Intro to R for Biologists

IBiS Special Topics, Fall 2021

Class 8: Oct. 18, 2021

(Bash/Quest Interlude)

Erik Andersen and Shelby Blythe

Today’s class

1. Illumina Sequencing and file formats
2. Fast Forward to the day you submit your paper
3. A Mapping Pipeline
4. Activity: Confirm that you can get a test script to run.

Before we begin

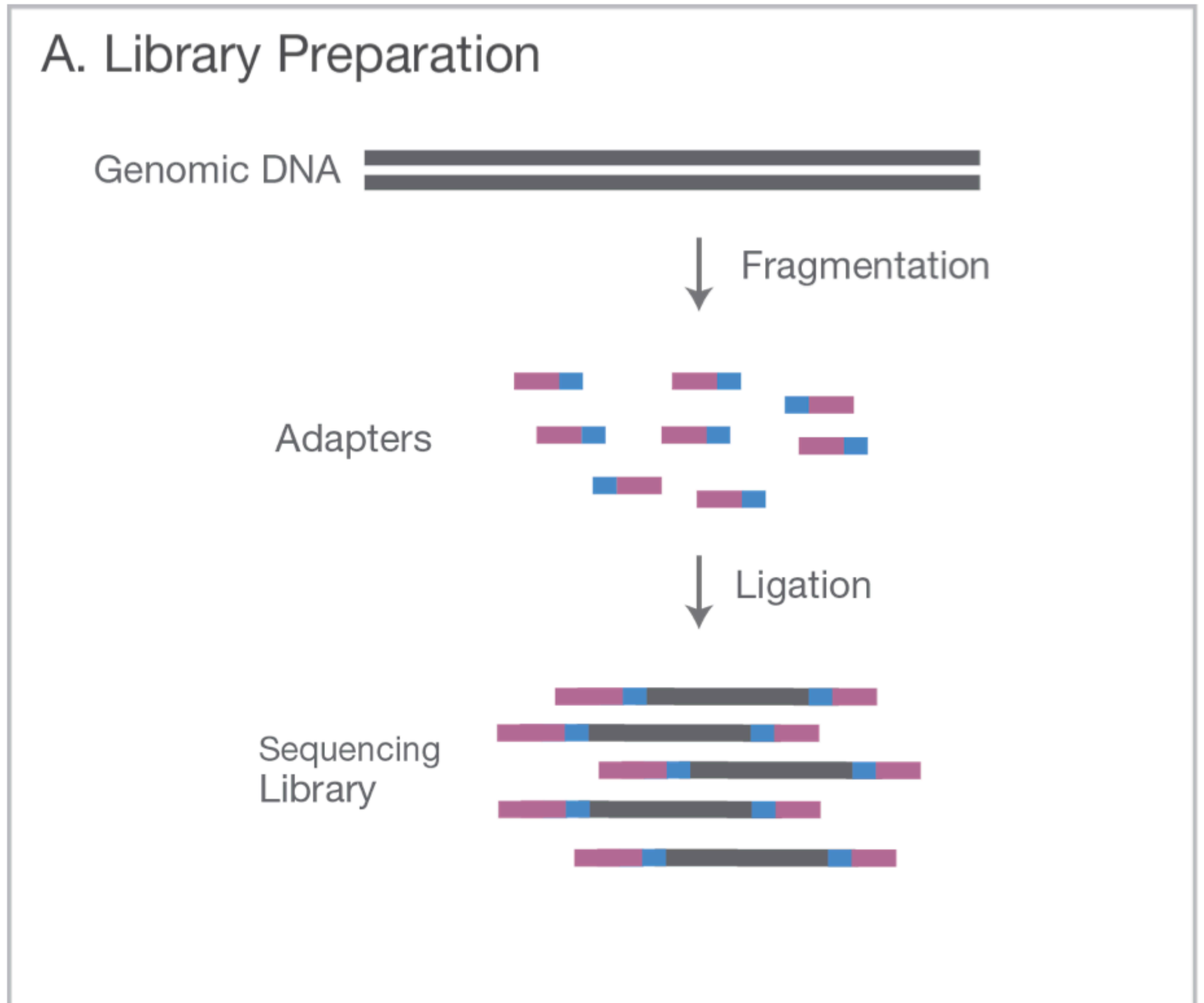
- Has everyone:

    1. Confirmed that they can log in to Quest with terminal and ssh and can navigate to the class allocation?

    2. Downloaded some kind of code-aware text editor (e.g., Atom or VSCode)?

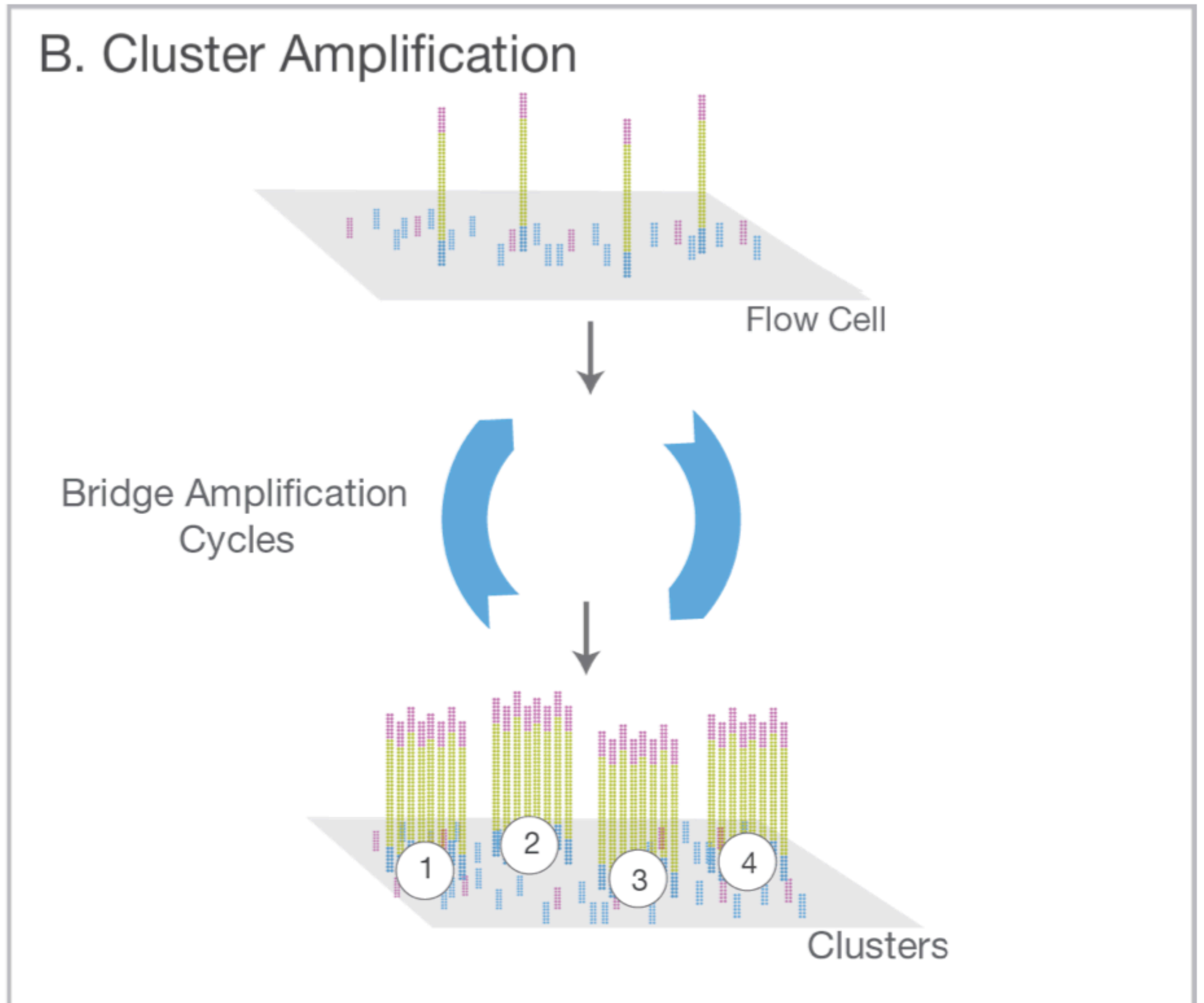
    3. Successfully set up Globus (or some other appropriate file-transfer system) to send/receive files from Quest?

