BIOL 343

Applied Bioinformatics I

FASTQ files

Dr. Nic Wheeler

Stages of sequencing data analysis

Primary, secondary, and tertiary

Primary

Occurs on the machine in real-time

Provides base calls and quality scores

Illumina Connected Analytics

Secondary

Demultiplexing and compression (DRAGEN)

Alignment

Assembly

Specialized tools selected by the end user

Tertiary

Questions of biological relevance

Differential expression, variant calling, etc

Specialized tools selected by the end user

Stages of sequencing data analysis

Primary, secondary, and tertiary

Primary

Occurs on the machine in real-time

Provides base calls and quality scores

Illumina Connected Analytics

Secondary

Demultiplexing and compression (DRAGEN)

Alignment

Assembly

Specialized tools selected by the end user

Tertiary

Questions of biological relevance

Differential expression, variant calling, etc

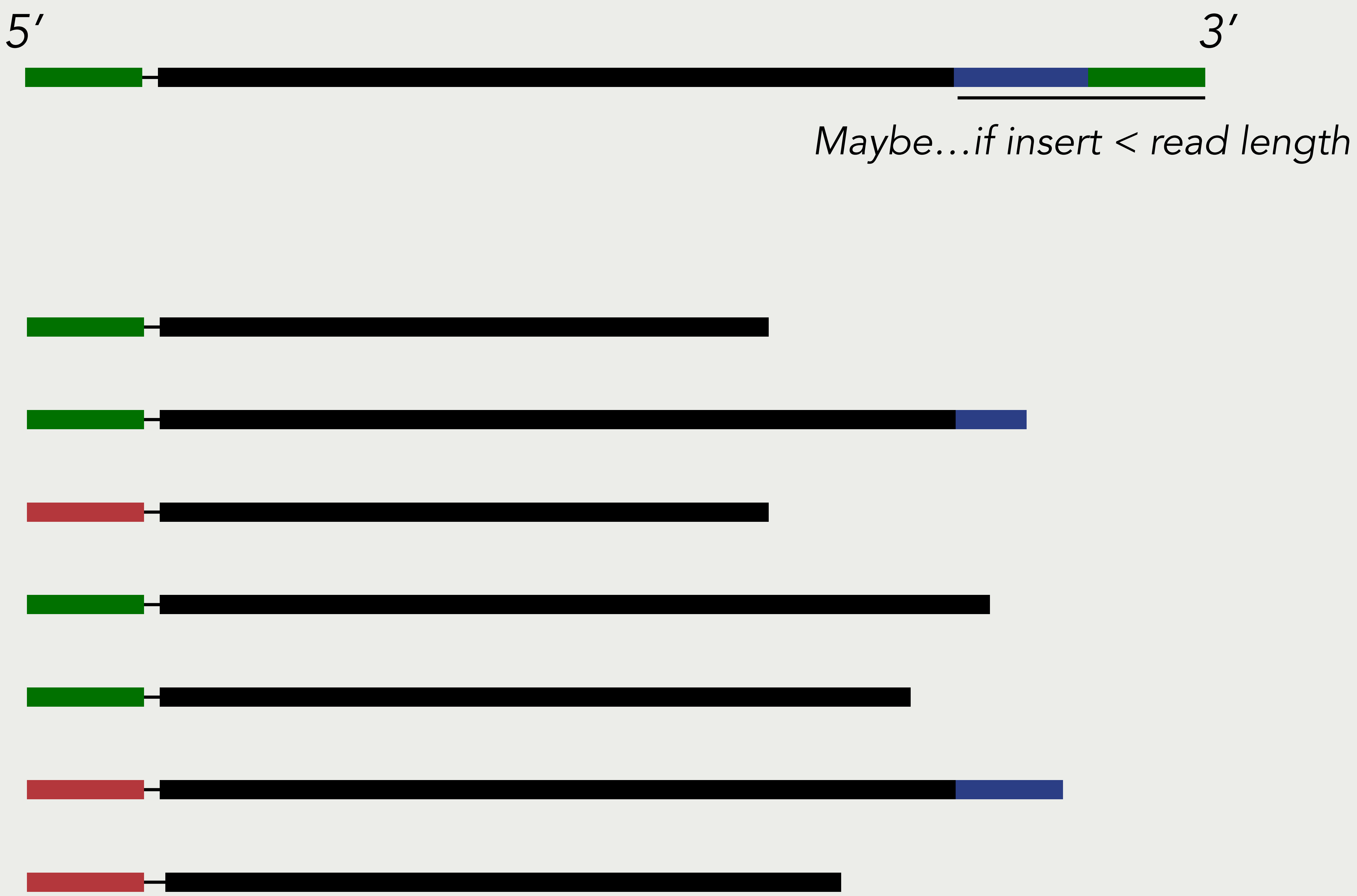
Specialized tools selected by the end user

