Séquençage et analyse haut débit

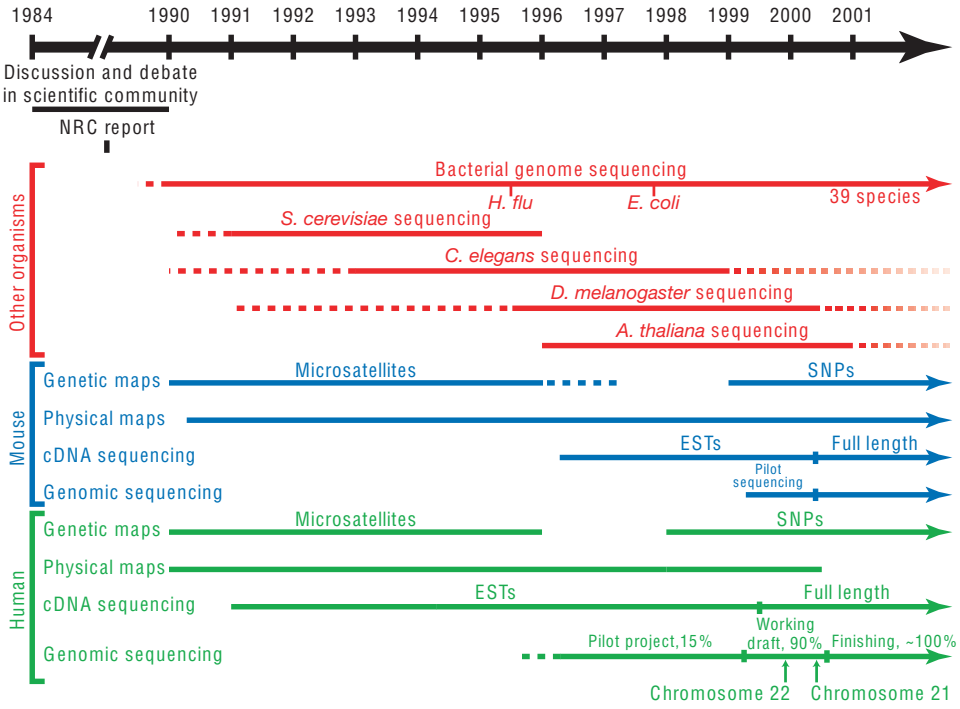
Génome humaine — 2001

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

860

NATURE | VOL 409 | 15 FEBRUARY 2001 | www.nature.com



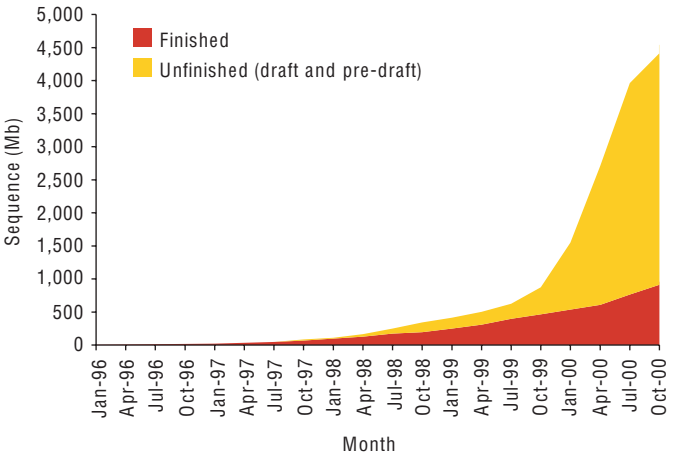


Table 2 Total genome sequence from the collection of sequenced clones, by sequence status

Sequence status	Number of clones	Total clone length (Mb)	Average number of sequence reads per kb*	Average sequence depth†	Total amount of raw sequence (Mb)
Finished	8,277	897	20–25	8–12	9,085
Draft	18,969	3,097	12	4.5	13,395
Predraft	2,052	267	6	2.5	667
Total					23,147

Début de la siècle

Mardis *Nature* 470 :198 (2011)

Nouvelles technologies

Rob Carlson, synthesis.cc (avril 2013)

%

Nouvelles technologies

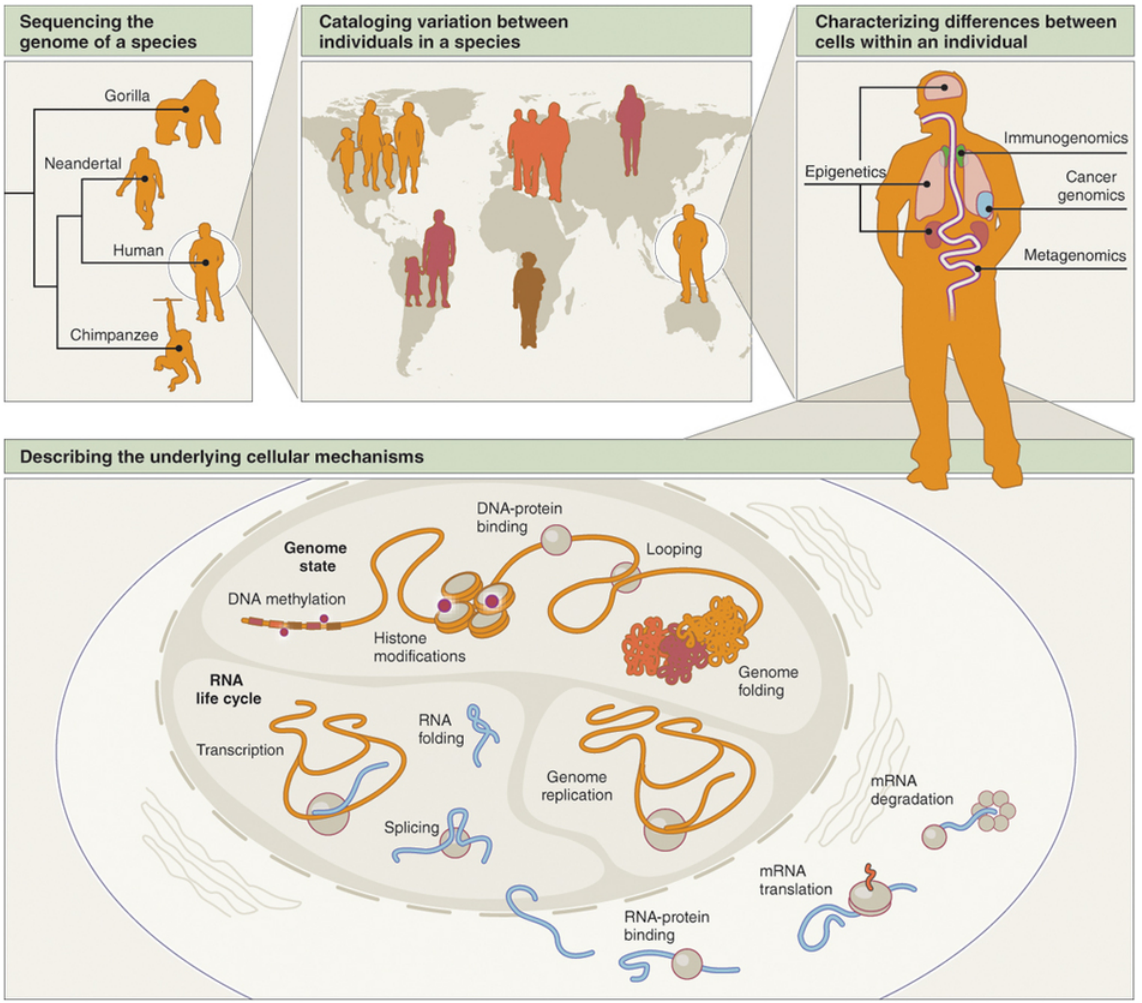
Table 2 Next-generation DNA sequencing instruments

