MeBaPiNa:
a **MetaBa**rcoding Analysis **Pi**peling for **Na**nopore Datasets

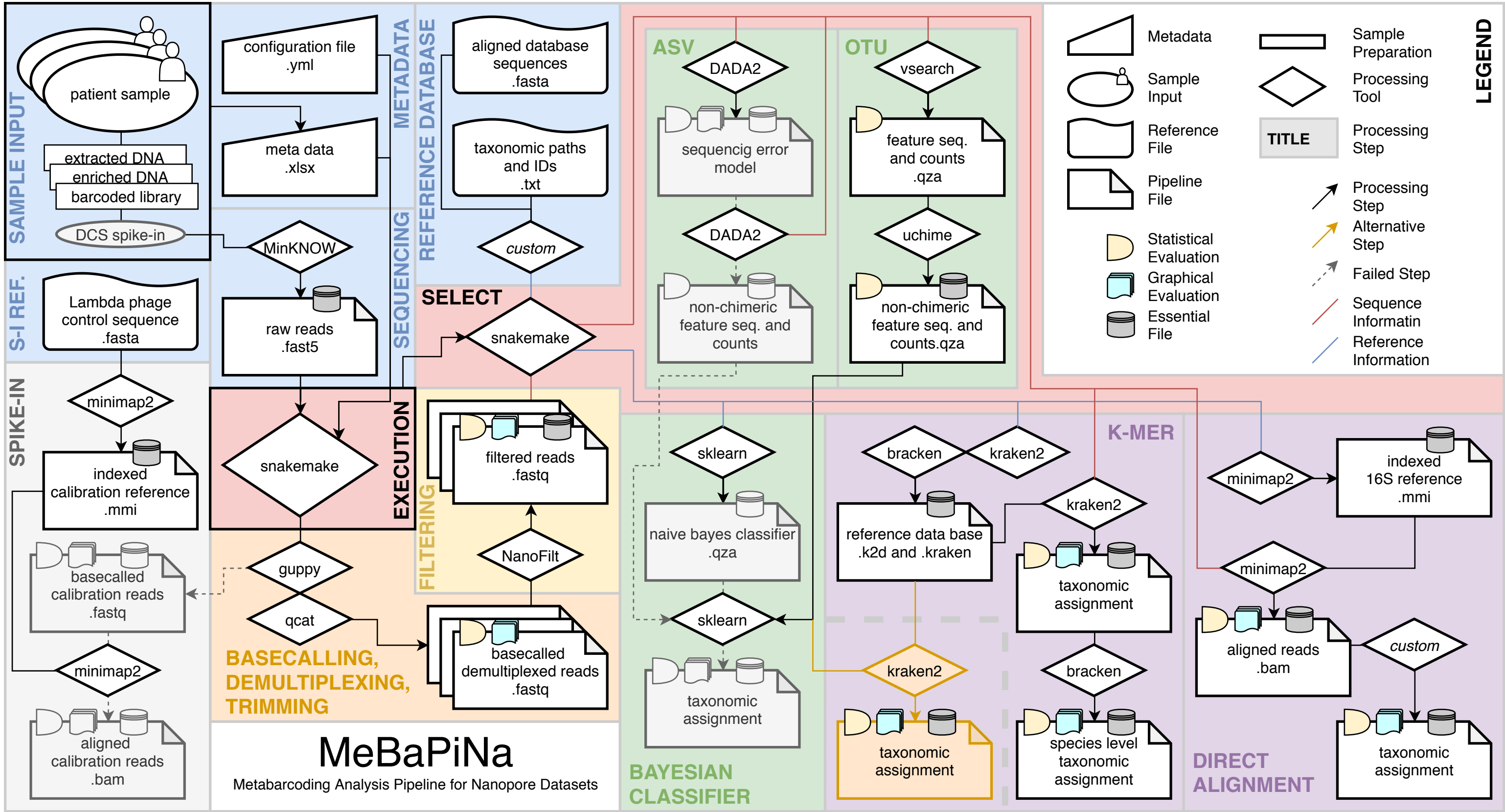
Thesis Title:
Comparing Metabarcoding Analysis Methodologies for Nanopore Sequencing in Clinical Application

Marc Rübsam

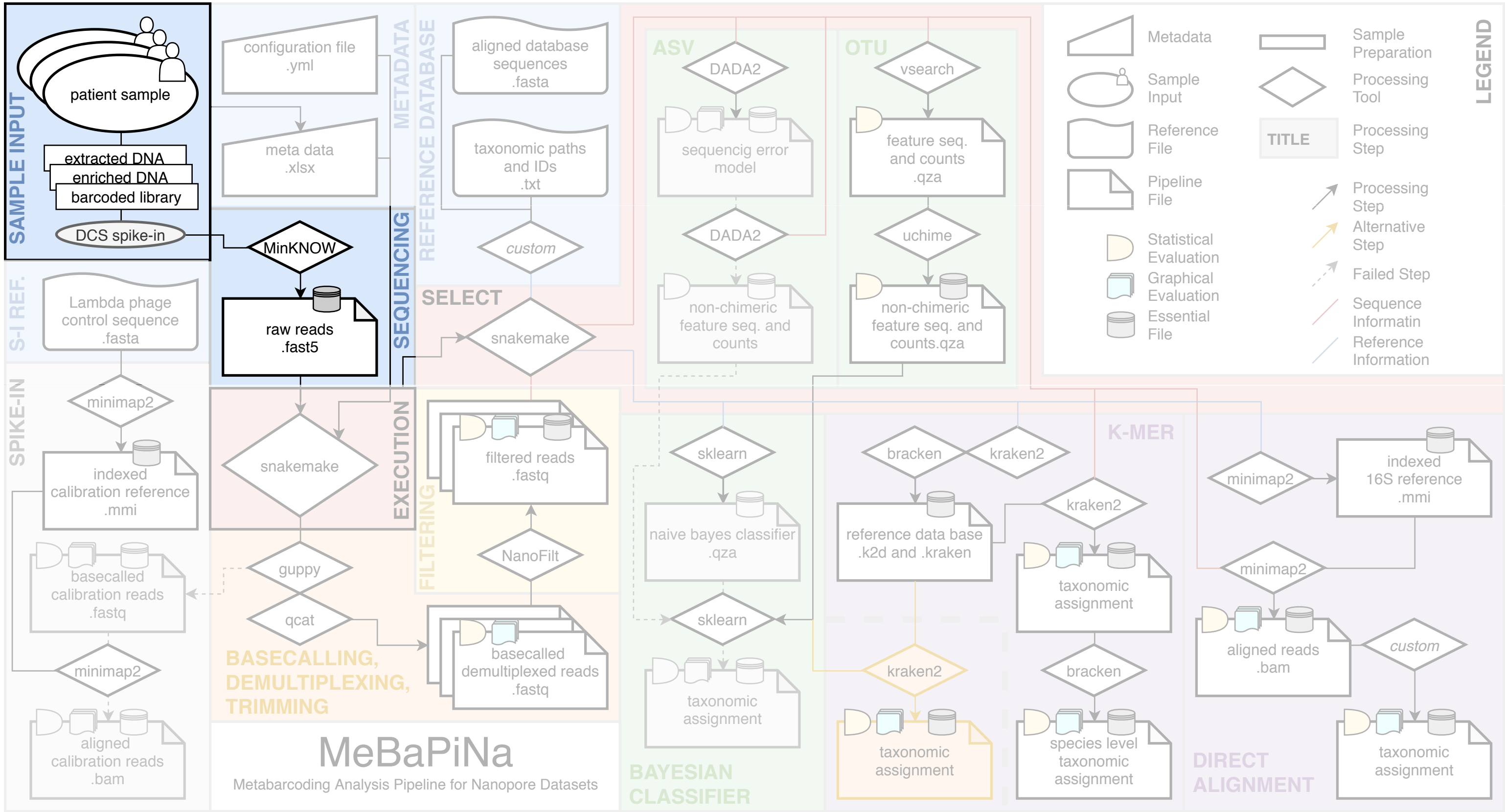
18th of May 2020

GERMAN
CANCER RESEARCH CENTER
IN THE HELMHOLTZ ASSOCIATION

Metabarcoding analysis pipeline for Nanopore datasets



Sample Input and Sequencing



SAMPLE INPUT
patient sample
extracted DNA
enriched DNA
barcoded library
DCS spike-in

