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 ?

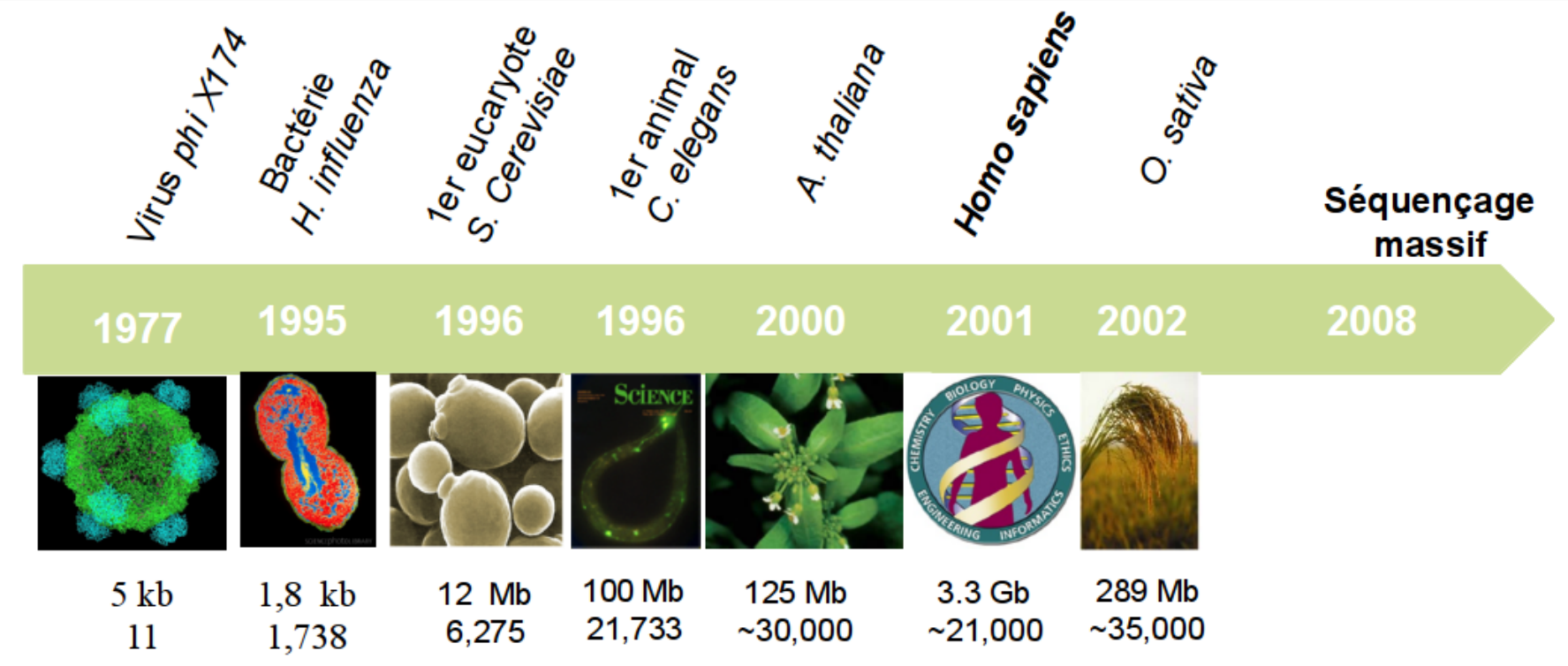
- Each afternoon, teaching and practices altogether
- Each evening & morning,on your own exercices
  - Some exercices will be done alone
  - Most can be done by groups
  - Debrief and solution every afternoon

How will it works ?

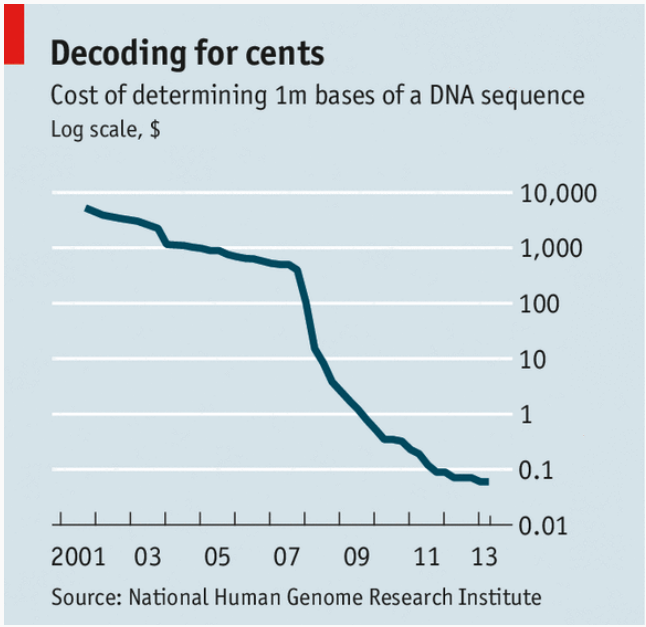
- Each afternoon, teaching and practices altogether
- Each evening & morning,on your own exercices
  - Some exercices will be done alone
  - Most can be done by groups
  - Debrief and solution every afternoon
- We will be on Slack all the time (almost)
- One speaks, the other doing the "*service après-vente*"

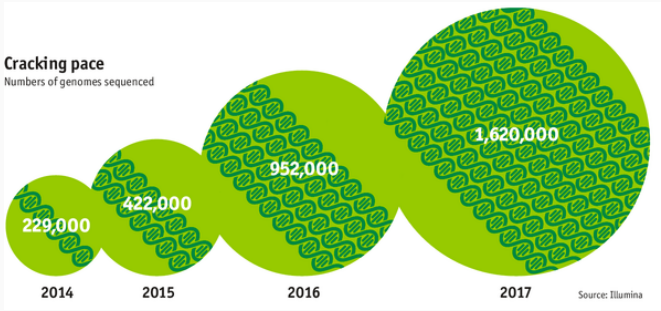
The NGS in themselves

A little history of sequencing...

