Microbiome Day

Working Discussion Topics

General Outline of the Discussion Topics

- Topics:
  1) **Study Design**
  2) **Sample Collection and Metadata**
  3) **Sample Processing and Sequencing**
  4) **Bioinformatic Analysis**

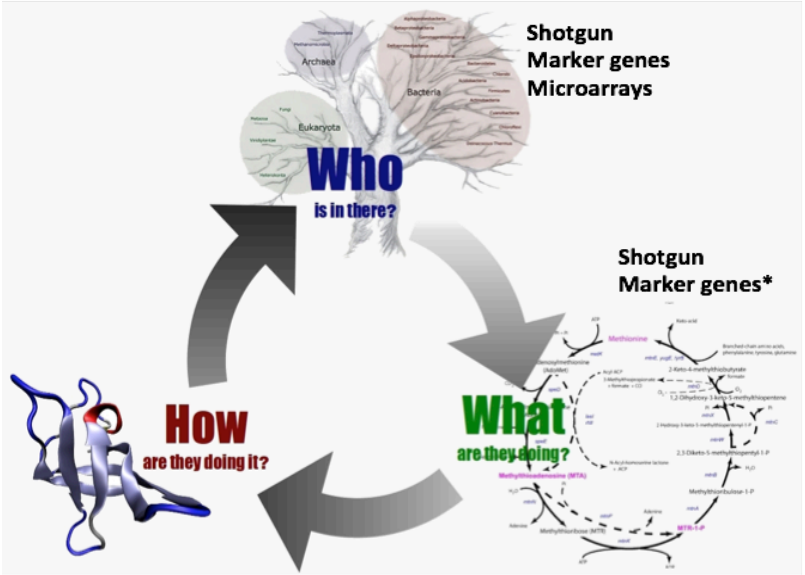
- We’re considering writing a paper with general recommendations discussed within each topic

- We’re looking for people (1 or 2) interested in writing and coordinating the recommendations of each topic

Discussion 1 – Study Design

- **The age old question: Whole genome shotgun or 16S rRNA sequencing?**

  Factors you should consider:
  - Number of samples
  - Cost
  - Presence of host DNA in your sample
  - What's the question you're trying to answer? (Who's there? What are they doing?)
  - How much time do I have to get results? (deadlines)



Discussion 1 – Study Design

- **Removing human host DNA from samples:**
  https://www.qiagen.com/shop/sample-technologies/dna-sample-technologies/genomic-dna/qiaamp-dna-microbiome-kit/
  https://www.neb.com/products/e2612-nebnext-microbiome-dna-enrichment-kit

- **Studying human or animal samples?**
  Depending on your hypothesis, animal studies may be more suitable.
  Mice microbiome =! Human microbiome.
  How animals are bred and raised is the most important source of confounding factors.
  The maternal effect determines the specific suite of microbes available to colonize a host.
  Randomization of treatments across litters/cages becomes an important aspect of experimental design.

- **What is a good sample size for my study?**
  No one knows, but the more, the better. The human microbiome has high inter-individual variability. Take into account different variables: smoking, drinking, diet, health status, etc.
  Estimating the study size:
  https://github.com/biocore/Evident

- **Using the HMP data as a control group – why not?**
  If protocols are compatible and the effect size (i.e., differences between controls and subjects) of the study is large.

