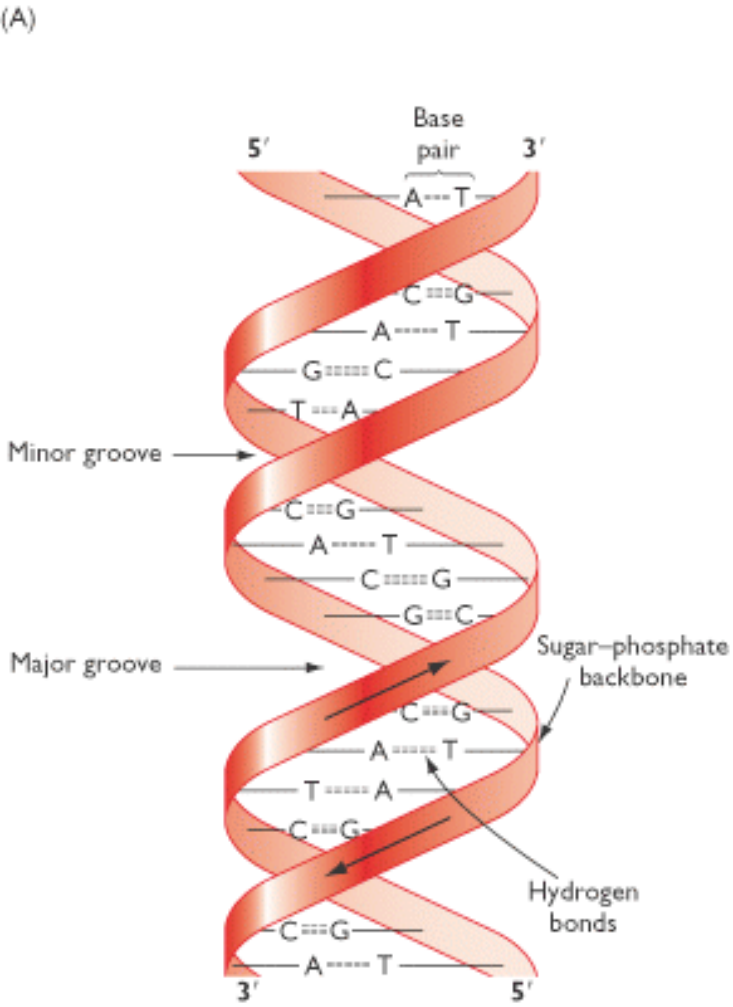
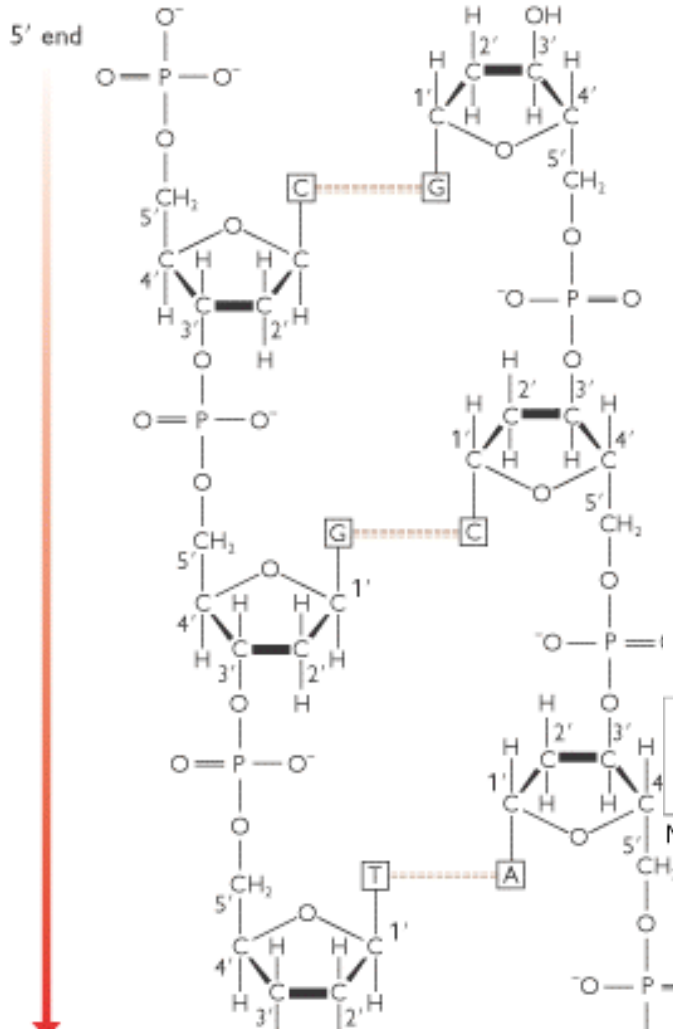
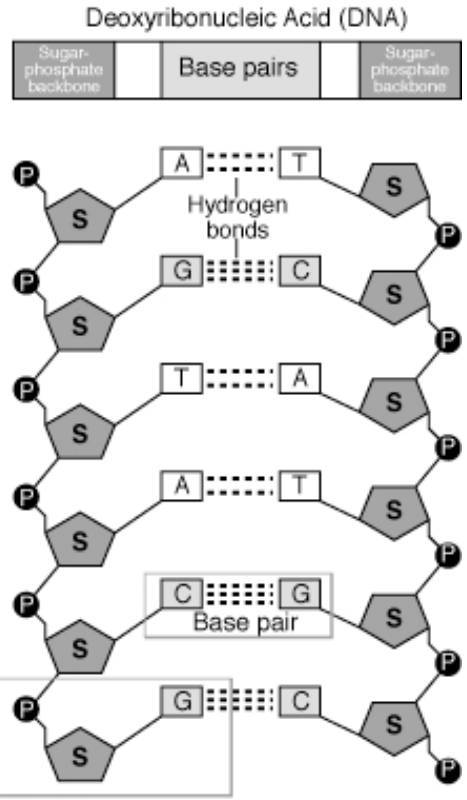
Nouvelles méthodes de séquencage d'ADN

Frédéric FLAMANT
Ecole Normale Supérieure de Lyon
Institut de Génomique Fonctionnelle de Lyon
UMR5242
Equipe neurodéveloppement
Frederic.flamant@ens-lyon.fr

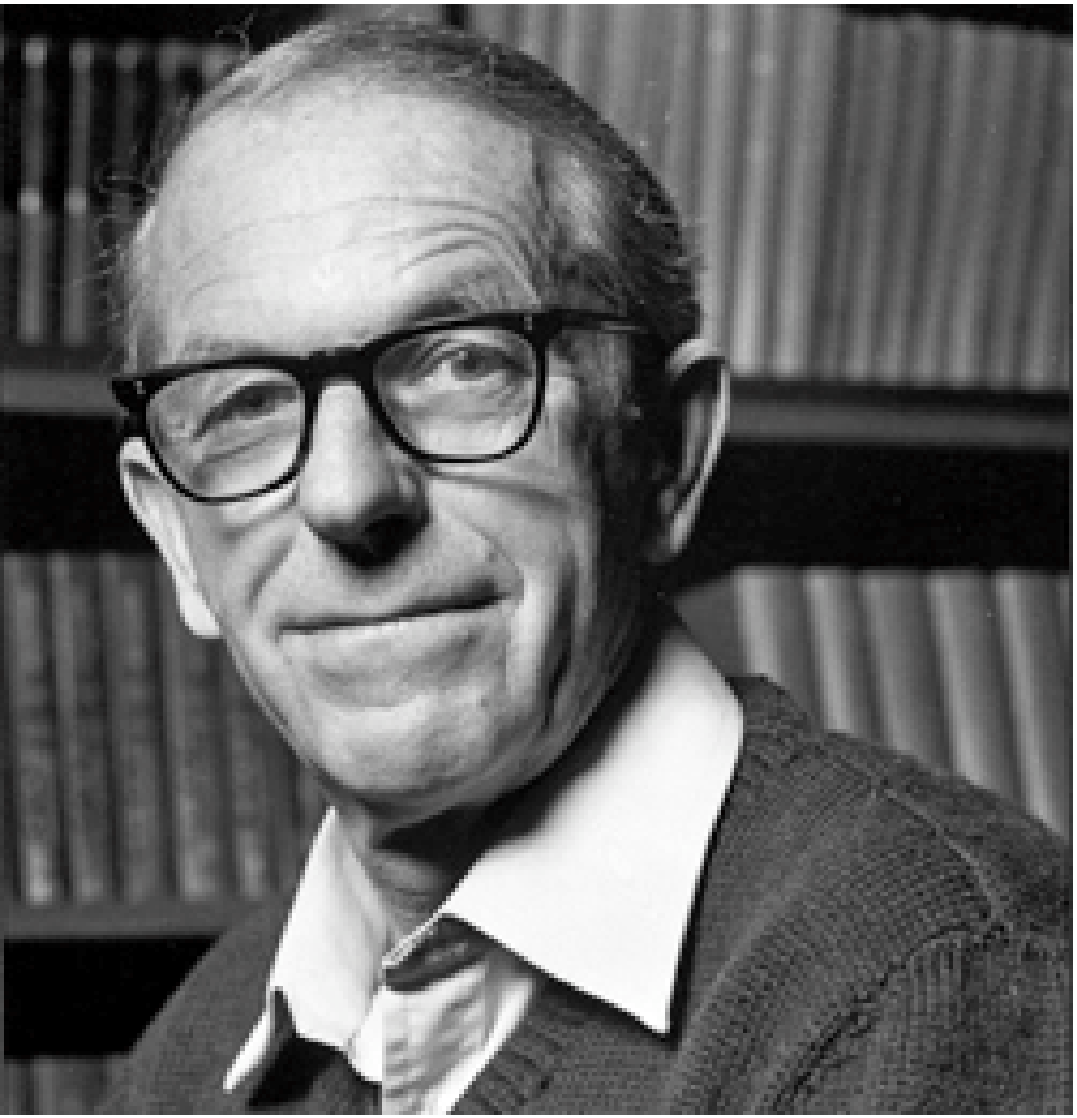
Comment déterminer l'ordre des nucléotides sur l'ADN?





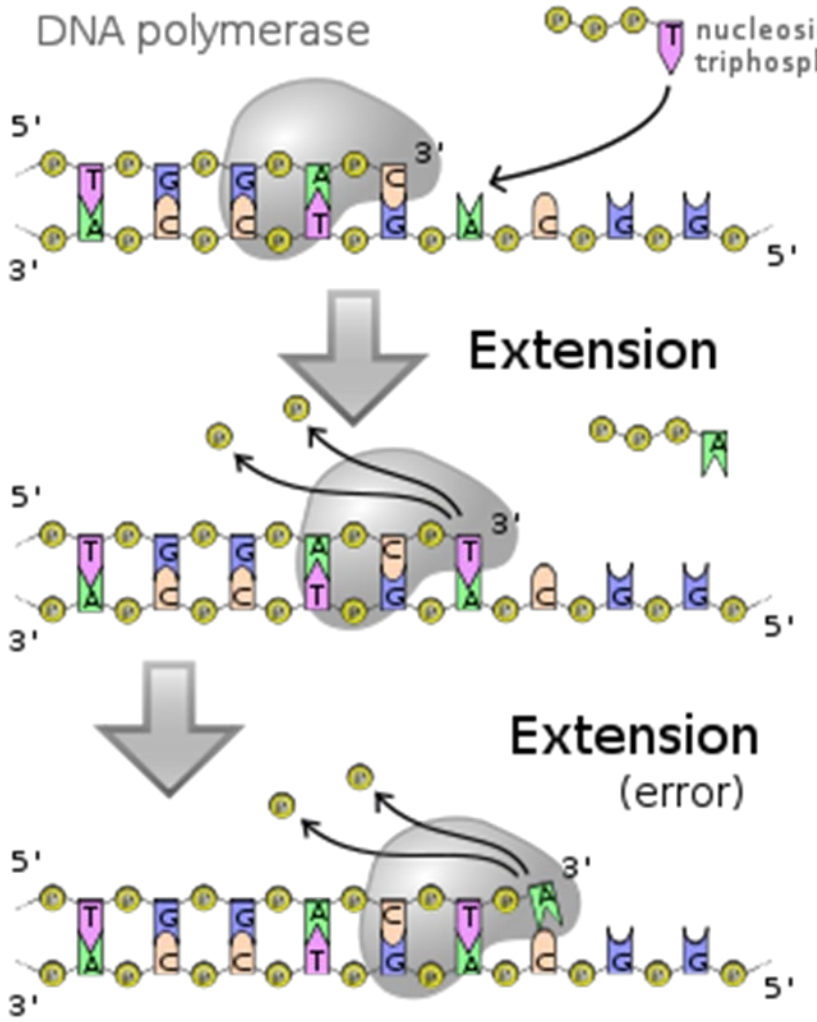


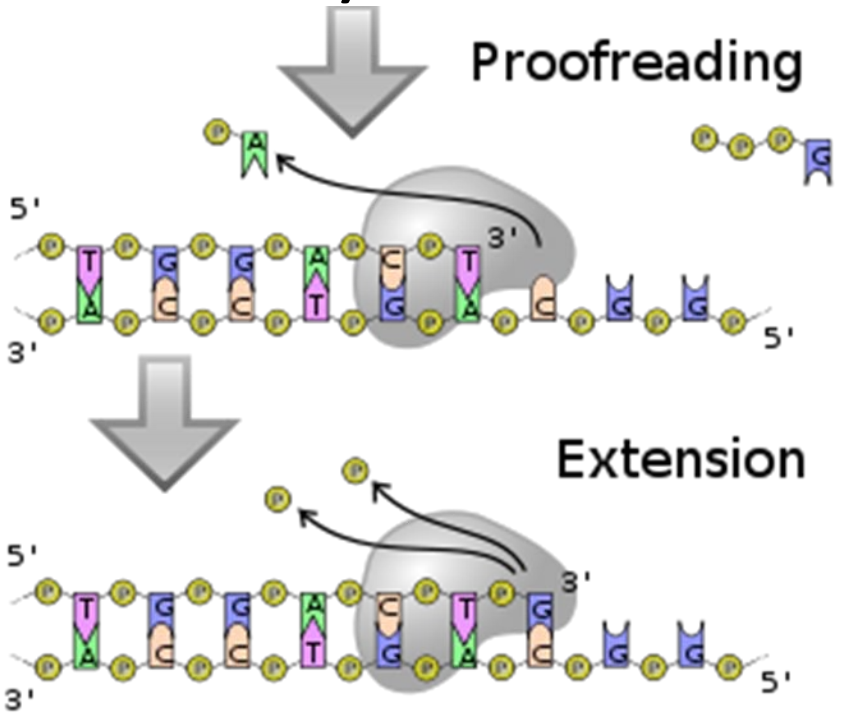
1)Le point de départ: la méthode de Sanger



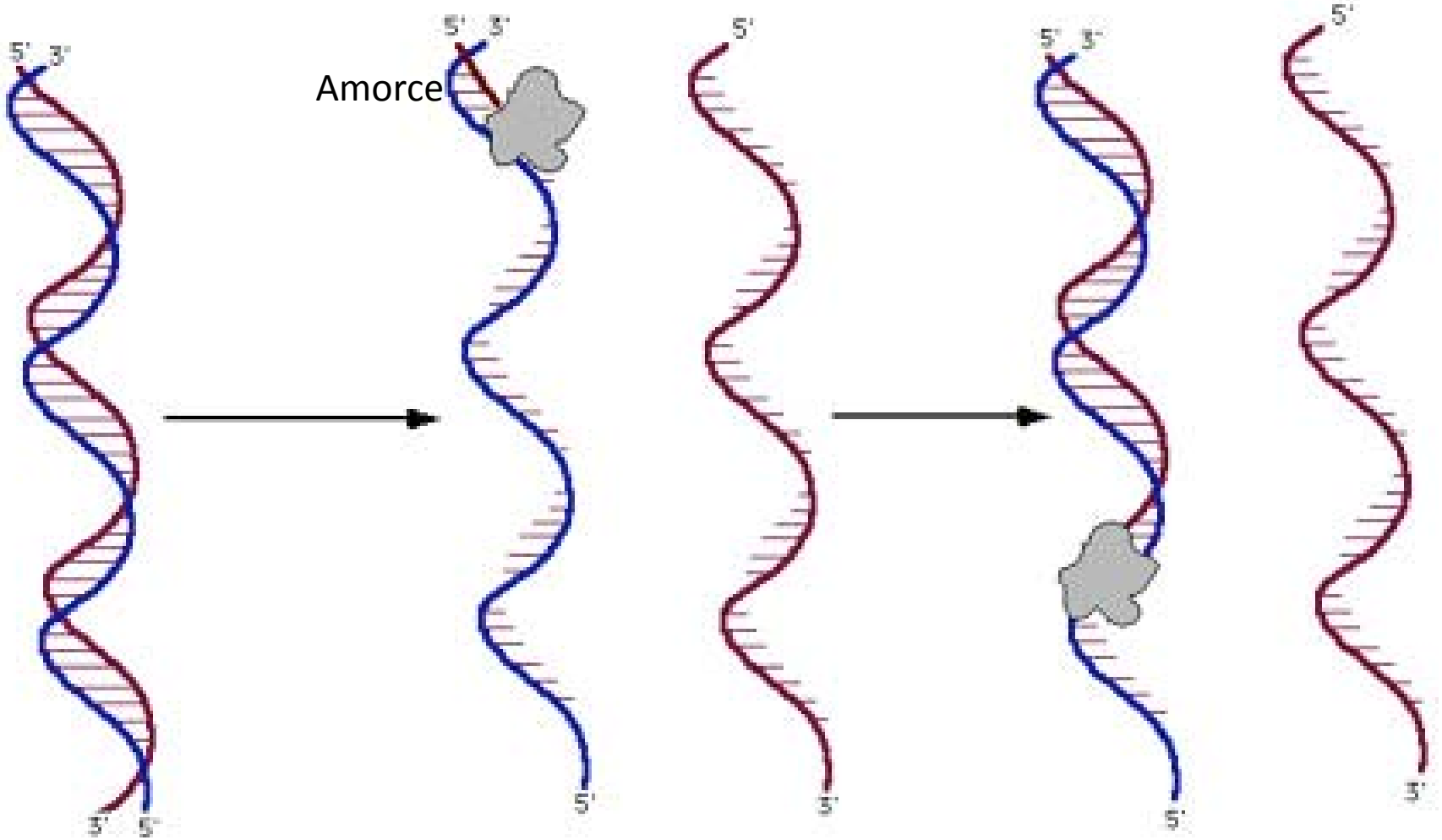
Prix Nobel de chimie
en 1958
et en 1980

Outil N°1: l'ADN polymérase de E coli (fragment de Klenow)

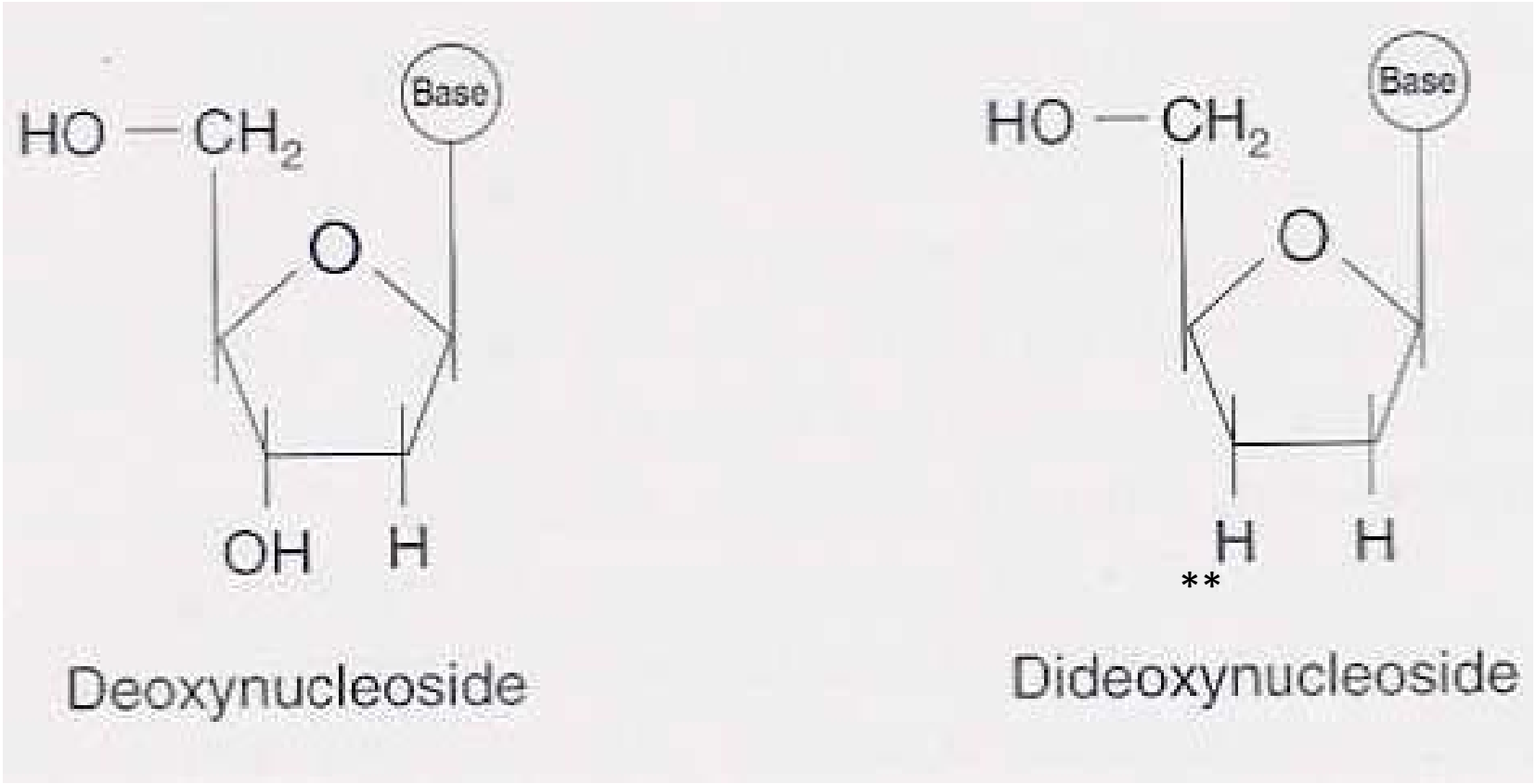




Outil n°2: la synthèse d'oligonucléotides et l'hybridation moléculaire

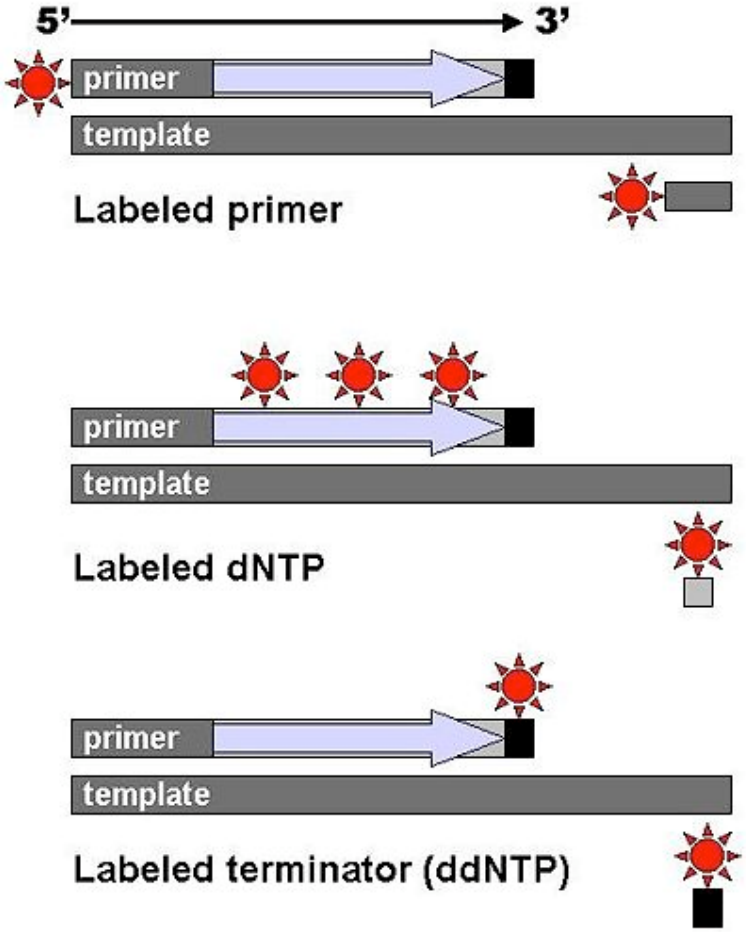


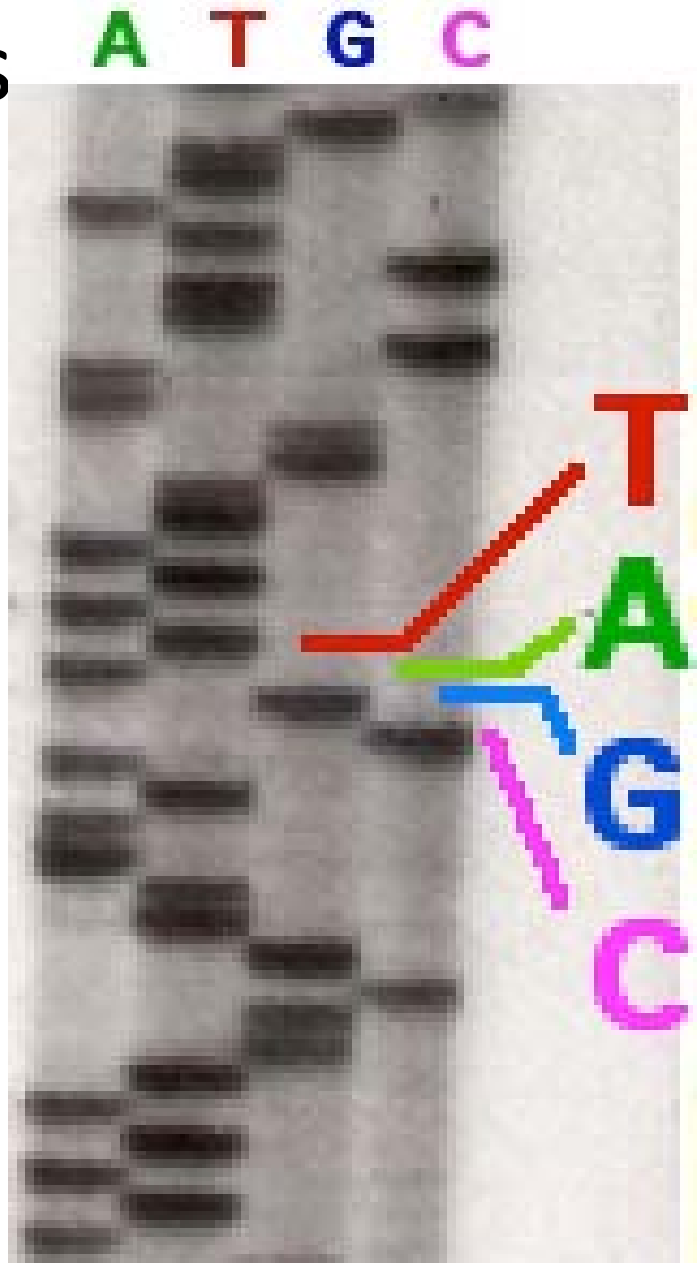
Outil N°3: les dideoxynucléotides empoisonneurs de polymérases

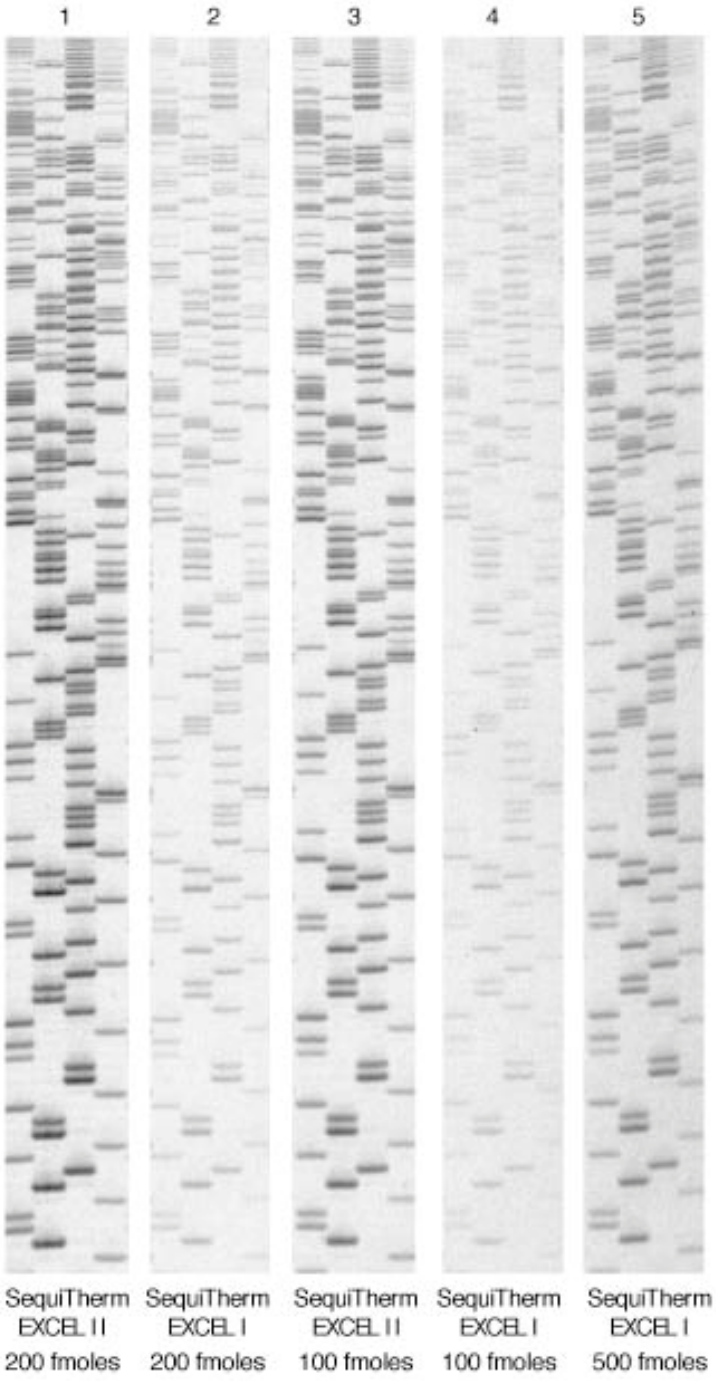


Et l’electrophorèse sur polyacrylamide à haute résolution

La méthode de séquencage d'ADN de Sanger en trois versions







2 jours de travail= 1 gel: 24 puits: 6x 4 réactions de séquencage

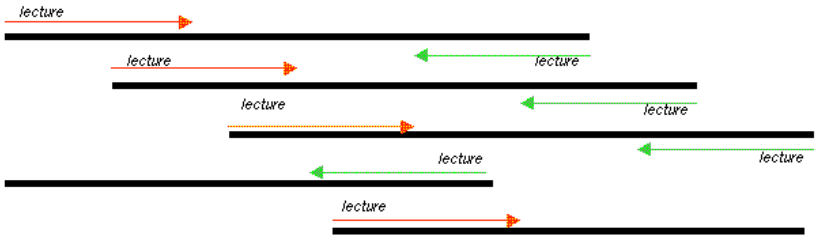
6x500 nuc= 3 kb

Nécessité de séquencer les deux brins, et de corriger les erreurs:

<0.5 kb/jour

1977 premier génome complet par le laboratoire de F Sanger:
5375 nucléotides d'ADN double brin.

