Freie Universität Berlin

Ess-BAR
Biologische Gefahren, Analyse und Resilienz der Lebensmittelwarenketten

Meta-Proteogenomics in KNIME

Kerstin Neubert

Ess-BAR project

Biologische Gefahren, **A**nalyse und **R**esilienz der Lebensmittelwarenketten

Biological 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 labaratory methods, epidemiological data and associated metadata (production chains, food delivery)

Ess-BAR project consortium

MPIMG
Max Planck Institute for Molecular Genetics

LC-MS proteomics

BfR
Bundesinstitut für Risikobewertung

Project coordination

Specialized diagnostics/
Epidemiology

FRIEDRICH-LOEFFLER-INSTITUT
FLI
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

Open for Innovation
KNIME

KNIME extensions
Data mining & Epidemiology

Three exemplified high risk pathogens

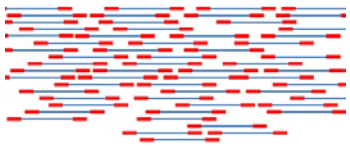
Pathogen	Food	CDC category
Brucella spp.	Milk products	High (B)
Francisella tularensis	Meat products	Very high (A)
Bacillus anthracis	Fruits and vegetables	Very high (A)

Experimental setup

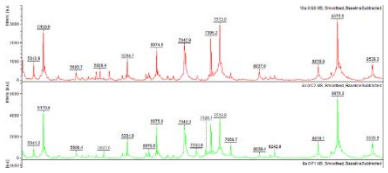
Lab cultivation
Isolation & Enrichment

Sequencing
MiSeq/ HiSeq

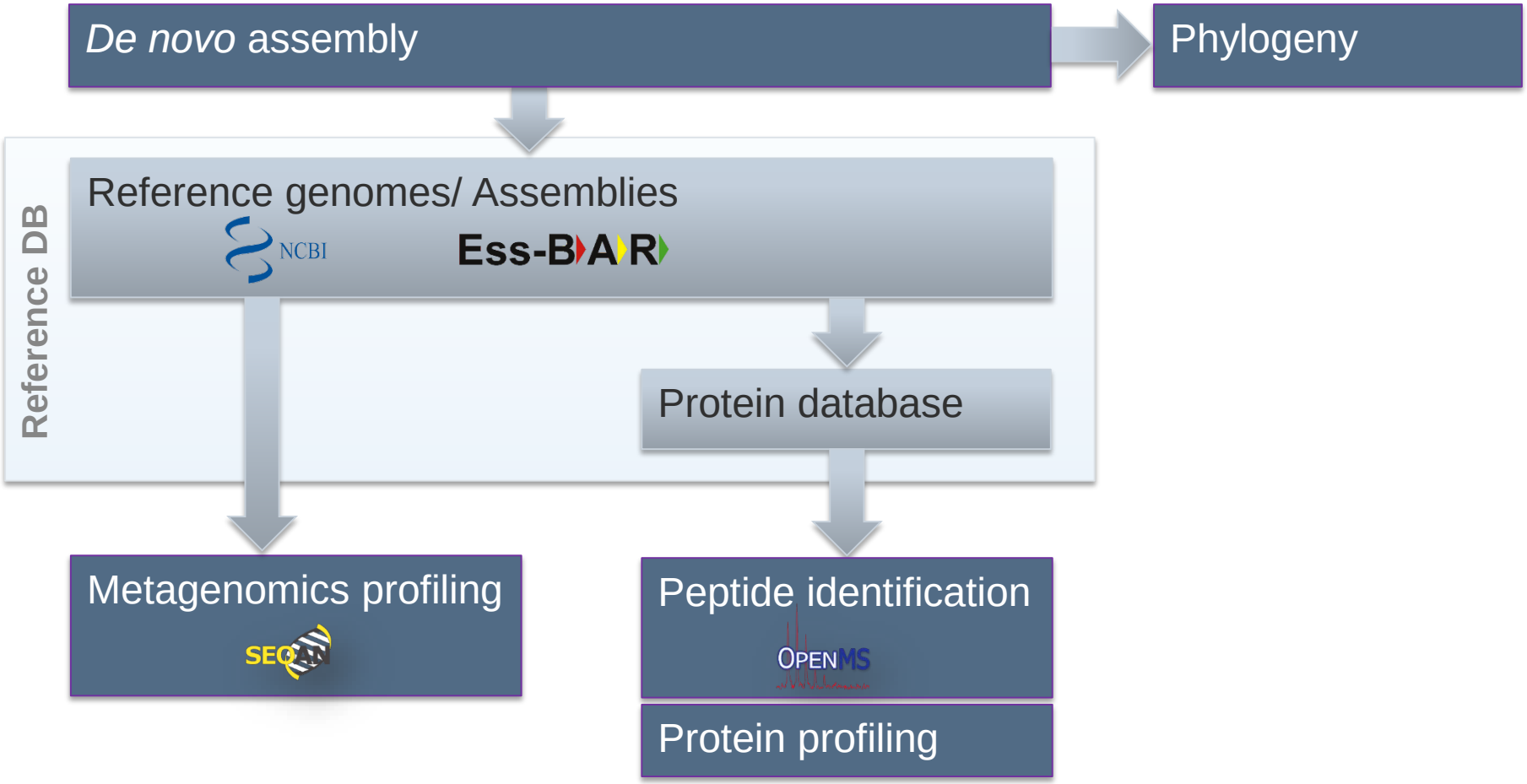
LC-MS/MS
Q Exactive HF



Data analysis
SeqAn/ OpenMS/
external tools
KNIME

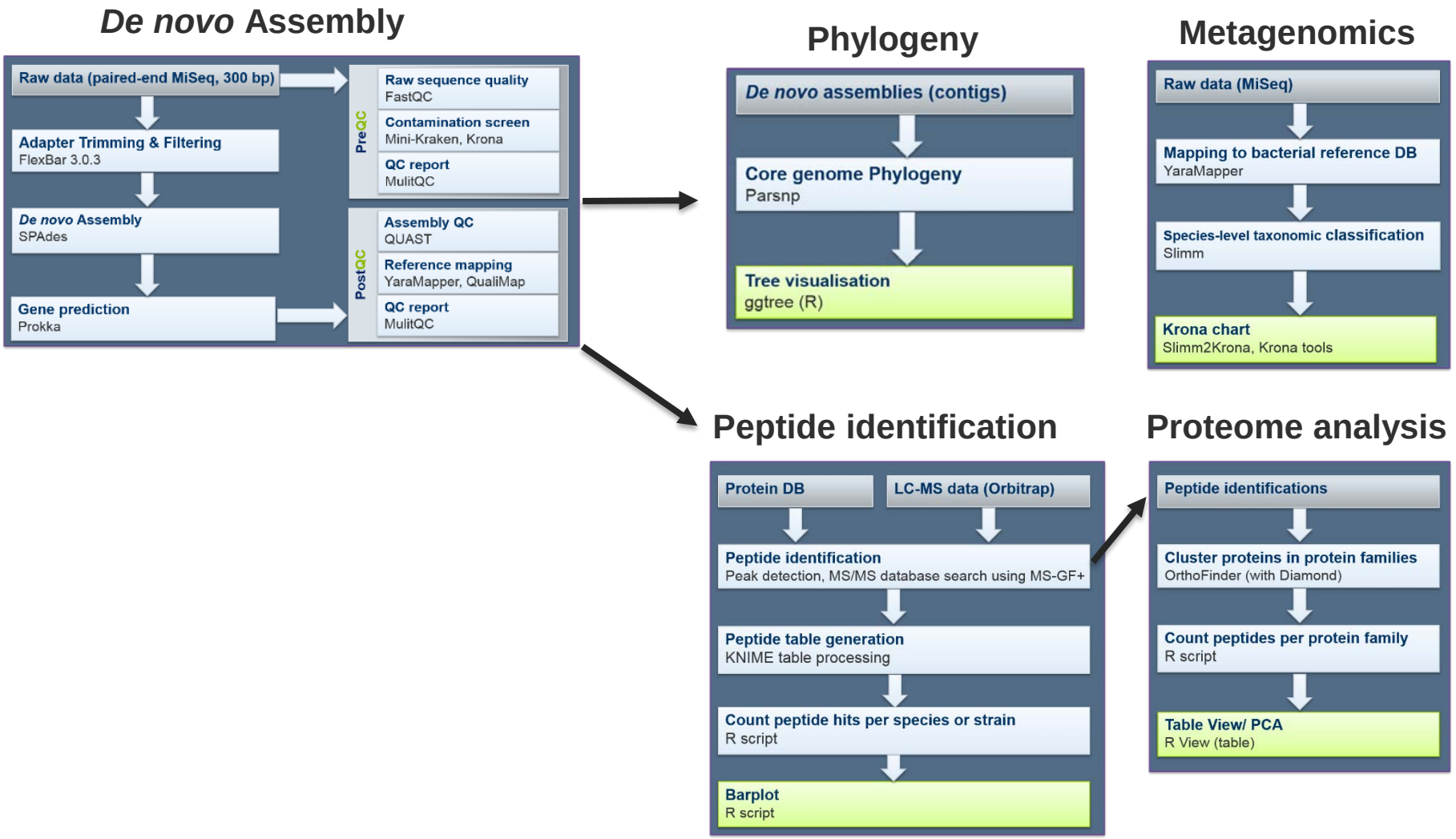


Data analysis pipelines

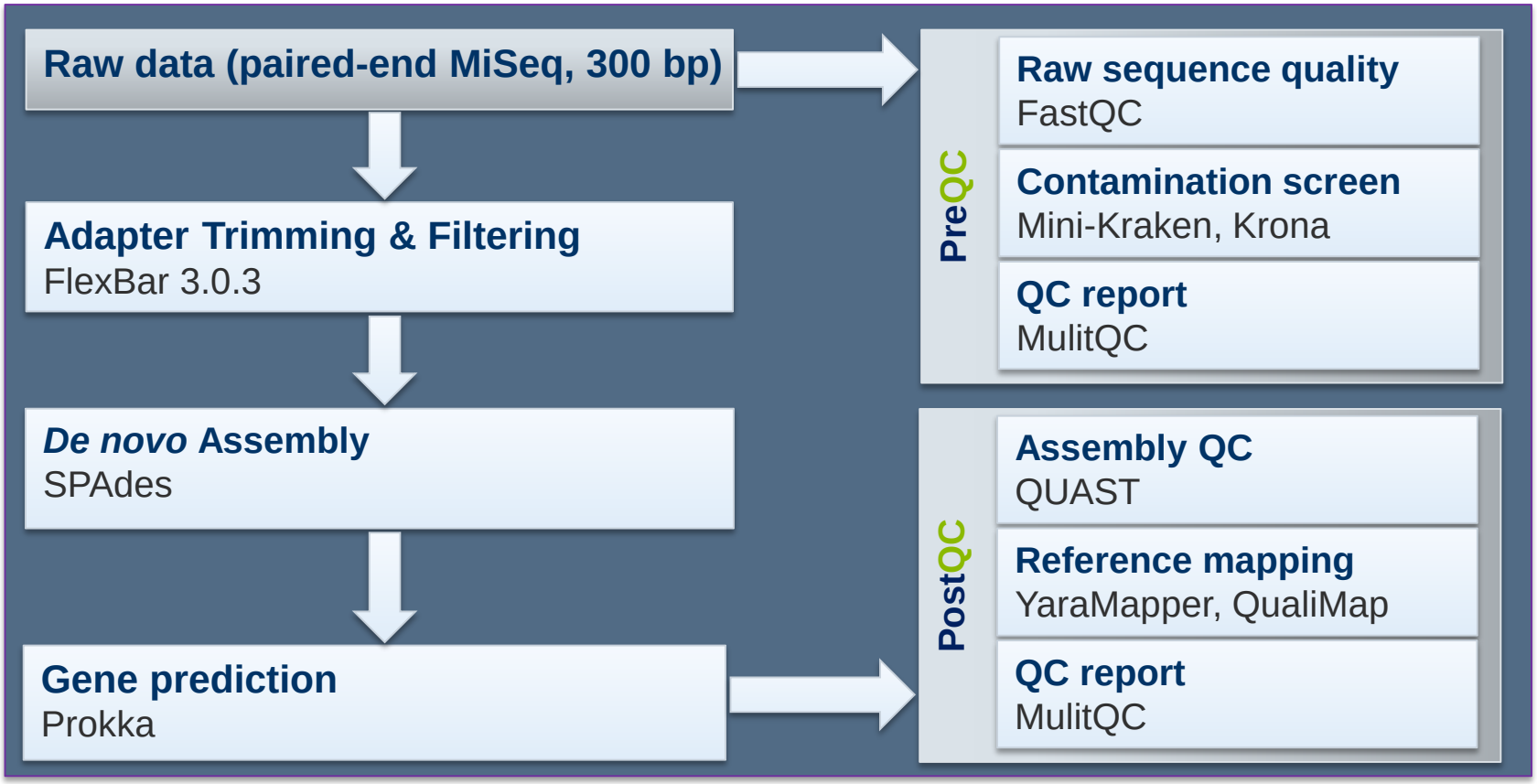


Open for Innovation
KNIME workflows

Pipeline overview



De novo assembly pipeline



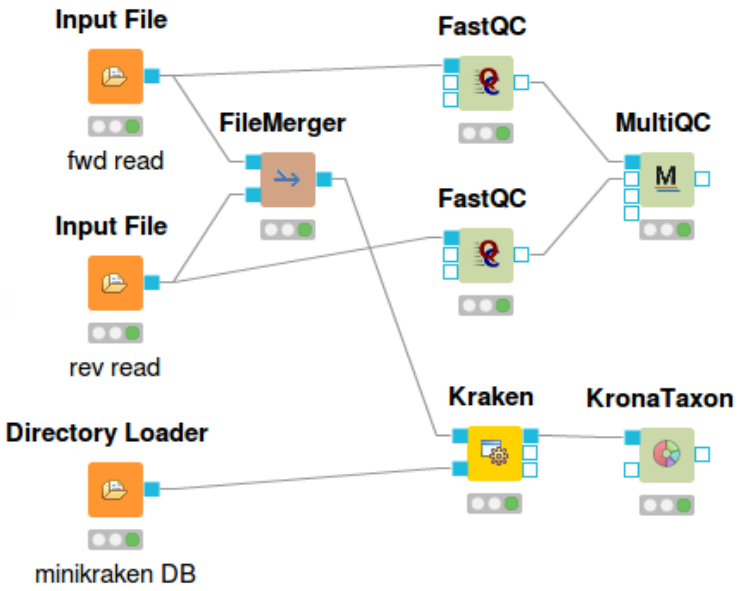
De novo assembly

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 *minikraken-DB*
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

