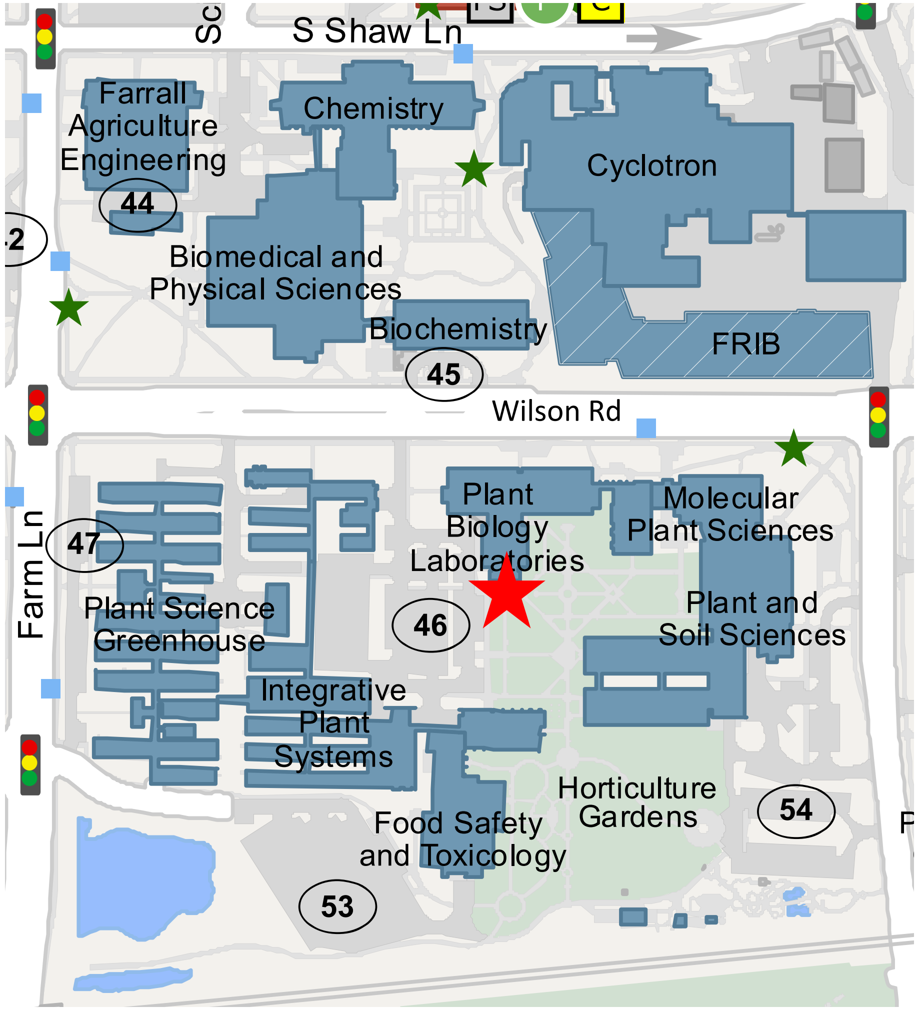
How Much Sequencing Do I Need?

Emily Crisovan – Operations Manager, Genomics Core

Genomics Core Location

Plant Biology Laboratories
S18 and S20
(in the basement)

Sample drop off in
the refrigerator/freezer
in the hallway

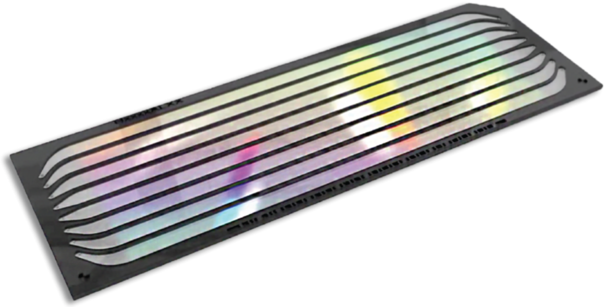


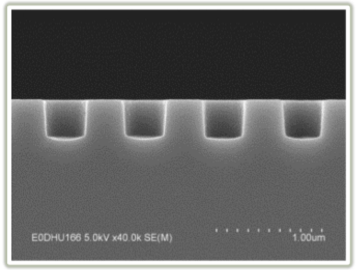
Next-Gen Library Preparation

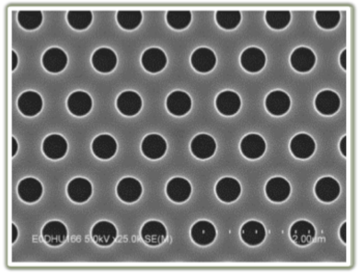
- DNA-seq libraries
  - Single end, paired end
  - Standard or Low input
  - Methylation-seq
- RNA-seq libraries
  - Stranded mRNA, total RNA, small RNA
  - Ribosomal RNA depletion
  - QuantSeq 3’ mRNA
- Amplicon libraries
  - 16S V4, 16S V3V4, custom targets (18S, ITS, etc)

