16S metagenomics training course

TRAINERS:
FALK HILDEBRAND, RAUL TITO

Research assistant (population genetics)

Master thesis (population genetics)

PhD in (metagenomics)

PostDoc in (metagenomics)

Thierry Wirth
Adam Eyre-Walker
Jeroen Raes
Peer Bork

2004 2008 2009 2014 2019
Ancient DNA and forensic genetics

Multi-domain Gut microbiota

Industrialization effect Gut microbiota

Human and population genetics

Ancient DNA and forensic genetics

Multi-domain Gut microbiota

Industrialization effect Gut microbiota

Bachelor
Universidad Mayor de San Marcos
Peru

Master
Universidad de Granada and University of Oklahoma

Phd
VUB-KUL-VIB

Post-Doc
KUL-VIB

Dr. Raul Y. Tito-Tadeo
Post-doctoral Scientist
KUL-VIB
What we will cover during the course

MORNING (9:30 – 12:30)

16s amplicon sequencing ← technical
- LotuS pipeline introduction (45 min)
  ◦ What is amplicon sequencing
  ◦ How to make sense of raw sequences?
- hands-on exercise (2 hours, linux):
  ◦ Demultiplex illumina data
  ◦ Quality filter appropriate for 16S experiments
  ◦ Use different taxonomic assignments
  ◦ Rarefy your data

AFTERNOON (13:30 – 17:00)

Numerical ecology ← biological
- Metagenomics introduction
  ◦ Applications, limits and opportunities
  ◦ How can we analyze the data reproducibly and statistically sound?
- hands-on exercise (2 hours, R):
  ◦ Loading data and normalizing it
  ◦ Testing for increased or decreased taxa
  ◦ Visualizing data
Amplicon sequencing

PROCESSING RAW READS WITH LOTUS AND OTHER PIPELINES
Overview

1. Introduction to metagenomics
2. Amplicon based metagenomics – common pipelines available
3. Discussion of technical details in LotuS amplicon pipeline
4. Exercise – LotuS pipeline to process raw sequence reads
The fascinating world of microbes..

.. AND HOW TO STUDY THEM
Microbiology + Ecology = Microecology

Ecology + Statistics = Numerical Ecology

Sequencing + Microbiology + Ecology = Metagenomics
Methods to study the microbiome

Microbial community

- Which microbes are there?
  - Nucleic acids
    - SSU rRNA approaches
  - RNA
    - Metatranscriptomics
  - Proteins
    - Metaproteomics
  - Metabolites
  - DNA
  - Metagenomics

- What are the microbes doing?
  - Metabolomics

- What is the genetic potential?
The metagenomic revolution

**Definition Metagenomics**

“Description of all microorganisms in an environment, using sequencing technology”
Metagenomics is about observing organisms, that are too small to see

**Bacteria**
- *Escherichia coli, Pseudomonas aeruginosa*
- Most diverse kingdom

**Archaea**
- Often very specialized (thermophile, cryophiles, acidophiles)

**Unicellular eukaryotes**
- Fungi, amoeba
Advantages of metagenomics

- Discover unculturable organisms
- Cheaper & higher throughput than classical methods
- High information content – diverse data analysis possibilities

**METATAXONOMICS (16S)**
- Streamlined analysis
- High taxonomic accuracy
- Most 16S reads can be classified

**METAGENOMICS PER SE**
- No PCR bias
- Function and taxonomy
- Genome assemblies
- Strain level resolution
16S metagenomics = amplicon sequencing
Shotgun / WGS metagenomics
16S amplicon sequencing

Established **affordable** and **high throughput** method for many labs to research bacterial communities

Many of the childhood diseases have been successfully tackled, some remain

**LotuS** is about tackling the bioinformatics side (crunching)

Bioinformatic analysis should **also be affordable** for every lab

- Use of desktop computers
- No specialized training
- Integrate the latest bioinformatics methods
- Automate user choices in an optimal workflow of informed choices programmed by experts
Marker gene: 16S rRNA gene

Phylogenetic Tree of Life

http://www.biochem.umd.edu
http://en.wikipedia.org/wiki/Carl_Woese
Types of amplicon sequencing

First we need a gene that is universally conserved (also called maker gene)

- Ribosomal RNA works well
  - SSU (16S, 18S)
  - LSU (23S, 28S)
  - ITS
- Universally conserved marker
  - E.g. ribosomal proteins
  - Diversity too high to get good primers
Ribosomal marker gene databases

**SILVA – SSU & LSU**
- [arb-silva.de](http://arb-silva.de)
- Frequent updates
- Most comprehensive database (SSU & LSU)
- 2,090,668 (SSU), 198,843 (LSU) sequences

**Greengenes – SSU**
- [greengenes.secondgenome.com/](http://greengenes.secondgenome.com/)
- Last update May 2013
- 1,262,986 sequences

