# Next Generation Sequencing

*Technologies, Application,*
*Hardware, Software*

# Outline

- Introduction, Presentation, Goals.
- Next generation sequencing technologies.
- Bioinformatics challenges.
- Some aspects of NGS data analysis.
- Conclusions and perspectives

# Introduction

# Why is NGS revolutionary?

- NGS has brought high speed not only to genome sequencing and personal medicine,
- it has also changed the way we do genome research
- Got a question on genome organization?
- SEQUENCE IT !!!

Vall d'Hebron
Institut de Recerca

# *Any DNA can be sequenced*

M Tuberculosis

S. cervisiae

C. elegans

Barley

Arabidopsis

Maccaca

Neanderthal

HIV

Tomato

Potato

Honeybee

Mammut

H5N1

James Watson

Grape wine

# *Sequenced genomes*

*Nature Reviews Genetics* 9, 303-313, 2008

Over the past years the genomes of some of the most important model organisms have been sequenced:

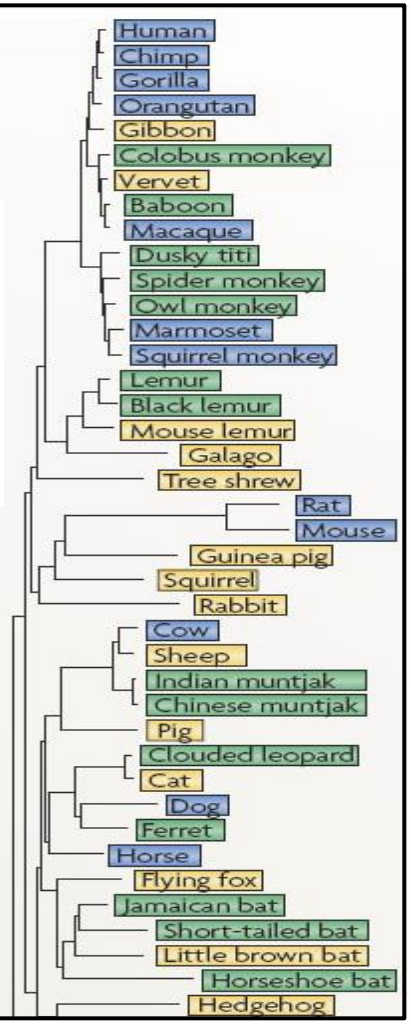
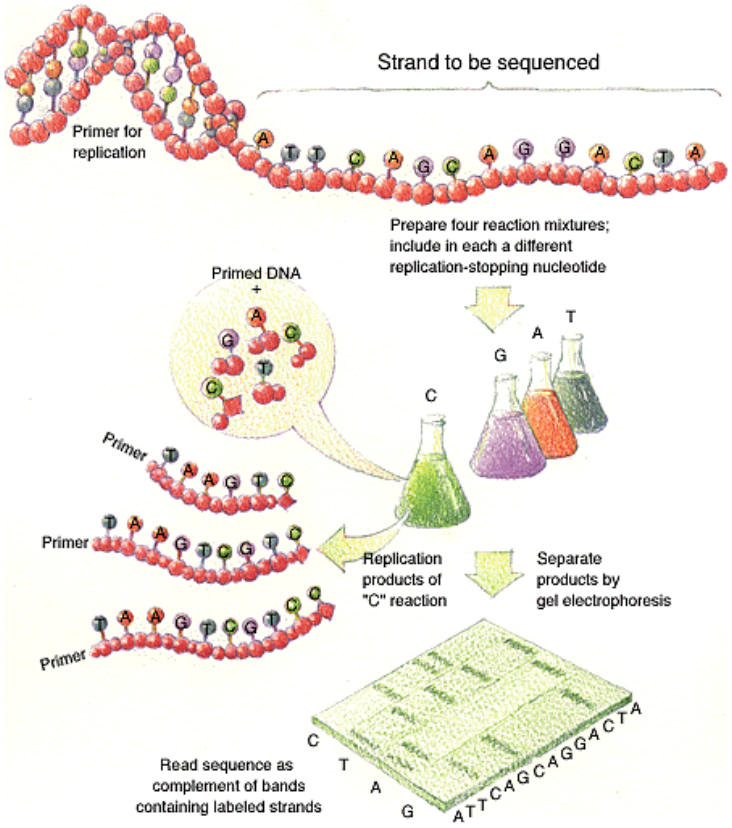
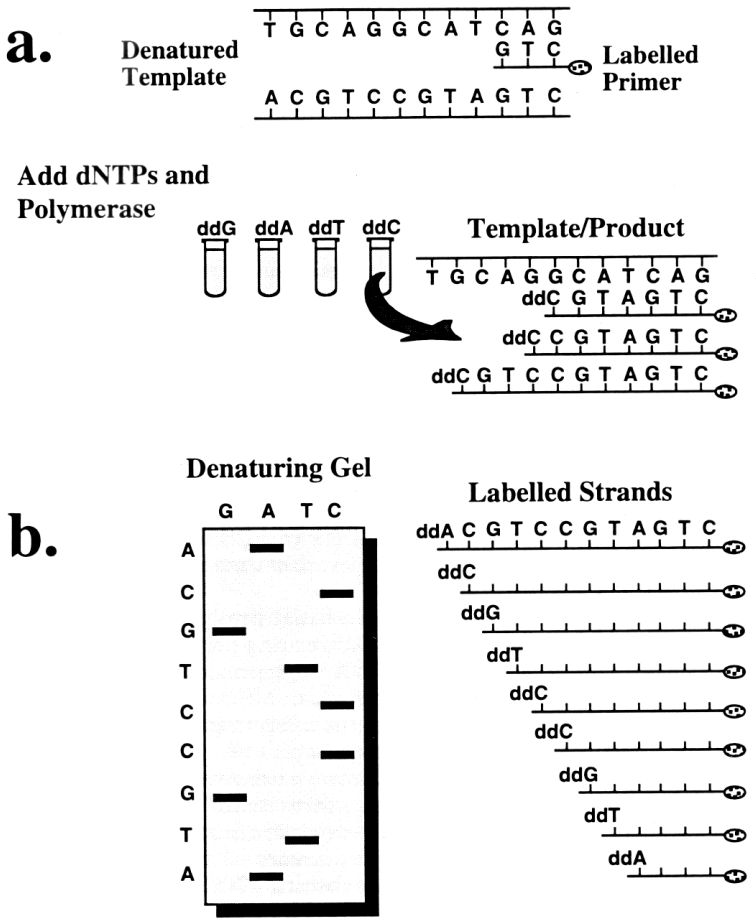



Figure 2. Vertebrate genomic sequence data. Phylogenetic tree representing species for which genomic sequence data are currently available Green indicates that BAC (bacterial artificial chromosome)-based sequence is available in targeted regions of the genome[11, 15]. Yellow represents 2X whole-genome shotgun assemblies[17], and blue represents full-shotgun or near-complete genomic sequence assemblies.

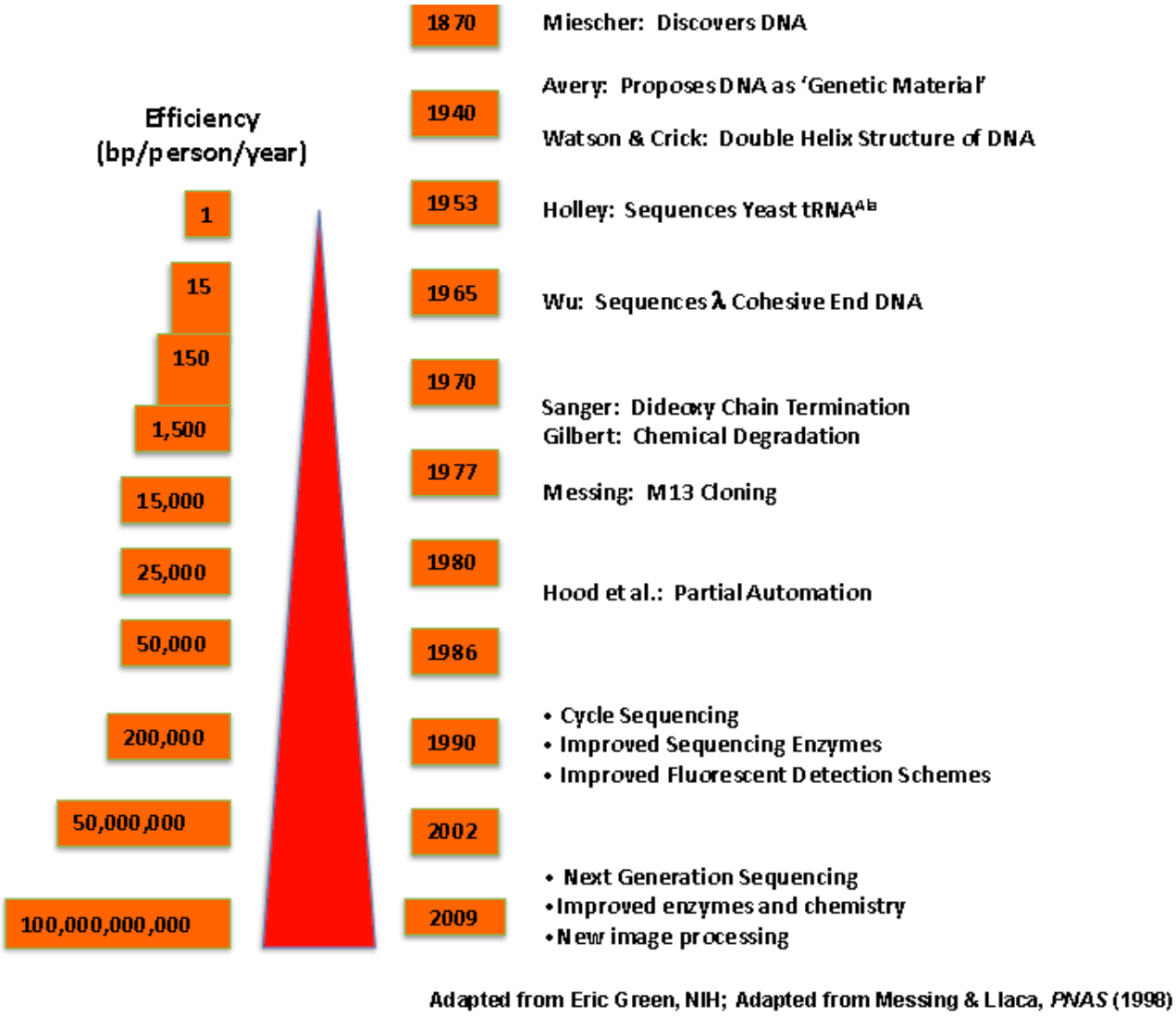
# Sequencing: the Sanger Method (1977)



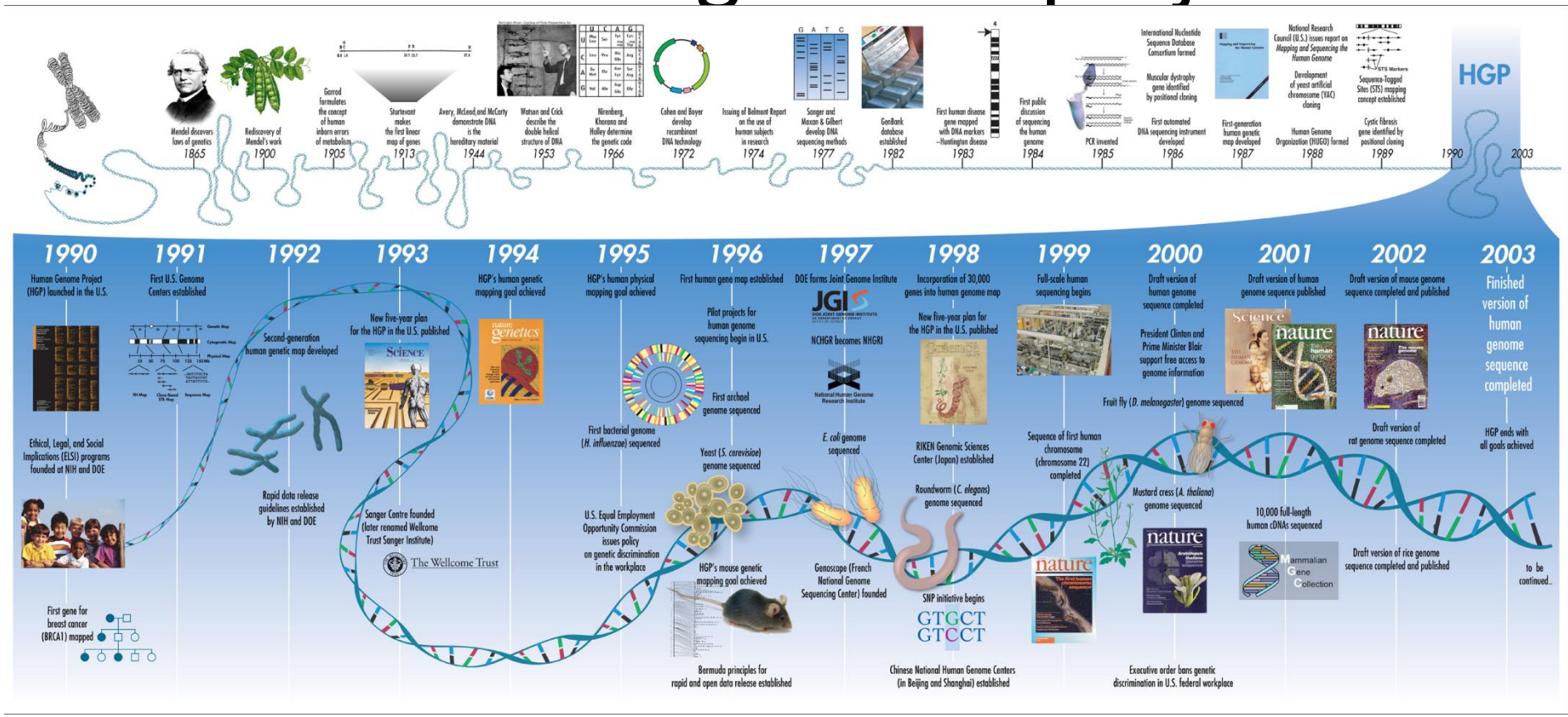

[Click here to see an animation]

