Le séquençage nouvelle génération (NGS) en génétique humaine

Solena Le Scouarnec
solena.lescouarnec@univ-nantes.fr

M2 Bioinformatique
Module Bioinformatique appliquée

23 septembre 2013

Objectifs

- **Partie 1**

Evolution des technologies de séquençage

Séquençage Illumina – obtention des données brutes

- **Partie 2 (Pierre Lindenbaum)**

Analyse bioinformatique

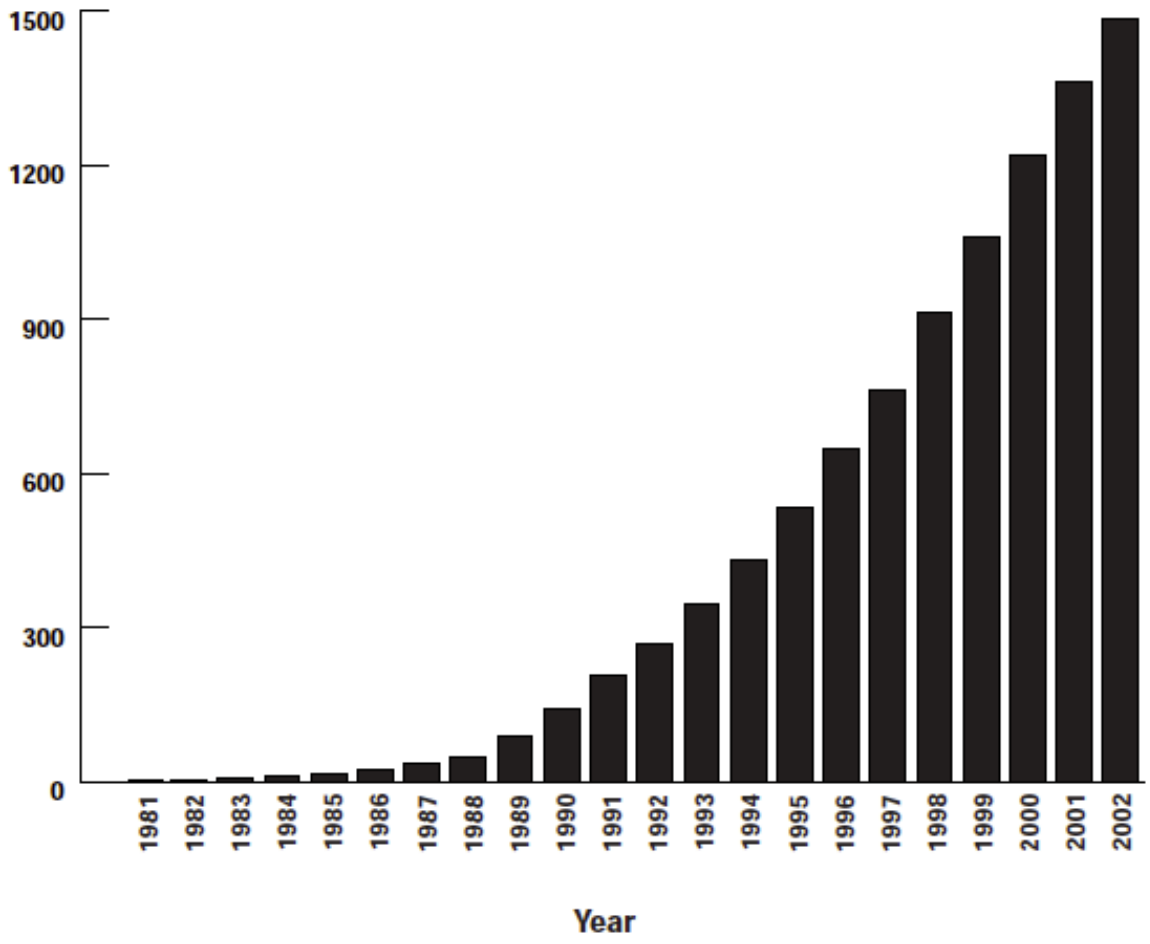
Format des fichiers

TP 9 & 16 octobre

Plan

- Avant le NGS (méthode Sanger)
- La révolution NGS
- Depuis le NGS
- Séquençage avec la plateforme Illumina
- Exemple

Maladies génétiques héréditaires



Cumulative Pace of Disease Gene Discovery (1981-2002). The number of disease genes identified so far is 1,485. Data provided by Online Mendelian Inheritance in Man.

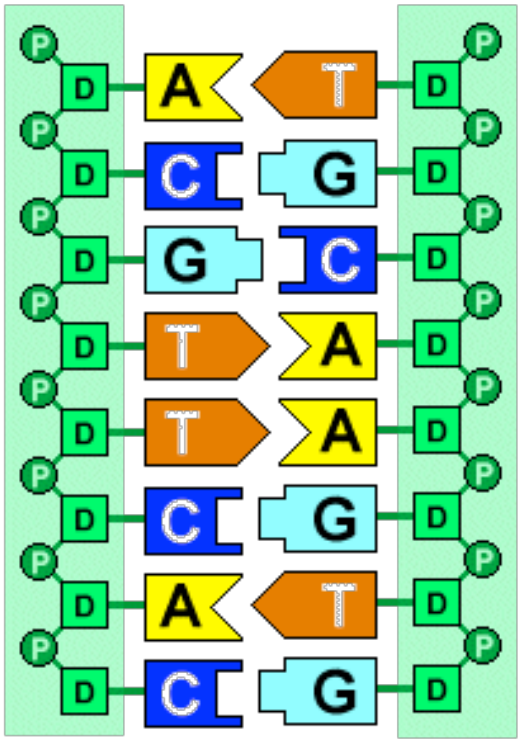
2002
1485 gènes identifiés

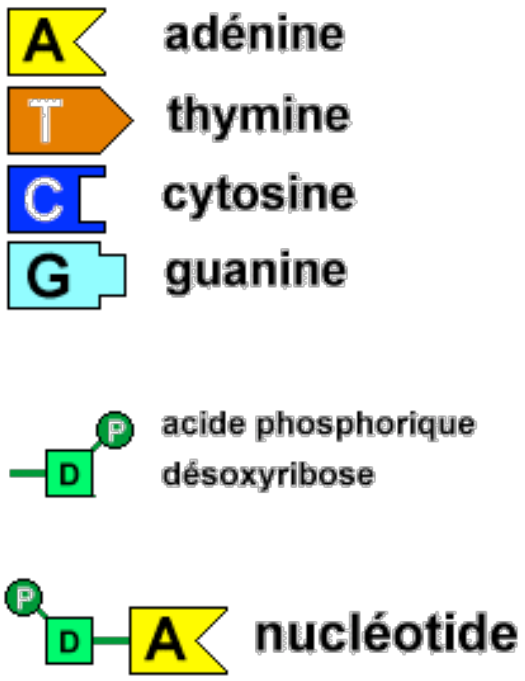
2013
3674 gènes identifiés

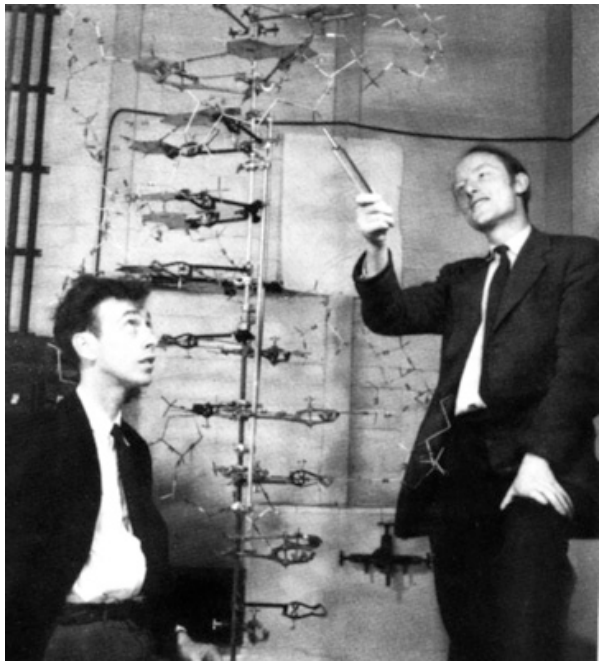
Base moléculaire inconnue pour >1000 maladies mendéliennes

http://www.nih.gov/news/pr/apr2003/Pace.pdf
http://www.ncbi.nlm.nih.gov/Omim/mimstats.html, January 7, 2013

L'ADN

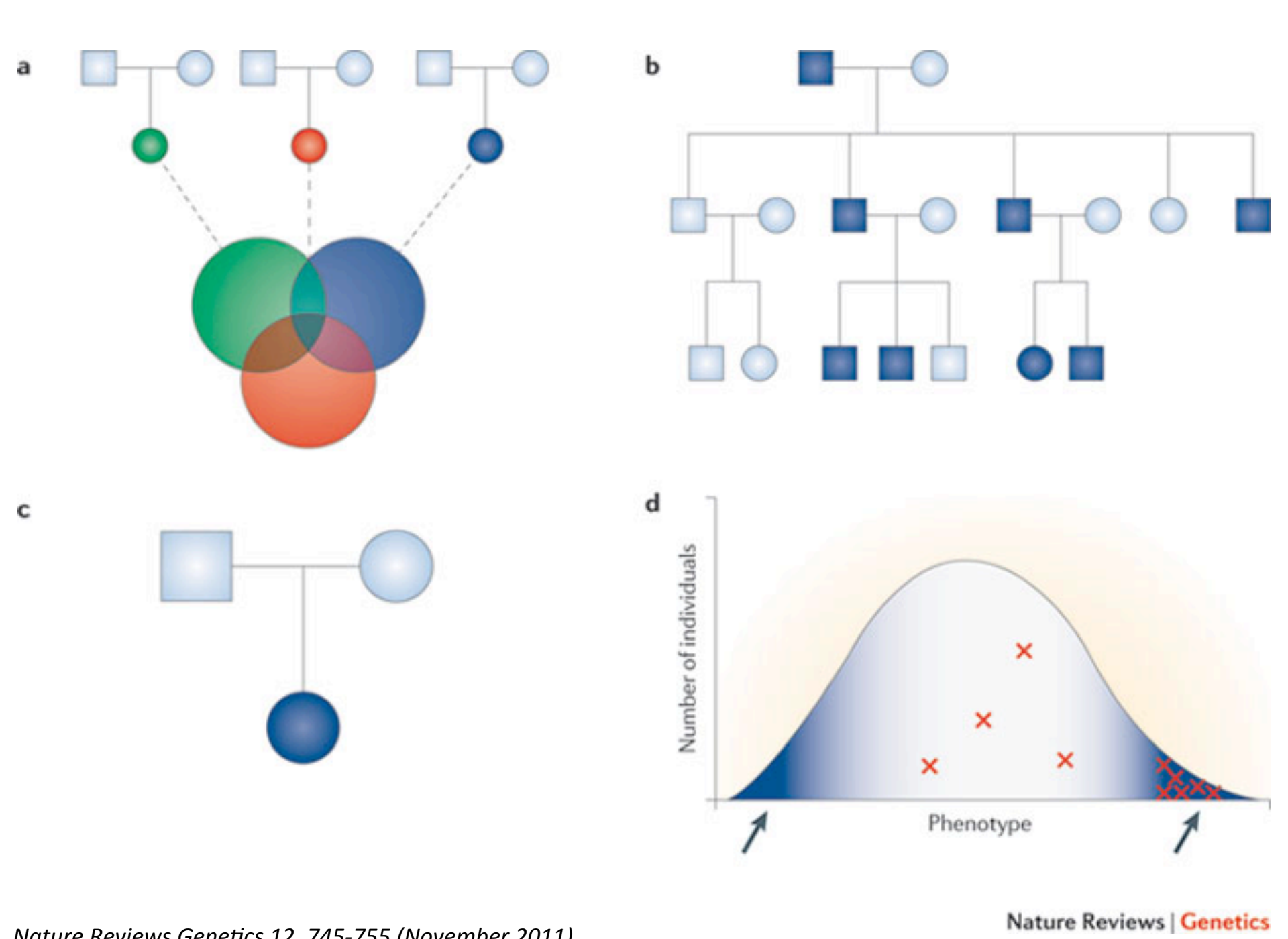






Séquençage d'ADN:
Déterminer l'ordre des nucléotides d'une molécule d'ADN

http://espace-svt.ac-rennes.fr/cartelec/cartelec_lyc/premiere_s/vegetal/adn/replic1.htm

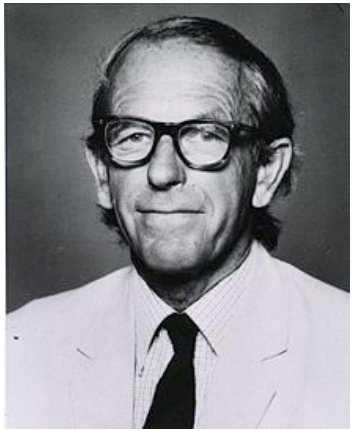


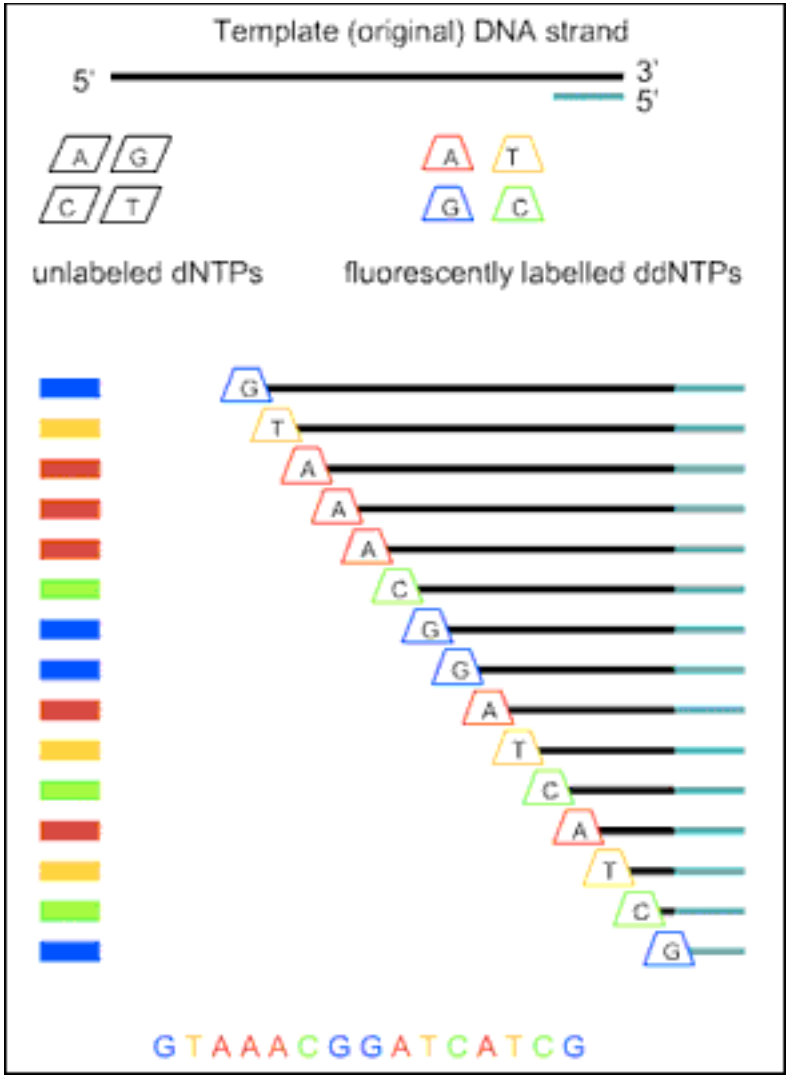
Nature Reviews Genetics 12, 745-755 (November 2011)

