From sample to fastq

Outline

- How library quality affects data recovery
- Sequencing costs
- Estimate cost for your own experiment

Requirements for library prep protocol for lcWGS

- To prepare libraries for hundreds of samples, we need a protocol that is
  - Cheap
  - Efficient
  - Reliable
- Sometimes robustness to sample degradation is also important

Sequencing-Ready Fragment

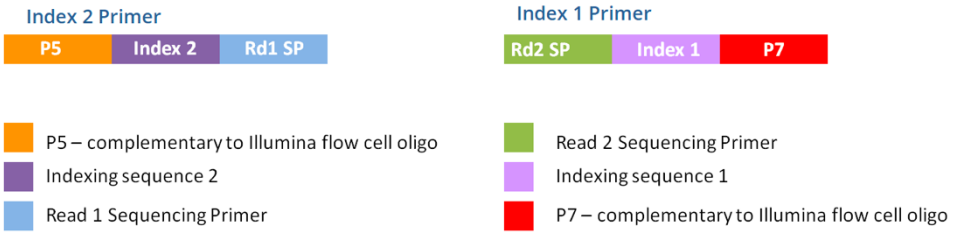
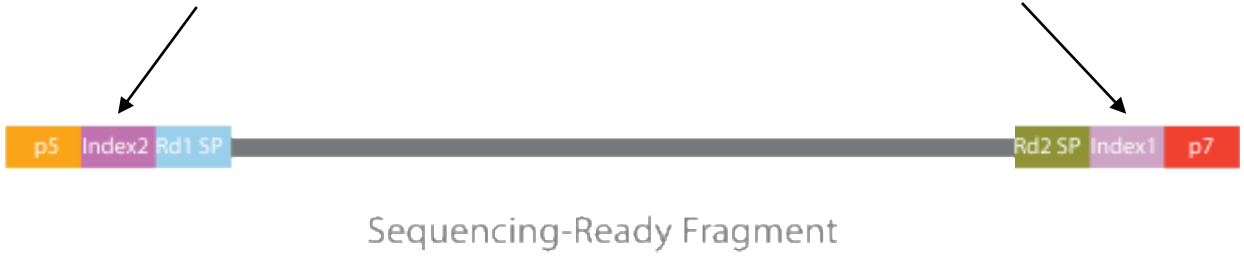


Image adapted from the Nextera XT DNA Library Prep Kit Reference Guide (© 2017 Illumina, Inc.)

