Summer School Bioinformatics

July 3, 2015 (Nantes)

Unité Inserm UMR 1087-CNRS UMR 6291

Inserm cnrs UNIVERSITÉ DE NANTES CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Program

10h-11h : Introduction to "Next Generation Sequencing" (NGS) and its applications in human genetics
S. Le Scouarnec

11h-12h : Visit of the Genomics and Bioinformatics core facility **(2nd floor)**
R. Redon & A. Bihouée

12h-14h : Lunch

14h-17h : Interactive session : From raw sequence data to disease causing variants
P. Lindenbaum
The demonstration will include :
- a description of a FASTQ file generated by a sequencer
- aligning the short reads on a reference genome
- processing the alignments (BAM)
- calling the mutations and generating a list of variations (VCF)
- adding some functional annotations to those variants

Introduction to "Next Generation Sequencing" (NGS) and its applications in human genetics

Unité Inserm UMR 1087-CNRS UMR 6291

Solena LE SCOUARNEC

Nantes

Inserm cnrs UNIVERSITÉ DE NANTES CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Outline

- Basic concepts in genomics
- Next generation sequencing (NGS)
- NGS applications

Questions are welcomed!

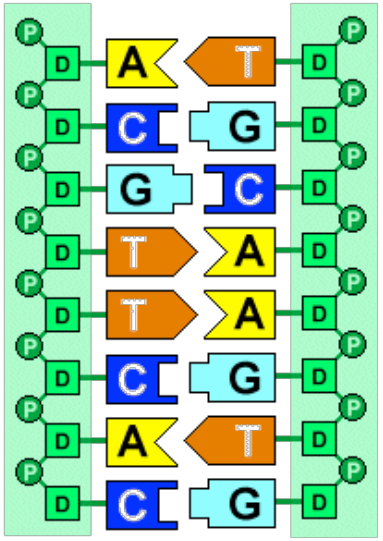
Genome

Genetic information in each organism cell

Each organism have their own genome

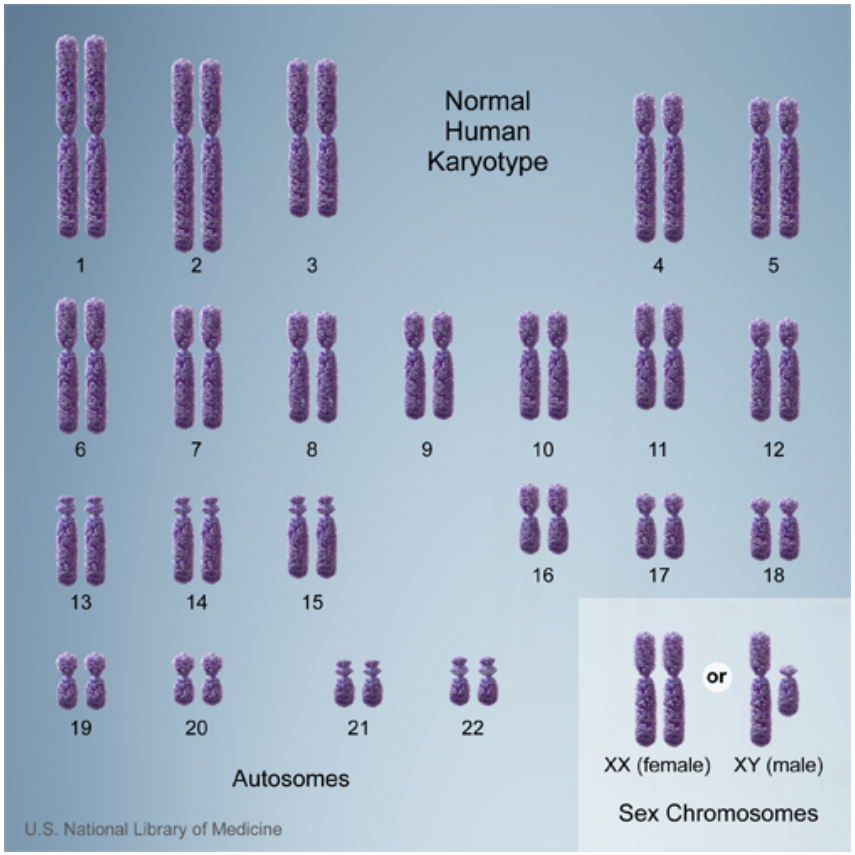
Encoded by DNA

Human genome: >3 billion nucleotides



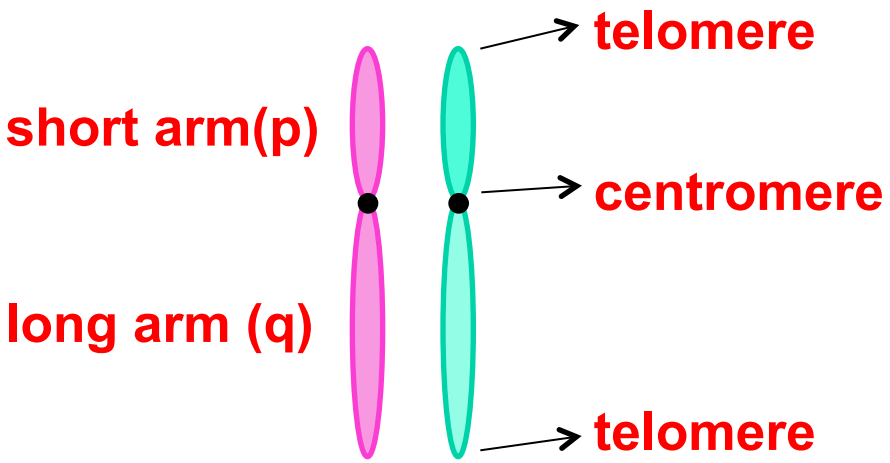
Genomics : study of genes and their functions

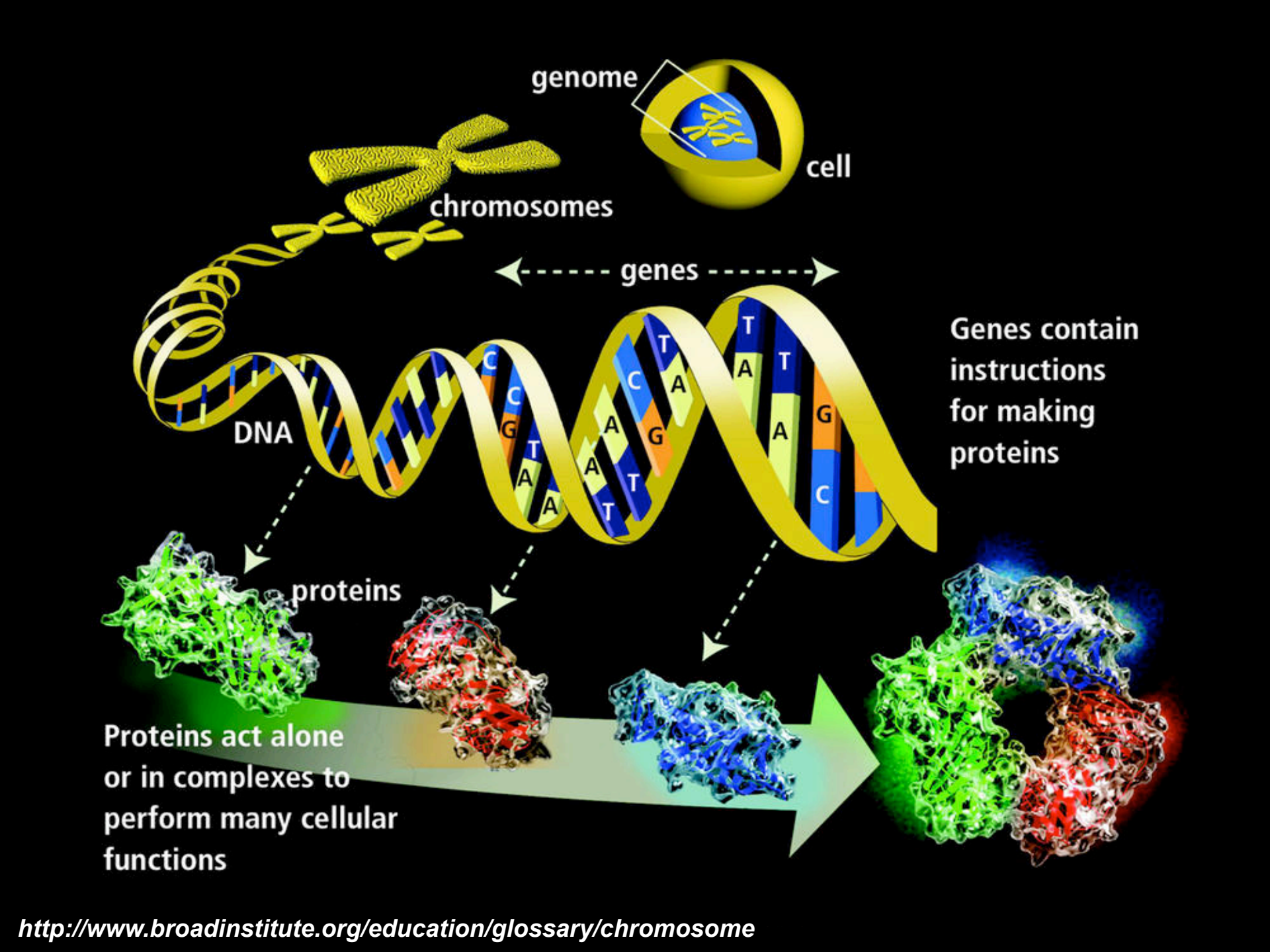
The human genome



23 pairs of chromosomes

- 22 pairs of autosomes
- One pair of sex chromosomes





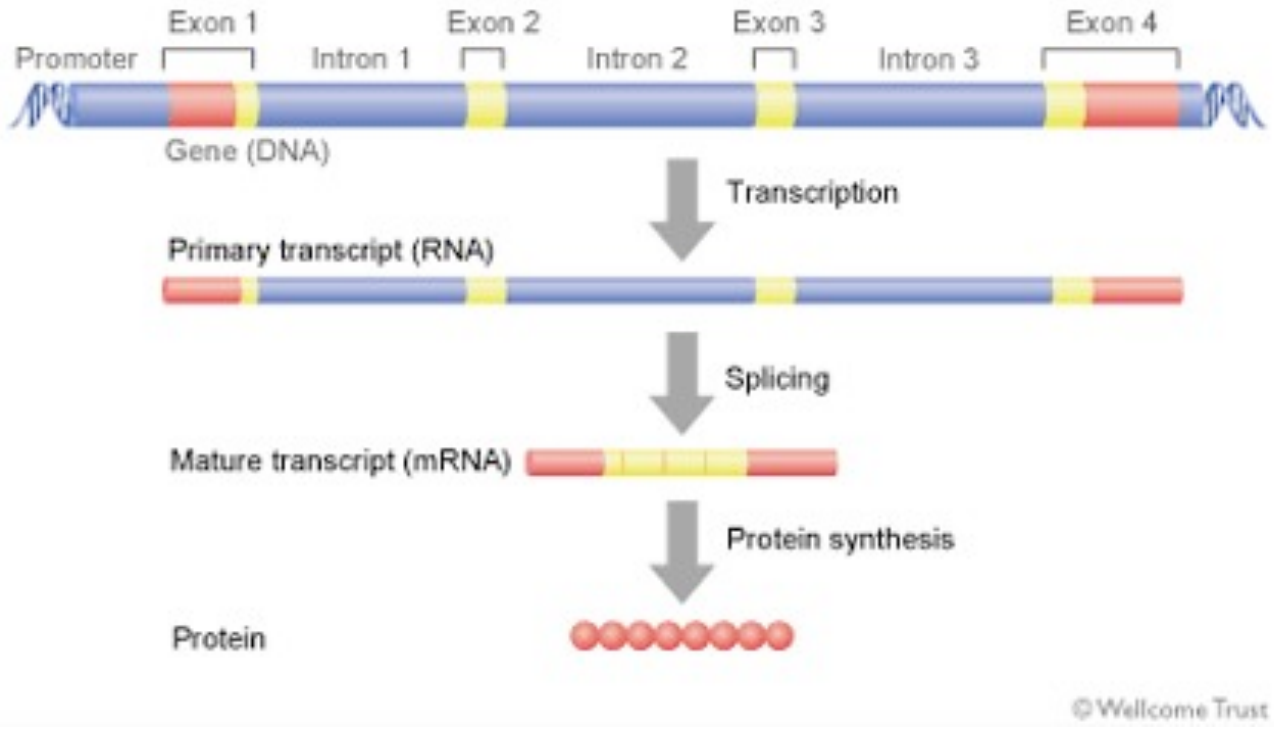
http://www.broadinstitute.org/education/glossary/chromosome

Gene

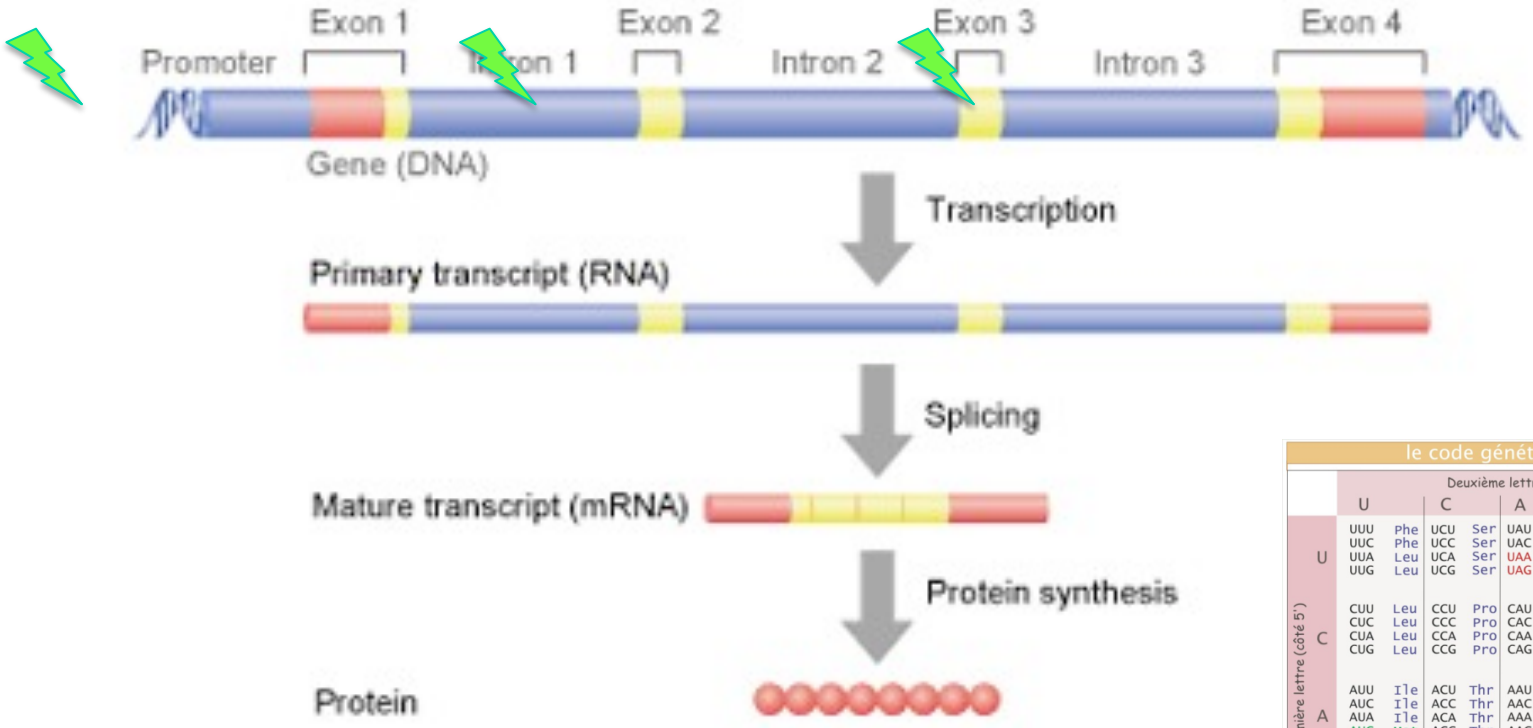
2 copies of each gene

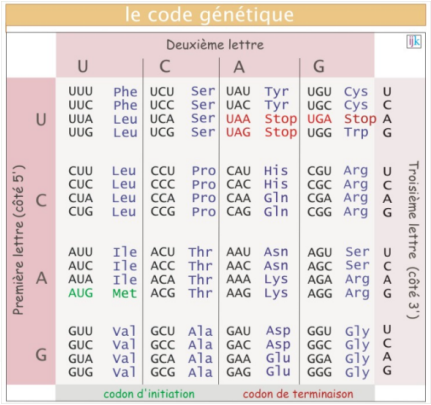
20 000 – 25 000 genes in the human genome

