How Much Sequencing Do I Need?

Kevin Childs – Director Genomics Core

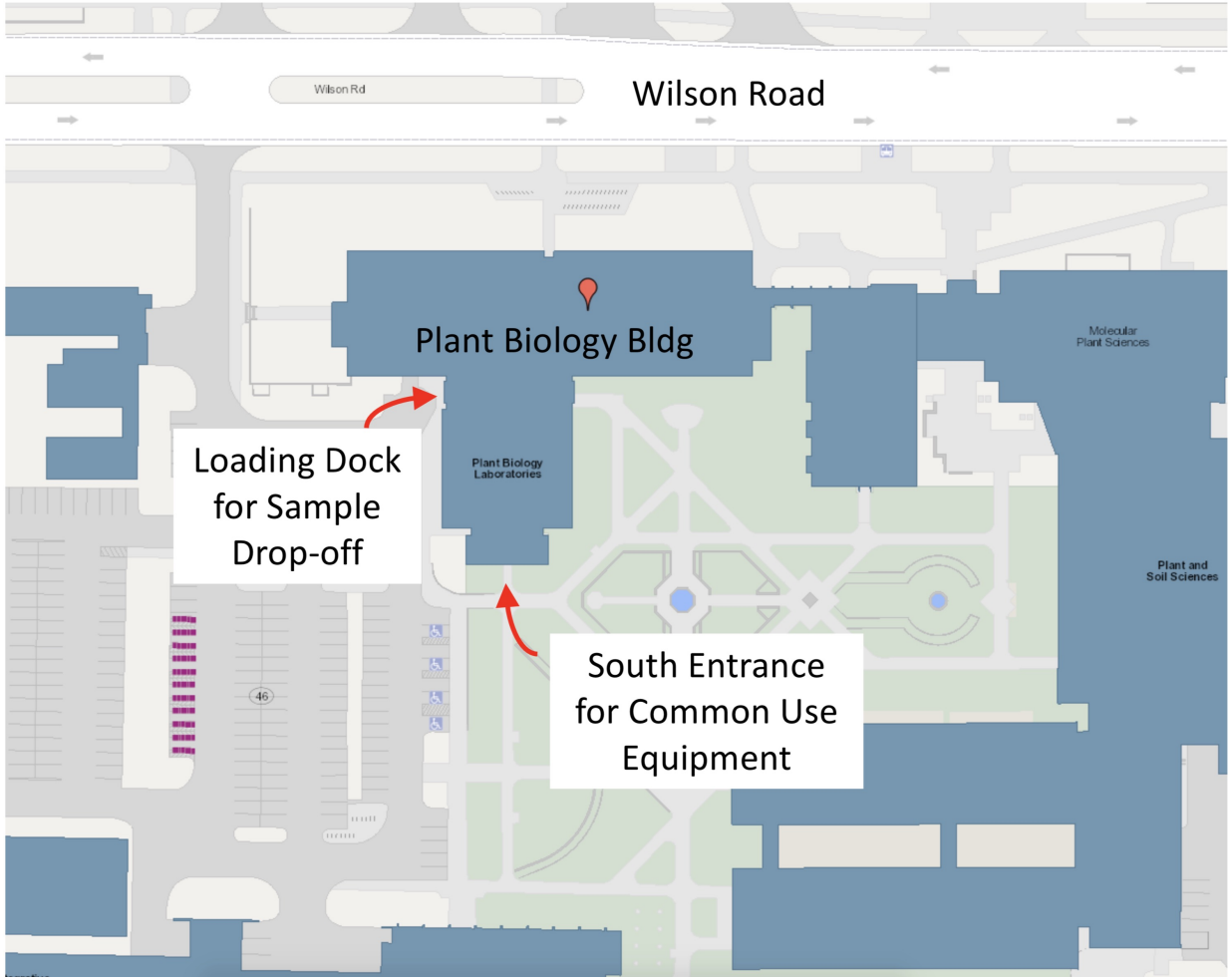
Overview

- Genomics Core details
- Sequencing services
- Illumina instruments
- Basic questions regarding how much sequencing you need
- Examples with solutions
  - Bacterial genome
  - Variant discovery
  - Gene expression – diploid, polyploid
  - Single cell sequencing