Illumina HiSeq 4000

- Most economical
- Eight lanes per flow cell
- Patterned flow cell
  - Biased towards small inserts
- SE50 and PE150 runs
- Average ~350 million reads per lane







Illumina NextSeq 500

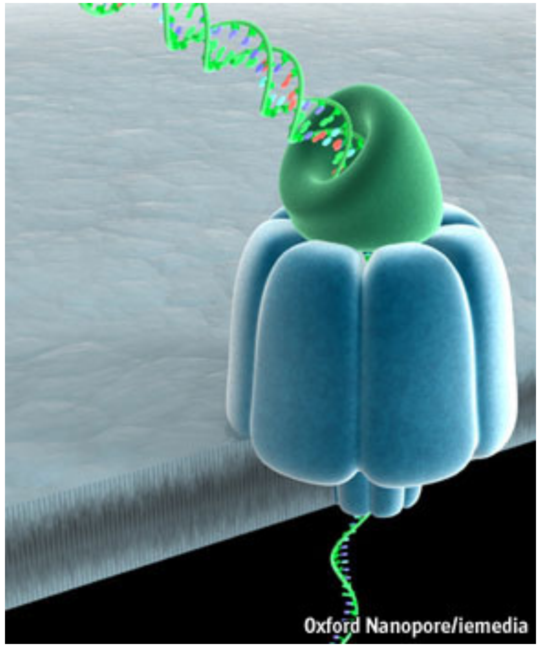
- Less economical than the HiSeq
- One sample per flow cell
- Not a patterned flow cell
  - Less likely to have biases
- Mid and High output flowcells
  - Mid output – 130 million reads
  - High output – 400+ million reads
  - SE75, SE150, PE35, PE75, PE150

Two Illumina MiSeqs

- Least economical, but versatile
- One sample per flow cell
- Not a patterned flow cell
- v2 chemistry
  - Standard, micro, nano outputs
  - 1 to 12 million reads
  - SE50, PE150, PE250
- v3 chemistry
  - 22 million reads
  - SE150, PE75, PE300

Oxford Nanopore GridION

- Long read sequencing
- Genome sequencing
  - Yields of 10 to 15 Gbp common
  - Read N50's of 15 to 40 kbp
- Transcriptome sequencing
  - Yields of 5 to 10 million reads common
  - Full-length and near-full length sequences


Oxford Nanopore/iemedia

How much Sequencing?

Really three questions:

1. How much sequence is required for good experimental design?
2. What type of sequencing run is best?
3. How many lanes of sequencing?

All based on Illumina sequencing options

Experimental Design

What are you sequencing?

Genome

- de novo assembly
- resequencing project
- variant discovery

Transcriptome

- de novo assembly
- gene expression analysis

Whole Meta-Genomes, Small RNAs, ChIP-Seq, Exome Capture, Amplicon Sequencing

What Type of Sequencing Run

Single end or paired end?

What read length?

35 bp, 50 bp, 75 bp, 150 bp, 250 bp, 300 bp
Not all read lengths available on all machines

Assembly of genome or transcriptome?

paired end reads: 150 bp, 250 bp, 300 bp

Counting experiment?

single end reads: 35 bp, 50 bp, 75 bp

https://rtsf.natsci.msu.edu/genomics/pricing/

How Many Lanes of Sequencing

For genome assembly

- answer depends on desired coverage
- new assembly 75X – 100X
- resequencing or variant discovery 10X – 30X
- long-read error correction 20X – 30X – 80X

# lanes required =

desired Gbp / expected Gbp per lane

How Many Lanes of Sequencing

For transcriptome assembly

- number of genes in the genome
- complexity of the transcriptome

# lanes required =

# reads per sample x # of samples / reads per lane

How many Lanes of Sequencing

For gene expression analysis

- counting experiment
    - Gbp not important
    - numbers of reads important
- what is typical in your field
- consider ploidy
- how many replicates

# lanes required =

$$\frac{\text{(minimum \# reads per sample x \# samples x \# replicates x “fudge factor”)}}{\text{\# reads per lane}}$$

Genome Sequencing Example #1

New eukaryotic genome assembly

- 1.2 Gbp genome
- target 80X coverage
- PE 150 reads
- HiSeq 4000 averages 350 million reads/lane

# lanes required =

desired Gbp / expected Gpb per lane

Genome Sequencing Example #1 Calculations

New eukaryotic genome assembly

- 1.2 Gbp genome
- Target 80x coverage
- PE 150
- HiSeq 4000 averages 350 million reads per lane

Desired Gbp?

