# Bioinformatics file formats

Jessen V. Bredeson
UC Berkeley

# Goals and outline

- Understand importance of standardized file formats

- Introduce you to commonly-used formats in bioinformatics

- Resources for manipulating or parsing them yourself

# Why are (standardized) file formats important?

**Syntactic and semantic interoperability**
"The capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units"[1]

"the capability of different programs to exchange data via a common set of exchange formats, to read and write the same file formats, and to use the same protocols.... the lack of interoperability can be a consequence of a lack of attention to standardization during the design of a program"[2]
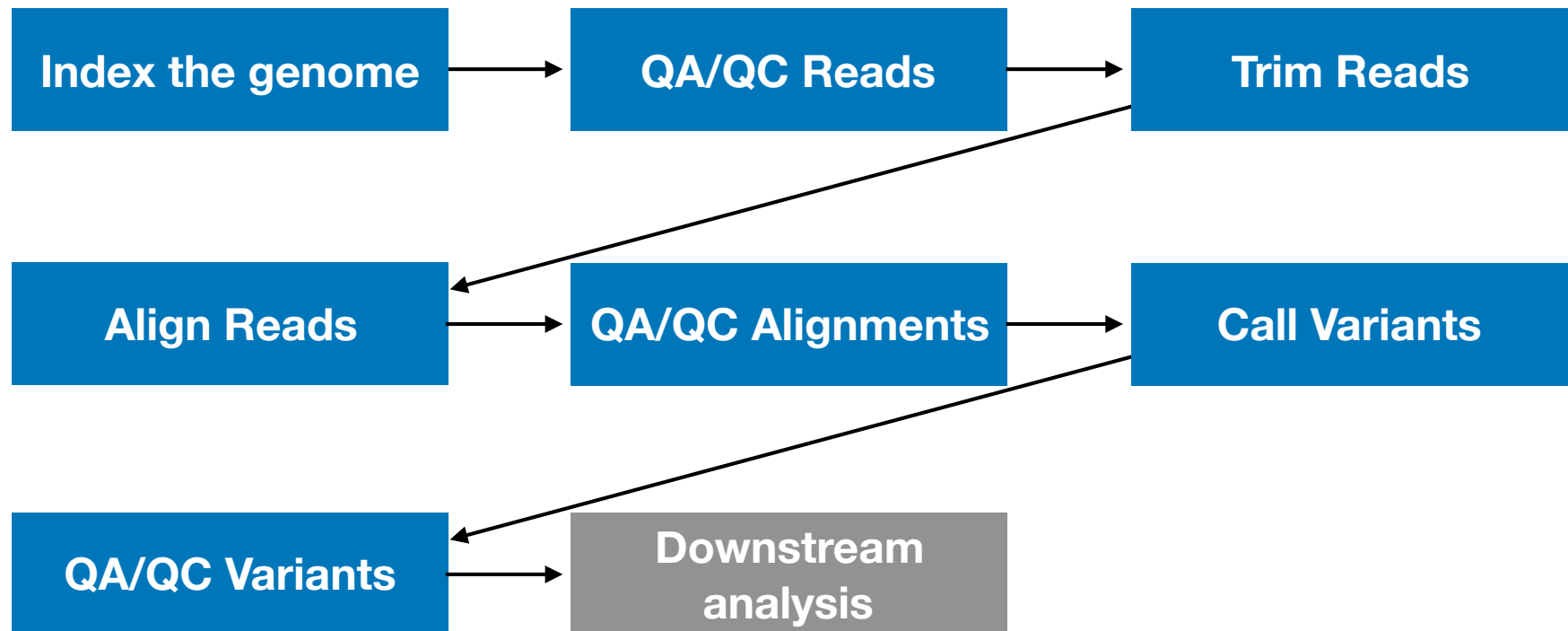
**Data sharing and collaboration**
File standards provide:
a common language for data sharing,
promote collaboration,
ensure data reusability,
reduce semantic and syntactic errors
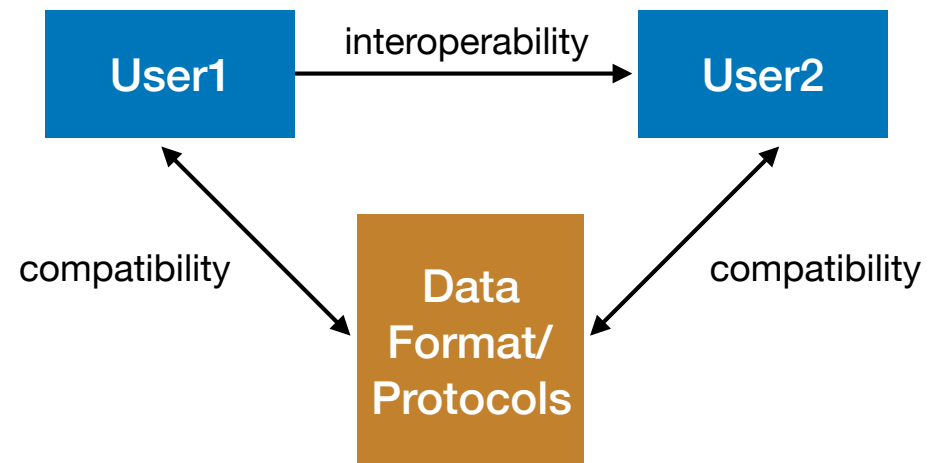


1. ISO/IEC 2382-01 *Information Technology Vocabulary, Fundamental Terms*
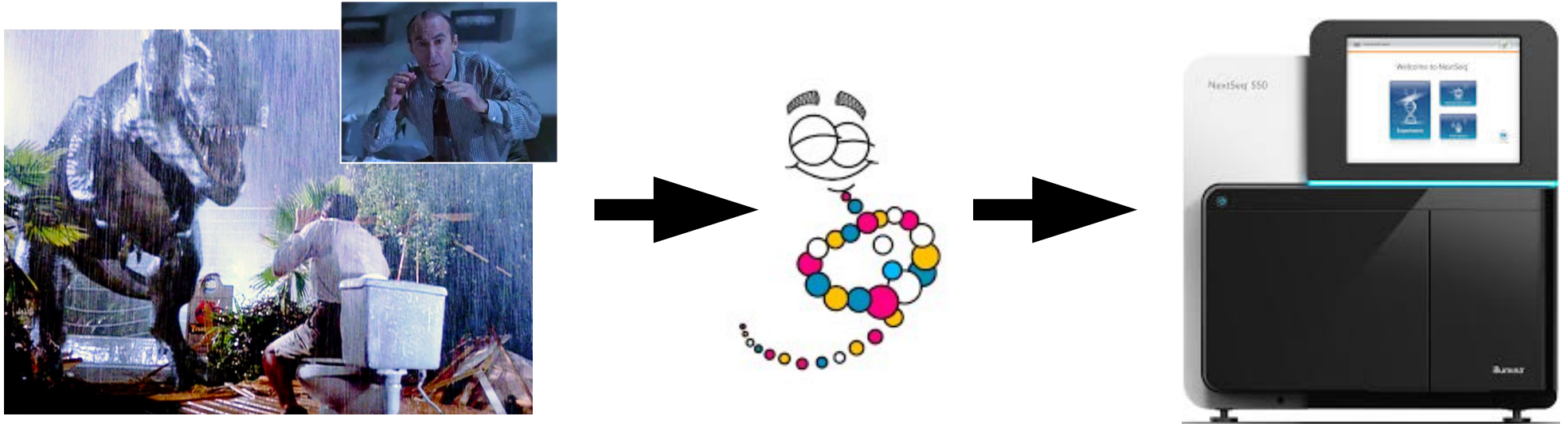2. Gordon and Hernandez, *The Official Guide to the SSCP Book*

# Variant-calling workflow

# We have a specimen of interest...

# We extract DNA...

# We have data, now what?

```
$ ls -1
Trex_genome.fasta
Trex_genome.annotation.gff3
SAMPLE_NoIndex_R1_001.fastq.gz
SAMPLE_NoIndex_R2_001.fastq.gz

$ head -3 Trex_genome.fasta
>Chr1
ATGCGTACGTATTCTGCGCTATCGCTATCAGCTATCTAGCTACGATCGATCAGCTATCGTACGATCACT
ACTAGCTGACTGATCGATATCTCTTCGAGATTCTCTCTCTCTTTTTTTTTCGATACGACGTACGATGCG

$ gunzip -c SAMPLE_NoIndex_R1_001.fastq.gz | head -4
@HS3:SAMPLE:1:1:1
ATCGCGAGCTTACGACTACGATCGATCGATTCGGCTACTATCGATCGTAGCTACGTAGTCTCGCGTATC
+
AAFFFJFJJFJJJJJJJJJJJJJJJAJAFFJJJJFJJAJ<<JJJJJJJFFJJFJ700+)'&%#####
```

# Index the genome

```
$ bwa index Trex_genome.fasta  # creates the following files:
$ ls -1 Trex_genome.fasta*
Trex_genome.fasta
Trex_genome.fasta.bwt
Trex_genome.fasta.pac
Trex_genome.fasta.ann
Trex_genome.fasta.amb
Trex_genome.fasta.sa

$ samtools faidx Trex_genome.fasta  # creates Trex_genome.fasta.fai
$ head -3 Trex_genome.fasta.fai
Chr1  217471166  141         100  101
Chr2  181034961  219646160   100  101
Chr3  153873357  402491612   100  101

$ samtools dict Trex_genome.fasta >Trex_genome.dict
$ head -3 Trex_genome.dict
@HD  VN:1.0   SO:unsorted
@SQ  SN:Chr1  LN:217471166  M5:56d95ce6647ea9087b857b1efa6d00dd
@SQ  SN:Chr2  LN:181034961  M5:20852c561ea38c67aa67e6d655cfebf2
```

