# Software and Pipelines for Taxonomic Assignment in Metagenomics

Gabriel Valiente

Algorithms, Bioinformatics, Complexity and Formal Methods Research Group
Technical University of Catalonia

Computational Biology and Bioinformatics Research Group
Research Institute of Health Science, University of the Balearic Islands

Centre for Genomic Regulation
Barcelona Biomedical Research Park

ELIXIR-IIB Workshop and School
Advanced Computational Metagenomics
Bari, Italy, 19–23 June 2017

# Abstract

The classification of next generation sequencing reads from a metagenomic sample using a reference taxonomy is usually based on first mapping the reads to the reference sequences and then, classifying each read at a node under the lowest common ancestor of the candidate sequences in the reference taxonomy with the least classification error. In this lecture, we will discuss potential biases of this approach, current software implementing taxonomic annotation, and their integration in some of the most widely used pipelines for metagenomic analysis

# Outline

- Taxonomic Assignment in Metagenomics
- Software for Taxonomic Assignment in Metagenomics
- Pipelines for Taxonomic Assignment in Metagenomics
- TANGO Workflow for Taxonomic Assignment

# Part I

# Taxonomic Assignment in Metagenomics

# QIIME Workflow for Microbial Community Analysis

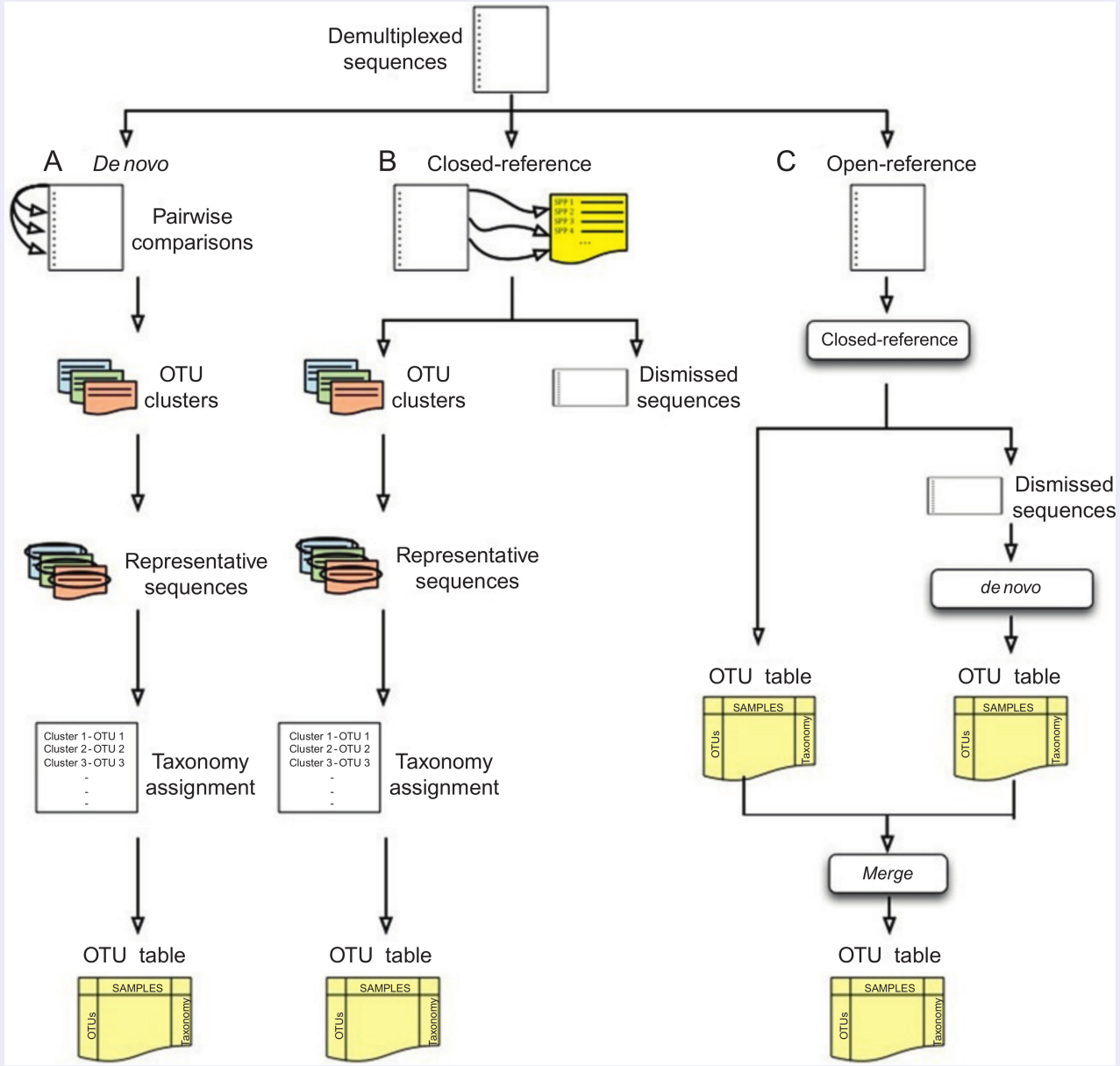
## Upstream analysis

- Demultiplexing and quality filtering
- OTU picking
- Chimeric sequences identification
- Taxonomy assignment
- Sequence alignment
- Phylogeny construction
- OTU table construction

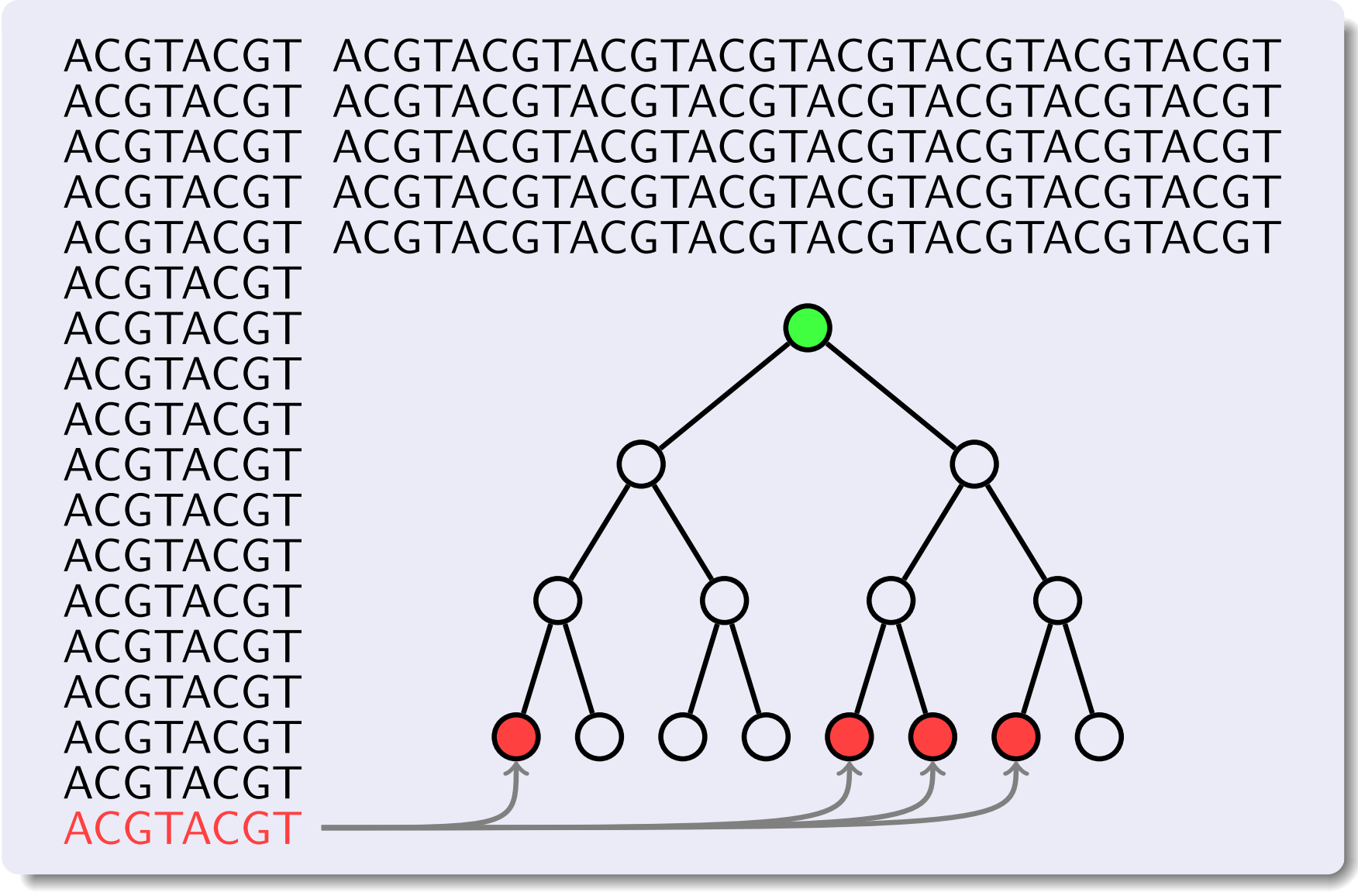
## Downstream analysis

- Taxa summaries
- Alpha-diversity analysis
- Beta-diversity analysis
- Statistical significance of differences in diversity

