6th Infectivology Today®

NGS
NEXT GENERATION SEQUENCING

Paestum (Sa) 15- 16 -17 maggio 2014

Relatore
Dr Cataldo Senatore
Dr.ssa Emilia Vaccaro

DEFINIZIONI

Sequenziamento

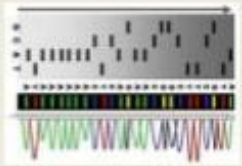
Il termine, in biologia molecolare, indica il processo per la determinazione esatta della struttura primaria di un biopolimero (basi nel caso di un acido nucleico, aminoacidi nel caso di proteine)

Sequenziamento del DNA

Determinazione dell'ordine dei diversi nucleotidi che costituiscono l'acido nucleico

Brief history of sequencing

- First Generation: Sanger sequencing



- Second Generation: amplified molecule sequencing

- Third Generation: single molecule sequencing

Sanger Sequencing Reactions

For given template DNA, it's like PCR except:

Uses only a single primer and polymerase to make new ssDNA pieces.

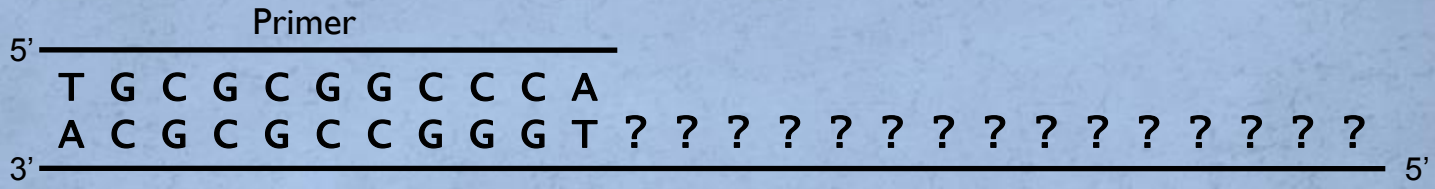
Includes regular nucleotides (A, C, G, T) for extension, but also includes dideoxy nucleotides.

