Dealing with 'raw reads'

Analysis of Next-Generation Sequencing Data

Friederike Dündar

Applied Bioinformatics Core

Slides at https://bit.ly/2T3sjRg[1]

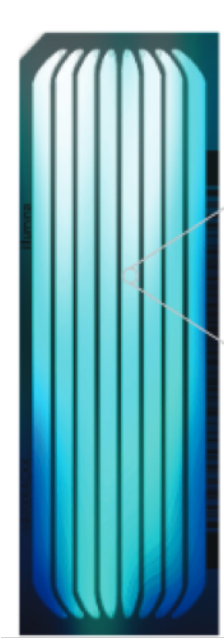
January 28, 2020

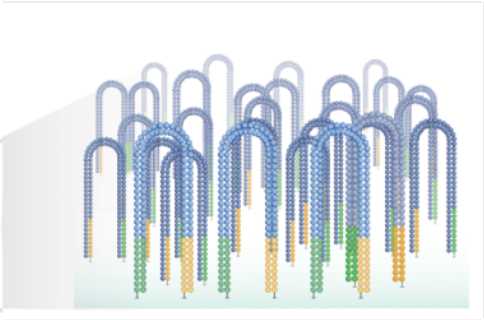
[1] https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/schedule_2020/

1. Fluorescence-based microscopy
2. Single and paired-end reads
3. Illumina's "raw reads"
4. Quality control of sequencing reads
5. Sequence Read Archive
6. References

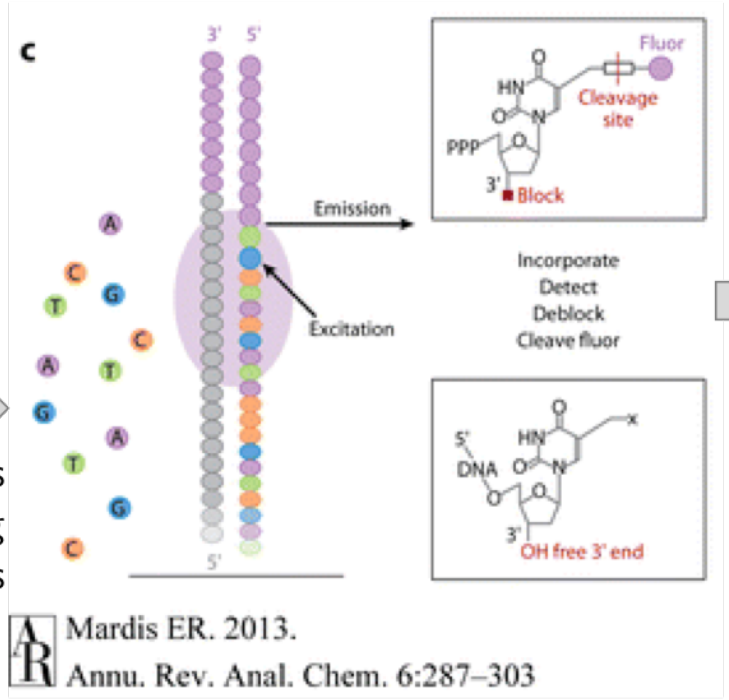
Fluorescence-based microscopy

Re-cap: Sequencing by synthesis after library preparation

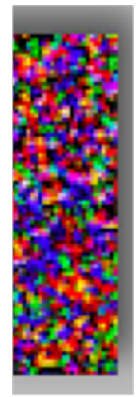




Clusters of identical fragments are sequenced using fluorophore-labeled dNTPs



Mardis ER. 2013.
Annu. Rev. Anal. Chem. 6:287–303



images of fluorescent signals from each cycle have to be analyzed to infer the order of the nucleotide within each cluster

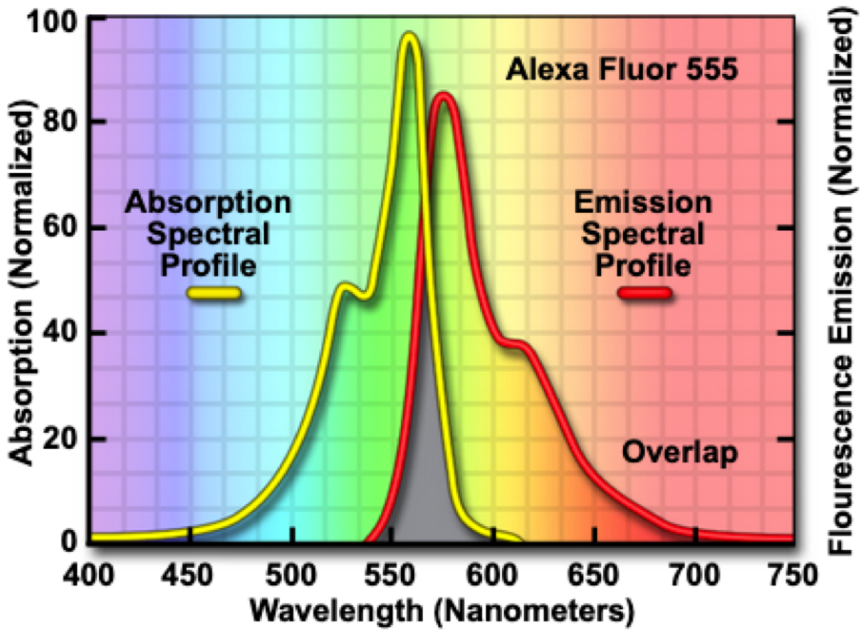
The number of **sequencing cycles**[2] determines the read **length**.

[2](1) Incorporate fluor-dNTP, (2) detect, (3) deblock, (4) cleave fluor

Fluorophores and fluorescence detection

Fluorophores: molecules that re-emit light upon absorption of light

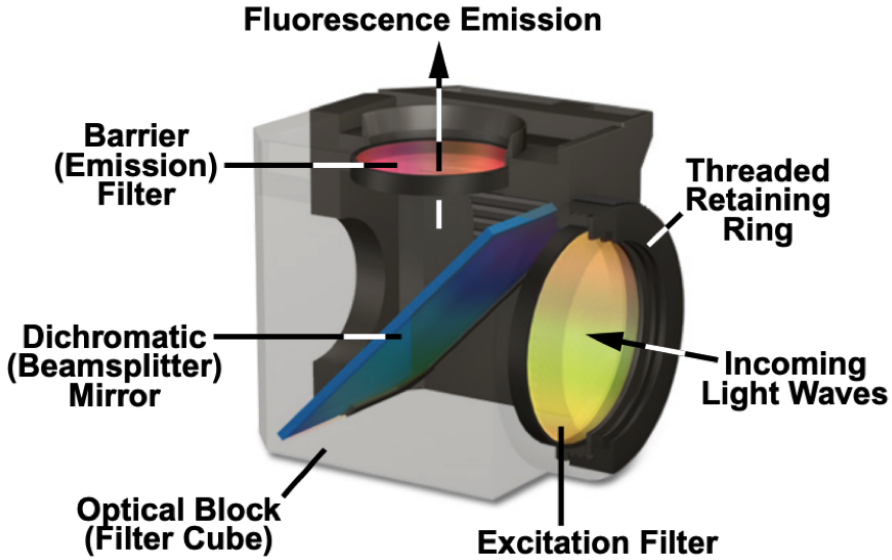
Figure 3 - Fluorophore Absorption and Emission Profiles



https://www.microscopyu.com/techniques/fluorescence/introduction-to-fluorescence-microscopy

Fluorescence microscopes separate emitted light (dim) from excitation light (bright).

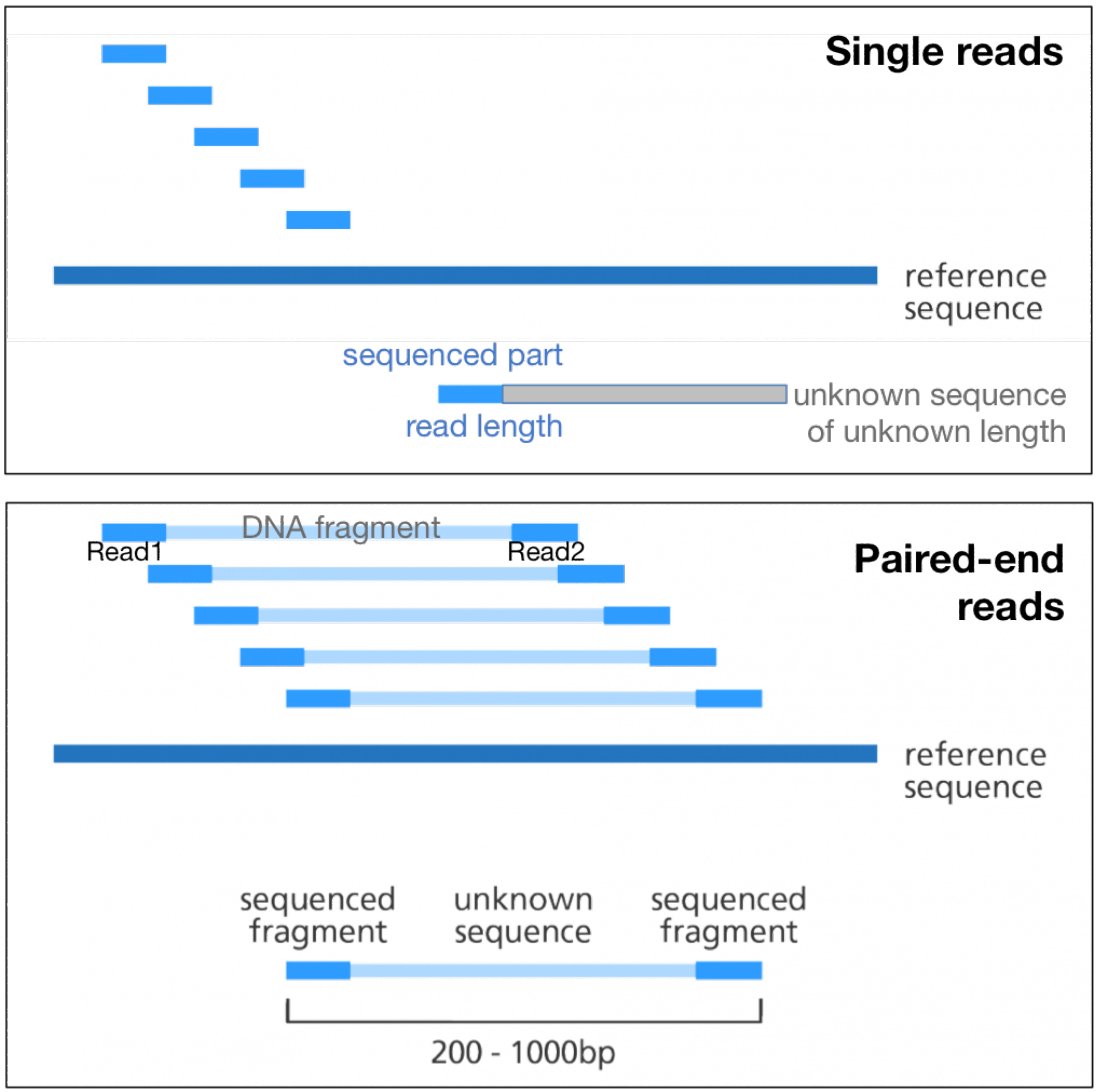
Figure 2 - Fluorescence Filters



See Sanderson et al. [2014] for an overview of fluorescence microscropy techniques (not just DNA-sequencing-related).

