

Omics or how to handle massive biologic data? Emergence of new sciences

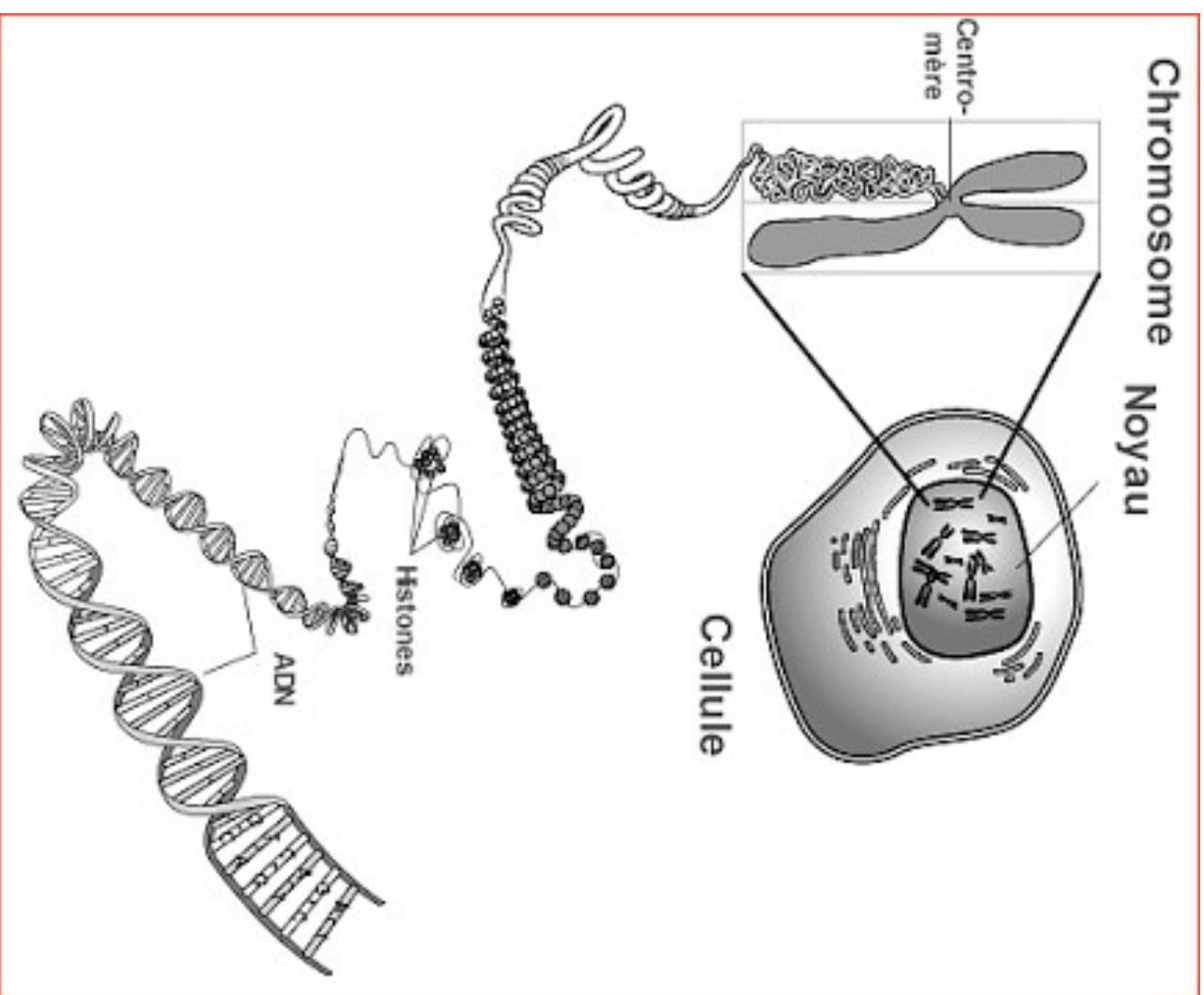
Anna-Sophie Fiston-Lavier

Background

(1) Where and how genetic information is organized

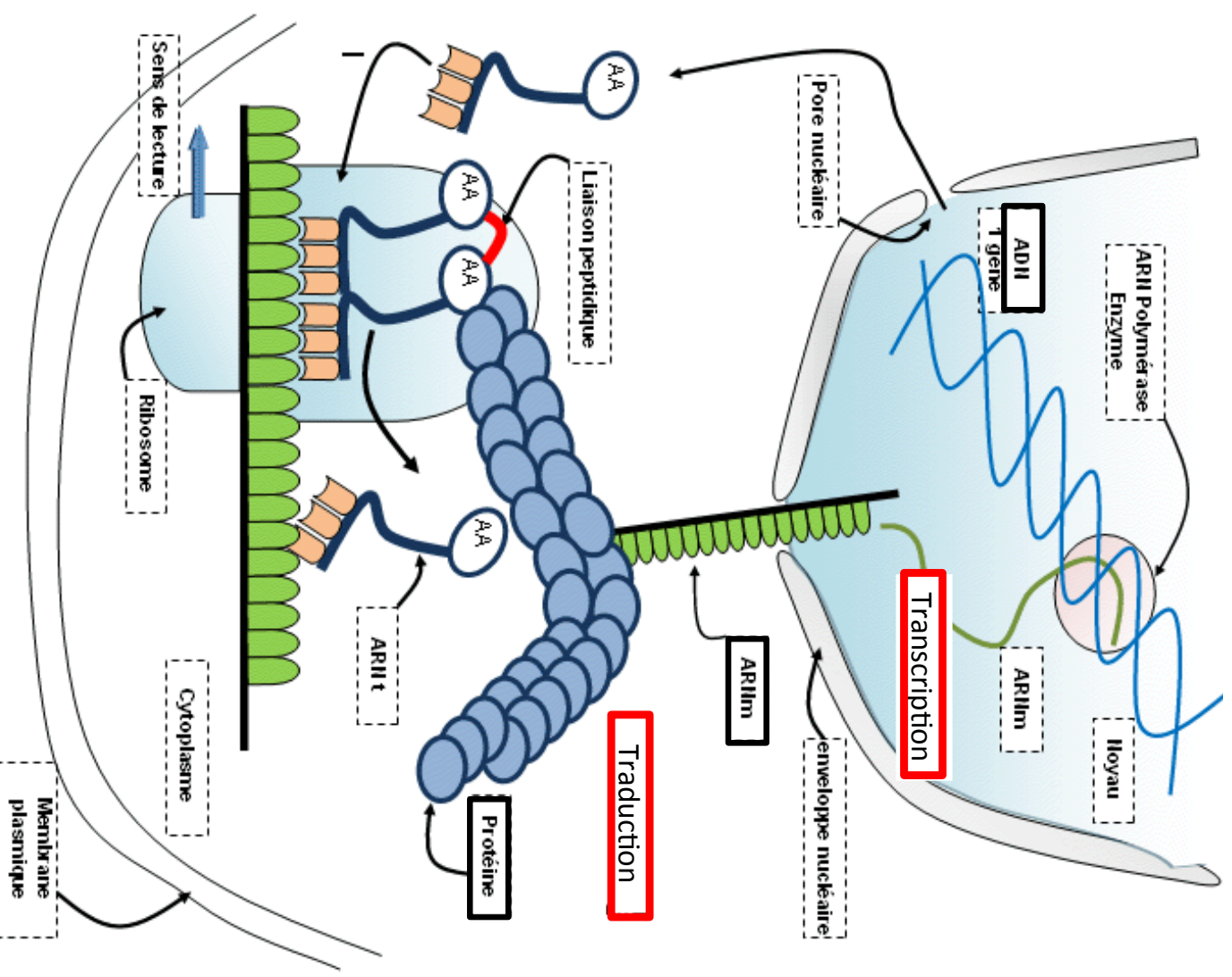
Cell = Unit of the organisms

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



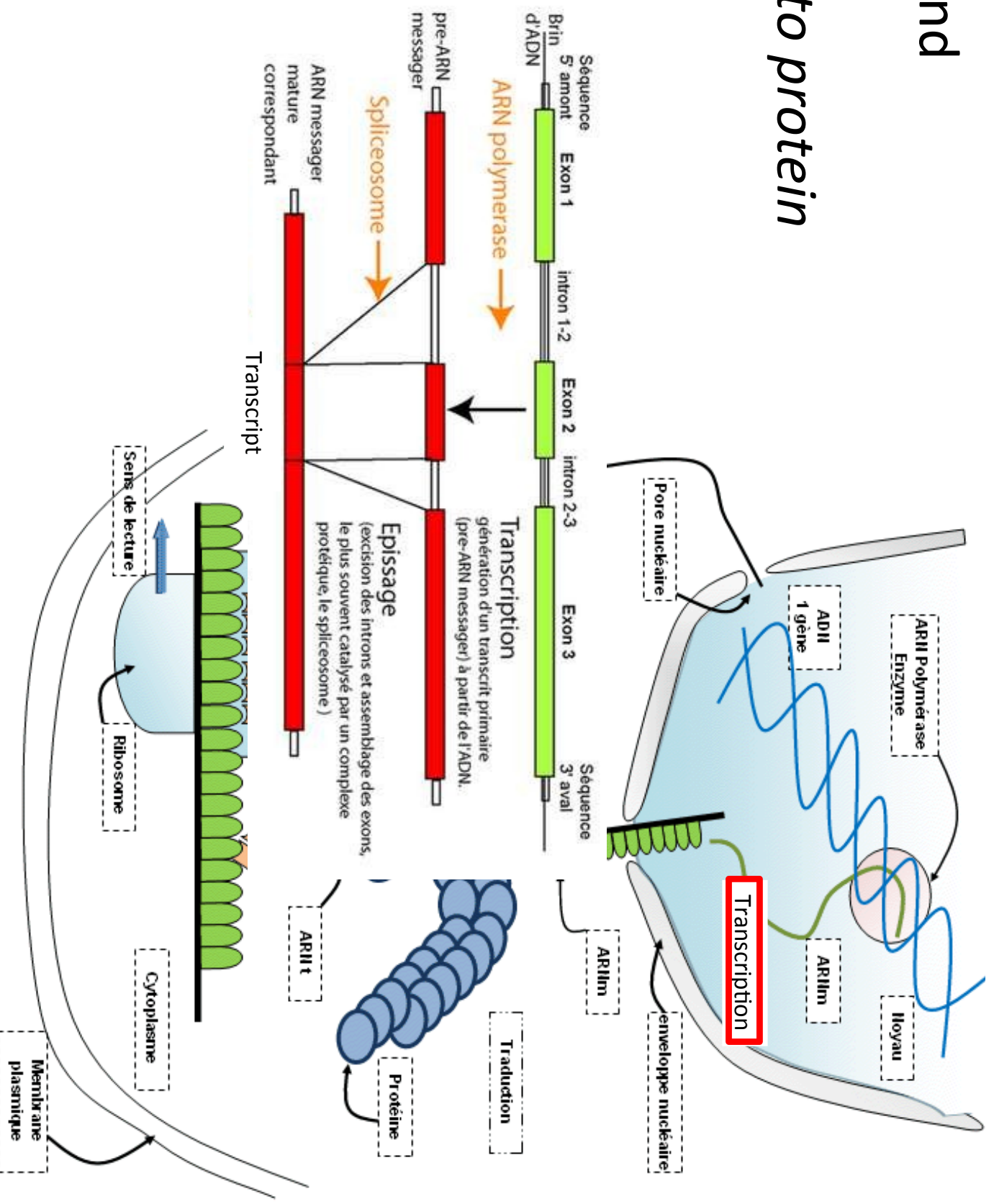
Background

(2) *Gene to protein*



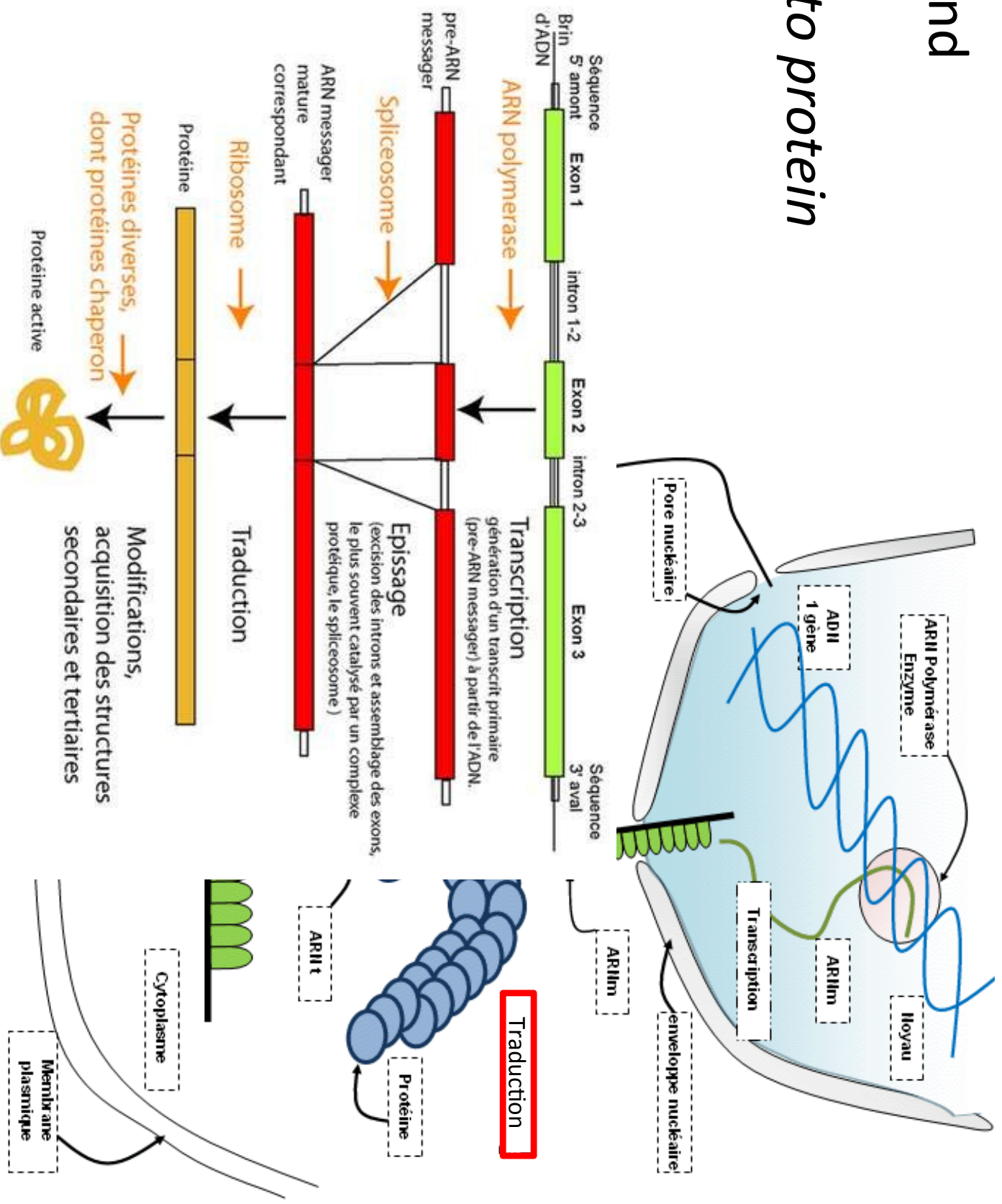
Background

(2) Gene to protein



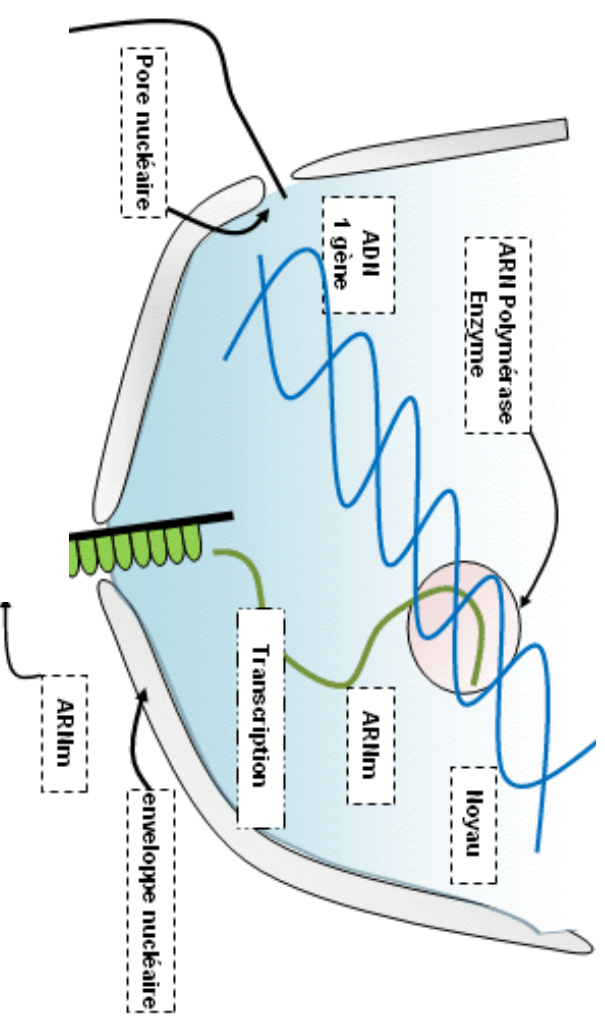
Background

(2) Gene to protein

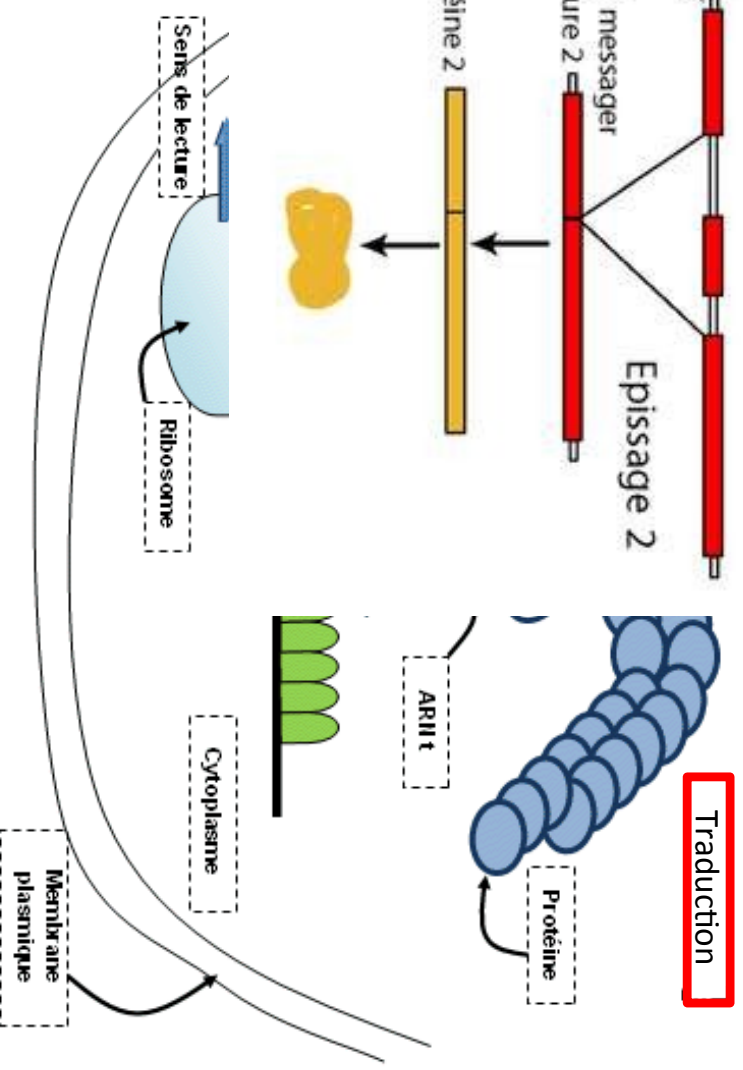
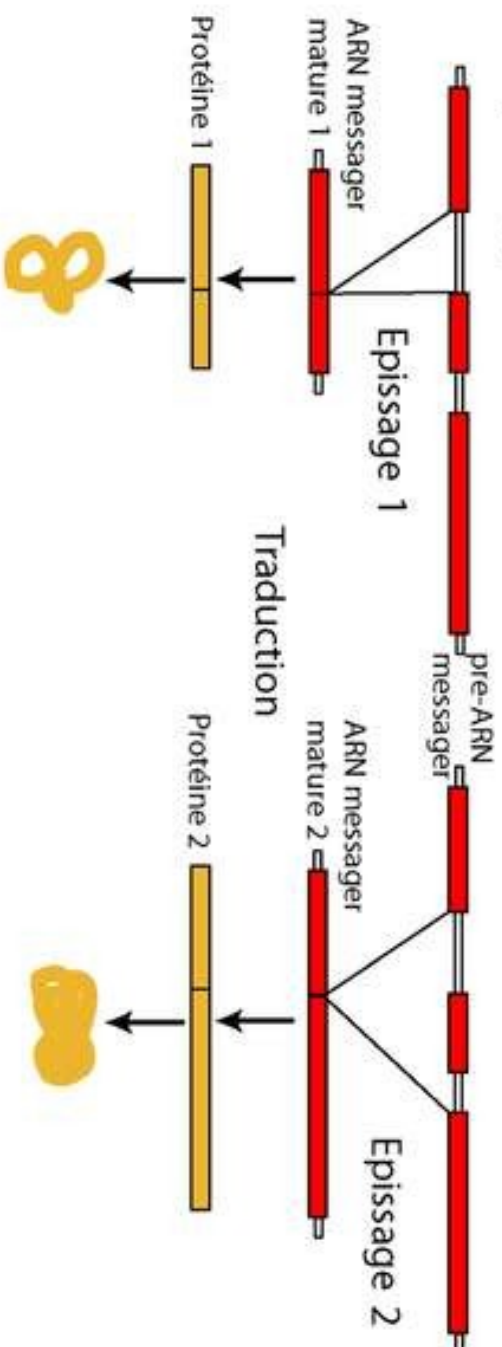


Background

(2) *Gene to protein*

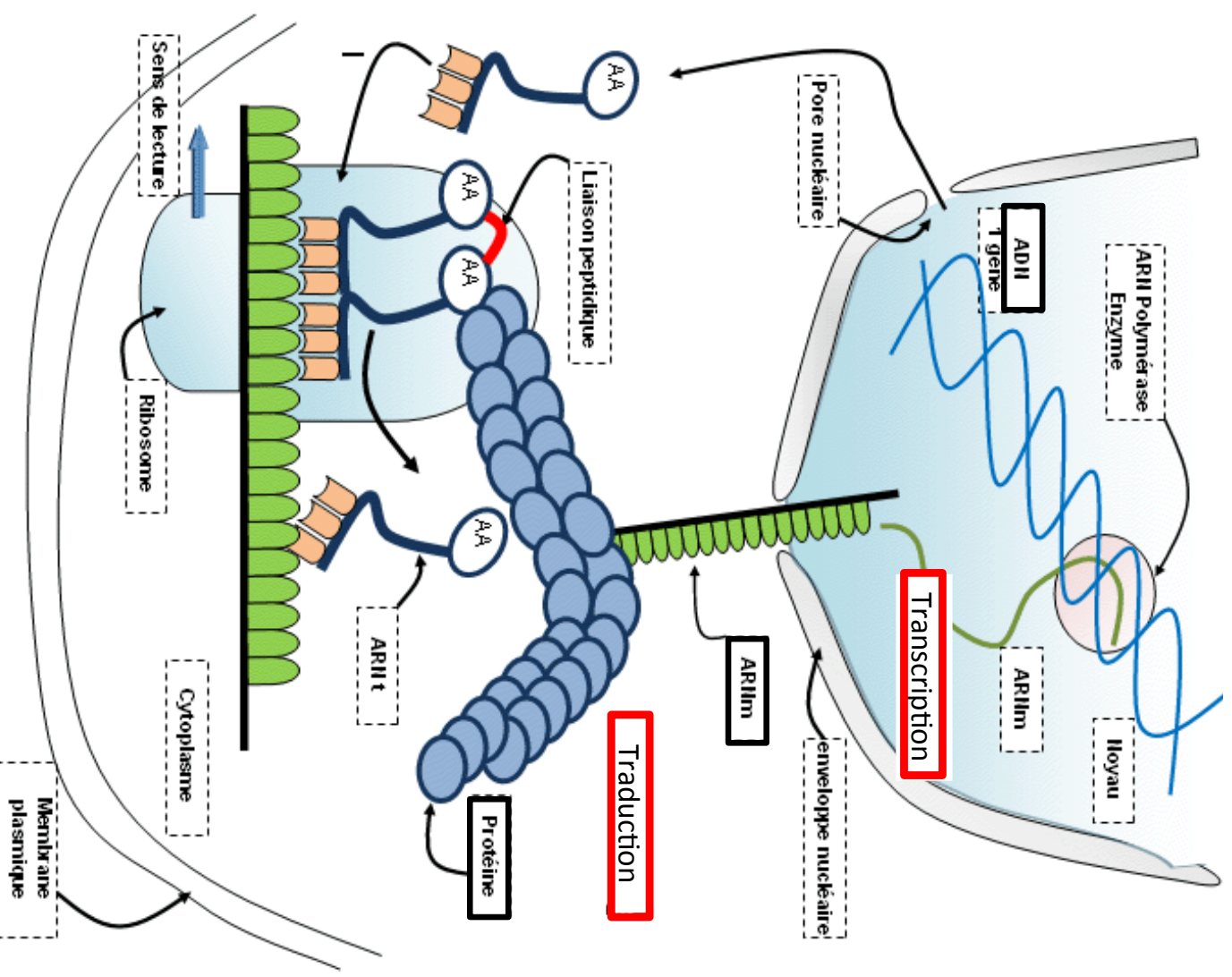


L'épissage alternatif

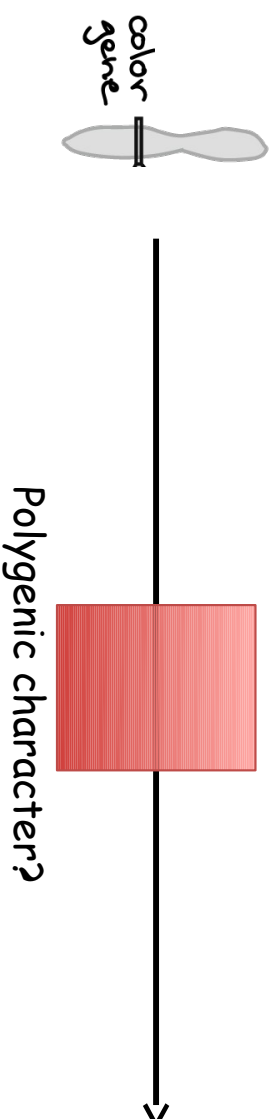
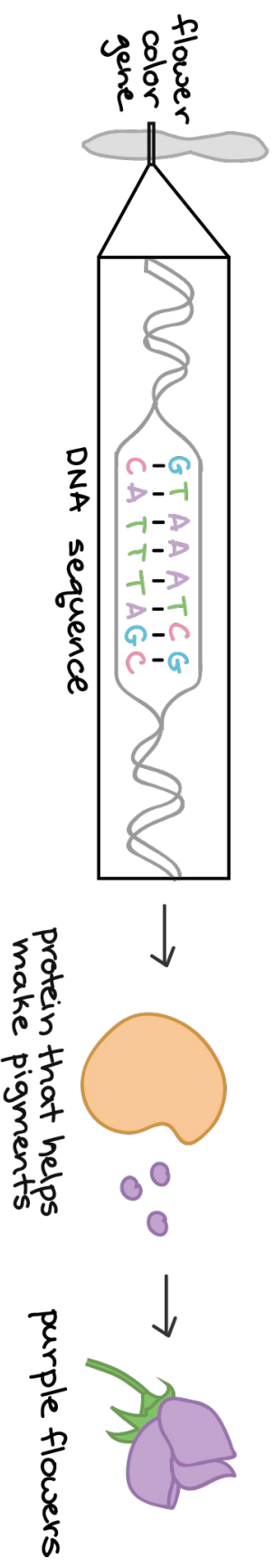


Background

(2) *Gene to protein*

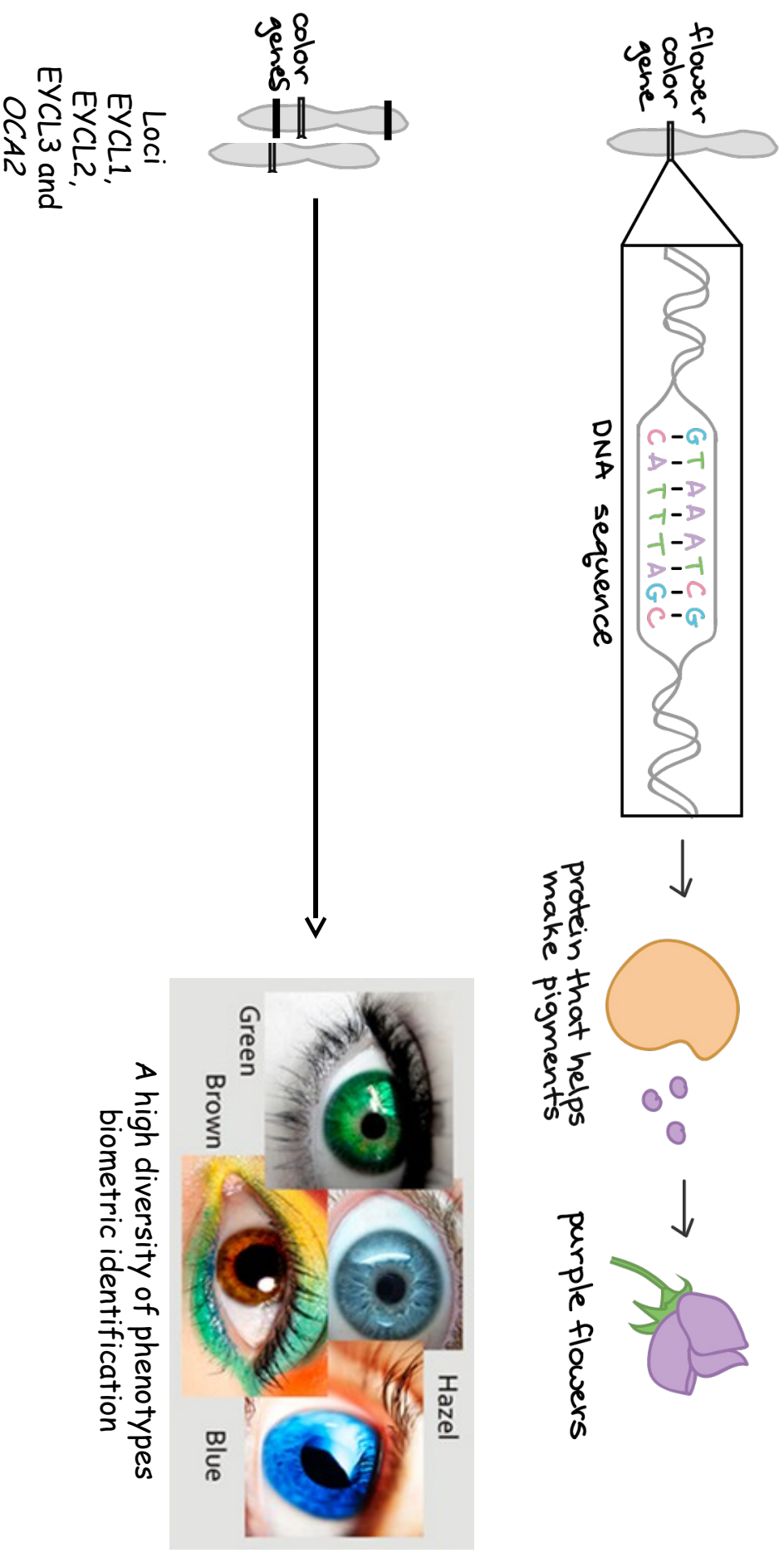


Dogma



A high diversity of phenotypes

Dogma



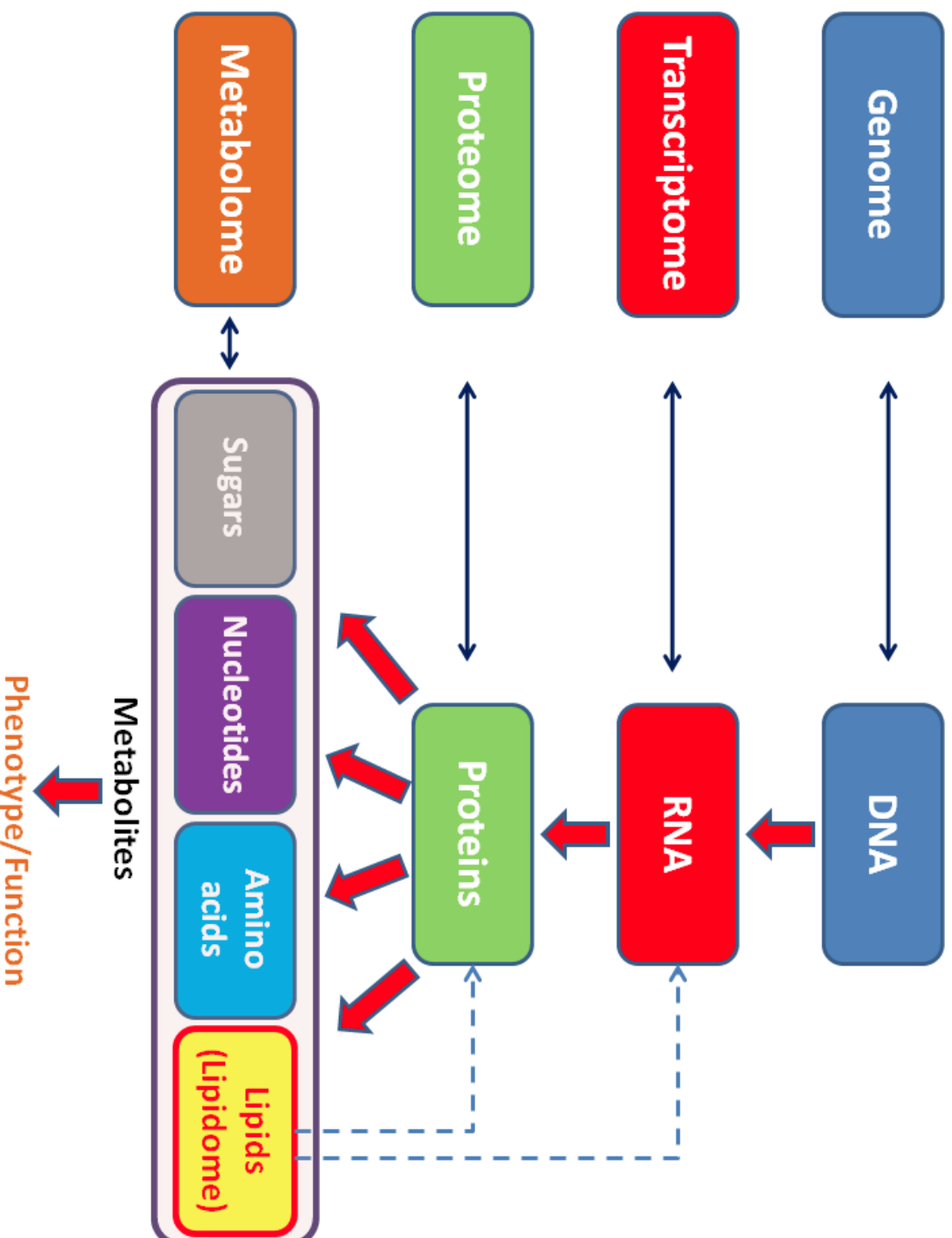
A high diversity of phenotypes
biometric identification

“Omics” a new way of thinking

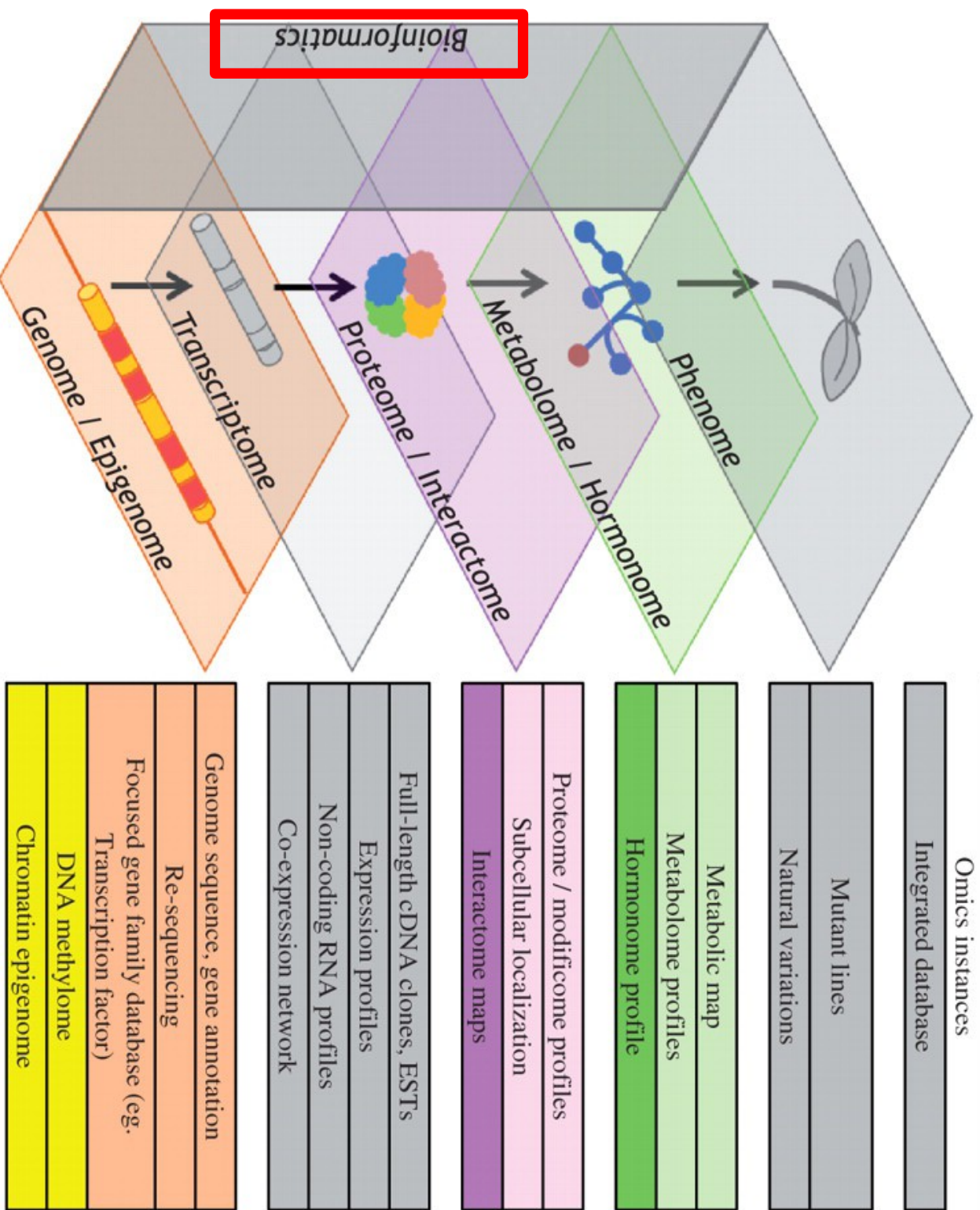
Ome = set of objects of study of such fields

Omics = field of study in biology ending in this suffix

The ‘Ome’s



The ‘Omics’



The ‘Omics’ for what?