# QIIME Workflow for Microbial Community Analysis

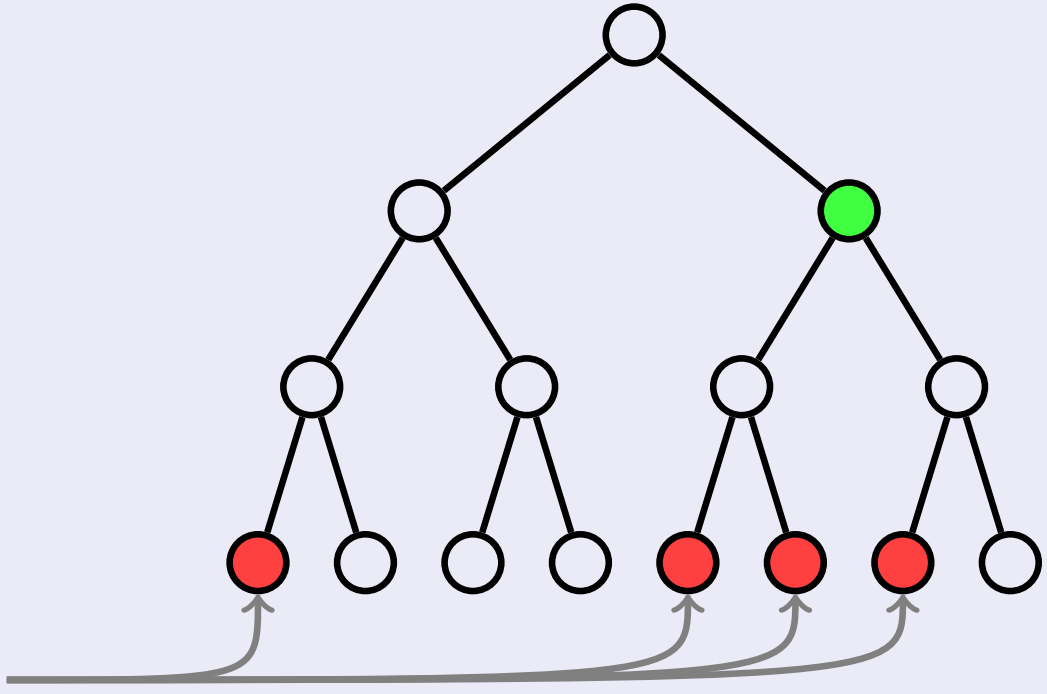


# MEGAN Workflow for Taxonomic Assignment



# TANGO Workflow for Taxonomic Assignment

ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT



# Part II

# Software for Taxonomic Assignment in Metagenomics

# MEGAN

- D. Huson and N. Weber. Microbial community analysis using MEGAN. In E. F. Delong, editor, *Methods in Enzymology*, volume 531, chapter 21, pages 465–485. Elsevier, 2013

# TANGO


- J. C. Clemente, J. Jansson, and G. Valiente. Flexible Taxonomic Assignment of Ambiguous Sequencing Reads. *BMC Bioinformatics*, 12:8, 2011
- D. Alonso, A. Barré, S. Beretta, P. Bonizzoni, M. Nikolski, and G. Valiente. Further Steps in TANGO: Improved Taxonomic Assignment in Metagenomics. *Bioinformatics*, 30(1):17–23, 2014
- B. Fosso, G. Pesolo, F. Rosselló, and G. Valiente. Unbiased Taxonomic Annotation of Metagenomic Samples. In Z. Cai, O. Daescu, and M. Li, editors, *Proc. 13th Int. Symp. Bioinformatics Research and Applications*, volume 10330 of *Lecture Notes in Bioinformatics*, pages 162–173. Springer, 2017

# Part III

# Pipelines for Taxonomic Assignment in Metagenomics

# MOTHUR

- P. D. Schloss and J. Handelsman. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.*, 71(3):1501–1506, 2005
- P. D. Schloss and J. Handelsman. Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. *Appl. Environ. Microbiol.*, 72(10):6773–6779, 2006
- P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. Introducing MOTHUR: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, 75(23):7537–7541, 2009

# QIIME

- Advancing our understanding of the human microbiome using QIIME. In E. F. Delong, editor, *Methods in Enzymology*, volume 531, chapter 19, pages 371–444. Elsevier, 2013

# RDP

- J. R. Cole, Q. Wang, B. Chai, and J. M. Tiedje. The Ribosomal Database Project: Sequences and Software for High-Throughput rRNA Analysis. In F. J. de Bruijn, editor, *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*, chapter 36, pages 313–324. Wiley, 2011

# MAP

- M. Huntemann, N. N. Ivanova, K. Mavromatis, H. J. Tripp, D. Paez-Espino, K. Tennessen, K. Palaniappan, E. Szeto, M. Pillay, I.-M. A. Chen, A. Pati, T. Nielsen, V. M. Markowitz, and N. C. Kyrpides. The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4). *Standards in Genomic Sciences*, 11:17, 2015

# MG-RAST

- K. P. Keegan, E. M. Glass, and F. Meyer. MG-RAST, a metagenomics service for analysis of microbial community structure and function. In F. Martin and S. Uroz, editors, *Microbial Environmental Genomics*, volume 1399 of *Methods in Molecular Biology*, chapter 13, pages 207–233. Springer, 2016