# FASTA/Pearson

https://en.wikipedia.org/wiki/FASTA_format

```
>U31202.1 Human noggin (NOGGIN) gene, complete cds
GAGCTCCGGCGGGTCAGCCGGACTGTCGGCTTCCCGGGGCATCTGGGTCCGGCGGGGCACAGCCCTGGGC
GCTGCCGAAGCCGCCGCCGCCGCCTCCGCGGCGAGTACAGGCGGCTTCCCCCGGAGCCTGTGCAGCTCCA
GCTCCTCGGGGGTGGAGAAGTGGGGGGTGGGGGTGATGTATGGGGGGAAGAAGGGGGAGGGGCCAACCCC
GAGAGAGTCAGTGGTTTCCATGGTGATGGAGCTGAAAGTGCAGGAAATTTAAAGGCTTGGACCCTGCGAG
ACAGACAAACCGGTGCCAACGTGCGCGGACGCCGCCGCCGCCGCCGCCGCTGGAGTCCGCCGGGCAGAGC
CGGCCGCGGAGCCCGGAGCAGGCGGAGGGAAGTGCCCCTAGAACCAGCTCAGCCAGCGGCGCTTGCACAG
AGCGGCCGGNCGAAGAGCAGCGAGAGGAGGAGGGGAGAGCGGCTCGTCCACGCGCCCTGCGCCGCCGCCG
GCCCGGGAAGGCAGCGAGGAGCCGGCGCCTCCCGCGCCCCGCGGTCGCCCTGGAGTAATTTCGGATGCCC
AGCCGCGGCCGCCTTCCCCAGTAGACCCGGGAGAGGAGTTGCGGCCAACTTGTGTGCCTTTCTTCCGCCC
CGGTGGGAGCCGGCGCTGCGCGAAGGGCTCTCCCGGCGGCTCATGCTGCCGGCCCTGCGCCTGCCCAGCC
TCGGGTGAGCCGCCTCCGGAGAGACGGGGGAGCGCGGCGGCGCCGCGGGCTCGGCGTGCTCTCCTCCGGG
GACGCGGGACGAAGCAGCAGCCCCGGGCGCGCGCCAGAGGCATGGAGCGCTGCCCCAGCCTAGGGGTCAC
CCTCTACGCCCTGGTGGTGGTCCTGGGGCTGCGGGCGACACCGGCCGGCGGCCAGCACTATCTCCACATC
CGCCCGGCACCCAGCGACAACCTGCCCCTGGTGGACCTCATCGAACACCCAGACCCTATCTTTGACCCCA
AGGAAAAGGATCTGAACGAGACGCTGCTGCGCTCGCTGCTCGGGGGCCACTACGACCCAGGCTTCATGGC
CACCTCGCCCCCCGAGGACCGGCCCGGCGGGGGCGGGGGTGCAGCTGGGGGCGCGGAGGACCTGGCGGAG
CTGGACCAGCTGCTGCGGCAGCGGCCGTCGGGGGGCCATGCCGAGCGAGATCAAAGGGCTAGAGTTCTCCG
AGGGCTTGGCCCAGGGCAAGAAGCAGCGCCTAAGCAAGAAGCTGCGGAGGAAGTTACAGATGTGGCTGTG
GTCGCAGACATTCTGCCCCGTGCTGTACGCGTGGAACGACCTGGGCAGCCGCTTTTGGCCGCGCTACGTG
AAGGTGGGCAGCTGCTTCAGTAAGCGCTCGTGCTCCGTGCCCGAGGGCATGGTGTGCAAGCCGTCCAAGT
CCGTGCACCTCACGGTGCTGCGGTGGCGCTGTCAGCGGCGCGGGGGCCAGCGCTGCGGCTGGATTCCCAT
CCAGTACCCCATCATTTCCGAGTGCAAGTGCTCGTGCTAGAACTCGGGGGCCCCCTGCCCGCACCCGGAC
ACTTGATCCTCGAGCTC
>lcl|BC064885.2_cds_AAH64885.1_1 [gene=mtpn] [protein=myotrophin] [protein_id=AAH64885.1]
ATGGGTGACAAGGAGTTCGTGTGGGCCATCAAGAACGGAGACCTGGATGCAGTGAAAGAATTCGTACTTG
GGGGCGAGGATGTGAACCGGACGCTGGATGGTGGAAGGAAACCTATGCACTACGCTGCCGACTGCGGGCA
GGATGAGGTCCTGGAGTTTCTTCTCTCGAAAGGAGCCAACATCAATGCTGCGGATAAACATGGCATCACC
CCCCTACTATCTGCCTGCTACGAGGGCCATCGCAAATGTGTCGAGTTGCTTTTTATCTAAGGGAGCCGACA
AGACGGTGAAGGGCCCAGACGGACTCAATGCTTTGGAATCTACAGACAACCAGGCTATCAAAGATTTGCT
CCATTAA
```

# FASTA/Pearson

**FASTA Defline:** Sequence ID + Description on same line, sequence string on the next

**Whitespace** only required if description present

**"greater than"**
Start of record

>U31202.1 Human noggin (NOGGIN) gene, complete cds

**Description/Comment**
Optional;
Free-form text

**Sequence ID**
Required;
Any printable
non-whitespace
characters:
[!-~]

```
GAGCTCCGGCGGGTCAGCCGGACTGTCGGCTTCCCGGGGCATCTGGGTCCGGCGGGGCACAGCCCTGGGC
GCTGCCGAAGCCGCCGCCGCCGCCTCCGCGGCGAGTACAGGCGGCTTCCCCCGGAGCCTGTGCAGCTCCA
GCTCCTCGGGGGTGGAGAAGTGGGGGGTGGGGGTGATGTATGGGGGGAAGAAGGGGGAGGGGCCAACCCC
GAGAGAGTCAGTGGTTTCCATGGTGATGGAGCTGAAAGTGCAGGAAATTTAAAGGCTTGGACCCTGCGAG
ACAGACAAACCGGTGCCAACGTGCGCGGACGCCGCCGCCGCCGCCGCTGGAGTCCGCCGGGCAGAGC
CGGCCGCGGAGCCCGGAGCAGGCGGAGGGAAGTGCCCCTAGAACCAGCTCAGCCAGCGGCGCTTGCACAG
AGCGGCCGGNCGAAGAGCAGCGAGAGGAGGAGGGGAGAGCGGCTCGTCCACGCGCCCTGCGCCGCCGCCG
GCCCGGGAAGGCAGCGAGGAGCCGGCGCCTCCCGCGCCCCGCGGTCGCCCTGGAGTAATTTCGGATGCCC
AGCCGCGGCCGCCTTCCCCAGTAGACCCGGGAGAGGAGTTGCGGCCAACTTGTGTGCCTTTCTTCCGCCC
CGGTGGGAGCCGGCGCTGCGCGAAGGGCTCTCCCGGCGGCTCATGCTGCCGGCCCTGCGCCTGCCCAGCC
TCGGGTGAGCCGCCTCCGGAGAGACGGGGGAGCGCGGCGGCGCCGCGGGCTCGGCGTGCTCTCCTCCGGG
GACGCGGGACGAAGCAGCAGCCCCGGGCGCGCGCCAGAGGCATGGAGCGCTGCCCCAGCCTAGGGGTCAC
CCTCTACGCCCTGGTGGTGGTCCTGGGGCTGCGGGCGACACCGGCCGGCGGCCAGCACTATCTCCACATC
CGCCCGGCACCCAGCGACAACCTGCCCCTGGTGGACCTCATCGAACACCCAGACCCTATCTTTGACCCCA
AGGAAAAGGATCTGAACGAGACGCTGCTGCGCTCGCTGCTCGGGGGCCACTACGACCCAGGCTTCATGGC
CACCTCGCCCCCCGAGGACCGGCCCGGCGGGGGCGGGGGTGCAGCTGGGGGCGCGGAGGACCTGGCGGAG
CTGGACCAGCTGCTGCGGCAGCGGCCGTCGGGGGCCATGCCGAGCGAGATCAAAGGGCTAGAGTTCTCCG
AGGGCTTGGCCCAGGGCAAGAAGCAGCGCCTAAGCAAGAAGCTGCGGAGGAAGTTACAGATGTGGCTGTG
GTCGCAGACATTCTGCCCCGTGCTGTACGCGTGGAACGACCTGGGCAGCCGCTTTTGGCCGCGCTACGTG
AAGGTGGGCAGCTGCTTCAGTAAGCGCTCGTGCTCCGTGCCCGAGGGCATGGTGTGCAAGCCGTCCAAGT
CCGTGCACCTCACGGTGCTGCGGTGGCGCTGTCAGCGGCGCGGGGGCCAGCGCTGCGGCTGGATTCCCAT
CCAGTACCCCATCATTTCCGAGTGCAAGTGCTCGTGCTAGAACTCGGGGGCCCCCTGCCCGCACCCGGAC
ACTTGATCCTCGAGCTC
```

**FASTA Body/
Sequence string**
Nucleotide,
amino acid,
IUPAC codes,
alignment
characters [-*]

May/may not be
wrapped flush

```
>lcl|BC064885.2_cds_AAH64885.1_1 [gene=mtpn] [protein=myotrophin] [protein_id=AAH64885.1]
ATGGGTGACAAGGAGTTCGTGTGGGCCATCAAGAACGGAGACCTGGATGCAGTGAAAGAATTCGTACTTG
GGGGCGAGGATGTGAACCGGACGCTGGATGGTGGAAGGAAACCTATGCACTACGCTGCCGACTGCGGGCA
GGATGAGGTCCTGGAGTTTCTTCTCTCGAAAGGAGCCAACATCAATGCTGCGGATAAACATGGCATCACC
CCCCTACTATCTGCCTGCTACGAGGGCCATCGCAAATGTGTCGAGTTGCTTTTTATCTAAGGGAGCCGACA
[faded text]
CCATTAA
```

FASTA files are *best* suffixed with ".fasta" or ".fa"; some tools *require* this.

