

# UBC Bioinformatics Class

Topic 3: Fastq files and quality  
checking and trimming

## Lecture outcomes

- Understand sequence file formats
- Identify the main steps for preparing NGS data for alignment/assembly

## NGS file formats: Fasta

- Sequences with a header (.fasta, .fa, .fas)
- Now mainly used for storing reference sequences (no qual scores) as either nucleotides or peptides
- Can have quality scores are stored in separate files (usually .fasta or .fa & .qual)
- 2 lines/sequence read:

Always begins with ">"

Sequence identifier (contig name, relevant info, etc.)

Sequence

```
>ctg7180038347536
CTTTGTGATCACATTACTATCATCGTTTTGAGCCTTGGCCGTGTTCTTACCATTACCTCCACCCTTTTAG
CCGATCATAACCTCCACTTAATTCTTTACCTTTTTGAGGAATAGCTGCGATGAGTAATTCTGTTAGCCA
CCTTCTTTTAACTGCCATTCTTGAAAAGTTTCAAACCTCAACTAGAACAGTTGCTACTTGAAAACATCAC
CCATTCTTAAAAAATGAGTCTCTTTTAAAGCTCTTTTTAGAACTCTAAAATATGAAAATATTGCCAAGCTA
CTGGCCTTTCCAGCTTGTTAA
>ctg7180038347539
TAAACGAAAGGCTCTTAAACCCCTAAAAGTGTTGCTTCATACCCTAGAGGATCAAGGTCAAATAACTACA
TCATTTCTAGAAAGTTCTCCCTAAAAAACTGCTCAGAACTGGTCAAAATTGGACCATACAGATTGCTCCA
```

# NGS file formats: Fastq

## FASTQ:

- Sequence and quality scores are stored in the same file (usually .fq or .fastq)
- Most common format for short read data returned from the sequencer
- 4 lines/sequence read:

Always begins with “@”

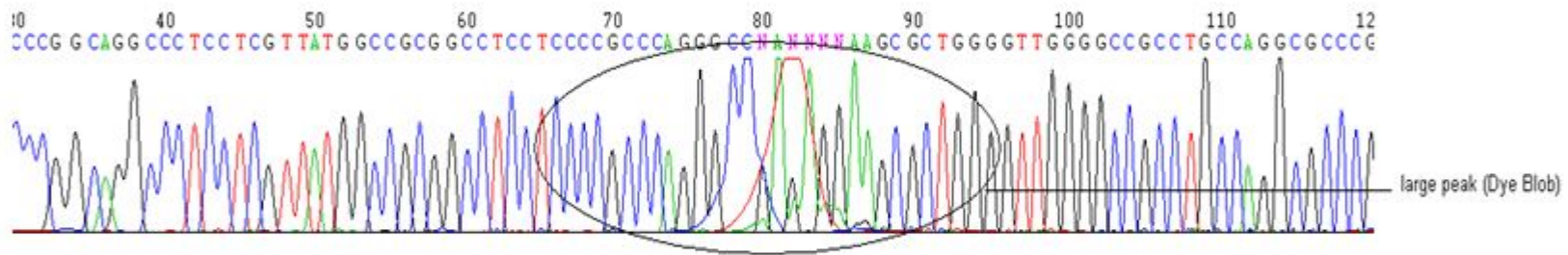
Sequence identifier (sequencer, lane, location info, etc.)

```
@HWI-ST521:81:C0HKCACXX:5:1101:1124:1158 1:N:0:GTCCGC  
GTGACTATTTTGTCAAAGCTATGGGTGAAGATTTTCAAGACGCTGGAAATGTATTCAAAG  
+  
CB@DFFFFHHHHFIIJIIJIEHIIJ<CGHGBHIIJIIJJJFGGHGHGHHHHIHHIGHJGIH
```

