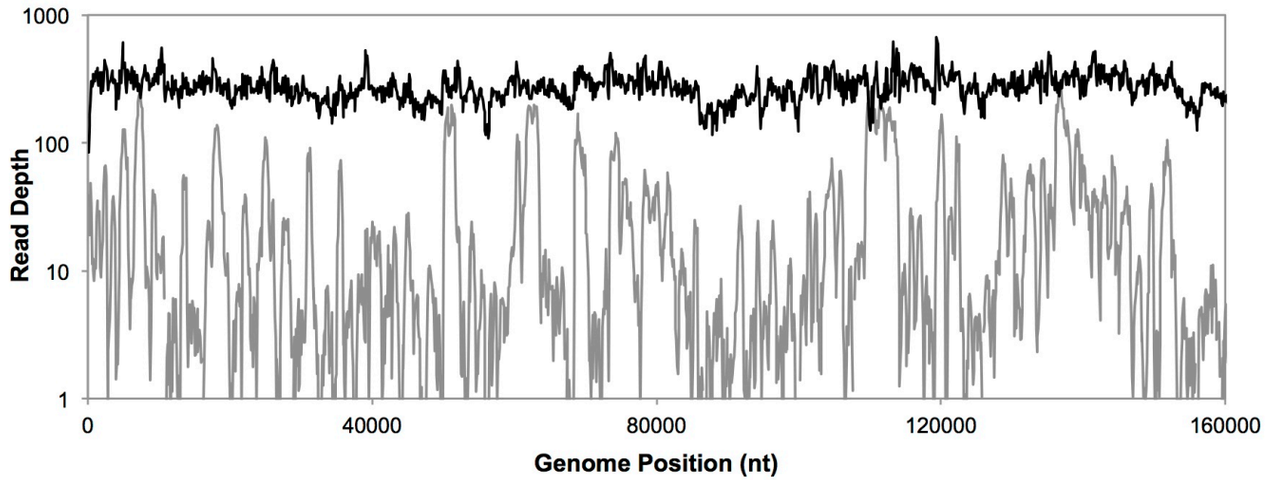
Practical Choices in Sequencing Projects



Frequently Asked Questions in NGS Library Construction

- **Platform**. Which Illumina sequencing platform is best? Or should I be using long-read technologies?
- **Read-Lengths**. How many sequencing cycles should I run?
- **Paired-End.** Should I do paired-end or single-read sequencing?
- **Read Number.** How many reads should I generate?
- **PCR**. How many PCR cycles should I do and which polymerase should I use?

Short-Read vs Long-Read Next-Generation Sequencing Techniques

Short-Read Sequencing

Illumina

Long-Read (Single Molecule) Sequencing

Oxford Nanopore

PacBio

Under what circumstances would you want to use short-read vs. long-read sequencing?

Illumina Sequencing Platforms

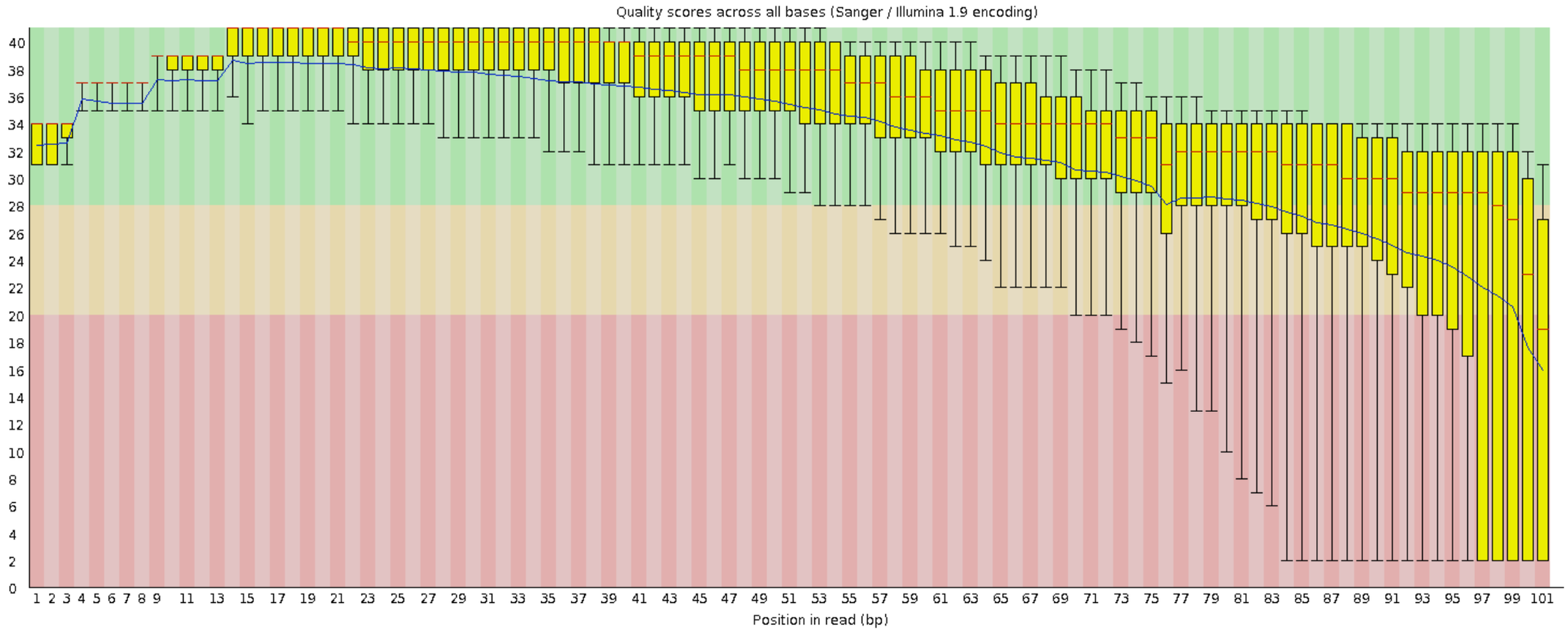
	Clusters (millions)	Max Read-Length	Max Output (Gb)	Cost	Bacterial Genomes	Eukaryotic Transcriptomes
MiniSeq	25	150 bp	7.5	$1,500	15	1.5
MiSeq	25	300 bp	15	$1,530	30	3
NextSeq 500 (mid)	130	150 bp	40	$1,650	80	8
NextSeq 500 (high)	400	150 bp	120	$4,240	240	24
HiSeq 4000 Lane	300	150 bp	90	$1,925	180	18
NovaSeq S4 Lane	2500	150 bp	750	$9,100	1500	150

