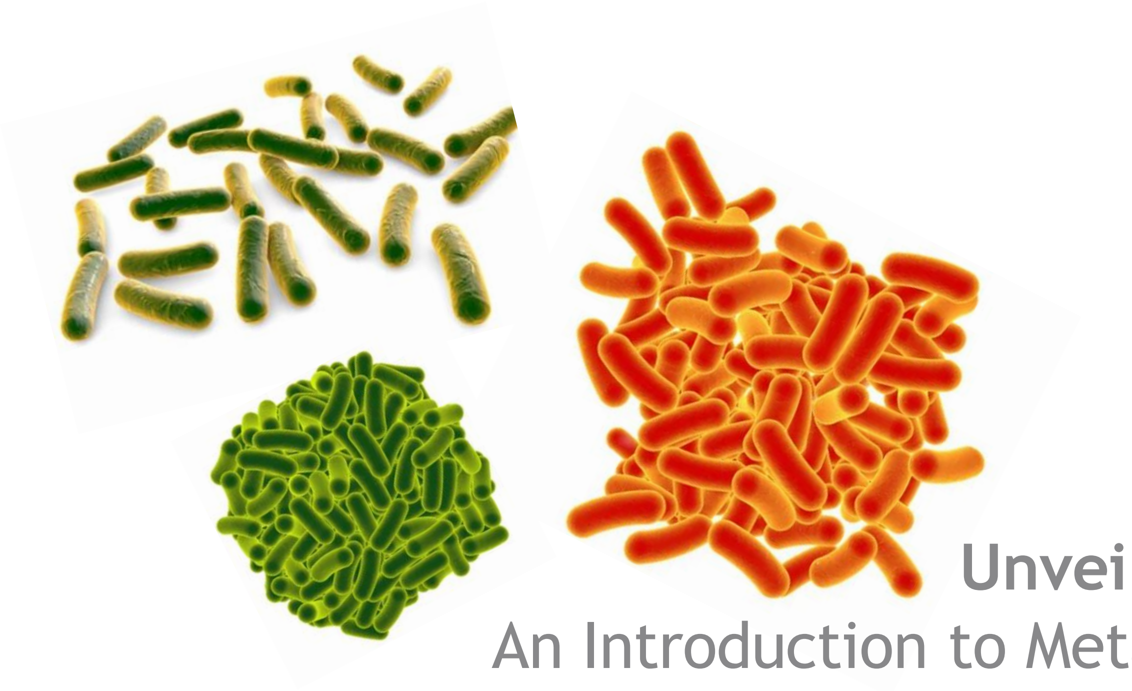
BIOSTATISTICS GROUP
WROCLAW, POLAND

THETA
Seminar



Unveiling the Mystery:
An Introduction to Metatranscriptome Analysis

MSc Eng. Krzysztof Kotlarz

Introduction - objectives

- **Opportunities and challenges of metatranscriptomics**
  - Understanding the capabilities of metatranscriptome
  - Learning important steps in data processing

- **Overview**
  - What is metatranscriptomics – how it is related to RNA-Seq
  - Experimental design
  - Processing of reads
    - Filtering
    - Assembly
    - Functional/taxonomical annotation
    - Statistical analysis
    - Visualisation

Introduction – Meta-genomics/transcriptomics

- 16S rRNA **(„Who is there”)**
  - Widely applied
  - Limited insight: cause or consequence