# History of DNA sequencing is related to the combination of new technologies.



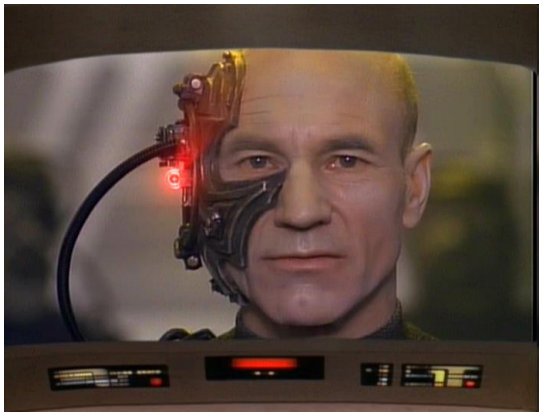

Adapted from Eric Green, NIH; Adapted from Messing & Llaca, *PNAS* (1998)

# The human genome project



# To know more about HGP

- http://exploreable.wordpress.com/2011/05/03/the-story-of-the-human-genome-project-a-short-narration/
- All about the human genome project
- http://e2013.wordpress.com/2013/03/11/les-estadistiques-del-genoma/

# Next generation sequencing

The future is here, now

# Next generation Sequencing

- Improvements in enzymes, chemistry and image analysis, mature by the middle of last decade dramatically increased sequencing capabilities.
- The newest type of technology, called “next-generation sequencing”, appeared with the potential to accelerate biological and biomedical research
  - *by enabling the comprehensive analysis of genomes, transcriptomes and interactomes,*
  - *by tending to become inexpensive, routine and widespread, rather than requiring very costly production-scale efforts.*

# NGS technologies

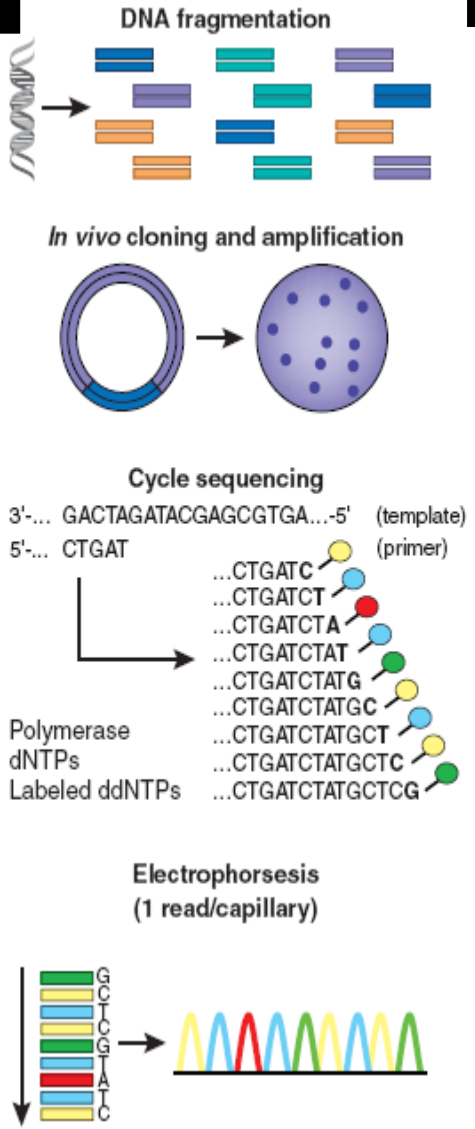

Cost-effective
Fast
Ultra throughput
Cloning-free
Short reads



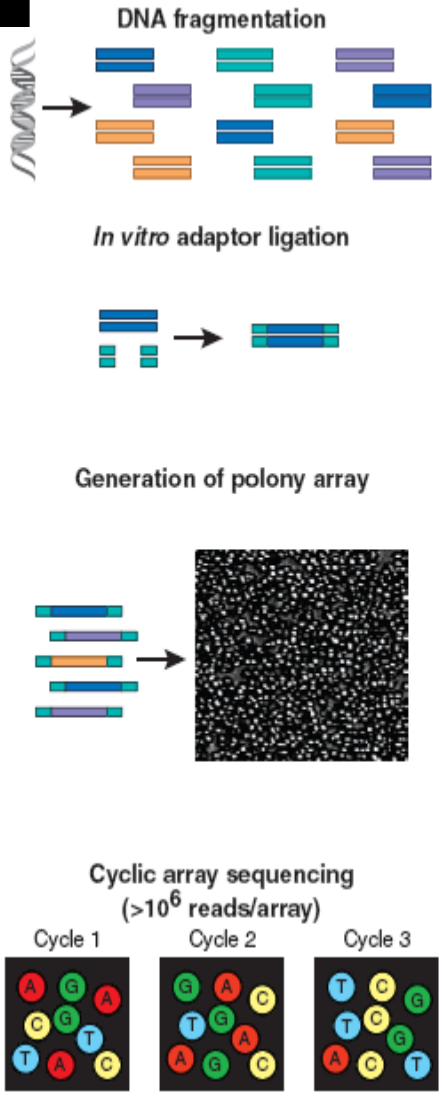


# Next-generation DNA sequencing

## Sanger sequencing

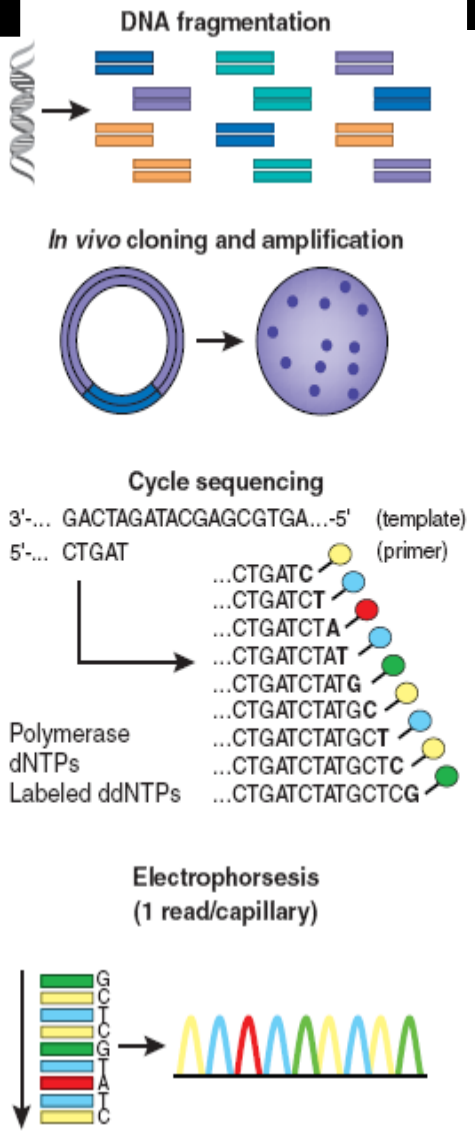



## Cyclic-array sequencing

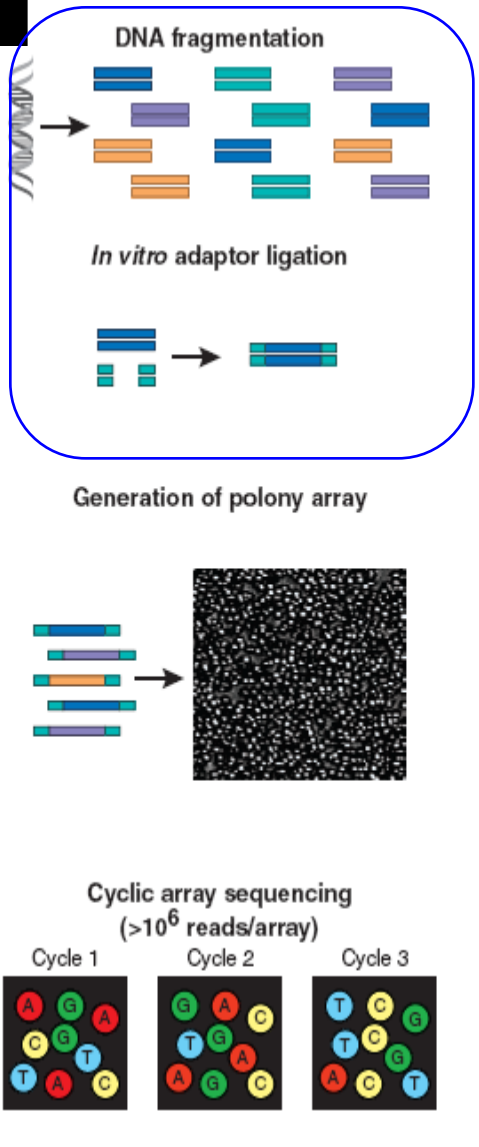


# Next-generation DNA sequencing

Sanger sequencing




Next-generation sequencing

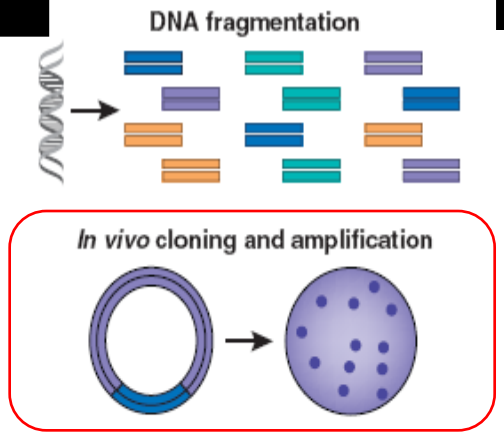


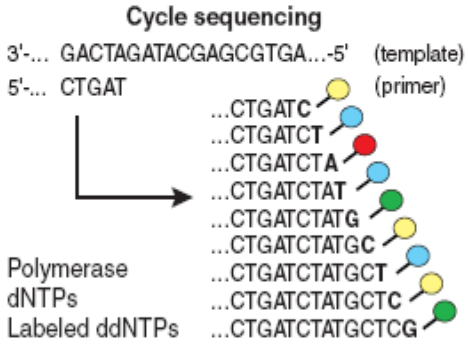

## Advantages of NGS

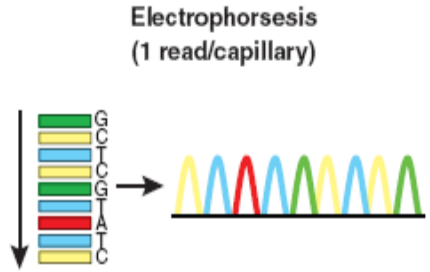
- Construction of a sequencing library → clonal amplification to generate sequencing features

