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

PLOS | ONE

RESEARCH ARTICLE

Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes

Michael Baym[1☯], Sergey Kryazhimskiy[2,3☯], Tami D. Lieberman[1☯], Hattie Chung[1☯], Michael M. Desai[2,3,4]*, Roy Kishony[1,5]*

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 (MB); roy_kishony@hms.harvard.edu (RK)

CrossMark
click for updates

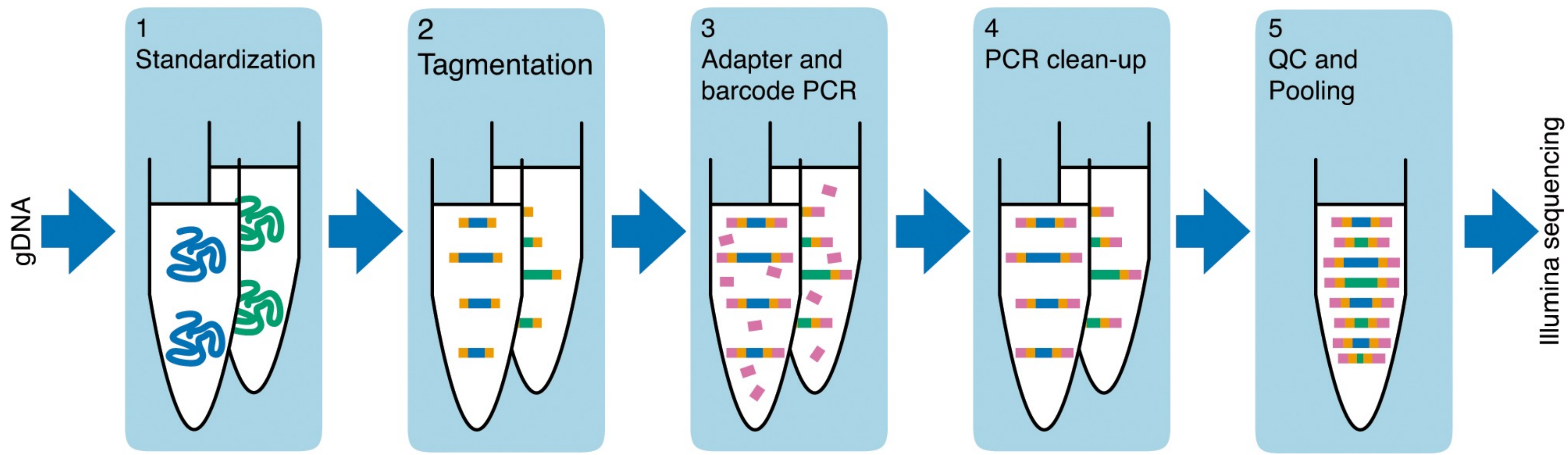
OPEN ACCESS

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 indispensible 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

