MG Rast workshop
February 7, 2013

February 6-8th, 2013: NBIC Advanced NGS course: Metagenomics approaches and data analysis

**MG Rast workshop outline**

**General Introduction**
- How to find public metagenomes in MG-RAST

**Single metagenome**
- Assess the quality of the data
- Extract species information from MG-RAST based of shotgun reads (non-16s)
- Extract functional annotation

**Multiple metagenomes**
- Compare metagenomes using the analysis toolkit of MG-RAST
- Interact with MG-RAST using R

**Introduction**

In this workshop we will a short overview of the MG-RAST metagenomics annotation pipeline. The pipeline provides insight in quantitative microbial population information and functional annotation of metagenomics shotgun sequence data.

We will NOT discuss how to upload your own data, however links will be given to support pages and videos.

The workshop concentrates on the data from a recent paper on soil metagenomics: Cross-biome metagenomic analyses of soil microbial communities and their functional attributes (Fierer et al., 2012, PNAS vol 109,21390-21395). We will try to partially reconstruct some conclusions of this paper.

How to find public metagenomes in MG-RAST

Please navigate to the following link:

http://metagenomics.anl.gov

Public metagenomes can be found without logging in into MG-RAST. The main screen contains access to the search interface using the magnifier glass in the right top corner:

The search box contains a simple search box with a link to the advanced search box. Simple searches are simple text query applied against multiple database fields. With advanced searches you can specify complex boolean queries combining multiple database fields. Both types of queries return a list of samples that match the query criteria.

Use this box if you know the MG-RAST name, biome or project name. Please note that you have to select a Match category.

Another option is to search using more specific functions on the ‘simple’ page:

Here you can enter terms from the metadata, like Fish gut, Human Microbiome Project (HMP), Twins etc. using the “Advanced Search” lets you search for combinations of parameters.

Note: these combined searches may take long.
Task 1: Find Metagenomes

a. Play around with the simple and advanced search tools.

For the next section on quality control we will use one of the metagenomes below, try to find them.

Table 1, A selection of metagenomes to be used in this tutorial.

<table>
<thead>
<tr>
<th></th>
<th>Metagenome</th>
<th>Biome</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Wright Valley, Antarctica</td>
<td>Deserts and xeric shrubland biome</td>
<td>Cold</td>
</tr>
<tr>
<td>2</td>
<td>Lake Hoare Valley, Antarctica</td>
<td>Deserts and xeric shrubland biome</td>
<td>Cold</td>
</tr>
<tr>
<td>3</td>
<td>Chihuahuan Desert, Sevilleta LTER, New Mexico, USA</td>
<td>Deserts and xeric shrubland biome</td>
<td>Hot</td>
</tr>
<tr>
<td>4</td>
<td>Misiones, Argentina</td>
<td>Tropical and subtropical moist broadleaf forest biome</td>
<td>Hot</td>
</tr>
<tr>
<td>5</td>
<td>Duke Forest, North Carolina, USA</td>
<td>Temperate coniferous forest biome</td>
<td>Temperate</td>
</tr>
<tr>
<td>6</td>
<td>Konza Prairie LTER, Kansas, USA</td>
<td>Temperate grasslands, savannas, and shrubland biome</td>
<td>Temperate</td>
</tr>
</tbody>
</table>

b. Note the project name to which these metagenomes belong (you will need it later on).

c. How many metagenome datasets belong to this project?

d. Note the name of the metagenomes above.

e. (Try to find the Guerrero Negro Metagenomes)

Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat.


Microbial Ecology Program, DOE Joint Genome Institute, Walnut Creek, CA, USA.

Abstract

To investigate the extent of genetic stratification in structured microbial communities, we compared the metagenomes of 10 successive layers of a phylogenetically complex hypersaline mat from Guerrero Negro, Mexico. We found pronounced millimeter-scale genetic gradients that were consistent with the physicochemical profile of the mat. Despite these gradients, all layers displayed near-identical and acid-shifted isoelectric point profiles due to a molecular convergence of amino-acid usage, indicating that hypersalinity enforces an overriding selective pressure on the mat community.
Assess the quality of the data

The MG-RAST pipeline has extensive capabilities to quality control your input data. It is recommended to submit raw data in Fasta, Fastq (or SFF, but that will be converted to FastQ anyway).

