Bioinformatics file formats

Jessen V. Bredeson
DOE Joint Genome Institute

- and -

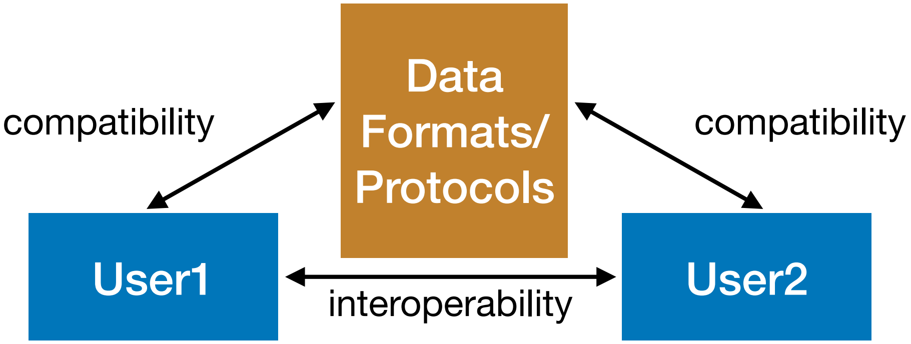
Eric Ross
Stowers Institute for Medical Research

Goals

- Understand importance of standardized file formats
- Introduce commonly-used formats in bioinformatics
- Resources for manipulating or parsing them yourself

Why are (standardized) file formats important?

FAIR - Findable, Accessible, Interoperable, Reusable



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 lack of interoperability can be a consequence of a lack of attention to standardization during the design of a program"[2]

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

Text vs Binary

Computers represent characters as a series of 0s and 1s (bits), in multiples of 8 (bytes), which get appropriately encoded/decoded by software able to read/write that encoding

Text (ASCII)

```
less username.tsv

Username        Identifier      First
name      Last name
booker12        9012    Rachel  Booker
grey07  2070    Laura   Grey
johnson81       4081    Craig   Johnson
jenkins46       9346    Mary    Jenkins
smith79 5079    Jamie   Smith
```

Binary

```
less username.tsv.gz

^_<8B>^H^H<A2><FD>/e^@^Cusername.tsv^@—
<CB>A^N<82>0^P<85><E1><F5><CC>)<BA<B4>v<
AB><89>F<A2>ESC<8D>^G^Xu<84>^A<DB>&S\p{+
<B8><F9><F3><F2>%<EF><96>X^^Cy<86><E3><9
3><C3> /
a<85><BD>h^Z^V<93><9E><E8><BF><F0>^^c<CF
>Z<94><E0>L̋^^—<BF>a;!
6ʣ<B1>P^Zk<F2><E3><A3>^D̲<87>,<D8><C5>6<A
4>^X6^ET&g<A7>$^M<D4>3bā<97><90><AA>5<B8
>eIt<84>z6L^<86><D6>:X<99><9C><9A><BC>0\
^?<84><F8>^E<99><A3><88>^?<AF>^@^@^@
```

Common text formats

- Single-line records
  - SAM
  - VCF
  - BED & BEDGRAPH
  - GFF/GTF
  - GFA
  - NEWICK
- Multi-line records
  - FASTA
  - FASTQ
  - GENBANK
  - JSON
  - YAML
  - XML

Single-line records

SAM

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

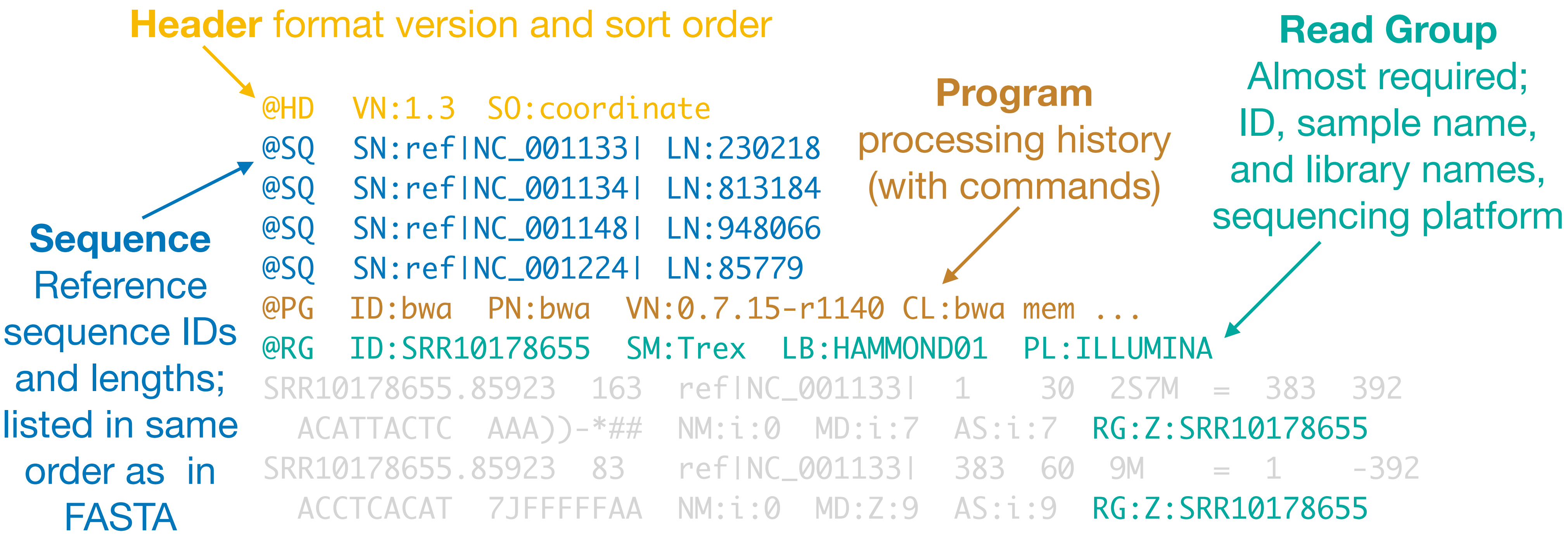
SAM: Sequence Alignment/Map format (file suffix: .sam)

```
@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
```

