Module XIV – Functional assignment - What are they doing?

ANAEROBICS

Overview of this talk

What is functional assignment ?

Experimental setup and Assembly - What is important ?

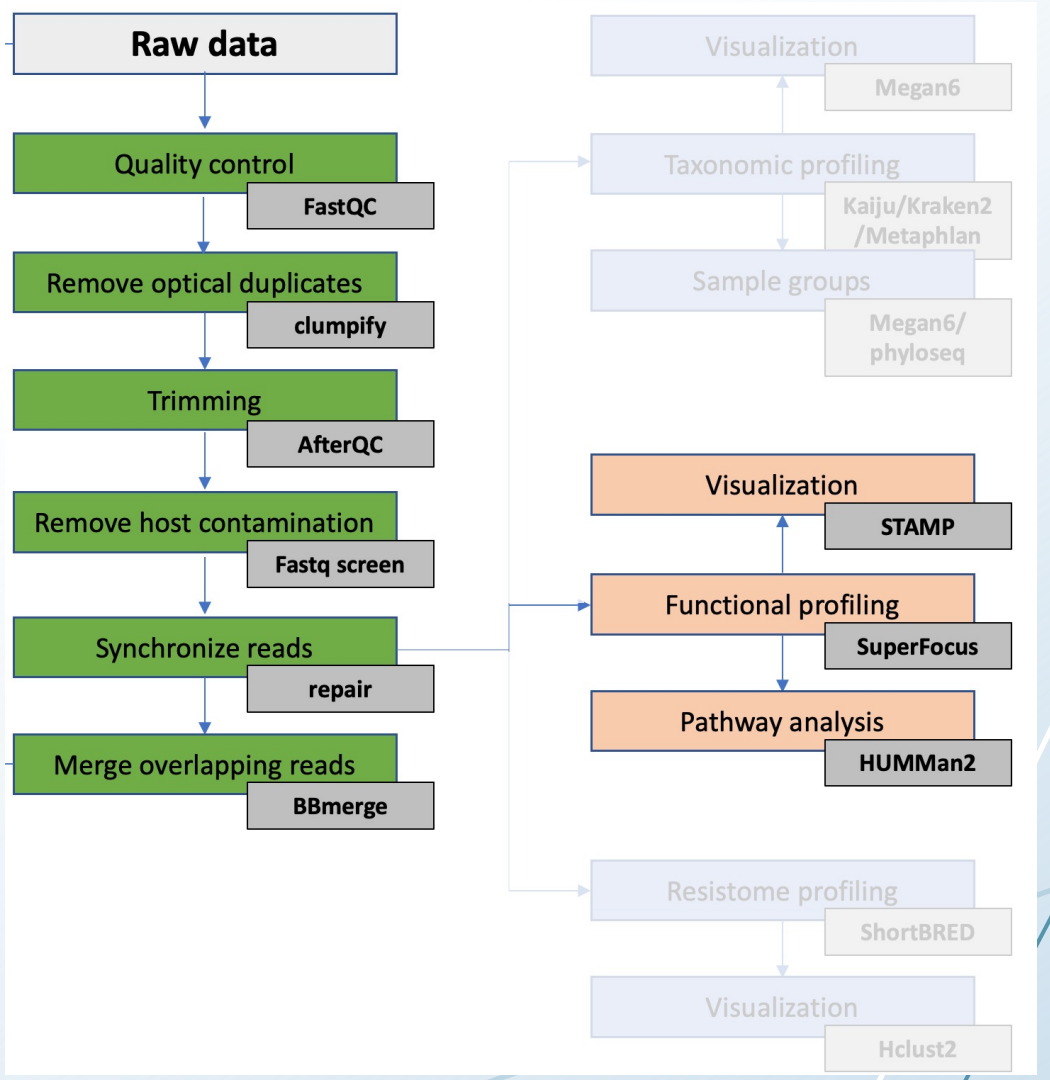
Overview of Functional assignment / Databases

Functional profiling of assembled sequences

Functional profiling of sequence reads

Visualization and comparison of functional profiles

Pathway profiling



The humane microbiome – the genetic repertoire changes

We focus a lot on which microbes are present (or absent)

DNA exchange occurs between microbes, and can potentially alter their functional capabilities

nature International weekly journal of science

Nature **464**, 908-912 (8 April 2010) | doi:10.1038/nature08937; Received 9 November 2009; Accepted 19 February 2010

Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota

Jan-Hendrik Hehemann[1,2,3], Gaëlle Correc[1,2], Tristan Barbeyron[1,2], William Helbert[1,2], Mirjam Czjzek[1,2] & Gurvan Michel[1,2]

Functional assignment - What are they doing?

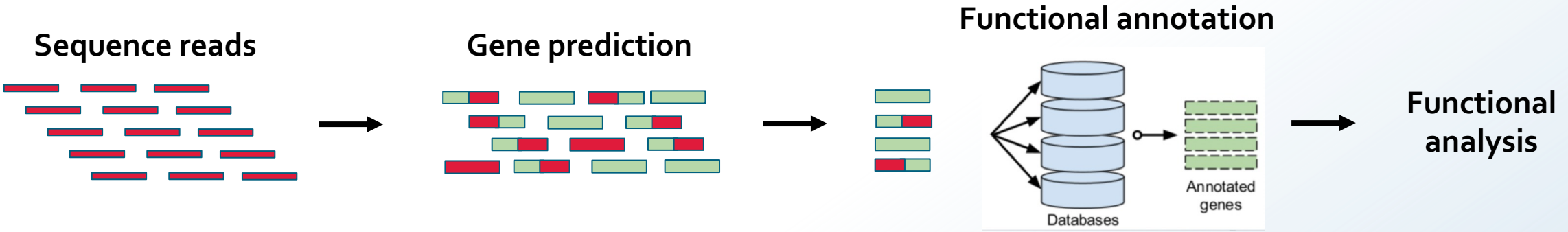
Annotation of all **available** genes

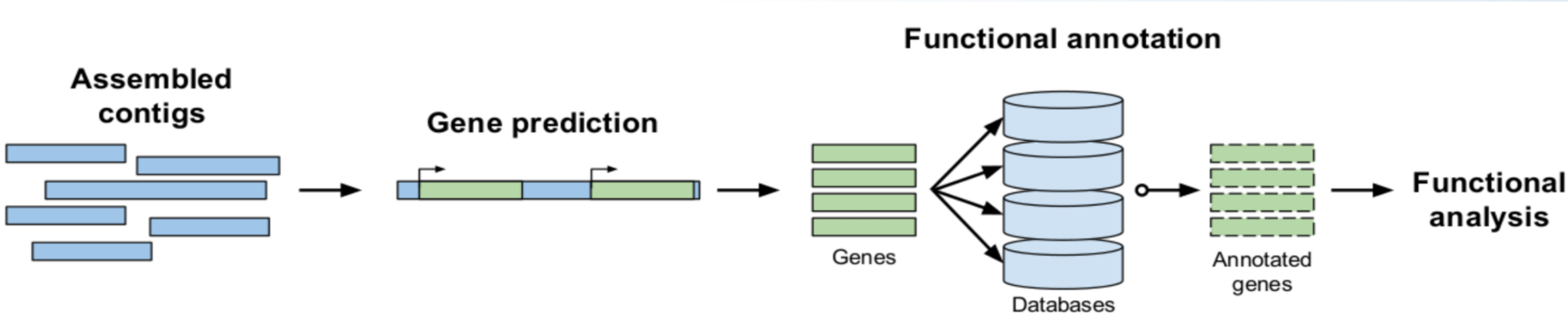
- We can only see part of the picture

Provides functional overview and novel genes

Functional assignment of genes – Two approaches

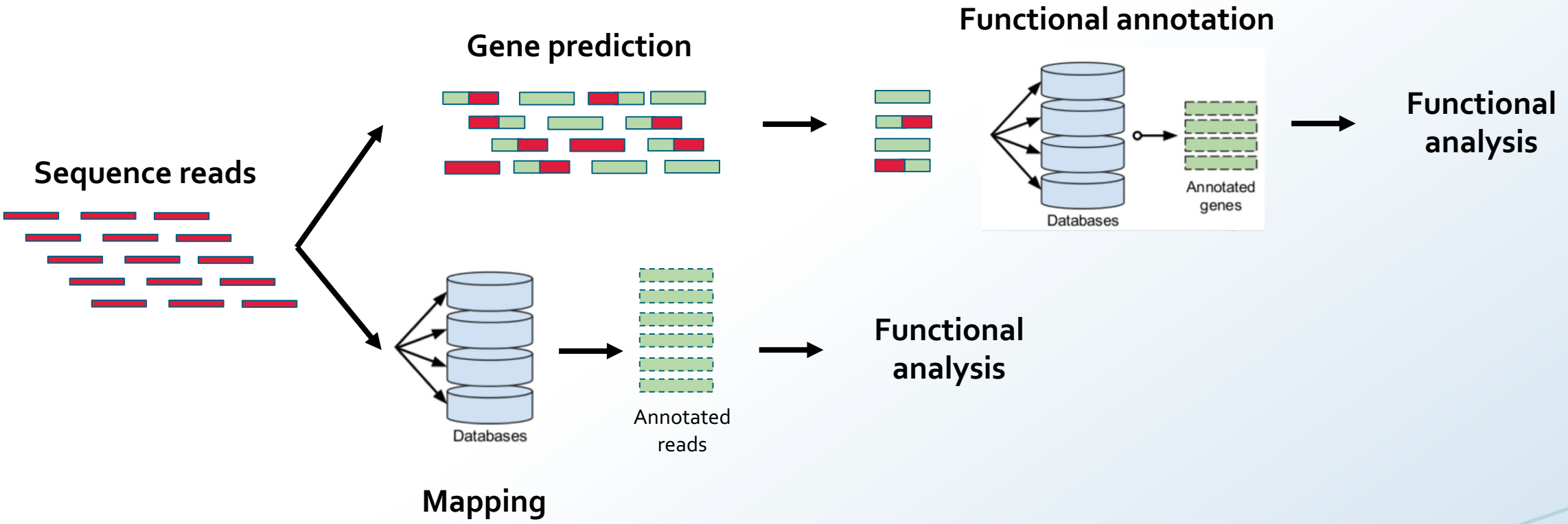
Genes are predicted from either 1) raw reads or 2) assembled contigs and annotated using a tool and a database