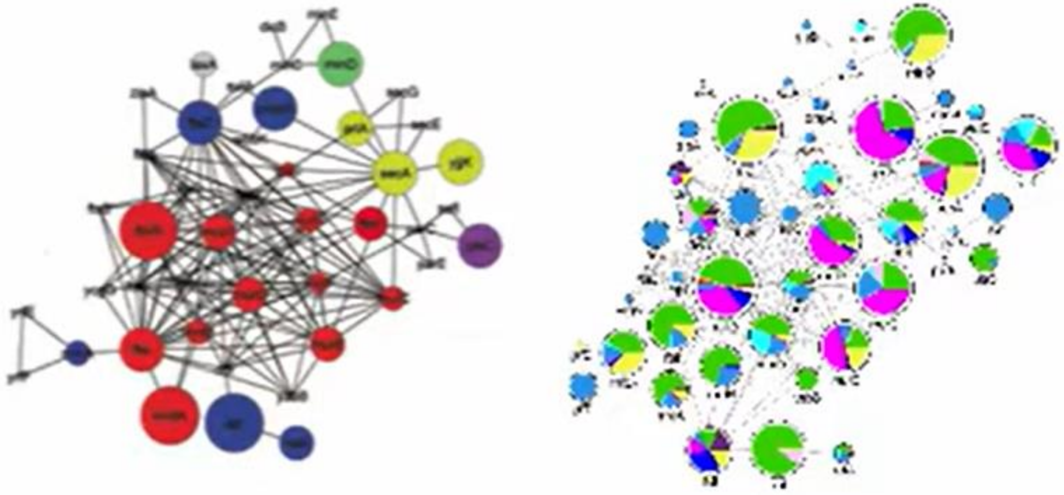
- Metagenomics **(„What can they do”)**
  - Differences in community composition but with conserved microbiome function

- Metatranscriptomics **(„Who is doing what”)**
  - Microbiome activity

Metatranscriptomics

- Focus on community activity
- Explore RNA-Seq to determine which genes and pathways are being actively expressed within the community

- **Reveal active functions**
- **Reveal which taxa are responsible for the active functions**



Relative abundance

Relative abundance and contributing taxa

Metatranscriptomics through RNA-Seq

- RNA-Seq is unbiased sequencing of RNA sample
  - Relative expression of transcripts within the sample
  - Typically applied to organisms with reference (sequenced) genome

Microbiome face a number of challenges

RNA-Seq

mRNA isolation

Fragmntation and sequencing

Mapping to the reference genome

Relative abundance of the transcript

Metatranscriptomics through RNA-Seq

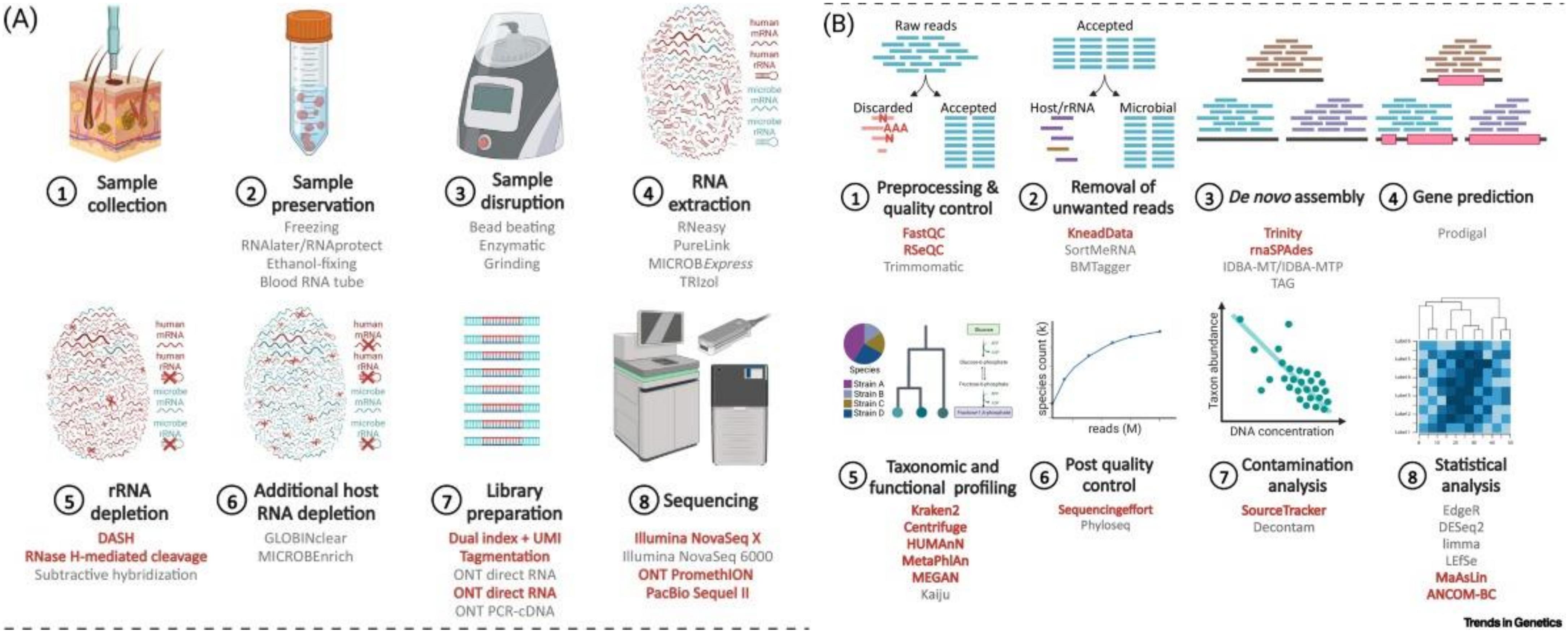
- For microbiome samples:
  - Lack of polyA signals make it difficult to isolate bacterial mRNA
    - Resulting in (massive) rRNA contamination
  - Environemntal microbiome samples
    - Lack of reference genomes make it difficult to map reads to their source