The input processing steps consist of quality filtering, demultiplexing (split on barcodes), length filtering, dereplication, and removal of model organism sequences. MG-RAST works with reads longer than 75bp. If you have shorter reads like 50bp Illumina reads it is possible to submit assembled data, however that may skew analysis results. Assembly will be added in a later version of MG-RAST.

Your samples may contain a large number of near identical reads. A filter is available to remove all but a single representative of clusters of reads whose first 50 base pairs are identical. It is recommended to use the MG_RAST dereplication steps.

If your samples are host associated, i.e. a mouse gut sample, a screening against the model organism’s genome using Bowtie is recommended. Several model organism genomes are available in MG-RAST.

MG_RAST includes DRISEE, a duplicate read inferred sequencing error estimation pipeline. It provides a novel measure for sequencing error for whole genome shotgun metagenomic sequence data using ADR’s (artifactual/artificial duplicate reads) to generate internal sequence standards from which an overall assessment of sequencing error in a sample is derived. DRISEE values are reported as % error. However, this number only reflects the quality of the raw data. The idea with DRISEE is to produce an assessment of error that is independent of all our other QC procedures. At present, DRISEE results are not used to cull or trim reads in any way.

It is possible for DRISEE error values to be high while other estimates of sequencing error (e.g. Phred) are low.

As an example, this can occur if whole genome amplification was used to amplify template DNA, creating a number of imperfect copies of the original template. DNA produced from a procedure of this kind could be high quality, and would cause no observable problems that Phred could detect (Phred tells us how well the sequencer thinks it did calling a particular base -- it has no way to know if the base represents an imperfect amplification).

DRISEE values can also be high if there are populations of highly repetitive sequences - inclusion of adapters, primer binding sites etc. -- the sorts of artifacts that sequencers should (but don't always) remove.
Examples:

<table>
<thead>
<tr>
<th>Good</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
<td><strong>SRR061345, Human Microbiome Project (HMP)</strong></td>
</tr>
<tr>
<td><strong>Sequencing QC</strong></td>
<td></td>
</tr>
<tr>
<td><img src="example1.png" alt="Sequencing QC Diagram" /></td>
<td><img src="example2.png" alt="" /></td>
</tr>
<tr>
<td><strong>DRISSEE error</strong></td>
<td></td>
</tr>
<tr>
<td><img src="example3.png" alt="DRISSEE Error Diagram" /></td>
<td><img src="example4.png" alt="" /></td>
</tr>
<tr>
<td><strong>Cumulative plot</strong></td>
<td></td>
</tr>
<tr>
<td><img src="example5.png" alt="Cumulative Plot Diagram" /></td>
<td><img src="example6.png" alt="" /></td>
</tr>
<tr>
<td><strong>Kmer profiles</strong></td>
<td></td>
</tr>
<tr>
<td><img src="example7.png" alt="Kmer Profiles Diagram" /></td>
<td><img src="example8.png" alt="" /></td>
</tr>
<tr>
<td><strong>Nucleotide Position Histogram</strong></td>
<td></td>
</tr>
<tr>
<td><img src="example9.png" alt="Nucleotide Position Histogram Diagram" /></td>
<td><img src="example10.png" alt="Nucleotide Position Histogram Diagram" /></td>
</tr>
<tr>
<td><strong>Sequence length distribution before and after QC</strong></td>
<td><strong>Fixed 100 bp reads, no histogram available</strong></td>
</tr>
<tr>
<td><img src="example11.png" alt="Sequence Length Distribution Diagram" /></td>
<td><img src="example12.png" alt="Sequence Length Distribution Diagram" /></td>
</tr>
<tr>
<td><strong>Sequence GC distribution before and after QC</strong></td>
<td><strong>No raw GC histogram available</strong></td>
</tr>
<tr>
<td><img src="example13.png" alt="Sequence GC Distribution Diagram" /></td>
<td><img src="example14.png" alt="Sequence GC Distribution Diagram" /></td>
</tr>
</tbody>
</table>

