

MUFFIN pipeline, an application of Metagenomics and Nextflow scripting

Renaud Van Damme



Table of Content

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





typical genomics



Single organism Single genome

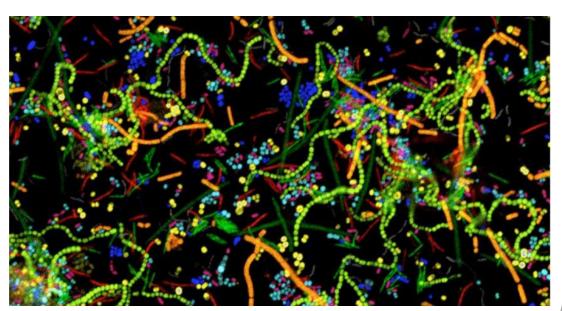


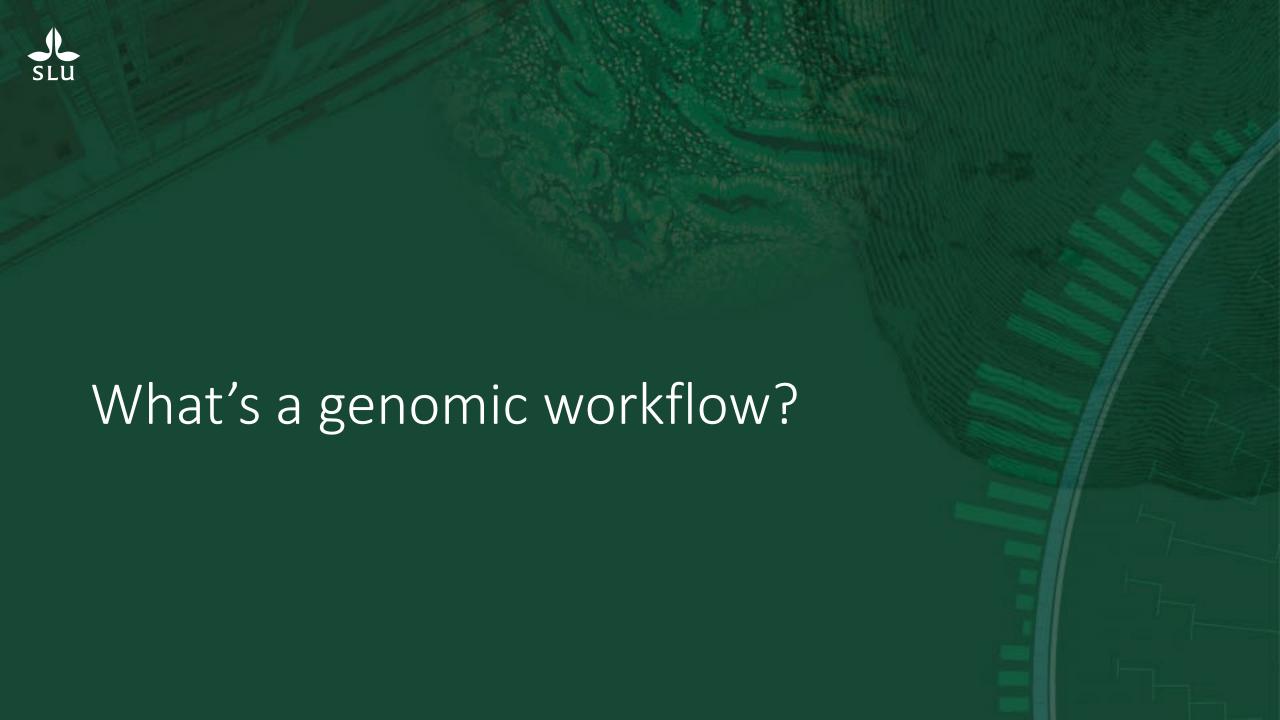
Meta-genomics



Many organisms

