




Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



Analisi di dati metatassonomici. Un flusso di lavoro in R

Prof. Eugenio Parente
Scuola SAFE



1



Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA





In questo webinar


- due parole sugli approcci omici
- due parole sulle piattaforme di sequenziamento
- due parole sulle regioni target (specialmente per batteri)
- flusso di lavoro per un'analisi metatassonomica
 - in generale
 - come organizzare le cartelle
 - importazione, controllo qualità, filtri
 - deduplicazione, stima del modello di errore, inferenza delle ASV


Analisi di dati metatassonomici. Un flusso di lavoro in R

2









Cosa cercare

Divisione
Review

Molecular BioSystems
Molecular BioSystems

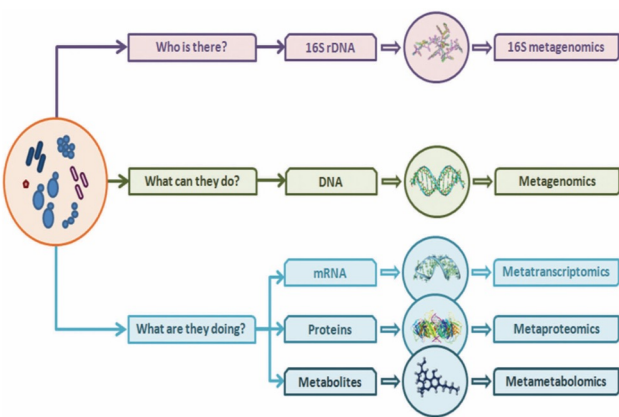






Fig. 1 Outline of the approaches available for studying the milk microbiota.

3

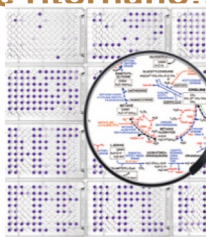
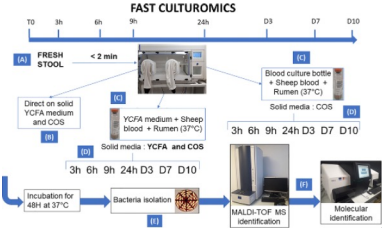








A volte ritornano...

- **fenomica**: lo studio ad alta capacità delle proprietà fenotipiche dei microrganismi e delle comunità microbiche
- **culturomica**: approccio basato sull'isolamento in un gran numero di condizioni diversi e sulla caratterizzazione ad alta capacità mediante spettrometria MALDI-ToF e sequenziamento

Analisi di dati metatranscriptomici. Un flusso di lavoro in 8

4

Finanziato dall'Unione europea
 NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Un po' di terminologia per le analisi metatassonomiche

- in letteratura sono identificate con molti nomi e sigle diverse: amplicon targeted metagenomics, 16S metagenomics o 18S ... o ITS2, etc.
- generalmente basate su PCR, target generalmente sequenze ribosomiali o spacer (ma possono essere altri geni)
- prevalentemente realizzate con short read (150-250 bp paired end) ma sempre più spesso long reads (fino a full length per 16S)
- se short reads ampliconi fra ca 200 e ca 450 bp (dipende dalla regione)

Analisi di dati metatassonomici. Un flusso di lavoro in R

5

Finanziato dall'Unione europea
 NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Popular Applications & Methods	Benchtop Sequencers				
	ISeq 100	MiSeq	MiSeq bench Q	NextSeq 250 Series Q	NextSeq 1000 & 2000
Large Whole-Genome Sequencing (human, plant, animal)					
Small Whole-Genome Sequencing (microbe, virus)					
Exome & Large Panel Sequencing (enrichment-based)					
Targeted Gene Sequencing (amplicon-based, gene panel)					
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging arrays)					
Transcriptome Sequencing (total RNA-Seq, bulk RNA-Seq, gene expression profiling)					
Targeted Gene Expression Profiling					
mRNA & Small RNA Analysis					
DNA-Protein Interaction Analysis (ChIP-Seq)					
Methylation Sequencing					
16S Metagenomic Sequencing					
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)					
Cell-Free Sequencing & Liquid Biopsy Analysis					

Benchtop Sequencer Sheds Light on Ebola Outbreak

Local scientists use the ISeq 100 Sequencing System to analyze transmission patterns and trace the origin of an Ebola outbreak in the Democratic Republic of the Congo.

