LEIDEN UNIVERSITY MEDICAL CENTER

Functional annotation of metagenomes

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

Introduction

Functional analysis

Objectives:

- Find the functional repertoire ...
  - of the identified species (taxonomic analysis).

Introduction

Functional analysis

Objectives:

- Find the functional repertoire ...
    - of the identified species (taxonomic analysis).

Challenges:

- Incomplete coverage.
- Abundance and diversity of species.
    - Homologies between species.
- NGS data:
    - Large volume of raw data.
    - Short reads.
- Proteins with unknown functions.
- Proteins with no known homologues.

Introduction

Alignment

One reference genome:

- Variant calling.
  - Strain identification (MLST).
- Functional consequences of a variant.

Introduction

Alignment

One reference genome:

- Variant calling.
  - Strain identification (MLST).
- Functional consequences of a variant.

Multiple reference genomes:

- Targeted identification.
- Related species.

Introduction

Alignment

One reference genome:

- Variant calling.
  - Strain identification (MLST).
- Functional consequences of a variant.

Multiple reference genomes:

- Targeted identification.
- Related species.

Other datasets:

- Shotgun datasets.
- 16S ribosomal RNA.
- Every known reference sequence (BLASTN).

Introduction

Alignment

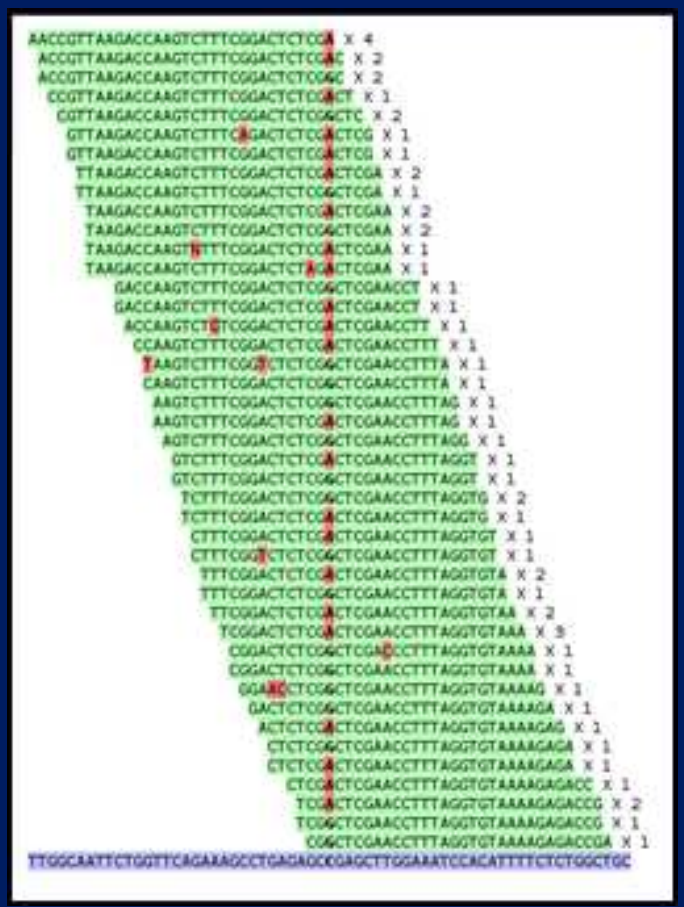

Figure 1: Alignment example.

Introduction

Alignment

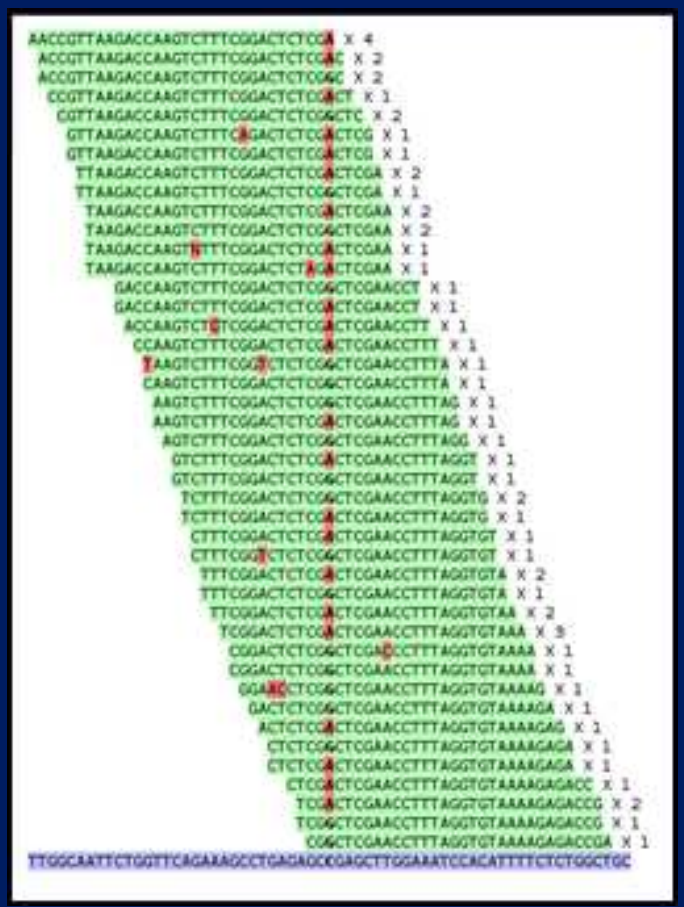

Figure 1: Alignment example.

Also useful for filtering:

- Remove contamination.
- Reduce the size of the dataset.

Introduction

Alignment

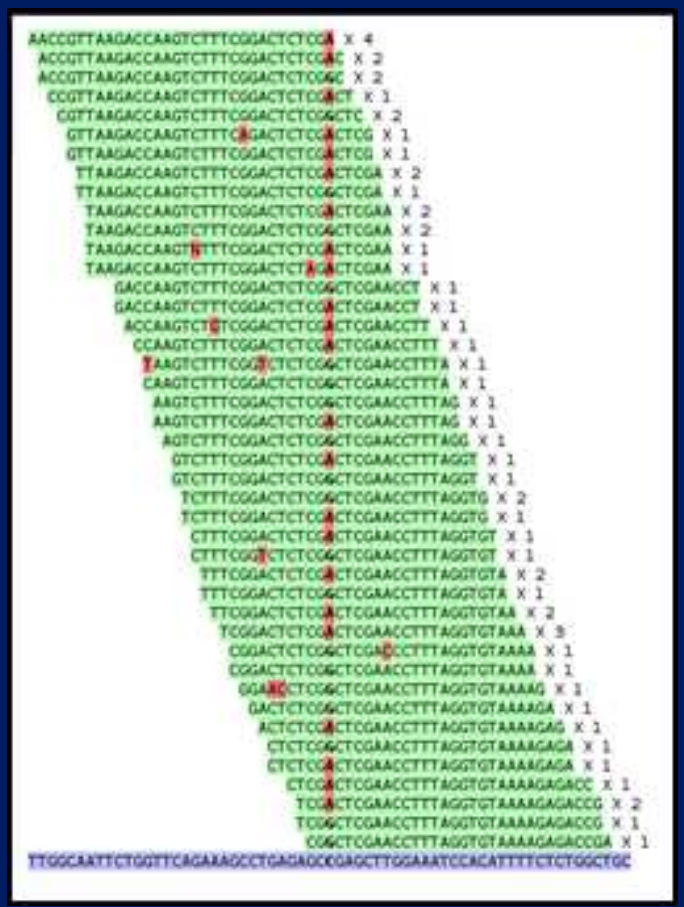

Figure 1: Alignment example.