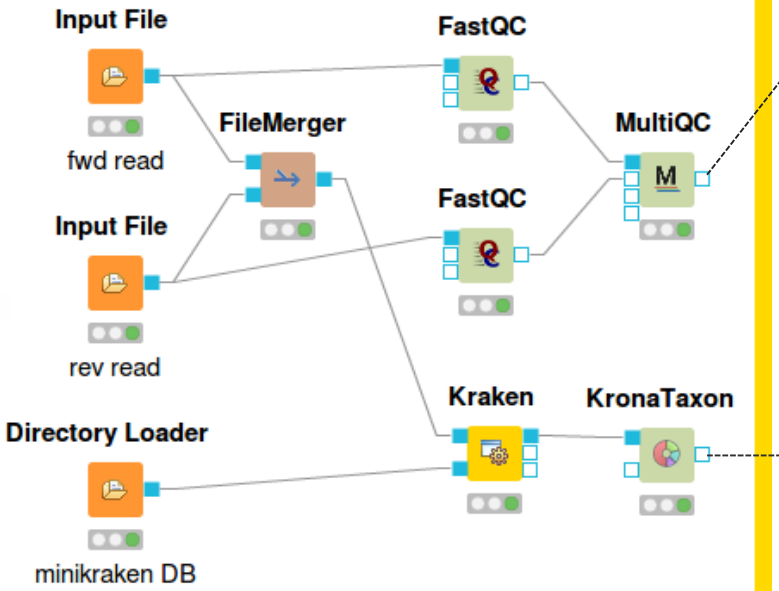
De novo assembly

PreQC

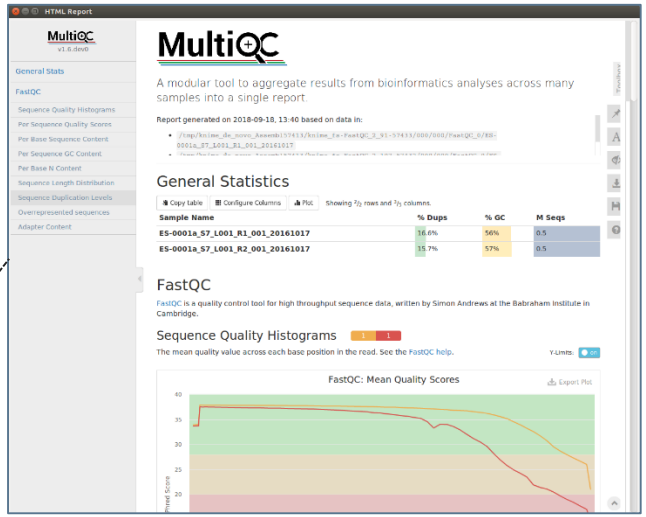
Quality of raw data (FastQC)

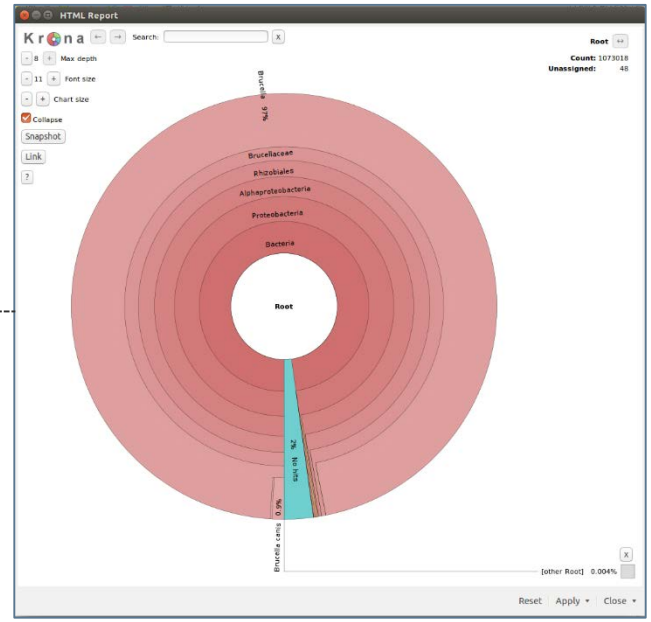
Screening for species (Kraken + Krona)

Summary (MultiQC)



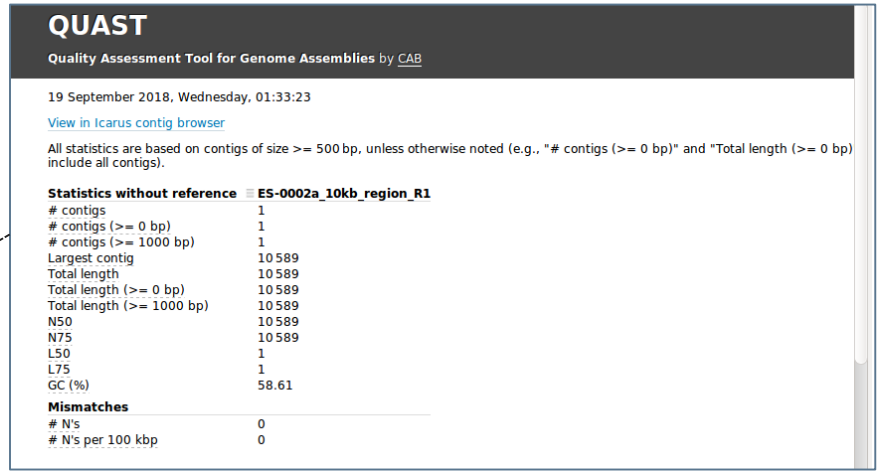
HTML Views





De novo assembly

De novo assembly



QUAST

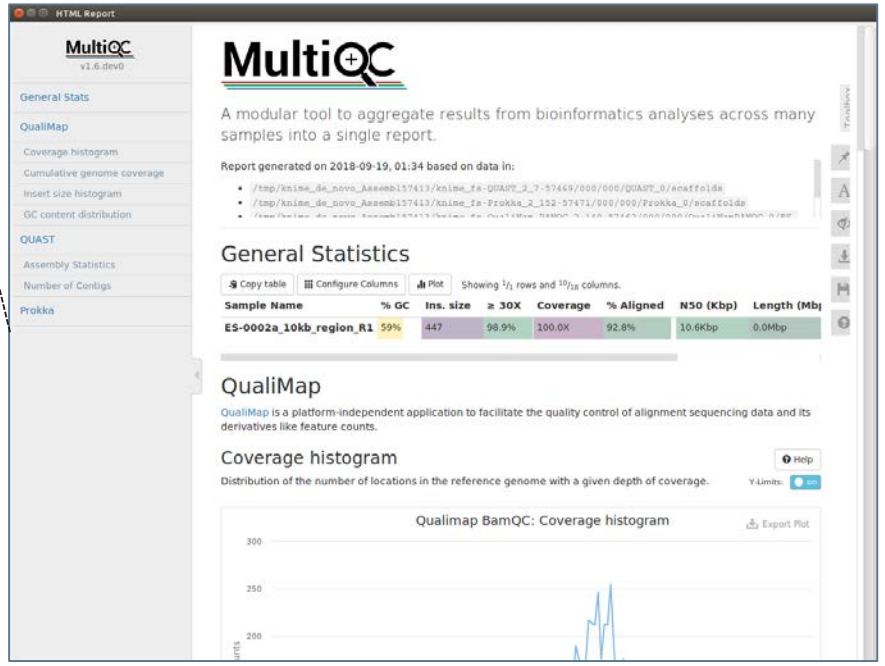
Quality Assessment Tool for Genome Assemblies by CAB

19 September 2018, Wednesday, 01:33:23

View in Icarus contig browser

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

Statistics without reference	ES-0002a_10kb_region_R1
# contigs	1
# contigs (>= 0 bp)	1
# contigs (>= 1000 bp)	1
Largest contig	10 589
Total length	10 589
Total length (>= 0 bp)	10 589
Total length (>= 1000 bp)	10 589
N50	10 589
N75	10 589
L50	1
L75	1
GC (%)	58.61
Mismatches	
# N's	0
# N's per 100 kbp	0



HTML Report

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

MultiQC

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/knime_de_novo_Assemb157413/knime_fs-QUAST_3_7-57469/000/000/QUAST_0/scaffolds
- /tmp/knime_de_novo_Assemb157413/knime_fs-Prokka_2_122-57471/000/000/Prokka_0/scaffolds

General Statistics

Copy table | Configure Columns | Plot | Showing 1/1 rows and 10/18 columns.

