



Institut de Recherche  
pour le Développement  
FRANCE

# Introduction & NGS methods

Briefly

---

Christine Tranchant-Dubreuil & Francois Sabot

18th of January, 2021

IRD

# Introduction

---

# The who's who



Christine Tranchant-Dubreuil  
Bioinformatics Engineer, PhD student  
Speciality: NGS, pangenomics, data analysis

# The who's who



Christine Tranchant-Dubreuil  
Bioinformatics Engineer, PhD student  
Speciality: NGS, pangenomics, data analysis



Francois Sabot  
Genomician, Senior Scientist  
Speciality: NGS, pangenomics, data analysis

# How will it works ?

- Each afternoon, teaching and practices altogether

# How will it works ?

- Each afternoon, teaching and practices altogether
- Each evening & morning, on your own exercices
  - Some exercices will be done alone
  - Most can be done by groups

# How will it works ?

- Each afternoon, teaching and practices altogether
- Each evening & morning, on your own exercices
  - Some exercices will be done alone
  - Most can be done by groups
  - Debrief and solution every afternoon

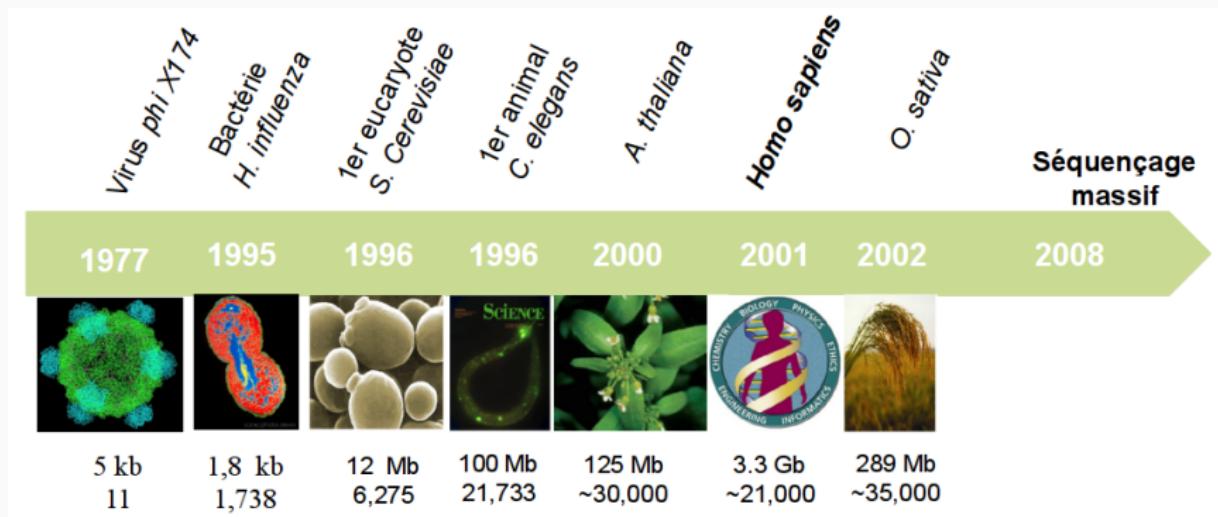
# How will it works ?

- Each afternoon, teaching and practices altogether
- Each evening & morning, on your own exercices
  - Some exercices will be done alone
  - Most can be done by groups
  - Debrief and solution every afternoon
- We will be on Slack all the time (almost)
- One speaks, the other doing the "*service après-vente*"