	Cost per base[a]	Read length (bp)[b]	Speed	Capital cost[c]
Minimum cost per base				
Complete Genomics	Low	Short	3 months	None (service)
HiSeq 2000 (Illumina)	Low	Mid	8 days	+++++++
SOLiD 5500xl (Life Technologies)	Low	Short	8 days	+++
Maximum read length				
454 GS FLX+ (Roche)	High	Long	1 day	+++++
RS (Pacific Biosciences)	High	Very long	<1 day	+++++++
Maximum speed, minimum capital cost and minimum footprint				
454 GS Junior (Roche)	High	Mid	<1 day	+
Ion Torrent PGM (Life Technologies)	Mid	Mid	<1 day	+
MiSeq (Illumina)	Mid	Long	1 day	+
Combined prioritization of speed and throughput				
Ion Torrent Proton (Life Technologies)	Low	Mid	<1 day	++
HiSeq 2500 (Illumina)	Low	Mid	2 days	++++++++

[a]'Low' is < $0.10 per megabase, 'mid' is in-between and 'high' is > $1 per megabase. [b]'Short' is < 200 bp, 'mid' is 200–400 bp, 'long' is > 400 bp and 'very long' is > 1,000 bp. [c]Each "+" corresponds to ~$100,000. We list only commercialized instruments that can be purchased and for which performance data are publically available (as opposed to a comprehensive list of companies developing next-generation sequencing technologies). The categorizations refer to the aspect of sequencing performance to which the technology and/or its implementation in a specific instrument are primarily geared. These estimates were made at the time of publication, and the pace at which the field is moving makes it likely that they will be quickly outdated.

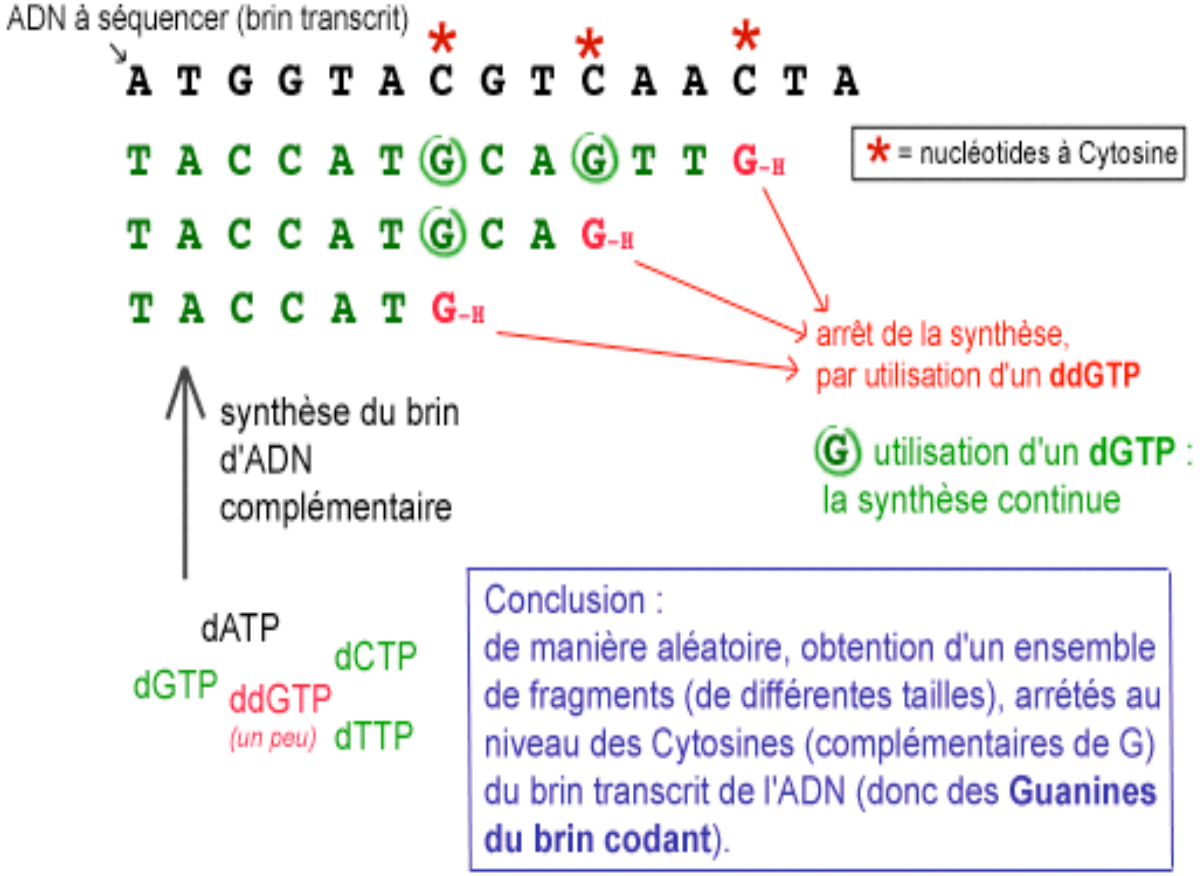
Shendure & Aiden *Nat Biotechnol* 30 :1084 (2012)

Nouvelles applications



Shendure & Aiden *Nat Biotechnol* 30 :1084 (2012)