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

QualiMap

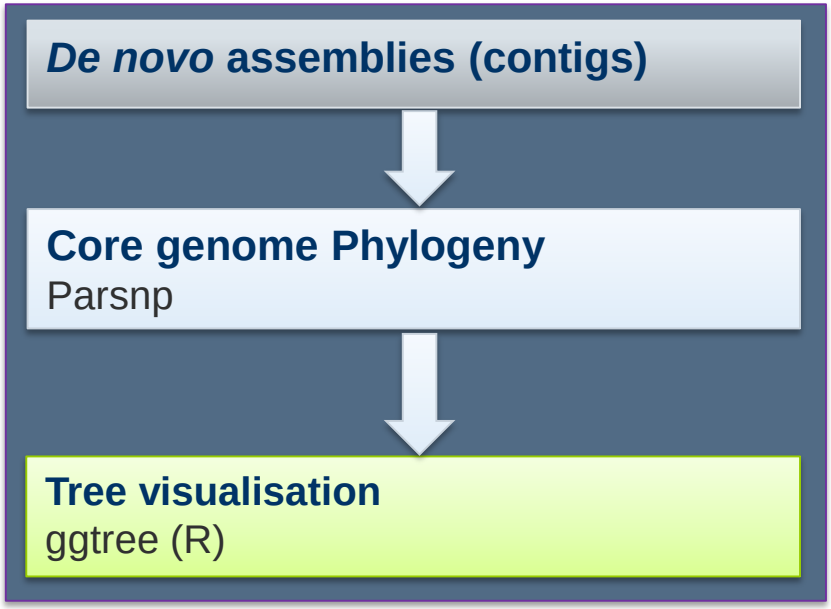
QualiMap is a platform-independent application to facilitate the quality control of alignment sequencing data and its derivatives like feature counts.

Coverage histogram

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

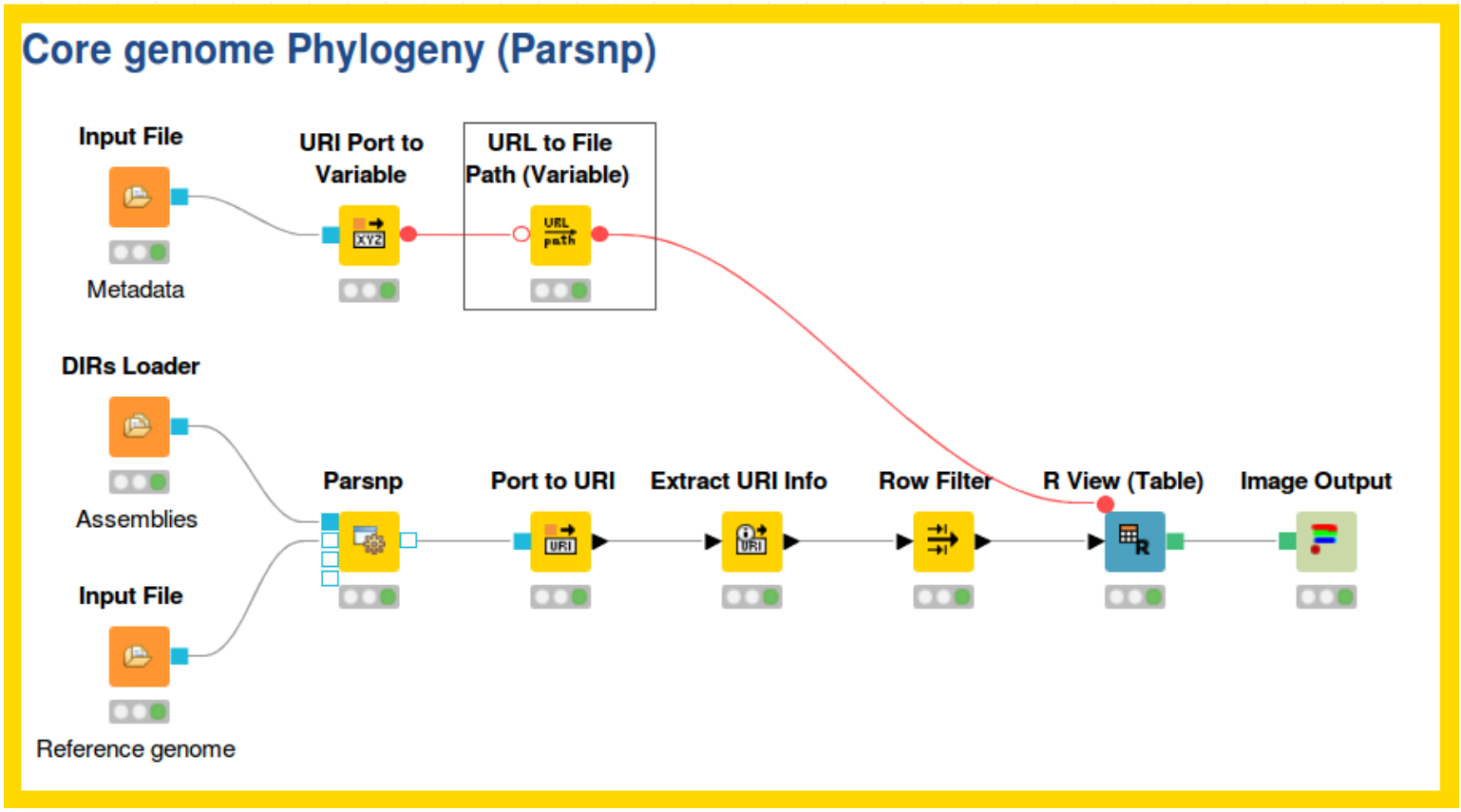
Qualimap BamQC: Coverage histogram

Phylogenetic pipeline

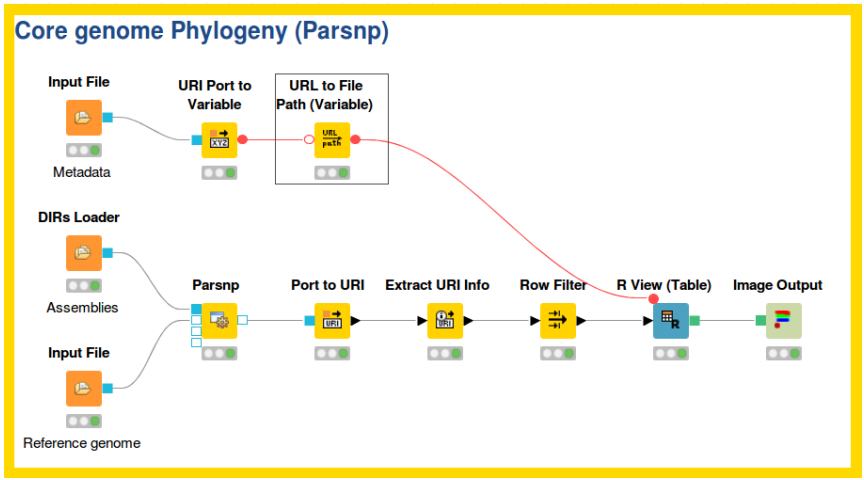


Treangen TJ*, Ondov BD*, Koren S, Phillippy AM: Rapid Core-Genome Alignment and Visualization for Thousands of Microbial Genomes. bioRxiv (2014). doi: http://dx.doi.org/10.1101/007351