Genome

Identification of candidate genes

Transcriptome

Identification of genes expressed

Proteome

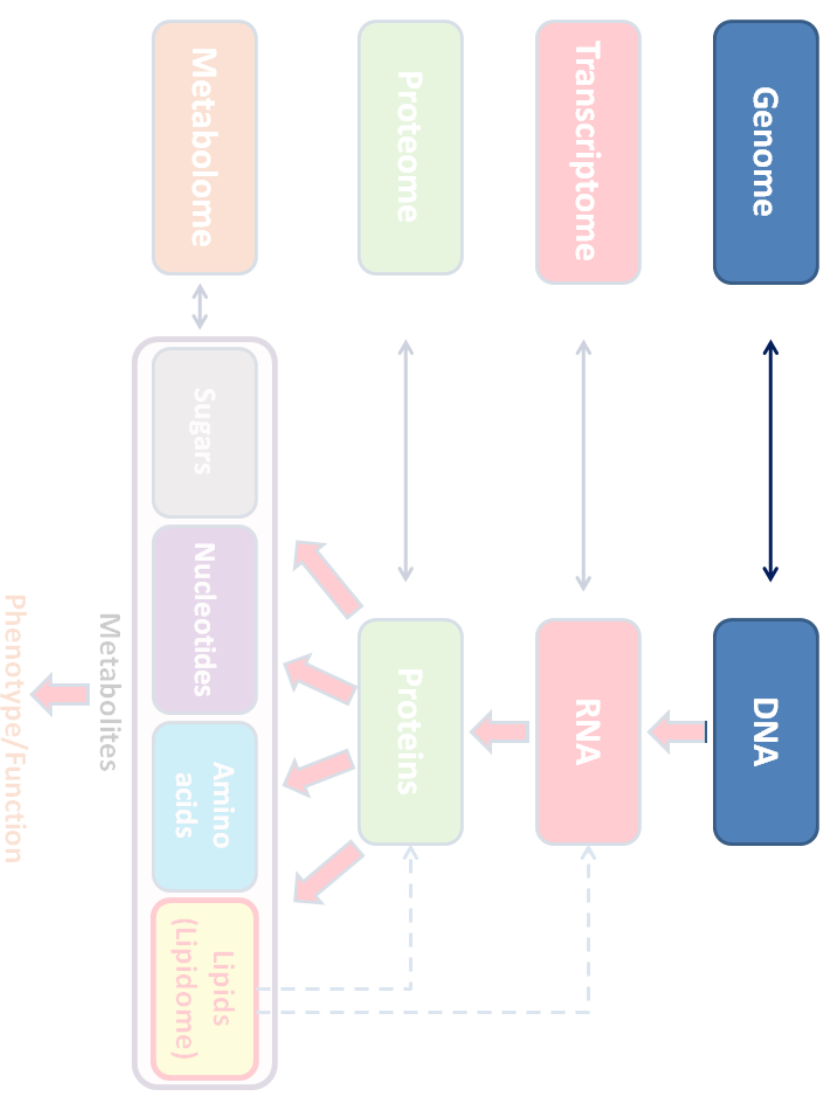
Identification of proteins produced

Metabolome

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.

Next in Genomics

- **Structural genomics** = generate new sequence assemblies, sequence organization
- **Comparative genomics** = identification conserved and unknown genomic sequences and interpreting their evolutionary history
- **Functional genomics** = function of all the gene sequences and their expressions in an organism
- **Metagenomics...**

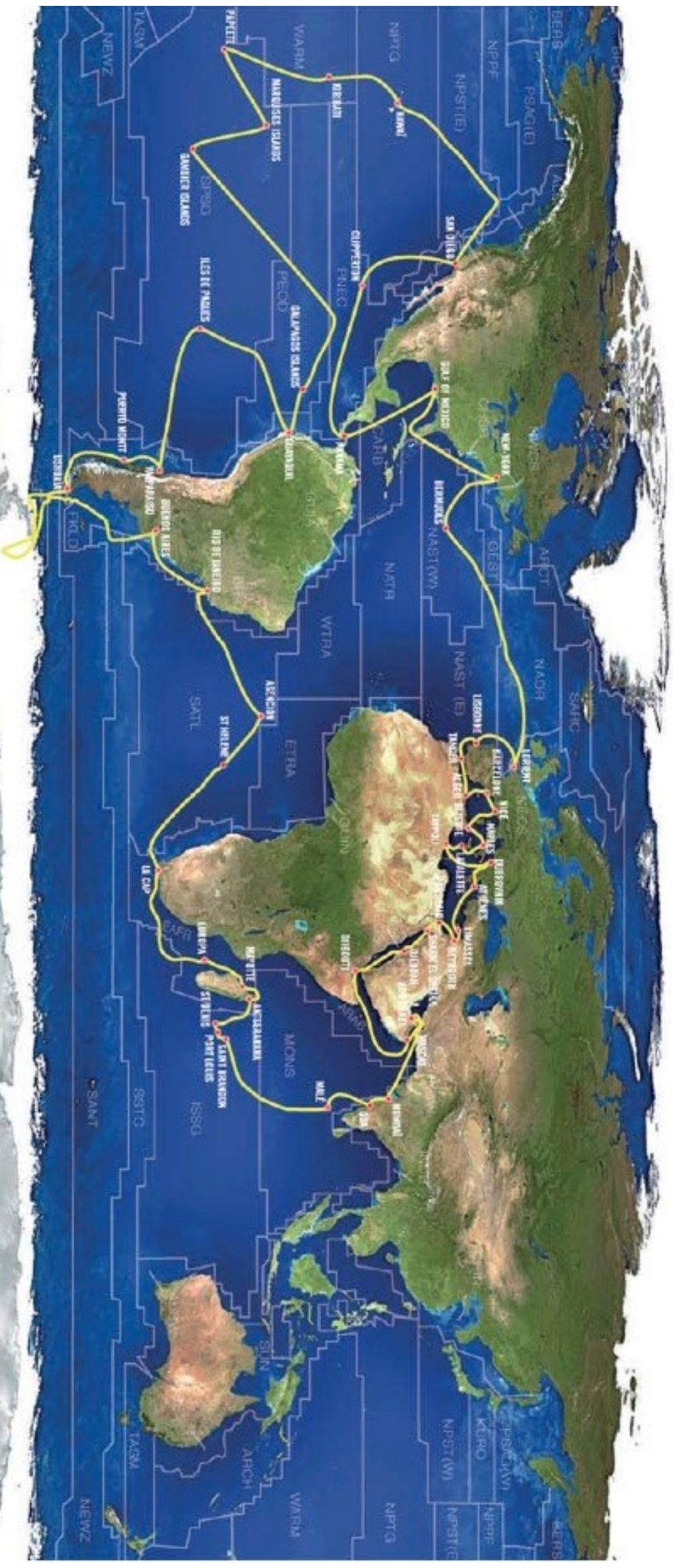
Metagenomics

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



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 project: a better understanding of marine, soil and

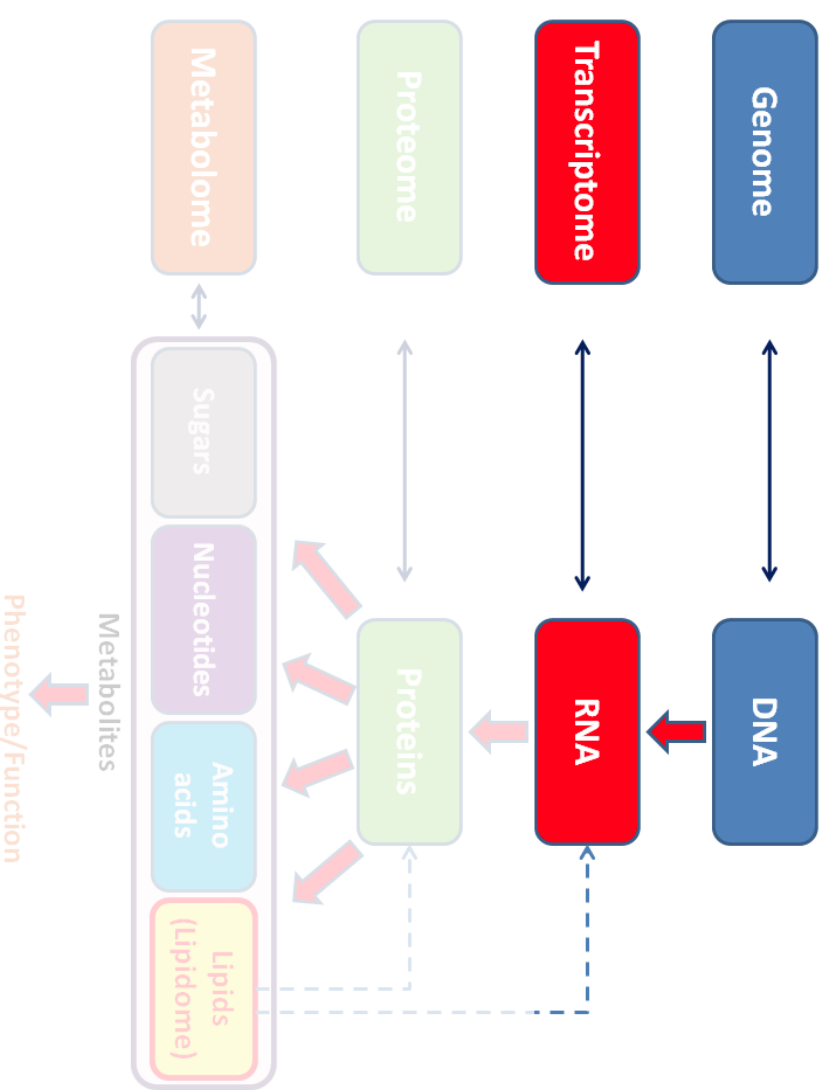


http://oceans.taraexpeditions.org/en/a-2-5-years-marine-and-scientific-expedition.php?id_page=1

Transcriptomics

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

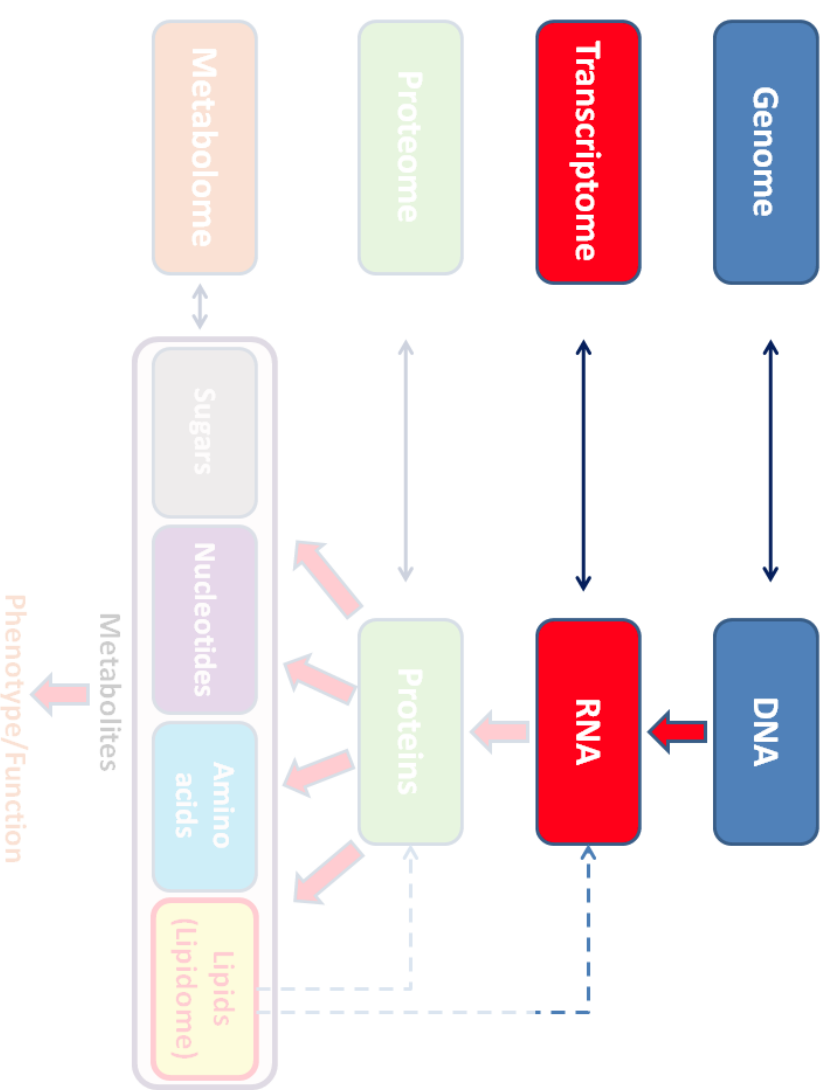


Transcriptome bypasses the need for exome enrichment

Transcriptomics

Transcriptome = complete set of all RNA molecules ("transcripts") produced from a genome.

specific subset of transcripts present in a particular cell type or under specific growth conditions



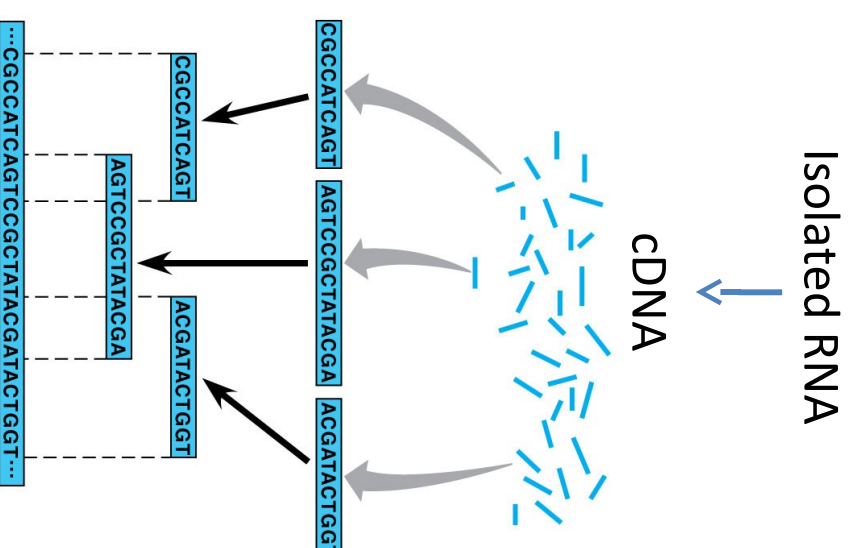
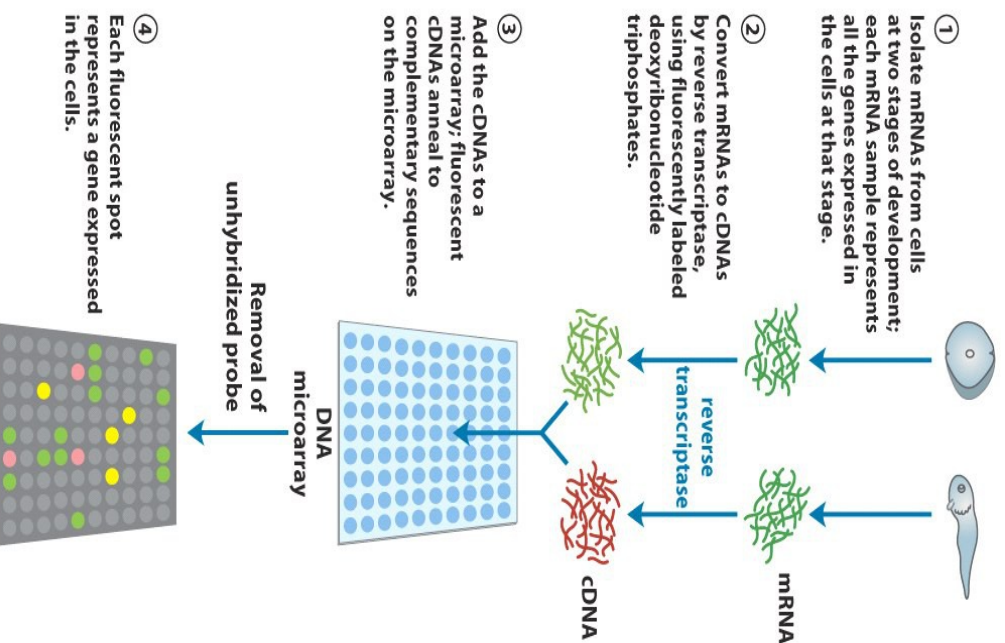
Transcriptomics involves large- scale analysis of RNAs to follow when, where, and under what conditions genes are expressed.

Transcriptomics: Expression profiling

High-throughput techniques based on

DNA microarray technology

NGS technology (RNA-seq)

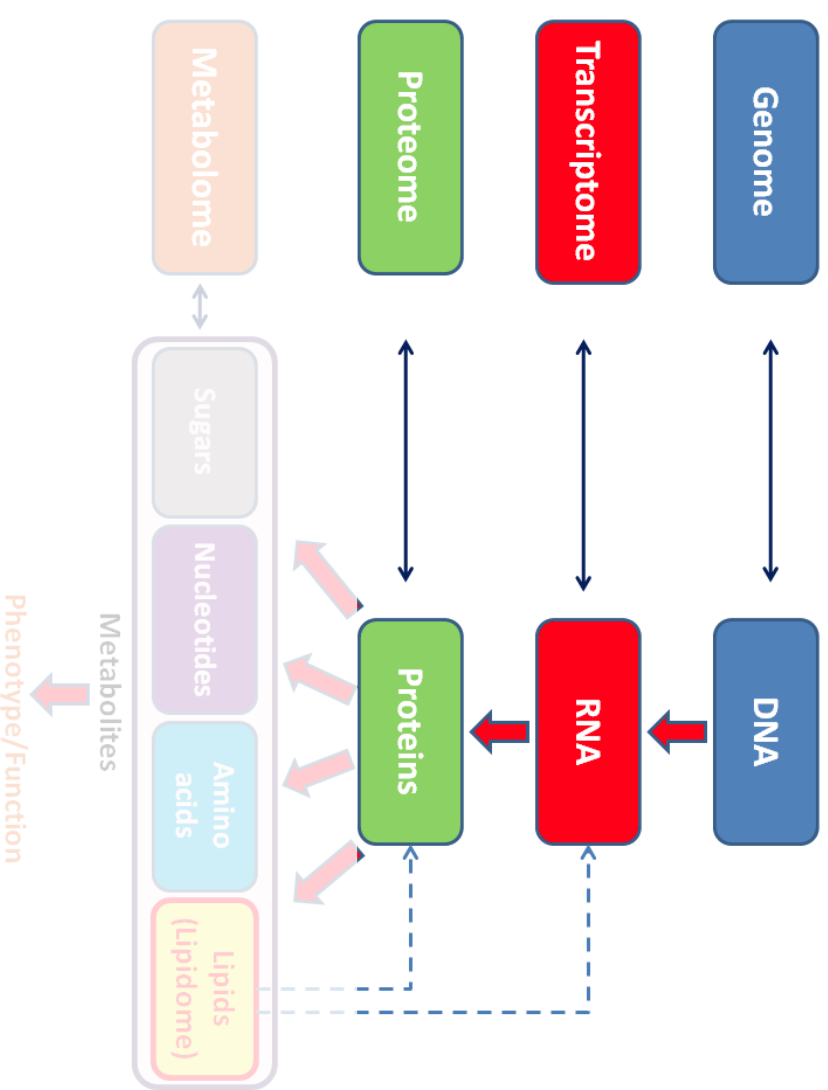


Proteomics

Proteome = complete set of proteins for a given organism

or

A complete set of protein produced under a given set of conditions



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

Proteomics an Extension of Genomics

Proteomics is the study of the structure and function of proteins, which is important in development of new diagnostic tests and drugs

Proteomics - Study of expressed proteins in a cell at a specific time under a particular set of circumstances

→ can bring researchers closer than gene expression studies to what's actually happening in the

Proteomics

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

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

Role of Proteomics

- 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

Transcriptomics vs. Proteomics

- Transcriptomics and proteomics are both powerful, but are used differently: transcriptomics is cheaper and more user friendly than proteomics

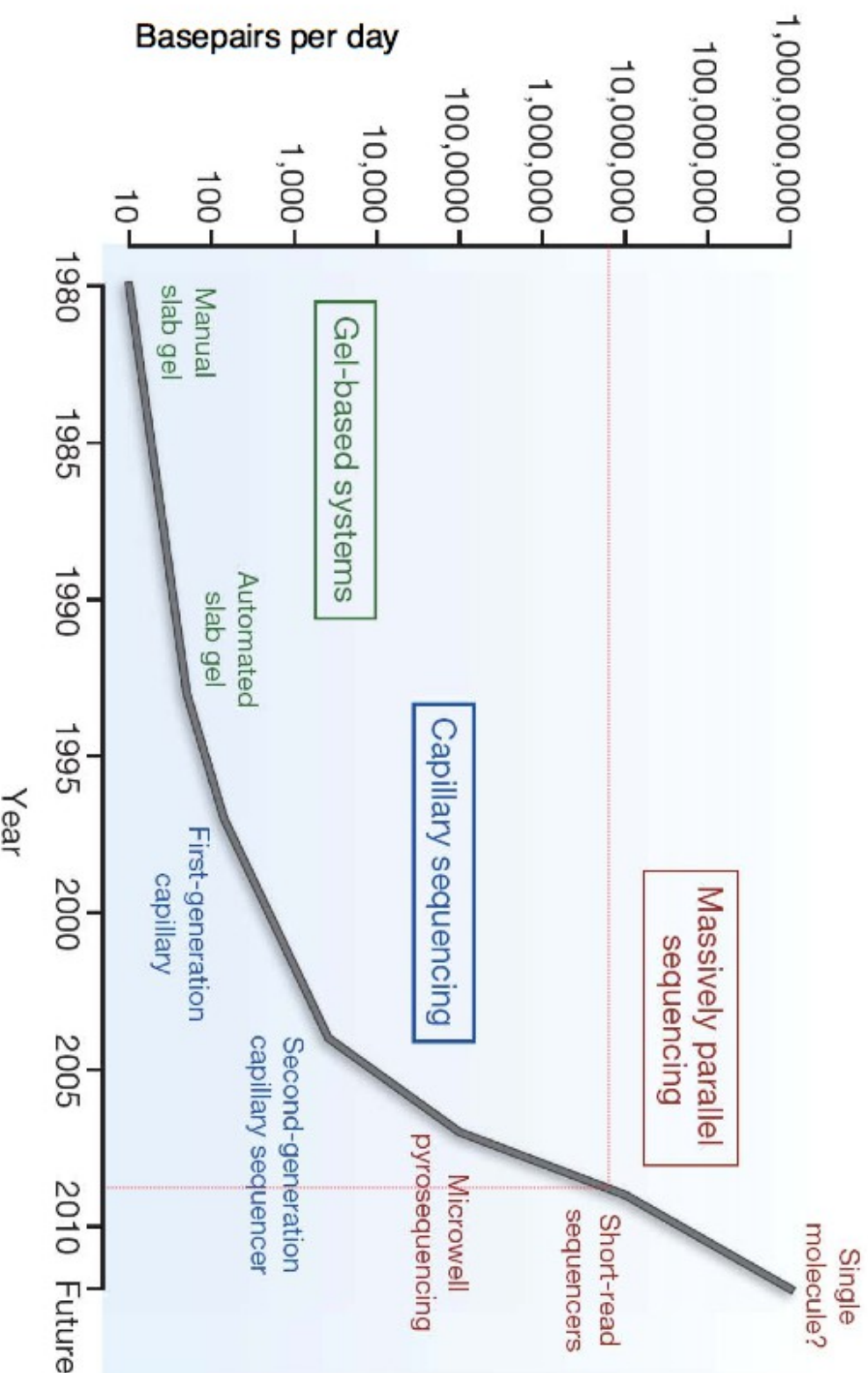
Differences in their practical application:

- Transcriptomics is robust, relatively cost-effective and user-friendly
- Proteomics still relatively limited – problems can remain with purification and stability of proteins

Sequencing technologies

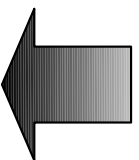
Omics

coincide with dramatic improvements in
different sequencing technologies

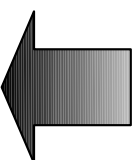


Maxam-Gilbert sequencing

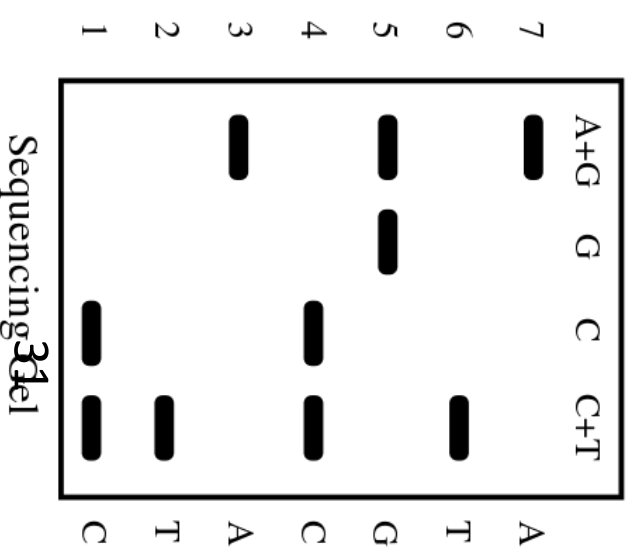
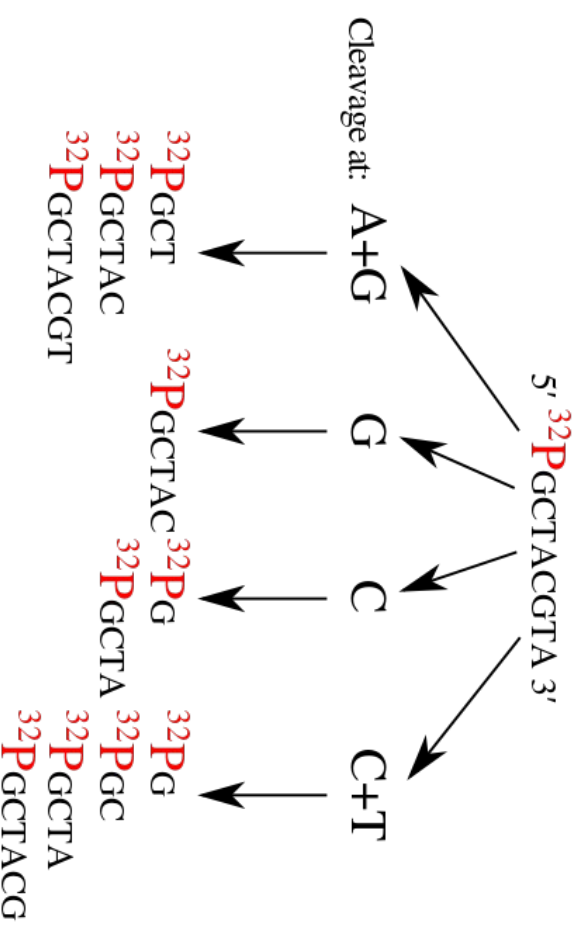
Chemical modification of DNA
Radioactive labelling in 5' end



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.

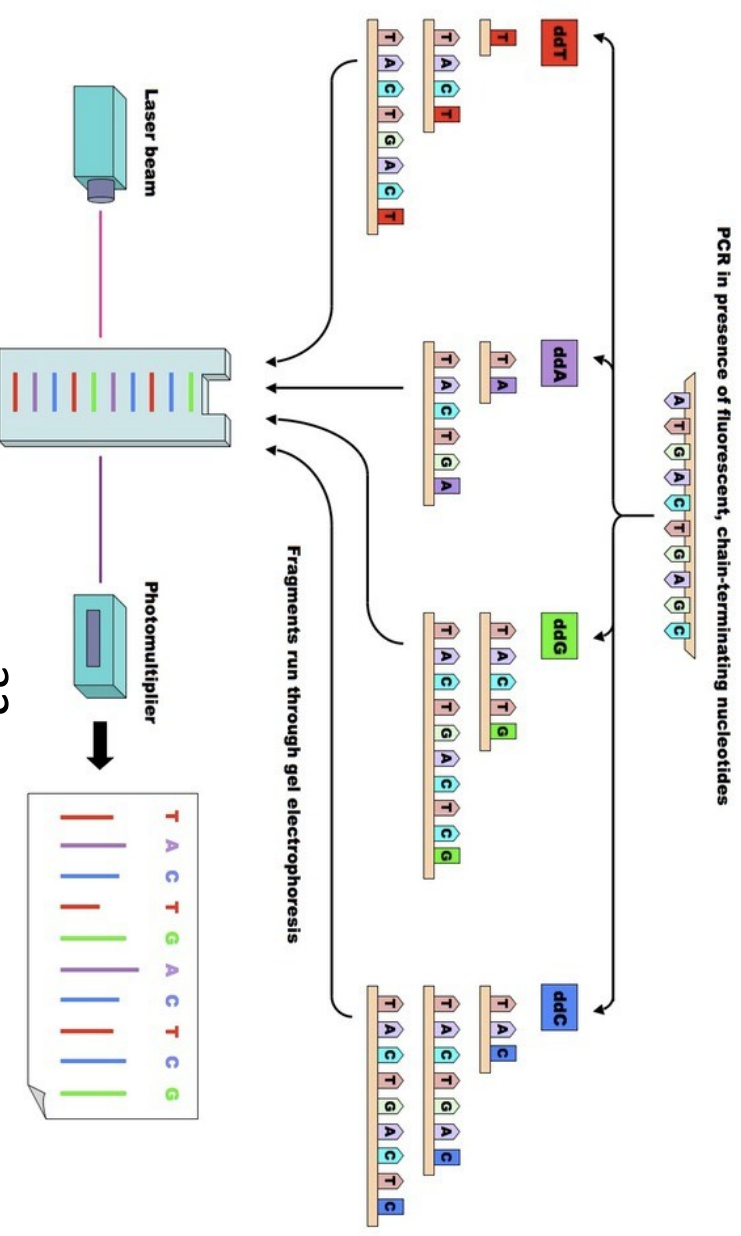


no longer in widespread use, having been
supplanted by next-generation sequencing.