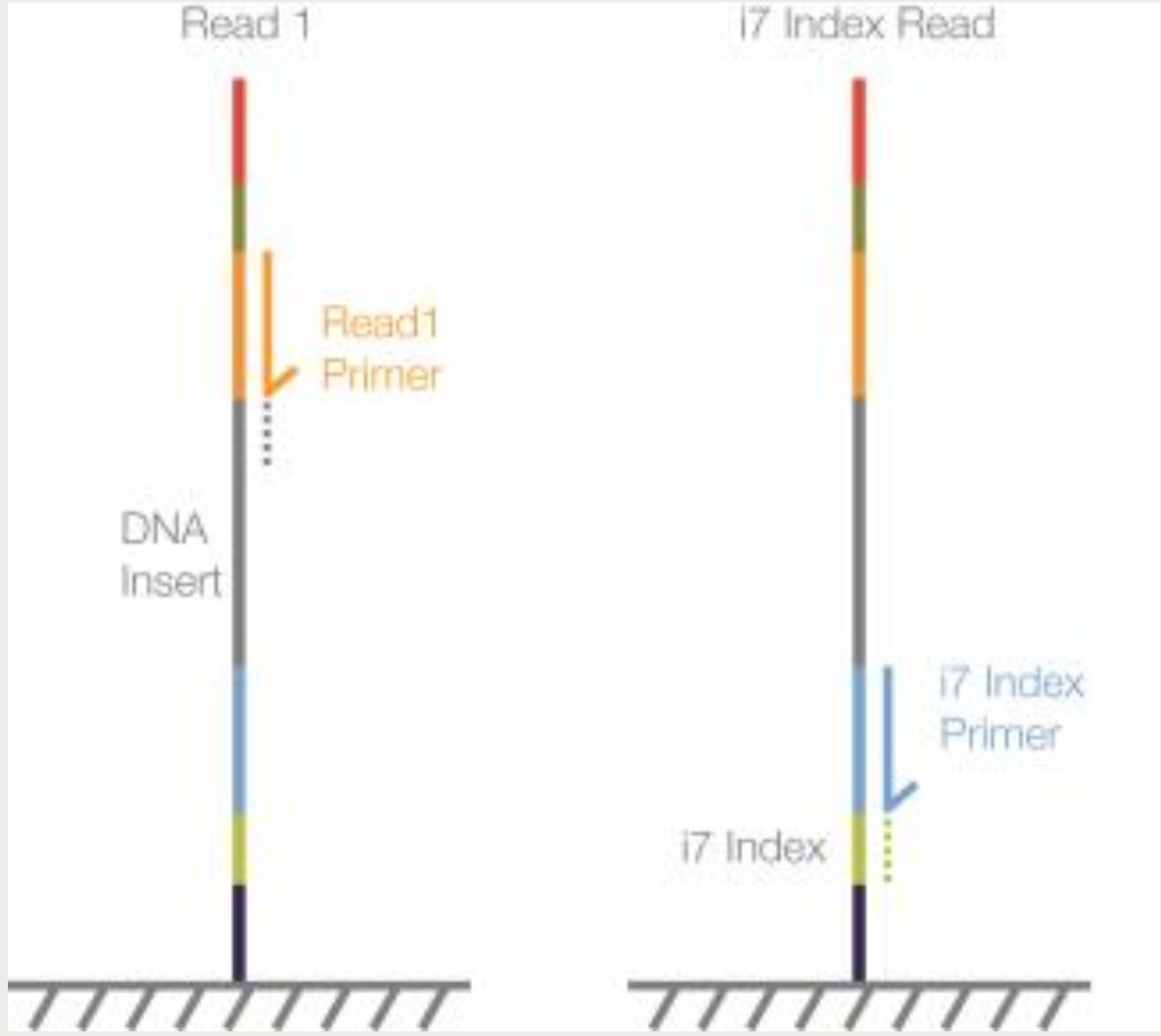
Produces FASTQ files

Stages of sequencing data analysis

Primary analysis produces FASTQ data and demultiplexes it

Read 1

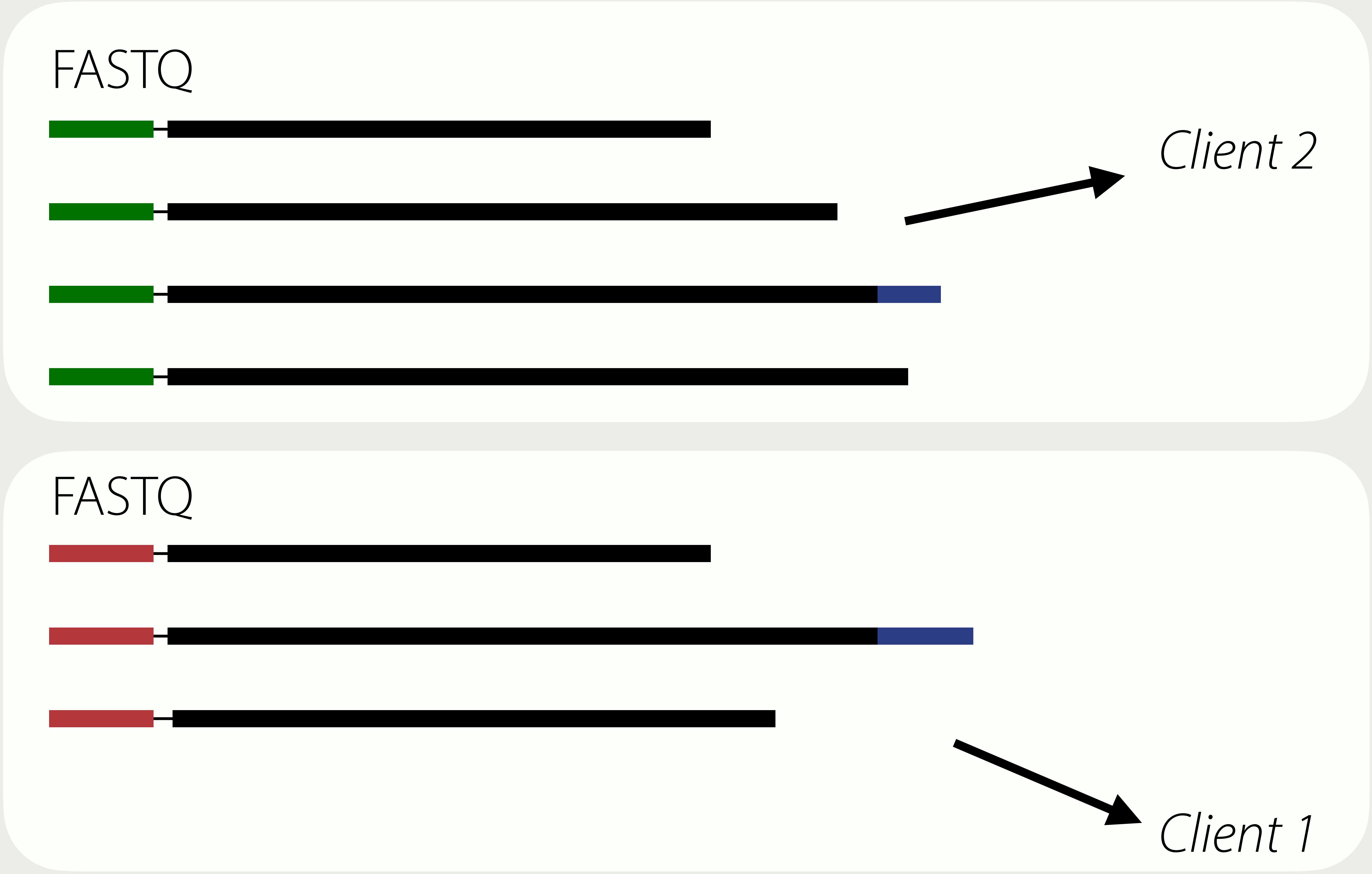




Stages of sequencing data analysis

Primary analysis produces FASTQ data and demultiplexes it

Read 1

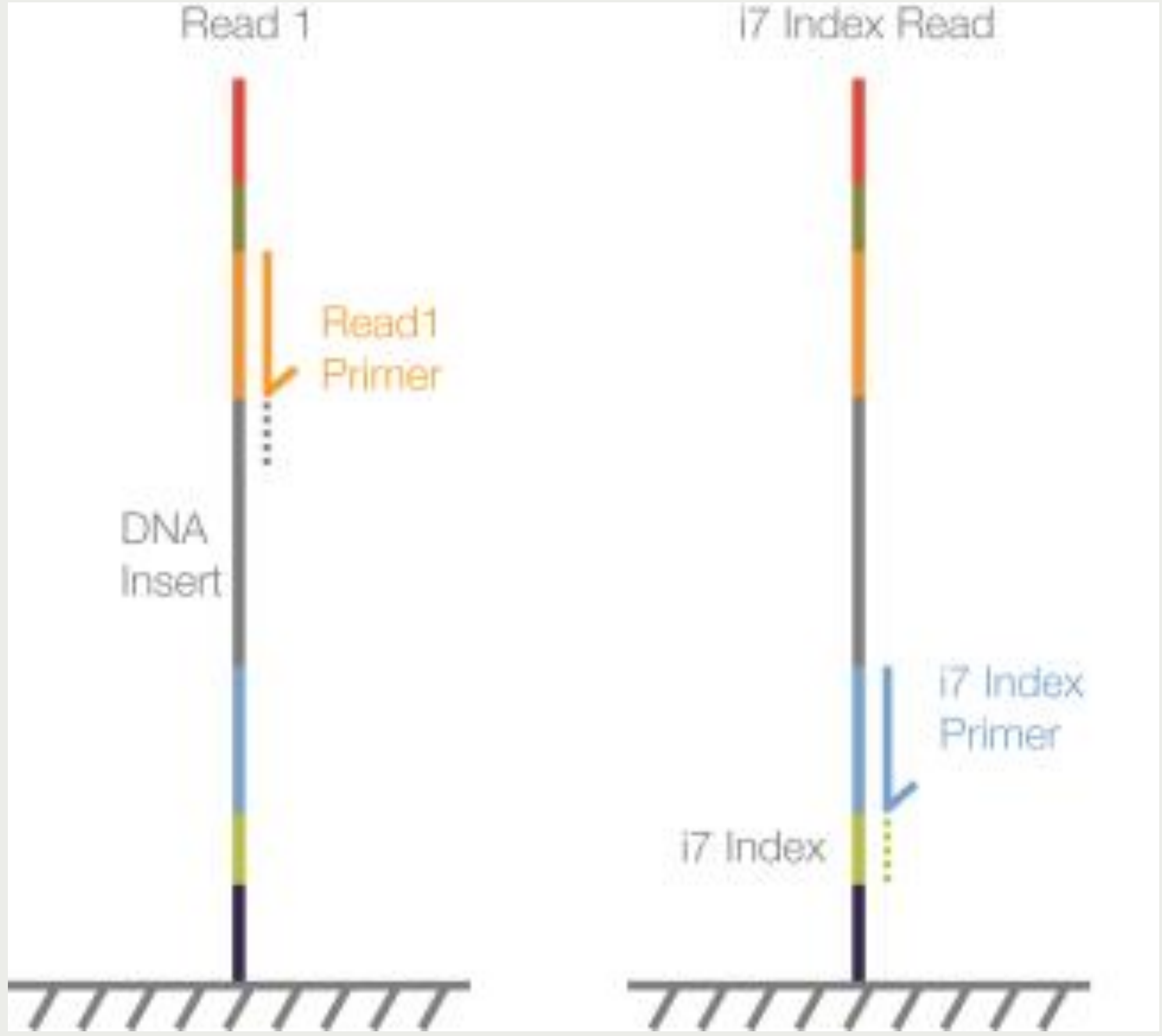


FASTQ

Client 2

FASTQ

Client 1



Types of sequence storage files

FASTA and FASTQ