Lack of polyA

Massive rRNA contimination

Lack of reference genome

Metatranscriptomic pipeline – from scratch


(A)
① Sample collection
② Sample preservation
Freezing
RNAlater/RNAprotect
Ethanol-fixing
Blood RNA tube
③ Sample disruption
Bead beating
Enzymatic
Grinding
④ RNA extraction
RNeasy
PureLink
MICROB*Express*
TRIzol
human mRNA
human rRNA
microbe mRNA
microbe rRNA
⑤ rRNA depletion
DASH
RNase H-mediated cleavage
Subtractive hybridization
⑥ Additional host RNA depletion
GLOBINclear
MICROBEnrich
⑦ Library preparation
Dual index + UMI
Tagmentation
ONT direct RNA
ONT direct RNA
ONT PCR-cDNA
⑧ Sequencing
Illumina NovaSeq X
Illumina NovaSeq 6000
ONT PromethION
PacBio Sequel II
(B)
Raw reads
Discarded
Accepted
N
AAA
N
Accepted
Host/rRNA
Microbial
① Preprocessing & quality control
FastQC
RSeQC
Trimmomatic
② Removal of unwanted reads
KneadData
SortMeRNA
BMTagger
③ *De novo* assembly
Trinity
rnaSPAdes
IDBA-MT/IDBA-MTP
TAG
④ Gene prediction
Prodigal
Species
Strain A
Strain B
Strain C
Strain D
species count (k)
reads (M)
Taxon abundance
DNA concentration
⑤ Taxonomic and functional profiling
Kraken2
Centrifuge
HUMAnN
MetaPhlAn
MEGAN
Kaiju
⑥ Post quality control
Sequencingeffort
Phyloseq
⑦ Contamination analysis
SourceTracker
Decontam
⑧ Statistical analysis
EdgeR
DESeq2
limma
LEfSe
MaAsLin
ANCOM-BC
Trends in Genetics

Pipeline – step by step (General)

Sample collection

- Cost ~300$ -400$ (2016)
- Number of replicates
- Power of analisis challenge

Preparing sample for sequencing

- Bacterial mRNA – remove abundant rRNA
- Host mRNA removal (informative)

Generating reads

Pipeline – step by step (General)

- Robust software standards and methods still need to be developed
- Metatranscriptomics is a relatively new field

- Diffrent alighment approaches:
  - Function annotation rely on the identification of sequence similarities
  - E.g BWA rely on perfect matches to the reference genomes – sequence diversity is huge!
  - Sequence diversity >> protein diversity -> solution BLASTX (DIAMOND/USEARCH) to work on peptide workspace
    - Issues over the quality and costs

Preprocessing

- Adaptor / Low quality trimming
- rRNA identification and removal
- Dereplication
- Host mRNA identification / removal

