Omics or how to handle massive biologic data?

Emergence of new sciences

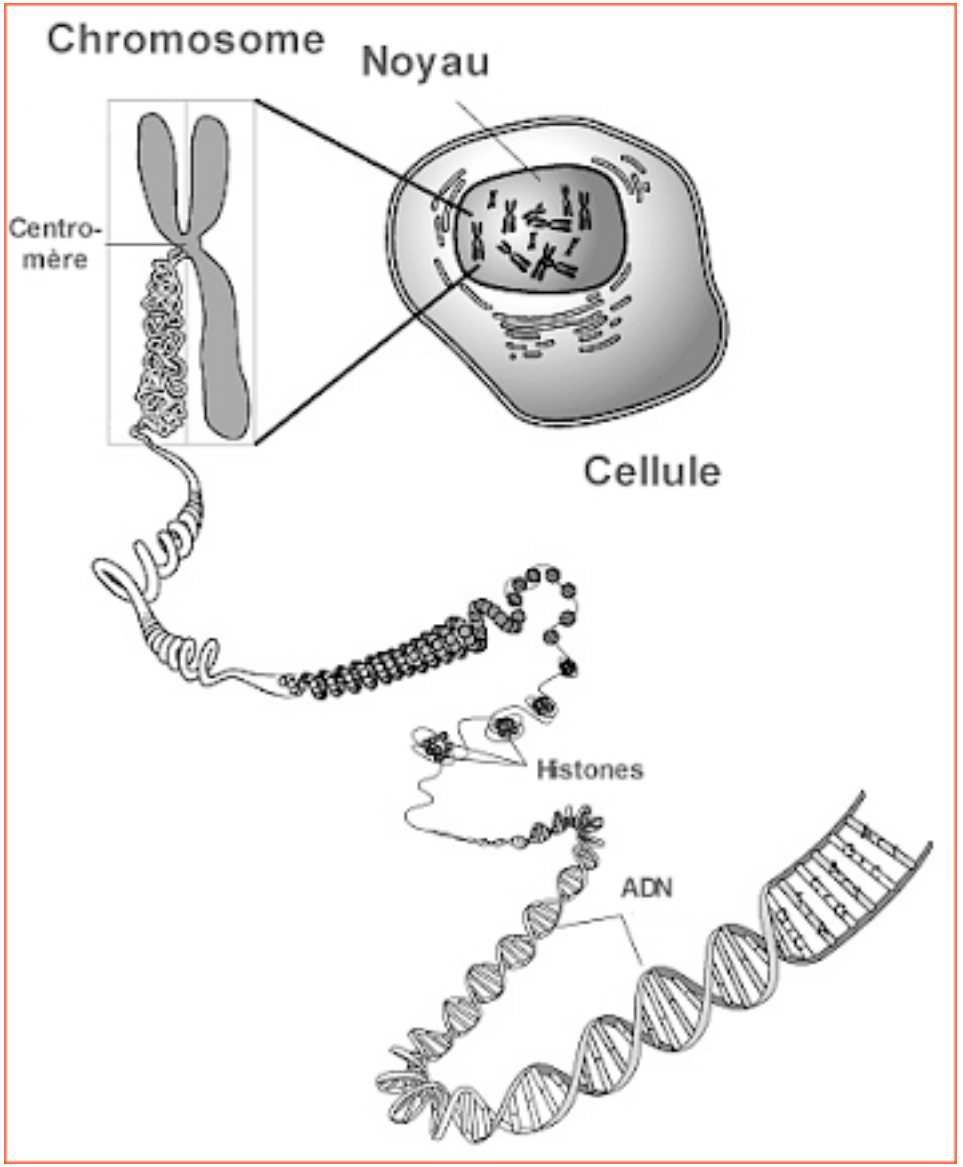
Anna-Sophie Fiston-Lavier

Background

(1) *Where and how genetic information is organized*

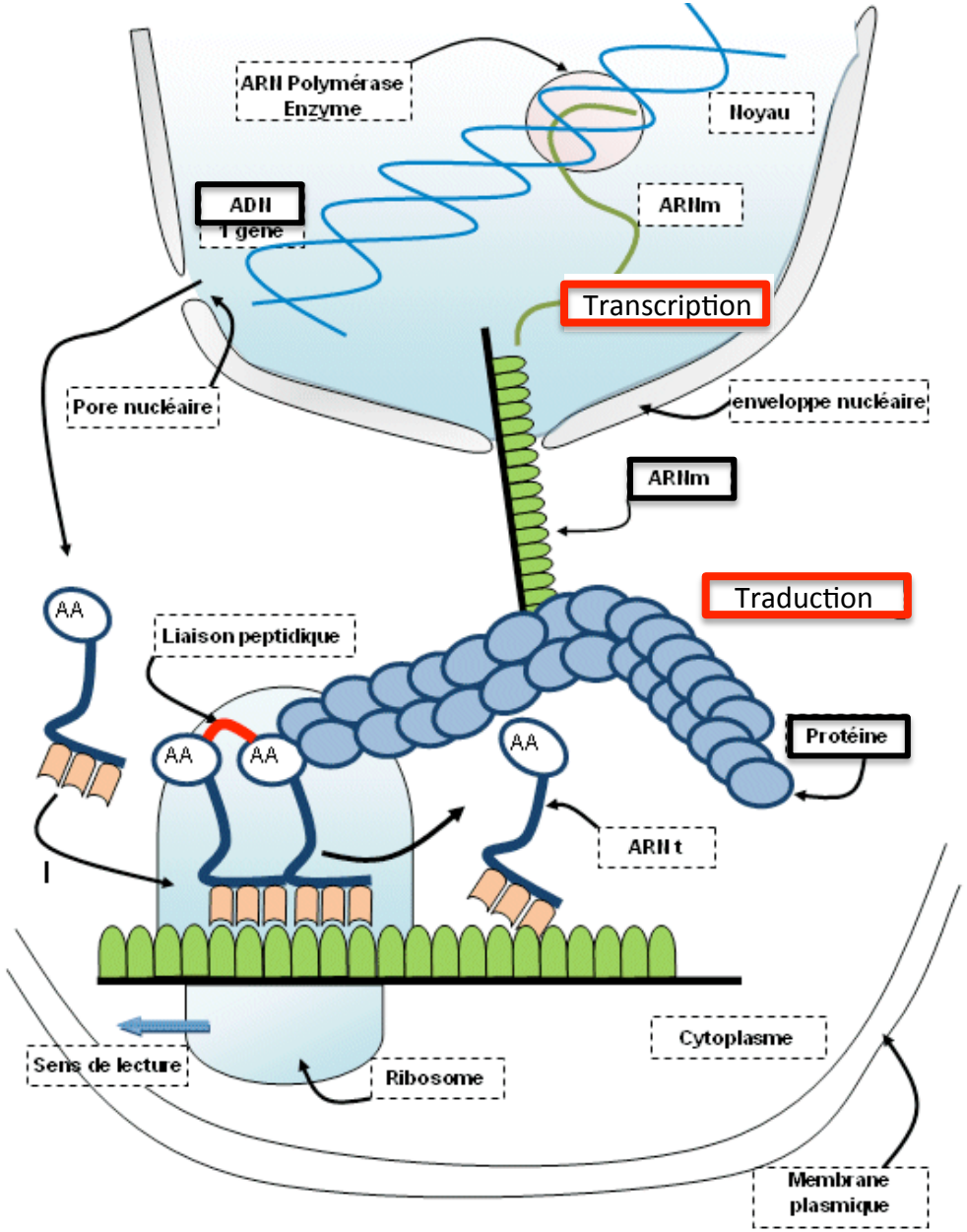
Cell = Unit of the organisms

DNA = Unit of the genetic information
String composed of [ACGT]



