MUFFIN pipeline, an application of Metagenomics and Nextflow scripting

Renaud Van Damme

Table of Content

1. Metagenomics
2. Nextflow
3. HTS
4. MUFFIN

What’s genomics?

typical genomics

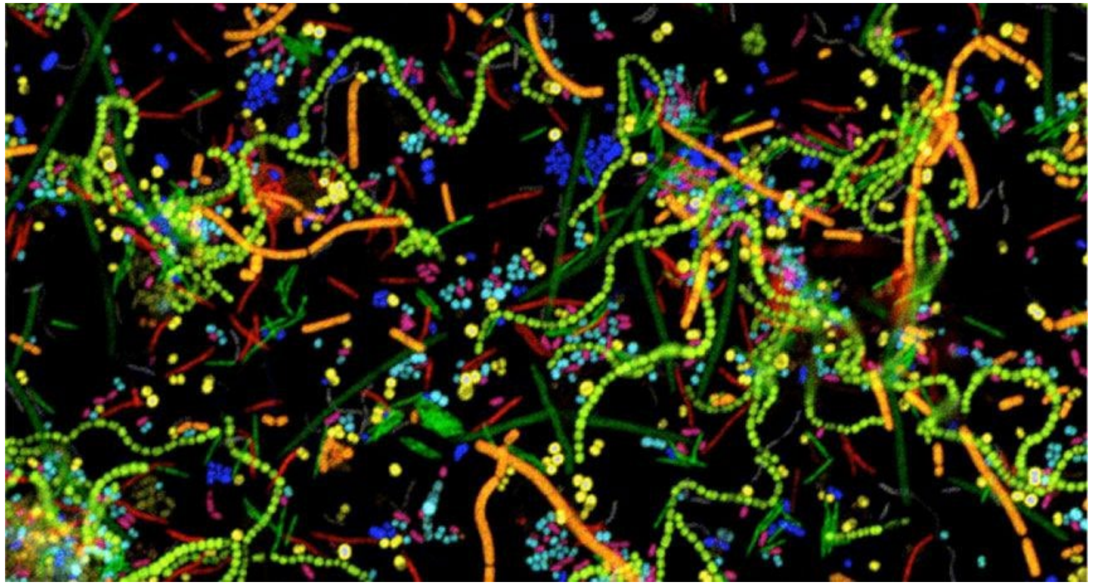
Single organism

Single genome

Meta-genomics

Many organisms

A Metagenome



What’s a genomic workflow?

Genomic (bioinformatics) workflows

- Data analysis apps to retrieve information from datasets
- Huge parallelisation, creating numerous job over a cluster
- Mix of tools and scripts
- Complex configurations and dependencies interactions between the said tools

Genomic workflows

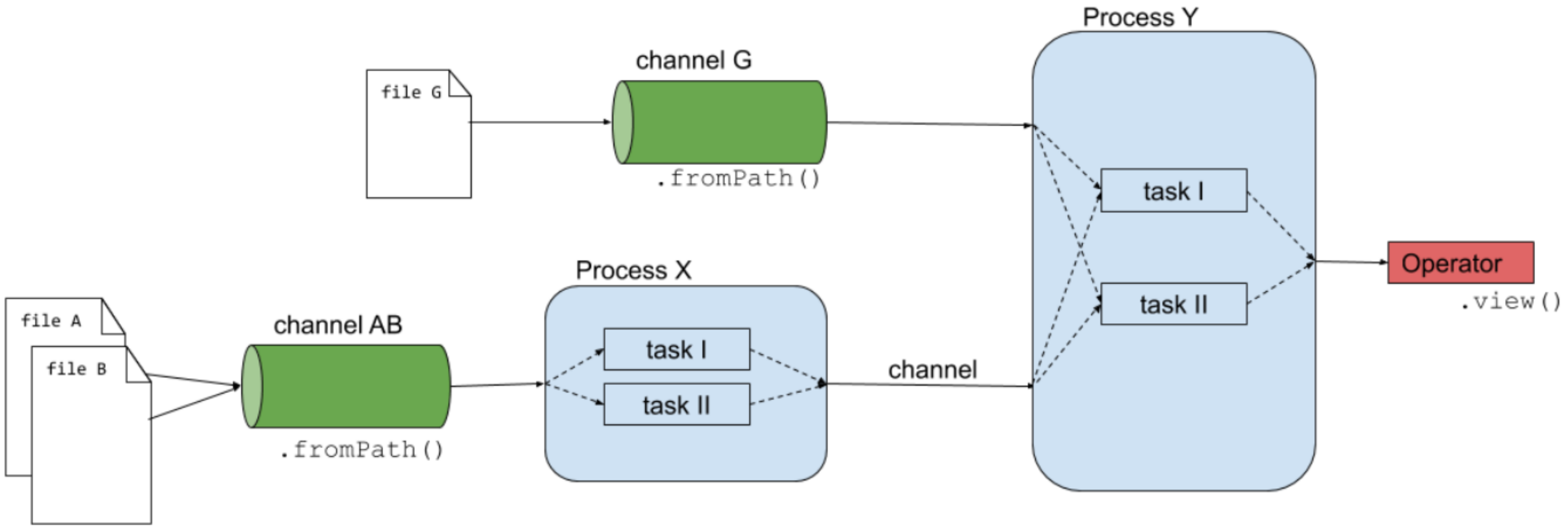
To reproduce a typical computational biology paper with minimal expertise it takes 280 Hours.
https://doi.org/10.1371/journal.pone.0080278

