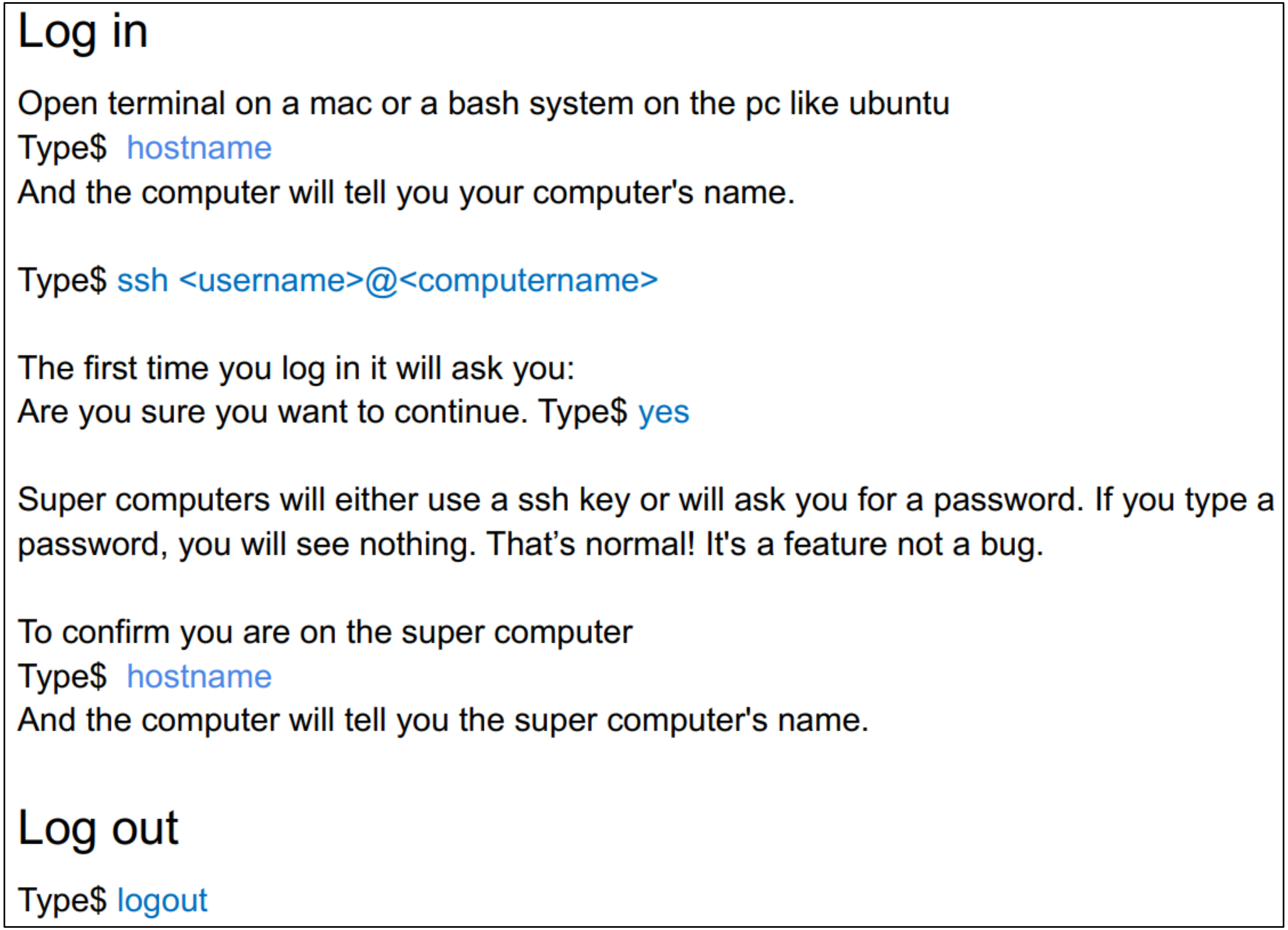
Logging on to a cluster



Log in

Open terminal on a mac or a bash system on the pc like ubuntu
Type$ hostname
And the computer will tell you your computer's name.

Type$ ssh <username>@<computername>

The first time you log in it will ask you:
Are you sure you want to continue. Type$ yes

Super computers will either use a ssh key or will ask you for a password. If you type a password, you will see nothing. That's normal! It's a feature not a bug.

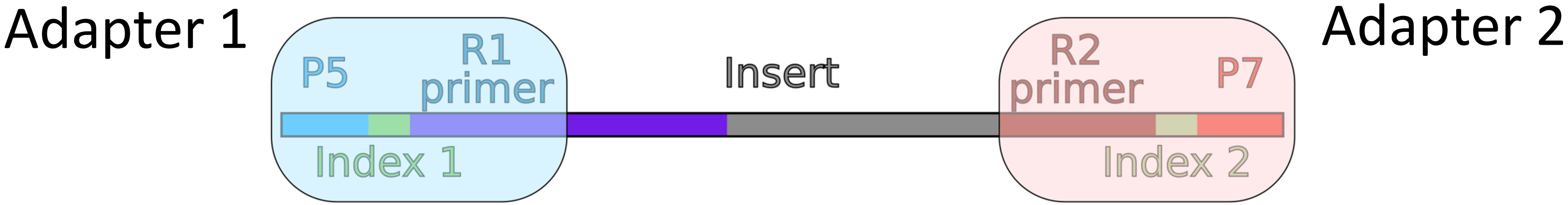
To confirm you are on the super computer
Type$ hostname
And the computer will tell you the super computer's name.

