Microbes and Microbiome
Julie Segre, PhD
Senior Investigator, National Human Genome Research Institute, NIH

Current Topics in Genome Analysis 2016
Julia Segre

No Relevant Financial Relationships with Commercial Interests
Why the Human Microbiome?

Each human cell has the same protein-encoding potential. Microbes are more diverse and dynamic than human genome.

Human Microbiome

- Humans are hosts to many microbes (bacteria, fungi, viruses)
- Microbiome is totality of microbial community DNA
- Microbial cells outnumber human cells
- Many unknown functions of microbes
- Many microbes are often considered pathogenic
  - Mycobacterium tuberculosis
  - Staphylococcus aureus
Not all microbes are bad: Beneficial microbes perform functions essential for human health

- Vitamin synthesis
- Digestion
- Education and activation of immune system
- Inhibition of skin colonization by pathogens

Many microbial-host and microbial-microbial interactions remain unknown

Elucidating the diversity of the human microbiome

- Traditional approaches rely on isolating bacteria in pure culture
- The majority of bacterial species do not grow in culture = “the great plate count anomaly”
- Culturing favors lab weeds--not necessarily the most dominant or influential species
- Excludes microbes that rely on community interactions
Direct sequencing vs. culture-based methods

Direct sequencing vs. culture data

Survey Culture Survey Culture

Actinobacteria
Cyanobacteria
Proteobacteria
Firmicutes
Others
Unidentified

Division contributing < 1%
Unclassified
Topics for today’s talk

1. Bacterial diversity studies: 16S rRNA
2. Fungal diversity studies: ITS1
3. Bacterial genomes: Shotgun sequencing
4. Metagenomics
5. Where is the technology going?

Core marker genes are specific to taxonomic clades.

- Eukarya, 18S/ITS1/28S
- Bacteria, 16S rRNA
- Archaea, 16S rRNA

Orange = rRNA;
Blue = small subunit proteins
Green = large subunit proteins
TOPIC 1. Bacterial Diversity: 16S rRNA gene

Universal and variable regions of 16S rRNA used for PCR amplification & classification

Peterson et al., Cell Host & Microbe (2008)
Siqueira et al., Journal of Oral Microbiology (2012)

Basic workflow

Sample microbiome → Harvest gDNA → Amplify 16S rRNA → Sequence → Taxonomic classification → Population analysis

Kong, Trends in Mol Medicine, 2011
Important Issues to Consider Before Initiating Experiment

1. Study Design. Define the question as precisely as possible; e.g. ‘I want to compare wild-type with knock-out mice.’ Are these mice littermates? Because there is a lot of variation between individuals, cages and facilities. What controls do you need?
2. What sequencing platform will you use?
3. What region of the 16S rRNA gene will you amplify?
4. How many reads do you need per sample?
5. What are hidden technical issues? CHIMERAS
6. What analysis tool will you use?
7. How will you display your data?
8. How will you compare your results with other published studies?
9. What information will yield a testable hypothesis?

Calculating Bacterial Load: qPCR with primers in conserved region of 16S rRNA gene

<table>
<thead>
<tr>
<th>Human DNA</th>
<th>309 pg</th>
<th>Bacterial DNA</th>
<th>39 pg</th>
<th>3 pg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ct</td>
<td>copy #</td>
<td>Ct</td>
<td>copy #</td>
</tr>
<tr>
<td>0.1 ng</td>
<td>17.85</td>
<td>54924.50</td>
<td>20.92</td>
<td>6951.93</td>
</tr>
<tr>
<td>0.3 ng</td>
<td>17.78</td>
<td>57573.03</td>
<td>20.93</td>
<td>6953.23</td>
</tr>
<tr>
<td>0.5 ng</td>
<td>17.76</td>
<td>57655.09</td>
<td>20.93</td>
<td>6953.30</td>
</tr>
<tr>
<td>1.0 ng</td>
<td>17.76</td>
<td>57655.09</td>
<td>20.93</td>
<td>6953.30</td>
</tr>
</tbody>
</table>

Ct of qPCR of bacterial DNA to calculate relative bacterial counts of each sampling method. Must also consider how to normalize sample. /cm² or /g stool?

- Swab yields 10,000 bacteria/cm²
- Scrape yields 50,000 bacteria/cm²
- Biopsy yields 1,000,000 bacteria/cm²

Grice et al, Genome Research 2008
Castillo M...Gasa J...2006
DNA Sequencing to assess bacterial diversity

Illumina Mi-Seq (2 x 300 bp paired-end reads)

– 2 runs/week on one instrument.
– Costs $2K, which is $4/sample if you multiplex 500 samples.
– Scale is the issue. Need to dual-index bar-code primers for multiplexing since platform generates >10 million reads per lane. Assume 10,000 reads is more than enough per sample, you can multiplex 500+ samples together in one lane.
– Short reads, but can link paired reads.
Primer: 8F______________ __________________505R primer

For a SMALL study, SEQUENCE is limiting;
For a LARGE study, BIOINFORMATICS is limiting.

