Towards understanding the functional and taxonomic repertoire of a metagenome

Rob Finn
JOBIM, Clermont-Ferrand,
8th July 2015
Metagenomics: a broad range of applications
Metagenomics: a broad range of applications

ARTICLE

Structure, function and diversity of the healthy human microbiome

The Human Microbiome Project Consortium

Targeted Restoration of the Intestinal Microbiota with a Simple, Defined Bacteriotherapy Resolves Relapsing Clostridium difficile Disease in Mice

Trevor D. Lawley1, Simon Clare1, Alan W. Walker1, Mark D. Stares1, Thomas R. Connor1, Claire Raisen1, David Goulding1, Roland Rad1, Fernanda Schreiber1, Cordelia Brandt1, Laura J. Deakin1, Derek J. Pickard1, Sylvia H. Duncan2, Harry J. Flint2, Taane G. Clark2, Julian Parkhill1, Gordon Dougan1

LETTER

Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease

Novel Gut-Based Pharmacology of Metformin in Patients with Type 2 Diabetes Mellitus

Antonella Napolitano2, Sam Miller3, Andrew W. Nicholls3, David Baker3, Stephanie Van Horn3, Elizabeth Thomas4, Deepak Rajpal4, Aaron Spivak5, James R. Brown6, Derek J. Nunez2

1 Immune Inflammation Unit, GSK R&D; 2 Stevanovic, Nis, Serbia; 3 Quantitative Sciences, GSK R&D; 4 Target and Pathways Validation, GSK R&D; 5 Upper Providence, Pennsylvania, United States of America; 6 Computational Biology, GSK R&D; 7 Research Triangle Park, North Carolina, United States of America.
From environment to DNA sequence

Sample

Size Fractioning

mRNA extraction
Metatranscriptome

DNA extraction
Metagenome

PCR
Amplicon

Sequencing

Analysis?
From environment to DNA sequence

Sample

Size Fractioning

Big Data

Metatranscriptome

Metagenome

Amplicon

Sequencing

Analysis?
Taxonomic Profiling

‘Classical’ approach:
- Identify a marker gene that differs between species
- Survey the marker gene in a given sample
- Compare to reference database
- Resolve the community composition
16s rRNA amplification

‘Classical’ Gene - 16S rRNA:

• prokaryotic community analysis
• Contains highly conserved primer binding sites and species-specific hypervariable regions

Even so, 16S drawbacks include:
+ low resolving power at species level
+ poor discriminatory power for some genera
+ copy number variation
Hiding in Plain Sight

Researchers using metagenomics and single-cell sequencing identify a potential new bacterial phylum.

By Kim Smuga-Otto | March 31, 2015

Studies on 16s ribosomal RNA (rRNA) sequences have opened scientists’ eyes to the complexity of microbial communities, but some bacteria evade detection. At the US Department of Energy (DOE) Joint Genome Institute User Meeting held in Walnut Creek, California, last week, researchers announced the genomic identification of a potential new bacterial phylum, Candidatus Kryptonia, based on their study of samples isolated from four hot springs located in North America and Asia. Altogether, the DOE team sequenced 22 Kryptonia genomes.

said microbial ecologist Jack Gilbert of Argonne National Laboratory was not involved with the study, “but, genomics-wise, I think we have a handle on it.”

Genomic analyses place Kryptonia in the Bacteroidetes and in marine environments. If confirmed, Kryptonia will join the SAR11 clade. Kryptonia appears to have acquired this characteristic from Archaea.

Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton

Amy Apprill1,*, Sean McNally1,2, Rachel Parsons2, Laura Weber1
The EBI metagenomics portal

- Free for all to use

Submission of sequence data for archiving and analysis

Powerful analysis using selected EBI and external software tools

Visualisation

Data analysis and visualization through web interface

- Secure archiving, intensive analysis and public display of results
Metagenomics - Big Data

- Speed is really important!
  - Submitted nucleotide sequences: **62 billion (62,595,667,453)**
  - Average length per sequence: **120 nt**
  - Number of different runs: **4,876**
  - Number of samples: **4,145**
  - Studies: **183 (115 public, 68 private)**
Metagenomics - Big Data

• Speed is really important!
  • Submitted nucleotide sequences: **62 billion (62,595,667,453)**
  • Average length per sequence: **120 nt**
  • Number of different runs: **4,876**
  • Number of samples: **4,145**
  • Studies: **183 (115 public, 68 private)**

• MG-RAST
  • Submitted nucleotide sequences: **93 billion**
  • Number of samples: **27,310 (19,983 amplicon, 6,059 WGS)**
  • Studies: **893**
Submitting to EBI Metagenomics

• EBI Metagenomics want to encourage people to supply as much detailed **metadata** as possible, but with the lowest possible overhead

  who  where, when, what  how

name  institute  country  ...  contact
latitude and longitude, depth, salinity, temperature, time ...
quantity, conservation process, storage conditions, treatments, extraction methods ...
platform, protocol, filtering and QC, analysis, tool versions ...

• Development of web-based tools: **ENA Webin**
• Use of templates and check-lists (MIGS/MIXS standards)
• Tutorial and direct support
Overview: EMG Portal analysis

• Provide robust sequence analysis services to all metagenomic researchers

○ Understand species diversity and functional potential of a community

Data analysis using selected EBI and external software tools
Metagenomics assembly?

- Metagenomics: Not clear how you avoid assembling sequences from different species together: chimaera

- No reference sequence to align against
raw reads $\rightarrow$ ENA

**QC**

discarded reads

processed reads

**rRNAselector**

reads with rRNA

reads without rRNA

FragGeneScan

predicted CDS

**InterProScan**

Unknown function

**pCDS**

**Taxonomic analysis**

**Function assignment**

Amplicon-based data

**Qiime**

**Amplicon**

**discarded reads**

**reads**

**with rRNA**

**reads without rRNA**

**FragGeneScan**

**predicted CDS**

**InterProScan**

**Unknown function**

**pCDS**

**Taxonomic analysis**

**Function assignment**
EMG portal does *NOT* use BLAST based homology methods

Instead reads are compared to models (signatures) generated from multi-sequences alignments:

- more divergent annotations, hence meaningful
- faster annotation

- rRNASelector identify 5, 16 and 28s rRNA (profile HMM models) => 16s-based Qiime taxonomy annotations

- FragGenScan predict CDSs (HMM models) => InterProScan functional annotations (profiles and models)
Growth of informatics resources

- Most grow exponentially
- Protein families do NOT!
- Great for scalability
Profile Hidden Markov Models

<table>
<thead>
<tr>
<th>seq1</th>
<th>ACG-LD</th>
<th>Consensus columns assigned, Defining inserts and deletes:</th>
</tr>
</thead>
<tbody>
<tr>
<td>seq2</td>
<td>SCG--E</td>
<td></td>
</tr>
<tr>
<td>Seq3</td>
<td>NCGgFD</td>
<td></td>
</tr>
<tr>
<td>Seq4</td>
<td>TCG-WQ</td>
<td></td>
</tr>
<tr>
<td></td>
<td>123-45</td>
<td></td>
</tr>
</tbody>
</table>
Profile Hidden Markov Models

<table>
<thead>
<tr>
<th>seq1</th>
<th>seq2</th>
<th>seq3</th>
<th>seq4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACG-LD</td>
<td>SCG--E</td>
<td>NCGgFD</td>
<td>TCG-WQ</td>
</tr>
<tr>
<td>123-45</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Input multiple alignment:

<table>
<thead>
<tr>
<th>seq1</th>
<th>ACG-LD</th>
</tr>
</thead>
<tbody>
<tr>
<td>seq2</td>
<td>SCG--E</td>
</tr>
<tr>
<td>Seq3</td>
<td>NCGgFD</td>
</tr>
<tr>
<td>Seq4</td>
<td>TCG-WQ</td>
</tr>
<tr>
<td></td>
<td>123-45</td>
</tr>
</tbody>
</table>

