

From sample to fastq

# Outline

- Brief overview of library preparation procedure
- Sequencing costs
- Estimate cost for your own experiment

# Requirements for library preparation protocol

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

# One example of a library preparation technique



## RESEARCH ARTICLE

### Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes

Michael Baym<sup>1☯</sup>, Sergey Kryazhimskiy<sup>2,3☯</sup>, Tami D. Lieberman<sup>1☯</sup>, Hattie Chung<sup>1☯</sup>, Michael M. Desai<sup>2,3,4\*</sup>, Roy Kishony<sup>1,5\*</sup>

**1** Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, United States of America, **2** Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, United States of America, **3** FAS Center for Systems Biology, Harvard University, Cambridge, Massachusetts, United States of America, **4** Department of Physics, Harvard University, Cambridge, Massachusetts, United States of America, **5** Faculty of Biology and Department of Computer Science, Technion-Israel Institute of Technology, Haifa, Israel

☯ These authors contributed equally to this work.

\* [mmdesai@fas.harvard.edu](mailto:mmdesai@fas.harvard.edu) (MB); [roy\\_kishony@hms.harvard.edu](mailto:roy_kishony@hms.harvard.edu) (RK)



CrossMark  
click for updates

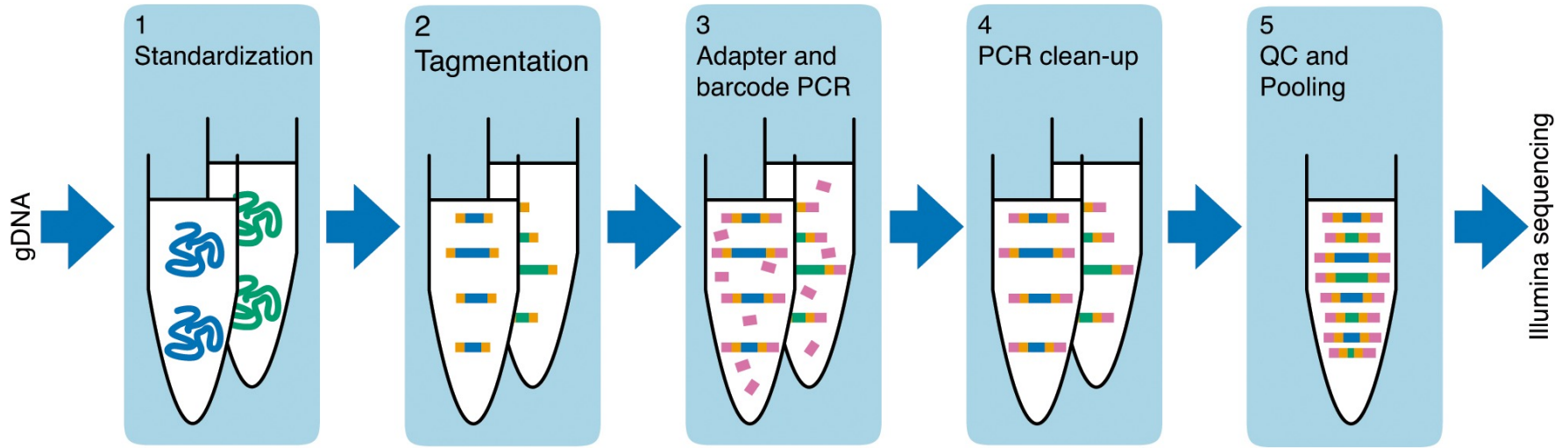
## OPEN ACCESS

**Citation:** Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R (2015) Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes. PLoS ONE 10(5): e0128036. doi:10.1371/journal.pone.0128036

## Abstract

Whole-genome sequencing has become an indispensable tool of modern biology. However, the cost of sample preparation relative to the cost of sequencing remains high, especially for small genomes where the former is dominant. Here we present a protocol for rapid and inexpensive preparation of hundreds of multiplexed genomic libraries for Illumina sequencing. By carrying out the Nextera tagmentation reaction in small volumes, replacing costly re-