That include:

- Understanding the paper and material used
- Installation, set up
- Finding parameters
- Workflow validation
- computing

Nextflow

- A genomic (bioinformatics) workflow manager
- Portable
- Scalable
- Reproducible
- Consistent
- Easy to use



Source: https://github.com/vibbits/nextflow-jnj/blob/master/presentation/slidedeck.pdf

HTS/NGS ?

High Throughput Sequencing / Next Generation Sequencing?

HTS short read

- Highly accurate on the base level
- Expensive cost per gigabase
- Low completeness and higher contamination risk
- Huge variety of tools and software available

HTS long read

- Lower accuracy on the base level (depending the technology)
- Cheaper cost per gigabase sequenced
- High completeness and lower contamination risk
- Fewer tools and software are available

MUFFIN

Background – Hybrid assembly

- High accuracy on base level
- High completeness and low contamination
- Require both sequencing technologies

Background - Aim

Create a metagenomics and transcriptomics pipeline

- Using Hybrid assembly
- Ergonomic, automated and reproducible
- Producing high quality metagenome-assembled genomes (MAGs)
- Outputting Valuable summary files

Methods

The pipeline Consist of 3 steps:

❖Assemble

Quality control, (hybrid) assembly, binning, optional re-assembly

❖Classify

Assess the bins quality and do a Taxonomic classification

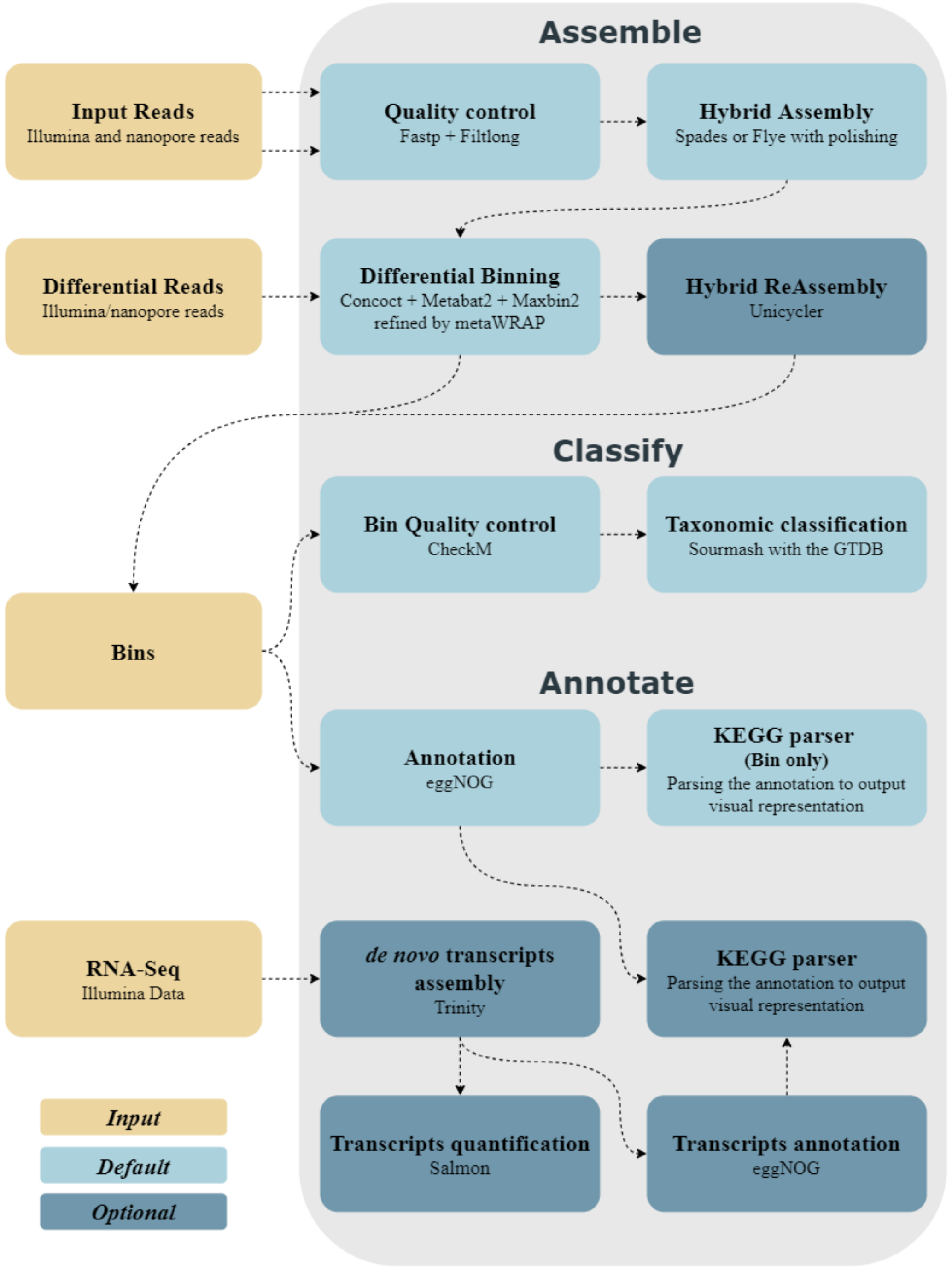
❖Annotate

Annotate the bins and if provided RNAseq data, do *de novo* transcriptoms assembly, quantification and annotation