Unique vs. combinatorial dual index barcodes



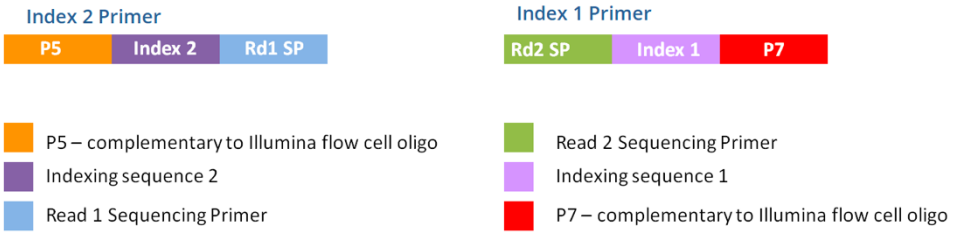
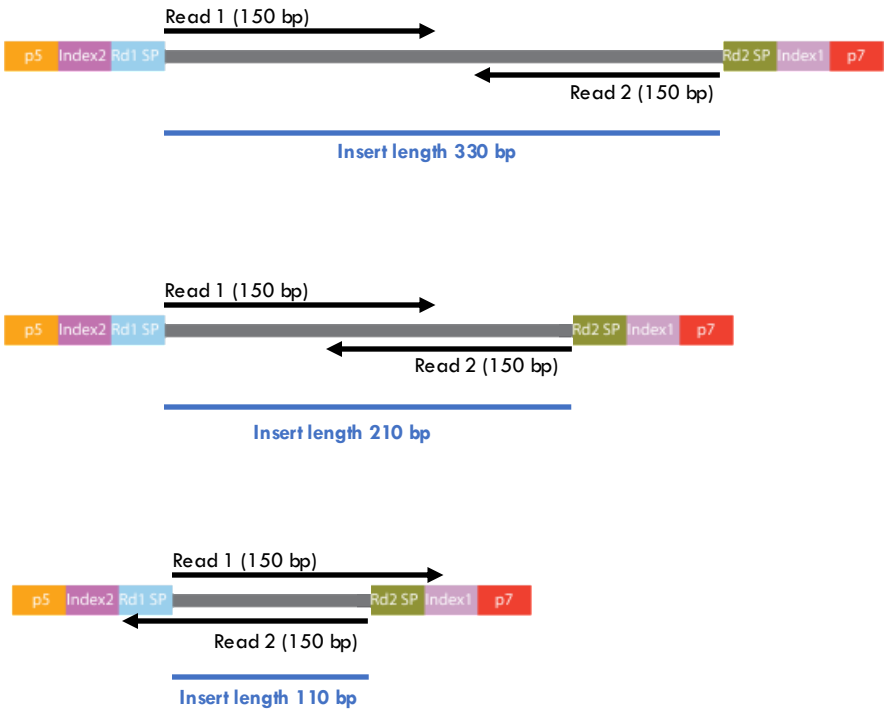
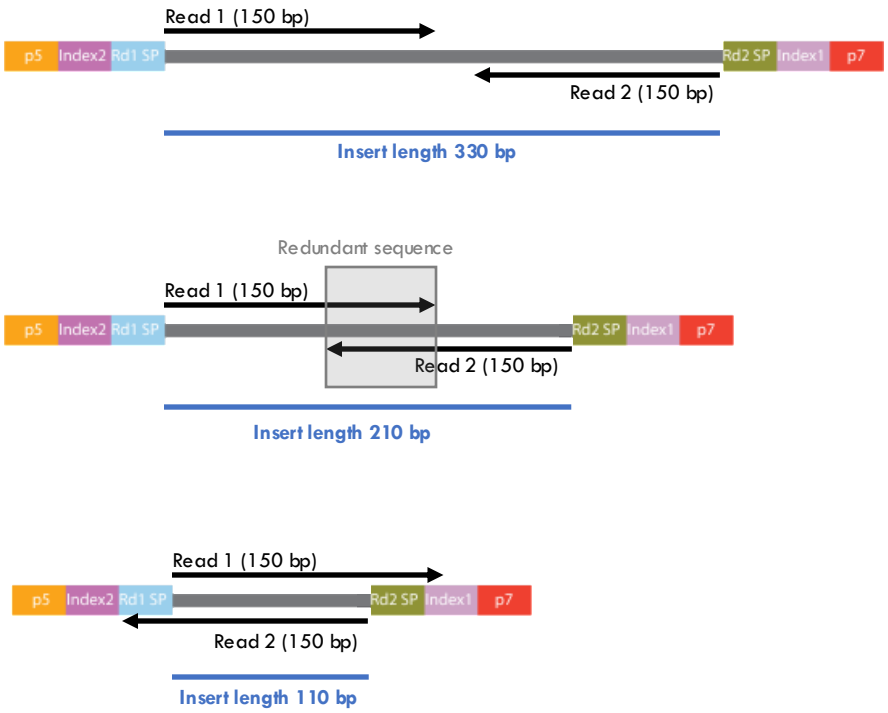


Image adapted from the Nextera XT DNA Library Prep Kit Reference Guide (© 2017 Illumina, Inc.)

