



# Meta-Proteogenomics in KNIME

**Kerstin Neubert**



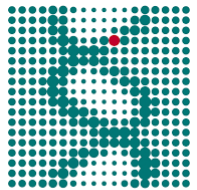
# Ess-BAR project

**B**iologische Gefahren, **A**nalyse und **R**esilienz der Lebensmittelwarenketten

**B**iological hazards, **A**nalysis and **R**esilience of the food supply chain

- I. Development of methods for the early identification of highly pathogenic species in the food supply chain using Meta-genomics (NGS) and high-resolution mass spectrometry (Orbitrap)
- II. Development of tools for data mining to integrate results from different laboratory methods, epidemiological data and associated metadata (production chains, food delivery)

# Ess-BAR project consortium



**MPIMG**

Max Planck Institute for Molecular Genetics

LC-MS proteomics



Bundesinstitut für Risikobewertung

**Project coordination**

Specialized diagnostics/  
Epidemiology

FRIEDRICH-LOEFFLER-INSTITUT



Bundesforschungsinstitut für Tiergesundheit  
Federal Research Institute for Animal Health

Specialized diagnostics/  
Forensic microbiology

**PolyAn**

molecular  
surface  
engineering

Rapid diagnostics  
(phage-based)

Freie Universität  Berlin

Software & Pipeline  
development



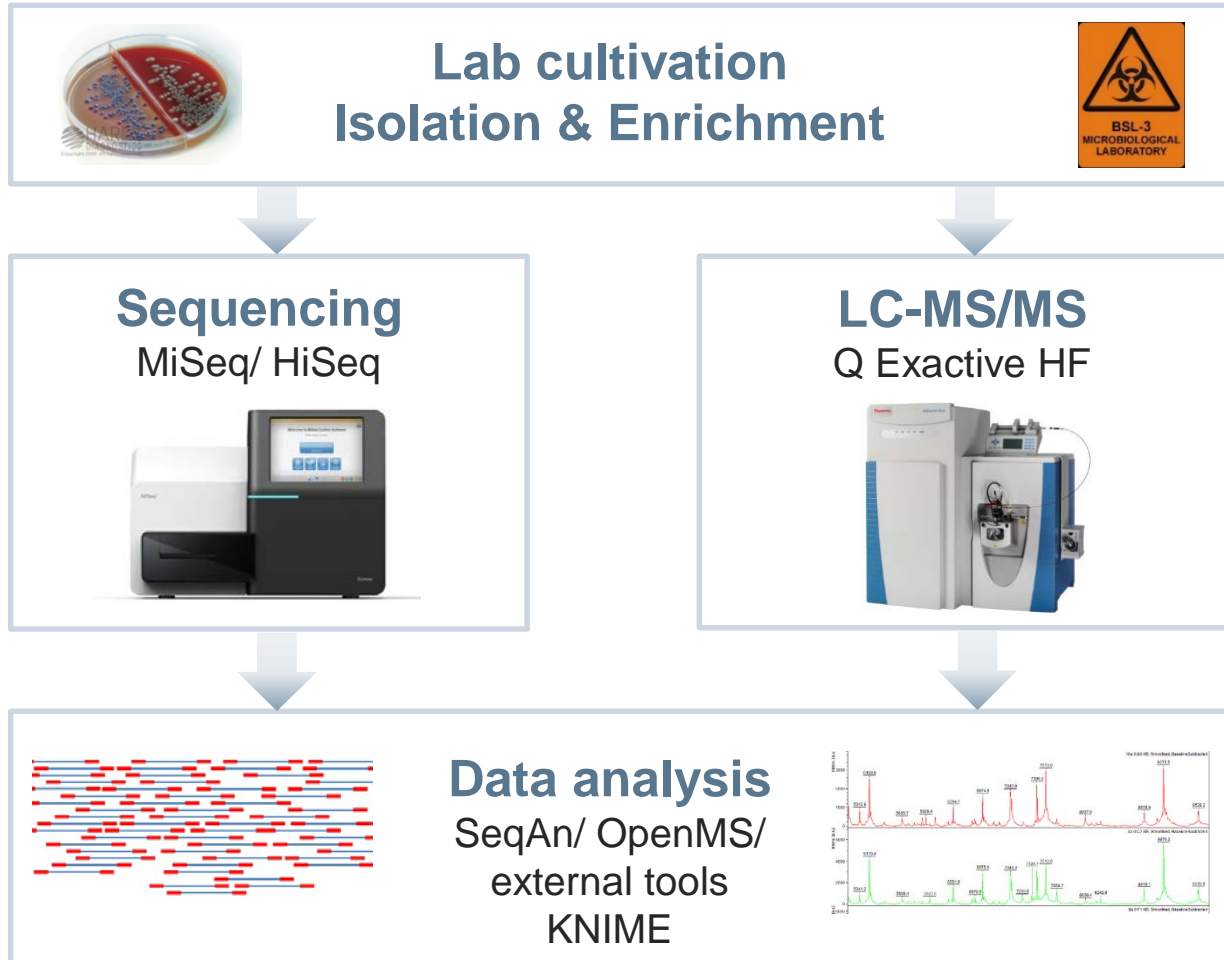
KNIME extensions  
Data mining & Epidemiology

# Three exemplified high risk pathogens

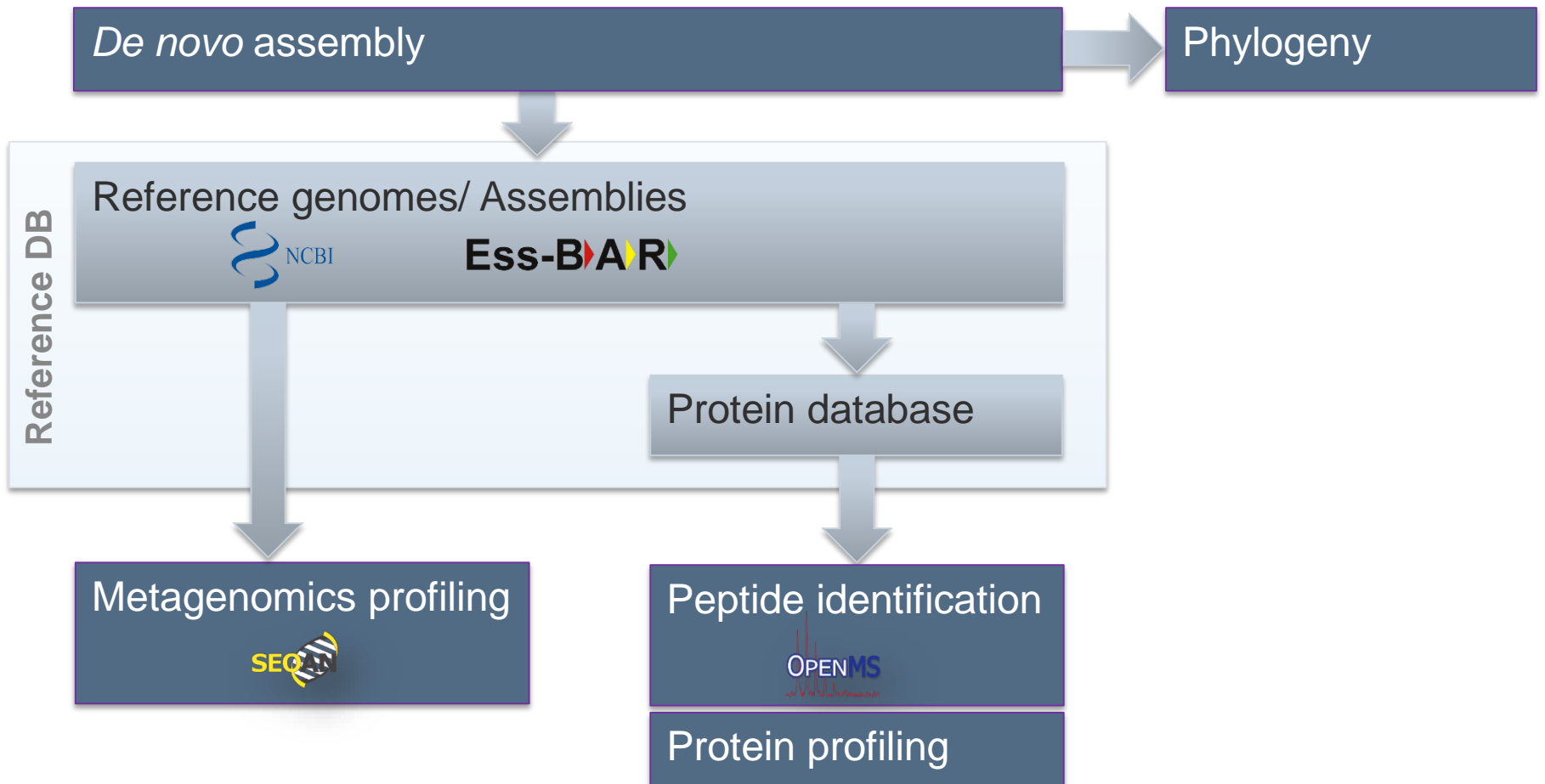
Pathogen	Food	CDC category
<i>Brucella spp.</i>	Milk products	High (B)
<i>Francisella tularensis</i>	Meat products	Very high (A)
<i>Bacillus anthracis</i>	Fruits and vegetables	Very high (A)