Fadrosh DW…Ravel J Microbiome 2014;
Kozich JJ….Schloss PD Appl Environ Microbiol 2013;
Caporaso JG…Knight R ISME J 2012

Other means of sequence data acquisition

• Phylochip (16S rRNA microarray)
  – Limited to known taxa, but can get species-level designations
  – More expensive.
  – will never find unique or novel species

Hi-Seq Illumina (2 x 100 bp paired-end reads)

– Production sequencing. High output mode (TruSeq v3 chemistry) runs for 10 days and produces 4 billion clusters.
How to identify a bacterial sequence and align sequences?

Matches MANY sequences. And many of them are not type strains, but UNCULTURED from 16S rRNA sequencing study.

Pipeline tools

-mothur
-qime
-CloVR
Alignment & Classification

- Reference-dependent
  - Ribosomal Database Project (RDP), SILVA, Greengenes
- But what about species?
  - Amplify the appropriate region of 16S rRNA gene (V1-3 for Staphylococcus\(^1\); or Lactobacillus\(^2\)) and use custom database.
- Sequences with no reference? Not so many of those, might have to consider other explanations

\(^1\)Conlan, PLoS One 2012; \(^2\)Ravel PNAS 2011

RDP Database  http://rdp.cme.msu.edu/

- RDP 10.18 consists of 920,643 aligned and annotated 16S rRNA sequences. Naïve Baysian classifier based on Bergey’s taxonomy. (Note: other taxonomies such as Euzebey and NCBI exist).
- Tools: RDP classifier, Seqmatch, Probematch
Silva Database (ARB): http://www.arb-silva.de/

Build a Phylogenetic Tree and Calculate Branch Length


SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.

Nuc. Acids Res. 2007; Vol. 35, No. 21, p. 7188-7196

Chimeras: PCR generated (template switching)

Evaluate Accuracy:
– True Positives (TP): artificial chimeras flagged
– False Positives (FP): reference (non-chimera) flagged
How Do Chimeras Occur?
Incomplete extension of PCR, Template Switching at Conserved Regions

ChimeraSlayer Detection Program
http://microbiomeutil.sourceforge.net

Do not underestimate primer bias or chimeras! Mock community contains 20 bacteria. Amplified and sequenced with various methods

NIH Common Fund: Human Microbiome Project

Longitudinally assess microbial diversity of 250 healthy subjects at 5 major body sites

http://commonfund.nih.gov/hmp/

Signature Taxa for each major body site, large variation between subjects

Grice and Segre, Annual Review of Human Genetics and Genomics, 2012
Skin microbiome changes with the transition through puberty

Figure 3A legend
- Actinobacteria (other)
- Corynebacteriaceae
- Propionibacteriaceae
- Firmicutes (other)
- Staphylococcaceae
- Streptococcaceae
- Betaproteobacteria
- Gammaproteobacteria
- Proteobacteria (other)
- Alphaproteobacteria
- Bacteroides
- Clostridia
- Other

Ecological Measures, OTU: Operational Taxonomic Unit
Cluster Sequences Based on Furthest Joining Method; i.e. Every sequence is at most X% different from every other sequence in the group

Calculate all pairwise distances between sequences

% identity within group determines the number of OTUs produced. This should be done on the TOTAL dataset. Most experiments classify at the 97% or 99% identity.

Adapted from Robert Edgar
Comparing Bacterial Diversity: Community Membership & Structure

<table>
<thead>
<tr>
<th>Grp A</th>
<th>Grp B</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>50</td>
</tr>
<tr>
<td>34</td>
<td>50</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Community Membership
(Categories of fruit in common)
= 2/5 = 0.4

Community Structure
(Pieces of fruit in common)
= ~ 0.9

Community Membership:
Pups are most like their mothers

Animal Name | Sex | Genotype
-------------|-----|---------
M3-3         | Female | +/-  
M3-1         | Male   | +/-   
M3-2         | Male   | ob/+  
M1-1         | Female | ob/+ob 
M1-2         | Male   | ob/+ob 
Mother 1     | Female | ob/+ob 
M1-4         | Female | +/-   
M1-5         | Female | ob/-  
M2A-1        | Female | ob/-ob 
M2B-2        | Male   | ob/-ob 
M2A-2        | Male   | ob/-ob 
M2A-3        | Female | ob/-ob 
M2A-4        | Female | ob/-ob 
M2B-1        | Male   | +/-   
M2B-3        | Male   | ob/+ob 
M2B-4        | Male   | ob/+ob 
Mother 2     | Female | ob/+ob 