Also useful for filtering:

- Remove contamination.
- Reduce the size of the dataset.

But beware:

- It also removes homologous areas in other species.

Targeted identification

Use case: E. coli plasmid and gene identification

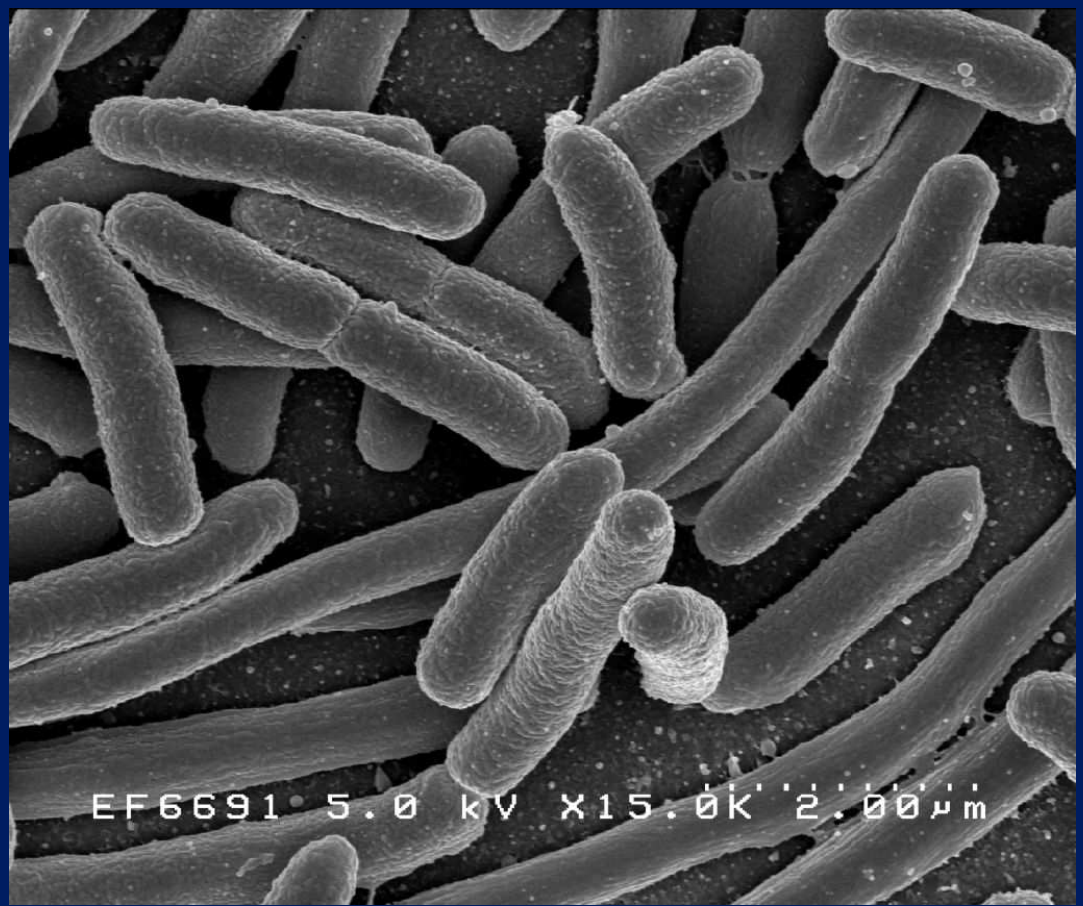

Figure 2: Escherichia coli.

Targeted identification

Some figures on the E. coli

Genome published in 1997.

- Genome size 4.6×10^6 basepairs.
- 4,288 genes in the assembly.
- 2,584 operons in the assembly.

Targeted identification

Some figures on the E. coli

Genome published in 1997.

- Genome size 4.6×10^6 basepairs.
- 4,288 genes in the assembly.
- 2,584 operons in the assembly.

However, per individual strain:

- Between 4,000 and 5,500 genes.
- 16,000 genes in total (pangenome).

Targeted identification

Some figures on the E. coli

Genome published in 1997.

- Genome size 4.6×10^6 basepairs.
- 4,288 genes in the assembly.
- 2,584 operons in the assembly.

However, per individual strain:

- Between 4,000 and 5,500 genes.
- 16,000 genes in total (pangenome).

Very diverse, only 20% of the genome is shared between all strains.

We could view this as a simple metagenome.

Targeted identification

Plasmids

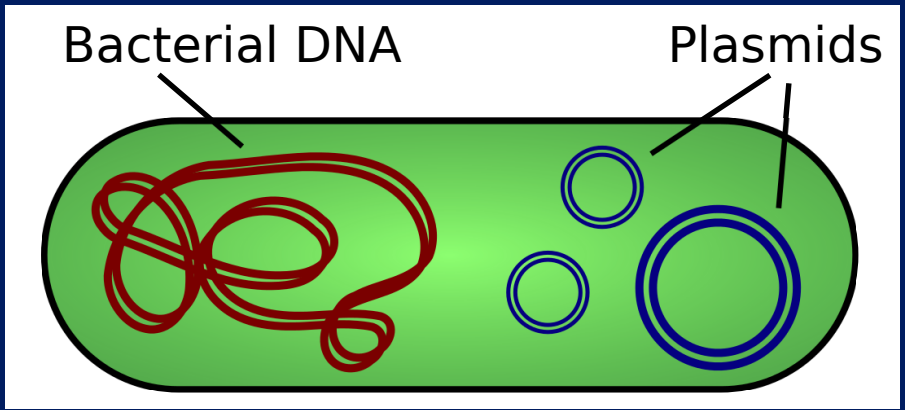

Figure 3: Schematic overview of a cell containing plasmids.

Targeted identification

Plasmids

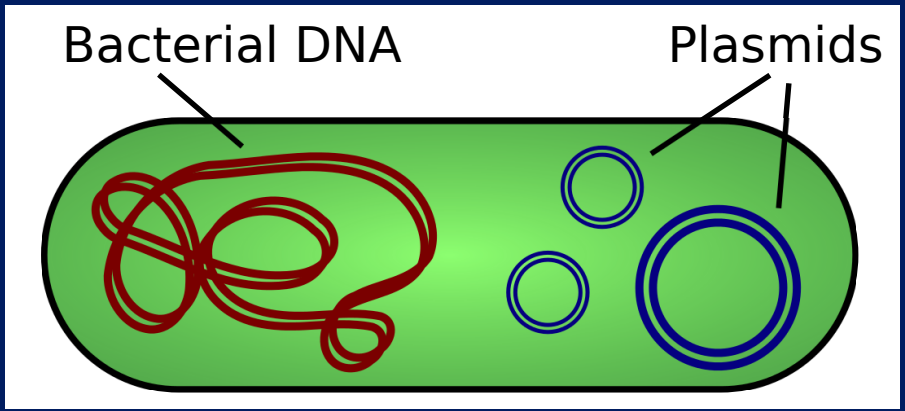

Figure 3: Schematic overview of a cell containing plasmids.