FASTA

- Used for storing generic primary sequence data
- Nucleotide or amino acids (ASCII format)
- Usually KB to MB in size
- Header defined by >
  - Sequence name and attributes
  - Spaces discouraged
- Sequence in the line immediately after
  - Can have new line characters (\n)

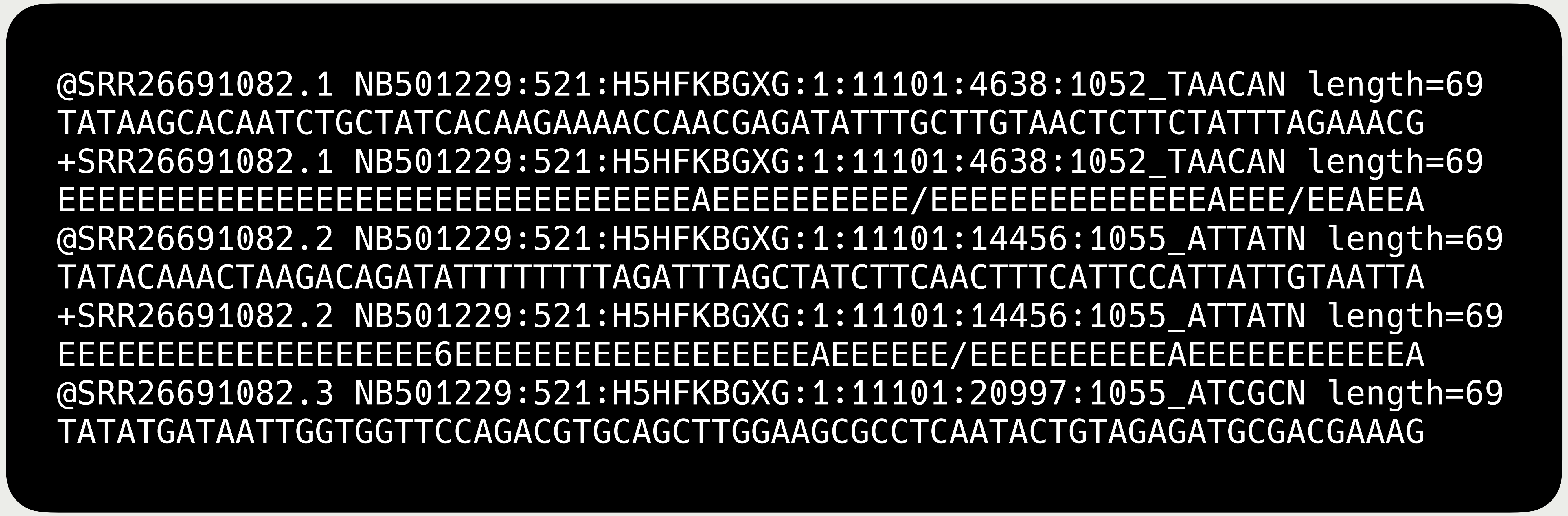
```
>Smp_104210.1 RHO cdna:protein_coding
TACTTCATTTATCATTCTGGTAAGTAATGAGTTAACTAAA
TCTGTTCATTTGTTTCACTAAATTTTAAATCAGAAATTTC
TTTTTTTTAACACTATTTCTAAACTGTTAAATGCACATTT
ATTTTTCAATTTCGTTTAATATCTAGTAGAGTAATCAGTC
TATGTTATTTTAATGAGAATCCTCATTAAAATACATTTCA
GATACTTGTTGAGTTCAATTGAAAAACATTCTCAGAAGGG
GTTTTGTGGAGATTTCAGTATTTTCATAGTTGAAATCATG
AGTCATTTGAAGCTAAACCCCCATGGAAAACCTAGGAGCA
ATGGACGGCCGTCTCGTTGTATTGTGAGACTCCTCAGCAG
TACCCATCCACGATCCCGCCTCGTGAGATTCGAACCCAGG
ATCTACCAGTCTCGCGCCAGAGCGCTTAACCACTAGATAT
TCTTTTATTATGTTAGGAAGTAATAAAAGTTTTCTTGAG
```

