

# XLVII Congresso Nazionale AMCLI

## Sessione 14

### LO STUDIO DEL MICROBIOTA: IL RUOLO DELL'AMCLI

# Come interpretare i dati a livello computazionale

**Dott.ssa Roberta De Grandi**

Laboratorio di Analisi Cliniche e Microbiologiche  
IRCCS Istituto Ortopedico Galeazzi, Milano



I.R.C.C.S. ISTITUTO ORTOPEDICO  
GALEAZZI

# Contenuti

- 1 Introduzione
- 2 Prima del sequenziamento
- 3 Come sequenziare
- 4 Analisi della diversità microbica
- 5 Annotazione e condivisione dei risultati

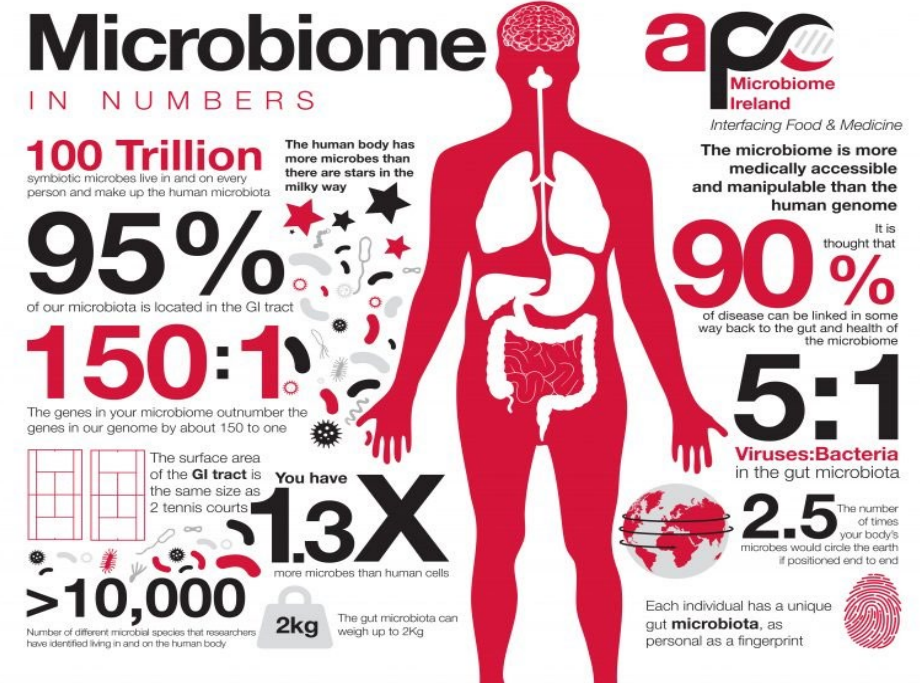
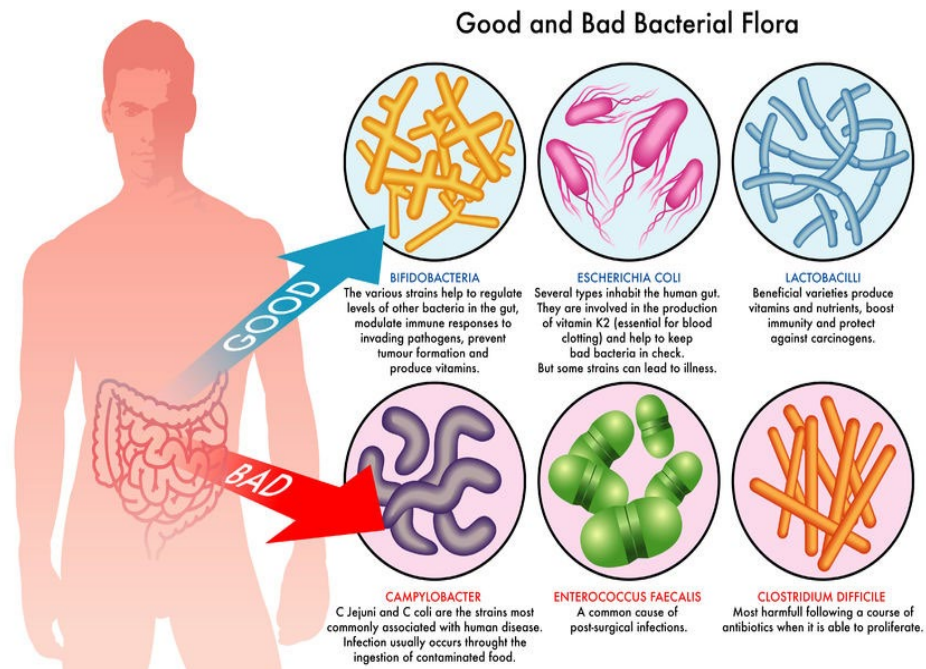
# Microbiota

Insieme di tutti i  
microorganismi colonizzanti il  
nostro corpo



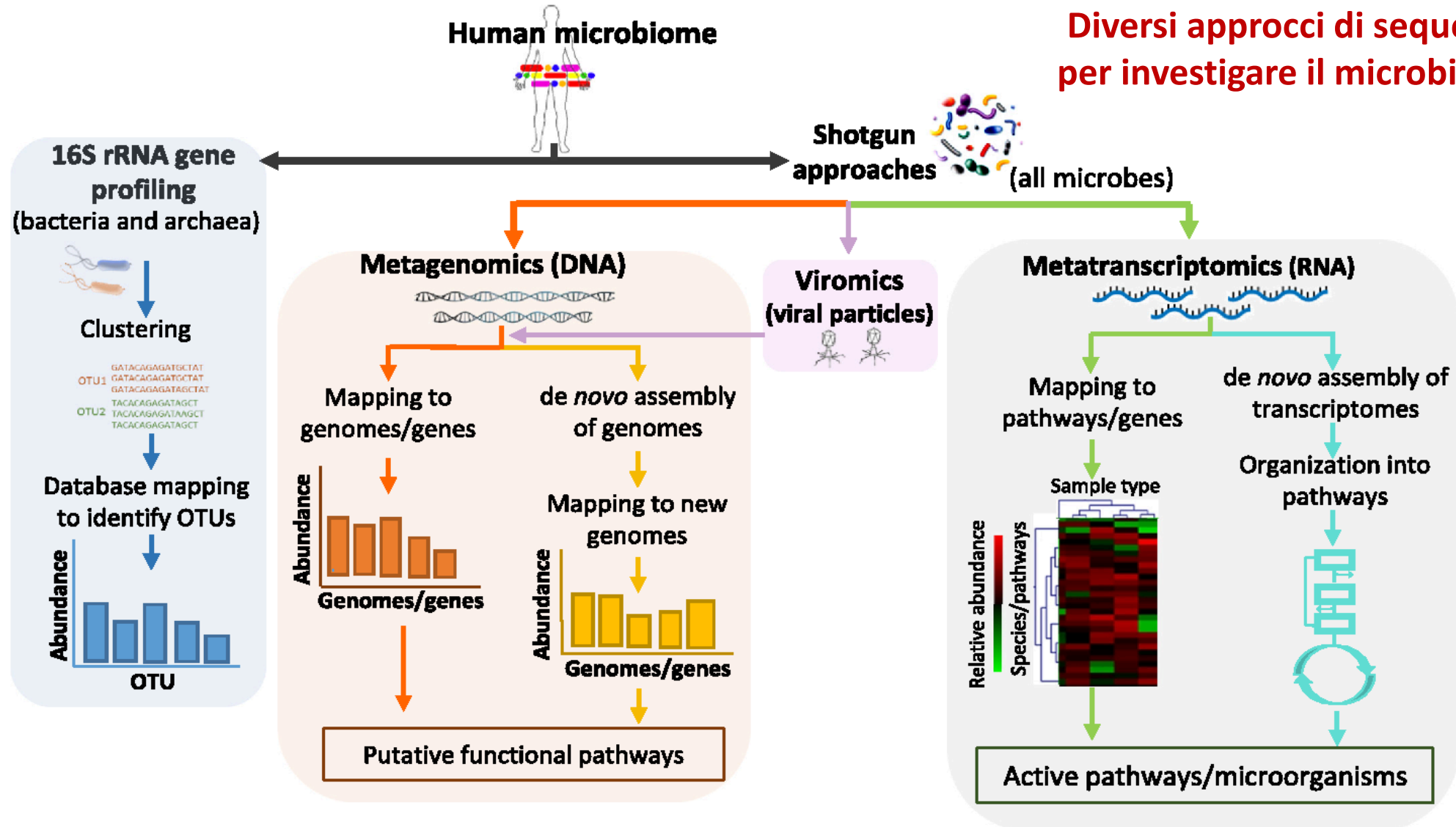
# Microbioma

Insieme dei microorganismi simbiotici,  
delle loro informazioni genetiche e delle  
loro interazioni con l'organismo ospite