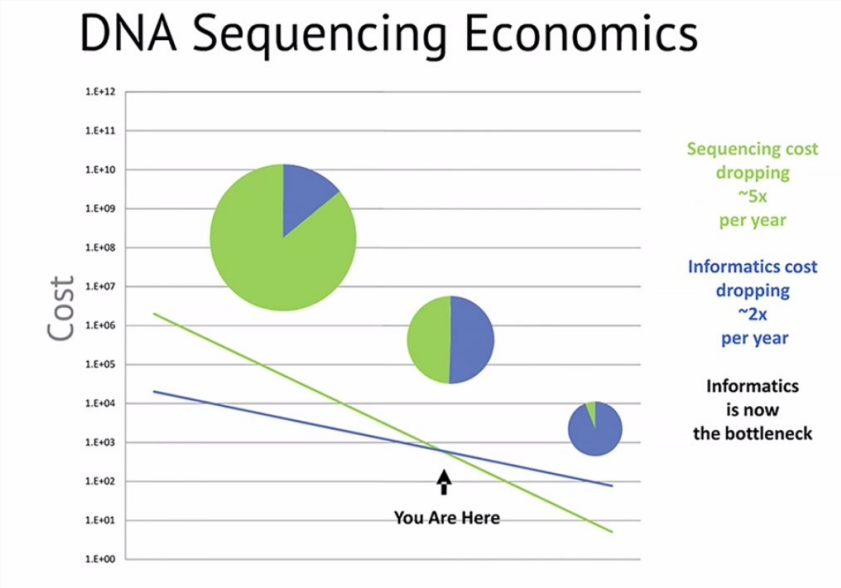


...From Data Rarity to Data Deluge



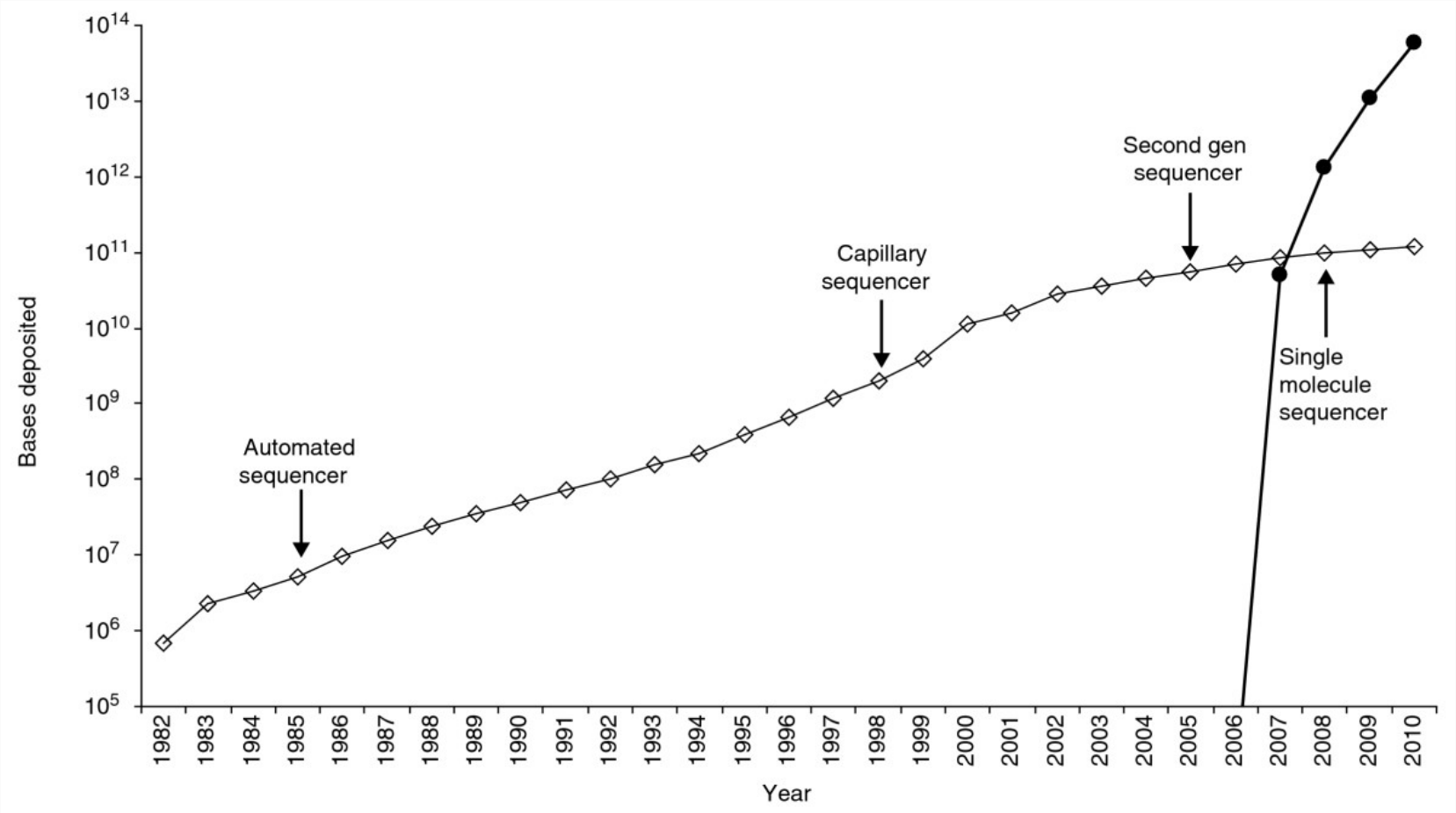


From The economist



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

Current Technologies

2^{nd} Generation Sequencing

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

Current Technologies

2^{nd} Generation Sequencing

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

3^{rd} Generation Sequencing

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

Current Technologies

2^{nd} Generation Sequencing

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

454
IonTorrent
Illumina

3^{rd} Generation Sequencing

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

Current Technologies

2^{nd} Generation Sequencing

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

454
IonTorrent
Illumina

3^{rd} Generation Sequencing

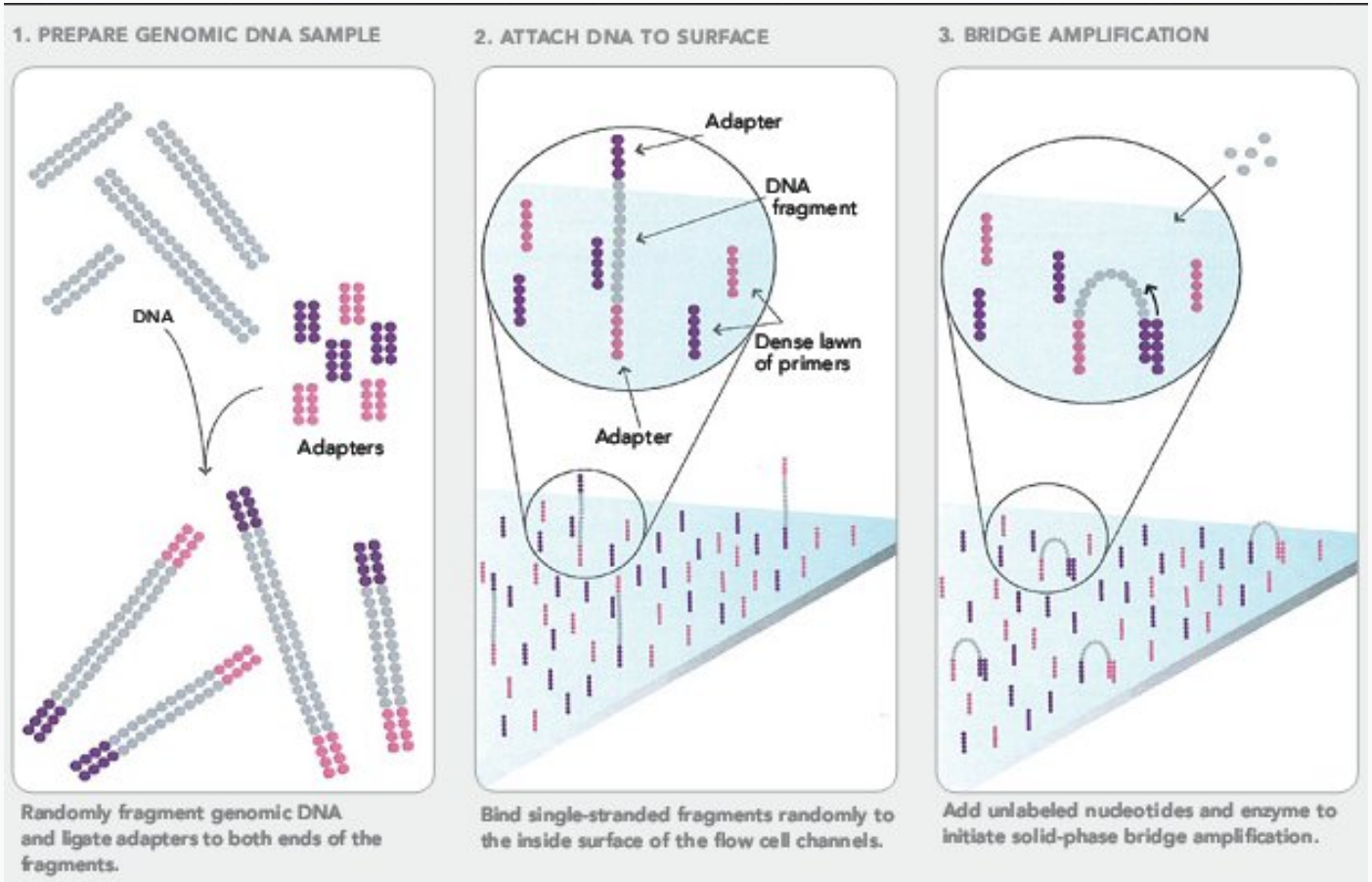
- DNA fragmentation (long)
- NO MATRIX AMPLIFICATION
- Long reads
- Important error rate
- Medium throughput

