Sequence Analysis '16 -- Lecture 23

Metagenomics

Genomic analysis of environmental DNA samples.
Metagenomics

• Introduction
  • The “extended phenotype”
  • 16s rDNA metagenomics
  • reference sequence alignment
  • Multi-genome microbial assembly w/o template
• Discovery metagenomics
• Comparative metagenomics
• Principle component analysis
Metagenomics --

is the analysis of sequences from environmental samples which may contain any number of microorganisms.

“Microbes run the world. It’s that simple. Although we cannot usually see them, microbes are essential for every part of human life—indeed all life on Earth. Every process in the biosphere is touched by the seemingly endless capacity”


“Everything is everywhere, the environment selects”

--The Baas-Becking hypothesis, in answer to a question about geographic locations of microbial species.
Importance of metagenomics

- **Agriculture**
  A community of microorganisms is required to recycle nutrients. Farmers want to know what is there.

- **Water pollution**
  Microorganism populations respond to nutrients, pollution.

- **Human health and nutrition**
  The microbial community of the gut, nose/throat, skin, vagina, are indicators of infected state, risk, and may be biomarkers of cancer.

- **Paleobiology**
  DNA from frozen mammoth, iceman reveal their diet, phylogeny.

- **Forensic science.**
  Barcoding. eDNA.

- **Biodiversity**
  Catalog it before it’s gone.
The rise of Metagenomics

Metagenomics: Application of Genomics to Uncultured Microorganisms
Microbiology and Molecular Biology Reviews, December 2004, p. 669-685, Vol. 68, No. 4
Jo Handelsman

Abstract
Metagenomics (also referred to as environmental and community genomics) is the genomic analysis of microorganisms by direct extraction and cloning of DNA from an assemblage of microorganisms. The development of metagenomics stemmed from the ineluctable evidence that as-yet-uncultured microorganisms represent the vast majority of organisms in most environments on earth. This evidence was derived from analyses of 16S rRNA gene sequences amplified directly from the environment, an approach that avoided the bias imposed by culturing and led to the discovery of vast new lineages of microbial life. Although the portrait of the microbial world was revolutionized by analysis of 16S rRNA genes, such studies yielded only a phylogenetic description of community membership, providing little insight into the genetics, physiology, and biochemistry of the members. Metagenomics provides a second tier of technical innovation that facilitates study of the physiology and ecology of environmental microorganisms. Novel genes and gene products discovered through metagenomics include the first bacteriorhodopsin of bacterial origin; novel small molecules with antimicrobial activity; and new members of families of known proteins, such as an Na\(^+\)(Li\(^+\))/H\(^+\) antiporter, RecA, DNA polymerase, and antibiotic resistance determinants. Reassembly of multiple genomes has provided insight into energy and nutrient cycling within the community, genome structure, gene function, population genetics and microheterogeneity, and lateral gene transfer among members of an uncultured community. The application of metagenomic sequence information will facilitate the design of better culturing strategies to link genomic analysis with pure culture studies.

Articles containing “metagenomics”
Nov 2011: 8,990 (Google Scholar)
Nov 2015: 26,500 (Google Scholar)
Freshwater Metagenomics

**Abstract**

The bacterial communities of seven freshwater lakes in the Adirondack Mountains of New York state were examined using culture-independent methods. β-Proteobacteria 16S rRNA sequences were recovered from all seven lakes and their presence was confirmed by direct DNA hybridization. The results are consistent with phylogenetic and in situ hybridization-based studies in other freshwater environments, but are significantly different than the results of marine oceanic studies, where β-Proteobacteria are noticeably absent. This relationship between evolutionary history and environmental distribution is striking, since these phylogenetic clades have not been correlated with consistent physiological features or biochemical capabilities, and there is no a priori reason to expect differences in phylogenetic composition between the environments. In contrast, freshwater relatives to marine phylogenetic clusters, in particular the SAR 11 cluster of the α-Proteobacteria, were identified. The data imply an underlying physiological distinction between the β- and other Proteobacteria groups and potentially an important difference between the composition of bacterial communities in marine and fresh-water environments.