Genomics Core Location

Plant Biology Laboratories
S18 and S20
(in the basement)

Sample drop off in the refrigerators at loading dock and in the hallway outside the lab



Contacting the Genomics Core

Email - gtsf@msu.edu

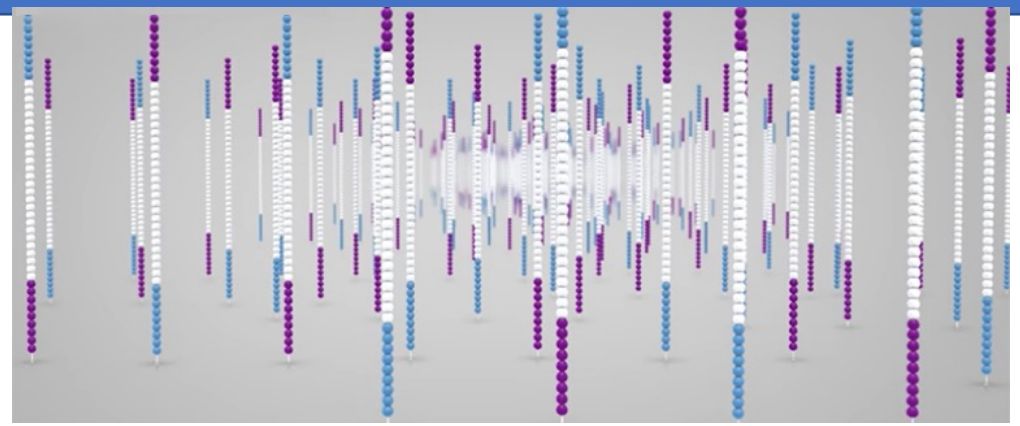
Daily Zoom Office Hour @ 3 PM -
https://bit.ly/3DuBYoY

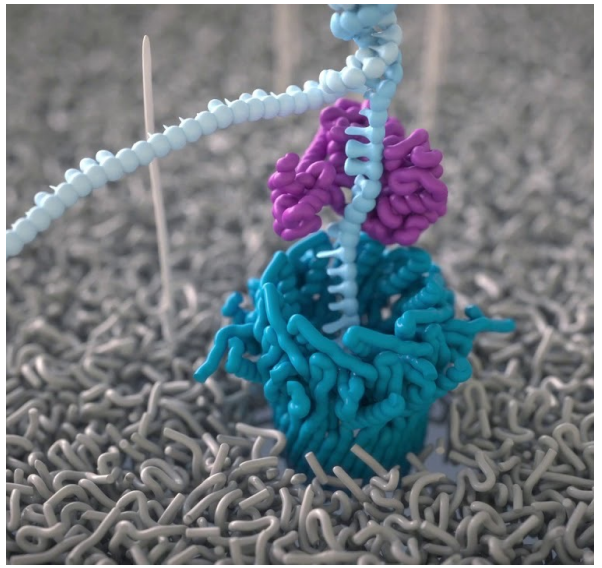
Website -
https://rtsf.natsci.msu.edu/genomics/

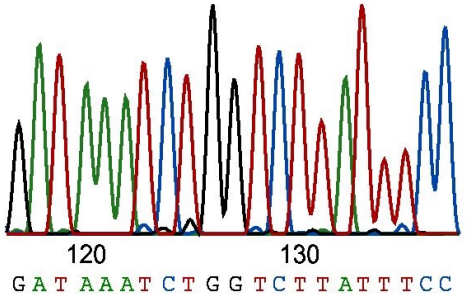
Email List -
https://rtsf.natsci.msu.edu/genomics/email-list/