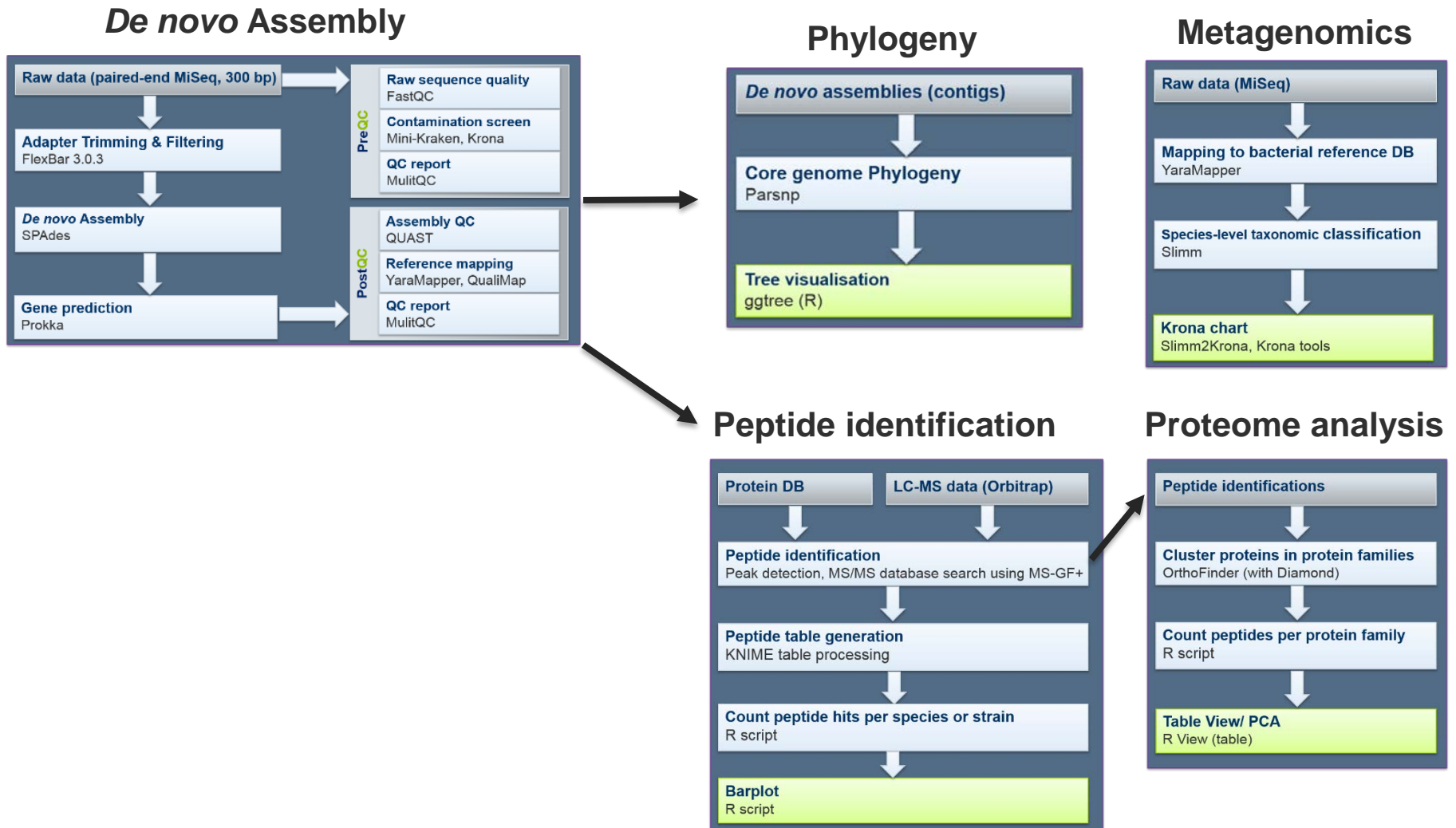
# Experimental setup



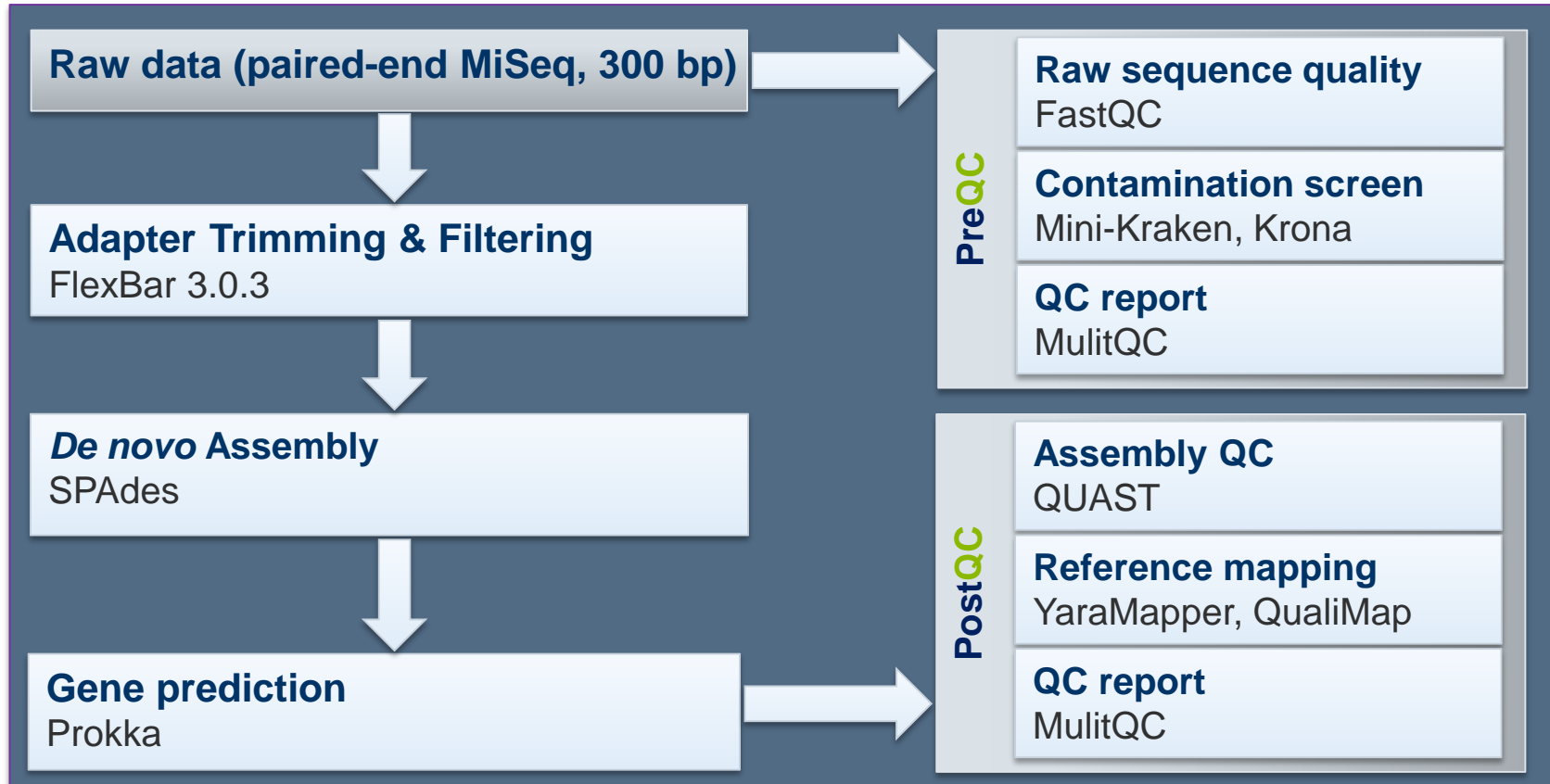
# Data analysis pipelines



# Pipeline overview



# De novo assembly pipeline





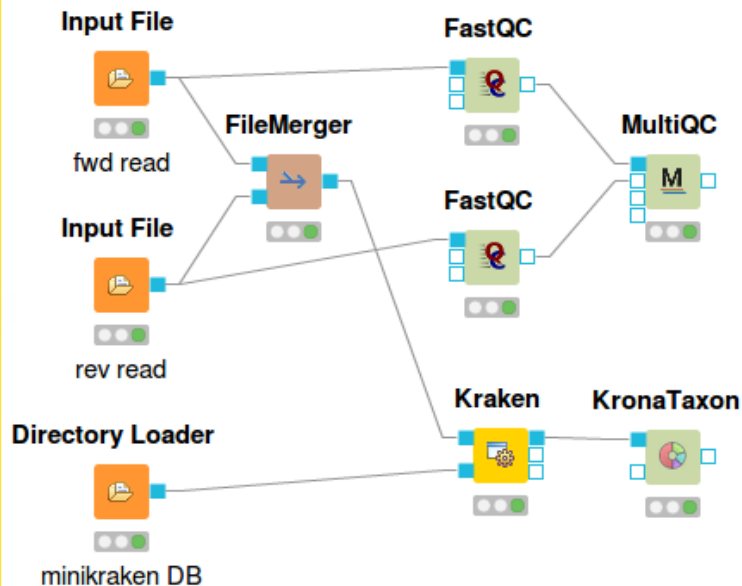
# De novo assembly

## PreQC

Quality of raw data (FastQC)

Screening for species (Kraken + Krona)

Summary (MultiQC)

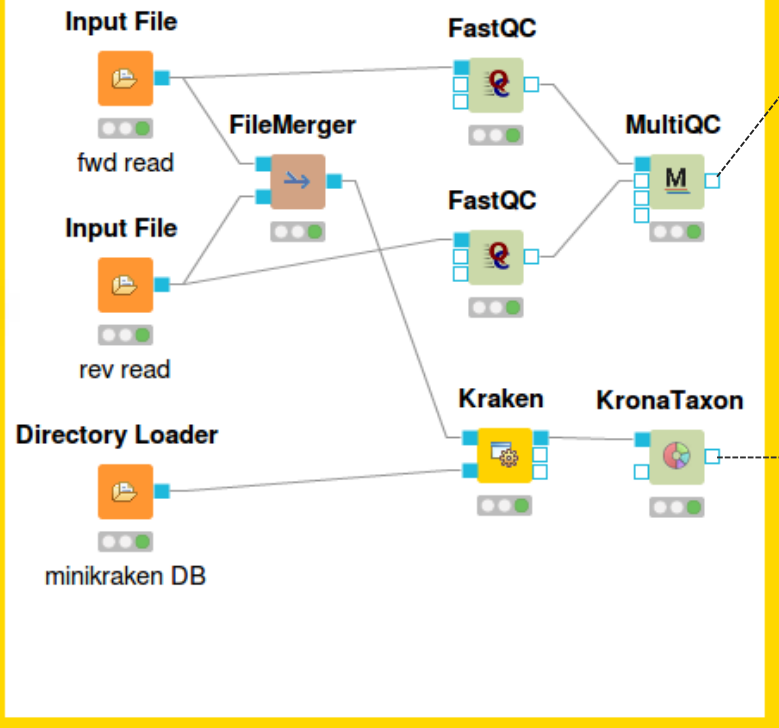


Node	Function
FastQC	Raw data QC: Sequence quality
Kraken	Raw data QC: contamination screen with <i>minikraken-DB</i>
KronaTaxon	Krona plot visualisation of taxonomic classification
MultiQC	Summary of QC results
Flexbar3	Preprocessing of NGS reads
SPAdes	De novo Assembly
Prokka	Assembly annotation
QUAST	Assembly QC

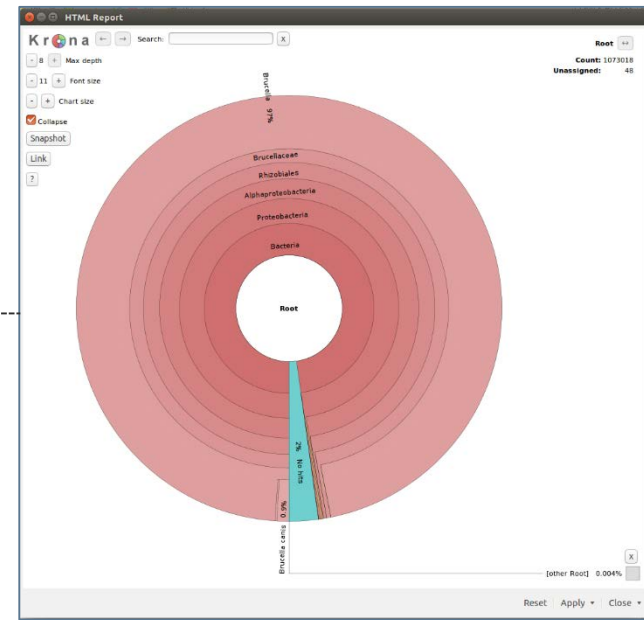
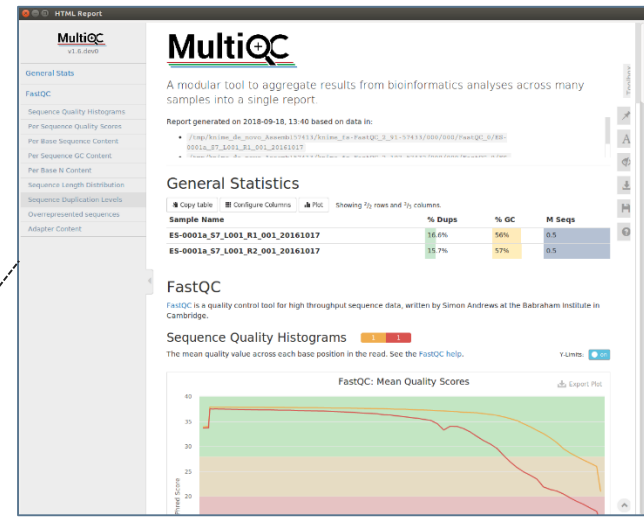
# De novo assembly

## PreQC

Quality of raw data (FastQC)  
 Screening for species (Kraken + Krona)  
 Summary (MultiQC)



## HTML Views



# De novo assembly



# De novo assembly



### QUAST

Quality Assessment Tool for Genome Assemblies by CAB

19 September 2018, Wednesday, 01:33:23

[View in Icarus contig browser](#)