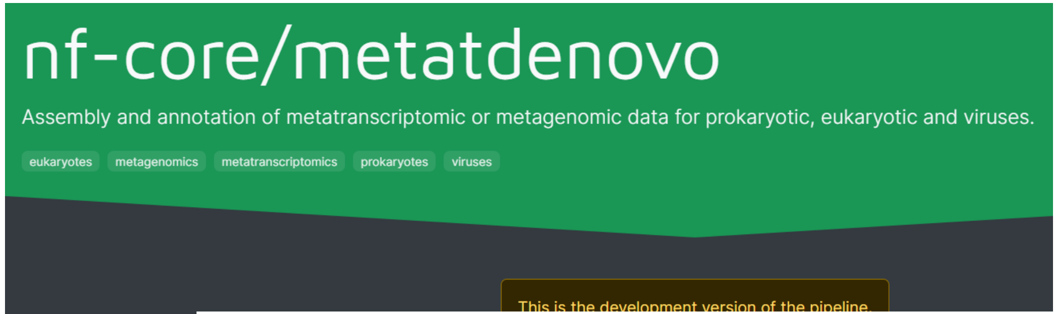
Assembly and annotation

- BWA & BLAT Searches
- BLASTX Searches
- De-novo assembly

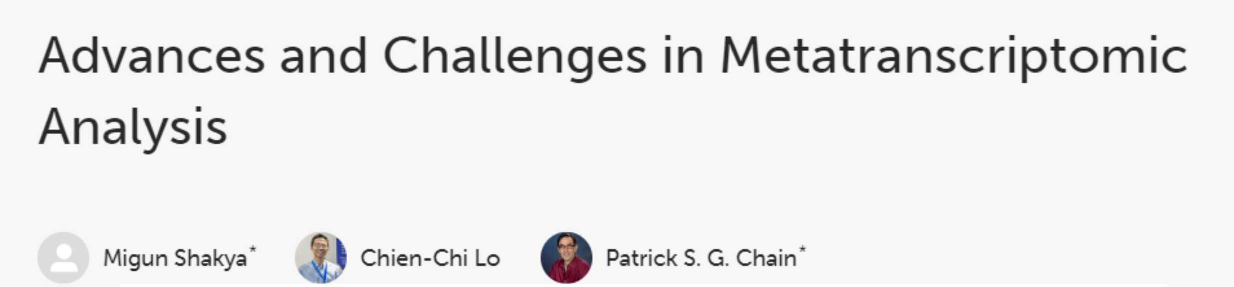
Analysis

- Taxonomy annotation
- Functional annotation
  - Differential expression
- Enzyme prediction
- Protein interaction data

As many papers…

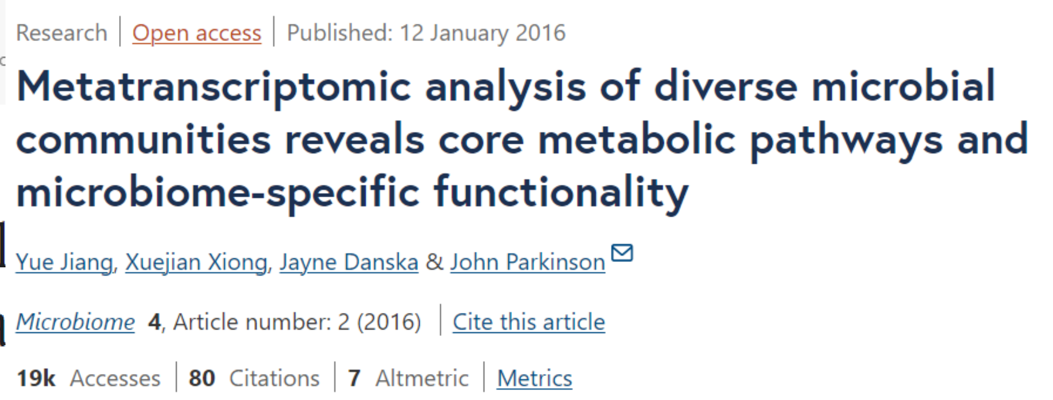

nf-core/metatdenovo

Assembly and annotation of metatranscriptomic or metagenomic data for prokaryotic, eukaryotic and viruses.