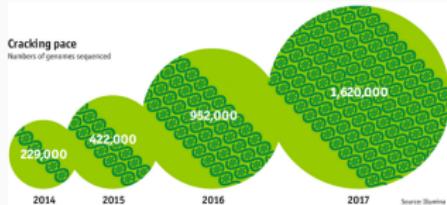
## **The NGS in themselves**

---

# A little history of sequencing...

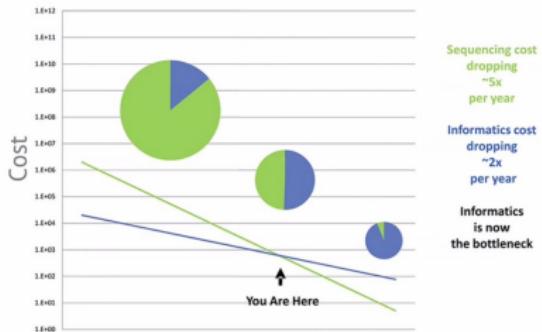


# ...From Data Rarity to Data Deluge



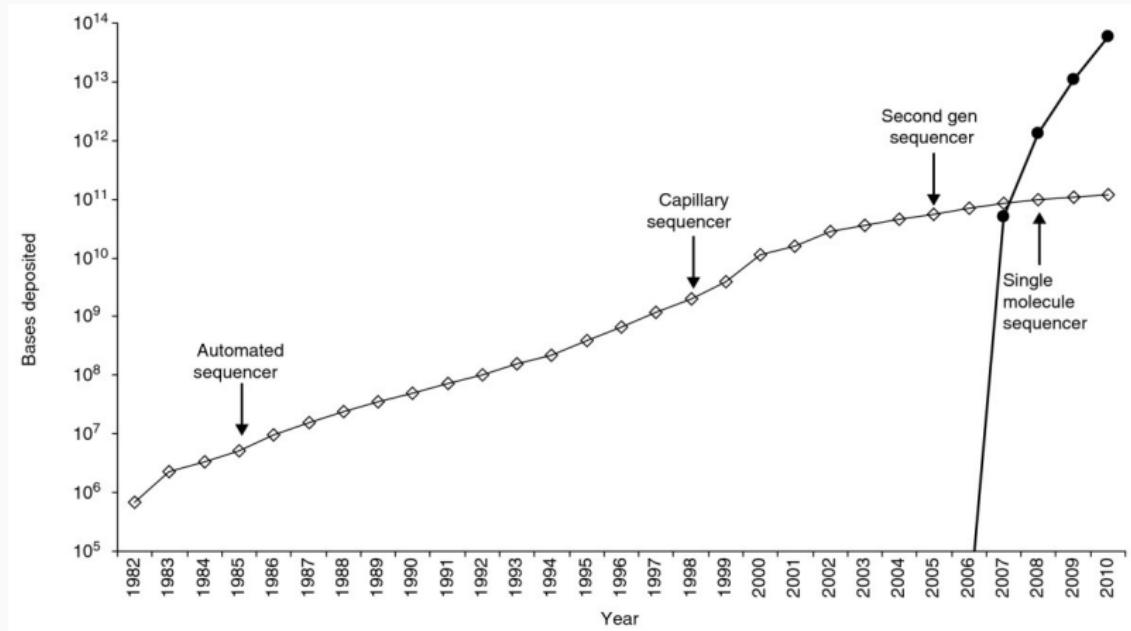
From The economist

## DNA Sequencing Economics



From Bussiness Insider

# ...From Data Rarity to Data Deluge



# What can we do with it ?

- Genetic diversity
- Gene discovery
- Genomic structure
- Contamination/pathogen detection
- Metagenomic
- Pangenomic
- And many other things...

## Methods

---

## 2<sup>nd</sup> Generation Sequencing

- DNA fragmentation (short)
- Matrix amplification
- Short reads
- Limited error rate
- High throughput

## 2<sup>nd</sup> Generation Sequencing

- DNA fragmentation (short)
- Matrix amplification
- Short reads
- Limited error rate
- High throughput

## 3<sup>rd</sup> Generation Sequencing

- DNA fragmentation (long)
- NO MATRIX AMPLIFICATION
- Long reads
- Important error rate
- Medium throughput

## 2<sup>nd</sup> Generation Sequencing

- DNA fragmentation (short) 454
- Matrix amplification Ion Torrent
- Short reads Illumina
- Limited error rate
- High throughput

## 3<sup>rd</sup> Generation Sequencing

- DNA fragmentation (long)
- NO MATRIX AMPLIFICATION
- Long reads
- Important error rate
- Medium throughput

## 2<sup>nd</sup> Generation Sequencing

- DNA fragmentation (short) 454
- Matrix amplification Ion Torrent
- Short reads Illumina
- Limited error rate
- High throughput

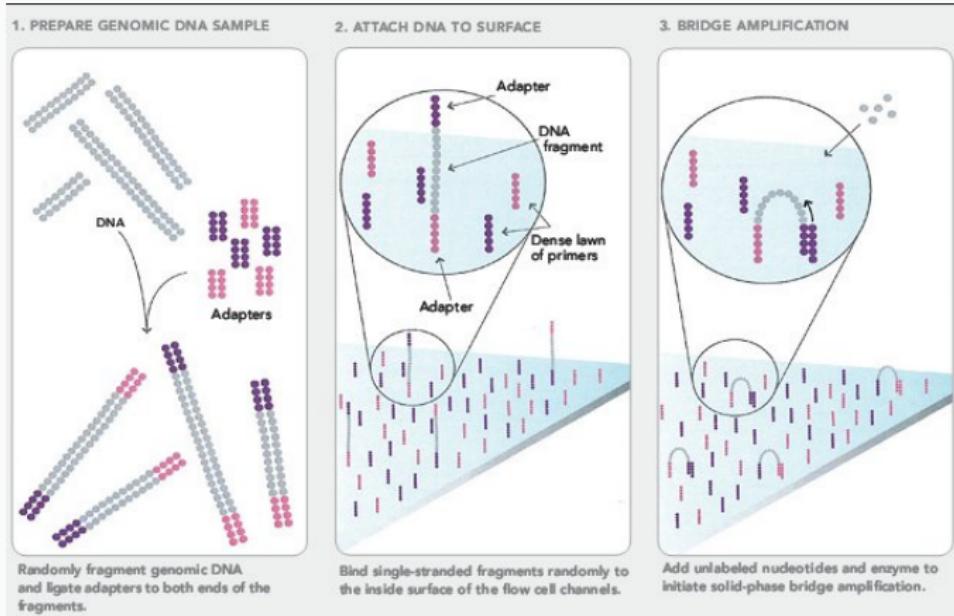
## 3<sup>rd</sup> Generation Sequencing

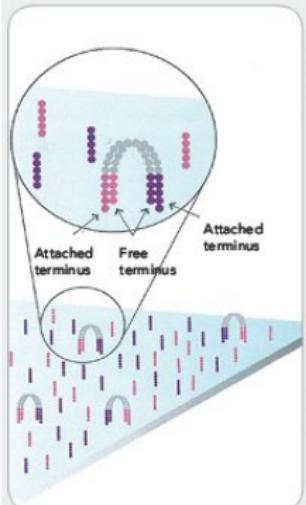
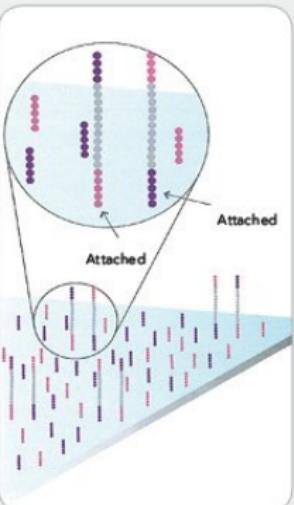
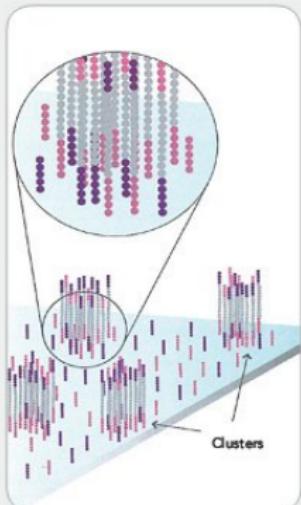
PacificBiosciences  
**Oxford Nanopore**

- DNA fragmentation (long)
- NO MATRIX AMPLIFICATION
- Long reads
- Important error rate
- Medium throughput