# Next-generation DNA sequencing

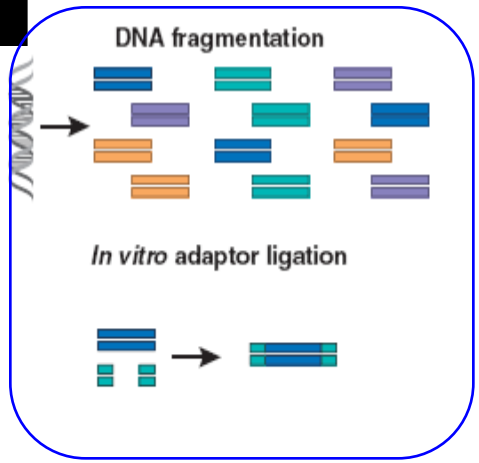
Sanger sequencing

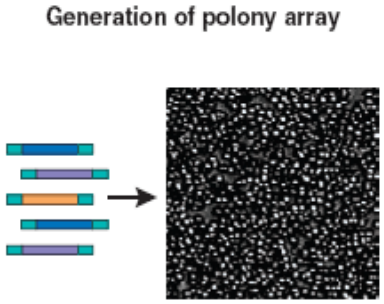


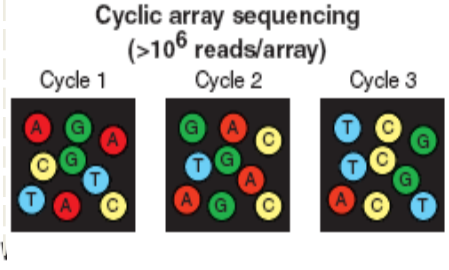




Next-generation sequencing



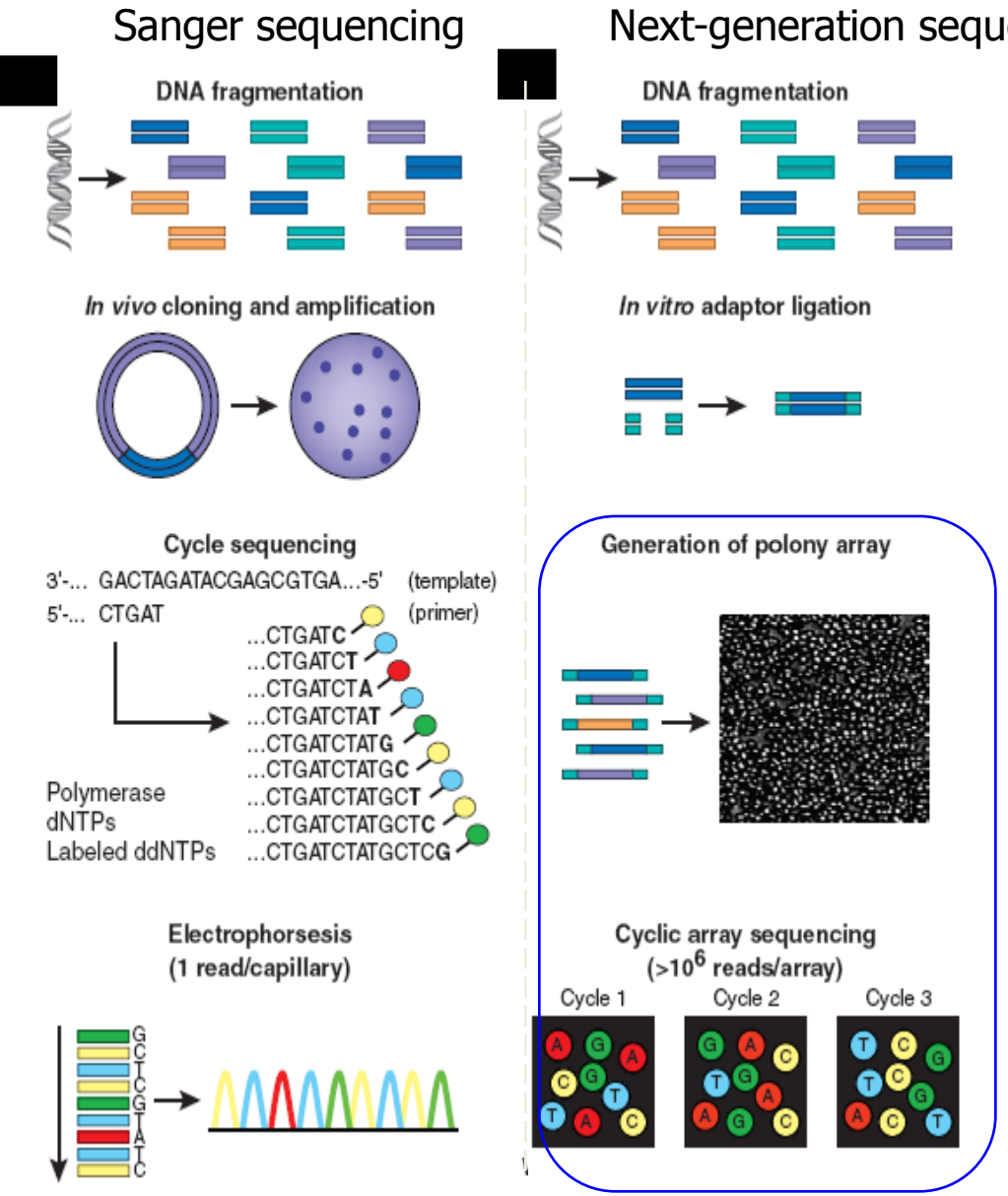




## Advantages:

- Construction of a sequencing library → clonal amplification to generate sequencing features

✓ No in vivo cloning, transformation, colony picking...

# Next-generation DNA sequencing




## Advantages:

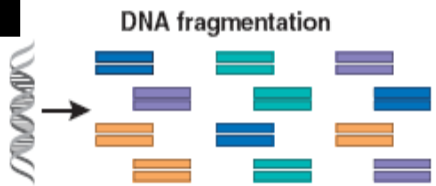
Construction of a sequencing library → clonal amplification to generate sequencing features

✓ No in vivo cloning, transformation, colony picking...

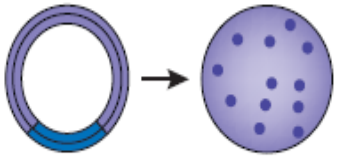
- Array-based sequencing

# Next-generation DNA sequencing

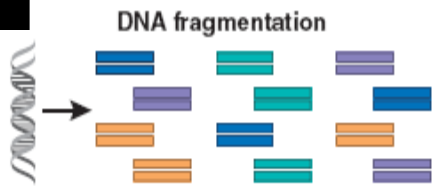
Sanger sequencing



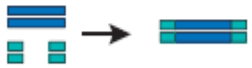

*In vivo* cloning and amplification

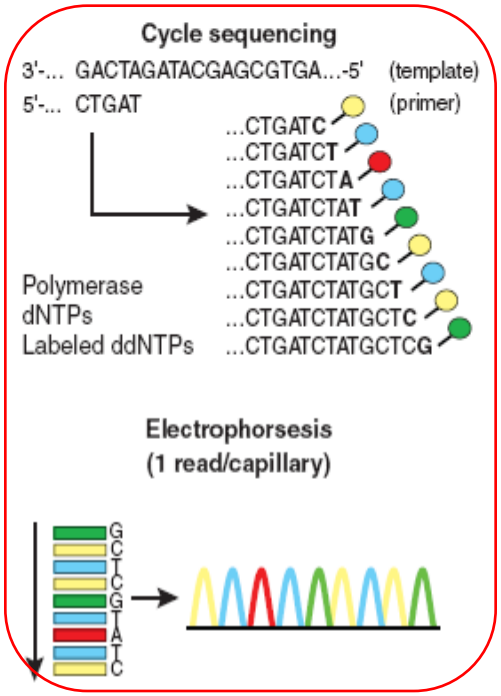


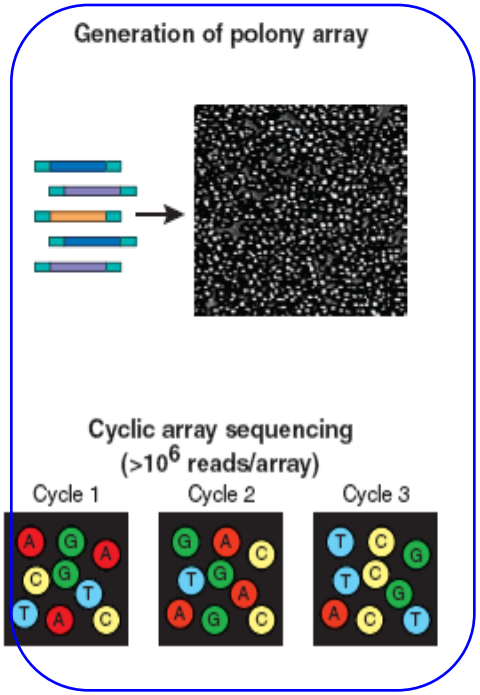
Next-generation sequencing




*In vitro* adaptor ligation






## Advantages:

- Construction of a sequencing library → clonal amplification to generate sequencing features

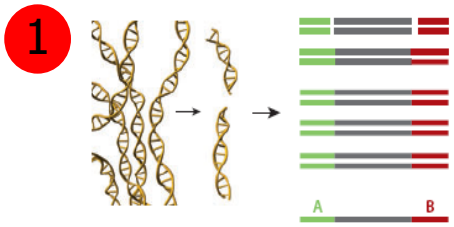
  ✓ No in vivo cloning, transformation, colony picking...

- Array-based sequencing

  ✓ Higher degree of parallelism than capillary-based sequencing

# The sequencing process, in detail

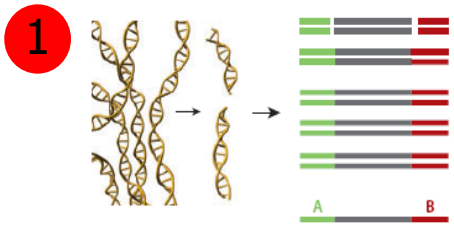
1. Library preparation




DNA fragmentation and in vitro adaptor ligation

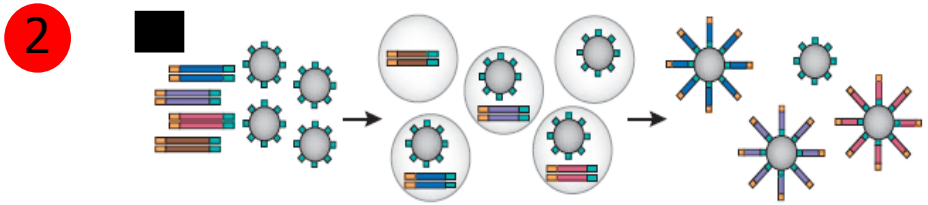
# Next-generation DNA sequencing

1. Library preparation
2. Clonal amplification



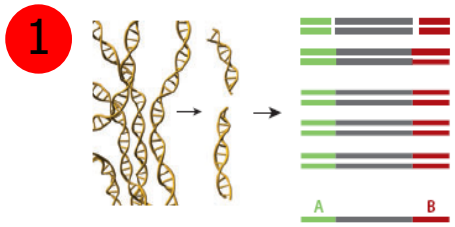

DNA fragmentation and in vitro adaptor ligation

emulsion PCR



# Next-generation DNA sequencing

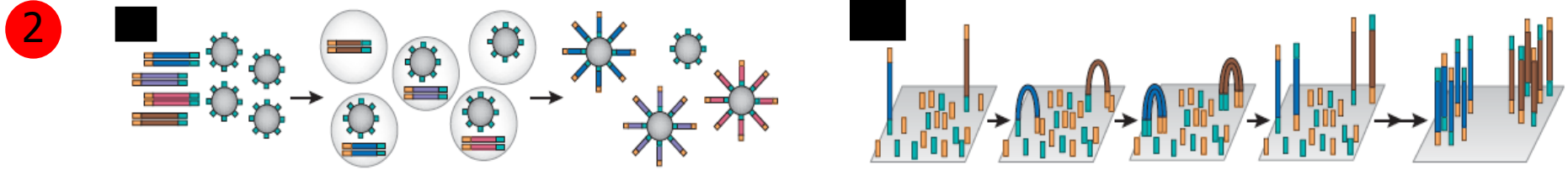
1. Library preparation
2. Clonal amplification




DNA fragmentation and in vitro adaptor ligation

emulsion PCR

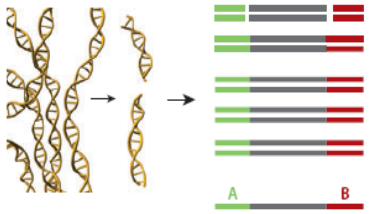
bridge PCR



# Next-generation DNA sequencing

1. Library preparation
2. Clonal amplification
3. Cyclic array sequencing

1

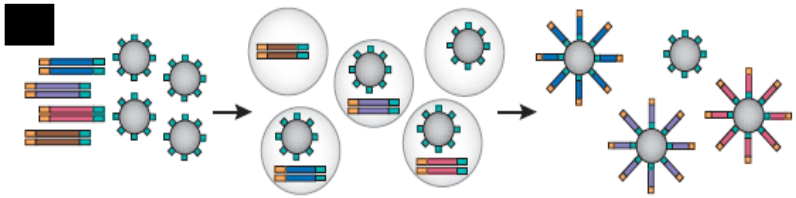


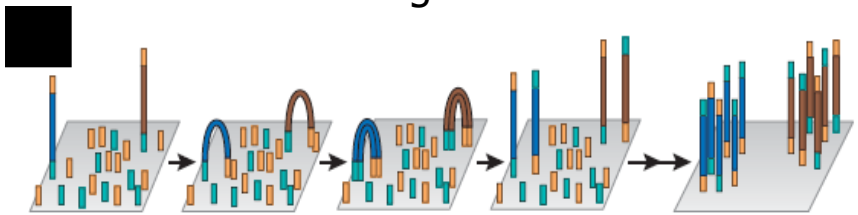

DNA fragmentation and in vitro adaptor ligation

emulsion PCR

bridge PCR

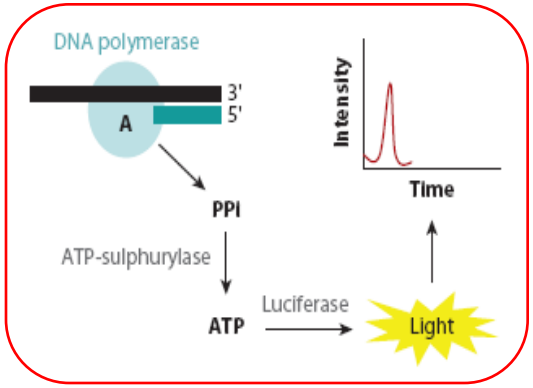
2





3

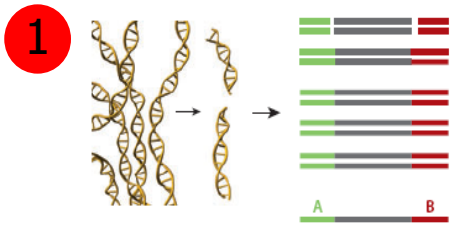
Pyrosequencing




454 sequencing

# Next-generation DNA sequencing

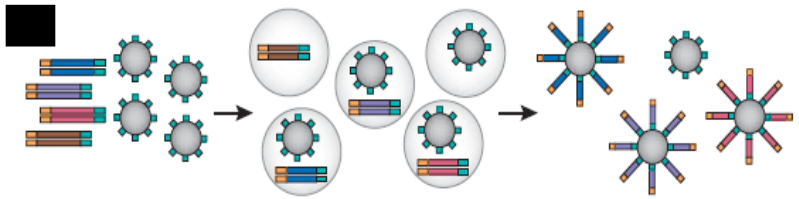
1. Library preparation
2. Clonal amplification
3. Cyclic array sequencing