Regular Nucleotides

Dideoxy Nucleotides

1. Labeled
2. Terminators

Sanger Sequencing

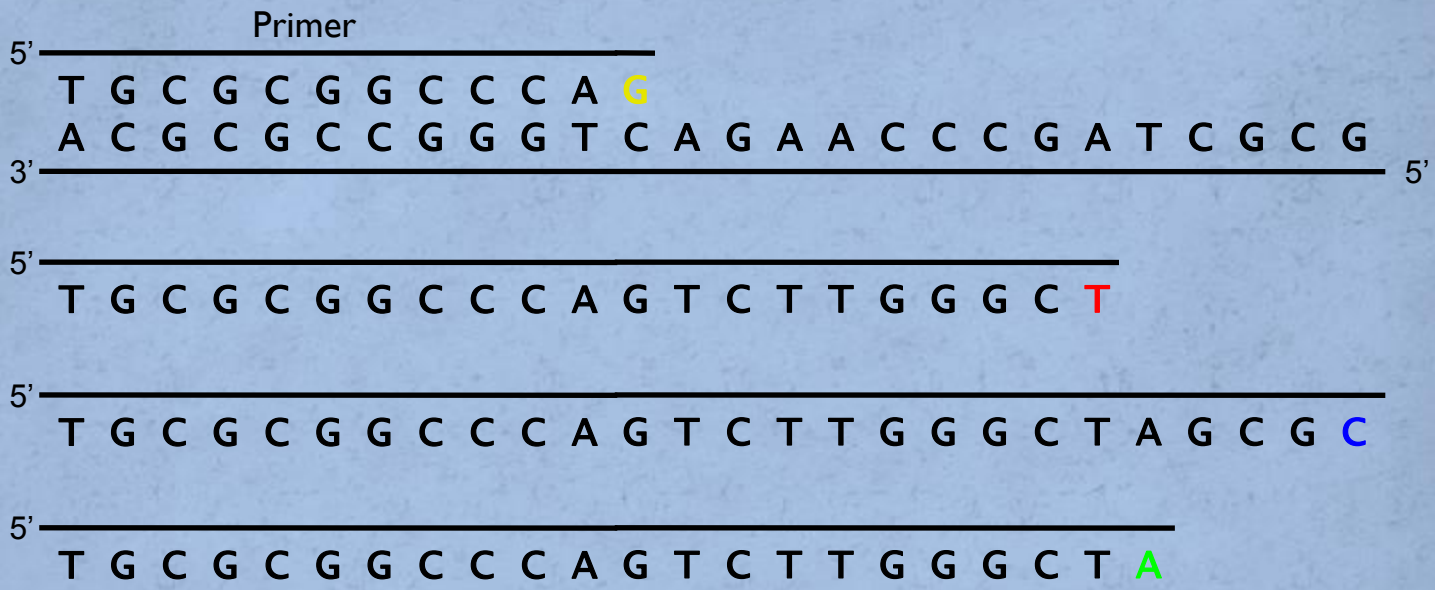


Sanger Sequencing

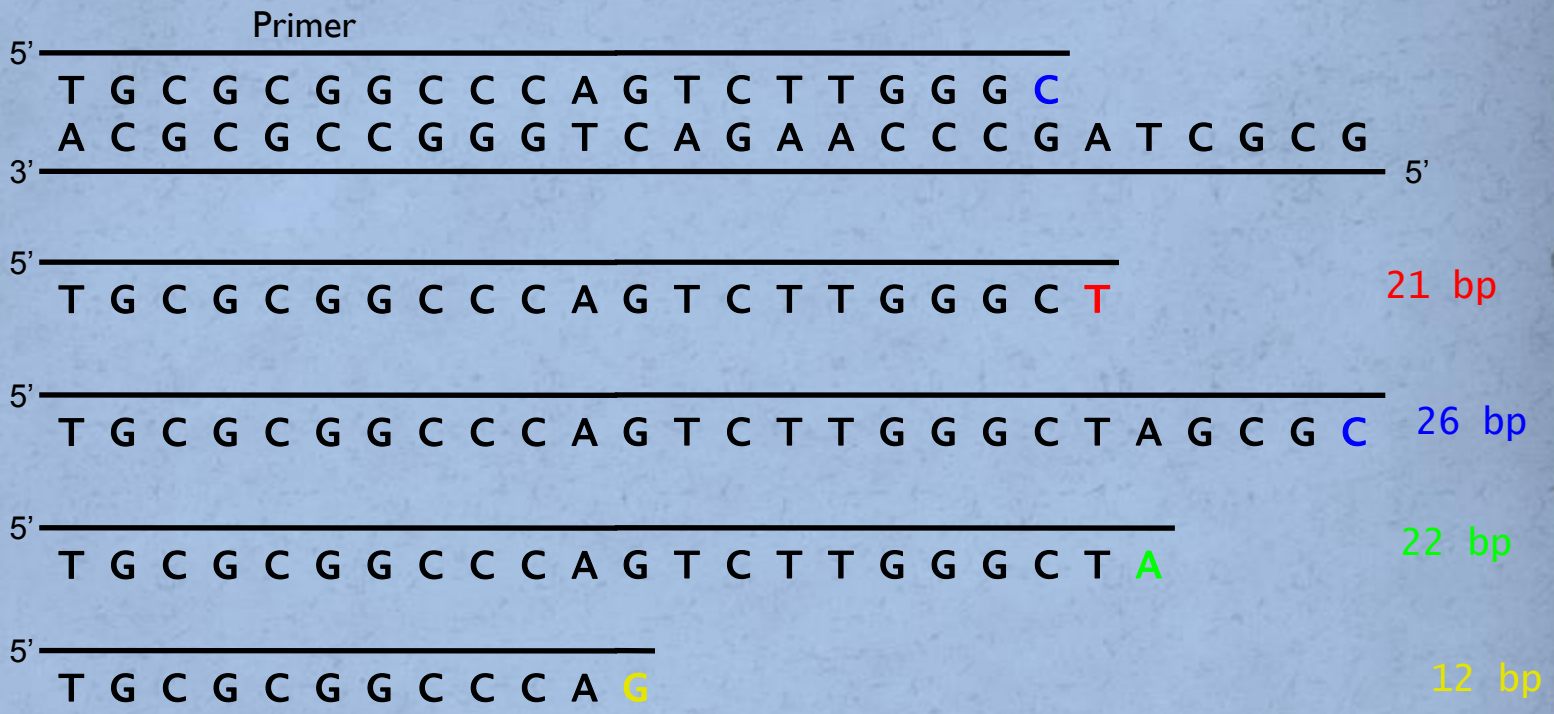
Sanger Sequencing

Sanger Sequencing

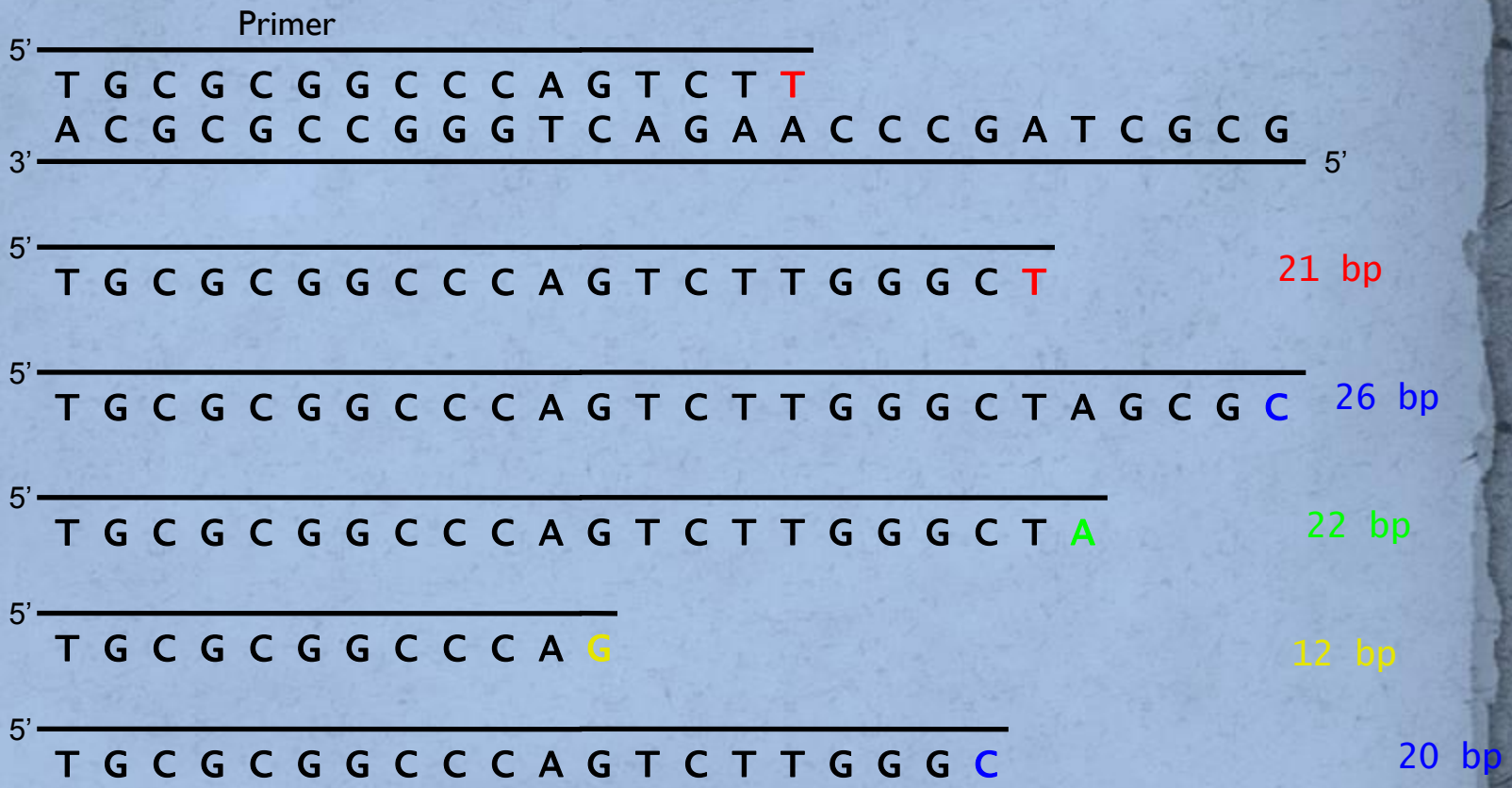
Sanger Sequencing



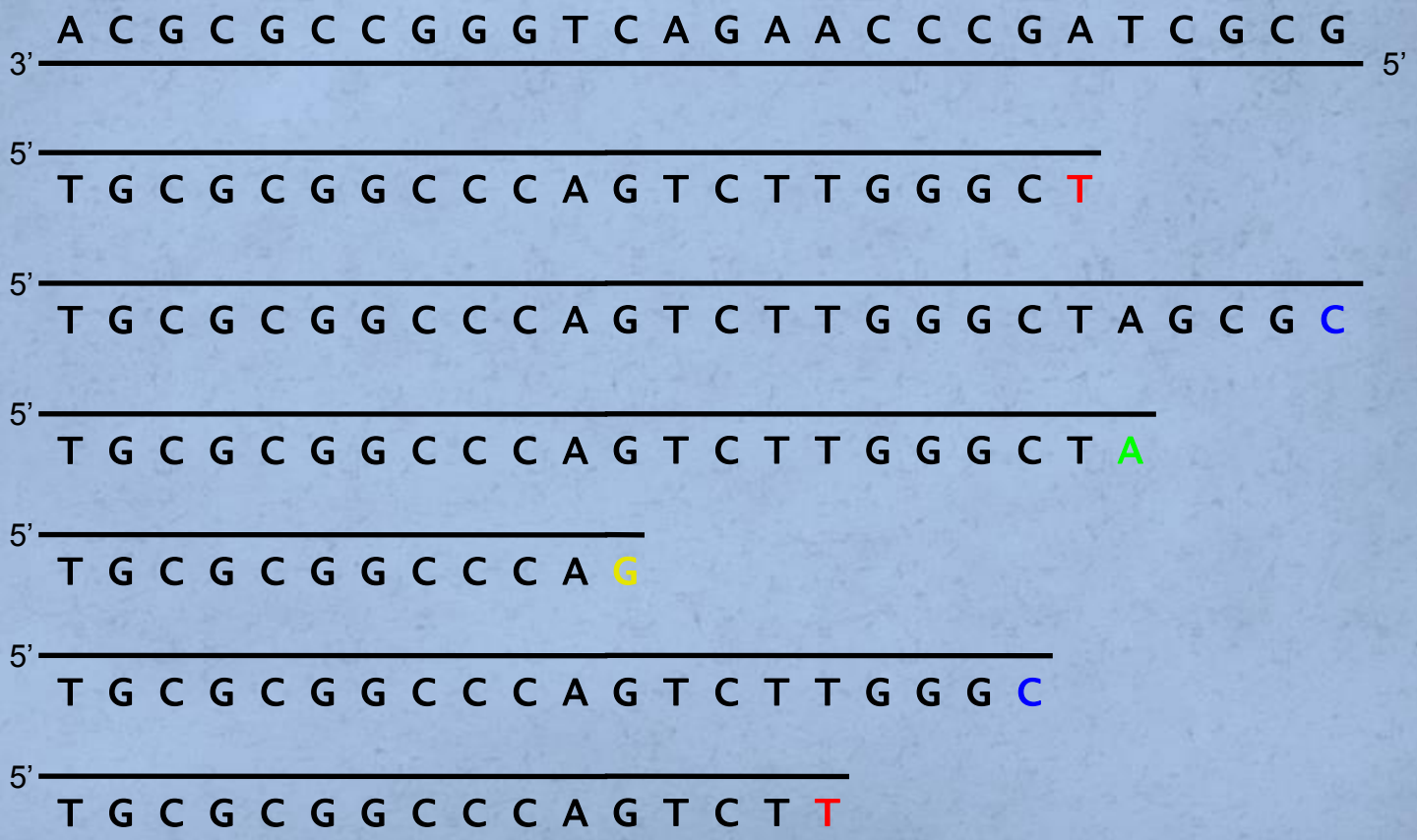
Sanger Sequencing



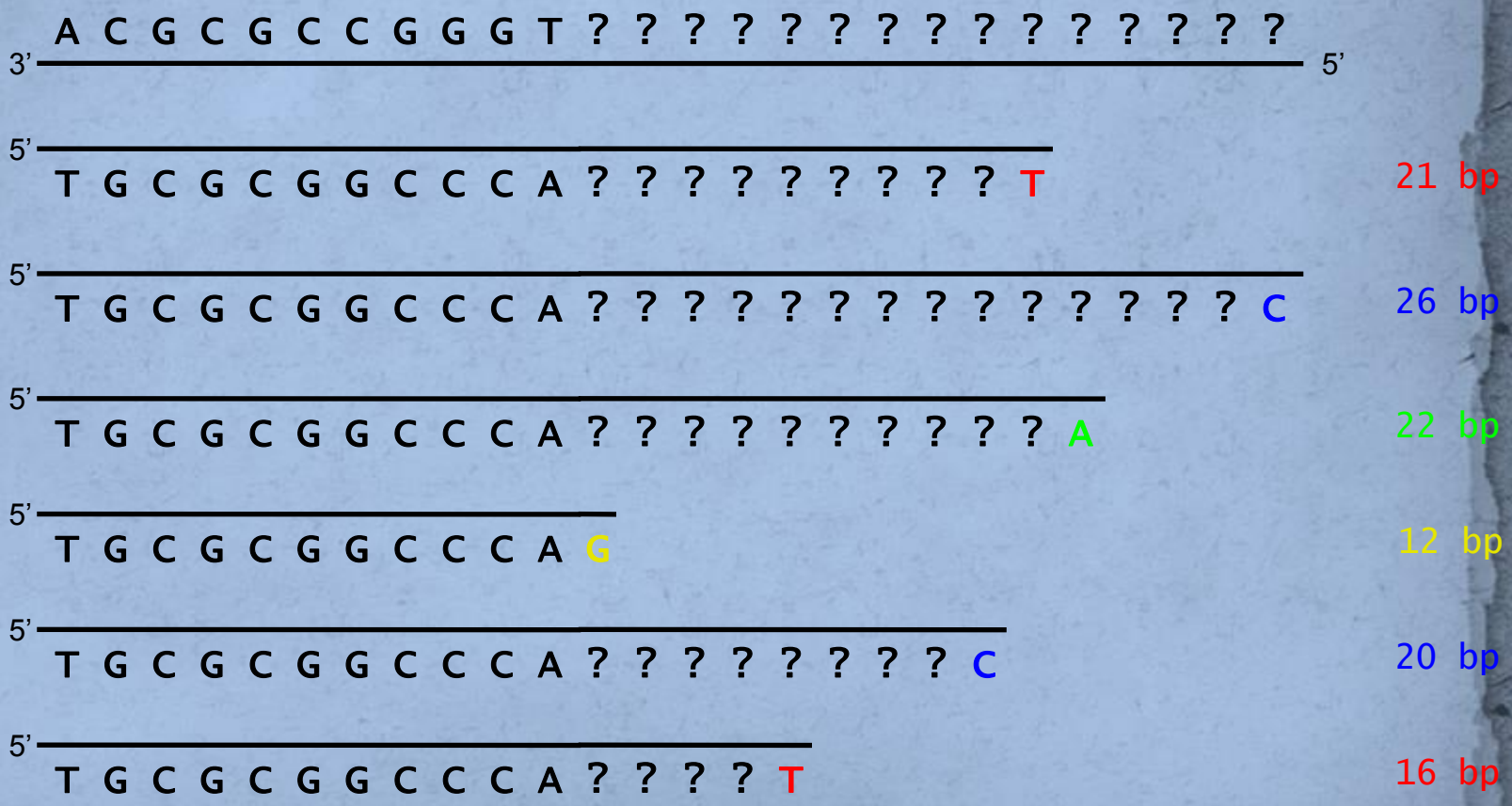
Sanger Sequencing



Sanger Sequencing



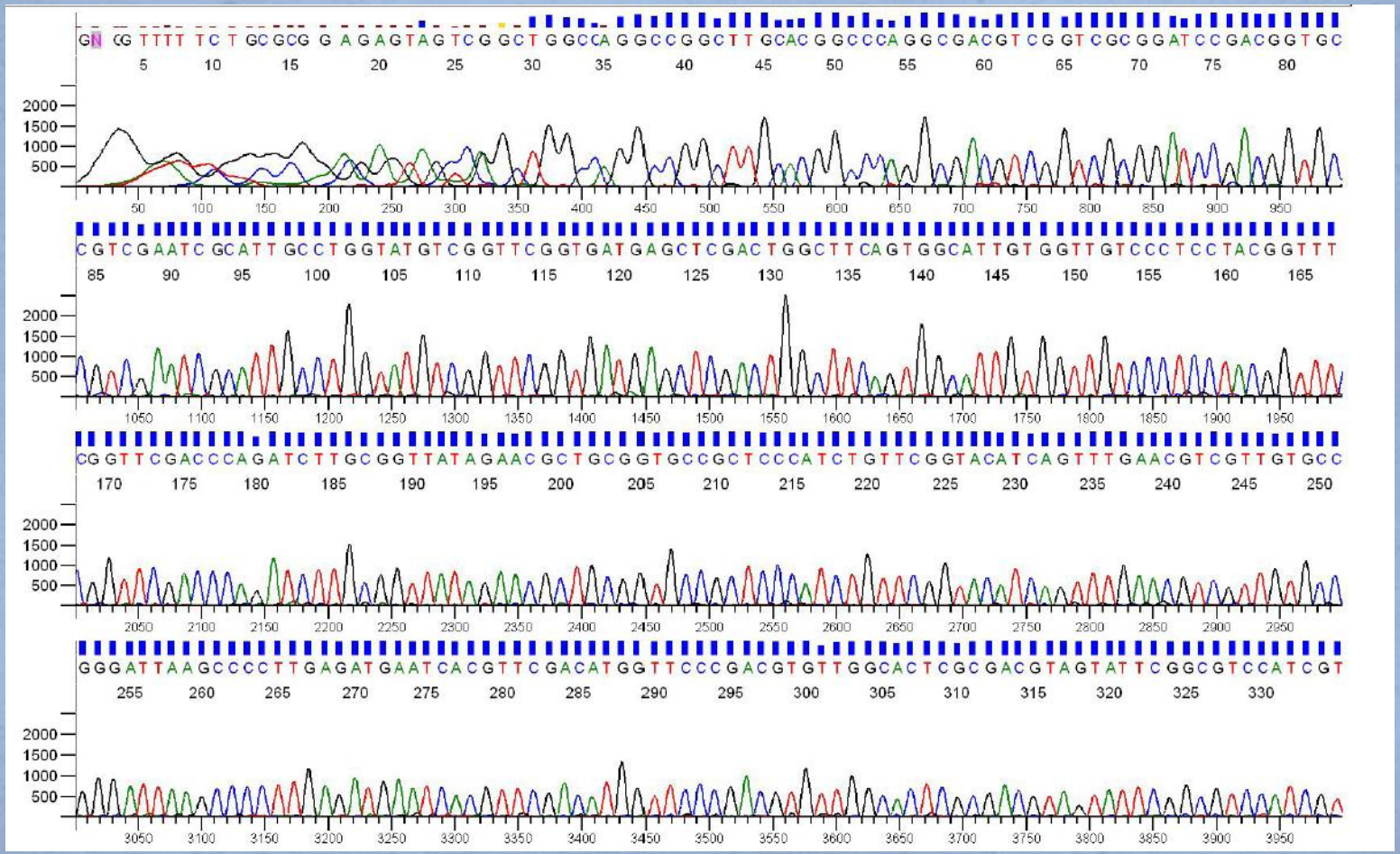
Sanger Sequencing



Sanger Sequencing

Sanger Sequencing Output

Each sequencing reaction gives us a **chromatogram**, usually ~600-1000 bp:



Central Dogma of Molecular Biology

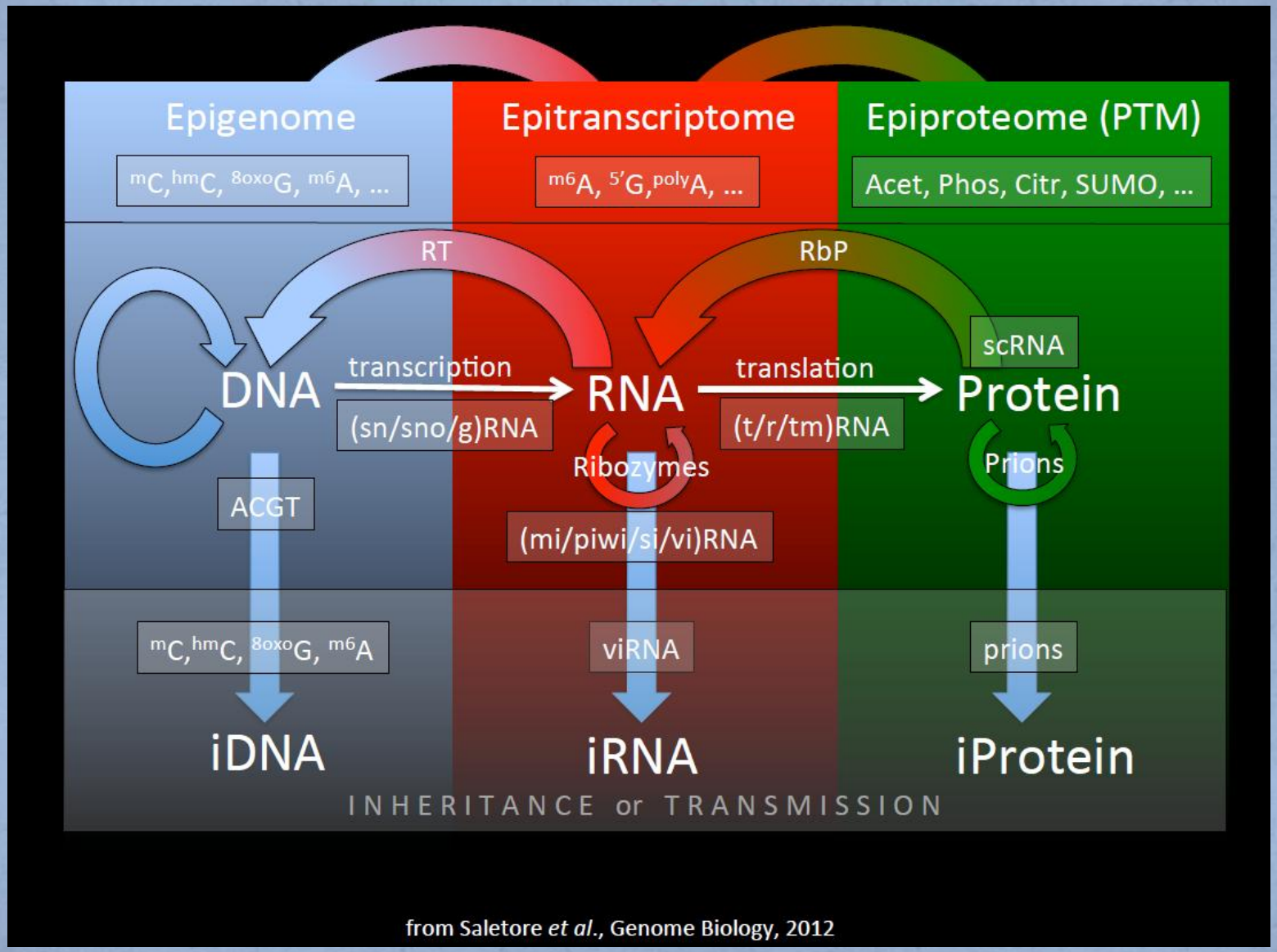
James Watson version - 1965

DNA → RNA → Protein

So once we have the genomic DNA sequence of a species we have all of the information there is?

Really?