PacificBiosciences
Oxford Nanopore

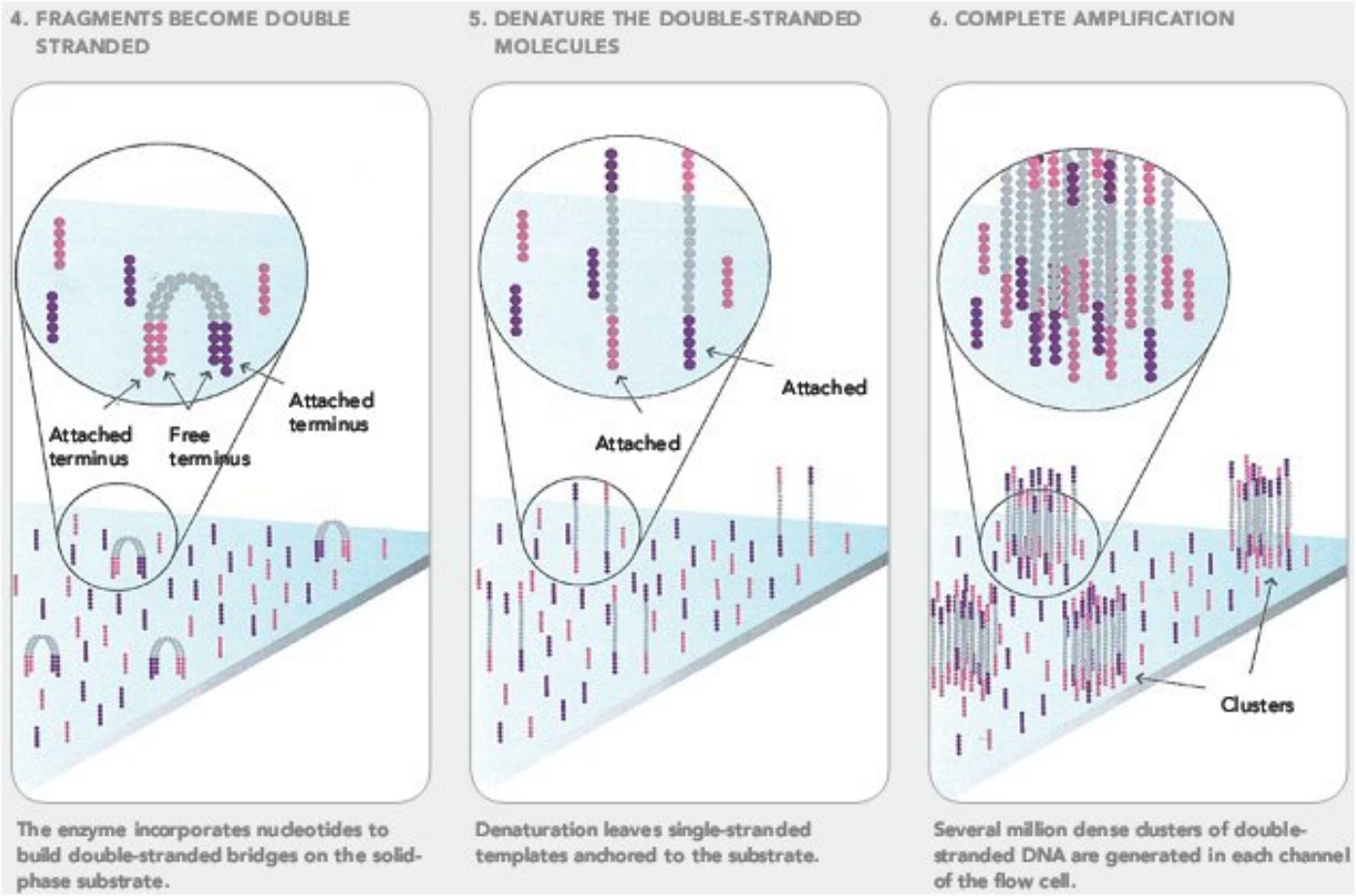
Illumina

Illumina

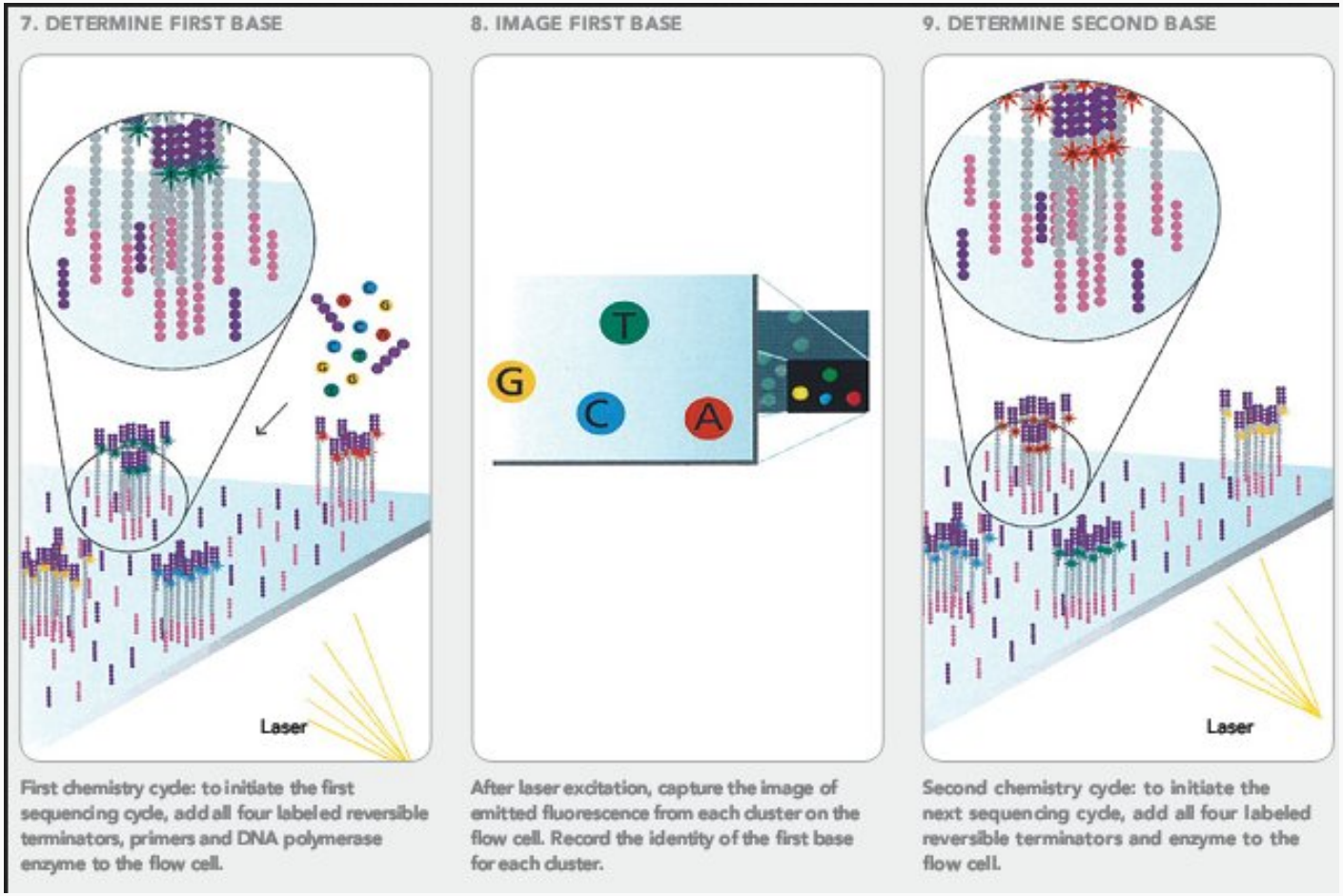
Illumina


1. PREPARE GENOMIC DNA SAMPLE
DNA
Adapters
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
2. ATTACH DNA TO SURFACE
Adapter
DNA fragment
Dense lawn of primers
Adapter
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
3. BRIDGE AMPLIFICATION
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

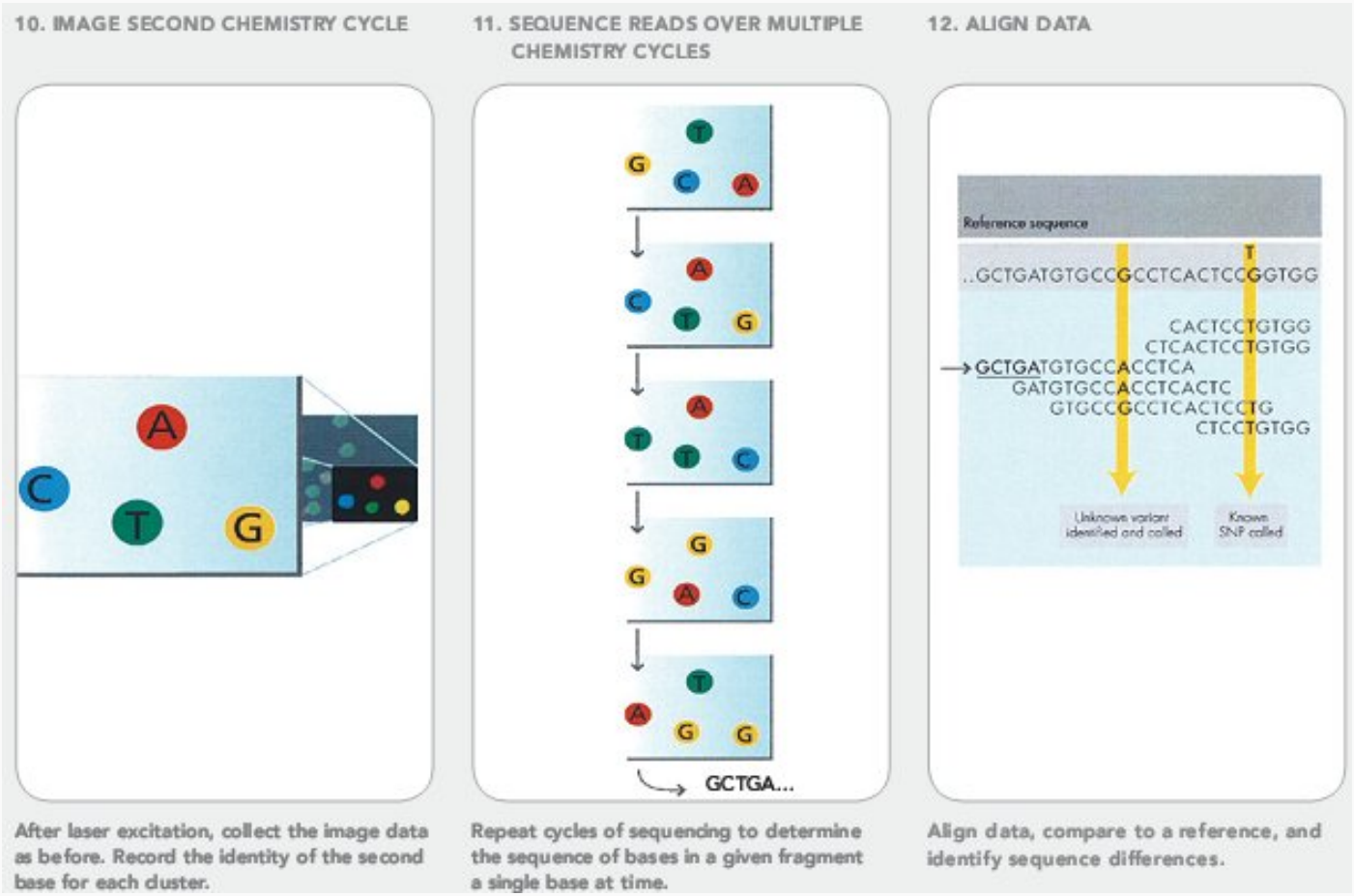
Illumina


4. FRAGMENTS BECOME DOUBLE STRANDED
5. DENATURE THE DOUBLE-STRANDED MOLECULES
6. COMPLETE AMPLIFICATION
Attached terminus
Free terminus
Attached terminus
Attached
Attached
Clusters
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.

Illumina


7. DETERMINE FIRST BASE
8. IMAGE FIRST BASE
9. DETERMINE SECOND BASE
T
G
C
A
Laser
Laser
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.
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.
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

Illumina


10. IMAGE SECOND CHEMISTRY CYCLE
11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES
12. ALIGN DATA
A
C
T
G
GCTGA...
Reference sequence
..GCTGATGTGCCGCCTCACTCCGGTGG
CACTCCTGTGG
CTCACTCCTGTGG
GCTGATGTGCCACCTCA
GATGTGCCACCTCACTC
GTGCCGCCTCACTCCTG
CTCCTGTGG
Unknown variant identified and called
Known SNP called
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.
Align data, compare to a reference, and identify sequence differences.

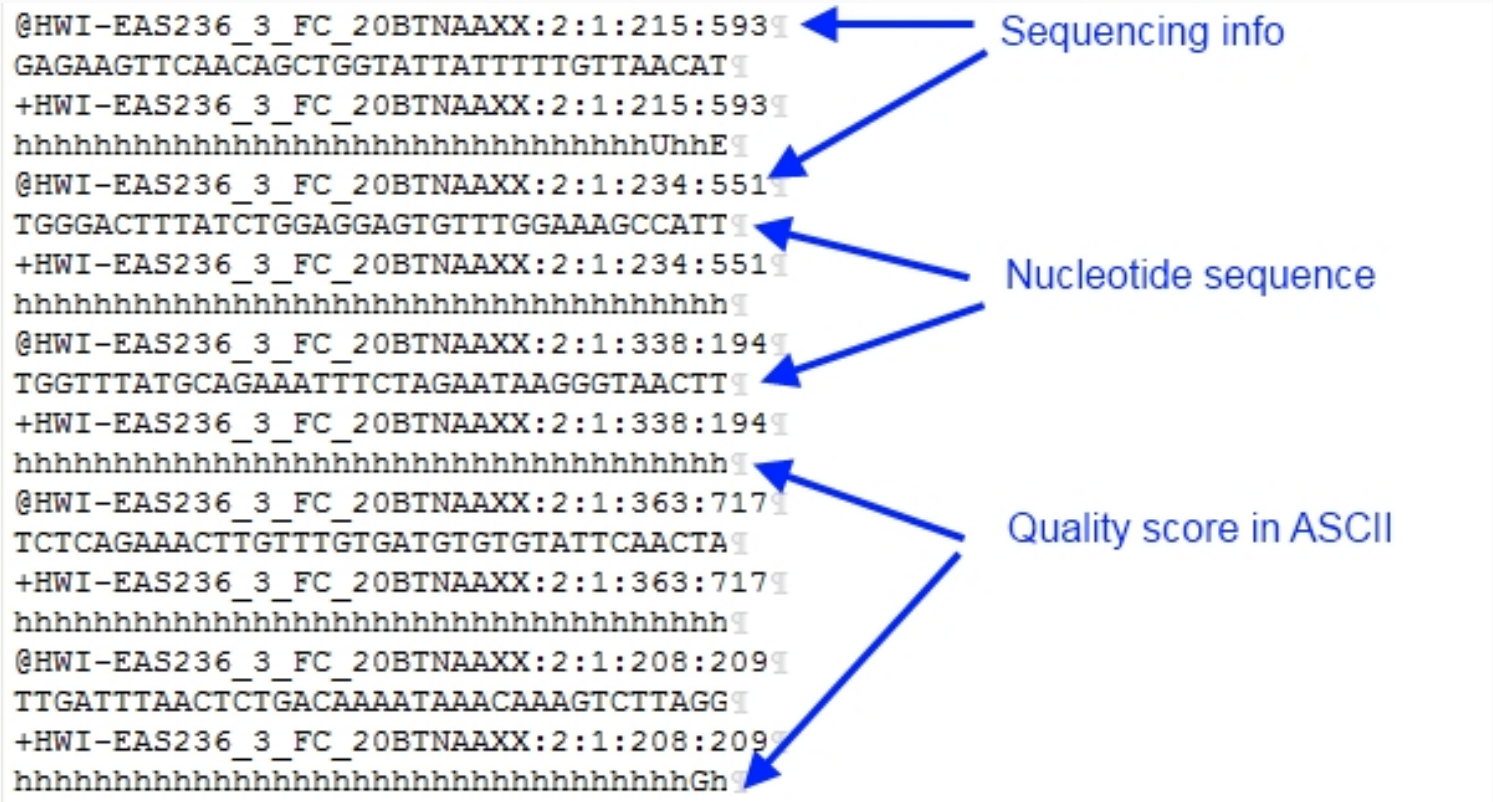
Illumina - Advantages and Limits

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)