Genome size * coverage

1.2 Gbp * 80 = 96 Gbp

Expected Gbp per lane?

(# reads * read length) / 1,000,000,000 bp

(350,000,000 reads * 300 bp) / 1,000,000,000 bp = 105 Gbp

How many lanes of sequencing are needed?

desired Gbp / expected Gbp = # of lanes

96 Gbp / 105 Gbp = 0.91 lanes ➔ 1 lane

Genome Sequencing Example #2

New prokaryotic genomes

- 16 different isolates
- 8 Mbp genome
- target 40X coverage
- PE 150 reads
- HiSeq 4000 averages 350 million reads/lane

# lanes required =

desired Gbp / expected Gpb per lane

Genome Sequencing Example #2 Calculations

New prokaryotic genome

- 16 different bacterial isolates
- 8 Mbp genome
- Target 40x coverage
- PE 150
- HiSeq 4000 averages 350 million reads per lane

Desired Gbp?

(Genome size * coverage) * # of samples

8 Mbp * 40 = 320 Mbp (0.32 Gbp) per sample

0.32 Gbp * 16 samples = 5.12 Gbp total

Expected Gbp per lane?

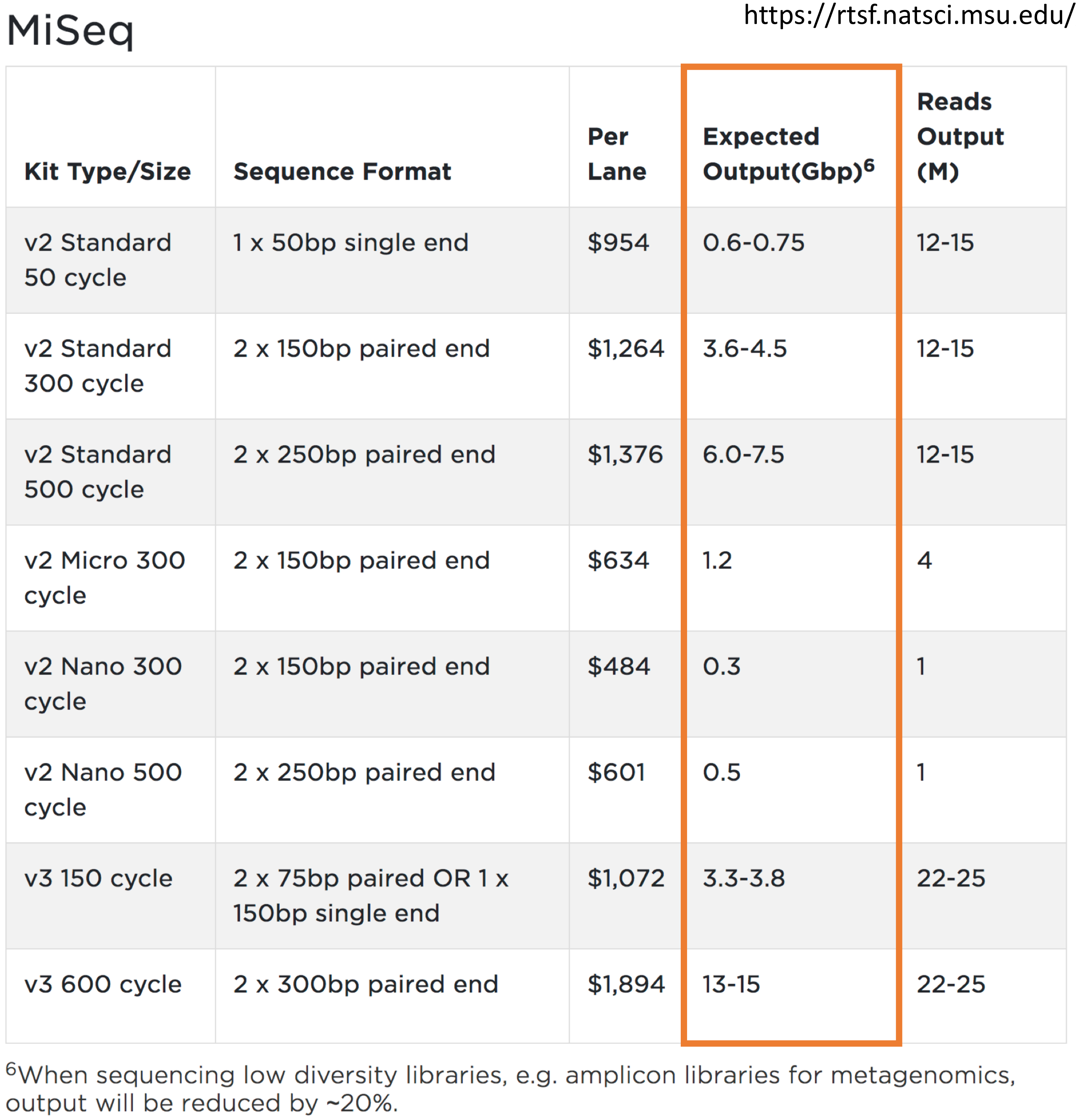
(# reads * read length) / 1,000,000,000 bp

