Abundance-based reconstitution of microbial pan-genomes from whole-metagenome shotgun sequencing data

Florian PLAZA OÑATE
fplaza-onate@enterome.com

RCAM 2017
09/10/2017
Background
The strain-level metagenomics area

- Shotgun metagenomics: whole-community DNA randomly sequenced
- Achieves strain-level characterization

Available tools hampered by reference genomes databases:
  - Few strains available for many species: unknown accessory genes
  - Many species lack reference genomes.

→ Metagenomic samples are an untapped reservoir of information.
Building microbial gene catalog by *de novo* assembly

Millions of genes from > 1K species
Quantitative metagenomics pipeline

- Sampling
- DNA extraction
- Sequencing
- Mapping
- Counting
- Statistics

Gene catalog
Need for structured gene catalogs

Problem
Test all genes to investigate core, accessory genes and mobile elements
• Huge number of tests → low statistical power

Proposed solution
• Many genes from the same microbe are highly correlated
• Group co-abundant genes across samples (Nielsen et. al, 2014):
  → reconstitute the gene repertoire of microbial species
  → reduce the number of variables to test

Limitations
• Current tools group species core parts
• But miss accessory genes or assign them to small separated clusters
Achievements

• Reconstitution of Metagenomic Species Pan-genomes (MSPs) by binning co-abundant genes across metagenomic samples
• Classification of genes as core, accessory or shared.
• Use of a new robust measure to detect relations between core and accessory genes.
Scientific rationale
<table>
<thead>
<tr>
<th></th>
<th>sample 1</th>
<th>sample 2</th>
<th>sample 3</th>
<th>sample 4</th>
<th>sample 5</th>
<th>sample 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
<tr>
<td>Shotgun sequencing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mapping</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Counting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene Type</th>
<th>sample 1</th>
<th>sample 2</th>
<th>sample 3</th>
<th>sample 4</th>
<th>sample 5</th>
<th>sample 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core gene 1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Core gene 2</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Core gene 3</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Accessory gene 1</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Accessory gene 2</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

**Co-abundant genes**

**Partially co-abundant with the core genes**

Shotgun sequencing → Mapping → Counting
Overview of the Method
Workflow – Step 1

Reconstitution of clusters of co-abundant and co-occurring genes = seeds

black point = gene | grey circle = seed
Workflow – Step 2

Detection of associated seeds

black point = gene | circle = seed | arrow = associated seed
Election of core seeds
→ Among associated clusters, the largest is the core

black point = gene | circle = seed | arrow = associated seed
Worflow – Step 4
Reconstitution of Metagenomic Species Pangenomes
Detection and classification of genes associated to the core seeds.

Core genes

Shared module

Accessory module

MSP 1

MSP 2

MSP 3
Core gene

Gene detected in all the 701 samples where the MSP core is detected
Accessory gene

Gene detected in 88.8% of samples where the MSP core is detected

Gene missing in 11.2% of samples where the MSP core is detected
Shared core gene

- Gene present in 286 samples where the MSP core is not detected
- Gene detected in all the samples where the MSP core is detected
Shared accessory gene

Gene present in 28 samples where the MSP core is not detected

Gene present in 33.5% of samples where the MSP core is detected

Gene missing in 66.5% of samples where the MSP core is detected
Implementation

• MSPminer: multithreaded program implemented in C++

• Process large datasets (millions of genes, hundreds of samples) in a few hours on a single node server.

• Map/Reduce strategy to avoid brute-force comparison of all pairs of genes.

• Pre-clustering: genes with greatest count in the same sample are grouped at first

• more samples $\rightarrow$ faster clustering
Detection of co-abundant genes
Data transformation

- Asymmetric distribution of gene counts
  → High variability of species abundance + sequencing depth
- \( \log_{10} \) and \( \sqrt{\text{compress}} \) the right tail of the distribution
Data transformation

- heteroscedasticity with raw counts → variance grows faster than the mean
- log10: performs poorly with low counts
- sqrt: variance-stabilizing transformation
Relation between co-abundant genes

• Assessment of direct proportionnality:
  \[ g_2 = \alpha \cdot g_1 \text{ with } \alpha > 0 \]