Plan

- Avant le NGS (méthode Sanger)
- La révolution NGS
- Depuis le NGS
- Séquençage avec la plateforme Illumina
- Exemple

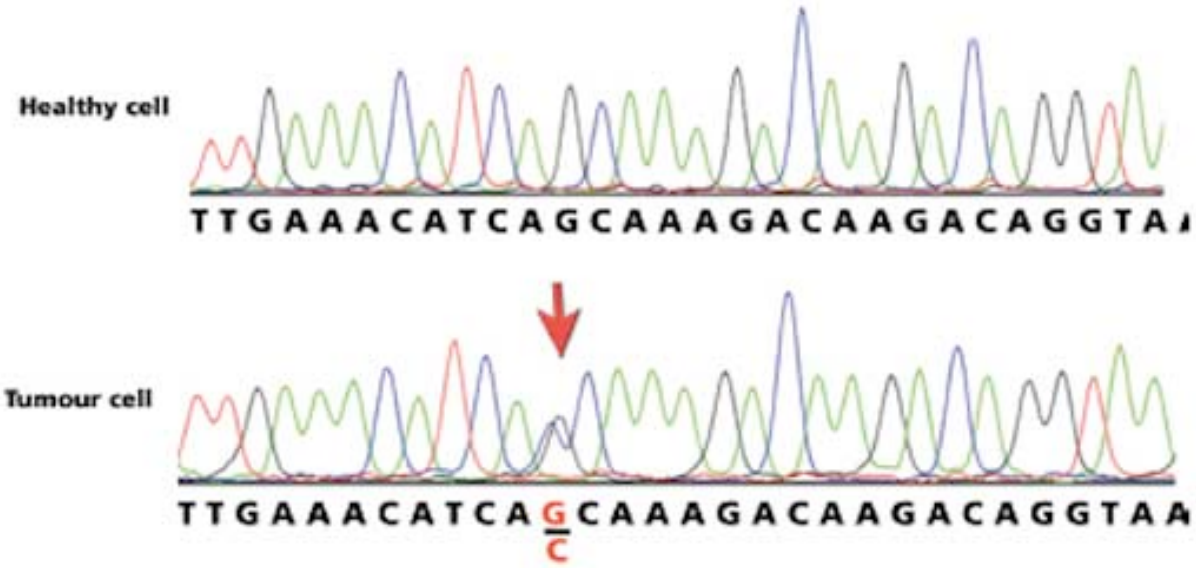
Séquençage Sanger
- Méthode -





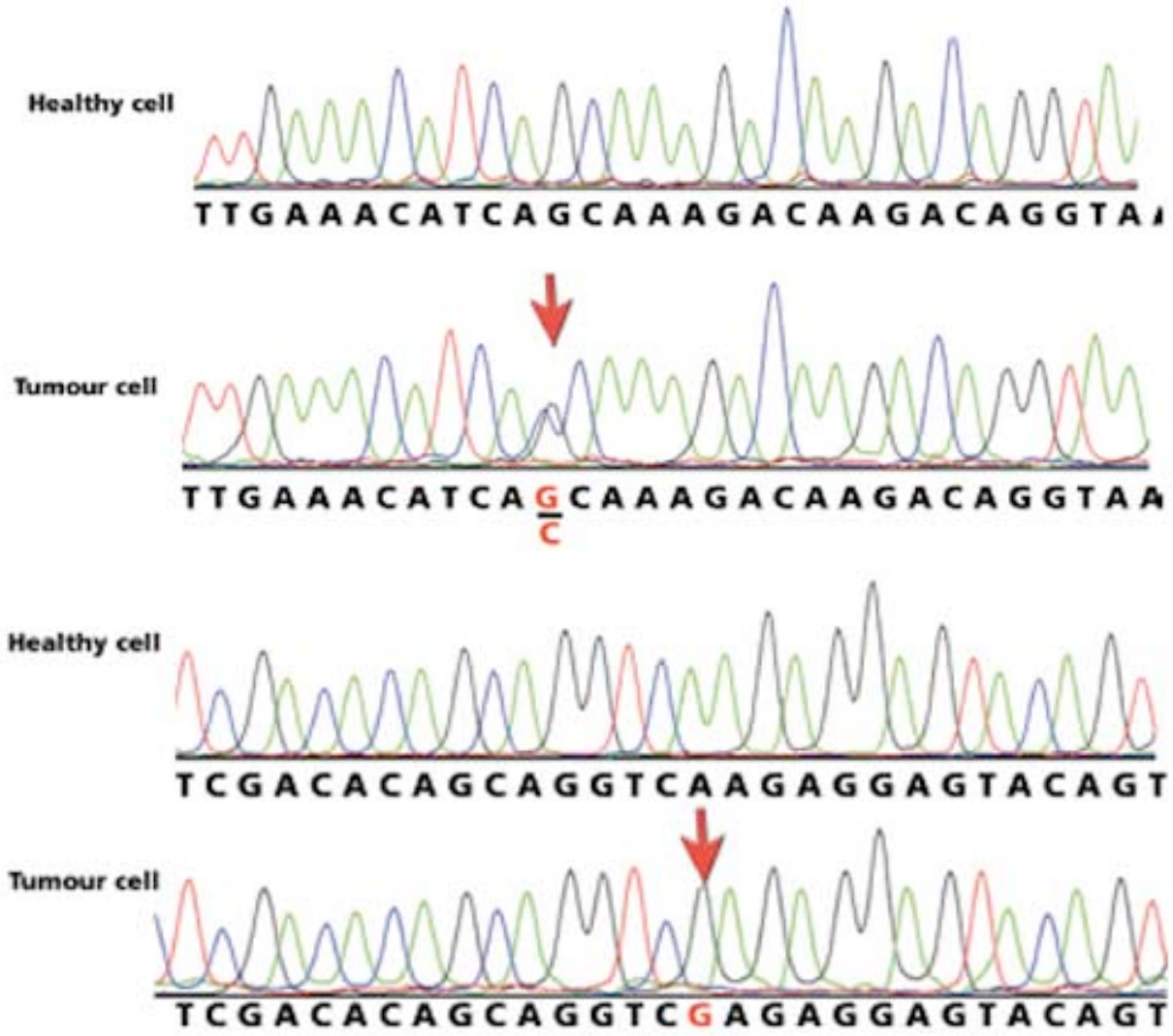
http://www.genomebc.ca/education/articles/sequencing/

Séquençage Sanger
- Analyse -



http://www.scienceinschool.org/2010/issue16/cancer/french

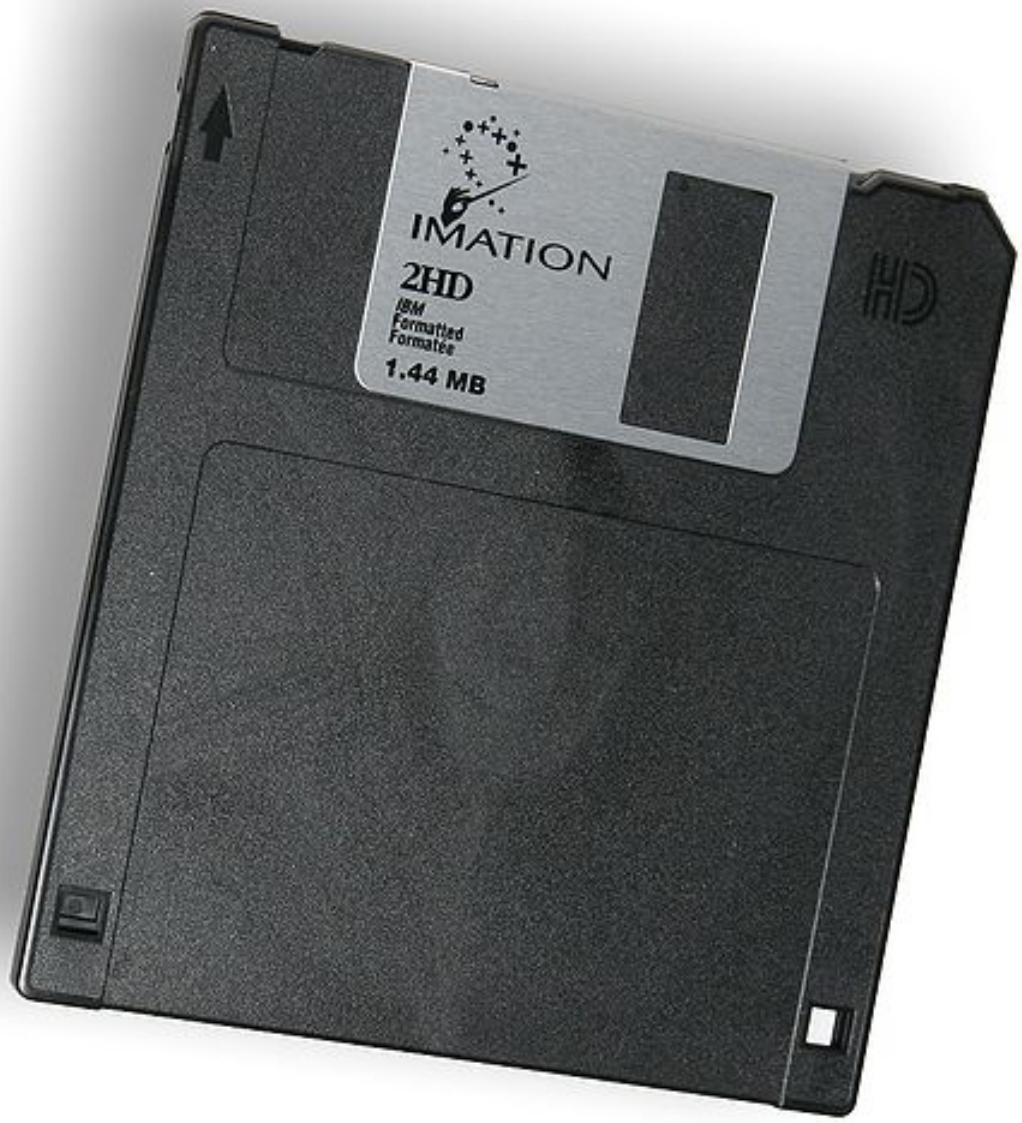
Séquençage Sanger
- Analyse -



Mutation hétérozygote

Mutation homozygote

http://www.scienceinschool.org/2010/issue16/cancer/french


IMATION
2HD
IBM
Formatted
Formatée
1.44 MB
HD

Séquençage Sanger

Avantages

- Reads longs (1000 pb)
- Efficace pour petits projets
- Peu d'erreurs

Limites

- Bas débit
- Coûteux

Projet génome humain

Version la plus récente: GRCh37

Plan

- Avant le NGS (méthode Sanger)
- La révolution NGS
- Depuis le NGS
- Séquençage avec la plateforme Illumina
- Exemple

Séquençage nouvelle génération (NGS)

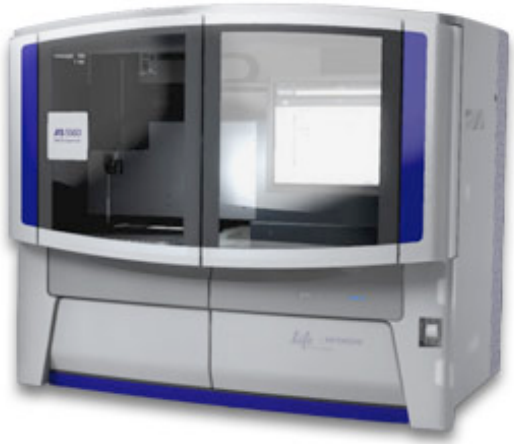
2005

Roche 454 SEQUENCING

Roche GS FLX
(« pyroséquençage »)

2006-2007

life technologies™



SOLiD 5500

2006-2007

illumina®

HiSeq

Séquençage Sanger / NGS

Avantages

- Reads longs (1000 pb) / Reads plus courts
- Efficace pour petits projets / Petits & grands projets
- Peu d'erreurs / Erreurs de séquençage

Limites

- Bas débit
- Coûteux

Séquençage nouvelle génération (NGS)

Table 2 Next-generation DNA sequencing instruments

	Cost per base[a]	Read length (bp)[b]	Speed
Minimum cost per base			
Complete Genomics	Low	Short	3 months
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

- coûteux

séquences + longues

Jay Shendure & Erez Lieberman Aiden, Nature Biotechnology 30, 1084–1094 (2012)


IMATION
2HD
IBM
Formatted
Formatée
1.44 MB

http://www.flickr.com/photos/esquimo_2ooo/5241744434/

Plan

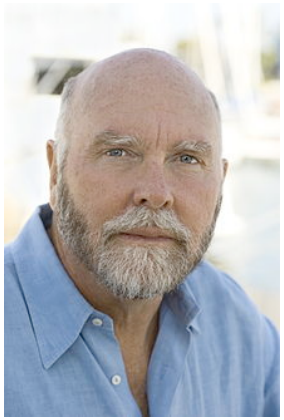
- Avant le NGS (méthode Sanger)
- La révolution NGS
- Depuis le NGS
- Séquençage avec la plateforme Illumina
- Exemple

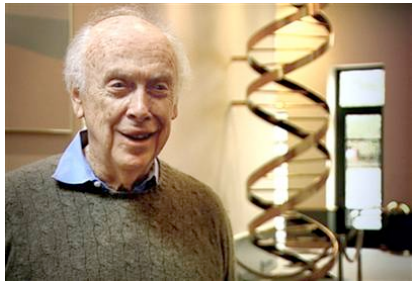
Séquençage de génomes personnels

QUICKER, SMALLER, CHEAPER

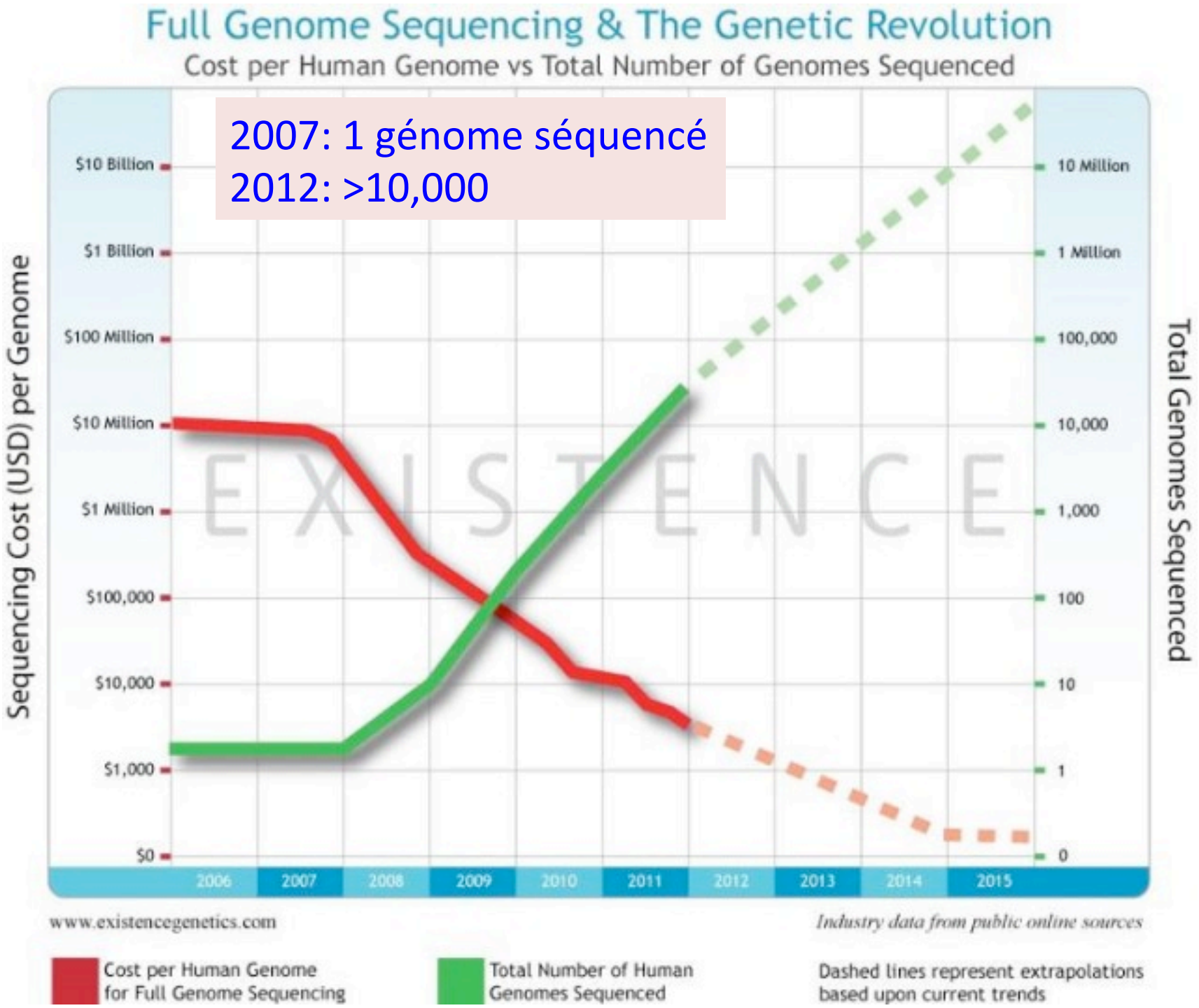
Genome sequenced (publication year)	HGP (2003)	Venter (2007)	Watson (2008)
Time taken (start to finish)	13 years	4 years	4.5 months
Number of scientists listed as authors	> 2,800	31	27
Cost of sequencing (start to finish)	$2.7 billion	$100 million	< $1.5 million
Coverage	8–10 ×	7.5 ×	7.4 ×
Number of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