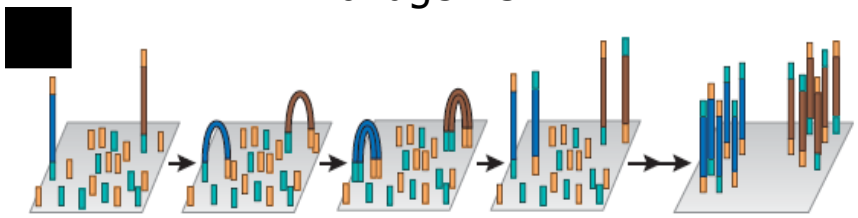



DNA fragmentation and in vitro adaptor ligation

emulsion PCR

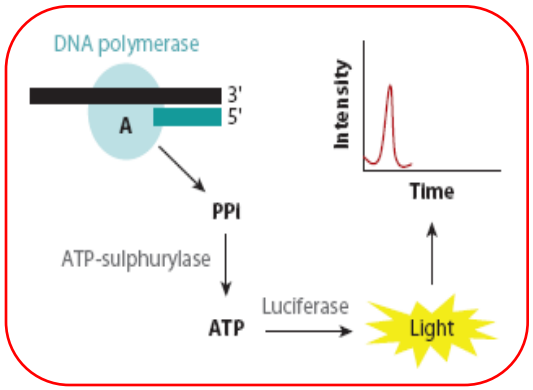
bridge PCR



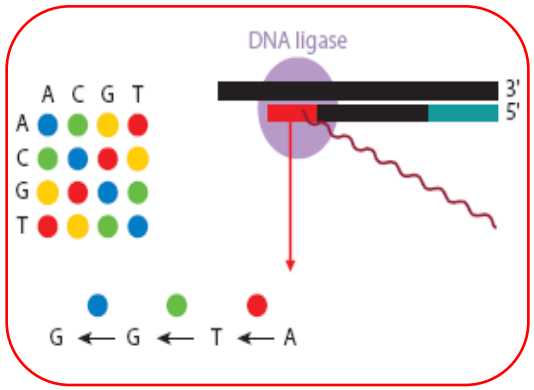
3

Pyrosequencing




454 sequencing

Sequencing-by-ligation

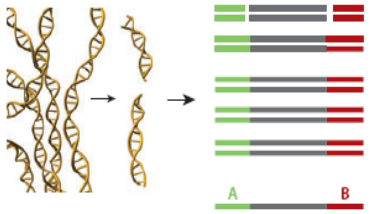



SOLiD platform

# Next-generation DNA sequencing

1. Library preparation
2. Clonal amplification
3. Cyclic array sequencing

1

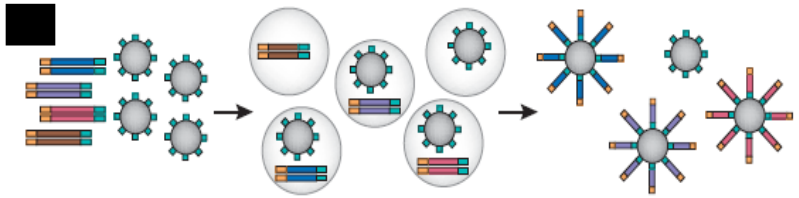


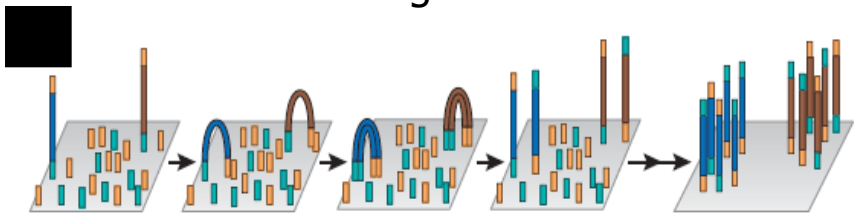

DNA fragmentation and in vitro adaptor ligation

emulsion PCR

bridge PCR

2



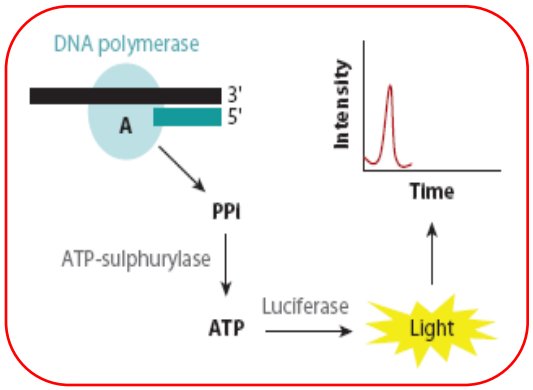


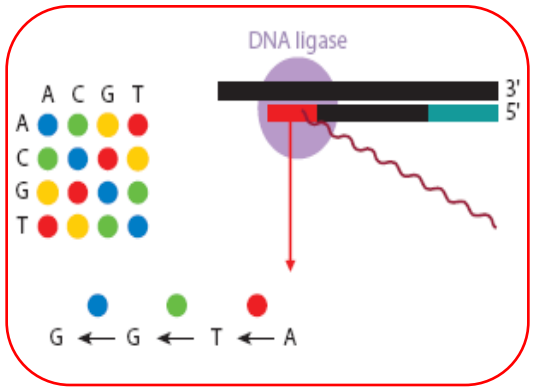
3

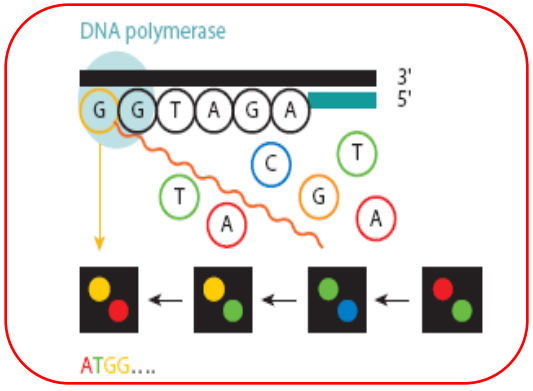
Pyrosequencing

Sequencing-by-ligation

Sequencing-by-synthesis



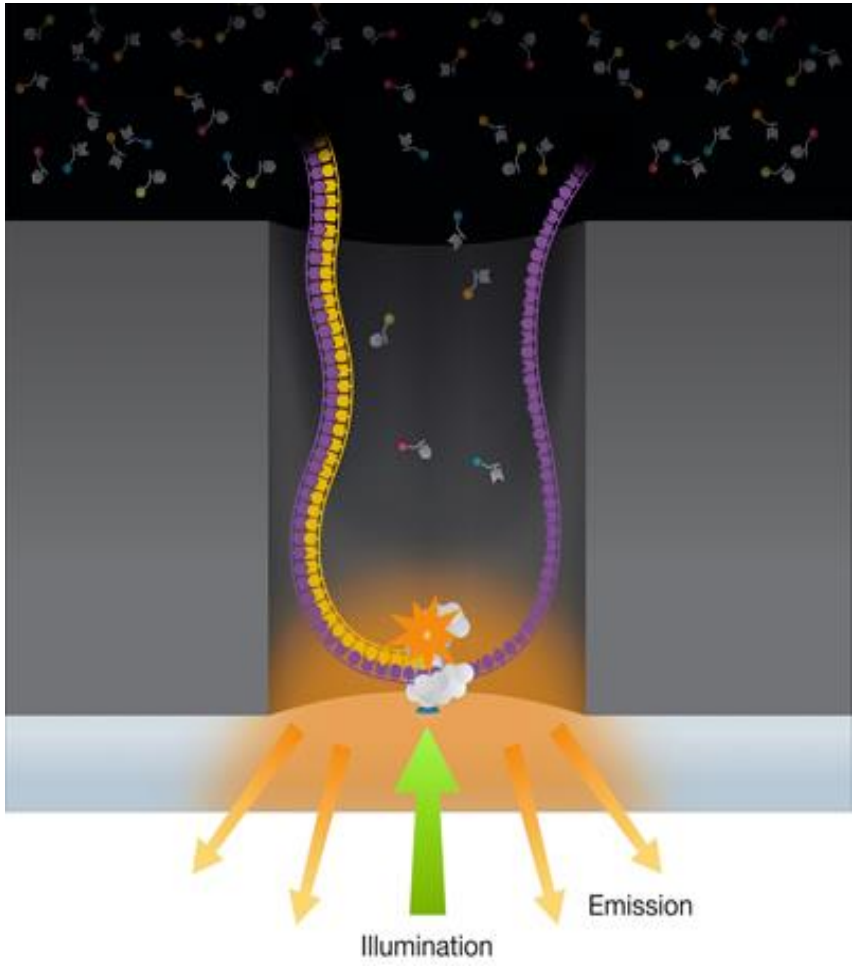

454 sequencing

SOLiD platform

Solexa technology

# Next next generation sequencing

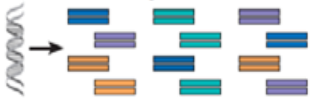
- Pacific Biosystems
  - –Real time DNA synthesis
  - –Up to 12000nt (?)
  - –50 bases/second (?)
- Promises delivery of human genome in minutes?
  - –Company on track for 2013



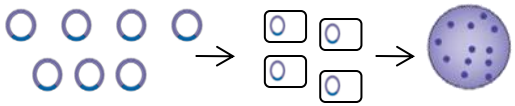
# Sanger sequencing vs. NGS (2nd and 3rd generatio

## Sanger

**1. Fragmentación de DNA**



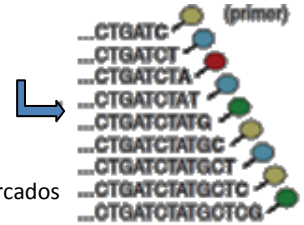
**2.Clonaje en Vectores; Transformación Bacterias; crecimiento y aislamiento vector DNA**



**3. Ciclo Secuenciación**

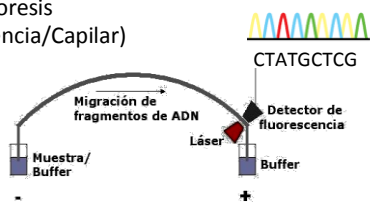
Secuencia:
Primer:

Polimerasa
dNTPs
ddNTPs marcados



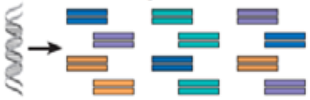

**4. Procesamiento imagen**

Electroforesis
(1 Secuencia/Capilar)



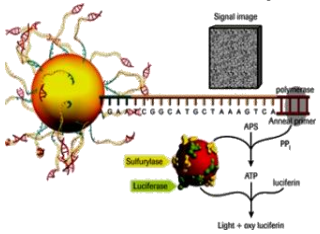

## 2ªNGS

**1. Fragmentación de DNA**

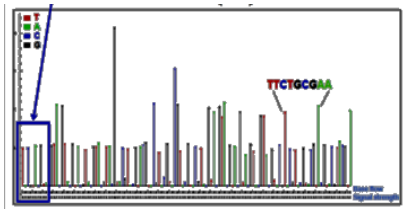


**2. Ligación de adaptadores in vitro y Amplificación clonal**


**3. Secuenciación masiva en paralelo**



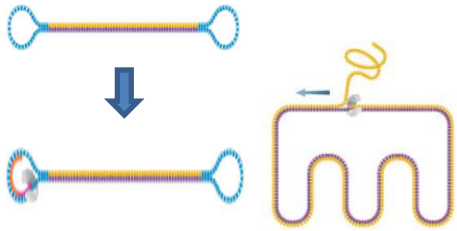

**4. Procesamiento imagen y análisis de datos**



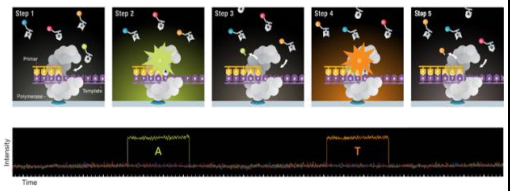

## 3ªNGS

**1. Fragmentación de DNA**

**2. y 3. Ligación de adaptadores in vitro y Secuenciación masiva SIN Amplificación**



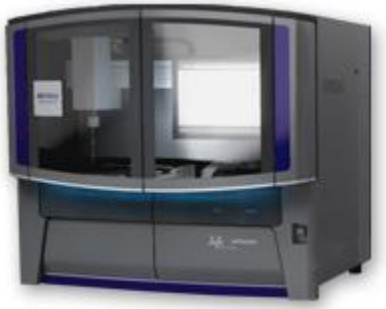
**4. Procesamiento imagen y análisis de datos**



# NGS means high sequencing capacity

GS FLX 454
(ROCHE)

HiSeq 2000
(ILLUMINA)