Sequencing Services at the Genomics Core

- Illumina sequencing
  - NovaSeq 6000
  - MiSeq
- Oxford Nanopore sequencing
  - GridION
  - PromethION
- Sanger sequencing





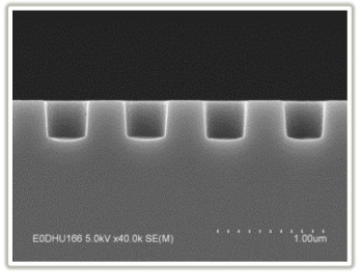


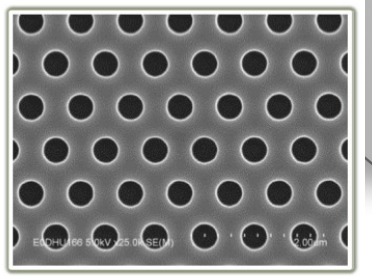
Illumina Library Preparations

- DNA-seq libraries
  - Genomic
  - Low input
  - Methylation-seq
- RNA-seq libraries
  - Stranded mRNA, total RNA, small RNA
  - 10X Genomics single cell seq
  - Ribosome depletion
  - QuantSeq 3’ mRNA
- Amplicon libraries
  - 16S, ITS, custom targets

Illumina NovaSeq 6000

- Most economical
- Two or four lanes per flowcell
- Patterned flowcell
  - Biased towards small inserts
- SE100 and PE150 runs most common
- Output dependent on flow cell type





NovaSeq Run Options

Flow Cell Type	Number lanes per flow cell	Kit Type/Size	Average Gbp per lane	Average number of reads per lane	Equivalent Number HiSeq Lanes	Pricing per Lane	Cost per Gbp
1/10th S4 lane		PE150	60	200,000,000		$641	$10.68
SP	2	100 cycles	36.25	362,500,000	1	$1,656	$45.68
SP	2	200 cycles	75.25	362,500,000	1	$2,107	$28.00
SP	2	300 cycles	112.5	362,500,000	1	$2,424	$21.55
SP	2	500 cycles	181.25	362,500,000	1	$3,405	$18.79
S1	2	100 cycles	75.25	725,000,000	2	$2,717	$36.11
S1	2	200 cycles	149.75	725,000,000	2	$3,620	$24.17
S1	2	300 cycles	225	725,000,000	2	$4,204	$18.68
S2	2	100 cycles	187.5	1,850,000,000	4	$4,651	$24.81
S2	2	200 cycles	375	1,850,000,000	4	$6,049	$16.13
S2	2	300 cycles	562.5	1,850,000,000	4	$6,832	$12.15
S4	4	200 cycles	450	2,250,000,000	6.5	$4,897	$10.88
S4	4	300 cycles	675	2,250,000,000	6.5	$5,846	$8.66
S4	4	35 cycles	78.75	2,250,000,000	6.5	$3,528	$44.80

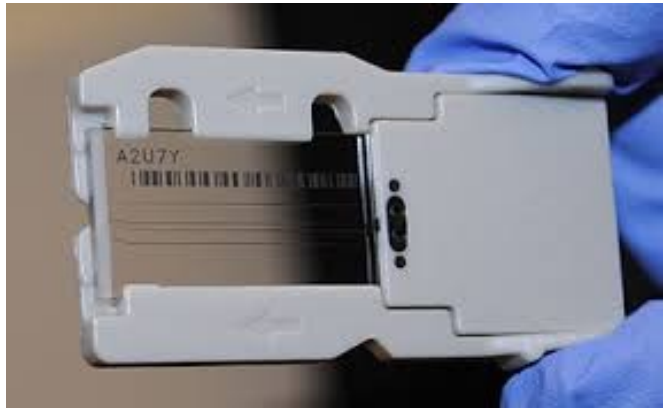
NovaSeq Run Options (Commonly Run)