Sanger method

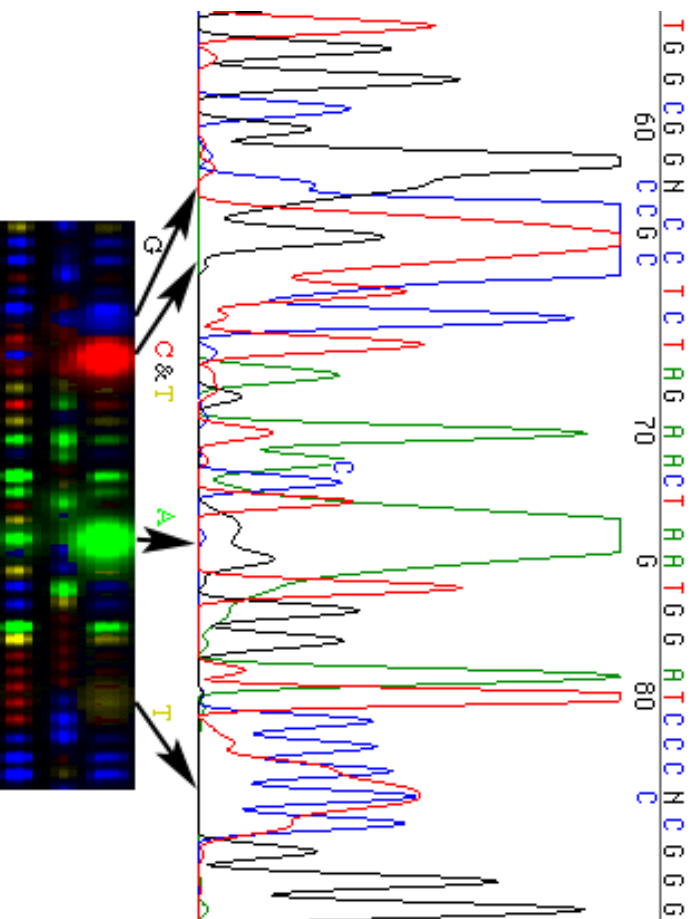
The Sanger method involves four PCR reactions. Each reaction contains the four normal nucleotides plus one dideoxynucleotide stock. As a typical PCR reaction generates over 1 billion DNA molecules, each of the four PCR reactions will generate all of the possible terminating fragments for that particular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



<https://youtu.be/KTstRrDTmWI>

chromatogram



Chromatogram Viewers



4 Peaks
[Mac]



BioEdit
[Windows]



Chromas
[Windows]



Trace Viewer
[Windows / Mac]



Finch TV
[Windows / Mac]



Sequence Scanner
[Windows]

Limitations of Sanger Sequencing

Low throughput
Inconsistent base quality
Expensive
Not quantitative

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)

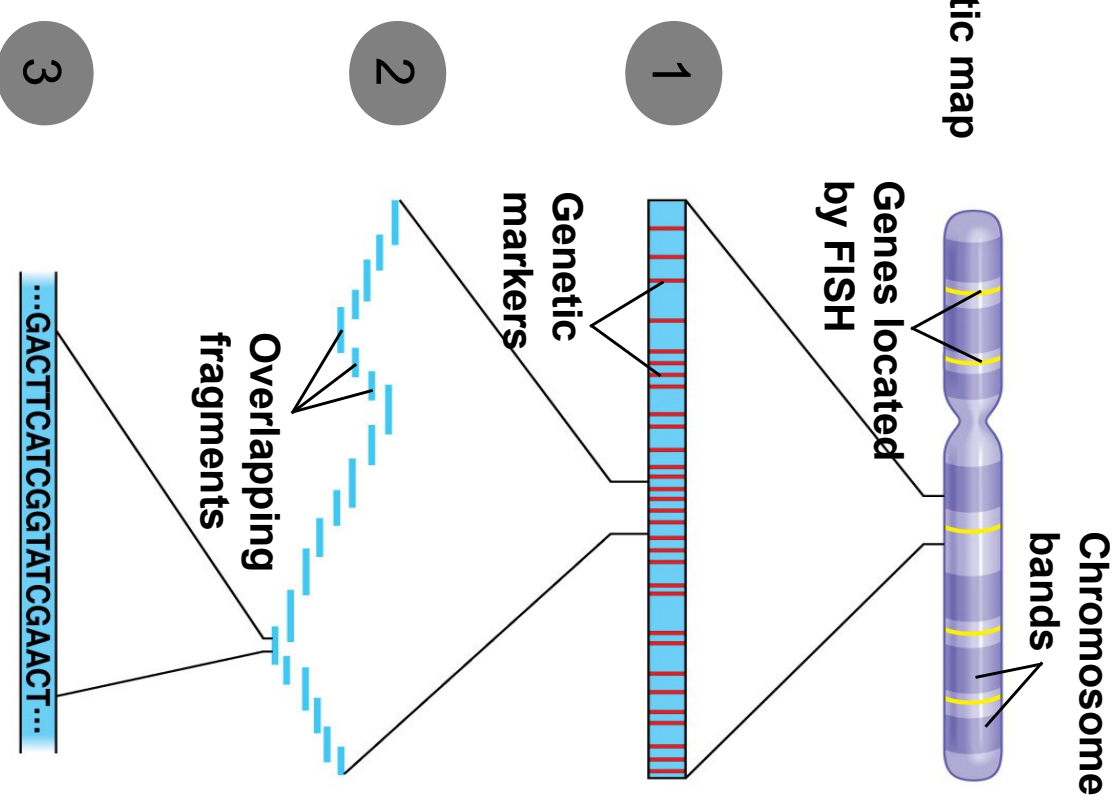
Clone-by-Clone (CBC)

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 \approx \pm 1 mégabase in plants 1 cM \approx \pm 200 kilobases)

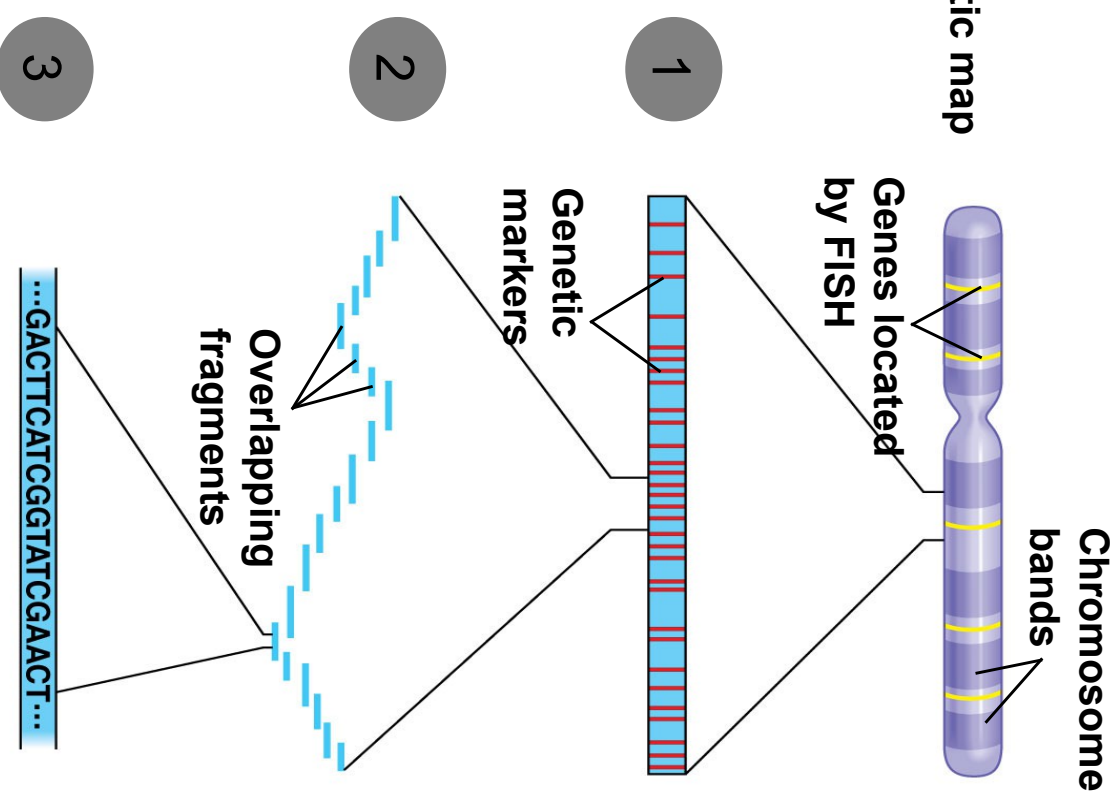


Clone-by-Clone (CBC)

Three-Stage Approach to Genome Sequencing

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

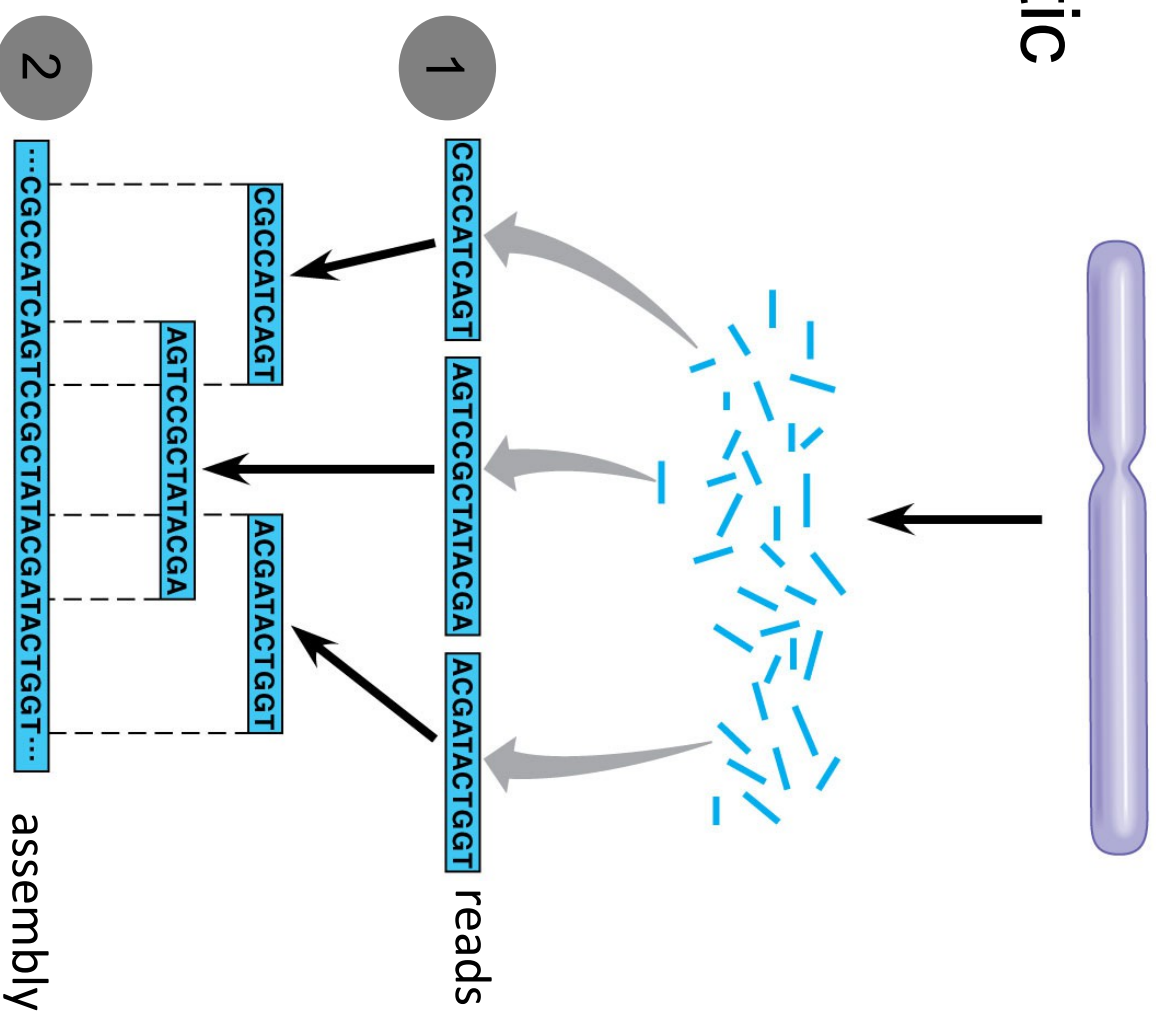
The clones have been arranged to cover an entire chromosome



Whole-Genome Shotgun (WGS)

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

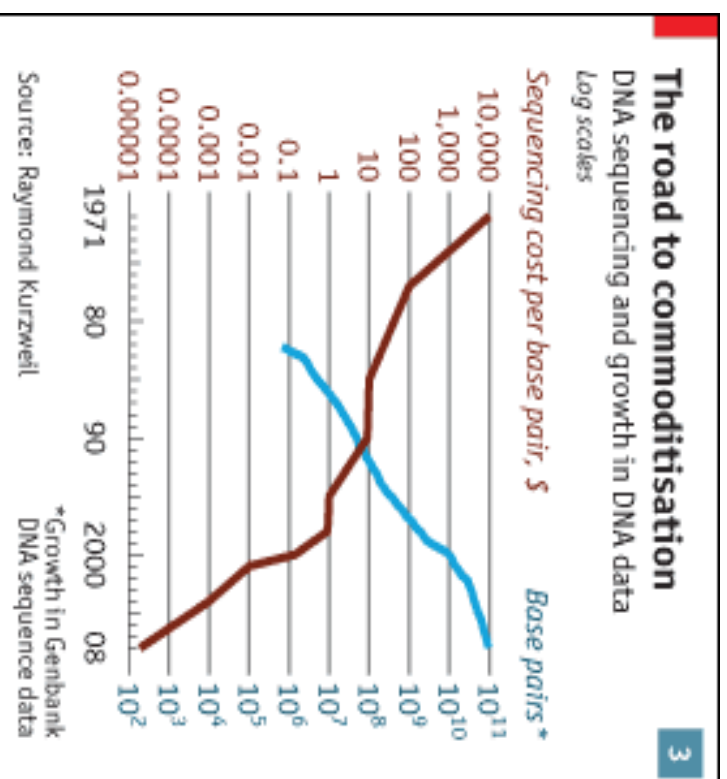
1. DNA sequencing of **random** clones
2. Assembly (order fragments into a continuous sequence)



CBC vs. WGS

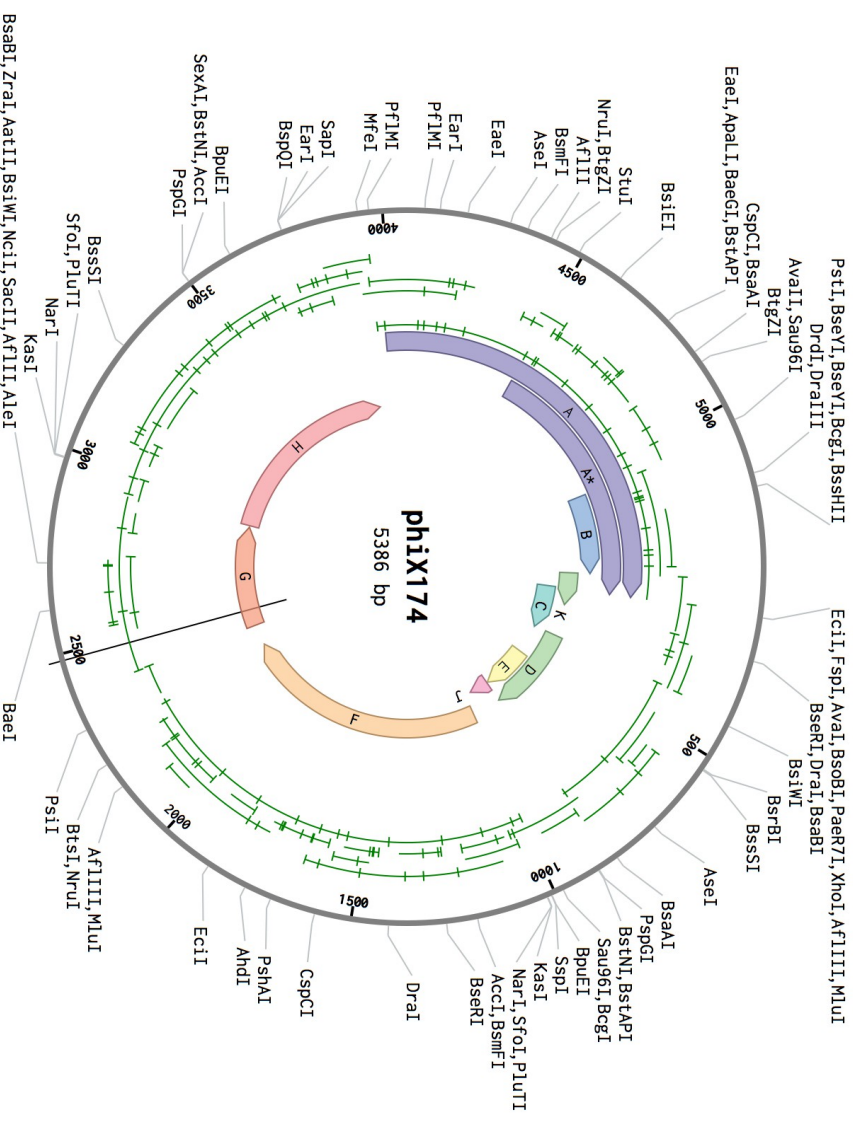
- CBC is time-consuming, expensive and does not allow resolving repeats
- WGS is now widely used as the sequencing method of choice.

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

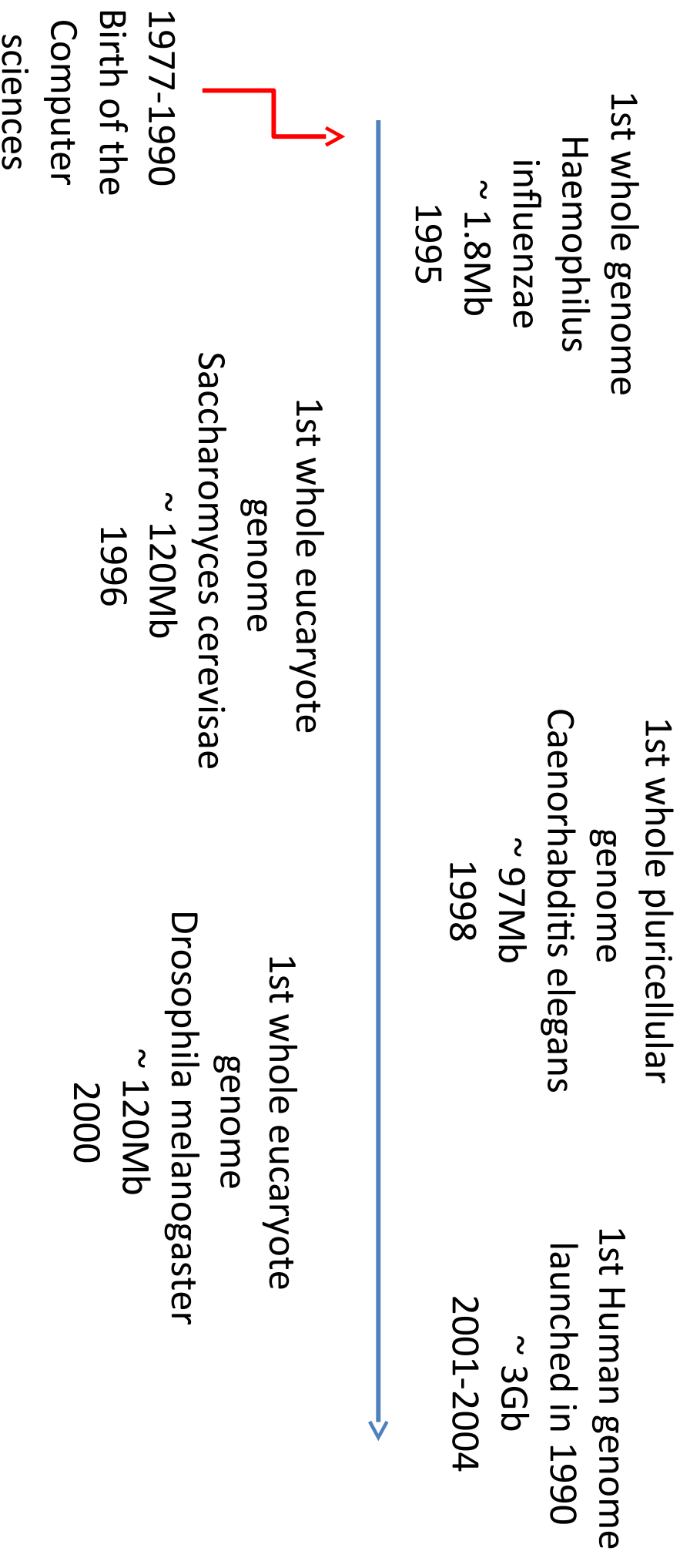


First genome sequenced

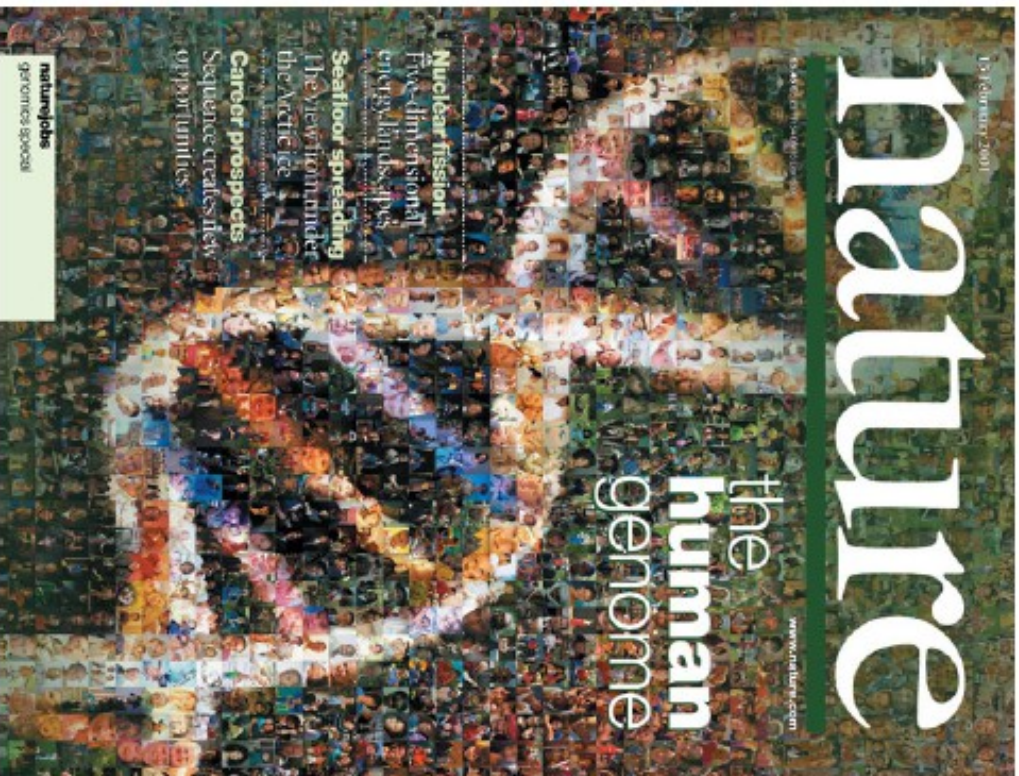
First genome sequenced by
Sanger sequencing (enzyme
synthesis) in 1977
bacteriophage X174 single
strand of 5,375 bp



Genomes sequenced



Human Genome Project... expensive



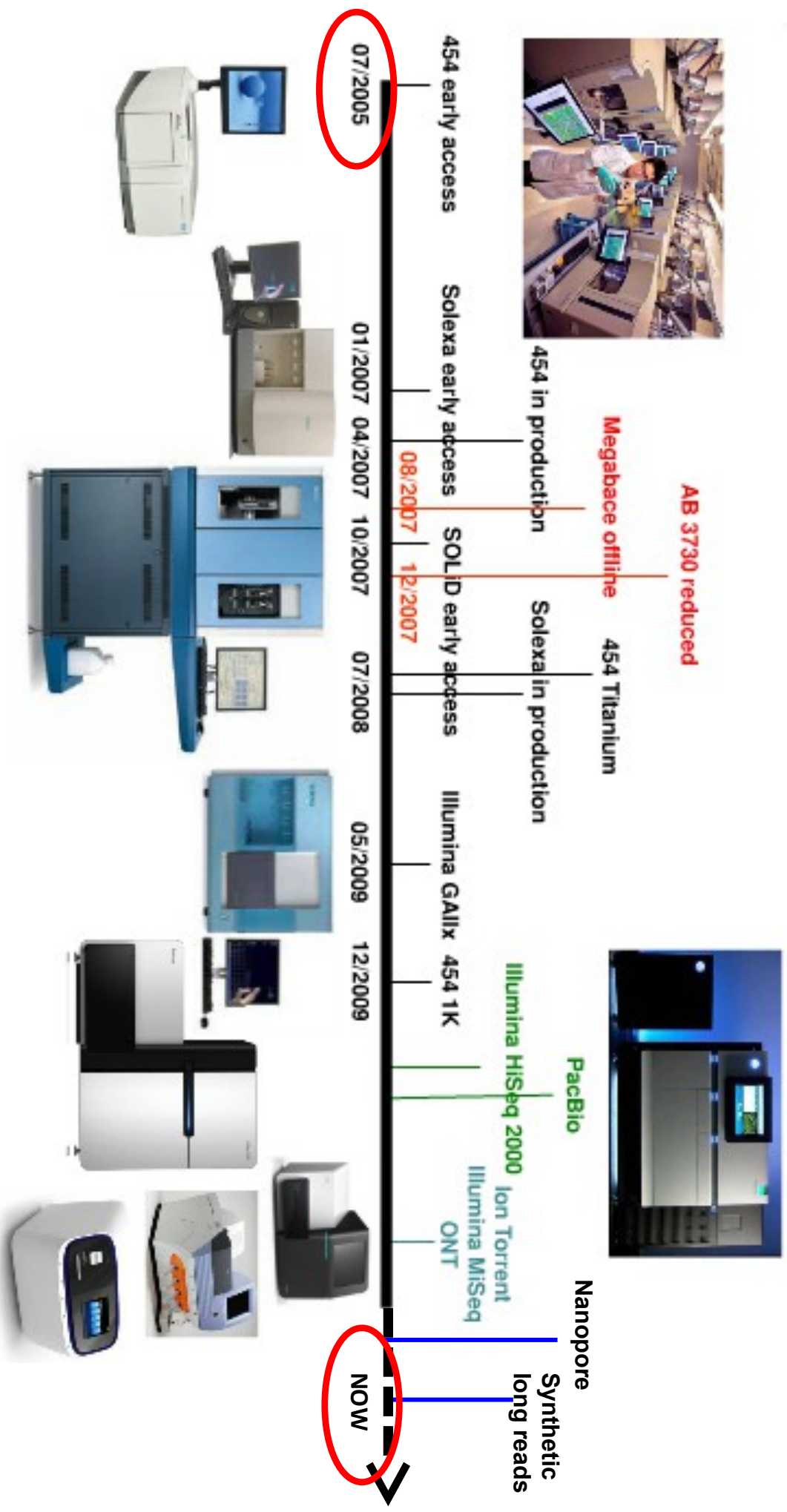
1988 - 2004

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.

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

Short reads



**Long
reads**

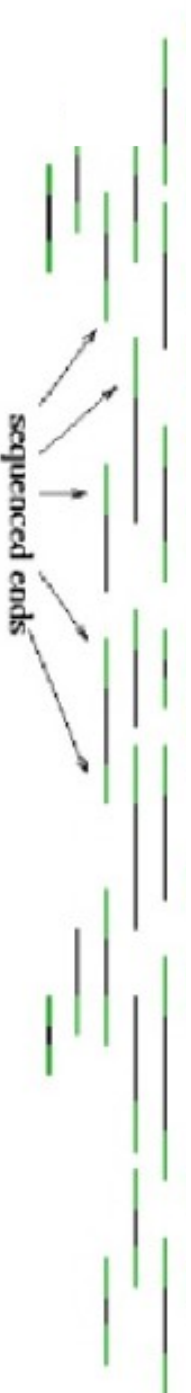


Sequencing ==> reads

DNA Sequence



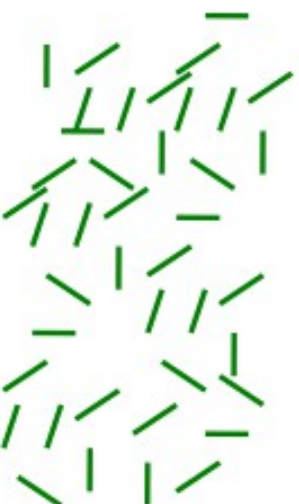
fragments



Size selection

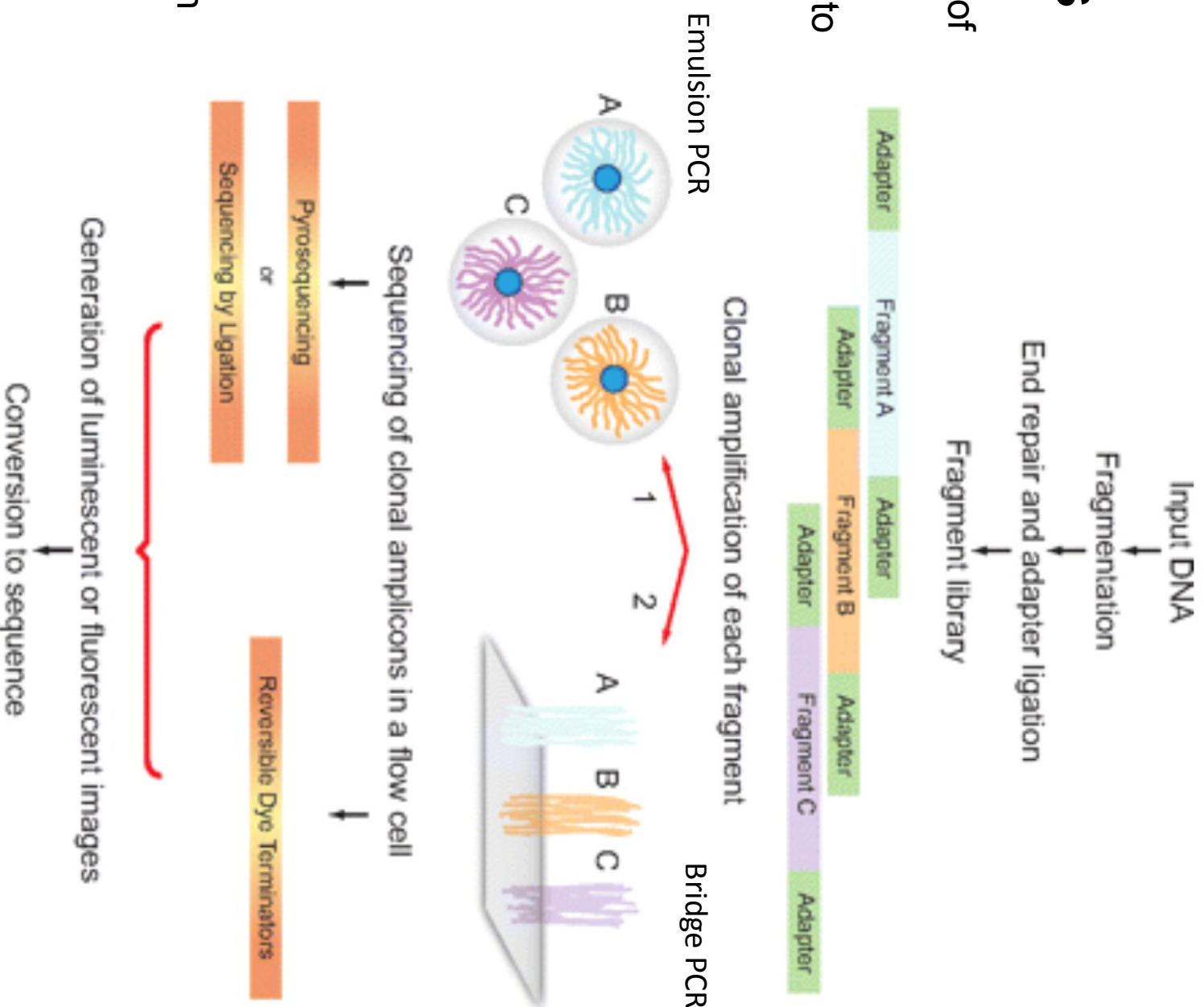
Sequencing

reads

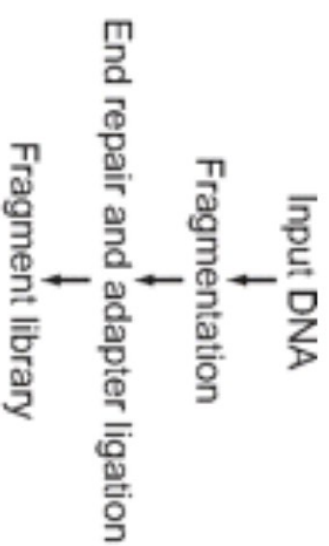


2nd NGS technologies

- 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



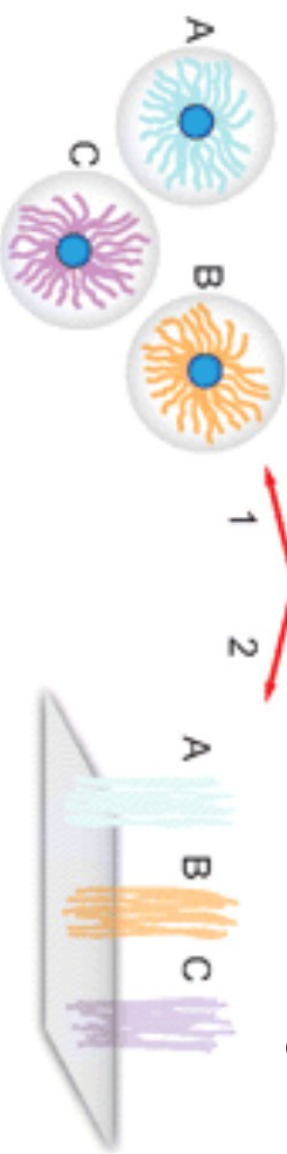
NGS process : illumina



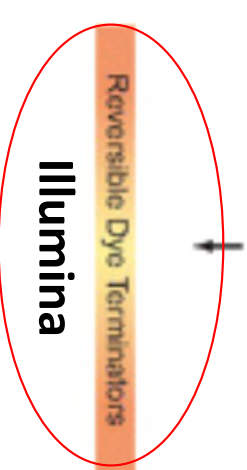
Clonal amplification of each fragment

Emulsion PCR

Bridge PCR



Sequencing of clonal amplicons in a flow cell



Videos :