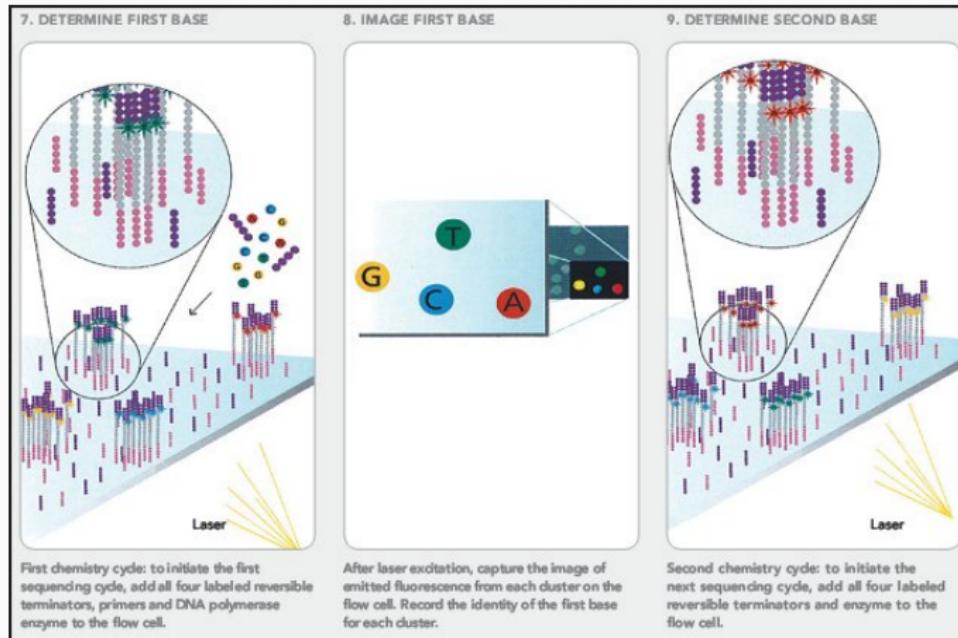


**4. FRAGMENTS BECOME DOUBLE STRANDED****5. DENATURE THE DOUBLE-STRANDED MOLECULES****6. COMPLETE AMPLIFICATION**

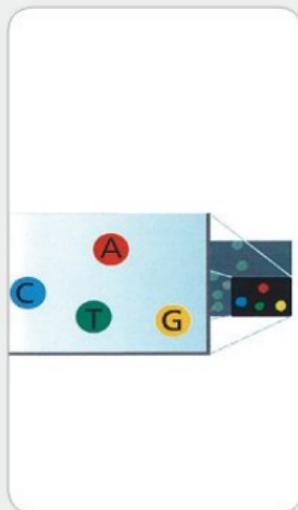
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

Denaturation leaves single-stranded templates anchored to the substrate.

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

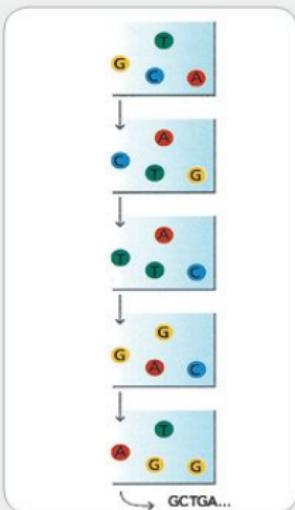


## 10. IMAGE SECOND CHEMISTRY CYCLE



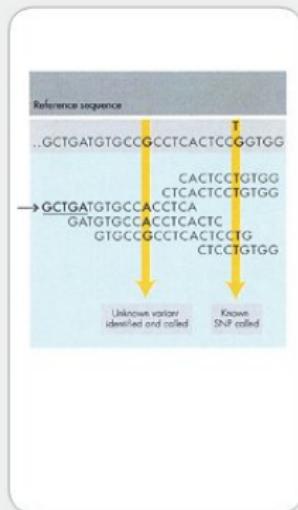
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

## 11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES



Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

## 12. ALIGN DATA



Align data, compare to a reference, and identify sequence differences.

## Advantages :

- Output volume (20 billions of 150b reads/6Tb, NovaSeq6000)
- Accuracy (99.99 % - but questionable)
- Run is cheap
- MySeq is cheap (around 60 000 USD per machine)

**Limits :** Size (150 + 150 in NovaSeq, but 400 for MySeq)

# The FASTQ Format

```
@HWI-EAS236_3_FC_20BTNAAXX:2:1:215:5931 ← Sequencing info
GAGAAGTTCAACAGCTGGTATTATTTTGTAAACAT1
+HWI-EAS236_3_FC_20BTNAAXX:2:1:215:5931
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhUhhE1
@HWI-EAS236_3_FC_20BTNAAXX:2:1:234:5511
TGGGACTTTATCTGGAGGAGTGTGGAAAGCCATT1 ← Nucleotide sequence
+HWI-EAS236_3_FC_20BTNAAXX:2:1:234:5511
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh1
@HWI-EAS236_3_FC_20BTNAAXX:2:1:338:1941
TGGTTTATGCAGAAAATTCTAGAATAAGGGTAACCT1
+HWI-EAS236_3_FC_20BTNAAXX:2:1:338:1941
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh1
@HWI-EAS236_3_FC_20BTNAAXX:2:1:363:7171
TCTCAGAAAATTGTTGTGATGTGTGTATTCAACTA1 ← Quality score in ASCII
+HWI-EAS236_3_FC_20BTNAAXX:2:1:363:7171
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh1
@HWI-EAS236_3_FC_20BTNAAXX:2:1:208:2091
TTGATTTAACTCTGACAAAATAACAAAGCTTAGG1
+HWI-EAS236_3_FC_20BTNAAXX:2:1:208:2091
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhGh1
```

## The QPHRED Scale



S - Sanger Phred+33, raw reads typically (0, 40)

X - Solexa Solexa+64, raw reads typically (-5, 40)