[Read Article](#)

Run Time	9.5-19 hrs	4-24 hours	4-45 hours	12-30 hours	11-48 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	100 Gb	330 Gb*
Maximum Reads Per Run	4 million	25 million	25 million†	600 million	1.1 billion*
Maximum Read Length	2 x 150 bp	2 x 150 bp	2 x 300 bp	2 x 150 bp	2 x 150 bp

Popular Applications & Methods	Production-Scale Sequencers		
	NextSeq 1000 & 2000	NextSeq 8000 Series Q	NextSeq X Series
Large Whole-Genome Sequencing (human, plant, animal)			
Small Whole-Genome Sequencing (microbe, virus)			
Exome & Large Panel Sequencing (enrichment-based)			
Targeted Gene Sequencing (amplicon-based, gene panel)			
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging arrays)			
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)			
Chromatin Analysis (ATAC-Seq, ChIP-Seq)			
Methylation Sequencing			
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)			
Cell-Free Sequencing & Liquid Biopsy Analysis			

See what's possible with the NovaSeq X series

Larger projects, deeper sequencing, faster than ever. Breakthrough innovations for groundbreaking discoveries.

[Contact an Instrument Specialist](#)

Run Time	11-48 hours	-13-38 hours (ISeq SP flow cells) -13-25 hours (ISeq S1 flow cells) -18-24 hours (ISeq S2 flow cells) -16-26 hours (ISeq S2 flow cells) -42 hours (ISeq S4 flow cells)	-13-21 hours (1.5B flow cells) -18-24 hours (1.05B flow cells) -48 hours (2.5B flow cells)*
Maximum Output	360 Gb*	6000 Gb	16 Tb
Maximum Reads Per Run	1.2 billion*	20 billion	28 billion (single flow cells) 52 billion (ISeq flow cells)
Maximum Read Length	2 x 150 bp	2 x 250 bp**	2 x 150 bp


paired-end
150-250 bp
alta qualità
benchtop:


- runtime 9,5-48 h
- 1,2-330 Gb
- 4x10⁶-1,1x10⁹ reads/run


Un breve confronto fra le più recenti piattaforme Illumina


Analisi di dati metatassonomici. Un flusso di lavoro in R


6












MiniON and Flongle Flow Cell compatible




PromethION Flow Cell compatible

Configuration	Platform				Techniques				Tech specifications			
Number of flow cells per device	1	1	1	5	2	2	24	48				
Maximum number of channels per flow cell	512	512	512	512	2,675	2,675	2,675	2,675				
Run time	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours				
Device TMO†	50 Gb	50 Gb	50 Gb	250 Gb	580 Gb	580 Gb	~7 Tb	~14 Tb				
Maximum number of flow cells per year*	104	104	104	520	208	208	2,596	4,992				
Offer sequencing as a service	No	No	No	Yes	Yes	Yes	Yes	Yes				


bassa qualità (Q 10,7, in miglioramento fino a Q20+)
 lunghezze anche >10 kb
 benchtop:
 • runtime 24-72 h
 • 50-14000 Gb
 • ? reads/run


Un breve confronto fra le più recenti piattaforme Oxford Nanopore





Analisi di dati metatassonomici. Un flusso di lavoro in R

7





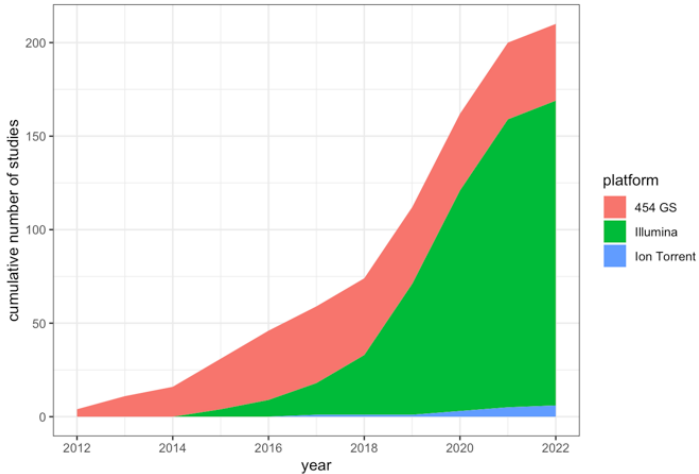




Le piattaforme per 16S metagenomics

Dati da FoodMicrobionet 4.2 (che non include dati da PacBio e Nanopore)
 Il grafico mostra il numero cumulativo di studi per ciascuna piattaforma
 L'anno è quello della pubblicazione, non quello del sequenziamento

The growth of FMBN, studies, by platform



Analisi di dati metatassonomici. Un flusso di lavoro in R

8

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI SICUREZZA E RESILIENZA

SUS-MIRRI.IT

Le regioni iù usate come target per 16S metagenomics

Dati da FoodMicrobionet 4.2 (che non include dati da PacBio e Nanopore)

Il grafico mostra il numero cumulativo di studi per ciascuna regione o combinazione di regioni

L'anno è quello della pubblicazione, non quello del sequenziamento

The growth of FMBN, studies, by 16S region

Analisi di dati metagenomici. Un flusso di lavoro in R

9

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI SICUREZZA E RESILIENZA

SUS-MIRRI.IT

Efficacia nell'identificazione a livello di specie

- le regioni target più corte (V4 e in minor misura V3) possono essere inadeguate persino per l'attribuzione del genere
- con V4 è sostanzialmente impossibile assegnare la specie, in particolare per generi con specie filogeneticamente molto vicine
- con V1-V3 risultati migliori che con V4
- con full length???????

