

# Experimental Design Microbial Sequencing

Matthew L. Settles

Genome Center Bioinformatics Core

University of California, Davis

[settles@ucdavis.edu](mailto:settles@ucdavis.edu); [bioinformatics.core@ucdavis.edu](mailto:bioinformatics.core@ucdavis.edu)

# General rules for preparing and experiment/ samples

- Prepare more samples than you are going to need, i.e. expect some will be of poor quality, or fail
- Preparation stages should occur across all samples at the same time (or as close as possible) and by the same person
- Spend time practicing a new technique to produce the highest quality product you can, reliably
- Quality should be established using Fragment analysis traces (pseudo-gel images, RNA RIN > 7.0)
- DNA/RNA should not be degraded
  - 260/280 ratios for RNA should be approximately 2.0 and 260/230 should be between 2.0 and 2.2. Values over 1.8 are acceptable
- Quantity should be determined with a Fluorometer, such as a Qubit.

# Sample preparation

In high throughput biological work (Microarrays, Sequencing, HT Genotyping, etc.), what may seem like small technical details introduced during sample extraction/preparation can lead to large changes, or technical bias, in the data.

Not to say this doesn't occur with smaller scale analysis such as Sanger sequencing or qRT-PCR, but they do become more apparent (seen on a global scale) and may cause significant issues during analysis.

# Be Consistent

**BE CONSISTENT ACROSS ALL SAMPLES!!!**

<https://www.illumina.com/systems/sequencing-platforms/miseq/specifications.html>

# Illumina MISEQ SEQUENCING

## Cluster Generation and Sequencing

	MiSeq Reagent Kit v2					MiSeq Reagent Kit v3	
Read Length	1 x 36 bp	2 x 25 bp	2 x 150 bp	2 x 250 bp	2 x 75 bp	2 x 300 bp	
Total Time*	~4 hrs	~5.5 hrs	~24 hrs	~39 hrs	~21 hrs	~56 hrs	
Output	540–610 Mb	750–850 Mb	4.5–5.1 Gb	7.5–8.5 Gb	3.3–3.8 Gb	13.2–15 Gb	

	MiSeq Reagent Kit v2 Micro		MiSeq Reagent Kit v2 Nano	
Read Length	2 x 150 bp	2 x 250 bp	2 x 150 bp	2 x 150 bp
Total Time*	~19 hrs	~28 hrs	~17 hrs	~17 hrs
Output	1.2 Gb	500 Mb	300 Mb	300 Mb

\* Total time includes cluster generation, sequencing, and base calling on a MiSeq System enabled with dual-surface scanning.

## Reads Passing Filter\*\*

	MiSeq Reagent Kit v2	MiSeq Reagent Kit v3	MiSeq Reagent Kit v2 Micro	MiSeq Reagent Kit v2 Nano
Single Reads	12–15 million	22–25 million	4 million	1 million
Paired-End Reads	24–30 million	44–50 million	8 million	2 million

\*\* Install specifications based on Illumina PhiX control library at supported cluster densities (865–965 k/mm<sup>2</sup> clusters passing filter for v2 chemistry and 1200–1400 k/mm<sup>2</sup> clusters passing filter for v3 chemistry). Actual performance parameters may vary based on sample type, sample quality, and clusters passing filter.

## Quality Scores†

MiSeq Reagent Kit v2	MiSeq Reagent Kit v3
> 90% bases higher than Q30 at 1 x 36 bp	> 85% bases higher than Q30 at 2 x 75 bp
> 90% bases higher than Q30 at 2 x 25 bp	> 70% bases higher than Q30 at 2 x 300 bp
> 80% bases higher than Q30 at 2 x 150 bp	
> 75% bases higher than Q30 at 2 x 250 bp	

† A quality score (Q-score) is a prediction of the probability of an error in base calling. The percentage of bases > Q30 is averaged across the entire run.



<https://www.illumina.com/systems/sequencing-platforms/hiseq-3000-4000/specifications.html>

# Illumina HiSeq Sequencing

## Performance Parameters

	HiSeq 3000 System	HiSeq 4000 System
No. of Flow Cells per Run	1	1 or 2
Data Yield - 2 × 150 bp	650–750 Gb	1300–1500 Gb
Data Yield - 2 × 75 bp	325–375 Gb	650–750 Gb
Data Yield - 1 × 50 bp	105–125 Gb	210–250 Gb
Clusters Passing Filter (8 lanes per flow cell)	up to 2.5B single reads or 5B paired end reads	up to 5B single reads or 10B PE reads
Quality Scores - 2 × 50 bp	≥ 85% bases above Q30	≥ 85% bases above Q30
Quality Scores - 2 × 75 bp	≥ 80% bases above Q30	≥ 80% bases above Q30
Quality Scores - 2 × 150 bp	≥ 75% bases above Q30	≥ 75% bases above Q30
Daily Throughput	> 200 Gb	> 400 Gb
Run Time	< 1–3.5 days	< 1–3.5 days
Human Genomes per Run*	up to 6	up to 12
Exomes per Run†	up to 48	up to 96
Transcriptomes per Run‡	up to 50	up to 100

Install specifications based on Illumina PhiX control library at supported cluster densities (between 1310–1524 K/mm<sup>2</sup> passing filter). Run times correspond to sequencing only. Performance may vary by sample quality, cluster density, and other experimental factors.