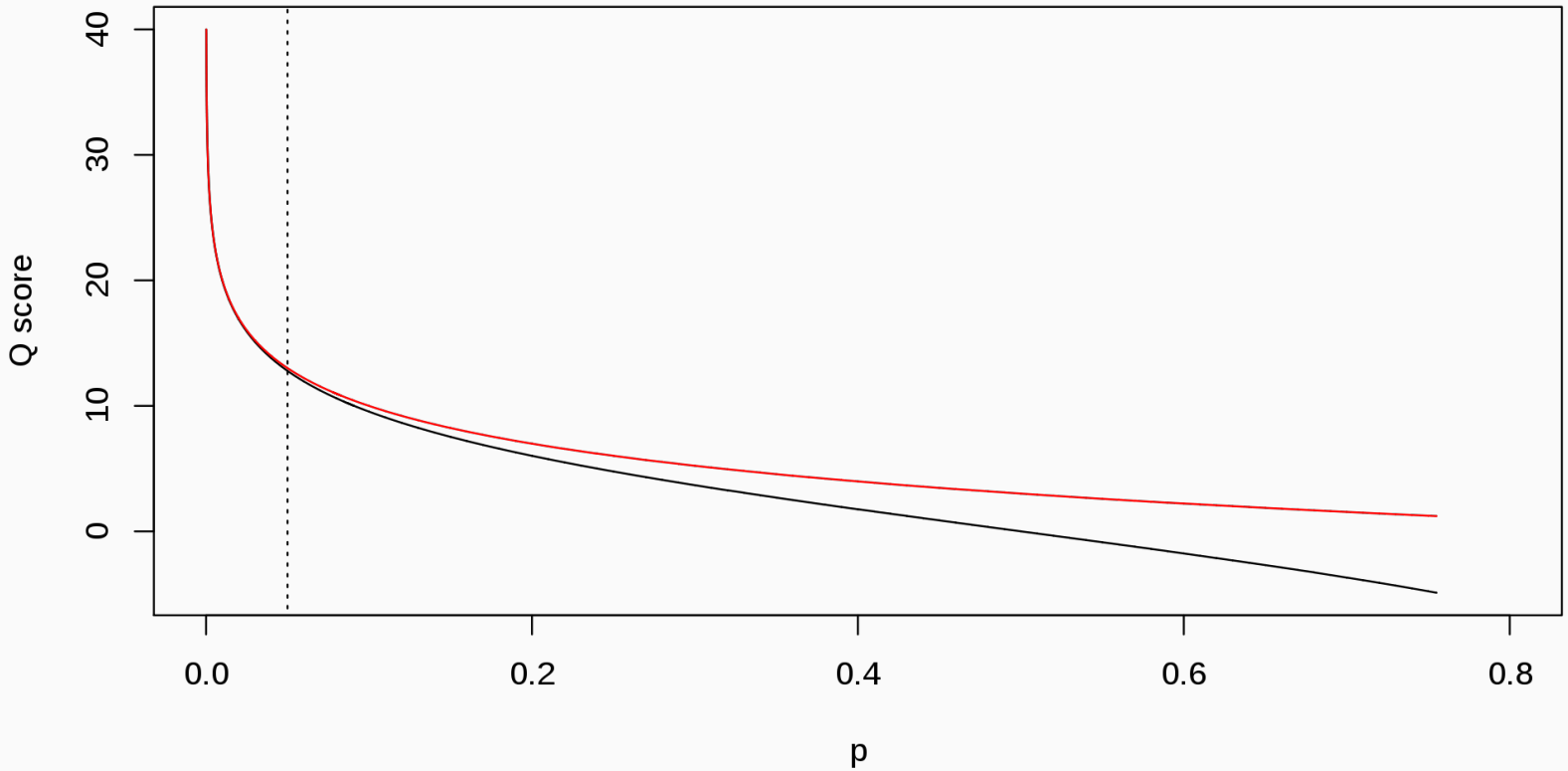
Limits : Size (150 + 150 in NovaSeq, but 400 for MySeq)