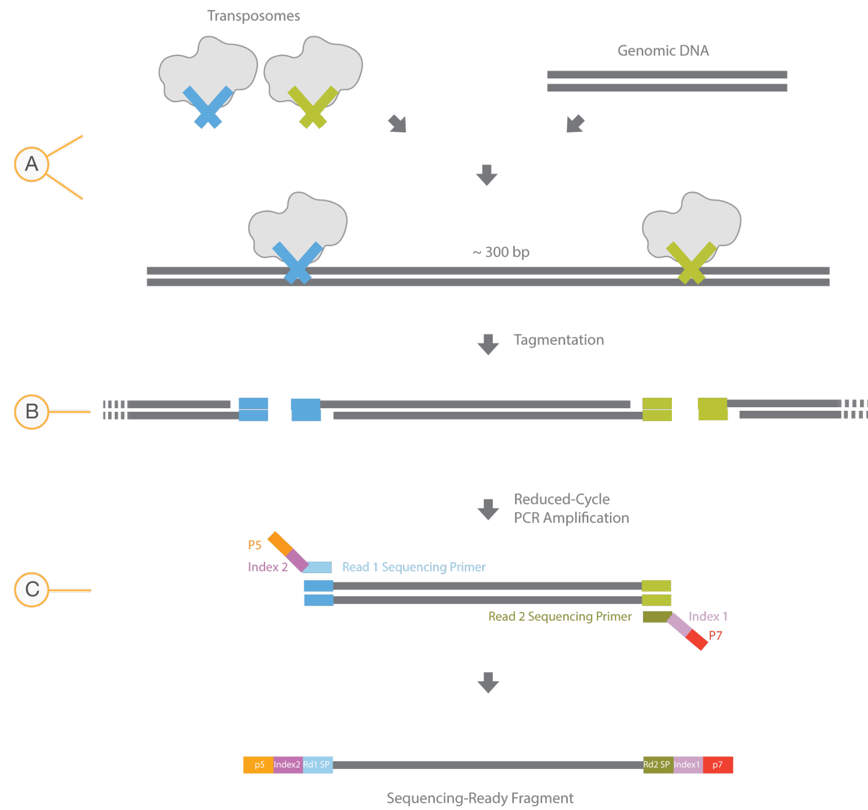


Image from the Nextera XT DNA Library Prep Kit Reference Guide (© 2019 Illumina, Inc.)

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

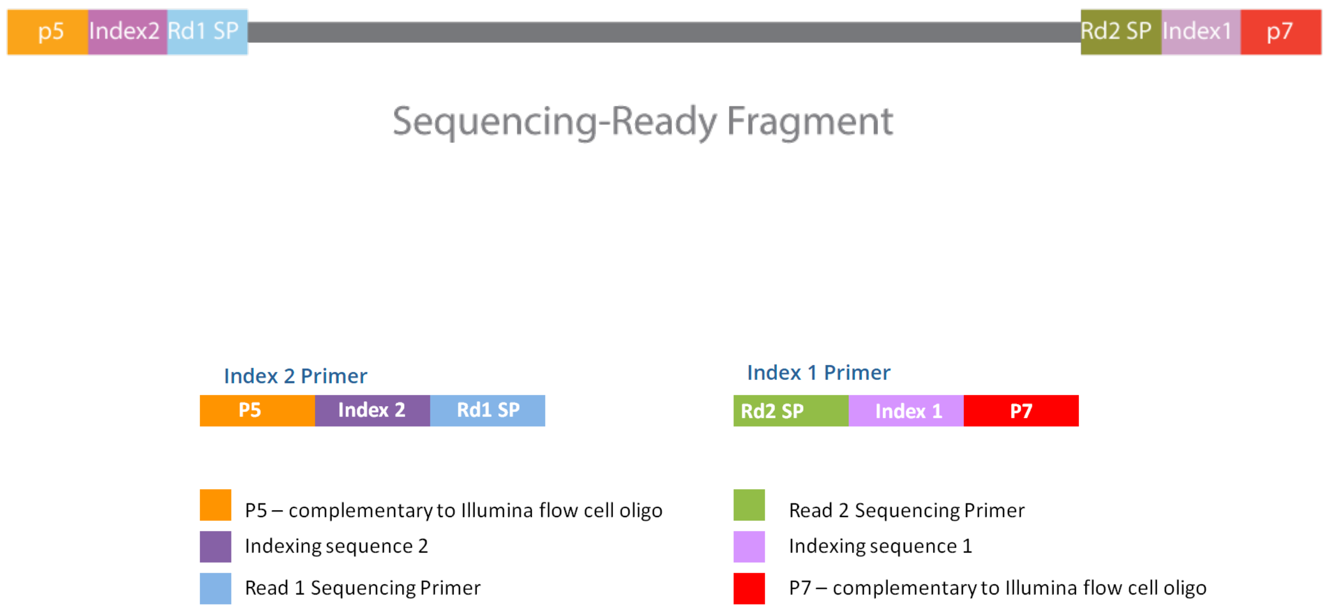


Image adapted from the Nextera XT DNA Library Prep Kit Reference Guide (© 2017 Illumina, Inc.)