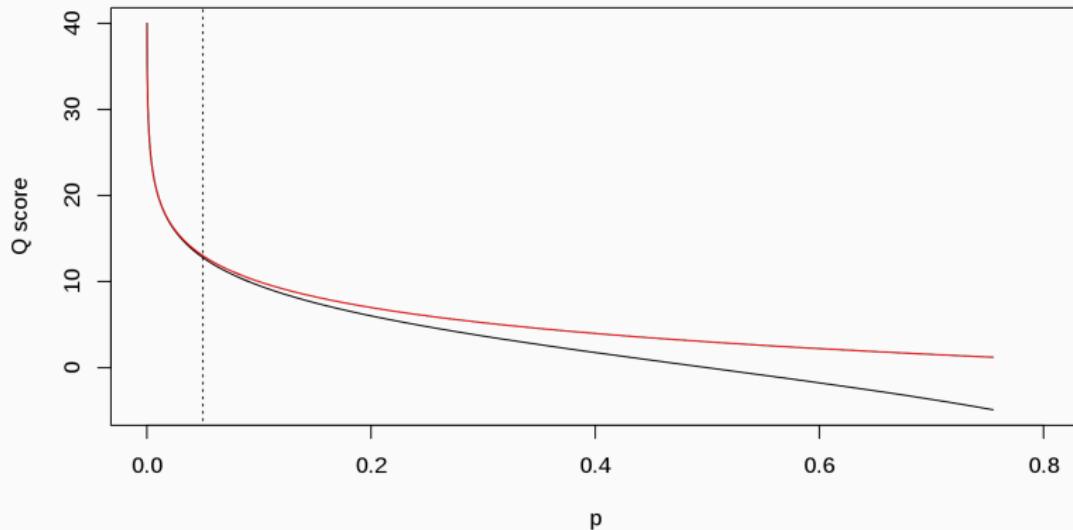
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)

$\delta$  = Illumina 1.5+ Phred $\delta$ 64, raw reads typically (3, 40)

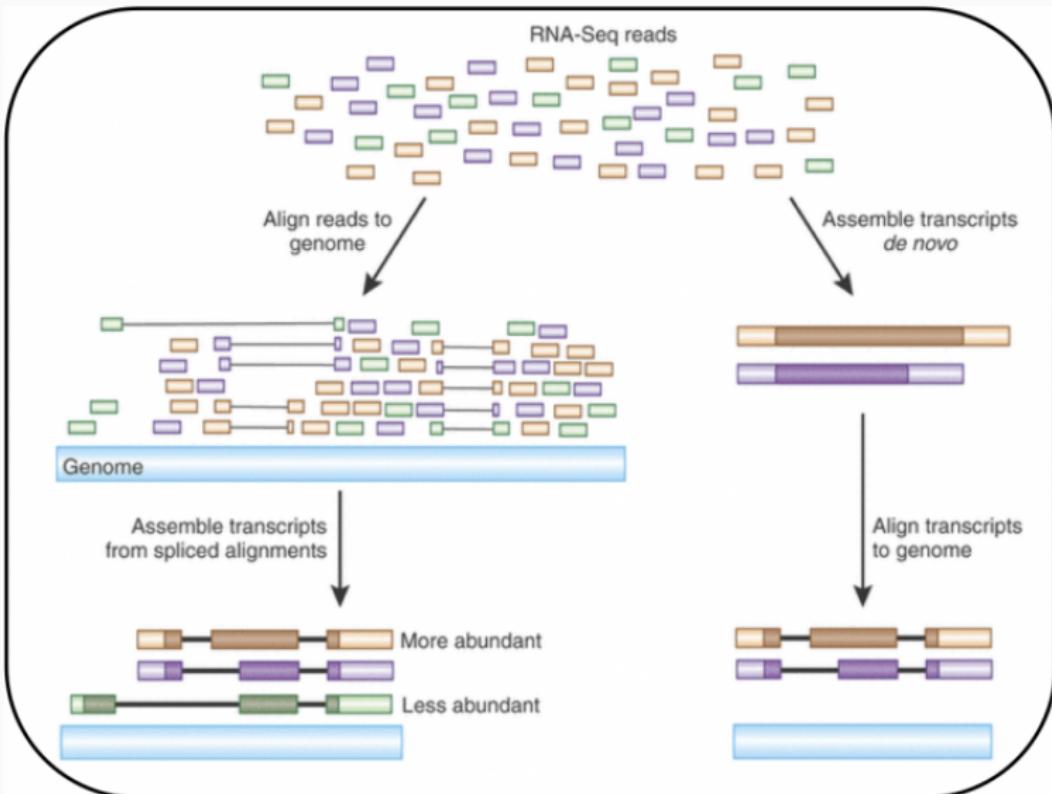
with 0=unused, 1=unused, 2=**Read Segment Quality Control Indicator** (**bold**)  
(Note: See discussion above).