- No, not really.



from Saletore *et al.*, Genome Biology, 2012

Overview

- The Past: Sanger
- The Present: **Next-Gen (454, Illumina, …)**
- The Future: ? (Nanopore, MinION, Single-molecule)

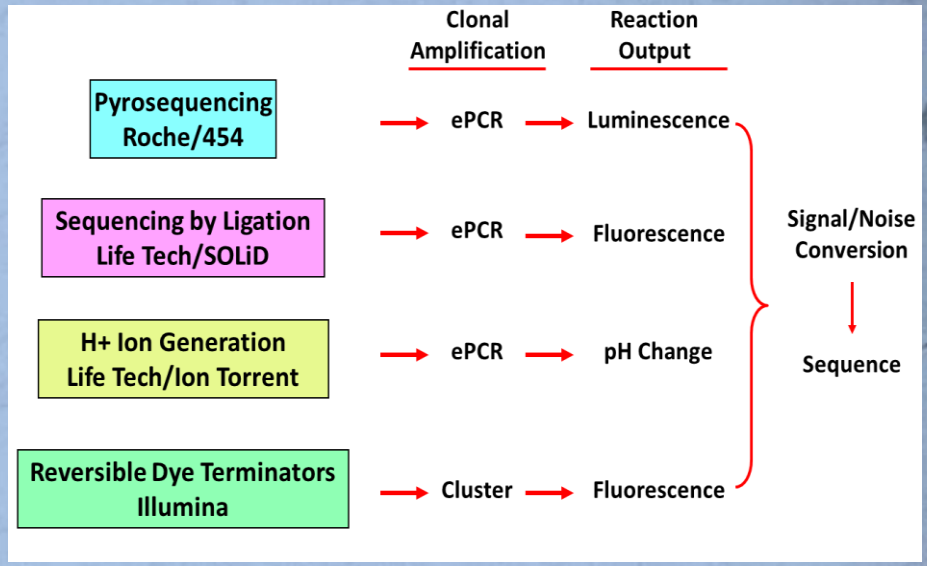
- 454 Sequencing / Roche
  - GS Junior System
  - GS FLX+ System
- Illumina (Solexa)
  - HiSeq System
  - Genome analyzer IIx
  - MySeq
- Applied Biosystems - Life Technologies
  - SOLiD 5500 System
  - SOLiD 5500xl System
- Ion Torrent - Life Technologies
  - Personal Genome Machine (PGM)
  - Proton
- Helicos
  - Helicos Genetic Analysis System
- Pacific Biosciences
  - PacBio RS
- Oxford Nanopore Technologies
  - GridION System
  - MinION

Next Generation Sequencing
Amplified Single Molecule Sequencing

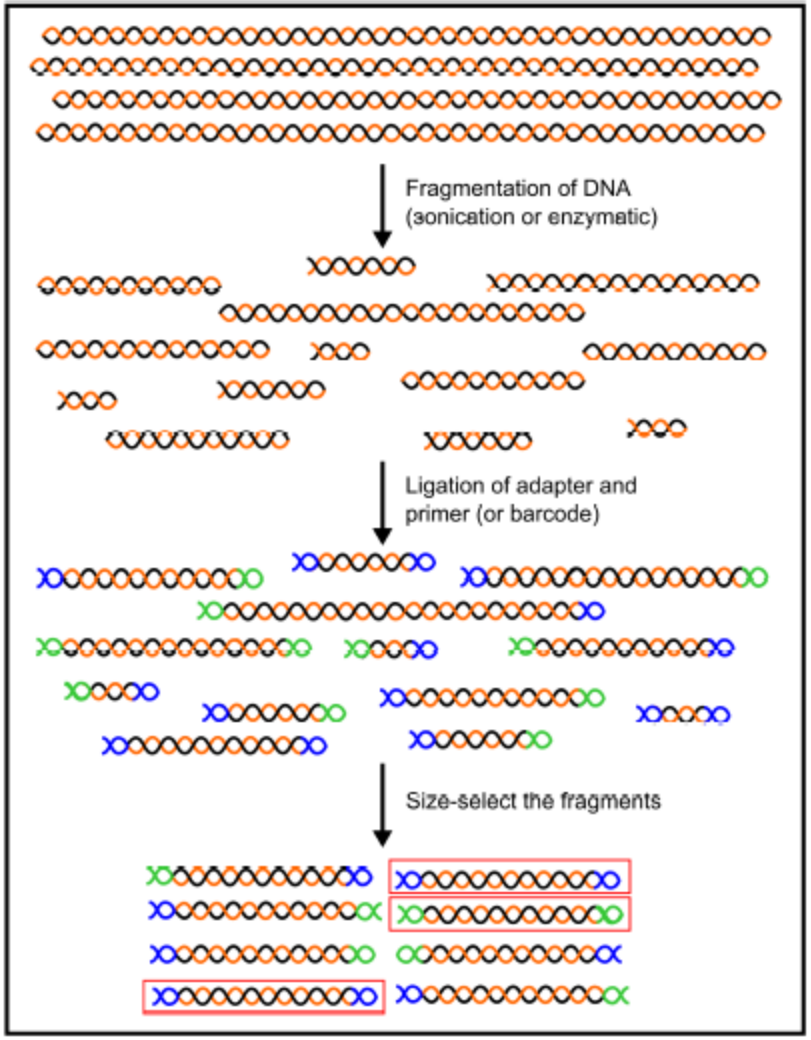
Third Generation Sequencing,
Next Next Generation Sequencing,
Single Molecule Sequencing

NGS Platforms

- Differ in design and chemistries
- Fundamentally related-sequencing of thousands to millions of clonally amplified molecules in a massively parallel manner
- Orders of magnitude more information-will continue to evolve
- Attractive for clinical applications – individual sequencing assays costly and laborious- serial “gene by gene” analysis

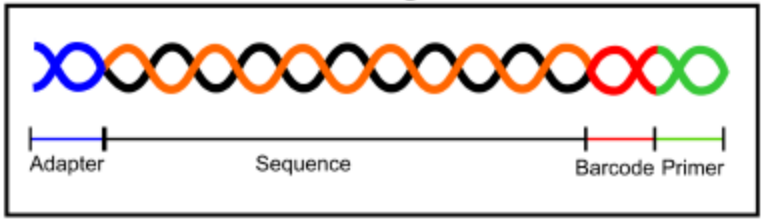


Next Generation Sequencing : Amplified Single Molecule Sequencing

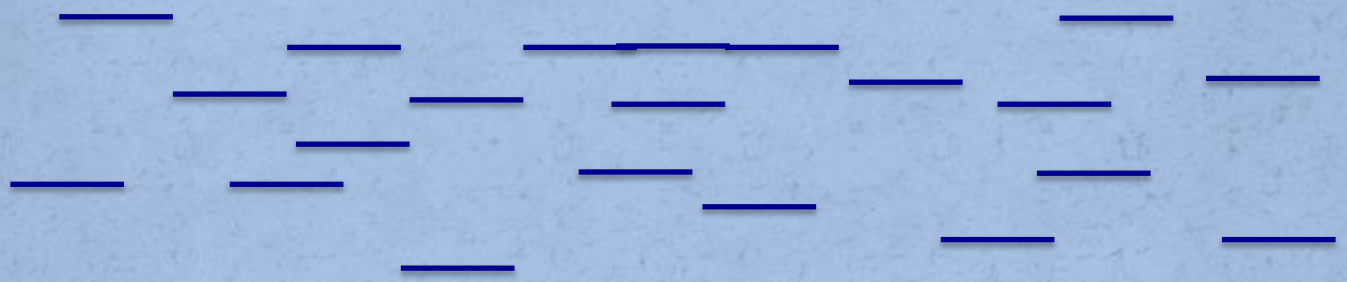


Library preparation

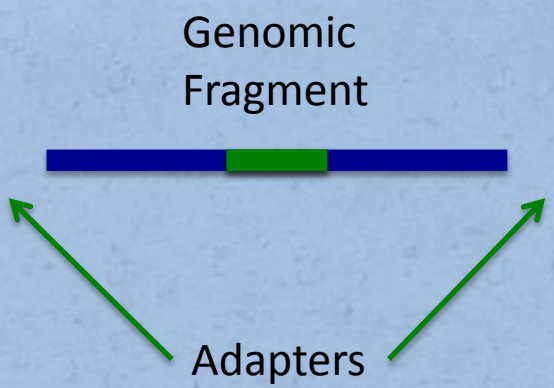
Good fragments :



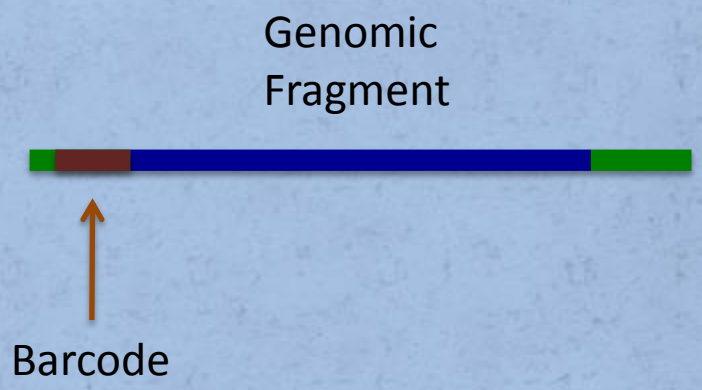
Shotgun sequencing by Ion Torrent Personal Genome Machine and 454