Flow Cell Type	Number lanes per flow cell	Kit Type/Size	Average Gbp per lane	Average number of reads per lane	Equivalent Number HiSeq Lanes	Pricing per Lane	Cost per Gbp
1/10th S4 lane		PE150	60	200,000,000		$641	$10.68
SP	2	100 cycles	36.25	362,500,000	1	$1,656	$45.68
SP	2	300 cycles	112.5	362,500,000	1	$2,424	$21.55
S1	2	100 cycles	75.25	725,000,000	2	$2,717	$36.11
S4	4	300 cycles	675	2,250,000,000	6.5	$5,846	$8.66

Shared S4 lanes may be used with Core prepared DNA-seq or RNA-seq libraries

Two Illumina MiSeqs

- Much less 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



MiSeq Run Options

Flow Cell Type	Sequence Format	Approximate Output in Gbp	Approximate Output in Reads (Millions)	Cost
V2 Standard 50 cycle	SE50	0.6 to 0.75	12 to 15	$1,064
V2 Standard 300 cycle	PE150	3.6 to 4.5	12 to 15	$1,398
V2 Standard 500 cycle	PE250	6.0 to 7.5	12 to 15	$1,519
V2 Micro 300 cycle	PE150	1.2	4	$696
V2 Nano 300 cycle	PE150	0.3	1	$524
V2 Nano 500 cycle	PE250	0.5	1	$647
V3 150 cycle	PE75 or SE150	3.3 to 3.8	22 to 25	$1,196
V3 600 cycle	PE300	13 To 15	22 to 25	$2,043

How Much Illumina Sequencing?

- What are you sequencing?
  - DNA
  - RNA
- How much sequence is required for good experimental design?
  - What is required for your organism?
  - What is expected by your community?
- Find the sequencing run that meets your needs
  - Sufficient sequence with a bit of room to spare
  - Price vs budget

16S Metagenomics Project

- 200 environmental samples
- 16S V4 amplicons
- One MiSeq V2 Standard PE250 run
- 12 to 15 million read pairs

What are you sequencing?
How much sequence is required?

- Genome
- Think in terms of fold-coverage or total Gbp
- De novo assembly
  - Bacteria only; eukaryotes use long read tech
  - 30 to 50 X the genome size
- Variant discovery from a few individuals
  - 10 to 30 X the genome size
  - More coverage = more variants identified
- Variant discovery from a population
  - Pool-seq
  - As little as 1 X per individual with many samples (~100)

De novo assembly bacterial genome

- 30 to 50X coverage
- Bacterial genomes - 0.1 to 15 Mbp
- *Mycobacterium leprae*
  - 3.2 Mbp
  - 3.2 Mbp x 30X coverage = 96 Mbp

MiSeq Run Options – Target 96 Mbp

Flow Cell Type	Sequence Format	Approximate Output in Gbp	Approximate Output in Reads (Millions)	Cost
V2 Standard 50 cycle	SE50	0.6 to 0.75	12 to 15	$1,064
V2 Standard 300 cycle	PE150	3.6 to 4.5	12 to 15	$1,398
V2 Standard 500 cycle	PE250	6.0 to 7.5	12 to 15	$1,519
V2 Micro 300 cycle	PE150	1.2	4	$696
V2 Nano 300 cycle	PE150	0.3	1	$524
V2 Nano 500 cycle	PE250	0.5	1	$647
V3 150 cycle	PE75 or SE150	3.3 to 3.8	22 to 25	$1,196
V3 600 cycle	PE300	13 To 15	22 to 25	$2,043

NovaSeq Run Options – Target 96 Mbp

Flow Cell Type	Number lanes per flow cell	Kit Type/Size	Average Gbp per lane	Average number of reads per lane	Equivalent Number HiSeq Lanes	Pricing per Lane	Cost per Gbp
1/10th S4 lane		PE150	60	200,000,000		$641	$10.68
SP	2	100 cycles	36.25	362,500,000	1	$1,656	$45.68
SP	2	300 cycles	112.5	362,500,000	1	$2,424	$21.55
S1	2	100 cycles	75.25	725,000,000	2	$2,717	$36.11
S4	4	300 cycles	675	2,250,000,000	6.5	$5,846	$8.66