L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

# The QPHRED Value



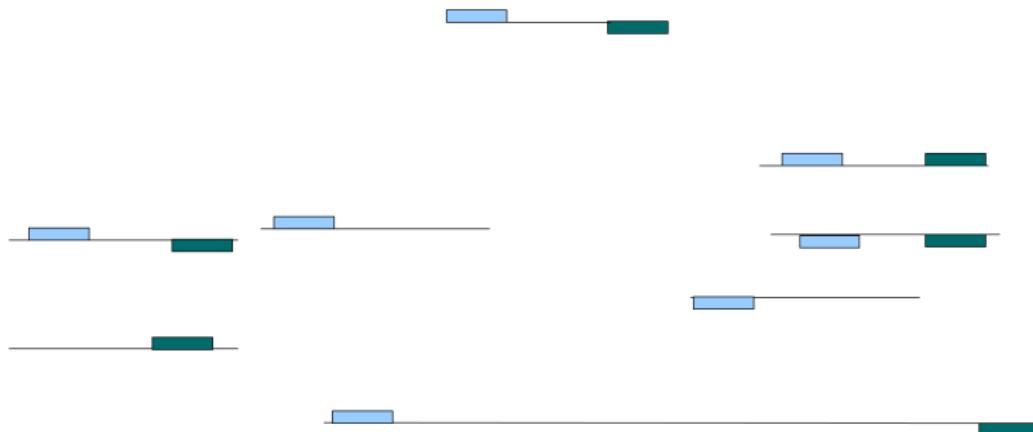
# Mapping



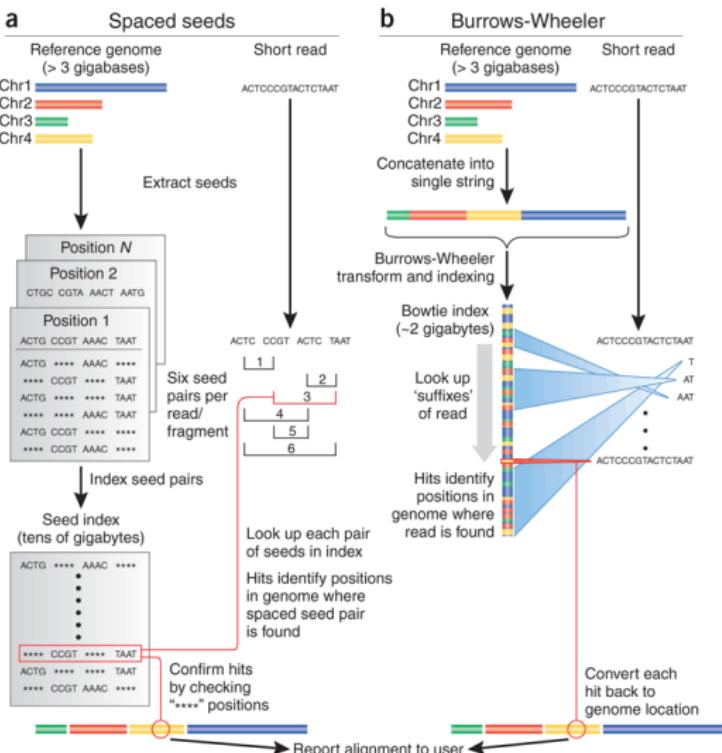
## Generally with **Pair-End** data

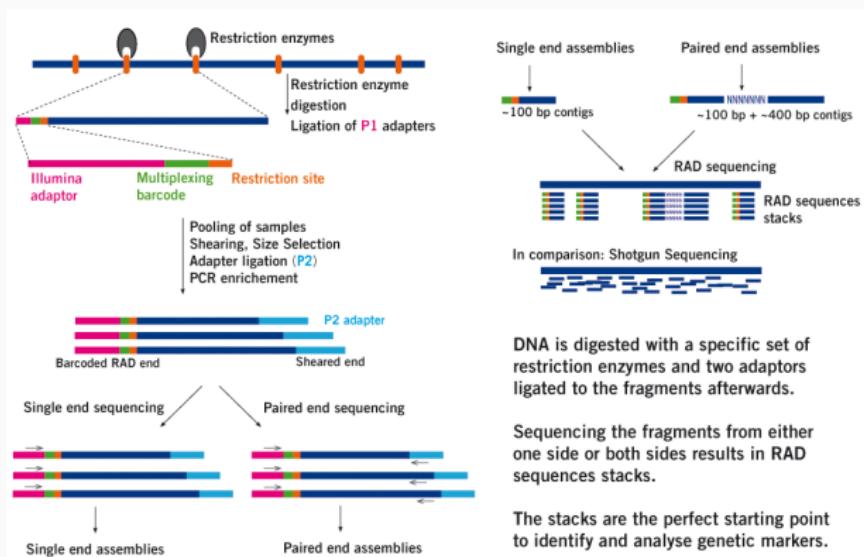


## Generally with **Pair-End** data

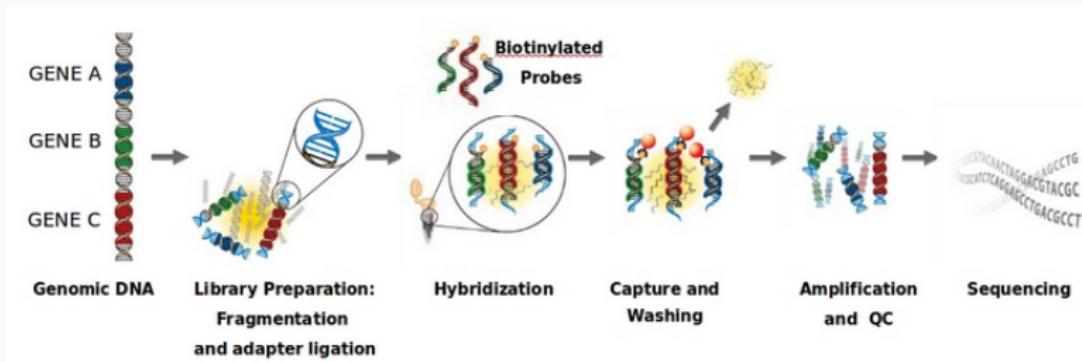


# Mapping



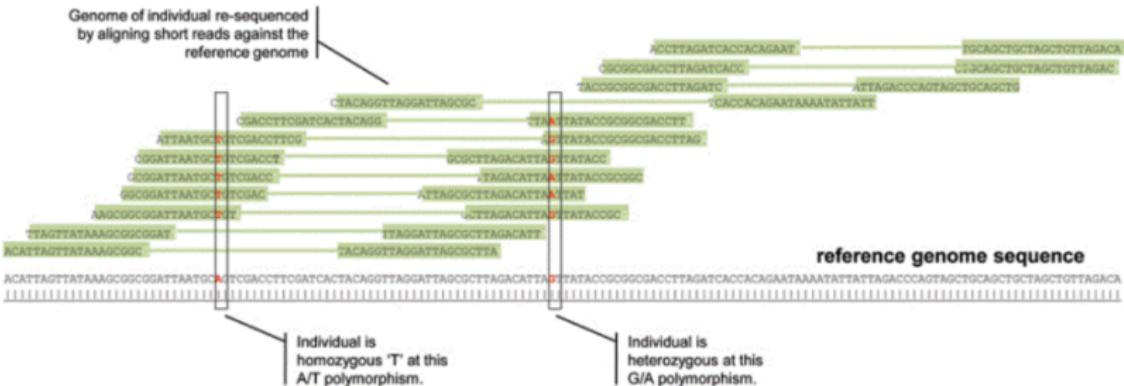


From Eurofins

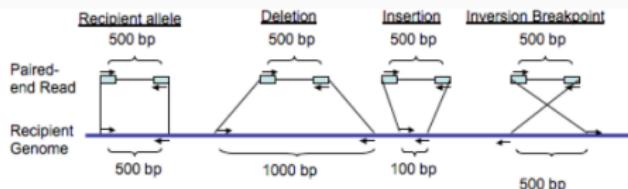
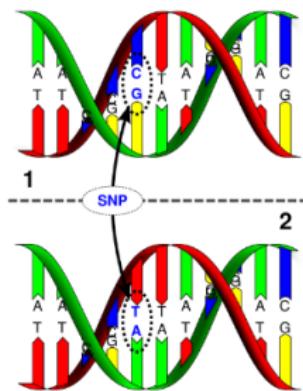


From CGFB, Bordeaux, France

# SNP and InDel Detection



# SNP and InDel Detection



## Types of variants

### SNPs

Alignment	VCF representation
ACGT	POS REF ALT
ATGT	2 C T

### Insertions

Alignment	VCF representation
AC-GT	POS REF ALT
ACTGT	2 C CT

### Deletions

Alignment	VCF representation
ACGT	POS REF ALT
A--T	1 ACG A

### Complex events

Alignment	VCF representation
ACGT	POS REF ALT
A-TT	1 ACG AT

## Large structural variants

VCF representation  
POS REF ALT INFO  
100 T <DEL> SVTYPE=DEL;END=300

# Common File for all Variations, the VCF

## Example

VCF header	##fileformat=VCFv4.0 ##fileDate=20100707 ##source=VCFtools ##reference=NCBI36 ##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele"> ##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership"> ##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype"> ##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)"> ##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)"> ##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth"> ##ALT=<ID=DEL,Description="Deletion"> ##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant"> ##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">	Mandatory header lines																																																							
Body	<table border="1"><thead><tr><th>CHROM</th><th>POS</th><th>ID</th><th>REF</th><th>ALT</th><th>QUAL</th><th>FILTER</th><th>INFO</th><th>FORMAT</th><th>SAMPLE1</th><th>SAMPLE2</th></tr></thead><tbody><tr><td>1</td><td>1</td><td>.</td><td>ACG</td><td>A,AT</td><td>.</td><td>PASS</td><td>.