High-throughput aligners, such as BWA, STAR, bowtie2, minimap2

SAM

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

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

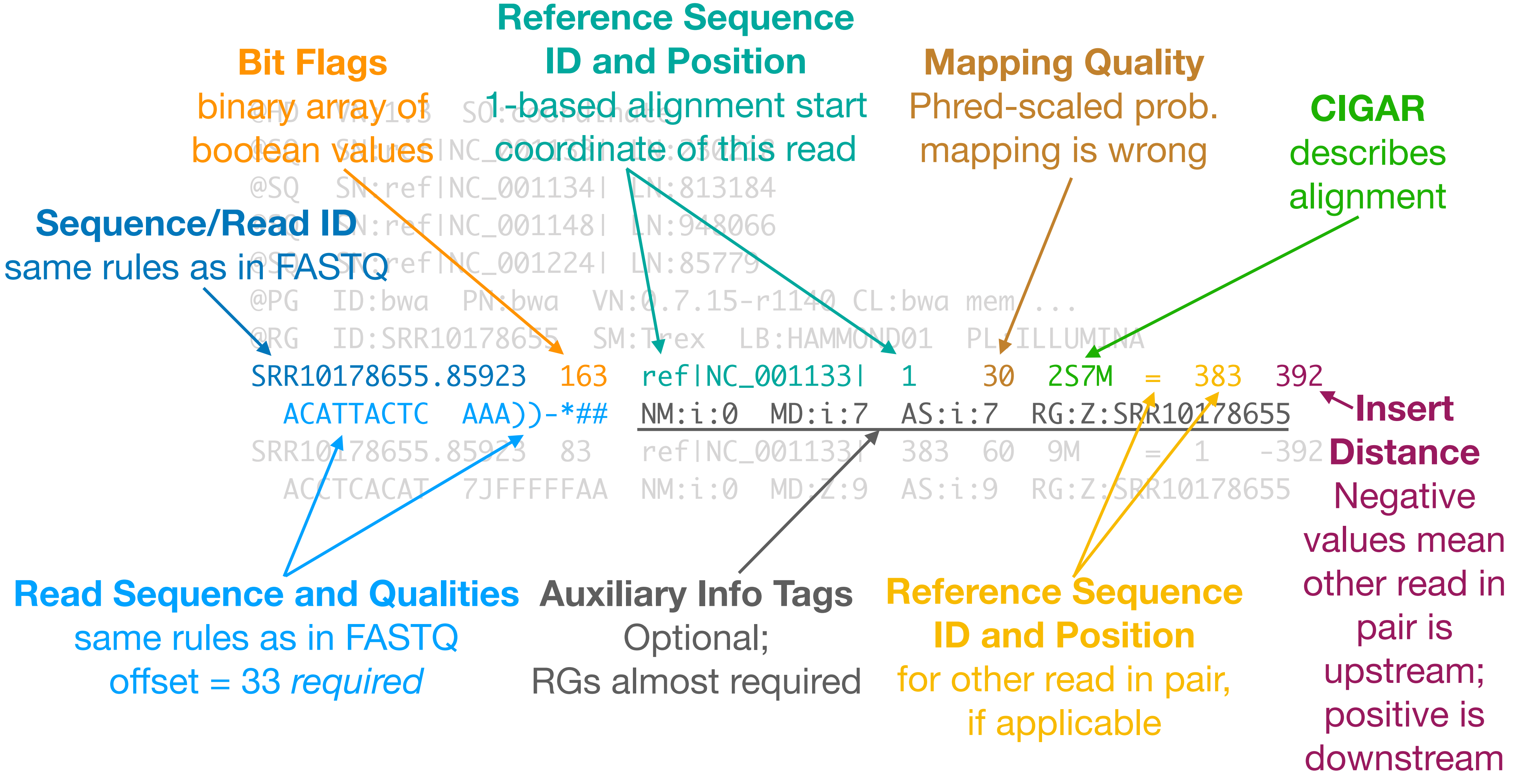
Program
processing history
(with commands)

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

```
@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

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



```
@HD   VN:1.5   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
```

Sequence/Read ID
same rules as in FASTQ

Bit Flags
binary array of boolean values

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

Mapping Quality
Phred-scaled prob. mapping is wrong

CIGAR
describes alignment

Insert Distance
Negative values mean other read in pair is upstream; positive is 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

Bitwise flag and CIGAR field details

Useful with `samtools view -f -F` filtering flags; See also `samtools flags`

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 bit flag(s)

```
flags |= 2**0
flags |= 2**1
flags |= 2**6
```

Test for presence of bit flag(s)

```
flags & 1024   # correct
flags > 1024   # incorrect!!
```

CIGAR operators

Op	Meaning
M :	Match
I :	Insertion
D :	Deletion
= :	Sequence match
X :	Sequence mismatch
N :	Forward-skip query on reference (intron)
H :	Hard-clipped unaligned query sequence end
S :	Soft-clipped unaligned query sequence end
P :	Padded reference
B :	Backward-skip query on reference

Example:

For the following alignment:

```
                     GG
Q: ATGACAGGACAGAT-GA
   ||| |||| ||||| ||
R: ATG-CAGGCCAGATTGATA
```

The standard CIGAR string:

```
3M 1I 10M 1D 2S
```

describes same alignment as this, but with mismatches:

```
3= 1I 4= 1X 5= 1D 2S
```

VCF

http://samtools.github.io/hts-specs/VCFv4.3.pdf

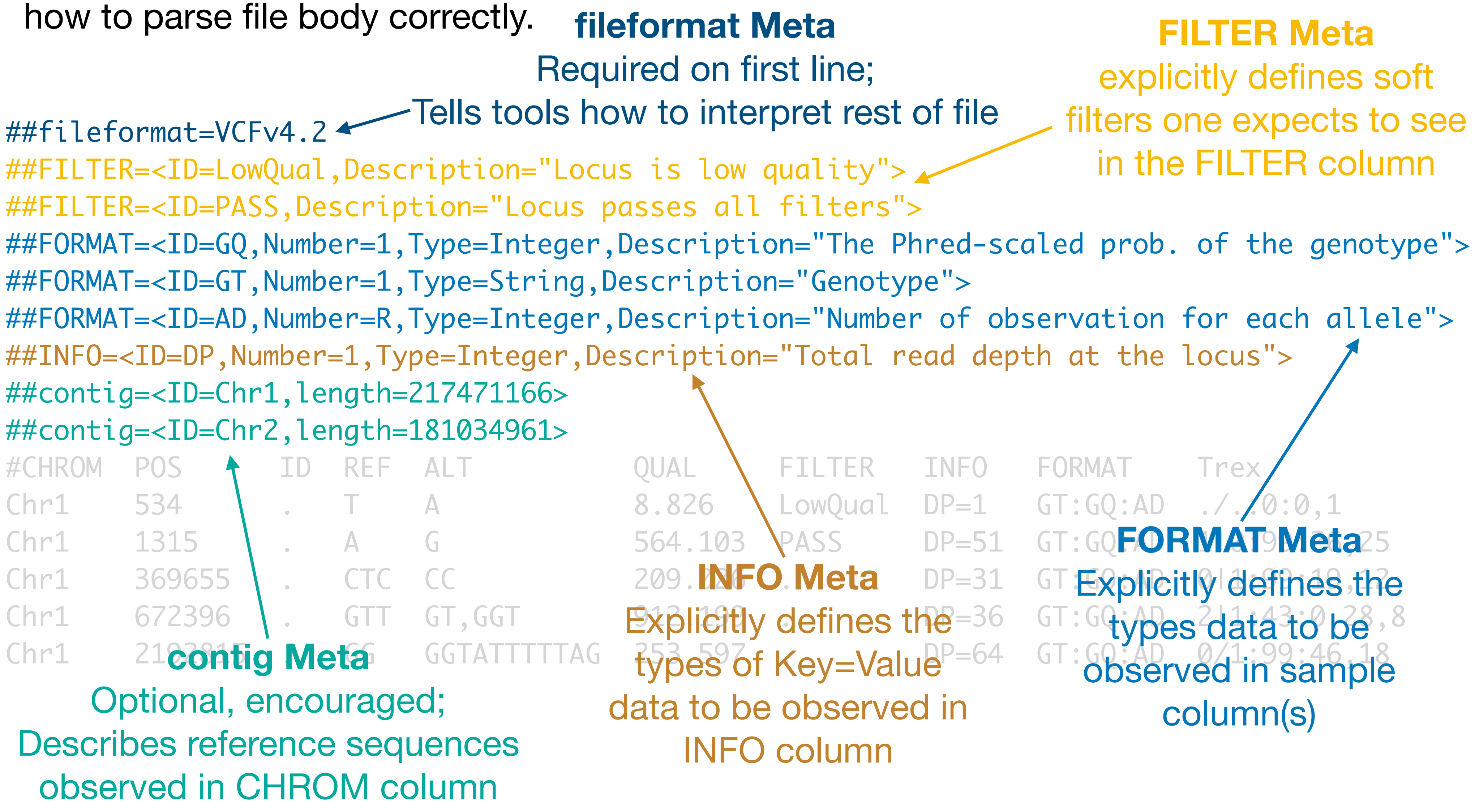
VCF: Variant Call Format (file suffix: .vcf)