Log out

Type$ logout

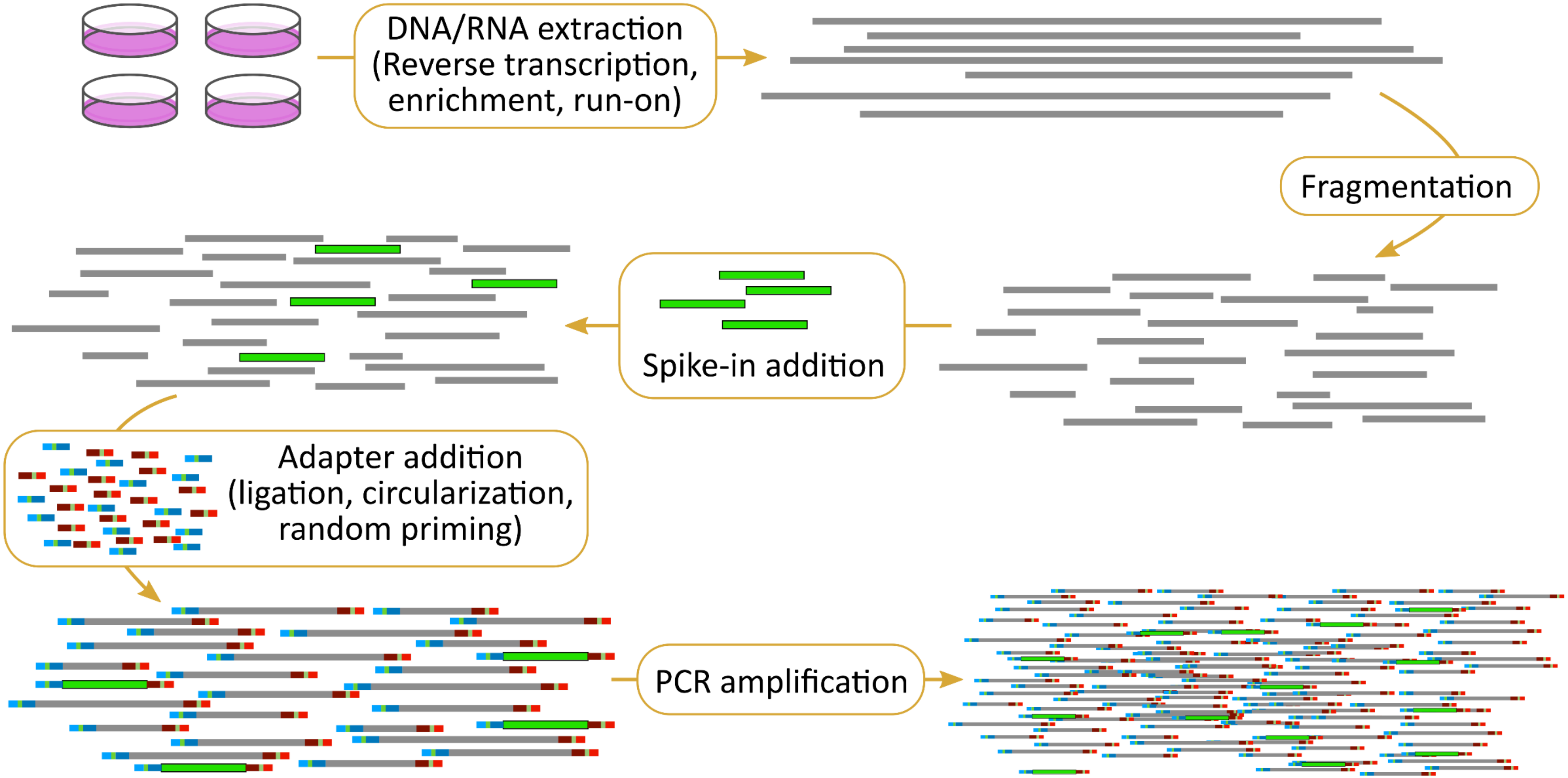
Library prep and QC

Anatomy of a library



P5/P7	Ends that attach to flow cell
Index 1/2	ID sequences for multiplexing samples
R1/R2 primers	Sequencing primers
Insert	Fragment of sample DNA/cDNA
Read	Sequenced portion of fragment

Creating libraries



DNA/RNA extraction (Reverse transcription, enrichment, run-on)

Fragmentation

Spike-in addition

Adapter addition (ligation, circularization, random priming)

PCR amplification

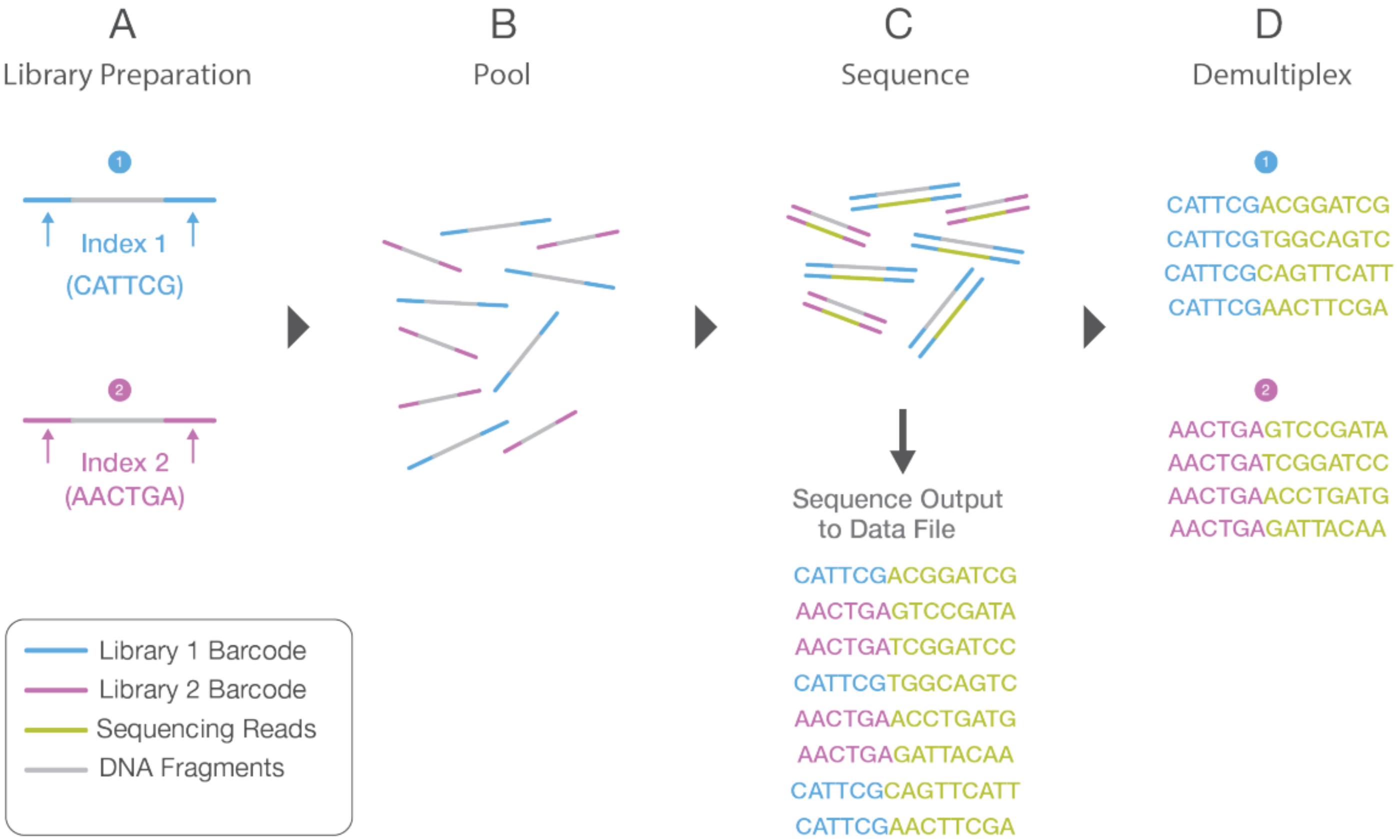
Library kits

- Your protocol will determine whether you use a kit
  - Whole genome/RNA sequencing mostly use kits
  - ChIP-seq, ATAC-seq, more specialized protocols do many steps outside of kits