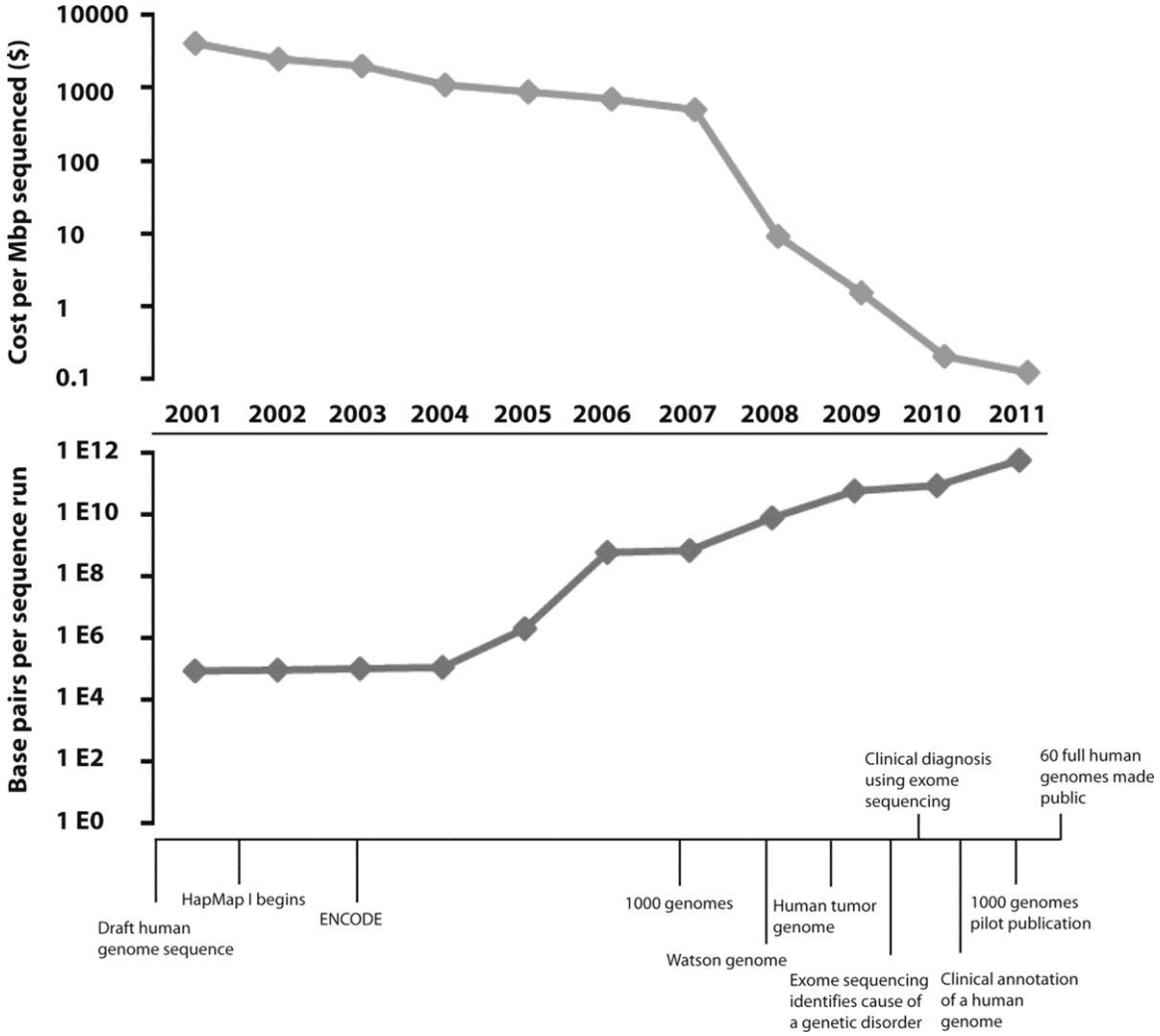
5500xl SOLiD
(ABI)


GS Junior

MiSeq
(ILLUMINA)

Ion TORRENT

# NGS: (much) More per (much) less



# NGS platforms comparison

| PLATFORM | ROCHE GS FLX+ 454 | ILLUMINA HISEQ 2500 | ION PROTON |
|---|---|---|---|
| Library preparation | emPCR | Bridge amplification | emPCR |
| Sequencing chemistry | Pyrosequencing | Reversible dye terminators | pH change |
| Read length | Up to 1000bp | From 2x125 bp to 2x300 bp | Up to 200 bp |
| Run time | 22 hrs | 7 hrs-6 days | From 2 to 4 hrs |
| Throughput/run | Up to 700 Mb | 500-1000Gb | Up to 60 Gb |
| Equipment Cost | 500.000 $ | 750.000 $ | 250.000 $ |
| Reagents Cost/run | 8.000 $ | 5.500 $ | 1.000 $ |
| GOOD! | Longest read length | High throughput/low cost per base/ease of use | Quick, easy to use and cheap |
| BAD! | High error rate in homopolymers (>6); very expensive; low throughput; not automatized at all | Short sequences<br>Strand-specific errors, substitutions towards the end of the read, base substitution errors (sistematic error GGT >GGG) | Errors in homopolymers<br>Higher bias than Illumina |

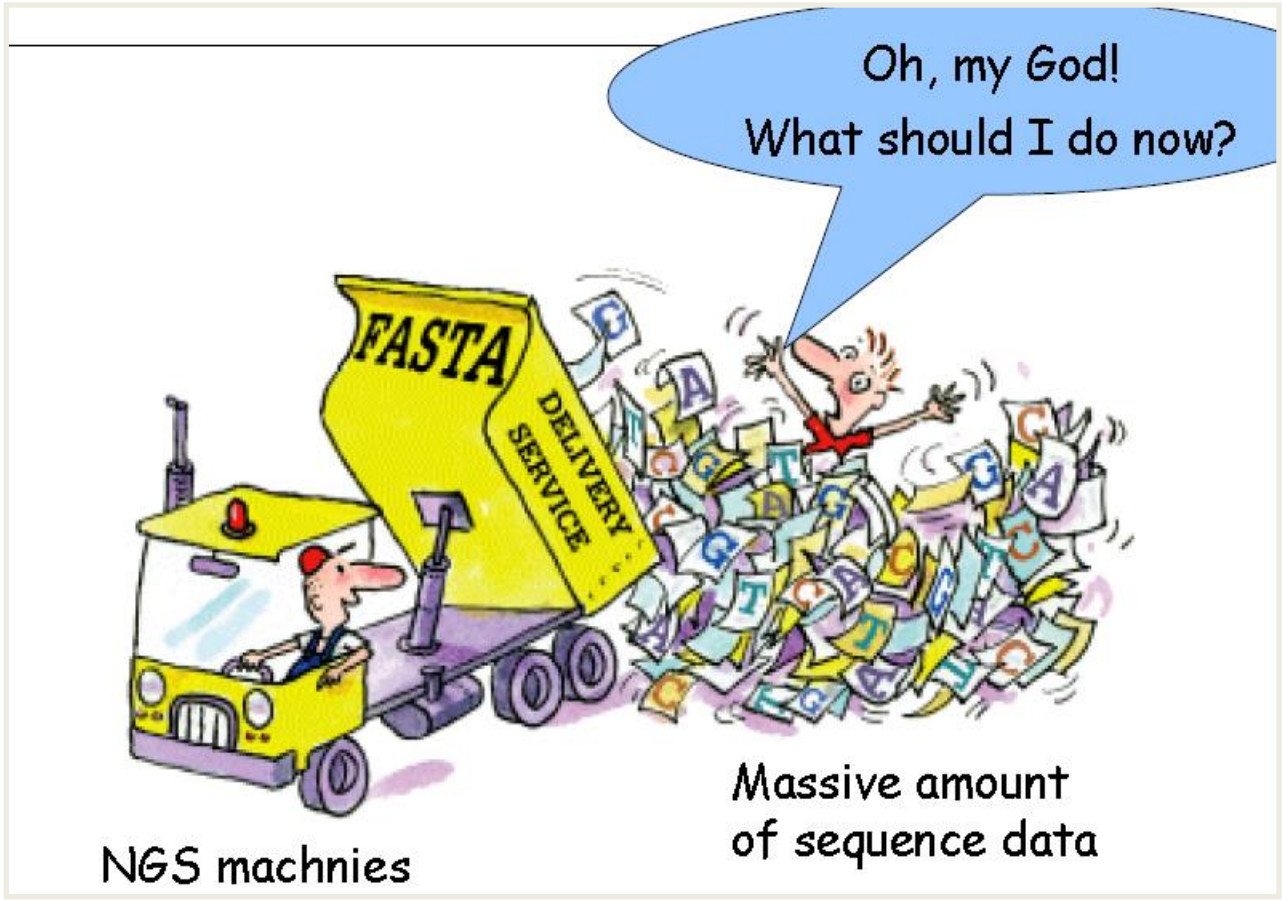
Ross *et al. Genome Biology* 2013, **14**:R51
http://genomebiology.com/2013/14/5/R51

# Comparison between small sequencers

| *Next-Gen Sequencer* | *Machine Cost* | *Cost per run* | *Minimum Throughput* | *Sequencing Run Time* | *Cost per Mb* |
|---|---|---|---|---|---|
| **Illumina MiSeq** | $125,000 | $750 | 1500 Mb (2 x 150 Bases) | 27 Hours | $0.5 |
| **454 GS Junior** | $108,000 | $1,100 | 35 Mb (400 Bases) | 8 Hours | $31 |
| **Ion Torrent PGM- 314 Chip** | $80,490 | $225 | 10Mb (100 Bases) | 3 Hours | $22.5 |
| **Ion Torrent PGM - 316 Chip** | $80,490 | $425 | 100Mb | 3 Hours | $4.25 |
| **Ion Torrent PGM - 318 Chip** | $80,490 | $625 | 1000Mb | 3 Hours | $0.63 |

# Bioinformatics challenges of NGS

# I have my sequences/images. Now what?



# NGS pushes (bio)informatics needs up

- Need for large amount of CPU power
  - Informatics groups must manage computer clusters
  - Challenges in parallelizing existing software or redesign of algorithms to work in a parallel environment
  - Another level of software complexity and challenges to interoperability
- VERY large text files (~10 million lines long)
  - Can’t do ‘business as usual’ with familiar tools such as Perl/Python.
  - Impossible memory usage and execution time
  - Impossible to browse for problems
- Need sequence Quality filtering

# Data management issues

- Raw data are large. How long should be kept?
- Processed data are manageable for most people
  - 20 million reads (50bp) ~1Gb
- More of an issue for a facility: HiSeq recommends 32 CPU cores, each with 4GB RAM
- Certain studies much more data intensive than other
  - Whole genome sequencing
    - A 30X coverage genome pair (tumor/normal) ~500 GB
    - 50 genome pairs ~ 25 TB

# So what?

- In NGS we have to process really big amounts of data, which is not trivial in computing terms.
- Big NGS projects require supercomputing infrastructures.
- Or put another way: it's not the case that anyone can do everything.
  - Small facilities must carefully choose their projects to be scaled with their computing capabilities.

# Computational infrastructure for NGS

- Infrastructures of different size
  - *Small,*
  - *Medium,*
  - *Big*

  provide diverse capabilities

  for a wide range of project types.

# Small infrastructure

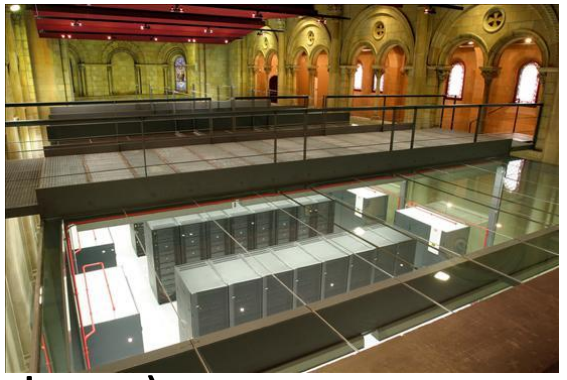
- Recommended at least 2 machines
  - 8 or 12 cores each machine.
  - 48Gb ram minimum each machine.
  - BIG local disk. At least 4TB each machine
    - As much local disks as we can afford
- Price range: starting at 8.000€ - 10.000€ (2 machines)
- Can do
  - RNA-seq analysis
  - Small metagenomic projects
  - A few exomes

# Middle size infrastructure

- "Small" distributed filesystem ( around 50TB).
- "Small" cluster (10 nodes, 80 to 120 cores).
- Gigabit ethernet network.
- Price : 50.000 – 100.000 € (just hardware) plus data center and informaticians salary
- Can do
  - Genome assembly
  - More than-a-few exomes
  - Middle-sized metagenomics.

# Big computing infrastructure

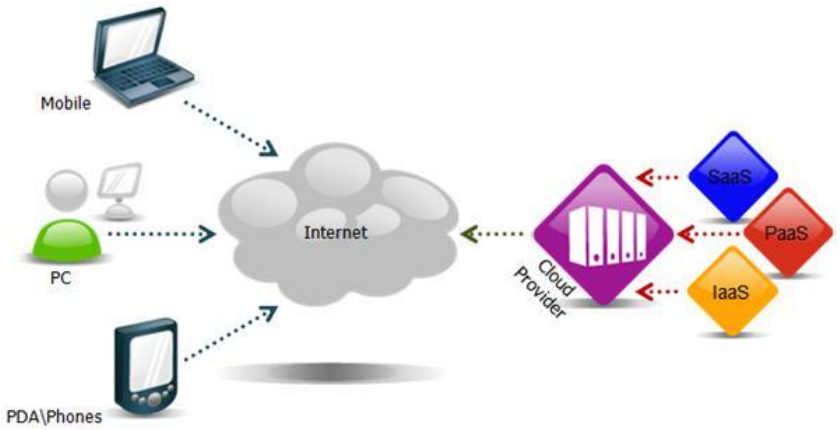
- Distributed memory cluster
  - Starting at 20 computing nodes
  - 160 to 240 cores
  - At least 48GB ram per node
- Fast networks (10Gbit or Infiniband



- Batch queue system (sge, condor, pbs, slurm)
- Optional MPI and GPUs environment
- Good for large-scale projects (**big data** at last)
  - Quick assembly of human genomes
  - Hundreds/Thousands of exomes
  - Large metagenomic projects
- Very expensive to acquire and maintain

# Alternatives (1): Cloud Computing

- Pros
  - –Flexibility.
  - –You pay what you use.
  - –Don´t need to maintain a data center.
- Cons
  - –Transfer big datasets over internet is slow.
  - –You pay for consumed bandwidth. That is a problem with big datasets.
  - –Lower performance, specially in disk read/write.
  - –Privacy/security concerns.
  - –More expensive for big and long term projects.