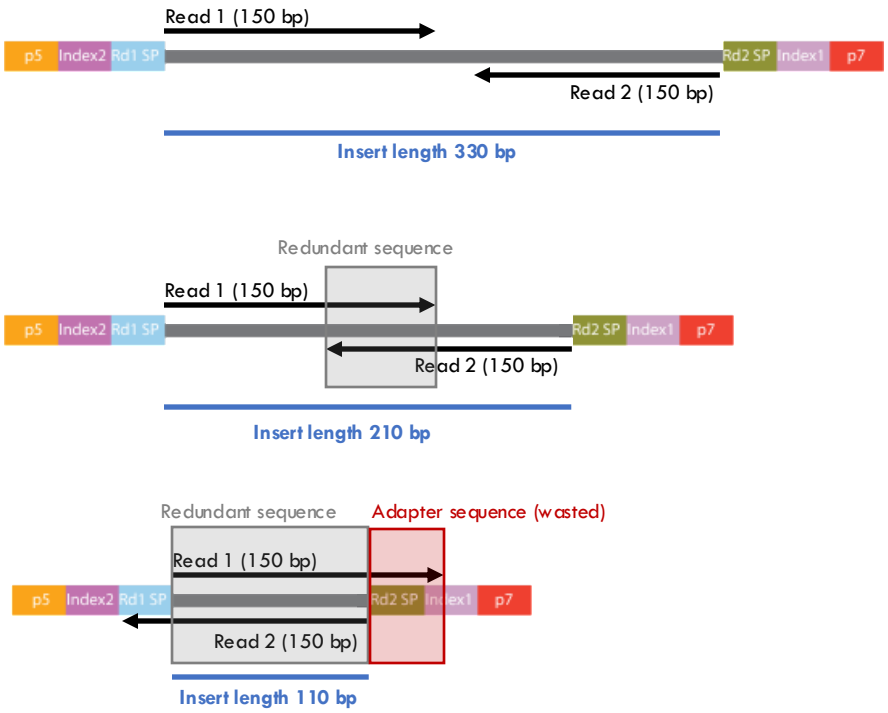
Insert length relative to read length



Insert length relative to read length

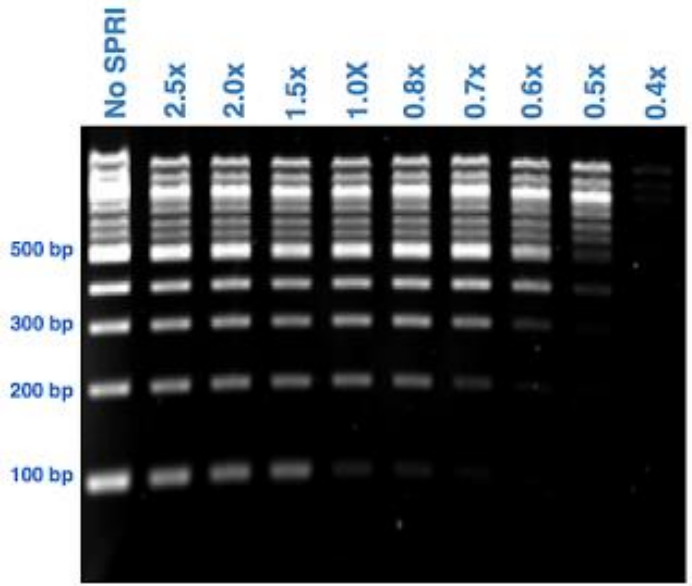


Insert length relative to read length



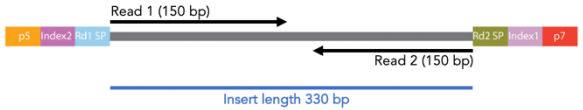
Size selection with Ampure beads

Tune the size distribution of your library fragments to minimize "waste" of sequence due to paired-end overlap and adapter read-through



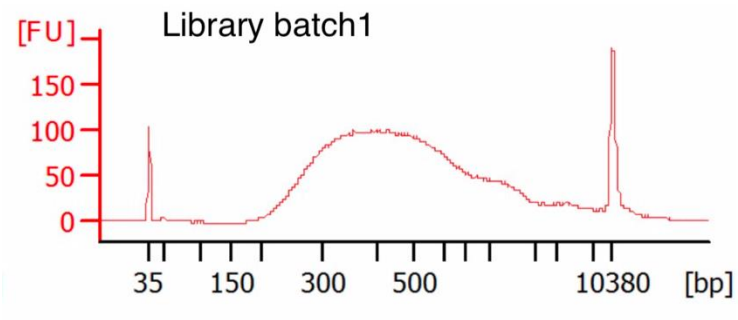
http://enseqlopedia.com/2012/04/how-do-spri-beads-work/