METADATA
configuration file .yml
meta data .xlsx

REFERENCE DATABASE
aligned database sequences .fasta
taxonomic paths and IDs .txt
custom

SEQUENCING
MinKNOW
raw reads .fast5

S-I REF.
Lambda phage control sequence .fasta

SPIKE-IN
minimap2
indexed calibration reference .mmi
basecalled calibration reads .fastq
minimap2
aligned calibration reads .bam

EXECUTION
snakemake

BASECALLING, DEMULTIPLEXING, TRIMMING
guppy
qcat
basecalled demultiplexed reads .fastq

FILTERING
NanoFilt
filtered reads .fastq

SELECT
snakemake

ASV
DADA2
sequencig error model
DADA2
non-chimeric feature seq. and counts

OTU
vsearch
feature seq. and counts .qza
uchime
non-chimeric feature seq. and counts.qza

BAYESIAN CLASSIFIER
sklearn
naive bayes classifier .qza
sklearn
taxonomic assignment

K-MER
bracken
kraken2
reference data base .k2d and .kraken
kraken2
taxonomic assignment
kraken2
taxonomic assignment
bracken
species level taxonomic assignment

DIRECT ALIGNMENT
minimap2
indexed 16S reference .mmi
minimap2
aligned reads .bam
custom
taxonomic assignment

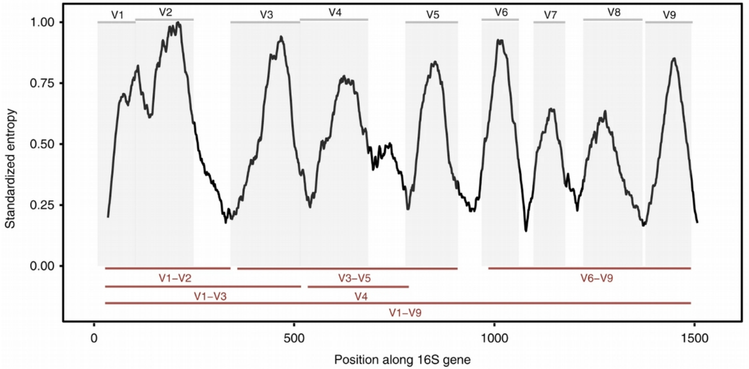
MeBaPiNa
Metabarcoding Analysis Pipeline for Nanopore Datasets

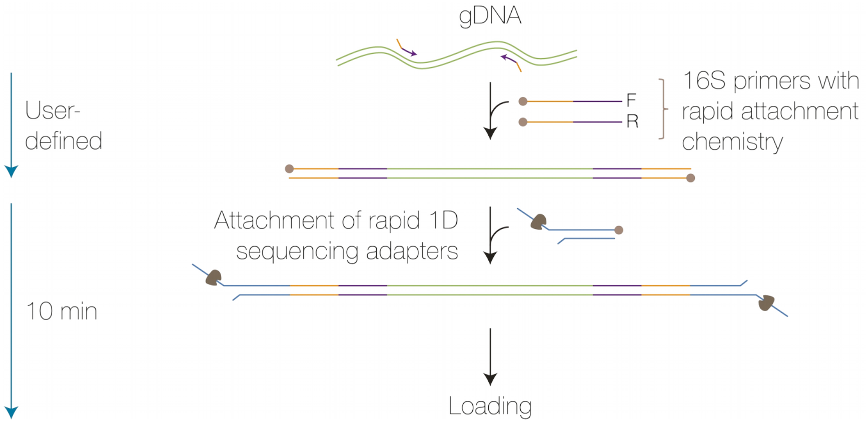
LEGEND
Metadata
Sample Input
Reference File
Pipeline File
Statistical Evaluation
Graphical Evaluation
Essential File
Sample Preparation
Processing Tool
TITLE Processing Step
Processing Step
Alternative Step
Failed Step
Sequence Informatin
Reference Information

Full-length 16S rRNA metabarcoding with Nanopore

Sequencing library

- Samples
  - mock community
  - clinical stool/tissue samples
  - controls
- 16S rRNA metabarcoding
  - Conserved regions → primers
  - Variable regions → distinction