**Unite – ITS**
- [unite.ut.ee/](http://unite.ut.ee/)
- Relatively frequent updates
- 690,548 sequences

**PR2 – 18S**
- specialized on protists

---

**Unite taxa distribution**

---

**SILVA SSU / LSU 132 - full release**

<table>
<thead>
<tr>
<th></th>
<th>SSU Parc</th>
<th>SSU Ref</th>
<th>SSU Ref NR 99</th>
<th>LSU Parc</th>
<th>LSU Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimal length</td>
<td>300</td>
<td>1200/900</td>
<td>1200/900</td>
<td>300</td>
<td>1900</td>
</tr>
<tr>
<td>Quality filtering</td>
<td>basic</td>
<td>strong</td>
<td>strong</td>
<td>basic</td>
<td>strong</td>
</tr>
<tr>
<td>Guide Tree</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Aligned rRNA</td>
<td>6,073,181</td>
<td>2,090,668</td>
<td>695,171</td>
<td>907,382</td>
<td>198,843</td>
</tr>
</tbody>
</table>
Sequencing costs are rapidly decreasing

End of human genome project (1st gen., Sanger)

Craig Venter’s personal genome (10 M$)

Adoption of 2nd gen. technology

PacBio (3rd gen.)

Illumina (2nd gen.)
Lots of tiny sequences!

[http://flxlexblog.wordpress.com/]
Chimeric Reads

Common problem of PCR to amplify 16S gene

Results in 16S sequences, that belong to two or more Target organism

Bioinformatic solution: Chimera detection
- De novo (uchime2, Perseus, vsearch, CATCh)
- Ref based (uchime2, ChimeraSlayer, Pintail, Mallad, DECIPHER)
Chimeric Reads

Common problem of PCR to amplify 16S gene

Results in 16S sequences, that belong to two or more Target organisms

Bioinformatic solution: Chimera detection
  ◦ De novo (uchime2, Perseus, vsearch, CATCh)
  ◦ Ref based (uchime2, ChimeraSlayer, Pintail, Mallad, DECIPHER)
Metagenomic amplicon pipelines
QIIME / QIIME2

(quantitative insights into microbial ecology)

Comprehensive suite of tools

Driven by Rob Knight lab

Medium speed

Outdated OTU calling, infamous for overestimating diversity

Relatively stable phylogentic tools

Nice suite of visualizations

..and basic numerical ecology tools

www.qiime.org/
mothur

Comprehensive suite of tools
Driven by Pat Schloss lab
Very slow, but stable clustering
Extreme resource requirements (careful if you have a big dataset)
Has a GUI (graphical user interface) !!
No visualizations
Basic numerical ecology functions

http://www.mothur.org/
DADA2

(Divisive Amplicon Denoising Algorithm)

Most recent clustering algorithm
Can define clusters with 1 nucleotide difference
Finds fine scaled variants
Otherwise similar performance to uparse

Incomplete pipeline, as it requires
- Non-biological nucleotides have been removed (primers/adapters/barcodes...)
- Samples are demultiplexed (split into individual per-sample fastqs)

https://github.com/benjjneb/dada2
LotuS pipeline

Ease of use

~10 X faster than Qiime/ Mothur / dada2
- 2x 454 runs (70 samples) completely processed in 1 minute on a macbook

Dual quality read filtering: **high quality OTUs** with **deeper coverage**
- State-of-the-art clustering, chimera removal, denoising

Extensive LCA algorithm to assign taxonomy

Standardized file formats

Simple installation and updating

http://psbweb05.psb.ugent.be/lotus/
## Common pipelines to analyze amplicon metagenomics

<table>
<thead>
<tr>
<th>Pipeline task</th>
<th>Qiime</th>
<th>Mothur</th>
<th>DADA2</th>
<th>LotuS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic sequence filtering / demultiplexing</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Probabilistic sequence filtering</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Denovo OTUs</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Reference based OTUs</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Taxonomic classification</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Phylogenetic tree</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Visualization</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic numerical ecology</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
LotuS pipeline

ALGORITHMIC DETAILS
Pipeline layout

Mix of Perl script (glue) and C++ program (heavy lifting)

> 10 external program potentially used

\textbf{LotuS} = less OTU scripts

\textbf{sdm} = simple demultiplexer
Sequence clustering algorithms

LotuS relies on three clustering algorithms

1. UPARSE
2. SWARM
3. CD-HIT

Planned / incompletely integrated

1. DNACLUST
2. SORTMERNA


How to use? (User interface)

Design principle was to automate as much as possible and reduce user intervention, while keeping fine tuning still available to expert users

Installation
- `./autoinstall.pl` → answer 4 questions → install usearch → Done

Run LotuS
1. Create mapping file
2. Configure sdm quality filter (optional)
3. `./lotus.pl` –i r1.fq.gz,r2.fq.gz –m Map.txt –s sdm_opt.txt –o Outputfolder/
4. Analyze data (in R / Matlab etc.)