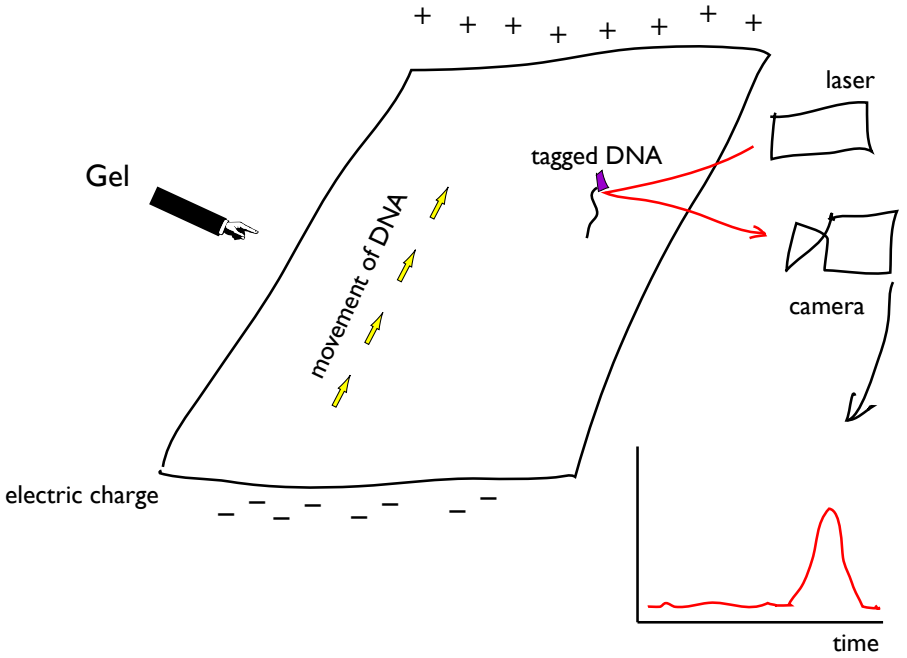
Séquençage — méthode Sanger



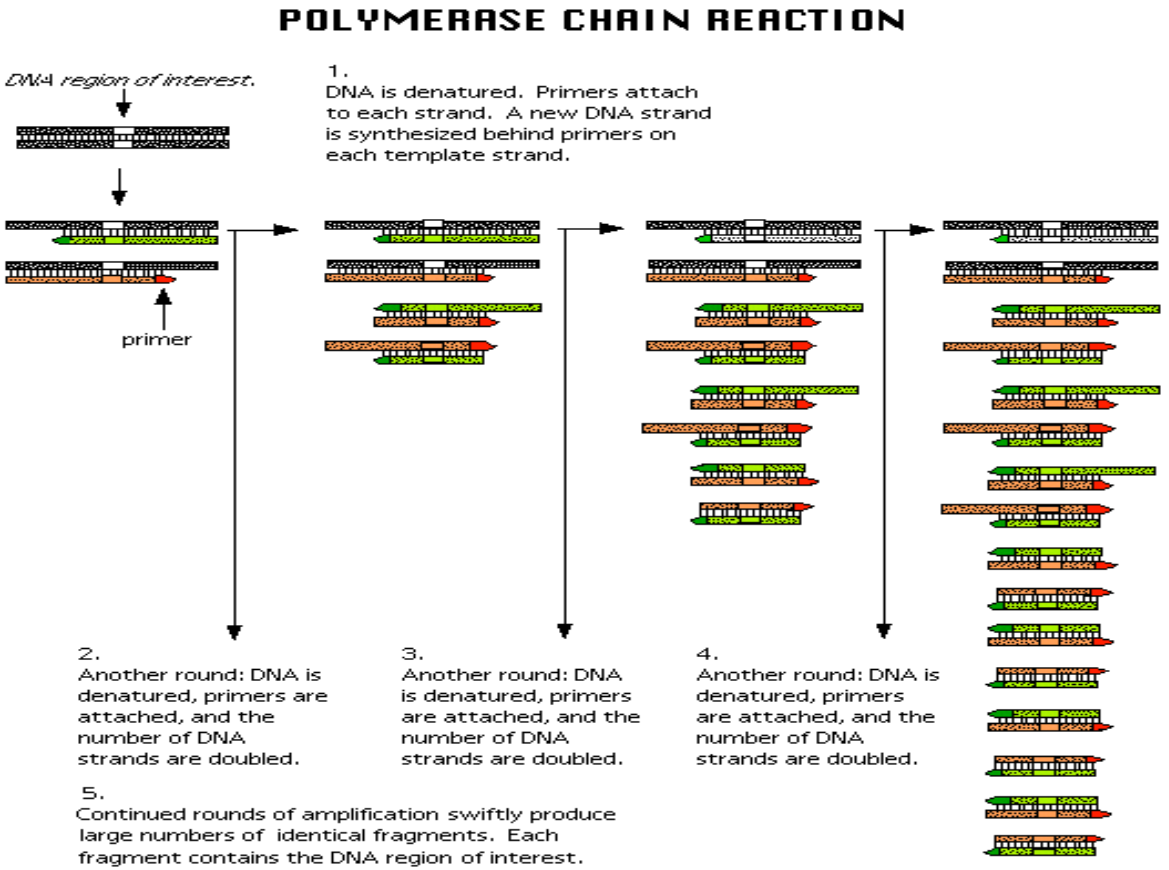
Lodish et al. *Molecular Cell Biology*, 4th ed., W. H. Freeman 1999 ; Delarue et Furelaud, Jussieu

