Metatranscriptomics

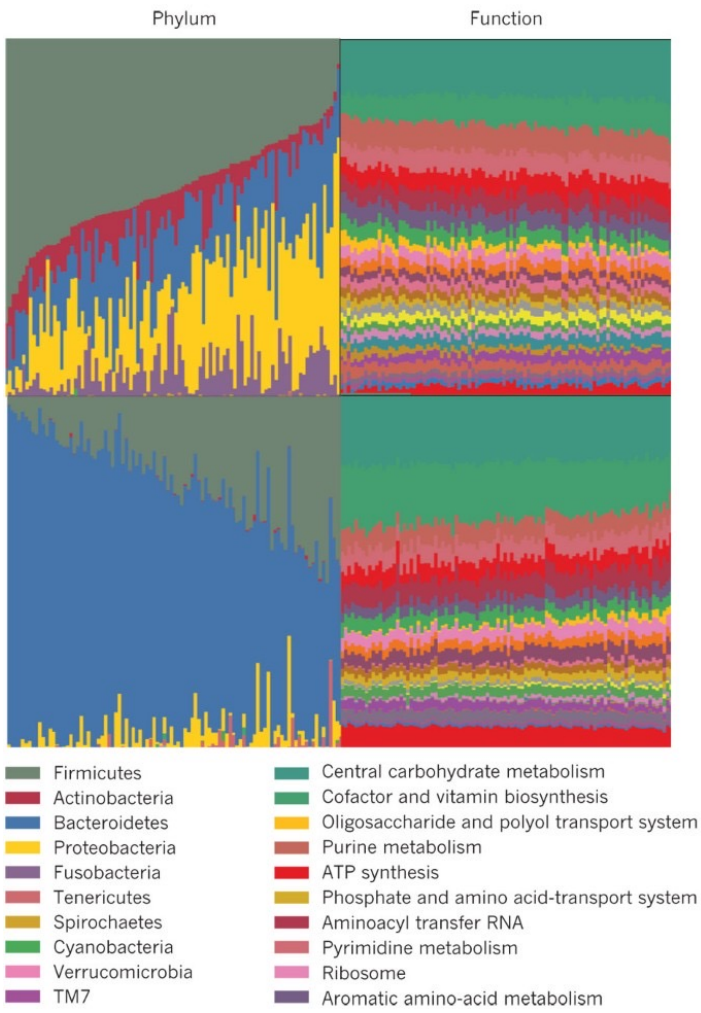
Dr. Natalia Zajac

Metagenomics, 03.2023

Overview

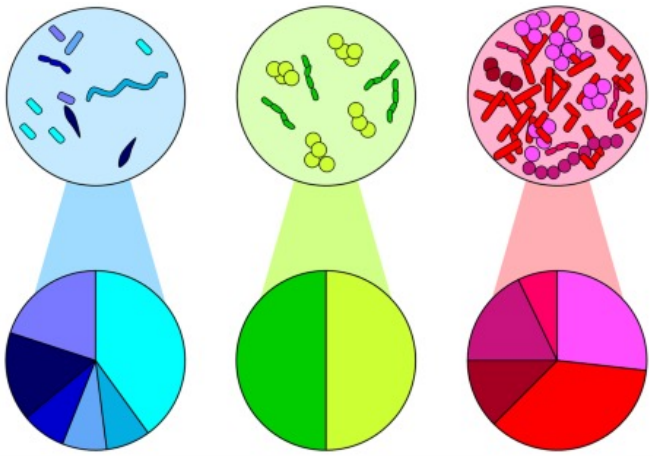
- What is metatranscriptomics - how does it relate to RNAseq and how does it differ from metagenomics?
- Processing of reads and statistical analysis
  - Quality filtering
  - Assembly/Mapping to databases
  - Functional and taxonomic annotation
- Visualization
- Research questions and case studies

What is metatranscriptomics?



https://www.nature.com/articles/nature11550

16S rRNA can tell us **WHO IS THERE?** But is what you see a cause or consequence?

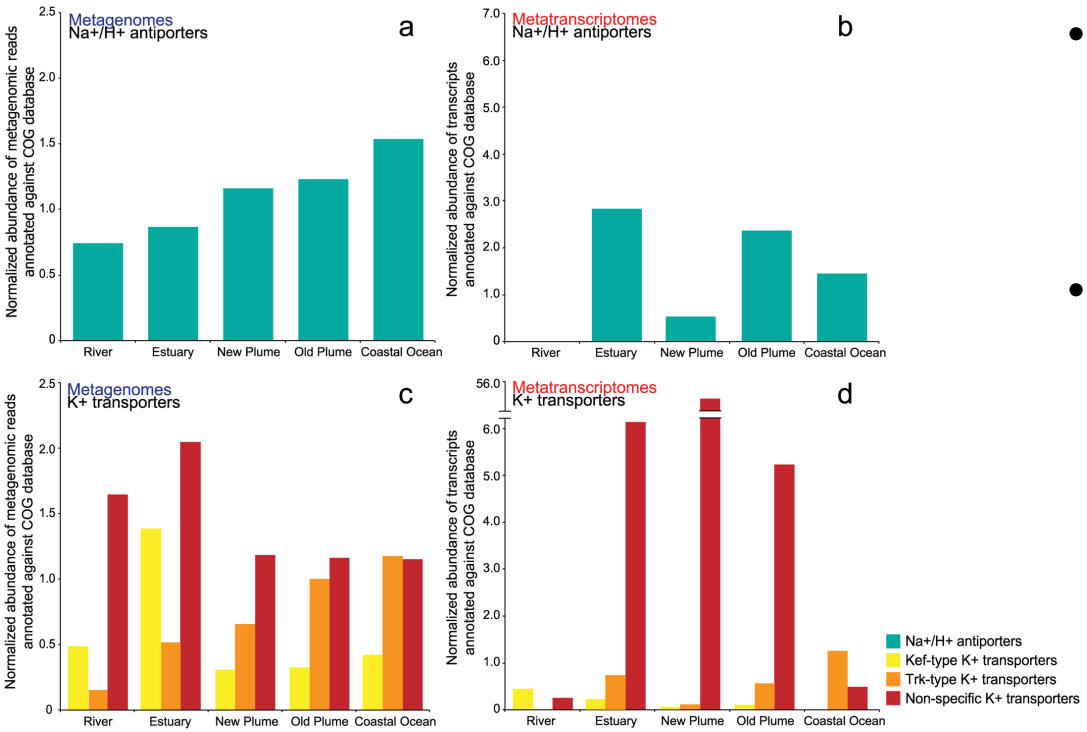


Claassen-Weitz et al. 2020

Metatranscriptomics can tell us **WHO IS DOING WHAT? – microbiome activity**

Shotgun metagenomics can tell us **WHO IS THERE? And WHAT IS THEIR FUNCTIONAL POTENTIAL?**

What is metatranscriptomics?



- FUNCTION: Which genes and pathways are actively expressed within a community
- TAXONOMY: Which taxa are responsible for the metabolic activity BUT NOT WHO IS THERE

Metabolic potential - highly similar across samples (with few differences in functional gene abundance from river to ocean)
Gene expression - highly variable and generally was independent of changes in salinity

Fortunato et al. 2015 PLOS ONE

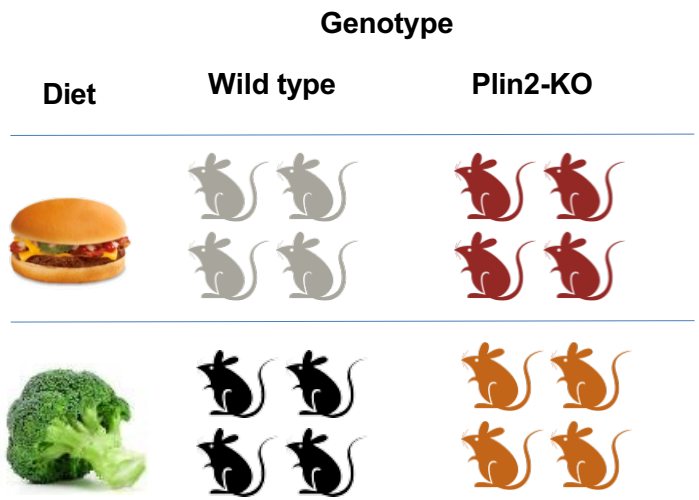
EXAMPLE

Research | Open Access | Published: 06 September 2017

Perilipin-2 modulates dietary fat-induced microbial global gene expression profiles in the mouse intestine

Xuejian Xiong, Elise S. Bales, Diana Ir, Charles E. Robertson, James L. McManaman, Daniel N. Frank & John Parkinson

Microbiome 5, Article number: 117 (2017) | Cite this article

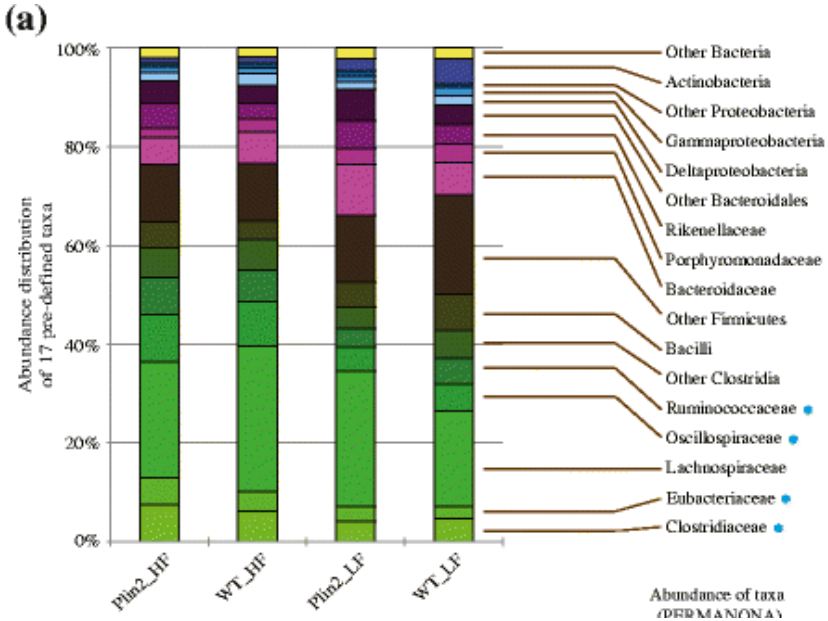


- Plin2 (Perilipin2) – involved in lipid uptake, interacts with lipid droplets
- Study comparing WT and Plin2-null mice - exposure to high-fat/low-carb (HF) or low-fat/high-carb (LF) diets