A Metagenome





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

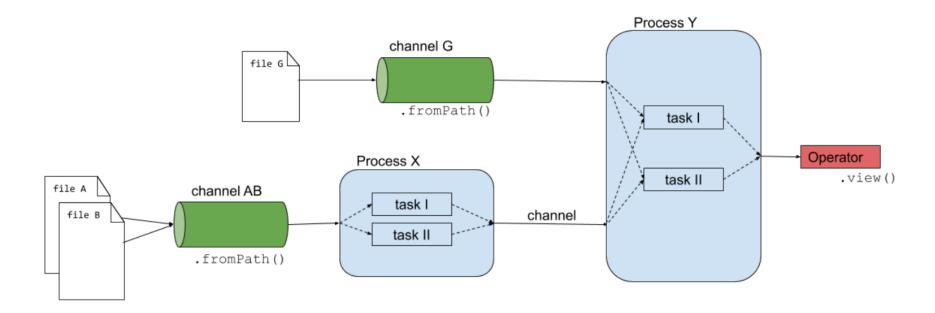
That include:

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

Nextflow

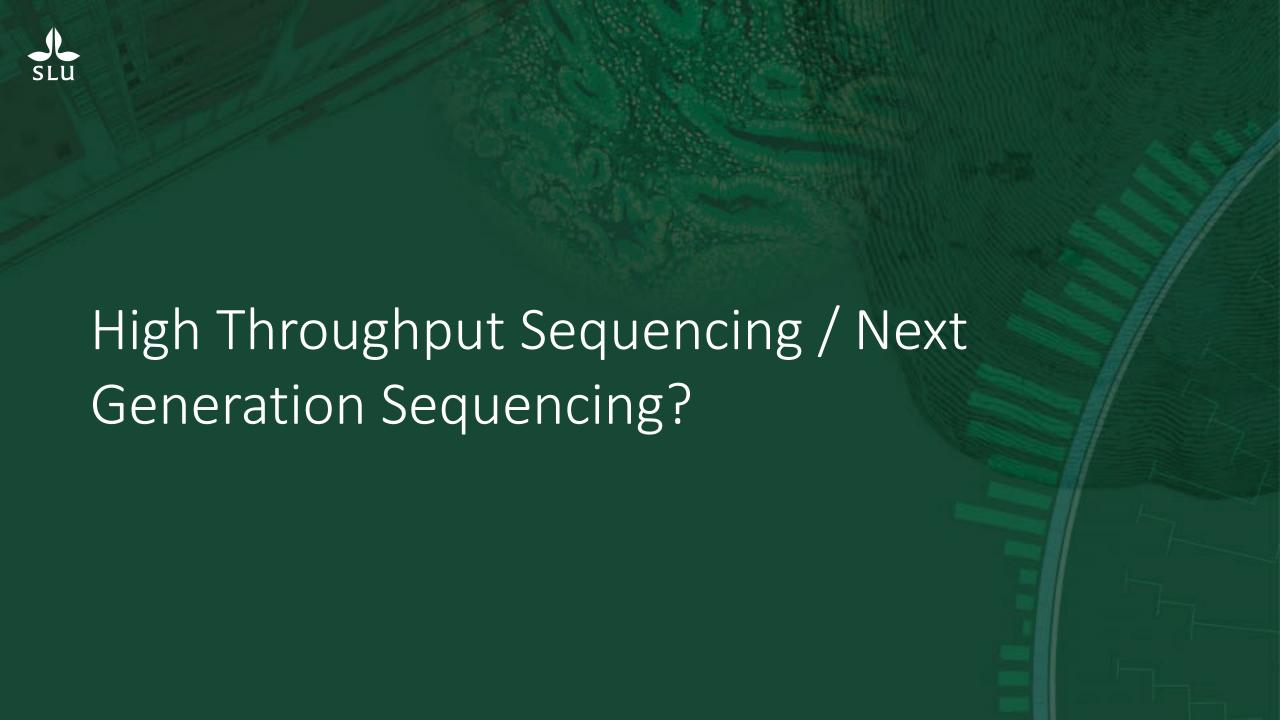
- A genomic (bioinformatics) workflow manager
- Portable
- Scalable
- Reproducible
- Consistent
- Easy to use