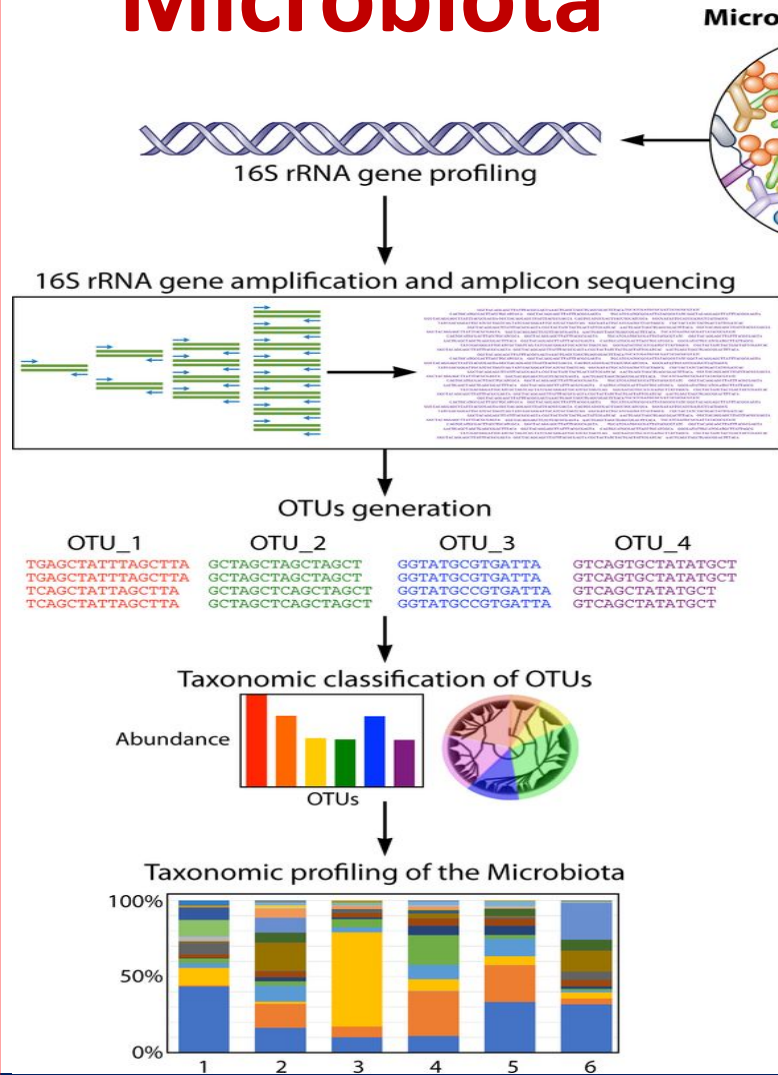


# Approccio di Sequenziamento

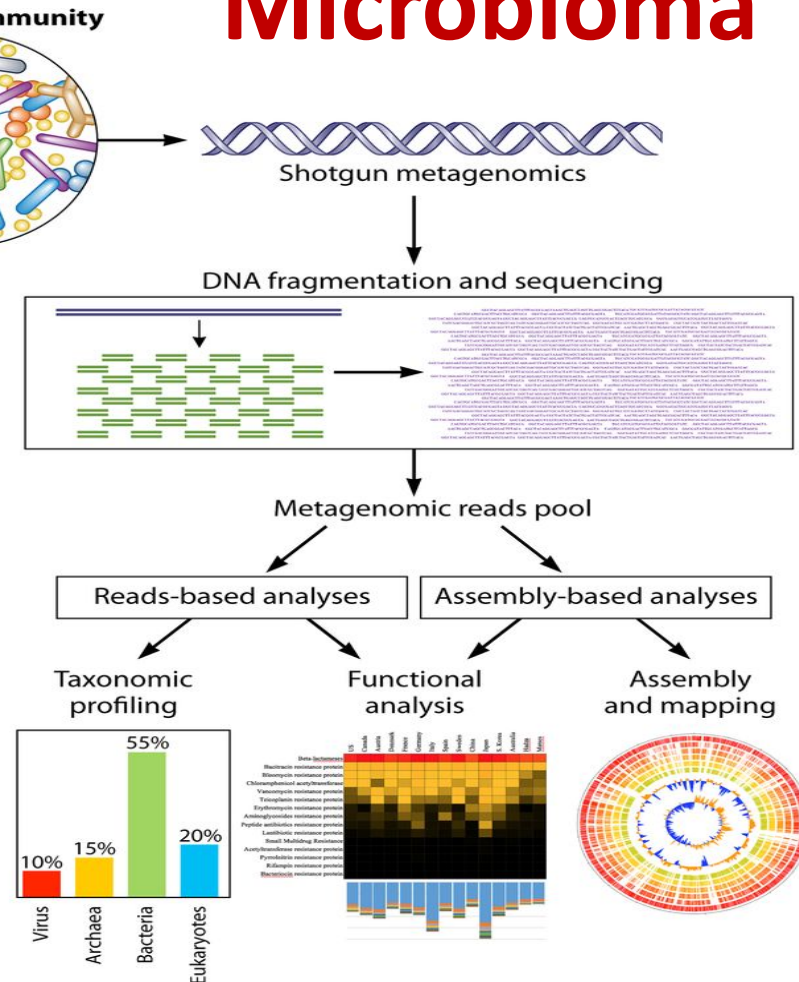
Diversi approcci di sequenziamento per investigare il microbioma umano