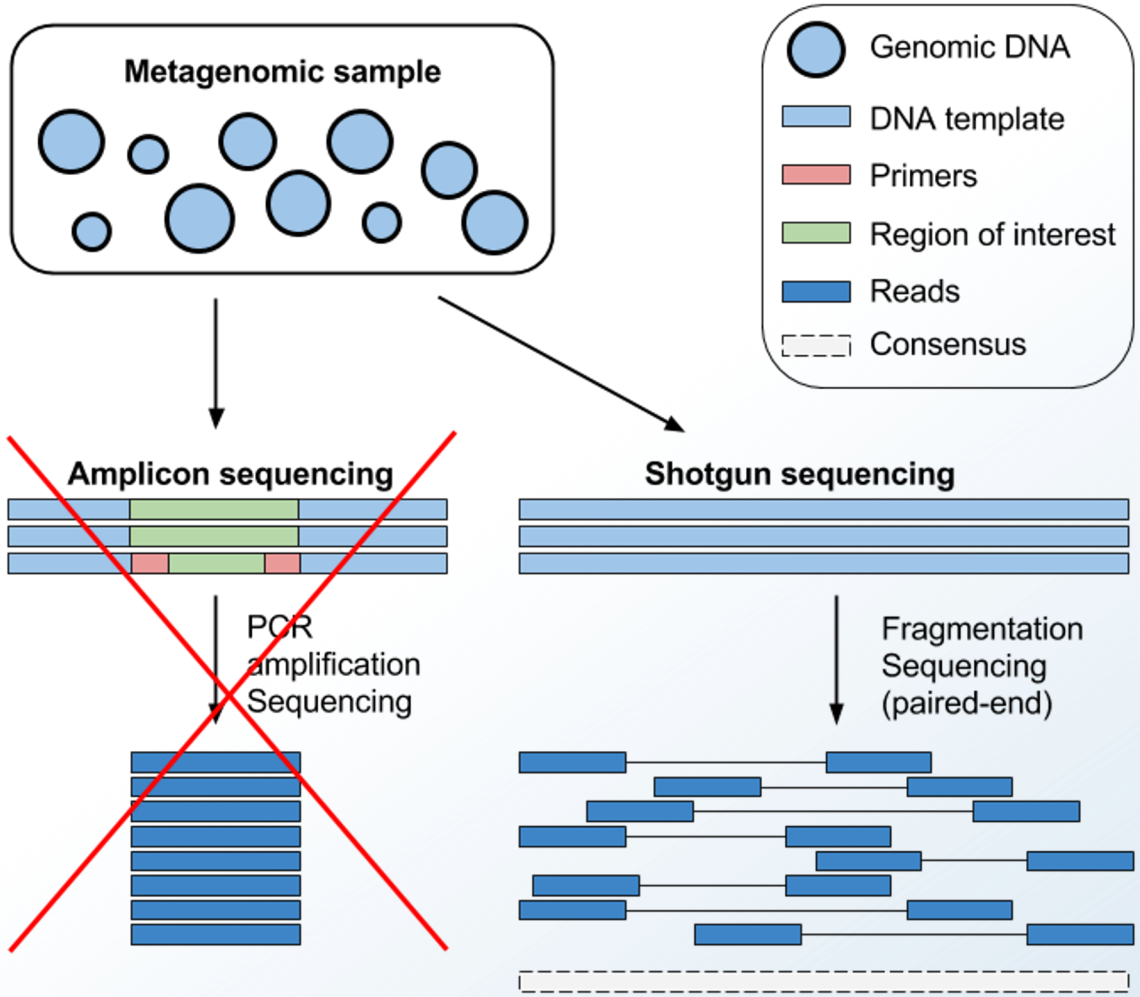


Functional assignment of reads – Two approaches

Functional assignment can be done either on predicted genes or by mapping reads against a database



Experimental setup - Just to avoid any confusion…



Experimental setup - What is important?

Consider diversity, how much coverage do you need?

Consider sequencing technology, coverage, cost, multiplexing…

What is the question you want to answer?

Try to keep financial considerations out!

Assembly versus no assembly - Pros and cons

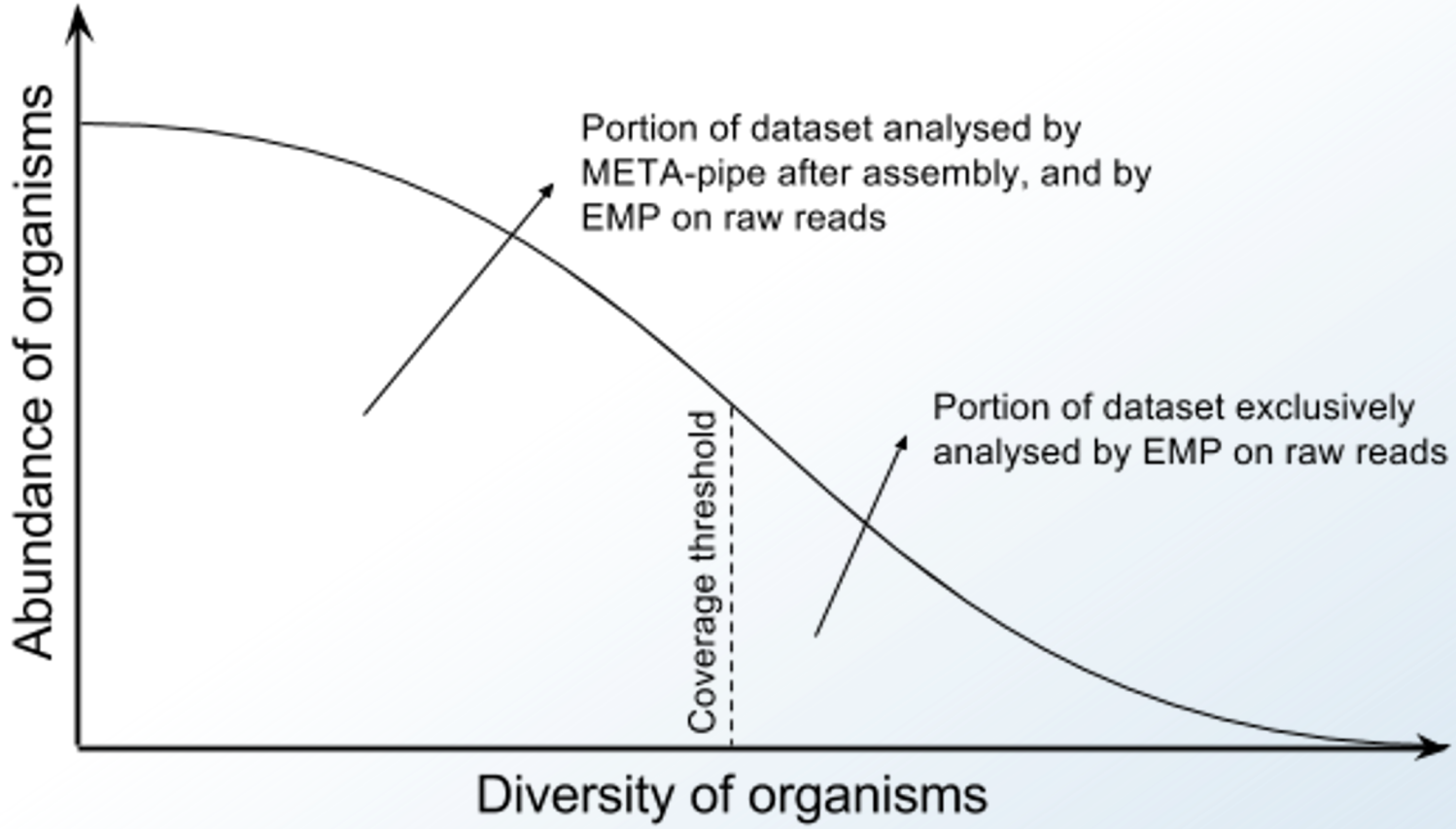
Functional annotation of full-length genes give more accurate assignments

Loose or complex to track abundance

Functional annotation assembled reads	Functional annotation sequence reads
Full-length genes give more accurate assignments	Gene fragments give more false positive matches
Assembly will remove abundance information	Counting reads will keep abundance measures