60 Gbp / 3.2 Mbp = 60,000 Mbp / 3.2 Mbp = 18,750X coverage

Limitations with number of available barcodes

Variant discovery from a few individuals

- 10 to 30 X the genome size
- Atlantic salmon (*Salmo salar*) – 3 Gbp
- 20 individuals x 15X coverage x 3 Gbp = 900 Gbp

NovaSeq Run Options – Target 900 Gbp

Flow Cell Type	Number lanes per flow cell	Kit Type/Size	Average Gbp per lane	Average number of reads per lane	Equivalent Number HiSeq Lanes	Pricing per Lane	Cost per Gbp
1/10th S4 lane		PE150	60	200,000,000		$641	$10.68
SP	2	100 cycles	36.25	362,500,000	1	$1,656	$45.68
SP	2	300 cycles	112.5	362,500,000	1	$2,424	$21.55
S1	2	100 cycles	75.25	725,000,000	2	$2,717	$36.11
S4	4	300 cycles	675	2,250,000,000	6.5	$5,846	$8.66

One S4 PE150 lane = 675 Gbp = 11.25X coverage for 20 specimens

NovaSeq Run Options – Target 900 Gbp

Flow Cell Type	Number lanes per flow cell	Kit Type/Size	Average Gbp per lane	Average number of reads per lane	Equivalent Number HiSeq Lanes	Pricing per Lane	Cost per Gbp
1/10th S4 lane		PE150	60	200,000,000		$641	$10.68
SP	2	100 cycles	36.25	362,500,000	1	$1,656	$45.68
SP	2	300 cycles	112.5	362,500,000	1	$2,424	$21.55
S1	2	100 cycles	75.25	725,000,000	2	$2,717	$36.11
S4	4	300 cycles	675	2,250,000,000	6.5	$5,846	$8.66

Two S4 PE150 lanes = 2 x 675 Gbp = 1350 Gbp =
22.5X coverage for 20 specimens

NovaSeq Run Options – Target 900 Gbp

Flow Cell Type	Number lanes per flow cell	Kit Type/Size	Average Gbp per lane	Average number of reads per lane	Equivalent Number HiSeq Lanes	Pricing per Lane	Cost per Gbp
1/10th S4 lane		PE150	60	200,000,000		$641	$10.68
SP	2	100 cycles	36.25	362,500,000	1	$1,656	$45.68
SP	2	300 cycles	112.5	362,500,000	1	$2,424	$21.55
S1	2	100 cycles	75.25	725,000,000	2	$2,717	$36.11
S4	4	300 cycles	675	2,250,000,000	6.5	$5,846	$8.66

One S4 PE150 lane + one SP PE150 lane = 675 + 112.5 Gbp = 787.5 Gbp = 13.1X coverage for 20 specimens

NovaSeq Run Options – Target 900 Gbp

Flow Cell Type	Number lanes per flow cell	Kit Type/Size	Average Gbp per lane	Average number of reads per lane	Equivalent Number HiSeq Lanes	Pricing per Lane	Cost per Gbp
1/10th S4 lane		PE150	60	200,000,000		$641	$10.68
SP	2	100 cycles	36.25	362,500,000	1	$1,656	$45.68
SP	2	300 cycles	112.5	362,500,000	1	$2,424	$21.55
S1	2	100 cycles	75.25	725,000,000	2	$2,717	$36.11
S4	4	300 cycles	675	2,250,000,000	6.5	$5,846	$8.66

One S4 PE150 lane + 4/10th S4 PE150 lane = 675 + 240 Gbp = 915 Gbp = 15.25X coverage for 20 specimens