Example mapping file

<table>
<thead>
<tr>
<th>#SampleID</th>
<th>CombineSamples</th>
<th>fastqFile</th>
<th>ForwardPrimer</th>
<th>ReversePrimer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample_001</td>
<td>sdh1/S1_L001_R1.fastq.gz,sdh1/S1_L001_R2.fastq.gz</td>
<td>CCTACGGGNGGCWGCAG</td>
<td>GACTACHVGGGTATCTAATCC</td>
<td></td>
</tr>
<tr>
<td>Sample_002</td>
<td>sdh1/S9_L001_R1.fastq.gz,sdh1/S9_L001_R2.fastq.gz</td>
<td>CCTACGGGNGGCWGCAG</td>
<td>GACTACHVGGGTATCTAATCC</td>
<td></td>
</tr>
<tr>
<td>Sample_003</td>
<td>sdh1/SS_L001_R1.fastq.gz,sdh1/SS_L001_R2.fastq.gz</td>
<td>CCTACGGGNGGCWGCAG</td>
<td>GACTACHVGGGTATCTAATCC</td>
<td></td>
</tr>
<tr>
<td>Sample_004</td>
<td>sdh1/S10_L001_R1.fastq.gz,sdh1/S10_L001_R2.fastq.gz</td>
<td>CCTACGGGNGGCWGCAG</td>
<td>GACTACHVGGGTATCTAATCC</td>
<td></td>
</tr>
<tr>
<td>Sample_005</td>
<td>sdh1/S11_L001_R1.fastq.gz,sdh1/S11_L001_R2.fastq.gz</td>
<td>CCTACGGGNGGCWGCAG</td>
<td>GACTACHVGGGTATCTAATCC</td>
<td></td>
</tr>
</tbody>
</table>
Barcode sequences assign unambiguously a read to each sample used – identifying and removing the Barcode is called “Demultiplexing”. Barcodes can contain heterogeneity Spacers\(^1\) to increase sequence quality – this is automatically detected and removed by LotuS.

The amplicon is the sole region that should be used for OTU building, tax assignments, OTU abundance and phylogeny. The amplicon primers are used to specifically amplify a gene region you are interested in – choose careful to avoid biases towards certain phylogenetic groups and to get an amplicon length appropriate to the sequencing technology used.

Illumina primers are used by the sequencer to detect the point from where to start sequencing and to physically bind the flowcell – illumina software should in all cases remove these.

These sequences can be in IUPAC code, but must be in read direction (i.e. all can be found by looking at the raw sequences without rev-complementing).

---

sdm (1) – quality filter & dereplication

**Initial Read pairs**

- Autodetect
  1. Fastq version
  2. Barcode orientation
  3. Reversed read pairs
  4. Reverse 16S Primer in 1st read (e.g. short amplicons)

- Check each read for
  1. PCR Primers
  2. Barcodes
  3. Heterogenity Spacers
  4. Cut outside sequence of any of these

- Quality filter
  1. Average Quality
  2. Quality Window
  3. Number of ambiguous bases
  4. Homonucleotide runs
  5. Min/Max Seq length

- Trim reads -> Dereplicate & count per Sample

- High quality, “ready for clustering” reduced read set

**Probabilistic filter**

1. Accumulated error
2. Estimated read error under binomial model
Quite a lot of logs – and these are important to fine tune your runs

- Overall statistics – how many reads passed & why some were filtered
- Read length histogram before & after filtering
- Avg Quality histogram before & after filtering
- Accepted reads per file (different runs OK?)
- Accepted reads per Sample (PCR OK?)
- Observed Q per Base (illumina has base specific biases)

Statistics of high quality reads

Reads processed: 54,762,653; 54,762,653 (pair 1; pair 2)
Rejected: 25,361,468; 29,143,967
Accepted: 29,401,185; 25,618,686 (0; 0 were 5' trimmed, with rev. primer: 0; 25,618,686)
Singletons among these: 8,202,546; 4,420,047

Min/Avg/Max stats Pair 1
- Seq Length : 190/43.9/190
- Quality : 27/34.3/38
- Median Seq Length : 0, Quality : 0
- Accum. Error 0.121

Filtered due to:
- < min Seq length (190) : 280721; 269931
  - after Quality trimming : 11553774; 19286867
- < avg Quality (28) : 4,856,114; 6074979
- < window (50 nt) avg. Quality (25) : 0; 0
- > max Seq length (1000) : 0; 0
- > (8) homo-nt run : 66,572; 4155
- > (0) amb. Bases : 2,058; 20310
- > (1) acc. errors : 0; 47643
- > (2.5) binomial est. errors : 5254466; 0
  - Failed to find fwd Primer (max 0 errors, required) : 3,347,763; 0
  - Failed to find rev Primer (max 0 errors, required) : 0; 3,452,225
  - Barcode unidentified (max 0 errors) : 0; 35,500 (0 pairs failed)

