

Introduction & NGS methods

Briefly

Christine Tranchant-Dubreuil & Francois Sabot

18th of January, 2021

IRD

Introduction



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



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



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

- Each afternoon, teaching and practices altogether

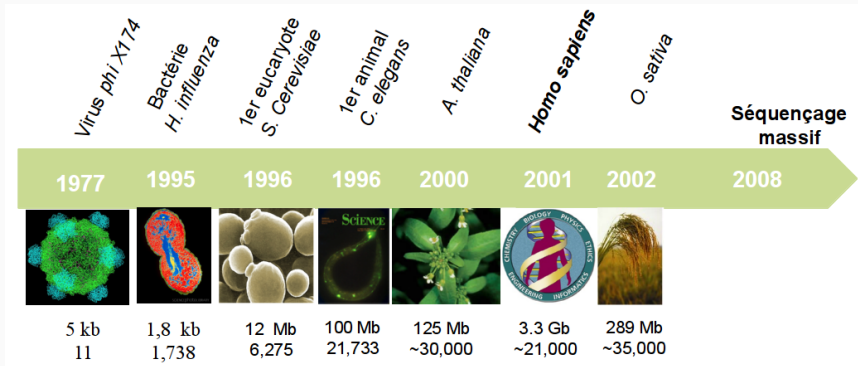
- Each afternoon, teaching and practices altogether
- Each evening & morning, on your own exercises
 - Some exercises will be done alone
 - Most can be done by groups

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

- Each afternoon, teaching and practices altogether
- Each evening & morning, on your own exercises
 - Some exercises 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

Decoding for cents

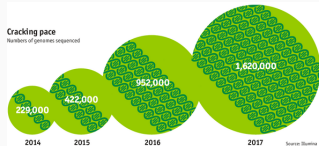
Cost of determining 1m bases of a DNA sequence
Log scale, \$