Illumina Sequencing



A. Library Preparation

NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.



B. Cluster Amplification

Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

The DNA fragments you sequence are capped on either end by Illumina sequencing adapters.

When you sequence the fragments, you recover X bp of sequence which may or may not be greater than the size of the original DNA fragment.

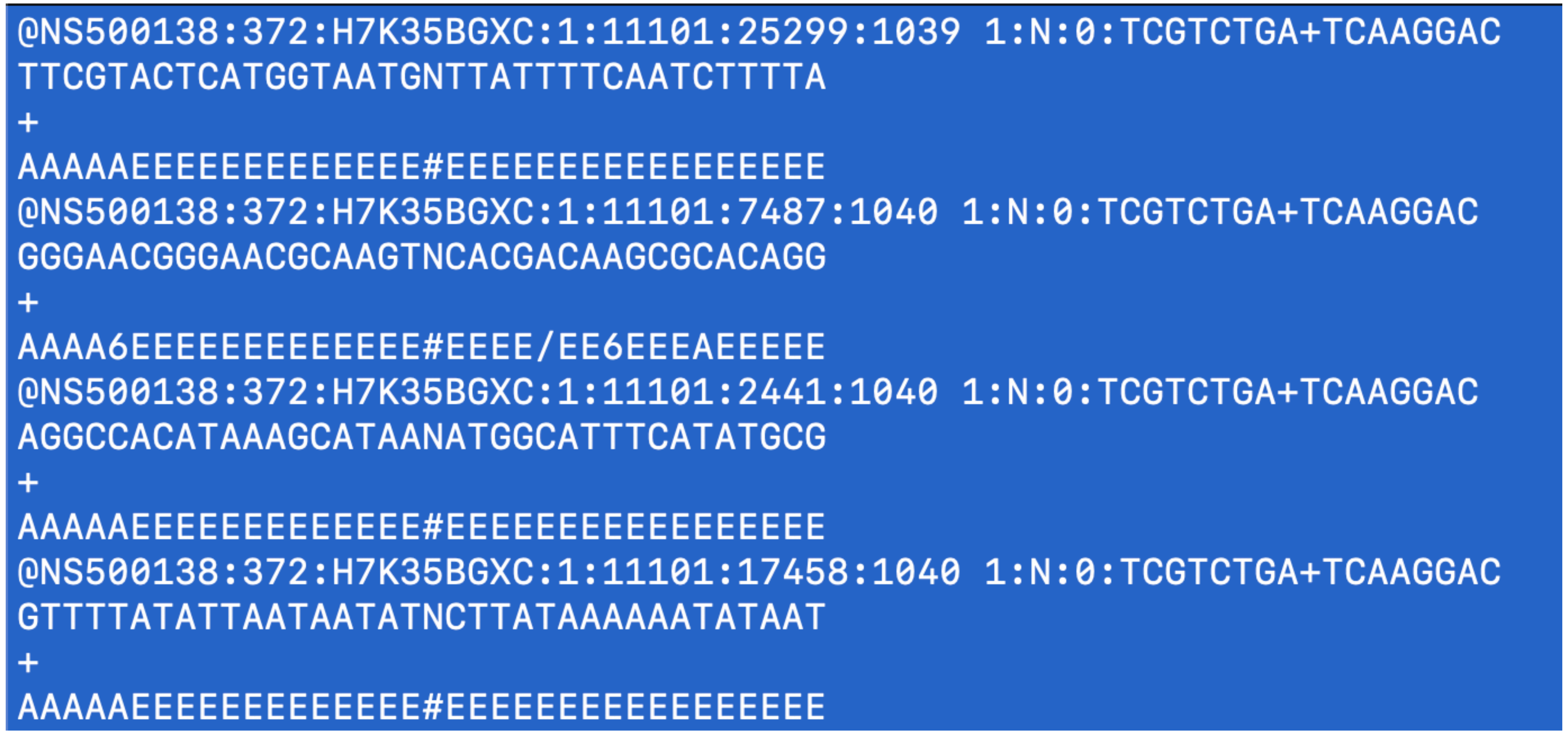
What does the data you get from the sequencer look like?

The FastQ file format:

- These are the first 4 sequences in our test data set for today’s activity. They are in FastQ format.

- A single sequence is represented by 4 lines in this format.

  1. Identifier
  2. Sequence
  3. +
  4. Quality Score



```
@NS500138:372:H7K35BGXC:1:11101:25299:1039 1:N:0:TCGTCTGA+TCAAGGAC
TTCGTACTCATGGTAATGNTTATTTTCAATCTTTTA
+
AAAAAEEEEEEEEEEEEE#EEEEEEEEEEEEEEEEE
@NS500138:372:H7K35BGXC:1:11101:7487:1040 1:N:0:TCGTCTGA+TCAAGGAC
GGGAACGGGAACGCAAGTNCACGACAAGCGCACAGG
+
AAAA6EEEEEEEEEEEEE#EEEE/EE6EEEAEEEEE
@NS500138:372:H7K35BGXC:1:11101:2441:1040 1:N:0:TCGTCTGA+TCAAGGAC
AGGCCACATAAAGCATAANATGGCATTTCATATGCG
+
AAAAAEEEEEEEEEEEEE#EEEEEEEEEEEEEEEEE
@NS500138:372:H7K35BGXC:1:11101:17458:1040 1:N:0:TCGTCTGA+TCAAGGAC
GTTTTATATTAATAATATNCTTATAAAAAATATAAT
+
AAAAAEEEEEEEEEEEEE#EEEEEEEEEEEEEEEEE
```