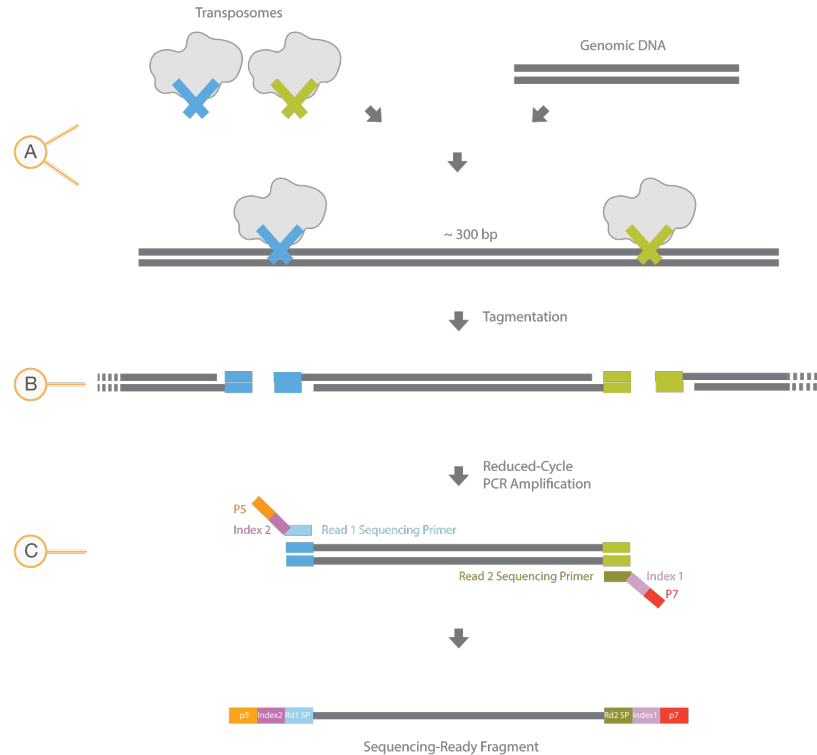
# Library preparation protocol



Transposome with adapters  
combined with template DNA

Tagmentation to fragment  
and add adapters

Limited-cycle PCR to add  
index adapter sequences



Transposome with adapters  
combined with template DNA

Tagmentation to fragment  
and add adapters

Limited-cycle PCR to add  
index adapter sequences

Other great library preparation methods  
work by adapter ligation  
(rather than tagmentation)






Sequencing-Ready Fragment



## Sequencing-Ready Fragment




### Index 2 Primer



-  P5 – complementary to Illumina flow cell oligo
-  Indexing sequence 2
-  Read 1 Sequencing Primer

### Index 1 Primer



-  Read 2 Sequencing Primer
-  Indexing sequence 1
-  P7 – complementary to Illumina flow cell oligo






# Unique vs. combinatorial dual index barcodes






## Index 2 Primer



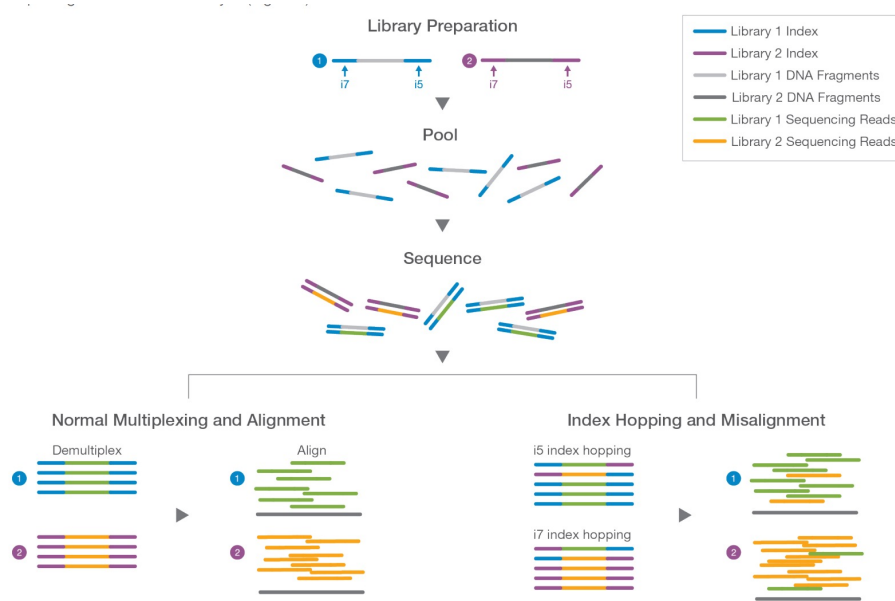
-  P5 – complementary to Illumina flow cell oligo
-  Indexing sequence 2
-  Read 1 Sequencing Primer

