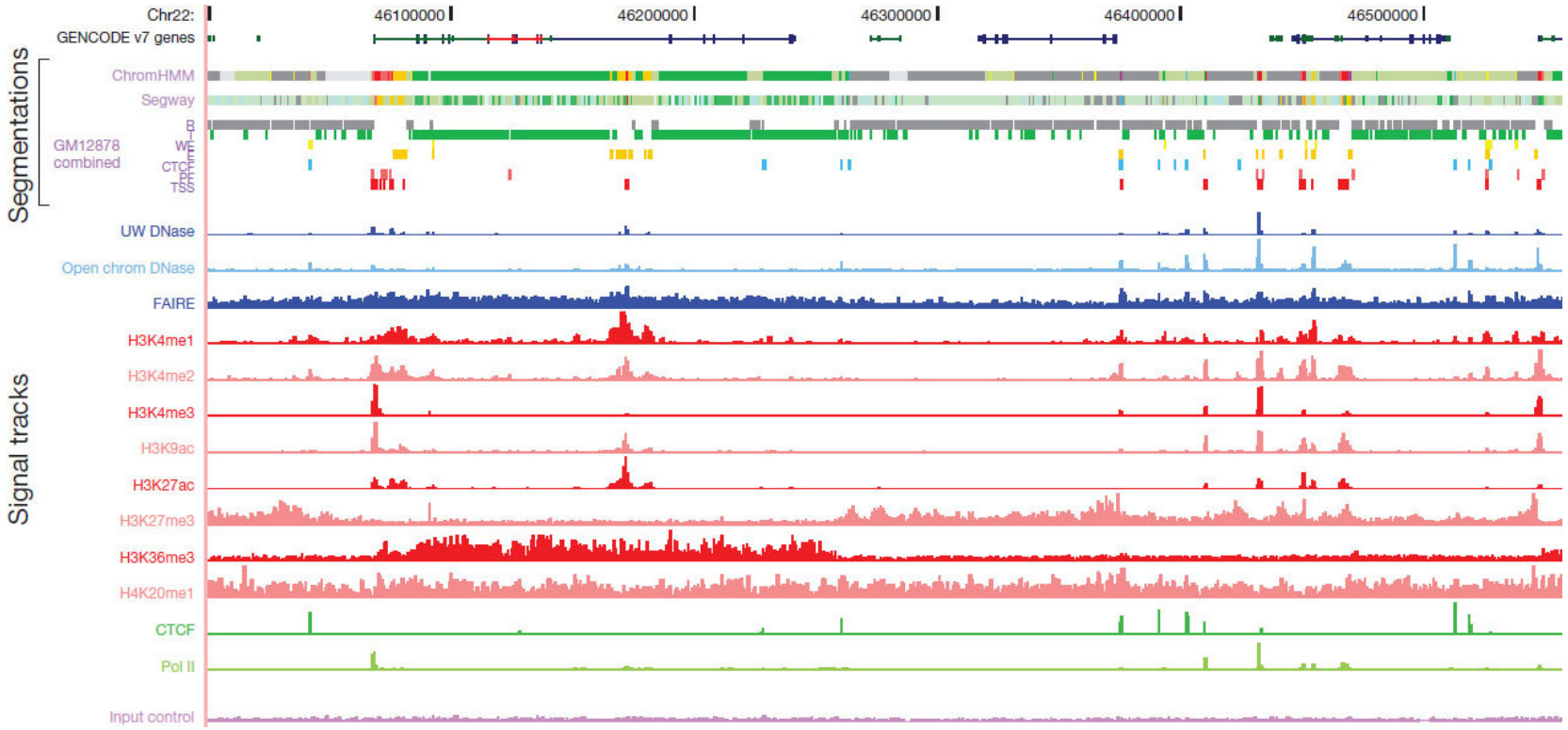
Chromato-immunoprécipatation

Séquençage

ChIP seq

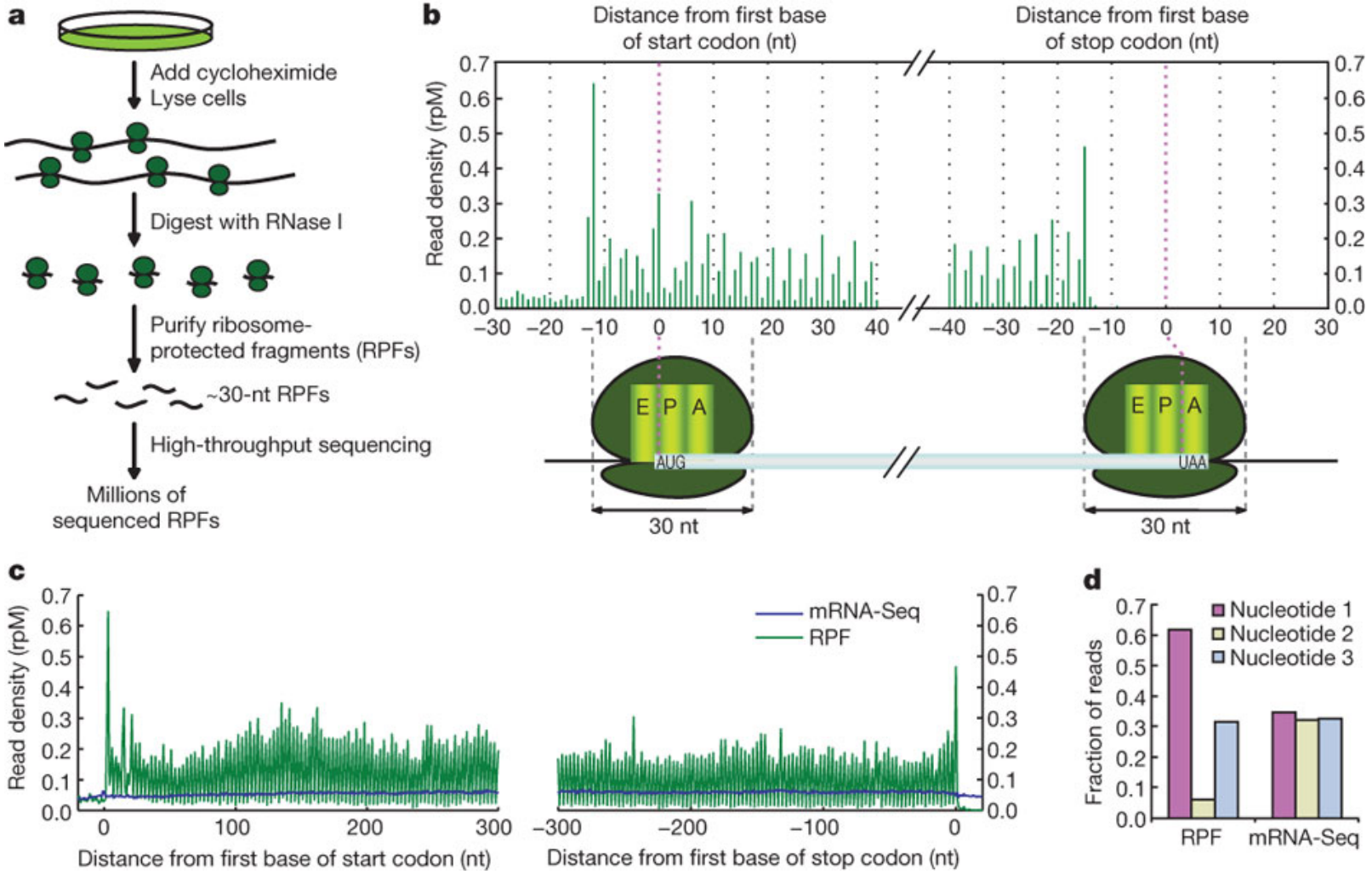
- Analyse des sites de liaison d'un régulateur : découverte de ses cibles
- Analyse des sites de liaison d'histones modifiées : identification des régions de la chromatine modifiée – active ou inactive

Intégration de l'ensemble des informations



- Analyse pour l'ensemble du génome:
  - Du niveau d'activité des gènes – des ARN non codants
  - De la méthylation de l'ADN
  - Des modifications des histones

Analyse de la traduction par ribosome ‘profiling’