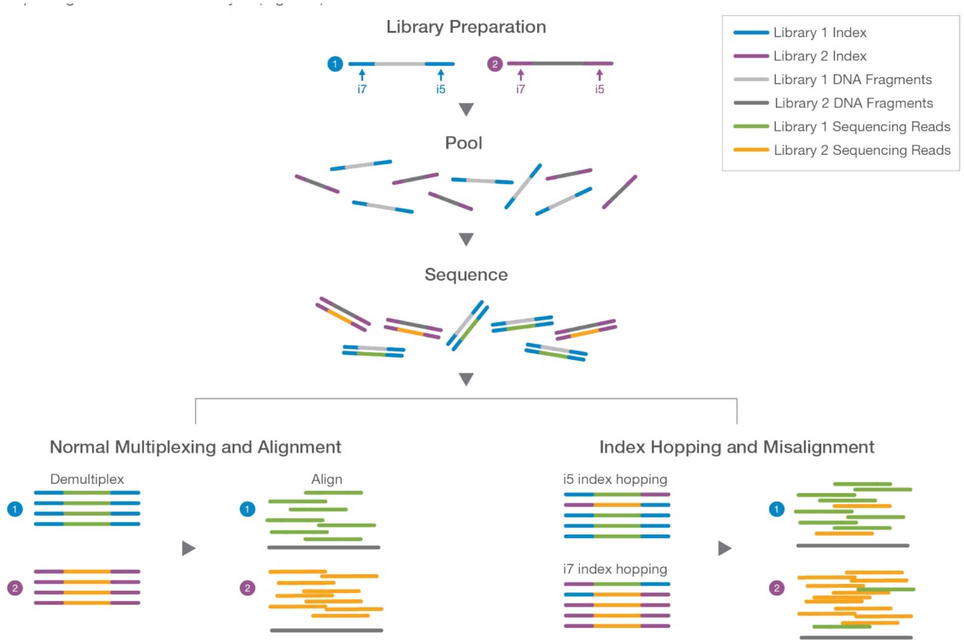
Unique vs. combinatorial dual index barcodes

Sequencing-Ready Fragment

- P5 – complementary to Illumina flow cell oligo
- Indexing sequence 2
- Read 1 Sequencing Primer
- Read 2 Sequencing Primer
- Indexing sequence 1
- P7 – complementary to Illumina flow cell oligo

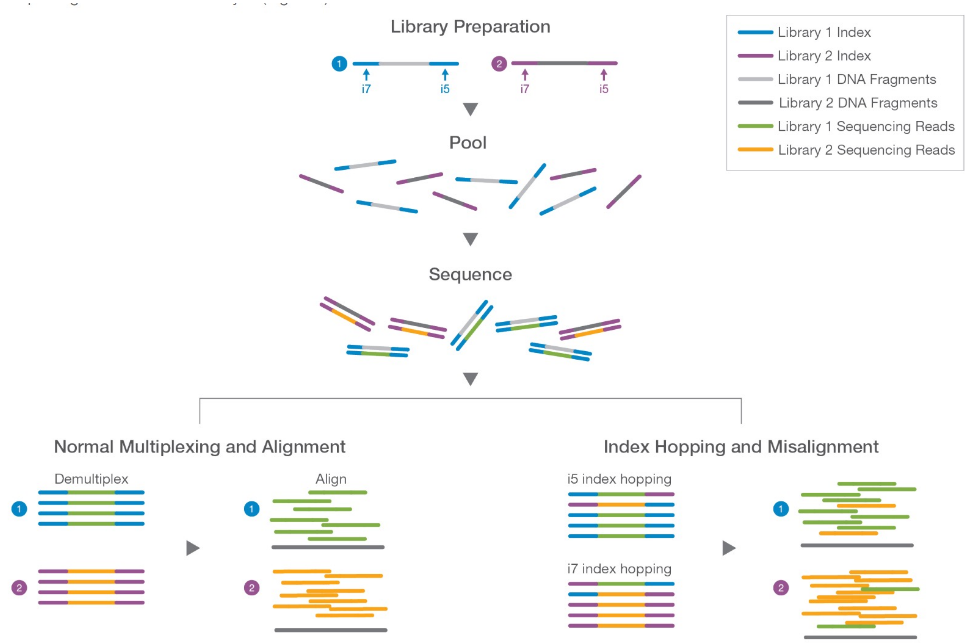
Image adapted from the Nextera XT DNA Library Prep Kit Reference Guide (© 2017 Illumina, Inc.)