Consensus columns assigned, Defining inserts and deletes:
Profile Hidden Markov Models

Input multiple alignment:
- seq1: ACG-LD
- seq2: SCG--E
- Seq3: NCGgFD
- Seq4: TCG-WQ
- 123-45

Consensus columns assigned, Defining inserts and deletes:

Plan7 core model
Profile Hidden Markov Models

Input multiple alignment:

- seq1: ACG-LD
- seq2: SCG--E
- Seq3: NCGgFD
- Seq4: TCG-WQ
- 123-45

Consensus columns assigned, Defining inserts and deletes:

Plan7 core model
EBI Metagenomics: functional analysis

reads without rRNA → FragGeneScan → predicted CDS → InterProScan → Function assignment

Unknown function pCDS
InterPro

- Hidden Markov Models
- Fingerprints
- Profiles
- Patterns

- Structural domains
- Functional annotation of families/domains
- Protein features (sites)
InterPro in the Metagenomics Portal
Annotations without assembly

Metagenomic Discovery of Biomass-Degrading Genes and Genomes from *Cow Rumen*

Comparison of the normalised number of genes / reads corresponding to CAZy Glycoside Hydrolase Family from the Hess et al paper and from the EMG pipeline.

**Hess et al**: genome assembly then gene prediction using a subset of Pfam.

**EMG pipeline**: no assembly and gene prediction using InterPro.

Discrepancies are due to the different ways in which significance cut-off are calculated.
Coverage of sequences independent of biome or technology

- But, there is huge variability between samples
Taxonomic analysis results

Switch to bar chart, column or Krona interactive views

Export charts
Functional analysis results

InterPro match summary
Most frequently found InterPro matches to this sample:
Export

InterPro matches summary (Total: 5452)

Export charts

Switch to bar chart, view

Links to InterPro website

Export

GO Terms annotation
A summary of Gene Ontology (GO) terms derived from InterPro matches to your sample is provided in the charts below.
Switch view: [ ] [ ] [ ] [ ]

Biological process

Molecular function

Cellular component

EMBL-EBI
**Downstream analysis: download options**

You can download in this section the full set of analysis results files and the original raw sequence reads.

### Sequence data
- Submitted nucleotide reads (ENA website)
- Processed nucleotide reads (FASTA) - 2 MB
- Processed reads with pCDS (FASTA) - 2 MB
- Processed reads with InterPro matches (FASTA) - 1 MB
- Processed reads without InterPro match (FASTA) - 836 KB
- Predicted CDS (FASTA) - 710 KB
- Predicted CDS with InterPro matches (FASTA) - 451 KB

### Functional Analysis
- InterPro matches (TSV) - 1 MB
- Complete GO annotation (CSV) - 44 KB
- GO slim annotation (CSV) - 7 KB

### Taxonomic Analysis
- Reads encoding 5S rRNA (FASTA) - 565 bytes
- Reads encoding 16S rRNA (FASTA) - 21 KB
- Reads encoding 23S rRNA (FASTA) - 37 KB
- OTUs and taxonomic assignments (BIOM) - 6 KB
- Phylogenetic tree (Newick format) - 289 bytes
- OTUs and taxonomic assignments (TSV) - 2 KB

relatively small result files: can be used for downstream analysis with other tools
Sample Comparisons

- [https://www.ebi.ac.uk/metagenomics/compare](https://www.ebi.ac.uk/metagenomics/compare)
Sample Comparisons

- https://wwwdev.ebi.ac.uk/metagenomics/compare
Sample Comparisons

- https://wwwdev.ebi.ac.uk/metagenomics/compare

- Phosphorylation
- Photosynthesis
- Protein metabolic process

Sample list (click to hide)
- SRS001909
- SRS001913

Biological process
- ATP metabolic process
- Biosynthetic process
- Carbohydrate metabolic process
- Catabolic process
- Cell communication
- Cell division
- Cell growth
- Cell morphogenesis
- Cell motility
- Cell projection assembly
- Cell wall organization or biosynthesis
- Cellular amino acid metabolic process
- Cellular component organization
- Cellular respiration
- Chemotaxis
- Cofactor metabolic process
- Composition

Cellular component
- ATP-binding cassette (ABC) transporter complex subunit
- Cell wall
- Chromosome
- Cytoplasm
- Two-component signal transduction system
- Viral reproduction
Comparison of two Marine Biomes - Taxonomic distributions

25m Depth

500m Depth

Larger archaeal proportion

Cyanobacteria

EMBL-EBI
Two of the many challenges......