Electrophorèse

mesure la taille d'un fragment ADN



Polymerase Chain Reaction

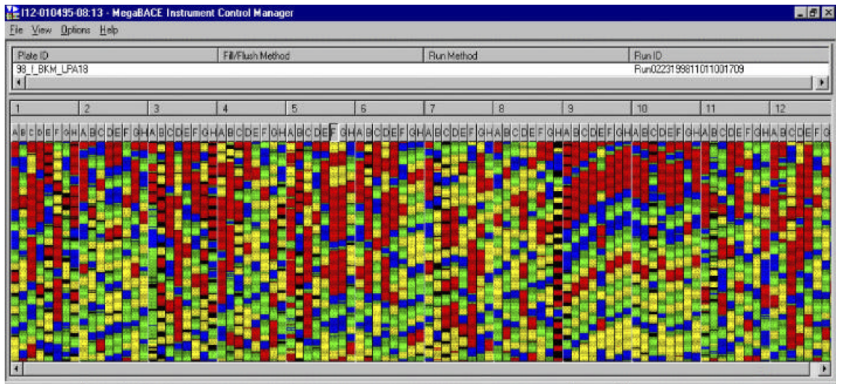


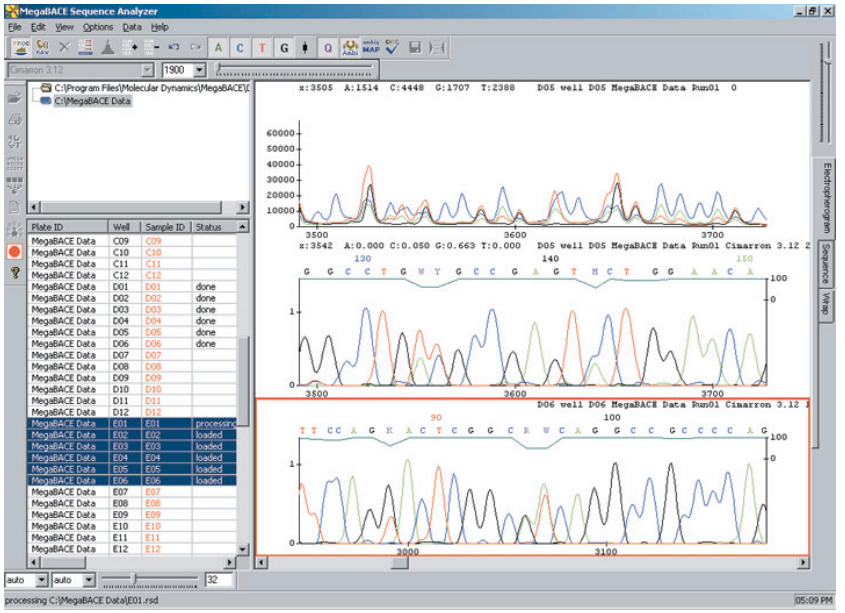
Access Excellence

Séquençage automatique

→ marquers fluorescents, multiplexage, capillaires, cycleurs thermiques

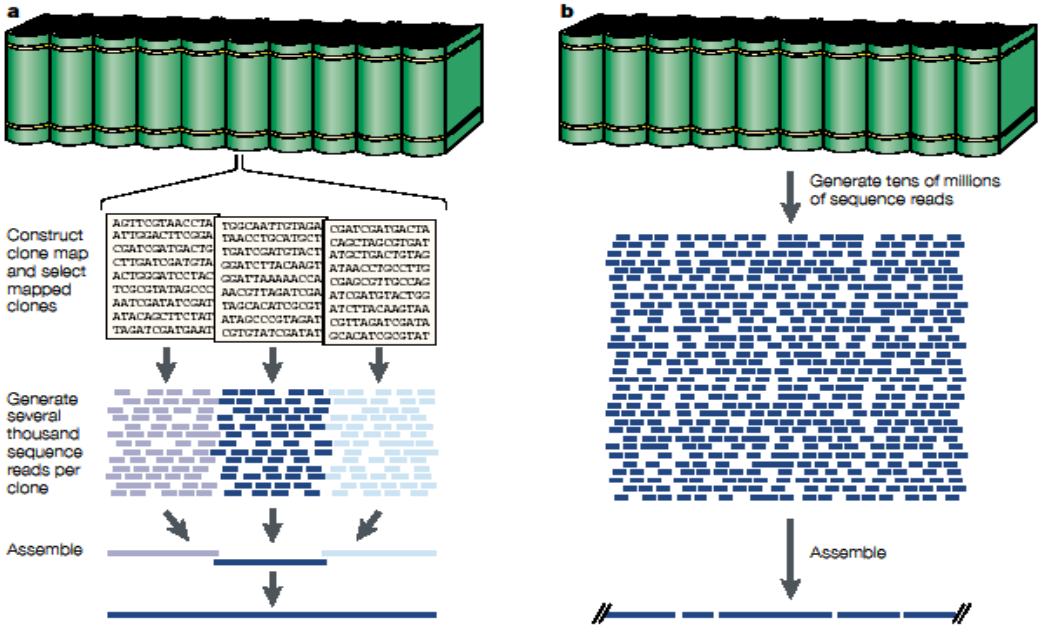
600–1000 pb, 96–384 capillaires, quelques heures





Molecular Dynamics → Amersham Biosciences → GE Life (Megabace)

Séquençage du génome humaine

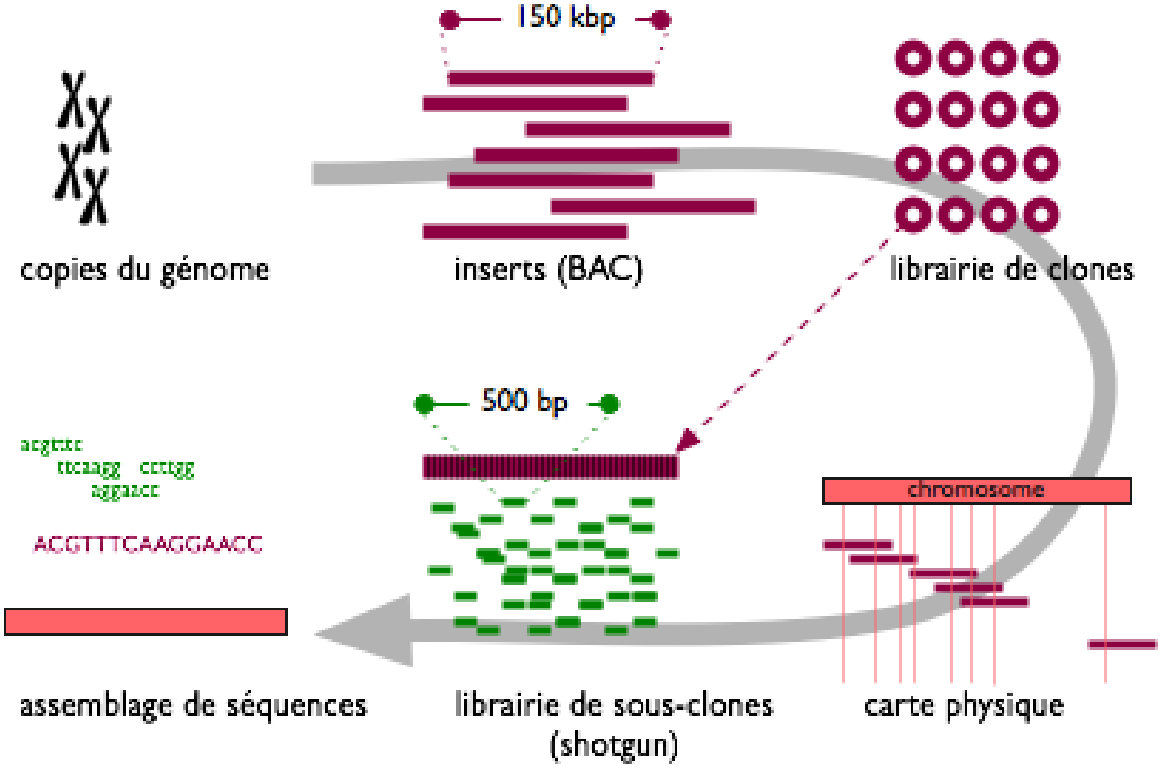


Clone-by-clone

Whole-genome shotgun

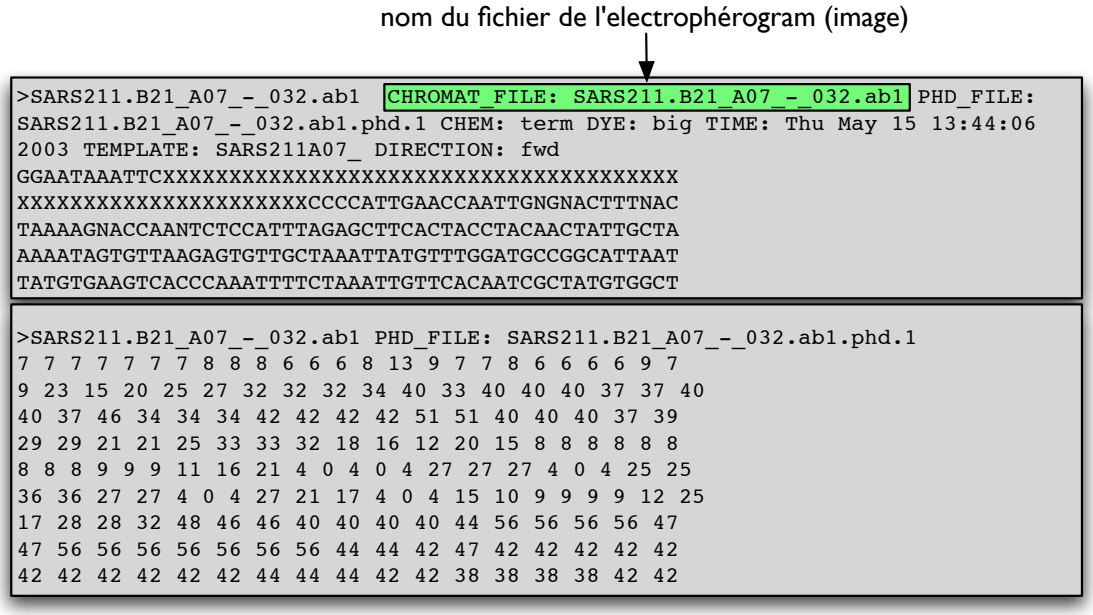
E. Green. Nature Reviews Genetics 2:573 (2001)

Approche hiérarchique



Phred

deux fichiers : séquences (fasta) et valeurs de qualité (`.qual`)



probabilité d'erreur pour qualité q : $10^{-q/10}$

Ewing & Green. *Genome Res* 8 :186.