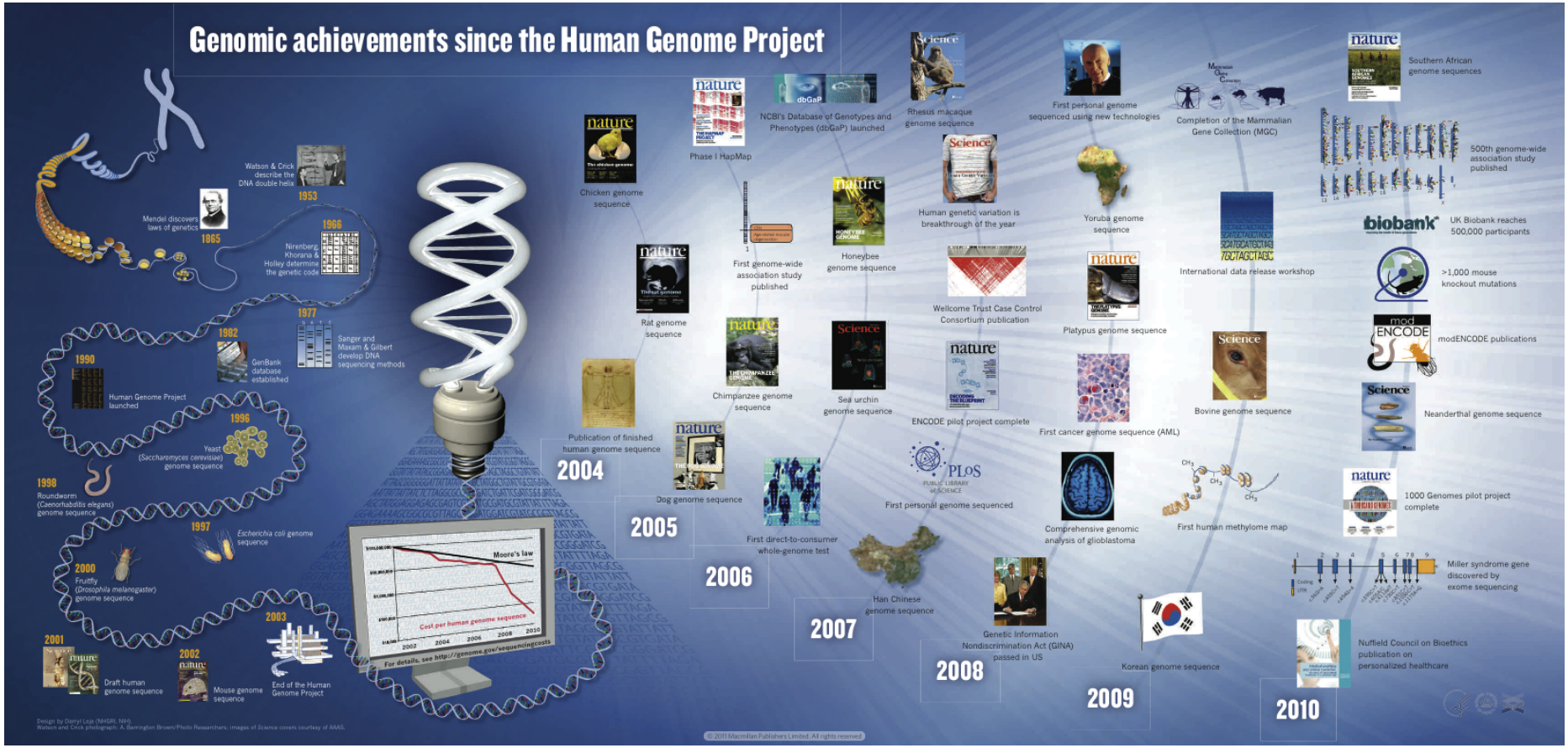
Sanger Sanger 454





Nature News, 16 April 2008 | 452, 788 (2008)


Full Genome Sequencing & The Genetic Revolution
Cost per Human Genome vs Total Number of Genomes Sequenced
2007: 1 génome séquencé
2012: >10,000
Sequencing Cost (USD) per Genome
$10 Billion
$1 Billion
$100 Million
$10 Million
$1 Million
$100,000
$10,000
$1,000
$0
Total Genomes Sequenced
10 Million
1 Million
100,000
10,000
1,000
100
10
1
0
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
EXISTENCE
www.existencegenetics.com
Industry data from public online sources
Cost per Human Genome for Full Genome Sequencing
Total Number of Human Genomes Sequenced
Dashed lines represent extrapolations based upon current trends



Eric D. Green, Mark S. Guyer & National Human Genome Research Institute
Nature 470, 204–213 (10 February 2011)

Exemples

- Génome (ADN)

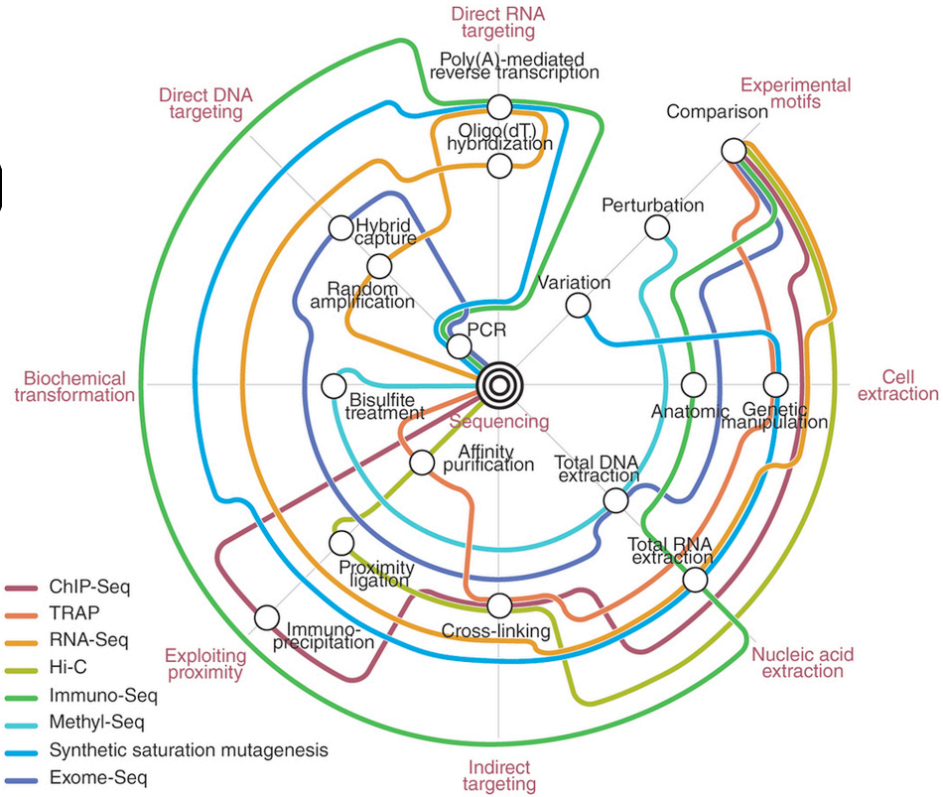
Séquençage *de novo*
Génome entier
Séquençage ciblé (exome, gènes…)

- Transcriptome (ARN)

mRNA-Seq
Small RNA-Seq (miRNA)

- Autres

ChIP-Seq
Methyl-Seq

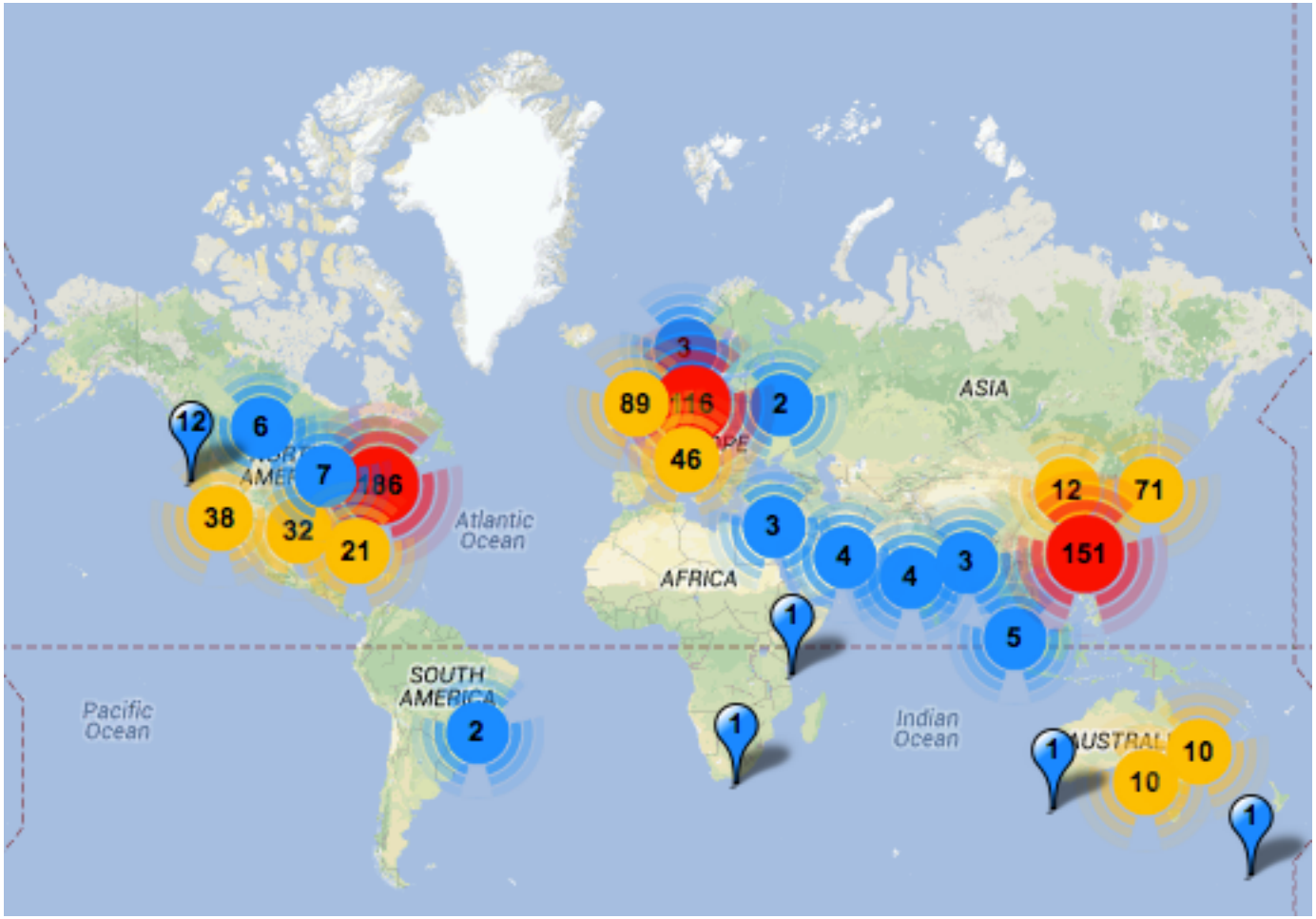


Jay Shendure & Erez Lieberman Aiden, Nature Biotechnology 30, 1084–1094 (2012)

Plan

- Avant le NGS (méthode Sanger)
- La révolution NGS
- Depuis le NGS
- Séquençage avec la plateforme Illumina
- Exemple

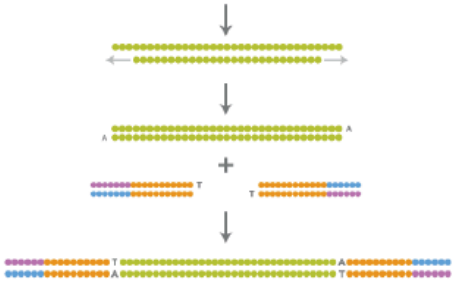
HiSeq (Illumina)



http://omicsmaps.com/

Overview

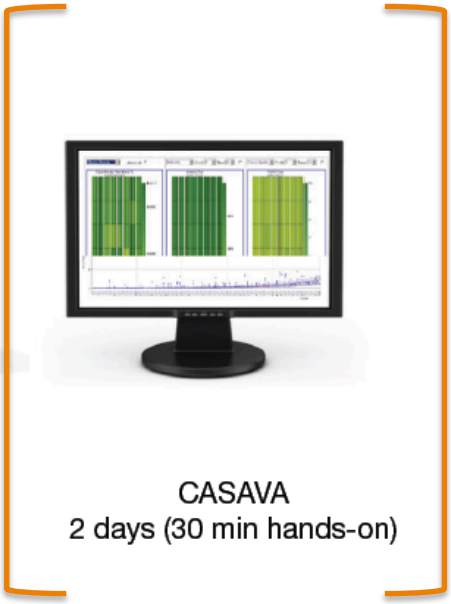
Figure 3: Next-Generation Sequencing Simplified



Library Preparation
~2 h [15 min hands-on (Nextera)]
< 6 h [< 3 h hands-on (TruSeq)]

Cluster Generation
~5 h (<10 min hands-on)

Sequencing by Synthesis
~1.5 to 11 days



CASAVA
2 days (30 min hands-on)

From simplified sample preparation kits and automated cluster generation, to streamlined sequencing by synthesis and complete data analysis, Illumina HiSeq sequencing systems offer the industry's simplest next-generation sequencing workflow.