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



https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf

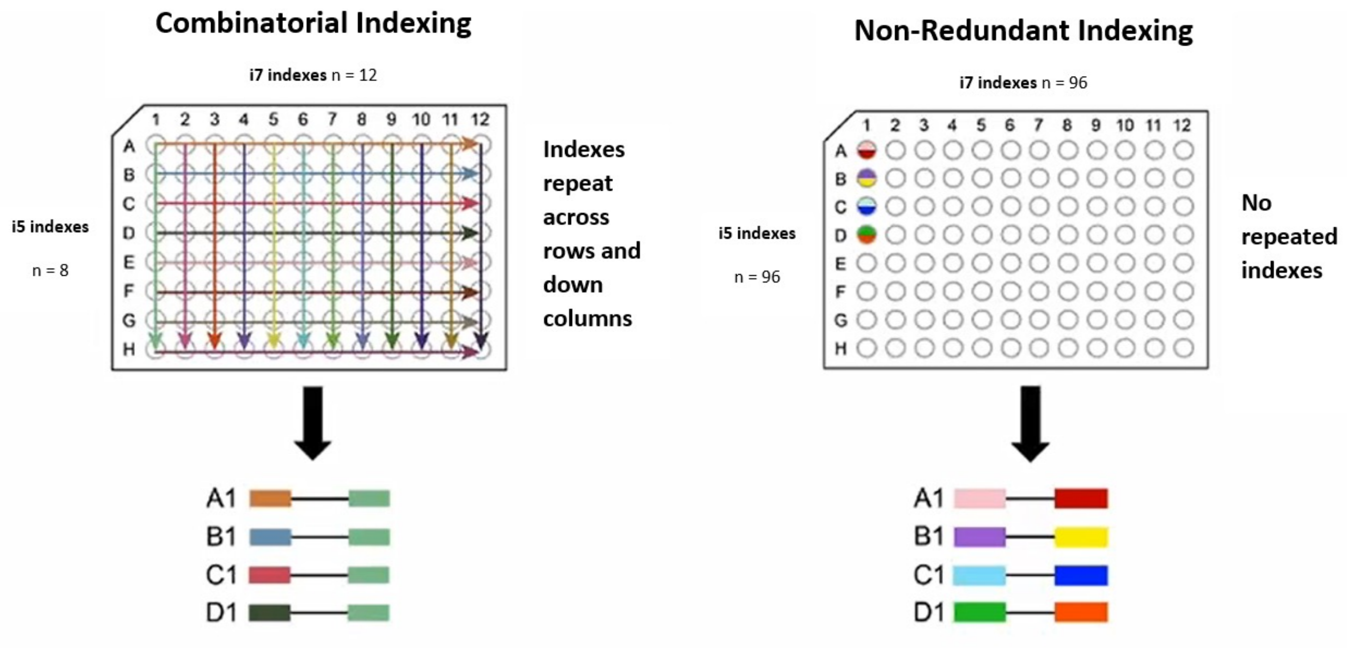
Beware that index hopping can cause misassigned sequence reads when using combinatorial index barcodes



Index hopping often affects 0.1-2% of reads!

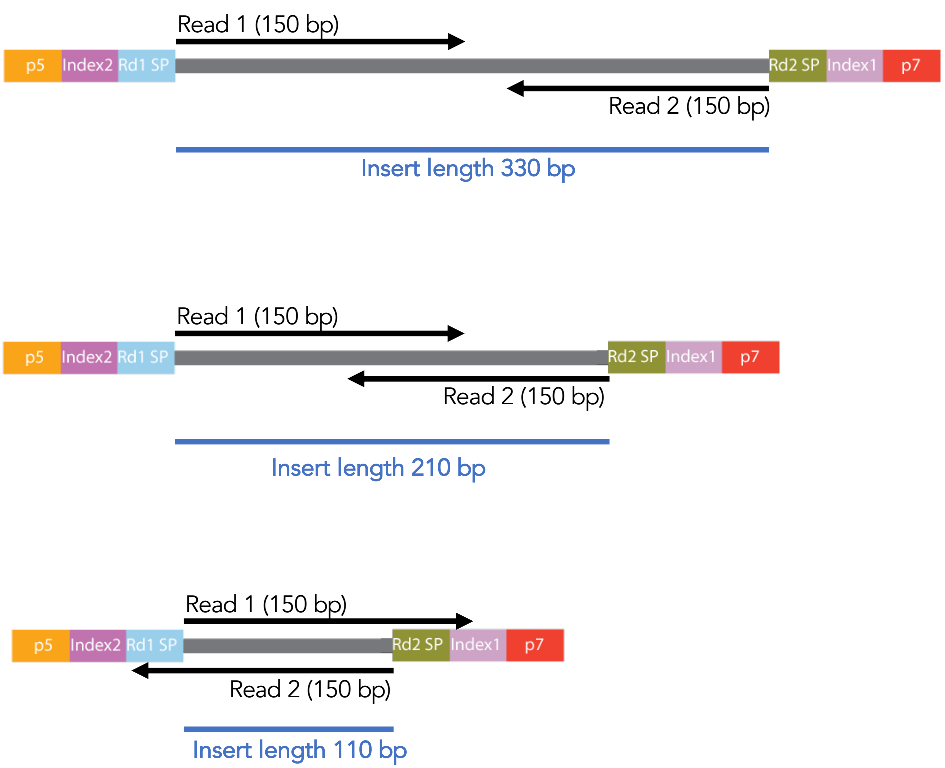
https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf