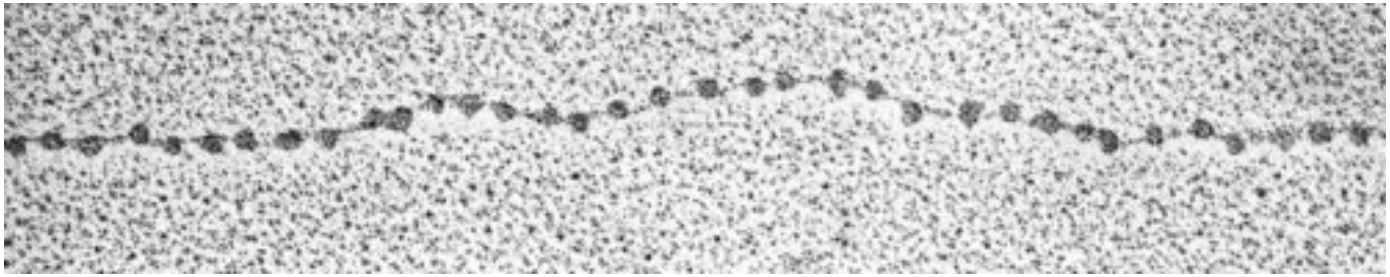


Applications

- Analyse de la réponse de l'hôte à l'infection
- Analyse du développement embryonnaire et la différenciation des cellules souches
- Analyse du développement tumoral et la dédifférenciation cellulaire
- Identification de marqueurs de pronostic pour les cancers (bio-marqueurs)
- Utilisation en médecine personnalisée

Message 6:

Le séquençage haut débit permet de comprendre comment fonctionne une cellule normale, comment le génétique et l'épigénétique interagissent, comment les cellules souches se différencient et comment une cellule devient cancéreuse.