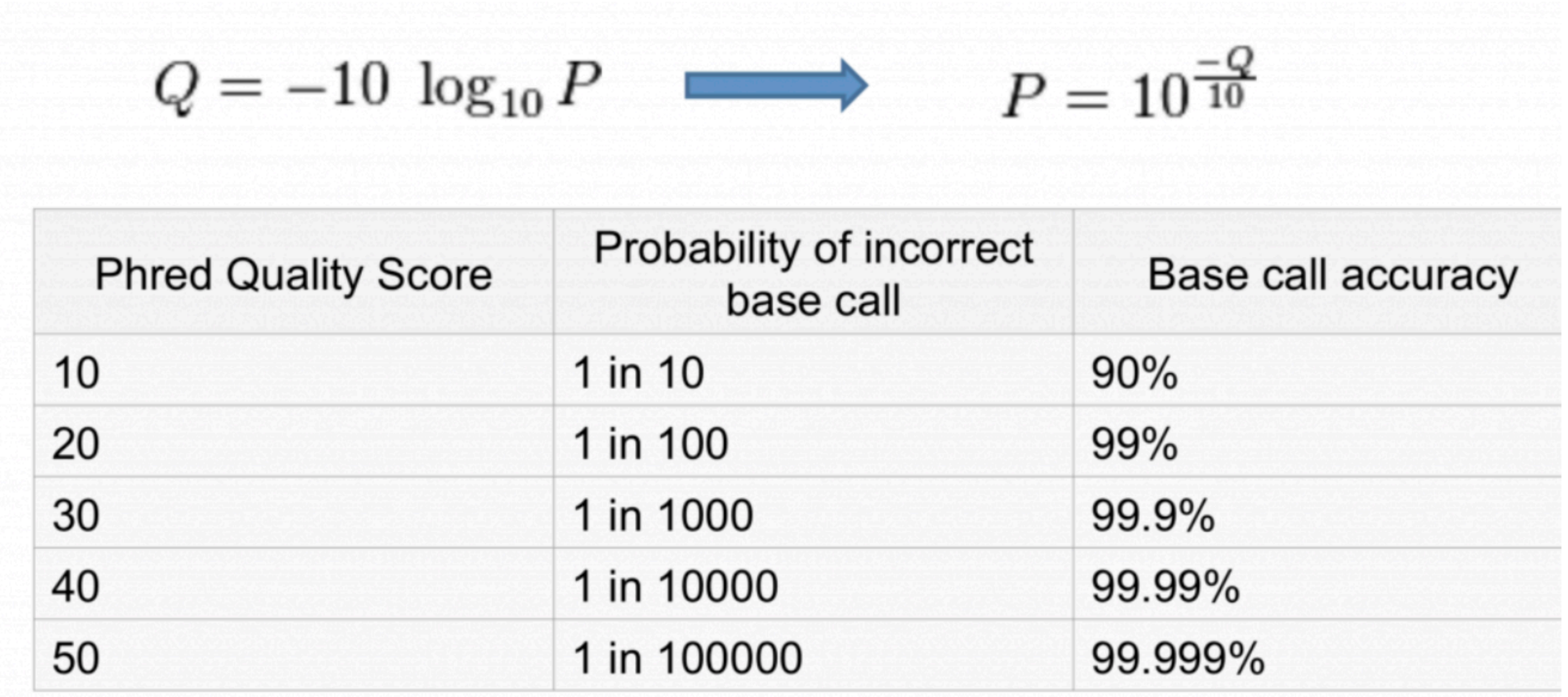
Identifier

```
@NS500138:372:H7K35BGXC:1:11101:25299:1039 1:N:0:TCGTCTGA+TCAAGGAC
```

- There is potentially useful information here:
  - The machine that did the sequencing (NS500138)
  - Run ID (372)
  - Flow Cell ID (H7K35BGXC)
  - Flow Cell Lane (1)
  - Tile (11101)
  - X-coordinate (25299)
  - Y-coordinate (1039)
  - Member of which pair? (1) *paired-end sequencing only*
  - Filtered? (N(o))
  - Control bits on? (0)
  - Barcodes (TCGTCTGA + TCAAGGAC)

Sequence and Quality Score

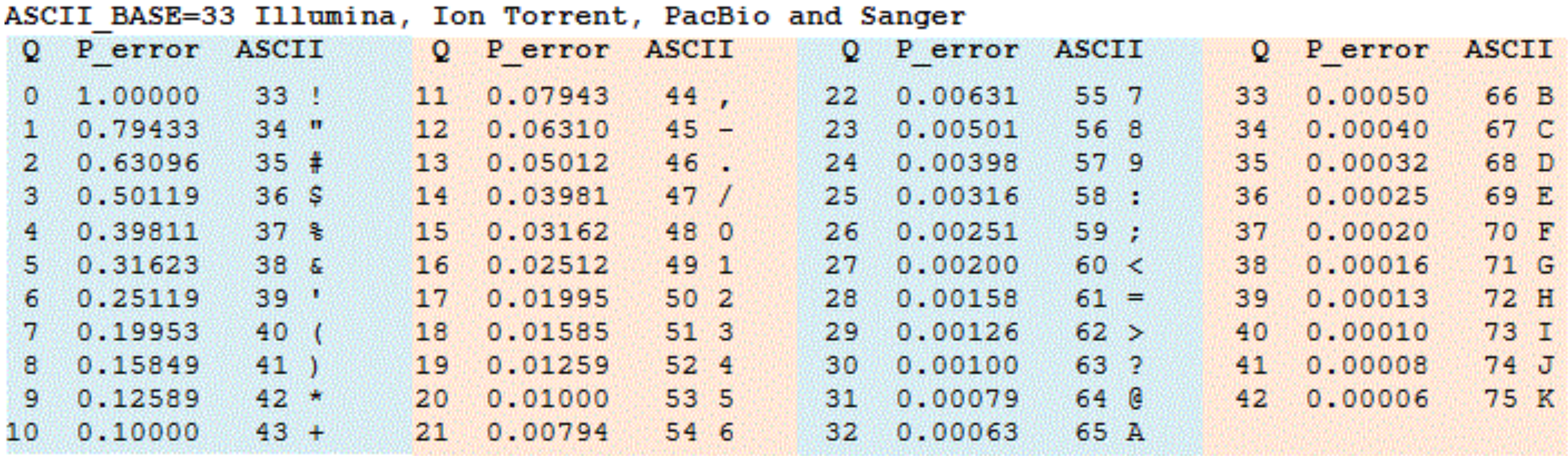
```
TTCGTACTCATGGTAATGNTTATTTTCAATCTTTTA
+
AAAAAEEEEEEEEEEEEE#EEEEEEEEEEEEEEEEE
```



$Q = -10\ \log_{10} P \quad \Rightarrow \quad P = 10^{\frac{-Q}{10}}$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

- Each base-call is identified and brings along with it an estimate of how confident the machine is that the base-call is correct.
- The current quality score format is referred to as "Phred33"



ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

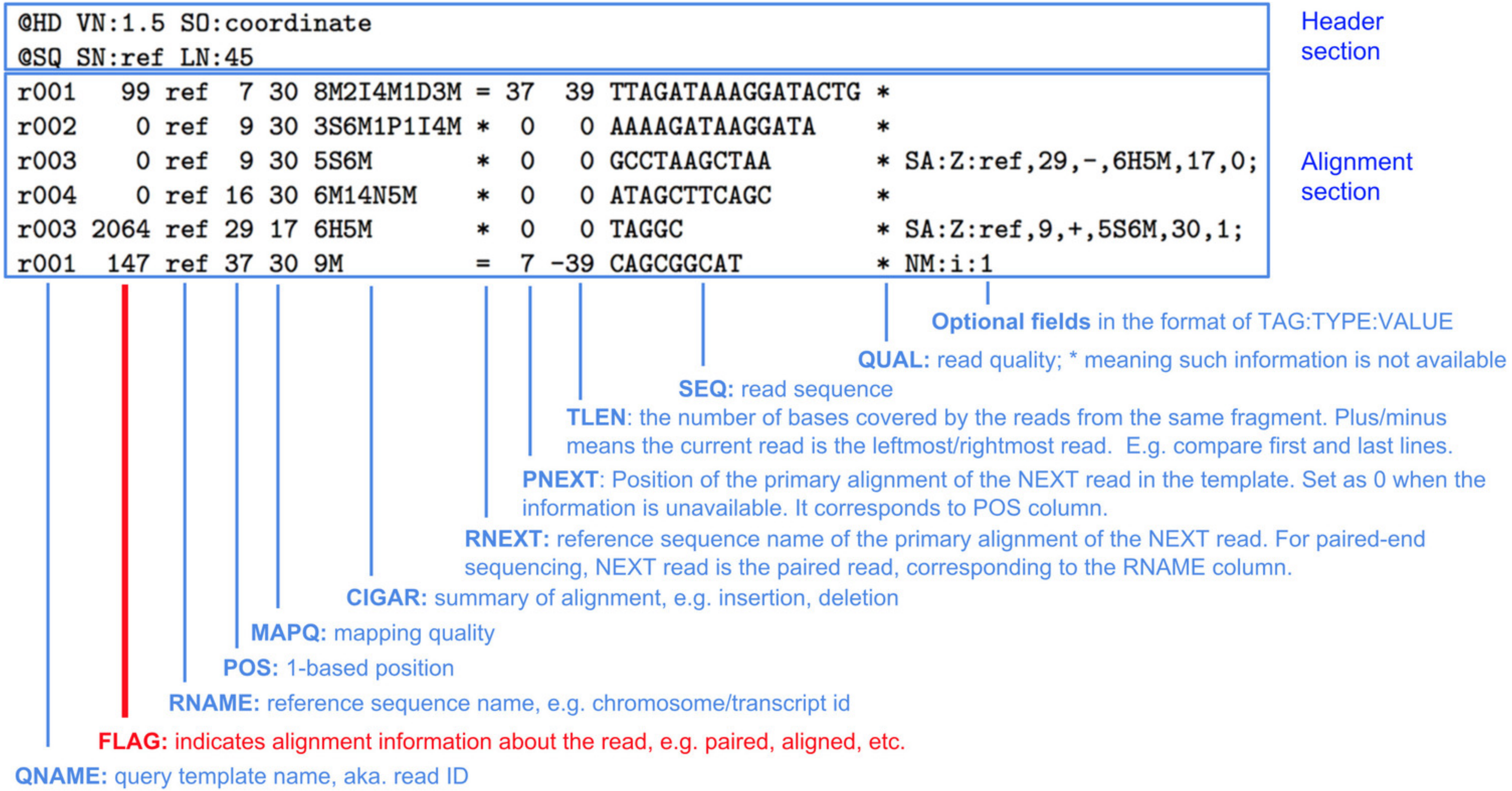
Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	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			

FastQ has no information about where the reads map to.

We have to process these reads and map them before they are informative.

- Typical steps in going from FastQ to “Mapped Reads”
    - Filter low-quality reads and trim adapter sequences (yields .fastq)
    - Map to reference genome (yields .sam)
    - Convert .sam to .bam (binarized .sam)
    - Sort reads (yields .bam)
    - Mark likely PCR or Optical Duplicates (yields .bam)
    - Index the .bam file (yields .bai)

The .sam/.bam format



```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Header section

Alignment section

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

The .sam/.bam format

Example: single end mapping to *Drosophila melanogaster*

“qname” = Name of the read from fastQ **flag** rname pos mapq cigar Sequence Phred Quality Scores

```
NS500138:372:H7K35BGXC:1:12111:21030:5973     16      chr2L    20634    31     36M     *       0       0       CTGGTCGGTACTAATATTTTTTCTAAGACTGTATAC    EEEE/EEEEEEEEEEEEEEEEEEEEEEEEEEAAAAA
NS500138:372:H7K35BGXC:1:11205:21596:8754     0       chr2L    20690    1      36M     *       0       0       CGTAGGTATGTATCCGTGAGTAAGCTGAGGTGCAAA    AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
NS500138:372:H7K35BGXC:1:11201:23277:9037     0       chr2L    20726    1      36M     *       0       0       ACCGTATCACCATTTGCTTTTCGGCTGATAACCAAT    AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
NS500138:372:H7K35BGXC:1:11211:25328:12582    0       chr2L    20760    1      33M     *       0       0       ATACCAAGTAAGTCAACTGTTTCCCAAATGTAT       AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

rnext pnext tlen

- Note, rnext, pnext, and tlen are not recorded for single-end mappings
- Are these good mappings? Two clues are the **Flag** and the **mapq** columns.

The .sam/.bam format

Example: single end mapping to *Drosophila melanogaster*

"qname" = Name of the read from fastQ **flag** rname pos mapq cigar Sequence Phred Quality Scores

```
NS500138:372:H7K35BGXC:1:12111:21030:5973    16   chr2L  20634  31   36M   *   0   0   CTGGTCGGTACTAATATTTTTTCTAAGACTGTATAC   EEEE/EEEEEEEEEEEEEEEEEEEEEEEEEEAAAAA
NS500138:372:H7K35BGXC:1:11205:21596:8754    0    chr2L  20690  1    36M   *   0   0   CGTAGGTATGTATCCGTGAGTAAGCTGAGGTGCAAA   AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
NS500138:372:H7K35BGXC:1:11201:23277:9037    0    chr2L  20726  1    36M   *   0   0   ACCGTATCACCATTTGCTTTTCGGCTGATAACCAAT   AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
NS500138:372:H7K35BGXC:1:11211:25328:12582   0    chr2L  20760  1    33M   *   0   0   ATACCAAGTAAGTCAACTGTTTCCCAAATGTAT      AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

rnext pnext tlen

- Bowtie2 'map quality': a complex subject.
    - 0 and 1 indicate that a read maps to multiple locations in the genome and the position cannot be uniquely assigned. Whether 0 or 1 depends on how many mismatches there are to the genome assembly.
    - A score of 42 is a perfect unique match to the genome assembly
    - Values in between are probably mapped to the right location, but there are other potential matches. Here, the score is roughly `-10 * log10(probability of mismatch)`
- For a great discussion about Bowtie2 map quality scoring, please see:
- http://biofinysics.blogspot.com/2014/05/how-does-bowtie2-assign-mapq-scores.html

The .sam/.bam format

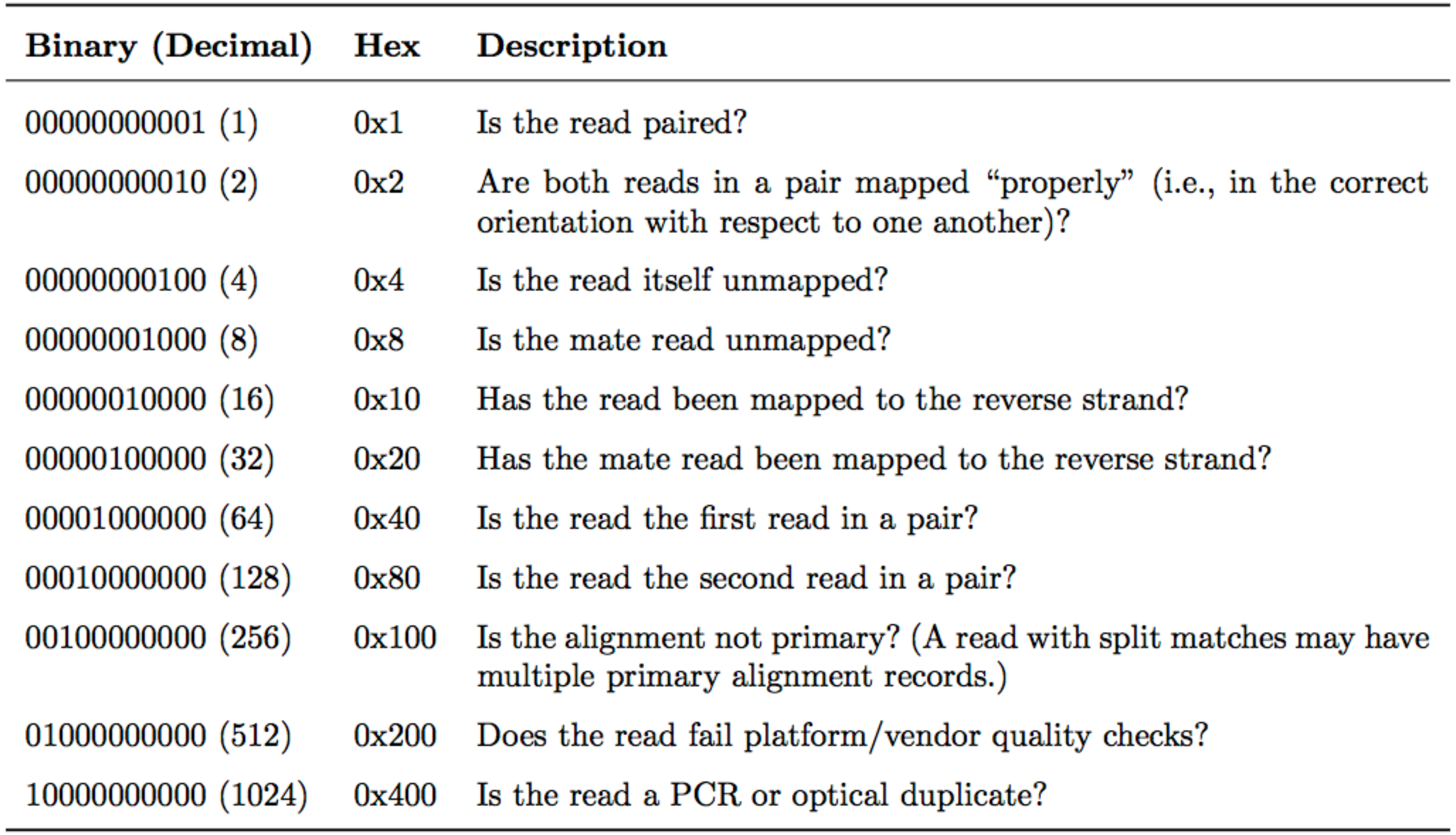
Example: single end mapping to *Drosophila melanogaster*

“qname” = Name of the read from fastQ **flag** rname pos mapq cigar Sequence Phred Quality Scores