Plasmids are small DNA molecules.

- Separate and independent from the chromosome.
- Can be transferred to other species.
- Size between 1×10^3 and 1×10^6 basepairs.
- Copy number between 1 and 1,000.
- Variable between strains and individuals.

Targeted identification

Profiling

Targeted identification

Profiling

Plasmids:

- May carry antibiotic resistance genes.
- Not all of them are known.
- May be highly similar to other plasmids.

Targeted identification

Profiling

Plasmids:

- May carry antibiotic resistance genes.
- Not all of them are known.
- May be highly similar to other plasmids.

Genes:

- Multi Locus Sequence Typing (MLST).
  - Uses household genes (genomic).
  - Fragments of 450 to 500 basepairs.

Targeted identification

Profiling

Plasmids:

- May carry antibiotic resistance genes.
- Not all of them are known.
- May be highly similar to other plasmids.

Genes:

- Multi Locus Sequence Typing (MLST).
  - Uses household genes (genomic).
  - Fragments of 450 to 500 basepairs.
- Antibiotic resistance.
  - The gene may be known, the plasmid may not be.

Targeted identification

Profiling

Plasmids:

- May carry antibiotic resistance genes.
- Not all of them are known.
- May be highly similar to other plasmids.

Genes:

- Multi Locus Sequence Typing (MLST).
  - Uses household genes (genomic).
  - Fragments of 450 to 500 basepairs.
- Antibiotic resistance.
  - The gene may be known, the plasmid may not be.
- Efflux pumps.
- ...

Targeted identification

Sequencers: Ion Torrent

Figure 4: Ion torrent.

Characteristics:

- 3 hours per run.
- 1 day sampleprep, 1 day emulsion PCR.
- 4×10^6 reads.
- Read length $\pm$300bp.
- 2 *E. coli* per run.

Targeted identification

Sequencers: Ion Torrent

Figure 4: Ion torrent.

Characteristics:

- 3 hours per run.
- 1 day sampleprep, 1 day emulsion PCR.
- 4×10^6 reads.
- Read length $\pm$300bp.
- 2 *E. coli* per run.

Fast and inexpensive.

Targeted identification

General overview

We screen for 130 known plasmids and 400 genes.

Targeted identification

General overview

We screen for 130 known plasmids and 400 genes.

Output:

- MLST.
- List of plasmids.
    - Otherwise, similar plasmids.
- List of genes of interest.

Targeted identification

General overview

We screen for 130 known plasmids and 400 genes.

Output:

- MLST.
- List of plasmids.
    - Otherwise, similar plasmids.
- List of genes of interest.

For the MLST, we need a list of variants

- Covered in the *NGS introduction course* . . .
- and the previous talk.

Targeted identification

Plasmid detection

Pipeline:

- Select all reads that do not map to the genome.
- Map these reads to each plasmid individually.
- Calculate the *horizontal coverage*.

Targeted identification

Plasmid detection

Pipeline:

- Select all reads that do not map to the genome.
- Map these reads to each plasmid individually.
- Calculate the *horizontal coverage*.

Some notes:

- This overestimates the number of plasmids.
- Should be used as an indication of presence.
  - E.g., 80% of a plasmid can be found.
- Homologies between plasmids should be known.
- Recombination can be an issue.

Targeted identification

Coverage

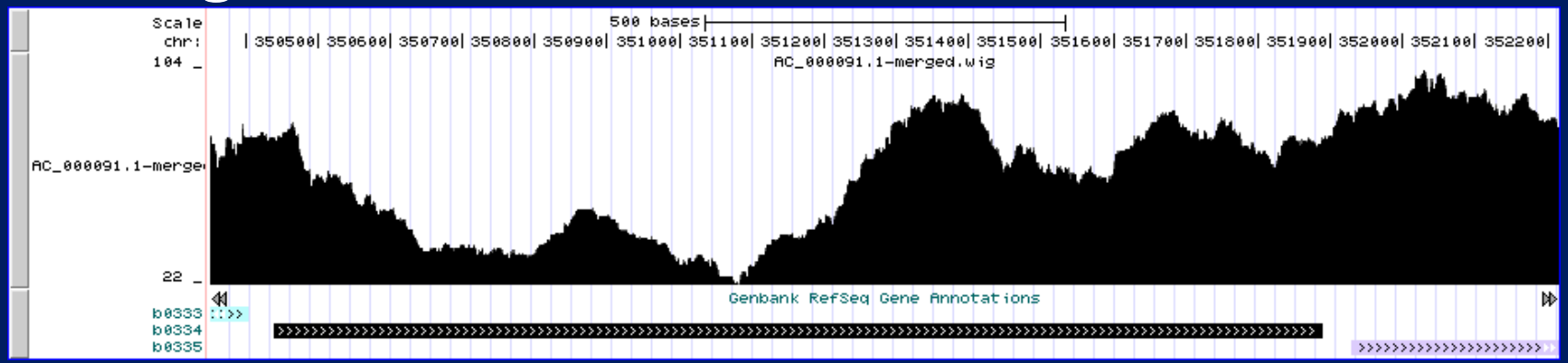

Figure 5: Coverage / depth histogram.

Targeted identification

Coverage

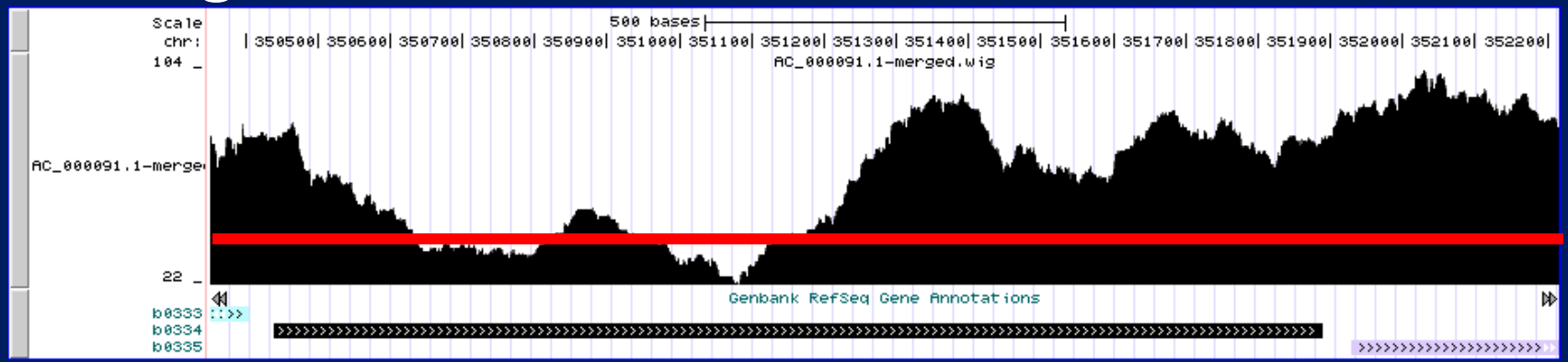

Figure 5: Coverage / depth histogram.