RNA-Seq
20-30 million sequence reads/mouse

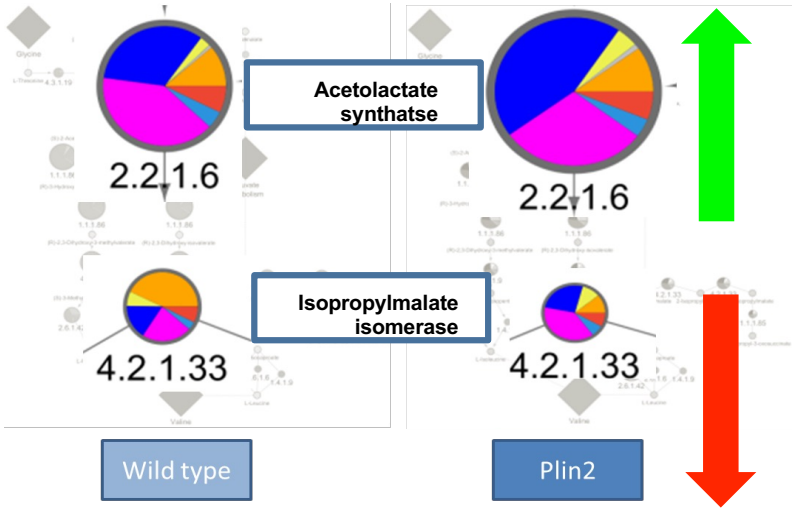
Xiong et al. 2017 BMC Microbiome

EXAMPLE



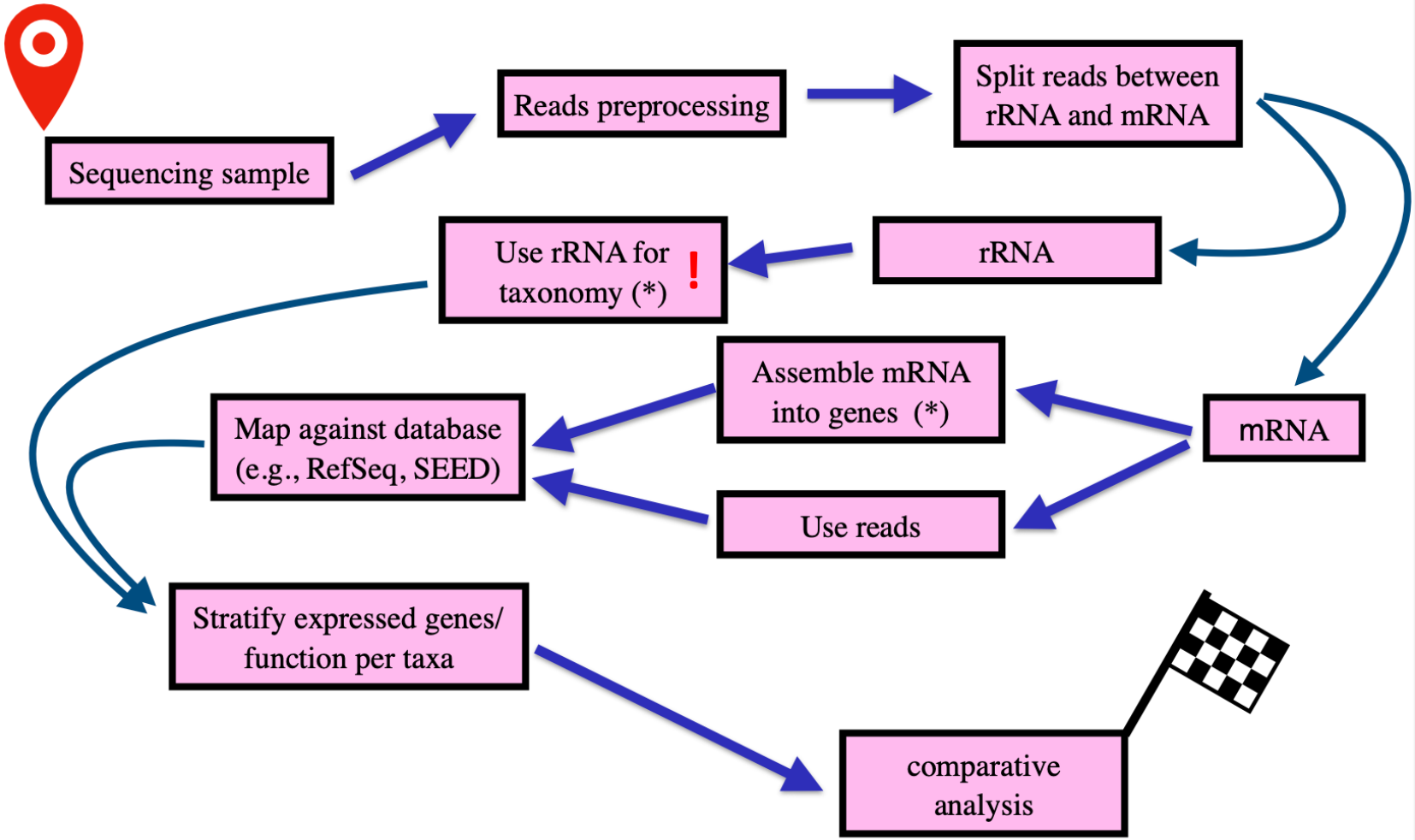
Perilipin2 KO mice fed a high fat diet exhibited a similar microbiome as WT mice fed the same diet….

…however metabolic pathways are differentially expressed – host genotype impacts microbiome function



Xiong et al. 2017 BMC Microbiome

WORKFLOW


Sequencing sample
Reads preprocessing
Split reads between rRNA and mRNA
Use rRNA for taxonomy (*)
!
rRNA
Assemble mRNA into genes (*)
mRNA
Map against database (e.g., RefSeq, SEED)
Use reads
Stratify expressed genes/ function per taxa
comparative analysis

CHALLENGES

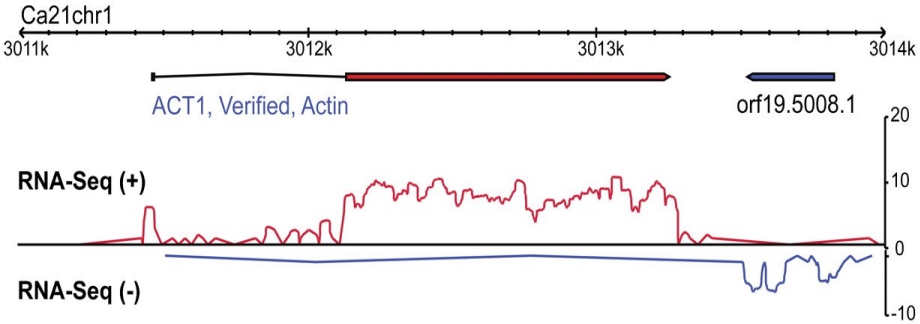
In a typical RNA-Seq experiment applied to a single eukaryotic organism, **mRNA** is **isolated**. After **fragmentation** and **sequencing**, reads are mapped to a reference genome using standard software such as **BWA and STAR** to provide: 1) support that the transcript is expressed; 2) the relative abundance of the transcript; and 3) the presence and abundance of isoforms

Resource

Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq

Vincent M. Bruno,[1] Zhong Wang,[2] Sadie L. Marjani,[3] Ghia M. Euskirchen,[4] Jeffrey Martin,[2] Gavin Sherlock,[4,5] and Michael Snyder[1,4,5]

[1]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA; [2]DOE Joint Genome Institute (JGI), Walnut Creek, California 94598, USA; [3]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA; [4]Department of Genetics, Stanford University Medical School, Stanford, California 94305-5120, USA



For microbiome samples we have the following problems:

a) **Lack of a polyA signal makes it difficult to isolate bacterial mRNA and resulting in (massive) rRNA contamination**

b) **Environmental microbiome samples lack reference genomes making it difficult to map reads back to their source transcripts**

c) **Host contamination**

CHALLENGES

RNA quality deteriorates rapidly – Method of storage and preparation can impact taxa recovered and has yet to be standardized

Best(?): Process immediately to extract RNA then store at -80
Next best(?): Snap freeze in liquid nitrogen and store at -80

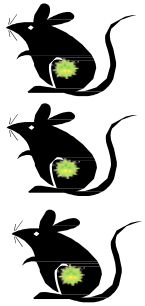
Avoid use of RNALater – it lyses some cells and can interfere with RNA extraction kits (e.g Ambion RiboPure Bacteria Kit / LifeTech Trizol Plus)

Standard Sequencing Recommendations

NovaSeq; 100bp single-end reads; 40M reads/sample
Packs of 200M reads

Costs for UZH and ETHZ research groups

Library preparation	190 CHF / sample
Sequencing 200M reads (5 samples)	1037 CHF
Bioinformatics (up to 15 samples, 3 comparisons)	880 CHF