```
##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
```

Variant callers, such as GATK, FreeBayes, DeepVariant

VCF

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



```
##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  1/1:99:0,25
Chr1    369655    .   CTC  CC           209.226  .        DP=31  GT:GQ:AD  0|1:99:10,12
Chr1    672396    .   GTT  GT,GGT       913.198  .        DP=36  GT:GQ:AD  2|1:43:0,28,8
Chr1    21028115  .   GG   GGTATTTTTAG  353.597  .        DP=64  GT:GQ:AD  0/1:99:46,18
```

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

contig Meta

Optional, encouraged;

Describes reference sequences observed in CHROM column

INFO Meta

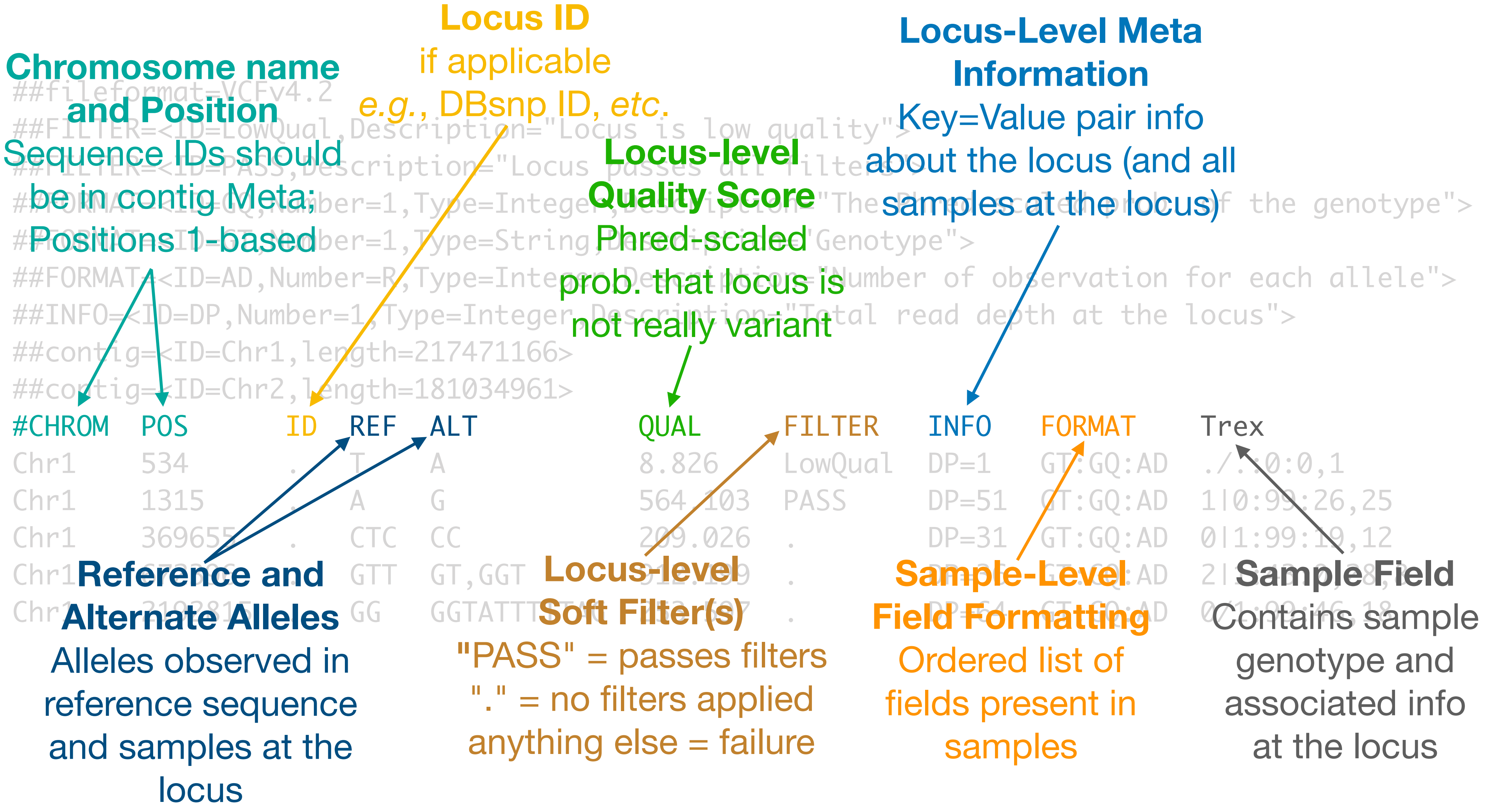
Explicitly defines the types of Key=Value data to be observed in INFO column

FORMAT Meta

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

VCF

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