**Google Scholar search**

“freshwater metagenomics”

Nov 2011: 2,480
Aug 2015: 8,790
Soil Metagenomics

The ecological role of biodiversity in agroecosystems
_Agriculture, Ecosystems & Environment, Volume 74, Issues 1-3, June 1999, Pages 19-31_
Miguel A. Altieri

Abstract

Increasingly research suggests that the level of internal regulation of function in agroecosystems is largely dependent on the level of plant and animal biodiversity present. In agroecosystems, biodiversity performs a variety of ecological services beyond the production of food, including recycling of nutrients, regulation of microclimate and local hydrological processes, suppression of undesirable organisms and detoxification of noxious chemicals. In this paper the role of biodiversity in securing crop protection and soil fertility is explored in detail. It is argued that because biodiversity mediated renewal processes and ecological services are largely biological, their persistence depends upon the maintenance of biological integrity and diversity in agroecosystems. Various options of agroecosystem management and design that enhance functional biodiversity in crop fields are described.

7,010 articles containing “soil metagenomics” (Nov 2011, Google Scholar)

22,600 articles containing “soil metagenomics” (Nov 2015, Google Scholar)
Phylogenetic diversity of termite gut spirochaetes.

Lilburn TG, Schmidt TM, Breznak JA.

Abstract

A molecular phylogenetic analysis was done of not-yet-cultured spirochaetes inhabiting the gut of the termite, Reticulitermes flavipes (Kollar). Ninety-eight clones of near-full-length spirochaetal 16S rDNA genes were classified by ARDRA pattern and by partial sequencing. All clones grouped within the genus Treponema, and at least 21 new species of Treponema were recognized within R. flavipes alone. Analysis of 190 additional clones from guts of Coptotermes formosanus Shiraki and Zootermopsis angusticollis (Hagen), as well as published data on clones from Cryptotermes domesticus (Haviland), Mastotermes darwiniensis Foggatt, Nasutitermes lujae (Wasmann) and Reticulitermes speratus (Kolbe), revealed a similar level of novel treponemal phylogenetic diversity in these representatives of five of the seven termite families. None of the clones was closely related (i.e. all bore < or = 91% sequence similarity) to any previously recognized treponeme. The data also revealed the existence of two major phylogenetic groups of treponemes: one containing all of the currently known isolates of Treponema and a large number of phylotypes from the human gingival crevice, but only a minority of the termite gut spirochaete clones; another containing the majority of termite spirochaete clones and two Spirochaeta (S. caldaria and S. stenostrepta), which, although free living, group within the genus Treponema on the basis of 16S rRNA sequence. Signature nucleotides that almost perfectly distinguished the latter group, herein referred to as the 'termite cluster', occurred at the following (E. coli numbering) positions: 289-G x C-311; A at 812; and an inserted nucleotide at 1273. The emerging picture is that the long-recognized and striking morphological diversity of termite gut spirochaetes is paralleled by their phylogenetic diversity and may reflect substantial physiological diversity as well.

645 articles containing “termite gut metagenomics” (Nov 2011, Google Scholar)

2430 articles containing “termite gut metagenomics” (Nov 2015, Google Scholar)
Fecal Metagenomics

Metagenomic Analyses of an Uncultured Viral Community from Human Feces
Mya Breitbart,¹ Ian Hewson,² Ben Felts,³ Joseph M. Mahaffy,³ James Nulton,³ Peter Salamon,³ and Forest Rohwer¹,4*

Abstract

Here we present the first metagenomic analyses of an uncultured viral community from human feces, using partial shotgun sequencing. Most of the sequences were unrelated to anything previously reported. The recognizable viruses were mostly siphophages, and the community contained an estimated 1,200 viral genotypes.