Single and paired-end reads

Types of reads

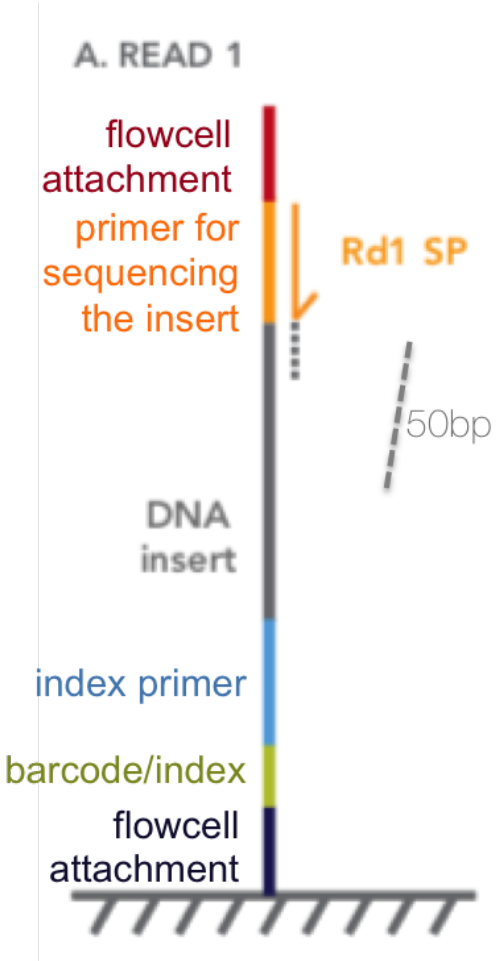


https://www.yourgenome.org/facts/how-do-you-put-a-genome-back-together-after-sequencing

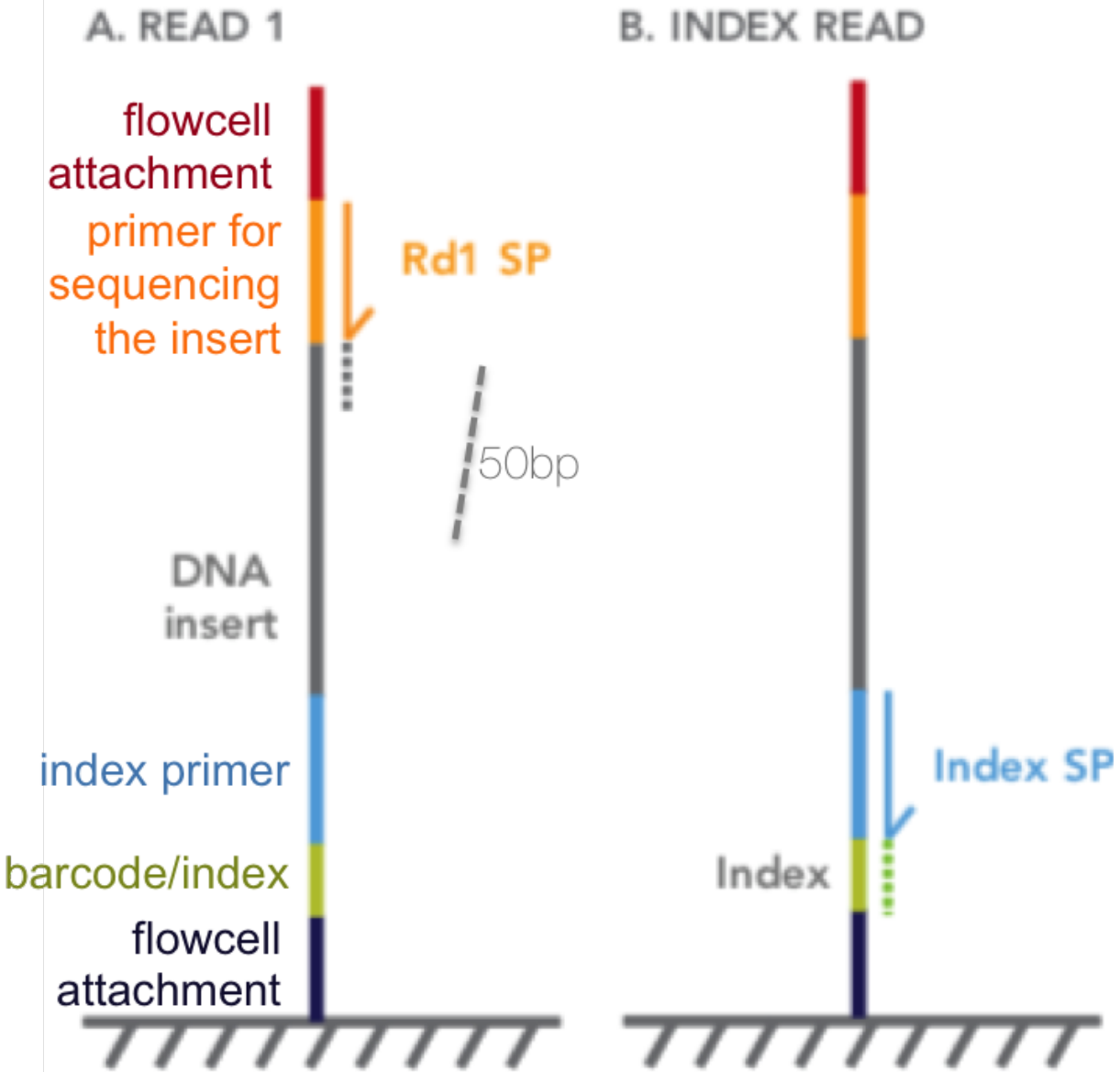
Single reads are cheaper. (why?)
Paired-end (PE) reads are helpful for:

- **alignment** along repetitive regions
- chromosomal **rearrangements** and gene fusion detection
- *de novo* genome and transcriptome **assembly**
- precise information about the size of the original fragment (**insert size**)
- PCR duplicate identification