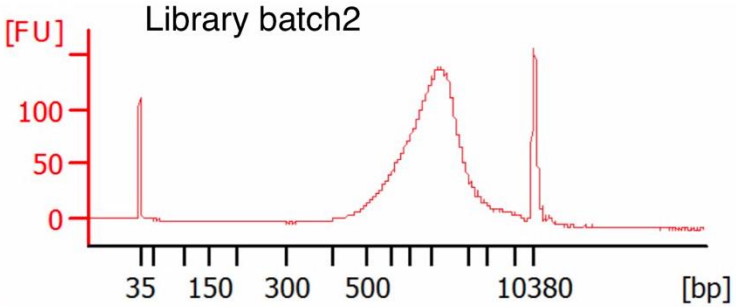
Ideally, we want all library fragments to be greater than the adapter length plus 2 x the read length (for PE)



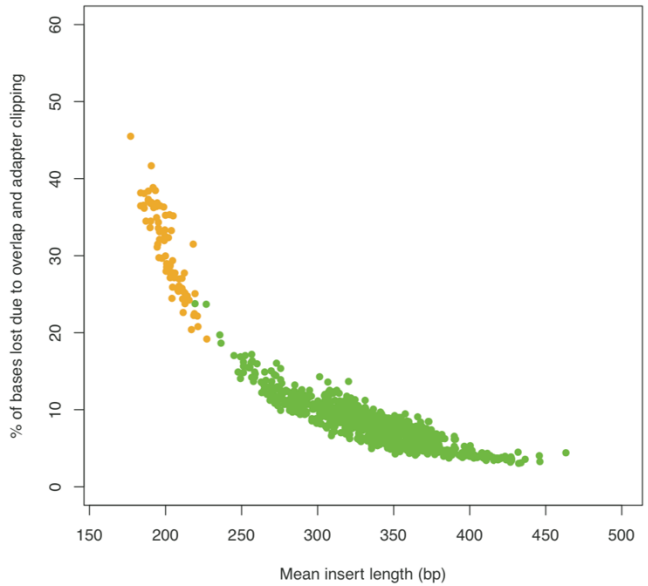
Ideal minimum fragment length

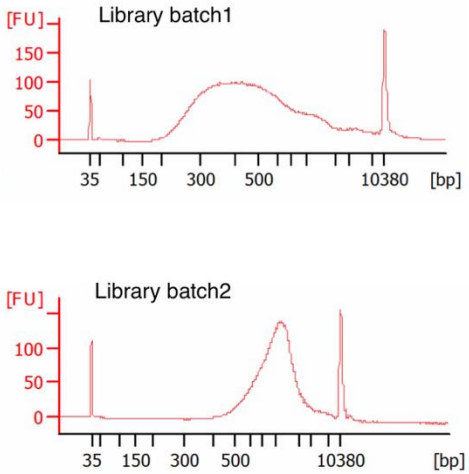
Two examples of our library pools



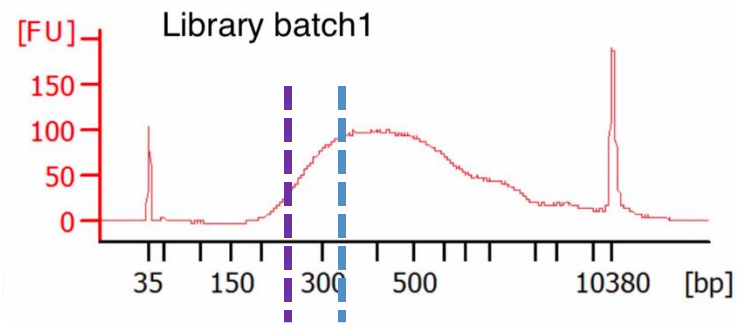


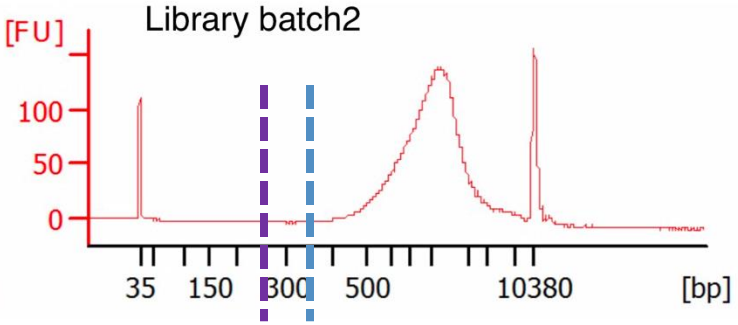
The library fragment size distribution can substantially influence the amount of data lost in data QC steps





