INTEGRATIVE MULTIOMIC IN GALAXY

24 APRIL, 2023
9:00 AM - 17:00 PM

Introduction to Microbiome Research and Metaproteomics
- Introduction to Basic Metaproteomics Workflow within Galaxy
- Introduction to the meta-omics / microbiome multi-omics
- Define the inputs for the microbiome multi-omics and start the workflow
- Discuss the multi-omics workflow outputs and visualization
- Overview of Informatics Platforms available for functional microbiome analysis
Workshop 1 – Integrative Multiomics in Galaxy

09:00 - 09:10 Welcome and quick survey
09:10 - 10:10 Introduction to Microbiome Research and Metaproteomics
10:10 - 10:30 Introduction to the meta-omics /microbiome multi-omics

10:30 - 10:50 Coffee Break

10:50 - 11:50 Introduction to the meta-omics /microbiome multi-omics (continued)

11:50 - 14:00 Lunch Break

14:00 - 14:30 Discuss the multi-omics workflow outputs
14:30 - 14:50 Case study - Biogas reactor
14:50 - 15:40 ViMO analysis

15:40 - 16:00 Coffee Break

16:00 - 16:30 Overview of Informatics Platforms available for functional microbiome analysis
16:30 - 16:50 Summary, Q&A and evaluation
Welcome and quick survey

https://www.surveylegend.com/s/4vym

QUESTIONS?

http://z.umn.edu/workshop1Qs
Introduction to Metaproteomics & Galaxy Platform

Pratik Jagtap
University of Minnesota

Learn more at galaxyp.org
z.umn.edu/itcrgalaxyvideo

International Metaproteomics Symposium
24th April, 2023

https://z.umn.edu/workshop1Part1
Workshop acknowledgements

• **Instructors**
  • Pratik Jagtap

• **Other contributors**
  • Tim Griffin
  • Subina Mehta
  • James Johnson
  • Andrew Rajczewski
  • Reid Wagner
  • Katherine Do
  • usegalaxy.eu team
  • Galaxy community

https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaproteomics/tutorial.html
Mass Spectrometry and Proteomics

1. Sample preparation
2. Separation
3. Data analysis
4. MS1
5. MS2

HPLC, Electro spray

Intensity vs. Retention time

m/z vs. Isolation window

MICROBIOME RESEARCH

Microbiome IN NUMBERS

100 Trillion
The human body has more microbes than there are stars in the milky way

95%
of our microbiota is located in the GI tract

150:1
The genes in your microbiome outnumber the genes in our genome by about 150 to one

5:1
Viruses:Bacteria in the gut microbiota

You have 1.3X more microbes than human cells

>10,000
Number of different microbial species that researchers have identified in and on the human body

2kg
The gut microbiota can weigh up to 2kg

https://worldmicrobiomeday.com/resources/

https://www.nature.com/articles/d41586-020-00193-3
Potential to unravel the mechanistic details of microbial interactions with host / environment by analyzing the functional dynamics of the microbiome.
"Through the application of metaproteomics to different microbial consortia over the past decade, we have learnt much about key functional traits in the various environmental settings where they occur."

microbial taxa vary while metabolic pathways remain stable within a healthy population

Metaproteomics Analytical Challenges

**SINGLE-ORGANISM PROTEOMICS**

- **SEARCH DATABASE SIZE**: SMALL TO MEDIUM SIZE (10 K TO 100K SEQUENCES)
- **SEARCH DATABASE COMPLEXITY**: SINGLE + CONTAMINANTS

**METAPROTEOMICS**

- **SEARCH DATABASE SIZE**: LARGE (1 MILLION AND ABOVE)
- **SEARCH DATABASE COMPLEXITY**: MULTI-ORGANISM DATABASE WITH HOMOLOGOUS PROTEINS

- **SEARCH ALGORITHMS BEING DEVELOPED TO ADDRESS LARGE AND COMPLEX DATABASE SEARCHES**
- **PROTEIN GROUPING AT MULTI-ORGANISM LEVEL**
- **IDENTIFICATION STATISTICS AFFECTED BY LARGE DATABASES**
- **TAXONOMY BASED ON UNIQUE PEPTIDE IDENTIFICATIONS**
- **FUNCTIONAL ANALYSIS BASED ON PROTEINS IDENTIFIED**

Disparate tools and multiple processing steps.
Metaproteomics Workflow

**DATABASE GENERATION**
- FASTQ

**DATABASE SEARCH & STRATEGIES**
- FAST
- Spectra
- Mass Spectrometry Data
- Search Algorithm

**FUNCTIONAL ANALYSIS**
- Peptides
- Known Function
- Proteins

**TAXONOMY ANALYSIS**
- Unique
- Shared
- Unassigned
Please Register for creating an account with a valid email ID and Password at usegalaxy.eu.

Once Registered, click on TIAAS to join the IMS 2023 Galaxy session. [https://usegalaxy.eu/join-training/ims-multiomics](https://usegalaxy.eu/join-training/ims-multiomics)
SHARED DATA HISTORIES & WORKFLOWS

Galaxy Europe

Tools
- search tools
- Upload Data

Get Data
- Send Data
- Collection Operations
- GENERAL TEXT TOOLS
- Text Manipulation

Saved Histories
- Advanced Search

History
- search datasets

Outputs from Metaproteomics-GTN
- Access
- 5.96 MB
- 20
- 2

Published Histories
- search name, annotation and tags

Published Workflows
- search name, annotation, own

Name
- Annotation
- Owner
- Community Rating

Inputs for Metaproteomics-GTN
- pratikjagtap

Metaproteomics_GTN_IMS2623
- pratikjagtap

Using 2.1 TB
RUN THE WORKFLOW ON AN ACTIVE HISTORY

Workflow: Metaproteomics_GTN_IMS2923

Sixgill generated protein FASTA File *
- 4: FASTA_Bering_Strait_Trimmed_metapeptides_cRAP.fasta

Dataset Collection of Bering Strait MGF Files *
- 9: Bering Strait

Gene Ontology Terms (Selected) *
- 5: Gene_Ontology_Terms_full_97.98.2917.tabular

Expand to full workflow form.
**HANDS-ON SESSION**

**Instructions**

Please **Register** for creating an account with a valid email ID and Password at usegalaxy.eu.