FASTQ

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- ⋆ Ligne $4k$ commence par '@' et introduit l'identificateur du morceau (*read*)
- ⋆ Ligne $4k + 1$ donne la séquence
- ⋆ Ligne $4k + 2$ commence par '+' et *peut* répéter l'identificateur de Ligne $4k$ ou non
- ⋆ Ligne $4k + 3$ donne les valeurs de qualité (encodage Sanger avec échelle Phred $q + 33$: $0 \rightarrow$!, $1 \rightarrow$ ", ..., $31 \rightarrow$ @, $32 \rightarrow$ A, ..., $40 \rightarrow$ I)

Méthodes de *next generation*

Principes :

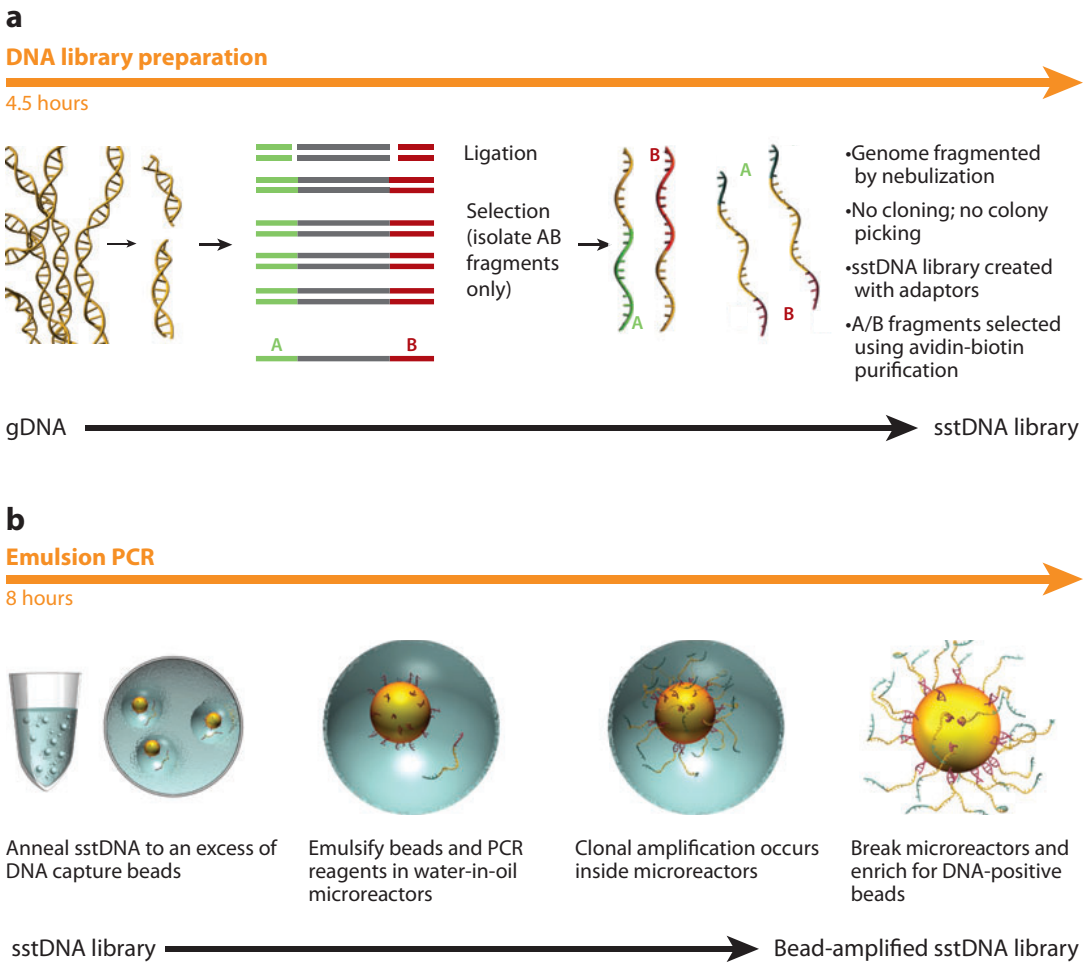
1. fixer fragments d'ADN ciblés sur une surface → arrangement en espace pour parallélisation («puce»)

amplification (si nécessaire)

2. séquençage de fragments d'ADN + production d'image

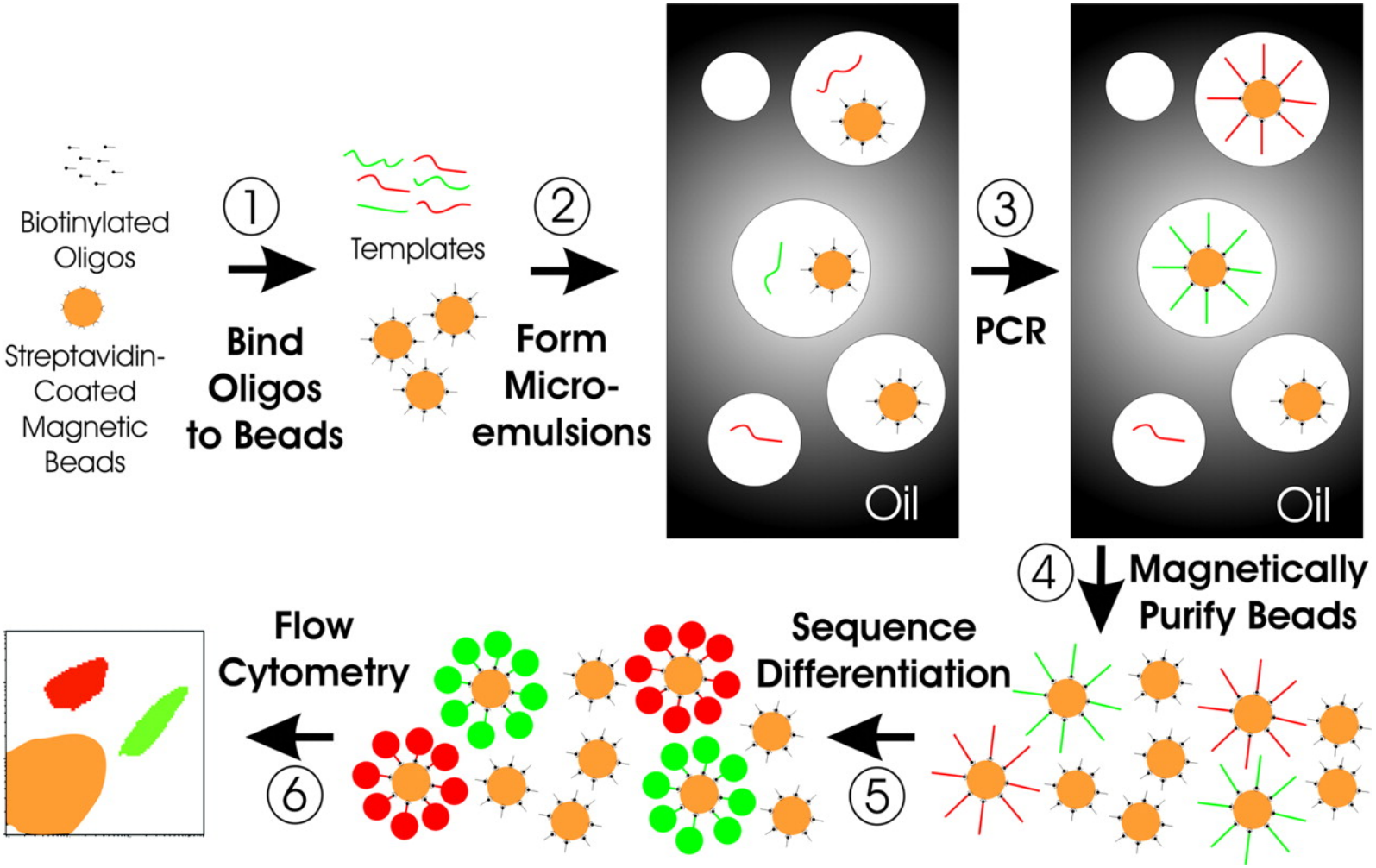
3. analyse informatique de l'image : production de séquences + estimation d'erreurs de séquençage