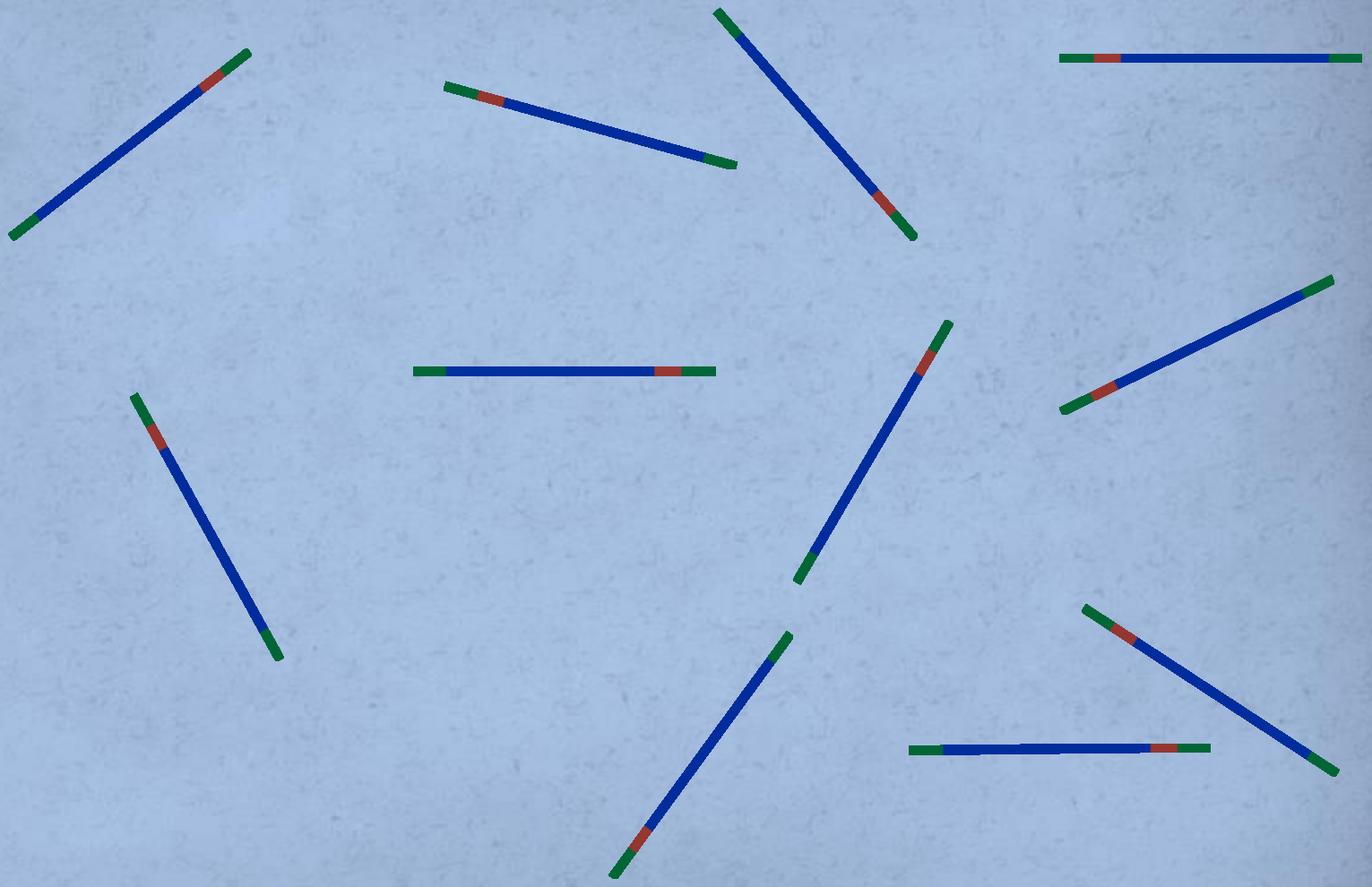
Shotgun sequencing by PGM/454



Shotgun sequencing by PGM/454

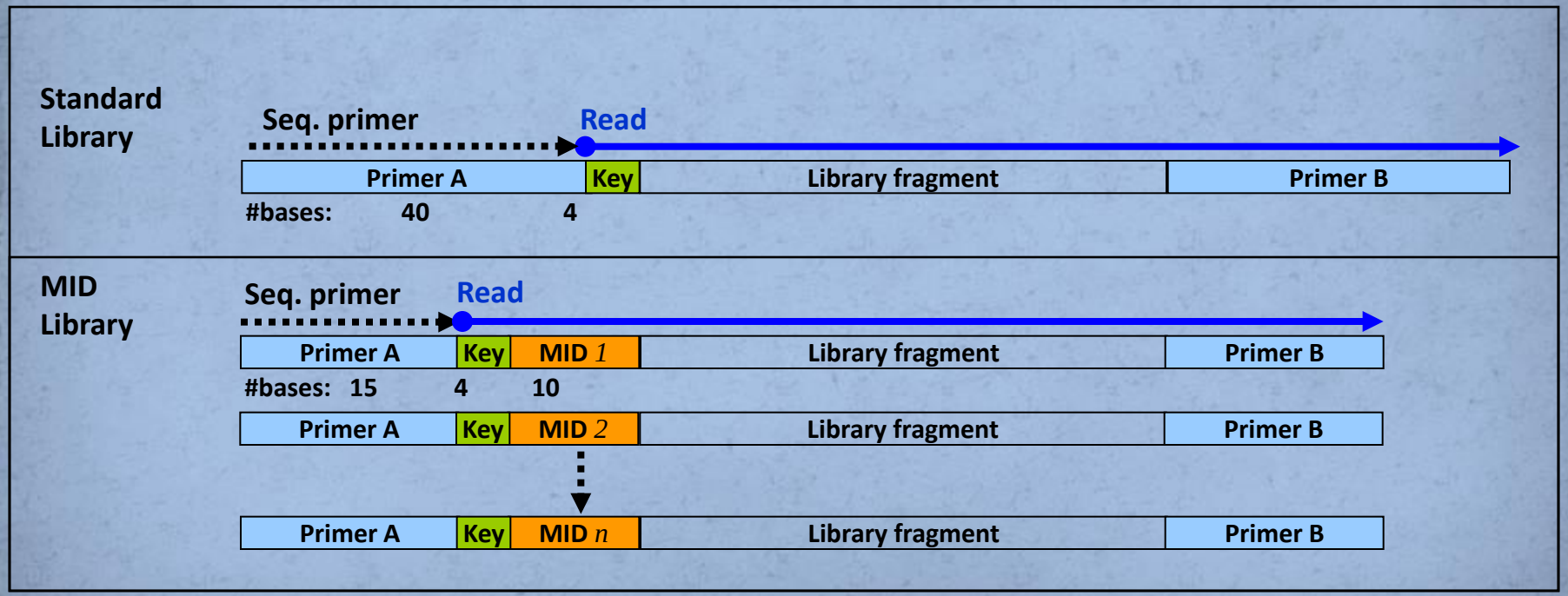


Shotgun sequencing by PGM/454



Multiplex Identifier *Basics*

- What is it?
  - Two **new kits**, each with 6 different library adapters (total of 12 adapters)
  - Each MID library adapter has an added, specially encoded 10-base region
  - Used to "bar-code" up to 12 different genomic library samples to be run in the same region of a single sequencing run



NGS = Next Generation Sequencing

After PCR,
THE new revolution
in Biology ?

NGS Synonym is : **High-throughput Sequencing (HTS)**

First Generation :
SANGER Sequencing

Second Generation :
NGS = Massively Parallel Sequencing

Third Generation :
NGS = HTS, Single Molecule Sequencing

Revolution in Throughput

One run, one sample ..

Capillary Sequencer

1 gene

1 page

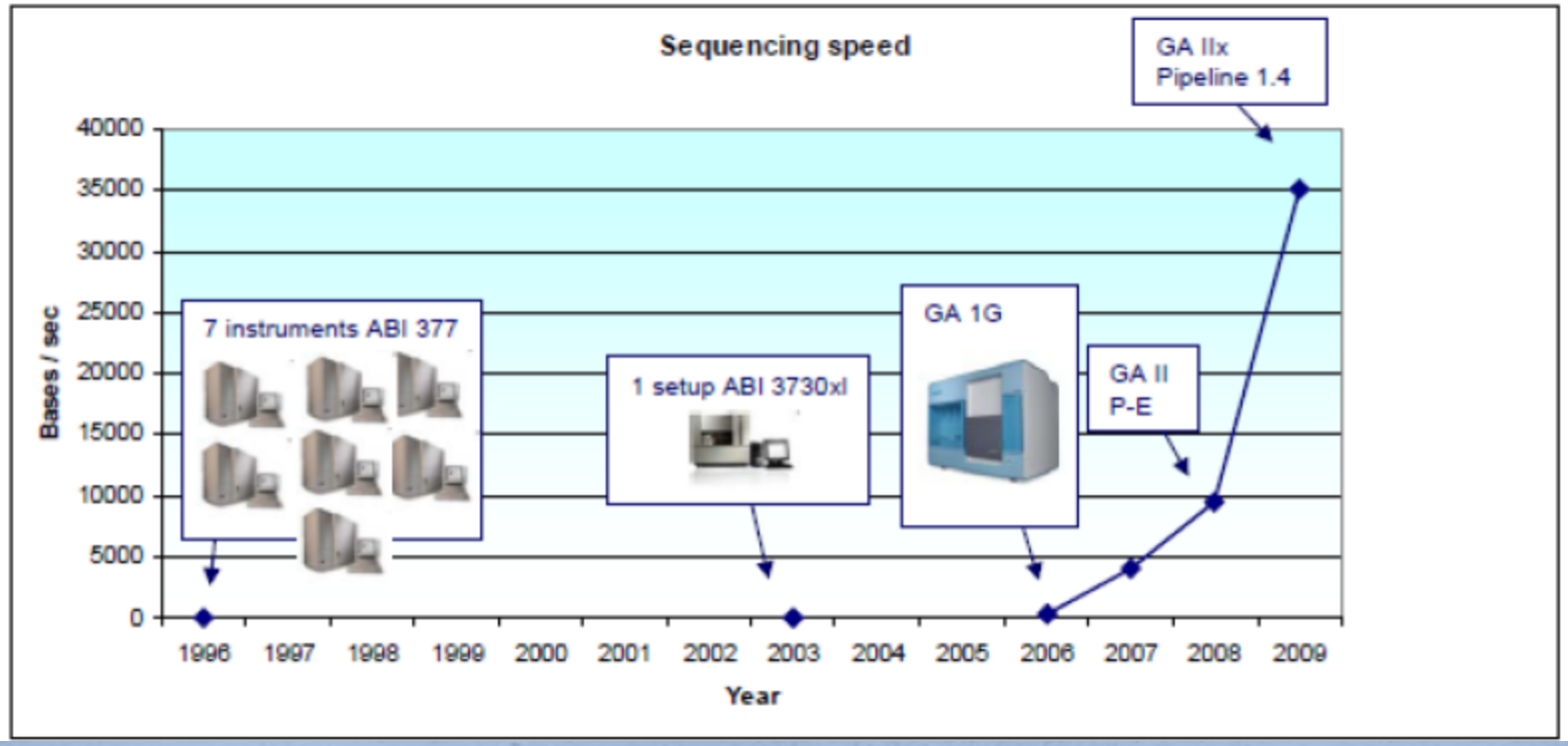
3'000'000x

Illumina Genome Analyzer

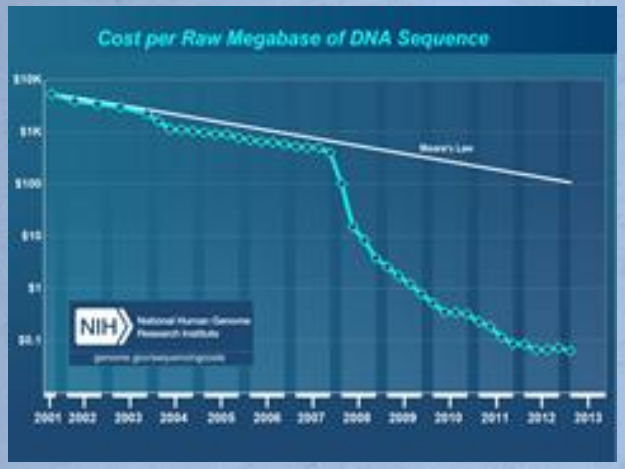
A genome

A library with 12'000 books of 500 pages each

Tremendous Speed Increase



Sequencing Cost

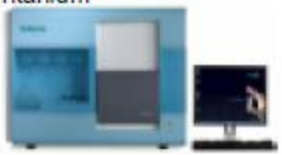


Date	Cost per Mb	Cost per Genome
Sep-01	$5,292.39	$95,263,072
Sep-02	$3,413.80	$61,448,422
Oct-03	$2,230.98	$40,157,554
Oct-04	$1,028.85	$18,519,312
Oct-05	$766.73	$13,801,124
Oct-06	$581.92	$10,474,556
Oct-07	$397.09	$7,147,571
Oct-08	$3.81	$342,502
Oct-09	$0.78	$70,333
Oct-10	$0.32	$29,092
Oct-11	$0.09	$7,743
Oct-12	$0.07	$6,618
Jan-13	$0.06	$5,671

Source - NHGRI : http://www.genome.gov/sequencingcosts/

What's Next ?

Roche, 454 GS-FLX Titanium



Illumina, GA2

Applied BioSys, Solid v3

Second Generation

NGS = Massively Parallel Sequencing (polony sequencing)

PACIFIC BIOSCIENCES

Helicos BioSciences Corporation

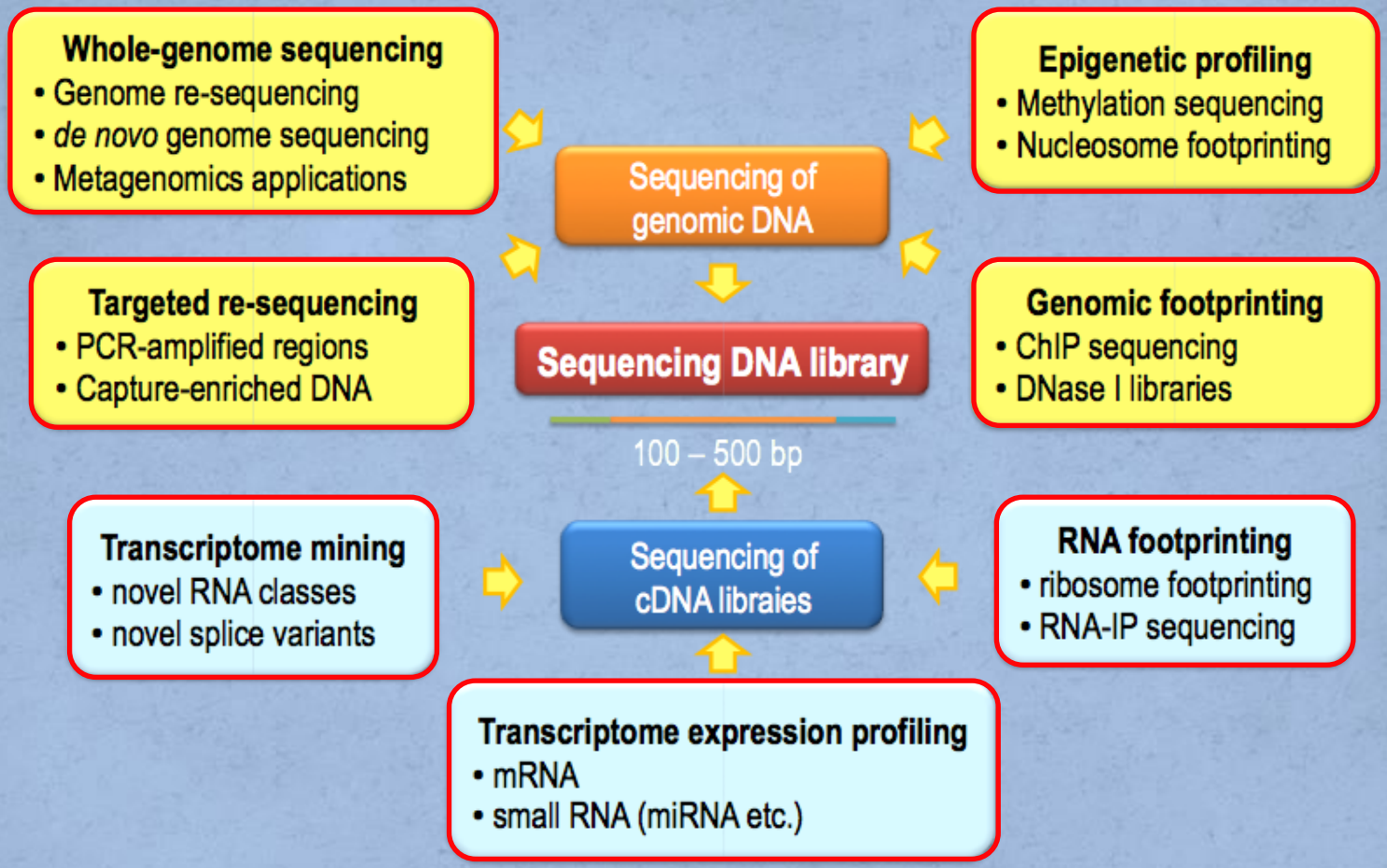
ion torrent

Oxford NANOPORE Technologies

Third Generation :

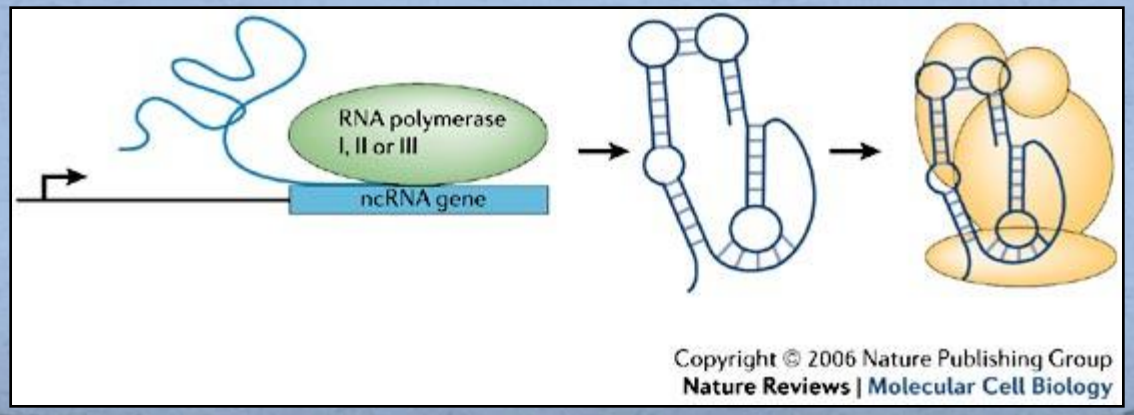
- Single Molecule Sequencing (no bias)
- Faster
- Cheaper (or not)
- 1000€ Human genome ?