Number of biological replicates

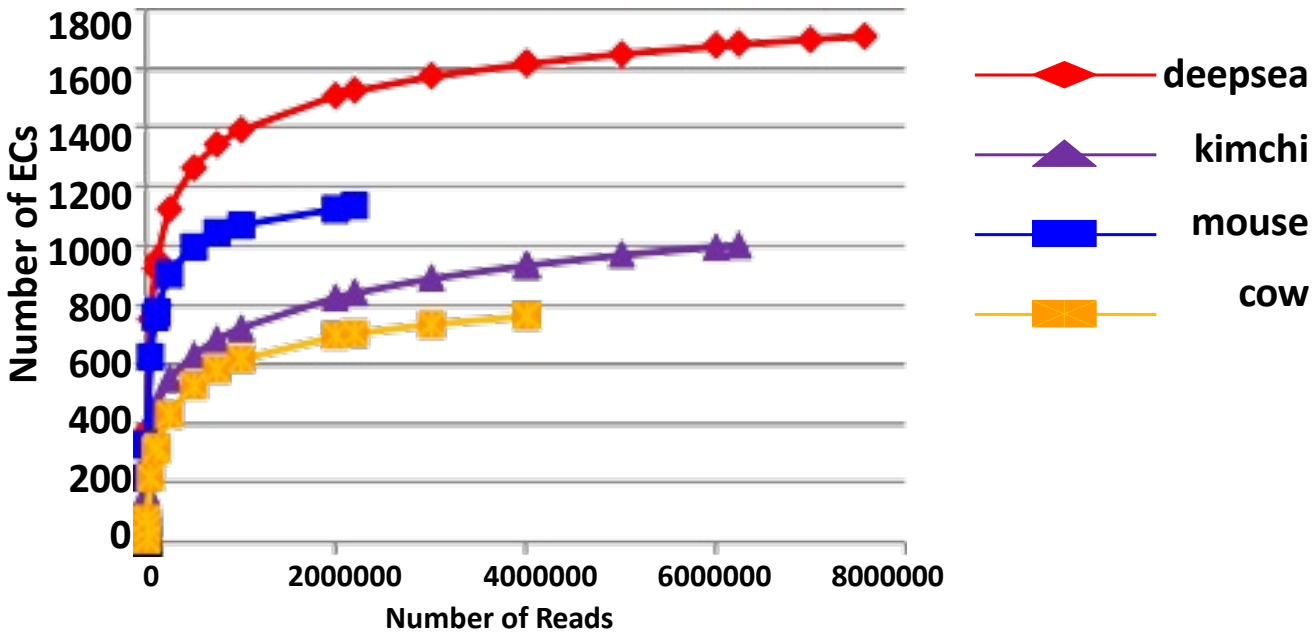
At least 2!

Power analyses challenging

https://bioinformatics.ca/

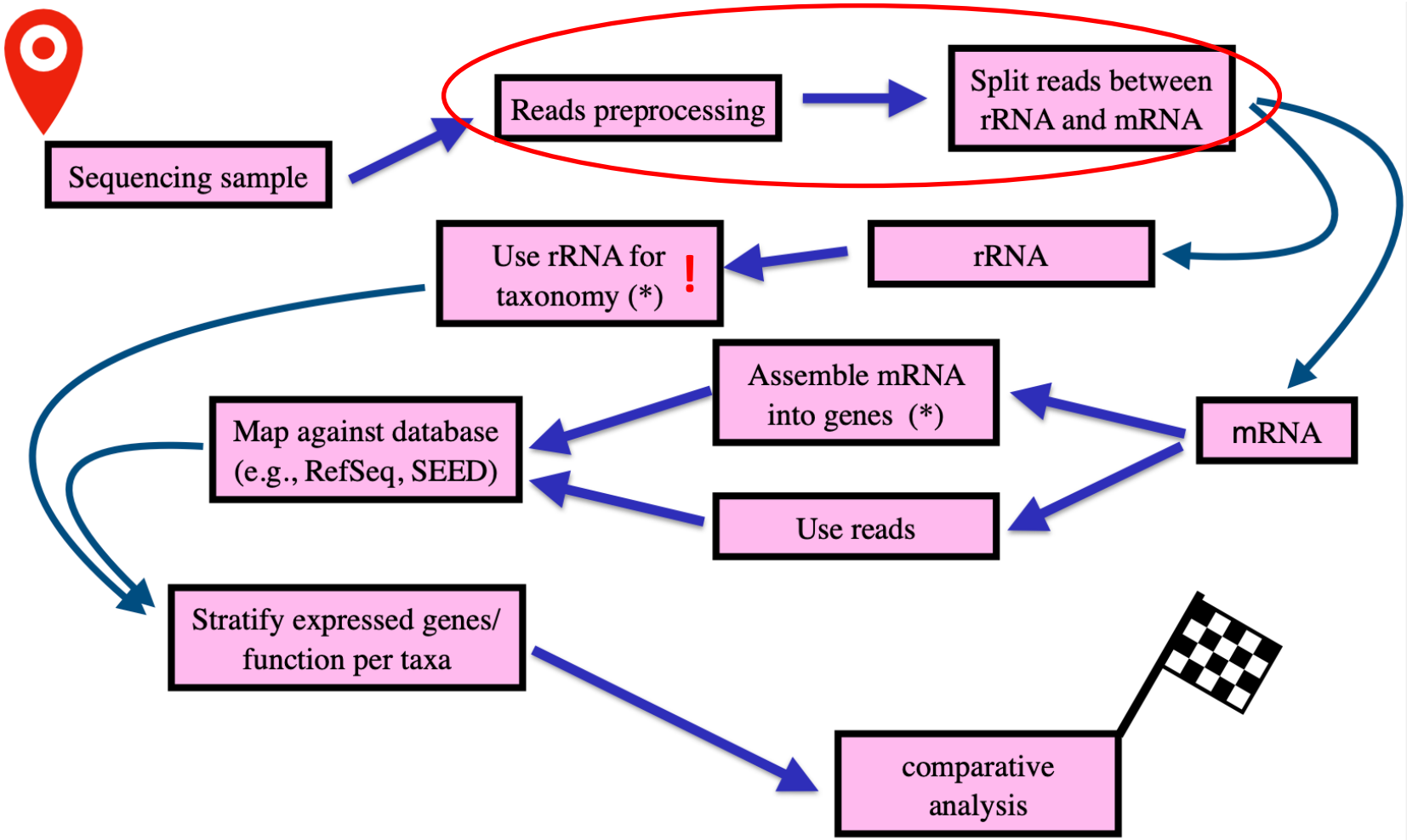
HOW MANY READS IS ENOUGH

https://bioinformatics.ca/



~5 million mRNA reads provide 90-95% of ECs in a microbiome
With kits yielding mRNA read rates of ~25%, this suggests 20 million/sample mRNA

WORKFLOW


Sequencing sample
Reads preprocessing
Split reads between rRNA and mRNA
Use rRNA for taxonomy (*)
!
rRNA
Assemble mRNA into genes (*)
mRNA
Map against database (e.g., RefSeq, SEED)
Use reads
Stratify expressed genes/ function per taxa
comparative analysis

Preprocessing

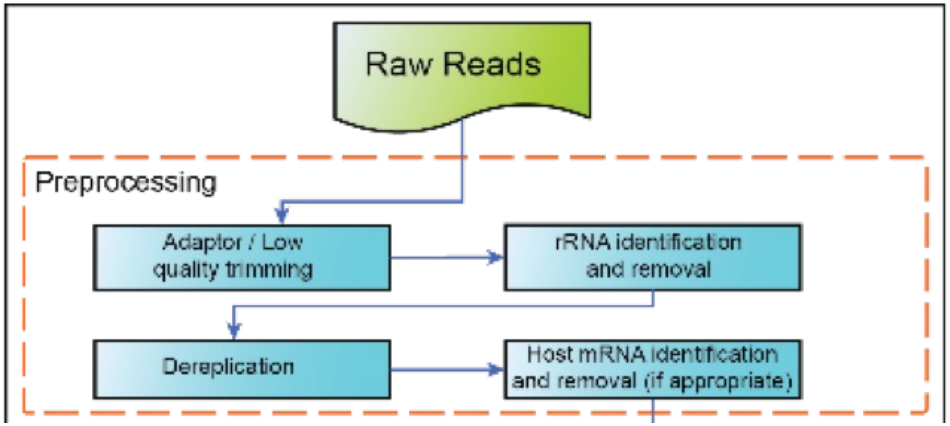
Trimmomatic - removal of:

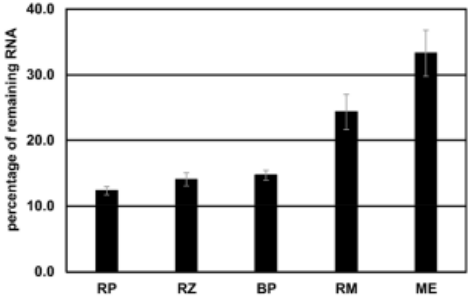
- low quality
- Adaptors

BWA/BLAT – alignment of reads to (and removal)