# FASTQ

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

```
@SRR10178655.1 0:N:0:
GGATCTATGGCCATGTAGGGACCATCTGAAGGCAGATCAAAATTTCGCTGAGCAAATTTAGGGTCCGGGTTTGTT
+
AA<A<F--FF<-F-A7FAF-F---A<F---<FF-<F--7F-----<-A7F-A----7FJ<-FF--<J<-7-FFFJ
@SRR10178655.2 0:N:0:
ATAAAAAAAAATTAATAATCTATTCTTTATTTAAAACTAATTTTTAAATTAATTGGTTTTTGTGGAATGGTATTA
+
AAFFFJFJAJJJFFA<JF-7FJF<JJJJ--<<FJ-J<A7---<-FFJFJFJJAJ-F<F<<-F-7---7-<-<FFA
@SRR10178655.3 0:N:0:
CATTATATACGTCGCCACTCTTAATTTCCTTTTCCATAAGAGCGTATAATCTTGTAATACAATGTCTTCTCCAAC
+
AAAFFJJJJAFJJJJAFJF<JFAFJJJF-<FAJJFJ-F-F-<7-7-<FJJJJF<JA-FF---<-7<F-F<-7-<F
@SRR10178655.4 0:N:0:
AAGTTATTCTGCCTCTAATGCGATAACTGTAATCTTTAATTGTGTAATTTCTTTTTCACAATCTGAGCCACGCCA
+
AAAAAJF<-A<-7FJFJJJFJFJJJJJ<FJ--7<FF-7-<--7-A<7FFJAJFFJJJAJ7FF-F7FA-7<-A-7-
@SRR10178655.5 0:N:0:
ATATATTAATTAATAATTAATTTATAATAAATATATGATATTAATTAATATATATATATAATATATTTAATAA
+
AAFFFJFJJJJJJ<FJJJAFJJ<FFAJFJ--FJJJJ-FFF-<FFJFA-FJJ-AJ-<<-FFFAFJJJJJAJ-7---
```

# FASTQ

**FASTQ Sequence Header:** Sequence ID + Description on same line, sequence string on the next

**Whitespace** only required if description present

**"At" symbol**
Start of sequence portion of record

**Description/Comment** optional

```
@SRR10178655.1 0:N:0:
GGATCTATGGCCATGTAGGGACCATCTGAAGGCAGATCAAAATTTCGCTGAGCAAATTTAGGGTCCGGGTTTGTT
+
AA<A<F--FF<-F-A7FAF-F---A<F---<FF-<F--7F-----<-A7F-A----7FJ<-FF--<J<-7-FFFJ
@SRR10178655.2 0:N:0:
ATAAAAAAAAATTAATAATCTATTCTTTATTTAAAACTAATTTTTAAATTAATTGGTTTTTGTGGAATGGTATTA
+
AAFFFJFJAJJJFFA<JF-7FJF<JJJJ--<<FJ-J<A7---<-FFJFJFJJAJ-F<F<<-F-7---7-<-<FFA
@SRR10178655.3 0:N:0:
CATTATATACGTCGCCACTCTTAATTTCCTTTTCCATAAGAGCGTATAATCTTGTAATACAATGTCTTCTCCAAC
+
AAAAFFJJJJAFJJJJAFJF<JFAFJJJF-<FAJJFJ-F-F-<7-7-<FJJJJF<JA-FF---<-7<F-F<-7-<F
@SRR10178655.4 0:N:0:
AAGTTATTCTGCCTCTAATGCGATAACTGTAATCTTTAATTGTGTAATTTCTTTTTCACAATCTGAGCCACGCCA
+
AAAAAJF<-A<-7FJFJJJFJFJJJJJ<FJ--7<FF-7-<--7-A<7FFJAJFFJJJAJ7FF-F7FA-7<-A-7-
@SRR10178655.5 0:N:0:
ATATATTAATTAATAATTAATTTATAATAAATATATGATATTAATTAATATATATATATATAATATATTTAATAA
+
AAFFFJFJJJJJJ<FJJJAFJJ<FFAJFJ--FJJJJ-FFF-<FFJFA-FJJ-AJ-<<-FFFAFJJJJJAJ-7---
```

**Sequence ID**
Required;
Any printable non-whitespace characters
[!-~]

**FASTQ Sequence**
Nucleotide, amino acid, IUPAC codes

Typically *not* wrapped flush (discouraged)

FASTQ files are *best* suffixed with ".fastq" or ".fq", some tools *require* this.

# FASTQ

**FASTQ Sequence Header:** Paired-end or mate-pair reads

Read 1

Type 1:  Read 2

@SRR10178655.1/1
GGATCTATGGCCATGTAGGGACCATCTGAAGGCAGATCAAAATTTCGCTGAGCAAATTTAGGGTCCGGGTTTGTT
+
AA<A<F--FF<-F-A7FAF-F---A<F---<FF-<F--7F-----<-A7F-A----7FJ<-FF--<J<-7-FFFJ
@SRR10178655.1/2
CATTTTTCCAAACATACCATGTCAAACCTGATTTTATCGCTAGGTCTCCTGGCAGAGTAAATCTGATTGGTGAGC
+
-AAFFJJJAF<F-FFFFJJFFJJ<FFFJFFAJFJJJJ-F-<FJ7JJFJJF<F-7A-7FJ-<FJJ<<FJFFJJFJ<

Type 2:

@SRR10178655.1 1:N:0:
GGATCTATGGCCATGTAGGGACCATCTGAAGGCAGATCAAAATTTCGCTGAGCAAATTTAGGGTCCGGGTTTGTT
+
AA<A<F--FF<-F-A7FAF-F---A<F---<FF-<F--7F-----<-A7F-A----7FJ<-FF--<J<-7-FFFJ
@SRR10178655.1 2:N:0:
CATTTTTCCAAACATACCATGTCAAACCTGATTTTATCGCTAGGTCTCCTGGCAGAGTAAATCTGATTGGTGAGC
+
-AAFFJJJAF<F-FFFFJJFFJJ<FFFJFFAJFJJJJ-F-<FJ7JJFJJF<F-7A-7FJ-<FJJ<<FJFFJJFJ<

Read 1

Read 2

Paired (or mated) reads may be interleaved into same file or separate files. If in separate files, Read 1 and Read 2 sequences *must* be in same order.

# FASTQ

**FASTQ Qualities Header:** (optional) Same as Sequence Header, or absent completely

**"Plus" symbol** → +
Start of qualities
portion of record

**Qualities ID**
Optional;
If present, typically
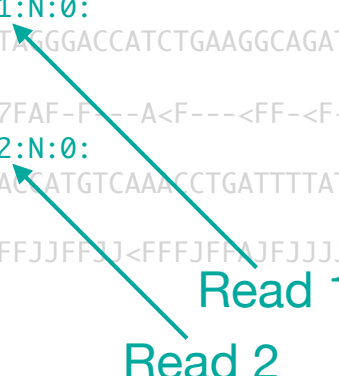same as Sequence
ID; Must follow
same rules

```
@SRR10178655.1 0:N:0:
GGATCTATGGCCATGTAGGGACCATCTGAAGGCAGATCAAAATTTCGCTGAGCAAATTTAGGGTCCGGGTTTGTT
+
AA<A<F--FF<-F-A7FAF-F---A<F---<FF-<F--7F-----<-A7F-A----7FJ<-FF--<J<-7-FFFJ
@SRR10178655.2 0:N:0:
ATAAAAAAAAATTAATAATCTATTCTTTATTTAAAACTAATTTTTAAATTAATTGGTTTTTGTGGAATGGTATTA
+
AAFFFJFJAJJJFFA<JF-7FJF<JJJJ--<<FJ-J<A7---<-FFJFJFJJAJ-F<F<<-F-7---7-<-<FFA
@SRR10178655.3 0:N:0:
CATTATATACGTCGCCACTCTTAATTTCCTTTTCCATAAGAGCGTATAATCTTGTAATACAATGTCTTCTCCAAC
+
AAAFFJJJJAFJJJJAFJF<JFAFJJJF-<FAJJFJ-F-F-<7-7-<FJJJJF<JA-FF---<-7<F-F<-7-<F
@SRR10178655.4 0:N:0:
AAGTTATTCTGCCTCTAATGCGATAACTGTAATCTTTAATTGTGTAATTTCTTTTTCACAATCTGAGCCACGCCA
+
AAAAAJF<-A<-7FJFJJJFJFJJJJJ<FJ--7<FF-7-<--7-A<7FFJAJFFJJJAJ7FF-F7FA-7<-A-7-
@SRR10178655.5 0:N:0:
ATATATTAATTAATAATTAATTTATAATAAATATATGATATTAATTAATATATATATATATAATATATTTAATAA
+
AAFFFJFJJJJJJ<FJJJAFJJ<FFAJFJ--FJJJJ-FFF-<FFJFA-FJJ-AJ-<<-FFFAFJJJJJAJ-7---
```

**FASTQ Qualities**
ASCII+*offset*
encoded "Phred"
scores.

Must be same
length as
sequence.