<table>
<thead>
<tr>
<th>SampleID</th>
<th>SampleGroup</th>
<th>Barcode</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample_001</td>
<td>Sample_001</td>
<td>68,964</td>
<td></td>
</tr>
<tr>
<td>Sample_002</td>
<td>Sample_002</td>
<td>41,365</td>
<td></td>
</tr>
<tr>
<td>Sample_003</td>
<td>Sample_003</td>
<td>95,870</td>
<td></td>
</tr>
<tr>
<td>Sample_004</td>
<td>Sample_004</td>
<td>76,443</td>
<td></td>
</tr>
</tbody>
</table>
Quite a lot of logs – and these are important to fine tune your runs

- Overall statistics – how many reads passed & why some were filtered
- Read length histogram before & after filtering
- Avg Quality histogram before & after filtering
- Accepted reads per file (different runs OK?)
- Accepted reads per Sample (PCR OK?)
- Observed Q per Base (illumina has base specific biases)

Note stark difference pair 1 vs pair 2!

*unpublished dataset
Clustering: Why shorten reads by default?

Seems at first counterintuitive: Why throw away nucleotides?

_Historically_, the biggest problem of 16S was and is overestimation of diversity ("rare biosphere")
- Due to sequencing errors

Tackle overestimation by
- Expensive clustering (denoising)
- Remove sequence errors

Lotus removes sequence errors with several methods, to **control diversity overestimation**

---

Binomial model of error distribution

Calculates probability to observe $X$ errors in sequence under binomial model

Extends UPARSE accumulated error approach with an explicit statistical model

Typical LotuS result: 50% of OTU’s while retaining 90% of reads

Clustering(2): Why use only first read?

1. Second read of too low quality
2. Merging reads does not resolve ambiguous bases
   - Is Read 1 or Read 2 correct?
   - Subsequent Quality reduction detrimental, as usually too many disagreeing positions
3. Computational complexity

---

First read

Second read

2\textsuperscript{nd} read erroneous despite high quality

---


Sdm(2): Seed extension - Recovering read length

Basic Idea: Find for each OTU the longest, highest quality representative read (pair)
- Restores full length reads, for FLASH read pair merging

All reads assigned to OTU are compared for
- Highest overall Quality
- Lowest accumulated error
- Highest similarity to OTU centroid

Step also used to
- Count OTU abundance / sample
- Distribute chimeric read counts (flag -count_chimeras)

97% OTU id

OTU seed used to
- Assign taxonomy
- Multiple Alignment
- Construct phylogenetic tree
Taxonomy assignment to OTU’s

RDP classifier
- Machine learning approach
- Robust performance & super fast
- Advantageous for unknown taxa

Similarity based taxonomy (LCA)
- Against greengenes / SILVA / UNITE / PR2
- 16S, 23S & ITS supported
- Blast / lambda searches
- Last Common Ancestor algorithm
  - Unknown species should be identified as being unknown genera / family / class etc

![Graph showing precision and specificity]

Default LotuS setting
Summary

Metagenomics is a powerful tool for describing bacterial community
- Simply put, this is ecology at a microscopic level

Amplicon based metagenomics focuses only on 16S rRNA, to retrieve phylogenetic composition of (bacterial) microbiome

Different pipelines available to process 16S amplicon metagenomics
- Qiime and mothur are established (old guard) pipelines
- DADA2 example of newer clustering methods
- LotuS is a slim pipeline reduced to essential parts, at a high quality

In this tutorial, LotuS will be used
Acknowledgements

VIB
- Jeroen Raes
- Raul Tito

EMBL
- Peer Bork

QIB/EI
- Andrea Telatin
- Stefano Romano

Others
- Julien Tap
- Mohammad Braham
- Bubba Brooks
- Fernando Puente Sánchez
- Alireza Kashani
- ... and many more for comments and bug reporting
Tutorial outline

Load genus abundance table into R

Analyze ecological aspects of matrix using numerical methods (Numerical ecology)
- Clustering
- Ordination
- Diversity

Lotus
Discussion

1. To pair or not? Merge read pairs before clustering OTUs?
2. Distribute Chimeric read counts on “mother” OTUs?

http://psbweb05.psb.ugent.be/lotus/
Clustering(2): Why use only first read?

1. Computational complexity

2. Second read of too low quality

3. Merging reads does not resolve ambiguous bases
   - Is Read 1 or Read 2 correct?
   - Subsequent Quality reduction detrimental, as usually too many disagreeing positions