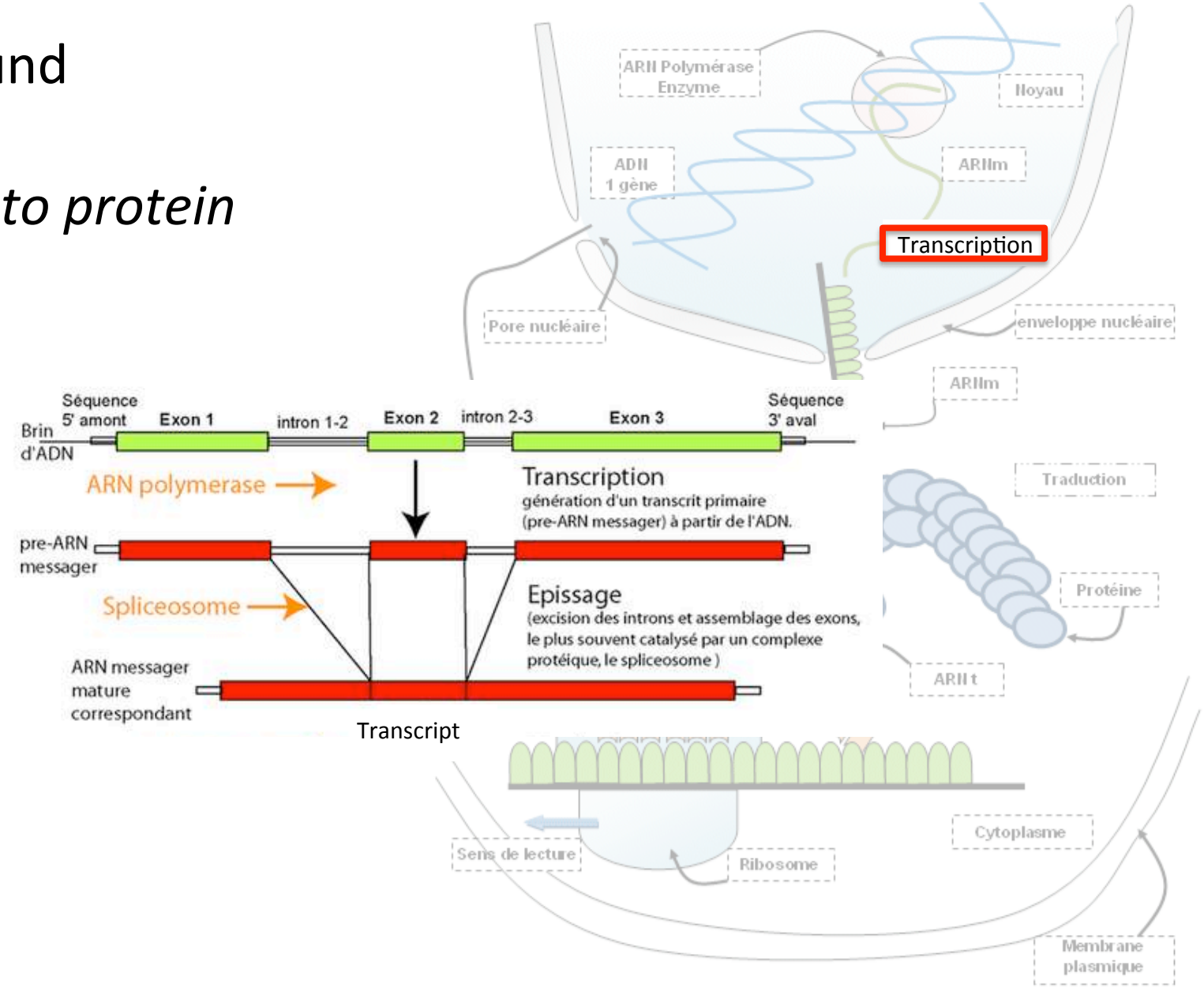
Background

(2) *Gene to protein*



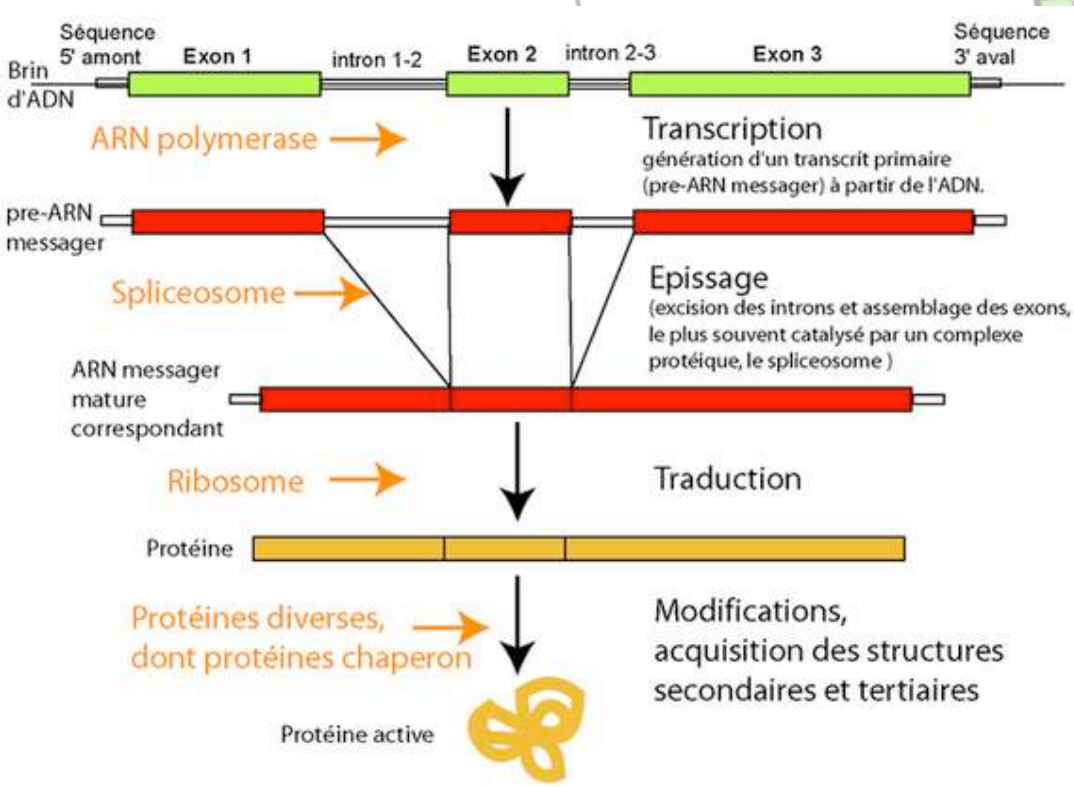
Background

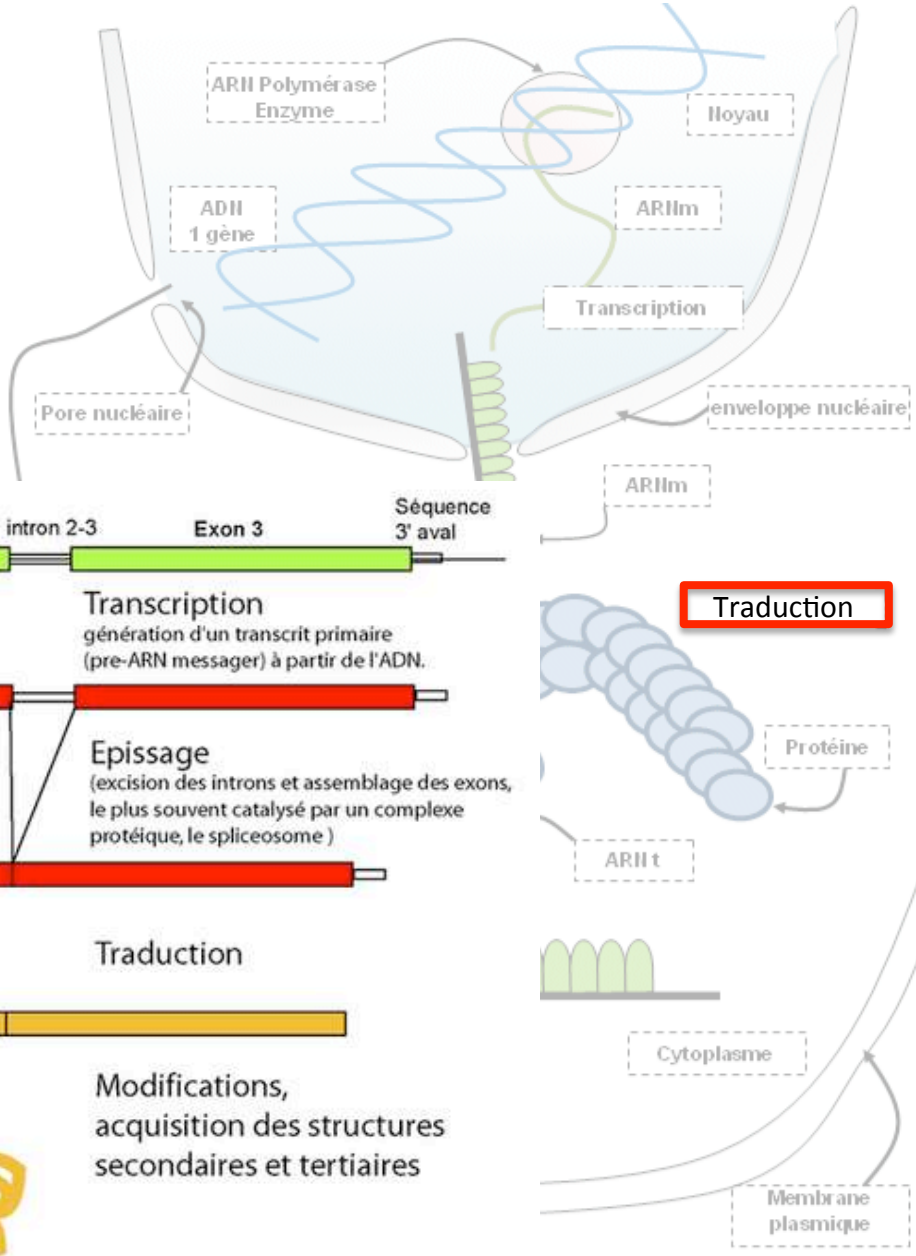
(2) *Gene to protein*



Background

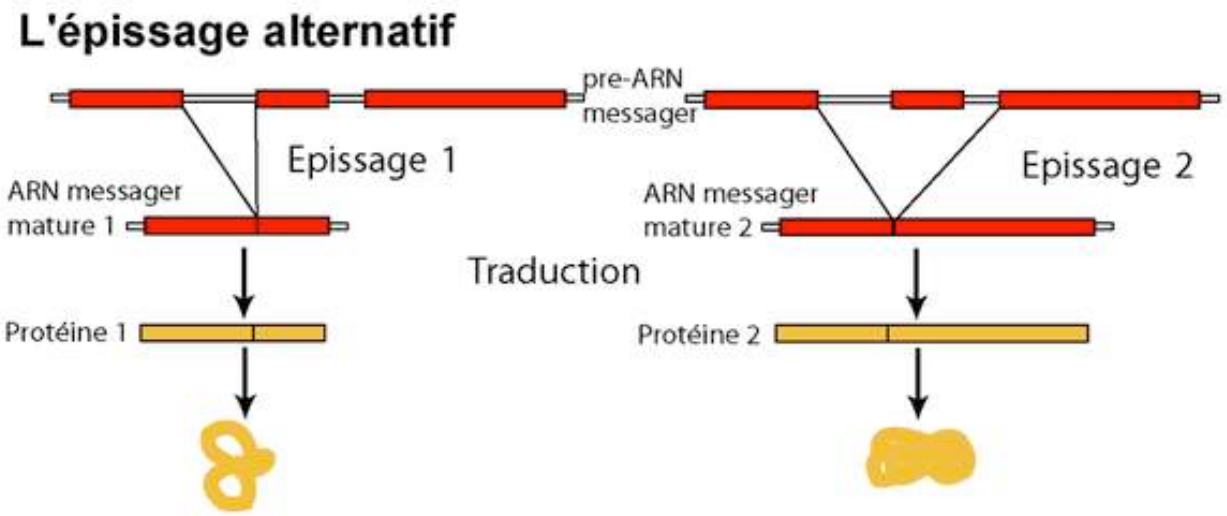
(2) *Gene to protein*

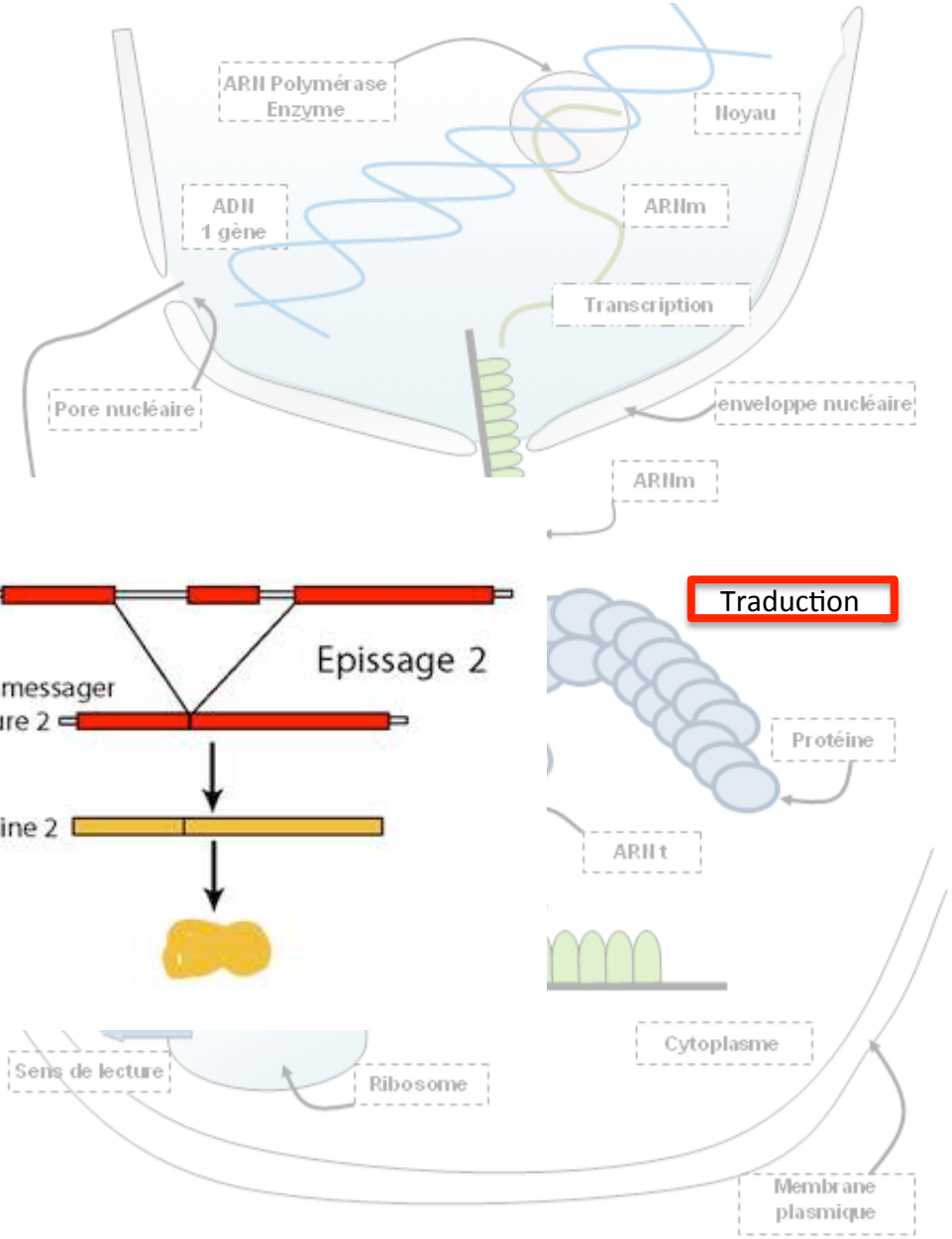




Background

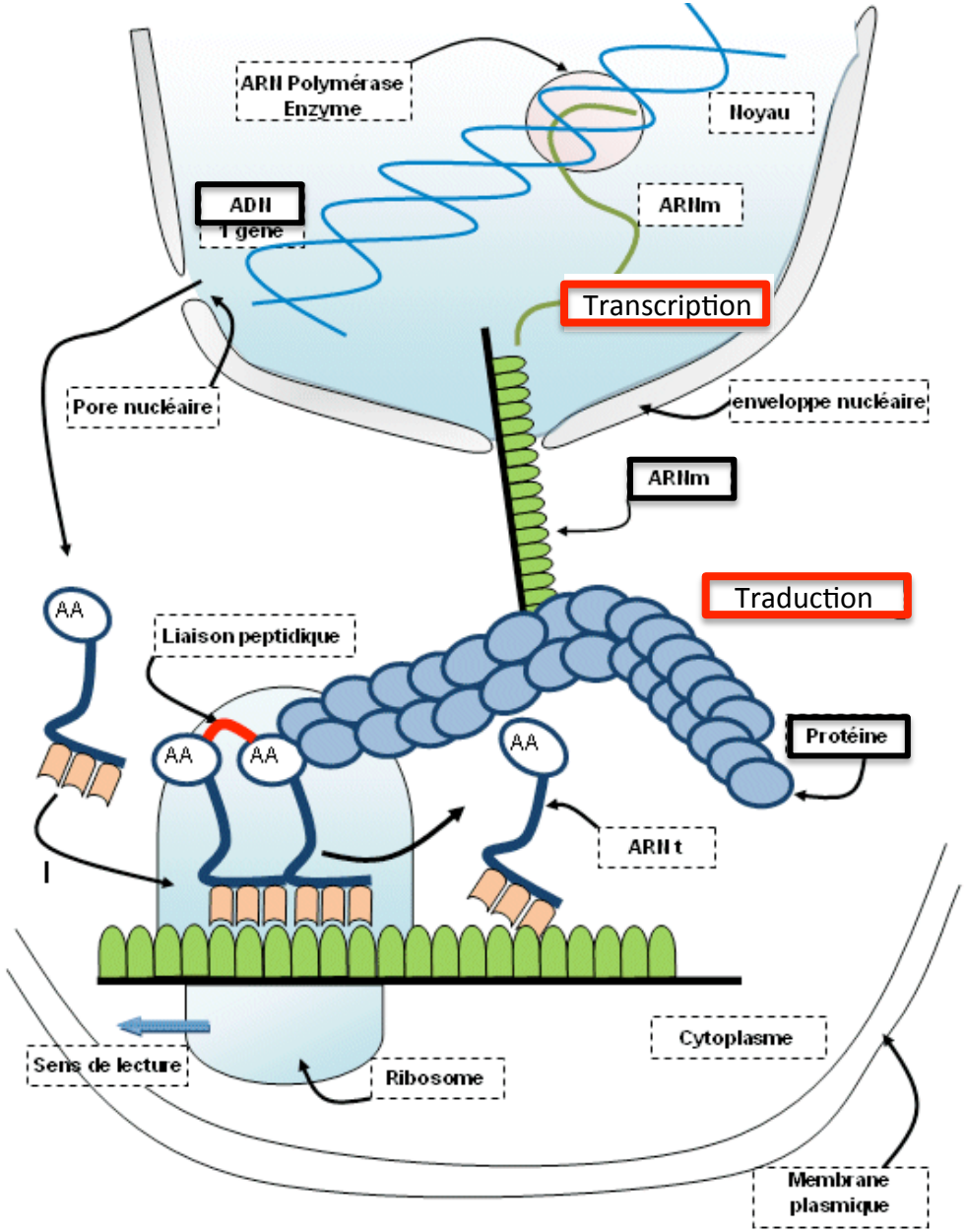
(2) *Gene to protein*



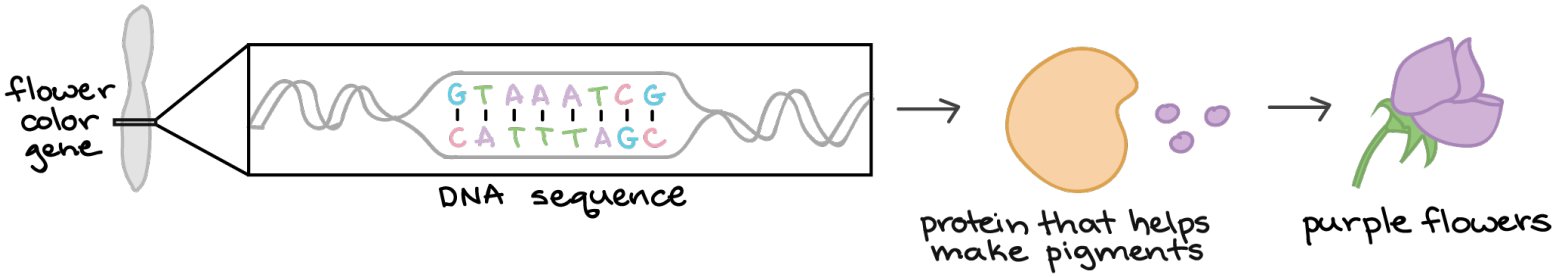


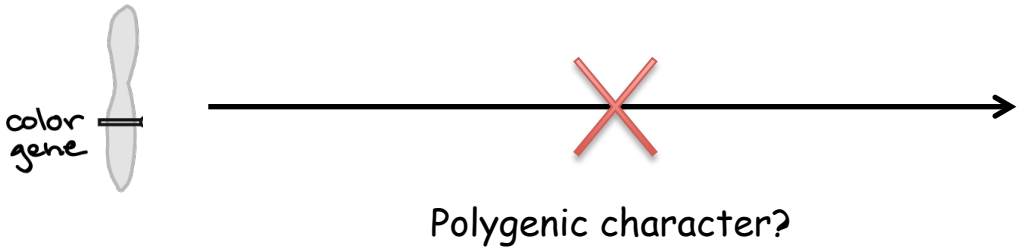
Background

(2) *Gene to protein*



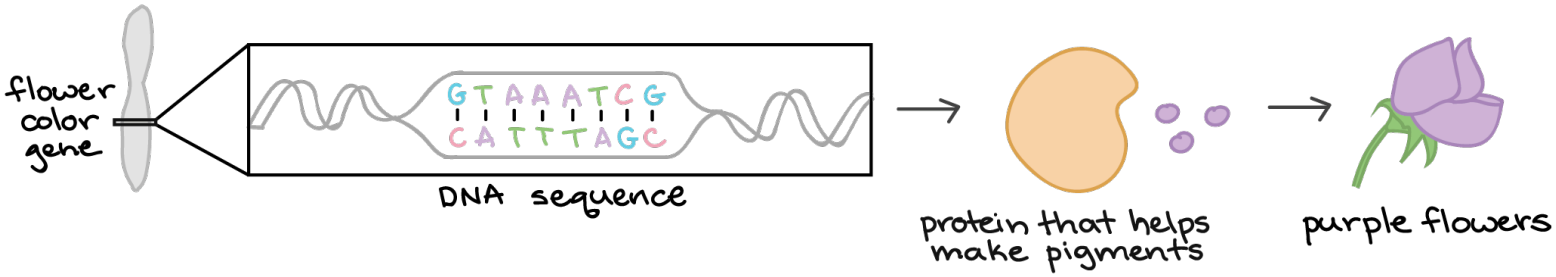
Dogma

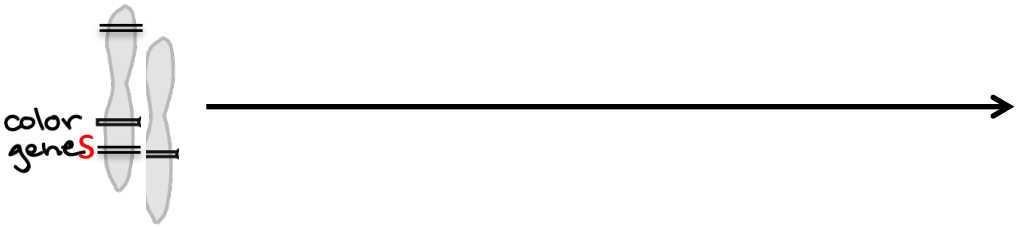




A high diversity of phenotypes

Dogma

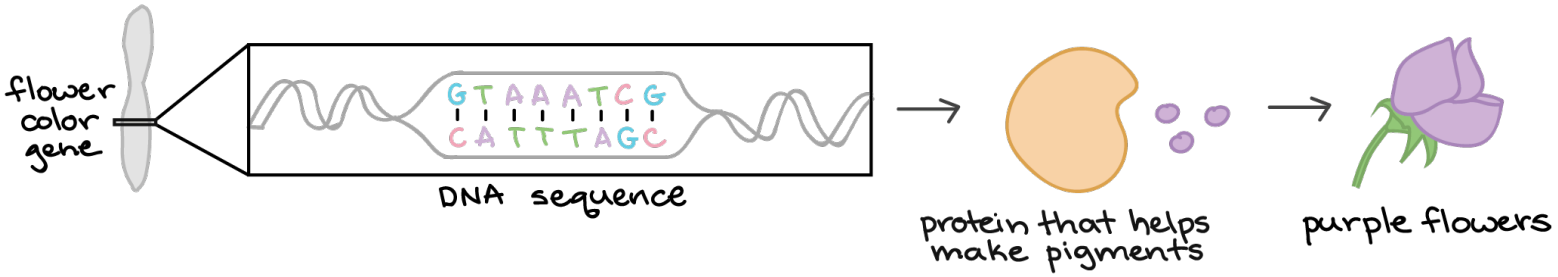


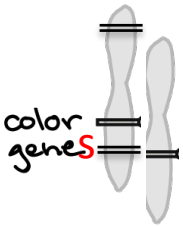


Loci
EYCL1,
EYCL2,
EYCL3 and
OCA2

A high diversity of phenotypes
biometric identification

Dogma





Loci EYCL1, EYCL2, EYCL3 and *OCA2*

Genetic is much more complex

A high diversity of phenotypes biometric identification

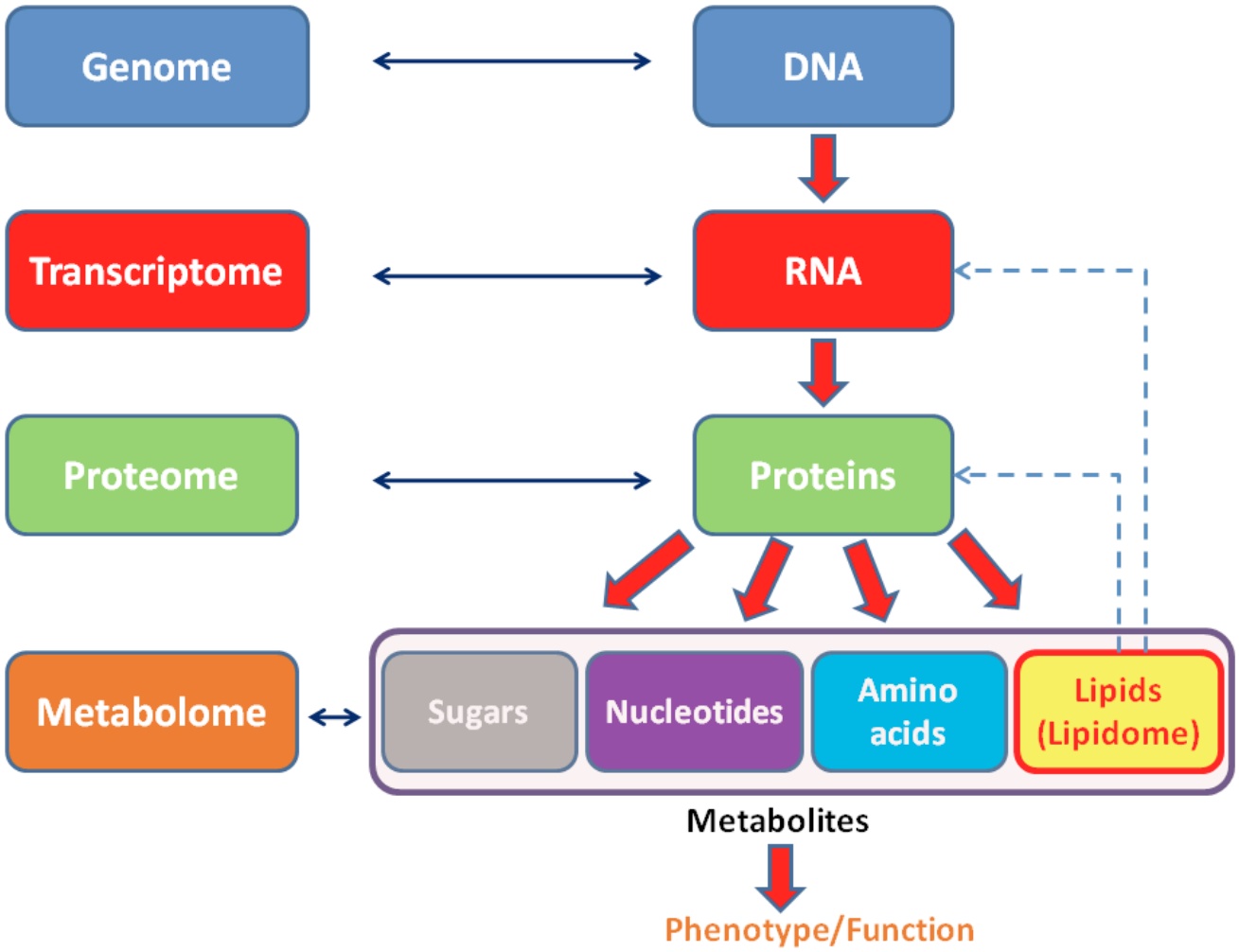
les Mélanésiens et les noirs de l'île Salomon ?
Another gene involved : **TYRP1**

"Omics" a new way of thinking

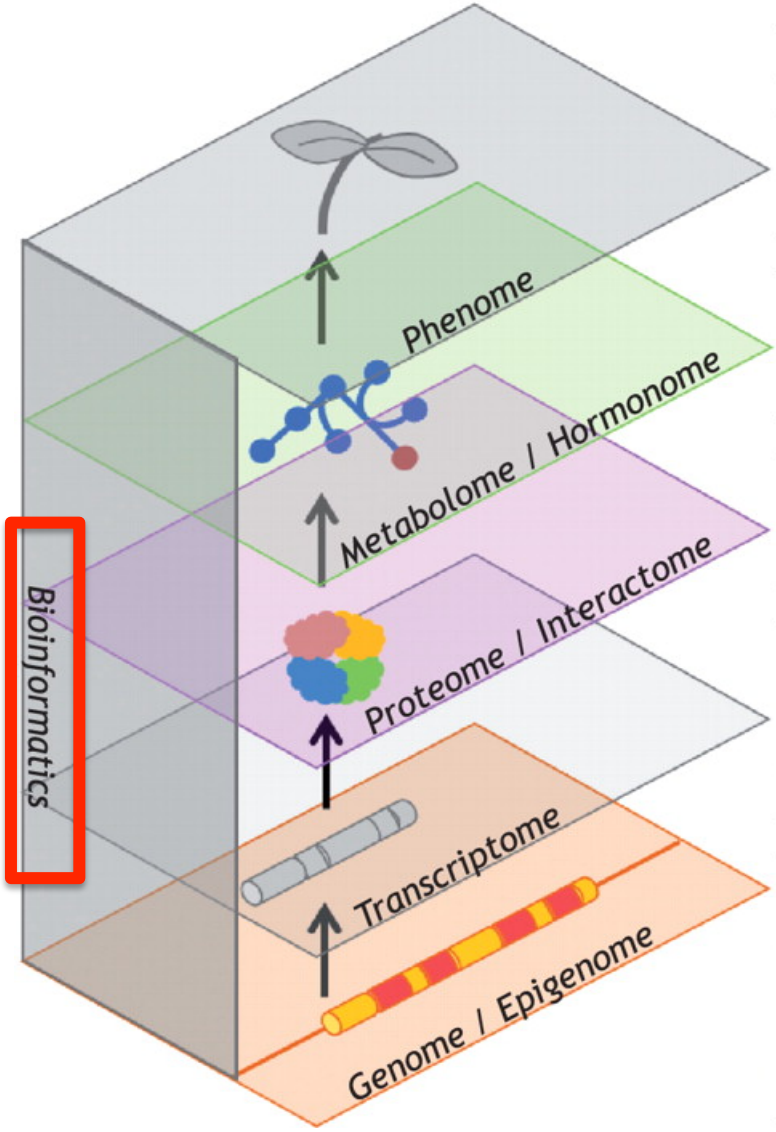
Ome = set of objects of study of such fields

Omics = field of study in biology ending in this suffix

The “Ome”s


Genome
DNA
Transcriptome
RNA
Proteome
Proteins
Metabolome
Sugars
Nucleotides
Amino acids
Lipids (Lipidome)
Metabolites
Phenotype/Function

The “Omics”



Omics instances
Integrated database
Mutant lines
Natural variations
Metabolic map
Metabolome profiles
Hormonome profile
Proteome / modificome profiles
Subcellular localization
Interactome maps
Full-length cDNA clones, ESTs
Expression profiles
Non-coding RNA profiles
Co-expression network
Genome sequence, gene annotation
Re-sequencing
Focused gene family database (eg. Transcription factor)
DNA methylome
Chromatin epigenome

From Mochida and Shinozaki 2011

The “Omics” for what?

Identification of candidate genes

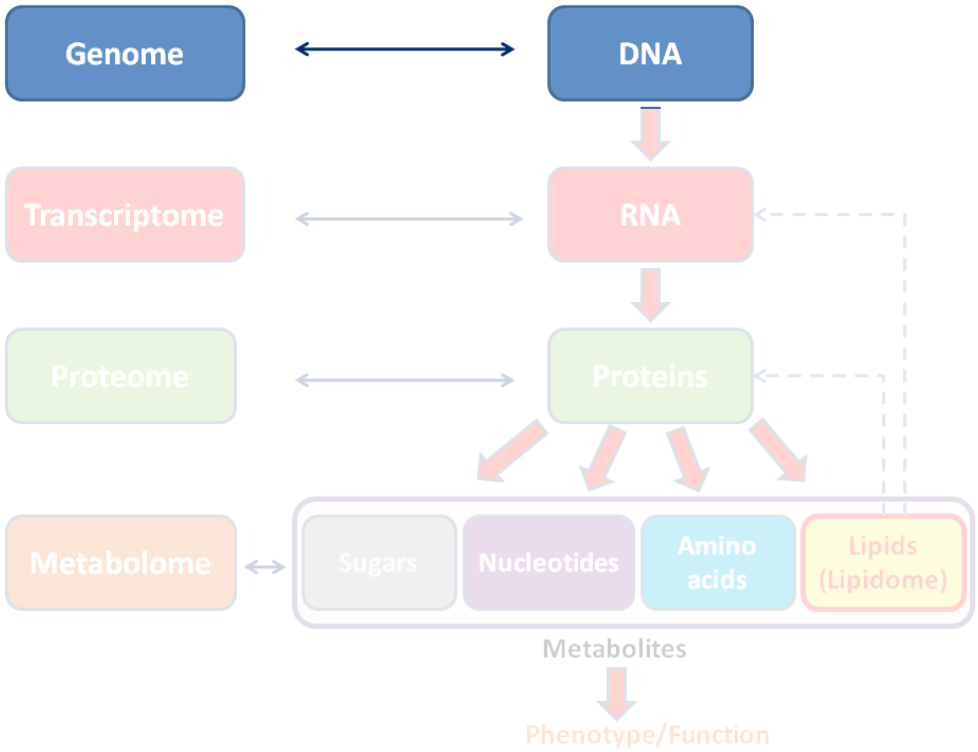
Identification of genes expressed

Identification of proteins producted

Identification of metabolites used in the cell

Genomics

Genome is a store of biological information.



Genomics is the study of whole sets of genes and their interactions.

Several fields in Gen**omics**

- **Structural genomics**
  Generate new sequence assemblies and study of the sequence organization

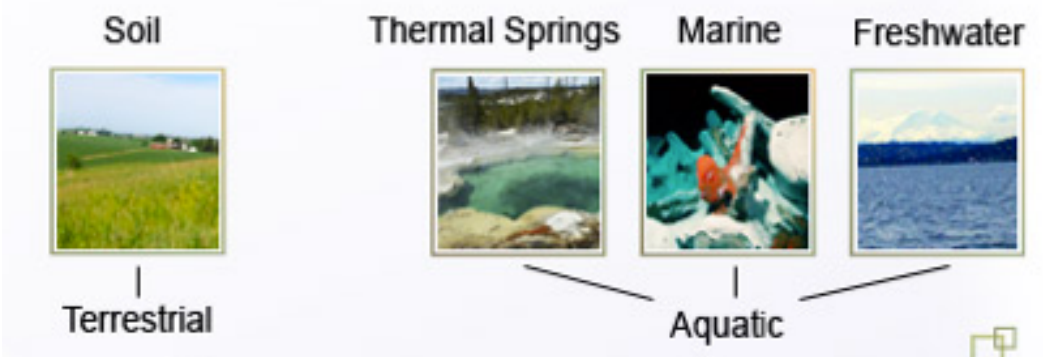
- **Comparative genomics**
  Identification conserved/specific genomic sequences among species and investigate their evolutionary history

- **Functional genomics**
  Link the known function of genes and their presence/absence and interaction

- **Metagenomics**
  Identification of species from a environmental sample (soil, water, gut…)

Metagen**omics**

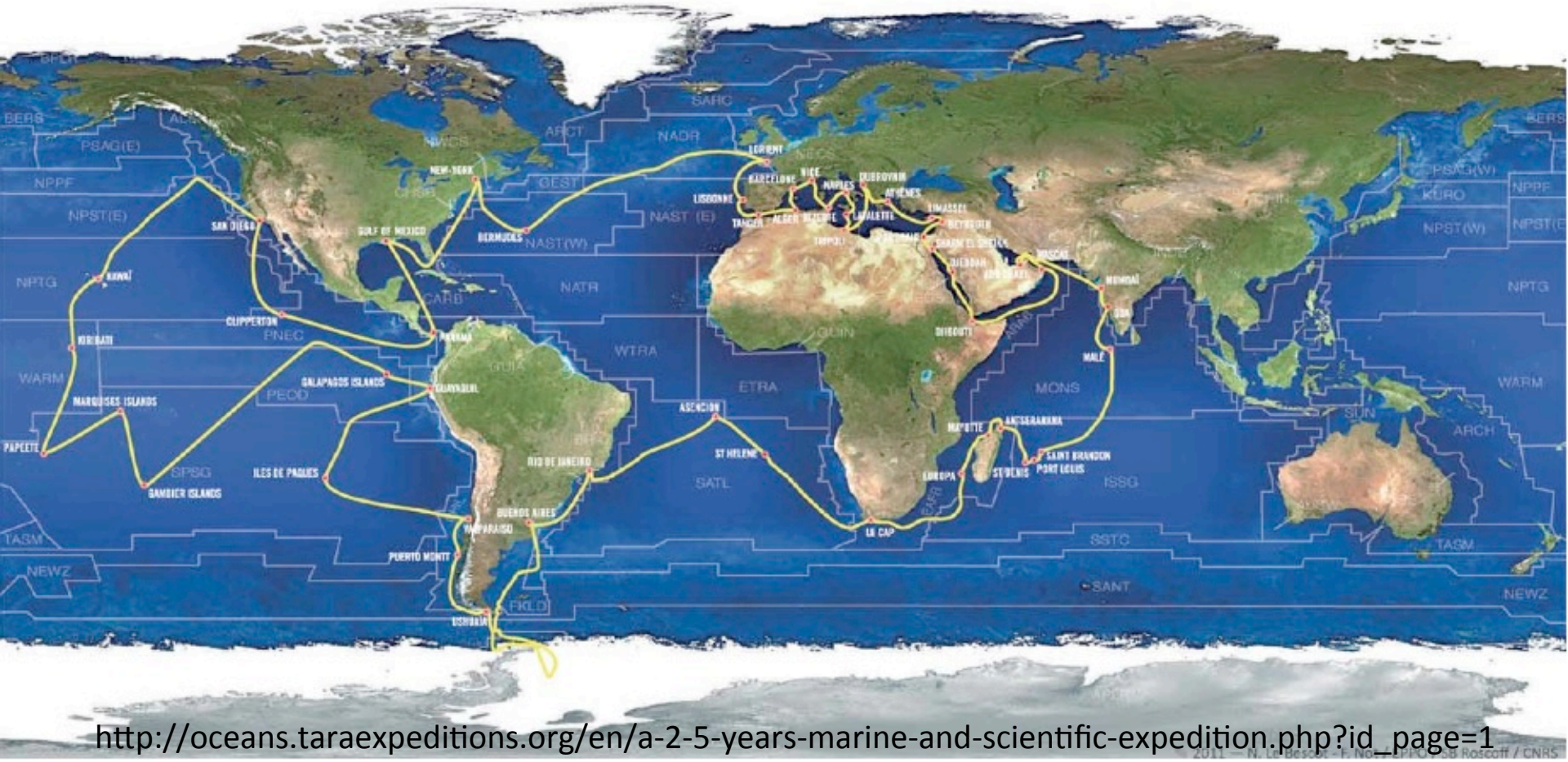
Technological advances have also facilitated **metagenomics**, in which DNA from a group of species (a **metagenome**) is collected from an environmental sample and sequenced.



Total DNA extracted - non-DNA impurities removed

This technique has been used on microbial communities, allowing the sequencing of DNA of mixed populations, and eliminating the need to culture species in the lab.

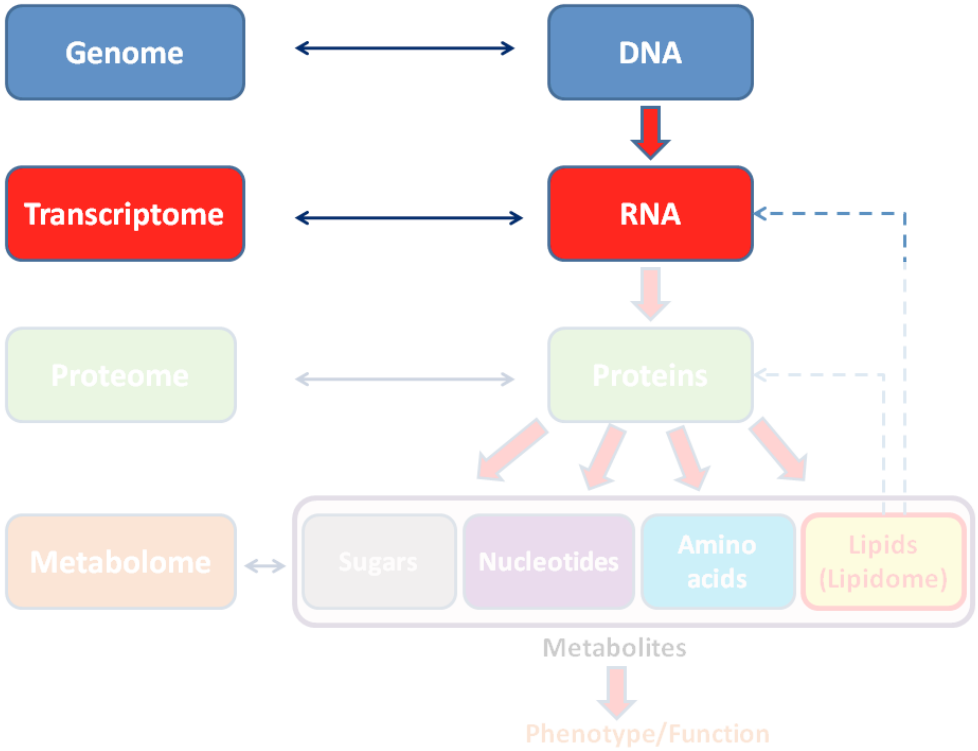
TARA, a metagenomic project to understand marine and soil ecosystems



http://oceans.taraexpeditions.org/en/a-2-5-years-marine-and-scientific-expedition.php?id_page=1

Transcript**omics**

Transcriptome = complete set of all RNA molecules ("transcripts") produced from a genome **OR** specific subset of transcripts present in a particular cell type or under specific growth conditions

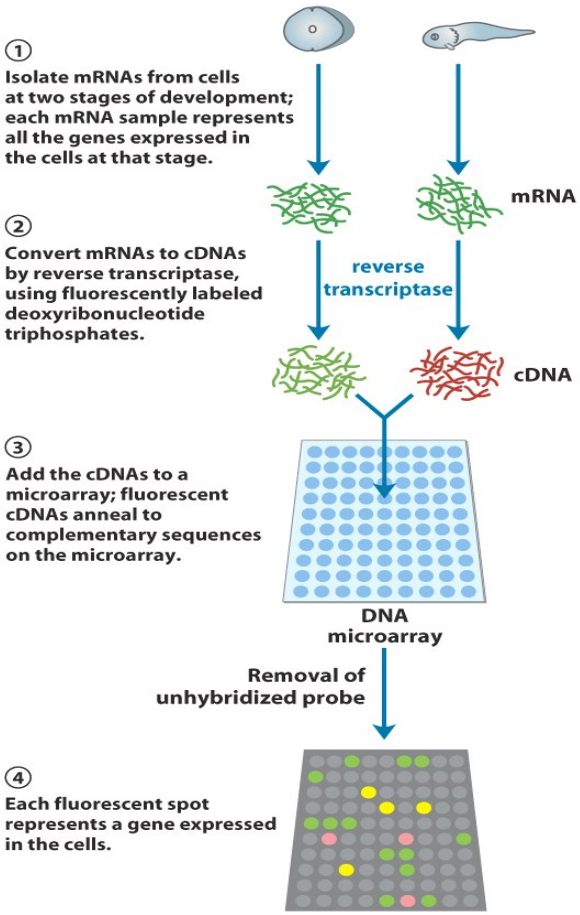


Transcriptomics involves large-scale analysis of RNAs to follow when, where, and under what conditions genes are expressed.