gDNA

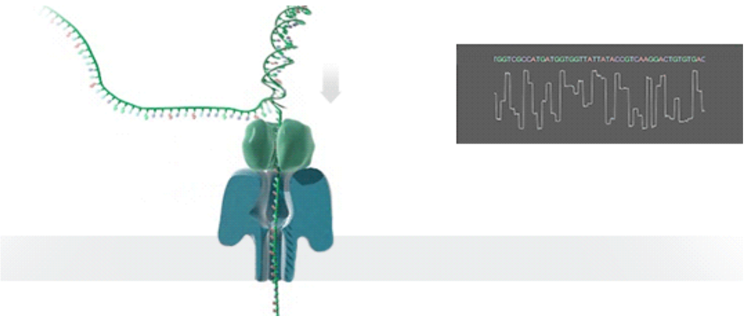
User-defined

16S primers with rapid attachment chemistry

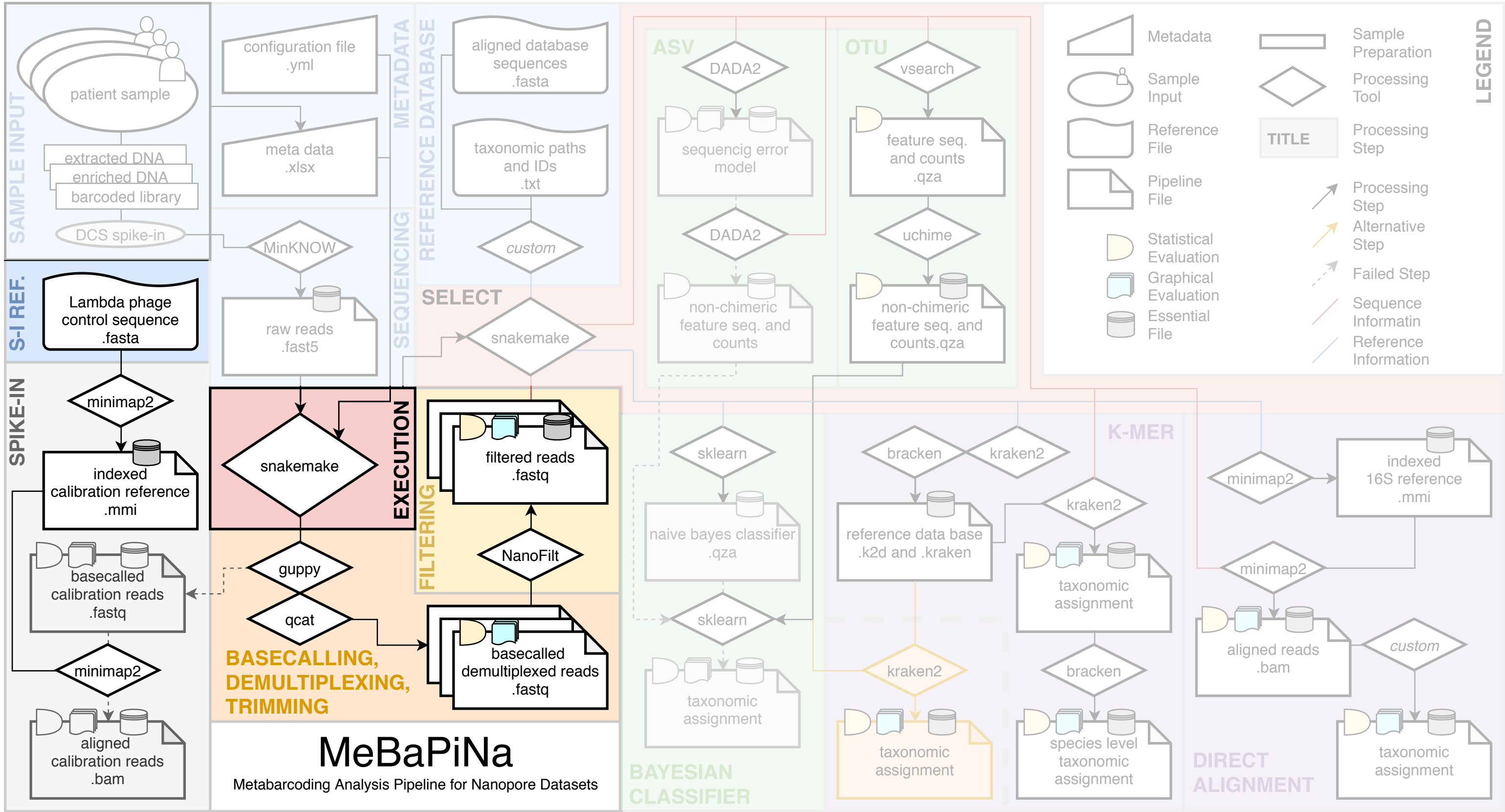
10 min

Attachment of rapid 1D sequencing adapters

Loading



Basecalling and read QC



SAMPLE INPUT
patient sample
extracted DNA
enriched DNA
barcoded library
DCS spike-in

S-I REF.
Lambda phage control sequence .fasta

SPIKE-IN
minimap2
indexed calibration reference .mmi
basecalled calibration reads .fastq
minimap2
aligned calibration reads .bam

METADATA
configuration file .yml
meta data .xlsx

SEQUENCING
MinKNOW
raw reads .fast5

REFERENCE DATABASE
aligned database sequences .fasta
taxonomic paths and IDs .txt
custom

SELECT
snakemake

EXECUTION
snakemake

BASECALLING, DEMULTIPLEXING, TRIMMING
guppy
qcat
basecalled demultiplexed reads .fastq

FILTERING
NanoFilt
filtered reads .fastq

MeBaPiNa
Metabarcoding Analysis Pipeline for Nanopore Datasets

ASV
DADA2
sequencig error model
DADA2
non-chimeric feature seq. and counts

OTU
vsearch
feature seq. and counts .qza
uchime
non-chimeric feature seq. and counts.qza

BAYESIAN CLASSIFIER
sklearn
naive bayes classifier .qza
sklearn
taxonomic assignment

K-MER
bracken
kraken2
reference data base .k2d and .kraken
kraken2
taxonomic assignment
kraken2
taxonomic assignment
bracken
species level taxonomic assignment

DIRECT ALIGNMENT
minimap2
indexed 16S reference .mmi
minimap2
aligned reads .bam
custom
taxonomic assignment

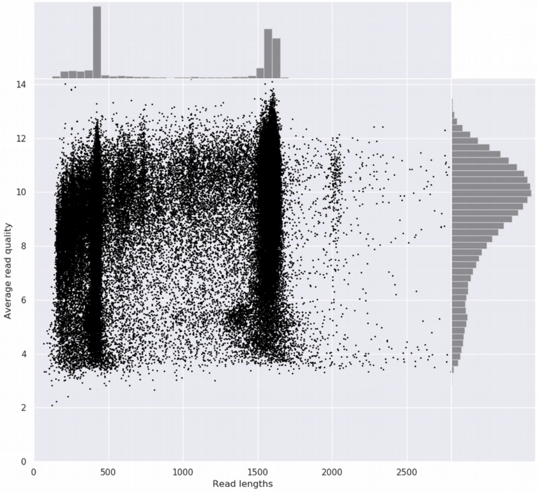
LEGEND
Metadata
Sample Preparation
Sample Input
Processing Tool
Reference File
TITLE
Processing Step
Pipeline File
Processing Step
Alternative Step
Failed Step
Statistical Evaluation
Graphical Evaluation
Essential File
Sequence Informatin
Reference Information

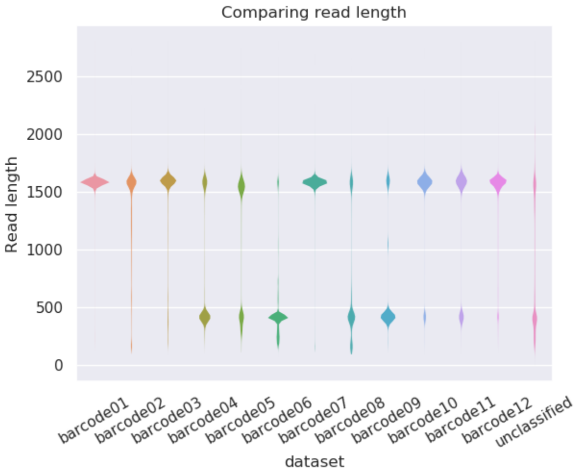
Removal of artifacts in nucleotide space

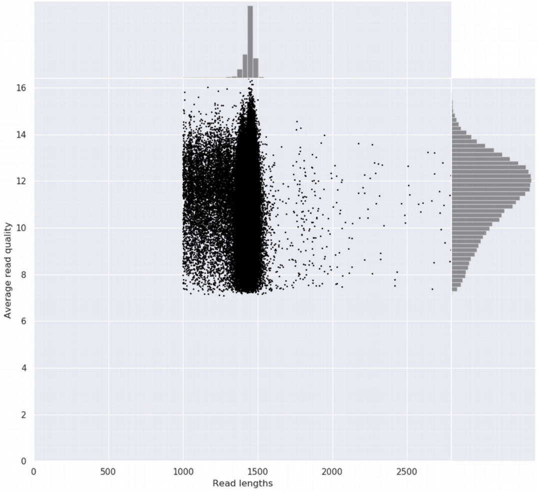
run	reads [10^6]		bases [Gb]		length [kb]	
	bac	flt	bac	flt	bac	flt
I	11.9	8.6	18.2	12.4	1.6	1.5
II	8.5	3.2	8.3	4.6	0.9	1.4