Targeted identification

Coverage

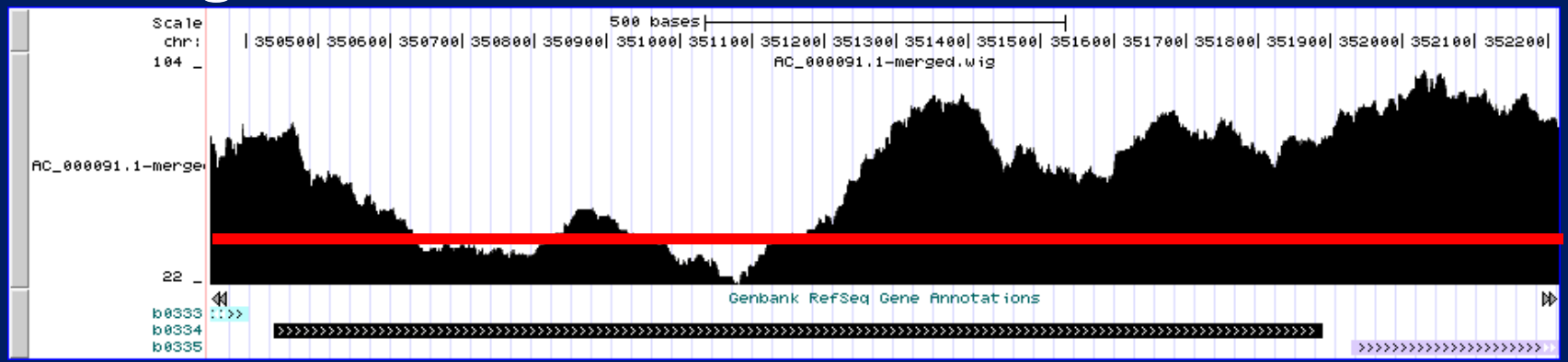

Figure 5: Coverage / depth histogram.

Figure 6: Coverage summary.

From this, we can easily calculate the percentage of the gene we found.

Targeted identification

Plasmid detection

Plasmid	Size	Reads	#3/#2	Cov	#5/#2
NC_001537	3895	18728	4.808	1418	0.364
NC_002119	9957	6130	0.615	789	0.079
NC_002127	3306	11749	3.553	1068	0.323
NC_002128	92721	11824	0.127	35783	0.385
NC_002142	68817	8163	0.118	15938	0.231
NC_002145	1549	46141	29.787	1549	1.000
NC_002487	5847	11669	1.995	1735	0.296
NC_002525	75582	420	0.005	1325	0.017
NC_004429	6349	961	0.151	1858	0.292

Table 1: Part of the plasmids table.

Targeted identification

Gene detection

Reference	Gene	Length	Cov	#4/#3
AB699171	CMY-87	959	90	0.093
AB715422	IMP-34	742	125	0.168
AB737978	ACT-16	1062	202	0.190
AB753456	IMP-42	739	417	0.564
AB753457	IMP-40	739	414	0.560
AB753458	IMP-41	731	364	0.497
AC_000091.1	accD	915	915	1.000
AC_000091.1	acrA	1194	1194	1.000
AC_000091.1	acrB	3150	3150	1.000

Table 2: Part of the genes table.

Semi-targeted approach.

Full genome analysis

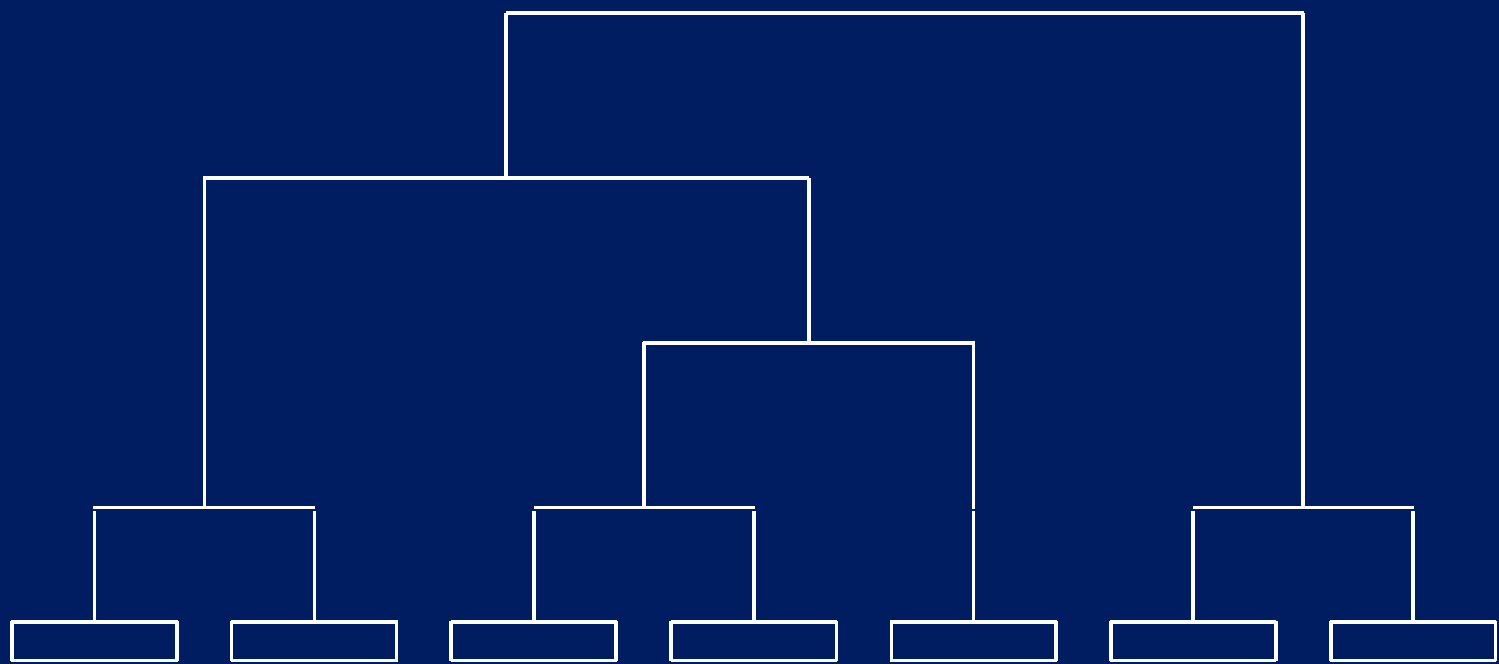

Figure 7: Horizontal coverage

Semi-targeted approach.