```
NS500138:372:H7K35BGXC:1:12111:21030:5973    16   chr2L  20634  31   36M   *   0   0   CTGGTCGGTACTAATATTTTTTCTAAGACTGTATAC  EEEE/EEEEEEEEEEEEEEEEEEEEEEEEEEAAAAA
NS500138:372:H7K35BGXC:1:11205:21596:8754    0    chr2L  20690  1    36M   *   0   0   CGTAGGTATGTATCCGTGAGTAAGCTGAGGTGCAAA  AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
NS500138:372:H7K35BGXC:1:11201:23277:9037    0    chr2L  20726  1    36M   *   0   0   ACCGTATCACCATTTGCTTTTCGGCTGATAACCAAT  AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
NS500138:372:H7K35BGXC:1:11211:25328:12582   0    chr2L  20760  1    33M   *   0   0   ATACCAAGTAAGTCAACTGTTTCCCAAATGTAT     AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

rnext pnext tlen

- Flags are equally confusing at first. Most are designated for describing paired-end data.
- In the example above, we have flags of 16 and 0
- 16 means “minus strand”, 0 means “plus strand” mapping.
- https://broadinstitute.github.io/picard/explain-flags.html



Binary (Decimal)	Hex	Description
00000000001 (1)	0x1	Is the read paired?
00000000010 (2)	0x2	Are both reads in a pair mapped “properly” (i.e., in the correct orientation with respect to one another)?
00000000100 (4)	0x4	Is the read itself unmapped?
00000001000 (8)	0x8	Is the mate read unmapped?
00000010000 (16)	0x10	Has the read been mapped to the reverse strand?
00000100000 (32)	0x20	Has the mate read been mapped to the reverse strand?
00001000000 (64)	0x40	Is the read the first read in a pair?
00010000000 (128)	0x80	Is the read the second read in a pair?
00100000000 (256)	0x100	Is the alignment not primary? (A read with split matches may have multiple primary alignment records.)
01000000000 (512)	0x200	Does the read fail platform/vendor quality checks?
10000000000 (1024)	0x400	Is the read a PCR or optical duplicate?

The .sam/.bam format

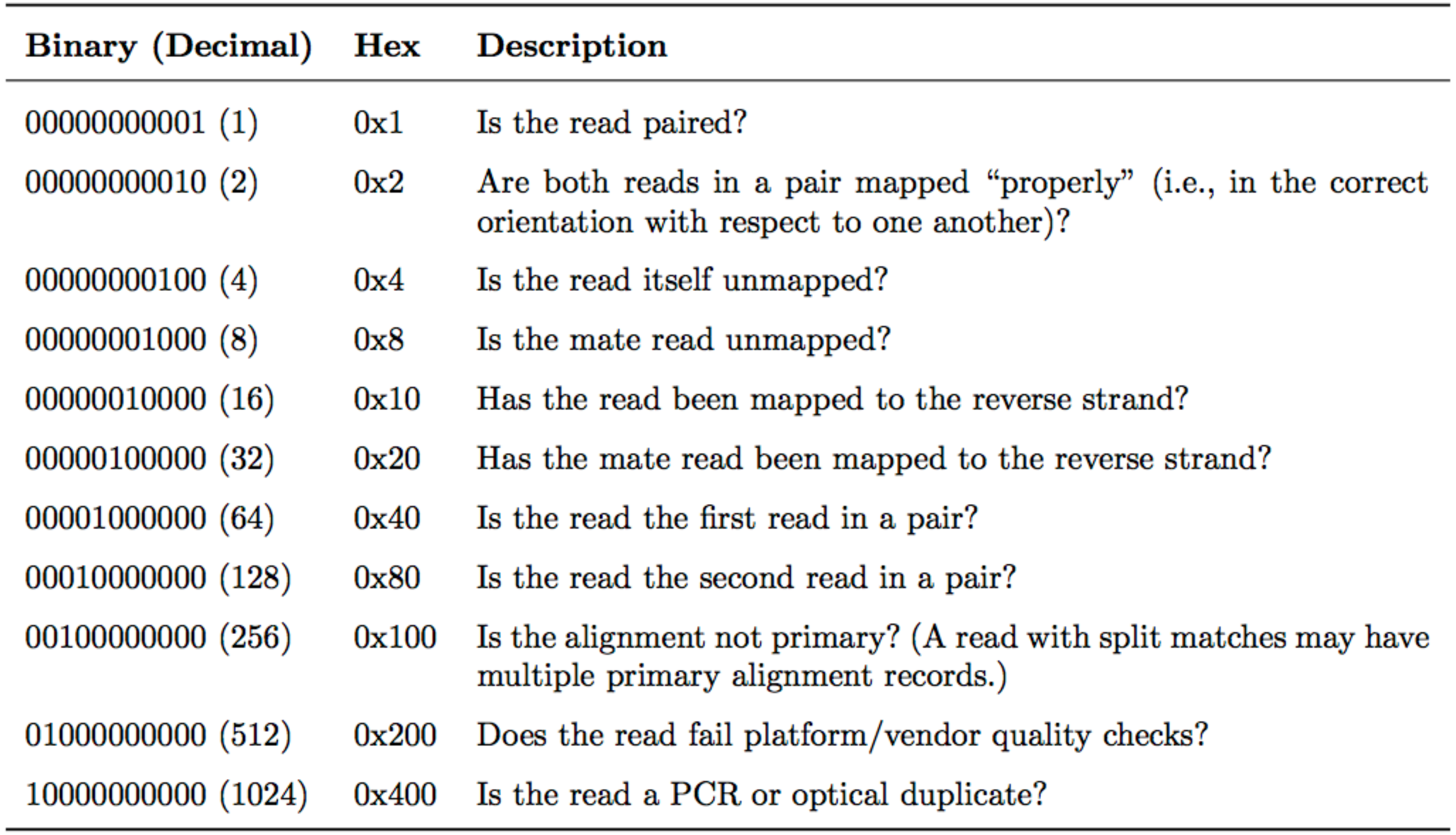
Example: single end mapping to *Drosophila melanogaster*