Assemblage manuel des séquences.



1985-1990: AppliedBiosystems développe les premiers séquenceurs automatiques

1985-1990: AppliedBiosystems développe les premiers séquenceurs automatiques

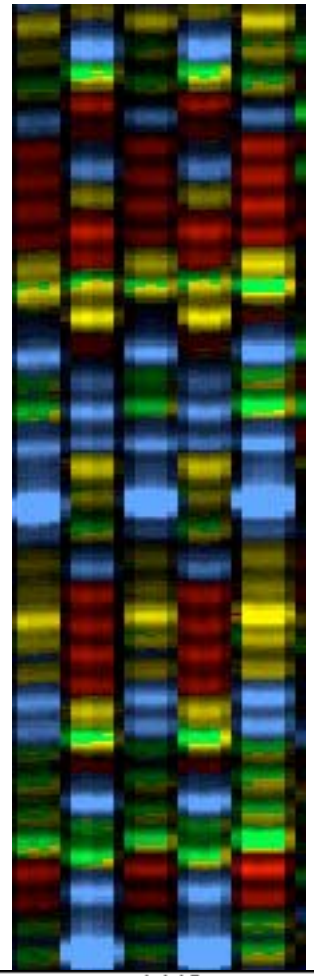
Les clés du succès:

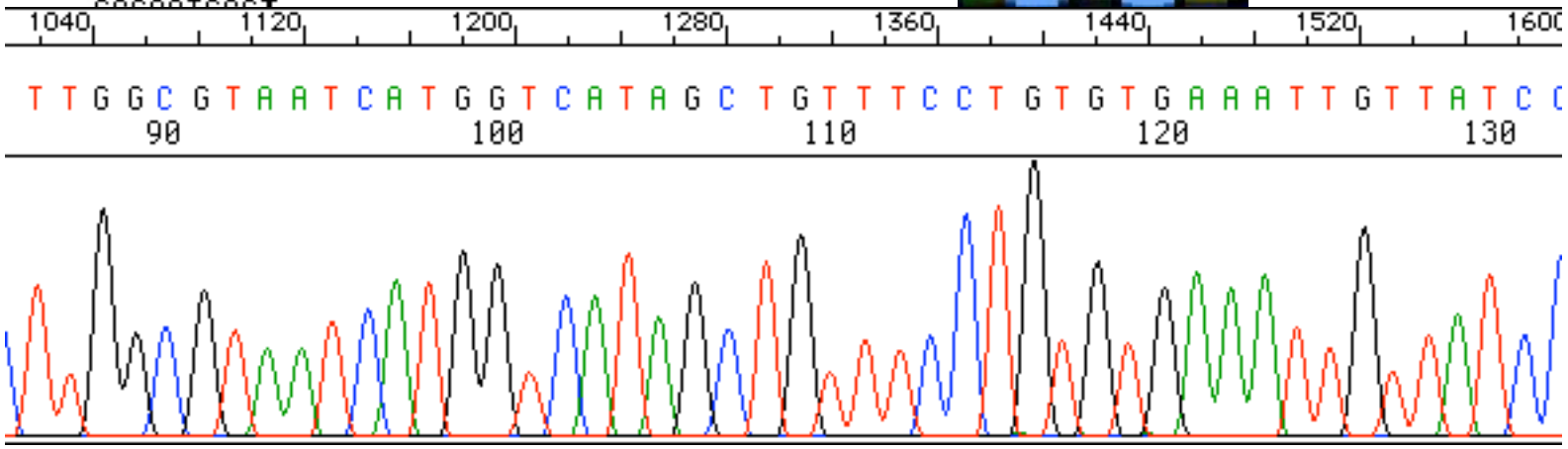
- Utilisation de 4 ddNTP marqués avec 4 fluorophores différents dans un même tube de réaction.
- La lecture laser du gel remplace l'autoradiographie
- L'ADN polymérase Taq thermostable remplace la polymérase de Klenow et permet d'effectuer 25 « cycles » de séquencage pour amplifier le signal
- Possibilité de charger 96 colonnes capillaires
- Lecture possible de 500 à 1000 nucléotides
- Une personne avec une machine produit 96x500 bp 3 fois par jour: 150 Kb/j (70 kb d'ADN double brin)

Gel:

GCGAATGCGTCCACAACGCTACAGGTG
GCGAATGCGTCCACAACGCTACAGGT
GCGAATGCGTCCACAACGCTACAGG
GCGAATGCGTCCACAACGCTACAG
GCGAATGCGTCCACAACGCTACA
GCGAATGCGTCCACAACGCTAC
GCGAATGCGTCCACAACGCTA
GCGAATGCGTCCACAACGCT
GCGAATGCGTCCACAACGC
GCGAATGCGTCCACAACG
GCGAATGCGTCCACAAC
GCGAATGCGTCCACAA
GCGAATGCGTCCACA
GCGAATGCGTCCAC
GCGAATGCGTCCA
GCGAATGCGTCC
GCGAATGCGTC
GCGAATGCGT

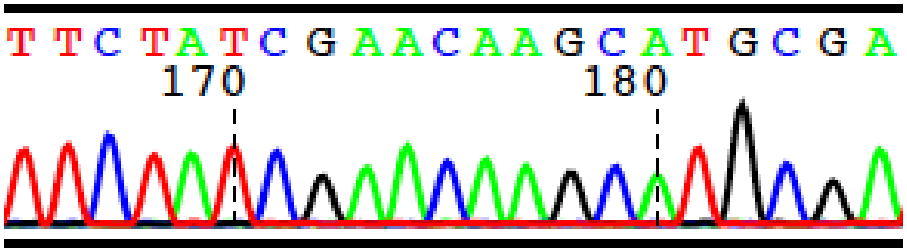
1€/nucléotide



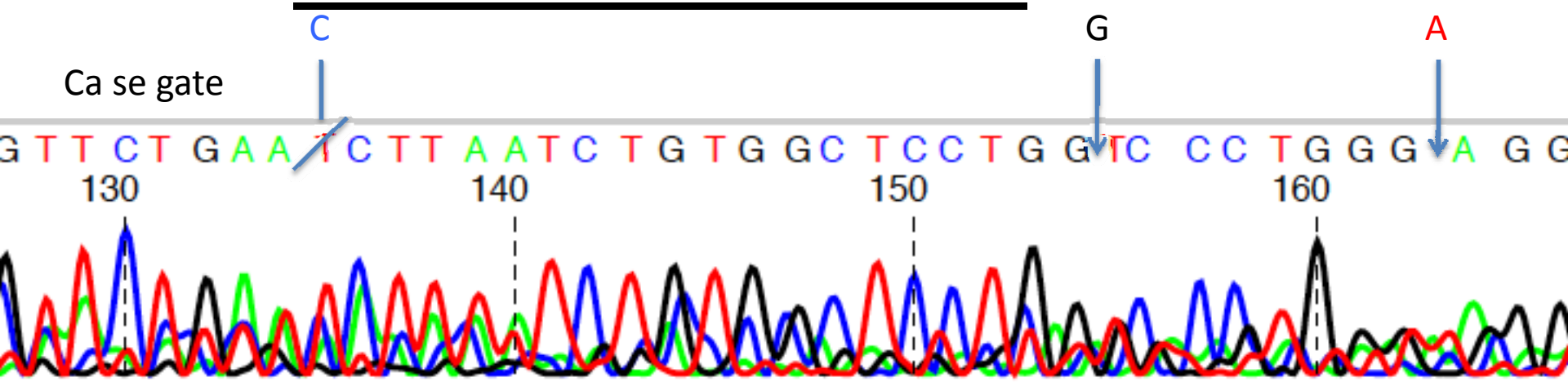


Les limites:
- La pureté des ADN
-La résolution de l'electrophorèse
- la « processivité » de la polymérase
- les séquences à problèmes

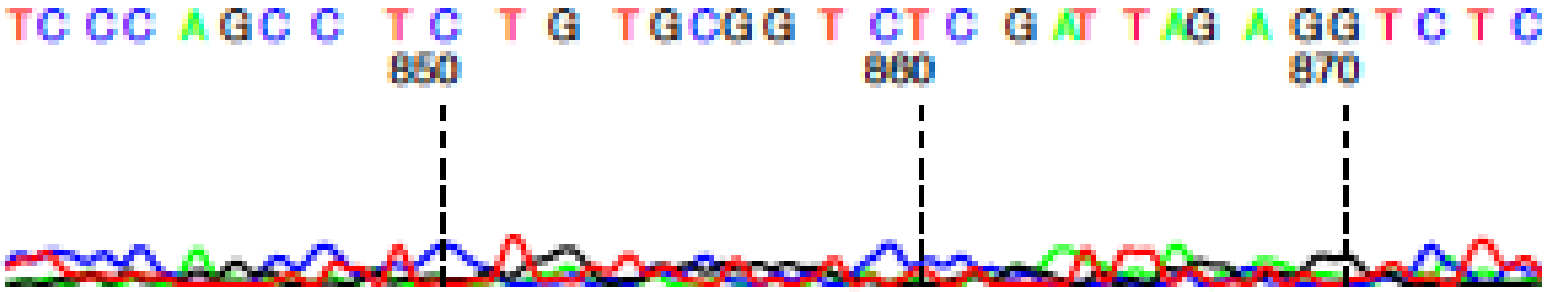
Parfait



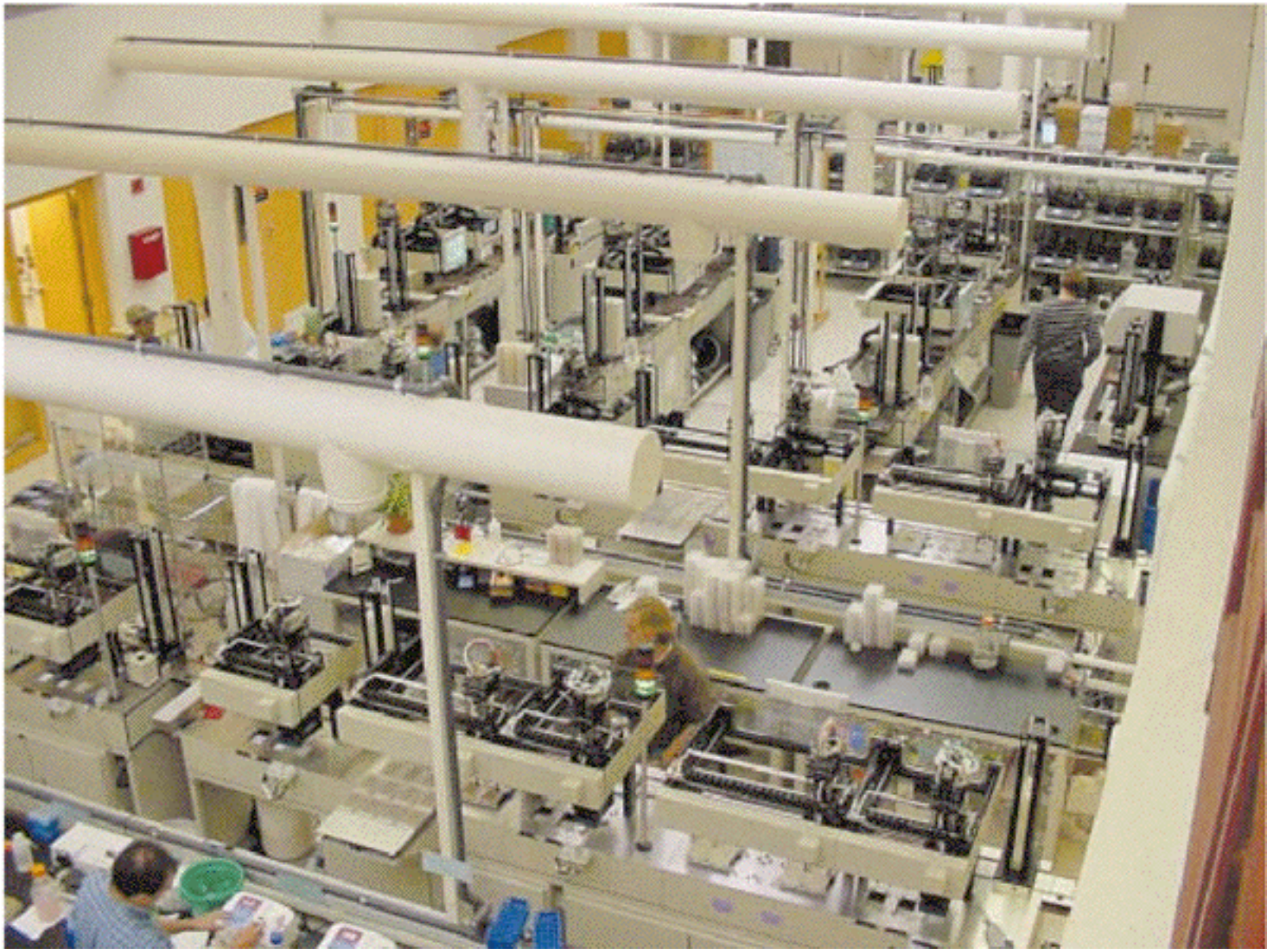
Ca se gate



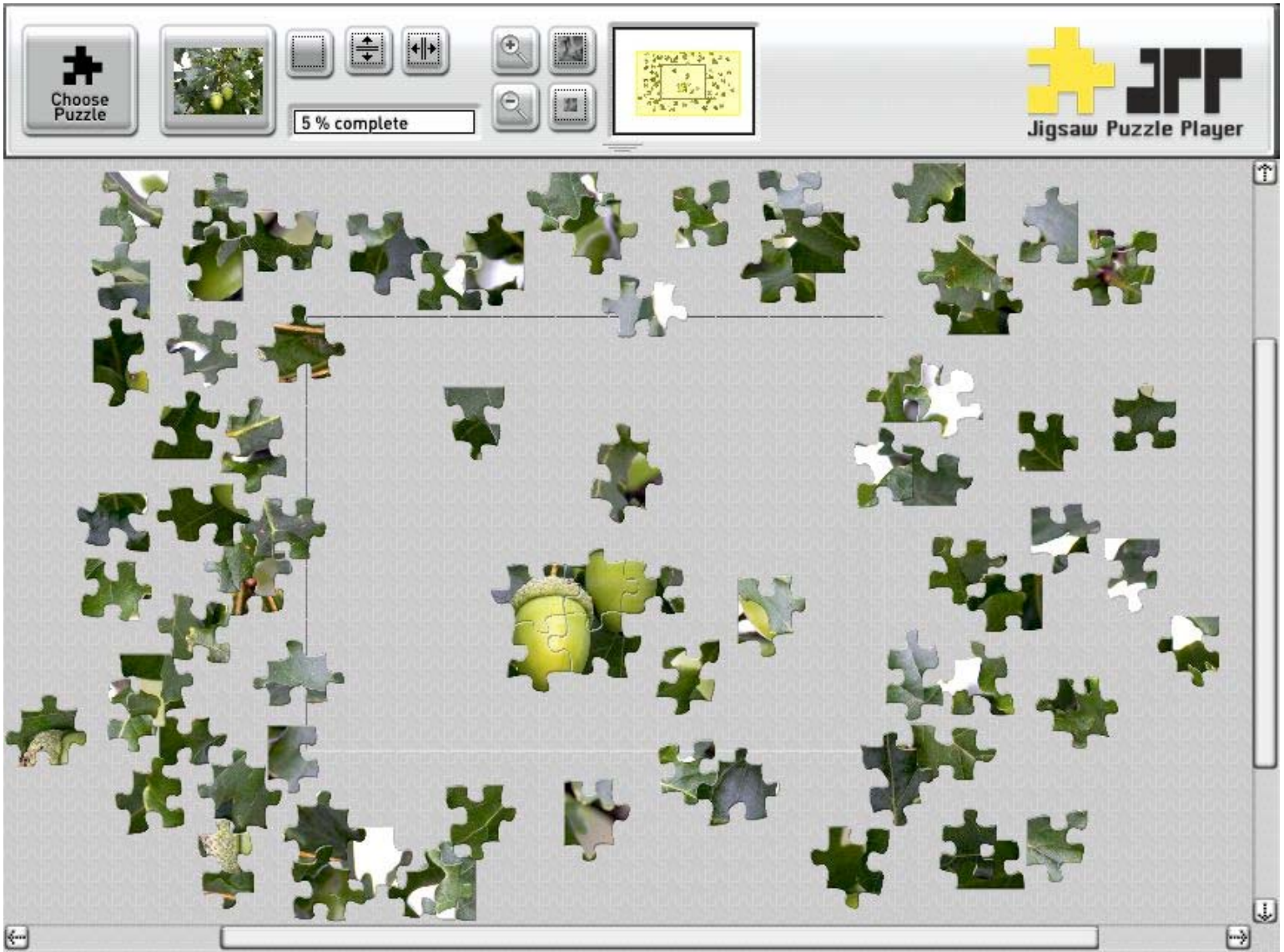
A la poubelle



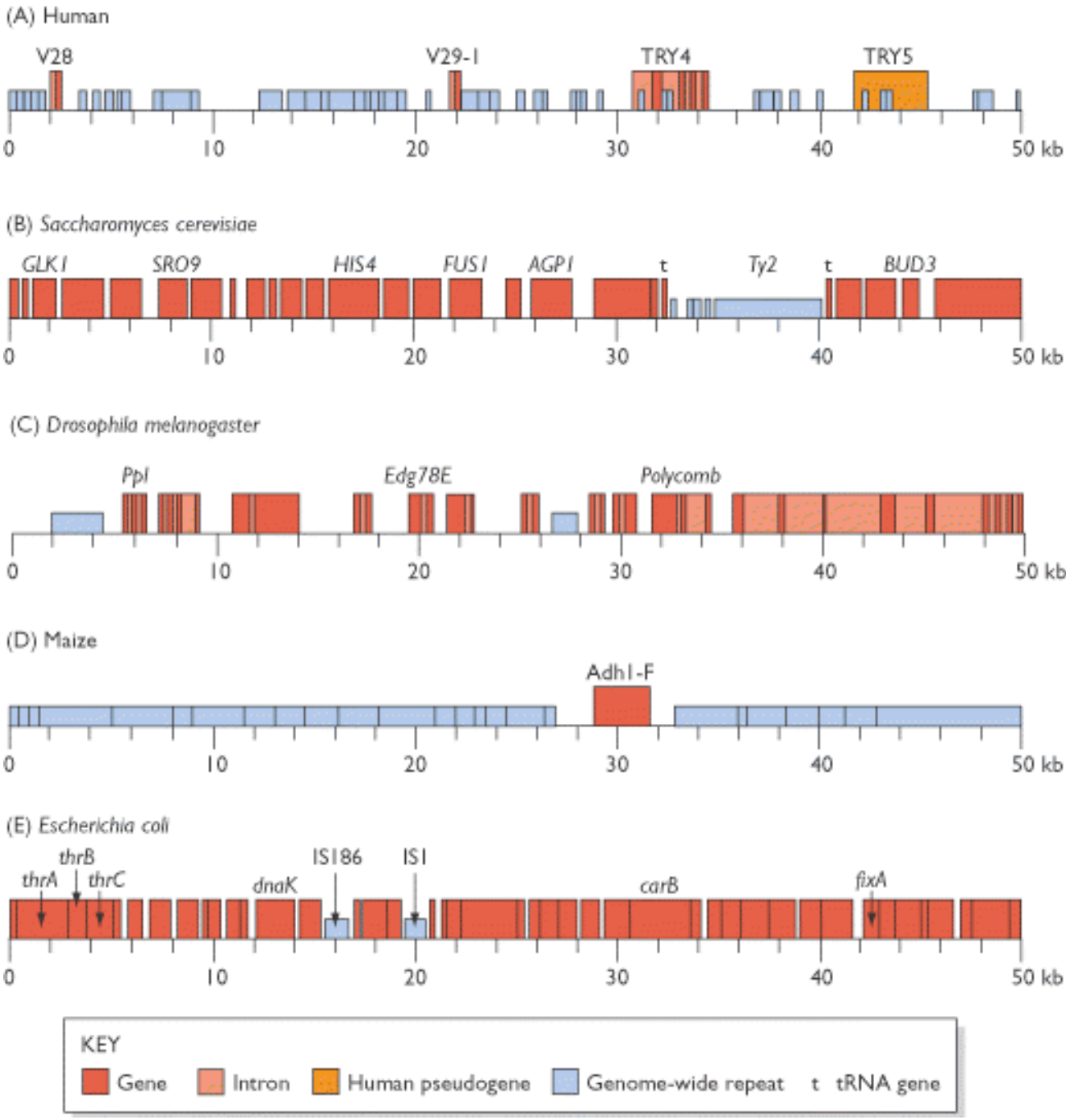
1990: début du projet du génome humain.



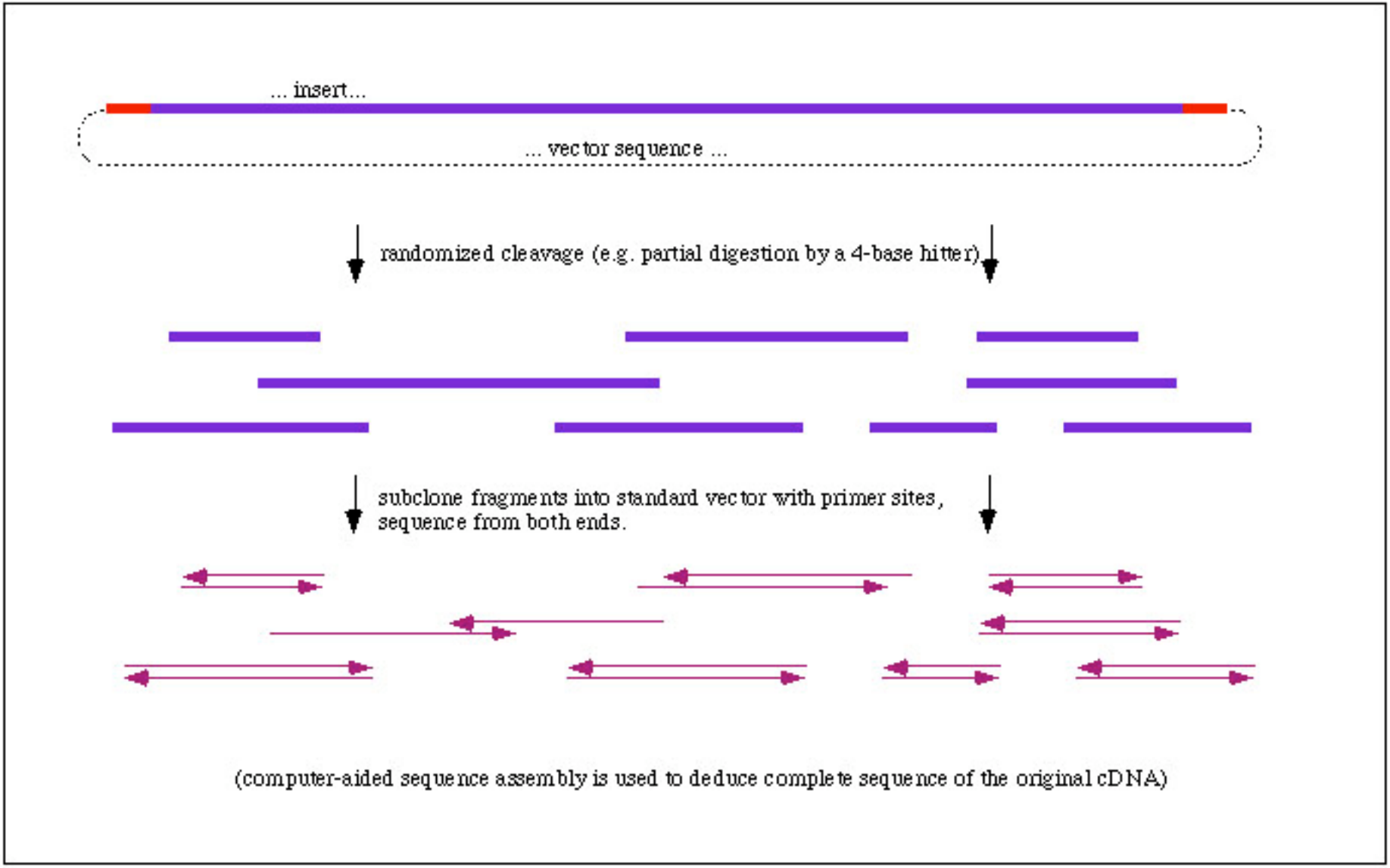
Comment résoudre le probleme de l’assemblage?
La réponse de Craig Venter: »Shotgunsequencing »
= En confiant tout le travail à des ordinateurs

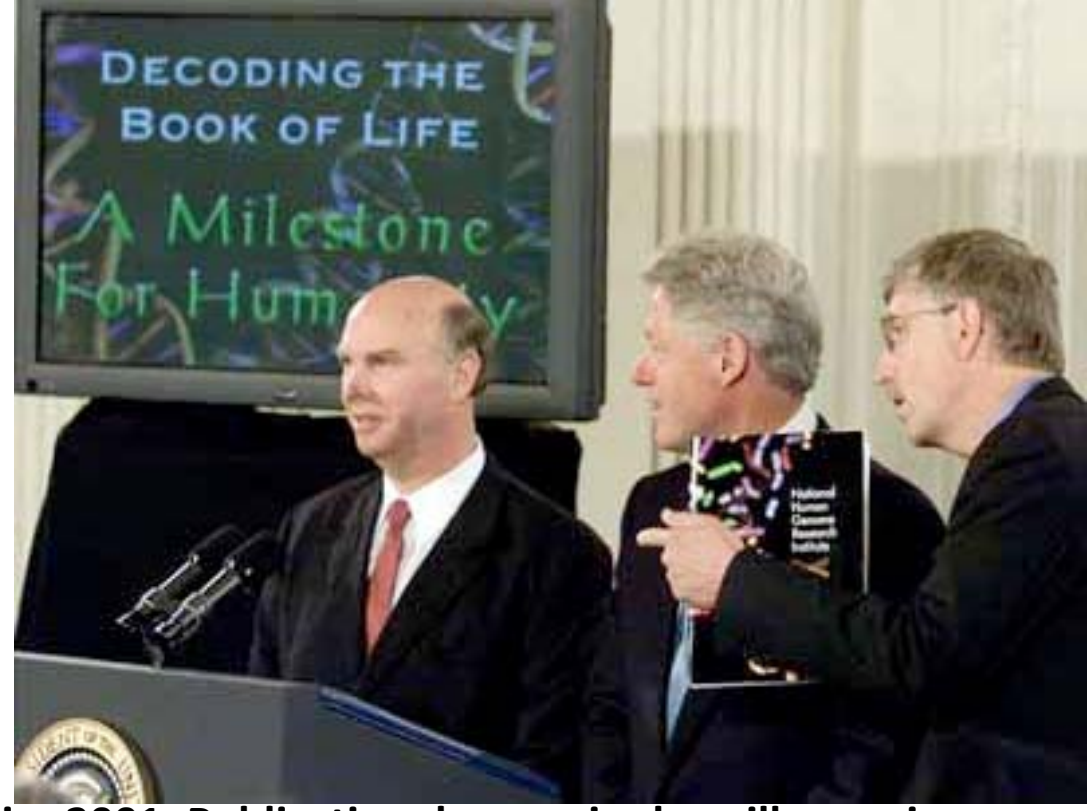


Le problème des séquences répétées



Comment résoudre le probleme de l'assemblage?
L a réponse du NIH « Paired-endsequencing » de BAC.