- Kit considerations:
  - How much input do you have? (> µg, < 10 ng, single-cell)
  - What quality input do you have?
  - Do you need to worry about fragmentation or amplification biases?
  - RNA: do you want total, poly-A, micro, or ribosomal-depleted RNA?
  - RNA: do you want a strand-specific library? (Yes)

www.illumina.com

Library multiplexing



A
Library Preparation

1

Index 1
(CATTCG)

2

Index 2
(AACTGA)

B
Pool

C
Sequence

Sequence Output
to Data File

CATTCGACGGATCG
AACTGAGTCCGATA
AACTGATCGGATCC
CATTCGTGGCAGTC
AACTGAACCTGATG
AACTGAGATTACAA
CATTCGCAGTTCATT
CATTCGAACTTCGA

D
Demultiplex

1

CATTCGACGGATCG
CATTCGTGGCAGTC
CATTCGCAGTTCATT
CATTCGAACTTCGA

2

AACTGAGTCCGATA
AACTGATCGGATCC
AACTGAACCTGATG
AACTGAGATTACAA

Library 1 Barcode
Library 2 Barcode
Sequencing Reads
DNA Fragments

www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Choosing indeces

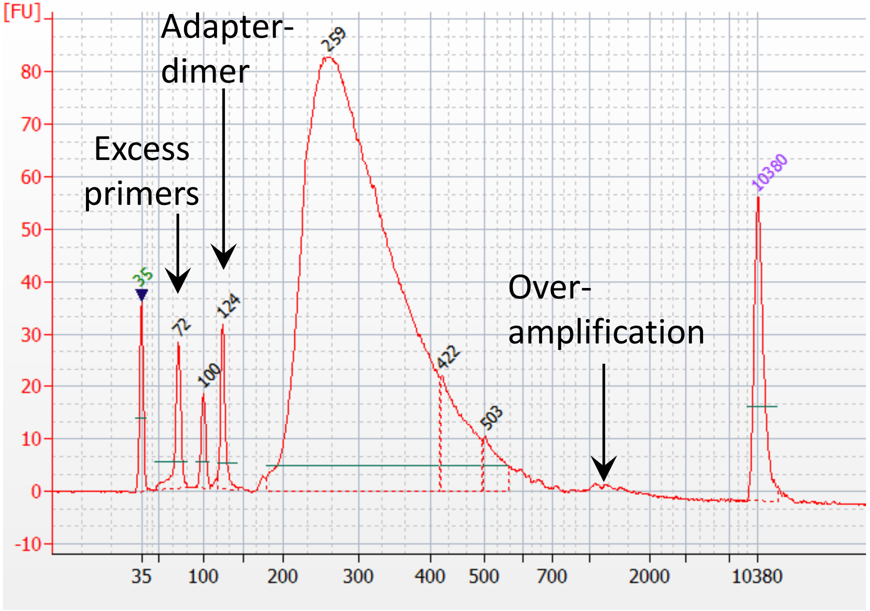
- Single indexing
- Combinatorial dual indexing
- Unique dual indexing
- Unique molecular identifiers
- Considerations:
    - Base diversity
    - Index hopping
    - Ease of deconvolution

Library quality control

Before sequencing		After sequencing
Bioanalyzer	Contamination: organism, nucleic acid, adapter	FastQC
	Base diversity	FastQC
Bioanalyzer	Insert size	picardtools
	Complexity	FastQC, preseq, picardtools
Qubit, qPCR	Concentration/quantity	Illumina stats

Pre-sequencing QC

- Size electrophoresis (Bioanalyzer)
- Fluorimeter (Qubit)
- qPCR for P5/P7
- Rarely see the same conc. among the three methods
- qPCR:Qubit molar ratios for well-performing libraries are 0.8-2.0



It is better to make a new library than to sequence a terrible library!