(350,000,000 reads * 300 bp) / 1,000,000,000 bp = 105 Gbp

Is the HiSeq appropriate for this project?

desired Gbp / expected Gbp = # of lanes

5.12 Gbp / 105 Gbp = 0.05 lanes ➔ HiSeq is not appropriate https://rtsf.natsci.msu.edu/genomics/pricing/



MiSeq

Kit Type/Size	Sequence Format	Per Lane	Expected Output(Gbp)[6]	Reads Output (M)
v2 Standard 50 cycle	1 x 50bp single end	$954	0.6-0.75	12-15
v2 Standard 300 cycle	2 x 150bp paired end	$1,264	3.6-4.5	12-15
v2 Standard 500 cycle	2 x 250bp paired end	$1,376	6.0-7.5	12-15
v2 Micro 300 cycle	2 x 150bp paired end	$634	1.2	4
v2 Nano 300 cycle	2 x 150bp paired end	$484	0.3	1
v2 Nano 500 cycle	2 x 250bp paired end	$601	0.5	1
v3 150 cycle	2 x 75bp paired OR 1 x 150bp single end	$1,072	3.3-3.8	22-25
v3 600 cycle	2 x 300bp paired end	$1,894	13-15	22-25

[6] When sequencing low diversity libraries, e.g. amplicon libraries for metagenomics, output will be reduced by ~20%.

Transcript Assembly Example

Goal is transcript assembly

- 25,000 genes, diploid
- target of 60 million reads per sample
- PE150
- HiSeq 4000 averages 350 million reads/lane

# lanes required =

# reads per sample x # of samples / reads per lane

How many different mRNA samples can be loaded into a single lane?

Transcript Sequencing Example Calculations

Transcriptome assembly

- Target 60 million read pairs per sample
- PE 150
- HiSeq 4000 averages 350 million reads per lane

How many different mRNA samples can be prepared and loaded on one lane?

# of read pairs per lane / # of read pairs per sample

350 M read pairs per lane / 60 M read pairs per sample

= 5.8 samples ➔ round down to 5 samples per lane

Gene Expression Example

Gather counts for differential expression analysis

- Mammals: 30 to 50 million reads per sample
- Plants: 25 million reads per sample
- Replicates: 3 to 5
- # samples is experiment-dependent
- SE 50
- HiSeq 4000 averages 350 million reads/lane

# lanes required =

$$\frac{\text{(minimum \# reads per sample x \# samples x \# replicates x “fudge factor”)}}{\text{\# reads per lane}}$$

Gene Expression Example

Gene expression of mammal:

- 6 samples
- 3 replicates
- Target = minimum of 30 million reads each
- SE 50
- HiSeq 4000 averages 350 million reads/lane

# lanes required =

$$\frac{\text{(minimum \# reads per sample x \# samples x \# replicates x “fudge factor”)}}{\text{\# reads per lane}}$$

Gene Expression Sequencing Example Calculations

Gene expression of mammal:

- 6 samples
- 3 replicates
- Target = minimum of 30 million reads each
- SE 50
- HiSeq 4000 averages 350 million reads per lane

# lanes required =

(minimum # reads per sample * # samples * # replicates x fudge factor) / # reads per lane

(30 M reads per sample * 6 samples * 3 replicates) / 350 million reads per lane
= 1.5 lanes ➔ round up to 2 lanes

Gene Expression Example – Fudge Factor

- When RNA-seq libraries are combined into one sequencing lane, libraries will not be sequenced equally
- There will be variation in the number of reads obtained from each library
- We must sequence more than if all libraries produced equal numbers of reads

# lanes required =
minimum reads per sample X # replicates X # samples / reads per lane

Add an extra 10 or 15%