# Análisis de datos ómicos: metagenómica, metatranscrictómica y RNA seq con GAIA y AIR

-

Data-Science para biociencias - 5 Julio 2023

Daniel Julián

djulian@sequentiabiotech.com

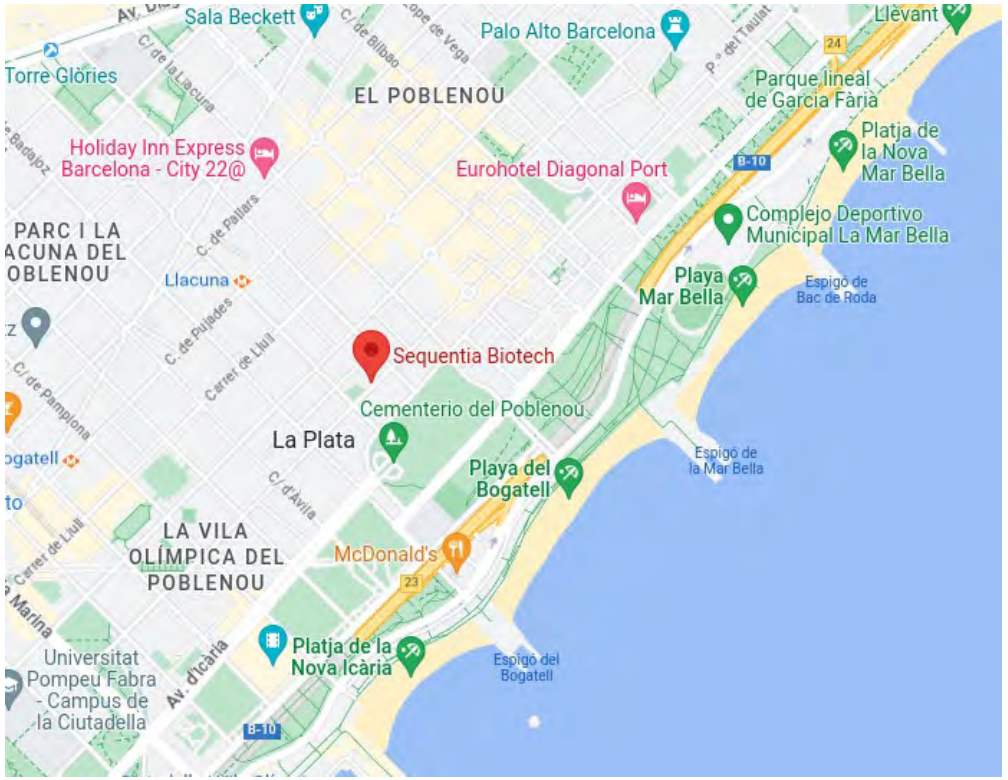
SEQUENTIA

DECODE • DISCOVER • DISRUPT

# Who are we?

SEQUENTIA
DECODE • DISCOVER • DISRUPT

SEQUENTIA
DECODE • DISCOVER • DISRUPT




# Index

SEQUENTIA
DECODE • DISCOVER • DISRUPT

1. **First part: omic data, metagenomics, metatranscriptomics and GAIA**
    1. Omic data
    2. A little bit of history
    3. Applications
    4. Strategies: amplicon-based and shotgun
    5. Bioinformatic approaches and limitations
    6. GAIA
    7. Future insights

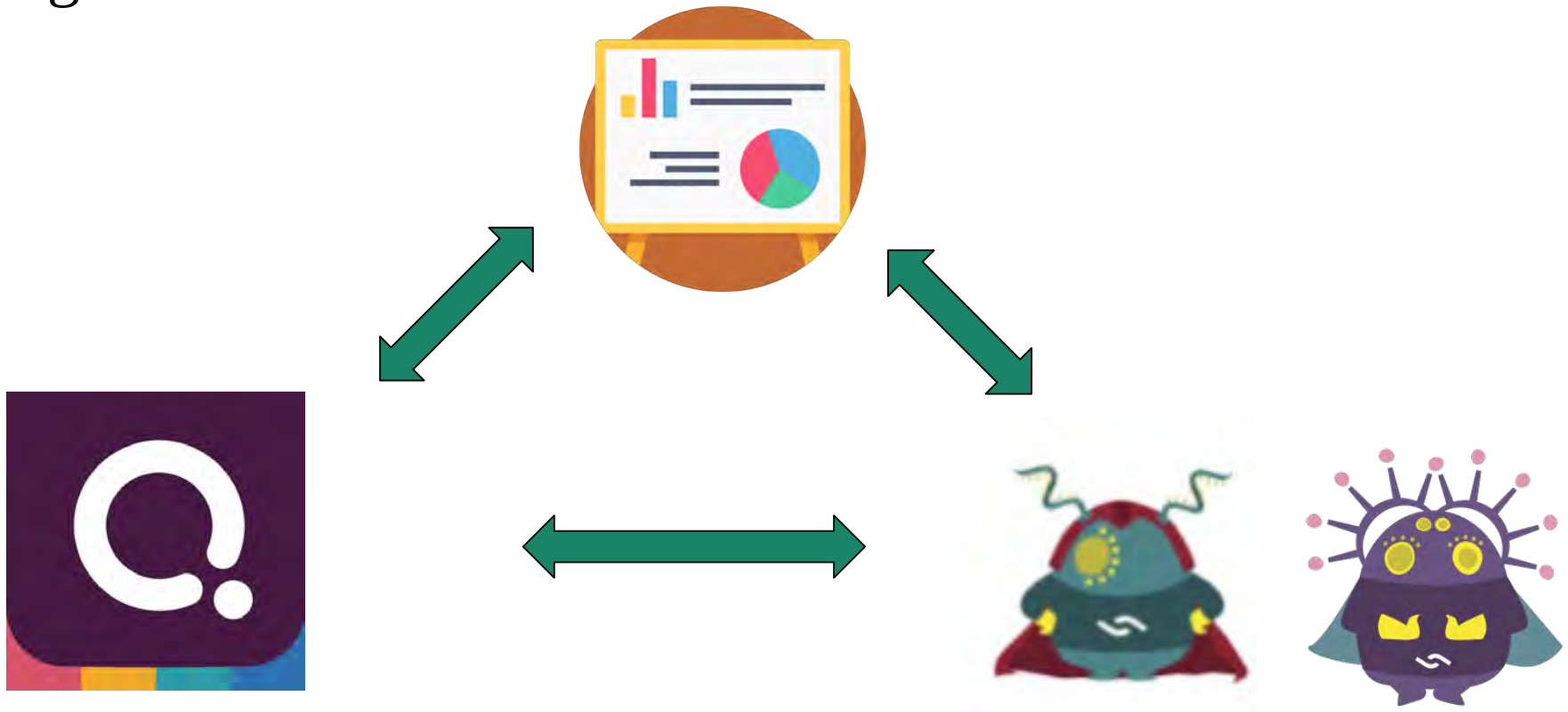
1. **Second part: RNA -seq and AIR**
    1. RNA-seq introduction
    2. Workflow
    3. Differential expression analysis
    4. AIR

# How are we going to do it?

SEQUENTIA
DECODE • DISCOVER • DISRUPT

Active learning

- questions
- ideas



# Quiziiz

SEQUENTIA
DECODE • DISCOVER • DISRUPT

## Rules

1. People are free to participate, it is not mandatory.
2. 45 seconds to answer each question
3. Answering well and quickly gives more points

   Think carefully your answer

# Question 1

https://quizizz.com/?lng=es-ES

# FIRST PART:
# omic data, metagenomics, metatranscriptomics and GAIA

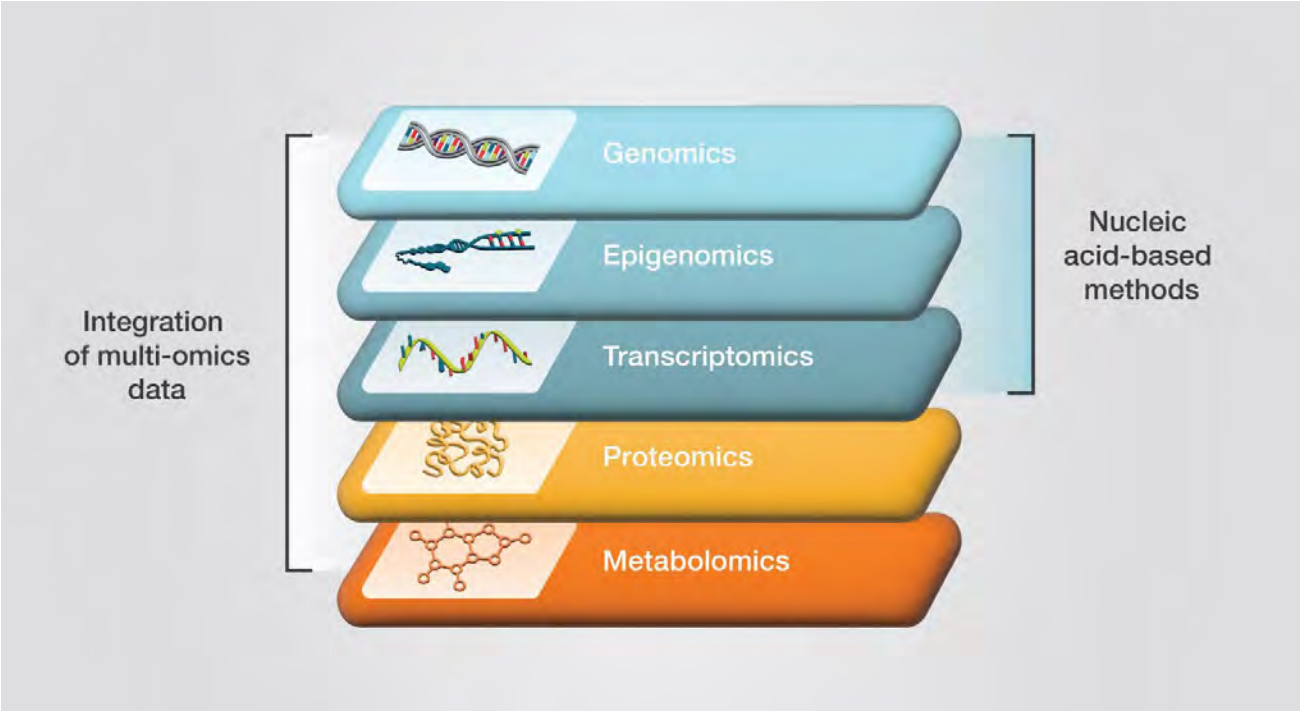
# Index

SEQUENTIA
DECODE • DISCOVER • DISRUPT

1. First part: omic data, metagenomics, metatranscriptomics and GAIA
    1. Omic data
    2. A little bit of history
    3. Applications
    4. Strategies: amplicon-based and shotgun
    5. Bioinformatic approaches and limitations
    6. GAIA
    7. Future insights

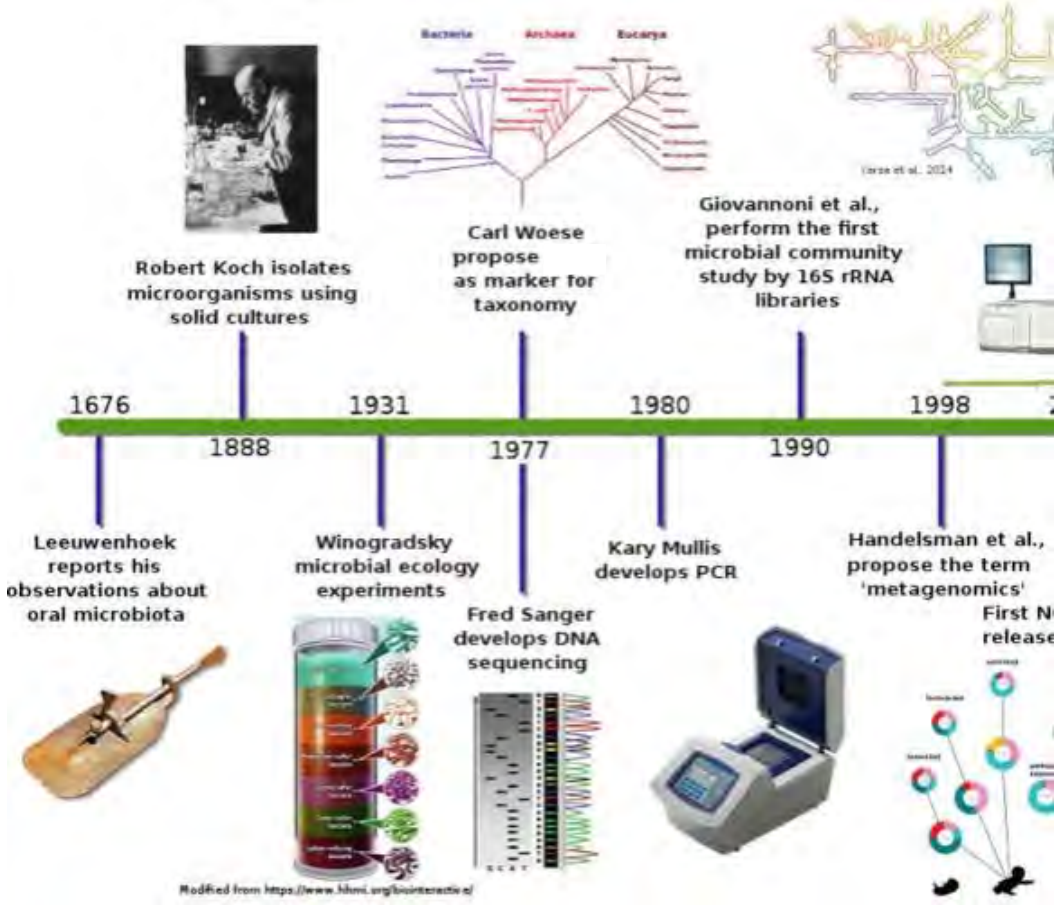
1. Second part: RNA -seq and AIR
    1. RNA-seq introduction
    2. Workflow
    3. Differential expression analysis
    4. AIR

# Omic data


Genomics
Epigenomics
Transcriptomics
Proteomics
Metabolomics
Integration of multi-omics data
Nucleic acid-based methods

# Metagenomics: Intro

# A little bit of history...




**Metagenomics** : application of genomics techniques without the need to do cell cultures in order to study a microbial community .