Full genome analysis

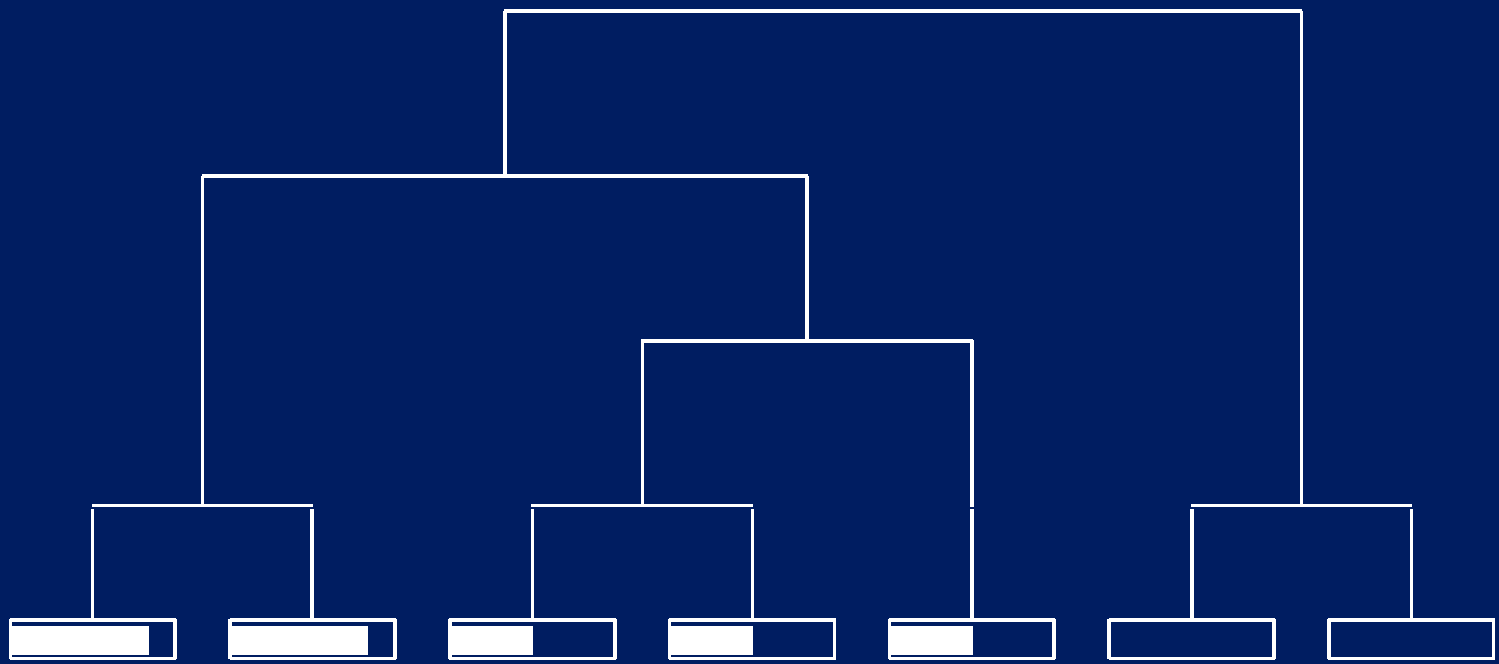

Figure 7: Horizontal coverage

Semi-targeted approach.

Full genome analysis

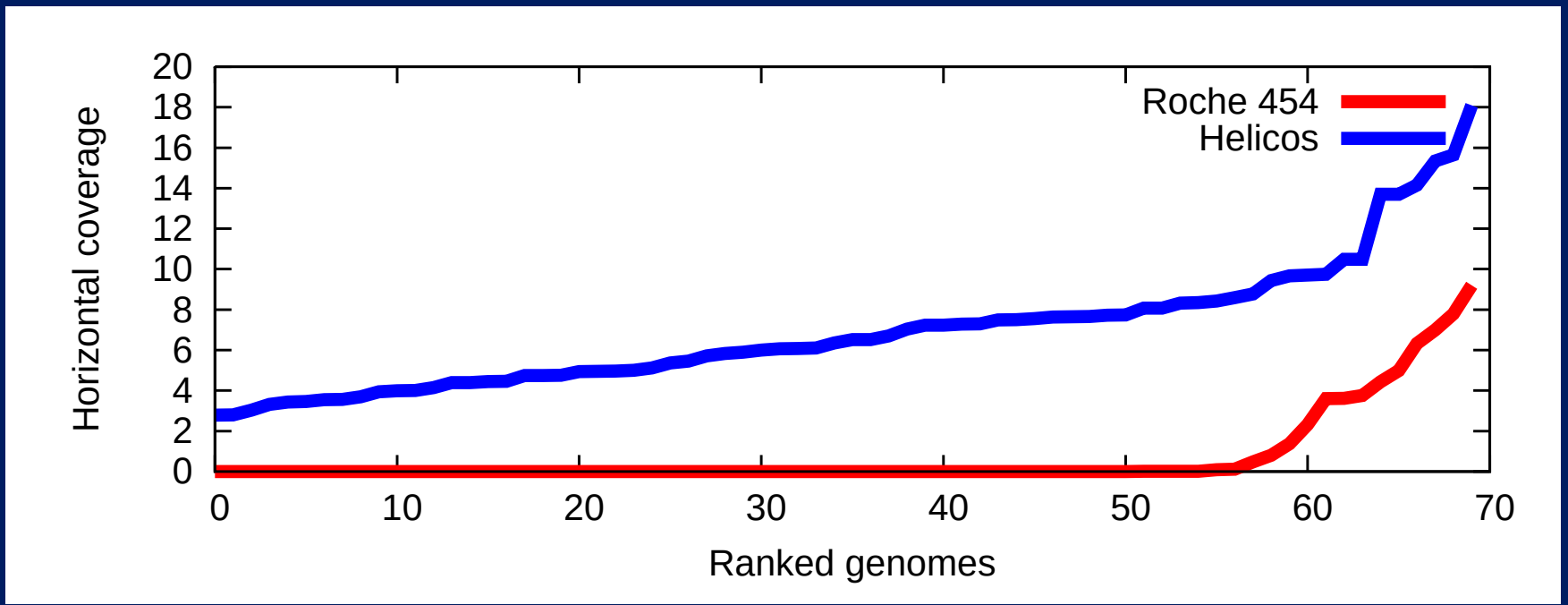

Figure 8: Horizontal coverage of ranked genomes

Functional analysis

An "unbiassed" approach

Use every available reference sequence.

- Focus on finding genes.
- Try to identify processes based on gene information.
    - The processes are not limited to one species.

Functional analysis

An "unbiassed" approach

Use every available reference sequence.

- Focus on finding genes.
- Try to identify processes based on gene information.
    - The processes are not limited to one species.

Identify genes.

- Looking at the best BLAST hist.
    - More sophisticated methods use weighed BLAST information.
- Do we have all components for a certain pathway?

Functional analysis

An "unbiassed" approach

Use every available reference sequence.

- Focus on finding genes.
- Try to identify processes based on gene information.
    - The processes are not limited to one species.

Identify genes.

- Looking at the best BLAST hist.
    - More sophisticated methods use weighed BLAST information.
- Do we have all components for a certain pathway?

Still biassed to the content of the databases used.

Functional analysis

De novo assembly

Assemble reads.

- Covered in the *De novo assembly course*.
- Can be optimised for *open reading frames*.

Functional analysis

De novo assembly

Assemble reads.

- Covered in the *De novo assembly course*.
- Can be optimised for *open reading frames*.

Find open reading frames.

- Glimmer.
- GeneMark.
- ORF-Finder.
- ...

Functional analysis

De novo assembly

Assemble reads.

- Covered in the *De novo assembly course*.
- Can be optimised for *open reading frames*.

Find open reading frames.

- Glimmer.
- GeneMark.
- ORF-Finder.
- ...

Blast these open reading frames.

- Longer sequences align easier.
- May find *homologous* genes.