Advances and Challenges in Metatranscriptomic Analysis

Migun Shakya*, Chien-Chi Lo, Patrick S. G. Chain*

Bioscienc

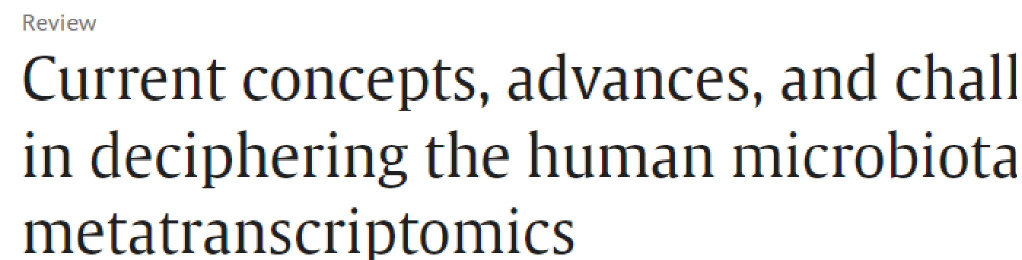

Research | Open access | Published: 12 January 2016

Metatranscriptomic analysis of diverse microbial communities reveals core metabolic pathways and microbiome-specific functionality

Yue Jiang, Xuejian Xiong, Jayne Danska & John Parkinson

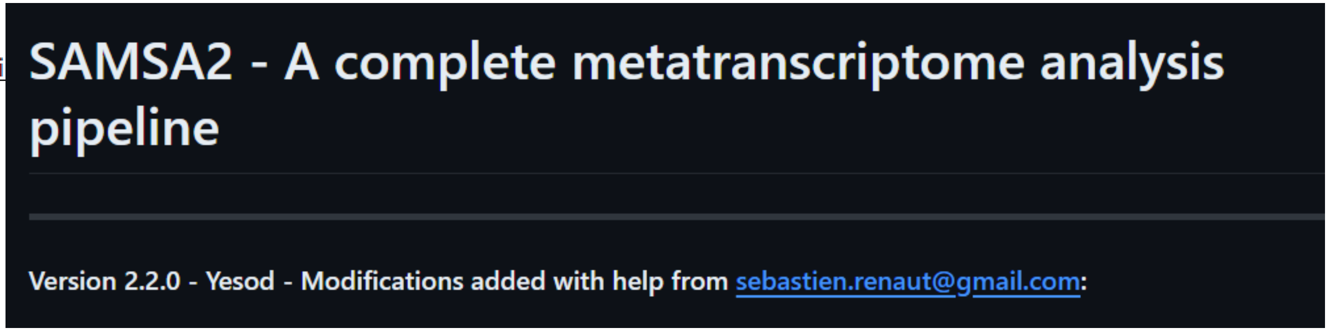
Microbiome 4, Article number: 2 (2016) | Cite this article

19k Accesses | 80 Citations | 7 Altmetric | Metrics

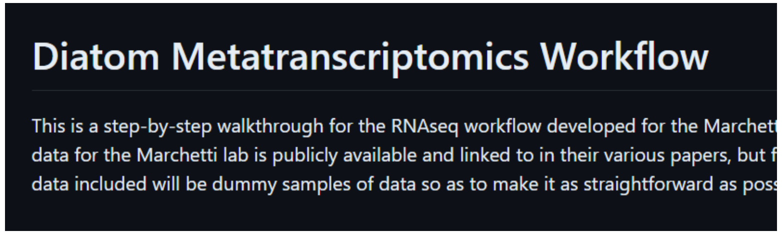

Review

Current concepts, advances, and chall in deciphering the human microbiota metatranscriptomics

Teija Ojala [1], Aino-Elina Häkkinen [2], Esko Kankuri [1], Matti Kankai


SAMSA2 - A complete metatranscriptome analysis pipeline

Version 2.2.0 - Yesod - Modifications added with help from sebastien.renaut@gmail.com:


Diatom Metatranscriptomics Workflow

This is a step-by-step walkthrough for the RNAseq workflow developed for the Marchett data for the Marchetti lab is publicly available and linked to in their various papers, but f data included will be dummy samples of data so as to make it as straightforward as poss

Pipeline – step by step (Suggested)