Once Registered, click on TIAAS to join the IMS 2023 Galaxy session. [https://usegalaxy.eu/join-training/ims-multiomics](https://usegalaxy.eu/join-training/ims-multiomics)

Go to Shared Data Published Histories

Go to Shared Data Published Workflows

Run the workflow on active history

---

**Input history:** [https://usegalaxy.eu/published/history?id=81074d4fed49eff1](https://usegalaxy.eu/published/history?id=81074d4fed49eff1)

**Import workflow:** [https://usegalaxy.eu/published/workflow?id=0e0be913c3e27022](https://usegalaxy.eu/published/workflow?id=0e0be913c3e27022)

**Backup Output history with results:** [https://usegalaxy.eu/published/history?id=f89e3be34ed92acf](https://usegalaxy.eu/published/history?id=f89e3be34ed92acf)
SHARED DATA HISTORIES & WORKFLOWS
Sample Collection
Water samples were collected from the Bering Strait and Chukchi Sea and oceanic marine bacteria retained on a 0.7 µM filter.
SearchGUI matches MS/MS spectra to peptide sequences
• SearchGUI allows for multiple search engines to run simultaneously
• Specific digestion conditions can be selected
• Mass spectrometer parameters can be selected to maximize the efficacy of spectral matches
• Post-Translational Modifications (PTMs) can be added to the search parameters
PeptideShaker filters SearchGUI results.

- Search GUI results are filtered by FDR to yield most confident peptide spectral matches (PSMs)
- Peptide Shaker generates outputs such as Protein Report, Peptide Report and mzIdentML files for subsequent analysis.

Unipept

**Functional Analysis**

- Known Function
- Proteins
- Peptides
- Unassigned
- Shared
- Unique

**Taxonomy Analysis**


Diagram showing the relationship between peptides, proteins, taxa, and taxonomy.
Unipept

**FUNCTIONAL ANALYSIS**

- Known Function
- Peptides
- Proteins
- Unassigned
- Shared
- Unique

**TAXONOMY ANALYSIS**

Taxonomic and Functional analysis with Unipept

https://unipept.ugent.be/publications
Who is there?

Get a taxonomy report from PSM report and Unipept pept2lca table

<table>
<thead>
<tr>
<th>peptide</th>
<th>superkingdom</th>
<th>genus</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADGHTMHFDVITGEK</td>
<td>Archaea</td>
<td>Nitrosopumilus</td>
</tr>
<tr>
<td>AALESFTGNVTSALK</td>
<td>Bacteria</td>
<td>Polaribacter</td>
</tr>
</tbody>
</table>

**SQLite Relational Data base**

**SQL** query joins PSM and LCA to report number of PSMs and Peptides per genus

<table>
<thead>
<tr>
<th>PSM#</th>
<th>Proteins</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EGGEDMFVHKSDV...</td>
<td>EGGEDMFVHK</td>
</tr>
<tr>
<td>2</td>
<td>GKRVAAAAVGTVPEQ...</td>
<td>VAAAVGTVPEQEWLK</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>genus</th>
<th>PSMs</th>
<th>DISTINCT PEPTIDES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planktomarina</td>
<td>161</td>
<td>20</td>
</tr>
<tr>
<td>Nitrosopumilus</td>
<td>122</td>
<td>27</td>
</tr>
</tbody>
</table>
Who is there? What are they doing?

How do we get taxonomy and function of a microbiome from a list of peptides?

<table>
<thead>
<tr>
<th>PSM#</th>
<th>Proteins</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EGGEDMFVHKSDVDGFINEGDK</td>
<td>EGGEDMFVHK</td>
</tr>
<tr>
<td>2</td>
<td>GKRVAANVTVPEQEWLK, KVAAAVTVPQEWLK, RVALAAVTV...</td>
<td>VAAAVTVPEQEWLK</td>
</tr>
</tbody>
</table>

**Pept2lca**
taxonomy lowest common ancestor for a peptide

**Pept2prot**
Uniprot entries for a peptide with GO terms

<table>
<thead>
<tr>
<th>peptide</th>
<th>superkingdom</th>
<th>genus</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADGHTMHFDVITGEK</td>
<td>Archaea</td>
<td>Nitrosopumilus</td>
</tr>
<tr>
<td>AALESFTGNVTSALK</td>
<td>Bacteria</td>
<td>Polaribacter</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>peptide</th>
<th>uniprot_id</th>
<th>...</th>
<th>go_references</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAEKLSAAQAR</td>
<td>W5T6F9</td>
<td></td>
<td>GO:0016021</td>
</tr>
<tr>
<td>AAEKLSAAQAR</td>
<td>A0A0Q6ZKK0</td>
<td></td>
<td>GO:0005524 GO:0016887 GO:0015833</td>
</tr>
</tbody>
</table>
Questions

• How can I match metaproteomic mass spectrometry data to peptide sequences derived from shotgun metagenomic data?

• How can I perform taxonomy analysis and visualize metaproteomics data?

• How can I perform functional analysis on this metaproteomics data?

https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaproteomics/tutorial.html
Hands On Session

https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaproteomics/tutorial.html

https://usegalaxy.eu

- Download and start workflow
- Observe outputs
- Tool Basics
META-OMICS APPROACH BY NMBU TEAM

METAGENOMICS

METATRANSCRIPTOMICS

METAPROTEOMICS

https://galaxyproject.eu/posts/2020/04/14/integrative-meta-omics/

Magnus Arntzen
Integrative meta-omics in Galaxy

Dr. Magnus Ø. Arntzen
Norwegian University of Life Sciences (NMBU), Norway
The meta-omics

Microbial community

DNA
RNA
Proteins
Metabolites
Stable/radioactive isotopes

16S rRNA gene
Metagenome
Metatranscriptome
Metaproteome
Metabolome
MAR/[Nano]SIMS

Gene expression profiles
Protein identification
Metabolic fluxes
Visualization of metabolically active cells

Who are there?
What can they do?

What are they actually doing?
Who is doing what and together with whom?