Exons (encoding for proteins) / introns

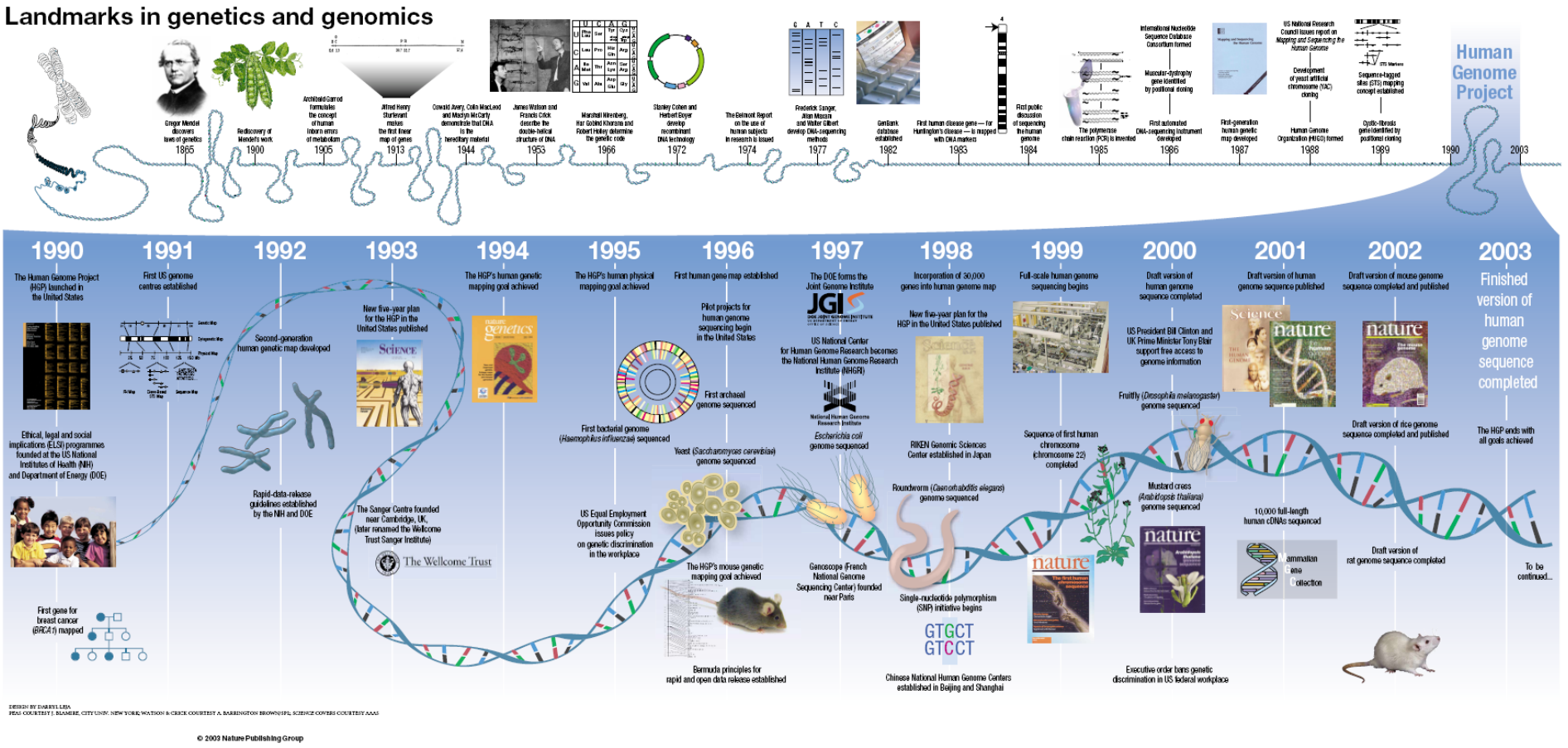


DNA Mutations





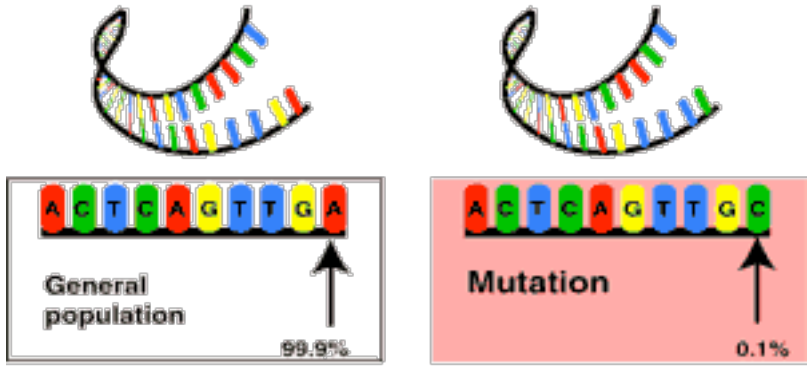
1865-2003 : From Mendel to the human genome sequence



http://www.nature.com/nature/journal/v422/n6934/pdf/timeline_01626.pdf

Genome variation

- Humans are 99.9% identical
- There are differences in the sequence of DNA between individuals (alleles)
- Variations can be common (polymorphisms), rare or private
- Variations can be benign, contribute to phenotypic traits, predispose to disease, or cause disease
- There are different types of variations : substitutions, small insertions/ deletions, copy number variants…



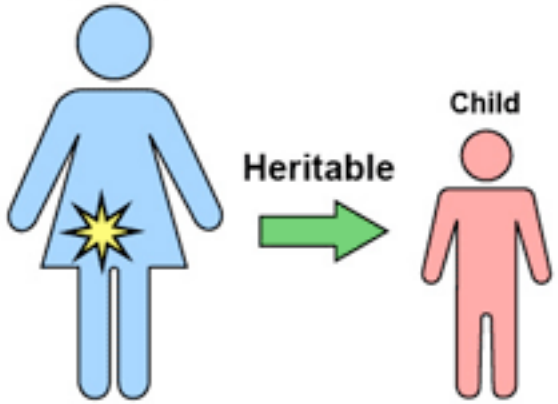
Mutations : somatic vs germline

Somatic mutations
- Occur in ***nongermline*** tissues
- Cannot be inherited

Nonheritable

Mutation in tumor only
(for example, breast)

Germline mutations
- Present in egg or sperm
- Can be inherited
- Cause cancer family syndrome

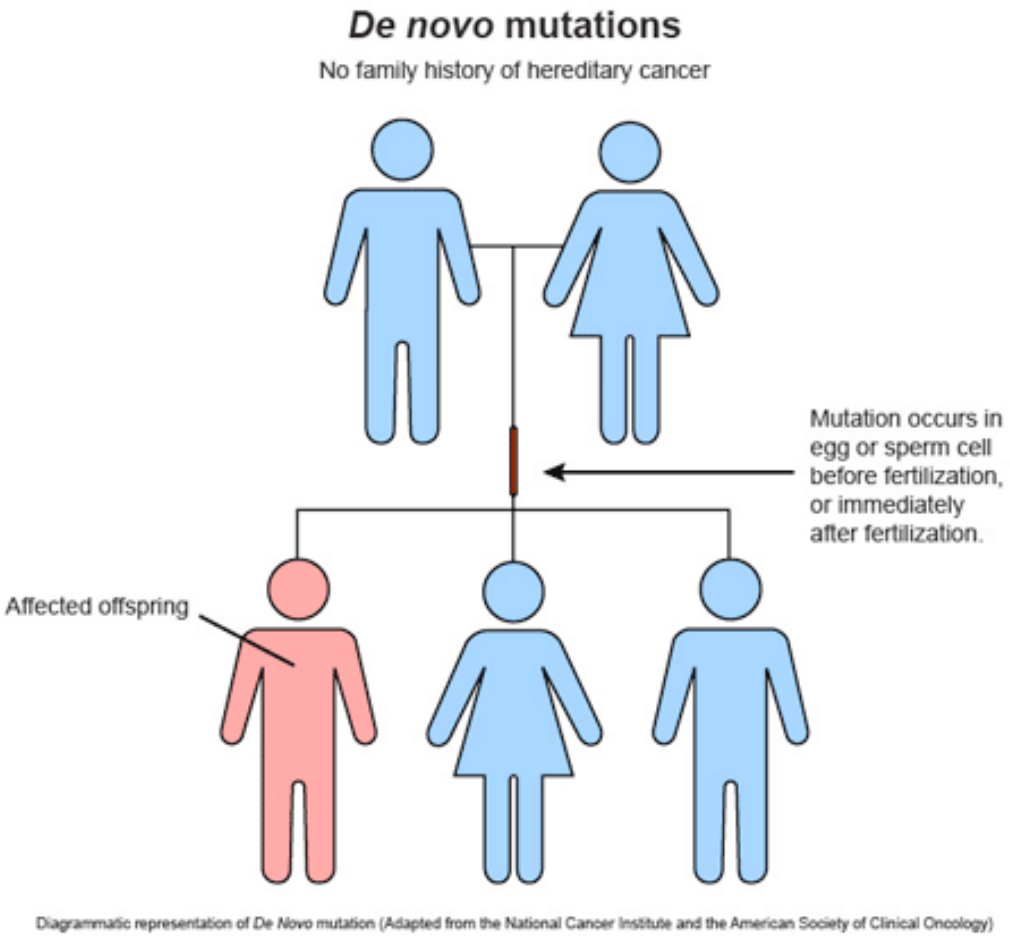


Mutation in
egg or sperm

All cells
affected in
offspring

Adapted from the National Cancer Institute and the American Society of Clinical Oncology

Mutations : *de novo*



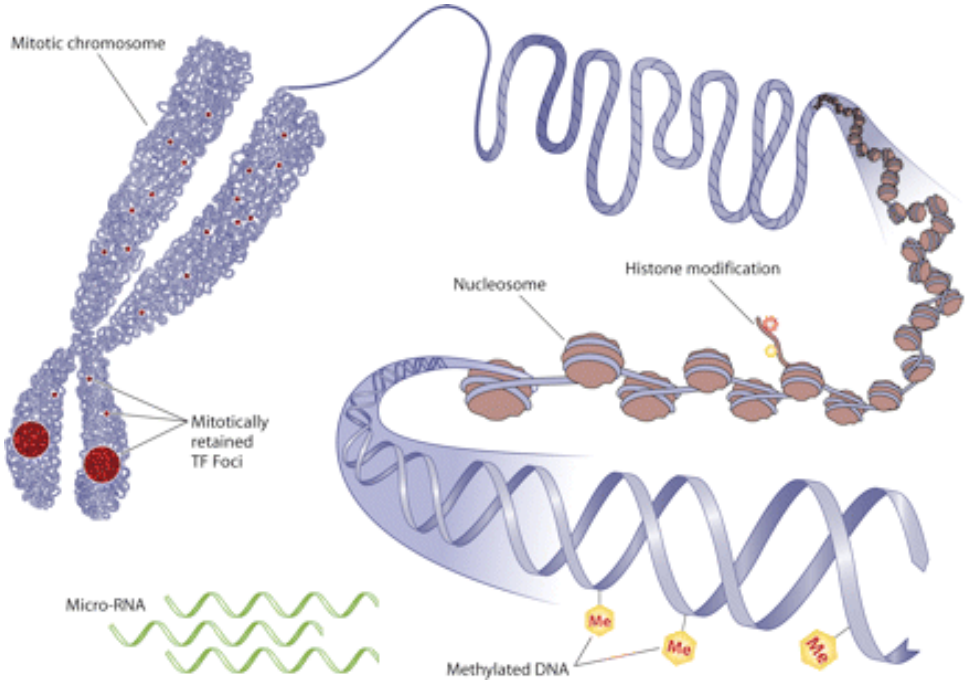
Diagrammatic representation of *De Novo* mutation (Adapted from the National Cancer Institute and the American Society of Clinical Oncology)

=> Interactive session

Epigenetics

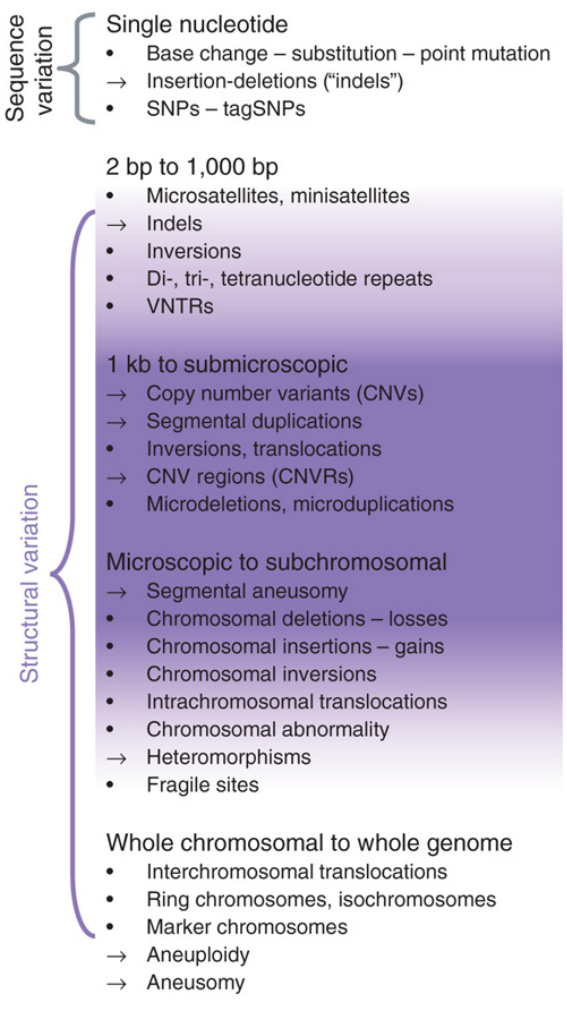
DNA methylation, chromatin remodeling, microRNAs…