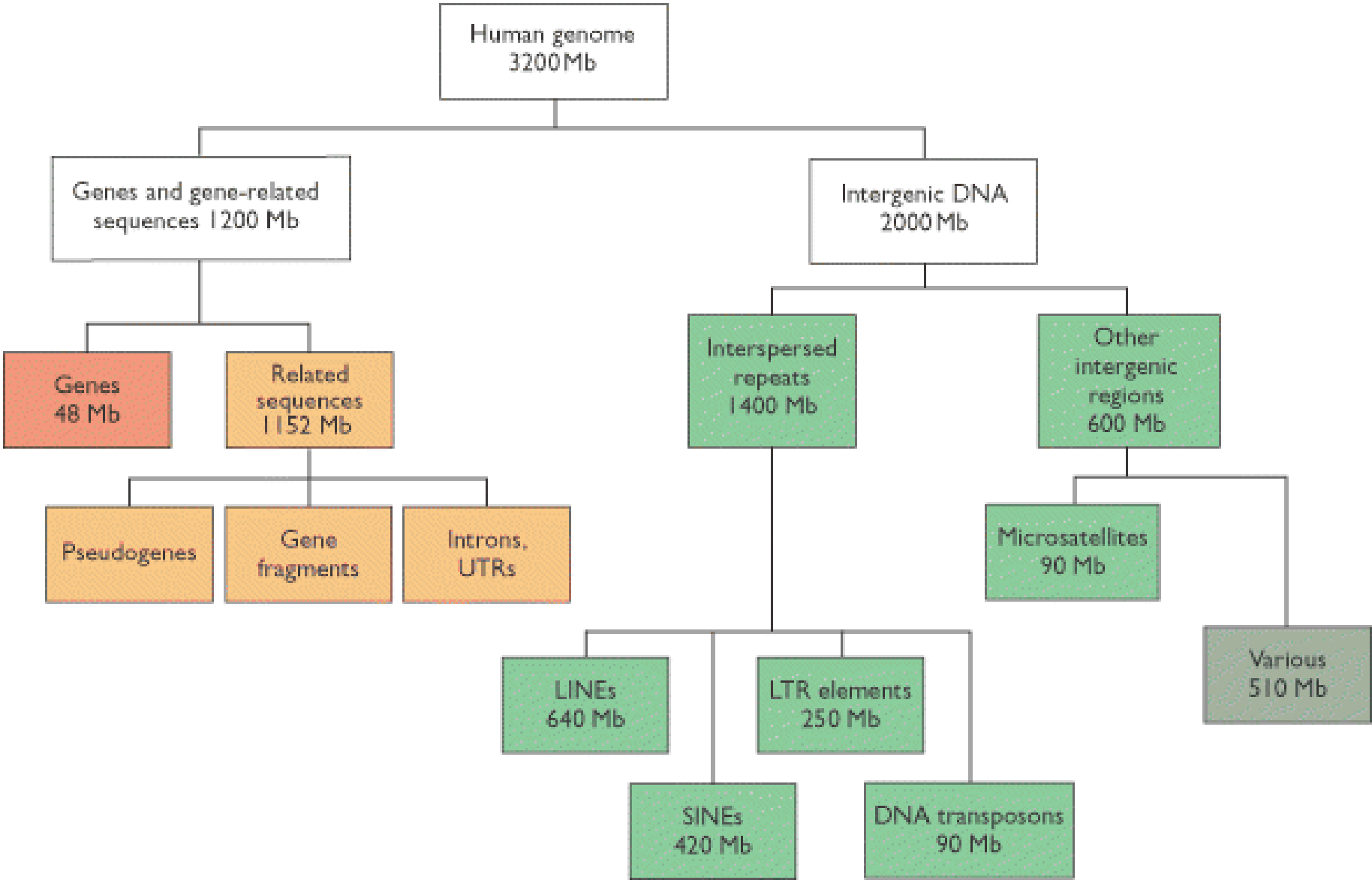




15 Février 2001: Publication du premier brouillon, qui aura couté 300 000 dollars à Celera et 3 milliards de dollars aux contribuables qui financent le NIH (70% du génome couvert, en moyenne chaque région est couverte 7 fois)

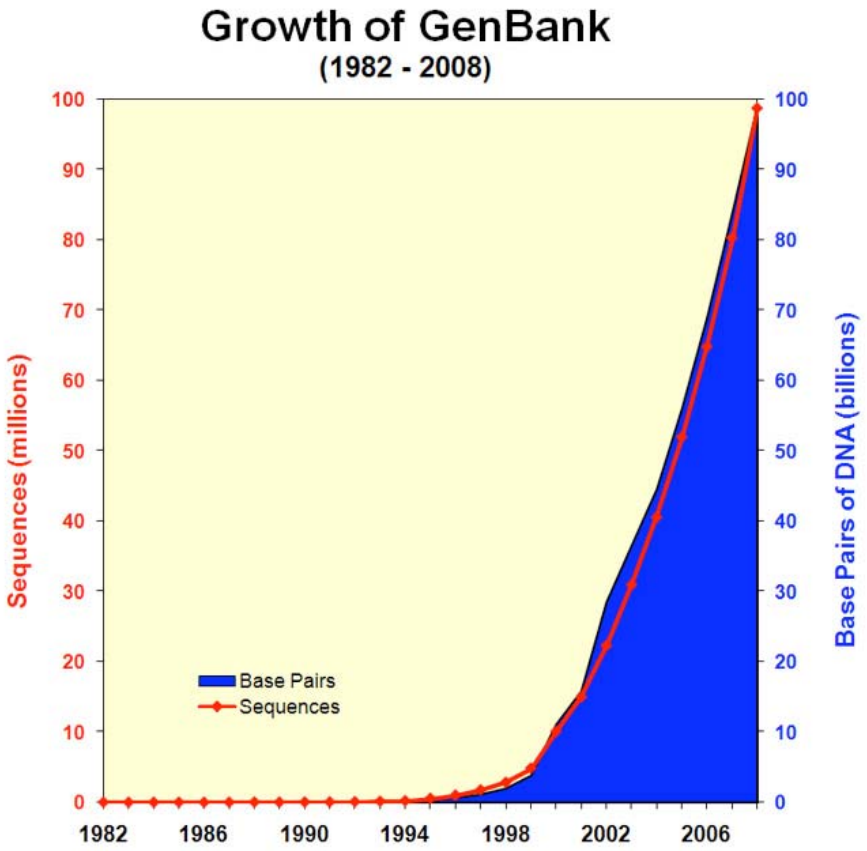
2006: fin officielle du programme:

Aujourd'hui:
-3,272,480,989 nucléotides connus en 2010 (98% génome)
- Taux d'erreur $<10^{-4}$: soit encore plus de 30000 erreurs….
-22,258 genes (?)
-183 328 782 « TRACES » dans les banques de données publiques


Human genome
3200 Mb
Genes and gene-related sequences 1200 Mb
Intergenic DNA
2000 Mb
Genes
48 Mb
Related sequences
1152 Mb
Interspersed repeats
1400 Mb
Other intergenic regions
600 Mb
Pseudogenes
Gene fragments
Introns, UTRs
Microsatellites
90 Mb
LINEs
640 Mb
LTR elements
250 Mb
Various
510 Mb
SINEs
420 Mb
DNA transposons
90 Mb

Des banques de données publiques en croissance permanente

- http://www.ncbi.nlm.nih.gov/Genbank/index.html
- http://www.ensembl.org/index.html
- 190 milliards de nucléotides (il faudrait 1(millions de pages pour imprimer l'ensemble, soit 10 kilomètres de rayonna dans une bibliothèque)



nature biotechnology

LETTERS

Single-molecule sequencing of an individual human genome

Dmitry Pushkarev[1,2], Norma F Neff[1,2] & Stephen R Quake[1]

Recent advances in high-throughput DNA sequencing technologies have enabled order-of-magnitude improvements in both cost and throughput. Here we report the use of single-molecule methods to sequence an individual human genome. We aligned billions of 24- to 70-bp reads (32 bp average) to ~90% of the National Center for Biotechnology Information (NCBI) reference genome, with 28× average coverage. Our results were obtained on one sequencing instrument by a single operator with four data collection runs. Single-molecule sequencing enabled analysis of human genomic information without the need for cloning, amplification or ligation. We determined ~2.8 million single nucleotide polymorphisms (SNPs) with a false-positive rate of less than 1% as validated by Sanger sequencing and 99.8% concordance with SNP genotyping arrays. We identified 752 regions of copy number variation by analyzing coverage depth alone and validated 27 of these using digital PCR. This milestone should allow widespread application of genome sequencing to many aspects of genetics and human health, including personal genomics.

on a surface can be extended asynchronously, thereby allowing substantial flexibility in the kinetics of sequencing chemistry. Previous reports of single-molecule sequencing have been proofs of principle[11–13], and their sequencing throughput has not been competitive with alternative approaches. Generally, read lengths have been relatively short and error rates have been dominated by deletions; it has not been clear whether the resulting sequence quality is suitable for human genome sequencing applications.

The Heliscope Single Molecule Sequencer (Helicos Biosciences) is the first commercial release of a single-molecule sequencing instrument. It allows one to follow ~1 billion individual molecules as they are sequenced over the course of a week—a throughput that is practical for human genome sequencing. There have been several technical improvements to the platform since the reported sequencing of a viral genome[12], including more than a 1,000-fold improvement in parallelism, a new generation of sequencing reagents that allows digital measurement of homopolymer sequences, and a new software algorithm, IndexDP, for performing alignments to the entire human genome.

We used two of the instrument's 50 flow-cell channels to resequence the *Staphylococcus aureus* genome as a calibration of sequencer perfor-

Le génome humain séquencé intégralement à nouveau en 2009:
Taux d'erreur 0.2% couverture moyenne x28
mais:
Avec une seule machine, par deux employés, en un mois,
et pour 50 000 dollars au lieu de 3 milliards

- « Nextgenerationsequencing »

= le séquencage d’aujourd’hui

454 Roche

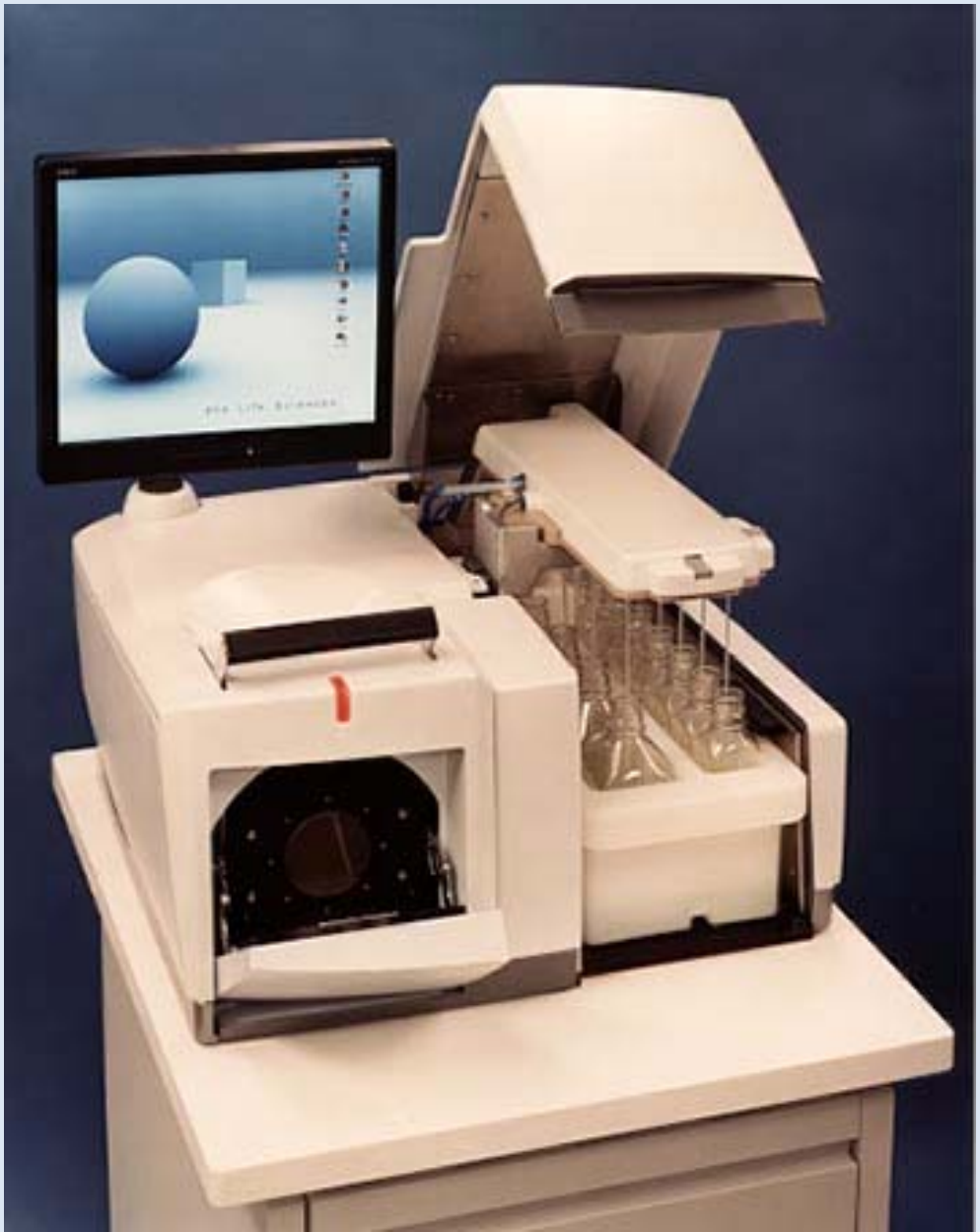
Solexa Illumina

SOLID AppliedBiosystems

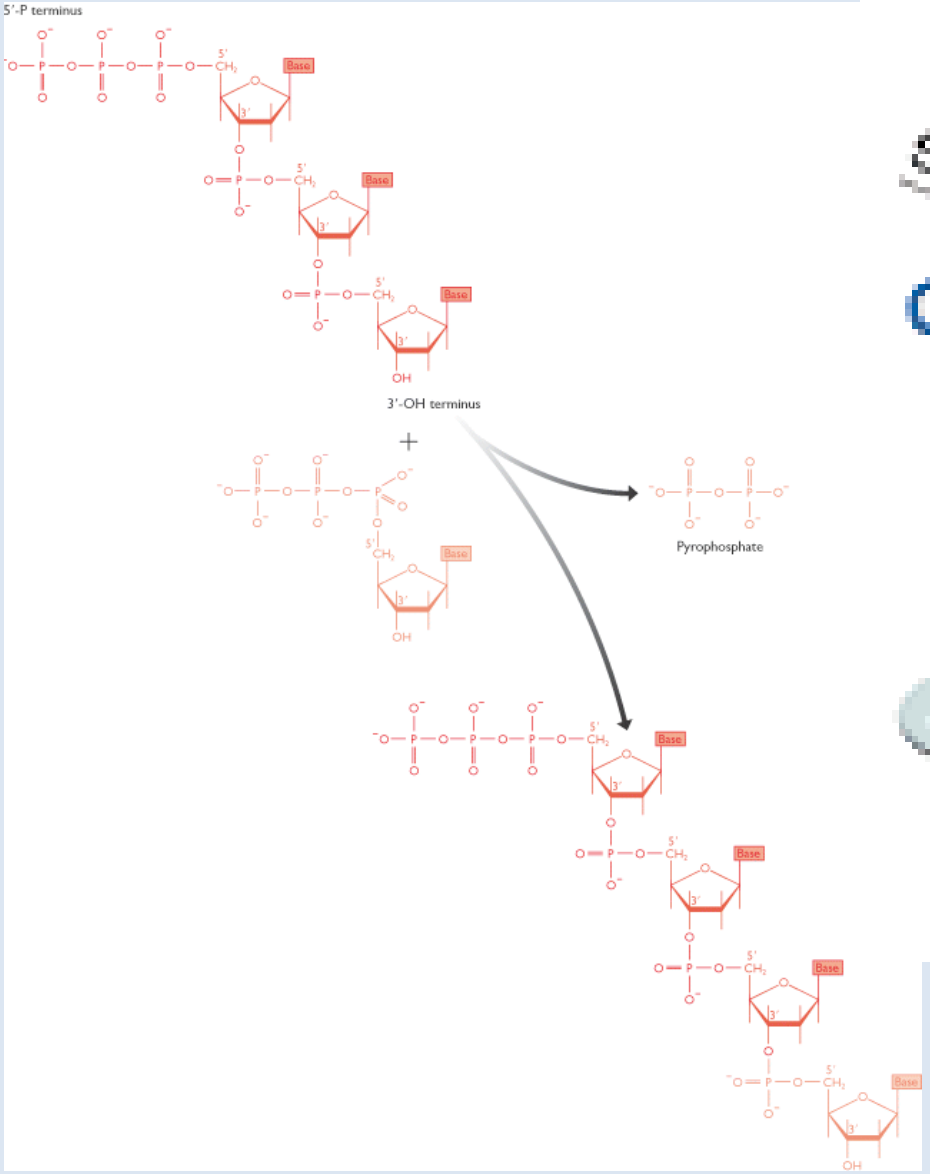
Helicos

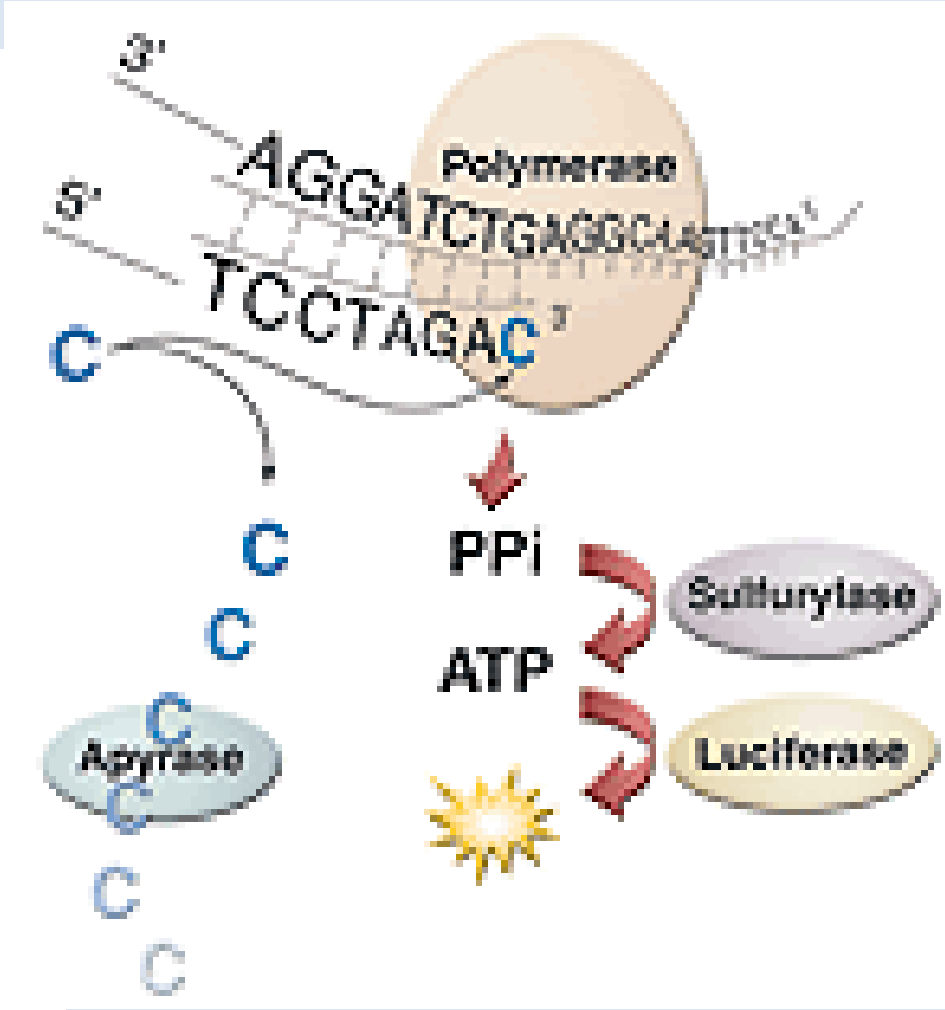
Pacific Biosciences

Pyroséquencageà haut débit: le séquenceur 454 de chez Roche



Principe du pyroséquencage





Emulsion Based Clonal Amplification

Adapter carrying library DNA

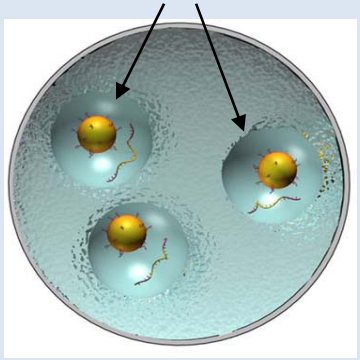
Mix DNA Library & capture beads (limited dilution)

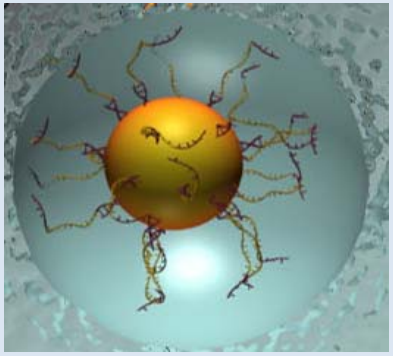
+ PCR Reagents

+ Emulsion Oil

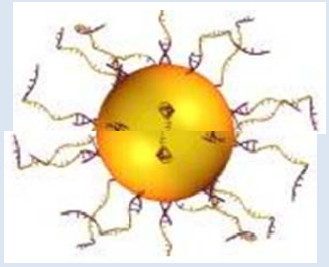
Create "Water-in-oil" emulsion

Micro-reactors





Perform emulsion PCR

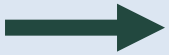


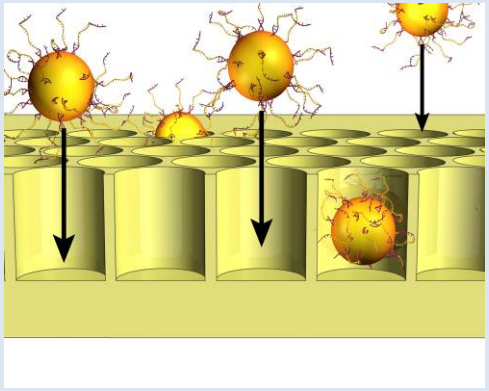
"Break micro-reactors" Isolate DNA containing beads

- Generation of millions of clonally amplified sequencing templates on each bead
- No cloning and colony picking

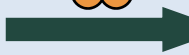
Depositing DNA Beads into the PicoTiter™Plate