Preprocessing

- **Trimmomatic**: Adaptor / Low quality trimming

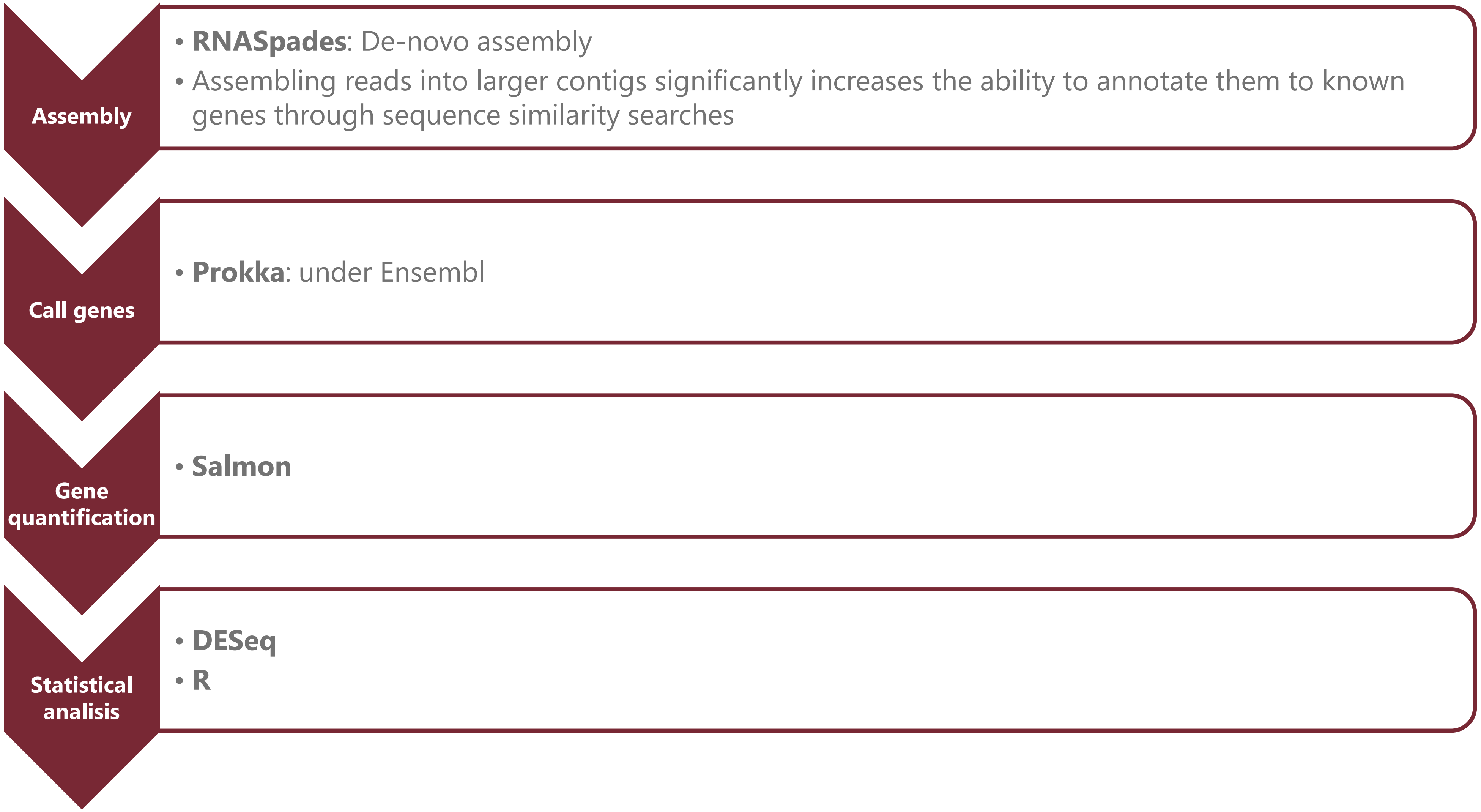
Merging overlapped reads

- **Pear**: Long reads are more likely to be processed in downastream analisis to increase the specificity among data with high variablity
- Long Illumina reads -> overalpped reads

rRNA identification and removal

- **SortMeRNA**: Host mRNA identification / removal

Pipeline – step by step (Suggested)



Assembly

- **RNASpades**: De-novo assembly
- Assembling reads into larger contigs significantly increases the ability to annotate them to known genes through sequence similarity searches

