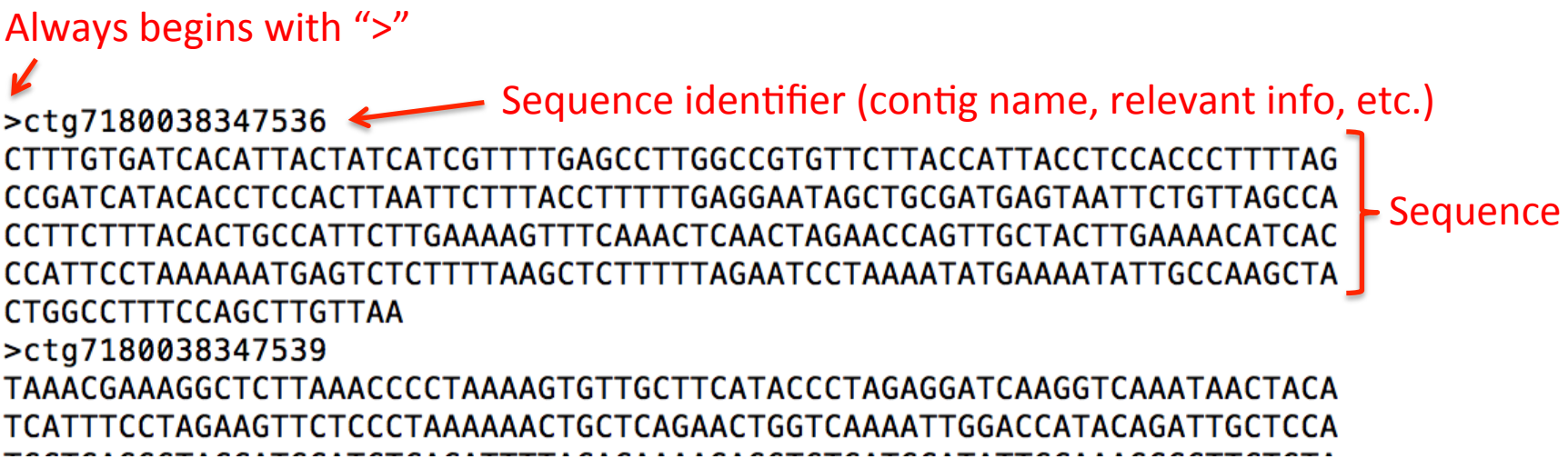
UBC Bioinformatics Class

Topic 2: Fastq files and quality checking and trimming

NGS File formats

FASTA:

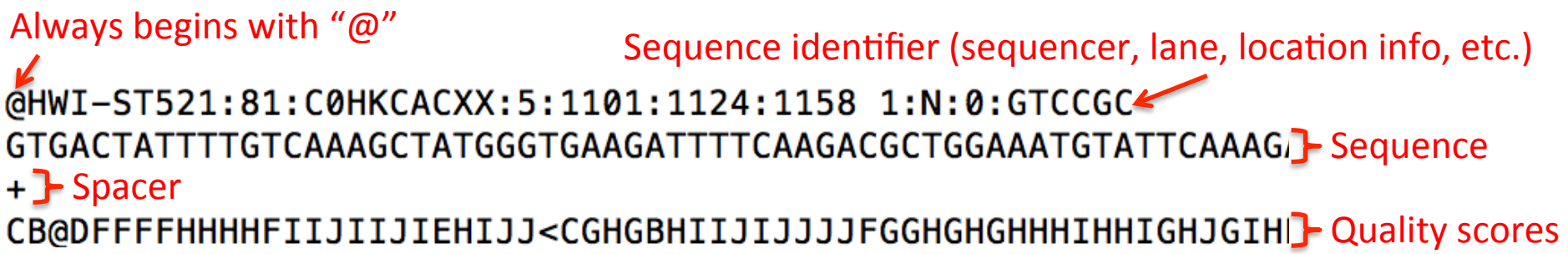
- Sequence and quality scores are stored in separate files (usually .fasta or .fa & .qual)
- Now mainly used for storing reference sequences (no qual scores) as either nucleotides or peptides
- 2 lines/sequence read:



NGS File formats

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:

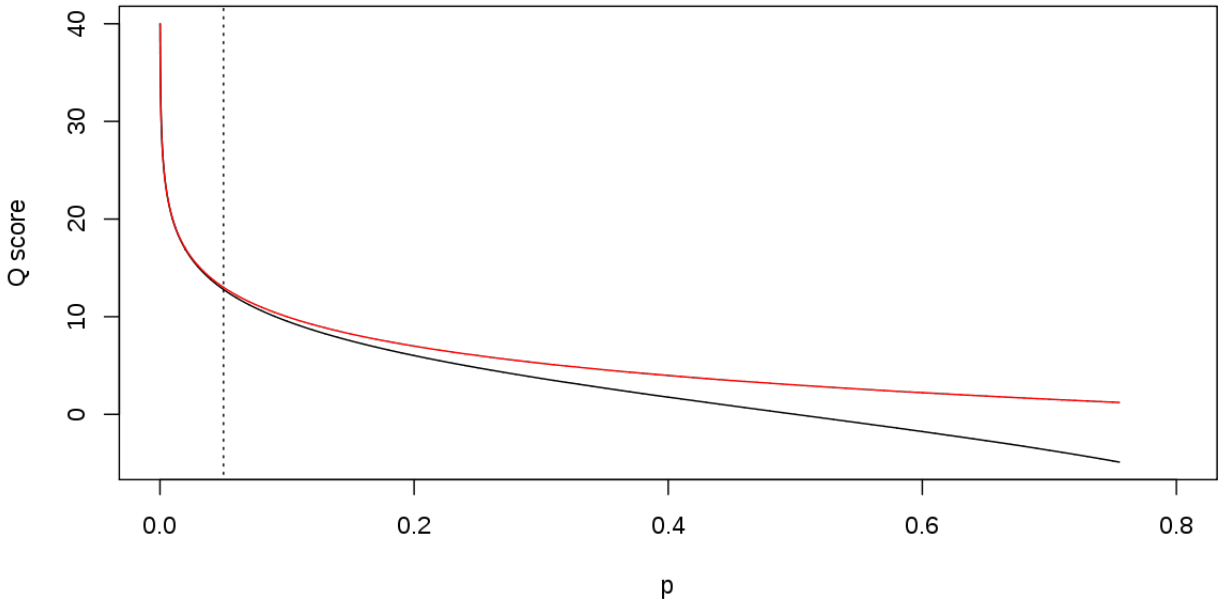


NGS Quality scores

Historically, two formats (now all are Sanger)

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

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]