Sequence

Quality scores

# NGS file formats: Quality scores

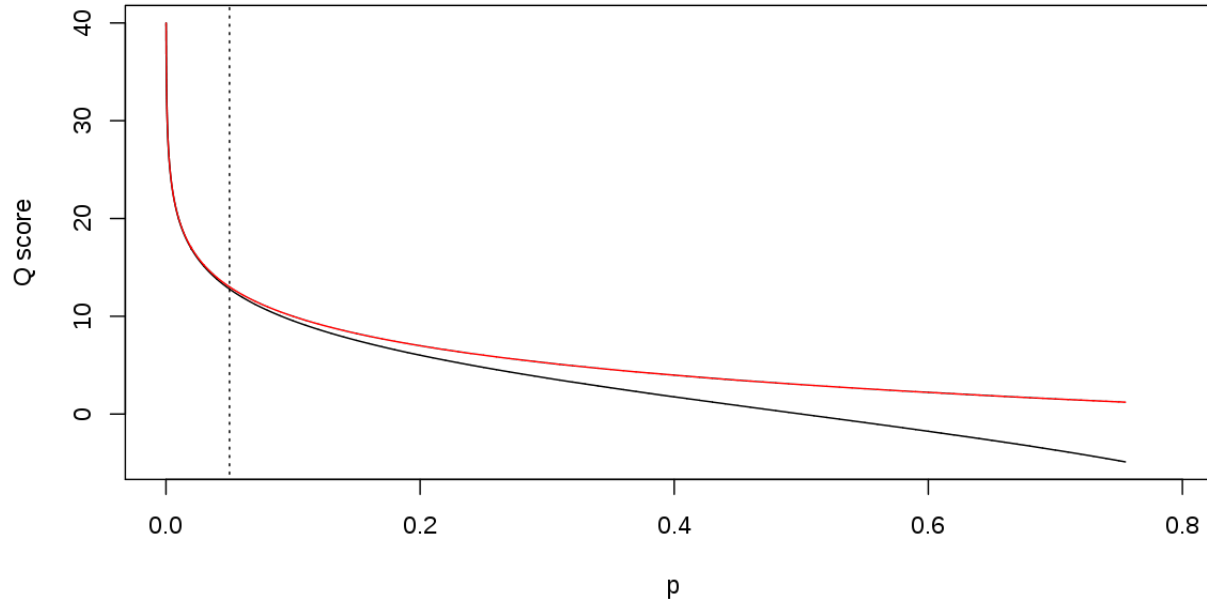


# NGS file formats: Quality scores

Historically, two formats (now all are Sanger)

- $Q_{\text{sanger}} = -10 * \log_{10}(p)$
- $Q_{\text{solexa}} = -10 * \log_{10}(p / (1 - p))$

where  $p$  is the probability that a base call is incorrect



High quality scores are good

To calculate  $p$  from  $Q$ :