Analisi di dati metagenomici. Un flusso di lavoro in R

10



Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



I passaggi di una tipica analisi metatassonomica - 1

- Fasi di pianificazione
 - disegno sperimentale
 - numero e tipo di campioni
 - bianchi, controlli, mock
- Esecuzione in laboratorio
 - campionamento
 - estrazione e purificazione del DNA
 - (preparazione delle library)

Analisi di dati metatassonomici. Un flusso di lavoro in R

11



Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



I passaggi di una tipica analisi metatassonomica - 2

- Sequenziamento (in house o service)
 - barcoding
 - sequenziamento
 - base calling, controllo qualità
 - demultiplexing (file fastq)

Analisi di dati metatassonomici. Un flusso di lavoro in R

12

I passaggi di una tipica analisi metatassonomica - 3

- **Analisi bioinformatica**
 - caricamento sequenze
 - (rimozione primer)
 - QC e filtri
 - OTU picking o **inferenza ASV**
 - ulteriori filtri
 - assegnazione della tassonomia
 - allineamento e calcolo di matrici di distanza
 - assemblaggio di tabelle di abbondanza, tabelle dei campioni, tabelle della tassonomia e dendrogrammi in un oggetto (phyloseq)

Analisi di dati metatassonomici. Un flusso di lavoro in R


13


I passaggi di una tipica analisi metatassonomica - 4


- **Analisi biostatistica**
 - eventuali filtri
 - calcolo di statistiche generali (nelle diverse fasi del processo)
 - analisi descrittiva
 - alpha diversity (indici e rappresentazione della composizione)
 - beta diversity (tecniche di ordinamento unconstrained e constrained, heatmaps, ...)
 - analisi differenziale
 - adonis / permanova
 - analisi differenziale di abbondanza
 - network analysis

Analisi di dati metatassonomici. Un flusso di lavoro in R

14









La pipeline

- utilizzeremo una pipeline ben collaudata, basata sull'inferenza di Amplicon Sequence Variants con DADA2 (disponibile anche in QIIME2), modificata da <https://f1000research.com/articles/5-1492/v2>
- la pipeline contiene alcune modificazioni per renderla multiplatforma e più flessibile e per produrre e salvare alcuni elementi utili
- l'assegnazione della tassonomia è fatta con la funzione assignTaxonomy() (che usa un naïve Bayesian classifier) con Silva v138.1 come database tassonomico di riferimento
- la pipeline si conclude con la creazione di un oggetto phyloseq



The DADA2 1.26 release is live, with native support for ARM architectures such as the Apple M1/M2 chips! [Release notes.](#)

Installation

Binaries for the current release version of DADA2 (1.26) are available from Bioconductor. Note that you must have R 4.2.0 or newer, and Bioconductor version 3.16, to install the most current release from Bioconductor.