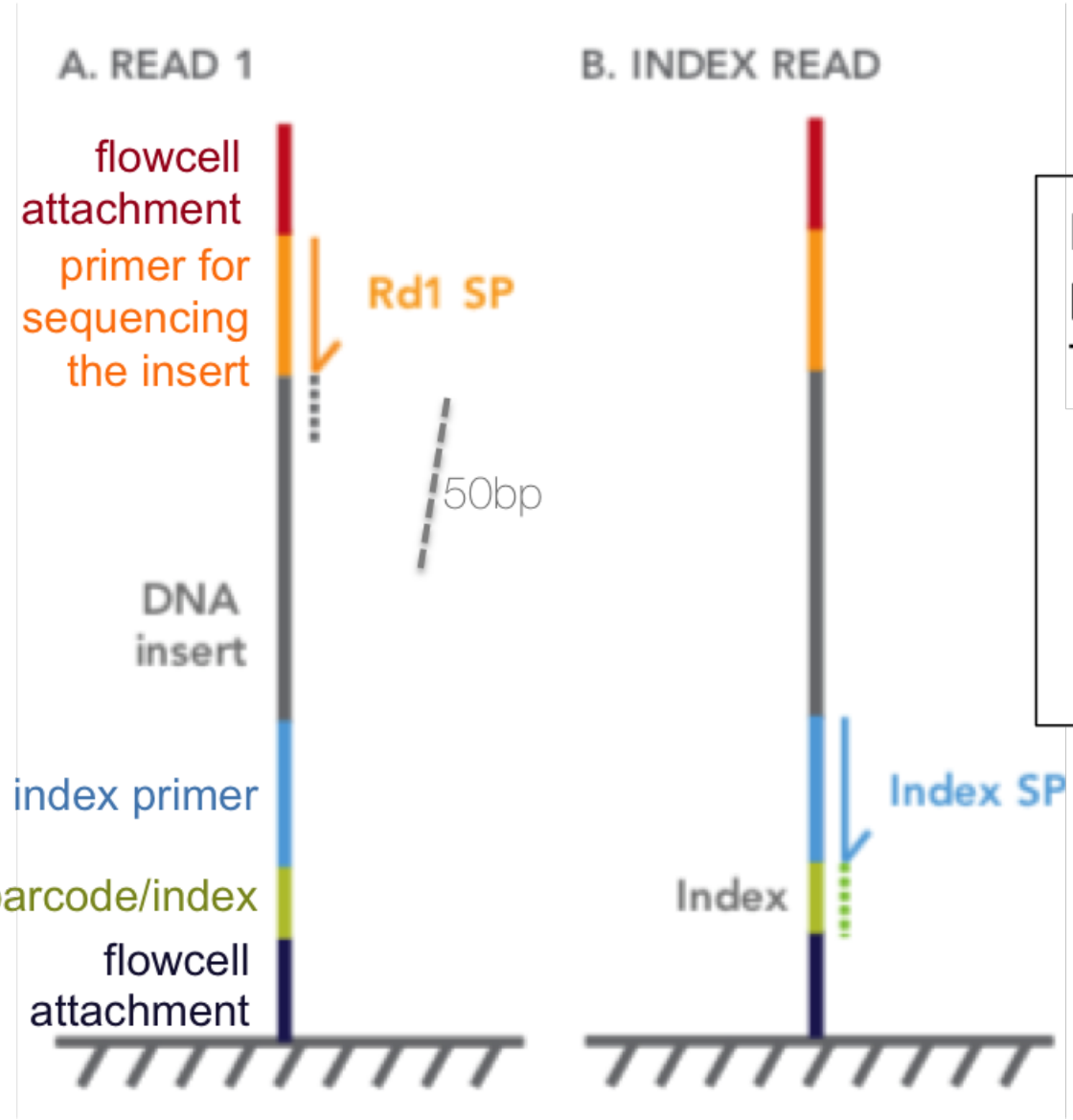
Paired-end read generation

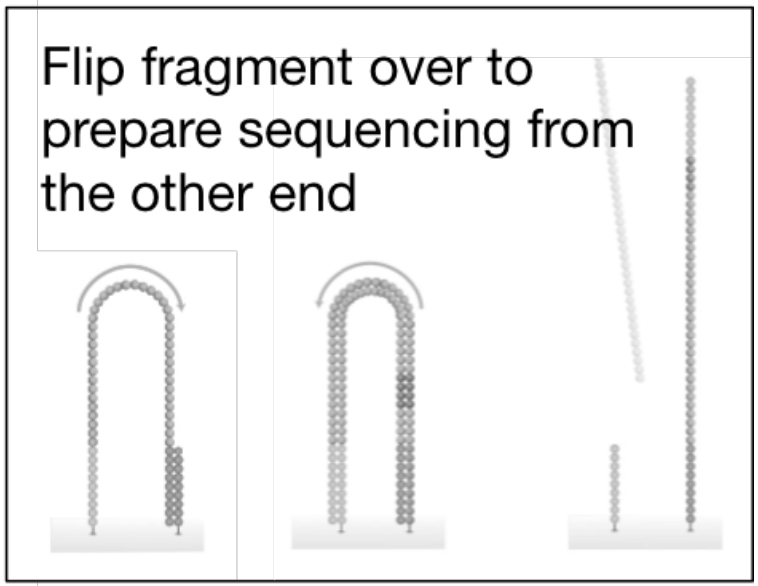


Paired-end read generation

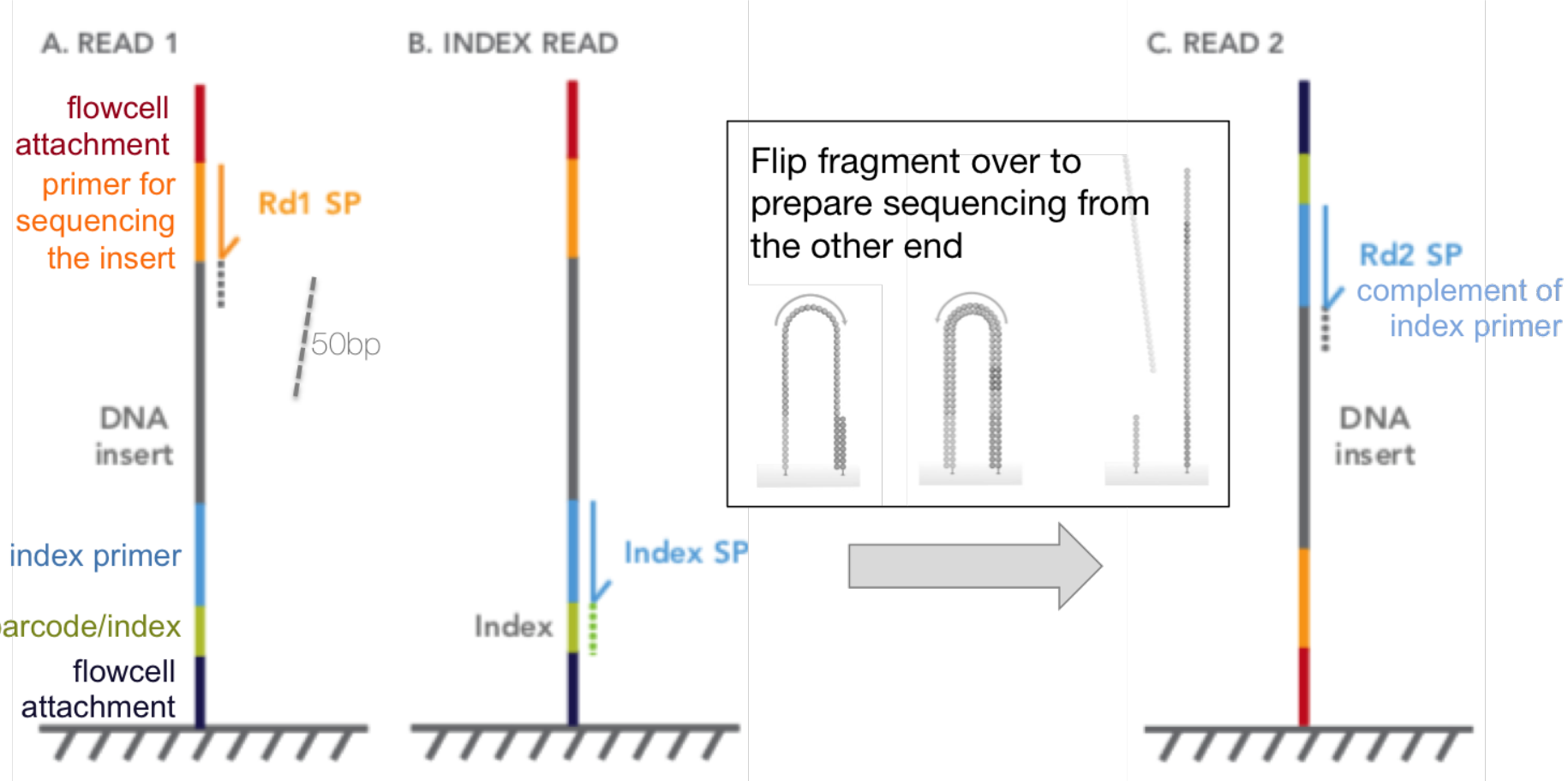


Paired-end read generation





Paired-end read generation



Illumina's "raw reads"

Illumina's read output: turning images into text files

TIFF

BCL file

basecall files (binary text files)

during sequencing, base calls for every location of the flowcell are added live for every **cycle**

FASTQ files

base calls are gathered per **read** rather than per cycle

reads are sorted into different files per sample as identified by the barcodes (demultiplexing)

All steps here are performed by Illumina's proprietory CASAVA software.
The file name usually includes some information about the sample:
`<sample name>_<barcode sequence>_<L(lane)>_<R(read number)>_<set number>.fastq.gz`, e.g. `MyExperiment_AGCTTGTTC_L001_R1_001.fastq.gz`

The FASTQ format: FASTA + quality score

1 read = 4 lines

```
1 @ERR459145.1 DHKW5DQ1:219:D0PT7ACXX:2:1101:1590:2149/1
2 GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC
3 +
4 @7<DBADDDBH?DHHI@DH>HHHEGHIIIGGIFFGIBFAAGAFHA'5?B@D
```

1. @Read ID and sequencing run information
2. sequence
3. + (additional description possible; usually an empty line)
4. quality scores

The read ID line is standardized by Casava 1.8

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

CAUTION

This will only be true if you receive FASTQ files fresh off the sequencer. If you download FASTQ files from public repositories, the read ID might have been changed significantly.

see https://en.wikipedia.org/wiki/FASTQ_format

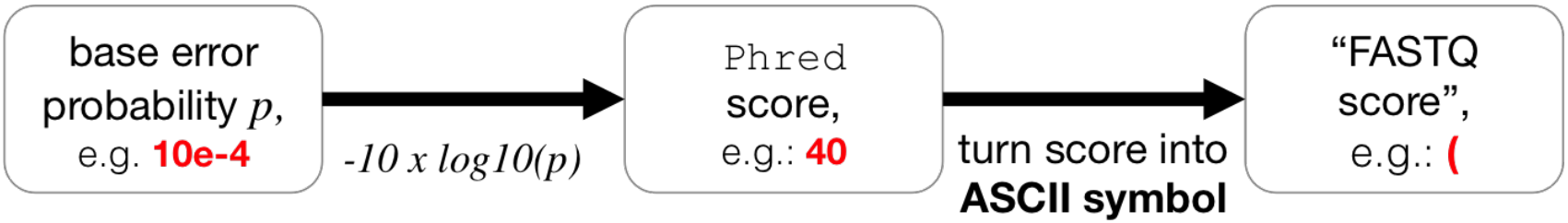
The quality scores: summarizing numerical scores into single-character representations

Illumina's `CASAVA` pipeline:

`BCL files`: Base calls (A/C/T/G) are immediately recorded with an *error probability*[3].

The error probabilites are translated into ASCII symbols in the `FASTQ` files.

```
@ERR459145.1 DHKW5DQ1:219:D0PT7ACXX:2:1101:1590:2149/1
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC
+
@7<DBADDDBH?DHHI@DH>HHHEGHIIIGGIFFGIBFAAGAFHA'5?B@D
```



[3] See the QC section for reasons for base call uncertainties.

ASCII symbols

DEC	OCT	HEX	BIN	Symbol
32	040	20	00100000	
33	041	21	00100001	!
34	042	22	00100010	"
35	043	23	00100011	#
36	044	24	00100100	$
37	045	25	00100101	%
38	046	26	00100110	&
39	047	27	00100111	'
40	050	28	00101000	(
41	051	29	00101001	)
42	052	2A	00101010	*
43	053	2B	00101011	+
65	101	41	01000001	A
66	102	42	01000010	B
67	103	43	01000011	C
68	104	44	01000100	D
69	105	45	01000101	E
70	106	46	01000110	F
71	107	47	01000111	G
72	110	48	01001000	H

`www.ascii-code.com`

ASCII encodes 128 specified characters into seven-bit integers, which is useful for digital communication.

The first 33 characters represent unprintable control codes (e.g. "Start of Text"), therefore the Phred scores were originally encoded by using an **offset of +33** (*Rightarrow* "!").

Printable ASCII symbols start at 33

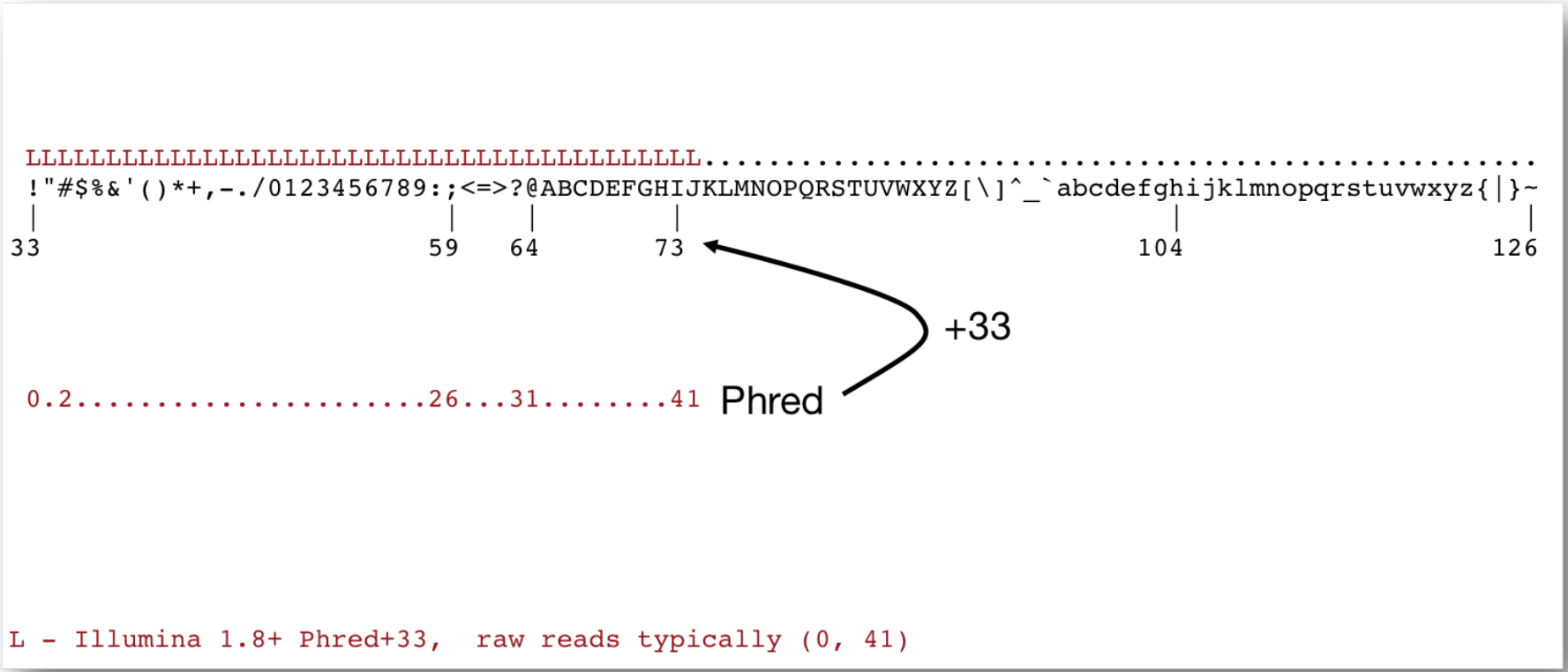


image from https://en.wikipedia.org/wiki/FASTQ_format

Different offsets have been used by different Casava versions

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                         59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

image from https://en.wikipedia.org/wiki/FASTQ_format

Different offsets have been used by different Casava versions

Both the **range of the base call score** as well as its translation via the ASCII code (**offset**) are somewhat arbitrary and have undergone numerous changes.