Bioinformatics analysis

MUFFIN pipeline

GitHub: RVanDamme/MUFFIN



Assemble

Input Reads
Illumina and nanopore reads

Quality control
Fastp + Filtlong

Hybrid Assembly
Spades or Flye with polishing

Differential Reads
Illumina/nanopore reads

Differential Binning
Concoct + Metabat2 + Maxbin2
refined by metaWRAP

Hybrid ReAssembly
Unicycler

Classify

Bins

Bin Quality control
CheckM

Taxonomic classification
Sourmash with the GTDB

Annotate

Annotation
eggNOG

KEGG parser
(Bin only)
Parsing the annotation to output
visual representation

RNA-Seq
Illumina Data

***de novo* transcripts assembly**
Trinity

KEGG parser
Parsing the annotation to output
visual representation

Transcripts quantification
Salmon

Transcripts annotation
eggNOG

Input

Default

Optional

Methods - Assemble

- Reads quality control

  Fastp and Filtlong

- Hybrid assembly / long read assembly

  Spades or Flye + racon +medaka + pilon

- Multiple Binning + bin refinement

  Metabat2, Concoct, Maxbin2 + metaWRAP refinement step

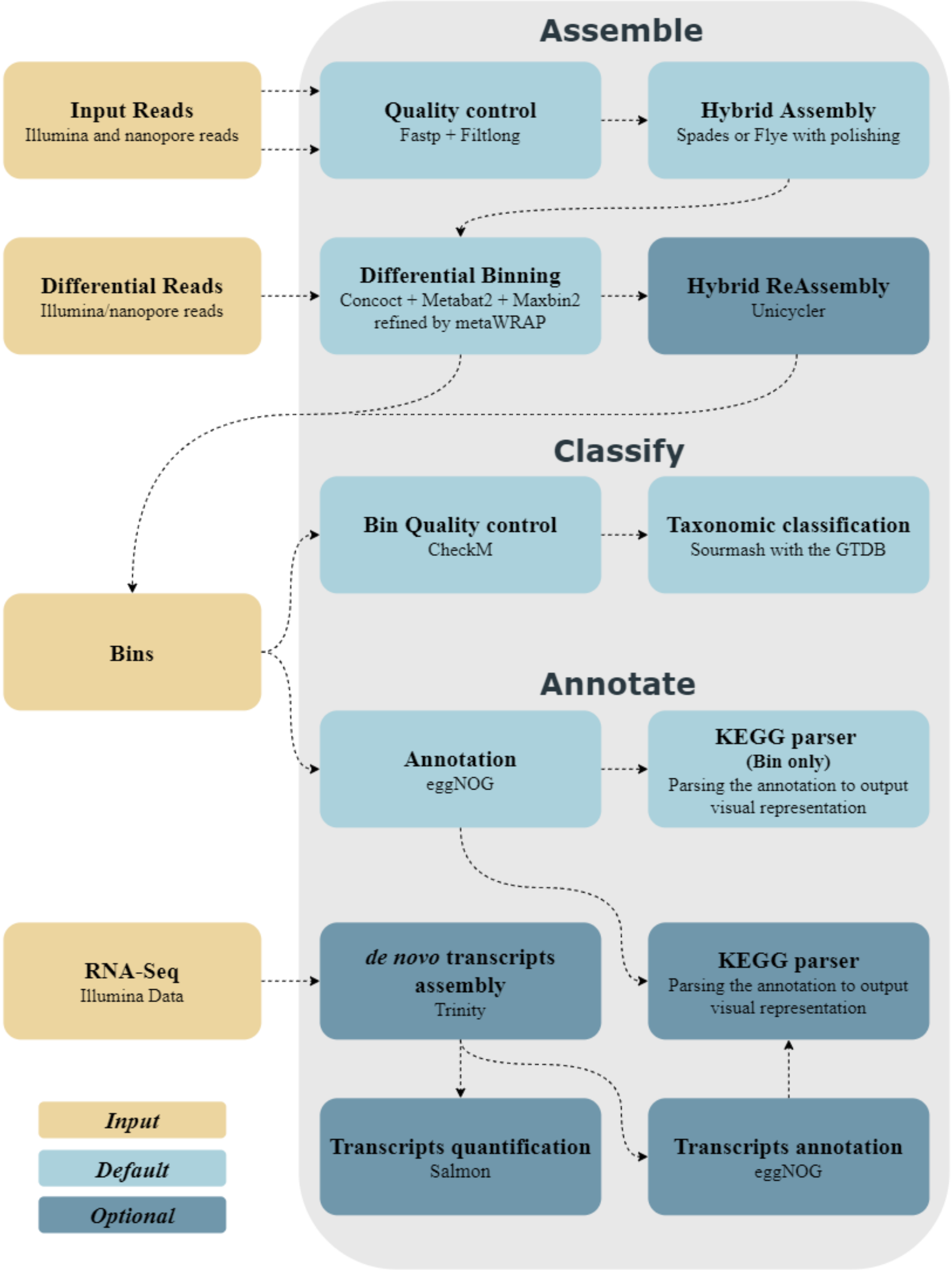
- Optional Re-assembly

  Unicycler

Bioinformatics analysis

MUFFIN pipeline

GitHub: RVanDamme/MUFFIN



Methods - classify

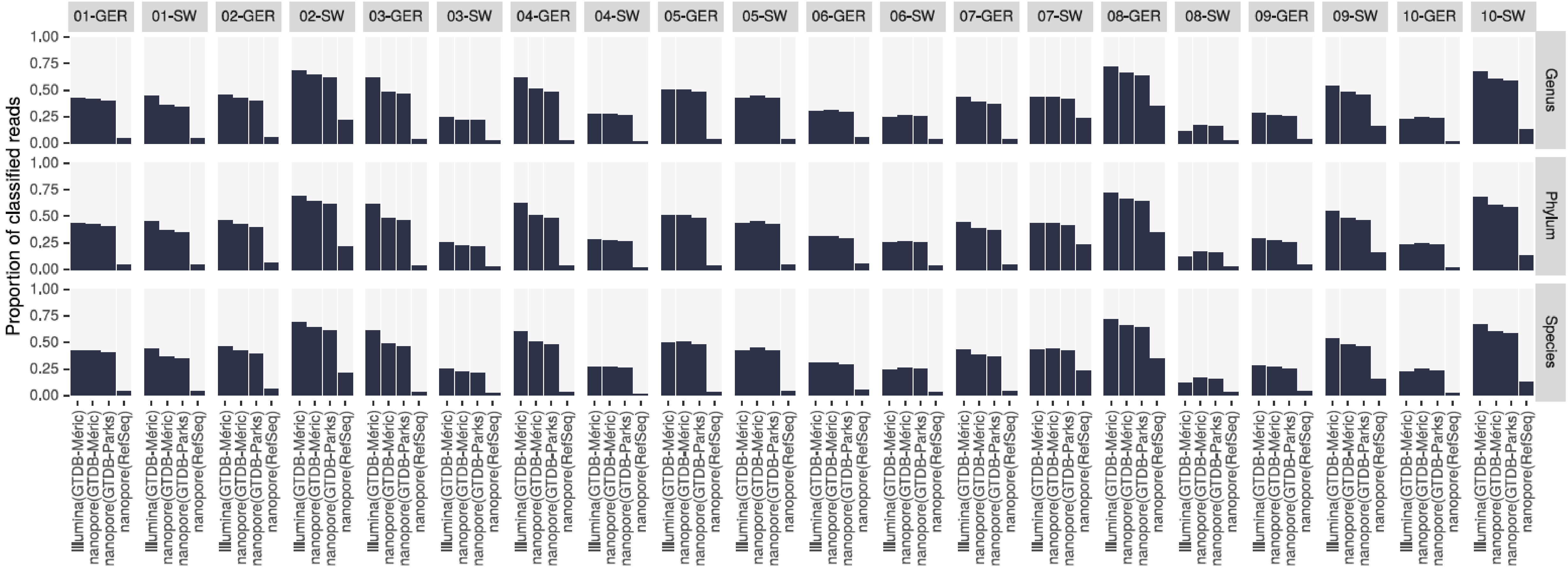
- Bins quality control

  CheckM

- Taxonomic Classification

  Sourmash with GTDB (Genome Taxonomy Database)