http://en.wikipedia.org/wiki/FASTQ_format

NGS Encoding of quality scores

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ..................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 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)
```

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

http://en.wikipedia.org/wiki/FASTQ_format

Preparing FASTQ data for analysis

- Check files for completeness, use md5 checksums if file corruption is suspected
- Inspect quality statistics
- Many possible steps to clean files, and the choice of steps to include depends on the application
  - De-multiplexing
  - Trimming adapters
  - Filtering low quality base calls
  - Removing contaminant sequences
  - Removing duplicate sequences
  - Removing sequences that are mainly adapter

Usually done by sequencing center

Genotyping and RNAseq

Reference assembly

Many programs to implement these steps!

Inspecting quality statistics

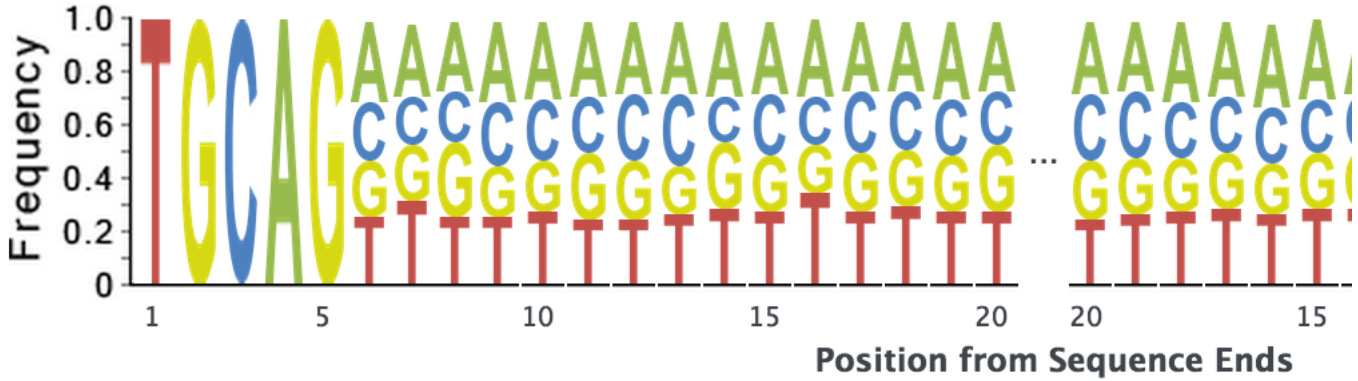
Many possible statistics to query:

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

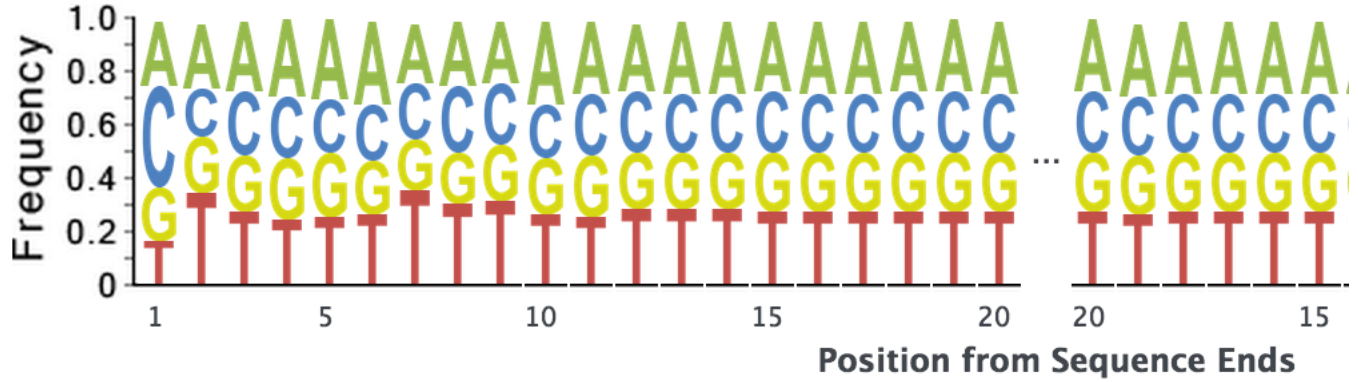
Recommended tools: prinseq, fastqc

Examples of quality metrics

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

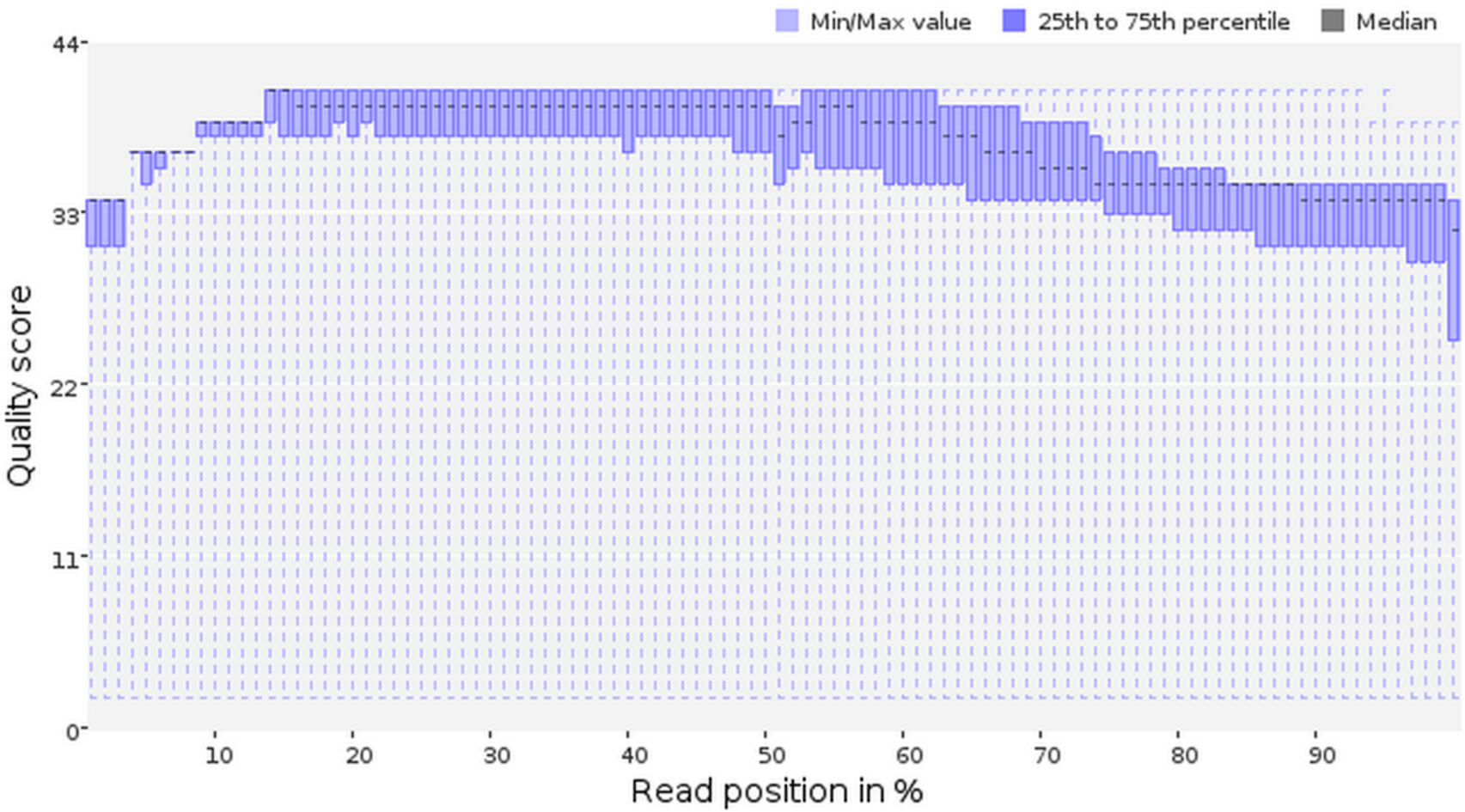


Distribution in RNAseq data, no adapters/tags:



Examples of quality metrics

A normal quality score distribution for Illumina reads:



De-multiplexing

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