**Name**: Duke Forest, North Carolina, USA

**Sequence QC**

- 5.5% Fasta QC
- 13.4% Unknown
- 43.4% Unknown Protein
- 20.2% Unannotated Protein
- 8% Reverse RNA

**DRISSEE error**

- Total DRISSEE Error = 36.24%
- Total DRISSEE Error = 43.53%

**Cumulative plot**

- Sequence length distribution
- Nucleotide position histogram

**Sequence length distribution before and after QC**

- Fixed 100 bp reads, no histogram available

**Sequence GC distribution before and after QC**

- No raw GC histogram available
Task 2

a. Assess the quality of the metagenomes from task 1.
b. Do the failed reads influence downstream analysis
c. How many reads can be annotated in the metagenomes of your choice?
d. In which metagenome do you find most functionally annotated predicted proteins and does this reflect your expectation?

Extract species information from MG-RAST based of shotgun reads (non-16s)

The Taxonomic Hits Distribution

The Taxonomic Hits Distribution pie charts on the metagenome page give a rough indication of what species are represented in the metagenome. Each slice indicates the percentage of reads with predicted proteins and ribosomal RNA genes annotated to the indicated taxonomic level. This information is based on all the annotation source databases used by MG-RAST. The slices are clickable to download the associated reads with annotation.

Rarefaction Curve & Alpha diversity index

The rarefaction curve plots annotated species richness. This curve is a plot of the total number of distinct species annotations as a function of the number of sequences sampled. On the left, a steep slope indicates that a large fraction of the species diversity remains to be discovered. If the curve becomes flatter to the right, a reasonable number of individuals is sampled: more intensive sampling is likely to yield only few additional species.

Sampling curves generally rise very quickly at first and then level off towards an asymptote as fewer new species are found per unit of individuals collected. These rarefaction curves are calculated from the table of species abundance. The curves represent the average number of different species annotations for subsamples of the complete dataset.

The above image shows the range of α-diversity values in project cross-site soil metagen. The min, max, and mean values are shown, with the standard deviation ranges (σ and 2σ) in different shades. The α-diversity of this metagenome is shown in red.

The α - diversity summarizes the diversity of organisms in a sample with a single number. The α diversity of annotated samples can be estimated from the distribution of the species-level annotations. Annotated species richness is the number of distinct species annotations in the combined MG-RAST dataset.

Task 3

a. Find two metagenome datasets within the project with a largely different taxonomy distribution at the domain, phylum, class and order level. Note their name
b. What do you note about the family and genus level slices. Do these assignments reflect true species diversity?
c. What can be concluded from the rarefaction curve and α-diversity index?
Extract functional annotation

The pie charts of the functional annotation distributions show functional categories for at the highest level supported by these functional hierarchies. Each slice indicates the percentage of reads with predicted protein functions annotated to the category for the given source. Sources are: KO (KEGG orthologous groups), COG (NCBI clusters of orthologous groups), NOG, and seed subsystems. The slices are clickable to download the associated reads with annotation.

Task 4

a. Download the SEED subsystem category counts from the two metagenomes chosen in task 3 and compare them in a spreadsheet.

b. What can you conclude from this data compared to the Taxonomic Hits Distribution (on the order level)

MG-RAST analysis tools

The analysis tools can be accessed from both the main page of MG-RAST and the individual metagenome pages. Clicking on the analysis icon opens the analysis page:
It is possible to analyse metagenomes within a project or across projects. In this part we will continue with comparing some of the metagenomes from the ‘cross-site soil metagenomics’ project.

