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

Also useful for filtering:

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

Figure 1: Alignment example.
Introduction

Alignment

Also useful for filtering:

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

But beware:

- It also removes homologous areas in other species.

Figure 1: Alignment example.
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 $4.6 \times 10^6$ 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 $4.6 \times 10^6$ 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 $4.6 \times 10^6$ 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.
Plasmids are small DNA molecules.

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

Figure 3: Schematic overview of a cell containing plasmids.
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

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

Figure 4: Ion torrent.
Targeted identification

**Sequencers: Ion Torrent**

Characteristics:

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

Figure 4: Ion torrent.

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

### Plasmid detection

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

Table 1: Part of the plasmids table.
### Gene detection

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

Table 2: Part of the genes table.
Semi-targeted approach.

Full genome analysis

Figure 7: Horizontal coverage
Full genome analysis

Figure 7: Horizontal coverage
Semi-targeted approach.

**Full genome analysis**

Figure 8: Horizontal coverage of ranked genomes
Functional analysis

An “unbiased” 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.
An “unbiased” 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.
De novo assembly

Assemble reads.

- Covered in the *De novo assembly course*.
- Can be optimised for *open reading frames*.
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.
Figure 9: Example pathway (Ye et al. 2009).
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**

The naïve mapping approach collects all pathways with one or more associated families annotated

MinPath keeps only the minimal set of pathways that explain all the functions annotated

---

**Figure 10: (Ye et al. 2009).**
Figure 11: Prakash et al. 2002.
Some examples:

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

Figure 11: Prakash et al. 2002.
**Functional analysis**

**HUMAnN: Human Microbiome**

Figure 12: Abucker et al. 2012.
Functional analysis

**HUMAnN: Human Microbiome**

This pipelines combines many tools:

- Data cleaning.
- Blasting (identify organisms).
- Functional translation / pathways.
- Taxonomic limitation.
- ...
MG-RAST pipeline overview

Figure 13: Simplified overview of the metagenomic pipeline.
**Functional analysis**

**MG-RAST pipeline overview**

Normalisation / QC:
- Deduplication, quality / length filtering ($\leq 75$bp).
- Model organism filtering.
Functional analysis

**MG-RAST pipeline overview**

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

**Search for genes:**
- Use BLASTX on the SEED database.
- Different alignments for specific databases:
  - Ribosomal: GREENGENSES, RDP-II, 16S (RNA).
  - Chloroplast, mitochondrial.
  - ACLAME (mobile elements).
**Functional analysis**

**MG-RAST pipeline overview**

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

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

**Phylogenetic reconstruction:**
- Combine the results from the previous step.
Questions?

Acknowledgements:

Bas Dutilh
Victor de Jager
Johan den Dunnen