</td><td>GT:DP</td><td>1/2:13</td><td>0/0:29</td></tr><tr><td>1</td><td>2</td><td>rs1</td><td>C</td><td>T,CT</td><td>.</td><td>PASS</td><td>H2;AA=T</td><td>GT:GQ</td><td>0 1:100</td><td>2/2:70</td></tr><tr><td>1</td><td>5</td><td>.</td><td>A</td><td>G</td><td>.</td><td>PASS</td><td>.</td><td>GT:GQ</td><td>1 0:77</td><td>1/1:95</td></tr><tr><td>1</td><td>100</td><td></td><td>T</td><td>&lt;DEL&gt;</td><td>.</td><td>PASS</td><td>SVTYPE=DEL;END=300</td><td>GT:GQ:DP</td><td>1/1:12:3</td><td>0/0:20</td></tr></tbody></table>	CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2	1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29	1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70	1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95	1	100		T	<DEL>	.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20	Optional header lines (meta-data about the annotations in the VCF body)
CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2																																															
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29																																															
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70																																															
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95																																															
1	100		T	<DEL>	.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20																																															
	Deletion	SNP	Large SV	Insertion	Other event			Reference alleles (GT=0)																																																	
							Phased data (G and C above are on the same chromosome)	Alternate alleles (GT>0 is an index to the ALT column)																																																	

VCF = Variant Call Format From 1000 Genomes Project

- Amount of original samples
- Choice of Sample
- Purity of Sample
- Size of sequenced unit
- Error rate
- Volume of Outputted data

All linked to technical constraints

- Cleaning data level
- Mapping Conditions
- Mapping Cleaning Conditions
- Variation Calling level

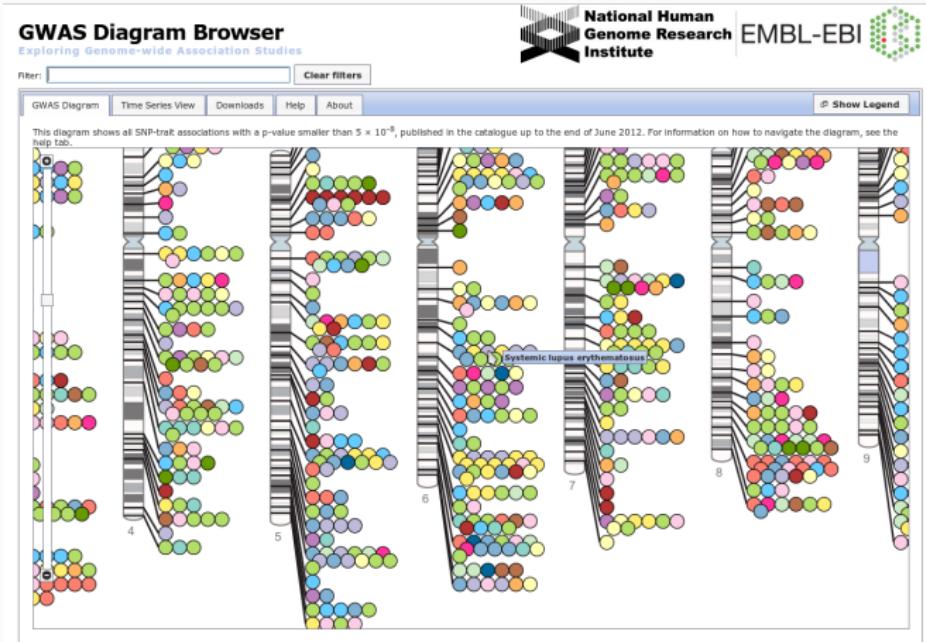
All linked to the Specificity/Sensitivity Informatics Paradox