\*Assumes >30x coverage of a human genome.

†Assumes 100x coverage with 80% on target using 2 × 75 bp reads.

‡Assumes 50 million reads per sample.



# Sequencing Depth

- The first and most basic question is how many base pairs of sequence data will I get

Factors to consider are:

- 1. Number of reads being sequenced
- 2. Read length (if paired consider them as individuals)
- 3. Number of samples being sequenced
- 4. Expected percentage of usable data

$$bpPerSample = \frac{readLength * readCount}{sampleCount} * 0.8$$

- The number of reads and read length data are best obtained from the manufacturer's website (search for specifications) and always use the lower end of the estimate.

# Genomic Coverage

**Once you have the number of base pairs per sample you can then determine expected coverage**

Factors to consider then are:

1. Length of the genome
2. Any extra-genomic sequence (ie mitochondria, virus, plasmids, etc.). For bacteria in particular, these can become a significant percentage

$$\frac{ExpectedCoverage}{sample} = \frac{\frac{(readLength * numReads) * 0.8}{numSamples} * num.lanes}{TotalGenomicContent}$$

# Metagenomics Sequencing

## Considerations (when a literature search turns up nothing)

- Proportion that is host (non-microbial genomic content)
- Proportion that is microbial (genomic content of interest)
- Number of species
- Genome size of each species
- Relative abundance of each species

### The back of the envelope calculation

$$\frac{\text{numReads}}{\text{sample}} = \frac{\text{Coverage} * (\text{AverageGenomeSize})}{\text{ReadLen} * \text{DilutionFactor} * (1 - \text{hostProportion})} * \frac{1}{0.8}$$

# Sequencing Depth – Counting based experiments

- Coverage is determined differently for "Counting" based experiments (RNAseq, amplicons, etc.) where an expected number of reads per sample is typically more suitable.
- The first and most basic question is how many reads per sample will I get  
Factors to consider are (per lane):
  1. Number of reads being sequenced
  2. Number of samples being sequenced
  3. Expected percentage of usable data
  4. Number of lanes being sequenced

$$\frac{\text{reads}}{\text{sample}} = \frac{\text{reads.sequenced} * 0.8}{\text{samples.pooled}} * \text{num.lanes}$$

- Read length, or SE vs PE, does not factor into sequencing depth.

# Amplicon Sequencing (Communities, genotyping)

## Considerations

- Number of reads being sequenced
- Proportion that is diversity sample (e.g. PhiX)
- Number of samples being pooled in the run

## The back of the envelope calculation

$$\frac{\text{reads}}{\text{sample}} = \frac{\text{reads\_sequenced} * (1 - \text{diversity\_sample})}{\text{num\_samples}}$$