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





HTS 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



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

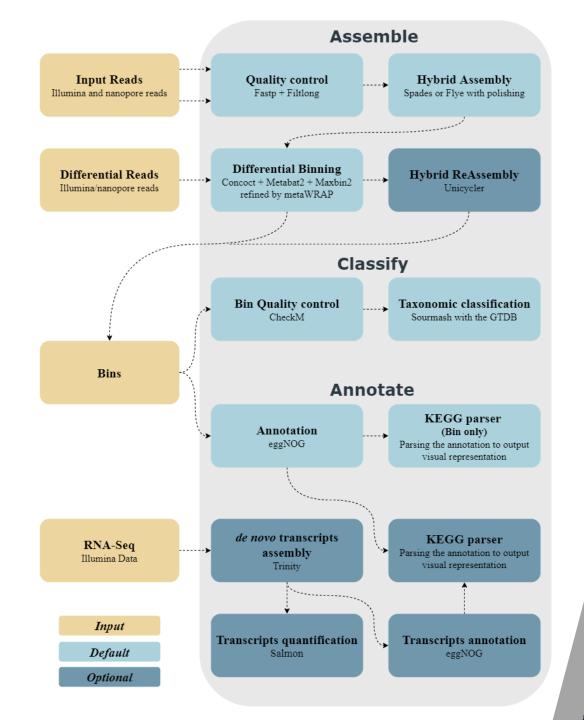
Annotate

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

Bioinformatics analysis

MUFFIN pipeline

GitHub: RVanDamme/MUFFIN



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

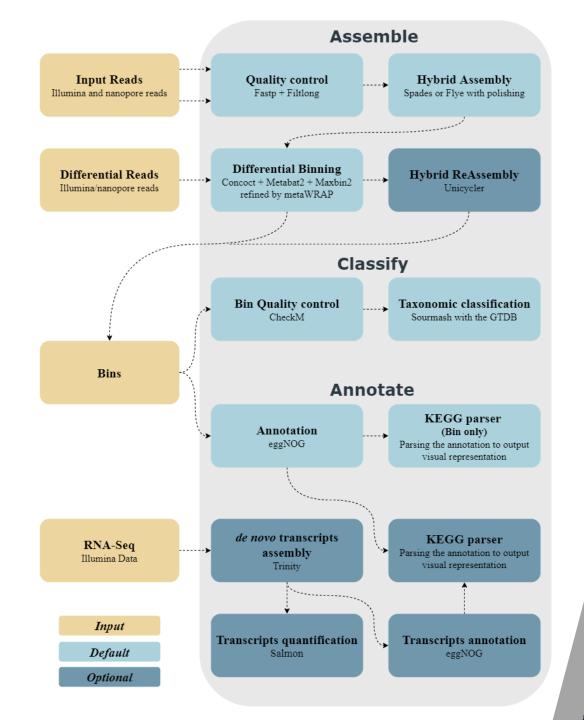
Optional Re-assembly

Unicycler