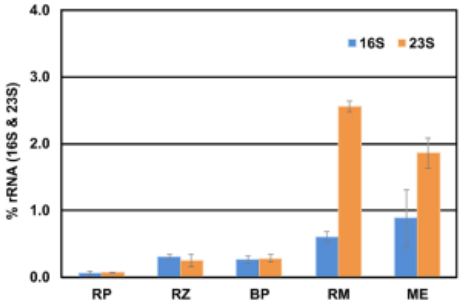
- Host transcriptome
- rRNA databases

SortMeRNA/Infernal – removal of rRNA





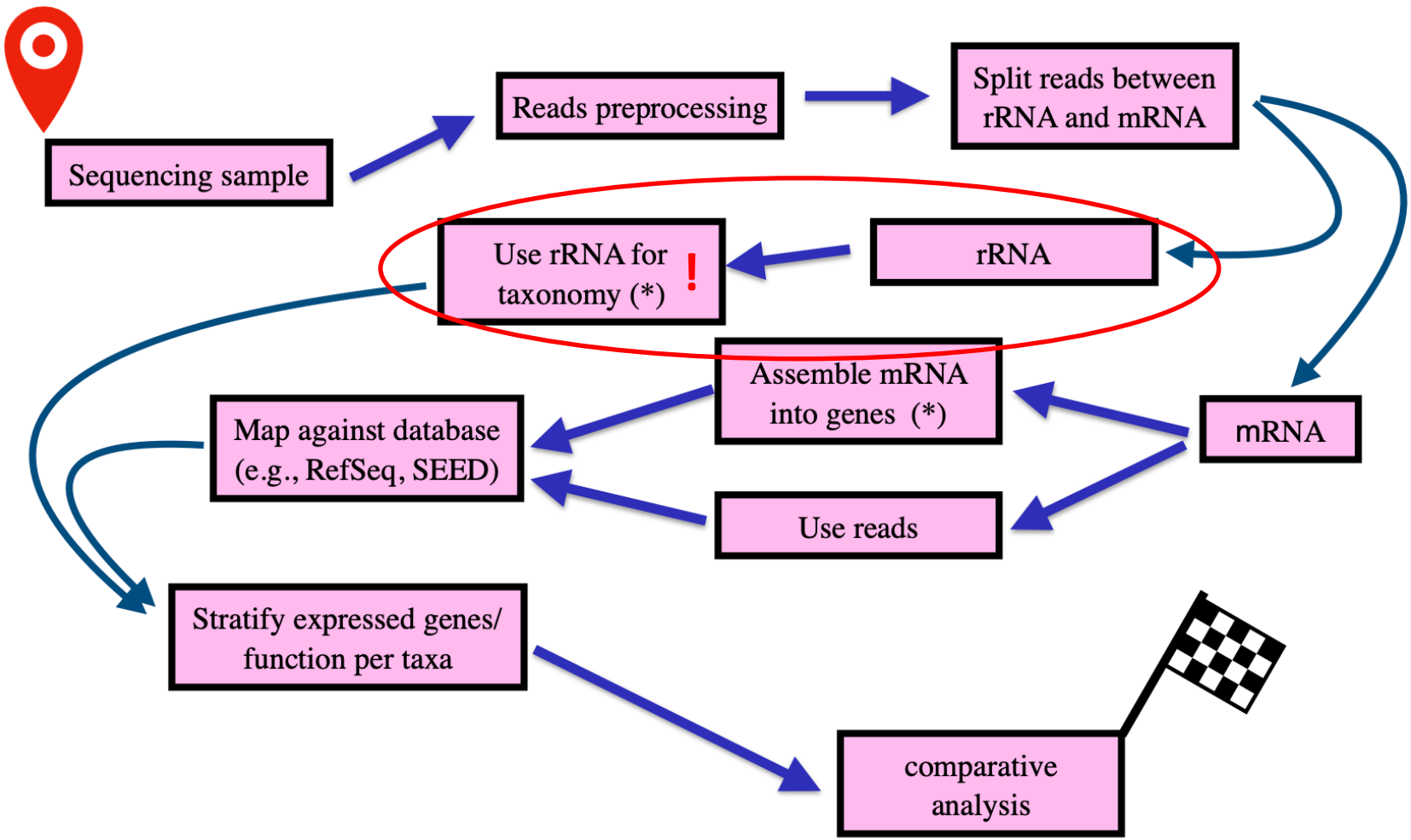
(A) Percentage of remaining RNA afer depletion using the diferent depletion kits compared. Te input of 2.5 µg total RNA was set to 100%.



(B) Remaining 16S and 23S rRNA afer depletion with the diferent kits compared to the untreated RNA determined by the Bioanalyzer electropherograms. RP, riboPOOLs; RZ, RiboZero; BP, biotinylated probes; RM, RiboMinus; ME, MICROBExpress.

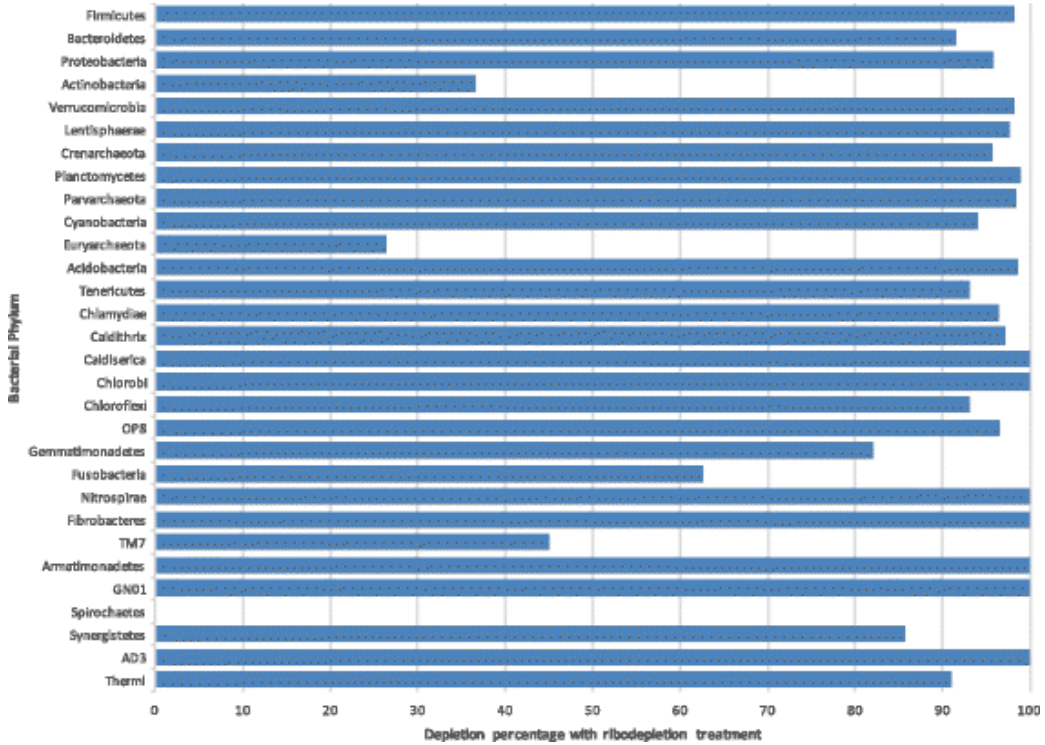
Wahl et al. 2022

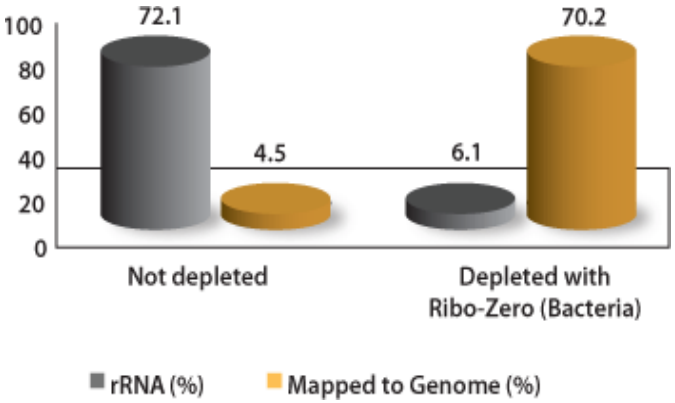
WORKFLOW


Sequencing sample
Reads preprocessing
Split reads between rRNA and mRNA
Use rRNA for taxonomy (*)
!
rRNA
Assemble mRNA into genes (*)
mRNA
Map against database (e.g., RefSeq, SEED)
Use reads
Stratify expressed genes/ function per taxa
comparative analysis

rRNA

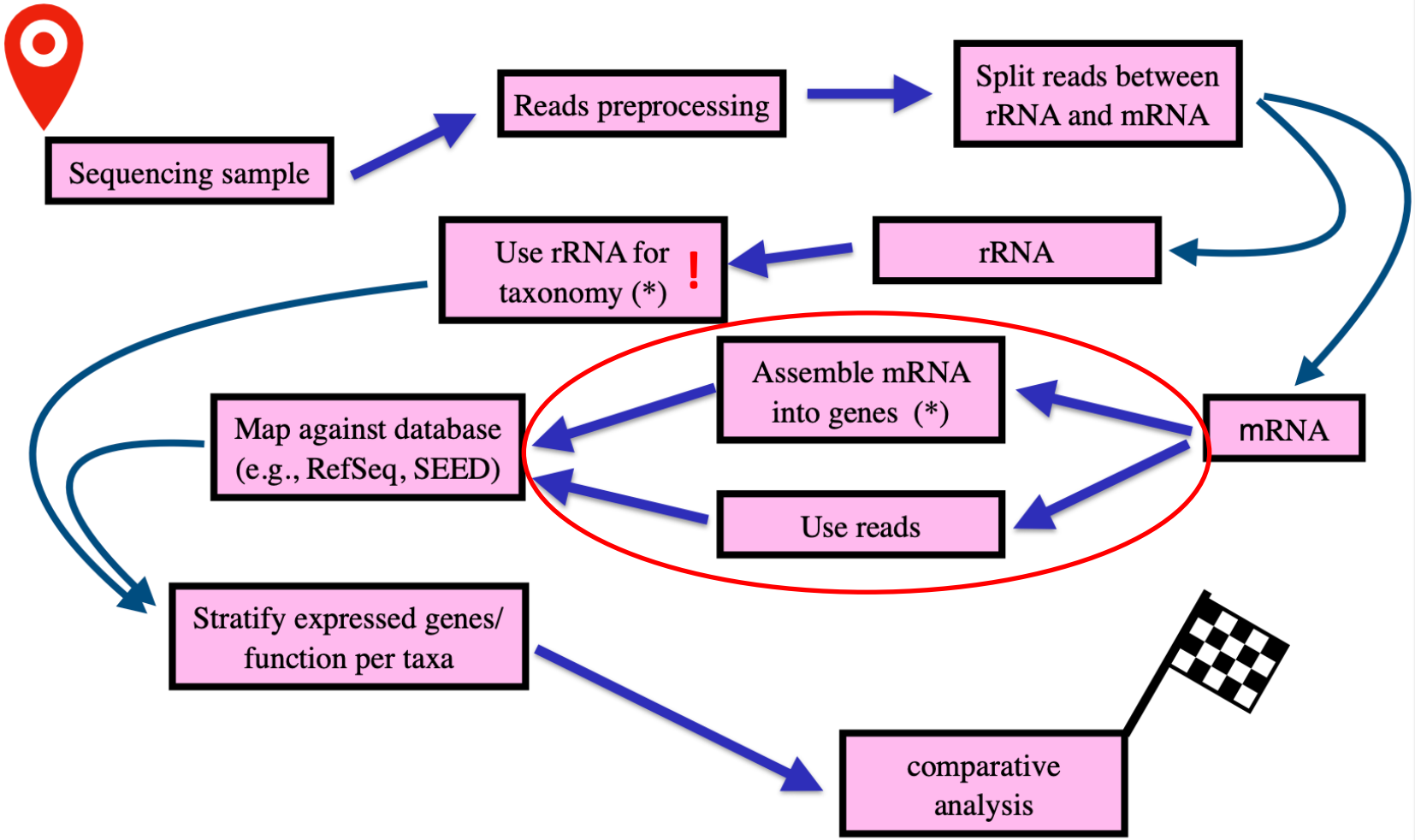
- Reads assigned to the rRNA database can be used for taxonomic classification
- BUT VERY OFTEN rRNA DEPLETION KITS ARE USED BEFORE EXTRACTION
- Not all species are equally depleted