Discussion 1 – Study Design

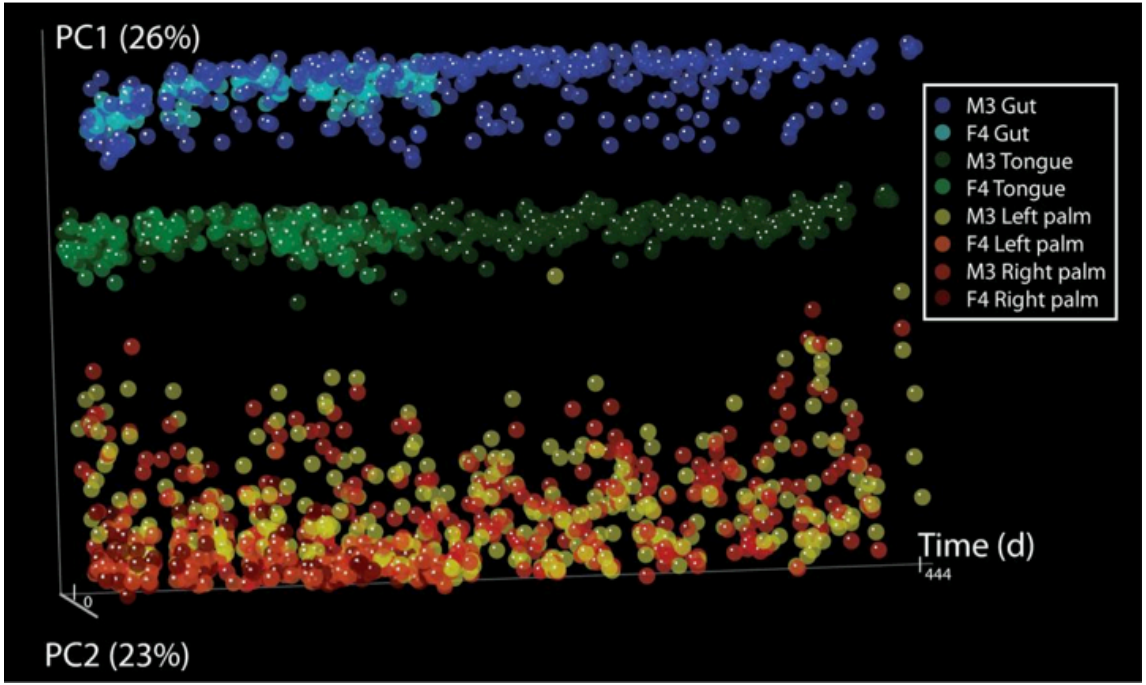
- **Time series samples or snapshot samples?**
  Obtain more samples with a single snapshot or less samples within a time series?

BioMed Central The Open Access Publisher | Genome Biology
this article | search | submit a manuscript | register

Genome Biol. 2011; 12(5): R50. PMCID: PMC3271711
Published online 2011 May 30. doi: 10.1186/gb-2011-12-5-r50

Moving pictures of the human microbiome

J Gregory Caporaso,[1] Christian L Lauber,[2] Elizabeth K Costello,[3] Donna Berg-Lyons,[2] Antonio Gonzalez,[4] Jesse Stombaugh,[1] Dan Knights,[4] Pawel Gajer,[5] Jacques Ravel,[5] Noah Fierer,[2,6] Jeffrey I Gordon,[7] and Rob Knight[1,8]



Discussion 2 – Sample Collection and Metadata

- **What type of sample should I work with? Tissue? Fluids? Feces? Using swabs? Vials?**
  Factors you should consider:
  - Your hypothesis
  - Sequencing approach
  - Host DNA content

- **What metadata should I collect?**
  As much as possible or as many as are pertinent to your research question.

- **Influences the microbiota composition of the gut and other body sites:**
  Antibiotics
  Diet
  Body mass index
  Age
  Pregnancy
  Ethnicity