Types of sequence storage files

FASTA and FASTQ

FASTQ

- Used for storing sequencing data and quality
- ASCII format
- Usually MB to GB in size
- Four lines per sequence:
    1. Header (@ instead of >)
    2. Sequence
    3. + (sometimes the seq id)
    4. Phred quality scores
- Regex for the block: `@<seqname>\n<seq>\n+[<seqname>]\n<qual>\n`



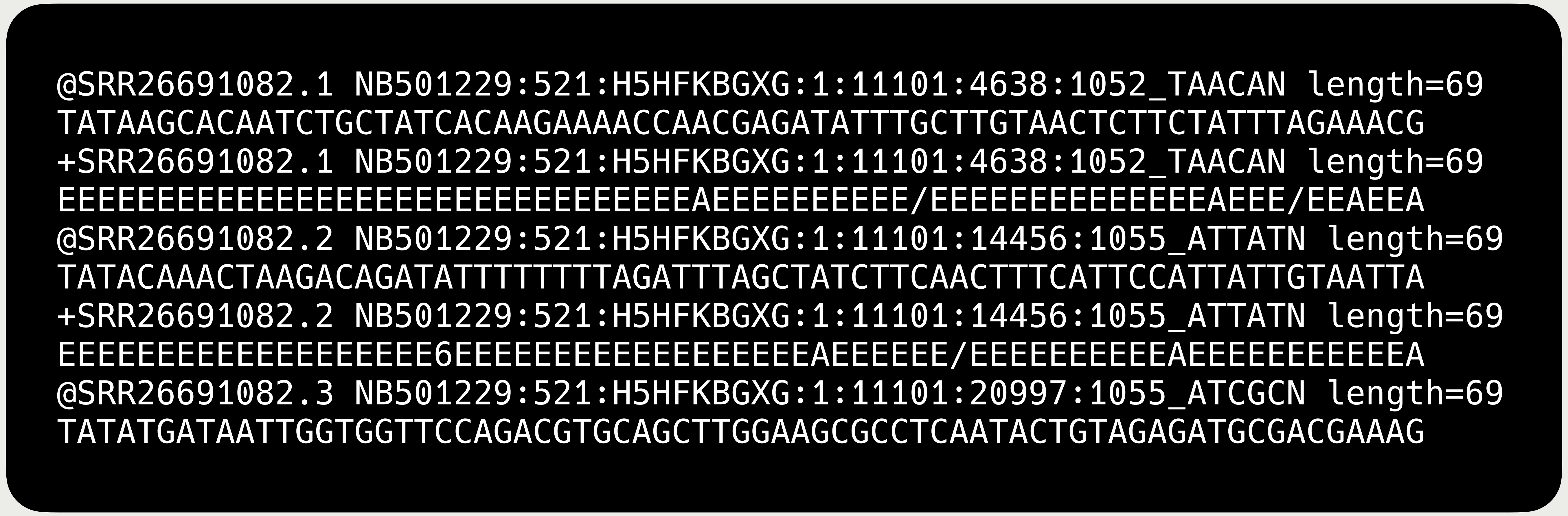
```
@SRR26691082.1 NB501229:521:H5HFKBGXG:1:11101:4638:1052_TAACAN length=69
TATAAGCACAATCTGCTATCACAAGAAAACCAACGAGATATTTGCTTGTAACTCTTCTATTTAGAAACG
+SRR26691082.1 NB501229:521:H5HFKBGXG:1:11101:4638:1052_TAACAN length=69
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEE/EEEEEEEEEEEEEEEAEEE/EEAEEA
@SRR26691082.2 NB501229:521:H5HFKBGXG:1:11101:14456:1055_ATTATN length=69
TATACAAACTAAGACAGATATTTTTTTTAGATTTAGCTATCTTCAACTTTCATTCCATTATTGTAATTA
+SRR26691082.2 NB501229:521:H5HFKBGXG:1:11101:14456:1055_ATTATN length=69
EEEEEEEEEEEEEEEEEE6EEEEEEEEEEEEEEEEEEEAEEEEEE/EEEEEEEEEAEEEEEEEEEEEEA
@SRR26691082.3 NB501229:521:H5HFKBGXG:1:11101:20997:1055_ATCGCN length=69
TATATGATAATTGGTGGTTCCAGACGTGCAGCTTGGAAGCGCCTCAATACTGTAGAGATGCGACGAAAG
```