Bioinformatics analysis

MUFFIN pipeline

GitHub: RVanDamme/MUFFIN



Methods - classify

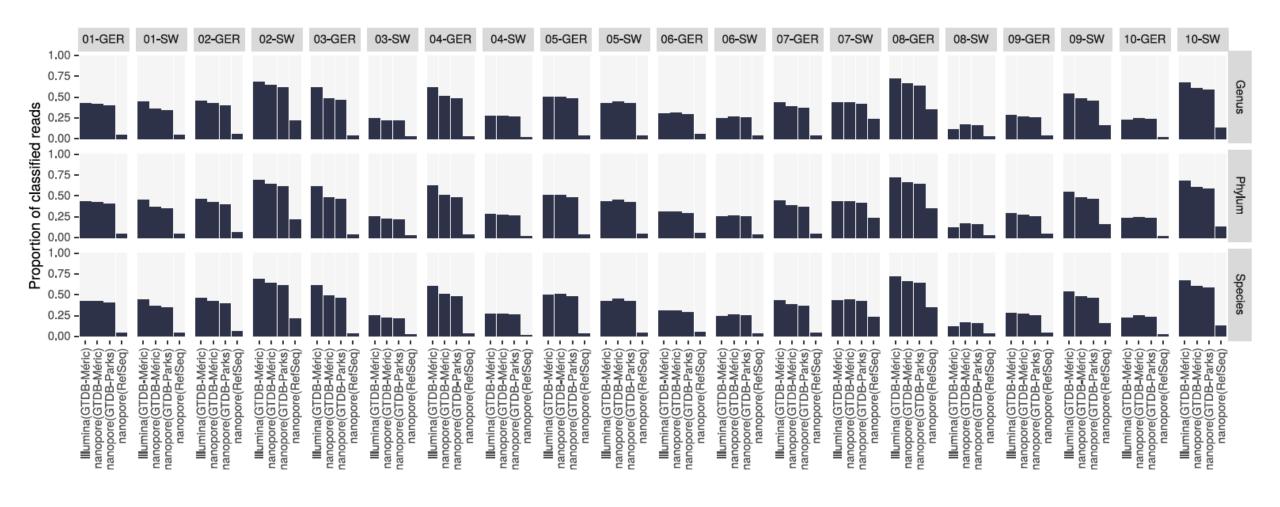
Bins quality control

CheckM

• Taxonomic Classification

Sourmash with GTDB (Genome Taxonomy Database)

Workflow - Classify



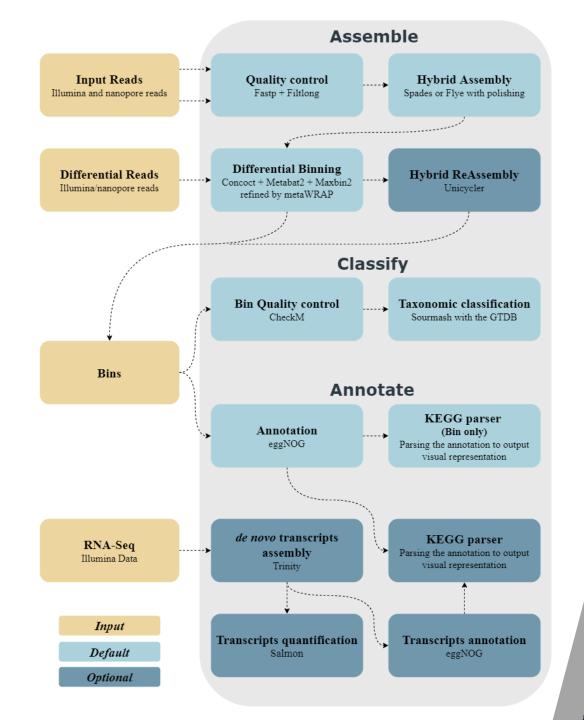
Source, Brandt et al., 2019, Abundance tracking by long-read nanopore sequencing of complex microbial communities in samples from 20 different biogas/wastewater plants.