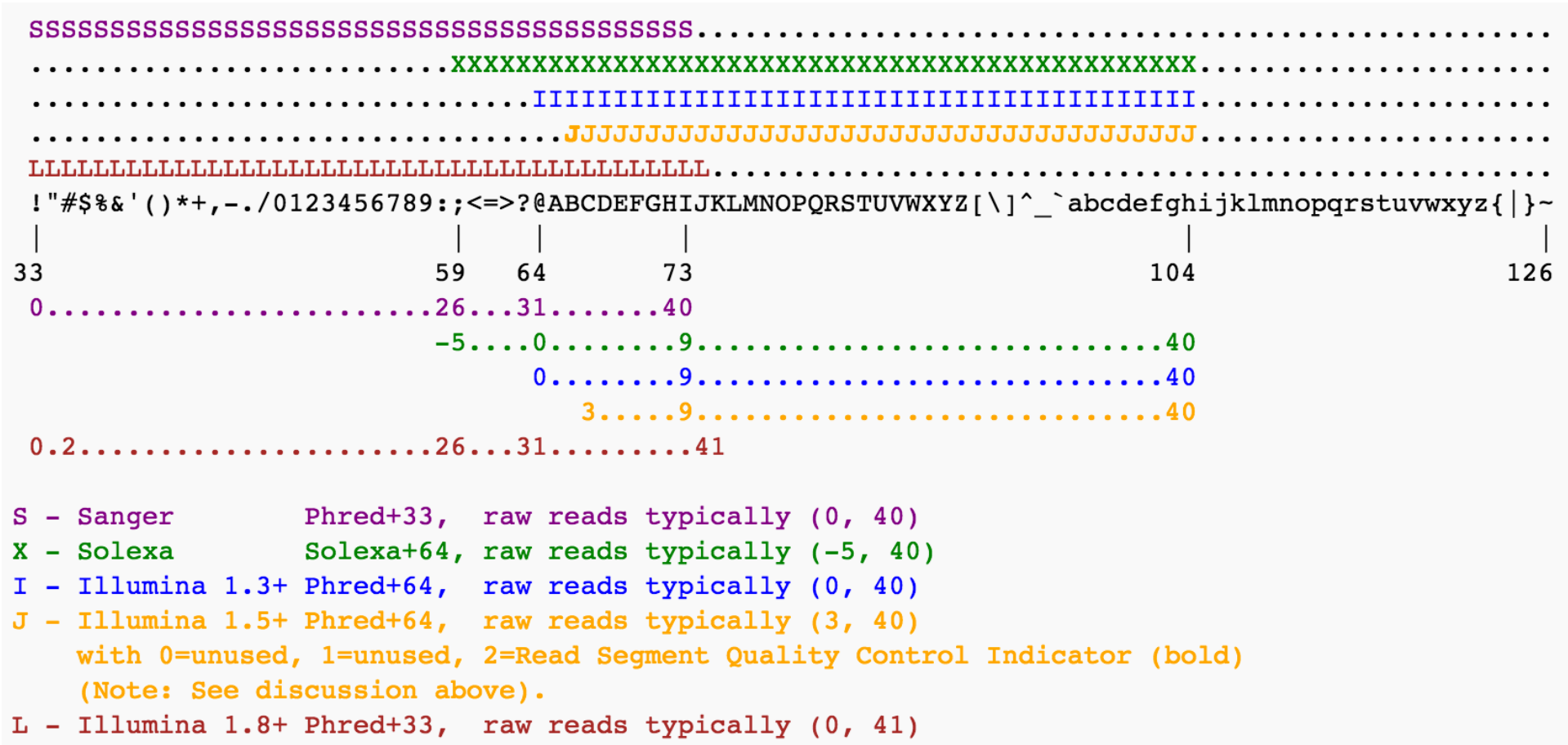
$$p = 10^{(-Q / 10)}$$

Q30 = 0.1%  $p$ [incorrect]

Q20 = 1%  $p$ [incorrect]

Q10 = 10%  $p$ [incorrect]

## NGS file formats: Quality scores



Fortunately, we seem to have settled on a standard in the community...for now!

[http://en.wikipedia.org/wiki/FASTQ\\_format](http://en.wikipedia.org/wiki/FASTQ_format)

# Code break

Move into directory ~/data/Topic3

1) How many sequences so you have in the file Pine\_reference\_rnaseq\_reduced.fa?

2) How many sequences do you have in the fastq file  
GBS12\_brds\_Pi\_197A2\_100k\_R1.fastq?

3) How many sequences contain a base with a Phred score of 2  
GBS12\_brds\_Pi\_197A2\_100k\_R1.fastq?

Note: there are more unix examples at the end of README\_quality\_trimming.txt file