Potential
Expression
Activity

Metabolic function and pathway reconstruction

The multi-omics toolbox

1. **Microbiome structure, 16S rRNA**

2. **Metagenomics**
   - Shotgun sequencing
   - Assembly
   - Binning -> generation of metagenome-assembled genomes (MAGs)
   - Taxonomic annotation
   - Functional annotation

3. **Metatranscriptomics**
   - Sequencing
   - Mapping to metagenome

4. **Metaproteomics**
   - Mass spectrometry
   - Mapping to metagenome

5. **Meta-metabolomics**
   - Targeted analysis of relevant components:
     - Polymers
     - Sugars
     - SCFAs
     - etc…

**Important aspects:**
- Preserving gene identifiers throughout all ‘omics
- Metagenomics is the base omics. Quality affects all other steps.
Integrated meta-omics (simplified workflow)

**metaG**
- Reads
- Contigs
- Quality control
- Trimming
- Assembly
- Gene calling
- Proteins
- Translating
- Genes
- Quality control
- Taxonomy

**metaT**
- Reads
- Non-ribosomal RNA
- Quality control
- Trimming
- Sorting
- Mapping
- Quantified mRNA
- Functional annotation
- Pfam, CAZy, EC, pathways

**metaP**
- MS spectra
- Pre-processing
- Accurate mass
- Retention time
- Fragment pattern
- Charge
- Quantified proteins
- Integrating
- & interpreting

15 software & bunch of scripts
+ alternative approaches
Metagenomics assembly and binning

Dr. Magnus Ø. Arntzen
Norwegian University of Life Sciences (NMBU), Norway
DNA extraction and shotgun sequencing

Metagenomics reads

Mapping

Reference genome 1

Reference genome 2

NCBI

No reference genome?
Unmapped and lost!
Microbiome

**DNA extraction and shotgun sequencing**

**Metagenomics reads**

Mapping

Reference genome 1

Reference genome 2

No reference genome? Unmapped and lost!

**De-novo assembly**

Reads provided to algorithm

Overlaps identified

Contigs

Binning

Metagenome-Assembled Genomes (MAGs)
Reconstruction of the genomes present in our sample

- Who’s there? Detect novel microbes
- What can they do? (genes → metabolic pathways)
- What are they doing (combining multi’omics)
Metagenomics – the binning

Contigs originating from the same genome
1. Have the tendency to exhibit similar sequence composition (k-mers)
2. Should have the same coverage across samples (e.g., time points)

$\begin{array}{ccccccccccc}
& AAAG & AAGG & AGGC & GCGC & GCGT & CGT & GGT & TTGA & TGAG & GAGG & AGGT & GGT
\\
contig1 & 1 & 1 & 1 & 2 & 2 & 1 & 0 & 2 & 1 & 2 & 1 & 2
\\
contig2 & 2 & 1 & 2 & 3 & 2 & 0 & 1 & 2 & 2 & 2 & 2 & 3
\\
contig3 & 1 & 2 & 0 & 0 & 2 & 3 & 1 & 2 & 9 & 9 & 2 & 5
\\
contig4 & 1 & 2 & 0 & 0 & 2 & 3 & 1 & 2 & 8 & 10 & 2 & 5
\\
contig5 & 3 & 1 & 5 & 6 & 0 & 3 & 1 & 2 & 0 & 0 & 2 & 4
\\
contig6 & 3 & 2 & 5 & 7 & 0 & 3 & 1 & 2 & 1 & 0 & 2 & 5
\\
\end{array}$

$k=4$ is often called tetranucleotide frequency

Short read to k-mers ($k=4$)

$k=4$ is often called tetranucleotide frequency

Coverage $\approx 2x$

Coverage $\approx 3x$

Coverage $\approx 4x$

Coverage $\approx 7x$
Mapping metatranscriptomics and metaproteomics to metagenomics

Dr. Magnus Ø. Arntzen
Norwegian University of Life Sciences (NMBU), Norway
Metatranscriptomics

RNA extraction, removal of residual DNA and tRNA, enrichment of mRNA and shotgun sequencing

Metatranscriptomics reads

By mapping to metaG, the gene_ids are preserved and makes the multi-omics easier.

Mapping always involves some level of ambiguity

Unique hit: The read matches one unique ORF
Shared hit: The read matches >1 ORF

Adding a quantitative «layer» on top of our metagenomics.

Which genes are expressed in each MAG and at what abundance?
Metaproteomics

Microbiome

Protein extraction, SDS-PAGE/fractionation, digestion to peptides and clean-up

Mass spectrometry

Peptide-to-spectrum matches (PSMs)

But what about taxonomic inference?
What about homologues proteins between species?

UniProt

Fasta
Database

Candidate peptides

Fragment ions

Mapping always involves some level of ambiguity
and one more level for proteomics:
MS/MS -> peptide
Peptide -> protein

Unique hit: The peptide matches one unique protein
Shared hit: The peptide matches >1 protein, also called razor peptide

Protein inference

Matched peptide sequence

Protein group 1

Protein group 2

Protein group 3

B1 B2 B3

MAGs

But what about taxonomic inference? What about homologues proteins between species?
**Meta-proteomics challenges**

1. **Database**
   - *Comprehensive*
     Contain the actual species in the sample, including strain variances
   - *Dedicated*
     Not all other species!

   ![UniProt-Bacteria 58M](image)

   **Large** – few significant hits
   **Incomplete** – only a few of the species have been sequenced
   **Low ID rate!!**

2. **Peptide-to-Protein-to-Species inference**

   **LAIVTTNSILSDLVK** – 15 AA
   Common lineage: **Bacteria** > **Firmicutes** > **Bacilli** > **Lactobacillales** > **Enterococcaceae** > **Enterococcus** > **Enterococcus faecalis**

   **DTVASNEK** – 8 AA
   Common lineage: **Bacteria** > **Firmicutes** > **Bacilli** > **Lactobacillales** > **Enterococcaceae** > **Enterococcus**

   **DVLVEK** – 6 AA
   Common lineage: **Nothing**

   **Short peptides are less informative**
Genome resolved meta-omics

Meta-proteome

Genome database

Isolates
Genome resolved meta-omics

Unmapped data is lost!!!
Meta-proteome

Meta(genome) database

Isolates

MAGs: Metagenome Assembled Genomes

Genome resolved meta-omics
Workshop: Integrative multi-omics in Galaxy

Valerie Schiml
valerie.schiml@nmbu.no
24.04.23
Metagenomics workflow
FastQC

<table>
<thead>
<tr>
<th>Label</th>
<th>Header of the tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step Annotation</td>
<td>Description</td>
</tr>
</tbody>
</table>

Add a step label.

Annotations are available when a workflow is viewed.

Conditionally skip step?

- No

Set to true and connect a boolean parameter that determines whether step will be skipped.

If on: reports show data for every base in the read.

Specify length of read group

If on: reports show data for every base in the read.

Galaxy standard is 7

If on: output deleted if not further needed in workflow.
FastQC

- **Name output**: Provide a short, unique name to describe the output.
- **Name dataset to follow along the workflow**: This action will rename the output dataset. Click here for more information. Valid input variables are:
  - `input_file` (Raw read data from your current history)
  - `contaminants` (Contaminant list)
  - `adapters` (Adapter list)
  - `limits` (Submodule and Limit specifying file)
- **Specify which datatype output should be**: This action will change the datatype of the output to the indicated datatype.
- **Can be useful for tracking and organizing multiple reports**: This action will set tags for the dataset.
- **Specify position of different output columns**: This action will remove tags for the dataset.
TrimGalore!