<https://youtu.be/fCd6B5HRaZ8>

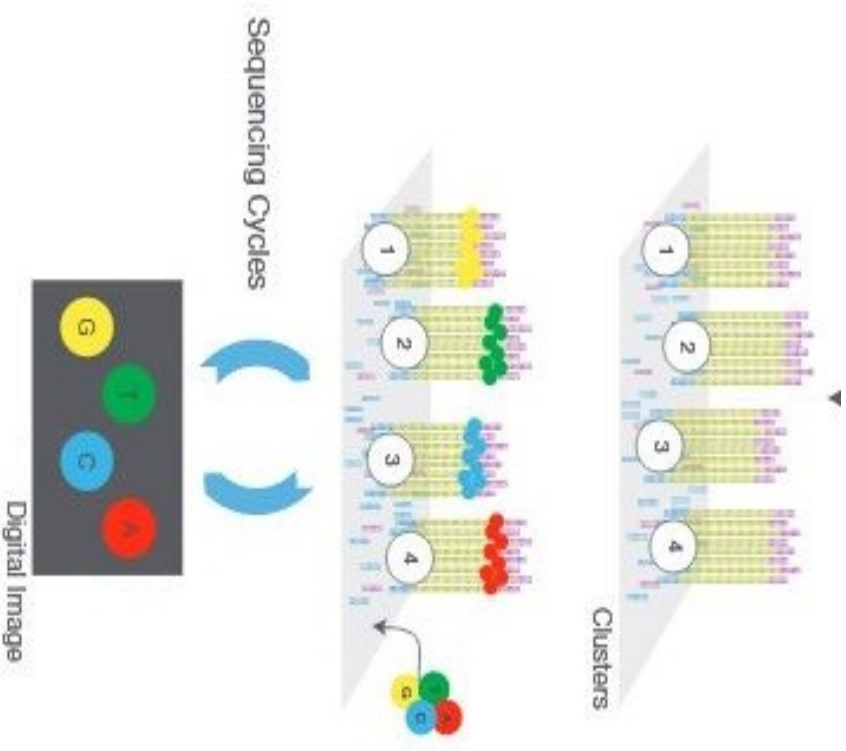
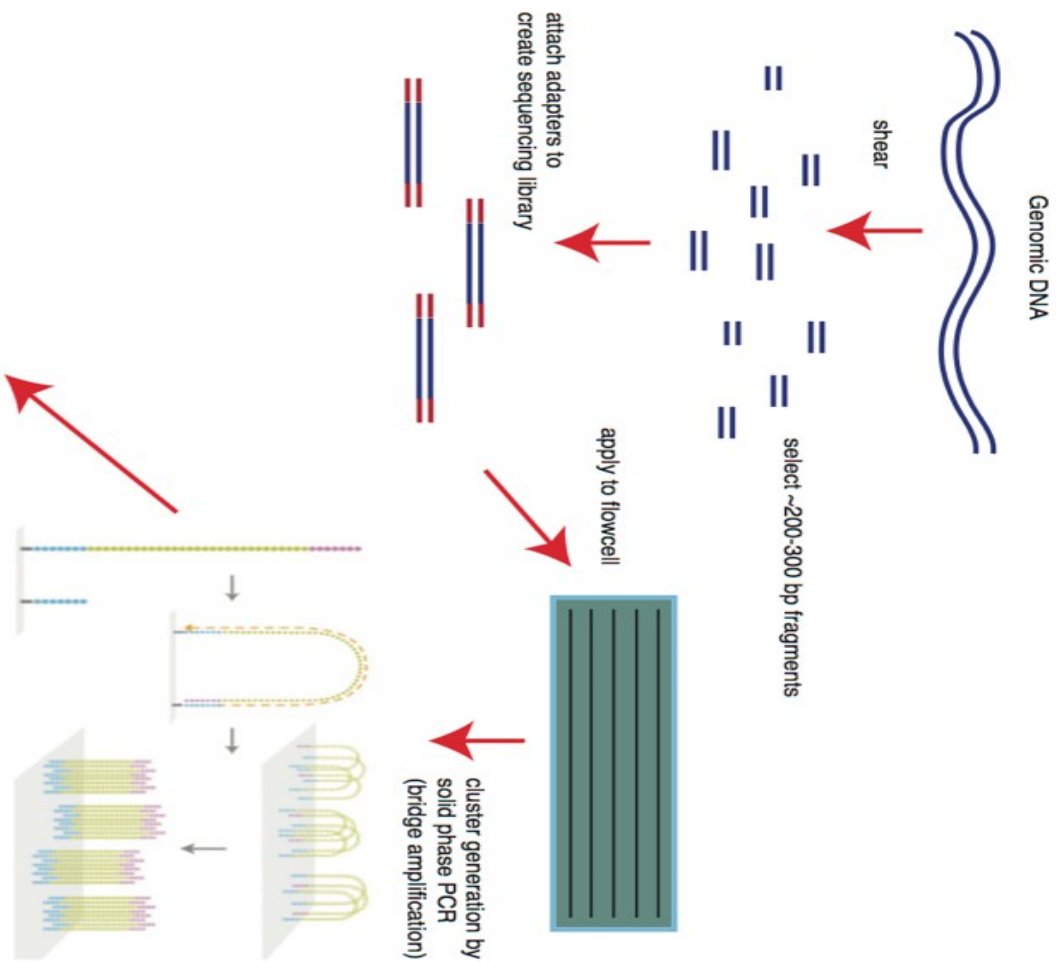
<https://www.youtube.com/watch?v=HMMyCqWhwB8E>

or fluorescent images

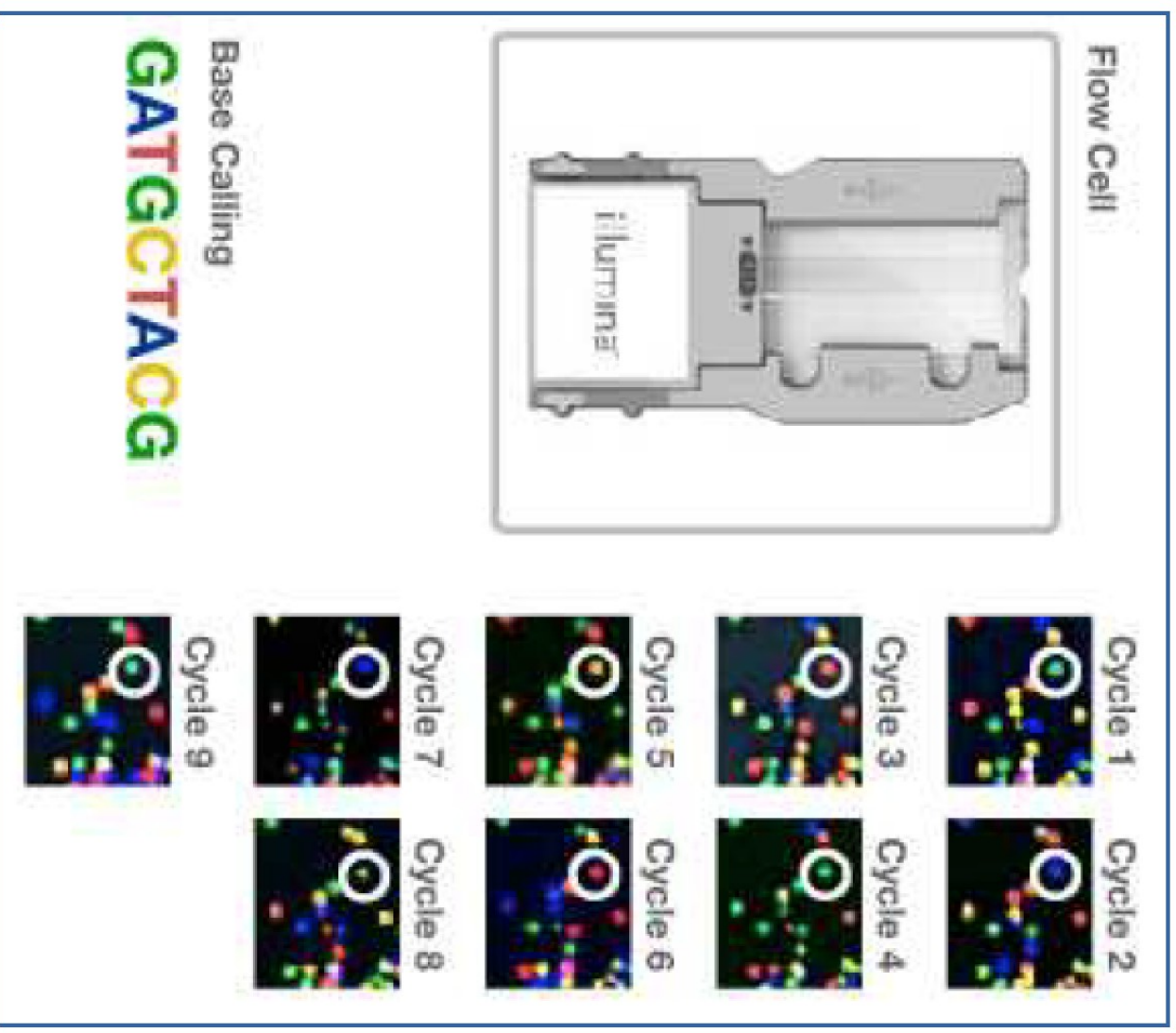
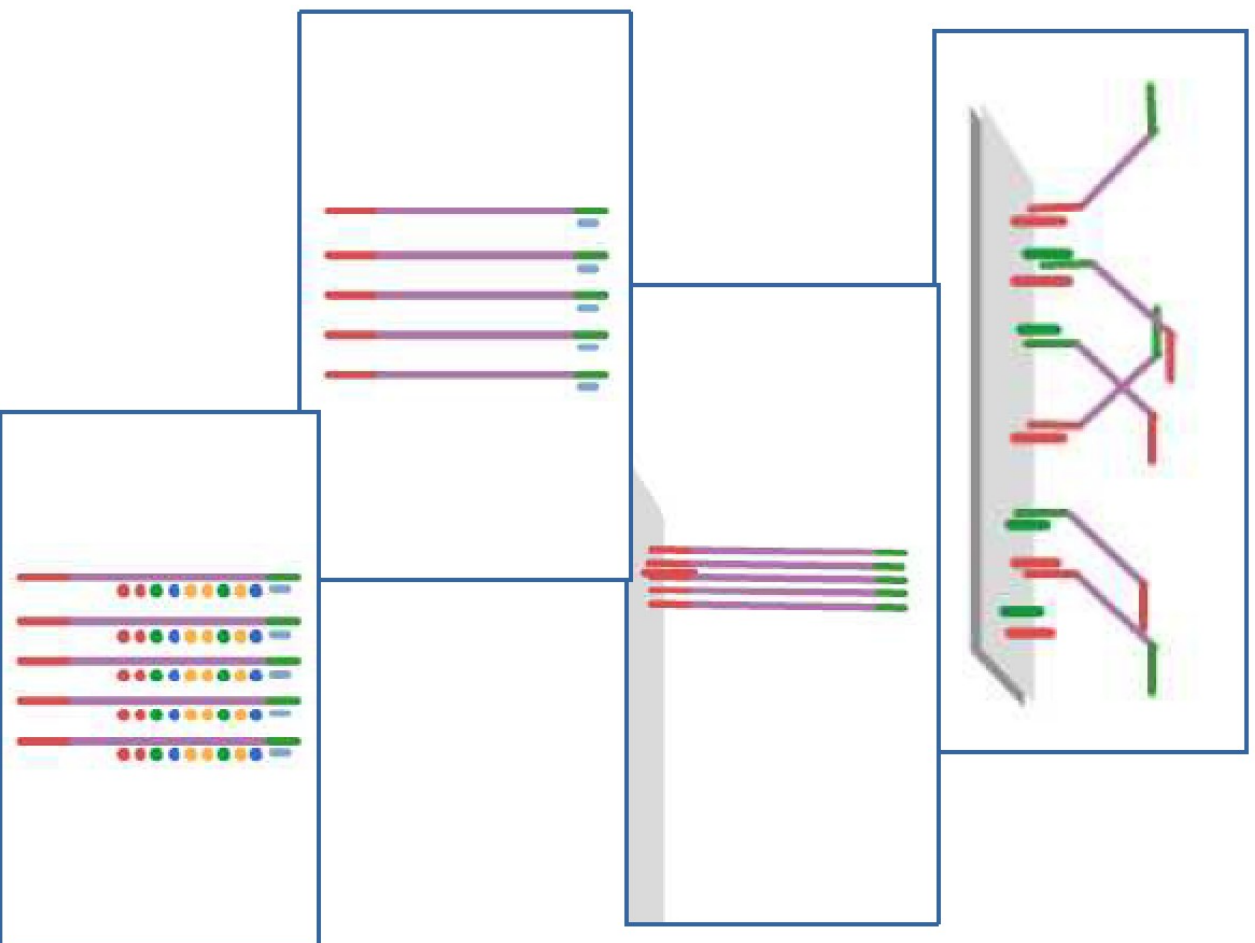
sequence

Sequencing technologies

Illumina



Reversible terminator sequencing - Illumina



Different read types

Single end read

- Only have sequence from one end of fragment

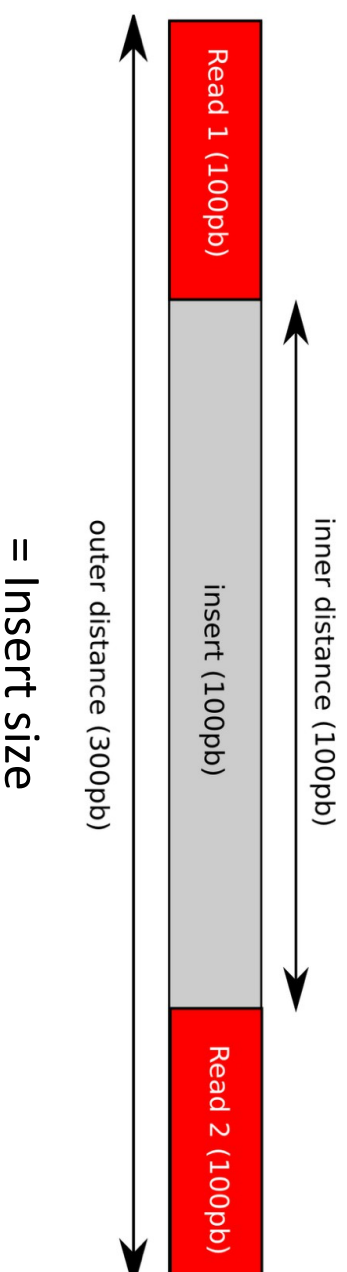


Paired / Mated read

- Have sequence from both ends of fragment



How to define a pair of reads?



	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

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 2^{de} et 3^{ème} génération.

Séquençage de 2^{de} 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.

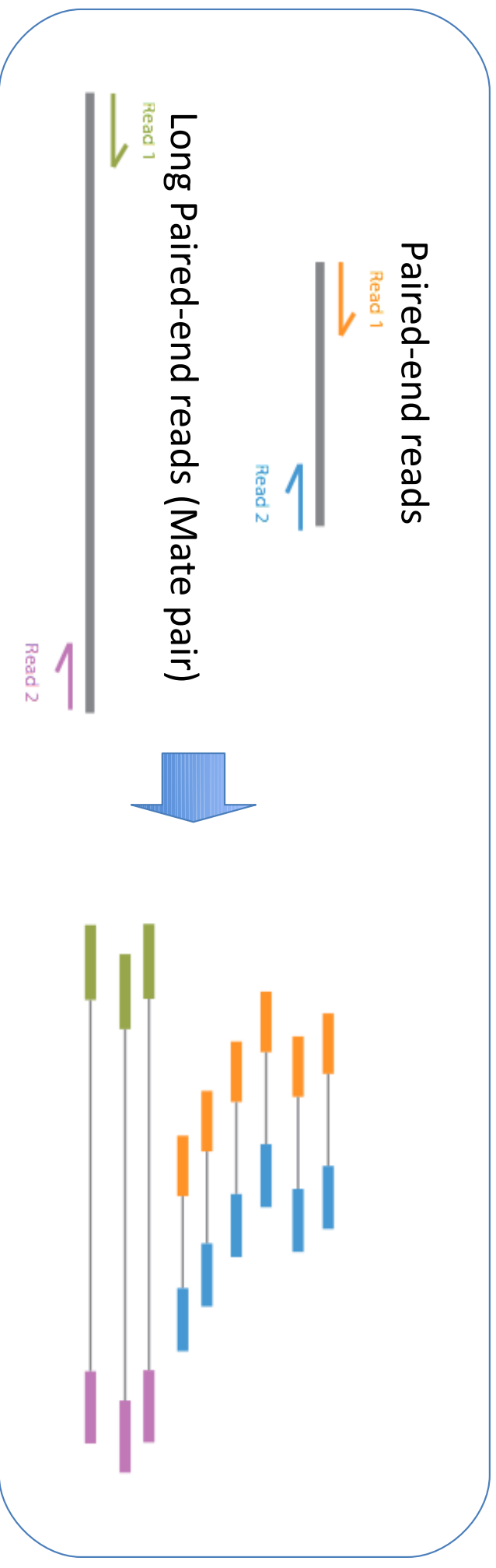
Definitions (3) (in french)

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 »

Short reads



**Sequencing data =
FASTQ file**

Génome de référence

Fasta format

Format of a FASTA definition line

```
>Seq1 [organism=Carpodacus mexicanus] [clone=6b] actin (act) mRNA, partial CDS
CCCTTAICTAATCTTGGAGCAYGAGCTGGCATAAGTTGGAACCGCCCTCAGCCTCCTCATC
```

entête → **○** → séquence

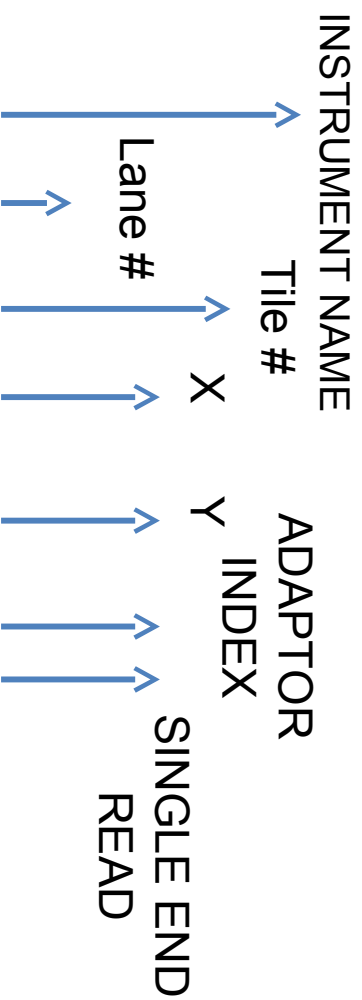
space space space hard return

Fichier multifasta

```
>Sequence_1 assembly1
CCCTAAACCCCTAAACCCCTAAACCCCTGAATCCTTAATCCCTAAATCCCTAAAT
CTTTAAATCCTACATCCATGAATCCCTAAATACCTAATTCCTAAACCCGAAACCGGTTT
CTCTGGTTGAAAAATCATTTGTATATAATGATAATTTTATCGTTTATGTAATGCTTA
TTGTGTGTAGATTTTAAATAATATCATTTGAGGTCATAACAATCCTATTTCTTGT
GGTTTCTTTCCTCAGCTAGCTATGGATGGTTTATCTTCATTTGTATATTGGATACAA
GCTTTGCTACGATCTACATTTGGGAATGTGAGTCTCTTATTTGTAACCTTAGGGTGGTTT
ATCTCAAGAATCTTATTAATTGTTGGACGTTTATGTTGGACATTTATTGTCATTCTT
>Sequence_2
CCCTAAACCCCTAAACCCCTAAACCCCTGAATCCTTAATCCCTAAATCCCTAAAT
CTTTAAATCCTACATCCATGAATCCCTAAATACCTAATTCCTAAACCCGAAACCGGTTT
CTCTGGTTGAAAAATCATTTGTATATAATGATAATTTTATCGTTTATGTAATGCTTA
TTGTGTGTAGATTTTAAATAATATCATTTGAGGTCATAACAATCCTATTTCTTGT
GGTTTCTTTCCTCAGCTAGCTATGGATGGTTTATCTTCATTTGTATATTGGATACAA
GCTTTGCTACGATCTACATTTGGGAATGTGAGTCTCTTATTTGTAACCTTAGGGTGGTTT
ATCTCAAGAATCTTATTAATTGTTGGACGTTTATGTTGGACATTTATTGTCATTCTT
```