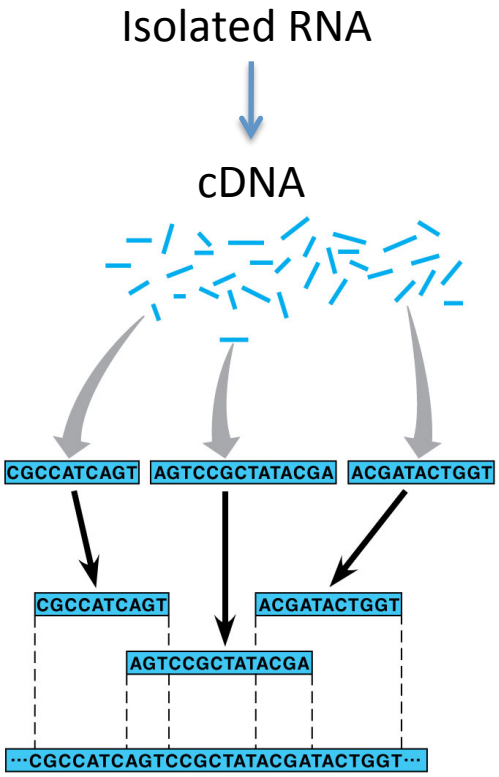
Expression profiling

Two High-throughput techniques

DNA microarray technology

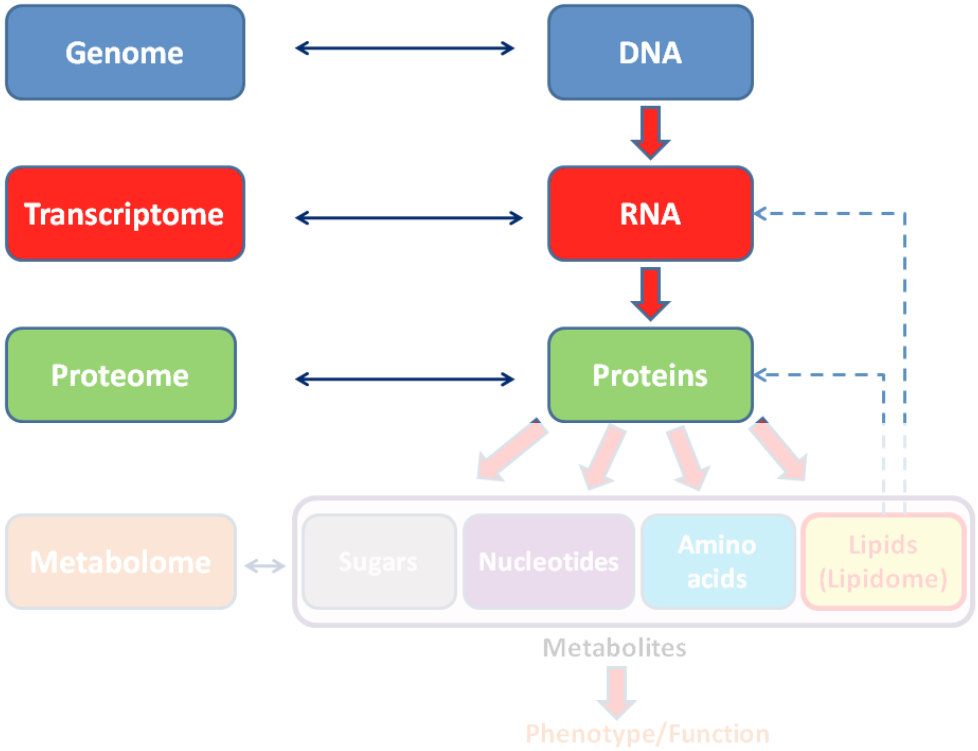


NGS technology (RNA-seq)



Proteomics

Proteome = complete set of proteins for a given organism **OR** a complete set of protein produced under a given set of conditions



Proteomics is :

- closer than gene expression studies to what's actually happening in the cell
- the study of the structure and function of proteins

Prote**omics**

- Understanding gene function and its changing role in development and aging
- Identifying proteins that are biomarkers for diseases; used to develop diagnostic tests
- Finding proteins for development of drugs to treat diseases and genetic disorders

- 2D-electrophoresis and mass spectrometry
- High-throughput, but less than transcriptomics

Prote**omic** limits

Proteome varies as it reflects **genes that are actively expressed at any given time**

- Advantages
  - Detect proteins not RNA (post transcriptional regulation)
- Limitations
  - Only the most highly expressed proteins are detected
  - Overlapping spots may be difficult to resolve
  - Not likely to be useful in metagenomics

Transcript**omics** *vs.* Prote**omics**

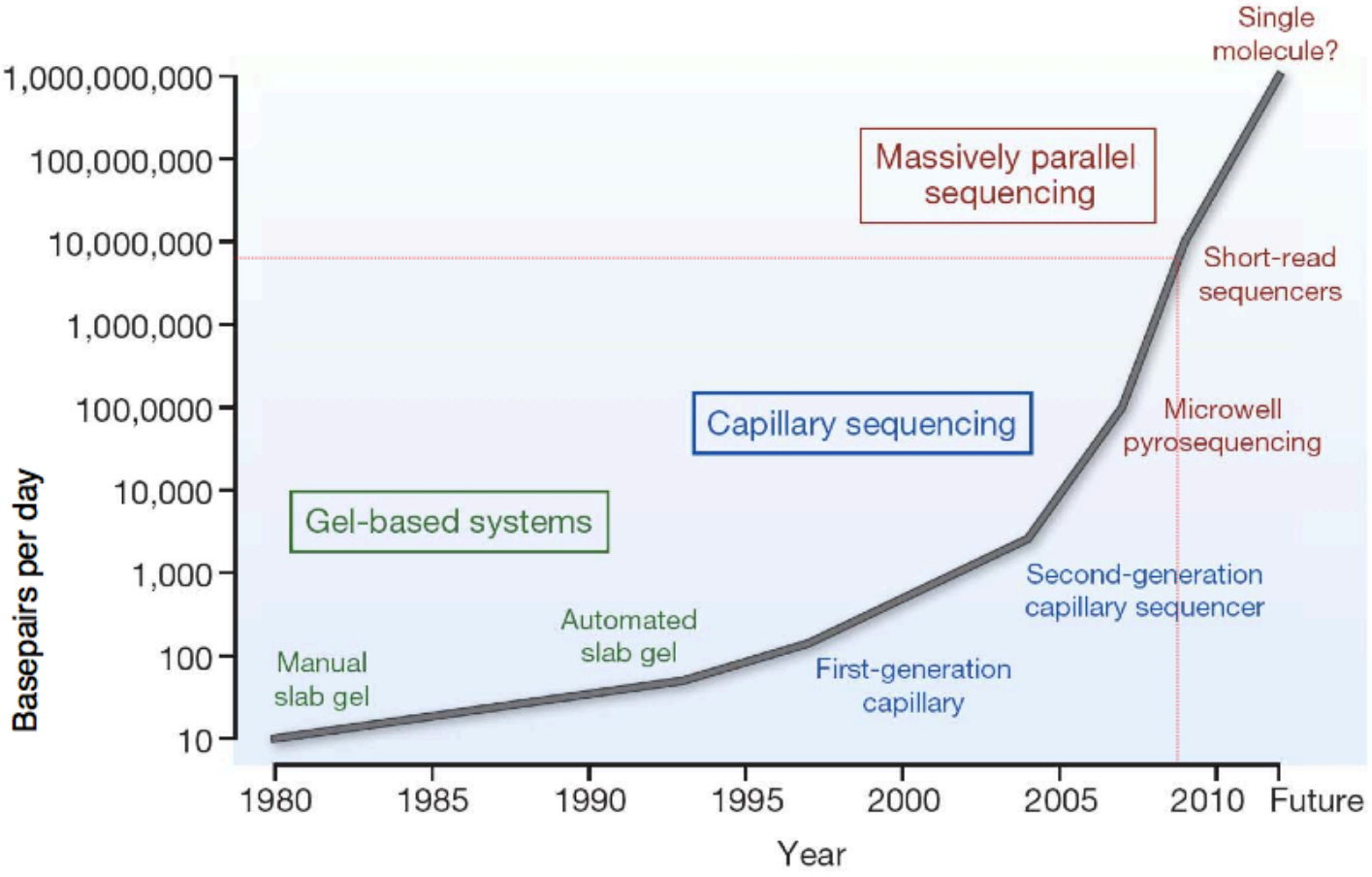
- Transcriptomics is robust, relatively cost-effective and user-friendly

- Proteomics still relatively limited – problems can remain with purification and stability of proteins

Sequencing technologies

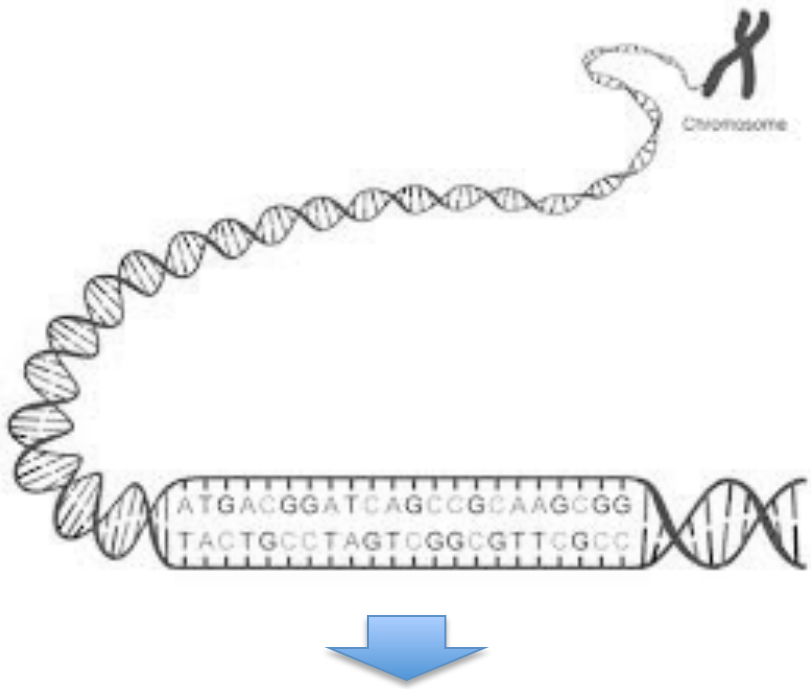
Omics

coincide with dramatic improvements in different sequencing technologies



adapted fron M. Stratton

DNA sequence



DNA Sequence

Sequencing or how to read a sequence

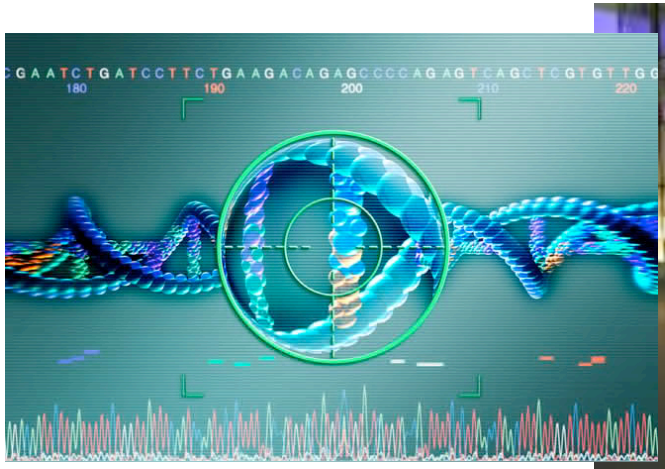
DNA Sequence

Set of strings based on
4 letters as DNA alphabet {A,C,G,T}

Sequencing or how to read a sequence

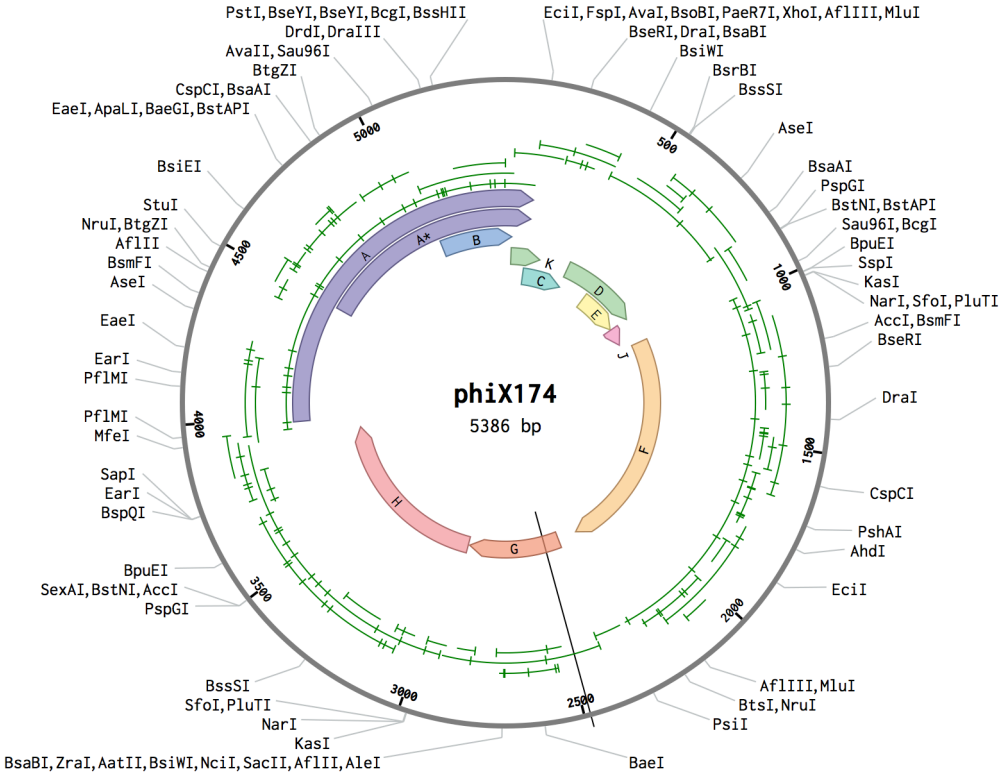
DNA Sequence

Set of strings based on
4 letters as DNA alphabet {A,C,G,T}
Source of information



First genome sequenced

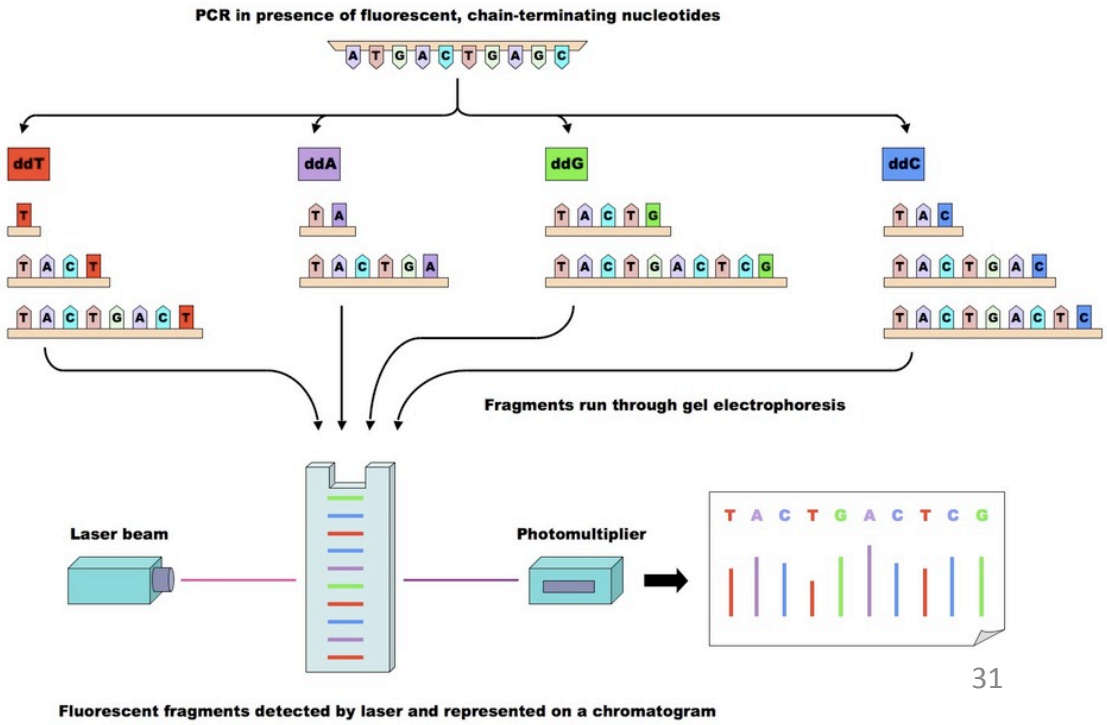
First genome sequenced by **Sanger sequencing** (enzyme synthesis) in 1977 bacteriophage X174 single strand of 5,375 bp



Sanger method

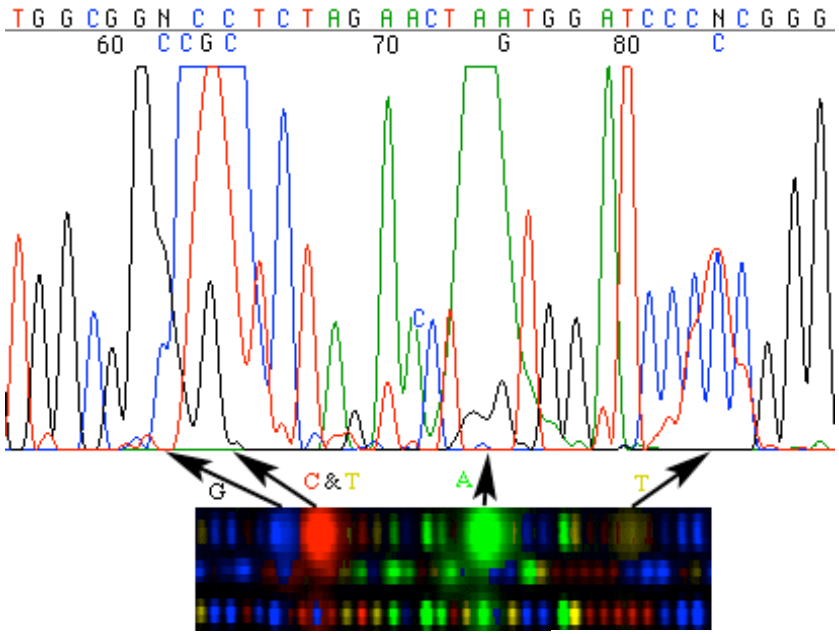
The Sanger method involves four PCR reactions. Each reaction contains the four normal nucleotides plus one dideoxynucleotide stock. As a typical PCR reaction generates over 1 billion DNA molecules, each of the four PCR reactions will generate all of the possible terminating fragments for that particular base.

Dideoxynucleotides are fluorescently labelled and so, when the four PCR samples are run through gel electrophoresis, the sequence of the fragments can be detected by a laser and represented via a chromatogram



Fluorescent fragments detected by laser and represented on a chromatogram

chromatogram



Chromatogram Viewers

4 Peaks
[Mac]

BioEdit
[Windows]

Chromas
[Windows]

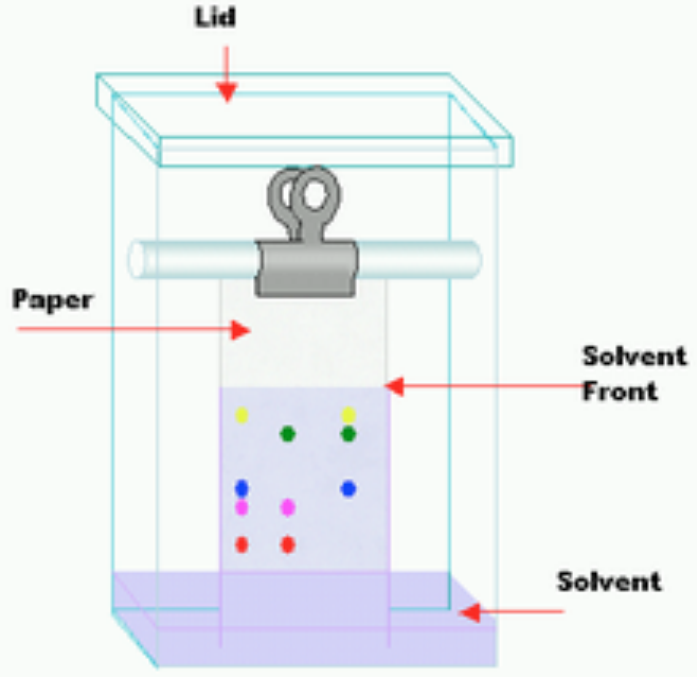
Trace Viewer
[Windows / Mac]

Finch TV
[Windows / Mac]

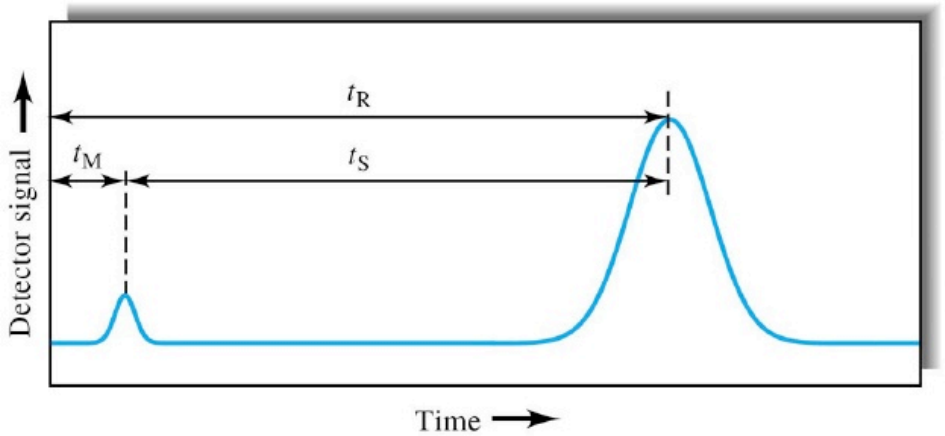
Sequence Scanner
[Windows]

How to read a chromatogram?

Retention = distance traveled by the compound/distance traveled by the solvent front

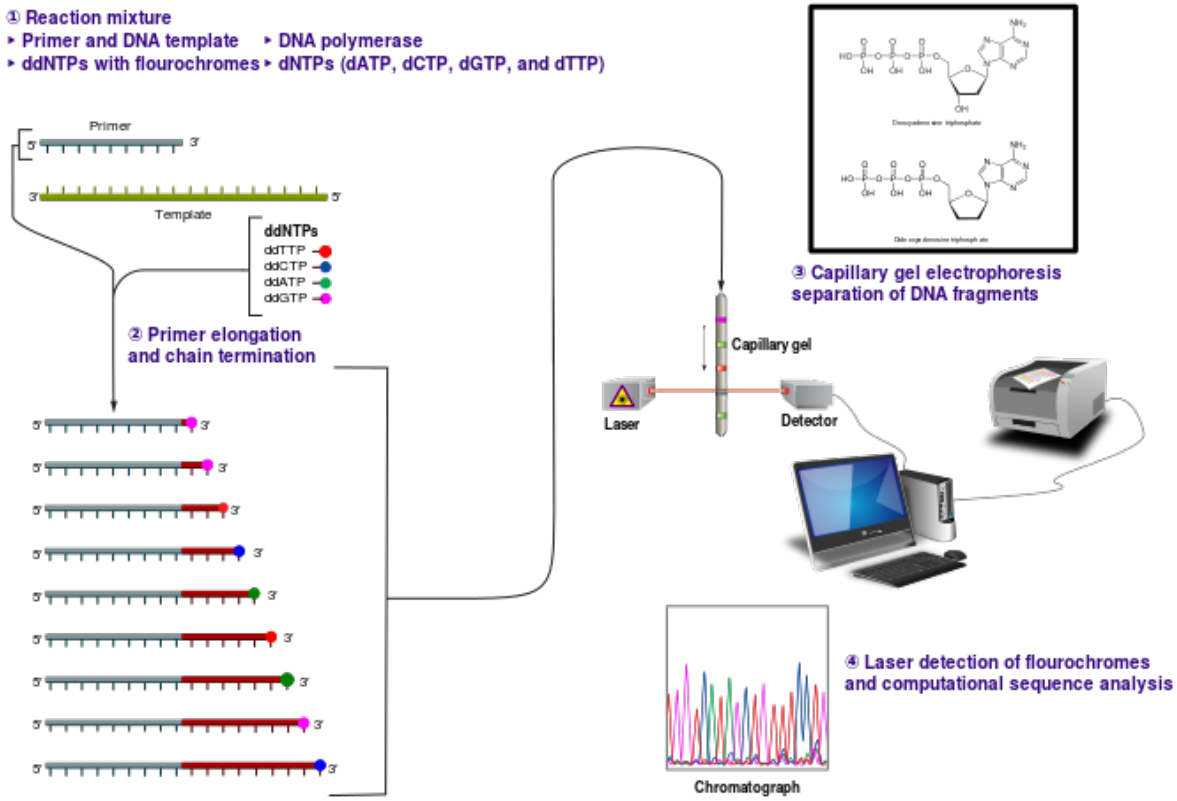


1.) Typical response obtained by chromatography (i.e., a chromatogram):



Where:

t_R = retention time
t_M = void time
W_b = baseline width of the peak in time units
W_h = half-height width of the peak in time units



Limitations of Sanger Sequencing

Low throughput

Inconsistent base quality

Expensive

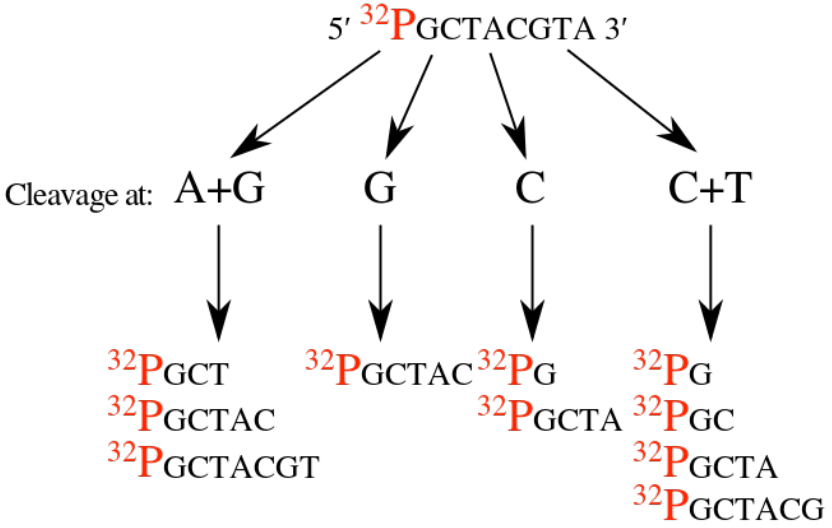
Not quantitative

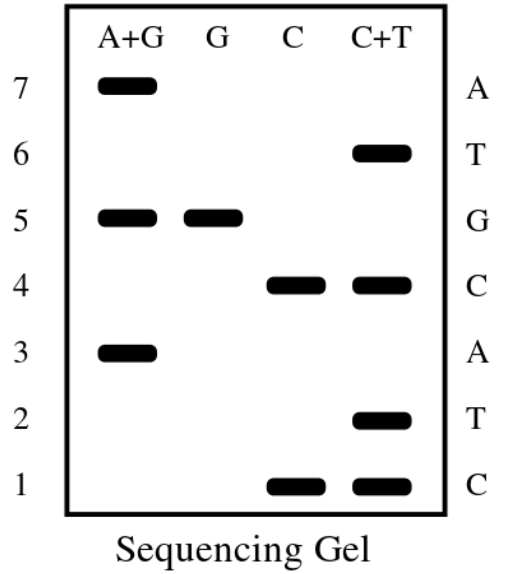
Maxam-Gilbert sequencing

Chemical modification of DNA
Radioactive labelling in 5’ end

DNA cleavage induced by the chemical treatment at a small proportion of four reactions (G, A+G, C, C+T).