All statistics are based on contigs of size  $\geq 500$  bp, unless otherwise noted (e.g. "# contigs ( $\geq 0$  bp)" and "Total length ( $\geq 0$  bp)" include all contigs).

**Statistics without reference**  ES-0002a\_10kb\_region\_R1

# contigs	1
# contigs ( $\geq 0$ bp)	1
# contigs ( $\geq 1000$ bp)	1
Largest contig	10589
Total length	10589
Total length ( $\geq 0$ bp)	10589
Total length ( $\geq 1000$ bp)	10589
N50	10589
N75	10589
L50	1
L75	1
GC (%)	58.61

**Mismatches**

# N's	0
# N's per 100 kbp	0

### MultiQC

v1.6.dev0

General Stats

QualiMap

Coverage histogram

Cumulative genome coverage

Insert size histogram

GC content distribution

QUAST

Assembly Statistics

Number of Contigs

Prokka

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2018-09-19, 01:34 based on data in:

- /tmp/ksime\_de\_novo\_assemb157413/ksime\_fa-QUAST\_3\_7-57449/000/000/QUAST\_0/caffoldb
- /tmp/ksime\_de\_novo\_assemb157413/ksime\_fa-Prokka\_2\_152-57471/000/000/Prokka\_0/caffoldb
- /tmp/ksime\_de\_novo\_assemb157413/ksime\_fa-Prokka\_2\_152-57471/000/000/Prokka\_0/caffoldb

**General Statistics**

Copy table | Configure Columns | Plot | Showing 1/1 rows and 10/10 columns.