Phylogenetic pipeline

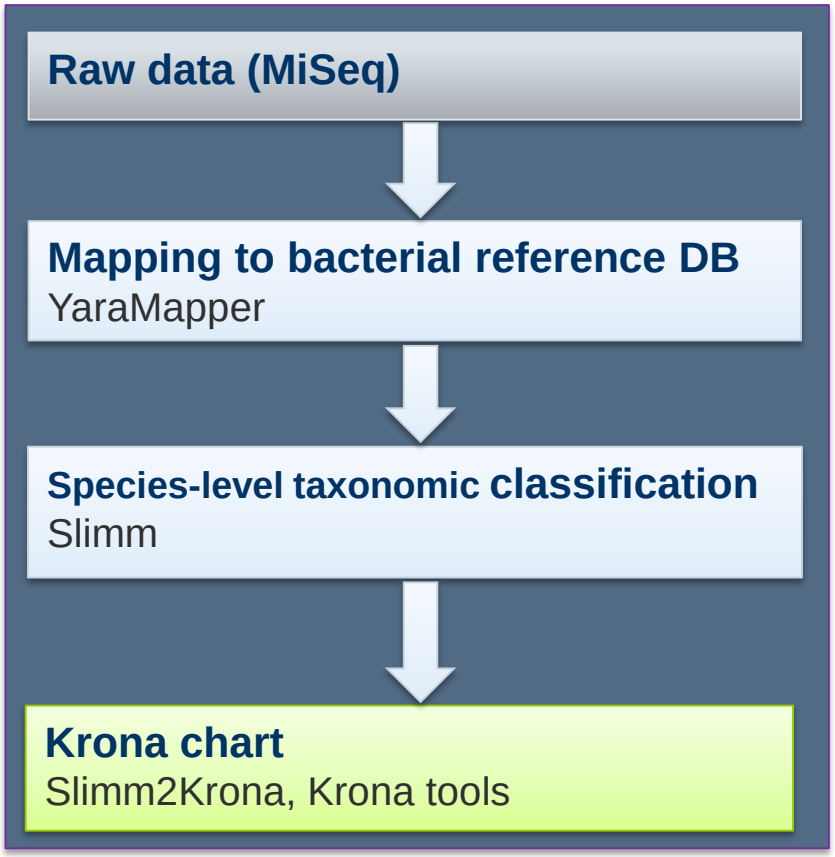


Phylogenetic pipeline



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



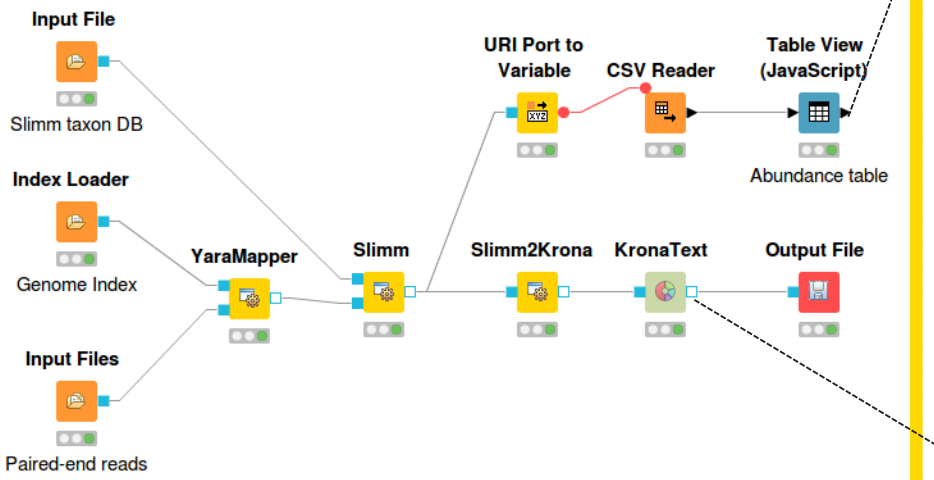
Dadi TH, Renard B, Wieler LH, Semmler T, Reinert K. (2016) SLIMM: Species level identification of microorganisms from metagenomes. PeerJ Preprints 4:e2378v1

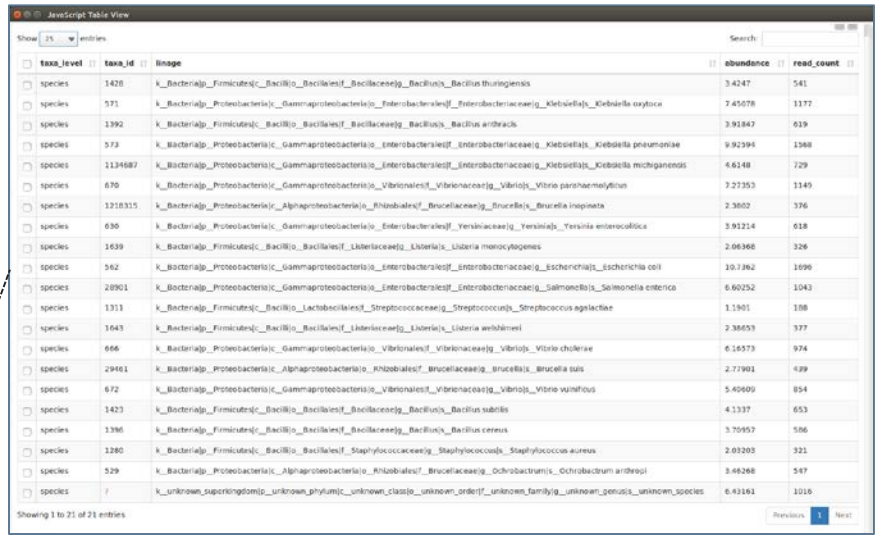
Ondov BD, Bergman NH, and Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011 Sep 30; 12(1):385.