## Microbiota

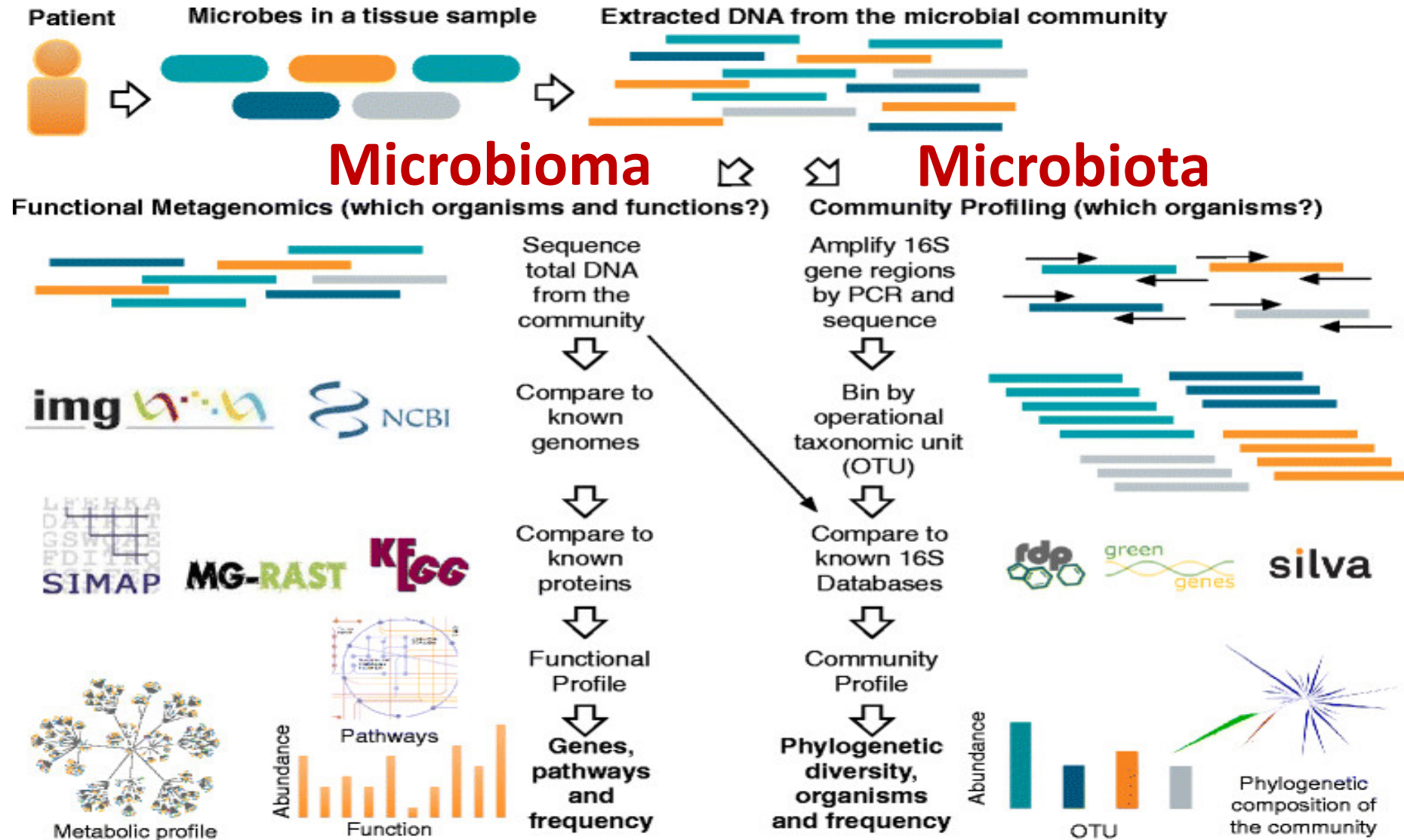





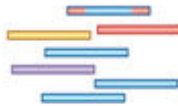

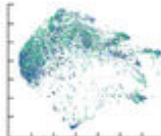
## Microbioma





# Approccio di Sequenziamento



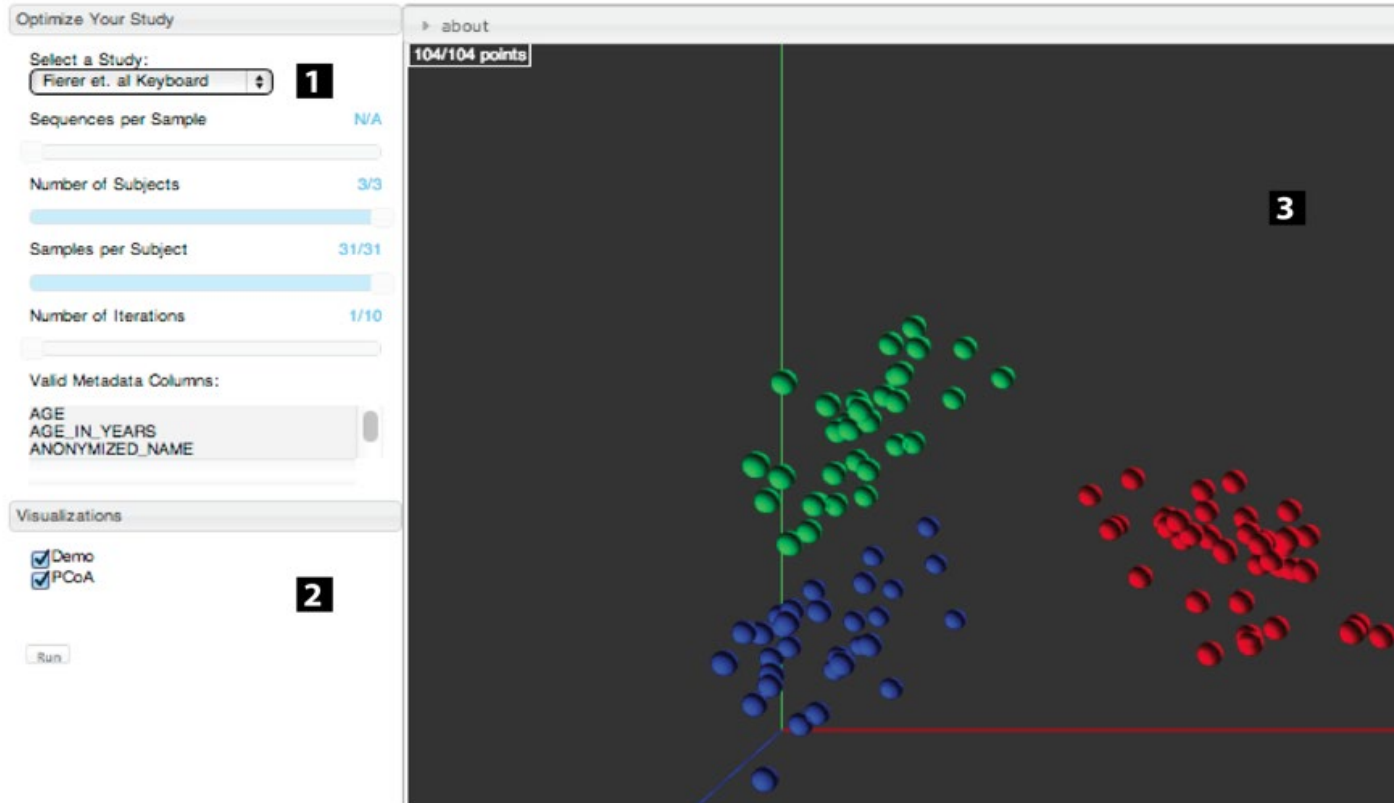
						
Sample processing step	Sample collection	Sample storage	DNA extraction	Sequencing library preparation	DNA sequencing	Computational analysis
Technical sources of error and bias	<ul style="list-style-type: none"><li>• Inadequate sampling</li><li>• Incomplete sample stabilization</li><li>• Sampling kit contamination</li><li>• Mislabeling of samples</li></ul>	<ul style="list-style-type: none"><li>• Change in community structure due to differential growth</li><li>• Degradation of DNA during freeze-thaw cycles</li></ul>	<ul style="list-style-type: none"><li>• Differential recovery of DNA from different strains</li><li>• Extraction kit contamination</li><li>• Sample swaps during transfer</li><li>• Sample cross-contamination</li></ul>	<ul style="list-style-type: none"><li>• Quantitative amplification bias (PCR efficiency)</li><li>• Qualitative amplification bias (primer mismatches)</li><li>• Amplification errors (PCR chimeras, substitution errors)</li><li>• Reagent contamination</li><li>• Sample cross-contamination</li></ul>	<ul style="list-style-type: none"><li>• Sequencing errors</li><li>• Run-to-run carryover</li><li>• Barcode swapping</li><li>• Demultiplexing errors</li></ul>	<ul style="list-style-type: none"><li>• Suboptimal quality control or filtering</li><li>• Alignment errors</li><li>• Database errors</li><li>• Database bias</li><li>• Batch effects</li><li>• Failure to flag contaminants</li></ul>
Costea <i>et al.</i> <sup>3</sup>			<ul style="list-style-type: none"><li>• Compare protocol for human fecal samples across 21 laboratories</li><li>• Recommend a standardized protocol</li></ul>			
Sinha <i>et al.</i> <sup>4</sup>			<ul style="list-style-type: none"><li>• Compare data generation processes across 15 laboratories and computational analyses across 9 groups</li><li>• Results inform future experimental design</li></ul>			

I metodi genomici per il sequenziamento di intere comunità presentano molte sfide tecniche e analitiche implicite nella metodica ma che possono anche variare ampiamente tra i diversi laboratori

# Disegno sperimentale

**Gli studi sul microbiota dipendono in maniera importante dal sample size**

e·vident



E' stato sviluppato per aiutare  
a stimare le dimensioni del  
campione richieste in base  
alla dimensione dell'effetto  
previsto utilizzando studi  
simili precedenti



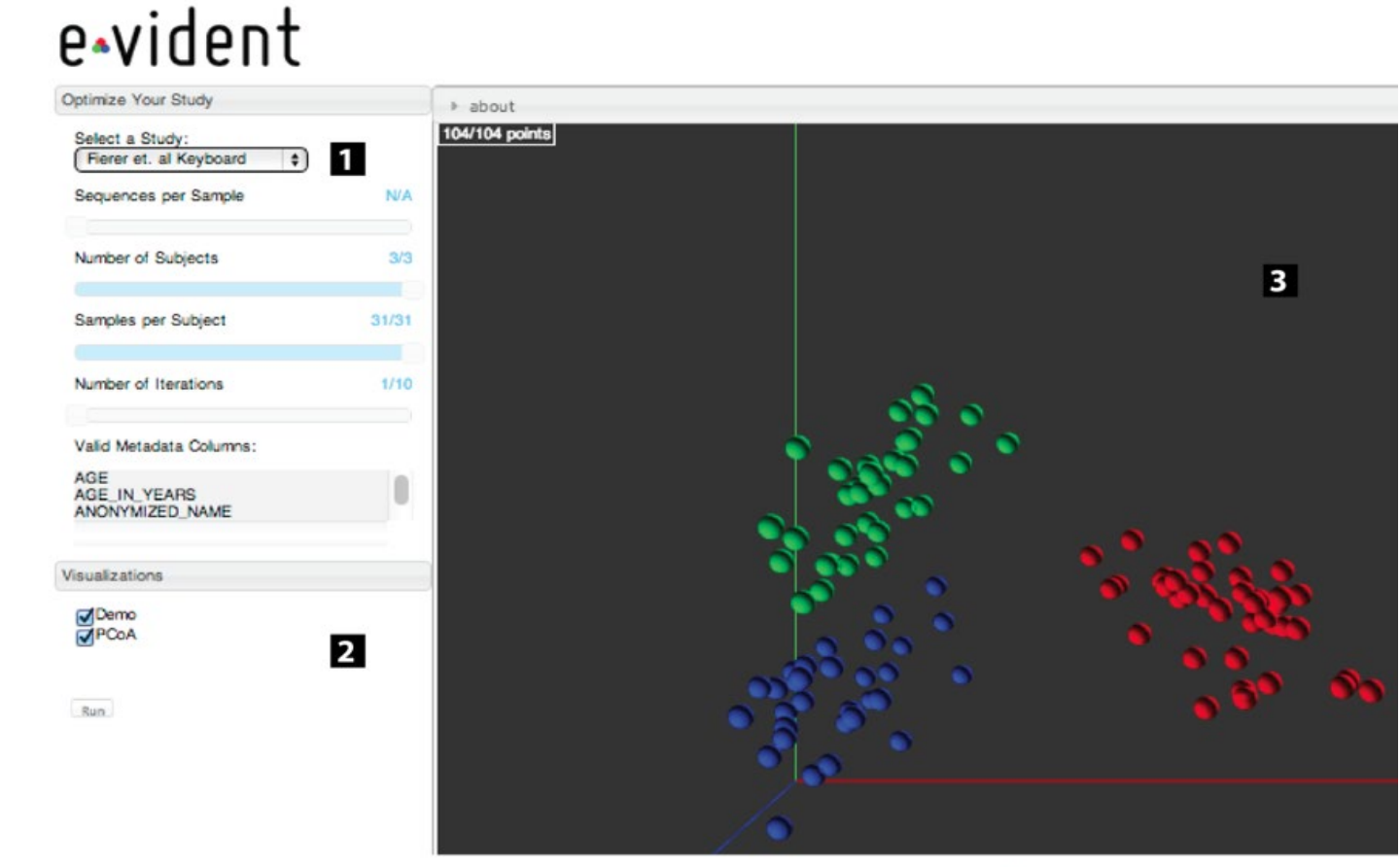
# Disegno sperimentale

**Gli studi sul microbiota dipendono in maniera importante dal sample size**



L'interfaccia di Evident fornisce :

- (1) la selezione dei parametri;
- (2) la selezione del tipo di visualizzazioni;
- (3) la visualizzazione del grafico WebGL.