## Index 1 Primer

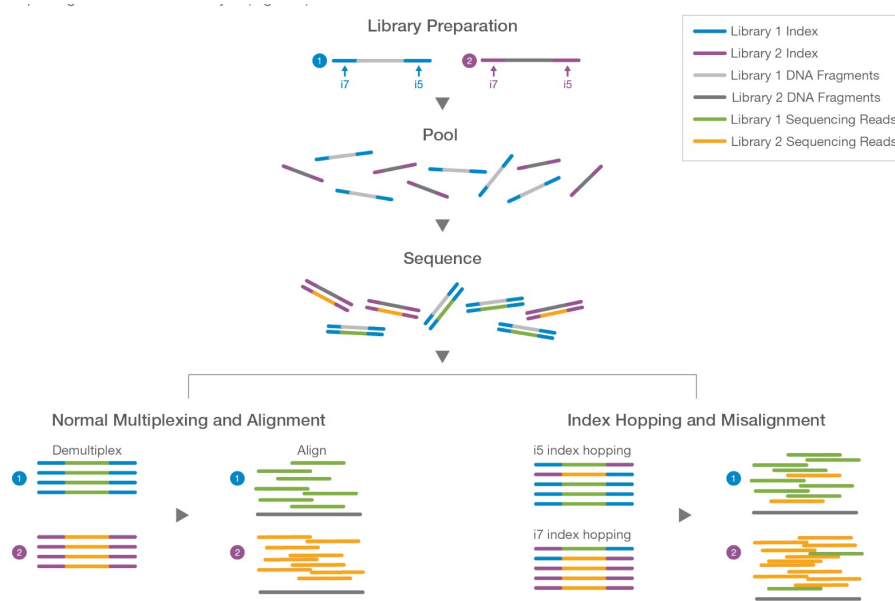


-  Read 2 Sequencing Primer
-  Indexing sequence 1
-  P7 – complementary to Illumina flow cell oligo

# Beware that index hopping can cause misassigned sequence reads when using combinatorial index barcodes

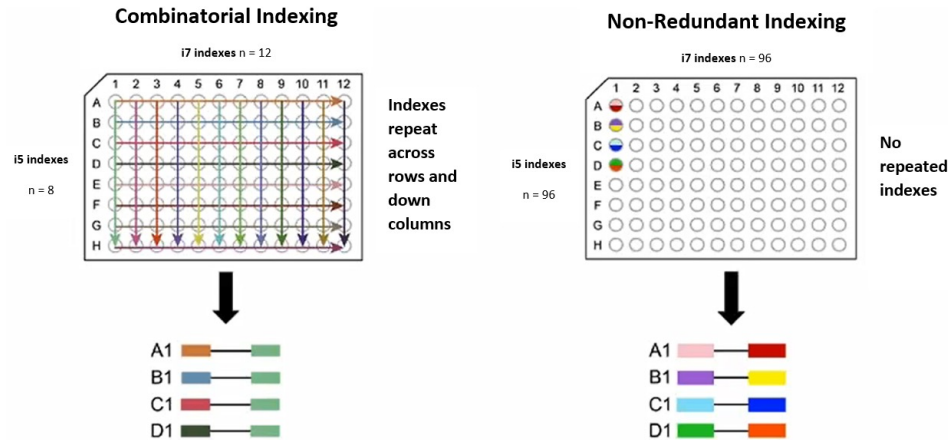


# Beware that index hopping can cause misassigned sequence reads when using combinatorial index barcodes

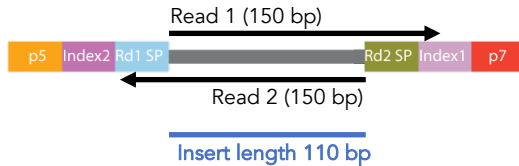
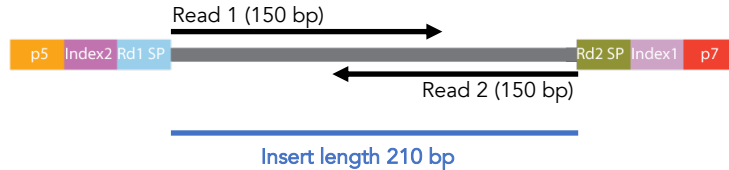
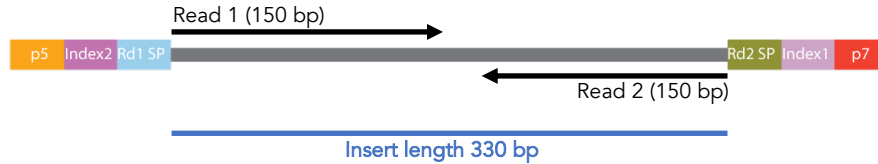


Index hopping often affects 0.1-2% of reads!