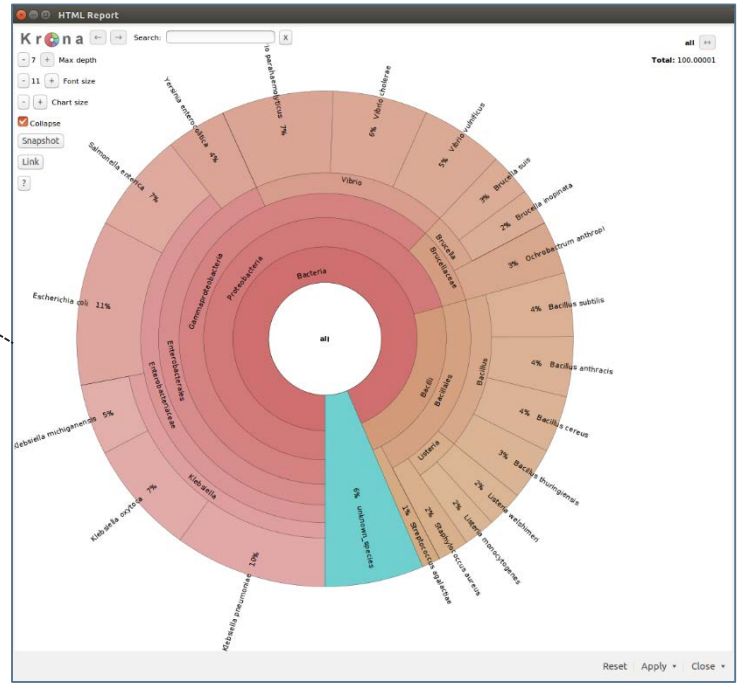
Metagenomics pipeline

Metagenomic profiling

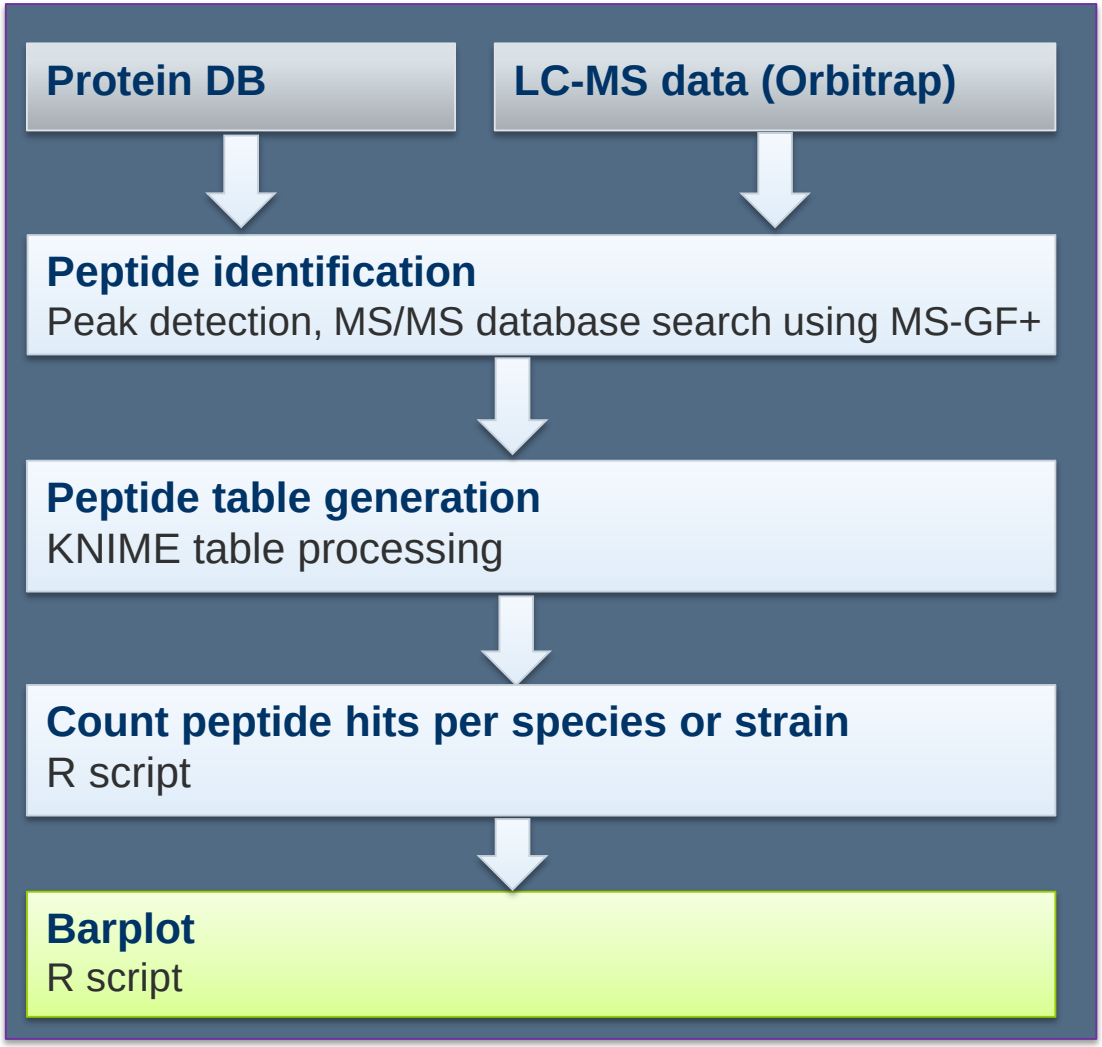
- Mapping to a reference DB (Yara)
- Species-level taxonomic classification (SLIMM)
- Visualisation (Krona Tools)