Assembly versus no assembly

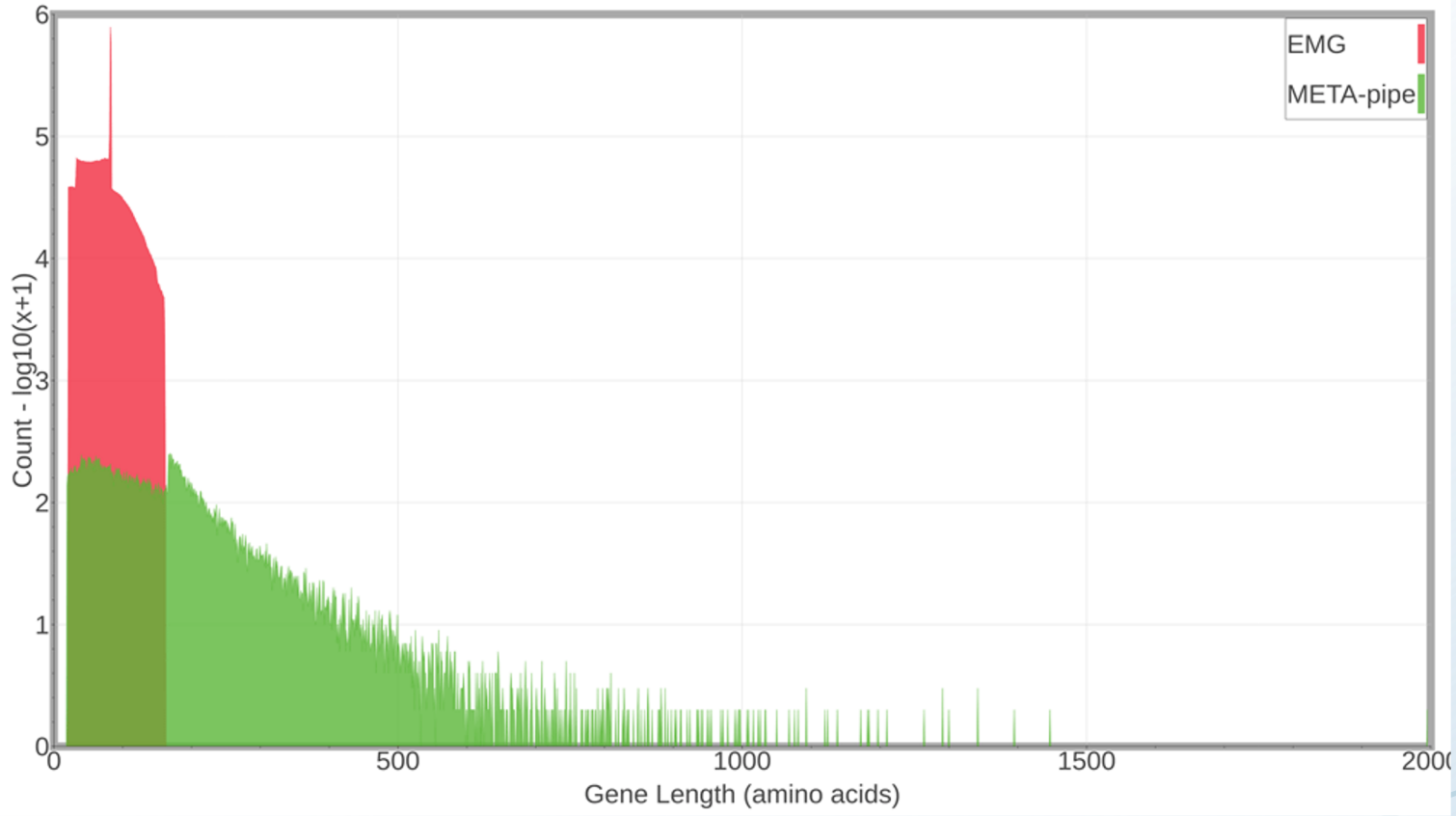
Functional comparison of assembled contigs against raw reads are problematic as some data is effectively discarded.



Functional comparison between EBI Metagenomics and Meta-pipe

Assembly VS. no assembly

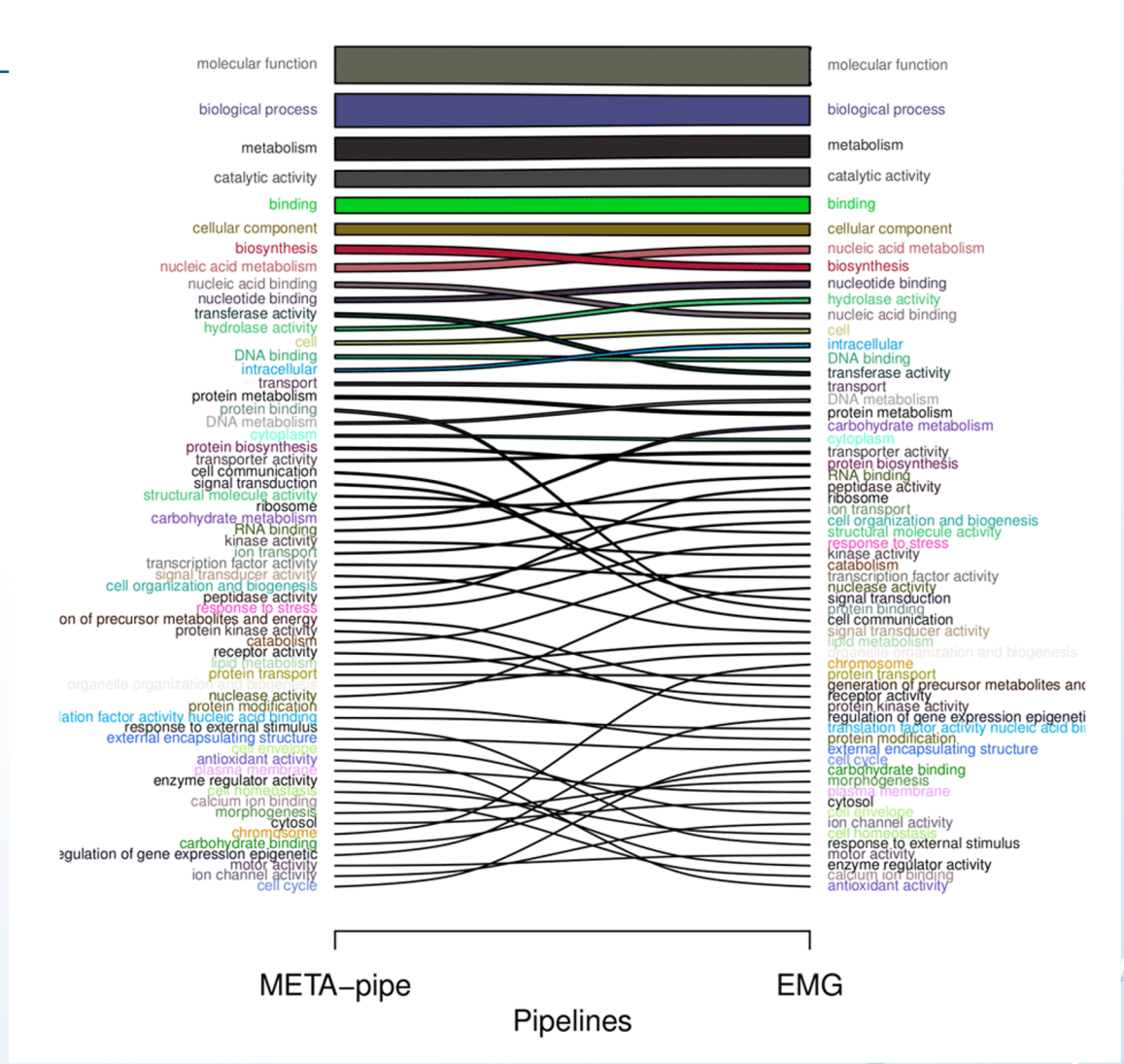
50.000 full length genes VS. 11.500.000 mostly fragmented genes



Functional comparison between EBI Metagenomics and Meta-pipe

Sorted counts of GO-terms to look at functional profiles

Huge differences in niche Gene Ontologies

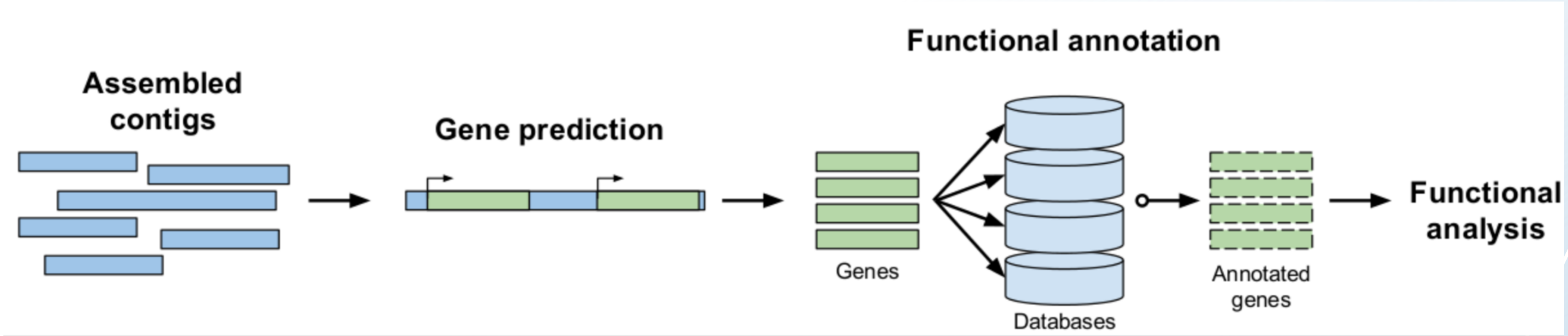


Robertsen EM, Denise H, Mitchell A et al. 2017 F1000Research 2017, 6:70 (doi: 10.12688/f1000research.10443.1)

Functional assignment

Annotation of all genes

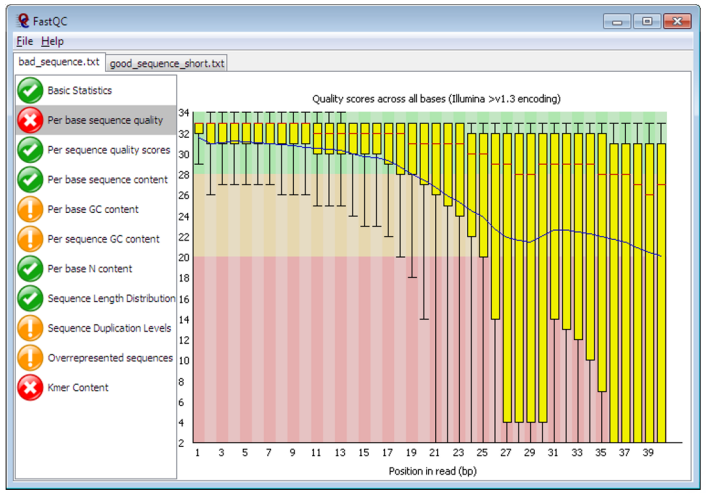
Provides functional overview and novel genes