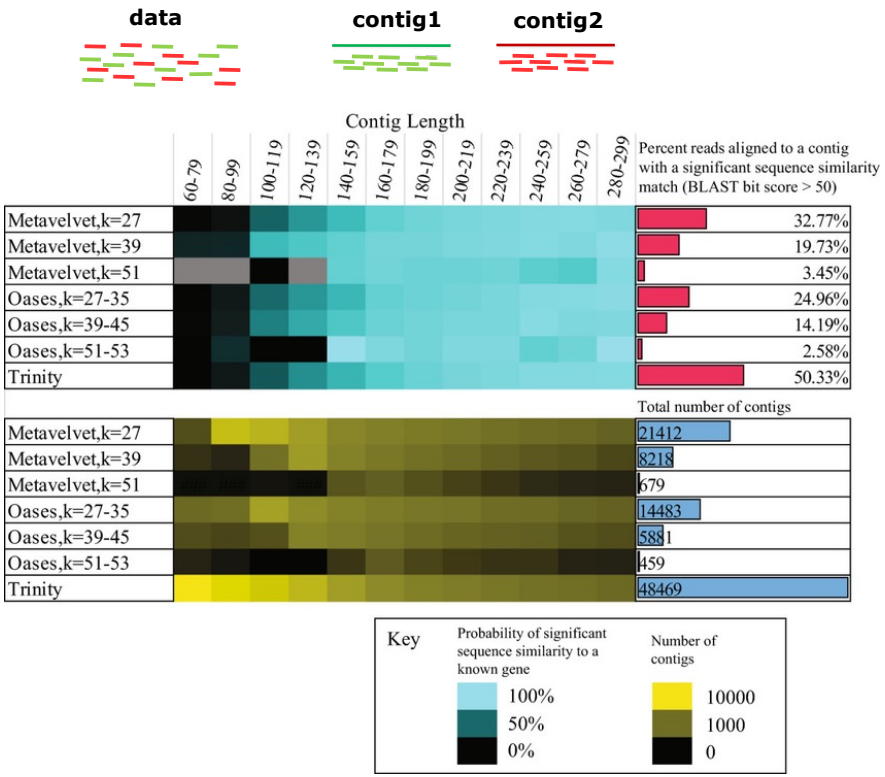


Westreich et al. BMC Bioinformatics 2016

WORKFLOW


Sequencing sample
Reads preprocessing
Split reads between rRNA and mRNA
Use rRNA for taxonomy (*)
!
rRNA
Assemble mRNA into genes (*)
mRNA
Map against database (e.g., RefSeq, SEED)
Use reads
Stratify expressed genes/ function per taxa
comparative analysis

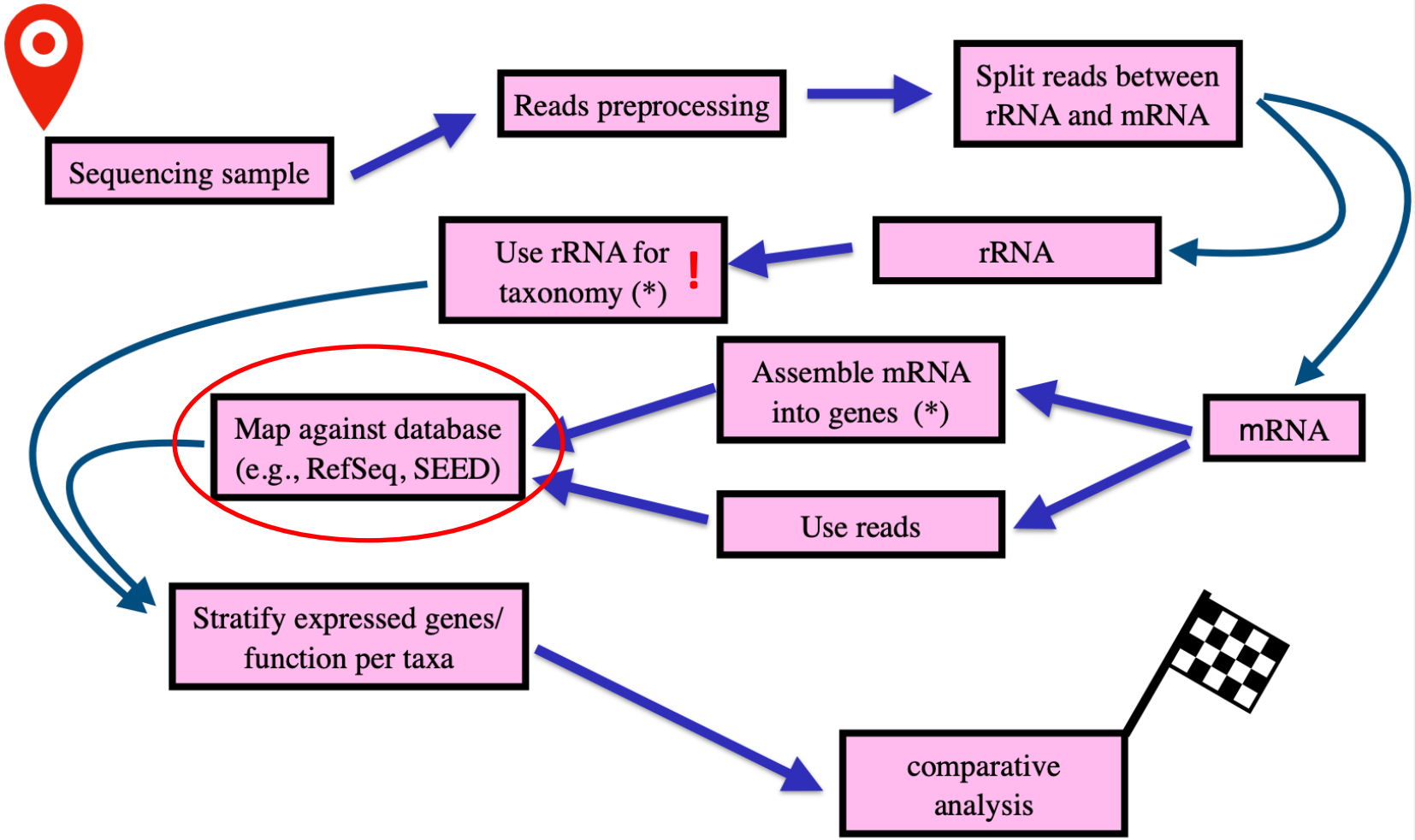
ASSEMBLY



- Reads can be assembled de novo into genes
- Assembly improves annotation accuracy
- Originally Trinity was used but now Spades
- Pros:
  - splicing better handled for eukaryotes
- Cons:
  - time consuming,
  - computationally intensive,
  - hardly needed for bacteria
  - Chimeras, missasembled contigs – there are tools for identification of those

Celaj et al. 2014 Microbiome

WORKFLOW


Sequencing sample
Reads preprocessing
Split reads between rRNA and mRNA
Use rRNA for taxonomy (*)
!
rRNA
Assemble mRNA into genes (*)
mRNA
Map against database (e.g., RefSeq, SEED)
Use reads
Stratify expressed genes/ function per taxa
comparative analysis

Annotation with databases

- Relies on sequence similarity searches with tools such as BLAT, BLAST, BWA
- BWA, BLAT – rely on near perfect matches
- Local alignment with Smith-Waterman algorithm
- BLAST – very time consuming → Diamond – in a blastx mode, much faster

Summary of BLASTX matches for a mouse metatranscriptome

% of Read Length

% ID of match	55	60	65	70	75	80	85	90	95	100
35	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
40	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
45	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
50	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.1
55	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.4	0.3
60	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.5	0.7	1.6
65	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.9	1.3	1.5
70	0.0	0.0	0.0	0.1	0.1	0.3	0.3	2.2	1.6	1.7
75	0.0	0.0	0.1	0.2	0.3	0.4	0.5	1.6	3.8	2.0
80	0.0	0.0	0.0	0.3	0.5	0.5	0.5	1.7	2.4	4.9
85	0.0	0.0	0.1	0.5	0.9	1.1	0.8	1.8	2.7	3.3
90	0.0	0.0	0.2	0.9	0.6	0.6	1.7	1.9	3.3	4.0
95	0.1	0.1	0.2	0.8	0.6	0.6	0.7	1.9	3.7	4.8
100	0.1	0.2	0.3	0.6	0.7	0.7	1.3	2.9	8.6	12.1

You want to filter based on % of read length but also % ID of match

MANY READS STILL COME OUT UNANNOTATED

Annotation with databases - taxonomy

Assigning RNA reads to taxa might reveal which critical functions are associated with whom

It could also help binning for the assembly.