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("dada2", version = "3.16")
```

If you wish to install the latest and greatest development version, or to install to earlier versions of R, see our [from-source installation instructions.](#)

Tutorials

Start here. The DADA2 tutorial goes through a typical workflow for paired-end Illumina MiSeq data: raw amplicon sequencing data is processed into the table of exact amplicon sequence variants (ASVs) present in each sample.


The DADA2 Workflow on Big Data goes through workflow optimized to run on large datasets (10s of millions to billions of reads).


An ITS-specific version of the DADA2 workflow identifies and verifiably removes primers on both ends of each ITS read, a key step due to the variable length of the ITS region.





Analisi di dati metatassonomici. Un flusso di lavoro in R

15



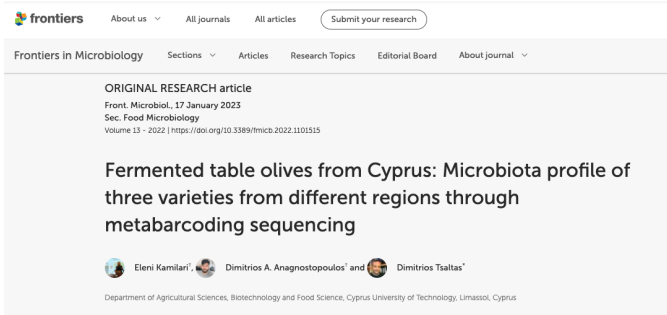






Il set di dati

- un piccolo set di dati su analisi metatassonomica (V3-V4, ma anche ITS1) di olive fermentate cipriote
- depositato in NCBI SRA con accession SRP399517
- numero limitato di campioni, qualità e numero delle sequenze abbastanza buone
- sequenze e metadata scaricabili da SRA e disponibili su Dropbox




frontiers About us All journals All articles Submit your research

Frontiers in Microbiology Sections Articles Research Topics Editorial Board About journal

ORIGINAL RESEARCH article
Front. Microbiol., 17 January 2023
Sec. Food Microbiology
Volume 13 - 2022 | <https://doi.org/10.3389/fmicb.2022.1101515>

Fermented table olives from Cyprus: Microbiota profile of three varieties from different regions through metabarcoding sequencing

Eleni Kamilari, Dimitrios A. Anagnostopoulos, and Dimitrios Tsaltas*
Department of Agricultural Sciences, Biotechnology and Food Science, Cyprus University of Technology, Limassol, Cyprus



Analisi di dati metatassonomici. Un flusso di lavoro in R

16

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI SICUREZZA E RESILIENZA

Lavorare organizzati

- usa la cartella di esempio per vedere come sono organizzate le cartelle che contengono i dati
- una cartella “madre” che contiene il progetto e gli script e dove verranno salvati alcuni dei file generati dall'analisi
- una sottocartella «data» contenente:
 - «fastq» con i file fastq (che devono avere nomi coerenti)
 - «filtered» dove verranno salvati i fastq dopo l'applicazione di filtri QC
 - «metadata» con i metadati
- una cartella «tax_db» con i database di riferimento al livello superiore della cartella madre

Nome	Modificato	Dimensione	Tipologia
BC_SRP399517_cypr.tab.olives.Rproj	Today at 16:00	205 bytes	R Project
bioconductor_pip_v6_3_3_0623_SRP399517.R	Today at 16:00	43 KB	R Source File
data	Today at 15:59	--	Folder
fastq	7 Mar 2023 at 14:41	--	Folder
SRR21703524.1.fastq	7 Mar 2023 at 14:41	12.4 MB	Document
SRR21703524.2.fastq	7 Mar 2023 at 14:41	12.3 MB	Document
SRR21703525.1.fastq	7 Mar 2023 at 14:41	12 MB	Document
SRR21703525.2.fastq	7 Mar 2023 at 14:41	12 MB	Document
SRR21703526.1.fastq	7 Mar 2023 at 14:41	16.7 MB	Document
SRR21703526.2.fastq	7 Mar 2023 at 14:41	16.7 MB	Document
SRR21703527.1.fastq	7 Mar 2023 at 14:41	22.5 MB	Document
SRR21703527.2.fastq	7 Mar 2023 at 14:41	22.5 MB	Document
SRR21703528.1.fastq	7 Mar 2023 at 14:41	20.6 MB	Document
SRR21703528.2.fastq	7 Mar 2023 at 14:41	20.6 MB	Document
SRR21703529.1.fastq	7 Mar 2023 at 14:41	49.1 MB	Document
SRR21703529.2.fastq	7 Mar 2023 at 14:41	49 MB	Document
SRR21703530.1.fastq	7 Mar 2023 at 14:41	13.8 MB	Document
SRR21703530.2.fastq	7 Mar 2023 at 14:41	13.8 MB	Document
SRR21703535.1.fastq	7 Mar 2023 at 14:41	20.5 MB	Document
SRR21703535.2.fastq	7 Mar 2023 at 14:41	20.4 MB	Document
SRR21703544.1.fastq	7 Mar 2023 at 14:41	28.6 MB	Document
SRR21703544.2.fastq	7 Mar 2023 at 14:41	28.5 MB	Document
SRR21703545.1.fastq	7 Mar 2023 at 14:41	15.3 MB	Document
SRR21703545.2.fastq	7 Mar 2023 at 14:41	15.2 MB	Document
SRR21703546.1.fastq	7 Mar 2023 at 14:41	19.7 MB	Document
SRR21703546.2.fastq	7 Mar 2023 at 14:41	19.6 MB	Document
SRR21703547.1.fastq	7 Mar 2023 at 14:41	50.3 MB	Document
SRR21703547.2.fastq	7 Mar 2023 at 14:41	50.2 MB	Document
filtered	Today at 15:59	--	Folder
metadata	7 Mar 2023 at 14:48	--	Folder
SraRunTable.txt	2 Mar 2023 at 11:27	5 KB	text
SRR_Acc_Lib.txt	2 Mar 2023 at 11:27	143 bytes	text

Analisi di dati metatranscriptomici. Un flusso di lavoro in R

17

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI SICUREZZA E RESILIENZA

Workflow, parte 1 organizzare i metadati

E' sempre bene documentare tutto quello che si fa!

Metadati importanti sono

- eventuali accession numbers
- il target
- la pubblicazione (se lo studio è pubblicato)
- la piattaforma e il tipo di sequenze
- i primer

```
# creating information for the study and sample data frames
Study <- "SRP399517"
target <- "16S RNA gene"
region <- "V3-V4"
seq_accn <- Study
DOI <- "10.3389/fmicb.2022.1101515"

