How much sequencing do I need?

Emily Crisovan

Genomics Core

September 26, 2018

How much sequencing?

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?

Based on Illumina sequencing options

Experimental Design

What are you sequencing?

- Genome
  - De novo assembly
  - Resequencing project
- Transcriptome
  - De novo assembly
  - Gene expression project
- Amplicon sequencing
- Whole meta-genomes
- Small RNAs
- ChIP-seq
- Exome capture

What type of sequencing run?

Single end (SE) or paired end (PE)?

What read length?

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

Assembly of genome or transcriptome?

- PE 150 bp, 250 bp, 300 bp

Counting experiment?

- SE 35 bp, 50 bp, 75 bp

How many lanes of sequencing?

Genome assembly

- Depends on the desired coverage
  - New assembly: 75x – 100x
  - Resequencing: 10x – 20x
  - Long-read error correction: 20x – 30x

lanes required =

desired Gbp / expected Gbp per lane

How many lanes of sequencing?

For gene expression analysis

- Counting experiment
- What is typical in you field?
- Consider ploidy
- How many replicates?
- Account for variability between samples

Transcriptome assembly

- Number of genes in the genome
- Complexity of the transcriptome

lanes required =

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

The “fudge factor”

- There will always be variation in the number of reads per sample per lane
  - Need to account for this when designing experiment
- It is difficult to assign a specific value to the fudge factor
- Call/e-mail us to discuss the fudge factor

Genome Sequencing Example #1

New eukaryotic genome assembly

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

How many lanes of sequencing do you need?

lanes required =

desired Gbp / expected Gbp per lane

What changes if this was a resequencing project?

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 * # bases) / 1,000,000,000

(350,000,000 * 300 bases) / 1,000,000,000 = 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 genome

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

How many lanes of sequencing do you need?

lanes required =

desired Gbp / expected Gbp 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 * # bases) / 1,000,000,000

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

How many lanes of sequencing are needed?

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%.

Genome Sequencing Example #3

Whole genome metagenomics sequencing

- Unknown number of fungal, bacterial & other species
- Unknown genome sizes
- PE 150
- HiSeq 4000 averages 350 million reads per lane

How many lanes of sequencing do you need?

lanes required =

desired Gbp / expected Gbp per lane

Perform experiment to determine what is present and then go forward from there

Transcript Sequencing Example

Transcript assembly

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

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

Transcript Sequencing Example Calculations

Transcriptome assembly

- 25,000 genes
- 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

Calculate the actual average to check if there is enough wiggle room:

350 M read pairs / 5 samples = 70 M read pairs per sample

Gene Expression Example

Gather counts for differential expression analysis

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

lanes required =

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

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

Calculate the actual average to check if there is enough wiggle room:

350 M reads per lane * 2 lanes = 700 M reads total

700 M reads / (6 samples * 3 replicates) = 38.8 M reads/sample

Amplicon Sequencing Example

Sequencing the same target from multiple samples

- Metagenomic survey
- Specific target from many individuals (ex. 16S V4)
- Barcoding required
- PE 250 MiSeq standard run
  - 8-10 million read pairs expected
- Coverage dependent on number of samples
- Variation between samples is very large

runs required =

read pairs desired * # samples * fudge factor / read pairs per run

Amplicon Sequencing Example

You have 96 samples and you would like 70,000 read pairs per sample:

9 M read pairs per run / 96 samples = ~93,750 read pairs per sample

With amplicon sequencing you will receive a wide range of reads per sample, for instance, 30,000 – 150,000 read pairs per sample would not be unreasonable for this example.

Will this be suitable for your experiment?

If not, reduce the number of samples per run or request multiple runs.

Small RNA Sequencing Example

Goal is to gather counts for differential expression analysis

- For miRNAs, 10 million reads are common
- 3 to 5 replicates
- SE 50

lanes required =

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