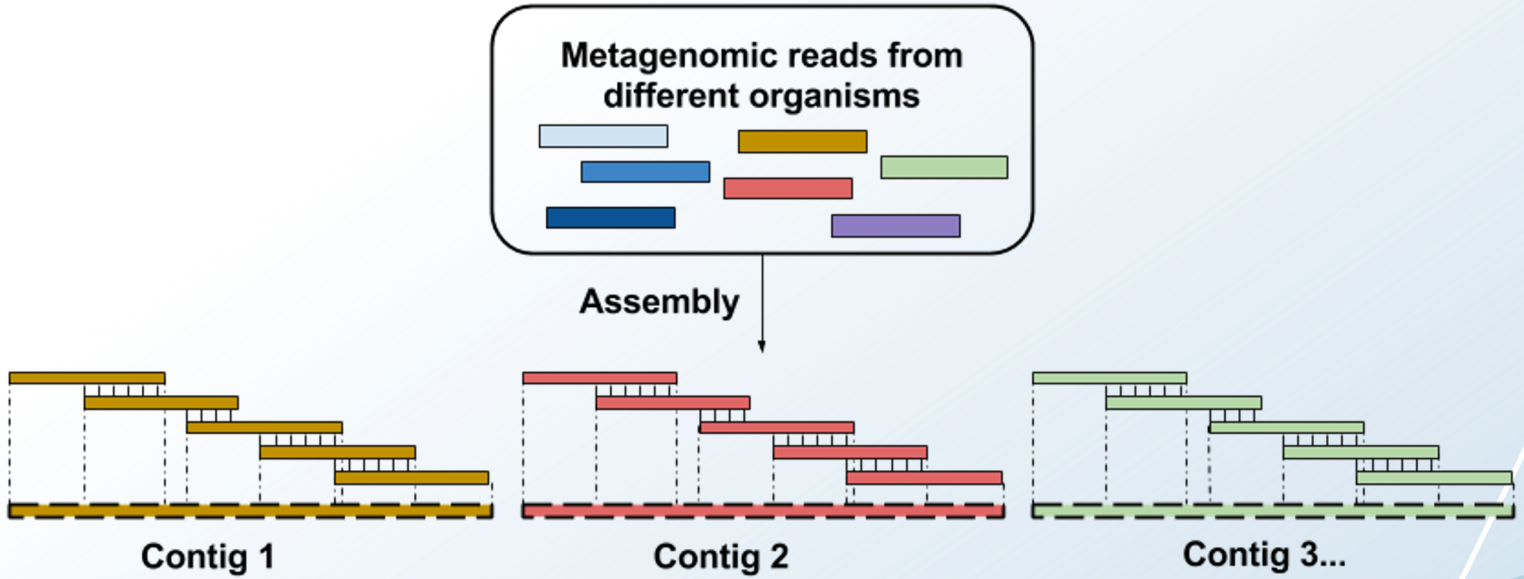


Pre-processing: QC and Assembly of metagenomic data

QC and filtering improves quality of assembly

Metagenomic assembly is challenging





Metagenomic gene prediction from metagenomic contigs or reads

Myriad of software available

Typically quick to run

Metagenomic contigs adds a layer of extra complexity (spurious contigs, codon usage, etc.)

Functional assignment tools

Used to query databases with predicted genes

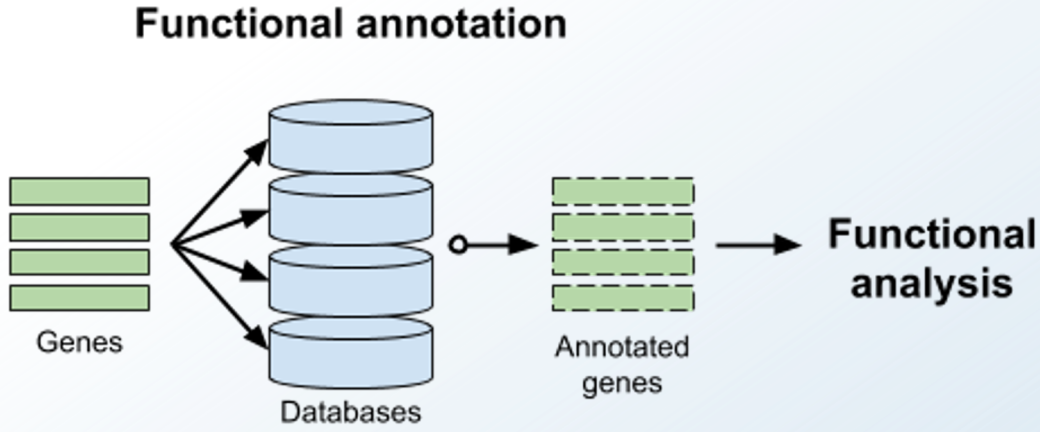
Most commonly tools used:

HMMer - Search sequence databases with Hidden Markov Models

BLAST - Basic Local Alignment Search Tool

Typical runtime for 30GB raw input (1 GB assembled):

1-2 days (functional annotation on 400 cores, supercomputer)



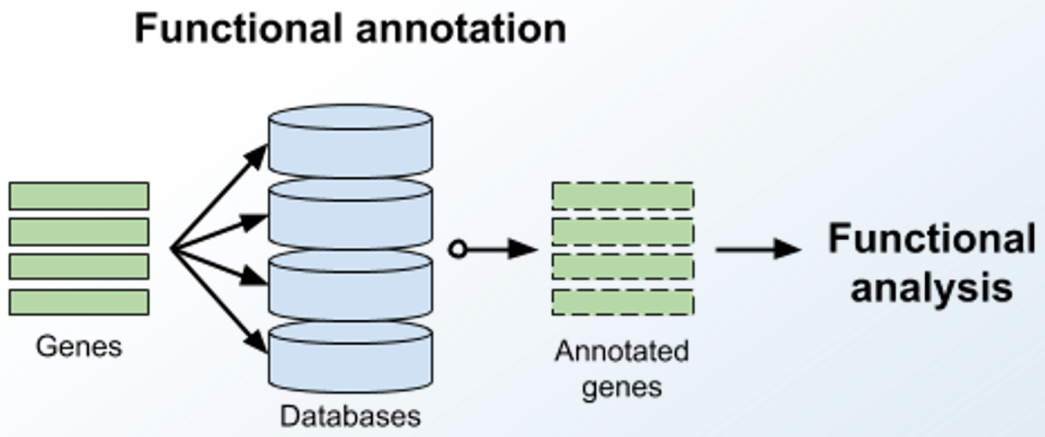
Functional assignment databases

Tons of databases available; some more generic, some specialized

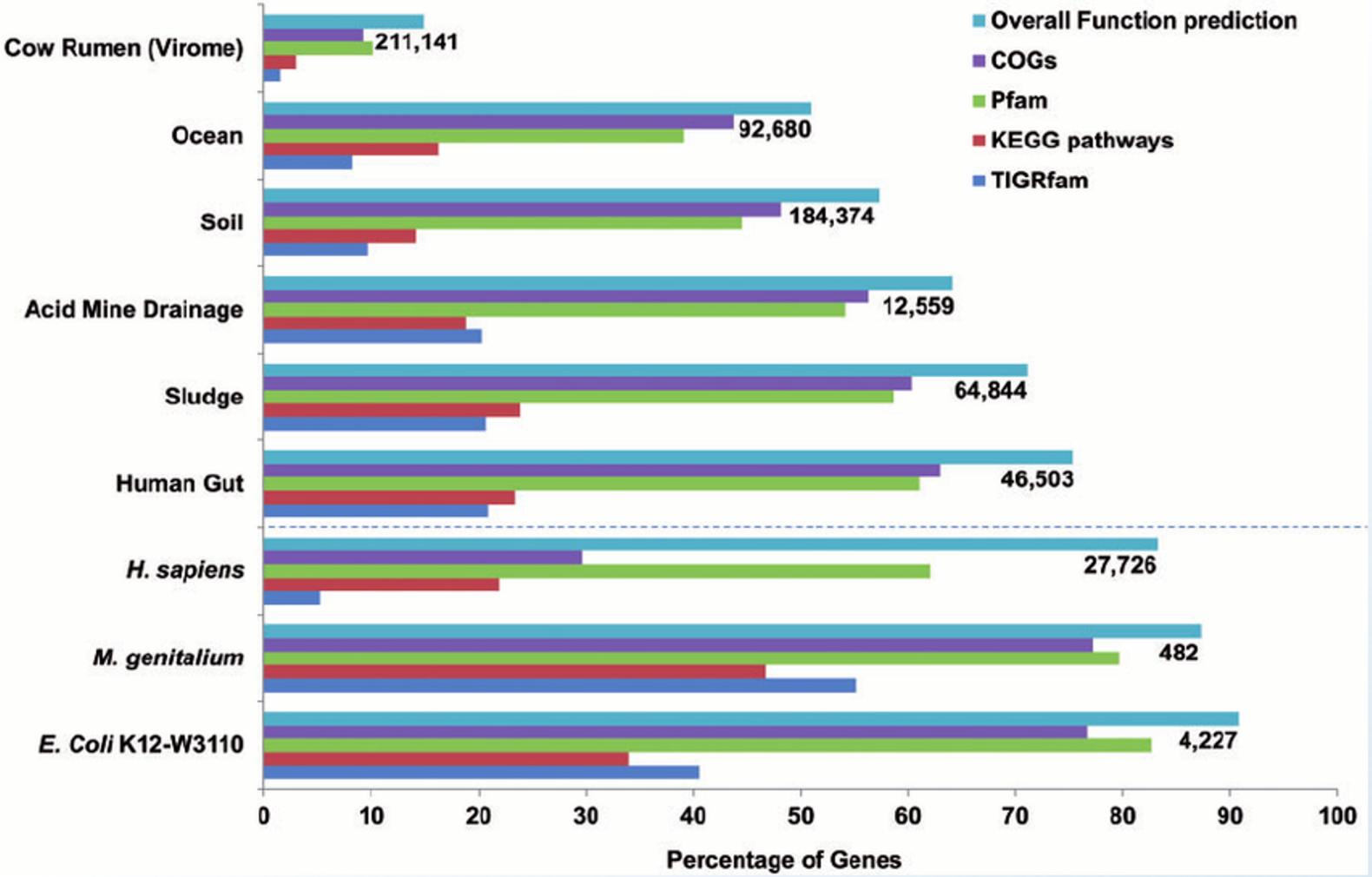
"Defines" the workload in metagenomic processing

Endless amounts of databases

Uniprot, NCBI, COG, KEGG



Functional assignment databases



Prakash, Tulika & Taylor, Todd. (2012). Functional assignment of metagenomic data: Challenges and applications. Briefings in bioinformatics. 13. . 10.1093/bib/bbs033.