0.10
Community Structure:
Pups cluster according to genotype

Scharschmidt et al. JID 2009

How many reads do you need? Ballpark 1,000 sequences for first pass analysis. With high throughput sequencing, no longer as relevant. How much diversity is there in the population? Have you sequenced enough to capture the diversity? Chao1 rarefaction curves
Richness, evenness, diversity: Shannon and Simpson diversity

**Richness:** Number of OTUs

**Evenness:** Shannon Equitability Index

Relative distribution of sequences among the OTUs.

0 is least even. 1 is most even distribution.

Shannon Diversity Index accounts for both richness and evenness of OTUs.

---

**Microbial community profiling for human microbiome projects: Tools, techniques, and challenges**

Microbial community profiling for human microbiome projects: Tools, techniques, and challenges

Microbiome Instability and Amy Knight

Genome Res. 2016 Mar; 11411-1152 originally published online April 21, 2009

Access the most recent version at doi:10.1101/gr.164513.116

**Experimental and analytical tools for studying the human microbiome**

Justin Kuczynski*, Christian L. Loafer*, William A. Walters*, Laura Wegener Parfrey*, José C. Clemente*, Dirk Gevers and Rob Knight*
TOPIC 2: Fungal Diversity

Skin swab

16S rRNA amplification → Ribosome Database
   Project classification

Custom ITS database

ITS1 marker amplification → Mine Genbank
   Resolve taxonomy
   Remove redundancy

Bacteria

Fungi

18S, ITS1, 5.8S, ITS2, 28S

Healthy skin bacterial survey (2009)

Grice et al., Science 2009
Healthy skin fungal survey (2013)

- (Gb) Glabella
- (Ea) External auditory canal
- (Na) Nare
- (Mb) Manubrium
- (Ac) Antecubital fossa
- (Vi) Volar forearm
- (Hp) Hypothenar palm
- (Ic) Inguinal crease
- (Tw) Toe web space

Core sites are low diversity; divergent at peripheral sites

Findley, Oh et al., Nature 2013
TOPIC 3. BACTERIAL GENOME

1. What is study objective? E.g. Determine if two hospital isolates are clonal? Or Determine what genes are encoded by diverse set of Staphylococcus epidermidis?
2. What reference genomes exist for phylogenetic comparison?
3. What sequencing platform will you use?
4. What depth of sequencing do you need for assembly?
5. What assembly tool will you use? What alignment tool will you use?
6. How will you display your data?
7. How will you compare your results with other published studies?
8. What information will yield a testable hypothesis?

TOPIC 3. BACTERIAL GENOME
How to Assemble a Bacterial Genome:
Gram-negative is ~6,000,000 base pair

Shotgun sequence 2x300 bp fragments on Illumina MiSeq at 30-fold redundancy.
Overlapping reads form large DNA contigs with N50 of ~100 kb.

Or very low coverage (3-5X) just to define species and strain
Assemblers (de novo)

- Phrap
- Celera
- Velvet
- SPAdes
- mira
- MaSuRCA
- ALL-PATHS

Velvet (Zerbino and Birney, 2008)

- Works in base-space and color-space
- Good for small genomes
- Agnostic of read length
  1. Construct k-mer hash
  2. Build De Bruijn graph
  3. Simplify graph
  4. Resolve
     1. Tips
     2. Bubbles
Evaluating Assemblies

- Coverage is a measure of how deeply a region has been sequenced
- The Lander-Waterman model predicts 8-10 fold coverage is needed to minimize the number of contigs for a 1 Mbp genome
- The N50 size is the point at which 50% of bases are in contigs this size or greater
PacBio Single Molecule Sequencing generates complete bacterial genomes as references

Genome Aligners: Compare sequences to identify sequence nucleotide variants, Insertion/Deletions
1. MumMER
2. MUGSY
3. MAUVE

Genome Annotation: Predicting and naming genes encoding proteins
1. PGAAP (NCBI)
2. IMG (JGI)
3. Glimmer, GeneMark
Is there a reference genome? Is the ‘pan-genome’ open or closed?

Whole Genome Sequence Comparison

GENOME 1 VERSUS GENOME 2 or REFERENCE

Single nucleotide variant (SNV) Mutation Deletion Insertion
Polyclonal outbreak of multi-drug resistant *Acinetobacter baumannii* (A,B, C). Clusters of variants = recombination

*ACICU is a previously sequenced strain of the EC II lineage from a 2005 outbreak in Rome.