# Disegno sperimentale

**Gli studi sul microbiota dipendono in maniera importante dal sample size**



L'interfaccia di Evident fornisce :

- (1) la selezione dei parametri;
- (2) la selezione del tipo di visualizzazioni;
- (3) la visualizzazione del grafico WebGL.



- ❖ Non c'è un metodo standardizzato per riportare le dimensioni degli effetti
- ❖ studi simili potrebbe non esistere

# Disegno sperimentale

## Studio pilota

per definire  
l'effetto sul sample size



## Campionamento ripetuto degli individui



Ricampionare gli stessi individui  
nel tempo o valutare più  
individui solo una volta?

## Non è definibile uno approccio analitico standard

Tuttavia, il campionamento ripetuto nel tempo  
fornirà una visione più completa della diversità.



## Perché condividere i tuoi dati e metadata in archivi pubblici?

- ❖ I dati senza metadata sono inutili;
- ❖ descrivere i propri dati in modo così dettagliato come vorreste trovare i dati di altri;
- ❖ usare descrittori standardizzati;
- ❖ gli archivi che presentando i dati danno credito ;
- ❖ tenere aggiornati i dati: allo stesso modo tenendo aggiornati i dati questi daranno valore alla ricerca e credito al ricercatore.

Please note that Project and Study have been merged into one common Study concept.

Home New Submission Studies Sample Groups Samples Experiments Runs Assemblies Variations

Start >> Study >> **Sample** >> Run >> Finish

Please create new samples by uploading a spreadsheet or by following the instructions below. If your data will refer only to previously submitted samples please skip to the 'Run' step.

Please select the checklist fields you would like to include with each sample. Recommended fields can be unselected from within the corresponding attribute group on the left-hand side panel. You may also add custom fields.

Filter fields... Add your own attribute... + Add

- + geography
- + non-sample terms
- + sample collection
- + host disorder
- + host description
- + local environment conditions
- + concentration measurement
- + other
- + host details
- + organism characteristics
- + User Fields

11 of 92 fields selected

+ Expand - Collapse Download Template

Please complete any fields that you would like to apply to all samples. This will act as a template for the rest of the samples.

**Template Basic Details**

Unique Name Prefix: TEST\_SAMPLE

\* Title: Dummy\_sample

Description: test sample created for demonstration purpose

**Organism Details**

If your organism is not found please go [here](#) and email [datasubs@ebi.ac.uk](mailto:datasubs@ebi.ac.uk) with the required details listed on the page in order for us to request a taxon Id for your organism.

Search: human met

\* Tax Id: [646099] human metagenome

\* Scientific Name: [646099] human metagenome

Common Name: [1621] human metagenome

\* Geography: [694067] Human metapneumovirus CAN97-83

Environmental package: host-associated

\* geographic location (latitude): DD

\* geographic location (longitude): DD

a) Enter the information common to all samples

The different components of the chosen checklist are displayed by section. Click to expand the view.

b) ESSENTIAL: make sure that you select a metagenomic tax Id!