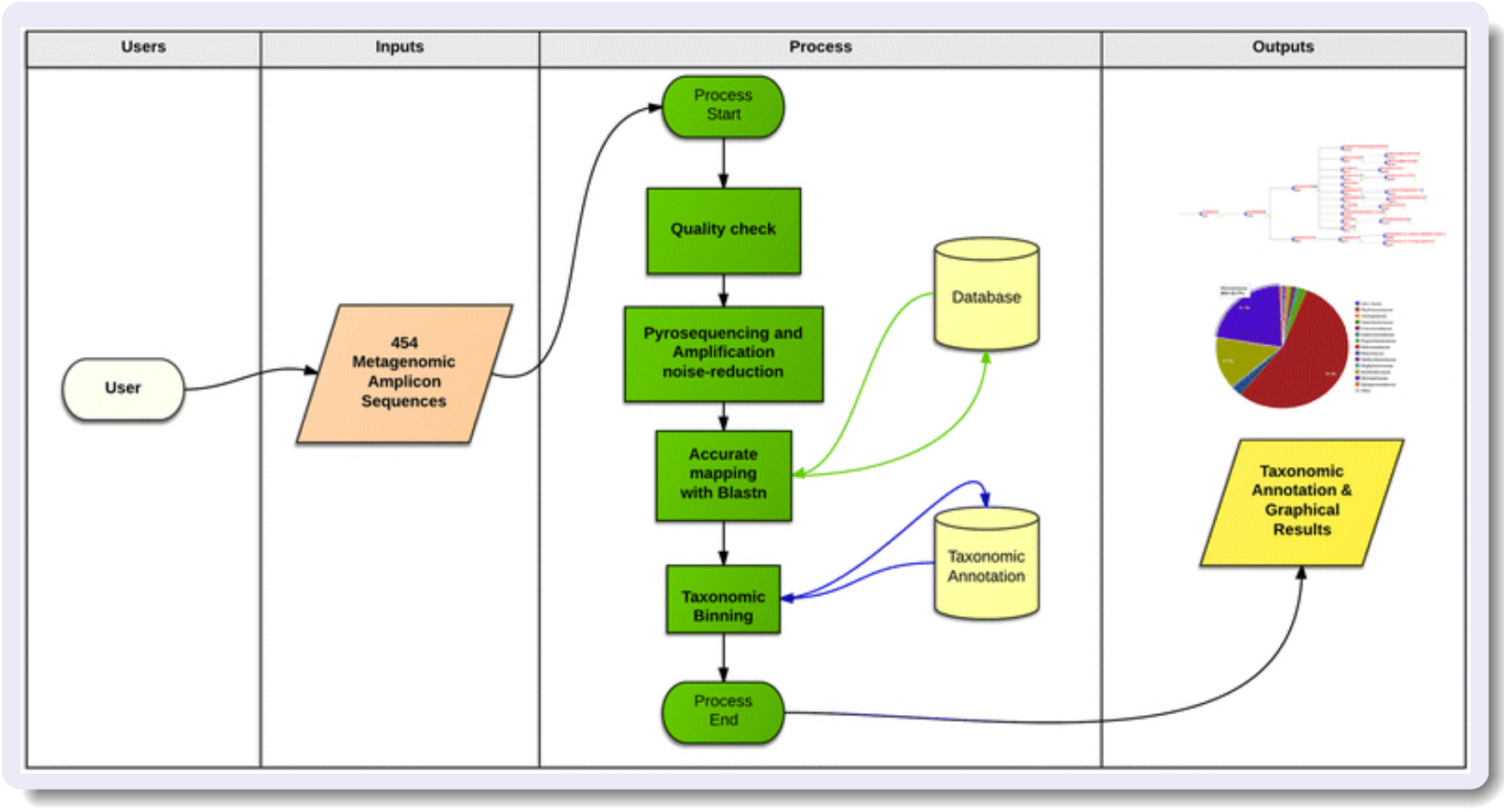
# EBI metagenomics

- A. Mitchell, F. Bucchini, G. Cochrane, H. Denise, P. ten Hoopen, M. Fraser, S. Pesseat, S. Potter, M. Scheremetjew, P. Sterk, and R. D. Finn. EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.*, 44(D1):D595–D603, 2016

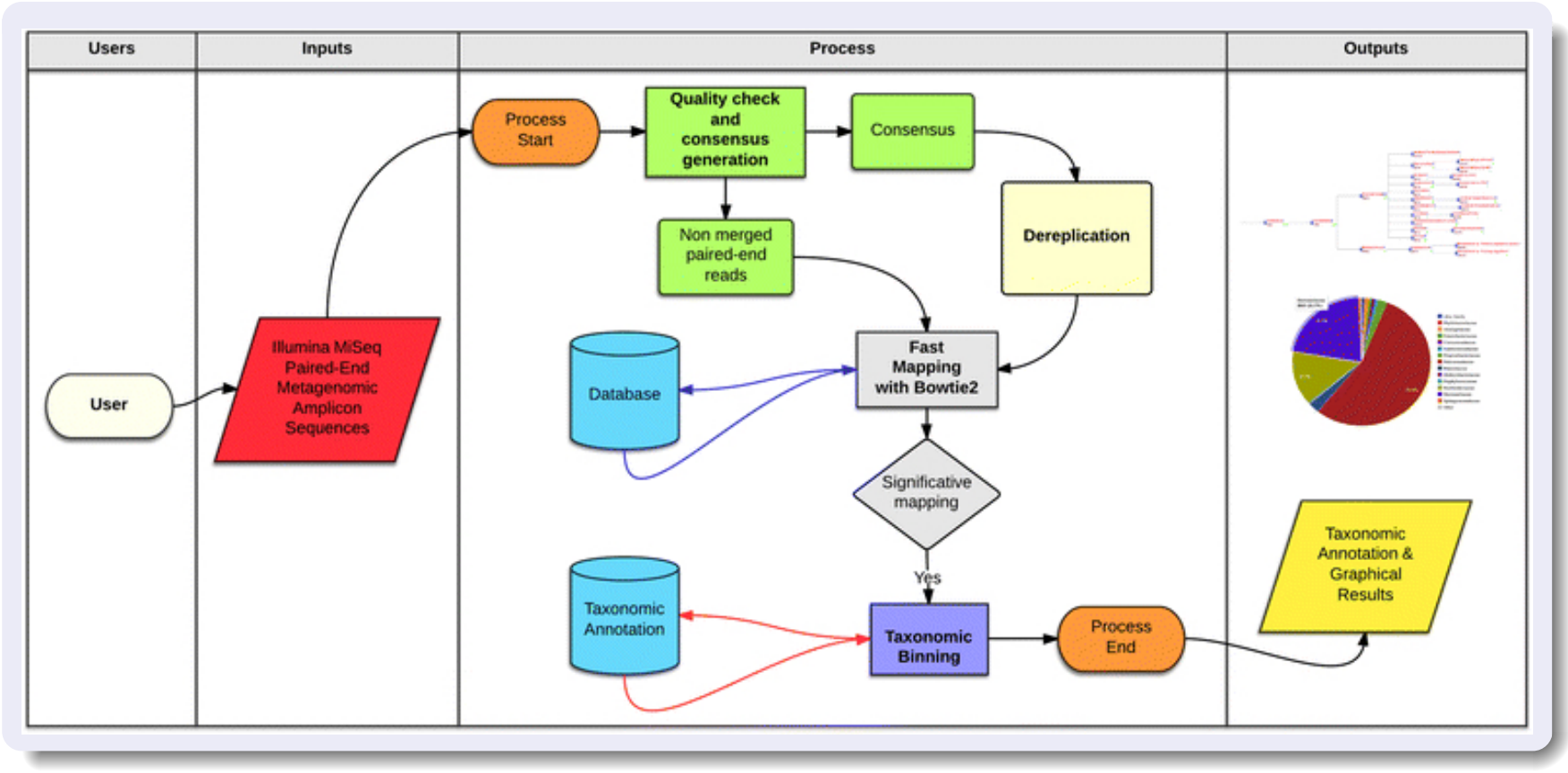
# BioMaS

- B. Fosso, M. Santamaria, M. Marzano, D. Alonso-Alemany, G. Valiente, G. Donvito, A. Monaco, P. Notarangelo, and G. Pesole. BioMaS: A Modular Pipeline for Bioinformatic Analysis of Metagenomic Amplicons. *BMC Bioinformatics*, 16:203, 2015

# BioMaS


Users
Inputs
Process
Outputs
User
454 Metagenomic Amplicon Sequences
Process Start
Quality check
Pyrosequencing and Amplification noise-reduction
Accurate mapping with Blastn
Taxonomic Binning
Process End
Database
Taxonomic Annotation
Taxonomic Annotation & Graphical Results

# BioMaS


Users
Inputs
Process
Outputs
User
Illumina MiSeq Paired-End Metagenomic Amplicon Sequences
Process Start
Quality check and consensus generation
Consensus
Dereplication
Non merged paired-end reads
Database
Fast Mapping with Bowtie2
Significative mapping
Yes
Taxonomic Annotation
Taxonomic Binning
Process End
Taxonomic Annotation & Graphical Results

# MetaShot

- B. Fosso, M. Santamaria, M. D'Antonio, D. Lovero, G. Corrado, E. Vizza, N. Passaro, A. R. Garbuglia, M. R. Capobianchi, M. Crescenzi, G. Valiente, and G. Pesole. MetaShot: A User-Friendly Workflow for Taxon Classification of Host-Associated Microbiome from Shotgun Metagenomic Data. *Bioinformatics*, 33(11):1730–1732, 2017