### example

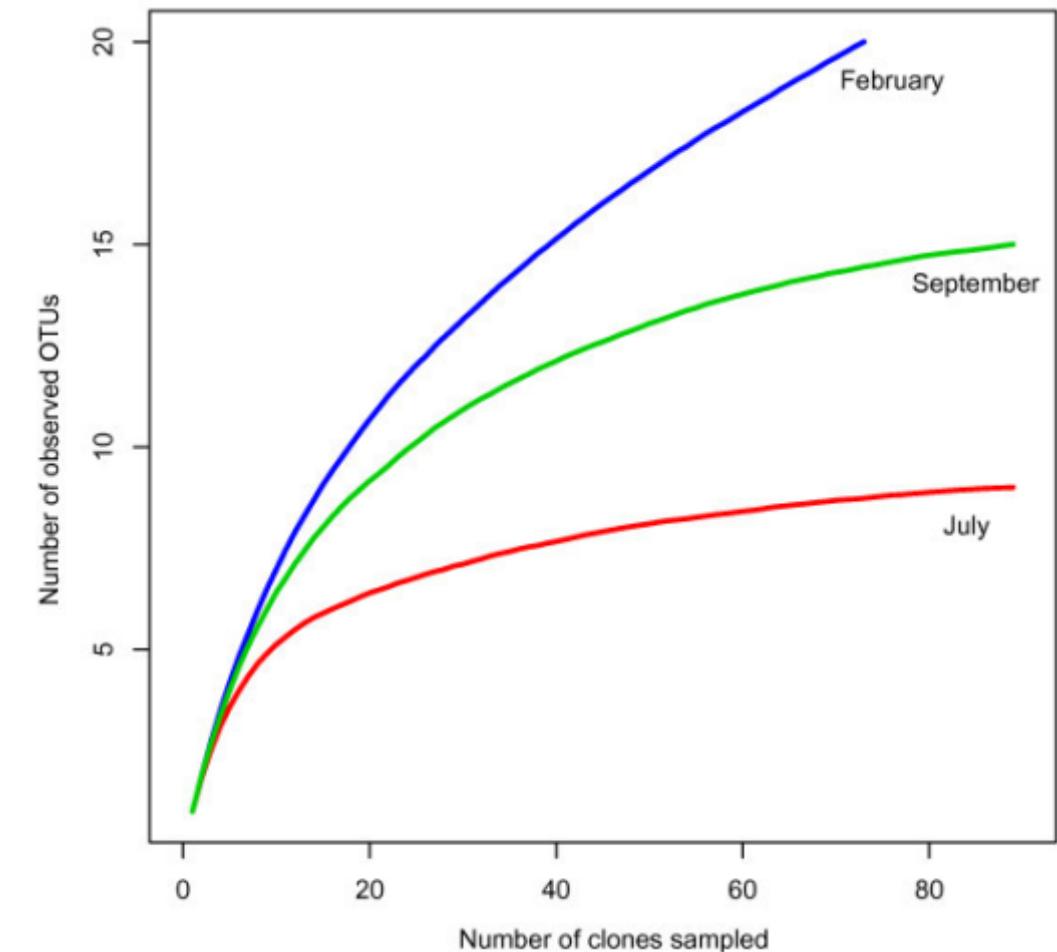
$$\frac{102,000}{\text{sample}} = \frac{18e6 * (1 - 0.15)}{150}$$

## Recommendations

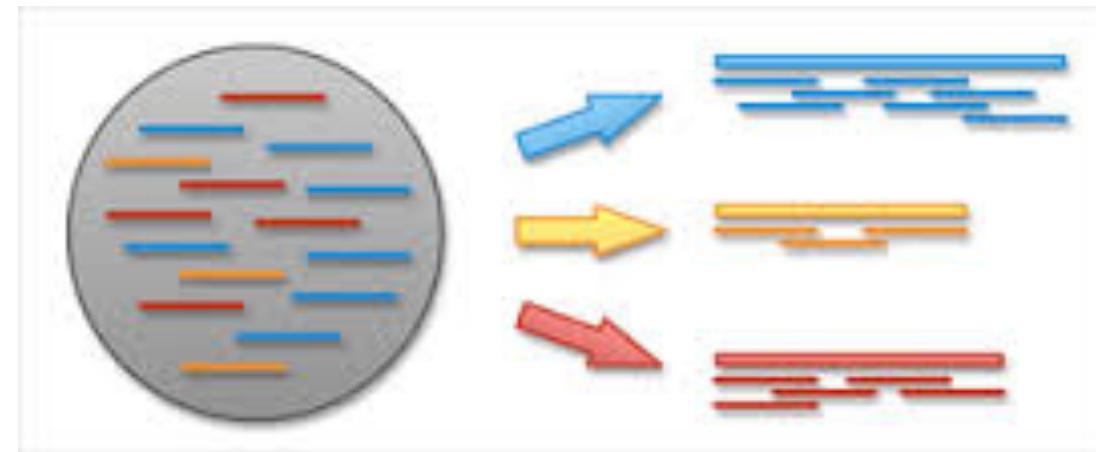
- Illumina ‘recommends’ 100K per sample
- I’ve used 30K per sample historically, others are fine with 3K per sample
- Really should have as many reads as your experiment needs

# How Much? Community Rarefaction curves

- 'Deep' sequence a number of test samples  
amplicons: ~ 1M+ reads.  
metagenomics: 1 full HiSeq lane
- Plot rarefactions curves of organism identification, to determine if saturation is achieved



# Metagenomics assembly



To determine if you've sequenced 'enough' to re-assemble 'most' of the community member's genetic content, look to what is left over - proportionally

# Amplicons vs. Metagenomics

- Metagenomics
  - Shotgun libraries intended to sequence random genomic sequences from the entire bacterial community.
  - Can be costly per sample (\$500 to multi thousands per sample)
  - Better resolution and sensitivity to characterize the sample
  - Due to cost, can only do relatively few samples
- Amplicon community profiling
  - Sequence only one regions of one gene (e.g. 16s, ITS, LSU)
  - Cheap per sample (at scale, down to \$20/sample)
  - Due to cost, can do many hundreds of samples make more global inferences