```
##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    672326   .   GTT  GT,GGT       512.199  .        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
```

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)

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

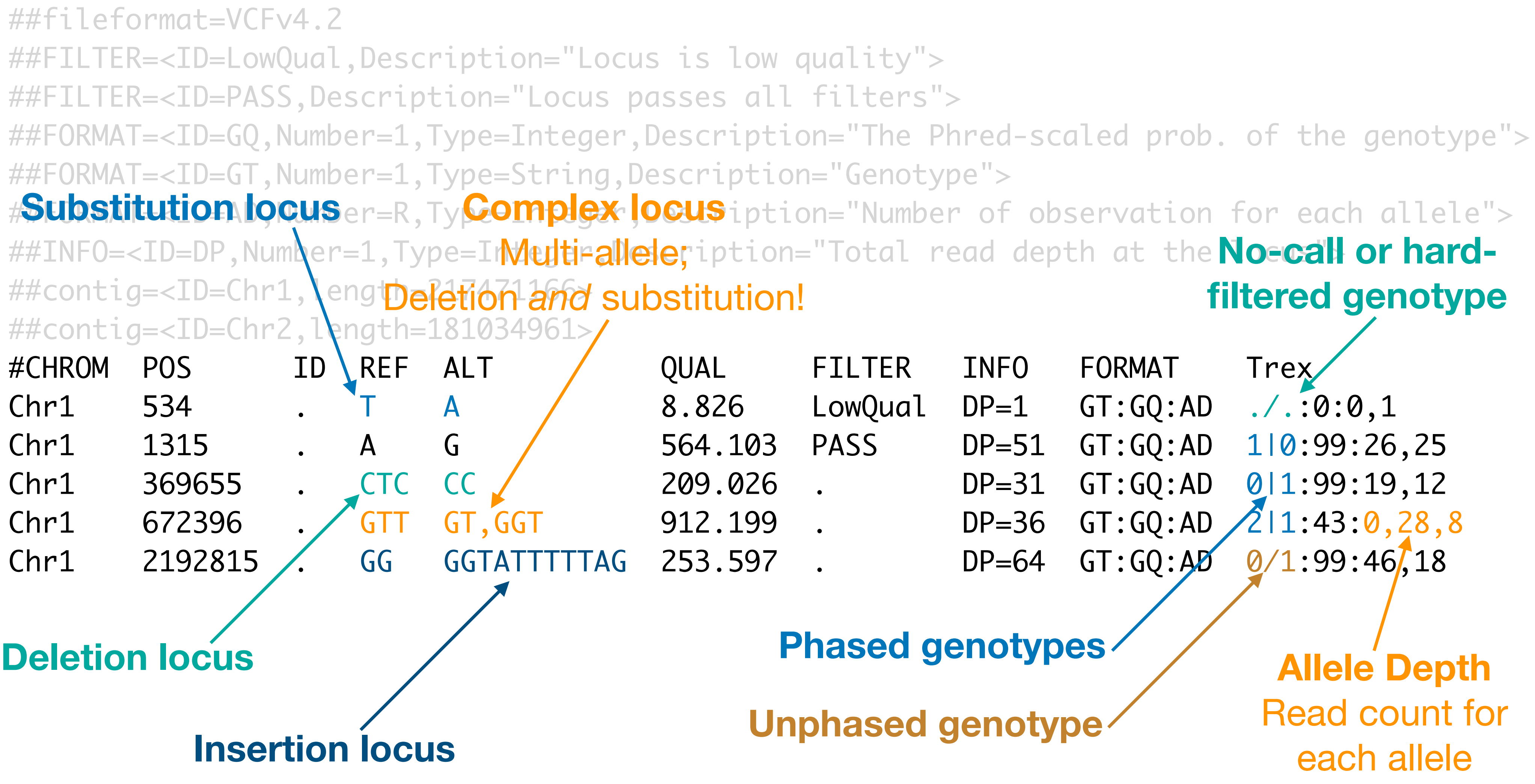
Locus-level Soft Filter(s)
"PASS" = passes filters
"." = no filters applied
anything else = failure

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

Sample Field
Contains sample genotype and associated info at the locus

VCF

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>
#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
```

Substitution locus

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

No-call or hard-filtered genotype

Deletion locus

Insertion locus

Phased genotypes

Unphased genotype

Allele Depth
Read count for each allele

BED

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

https://genome.ucsc.edu/goldenPath/help/bedgraph.html

BED: Browser Extensible Data format (file suffix: .bed)

```
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
```

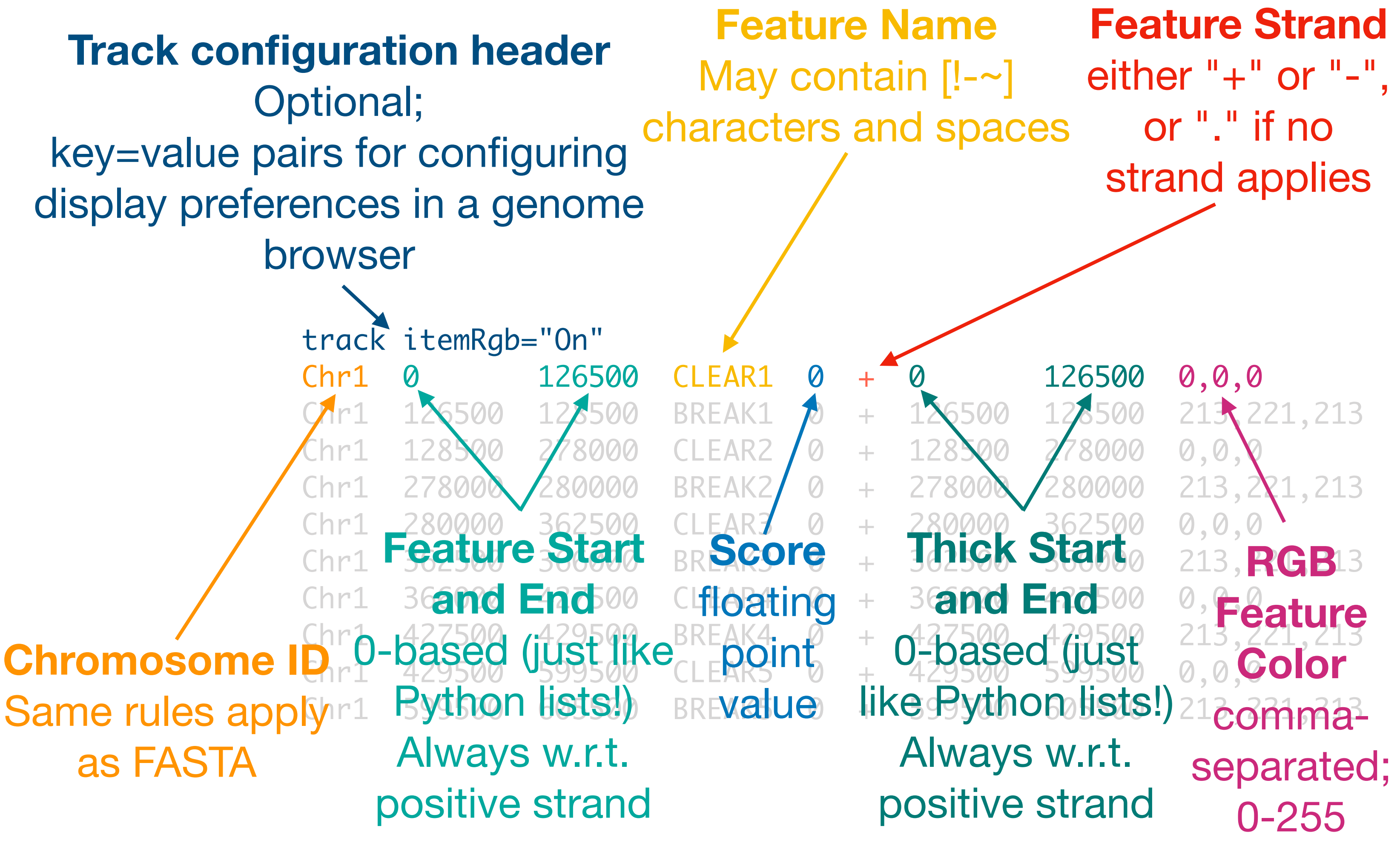
bedgraph: BED continuous graphing format (file suffix: .bedgraph)

```
Chr1  0       126500  2
Chr1  126500  128500  4
Chr1  128500  278000  5
Chr1  278000  280000  10
Chr1  280000  362500  13
Chr1  362500  366000  14
Chr1  366000  427500  13
Chr1  427500  429500  13
Chr1  429500  599500  15
Chr1  599500  605500  14
```