Library quality control

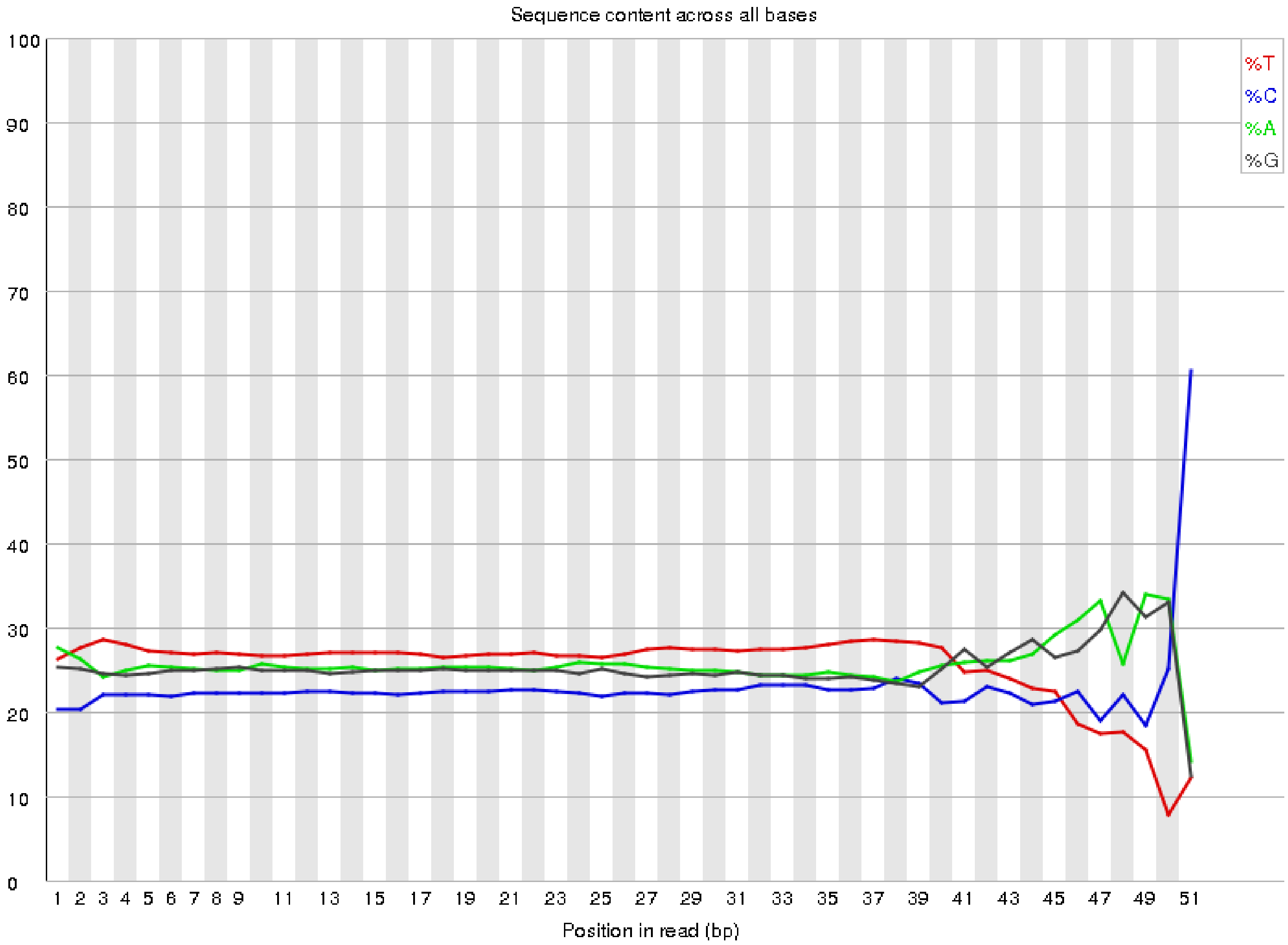
Before sequencing		After sequencing
Bioanalyzer	Contamination: organism, nucleic acid, adapter	FastQC
	Base diversity	FastQC
Bioanalyzer	Insert size	picardtools
	Complexity	FastQC, preseq, picardtools
Qubit, qPCR	Concentration/quantity	Illumina stats