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
NAAATTTCACATTGCGTTGGGAACAGTTGGCCCAACTCAGGTTCAGTAACTGTACACAATACCAATCTCCATCAACTTC
AAGAAATGTTCAACAAACACAC
+
@P\cceeeggggiihiiiiiihghhiiiiiiifghhhghfghiiifhiihfhiiiihiggggggeeeeddcdcdcbcdcdccccccc
```

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

[illegible]

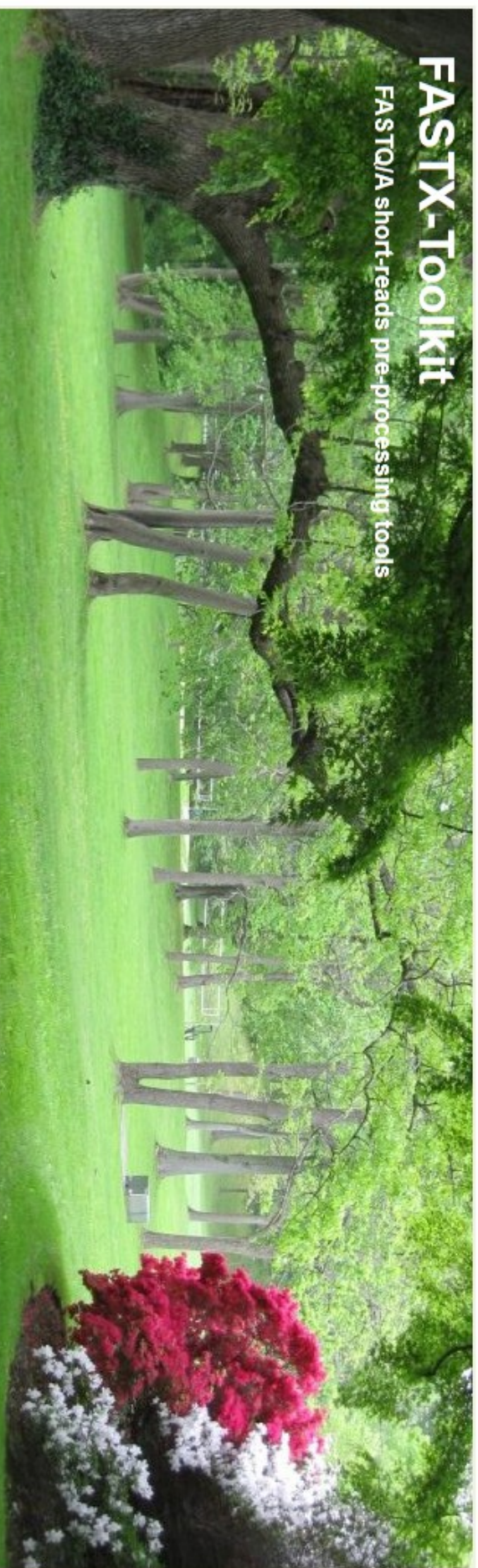
Differing in the format of the sequence identifier and in the valid range of quality scores. See:

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

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

<http://nar.oxfordjournals.org/content/early>

Pre-process: FASTQ data analysis



[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.

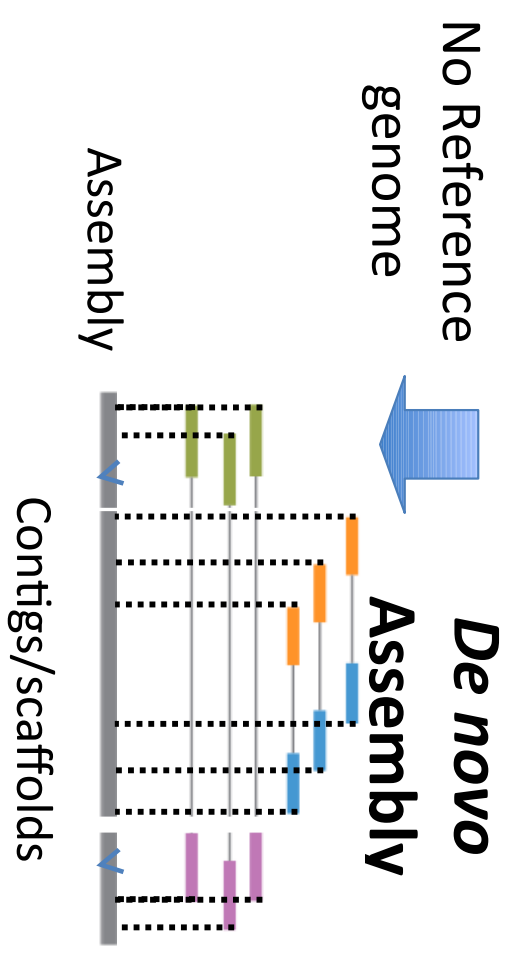
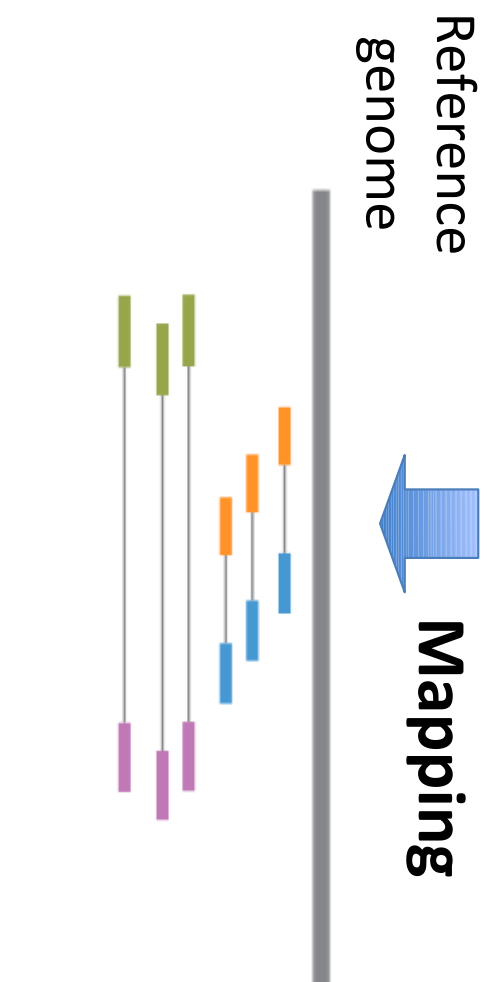
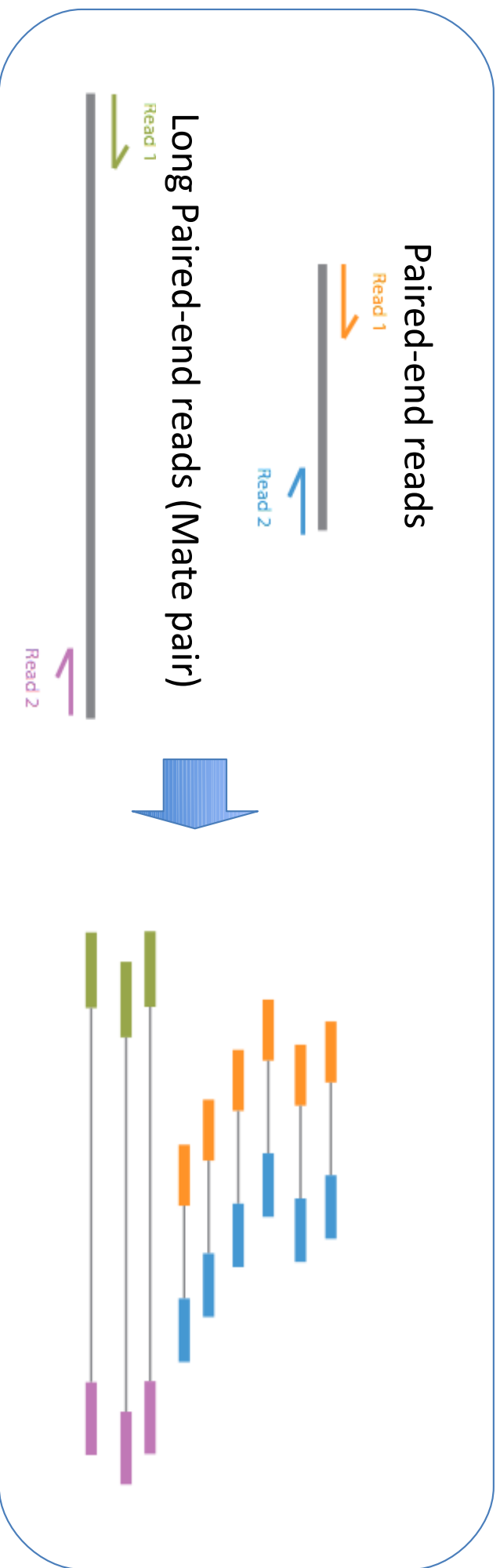
However,

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

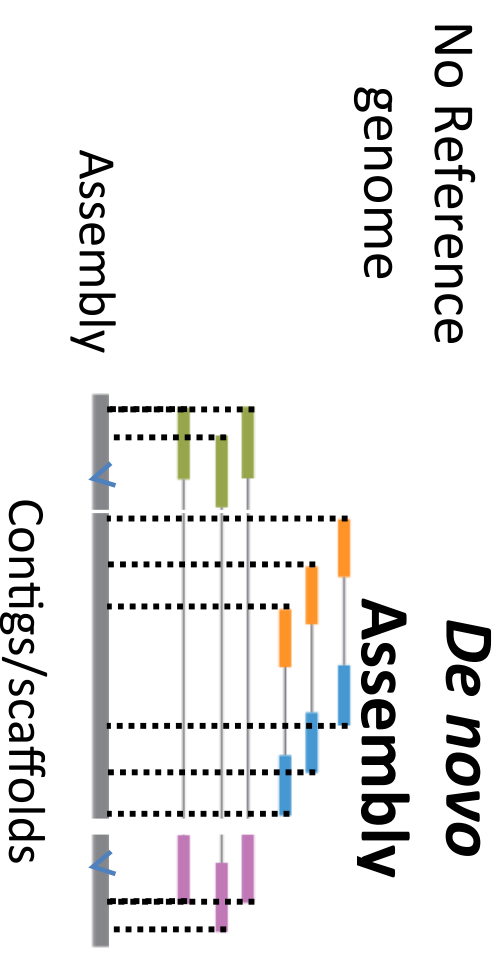
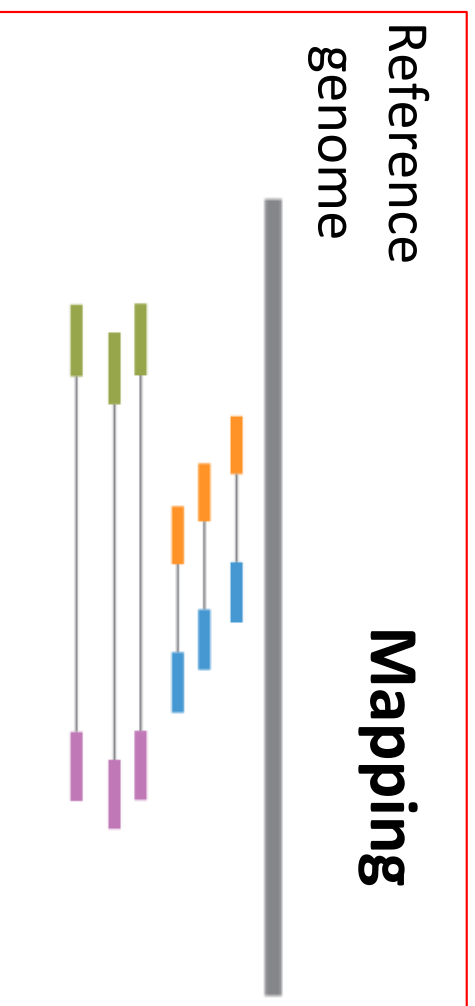
Mapping vs. Assembly



Mapping vs. Assembly

Mapping (re-sequencing):

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



The most famous NGS technologies

Short reads

Illumina - Genome Analyzer IIx (GAIIx), HiSeq2000, HiSeq2500, MiSeq

Long reads

PacBio RS - Pacific Bioscience

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

C2 chemistry – installed March 2012

- Long reads 6-10kb
- Median size of molecules 3kb
- Still 15% error rate
- No strobe sequencing

Software focus on:

De novo assembly

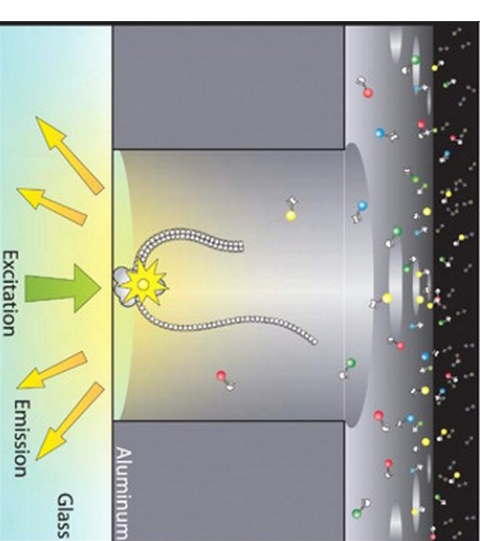
Hi quality CCS consensus reads

In preparation

- Load long molecules by magnetic beads
- Modified nucleotides detection

Videos :

<https://www.youtube.com/watch?v=RcP85JHLmnl>



LS – long sequencing reads

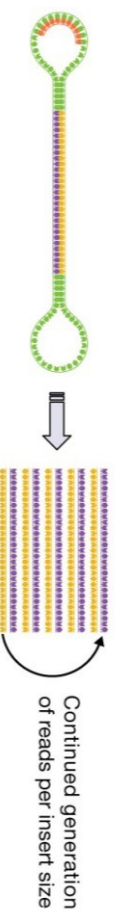


Standard

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

Sam
ple
Prep
arati
on

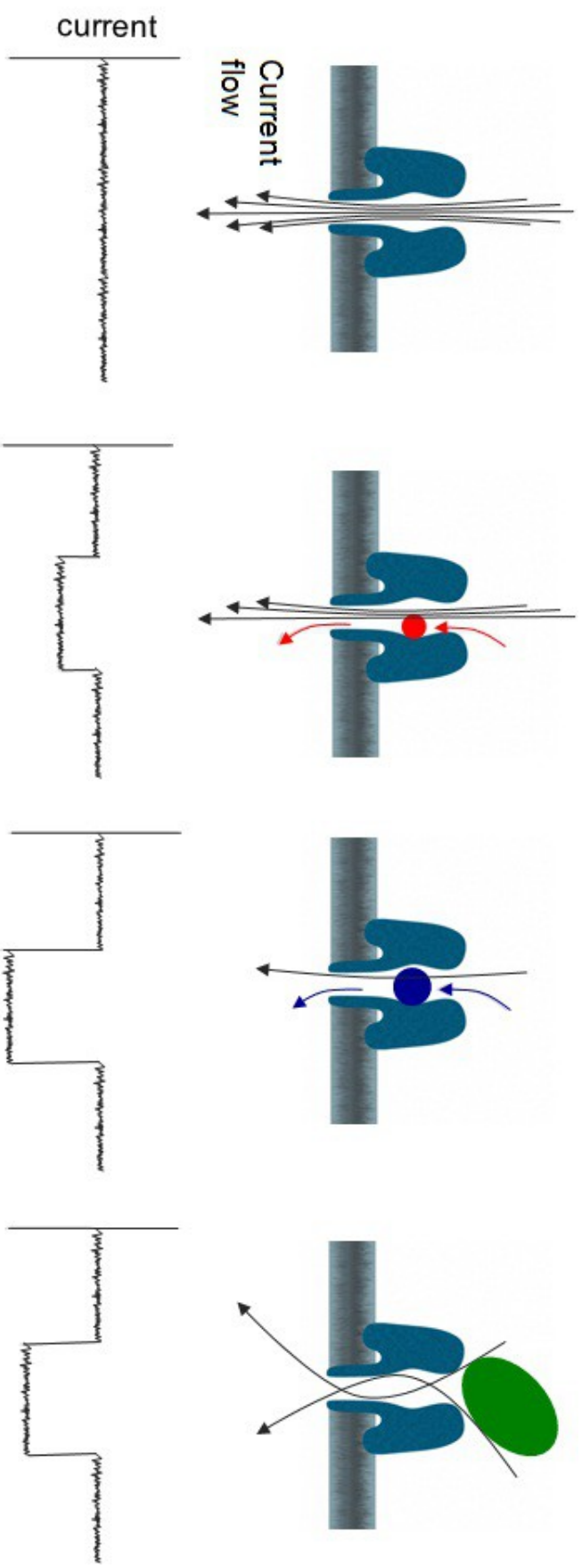
CCS – high quality sequencing reads



**Circular
Consensus**

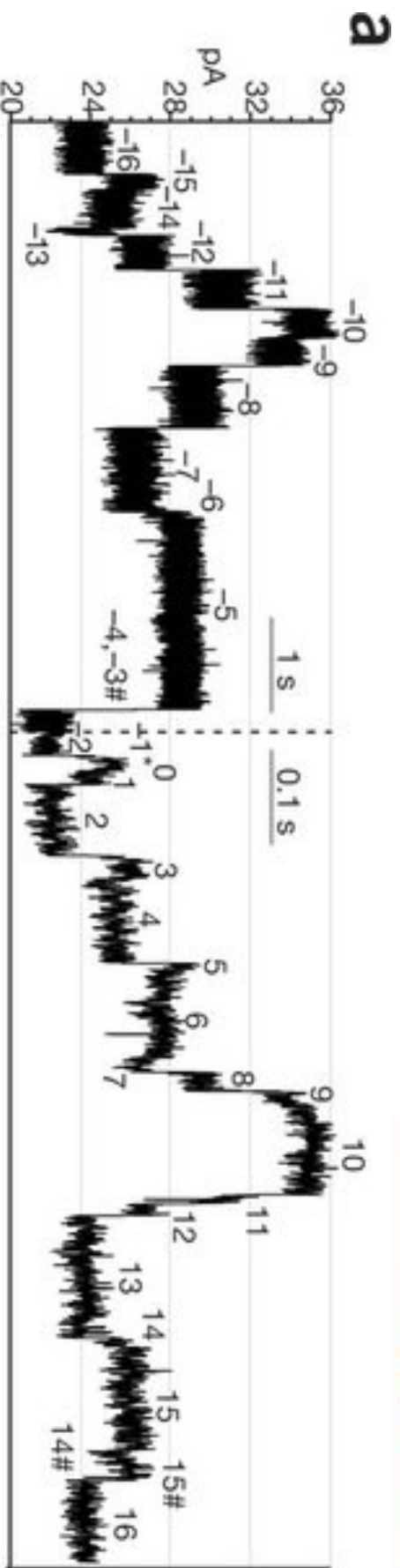
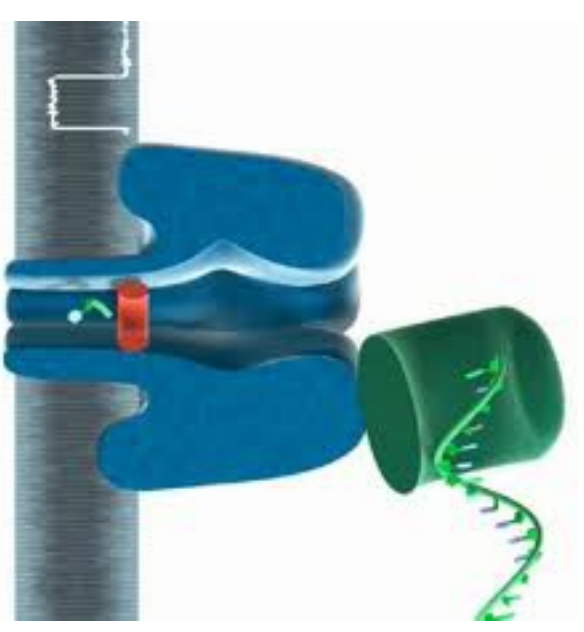
- Small insert sizes 500bp
- Generates multiple passes on each molecule sequenced

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

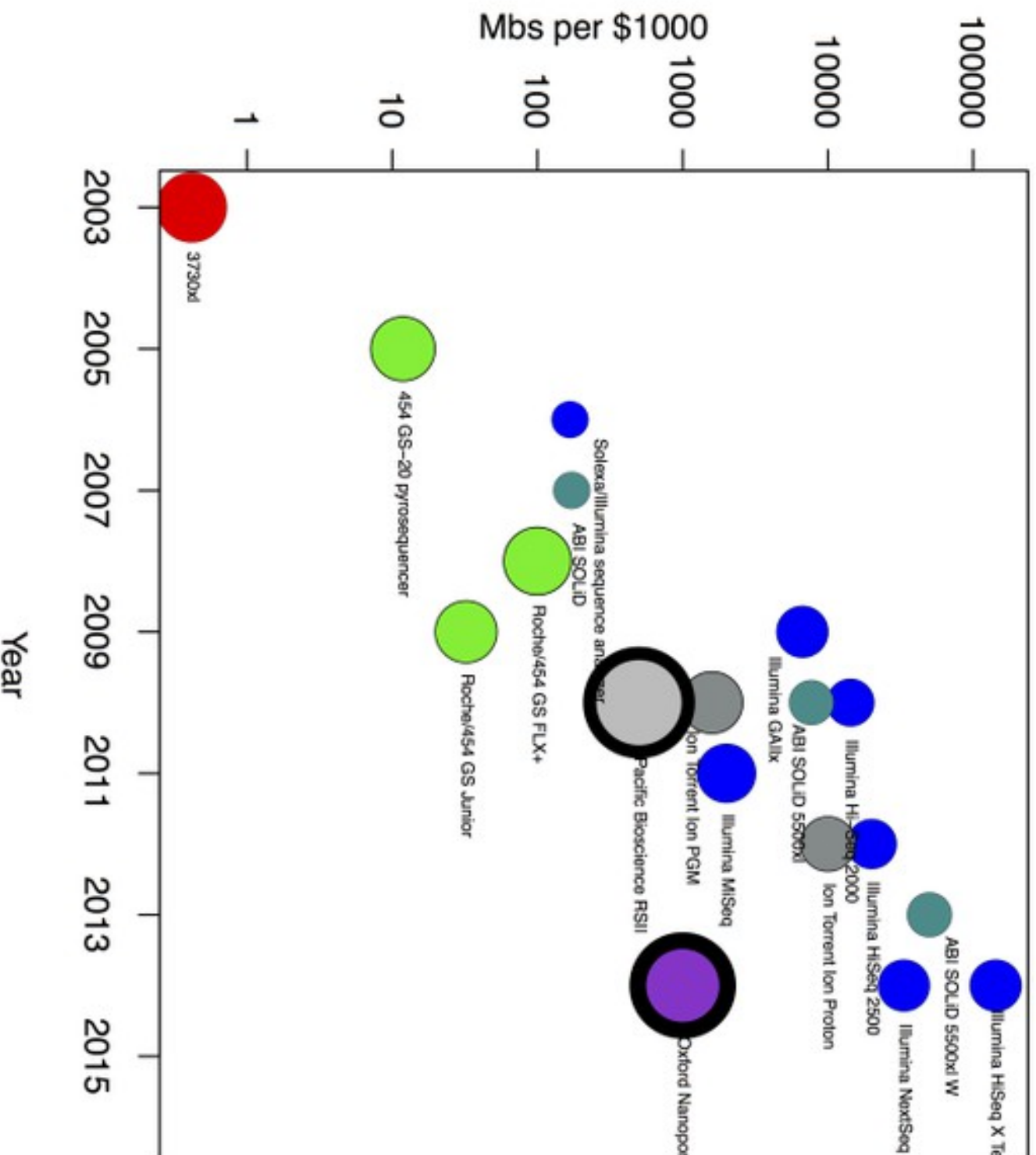
- 150Mb per run
- Tested 48kb read length
- \$900 per instrument
- 500 pores per device



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 (
@atfbd)

NGS technology comparison

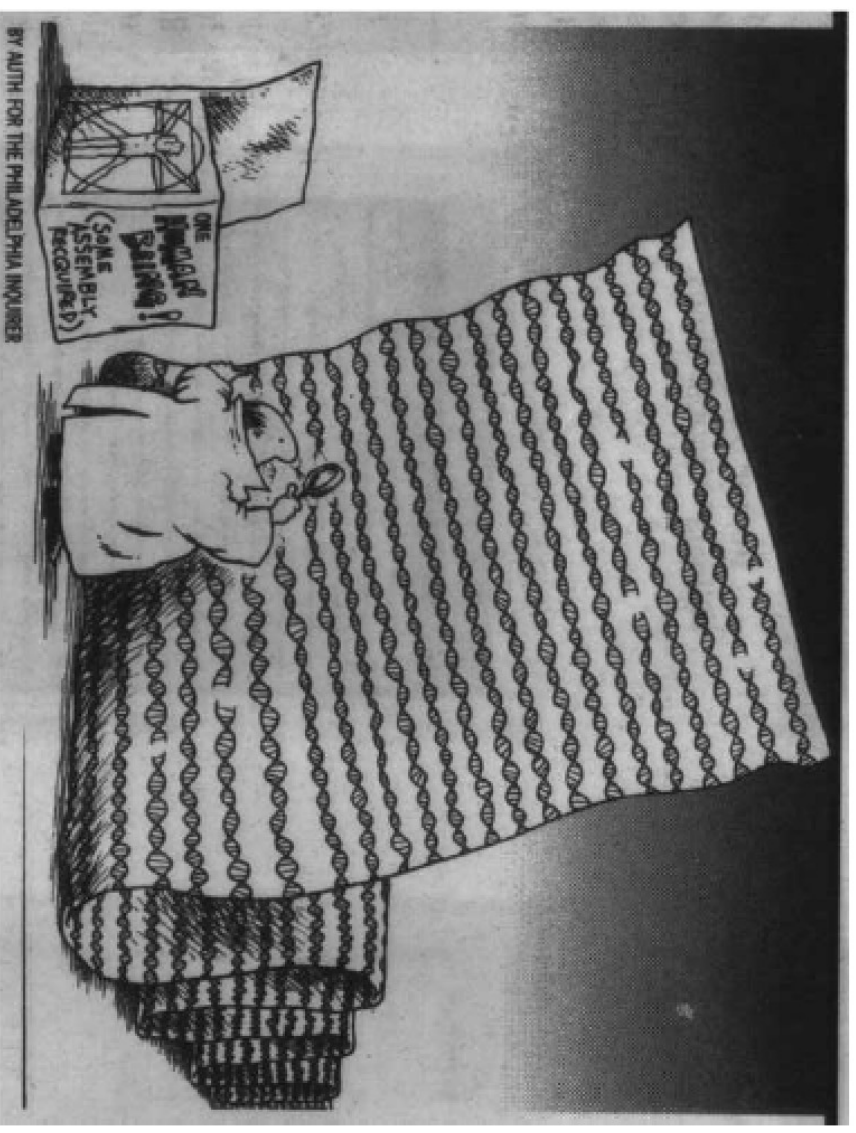
	Capacity	Speed	Read Length
454 Roche	35-700 Mb	10-23 hours	400-700 bp
SOLID	90-180 Gb	7-12 days	75 bp
Illumina*	6-600 Gb	2-14 days	100-250 bp
Ion Torrent	20 Mb- 1Gb	4,5 hours	200 bp
Helicos	35 Gb	8 days	35 bp
PacBio*	1Gb	30 minutes	3000 bp

High DNA quality and quantity
Low sequencing error rate

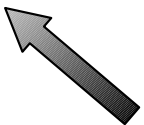
Challenges

“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



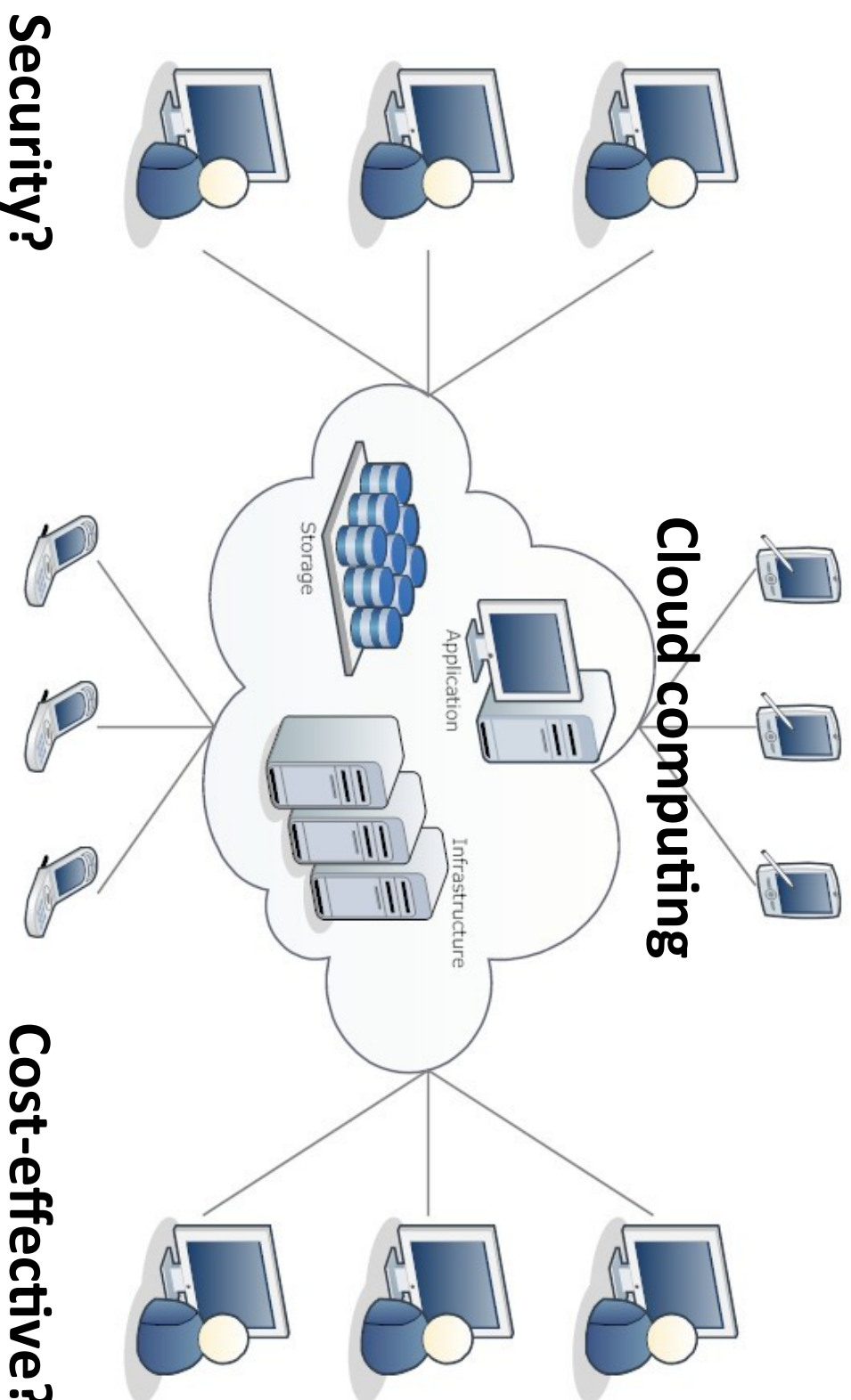
Expensive studies



**We need to deal
with large and
complex data**

How to handle Omics data?

- Tools needed to manage large amounts of data





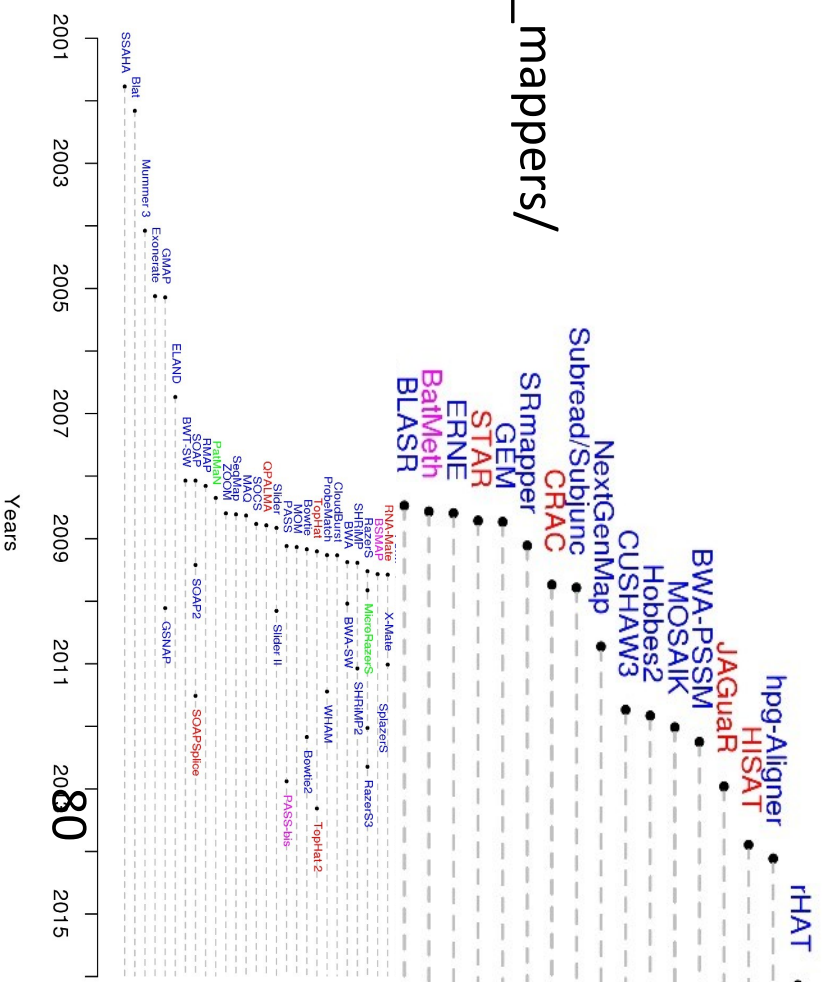
**We need specific
computational
tools that can deal
with these new
data**

How to handle Omics data?

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

Mapping tools

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

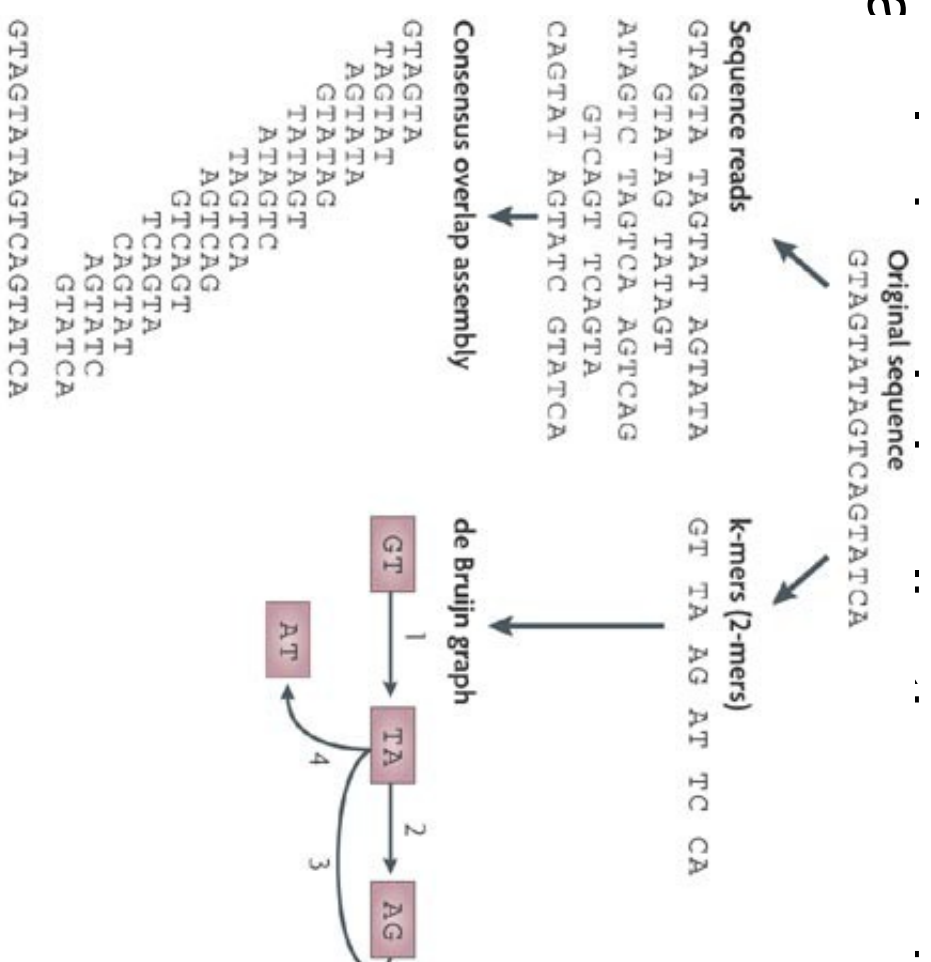


How to handle Omics data?

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

- New methods for ϵ and δ

De bruijn algo



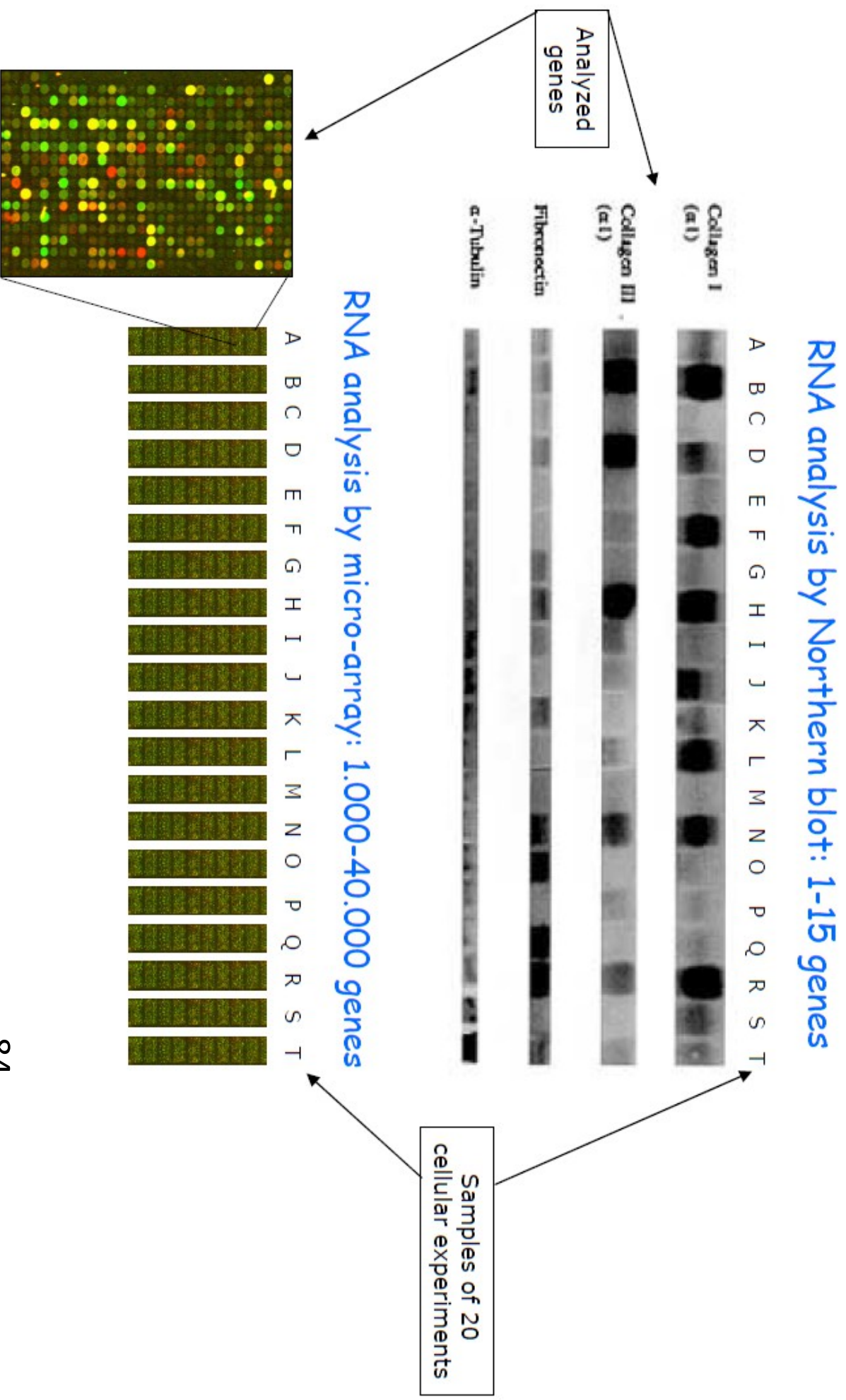
How to handle Omics data?

- 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?



Visualization of NGS Data -

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

**Integrative Genomics Viewer**

[Home](#)
[Downloads](#)
[Documents](#)
→ Hosted Genomes
→ FAQ
⊕ IGV User Guide
⊕ File Formats
⊕ Release Notes
→ Credits
[@ Contact](#)

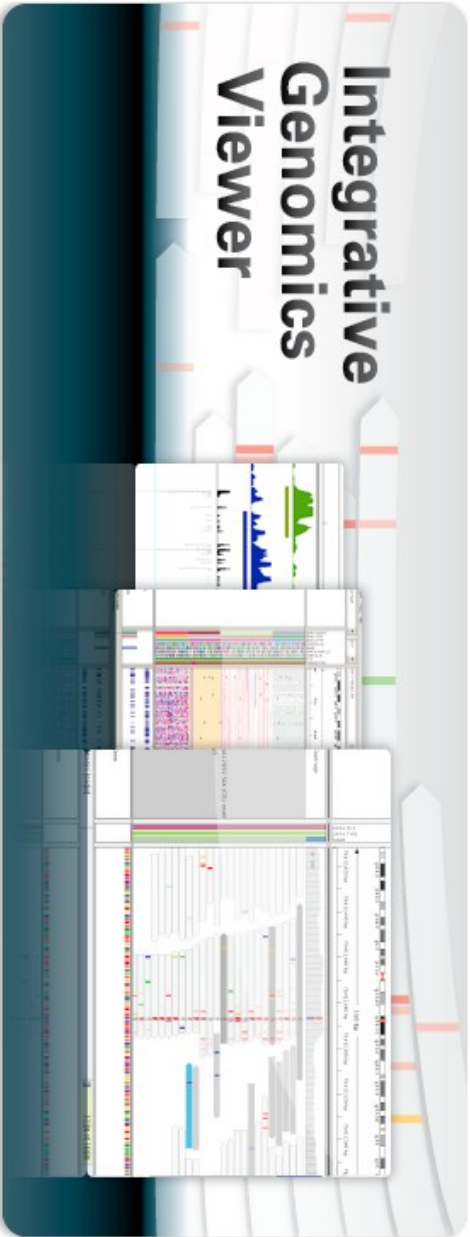
Search website

search


[Broad Home](#)
[Cancer Program](#)
**BROAD INSTITUTE**
© 2011 Broad Institute

Home

Integrative Genomics Viewer



What's New


**NEWS**
Get Latest Data

October 17, 2011. Data from the [29 mammals paper](#) is now available from the hg18 "Load from server..." menu.

September 24, 2011. IGV 2.0.10 is released with bug fixes and support for loading indexed fasta files. See [notes](#) for more details.


May 31, 2011. IGV version 2.0 is now available on the [downloads](#) page. See the IGV 2.0 [Feature Guide](#) for an overview of changes and new features.

[More...](#)

 [Subscribe](#)

Overview

Downloads



Please [register](#) to download IGV. After registering, you can log in at any time using your email address. Permission to use IGV is granted under the GNU [LGPL](#) license.

Citation

To cite your use of IGV in your publication:

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. [Integrative Genomics Viewer](#). *Nature Biotechnology* 29, 24-26 (2011)

Funding

Development of IGV is made possible by funding from the

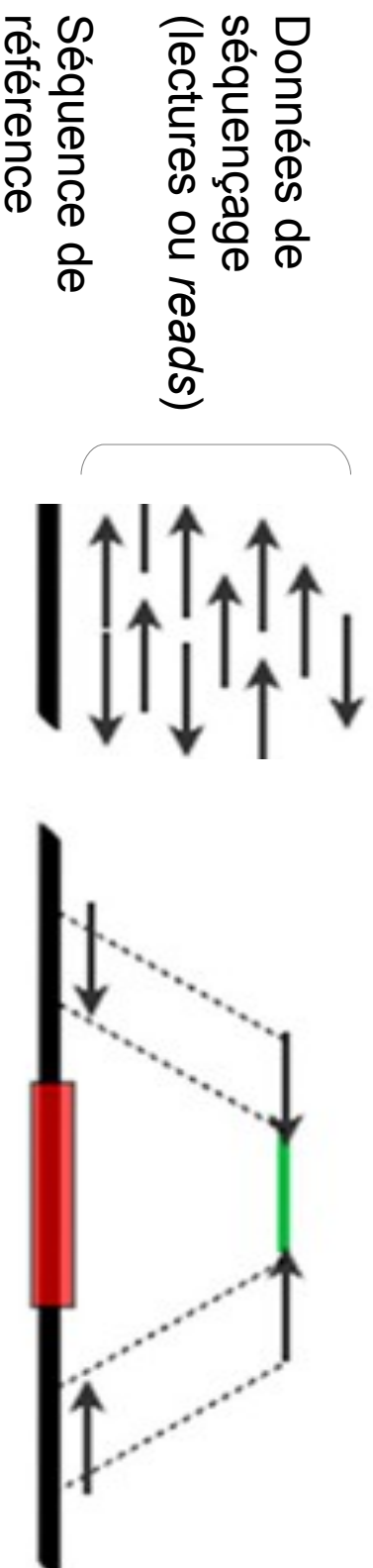