DOI: <u>10.21203/rs.2.17734/v1</u>

Bioinformatics analysis

MUFFIN pipeline

GitHub: RVanDamme/MUFFIN



Methods - annotate

Annotation

eggNOG

• De novo transcripts assembly and quantification

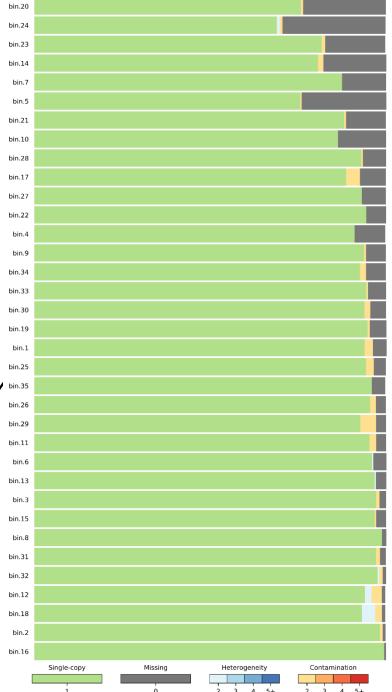
Trinity and Salmon

Summarizing the annotation

PANKEGG

Results Binning

- 35 Bins
- 71.16% to 99.60% completeness
- 0% to 6,78% contamination
- Mean 90,99% completeness and 1,389 bin.35 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 | cClostridia | found | pFirmicutes_B | cSyntrophomonadia |
| bin.02 | kBacteria | found | pFirmicutes | cBacilli |
| bin.03 | pFirmicutes | found | pFirmicutes_G | cUBA4882 |
| bin.04 | k Bacteria | found | n Thormotogota | c. Thermetegae |
| biii.04 | Kbacteria | Touriu | pThermotogota | cThermotogae |
| bin.05 | oClostridiales | found | pFirmicutes_A | cClostridia |
| 11.00 | -· · · | | E | |
| bin.06 | pFirmicutes | found | pFirmicutes_B | cSyntrophomonadia |
| bin.07 | cClostridia | found | pFirmicutes_A | cClostridia |
| bin.08 | o Clostridiales | nomatch | | |
| bin.09 | p Firmicutes | found | pFirmicutes_G | c SHA-98 |
| | | | _ | |
| bin.10 | pEuryarchaeota | found | pHalobacterota | cMethanomicrobia |
| bin.11 | p Firmicutes | found | p Firmicutes G | c Limnochordia |
| | <u> </u> | | <u>p</u> | <u></u> |
| bin.12 | kBacteria | found | pThermotogota | cThermotogae |
| bin.13 | k Bacteria | disagree | p Bacteroidota | c Bacteroidia |
| | | , and the second | | |
| bin.14 | oClostridiales | found | pFirmicutes_A | cClostridia |
| bin.15 | pFirmicutes | found | pDTU030 | cDTU030 |
| | | | | |
| bin.16 | pEuryarchaeota | found | pThermoplasmatota | cThermoplasmata |
| bin.17 | pFirmicutes | nomatch | | |
| bin.18 | kBacteria | found | pCaldatribacteriota | cCaldatribacteriia |

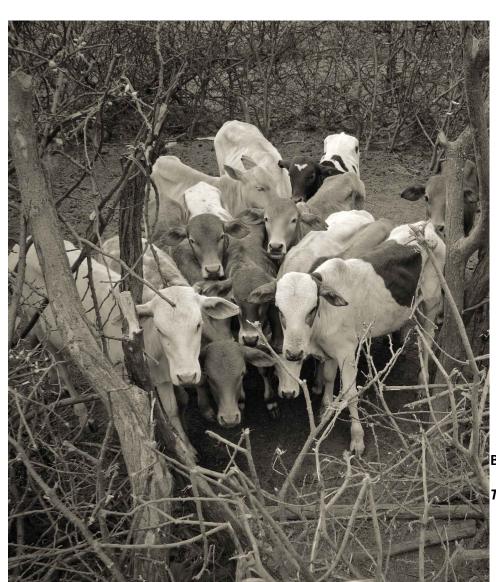
| Bin ID | Checkm Marker lineage | Sourmash Status | Sourmash phylum | Sourmash Class |
|------------------|----------------------------|--------------------|---------------------|--------------------|
| bin.19 | pBacteroidetes | found | pBacteroidota | cBacteroidia |
| bin.20 | kBacteria | nomatch | | |
| bin.21 bin.22 | pFirmicutes pFirmicutes | found nomatch | pFirmicutes_G | cLimnochordia |
| bin.23 | kBacteria | found | pCaldatribacteriota | cCaldatribacteriia |
| bin.24 | kBacteria | found | pFirmicutes | cBacilli |
| bin.25 | pFirmicutes | found | pFirmicutes_G | cSHA-98 |
| bin.26 | pFirmicutes | disagree | pFirmicutes_G | |
| bin.27 | pFirmicutes | found | pFirmicutes_E | cDTU015 |
| bin.28 | pFirmicutes | nomatch | | |
| bin.29 | pFirmicutes | found | pFirmicutes_A | cThermovenabulia |
| bin.30 | pFirmicutes | found | pFirmicutes_G | cLimnochordia |
| bin.31 | pFirmicutes | nomatch | | |
| bin.32 | pFirmicutes | found | pFirmicutes_F | cHalanaerobiia |
| bin.33 | pFirmicutes | found | pFirmicutes_D | cDethiobacteria |
| bin.34 | pFirmicutes | found | pFirmicutes_G | cDTU065 |
| bin.35 | kBacteria | nomatch | | 27 |