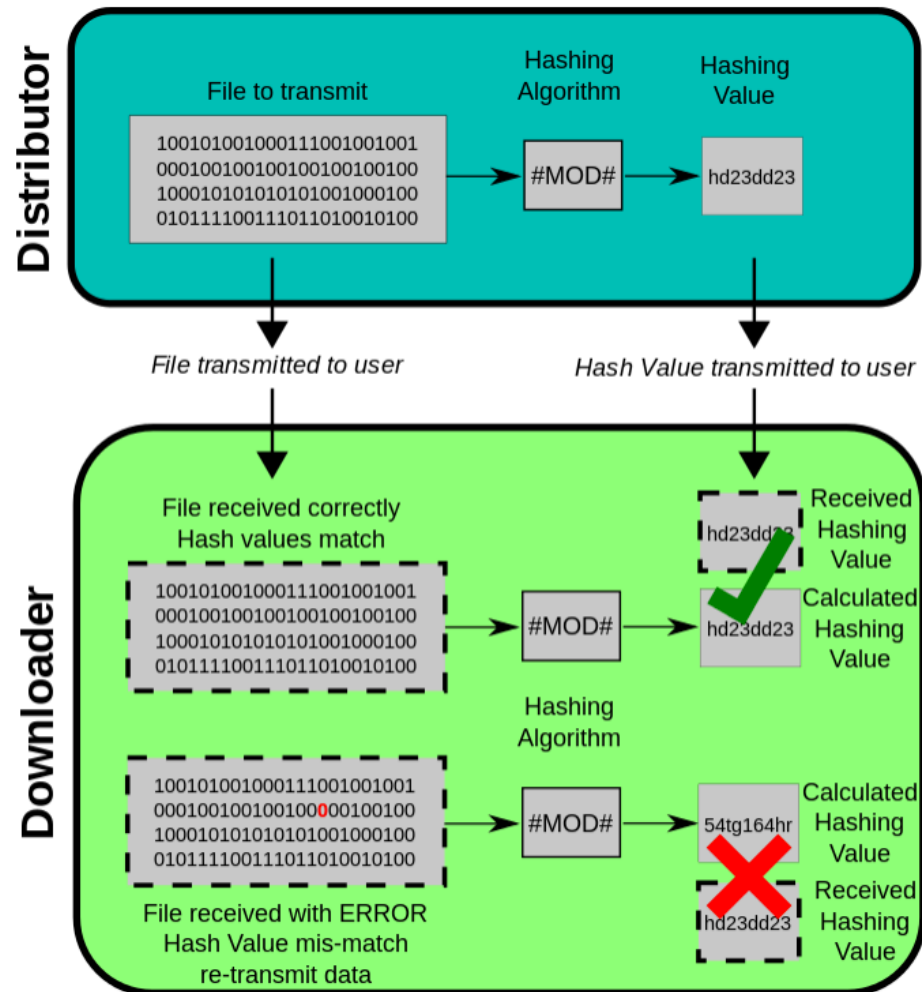
# Preparing Fastq for analysis

- 1) Check files for completeness, use md5 checksums if file corruption is suspected
- 2) Inspect quality statistics
- 3) Possible steps to clean files (choice of steps depends on the application)

- De-multiplex
  - Trim adapters
  - Filter low quality base calls
  - Remove duplicate sequences
  - Remove contaminant sequences
  - Remove sequences that are mainly adapter
- 
- The diagram uses colored brackets to group the steps in the list:
- A green bracket groups "De-multiplex" and "Trim adapters", with the text "Usually done by sequencing center" in green to its right.
  - A red bracket groups "Filter low quality base calls", "Remove duplicate sequences", and "Remove contaminant sequences", with the text "Genotyping and RNAseq" in red to its right.
  - A purple bracket groups "Remove duplicate sequences", "Remove contaminant sequences", and "Remove sequences that are mainly adapter", with the text "Reference assembly" in purple to its right.

Many programs to implement these steps!

# Preparing Fastq: md5 checksum



# Preparing Fastq: Quality metrics

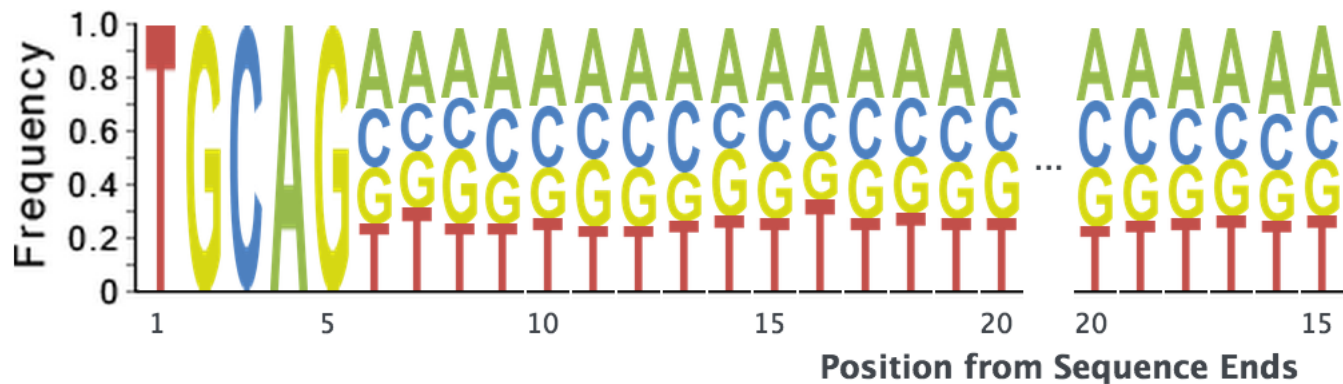
Many possible statistics to query:

- Number and length of sequences
- Base qualities
- Poly A/T tails
- Presence of tag sequences (stuff you added during preparation)
- Sequence complexity (e.g. ATATATATATA...)

