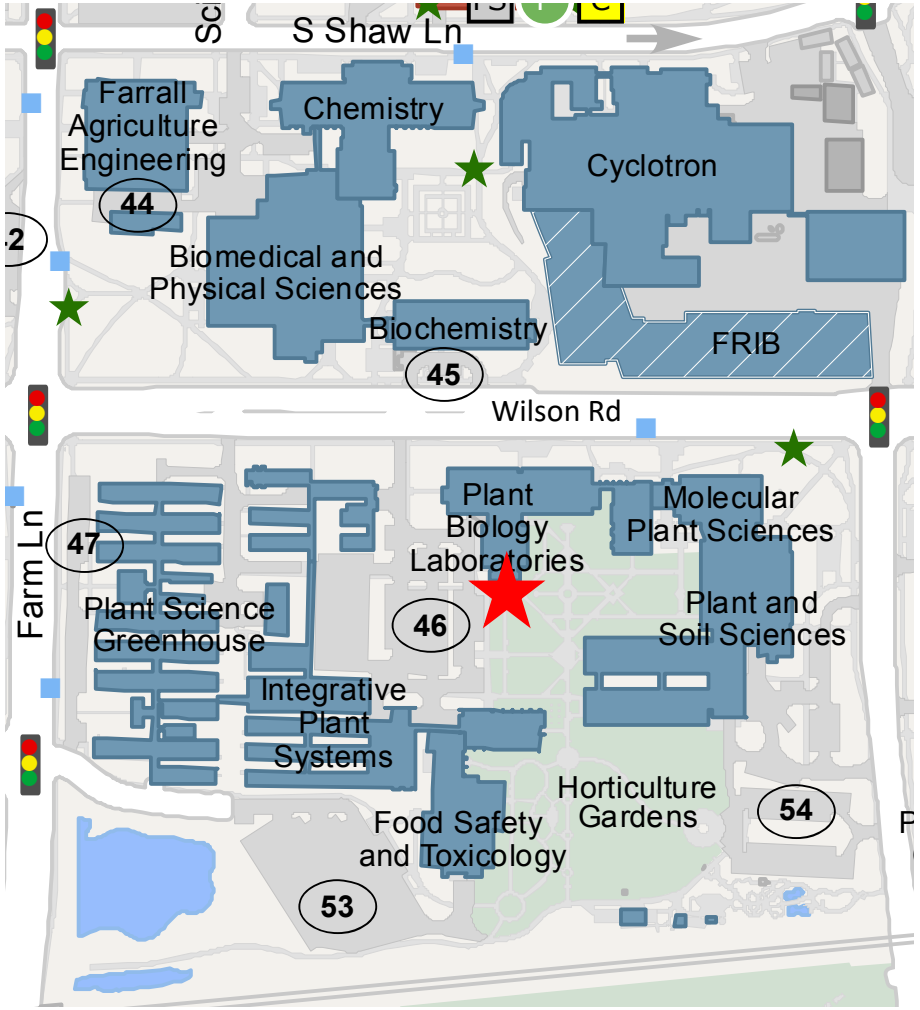
How Much Sequencing Do I Need?

Kevin Childs – Director Genomics Core

Genomics Core Location

Plant Biology Laboratories
S18 and S20
(in the basement)

Sample drop off in
the refrigerator in the
hallway

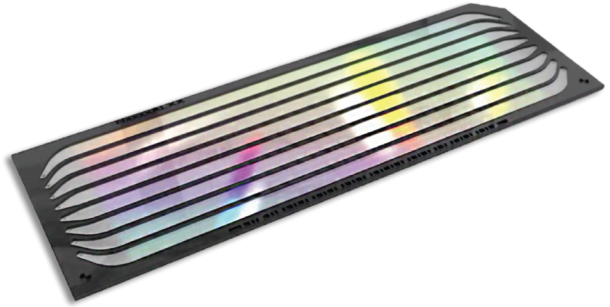


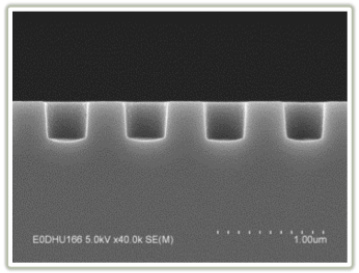
Next-Gen Library Preparation

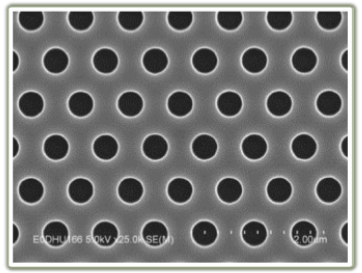
- DNA-seq libraries
  - Single end, paired end
  - Low input
  - Methylation-seq
- RNA-seq libraries
  - Stranded mRNA, total RNA, small RNA
  - Ribosome depletion
  - QuantSeq 3’ mRNA
- Amplicon libraries
  - 16S, 18S, ITS, custom targets