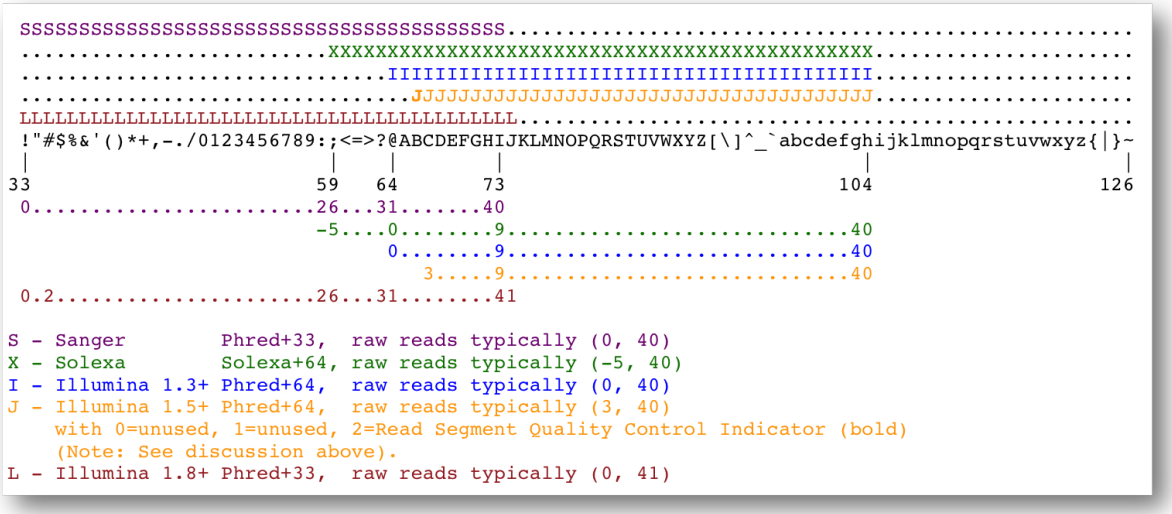


image from https://en.wikipedia.org/wiki/FASTQ_format

Today's standard:

- min. score: 0
- max. score: 41
- ASCII offset: 33

Make sure you know which version you're dealing with.

Quality control of sequencing reads

Two basic QC questions

1. Did our **library prep** generate a **faithful representation** of the DNA/RNA molecules our our samples?
   - ideally, the entire universe of nucleotides was captured (diverse library)
   - no contaminations
   - no degradation
   - no bias towards fragments of certain GC contents and/or sizes
2. How successful was the actual **sequencing**?
   - consistently high base call confidence
   - uniform nucleotide frequencies

Biases

QC should help identify **systematic distortions** of data and their possible sources.

FastQC

http://www.bioinformatics.babraham.ac.uk/projects/fastqc

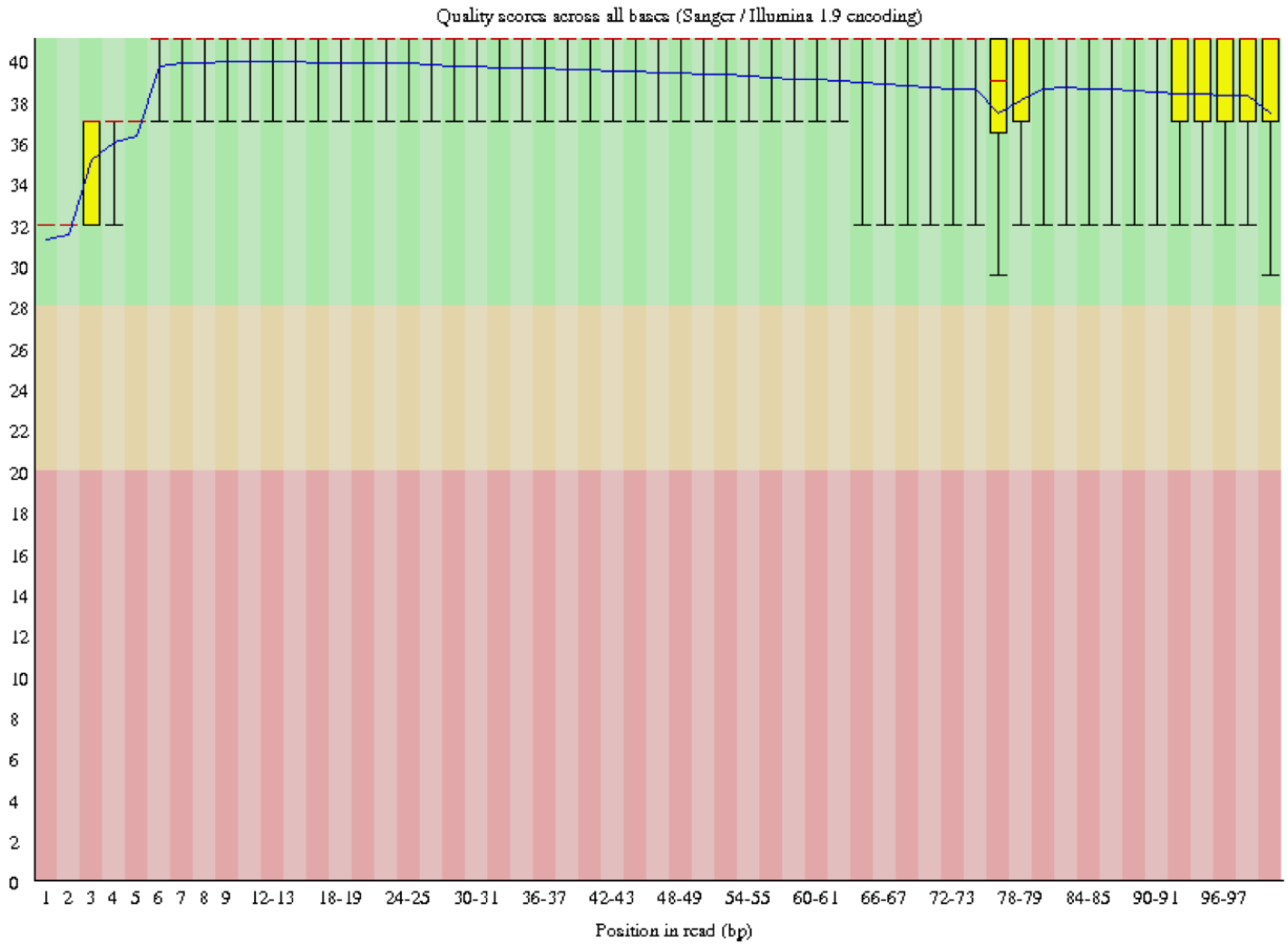
- unpublished, but most widely used QC tool
- supports all NGS technologies
- continuously developed and maintained by long-time bioinformatics experts
- will only use the first 200K reads for the diagnosis!

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later.
Initial Contact	Simon Andrews

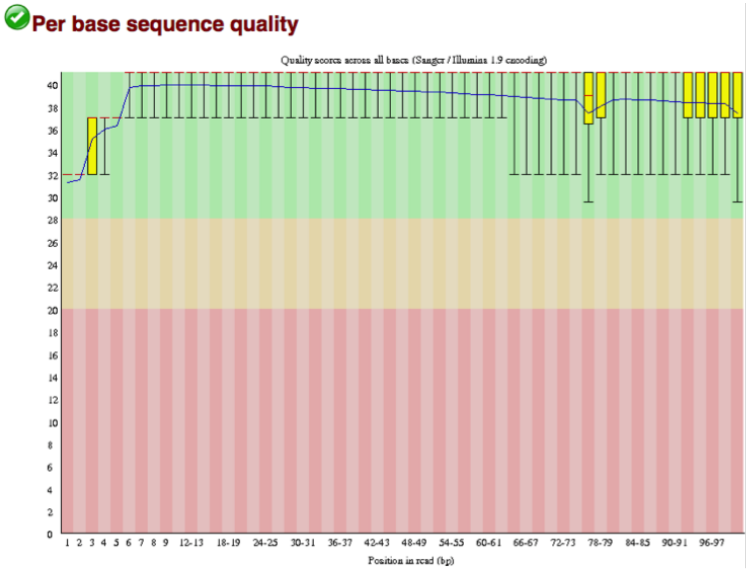
Sequencing quality

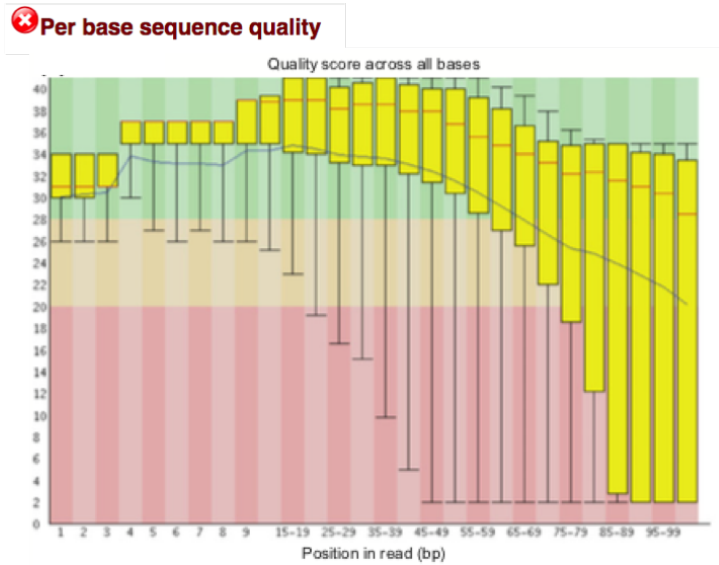
Based on ASCII-endoced Phred scores within the `fastq` file.