1,620 articles containing “human feces metagenomics”
(Nov 2011, Google Scholar)

13,200 articles containing “human feces metagenomics”
(Nov 2015, Google Scholar)
Viral Metagenomics

Viral metagenomics
Reviews in Medical Virology. Volume 17 Issue 2, Pages 115 - 131
Eric L. Delwart
Abstract

Characterisation of new viruses is often hindered by difficulties in amplifying them in cell culture, limited antigenic/serological cross-reactivity or the lack of nucleic acid hybridisation to known viral sequences. Numerous molecular methods have been used to genetically characterise new viruses without prior in vitro replication or the use of virus-specific reagents. In the recent metagenomic studies viral particles from uncultured environmental and clinical samples have been purified and their nucleic acids randomly amplified prior to subcloning and sequencing. Already known and novel viruses were then identified by comparing their translated sequence to those of viral proteins in public sequence databases. Metagenomic approaches to viral characterisation have been applied to seawater, near shore sediments, faeces, serum, plasma and respiratory secretions and have broadened the range of known viral diversity. Selection of samples with high viral loads, purification of viral particles, removal of cellular nucleic acids, efficient sequence-independent amplification of viral RNA and DNA, recognisable sequence similarities to known viral sequences and deep sampling of the nucleic acid populations through large scale sequencing can all improve the yield of new viruses. This review lists some of the animal viruses recently identified using sequence-independent methods, current laboratory and bioinformatics methods, together with their limitations and potential improvements. Viral metagenomic approaches provide novel opportunities to generate an unbiased characterisation of the viral populations in various organisms and environments. Copyright © 2007 John Wiley & Sons, Ltd.

4,020 articles containing “viral metagenomics”
(Nov 2011, Google Scholar)

19,900 articles containing “viral metagenomics”
(Nov 2015, Google Scholar)
Abstract

We sequenced 28 million base pairs of DNA in a metagenomics approach, using a woolly mammoth (Mammutus primigenius) sample from Siberia. As a result of exceptional sample preservation and the use of a recently developed emulsion polymerase chain reaction and pyrosequencing technique, 13 million base pairs (45.4%) of the sequencing reads were identified as mammoth DNA. Sequence identity between our data and African elephant (Loxodonta africana) was 98.55%, consistent with a paleontologically based divergence date of 5 to 6 million years. The sample includes a surprisingly small diversity of environmental DNAs. The high percentage of endogenous DNA recoverable from this single mammoth would allow for completion of its genome, unleashing the field of paleogenomics.
Ocean Metagenomics

Sorcerer II Global Ocean Sampling Expedition

Sorcerer II

J Craig Venter Institute is sequencing of the world’s oceans.
Metagenomics

- Introduction
- The “extended phenotype”
- 16s rDNA metagenomics
- reference sequence alignment
- Multi-genome microbial assembly w/o template
- Discovery metagenomics
- Comparative metagenomics
- Principle component analysis
RICHARD DAWKINS
the
extended phenotype
The long reach of the gene
OXFORD
Host/microbiome

Animal/gut

Plant/rhizosphere
Microbial communities of plants

- Phyloosphere (above ground)
- Anthosphere (flowers)
- Phylloplane (leaves)
- Caulosphere (stems)
- Carposphere (fruit)
Animal microbiome
Microbes make enzymes that break down wood into fermentable sugars.
Metagenomics

- Introduction
- The “extended phenotype”
- 16s rRNA metagenomics
  - reference sequence alignment
  - Multi-genome microbial assembly w/o template
- Discovery metagenomics
- Comparative metagenomics
- Principle component analysis
16s rRNA procaryotic metagenomics

1. Collect sample
2. PCR using 16s RNA universal primers
3. NGS
4. Phylogenetic analysis.
# 16s rRNA

Contains highly conserved regions AND hypervariable regions. Can be used to compare distant and close homolog species. Can be used to identify new species.