# So what?

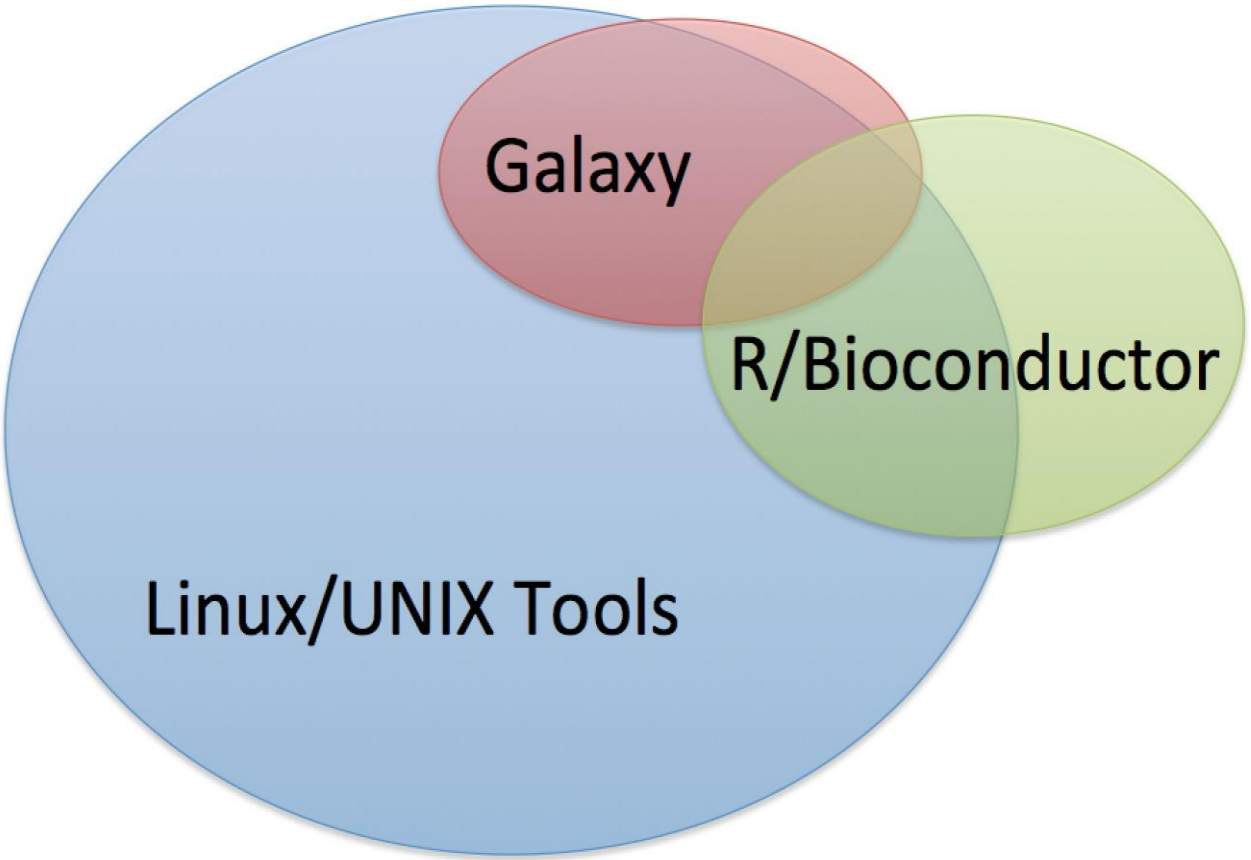
- Think before you NGS
- Decide what you …
  - want to do,
  - can afford
  - know how to do
- Consider all alternatives
- Look for expert advice ..

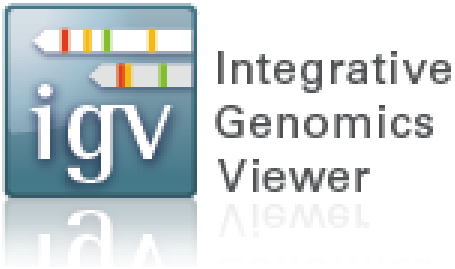
# SOFTWARE FOR NGS DATA ANALYSIS

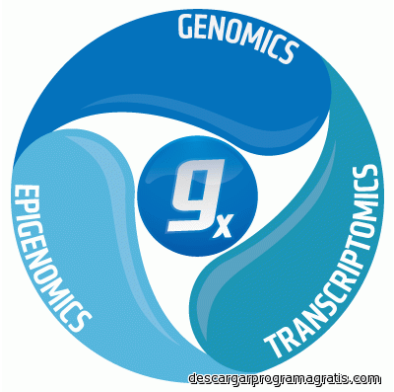
# Which software for NGS (data) analysis?

- Answer is not straightforward.
- Many possible classifications
  - Biological domains
    - SNP discovery, Genomics, ChIP-Seq, De-novo assembly, …
  - Bioinformatics methods
    - Mapping, Assembly, Alignment, Seq-QC,…
  - Technology
    - Illumina, 454, ABI SOLID, Helicos, …
  - Operating system
    - Linux, Mac OS X, Windows, …
  - License type
    - GPLv3, GPL, Commercial, Free for academic use,…
  - Language
    - C++, Perl, Java, C, Phyton
  - Interface
    - Web Based, Integrated solutions, command line tools, pipelines,…

http://seqanswers.com/wiki/Software/list


Galaxy
R/Bioconductor
Linux/UNIX Tools

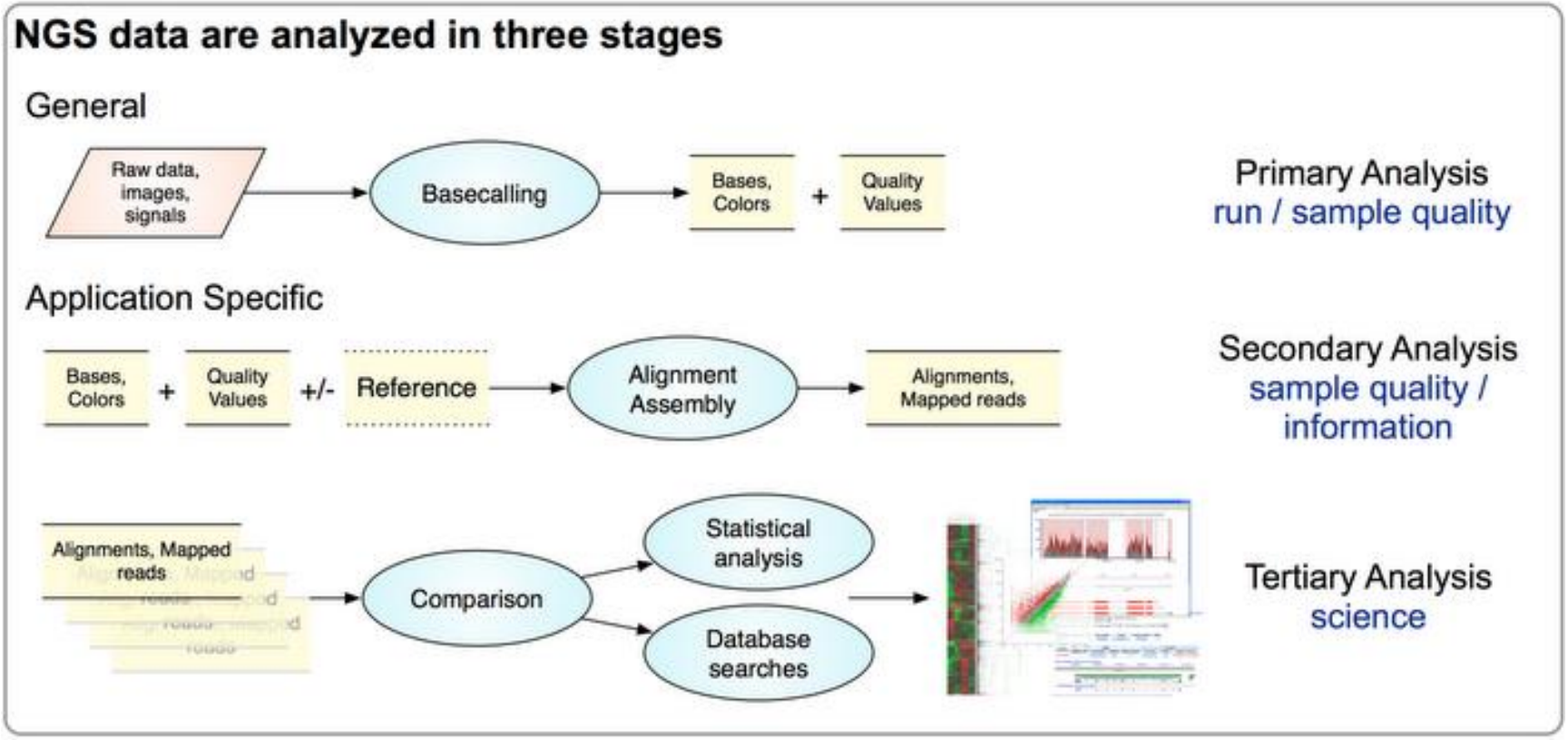

igv
Integrative
Genomics
Viewer


GENOMICS
EPIGENOMICS
TRANSCRIPTOMICS
descargarprogramagratis.com


Welcome to push-button bioinformatics.
BaseSpace
Data storage, analysis, and collaboration made easy.

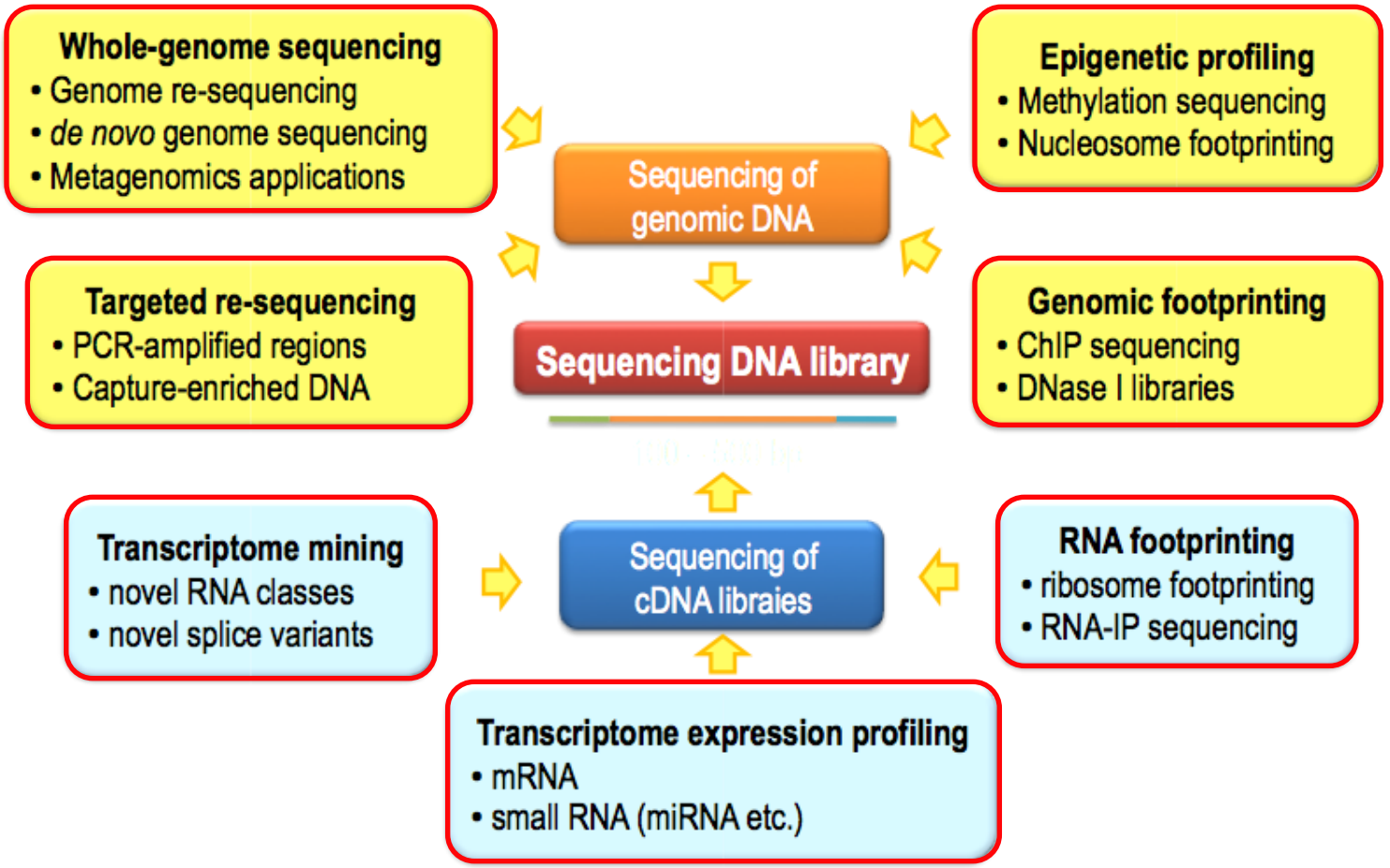
# NGS data analysis

# NGS data analysis stages



# NGS aplications

# Applications of Next-Generation Sequencing



# Metagenomics

Metagenomics is the study of a collection of genetic material (genomes) from a mixed community of organisms. Metagenomics usually refers to the study of microbial communities.

•The biosphere contains between $10^{30}$ and $10^{31}$ microbial genomes, at least 2–3 orders of magnitude more than the number of plant and animal cells combined.
•Microbes associated with the human body outnumber human cells by at least a factor of ten.
•The vast majority cannot be cultured.