The FASTQ Format

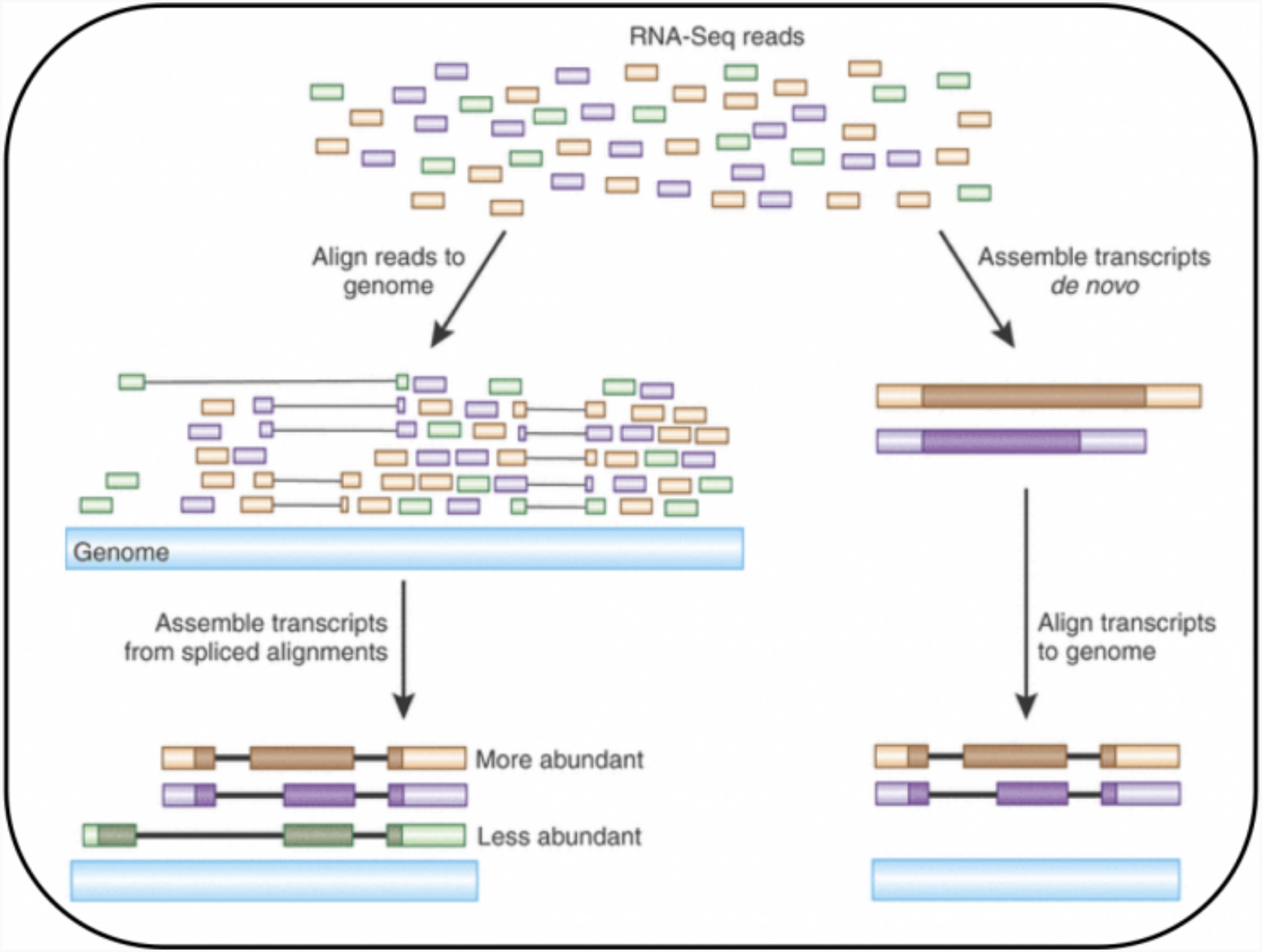


The QPHRED Scale

The QPHRED Value



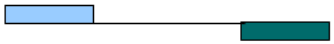
Mapping



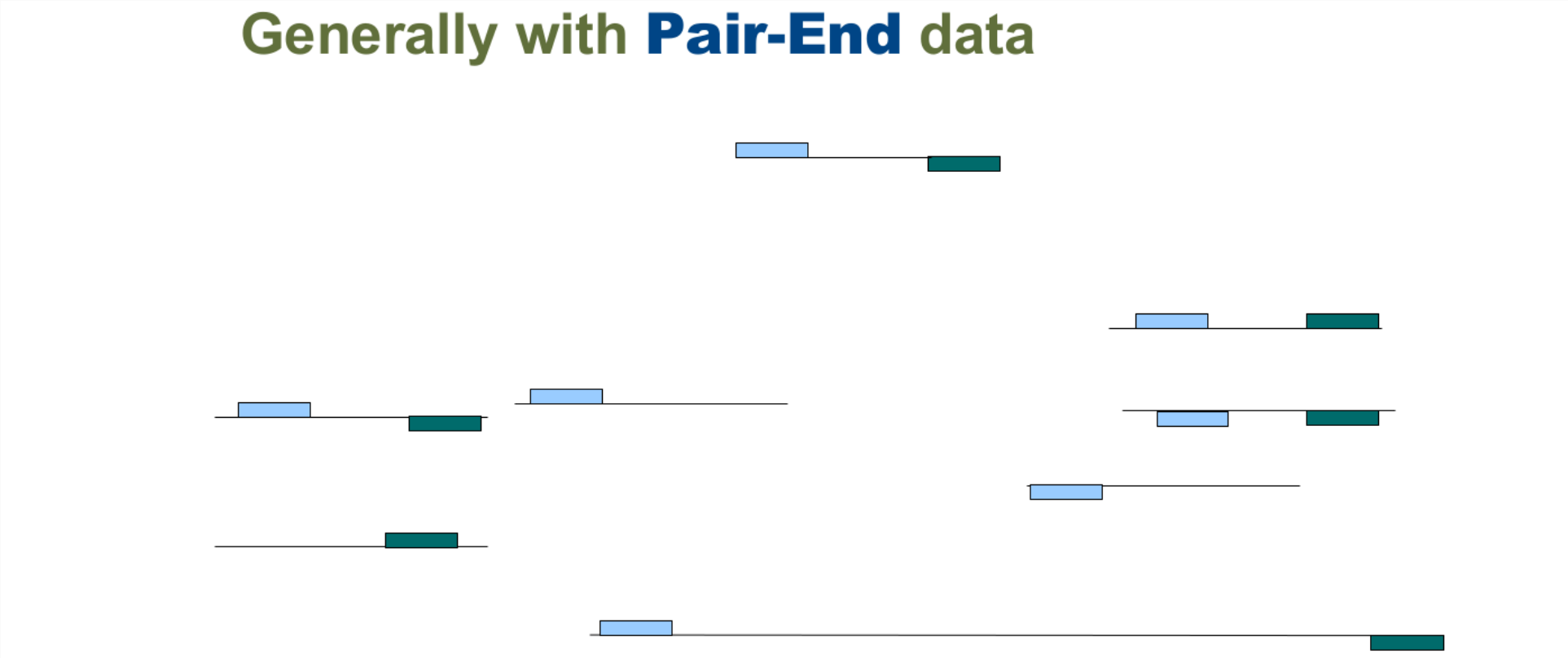
From EMBL-EBI

Mapping approaches for differences

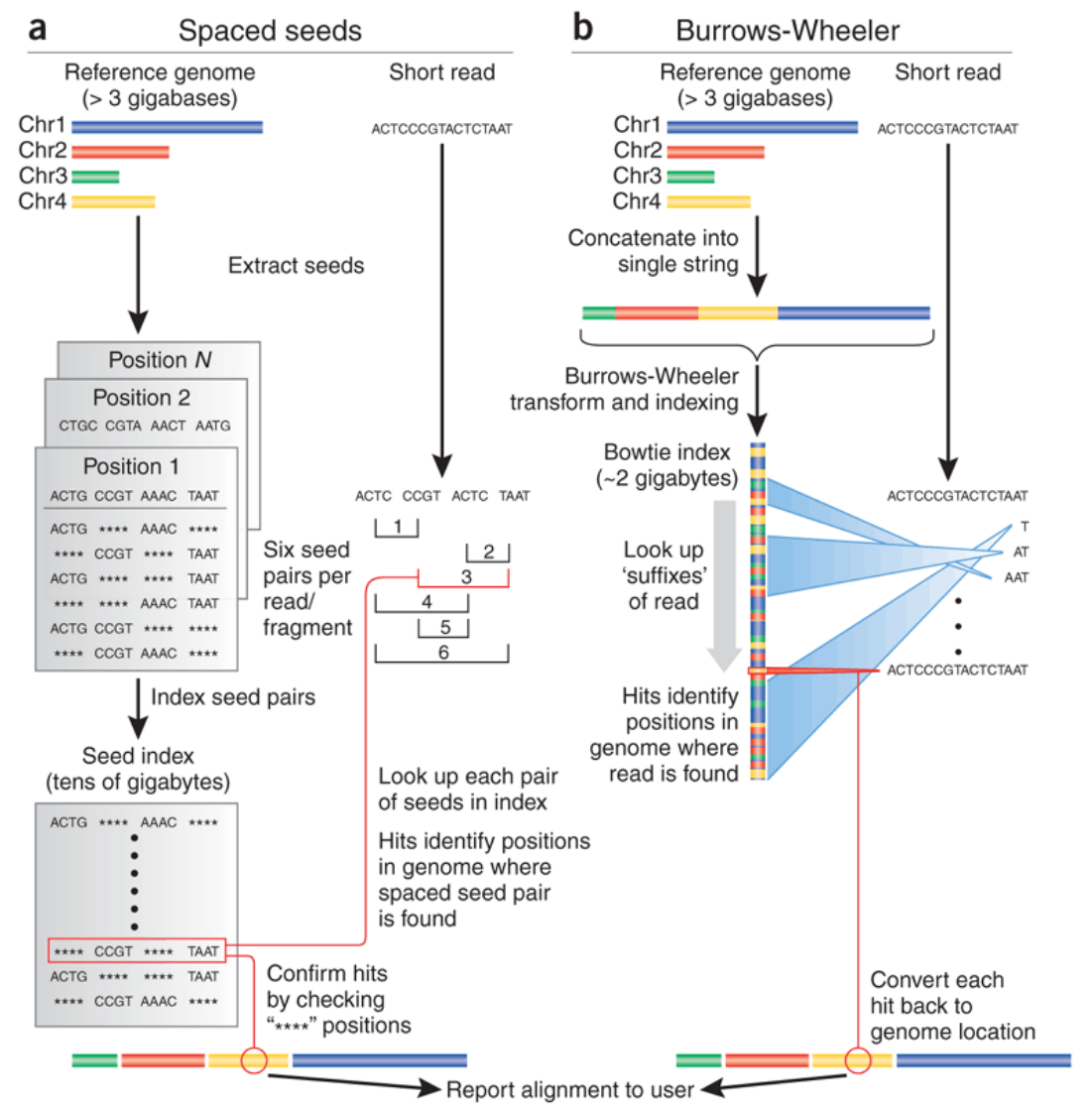
Generally with **Pair-End** data



Mapping approaches for differences

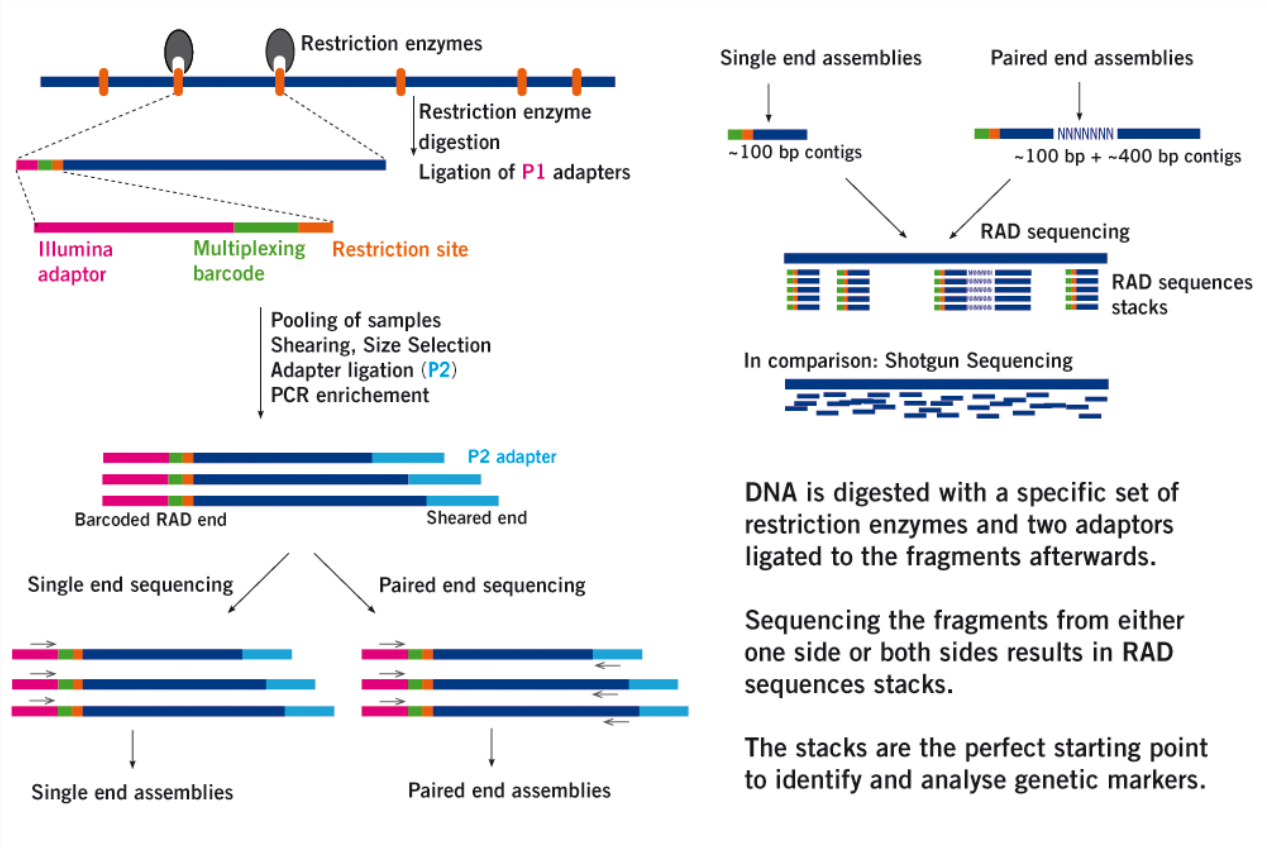


Mapping



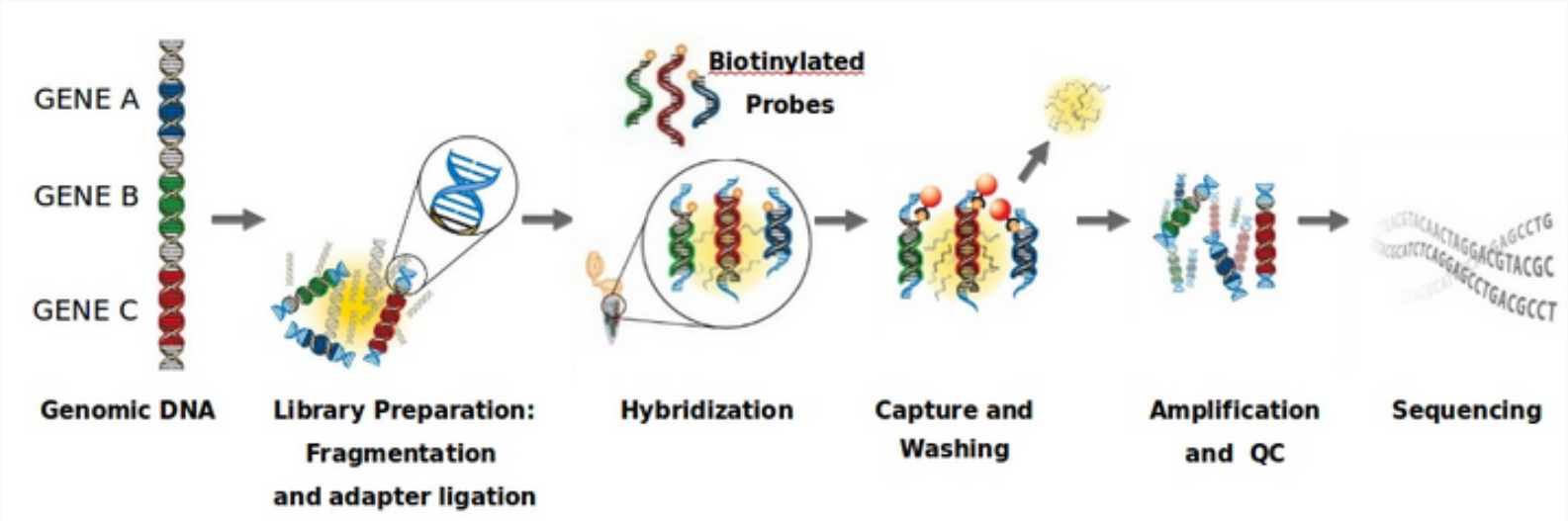
From Trapnell & Salzberg, 2006

RAD-Seq



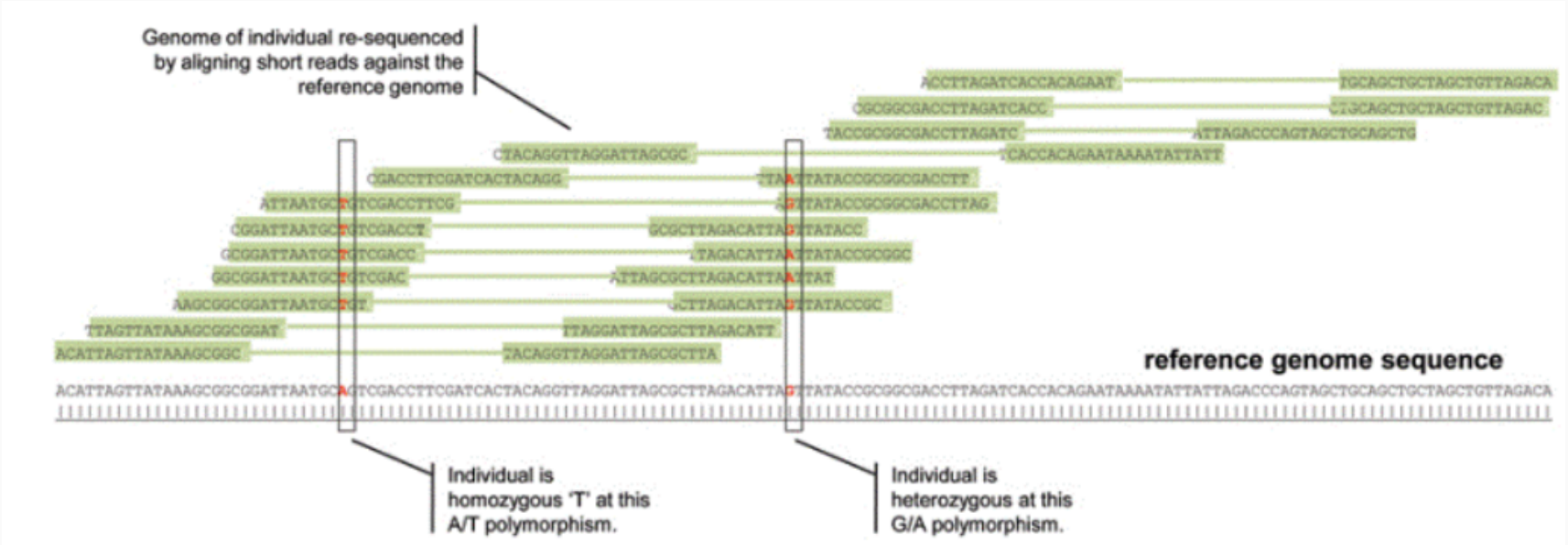
From Eurofins

CaptureSeq

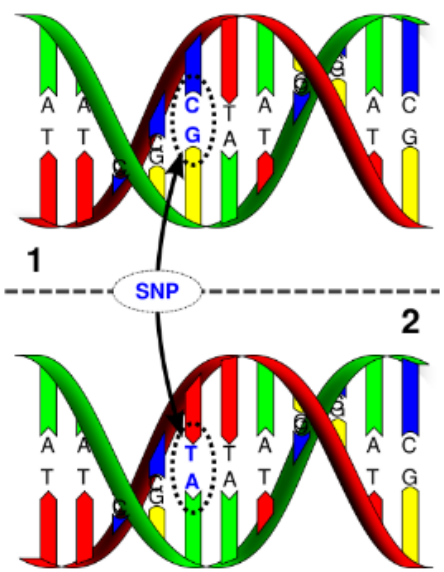