Genome browsers, MACS ChIPseq peak caller, BEDtools

BED

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



```
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
```

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

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

GFF3: Generic Feature Format, version 3 (file suffix: .gff3, .gff)

```
##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
```

Genome annotations tools and browsers, such as Augustus, MAKER, Helixer, miniprot

GFF3

GFF Header: Pragma begin with "##", comments with "#". Format version 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   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
```

Format Version Pragma/Directive

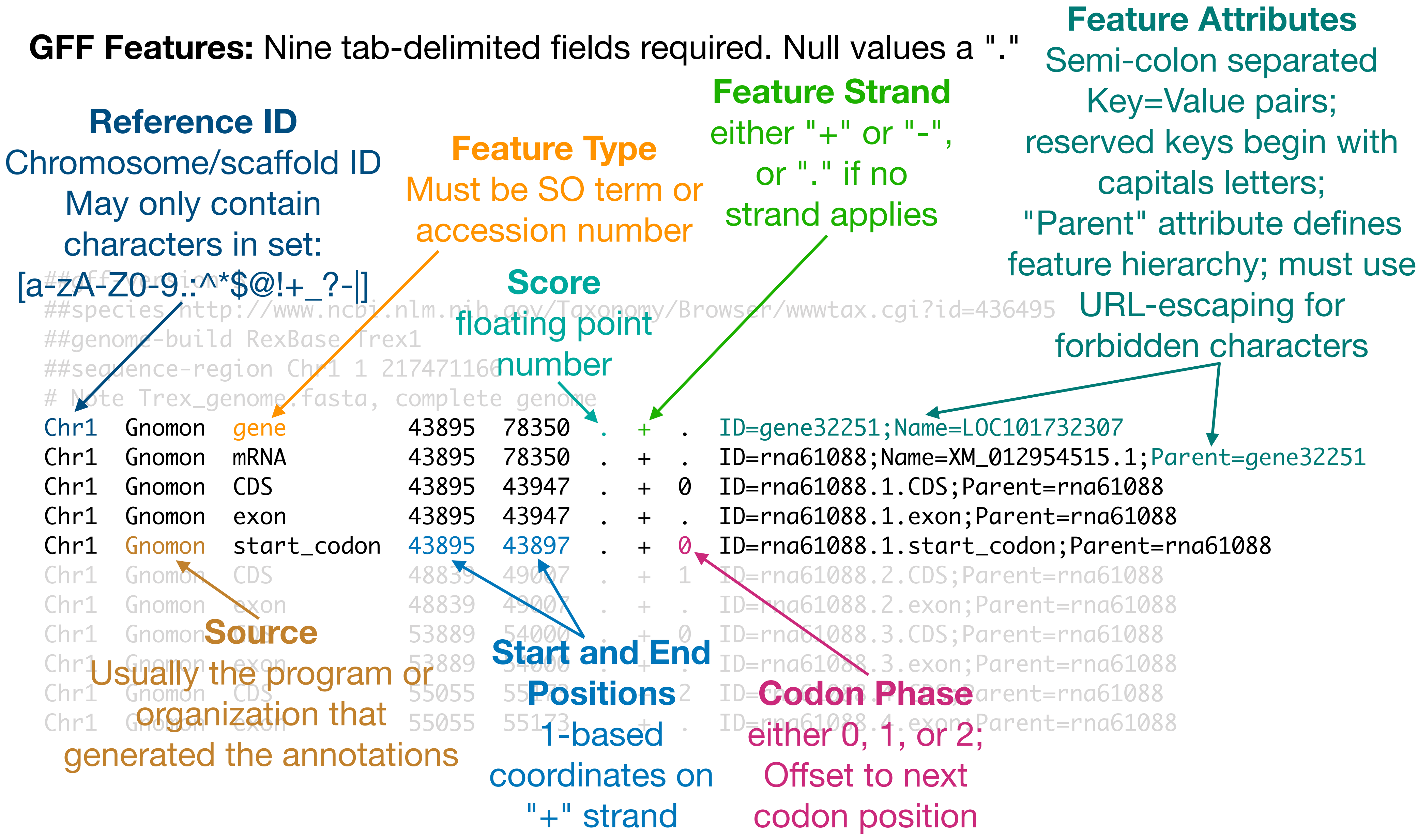
Required for GFF3, highly-recommended for GFF2/GTF formats

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

Score
floating point 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

```
##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
```

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

GFA

https://gfa-spec.github.io/GFA-spec/GFA1.html

GFA: Graphical Fragment Assembly format (file suffix: .gfa)

Assembly graph:

