ECOGEOR Workshop 2: 
Introduction to Env ‘Omics
Metagenomic Experimental Design
Ben Tully
Bioinformatic Specialist
Center for Dark Energy Biosphere Investigations
University of Southern California
Background - Terms

- Metagenomics - the random sampling and sequencing of genomic material from an environment
- Contigs - contiguous stretches of DNA composed of the assembly of smaller fragments
The First Metagenomes

Inserted DNA into different sized carrier molecules ranging from small (plasmids) to medium (fosmids) and large (BACs)
Sargasso Sea (2004)

Venter et al 2004
Acid Mine Drainage

Tyson et al 2004

FeS₂ + 14 Fe³⁺ + 8 H₂O → 15 Fe²⁺ + 2 SO₄²⁻ + 16 H⁺
High Throughput Sequencing Methods

Pyrosequencing provides more sequence for less money

- 15% of the Sargasso Sea metagenome assembled into contigs >10kb
- Pyrosequencing metagenomes struggled with assemblies
Peru Margin Sediments (2008)

Biddle et al 2008
Better High Throughput Assemblers and Binning Methods

Assemblers designed for small read inputs = better contigs

Algorithms that incorporated as much data as possible to cluster contigs with similar features

• Tetranucleotide frequencies, %GC, Amino acid usage, Coverage
Group II Euryarchaeaea (2012)
Current Results (2015)

Brown et al 2015
Current Results (2015)

Baker et al 2015

Hug et al 2015
Protocol Design

- High quality DNA
- Size of insert
- Amplification
- Depth of sequencing
Design Considerations - 1

- High quality DNA
  - Crucial step - especially from low biomass or rare samples
  - Clean – no humics or metals
  - High molecular weight

- Size of insert
  - Trade off between quality assessment and insert size
  - Larger insert sizes increase potential of spanning repeat regions
    - Increasing contig fidelity
    - Creating scaffolds from paired contigs
Design Considerations - 2

- Amplification
  - Acceptable: Random linear amplification
  - Ideal: Sufficient DNA extraction
    - Illumina still recommends 1µg
  - DO NOT USE Rolling Circle Amplification (MDA)
Design Considerations - 3

Depth of Sequencing

1. Low depth (1 MiSeq run / portion of 1 HiSeq lane)
   a. Good for supporting extensive geochemical data
   b. Provides genes of interest, full-length 16S, abundance measures
   c. Taxonomy/phylogeny of gene, but not organism

2. High depth (Multiple MiSeq runs / ½-1 HiSeq lane)
   a. Good for genomic reconstructions
   b. Binning
   c. Variable coverage between samples