# Part IV

# TANGO Workflow for Taxonomic Assignment

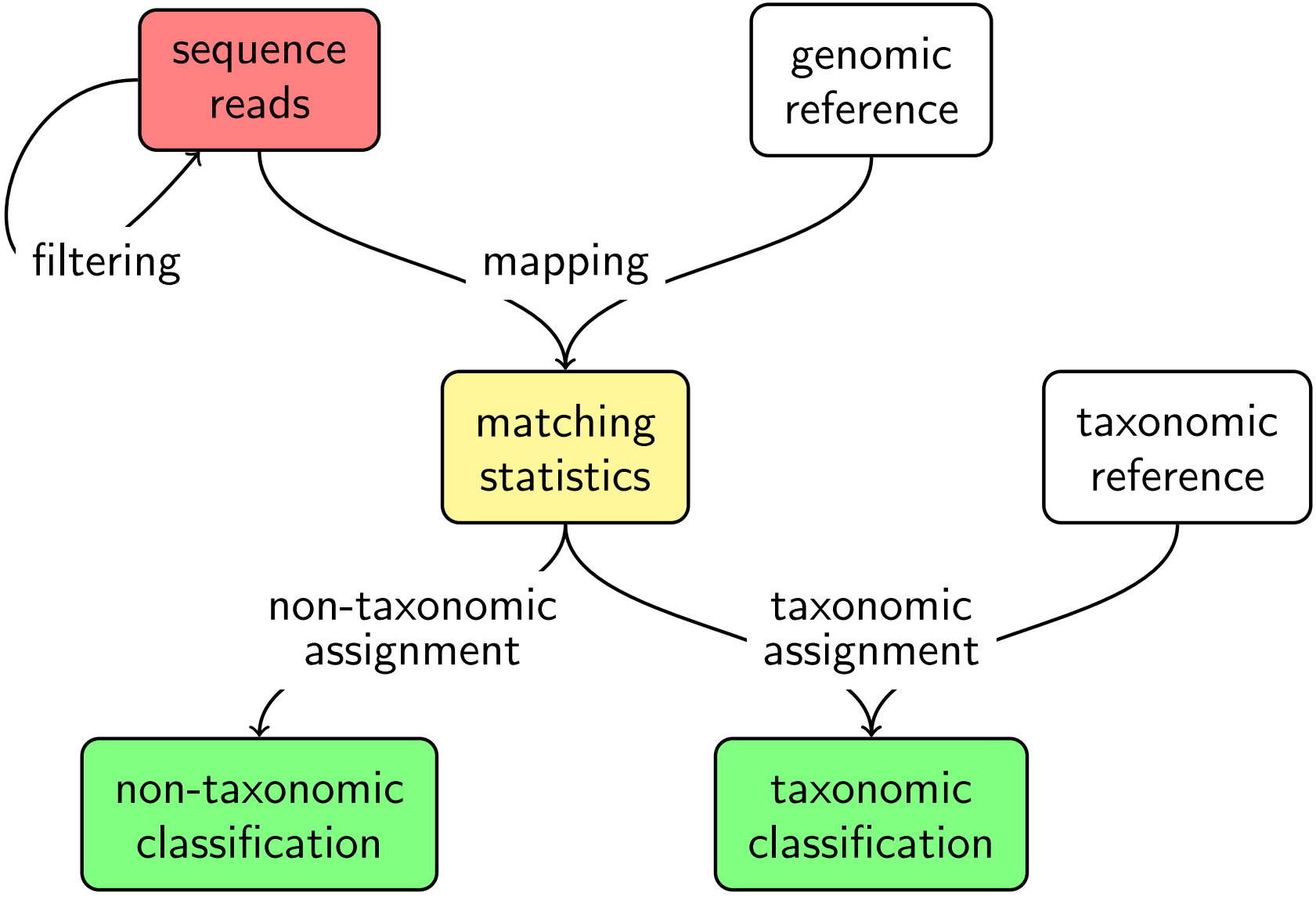
# Motivation

## Classification of Reads from a Metagenomic Sample Using a Reference Taxonomy

- Mapping the reads to the reference sequences
- Classifying each read at a node under the LCA of the candidate sequences in the reference taxonomy with the least classification error

## Potential Sources of Bias

- Multiple nodes in the reference taxonomy with the least classification error for a given read


sequence reads
genomic reference
filtering
mapping
matching statistics
taxonomic reference
non-taxonomic assignment
taxonomic assignment
non-taxonomic classification
taxonomic classification

## Example

| | $R_1$ | $R_2$ | $R_3$ | $R_4$ | $R_5$ | $R_6$ | $R_7$ | $R_8$ | $R_9$ | $R_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $R_1$ | x | x | | | | | | x | | |
| $R_2$ | x | x | | | | | | | x | |
| $R_3$ | | | x | x | | | | | | |
| $R_4$ | | | x | x | | | | | | |
| $R_5$ | | | | | x | x | x | | | x |
| $R_6$ | | | | | x | x | x | | | |
| $R_7$ | | | | | x | x | x | | | |
| $R_8$ | x | | | | | | | x | | |
| $R_9$ | | x | | | | | | | x | |
| $R_{10}$ | | | | | x | | | | | x |
| | | | | | | | | | | |

- Reads $R_1, \ldots, R_{10}$ match reads $R_1, \ldots, R_{10}$

## Example

| | $R_1$ | $R_2$ | $R_3$ | $R_4$ | $R_5$ | $R_6$ | $R_7$ | $R_8$ | $R_9$ | $R_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $R_1$ | x | x | | | | | | x | | |
| $R_2$ | x | x | | | | | | | x | |
| $R_3$ | | | x | x | | | | | | |
| $R_4$ | | | x | x | | | | | | |
| $R_5$ | | | | | x | x | x | | | x |
| $R_6$ | | | | | x | x | x | | | |
| $R_7$ | | | | | x | x | x | | | |
| $R_8$ | x | | | | | | | x | | |
| $R_9$ | | x | | | | | | | x | |
| $R_{10}$ | | | | | x | | | | | x |
| | A | A | B | B | C | C | C | A | A | C |

- Reads $R_1, \ldots, R_{10}$ match reads $R_1, \ldots, R_{10}$

## Example

| | $S_1$ | $S_2$ | $S_3$ | $S_4$ |
|---|---|---|---|---|
| $R_1$ | x | | | |
| $R_2$ | | x | | |
| $R_3$ | | | x | |
| $R_4$ | | | | x |
| $R_5$ | x | x | | |
| $R_6$ | | x | x | |
| $R_7$ | | | x | x |
| $R_8$ | x | x | x | |
| $R_9$ | | x | x | x |
| $R_{10}$ | x | x | x | x |
| | | | | |

- Reads $R_1, \ldots, R_{10}$ match sequences $S_1, \ldots, S_4$

## Example (Statistical mapping)

| | $S_1$ | $S_2$ | $S_3$ | $S_4$ |
|---|---|---|---|---|
| $R_1$ | 1 | | | |
| $R_2$ | | 1 | | |
| $R_3$ | | | 1 | |
| $R_4$ | | | | 1 |
| $R_5$ | $1/2$ | $1/2$ | | |
| $R_6$ | | $1/2$ | $1/2$ | |
| $R_7$ | | | $1/2$ | $1/2$ |
| $R_8$ | $1/3$ | $1/3$ | $1/3$ | |
| $R_9$ | | $1/3$ | $1/3$ | $1/3$ |
| $R_{10}$ | $1/4$ | $1/4$ | $1/4$ | $1/4$ |
| | | | | |

- Reads $R_1, \ldots, R_{10}$ match sequences $S_1, \ldots, S_4$

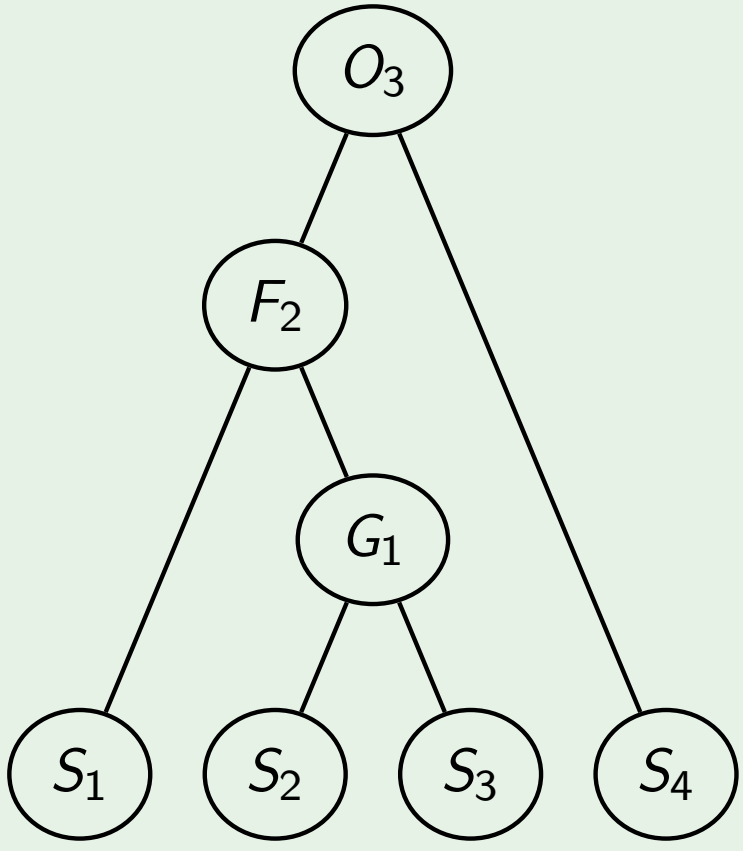
## Example (Statistical mapping)