- **Set input type**
- **Set adapter sequence**
- **Set specific 3’ end trimming for read 1 and read 2:**
  - e.g. useful for remove biased residues from start sequence
- **Trim low-quality ends from reads in addition to adapter removal**
- **A single base pair of overlapping sequence will be trimmed of the 3’ end of any read**
- **Number of errors divided by the length of the matching region**
- **Min. read length**
- **Set specific 5’ end trimming for read 1 and read 2:**
  - e.g. useful for shortening reads to specific length
TrimGalore!

**RRBS specific settings**

- **Default → no RRBS**
- **RRBS does not have to be specified when DNA digested with Msel**
- **If yes: removes illumina sequencing artifact poly-C or poly-G runs**

**Trimming settings**

- **Specific trimming length**
- **Specific trimming length**
- **Trimming in the way as used for Mouse Epigenetic Clock**
- **For rRNA trimming**
MetaG-collection-operation

Trimmed reads

Separate forward and reversed reads

Combines list collection to a single file

Flattens nested collection to a single list: used for MaxBin2
Minimum multiplicity for filtering:
- single base of sequencing error → false k-mer singletons
- more memory needed
- count and sort k-mers before De Bruijn graph construction

Large k-mer: help construction of the genome
Small k-mer: better chance for overlapping, loose information
MEGAHIT

For low abundant species sequenced with low depth → Strengthen coniguity for low-depth regions

Remove bubbles in De Bruijn graph Based on length and similarity (redundant edges)

Pruning spurious branches introduced by sequencing errors

Remove high-fidelity copies of a repeat, potential information loss but simplification of De Bruijn graph

Pruning for minimal connection strength; might still be connected to other edges

Delete edges based on minimal connection strength; When 0 less pruned

Less pruned assembly when low (usually shorter contigs)

If sequencing depth is low and too much memory used when build the graph of k_min

Set min. contig length for output

See in between steps
CoverM - contig

**Standard:** Short single-end reads without splicing

Removes potential bias for the genome

Disadvantage: loose potential information

If counting expressed genes for different treatments multiple alignments bias the analysis → false positives

Reduces reads to fasten up analysis

Disadvantage: loose potential information

Activate for contig: Average number of aligned reads overlapping each position on the contig
CoverM - genome

Choose input type

Collection of paired-reads:
Data input 'paired_reads' (fasta, fastq, fastaq.gz or fastagz)
One or more pairs of forward and reverse possibly gzipped FASTA/Q files for mapping in order.

Reference sequence mode

Choose input MAG after binning or contigs belonging to the same genome

Reference genome source

Raw reads coming from the workflow

Low quality, incompleteness, specific research question

Usually run in a separate tool like dRep or dasTool

Coverage calculation options

- Relative abundance (default)
- Mean
- No
Quast

Tool Parameters

- Use customized names for the input files?
  - No, use dataset names
  - They will be used in reports, plots and logs
- Contigs/scaffolds file:
  - Data Input 'inputs' (fasta)

Reads options

- Illumina paired-end reads
  - Currently, the supported read types: Illumina paired-end and mate-pair reads, PacBio SMRT, and Oxford Nanopore long reads.
- FASTQ/FASTA file #1
  - Data Input 'input_1' (fastq, fastaq, gz, fasta or fasta.gz)
- FASTQ/FASTA file #2
  - Data Input 'input_2' (fastq, fastaq, gz, fasta or fasta.gz)

Type of assembly

- In this case: Metagenome

Reference genome

- In this case: Reads are the mapping reference

Select input names

Read type

In this case: Metagenome

Saves time when on: avoids realigning assembly to reference

Default for MetaQUAST: 90% identity for alignment

Default min contig length: Too short contig contain more errors that may affect the accuracy of the assembly evaluation

In this case: only contigs used

In this case: Reads are the mapping reference

Minimum ID% considered as proper alignment

Default min contig length: Too short contig contain more errors that may affect the accuracy of the assembly evaluation

In this case: only contigs used
Quast

Espacially for fungi/eukarya

In this case: use different tools

In this case: prokaryots

Espacially for fungi/eukarya

Default: use only one alignment per contig

Filter out too short, less reliable alignments

Save time, loose information

For case of multiple alignments for contig to reference genome with same alignment score: S score calculation determines best contig alignment

Detect misassemblies from fragmentation
Use of simulated assemblies for evaluating against the reference genome

For upper bound contigs: realistic estimation of genome assembly fragmentation due to long reads

Min. size for misassembled bp to count as a separate misassembly
MaxBin2

In this case: co-assembly

Optional: abundances e.g. read depth from Minimap2 → Faster binning

Min of 500 bp
Small: more likely to detect small proteins
Large: better MAG quality

Estimation of abundance of each MAG based on coverage and composition of DNA sequences

Threshold for a minimum probability value for a contig to be assigned to a MAG → Specificity vs Sensitivity

Define output for marker genes
CAT bins – Taxonomic annotation of MAGs

- **Tool Parameters**
  - Metagenome assembled genomes (MAGs/bins)
  - Data Input ‘mag’ (fasta)

- **CAT database**
  - Local cached database
  - Use a built-in CAT database
  - Define reference db

- **Use previous prodigal gene prediction and diamond alignment**
  - No
  - predicted_proteins.faa and alignment.diamond from previous CAT run.

- **Contig range for analysis**
  - 10
  - Min. fraction of reads per contig
  - 6.5

- **Set advanced diamond options**
  - Run DIAMOND in sensitive mode (considerably slower)
  - DIAMOND block-size parameter
  - DIAMOND index-chunks parameter
  - DIAMOND top parameter

- **High**:
  - Fast search
  - Slow indexing

- **Low**:
  - Slow search
  - Fast indexing

- **Number of best hits for output**
  - Improves taxonomic annotation
  - Longer computer time

- **Number of sequences loaded at once in memory for db search**
  - Stricter alignment algorithm
  - Long computer time

- **Number of chunks ref db is split into during indexing**
  - High: fast search, slow indexing
  - Low: slow search, fast indexing
CAT bins – Taxonomic annotation of MAGs

- Taxonomic names listed
- Only MAGs in output annotated with an official taxonomic level
- Information about quality scores for MAGs: we use CheckM for that
- Contig and taxonomy information
- Select output
CheckM