The fragments in the four reactions are then electrophoresed side by side in denaturing acrylamide gels for size separation.





Sequencing Gel

no longer in widespread use, having been supplanted by next-generation sequencing.

Genome sequencing

Two genome sequencing strategies:

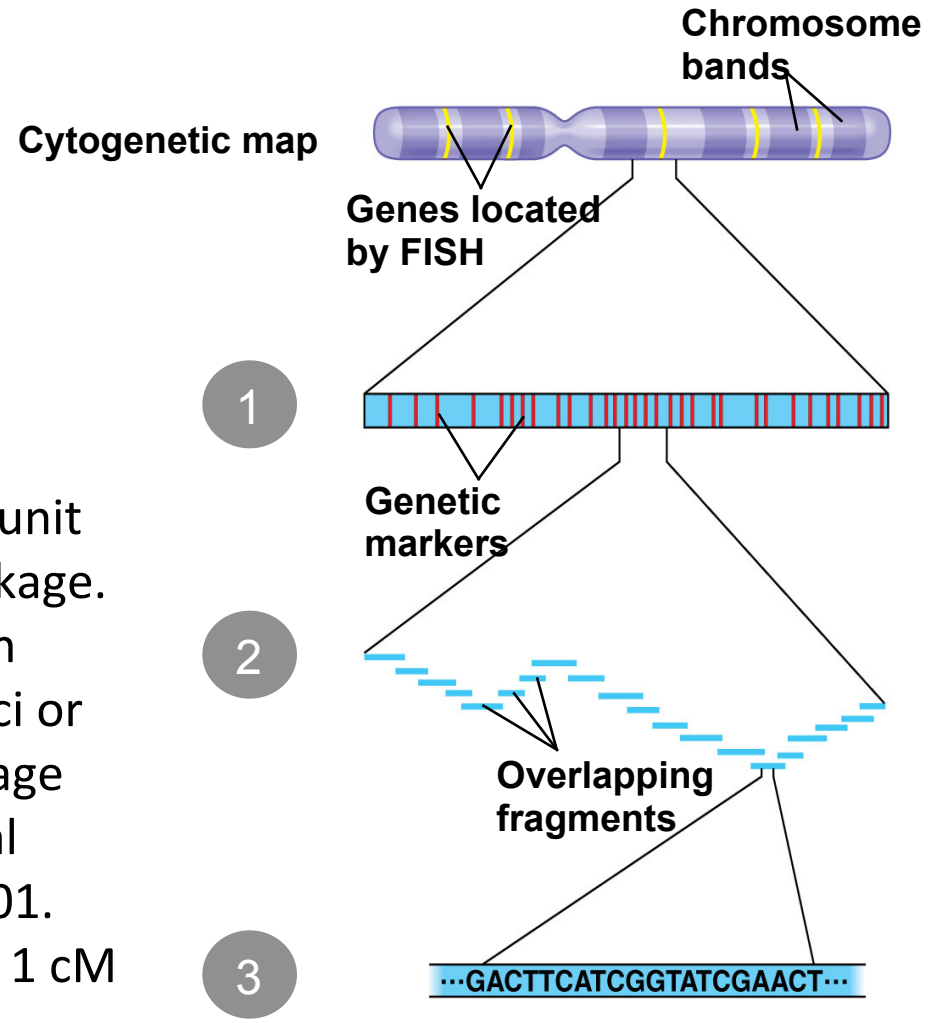
- **Clone-by-clone method (aka hierarchical shotgun or BAC by BAC sequencing)** (government's genome project)
- **Whole Genome Shotgun method** (privately-funded Celera genome project)

Clone-by-Clone (CBC)

Three-Stage Approach to Genome Sequencing

1. Genetic mapping (cM)

centimorgan (abbreviated cM) or map unit (m.u.) is a unit for measuring genetic linkage. It is defined as the distance between chromosome positions (also termed loci or markers) for which the expected average number of intervening chromosomal crossovers in a single generation is 0.01. (in human: 1cM =± 1 mégabase in plants 1 cM =± 200 kilobases)

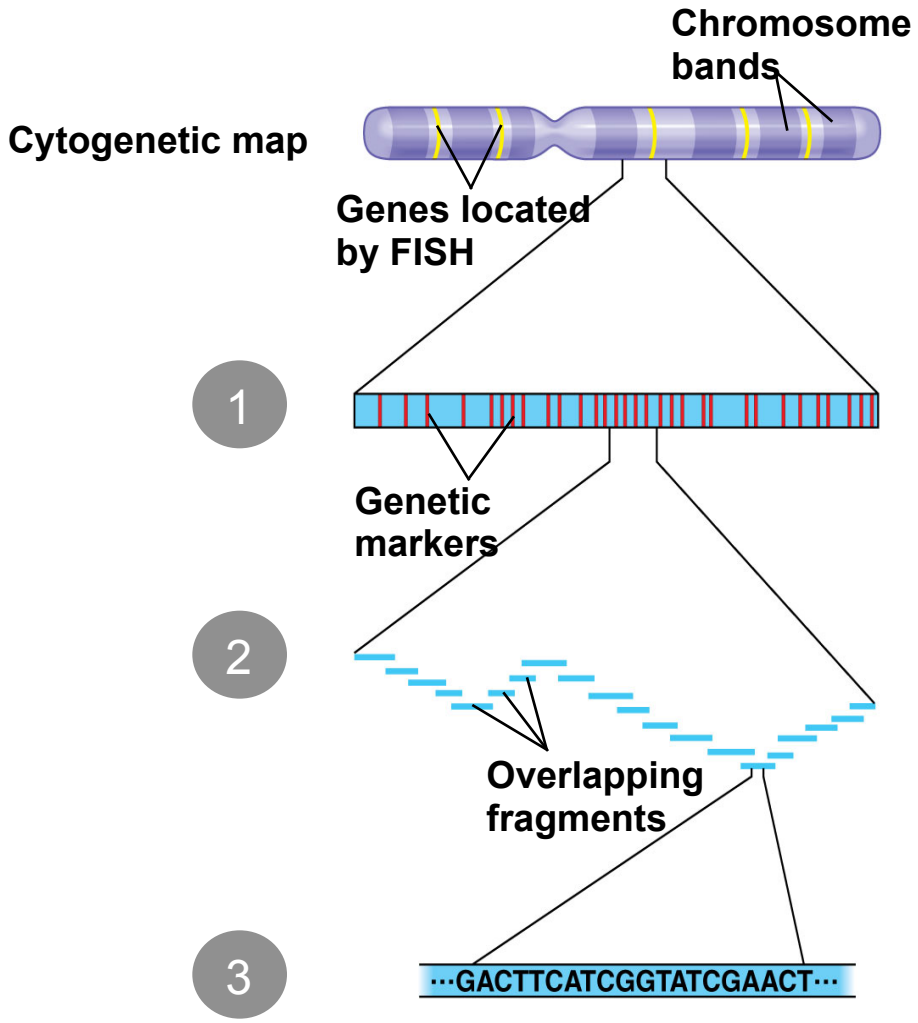


Clone-by-Clone (CBC)

Three-Stage Approach to Genome Sequencing

1. Genetic mapping (cM)
2. Physical maps (bp)
3. DNA sequencing of **ordered** clones

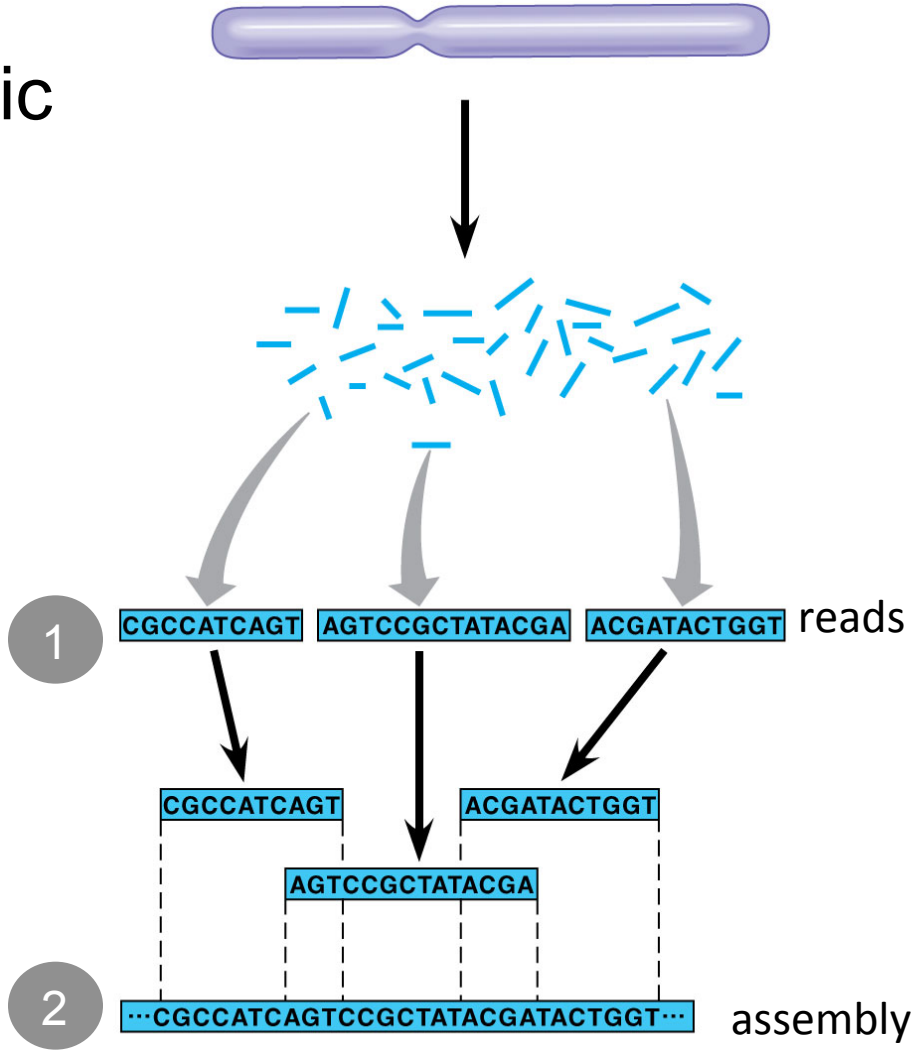
The clones have been arranged to cover an entire chromosome



Whole-Genome Shotgun (WGS)

This approach skips genetic and physical mapping and sequences random DNA fragments directly

1. DNA sequencing of **random** clones
2. Assembly (order fragments into a continuous sequence)

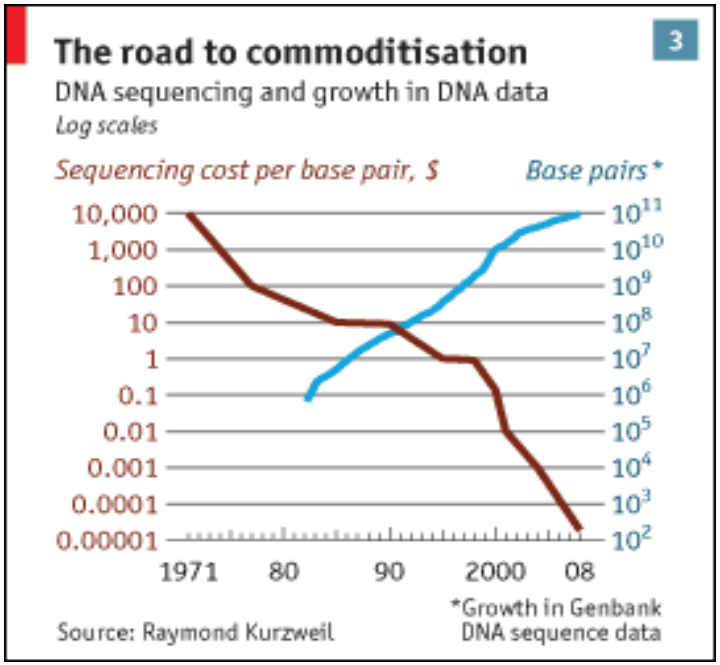


CBC *vs.* WGS

- CBC is time-consuming, expensive and does not allow resolving repeats

- WGS is now widely used as the sequencing method of choice.

The development of new WGS sequencing technologies has resulted in **massive increases in speed** and **decreases in cost**.



Genomes sequenced

1st whole genome
Haemophilus
influenzae
~ 1.8Mb
1995

1st whole pluricellular
genome
Caenorhabditis elegans
~ 97Mb
1998

1st Human genome
launched in 1990
~ 3Gb
2001-2004

1977-1990
Birth of the
Computer
sciences

1st whole eucaryote
genome
Saccharomyces cerevisae
~ 120Mb
1996

1st whole eucaryote
genome
Drosophila melanogaster
~ 120Mb
2000

Human Genome Project…expensive

1988 - 2004

soit 16 ans et 3 milliards de $

1 dollar par base

Objectif : décoder le génome humain pour accélérer les progrès en génétique, de la médecine à l'évolution de l'humain.

Definitions (1) (in french)

Séquençage haut débit (SHD) : terme générique et peu spécifique (utilisation à éviter).

Séquençage nouvelle génération (NGS) ou massif en parallèle : regroupe les technologies de 2nde et 3ème génération.

Séquençage de 2nde génération : séquençage d'un ensemble de molécules nucléotidiques à l'aide de techniques de “wash-and-scan” (ou cycles).

“Wash-and-scan” : technique basée sur des polymérases et réactifs qui doivent être enlevés à chaque cycle après l'incorporation des bases à lire.

Definitions (2) (in french)

Séquençage de 3ème génération : processus de séquençage de molécules uniques ne nécessitant pas de “wash-and-scan”.

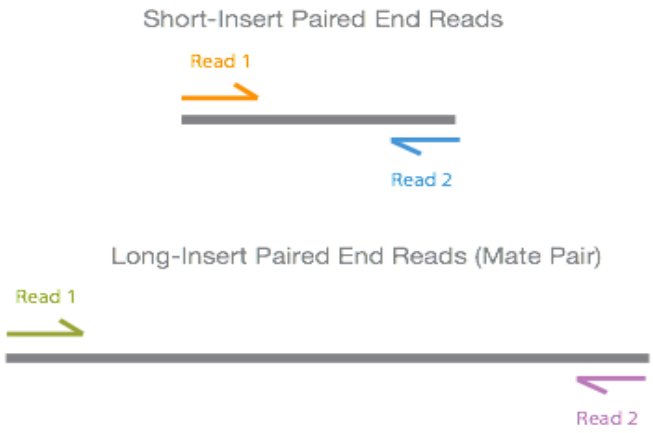
Lecture : fragment nucléotidique individuel dont la séquence est déterminée par un instrument.

Longueur de lecture : correspond au nombre de bases individuelles composant une lecture donnée.

Préparation de librairies : procédure expérimentale précédant le séquençage des fragments d'ADN d'intérêt. Varie en fonction de la technologie.

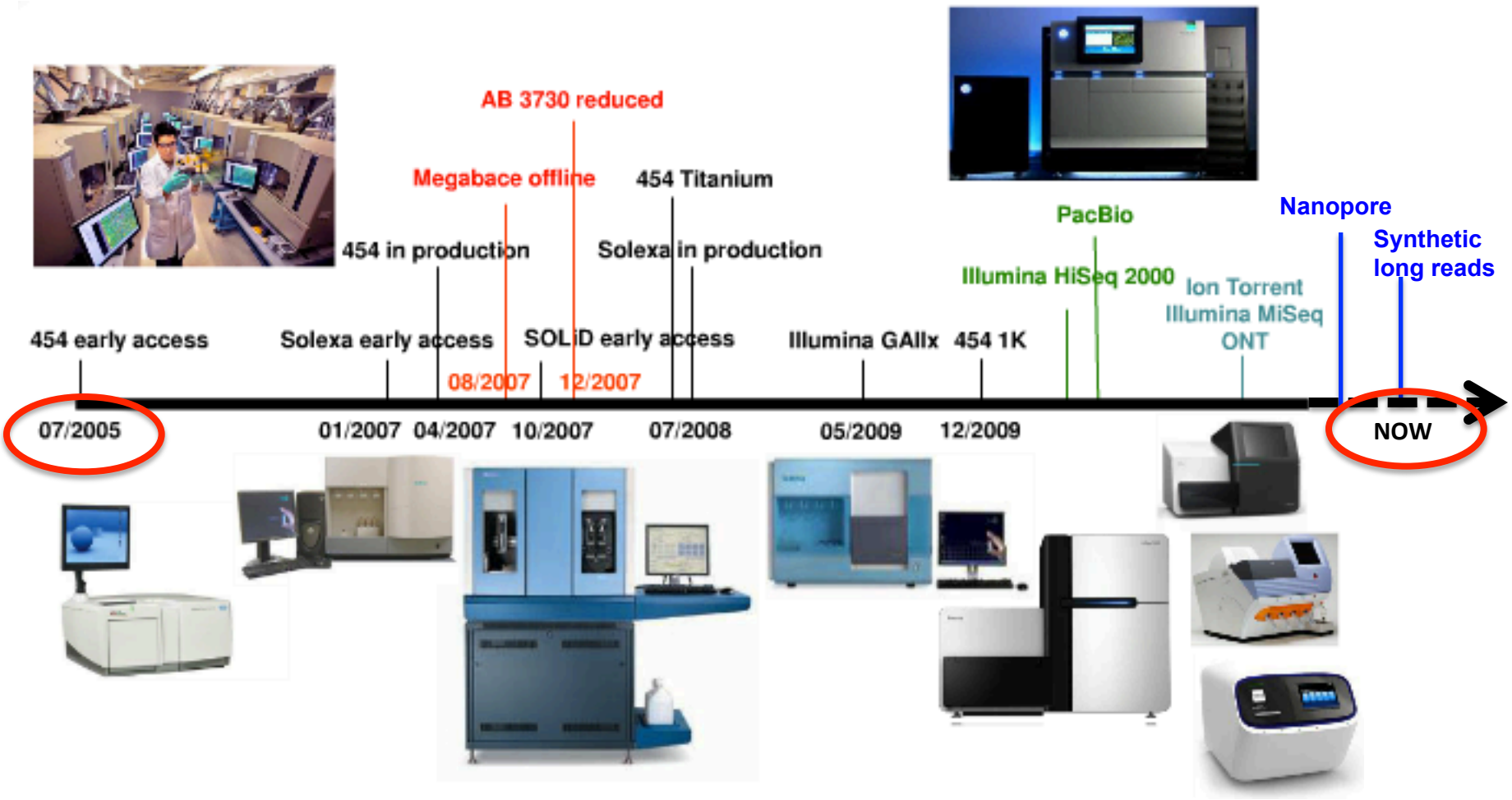
Definitions (3) (in french)

Paire de lecture: couple de deux lectures correspondant aux deux extrémités du fragment à séquencer. En fonction du protocole expérimental utilisé pour la préparation de la librairie, la taille et orientation des lectures varient.



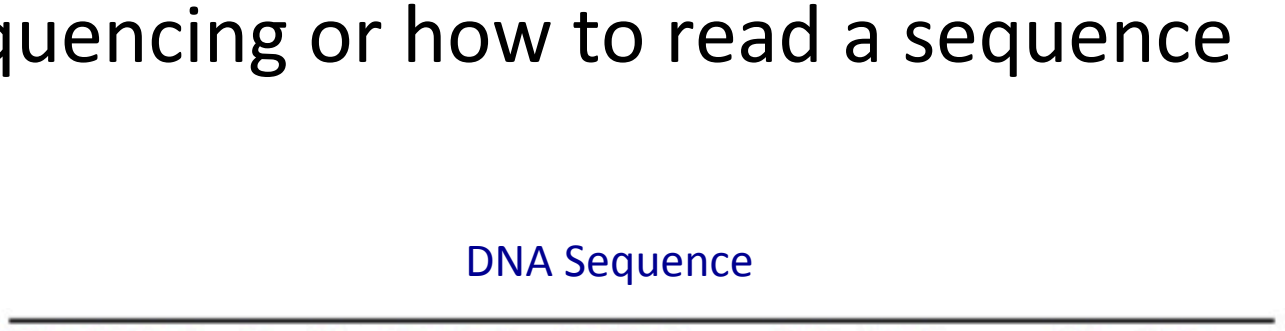
Taille d'insert (insert size): Distance entre deux lectures d'une paire en incluant leur longueur. Différent de la « outer size »

Next-generation sequencing (NGS) technologies



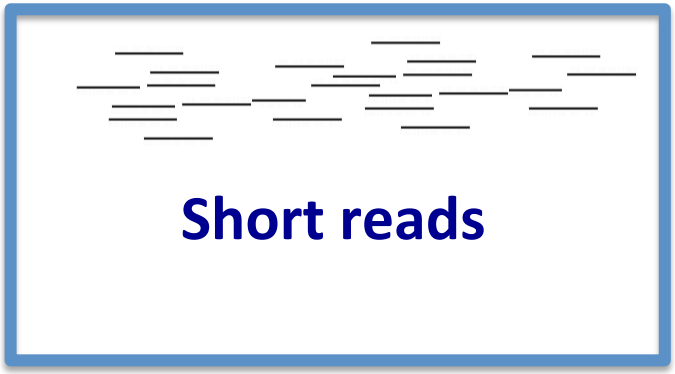
Sequencing or how to read a sequence

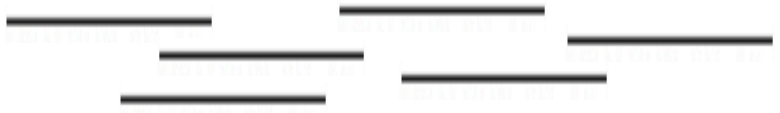
DNA Sequence



Set of strings based on
4 letters as DNA alphabet {A,C,G,T}

Different technologies for different data



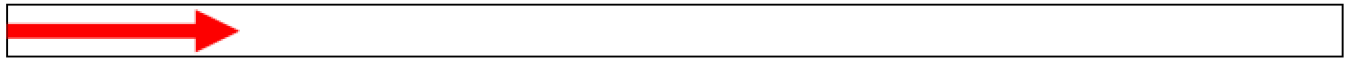


Long reads

Different read types

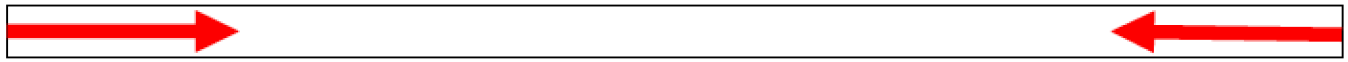
Single end read

– Only have sequence from one end of fragment

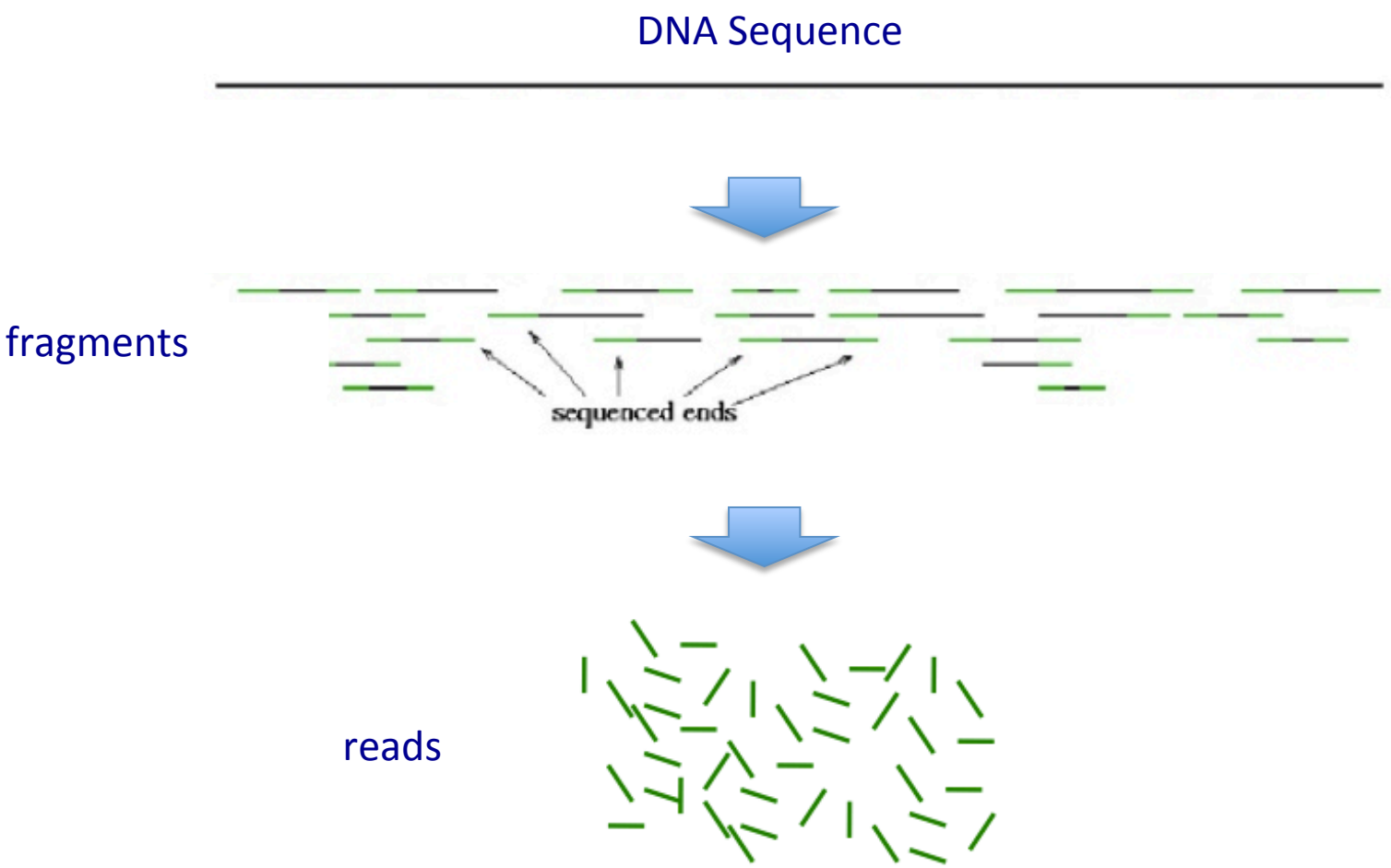


Paired / Mated read

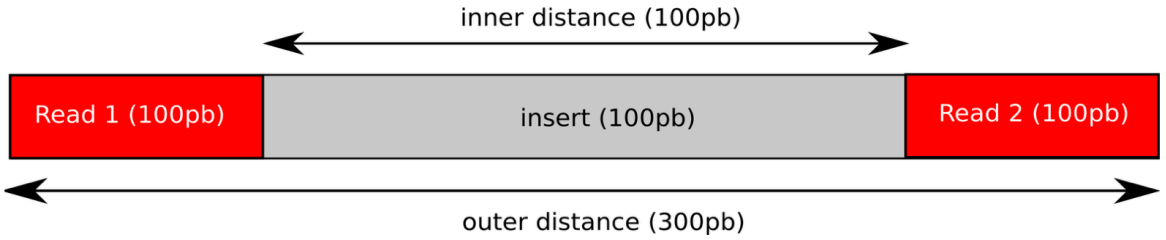
– Have sequence from both ends of fragment



Paired-end sequencing



How to define a pair of reads?