- Inherited or somatic
- Role in gene expression
- Tissue dependent



http://mcb.asm.org/content/30/20/4758/F1.expansion.html

Detection of mutations



Sequencing

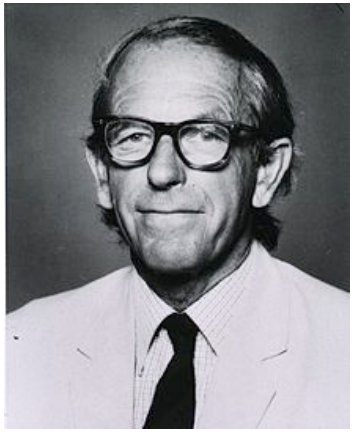
PCR
Southern Blot

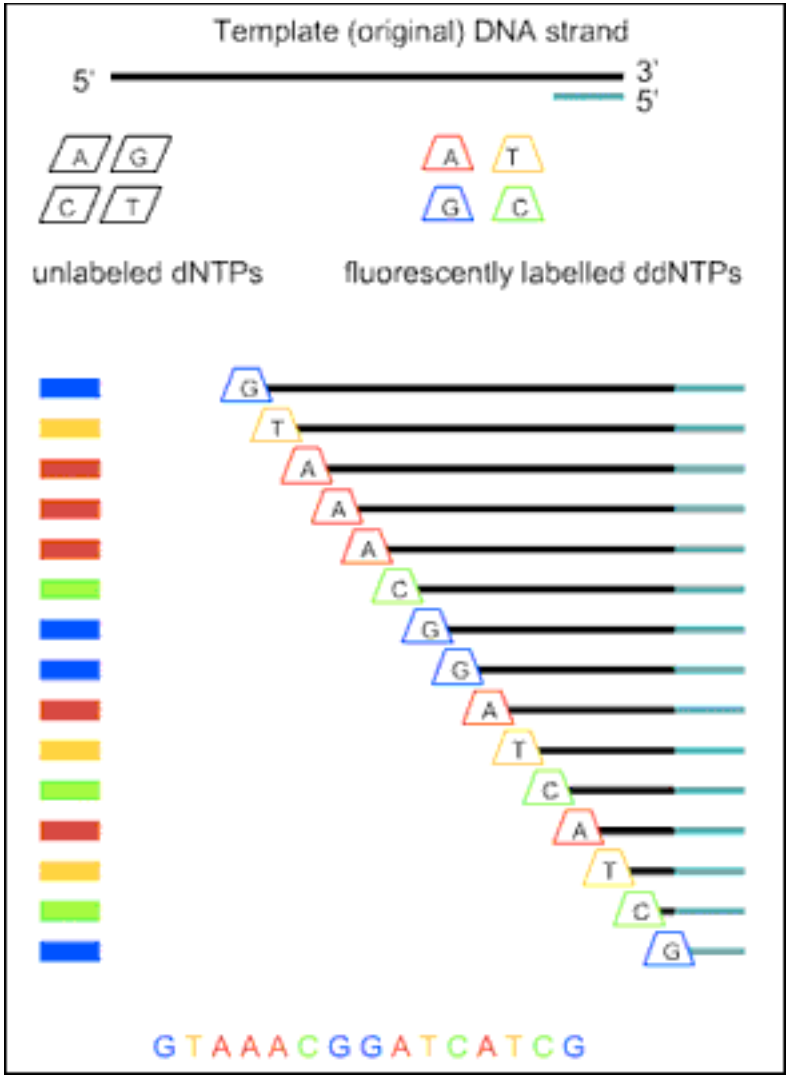
SNP array/array-CGH
Sequencing

Cytogenetics
NGS sequencing

Nature Genetics 39, S7 - S15 (2007)

Sanger sequencing





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

Next generation sequencing (NGS)

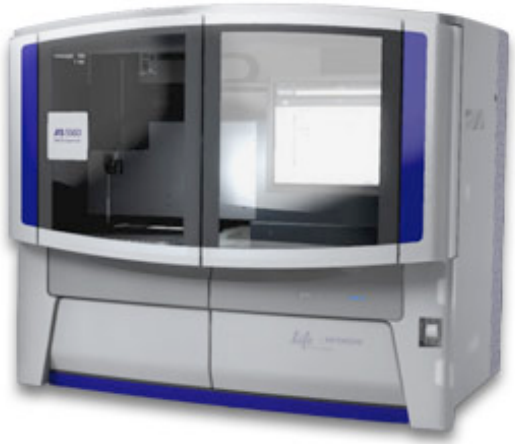
2005

Roche 454 SEQUENCING

Roche GS FLX
(« pyrosequencing »)

2006-2007

life technologies™



SOLiD 5500

2006-2007

illumina®

HiSeq

Next generation sequencing (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

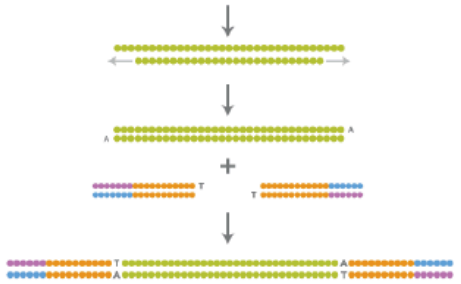
cheaper

longer sequences

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

Overview

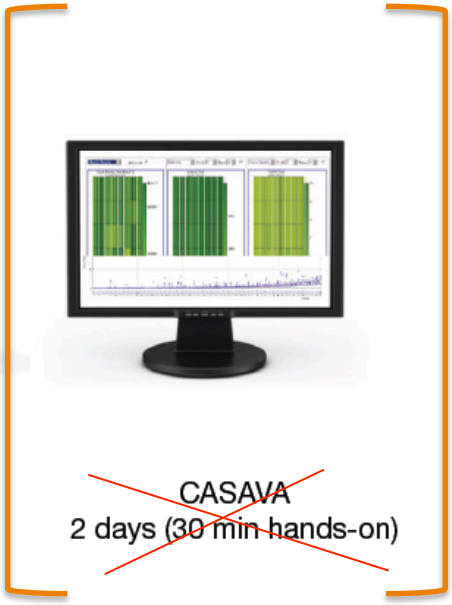
Figure 3: Next-Generation Sequencing Simplified



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)

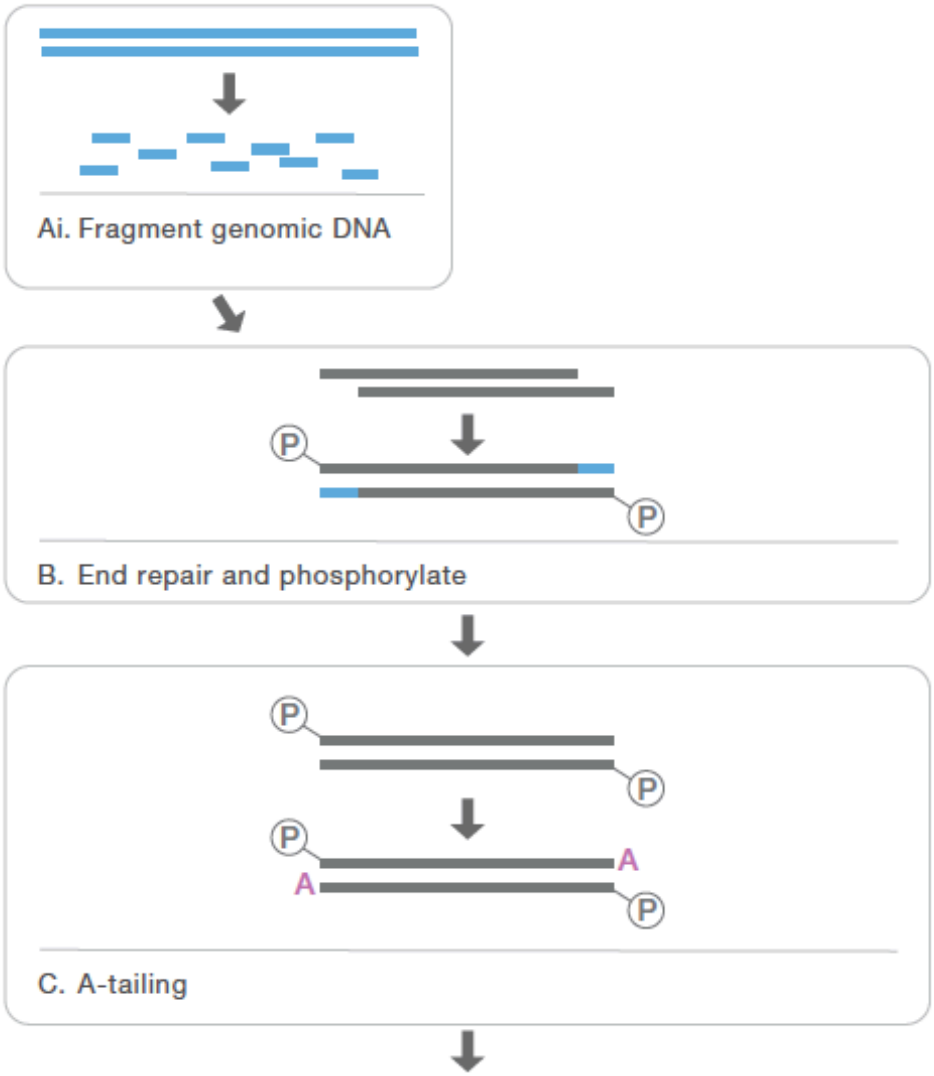
From simplified sample preparation kits and automated cluster generation, to streamlined sequencing by synthesis and complete data analysis, Illumina HiSeq sequencing systems offer the industry's simplest next-generation sequencing workflow.

Library preparation **Cluster generation** **Sequencing => reads** **Bioinformatics +++**

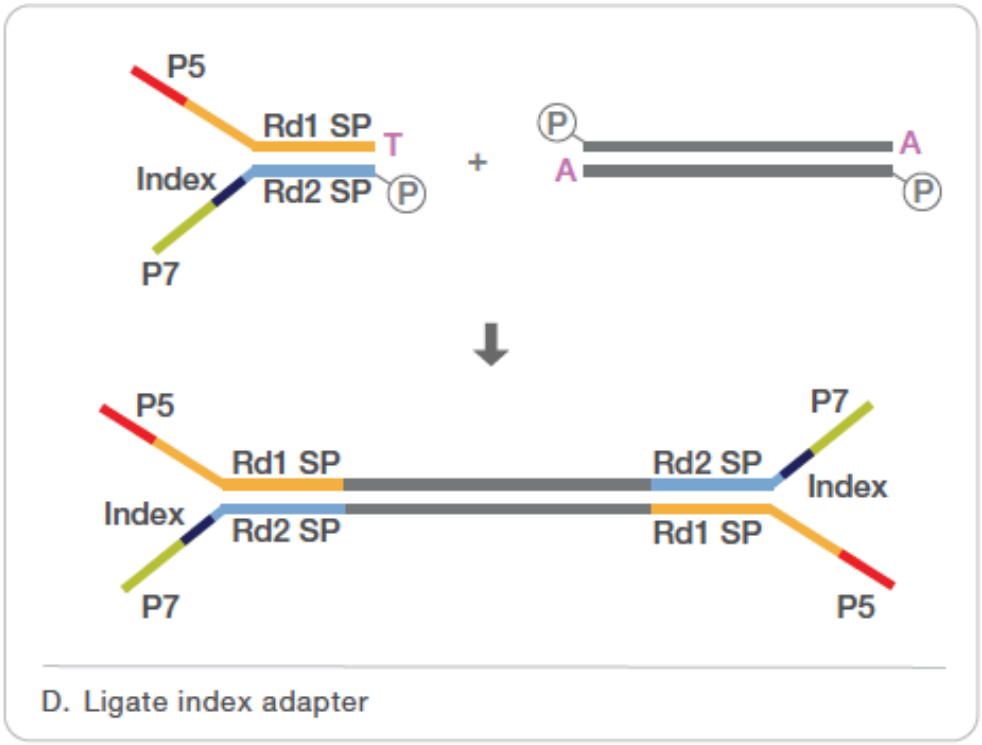
Inserm cnrs UNIVERSITÉ DE NANTES CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

