

# Quality trimming

Jenni Hultman

Department of Microbiology

University of Helsinki

[Jenni.hultman@helsinki.fi](mailto:Jenni.hultman@helsinki.fi)

# Raw data Illumina

- We have paired end data on the sequenced genomes

A004\_07004-B\_TTGCATGT\_GACTCGCA\_run20171107N\_S4\_R1\_001.fastq

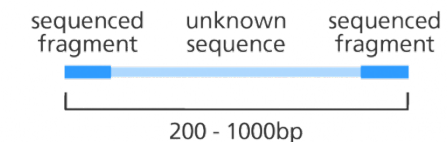
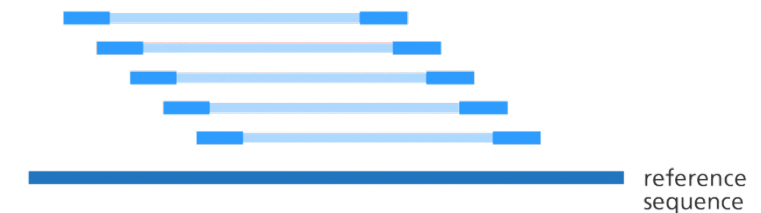
A004\_07004-B\_TTGCATGT\_GACTCGCA\_run20171107N\_S4\_R2\_001.fastq

- 150 + 150 bp
- Can contain
  - Sequence adapters
  - low quality sequence (usually in the end)
    - Occurrence: substitutions > indels
    - Quality scores: substitutions < indels
    - Overall quality: R1 > R2; beginning > end
- Need to check quality and trim the reads