“qname” = Name of the read from fastQ **flag** rname pos mapq cigar Sequence Phred Quality Scores

```
NS500138:372:H7K35BGXC:1:11111:25532:10062   16     chr2L   356378  42     35M    *     0     0     CTGTGGCCGTACTCTGACTGCATAAGCACTGTAAC     EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAAAAA
NS500138:372:H7K35BGXC:1:11203:22330:18504   1040   chr2L   356378  42     35M    *     0     0     CTGTGGCCGTACTCTGACTGCATAAGCACTGTAAC     EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAAAAA
NS500138:372:H7K35BGXC:1:11105:14496:15080   1024   chr2L   356397  42     36M    *     0     0     GCATAAGCACTGTAACTGCCGGATAGACAGACAGGG    AA6AA//AEE/EEAE/E//EA//EEE/AAEEEEE/E
NS500138:372:H7K35BGXC:1:11109:5988:6069     0      chr2L   356397  42     36M    *     0     0     GCATAAGCACTGTAACTGCCGGATAGACAGACAGGG    AAAAAEEEEEEEEEEEEEEEEEEEEEE6EEEEEEEEE
```

rnext pnext tlen

- Flags are equally confusing at first. Most are designated for describing paired-end data.
- Flag values **sum**, the above example shows two duplicated reads, one on the minus, the other on the plus strand.
- https://broadinstitute.github.io/picard/explain-flags.html



Binary (Decimal)	Hex	Description
00000000001 (1)	0x1	Is the read paired?
00000000010 (2)	0x2	Are both reads in a pair mapped “properly” (i.e., in the correct orientation with respect to one another)?
00000000100 (4)	0x4	Is the read itself unmapped?
00000001000 (8)	0x8	Is the mate read unmapped?
00000010000 (16)	0x10	Has the read been mapped to the reverse strand?
00000100000 (32)	0x20	Has the mate read been mapped to the reverse strand?
00001000000 (64)	0x40	Is the read the first read in a pair?
00010000000 (128)	0x80	Is the read the second read in a pair?
00100000000 (256)	0x100	Is the alignment not primary? (A read with split matches may have multiple primary alignment records.)
01000000000 (512)	0x200	Does the read fail platform/vendor quality checks?
10000000000 (1024)	0x400	Is the read a PCR or optical duplicate?

The .sam/.bam format

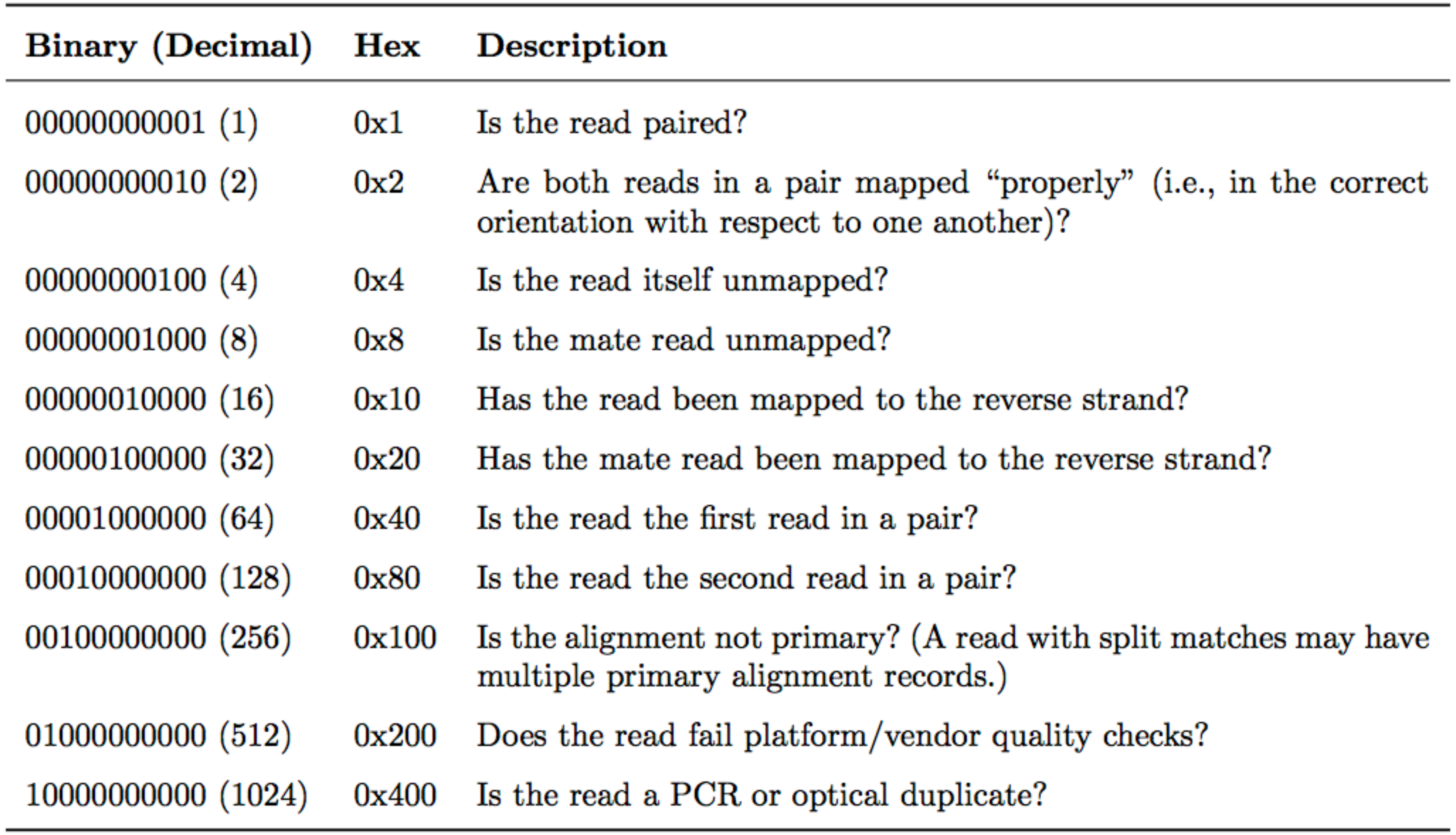
Example: paired-end mapping to *Drosophila melanogaster*

“qname” = Name of the read from fastQ **flag** rname pos mapq cigar Sequence Phred Quality Scores

```
NS500138:372:H7K35BGXC:1:12201:14062:6164    147    chr2L   83234   42    36M    =     83017   -253    CTAAAAAAGTAAGTTATCCACGAAACTGGACAGATT    EEEEEEEAAAEEE///AEEEEEEEEE6EEEEAAAAA
NS500138:372:H7K35BGXC:1:11106:22617:13759   147    chr2L   83253   42    36M    =     83155   -134    ACGAAACTGGACAGATTACTCGTCTTTCTTTCGTTT    EEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEAAAAA
NS500138:372:H7K35BGXC:1:11105:4862:1314     163    chr2L   83398   42    36M    =     83407   43      AGTATATAGGAACACGCTACCCGAAGAACGGAGAGT    AAAAAEEEEEAEEEEAEEEEEEEEEEE/EEEEEEEEE
NS500138:372:H7K35BGXC:1:11105:4862:1314     83     chr2L   83407   42    34M    =     83398   -43     GAACACGCTACCCGAAGAACGGAGAGTTTGGAAA      EEEEEEAEEEEEEEEEEEEEEEEEEE6EEAAAAAA
```

rnext pnext tlen

- This paired-end example fills in more columns. The flags are also more complex.
- 147 = read paired, properly, reverse strand, second read.
- 163 = read paired, properly, plus strand, second read.
- 83 = read paired, properly, minus strand, first read