Sample Name	% GC	Ins. size	$\approx$ 30X	Coverage	% Aligned	N50 (Kbp)	Length (Mb)
ES-0002a_10kb_region_R1	59%	447	98.9%	100.0X	92.8%	10.6Kbp	0.0Mb

**QualiMap**

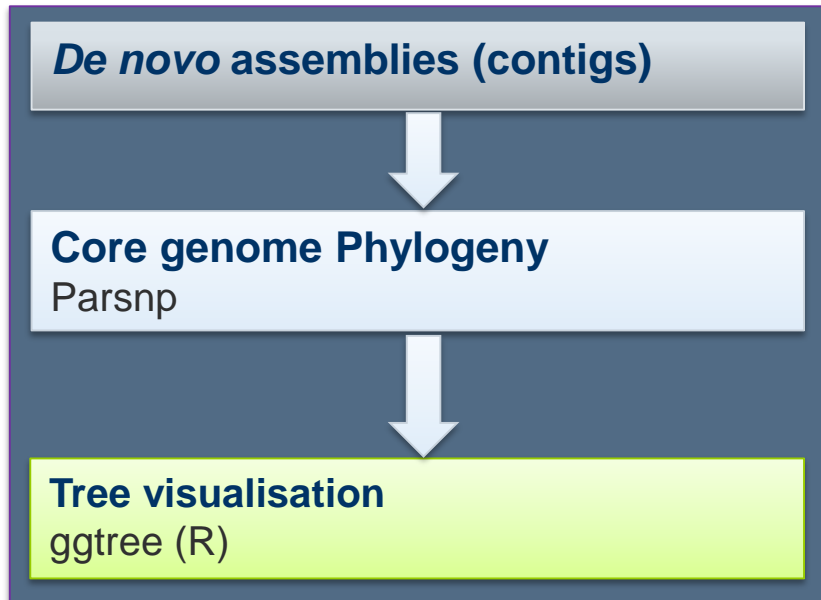
QualiMap is a platform-independent application to facilitate the quality control of alignment sequencing data and its derivatives like feature counts.

**Coverage histogram**

Distribution of the number of locations in the reference genome with a given depth of coverage.

Qualimap BamQC: Coverage histogram

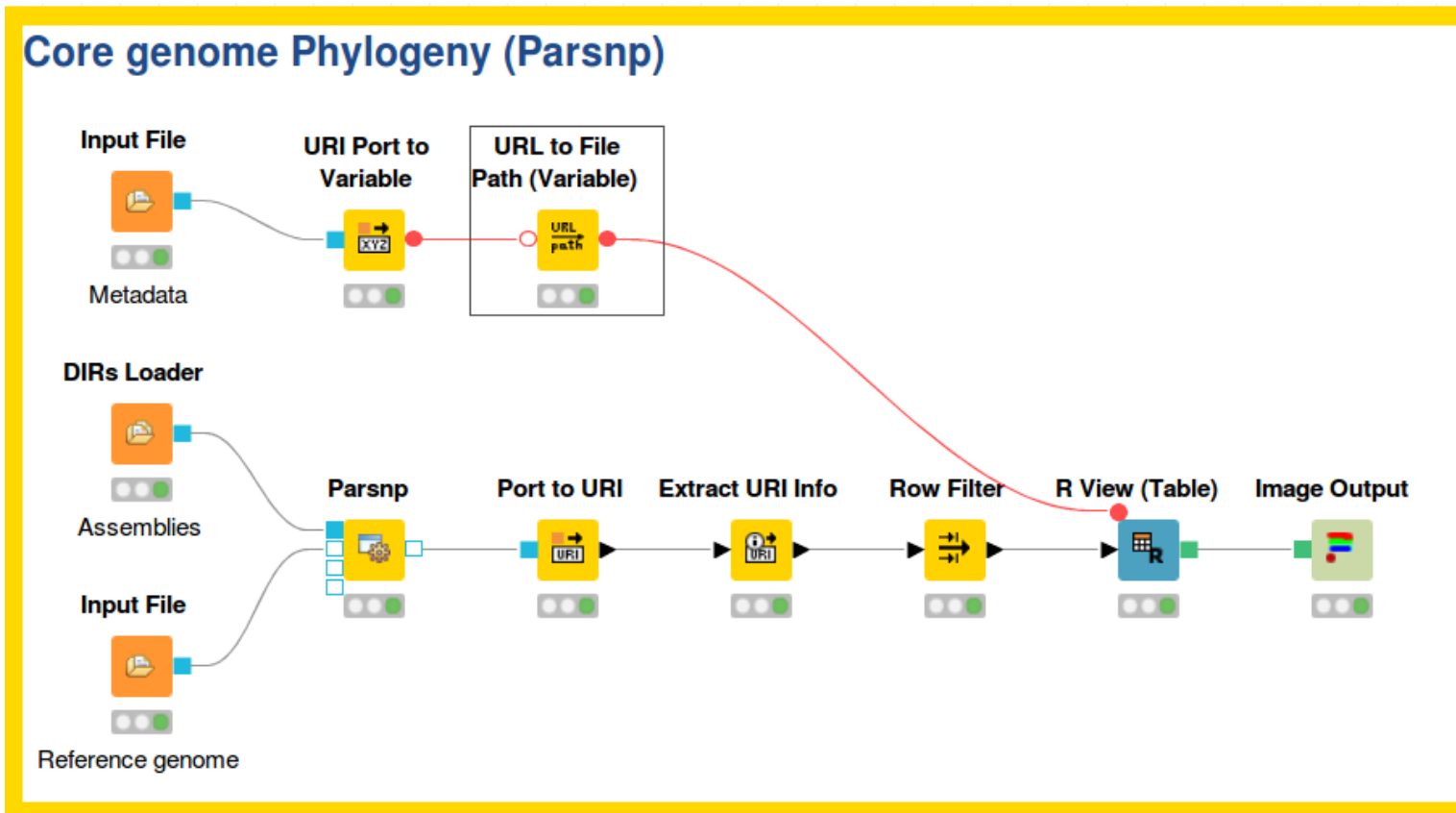
# Phylogenetic pipeline



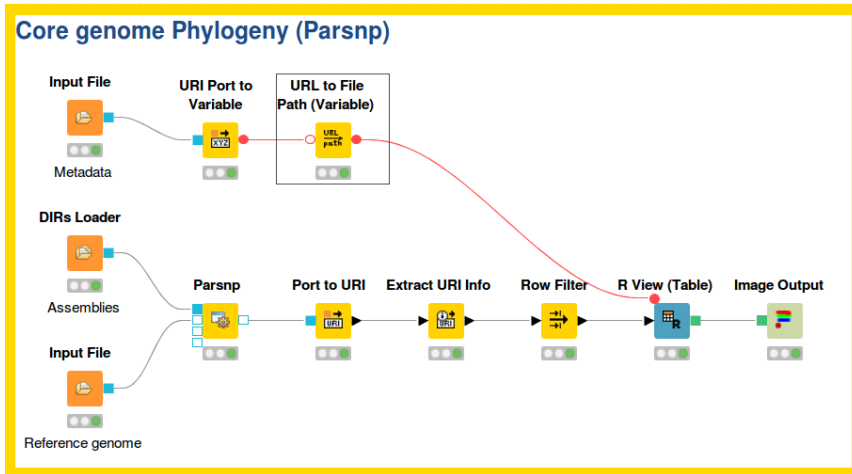
Treangen TJ\*, Ondov BD\*, Koren S, Phillippy AM: Rapid Core-Genome Alignment and Visualization for Thousands of Microbial Genomes. bioRxiv (2014). doi: <http://dx.doi.org/10.1101/007351>