| | $S_1$ | $S_2$ | $S_3$ | $S_4$ |
|---|---|---|---|---|
| $R_1$ | 1 | | | |
| $R_2$ | | 1 | | |
| $R_3$ | | | 1 | |
| $R_4$ | | | | 1 |
| $R_5$ | 1/2 | 1/2 | | |
| $R_6$ | | 1/2 | 1/2 | |
| $R_7$ | | | 1/2 | 1/2 |
| $R_8$ | 1/3 | 1/3 | 1/3 | |
| $R_9$ | | 1/3 | 1/3 | 1/3 |
| $R_{10}$ | 1/4 | 1/4 | 1/4 | 1/4 |
| | 21% | 29% | 29% | 21% |

- Reads $R_1, \ldots, R_{10}$ match sequences $S_1, \ldots, S_4$

## Example

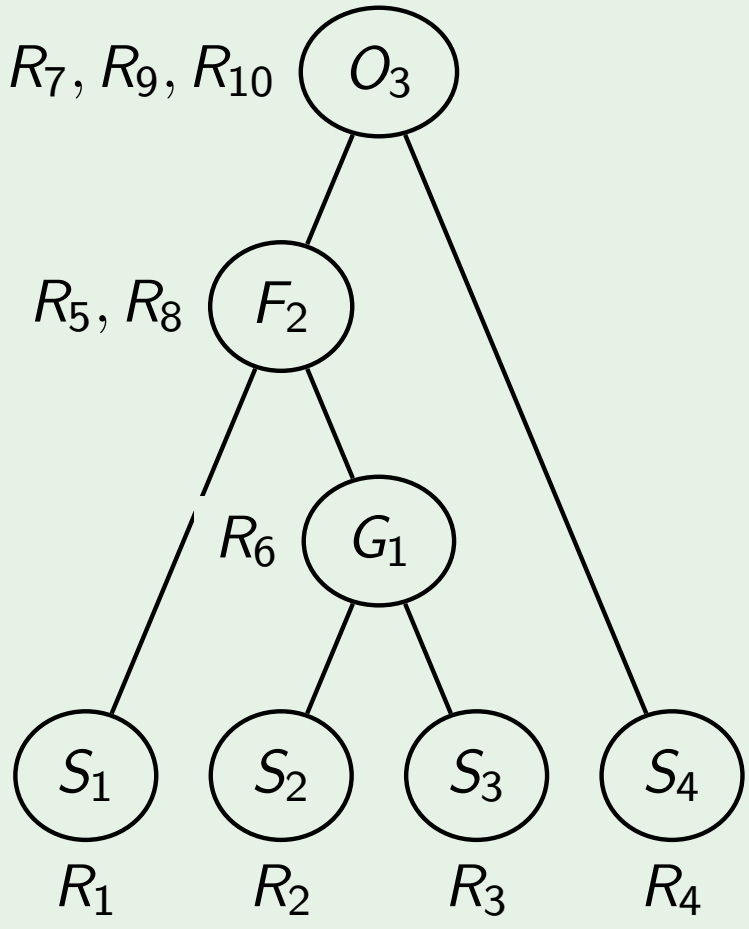
| | $S_1$ | $S_2$ | $S_3$ | $S_4$ |
|---|---|---|---|---|
| $R_1$ | x | | | |
| $R_2$ | | x | | |
| $R_3$ | | | x | |
| $R_4$ | | | | x |
| $R_5$ | x | x | | |
| $R_6$ | | x | x | |
| $R_7$ | | | x | x |
| $R_8$ | x | x | x | |
| $R_9$ | | x | x | x |
| $R_{10}$ | x | x | x | x |
| | | | | |




- Reads $R_1, \ldots, R_{10}$ match sequences $S_1, \ldots, S_4$

## Example (LCA mapping)

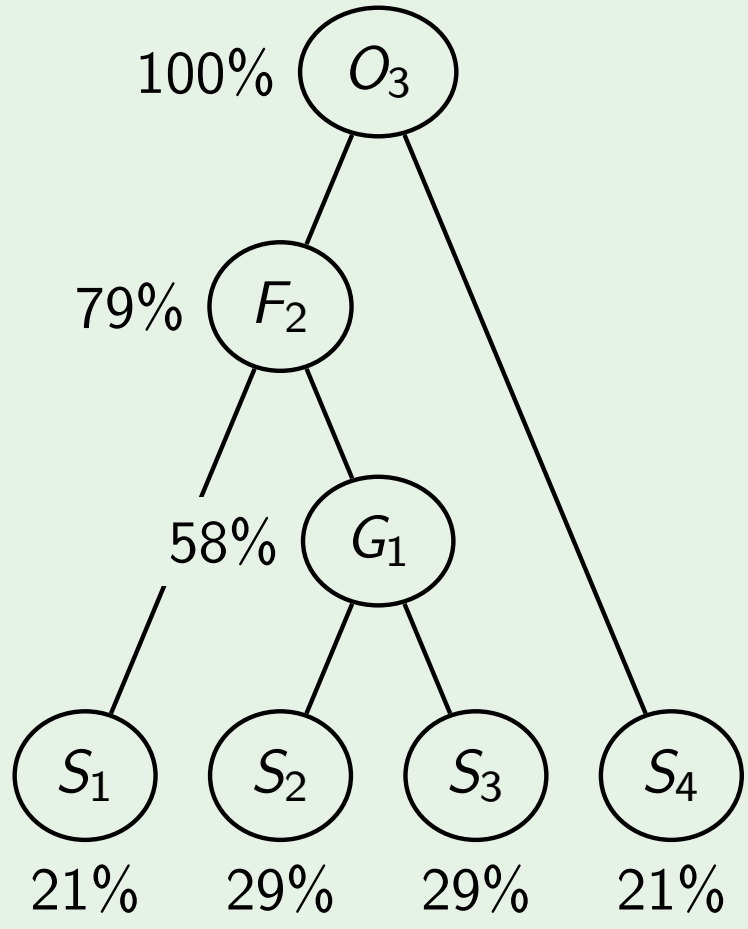
| | $S_1$ | $S_2$ | $S_3$ | $S_4$ |
|---|---|---|---|---|
| $R_1$ | x | | | |
| $R_2$ | | x | | |
| $R_3$ | | | x | |
| $R_4$ | | | | x |
| $R_5$ | x | x | | |
| $R_6$ | | x | x | |
| $R_7$ | | | x | x |
| $R_8$ | x | x | x | |
| $R_9$ | | x | x | x |
| $R_{10}$ | x | x | x | x |
| | | | | |




- Reads $R_1, \ldots, R_{10}$ match sequences $S_1, \ldots, S_4$

## Example (Statistical mapping)

| | $S_1$ | $S_2$ | $S_3$ | $S_4$ |
|---|---|---|---|---|
| $R_1$ | 1 | | | |
| $R_2$ | | 1 | | |
| $R_3$ | | | 1 | |
| $R_4$ | | | | 1 |
| $R_5$ | 1/2 | 1/2 | | |
| $R_6$ | | 1/2 | 1/2 | |
| $R_7$ | | | 1/2 | 1/2 |
| $R_8$ | 1/3 | 1/3 | 1/3 | |
| $R_9$ | | 1/3 | 1/3 | 1/3 |
| $R_{10}$ | 1/4 | 1/4 | 1/4 | 1/4 |
| | 21% | 29% | 29% | 21% |




- Reads $R_1, \ldots, R_{10}$ match sequences $S_1, \ldots, S_4$

## Assumption

The reads in a metagenomic sample to be classified come from known sequences in a reference taxonomy