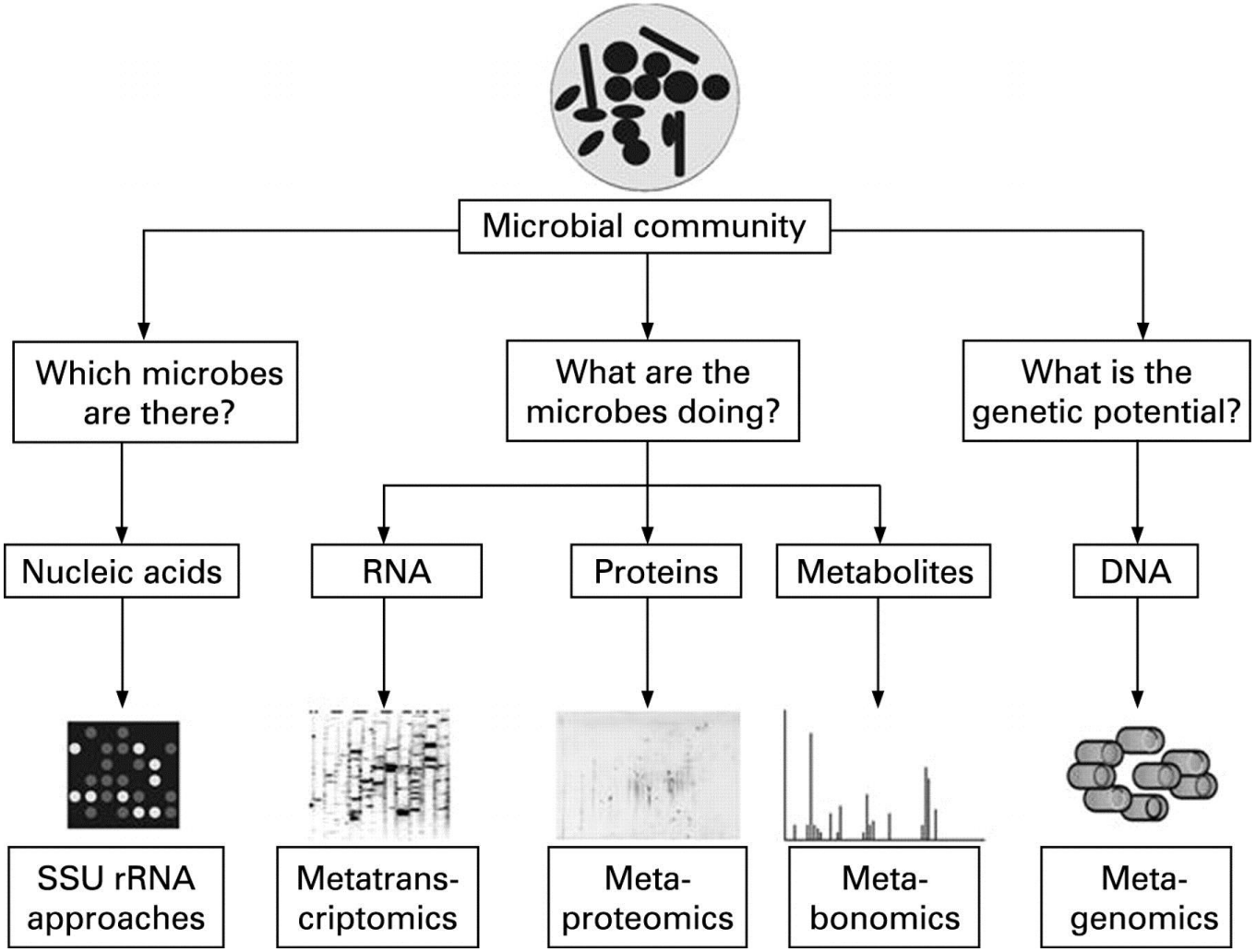
## What can we study?

- A community in a natural environment (such as seawater or soil),

- A host-associated community (such as the microbes in the human gut or mouth), and

- A "managed" environment community (such as a sewage treatment facility or bioremediation site).

# Metagenomics and other community-based “omics”




Zoetendal E G et al.
Gut 2008;57:1605-1615

GUT

# Metagenomic Studies


NIH Human Microbiome Project

**Program Snapshot**

The Common Fund's Human Microbiome Project (HMP) aims to characterize the microbial communities found at several different sites on the human body, including nasal passages, oral cavities, skin, gastrointestinal tract, and urogenital tract, and to analyze the role of these microbes in human health and disease. HMP includes the following initiatives.

- Development of a reference set of microbial genome sequences and preliminary characterization of the human microbiome
- Elucidation of the relationship between disease and changes in the human microbiome
- Development of new technologies for computational analysis
- Development of new tools for computational analysis
- Establishment of a data analysis and coordinating center (DACC)
- Establishment of resource repositories
- Examination of the ethical, legal and social implications (ELSI) of HMP research

Read More...

**Program Highlights**

NIH Human Microbiome Project Completes Seminal Study of Microbial Diversity in Healthy Volunteers

Read more...

Listen to the BBC Frontiers radio program, 'Human Microbes', featuring HMP researchers as they talk about the research and its potential impact on understanding how microbes affect health and disease. EXIT Disclaimer

**Access the HMP data:**

View the genomes of 100s of HMP reference strains in Gen Bank:
http://www.ncbi.nlm.nih.gov/bioproject/28331

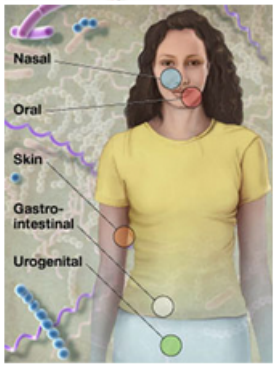
Order an HMP reference strain:
http://www.beiresources.org/

View the HMP BioProjects page at NCBI with sequence and phenotype data:
http://www.ncbi.nlm.nih.gov/genomeprj/43021

Visit the HMP Data Analysis and Coordination Center (DACC) site:
http://www.hmpdacc.org

**WANTED: DEAD or ALIVE!**

THE HUMAN MICROBIOME PROJECT (HMP) NEEDS YOUR HELP!!!


Researchers in the HMP are sampling and analyzing the genome of microbes from five sites on the human body: nasal passages, oral cavities, skin, gastrointestinal tract, and urogenital tract.

Trends Genet. 2012 Nov 7. pii: S0168-9525(12)00145-X. doi: 10.1016/j.tig.2012.09.005. [Epub ahead of print]

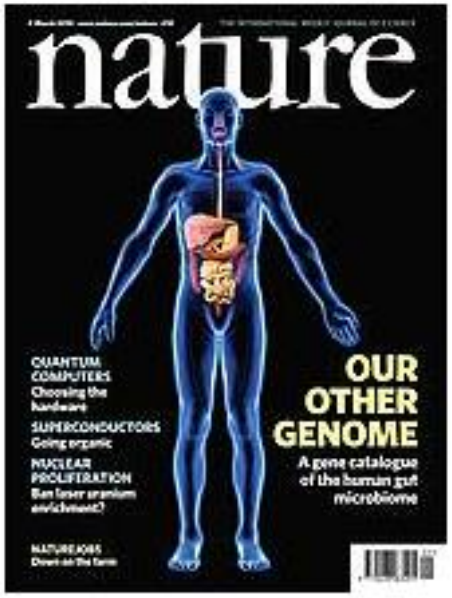
**Biodiversity and functional genomics in the human microbiome.**

Morgan XC, Segata N, Huttenhower C.

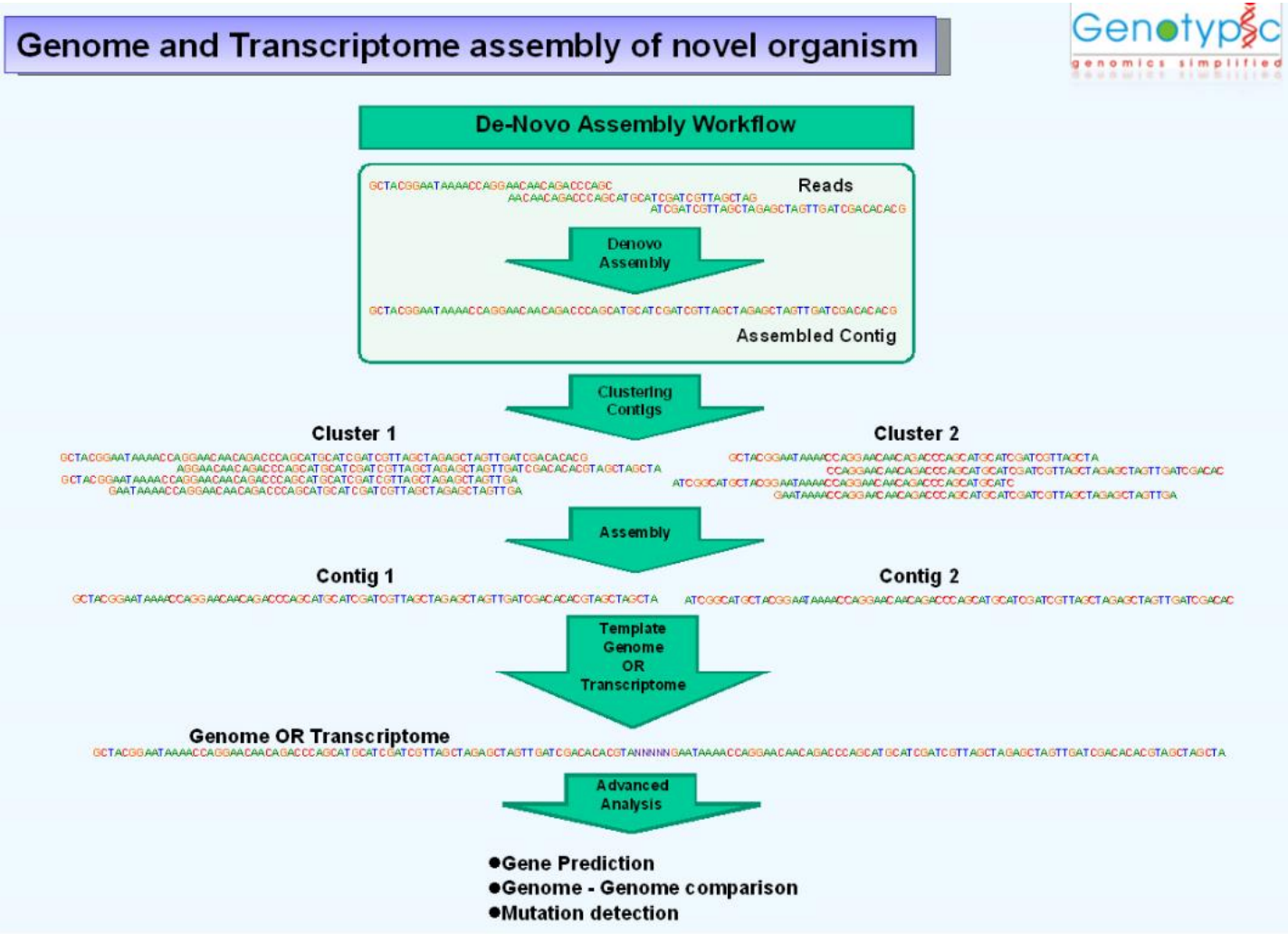
Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA.

# [2]Metagenomic Studies


- To establish associations between the genes of the human intestinal microbiota and our health and disease.
- Focused on two disorders of increasing importance in Europe, Inflammatory Bowel Disease (IBD) and obesity.

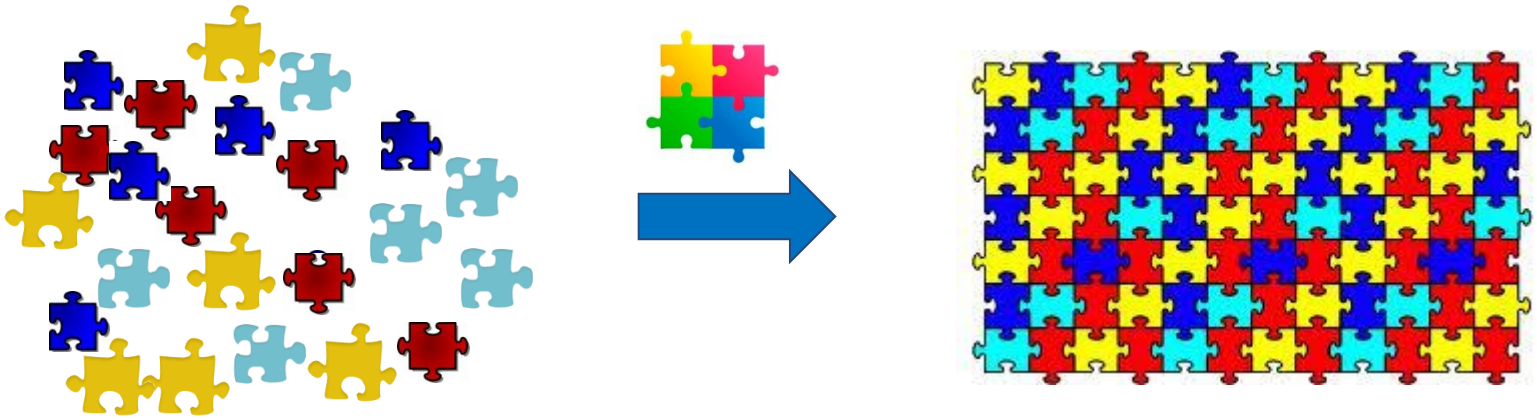


# *De novo* sequencing

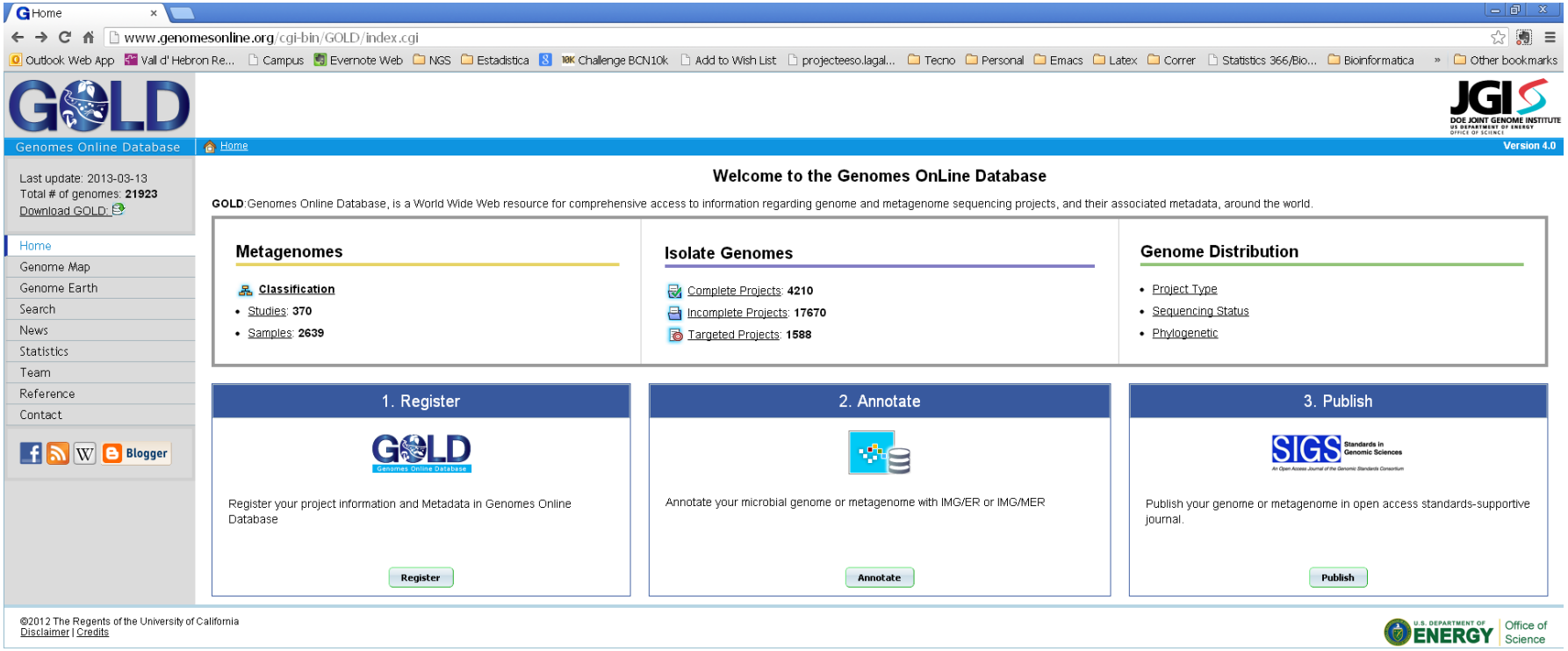
# Sequencing vs re-sequencing

Case A: There is no reference geneome available ( *“de novo” sequencing*)