454 : perles (*beads*)



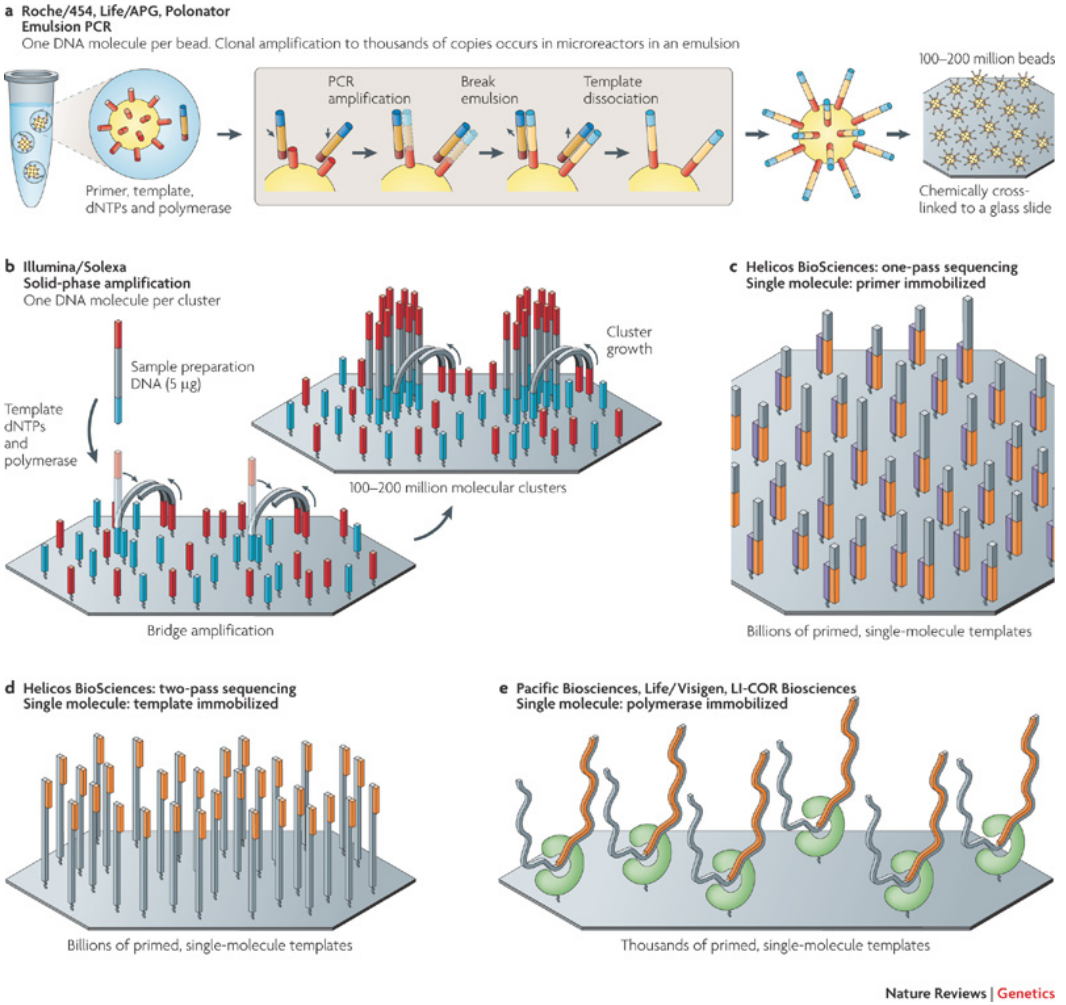
Mardis *Annu Rev Genomics Hum Genet* 9 :387 (2008)

PCR en emulsion



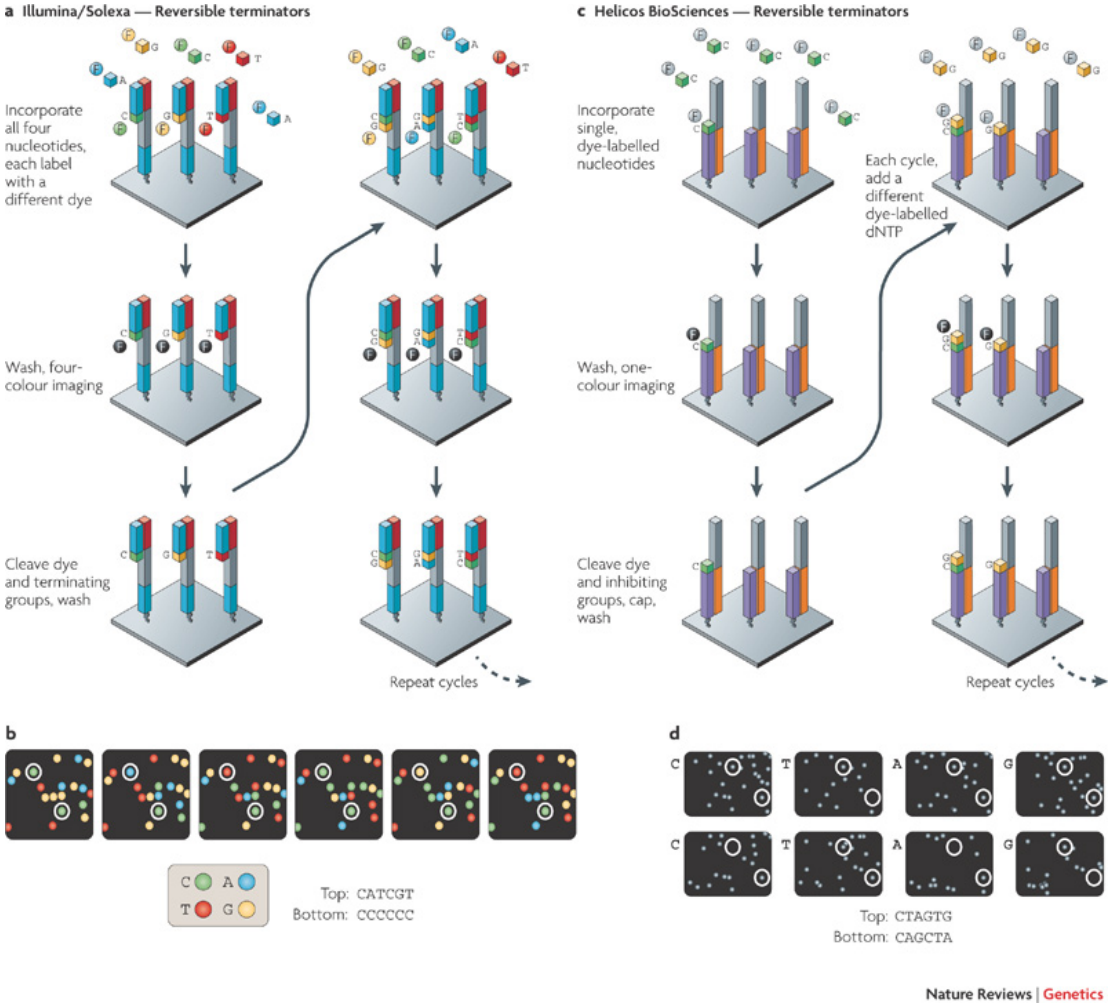
Dressman & al *PNAS* 100 :8817 (2003) ; → YouTube Ion Torrent