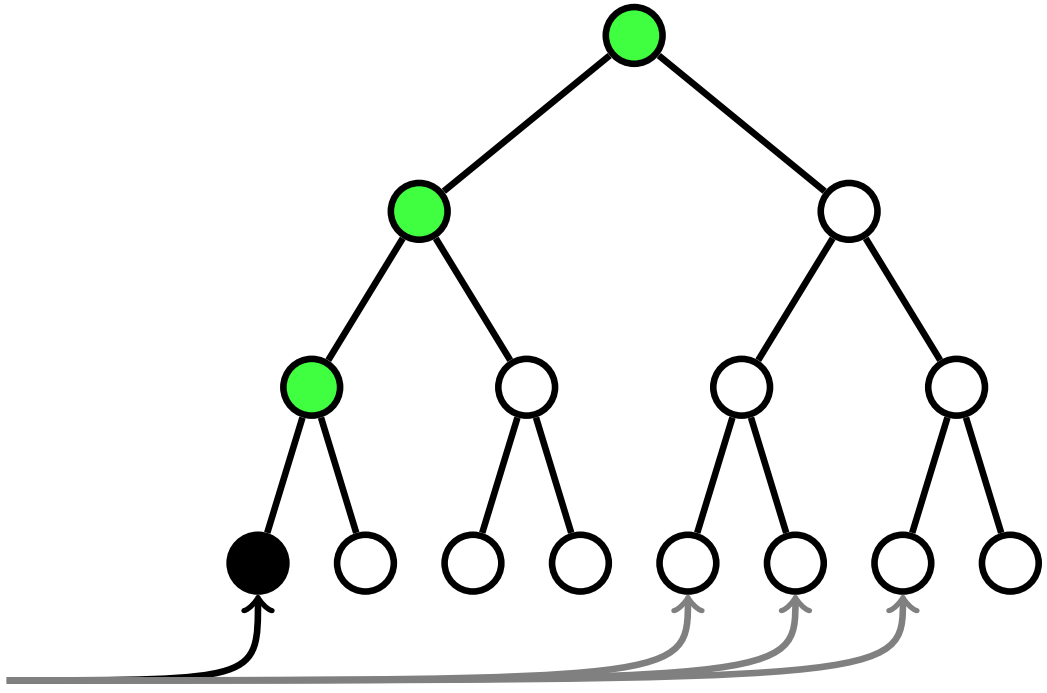
## Definition

The taxonomic annotation of the read at a certain node in the clade of the LCA in the reference taxonomy of the set of candidate sequences is correct if the candidate sequence that the read comes from lies in the clade of the node at which it is annotated

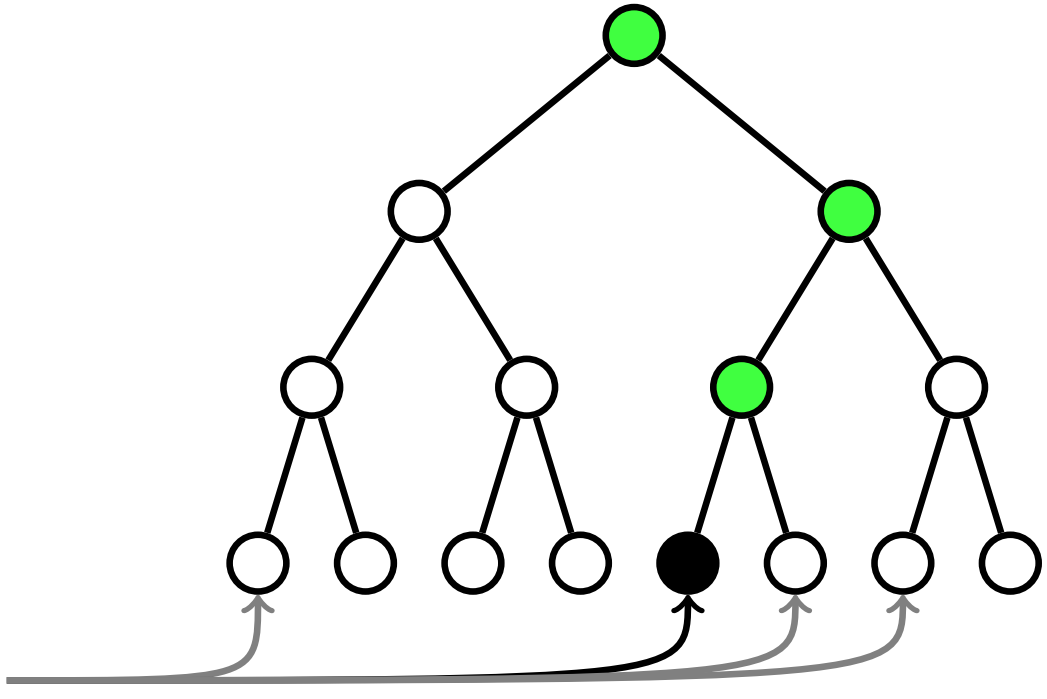
## Question

What is the best indicator of classification error for the taxonomic annotation of metagenomic samples?

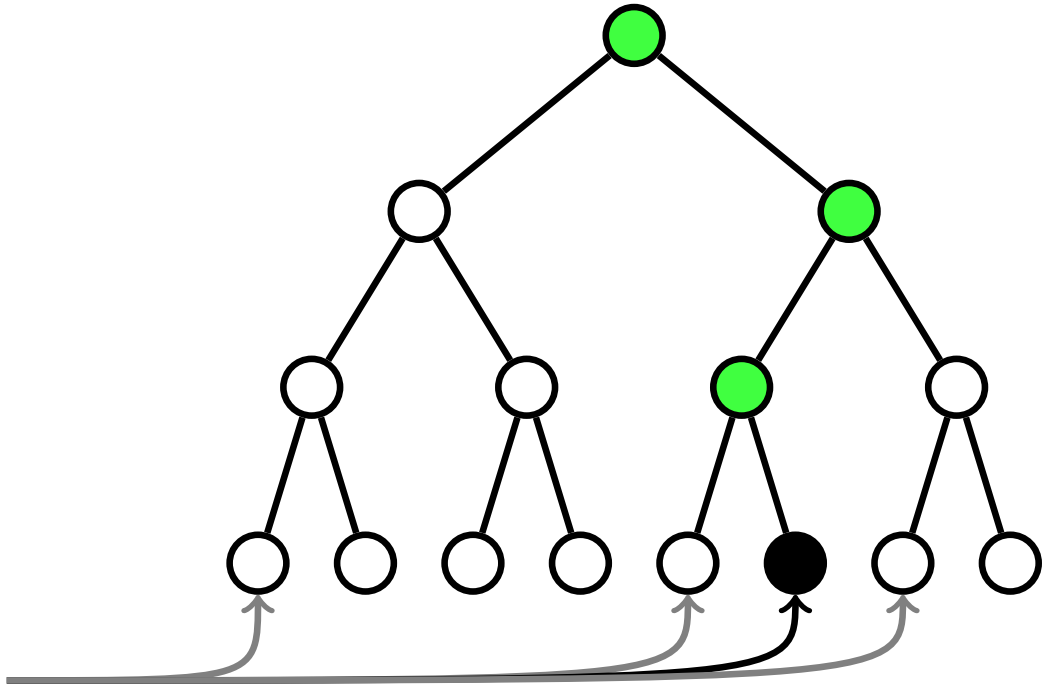
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT



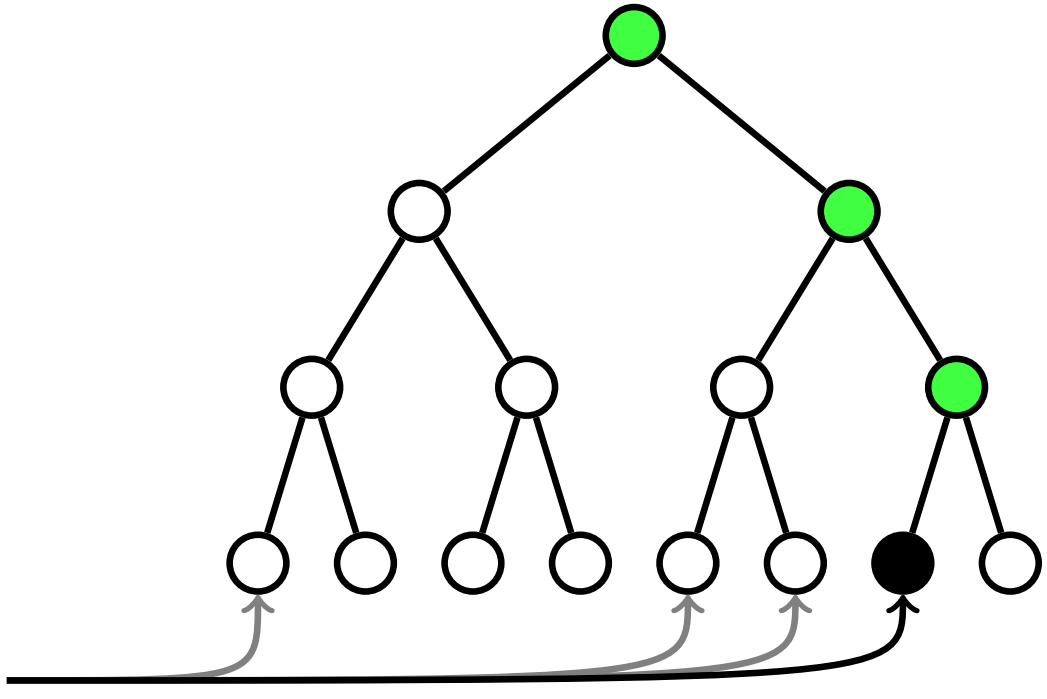
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT



ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT



ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT



## Indicators of classification error

- Yule $\phi$ (Matthews correlation coefficient)
- Youden $J$
- Area under the ROC curve
- $F$-measure
- Jaccard similarity coefficient
- Rand index

## Method

Compute the value of all these indicators of classification error for each possible set of candidate sequences in a reference taxonomy and for each possible candidate node for the taxonomic annotation of a read coming from each of the candidate sequences

## Total number of correct taxonomic annotations

| Complete binary tree | | | | | | | |
|---|---|---|---|---|---|---|---|
| $n$ | 4 | 6 | 8 | 10 | 12 | 14 | 16 |
| Yule $\phi$ | 40 | 262 | 824 | 4,318 | 17,064 | 63,378 | 270,448 |
| AUC | 40 | 262 | 920 | 4,726 | 22,056 | 79,322 | 352,496 |
| Jaccard | 32 | 220 | 984 | 5,188 | 24,844 | 112,812 | 493,856 |
| Rand | 48 | 344 | 1,544 | 8,308 | 37,764 | 154,012 | 672,416 |
| Rooted caterpillar | | | | | | | |
| $n$ | 4 | 6 | 8 | 10 | 12 | 14 | 16 |
| Yule $\phi$ | 38 | 203 | 945 | 4,344 | 20,152 | 88,063 | 398,700 |
| AUC | 38 | 211 | 973 | 4,628 | 22,230 | 94,962 | 421,697 |
| Jaccard | 32 | 195 | 1,024 | 5,104 | 24,491 | 113,305 | 518,937 |
| Rand | 36 | 222 | 1,191 | 5,949 | 28,459 | 132,263 | 602,076 |

**Remark**

The taxonomic annotation of a metagenomic sample involves obtaining the candidate nodes in the LCA skeleton tree of a reference taxonomy with the least classification error (for a given indicator) for each of the reads in the metagenomic sample

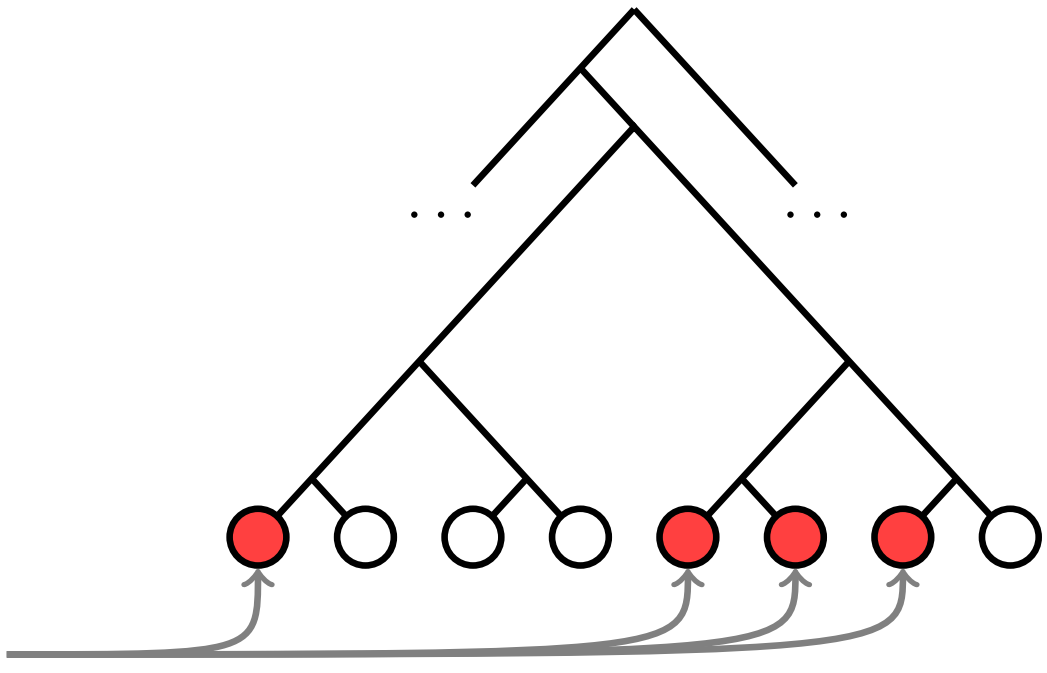
**Theorem**

*For each candidate node in a reference taxonomy there exists a node in the LCA skeleton tree of the candidate sequences with at most the same classification error*

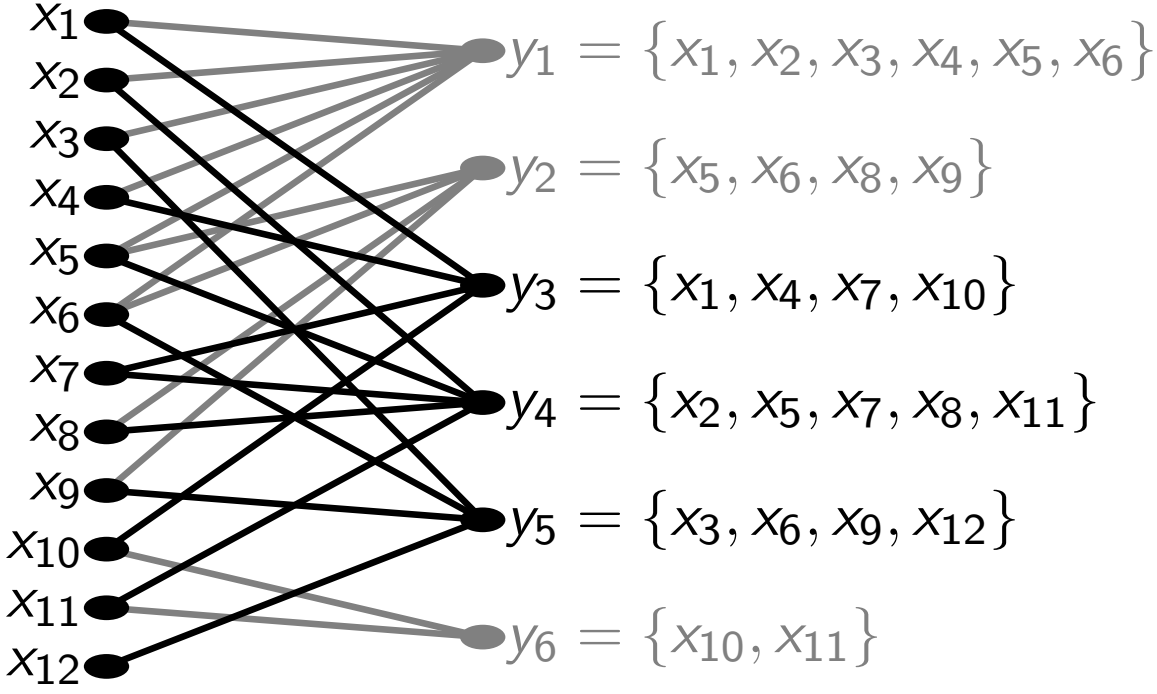
**Proof.**

$F$-measure (2011). Youden $J$, AUC, Jaccard similarity coefficient, Rand index (2017). Yule $\phi$ (left to the reader) □