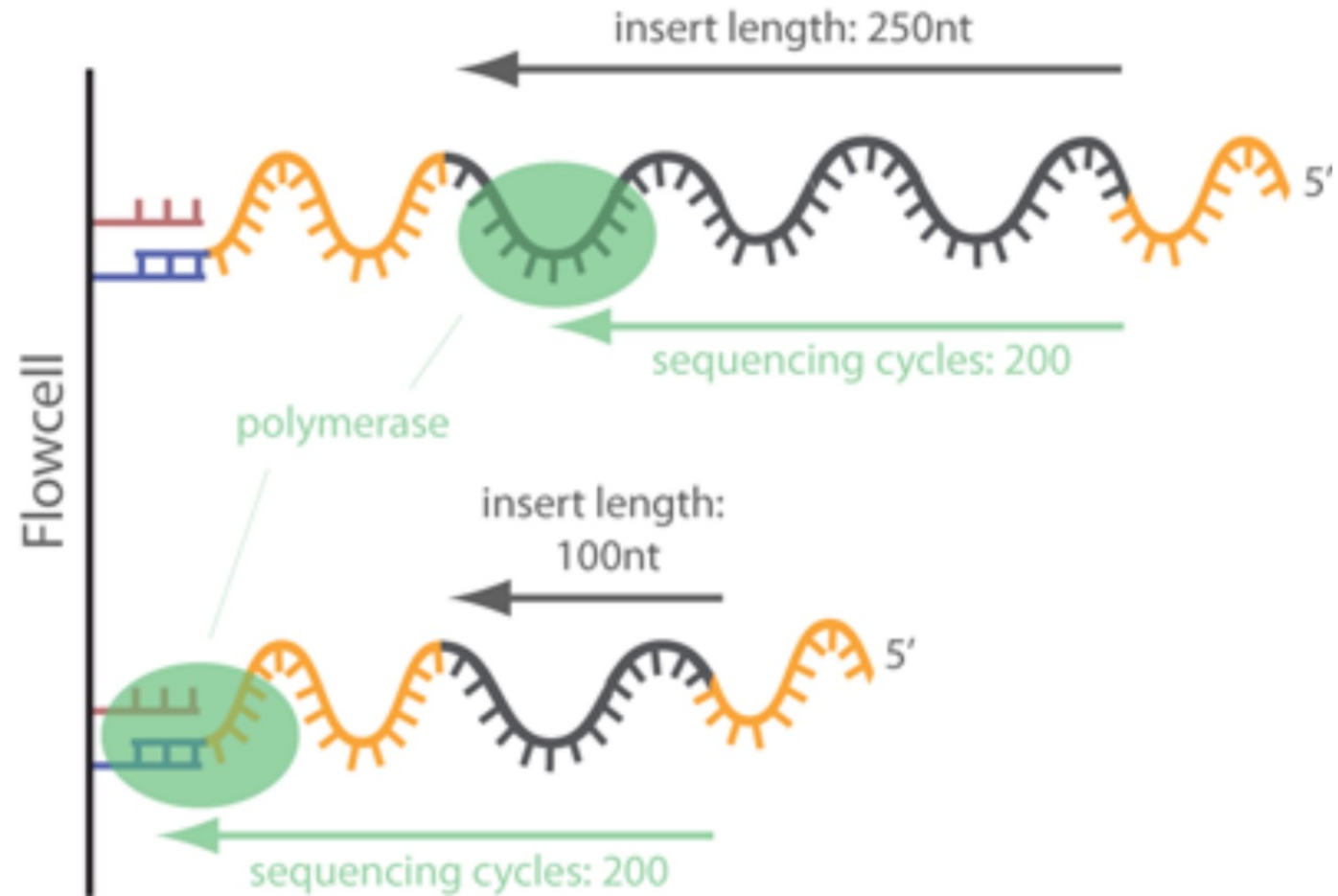
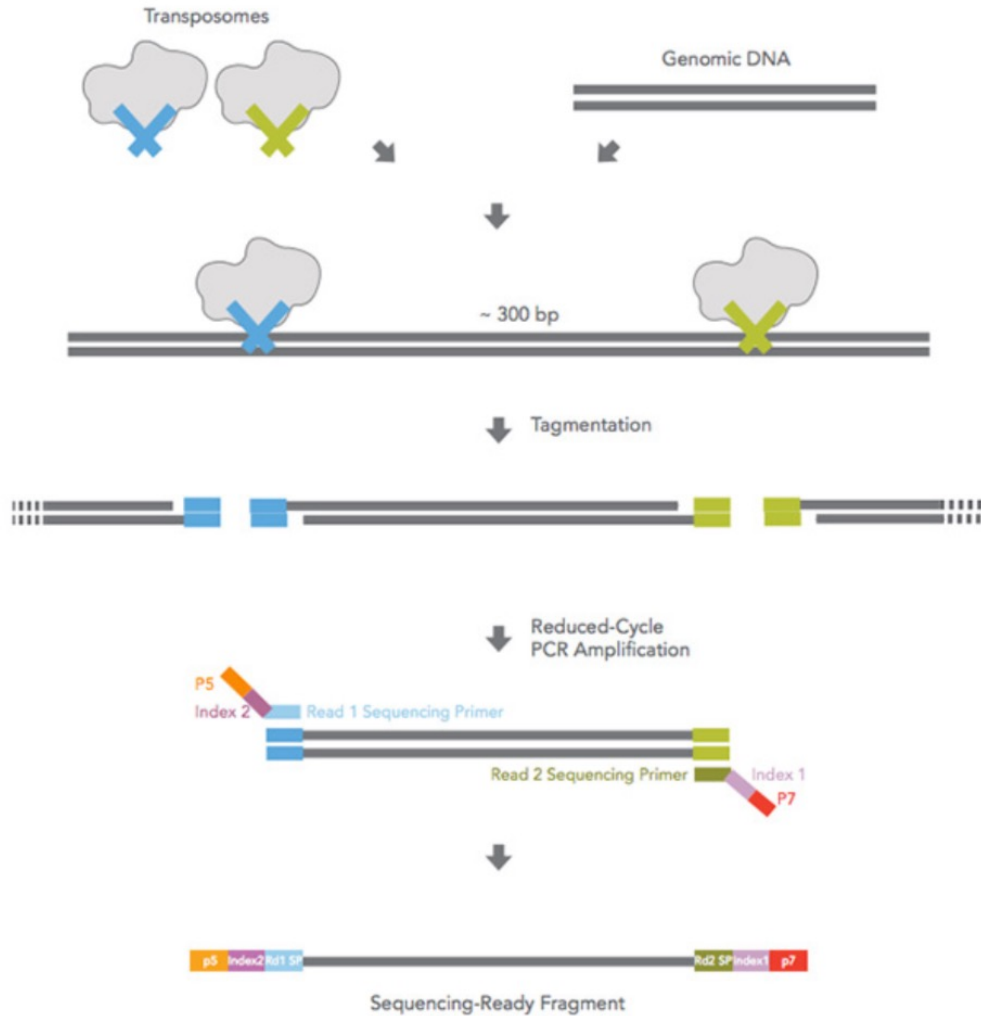
Single-end reads