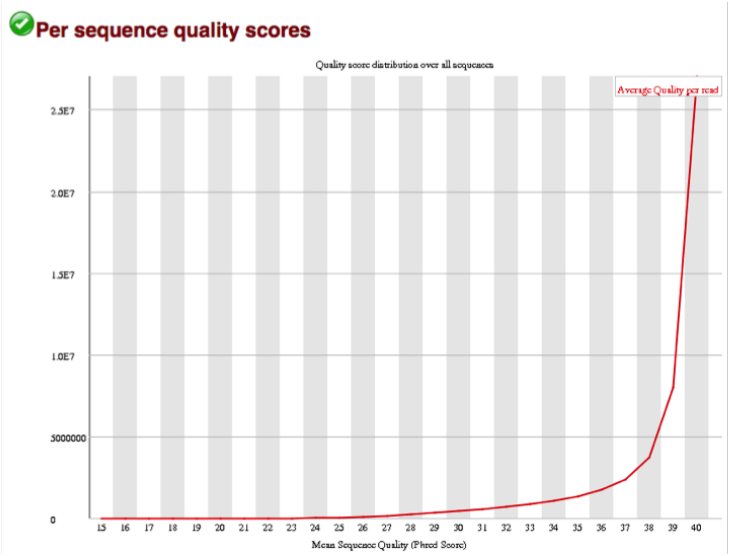
Per base sequence quality

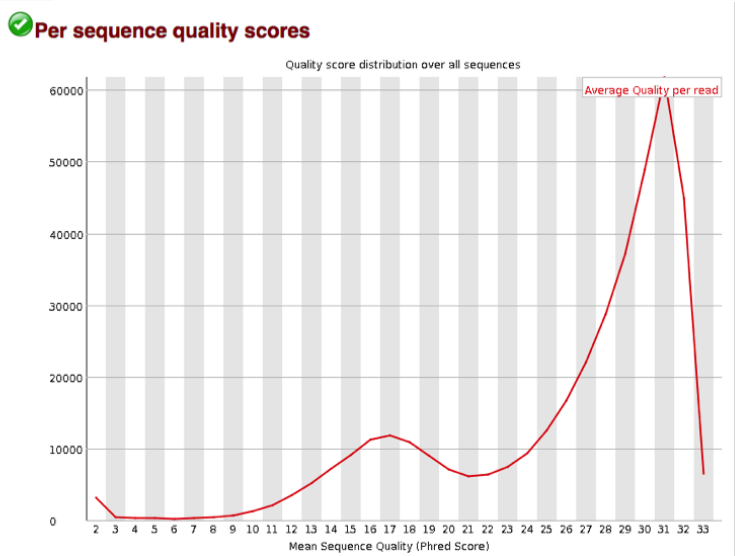


Sequencing quality







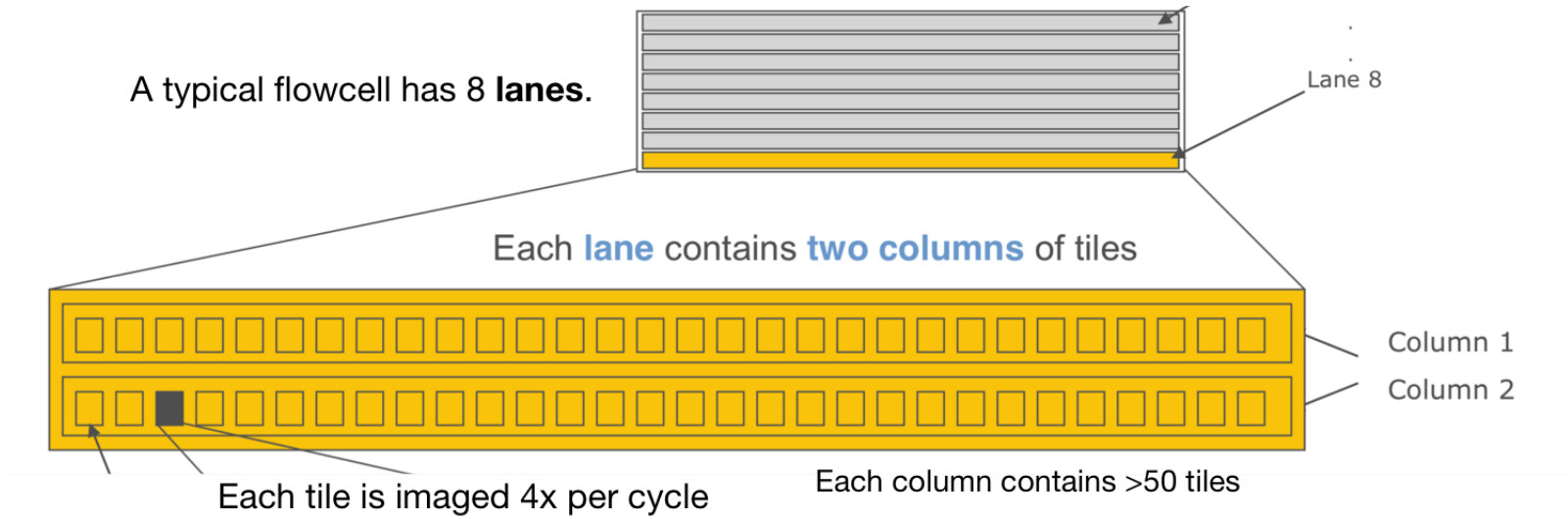


Sequencing quality: reasons for sequencing noise

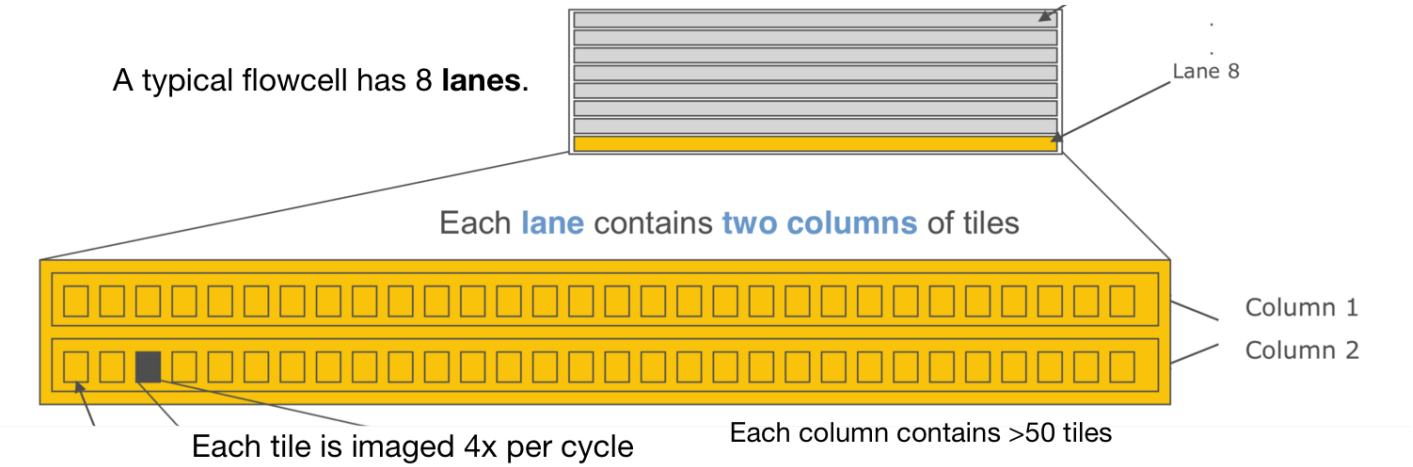
Noise = fluorophore intensity signal is not as strong and clear as expected.

- laser not well **calibrated**
- **interfering signals** from neighbouring clusters or bases with similar emission spectra
- **unsynchronized fragments** in each cluster:
  - *phasing*: small fraction of fragments in each cluster fails to incorporate any base
  - *prephasing*: more than one base is incorporated
- **decaying** chemicals (runs often last several days to a week!)
- **extraneous objects** on the flow cell (e.g. dust, air bubbles)

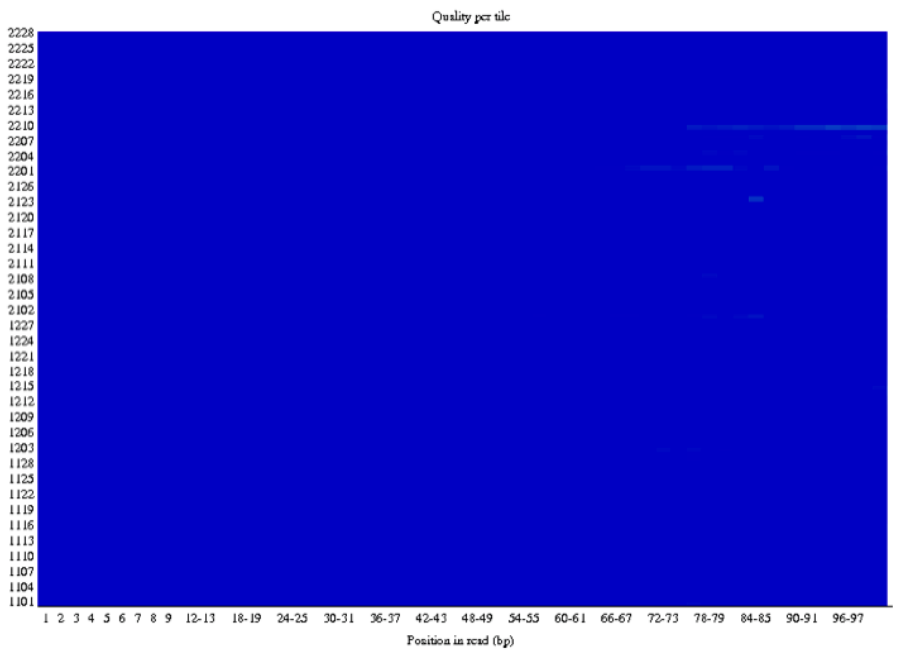
Physically localized error rates: tiles vs. time



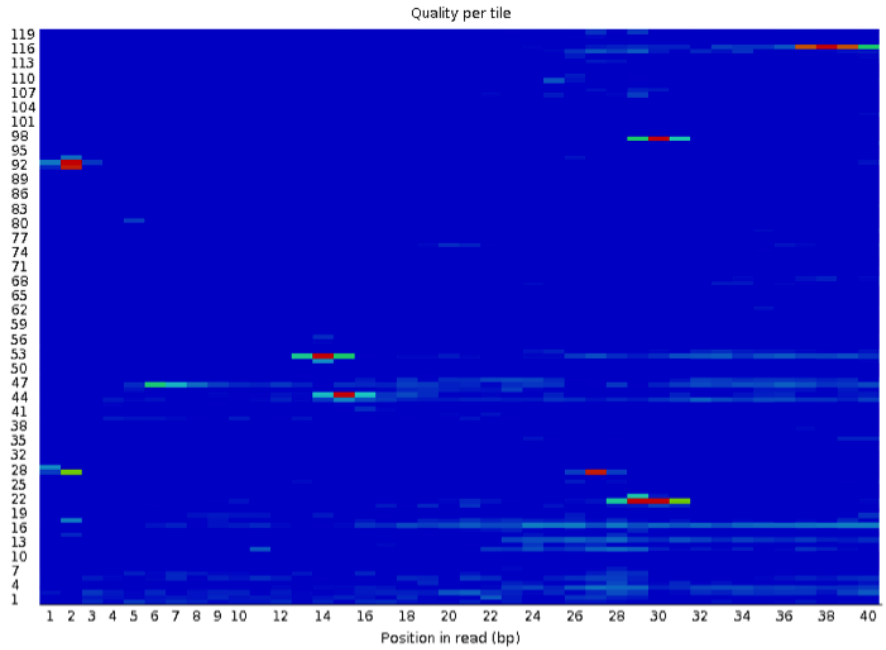
Physically localized error rates: tiles vs. time



Per tile sequence quality



Per tile sequence quality



Contaminations: threats to the full representation of our original fragment pool (and waste of seq. reads)

- Sources:
  - **primer** contamination
  - **adapter** contamination
    - sequence read length larger than the fragment size (3' contamination)
    - adapter dimers without insert
  - DNA from other species/libraries
- Consequences:
  - noise
  - reduced alignment rates

Can be identified by examining **sequence composition** and **overrepresented sequences/k-mers**.

Contaminations: threats to the full representation of our original fragment pool (and waste of seq. reads)

- Sources:
  - **primer** contamination
  - **adapter** contamination
    - sequence read length larger than the fragment size (3' contamination)
    - adapter dimers without insert
  - DNA from other species/libraries
- Consequences:
  - noise
  - reduced alignment rates

Can be identified by examining **sequence composition** and **overrepresented sequences/k-mers**.