# information on the platform and arrangement
platform <- "Illumina" # (or set to "Illumina" or "Ion_Torrent" or "F454")
paired_end <- T # set to true for paired end, false for single end
if (!paired_end) overlapping <- T # needed to run species assignment for SILVA

# expected amplicon length ? including primers
primer_f <- "Bakt_341F"
primer_r <- "Bakt_805R"
target1 <- "16S_DNA"
target2 <- region
```

Analisi di dati metatranscriptomici. Un flusso di lavoro in R

18

Workflow, parte 2 organizzare i file e i nomi dei campioni

La coerenza fra i nomi dei campioni nel file dei metadati e i nomi delle sequenze è essenziale. In genere è necessario automatizzare le seguenti operazioni:

- estrazione dei nomi dei campioni
- creazione dei vettori dei nomi dei file delle sequenze (forward e reverse, se disponibili entrambe)

```
# sub-directory data must already be in the wd

fastq_path <- file.path("data", "fastq")
filt_path <- file.path("data", "filtered")

# if(!file_test("-d", fastq_path)) {
#   dir.create(fastq_path)
# only needed if you want to create the directory and move the data from somewhere else
file_list <- list.files(fastq_path)
# filenames for forward sequences, sorted
fns <- sort(list.files(fastq_path, full.names = TRUE))
# this assigns the file names (with paths) for forward and reverse to two vectors
# may need adaptations if coding of the filenames is different
if(paired_end){
  fnFs <- fns[grepl("_1", fns)]
  fnRs <- fns[grepl("_2", fns)]
} else {
  fnFs <- fns
}
```

Analisi di dati metatassonomici. Un flusso di lavoro in R

19

Workflow, parte 3 rimuovere i primer

I primer contengono basi degenerate che impediscono l'assegnazione corretta della tassonomia a livello di genere e, quando possibile, di genere. Per rimuovere i primer si può

- chiedere al fornitore di servizi di farlo per noi (insieme alla rimozione degli adattatori e indici)
- usare cutadapt dal terminale (vedi QRcode)
- operare manualmente il fase di filtraggio dopo aver ispezionato le sequenze

```
if(keep_time) tic("\nReading sequences")
# check for occurrence of primers and adapters on a sample of forward and
# reverse sequences
# get a sample of 6 sequences
sampleFs <- if(length(fnFs)>=6) sample(fnFs,6) else fnFs[1:length(fnFs)]
# works with .fasta and fastq.gz
myFwsample <- ShortRead::readFastq(sampleFs)
head(sread(myFwsample),10)
tail(sread(myFwsample),10)
ave_seq_length_f <- round(mean(sread(myFwsample)@ranges@width))
ave_seq_length <- ave_seq_length_f


cat("\naverage sequence length for forward sequences is", ave_seq_length_f, "bp\n")