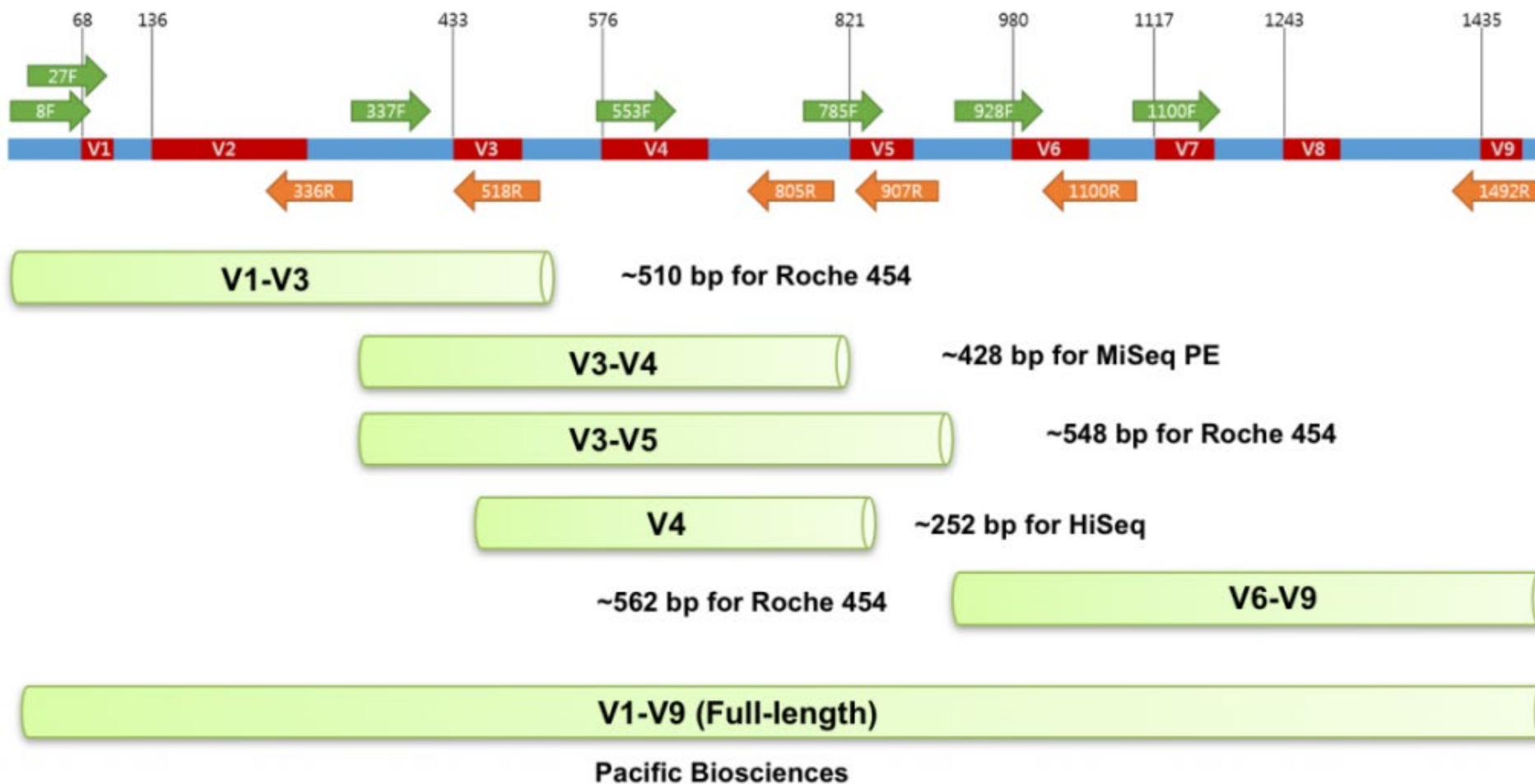
c) Enter the metadata on the right side

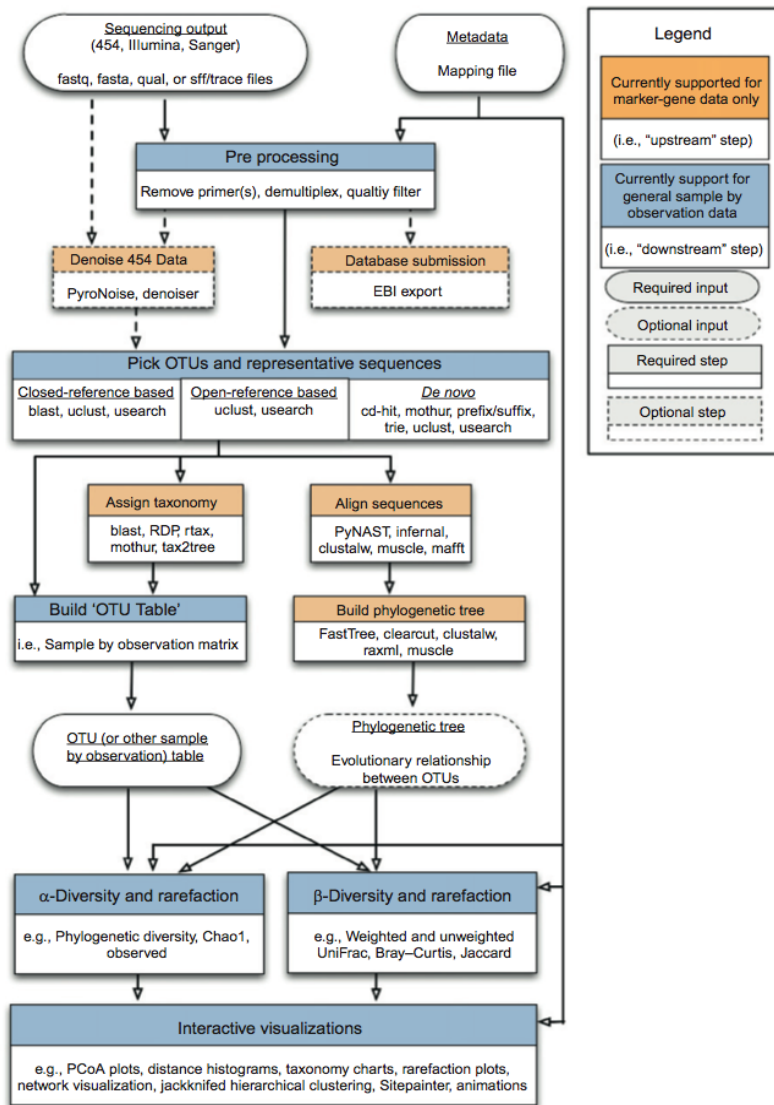
A red icon denotes a mandatory field

<< Previous Next >>



# Selezione dei primer



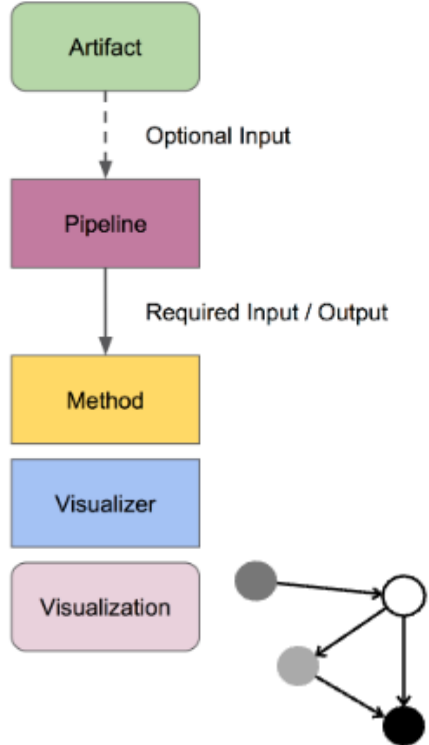


- ❖ Consente diverse modalità di data mining tra cui: il demultiplexing e il filtraggio di qualità, la selezione OTU, l'assegnazione tassonomica e la ricostruzione filogenetica, e analisi e visualizzazioni di diversità.
- ❖ Raccolta di molti programmi elaborazione dei dati grezzi mediante input da riga di comando.

3

Come sequenziare:

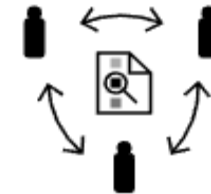
# Quali strumenti di analisi preferire



**Core concept semplificato**  
**Implementato di una interfaccia grafica**



Esplora in modo  
interattivo i tuoi dati con  
bellissime visualizzazioni  
che offrono nuove  
prospettive.



Facile condivisione dei  
risultati con il tuo  
team, anche quelli  
senza QIIME 2  
installato.



Sistema basato su plugin:  
i tuoi metodi di analisi  
preferiti in un unico  
posto.

Monitora automaticamente le tue  
analisi con la provenienza dei dati  
decentralizzata - non è più  
necessario fare ipotesi su quali  
comandi sono stati eseguiti!

**mothur****Download****Wiki****Forum****facebook**

- ❖ Consente l'analisi dei dati del gene16S rRNA;
- ❖ Consente di analizzare i dati utilizzando una pipeline molto simile a QIIME;
- ❖ Ha il vantaggio di essere un singolo programma da eseguire;
- ❖ Diversamente da QIIME manca un po' di una facile personalizzazione.

**Metriche utilizzate per descrivere l' alfa diversità:****Diversità  $\alpha$** **Shannon:**

- ❖ misura la probabilità di predire l'estrazione di una specie batterica tra quelle presenti nell'elenco delle specie identificate;
- ❖ utile a stimare la ricchezza di specie batteriche presenti nel campione.

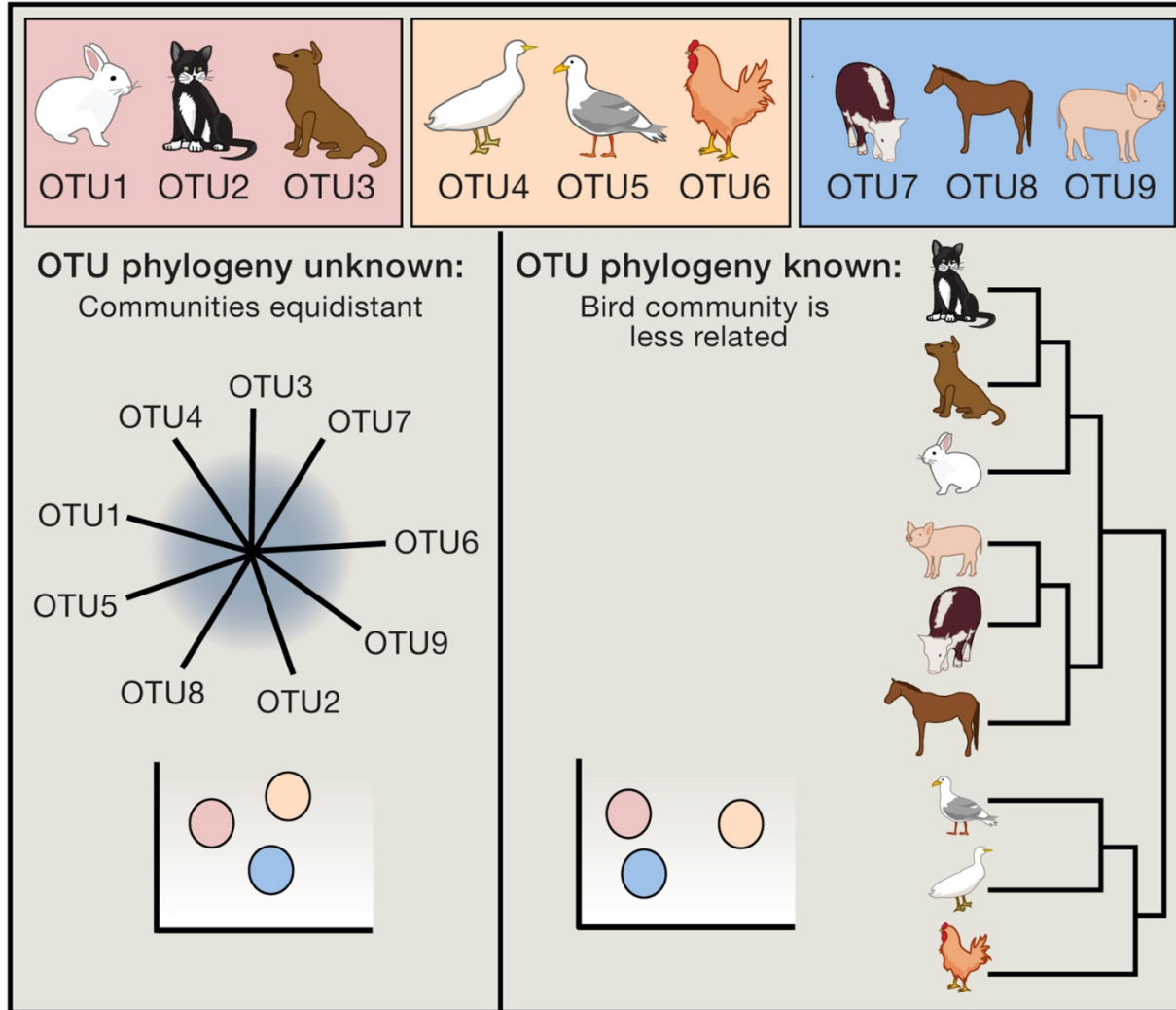
**Chao1:**

- ❖ stima il numero di OTU che non sono rilevati dal sequenziamento in base al numero di taxa batterici rappresentati da una singola lettura, definita singletons;
- ❖ utile ad avere una stima indiretta dell'abbondanza dei taxa più rari.

**Simpson:**

- ❖ assume valore 0 nel caso di una popolazione perfettamente omogenea e valore 1 nel caso di una popolazione eterogenea al massimo (ogni specie rappresentata nella stessa proporzione) ;
- ❖ rappresenta l'eterogeneità di distribuzione delle specie all'interno di un habitat.