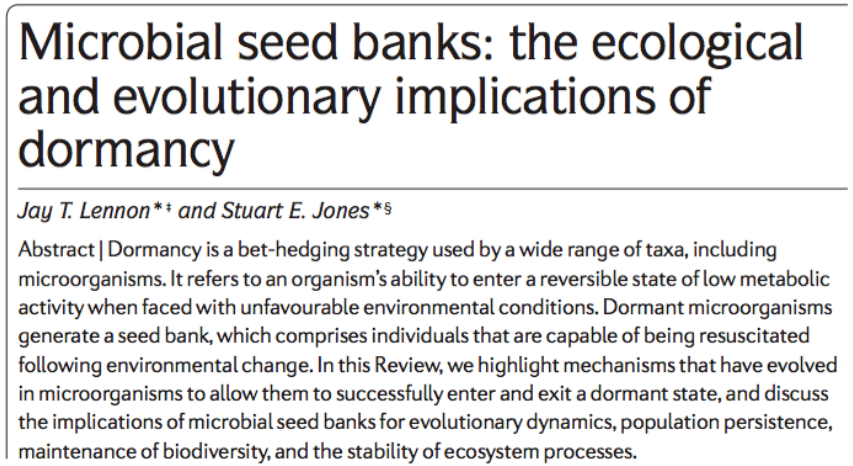
- **Metadata may be study specific, but we may want to have a central core, for all projects (making the data/ samples useful as controls for other studies)**

Discussion 3 – Sample Processing and Sequencing

- **Using mock communities to validate 16S rRNA experimental procedures**
  http://www.beiresources.org/Catalog/otherProducts/HM-782D.aspx

- **Contamination:** Critical for samples with low microbial DNA concentrations.

- **Choice of RNA vs DNA:**
  Microbial seed banks -> Active members of the community vs all members



Microbial seed banks: the ecological and evolutionary implications of dormancy

Jay T. Lennon‡ and Stuart E. Jones*§*

Abstract | Dormancy is a bet-hedging strategy used by a wide range of taxa, including microorganisms. It refers to an organism's ability to enter a reversible state of low metabolic activity when faced with unfavourable environmental conditions. Dormant microorganisms generate a seed bank, which comprises individuals that are capable of being resuscitated following environmental change. In this Review, we highlight mechanisms that have evolved in microorganisms to allow them to successfully enter and exit a dormant state, and discuss the implications of microbial seed banks for evolutionary dynamics, population persistence, maintenance of biodiversity, and the stability of ecosystem processes.

Discussion 3 – Sample Processing and Sequencing

- **What is the best DNA/RNA extraction method?**

  Mechanical Lysis

  Chemical Lysis (New enzyme kits: Lysozyme, Mutanolysin, achromopepdiase, Lysostaphin, Liabase, chitinase, lyticase, and potentially proteinase K)

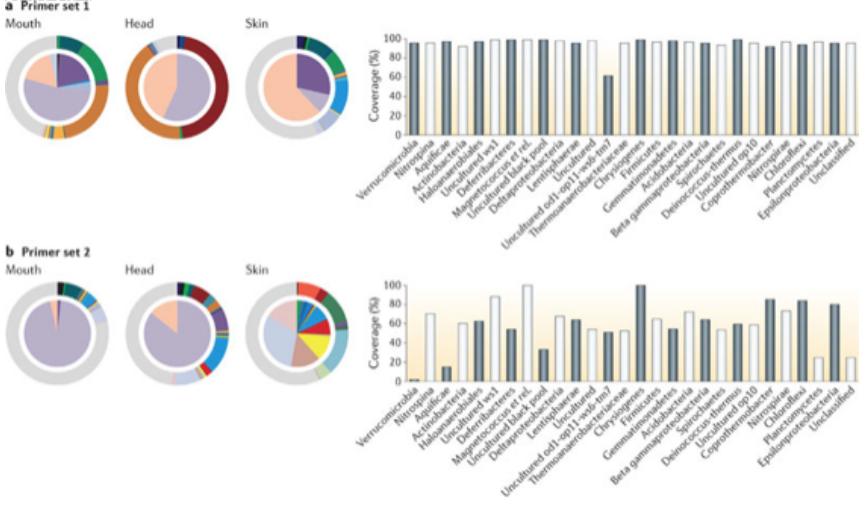
- **For 16s rRNA sequencing - Source of PCR amplification biases:**