# Phylogenetic pipeline

## Core genome Phylogeny (Parsnp)

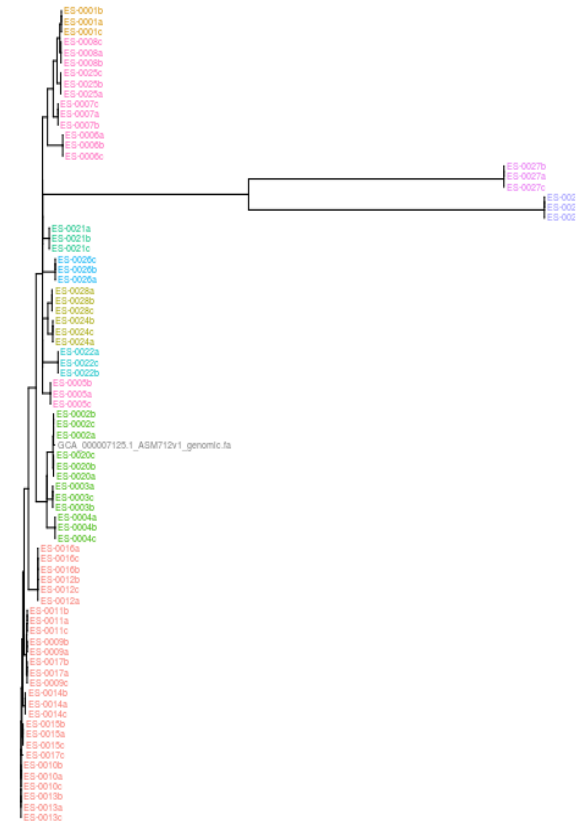


# Phylogenetic pipeline



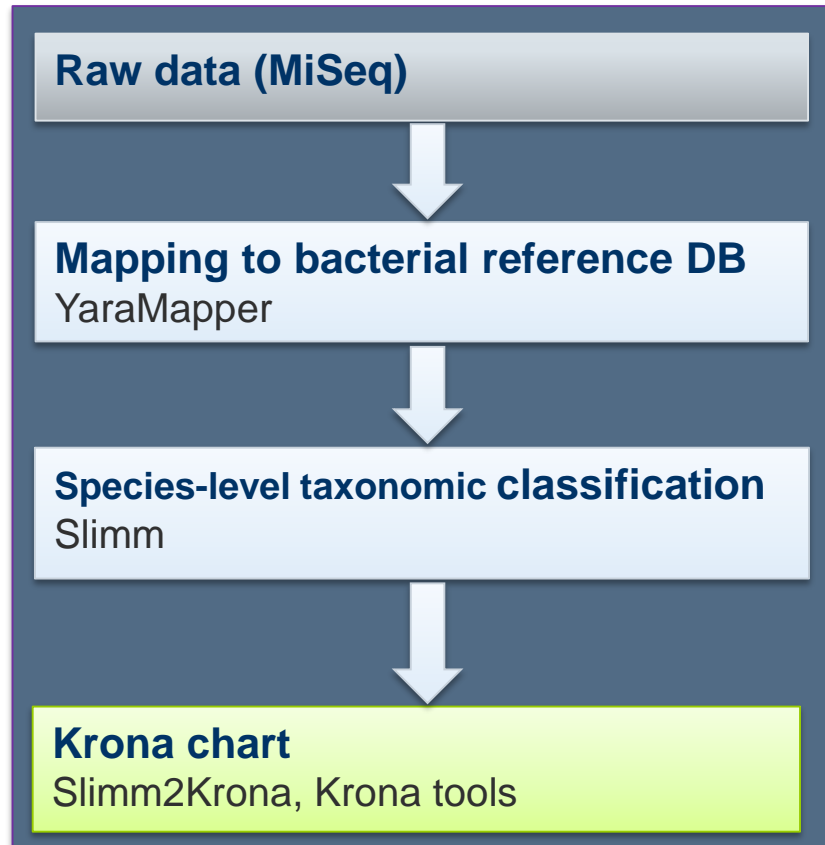
## species

- a Brucella abortus
- a Brucella canis
- a Brucella ceti
- a Brucella melitensis
- a Brucella microti
- a Brucella neotomae
- a Brucella papionis
- a Brucella pyxicephali
- a Brucella rodentia
- a Brucella suis
- a NA



node	Function
DIRs Loader	Load directories into a URI port object
Parsnp	Core genome alignment of assemblies and phylogeny
R View (Table)	Visualization of tree with ggtree

# Metagenomics pipeline



**Dadi TH, Renard B, Wieler LH, Semmler T, Reinert K. (2016)** SLIMM: Species level identification of microorganisms from metagenomes. PeerJ Preprints 4:e2378v1

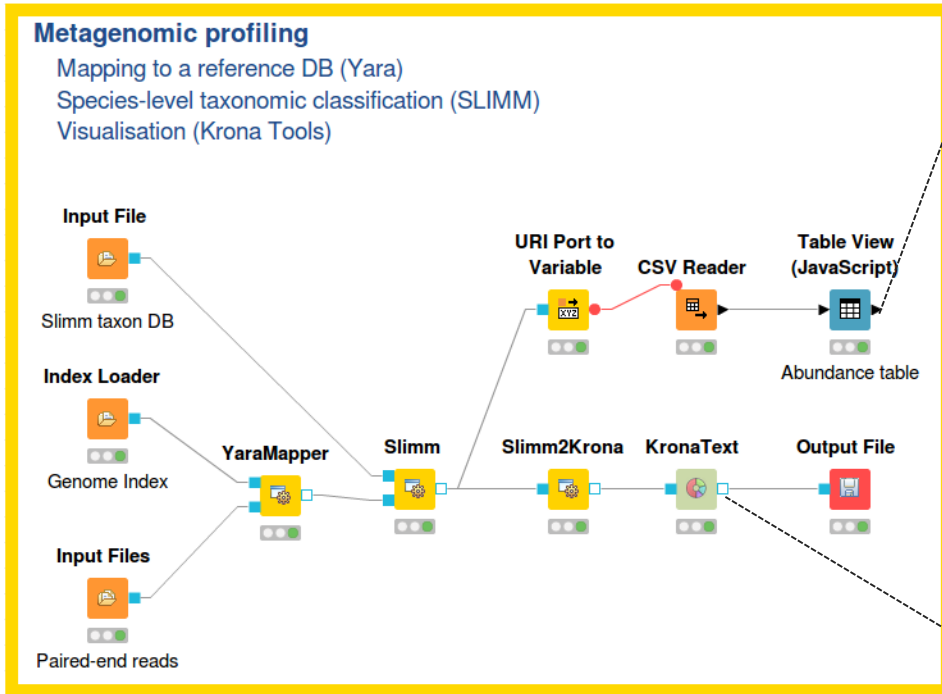
**Ondov BD, Bergman NH, and Phillippy AM.** Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011 Sep 30; 12(1):385.



# Metagenomics pipeline

## Metagenomic profiling

Mapping to a reference DB (Yara)  
 Species-level taxonomic classification (SLIMM)  
 Visualisation (Krona Tools)