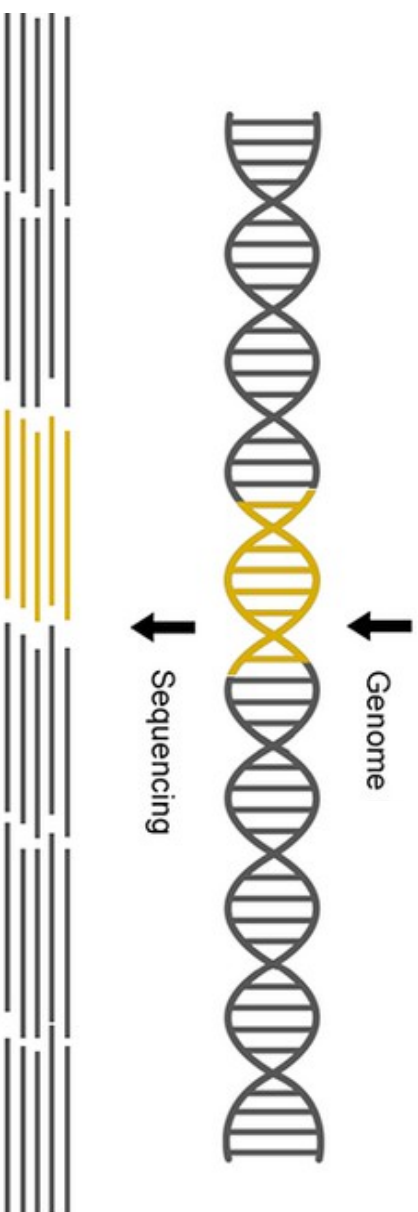
Application examples

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 disulfite-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

The Future of Omic Research

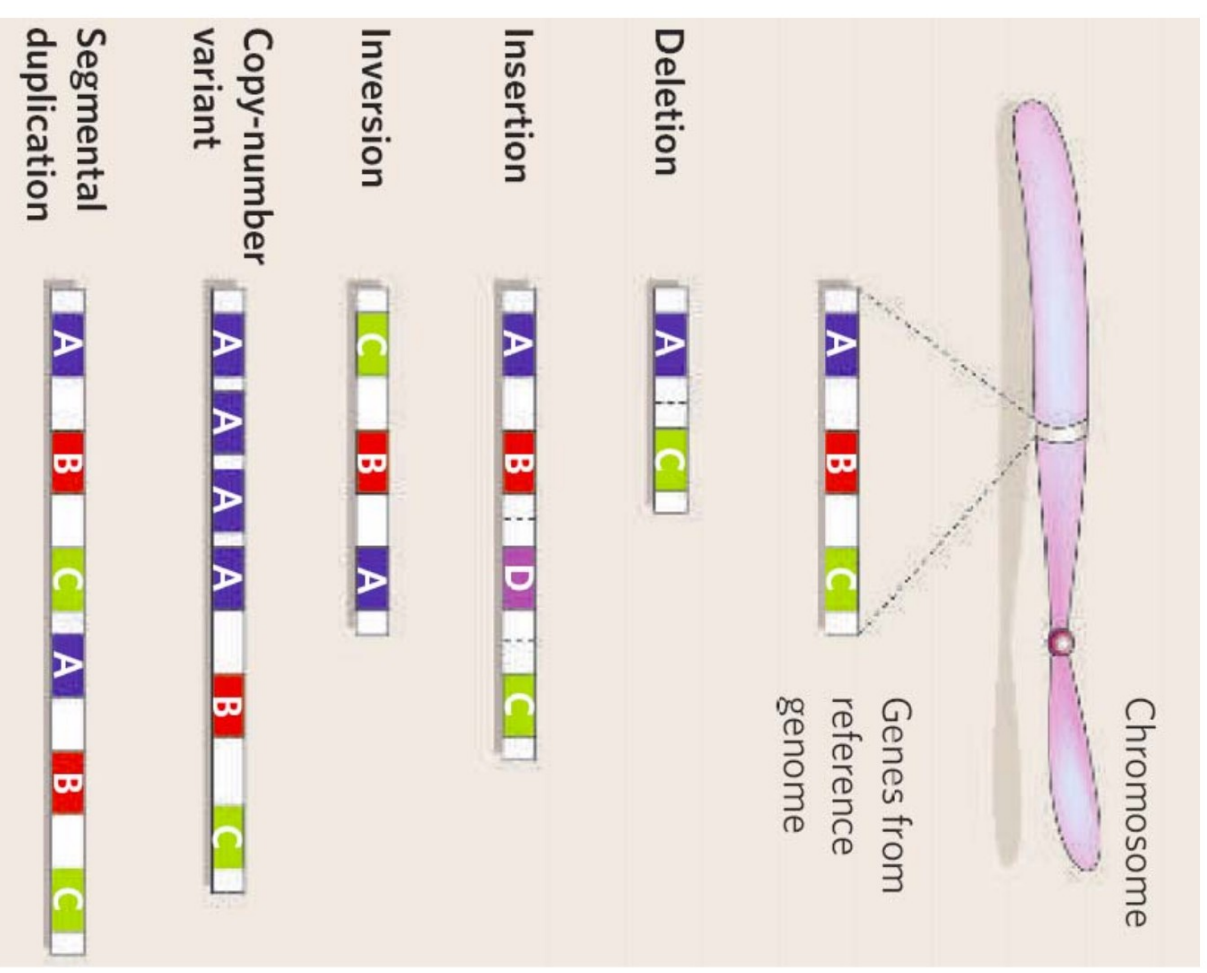
- Six fields were targeted for development as **Omic** information grows
 - Resources: Genome sequences and libraries/DB
 - Technology such as new sequencing methods
 - Software for computational biology
 - Training professionals in interdisciplinary skills
 - Ethical, legal, and social implications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

“Mapping”

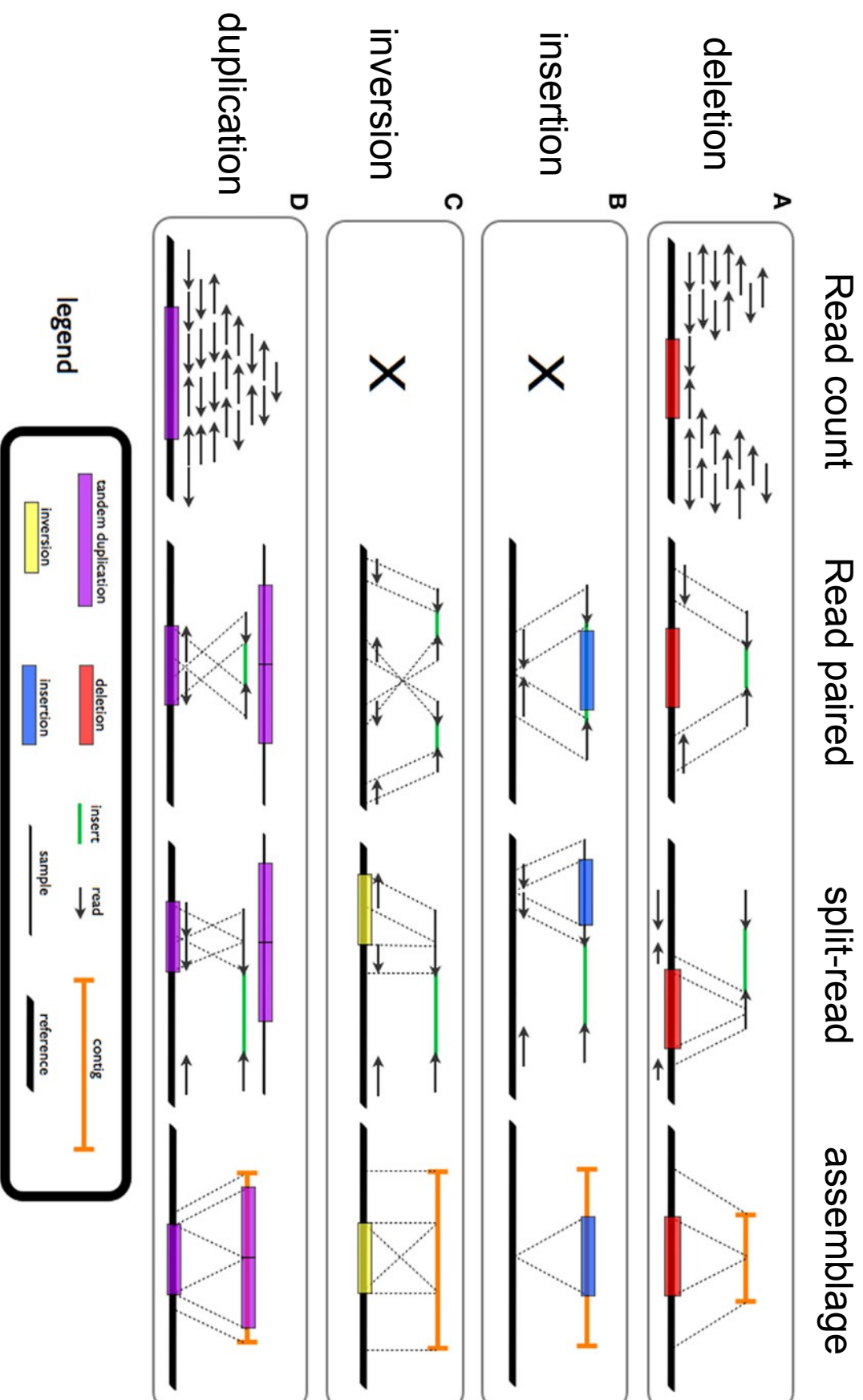


Structural variants (SV)

- 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

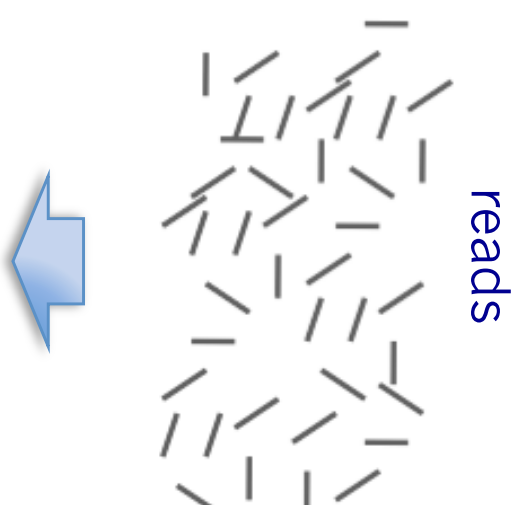


Les variants nucléotidiques (SNP) et structuraux



An assembly

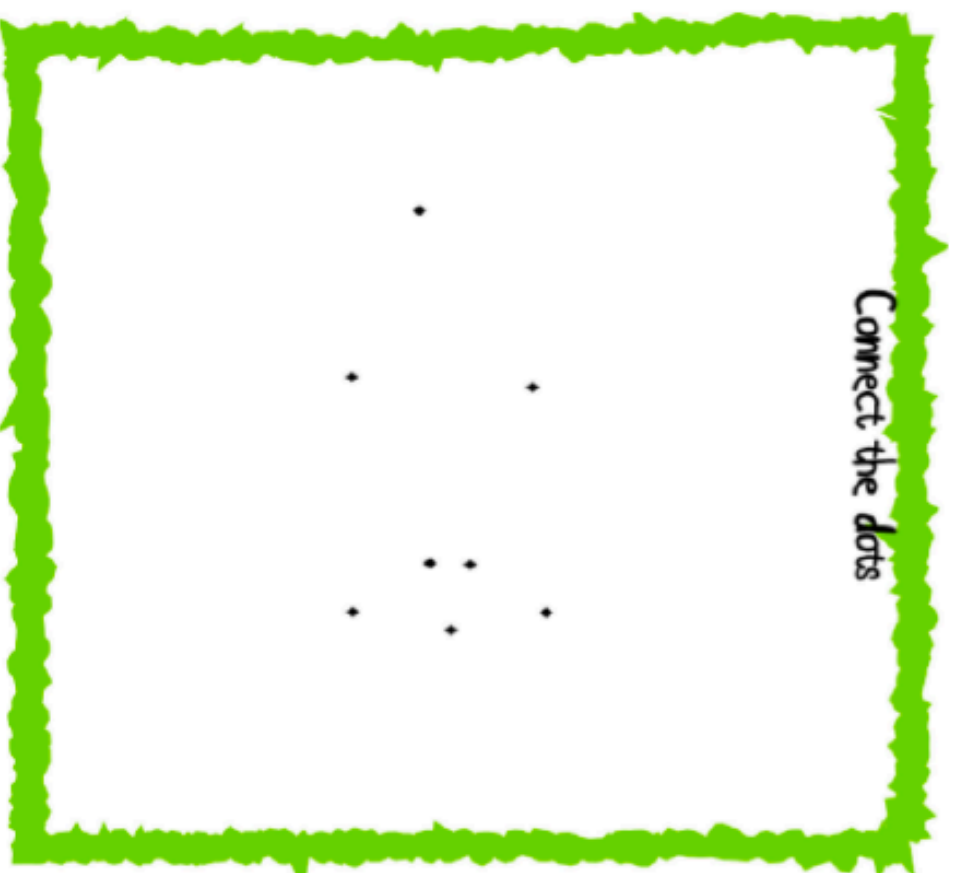
What you get from
the sequencing



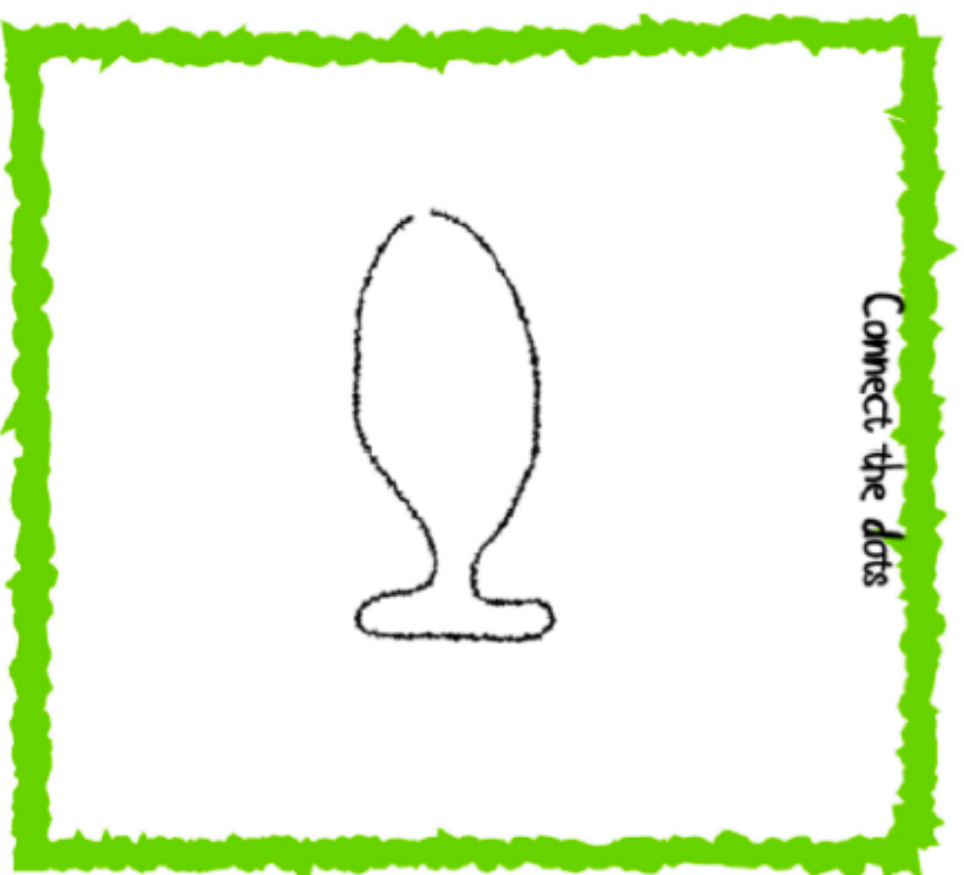
What you want



Assembly = Solving Puzzles Without a Picture



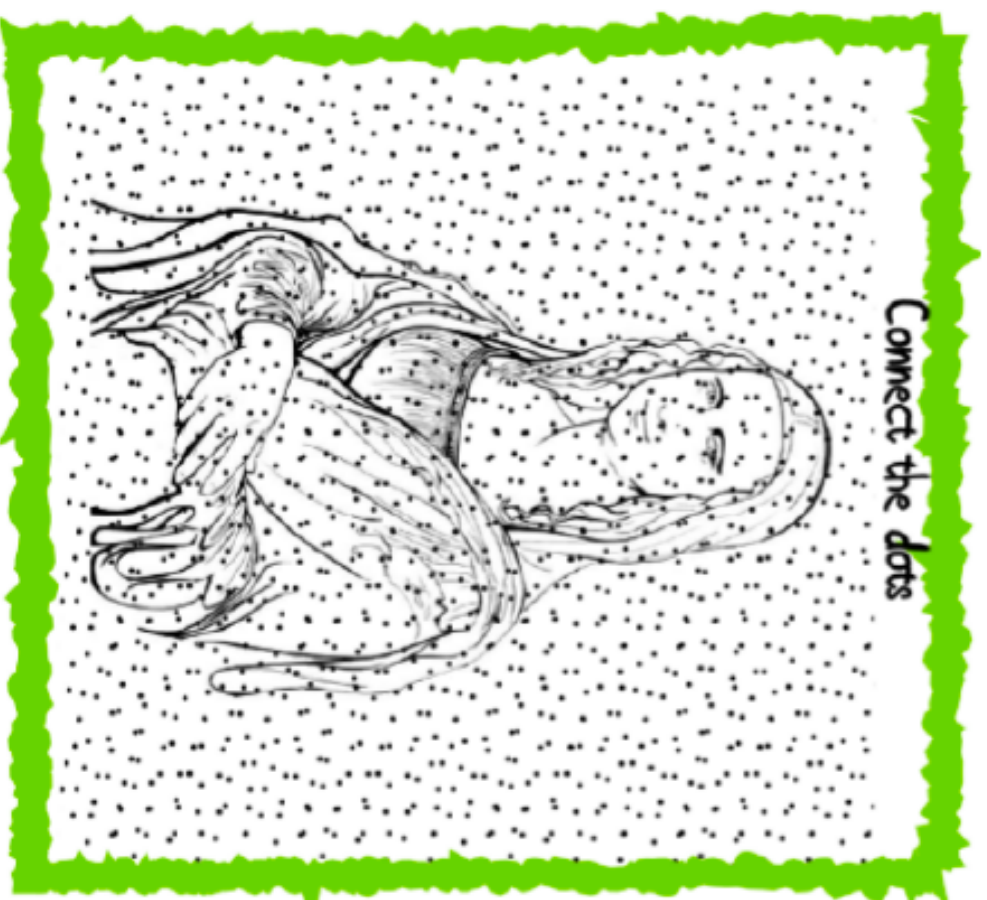
Assembly = Solving Puzzles Without a Picture



Assembly = Solving Puzzles Without a Picture



Assembly = Solving Puzzles Without a Picture



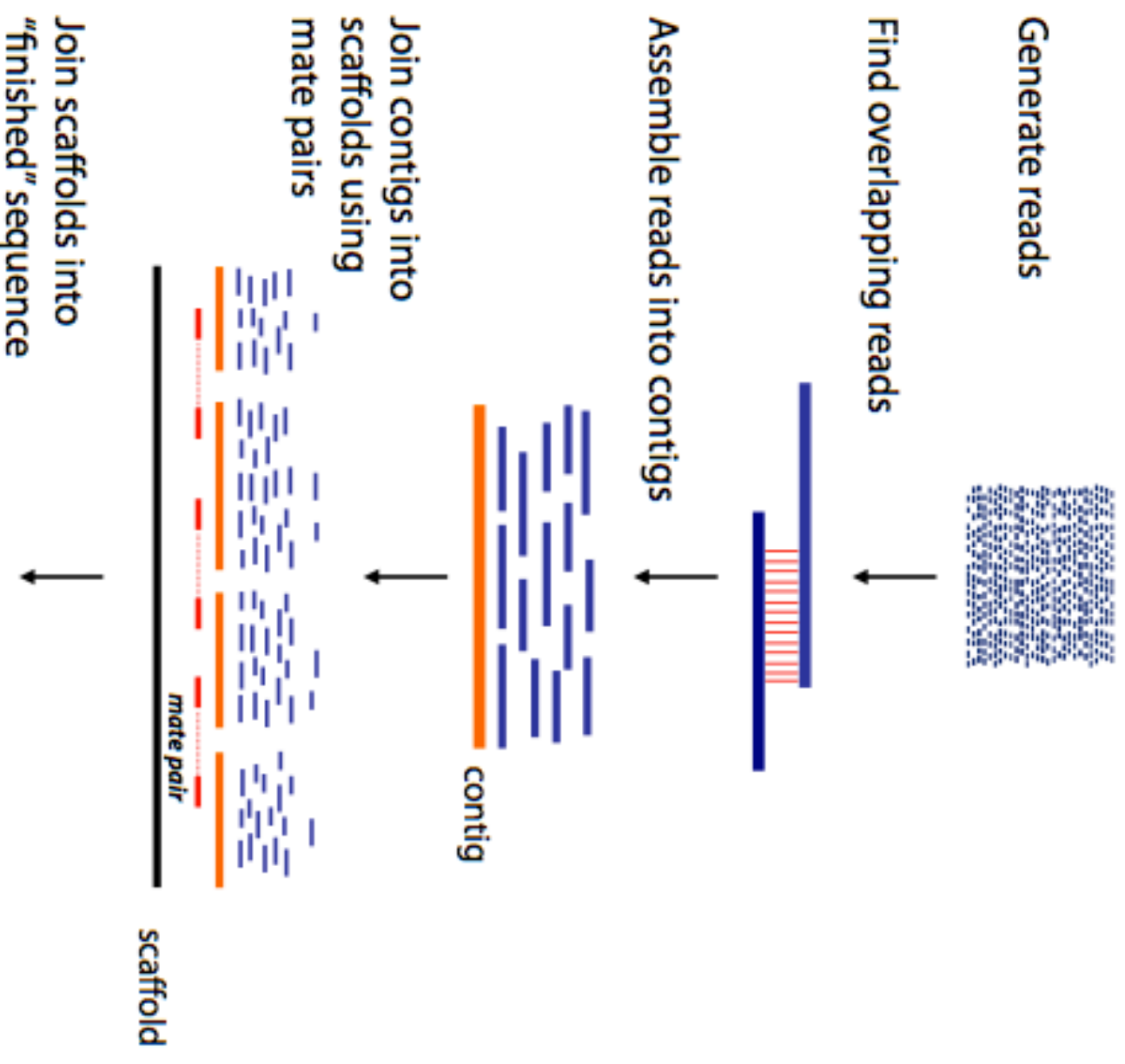
The assembly process is complicated by the fact that underlying assumption is incorrect.

Assembly process

Contig = a set of reads

Unitig = a contig formed from overlapping unambiguously unique sequences (*i.e.*, a high-confidence contig)

Scaffold = an ordered and oriented set of one or more contigs with distances assigned to the gaps between contigs



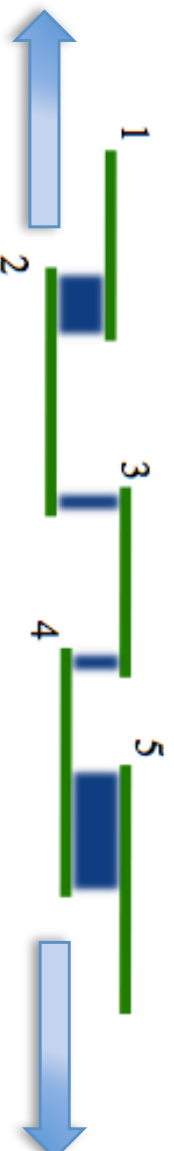
Assembly paradigms

Assemblers are based on one of several different paradigms:

1. Greedy
2. Overlap-Layout-Consensus (OLC)
3. De Bruijn Graph

1/Greedy

- While there are sequences with overlap:
 - Find sequences with largest overlap
 - Merge those sequences



The choices made by the assembler are inherently local and do not take into account the global relationship between the reads.

1/Greedy

- **Advantage:**
 - Simple
- **Disadvantage:**
 - Early mistakes create bad assemblies

2/OLC:

Overlap-Layout-Consensus

2/OLC:

Overlap-Layout-Consensus

- **Clean your input**

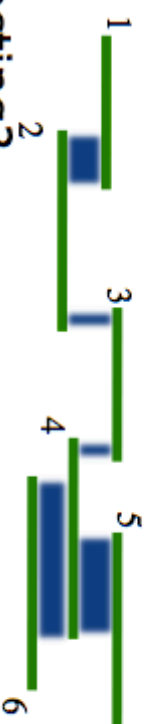
Remove "vector sequence", low quality, etc

2/OLC:

Overlap-Layout-Consensus

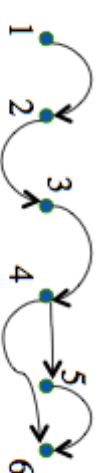
- **Clean your input**

Remove “vector sequence”, low quality, etc



- **Overlap:** What reads are intersecting?

- Create a node for each read
- Create directed edge for each overlap

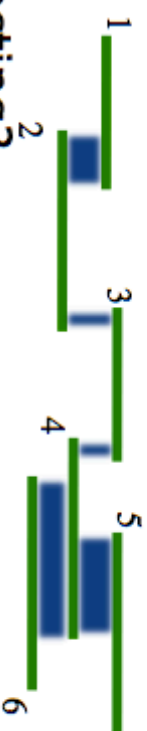


2/OLC:

Overlap-Layout-Consensus

- **Clean your input**

Remove “vector sequence”, low quality, etc

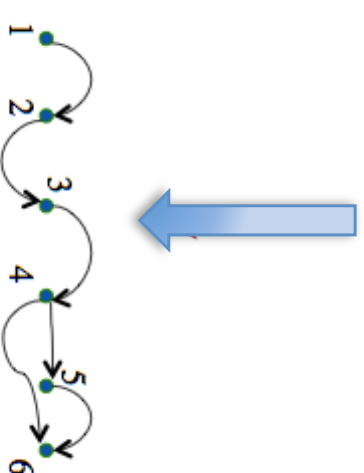


- **Overlap:** What reads are intersecting?

- Create a node for each read
- Create directed edge for each overlap

- **Layout:** How combine the reads?

- Simplify graph
- Find suitable paths in the graph

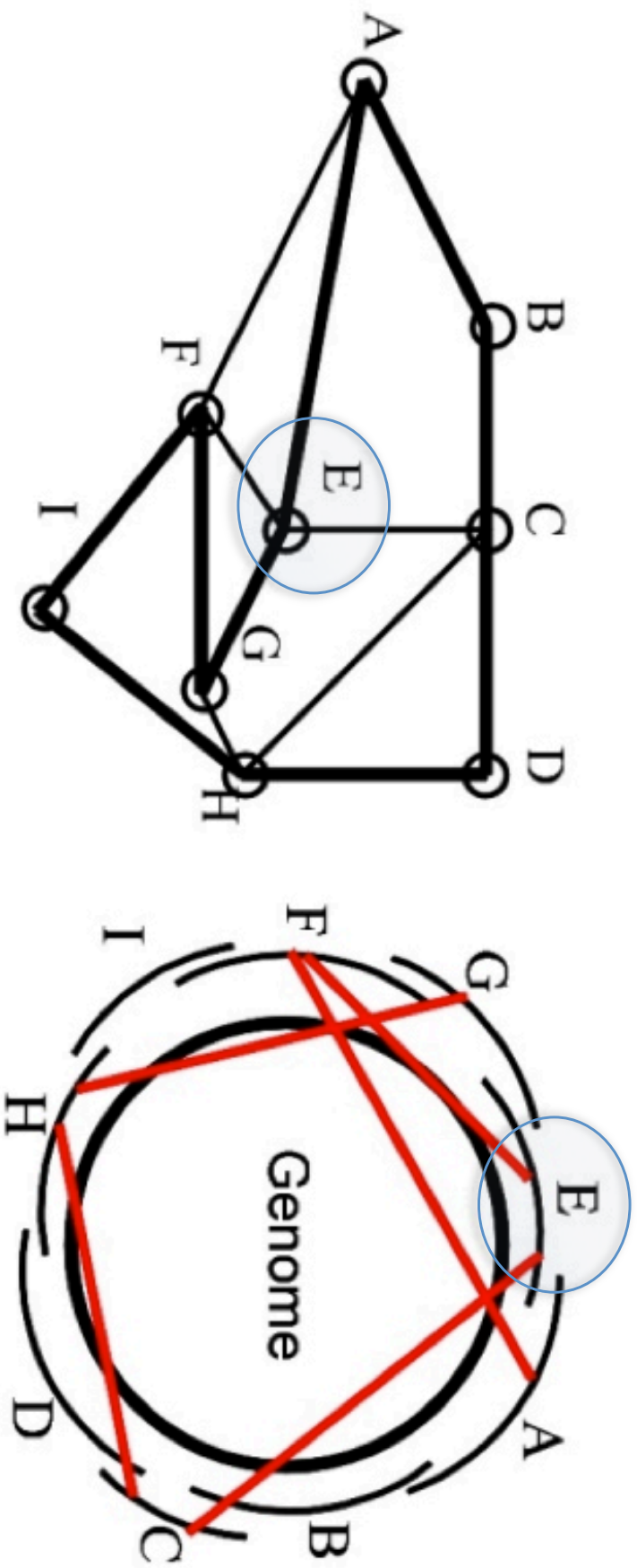


Determine Hamiltonian path

« In the mathematical field of graph theory, a *Hamiltonian path* (or *traceable path*) is a path in an undirected or directed graph that visits each node exactly once. »

2/OLC:

Overlap-Layout-Consensus



2/OLC:

Overlap-Layout-Consensus

Example:

- True sequence (7bp) AGTCTAT
- 3 Reads (4bp each) AGTC (A), GTCT (B), CTAT (C)
- Alignments

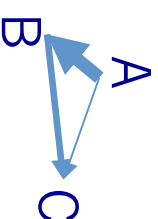
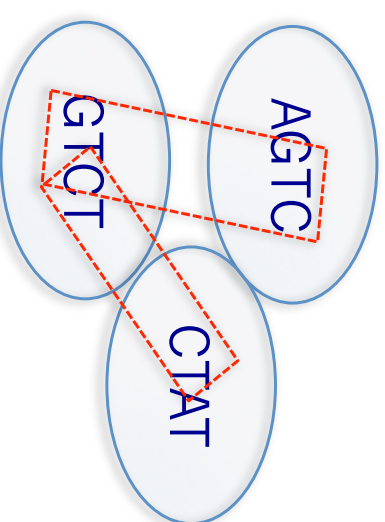
AGTC-	AGTC---	GTCT--
-GTCT	---CTAT	--CTAT

2/OLC:

Overlap-Layout-Consensus

- Nodes are the 3 read sequences
- Edges are the overlap alignment with orientation
- Edge thickness represents score of overlap

Optimal path: A -> B -> C

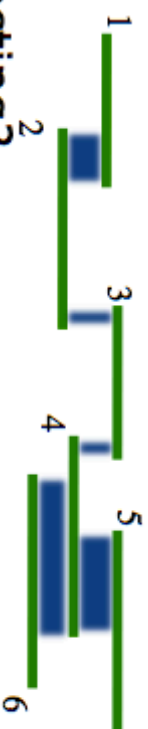


2/OLC:

Overlap-Layout-Consensus

- **Clean your input**

Remove “vector sequence”, low quality, etc



- **Overlap:** What reads are intersecting?

- Create a node for each read
- Create directed edge for each overlap



- **Layout:** How combine the reads?

- Simplify graph
- Find suitable paths in the graph

- **Consensus:** Derive contigs from layout

2/OLC:

Overlap-Layout-Consensus



		↓		↓	
Seq4	TTCACACACCTATACCAATAGTTT	CTGGCTCCTGACCA	TCAAACTG		
Seq5		TTTCTGGCTCCTGACCT	TCAAACTGCCCTCCATATGACTGTGCTCT		
Seq6			TACCAATAGTTTA	CTGGCTCCTGACCT	TCAAACTGCCCTCC
Seq7			ATAGTTTCTGGCTCCTGACCG	TCAAACTGCCCTCCATATGA	
Cons	TTCACACACCTATACCAATAGTT	TTCTGGCTCCTGACCN	TCAAACTGCCCTCCATATGACTGTGCTCT		

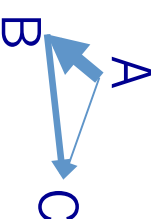
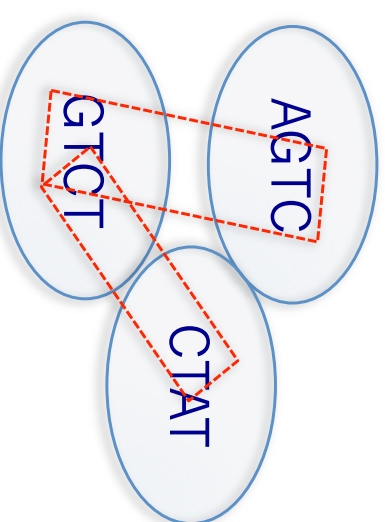
2/OLC:

Overlap-Layout-Consensus

- Nodes are the 3 read sequences
- Edges are the overlap alignment with orientation
- Edge thickness represents score of overlap

Optimal path: A -> B -> C

Consensus aGTCTat



2/OLC:

Overlap-Layout-Consensus

The OLC paradigm was made **popular** by the work of Gene Myers, embodied in Celera Assembler and **dominated the assembly world until the emergence of the new generation of short-read sequencing technologies.**

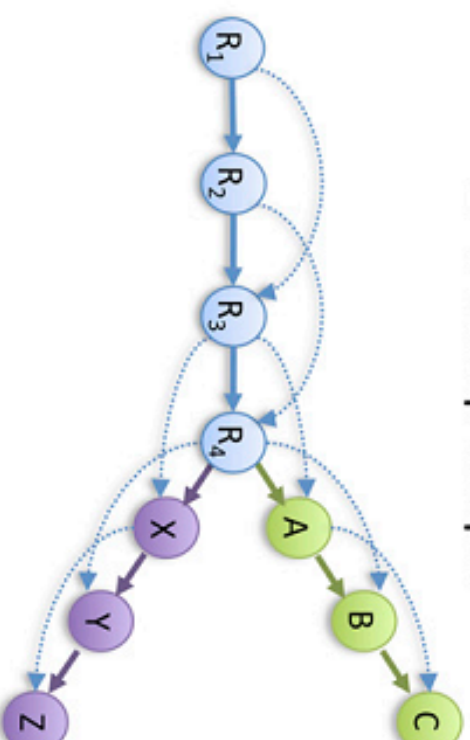
3/De Bruijn Graphs

A Read Layout