l'institut du thorax

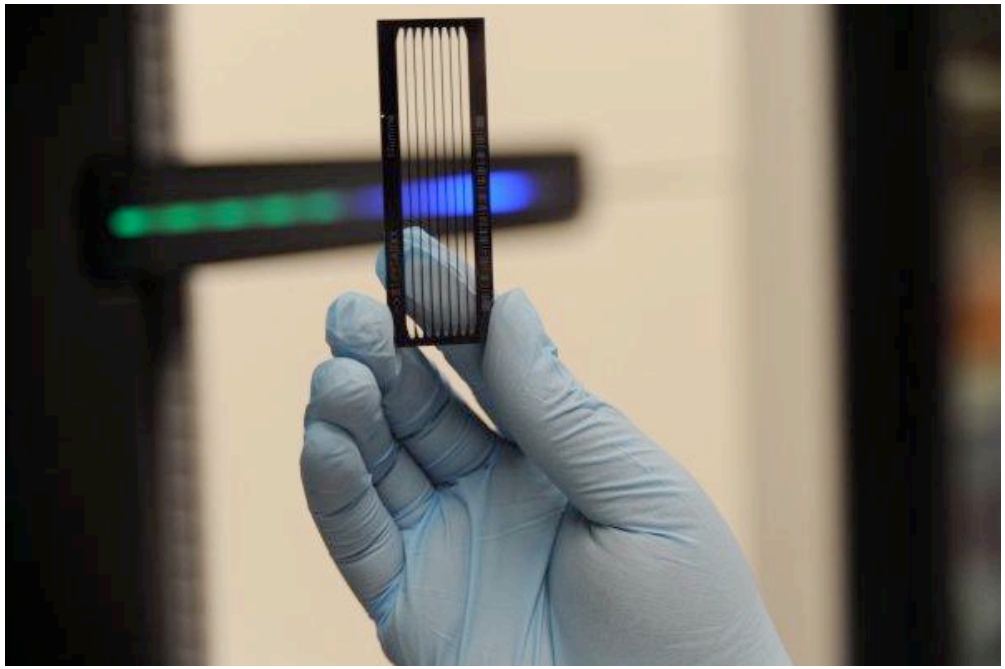
1/ Library preparation



1/ Library preparation

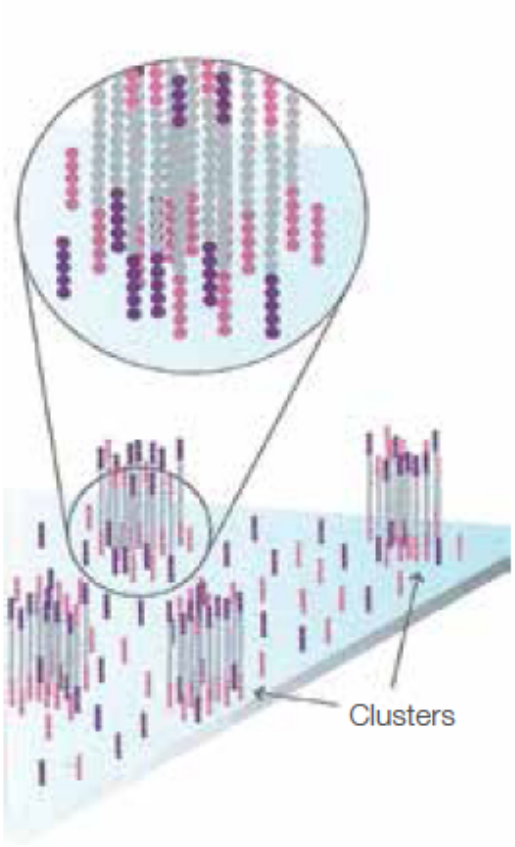


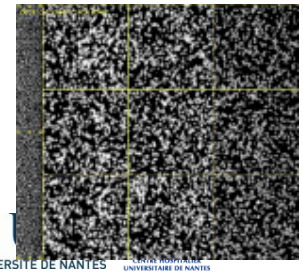
Flow cell



1 flow cell
=8 lanes

2/ Cluster generation





100-200 millions of clusters/lane

Inserm cnrs UNIVERSITÉ DE NANTES

l'institut du thorax

2/ cBOT

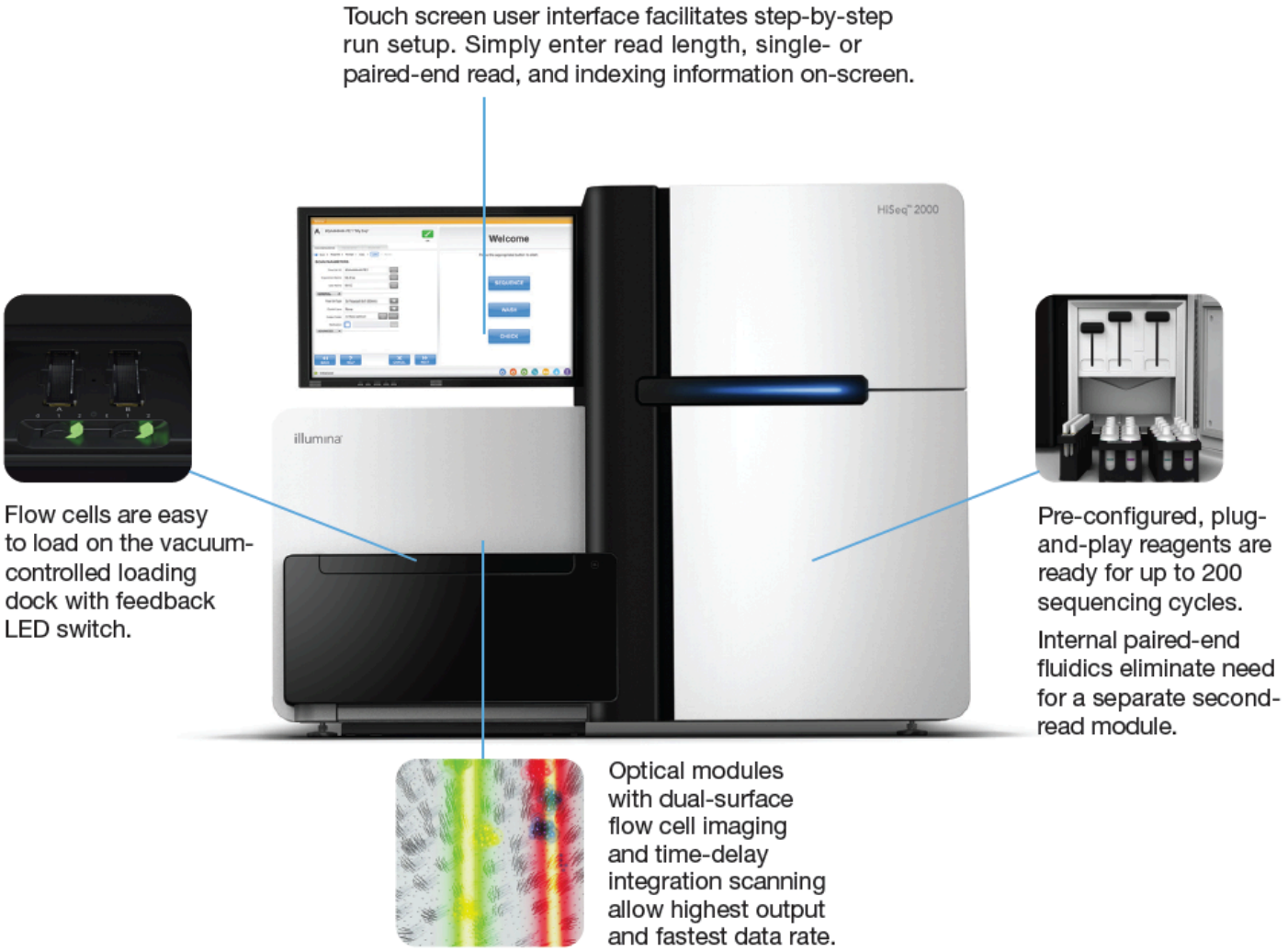
Figure 2: New cBot Features Enables Rapid and Streamlined Cluster Generation

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.

Inserm cnrs UNIVERSI

l'institut du thorax

3/ HiSeq

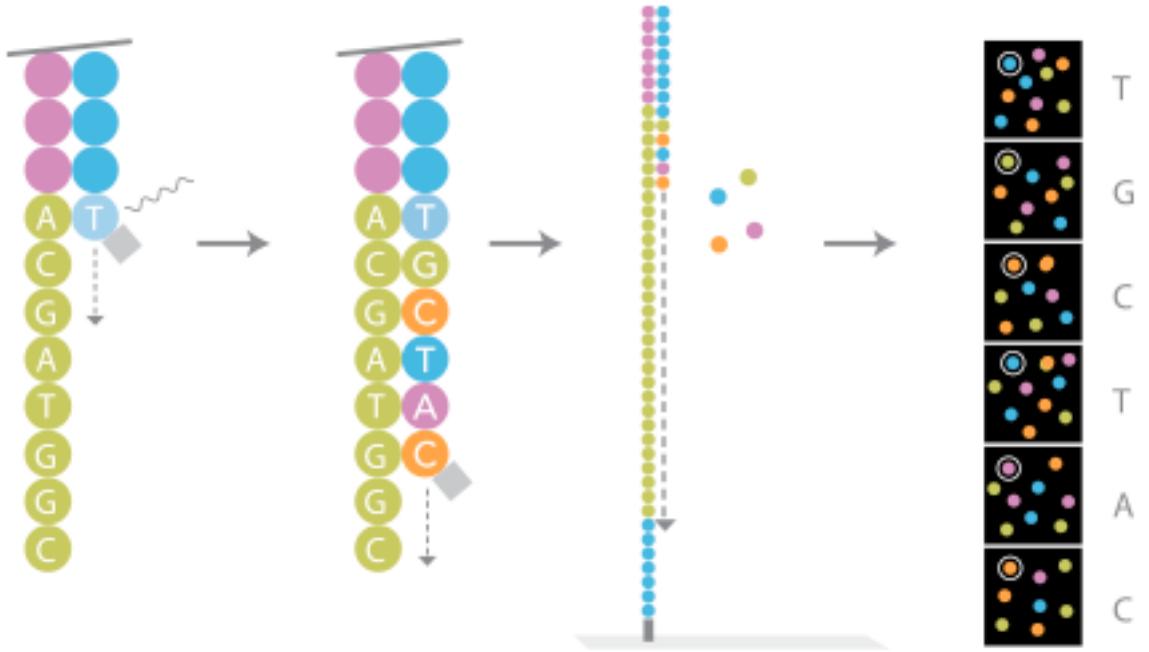

Touch screen user interface facilitates step-by-step run setup. Simply enter read length, single- or paired-end read, and indexing information on-screen.
Welcome
HiSeq™ 2000
illumina
Flow cells are easy to load on the vacuum-controlled loading dock with feedback LED switch.
Pre-configured, plug-and-play 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/ Sequencing

Cycle 1

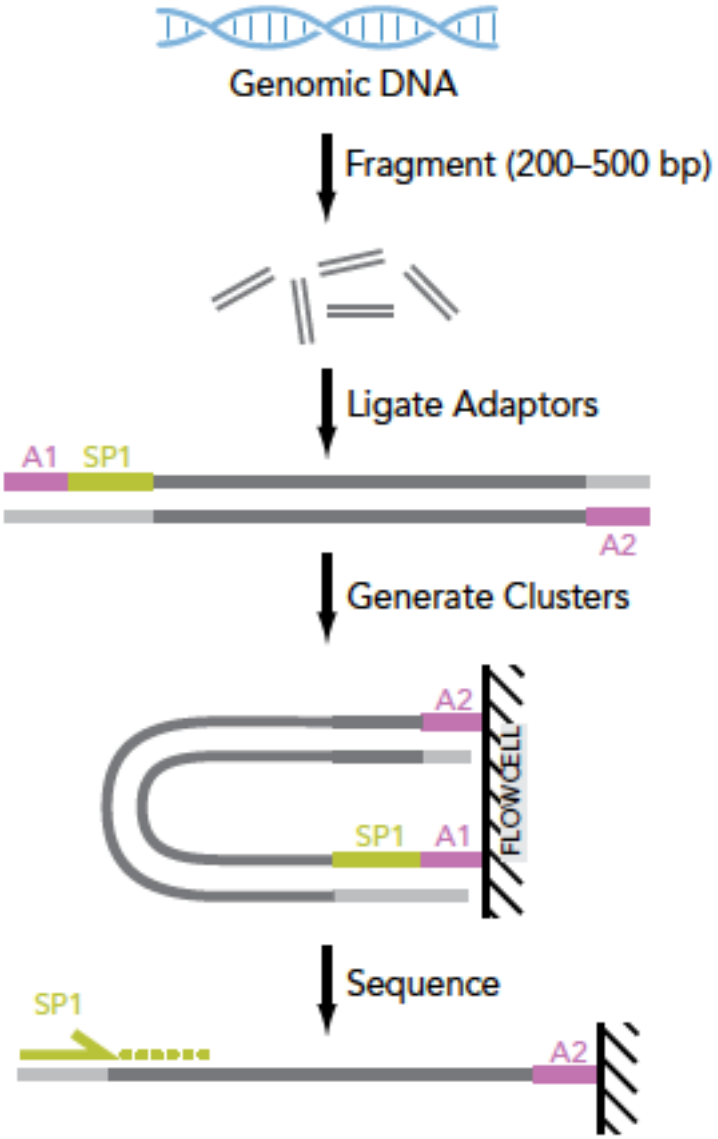
3/ Sequencing

Cycle 1 … Cycle 6 …

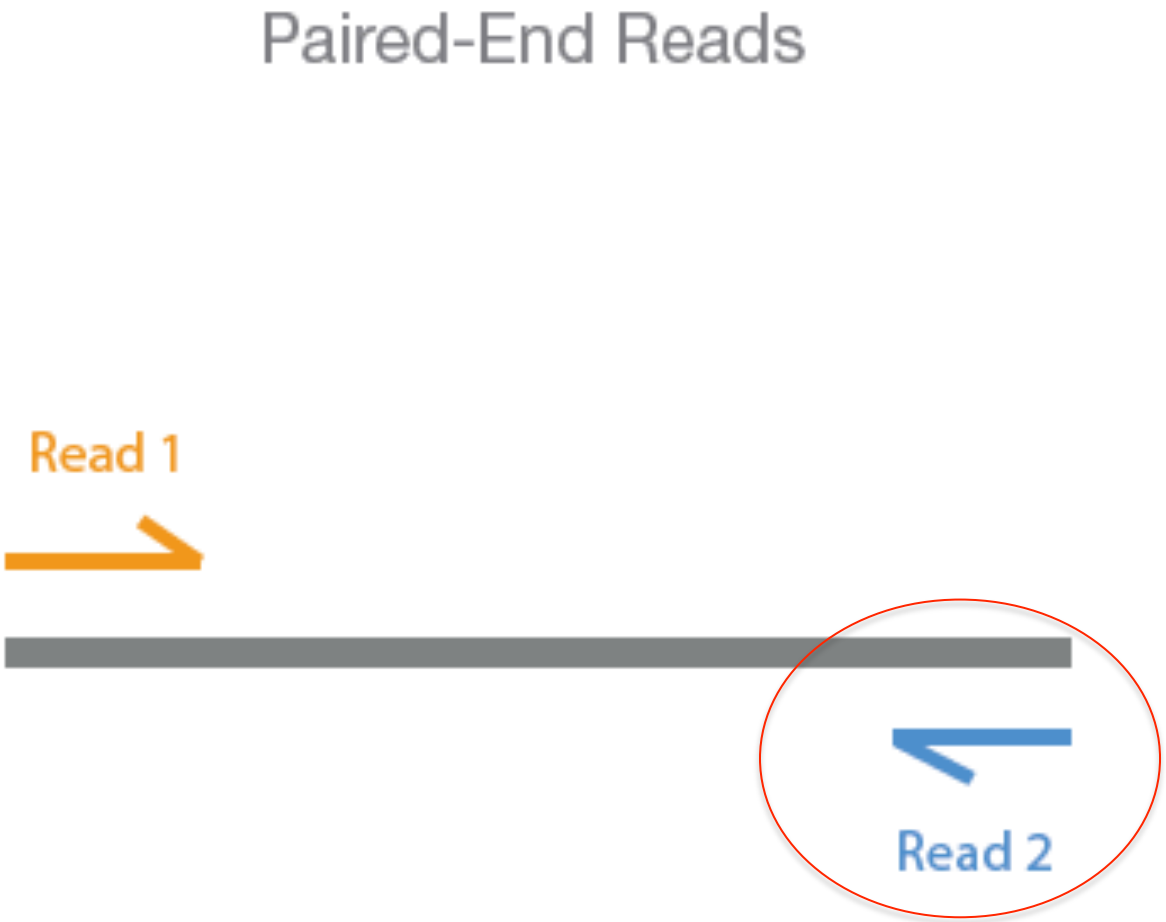


➔ reads
(e.g. 100 bp)
Each base is assigned a quality score

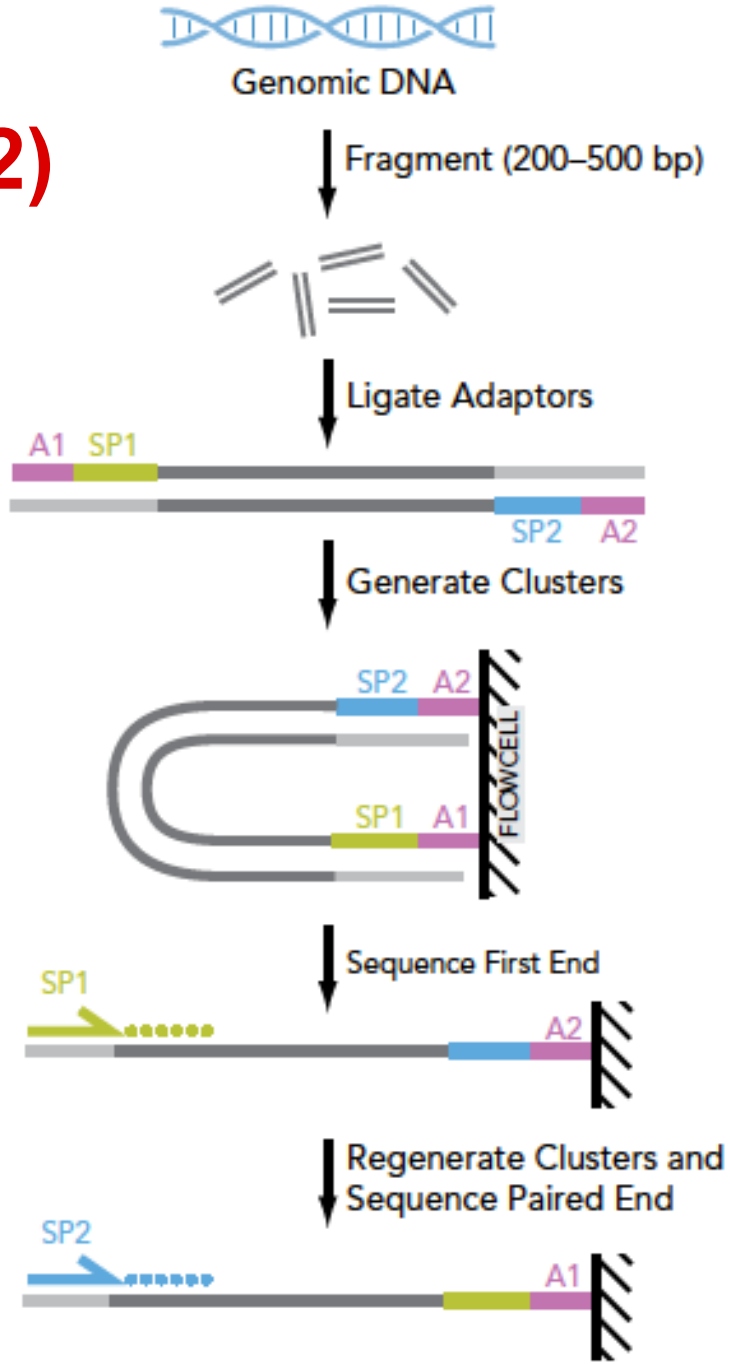
Single-end sequencing



Paired-end sequencing (1)

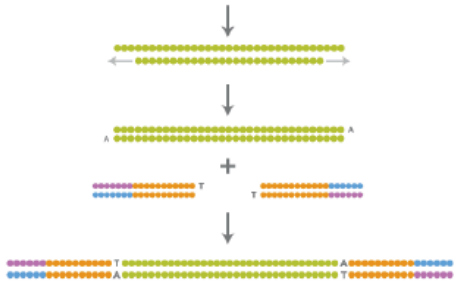


Paired-end sequencing (2)