= Insert size

NGS technologies

Short reads

Genome Analyzer IIx (GAIIx), HiSeq2000, HiSeq2500, MiSeq – **Illumina**

SOLiD 5500xl System – Applied Biosystem

HeliScope™ Single Molecule Sequencer **-** Helicos

Long reads

Synthetic long reads – Illumina

Genome Sequencer FLX System (454) – Roche

PacBio RS - Pacific Bioscience

Personal Genome Machine, Ion Proton - Ion Torrent

GridION – Oxford Nanopore

2nd NGS technologies

1) Fragmentation and tagging of genomic/cDNA fragments – provides universal primer allowing complex genomes to be amplified with common PCR primers

2) Template immobilization – DNA separated into single strands and captured onto beads (1 DNA molecule/ bead)

3) Clonal Amplification – Solid Phase Amplification

4) Sequencing and Imaging – Cyclic reversible termination (CRT) reaction

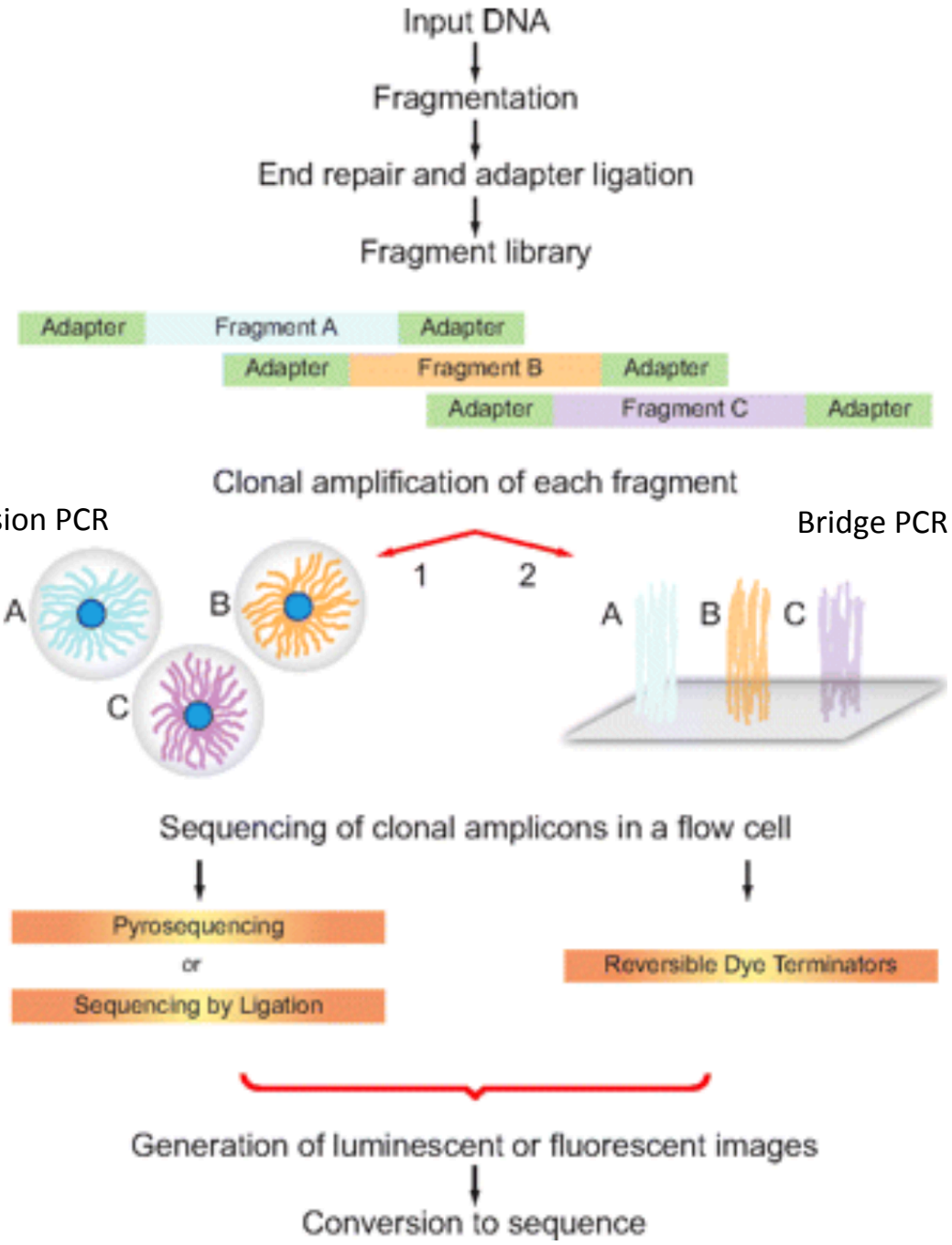
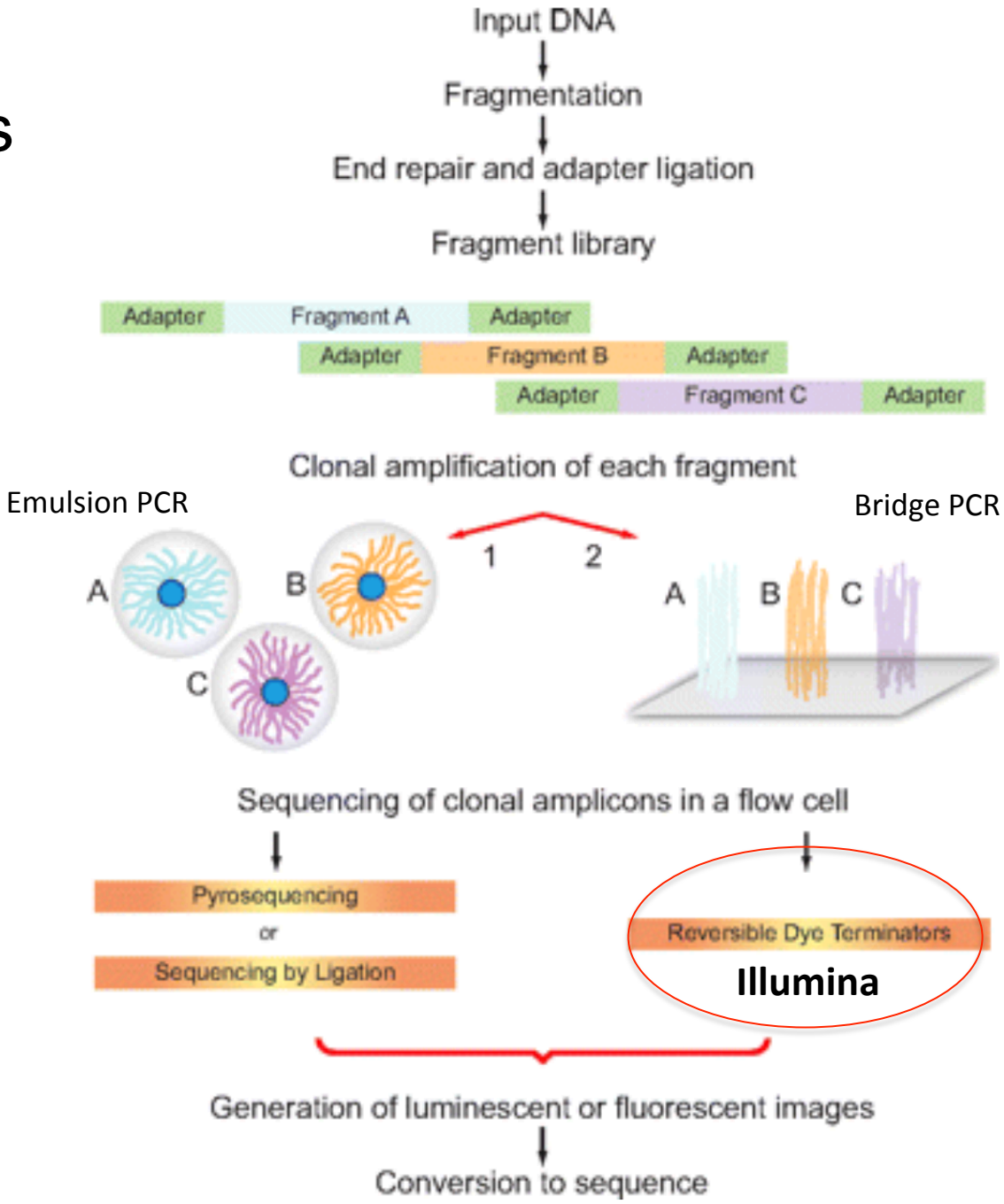


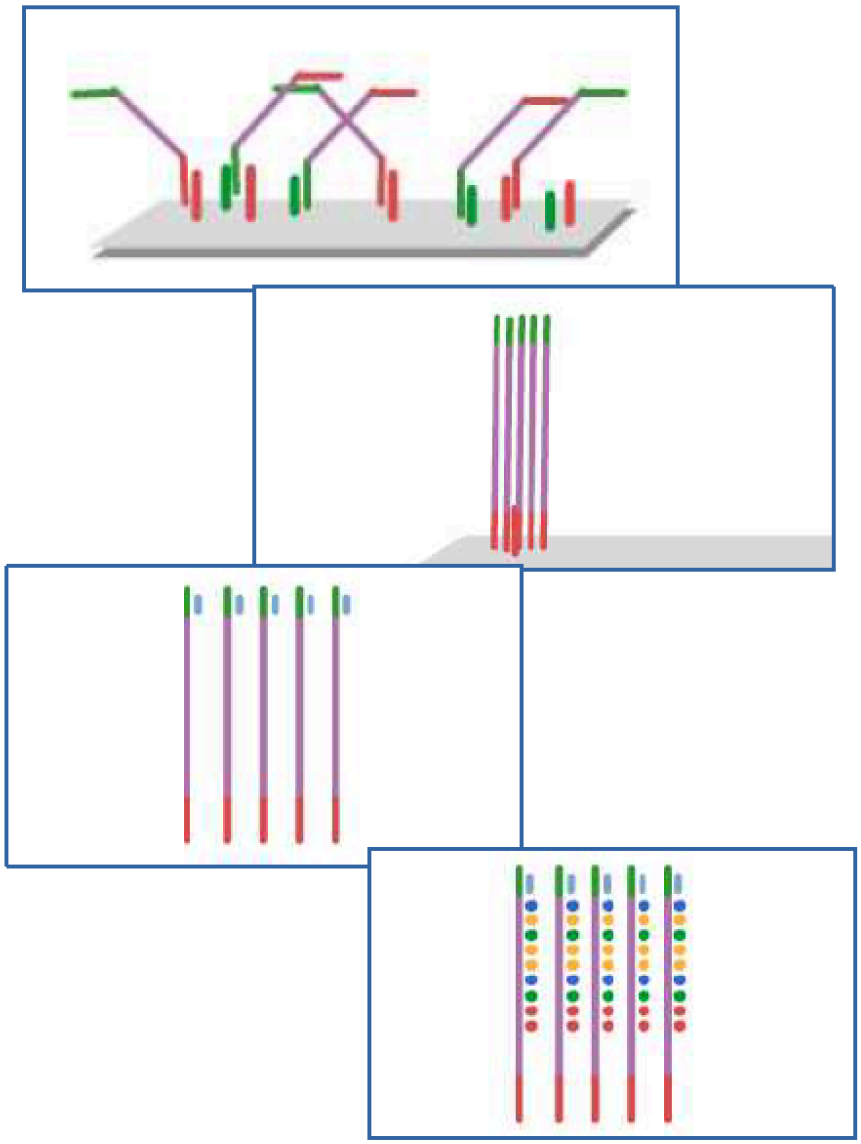
Table 1. 2^{nd} and 3^{rd} Generation DNA sequencing platforms listed in the order of commercial availability

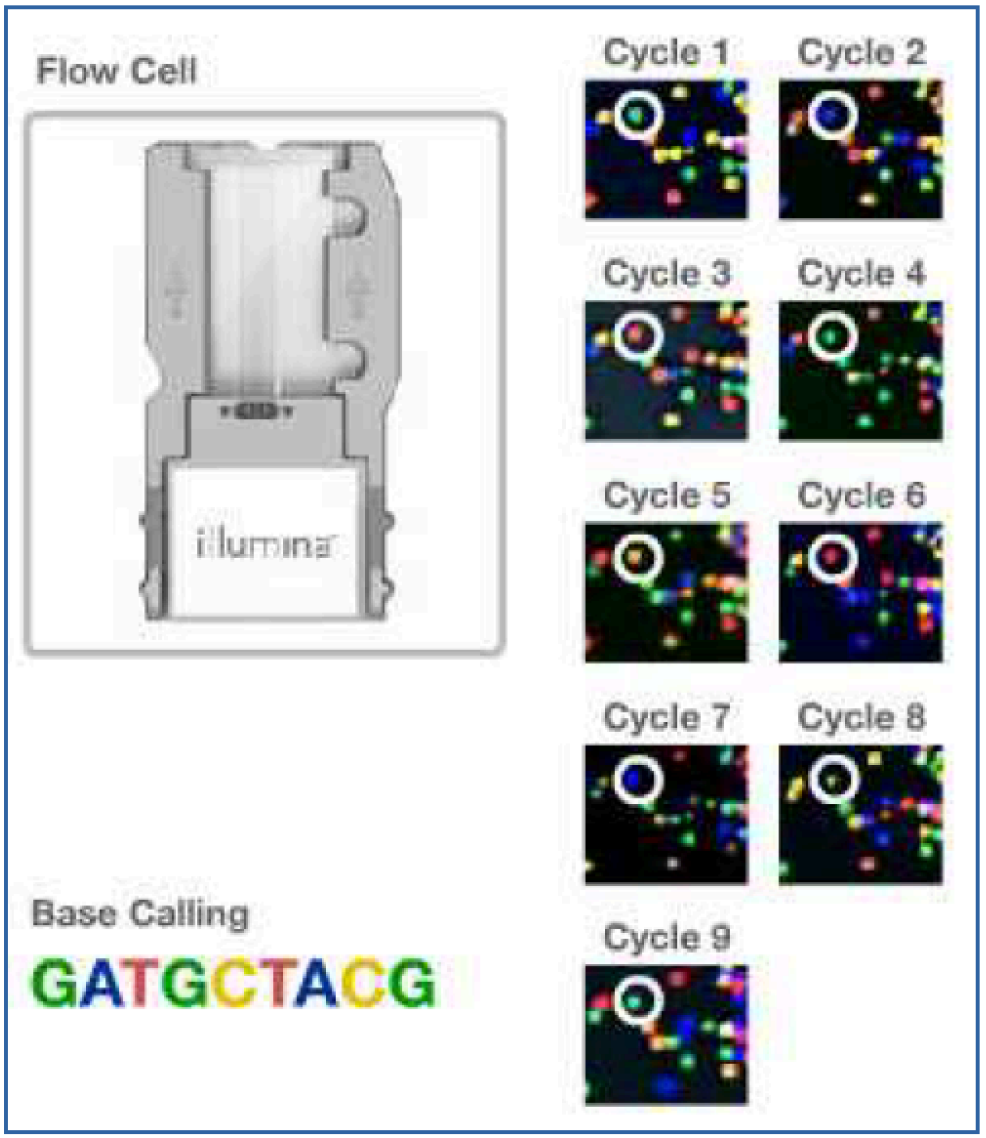
Platform	Current company	Former company	Sequencing method	Amplification method	Claim to fame
454	Roche	454	Synthesis (pyrosequencing)	emPCR	First Next-Gen Sequencer, Long reads
Illumina	Illumina	Solexa	Synthesis	BridgePCR	First short-read sequencer; current leader in advantages†
SOLiD	Life Technologies	Applied Biosystems	Ligation	emPCR	Second short-read sequencer; low error rates
HeliScope	Helicos	N/A	Synthesis	None	First single-molecule sequencer
Ion Torrent	Life Technologies	Ion Torrent	Synthesis (H^+ detection)	emPCR	First Post-light sequencer; first system <$100 000
PacBio	Pacific Biosciences	N/A	Synthesis	None	First real-time single-molecule sequencing
Starlight‡	Life Technologies	N/A	Synthesis	None	Single-molecule sequencing with quantum dots

NGS process


Input DNA
Fragmentation
End repair and adapter ligation
Fragment library
Adapter
Fragment A
Adapter
Adapter
Fragment B
Adapter
Adapter
Fragment C
Adapter
Clonal amplification of each fragment
Emulsion PCR
Bridge PCR
A
B
C
1
2
A
B
C
Sequencing of clonal amplicons in a flow cell
Pyrosequencing
or
Sequencing by Ligation
Reversible Dye Terminators
Illumina
Generation of luminescent or fluorescent images
Conversion to sequence

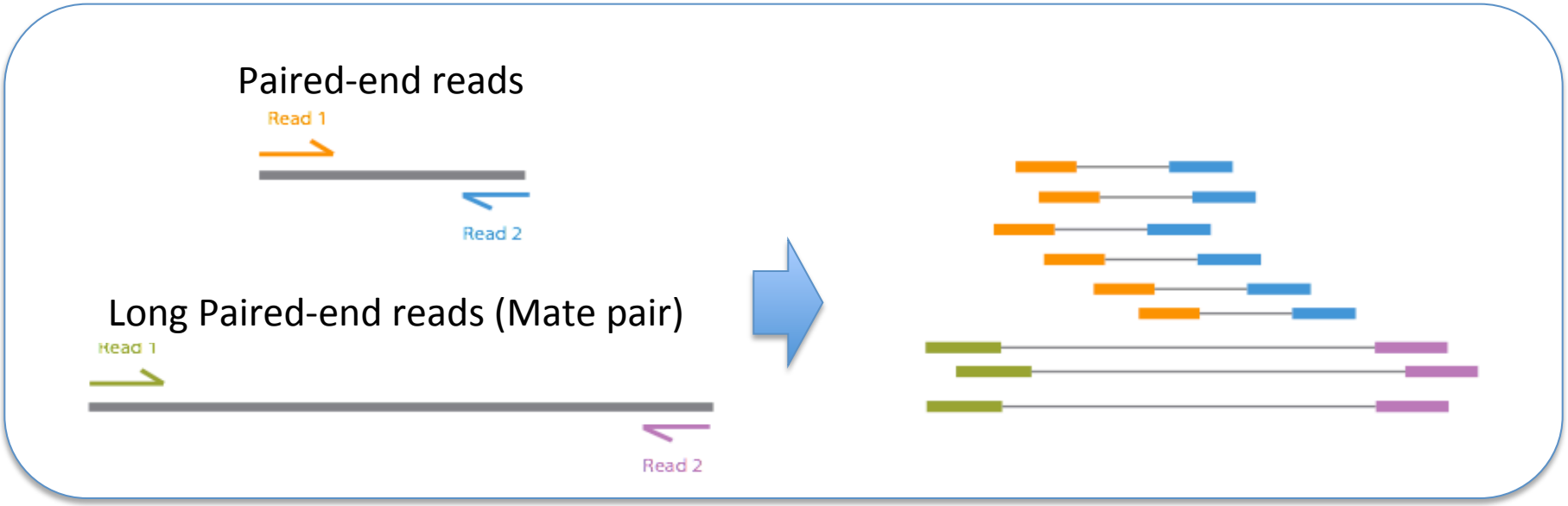
Reversible terminator sequencing - Illumina





	HiSeq 2000/2500	HiScan SQ	Genome Analyzer IIx	MiSeq
Lectures	2x100 pb	2x100 pb	2x150 pb	2x250 pb
Débit	600 Gb	140 Gb	96 Gb	7,5 Gb
Lectures/run	3 milliards	700 millions	320 millions	15 millions
Précision	99,9%	99,9%	99,9%	99,9%
Temps d'exécution	11 jours	8 jours	14 jours	39h

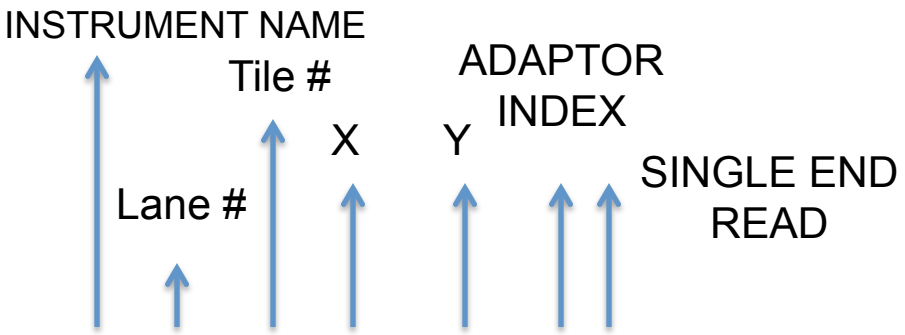
Short reads



Sequencing data =
FASTQ file

Converting RAW data to FASTQ

FASTQ File

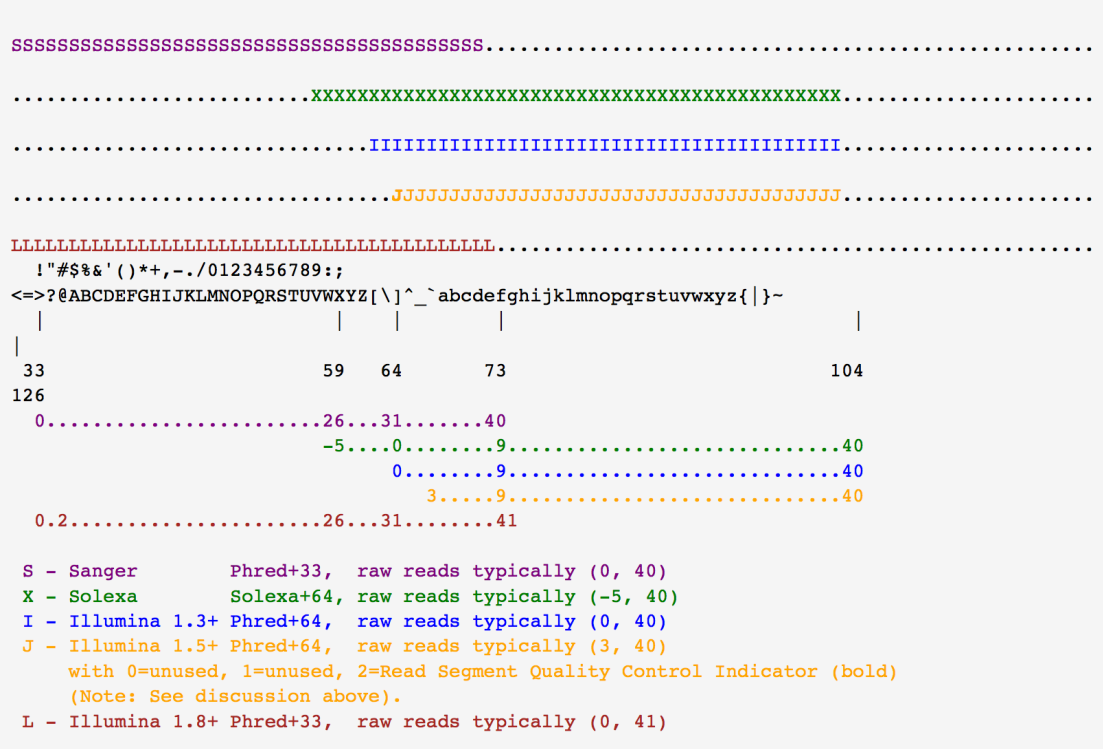


FASTQ – FASTA "with an attitude" (embedded quality scores). Originally developed at the Sanger to couple (Phred) quality data with sequence, it is now common to specify raw read output data from NGS machines in this format.