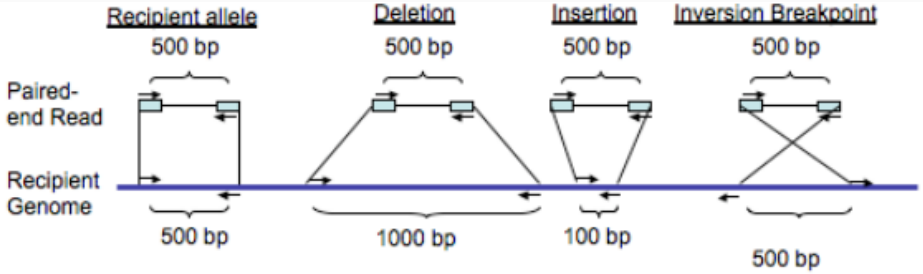
From CGFB, Bordeaux, France

SNP and InDel Detection



SNP and InDel Detection





Types of variants

SNPs

Alignment
```
ACGT
ATGT
```
VCF representation
```
POS REF ALT
2   C   T
```

Insertions

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

Deletions

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

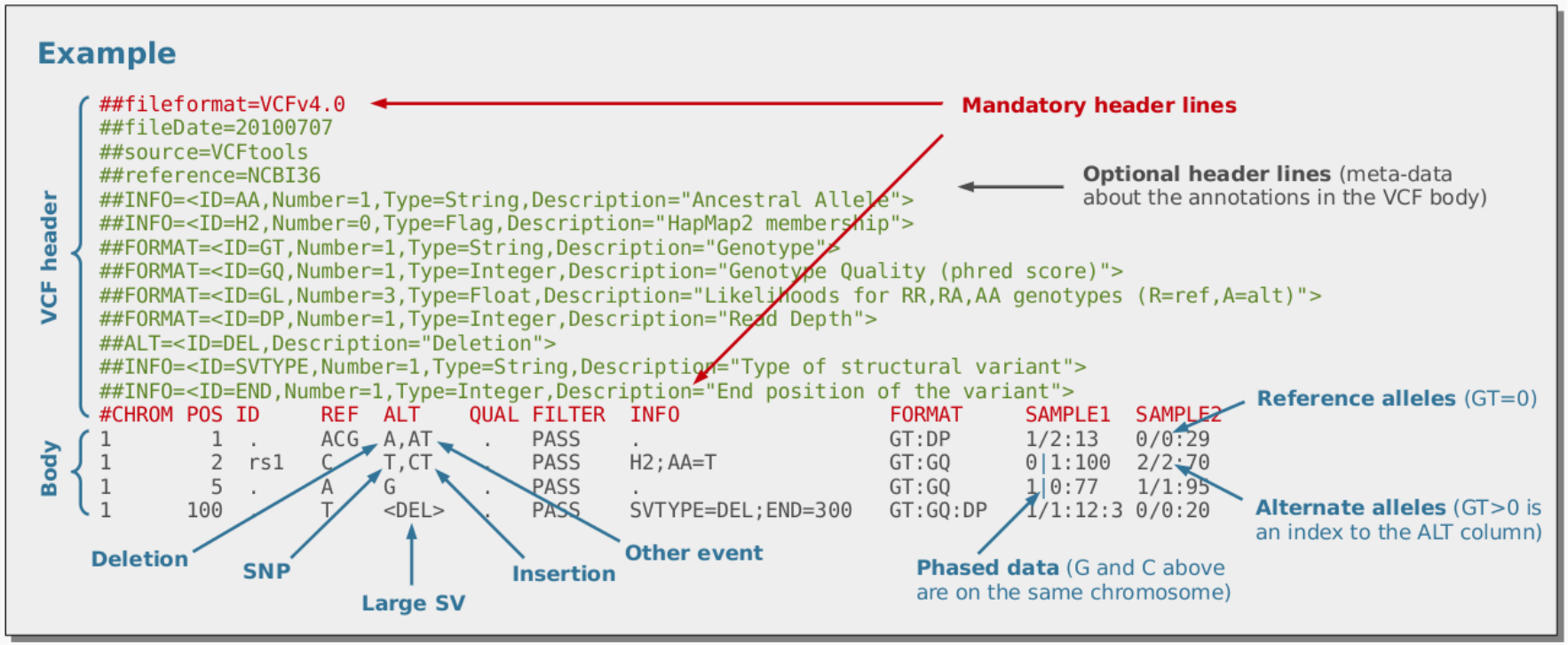
Complex events

Alignment
```
ACGT
A-TT
```
VCF representation
```
POS REF  ALT
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



VCF = Variant Call Format From 1000 Genomes Project

Limits, Wetlab