Types of sequence storage files

FASTA and FASTQ

FASTQ Header

- @SRR26691082.1 - SRR ID and read #
- NB501229 - instrument name
- 521 - Run ID
- H5HFKBGXG - flow cell ID
- 1 - flow cell lane
- 11101 - flow cell tile #
- 4638 - x coordinate in the tile
- 1052 - y coordinate in the tile
- TAACAN - index sequence



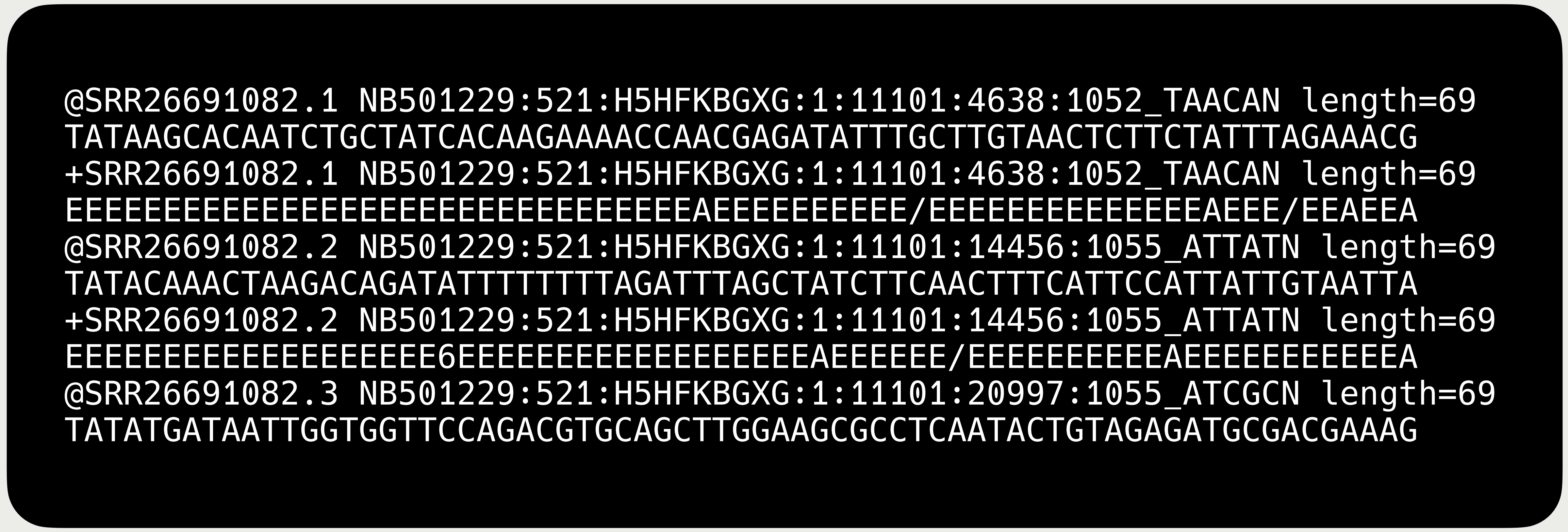
```
@SRR26691082.1 NB501229:521:H5HFKBGXG:1:11101:4638:1052_TAACAN length=69
TATAAGCACAATCTGCTATCACAAGAAAACCAACGAGATATTTGCTTGTAACTCTTCTATTTAGAAACG
+SRR26691082.1 NB501229:521:H5HFKBGXG:1:11101:4638:1052_TAACAN length=69
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEE/EEEEEEEEEEEEEEAEEE/EEAEEA
@SRR26691082.2 NB501229:521:H5HFKBGXG:1:11101:14456:1055_ATTATN length=69
TATACAAACTAAGACAGATATTTTTTTTAGATTTAGCTATCTTCAACTTTCATTCCATTATTGTAATTA
+SRR26691082.2 NB501229:521:H5HFKBGXG:1:11101:14456:1055_ATTATN length=69
EEEEEEEEEEEEEEEEEEE6EEEEEEEEEEEEEEEEEAEEEEEE/EEEEEEEEEEEAEEEEEEEEEEEA
@SRR26691082.3 NB501229:521:H5HFKBGXG:1:11101:20997:1055_ATCGCN length=69
TATATGATAATTGGTGGTTCCAGACGTGCAGCTTGGAAGCGCCTCAATACTGTAGAGATGCGACGAAAG
```