## Universal primers

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Sequence (5'-3')</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>8F</td>
<td>AGA GTT TGA TCC TGG CTC AG</td>
<td>[10][11]</td>
</tr>
<tr>
<td>U1492R</td>
<td>GGT TAC CTT GTT ACG ACT T</td>
<td>same as above</td>
</tr>
<tr>
<td>928F</td>
<td>TAA AAC TYA AAK GAA TTG ACG GG</td>
<td>[12]</td>
</tr>
<tr>
<td>336R</td>
<td>ACT GCT GCC YCC CGT AGG AGT CT</td>
<td>as above</td>
</tr>
<tr>
<td>1100F</td>
<td>YAA CGA GCG CAA CCC</td>
<td></td>
</tr>
<tr>
<td>1100R</td>
<td>GGG TTG CGC TCG TTG</td>
<td></td>
</tr>
<tr>
<td>337F</td>
<td>GAC TCC TAC GGG AGG CWG CAG</td>
<td></td>
</tr>
<tr>
<td>907R</td>
<td>CCG TCA ATT CCT TTR AGT TT</td>
<td></td>
</tr>
<tr>
<td>785F</td>
<td>GGA TTA GAT ACC CTG GTA</td>
<td></td>
</tr>
<tr>
<td>805R</td>
<td>GAC TAC CAG GGT ATC TAA TC</td>
<td></td>
</tr>
<tr>
<td>533F</td>
<td>GTG CCA GCM GCC GCG GTA A</td>
<td></td>
</tr>
<tr>
<td>518R</td>
<td>GTA TTA CCG CGG CTG CTG G</td>
<td></td>
</tr>
<tr>
<td>27F</td>
<td>AGA GTT TGA TCM TGG CTC AG</td>
<td>[13]</td>
</tr>
<tr>
<td>1492R</td>
<td>CGG TTA CCT TGT TAC GAC TT</td>
<td>as above</td>
</tr>
</tbody>
</table>

Isolation and characterization of a novel Bacillus strain from coffee phyllosphere showing antifungal activity
J.R. Nair · G Singh · V Sekar ·
Metagenomics

- Introduction
- The “extended phenotype”
- 16s rRNA metagenomics
- Reference sequence alignment
- Multi-genome microbial assembly w/o template
- Discovery metagenomics
- Comparative metagenomics
- Principle component analysis
16s RNA versus metagenomics


- metagenomics found 1.5x phylla and 10x genera as compared to 16s rRNA sequencing.

Too bad, because 16s sequencing is easier…
Reference sequence alignment

1. Collect samples.
2. Shotgun sequencing or NGS.
3. Align to reference genomes.
4. Analysis of abundances.

- Assumes all (relevant) genomes have already been sequenced.
- Requires less DNA than “ab initio” metagenomics
Estimating number of species

The probability of finding a species is proportional to the abundance. The total number can be estimated by extrapolating to infinite sampling. The distribution of abundances can be expressed as the Gini coefficient.
Gini coefficient

- Simple measure of abundance distribution.
- Invented to describe income inequality in a population.
- Used as a measure of (relative) biodiversity.
- $0 \leq G \leq 1$
  
  $G = 1 - 2 \int_0^1 L(X) \, dX$

- $G=0$ represents a perfectly even distribution
- $G=1$ represents a perfectly uneven distribution.

...where $L(X)$ is the cumulative share of total abundances, given $X$ the cumulative share of total species.
Relative abundances as box-whisker plots in log scale

Analysis of 576.7 gigabases of sequence, from faecal samples of 124 European individuals
What is a box&whisker plot?

A way to express how data is distributed in one dimension.
Swarm plot

Shows distribution (unimodal, bimodal, ...) that would be hidden in a box&whisket plot. Allows other visual cues (color, spot size).
Metagenomics

- Introduction
- The “extended phenotype”
- 16s rRNA metagenomics
- Reference sequence alignment
- Multi-genome microbial assembly w/o template
- Comparative metagenomics
- Principle component analysis
De Bruijn graph

Example

nodes = sequence reads
edges = sequence overlap

eetc...
Example

**nodes with simple connections merged**

After simplification…

Low abundance tips and bubbles are most likely sequencing errors.
Example

Tips removed...
Example

Bubbles removed... by TourBus

TAGTCGA  CGAG
    ↓  ↓
GAGGCT  GGCT
    ↓  ↓
GATCCGATGAG
    ↓  ↓
AGAT
    ↓  ↓
GCTTTAG
    ↓  ↓
TAGA
    ↓  ↓
AGAGA
    ↓  ↓
AGACAG
Tour Bus algorithm
Dijkstra’s algorithm with weights

• Find the shortest path from point to point in a graph.
• Nodes with greater abundance count for shorter distance.
• Eliminates frayed ends and bubbles.
Read variety