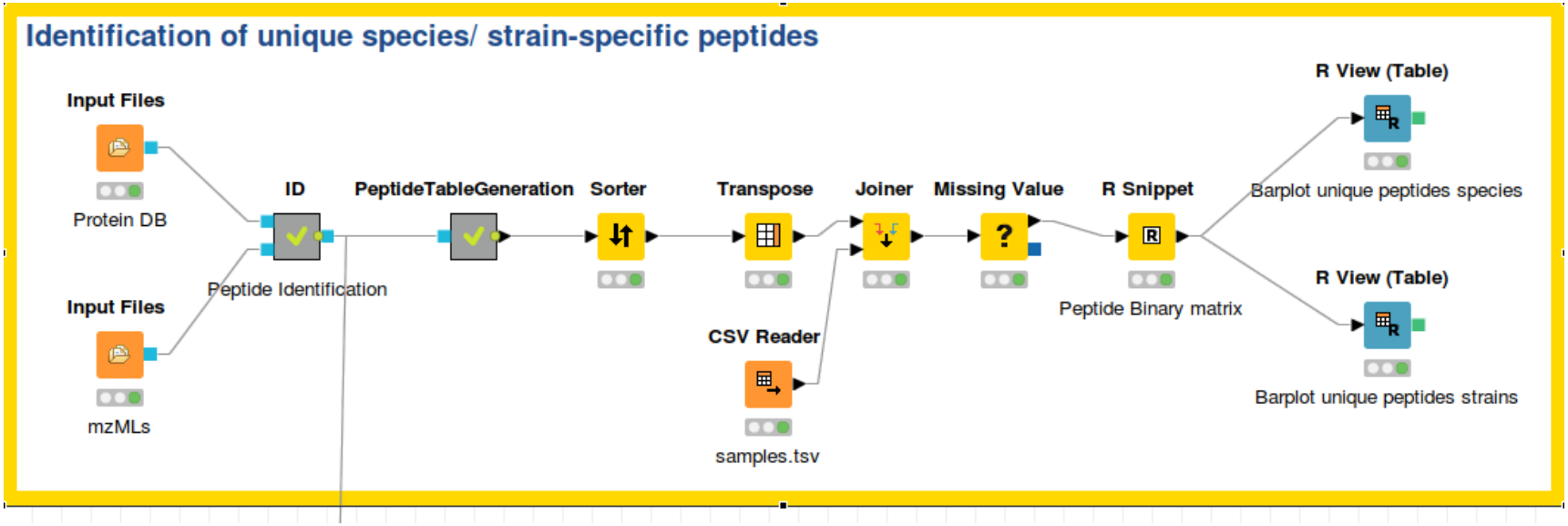




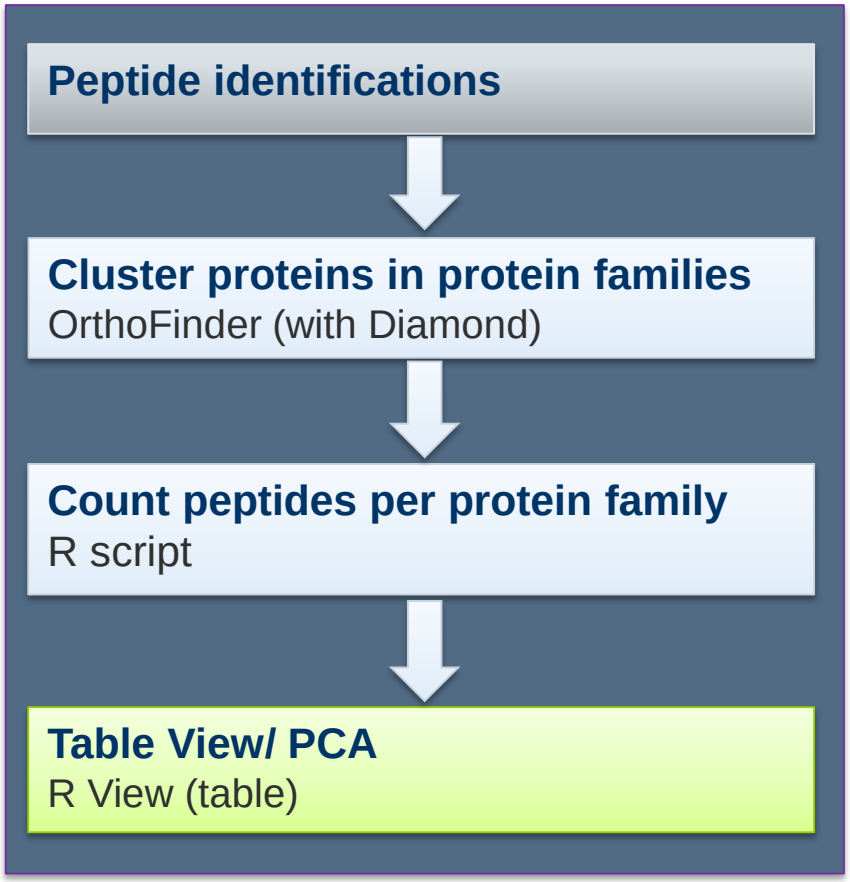
Identification of species-specific peptides