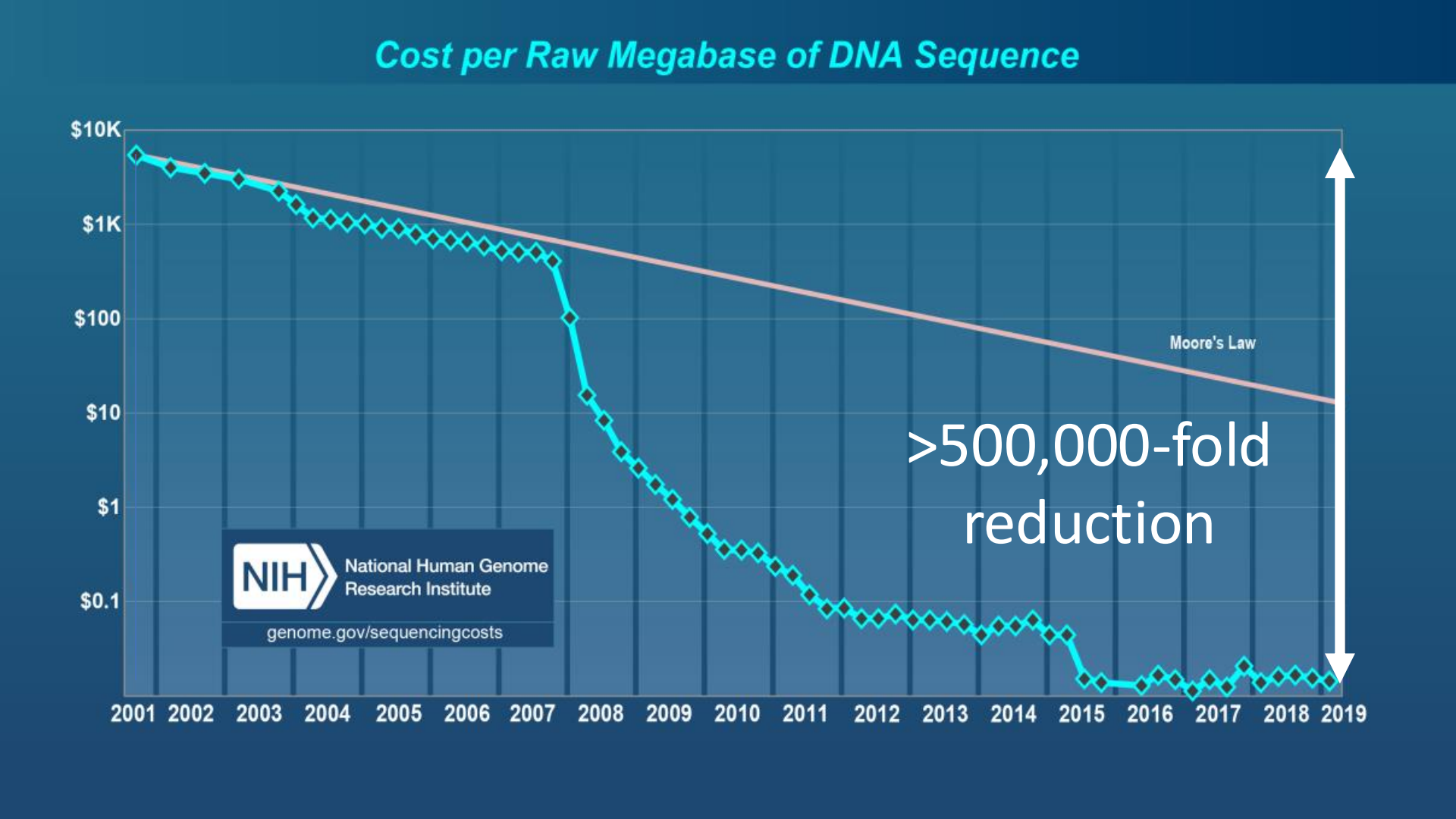
Two examples of our library pools





The length of Nextera adapters is 138 bp and libraries were sequenced with 2*125bp reads

→ Minimum fragment length to avoid overlap 388bp
→ Minimum fragment length to avoid adapter read-through 263bp


Cost per Raw Megabase of DNA Sequence
$10K
$1K
$100
$10
$1
$0.1
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Moore's Law
>500,000-fold reduction
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts

What is the current price for 2x sequencing of an Atlantic silverside (including library preparation)?

Genome size ~650 Mb

- $4
- $16
- $36
- $72
- $113

1 USD ≈ 1 EURO

What is the current price for 2x sequencing of an Atlantic silverside?

Genome size ~650 Mb

- $4
- $16
- $36
- $72
- $113

1 USD ≈ 1 EURO

Example costs for other genome sizes

Incl. library preparation and sequencing to 2x genome coverage*

Genome size (Gb)	Cost per sample (USD)[a] 1× coverage	2× coverage	Example organisms
0.2	11 (3)	13 (5)	Fruit fly, honeybee, arabidopsis
0.65	16 (8)	25 (17)	Atlantic silverside, stickleback, eastern oyster
1	21 (13)	34 (26)	Zebra finch, chicken, purple sea urchin
3	47 (39)	86 (78)	Human, Atlantic salmon, African clawed frog

*Cost estimates do not include labor and assume sequencing costs ~13 USD per Gb in shared S4 lanes on an Illumina NovaSeq and 8 USD per sample for library preparation

Lou et al. 2021. Mol. Ecol

Example costs for other genome sizes