@SN971:3:2304:20.80:100.00#0/1
NAAATTTCACATTGCGTTGGGAACAGTTGGCCCAAACTCAGGTTGCAGTAACTGTCACAATACCATTCTCCATCAACTTC
AAGAAATGTTCAACAAAACAC
+
@P\cceeeggggiihhiiiiiiihighiiiiiiiiiiiiifghhhhgfghiifihihfhhiiiihiggggggeeeeeeddcdddccbcdddccccccc

Line 1: begins with '@' followed by sequence identifier
Line 2: raw sequence
Line 3: +
Line 4: base quality values for sequence in Line 2

FASTQ file



Differing in the format of the sequence identifier and in the valid range of quality scores. See:

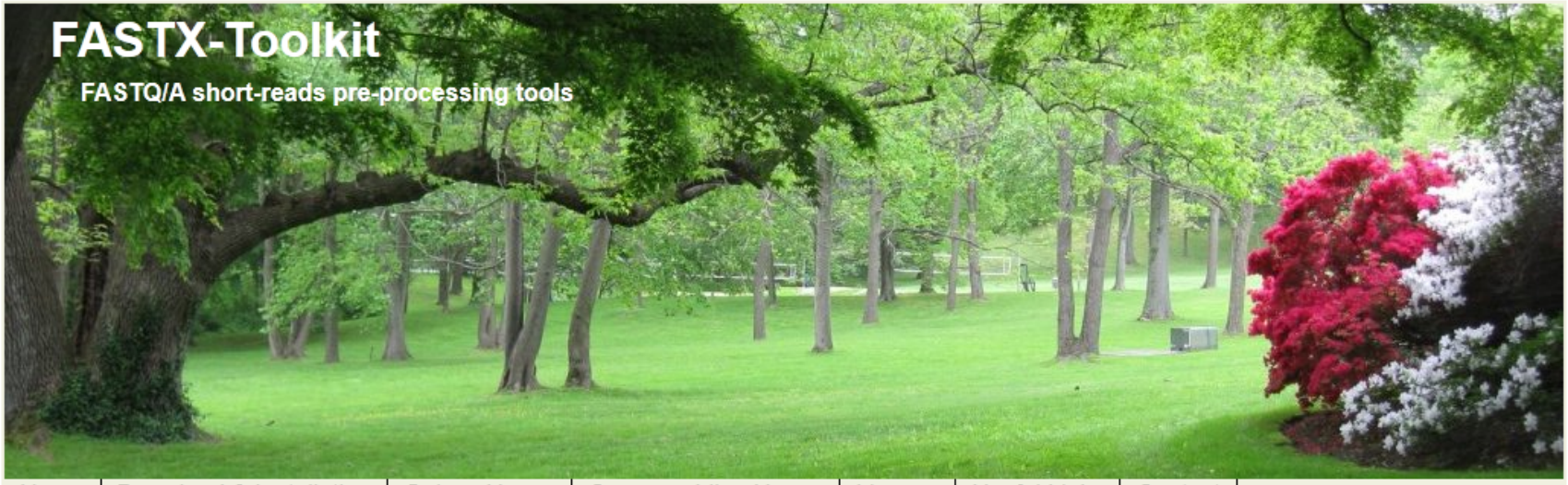
http://en.wikipedia.org/wiki/FASTQ_format

http://maq.sourceforge.net/fastq.shtml

http://nar.oxfordjournals.org/content/early/2009/12/16/nar.gkp1137.full

Pre-process: FASTQ analysis

http://hannonlab.cshl.edu/fastx_toolkit/



Home | Download & Installation | Galaxy Usage | Command-line Usage | License | Useful Links | Contact

Introduction

The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.

Next-Generation sequencing machines usually produce FASTA or FASTQ files, containing multiple short-reads sequences (possibly with quality information).

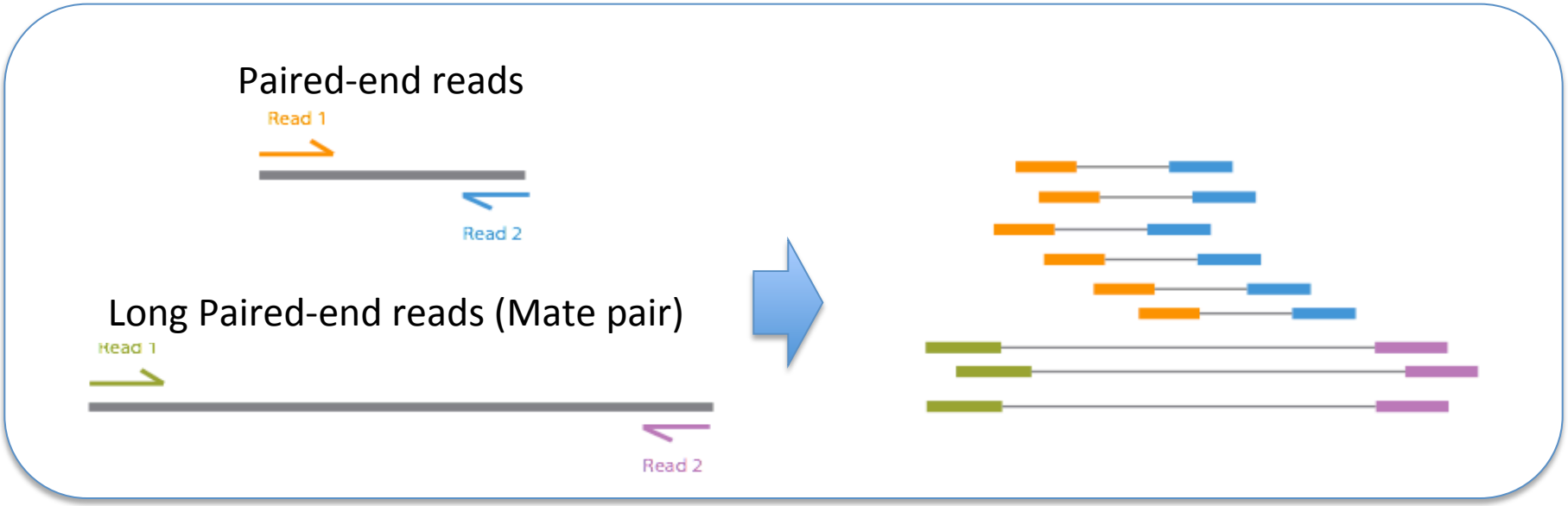
The main processing of such FASTA/FASTQ files is mapping (aka aligning) the sequences to reference genomes or other databases using specialized programs. Example of such mapping programs are: Blat, SHRiMP, LastZ, MAQ and many many others.

However,
It is sometimes more productive to preprocess the FASTA/FASTQ files before mapping the sequences to the genome - manipulating the sequences to produce better mapping results.

The FASTX-Toolkit tools perform some of these preprocessing tasks.

Linux, MacOSX or Unix only

Mapping *vs.* Assembly

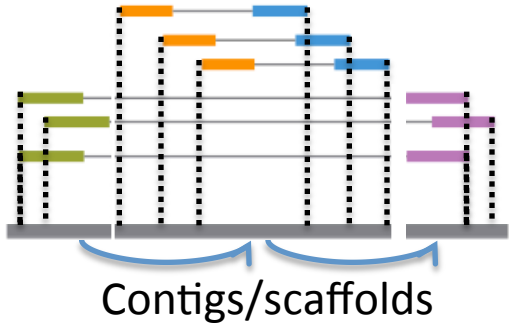


Reference genome

Mapping

No Reference genome

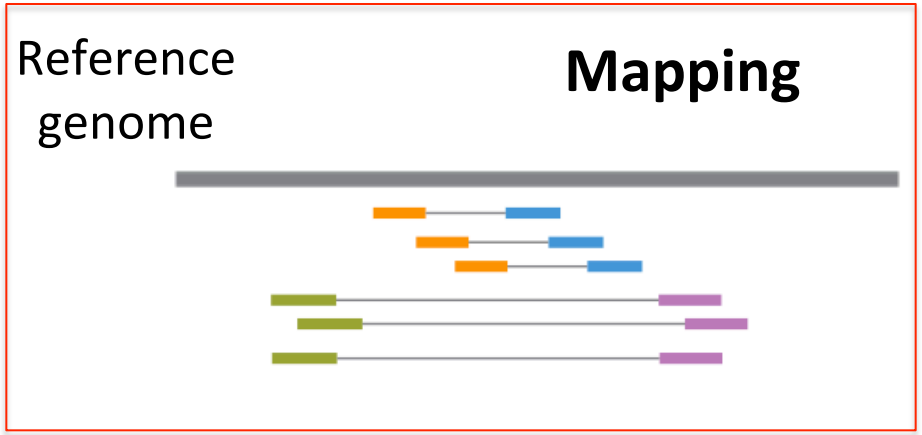
***De novo* Assembly**

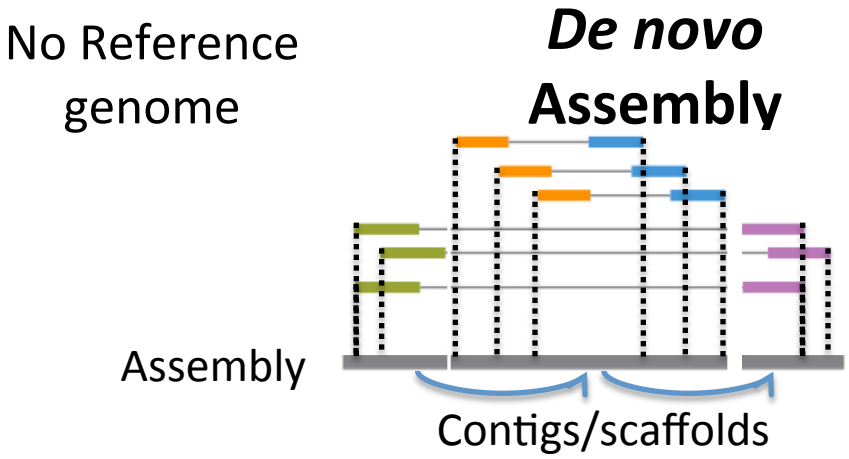


Mapping *vs.* Assembly

Mapping (re-sequencing):

- Will miss genome rearrangements
- Only as good as the reference





NGS technologies

Short reads

Genome Analyzer IIx (GAIIx), HiSeq2000, HiSeq2500, MiSeq – **Illumina**

SOLiD 5500xl System – Applied Biosystem

HeliScope™ Single Molecule Sequencer **-** Helicos

Long reads

Synthetic long reads – Illumina

Genome Sequencer FLX System (454) – Roche

PacBio RS - Pacific Bioscience

Personal Genome Machine, Ion Proton - Ion Torrent

GridION – Oxford Nanopore

PacBio sequencing

Single molecule resolution in real time

- Short waiting time for result and simple workflow
  - Generate basecalls in <1 day
  - Polymerase speed ≥1 base per second
- No amplification required
  - Bias not introduced
  - More uniform coverage
- Direct observation
  - Distinguish heterogeneous samples
  - Simultaneous kinetic measurements
- Long reads
  - Identify repeats and structural variants
  - Less coverage required
- Information content
  - One assay, multiple applications
    - Genetic variation (SVs to SNPs)
    - Methylation
    - Enzymology

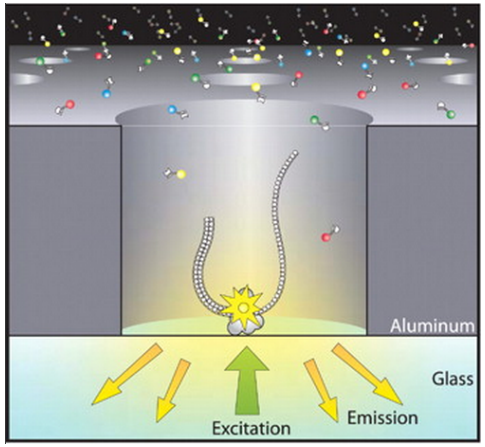
PacBio introduction

C2 chemistry – installed March 2012
-Long reads 6-10kb
-Meidan size of molecules 3kb
-Still 15% error rate
-No strobe sequencing

Software focus on:
De novo assembly
Hi quality CCS consensus reads

In preparation

-Load long molecules by magnetic beads
-Modified nucleotides detection



LS – long sequencing reads

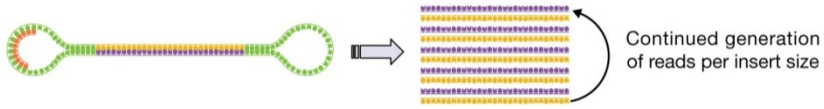
Sample Preparation

Standard

- Large insert sizes (2kb-10kb)
- Generates one pass on each molecule sequenced

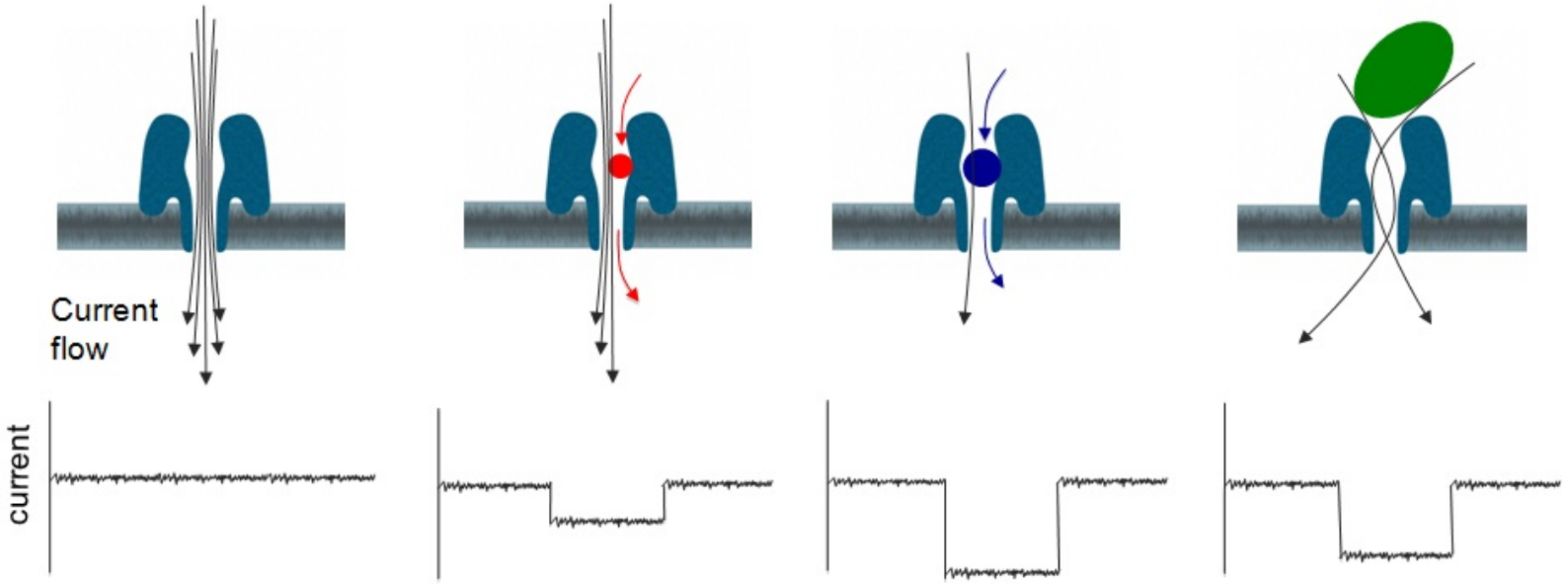
CCS – high quality sequencing reads

Circular Consensus



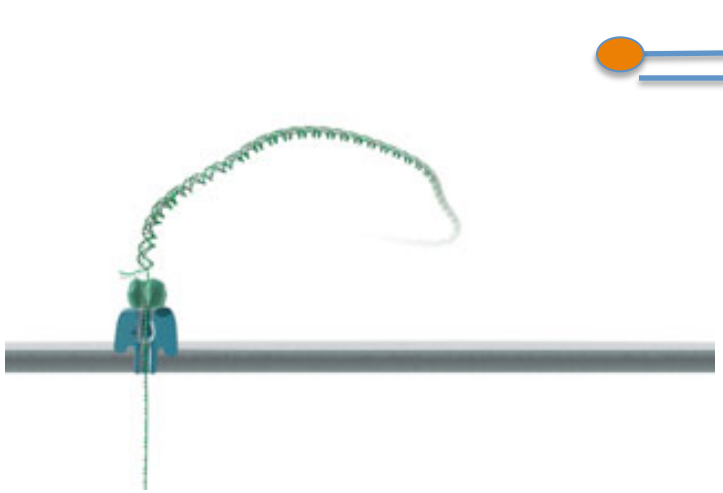
- Small insert sizes 500bp
- Generates multiple passes on each molecule sequenced

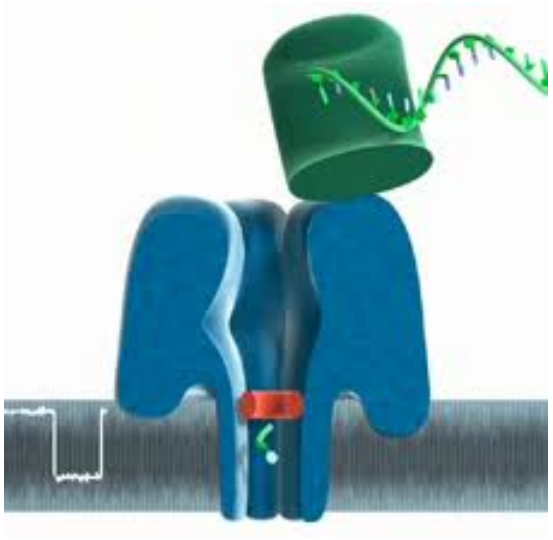
Oxford Nanopore – new view on sequencing

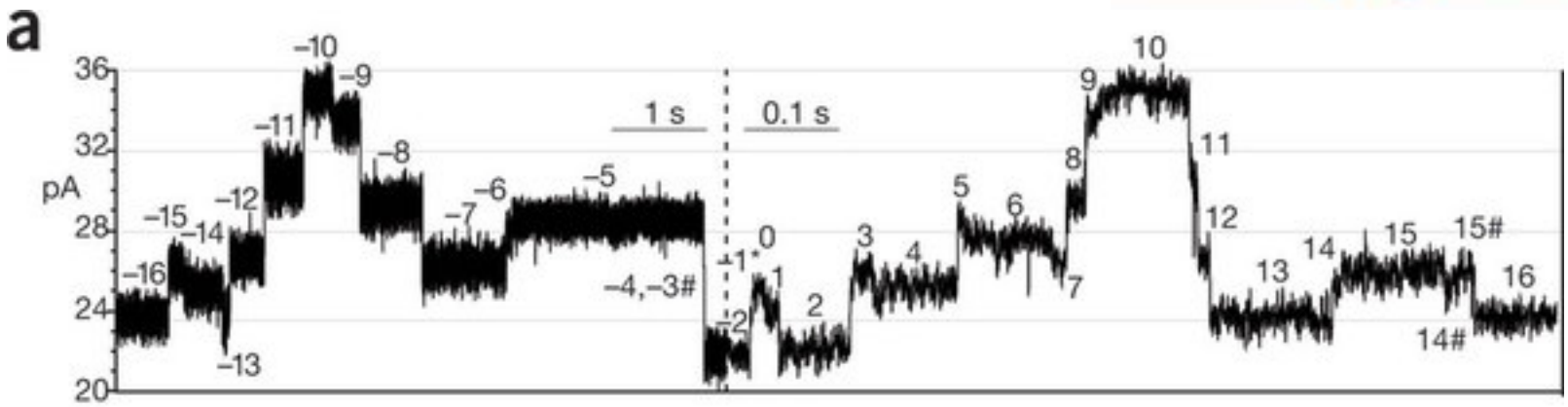


Hemolysin – pore - inner diameter of 1nm, about 100,000 times smaller than that of a human hair.

Oxford Nanopore







DNA sequencing

Error rate 4%, prediction for end of the year 0.1 – 2%.

Oxford Nanopore – new concepts

MinION

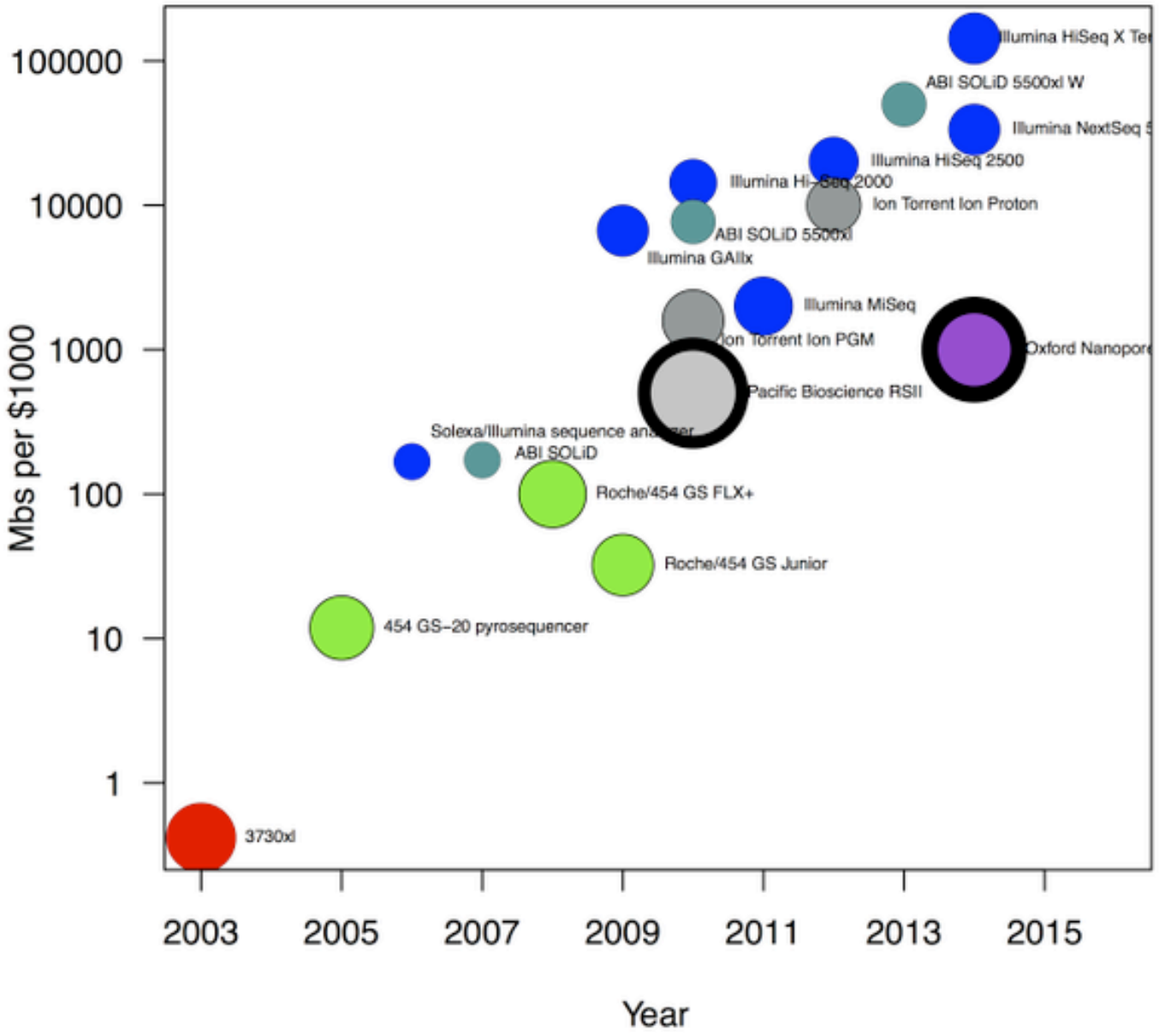
- 150Mb per run
- Tested 48kb read length
-$900 per instrument
-500 pores per device

GridION

- Tested 48kb read length
- 2000 pores per device, soon 8000 pores

-Cost per human genome $1500.

Cost versus error rate



From Lior Patcher lab (**@atifbd**)

NGS technology comparison

	Capacity	Speed	Read Length
454 Roche	35-700 Mb	10-23 hours	400-700 bp
SOLiD	90-180 Gb	7-12 days	75 bp
Illumina*	6-600 Gb	2-14 days	100-250 bp
Ion Torrent	20 Mb- 1Gb	4,5 hours	200 bp
Helicos	35 Gb	8 days	35 bp
PacBio*	1Gb	30 minutes	3000 bp

High DNA quality and quantity

Low sequencing error rate

Omics for what?

Study design

Optimize your experimental design.

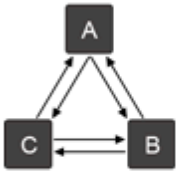


Image analysis

Analyse protein, RNA, DNA microarray images.

Mapping

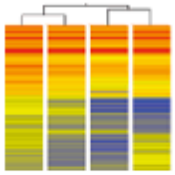
Align short reads to reference sequences.

***De novo* assembly**

Compute *de novo* assemblies.

Expression profiling

Discover potential biomarker panels by extracting errors and confidence levels.

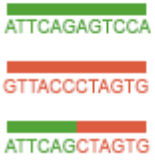


Pathway analysis

Get a better understanding of the system.

Gene fusions

Identify fusion transcripts from RNA-seq data.

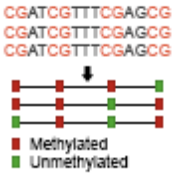


Disease-causing variants

Implement strategies for identifying rare variants underlying complex traits.

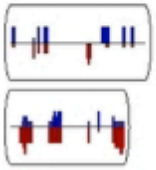
Methylation profiles

Analyse DNA methylation data.



Sequence variation

Discover sequence variation (SNPs, indels, CNVs).



OTU clustering

Group reads or contigs and assign them to operational taxonomic units.