Load beads into PicoTiter™Plate

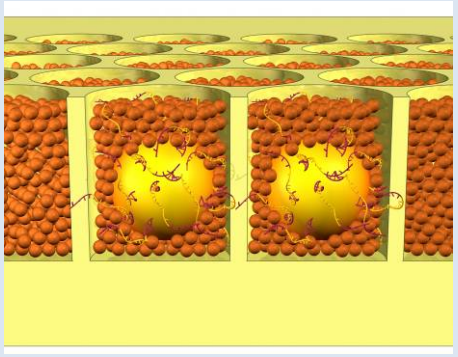


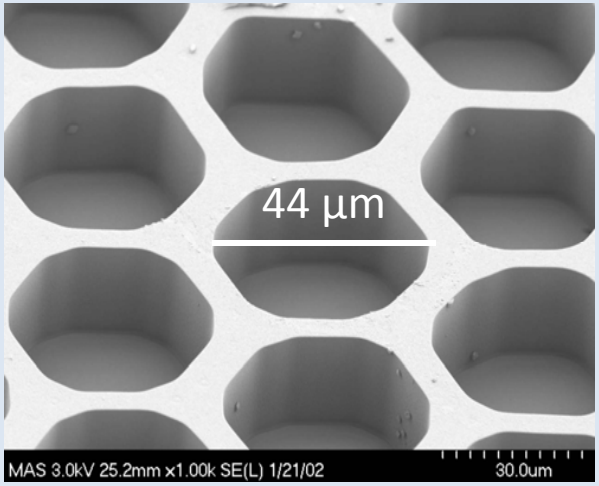


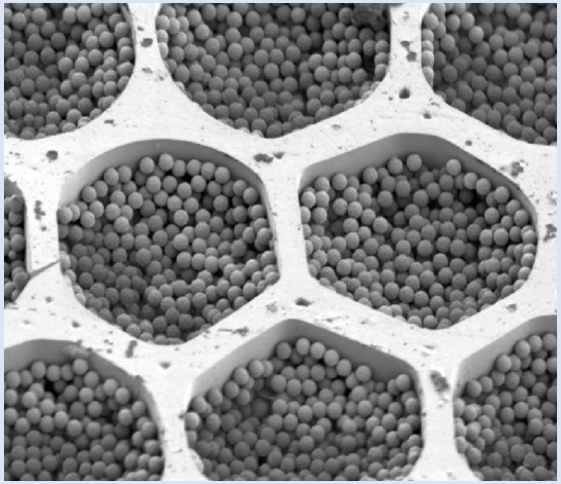
Load Enzyme Beads



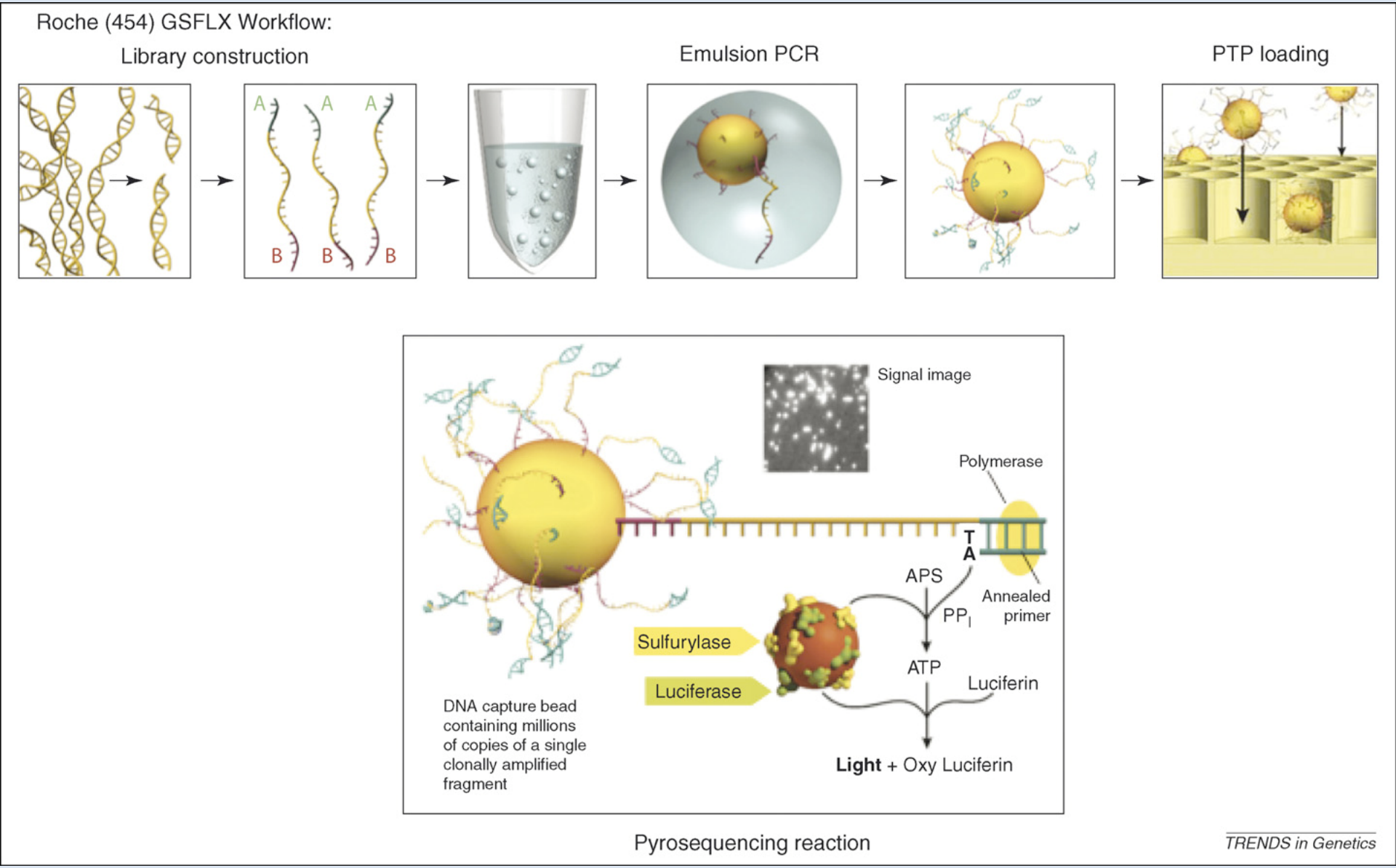
Centrifuge Step



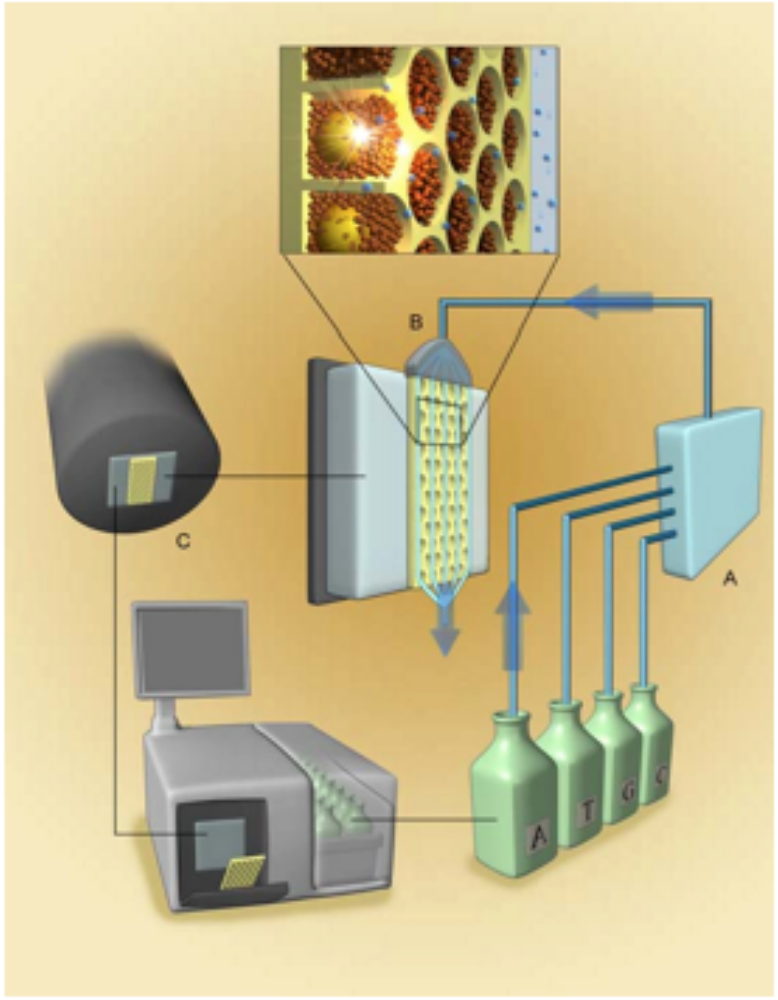


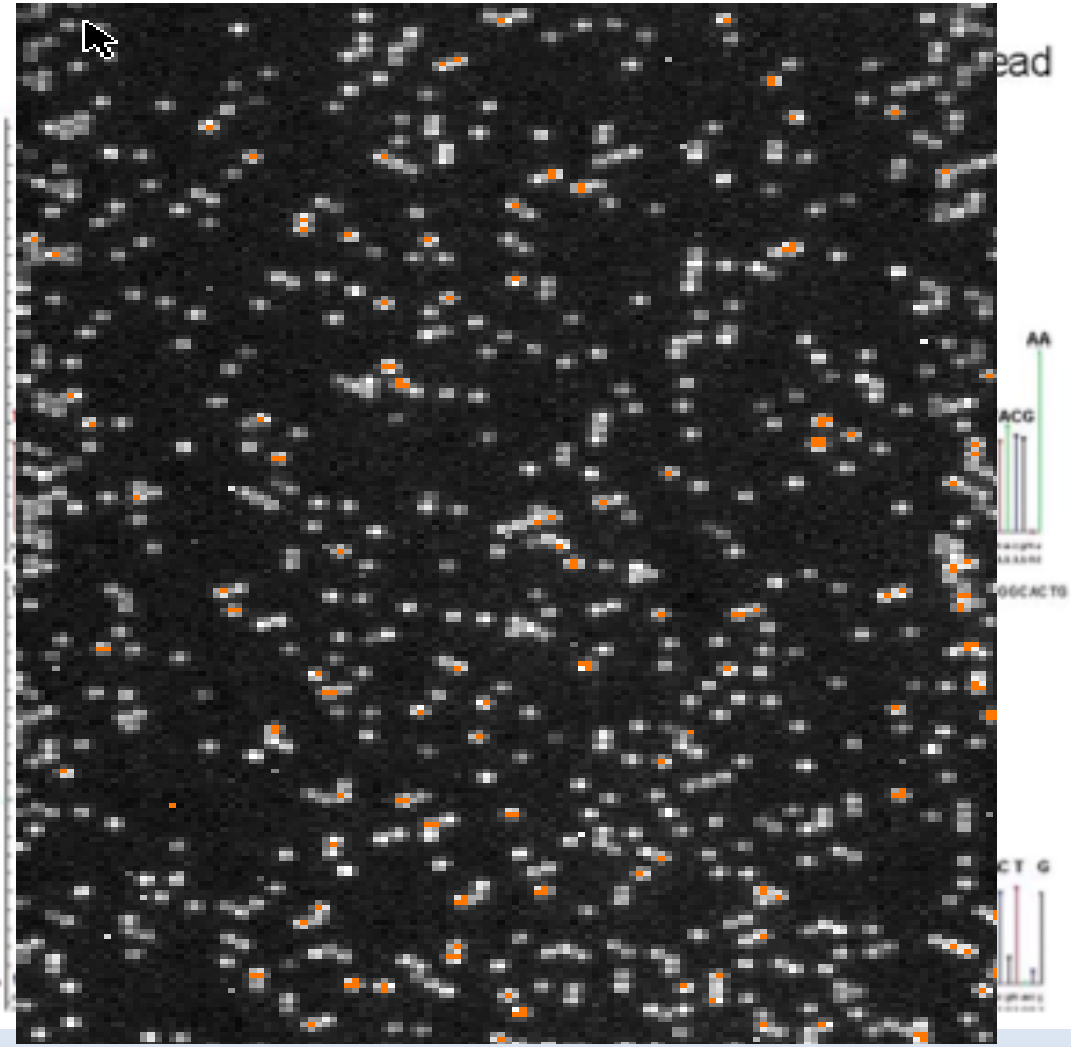


Pyroséquencage 454 Roche

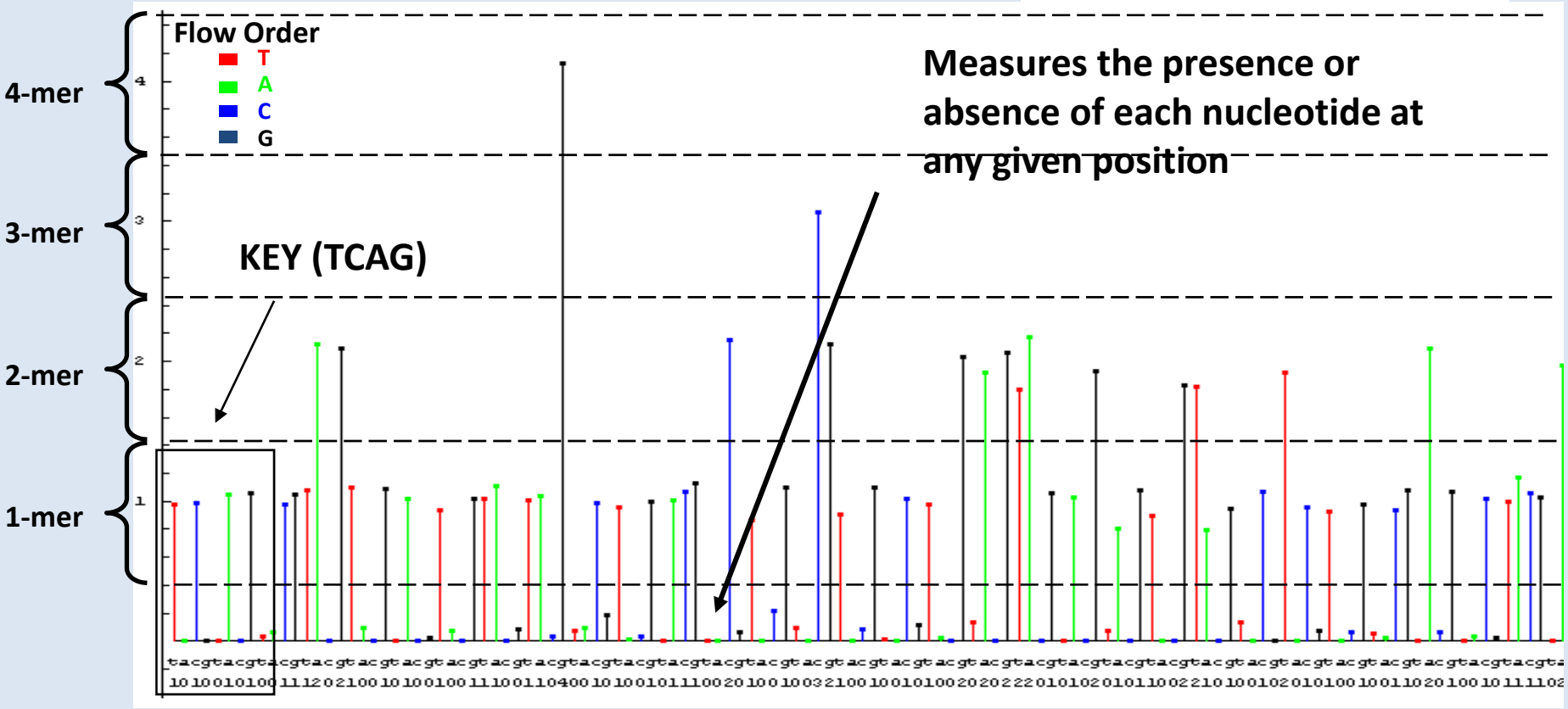


454 Technology - Sequencing Instrument





454 Sequencing: BaseCalling

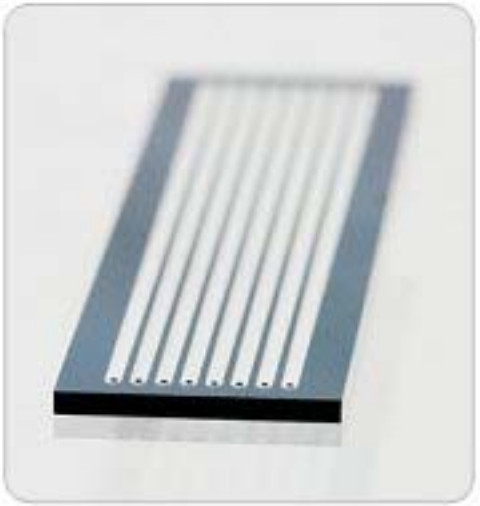


GS Evolution

GS Evolution	Genome Sequencer 20	Genome Sequencer FLX	Genome Sequencer FLX 2008
Read length	100 bases	>200 bases	> 400 bp
# of clonal reads	>200,000	> 400,000	> 2,000,000
System throughput /8hr shift	20-30mb	100mb	1gb
Cost	$6,000-9,000.00	$3,000-9,000	$10,000
Accuracy	99.99%	99.99%	99.99%

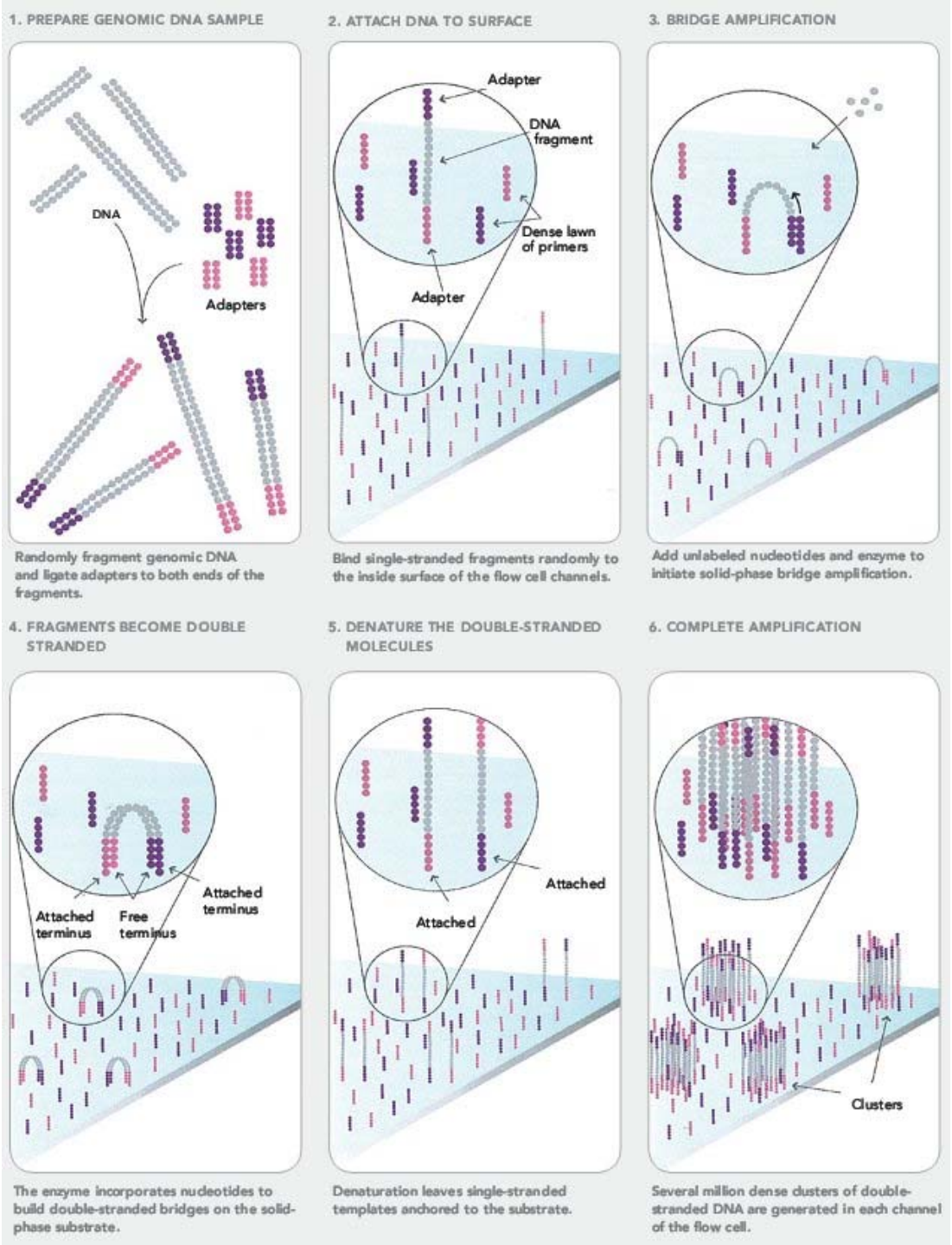
Méthode Solexa Illumina

FIGURE 1: ILLUMINA GENOME ANALYZER FLOW CELL



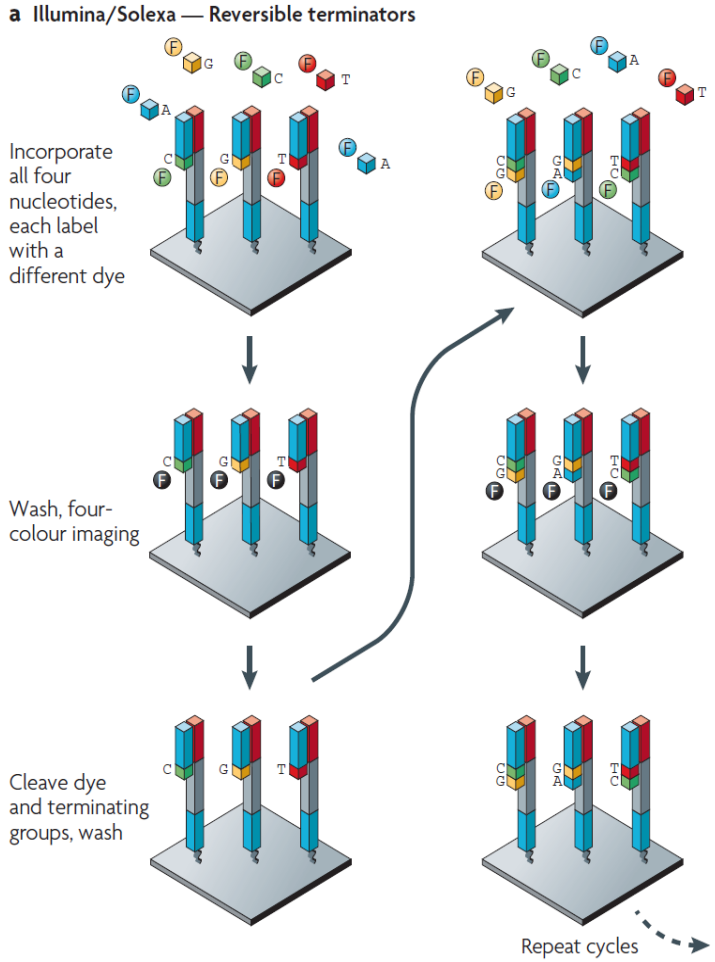
Up to eight samples can be loaded onto the flow cell for simultaneous analysis on the Illumina Genome Analyzer.

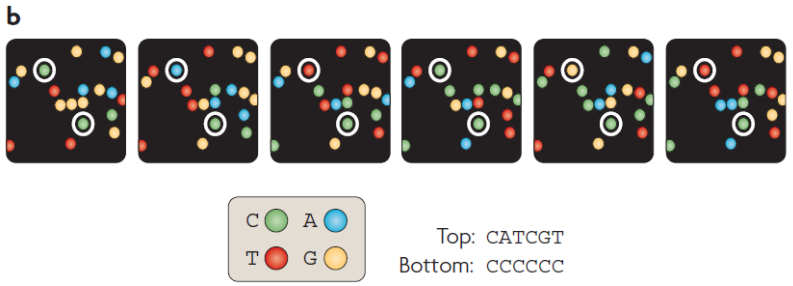
1) PCR in situ


1. PREPARE GENOMIC DNA SAMPLE
DNA
Adapters
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
2. ATTACH DNA TO SURFACE
Adapter
DNA fragment
Dense lawn of primers
Adapter
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
3. BRIDGE AMPLIFICATION
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
4. FRAGMENTS BECOME DOUBLE STRANDED
Attached terminus
Free terminus
Attached terminus
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.
5. DENATURE THE DOUBLE-STRANDED MOLECULES
Attached
Attached
Denaturation leaves single-stranded templates anchored to the substrate.
6. COMPLETE AMPLIFICATION
Clusters
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Sequencing Workflow

Sequencing-by-synthesis (Illumina)





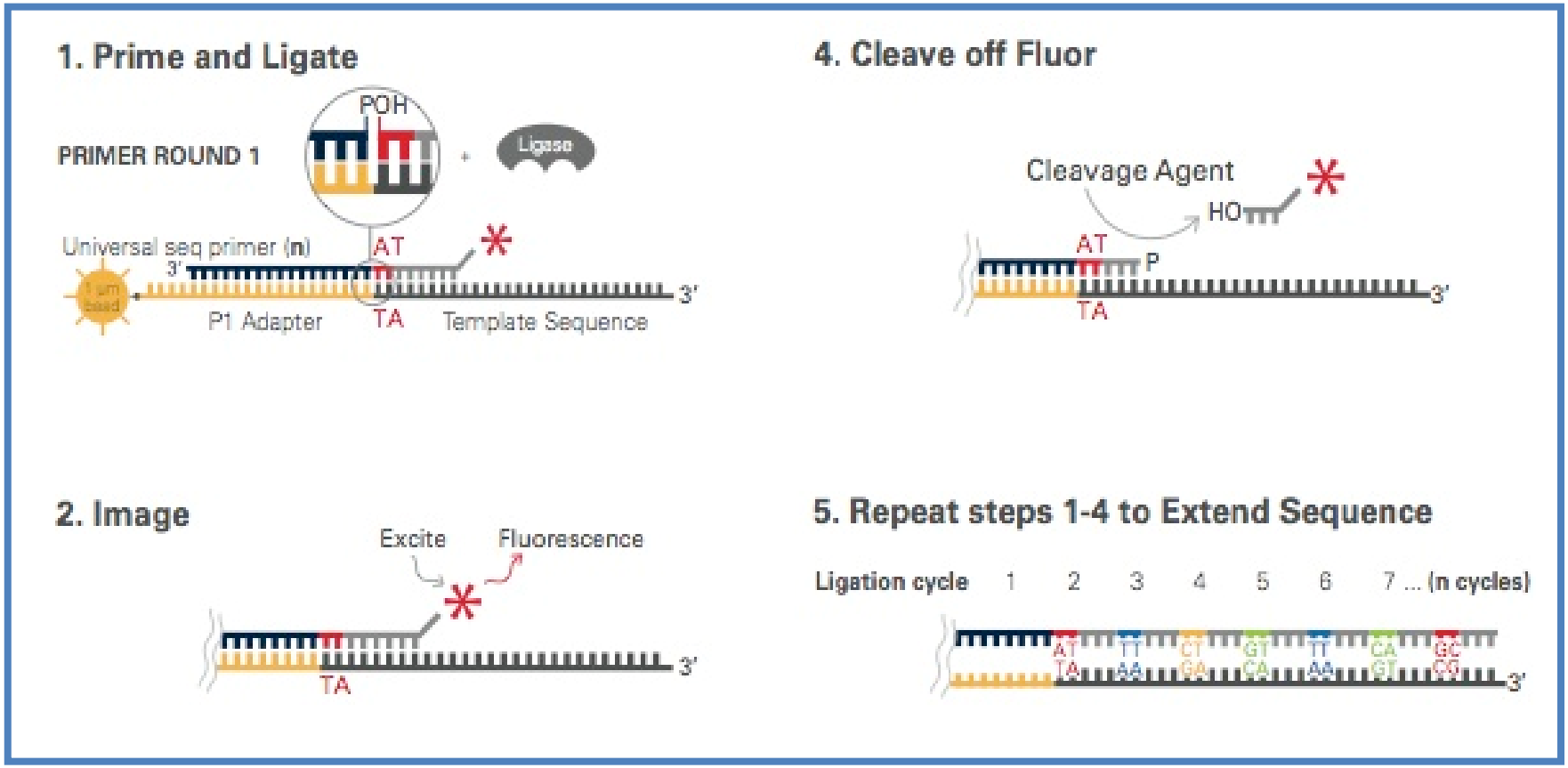
DNA library preparation and titration — 4.5 h and 10.5 h → Amplification — 8.0 h → Sequencing — 7.5 h

Read Length	Run Time	Output
1 x 35 bp	~1.5 days	26-35 Gb
2 x 50 bp	~4 days	75-100 Gb
2 x 100 bp	~8 days	150-200 Gb

SOLID AppliedBiosystems

Sequencing Workflow

Sequencing-by-ligation (SOLiD)



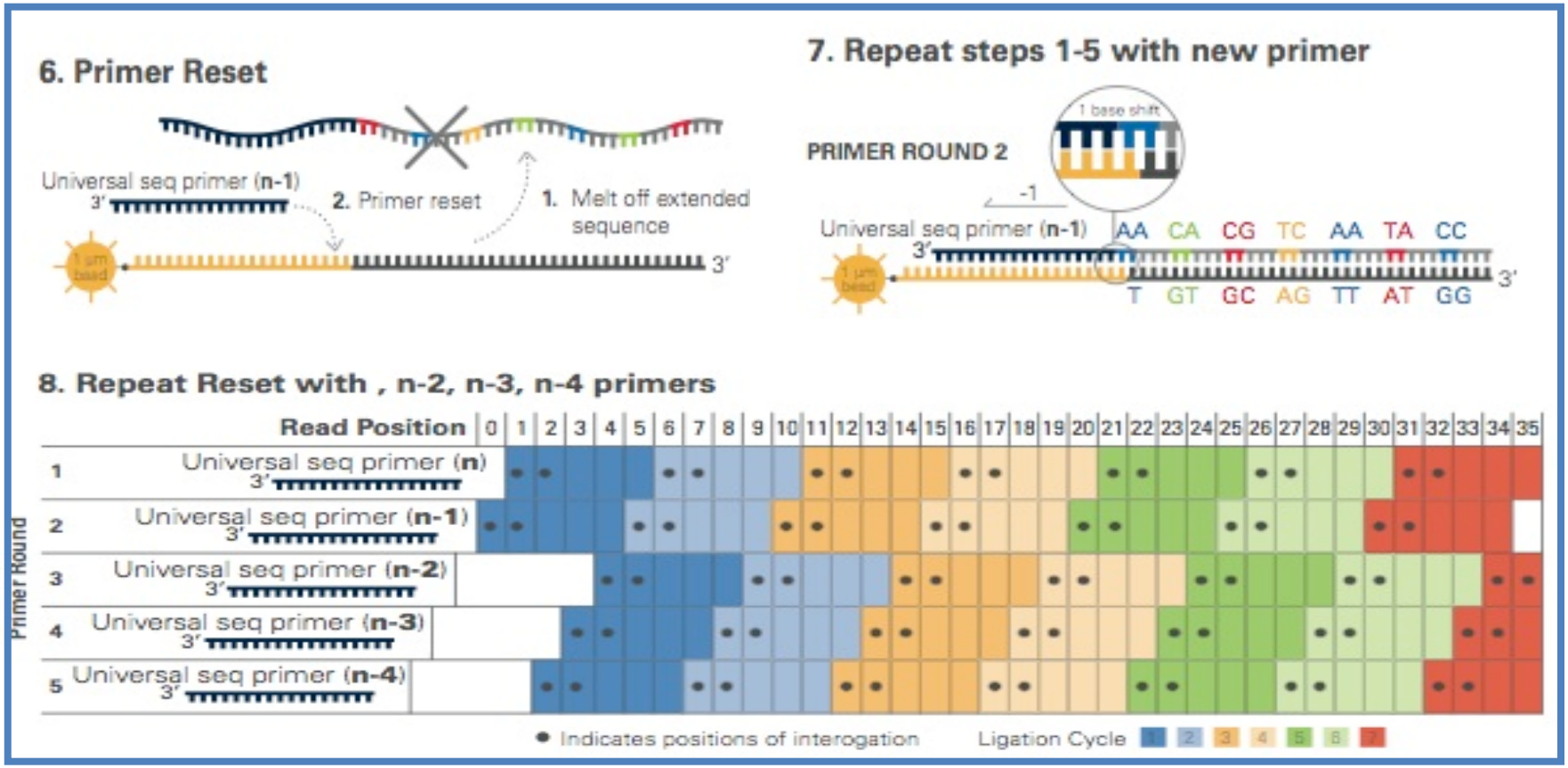
DNA library preparation and titration — 4.5 h and 10.5 h

Amplification — 8.0 h

Sequencing — 7.5 h

Sequencing Workflow

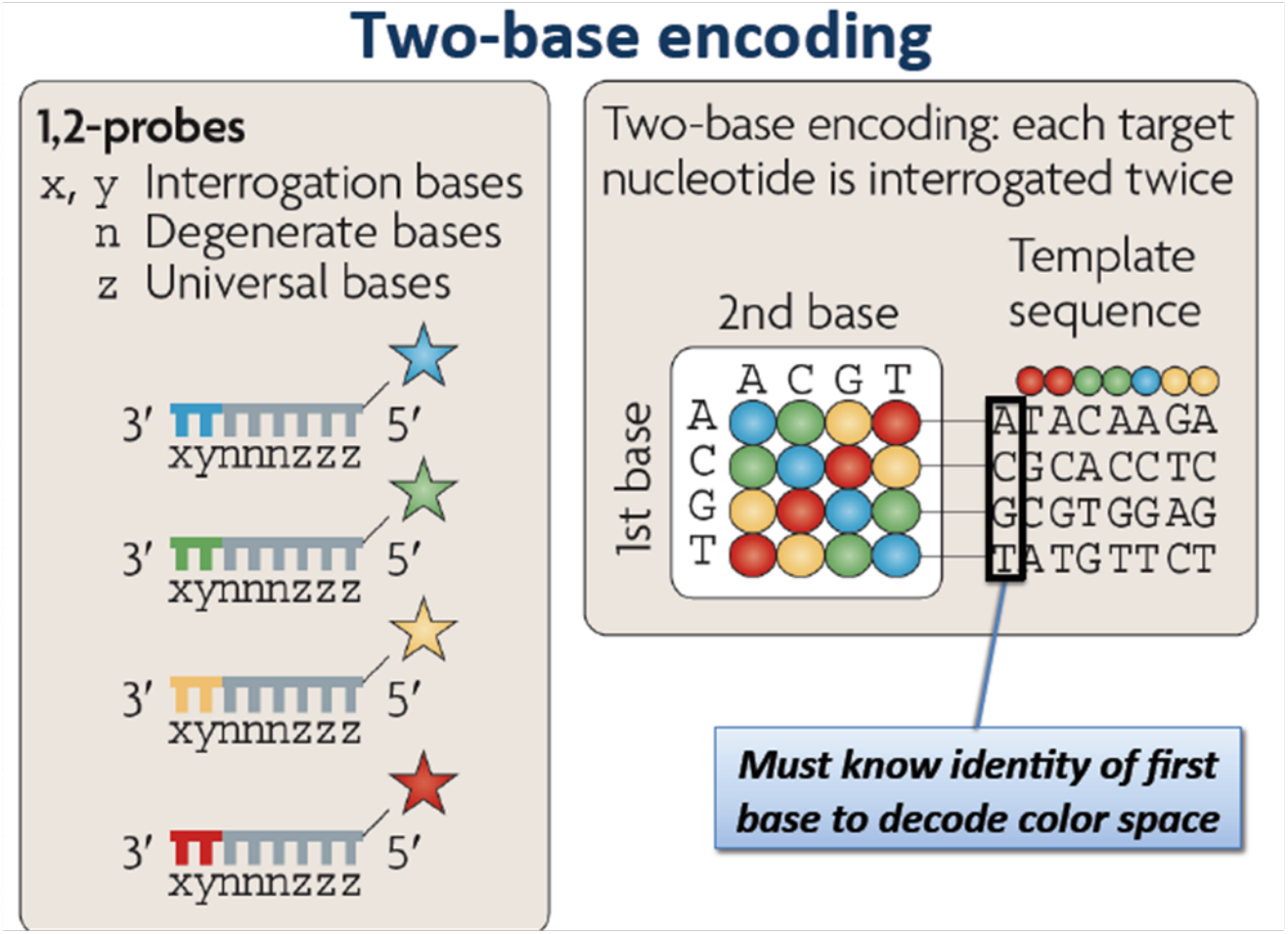
Sequencing-by-ligation (SOLiD)



DNA library preparation and titration — 4.5 h and 10.5 h → Amplification — 8.0 h → Sequencing — 7.5 h

Sequencing Workflow

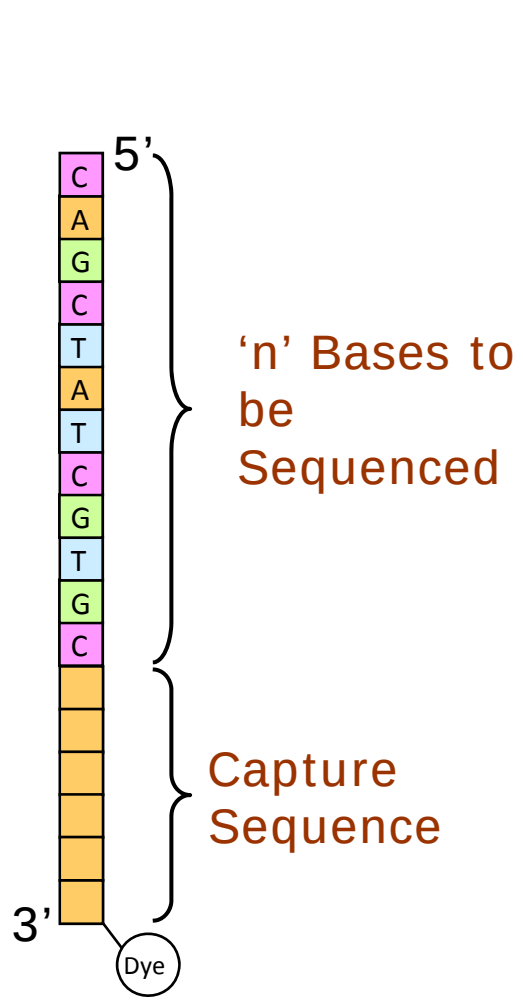
Sequencing-by-ligation (SOLiD)

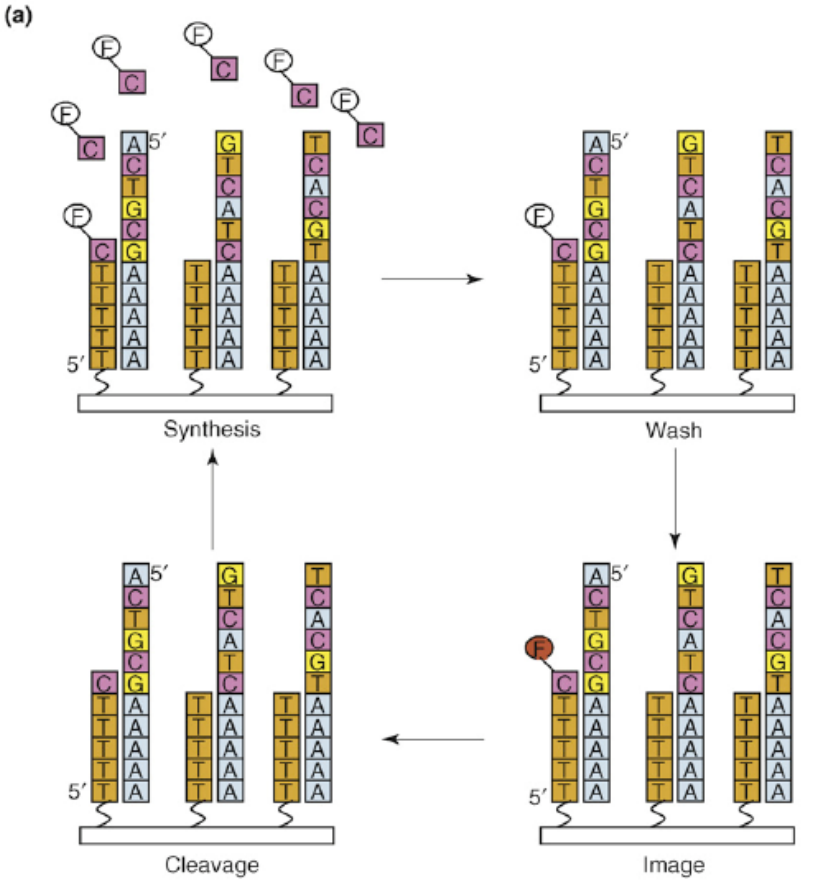


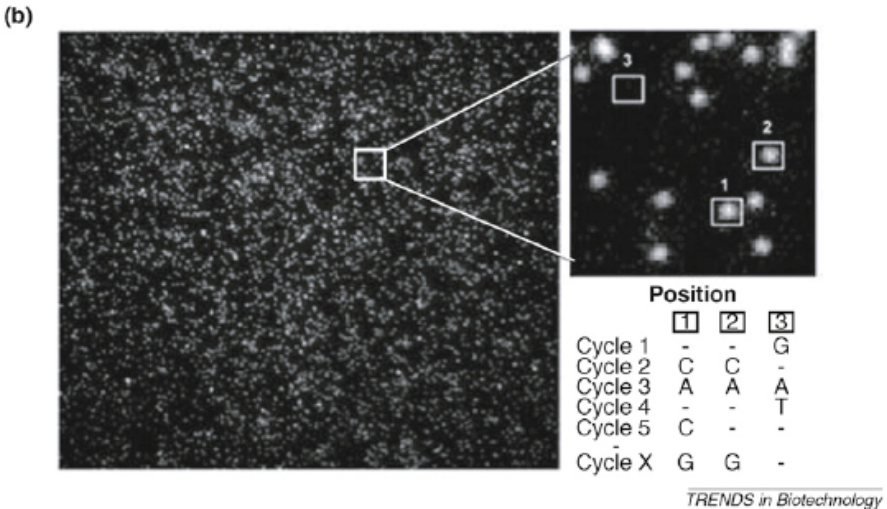
Helicos: séquencage de molécules individuelles