Challenges in functional assignment

Databases are lacking reference data to a varying degree

Requires extensive hardware resources

Backed by cluster computer

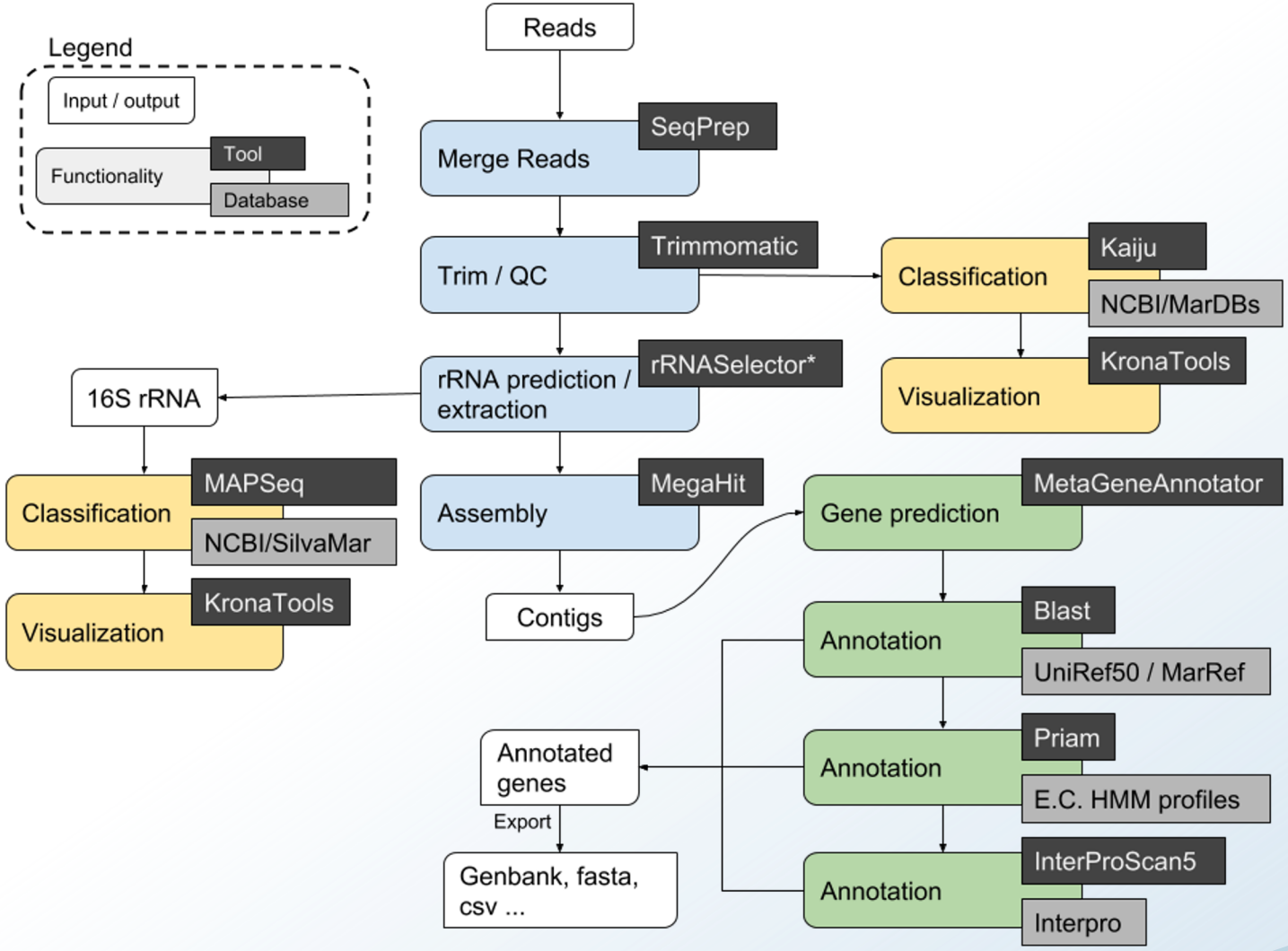
Typical runtime for 30GB raw input:

5-6 hours (preprocessing)

1-2 days (functional annotation)

Not feasible on a laptop!

Functional analysis with Meta-pipe



Brief overview of available tools and databases

MetaGeneAnnotator – Prediction of genes

PRIAM – Annotation of EC-numbers with RPS-blast

Uniref50 – Clustered UniProt using Blast

InterPro – Collection of 14 databases queried with InterProScan

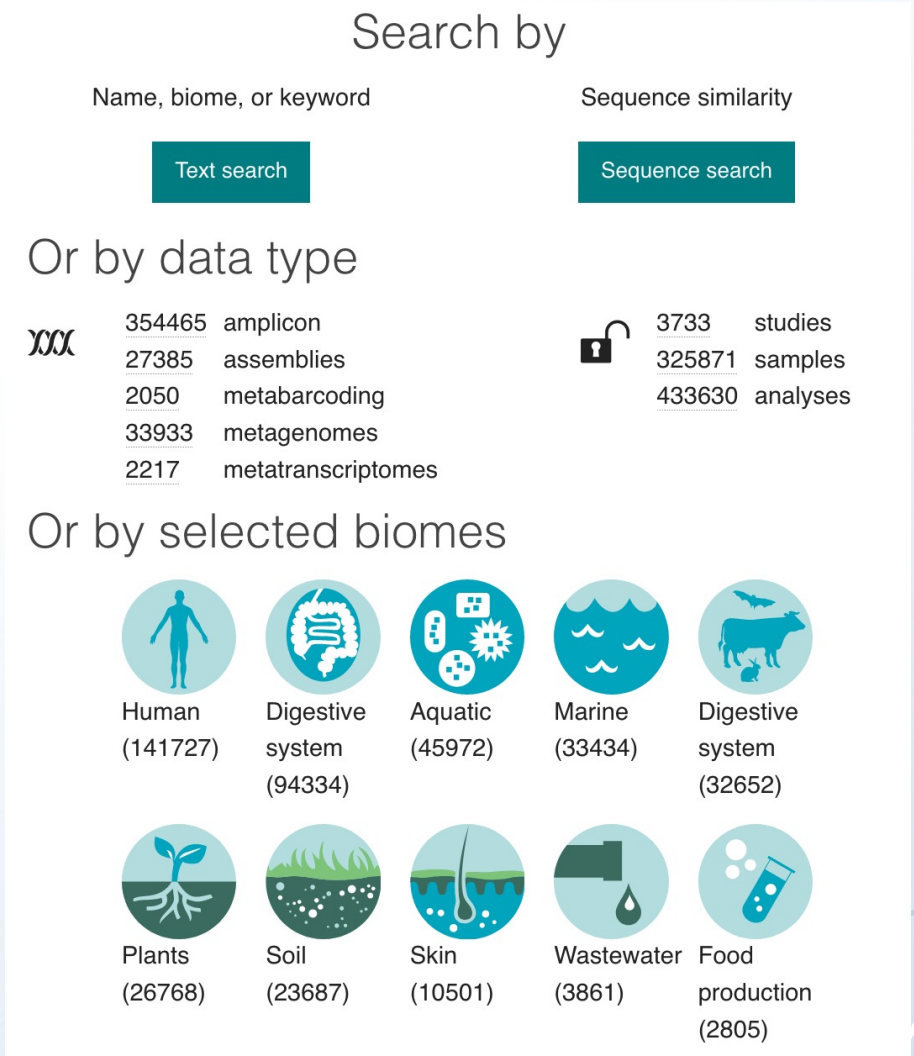
MarDBs – Marine specific annotation

EBI Metagenomics - MGnify

MGnify
Submit, analyse, discover and compare microbiome data

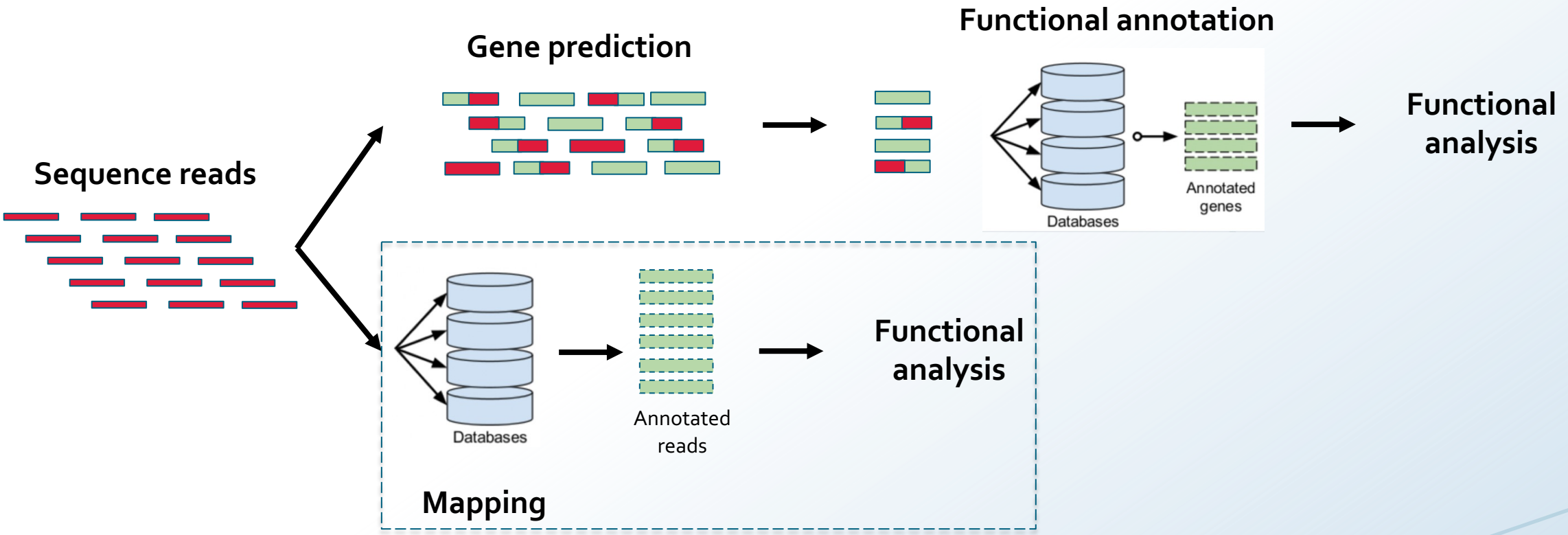
Automated pipeline for the analysis and archiving of microbiome data