• \( \alpha \) expected to be equal to the ratio of the genes length

\[
\frac{g_2}{g_1} = \frac{2}{1} \quad \Rightarrow \quad \frac{l_{g_1}}{l_{g_2}} = \frac{1}{3}
\]
Coefficient of proportionality

- Estimation of $\alpha$ directly from gene counts
- Deals with:
  - non uniform coverage
  - multiple-copy genes
- Discard samples missing in one or both genes
- Discard false positives
- Median to tolerate some outliers
- $\sqrt{\alpha} = \text{median} \left( \frac{\sqrt{g_2}}{g_1} \right)$
Measure of proportionnality

• Based on the Lin’s concordance correlation coefficient

• Calculated only on samples where both genes are detected

• Detect and discard outliers

Concordance with outliers = 0.13
Concordance without outliers = 0.93
Results
Application to a real large-scale metagenomic dataset

- Integrated Gene Catalog of the human gut microbiome: 9.9M genes x 1267 samples
- 7M genes with enough signal
- 3.3M genes organized in 1677 MSPs with > 200 core genes
- Majority of small MSPs: median 1784 genes
- 53 MSPs with > 5K genes
- Few prevalent MSPs: 93 (5.5%) detected in > 50% of samples
- 1257 (75%) detected in less than 10% of samples
Validation by census of Conserved Single Copy Genes (CSCG)

- Orthologous gene families: 161 in Archaea, 139 in Bacteria (Rinke et al. 2013)
- 829 MSPs have 80% of them
- 86% of CSCGs classified as core genes
- 87% of non-core CSCGs are highly prevalent accessory (prevalence > 90%)
Validation by taxonomic annotation

• Highly consistent taxonomical annotation for core genes (median 97%)

• Lower consistency for accessory genes (median 62%) mainly because of genes not annotated.
Accessory genes in MSPs

- MSP size correlated with the number of accessory genes
  Pearson's $r = 0.8$
- 4 outliers: unicellular eukaryotes (Blastocystis sp.)
- Many rare and prevalent accessory genes
- Few with intermediate prevalence
  → Importance of large cohorts for the exhaustiveness of the MSPs
Comparison to the Canopy clustering algorithm

- **Best CAG**: Green
- **Second best CAG**: Pink
- **Accessory genes not in CAGs**: Orange
- **Other CAGs**: Blue

---

<table>
<thead>
<tr>
<th>Genus</th>
<th>Accessory genes not in CAGs</th>
<th>Other CAGs</th>
<th>Second best CAG</th>
<th>Best CAG</th>
<th>Not in CAGs</th>
<th>In CAGs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteroides cellulositycus</td>
<td>msp_0003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parabacteroides distasonis</td>
<td>msp_0011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ruminococcus bicirculans</td>
<td>msp_0012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacteroides intestinalis</td>
<td>msp_0007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacteroides plebeius</td>
<td>msp_0015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roseburia faecis</td>
<td>msp_0014</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacteroides thetaiotaomicron</td>
<td>mss_0010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roseburia intestinalis</td>
<td>msp_0016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paraprevotella clara</td>
<td>msp_0024</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>msp_0005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Focus on Parabacteroides distasonis (msp_0011)
Applications and perspectives

• Reconstitution and reconstitution of microbial species pan-genomes.
• Better quantification of species by automatic selection of the best representative core genes.
• Access to accessory genes → strain-level analysis.
• Highlights potential inconsistencies in the taxonomy:
  • distant strains assigned to the same species (e.g. Faecalibacterium prausnitzii: 5 MSPs, Bacteroides fragilis: 2 MSPs)
  • close strains annotated as different species (e.g. Bifidobacterium saeculare and Bifidobacterium gallinarum)
For more information

• Pre-print available on bioRxiv

New Results

Abundance-based reconstitution of microbial pan-genomes from whole-metagenome shotgun sequencing data

Florian Plaza Oñate, Alessandra C. L. Cervino, Frédéric Magoulès, S. Dusko Ehrlich, Matthieu Pichaud

doi: https://doi.org/10.1101/173203
Acknowledgments

S. Dusko Ehrlich, INRA MetaGenoPolis

Frédéric Magoules, MAS Laboratory, CentraleSupélec

Matthieu Pichaud, Alessandra Cervino, Enterome