Binary (Decimal)	Hex	Description
00000000001 (1)	0x1	Is the read paired?
00000000010 (2)	0x2	Are both reads in a pair mapped “properly” (i.e., in the correct orientation with respect to one another)?
00000000100 (4)	0x4	Is the read itself unmapped?
00000001000 (8)	0x8	Is the mate read unmapped?
00000010000 (16)	0x10	Has the read been mapped to the reverse strand?
00000100000 (32)	0x20	Has the mate read been mapped to the reverse strand?
00001000000 (64)	0x40	Is the read the first read in a pair?
00010000000 (128)	0x80	Is the read the second read in a pair?
00100000000 (256)	0x100	Is the alignment not primary? (A read with split matches may have multiple primary alignment records.)
01000000000 (512)	0x200	Does the read fail platform/vendor quality checks?
10000000000 (1024)	0x400	Is the read a PCR or optical duplicate?

Next week, we will learn how to import these data to R

These features (flag, mapq) can be used to filter out bad stuff.

- First, we need to learn how to generate .bam files from .fastq data.
  - Lots of work to set up the first time… trial and error… once you have the data, you won’t think about this part very often.
  - But you need to be systematic and the mapping needs to be reproducible.

Fast-forward to the day you submit your paper

Public Database Submission

- Every published genomics dataset must be deposited in a public repository (e.g., Gene Expression Omnibus/Sequence Read Archive).

- What does GEO/SRA want?
  - Original FastQ files (trimmed reads might be OK), gzipped.
  - All pertinent information about how you processed the files
  - None of the .bam files.

Fast-forward to the day you submit your paper

Materials and Methods

- We all know there are good and bad Materials and Methods sections in papers. **You should all commit to writing really good ones.**

- Your analysis should be perfectly reproducible. Because GEO doesn't take .bam files, and this is the centerpiece of an analysis, you also need to document exactly how you got your reads mapped.

- It is also good (and sometimes necessary) to enumerate how many reads you started with, what the mapping rate was, what the duplication rate was, and so on…

Reproducible genomics begins with reproducible code

- I have made two ‘test scripts’ that will map a test dataset *if* the computing environment is set up properly.
- It also highlights one possible way to set up a single script that you could use for all your mapping needs for either paired- or single-end data

Reproducible approaches require reproducible directory structures.

- My suggested directory structure for one sequencing run is:

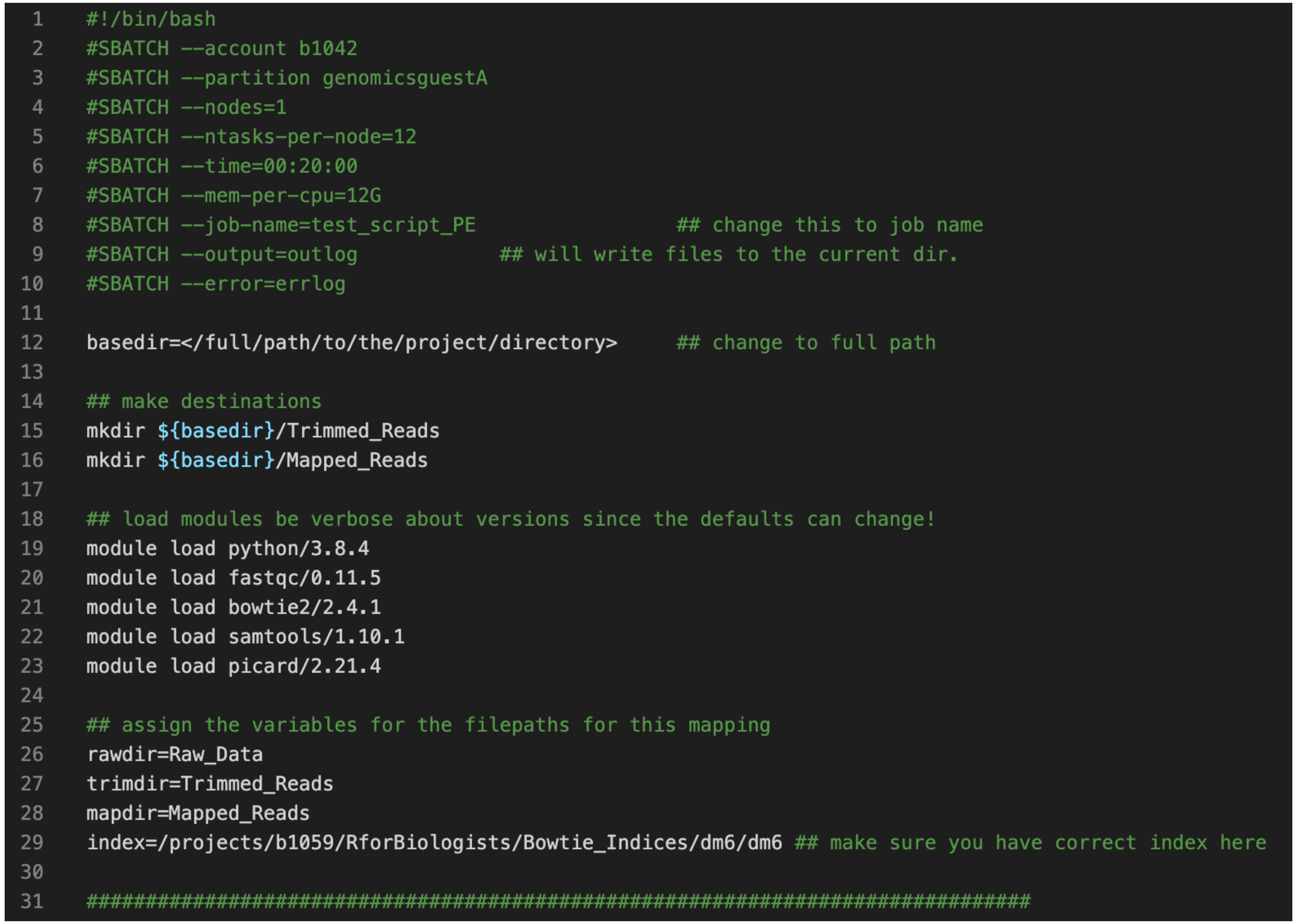
```
>base_directory (e.g., 211018_pilot_experiment)

        >Raw_Data
        >Trimmed_Reads
        >Mapped_Reads
        - outlog
        - errlog
        - mapping_script.sh (not the master version, but the version that you use for this particular mapping)
```

The Test Script

- I have created test scripts for both paired- and single-end data. They are in the “Quest” directory in the class Github Repo.

- I have created an annotated markdown that works through the paired-end script line by line.