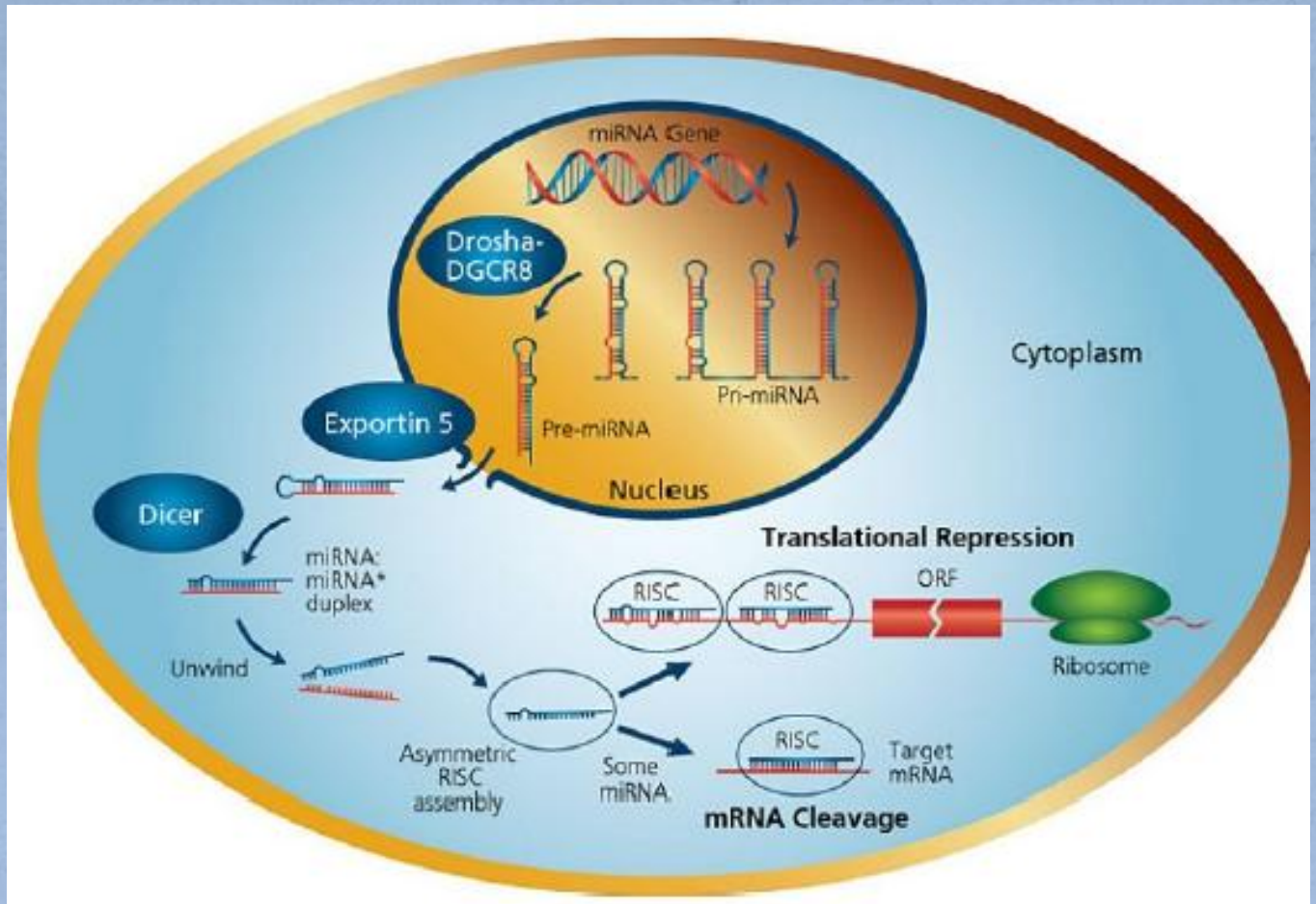
Applications of Next-Generation Sequencing



Discovering noncoding RNAs

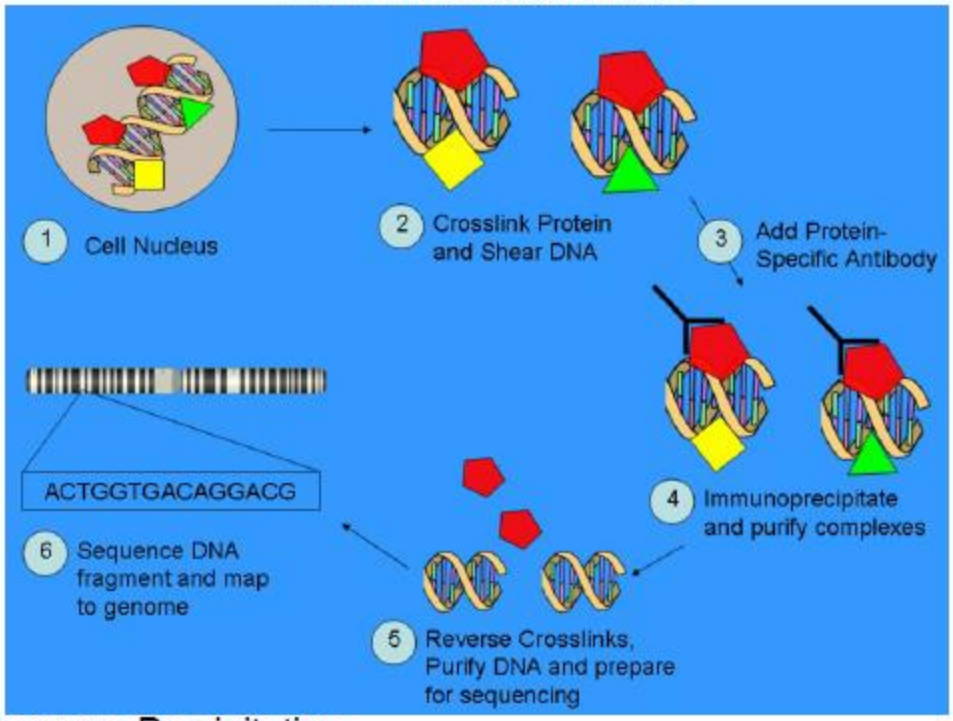
- ncRNA presence in genome difficult to predict by computational methods with high certainty because the evolutionary diversity
- Detecting expression level changes that correlate with changes in environmental factors, with disease onset and progression, complex disease set or severity
- Enhance the annotation of sequenced genomes (impact of mutations more interpretable)





microRNA is approx 21 bases long

- Genome Sequencing
  - De novo sequencing
  - Resequencing
  - Targeted (re)sequencing
  - Mitochondrial sequencing
  - Mutation detection
  - Amplicon sequencing
  - Amplicon Cancer Panel
  - ...
- Transcript Expression Profiling
  - RNA sequencing
  - miRNA sequencing
  - Deep-SAGE
  - Deep-CAGE
  - PAS : polyadenylation site
  - ...
- Transcription factor binding
  - ChIP sequencing : Chromatin Immuno Precipitation
- Structural variation
- Metagenomics
- ...



Metagenomics

- Characterizing the biodiversity found on Earth
- The growing number of sequenced genomes enables us to interpret partial sequences obtained by direct sampling of specif environmental niches.
- Examples: ocean, acid mine site, soil, coral reefs, human microbiome which may vary according to the health status of the individual

THE METAGENOMICS PROCESS

Extract all DNA from microbial community in sampled environment

DETERMINE WHAT THE GENES ARE
(Sequence-based metagenomics)

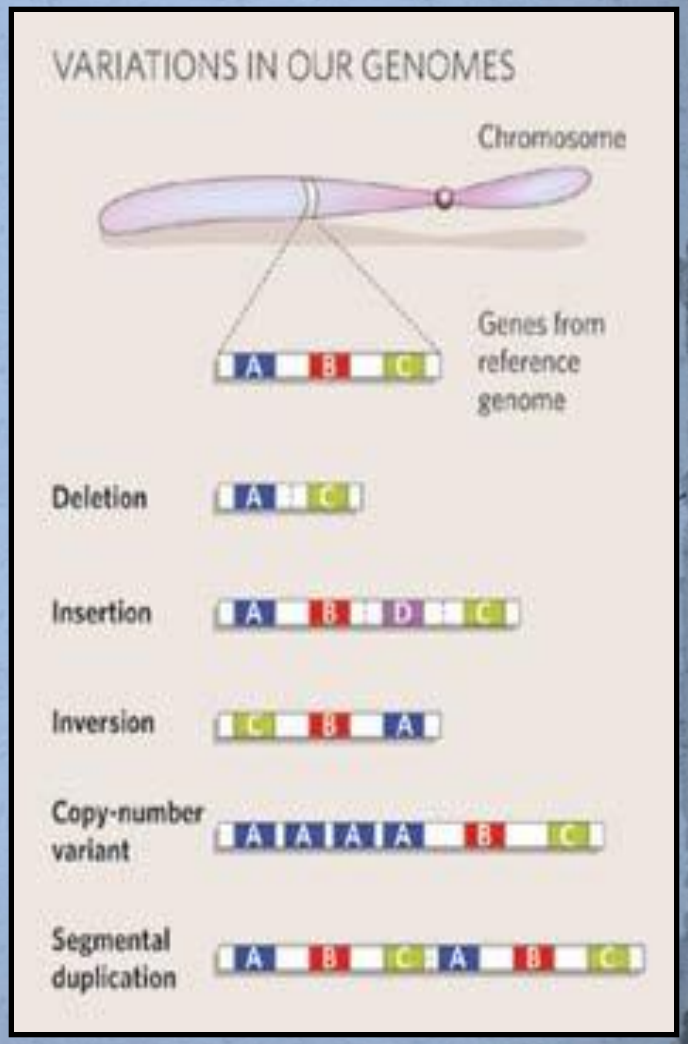
- Identify genes and metabolic pathways
- Compare to other communities
- and more…

DETERMINE WHAT THE GENES DO
(Function-based metagenomics)

- Screen to identify functions of interest, such as vitamin or antibiotic production
- Find the genes that code for functions of interest
- and more…

Defining variability in many human genomes

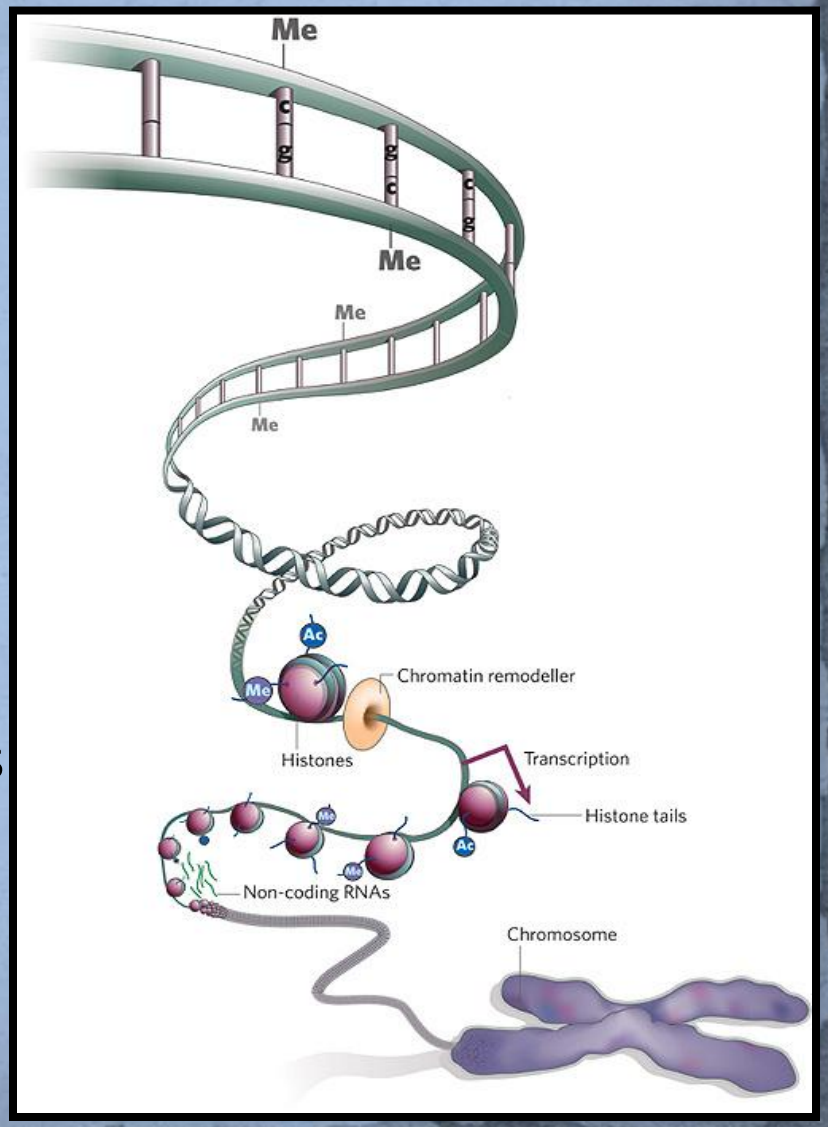
- Common variants have not yet completly explained complex disease genetics →rare alleles also contribute
- Also structural variants, large and small insertions and deletions
- Accelerating biomedical research