## Genetic diversity in Sargasso Sea bacterioplankton

**Stephen J. Giovannoni, Theresa B. Britschgi, Craig L. Moyer & Katharine G. Field**

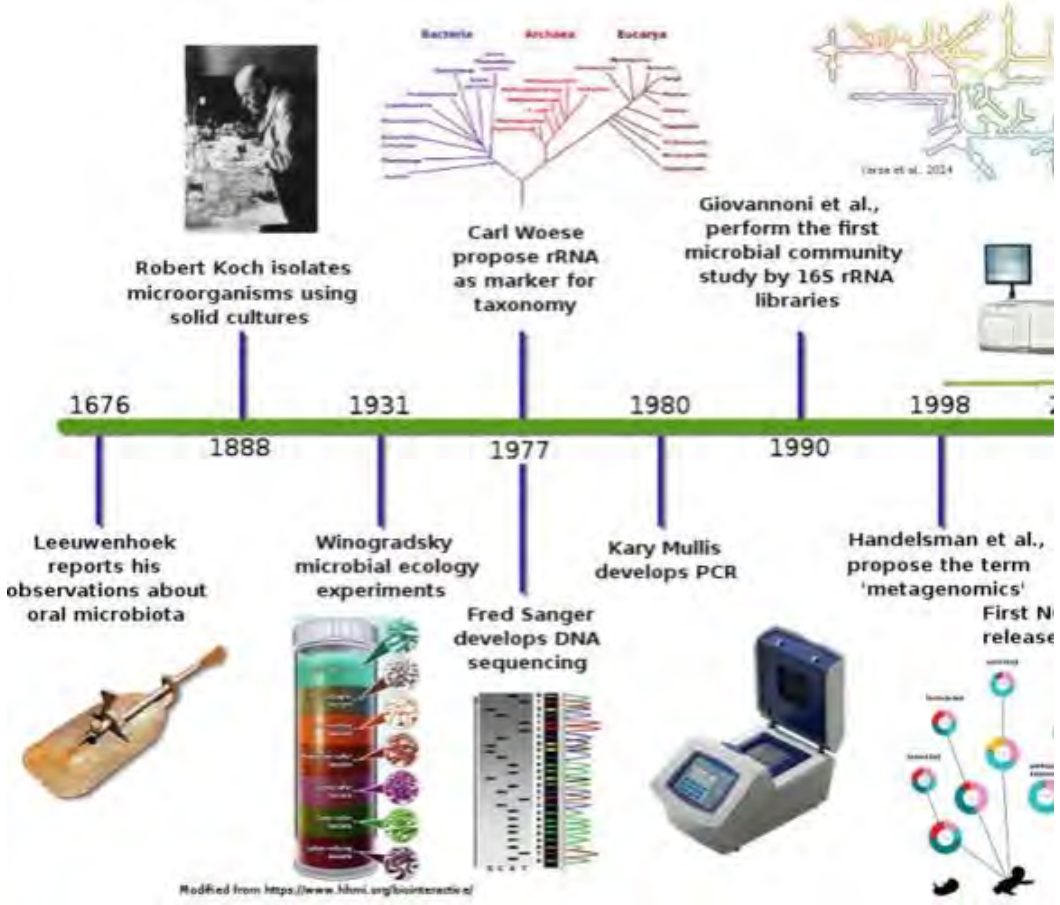
Department of Microbiology, Oregon State University, Corvallis, Oregon 97331, USA

BACTERIOPLANKTON are recognized as important agents of biogeochemical change in marine ecosystems, yet relatively little is known about the species that make up these communities.

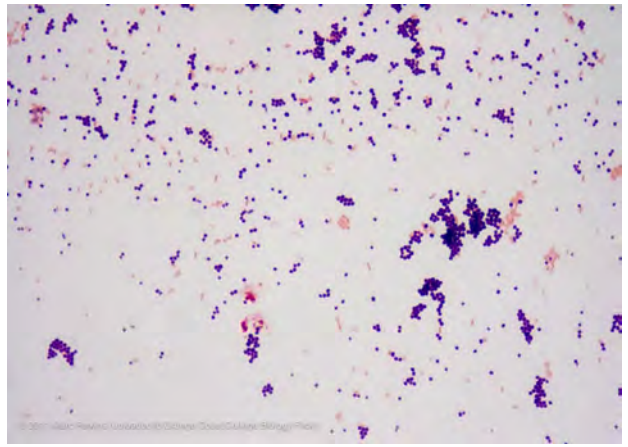
# Question 2

SEQUENTIA
DECODE • DISCOVER • DISRUPT

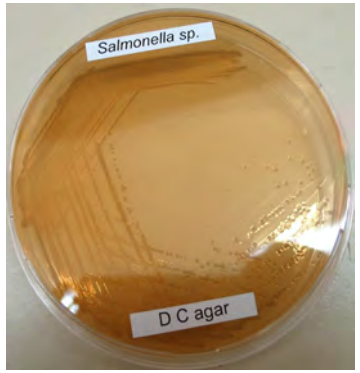
# Un poco de historia…



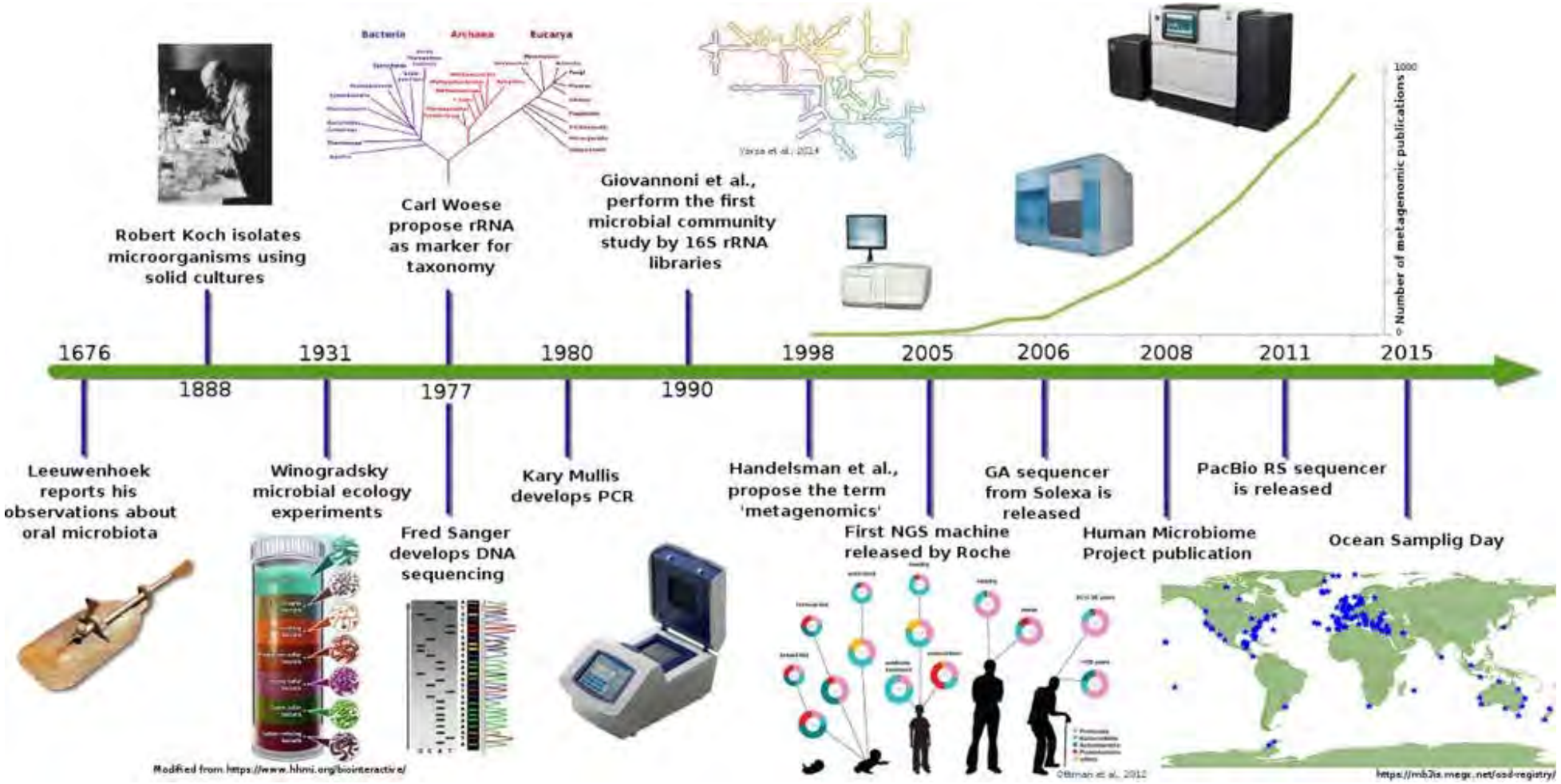

Microscope and staining (e.g. Gram)



Cell cultures



# Un poco de historia…

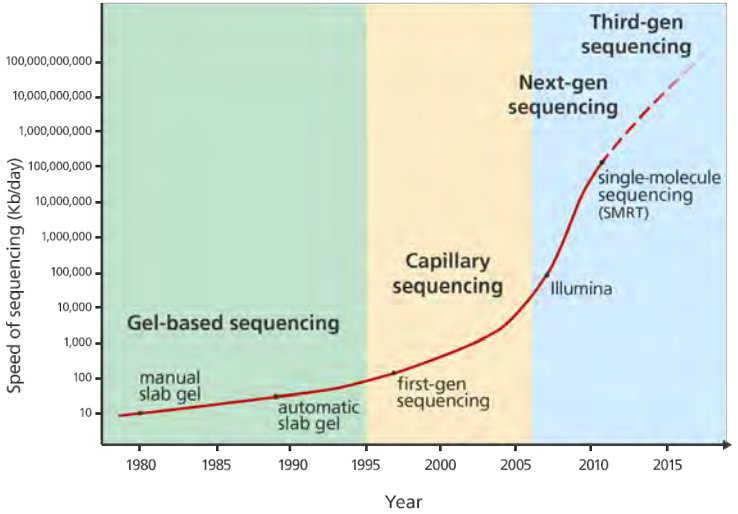


# What is Next -Generation Sequencing?

SEQUENTIA
DECODE • DISCOVER • DISRUPT

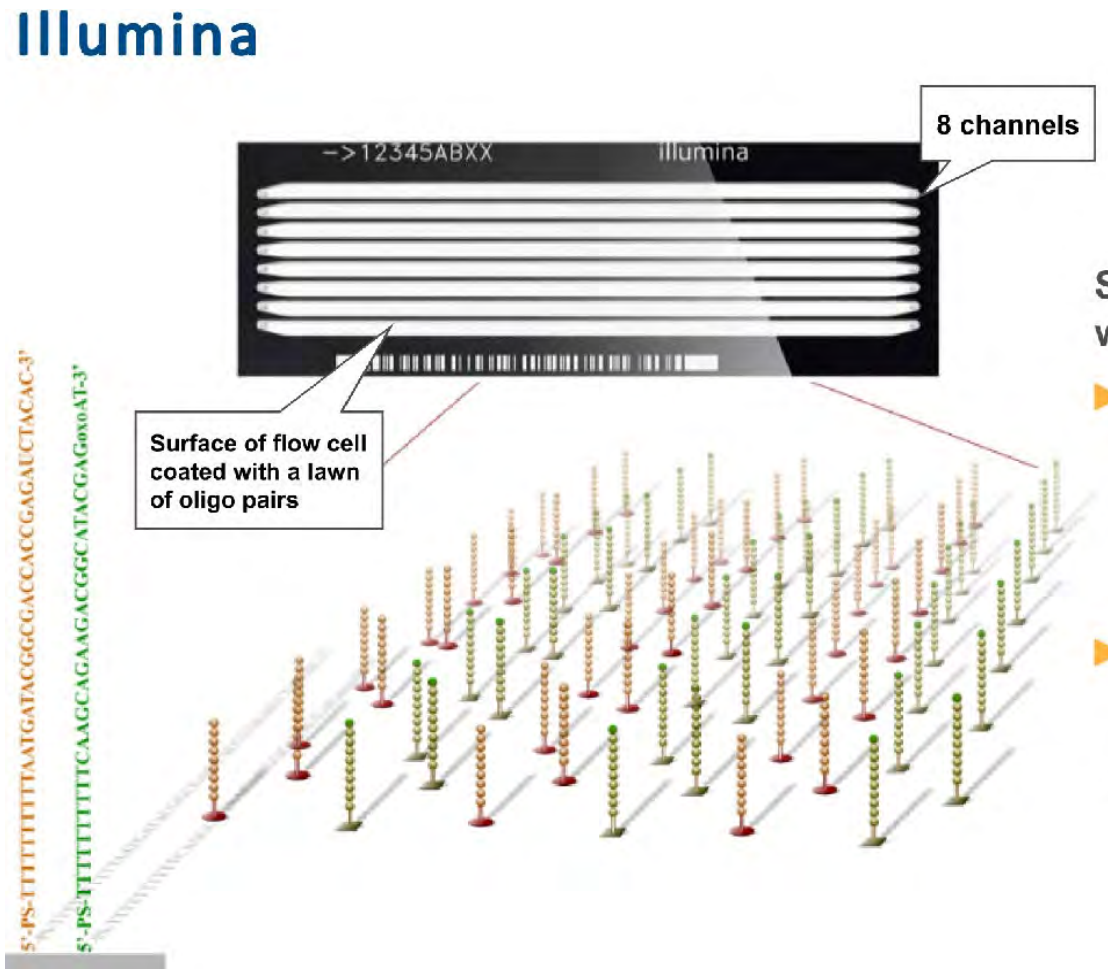
- The high demand for low-cost sequencing has driven the development of high-throughput sequencing (or next-generation sequencing) technologies that parallelize the sequencing process, producing thousands or millions of sequences concurrently.

- High-throughput sequencing technologies are intended to lower the cost of DNA sequencing beyond what is possible with standard dye-terminator methods.