Variant discovery from many individuals

- 1X the genome size per individual
- *Arabidopsis lyrata* – ~207 Mbp
- 100 individuals x 1X coverage x 207 Mbp = 20.7 Gbp

NovaSeq Run Options – Target 20.7 Gbp

Flow Cell Type	Number lanes per flow cell	Kit Type/Size	Average Gbp per lane	Average number of reads per lane	Equivalent Number HiSeq Lanes	Pricing per Lane	Cost per Gbp
1/10th S4 lane		PE150	60	200,000,000		$641	$10.68
SP	2	100 cycles	36.25	362,500,000	1	$1,656	$45.68
SP	2	300 cycles	112.5	362,500,000	1	$2,424	$21.55
S1	2	100 cycles	75.25	725,000,000	2	$2,717	$36.11
S4	4	300 cycles	675	2,250,000,000	6.5	$5,846	$8.66

1/10th S4 PE150 lane = 60 Gbp = 2.9X coverage for 100 specimens

What are you sequencing?
How much sequence is required?

- Transcriptome
- De novo assembly
  - Generally, suggest using long read sequencing
- Gene expression analysis
  - How many genes in the genome?
  - What is your community standard?
  - Think in terms of millions of reads per sample
- Single cell gene expression
  - How many cells were sequenced?
  - How many reads per cell desired?
  - Sequence saturation attained?

Gene expression analysis

- *Eisenia fetida* – 29,552 genes
- Replicates – check your literature
  - 3 for this example
- Number of treatments
  - 2 for this example
- Community accepted reads per sample
  - 25 million reads
- 25 M reads x 3 x 2 = 150 M reads

NovaSeq Run Options – Target 150 M Reads

Flow Cell Type	Number lanes per flow cell	Kit Type/Size	Average Gbp per lane	Average number of reads per lane	Equivalent Number HiSeq Lanes	Pricing per Lane	Cost per Gbp
1/10th S4 lane		PE150	60	200,000,000		$641	$10.68
SP	2	100 cycles	36.25	362,500,000	1	$1,656	$45.68
SP	2	300 cycles	112.5	362,500,000	1	$2,424	$21.55
S1	2	100 cycles	75.25	725,000,000	2	$2,717	$36.11
S4	4	300 cycles	675	2,250,000,000	6.5	$5,846	$8.66

1/10th S4 PE150 lane = 200 M reads

One SP SE100 lane = 362 M reads

Gene expression analysis

- *Brassica rapa* – 45,985 genes
- Replicates – check your literature
  - 3 for this example
- Number of treatments
  - 4 for this example
- Community accepted reads per sample
  - 60 million reads
- 60 M reads x 3 x 4 = 720 M reads

NovaSeq Run Options – Target 720 M Reads

Flow Cell Type	Number lanes per flow cell	Kit Type/Size	Average Gbp per lane	Average number of reads per lane	Equivalent Number HiSeq Lanes	Pricing per Lane	Cost per Gbp
1/10th S4 lane		PE150	60	200,000,000		$641	$10.68
SP	2	100 cycles	36.25	362,500,000	1	$1,656	$45.68
SP	2	300 cycles	112.5	362,500,000	1	$2,424	$21.55
S1	2	100 cycles	75.25	725,000,000	2	$2,717	$36.11
S4	4	300 cycles	675	2,250,000,000	6.5	$5,846	$8.66

Two SP SE100 lanes = 725 M reads

NovaSeq Run Options – Target 720 M Reads

Flow Cell Type	Number lanes per flow cell	Kit Type/Size	Average Gbp per lane	Average number of reads per lane	Equivalent Number HiSeq Lanes	Pricing per Lane	Cost per Gbp
1/10th S4 lane		PE150	60	200,000,000		$641	$10.68
SP	2	100 cycles	36.25	362,500,000	1	$1,656	$45.68
SP	2	300 cycles	112.5	362,500,000	1	$2,424	$21.55
S1	2	100 cycles	75.25	725,000,000	2	$2,717	$36.11
S4	4	300 cycles	675	2,250,000,000	6.5	$5,846	$8.66