Identification of species-specific peptides

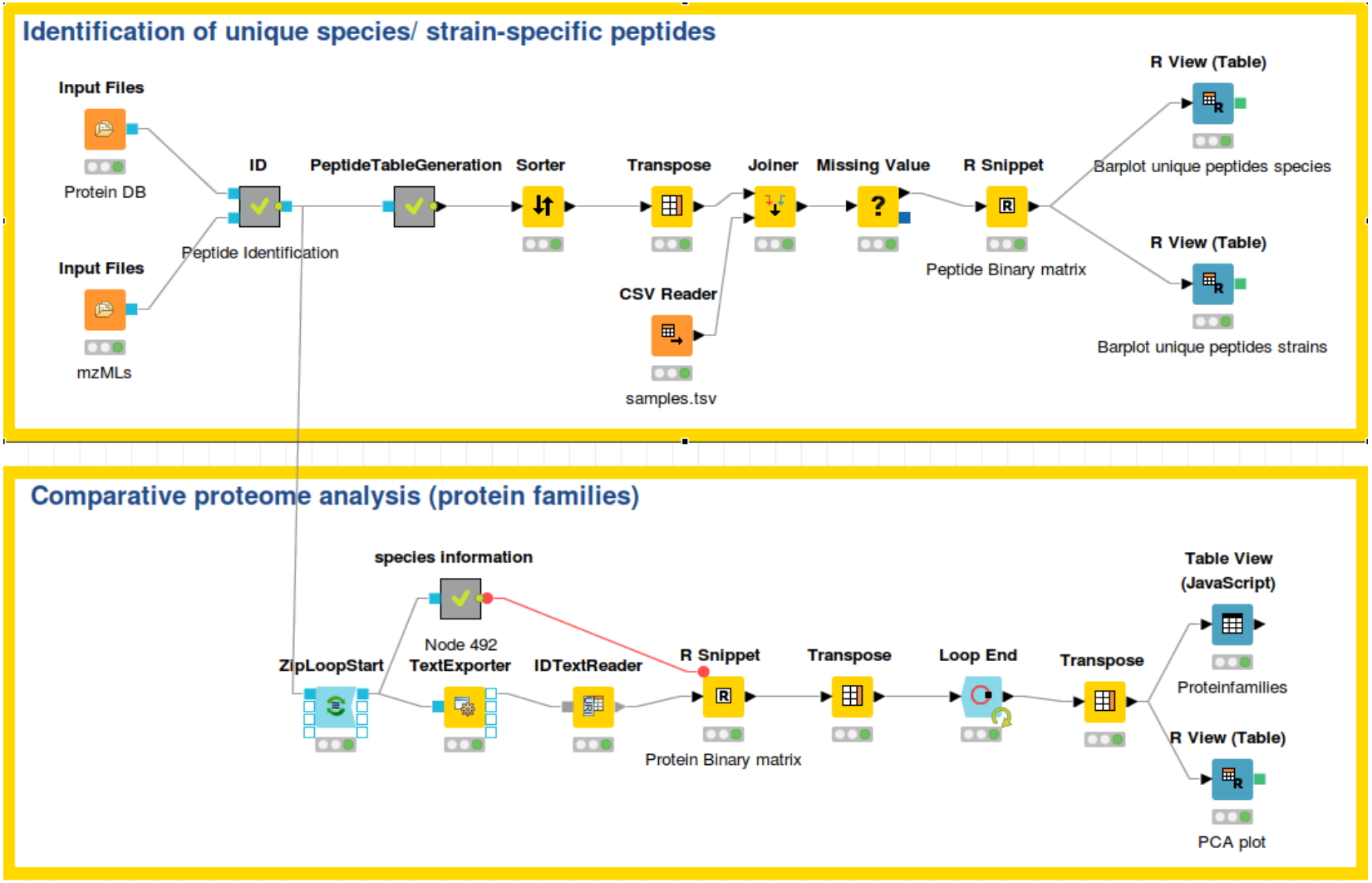


Proteome analysis

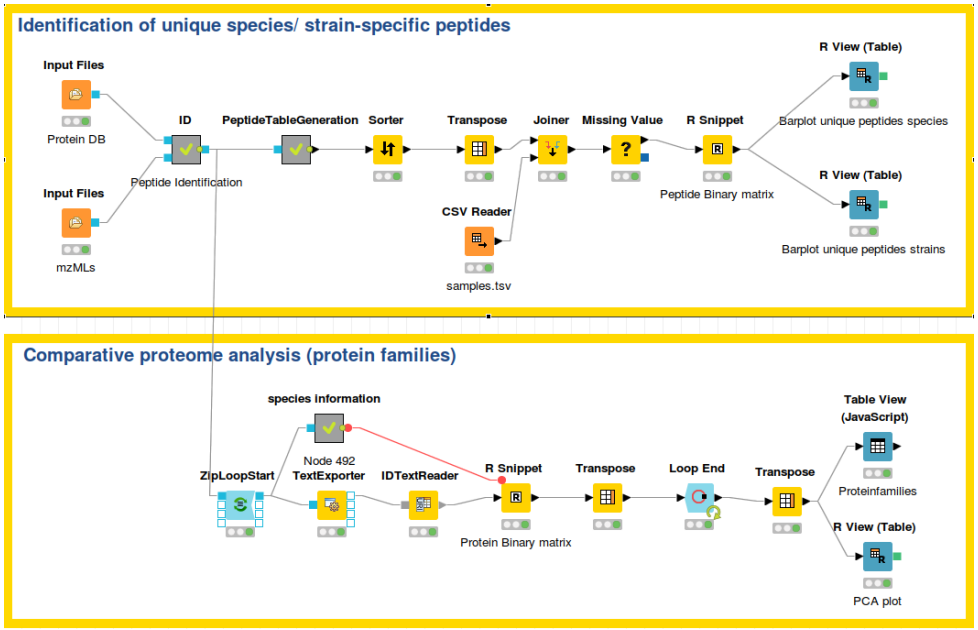


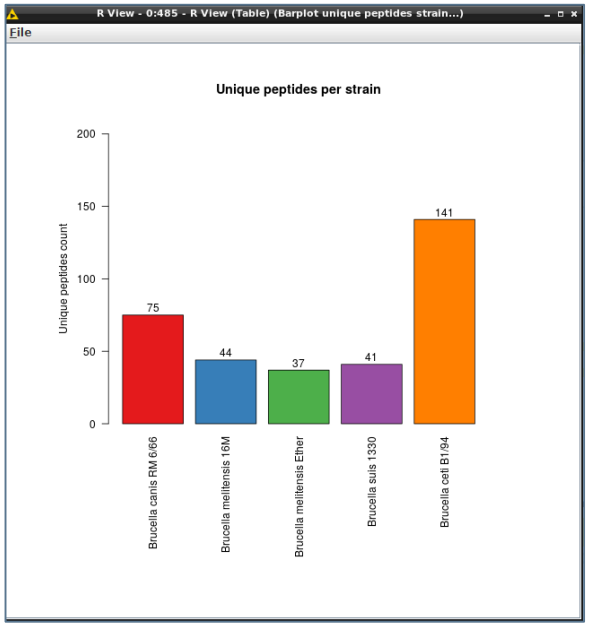
Emms, D.M. and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology 16:157

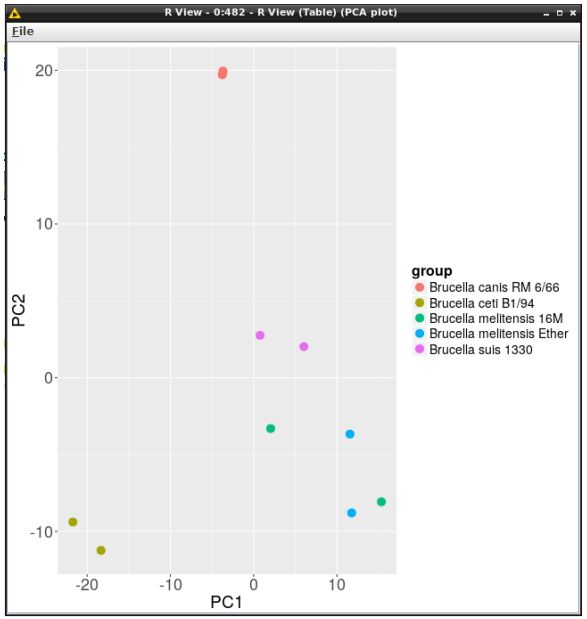
Proteome pipelines

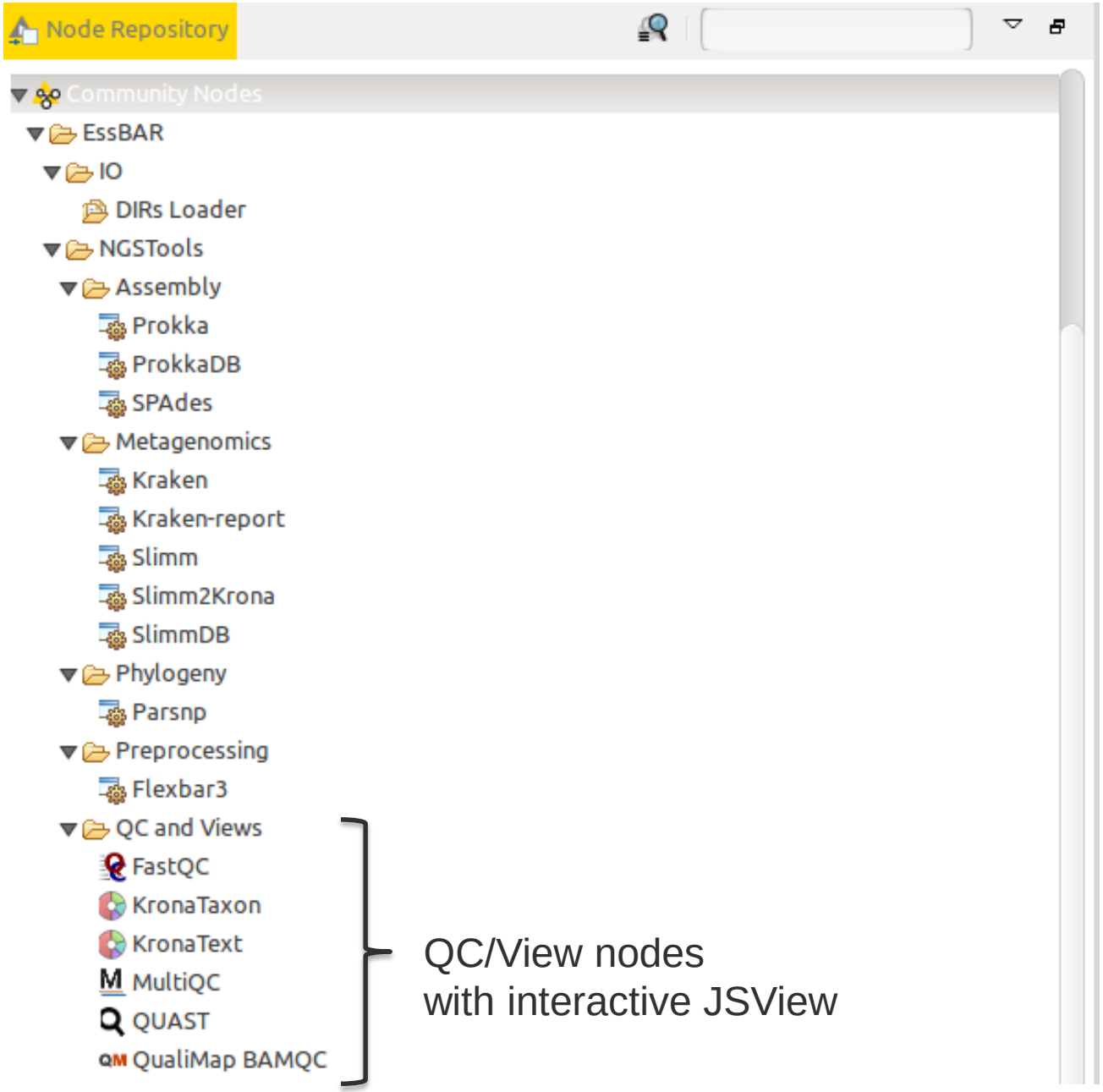


Proteome pipelines








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