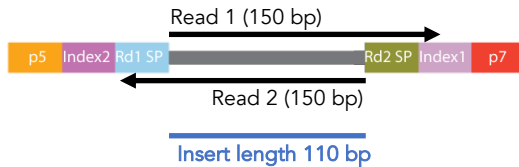
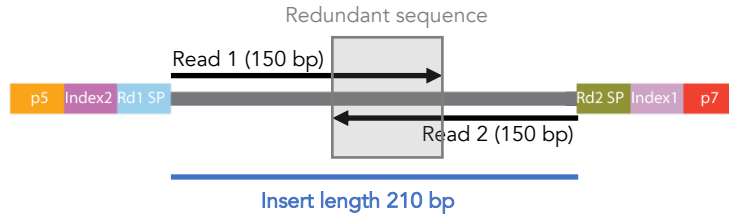
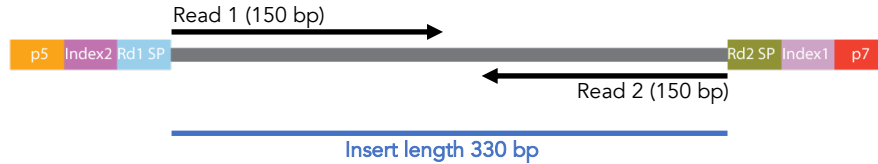
Unique dual index recommended even though they are more expensive than combinatorial dual index adapters



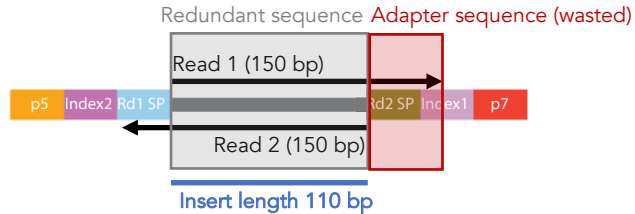
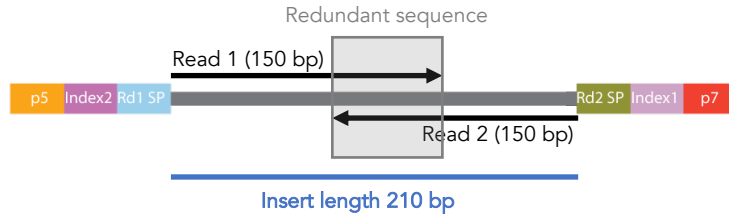
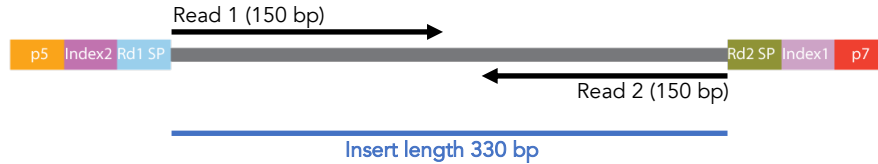
# Insert length relative to read length