Source: National Human Genome Research Institute

Cracking pace

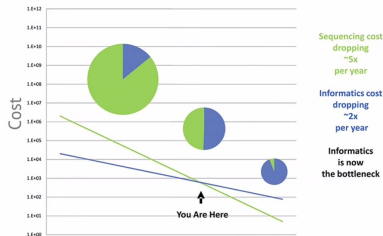
Numbers of genomes sequenced



Source: Statista

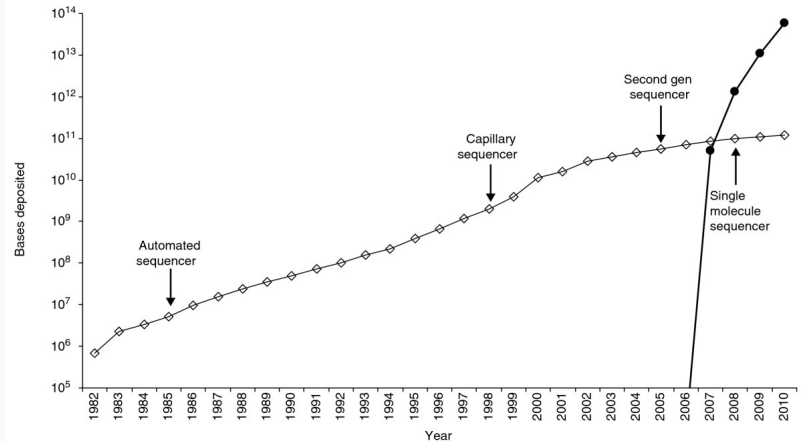
From The economist

DNA Sequencing Economics



From Bussiness Insider

...From Data Rarity to Data Deluge



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

Methods

2nd Generation Sequencing

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

2nd Generation Sequencing

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

3rd Generation Sequencing

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

2nd Generation Sequencing

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

454

IonTorrent

Illumina

3rd Generation Sequencing

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

2nd Generation Sequencing

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

454

IonTorrent

Illumina

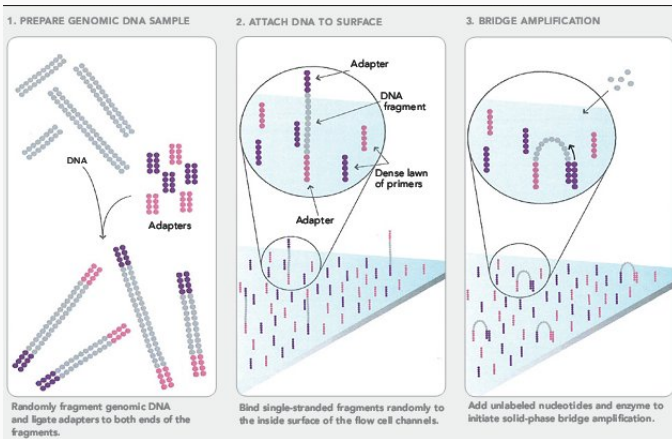
3rd Generation Sequencing

PacificBiosciences
Oxford Nanopore

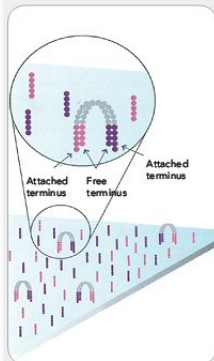
- DNA fragmentation (long)
- NO MATRIX AMPLIFICATION
- Long reads
- Important error rate
- Medium throughput





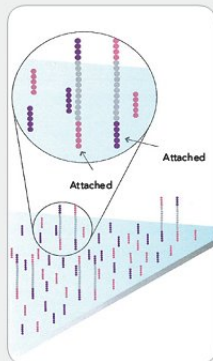


4. FRAGMENTS BECOME DOUBLE STRANDED



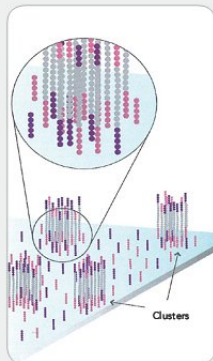
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES



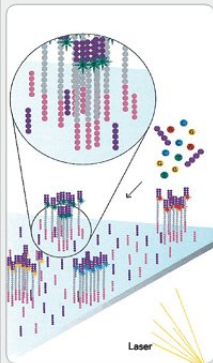
Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION



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

7. DETERMINE FIRST BASE



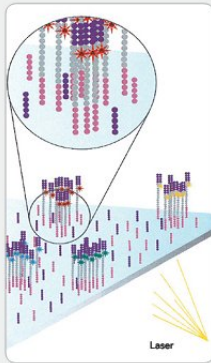
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE



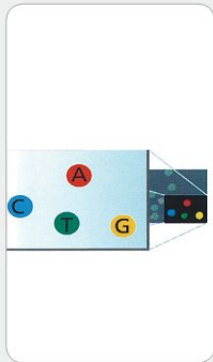
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE



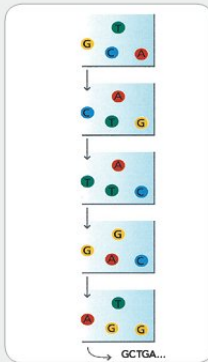
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

10. IMAGE SECOND CHEMISTRY CYCLE



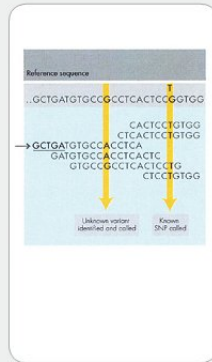
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

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)