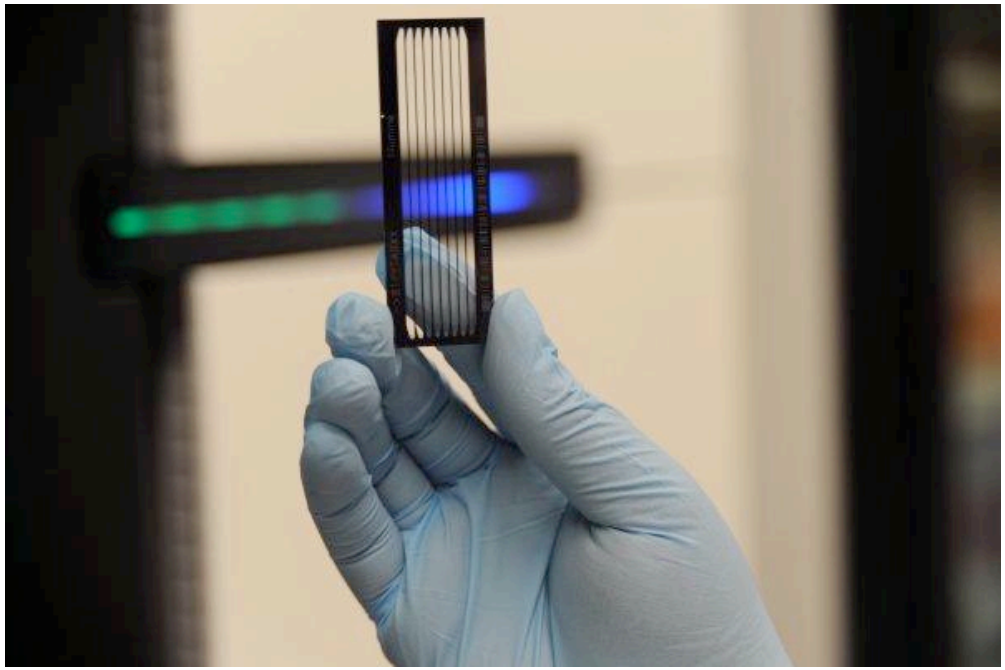
Préparation de la librairie

Génération de clusters

Séquençage => reads

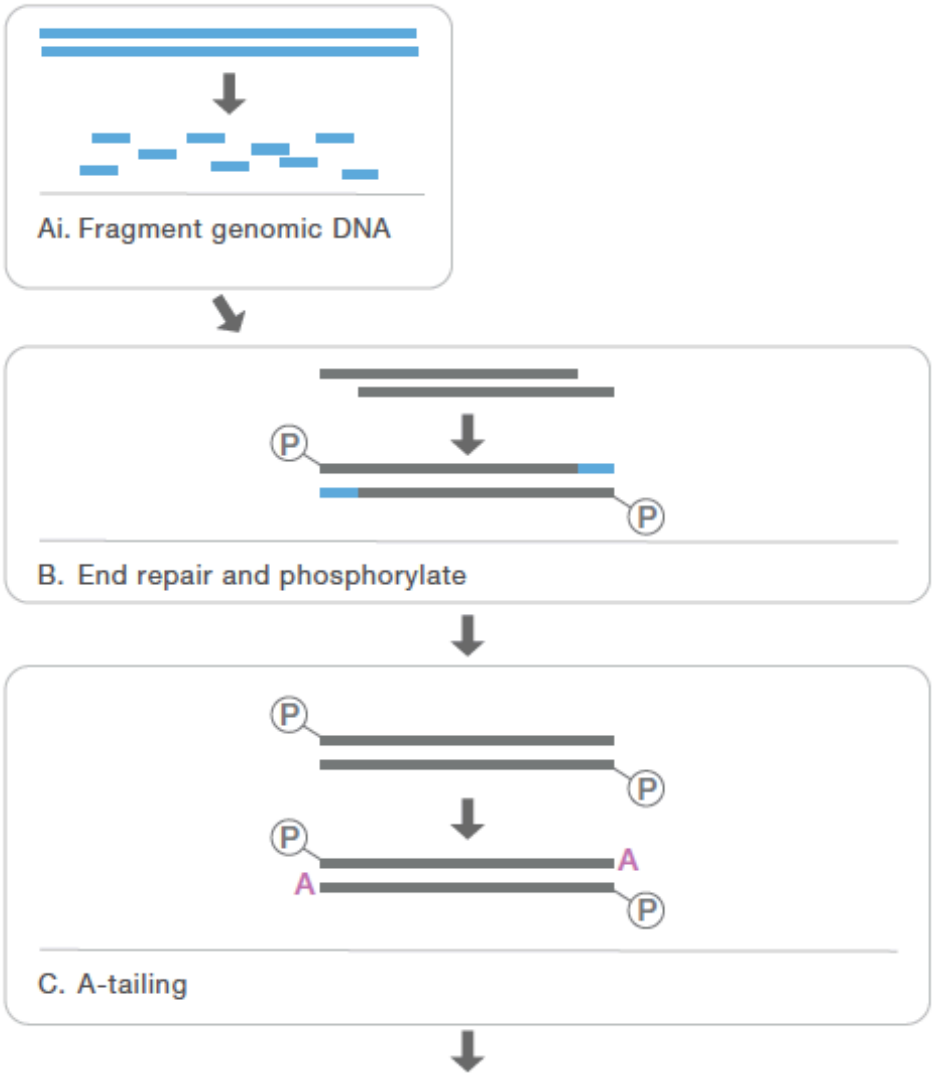
Bioinformatique +++

Flow cell

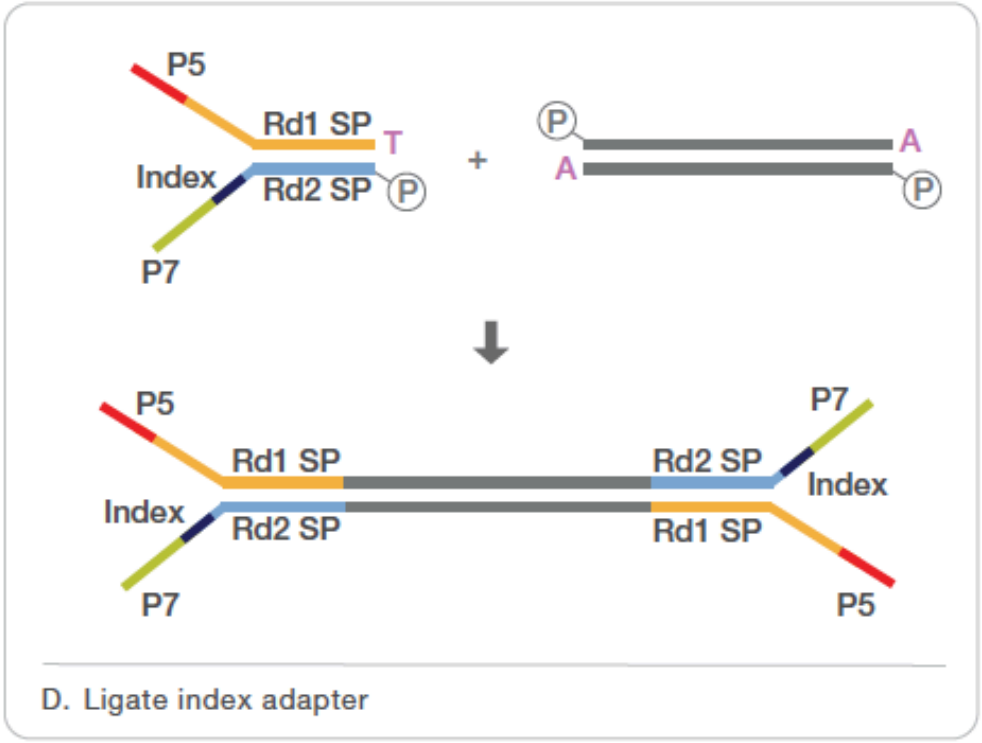


1 flow cell
=8 lanes

1/ Préparation de librairie



1/ Préparation de librairie



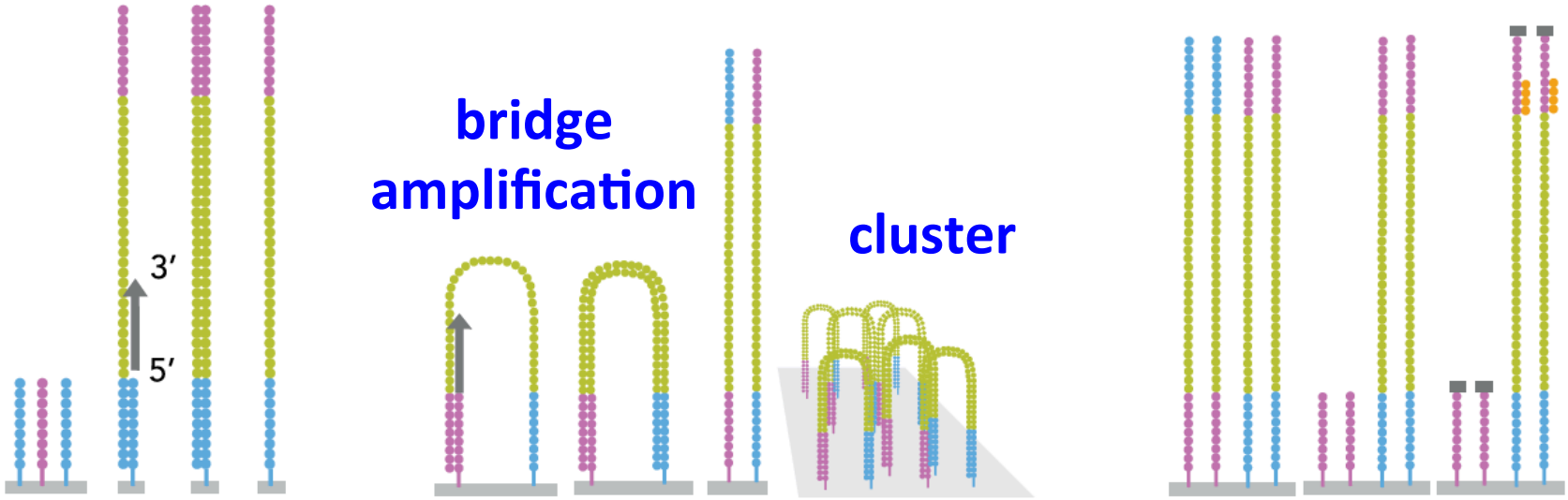
2/ cBOT

Figure 2: New cBot Features Enables Rapid and Streamlined Cluster Generation

The cBot cluster generation system is the next generation of workflow improvements for Illumina sequencing. Novel innovations include pre-packaged reagents, a single manifold, advanced fluidics and thermal stage features, integrated sensors, remote monitoring capabilities, and simplified data entry and tracking with the touch screen and barcode scanner.

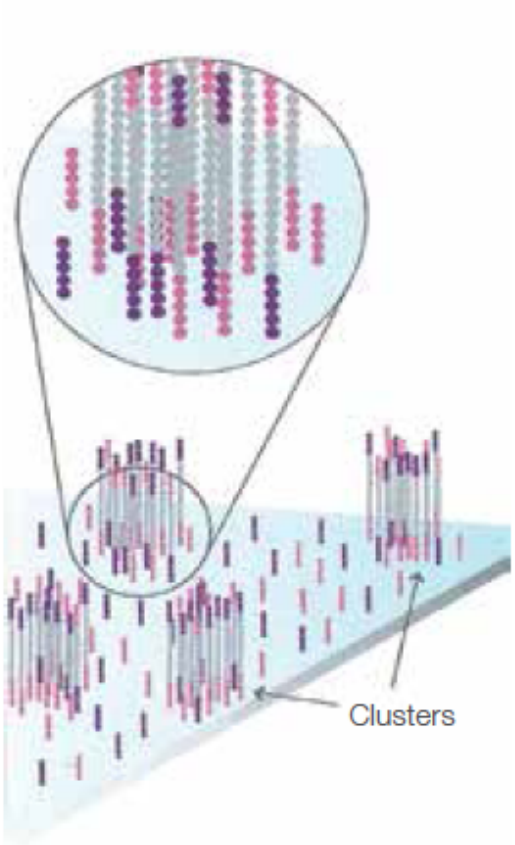
2/ Génération de clusters

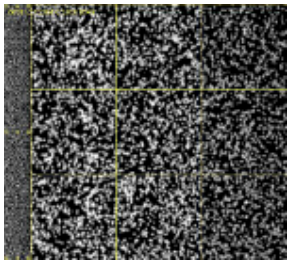
Figure 3: Cluster Generation by Isothermal Bridge Amplification



Cluster generation from single-molecule DNA templates occurs within the sealed Illumina flow cell on the cBot instrument, and involves immobilization and 3' extension, bridge amplification, linearization, and hybridization.