Epigenomic variation

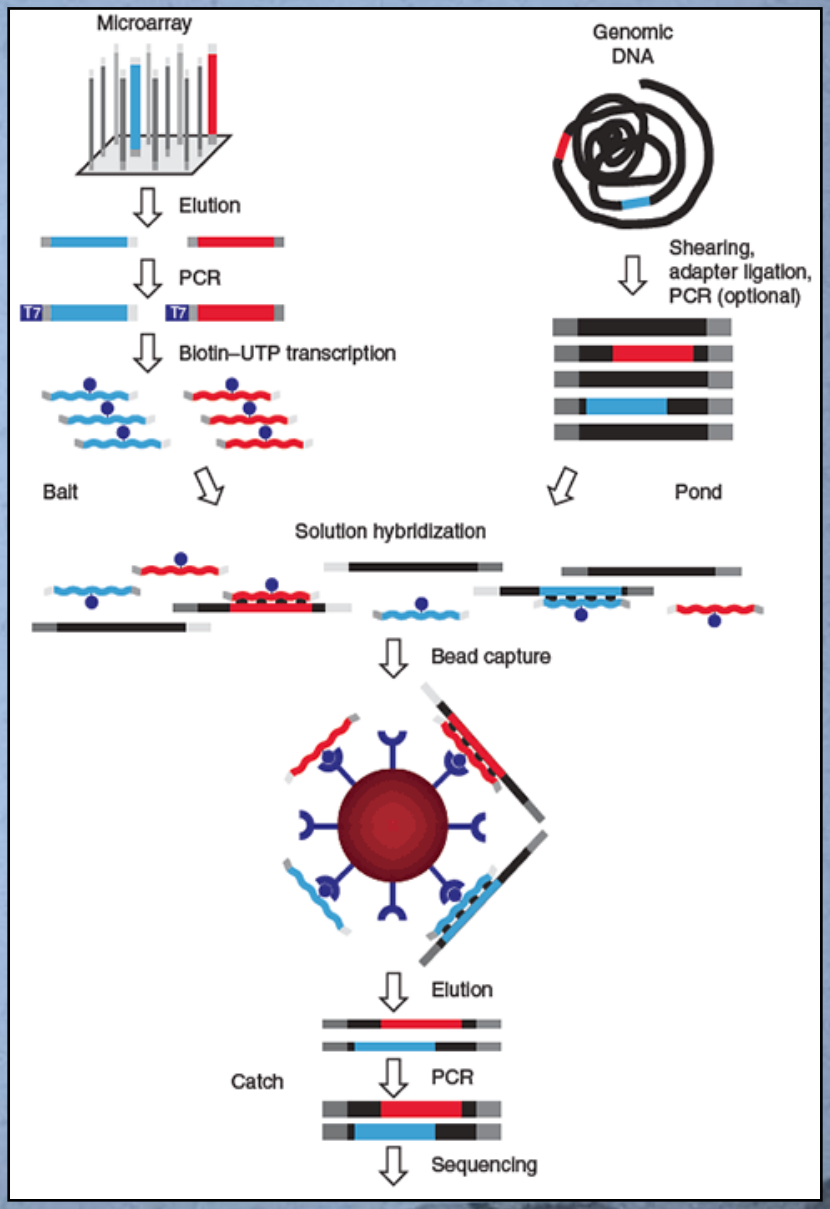
- Enable of genome-wide patterns of methylation and how this patterns change through the course of an organism's development.

- Enhanced potential to combine the results of different experiments, correlative analyses of genome-wide methylation, histone binding patterns and gene expression, for example.



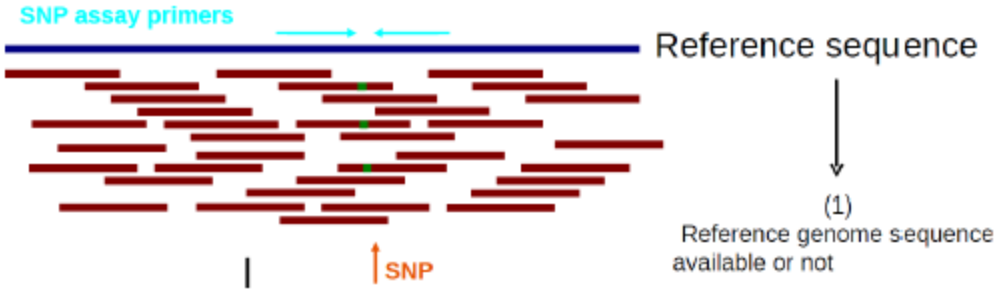
Mutation discovery

- Extreme example: multiplexing the amplification of 10 000 human exons using primers from a programmable microarray and sequencing them using NGS.



- Genome Sequencing
  - De novo sequencing
  - Resequencing
  - Targeted (re)sequencing
  - Mitochondrial sequencing
  - Mutation detection
  - Amplicon sequencing
  - Amplicon Cancer Panel
  - ...
- Transcript Expression Profiling
  - RNA sequencing
  - miRNA sequencing
  - Deep-SAGE
  - Deep-CAGE
  - PAS : polyadenylation site
  - ...
- Transcription factor binding
  - ChIP sequencing
- Structural variation
- Metagenomics
- ...

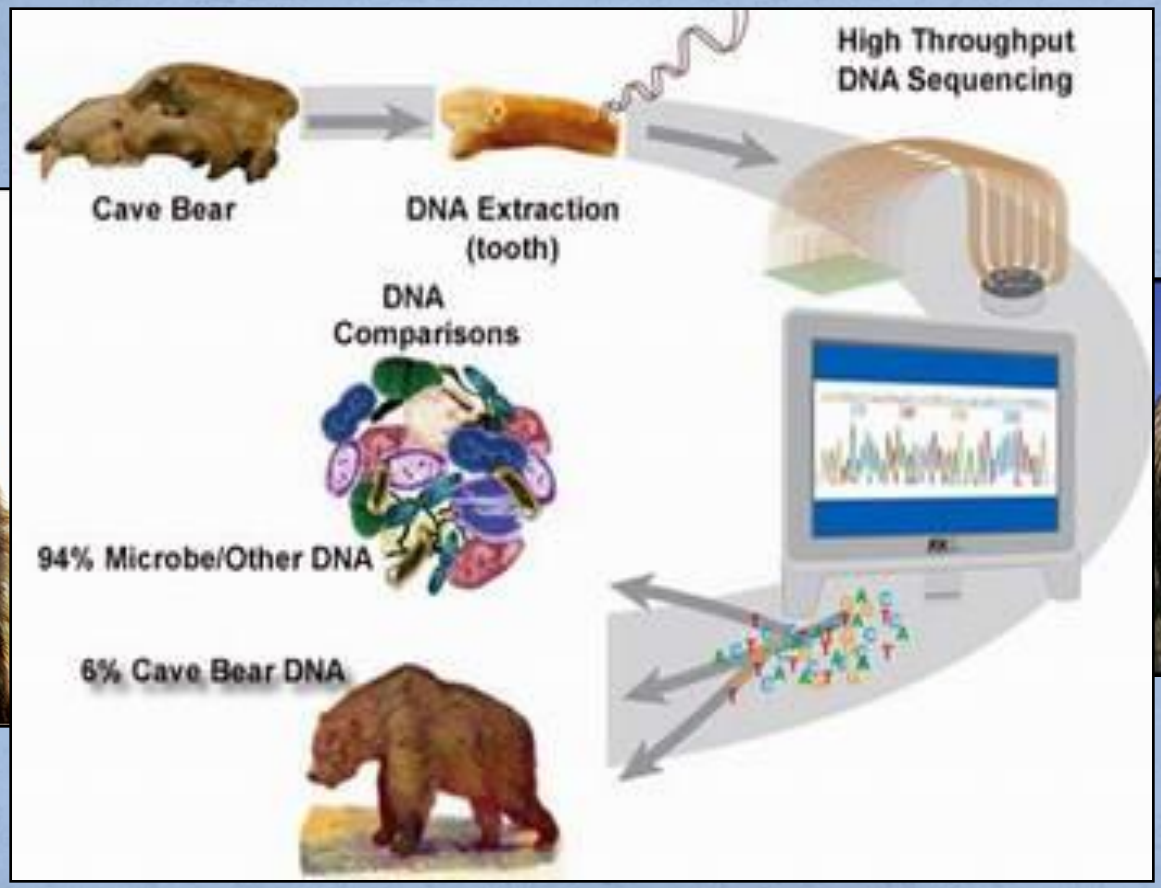
SNP identification: Principle




```
Individual_1_haplo1  AACGATTATCGCAATAACGAGGATTATCCCAGTTA
Individual_1_haplo2  AACGATTATCGCAATGACGAGGATTATCTCAGTTA

Individual_2_haplo1  AACGACTATCGCAATAACGAGGATTATCCCAATTA
Individual_2_haplo2  AACGATTATCGCAATAACGAGGATTATCCCAGTTA

Individual_3_haplo1  AACGACTATCGCAATAACGAGGATTATCCCAATTA
Individual_3_haplo2  AACGATTATCGCAATGACGAGGATTATCTCAGTTA

Individual_4_haplo1  AACGATTATCGCAATAACGAGGATTATCCCAGTTA
Individual_4_haplo2  AACGATTATCGCAATGACGAGGATTATCTCAGTTA
```

Ancient Genomes Resurrected

- Degraded state of the sample → mitDNA sequencing
- Nuclear genomes of ancient remains: cave bear, mommoth, Neanderthal (10^6 bp)

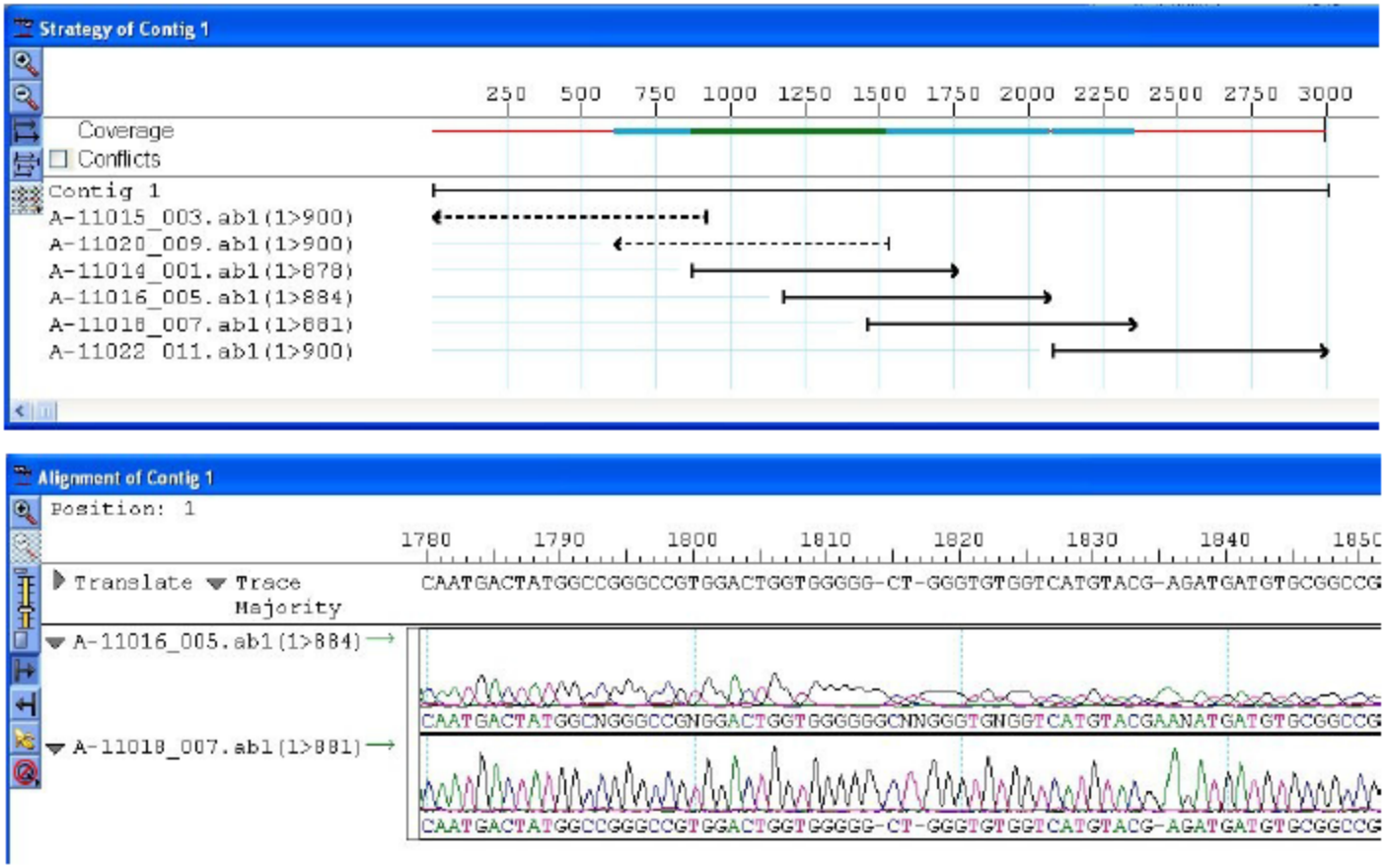


Problems: contamination modern humans and coisolation bacterial DNA


Oh, my God!
What should I do now?
FASTA
DELIVERY SERVICE
NGS machnies
Massive amount
of sequence data