Overview

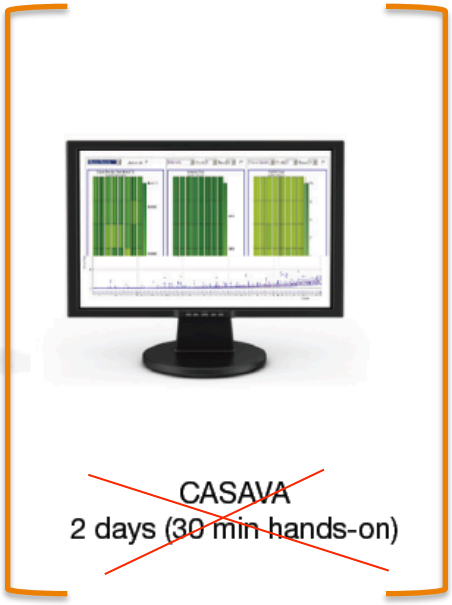
Figure 3: Next-Generation Sequencing Simplified



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



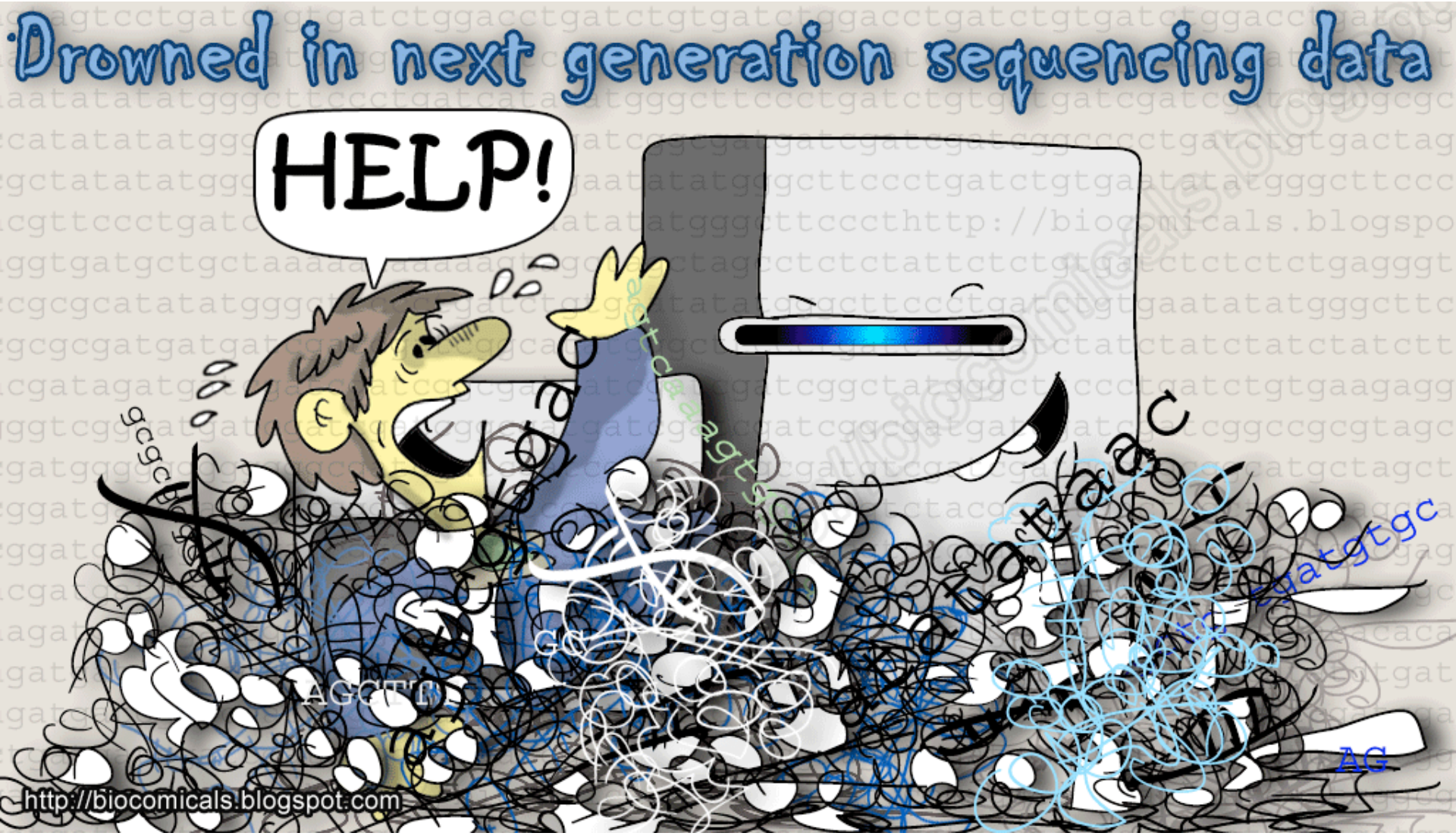
~~CASAVA~~
2 days (~~30 min~~ hands-on)

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.

Library preparation **Cluster generation** **Sequencing => reads** **Bioinformatics +++**

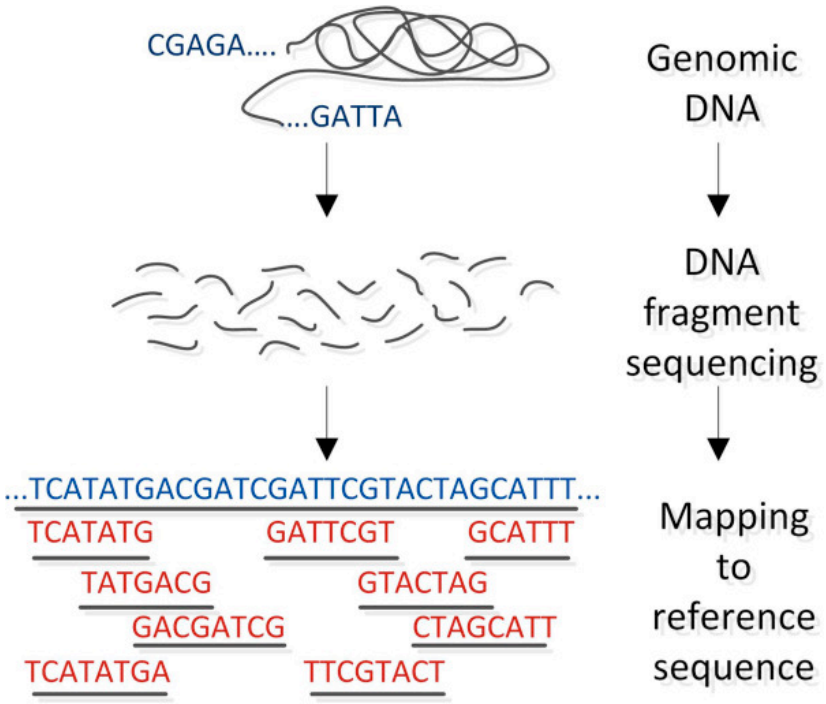
Inserm cnrs UNIVERSITÉ DE NANTES CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

l'institut du thorax

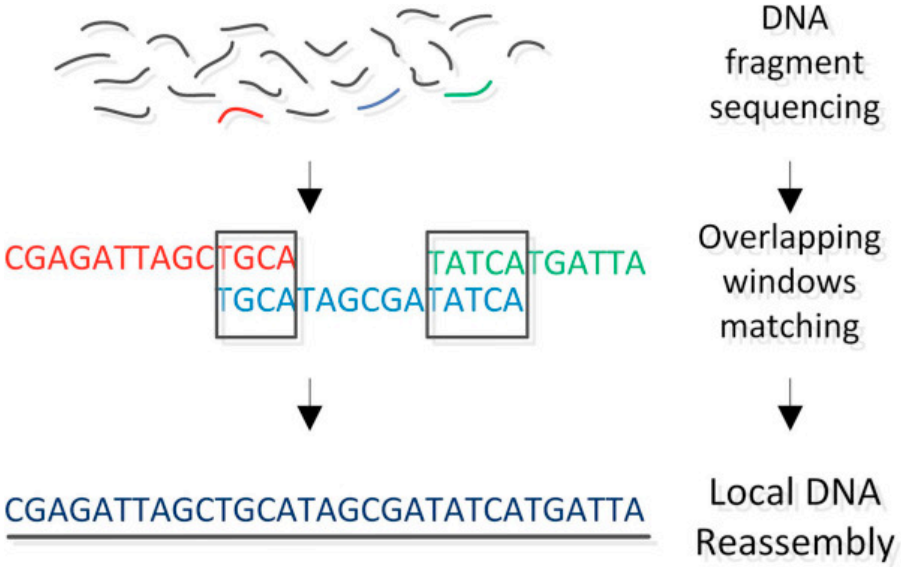

Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com

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

Reference genome => alignment

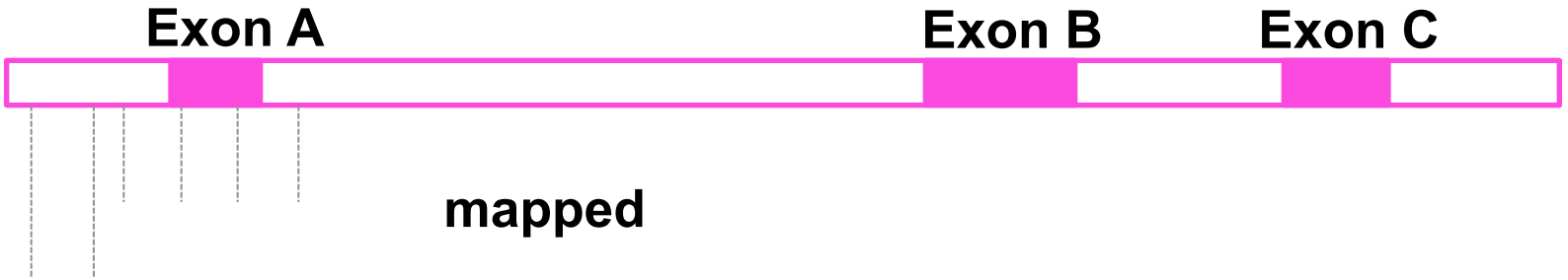


de novo assembly

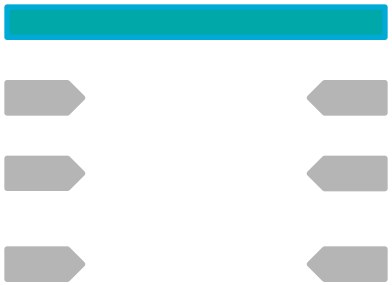


Pavlopoulos et al., BioData Mining 2013, 6:13

? unmapped



Captured fragment



Fastq file (raw data)

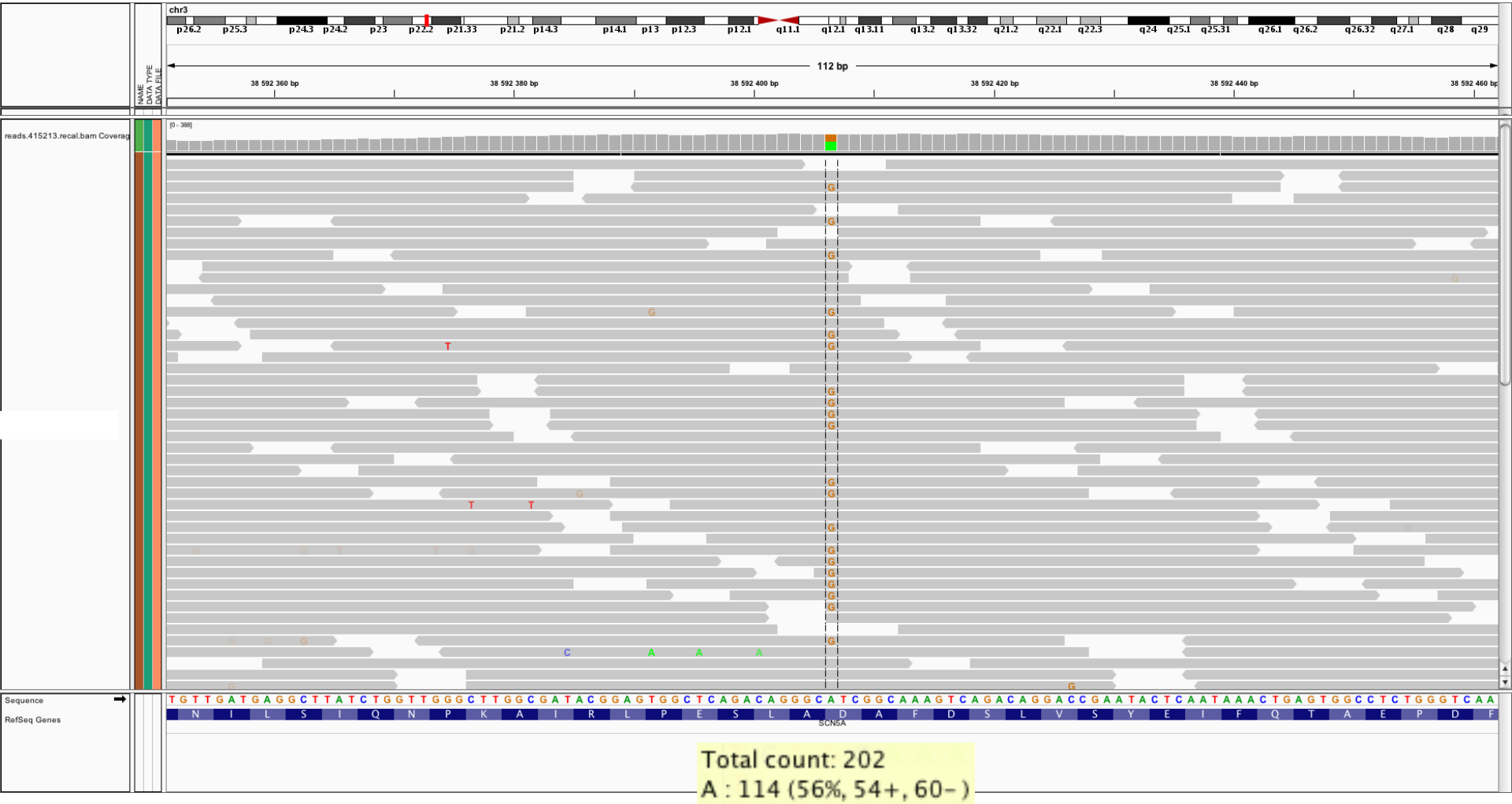
@SEQ_ID

GATCTTCTGTGACTGGAAGAAAATGTGTTACATATTACATTTCTGTCCCCATTG

+

E?=EECE<EEEE98EEEEAEEBD??BE@AEAB><EEABCEEDEC<<EBDA=DEE

IGV



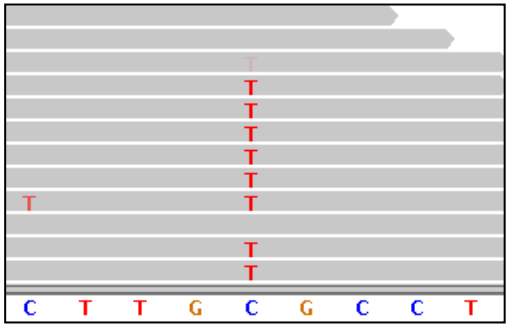
Inserm cnrs UNIVERSITÉ DE NANTES CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

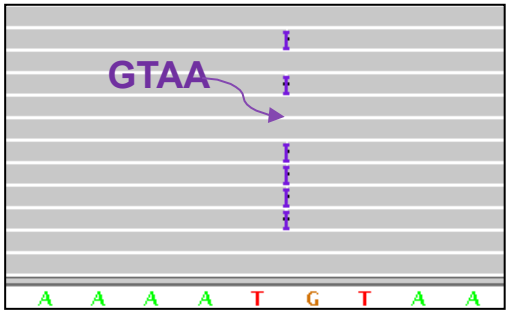
l'institut du thorax

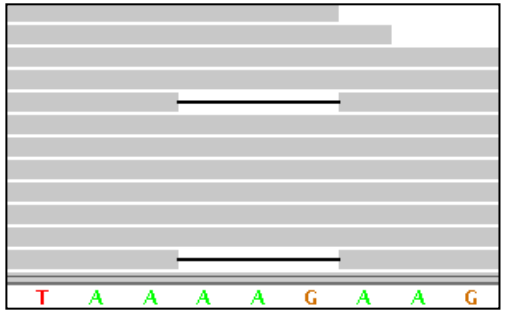
Table 1: Flexible Paired Sequencing Provides Optimal Detection Of Any Variant

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.