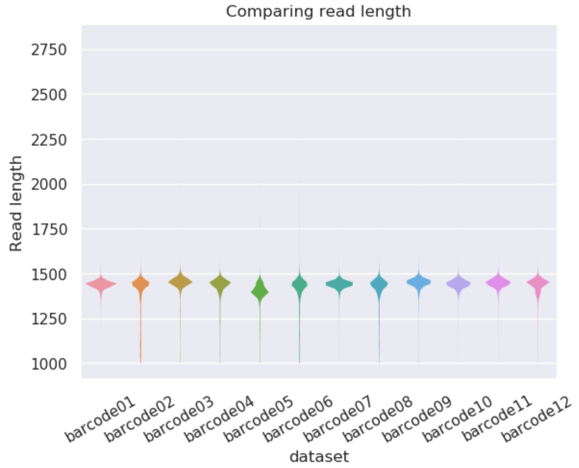
Extract amplicon reads

- Barcode detected
- Phred score >7
- Length 1000-2800 bp

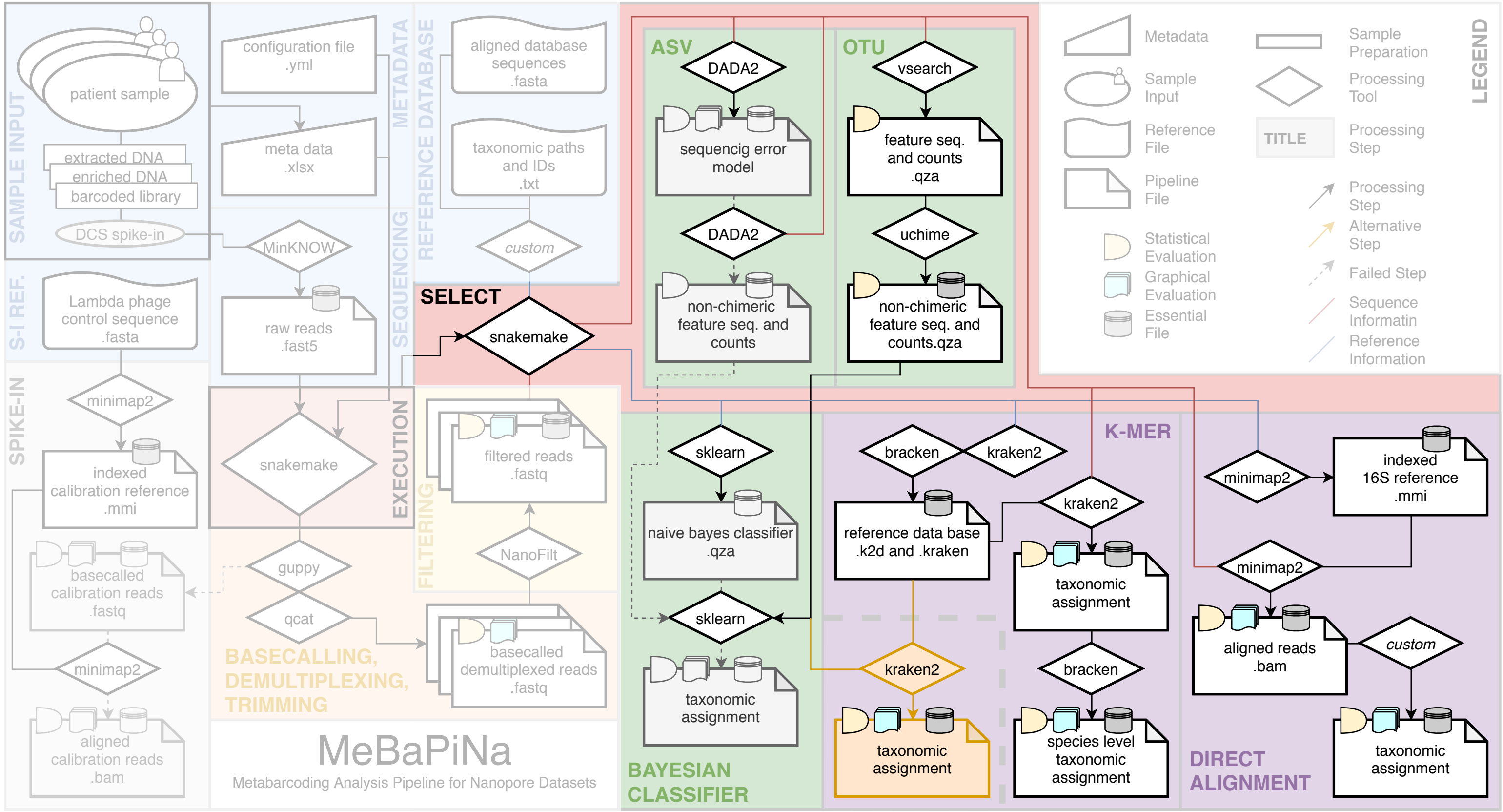



Comparing read length




Comparing read length

Alternative Methodologies



SAMPLE INPUT
patient sample
extracted DNA
enriched DNA
barcoded library
DCS spike-in

S-I REF.
Lambda phage control sequence .fasta

SPIKE-IN
minimap2
indexed calibration reference .mmi
basecalled calibration reads .fastq
minimap2
aligned calibration reads .bam

METADATA
configuration file .yml
meta data .xlsx

REFERENCE DATABASE
aligned database sequences .fasta
taxonomic paths and IDs .txt
custom

SEQUENCING
MinKNOW
raw reads .fast5

EXECUTION
snakemake

BASECALLING, DEMULTIPLEXING, TRIMMING
guppy
qcat
basecalled demultiplexed reads .fastq

FILTERING
NanoFilt
filtered reads .fastq

SELECT
snakemake

ASV
DADA2
sequencig error model
DADA2
non-chimeric feature seq. and counts

OTU
vsearch
feature seq. and counts .qza
uchime
non-chimeric feature seq. and counts.qza

BAYESIAN CLASSIFIER
sklearn
naive bayes classifier .qza
sklearn
taxonomic assignment

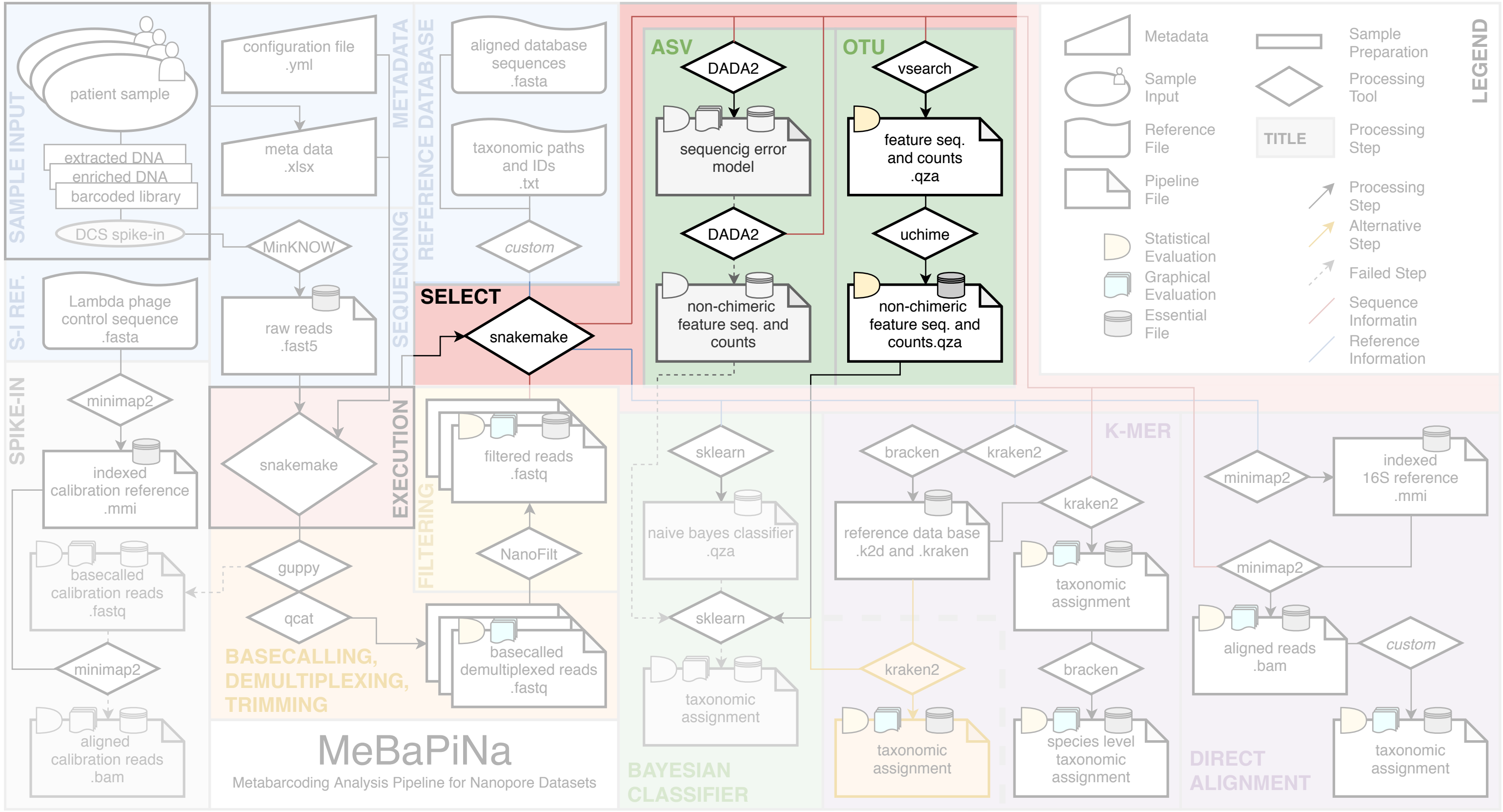
K-MER
bracken
kraken2
reference data base .k2d and .kraken
kraken2
taxonomic assignment
kraken2
taxonomic assignment
bracken
species level taxonomic assignment