Data analyses

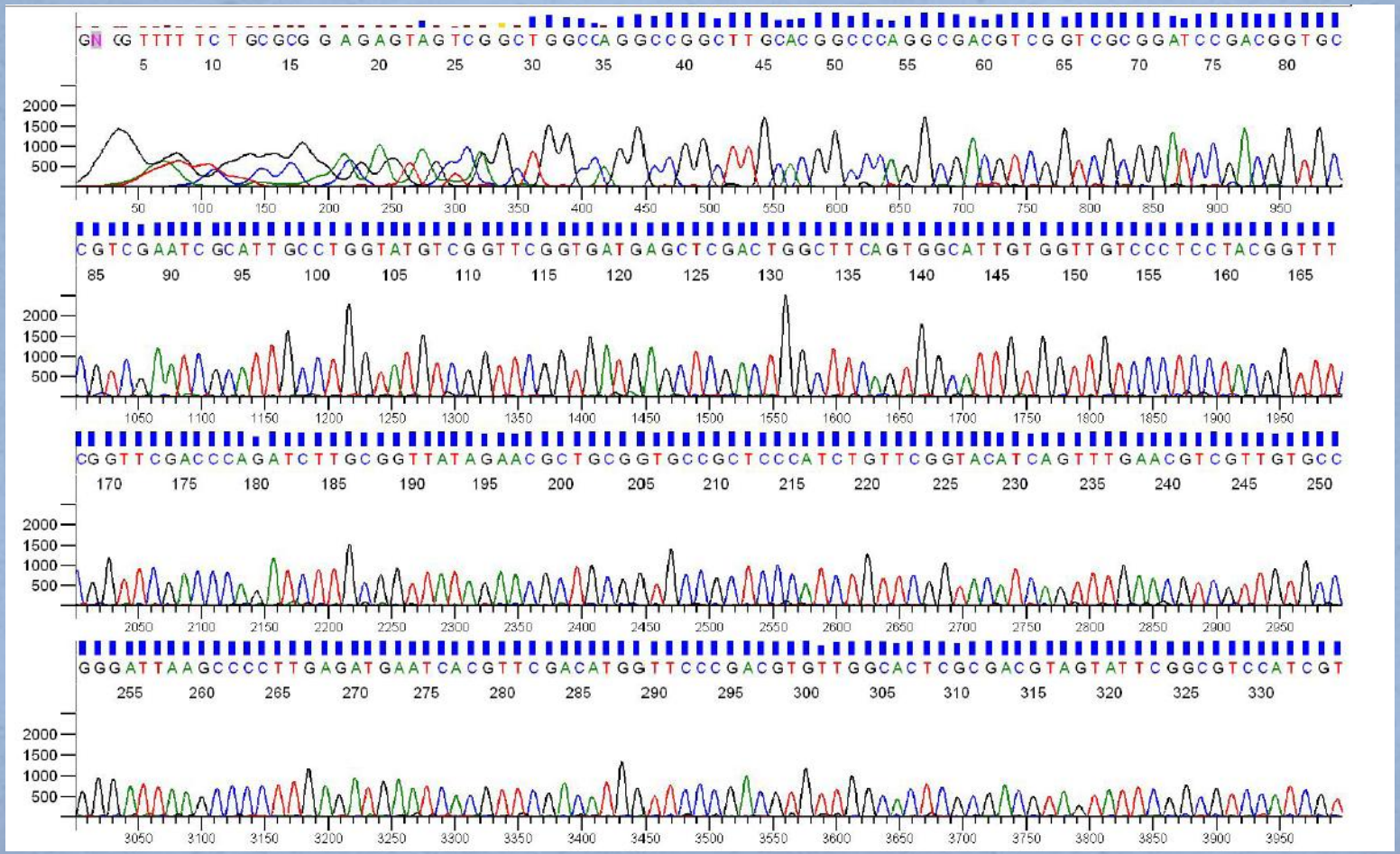
Sanger sequencing : e.g. : one gene sequenced with 6 primers



Manually check the assembly and correct errors.
2000 bp takes 5-10 minutes.

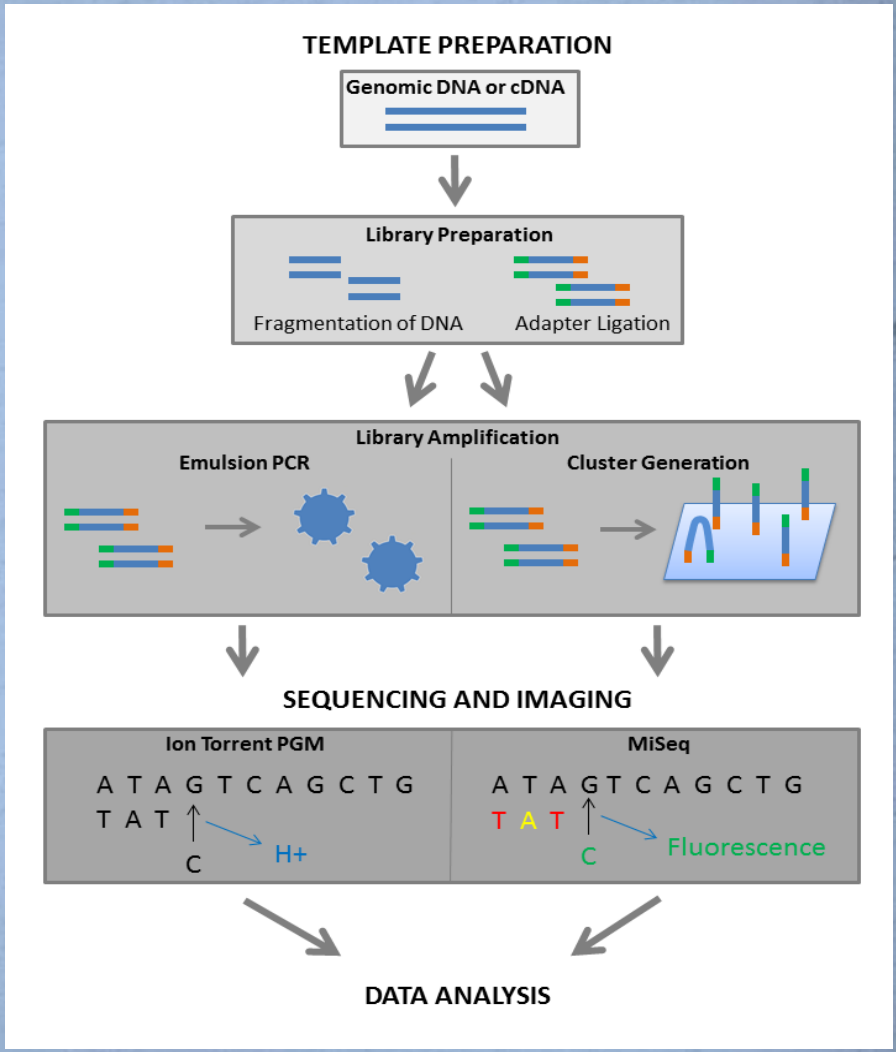
Sanger Sequencing Output

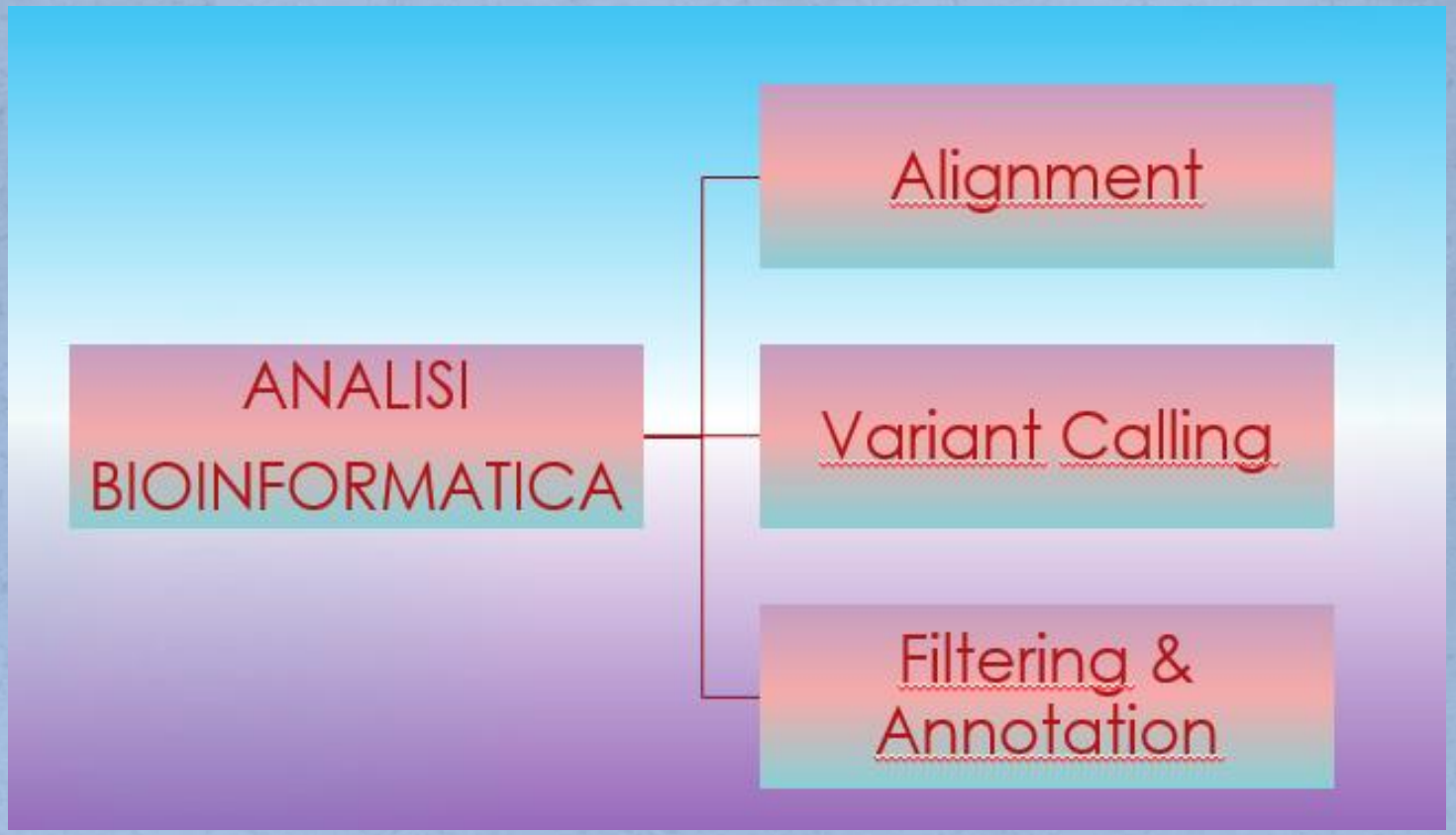
Each sequencing reaction gives us a **chromatogram**, usually ~600-1000 bp:



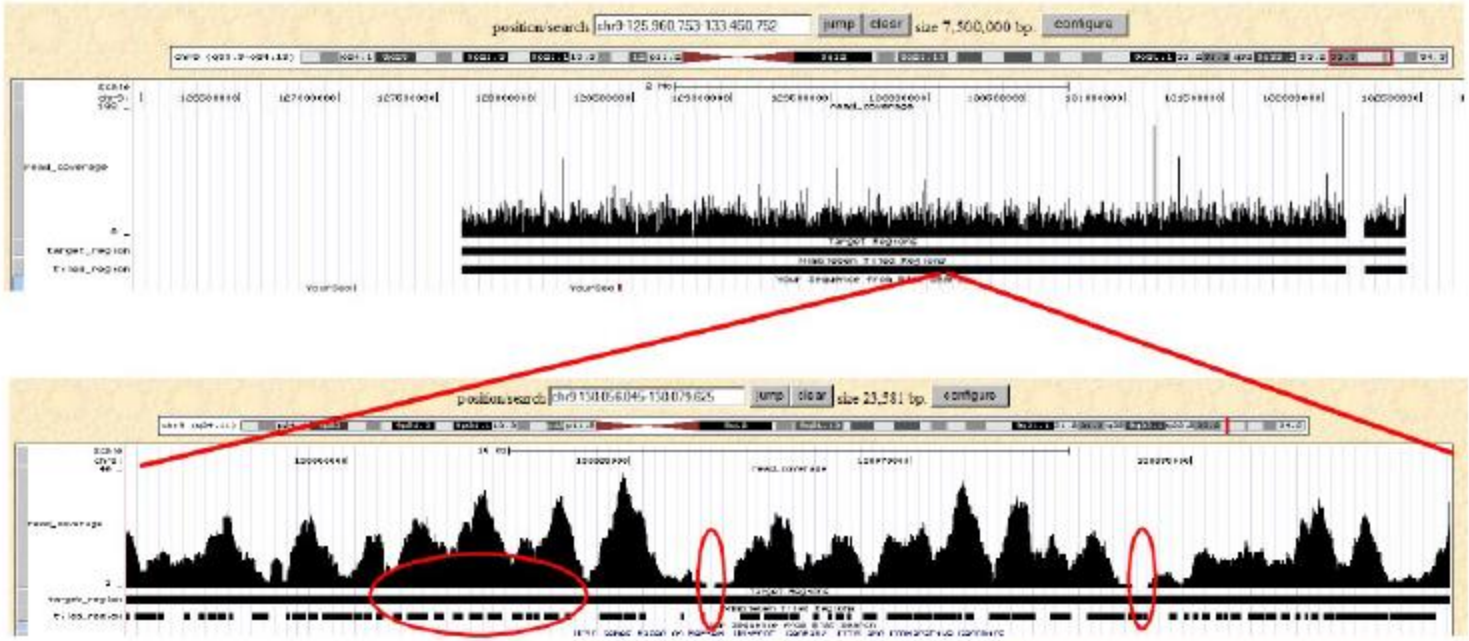
Data Analysis

- Raw sequence data must undergo several analysis steps
  - Preprocessing to remove adapter sequences and low-quality reads
  - Alignment to a reference sequence or de novo alignment
  - Analysis of compiled sequence