Principles when Choosing Read Lengths

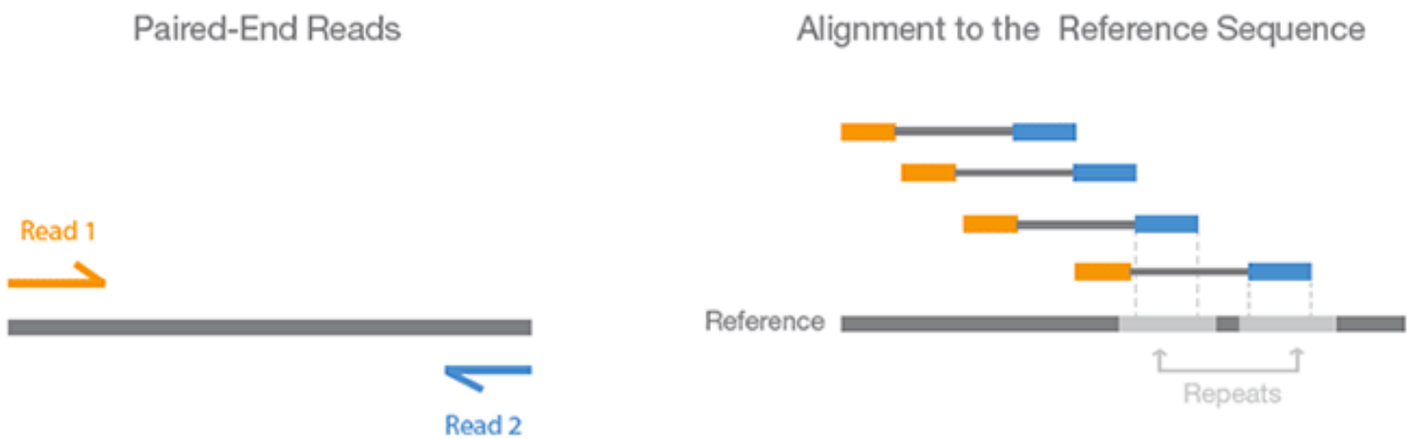
- Read lengths are defined by the cycle number on an Illumina run (1 bp per cycle).
- Advantages of longer reads
    - Cost per bp declines
        - MiSeq V2 50 cycles -- $747 ($996/Gb)
        - MiSeq V2 300 cycles -- $958 ($213/Gb)
        - MiSeq V2 500 cycles -- $1066 ($142/Gb)
    - Better for distinguishing among repetitive sequences in assembly/mapping
- Disadvantages
    - Worthless if your inserts are short
    - Additional sequence is not “independent” (e.g., for quantifying gene expression)
    - Basecall quality diminishes with read length.

Read Quality

Quality declines with increasing cycle number because amplicons within clusters get out of phase.



Single-Read vs. Paired-End



Advantages of paired-end runs

- Cost per bp declines
  - NextSeq V2 1x150 cycles -- $2650 ($44/Gb)
  - NextSeq V2 2x150 cycles -- $4240 ($35/Gb)

- Better for distinguishing among repetitive sequences in assembly/mapping

Disadvantages

- Worthless/redundant if your inserts are short

- Additional sequence is not “independent” (e.g., for quantifying gene expression)

How Many Reads Do I Need?

de novo genome assembly

- 100x sequence coverage (e.g., 5 Mb genome → 500 Mb total sequence data)
- Longest paired-end reads available
- But consider PacBio/Nanopore

Genome re-sequencing (SNP and indel variant calling)

- 20x and 35x for haploid and diploid genomes, respectively
- Longest paired-end reads available
- Low error rate technologies

RNA-seq for measuring gene expression (with ref genome)

- 36 million reads to get reliable quantification for 80% of human genes with FPKM > 10. (ENCODE 2011 PLoS Biol. e1001046)
- Short single-end reads

Table 2 | **Representative read counts for location-based approaches**

Techniques	Read counts in representative studies	Refs
DNaseI-seq and FAIRE–seq	20–50 million	79
CLIP–seq	7.5 million; 36 million	89, 90
iCLIP and PAR–CLIP	8 million; 14 million	105, 106
CHiRP and CHART	26 million	72
4C	1–2 million	92
ChIA–PET	20 million	107
5C	25 million	108
Hi-C	>100 million	94
MeDIP–seq	60 million	109
CAP–seq	>20 million	110
ChIP–seq	>10 million per sample (point source); >20 million per sample (broad source)	79

Sims et al. 2014 (Nature Reviews Genet. 15: 121-132)

Minimizing Bias

Generally try to use a minimal number of PCR cycles during library prep.

Use high-fidelity polymerases that exhibit a low amplification bias (KAPA HiFi or NEB Q5)

Discussion of additional sources of bias: van Dijk et al. 201
https://www.ncbi.nlm.nih.gov/pubmed/24440557