Contaminations: threats to the full representation of our original fragment pool (and waste of seq. reads)

- Sources:
  - **primer** contamination
  - **adapter** contamination
    - sequence read length larger than the fragment size (3' contamination)
    - adapter dimers without insert
  - DNA from other species/libraries
- Consequences:
  - noise
  - reduced alignment rates

Can be identified by examining **sequence composition** and **overrepresented sequences/k-mers**.

Contaminations: threats to the full representation of our original fragment pool (and waste of seq. reads)

- Sources:
  - **primer** contamination
  - **adapter** contamination
    - sequence read length larger than the fragment size (3' contamination)
    - adapter dimers without insert
  - DNA from other species/libraries
- Consequences:
  - noise
  - reduced alignment rates

Can be identified by examining **sequence composition** and **overrepresented sequences/k-mers**.

Contaminations: threats to the full representation of our original fragment pool (and waste of seq. reads)

- Sources:
  - **primer** contamination
  - **adapter** contamination
    - sequence read length larger than the fragment size (3' contamination)
    - adapter dimers without insert
  - DNA from other species/libraries
- Consequences:
  - noise
  - reduced alignment rates

Can be identified by examining **sequence composition** and **overrepresented sequences/k-mers**.

Contaminations: threats to the full representation of our original fragment pool (and waste of seq. reads)

- Sources:
  - **primer** contamination
  - **adapter** contamination
    - sequence read length larger than the fragment size (3' contamination)
    - adapter dimers without insert
  - DNA from other species/libraries
- Consequences:
  - noise
  - reduced alignment rates

Can be identified by examining **sequence composition** and **overrepresented sequences/k-mers**.

Contaminations: threats to the full representation of our original fragment pool (and waste of seq. reads)

- Sources:
  - **primer** contamination
  - **adapter** contamination
    - sequence read length larger than the fragment size (3' contamination)
    - adapter dimers without insert
  - DNA from other species/libraries
- Consequences:
  - noise
  - reduced alignment rates

Can be identified by examining **sequence composition** and **overrepresented sequences/k-mers**.

Contaminations: threats to the full representation of our original fragment pool (and waste of seq. reads)

- Sources:
  - **primer** contamination
  - **adapter** contamination
    - sequence read length larger than the fragment size (3' contamination)
    - adapter dimers without insert
  - DNA from other species/libraries
- Consequences:
  - noise
  - reduced alignment rates

Can be identified by examining **sequence composition** and **overrepresented sequences/k-mers**.

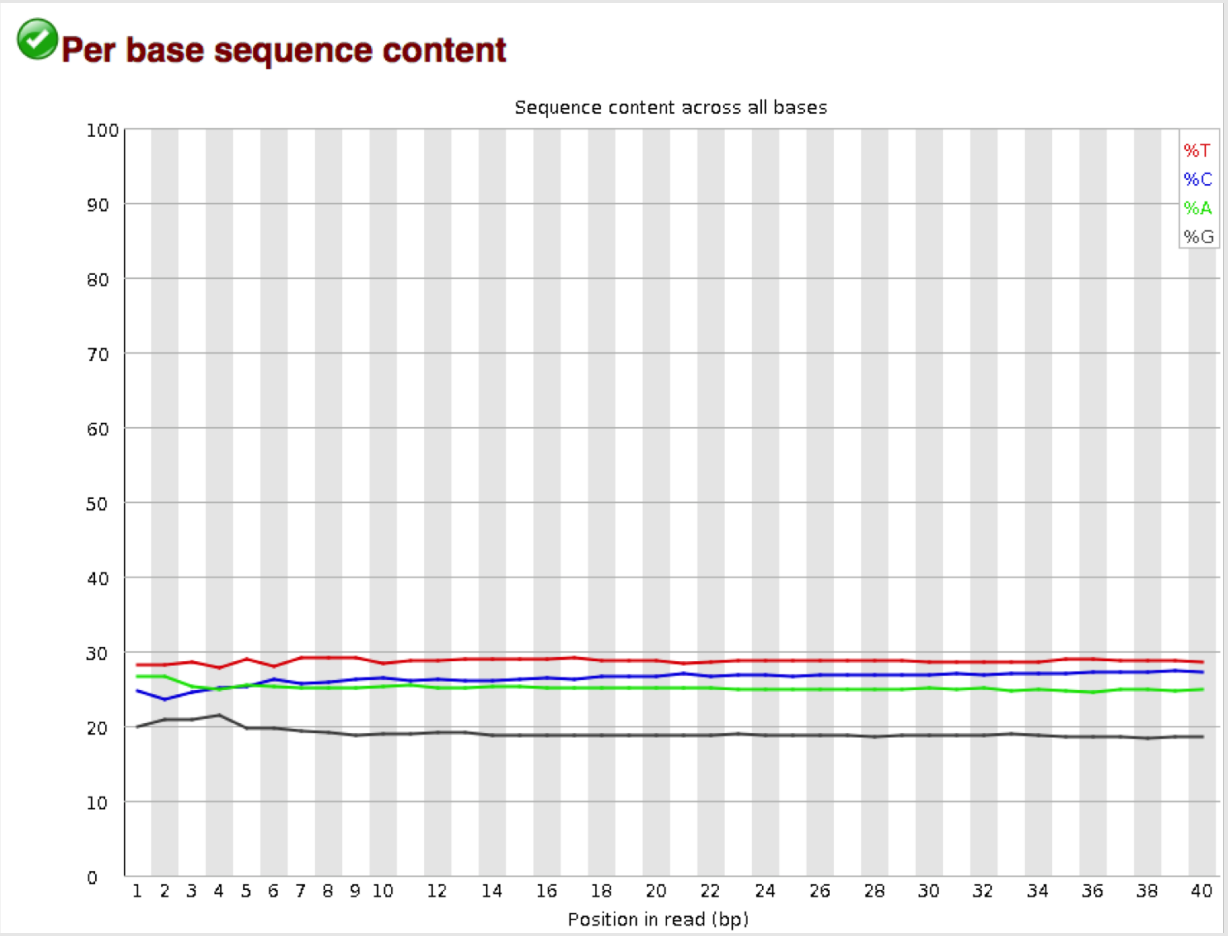
Contaminations: threats to the full representation of our original fragment pool (and waste of seq. reads)

- Sources:
  - **primer** contamination
  - **adapter** contamination
    - sequence read length larger than the fragment size (3' contamination)
    - adapter dimers without insert
  - DNA from other species/libraries
- Consequences:
  - noise
  - reduced alignment rates

Can be identified by examining **sequence composition** and **overrepresented sequences/k-mers**.

Detecting contaminations

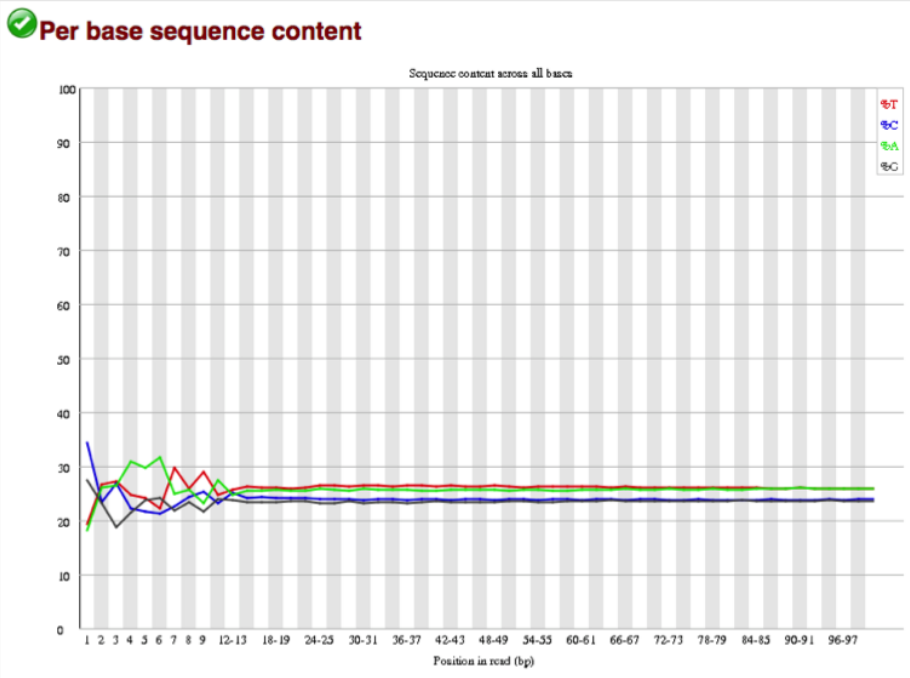
Per Base Sequence Content

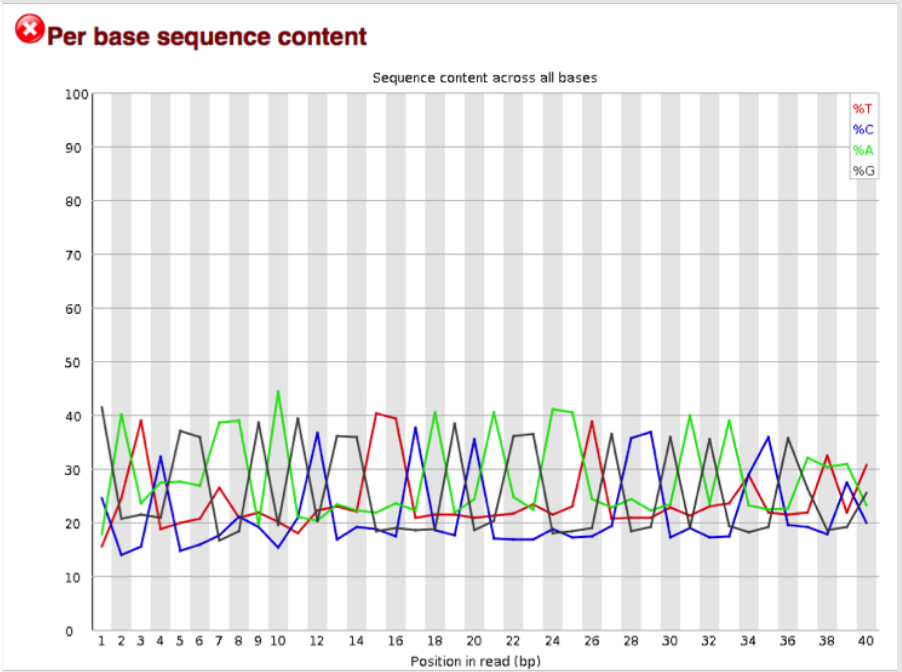


If the fragments represent a random and diverse representation of the entire genome, there should be a **uniform distribution** of all four bases across all cycles.