- Les chromosomes sont constitués d'ADN et de protéines quel est le support de l'hérédité?
  - Réponse 1: les protéines
  - Réponse 2: l'ADN
  - Aujourd'hui: ce n'est pas si simple: le génétique, l'épigénétique et l'interaction avec le microbiome

Le microbiome humain

Estomac

Intestin grêle

Colon

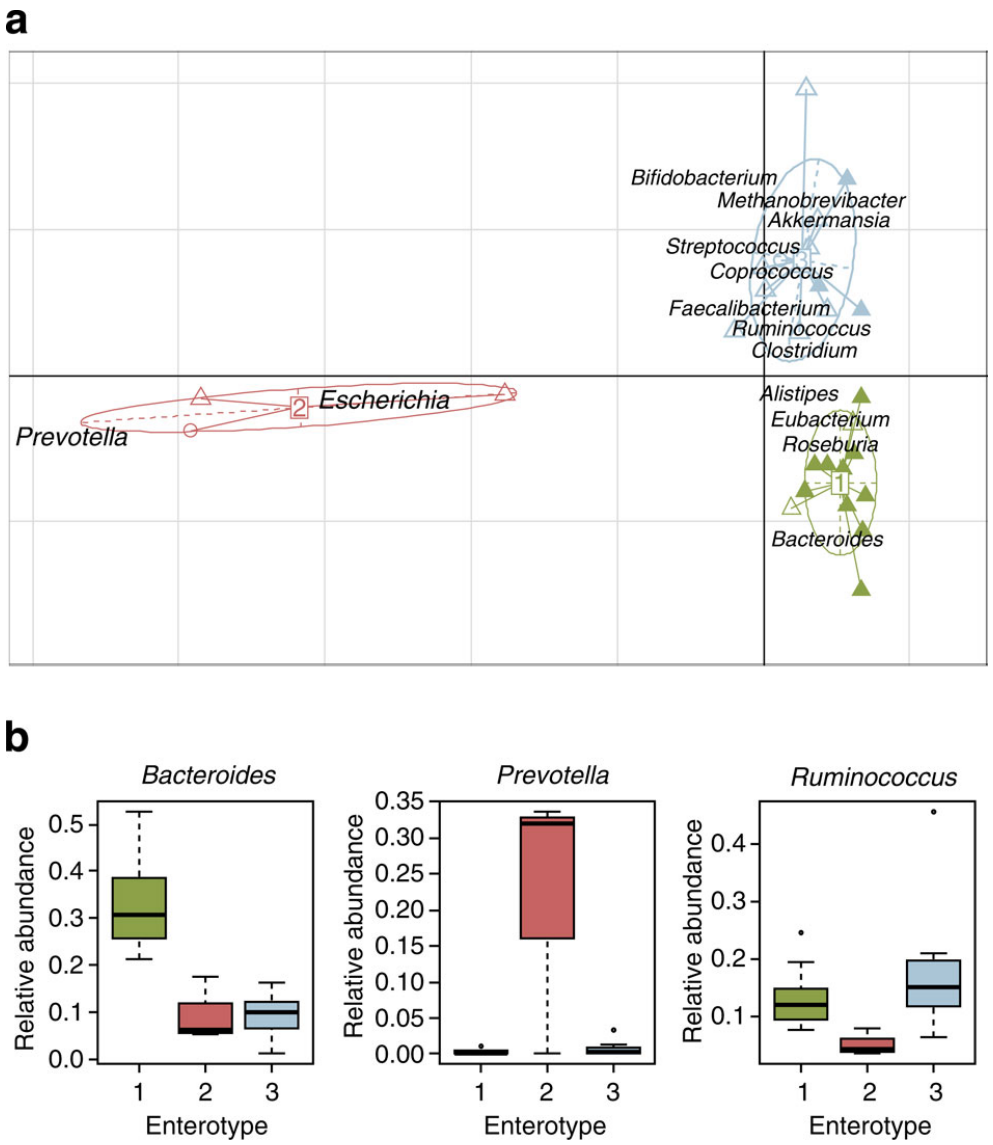
- 10^{14} bactéries (10 fois le nombre de cellule humaine)
- Plus de 1000 espèces
- Métabolisme très complexe et critique pour la nutrition (plus d'un million de gènes découverts)
- Interaction complexe avec le système immunitaire
- Modification de la flore avec l'âge

=> Comprendre l'association entre type de flore et l'état sain ou pathologique (maladie de Crohn, diabète)

INRA

Caractérisation du microbiome

définition des entéro-types



Caractérisation du microbiome

définition des entéro-types

- Plus de 1000 espèces bactériennes avec des abondances très diverses
- Compréhension du rôle de la flore digestive et des interaction avec l'hôte
- Identification d'espèces ou d'activités liées aux bénéfices ou aux pathologies associée à la flore digestive
- Des fonctions multiples ont été associées au microbiote digestif

metagenopolis mgps.eu

INRA

SECOND GENOME
THE MICROBIOME COMPANY

Conclusions

- Le séquençage haut débit apporte une nouvelle dimension en recherche en permettant de transposer les analyses à l'échelle du génome
- Il a permis des découvertes fondamentales dans de multiples domaines
- Il commence a être utilisé dans le domaine de la clinique