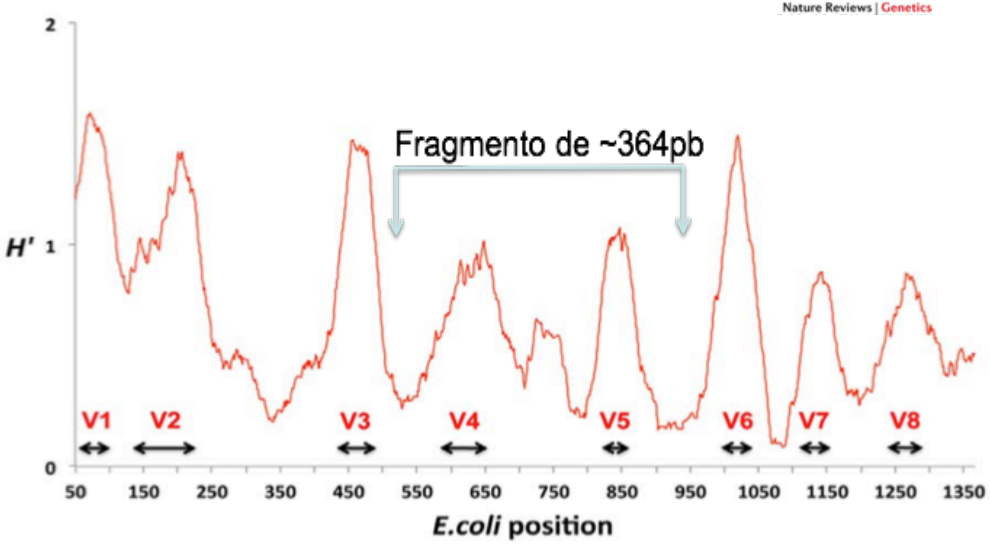
  Number of cycles

  Low DNA template concentration

  Choice of primers:

  - compatibility with other studies
  - primer specificity
  - the choice of region is more important than the length of the amplicon





	SILVA Database	RDP Database
Taxonomy	**Coverage (%)**	**Coverage (%)**
Eubacteria	*84.4*	*52.1*
Acidobacteria	*91.9*	*32.2*
Actinobacteria	*87.4*	*64.0*
Bacteroidetes	*84.1*	*46.8*
Firmicutes	*86.5*	*62.2*
Lentisphaerae	*46.3*	*62.2*
Proteobacteria	*86.1*	*48.6*
Spirochaetes	*76.1*	*59.3*
Verrucomicrobia	*21.6*	*10.9*

Discussion 3 – Sample Processing and Sequencing

- **What sequencing platform to use?**

  1st (Sanger), 2nd (Roche/454), 3rd (Illumina/PGM) and 4th (PacBio) generation sequencing platforms.

  Factors to consider:

  - Long reads vs short reads
  - Sequencing depth -> How much is good enough?
  - Methylation analysis using PacBio SMRT
  - Cost
  - Throughput
  - Sequencing Errors (indels)