Alternate O-antigen biosynthetic clusters

**Circos**: an information aesthetic for comparative genomics. Marra MA, Genome Res. 2009 Sep; 19(9):1639-45
2011 outbreak at NIH of carbapenem-resistant *Klebsiella pneumoniae*:

Evidence for transmissions originating from distinct sites on patient 1
TOPIC 4. METAGENOMICS: DNA sequence from multiple organisms

Fungal, Bacterial, Viral, Archaeal DNA all together (with human DNA).
Very Complex mixture and very complex computationally.

Metagenomics

Ten years after the term metagenomics was coined, the approach continues to gather momentum. This culture-independent, molecular way of analysing environmental samples of coexisting microbial populations has opened up fresh perspectives on microbiology.
Goals of whole genome shotgun metagenomic analysis

1. Want to know who’s there & abundance
2. Want to know what they do (function)
   – Want to know what genes are present
   – Can we identify pathways?
   – Can we identify strains?
3. Can we recover genomes?
4. Can we find novel pathogenic organisms?

Whole genome shotgun sequencing vs. amplicon

- Amplify marker region using universal primers
- Fragment genomic DNA
- Sequence amplicons; database-based classification
- Amplification & library construction
- Shotgun sequencing
Metagenomics: types of bacteria similar between 2 populations, but pink genes enriched in top population

Using metagenomic sequencing to find new metabolic enzymes


Metagenomics: computational infancy to develop pipelines for analysis

Paired-end Illumina data

Read-based

Assembly-based

Map to references

Completeish genomes

Functional (KEGG)

Strain heterozygosity

Pathoscope

Segata et al., Nature Methods (2012)

Sunagawa et al., Nature Methods (2013)

Owen et al., Genome Research (2013)

Adaptive iterative de novo assembly to generate contigs

Looking for function

- Leverage functional databases like

  - KeGG

  - MetaCyc

  - The Gene Ontology

- Generally, use blastx-like programs to map reads to these databases

  - eggNOG 4.0
HUMANN (The HMP Unified Metabolic Analysis Network): post-processes BLASTx hits into pathways

Example output

Human Microbiome Project Nature. 2012
Calling genomes

- Binning methods
  - GC content/ tetranucleotide frequencies

Ghai et al., Scientific Reports 2011

Metagenomic linkage groups to aggregate contigs

Karlsson et al., Nature (2013)
Strain Tracking: Noncore regions confer important functional differences

Non-human metagenomics reads

P. acnes genome sequences

Strain tracking with ClinPathoscope
Step 1: Align reads to strain sequences
Strain tracking

Step 2:
Reassign ambiguous reads with PathoScope

Strain variation in *P. acnes* is driven by host-differences

Oh, Byrd et al., Nature 2014
Strain variation in S. epi is driven by host-differences and site characteristic.

Different species can be differentially shaped by host and environment.

Oh, Byrd et al., Nature 2014
Strain Tracking with read assignment, finds core and accessory genes

Zhu, Sunagawa, ..., and Bork, Genome Biology 2015

Why are strains important?

- Accessory genes determine much of a bacteria’s function
- Strain stability determines whether prebiotics or probiotics can have a lasting effect;
- Understand the mechanism underlying new treatment modalities, such as fecal microbial transplant
The SURPI pipeline for pathogen detection.

Samia N. Naccache et al. Genome Res. 2014;24:1180-1192

© 2014 Naccache et al.; Published by Cold Spring Harbor Laboratory Press

Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing

Wilson,...Chiu

NEJM 2014
Human DNA Admixture

- Important when dealing with human-derived samples
- Ethically, projects should attempt to filter human subject sequences before submission to public databases
- This is actually harder than it sounds

Topic 5: Where is sequencing technology now?

Now: Illumina MiSeq generates 2x300 bp paired end for amplicon and bacterial whole-genome sequencing. HiSeq generates 200,000,000 reads/lane for metagenomics. PacBio for long reads both for complete microbial genome assembly and shotgun metagenomics to scaffold reads.
Any new technology on the horizon before you give this talk again in 2 years?

MinION from Oxford Nanopore
Long read sequencing, Portable small cell, Still high error rate, Requires internet hookup (currently). Could be used for fast diagnosis (think Ebola) in the field.

Sequencing is just the start...
Koch’s Postulates: The basis for assigning causality to an infectious disease.
1 microbe => 1 disease

- Microorganism abundant in diseased hosts and absent in healthy hosts.
- Microorganism isolated from diseased host and grown in pure culture.
- Cultured microorganism should cause disease when introduced into a healthy host.
- Microorganism must be re-isolated from diseased experimental host.
Adapting Koch’s postulates to include microbial community

Byrd and Segre
Science 2016