- Repose sur la spFRET: une méthode de microscopie qui permet de voir des sources de lumière de 5 nm (50x plus petites que la limite de résolution théorique d'un microscope)

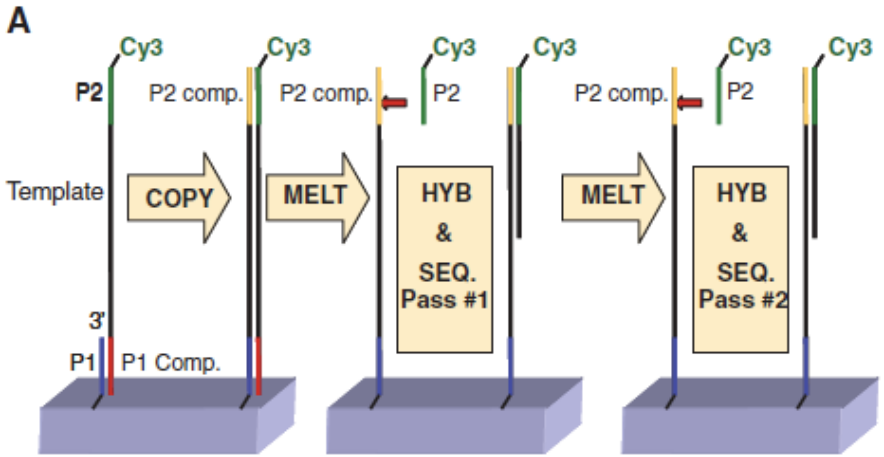
Pas d'amplification: utilisation de la transferase pour ajouterune "queue polyA"

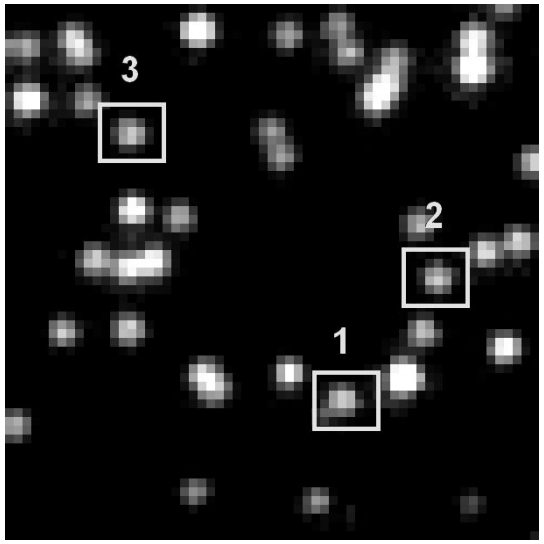


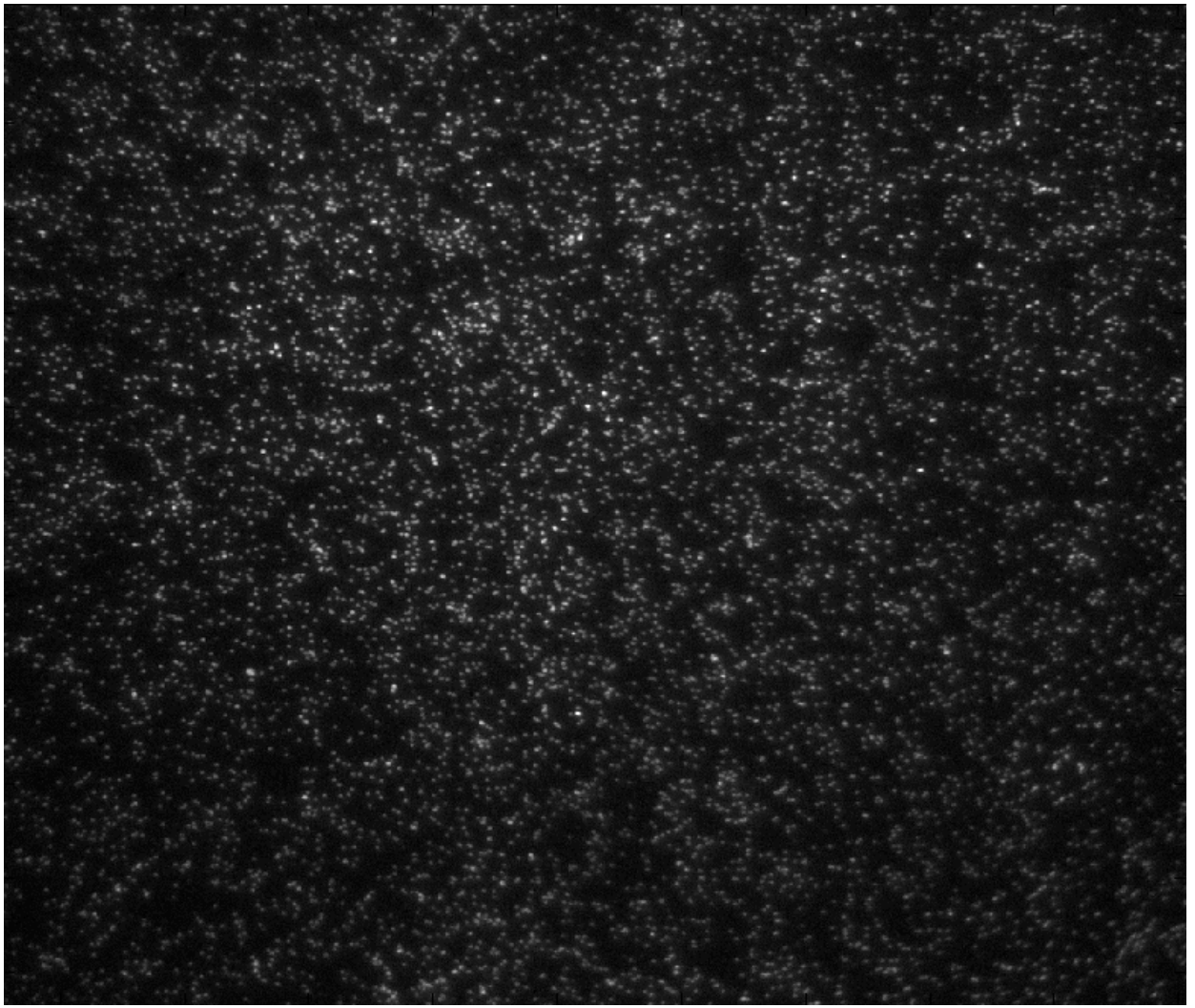

(a)
Synthesis
Wash
Image
Cleavage


(b)
Position
1 2 3
Cycle 1 - - G
Cycle 2 C C -
Cycle 3 A A A
Cycle 4 - - T
Cycle 5 C - -
Cycle X G G -
TRENDS in Biotechnology

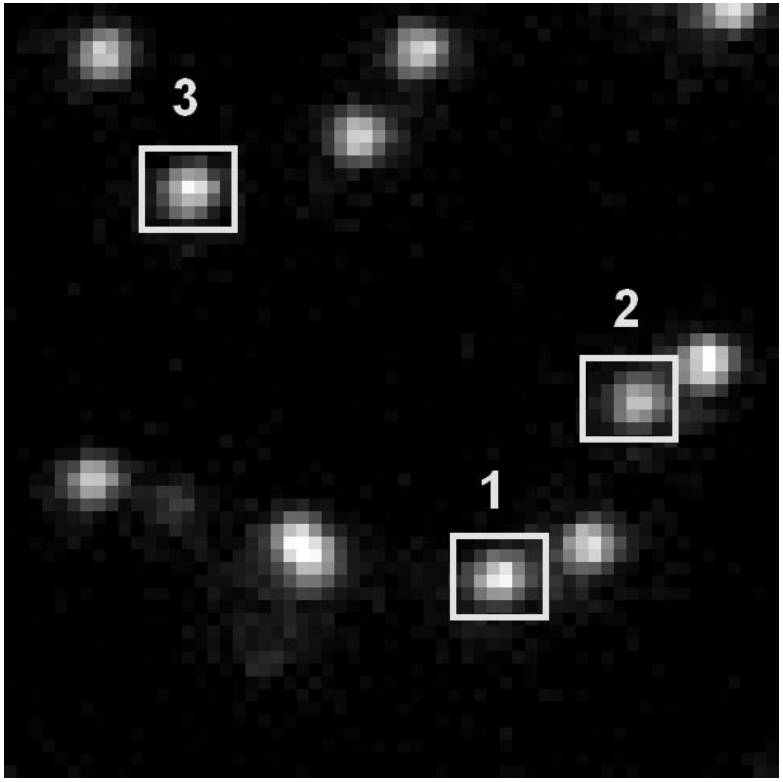
Helicos


A
Cy3
P2
P2 comp.
Template
COPY
MELT
HYB & SEQ. Pass #1
MELT
HYB & SEQ. Pass #2
3'
P1
P1 Comp.


3
2
1



Cycle: G C A G T C A



Position		
1	2	3
-	-	G
C	C	-
-	A	A
G	G	-
-	-	T
C	-	-
A	A	A

Une machine qui n’est plus commercialisée mais utilisée uniquement chez Helicos

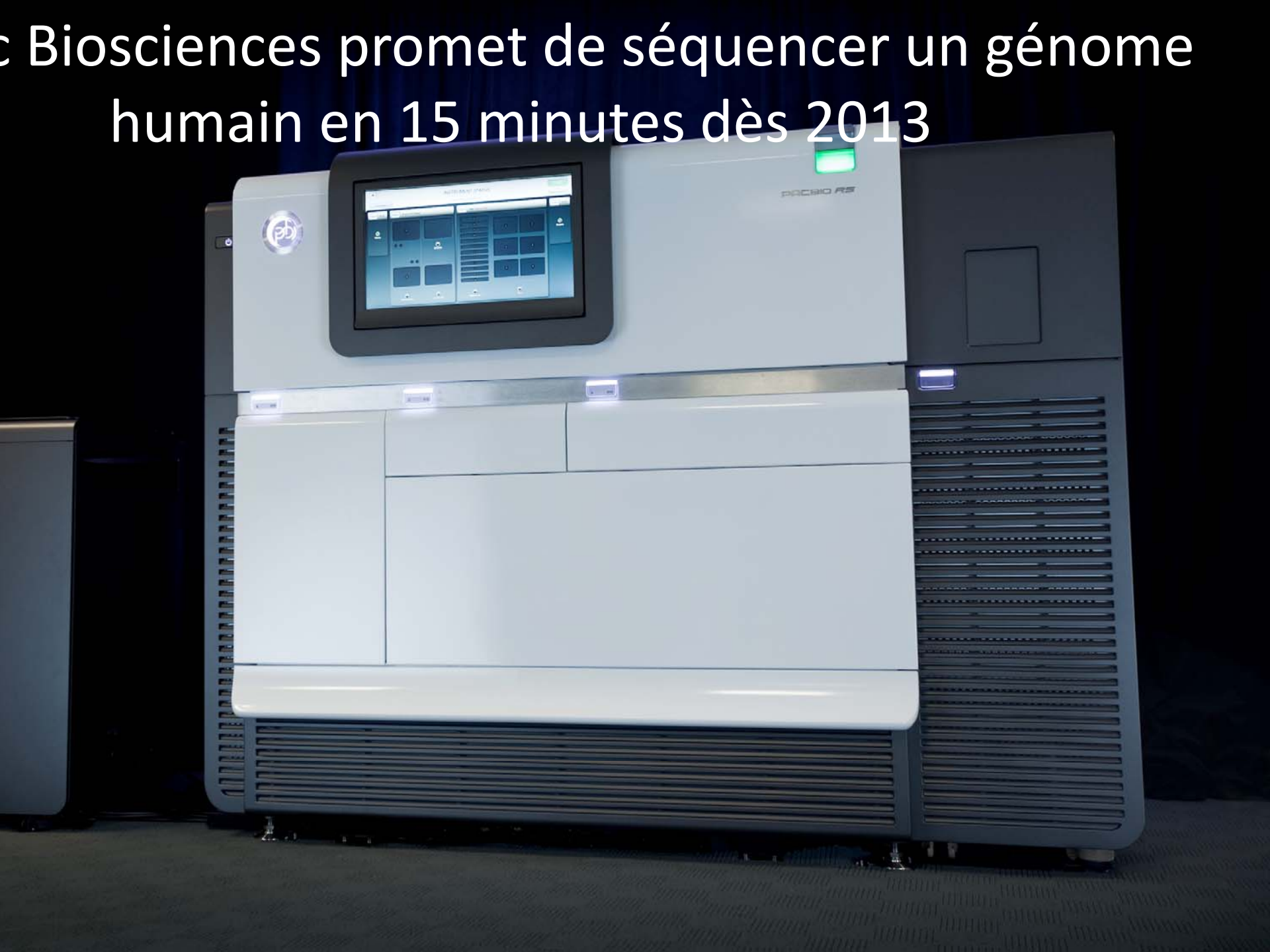
Lecture de 35 nucléotides maximum (ou 2x35)

Taux d’erreur de 0.5%

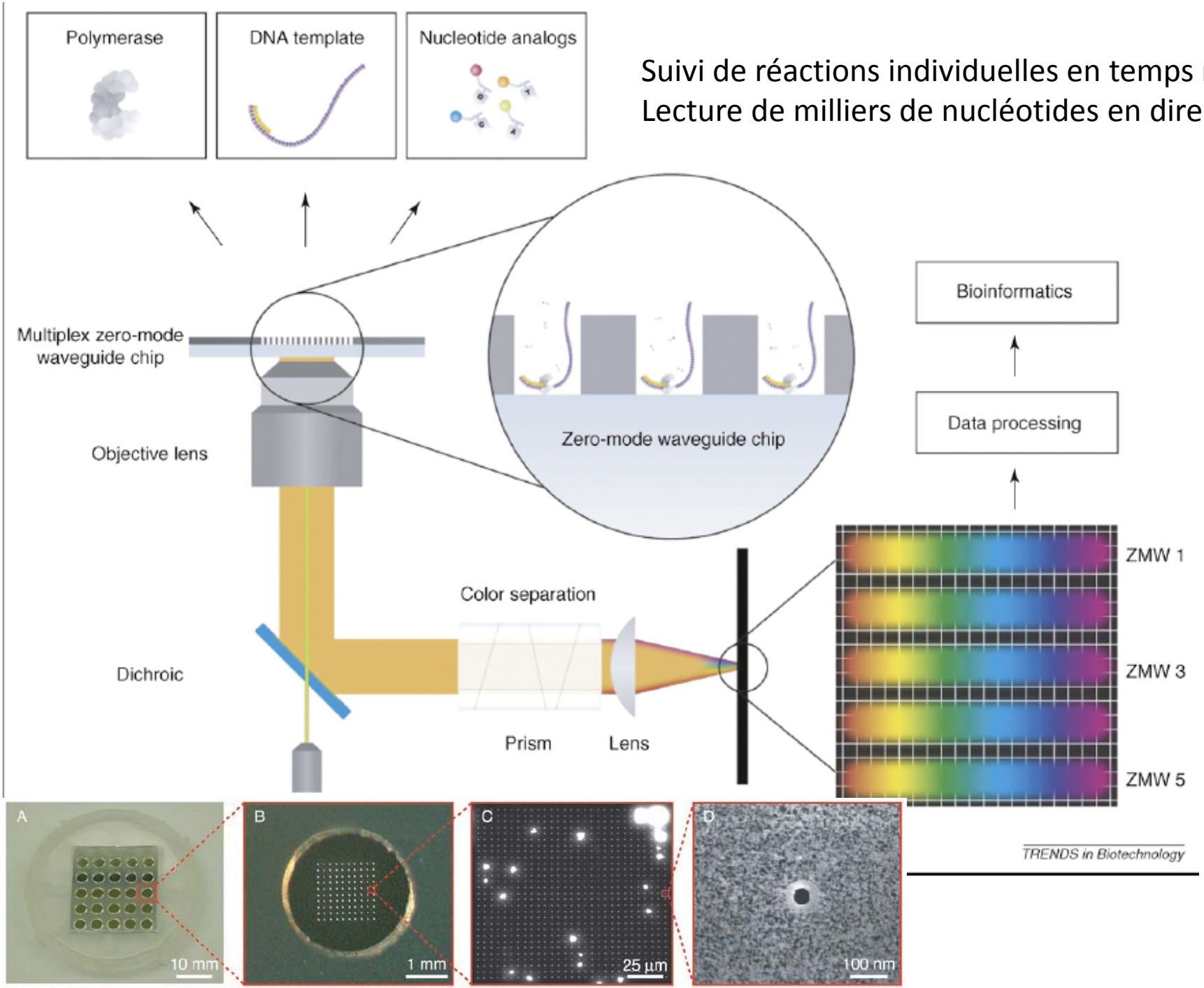
- 1Gb/heure

- (soit 100 heures pour un génome humain avec une profondeur de séquencage de x30)

c Biosciences promet de séquencer un génome humain en 15 minutes dès 2013



Suivi de réactions individuelles en temps réel
Lecture de milliers de nucléotides en direct.

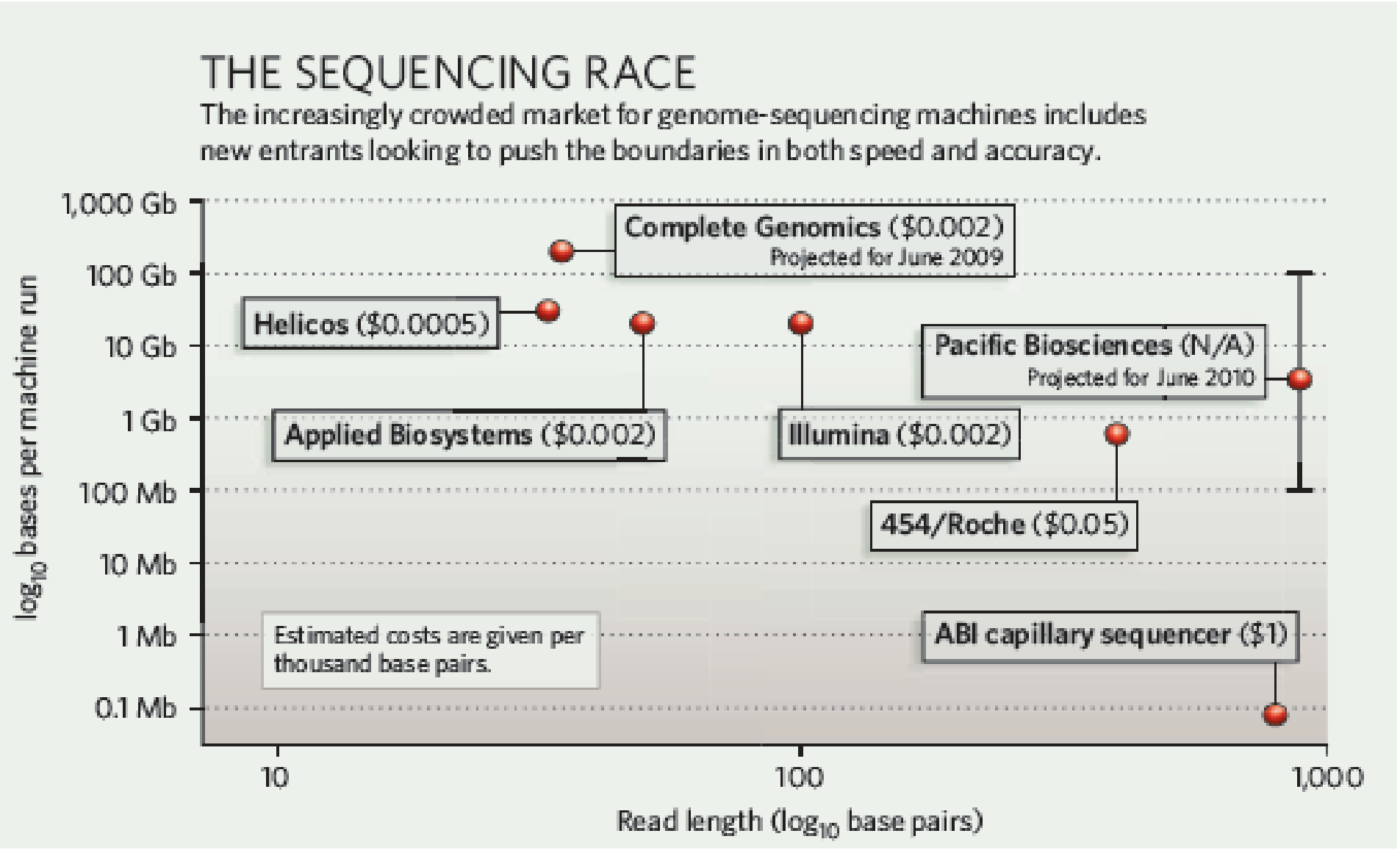


Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA$_{II}$	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	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‡, 14§	30‡, 50§	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§	12§	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‡	37‡	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. ‡Fragment run. §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.

Le séquencage de nouvelle génération: un domaine en constante évolution

Prochaines étapes: 1000 génomes humains et le génome à 1000 dollars



Préparez vous au déluge!

100 0000

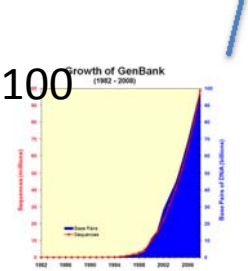
100000

10000

1000

100

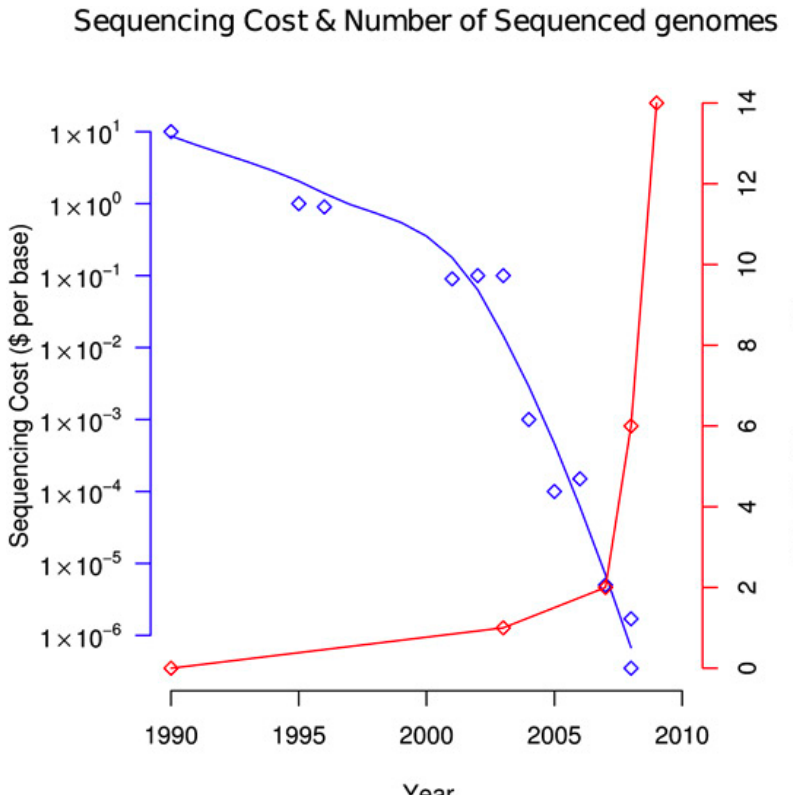
?



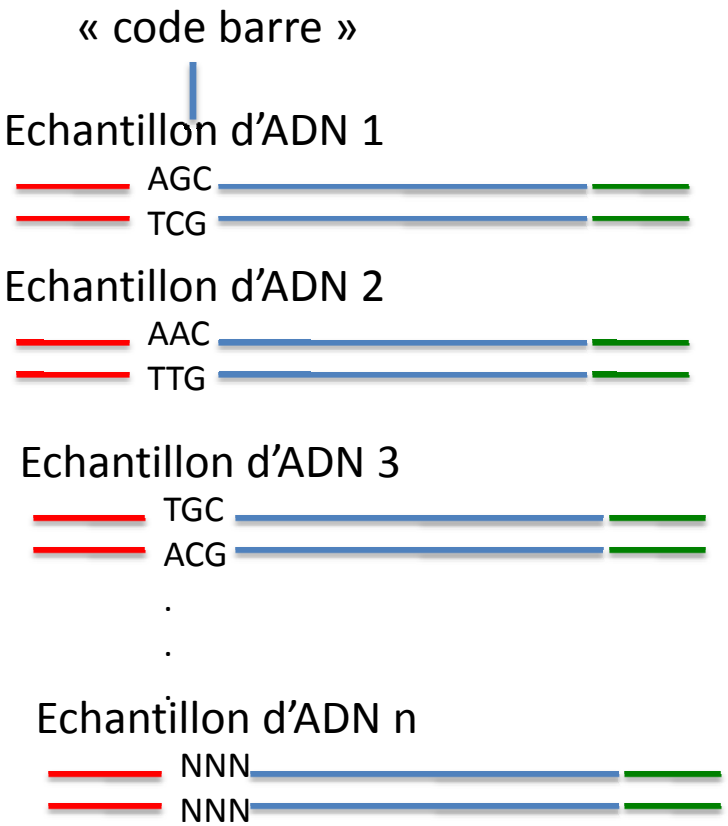
1998 2008 2018

2008: 100 Gb dans Genebank
1 machine nouvelle génération: 10 Gb par jour
Des centaines de machines vendues en 2009
1000 genomes à 1000 $?