```
H   VN:Z:1.0
S   11  ACCTT
S   12  TCAAGG
S   13  CTTGATT
L   11  +   12  -   4M
L   12  -   13  +   5M
L   11  +   13  +   3M
P   14  11+,12-,13+4M,5M
```

Pangenome graph:

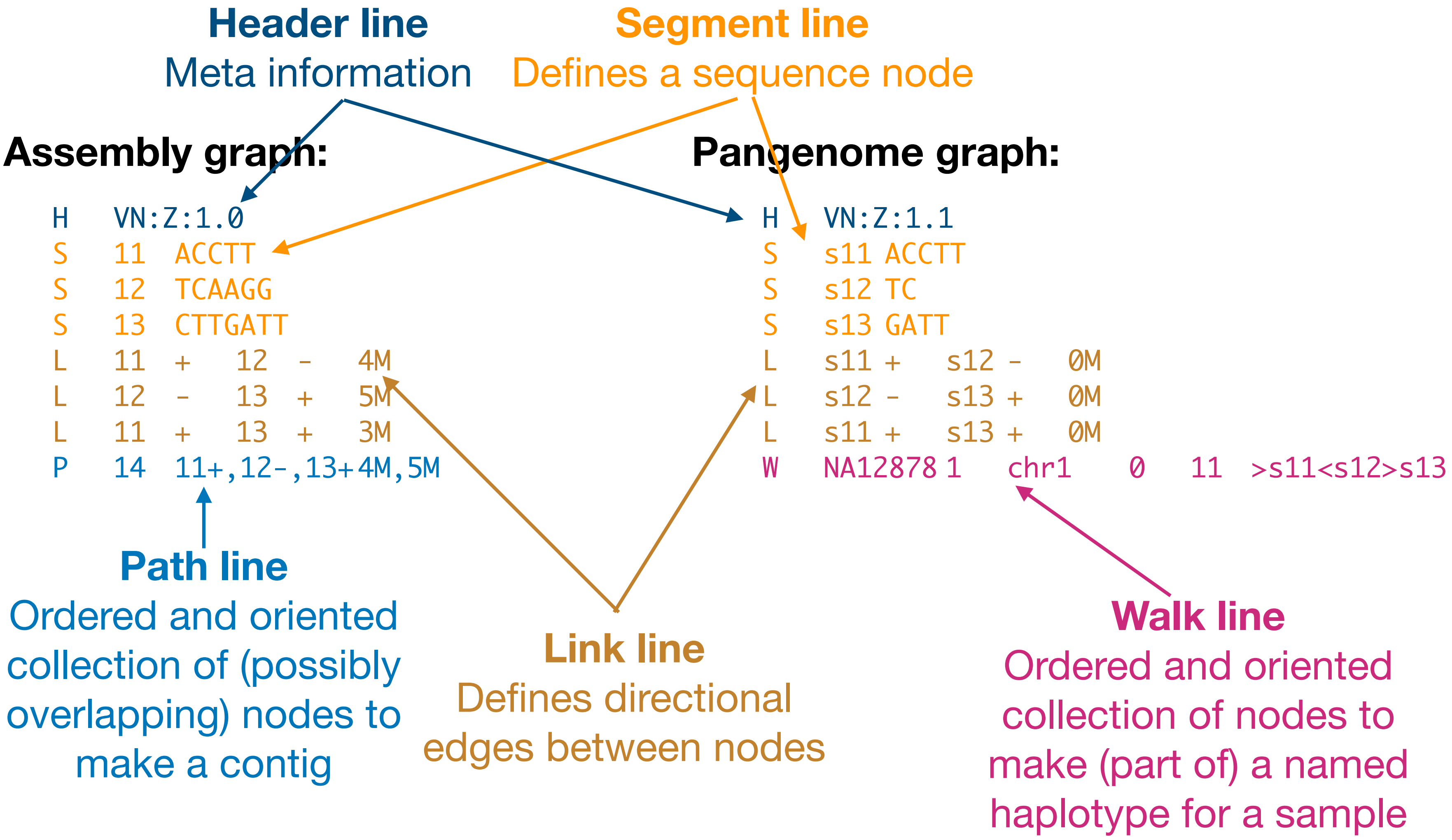
```
H   VN:Z:1.1
S   s11 ACCTT
S   s12 TC
S   s13 GATT
L   s11 +   s12 -   0M
L   s12 -   s13 +   0M
L   s11 +   s13 +   0M
W   NA12878 1   chr1    0   11
>s11<s12>s13
```

Genome assemblers, such as hifiasm, Canu; Pangenome graph tools, VG, ODGI

GFA

https://gfa-spec.github.io/GFA-spec/GFA1.html

GFA: Graphical Fragment Assembly format (file suffix: .gfa)



Header line
Meta information

Segment line
Defines a sequence node

Assembly graph:

```
H   VN:Z:1.0
S   11  ACCTT
S   12  TCAAGG
S   13  CTTGATT
L   11  +   12  -   4M
L   12  -   13  +   5M
L   11  +   13  +   3M
P   14  11+,12-,13+4M,5M
```

Pangenome graph:

```
H   VN:Z:1.1
S   s11 ACCTT
S   s12 TC
S   s13 GATT
L   s11 +   s12 -   0M
L   s12 -   s13 +   0M
L   s11 +   s13 +   0M
W   NA12878 1   chr1    0   11  >s11<s12>s13
```

Path line
Ordered and oriented collection of (possibly overlapping) nodes to make a contig

Link line
Defines directional edges between nodes

Walk line
Ordered and oriented collection of nodes to make (part of) a named haplotype for a sample

GFA

https://gfa-spec.github.io/GFA-spec/GFA1.html

GFA: Graphical Fragment Assembly format (file suffix: .gfa)

Segment overlap CIGAR
SAM-like CIGAR of
overlapped segments

Assembly graph:

```
H   VN:Z:1.0
S   11  ACCTT
S   12  TCAAGG
S   13  CTTGATT
L   11  +   12  -   4M
L   12  -   13  +   5M
L   11  +   13  +   3M
P   14  11+,12-,13+  4M,5M
```

Pangenome graph:

```
H   VN:Z:1.1
S   s11 ACCTT
S   s12 TC
S   s13 GATT
L   s11 +   s12 -   0M
L   s12 -   s13 +   0M
L   s11 +   s13 +   0M
W   NA12878 1   chr1   0   11
>s11<s12>s13
```

Strand/orientation
Defines orientation relative to
sequence strand in Segment lines

```
11+ ACCTT
12-  CCTTGA
13+   CTTGATT
14  ACCTTGATT
```

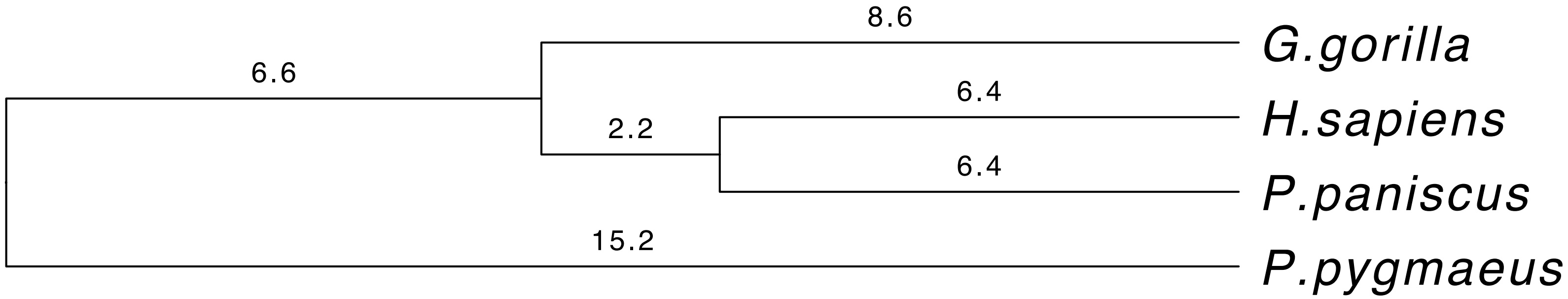
```
                   s11+   s13+
NA12878 chr1  ACCTTGAGATT
                        s12-
```

NEWICK

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

NEWICK: Newick tree format (file suffix: .nwk, .newick)