Select data structure

Speed up analysis, evtl less accurate

Alignment can help functional annotation

Nucleotide sequences for each predicted gene in MAG

In this case: MaxBin2 output is nucleotide contigs

Classify and use for calculation of completeness, contamination and heterogeneity

If lineage-specific marker set too low

Reduce analysis time but less specific

Refinement improves accuracy of classification
Co-located marker genes may overrepresent functions

Exclusion for contigs come from different species

Skipping causes inaccurate calculation of contamination and completeness

Cut-off for similarity between genomes

Reduce false positives

Cut-off for marker gene identification with BLAST

Overlap of sequence with db to count as true positive

Select additional output
Workflow for functional annotation

Gene functional annotation

Function analysis of proteins

Combining and formatting annotations to one input file for ViMO

Gene prediction

Compare putative genes to protein db: CAZy db
FragGeneScan

- **Tool Parameters**
  - Input sequence file
  - Data Input: genome (fasta)
  - Does the sequence file have complete genomic sequences?
    - No
  - Model
    - 454 pyrosequencing reads with ...
    - (train)

- **Additional Options**
  - Email notification
    - No
    - An email notification will be sent when the job has completed.
  - Output cleanup
    - No
    - Upon completion of this step, delete unchecked outputs from completed workflow steps if they are no longer required as inputs.

Define output names and type:
- Configure Output: "coind"
- Configure Output: "hit_seq"
- Configure Output: "prot_seq"
- Configure Output: "gif"
InterProScan

Tool Parameters

- **Protein FASTA File**
- **Data Input**
  - **Type of the input sequences**
    - Protein
  - **InterProScan database**
    - InterProScan 5.59-91.0
  - **Applications to run**
    - Select/Unselect all

Select your program

- **TIGERAM**
  - Protein families based on hidden Markov models (HMMs)
- **HAMAP**
  - High-quality automated Annotation of Microbial Proteomes
- **PIGRA**
  - Pfam protein families, each represented by multiple sequence alignments and hidden Markov models

Select output

- **Include pathway information**
- **Include Gene Ontology (GO) mappings**
- **Provide additional mappings**

Select output format

- **Tab-separated values format** (TSV)

Additional annotations that must be installed manually by admin of this Galaxy instance

The corresponding tools must be installed manually by the administrator of this Galaxy instance
Hmmscan and CAZy annotation

Choose output tables

Have acc number in first column and name in second column

Choose if alignment information needed

Output in a continuous line of tab separated text

Define output criteria

Choose output tables

Have acc number in first column and name in second column

Choose if alignment information needed

Output in a continuous line of tab separated text

Define output criteria

Define analysis criteria via E-value or score threshold

Set gathering cutoff for true positives

Set noise cutoff for false positives

Set trusted cutoff for true positives

(see settings above); filters may bias analysis
Hmmscan and CAZy annotation

Multiple segment viterbi: measure similarity between sequence an HMM profile

Viterbi: calculate highest probability path through HMM profile

Forward threshold: calculate probability of sequence generated by HMM profile

Remove sequences having biased AA/nucleotide composition → cause false positives
On for large db or sequences with high biased composition

Adjusting similarity scores of sequence to account for the bias

Per sequence E-value calculation instead of only for the targets passing the reporting threshold

Per-domain conditional E-value calculation instead of only for the targets passing the reporting threshold

Reproducibility
Choose input

Likelihood of match between a query AA sequence and protein profile

Adjust E-value cut-offs for each functional category based on their confidence and specificity
Metatranscriptomics workflow
Workflow

1. Collection of paired reads
2. FastQC
   - Short read data from your current history
   - Contaminant list
   - Submodule and Limit specifying file
     - html.txt (htm)
     - text_file (txt)
3. Trim Galore!
   - Select a paired collection
     - trimmed_reads_paired_collection (input)
4. Linseq Collection
5. Unmap collection
   - Paired input to unmap
     - forward (input)
     - reverse (input)
6. Filter with SortMeRNA
   - Forward reads
     - aligned_forward (input)
     - unaligned_forward (input)
   - Reverse reads
     - aligned_reverse (input)
     - unaligned_reverse (input)
   - output_log (txt)
7. Kallisto quant
   - FAS7E reference transcriptome
   - Forward reads
   - Reverse reads
   - abundance (h5)
   - abundance (tab)
   - output (tabular)
8. Column join
   - Tabular files
   - tabular_output (tabular)

Input

Adapter trimming

Split collection into forward and reversed reads

Quality control

Filter rRNA (not informative)

Predicted nucleotide sequences from FragGeneScan (MetaG)

mRNA quantification

Data formatting
In this case: paired reads

If not: decreased sensitivity and specificity

Information about sorting process

Alternative: best or first alignment per read

Sequence similarity, low → more stringent

Alignment score for query to ref sequence <

Score penalty for mismatch

Score penalty for gap in aligned sequence

Score penalty for extending gap

Score penalty for uncertain nucleotides

Default: SILVA, different db usable based on research question
Putative genes from MetaG

In this case: paired

Individual forward/reverse files from SortMeRNA

Sorting transcripts into classes based on De Bruijn graph

→ bias correction on each class
→ accurate estimation of transcript abundance

Repeats of bootstrapping

Reproducability

Detection and information of chimeric transcripts: e.g. for cancer research

Library can be generated from forward or reversed strand

Select if output in BAM format
Metaproteomics workflow
MaxQuant

Parameter file can be reused for running MQ

Information about identified proteins

Parameter file can be reused for running MQ

Information about identified peptides

Additional information about identified peptides

Describes parameters used

Information about peptide modifications

Information about MS/MS scans

Quality of raw file data
ViMO
β-Lactam resistance

Hussain, 2021
Resistance strategies

Deactivation by β-lactamases

- NagZ: muropeptide cleavage;
  GlcNAc linked to aMurNAc-peptides cleaving murein for recycling; cleavage product represses BlaI
- BlaZ: β-lactamase
  → β-lactamases class A
Resistance strategies

Modified PBP side: lower affinity for most β-lactam antibiotics

- β-lactam mimicks the natural D-Ala-D-Ala substrate of the enzyme family of penicillin-binding proteins (PBP)
- PBP cross-links peptidoglycan of bacterial cell wall
- β-lactam AB form complex with PBPs
  → inhibiting their transpeptidation activity
  → Disrupting cell wall integrity
    → Cell lysis
Resistance strategies

Modified PBP side: lower affinity for most β-lactam antibiotics

Expressed:
- PBP2: Catalyzes cross-linking of peptidoglycan cell wall
- PBP1a: Synthesis of cross-linking of peptidoglycan from lipid intermediates
- FtsI: Cross-linking of peptidoglycan in the cell wall (division septum)