Janvier 2011
1000 génomes bactériens,
628 espèces animales ou végétales dont:
39 génomes complets
256 « brouillons » assemblés
333 en cours d'assemblage



Multiplexage Souplesse et baisse des prix

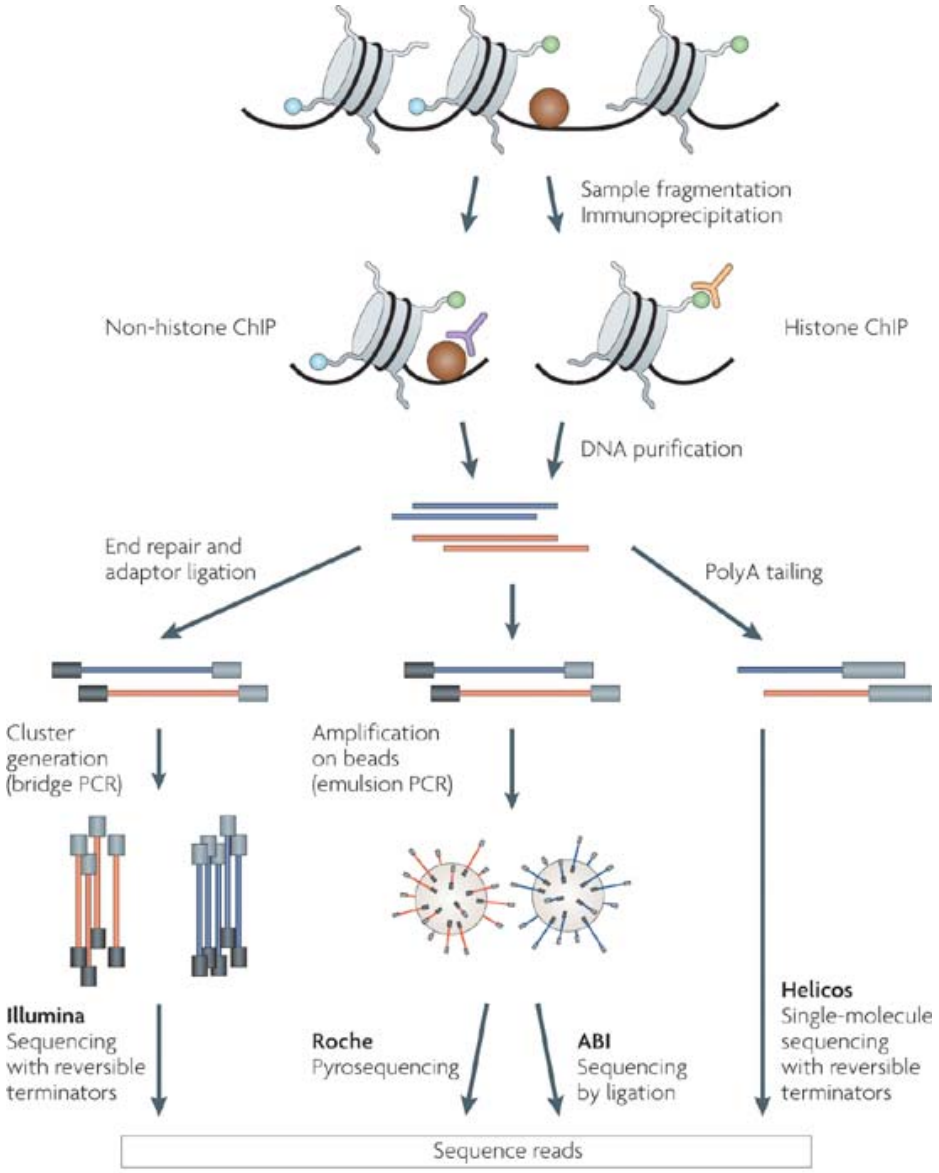


- Amplification par PCR
- mélange (jusqu 'à 96 banques)
-Séquencage massif
-Utilisation du code barre pour séparer les données

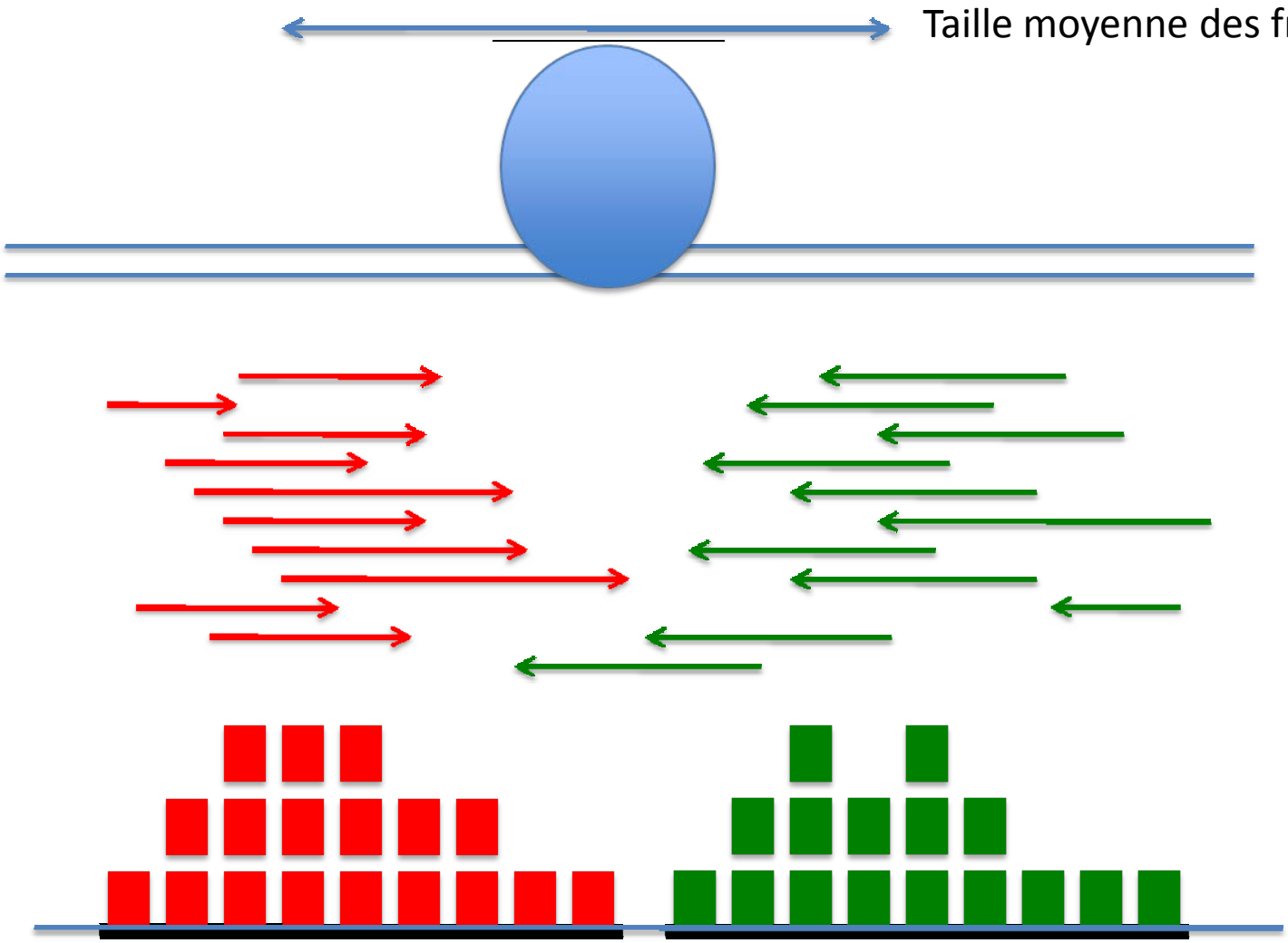
Le séquencagehaut-débit ne sert pas seulement à séquencer des génomes:

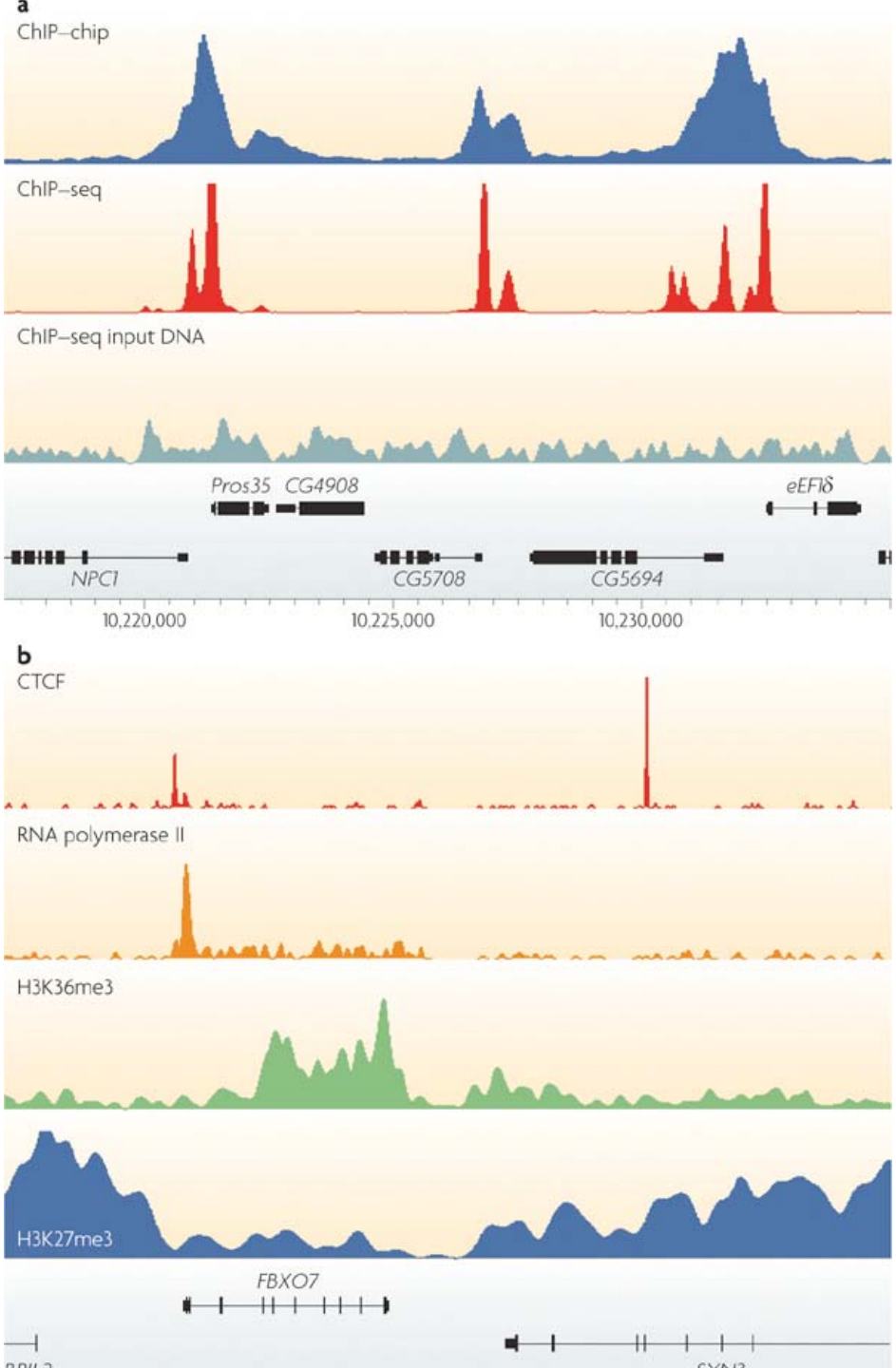
Epigénome
Métagénome
Transcriptome

2) ChipSeq: localiser les sites de fixation des protéines dans le génome

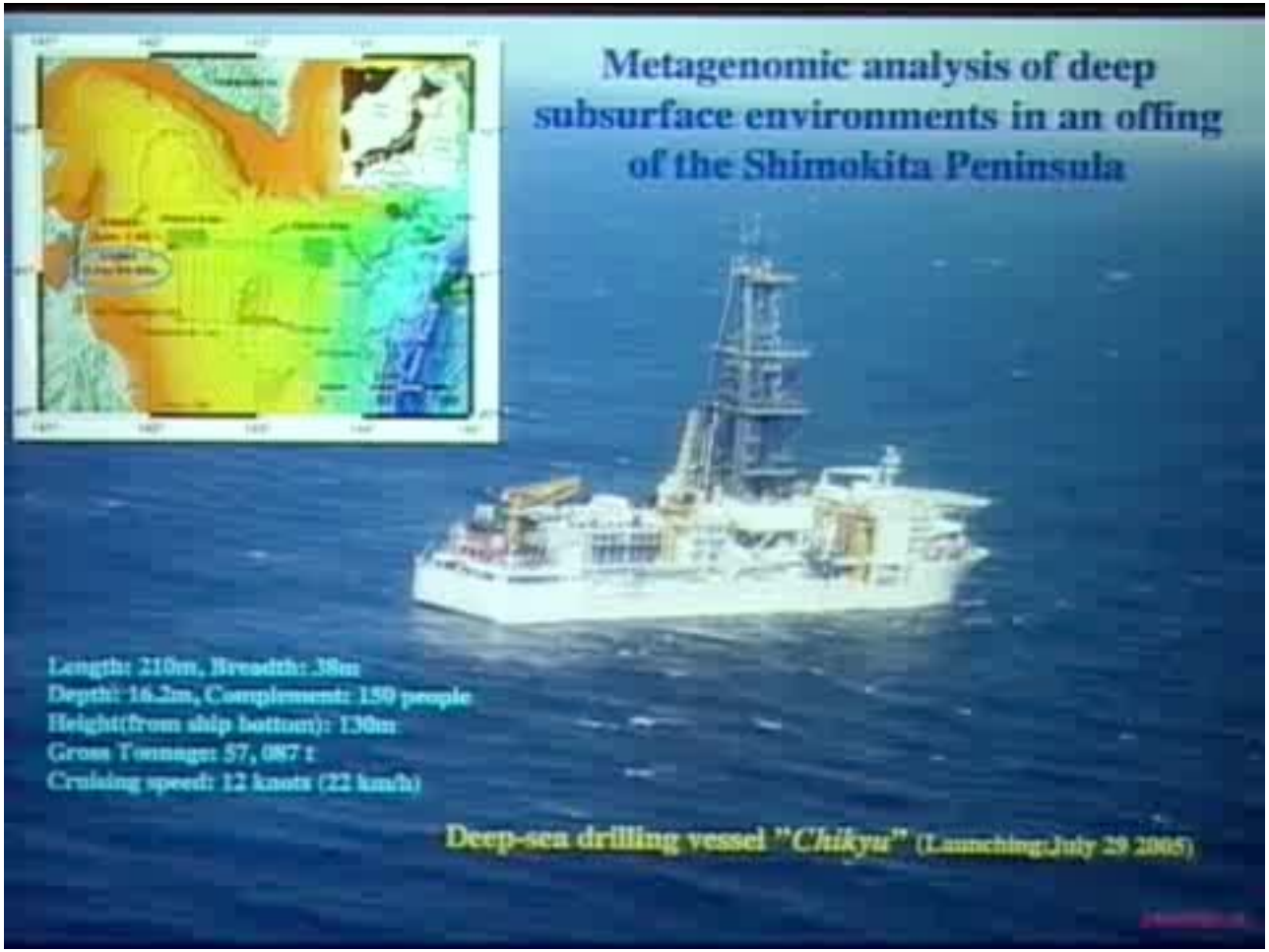


Exemple de localisation d'un site de fixation génomique




a
ChIP–chip
ChIP–seq
ChIP–seq input DNA
Pros35
CG4908
eEF1δ
NPC1
CG5708
CG5694
10,220,000
10,225,000
10,230,000
b
CTCF
RNA polymerase II
H3K36me3
H3K27me3
FBXO7

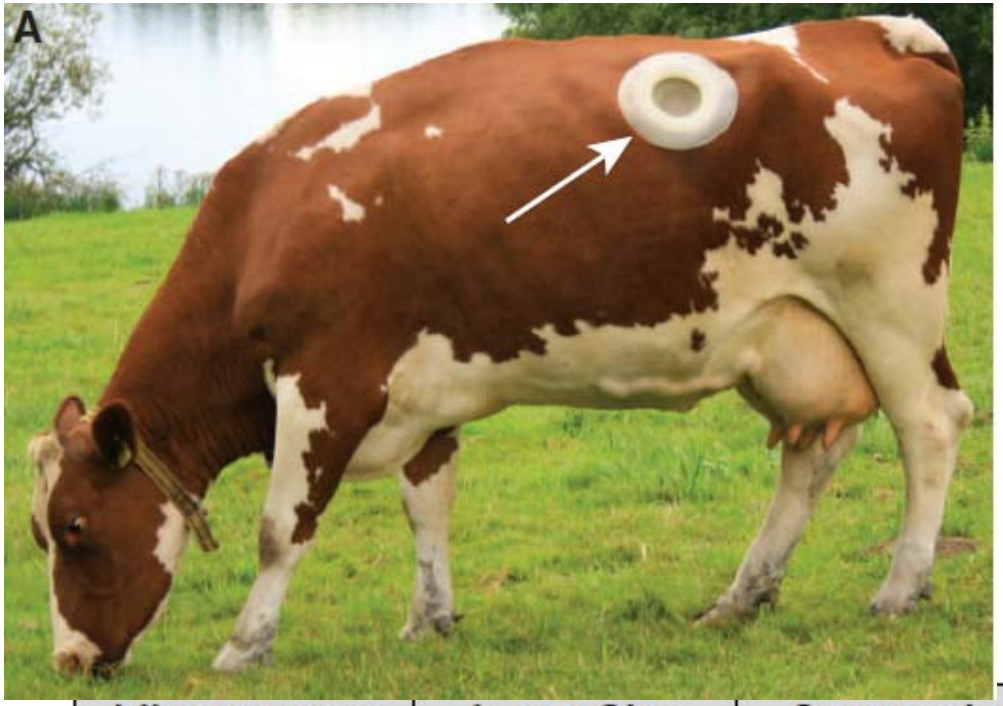
3) Métagénome



Analyse de la biodiversité des profondeurs océaniques par séquencage de l'ADN extrait des fonds océaniques:

-Description de nouvelles especes de microorganismes

-Estimation de la composition des populations de microorganismes



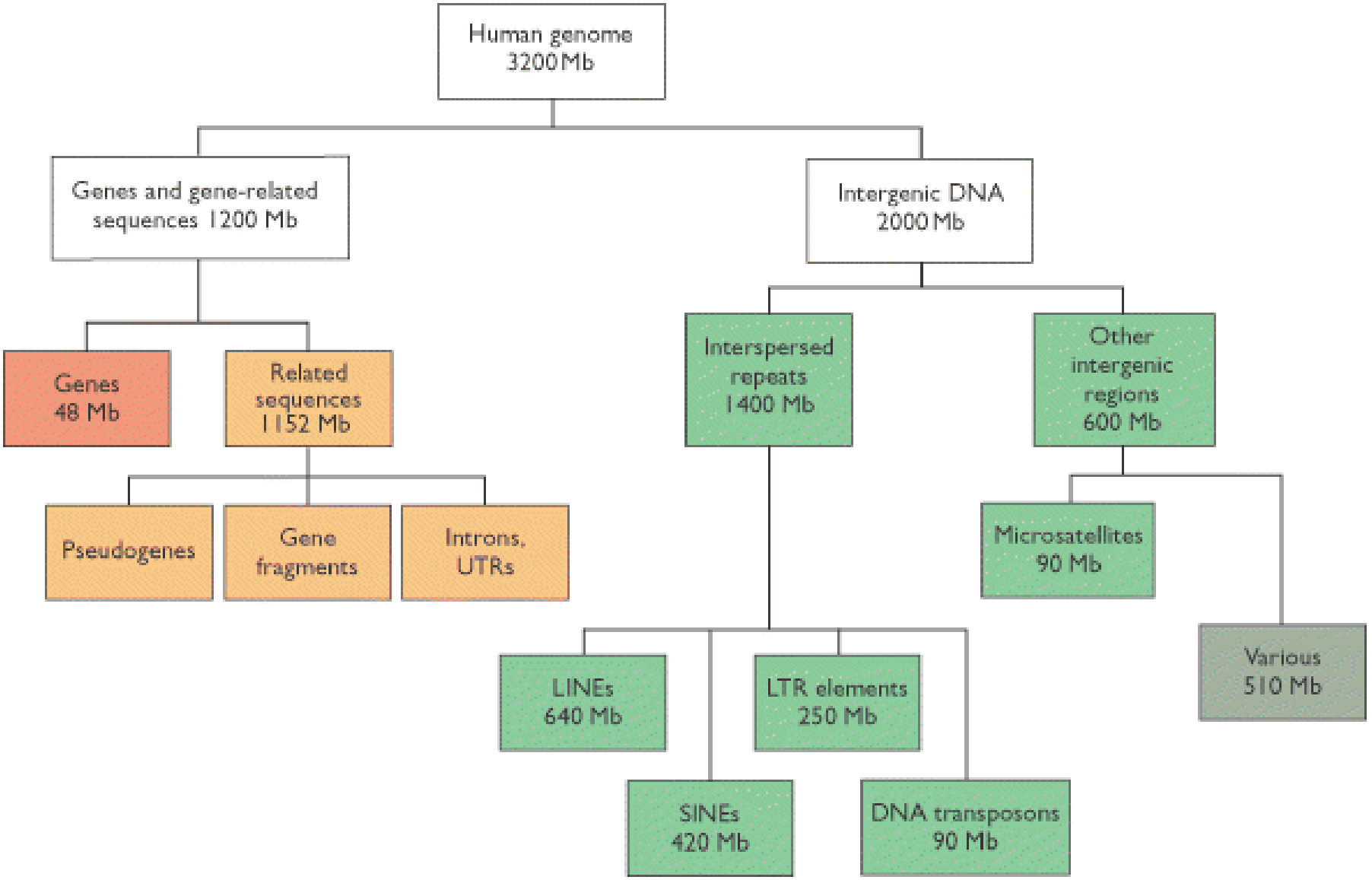
Analyse des bactéries du rumen de la vache:
260 Gb

15 nouveaux génomes assemblés
27 755 genes induits par les carbohydrates
Hess *Science* ***331, 463 (2011)***

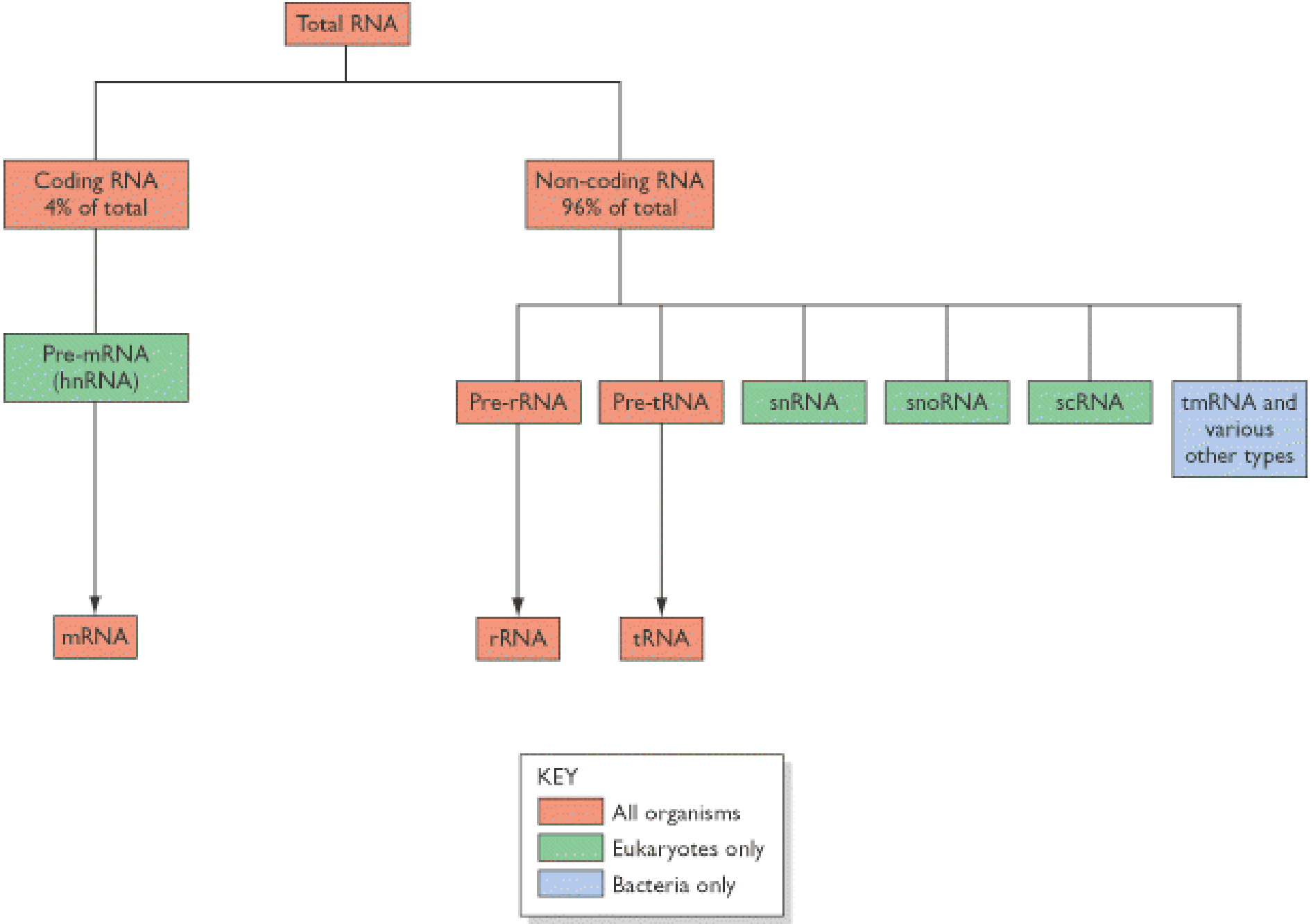
Library name	Insert Size	Sequencing	Bases (Gb)	Reads (M)
GAII_200bp	200 bp	2x125bp	17.3	138.3
GAII_3kb[1)]	3 kbp	2x36bp	4.5	124.5
GAII_3kb		2x75bp	27.2	357.6
GAII_5kb[1)]	5 kbp	2x36bp	4.0	110.8
GAII_5kb		2x75bp	22.8	300.6
HiSeq_300bp	300 bp	2x36bp	3.9	108.9
		2x100bp	188.2	1,863.6
Total			**267.9**	**3,004.3**

Table S3. Description of generated libraries and employed sequencing technologies

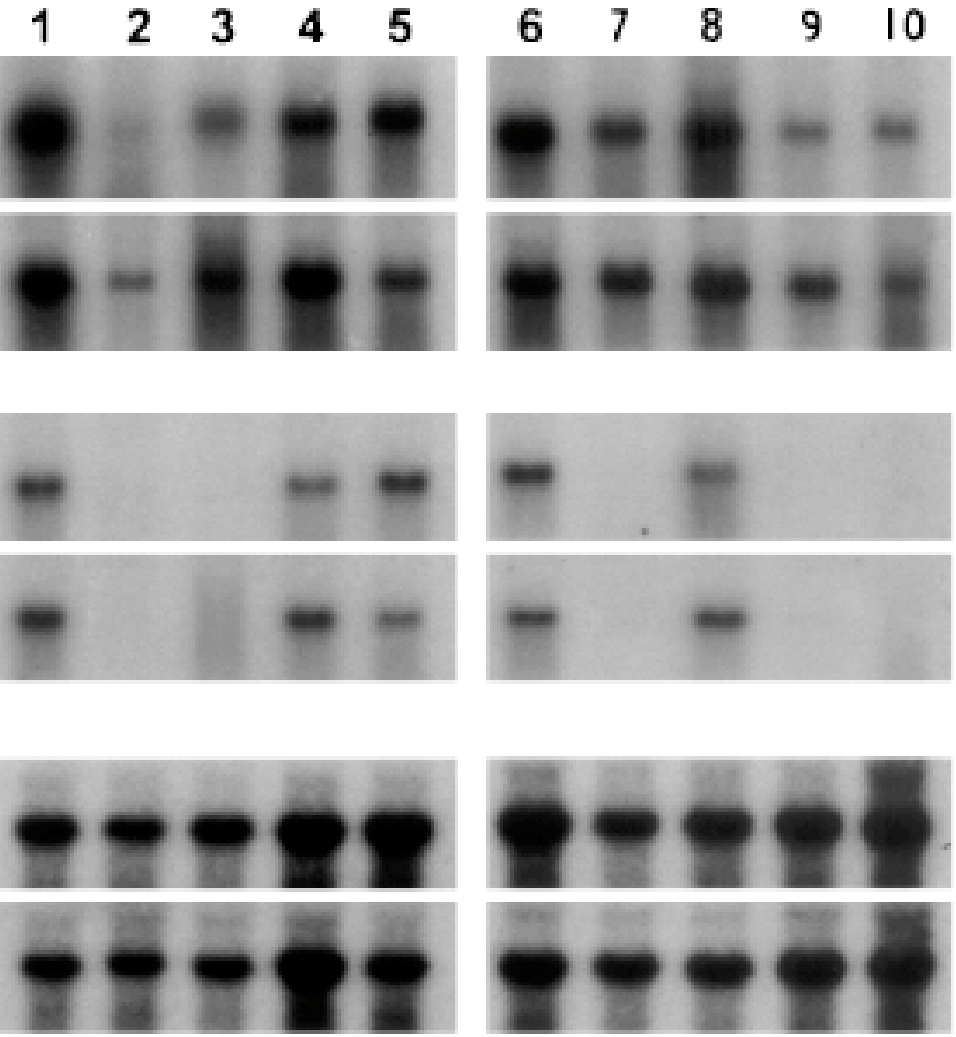
Le transcriptome



Les ARN messagers: 1% du génome

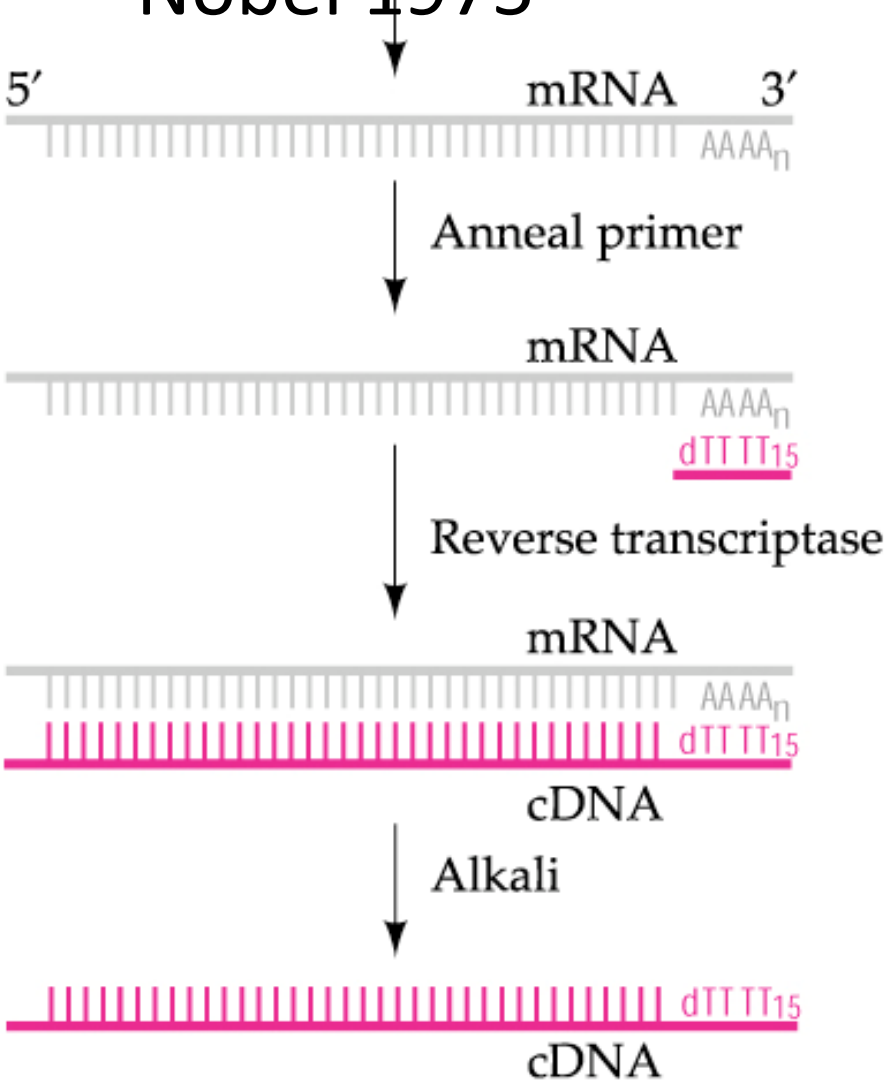


Northernblotting



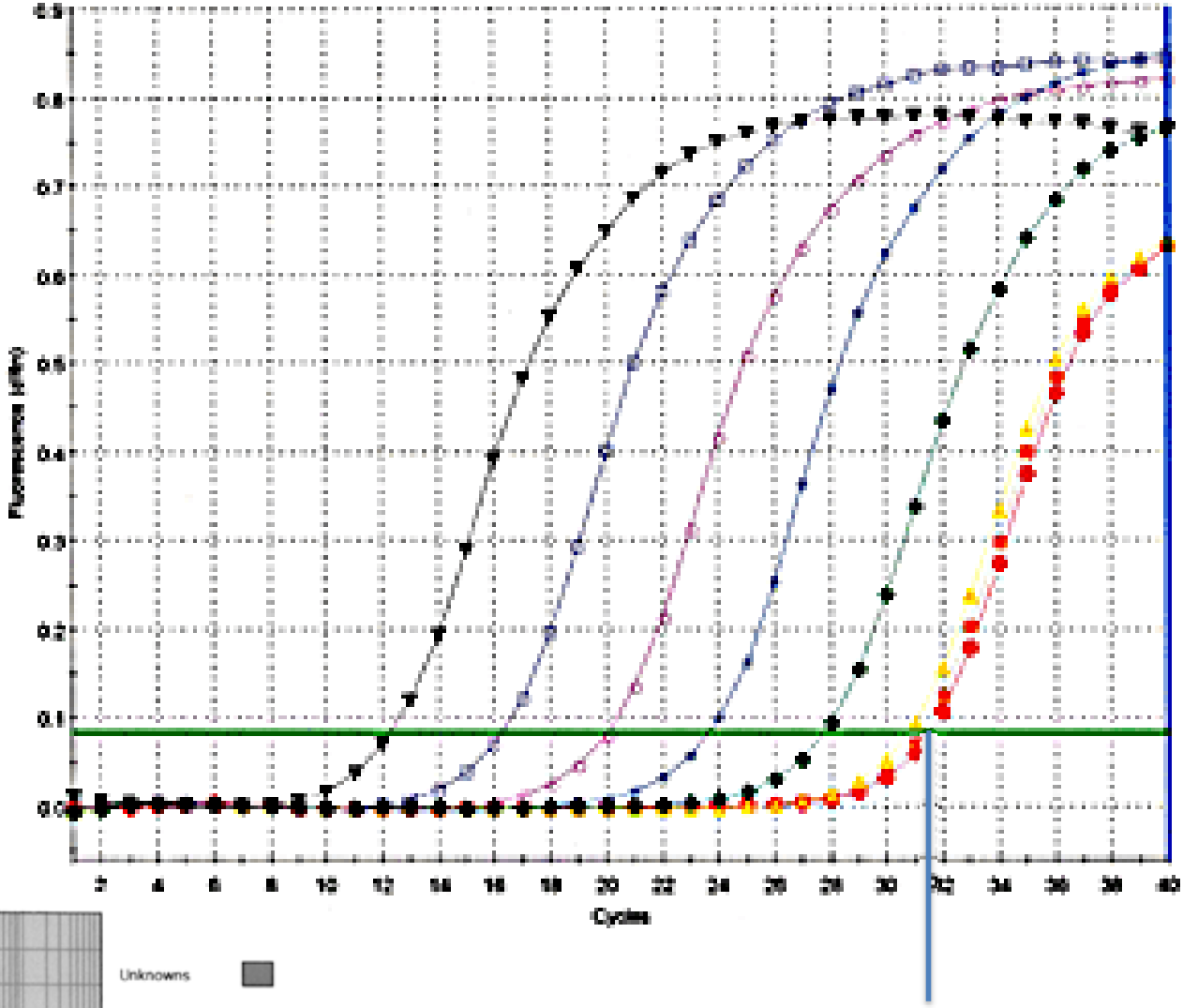
1) Un débit très faible (une dizaine de mesures par jour)
2) Une sensibilité limitée
3) Une information imprécise sur la taille et l'abondance des ARN