Workflow - Classify



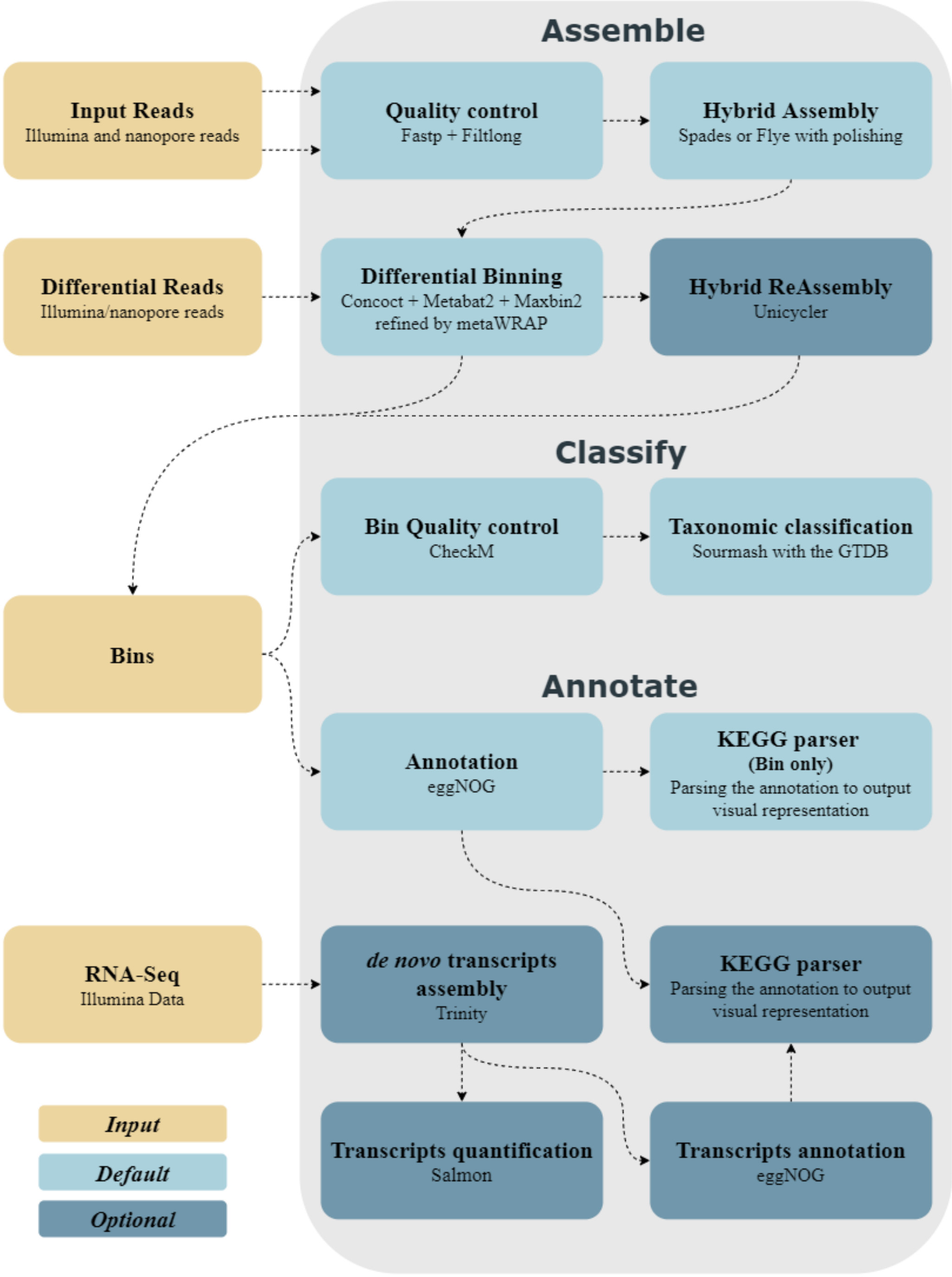
Source, Brandt et al., 2019, ***Abundance tracking by long-read nanopore sequencing of complex microbial communities in samples from 20 different biogas/wastewater plants.***