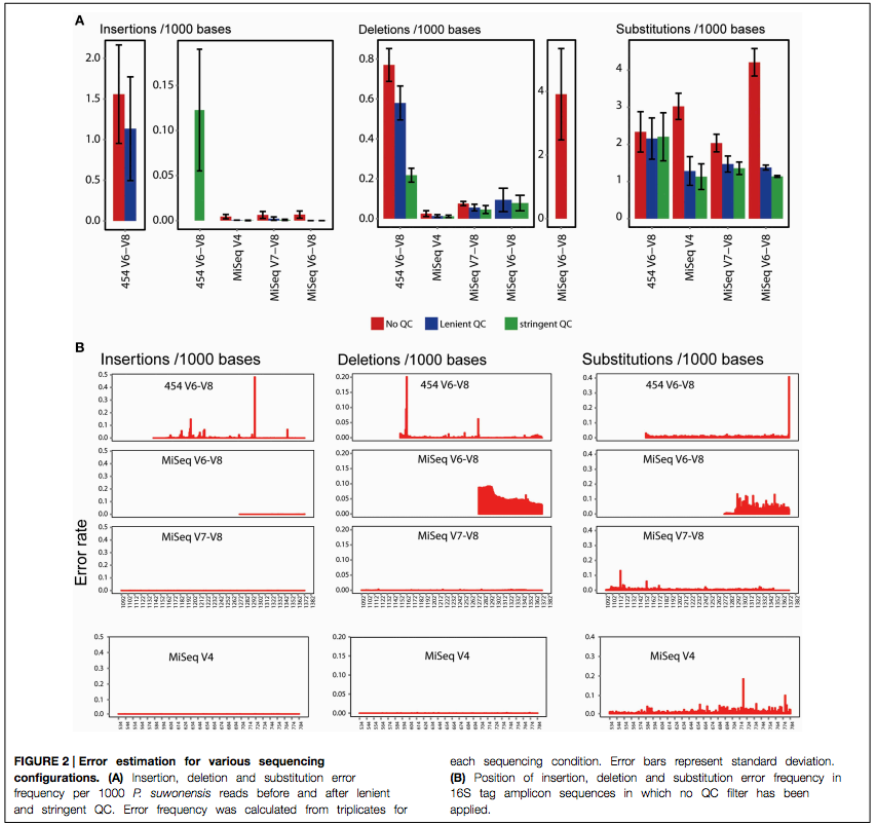


FIGURE 2 | Error estimation for various sequencing configurations. (A) Insertion, deletion and substitution error frequency per 1000 *P. suwonensis* reads before and after lenient and stringent QC. Error frequency was calculated from triplicates for each sequencing condition. Error bars represent standard deviation. **(B)** Position of insertion, deletion and substitution error frequency in 16S tag amplicon sequences in which no QC filter has been applied.

Discussion 4 – Bioinformatic Analysis

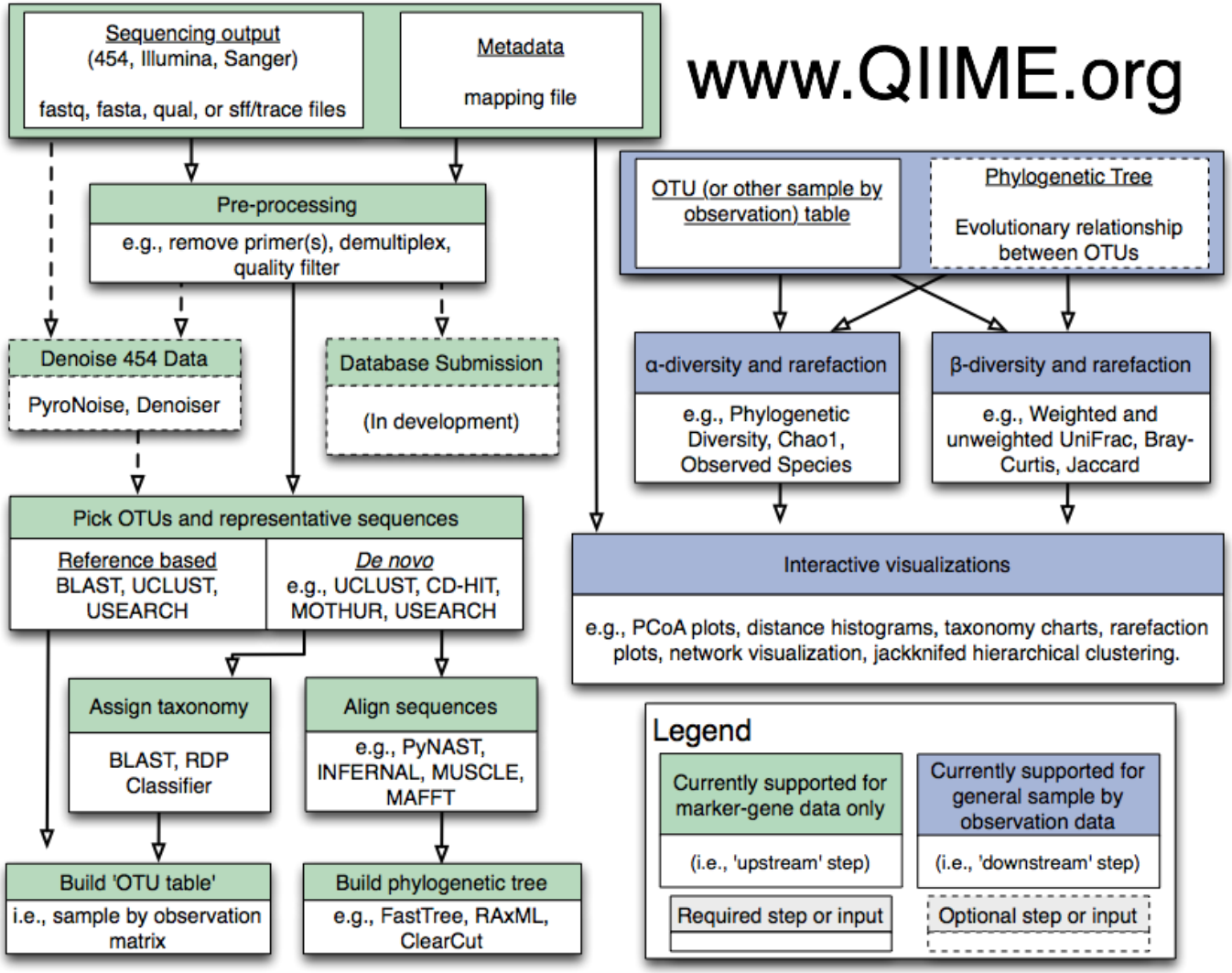
- **Which software to use (16s rRNA):**
  Qiime
  Mothur