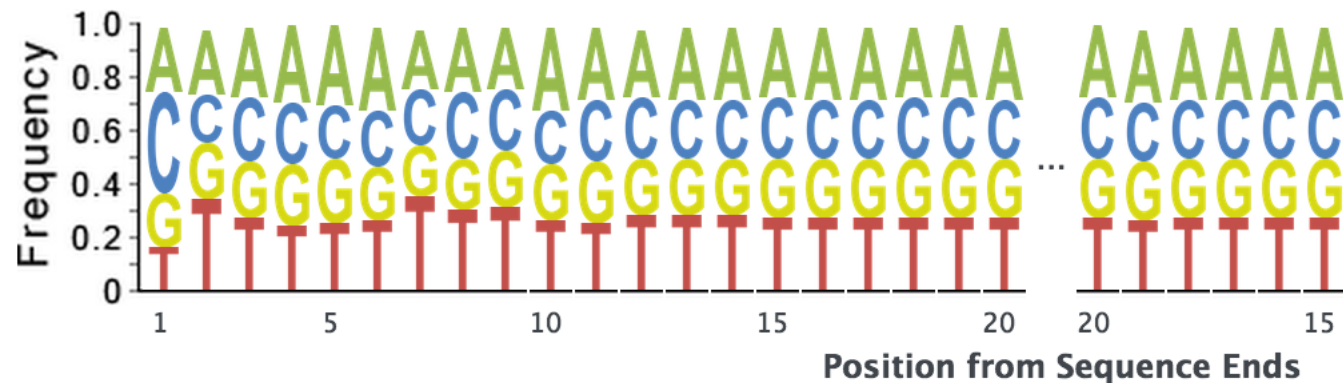
Recommended tools: prinseq, fastqc

# Preparing Fastq: Quality metrics

Distribution of base frequencies in GBS reads with enzyme cut site:

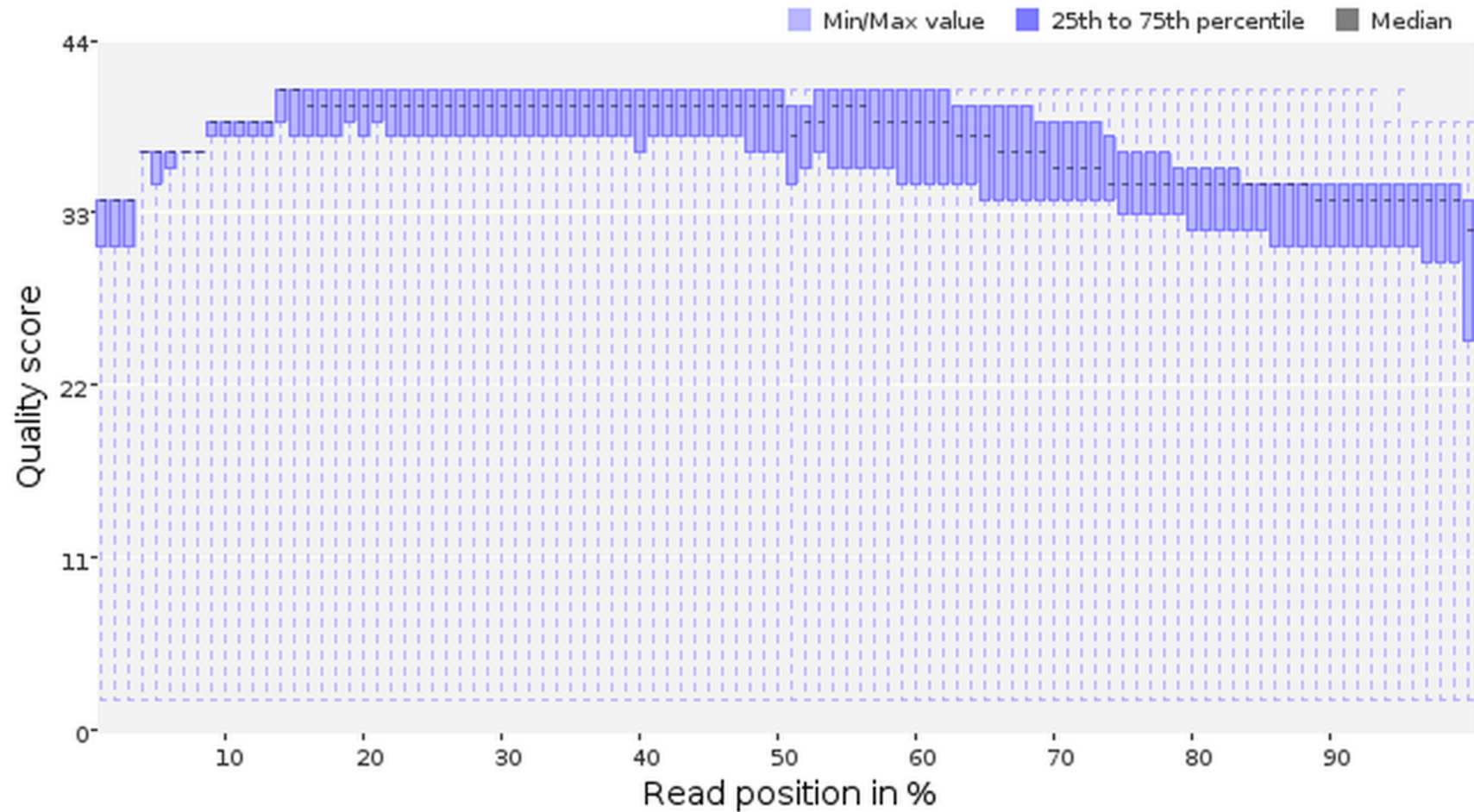


Distribution in RNAseq data, no adapters/tags:



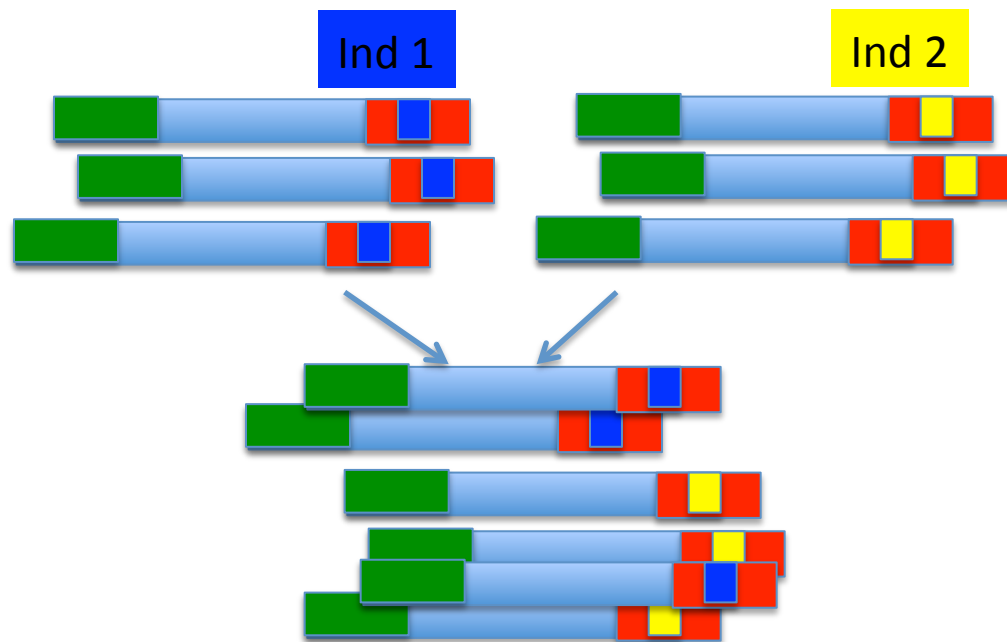
# Preparing Fastq: Quality metrics

A normal quality score distribution for Illumina reads:



# Preparing Fastq: De-multiplexing

Multiplexing is when several libraries are barcoded and sequenced on the same lane



- Most sequencing centers will de-multiplex the data
- Casava can be used for de-multiplexing and trimming barcodes from standard Illumina library preps