```
((G.gorilla:8.6,(H.sapiens:6.4,P.paniscus:6.4)'14':2.2)'13':6.6,P.pygmaeus:15.2):0.0;
```

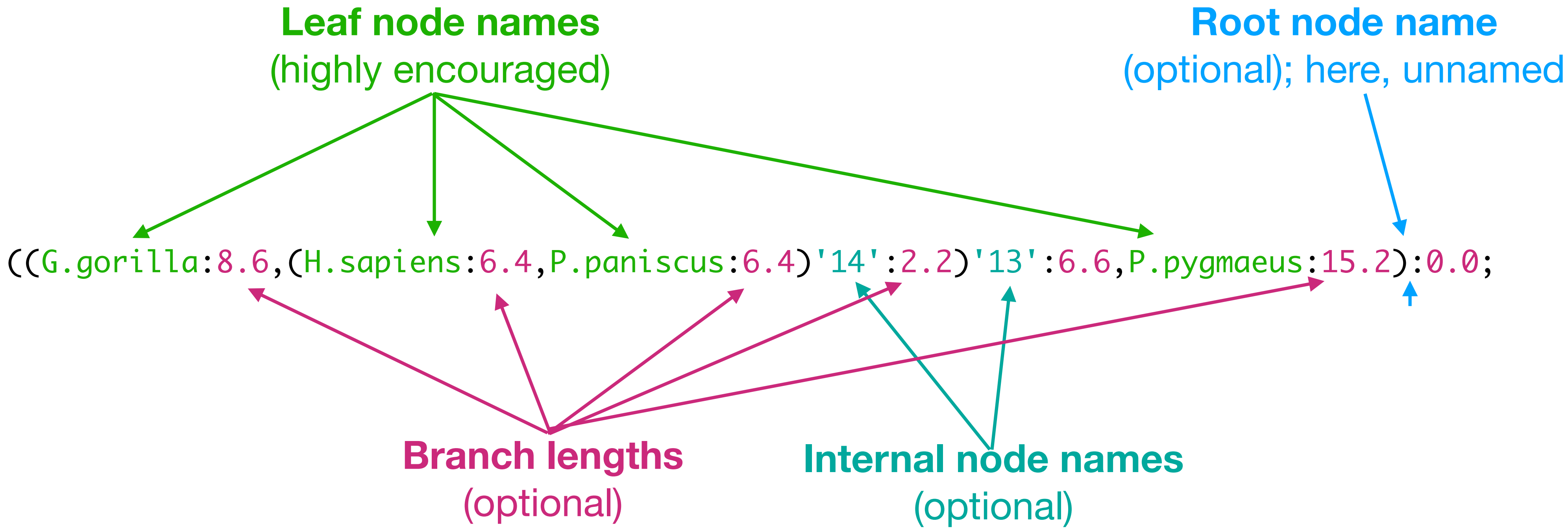


Phylogenetic tree estimators: PhyML, FastTree, RAxML, mashtree

NEWICK

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

NEWICK: Newick tree format (file suffix: .nwk, .newick)



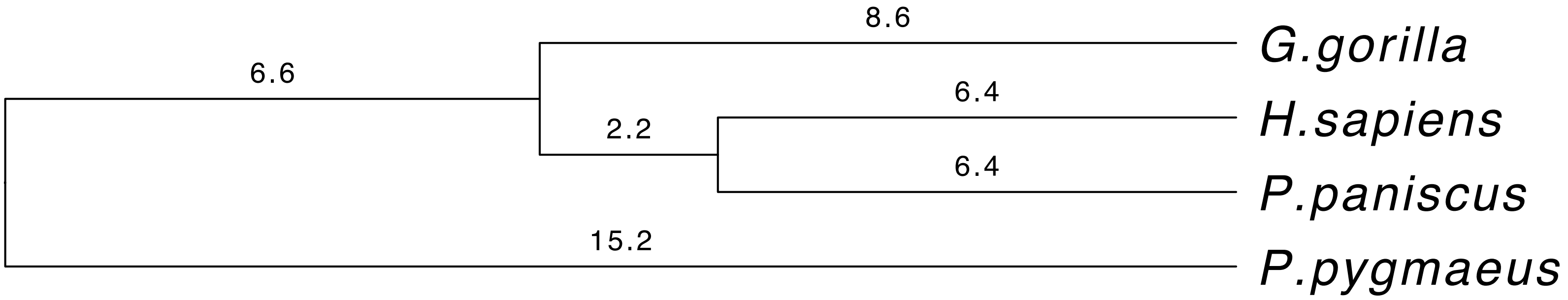
Leaf node names
(highly encouraged)

Root node name
(optional); here, unnamed