# Insert length relative to read length

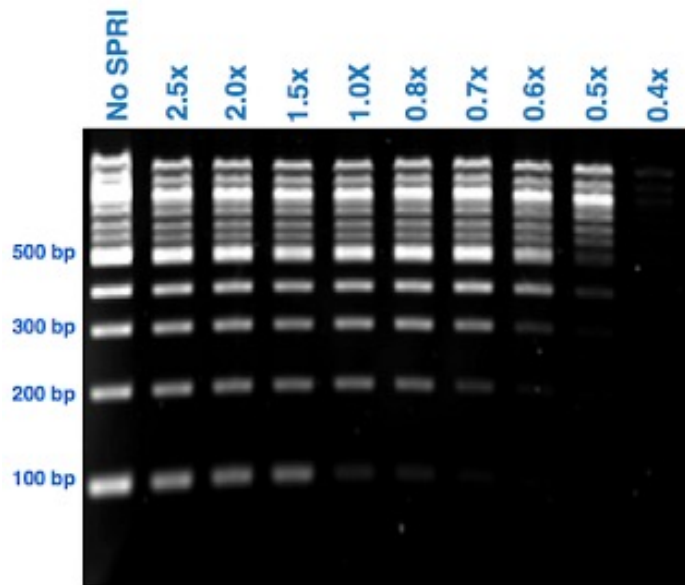


# Insert length relative to read length

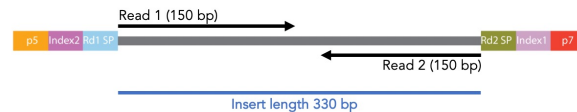


# Size selection with Ampure beads

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



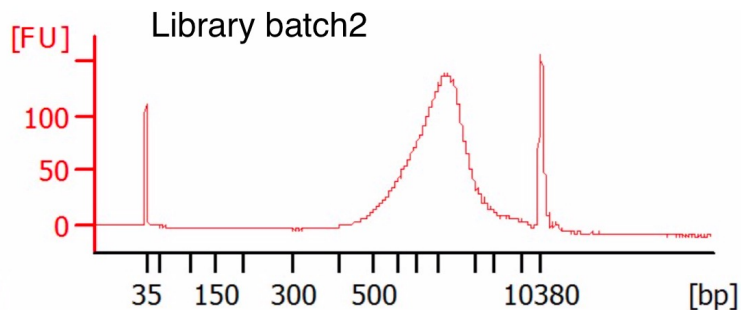
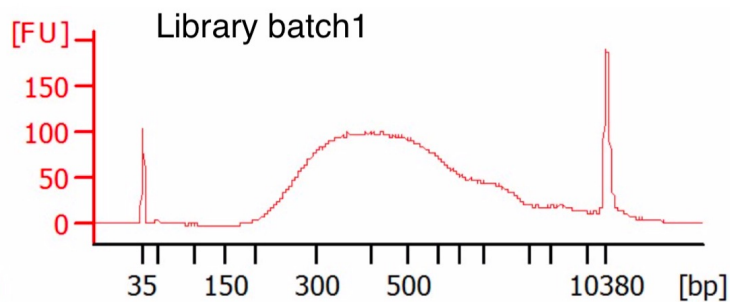
Ideally, we want all library fragments to be greater than the adapter length plus 2 x the read length (for PE)



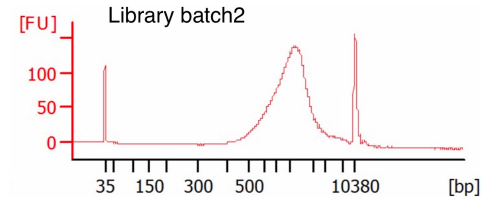
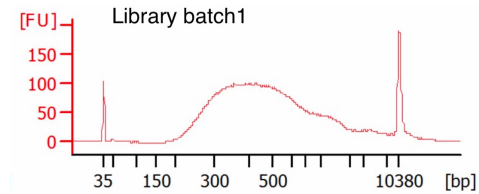
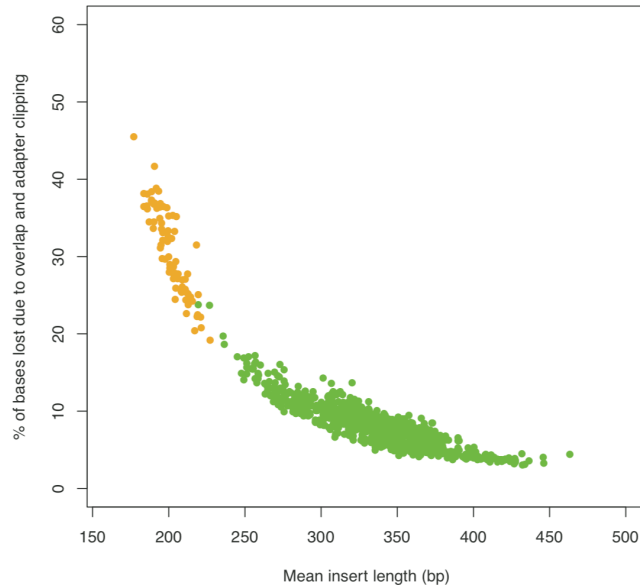
Ideal minimum fragment length