Vigouroux, 2020
Wissel, 2004
References


The example dataset

Dr. Magnus Ø. Arntzen
Norwegian University of Life Sciences (NMBU), Norway
Case study – Biogas reactor

Biogas-plant (60°C) Fredrikstad, Norway

Lab-scale reactor (55°C)

Anaerobic bottles (65°C)

Food waste Manure

Food waste Manure

Cellulose

Serial dilution

0h
8h
13h
18h
23h
28h
33h
38h
43h

Triplicate s of all time points

Metagenomics Illumina HiSeq 3000, 2 x 125 bp

Metatranscriptomics Metaproteomics
Integrating meta-omics data, the various types of data tables

Dr. Magnus Ø. Arntzen
Norwegian University of Life Sciences (NMBU), Norway
The multi-omics toolbox

1. Microbiome structure, 16S rRNA
2. Metagenomics
3. Metatranscriptomics
4. Metaproteomics
5. Meta-metabolomics

What do we have:
- FASTA-files of genes & proteins per MAG
- Table of all MAGs present in sample with taxID
- Table of all genes per MAG with
  - Functional annotation (KEGG, Pfam, InterPro, etc…)
  - Expression levels (transcripts)
  - Protein amount
- Table of chemicals (metabolites) and their abundance throughout experiment
Data integration with plotting

For one microbe, *Hungateclostridium thermocellum*

Ploting **two functional categories**, here enzymes: glycoside hydrolase (GH; red) and glycosyl transferase (GT; blue)
Data integration with ViMO: Visualizer for Meta-Omics

Uses the previous tables to integrate and visualize the data automatically.
Explore in ViMO before showing last slide 😊
Connecting the dots...

- Metabolic reconstruction of all MAGs
  - Which enzymes and pathways do they have?
  - What can each microbe do?

- Add a functional layer (metaT / metaP)
  - Which enzymes and pathways are expressed and in what abundance?
  - What are they actually doing?

- Using the known biology of the system + meta-metabolomics
  - Which MAGs are active when?
  - Infer carbon-flow and microbial interactions

Hydrogenotrophic methanogen METH1

Saccharolytic bacteria RCLO1 and CLOS1

Sugar fermenter TISS1

Generalist bacteria Coprothermobacter proteolyticus COPR1

Syntrophic acetate-oxidizing bacteria TEPI1 and TEPI2

Andersen et al., Methods (2020)
OVERVIEW OF INFORMATICS PLATFORMS AVAILABLE FOR FUNCTIONAL MICROBIOME ANALYSIS

Avignon
Jun.-Prof. Robert Heyer (ISAS Dortmund/ Bielefeld University)
24.04.2023
• **Complexity and heterogeneity:**
  Microbiomes may contain hundreds of different species and functions

• **Lack of species information:**
  Only a small proportion microbial species is isolated, sequences and studied

• **Huge data volumes:**
  1x measurements $\rightarrow$ 20,000-200,000 MS/MS spectra
  $\rightarrow$ 1,000-5,000 protein groups
  $\rightarrow$ 1.4 Tb data per week
Recap Experimental WOrkflow

Introduction/Recap

Workflows

Other tools

Knowledge repositories

Summary

Sample

A. Protein extraction and purification
   - Cell lysis
   - Centrifugation
   - Protein extraction
   - Precipitation
   - Protein quantification

B. Tryptical digestion

C. Separation
   - Gel-based: SDS-PAGE or 2D-PAGE of proteins
   - Gel-free: 1D or 2D HPLC of peptides coupled directly to MS/MS

D. LC-MS/MS
   - Chromatogram
   - MS-spectra
   - MS/MS-spectra
Recap Bioinformatic Workflow

Introduction/Recap
- Skyline/Panorama

Workflows
- MS-Spectra
- Database search algorithms
  - Xtandem
  - OMMSA
  - MS-Amanda
  - Mascot
  - BLAZMASS

- Protein list
- Taxonomical and functional evaluation
  - Protein grouping
  - UniPept
  - Prophane
  - Pipasic
  - Megan

- Data visualization
  - Krona plots
  - KEGG pathways
  - Reactome

- Long term storage
  - PRIDE

Other tools
- Knowledge repositories
- Summary
- Software platforms
  - MetaProteomeAnalyzer
  - Galaxy-P
  - MetaPro-IQ
Software concepts

- RegEx
- Command line tools (e.g., BLAST, Diamond)
- Notebooks (e.g., Jupyter)
- Container (e.g., Bioconda)
- Workflows (e.g., KNIME, nextflow)
- Graphical user interface
- Server-client architecture
- Webtools
- Repositories (e.g., GIT)
• Data Management Planning
• Creating data
• Process data
• Analyze / use data
• Preserving / store data
• Publish / share data
• Reuse data
About iMetalab

MetaLab was originally designed for human/mouse gut metaproteomics protein identification and quantification, rooted from the MetaPro-IQ workflow. Upon the rising request, we wrapped up the workflow into a desktop standalone version, with taxon and function analysis module built-in, and share the tool to the community for free.

iMetalab represents the whole framework, including the Metalab Desktop version, automatic iMetaReport, iMetaShiny apps for data visualization and analysis. We aim to make iMetaLab platform a free and one-stop toolset for metaproteomics, with increasing amount of tools under active development. See publications about MetaLab.

EVOLUTION OF METALAB SUITE
Metalab

1. Database construction
   - Raw files
   - Spectra cluster
   - Iterative search
   - Peptide list
   - Specific database

2. Peptide identification & quantification
   - Specific database
   - Peptide list

3. Taxonomy analysis
   - Peptide list
   - Built-in database
   - Taxonomy information
   - Unipept
MaxQuant

MaxQuant is a quantitative proteomics software package designed for analyzing large mass-spectrometric data sets. It is specifically aimed at high-resolution MS data. Several labeling techniques as well as label-free quantification are supported. MaxQuant is freely available and can be downloaded from this site. The download includes the search engine andromeda, which is integrated into MaxQuant as well as the viewer application for inspection of raw data and identification and quantification results. For statistical analysis of MaxQuant output, we offer the Perseus framework.
MetaproteomeAnalyzer

Introduction/Recap | Workflows | Other tools | Knowledge repositories | Summary

![Browser screen with MetaproteomeAnalyzer interface]


The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation

Thilo Muth,1 Alexander Behne, Robert Heyer, Fabian Kohrs, Dirk Benndorf, Marcus Hoffmann, Miro Lehteva, Udo Reichl, Lennart Martens, Erdmann Rapp
• MPA Cloud release in summer 2023
Deepsival