Typically *not*
wrapped flush
(discouraged)

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

# FASTQ

$$\text{Phred} = -10 \cdot \log_{10}(P)$$

$P$ = fractional probability that the base call is wrong

ascii_char = chr(Phred + offset);     Phred = ord(ascii_char) - offset

| Dec | Hx | Oct | Char | | | | | Dec | Hx | Oct | Html | Chr | Dec | Hx | Oct | Html | Chr | Dec | Hx | Oct | Html | Chr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 000 | NUL | (null) | | | | 32 | 20 | 040 | &#32; | Space | 64 | 40 | 100 | &#64; | @ | 96 | 60 | 140 | &#96; | ` |
| 1 | 1 | 001 | SOH | (start of heading) | | | | 33 | 21 | 041 | &#33; | ! | 65 | 41 | 101 | &#65; | A | 97 | 61 | 141 | &#97; | a |
| 2 | 2 | 002 | STX | (start of text) | | | | 34 | 22 | 042 | &#34; | " | 66 | 42 | 102 | &#66; | B | 98 | 62 | 142 | &#98; | b |
| 3 | 3 | 003 | ETX | (end of text) | | | | 35 | 23 | 043 | &#35; | # | 67 | 43 | 103 | &#67; | C | 99 | 63 | 143 | &#99; | c |
| 4 | 4 | 004 | EOT | (end of transmission) | | | | 36 | 24 | 044 | &#36; | $ | 68 | 44 | 104 | &#68; | D | 100 | 64 | 144 | &#100; | d |
| 5 | 5 | 005 | ENQ | (enquiry) | | | | 37 | 25 | 045 | &#37; | % | 69 | 45 | 105 | &#69; | E | 101 | 65 | 145 | &#101; | e |
| 6 | 6 | 006 | ACK | (acknowledge) | | | | 38 | 26 | 046 | &#38; | & | 70 | 46 | 106 | &#70; | F | 102 | 66 | 146 | &#102; | f |
| 7 | 7 | 007 | BEL | (bell) | | | | 39 | 27 | 047 | &#39; | ' | 71 | 47 | 107 | &#71; | G | 103 | 67 | 147 | &#103; | g |
| 8 | 8 | 010 | BS | (backspace) | | | | 40 | 28 | 050 | &#40; | ( | 72 | 48 | 110 | &#72; | H | 104 | 68 | 150 | &#104; | h |
| 9 | 9 | 011 | TAB | (horizontal tab) | | | | 41 | 29 | 051 | &#41; | ) | 73 | 49 | 111 | &#73; | I | 105 | 69 | 151 | &#105; | i |
| 10 | A | 012 | LF | (NL line feed, new line) | | | | 42 | 2A | 052 | &#42; | * | 74 | 4A | 112 | &#74; | J | 106 | 6A | 152 | &#106; | j |
| 11 | B | 013 | VT | (vertical tab) | | | | 43 | 2B | 053 | &#43; | + | 75 | 4B | 113 | &#75; | K | 107 | 6B | 153 | &#107; | k |
| 12 | C | 014 | FF | (NP form feed, new page) | | | | 44 | 2C | 054 | &#44; | , | 76 | 4C | 114 | &#76; | L | 108 | 6C | 154 | &#108; | l |
| 13 | D | 015 | CR | (carriage return) | | | | 45 | 2D | 055 | &#45; | - | 77 | 4D | 115 | &#77; | M | 109 | 6D | 155 | &#109; | m |
| 14 | E | 016 | SO | (shift out) | | | | 46 | 2E | 056 | &#46; | . | 78 | 4E | 116 | &#78; | N | 110 | 6E | 156 | &#110; | n |
| 15 | F | 017 | SI | (shift in) | | | | 47 | 2F | 057 | &#47; | / | 79 | 4F | 117 | &#79; | O | 111 | 6F | 157 | &#111; | o |
| 16 | 10 | 020 | DLE | (data link escape) | | | | 48 | 30 | 060 | &#48; | 0 | 80 | 50 | 120 | &#80; | P | 112 | 70 | 160 | &#112; | p |
| 17 | 11 | 021 | DC1 | (device control 1) | | | | 49 | 31 | 061 | &#49; | 1 | 81 | 51 | 121 | &#81; | Q | 113 | 71 | 161 | &#113; | q |
| 18 | 12 | 022 | DC2 | (device control 2) | | | | 50 | 32 | 062 | &#50; | 2 | 82 | 52 | 122 | &#82; | R | 114 | 72 | 162 | &#114; | r |
| 19 | 13 | 023 | DC3 | (device control 3) | | | | 51 | 33 | 063 | &#51; | 3 | 83 | 53 | 123 | &#83; | S | 115 | 73 | 163 | &#115; | s |
| 20 | 14 | 024 | DC4 | (device control 4) | | | | 52 | 34 | 064 | &#52; | 4 | 84 | 54 | 124 | &#84; | T | 116 | 74 | 164 | &#116; | t |
| 21 | 15 | 025 | NAK | (negative acknowledge) | | | | 53 | 35 | 065 | &#53; | 5 | 85 | 55 | 125 | &#85; | U | 117 | 75 | 165 | &#117; | u |
| 22 | 16 | 026 | SYN | (synchronous idle) | | | | 54 | 36 | 066 | &#54; | 6 | 86 | 56 | 126 | &#86; | V | 118 | 76 | 166 | &#118; | v |
| 23 | 17 | 027 | ETB | (end of trans. block) | | | | 55 | 37 | 067 | &#55; | 7 | 87 | 57 | 127 | &#87; | W | 119 | 77 | 167 | &#119; | w |
| 24 | 18 | 030 | CAN | (cancel) | | | | 56 | 38 | 070 | &#56; | 8 | 88 | 58 | 130 | &#88; | X | 120 | 78 | 170 | &#120; | x |
| 25 | 19 | 031 | EM | (end of medium) | | | | 57 | 39 | 071 | &#57; | 9 | 89 | 59 | 131 | &#89; | Y | 121 | 79 | 171 | &#121; | y |
| 26 | 1A | 032 | SUB | (substitute) | | | | 58 | 3A | 072 | &#58; | : | 90 | 5A | 132 | &#90; | Z | 122 | 7A | 172 | &#122; | z |
| 27 | 1B | 033 | ESC | (escape) | | | | 59 | 3B | 073 | &#59; | ; | 91 | 5B | 133 | &#91; | [ | 123 | 7B | 173 | &#123; | { |
| 28 | 1C | 034 | FS | (file separator) | | | | 60 | 3C | 074 | &#60; | < | 92 | 5C | 134 | &#92; | \ | 124 | 7C | 174 | &#124; | | |
| 29 | 1D | 035 | GS | (group separator) | | | | 61 | 3D | 075 | &#61; | = | 93 | 5D | 135 | &#93; | ] | 125 | 7D | 175 | &#125; | } |
| 30 | 1E | 036 | RS | (record separator) | | | | 62 | 3E | 076 | &#62; | > | 94 | 5E | 136 | &#94; | ^ | 126 | 7E | 176 | &#126; | ~ |
| 31 | 1F | 037 | US | (unit separator) | | | | 63 | 3F | 077 | &#63; | ? | 95 | 5F | 137 | &#95; | _ | 127 | 7F | 177 | &#127; | DEL |

| $P$ | Phred |
|---|---|
| $1\times10^{0}$ | 0 |
| $1\times10^{-1}$ | 10 |
| $1\times10^{-2}$ | 20 |
| $1\times10^{-3}$ | 30 |
| $1\times10^{-4}$ | 40 |
| $1\times10^{-5}$ | 50 |
| $1\times10^{-6}$ | 60 |

Source: www.LookupTables.com

# FASTQ

$$\text{Phred} = -10 \cdot \log_{10}(P)$$

$P$ = fractional probability that the base call is wrong

```
ascii_char = chr(Phred + offset);      Phred = ord(ascii_char) - offset
```

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS................................
  ..........................................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX..............
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
  .............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                              |    |         |                                        |                    |
  33                             59   64        73                                       104                  126
  0........................26...31.......40
                           -5....0.......9................................40
                                 0.......9............................40
                                  3.....9..............................41
  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, 41)
    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)
```
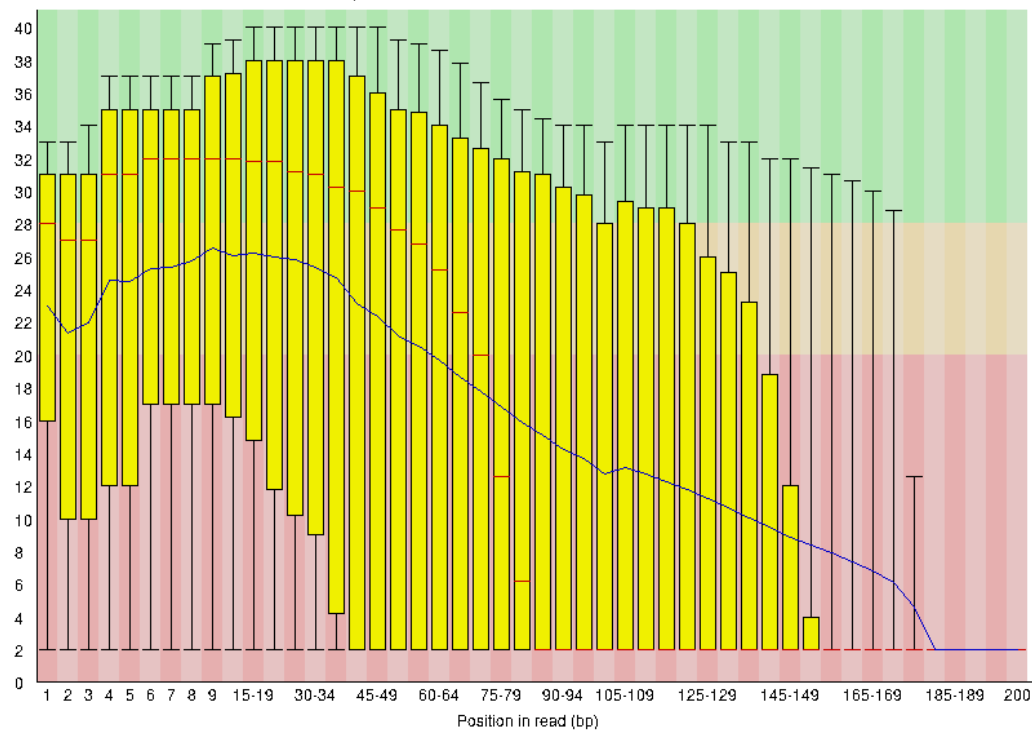
https://en.wikipedia.org/wiki/FASTQ_format

# QA/QC'ing Illumina Reads
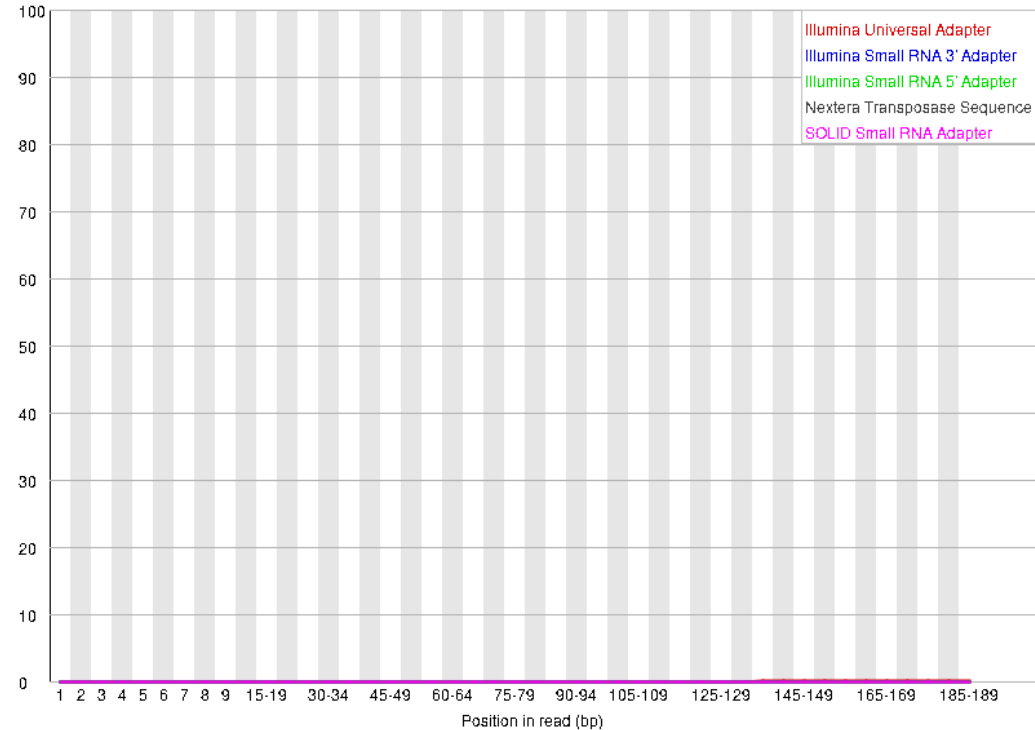
```
$ fastqc --threads 2 --extract SRR10178655_1.fastq.gz SRR10178655_2.fastq.gz