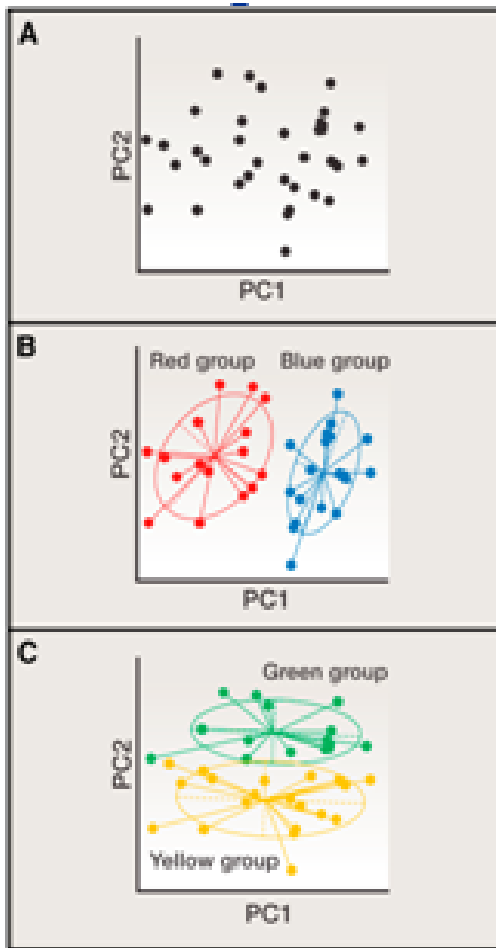


Diversità  $\beta$ 

- ❖ E' la misura della diversità tra i campioni.
- ❖ Le metriche di diversità beta possono essere raggruppate in diversi modi :
  1. **possono essere quantitativi** (usando abbondanza di sequenza, ad es. **Bray-Curtis o ponderata UniFrac**);
  2. **qualitativo** (considerando solo la presenza-assenza di sequenze, ad esempio, **Jaccard binario o UniFrac non pesato**);
  3. **possono essere basati sulla filogenia** (le metriche di UniFrac) **o meno** (Bray-Curtis, ecc.).
- ❖ Le metriche basate su filogenia possono risultare più potenti delle altre metriche nei confronti delle comunità.

## PCA e PCoA

## HeatMap

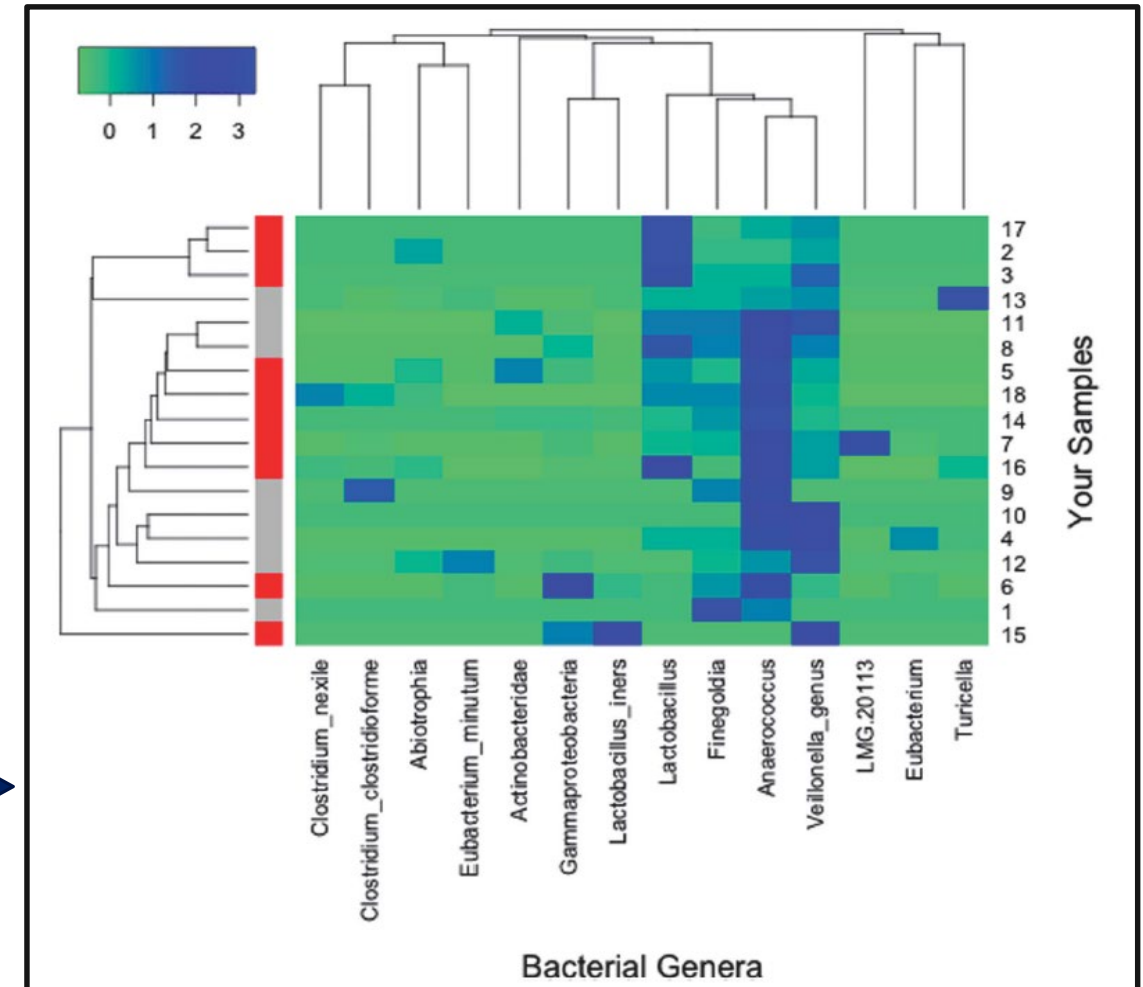


## Riduzione di dati

per organizzare le features  
costituenti il mio set di dati

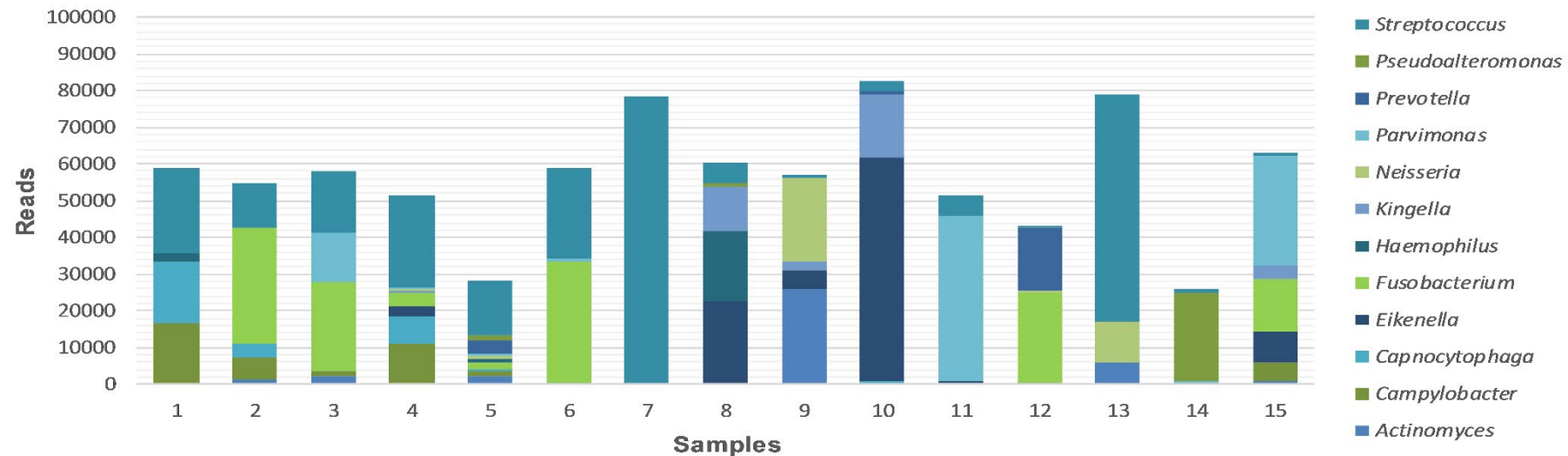
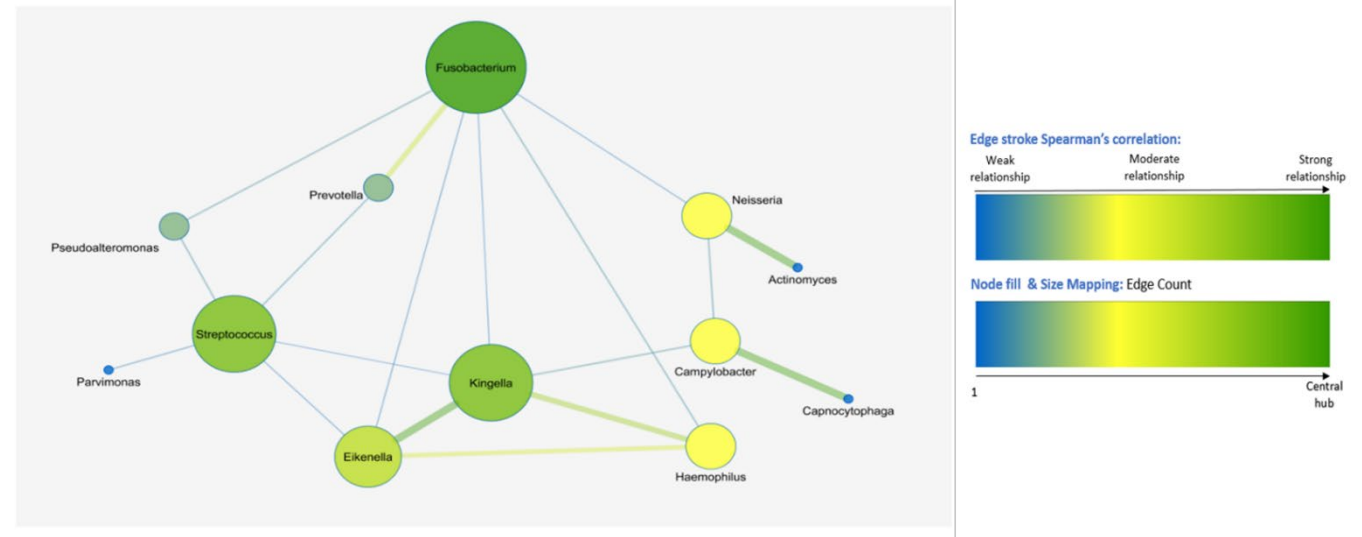
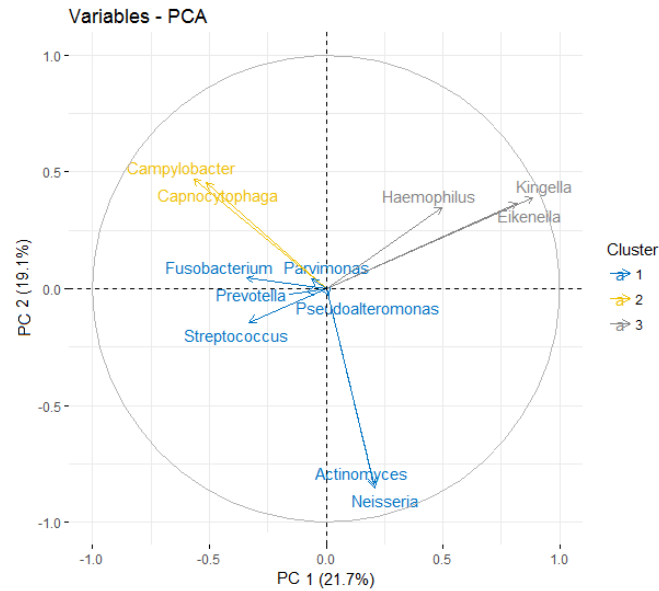
## Clustering