**Compare metagenomes on the species level**
Using the Organism Abundance tab you can create barcharts, trees, tables, heatmaps, PCA plots and rarefaction curves for selected metagenome datasets.

**Task 5**
- a. Go to the analyses page and select three distinct metagenomes. (e.g. Antarctic desert, temperate forest and rain forest) Create a circular species tree, bar chart and table.
- b. Create a PCoA plot for the whole ‘cross site soil metagenome’ project. Do the Antarctic metagenomes group together? Do the temperate forest metagenomes group together.
- c. Create a species heatmap at the class level.

**Compare metagenomes on the functional annotation level**
Using the Organism Abundance tab you can create barcharts, trees, tables, heatmaps, PCoA plots and use the KEGG mapper for selected metagenome datasets.

**Task 6**
- a. Go to the analyses page and select the cross site soil metagenome’ project. ‘ Create a functional annotation heatmap using three of the 16 metagenomes (otherwise the process will take long.)
- b. Use the KEGG mapper to visualize the data of two available metagenomes. Unfortunately, the cross-site soil metagenomes are not available in this tool. Try to select Chicken Cecum A & B

**Interact with MG-RAST using R**
matR: Using R with MG-RAST [https://github.com/MG-RAST/matR](https://github.com/MG-RAST/matR)

**Other resources**
The extensive help pages of MG-RAST - [http://blog.metagenomics.anl.gov/howto/](http://blog.metagenomics.anl.gov/howto/)

The MG-RAST youtube channel - [http://www.youtube.com/channel/UCoBKJ-yYJu5HK1szy3my91A](http://www.youtube.com/channel/UCoBKJ-yYJu5HK1szy3my91A)

Titus Brown on metagenome assembly: [http://www.slideshare.net/c.titus.brown/2012-stampsmbi2](http://www.slideshare.net/c.titus.brown/2012-stampsmbi2)

Answers

Task 1

a. the metagenomes can be found on http://metagenomics.anl.gov/linkin.cgi?project=2997
b. The name of the project is “cross site metagenomes”, PI Noah Fierer
c. There are 16 metagenomes in this study
d. The metagenome names are:
   EB024_r1, EB021_r1, EB020_r1, EB019_r1, EB017_r1, DF1_r1, CL1_r1, BZ1_r1, AR3_r1,
   TL1_r1, SV1_r1, SF2_r1, PE6_r1, MD3_r1, KP1_r1, EB026_read1

Task 2

a. No, DRISEE is not used in the QC step of the reads, it is just an indication of the sequence
good quality. Only reads with good quality are propagated to the annotation steps.
More information can be found on the MG-RAST website

b. Most annotated reads are found in species rich metagenomes like those from temperate
forests and rain forests. The cold and hot desert metagenomes represent less species
probably due to the absence or low amounts of plant material (carbon sources)

Task 3

a. Metagenomes with different taxonomic distributions are for example EB024_r1 and
   DF1_r1
b. The family and genus levels contain many non-assigned reads i.e. the reads cannot be
   mapped to an isolated family but may reflect more common sequences. Even reads that
   can be assigned to a species may not actually be from that species as a result that from
   the fact that reference databases are biased towards culturable species.
   Although the rarefaction curve of the cross site soil metagenomes suggest that more
   species can be recovered with deeper sequencing this estimate may be too high.

Task 4

a. You should get a spread sheet with two columns with raw counts per subsystem.
b. For these metagenomes the distribution of counts for the SEED subsystems is highly
   comparable despite the differences in taxonomic distribution.
Task 5

a. Just try, you should be able to get a circular species tree and barchart, similar to:

![Circular species tree and barchart](image)

b. The antarctic desert metagenomes do not really group together, the temperate forests do. Try the different distance metrics, they influence the PCoA clustering.

c. Just try, you should be able to get a species heatmap similar to:

![Species heatmap](image)

Task 6

a. Just try, you should get a heatmap like this:

![Heatmap](image)

b. The KEGG mapper uses a subset of the metagenomes. Not all metagenomes are available.