## Applications

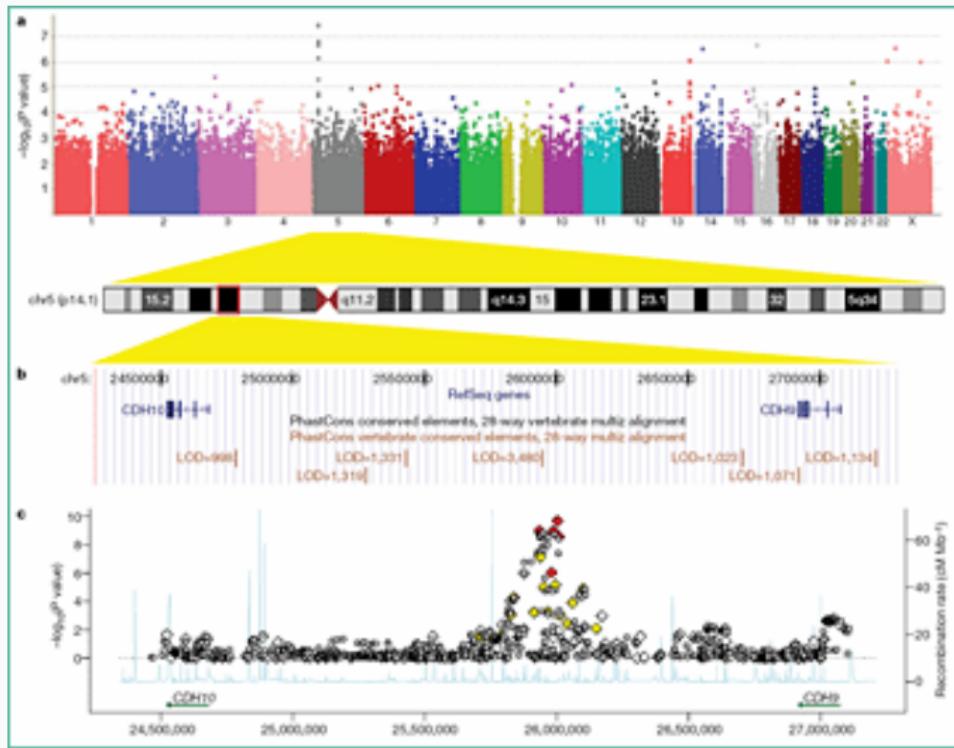
---

- Gene discovery/GWAs
- Species Definition
- Subspecies/specific subgroup definition
- Global genotyping (for breeding in agriculture e.g.)
- Genomic Ecology (Transposable elements, etc...)

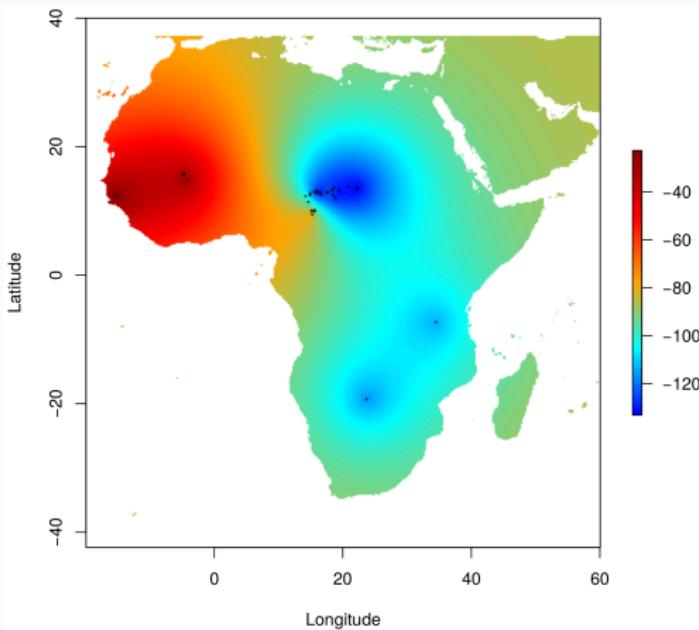
# Example in GWAs & Population Genomics



# Example in GWAs & Population Genomics



# Example in Global Genotyping & Population Genomics



From Cubry et al, 2018

# Large Projects

The image displays two side-by-side screenshots of large-scale genomic projects. On the left is the "1000 Genomes" project, featuring a dark header with the title "1000 Genomes" and "A Deep Catalog of Human Genetic Variation". Below the header is a navigation bar with links to Home, About, Data, Analysis, Participants, Contact, and Help. A "LATEST ANNOUNCEMENTS" section highlights "WEDNESDAY FEBRUARY 16, 2011 February 2011 Data Up Full Project Indel Release". It also mentions "Indels calls from Dindel. These calls genome project. This release is ba" and "Data access links: EBI / NCBI". On the right is the "1001 Genomes" project, which is a catalog of *Arabidopsis thaliana* genetic variation. Its header includes the title "1001 Genomes" and "A Catalog of *Arabidopsis thaliana* Genetic Variation", along with a decorative graphic of a flower. The navigation bar for "1001 Genomes" includes Home, Collaborators, Accessions, Tools, Software, Data Center, Gallery, About, and Help desk. Below the navigation is a "Welcome to the 1001 Genomes Project" message. The main content area features a large banner for "GENOME 10K" with the tagline "Unveiling animal diversity" and a search bar. The banner is set against a background of blue DNA helixes and various animal illustrations. At the bottom of the page, there are sections for "Join us" (with a link to "Become a G10K affiliate") and "Genome assembly".

# Sample types

- DNA from plant, animal, microbial...
- Organite DNA (mitochondria, chloroplast)
- Subsample DNA (exon capture, 16S capture for Barcoding)
- Viral sample from infected tissue
- Environmental sample: water, feces, cloud...