# View the FastQC results in Safari (Mac only):
$ open -a Safari.app SRR10178655_1_fastqc/fastqc_report.html
```

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
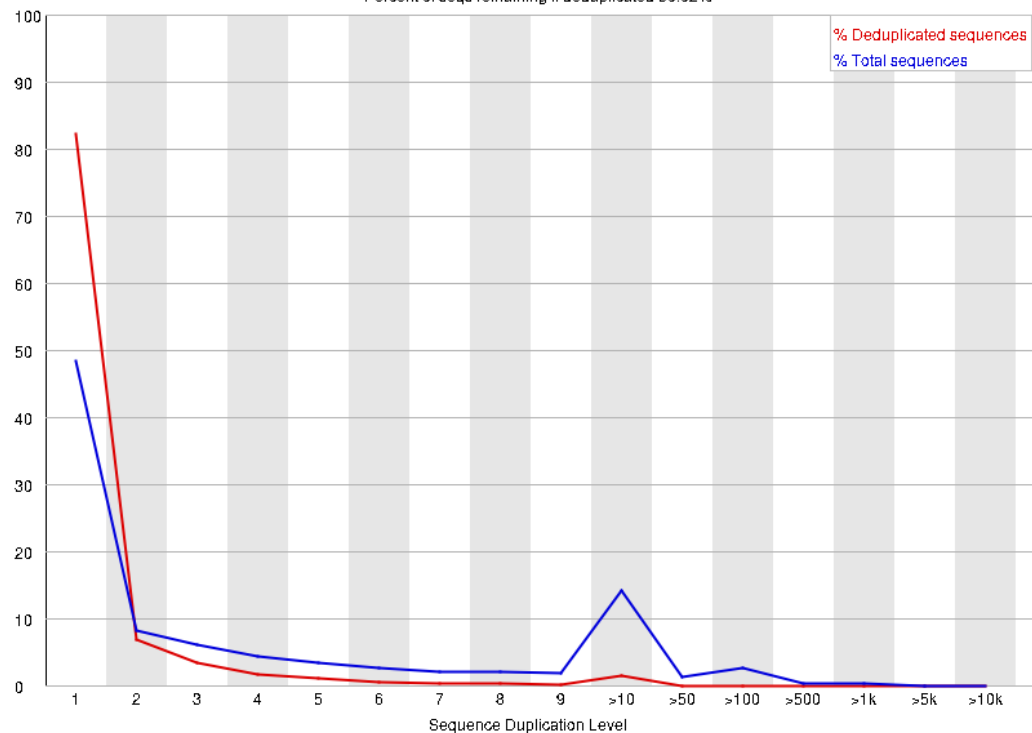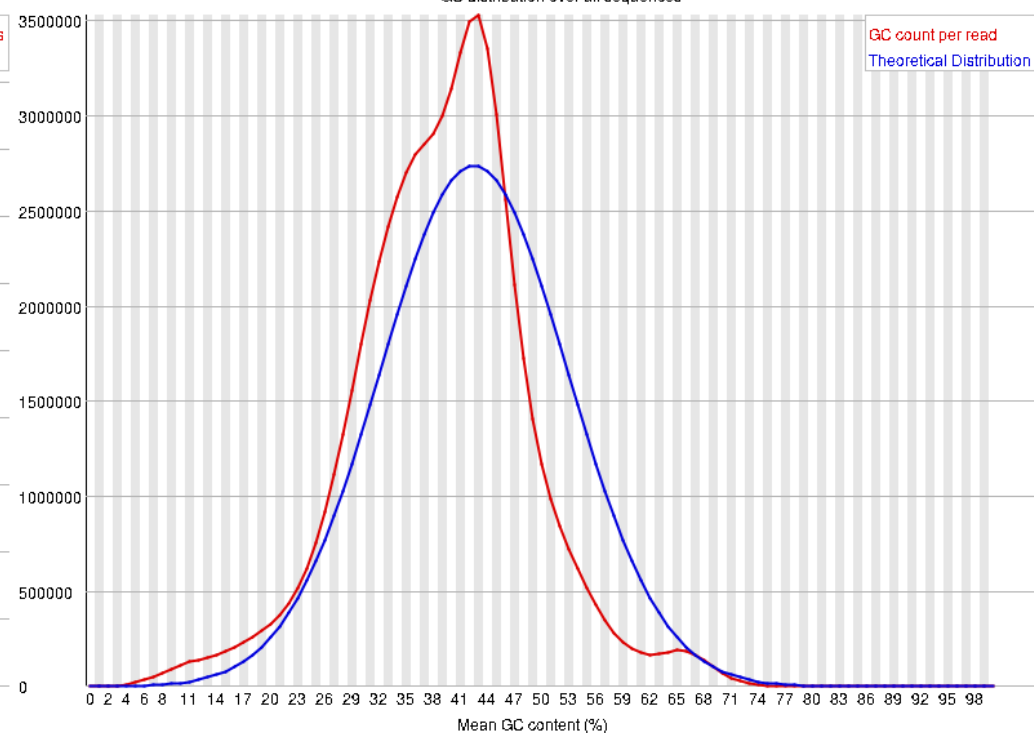
# Trim and align Reads

```
$ java -Xmx500m -jar ./Trimmomatic-0.39/trimmomatic-0.39.jar PE -phred33 \
    -summary SRR10178655.summary SRR10178655_1.fastq.gz SRR178655_2.fastq.gz \
    SRR10178655_1_passed.fastq.gz SRR10178655_1_failed.fastq.gz \
    SRR10178655_1_passed.fastq.gz SRR10178655_2_failed.fastq.gz MINLEN:100 \
    ILLUMINACLIP:./Trimmomatic-0.39/adapters/NexteraPE-PE.fa:2:30:10:2:keepBothReads

$ bwa mem -R '@RG\tID:SRR10178655\tSM:Trex\tLB:HAMMOND01\tPL:ILLUMINA' \
    Trex_genome.fasta SRR10178655_1_passed.fastq.gz SRR10178655_2_passed.fastq.gz | \
    samtools view -b - >SRR10178655.bam

$ samtools sort -m 1g -o SRR10178655.srt.bam SRR10178655.bam

$ samtools index SRR10178655.srt.bam  # creates SRR10178655.srt.bam.bai
```

# SAM/BAM/CRAM

http://samtools.github.io/hts-specs/SAMv1.pdf

http://samtools.github.io/hts-specs/SAMtags.pdf

**SAM:** Sequence Alignment/Map format
**BAM:** Binary SAM
**CRAM:** Reference-Compressed SAM (also binary)

```
@HD   VN:1.3  SO:coordinate
@SQ   SN:ref|NC_001133|  LN:230218
@SQ   SN:ref|NC_001134|  LN:813184
@SQ   SN:ref|NC_001148|  LN:948066
@SQ   SN:ref|NC_001224|  LN:85779
@PG   ID:bwa  PN:bwa  VN:0.7.15-r1140 CL:bwa mem ...
@RG   ID:SRR10178655  SM:Trex  LB:HAMMOND01  PL:ILLUMINA
SRR10178655.85923  163  ref|NC_001133|  1    30  2S7M  =  383  392
  ACATTACTC  AAA))-*##  NM:i:0  MD:i:7  AS:i:7  RG:Z:SRR10178655
SRR10178655.85923  83   ref|NC_001133|  383  60  9M    =  1    -392
  ACCTCACAT  7JFFFFFAA  NM:i:0  MD:Z:9  AS:i:9  RG:Z:SRR10178655
```

# SAM/BAM/CRAM

**SAM Header:** Meta information describing file format and data within. Header lines must start with "@" symbol (and read IDs must not). Tab separated. Reference IDs cannot be "*", "0", or "="; they have special meaning.

**Header** format version and sort order

**Read Group**
Almost required; ID, sample name, and library names, sequencing platform

**Program**
processing history (with commands)

**Sequence**
Reference sequence IDs and lengths; listed in same order as in FASTA