Detecting contaminations

Per Base Sequence Content – more examples





- irregularities in the first ca. 8 bp are often seen for RNA-seq and ATAC-seq and indicate a bias for certain sequences at the fragment beginning
- more severe deviations from uniformity often indicate contaminations and/or lack of library diversity

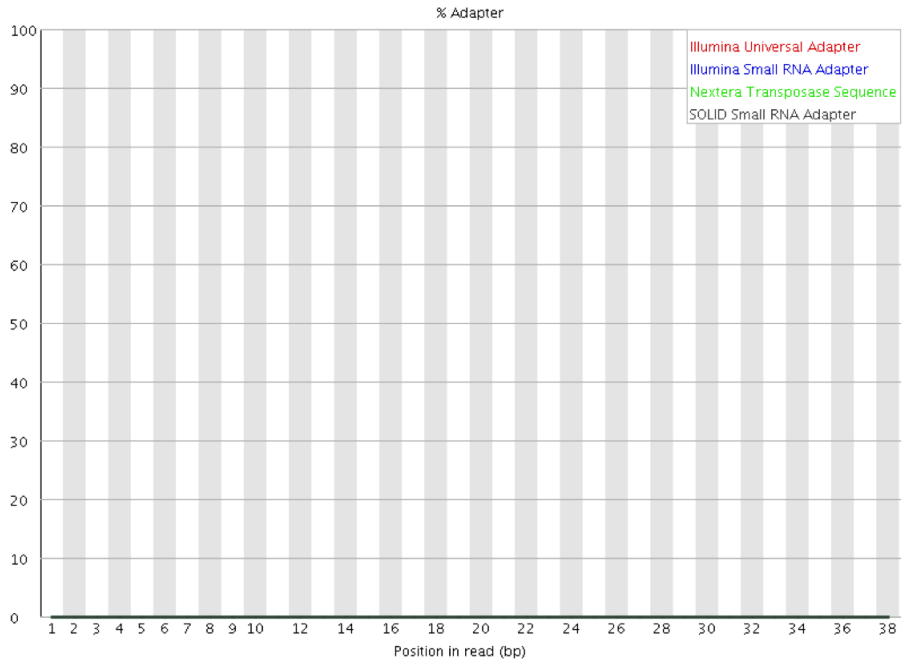
Detecting contaminations

Overrepresented sequences & adapter sequence frequencies

Overrepresented sequences

No overrepresented sequences

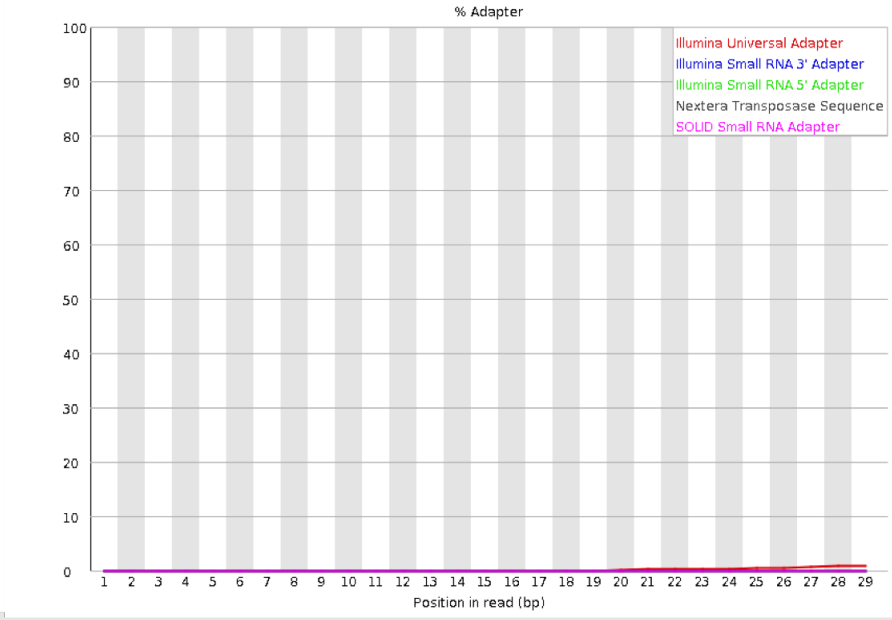
Adapter Content



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.4670012750197325	No Hit

Adapter Content



Trimming contaminations & low-quality bases

Mostly done to improve alignment.

- Can be done before alignment or, if contaminations/low-quality bases are low in number, might be left to the "soft-clipping" function[4] of read aligners.
- There are numerous tools out there to do the job, e.g. `Cutadapt` [Martin, 2011] and `TrimGalore`.
- For *de novo* assemblies, it is probably more meaningful to perform some error-correction based on overlapping reads rather than trimming the reads [Salzberg et al., 2012, Yang et al., 2013]

[4] ignoring mis-matched bases at the beginning/end of a read

Duplicate read: types

- **optical duplicates** (same DNA cluster erraneously reported as separate clusters)
- **natural duplicates** (multiple independent original fragments with very similar sequence)
  - more likely to occur for small(ish) genomes/transcriptomes and experiments that enrich for relatively few and small regions of the genome
- **PCR duplicates** (1 original fragment)
  - often sample-specific and very difficult to correct *in silico*
  - can be reduced by avoiding excessive PCR

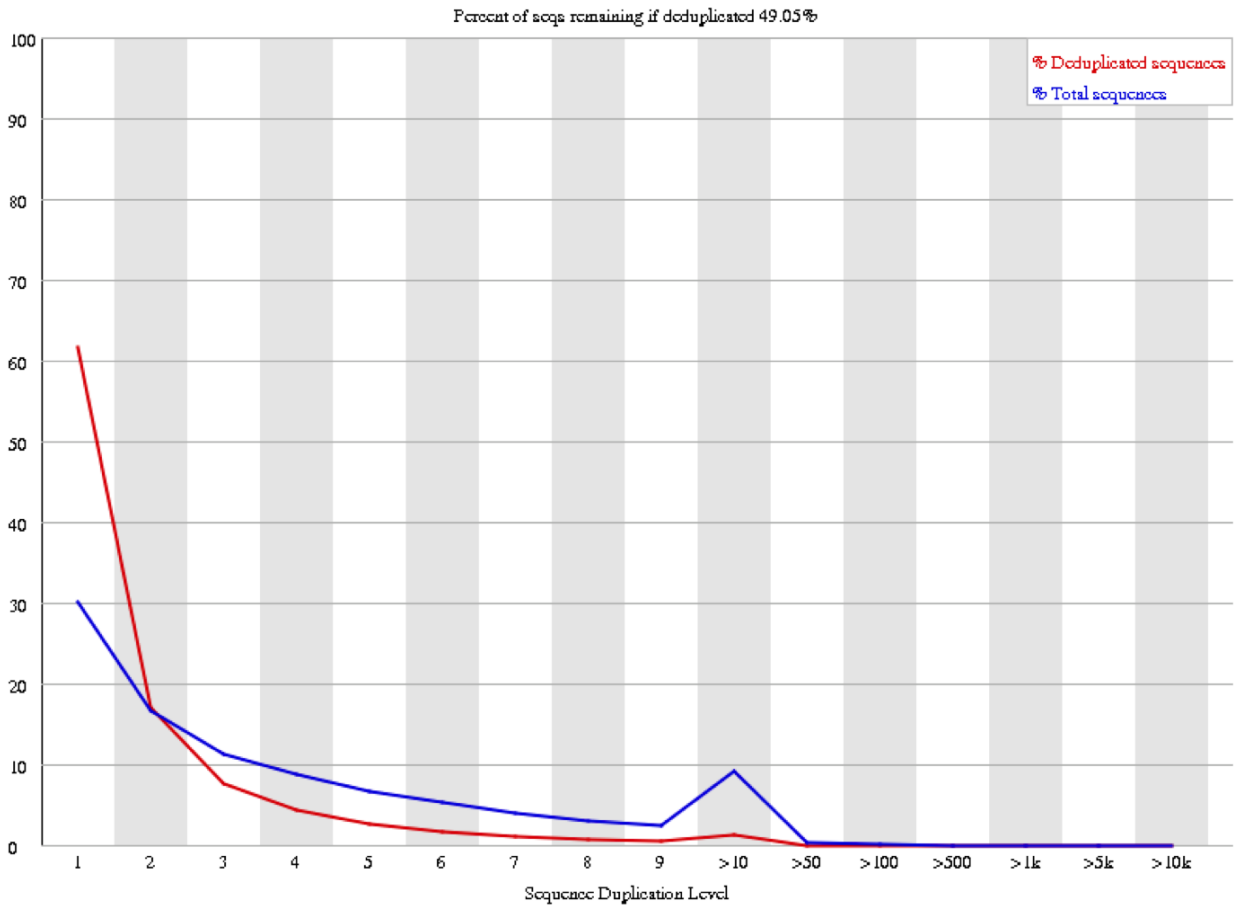
The Problem

There is no way to distinguish natural from PCR duplicates!

Duplicate reads: FastQC assessment

Proportion of reads (y-axis) that contain sequences in each of the different duplication level bins (x-axis).

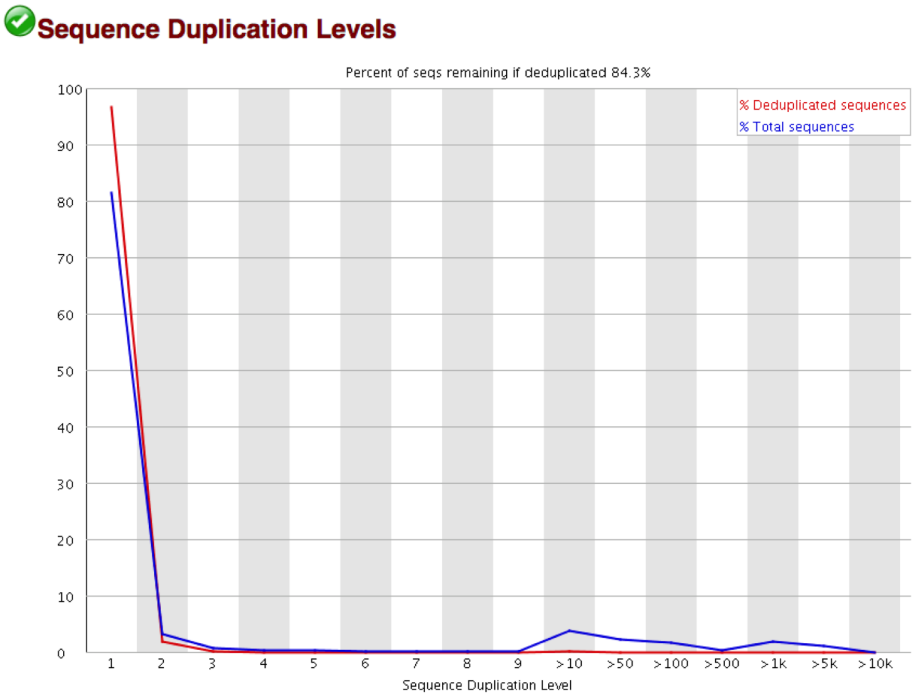
Sequence Duplication Levels

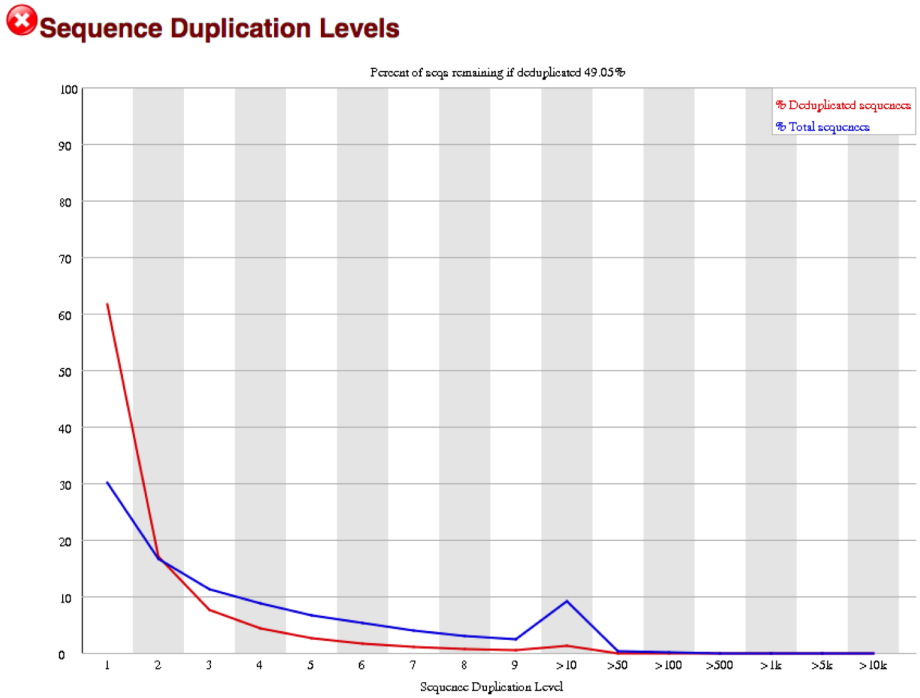


Blue line: all reads (= first 100K!) – how many times are individual sequences found?