# Illumina sequencing

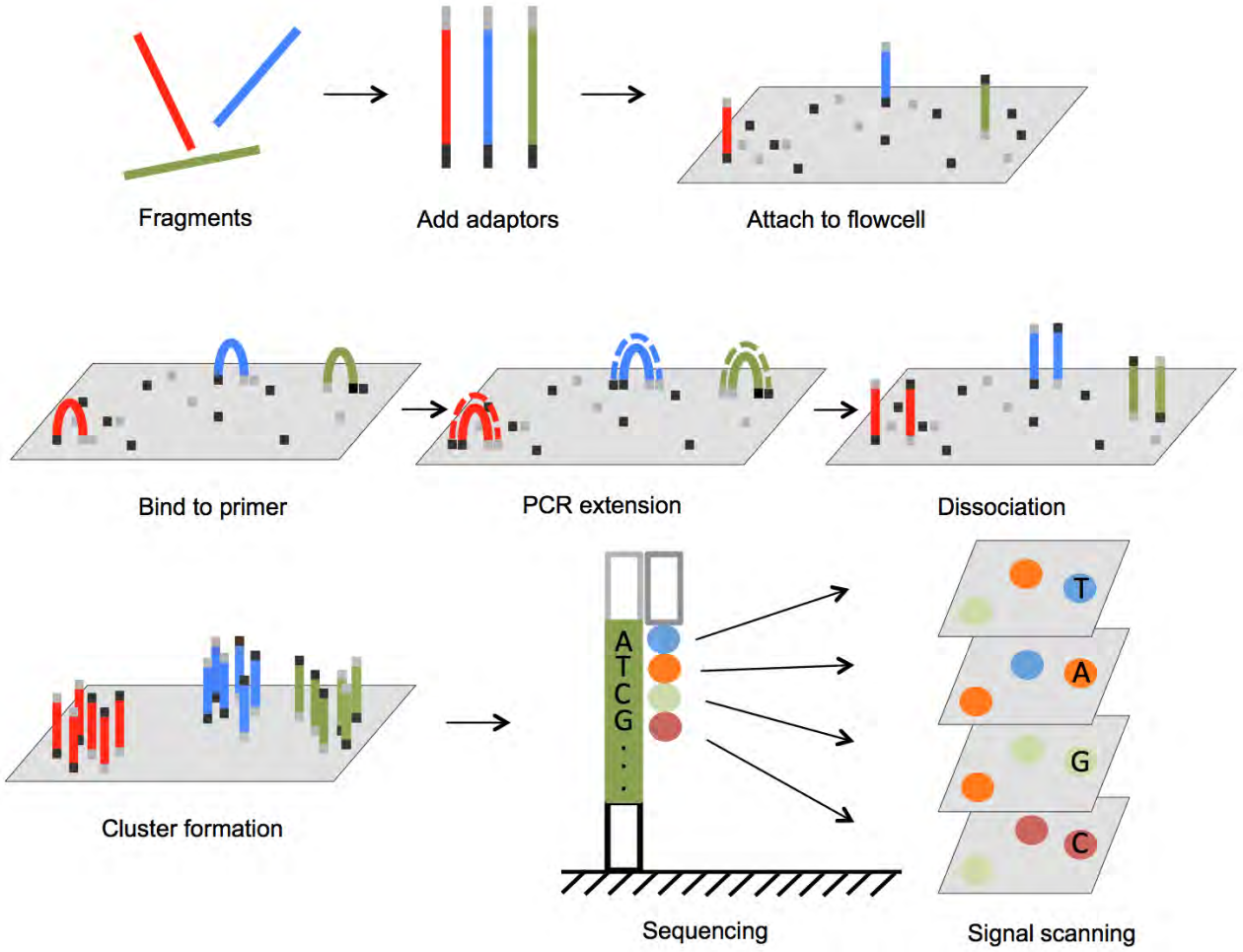
SEQUENTIA
DECODE • DISCOVER • DISRUPT



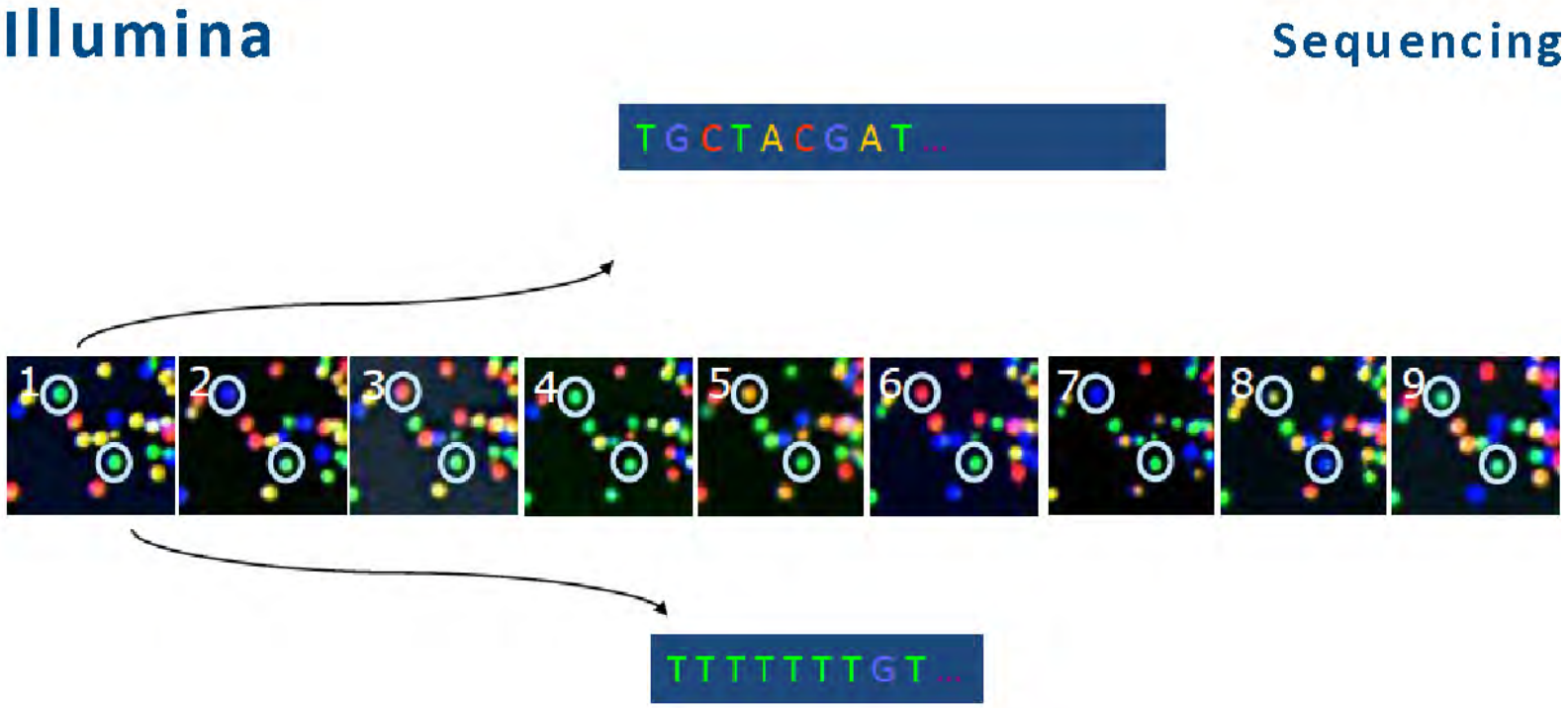

**Simplified workflow**

- Clusters in a contained environment (no need for clean rooms)
- Sequencing performed in the flow cell on the clusters

# Illumina sequencing



# Illumina sequencing




The identity of each base of a cluster is read off from sequential images.

# Fastq format

SEQUENTIA
DECODE • DISCOVER • DISRUPT

## Reads. What do they look like?

*FASTQ*

```
1  @SOLEXA2_0414:3:1:19459:1418#0/1
2  NTGCGATCTCATGGACAAACCAGACCTTACAACTGTTACTCTGAAT...
3  +SOLEXA2_0414:3:1:19459:1418#0/1
4  BGGIFMRPOO__________P__T_YYYYRYYYYM[[[[__Y___...
```

**Repeated blocks, four lines each:**

**1** … header, starting with “@”
**2** … sequence
**3** … header, starting with “+” (often left blank)
**4** … base qualities (same length as sequence)

# Fastq format

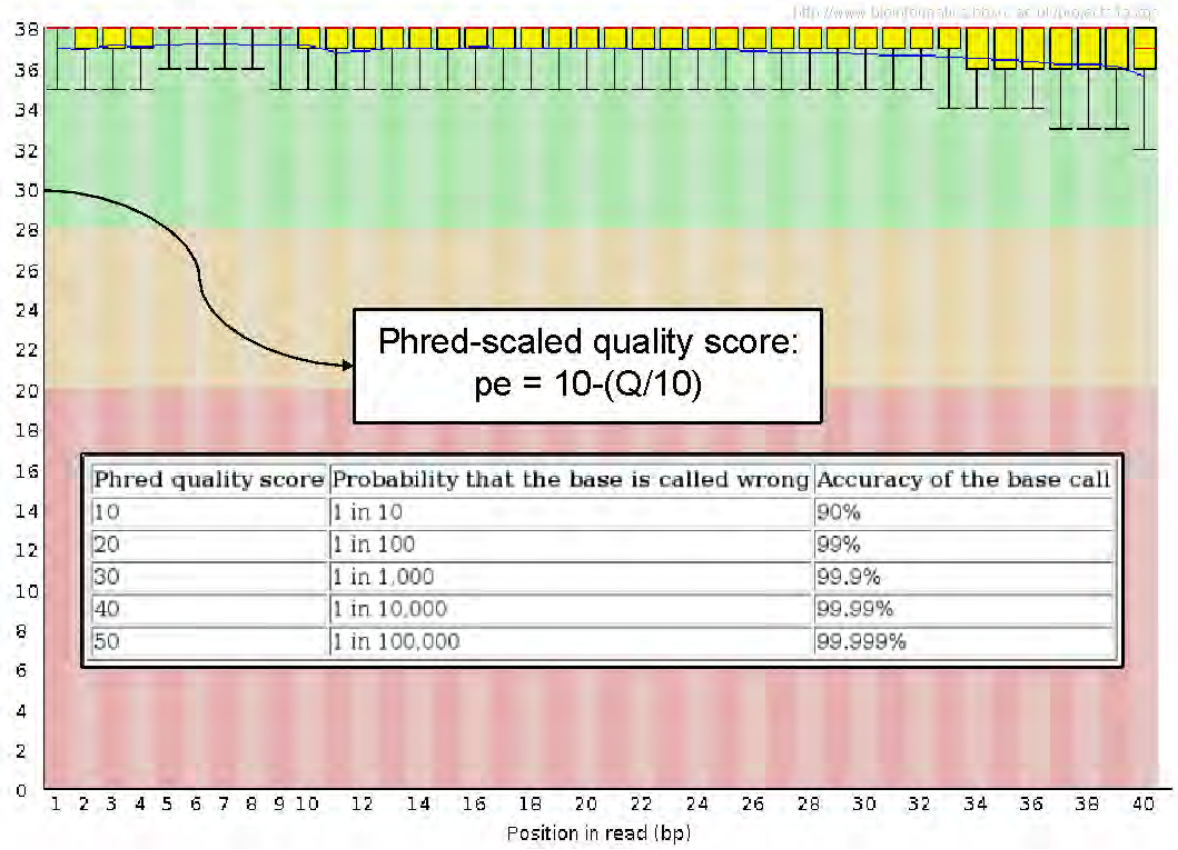
SEQUENTIA
DECODE • DISCOVER • DISRUPT

## Quality Score

- Each base position in a sequence comes with a “quality score”.
- This measures the probability that a base is called incorrectly, by a phred-like algorithm similar to that originally developed for Sanger sequencing experiments.
- The quality score of a given base, Q, is defined by $Q = -10*\log_{10}(e)$ where e is the estimated probability of the base call being wrong.
- A quality score of 20 represents an error rate of 1 in 100, with a corresponding call accuracy of 99%.

# Fastq format

## Basecall Qualities




| Phred quality score | Probability that the base is called wrong | Accuracy of the base call |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1,000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 99.999% |

# GAIA

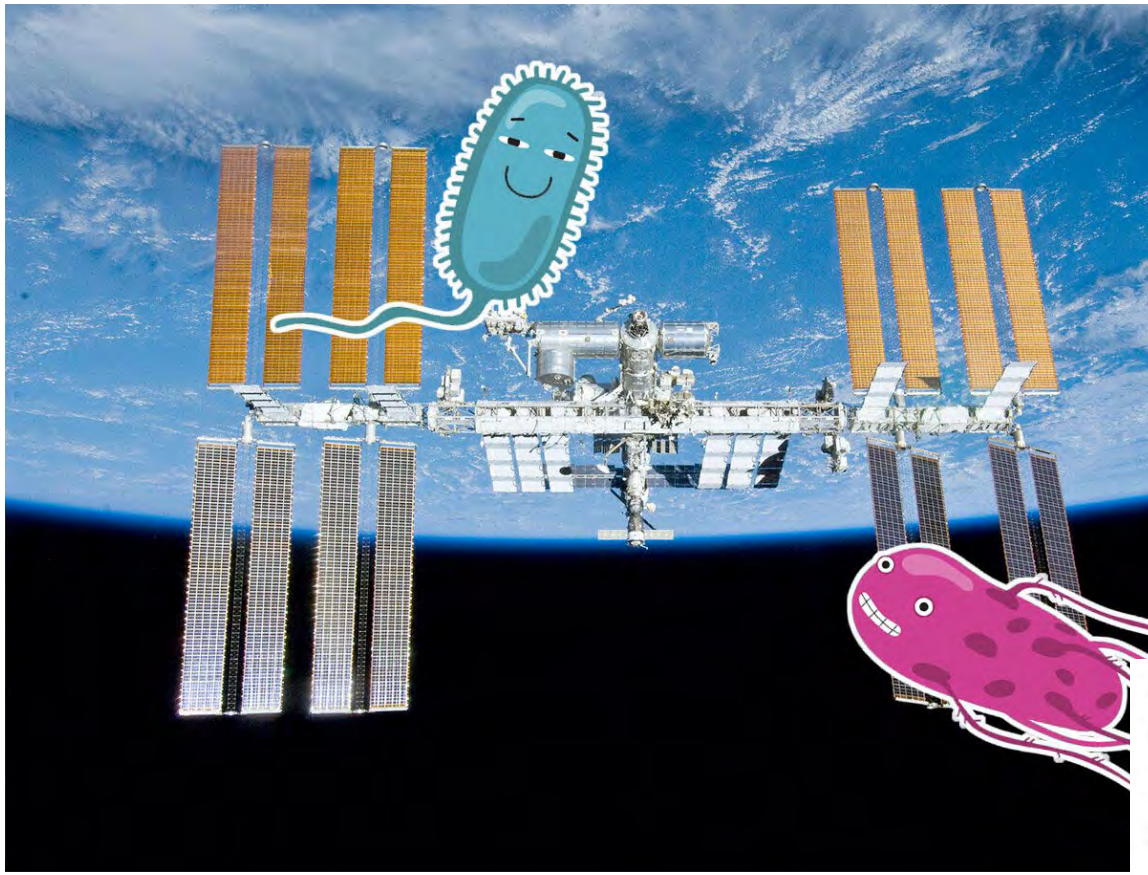
SEQUENTIA
DECODE • DISCOVER • DISRUPT

https://metagenomics.sequentiabiotech.com/gaia/

# Applications

SEQUENTIA
DECODE • DISCOVER • DISRUPT

# Applications



## *Methylobacterium ajmalii* sp. nov., Isolated From the International Space Station

*Swati Bijlani[1], Nitin K. Singh[2], V. V. Ramprasad Eedara[3], Appa Rao Podile[3], Christopher E. Mason[4], Clay C. C. Wang[1*] and Kasthuri Venkateswaran[2*]*

[1] Department of Pharmacology and Pharmaceutical Sciences, School of Pharmacy, University of Southern California, Los Angeles, CA, United States, [2] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, United States, [3] Department of Plant Science, School of Life Sciences, University of Hyderabad, Hyderabad, India, [4] WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, United States