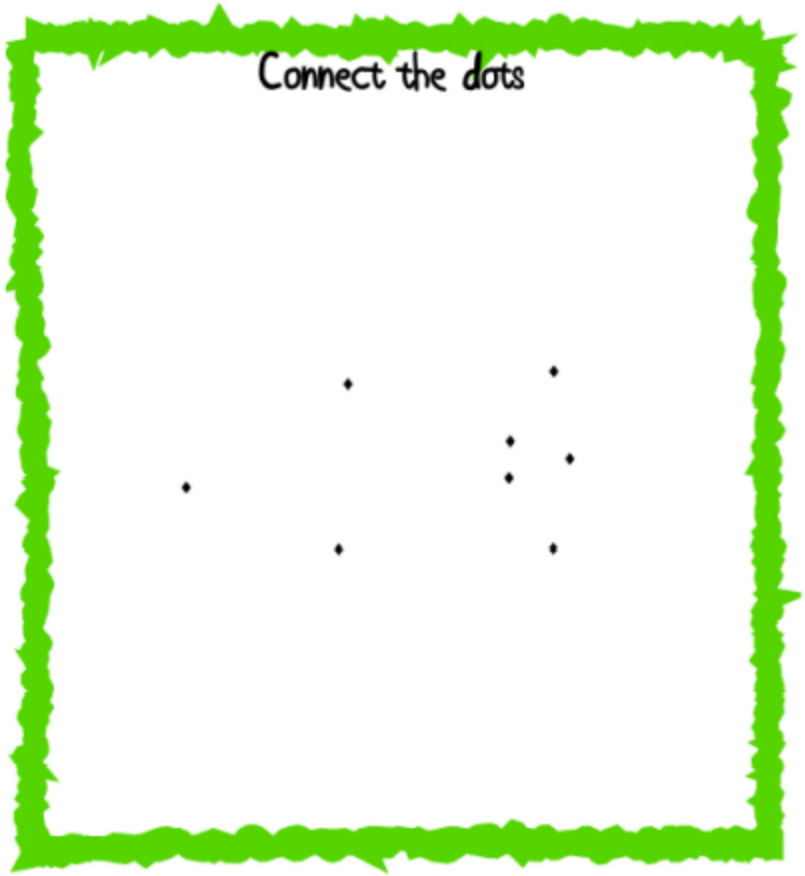



ANALISI BIOINFORMATICA
Alignment
Variant Calling
Filtering & Annotation

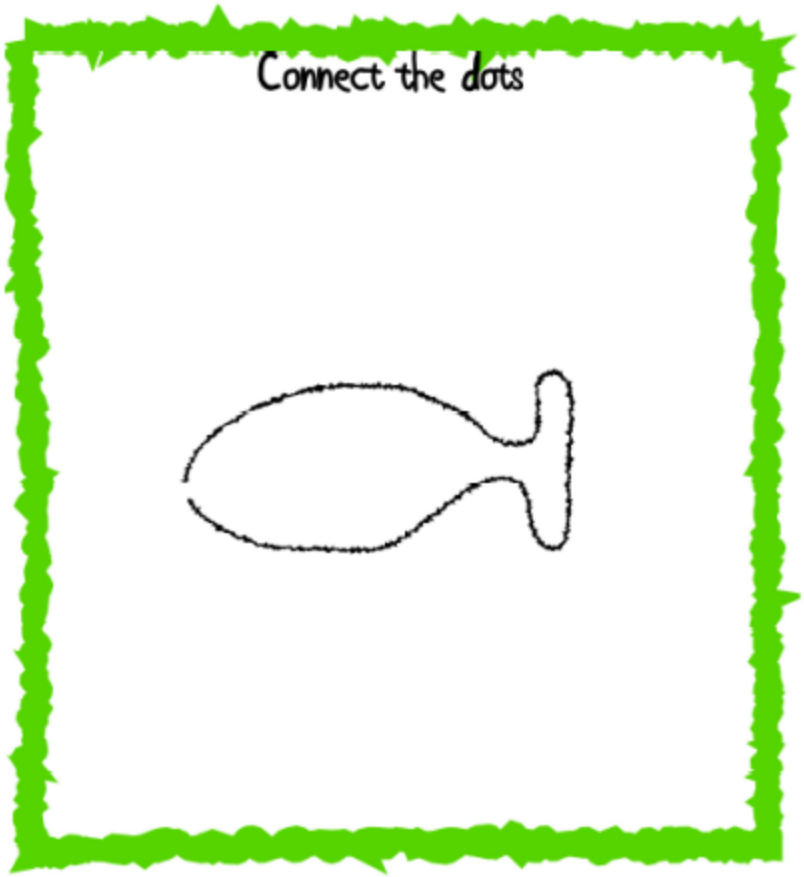
Next Generation sequencing : Impossible to check manually



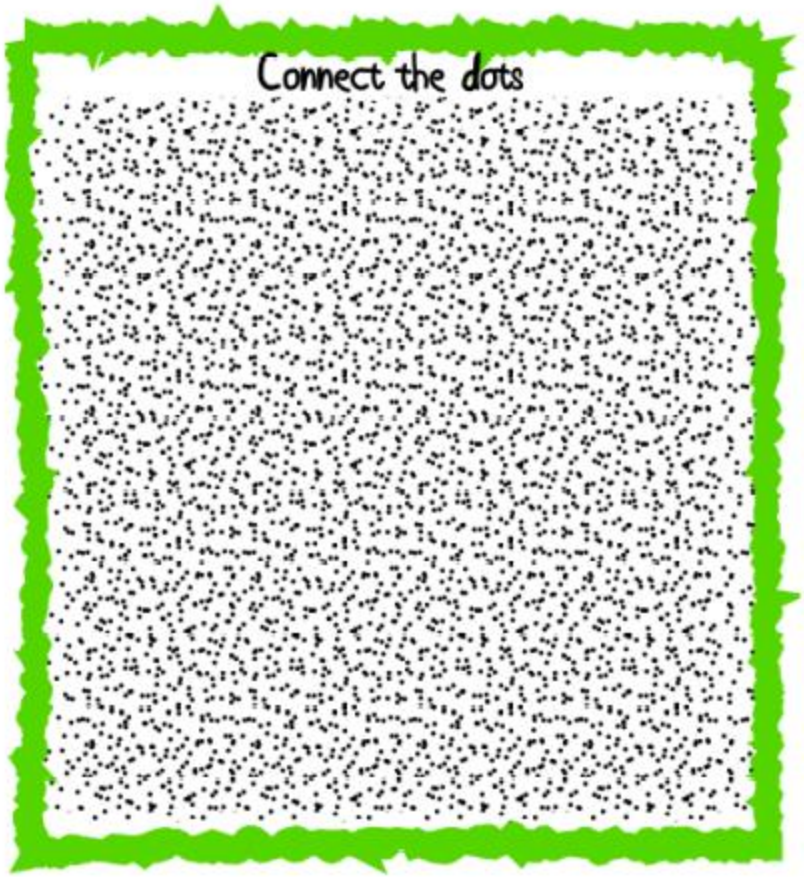
Sanger sequencing : simplified :



Sanger sequencing : simplified :

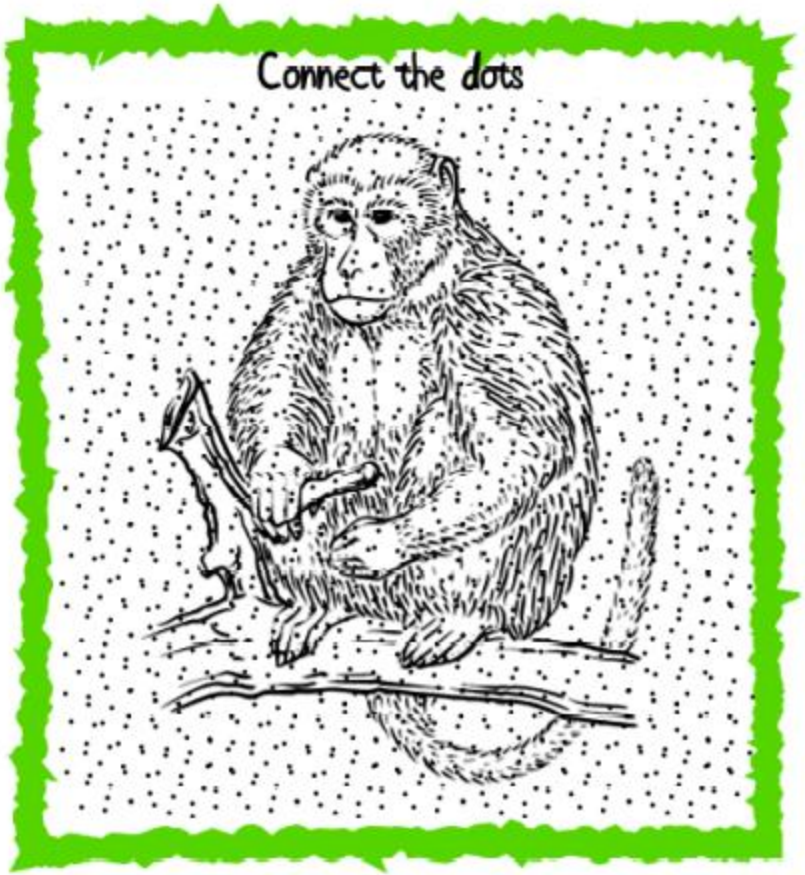


Next Generation sequencing : simplified :

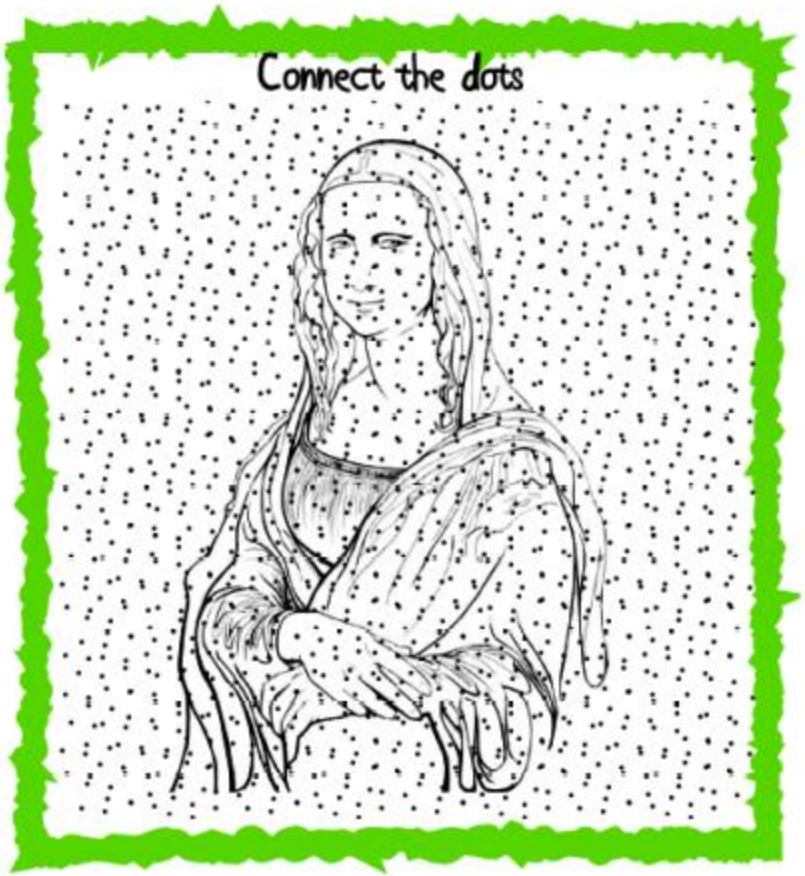


Impossible to assemble manually

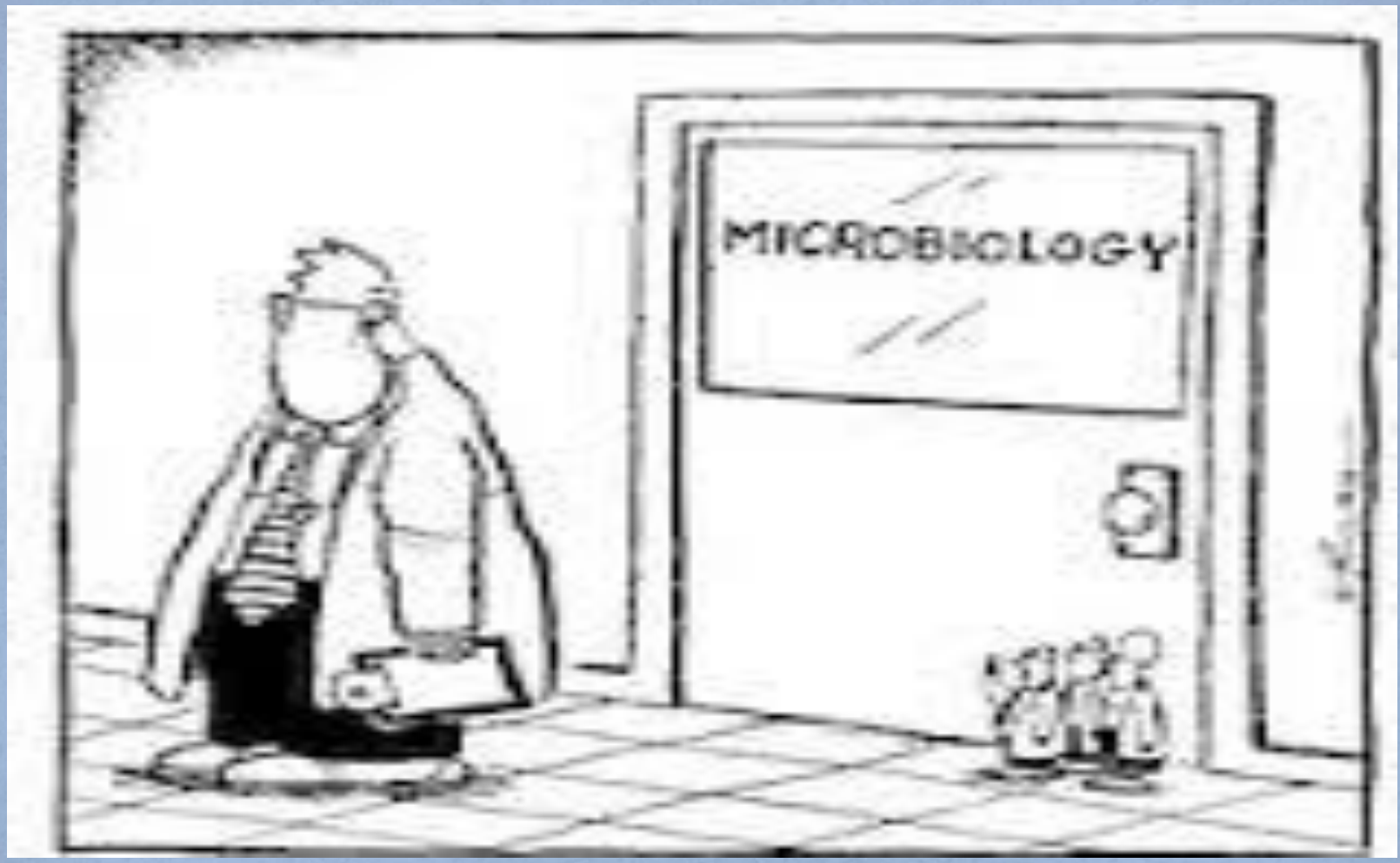
Next Generation sequencing : simplified :



Next Generation sequencing : simplified :



Same dataset, different parameters


MICROBIOLOGY

6th Infectivology Today®

L'infettivologia del 3° millennio: AIDS ed altro

VI Convegno Nazionale

15- 16 -17 maggio 2014

Who's next

1971

Centro Congressi Hotel Ariston

Paestum (SA)