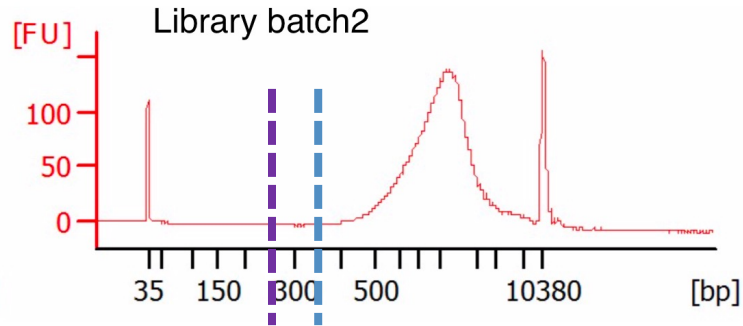
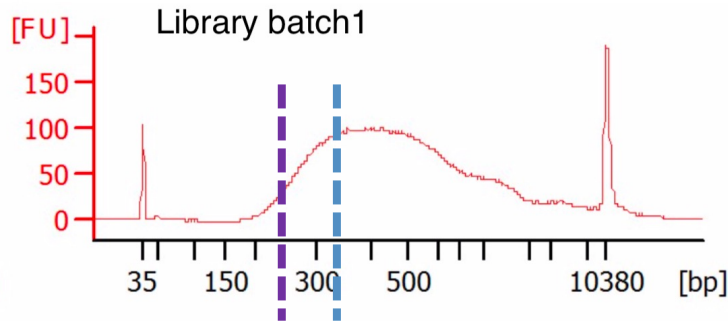
# Two examples of our library pools