Red line: sequences after de-deduplication – how many *different* sequences were found to be duplicated?

Duplicate reads: FastQC assessment





Check that the red line is flat and that the number of remaining reads after de-duplication is acceptable.

Two basic QC questions

1. Did our **library prep** generate a **faithful representation** of the DNA/RNA molecules our our samples?
   - ideally, the entire universe of nucleotides was captured (diverse library)
   - no contaminations
   - no bias towards fragments of certain GC contents and/or sizes
   - no degradation
2. How successful was the actual **sequencing**?
   - consistently high base call confidence
   - uniform nucleotide frequencies

QC summary

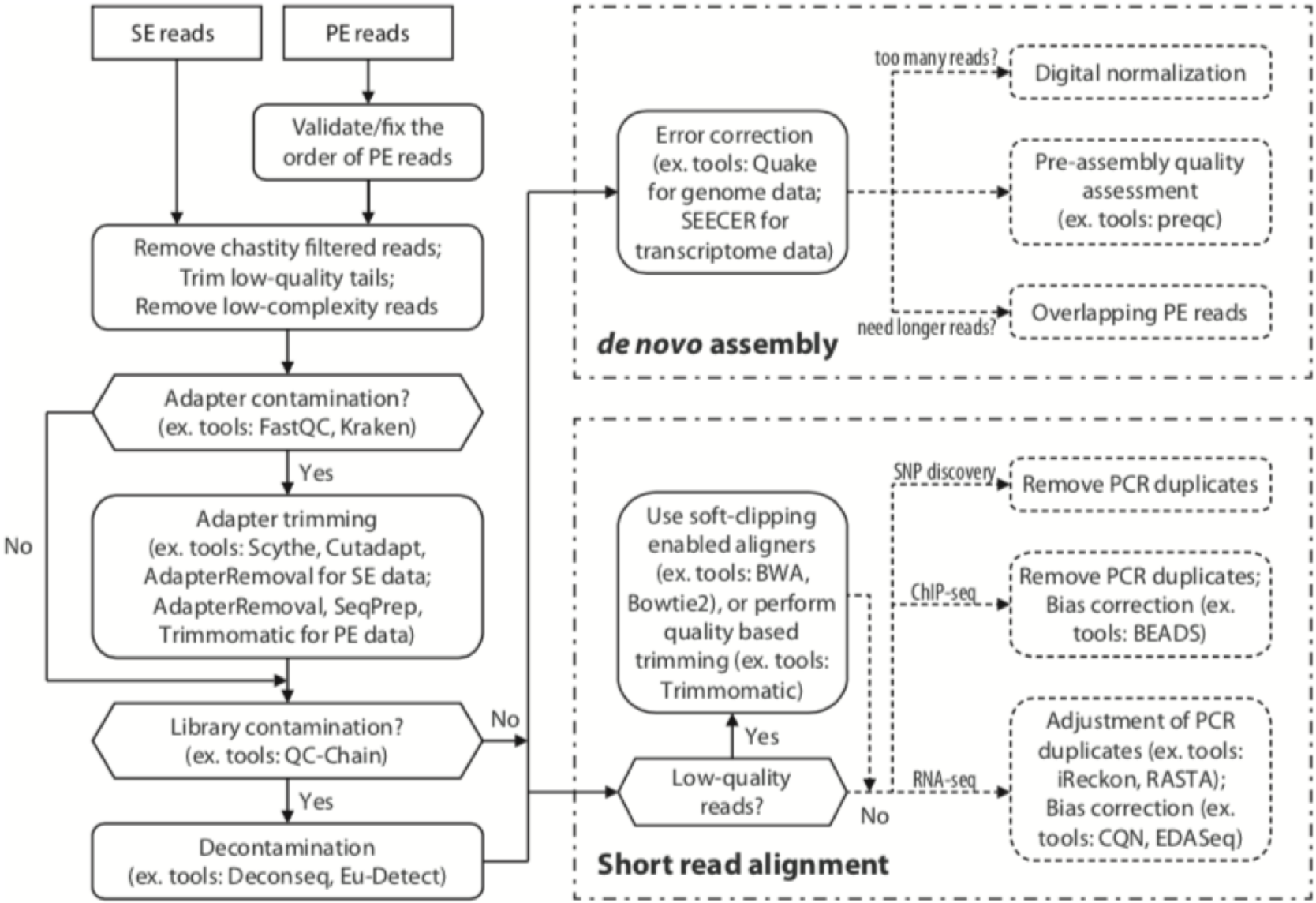
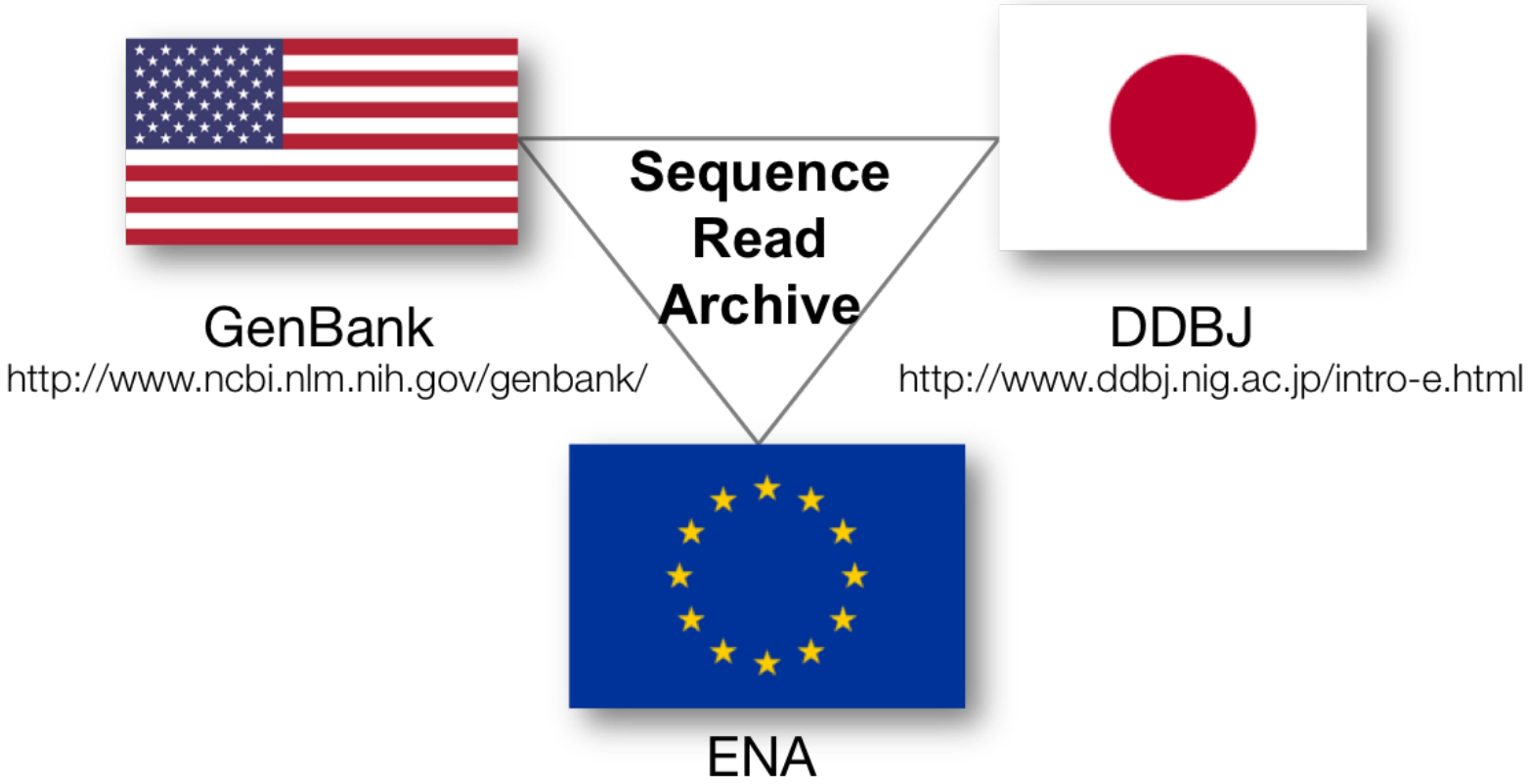


Figure from Zhou and Rokas [2014] (highly recommended reading!)

Sequence Read Archive

Where are all the reads?

SRA = main repository for publicly available DNA and RNA sequencing data of which three instances are maintained world-wide.
GEO (https://www.ncbi.nlm.nih.gov/geo/) can be used to find SRA data, too.



See O'Sullivan et al. [2017] for many more details.

References

Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 2011. doi: 10.14806/ej.17.1.200.

Christopher O'Sullivan, Benjamin Busby, and Ilene Karsch Mizrachi. In Jonathan M. Keith, editor, *Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution*, chapter Managing Sequence Data. Humana Press, 2017. doi: 10.1007/978-1-4939-6622-6_4.

Steven L. Salzberg, Adam M. Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J. Treangen, Michael C. Schatz, Arthur L. Delcher, Michael Roberts, Guillaume Marcxais, Mihai Pop, and James A. Yorke. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, 2012. doi: 10.1101/gr.131383.111.

Michael J. Sanderson, Ian Smith, Ian Parker, and Martin D. Bootman. Fluorescence microscopy. *Cold Spring Harbor Protocols*, 2014. doi: 10.1101/pdb.top071795.

Xiao Yang, Sriram P. Chockalingam, and Srinivas Aluru. A survey of error-correction methods for next-generation sequencing. *Briefings in Bioinformatics*, 2013. doi: 10.1093/bib/bbs015.

Xiaofan Zhou and Antonis Rokas. Prevention, diagnosis and treatment of high-throughput sequencing data pathologies. *Molecular Ecology*, 23(7): 1679–1700, 2014. doi: 10.1111/mec.12680.