FastQC

Base diversity

Complexity

Per base sequence content

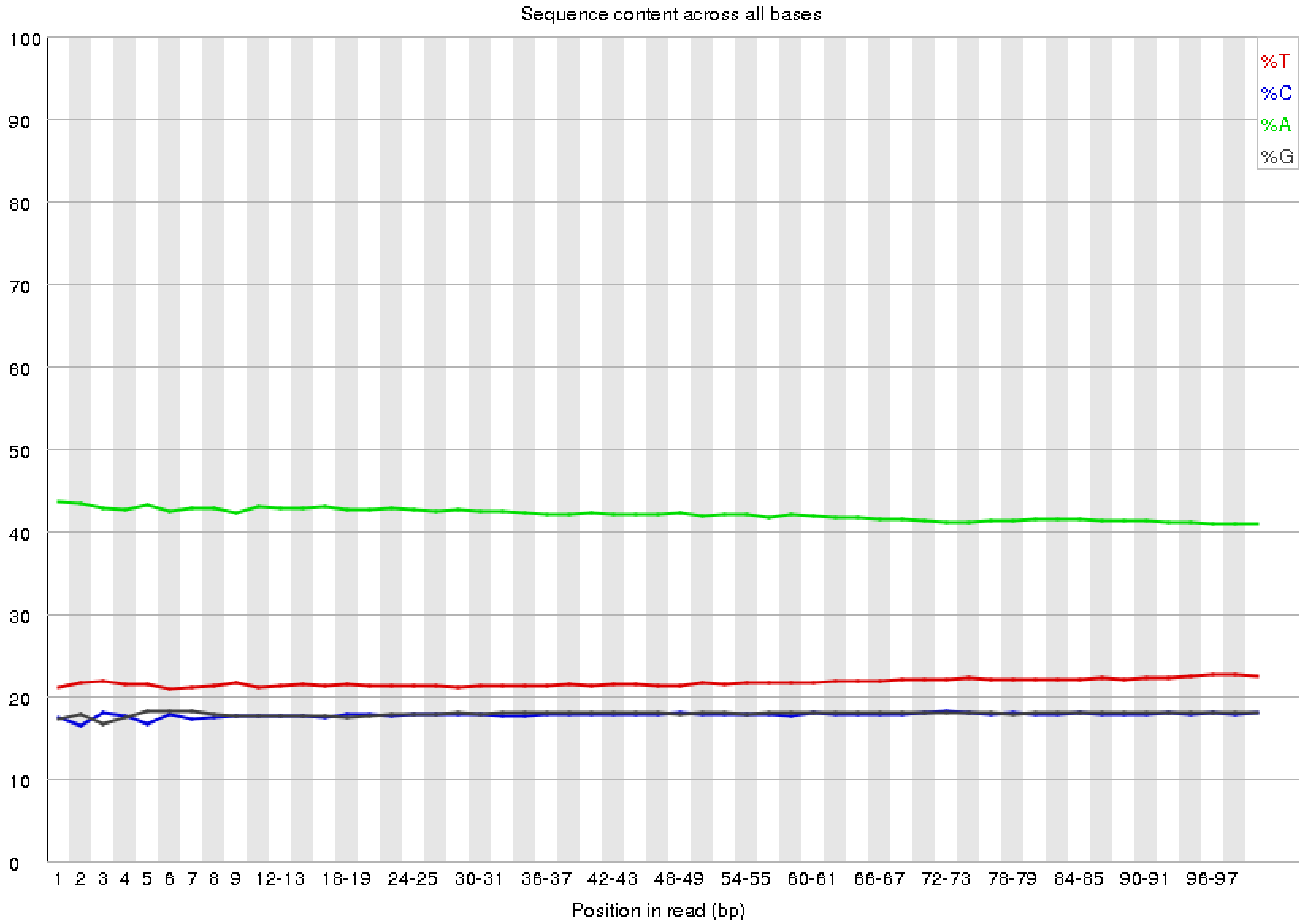


FastQC

Base diversity

Complexity

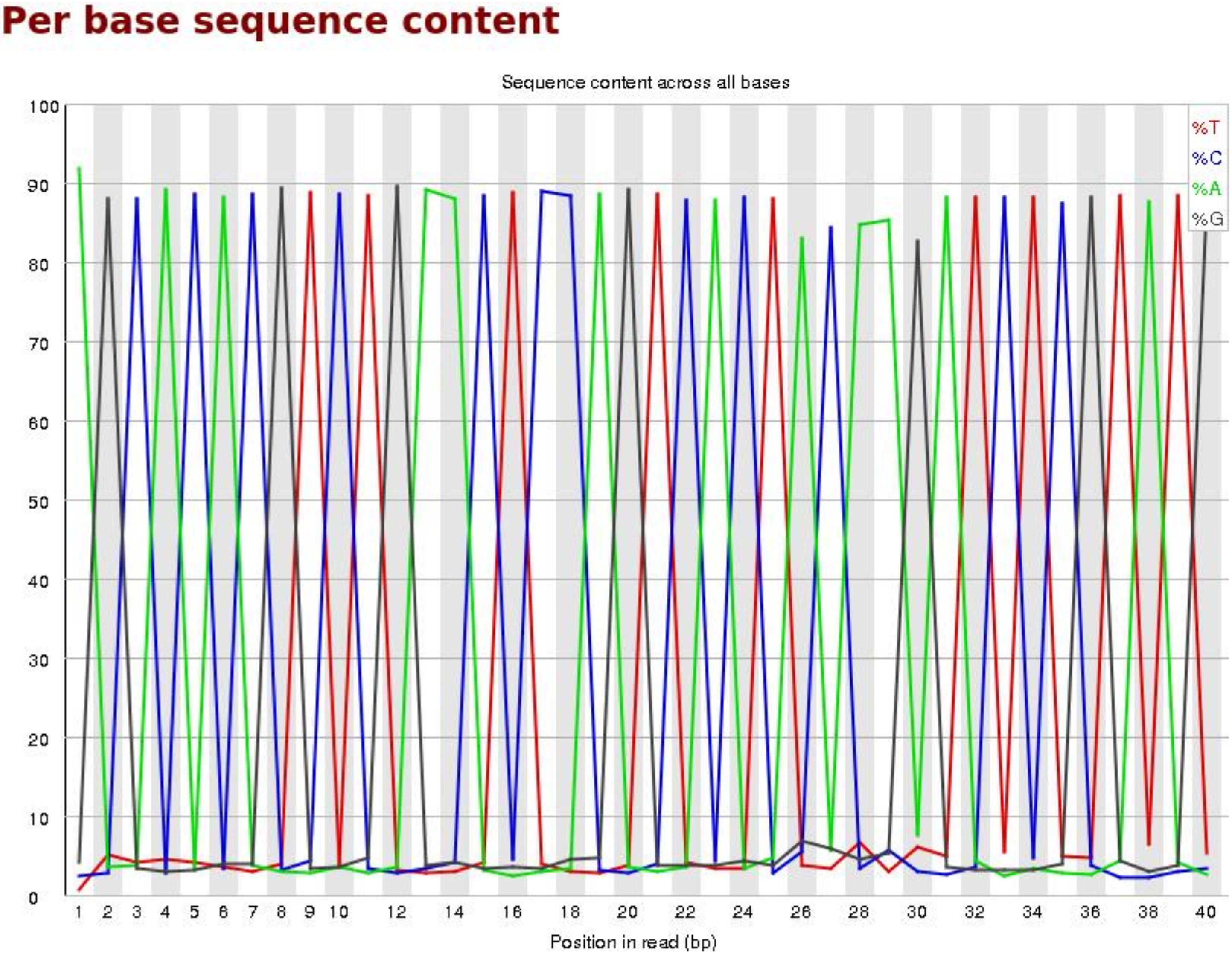
Per base sequence content