# Question 3

SEQUENTIA
DECODE • DISCOVER • DISRUPT

# Strategy: amplicon -based

SEQUENTIA
DECODE • DISCOVER • DISRUPT

**Amplicon metagenomics**

computationally inexpensive

economically cheap

kingdom - limited

lack of functional identification

family/genus level

# Strategy: amplicon -based

SEQUENTIA
DECODE • DISCOVER • DISRUPT

Amplicon metagenomics

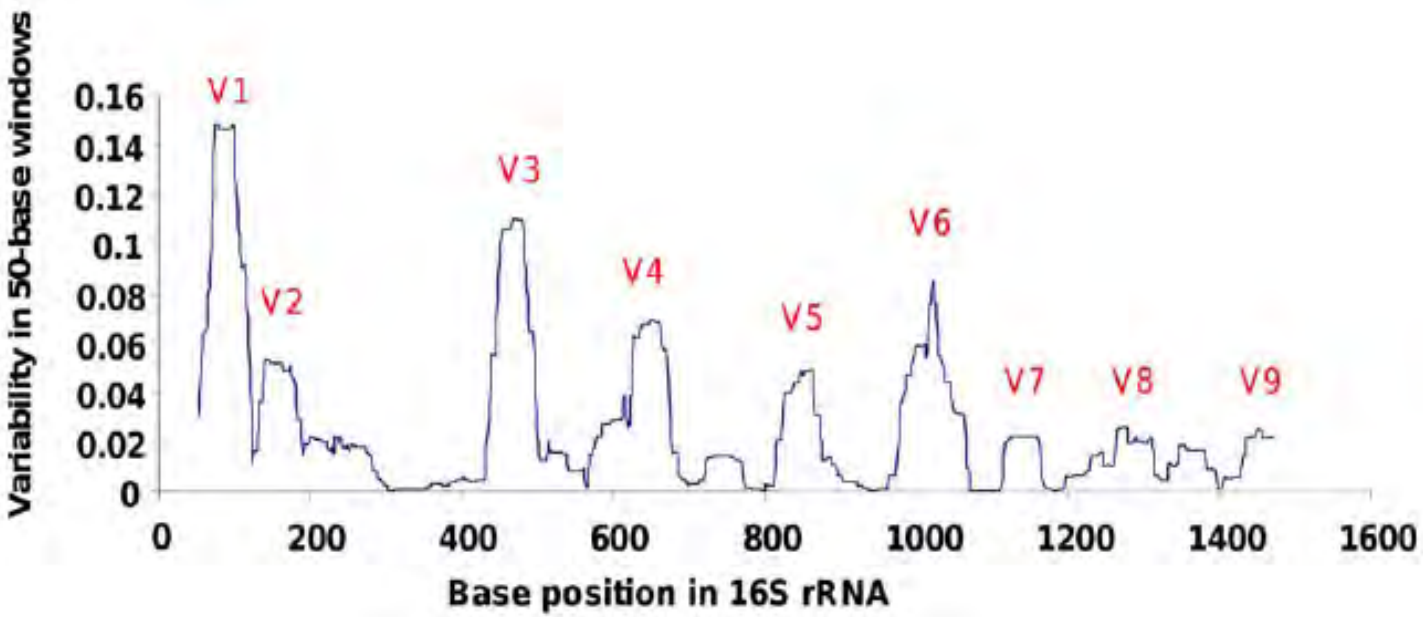
computationally inexpensive

economically cheap

kingdom - limited

lack of functional identification

family/genus level



# Strategy: amplicon -based

SEQUENTIA
DECODE • DISCOVER • DISRUPT

**Amplicon metagenomics**

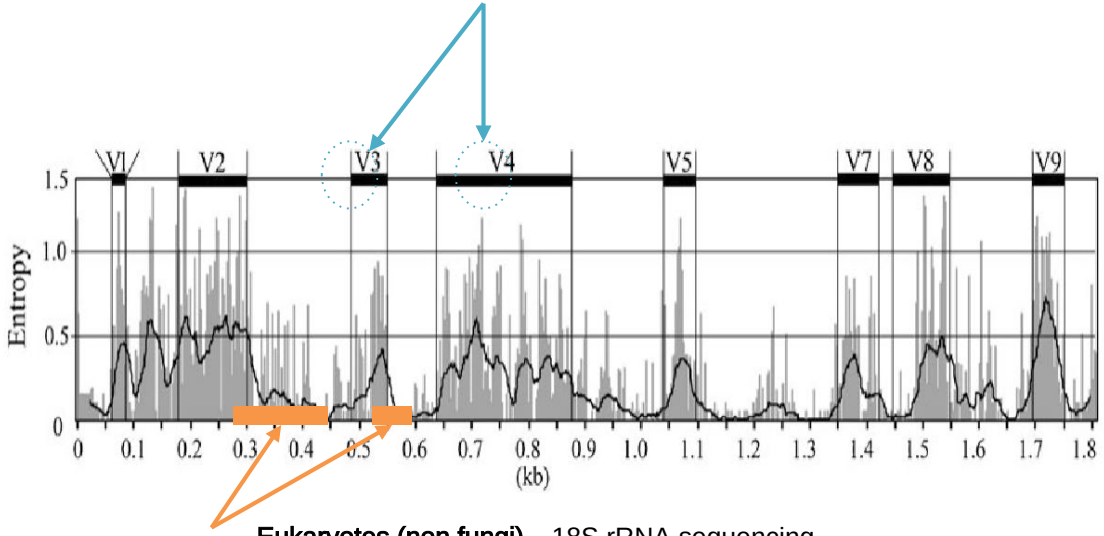
computationally inexpensive

economically cheap

kingdom - limited

lack of functional identification

family/genus level




**Eukaryotes (non fungi)** . 18S rRNA sequencing.
Source: Jang Seu Ki. 2012. doi: 10.1007/s10811-011-9730-z

# Strategy: amplicon -based

SEQUENTIA
DECODE • DISCOVER • DISRUPT

**Amplicon metagenomics**

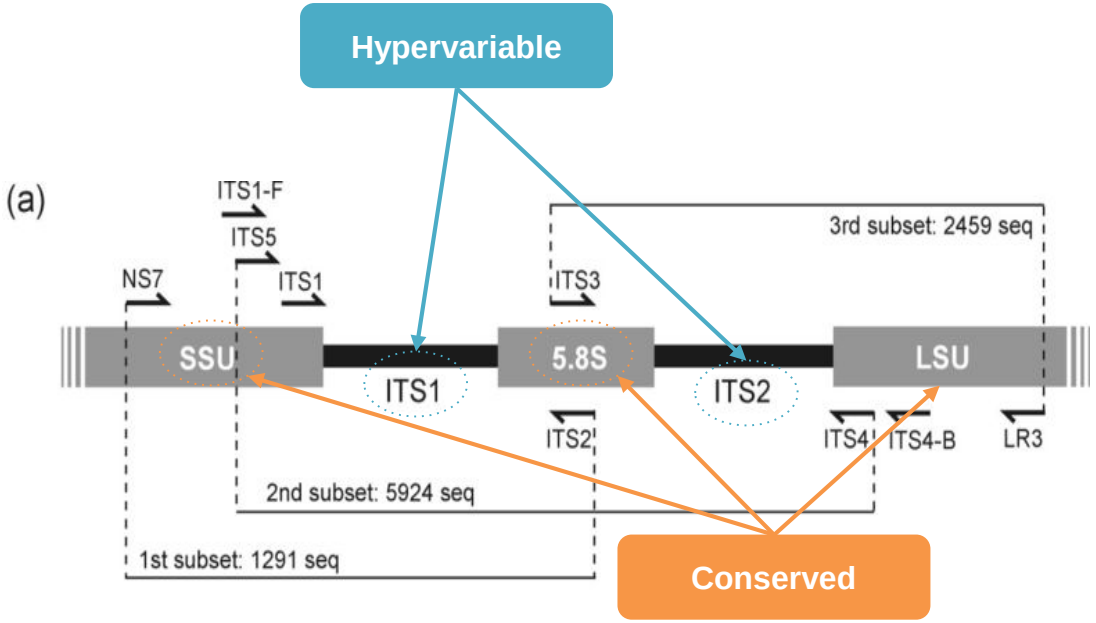
computationally inexpensive

economically cheap

kingdom - limited

lack of functional identification

family/genus level




**Fungi** . ITS sequencing.
Source: Bellemain, *et al.* 2010. doi: 10.1186/1471-2180-10-189

# Strategy: amplicon -based

SEQUENTIA
DECODE • DISCOVER • DISRUPT

**Amplicon metagenomics**

computationally inexpensive

economically cheap

kingdom - limited

lack of functional identification

family/genus level

Sequencing platforms commonly used

Roche 454 → 400-700 bp (obsolete)

Illumina MiSeq v3 →2 x 300 bp

Ion PGM →400 bp

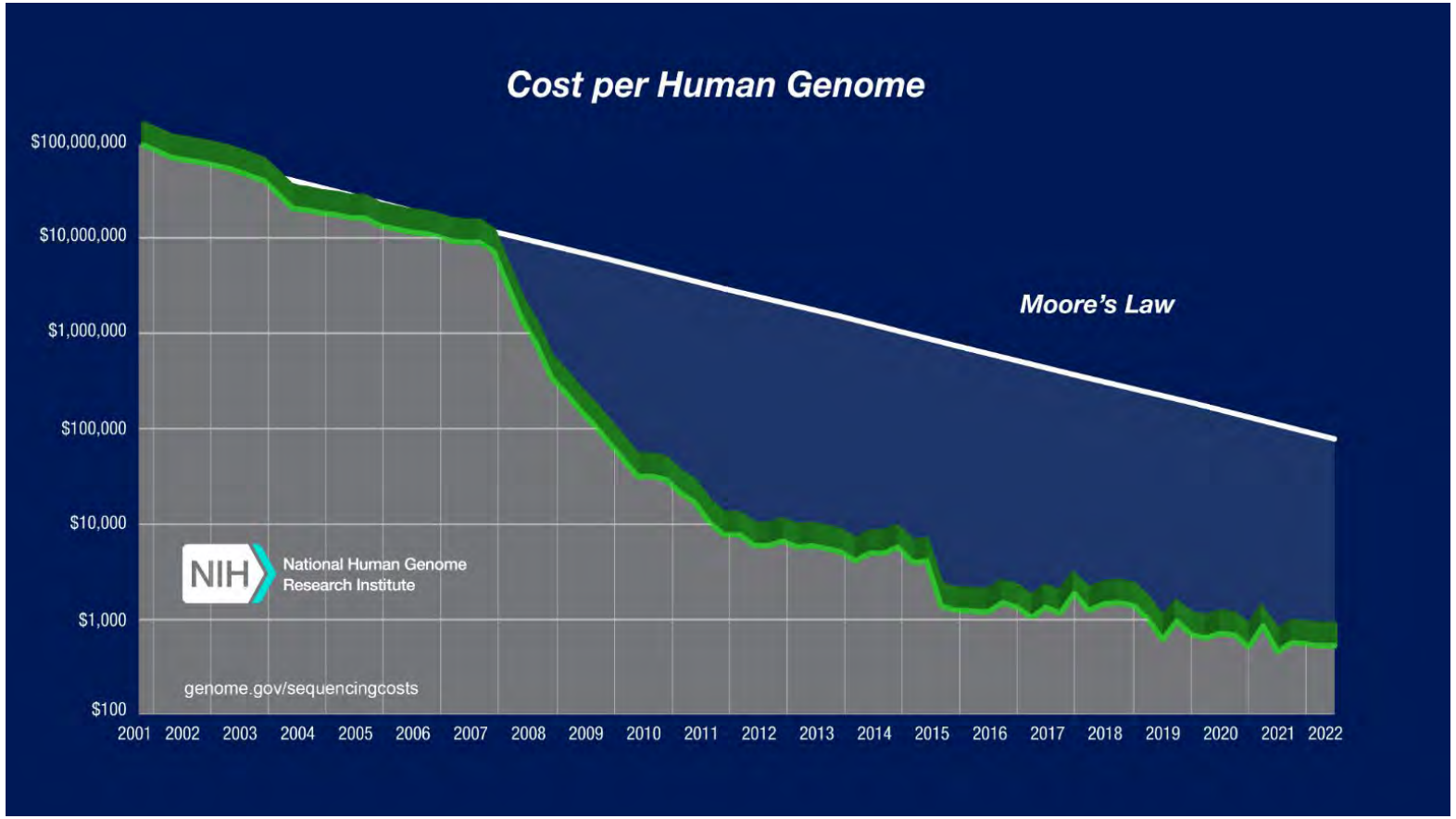
ion torrent by life technologies™

illumina®

# Question 4

SEQUENTIA
DECODE • DISCOVER • DISRUPT

# Moore's Law


Cost per Human Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
$100
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts

# Strategy: shotgun

SEQUENTIA
DECODE • DISCOVER • DISRUPT

## Amplicon metagenomics

computationally inexpensive

economically cheap

kingdom - limited

lack of functional identification

family/genus level

## Shotgun metagenomics

### Whole Genome Sequencing (WGS) metagenomics

strain level

functional identification

unable to know the gene expression abundance

computationally expensive

economically expensive (+)

### Metatranscriptomics

strain level

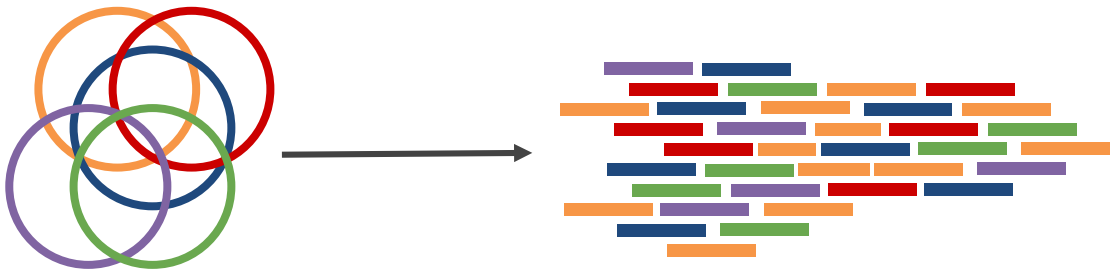
functional identification of expressing genes

differential expression between conditions

computationally expensive

economically expensive (++)