# Preparing Fastq: Trimming

**Table 1.** Availability and characteristics of the trimming tools investigated in the current work.

Tool	Version	Link	Language	Algorithm family	Can work directly on gzip	Can work on paired end	PHRED format autodetection	Works on both read ends	Notes
Cutadapt	1.1	<a href="http://code.google.com/p/cutadapt/downloads/list">code.google.com/p/cutadapt/downloads/list</a>	Python and C	Running sum	yes	no	no	no	Can also remove adapters, multi-threaded
ConDeTri	2.2	<a href="http://code.google.com/p/condetri/">code.google.com/p/condetri/</a>	Perl	Window based	yes (since v2.2)	yes	no	no	
ERNE-FILTER	1.2	<a href="http://sourceforge.net/projects/erne/files/">sourceforge.net/projects/erne/files/</a>	C++	Running sum	yes	yes	yes	yes	Can be combined with contaminant removal, multi-threaded
FASTX quality trimmer	0.0.13.2	<a href="http://hannonlab.cshl.edu/fastx_toolkit/download.html">hannonlab.cshl.edu/fastx_toolkit/download.html</a>	C++	Window based	no	no	no	no	The default minimum read length parameter (-p) is set to zero
PRINSEQ	0:19:05	<a href="http://sourceforge.net/projects/prinseq/files/">sourceforge.net/projects/prinseq/files/</a>	Perl	Window based	no	no	no	yes	Also web interface for medium-size data
Trimmomatic	0.22	<a href="http://www.usadellab.org/cms/index.php?page=trimmomatic">www.usadellab.org/cms/index.php?page=trimmomatic</a>	Java	Window based	yes	yes	no	yes	Can also remove adapters
SolexaQA	1.13	<a href="http://sourceforge.net/projects/solexaqa/files/">sourceforge.net/projects/solexaqa/files/</a>	Perl	Window based (Running sum with -bwa option)	no	no	yes	no	Cannot specify minimum read length to keep
Sickle	1.2	<a href="https://github.com/ucdavis-bioinformatics/sickle">github.com/ucdavis-bioinformatics/sickle</a>	C	Window based	yes	yes	no	yes	

doi: 10.1371/journal.pone.0085024.t001

# Preparing Fastq: Trimming

- Adapters are short sequences that are added to the beginning and end of DNA molecules to prepare them for sequencing



- Can compromise how well the reads align to a reference if not removed
- Detect during the quality control phase
- Removed by a range of tools (most sequencing centers will already have removed the adapters)

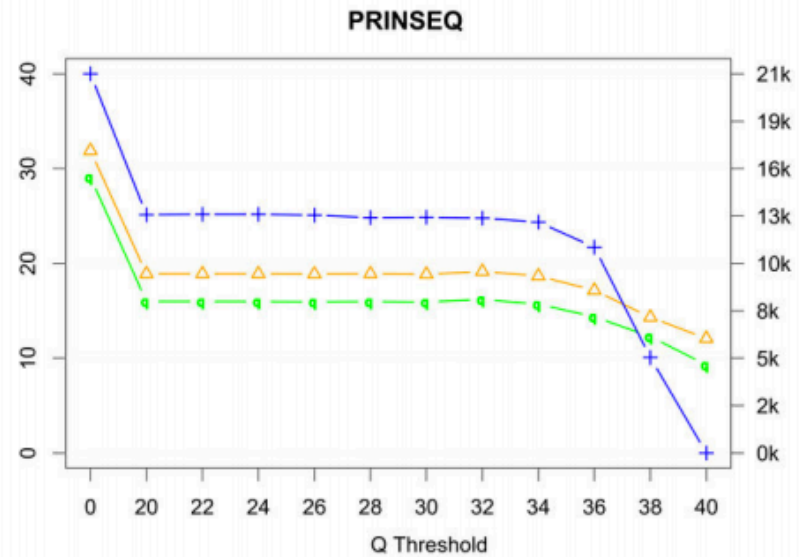
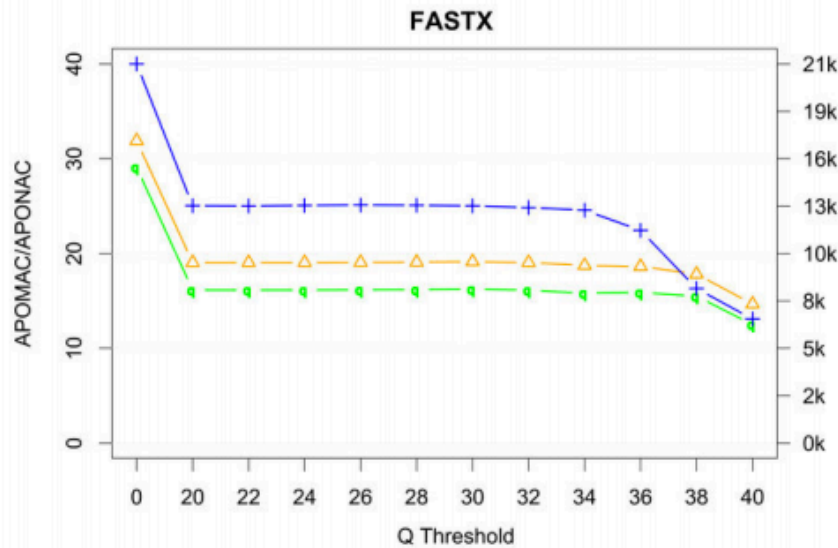