```

R1:: GACCTACA
R2::  ACCCTACAA
R3::   CCTACAAG
R4::    CTACAAGT
A:    TACAAGTT
B:    ACAAGTTA
C:    CAAGTTAG
X:    TACAAGTC
Y:    ACAAGTCC
Z:    CAAGTCCG
    
```

B Overlap Graph



3/De Bruijn Graphs

- Divide reads into smaller strings of size k (k -mers)
 - Break reads of L bp into $L-k+1$ k -mers per read

3/De Bruijn Graphs

- Divide reads into smaller strings of size k (k -mers)
 - Break reads of L bp into $L-k+1$ k -mers per read
- If $L=36$ and $k=31$, we will get $36-31+1=6$ k -mers

3/De Bruijn Graphs

- Divide reads into smaller strings of size k (k -mers)
 - Break reads of L bp into $L-k+1$ k -mers per read
 - If $L=36$ and $k=31$, we will get $36-31+1=6$ k -mers
 - Why create even smaller segments?

3/De Bruijn Graphs

- Divide reads into smaller strings of size k (k -mers)
 - Break reads of L bp into $L-k+1$ k -mers per read
If $L=36$ and $k=31$, we will get $36-31+1=6$ k -mers
 - Why create even smaller segments?
 - Smaller chance of containing erroneous base
 - But the tradeoff is that repetitive sequences are more common and harder to resolve
- 1. Construct a de Bruijn graph (DBG)
 1. Nodes = one for each unique k -mer
 2. Edges = $k-1$ exact overlap between two nodes
- 2. Graph simplification
 1. Merge chains, remove bubbles and tips
- 3. Find a Eulerian path through the graph
 1. Linear time algorithm, unlike Hamiltonian

1. Construct a de Bruijn graph (DBG)

- Sequence CAATATG
- K-mers (k=3) CAA AAT ATA TAT ATG
- Graph

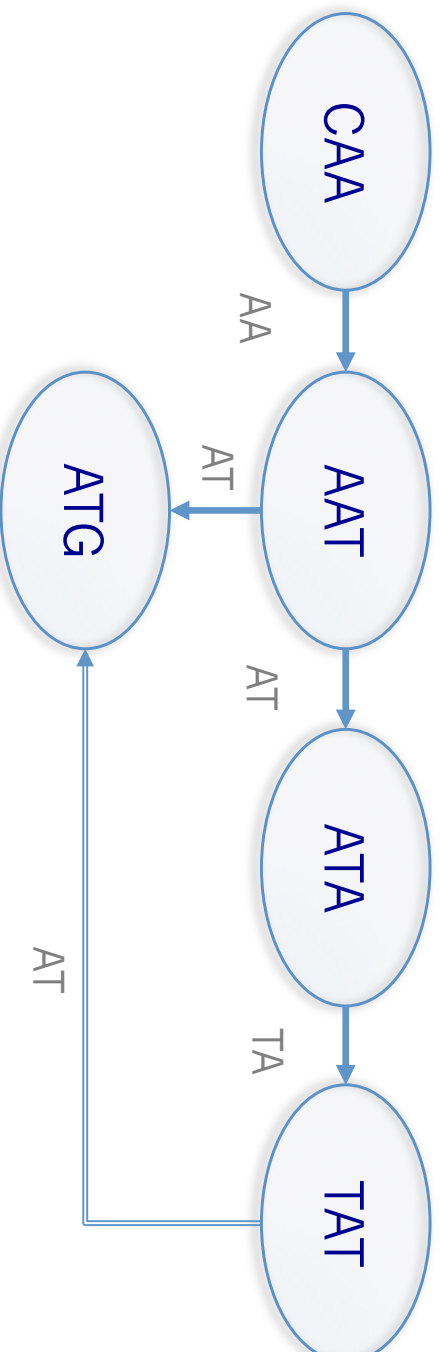
Nodes = one for each unique k-mer

Edges = k-1 exact overlap between two nodes

1. Construct a de Bruijn graph (DBG)

- Sequence CAATATG
- K-mers (k=3) CAA AAT ATA TAT ATG
- Graph

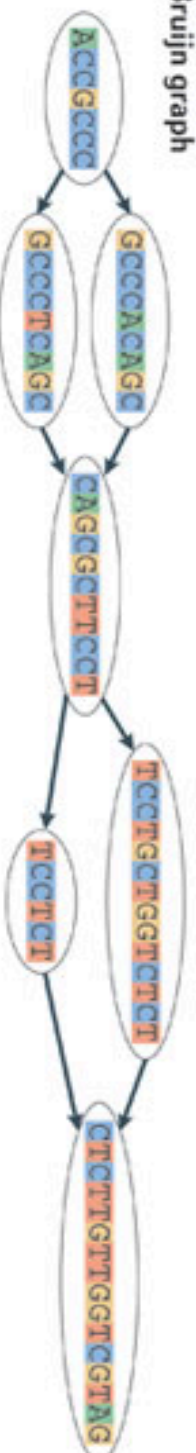
Nodes = one for each unique k-mer
Edges = k-1 exact overlap between two nodes



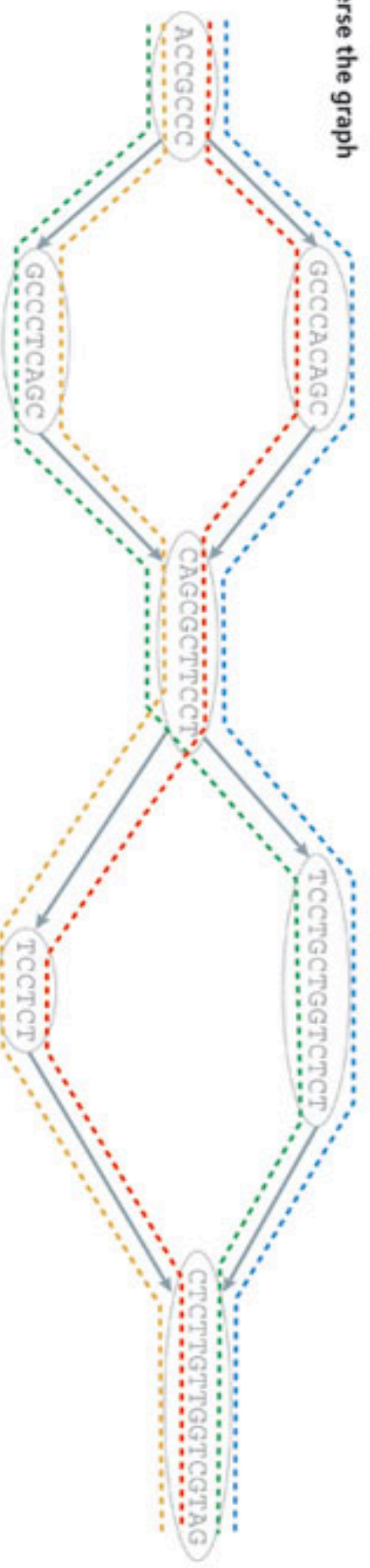
2. graph simplification

- Remove tips or spurs
 - Dead ends in graph due to errors at read end
 - Collapse bubbles
 - Errors in middle of reads
 - But could be true SNPs or diploidity
 - Remove low coverage paths
 - Possible contamination
- Makes final Eulerian path easier and hopefully more accurate contigs

c Collapse the De Bruijn graph



d Traverse the graph



e Assembled isoforms



« In graph theory, an Eulerian trail (or Eulerian path) is a trail in a graph which visits every edge exactly once. »

3/De Bruijn Graphs

De Bruijn approach was popularized by the assembler Euler17 and has dominated the design of **modern assemblers** targeted at short-read sequencing data, such as **Velvet**, **SOAPdenovo** and **ALLPATHS**.

De Bruijn-graph-based approaches have been successful in assembling **highly accurate short reads** (<~100 bp, such as those generated by the Illumina Solexa technology), whereas overlap-based approaches (such as OLC or string graph) are mostly used for longer, more inaccurate data (>200 bp, such as Roche 454 and Sanger sequencing data)

Modern Sequence Assemblers

Assemblers	Technology	Availability	Notes	Refs
Genome assemblers				
ALLPATHS-LG	Illumina, Pacific Biosciences	http://ftp.broadinstitute.org/pub/crd/ALLPATHS/Release-LG	Requires a specific sequencing recipe (BOX 3)	40
SOAPdenovo	Illumina	http://soap.genomics.org.cn/soapdenovo.html	Also used for transcriptome and metagenome assembly	22
Velvet	Illumina, SOLiD, 454, Sanger	http://www.ebi.ac.uk/~zerbino/velvet	May have substantial memory requirements for large genomes	20
ABYSS	Illumina, SOLiD, 454, Sanger	http://www.bcgsc.ca/platform/bioinfo/software/abyss	Also used for transcriptome assembly	21
Metagenome assemblers				
Genovo	454	http://cs.stanford.edu/group/genovo	Uses a probabilistic model for assembly	66
MetaVelvet	Illumina, SOLiD, 454, Sanger	http://metavelvet.dna.bio.keio.ac.jp	Based on Velvet	4
Meta-IDBA	Illumina	http://i.cs.hku.hk/~alse/hkubrgl/projects/metaidba	Based on IDBA	5
Transcriptome assemblers				
Trinity	Illumina, 454	http://trinityrnaseq.sourceforge.net	Tailored to reconstruct full-length transcripts; may require substantial computational time	8
Oases	Illumina, SOLiD, 454, Sanger	http://www.ebi.ac.uk/~zerbino/oases	Based on Velvet	72
Single-cell assemblers				
SPAdes	Illumina	http://bioinf.spbau.ru/en/spades		7
IDBA-UD	Illumina	http://i.cs.hku.hk/~alse/hkubrgl/projects/idba_ud	Based on IDBA	6

Note that only a few of the popular and freely available assemblers are included here for each application (a more complete list is provided in Supplementary information S1 (table)), and all of the listed assemblers (except Genovo) are based on de Bruijn graph construction. IDBA, Iterative De Bruijn graph short read Assembler.

Modern Sequence Assemblers

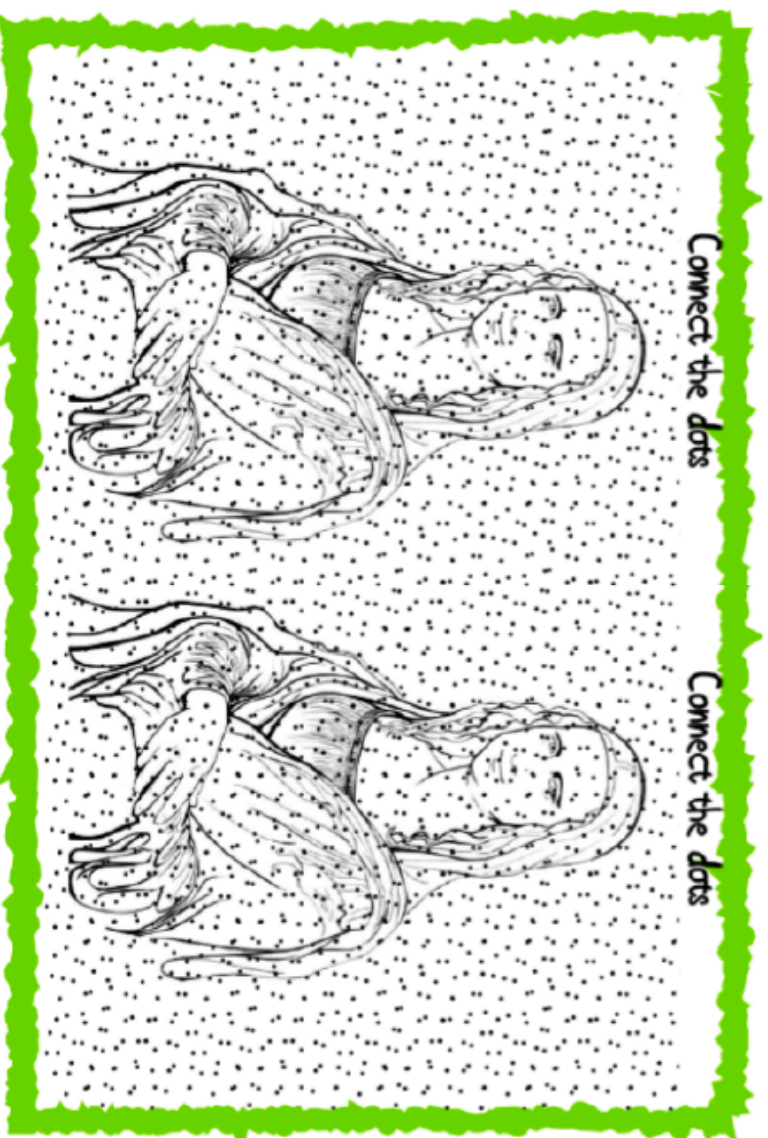
Assemblers	Technology	Availability	Notes	Refs
Genome assemblers				
ALLPATHS-LG	Illumina, Pacific Biosciences	ftp://ftp.broadinstitute.org/pub/crd/ALLPATHS/Release-LG	Requires a specific sequencing recipe (BOX 3)	40
SOAPdenovo	Illumina	http://soap.genomics.org.cn/soapdenovo.html	Also used for transcriptome and metagenome assembly	22
Velvet	Illumina, SOLiD, 454, Sanger	http://www.ebi.ac.uk/~zerbino/velvet	May have substantial memory requirements for large genomes	20
ABYSS	Illumina, SOLiD, 454, Sanger	http://www.bcgsc.ca/platform/bioinfo/software/abyss	Also used for transcriptome assembly	21
Metagenome assemblers				
MetaVelvet	Illumina, SOLiD, 454, Sanger	http://metavelvet.sourceforge.net/	Based on velvet	4
Meta-IDBA	Illumina	http://l.cs.hku.hk/~alse/hkubrg/projects/metaidba	Based on IDBA	5
Transcriptome assemblers				
Trinity	Illumina, 454	http://trinityrnaseq.sourceforge.net	Tailored to reconstruct full-length transcripts; may require substantial computational time	8
Oases	Illumina, SOLiD, 454, Sanger	http://www.ebi.ac.uk/~zerbino/oases	Based on Velvet	72
Single-cell assemblers				
SPAdes	Illumina	http://bioinf.spbau.ru/en/spades		7
IDBA-UD	Illumina	http://l.cs.hku.hk/~alse/hkubrg/projects/idba_ud	Based on IDBA	6

Note that only a few of the popular and freely available assemblers are included here for each application (a more complete list is provided in Supplementary Information S1 (table)), and all of the listed assemblers (except Genovo) are based on de Bruijn graph construction. IDBA, Iterative De Bruijn graph short read Assembler.

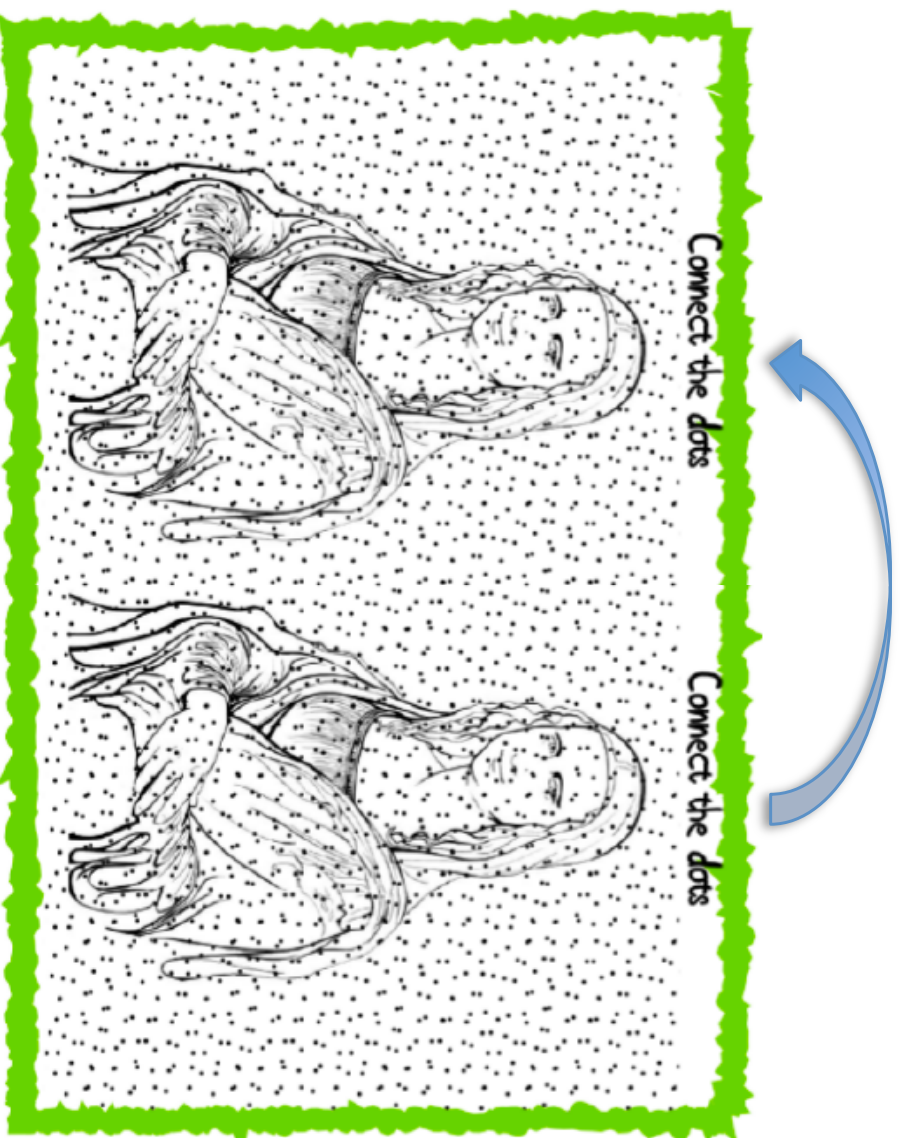
Genome Assembly Gold-Standard Evaluations

<http://gage.cbcb.umd.edu>

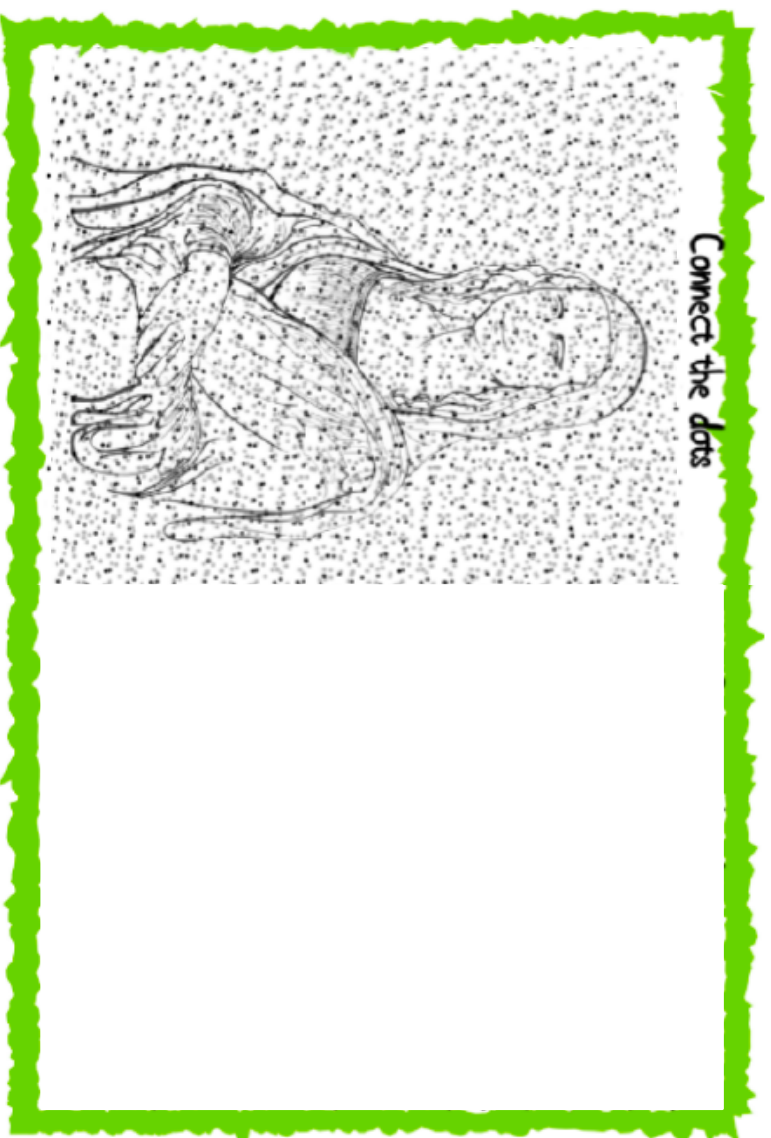
The main source of mis-assembly: Repeats



The main source of mis-assembly: Repeats

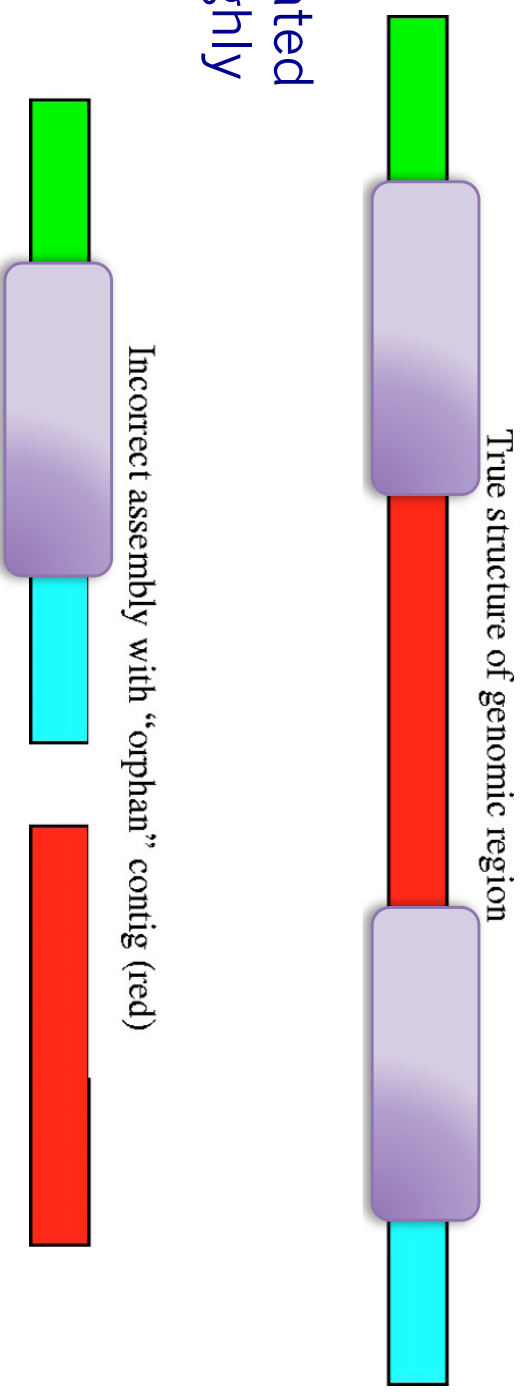


The main source of mis-assembly: Repeats



The main source of mis-assembly: Repetitive DNA

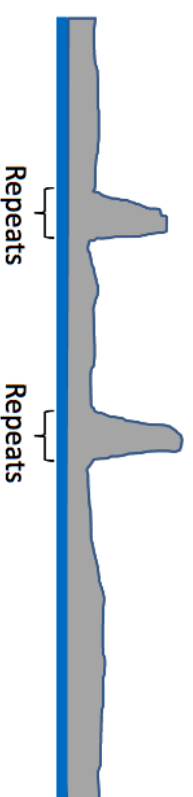
Segments of DNA repeated
yield fragments with highly
similar sequences that
originate from different
places in the genome.



Omic impact

Genome assemblers expect even coverage.

Assembled regions with high coverage are assumed to be repeats.



Transcriptome
→ nearly identical
sequences may originate
from different transcripts

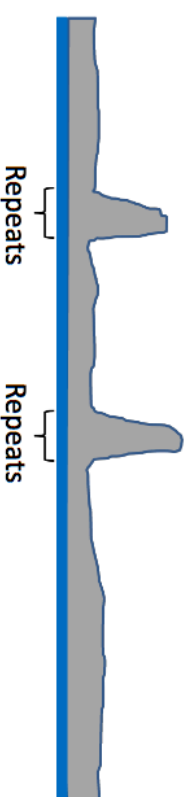
But individual genes within a transcriptome will have very
different amounts of coverage...



Omic impact

Genome assemblers expect even coverage.

Assembled regions with high coverage are assumed to be repeats.



Transcriptome
→ nearly identical sequences may originate from different transcripts

But individual genes within a transcriptome will have very different amounts of coverage...



Metagenomic samples

→ nearly identical sequences may originate from genomes within the sample.

Sequencing to Assembly

In:

A set of reads



Out:

A genome model

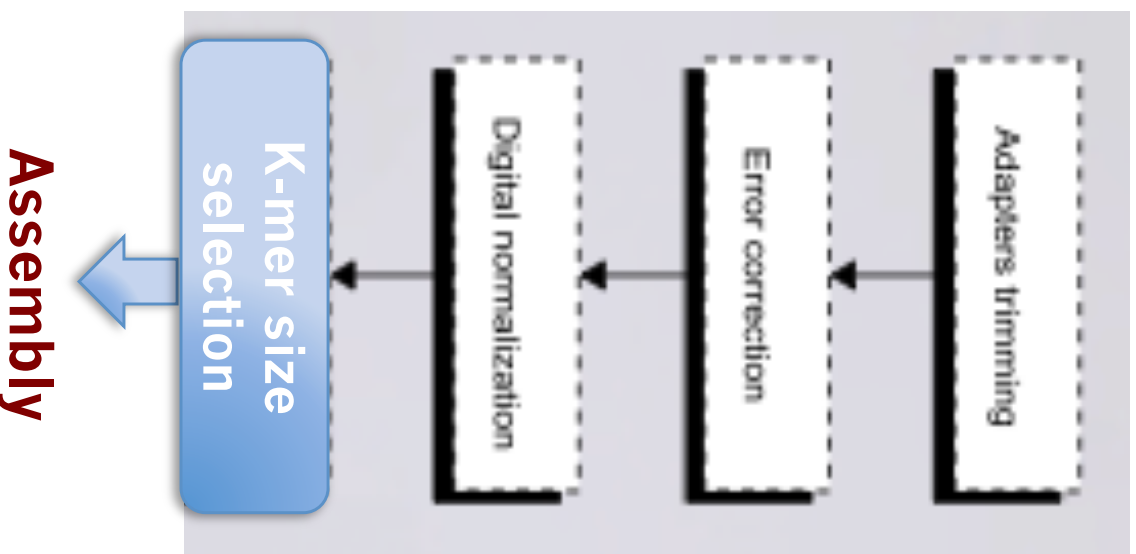


In practice:

A set of contigs

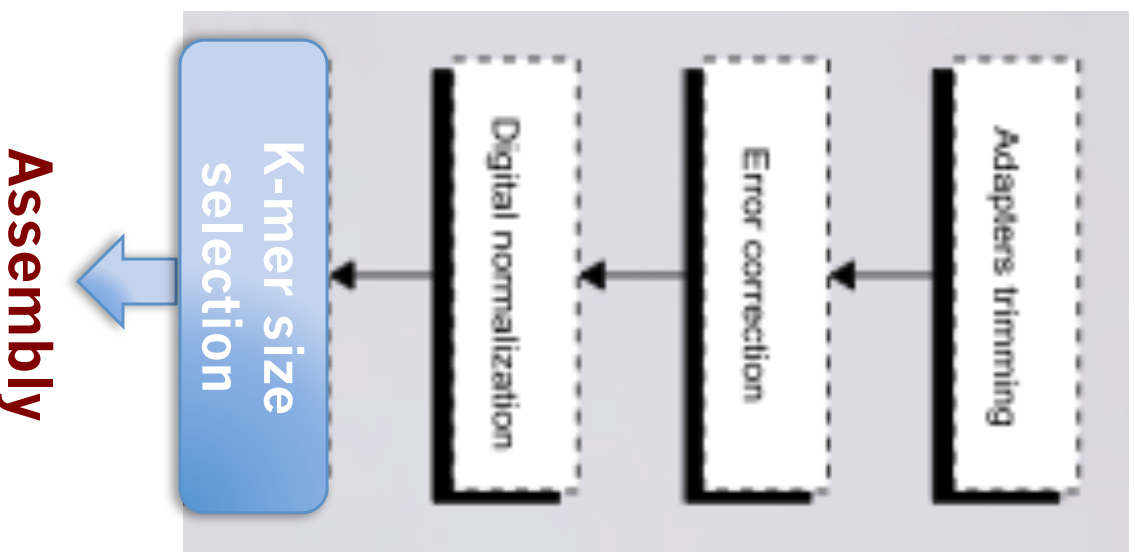


From reads to assembly



From reads to assembly

Most assemblers cut reads into k-mers

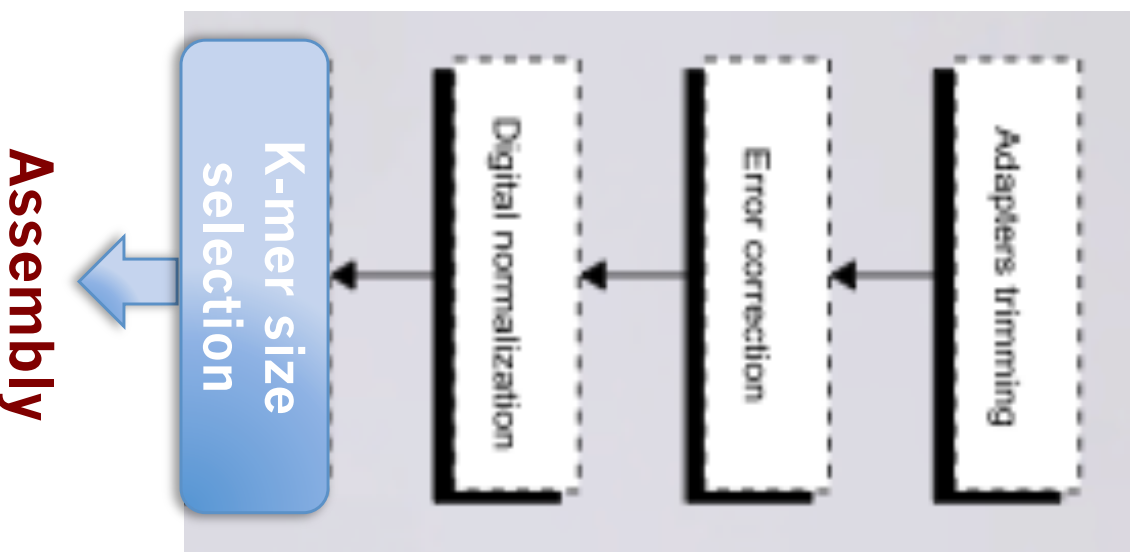


From reads to assembly

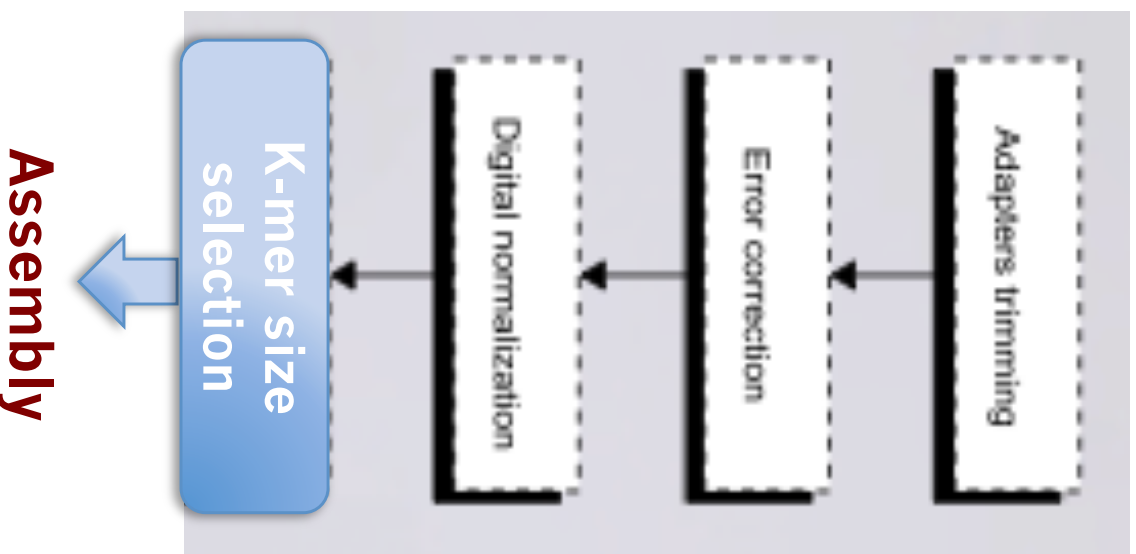
Most assemblers cut reads into k-mers



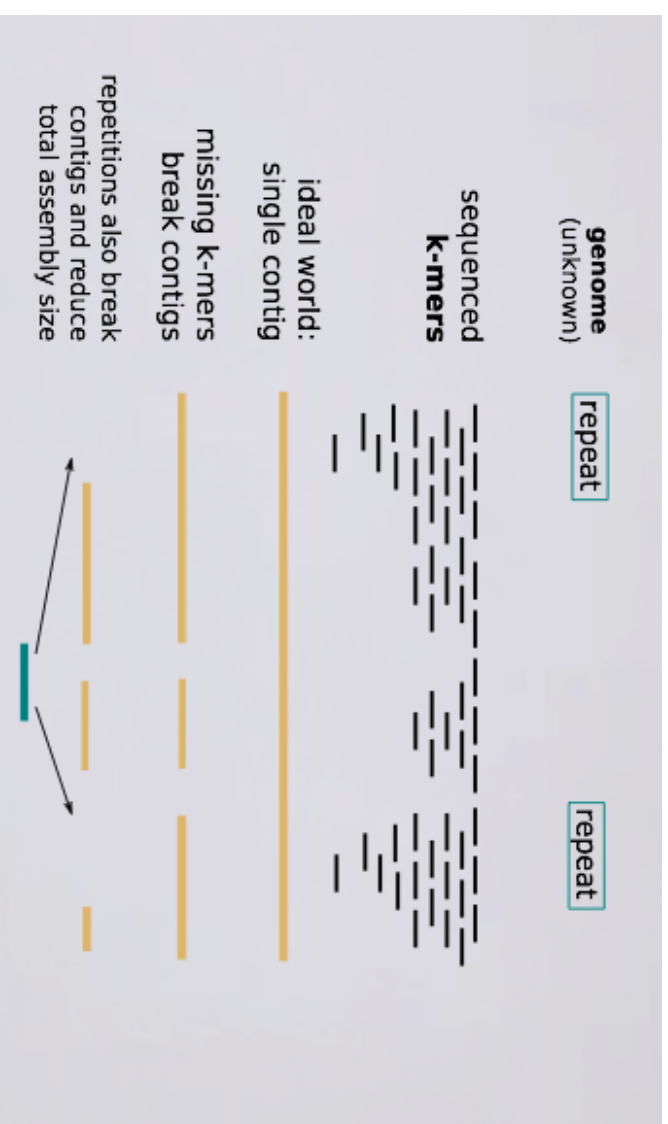
Why this is so important?



From reads to assembly



Why this is so important?
Again repeats!!!