# Strategy: shotgun



Sequencing platforms commonly used

Illumina MiSeq v3 → 2 x 300 bp

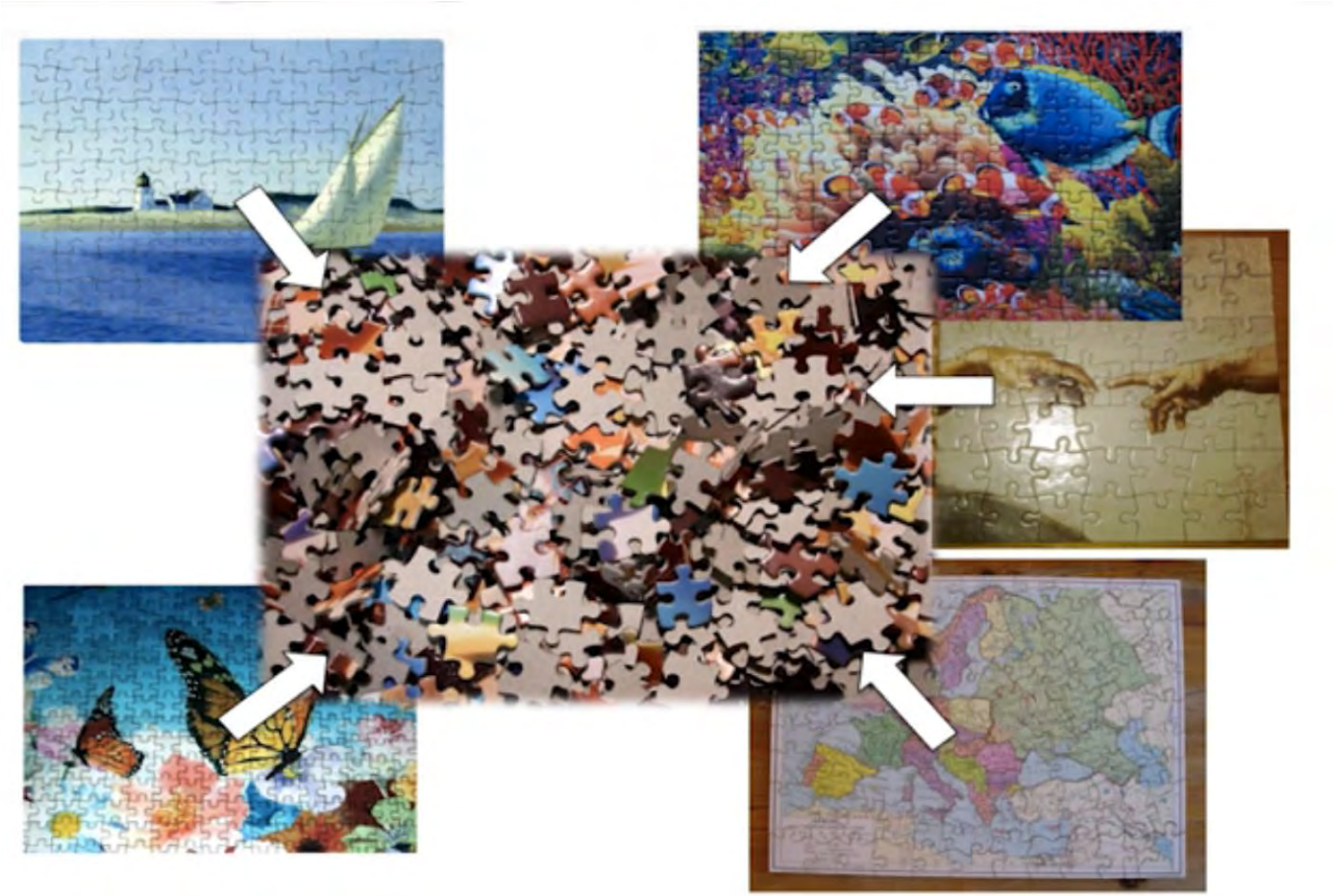
Illumina NextSeq 500 →2 x 150 bp

Illumina HiSeq 3000 →2 x 150 bp

# Question 5

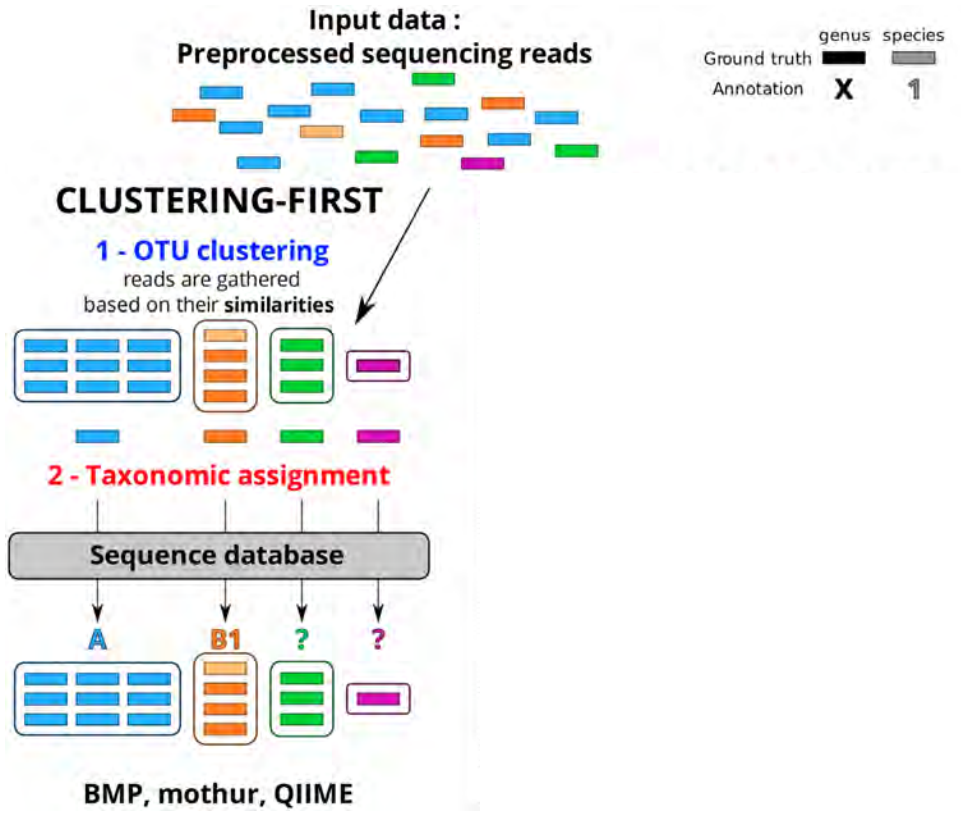
SEQUENTIA
DECODE • DISCOVER • DISRUPT

# Strategy: shotgun



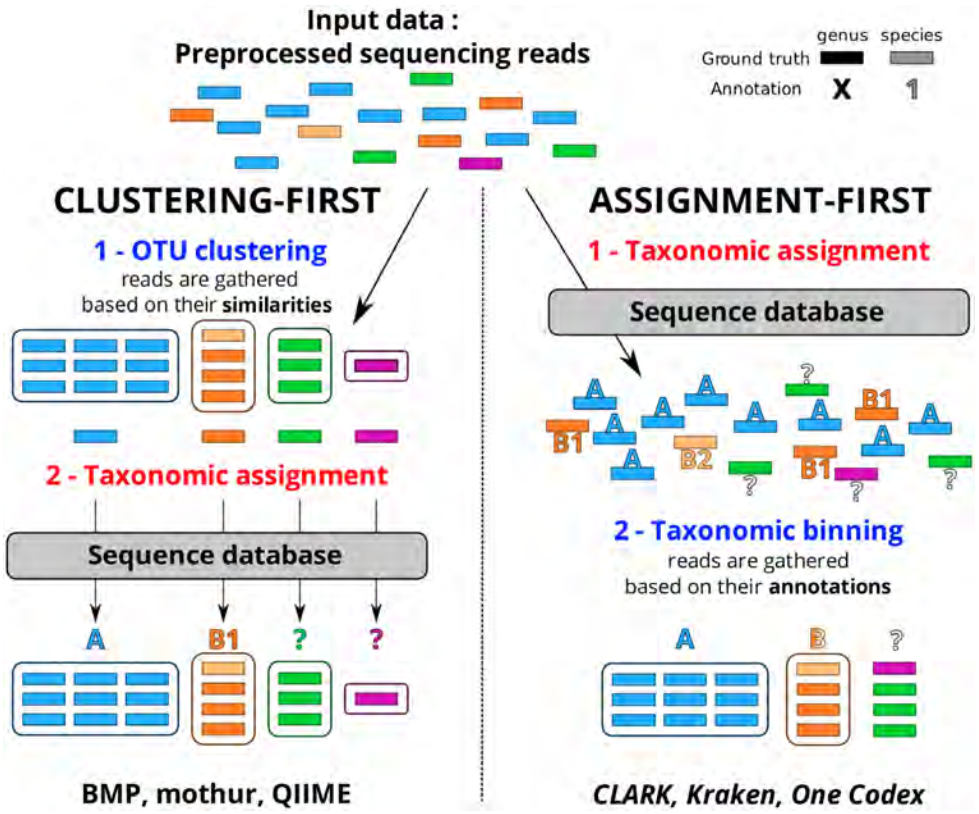
# Bioinformatic approaches

Suitable for
**amplicon - based**
metagenomics



# Bioinformatic approaches

Suitable for **amplicon - based** metagenomics




Suitable for **amplicon - based** and **shotgun - based** metagenomics

# Question 6

SEQUENTIA
DECODE • DISCOVER • DISRUPT

# Bioinformatic limitations

1. There is still room for improvement in terms of **accuracy**
2. The vast majority of tools only work on **prokaryotes**
3. Lack of true **user-friendly** interfaces for the vast majority of the pipelines
4. **Computational power** and **data storage** required
5. **High** processing time

solving it!

# GAIA: pipeline

SEQUENTIA
DECODE • DISCOVER • DISRUPT

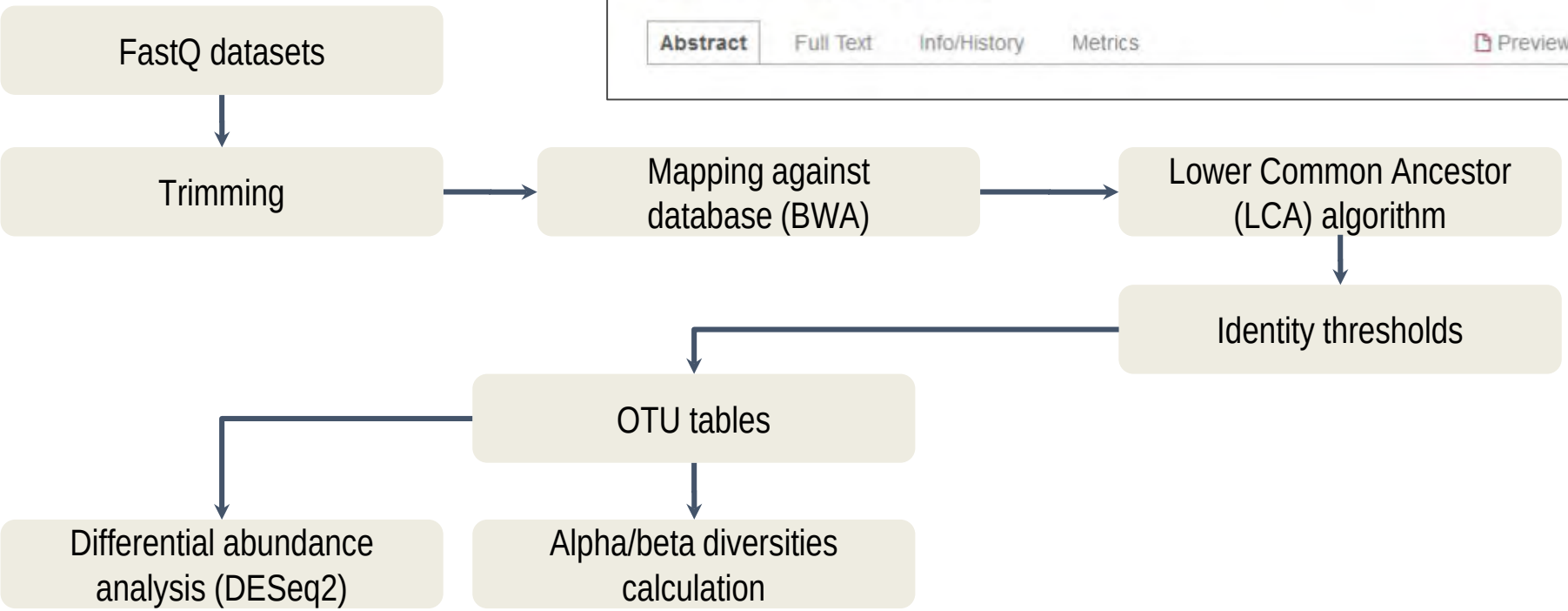
New Results

**GAIA: an integrated metagenomics suite**

A. Paytuvi, E. Battista, F. Scippacercola, R. Aiese Cigliano, W. Sanseverino

**doi:** https://doi.org/10.1101/804690

**Abstract** Full Text Info/History Metrics Preview PDF



# GAIA: OTU table

SEQUENTIA
DECODE • DISCOVER • DISRUPT

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | otu | X10n | X10p | X11n | X11p | X120n | X120p | X121n | X121p | X122n | X122p | X125n | X125p | X126n | X126p | X127n | X13n | X13p | X140n | X140p | X141n |
| 2 | Otu001 | 13679 | 6292 | 42 | 2500 | 18850 | 5 | 43 | 7138 | 9432 | 10541 | 9 | 9772 | 1388 | 7 | 31538 | 38 | 2338 | 23 | 9 | 1358 |
| 3 | Otu002 | 18 | 7134 | 38 | 9830 | 45 | 61420 | 182 | 23751 | 36 | 11 | 4535 | 3502 | 11018 | 5473 | 26 | 14411 | 38 | 19018 | 12 | 3080 |
| 4 | Otu003 | 9939 | 8983 | 31 | 13 | 24620 | 19 | 19 | 16 | 12502 | 3831 | 4621 | 2240 | 9924 | 4052 | 9292 | 18 | 0 | 37 | 7 | 3680 |
| 5 | Otu004 | 3675 | 4234 | 24 | 22 | 11 | 16 | 32967 | 35 | 6 | 18 | 6908 | 5 | 16 | 8702 | 24 | 11 | 37717 | 0 | 25 | 4196 |
| 6 | Otu005 | 0 | 5 | 0 | 7 | 0 | 8 | 0 | 16 | 20166 | 0 | 0 | 2 | 5 | 8 | 2 | 16 | 0 | 13 | 0 | 0 |
| 7 | Otu006 | 0 | 8 | 0 | 0 | 0 | 8 | 0 | 0 | 5 | 3 | 3 | 0 | 0 | 9 | 0 | 5 | 4 | 0 | 0 | 3 |
| 8 | Otu007 | 4587 | 518 | 4 | 386 | 8775 | 5 | 6 | 1102 | 14336 | 0 | 0 | 3626 | 51 | 0 | 6 | 12 | 0 | 10 | 0 | 395 |
| 9 | Otu008 | 1 | 8 | 2 | 4408 | 3 | 29 | 6 | 12355 | 0 | 0 | 0 | 0 | 0 | 9 | 3 | 1588 | 0 | 6 | 3 | 3 |
| 10 | Otu009 | 115 | 914 | 3 | 325 | 0 | 629 | 1 | 834 | 5 | 0 | 1354 | 2108 | 1117 | 67 | 0 | 2010 | 1897 | 11227 | 1 | 3 |
| 11 | Otu010 | 780 | 8 | 23810 | 12 | 3279 | 0 | 12 | 7 | 3027 | 0 | 2 | 4156 | 0 | 0 | 18 | 0 | 0 | 0 | 0 | 0 |
| 12 | Otu011 | 0 | 3 | 2 | 2 | 0 | 13 | 5 | 5 | 4 | 7 | 3081 | 11 | 4 | 6804 | 0 | 3 | 11 | 0 | 5 | 0 |
| 13 | Otu012 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 16 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 0 | 6 | 0 | 0 |
| 14 | Otu013 | 6321 | 2471 | 2 | 0 | 12 | 3 | 0 | 0 | 4 | 20272 | 0 | 15 | 9 | 0 | 5 | 0 | 11 | 0 | 14 | 0 |
| 15 | Otu014 | 0 | 82 | 4 | 3304 | 1 | 1667 | 4 | 9233 | 13 | 3 | 0 | 2707 | 0 | 0 | 3 | 4806 | 9 | 3 | 5 | 0 |
| 16 | Otu015 | 0 | 12 | 0 | 3 | 7 | 25 | 1 | 6 | 10 | 0 | 4 | 2772 | 1 | 3 | 0 | 2 | 0 | 10 | 13 | 8052 |
| 17 | Otu016 | 1 | 0 | 0 | 9 | 5 | 0 | 0 | 14 | 0 | 0 | 0 | 0 | 2654 | 0 | 0 | 6 | 1 | 1 | 0 | 0 |
| 18 | Otu017 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 8 | 0 | 0 | 0 | 0 | 0 | 17 | 24 | 48 | 35210 | 4 |
| 19 | Otu018 | 1 | 0 | 9 | 911 | 0 | 0 | 15 | 2702 | 6 | 4 | 342 | 2217 | 606 | 0 | 13 | 3846 | 4 | 6 | 8513 | 1 |
| 20 | Otu019 | 0 | 0 | 13 | 0 | 0 | 0 | 29 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 0 | 0 | 5 | 4 | 0 |
| 21 | Otu020 | 425 | 0 | 1 | 0 | 1706 | 0 | 8447 | 1 | 0 | 0 | 0 | 0 | 0 | 26 | 0 | 0 | 3490 | 0 | 2620 | 0 |
| 22 | Otu021 | 0 | 4 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 4 |
| 23 | Otu022 | 0 | 0 | 0 | 4987 | 0 | 0 | 0 | 6 | 90 | 1 | 1 | 524 | 0 | 467 | 0 | 4 | 8 | 6198 | 0 | 1 |
| 24 | Otu023 | 4 | 0 | 1 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 3351 | 3 | 0 | 3910 | 1 | 2 | 3 | 0 |
| 25 | Otu024 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 |
| 26 | Otu025 | 69 | 0 | 0 | 0 | 290 | 0 | 0 | 0 | 21 | 0 | 118 | 2 | 9 | 513 | 2 | 0 | 0 | 2 | 0 | 0 |
| 27 | Otu026 | 0 | 2 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 28 | Otu027 | 6 | 2304 | 0 | 0 | 5 | 0 | 0 | 0 | 57 | 4 | 0 | 14529 | 9597 | 2 | 6 | 0 | 0 | 0 | 0 | 0 |

# GAIA: benchamrk

SEQUENTIA
DECODE • DISCOVER • DISRUPT

Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics

Léa Siegwald, Hélène Touzet, Yves Lemoine, David Hot, Christophe Audebert, Ségolène Caboche

Published: January 4, 2017 • https://doi.org/10.1371/journal.pone.0169563

6,535 View | 10 Share

BMC | Explore journals | Get published | About BMC | Login | Search

Genome Biology

Home | About | Articles | Submission Guidelines

Abstract
Background
Results
Discussion
Conclusions
Methods
Declarations
References

Research | Open Access

Comprehensive benchmarking and ensemble approaches for metagenomic classifiers

Alexa B. R. McIntyre, Rachid Ounit, Ebrahim Afshinnekoo, Robert J. Prill, Elizabeth Hénaff, Noah Alexander, Samuel S. Minot, David Danko, Jonathan Foox, Sofia Ahsanuddin, Scott Tighe, Nur A. Hasan, Poorani Subramanian, Kelly Moffat, Shawn Levy, Stefano Lonardi, Nick Greenfield, Rita R. Colwell, Gail L. Rosen and Christopher E. Mason

*Genome Biology* 2017 **18**:182

https://doi.org/10.1186/s13059-017-1299-7 | © The Author(s). 2017

Received: 7 February 2017 | Accepted: 16 August 2017 | Published: 21 September 2017

Download PDF

Export citations

IN THESE COLLECTIONS

Microbiomes and Metagenomics

Metrics

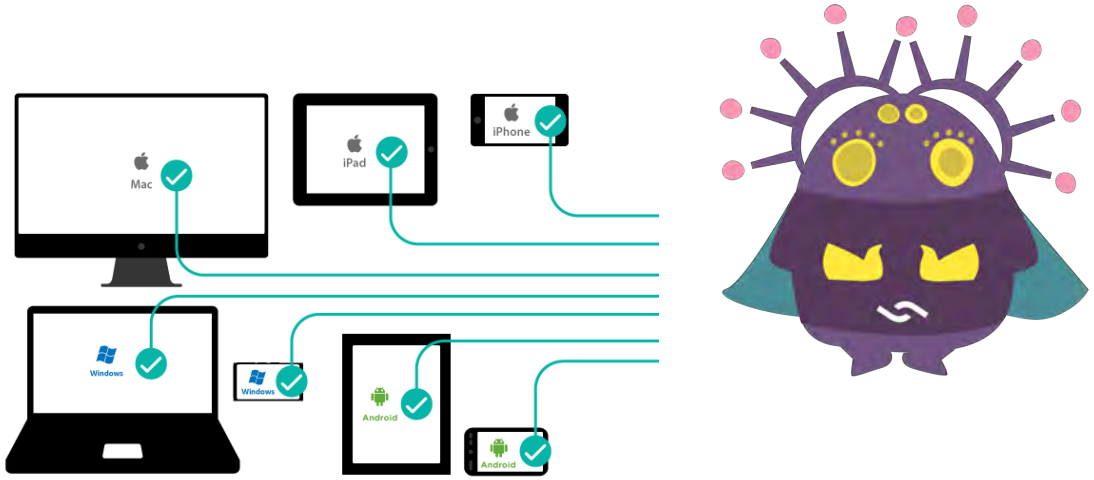
**precision** = # classified reads correctly / # classified reads

**recall** = # classified reads correctly / # total number of reads

**F-measure** = harmonic mean of precision and recall

# GAIA

SEQUENTIA
DECODE • DISCOVER • DISRUPT




gaia.sequentiabiotech.com

# Question 7

SEQUENTIA
DECODE • DISCOVER • DISRUPT

# GAIA

SEQUENTIA
DECODE • DISCOVER • DISRUPT

1. There is still room for improvement in terms of **accuracy** ✓
2. The vast majority of tools only work on **prokaryotes** ✓
3. Lack of true **user-friendly** interfaces for the vast majority of the pipelines ✓
4. **Computational power** and **data storage** required ✓
5. **High** processing time ✓

# Future

SEQUENTIA
DECODE • DISCOVER • DISRUPT

Oxford NANOPORE Technologies™

- **Portable** everywhere
- **10 min** library preparation
- Reads of up to **tens of kbp**

Drawbacks of Oxford Nanopore

- High error rate (~15% for R9 release)
- Low-coverage
- ~250 bp/s (0.9 Mbp/h)

~2000 times!