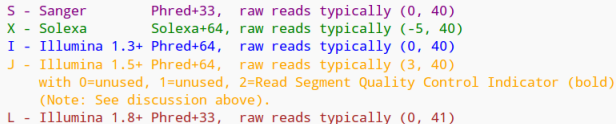
```
@HWI-EAS236_3_FC_20BTNAAXX:2:1:215:593
GAGAAGTTCAACAGCTGGTATTATTTTGTTAACAT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:215:593
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhUhhE
@HWI-EAS236_3_FC_20BTNAAXX:2:1:234:551
TGGGACTTTATCTGGAGGAGTGTGGAAAGCCATT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:234:551
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:338:194
TGGTTTATGCAGAAATTTCTAGAATAAGGGTAACTT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:338:194
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:363:717
TCTCAGAAACTTGTGTGTGTGTATTCAACTA
+HWI-EAS236_3_FC_20BTNAAXX:2:1:363:717
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:208:209
TTGATTTAACCTCTGACAAAATAAACAAAGTCTTAGG
+HWI-EAS236_3_FC_20BTNAAXX:2:1:208:209
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhGh
```

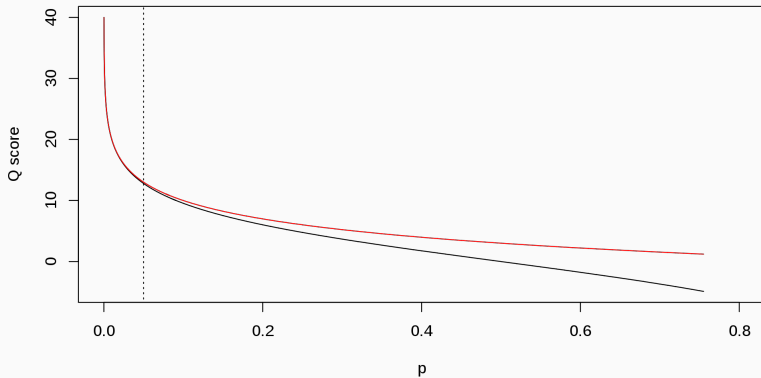
Sequencing info

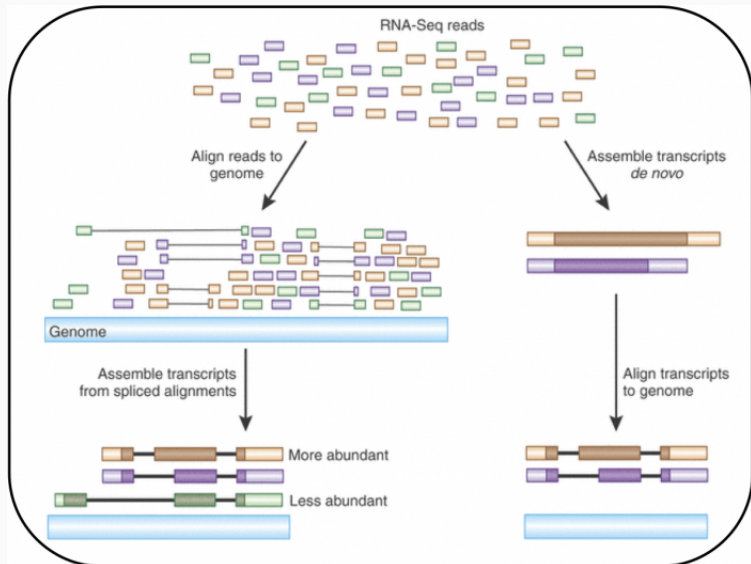
Nucleotide sequence

Quality score in ASCII

Institut de Recherche
 pour le Développement
 FRANCE



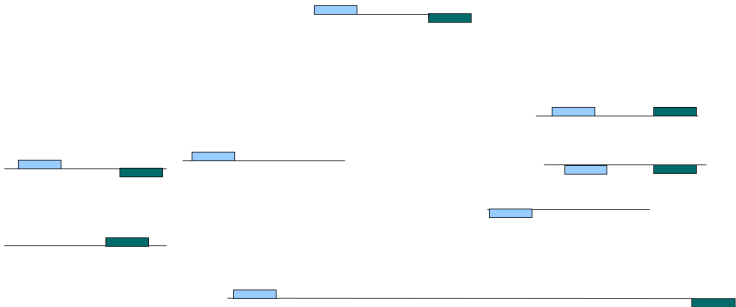


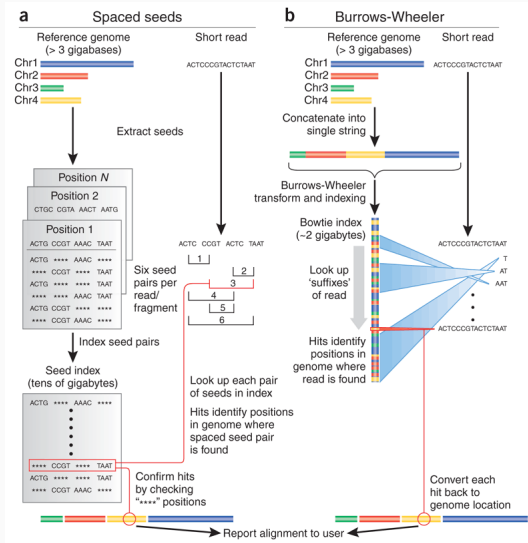


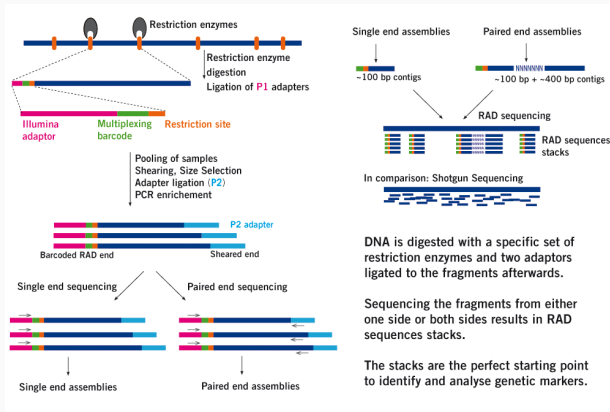
Generally with **Pair-End** data



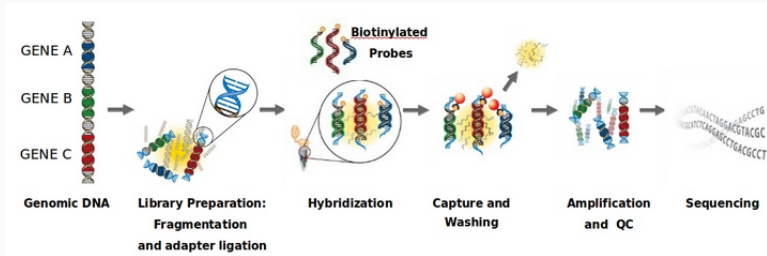
Generally with **Pair-End** data



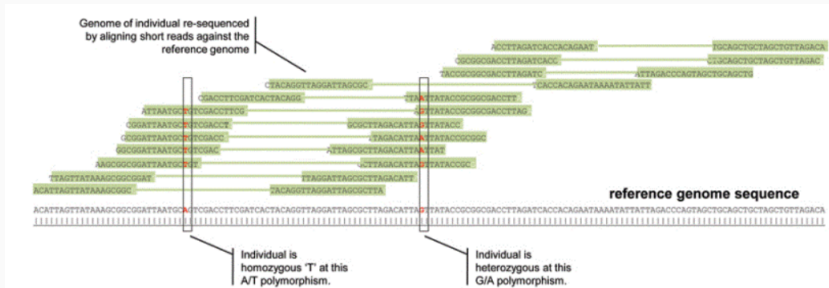


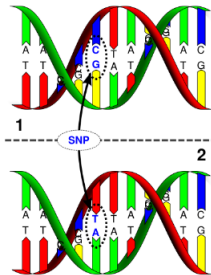
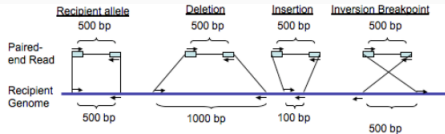


From Eurofins



From CGFB, Bordeaux, France





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		SVTYPE=DEL;END=300

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

Optional header lines (meta-data about the annotations in the VCF body)