DIRECT ALIGNMENT
minimap2
indexed 16S reference .mmi
minimap2
aligned reads .bam
custom
taxonomic assignment

LEGEND
Metadata
Sample Input
Reference File
Pipeline File
Statistical Evaluation
Graphical Evaluation
Essential File
Sample Preparation
Processing Tool
TITLE Processing Step
Processing Step
Alternative Step
Failed Step
Sequence Informatin
Reference Information

MeBaPiNa
Metabarcoding Analysis Pipeline for Nanopore Datasets

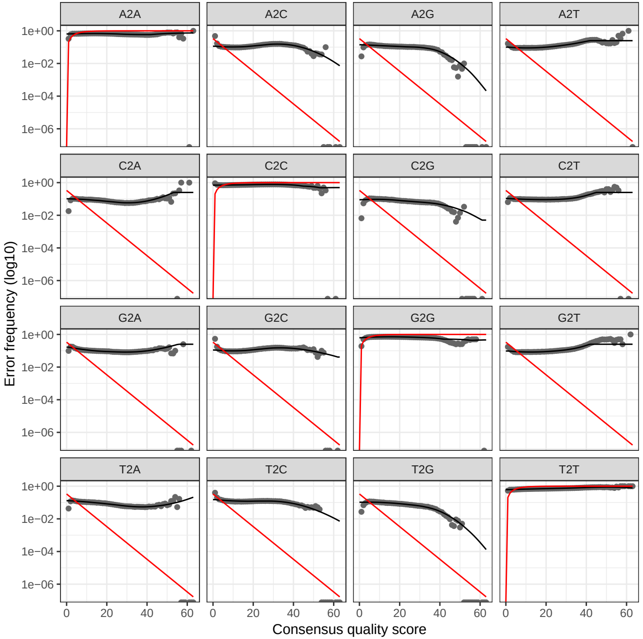
Feature Extraction



High error-rates permit feature extraction

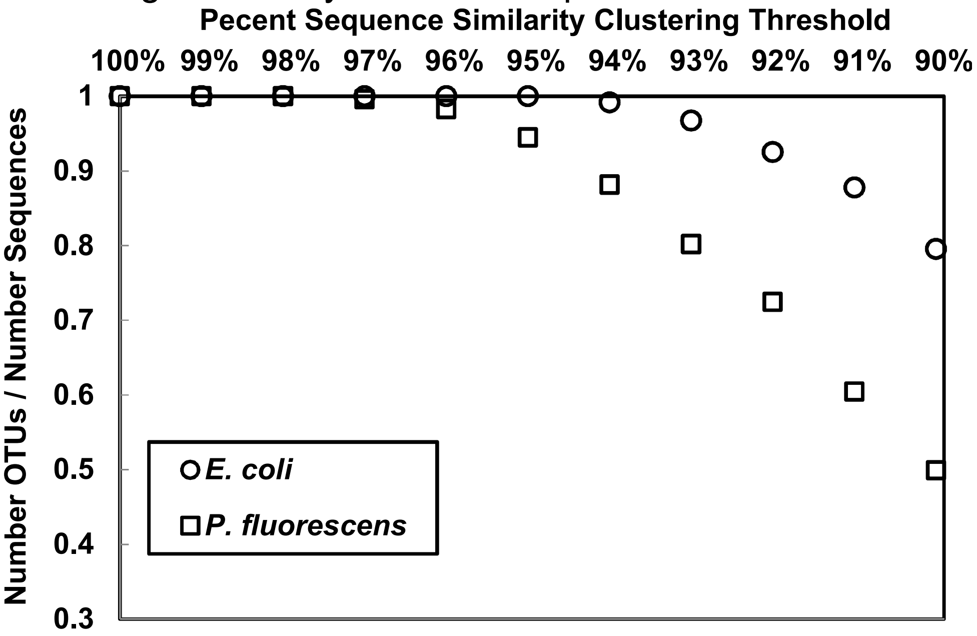
ASV recovery

- Incorrect error-model

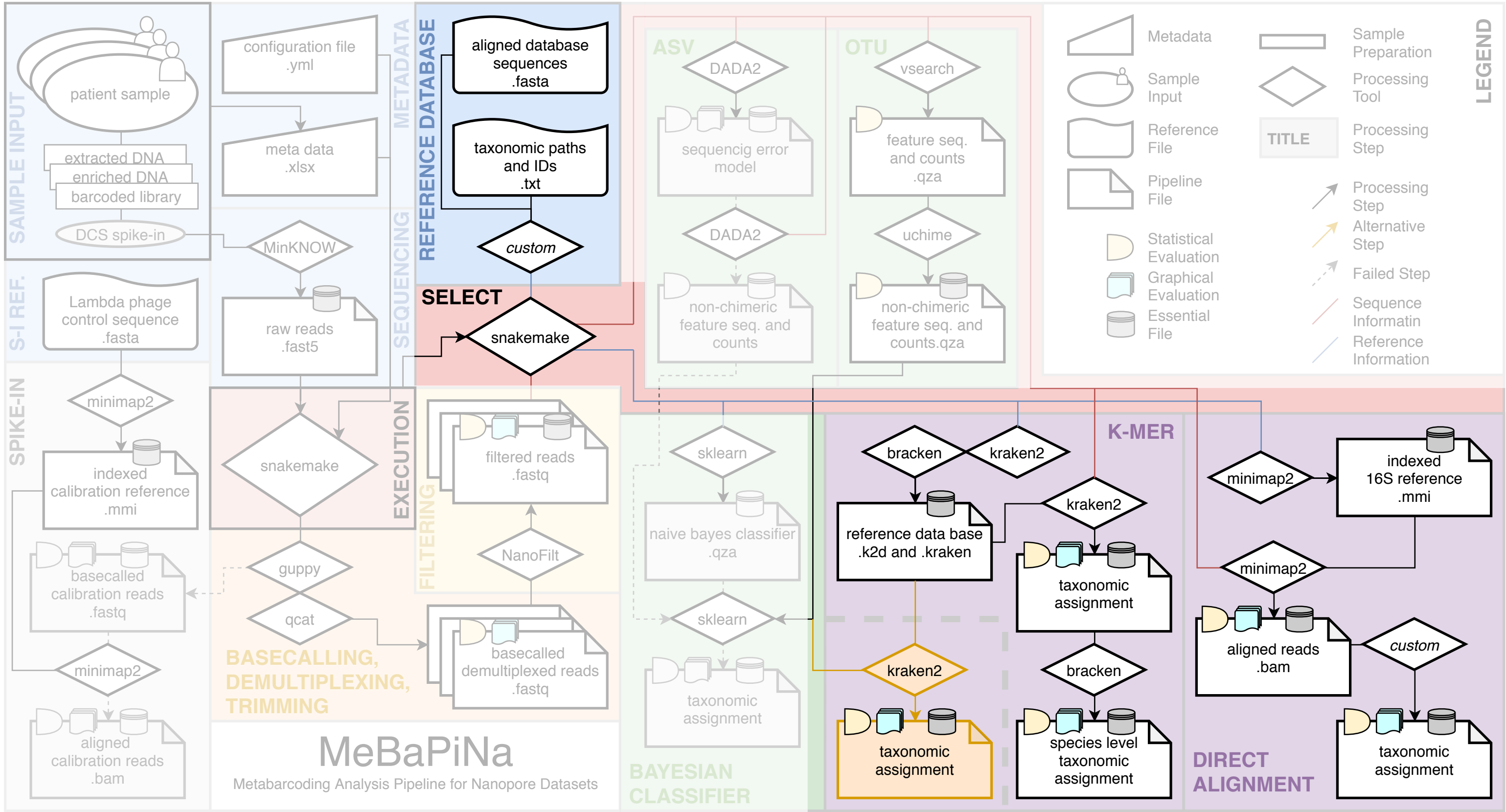


OTU picking

- Stringent identity threshold required



Taxonomic Classification



Comparison in numbers

OTU (+k-mer map.)