Functional analysis

Pathways

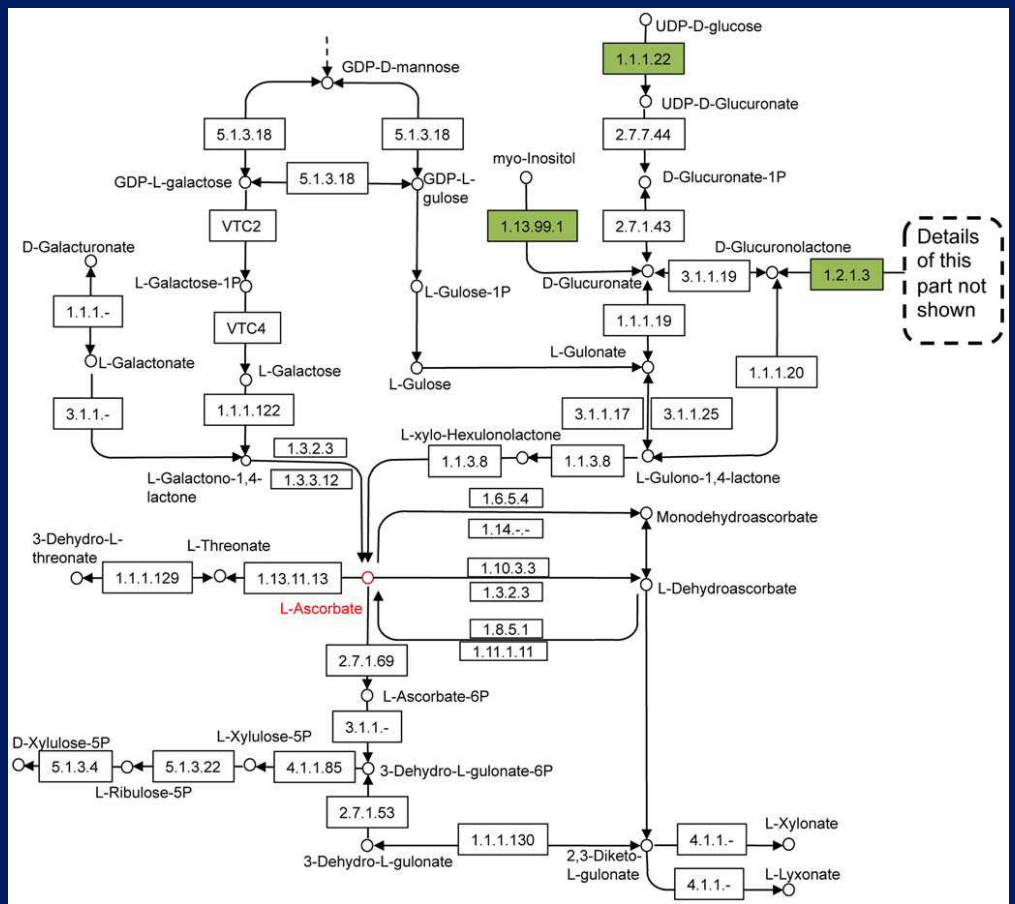

Figure 9: Example pathway (Ye et al. 2009).

Functional analysis

Identifying pathways

In general, a pathway has been found if all the genes involved in that pathway have been found.

Functional analysis

Identifying pathways

In general, a pathway has been found if all the genes involved in that pathway have been found.

This approach may lead to overestimation of:

- The number of pathways.
- The size of the pathways.

But also the underestimation of the size of a pathway.

Functional analysis

Identifying pathways

In general, a pathway has been found if all the genes involved in that pathway have been found.

This approach may lead to overestimation of:

- The number of pathways.
- The size of the pathways.

But also the underestimation of the size of a pathway.

Several approaches to solve these issues:

- Find the minimum number of pathways that explain the observed genes (MinPath).
- Smoothing or “gap filling”.
- Taxonomic limitation.

Functional analysis

Minpath

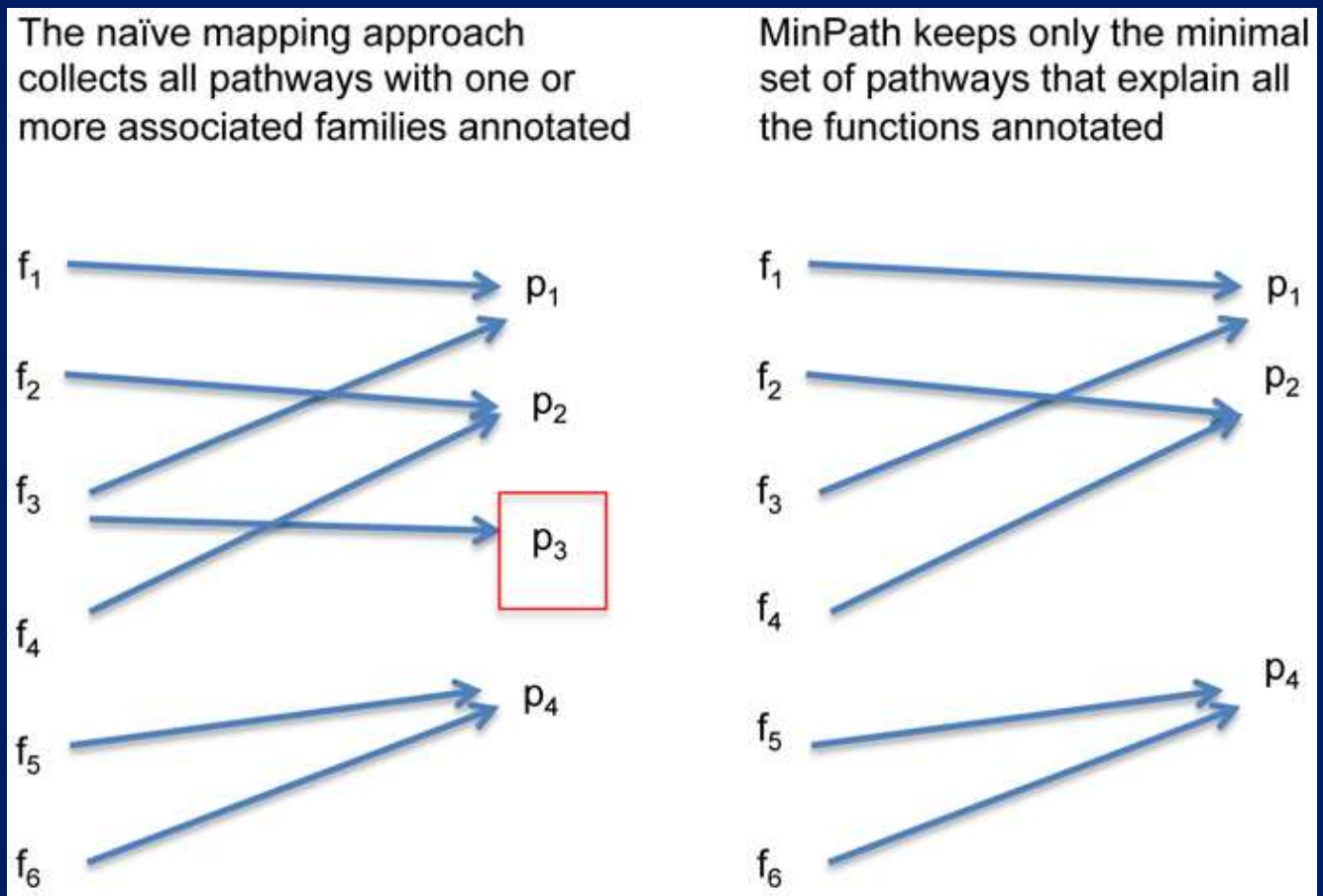

Figure 10: (Ye et al. 2009).