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
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		.	PASS	SVTYPE=DEL;END=300	.	GT:GQ:DP	1/1:12:3	0/0:20

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Deletion

SNP

Large SV

Insertion

Other event

Phased data (G and C above are on the same chromosome)

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

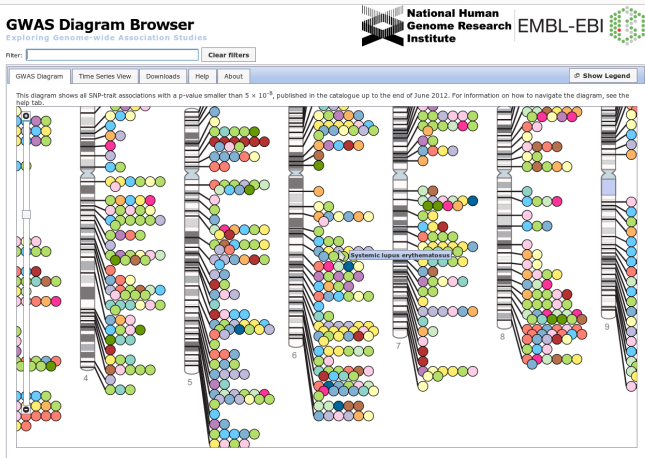
- Cleaning data level
- Mapping Conditions
- Mapping Cleaning Conditions
- Variation Calling level

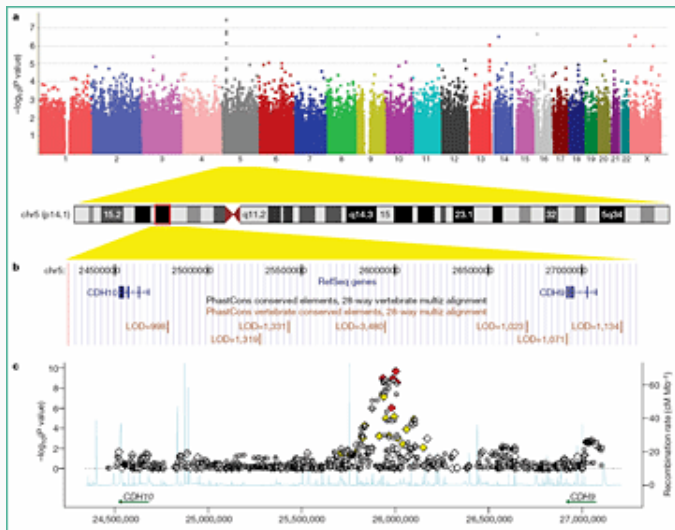
All linked to the Specificity/Sensitivity Informatics Paradox

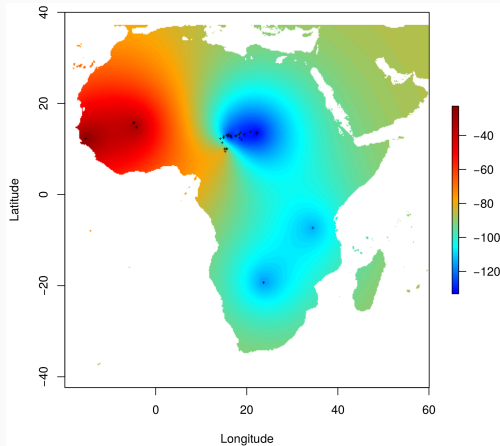
Applications

- 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







From Cubry et al, 2018

1000 Genomes

A Deep Catalog of Human Genetic Variation

[Home](#) [About](#) [Data](#) [Analysis](#) [Participants](#) [Contact](#)

LATEST ANNOUNCEMENTS

WEDNESDAY FEBRUARY 16, 2011
February 2011 Data Up
Full Project Indel Release
Indels calls from [Dindel](#). These call genomes project. This release is ba
Data access links: [EBI](#) / [NCBI](#)