FastQC

Base diversity

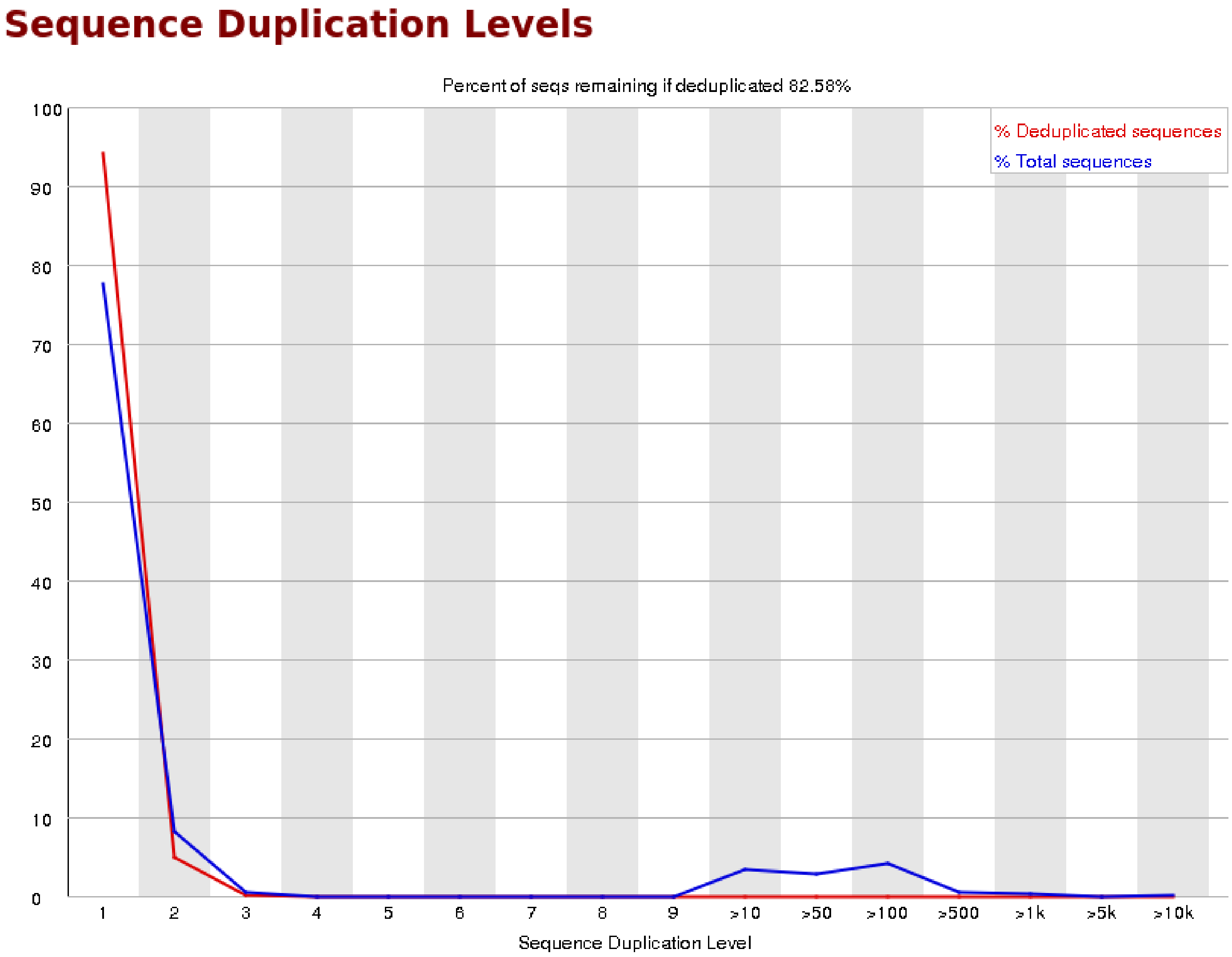
Complexity



FastQC

Complexity

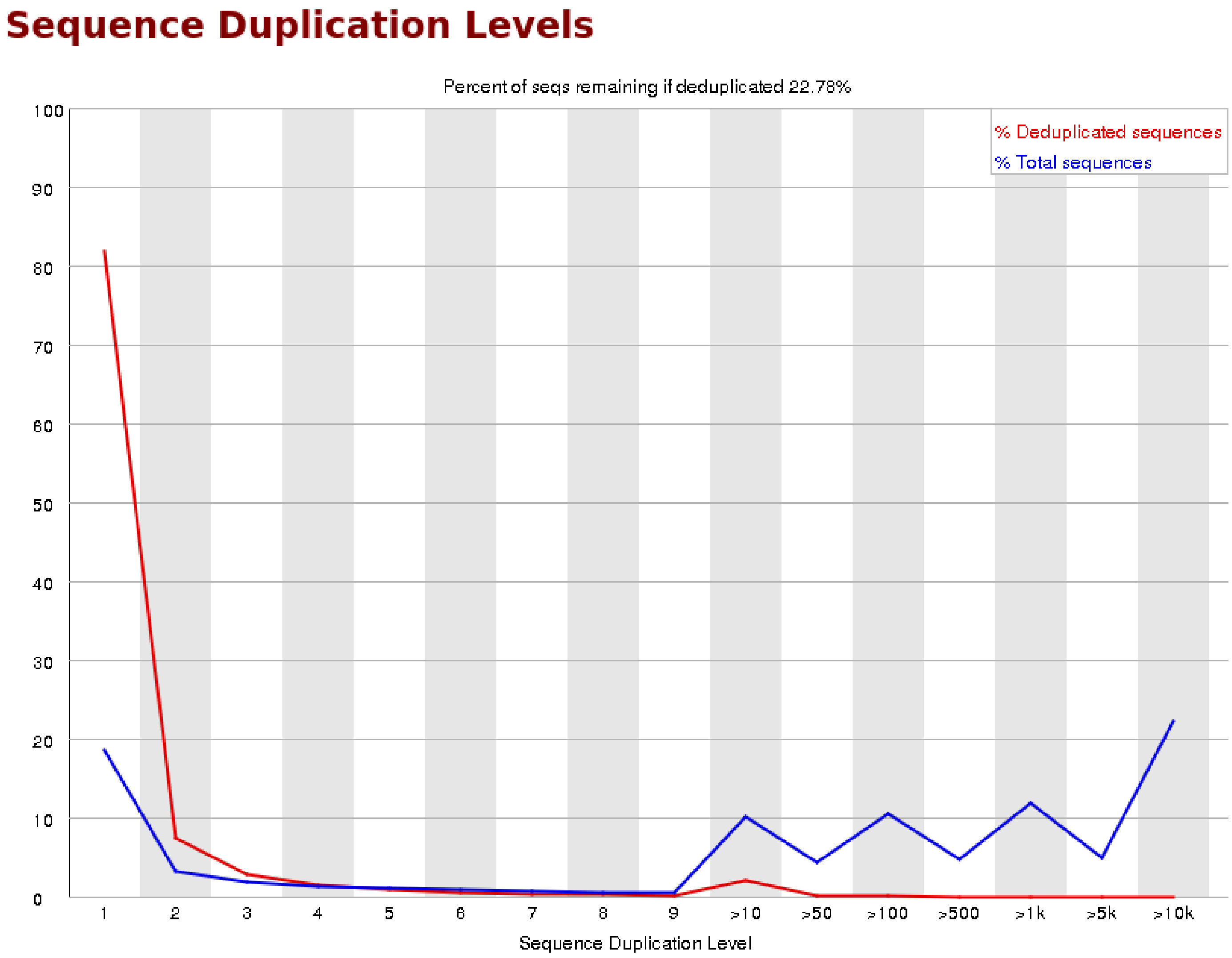
Duplication



FastQC

Complexity

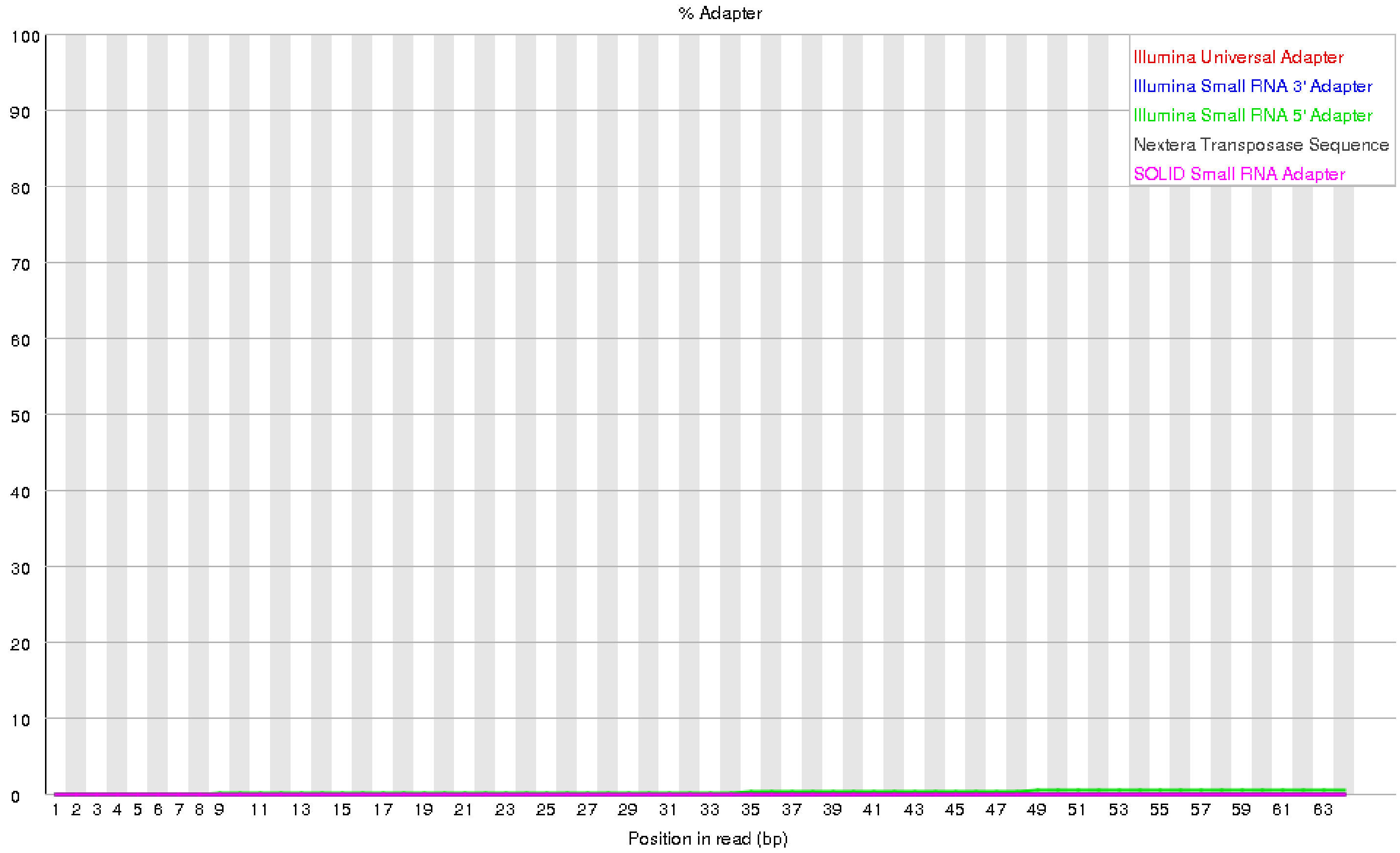