# Preparing Fastq: Filtering

Choice of quality score to filter to depends upon the application:

- Too low a quality score cutoff:
  - 1) increase run times and RAM usage
  - 2) bad results (e.g. false SNP calls)
- Too high a quality score cutoff:
  - 1) faster run times
  - 2) lose useful data (e.g. more fragmented assemblies, missing SNPs)
- Usually Q20, but sometimes lower or higher

# Preparing Fastq: Filtering



# Preparing Fastq: Assembly

- Remove sequences consisting of adapter dimers (otherwise, they may be included as contigs). (e.g. tagdust)
- Clean out contaminants by blasting to known databases (can also be conducted post-assembly)
- Remove duplicate sequences: for *de novo* assembly, sequences that are exact copies will slow down the assembly without adding anything (e.g. fastx\_collapser)

## Preparing Fastq: Pairing

- With paired-end reads, if one read direction is removed but the other is not, then the \_R1 and \_R2 files are mismatched
- Need to run a script to eliminate unpaired reads from each \_R1 and \_R2 file

Some programs output reads in paired and unpaired files (e.g. prinseq, Trimmomatic). Others do not and custom scripts are required to re-pair data.

# Preparing Fastq: GBS-specific filtering

- GBS / RAD use enzymes to cleave the DNA, so all reads will begin with the recognition sequence:

```
TGCAGTCCAACGCCACGGTCAAAGAATACCAGCTTTTAAATTAACTTTGCCCCGGTCTTCC/  
TGCAGTCCTCGGTGTCAGGAGTATAACTGCATTGTGTCATCTTCATGGTGAAGATCTCTGCT`  
TGCAGCATCCTATTTCTAATTTGGATTTAAATAAAACTGGAAGCTATTGTAAGTCCCCGGCC`  
TGCAGTGTTACTCTTACCTCCTGAATTGAACGGAAAACGATCTAGCAAACTGAACTGCCAT`  
TGCAGGTGAAATGAGAGAGGAAGATTGGGGTCAAATAAATTTTCCTAAAGTGGAAGCTTTGAI  
TGCAGAGAAGGGAAATGCAGAGTCTGTGCTGAAGGCCATTGGCGATTTTAAATAGCCATACCT(  
TGCAGGGTATTTAGTTTTTTGAATGAGAATTTTCTGACTTGAGATTTTTTTACTGTTTCAGTATC(  
TGCAGCAGTTTGAGTAAGAGGAAAATGGTTTTCCAAAATTCACA ACTTAAAGAAACATCCATC
```

- Will need to de-multiplex using Stacks or custom script
- Clean GBS-specific adapters or other home-brew sequences that sequencing centers didn't remove

## Further reading

- Del Fabbro et al. 2013. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. PLoSOne. 8:e85024.
- [http://prinseq.sourceforge.net/Data\\_preprocessing.pdf](http://prinseq.sourceforge.net/Data_preprocessing.pdf)
- <http://prinseq.sourceforge.net/manual.html#STANDALONE>

# Tutorial

- Navigate to the directory ~/data/Topic3
- Open the README file there, and follow the directions.

Questions:

1) Compare the two .html files for the initial filtering options. What kinds of differences do you see in the files? Why do you think these differences are found (think about the types of data you are analyzing)?

2) Try different filtering options for the GBS data (see <http://prinseq.sourceforge.net/manual.html> for options) and plot QC graphs. Discuss in a group of four which options you would choose to implement if this was your data. Be prepared to share your findings with the class.