**~1.87 Gbp/h** (75 Gbp/40h) for **Illumina HiSeq**
(calculated using Rapid Run Mode with a single flow-cell, 2×150 bp)

Jain, M. *et al.* 2017. doi: 10.12688/f1000research.11354.1

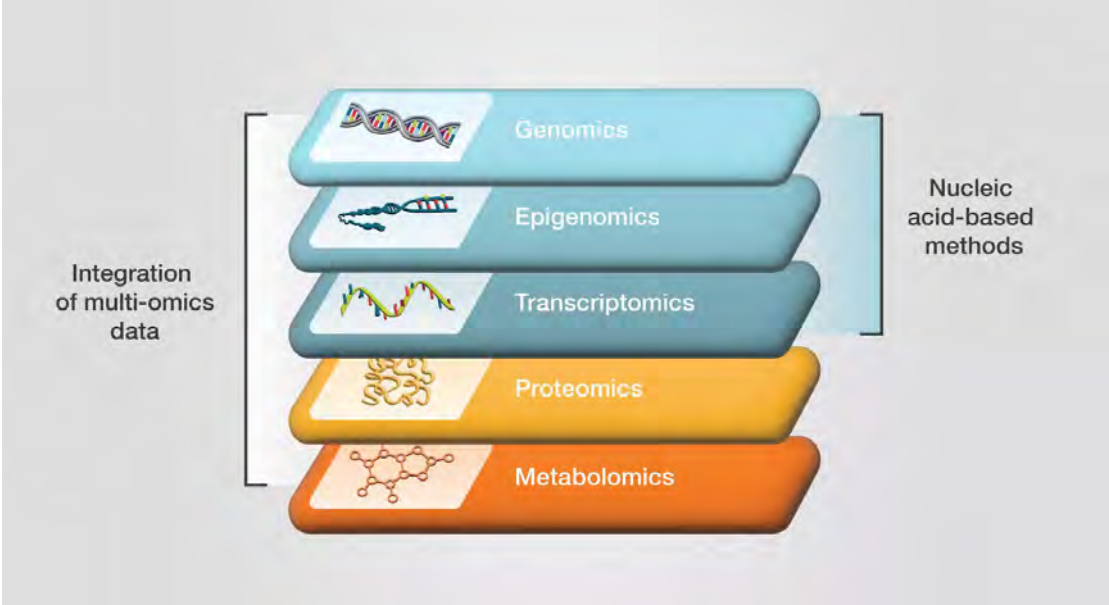
# Future

SEQUENTIA
DECODE • DISCOVER • DISRUPT

Oxford NANOPORE Technologies™

- **Portable** everywhere
- **10 min** library preparation
- Reads of up to **tens of kbp**

## Omics integration



# Future

SEQUENTIA
DECODE • DISCOVER • DISRUPT

Oxford NANOPORE Technologies™

- **Portable** everywhere
- **10 min** library preparation
- Reads of up to **tens of kbp**

Personalized medicine

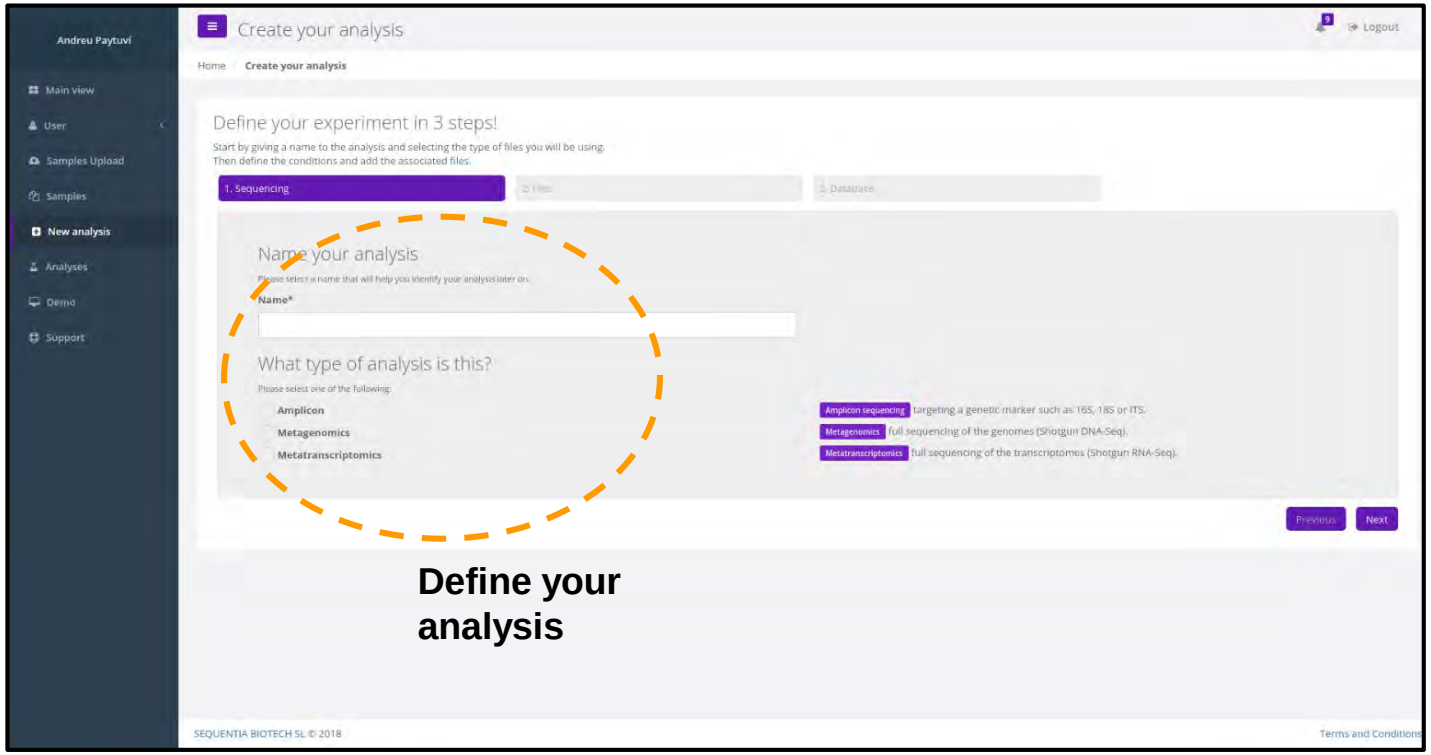
# Question 8

SEQUENTIA
DECODE • DISCOVER • DISRUPT

# GAIA

SEQUENTIA
DECODE • DISCOVER • DISRUPT

## Gaia user interface

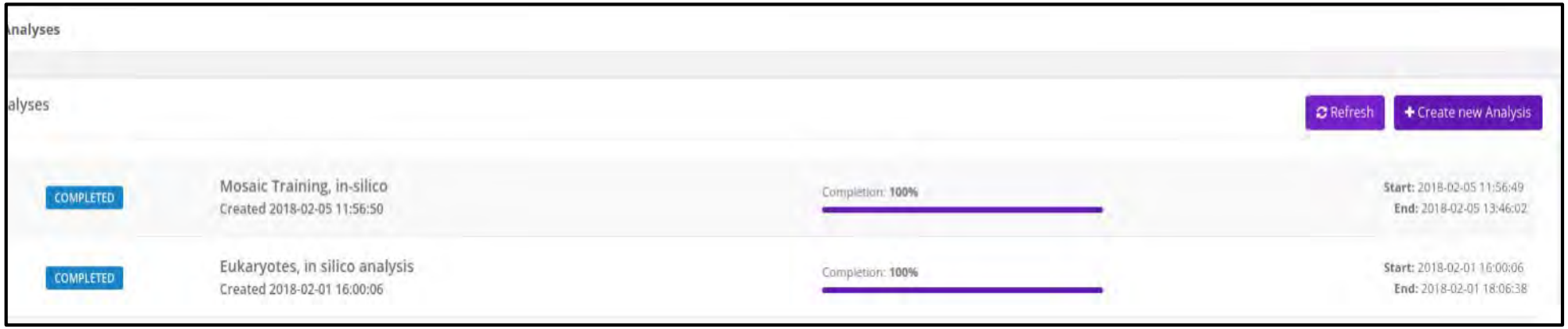



**Define your analysis**

# GAIA

SEQUENTIA
DECODE • DISCOVER • DISRUPT

**Gaia user interface**



# GAIA

SEQUENTIA
DECODE • DISCOVER • DISRUPT

## Alpha-diversities

Alpha-diversity (OTUs observed and Chao1, Shannon, Simpson, and Fisher indices) Fast Filtering

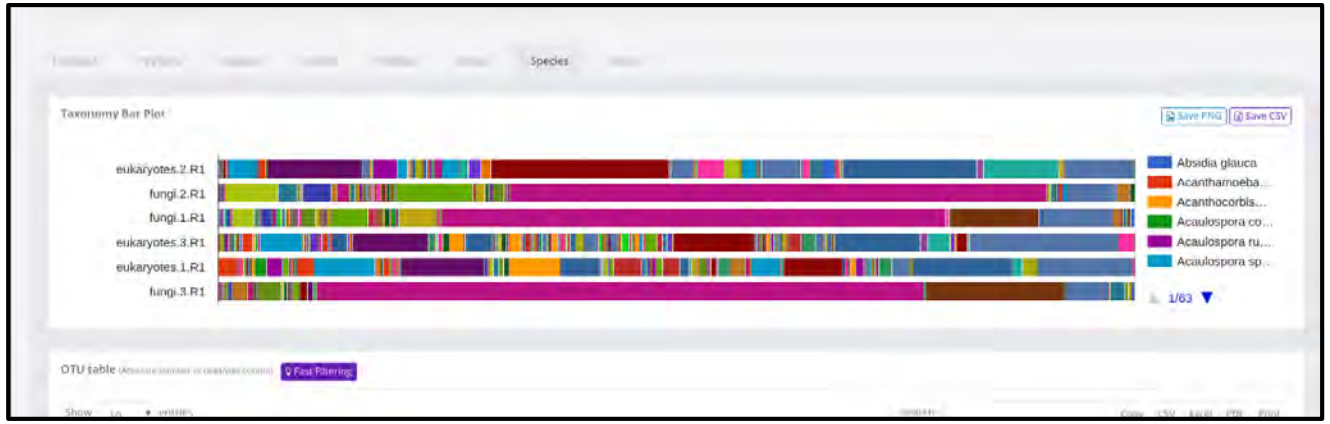
Show 10 entries  Search:  Copy CSV Excel PDF Print

Showing 1 to 6 of 6 entries

| Samples | Observed | Chao1 | Shannon | Simpson | Fisher |
|---|---|---|---|---|---|
| eukaryotes.1.R1 | 118 | 118 | 3.59561 | 0.95018 | 8.96918 |
| eukaryotes.2.R1 | 73 | 73 | 2.98961 | 0.91037 | 5.33176 |
| eukaryotes.3.R1 | 158 | 158 | 3.78232 | 0.94305 | 12.3204 |
| fungi.1.R1 | 87 | 87 | 2.07557 | 0.6774 | 6.43459 |
| fungi.2.R1 | 84 | 84 | 1.99539 | 0.64321 | 6.18383 |
| fungi.3.R1 | 51 | 51 | 1.41224 | 0.5353 | 3.61862 |
| Samples Search | Observed Search | Chao1 Search | Shannon Search | Simpson Search | Fisher Search |

Showing 1 to 6 of 6 entries  Previous 1 Next

## Taxonomy



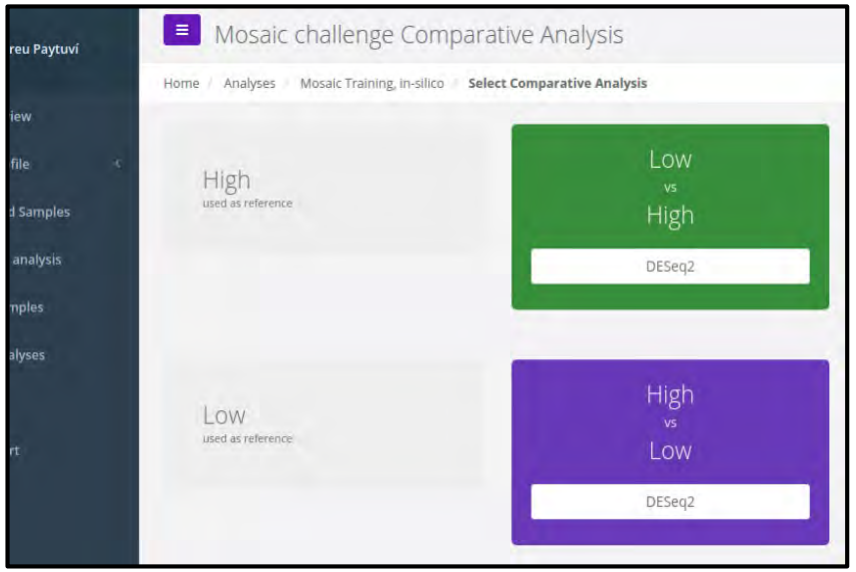

## Beta-diversities

Beta-diversity Bray Curtis measure  Download Excel

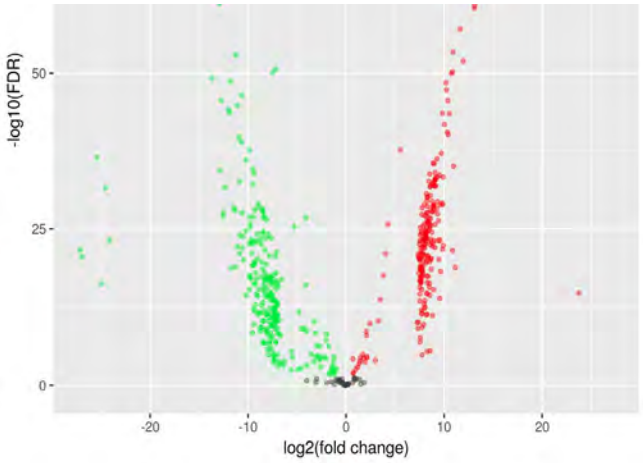
| Samples | eukaryotes.2.R1 | fungi.2.R1 | fungi.1.R1 | eukaryotes.3.R1 | eukaryotes.1.R1 | fungi.3.R1 |
|---|---|---|---|---|---|---|
| eukaryotes.2.R1 | 0 | 0.948 | 0.925 | 0.528 | 0.523 | 0.953 |
| fungi.2.R1 | 0.948 | 0 | 0.267 | 0.948 | 0.948 | 0.334 |
| fungi.1.R1 | 0.925 | 0.267 | 0 | 0.923 | 0.923 | 0.261 |
| eukaryotes.3.R1 | 0.528 | 0.948 | 0.923 | 0 | 0.428 | 0.953 |
| eukaryotes.1.R1 | 0.523 | 0.948 | 0.923 | 0.428 | 0 | 0.953 |
| fungi.3.R1 | 0.953 | 0.334 | 0.261 | 0.953 | 0.953 | 0 |

