Simulation and Analysis of Metagenomic Data

Jacob Moran-Gilad, MD MPH

School of Public Health, Faculty of Health Sciences, Ben-Gurion University, Beer-Sheva, Israel
Metagenomics vs. WGS

- Metagenomics: Targeted Gene Sequencing vs. Shotgun Metagenome Sequencing.
  - Targeted Gene Sequencing: Known genes (identity, diversity, distribution, dynamics).
  - Shotgun Metagenome Sequencing: ‘all’ genes (functional potential, metabolic reconstruction).

- Whole Genome Sequencing (WGS): Single-Cell Genome Sequencing
  - Genomes: Linking phylogeny & function, genome diversification.
Downstream analysis - WGS

Overview *de novo* sequencing of individual genome

Short / Long reads

Contig assembly

Scaffold assembly

Finished genome

Reference Genome Sequence

35 bp identified

330 - 430 bp unknown sequence

35 bp identified
Downstream analysis – Metagenomics

16S rRNA gene
Amplification: 16S rRNA

Gut microbiota
Extract DNA

Metagenome
Amplification: Functional gene

Clustering for OTUs

OTU Classification

Alpha / Beta Diversity

Gene Prediction

Annotation

PhyloSift
eggNOG 4.0
Classification

Assembly

Whole genome sequencing

OTU1
OTU2
OTU3
OTU4

rdp
silva

Fungene
NCBI
WGS as ‘One-Stop Shop’

**Traditional microlab**
- Prep: Sample processing, Culture or direct detection
- ID: Determine genus/species, Phenotype (e.g. serogroup)
- AST: Measure MIC or surrogate, PCR for specific genes
- Vir/Tox: Phenotypic assays / animal, PCR for specific genes
- Typing: DNA fingerprinting, Sanger seq – MLST

**WGS era**
- Prep: Sample processing, culture, Extraction, lib prep, seq
- ID: Genomospecies calling, Further taxonomy
- AST: Map resistome, infer, Detect new variants
- Vir/Tox: Map virulome, infer, Gene expression?
- Typing: Extract current schemes, Variant calling / extended MLST
Metagenomics vs. WGS

**Bacterial WGS**
- Prep
  - Sample processing, culture
  - Extraction, lib prep, seq
- ID
  - Genomospecies calling
  - Further taxonomy
- AST
  - Map resistome, infer
  - Detect new variants
- Vir/Tox
  - Map virulome, infer
  - Gene expression?
- Typing
  - Extract current schemes
  - Variant calling / extended MLST

**Clinical metagenomics**
- Prep
  - Sample processing
  - Extraction, lib prep, seq
- ID
  - Genomospecies calling
  - Population analysis
- AST
  - Map resistome, infer
  - Detect new variants
- Vir/Tox
  - Map virulome, infer
  - Metatranscriptomics?
- Typing
  - Extract current schemes
  - Variant calling / extended MLST

**Unbiased; Cx independence**
- All pathogens
- All microbes
- The host

**Patho-typing & Epi-typing but challenging assignment**

**Cost ↑**
- TAT (?)

**Tools ↓**
Metagenomics – Syndromic Dx
Challenges for Clinical Metagenomics

- Streamlining sample prep & sequencing
- Bioinformatics tools for analysis – require optimisation and standardisation
- Establish clinical correlates for metagenomic signals
- Automation of analysis – produce comprehensible and actionable lab reports
- Generate and analyse hq MAGs
- Resolving closely related species / strains
- A validation nightmare
- Need for datasets for tool development and validation – scarcity, cost, uncertainty
Usefulness of metagenomic sims

• Development of new software / pipelines
• Benchmarking of software
• Road testing infrastructure
• Produce datasets for training algorithms
• Training of personnel
• Proficiency testing / EQA
• Emergency preparedness
• Accreditation
Virtual Sample Generator user interface

<table>
<thead>
<tr>
<th>Sample name</th>
<th>Sample ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matrix1 (e.g. Stool sample)</td>
<td>100000</td>
</tr>
<tr>
<td>Pathogen1 (e.g. E. coli)</td>
<td>1000</td>
</tr>
<tr>
<td>Pathogen2 (e.g. E. faecium)</td>
<td>10</td>
</tr>
</tbody>
</table>

NGS raw sequence reads database
Genome assembly sequence database
Source files

Virtual Sample Generator module

Metagenomics software / pipeline
Analysis
### Table 1
List of scenarios used for generating virtual metagenomics samples