Case B: There exists a reference genome available (*resequencing*)

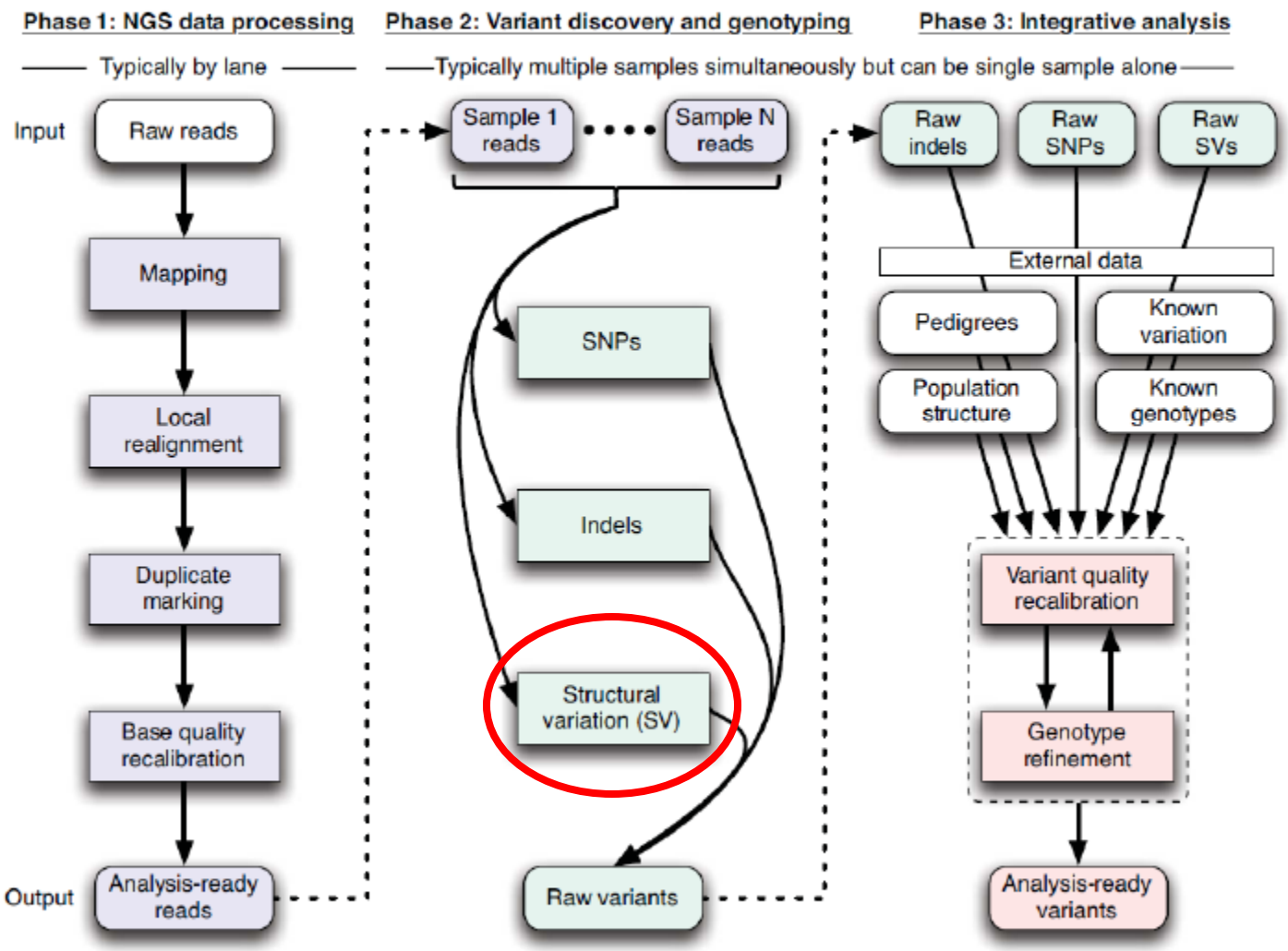
# Which genomes have been sequenced



# Exome Sequencing

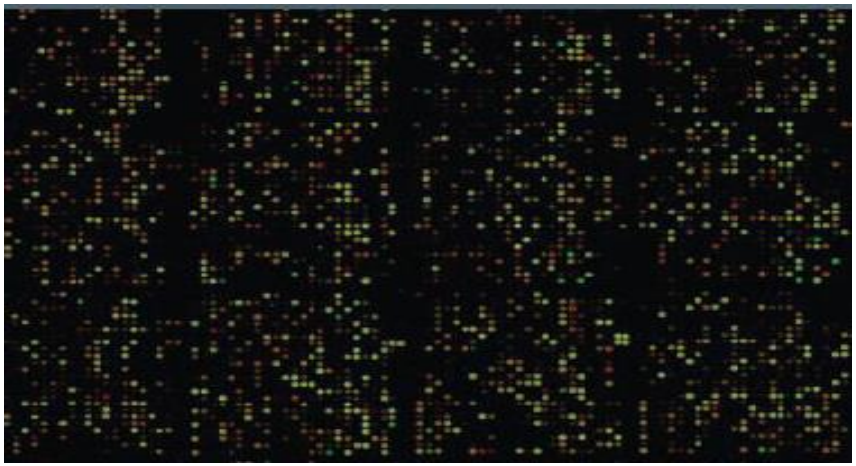
- Capturing protein coding portion of the genome
- ~85% of the disease-causing mutations occur in protein coding regions (exome)
- Exome constitutes 1% of the genome
- About 160,000-180,000 exons
- Time-saving and cost-effective

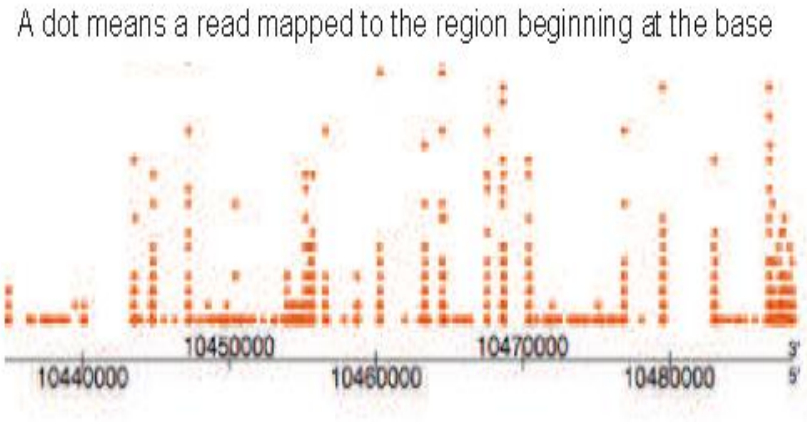
# General Workflow




Source:Nature Genetics 43, 491–498 (2011)

# Transcriptomics by NGS: RNASeq





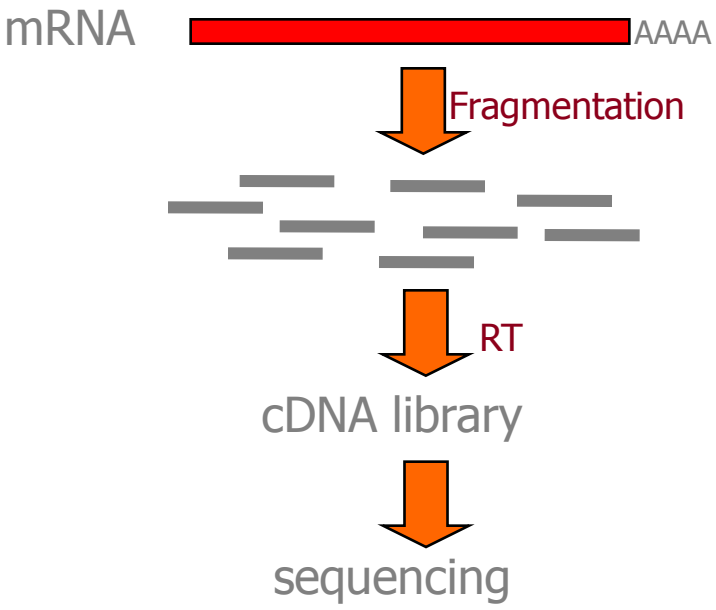

- Analog Signal

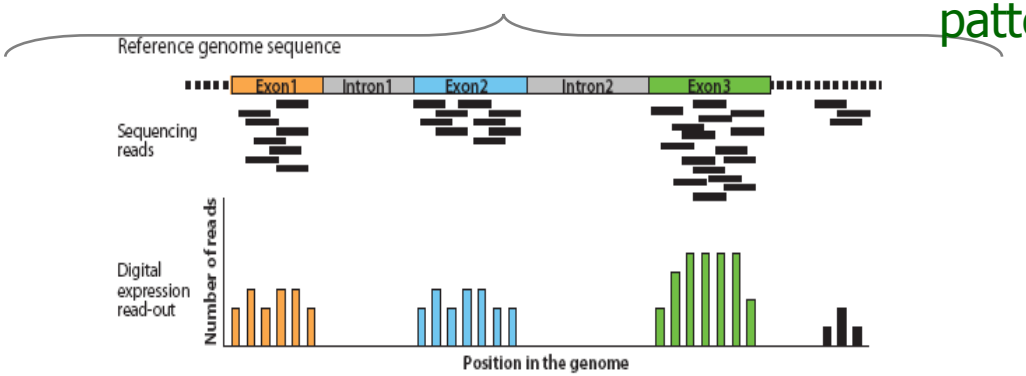
- Easy to convey the signal's information
- Continuous strength
- Signal loss and distortion

- Digital Signal

- Harder to achieve & interpret
- Reads counts: discrete values
- Weak background or no noise

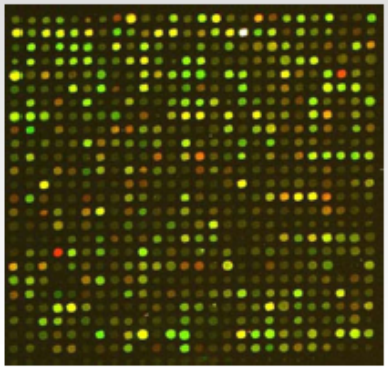
# Applications → Whole transcriptome analysis




- ✓ Detects (differential) expression of known and novel mRNAs
- ✓ Identification of alternative splicing events
- ✓ Detects expressed SNPs or mutations
- ✓ Identifies allele specific expression patterns

# Microarray vs RNA seq data analysis




Only see what's on the chip

One intensity value per probe

| Block | Colum | Row | Name | ID | X | Y | Dia. | F635 Median | F635 Mean | F635 SD | B635 Median | B635 Mean | F532 Median | F532 Mean | F532 SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | CA4018E08 | (2)CA4018E08 | 1880 | 6260 | 140 | 19292 | 20291 | 6153 | 6561 | 6553 | 49758 | 49045 | 7560 |
| 1 | 2 | 2 | CA4018G08 | (7)CA4018G08 | 2120 | 6250 | 150 | 6329 | 8061 | 3555 | 5785 | 5820 | 8338 | 9971 | 6402 |
| 1 | 3 | 2 | Contig2226 | (9)CA4018J10 | 2360 | 6250 | 150 | 6072 | 6839 | 2050 | 5214 | 5187 | 14865 | 16133 | 9977 |
| 1 | 4 | 2 | CA4018C10 | (7)CA4018C10 | 2640 | 6250 | 150 | 3188 | 3186 | 738 | 4907 | 4564 | 3895 | 5975 | 4474 |

Normalization

Differential expression

Unbiased detection

Millions of short sequences

>000590_2485_1157 length=49 uaccno=FGX3UK402GCGZ7
CGTGTCTCTAGACGTCGAAAGCTTCAAATACAAGGCGAAGTACAT

Mapping/Quantification

Normalization

Differential expression