DOI: 10.21203/rs.2.17734/v1

Bioinformatics analysis

MUFFIN pipeline

GitHub: RVanDamme/MUFFIN



Assemble

Input Reads
Illumina and nanopore reads

Quality control
Fastp + Filtlong

Hybrid Assembly
Spades or Flye with polishing

Differential Reads
Illumina/nanopore reads

Differential Binning
Concoct + Metabat2 + Maxbin2
refined by metaWRAP

Hybrid ReAssembly
Unicycler

Classify

Bins

Bin Quality control
CheckM

Taxonomic classification
Sourmash with the GTDB

Annotate

Annotation
eggNOG

KEGG parser
(Bin only)
Parsing the annotation to output
visual representation

RNA-Seq
Illumina Data

***de novo* transcripts**
assembly
Trinity

KEGG parser
Parsing the annotation to output
visual representation

Transcripts quantification
Salmon

Transcripts annotation
eggNOG

Input

Default

Optional

Methods - annotate

- Annotation

  eggNOG

- *De novo* transcripts assembly and quantification

  Trinity and Salmon

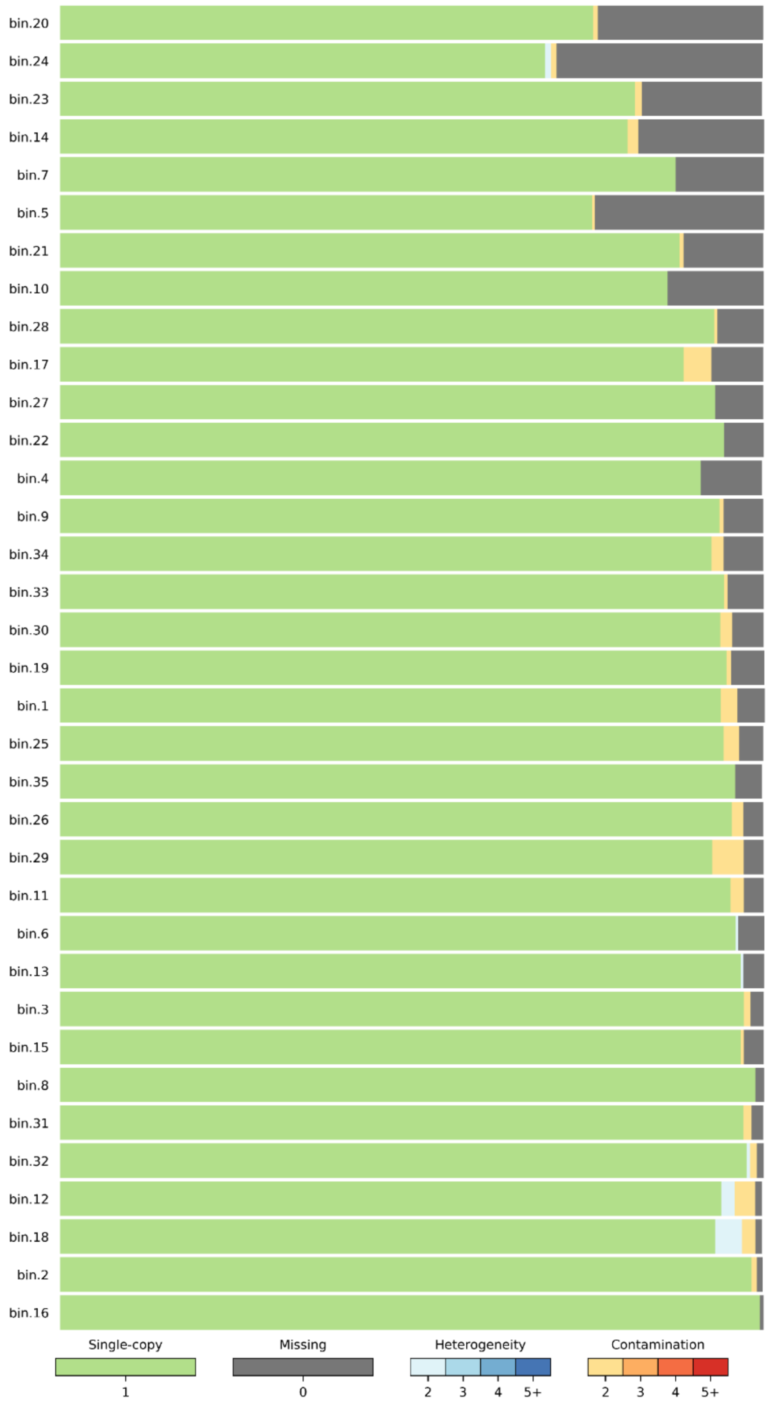
- Summarizing the annotation

  PANKEGG

Results

Binning

- 35 Bins
- 71.16% to 99.60% completeness
- 0% to 6,78% contamination
- Mean 90,99% completeness and 1,38% contamination
- Before the refinement
    - Maxbin2 had 51 bins
    - Metabat2 had 60 bins
    - Concoct had 138 bins