<table>
<thead>
<tr>
<th>Scenario name</th>
<th>Scenario description/purpose</th>
<th>Scenario composition (M = matrix, P = pathogen)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respiratory (sputum) sample</td>
<td>Rare occurrence of Legionnaires' disease, co-infecting strains</td>
<td>M1: Normal respiratory sample (human)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P1: <em>Legionella pneumophila</em> serogroup 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P2: <em>Legionella pneumophila</em> serogroup 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P3: Human adenovirus</td>
</tr>
<tr>
<td>Polymicrobial bacteraemia</td>
<td>Blood sample, two different ‘model’ resistant bacteria</td>
<td>M1: Normal human blood</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P1: <em>Acinetobacter baumanitii</em> (multidrug resistant)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P2: Vancomycin-resistant <em>Enterococcus faecium</em></td>
</tr>
<tr>
<td>Carriage of resistant pathogen</td>
<td>Faecal sample, representing low-level carriage)</td>
<td>M1: Normal human stool</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P1: <em>Klebsiella pneumoniae</em> ST258 KPC-producer</td>
</tr>
<tr>
<td>Genital tract infection</td>
<td>Vaginal sample, sexually transmitted pathogen</td>
<td>M1: Normal vaginal microbiota</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P1: <em>Neisseria gonorrhoea</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P2: <em>Chlamydia trachomatis</em></td>
</tr>
<tr>
<td>Poliovirus surveillance</td>
<td>Environmental sample (sewage), low level Poliovirus</td>
<td>M1: Different sewage samples (untreated)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P1: Poliovirus type 1 (Enterovirus C Mahoney)</td>
</tr>
<tr>
<td>Contaminated food</td>
<td>Sensitive food (chicken salad) contaminated by different foodborne pathogens</td>
<td>M1: Normal poultry</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M2: Tomato</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3: Lettuce</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P1: <em>Campylobacter jejuni</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P2: <em>Salmonella enterica</em></td>
</tr>
<tr>
<td>Ebola virus disease, post mortem</td>
<td>Post mortem sample (human blood), sensitivity analysis for detection of low-level virus</td>
<td>M1: Normal human blood</td>
</tr>
<tr>
<td>diagnosis</td>
<td></td>
<td>P1: Ebola virus Zaire</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M1: Drinking water</td>
</tr>
<tr>
<td>Intentional contamination of water</td>
<td>Water, select agent (<em>B. pseudomallei</em>) and competing non-fermenters</td>
<td>P1: <em>Burkholderia pseudomallei</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P2: <em>Pseudomonas aeruginosa</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P3: <em>Burkholderia thailandensis</em></td>
</tr>
<tr>
<td>Deliberate release of multiple agents</td>
<td>Blood sample, two select agents and a contaminant</td>
<td>M1: Normal human blood</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P1: <em>Bacillus anthracis</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P2: <em>Bacillus cereus</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P3: <em>Francisella tularensis</em></td>
</tr>
<tr>
<td>Vector carrying pathogen</td>
<td>Insect vector (mosquito), flavivirus</td>
<td>M1: <em>Aedes albopictus</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P1: West Nile virus</td>
</tr>
</tbody>
</table>
| Contaminated food | Sensitive food (chicken salad) contaminated by different foodborne pathogens | M1: Normal poultry  
M2: Tomato  
M3: Lettuce  
P1: Campylobacter jejuni  
P2: Salmonella enterica |
Metagenomics pipeline

Scenario Sample List

M3S3 simulator

Simulated Sample data

Customized DB

Metagenome analysis (species/strains ID)

MAGs isolation pipeline

AMR analysis

Metagenome analysis report

MAG recovery report

AMR analysis results report
Summary

- Sample simulation is useful for design and evaluation of bioinformatics solution for challenging scenarios.
- Simulation may be a cost-effective approach to introduce metagenomics into clinical labs (quality, regulatory, etc).
- The in house metagenomics pipeline identified the target select agent in 80% of samples.
- Matrix:pathogen ratio up to 10,000:1.
- Successful reconstruction of MAGs, to the level of discerning virulent vs. avirulent strains (against dB).
- hq MAGs achieved at M:P of 200:1 or lower.
- Improvement of MAG reconstruction is central to ensure feasibility and future integration.
- Extensive validation required to operationalise.
THANK YOU

Prof. Jacob Moran-Gilad
giladko@post.bgu.ac.il

Acknowledgements

Dr. Yair Motro, Bioinfo lead
MAGICAL Group, BGU