- 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

Limits, Technical

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

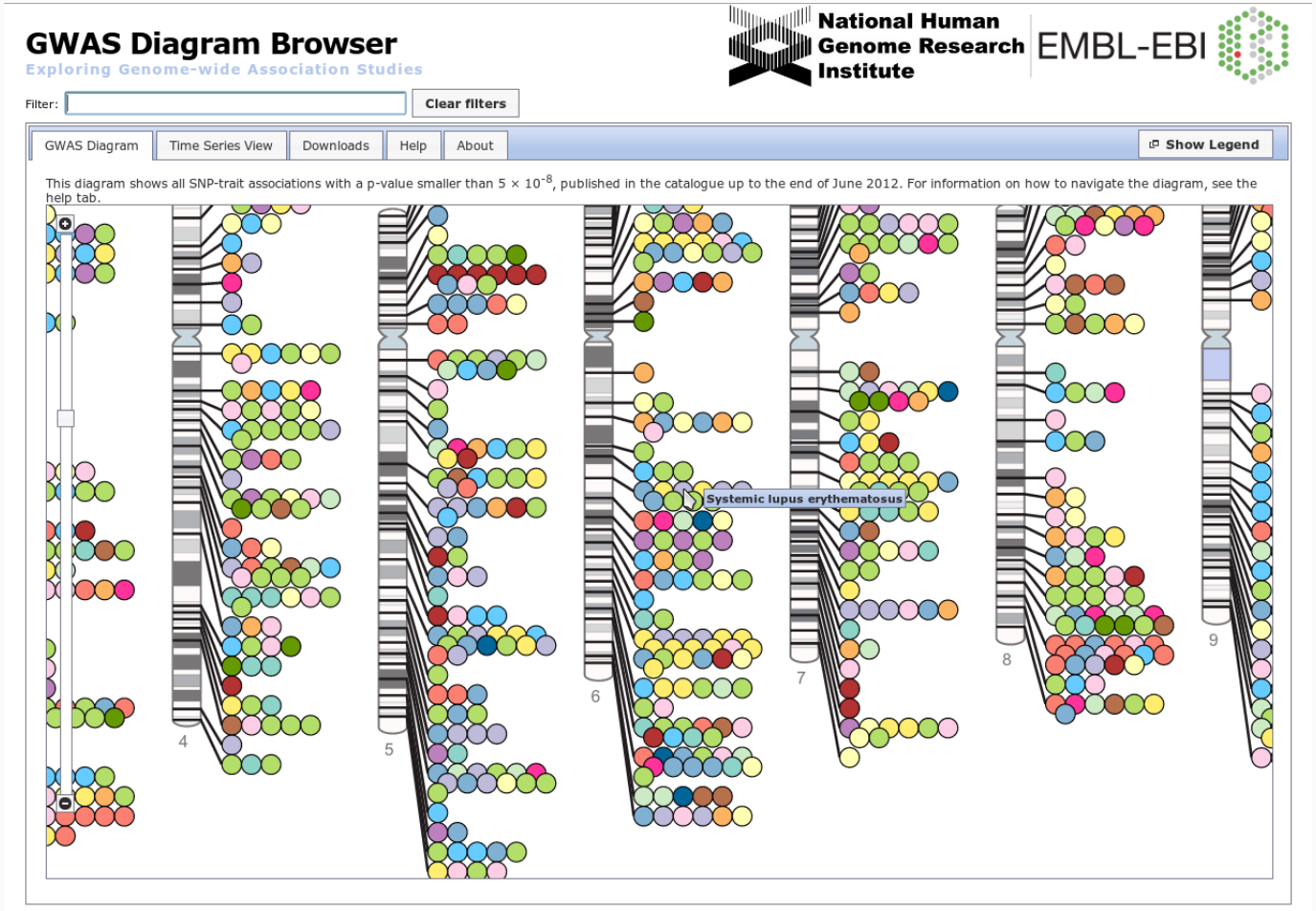
All linked to the Specificity/Sensitivity Informatics Paradox

Applications

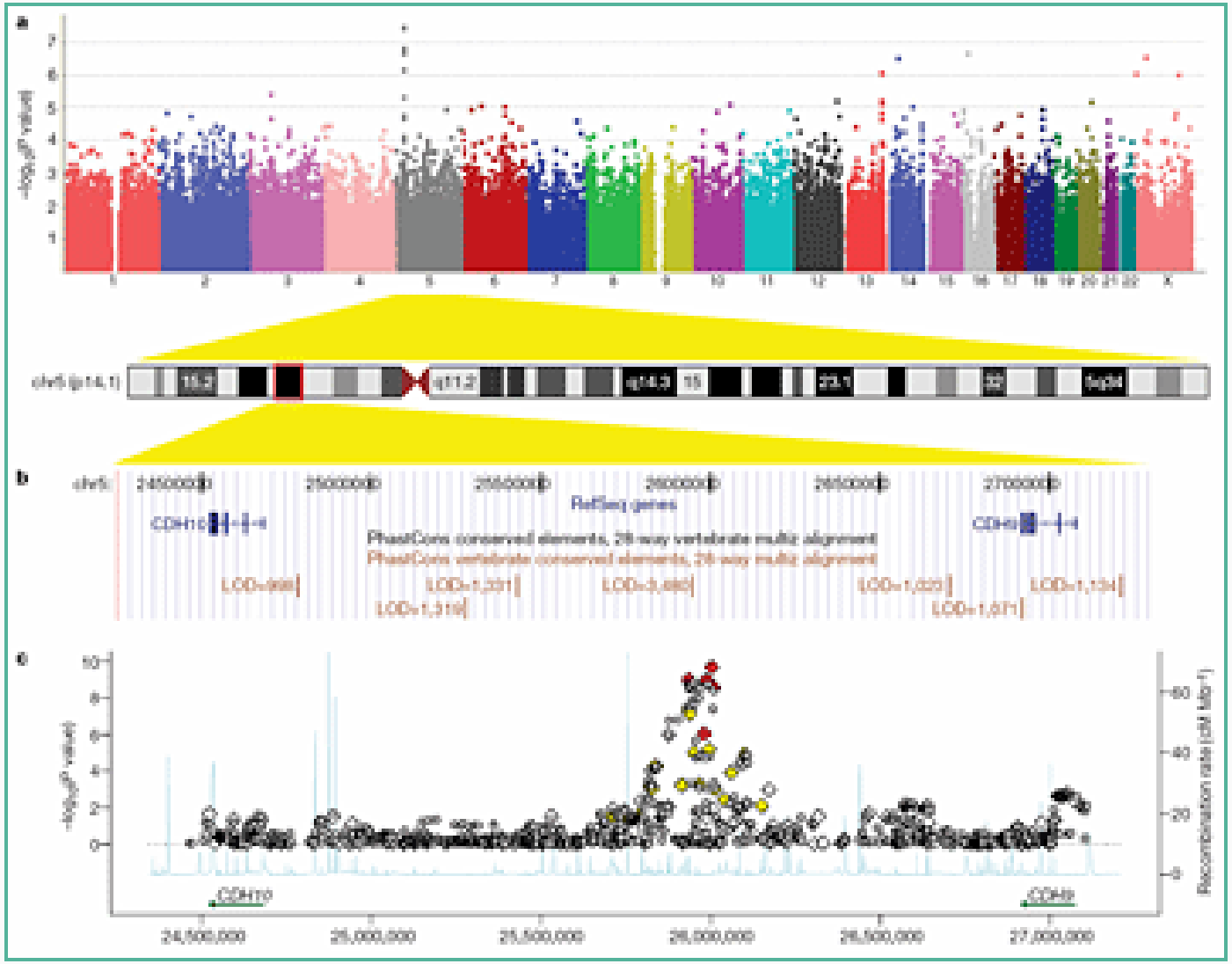
Population Genomics/Genomics

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

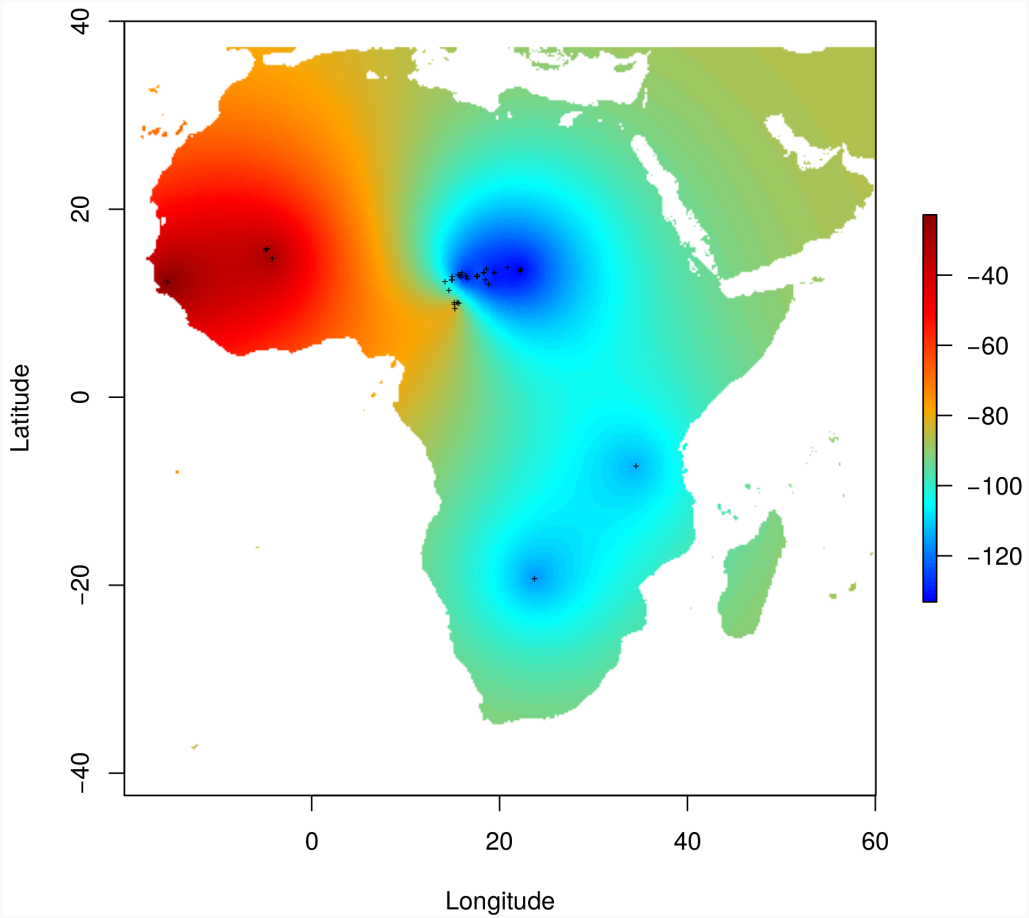
Example in GWAs & Population Genomics



Example in GWAs & Population Genomics



Example in Global Genotyping & Population Genomics



From Cubry et al, 2018

Large Projects

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