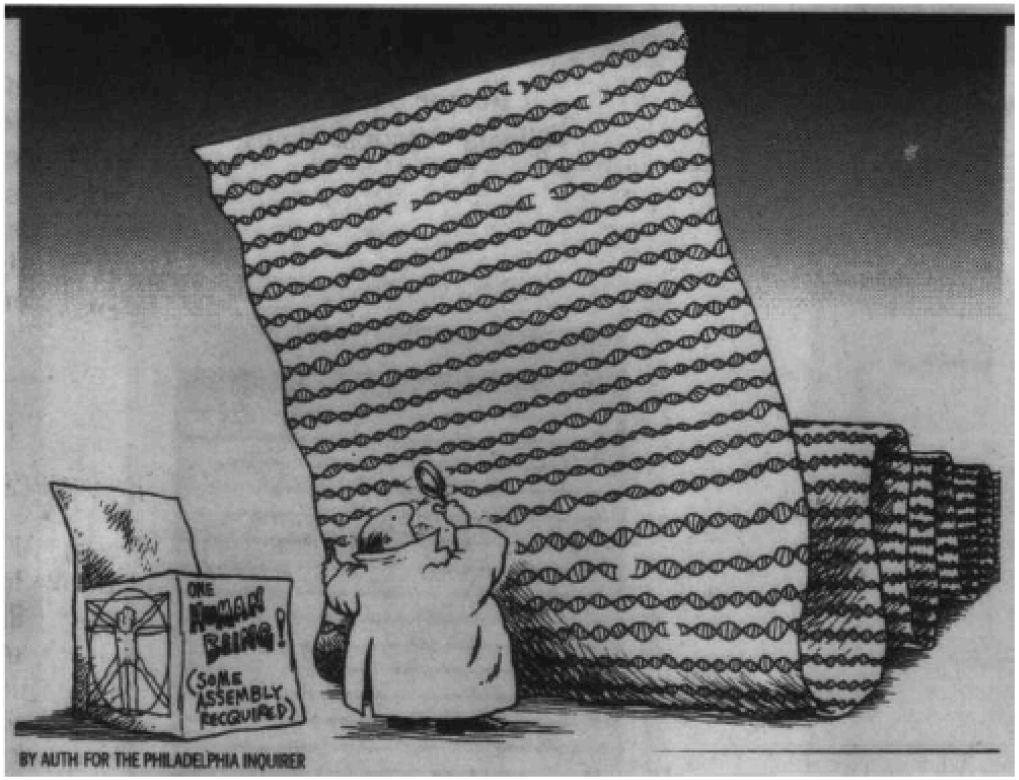
Enriched regions

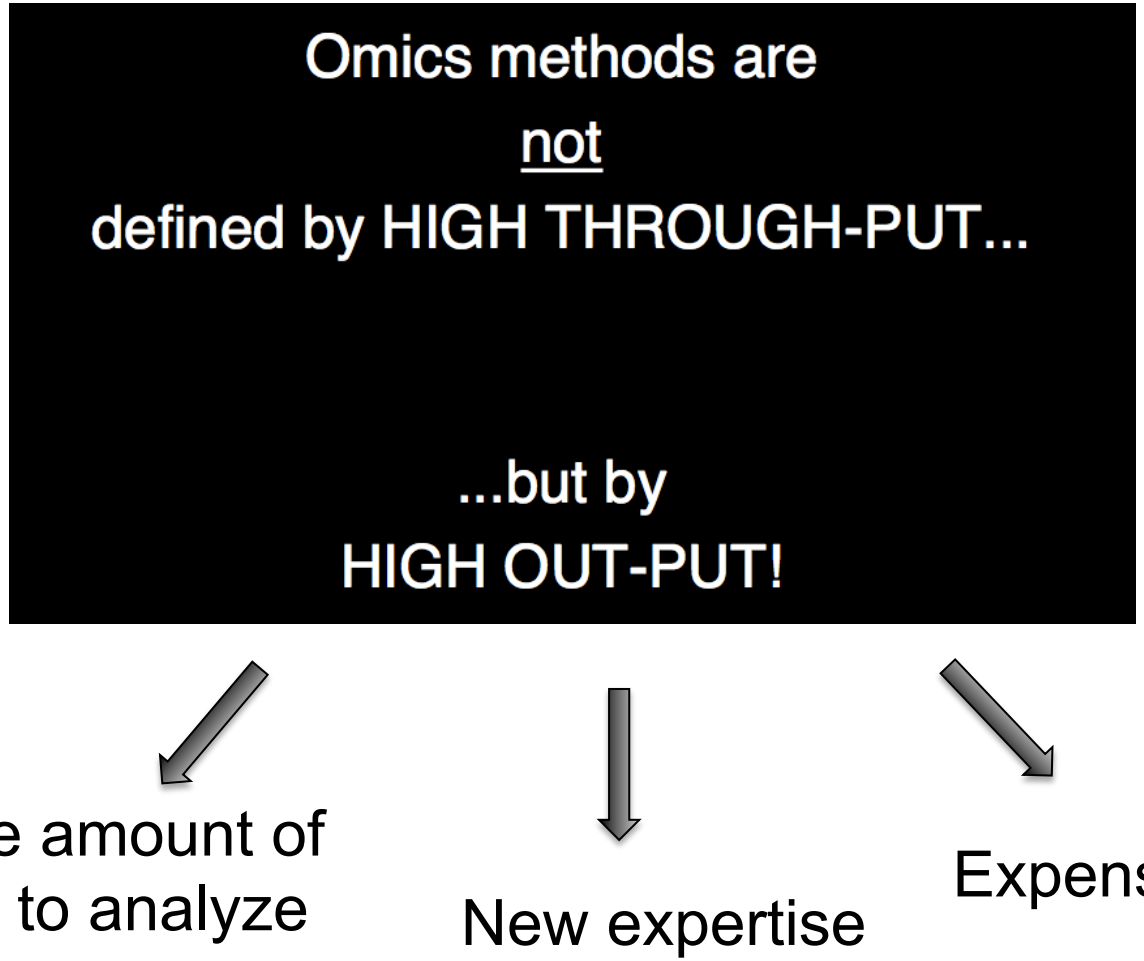
Identify regions that are enriched in ChIP samples.

Challenges

“There is a real disconnect between the ability to collect next-generation sequence data (easy) and the ability to analyze it meaningfully (hard)”

Dave O’Connor

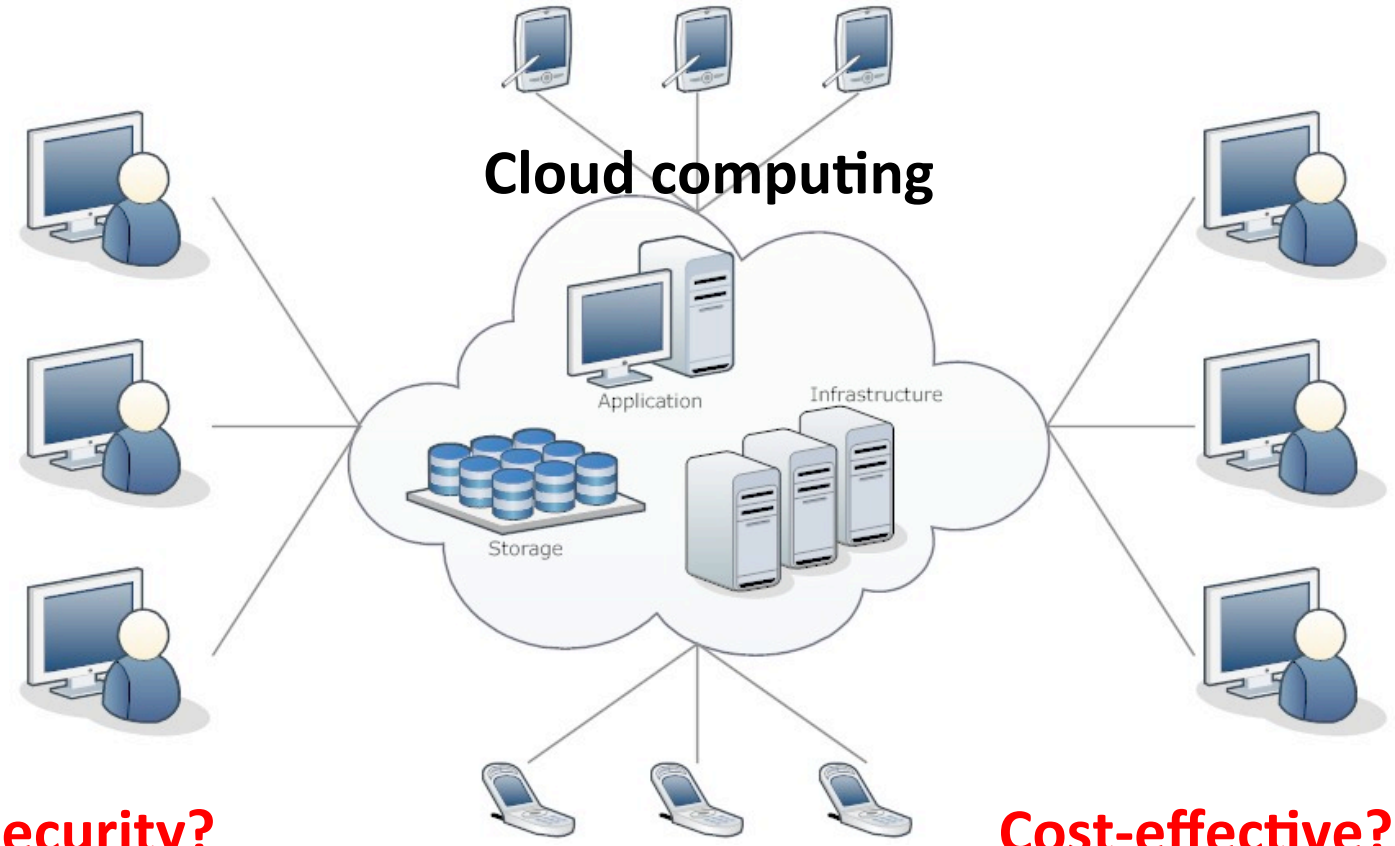



Omics methods are
not
defined by HIGH THROUGH-PUT...
...but by
HIGH OUT-PUT!
Large amount of data to analyze
New expertise
Expensive studies

We need to deal with large and complex data

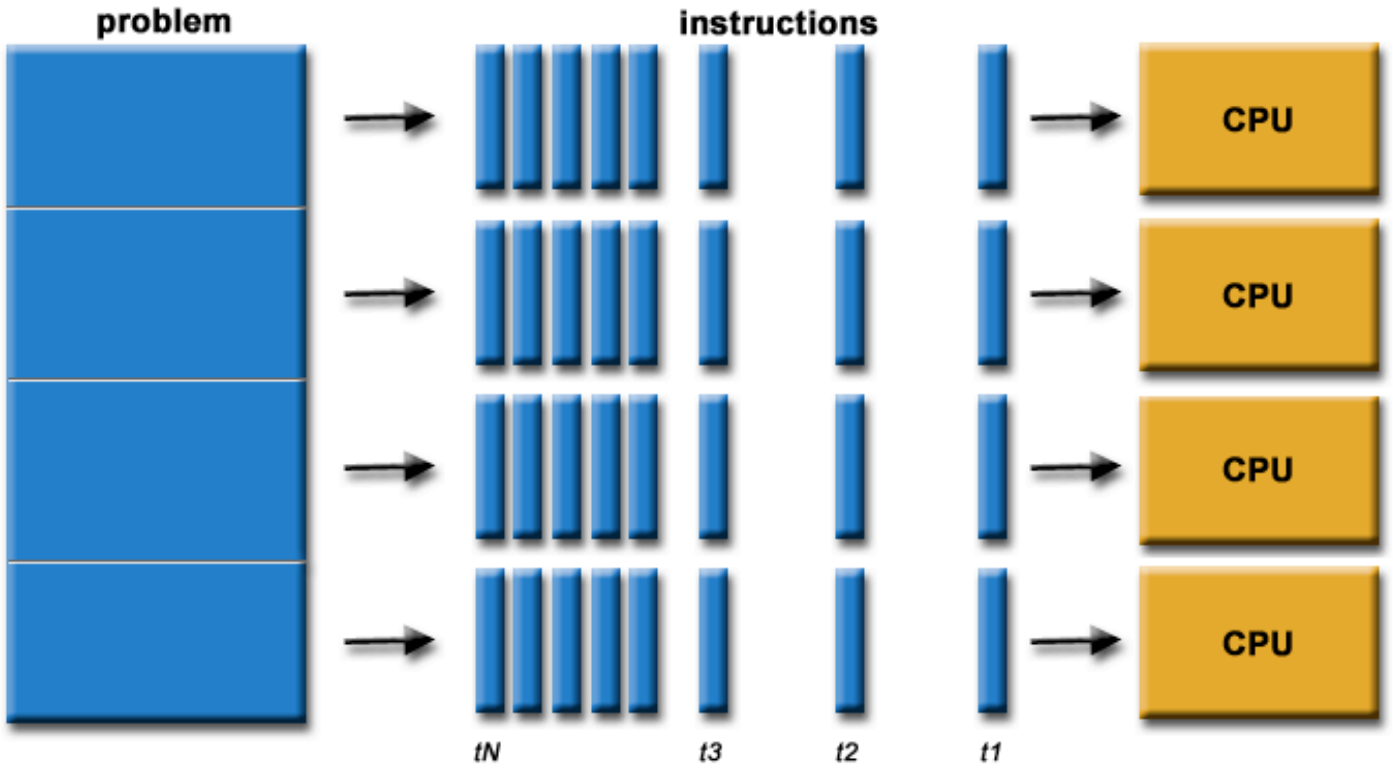
How to handle **Omics** data?

- Tools needed to manage large amounts of data



How to handle **Omics** data?

- Tools needed to manage large amounts of data
- New computational approaches needed



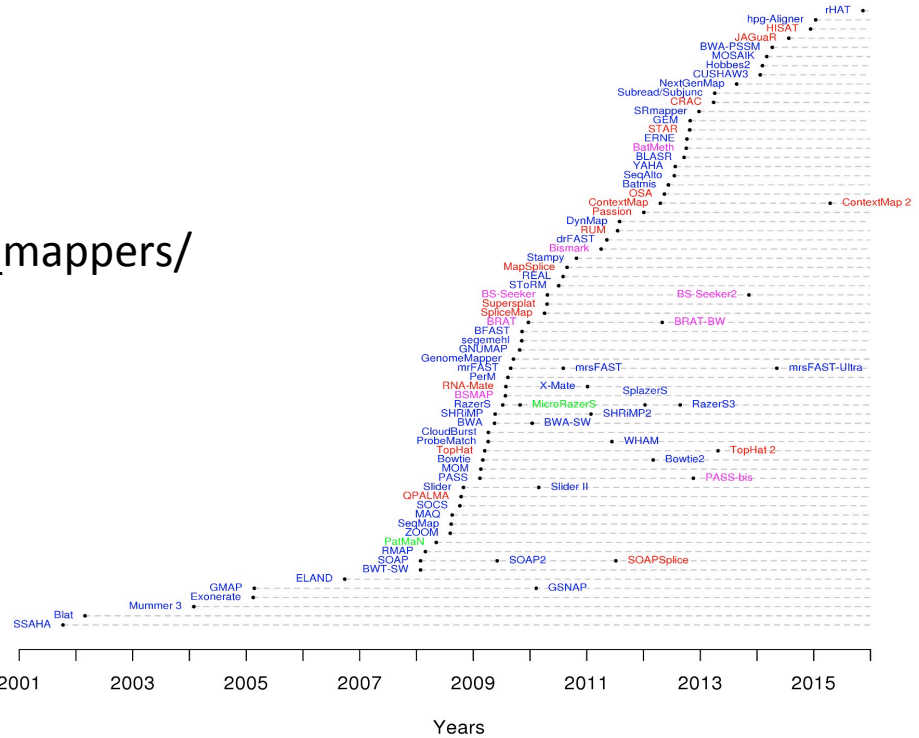
We need specific computational tools that can deal with these new data

How to handle **Omics** data?

- Tools needed to manage large amounts of data
- New computational approaches needed
- New methods for analysis and visualization needed

Mapping tools

http://www.ebi.ac.uk/~nf/hts_mappers/

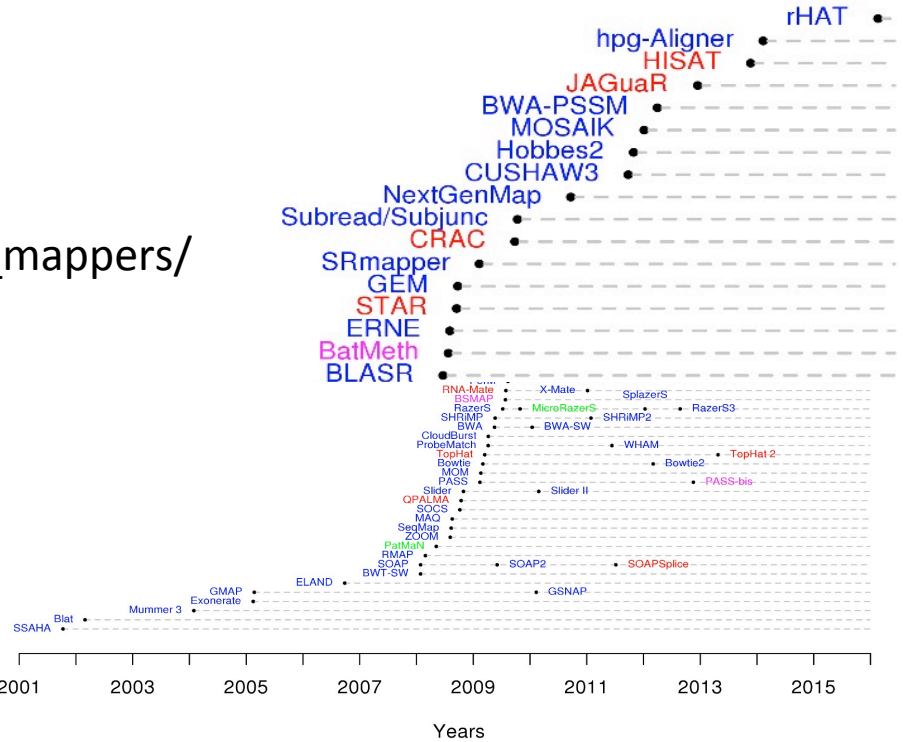


How to handle **Omics** data?

- Tools needed to manage large amounts of data
- New computational approaches needed
- New methods for analysis

Mapping tools

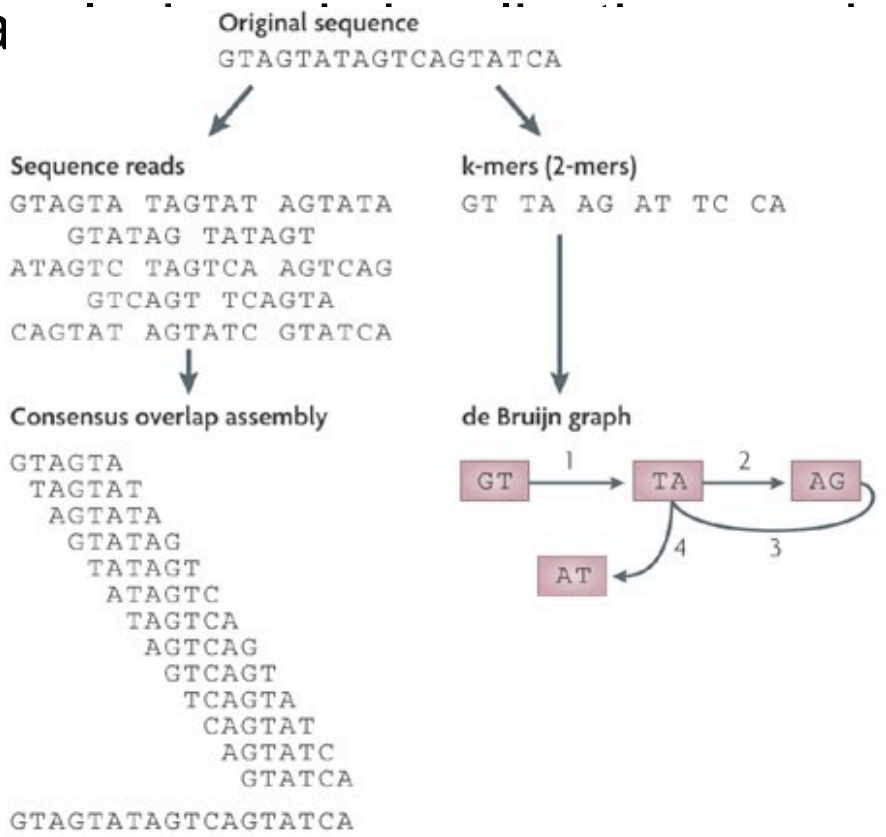
http://www.ebi.ac.uk/~nf/hts_mappers/



How to handle **Omics** data?

- Tools needed to manage large amounts of data
- New computational approaches needed
- New methods for a [illegible] ed

De bruijn algo



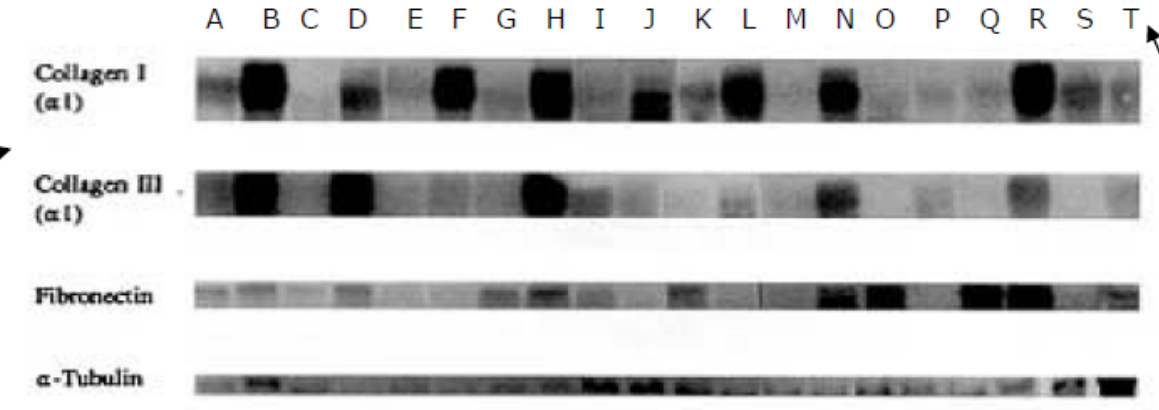
How to handle **Omics** data?

- Tools needed to manage large amounts of data
- New computational approaches needed
- New methods for analysis and visualization needed
- Experiments + theory needed for design for omics experimentation:
  - Sampling resolution?
  - Dosis concentration?
  - Study which (parts of) cells?
- New ideas and concepts about regulation of biological functions needed.

New ways of analyzing and showing

How did life change for a biologist?

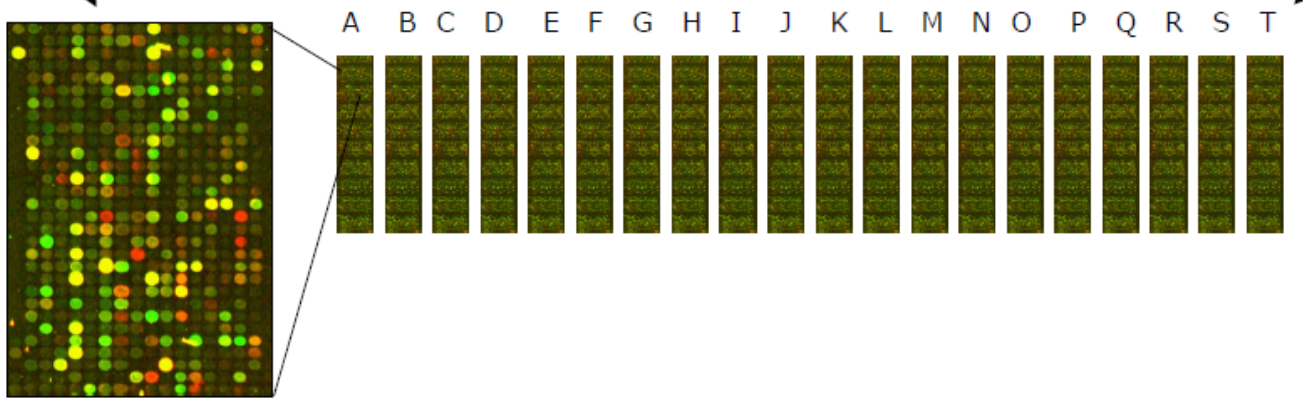
RNA analysis by Northern blot: 1-15 genes



Analyzed genes

Samples of 20 cellular experiments

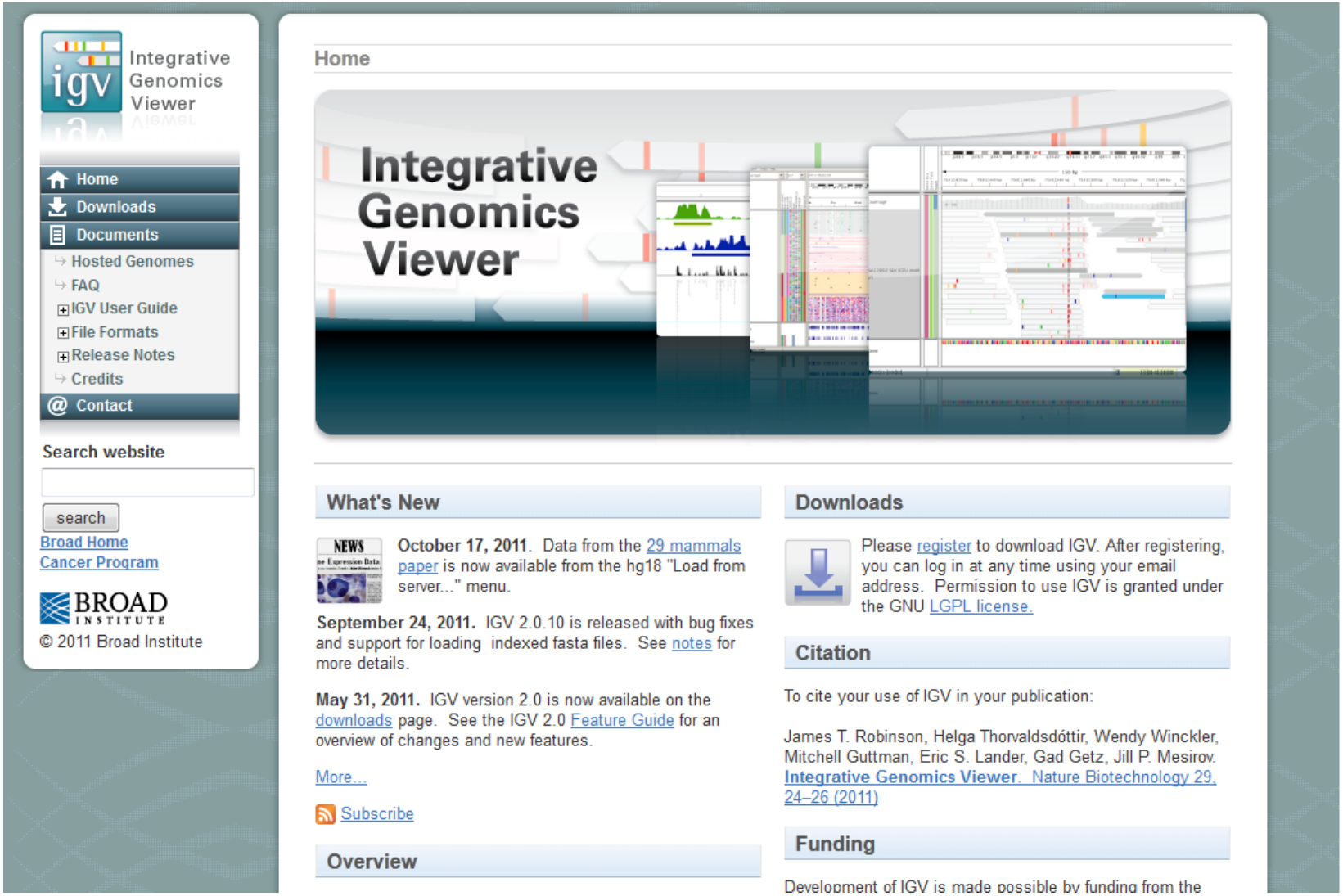
RNA analysis by micro-array: 1.000-40.000 genes