- **Introduction/Recap**
- **Workflows**
- **Other tools**
- **Knowledge repositories**
- **Summary**

### Training
- MS2 Spectra
- Labeled PSMs
- Peptide sequence
- Class label True/False

### Prediction
- PSMs
- MS2 Spectra
- Peptide sequence
- Class label True/False

### DeePSIVal Workflow
- **DeePSIVal Training**
- **DeePSIVal MODEL**

### Summary
- **CNN**
- 12 ion types
- 30 AA residues
- 10 bins per peak

### Accuracy Table

<table>
<thead>
<tr>
<th></th>
<th>Training vs Test</th>
<th>Ecoli</th>
<th>Hela</th>
<th>Proteinmix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ecoli</td>
<td>0.973</td>
<td>0.960</td>
<td>0.953</td>
<td></td>
</tr>
<tr>
<td>Hela</td>
<td>0.958</td>
<td>0.975</td>
<td>0.956</td>
<td></td>
</tr>
<tr>
<td>Proteinmix</td>
<td>0.939</td>
<td>0.947</td>
<td>0.984</td>
<td></td>
</tr>
</tbody>
</table>
## Protein grouping

<table>
<thead>
<tr>
<th>Rule</th>
<th>Principle</th>
<th>Explanation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein rule</td>
<td>1. UniRef-Cluster</td>
<td>Grouping of proteins when they have 50%, 90% or 100% sequence similarity. Protein clustering provided by UniRef Cluster [Suzek2007].</td>
<td>Lu et al. (2014), Suzek et al. (2007)</td>
</tr>
<tr>
<td></td>
<td>2. KEGG Ontologies</td>
<td>Grouping of proteins when they are similar to functional classified genes within KEGG Ontology [Mai 2005]. KEGG Ontologies are provided by UniProtKB databases [JAPI PAPER].</td>
<td>Gotelli et al. (2012), Kanehisa et al. (2016)</td>
</tr>
<tr>
<td>Peptide rule</td>
<td>1. Shared peptide set</td>
<td>Group proteins when they share the same peptides.</td>
<td>Keiblinger et al. (2012), Kolmедер et al. (2012), Schneider et al. (2011) Kohrs et al. (2014), Lu et al. (2014)</td>
</tr>
<tr>
<td></td>
<td>2. One shared peptide</td>
<td>Group proteins when they have one identified peptide in common</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. One shared peptide + Levenshtein, distance &lt; 2</td>
<td>Group proteins when they share the same peptides, but not if they have two similar peptides with less than 2 point mutations differences. This tracks the production of one protein by different microorganisms.</td>
<td></td>
</tr>
<tr>
<td>Taxonomy rule</td>
<td>1. Phylogenetic affiliation</td>
<td>Extends other rules by a certain phylogenetic affiliation.</td>
<td>Muth et al. (2015a)</td>
</tr>
</tbody>
</table>
Pout2Prot

Introduction/Recap

Workflows

Other tools

Knowledge repositories

Summary

Protein groups

Occam's Razor

- Group 1
  - Subgroup 1.1: Protein 1
    - 3 PSMs
  - Subgroup 2: Protein 4
    - 3.5 PSMs
- Subgroup 2.2: Protein 5
  - 2.5 PSMs

Protein assignment

- Peptide A
  - Protein 1
  - Peptide C
- Peptide A: Protein 2
- Peptide A: Protein 3
- Peptide B: Protein 4
  - Peptide D
- Peptide B: Protein 5
  - Peptide E

Protein groups

Anti-Occam's Razor

- Group 1
  - Subgroup 1.1: Protein 1
    - 2 PSMs
  - Subgroup 1.2: Protein 2
    - 2 PSMs
  - Subgroup 1.3: Protein 4
    - 3 PSMs
  - Subgroup 1.4: Protein 5

Related article:

Journal of Proteome Research

Pout2Prot: An Efficient Tool to Create Protein (Sub)groups from Percolator Output Files

Kay Schallert, Pieter Verschaeffelt, Bart Mesuere, Dirk Benndorf, Lennart Martens,* and Tim Van Den Bossche

pubs.acs.org/jpr
**MPA Pathway Tool**

Introduction/Recap  Workflows  Other tools  Knowledge repositories  Summary

![Diagram of metabolic pathways](image)

\[ \begin{align*}
R1 & \quad R2 & \quad R3 & \quad R4 & \quad R5 \\
-1 & \quad 0 & \quad 0 & \quad 0 & \quad 0 & \rightarrow S \\
1 & \quad -1 & \quad -1 & \quad 0 & \quad -1 & \rightarrow A \\
0 & \quad 1 & \quad 1 & \quad -1 & \quad 0 & \rightarrow B \\
0 & \quad 0 & \quad 0 & \quad 1 & \quad 0 & \rightarrow P1 \\
0 & \quad 0 & \quad 0 & \quad 0 & \quad 1 & \rightarrow P2
\end{align*} \]
**MPA Pathway Tool**

- **Introduction/Recap**
- **Workflows**
- **Other tools**
- **Knowledge repositories**
- **Summary**

**Flow rate in mmol/(gDW)$^{-1}$**

**Growth rate in 1/h**

- 0.12
- 0.2
- 0.35
- 0.5
- FBA Fluss

---

**The quantitative and condition-dependent *Escherichia coli* proteome**

Alexandet Schmidt, Karl Kochanowski, Silke Vedelaar, Erik Ahmä, Benjamin Volkmer, Luciano Callipo, Kévin Knoops, Manuel Bauer, Ruedi Aebersold & Matthias Heinemann

*Nature Biotechnology* 34, 104–110 (2016) | [Cite this article](#)
MPA Pathway Tool

Introduction/Recap

Workflows

Other tools

Knowledge repositories

Summary

The quantitative and condition-dependent *Escherichia coli* proteome

**Published:** 01 January 2016

Alexander Schmidt, Karl Kochanowski, Silke Vedelaar, Erik Ahmén, Benjamin Volkmer, Luciano Callipo, Kevin Knoop, Manuel Bauer, Ruedi Aebersold, Matthias Heinemann

*Nature Biotechnology* 34, 104–110 (2016) | Cite this article

22k Accesses | 370 Citations | 67 Altmetric | Metrics
AutoPACMEN

Selected growth scenarios → List of $k_{cat}$ values for calibration → Model Calibrator (MATLAB/Python) → calibrated $k_{cat}$ values and enzyme pool $P$ → SBML of original model → Model Generator (Python) → SBML of sMOMENT model → Analysis