One S1 SE100 lanes = 725 M reads

NovaSeq Run Options – Target 720 M Reads

Flow Cell Type	Number lanes per flow cell	Kit Type/Size	Average Gbp per lane	Average number of reads per lane	Equivalent Number HiSeq Lanes	Pricing per Lane	Cost per Gbp
1/10th S4 lane		PE150	60	200,000,000		$641	$10.68
SP	2	100 cycles	36.25	362,500,000	1	$1,656	$45.68
SP	2	300 cycles	112.5	362,500,000	1	$2,424	$21.55
S1	2	100 cycles	75.25	725,000,000	2	$2,717	$36.11
S4	4	300 cycles	675	2,250,000,000	6.5	$5,846	$8.66

4/10th S4 PE150 lane = 800 M reads @ $2,564

Single cell expression analysis

- Simple control + treatment experiment
- Replicates – probably not
- 10,000 target cells per sample
- 10X Genomics recommends initially sequencing 20,000 reads per cell
- 2 x 10,000 x 20,000 = 400 M reads

- 10X sequencing format
  - Read 1 – 28 bp
  - Read 2 – 90 bp
  - Enough chemistry in a 100 cycle run

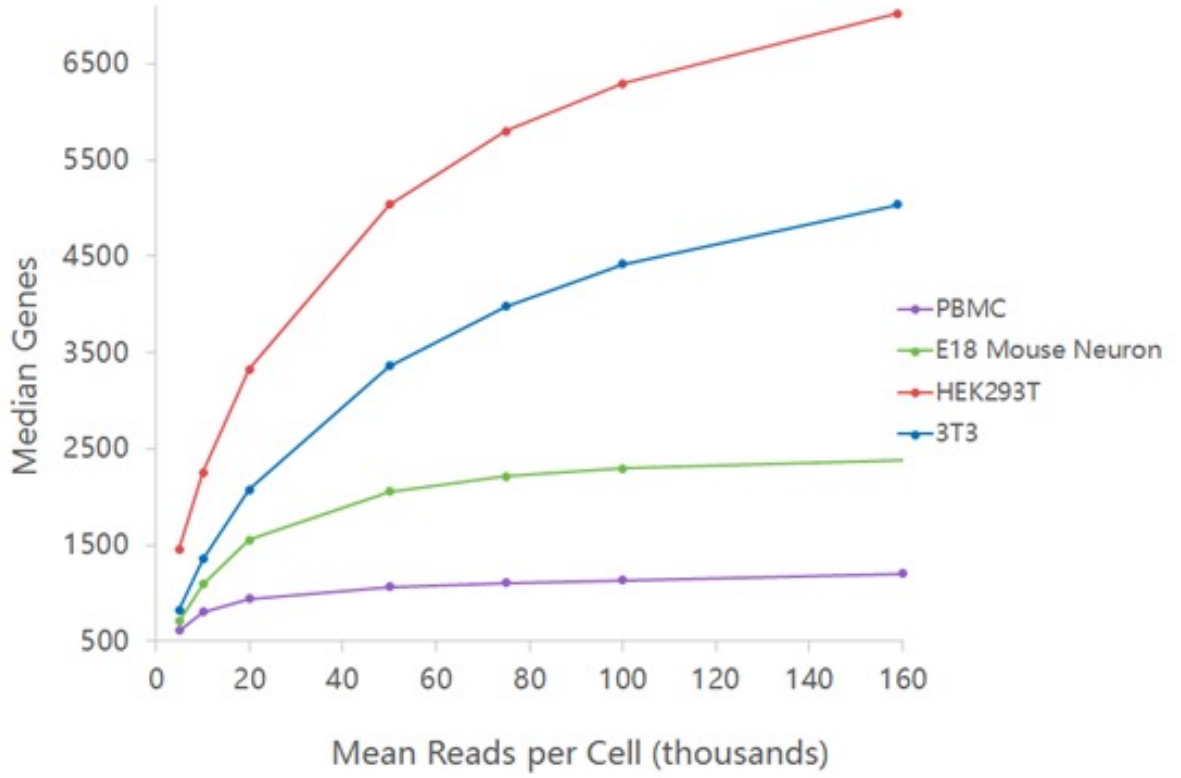
NovaSeq Run Options – Target 400 M Reads