```
@HD   VN:1.3  SO:coordinate
@SQ   SN:ref|NC_001133|  LN:230218
@SQ   SN:ref|NC_001134|  LN:813184
@SQ   SN:ref|NC_001148|  LN:948066
@SQ   SN:ref|NC_001224|  LN:85779
@PG   ID:bwa  PN:bwa  VN:0.7.15-r1140 CL:bwa mem ...
@RG   ID:SRR10178655  SM:Trex  LB:HAMMOND01  PL:ILLUMINA
SRR10178655.85923  163  ref|NC_001133|  1    30   2S7M  = 383  392
     ACATTACTC  AAA))-*##  NM:i:0  MD:i:7  AS:i:7  RG:Z:SRR10178655
SRR10178655.85923  83   ref|NC_001133|  383  60  9M    = 1    -392
     ACCTCACAT  7JFFFFFAA  NM:i:0  MD:Z:9  AS:i:9  RG:Z:SRR10178655
```

# SAM/BAM/CRAM

**SAM Body:** Describes mapping and alignment without the reference. Eleven required fields. Tab separated. Undefined values: "0" for numeric field, a "*" for non-numeric.

**Reference Sequence ID and Position**
1-based alignment start coordinate of this read

**Bit Flags**
binary array of boolean values

**Mapping Quality**
Phred-scaled prob. mapping is wrong

**CIGAR**
describes alignment

**Sequence/Read ID**
same rules as in FASTQ

```
@HD   VN:1.0   SO:coordinate    sample:Trex
@SQ   SN:ref|NC_001133|   LN:230218
@SQ   SN:ref|NC_001134|   LN:813184
@SQ   SN:ref|NC_001148|   LN:948066
@SQ   SN:ref|NC_001224|   LN:85779
@PG   ID:bwa   PN:bwa   VN:0.7.15-r1140  CL:bwa mem ...
@RG   ID:SRR10178655   SM:Trex   LB:HAMMOND01   PL:ILLUMINA
```

SRR10178655.85923   163   ref|NC_001133|   1   30   2S7M   =   383   392
ACATTACTC   AAA))-*##   NM:i:0   MD:i:7   AS:i:7   RG:Z:SRR10178655
```
SRR10178655.85923   83   ref|NC_001133|   383   60   9M   =   1   -392
ACCTCACAT   7JFFFFFAA   NM:i:0   MD:Z:9   AS:i:9   RG:Z:SRR10178655
```

**Insert Distance**
Negative values mean other read in pair is upstream, positive downstream

**Read Sequence and Qualities**
same rules as in FASTQ
offset = 33 required

**Auxiliary Info Tags**
Optional;
RGs almost required

**Reference Sequence ID and Position**
for other read in pair, if applicable

# SAM/BAM/CRAM

## CIGAR operators

| Op | Meaning |
|---|---|
| M : | Match |
| I : | Insertion |
| D : | Deletion |
| = : | Sequence match |
| X : | Sequence mismatch |
| N : | Forward-skip query on reference (intron) |
| H : | Query hard clipping |
| S : | Query soft clipping |
| P : | Padded reference |
| B : | Backward-skip query on reference |

**Example:**

```
Q: ATGACAGGACAGAT-GA GG

   ||| ||||| |||||| ||

R: ATG-CAGGCCAGATTGATA


   3M 1I 10M 1D 2S
   describes same alignment as
   3= 1I 4= 1X 5= 1D 2S
```

## Bit Flags

| n | $2^n$ | Meaning |
|---|---|---|
| 0 : | 1 : | Read is paired |
| 1 : | 2 : | Read is part of proper pair |
| 2 : | 4 : | Read is unmapped |
| 3 : | 8 : | Other read in pair is unmapped |
| 4 : | 16 : | Read is rev complemented |
| 5 : | 32 : | Other read is rev complemented |
| 6 : | 64 : | Read is R1 |
| 7 : | 128 : | Read is R2 |
| 8 : | 256 : | Alignment is a secondary hit |
| 9 : | 512 : | Read fails QA/QC |
| 10 : | 1024 : | Read is duplicate |
| 11 : | 2048 : | Alignment is split/supplementary |

To add or test for flags, use $2^n$ values with bitwise operations:

| **Add flag(s)** | **Test for flag(s)** | |
|---|---|---|
| flags |= 2**0 | flags & 1024 | # correct |
| flags |= 2**1 | flags > 1024 | # incorrect!! |
| flags |= 2**6 | | |

# QA/QC'ing Alignments

```
$ gatk MarkDuplicates --java-options '-Xmx1G' \
   -MAX_FILE_HANDLES 2000 \
   -I SRR10178655.srt.bam \
   -O SRR10178655.srt.mdup.bam \
   -M SRR10178655.metrics

$ samtools stats -@4 --ref-seq Trex_genome.fasta \
   SRR10178655.srt.mdup.bam >SRR10178655.stats

$ plot-bamstats -s Trex_genome.fasta >Trex_genome.gc

$ plot-bamstats -r Trex_genome.gc -p SRR10178655 SRR10178655.stats

$ open -a Safari.app SRR10178655.html
```
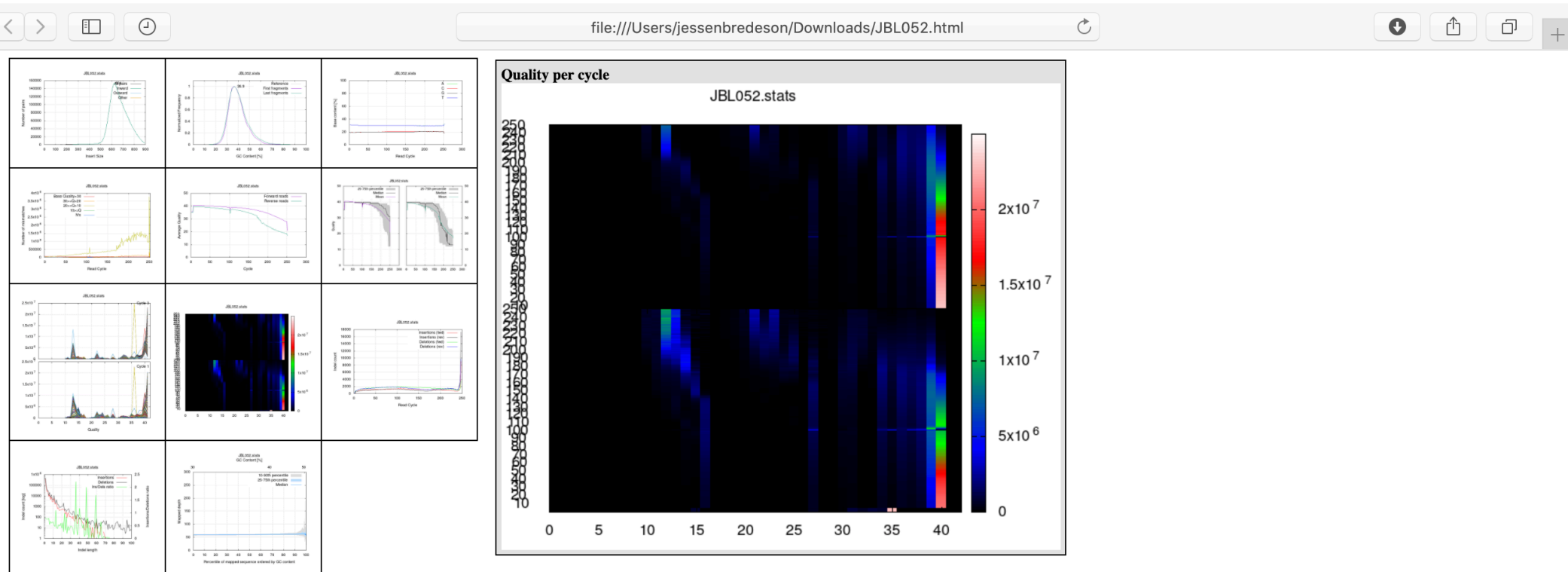
# QA/QC'ing Alignments

```
$ samtools tview SRR10178655.srt.mdup.bam Trex_genome.fasta
```

```
$ samtools view −b −f3 −F3852 SRR10178655.srt.mdup.bam > SRR10178655.srt.mdup.proper.bam
$ samtools index SRR10178655.srt.mdup.proper.bam
$ samtools tview SRR10178655.srt.mdup.proper.bam Trex_genome.fasta
```

```
$ samtools tview SRR10178655.srt.mdup.proper.bam Trex_genome.fasta
```

# Call Variants

```
gatk HaplotypeCaller \
    --minimum-mapping-quality 30 \
    --min-base-quality-score 20 \
    --read-validation-stringency SILENT \
    --reference Trex_genome.fasta \
    --input SRR10178655.srt.mdup.proper.bam \
    --output SRR10178655.vcf
```

# VCF/BCF

**VCF:** Variant Call Format
**BCF:** Binary VCF

```
##fileformat=VCFv4.2
##FILTER=<ID=LowQual,Description="Locus is low quality">
##FILTER=<ID=PASS,Description="Locus passes all filters">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="The Phred-scaled prob. of the genotype">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Number of observation for each allele">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
##contig=<ID=Chr1,length=217471166>
##contig=<ID=Chr2,length=181034961>
```

| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | FORMAT | Trex |
|--------|-----|-----|-----|-----|------|--------|------|--------|------|
| Chr1 | 534 | . | T | A | 8.826 | LowQual | DP=1 | GT:GQ:AD | ./.:0:0,1 |
| Chr1 | 1315 | . | A | G | 564.103 | PASS | DP=51 | GT:GQ:AD | 1\|0:99:26,25 |
| Chr1 | 369655 | . | CTC | CC | 209.026 | . | DP=31 | GT:GQ:AD | 0\|1:99:19,12 |
| Chr1 | 672396 | . | GTT | GT,GGT | 912.199 | . | DP=36 | GT:GQ:AD | 2\|1:43:0,28,8 |
| Chr1 | 2192815 | . | GG | GGTATTTTTAG | 253.597 | . | DP=64 | GT:GQ:AD | 0/1:99:46,18 |

# VCF/BCF

**VCF Metadata Lines:** For humans and computers. Required by most tools to pre-declare how to parse file body correctly.

**fileformat Meta**
Required on first line;
Tells tools how to interpret rest of file

**FILTER Meta**
explicitly defines soft filters one expects to see in the FILTER column