- Most sequencing centers will de-multiplex the data prior to returning it
- Otherwise, **casava** can be used for de-multiplexing and trimming the barcodes

Trimming adapters

- Adapters are short sequences that are added to the beginning and end of DNA molecules to prepare them for sequencing
- Failure to remove adapters from sequence will compromise how well the reads align to a reference
- Adapters can be detected during the quality control phase and then removed by a range of tools
- Most sequencing centers will already have removed the adapters before giving you the data

Recommended tools: fastx toolkit, trimmomatic, prinseq

Filtering low quality base calls

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

- Too low a quality score cutoff and bad sequence information will be included
  - May not be a problem for alignment and genotype calling, but it can slow things down
  - More of a problem for assembly
- Too high a quality score cutoff will result in loss of useful data
- Usually Q10-Q20, but sometimes lower or higher

Recommended tools: fastx toolkit, trimmomatic, prinseq

Additional filtering for assembly

- Remove sequences consisting of adapter dimers (otherwise, they may be included as contigs). Use **tagdust**
- Clean out contaminants by blasting to known databases (can also be conducted post-assembly)
- Remove duplicate sequences: if you are doing a de novo assembly, sequences that are exact copies will slow down the assembly without adding anything (remove using **fastx_collapser**)

SnoWhite is a good pipeline for aggressive cleaning and filtering prior to assembly

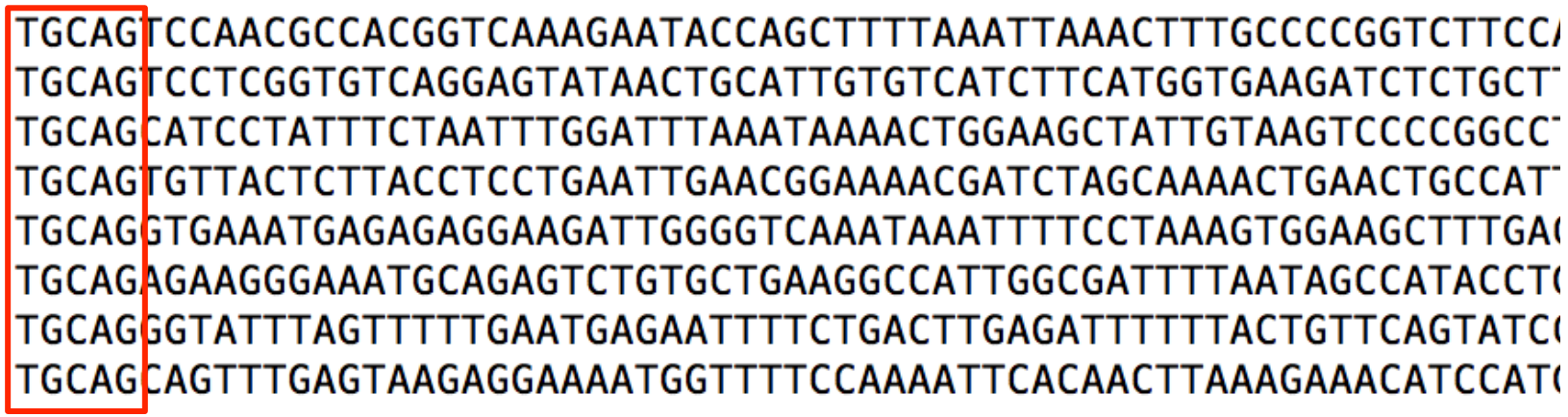
Re-pairing reads after filtering

- 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.

GBS-specific filtering

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



- Cut the first X-bases from each sequence (all filtering programs)
- May need to 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/manual.html#STANDALONE

Starting the exercise:

To make a new directory:
mkdir dir_name
mv filename ./dirname

- Start your VM in virtualbox
- Open “firefox” under the activities menu
- Navigate to www.zoology.ubc.ca/~yeaman/bioinformatics
- “right click” on the link that says “`data_cleaning_and_manipulation.tbz`” and choose “copy link location” [for mac, this should be command-mouseclick; not sure about PC keyboard]
- Open “terminal” under the activities menu, type “`wget` “ and then paste in the link that you copied. Alternatively, type in the link yourself and add the filename at the end. This will download today’s dataset onto your virtual machine.
- Type “`tar –xjvf data_cleaning_and_manipulation.tbz`”. This will unpack the archive and create a new directory. Navigate to the directory (type “cd data_cleaning_and_manipulation”) and open the README file there, and follow the directions. You can open the README file in a more friendly text editor by choosing the “files” icon from the activities menu and opening the README file there.