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

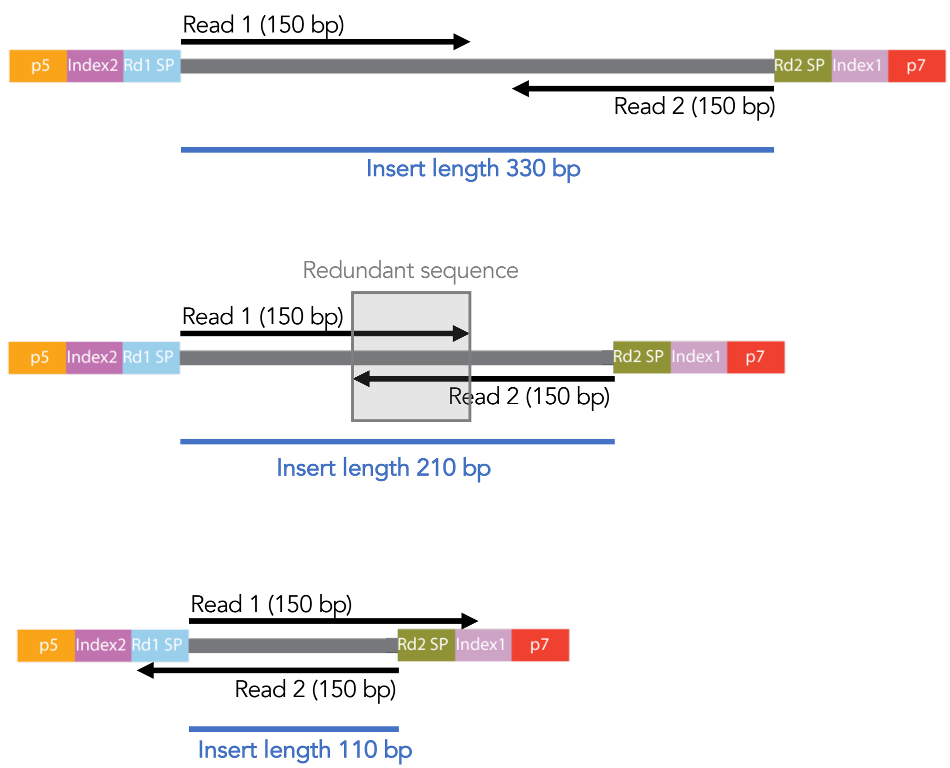


https://support.illumina.com/bulletins/2018/08/understanding-unique-dual-indexes--udi--and-associated-library-p.html