METHANE METABOLISM Ribulose-P pathway CoM-S-S-CoB Methane Formaldehyde **→**0 → D-Arabino-Hex-3-ulose-6P 4.1.2.43 D-Xylulose-5P 1.8.98.1 1.8.98 1.14.1325 1.14.183 D-Ribulose-5P 5.3.1.27 Methanogenesis D-Fructose-6P Pentose phosphate Coenzyme B Methanol MtaA 2.2.1.3 pathway #\$ Methyl-CoM 2.1.1.86 MtaB MtaC Coenzyme M 2.7.1.11 27.1.146 Trimethylamine metabolism Xylulose-P pathway Sulfur metabolism 1.1.2.7 MtbA MtbA MtbA 1.1.3.13 1.1.1.244 MtbB MtbC ItmB MtmC MttC Glycerone! Trimethylamine-N-oxide D-Fructose-1.6-P2 Glycerone-P 🟲 🗘 Methylamine Glycine, serine and Dime thylamine Trimethylamine Folate biosynthesis 4.1.2.32 2.1.1.21 3-Phosphono-L-Serine 3-Phosphoserine 1.4.9.1 N-Methyl-L-glutamate Formaldehyde 🔻 Te trahydrofolate 421.147 5.4.2.11 5.4.2.1 <u>∽⊸</u>Òthm(s)pt Carbon fixation 4.4.1.22 5,10-Methylene-te trahydrofolate 5,10-Methylene-THMP<u>T</u> S-Hydroxy-methyl-glutathione N5-Formyl-THMPT _ 1.12.982 Glycine Q Formy MFR Serine pathway 23.1.101 → ▶○ ◆ 1.5.98.1 1.5.98.2 2.7.1.165 Glyoxylate O 2.6.1.45 **—▶** Glycerate-2P 5,10-Methenyl-THMPT 1.2.98.1 1.1.1.284 D-Glycerate Methanofuran (MFR) 1.2.1.46 0-1.12.98.1 Glyoxylate and dicarboxylate metabolism F420 F420H2 Coenzyme F420H2 S-Formyl-4.1.3.24 Acetyl-CoA pathway 1.1.1.37 3.1.2.12 L-Malate Malyl-CoA Oxaloacetate Phosphoe nol-pyruvate 1.17.98.3 Acetyl-CoA CO2 Pyruvate Formate 4 Acetylphosphate Glyoxylate and dicarboxylate Nitrogen metabolism Coenzyme B biosynthesis 2-Oxoadipate cis-(Homo) 2ac onita te 2-Oxopime late 2-Oxosube rate cis-(Homo)3aconitate 7-Mercaptoheptanoate Coenzyme B Lysine biosynthesis -⊳o- 2.3.3.- -2.1.114 ▶○ 1.1.1.- ▶○ 2.3.3.- ▶○ 21.114 →○ 1.1.1.- →○ **▶**○ 2.7.1.- **▶**○ 7-Mercaptoheptanoyl-threonine (R)-(Homo)2citrate (-)-threo-Iso(homo)2citrate (R)-(Homo)3citrate (-)-threo-Iso(homo)3citrate 7-Oxoheptanoate Coenzyme M biosynthesis Coenzyme F420 biosynthesis Phosphoe nol-7,8-Didemethyl-8-hydroxy-5-deazariboflavin 5-Amino-(2R)-3-Sulfolactate Sulfoacetaldehyde Coenzyme F420-1 6-ribitylaminouracil O 4.4.1.19 ►O 3.1.3.71 ►O 1.1.1.337 - DO 251.147 DO 4.3.1.32 DO 2.7.8.28 ▶○ 6.3.2.31 ▶○ 6.3.2.34 ▶○ (2R)-Phosphosulfolactate Coenzyme M me tabolism 5-Amino-5-(4-hydroxybenzyl)-6-(D-ribitylimino)-5,6-dihydrouracil Coenzyme F420-0 6.3.2.32 mfnA 4.1.1.25 mfnD 6.3.4.24 Methanofuran biosynthesis Tyrosine metabolism → y-Glutamyltyramine O 2.7.7.68 → (2S)-Lac tyl-2-diphospho-Coenzyme F420-3 (2S)-2-Phospholactate 423.153 - - - ► Methanofuran Glycolysis 5-(Aminomethyl)-3-furanmethanol-2F Glyceraldehyde-3P 4-HFC-P 5-(Aminomethyl)-3-furanmethanol-F APMF-Glu 00680 2/28/19 (c) Kanehisa Laboratories