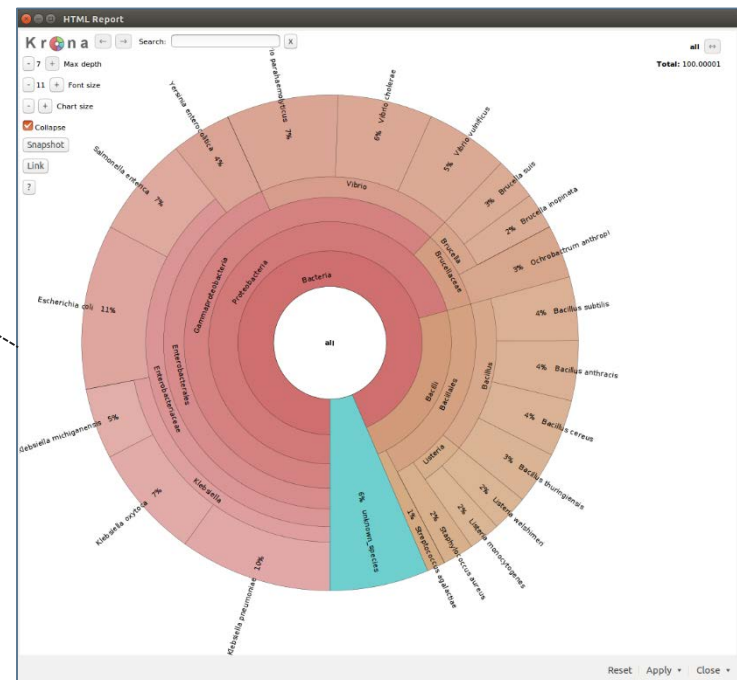


JavaScript Table View

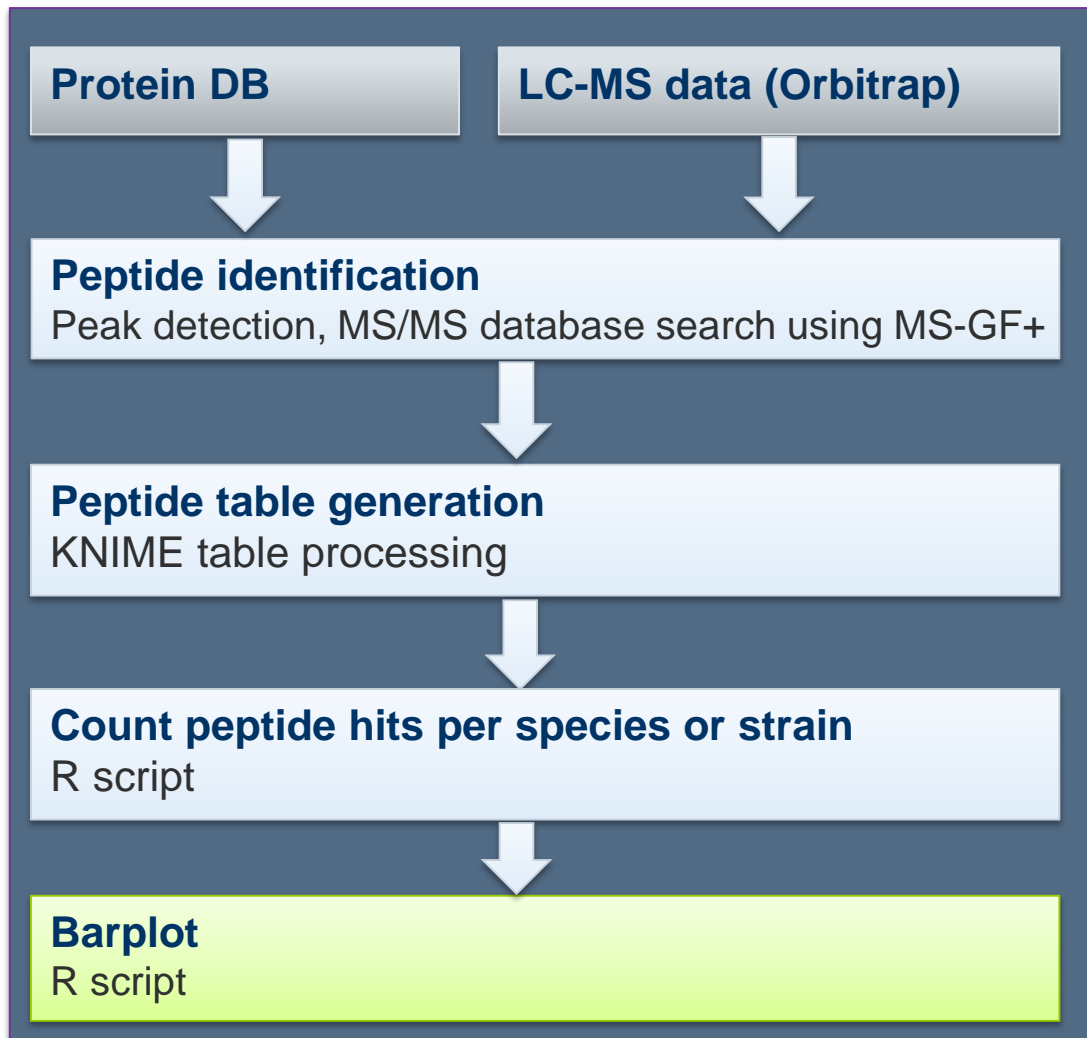
Show 25 entries

taxa_level	taxa_id	linage	abundance	read_count
species	1420	k__Bacteria p__Firmicutes c__Bacilli o__Bacillales f__Bacillaceae g__Bacillus s__Bacillus thuringiensis	3.4247	541
species	571	k__Bacteria p__Proteobacteria c__Gammaproteobacteria o__Enterobacteriales f__Enterobacteriaceae g__Klebsiella s__Klebsiella oxytoca	7.45078	1177
species	1392	k__Bacteria p__Firmicutes c__Bacilli o__Bacillales f__Bacillaceae g__Bacillus s__Bacillus anthracis	3.91847	619
species	573	k__Bacteria p__Proteobacteria c__Gammaproteobacteria o__Enterobacteriales f__Enterobacteriaceae g__Klebsiella s__Klebsiella pneumoniae	9.92394	1360
species	1134687	k__Bacteria p__Proteobacteria c__Gammaproteobacteria o__Enterobacteriales f__Enterobacteriaceae g__Klebsiella s__Klebsiella michiganensis	4.6148	729
species	670	k__Bacteria p__Proteobacteria c__Gammaproteobacteria o__Vibrionales f__Vibrionaceae g__Vibrio s__Vibrio parahaemolyticus	3.27353	1349
species	1210315	k__Bacteria p__Proteobacteria c__Alphaproteobacteria o__Rhizobiales f__Brucellales g__Brucellaceae s__Brucella s__Brucella inopinata	2.300	376
species	630	k__Bacteria p__Proteobacteria c__Gammaproteobacteria o__Enterobacteriales f__Yersiniaceae g__Yersinia s__Yersinia enterocolitica	3.93214	618
species	1639	k__Bacteria p__Firmicutes c__Bacilli o__Bacillales f__Liberiaceae g__Liberia s__Liberia monocytogenes	2.06368	326
species	342	k__Bacteria p__Proteobacteria c__Gammaproteobacteria o__Enterobacteriales f__Enterobacteriaceae g__Escherichia s__Escherichia coli	10.3362	1699
species	28901	k__Bacteria p__Proteobacteria c__Gammaproteobacteria o__Enterobacteriales f__Enterobacteriaceae g__Salmonella s__Salmonella enterica	6.60252	1043
species	1311	k__Bacteria p__Firmicutes c__Bacilli o__Lactobacillales f__Streptococcaceae g__Streptococcus s__Streptococcus agalactiae	1.1901	188
species	1643	k__Bacteria p__Firmicutes c__Bacilli o__Bacillales f__Liberiaceae g__Liberia s__Liberia amblyselae	2.38653	377
species	666	k__Bacteria p__Proteobacteria c__Gammaproteobacteria o__Vibrionales f__Vibrionaceae g__Vibrio s__Vibrio cholerae	6.16573	974
species	29401	k__Bacteria p__Proteobacteria c__Alphaproteobacteria o__Rhizobiales f__Brucellales g__Brucellaceae s__Brucella s__Brucella suis	2.77901	439
species	672	k__Bacteria p__Proteobacteria c__Gammaproteobacteria o__Vibrionales f__Vibrionaceae g__Vibrio s__Vibrio vulnificus	5.40409	834
species	1423	k__Bacteria p__Firmicutes c__Bacilli o__Bacillales f__Bacillaceae g__Bacillus s__Bacillus subtilis	4.1337	653
species	1336	k__Bacteria p__Firmicutes c__Bacilli o__Bacillales f__Bacillaceae g__Bacillus s__Bacillus cereus	3.70957	586
species	1280	k__Bacteria p__Firmicutes c__Bacilli o__Bacillales f__Staphylococcaceae g__Staphylococcus s__Staphylococcus aureus	2.03203	321
species	529	k__Bacteria p__Proteobacteria c__Alphaproteobacteria o__Rhizobiales f__Brucellales g__Brucellaceae s__Ochrobactrum s__Ochrobactrum anthropi	3.46268	547
species	?	k__unknown s__supernkingdom s__unknown p__phylum s__unknown c__class s__unknown o__order s__unknown f__family s__unknown g__genus s__unknown s__species	6.43161	1016