1001 Genomes

A Catalog of *Arabidopsis thaliana* Genetic Variation

[Home](#) [Collaborators](#) [Accessions](#) [Tools](#) [Software](#) [Data Center](#) [Gallery](#) [About](#) [Help desk](#)

Welcome to the 1001 Genomes Project

Links

	Database & Species lists	News	Events	Publications	Participants	For G10K Organizers (restricted)
---	--	----------------------	------------------------	------------------------------	------------------------------	--

Search: [Go](#)

GENOME 10K[®]

Unveiling animal diversity

Genome 10K Project

To understand how complex animal life evolved through changes in DNA and use this knowledge to become better stewards of the planet.

April 2009—The Genome 10K project aims to assemble a genomic zoo—a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately one for every vertebrate genus. The trajectory of cost reduction in DNA sequencing suggests that this project will be feasible within a few

[Join us](#)
Become a G10K affiliate

[Genome assembly](#)

- 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)

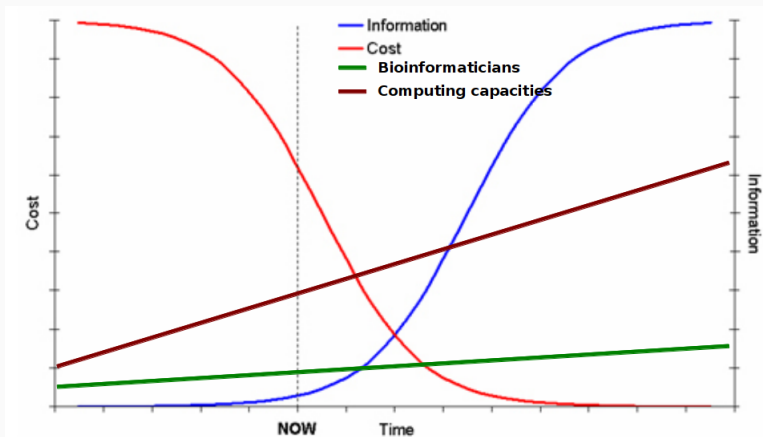
- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014
- Single-Cells Transcriptomics (and smallRNA) -> DONE in 2015

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

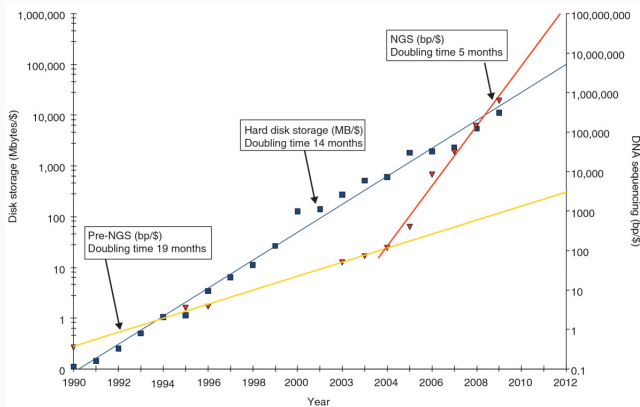
- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014
- Single-Cells Transcriptomics (and smallRNA) -> DONE in 2015
- Personal Genomics medicine (ethical problems...) -> Available

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...

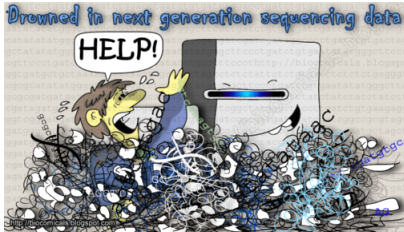
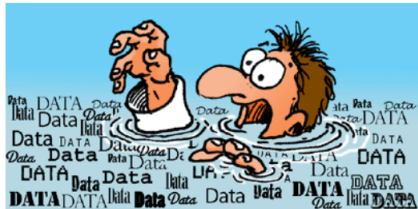


...From Data Rarity to Data Deluge



From L. Stein, 2010

Be Careful to data drowning!



Thanks for your attention