JOURNAL ARTICLE

The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes

Ross Overbeek, Tadhg Begley, Ralph M. Butler, Jomuna V. Choudhuri, Han-Yu Chuang, Matthew Cohoon, Valérie de Crécy-Lagard, Naryttza Diaz, Terry Disz, Robert Edwards ...
Show more

Nucleic Acids Research, Volume 33, Issue 17, 1 September 2005, Pages 5691–5702,
https://doi.org/10.1093/nar/gki866
Published: 01 January 2005 **Article history**

Subsystems

Ounit *et al. BMC Genomics* (2015) 16:236
DOI 10.1186/s12864-015-1419-2

BMC Genomics

RESEARCH ARTICLE Open Access

CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative *k*-mers

Rachid Ounit[1], Steve Wanamaker[2], Timothy J Close[2] and Stefano Lonardi[1*]

CLARK - 2015

Published online 31 August 2012 *Nucleic Acids Research, 2013, Vol. 41, No. 1 e3*
doi:10.1093/nar/gks828

Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms

Jiemeng Liu[1,2,3], Haifeng Wang[1,4], Hongxing Yang[1,4], Yizhe Zhang[5], Jinfeng Wang[6], Fangqing Zhao[6,*] and Ji Qi[1,4,*]

[1]State Key Laboratory of Genetic Engineering, [2]State Key Laboratory of Surface Physics, [3]The T-Life Research Center, [4]Institute of Plant Biology, School of Life Sciences, Fudan University, [5]School of Life Sciences, Shanghai Jiaotong University, Shanghai 200433, [6]Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, People's Republic of China

Received March 20, 2012; Revised July 27, 2012; Accepted August 9, 2012

NBC - 2011

JOURNAL ARTICLE

Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation

Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei ...
Show more

RefSeq NCBI

Wood and Salzberg *Genome Biology* 2014, **15**:R46
http://genomebiology.com/2014/15/3/R46

Genome Biology

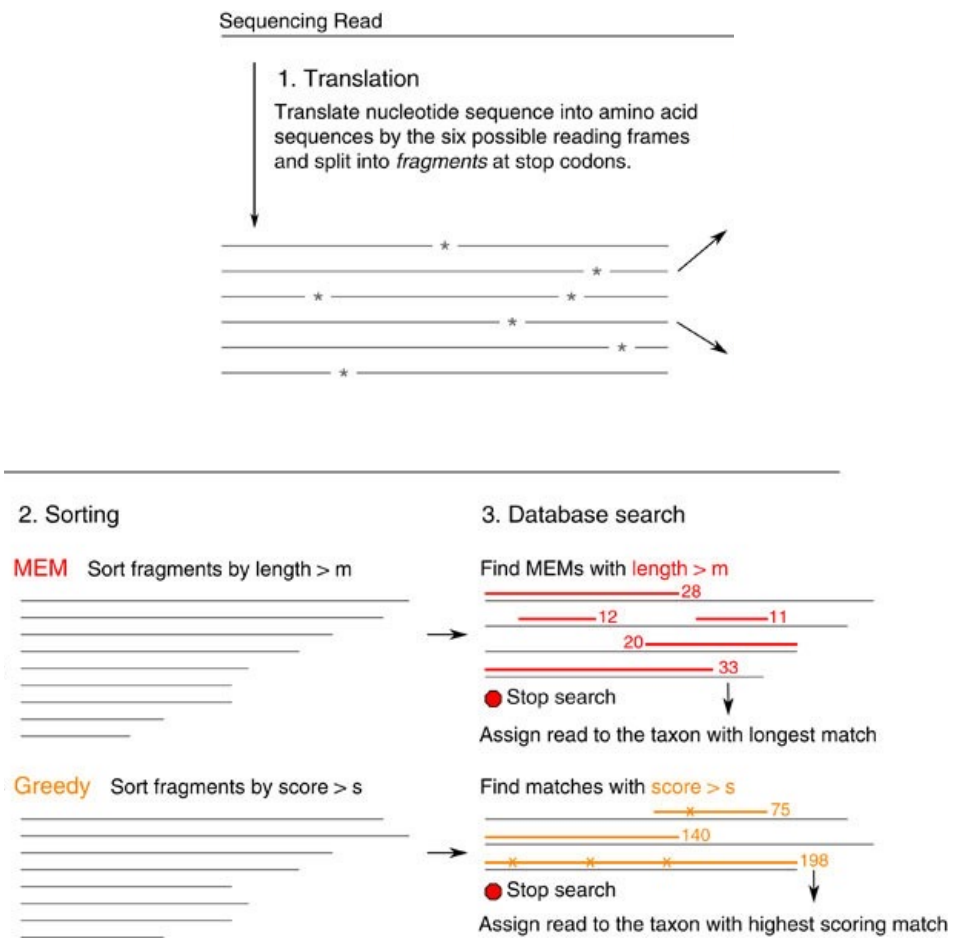
METHOD Open Access

Kraken: ultrafast metagenomic sequence classification using exact alignments

Derrick E Wood[1,2*] and Steven L Salzberg[2,3]

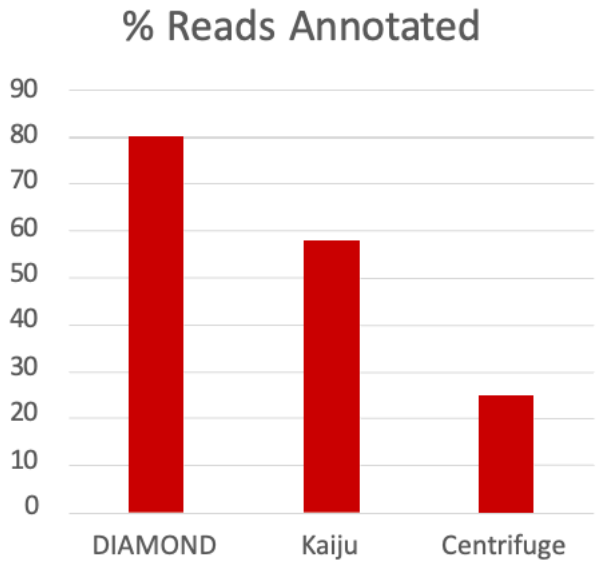
KRAKEN - 2014

Annotation with databases - taxonomy



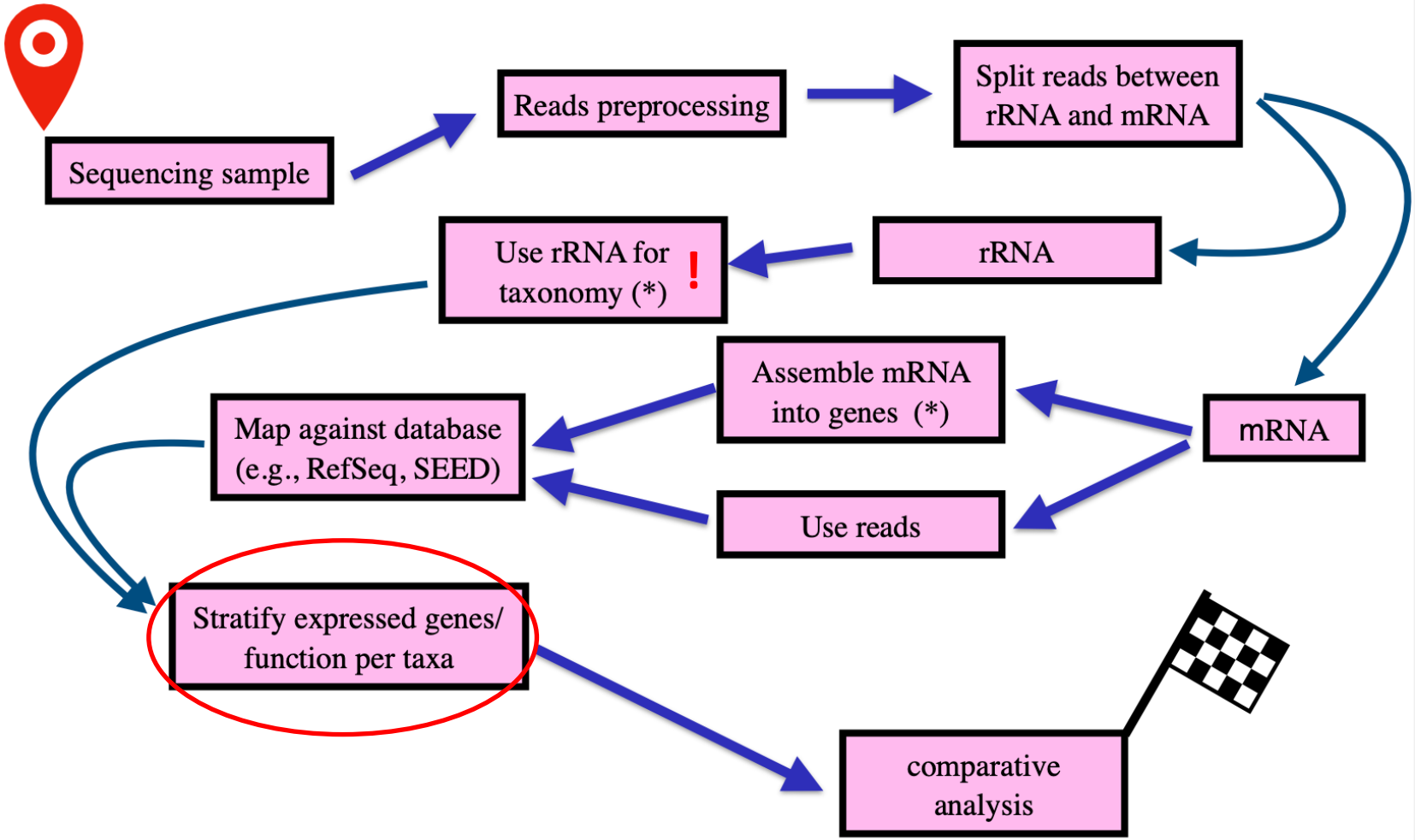
- How Kraken works already mentioned
- Mapping to Refseq or Subsystems done with Diamond (blastx type of search)
- A lot of the methods work on Nearest Neighbour to assign sequence to the genome
- Kaiju is based on Burrows Wheeler Transform (BWT) and also uses a database e.g. RefSeq
  - Fast
  - Accounts for sequencing errors
  - Needs large amount of memory
  - Features a GUI to explore results

Annotation with databases - taxonomy



- Diamond used in standardized workflows
- For a SPF mouse gut microbiome – DIAMOND annotates most reads

WORKFLOW


Sequencing sample
Reads preprocessing
Split reads between rRNA and mRNA
Use rRNA for taxonomy (*)
!
rRNA
Assemble mRNA into genes (*)
mRNA
Map against database (e.g., RefSeq, SEED)
Use reads
Stratify expressed genes/ function per taxa
comparative analysis