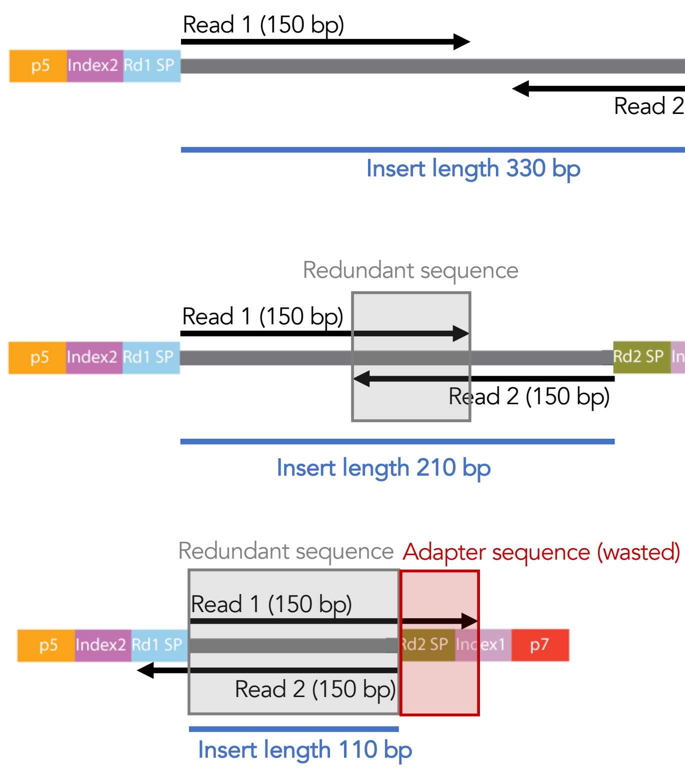
Insert length relative to read length



Insert length relative to read length

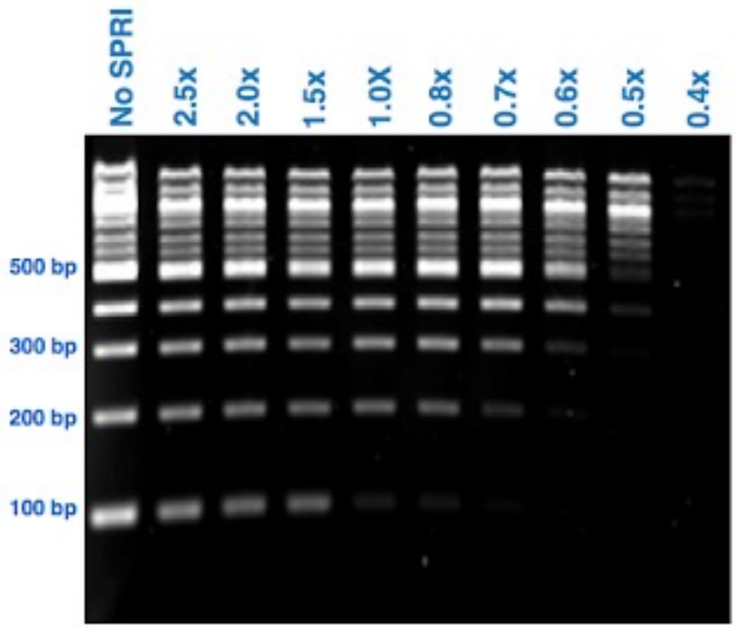


Insert length relative to read length



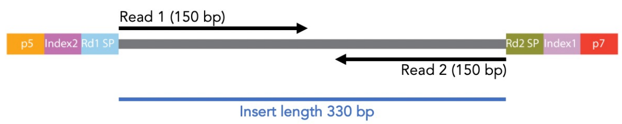
Size selection with Ampure beads

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



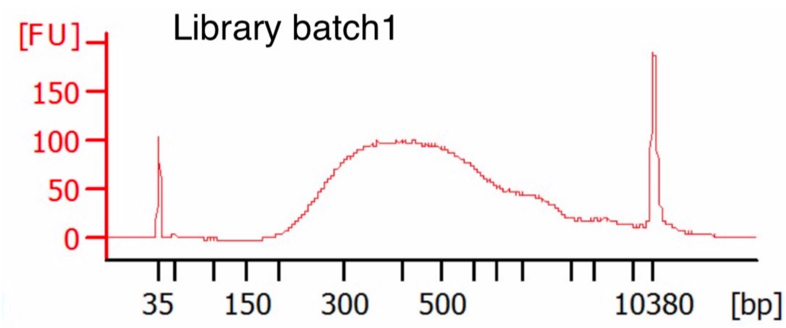
http://enseqlopedia.com/2012/04/how-do-spri-beads-work/

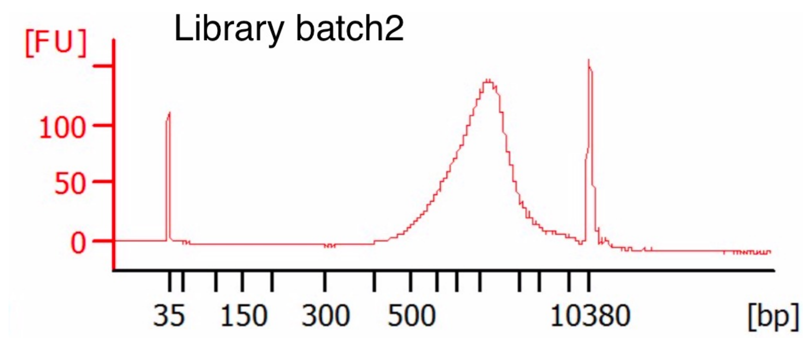
Ideally, we want all library fragments to be greater than the adapter length plus 2 x the read length (for PE)