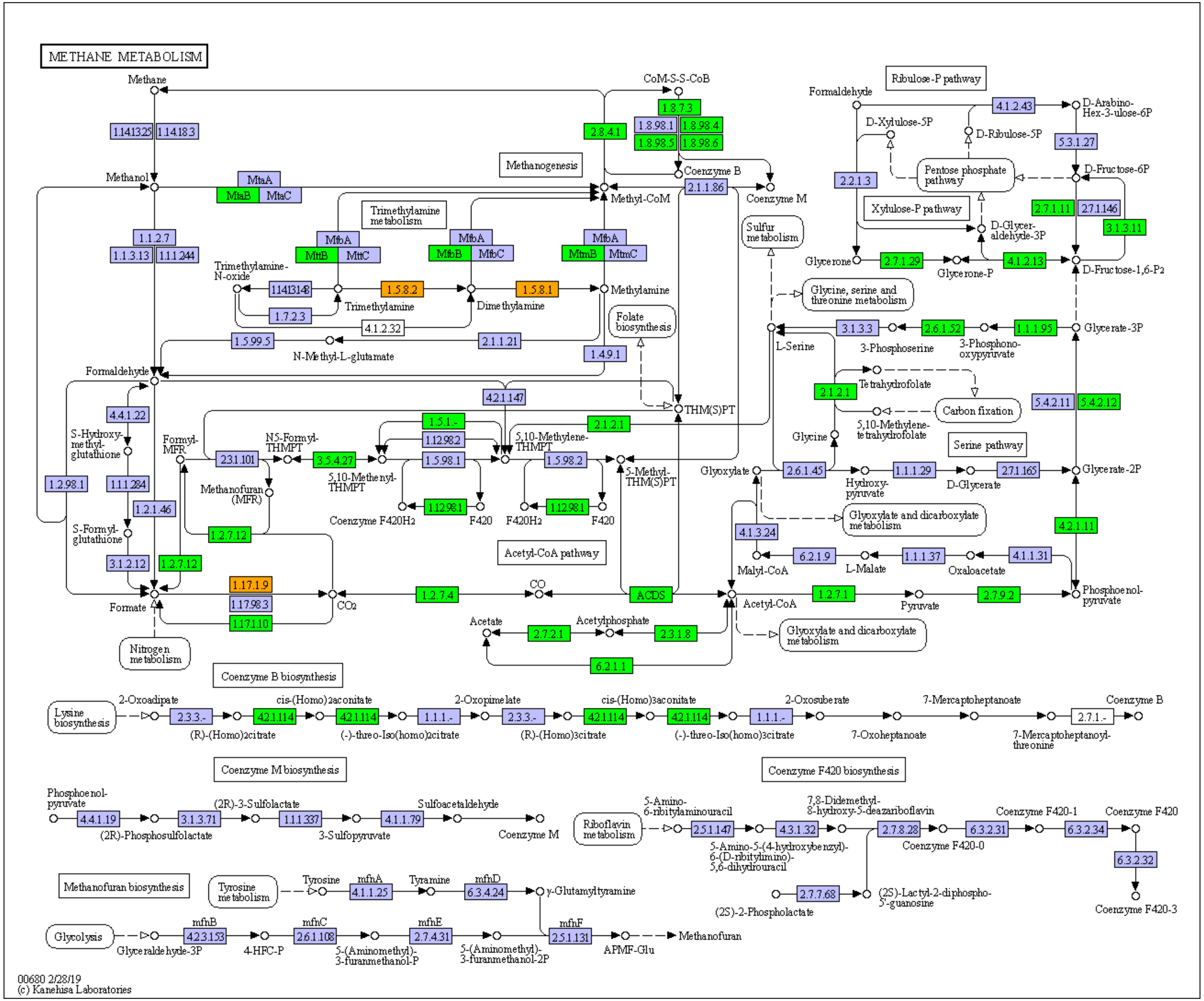


Results

Classification

Bin ID	Checkm Marker lineage	Sourmash Status	Sourmash phylum	Sourmash Class
bin.01	c__Clostridia	found	p__Firmicutes_B	c__Syntrophomonadia
bin.02	k__Bacteria	found	p__Firmicutes	c__Bacilli
bin.03	p__Firmicutes	found	p__Firmicutes_G	c__UBA4882
bin.04	k__Bacteria	found	p__Thermotogota	c__Thermotogae
bin.05	o__Clostridiales	found	p__Firmicutes_A	c__Clostridia
bin.06	p__Firmicutes	found	p__Firmicutes_B	c__Syntrophomonadia
bin.07	c__Clostridia	found	p__Firmicutes_A	c__Clostridia
bin.08	o__Clostridiales	nomatch		
bin.09	p__Firmicutes	found	p__Firmicutes_G	c__SHA-98
bin.10	p__Euryarchaeota	found	p__Halobacterota	c__Methanomicrobia
bin.11	p__Firmicutes	found	p__Firmicutes_G	c__Limnochordia
bin.12	k__Bacteria	found	p__Thermotogota	c__Thermotogae
bin.13	k__Bacteria	disagree	p__Bacteroidota	c__Bacteroidia
bin.14	o__Clostridiales	found	p__Firmicutes_A	c__Clostridia
bin.15	p__Firmicutes	found	p__DTU030	c__DTU030
bin.16	p__Euryarchaeota	found	p__Thermoplasmatota	c__Thermoplasmata
bin.17	p__Firmicutes	nomatch		
bin.18	k__Bacteria	found	p__Caldatribacteriota	c__Caldatribacteriia