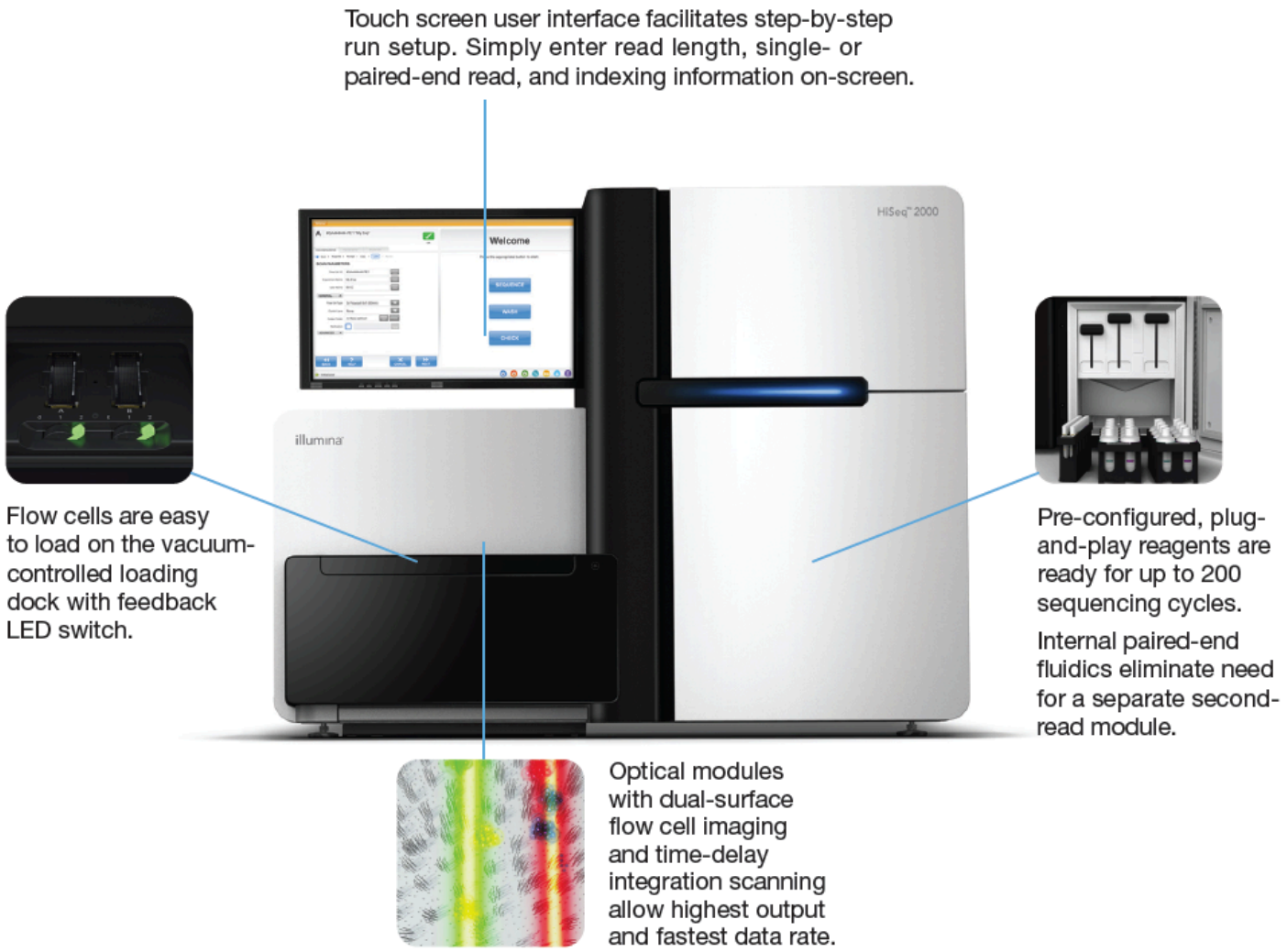
2/ Génération de clusters





100-200 millions de clusters/lane

3/ HiSeq

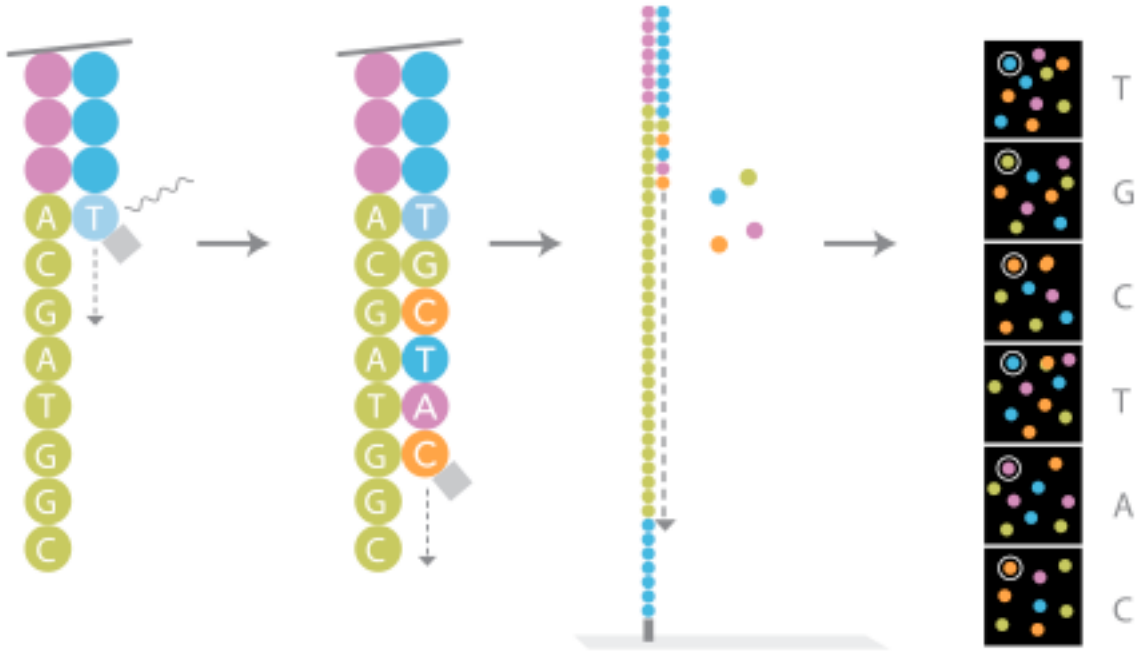

Touch screen user interface facilitates step-by-step run setup. Simply enter read length, single- or paired-end read, and indexing information on-screen.
Welcome
HiSeq™ 2000
illumina
Flow cells are easy to load on the vacuum-controlled loading dock with feedback LED switch.
Pre-configured, plug-and-play reagents are ready for up to 200 sequencing cycles.
Internal paired-end fluidics eliminate need for a separate second-read module.
Optical modules with dual-surface flow cell imaging and time-delay integration scanning allow highest output and fastest data rate.

3/ Séquençage

Cycle 1

3/ Séquençage

Cycle 1 … Cycle 6 …



➔ **reads**

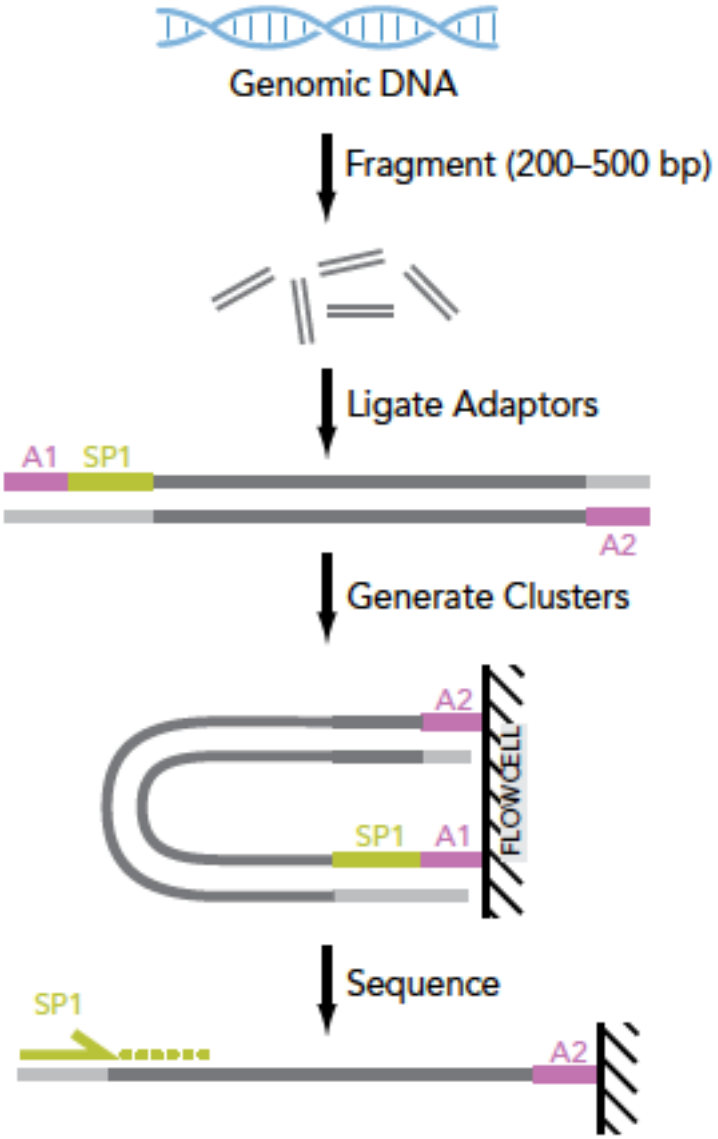
(ex. 50 pb, 75 pb ou 100 pb)

Chaque base d'un read est associée à une valeur de **qualité**

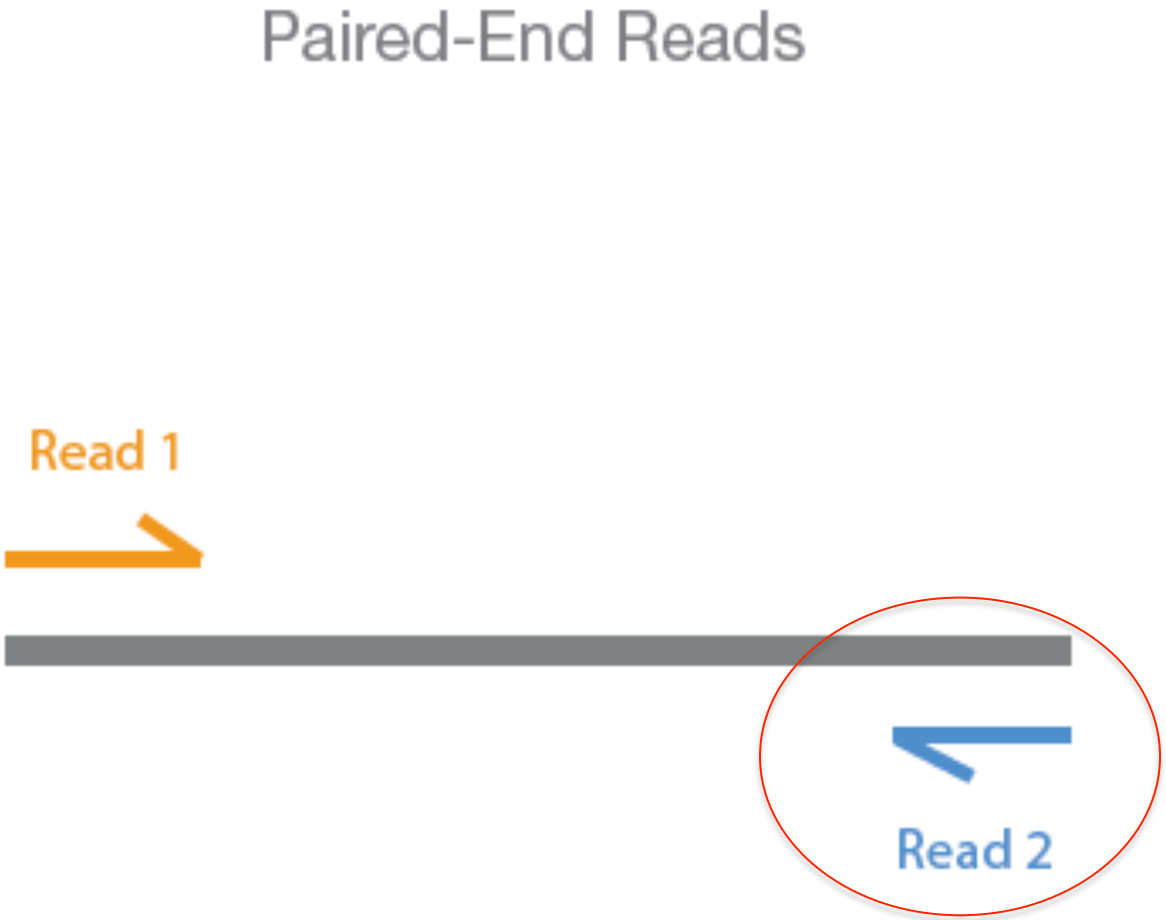
Animation

http://www.yourgenome.org/teachers/speed.shtml

Single-end sequencing



Paired-end sequencing (1)



Paired-end sequencing (2)

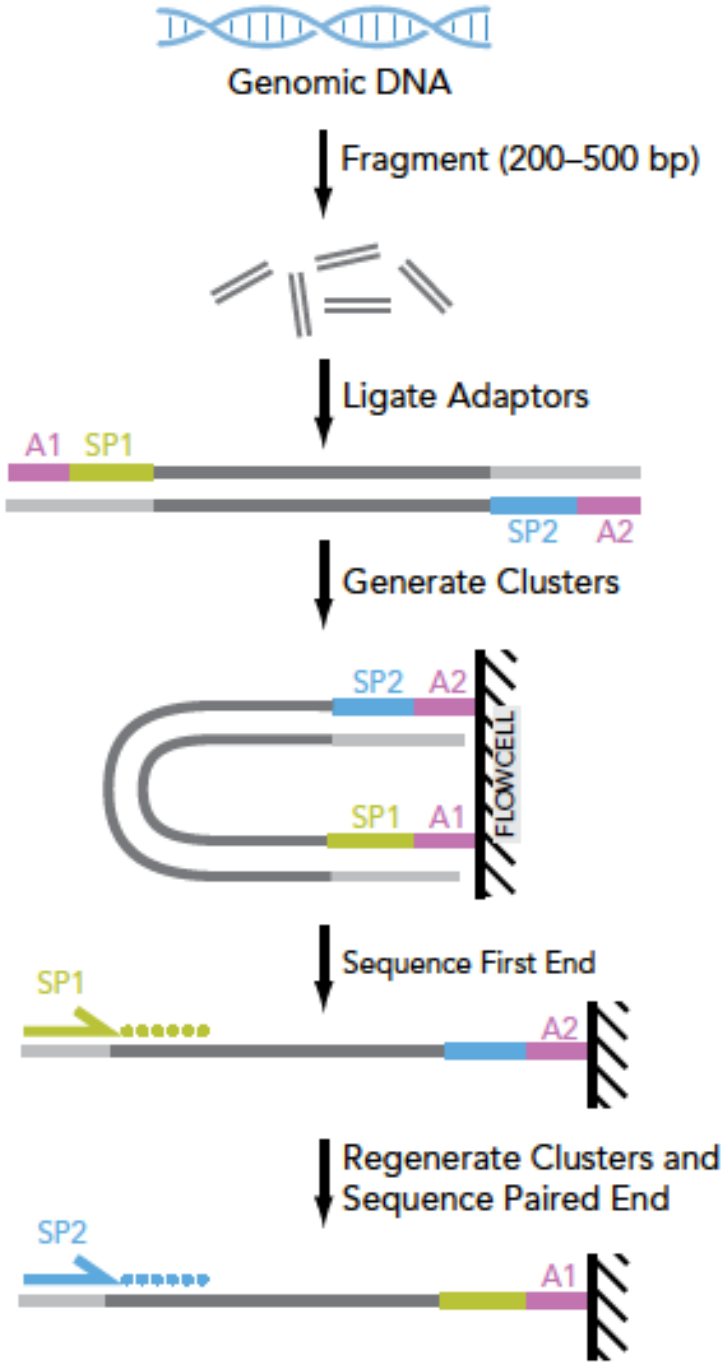
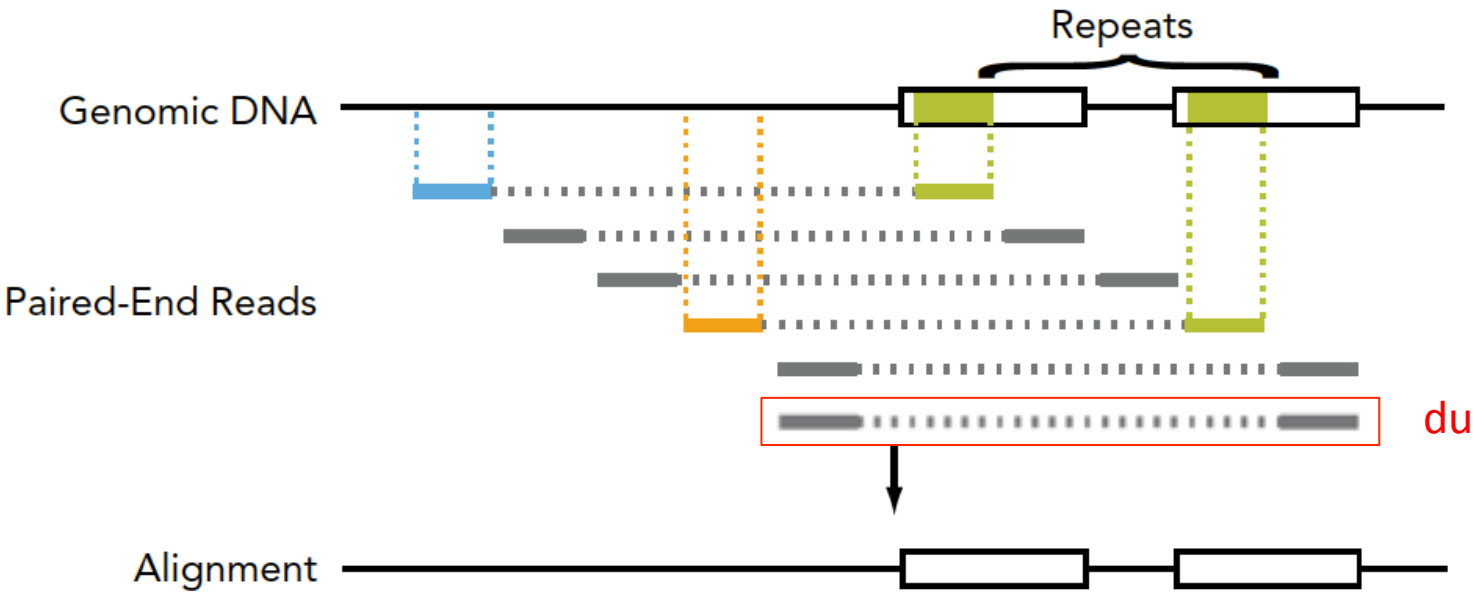


Figure 2: Unique Alignment Of Paired Reads In Repeats

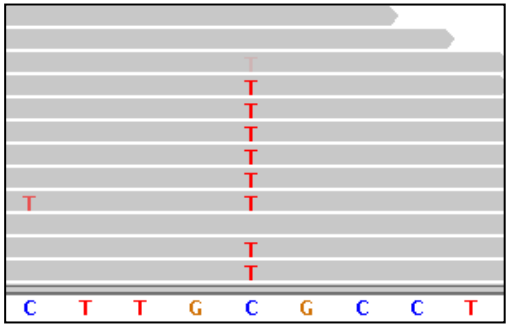


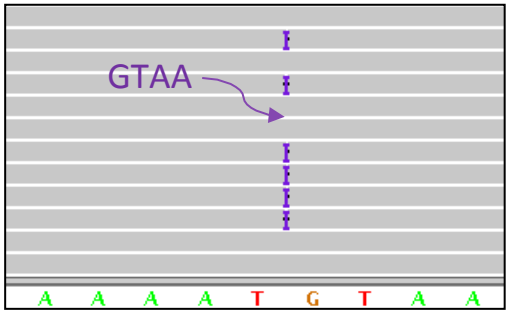
Reads in repeats (green) can be unambiguously aligned in complex genomes. Each read is associated with a paired read (blue or orange) and the separation between read pairs is known from the fragment size of the input DNA.