# same for reverse
if(paired_end){
  sampleRs <- if(length(fnRs)>=6) sample(fnRs,6) else fnRs[1:length(fnRs)]
  myRvsample <- ShortRead::readFastq(sampleRs)
  head(sread(myRvsample),10)
  tail(sread(myRvsample),10)
  ave_seq_length_r <- round(mean(sread(myRvsample)@ranges@width))
  cat("\naverage sequence length for reverse sequences is", ave_seq_length_f, "bp\n")
  ave_seq_length <- mean(c(ave_seq_length_f, ave_seq_length_r))
}
```




Analisi di dati metatassonomici. Un flusso di lavoro in R


20




Finanziato dall'Unione europea
NextGenerationEU



Ministero dell'Università e della Ricerca



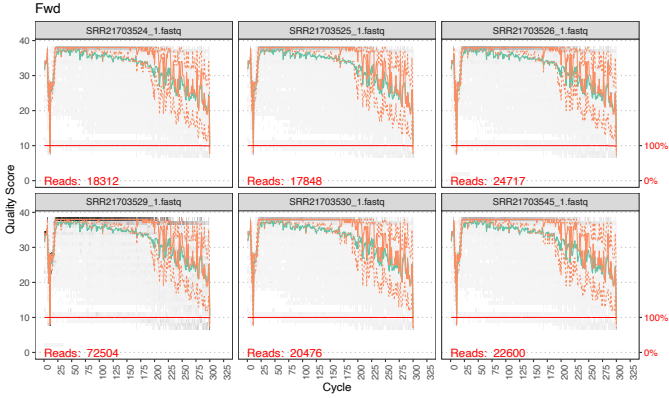
Italiadomani
PIANO NAZIONALE DI SICUREZZA E RESILIENZA



SUS-MIRRI.IT


Workflow, parte 4 quality profile plot

- 6 campioni estratti casualmente (forse sarebbe meglio farne di più)
- la regione di bassa qualità iniziale può dipendere dalle basi degenerate del primer
 - la **scala grigia** rappresenta il numero di sequenze di una data qualità in una data posizione
 - la **linea verde** è la qualità media, la **linea continua arancione** è la qualità mediana,
 - le **linee tratteggiate** sono il 25° e 75° percentile;
 - se le sequenze variano in lunghezza la linea continua rossa mostra la % di sequenze che si estendono ad una certa lunghezza




Analisi di dati metatranscriptomici. Un flusso di lavoro in R


21



Finanziato dall'Unione europea
NextGenerationEU



Ministero dell'Università e della Ricerca



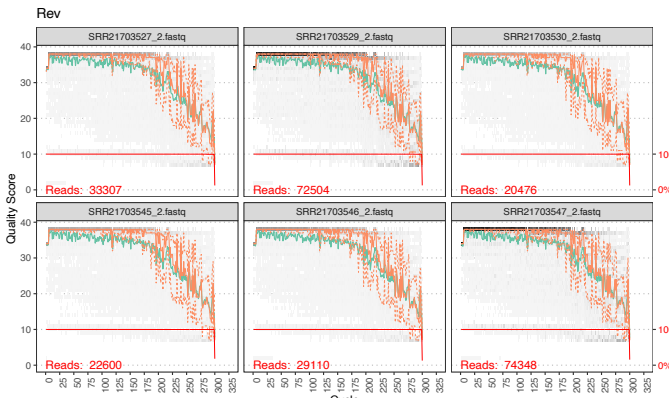
Italiadomani
PIANO NAZIONALE DI SICUREZZA E RESILIENZA



SUS-MIRRI.IT

Workflow, parte 4 quality profile plot

- 6 campioni estratti casualmente (forse sarebbe meglio farne di più)
- la regione di bassa qualità iniziale può dipendere dalle basi degenerate del primer
- le sequenze reverse sono tipicamente di qualità peggiore



Analisi di dati metatranscriptomici. Un flusso di lavoro in R

22

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Evidenze di tampering: qui sono state caricate sequenze pre-processate, non sequenze grezze

Fwd

Rev

Analisi di dati metatranscriptomici. Un flusso di lavoro in R

23

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Workflow, parte 5 controllo qualità e filtri

è una **fase delicata** che serve a:

- rimuovere sequenze iniziali e finali (primer, basi di cattiva qualità)
- rimuovere sequenze di cattiva qualità

La funzione usata da DADA2 ha molti parametri aggiustabili ed occorre essere pragmatici, per lasciare il maggior numero di sequenze per l'analisi senza lasciare troppe sequenze di cattiva qualità (che potrebbero causare problemi in molte fasi successive). Per paired end è inoltre importante che le sequenze residue abbiano un buon overlap (almeno 25 bp)

```

truncf-- 250
truncr-- 250 # NULL if not paired end
trim_left = c(17,21) # use a length 2 vector c(x,y) if paired end or a single number if not
if(platform == "Ion_Torrent") trim_left <- trim_left-15
maxEEF = 2 # with very high quality data can be reduced to 1
maxEEr = 5 # 2 very restrictive, 5 does well in most cases, not needed for single end
trunc_q = 2 # with very high quality data can be increased up to 10-11
max_length <- 999 # not needed for Illumina and Ion Torrent, modify to max. exp. seq. length for F454
filter_and_trim_par <- as.data.frame(cbind(truncf, truncr, trim_left,
                                           maxEEF, maxEEr, trunc_q, max_length))

# matchIDs = true if prefiltered in QIIME;
# paired end

out <- if(paired_end) {
  filterAndTrim(fnFs, filtFs, rev = fnRs, filt.rev = filtRs,
               truncQ=trunc_q,
               truncLen=c(truncf, truncr),
               trimLeft = trim_left,
               maxN=0, maxEE=c(maxEEF,maxEEr),
               rm.phix=TRUE,
               compress=TRUE,
               multithread=TRUE)
  # On Windows set multithread=FALSE
} else {
  # not paired end
  max_length <- ifelse(max_length>truncf, Inf, max_length)
  out <- filterAndTrim(fnFs, filtFs, rev = NULL, filt.rev = NULL,
                       truncQ=trunc_q,
                       truncLen=c(truncf,
                                   truncr),
                       trimLeft = trim_left,
                       maxLen = max_length,
                       maxN=0, maxEE=maxEEF,
                       rm.phix=TRUE,
                       compress=TRUE,
                       multithread=TRUE)
  # On Windows set multithread=FALSE
}
    