Immobilisation du cible



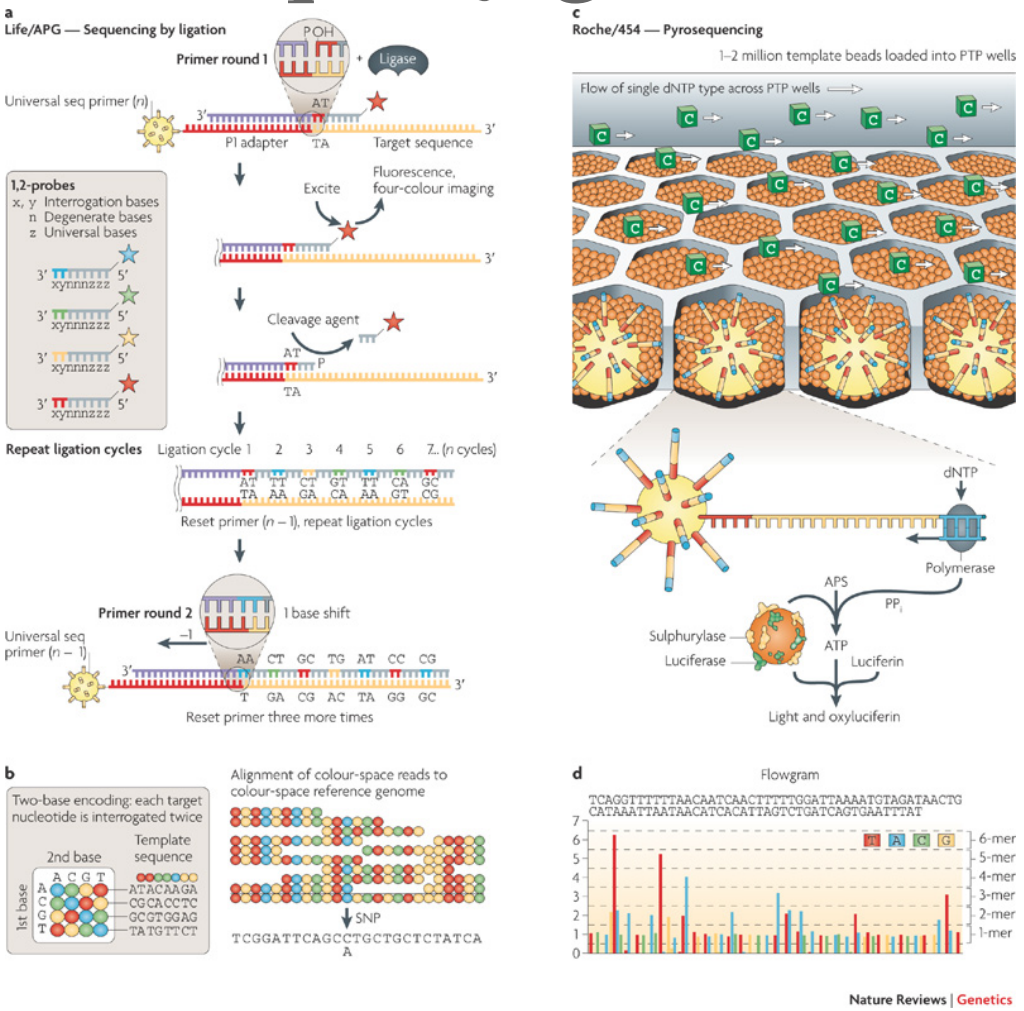
Metzker *Nat Rev Genet* 11 :31 (2010)

Termination cyclique (Illumina)



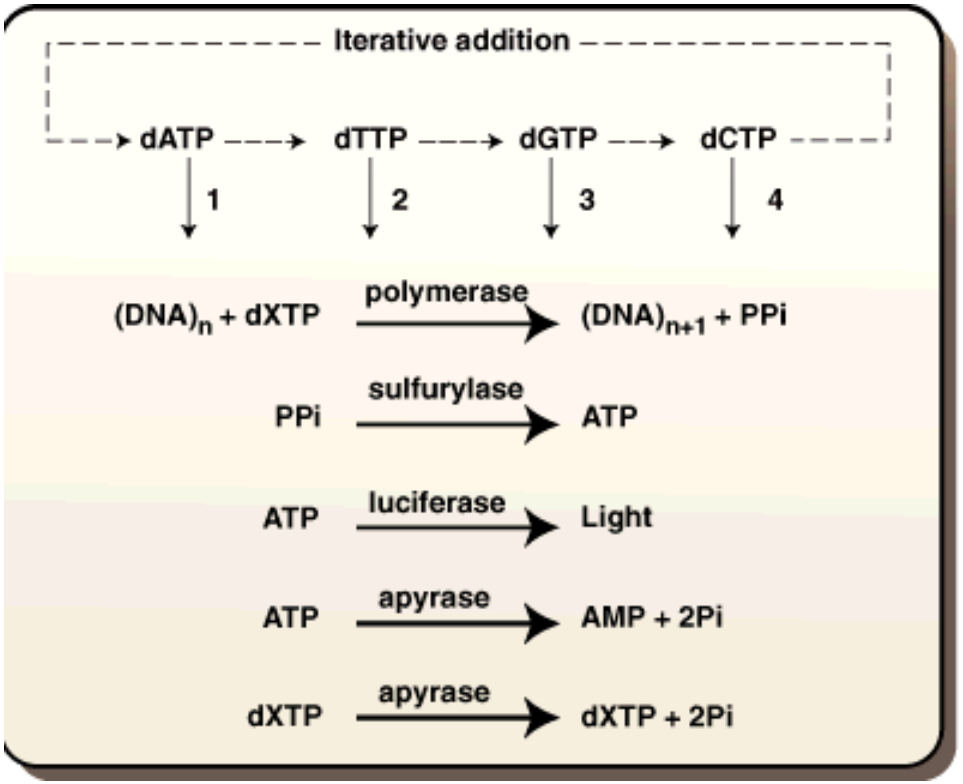
Metzker *Nat Rev Genet* 11 :31 (2010) ; → YouTube Illumina