Duplication



FastQC

Adapter Contamination

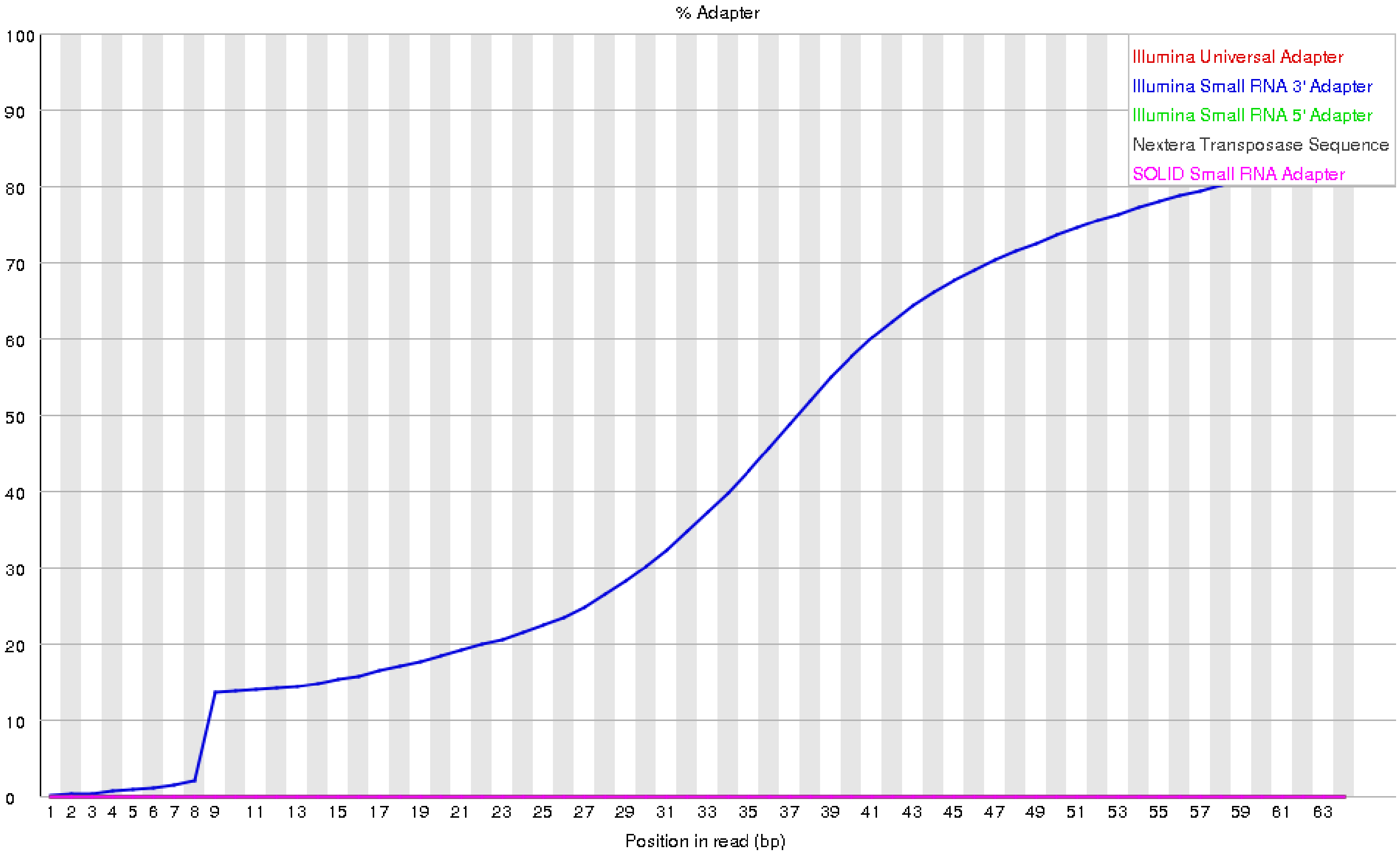
Adapter Content



FastQC

Adapter Contamination

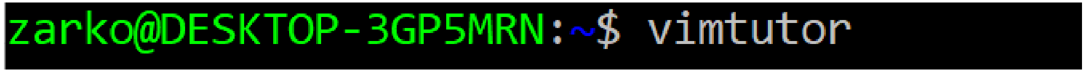
Adapter Content



Break!