1. Volume of data/scalability

2. Linking function and taxonomy
EBI Metagenomics Portal

- Two flagship marine projects
  - Ocean sampling day
    - 150 oceanographic metagenomic samples
    - Sample on the Solstice (June 21st) 2014
    - Identical DNA preparation and sequencing methods
      - 120GB sequence data, ~800MB per sample
  - TaraOceans
    - Significantly bigger, 35K samples, 210 stations
    - Focusing on the Global Ocean prokaryotic set (250 samples), size fractionated
      - 10TB sequence data, ~40GB per sample
      - 106/250 samples completed
Geographic comparison

<table>
<thead>
<tr>
<th></th>
<th>OSD</th>
<th>TaraOceans (to date)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processed nucleotide sequences</td>
<td>220 million</td>
<td>7,329 million</td>
</tr>
<tr>
<td>Predicted CDS</td>
<td>181 million</td>
<td>6,090 million</td>
</tr>
<tr>
<td>CDS with InterPro match</td>
<td>68 million</td>
<td>2,504 million</td>
</tr>
</tbody>
</table>
Unipept: an alternative approach

- Developed by Peter Dawyndt at Ghent University
- Tryptic peptide-based biodiversity approach for proteomic/metaproteomic data
- *In silico* trypsin digest of UniProtKB, providing searchable database of tryptic peptides linked to species

Allows simultaneous analysis of eukaryotic, prokaryotic and viral communities
Unipept analysis: OSD data
OSD results: Prochlorococcus distribution

Samples with Prochlorococcus counts >10k

Prochlorococcus: very small marine cyanobacteria with an unusual pigmentation. These bacteria belong to the photosynthetic picoplankton and are probably the most abundant photosynthetic organism on Earth…
OSD results: bacillariophyta

ERR771080 and ERR771082: 58k and 27k *Thalassiosira pseudonanana* counts

*Thalassiosira pseudonanana*: a species of marine centric diatom. It was chosen as the first eukaryotic marine phytoplankton for whole genome sequencing...
OSD results: chordata

ERR771053, ERR771054 & ERR771055
7.5k, 5.5k and 2.7k *Oikopleura dioica* counts

*Oikopleura dioica*: a species of small pelagic tunicate found in the surface waters of most of the world's oceans...
95% of all viruses in the OSD samples are:

- **Caudovirales** (tailed bacteriophages)

- **Phycodnaviridae** (dsDNA viruses infecting marine or freshwater eukaryotic algae)
OSD results: phycodnaviridae distribution

Samples with phycodnaviridae counts >1k
OSD results: chlorophyta distribution

Samples with chlorophyta (green algae) counts >20k
Conclusions

• EBI metagenomics provides a free, open access analysis platform

• Provides a standard, scalable analysis, the results of which form the basis for further analysis

• Metadata provides data longevity and increases interpretation

• Increasing need not to focus just prokaryotes

• Many more challenges on scale, reproducibility and sequencing technologies
Acknowledgements

The EMG team:

EMBL-EBI
Alex Mitchell
Hubert Denise
Matthew Fraser
Gift Nuka
Sebastien Pesseat
Maxim Scheremetjew

Guy Cochrane
Rasko Leinonen
Petra Ten Hoopen

Past Members
Dawn Field
Peter Sterk


Harvard
Sean Eddy

NCBI
Eric Nawrocki

WTSI
Trevor Lawley
Sam Forster

University Tromso
Nils-Peder Willassen

EBI metagenomics - a new resource for the analysis and archiving of metagenomic data
Hunter et al, NAR, 2014 42:D600-D606