Méthodes de séquençage



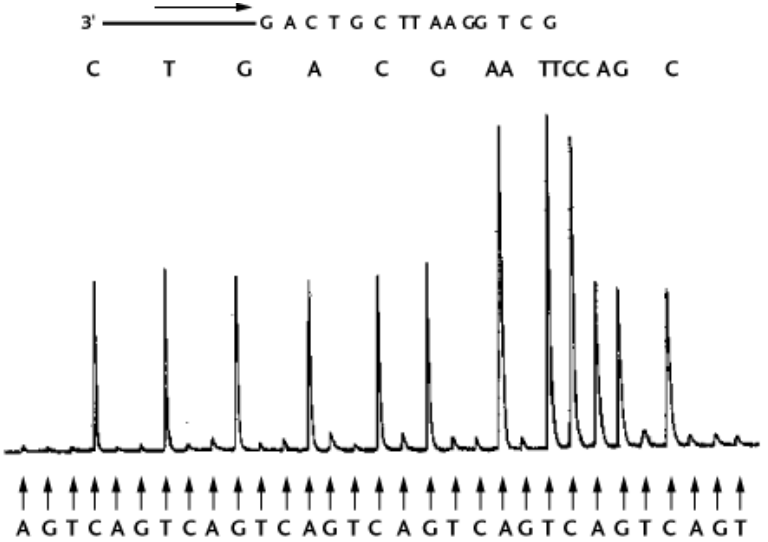
Metzker *Nat Rev Genet* 11 :31 (2010)

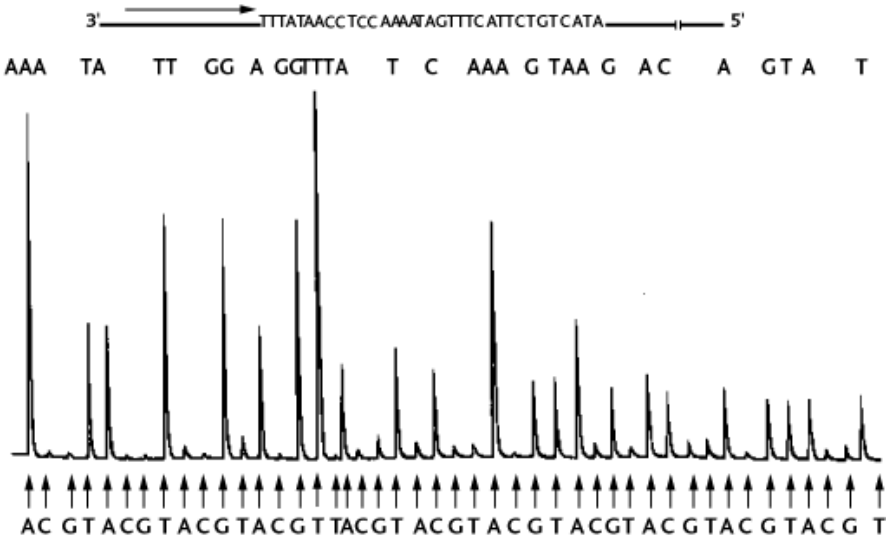
Pyroséquençage — principle



Ronaghi et al. *Science* 281 :363.

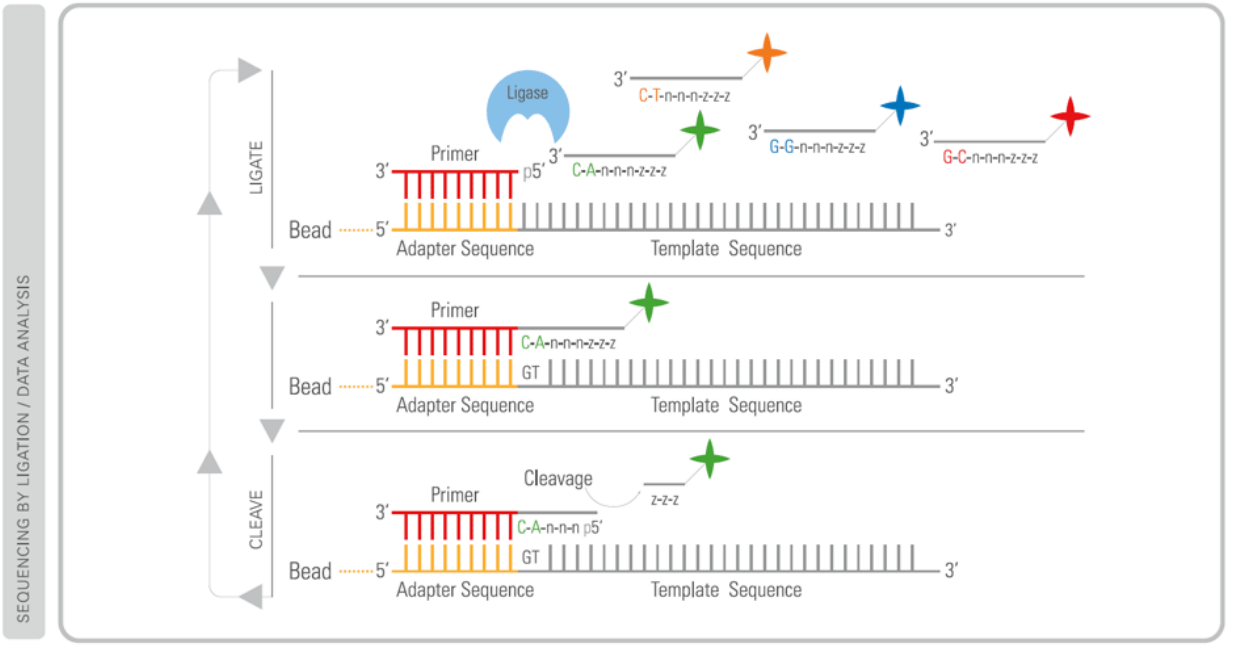
Pyrogram





Ronaghi et al. *Science* 281 :363.

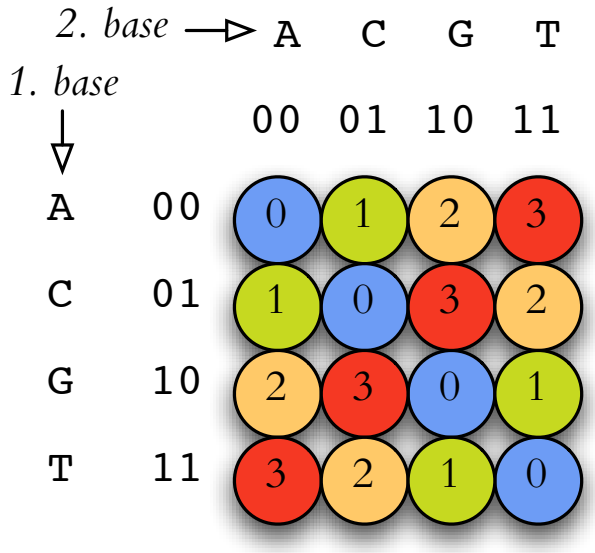
Solid : sequencing by ligation



Applied Biosystems

Solid : encodage dans l'espace de couleurs

Séquence $t[1..m+1]$ encodée par $t \cdot c[1..m]$, où $t = t[1]$, et $c[i] = t[i] \oplus t[i+1]$. L'$\oplus$ dénote l'OR exclusive dans l'encodage à 2 bits.



index	*dye*	*color*	*bitcode*
0	FAM	blue	00
1	Cy3	green	01
2	TXR	orange	10
3	Cy5	red	11

2. base → A C G T

00	01	10	11
A	C	G	T
C	A	T	G
G	T	A	C
T	G	C	A