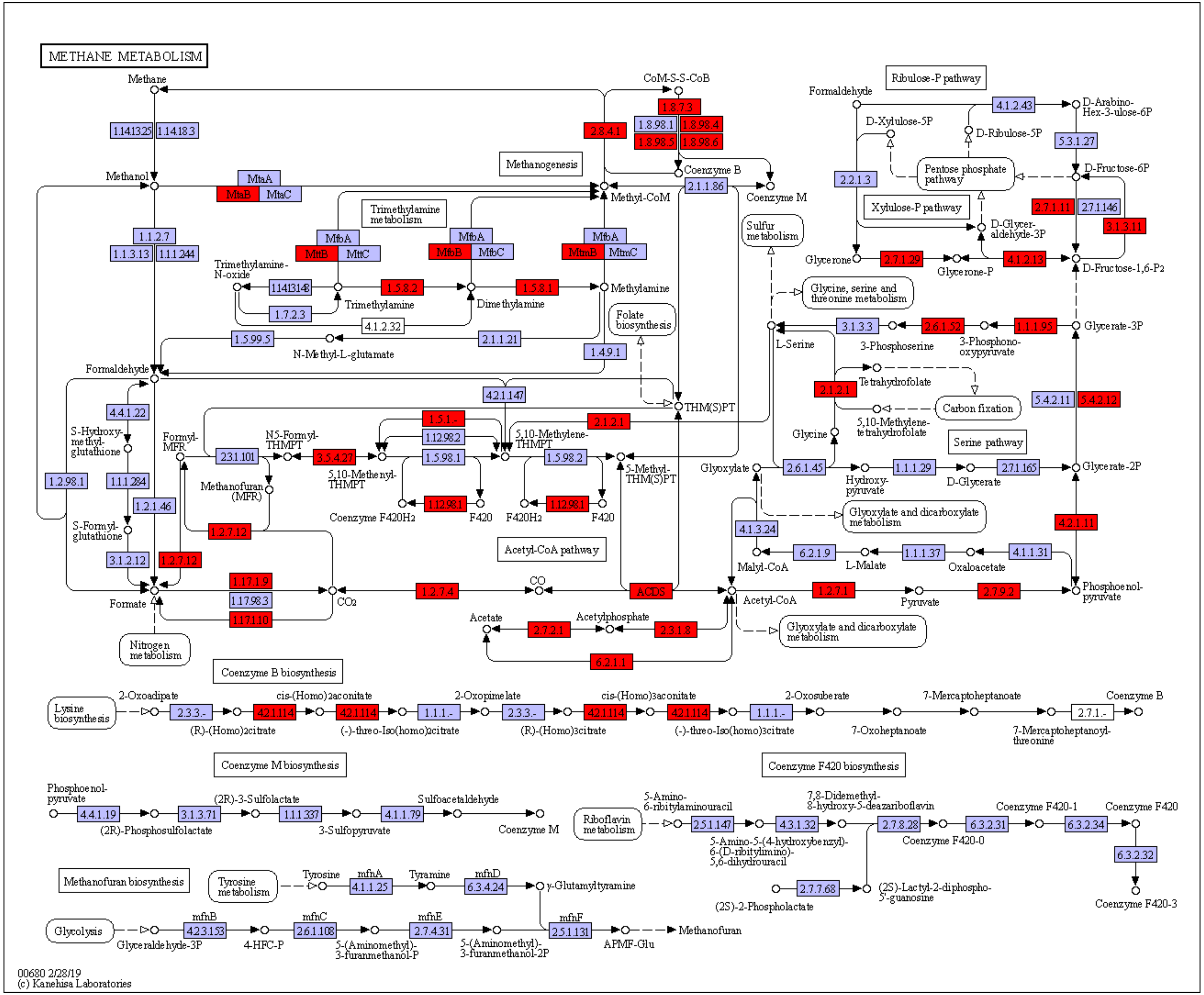
Bin ID	Checkm Marker lineage	Sourmash Status	Sourmash phylum	Sourmash Class
bin.19	p__Bacteroidetes	found	p__Bacteroidota	c__Bacteroidia
bin.20	k__Bacteria	nomatch		
bin.21	p__Firmicutes	found	p__Firmicutes_G	c__Limnochordia
bin.22	p__Firmicutes	nomatch		
bin.23	k__Bacteria	found	p__Caldatribacteriota	c__Caldatribacteriia
bin.24	k__Bacteria	found	p__Firmicutes	c__Bacilli
bin.25	p__Firmicutes	found	p__Firmicutes_G	c__SHA-98
bin.26	p__Firmicutes	disagree	p__Firmicutes_G	
bin.27	p__Firmicutes	found	p__Firmicutes_E	c__DTU015
bin.28	p__Firmicutes	nomatch		
bin.29	p__Firmicutes	found	p__Firmicutes_A	c__Thermovenabulia
bin.30	p__Firmicutes	found	p__Firmicutes_G	c__Limnochordia
bin.31	p__Firmicutes	nomatch		
bin.32	p__Firmicutes	found	p__Firmicutes_F	c__Halanaerobiia
bin.33	p__Firmicutes	found	p__Firmicutes_D	c__Dethiobacteria
bin.34	p__Firmicutes	found	p__Firmicutes_G	c__DTU065
bin.35	k__Bacteria	nomatch		



Results

Annotation - Methane Metabolism

Bins X RNAseq Data



Results

Annotation - Methane Metabolism

Bins Data

Thanks for your attention

Book: “The story of cattle in Africa:Why diversity matters”

Tadelle Dessie *and* ***Okeyo Mwai****, Livestock Genetics Research Program, ILRI*

SCIENCE AND EDUCATION FOR SUSTAINABLE LIFE