- General
  - ~24 h computation
  - 60%-78% assignment
- Mock community
  - **136-220 taxa**
  - 15%-16% correct assignment
- Clinical samples
  - stool: 345-560 taxa
  - lung: 142-196 taxa
  - ovary: 593 taxa

k-mer mapping

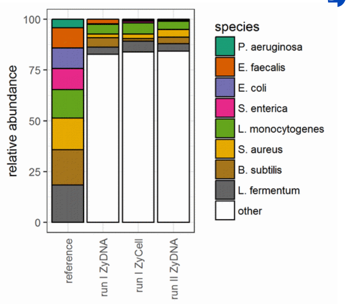
- General
  - **~10 min computation**
  - **97%-100% assignment**
- Mock community
  - 600-900 taxa
  - 21%-26% correct assignment
- Clinical samples
  - stool: 1100-1700 taxa
  - lung: 600-900 taxa
  - ovary: 1300 taxa

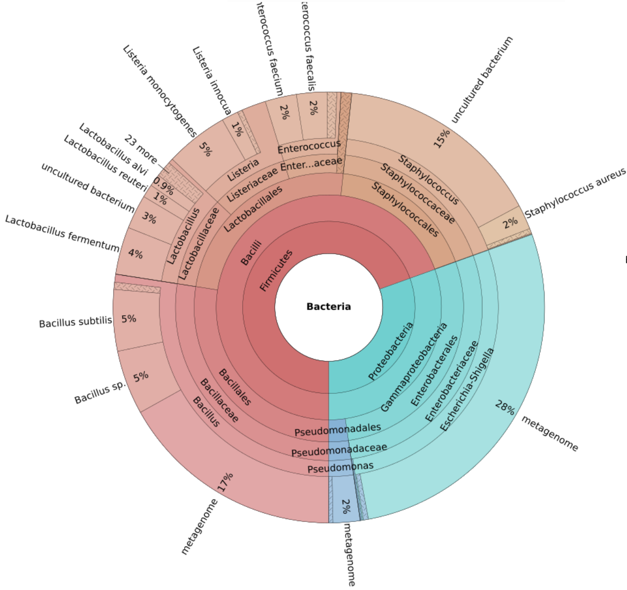
Full-length alignment

- General
  - ~2 h computation
  - 65%-86% assignment
- Mock community
  - **120-236 taxa**
  - **65%-78% correct assignment**
- Clinical samples
  - stool: 298-575 taxa
  - lung: 103-184 taxa
  - ovary: 353 taxa