Results Annotation Methane Metabolism Bins X RNAseq Data

METHANE METABOLISM CoM-S-S-CoB Ribulose-P pathway Methane Formaldehyde D-Arabino-Hex-3-ulose-6P 4.1.2.43 D-Xvlulose-5P 1.8.98.1 1.14.1325 1.14.183 D-Ribulose-5P 5.3.1.27 Methanogenesis D-Fructose-6P Coenzyme B Pentose phosphate Methanol MtaA 2.2.1.3 pathway ♣ Methyl-CoM 2.1.1.86 MtaC Coenzyme M 27.1.146 Xylulose-P pathway Trimethylamine me tabolism Sulfur metabolism D-Glycer-aldehyde-3P 1.1.2.7 MtbA MtbA MtbA 1.1.3.13 1.1.1.244 MtbC MtmC MttC Trimethylamine-N-oxide D-Fructose-1,6-P2 Glycerone-P 11413148 Methylamine Glycine, serine and threonine metabolism Dime thylamine Trimethylamine Folate biosynthesis 4.1.2.32 —Ò Glycerate-3P 2.1.1.21 3-Phosphono-oxypyruvate L-Serine 3-Phosphoserine 1.4.9.1 N-Methyl-L-glutamate Formaldehvde V Te trahydrofolate 421.147 5.4.2.11 OTHM(S)PT Carbon fixation 4.4.1.22 5,10-Methylene tetrahydrofolate 5,10-Methylene-THMPT S-Hydroxy-me thyl-glutathione N5-Formyl-THMPT Glycine Q 1.12.982 Serine pathway 231.101 1.5.98.1 1.5.98.2 27.1.165 5-Methyl-THM(S)PT Glyoxylate O → Glycerate-2P 5,10-Methenyl-THMPT 1.2.98.1 1.1.1 284 Hydroxy-D-Glycerate Methanofuran (MFR) pyruvate 1.2.1.46 Glyoxylate and dicarboxylate metabolism F420 F420H2 Coenzyme F420H2 S-Formyl-4.1.3.24 Acetyl-CoA pathway 3.1.2.12 L-Malate Malyl-CoA Oxaloacetate Phosphoe nol-1.17.983 🙀 Acetyl-CoA CO2 Pyruvate Formate Acetylphosphate Acetate Glyoxylate and dicarboxylate metabolism Nitrogen metabolism Coenzyme B biosynthesis 2-Oxoadipate 2-Oxopime late 2-Oxosuberate cis-(Homo)3aconitate 7-Mercaptoheptanoate Coenzyme B Lysine ►O 1.1.1.- ►O 2.3.3.- ►O ->O- 2.3.3.-**▶**0—1.1.1.- **▶**0-**▶**0 2.7.1.- **▶**0 biosynthesis (-)-threo-Iso(homo)2citrate (-)-threo-Iso(homo)3citrate 7-Mercaptoheptanoyl-threonine (R)-(Homo)2citrate (R)-(Homo)3citrate 7-Oxoheptanoate Coenzyme M biosynthesis Coenzyme F420 biosynthesis Phosphoe nol-7,8-Didemethyl-8-hydroxy-5-deazariboflavin 5-Amino-6-ribitylaminouracil (2R)-3-Sulfolactate Sulfoacetaldehyde Coenzyme F420-1 Coenzyme F420 O 4.4.1.19 →O 3.1.3.71 →O 111.337 →O 4.1.1.79 →O Riboflavin 2.7.8.28 ▶○ 6.3.2.31 ▶○ 6.3.2.34 ▶○ (2R)-Phosphosulfolactate 3-Sulfopyruvate Coenzyme M me tabolism 5-Amino-5-(4-hydroxybe nzyl)-6-(D-ribitylimino)-5,6-dihydrouracil Coenzyme F420-0 6.3.2.32 mfnA 4.1.1.25 Methanofuran biosynthesis Tyrosine metabolism 6.3.4.24 → y-Glutamyltyramine O 2.7.7.68 → O (2S)-Lac tyl-2-diphospho-Coenzyme F420-3 (2S)-2-Phospholactate 423.153 ►O———► Methanofuran Glycolysis 5-(Aminomethyl)-3-furanmethanol-2P APMF-Glu Glyce ralde hyde-3P 4-HFC-P 5-(Aminomethyl)-3-furanmethanol-P 00680 2/28/19 (c) Kanehisa Laboratories

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