K should be kept small to prevent the overuse of computer memory, while still large enough so that most k-mers are unique in the genome

k-mer size < mean repeat size

Assembly

How to estimate the optimum k-mer size?

Without looking at the data

$$k_{\text{opt}} = \lceil (N_k/G) - C \rceil$$

where N_k is the total number of k-mers in the reads,
 G is the estimated genome size,
 C is the desired k-mer coverage.

How to estimate the optimum k-mer size?

Looking at the data -> Compute the k-mer abundance histogram

- x axis: *abundance*
- y axis: number of k-mers having abundance x (seen x times)

Example reads dataset:

ACTCA
GTCA

Solution: 3-mers

Solution: Abundance of each
distinct 3-mer

How to estimate the optimum k-mer size?

Looking at the data -> Compute the k-mer abundance histogram

- x axis: *abundance*
- y axis: number of k-mers having abundance x (seen x times)

Example reads dataset:

ACTCA
GTCA

3-mers:

ACT
CTC
TCA
GTC
TCA

Solution: Abundance of each
distinct 3-mer

How to estimate the optimum k-mer size?

Looking at the data -> Compute the k-mer abundance histogram

<ul style="list-style-type: none"> - x axis: <i>abundance</i> - y axis: number of k-mers having abundance x (seen x times) 	
Example reads dataset:	
ACTCA	
GTCA	
3-mers:	
ACT	
CTC	
TCA	
GTC	
TCA	

Abundance of each distinct 3-mer:	
ACT: 1	
CTC: 1	
TCA: 2	
GTC: 1	
3-mer abundance:	
x y	
1 3	
2 1	
3 0	
4 0	

How to estimate the optimum k-mer size?

Looking at the data -> Compute the k-mer abundance histogram

- Several tools to build the k-mer histograms already exist (e.g. **k-mer counting**, **Jellyfish**, **DSK**...)
- Chikhi R., Medvedev P. designed one approach to estimate the optimum k-mer (**KmerGenie** Bioinformatics 2013)

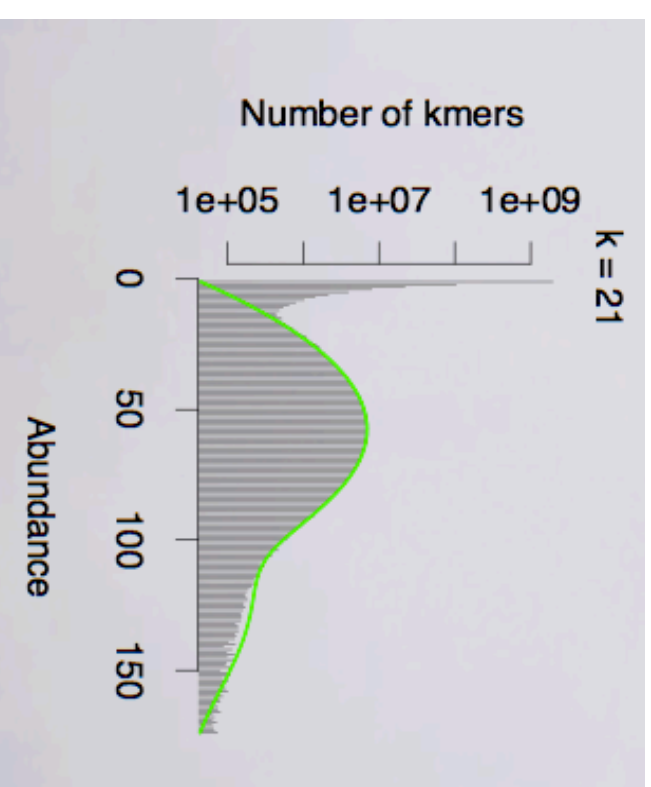
Understand the k-mer histogram

GAGE dataset

Human chr14 ~ 88 Mb

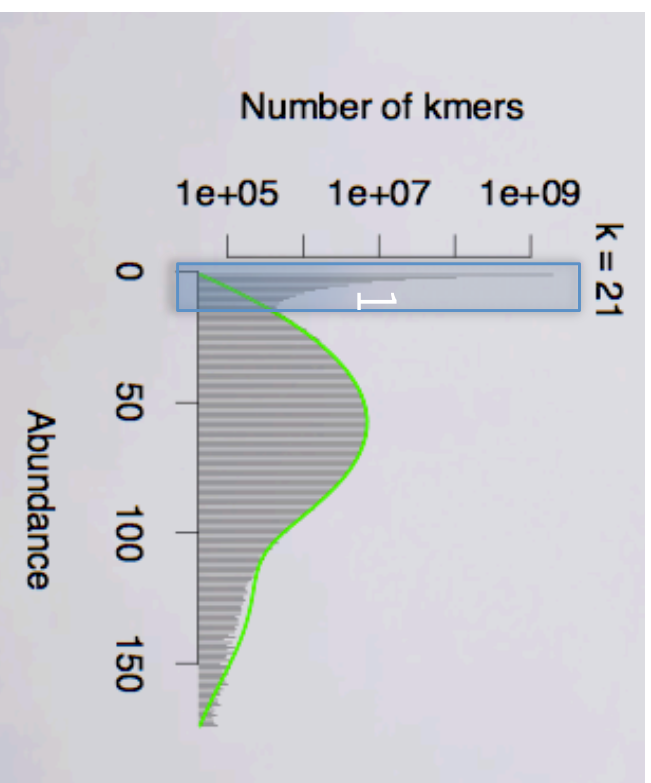
k=21

We expect to see multiple peaks, from different causes.



Understand the k-mer histogram

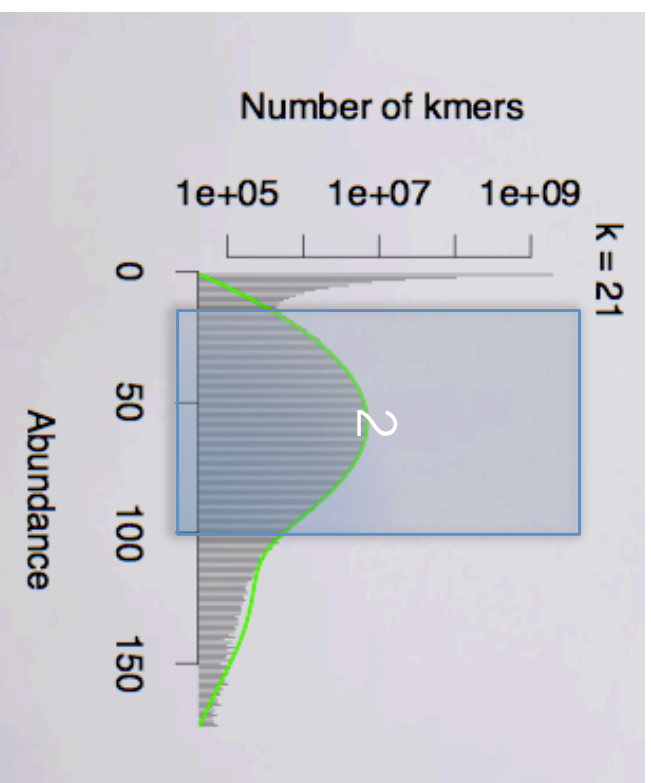
Low k-mer frequency = Erroneous k-mers



- Sequencing errors,
- PCR amplification errors,
- Polymorphisms...

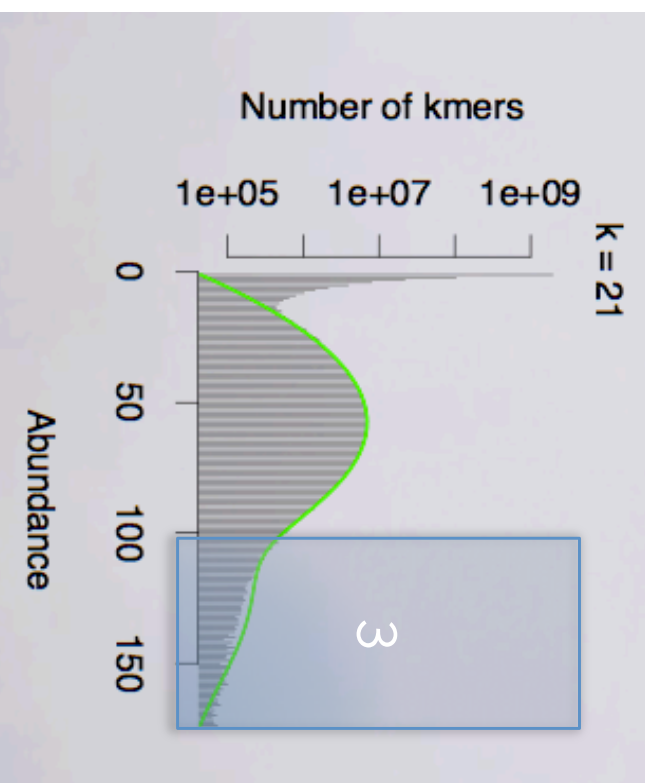
Note: For metagenomic libraries, that polymorphism may be a major contributor to the initial peak

Understand the k-mer histogram



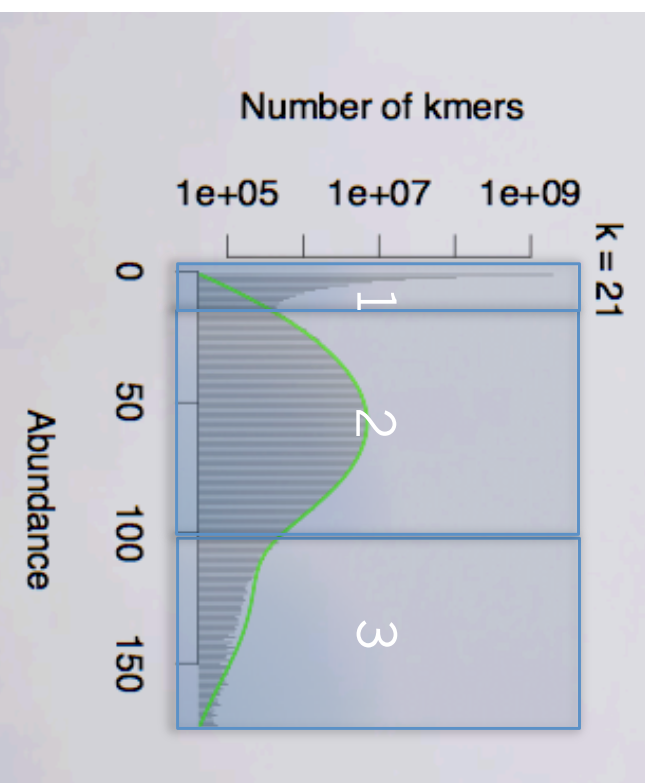
Genomic non-repeated k-mers

Understand the k-mer histogram



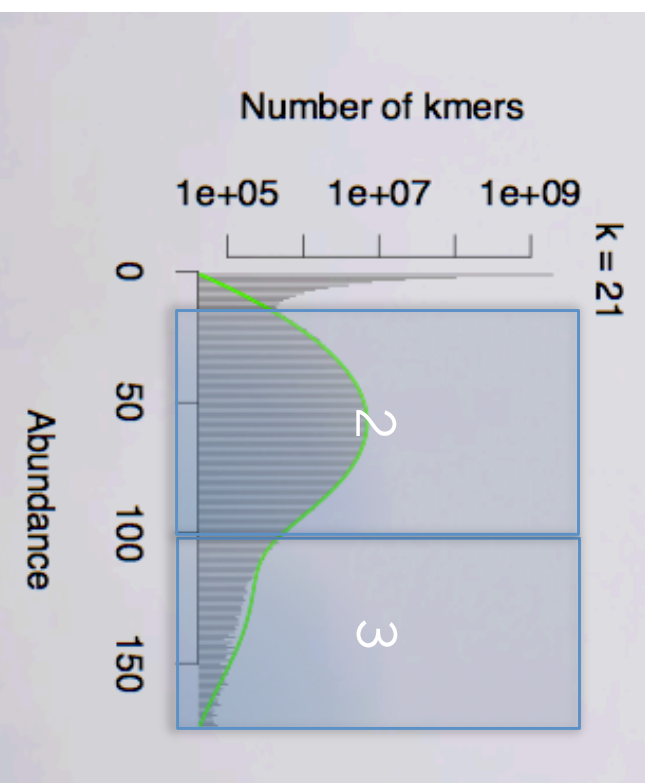
**Genomic repeated
k-mers or artifacts**

Understand the k-mer histogram



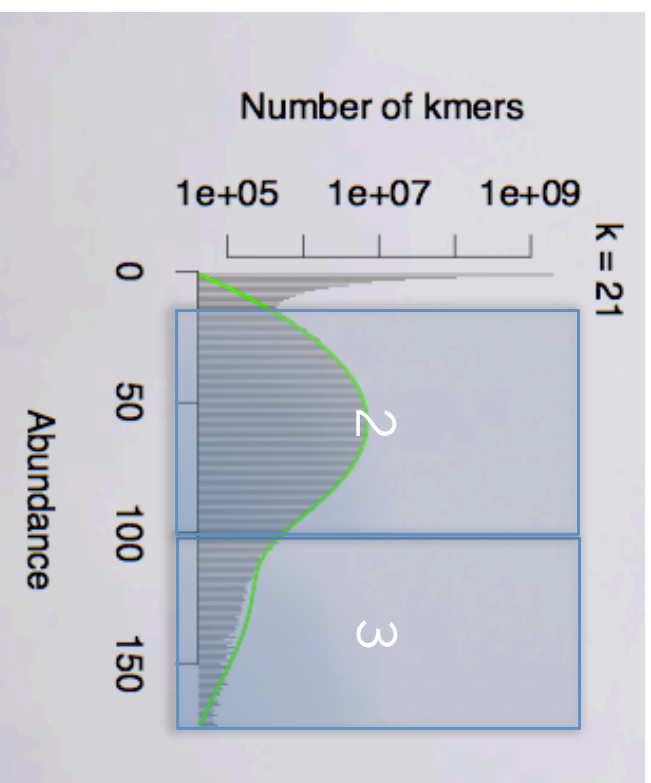
-> genomic and sequencing characteristics

Understand the k-mer histogram



**2 + 3 = total number
of distinct k-mers
covering the genome**

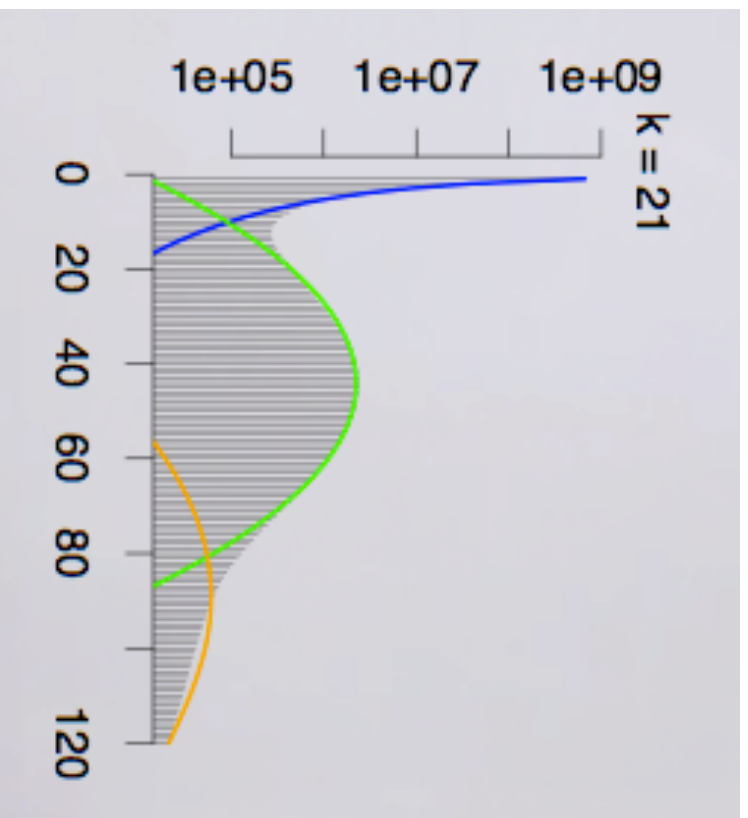
Understand the k-mer histogram



**2 + 3 = total number
of distinct k-mers
covering the genome**



How to determine exactly the area



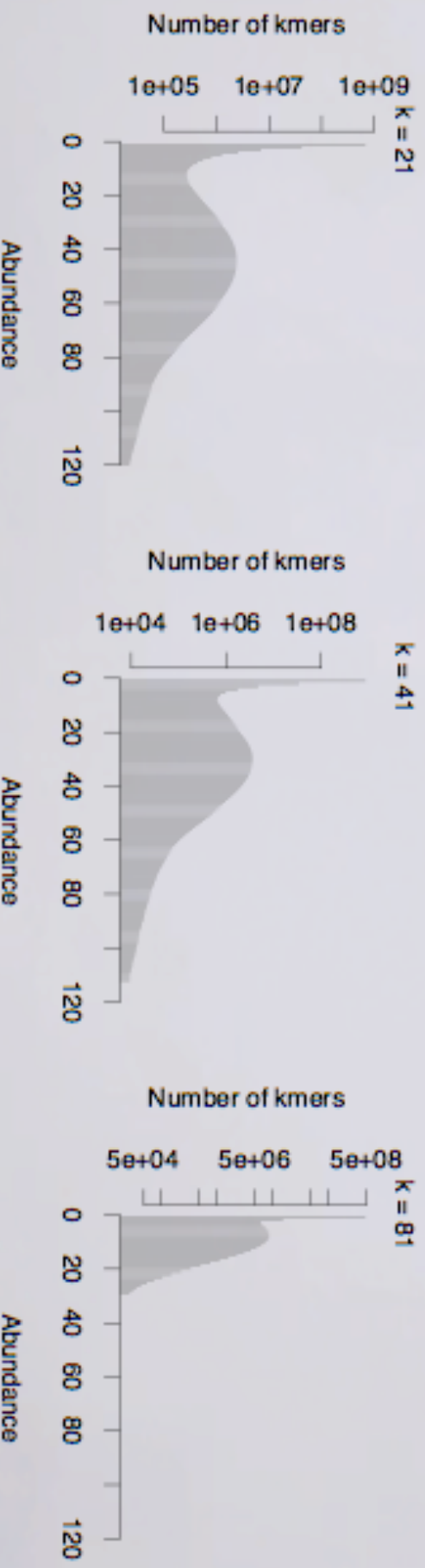
Quake's statistical model

They fit a Gamma distribution to the untrusted k-mers, a Gaussian for untrusted k-mers, a Gaussian for untrusted k-mers, a Gaussian for untrusted k-mers and a Zeta for untrusted k-mers and a Zeta for untrusted k-mers.

The distribution of the trusted reads is actually expected to be Poisson, but the variance is significantly larger than the mean due to sequencing biases

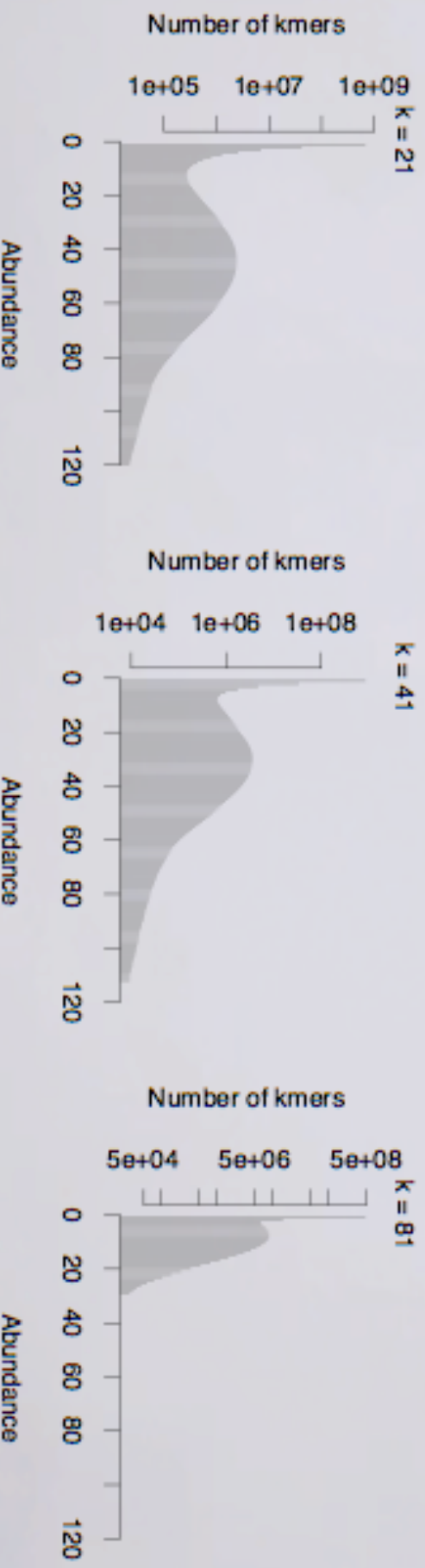
Now, read the k -mer histograms

To find the optimal k , one can **compare histograms** for different values of k .



Now, read the k-mer histograms

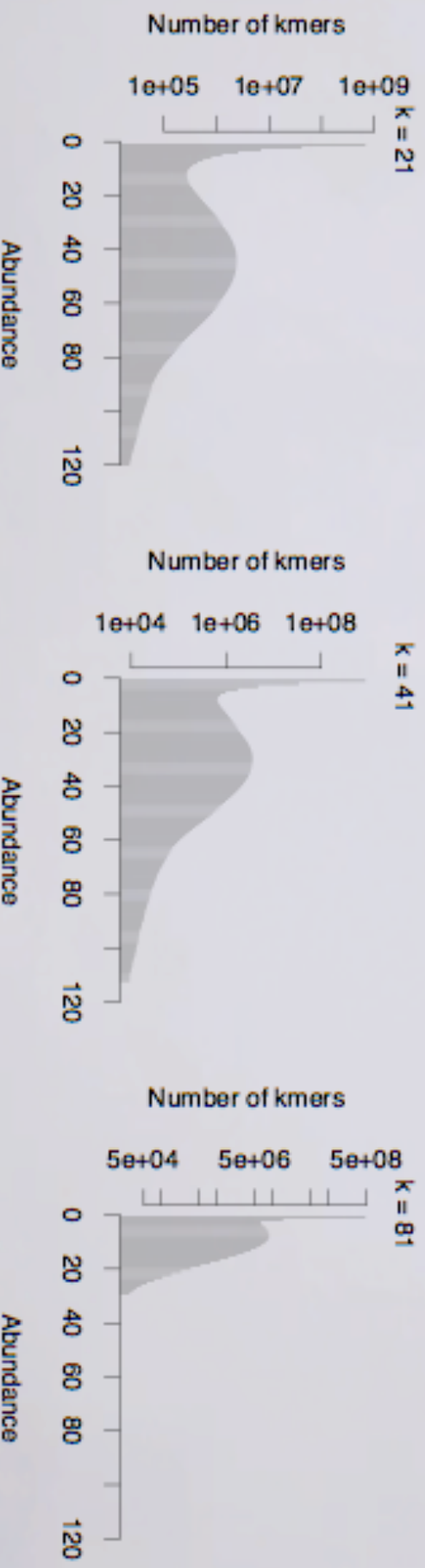
To find the optimal k , one can **compare histograms** for different values of k .



Choose the k value which maximizes the assembly size

Now, read the k -mer histograms

To find the optimal k , one can **compare histograms** for different values of k .

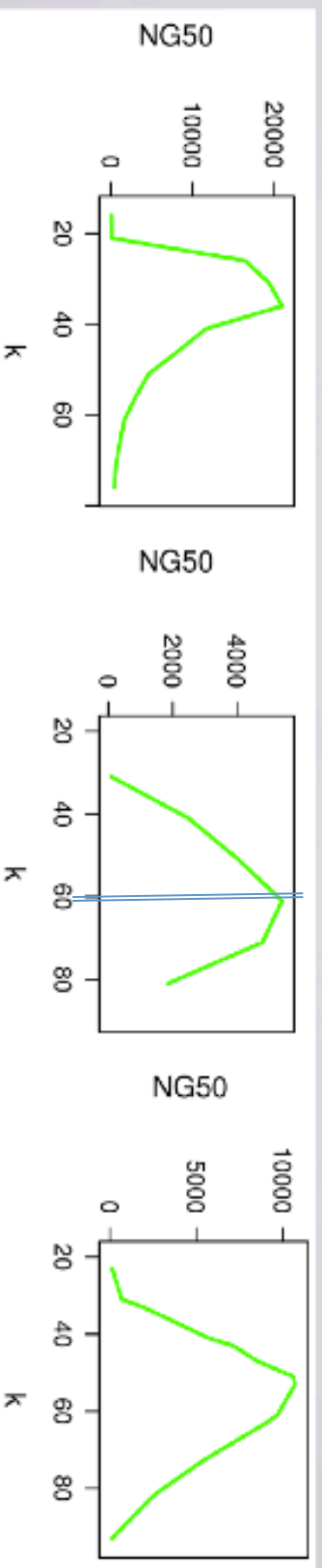


Computing a single histogram is time and memory expensive.
=> quickly estimates the histograms using sampling

The k -parameter

Assembly is not robust with respect to the parameter k . Because the ideal k -mer size depends on :

- sequencing coverage
- sequencing error rate
- genome complexity



k vs NG50 for 3 organisms : bacteria (*S. aureus*), human chr14, whole bumblebee genome (*B. impatiens*)

How to estimate the optimum k-mer size?

Velvet Advisor

Questions

I have million reads.

They are reads.

Each read is base-pairs long.

I estimate my genome size to be megabases (million bases).

I would like to have about fold k-mer coverage for my assembly (defined below, suggest between 10 and 30)

Answer

You have a yield of megabases.

You have about fold *nucleotide* coverage of your genome.

We recommend trying k= for your Velvet assembly.

The Velvet sequence type is:

http://dna.med.monash.edu.au/~torsten/velvet_advisor/

The main stats to estimate the quality of an assembly

- Number of contigs/scaffolds
- Total length of the assembly
- Length of the largest contig/scaffold
- Percentage of gaps in scaffolds ('N')
- N50/NG50 of contigs/scaffolds
- Number of predicted genes
- Number of core genes

The main stats to estimate the quality of an assembly

- Number of contigs/scaffolds
- Total length of the assembly
- Length of the largest contig/scaffold
- Percentage of gaps in scaffolds ('N')
- N50/NG50 of contigs/scaffolds
- Number of predicted genes
- Number of core genes

K-mer!!!

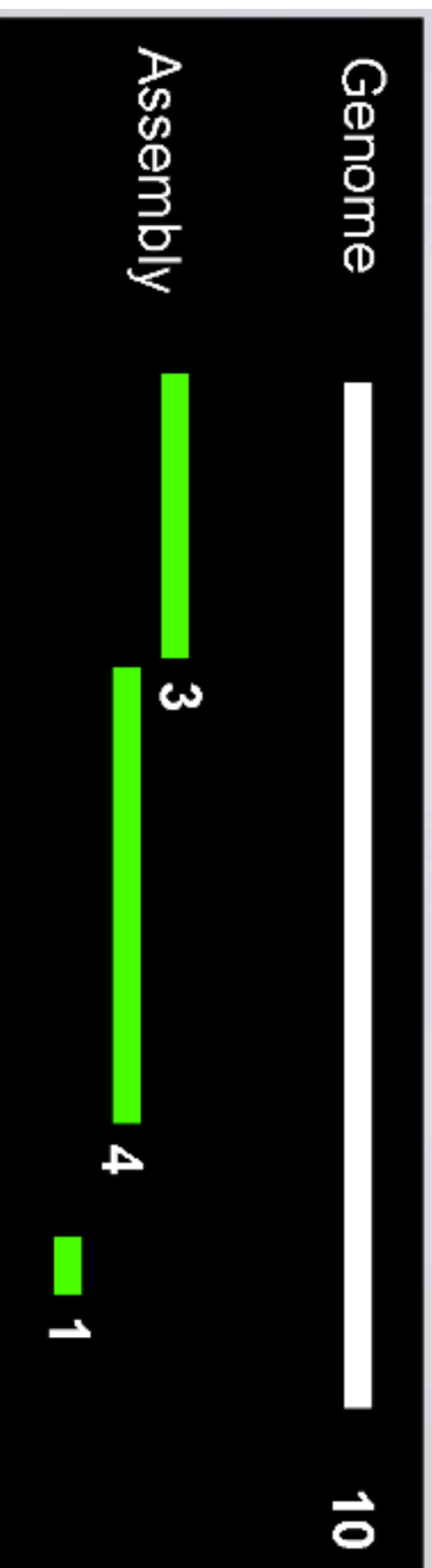
The main stats to estimate the quality of an assembly

- Number of contigs/scaffolds
- Total length of the assembly
- Length of the largest contig/scaffold
- Percentage of gaps in scaffolds ('N')
- N50/NG50 of contigs/scaffolds
- Number of predicted genes
- Number of core genes

N50/NG50

N50 = Largest contig length at which longer contigs cover 50% of the total **assembly** length

NG50 = Largest contig length at which longer contigs cover 50% of the total **genome** length



N50/NG50

N50 = Largest contig length at which longer contigs cover 50% of the total **assembly** length

NG50 = Largest contig length at which longer contigs cover 50% of the total **genome** length

Genome



10

Assembly



3



4



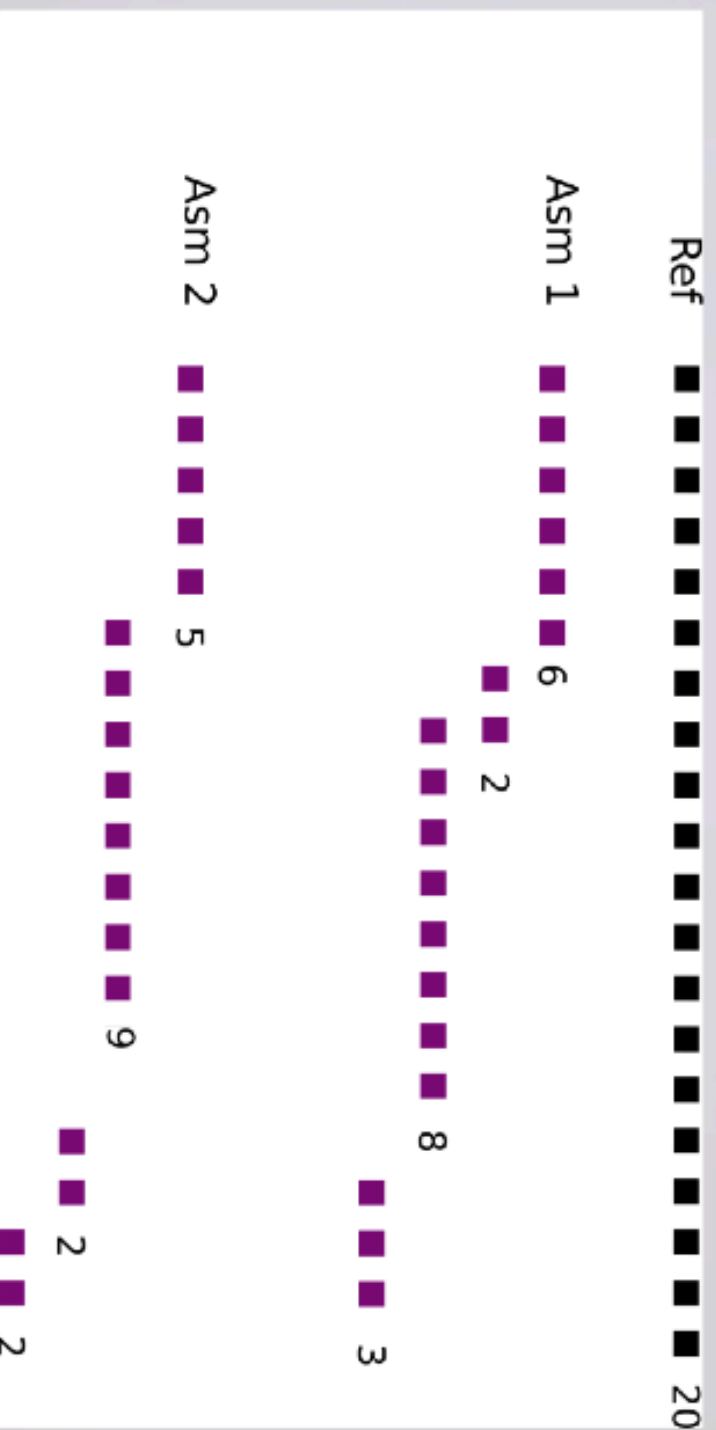
1

A practical way to compute N50 :

- Sort contigs by decreasing lengths
- Take the first contig (the largest) : does it cover 50% of the assembly ?
- If yes, this is the N50 value. Else, try the next one (the second largest), and so on..

Let's do it

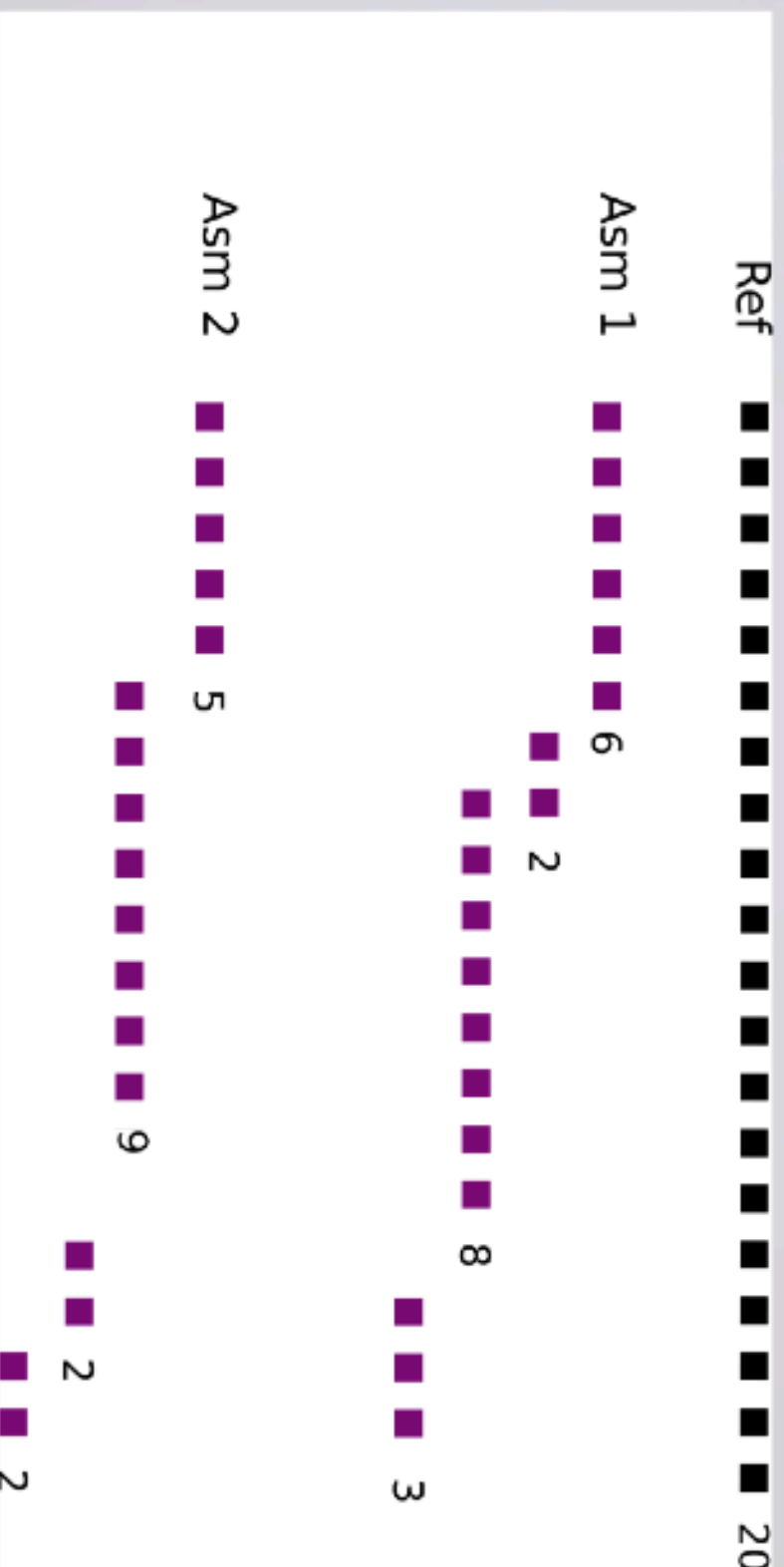
Here are two assemblies, aligned to the same reference :



- For each, compute the following metrics :
 - ▶ Total size of the assembly, N50, NG50 (bp)
 - ▶ Coverage (%)
- Which one is better than the other ?

Solution

Here are two assemblies, aligned to the same reference :



- For each, compute the following metrics :
 - ▶ Total size of the assembly (19 bp, 18 bp), N50 (6 bp, 9 bp), NG50 (6 bp, 5 bp)
 - ▶ Coverage (%) (90, 90)
- Which one is better than the other? (I would say first one)

High N50 but bad assembly

“The standard of judging assembly quality by size of contigs is questionable. Large contigs may simply reflect overly aggressive joining of contigs, thereby creating larger contigs with mis-assemblies. As a consequence, genome scientists who are not experts at assembly can be completely misled by statistics about contig sizes, and as a result might prefer the ‘larger’ but incorrect assembly when given a choice.”

Salzberg & Yorke, 2005

Assembly score

N_{50} is the N50 statistic of the assembly

G_S is the number of contigs contained in the assembly

M_S is the mean assembly contig length

then $G_S M_S := L_S$ is the estimated size of the assembled genome

L_E is the expected (actual) genome size

then $|L_E - L_S|$ is the error between the expected size of the (actual) genome and the estimated size of the assembled genome, and in non-pathological cases, $|L_S - (L_E + 1)| > 0$

\mathcal{A} is the Assembly Score

$$\mathcal{A} = \frac{N_{50}}{G_S |L_S - (L_E + 1)|} \left(\frac{1}{10} \right)$$

Note that the $\frac{1}{10}$ multiplier is only added for ease of interpretation of the scores. It has no biological meaning.

Mapping vs. Assembly