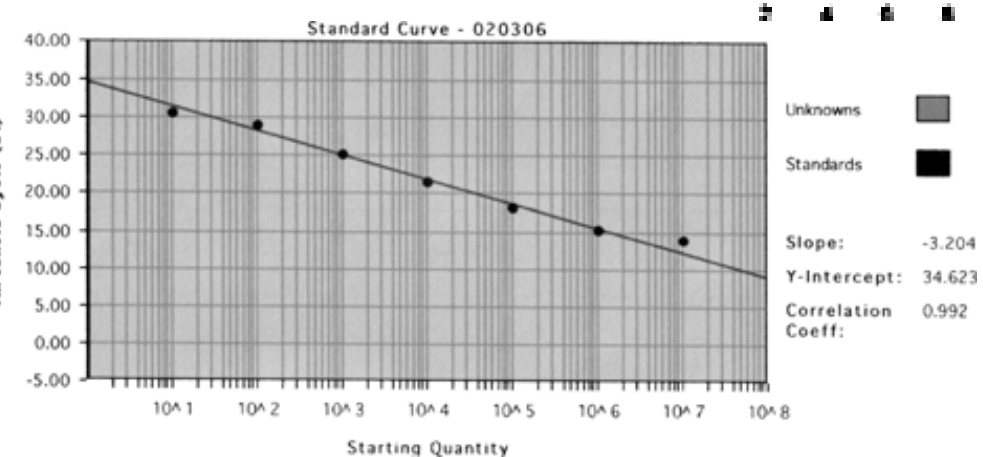
- Howard Temin
- David Baltimore Prix Nobel 1975



La RT-PCR quantitative:

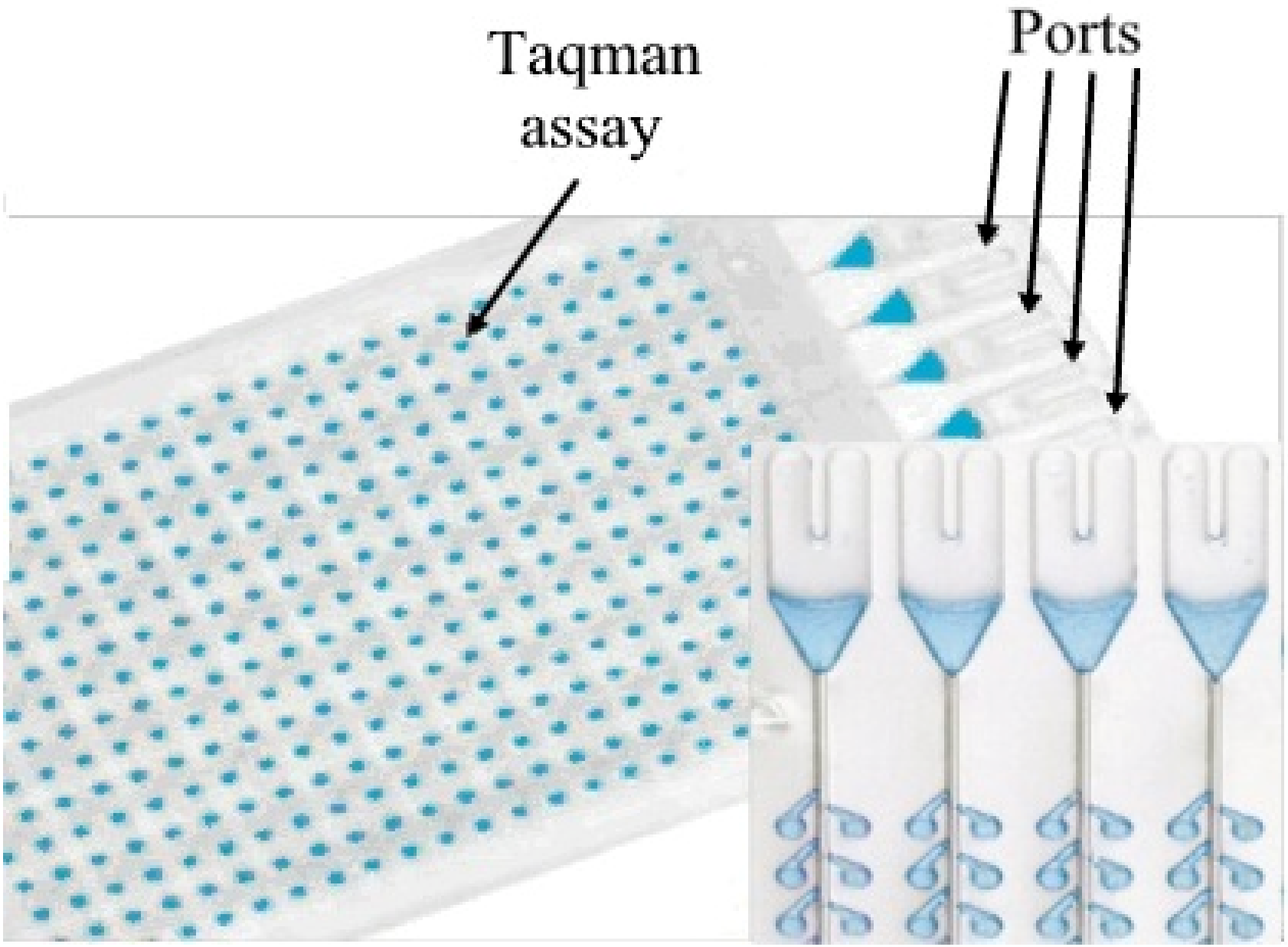
1) Extraction de l'ARN
2) Transcription inverse
3) PCR « en temps réél »



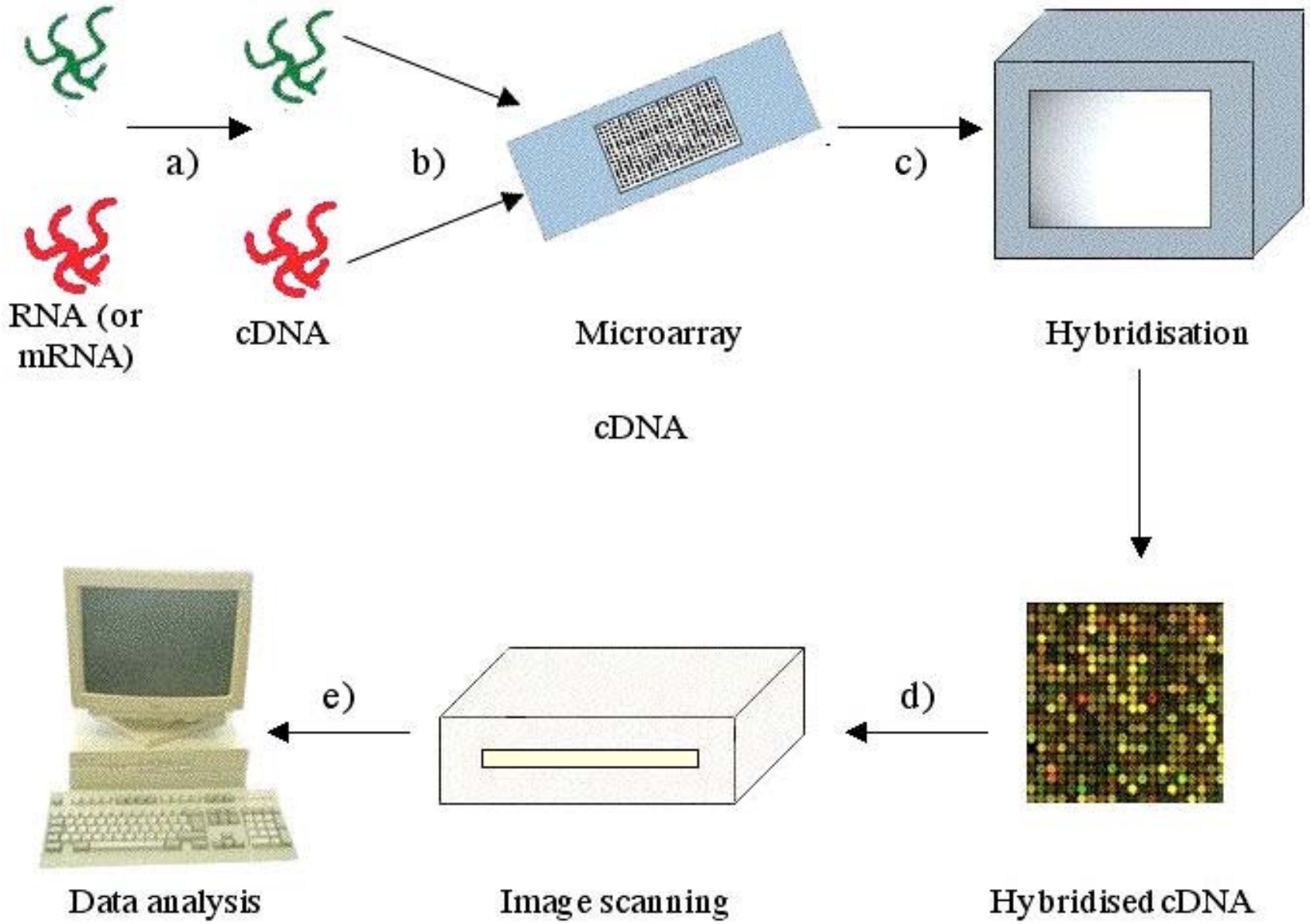


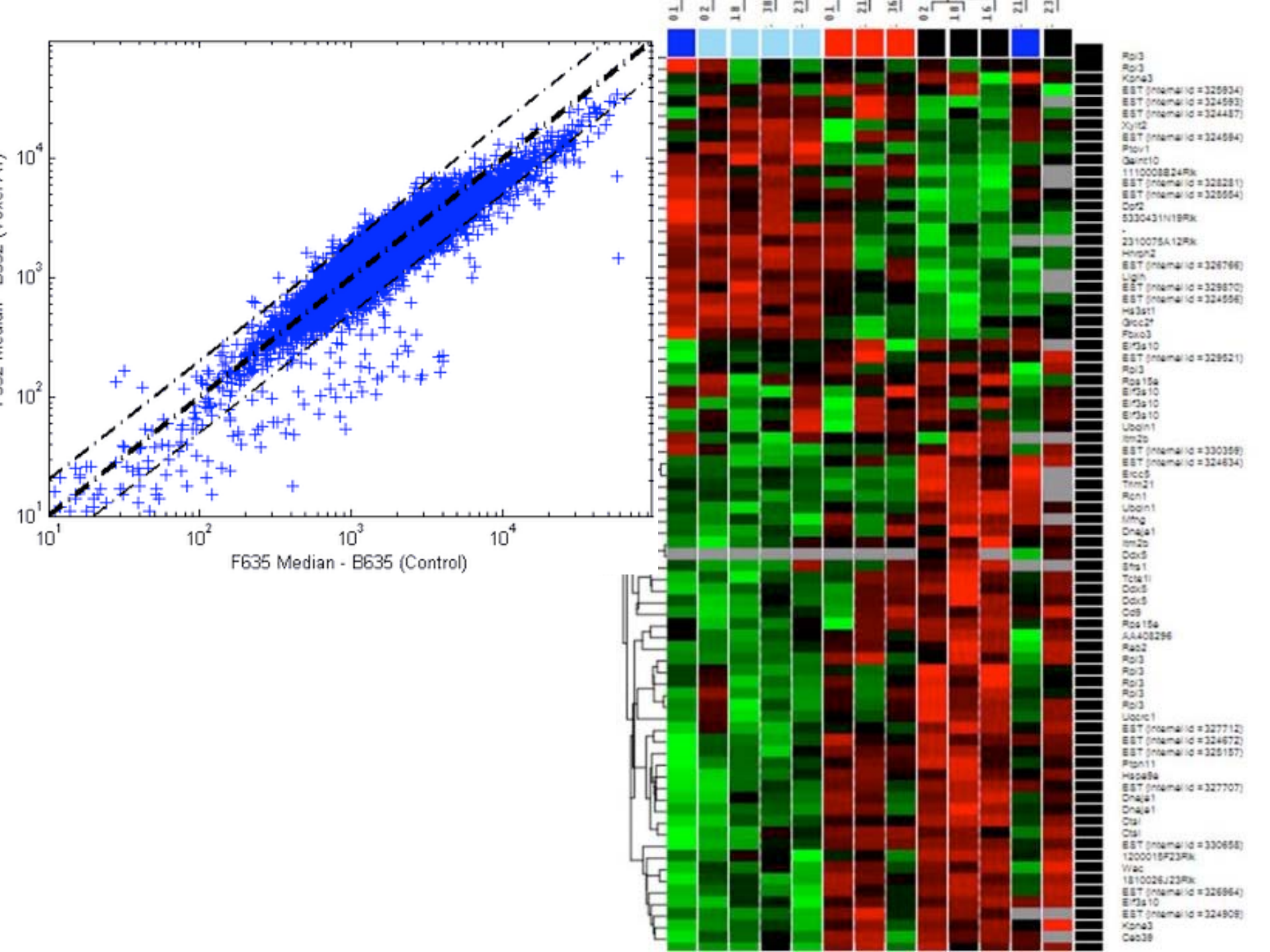
Ct: cycle de PCR à partir duquel l'ADN amplifié devient détectable

Les plaques microfluidiquesAppliedBiosystems: 384 mesures en deux heures pour 384 €

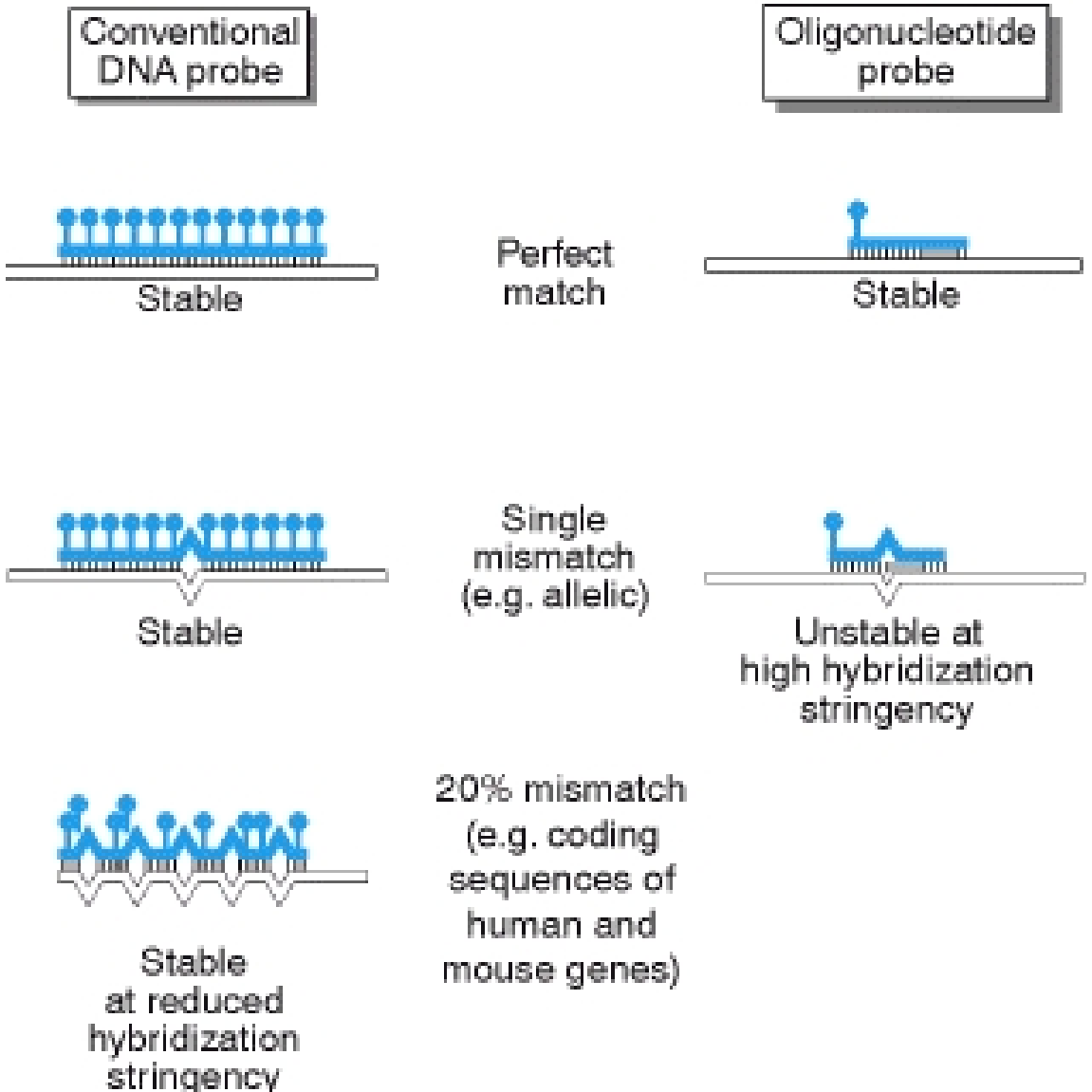


Les microarrays: mesurer l'abondance de 25000 ARNm en 24 heures par hybridation moléculaire




F635 Median - B635 (Control)
10^1
10^2
10^3
10^4

Principal problème: réactions croisées et hybridation non spécifiques

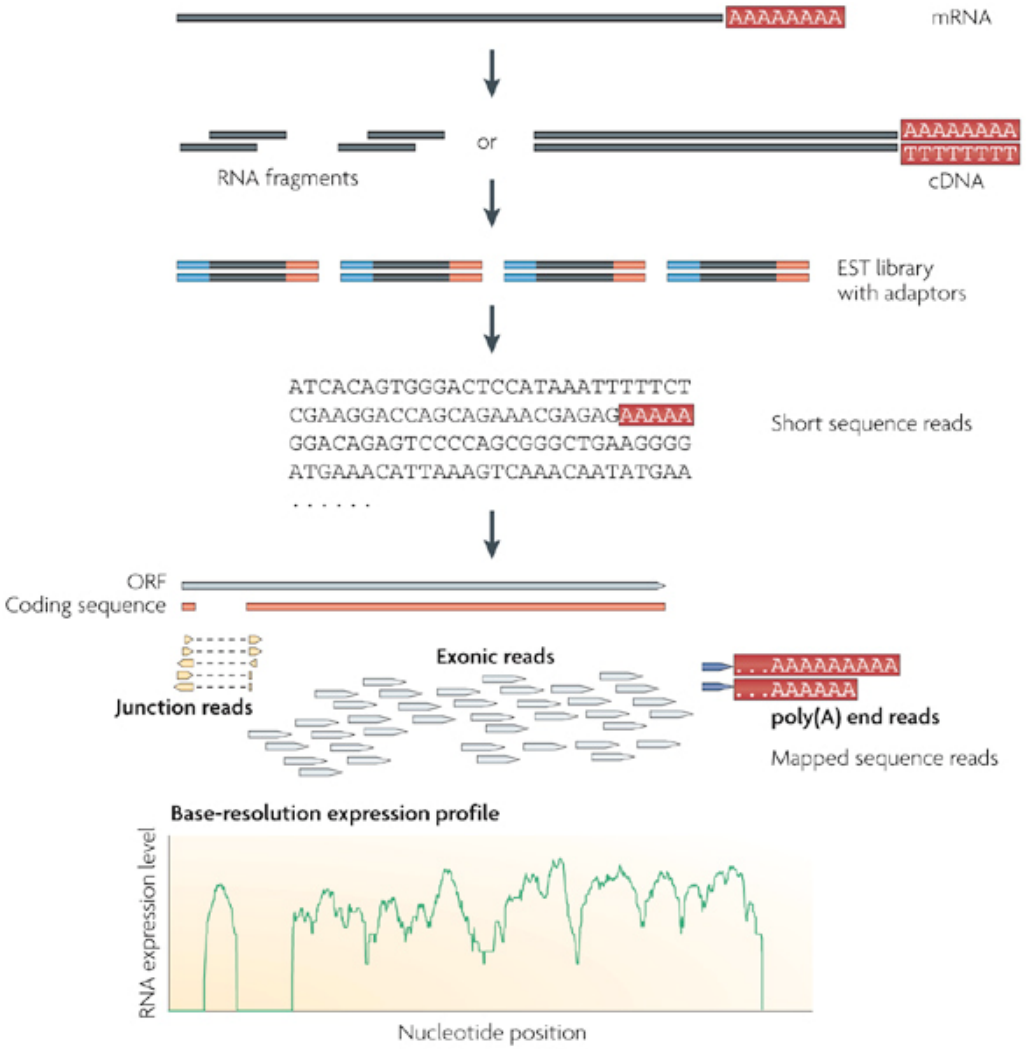


Bilan sur les microarrays

- Une méthode:
- Haut débit
- Sensible
- Pour une quantification relative
- Mais:
- Orientée vers des séquences connues
- Peu précise
- Coûteuse
- Difficile
- Peu fiable
- Sans avenir

RNAseq

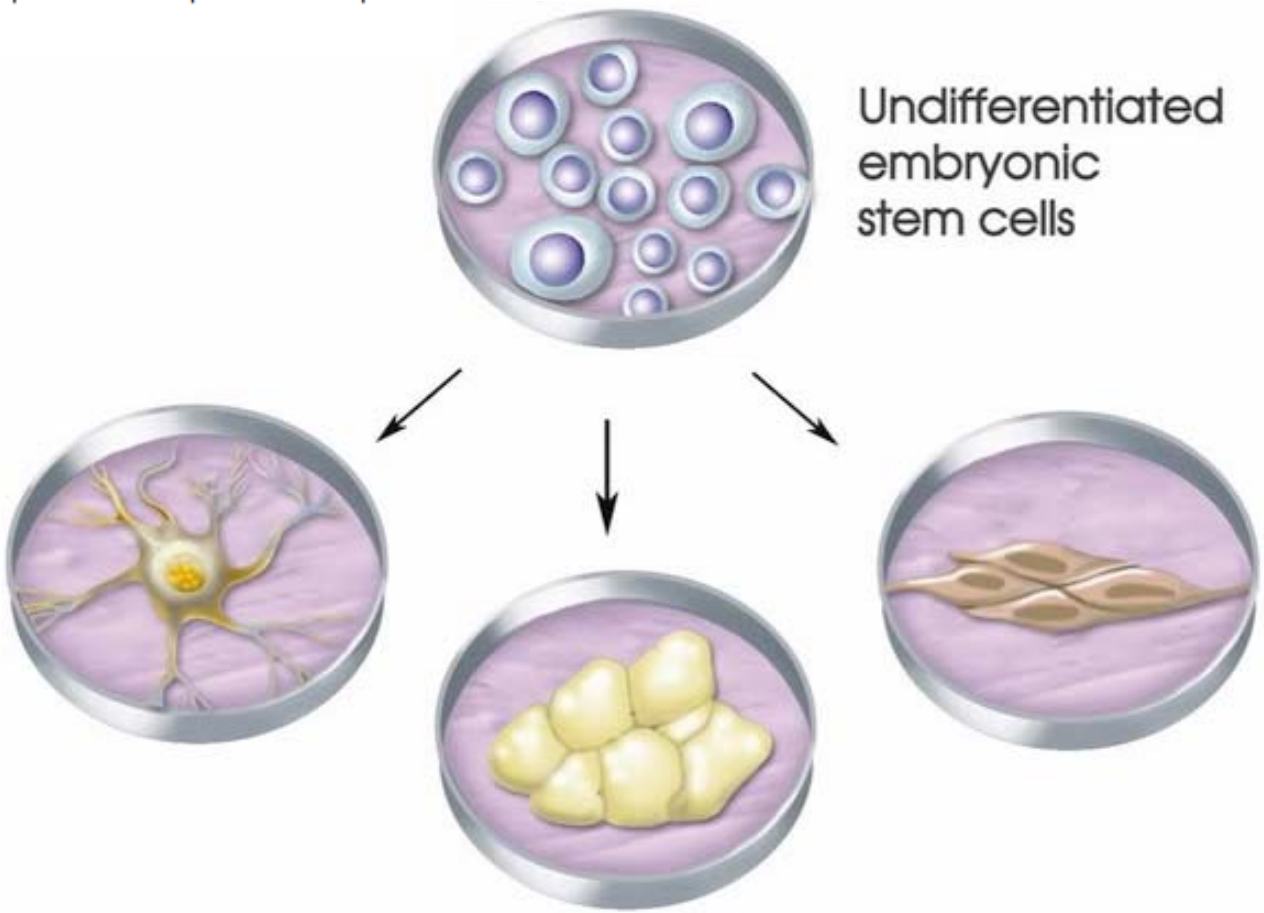
La fréquence avec laquelle un fragment d'un gène est retrouvé ne dépend que:
-De l'efficacité de la transcription inverse (qu'on suppose constante)
-De la taille du gène
- de l'abondance de l'ARNm dans les cellules au départ.



Stem cell transcriptome profiling via massive-scale mRNA sequencing

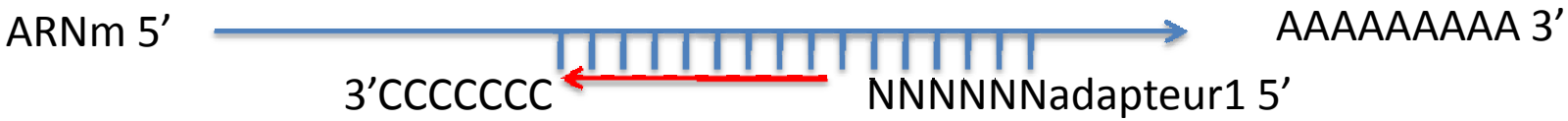
Nicole Cloonan[1,4], Alistair R R Forrest[1,3,4], Gabriel Kolle[1,4], Brooke B A Gardiner[1], Geoffrey J Faulkner[1], Mellissa K Brown[1], Darrin F Taylor[1], Anita L Steptoe[1], Shivangi Wani[1], Graeme Bethel[1], Alan J Robertson[1], Andrew C Perkins[1], Stephen J Bruce[1], Clarence C Lee[2], Swati S Ranade[2], Heather E Peckham[2], Jonathan M Manning[2], Kevin J McKernan[2] & Sean M Grimmond[1]

NATURE METHODS | VOL.5 NO.7 | JULY 2008 | 613



Protocole

- Culture de cellules souches dans un milieu assurant leur autorenouvellement ou pendant 24 heures dans un milieu déclenchant leur différenciation (expérience tripliquée)
- extraction des ARNm (6 échantillons)
- Synthèse d'ADNc simple brin



- Synthèse deuxieme brin (amorce 5'adapteur2GGGGGGG3')
- Amplification PCR (amorces complémentaires adapteur 1 + 2)(nombre de cycles=20)
- PCR en émulsion sur billes
- Séquencage (SOLID) >300 millions de séquencesx35 nt= 10Gb

Analyse des données

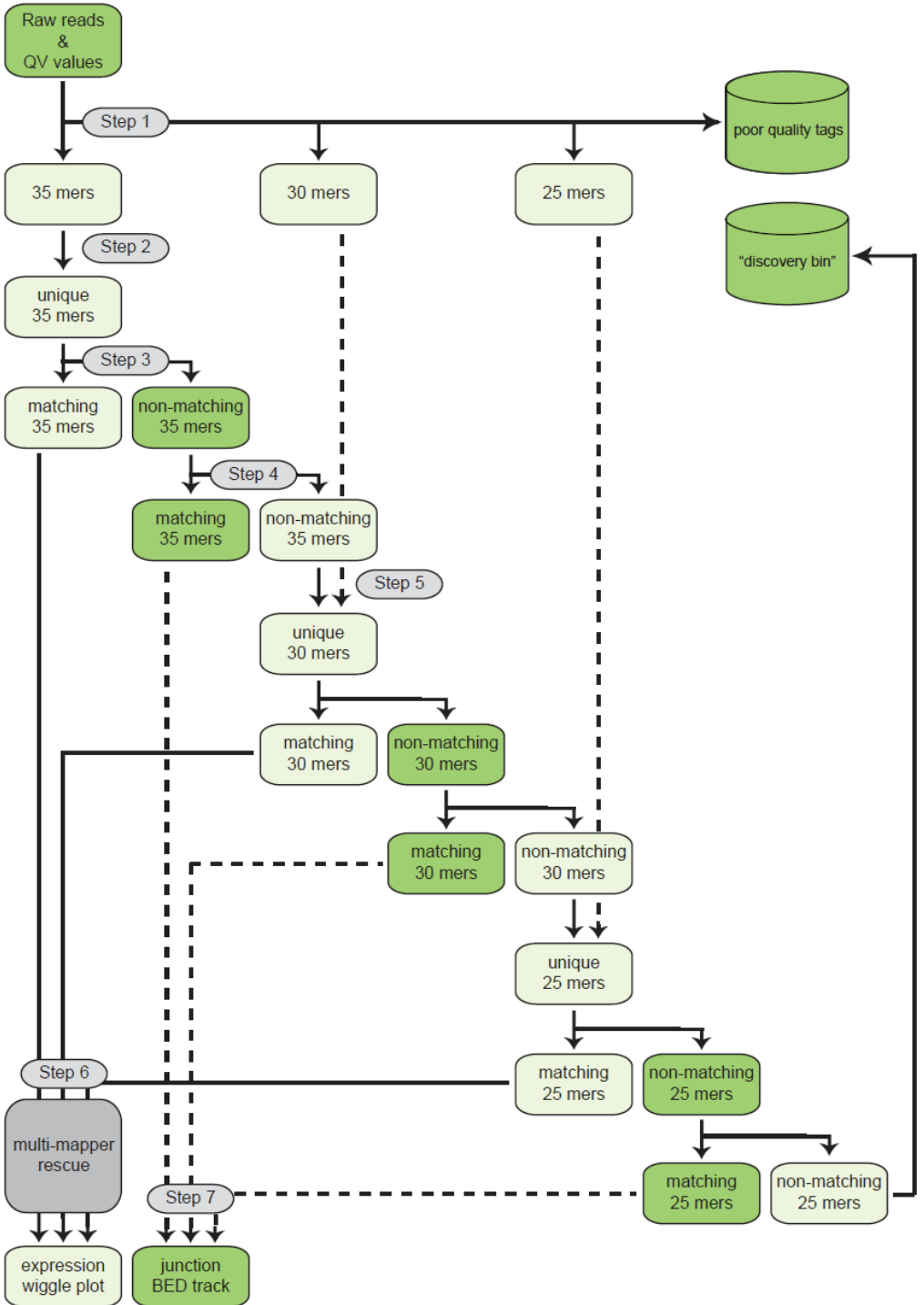
1) Séquencage 35 nt

Step1) Contrôle de qualité= accepter si moins de 5 nucléotides illisibles. Sinon élimination des 5 derniers nucléotides

Step2) Comptage des « tags » identiques.