Mappings -> gene expression

To normalize expression to account for differences in gene length - read counts are converted to *Reads per kilobase of transcript mapped (RPKM)*

Longer transcripts should have more reads mapping to them

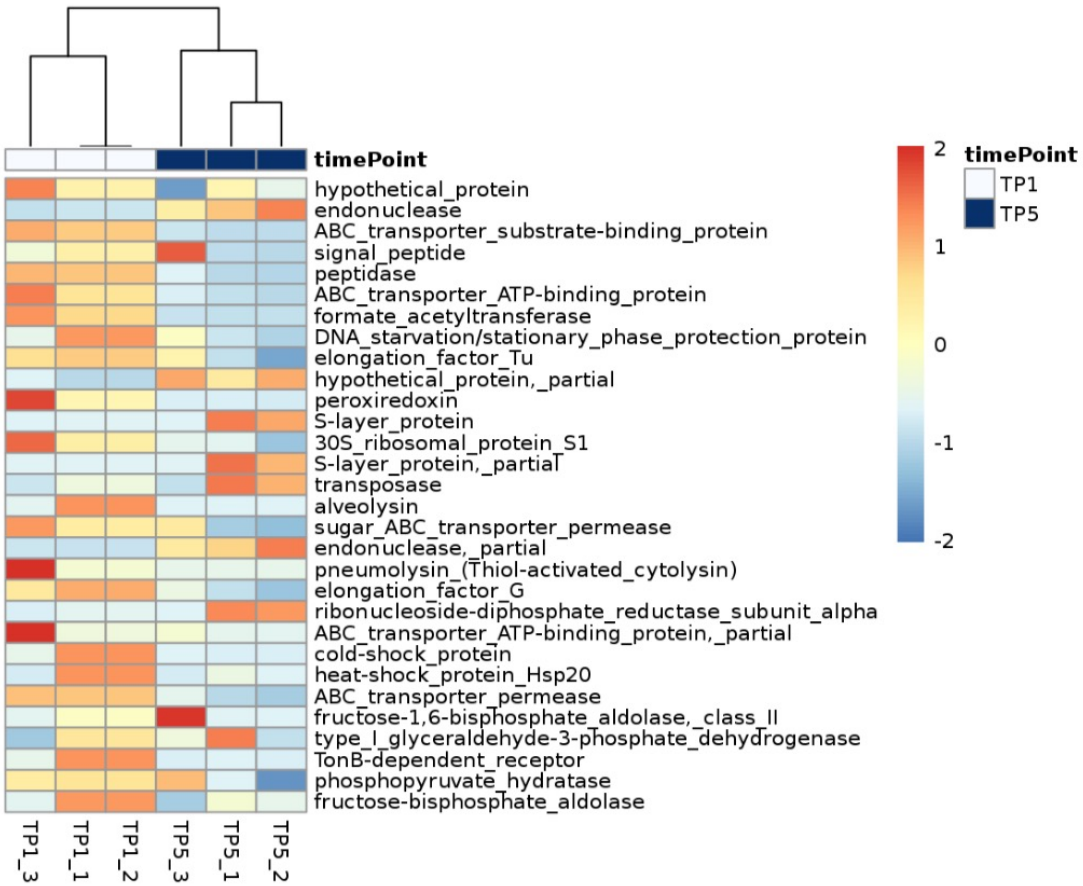
$$RPKM_{geneA} = 10^9 \, C_{geneA} / NL$$

C_{geneA} = number of reads mapped to geneA
N = total number of reads
L = length of transcript in units of Kb

Several softwares are available to do mapping and calculate normalized expression measurements across different samples including Bowtie and Cufflinks

But also implemented in several standardized workflows

Functional annotation/stratification



Mapping transcripts to more general functions

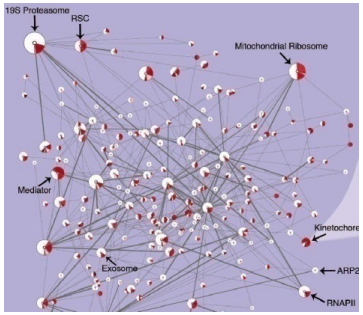
- Gene Ontology Annotation can help statify functions assigned to transcripts: EggNOG, OMA, Pannzer2

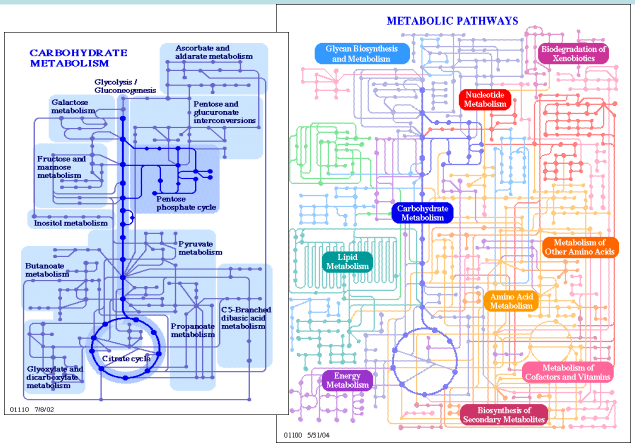
- Mapping transcripts into more general functions – transcripts might not be important but might code for similar functions

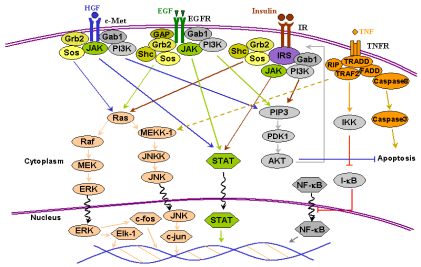
Functional categories

Genes and proteins form parts of interconnected functional modules

Grouping gene functions into larger categories can help explain which metabolic processes are most important in the sample/ different between samples.