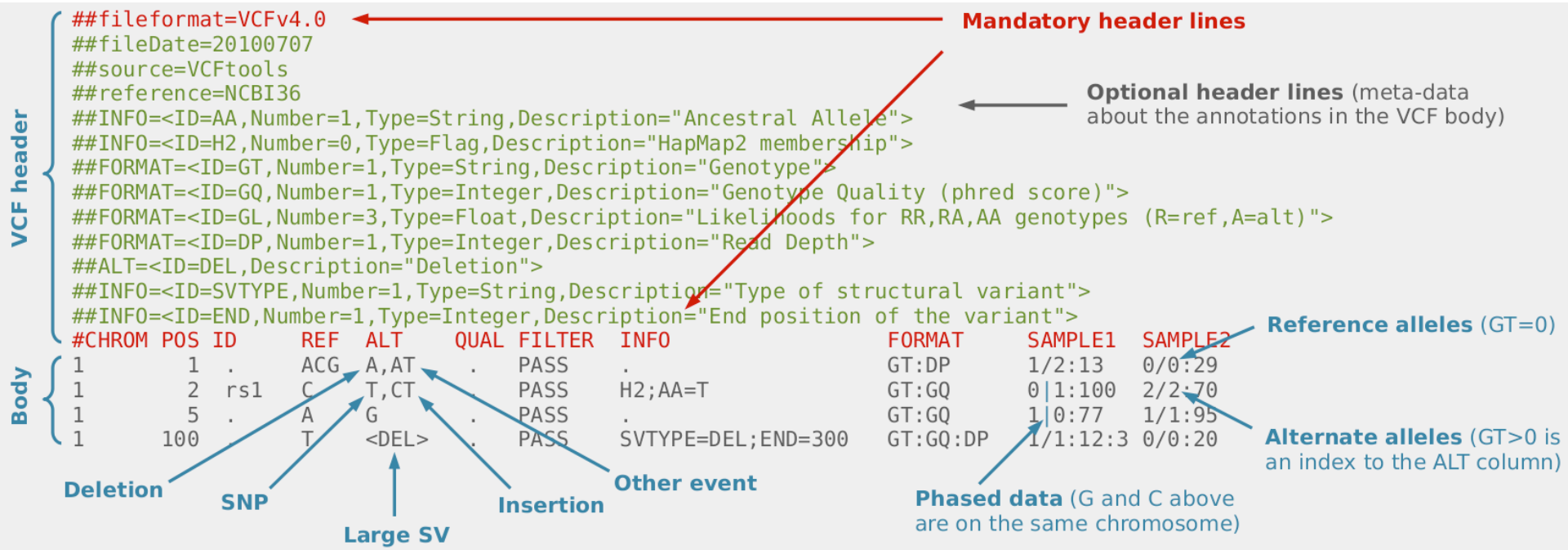




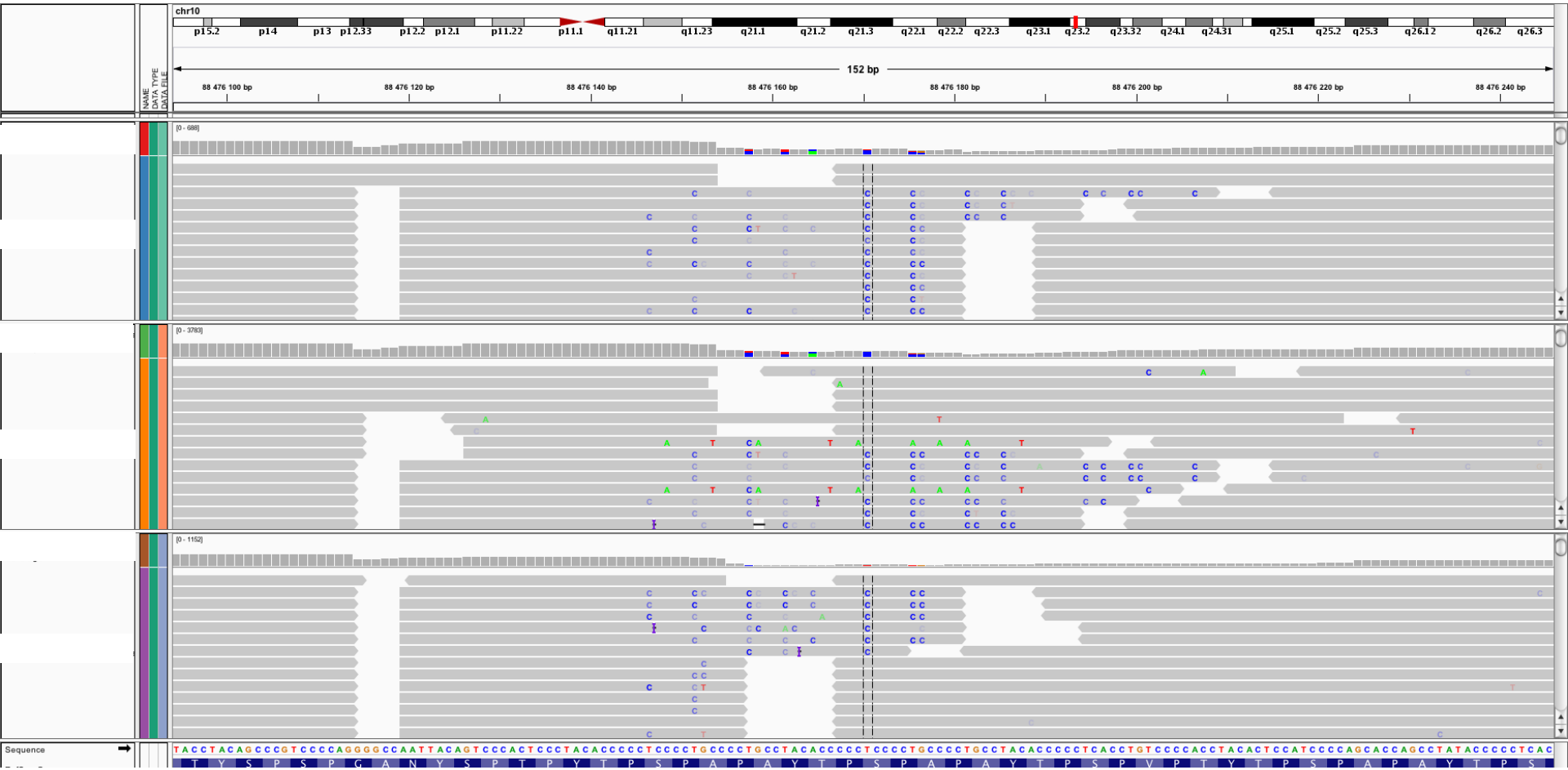
l'institut du thorax

List of genetic variants

VCF files



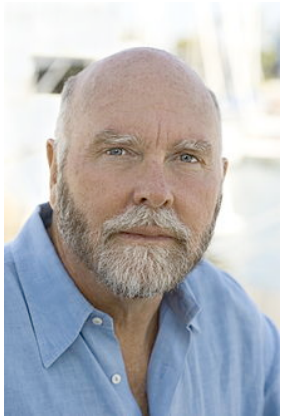
Example of false positive

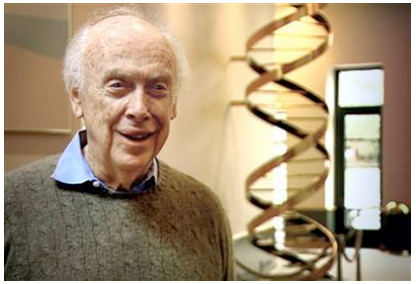


Sequencing personal genomes

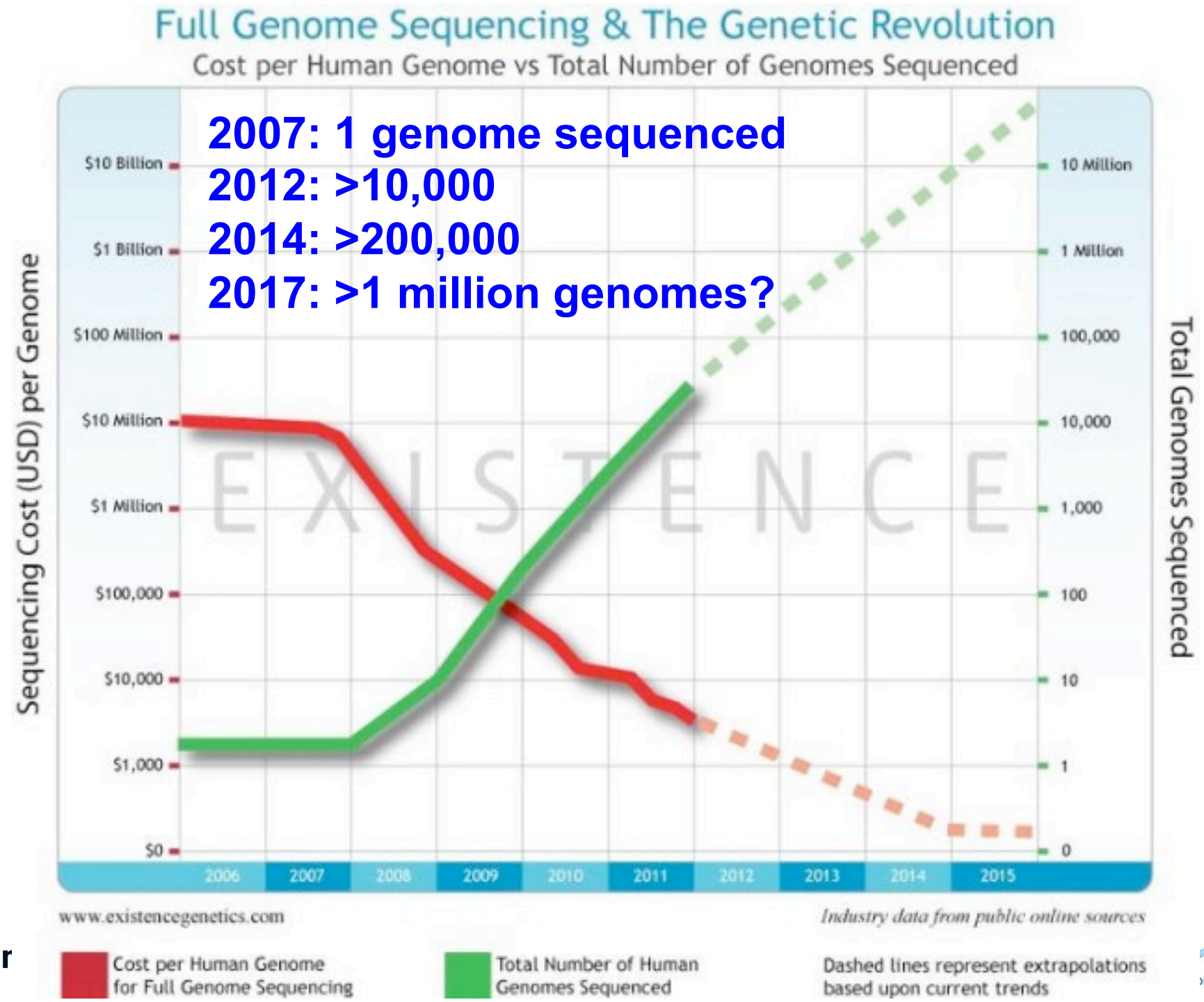
QUICKER, SMALLER, CHEAPER

Genome sequenced (publication year)	HGP (2003)	Venter (2007)	Watson (2008)
	Sanger	**Sanger**	**454**
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





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 genome sequenced
2012: >10,000
2014: >200,000
2017: >1 million genomes?
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

2014

1 HiSeq X Ten : 18,000 human genomes per year
« World's First $1000 Human Genome »

illumina

Log in to get personalized account information. Quick Order View Cart

Contact Us MyIllumina Tools

APPLICATIONS SYSTEMS INFORMATICS CLINICAL SERVICES SCIENCE SUPPORT COMPANY Search

Systems / HiSeq X Ten

Subscribe | Follow us: | Select Language

Overview System Specs Kits Support REQUEST PRICING

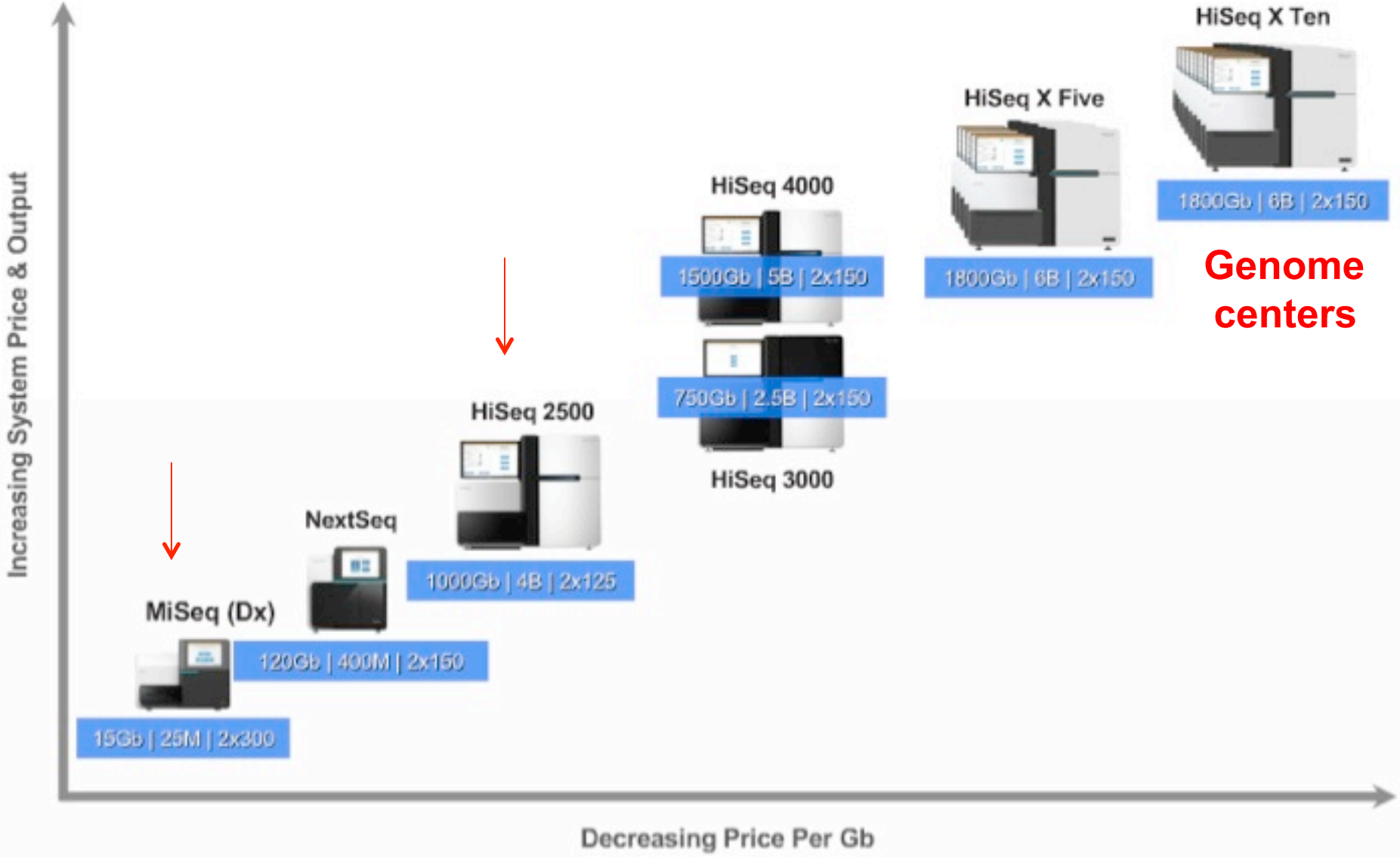
Population power. Extreme throughput. $1,000 human genome.