Call genes

- **Prokka**: under Ensembl

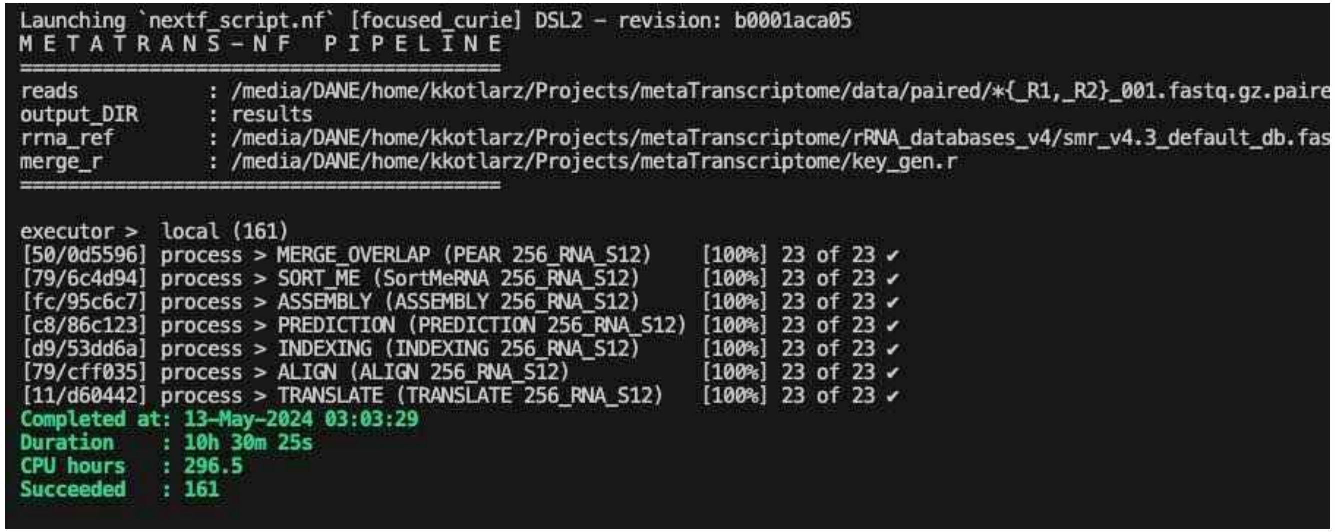
Gene quantification

- **Salmon**

Statistical analisis

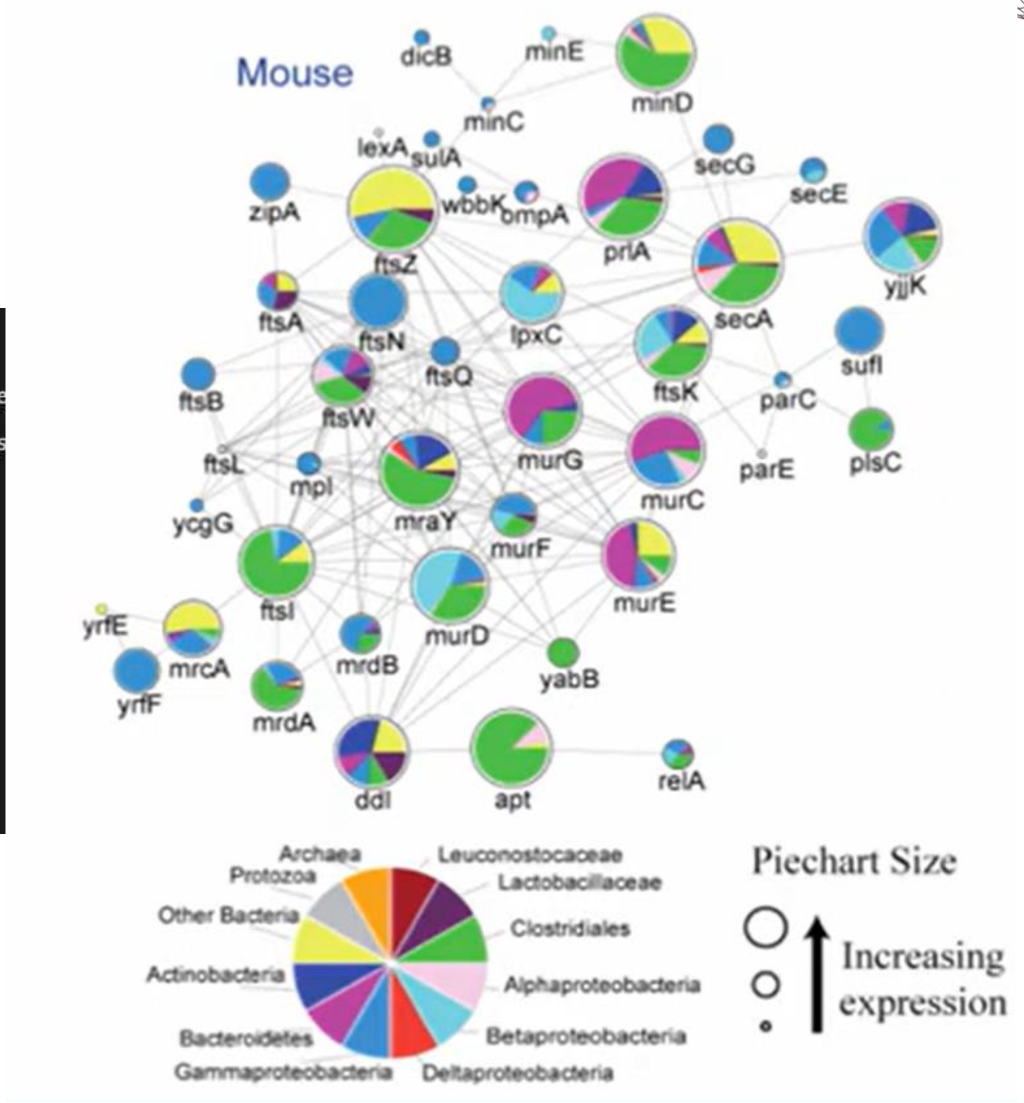
- **DESeq**
- **R**

Current and (expected) results



```
Launching `nextf_script.nf` [focused_curie] DSL2 - revision: b0001aca05
M E T A T R A N S - N F   P I P E L I N E
========================================
reads          : /media/DANE/home/kkotlarz/Projects/metaTranscriptome/data/paired/*{_R1,_R2}_001.fastq.gz.paire
output_DIR     : results
rrna_ref       : /media/DANE/home/kkotlarz/Projects/metaTranscriptome/rRNA_databases_v4/smr_v4.3_default_db.fas
merge_r        : /media/DANE/home/kkotlarz/Projects/metaTranscriptome/key_gen.r
========================================

executor >  local (161)
[50/0d5596] process > MERGE_OVERLAP (PEAR 256_RNA_S12)    [100%] 23 of 23 ✔
[79/6c4d94] process > SORT_ME (SortMeRNA 256_RNA_S12)     [100%] 23 of 23 ✔