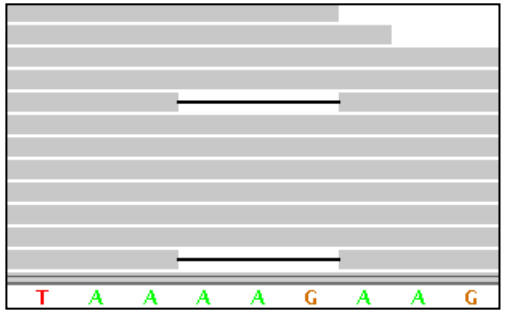
Table 1: Flexible Paired Sequencing Provides Optimal Detection Of Any Variant

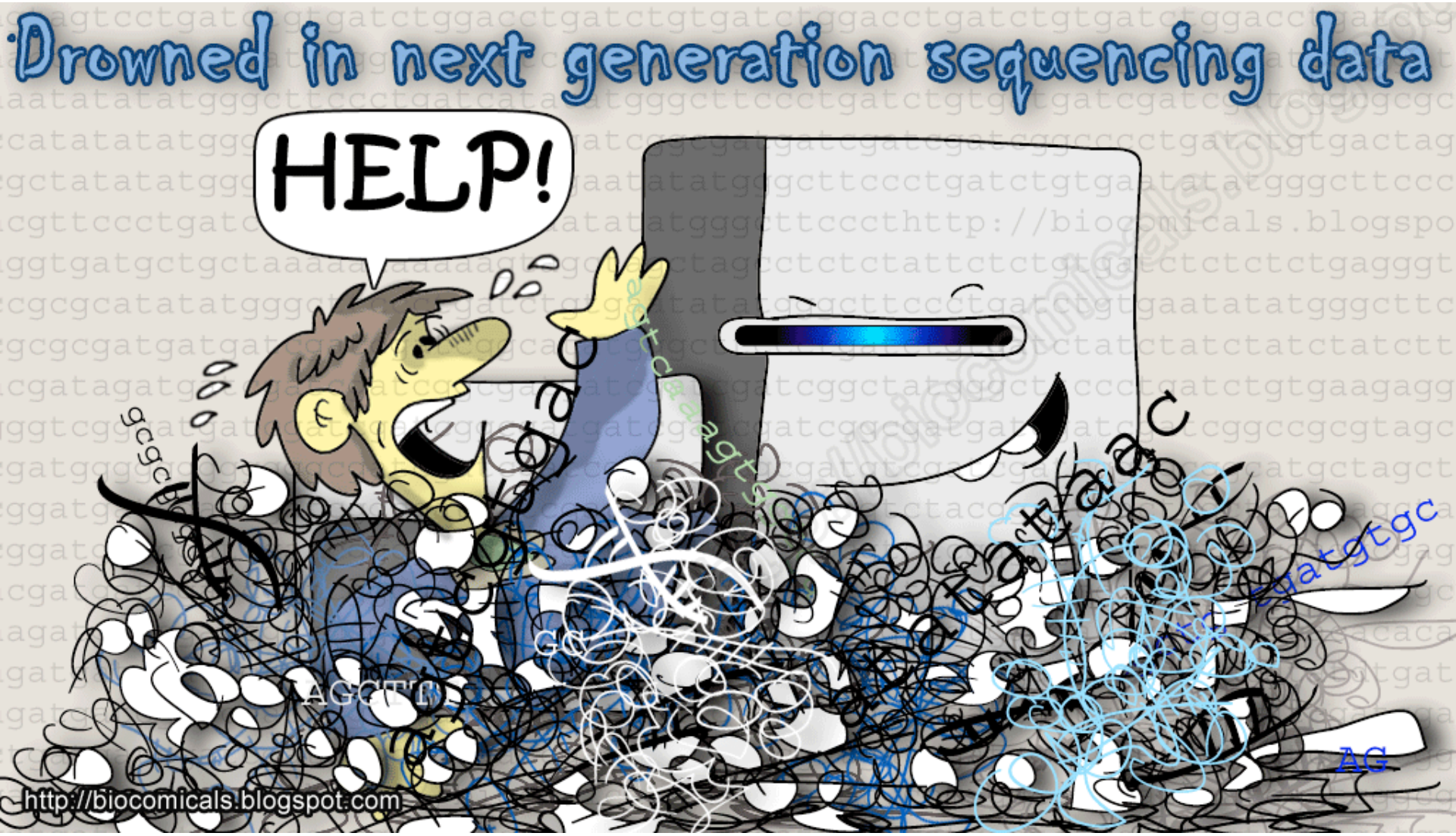
Variant	Single Read	Short Insert Paired-Ends (200–500 bp)	Long Insert Mate Pairs (2–5 kb)	Paired-End And Mate Pair Combined
SNP	++	++++	++	++++
Small indels	++	++++	++	++++
Insertion	+	+++	+++	++++
Amplification	++	+++	+++	++++
Deletion	+	+++	++	++++
Inversion	+	+++	++	++++
Complex rearrangement	+	+++	++	++++
Large rearrangement	+	++	+++	++++

Only by combing short and long inserts can researchers be certain to find all different sizes and types of variants. In particular, short inserts are essential to identifying small indels and mate pairs are essential for identifying the largest rearrangements.

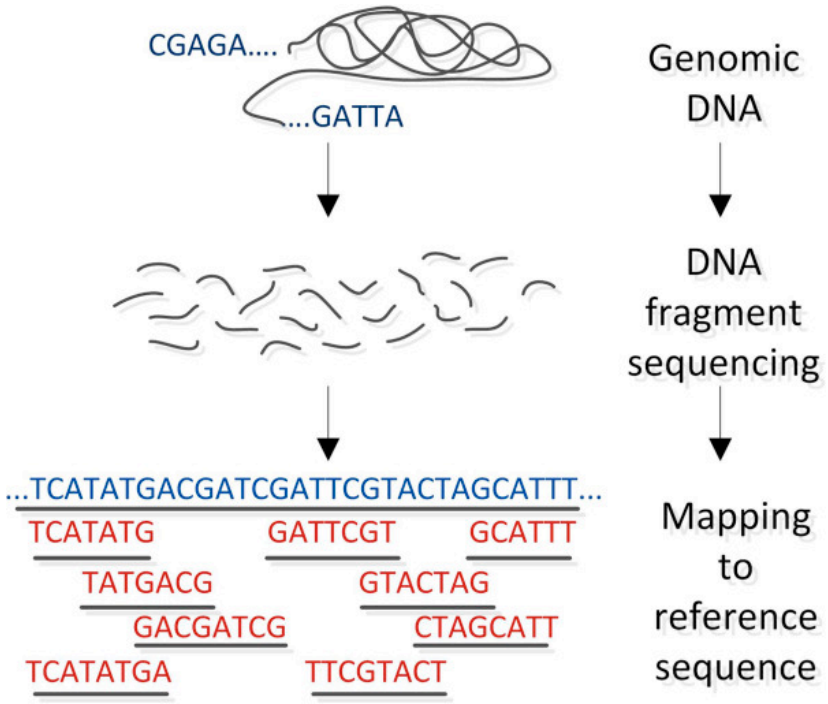




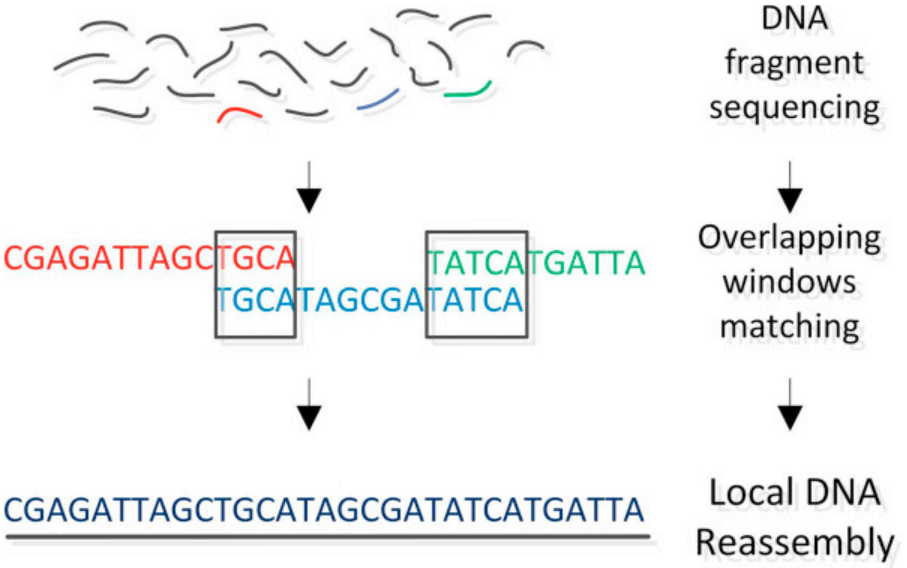



Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com

Génome de référence existant => alignement



Assemblage *de novo*



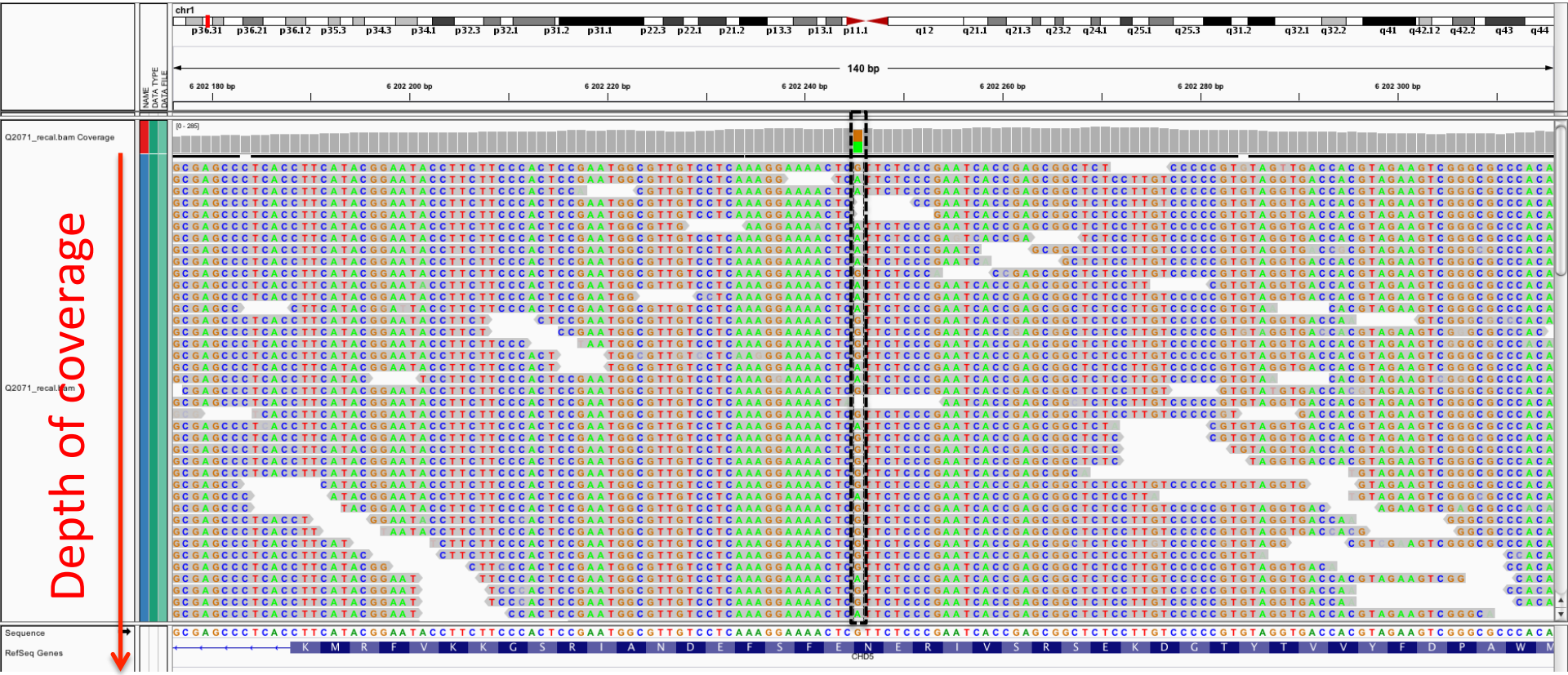
Pavlopoulos et al., BioData Mining 2013, 6:13

Fichiers **FASTQ** (reads +qualités)

Fichiers **BAM** (reads alignés)

Fichiers BAM améliorés

Fichiers **VCF** (variants)