Incl. library preparation and sequencing to 2x genome coverage*

Genome size (Gb)	Cost per sample (USD)[a]	
	1× coverage	2× coverage
0.2	11 (3)	13 (5)
0.65	16 (8)	25 (17)
1	21 (13)	34 (26)
3	47 (39)	86 (78)

Compare to:

$30 per sample for RADseq
$15 per sample for RADcapture

Meek and Larson. 2019. Mol Ecol Res

*Cost estimates do not include labor and assume sequencing costs ~13 USD per Gb in shared S4 lanes on an Illumina NovaSeq and 8 USD per sample for library preparation

Lou et al. 2021. Mol. Ecol

Exercise – how much will your experiment cost?

- Assumed costs:
  - Library preparation: $8 per sample
  - Sequencing: $4 per Gb (assuming a full NovaSeqX 25B lane)
  - Target coverage per sample: Expect to lose at least 30-50% of your data in filtering

Exercise – how much will your experiment cost?

- Assumed costs:
  - Library preparation: $8 per sample
  - Sequencing: $4 per Gb (assuming a full NovaSeqX 25B lane)
  - Target coverage per sample: Expect to lose at least 30-50% of your data in filtering

- Example: I would like to have 1x coverage for downstream analysis for 40 individuals from each of 5 populations (200 individuals total) of my favorite animal with a genome size of ~800 Mb

- Calculation: I will target 2x coverage raw sequencing. This means

  2 * 800 Mb/individual * 200 individuals = 320,000 Mb (320 Gb)

  My total cost is thus (320 Gb * $4/Gb) + (200 libraries * $8 per library) = **$2,880**