How much sequencing do I need?

Emily Crisovan

Genomics Core

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?

All 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

Transcriptome assembly

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

# 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

# lanes required =

(minimum # reads per sample x # replicates x # samples 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 hard 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

Calculate expected Gbp per lane of HiSeq 4000 PE150:

(# of reads x read length) / 1,000,000,000

(350,000,000 x 300) / 1,000,000,000 = 105 Gbp

Calculate desired Gbp:

1.2 Gbp x 80 = 96 Gbp

Calculate # of lanes required:

96 Gbp / 105 Gbp = 0.91 lanes → round up to 1 lane because we do not sell partial lanes

What changes if this was a resequencing project?

The desired coverage would be reduced to 10-20x

Genome Sequencing Example #2

New prokaryotic genome

- 12 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

Calculate expected Gbp per lane of HiSeq 4000 PE150:

(# of reads x read length) / 1,000,000,000

(350,000,000 x 300) / 1,000,000,000 = 105 Gbp

Calculate desired Gbp:

8 Mbp x 40x coverage x 12 isolates = 3.84 Gbp

Calculate # of lanes required:

3.84 Gbp / 105 Gbp = 0.04 lanes

The HiSeq 4000 is not the appropriate sequencer for this project.

Genome Sequencing Example #2

New prokaryotic genome

- 12 different bacterial isolates
- 8 Mbp genome
- Target 40x coverage
- PE 150
- ~~HiSeq 4000 averages 350 million reads per lane~~
- MiSeq v2 Standard PE 250 gives ~6.0-7.5 Gbp per run

How many lanes of sequencing do you need?

# lanes required =

desired Gbp / expected Gbp per lane

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

# of reads per lane / # of reads per sample = # of samples that can be loaded on one lane

350 M reads / 60 M reads per sample = 5.8 samples → round down to 5

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

350 M reads / 5 samples = 70 M reads per sample

Yes, this is enough wiggle room.

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 – Method 1

How many lanes of sequencing are needed if you have 6 samples with 3 replicates each and you would like a minimum of 30 million reads each?

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

30 M reads x 6 samples x 3 replicates x 1.25 fudge factor = 675 M reads

675 M reads / 350 M reads per lane = 1.9 lanes ➔ 2 lanes

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

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

Is this enough wiggle room?

Gene Expression Example – Method 2

How many lanes of sequencing are needed if you have 6 samples with 3 replicates each and you would like a minimum of 30 million reads each?

30 M reads x 6 samples x 3 replicates = 540 M reads

540 M reads / 350 M reads per lane = 1.5 lanes ➔ 2 lanes

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

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

Is this enough wiggle room?

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

9,000,000 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.

Will this be suitable for your experiment?

If not, reduce the number of samples per run.

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

Same calculations as a gene expression study