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



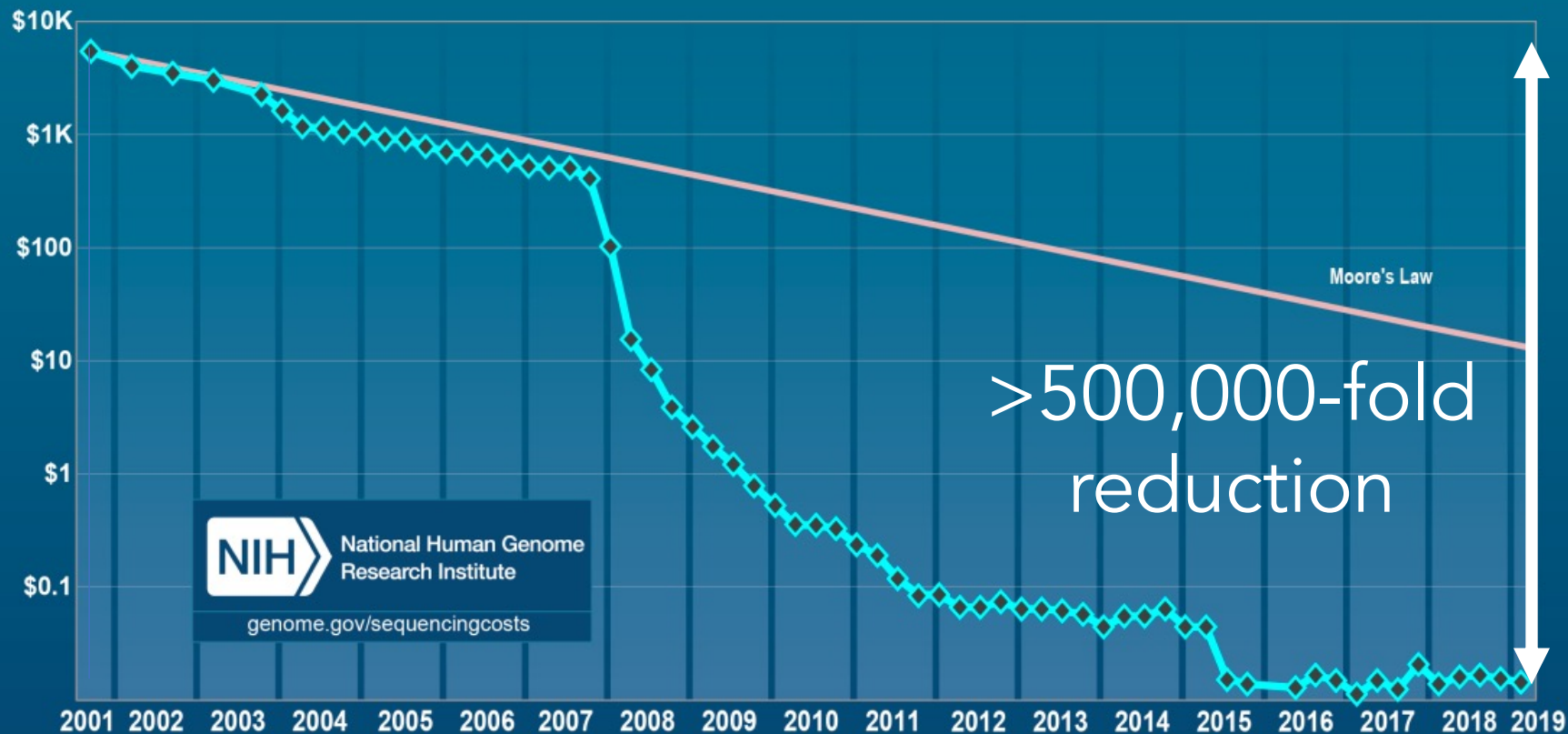
# Two examples of our library pools



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

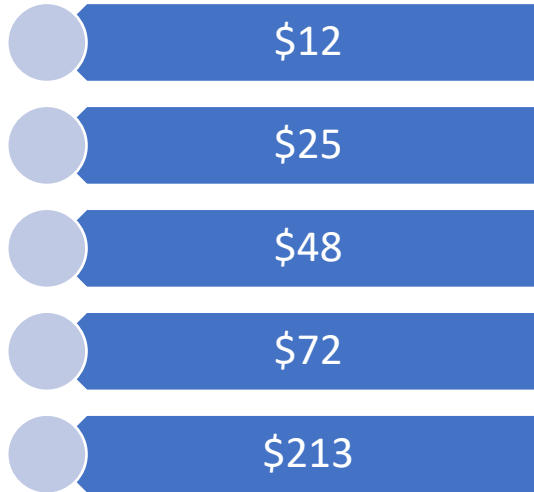
- Minimum fragment length to avoid overlap 383bp
- Minimum fragment length to avoid adapter read-through 250bp

## Cost per Raw Megabase of DNA Sequence



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

Genome size ~650 Mb

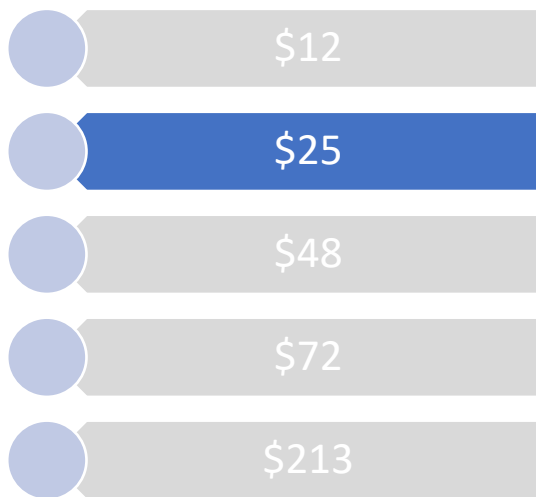


1 USD  $\approx$  1 EURO



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

Genome size ~650 Mb



1 USD  $\approx$  1 EURO



# Example costs for other genome sizes

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

| Genome size (Gb) | Cost per sample (USD) <sup>a</sup> |             | Example organisms                                |
|------------------|------------------------------------|-------------|--|
|                  | 1x coverage                        | 2x coverage |  |
| 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

# Example costs for other genome sizes

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

| Genome size (Gb) | Cost per sample (USD) <sup>a</sup> |             |
|------------------|------------------------------------|-------------|
|                  | 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



# Exercise – how much will your experiment cost?

- Assumed costs:
  - Library preparation: \$8 per sample
  - Sequencing: \$13 per Gb
  - 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: \$13 per Gb
  - 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
$$2 * 800 \text{ Mb/individual} * 200 \text{ individuals} = 320,000 \text{ Mb (320 Gb)}$$
My total cost is thus  $(320 \text{ Gb} * \$13/\text{Gb}) + (200 \text{ libraries} * \$8 \text{ per library}) = \textbf{\$5,760}$