Protein complexes

InterProScan

UniProt

Metabolic pathways

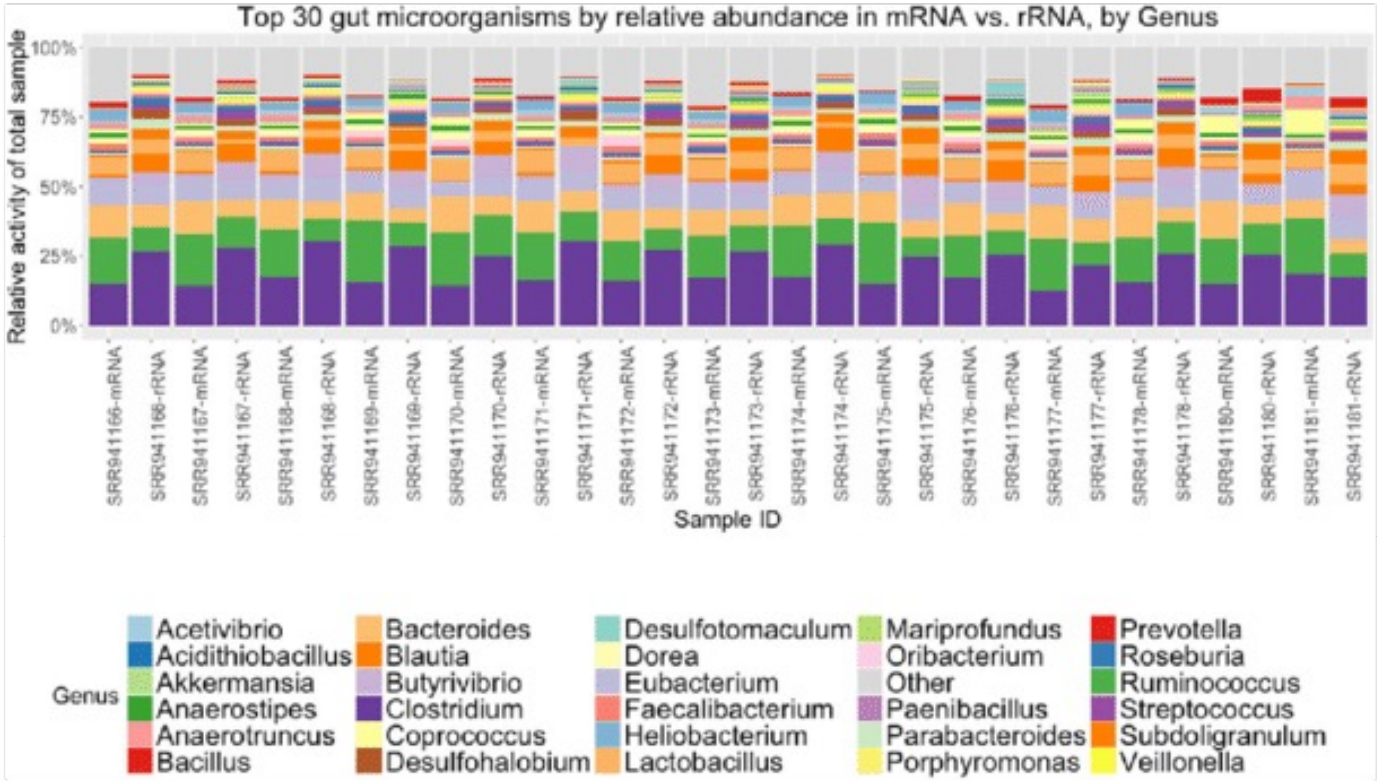
KEGG

Signalling pathways

KEGG/IPath

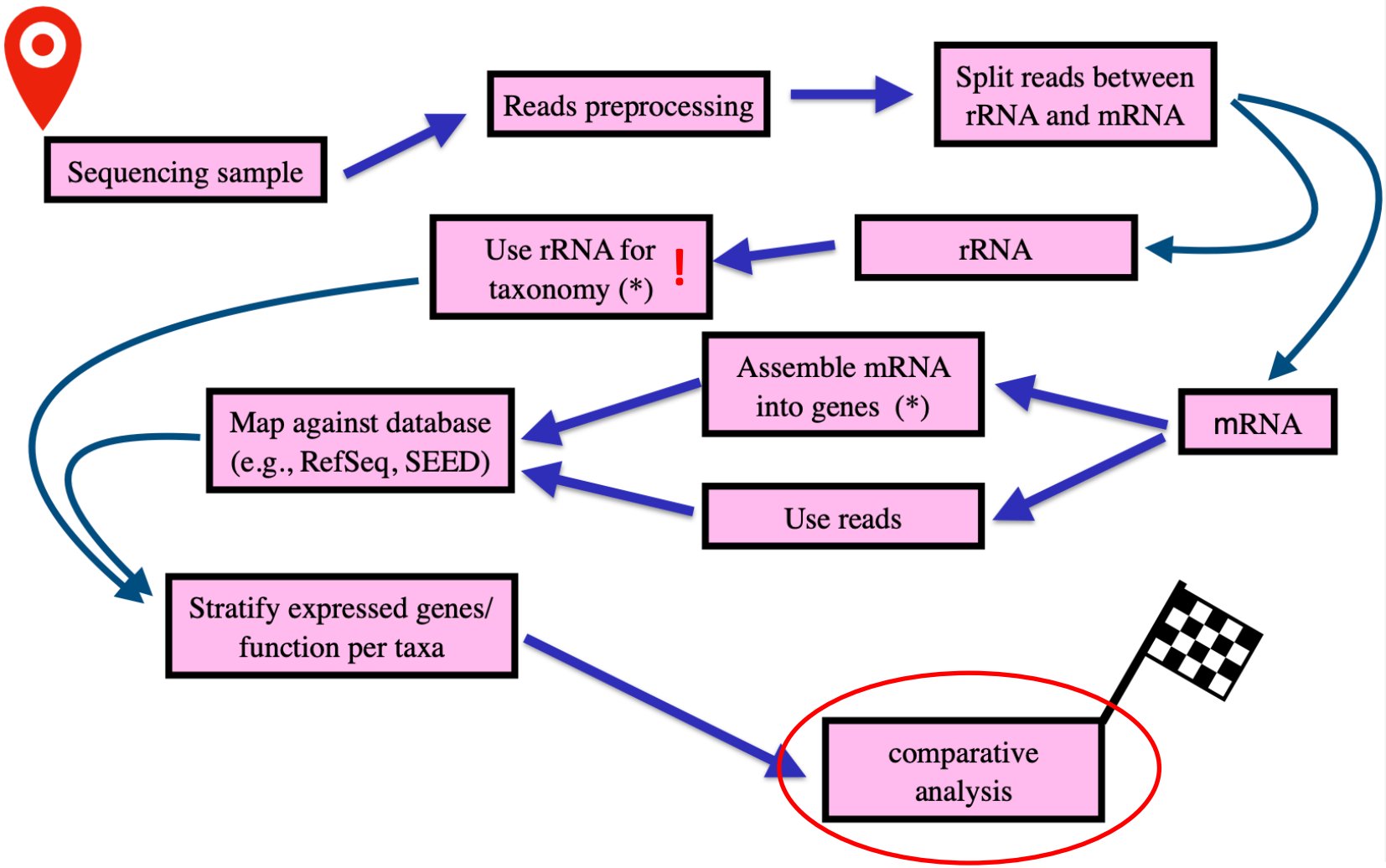
Taxonomy from mRNA expression ≠ rRNA abundance

- Comparison of mRNA vs. rRNA based abundance estimates
- Abundance not the same as expression!!
- Differences can be subtle but also can be big

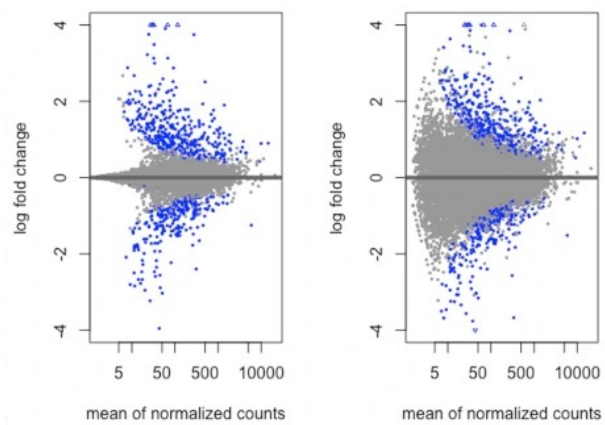


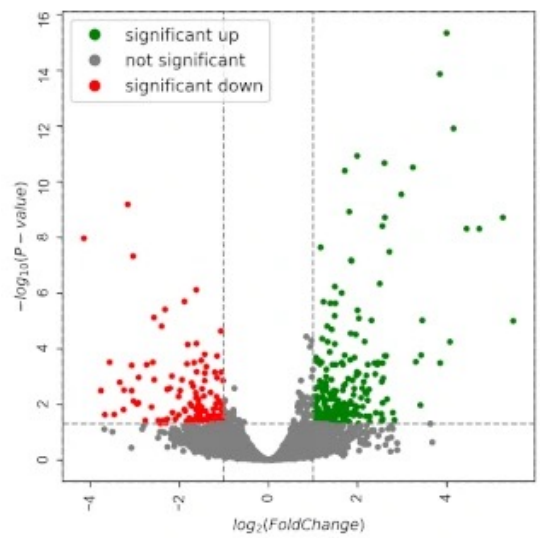
Westreich et al 2016

WORKFLOW


Sequencing sample
Reads preprocessing
Split reads between rRNA and mRNA
Use rRNA for taxonomy (*)
!
rRNA
Assemble mRNA into genes (*)
mRNA
Map against database (e.g., RefSeq, SEED)
Use reads
Stratify expressed genes/ function per taxa
comparative analysis

Statistical/ comparative analysis





- DESeq2 takes in a count table

1. Normalisation
   - **geometric mean** is calculated for **each gene** across all samples,
   - **counts for a gene in each sample** is **divided** by this mean,
   - **median** of these ratios in a sample is the size factor for that sample
2. Variance estimation (dispersion and fold changes)
   - within-group variance calculated between replicates
   - shrinkage estimation for dispersions and fold changes
   - dispersion value is estimated for each gene through a model fit procedure.
3. Differential expression
   - negative binomial generalized linear models fitting for each gene and the Wald test for significance testing

Samsa2 (https://github.com/transcript/samsa2

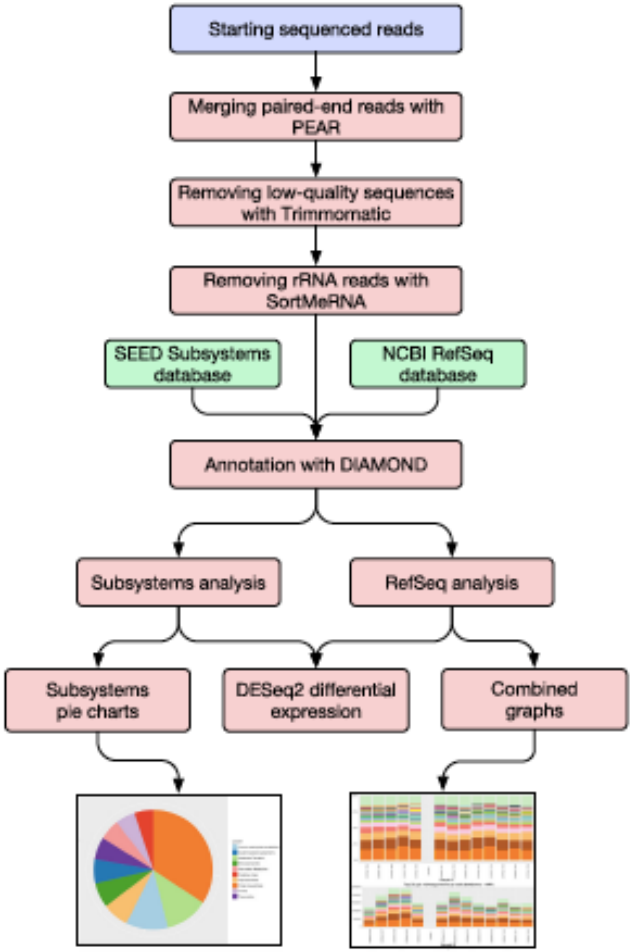
README.md

SAMSA2 - A fork of the complete metatranscriptome analysis pipeline

Version 2 of the SAMSA pipeline - faster! Lighter! More options! Less waiting!

New in version 2:

- DIAMOND integration, allowing for SAMSA2 to be run without ever needing an MG-RAST account.
- Option to annotate against custom databases.
- Better, more polished R scripts that can be executed from the command line.
- PCA plots and other graphical outputs.
- Filtering of ribosomes for even more speed.
- And more!

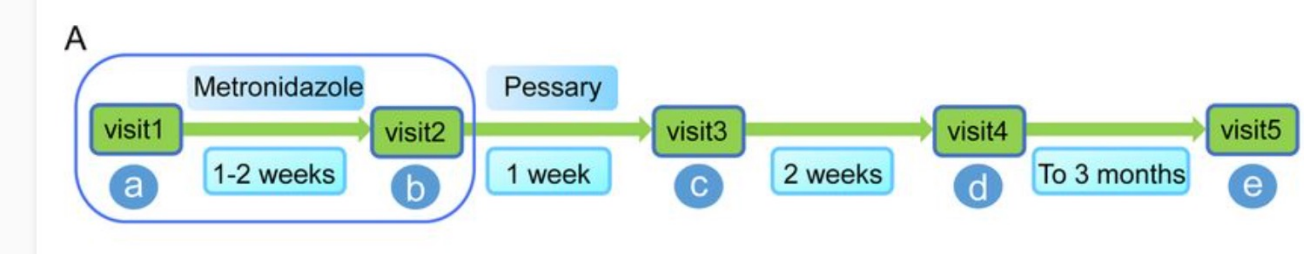


Remarks

•All tools nicely wrapped up

Tutorial – Bacterial Vaginosis

- multifactorial disease characterized by a shift from the *Lactobacillus* species-dominated microbial community toward a taxonomically diverse anaerobic community
- 8 women diagnosed with BV
- 4 successfully treated, 4 not
- Paired-end seq performed on a HiSeq 2500 Sequencer to yield 2 × 110-bp



https://journals.asm.org/doi/full/10.1128/mSphereDirect.00262-18