Types of sequence storage files

FASTA and FASTQ

Phred quality

- String of the same length as <seq>
- Each character represents the Phred quality of the corresponding nt
- Represents the likelihood that the base call was correct



```
@SRR26691082.1 NB501229:521:H5HFKBGXG:1:11101:4638:1052_TAACAN length=69
TATAAGCACAATCTGCTATCACAAGAAAACCAACGAGATATTTGCTTGTAACTCTTCTATTTAGAAACG
+SRR26691082.1 NB501229:521:H5HFKBGXG:1:11101:4638:1052_TAACAN length=69
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEE/EEEEEEEEEEEEEEAEEE/EEAEEA
@SRR26691082.2 NB501229:521:H5HFKBGXG:1:11101:14456:1055_ATTATN length=69
TATACAAACTAAGACAGATATTTTTTTTAGATTTAGCTATCTTCAACTTTCATTCCATTATTGTAATTA
+SRR26691082.2 NB501229:521:H5HFKBGXG:1:11101:14456:1055_ATTATN length=69
EEEEEEEEEEEEEEEEEEE6EEEEEEEEEEEEEEEEEEAEEEEEE/EEEEEEEEEEAEEEEEEEEEEEA
@SRR26691082.3 NB501229:521:H5HFKBGXG:1:11101:20997:1055_ATCGCN length=69
TATATGATAATTGGTGGTTCCAGACGTGCAGCTTGGAAGCGCCTCAATACTGTAGAGATGCGACGAAAG
```

```
$Q = -10log₁₀($e)
```

Where $e is the error probability

- If the quality of a base call is 30, the probability that it is wrong is 0.001. In other words, given 1000 base calls with Q=30, one of them is wrong in average.
- Minimum of 0 (!) and maximum of 42 (K)

Phred+33 encoding

- $Q + 33, take the corresponding ASCII character
- Use this chart to get $Q for E, /, and A