[fc/95c6c7] process > ASSEMBLY (ASSEMBLY 256_RNA_S12)     [100%] 23 of 23 ✔
[c8/86c123] process > PREDICTION (PREDICTION 256_RNA_S12) [100%] 23 of 23 ✔
[d9/53dd6a] process > INDEXING (INDEXING 256_RNA_S12)     [100%] 23 of 23 ✔
[79/cff035] process > ALIGN (ALIGN 256_RNA_S12)           [100%] 23 of 23 ✔
[11/d60442] process > TRANSLATE (TRANSLATE 256_RNA_S12)   [100%] 23 of 23 ✔
Completed at: 13-May-2024 03:03:29
Duration    : 10h 30m 25s
CPU hours   : 296.5
Succeeded   : 161
```


Mouse
dicB
minE
minD
minC
lexA
sulA
zipA
ftsZ
wbbK
ompA
prlA
secG
secE
yjjK
secA
ftsA
ftsN
lpxC
ftsQ
ftsK
sufI
ftsB
ftsW
murG
parC
ftsL
mpl
mraY
murC
parE
plsC
ycgG
murF
ftsI
murE
yrfE
mrcA
mrdB
murD
yrfF
mrdA
yabB
ddl
apt
relA
Archaea
Protozoa
Other Bacteria
Actinobacteria
Bacteroidetes
Gammaproteobacteria
Leuconostocaceae
Lactobacillaceae
Clostridiales
Alphaproteobacteria
Betaproteobacteria
Deltaproteobacteria
Piechart Size
Increasing expression

Thanks!