VIM and vimtutor

- What is VIM?
  - Text editor – read, write and save text files
  - Entirely keyboard-based
  - You cannot use your mouse to move the cursor!
- vimtutor is on every linux system and teaches you how to use vim – open it now



```
zarko@DESKTOP-3GP5MRN:~$ vimtutor
```

Illumina sequencing

Sequencing technologies

Short read sequencing

(37 to 250 bases)

Illumina

Roche 454

Applied Biosystems SOLiD

Complete genomics Nanoball

Thermo Fisher Ion Torrent

Long read sequencing

(10 to >50 kb)

Pacific Biosciences SMRT

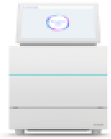
Oxford Nanopore

Illumina sequencing technology

Imaging a slide (flow cell) with millions/billions of DNA clusters by cycling in fluorescent nucleotides

https://www.youtube.com/watch?v=fCd6B5HRaZ8

Illumina sequencing platforms

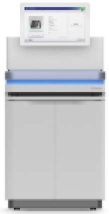


iSeq 100 System

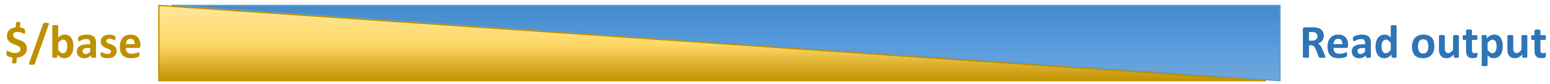
MiSeq Series ⊕

NextSeq Series ⊕

HiSeq 4000 System

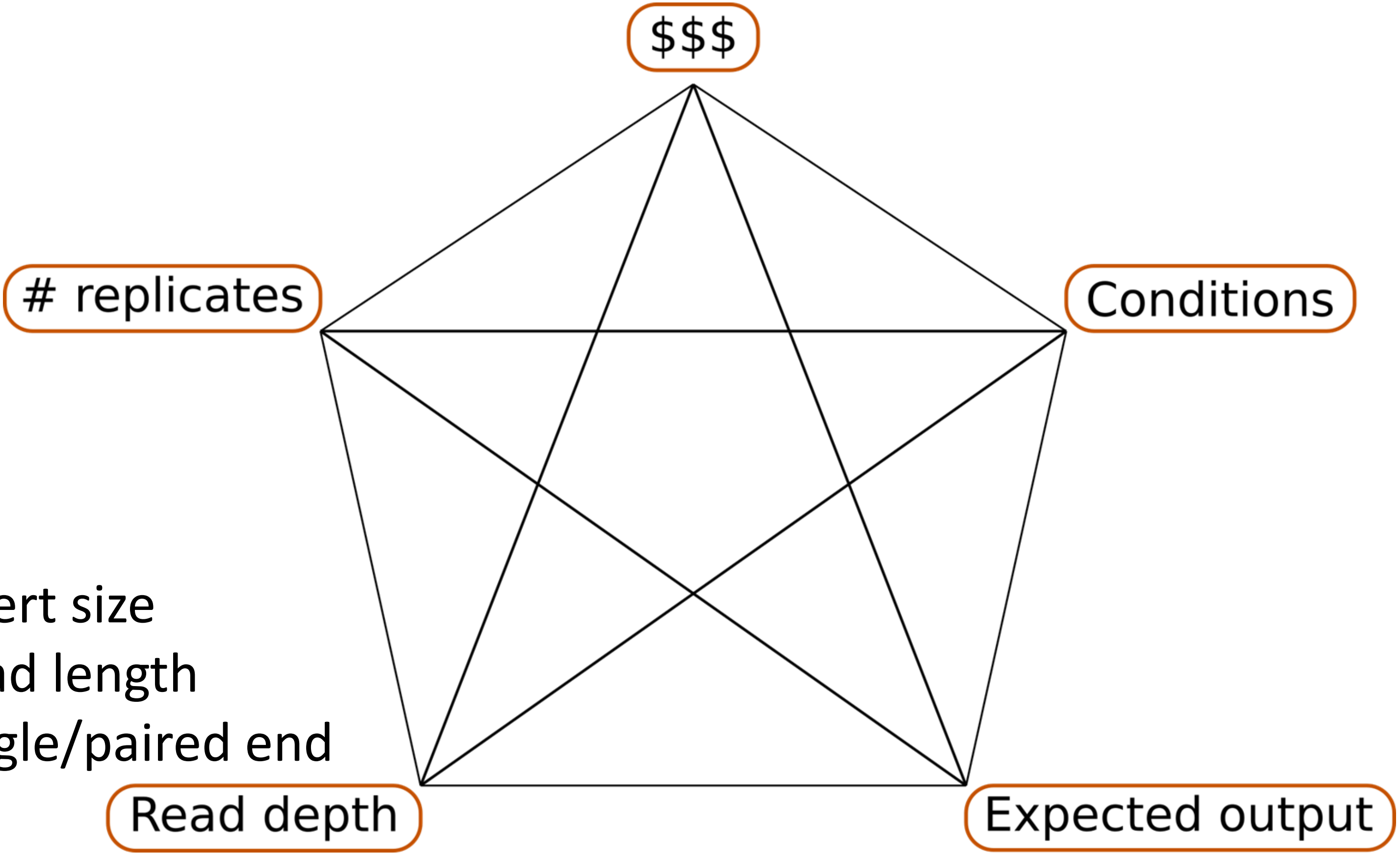


NovaSeq 6000 System



	iSeq 100	MiSeq	NextSeq	HiSeq 4000	NovaSeq
Run time	9-17.5 hr	9-55hr	12-30hr	1-3.5 days	13-44hrs
Throughput	1.2Gb	7.5-15Gb	120Gb	1500Gb	6000Gb
Read output	4M	12-25M	130-400M	6B	20B
Color system	1 channel	4 channel	2 channel	4 channel	2 channel
Flowcell	Patterned	Non-Patterned	Non-Patterned	Patterned	Patterned

Designing a sequencing experiment



$$$

replicates

Conditions

Read depth

Expected output

Insert size

Read length

Single/paired end

Read depth and expected outcomes

	Min. depth (in mammals)	Other specs
• RNA-seq DEA	20 x 10^6 reads/sample	SE/PE, insert size 100s
• Low-abundance RNA	50 x 10^6 reads/sample	SE/PE, insert size 100s
• Isoform analysis	50 x 10^6 reads/sample	PE, longer reads, insert size 100s
• Whole genome seq	15x coverage/sample	SE/PE, insert size 100s
• Heterozygous SNPs	30x coverage/sample	SE/PE, insert size 100s
• Indels	60x coverage/sample	PE, insert size 100s
• ChIP-seq	12 x 10^6 reads/sample	SE, insert size 100s
• Broad peaks	30 x 10^6 reads/sample	SE, insert size 100s
• microRNA	5-10 x 10^6 reads/sample	SE, short reads, small insert size

https://genohub.com/recommended-sequencing-coverage-by-application/

Questions?