Functional analysis

Pipelines

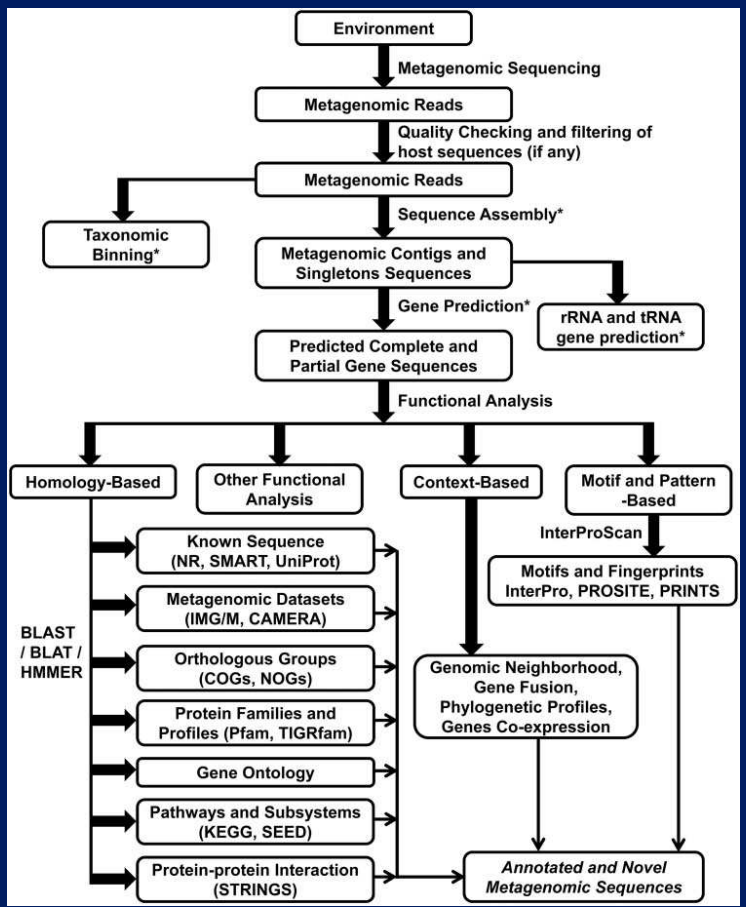

Figure 11: Prakash et al. 2002.

Functional analysis

Pipelines

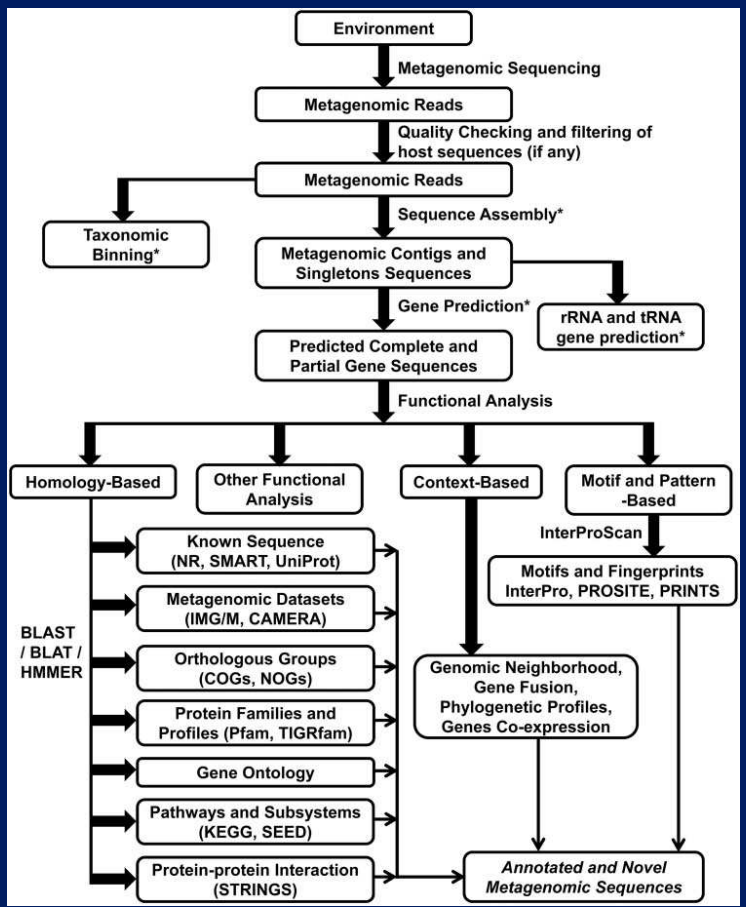

Figure 11: Prakash et al. 2002.

Some examples:

- HMP Unified Metabolic Analysis (HUMAnN).
- MetaGenomics Rapid Annotation using Subsystems Technology (MG-RAST).

Functional analysis

HUMAnN: Human Microbiome

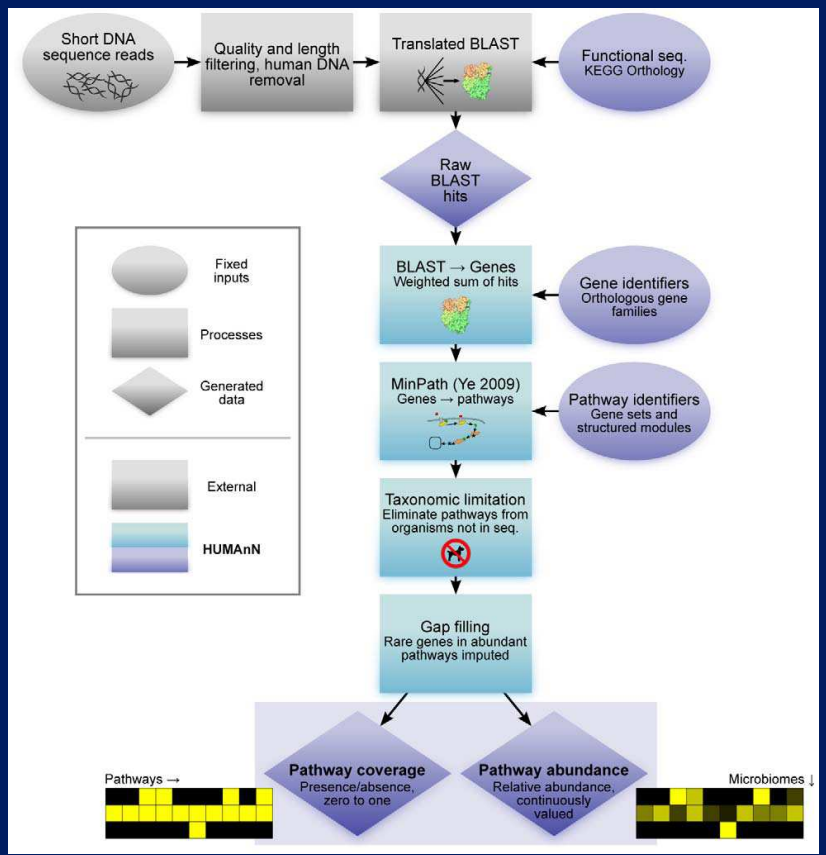

Figure 12: Abucker et al. 2012.