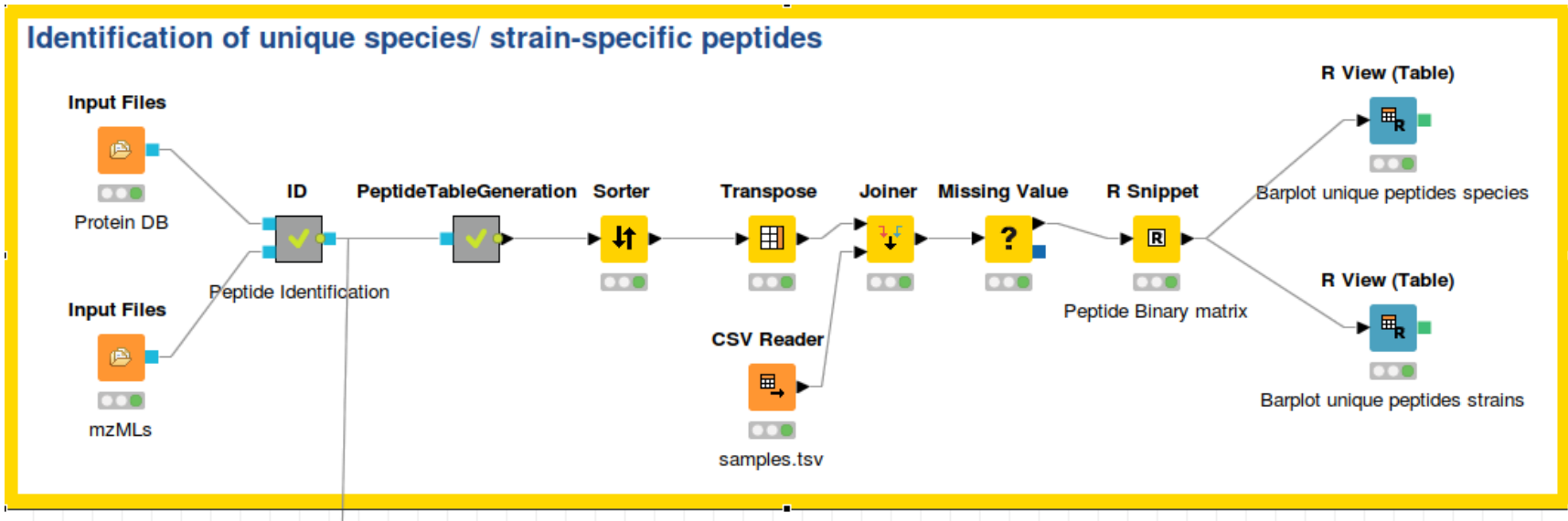
Showing 1 to 21 of 21 entries



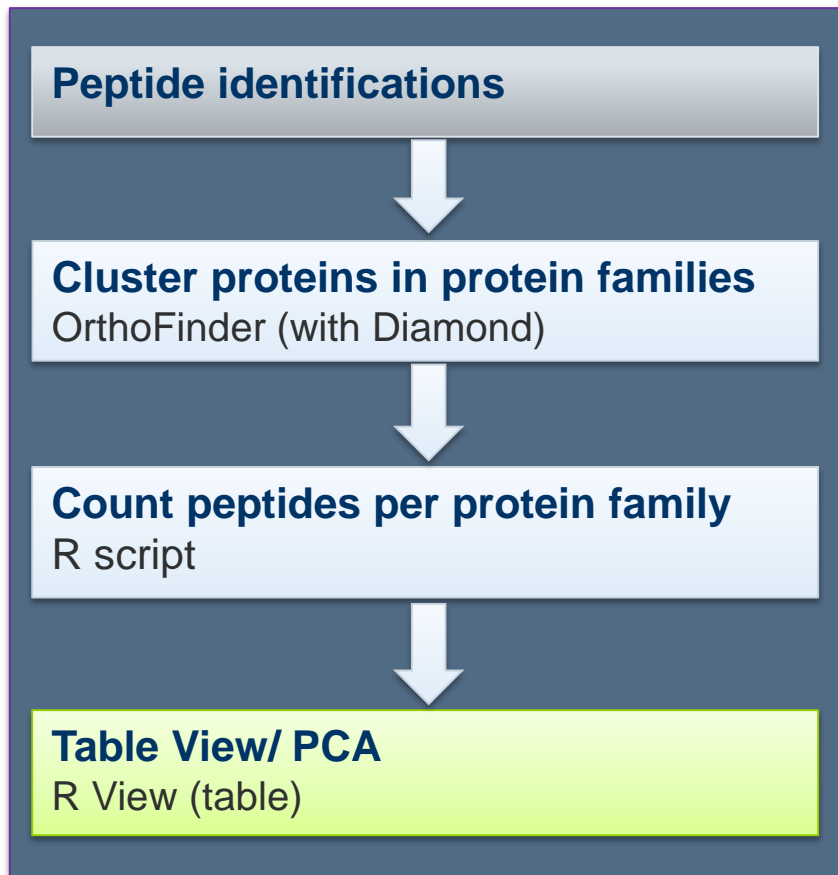
# Identification of species-specific peptides



# Identification of species-specific peptides



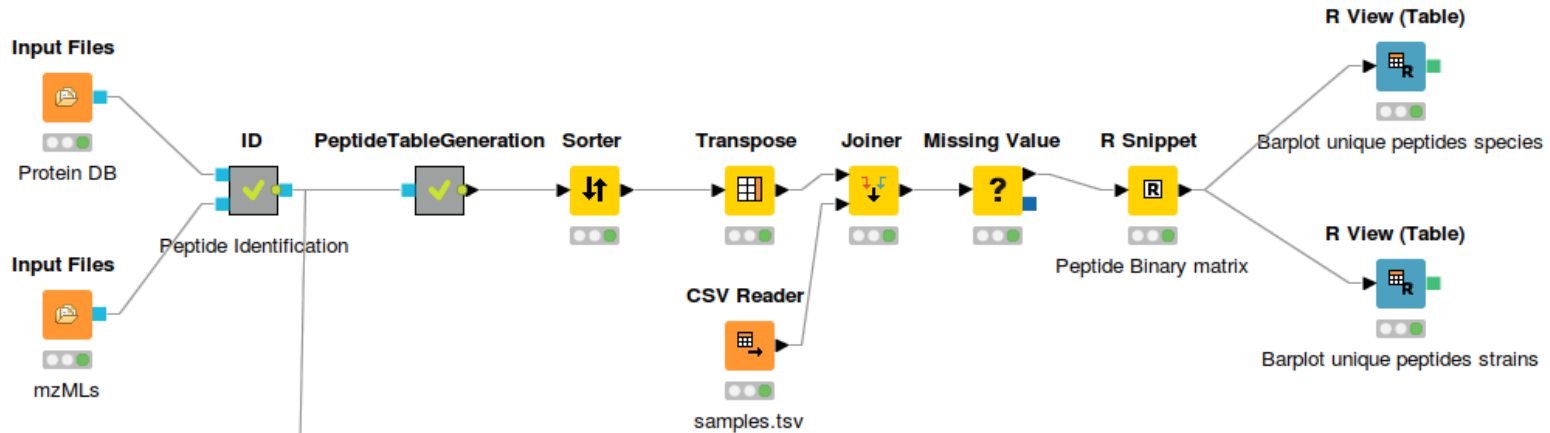
# Proteome analysis



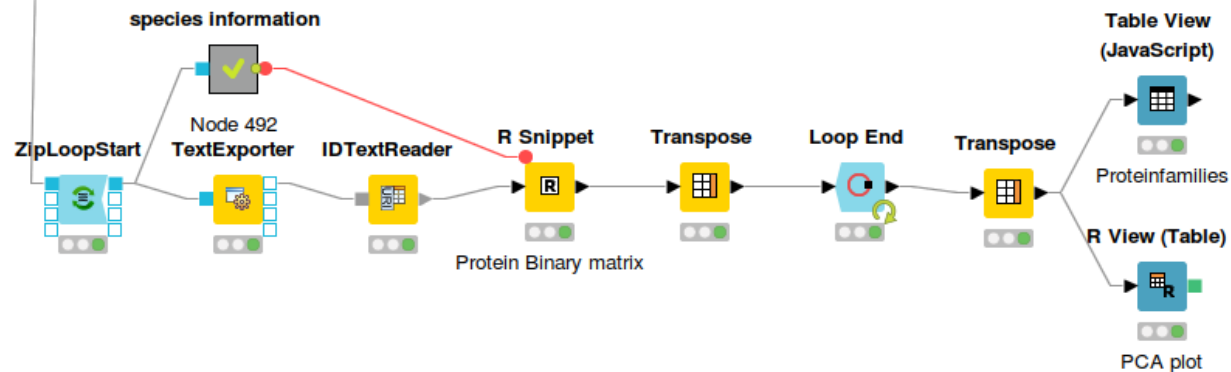
Emms, D.M. and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16:157