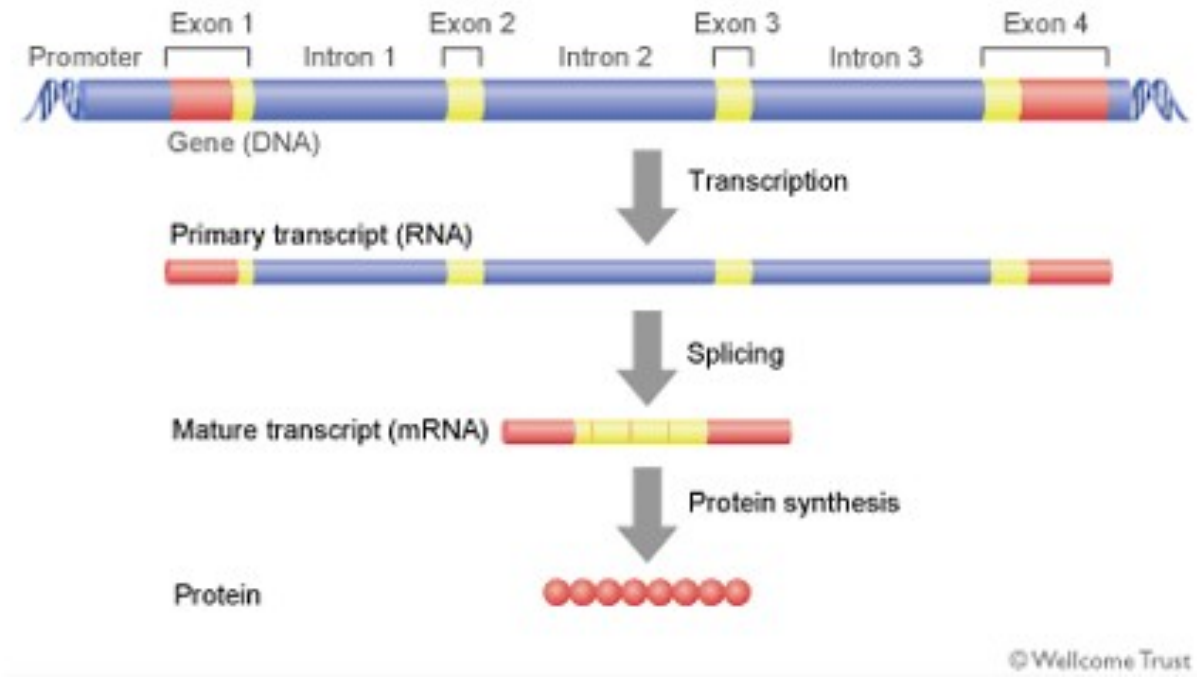


Plan

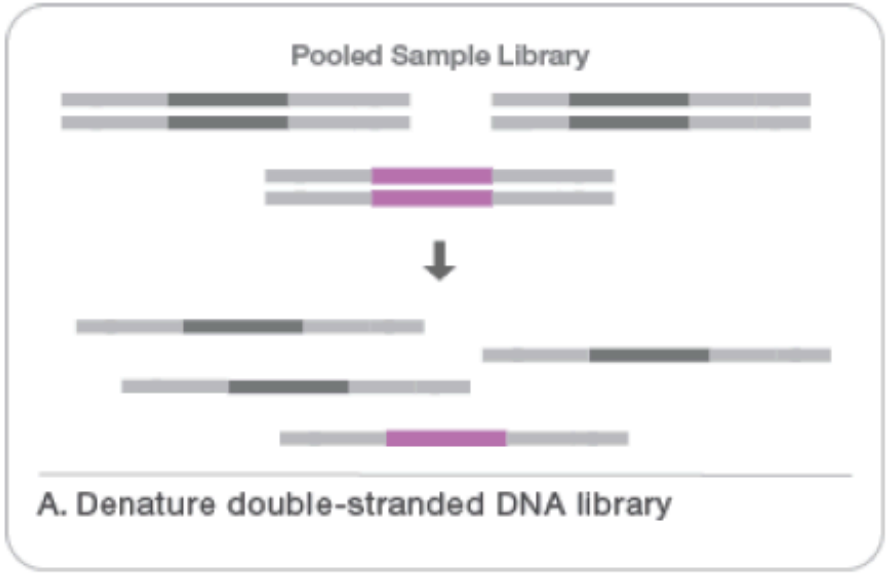
- Avant le NGS (méthode Sanger)
- La révolution NGS
- Depuis le NGS
- Séquençage avec la plateforme Illumina
- Exemple

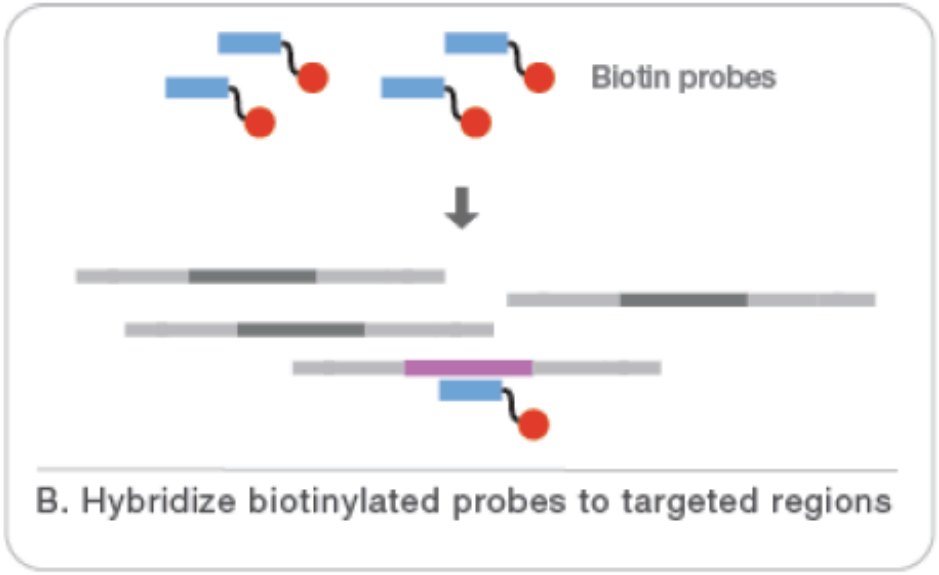
Identification de la mutation génétique responsable d'une maladie chez le patient X

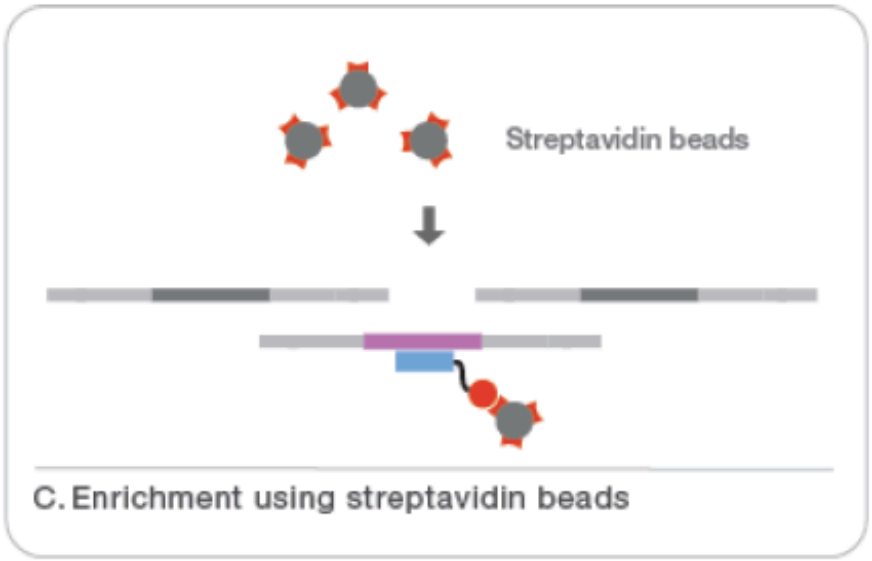
Séquençage du génome entier ou de l'exome?



Séquençage d'exome







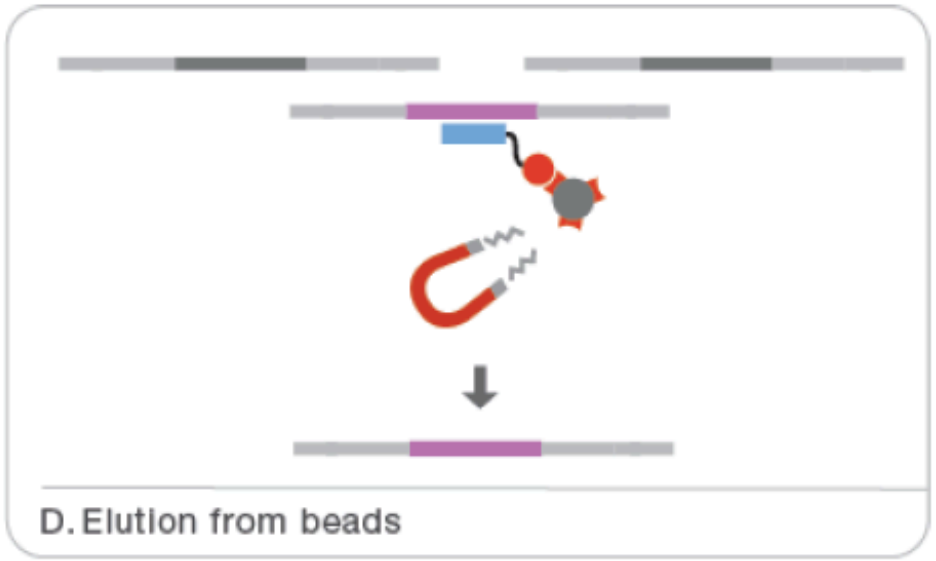
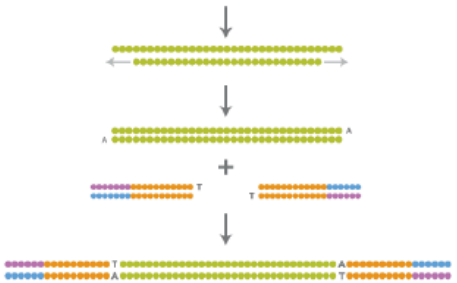


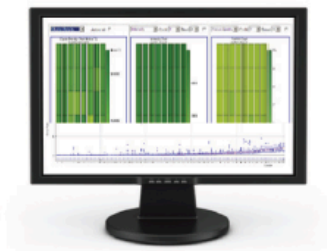
Figure 3: Next-Generation Sequencing Simplified



Library Preparation
~2 h [15 min hands-on (Nextera)]
< 6 h [< 3 h hands-on (TruSeq)]

Cluster Generation
~5 h (<10 min hands-on)

Sequencing by Synthesis
~1.5 to 11 days



CASAVA
2 days (30 min hands-on)

From simplified sample preparation kits and automated cluster generation, to streamlined sequencing by synthesis and complete data analysis, Illumina HiSeq sequencing systems offer the industry's simplest next-generation sequencing workflow.

Préparation de la librairie

Génération de clusters

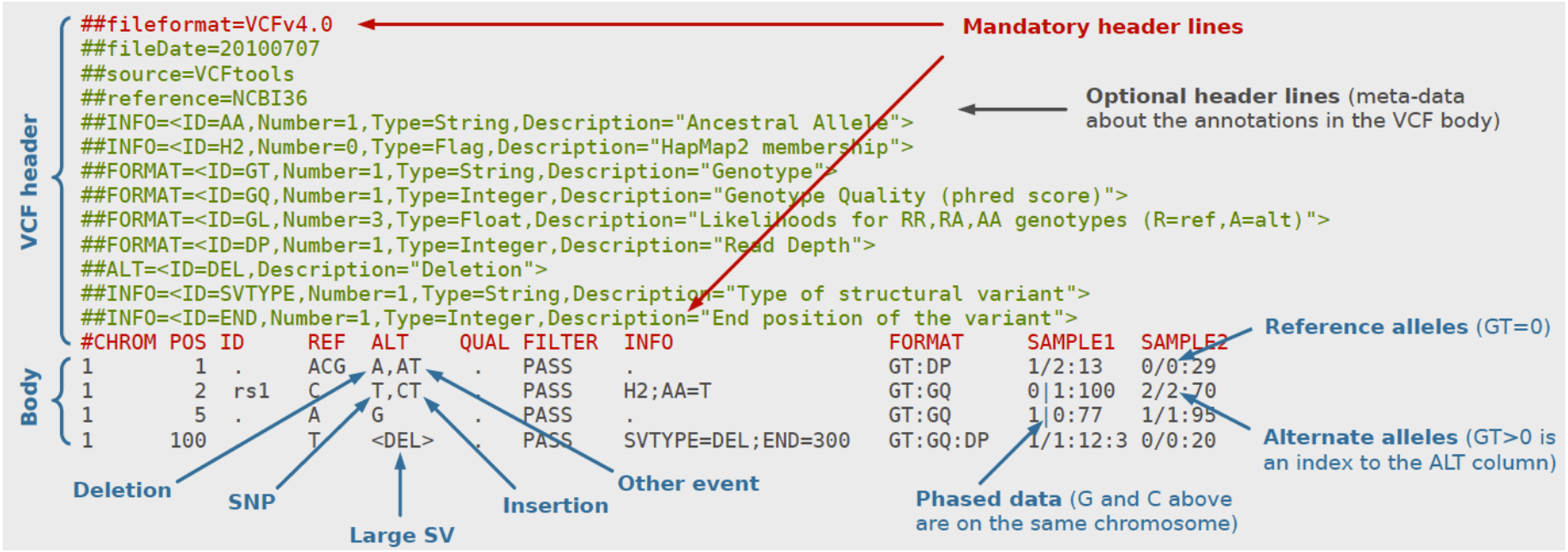
Séquençage => reads

Bioinformatique +++

Fichiers FASTQ (reads +qualités) → **Fichiers BAM (reads alignés)** → **Fichiers BAM améliorés** → **Fichiers VCF (variants)**

Le format VCF

VCF = Variant Call Format



http://digitheadslabnotebook.blogspot.fr/2013/01/vcf-variant-call-format.html

Séquençage d'exome

Variant type	Mean number of variants (± sd) in African Americans	Mean number of variants (± sd) in European Americans
Novel variants		
Missense	303 (± 32)	192 (± 21)
Nonsense	5 (± 2)	5 (± 2)
Synonymous	209 (± 26)	109 (± 16)
Splice	2 (± 1)	2 (± 1)
Total	520 (± 53)	307 (± 33)
Non-novel variants		
Missense	10,828 (± 342)	9,319 (± 233)
Nonsense	98 (± 8)	89 (± 6)
Synonymous	12,567 (± 416)	10,536 (± 280)
Splice	36 (± 4)	32 (± 3)
Total	23,529 (± 751)	19,976 (± 505)
Total variants		
Missense	11,131 (± 364)	9,511 (± 244)
Nonsense	103(± 8)	93 (± 6)
Synonymous	12,776 (± 434)	10,645 (± 286)
Splice	38 (± 5)	34 (± 4)
Total	24,049 (± 791)	20,283 (± 523)

Nature Reviews Genetics 12, 745-755 (November 2011)

BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 15 2011, pages 2156–2158
doi:10.1093/bioinformatics/btr330

Sequence analysis

Advance Access publication June 7, 2011

The variant call format and VCFtools

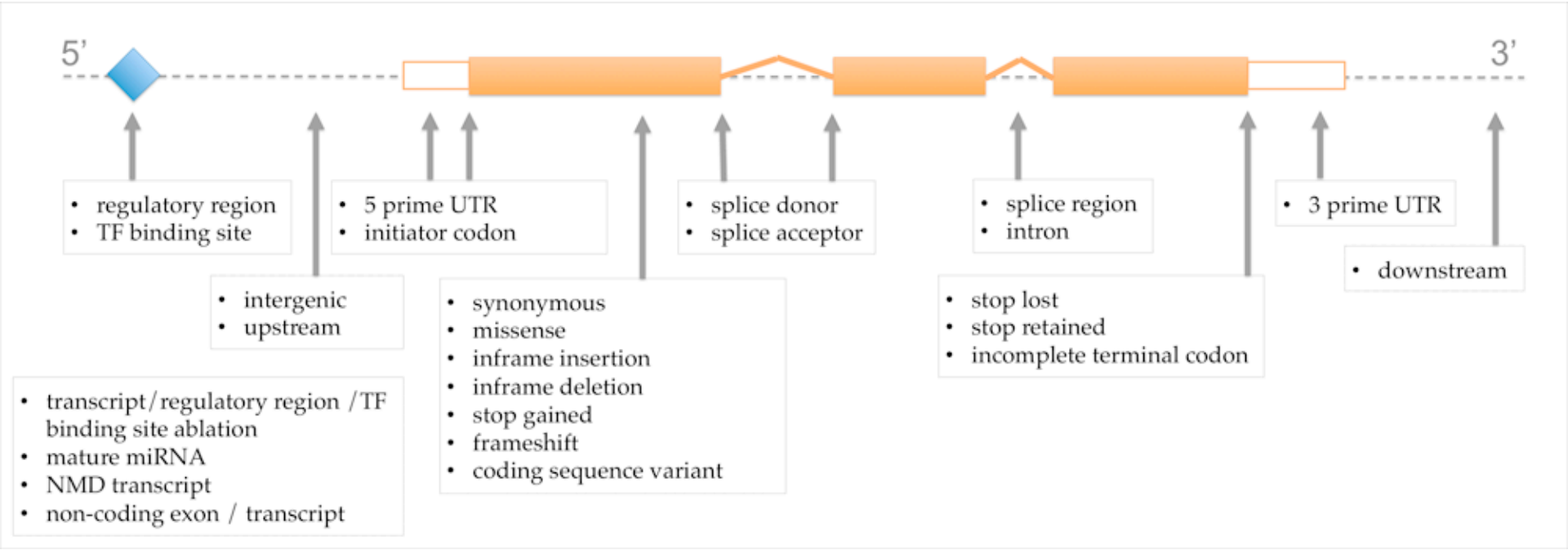
Petr Danecek[1,†], Adam Auton[2,†], Goncalo Abecasis[3], Cornelis A. Albers[1], Eric Banks[4], Mark A. DePristo[4], Robert E. Handsaker[4], Gerton Lunter[2], Gabor T. Marth[5], Stephen T. Sherry[6], Gilean McVean[2,7], Richard Durbin[1,*] and 1000 Genomes Project Analysis Group[‡]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK, [3]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02141, [5]Department of Biology, Boston College, MA 02467, [6]National Institutes of Health National Center for Biotechnology Information, MD 20894, USA and [7]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK

Associate Editor: John Quackenbush

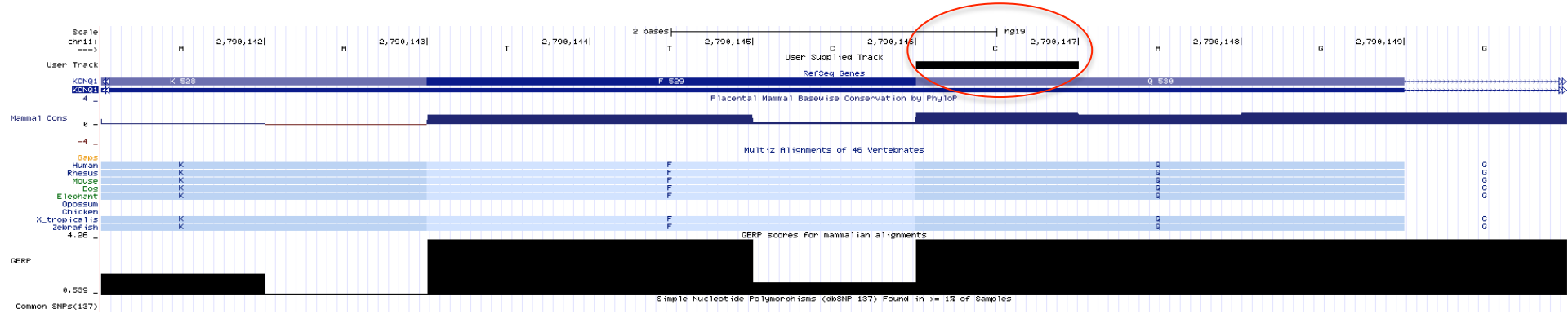
https://github.com/amarcket/vcf_spec

Annotation fonctionnelle (1)



Annotation fonctionnelle (2)

CAG => AAG



#CHROM	POS	REF	ALT
Chr11	2790147	C	A

le code génétique

Première lettre (côté 5')	Deuxième lettre U		Deuxième lettre C		Deuxième lettre A		Deuxième lettre G		Troisième lettre (côté 3')
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

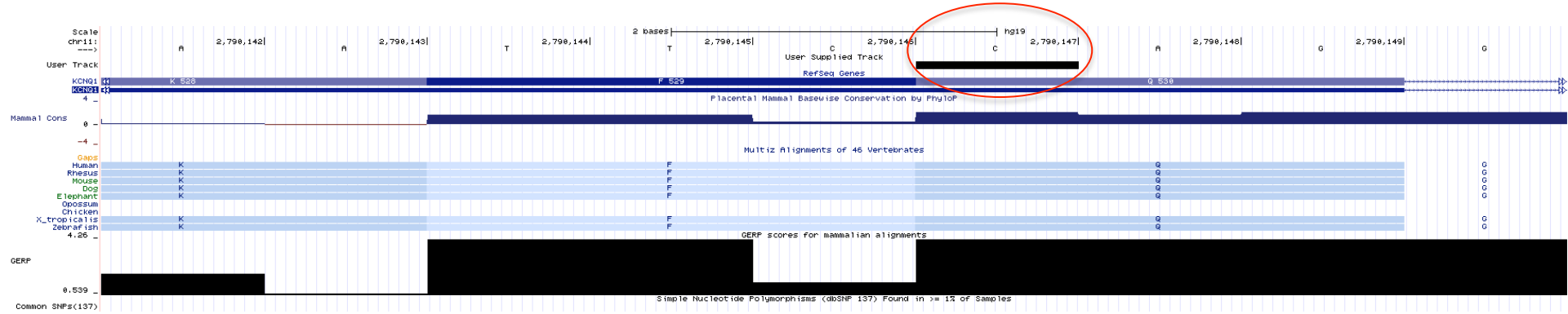
codon d'initiation

codon de terminaison

Annotation fonctionnelle (3)

Q	K
Gln	Lys

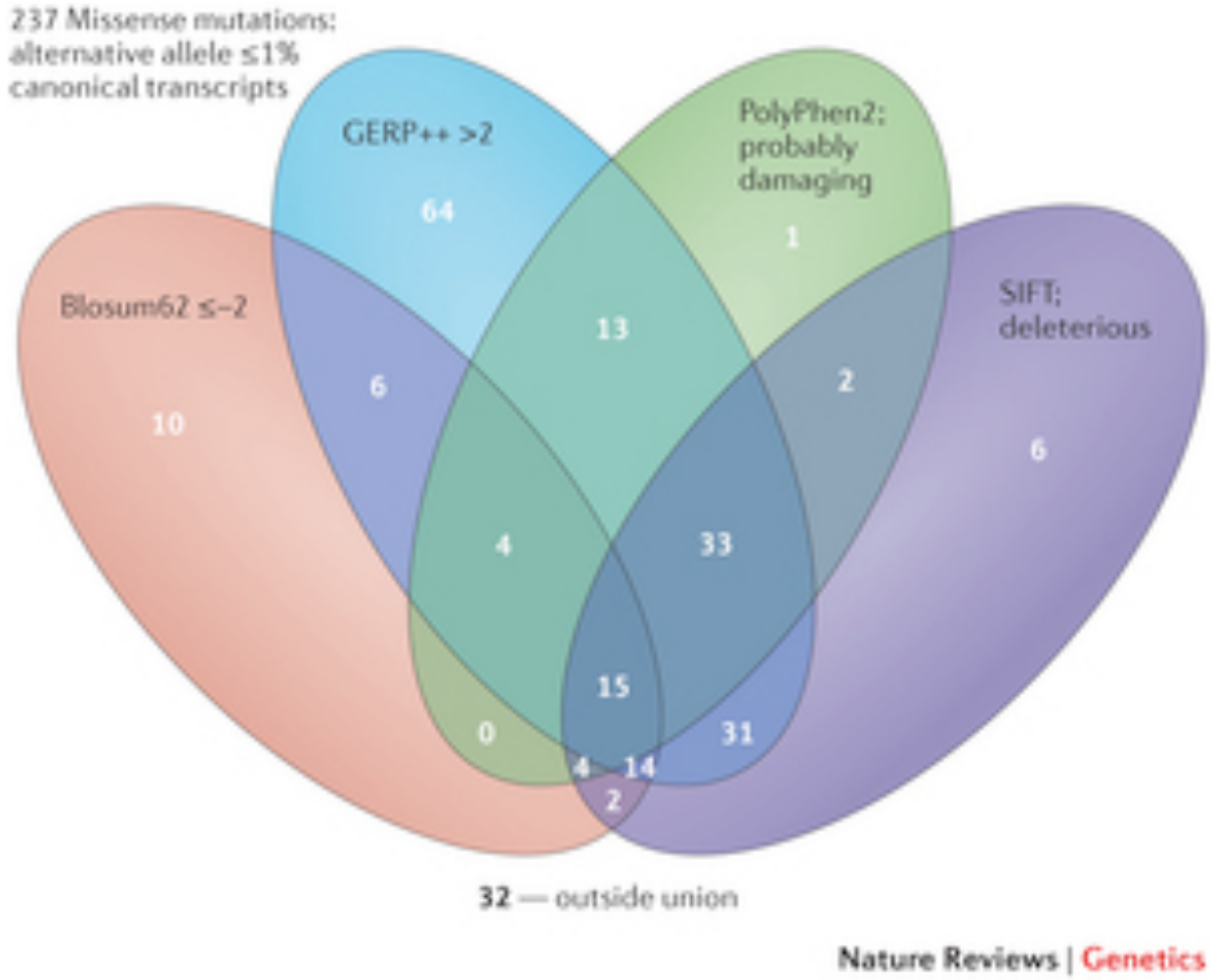
CAG => AAG



#CHROM	POS	REF	ALT
Chr11	2790147	C	A

Mutation faux-sens Q530K

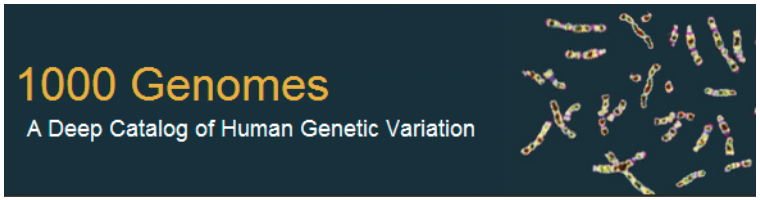
Outils de prédiction



Nature Reviews Genetics 14, 460–470 (2013)

Prioritisation des variants

Focus sur les variants rares?



NHLBI Exome Sequencing Project (ESP)

Exome Variant Server

Variant Pos	rs ID	Alleles	EA Allele #	AA Allele #	All Allele #	EA Genotype #	AA Genotype #	All Genotype #	MAF (%) (EA/AA/All)
3:38595746	rs45440596	C>T	T=1/C=8327	T=1/C=3991	T=2/C=12318	TT=0/TC=1 /CC=4163	TT=0/TC=1 /CC=1995	TT=0/TC=2 /CC=6158	0.012/0.0251 /0.0162
3:38595747	rs369559182	G>A	A=8/G=8318	A=0/G=4006	A=8/G=12324	AA=0/AG=8 /GG=4155	AA=0/AG=0 /GG=2003	AA=0/AG=8 /GG=6158	0.0961/0.0 /0.0649
3:38595795	rs141789366	G>A	A=0/G=8444	A=5/G=4161	A=5/G=12605	AA=0/AG=0 /GG=4222	AA=0/AG=5 /GG=2078	AA=0/AG=5 /GG=6300	0.0/0.12/0.0397
3:38595795	rs141789366	G>A	A=0/G=8444	A=5/G=4161	A=5/G=12605	AA=0/AG=0 /GG=4222	AA=0/AG=5 /GG=2078	AA=0/AG=5 /GG=6300	0.0/0.12/0.0397
3:38595801	rs376720757	G>A	A=1/G=8453	A=1/G=4177	A=2/G=12630	AA=0/AG=1 /GG=4226	AA=0/AG=1 /GG=2088	AA=0/AG=2 /GG=6314	0.0118/0.0239 /0.0158
3:38595801	rs376720757	G>A	A=1/G=8453	A=1/G=4177	A=2/G=12630	AA=0/AG=1 /GG=4226	AA=0/AG=1 /GG=2088	AA=0/AG=2 /GG=6314	0.0118/0.0239 /0.0158
3:38595989	rs199473618	C>T	T=0/C=8396	T=7/C=4079	T=7/C=12475	TT=0/TC=0 /CC=4198	TT=0/TC=7 /CC=2036	TT=0/TC=7 /CC=6234	0.0/0.1713 /0.0561
3:38595990	rs372507927	G>A	A=0/G=8398	A=1/G=4073	A=1/G=12471	AA=0/AG=0 /GG=4199	AA=0/AG=1 /GG=2036	AA=0/AG=1 /GG=6235	0.0/0.0245 /0.008
3:38596010	rs199473269	C>T	T=1/C=8369	T=0/C=4026	T=1/C=12395	TT=0/TC=1 /CC=4184	TT=0/TC=0 /CC=2013	TT=0/TC=1 /CC=6197	0.0119/0.0 /0.0081

http://www.1000genomes.org/
http://evs.gs.washington.edu/EVS/

Jay Shendure & Erez Lieberman Aiden, Nature Biotechnology 30, 1084–1094 (2012)

Populations séquencées

Go • NL
GENOMEoftheNETHERLANDS

estonian genome center
university of tartu

UK 10K
RARE GENETIC VARIANTS IN HEALTH AND DISEASE

People of the *British Isles*

deCODE genetics

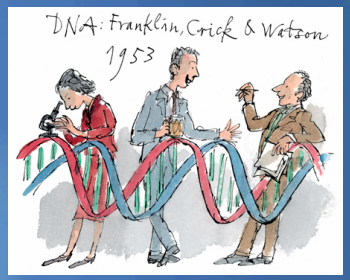
Tiny Faroe Islands to Begin Sequencing Genomes of All 50,000 Residents in Ambitious Effort to Advance Personalized Medicine

The SardiNIA Project

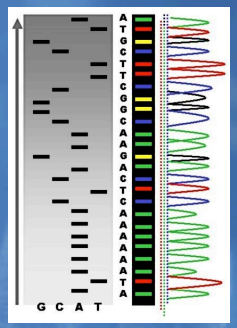
generation SCOTLAND

VACARME

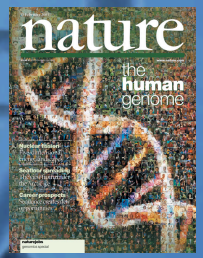
1953: Structure en double hélice (Watson & Crick)

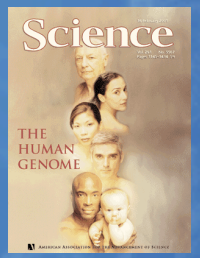


1977: Séquençage Sanger (« première génération »)

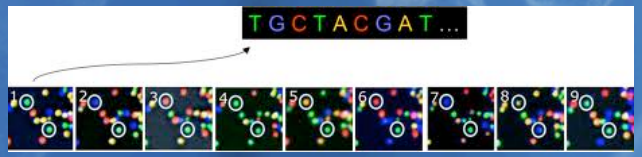


2001-2003: HGP





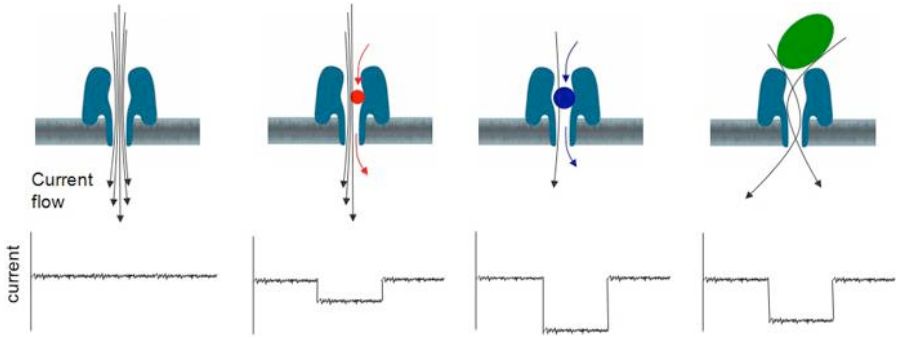
2005: Séquençage nouvelle génération (« seconde génération »)



Next-next generation sequencing

MinION: A complete DNA sequencer on a USB stick

By John Hewitt on March 29, 2013 at 10:07 am | 4 Comments



Helicos Bioxciences
Pacific Biosciences
Oxford Nanopore
…

« An exciting time for genomics »

http://www.extremetech.com/extreme/151086-minion-a-complete-dna-sequencer-on-a-usb-stick