Paired-end reads

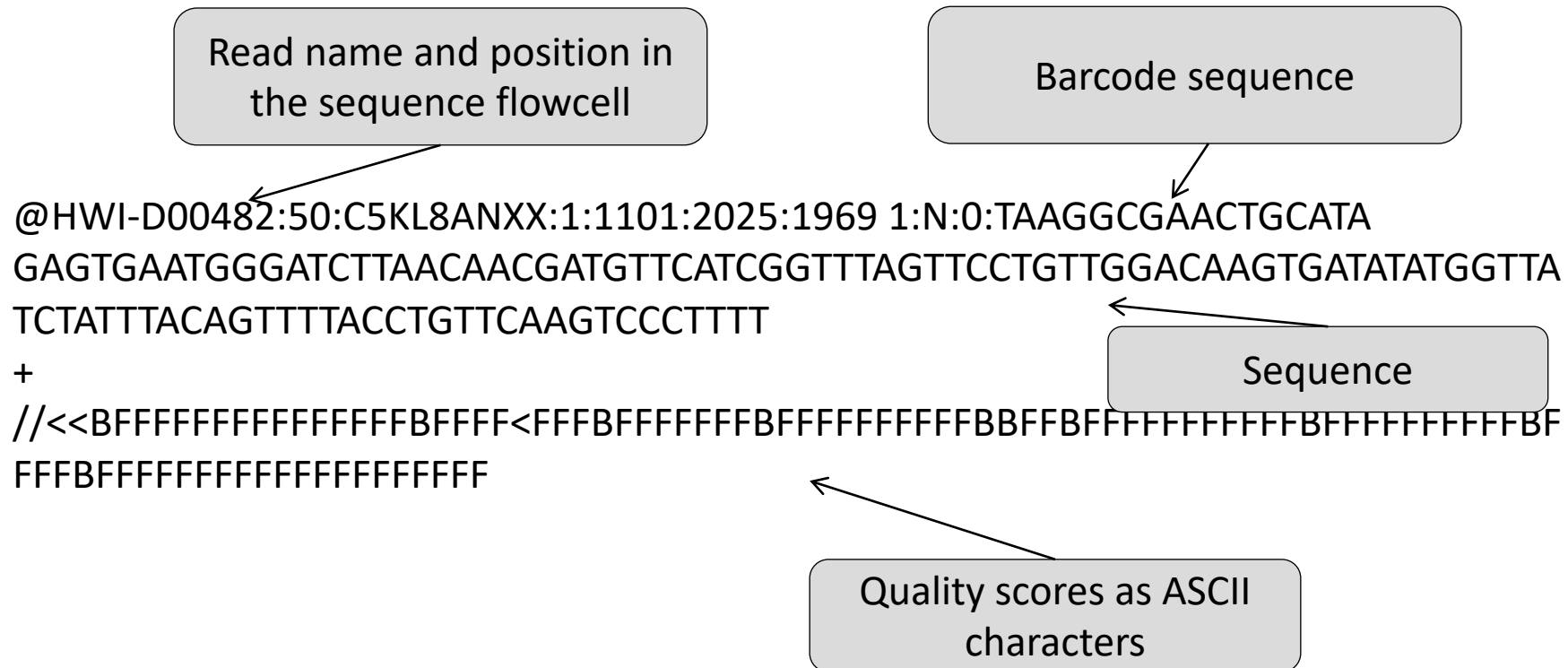


# Adapter contamination



# Fastq

- Sequence data is commonly delivered in FASTQ format. No chromatograms!



# Quality scores

- measure of the quality of the identification of the bases generated by sequencer
- Phred-score

Phred Quality Score	Probability of incorrect base call	Base call accuracy	ASCII
10	1 in 10	90%	+
20	1 in 100	99%	5
30	1 in 1000	99.9%	?
40	1 in 10000	99.99%	I

- Phred score above 20-25 considered as acceptable
  - 1 mistake in 100

```
@HWI-D00482:50:C5KL8ANXX:1:1101:2025:1969 1:N:0:TAAGGCGAACTGCATA
GAGTGAATGGGATCTTAACAACGATGTTTCATCGGTTTAGTTCCTGTTGGACAAGTGATATATGGTTATCT
ATTACAGTTTTACCTGTTCAAGTCCCTTTT
+
//<<BFFFFFFFFFFFFFFFFFBFFFF<FFFBFFFFFFFFBFFFFFFFFFBBFFBFFFFFFFFBFFFFFFFFBFFFF
BFFFFFFFFFFFFFFFFFFFFF
```

ASCII\_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 \$	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41 )	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

# FASTQC

- Quality assessment program
  - How the data looks like. No trimming.
  - <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Output of FASTQC is a zip archive and an HTML document
- Combine files with **multiqc**
- View the HTML in web browser

# How does the data look like?

- Where is the best quality sequence?
  - Begin, middle, end?
- Are there adapters?
  - What are adapters? Why to remove?
- Differences in R1 and R2?
  - Forward and reverse reads

**What kind of trimming do you think should be done?**



# Quality filtering

- Removal of low-quality regions and adapters
- Several programs available, we will use **cutadapt**  
<http://cutadapt.readthedocs.io/en/stable/>

# Cutadapt

- When looking at the cutadapt manual, which flags (=“-letter”) are for
  - Length trimming \_\_\_\_\_
  - 3’ adapter \_\_\_\_\_
  - Paired end 3’adapter \_\_\_\_\_
  - Quality score \_\_\_\_\_
  - Output name \_\_\_\_\_
  - Paired end output \_\_\_\_\_

# Decontamination

- How about host associated metagenomes?
- Removal of host DNA from the samples
- Map the reads with Bowtie or BWA to the host genome
  - What to do with the reads that map?

Garbage in – garbage out