# GAIA

SEQUENTIA
DECODE • DISCOVER • DISRUPT

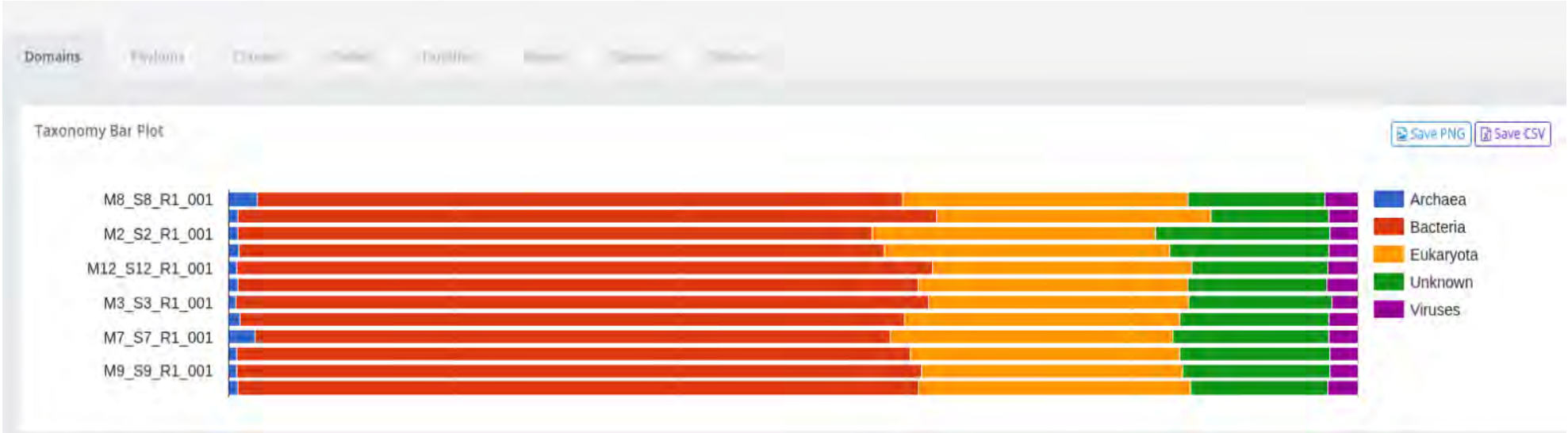







| OTUs | FDR | P-value | logFC (Low) |
|---|---|---|---|
| Ablotrophia defectiva | 1.31456421060578e-74 | 8.43846895017686e-75 | 15.0945637502851 |
| Acinetobacter johnsonii | 0.019325020944706 | 0.0186499110427075 | 0.82490391369906 |
| Acinetobacter radioresistens | 5.03462724408229e-65 | 3.53962876112336e-65 | 15.0953432460206 |
| Actinomyces odontolyticus | 4.73288969414525e-70 | 3.18281665021121e-70 | 14.9689216193574 |
| Bifidobacterium longum | 0.000396611873521357 | 0.000374096788998311 | 5.04567188957837 |
| Escherichia fergusonii | 2.99680147297891e-81 | 1.74050041880435e-81 | 15.4810346246184 |
| Megamonas hypermegale | 0.0355419504827427 | 0.0344555153151479 | 1.30031883233197 |
| Parabacteroides johnsonii | 2.27290520507079e-73 | 1.48880253607257e-73 | 14.7847444643296 |

# GAIA




Microorganisms are not only bacteria or archaea… we want to see fungi, algae and other eukaryotes!

# FIN DE LA PRIMERA PARTE

# SECOND PART: RNA-Seq and AIR

# Index

SEQUENTIA
DECODE • DISCOVER • DISRUPT

1. First part: omic data, metagenomics, metatranscriptomics and GAIA
    1. Omic data
    2. A little bit of history
    3. Applications
    4. Strategies: amplicon-based and shotgun
    5. Bioinformatic approaches and limitations
    6. GAIA
    7. Future insights

1. Second part: RNA -seq and AIR
    1. RNA-seq introduction
    2. Workflow
    3. Differential expression analysis
    4. AIR

# RNA -seq

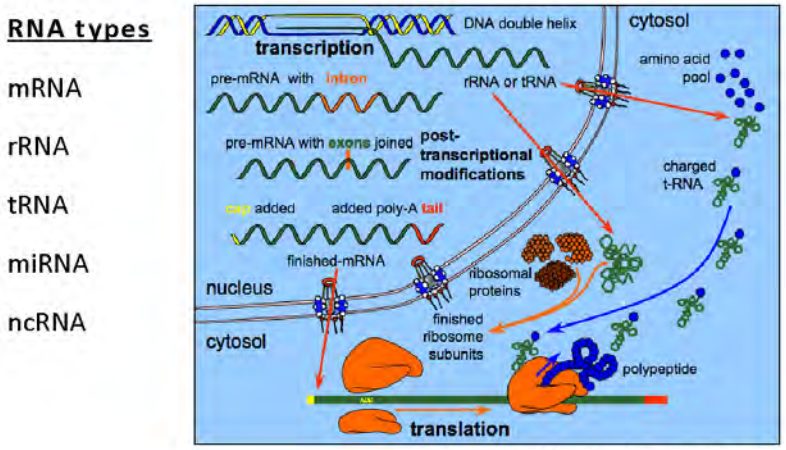
When the NGS technology is applied to RNA molecules we talk about

RNA-seq

Respect to genomic DNA sequencing, RNA-seq requires specific steps:

- Conversion of RNA to cDNA
- Selection of polyA transcripts or removal of rRNA
- Strand-specific libraries

RNA Transcription and Processing



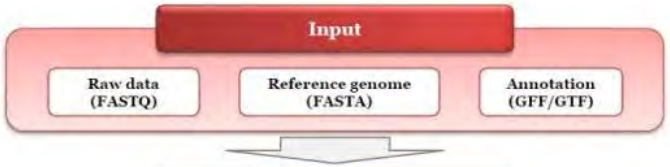

Koning, Plant Physiology Information Website

# mRNA -seq: applications

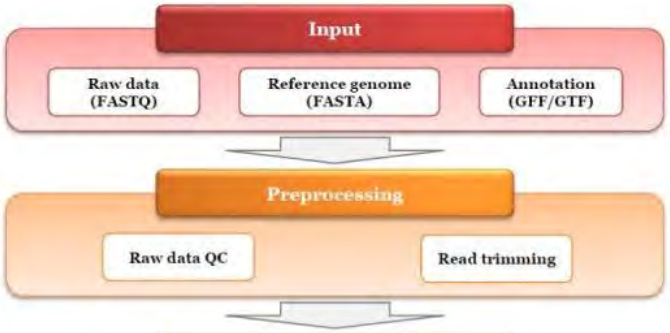
SEQUENTIA
DECODE • DISCOVER • DISRUPT

- Differential gene expression analysis
  - Healthy vs. diseased
  - Time course experiments
  - Different genotypes
- Transcriptional profiling
  - Tissue-specific expression
- Novel gene identification/transcriptome assembly
- Identification of splice variants
- SNP finding
- RNA editing

# RNA -seq: workflow



# RNA -seq: workflow

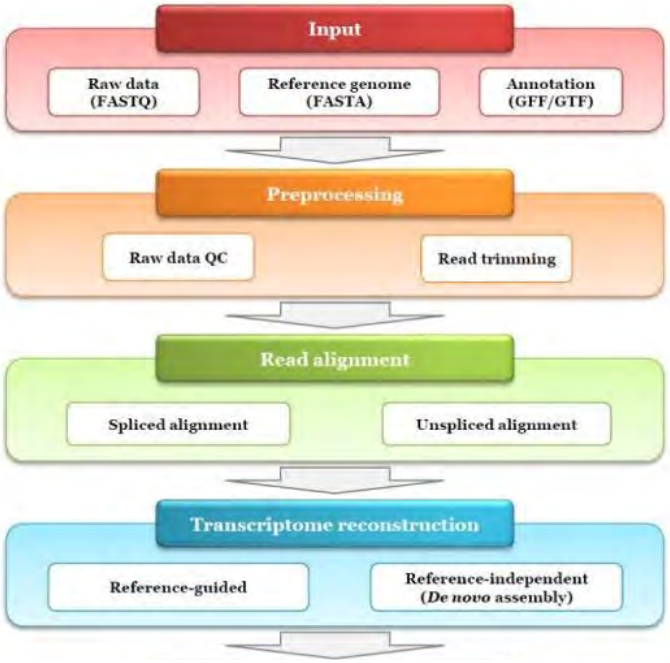


# Question 9

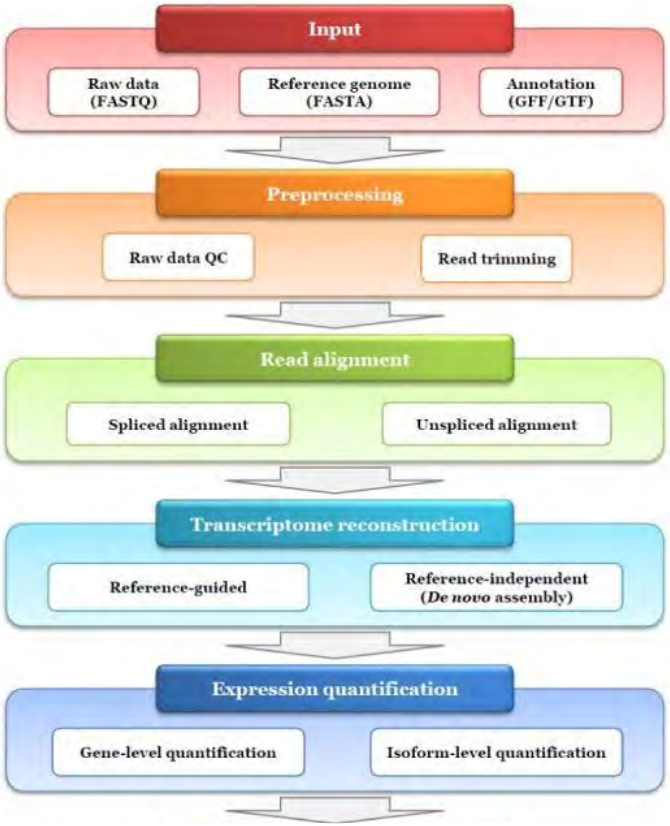
SEQUENTIA
DECODE • DISCOVER • DISRUPT

# RNA -seq: workflow

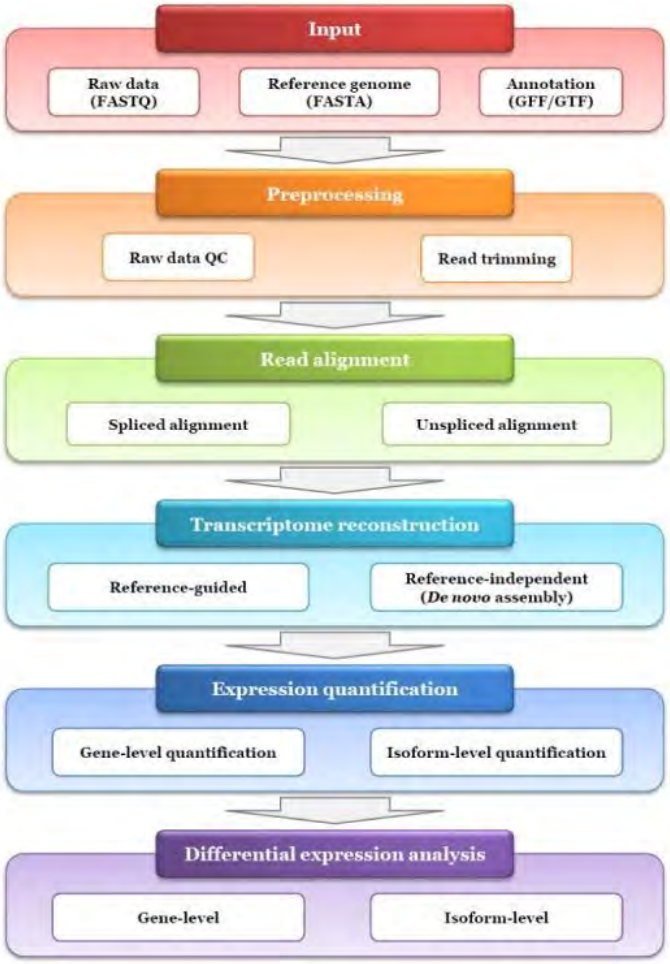
# RNA -seq: workflow



# RNA -seq: workflow



# RNA -seq: workflow

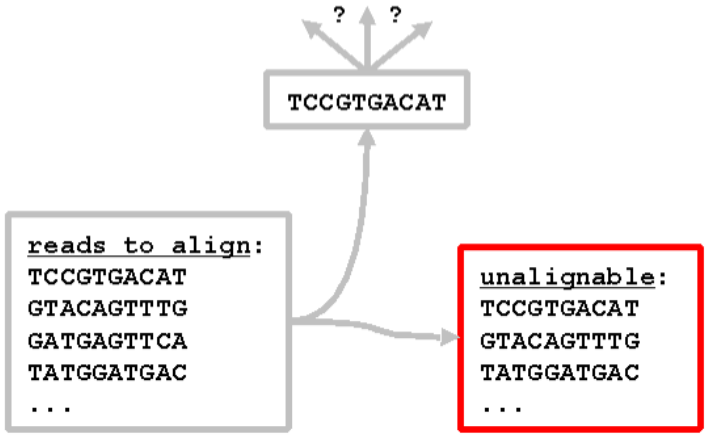


# RNA -seq: alignment and assembly

SEQUENTIA
DECODE • DISCOVER • DISRUPT

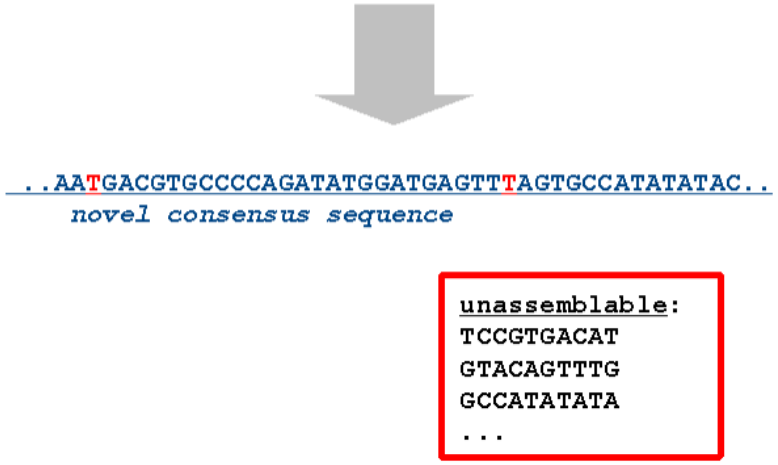
## Alignment

```
reference
..AA*GACGTGCCCCAGATATGGATGAGTTCAGTGCCATATATAC..
    TGACGTGCCC    tatggatgag       CCATATATAC
     gacgtgcccc  ATATGGATGA TTCAGTGCCA    TAC..
  AATGACGTGC   AGATATGGAT    ttaggccct
      ACGTGCCCCA       atgagtttag GCCATA*ATA
        GTGC*CCAGA
     GACGTGCCCC                        reads
        GTGCCCCACA
```