- Short reads ~75bp
  - Illumina / Solexa
  - SOLiD (colour space)
- Long reads 500-1000 bp
  - 454 read
  - Sanger capillary reads
- Paired-end reads
  - Short reads
  - short insert length
- Mate pair reads
  - Short reads
  - long insert length
Skaffolding by paired-ends, and mate-pairs
Assembly algorithm w/scaffolds
First used for the drosophila genome, 2000

Sequence placement order:
1. “Unitigs” = contiguous confidently assembled reads
2. “Scaffold” = 2 or more Unitigs connected by bundles of re-enforcing BAC-ends
3. “Rocks” = unitigs connected by 2 or more BAC-ends
4. “Stones” = unitigs linked by one BAC-end to a Scaffold.
5. “Pebbles” = un-linked Unitigs.

Myers et al. Science 24 March 2000:Vol. 287. no. 5461, pp. 2196 - 2204
Abundance data can resolve ambiguities in assembly

Species A has 1000 copies

ACGGCTCGCTAATTGCGACTGCAAAAGGCTAGACCG

ACGGCTCGCTAATTGGACTGCAAAAGGCTAGACCG

Species B has 100 copies

Relative abundance of mutations should match relative abundance of species, and can be used to resolve ambiguous assemblies.

ACGGCTCGCTAATTG

ACGGCTCGCTAATTG

ACGGCTCGCTAATTG

GACTGCAAAGGCTAGAC

GACTGCAAAGGCTAGAC

ACGGCTCGCTAATTG

GACTGCAAAGGCTAGAC

ACGGCTCGCTAATTG

GACTGCAAAGGCTAGAC

Widely separated differences. How do we know which characters go together?
Metagenomics

- Introduction
- The “extended phenotype”
- 16s rRNA metagenomics
- Reference sequence alignment
- Multi-genome microbial assembly w/o template
- Comparative metagenomics
- Principle component analysis
Functional Metagenomics

assembly → contigs
reads → Find ORFs
contigs → Find ORFs
Find ORFs → ORFs
ORFs → Search database
Search database → genes
genes → Search GO
Search GO → GO terms
GO terms → statistics
statistics → functional composition
functional composition → reads
**Rows** are grouped by the similarity in their enrichment/depletion pattern across all biological processes. The nodes in the NJ tree represent groupings of samples that preserve the distribution of biological processes.

**Columns** are grouped by the similarity in their enrichment/depletion pattern across all samples. The nodes in the NJ tree represent groupings of biological processes that best indicate (correlate with) the source of the sample.

Enrichment (yellow) or depletion (blue) of genes annotated with a certain biological process (column).
Fecal metagenomics before and after weight-loss supplement Ephedra sinica

“Ma Huang” has been used in traditional Chinese medicine for more than 5000 years

Bacteroides play a fundamental role in processing of complex molecules to simpler ones in the host intestine. Bacteroides was significantly depleted after Ephedra.

Cooking show: BLAST2GO

- Download processed sequences from linked file on web site (dhfrs.b2g)
- In amigo.geneontology.org, run BLAST search. Choose closest homolog.
- Display GO terms for Biological Process
  - What is the function?
Metagenomics

- Introduction
- The “extended phenotype”
- 16s rRNA metagenomics
- Reference sequence alignment
- Multi-genome microbial assembly w/o template
- Comparative metagenomics
- Principle component analysis
Principle component analysis (PCA)

- Dimension reduction.
- Prelude to cluster analysis.
- Helps solve the problem of #variables >> #samples
- Requires choosing “response variables” (the space of the data)
- Finds linear combinations of response variables that best explain the variance in the data.
2D example

PC#1 is a line in the direction of greatest variance.

Data are projected onto the PC#1 line.

PC#2 is a perpendicular line that explains the remaining variance.

If two variables are highly correlated, then there exists a linear combination of the two that explains the variance in those directions.