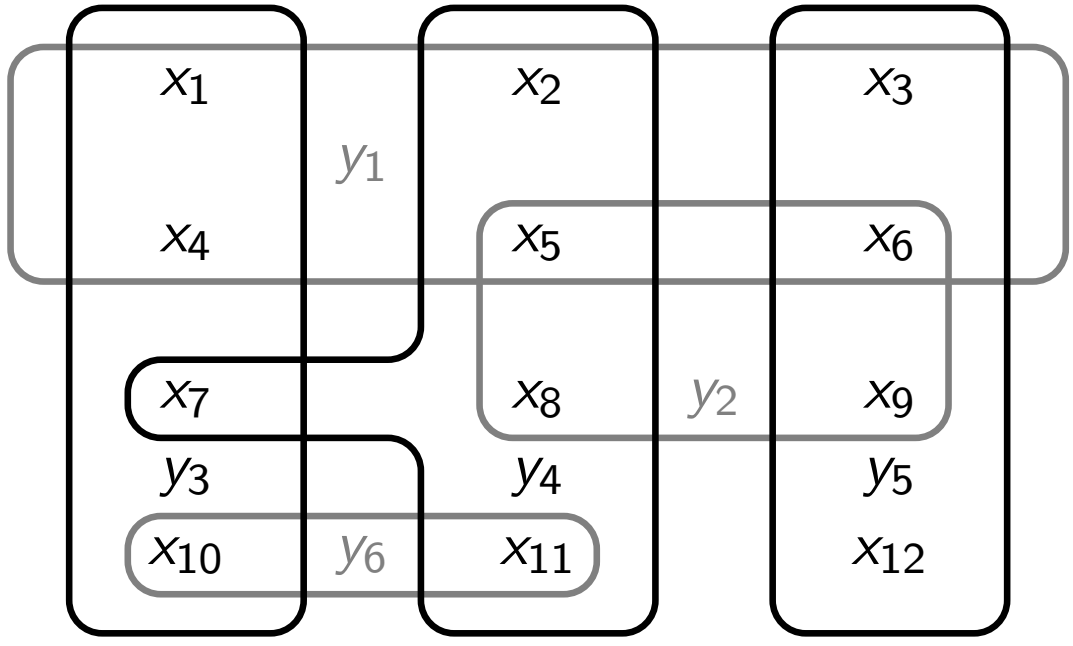
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT



ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT



ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT



**Definition**

Let $X$ be a finite set and let $C$ be a collection of subsets of $X$ whose union is $X$. The overlap of a set cover $C' \subseteq C$ is the total size of the subsets minus the size of $X$

**Corollary**

*A set cover with the least total size of subsets has the least overlap*

## Claim

*A set cover with the least number of subsets does not necessarily have the least overlap*

## Proof.

Let $X = \{1, \ldots, n\}$ and assume, without loss of generality, that $n = 2k$ for $k \geqslant 3$. Let $S$ be the following collection of subsets of $X$:

$$\{1, 2\}, \{3, 4\}, \ldots, \{n-1, n\}, \{1, \ldots, n-1\}, \{2, \ldots, n\}$$

The set cover $\{1, \ldots, n-1\}, \{2, \ldots, n\}$ has size 2, which is the smallest possible for $S$ and $X$, and overlap $n$. The set cover $\{1, \ldots, n-1\}, \{n-1, n\}$ also has size 2, but it has overlap 1. Same for the set cover $\{1, 2\}, \{2, \ldots, n\}$, and $S$ and $X$ have no other set cover of size 2. However, the set cover $\{1, 2\}, \{3, 4\}, \ldots, \{n-1, n\}$ has size $n/2$ and overlap 0, which is the least possible overlap $\square$

## Example (Abundance profile of a metagenomic sample)

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ |
|---|---|---|---|---|---|---|
| $x_1$ | ✓ | | ✓ | | | |
| $x_2$ | ✓ | | | ✓ | | |
| $x_3$ | ✓ | | | | ✓ | |
| $x_4$ | ✓ | | ✓ | | | |
| $x_5$ | ✓ | ✓ | | ✓ | | |
| $x_6$ | ✓ | ✓ | | | ✓ | |
| $x_7$ | | | ✓ | ✓ | | |
| $x_8$ | | ✓ | | ✓ | | |
| $x_9$ | | ✓ | | | ✓ | |
| $x_{10}$ | | | ✓ | | | ✓ |
| $x_{11}$ | | | | ✓ | | ✓ |
| $x_{12}$ | | | | | ✓ | |
| | 22.2% | 13.9% | 16.7% | 19.4% | 19.4% | 8.3% |

## Example (Abundance profile of a metagenomic sample)

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ |
|---|---|---|---|---|---|---|
| $x_1$ | ✓ | | ✓ | | | |
| $x_2$ | ✓ | | | ✓ | | |
| $x_3$ | ✓ | | | | ✓ | |
| $x_4$ | ✓ | | ✓ | | | |
| $x_5$ | ✓ | ✓ | | ✓ | | |
| $x_6$ | ✓ | ✓ | | | ✓ | |
| $x_7$ | | | ✓ | ✓ | | |
| $x_8$ | | ✓ | | ✓ | | |
| $x_9$ | | ✓ | | | ✓ | |
| $x_{10}$ | | | ✓ | | | ✓ |
| $x_{11}$ | | | | ✓ | | ✓ |
| $x_{12}$ | | | | | ✓ | |
| | | | 29.2% | 37.5% | 33.3% | |

## Example (LP formulation of the set cover approach)

| $a_{ij}$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ | $m_i$ |
|---|---|---|---|---|---|---|---|
| $x_1$ | 1 | 0 | 1 | 0 | 0 | 0 | 2 |
| $x_2$ | 1 | 0 | 0 | 1 | 0 | 0 | 2 |
| $x_3$ | 1 | 0 | 0 | 0 | 1 | 0 | 2 |
| $x_4$ | 1 | 0 | 1 | 0 | 0 | 0 | 2 |
| $x_5$ | 1 | 1 | 0 | 1 | 0 | 0 | 3 |
| $x_6$ | 1 | 1 | 0 | 0 | 1 | 0 | 3 |
| $x_7$ | 0 | 0 | 1 | 1 | 0 | 0 | 2 |
| $x_8$ | 0 | 1 | 0 | 1 | 0 | 0 | 2 |
| $x_9$ | 0 | 1 | 0 | 0 | 1 | 0 | 2 |
| $x_{10}$ | 0 | 0 | 1 | 0 | 0 | 1 | 2 |
| $x_{11}$ | 0 | 0 | 0 | 1 | 0 | 1 | 2 |
| $x_{12}$ | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| $n_j$ | 6 | 4 | 4 | 5 | 4 | 2 | 25 |

## LP formulation of the set cover approach

- $X = \{x_1, x_2, \ldots, x_{12}\}$ (reads)
- $Y = \{y_1, y_2, \ldots, y_6\}$ (candidate sequences) where
  - $y_1 = \{x_1, x_2, x_3, x_4, x_5, x_6\}$
  - $y_2 = \{x_5, x_6, x_8, x_9\}$
  - $y_3 = \{x_1, x_4, x_7, x_{10}\}$
  - $y_4 = \{x_2, x_5, x_7, x_8, x_{11}\}$
  - $y_5 = \{x_3, x_6, x_9, x_{12}\}$
  - $y_6 = \{x_{10}, x_{11}\}$
- Minimize $\sum_j n_j y_j$
- Subject to $\sum_j a_{ij} y_j \geqslant 1$ for all $i$

  and $y_j \geqslant 0$ for all $j$

  and $y_j \leqslant 1$ for all $j$

# Conclusion

The Rand index is a better indicator of classification error than the often used area under the ROC curve and $F$-measure

The taxonomic annotation problem for a whole metagenomic sample can be reduced to a set cover problem, for which a logarithmic approximation can be obtained in linear time and an exact solution can be obtained by linear programming

A solution to the set cover problem with the least total size of subsets minimizes the ambiguity in the taxonomic annotation of the reads in a metagenomic sample

# PhD in Bioinformatics

- Omics and Molecular Bioinformatics
- Biomolecular Modelling and Simulation
- Systems and Synthetic Biology
- Data Science in Bioinformatics
- Biostatistics and Mathematical Modelling in Bioinformatics

BIOINFORMATICS BARCELONA
http://www.bioinformaticsbarcelona.eu/

Application Deadline: 30 June 2017