- I have also included documentation of how I set up the class Quest allocation to run this code.



```
#!/bin/bash
#SBATCH --account b1042
#SBATCH --partition genomicsguestA
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=12
#SBATCH --time=00:20:00
#SBATCH --mem-per-cpu=12G
#SBATCH --job-name=test_script_PE                 ## change this to job name
#SBATCH --output=outlog            ## will write files to the current dir.
#SBATCH --error=errlog

basedir=</full/path/to/the/project/directory>     ## change to full path

## make destinations
mkdir ${basedir}/Trimmed_Reads
mkdir ${basedir}/Mapped_Reads

## load modules be verbose about versions since the defaults can change!
module load python/3.8.4
module load fastqc/0.11.5
module load bowtie2/2.4.1
module load samtools/1.10.1
module load picard/2.21.4

## assign the variables for the filepaths for this mapping
rawdir=Raw_Data
trimdir=Trimmed_Reads
mapdir=Mapped_Reads
index=/projects/b1059/RforBiologists/Bowtie_Indices/dm6/dm6 ## make sure you have correct index here

################################################################################
```

Goals/Plans

- **Today, Monday October 18:** Set up directory structures for running the test scripts. Upload data to your Quest folder. Modify the test script(s) to run correctly. See if you can get them to run.

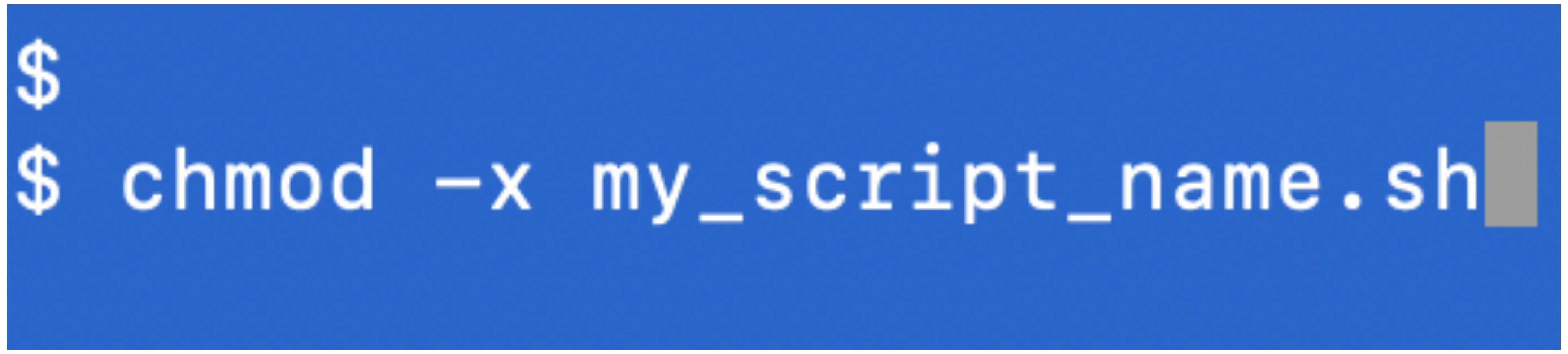
- **By next Monday**: Upload your own raw fastQ data to project directories you create in your Quest folder. Modify the test script appropriately, and map your data to the appropriate reference genome.

- **Next class (Thursday) will be entirely project-time, figuring out how to do the mapping. I will be available for consultations.**

Big hints:

- Please read the script annotation document as well as the comments I’ve added to the script itself.

- When you initially upload the script to Quest, **it will not be executable.** This is because the correct permission has not been set. This is a frequent thing to encounter. **You fix this by running the following bash command**:



```
$
$ chmod -x my_script_name.sh
```

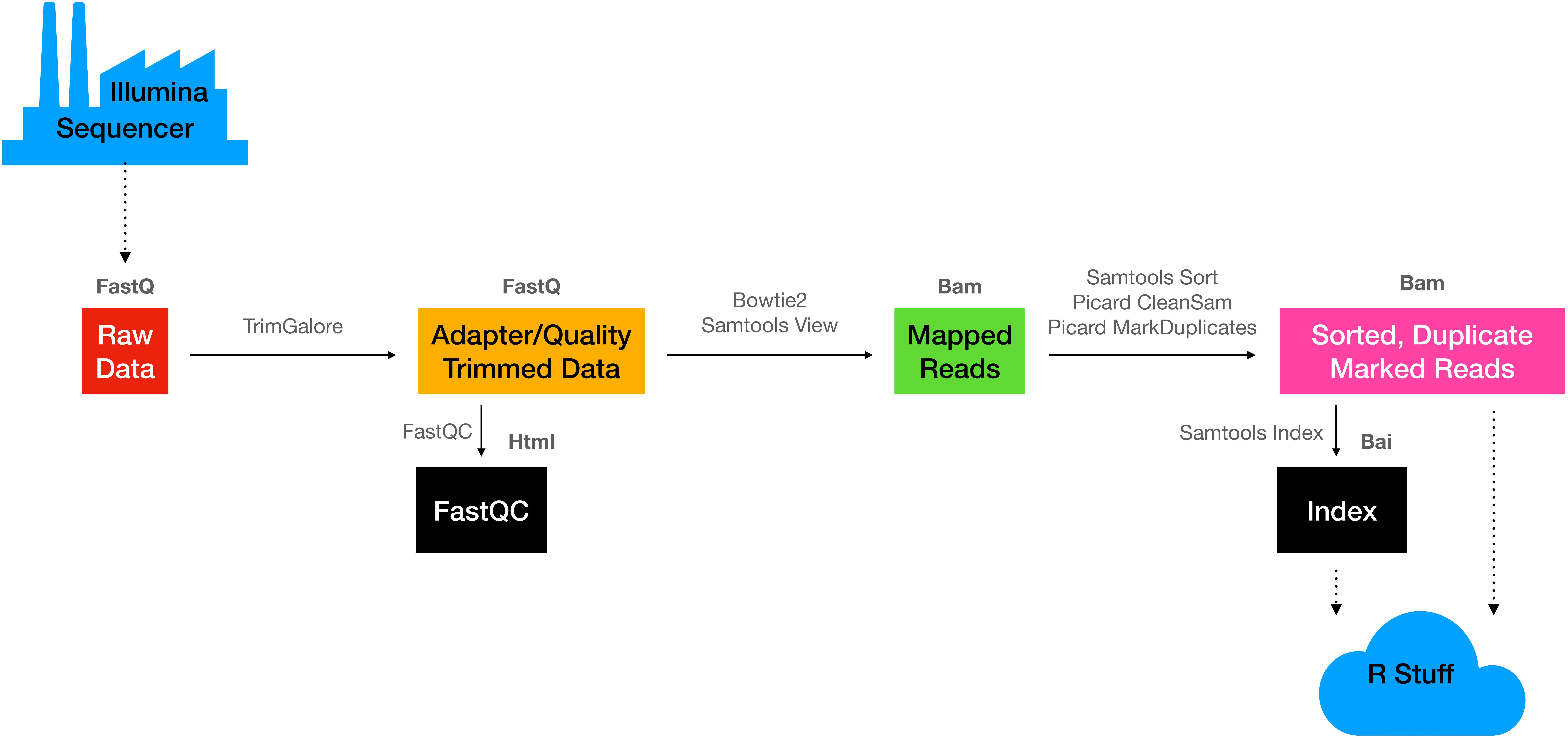
- (On the assumption that you are *in the directory that contains the script* and that the name of the script is “my_script_name.sh”).

Take the time now to find the scripts and download them to your computer.

Open them in your code-aware text editor.

I will then give you an overview of the workflow.

General Workflow



Illumina Sequencer

FastQ

Raw Data

TrimGalore

FastQ

Adapter/Quality Trimmed Data

FastQC

Html

FastQC

Bowtie2
Samtools View

Bam

Mapped Reads

Samtools Sort
Picard CleanSam
Picard MarkDuplicates

Bam

Sorted, Duplicate Marked Reads

Samtools Index

Bai

Index

R Stuff