Gbrowse/ Jbrowse

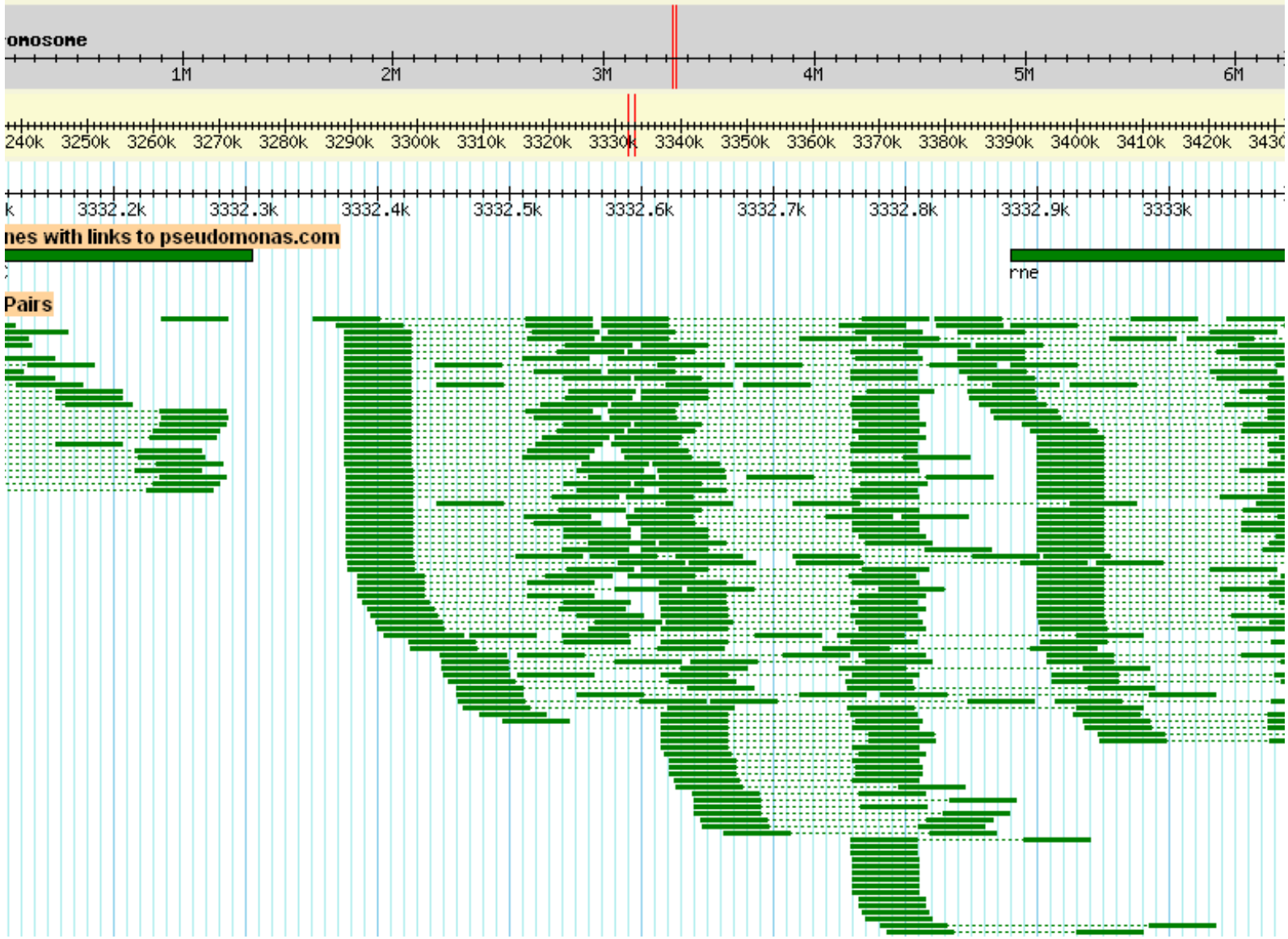
Visualization of NGS Data - Standalone

http://www.broadinstitute.org/igv/



Visualization of NGS Data – Web Site

http://gmod.org/wiki/GBrowse_NGS_Tutorial



Application exemples

Category	Examples of applications
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commensal flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals

Statistique et données massives : enjeux et perspectives

UPMC, campus de Jussieu (amphi 25), Paris, mardi 13 octobre 2015

ACCUEIL COMITÉ D'ORGANISATION PROGRAMME INSCRIPTION INFORMATIONS PRATIQUES

SOCIETE FRANÇAISE DE STATISTIQUE

Présentation de la manifestation

Après le succès de la journée « Horizons de la Statistique » en 2014, la Société Française de Statistique (SFdS) organise le mardi 13 octobre 2015 sur le campus de Jussieu une manifestation intitulée **Statistique et données massives : enjeux et perspectives**.

L'objectif de la journée est de discuter des enjeux nouveaux que soulèvent les données massives sur nos sociétés, et les perspectives qu'elles ouvrent aussi bien pour les entreprises, les institutions et les statisticiens qui y seront confrontés dans les prochaines années. Pour ce faire, la SFdS propose une belle affiche, équilibrée entre mathématiques, informatique, industrie et monde académique, et choisit de donner la parole à des orateurs prestigieux :

http://bigdata2015.sfds.asso.fr/

L'événement sera également retransmis en direct (streaming) à l'adresse : http://video.upmc.fr

The Future of **Omic** Research

- Six fields were targeted for development as **Omic** information grows
  - Resources: Genome sequences and libraries/DB
  - Technology such as new sequencing methods
  - Software for computational biology
  - Training professionals in interdisciplinary skills
  - Ethical, legal, and social implications
  - Education of health professionals and public

How to store inputs/ outputs sequencing data?

Benchmarking (Performance comparisons)

Tools and packages dedicated to sequencing data

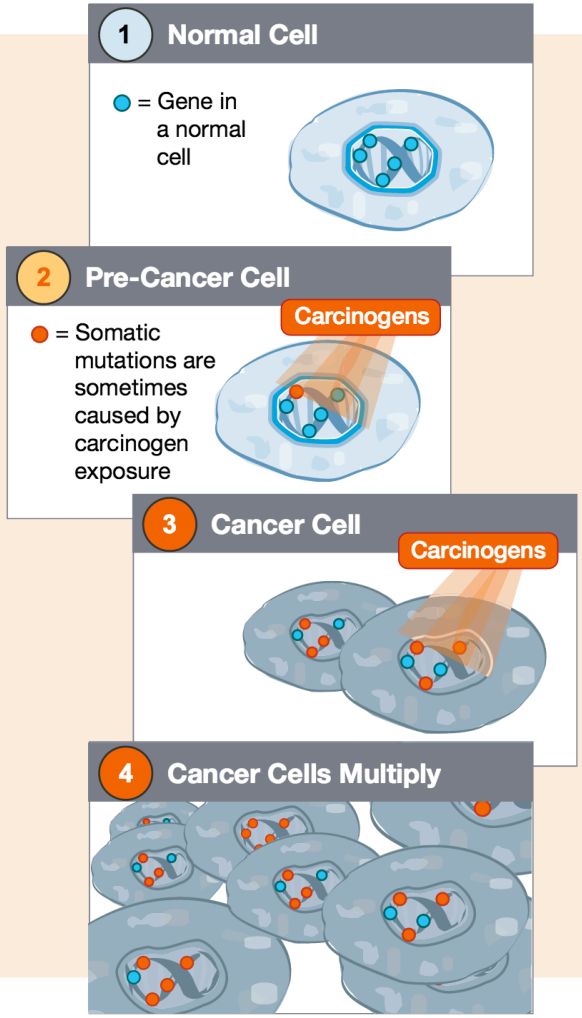
New way to analysis and visualize

One example: Cancer
A Disease of the Genome

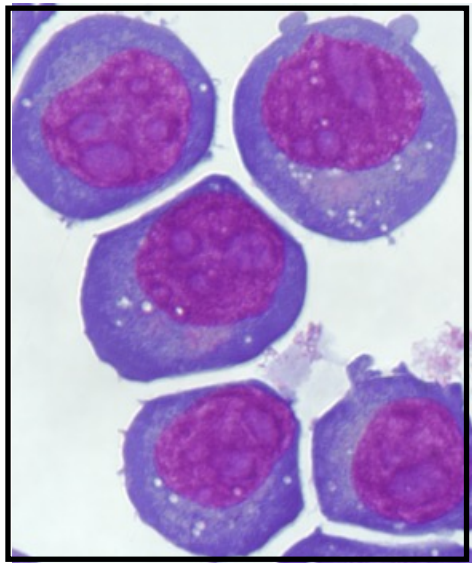
What Causes Cancer?

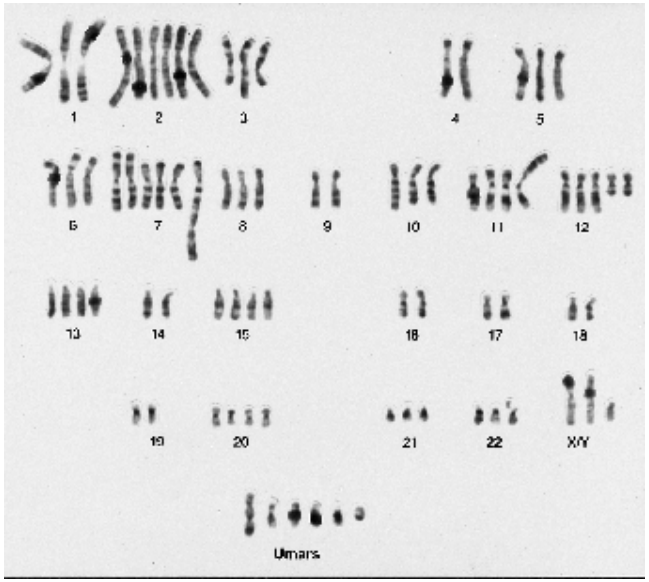
The development of cancer is a multi-step process in the life of a cell. Cancer can be initiated by mutations in a cell's DNA. DNA is the molecule in our body that carries genetic information we inherit from our parents. A mutation in DNA is simply a change in DNA, but these changes can occur through a variety of mechanisms, some that are inherited and some that are acquired after birth.

All children are born with some genetic mutations that they inherit from their parents. These are called germline mutations. Other DNA mutations occur sporadically throughout our lifetime. These mutations are called somatic mutations. They are sometimes caused by exposure to carcinogens from the environment or from lifestyle choices, such as tobacco use. Somatic mutations are responsible for the vast majority of cancers. The purpose of TCGA is to create an "atlas" of the significant somatic mutations associated with most cancers.



One example: Cancer
A Disease of the Genome



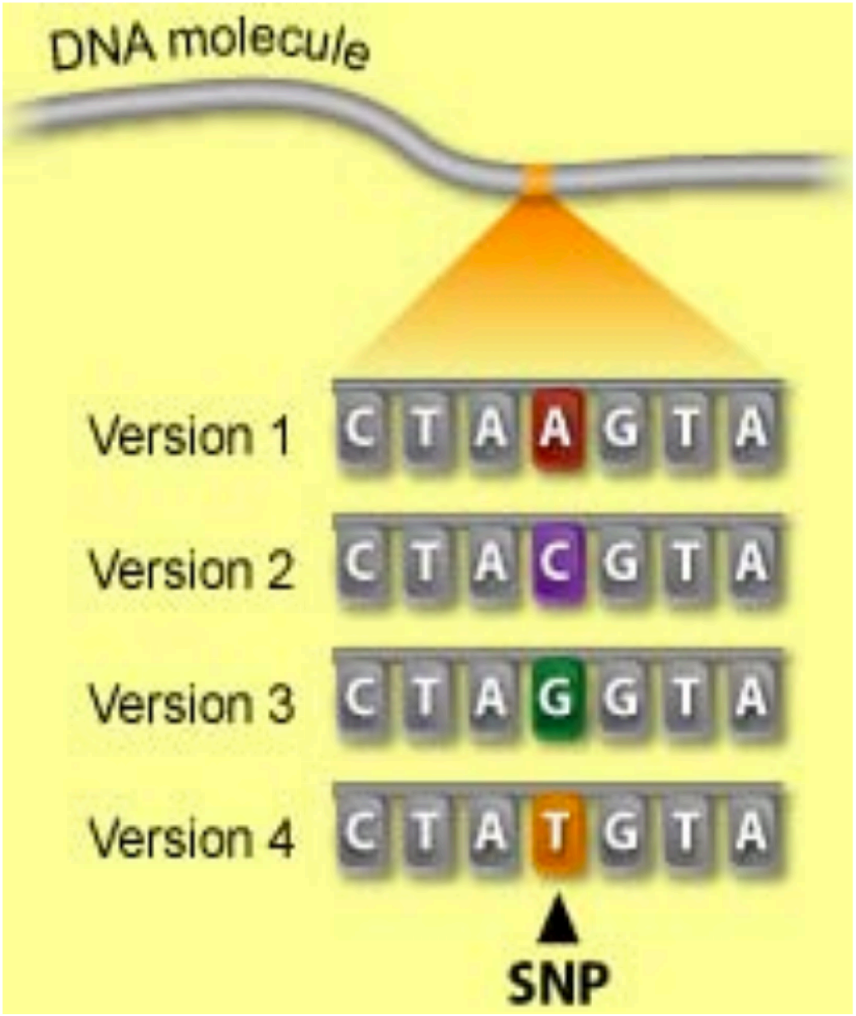


Goal

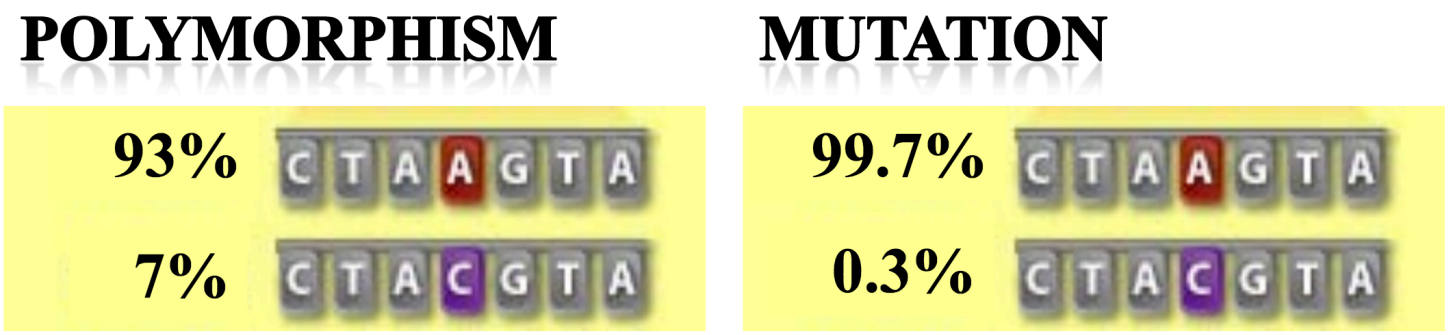
- Identify **changes** in the genomes of tumors that drive cancer progression
- Identify new targets for therapy
- Select drugs based on the genomics of the tumor

One type of change: Single nucleotide polymorphism (SNP)

- Single base substitution
- Four versions possible but generally only two (major and minor) exist
- About 1 in every 300 bp in human genome
- Translates to ~10 million SNPs across the genome



SNP *vs.* mutation



Polymorphism: Single DNA base change found in >1% of population

Mutation: Single DNA base change found in <1% of population (somatic)

Genetic mutations are one type of genetic polymorphism

Beyond SNPs

- Small insertion/deletions (Indels)

```
Ref TGCATT---TAGGC
    TGCATTCCGTAGGC
      Insertion
```

```
Ref TGCATTCCGTAGGC
    TGCATT---TAGGC
      Deletion
```

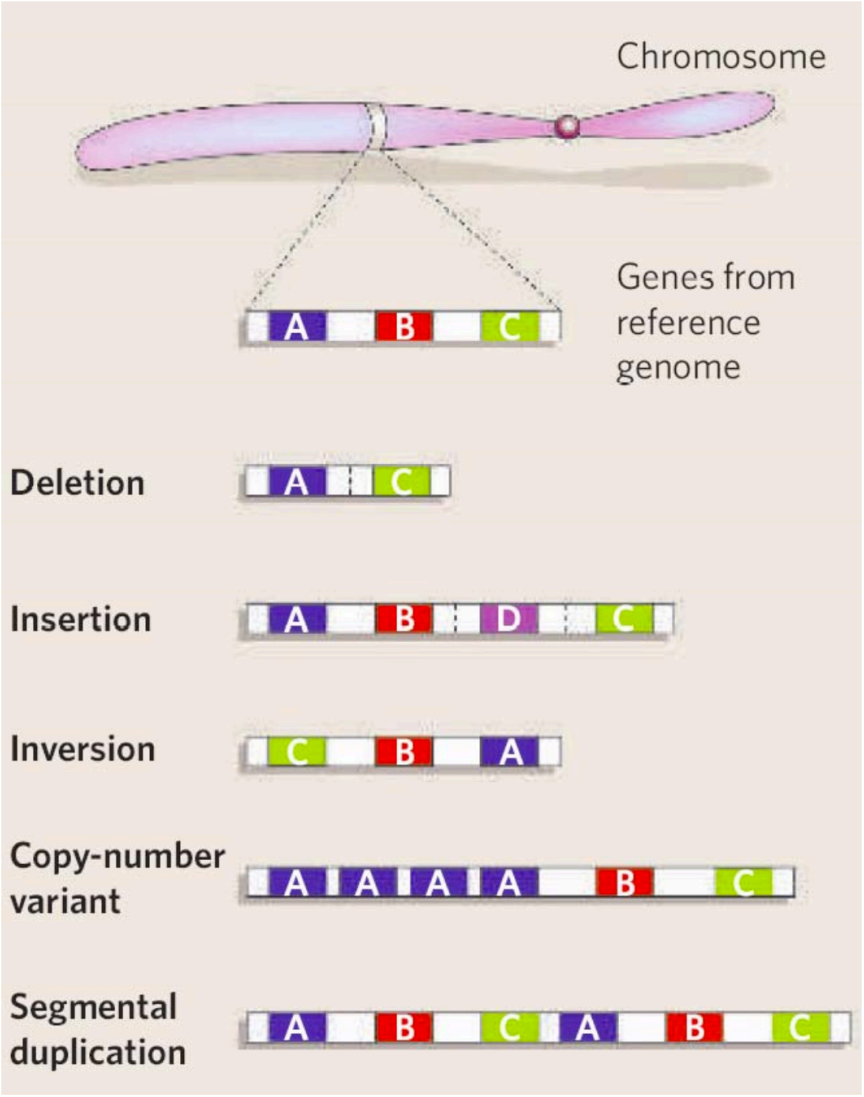
- Structural Variation (SV)
  - Variable number tandem repeat (VNTR) polymorphisms
    - very common, >600,000 simple repeats in the human genome
    - Microsatellites: 2-6 bp repeat units

```
Ref TGCTCATCATCATCAGC
    TGCTCATCA------GC
```

    - Minisatellites: 10-100s bp repeat units
  - Unbalanced structural variants or copy number variants (CNVs)

Why have SVs been ignored?

- SV traditionally defined as deletions, insertions, or inversions > 1 kb
- Often involves repetitive regions of the genome and complex rearrangements
- Importance not recognized
- No optimal method for SV discovery



One example: Cancer
A Disease of the Genome

Next-generation sequencing is enabling worldwide collaborative efforts, such as the International Genome Consortium (ICGC) and The **Cancer Genome Atlas (TCGA) project**, to catalogue the genomic landscape of thousands of cancer genomes across many disease types.

DATA

The Cancer Genome Atlas (TCGA) project has analyzed mRNA expression, miRNA expression, promoter methylation, and DNA copy number in 489 high-grade serous ovarian adenocarcinomas (HGS-OvCa) and the DNA sequences of exons from coding genes in 316 of these tumors.

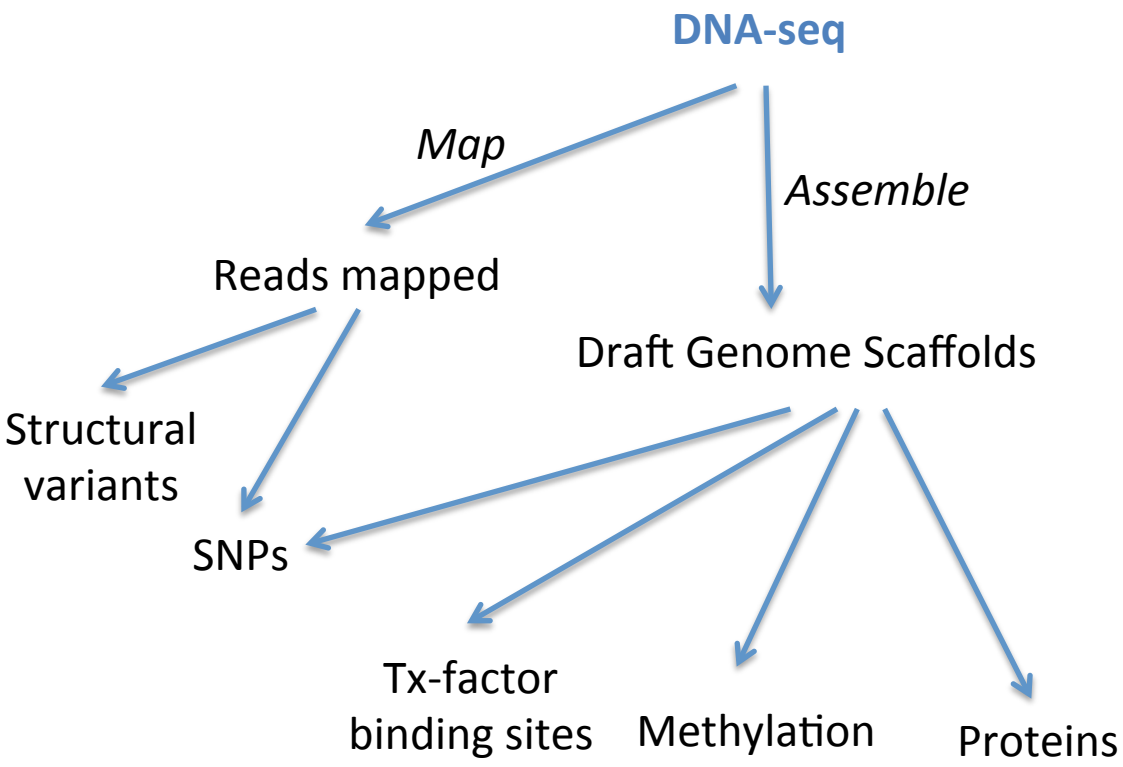
Results

- High grade serous ovarian cancer is characterized by *TP53* mutations in almost all tumors (96%)
- Low prevalence but statistically recurrent somatic mutations in 9 additional genes including *NF1, BRCA1, BRCA2, RB1*, and *CDK12*
- 113 significant focal DNA copy number aberrations and
- promoter methylation events involving 168 genes.

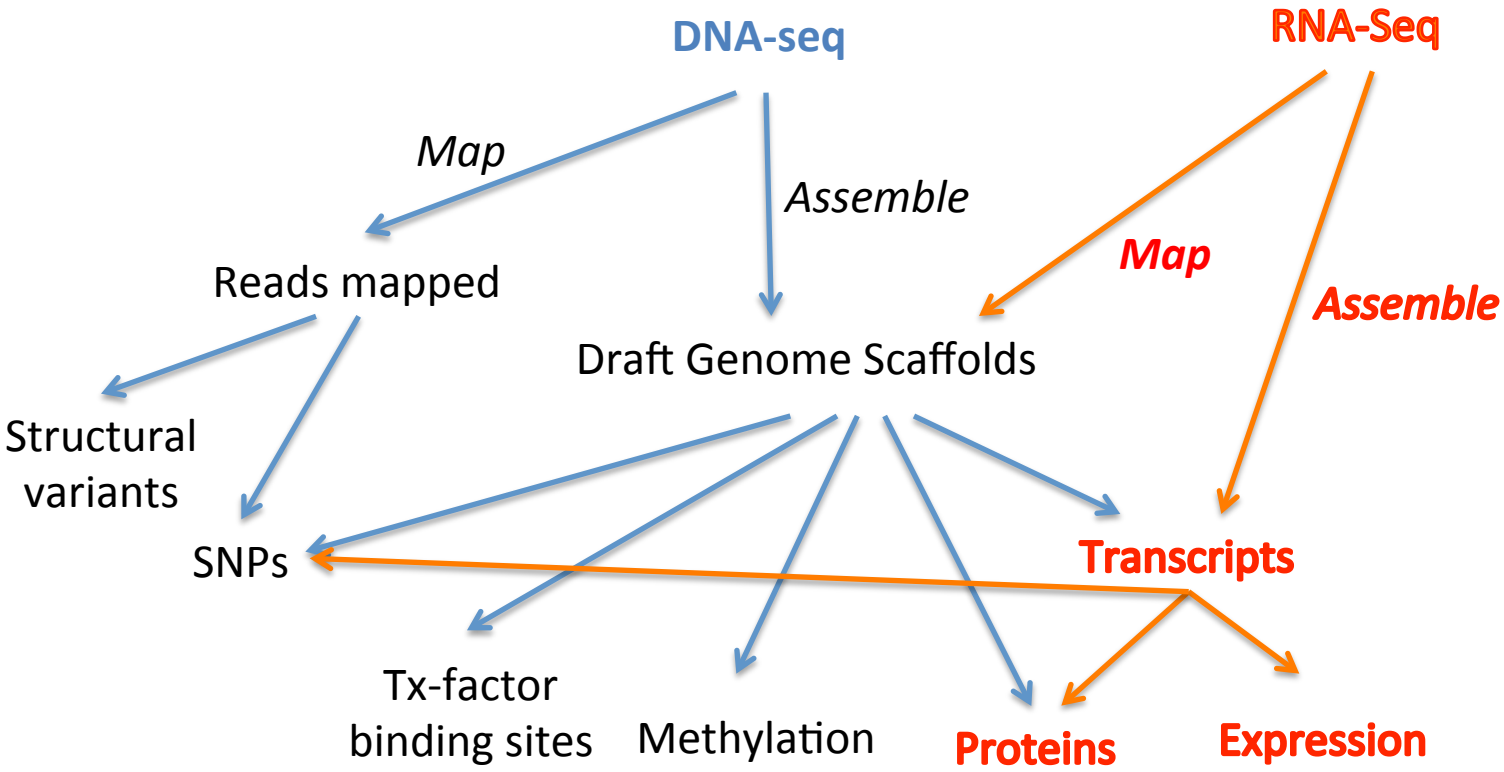
➔ Pathway analyses suggested that homologous recombination is defective in about half of tumors, and that Notch and FOXM1 signaling are involved in serous ovarian cancer pathophysiology.

The Cancer Genome Atlas Research Network, 2011

Applications



Applications



Applications