Functional analysis

HUMAnN: Human Microbiome

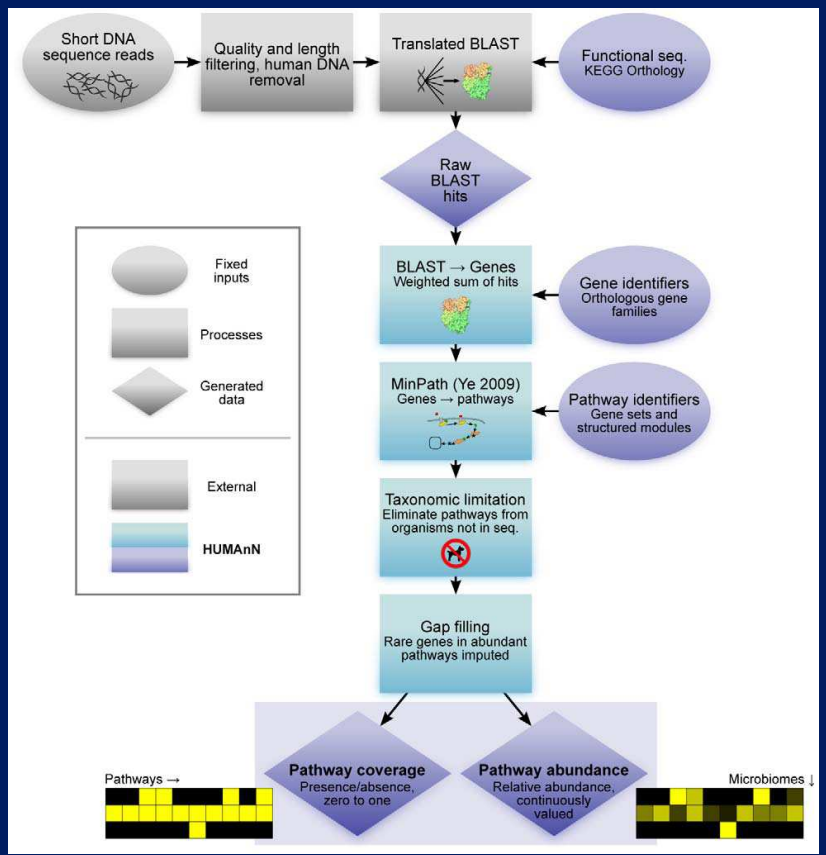

Figure 12: Abucker et al. 2012.

This pipelines combines many tools:

- Data cleaning.
- Blasting (identify organisms).
- Functional translation / pathways.
- Taxonomic limitation.
- ...

Functional analysis

MG-RAST pipeline overview

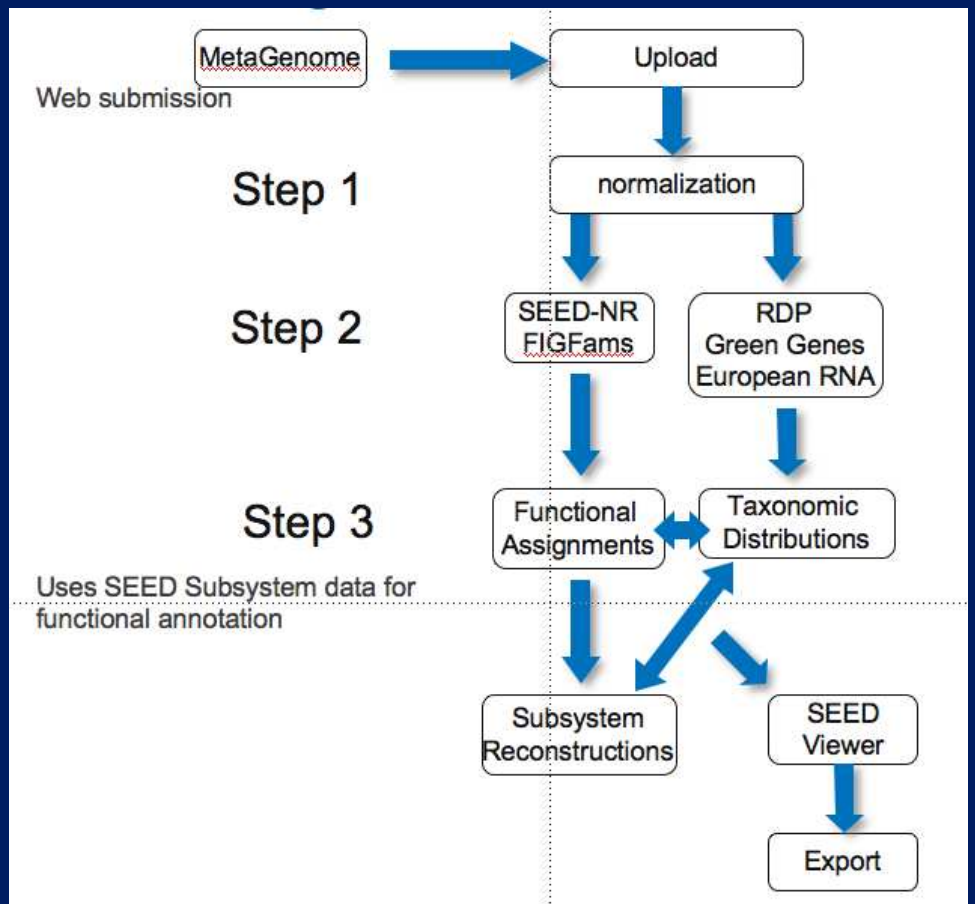

Figure 13: Simplified overview of the metagenomic pipeline.

Functional analysis

MG-RAST pipeline overview

Normalisaton / QC:

- Deduplication, quality / length filtering ($\leq$ 75bp).
- Model organism filtering.

Functional analysis

MG-RAST pipeline overview

Normalisaton / QC:

- Deduplication, quality / length filtering ($\leq$ 75bp).
- Model organism filtering.

Search for genes:

- Use BLASTX on the SEED database.
- Different alignments for specific databases:
  - Ribosomal: GREENGENES, RDP-II, 16S (RNA).
  - Cloroplast, mitochondrial.
  - ACLAME (mobile elements).

Functional analysis

MG-RAST pipeline overview

Normalisaton / QC:

- Deduplication, quality / length filtering ($\leq$ 75bp).
- Model organism filtering.

Search for genes:

- Use BLASTX on the SEED database.
- Different alignments for specific databases:
  - Ribosomal: GREENGENES, RDP-II, 16S (RNA).
  - Cloroplast, mitochondrial.
  - ACLAME (mobile elements).

Phylogenetic reconstruction:

- Combine the results from the previsous step.

Questions?

Acknowledgements:

Bas Dutilh
Victor de Jager
Johan den Dunnen