Types of sequence storage files

FASTA and FASTQ

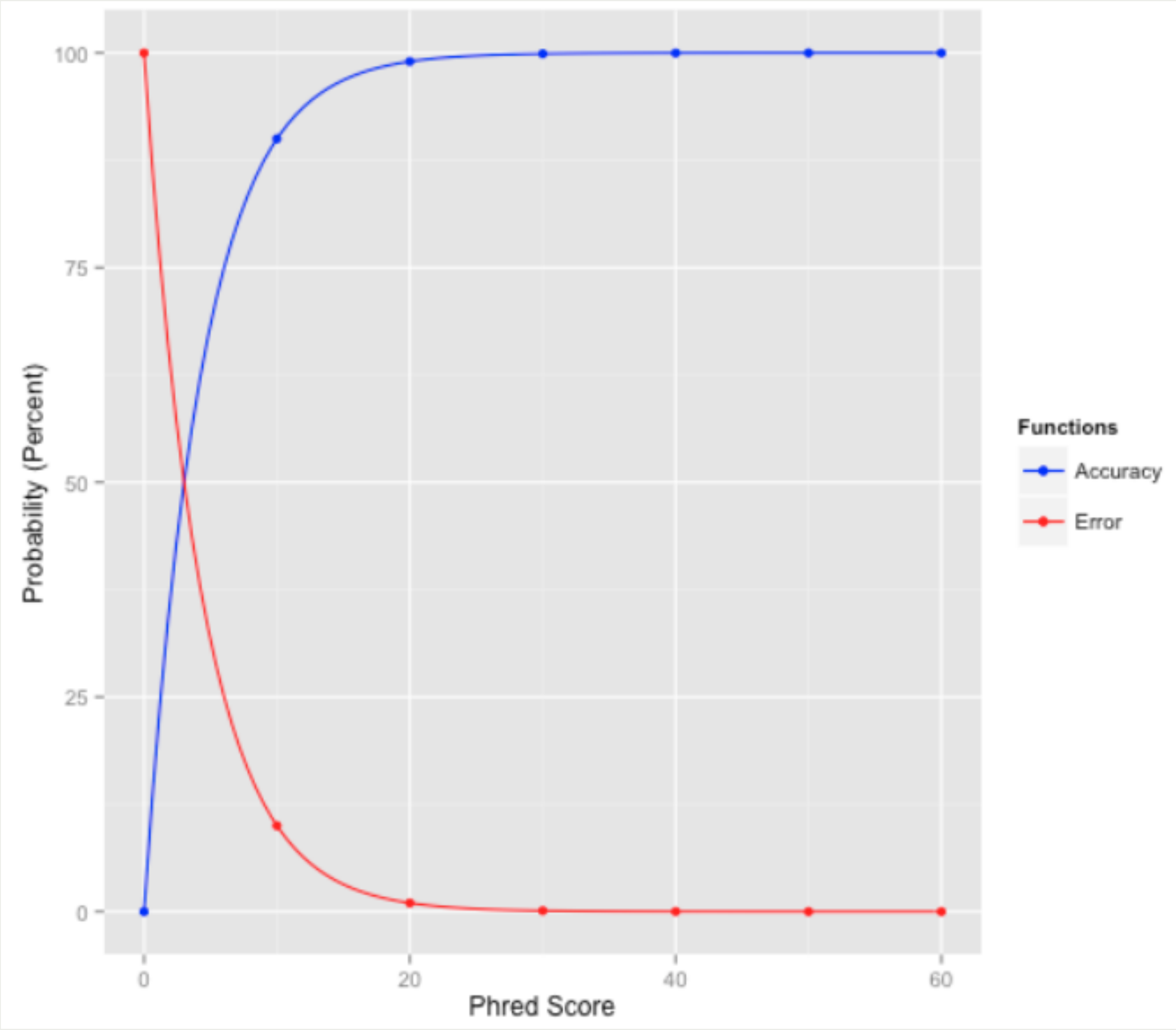
Phred quality

- String of the same length as <seq>
- Each character represents the Phred quality of the corresponding nt
- Represents the likelihood that the base call was correct

$$\$Q = -10\log_{10}(\$e)$$

Where $e is the error probability

- If the quality of a base call is 30, the probability that it is wrong is 0.001. In other words, given 1000 base calls with Q=30, one of them is wrong in average.
- Minimum of 0 (!) and maximum of 42 (K)



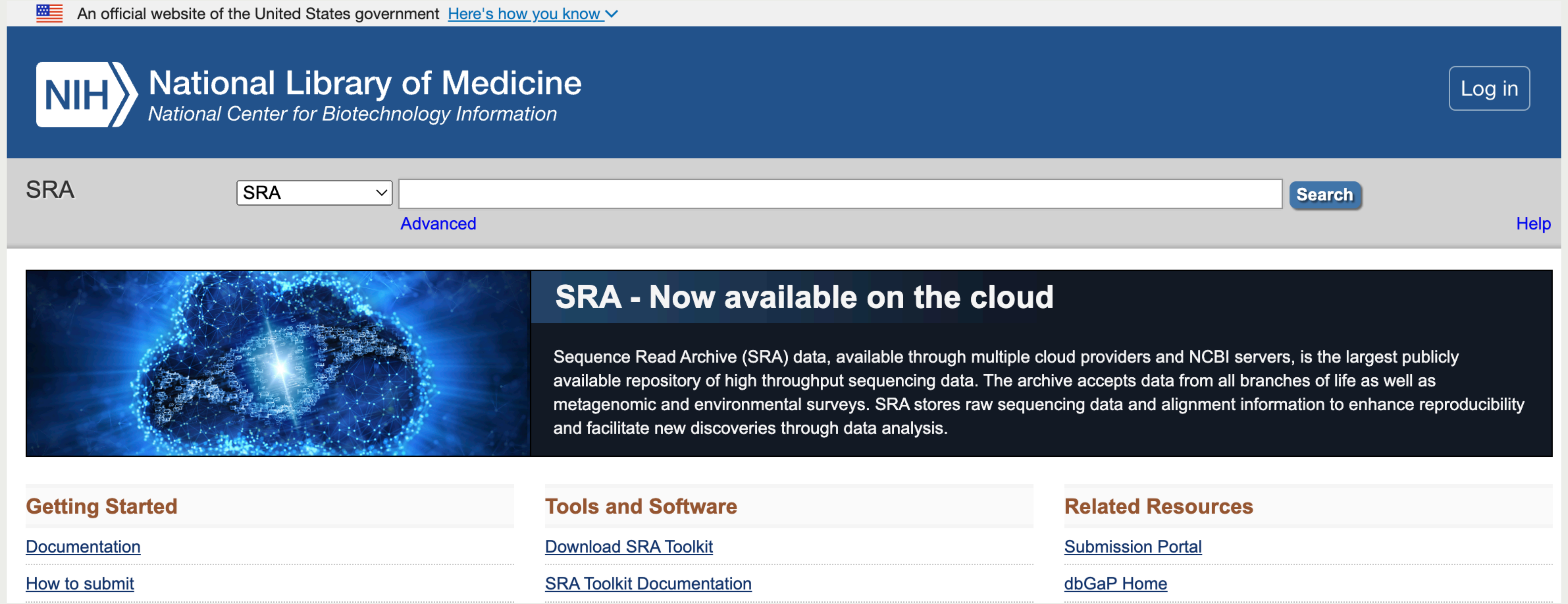
https://gatk.broadinstitute.org/hc/en-us/articles/360035531872-Phred

FASTQ databases

NCBI SRA & ENA

NCBI SRA

- https://www.ncbi.nlm.nih.gov/sra
- Sequence Read Archive
- Free database of FASTQ (or BAM) files
- Most public sequencing data can be found on SRA
  - Europeans post to ENA
- Dedicated software: SRA Toolkit (`conda install bioconda::sra-tools`)
  - Already installed in `biol343` environment


An official website of the United States government Here's how you know
NIH
National Library of Medicine
National Center for Biotechnology Information
Log in
SRA
SRA
Search
Advanced
Help

SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

Getting Started
- Documentation
- How to submit

Tools and Software
- Download SRA Toolkit
- SRA Toolkit Documentation

Related Resources
- Submission Portal
- dbGaP Home