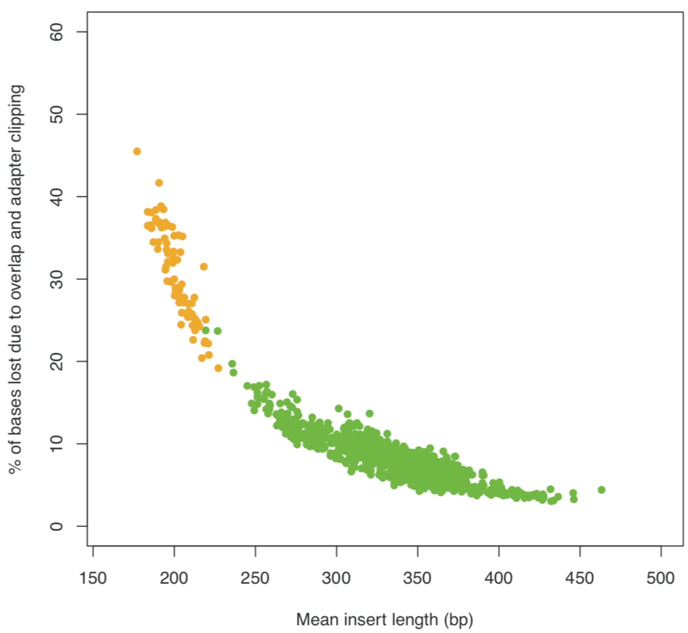
Ideal minimum fragment length

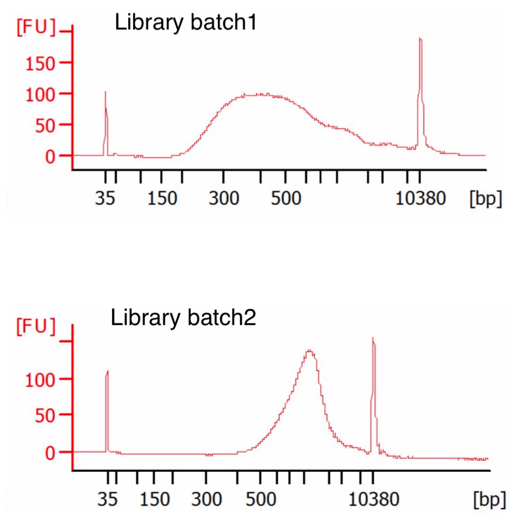
Two examples of our library pools



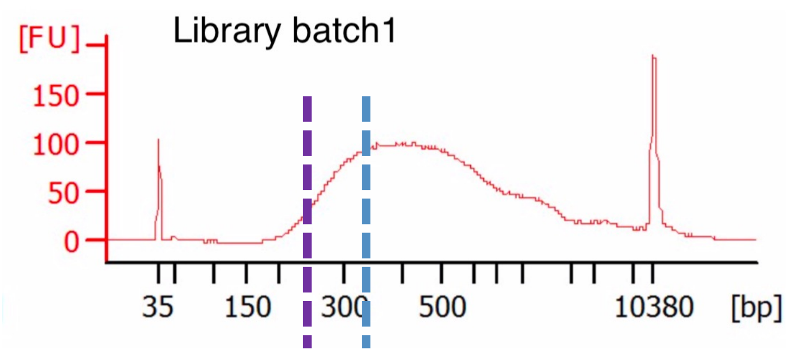


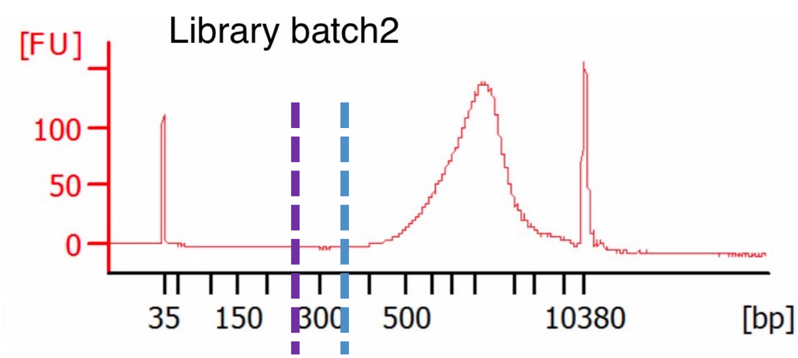
The library fragment size distribution can substantially influence the amount of data lost in data QC steps