```

Analisi di dati metatranscriptomici. Un flusso di lavoro in R

24

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Workflow, parte 6

stima del modello degli errori e delle ASV

la stima degli errori avviene con DADA2 (Divisive Amplicon Denoising Algorithm) che permette poi l'inferenza di Amplicon Sequence Variants

si tratta di un processo che può essere pesante da un punto di vista computazionale

Il diagnostico principale è l'error plot; stimato il modello vengono inferite le ASV e se paired end si provvede a unire le due estremità

Analisi di dati metatranscriptomici. Un flusso di lavoro in R

25

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Workflow, parte 7

filtrare per lunghezza, rimuovere singleton, doubleton, chimere

dopo questa fase è possibile creare la tabella delle sequenze e:

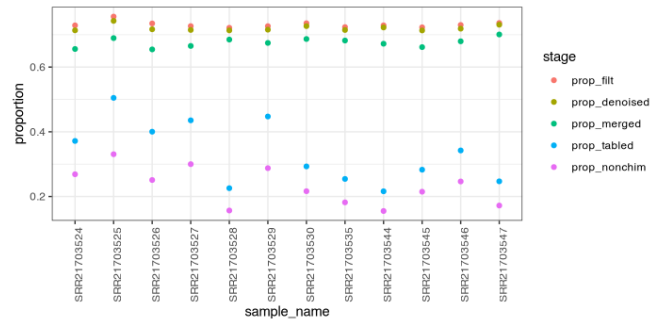
- rimuovere sequenze troppo più lunghe o più corte (spesso se. mitocondriali) della lunghezza attesa
- rimuovere singleton (spesso bimere) e doubleton (non strettamente necessario)
- rimuovere bimere (formate dall'unione esatta di due estremità presenti in 2 sequenze diverse)

Analisi di dati metatranscriptomici. Un flusso di lavoro in R

26

Workflow, parte 8 ma quante sequenze ho perso?

- ogni fase del processo può determinare una perdita di sequenze
- un grafico che tiene conto dei diversi passaggi può aiutare a prendere decisioni informate sulla necessità di variare i parametri in alcuni passaggi chiave
- qui la perdita importante è dovuta al filtro per dimensione (probabilmente per rimozione di sequenze di cloroplasti, abbondanti in questa matrice)



Analisi di dati metatassonomici. Un flusso di lavoro in R

27

Workflow, parte 10 Assegnare la tassonomia

- **assignTaxonomy()** usa il naïve Bayes classifier per assegnare prima il genere e poi, se possibile, la specie, sulla base della frequenza di ottameri
- **addSpecies()** assegna, se possibile, le specie sulla base del matching esatto delle sequenze con quelle di riferimento, se uniche (ma è possibile far produrre match multipli)
- per i batteri, il database di riferimento migliore è attualmente Silva v138.1
- è possibile usare **DECIPHER::idTaxa()**, più veloce e accurato, ma attualmente non è disponibile una versione formattata correttamente di Silva v138.1

```

598 # SILVA
599 ref_fasta <- file.path(taxdb_dir, "silva_nr99_v138.1_train_set.fa")
600 taxtab <- assignTaxonomy(seqtab.nochim, refFasta = ref_fasta, multithread = TRUE)
601 # setting tryRC to T if there are too many sequences not identified at the phylum level
602 if(mean(is.na(taxtab[,2]))>0.2) {
603   RC<-T
604   taxtab <- assignTaxonomy(seqtab.nochim, refFasta = ref_fasta, multithread = TRUE, tryRC =
605   }
606
607 # optionally, if the sequences do not get identified, add the option tryRC=T
608 # which also tries the reverse complement
609 if(keep_time) toc()
610
611 # do species assignment, DOES NOT WORK WITH JUST CONCATENATE
612
613 if(!paired_end | overlapping){
614   if(keep_time) tic("nassign taxonomy, species")
615   sp_ass_SILVA <- file.path(taxdb_dir, "silva_species_assignment_v138.1.fa")
616   taxtab <- addSpecies(taxtab, sp_ass_SILVA, tryRC = RC)
617   # optionally, if the sequences do not get identified, add the option tryRC=T
618   # which also tries the reverse complement
619   if(keep_time) toc()
620 }

```

Analisi di dati metatassonomici. Un flusso di lavoro in R

28

Workflow, parte 11 Costruire l'albero filogenetico

L'albero filogenetico è necessario solo se si vogliono usare metodi di ordinamento e analisi basati sulla distanza filogenetica. La pipeline una combinazione di comandi da dada2, phangorn e DECIPHER per

- estrarre le sequenze
- allinearle
- creare una matrice di distanza
- creare un albero guida con NJ
- creare un albero filogenetico

```

seqs <-
dada2::getSequences(seqtab.nochim) # the collapse option is very interesting
names(seqs) <- seqs # This propagates to the tip labels of the tree
alignment <-
DECIPHER::AlignSeqs(DNAStringSet(seqs),
  anchor = NA,
  processors = nc)