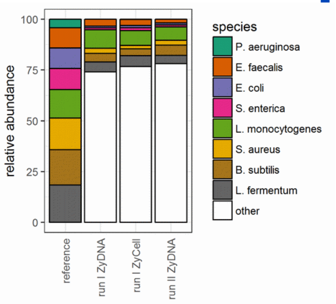
Mock Community Species-Level Comparison

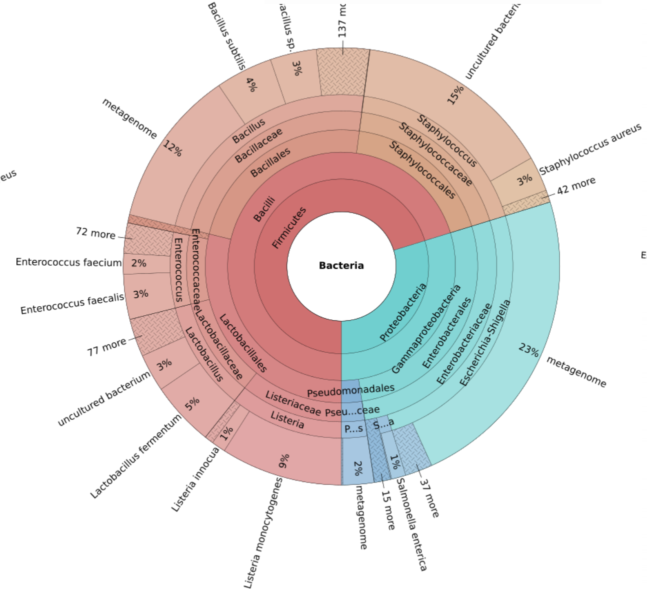
OTU



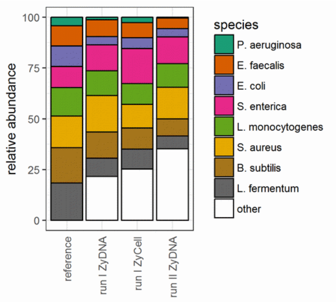


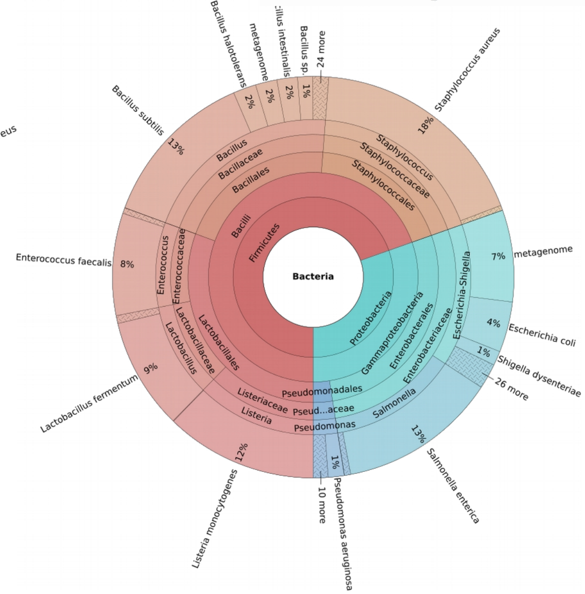
k-mer





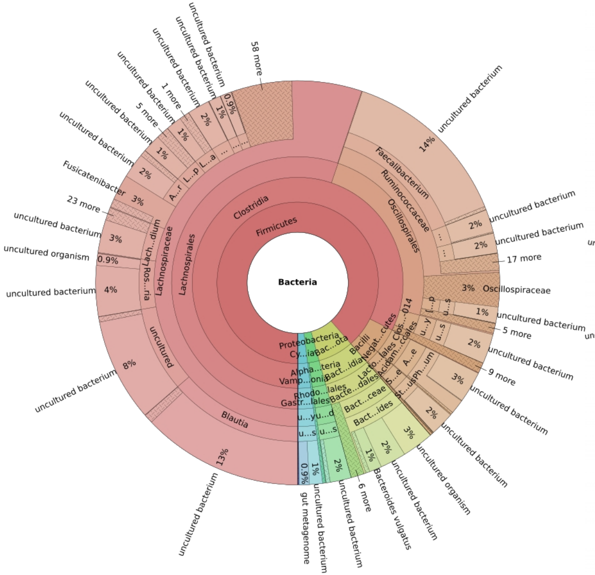
align



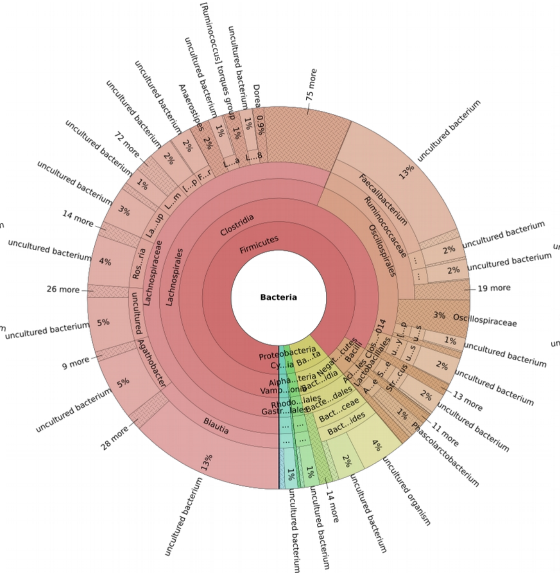


Different diversity in clinical sample

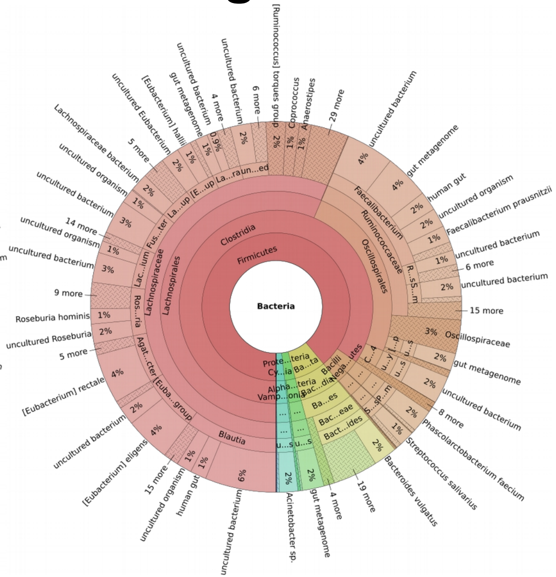
OTU



k-mer



align



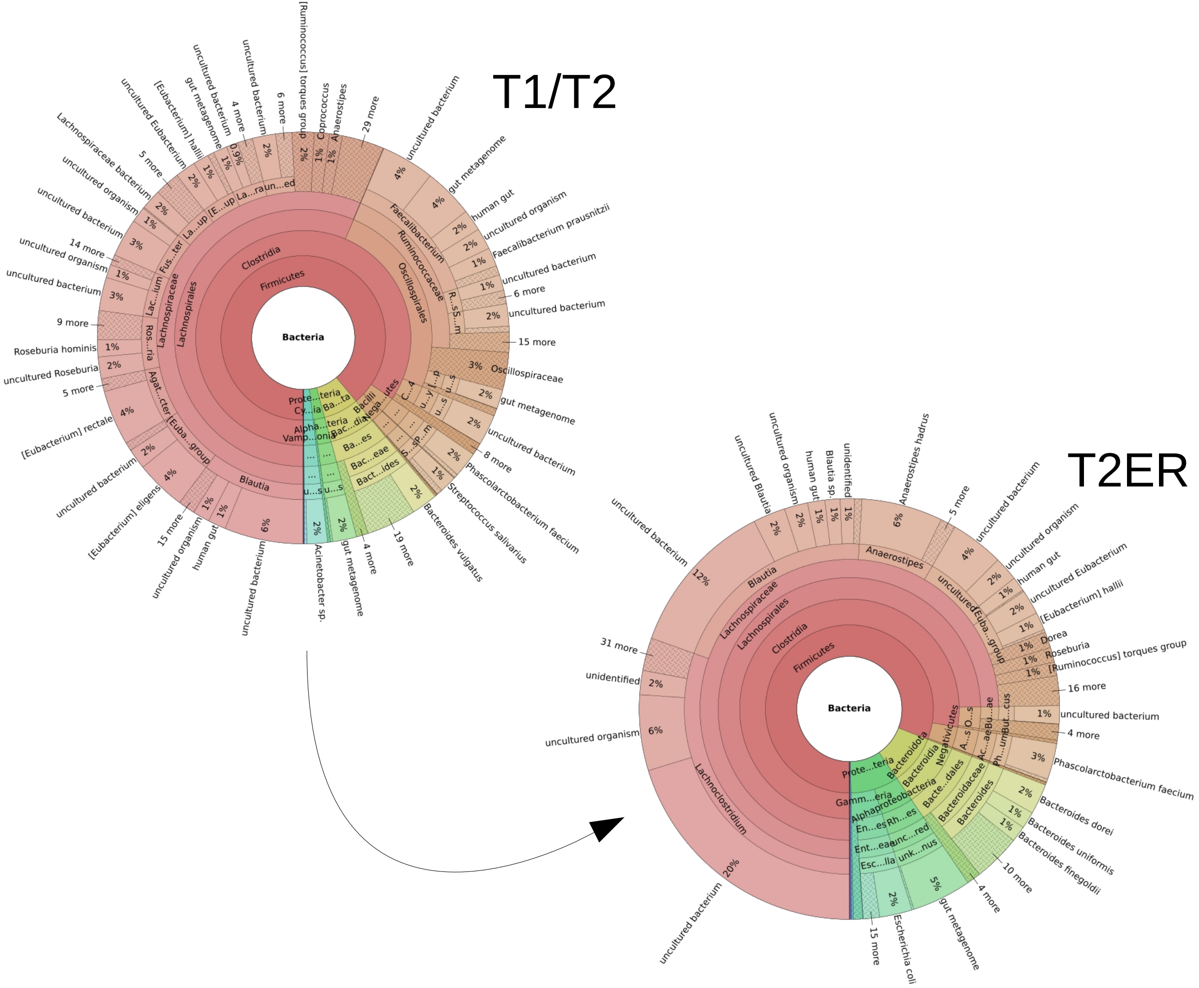
Time course

Observable dynamic

- In one of two patients
- Decrease diversity
  - Richness: 526 → 298
  - Shannon entropy: 4.49 → 3.46

- Constant in other patient

→ higher sample size required



Conclusion

Samples and Sequencing

- Successful representation of PROMISE samples
- Good representation of microbial composition and its dynamic

Pipeline

- Working pipeline, reproducible, configurable
- Full-length alignment performs best, but may come with biases
- K-mer mapping is faster, is maintained standalone package
- Feature extraction not beneficial

Outlook

Pipeline

- Optimize parameter (especially k-mer mapping)
- Evaluate performance with other reference database
- Scale up sample set
- Resequence a sample to asses deviation

Wet-lab

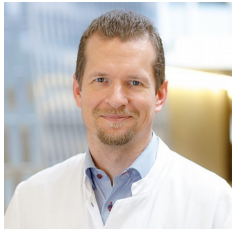
- Different primer sets
- Optimize library preparation to minimize artifacts (e.g. adjust bead ratio)

Further analysis – the fun just started

- Normalization, removal of kit contamination
- Clustering analysis
- Deferentially abundant taxa analysis

Thank you!