Users can submit their own data for analysis or freely browse all of the analysed public datasets held within the repository



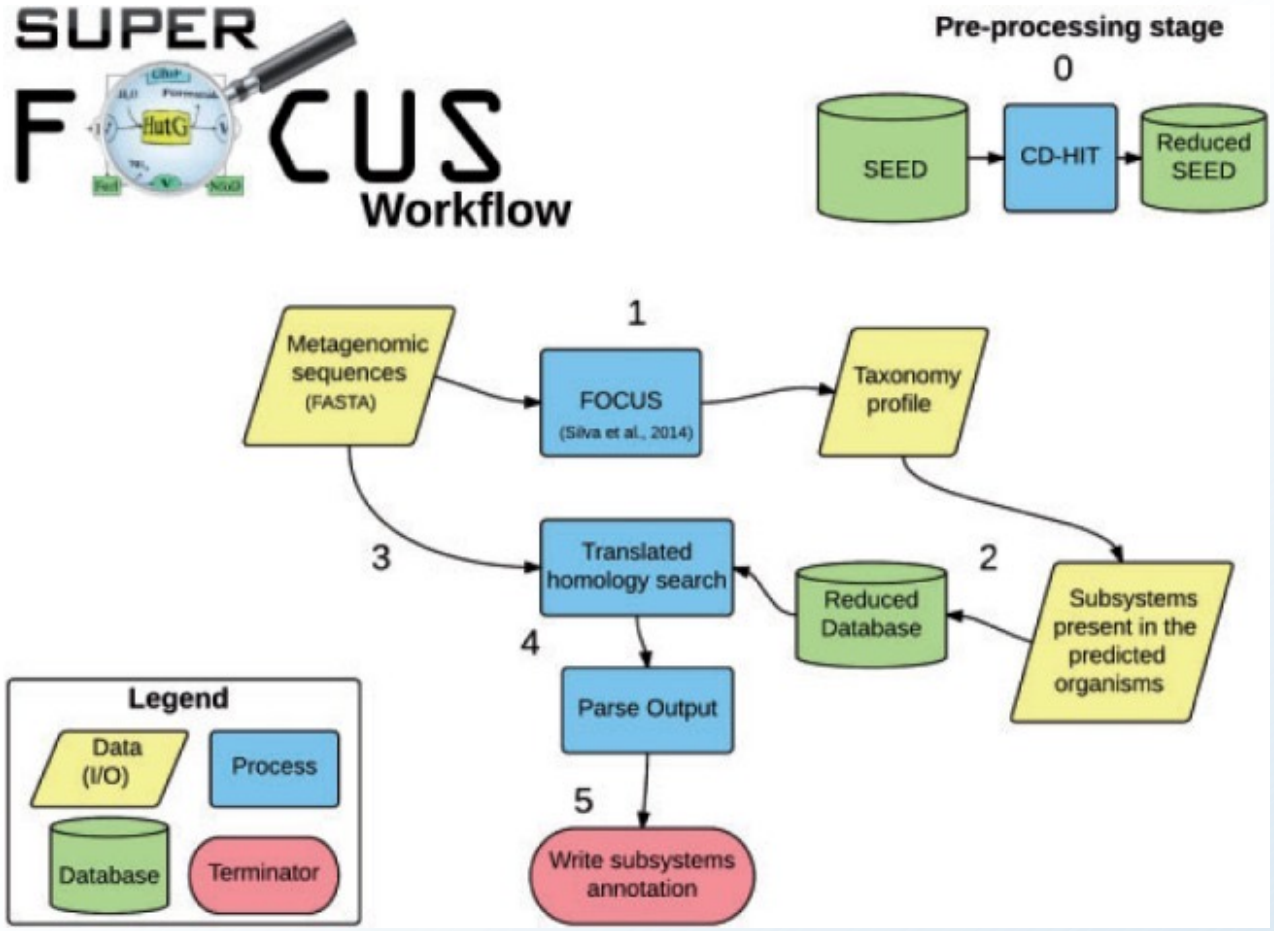
Functional assignment on sequence reads

Mapping or gene prediction



SUbsystems Profile by databasE Reduction using FOCUS

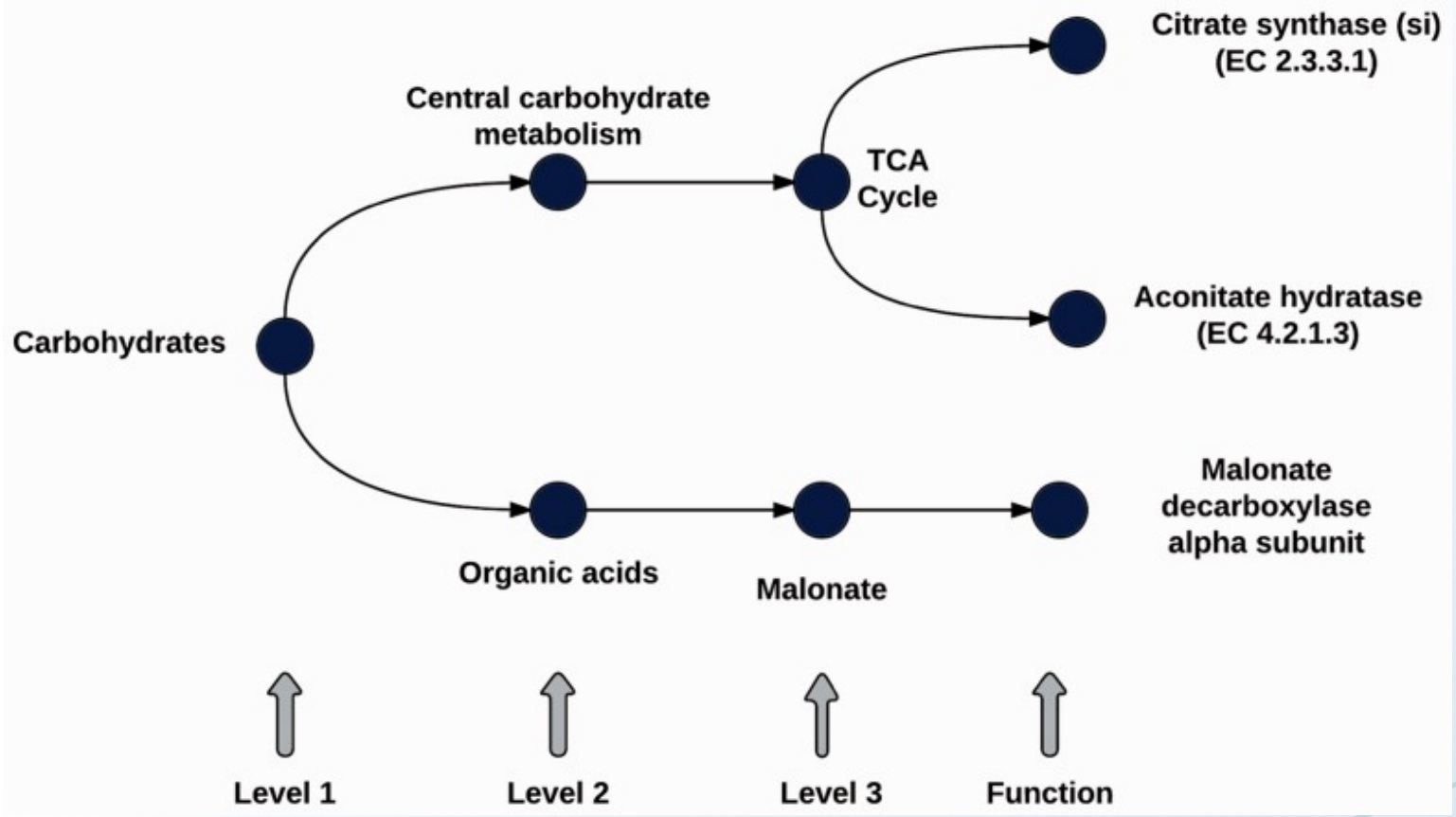
Classifies each sequence in the metagenome into a subsystem



SEED database

SEED houses subsystems - collections of functionally related protein families

Composed of subsystems structured into three levels



SUPER-FOCUS outputs tables for each subsystem

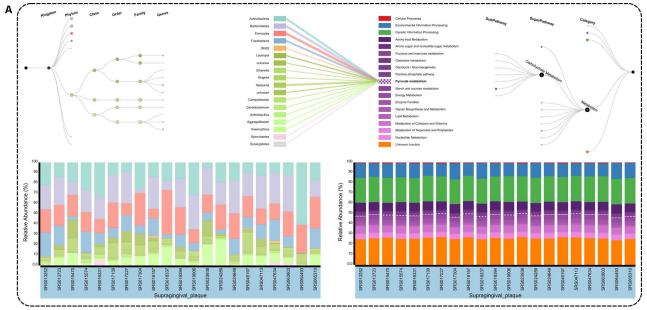
Two columns per sample:

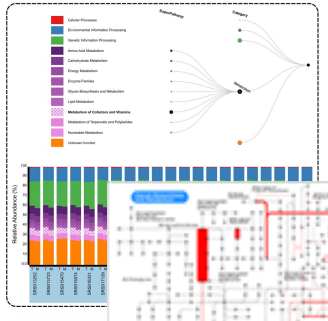
Read counts - how many reads maps to each subsystem

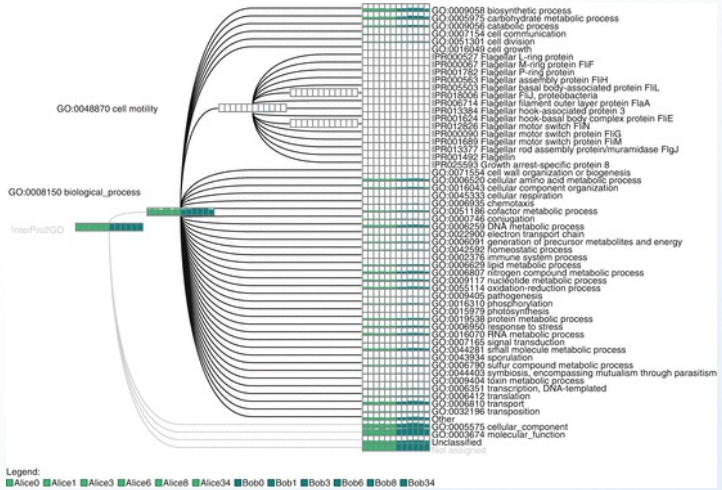
Fractional abundance - how many percent of the total reads maps to each subsystem

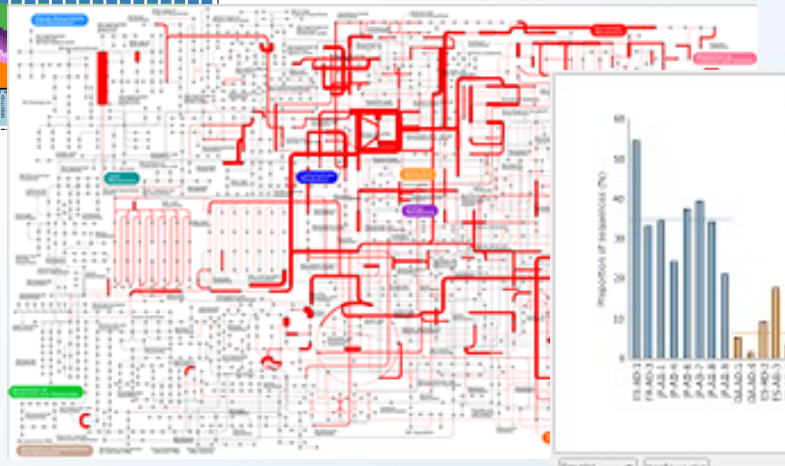
Subsystem 1	Sample1 counts	Sample1 %	Sample2 counts	Sample2 %
Amino Acids and Derivatives	100000	10	500000	50
Carbohydrates	200000	20	100000	10
Cell Division and Cell Cycle	300000	30	200000	20
Cell Wall and Capsule	300000	30	100000	10
Central metabolism	100000	10	100000	10
…	…	…	…	…
…	…	…	…	…

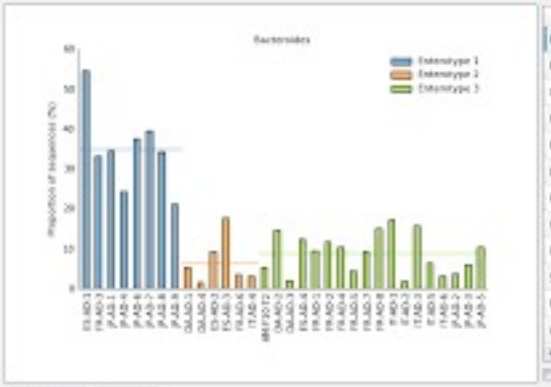
Visualizing taxonomic or metabolic profiles - it is a jungle out there….











Many of these tools also include statistical methods

MEGAN

MG-RAST

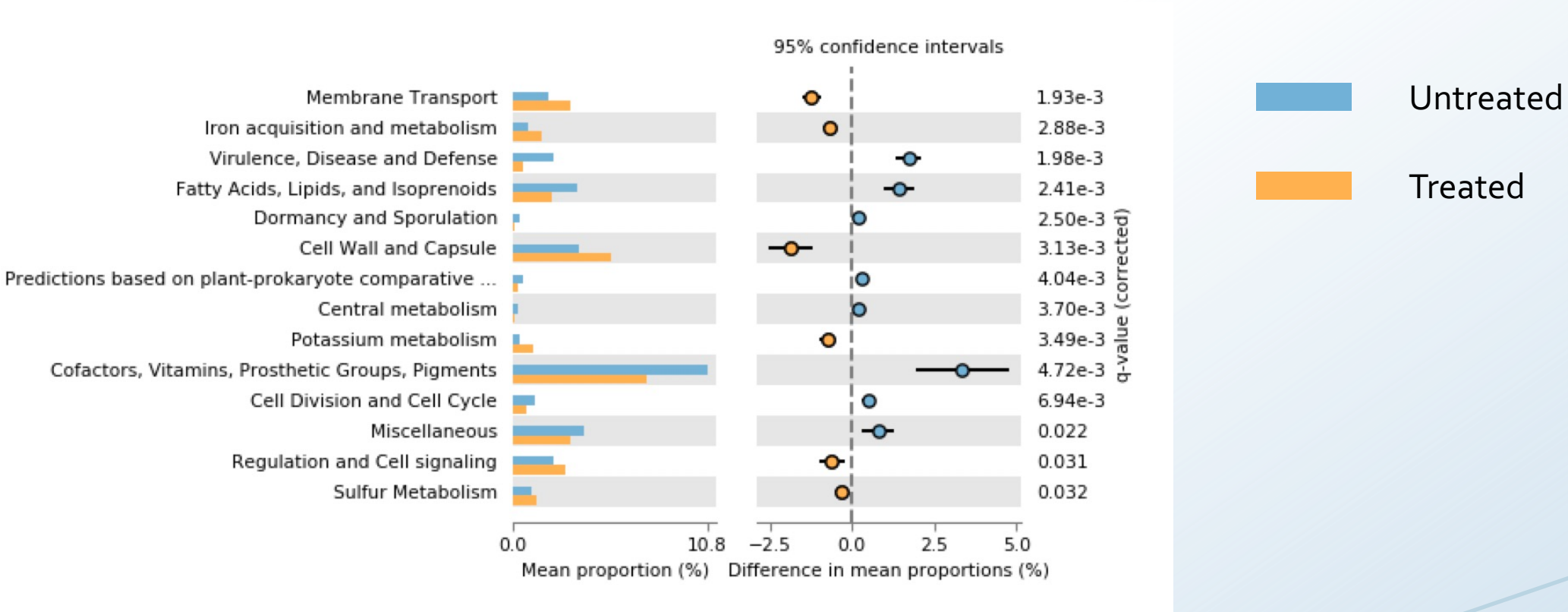
STAMP

For doing more advanced statistical tests – R

Online tools: http://elbo.gs.washington.edu/software_burrito.html

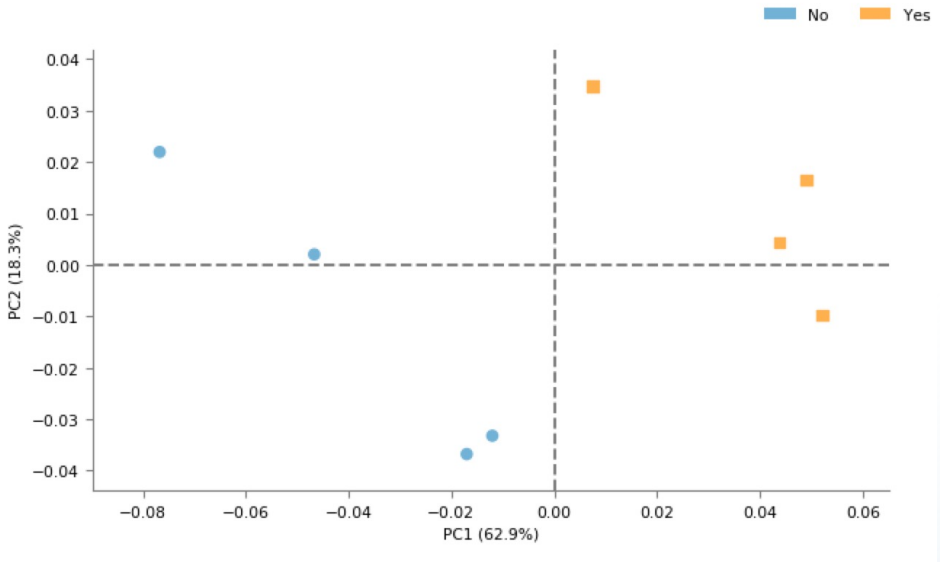
STAMP: Statistical analysis of taxonomic and functional profiles