# Community Sequencing Designs

- Taxonomic Identification
  - Amplicon based (e.g. 16s variable regions)
  - Shotgun Metagenomics
- Functional Characterization
  - Shotgun Metagenomics
  - Shotgun Metatranscriptomics (active)
- Genome Assembly, Function and Variation
  - Shotgun Metagenomics
  - Shotgun Metatranscriptomics

# Cost Estimation

- DNA/RNA extraction and QA/QC (Bioanalyzer/Gels)
- Metatranscriptomes: Enrichment of RNA of interest and RNA library preparation
  - Library QA/QC (Bioanalyzer and Qubit)
  - Pooling (\$10/library)
- Metagenomes: DNA library preparation
  - Library QA/QC (Bioanalyzer and Qubit)
  - Pooling (\$10/library)
- Community Profiling: PCR reactions
  - Library QA/QC (Bioanalyzer and Qubit/microplate reader)
  - Pooling
- Sequencing (Number of Lanes / runs)
- Bioinformatics (General rule is to estimate the same amount as data generation, i.e. double your budget)

<http://dnatech.genomecenter.ucdavis.edu/prices/>

# Bioinformatics Costs

**Bioinformatics includes:**

1. Storage of data
2. Access and use of computational resources and software
3. System Administration time
4. Bioinformatics Data Analysis time
5. Back and forth consultation/analysis to extract biological meaning

Rule of thumb:

Bioinformatics can and should cost as much (sometimes more) as the cost of data generation.

# Cost Estimation

- Amplicons
  - 384 Samples
    - Amplicon generation (\$20/sample)= \$383/sample = \$4,596
  - Sequencing PE300, target 30K reads per sample
  - Bioinformatics
- Metagenome
  - 12 samples (DNA)
  - Expectations: Host Proportion 40%, use average genome size of eColi,  
Target the 1% and coverage of 20
  - Sequencing PE150
  - Bioinformatics

# Taxonomic Assignment

## KRAKEN

- A taxonomic classifier using k-mers, current db contains > 75Gb of microbial genome data (unique kmers).
- Requires a large server, 128Gb to 256Gb of memory
- Assigns each read to its lowest common ancestor in the tree in a taxonomic tree based on the set of kmers in a read
- Accepts ‘single’ read fasta format (flags for pairs and fastq)
  - Output is ‘unusable’, meant for additional processing
  - Kraken-translate to output taxonomic assignment for every read, output can then be used to build abundance tables
- Kraken-filter will move a read up the tree based on confidence of mapping (loosely based on proportion of kmers)
- Can build your own database
- Kraken-report and kraken-mpa-report for abundance table construction

## MetaPhlAn

- Classifies by using a set of marker genes – measures species abundances

# Assembly

- Many assemblers to choose from and more each day
  - Relatively recent tutorial using cloud computing  
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4496567/>
- Most metagenomics assemblers use kmers
  - Either normalize reads by kmers (remove what appears to be redundant information)
  - Or first bin by kmers (each bin is assumed to be a unique species), then assemble each bin (first normalizing by kmers).
- Map reads back to assembly to estimate coverage/count
- BUT then you have to do some with ambiguous contigs
  - Identify ORFs, marker genes, etc. to characterize gene/taxon content
  - IT IS all about the databases!!

# MG-RAST

The MG-RAST system provides answers to the following scientific questions:

- Who is out there? Identifying the composition of a microbial community either by using amplicon data for single genes or by deriving community composition from shotgun metagenomic data using sequence similarities.
- What are they doing? Using shotgun data (or metatranscriptomic data) to derive the functional complement of a microbial community using similarity searches against a number of databases.
- Who is doing what? Based on sequence similarity searches, identifying the organisms encoding specific functions.
- Finally compare samples to each other

# MG-RAST

- Can upload
  - 16s amplicons
  - Metagenomes/Metatranscriptomes
  - Assembled contigs
  - Raw reads
- Use their resources for analysis, don't have to have your own computational resources
- More of a black box, but can download many of output data options
- Subjected to their philosophy for analysis of metagenomic/transcriptomic data

A downloadable alternative to MG-RAST is MEGAN5

# Take Homes

- Experience and/or literature searches (other peoples experiences) will provide the best justification for estimates on needed depth.
- ‘Longer’ reads are better than short reads.
- Paired-end reads are more useful than single-end reads
- Libraries can be sequenced again, so do a pilot, perform a preliminary analysis, then sequence more accordingly.