Dr. Matthias Schlesner



PD Dr. Niels Halama

Pornpimol Charoentong, PhD

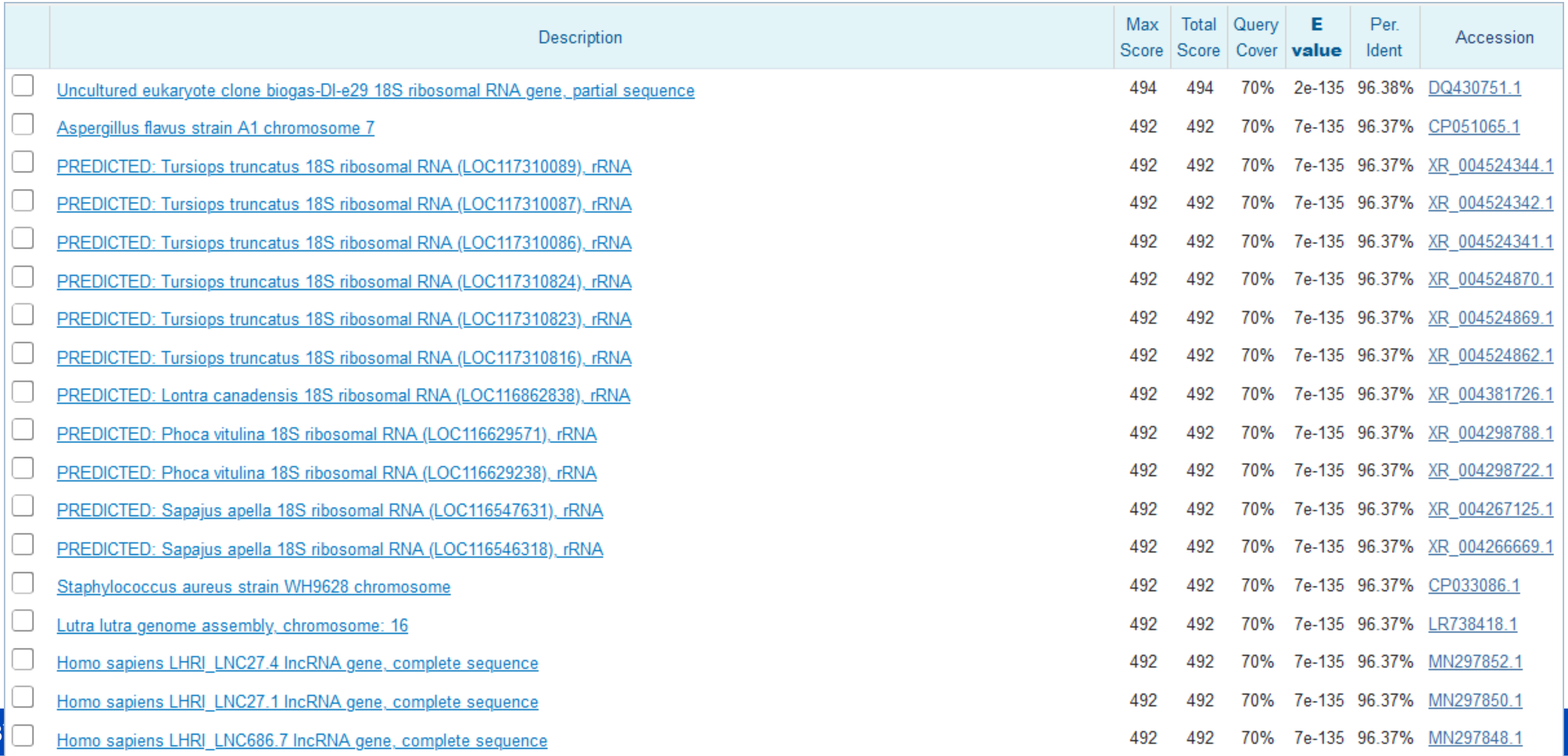
Dr. Silke Grauling-Halama

Daniel Browne

NCT and DKFZ Team

18S contamination

Quick BLAST



	Description	Max Score	Total Score	Query Cover	**E value**	Per. Ident	Accession
□	Uncultured eukaryote clone biogas-DI-e29 18S ribosomal RNA gene, partial sequence	494	494	70%	2e-135	96.38%	DQ430751.1
□	Aspergillus flavus strain A1 chromosome 7	492	492	70%	7e-135	96.37%	CP051065.1
□	PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310089), rRNA	492	492	70%	7e-135	96.37%	XR_004524344.1
□	PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310087), rRNA	492	492	70%	7e-135	96.37%	XR_004524342.1
□	PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310086), rRNA	492	492	70%	7e-135	96.37%	XR_004524341.1
□	PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310824), rRNA	492	492	70%	7e-135	96.37%	XR_004524870.1
□	PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310823), rRNA	492	492	70%	7e-135	96.37%	XR_004524869.1
□	PREDICTED: Tursiops truncatus 18S ribosomal RNA (LOC117310816), rRNA	492	492	70%	7e-135	96.37%	XR_004524862.1
□	PREDICTED: Lontra canadensis 18S ribosomal RNA (LOC116862838), rRNA	492	492	70%	7e-135	96.37%	XR_004381726.1
□	PREDICTED: Phoca vitulina 18S ribosomal RNA (LOC116629571), rRNA	492	492	70%	7e-135	96.37%	XR_004298788.1
□	PREDICTED: Phoca vitulina 18S ribosomal RNA (LOC116629238), rRNA	492	492	70%	7e-135	96.37%	XR_004298722.1
□	PREDICTED: Sapajus apella 18S ribosomal RNA (LOC116547631), rRNA	492	492	70%	7e-135	96.37%	XR_004267125.1
□	PREDICTED: Sapajus apella 18S ribosomal RNA (LOC116546318), rRNA	492	492	70%	7e-135	96.37%	XR_004266669.1
□	Staphylococcus aureus strain WH9628 chromosome	492	492	70%	7e-135	96.37%	CP033086.1
□	Lutra lutra genome assembly, chromosome: 16	492	492	70%	7e-135	96.37%	LR738418.1
□	Homo sapiens LHRI_LNC27.4 lncRNA gene, complete sequence	492	492	70%	7e-135	96.37%	MN297852.1
□	Homo sapiens LHRI_LNC27.1 lncRNA gene, complete sequence	492	492	70%	7e-135	96.37%	MN297850.1
□	Homo sapiens LHRI_LNC686.7 lncRNA gene, complete sequence	492	492	70%	7e-135	96.37%	MN297848.1

A word on reference databases and available packages

The agony of choice

- Several available
  - SILVA
  - RDP
  - RefSeq
  - ...
- Choices have to be made
  - completeness vs. duplications vs. curation
  - update frequency / still maintained?
  - only higher taxa vs. manual adaptation
  - personal preference and missing benchmarking
- General problems
  - "Legacy" classification / intermediate ranks
  - Different ID systems
  - Incomplete Information

Our Reference Database

- 510’984 reference sequences in SILVA 138 NR99

  → reduce to amplicon range
  - 8’387 below min length
  - 44 above max length
  - 50‘707 deduplication
  451‘846 reference sequences

- 10’259 higher taxa in SILVA 138 NR99

  + 2‘432 simplify taxonomic rank system
  + 78’191 unique species
  - 684 no reference after deduplication
  90‘199 taxa

- → and still ambiguos classifications (people BLAST)

Comparison in numbers

OTU

- General
  - ~24 h computation
  - 60%-78% assignment
- Mock community
  - **136-220 taxa**
  - 15%-16% correct assignment
  - diversity H: 2.53-2.67
  - diversity λ: 0.86
  - Eveness: 0.49-0.50
- Clinical samples
  - stool: 345-560 taxa
  - lung: 142-196 taxa
  - ovary: 593 taxa

k-mer

- General
  - **~10 min computation**
  - **97%-100% assignment**
- Mock community
  - 600-900 taxa
  - 21%-26% correct assignment
  - diversity H: 2.96-3.05
  - diversity λ: 0.88-0.89
  - Eveness: 0.44-0.46
- Clinical samples
  - stool: 1100-1700 taxa
  - lung: 600-900 taxa
  - ovary: 1300 taxa

align

- General
  - ~2 h computation
  - 65%-86% assignment
- Mock community
  - **120-236 taxa**
  - **65%-78% correct assignment**
  - diversity H: 2.60-2.93
  - diversity λ: 0.90-0.92
  - **Eveness: 0.53-0.54**
- Clinical samples
  - stool: 298-575 taxa
  - lung: 103-184 taxa
  - ovary: 353 taxa