MetaPhlAn

Metagenomic Phylogenetic Analysis
### Where are we?

<table>
<thead>
<tr>
<th>Time</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 weeks</td>
<td>From 16S sequencing to taxonomic profiles</td>
</tr>
<tr>
<td>2 weeks</td>
<td>From shotgun metagenomics to species- and strain-level taxonomic profiles</td>
</tr>
<tr>
<td>2 weeks</td>
<td>Comparing Taxonomic Profiles: Beta Diversity and Balance Trees</td>
</tr>
<tr>
<td>2 weeks</td>
<td>Inferring Species Interactions</td>
</tr>
</tbody>
</table>

**Tools used:**
- **MetaPhlAn** (2012)
- **SWARM** (2014)
- **DADA2** (2016)
- **Deblur** (2017)
Outline

- General motivation
- 16S rRNA sequencing
- Full shotgun sequencing
- MetaPhlAn
- Results
- Conclusions & Discussion
Who is there?

What do they do?
16S rRNA gene

- Encodes part of the bacterial ribosome
- Present in all bacteria
- Very short usually (~1.5Kbp)
16S rRNA gene

- Has conserved and **variable** regions
- Conserved areas → relationship among species
- Highly variable areas → differences between species
16S rRNA sequencing
16S rRNA analysis

- We saw different pipelines to determine:
  - Taxonomy
  - Relative abundances (RA)
Who is there? Solved!

- 16S technique seems to work great
- It’s solid, based on biological logic, and widely used
Well not entirely...

- Copy number variation
Well not entirely...

- Copy number variation
- PCR amplification bias
Well not entirely…

- Copy number variation
- PCR amplification bias
- PCR chimera formation
Well not entirely...

- Copy number variation
- PCR amplification bias
- PCR chimera formation
- Limited resolution under genus
Who is there?
What do they do?
Full shotgun sequencing

From sequencing **one gene**

Into sequencing the **entire genome**
Full shotgun sequencing

**Motivation**

1. **Functional** information (e.g. gene 1 is present)
2. Detecting **non-bacterial** species (can’t use 16S rRNA gene)
3. Handling “**unseen**” genomes better
Full shotgun sequencing

Reality is complex....

How can we tell “who is there”? 
MetaPhlAn (2012)

Curtis HUTTENHOWER
Associate Professor of Computational Biology and Bioinformatics, Department of Biostatistics, Chan School of Public Health, Harvard University, USA
MetaPhlAn - General intuition

● We’re no longer limited to a single special gene

● Given the **full genomes** of our target species…

● We could look for “unique regions”
MetaPhlAn - General intuition

- The region is “hit” by a read? → the specie is there!

- We’ll call those regions *clade-specific marker-genes*
Some definitions before we move on

- **Def:**
  **Clades** are groups of genomes (organisms) believed to have evolved from a common ancestor.

  It can be as specific as **species** or as broad as **phyla**.
Some definitions before we move on

- **Def:**
  
  **Clade-specific marker-genes** are sequences that satisfy:

  - **Being strongly conserved** within the clade’s genomes.
  - **Not similar** to any sequence in other clades (of the same level)
Clade-specific marker-genes

- Unique markers change as the clade level grows
- They also accumulate in a way (direct vs indirect)...
MetaPhlAn - High level algorithm

- Use a set of reference genomes
- And a reference taxonomy
- Find clade-specific marker-genes
- Sequence sample
- Map reads to unique marker-genes
- Calculate “who is there”
MetaPhlAn - High level algorithm

- Use a set of reference genomes
- And a reference taxonomy
- Find clade-specific marker-genes
- Sequence sample
- Map reads to unique marker-genes
- Calculate “who is there”
Step 1: Acquire reference genomes

- **2,887 genomes** available from the Integrated Microbial Genomes (IMG) system

- The genomes are classified by quality measures

- They are based on many different public data sources (non-homogenic!)
Step 1: Acquire reference genomes

- The methods in which the genomes are acquired vary

`de-novo assembly`
Step 2: Acquire reference taxonomy

- *The basic idea:* hierarchical clustering tree based on genome similarity
Step 2: Acquire reference taxonomy

- *Wait...*
  Similarity of what sequences?

- Whole genomes are too long, too diverged

- Most databases still look for evolutionary justifications → distances are not based on the entire genome

- The 16S rRNA gene still plays an important part (!)
Step 1 & 2: Acquire reference genomes & taxonomy

- 2,887 genomes
- 2,834 "quality" genomes
- 2,383 taxonomic clades
- 2 domains ➔ 33 phyla ➔ 66 classes ➔ 130 orders ➔ 278 families ➔ 652 genera ➔ 1,221 species
Step 3: Locate unique marker-genes

- The general process:
  - Each genome → bag-of-genes representation
  - Only conserved genes in the clade are saved
  - Inter-clade uniqueness index elimination
  - Single-copy genes were preferred over multi-copy genes
Step 3: Locate unique marker-genes

- Properties of the markers
  - **Gene** level
  - 400,141 filtered genes (out of original ~2M possible)
  - Not necessarily continuous (bag-of-genes)
  - ~4% of the total genome length
  - 231 (± 107) markers per specie
  - Only 12 species with < 15 markers
Step 3: Locate unique marker-genes

- Properties of the markers
  - As the clade grows (higher taxonomic levels) → it is usually **well covered**
  - Allowing MetaPhlAn to recover relative abundances within broader clades even in the absence of sequenced genomes
Calculate relative abundances (who is there)

- **Normalization**
  - Sum the total reads mapped to clade markers
  - Divide by marker’s total length
  - Abundances in every clade-level sum up to 100%

![Diagram showing specie A and specie B with relative abundances of 72% and 28% respectively.](image)
Map reads to unique marker-genes

- The unclassified case:
  - Clade abundances obtained by direct read mapping
Map reads to unique marker-genes

- The unclassified case:
  - The same is done to the clade at the next level
Map reads to unique marker-genes

- The unclassified case:
  - What if there is a contradiction?
Limitations

- Balls into bins problem
- Happens in low abund. (theoretical prob. can be calculated)
- MetaPhlAn is known to operate poorly at low abundances
Estimating performance

- Synthetic communities
  - 2 samples with 100 genomes each (high-complexity)
  - 8 samples with 25 genomes each (low-complexity)
  - Total of 4M synthetic noisy reads
Estimating performance
Estimating performance

C

- MetaPhIAn ($r = 0.997$)
- PhymmBL ($r = 0.887$)
- Phymm ($r = 0.888$)
- RITA ($r = 0.812$)
- NBC ($r = 0.880$)

Predicted relative abundance (%) vs. Expected relative abundance (%)

Graph showing a scatter plot with lines and markers for different methods.
Estimating performance
higher resolution compared to 16S
Conclusion

- Reality is complex. There is no one solution.
- All methods have internal biases and limitations.
- MetaPhlAn sets a good standard in many settings.
- It is commonly used, and very well known.
- These are still areas of active research, and progress is still at a steep curve.
Points for discussion

● What are the biggest generators of noise in the process?
● What happens when the true species skewes from the ref. genome?
● How can I know what sequencing depth I should use?...
Thanks!

Shahar Azulay
shahar4@gmail.com