per capire quale sia l'impatto  
delle nostre OTU e della loro  
organizzazione nei campioni



- ❖ un modo per visualizzare quanto differenti specie batteriche o altri fattori (es. acidi grassi, fattori dell'ospite, metaboliti, ecc.) interagiscono tra di loro .

- ❖ un metodo utile per confrontare i diversi gruppi, evidenziando quali batteri / fattori coesistono e quale no

Esempio di interpretazione  
computazionale dei dati

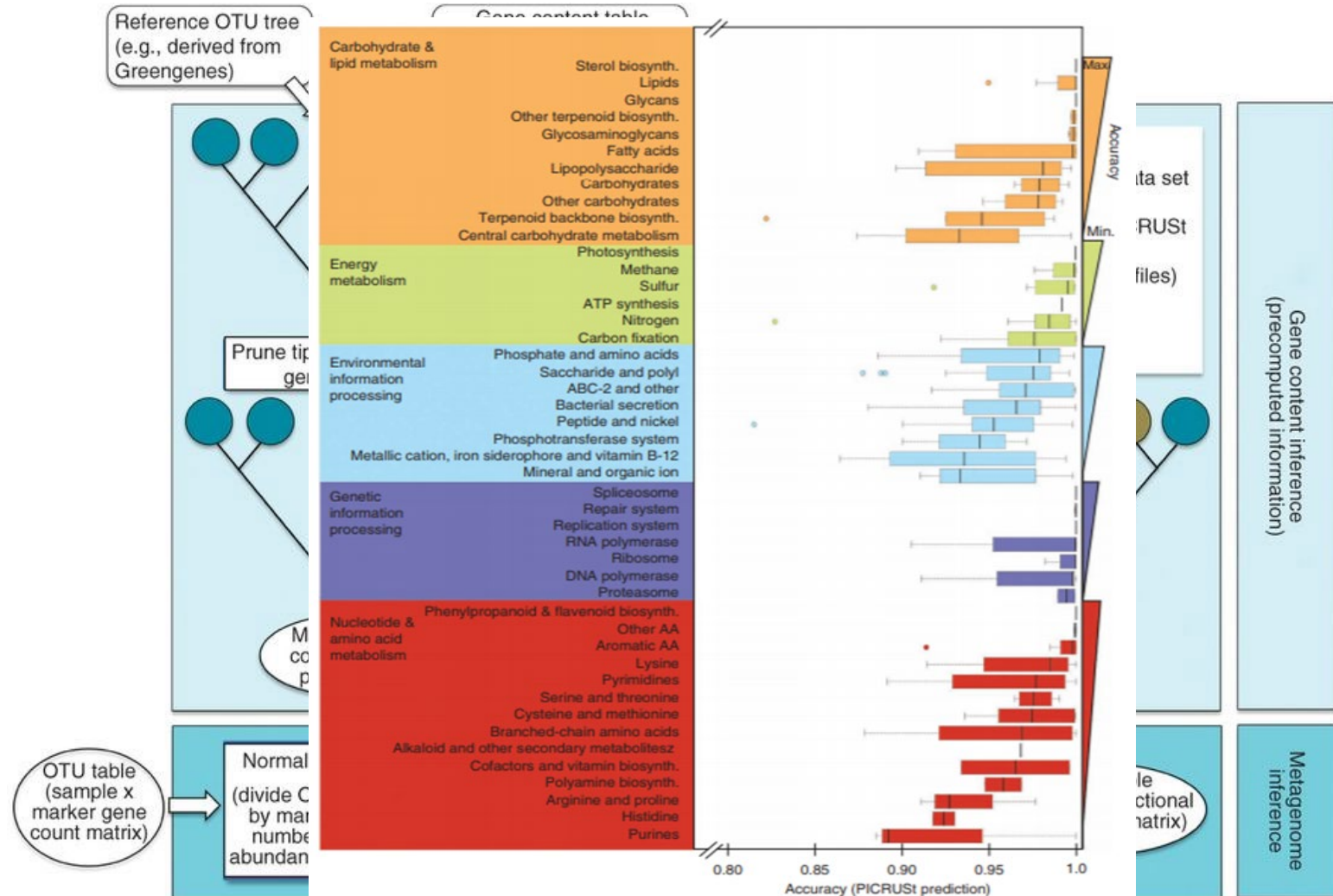


# Dedurre l'impatto funzionale di una comunità batterica

## PICRUSt

(Phylogenetic Investigation of Communities by Reconstruction of Unobserved States)

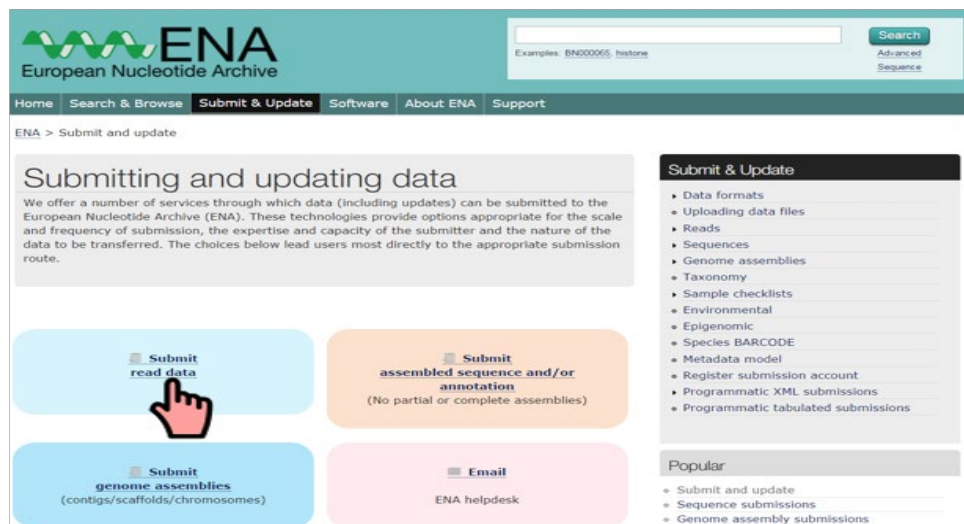
Consente previsioni funzionali su una comunità batterica, per cui è stato sequenziato il gene 16S rRNA.