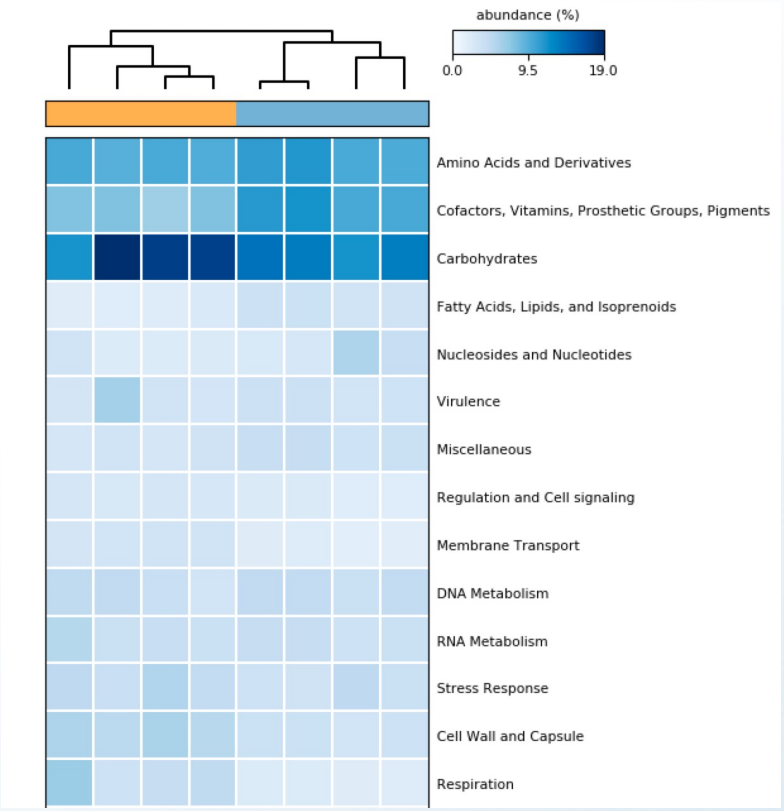
Software package for analysing taxonomic or metabolic profiles

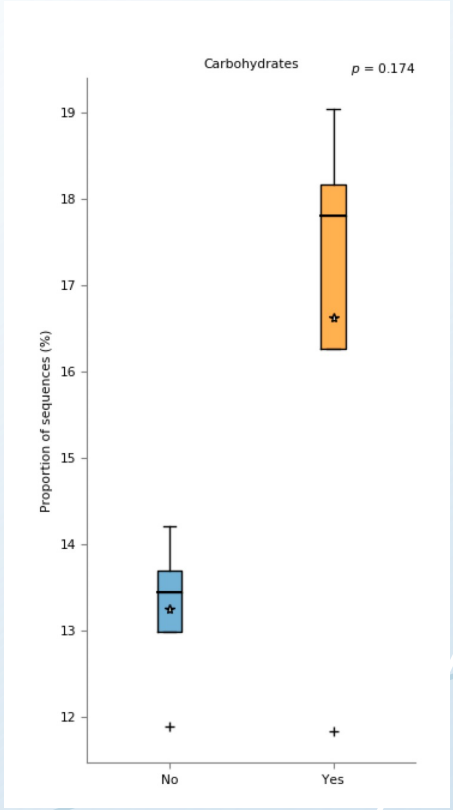


STAMP: Statistical analysis of taxonomic and functional profiles

Statistical hypothesis tests for pairs of samples or groups of samples is support along with a wide range of exploratory plots





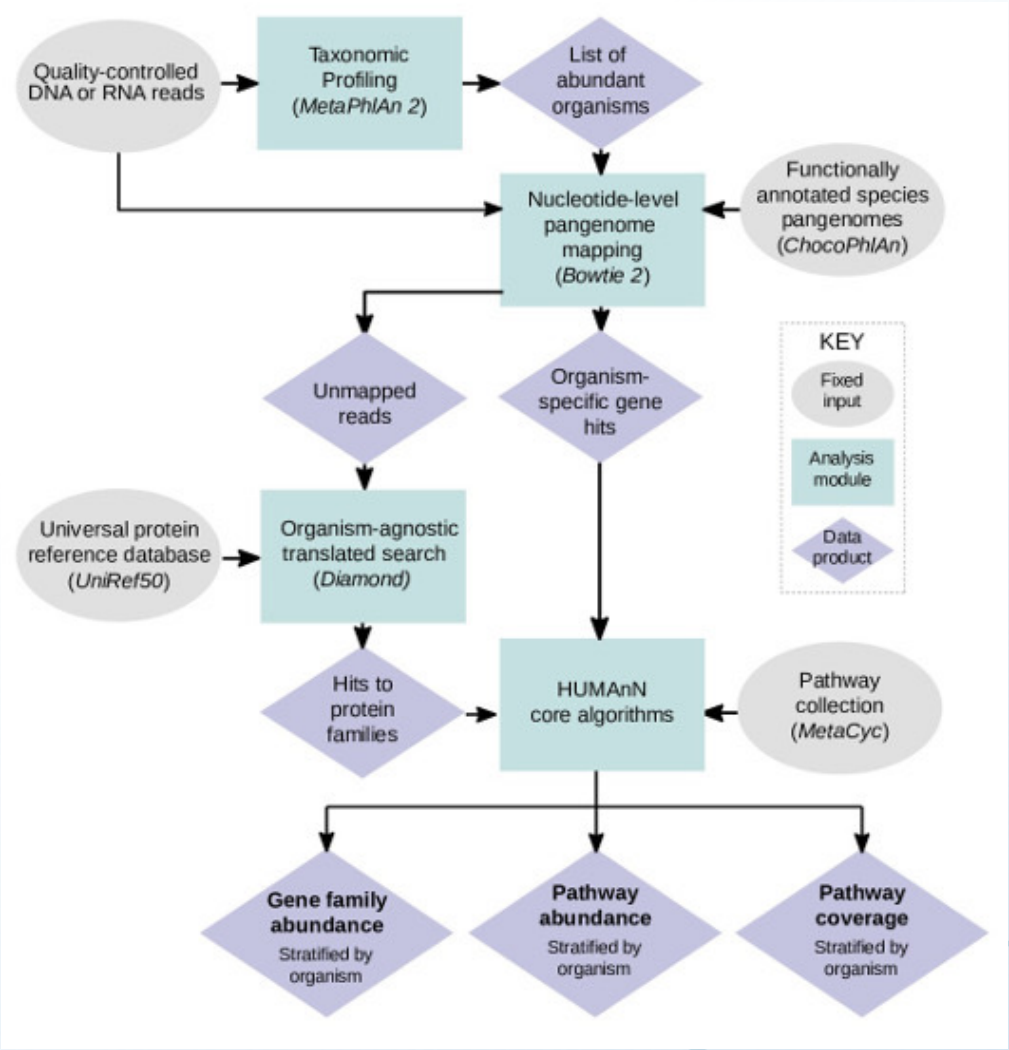


HUMAnN2 - The HMP Unified Metabolic Analysis Network 2

Analysis pipeline for abundance of microbial pathways

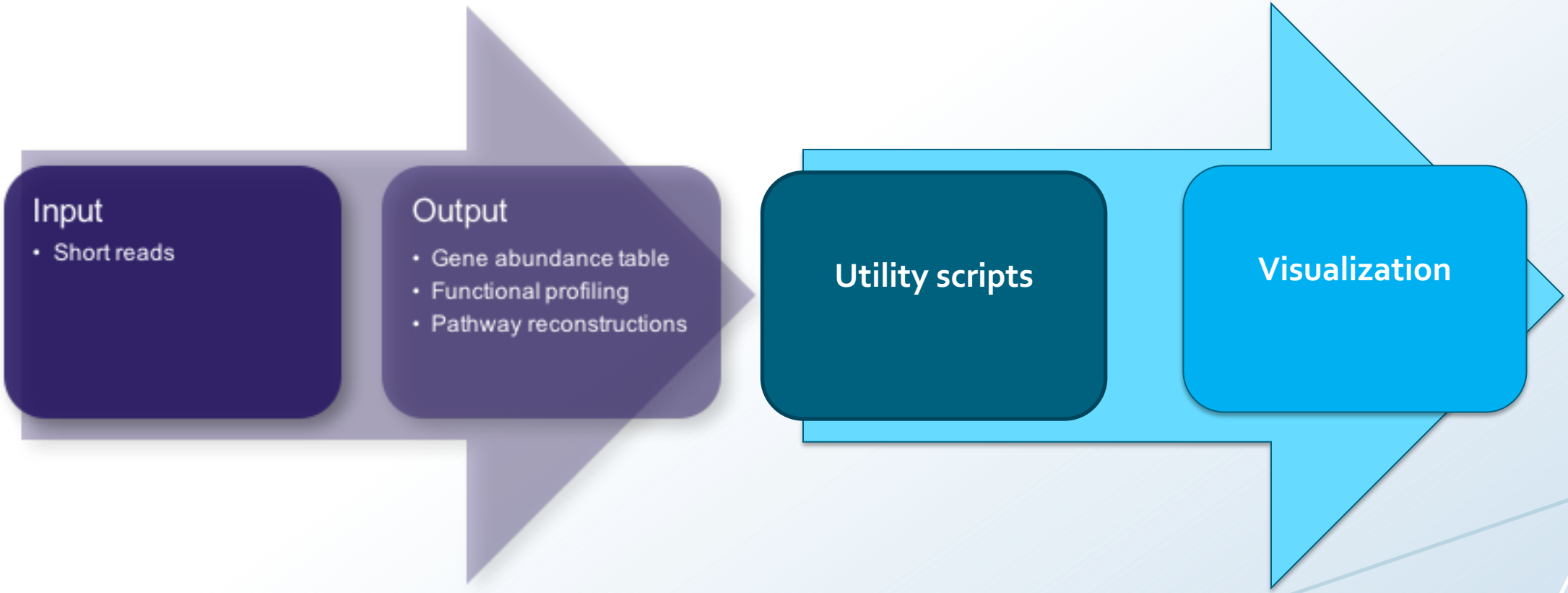
Allows comparisons of multiple samples

Combine taxonomic information with functional information



HUMAnN2 - The HMP Unified Metabolic Analysis Network 2

Utility scripts for downstream analysis and visualization



HUMAnN2 visualization

Bar plot that combine functional information with taxonomic contribution