The HiSeq X Ten is a set of ten ultra-high-throughput sequencers, purpose-built for large-scale human whole-genome sequencing.

Inserm cnrs UNIVERSITÉ DE NANTES CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

l'institut du thorax

Sequencing Power For Every Scale.

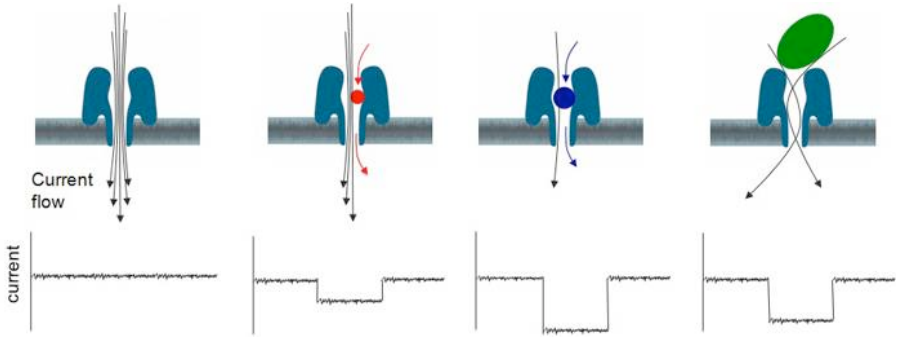


http://blog.genohub.com/illuminas-latest-release-hiseq-3000-4000-nextseq-550-and-hiseq-x5/

3rd generation sequencing

MinION: A complete DNA sequencer on a USB stick

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

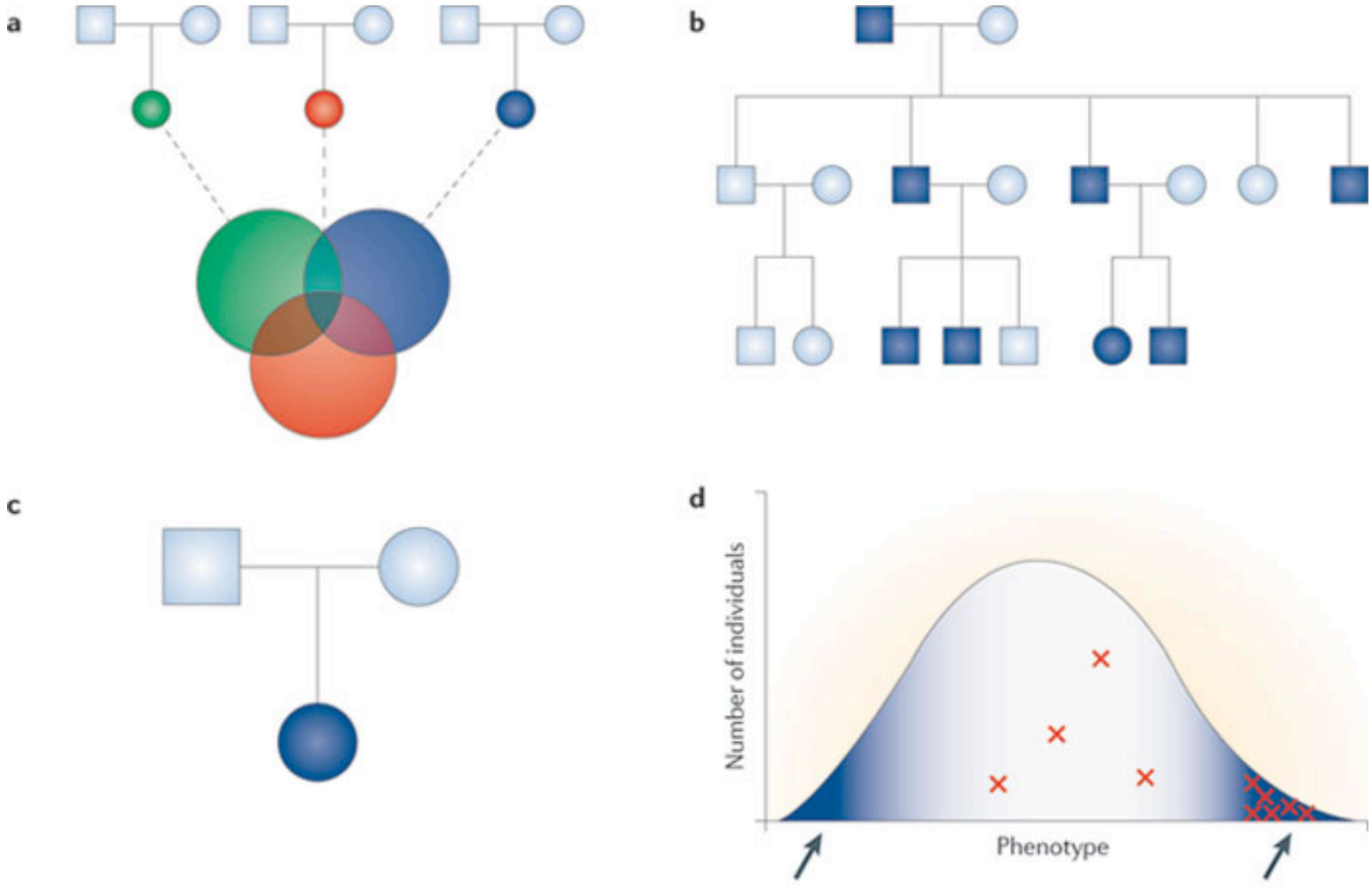


Helicos Bioxciences
Pacific Biosciences
Oxford Nanopore
…

« An exciting time for genomics »

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

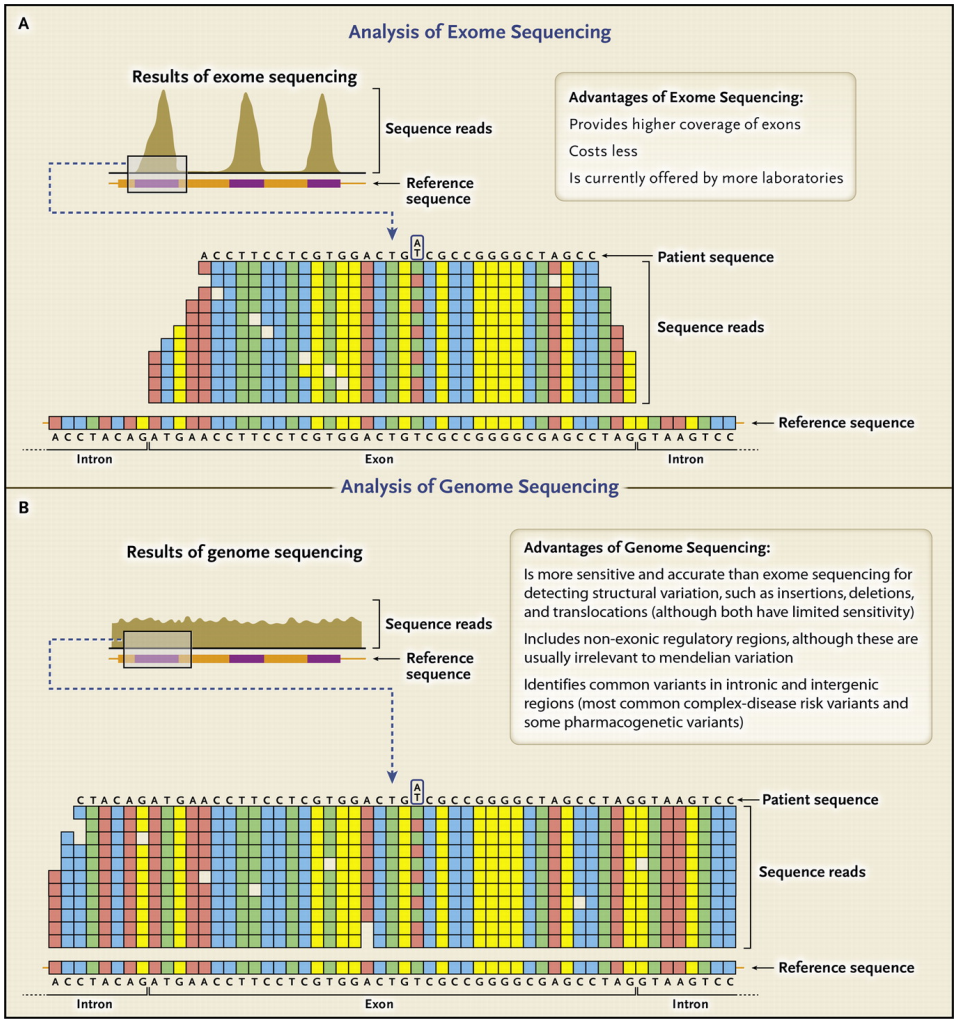
Disease causing variants?



Nature Reviews | Genetics

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

Exome vs whole genome sequencing



Biesecker LG, Green RC. N Engl J Med 2014;370:2418-2425.

Exome sequencing : cheaper, but…

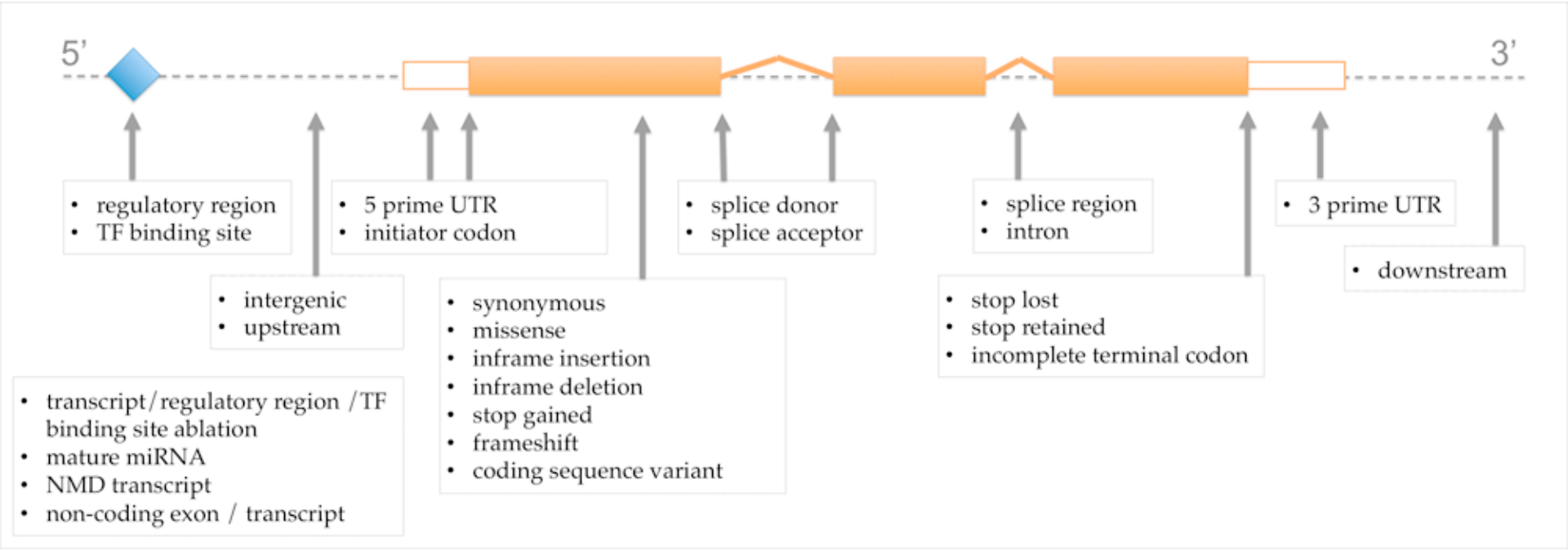
Table 1. DNA Variant Types Currently Not Well Detected or Undetectable by Clinical Genome and Exome Sequencing (CGES), with Examples of Phenotypes Associated with Such Variants.

Variant Type	Associated Phenotype or Phenotypes
Repetitive DNA, including tri-nucleotide repeats	Fragile X syndrome, Huntington's disease
Copy-number variants	DiGeorge syndrome (22q11.2 deletion syndrome), Charcot–Marie–Tooth disease type 1A
Long insertion–deletion variants*	Resistance to human immunodeficiency virus infection
Structural variants	Chromosomal translocations associated with spontaneous abortions
Aneuploidy	Down's syndrome, Turner's syndrome
Epigenetic alterations	Prader–Willi syndrome, Beckwith–Wiedemann syndrome

* There is a spectrum of genomic variants, from nucleotide insertions and deletions of at least 8 to 10 bp through copy-number variants, that are less effectively assayed by current CGES technology.

Biesecker LG, Green RC. N Engl J Med 2014;370:2418-2425.

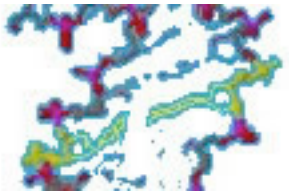
Functional annotations



For missense variants : prediction tools (SIFT, PolyPhen2…)

Databases of genetic variants

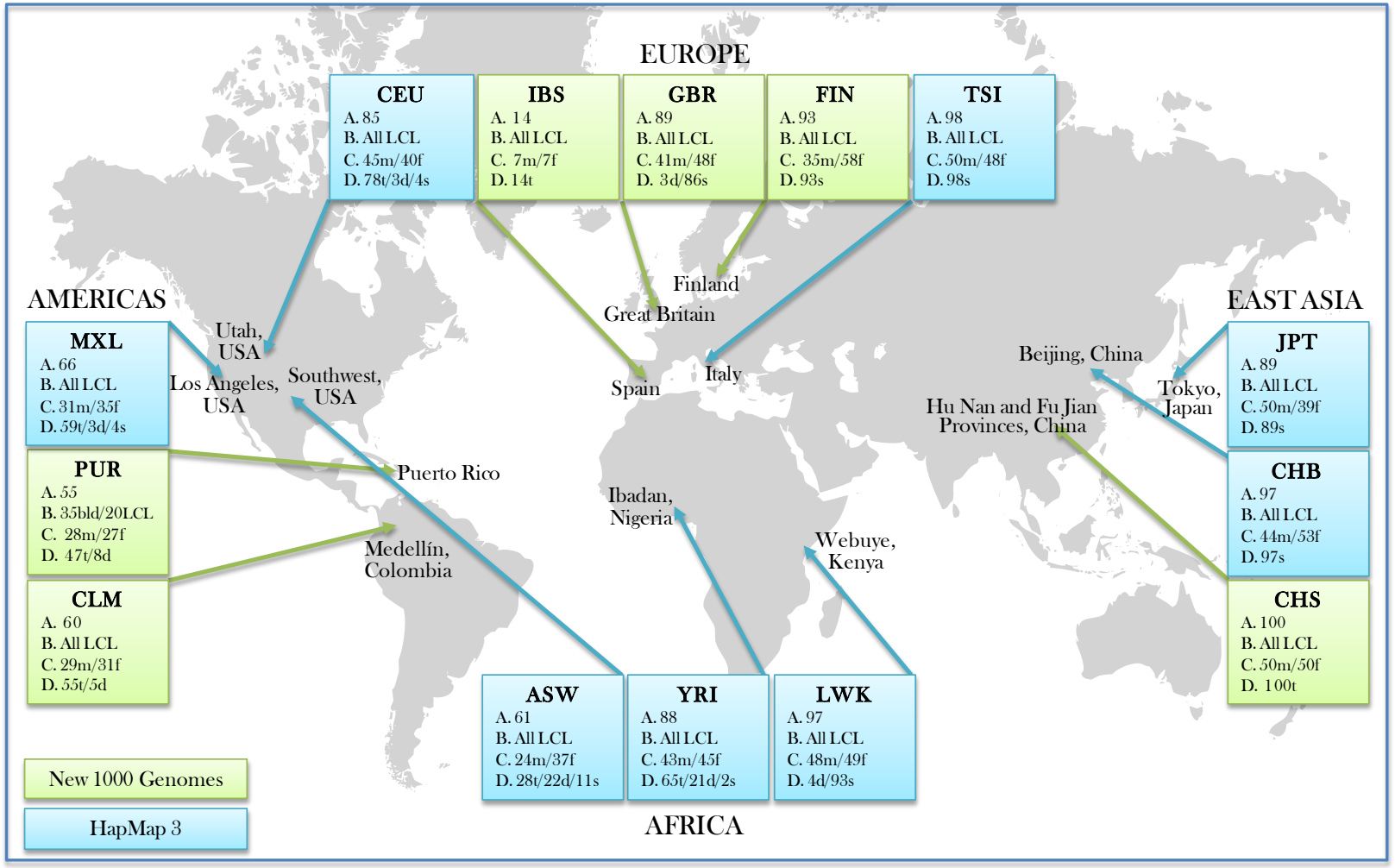
dbSNP
Short Genetic Variations



1000 Genomes
A Deep Catalog of Human Genetic Variation

NHLBI Exome Sequencing Project (ESP)
Exome Variant Server

Phase 1 populations

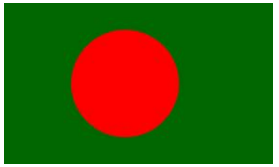


Phase 2/3 populations

Barbados

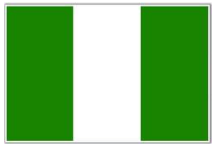
Ghana

Pakistan



Bangladesh

Peru



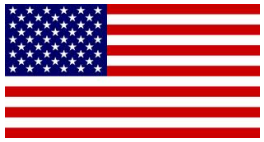
Nigeria

India

Sierra Leone

Sri Lanka

Vietnam



USA

1000 Genomes

A Deep Catalog of Human Genetic Variation

Human ▼ | Location: 6:74,125,388-74,126,388 | Variation: rs311685

Variation displays

- Explore this variation
- Genomic context
  - Gene/Transcript (3)
- **Population genetics (51)**
- Individual genotypes
  - Ensembl (3673)
  - 1000Genomes (1092)
- Linkage disequilibrium
- Phenotype Data
- Phylogenetic Context
- Flanking sequence
- External Data