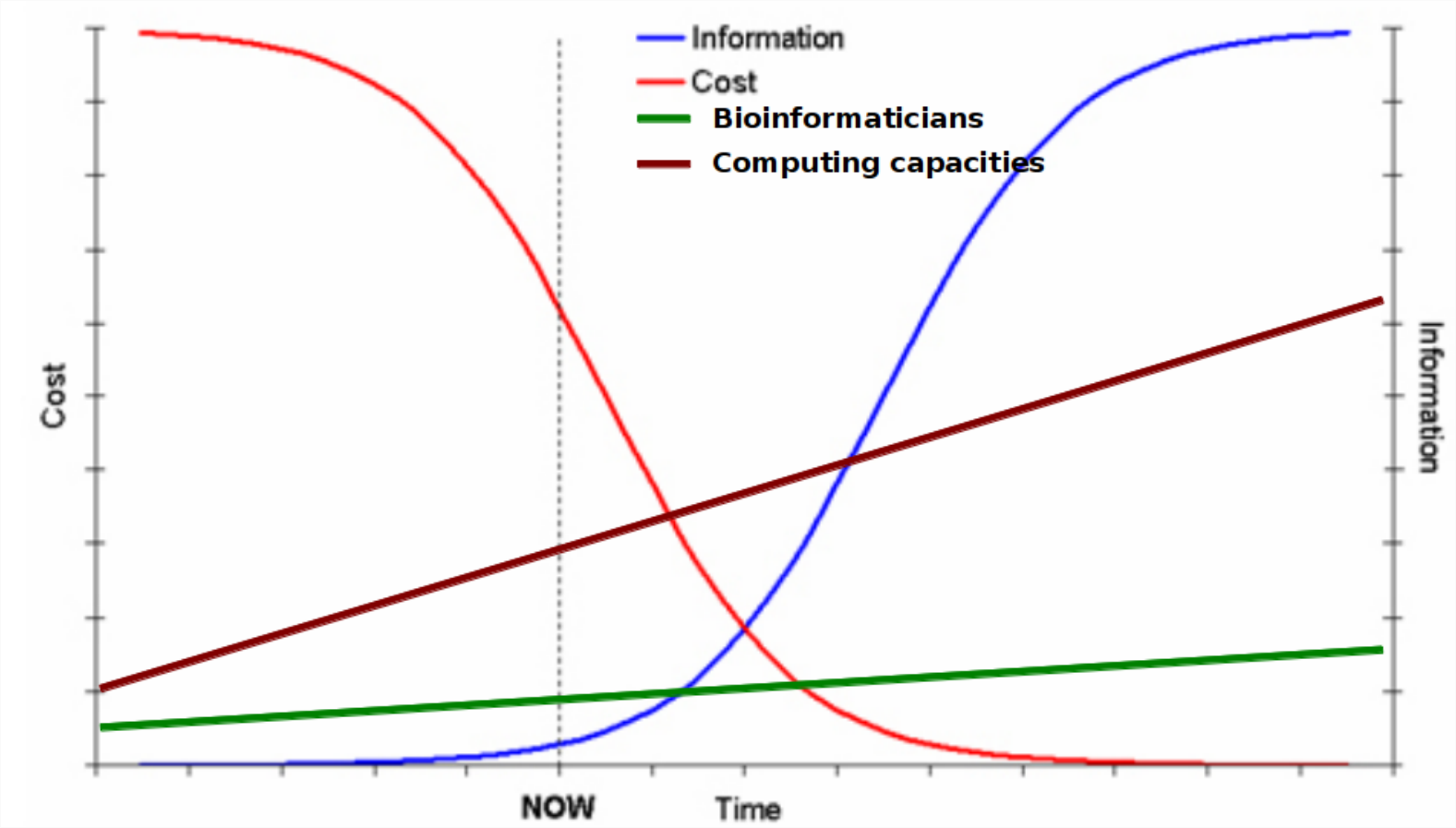
Possibilities in the next 5-10 years (From a presentation in 2013)

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

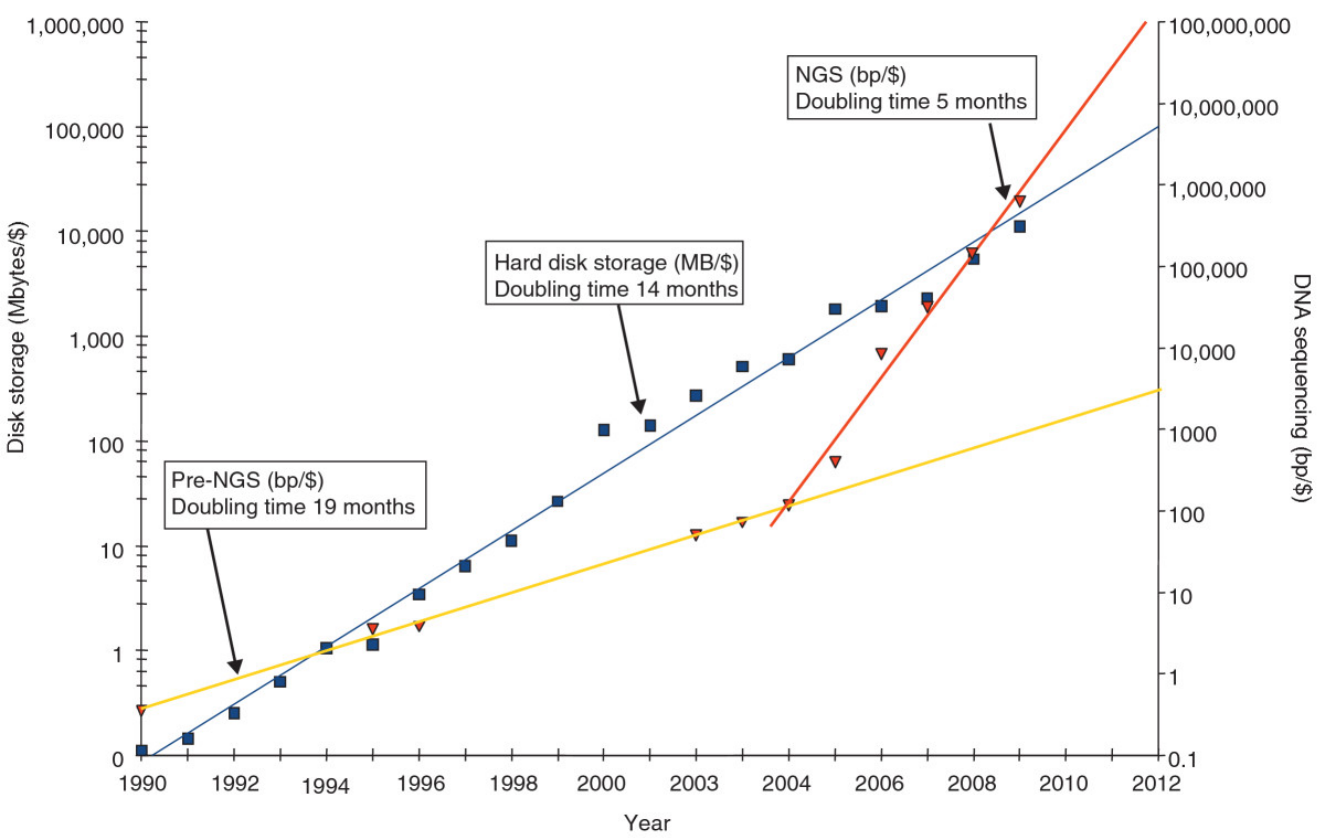
Possibilities in the next 5-10 years (From a presentation in 2013)

- 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!

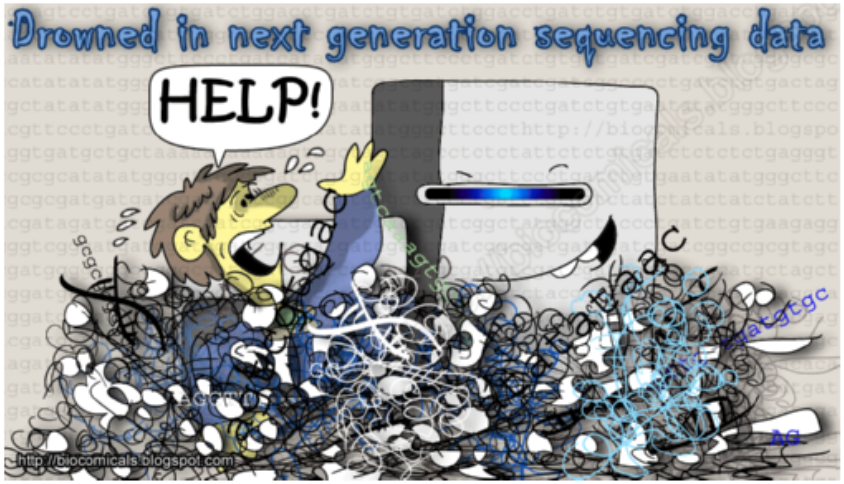

Information
Cost
Bioinformaticians
Computing capacities
Cost
Information
NOW
Time

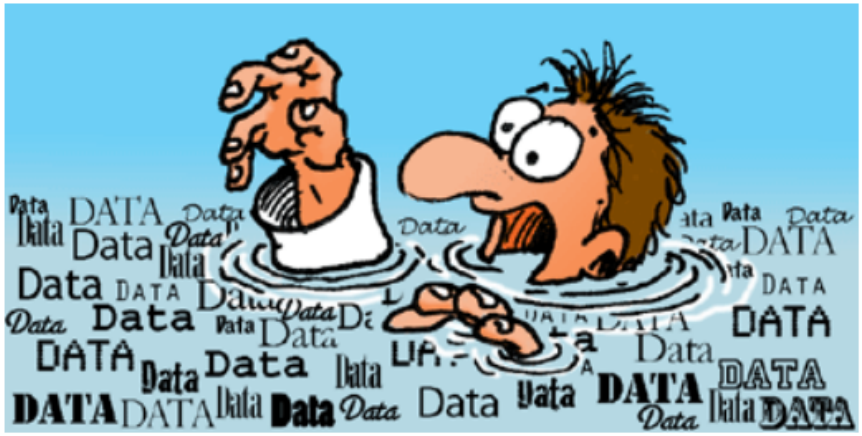
...From Data Rarity to Data Deluge



From L. Stein, 2010

Be Careful to data drowning!





Thanks for your attention