```
##fileformat=VCFv4.2
##FILTER=<ID=LowQual,Description="Locus is low quality">
##FILTER=<ID=PASS,Description="Locus passes all filters">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="The Phred-scaled prob. of the genotype">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Number of observation for each allele">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
##contig=<ID=Chr1,length=217471166>
##contig=<ID=Chr2,length=181034961>
#CHROM   POS       ID    REF   ALT         QUAL      FILTER    INFO    FORMAT      Trex
Chr1     534       .     T     A           8.826     LowQual   DP=1    GT:GQ:AD    ./.:0:0,1
Chr1     1315      .     A     G           564.103   PASS      DP=51   GT:GQ       .25
Chr1     369655    .     CTC   CC          209.      PASS      DP=31   GT:GQ:AD    18,2
Chr1     672396    .     GTT   GT,GGT      513.      PASS      DP=36   GT:GQ:AD    2.11:43:0,28,8
Chr1     210       .     GG    GGTATTTTTAG 253.597   PASS      DP=64   GT:GQ:AD    0/1:99:46,18
```

**contig Meta**
Optional, encouraged;
Describes reference sequences observed in CHROM column

**INFO Meta**
Explicitly defines the types of Key=Value data to be observed in INFO column

**FORMAT Meta**
Explicitly defines the types data to be observed in sample column(s)

# VCF/BCF

**VCF Header Line:** Defines columns, including the sample names. Required by most tools to parse file correctly; undefined fields set to " . "

**Chromosome name and Position**
Sequence IDs should be in contig Meta; Positions 1-based

**Locus ID**
if applicable
*e.g.*, DBsnp ID, *etc.*

**Locus-level Quality Score**
Phred-scaled prob. that locus is not really variant

**Locus-Level Meta Information**
Key=Value pair info about the locus (and all samples at the locus)

```
##fileformat=VCFv4.2
##FILTER=<ID=LowQual,Description="Locus is low quality">
##FILTER=<ID=PASS,Description="Locus passes all filters">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="The Phred-scaled prob of the genotype">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Number of observation for each allele">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
##contig=<ID=Chr1,length=217471166>
##contig=<ID=Chr2,length=181034961>
#CHROM  POS       ID    REF   ALT           QUAL      FILTER    INFO    FORMAT     Trex
Chr1    534       .     T     A             8.826     LowQual   DP=1    GT:GQ:AD   ./.:0:0,1
Chr1    1315      .     A     G             564.103   PASS      DP=51   GT:GQ:AD   1|0:99:26,25
Chr1    369655    .     CTC   CC            209.026   .         DP=31   GT:GQ:AD   0|1:99:19,12
Chr1              .     GTT   GT,GGT        512.199   .               :GQ:AD    2|          8,
Chr1    3192815   .     GG    GGTATTTT...   113.047   .         DP=54   GT:GQ:AD   0|1:99:46,18
```

**Reference and Alternate Alleles**
Alleles observed in reference sequence and samples at the locus

**Locus-level Soft Filter(s)**
"PASS" = passes filters
"." = no filters applied
anything else = failure

**Sample-Level Field Formatting**
Ordered list of fields present in samples

**Sample Field**
Contains sample genotype and associated info at the locus

# VCF/BCF

**VCF Loci:** Tab-delimited columns. Alleles indexed from 0 (REF) to N (ALT) alleles. Genotypes represented with those indices

```
##fileformat=VCFv4.2
##FILTER=<ID=LowQual,Description="Locus is low quality">
##FILTER=<ID=PASS,Description="Locus passes all filters">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="The Phred-scaled prob. of the genotype">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Number of observation for each allele">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
##contig=<ID=Chr1,length=217471166>
##contig=<ID=Chr2,length=181034961>
```

**Substitution locus**

**Complex locus**
Multi-allele;
Deletion *and* substitution!

**No-call or hard-filtered genotype**

| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | FORMAT | Trex |
|--------|-----|-----|-----|-----|------|--------|------|--------|------|
| Chr1 | 534 | . | T | A | 8.826 | LowQual | DP=1 | GT:GQ:AD | ./.:0:0,1 |
| Chr1 | 1315 | . | A | G | 564.103 | PASS | DP=51 | GT:GQ:AD | 1\|0:99:26,25 |
| Chr1 | 369655 | . | CTC | CC | 209.026 | . | DP=31 | GT:GQ:AD | 0\|1:99:19,12 |
| Chr1 | 672396 | . | GTT | GT,GGT | 912.199 | . | DP=36 | GT:GQ:AD | 2\|1:43:0,28,8 |
| Chr1 | 2192815 | . | GG | GGTATTTTTAG | 253.597 | . | DP=64 | GT:GQ:AD | 0/1:99:46,18 |

**Deletion locus**

**Insertion locus**

**Phased genotypes**

**Unphased genotype**

**Allele Depth**
Read count for each allele

# Annotation files

# BED

https://genome.ucsc.edu/FAQ/FAQformat.html#format1

**BED:** Browser Extensible Data format

```
track itemRgb="On"
Chr1  0       126500  CLEAR1  0  +  0       126500  0,0,0
Chr1  126500  128500  BREAK1  0  +  126500  128500  213,221,213
Chr1  128500  278000  CLEAR2  0  +  128500  278000  0,0,0
Chr1  278000  280000  BREAK2  0  +  278000  280000  213,221,213
Chr1  280000  362500  CLEAR3  0  +  280000  362500  0,0,0
Chr1  362500  366000  BREAK3  0  +  362500  366000  213,221,213
Chr1  366000  427500  CLEAR4  0  +  366000  427500  0,0,0
Chr1  427500  429500  BREAK4  0  +  427500  429500  213,221,213
Chr1  429500  599500  CLEAR5  0  +  429500  599500  0,0,0
Chr1  599500  605500  BREAK5  0  +  599500  605500  213,221,213
```

# BED

**BED:** Columns tab-delimited. First three required, all others optional (first 6 typical).

**Track configuration header**
Optional;
key=value pairs for configuring display preferences in a genome browser

**Feature Name**
May contain [!-~] characters and spaces

**Feature Strand**
either "+" or "-", or "." if no strand applies
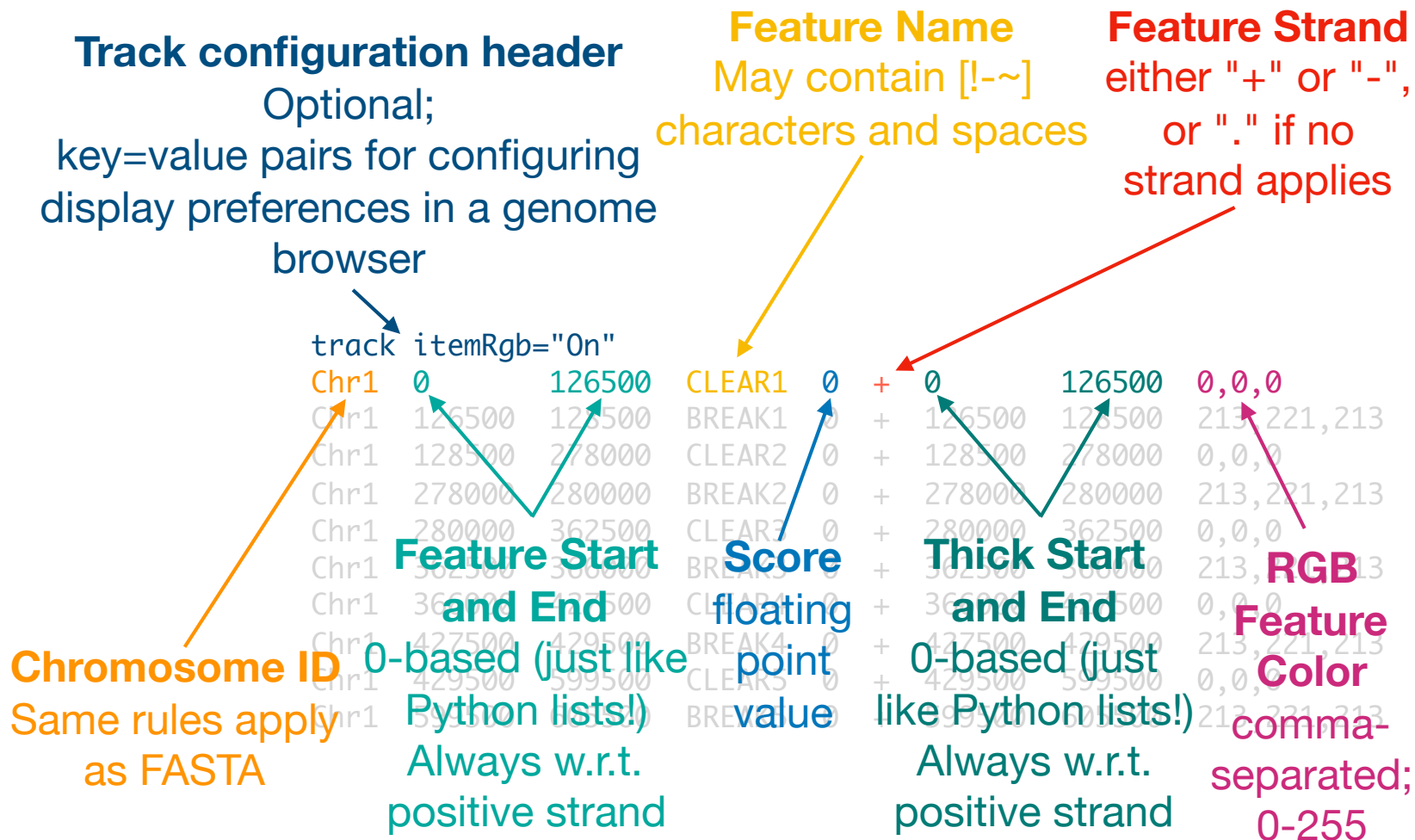
```
track itemRgb="On"
Chr1    0       126500  CLEAR1  0   +   0       126500  0,0,0
Chr1    126500  126500  BREAK1  0   +   126500  126500  213,221,213
Chr1    128500  278000  CLEAR2  0   +   128500  278000  0,0,0
Chr1    278000  280000  BREAK2  0   +   278000  280000  213,221,213
Chr1    280000  362500  CLEAR3  0   +   280000  362500  0,0,0
Chr1    362500  366000  BREAK3  0   +   362500  366000  213,221,213
Chr1    366000  427500  CLEAR4  0   +   366000  427500  0,0,0
Chr1    427500  429500  BREAK4  0   +   427500  429500  213,221,213
Chr1    429500  599500  CLEAR5  0   +   429500  599500  0,0,0
Chr1    599500  603000  BREAK5  0   +   599500  603000  213,221,213
```

**Chromosome ID**
Same rules apply as FASTA

**Feature Start and End**
0-based (just like Python lists!)
Always w.r.t. positive strand

**Score**
floating point value

**Thick Start and End**
0-based (just like Python lists!)
Always w.r.t. positive strand

**RGB Feature Color**
comma-separated; 0-255

# GFF3

https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md

**GFF:** Generic Feature Format