Configure this page

Manage your data

Export data

Get VCF data

Bookmark this page

View in Ensembl

Download view as CSV

Source	dbSNP 134 - Variants (including SNPs and indels) imported from dbSNP
Alleles	Reference/Alternative: **A/G** | Ancestral: **A** | Ambiguity code: **R** | MAF: **0.48** (G)
Location	Chromosome **6:74125888** (forward strand) | View in location tab
Validation status	This variation is validated by **1000 Genomes**, **HapMap** and also cluster, doublehit, frequency
Synonyms ⊞	This feature has **12** synonyms - click the plus to show
HGVS names ⊞	This feature has 4 HGVS names - click the plus to show

1000 genomes alleles frequencies

ALL	AFR	AMR	ASN	EUR
A: 51%	A: 69%	A: 54%	A: 45%	A: 42%
G: 49%	G: 31%	G: 46%	G: 55%	G: 58%

Different allele frequencies

1000 genomes

Show/hide columns

Population	Alleles A	Alleles G	Genotypes A|A
1000GENOMES:AFR	0.689	0.311	0.463
1000GENOMES:ALL	0.507	0.493	0.268
1000GENOMES:AMR	0.539	0.461	0.293
1000GENOMES:ASN	0.446	0.554	0.199
1000GENOMES:EUR	0.420	0.580	0.182

Inserm cnrs UNIVERSITÉ DE NANTES CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

l'institut du thorax

Exome Aggregation Consortium (ExAC)

October 15, 2014

What populations are represented in the ExAC data?

Population	Male Samples	Female Samples	Total
African/African American (AFR)	1,888	3,315	5,203
Latino (AMR)	2,254	3,535	5,789
East Asian (EAS)	2,016	2,311	4,327
Finnish (FIN)	2,084	1,223	3,307
Non-Finnish European (NFE)	18,740	14,630	33,370
South Asian (SAS)	6,387	1,869	8,256
Other (OTH)	275	179	454
Total	33,644	27,062	**60,706**

Exome Aggregation Consortium (ExAC)

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
European (Finnish)	15	6614	0	0.002268
European (Non-Finnish)	124	63230	1	0.001961
Other	1	780	0	0.001282
African	4	9314	0	0.0004295
South Asian	1	9006	0	0.000111
Latino	1	11476	0	8.714e-05
East Asian	0	8474	0	0
Total	**146**	**108894**	**1**	**0.001341**

First gene identified by exome sequencing

Miller syndrome
(rare condition that mainly affects the development of the face and limbs)

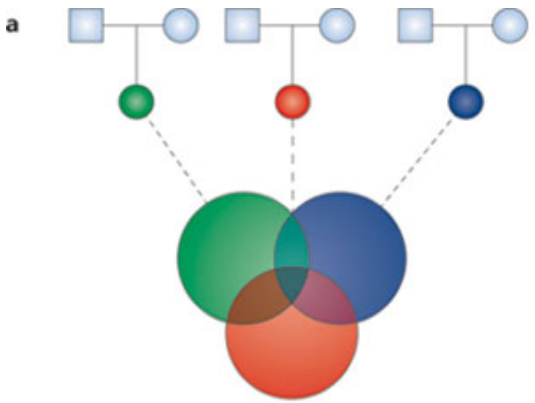
Exome sequencing identifies the cause of a mendelian disorder

Sarah B Ng[1,10], Kati J Buckingham[2,10], Choli Lee[1], Abigail W Bigham[2], Holly K Tabor[2,3], Karin M Dent[4], Chad D Huff[5], Paul T Shannon[6], Ethylin Wang Jabs[7,8], Deborah A Nickerson[1], Jay Shendure[1] & Michael J Bamshad[1,2,9]

Family 1
2 patients

Family 2
1 patient

Family 3
1 patient



Ng et al, Nat Genet. 42(1):30-5 (2010)

Data analysis

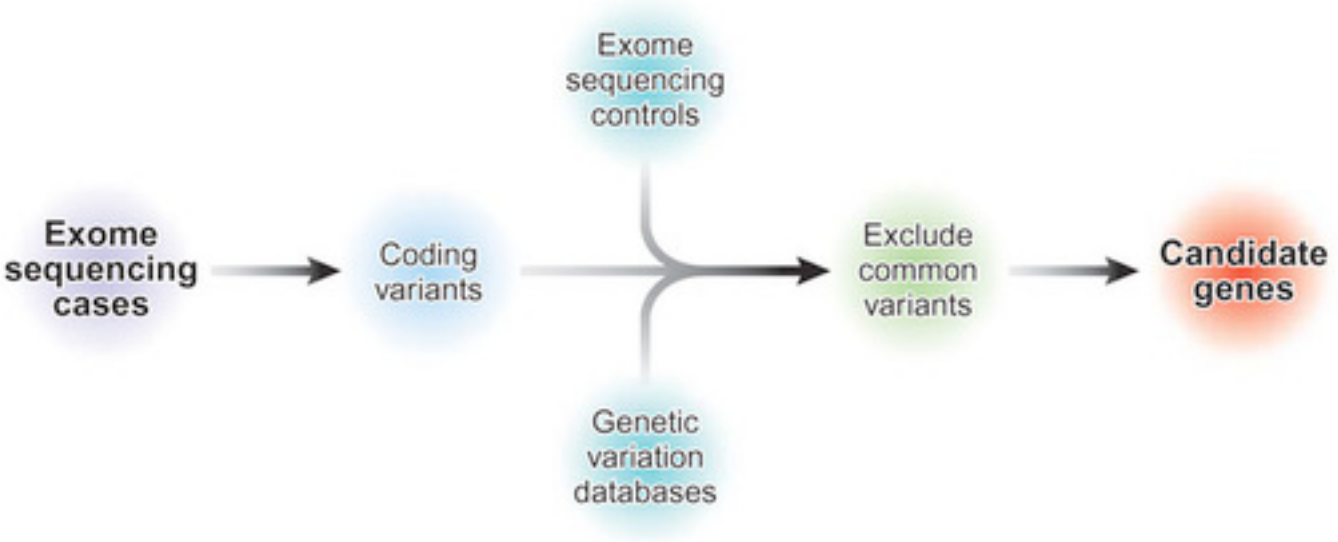


Table 1 Direct identification of the gene for a mendelian disorder by exome resequencing

	Kindred 1-A		Kindred 1-B		Kindred 1 (A+B)		Kindreds 1+2		Kindreds 1+2+3	
Filter	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive
NS/SS/I	4,670	2,863	4,687	2,859	3,940	2,362	3,099	1,810	2,654	1,525
Not in dbSNP129	641	102	647	114	369	53	105	25	63	21
Not in HapMap 8	898	123	923	128	506	46	117	7	38	4
Not in either	456	31	464	33	228	9	26	1*	8	1*
Predicted damaging	204	6	204	12	83	1	5	0	2	0

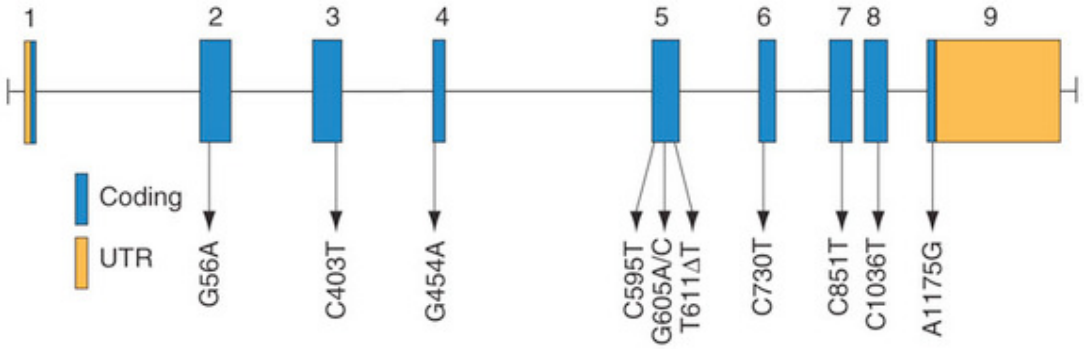
=> ***DHODH***: Dihydroorotate Dehydrogenase

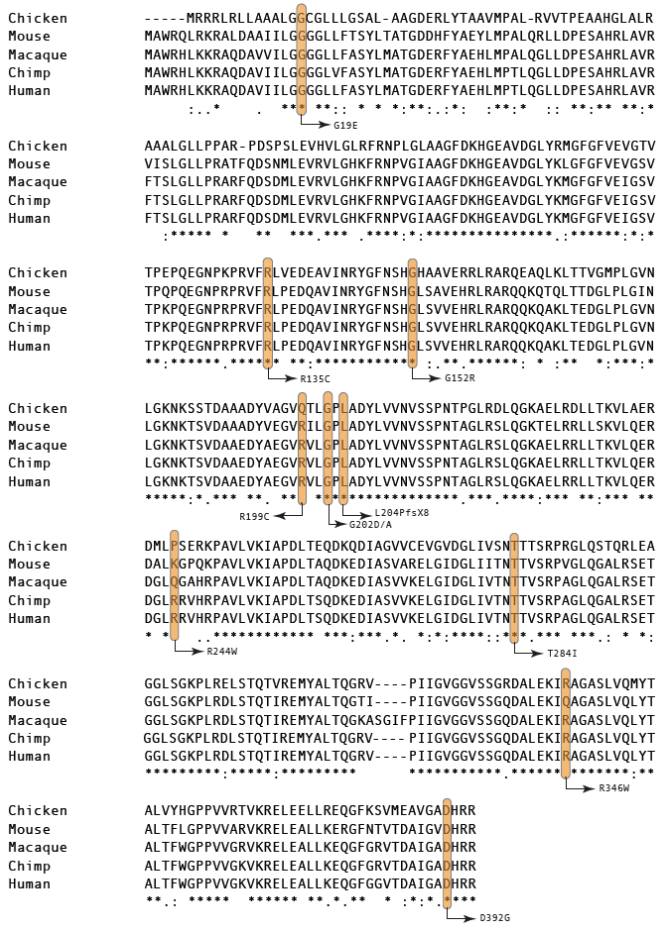
Mutations in the *DHODH* gene

Table 4 Summary of *DHODH* mutations in kindreds with Miller syndrome

Kindred	Mutation	Exon	Amino acid change	Location[b]
1[a]	G454A	4	G152R	chr16: 70608443
	G605C	5	G202A	chr16: 70612611
2[a]	C403T	3	R135C	chr16: 70606041
	C1036T	8	R346W	chr16: 70614936
3[a]	C595T	5	R199C	chr16: 70612601
	611ΔT	5	L204PfsX8	chr16: 70612617
4	G605A	5	G202D	chr16: 70612611
	C730T	6	R244W	chr16: 70613786
5	G56A	2	G19E	chr16: 70603484
	C1036T	8	R346W	chr16: 70614936
6	C851T	7	T284I	chr16: 70614596
	A1175G	9	D392G	chr16: 70615586

[a]Kindreds in which mutations were originally identified by exome resequencing. [b]Chromosomal position was determined using the March 2006 assembly from UCSC (hg18).





Other applications

- **Genome (DNA)**

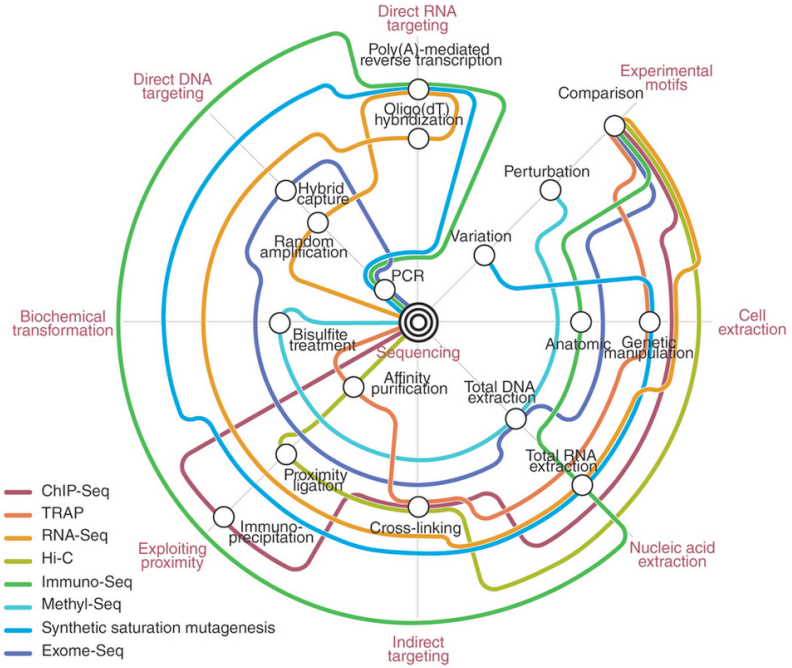
de novo sequencing
Whole genome sequencing
Targeted sequencing (exome, genes…)

- **Transcriptome (RNA)**

mRNA-Seq
Small RNA-Seq (miRNA)

- **Other**

ChIP-Seq
Methyl-Seq



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

Program

10h-11h : Introduction to "Next Generation Sequencing" (NGS) and its applications in human genetics
S. Le Scouarnec

11h-12h : Visit of the Genomics and Bioinformatics core facility **(2nd floor)**
R. Redon & A. Bihouée

12h-14h : Lunch

14h-17h : Interactive session : From raw sequence data to disease causing variants
P. Lindenbaum
The demonstration will include :
- a description of a FASTQ file generated by a sequencer
- aligning the short reads on a reference genome
- processing the alignments (BAM)
- calling the mutations and generating a list of variations (VCF)
- adding some functional annotations to those variants