Flow Cell Type	Number lanes per flow cell	Kit Type/Size	Average Gbp per lane	Average number of reads per lane	Equivalent Number HiSeq Lanes	Pricing per Lane	Cost per Gbp
1/10th S4 lane		PE150	60	200,000,000		$641	$10.68
SP	2	100 cycles	36.25	362,500,000	1	$1,656	$45.68
SP	2	300 cycles	112.5	362,500,000	1	$2,424	$21.55
S1	2	100 cycles	75.25	725,000,000	2	$2,717	$36.11
S4	4	300 cycles	675	2,250,000,000	6.5	$5,846	$8.66

One S1 SE100 lane = 725 M reads @ $2,717

Single cell sequence saturation

- Sequencing saturation occurs when additional sequence does not reveal new genes in a cell
- 10X Genomics suggests reaching sequence saturation
- Researchers should assess for themselves



Single cell expression analysis

- Simple control + treatment experiment
- Replicates – probably not
- 10,000 target cells per sample
- 10X Genomics recommends initially sequencing 60,000 reads per cell
- 2 x 10,000 x 60,000 = 1200 M reads

NovaSeq Run Options – Target 1,200 M Reads

Flow Cell Type	Number lanes per flow cell	Kit Type/Size	Average Gbp per lane	Average number of reads per lane	Equivalent Number HiSeq Lanes	Pricing per Lane	Cost per Gbp
1/10th S4 lane		PE150	60	200,000,000		$641	$10.68
SP	2	100 cycles	36.25	362,500,000	1	$1,656	$45.68
SP	2	300 cycles	112.5	362,500,000	1	$2,424	$21.55
S1	2	100 cycles	75.25	725,000,000	2	$2,717	$36.11
S4	4	300 cycles	675	2,250,000,000	6.5	$5,846	$8.66

Two S1 SE100 lanes = 1,450 M reads

Single Cell Runs on S2 Lanes

Flow Cell Type	Number lanes per flow cell	Kit Type/Size	Average Gbp per lane	Average number of reads per lane	Equivalent Number HiSeq Lanes	Pricing per Lane	Cost per Gbp
1/10th S4 lane		PE150	60	200,000,000		$641	$10.68
SP	2	100 cycles	36.25	362,500,000	1	$1,656	$45.68
SP	2	200 cycles	75.25	362,500,000	1	$2,107	$28.00
SP	2	300 cycles	112.5	362,500,000	1	$2,424	$21.55
SP	2	500 cycles	181.25	362,500,000	1	$3,405	$18.79
S1	2	100 cycles	75.25	725,000,000	2	$2,717	$36.11
S1	2	200 cycles	149.75	725,000,000	2	$3,620	$24.17
S1	2	300 cycles	225	725,000,000	2	$4,204	$18.68
S2	2	100 cycles	187.5	1,850,000,000	4	$4,651	$24.81
S2	2	200 cycles	375	1,850,000,000	4	$6,049	$16.13
S2	2	300 cycles	562.5	1,850,000,000	4	$6,832	$12.15
S4	4	200 cycles	450	2,250,000,000	6.5	$4,897	$10.88
S4	4	300 cycles	675	2,250,000,000	6.5	$5,846	$8.66
S4	4	35 cycles	78.75	2,250,000,000	6.5	$3,528	$44.80