Step3) Blast refseq. Si = >100 sites éliminer

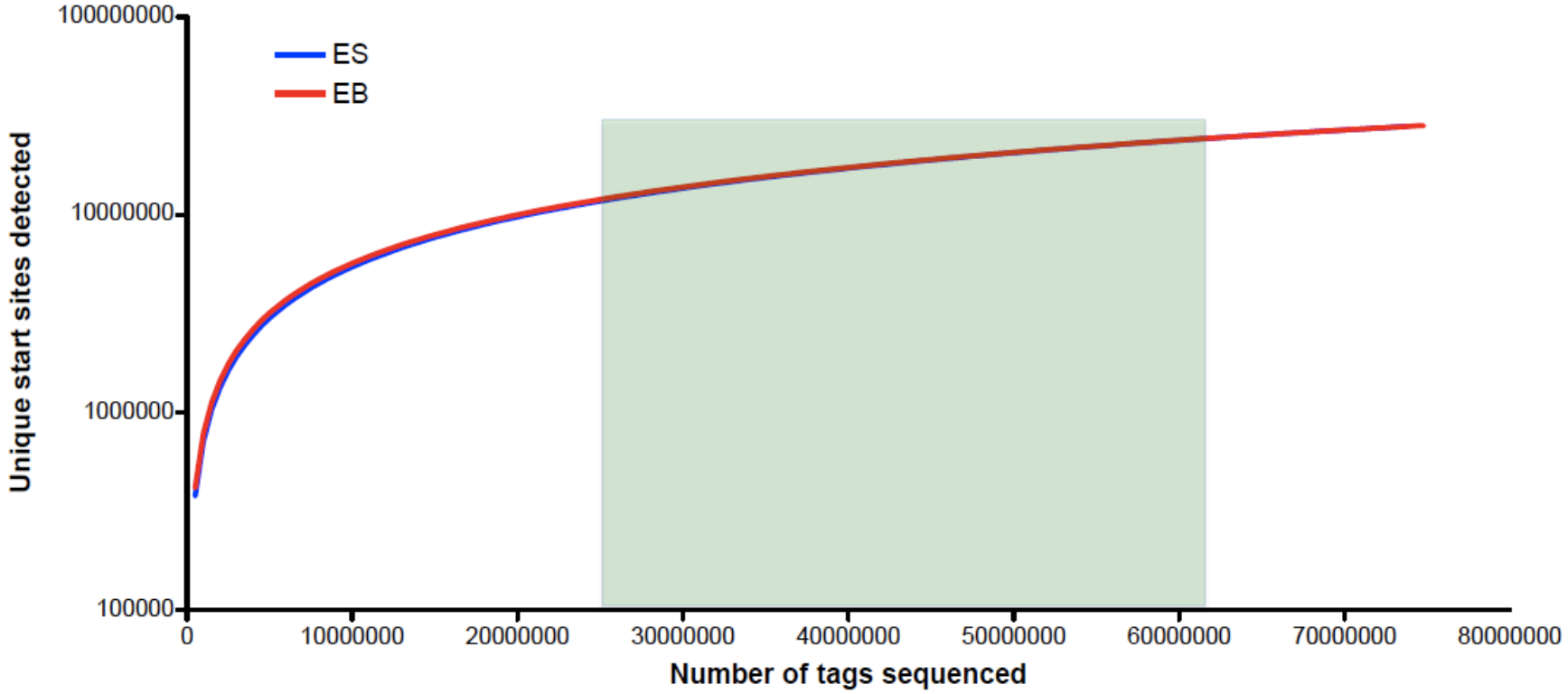
Step4) identification des jonctions



Reste 187 millions de séquences (60% des données) exploitables pour mesurer le niveau d'expression

Sample	Library	Number (volume) good quality tags	Number (Volume) tags mapping genome	Number (Volume) tags mapping junctions
ES	ES-2	32233538 (1.00 Gb)	17188745 (0.52 Gb)	467194 (0.02 Gb)
	ES-3	76779322 (2.45 Gb)	56799108 (1.75 Gb)	1370413 (0.05 Gb)
	ES-5	46597803 (1.46 Gb)	22314735 (0.66 Gb)	603597 (0.02Gb)
	ES all	155610663 (4.91 Gb)	96334528 (2.93 Gb)	2441204 (0.08 Gb)
EB	EB-1	95893107 (2.98 Gb)	57876542 (1.76 Gb)	2407871 (0.08 Gb)
	EB-2	12783982 (0.39 Gb)	6807410 (0.20 Gb)	205224 (0.01 Gb)
	EB-5	51007013 (1.60 Gb)	26466027 (0.80 Gb)	767657 (0.02 Gb)
	EB all	159684102 (4.98 Gb)	91202673 (2.77 Gb)	3380752 (0.11 Gb)

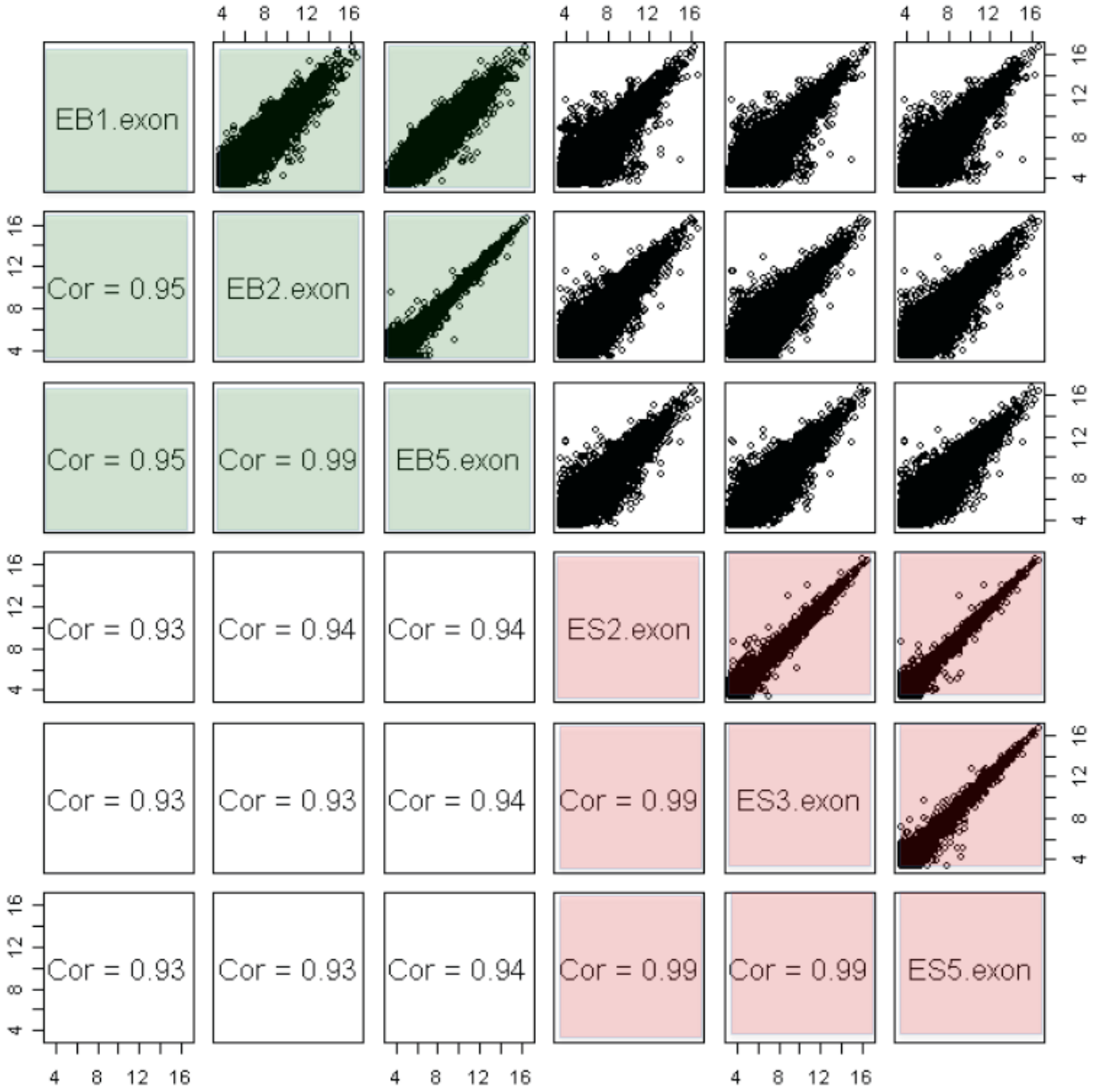
25 à 60 millions de « tag » utilisables par échantillons: Une couverture complète du génome.



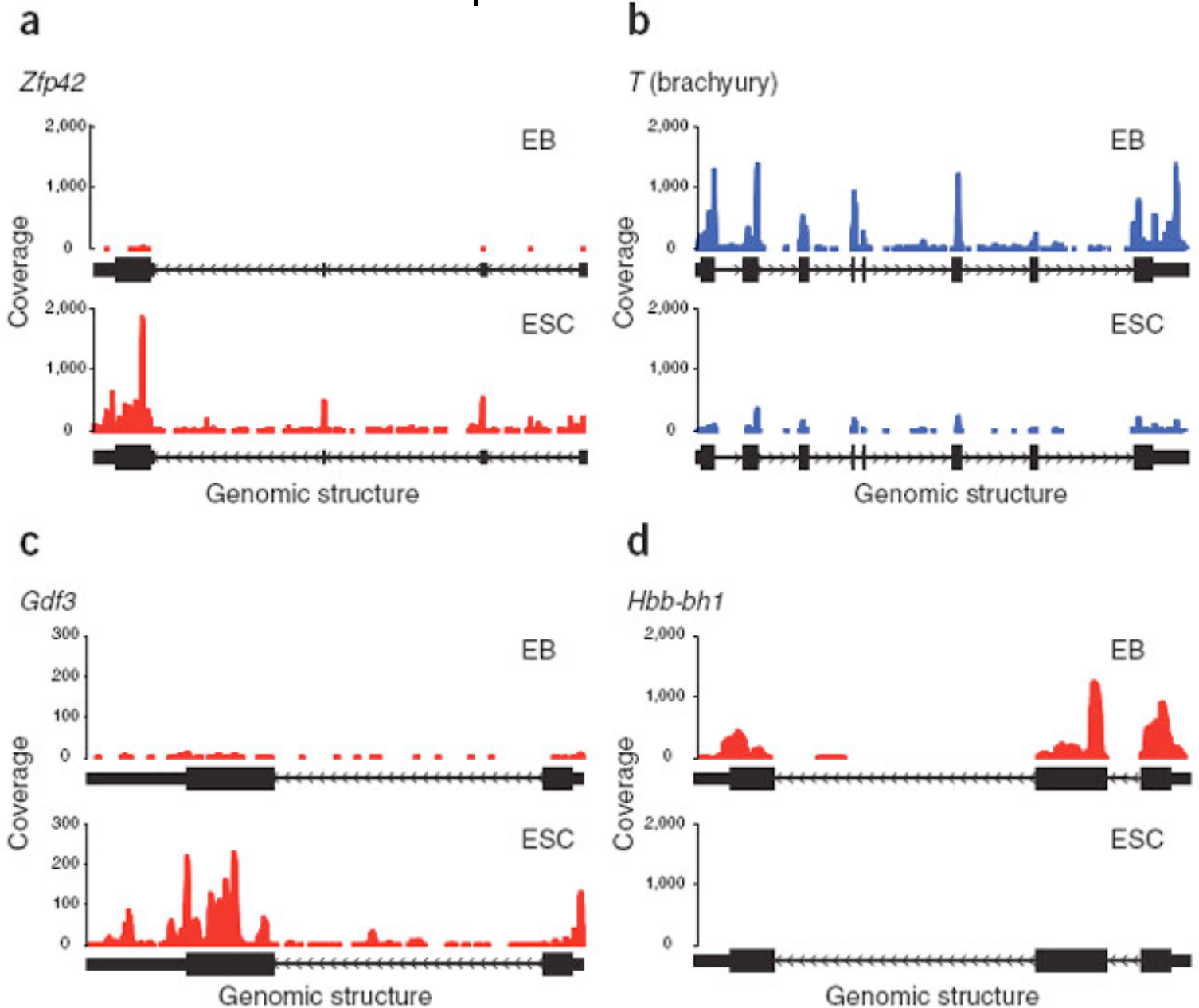
Analyse de la variabilité intraclasse et interclasse

Cellules Différenciées

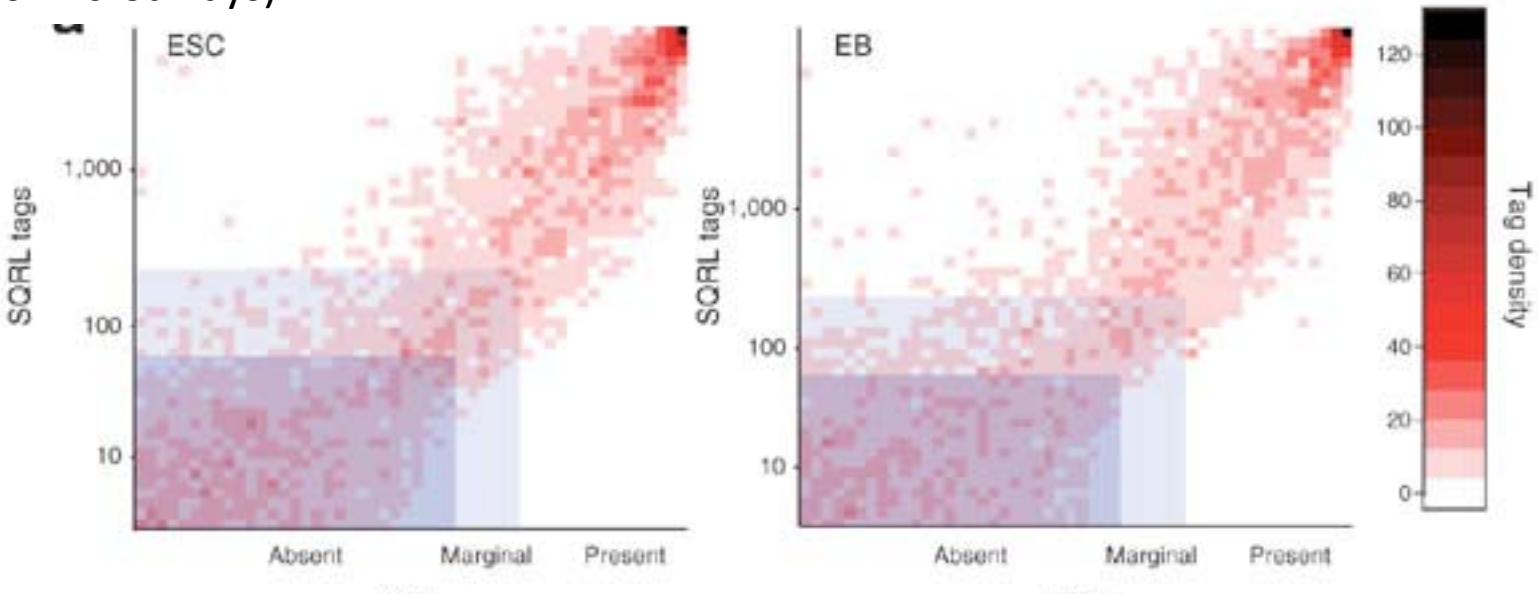
Cellules Non Différenciées



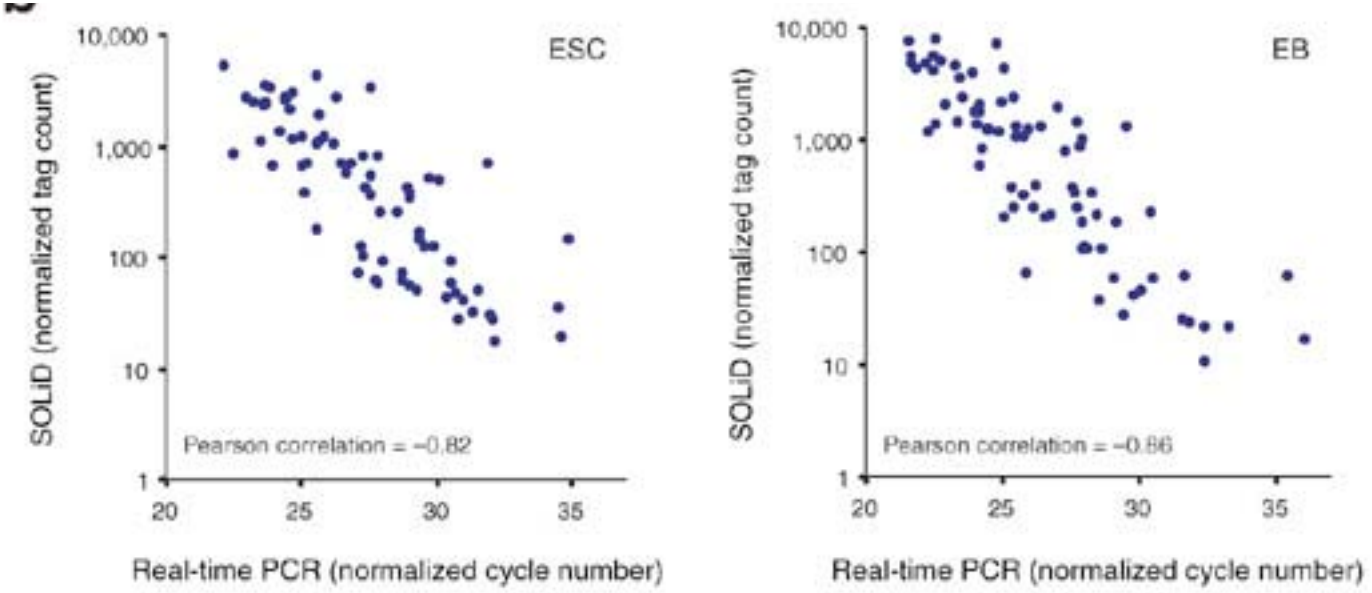
1136 gènes avec une niveau d'expression significativement modifié par la différenciation



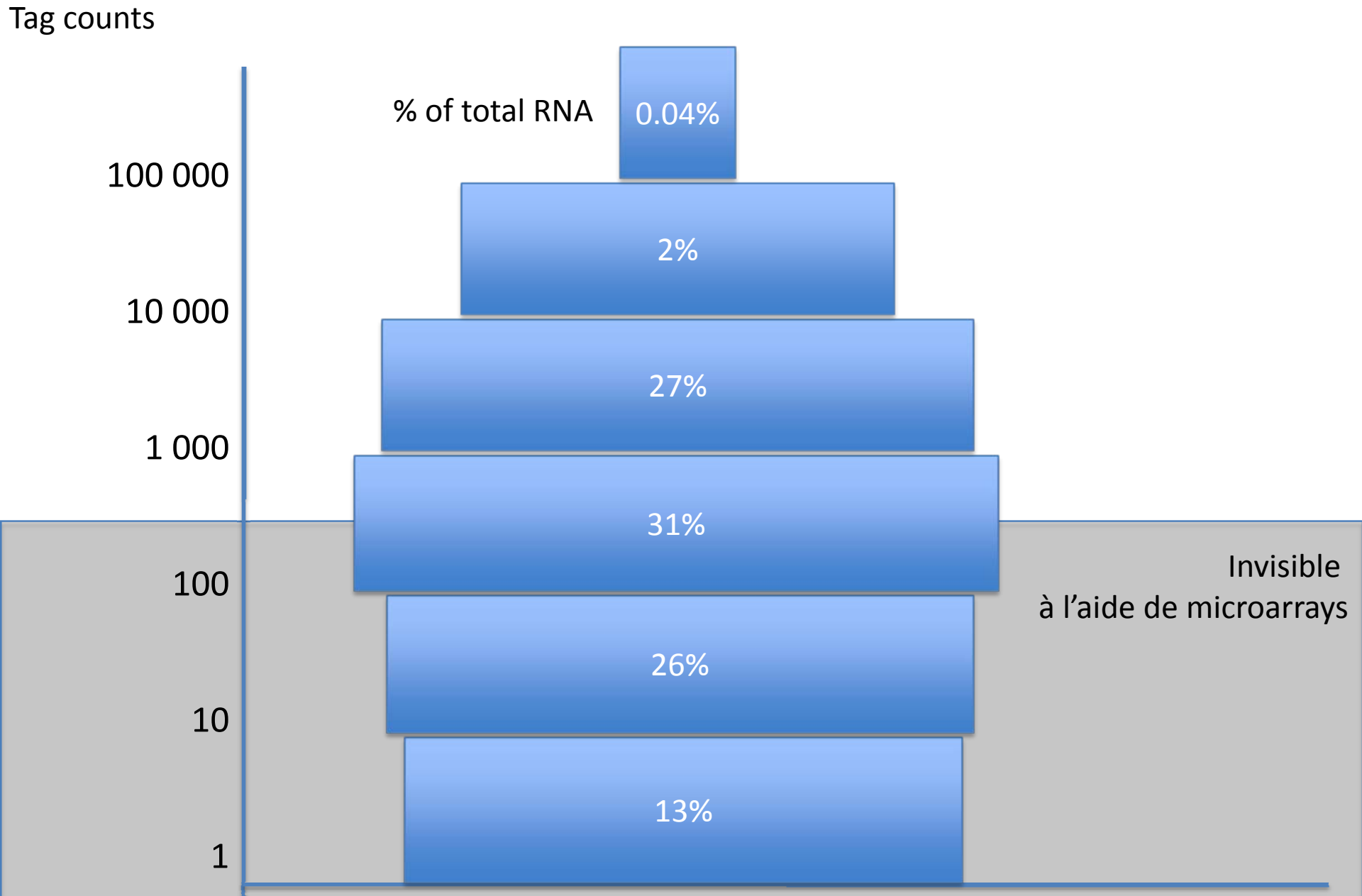
Plus sensible que les microarrays. Seuil 50 tags: 11436 genes détectés
(7000 pour les microarrays)



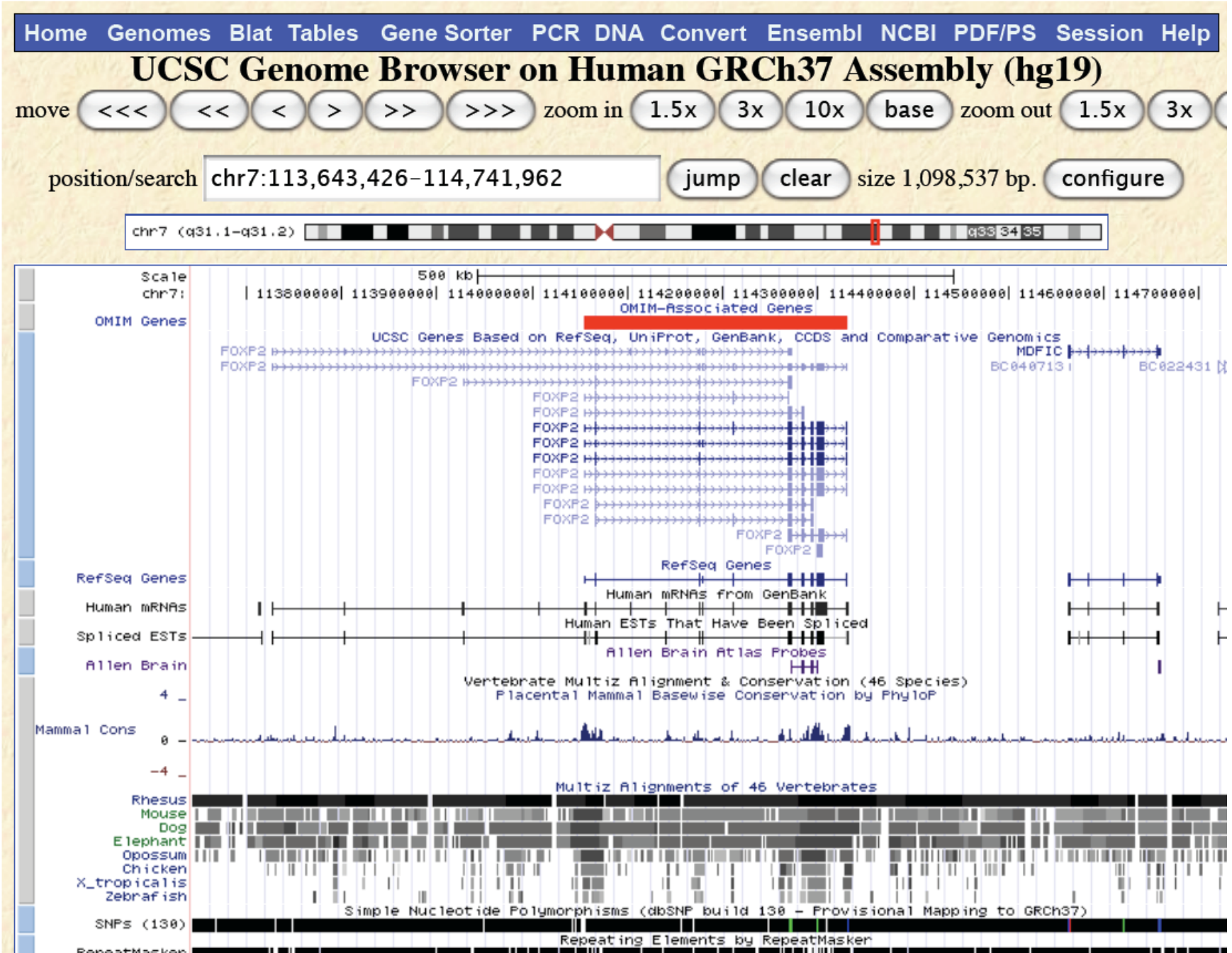
Corrélé aux résultats de Q-RT-PCR meme quand le niveau d'expression est faible.



RNA seq: pour voir la partie immergée de l'iceberg

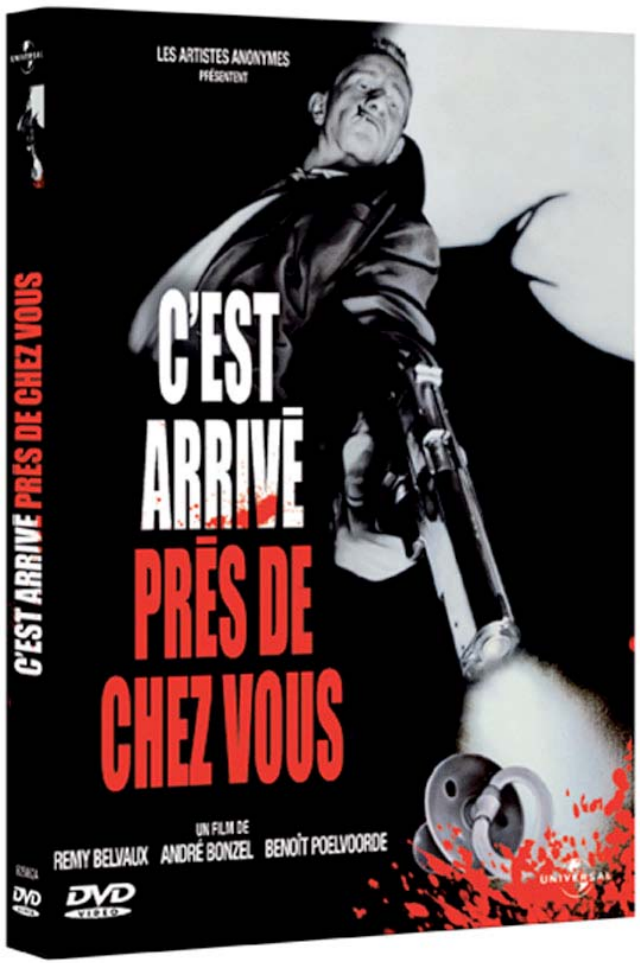


Des bases de données qui s'enrichissent rapidement



Les prix à la baisse

- Reséquencage du génome humain: 8000 €
- ChipSeq pour un facteur de transcription: 1500€
- Génome bactérien: 1000 €
- RNAseq tissu humain: 1000 €
- Digital gene expression: 400 €
- Et des logiciels libres pour analyser les données



Roche 454 Junior
Plateau Palgene ENS Lyon
Catherine Hänni
10^5reads/run
500 bases/read

Hiseq2000 Solexa
ProfilExpert IFR neurosciences
JoelLachuer
10^9reads/run
30-100 bases/read