# Proteome pipelines

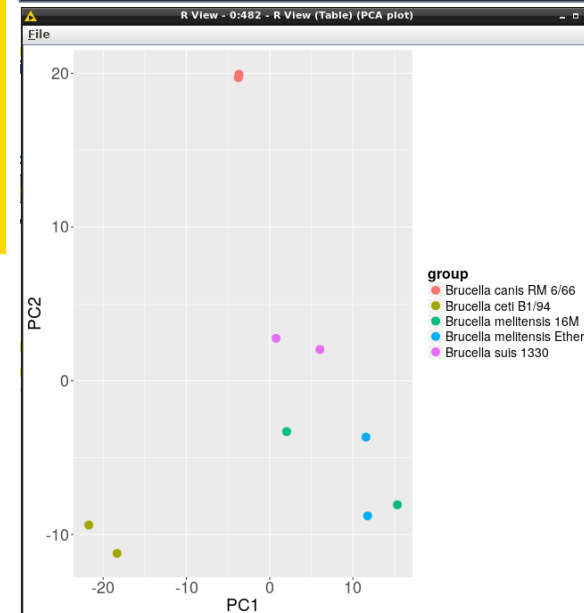
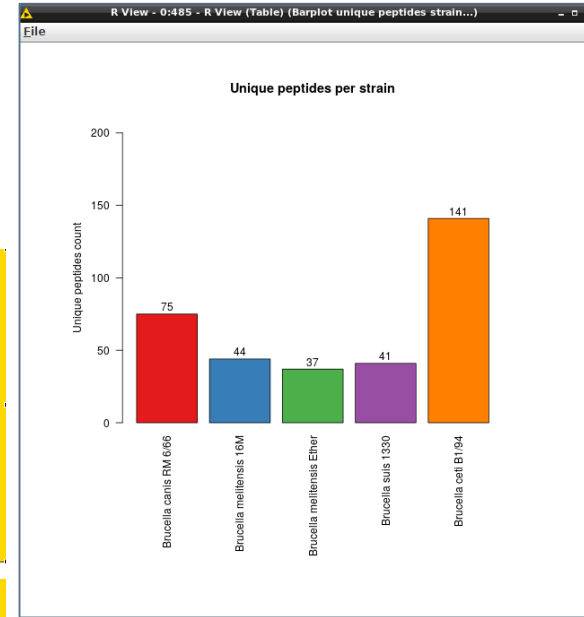
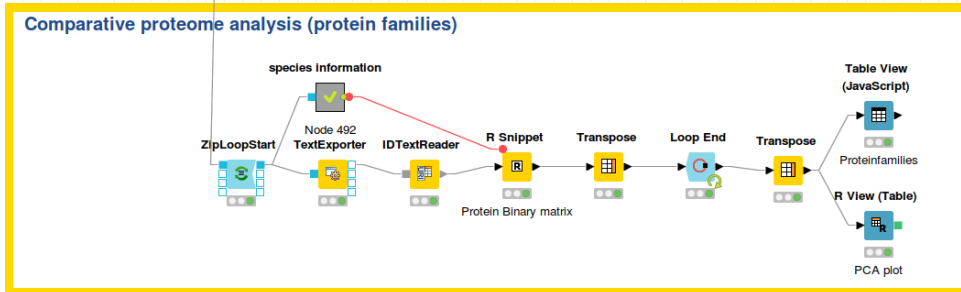
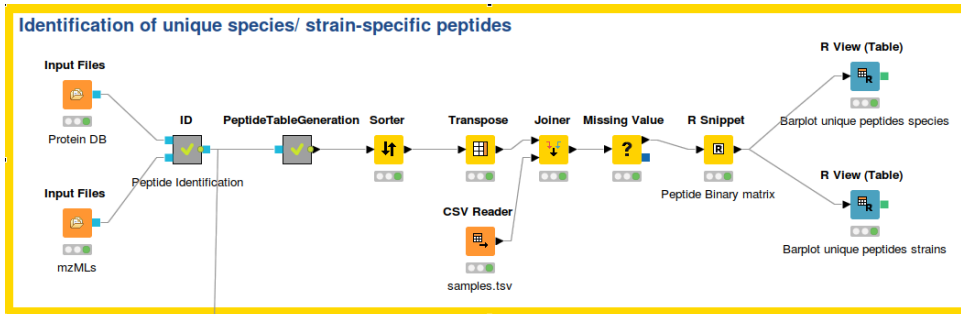
## Identification of unique species/ strain-specific peptides

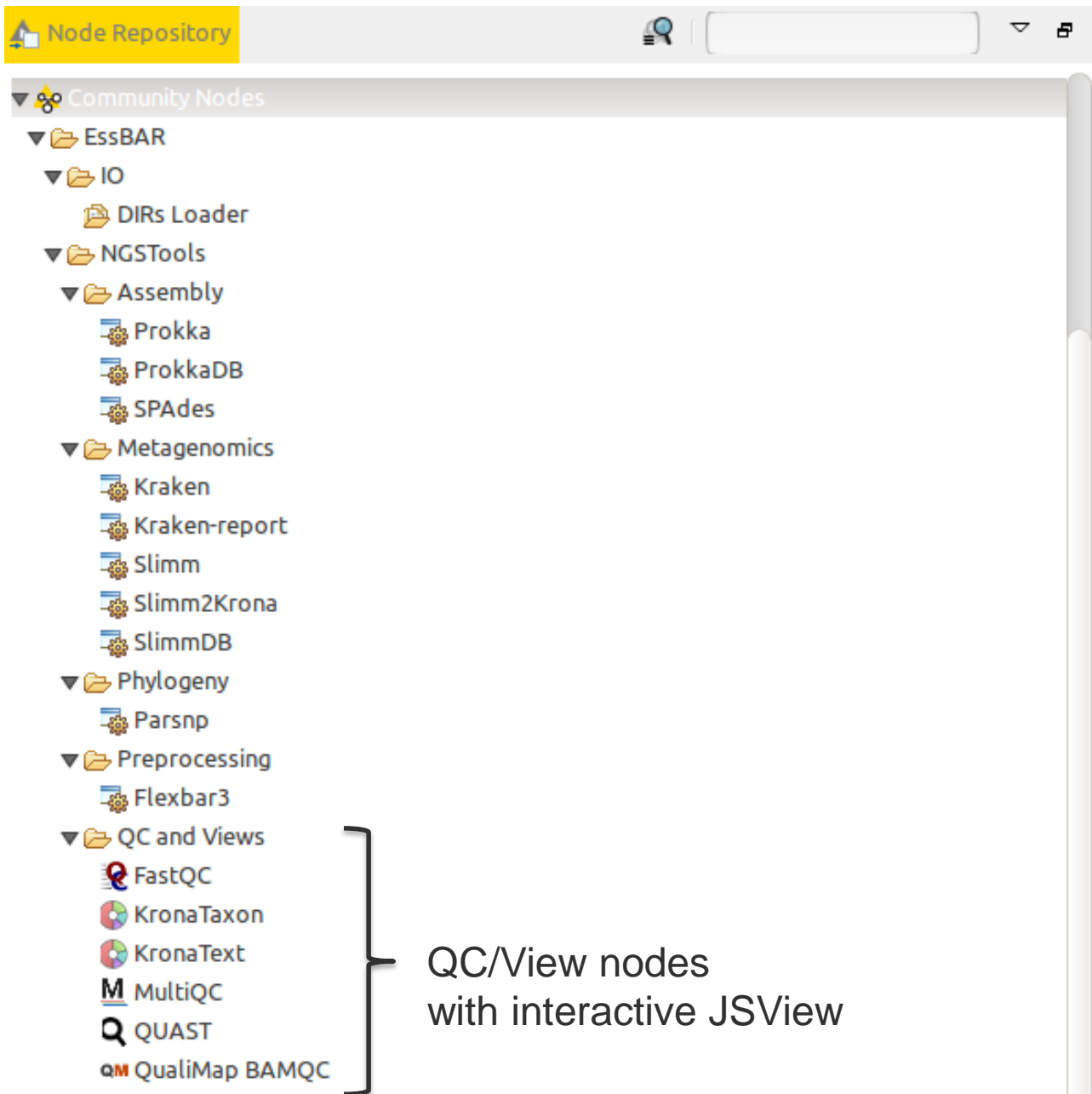


## Comparative proteome analysis (protein families)



# Proteome pipelines





Node Repository

- Community Nodes
  - EssBAR
    - IO
      - DIRs Loader
    - NGSTools
      - Assembly
        - Prokka
        - ProkkaDB
        - SPAdes
      - Metagenomics
        - Kraken
        - Kraken-report
        - Slimm
        - Slimm2Krona
        - SlimmDB
      - Phylogeny
        - Parsnp
      - Preprocessing
        - Flexbar3
      - QC and Views
        - FastQC
        - KronaTaxon
        - KronaText
        - MultiQC
        - QUAST
        - QualiMap BAMQC

} QC/View nodes  
with interactive JSView