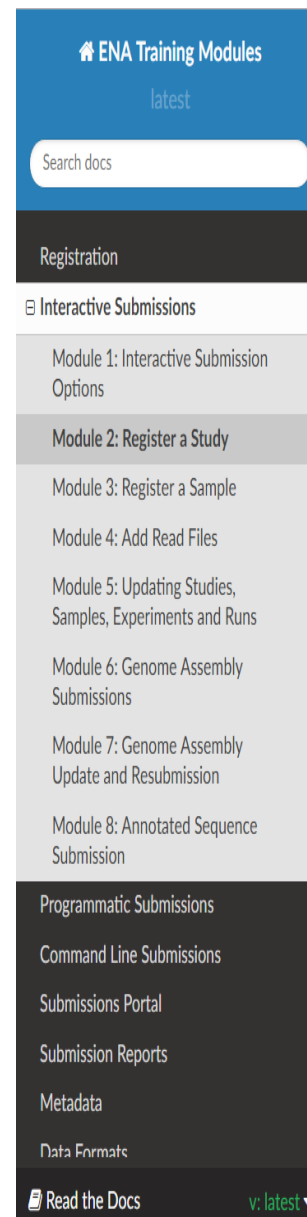
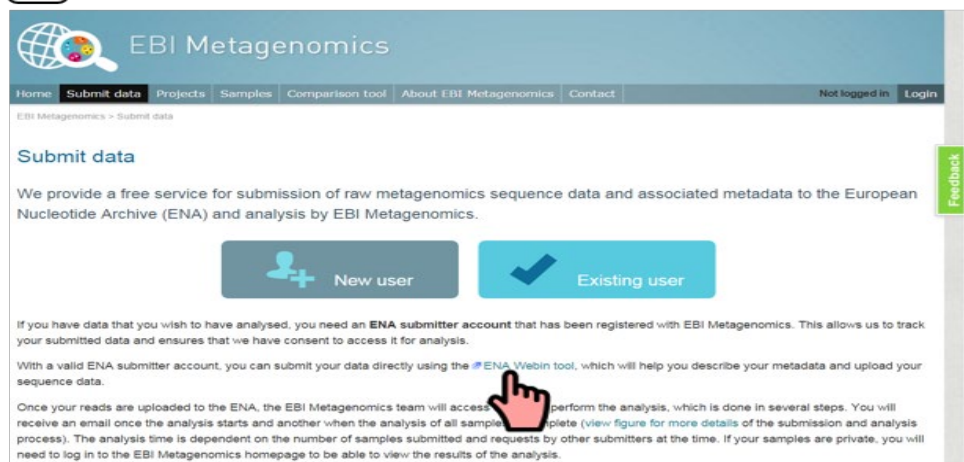


# 5 Annotazione e condivisione dei risultati:

A



B



Docs » Interactive Submissions » Module 2: Register a Study

Edit on GitHub

## Module 2: Register a Study

This form is used to register a study (also referred to as a project). Studies are typically registered before any data is submitted. Data can be added to the study at any time. Please see [Interactive Submission Options](#) to learn how to access it. Fill it out and click 'Submit' to register your study and receive accession numbers. The details can be edited and updated later, but you need to have registered a study to submit your data.

Please specify the release date of your study:  
This is when your study will be made public.

Please provide attributes to add a deeper description of the study:

Tag	FieldType
	Add

Please provide a short name for the study:

Please provide PubMed IDs of publications you want to associate with the study:  
(numeric value)

PubMed IDs

Add

Please provide a short descriptive title for the study: (\*)

For genome assembly projects only: In this study, will you provide functional genome annotation? (\*)  
PLEASE ANSWER WITH YES IF YOU HAVE ANNOTATION: Locus tag prefixes are only associated to studies providing functional genome annotation.

Yes  
No

Please provide an abstract to describe the study in detail: (\*)

<< Previous Submit

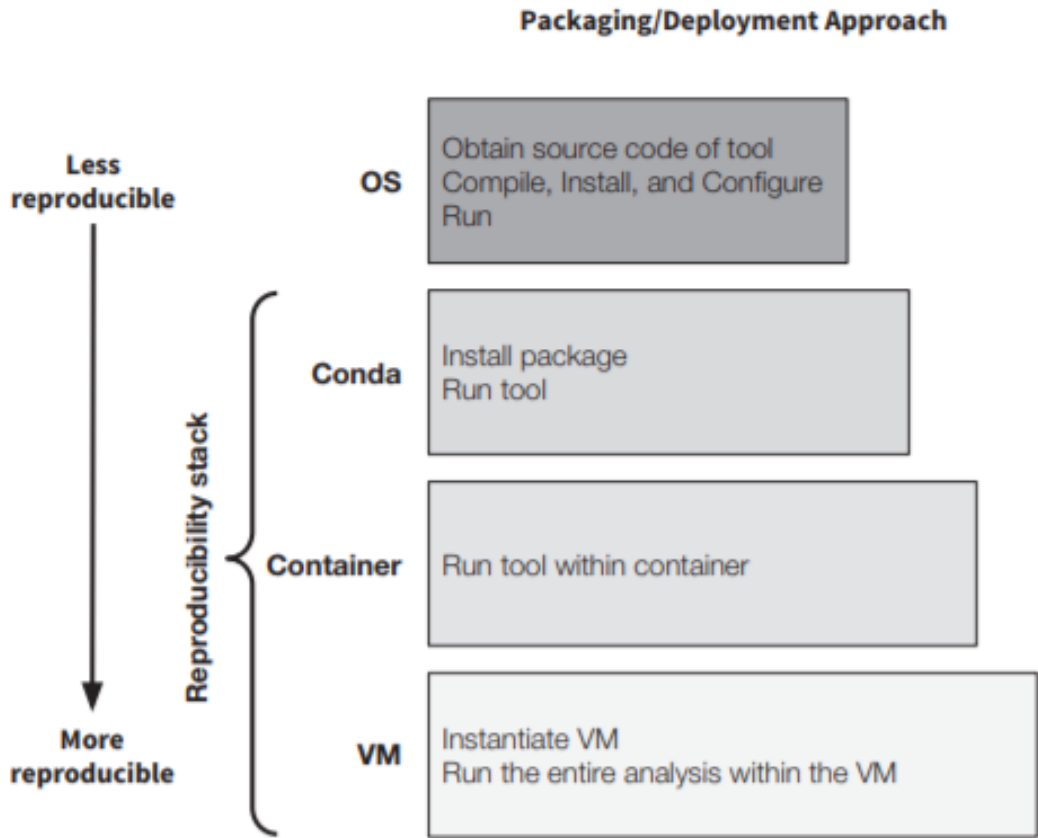
Webin will report two accession numbers for the study. The first starts with PRJEB and is called the BioProject accession. This is typically used in journal publications. The study will also be assigned an alternative accession number that starts with ERP. This accession number is called the SRA (Sequence Read Archive) study accession.

## Practical Computational Reproducibility in the Life Sciences

Björn Grüning,<sup>1</sup> John Chilton,<sup>2</sup> Johannes Köster,<sup>3</sup> Ryan Dale,<sup>4</sup> Nicola Soranzo,<sup>5</sup> Marius van den Beek,<sup>6</sup> Jeremy Goecks,<sup>7</sup> Rolf Backofen,<sup>1,\*</sup> Anton Nekrutenko,<sup>2,\*</sup> and James Taylor<sup>8,\*</sup>

- <sup>1</sup>Albert Ludwigs University, Freiburg, Germany
- <sup>2</sup>The Pennsylvania State University, University Park, PA, USA
- <sup>3</sup>University of Duisburg-Essen, Essen, Germany
- <sup>4</sup>National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, USA
- <sup>5</sup>Earlham Institute, Norwich, UK
- <sup>6</sup>Institut Curie, Paris, France
- <sup>7</sup>Oregon Health & Sciences University, Portland, OR, USA
- <sup>8</sup>Johns Hopkins University, Baltimore, MD, USA

\*Correspondence: [backofen@informatik.uni-freiburg.de](mailto:backofen@informatik.uni-freiburg.de) (R.B.), [anton@nekrut.org](mailto:anton@nekrut.org) (A.N.), [james@taylorlab.org](mailto:james@taylorlab.org) (J.T.)  
<https://doi.org/10.1016/j.cels.2018.03.014>



# Conclusioni

- ❖ Indipendentemente dalle metodologie utilizzate per investigare il microbioma umano, alcune linee guida sono fondamentali per la realizzazione di studi che contribuiscano efficacemente al conseguimento di un' approfondita caratterizzazione delle comunità batteriche associate a differenti condizioni fisio-patologiche.
- ❖ Queste raccomandazioni interessano principalmente:
  - un'attenta progettazione dello studio;
  - la raccolta e conservazione di dettagliati metadati;
  - l'applicazione di metodi sperimentali e analitici che siano tra loro coerenti;
  - la conservazione di buona parte della documentazione raccolta perché possa essere utilizzata per la progettazione di modelli statistici;
  - l'utilizzo di software e strumenti statistici che consentano di generare registri di analisi e tengano conto delle versioni degli strumenti bioinformatici applicati;
  - il deposito in database pubblici di tutti i dati raccolti attraverso l'utilizzo di formati standard fruibili anche per ulteriori indagini.



I.R.C.C.S. ISTITUTO ORTOPEDICO  
GALEAZZI



**Laboratorio di Analisi Cliniche**

**I.R.C.C.S. ISTITUTO ORTOPEDICO GALEAZZI**

**tel: 0039 02 6621 4982**

**Responsabile: Dott.ssa Elena De Vecchi**

**Research Team: Dott.ssa Roberta De Grandi**

**([roberta.degrandi@grupposandonato.it](mailto:roberta.degrandi@grupposandonato.it))**

**Dott. Alessandro Bidossi**

**Dott.ssa Marta Bottagisio**

**Grazie  
dell'attenzione**