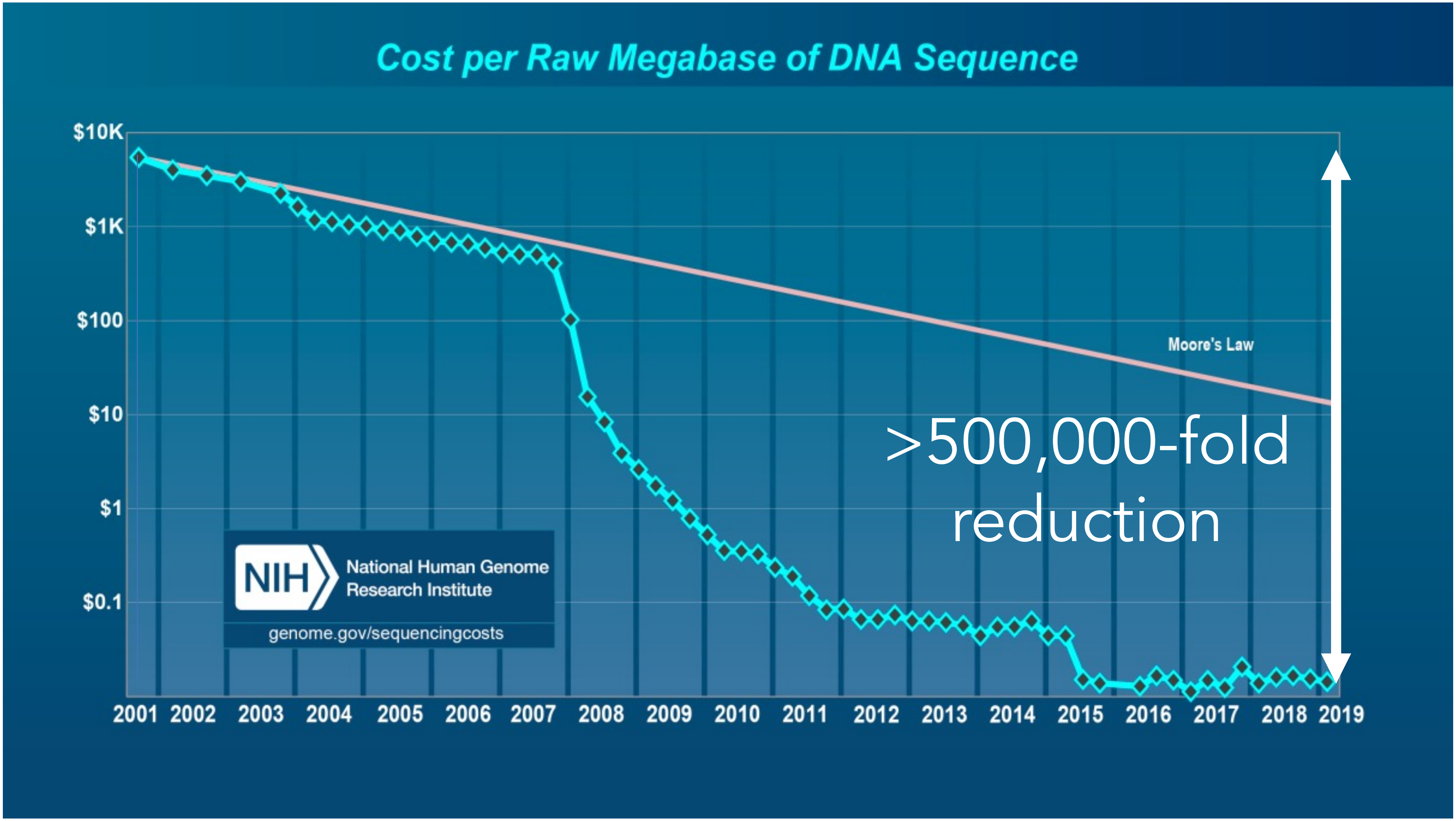
Two examples of our library pools





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**