# The phangorn R package is then used to construct a phylogenetic tree.
# Here we first construct a neighbor-joining tree, and then fit a GTR+G+I
# (Generalized time-reversible with Gamma rate variation)
# maximum likelihood tree using the neighbor-joining tree as a starting point.
# transform in phydot object
phang.align <- phangorn::phyDat(as(alignment, "matrix"), type = "DNA")
# create distance matrix
cat("creating distance matrix...","\n")
dm <- phangorn::dist.ml(phang.align)
# perform Neighbor joining
cat("creating tree...","\n")
treeNJ <- phangorn::NJ(dm) # Note, tip order != sequence order
# internal maximum likelihood for tree
cat("estimating internal ML for tree...","\n")
fit <- phangorn::pml(treeNJ, data = phang.align)
Sys.sleep(5)
fitGTR <- update(fit, k = 4, inv = 0.2)
Sys.sleep(5)
# this is the step taking the longest time
cat("Optimization, please be patient (with >1000 seqs you are better off doing this overnight)...",
fitGTR <- optim.pml(
  fitGTR,
  model = "GTR",
  optInv = TRUE,
  optGamma = TRUE,
  rearrangement = "stochastic",
  control = pml.control(trace = 0)
)
detach("package:phangorn", unload = TRUE)

```

Analisi di dati metatassonomici. Un flusso di lavoro in R

29

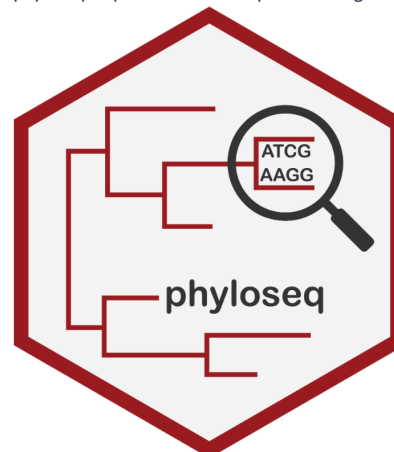
Workflow, parte 12 Creare un oggetto phyloseq

phyloseq è un eccellente pacchetto che fornisce funzioni per molte fasi dell'analisi metatassonomica:

- ha una specifica classe S4 di oggetti che mantiene ben organizzate tutte le tabelle
- fornisce funzioni per filtrare, estrarre, modificare i dati
- fornisce funzioni (spesso derivate da altri pacchetti) per condurre analisi di alpha e beta diversity

phyloseq [Installation](#) [Articles](#) [Courses](#) [Tutorials](#) [Issues](#)

phyloseq: Explore microbiome profiles using R



Analisi di dati metatassonomici. Un flusso di lavoro in R

30



Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



Extra: un flusso per big data

- uno dei problemi principali di R è che lavora usando la RAM
- la dimensione dei data set metatassonomici aumenta rapidamente e non è infrequente trovare studi con target V3-V4 di >100 campioni e V4 > 500 campioni, con >10⁵ sequenze per campione
- per set di grandi dimensioni il flusso di lavoro per big data modificato lavora in questo modo
 - analisi della tabella di sequenze per capire se sono state ottenute su macchine diverse
 - prima parte dell'analisi fino alla creazione della tabella di sequenze per sottoinsiemi di dati
 - seconda parte dell'analisi con creazione di una tabella di sequenze fusa e fasi successive dell'analisi (rimozione bimere, assegnazione della tassonomia, etc.)

Analisi di dati metatassonomici. Un flusso di lavoro in R

31



Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



Extra – analisi grafica e statistica di dati metatassonomici

- il numero di strumenti per l'analisi grafica e statistica di dati metatassonomici è immenso
- un buon articolo per orientarsi è Wen, T., Niu, G., Chen, T., Shen, Q., Yuan, J., Liu, Y.-X., 2023. The best practice for microbiome analysis using R. Protein Cell. <https://doi.org/10.1093/procel/pwad024>
- fra i pacchetti / risorse più completi
 - phyloseq (anche disponibile come app interattiva, Shiny phyloseq)
 - MicrobiomeAnalist (on line e come pacchetto R)
 - animalcules (app interattiva basata su R)
 - microeco

Analisi di dati metatassonomici. Un flusso di lavoro in R

32

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI SICUREZZA E RESILLENZA

Caratteristiche dei principali pacchetti di R per l'analisi metatassonomica

- alpha diversity
- beta diversity
- struttura delle comunità
- analisi differenziale
- identificazione biomarker
- alberi filogenetici
- network analysis
- predizione di funzioni

The best practice for microbiome analysis using R
Tao Wei^{1,2,†}, Gaoqing Nie^{1,†}, Tong Chen¹, Qiong Shen², Jun Yuan^{1*}, Yong-Xin Liu^{1*}

Analisi di dati metatassonomici. Un flusso di lavoro in R

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI SICUREZZA E RESILLENZA

Some rights reserved

This presentation was created by Eugenio Parente, 2017 and modified in 2018, 2019, 2020, 2022, 2023. With the exception of figures and tables taken from published articles/books (for which different copyrights apply), the material included in this presentation is covered by Creative Commons Public License “by-nc-sa” (<http://creativecommons.org/licenses/by-nc-sa/4.0/>). If you decide to use/modify this lecture please let me know.

Analisi di dati metatassonomici. Un flusso di lavoro in R