UniProt
SABIO-RK
BRENDA
Optional custom $k_{cat}$ database
Optional enzyme concentration data

Automatic construction of metabolic models with enzyme constraints

Pavlos Stephanos Bekiaris 1, Steffen Klamt 2

Affiliations: + expand

PMID: 31937255  PMCID: PMC6961255  DOI: 10.1186/s12859-019-3329-9

Free PMC article
**MPA Pathway Tool**

**Introduction/Recap**

**Workflows**

**Other tools**

**Knowledge repositories**

**Summary**

### Model % of total variability

<table>
<thead>
<tr>
<th>Model</th>
<th>% of total variability</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em> core</td>
<td>100.00%</td>
</tr>
<tr>
<td><em>E. coli</em> core kcat</td>
<td>33.33%</td>
</tr>
<tr>
<td><em>E. coli</em> core proteins</td>
<td>17.50%</td>
</tr>
</tbody>
</table>
MPA Pathway Tool

Introduction/Recap  Workflows  Other tools  Knowledge repositories  Summary

MPA Pathway Tool

PATHWAY-CREATOR

FLUX-ANALYSIS

PATHWAY-CALCULATOR

Upload
Help
node configurations
Add Kegg Reaction
Add User-defined Reaction
import multiple reactions
Download
Network Anotation
Welcome

Unipept is an open source web application developed at Ghent University that is designed for metaproteomics data analysis with a focus on interactive datavisualizations. Unipept is powered by an index containing all UniProt entries, a tweaked version of the NCBI taxonomy and a custom lowest common ancestor algorithm. This combination enables a blazingly fast biodiversity analysis of large and complex metaproteome samples. This functionality is also available via an API and a set of command line tools. Next to these core functions, Unipept also has a tool for selecting unique peptides for targeted proteomics and for comparing genomes based on peptide similarity.

With tryptic peptide analysis, you can submit a single tryptic peptide that can be 5 to 80 residues long. The application will respond with a list of all UniProt entries wherein the peptide was found along with a complete taxonomic lineage derived from the NCBI taxonomy.

Metaproteomics analysis helps you analyze lists of tryptic peptides, e.g. extracted from an environmental sample using shotgun tandem mass spectrometric methods. Of these peptides, the lowest common ancestors (LCA) will be calculated. These LCAs will be...
KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. See Release notes (April 1, 2023) for new and updated features.

Main entry point to the KEGG web service
- KEGG2
- KEGG Table of Contents
- [Update notes | Release history]

Data-oriented entry points
- KEGG PATHWAY
- KEGG BRITE
- KEGG MODULE
- KEGG ORTHOLOGY
- KEGG GENES
- KEGG GENOME
- KEGG COMPOUND
- KEGG GLYCAN
- KEGG REACTION
- KEGG ENZYME
- KEGG NETWORK
- KEGG PATHWAY maps
- BRITE hierarchies and tables
- KEGG modules
- KO functional orthologs
- Genes and proteins
- Genomes
- Small molecules
- Glycans
- Biochemical reactions
- Enzyme nomenclature
- Disease-related network variations
<table>
<thead>
<tr>
<th>Issue</th>
<th>Name/principle</th>
<th>Explanation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Taxonomic classification</strong></td>
<td>1. Lowest common ancestor</td>
<td>Define taxonomy as the lowest common ancestor into the phylogenetic tree.</td>
<td>Huson et al. (2011), Jagtap et al. (2012)</td>
</tr>
<tr>
<td></td>
<td>2. Weighted lowest common ancestor</td>
<td>Adjust the lowest common ancestor by unique identification for the single taxon.</td>
<td>Huson et al. (2016)</td>
</tr>
<tr>
<td></td>
<td>3. Peptide similarity estimation and expression level weighting</td>
<td>Weight taxonomy of identified peptides by their spectra abundance and their occurrence in a reference proteome.</td>
<td>Peralta et al. (2014)</td>
</tr>
<tr>
<td></td>
<td>4. Unique peptides</td>
<td>Define taxonomy and taxonomy profiles only based on unique peptides.</td>
<td>Rooijers et al. (2011), Karlsson et al. (2012)</td>
</tr>
<tr>
<td><strong>Functional classification</strong></td>
<td>1. KEGG Orthologs (KO)</td>
<td>Grouping of genes with same function by sequence similarity.</td>
<td>Kanehisa et al. (2016)</td>
</tr>
<tr>
<td></td>
<td>2. Cluster of orthologous genes (COG)</td>
<td>Grouping of genes with same function by sequence similarity.</td>
<td>Tatusov et al. (2000)</td>
</tr>
<tr>
<td></td>
<td>4. PFAM</td>
<td>Database of conserved functional units, represented by a set of aligned sequences with their probabilistic representation (hidden Markov model).</td>
<td>Finn et al. (2016)</td>
</tr>
<tr>
<td></td>
<td>5. TIGRFAM</td>
<td>Database of conserved functional units, represented by a set of aligned sequences with their probabilistic representation (hidden Markov model).</td>
<td>Hafit et al. (2013)</td>
</tr>
<tr>
<td></td>
<td>7. InterPro</td>
<td>Functional analyses of protein sequences by classifying them into families and predicting the presence of domains and important sites. Signatures are provided by 14 different member databases (among others PFAM, TIGRFAM, SMART).</td>
<td>Finn et al. (2017)</td>
</tr>
<tr>
<td><strong>Pathway mapping</strong></td>
<td>1. MetaCyc</td>
<td>Curated database of experimentally confirmed metabolic pathways.</td>
<td>Cerq et al. (2016)</td>
</tr>
<tr>
<td></td>
<td>3. Reactome</td>
<td>Pathway database.</td>
<td></td>
</tr>
<tr>
<td><strong>Calculation of sequence similarity</strong></td>
<td>1. BLAST</td>
<td>Calculation of sequence similarities.</td>
<td>Altschul et al. (1990)</td>
</tr>
<tr>
<td></td>
<td>2. DIAMOND</td>
<td>Calculation of sequence similarities. Up to 20,000 faster as BLAST.</td>
<td>Buchfink et al. (2015)</td>
</tr>
<tr>
<td></td>
<td>3. MS-BLAST</td>
<td>Calculation of sequence similarities optimized for peptides.</td>
<td>Shevchenko et al. (2001)</td>
</tr>
</tbody>
</table>
Summary

• Start point: reproduce known dataset

• Document applied steps

• Plan data and tool integration before experiments

• Care about data size
THANK YOU