- **Using web based-software:**
  MG-Rast
  IMG/M

- **Which assembler to use (whole genome shotgun):**
  IDB-Meta
  Soap *de novo*

- **Which taxonomic classifier to use (whole genome shotgun):**
  MyTaxa
  Kraken

- **How to quality filter your sequences:**
  Size
  Phred quality


Sequencing output
(454, Illumina, Sanger)
fastq, fasta, qual, or sff/trace files
Metadata
mapping file
www.QIIME.org
Pre-processing
e.g., remove primer(s), demultiplex, quality filter
OTU (or other sample by observation) table
Phylogenetic Tree
Evolutionary relationship between OTUs
Denoise 454 Data
PyroNoise, Denoiser
Database Submission
(In development)
α-diversity and rarefaction
e.g., Phylogenetic Diversity, Chao1, Observed Species
β-diversity and rarefaction
e.g., Weighted and unweighted UniFrac, Bray-Curtis, Jaccard
Pick OTUs and representative sequences
Reference based
BLAST, UCLUST, USEARCH
De novo
e.g., UCLUST, CD-HIT, MOTHUR, USEARCH
Interactive visualizations
e.g., PCoA plots, distance histograms, taxonomy charts, rarefaction plots, network visualization, jackknifed hierarchical clustering.
Assign taxonomy
BLAST, RDP Classifier
Align sequences
e.g., PyNAST, INFERNAL, MUSCLE, MAFFT
Legend
Currently supported for marker-gene data only
(i.e., 'upstream' step)
Currently supported for general sample by observation data
(i.e., 'downstream' step)
Required step or input
Optional step or input
Build 'OTU table'
i.e., sample by observation matrix
Build phylogenetic tree
e.g., FastTree, RAxML, ClearCut

Discussion 4 – Bioinformatic Analysis

- **Data reproducibility:**
  Creating standard operating procedures (SOPs)
  IPython notebooks for analysis and publications

- **Data sharing:**
  Creating a sequence database

- **How do I validate my results?**
  Technical validation
  Biological validation

- **Any other suggestions?**