## Assembly

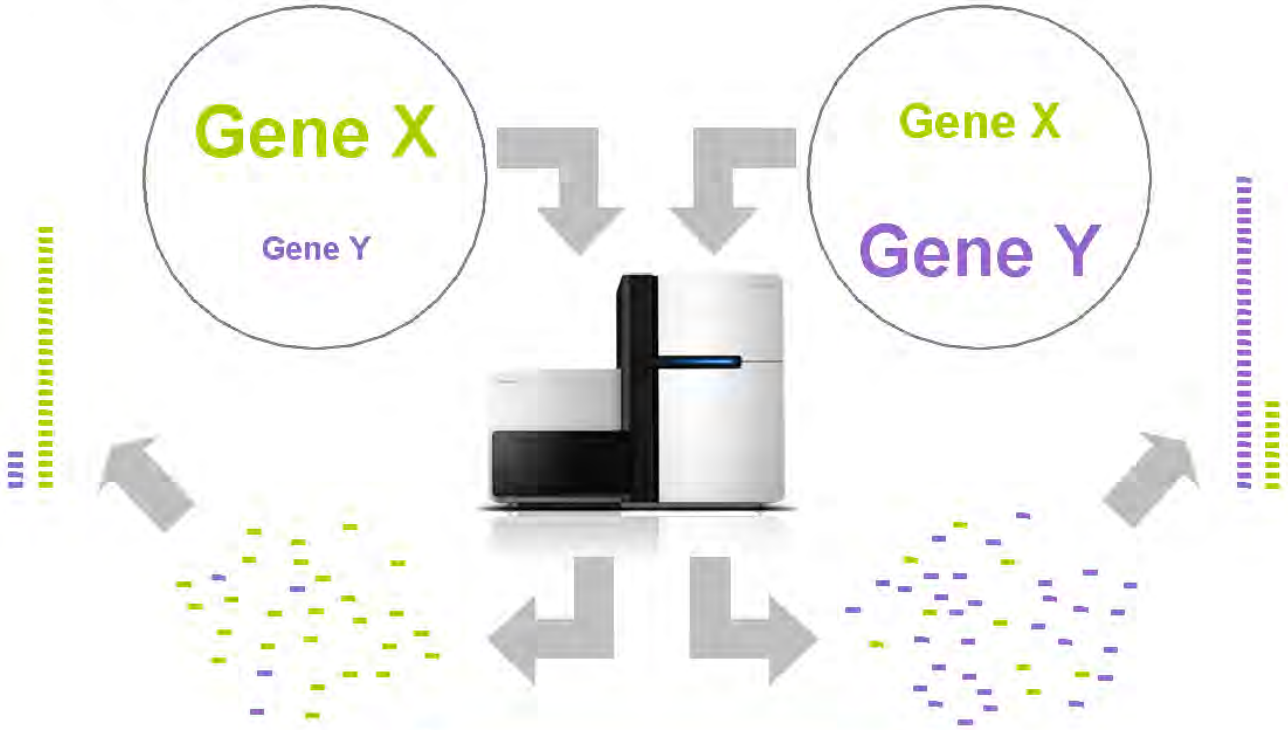
```
   TGACGTGCCC    tatggatgag       CCATATATAC
    gacgtgcccc  ATATGGATGA TTCAGTGCCA    TAC..
AATGACGTGC   AGATATGGAT    ttaggccct
    ACGTGCCCCA       atgagtttag GCCATA*ATA
      GTGC*CCAGA
   GACGTGCCCC                        reads
      GTGCCCCACA
```



# RNA -seq: alignment and assembly

SEQUENTIA
DECODE • DISCOVER • DISRUPT

## Why Do We Align?




To infer abundance.

# RNA -seq: gene counts

SEQUENTIA
DECODE • DISCOVER • DISRUPT

Given a file with aligned sequencing reads and a list of genomic features, a common task is to count how many reads map to each feature. A feature is here an interval (i.e., a range of positions) on a chromosome or a union of such intervals. In the case of RNA-Seq, the features are typically genes, where each gene is considered here as the union of all its exons.

| Locus | Sample A | Sample B | Sample C | Sample D |
|---|---|---|---|---|
| ENSMUSG00000090025 | 0 | 0 | 0 | 0 |
| ENSMUSG00000064842 | 0 | 0 | 0 | 0 |
| ENSMUSG00000051951 | 2637 | 3201 | 2180 | 364 |
| ENSMUSG00000089699 | 0 | 0 | 1 | 0 |
| ENSMUSG00000088390 | 0 | 0 | 0 | 0 |
| ENSMUSG00000089420 | 0 | 0 | 0 | 0 |
| ENSMUSG00000025900 | 1 | 1 | 0 | 0 |
| ENSMUSG00000025902 | 0 | 0 | 2 | 0 |
| ENSMUSG00000096126 | 0 | 0 | 0 | 0 |
| ENSMUSG00000098104 | 2 | 7 | 7 | 0 |
| ENSMUSG00000088000 | 0 | 0 | 0 | 0 |
| ENSMUSG00000033845 | 872 | 878 | 952 | 1875 |
| ENSMUSG00000025903 | 865 | 938 | 927 | 535 |
| ENSMUSG00000033813 | 2493 | 2669 | 2441 | 1561 |
| ENSMUSG00000062588 | 54 | 98 | 53 | 60 |
| ENSMUSG00000002459 | 144 | 174 | 152 | 70 |

# Question 10

SEQUENTIA
DECODE • DISCOVER • DISRUPT

# RNA -seq: gene count normalization

SEQUENTIA
DECODE • DISCOVER • DISRUPT

Raw read counts correspond to the number of reads associated to each gene in a given sample. In order to compare the expression levels in different samples or between different genes a normalization procedure is required. This is due to:

- Different number of reads in different samples (library size effect)
- Different length of transcripts
- Amplification bias
- %GC content

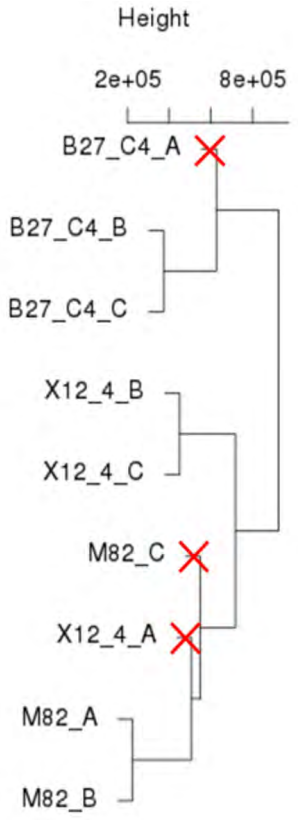
The most common normalization methods are:

- CPM (Counts Per Million)
- RPKM/FPKM (Reads/Fragment Per Kilobase Per Million)
- TPM (Transcript Per Million)
- TMM (Trimmed Mean Normalization of M-values)

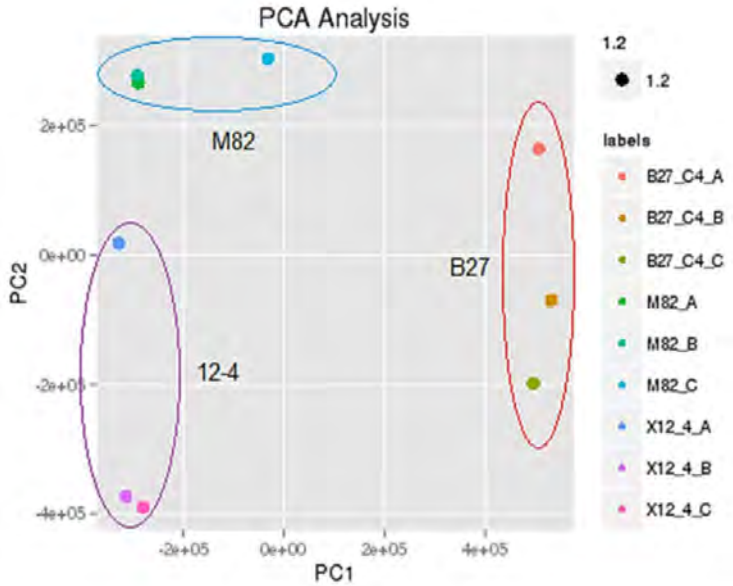
For differential expression analysis TMM was proved to be the most accurate method

# RNA -seq: evaluation of similarity

Once data has been normalized, a quality control of the experiment can be performed by evaluating the similarity/distance between replicates

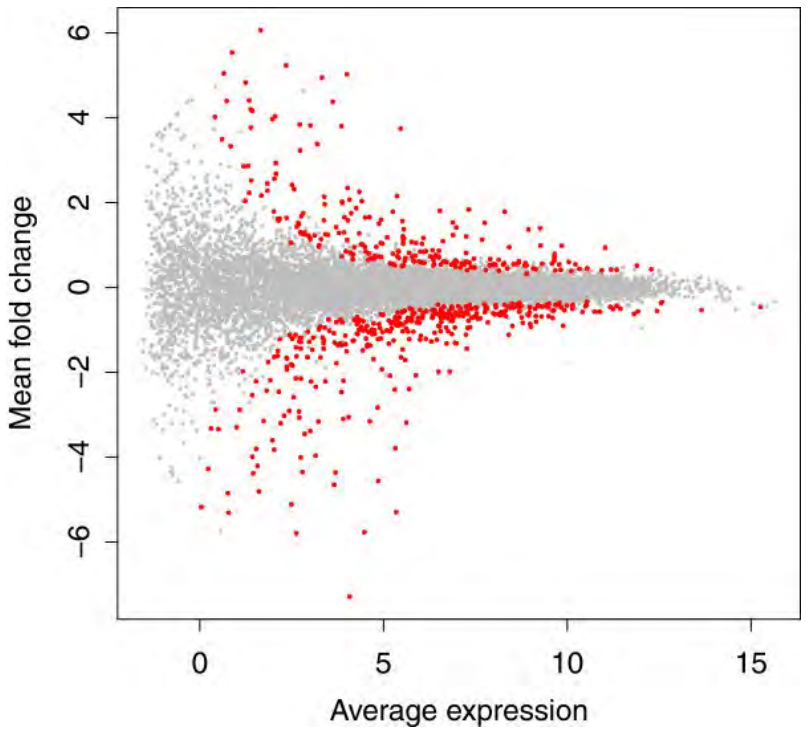


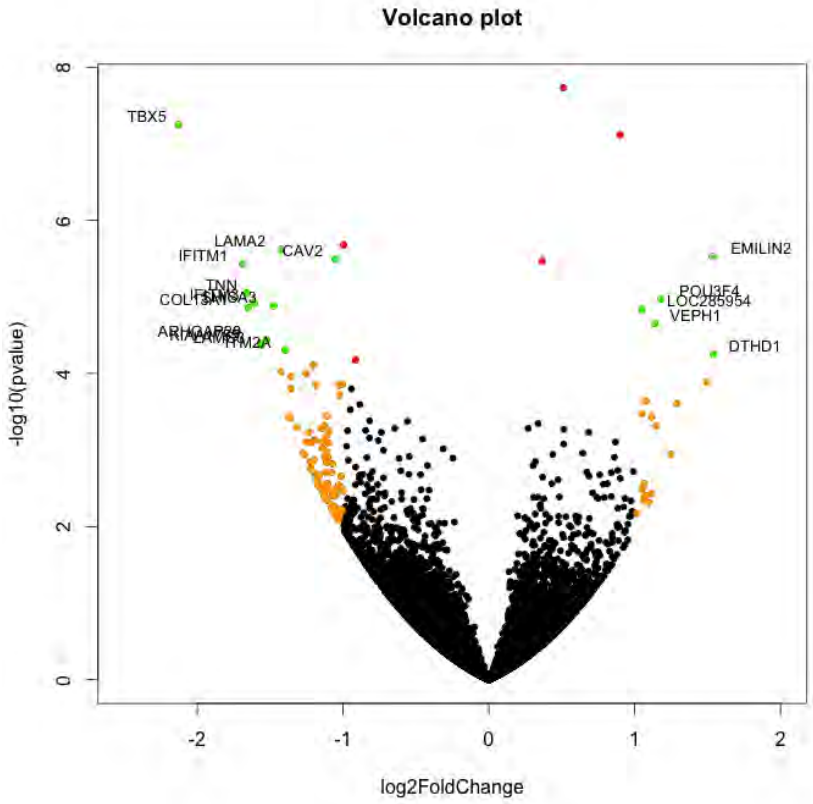

A hierarchical clusterization based on Euclidean distance or other specific measures of distance can be used to evaluate replicate homogeneity



# RNA -seq: differential expression

EdgeR and DESeq2 (parametric)
Bayesian (such as EBSeq) or permutation based (such as NOIseq)



# What else?

SEQUENTIA
DECODE • DISCOVER • DISRUPT

- An obvious way to gain biological insight is to assess the differentially expressed genes in terms of their known function(s)
- Required an automated and objective (statistical) approach
- Functional profiling or pathway analysis

GENEONTOLOGY
Unifying Biology

KEGG
Kyoto Encyclopedia of Genes and Genomes

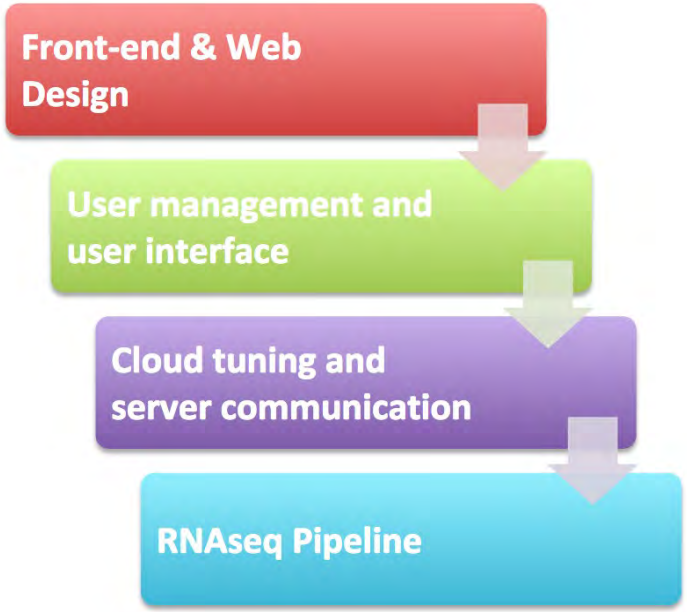
REACTOME
A CURATED PATHWAY DATABASE

# AIR

SEQUENTIA
DECODE • DISCOVER • DISRUPT

Artificial Intelligence RNA-seq (AIR)

We started a project to create a system that combining cutting-edge cloud technology together with the most used bioinformatics pipelines would be able to offer a completely user-friendly platform for RNA-seq analysis.




A coordinated effort of informatics engineers, bioinformaticians, web designers and web marketing experts.

Our informatics engineers developed a private cloud configuration system named “Orchestrator” which can use any cloud based computing platform (Google, Amazon, etc.) with an efficient resource management.

On top of this system we mounted our RNA-seq bioinformatics pipeline to perform differential gene expression analysis starting from raw Illumina reads.

# AIR

SEQUENTIA
DECODE • DISCOVER • DISRUPT

Artificial Intelligence RNA-seq (AIR)

http://www.transcriptomics.cloud

# THANK YOU

SEQUENTIA
DECODE • DISCOVER • DISRUPT

www.sequentiabiotech.com

info@sequentiabiotech.com

/SequentiaBiotech

/SequentiaBio