# Possibilities in the next 5-10 years (From a presentation in 2013)



- Real-time Transcriptomics

# Possibilities in the next 5-10 years (From a presentation in 2013)



- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014

# Possibilities in the next 5-10 years (From a presentation in 2013)



- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014
- Single-Cells Transcriptomics (and smallRNA) -> DONE in 2015

# Possibilities in the next 5-10 years (From a presentation in 2013)



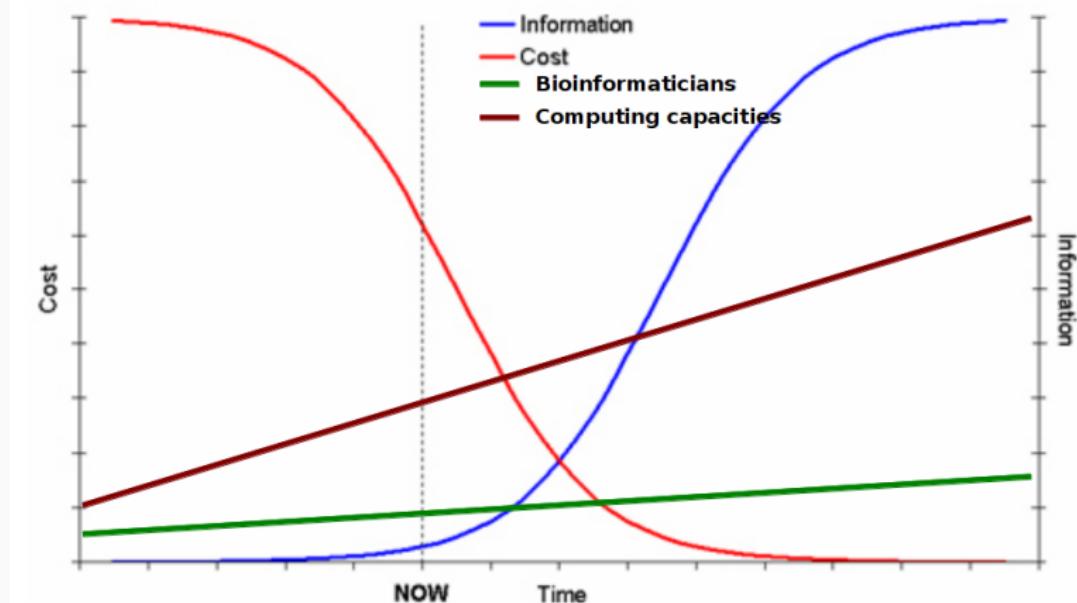
- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014
- Single-Cells Transcriptomics (and smallRNA) -> DONE in 2015
- Personal Genomics medicine (ethical problems...) -> Available

# Possibilities in the next 5-10 years (From a presentation in 2013)

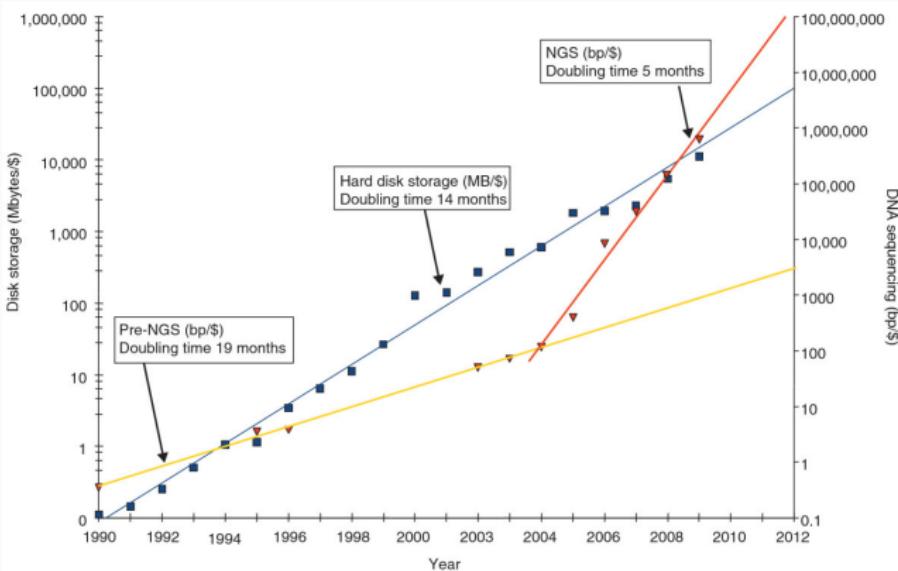


- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014
- Single-Cells Transcriptomics (and smallRNA) -> DONE in 2015
- Personal Genomics medicine (ethical problems...) -> Available
- And any new ideas you will have...

# Keep in mind!

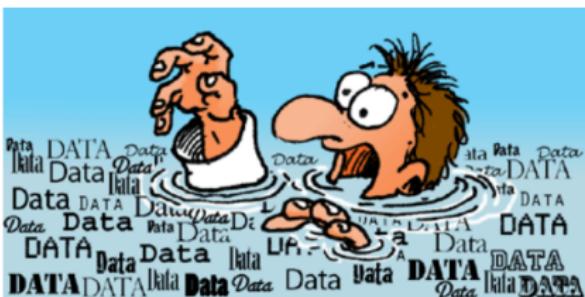
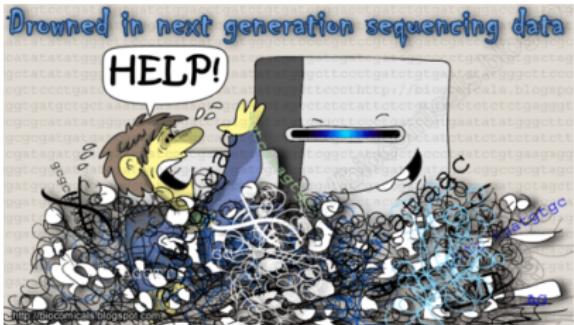


# ...From Data Rarity to Data Deluge



From L. Stein, 2010

# Be Careful to data drowning!



Thanks for your attention