PC = Principle Component
Solving for Principle Components: The Covariance matrix

Each off-diagonal element is the correlation between variable $i$ and variable $j$. 

\[ \begin{bmatrix}
1 & \frac{<(x_1x_2)>}{<x_1><x_2>} & \cdots & \frac{<(x_1x_n)>}{<x_1><x_n>} \\
\frac{<(x_1x_2)>}{<x_1><x_2>} & 1 & \ddots & \\
\vdots & \ddots & \ddots & \\
\frac{<(x_1x_n)>}{<x_1><x_n>} & \cdots & 1 & \\
\end{bmatrix} \begin{bmatrix}
\eta_{1,1} \\
\eta_{1,2} \\
\vdots \\
\eta_{1,n} \\
\end{bmatrix} = \lambda_1 \begin{bmatrix}
\eta_{1,1} \\
\eta_{1,2} \\
\vdots \\
\eta_{1,n} \\
\end{bmatrix} \]

\[ K_T \vec{\eta}_\alpha = \lambda_\alpha \vec{\eta}_\alpha \]
Principle components are eigenvectors

- Eigenvectors may be solved two ways: (1) by setting the determinant of matrix $K = 0$, then finding all of the solutions of that equation. This is called *eigendecomposition*.

- Or....More easily, (2) *recycle-to-convergence*
  1. Choose a random vector $\eta$
  2. Multiply $\eta$ by $K$ to get $\eta'$.
  3. Recycle. Set $\eta = \eta'$
  4. Repeat (2,3) until $\eta == \eta'$

   The resulting vector $\eta$ is the first PC $\eta_1$.

  5. "Project" the data onto a hyperplane normal to $\eta_1$.
  6. Repeat 1-4 to get $\eta_2$, etc.

$$K_I \vec{\eta}_\alpha \rightarrow \lambda_\alpha \vec{\eta}_\alpha$$

vector recycling
Axes are principle components in word frequency space. Colors indicate classifications.

PCA clusters the classes.

http://journals.plos.org/plosone/article?id=info:doi/10.1371/journal.pone.0009773
Example word analysis for one read, using 2-mer words.

<table>
<thead>
<tr>
<th># Word</th>
<th>Obs Count</th>
<th>Obs Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>194</td>
<td>0.0757221</td>
</tr>
<tr>
<td>AC</td>
<td>129</td>
<td>0.0503513</td>
</tr>
<tr>
<td>AG</td>
<td>203</td>
<td>0.0792350</td>
</tr>
<tr>
<td>AT</td>
<td>185</td>
<td>0.0722092</td>
</tr>
<tr>
<td>CA</td>
<td>212</td>
<td>0.0827479</td>
</tr>
<tr>
<td>CC</td>
<td>174</td>
<td>0.0679157</td>
</tr>
<tr>
<td>CG</td>
<td>49</td>
<td>0.0191257</td>
</tr>
<tr>
<td>CT</td>
<td>174</td>
<td>0.0679157</td>
</tr>
<tr>
<td>GA</td>
<td>175</td>
<td>0.0683060</td>
</tr>
<tr>
<td>GC</td>
<td>142</td>
<td>0.0554254</td>
</tr>
<tr>
<td>GG</td>
<td>131</td>
<td>0.0511319</td>
</tr>
<tr>
<td>GT</td>
<td>118</td>
<td>0.0460578</td>
</tr>
<tr>
<td>TA</td>
<td>130</td>
<td>0.0507416</td>
</tr>
<tr>
<td>TC</td>
<td>163</td>
<td>0.0636222</td>
</tr>
<tr>
<td>TG</td>
<td>184</td>
<td>0.0718189</td>
</tr>
<tr>
<td>TT</td>
<td>199</td>
<td>0.0776737</td>
</tr>
</tbody>
</table>

This represents a single point in a 16-dimensional space that is all-positive, bounded.

PCA is dimensional reduction. 16D -> 3D
Usually 2 or 3 dimensions is enough visualize clusters.
Review

1. What does BLAST2GO do?
2. What are 16s sequences? Why use them?
3. What kind of experimental techniques are used to assemble genomes from short read data?
4. What kind of data is used to resolve ambiguous assemblies?
5. How is relative biodiversity measured?
6. Why study fecal metagenomics?
7. How are the number of species estimated?
8. What good does it do to cluster GO terms from metagenomic data?
9. What good does it do to cluster samples from metagenomic data?
10. What good principle component analysis?