```
##gff-version 3
##species http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=436495
##genome-build RexBase Trex1
##sequence-region Chr1 1 217471166
# Note Trex_genome.fasta, complete genome
Chr1   Gnomon   gene          43895   78350   .   +   .   ID=gene32251;Name=LOC101732307
Chr1   Gnomon   mRNA          43895   78350   .   +   .   ID=rna61088;Name=XM_012954515.1;Parent=gene32251
Chr1   Gnomon   CDS           43895   43947   .   +   0   ID=rna61088.1.CDS;Parent=rna61088
Chr1   Gnomon   exon          43895   43947   .   +   .   ID=rna61088.1.exon;Parent=rna61088
Chr1   Gnomon   start_codon   43895   43897   .   +   0   ID=rna61088.1.start_codon;Parent=rna61088
Chr1   Gnomon   CDS           48839   49007   .   +   1   ID=rna61088.2.CDS;Parent=rna61088
Chr1   Gnomon   exon          48839   49007   .   +   .   ID=rna61088.2.exon;Parent=rna61088
Chr1   Gnomon   CDS           53889   54000   .   +   0   ID=rna61088.3.CDS;Parent=rna61088
Chr1   Gnomon   exon          53889   54000   .   +   .   ID=rna61088.3.exon;Parent=rna61088
Chr1   Gnomon   CDS           55055   55173   .   +   2   ID=rna61088.4.CDS;Parent=rna61088
Chr1   Gnomon   exon          55055   55173   .   +   .   ID=rna61088.4.exon;Parent=rna61088
```

# GFF3

**GFF Header:** Pragma begin with "##", comments with "#". Format pragma required for GFF3, highly-recommended for GFF2/GTF.

**Pragma/Directives**

Pre-declared set of pragma with specific formats/definitions.
Mostly for computers/browsers.

```
##gff-version 3
##species http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=436495
##genome-build RexBase Trex1
##sequence-region Chr1 1 217471166
# Note Trex_genome.fasta, complete genome
Chr1   Gnomon   gene          43895   78350   .   +   .   ID=gene32251;Name=LOC101732307
Chr1   Gnomon                 43895   78350   .   +   .   ID=rna61088;Name=XM_012954515.1;Parent=gene32251
Chr1   Gnomon   CDS           43895   43947   .   +   0   ID=rna61088.1.CDS;Parent=rna61088
Chr1   Gnomon   exon          43895           .       .   ID=rna61088.1.exon;Parent=rna61088
Chr1   Gnomon   start_codon   43895   43897   .   +   0   ID=rna61088.1.start_codon;Parent=rna61088
Chr1   Gnomon   CDS           48839   49007   .   +   1   ID=rna61088.2.CDS;Parent=rna61088
Chr1   Gnomon   exon          48839   49007   .   +   .   ID=rna61088.2.exon;Parent=rna61088
Chr1   Gnomon   CDS           53889   54000   .       0   ID=rna61088.3.CDS;Parent=rna61088
Chr1   Gnomon   exon          53889   54000   .   +   .   ID=rna61088.3.exon;Parent=rna61088
Chr1   Gnomon   CDS           55055   55173   .   +   2   ID=rna61088.4.CDS;Parent=rna61088
Chr1   Gnomon   exon          55055   55173   .   +   .   ID=rna61088.4.exon;Parent=rna61088
```

**Format Version Pragma/Directive**
Required for GFF3, highly-recommended for GFF2/GTF formats

**Comments**
Free-form text for humans, ignored by parsers.

# GFF3

**GFF Features:** Nine tab-delimited fields required. Null values a ".".

**Reference ID**
Chromosome/scaffold ID
May only contain
characters in set:
[a-zA-Z0-9.:^*$@!+_?-|]

**Feature Type**
Must be SO term or
accession number

**Feature Strand**
either "+" or "-",
or "." if no
strand applies

**Feature Attributes**
Semi-colon separated
Key=Value pairs;
reserved keys begin with
capitals letters;
"Parent" attribute defines
feature hierarchy; must use
URL-escaping for
forbidden characters

**Score**
floating point
number

```
##gff-version 3
##species http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=436495
##genome-build RexBase Trex1
##sequence-region Chr1 1 21747116
# Note Trex_genome.fasta, complete genome
Chr1   Gnomon   gene          43895   78350   .   +   .   ID=gene32251;Name=LOC101732307
Chr1   Gnomon   mRNA          43895   78350   .   +   .   ID=rna61088;Name=XM_012954515.1;Parent=gene32251
Chr1   Gnomon   CDS           43895   43947   .   +   0   ID=rna61088.1.CDS;Parent=rna61088
Chr1   Gnomon   exon          43895   43947   .   +   .   ID=rna61088.1.exon;Parent=rna61088
Chr1   Gnomon   start_codon   43895   43897   .   +   0   ID=rna61088.1.start_codon;Parent=rna61088
Chr1   Gnomon   CDS           48839   49007   .   +   1   ID=rna61088.2.CDS;Parent=rna61088
Chr1   Gnomon   exon          48839   49007   .   +   .   ID=rna61088.2.exon;Parent=rna61088
Chr1   Gnomon   CDS           53889   54000   .   +   0   ID=rna61088.3.CDS;Parent=rna61088
Chr1   Gnomon   exon          53889   54000   .   +   .   ID=rna61088.3.exon;Parent=rna61088
Chr1   Gnomon   CDS           55055   55173   .   +   2   ID=rna61088.4.CDS;Parent=rna61088
Chr1   Gnomon   exon          55055   55173   .   +   .   ID=rna61088.4.exon;Parent=rna61088
```

**Source**
Usually the program or
organization that
generated the annotations

**Start and End
Positions**
1-based
coordinates on
"+" strand

**Codon Phase**
either 0, 1, or 2;
Offset to next
codon position

# Resources

## File manipulation/filtering

| | | |
|---|---|---|
| **pysam (API)** | FASTA/Q, BED, B/CR/SAM, B/VCF | https://pysam.readthedocs.io/en/latest/api.html#sam-bam-cram-files |
| **pyFaidx (API)** | FASTA | https://doi.org/10.7287/peerj.preprints.970v1 |
| **Seqtk** | FASTA/Q | https://github.com/lh3/seqtk |
| **Seqkit** | FASTA/Q | https://doi.org/10.1371/journal.pone.0163962 |
| **seqmagick** | Many | https://seqmagick.readthedocs.io |
| **bedtools** | BAM, BED, GFF, VCF | https://bedtools.readthedocs.io |
| **bcftools** | VCF/BCF | https://samtools.github.io/bcftools |
| **genometools** | FASTA/Q, GFF3 | http://genometools.org |
| **samtools** | FASTA/Q, B/SAM | https://github.com/samtools/samtools |
| **vcftools** | VCF/BCF | https://vcftools.github.io/man_latest.html |

# Resources

## QA/QC, Adapter and Quality trimming

| | | |
|---|---|---|
| **trimmomatic** | FASTQ | http://usadellab.org/cms/?page=trimmomatic |
| **FastQC** | FASTQ, B/SAM | https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ |
| **Sickle** | FASTA/Q | https://github.com/ucdavis-bioinformatics/sickle |
| **Scythe** | FASTA/Q | https://github.com/ucdavis-bioinformatics/scythe |
| **Sabre** | FASTA/Q | https://github.com/najoshi/sabre |
| **cutadapt** | FASTA/Q | https://cutadapt.readthedocs.io/en/stable/ |

## Alignment

| | | |
|---|---|---|
| **minimap2** | FASTA/Q | https://github.com/lh3/minimap2 |
| **BWA** | FASTA/Q | https://github.com/lh3/bwa |
| **bowtie2** | FASTA/Q | http://bowtie-bio.sourceforge.net/bowtie2/index.shtml |
| **STAR** | FASTQ | https://github.com/alexdobin/STAR |
| **GMAP** | FASTA/Q | http://research-pub.gene.com/gmap/ |
| **exonerate** | FASTA | https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate |

# Resources

**Variant calling**

| | | |
|---|---|---|
| **FreeBayes** | BAM, VCF | https://github.com/ekg/freebayes |
| **GATK4** | FASTA/Q, B/CRAM, VCF | https://software.broadinstitute.org/gatk/documentation |
| **DeepVariant** | FASTA/Q | https://github.com/google/deepvariant |
| **vg** | FASTA/Q | https://github.com/vgteam/vg |

# Common file issues

- Non-printable characters

- Non-ASCII encoded characters

- Incorrect formatting (spaces instead of tabs)

- Truncated files

# od -c