Illumina HiSeq 4000

- Most economical
- Eight lanes per flowcell
- Patterned flowcell
  - Biased towards small inserts
- SE50 and PE150 runs
- Typically 350 million reads per lane





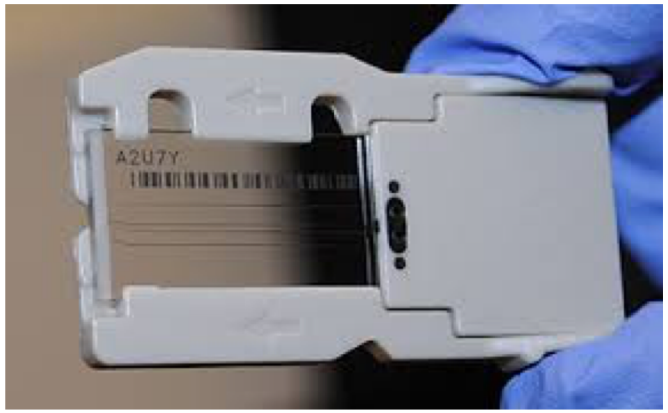


Illumina NextSeq 500

- Less economical than the HiSeq
- One sample per flowcell
- Not a patterned flowcell
  - No biases
- Mid and High output flowcells
  - Mid output – 130 million reads
  - High output – 400+ million reads
  - SE75, SE150, PE35, PE75, PE150

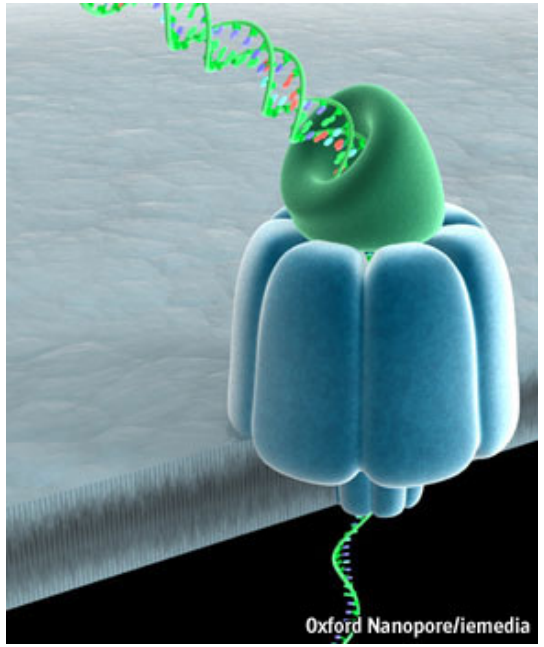
Two Illumina MiSeqs

- Least economical, but versatile
- One sample per flowcell
- Not a patterned flowcell
- 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



How much Sequencing?

- Really three questions
- How much sequence is required for good experimental design?
- What type of sequencing run is best?
- 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 / 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 =

minimum reads per sample X # replicates X # samples X fudge factor / reads per lane

Genome Sequencing Example #1

New eukaryotic genome (reference-guided 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 #2

New prokaryotic genomes

- 12 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

New prokaryotic genomes

- 12 different bacterial isolates
- 8 Mbp genome
- target 40X coverage
- ~~PE 150 reads~~
- ~~HiSeq 4000 averages 350 million reads/lane~~
- MiSeq v2 Standard PE 250 give ~12-15 Gbp

lanes required =

desired Gbp / expected Gpb per lane

Transcript Assembly Example

Goal is transcript assembly

- 25,000 genes
- target of 60 million reads per sample
- PE 150
- HiSeq 4000 averages 350 million reads/lane

lanes required =

samples X # reads desired per sample/ expected # reads per lane

Bonus – How many different mRNA samples can be prepared and loaded into a single 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 =

minimum reads per sample X # replicates X # samples / reads per lane

Gene Expression Example

- 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%