```
((G.gorilla:8.6,(H.sapiens:6.4,P.paniscus:6.4)'14':2.2)'13':6.6,P.pygmaeus:15.2):0.0;
```

Branch lengths
(optional)

Internal node names
(optional)



Multi-line records

FASTA/Pearson

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

FASTA: Pearson FASTA format (file suffix: .fasta, .fa, .fna, .faa, .fas, .ffn, .frn, .mpfa)

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

Aligners: FASTA, BLAST, MUSCLE, BWA; Genome browsers; Your code!

FASTA/Pearson

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

"greater than"
Start of record

Whitespace only required if description present

Description/Comment
Optional;
Free-form text

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

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

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

Should be
wrapped flush,
but sometimes is
not

FASTQ

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

FASTQ: Pearson FASTA format (file suffix: .fastq, .fq, .fnq)

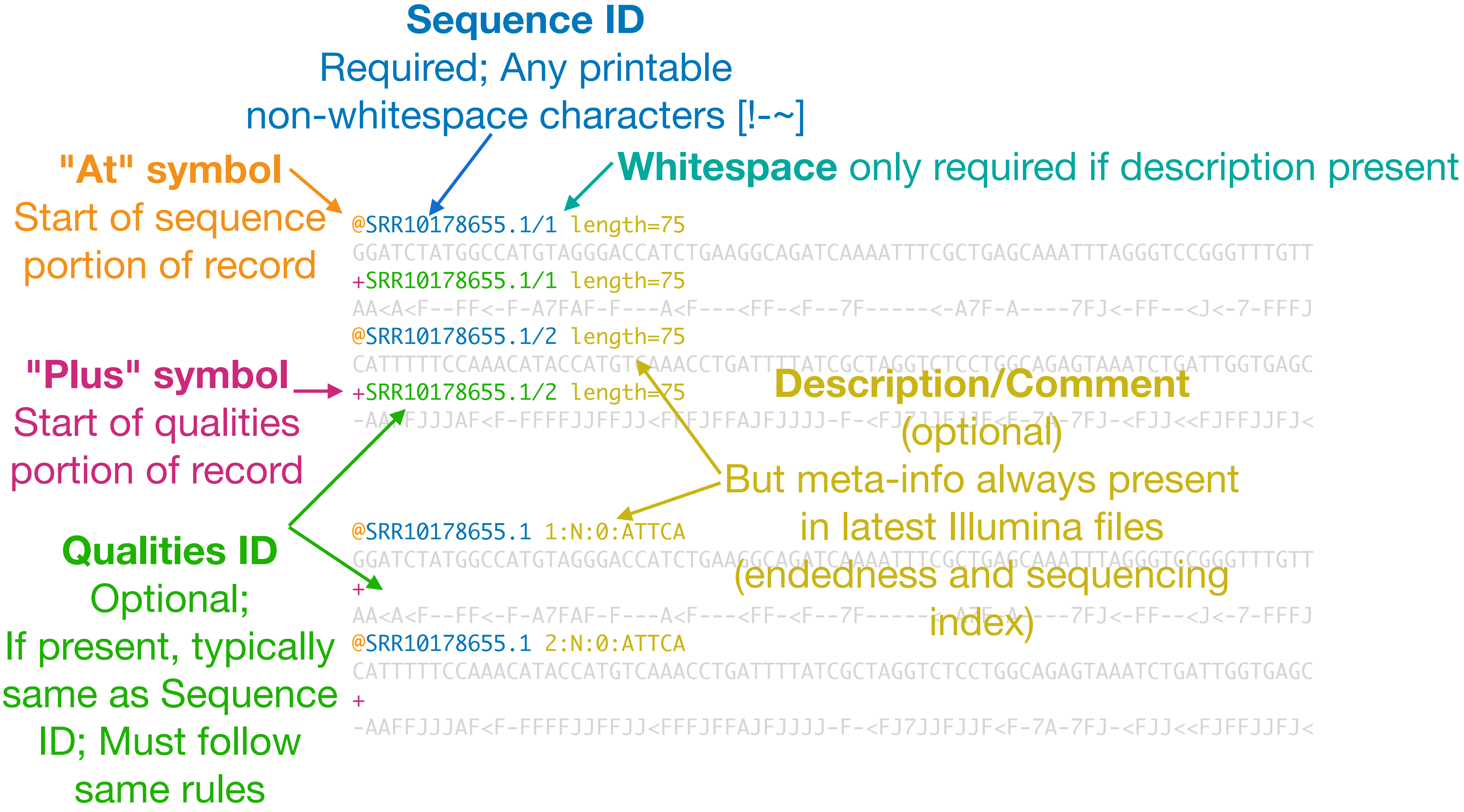
```
@SRR10178655.1/1
GGATCTATGGCCATGTAGGGACCATCTGAAGGCAGATCAAAATTTCGCTGAGCAAATTTAGGGTCCGGGTTTGTT
+
AA<A<F--FF<-F-A7FAF-F---A<F---<FF-<F--7F-----<-A7F-A----7FJ<-FF--<J<-7-FFFJ
@SRR10178655.1/2
CATTTTTCCAAACATACCATGTCAAACCTGATTTTATCGCTAGGTCTCCTGGCAGAGTAAATCTGATTGGTGAGC
+
-AAFFJJJAF<F-FFFFJJFFJJ<FFFJFFAJFJJJJ-F-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

```
@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<
```

High-throughput aligners, such as BWA, STAR, bowtie2

FASTQ

FASTQ Sequence & Quality Headers: Sequence ID & Description on same line; first and third lines of a record.



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

Whitespace only required if description present

"At" symbol
Start of sequence
portion of record

"Plus" symbol
Start of qualities
portion of record

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

Description/Comment
(optional)
But meta-info always present
in latest Illumina files
(endedness and sequencing
index)

```
@SRR10178655.1/1 length=75
GGATCTATGGCCATGTAGGGACCATCTGAAGGCAGATCAAAATTTCGCTGAGCAAATTTAGGGTCCGGGTTTGTT
+SRR10178655.1/1 length=75
AA<A<F--FF<-F-A7FAF-F---A<F---<FF-<F--7F-----<-A7F-A----7FJ<-FF--<J<-7-FFFJ
@SRR10178655.1/2 length=75
CATTTTTCCAAACATACCATGTCAAACCTGATTTTATCGCTAGGTCTCCTGGCAGAGTAAATCTGATTGGTGAGC
+SRR10178655.1/2 length=75
-AAFFJJJAF<F-FFFFJJFFJJ<FFFJFFAJFJJJJ-F-<FJ7JJFJJF<F-7A-7FJ-<FJJ<<FJFFJJFJ<


@SRR10178655.1 1:N:0:ATTCA
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:ATTCA
CATTTTTCCAAACATACCATGTCAAACCTGATTTTATCGCTAGGTCTCCTGGCAGAGTAAATCTGATTGGTGAGC
+
-AAFFJJJAF<F-FFFFJJFFJJ<FFFJFFAJFJJJJ-F-<FJ7JJFJJF<F-7A-7FJ-<FJJ<<FJFFJJFJ<
```

FASTQ

Paired/Mated FASTQ files: Paired-end or mate-pair reads share same ID, but may have endedness appended to ID or in comment/description.

Older format:

Read 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<
```

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

Newer format:

```
@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

FASTQ

FASTQ Sequence and Qualities: sequence string on second line of record, qualities on fourth line of record.

```
@SRR10178655.1/1 length=75
GGATCTATGGCCATGTAGGGACCATCTGAAGGCAGATCAAAATTTCGCTGAGCAAATTTAGGGTCCGGGTTTGTT
+
AA<A<F--FF<-F-A7FAF-F---A<F---<FF-<F--7F-----<-A7F-A----7FJ<-FF--<J<-7-FFFJ
@SRR10178655.1/2 length=75
CATTTTTCCAAACATACCATGTCAAACCTGATTTTATCGCTAGGTCTCCTGGCAGAGTAAATCTGATTGGTGAGC
+
-AAFFJJJAF<F-FFFFJJFFJJ<FFFJFFAJFJJJJ-F-<FJ7JJFJJF<F-7A-7FJ-<FJJ<<FJFFJJFJ<
```

```
@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<
```

FASTQ Sequence
Nucleotide, amino acid, IUPAC codes

Should *NOT* be wrapped flush

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

Must be same length as sequence.

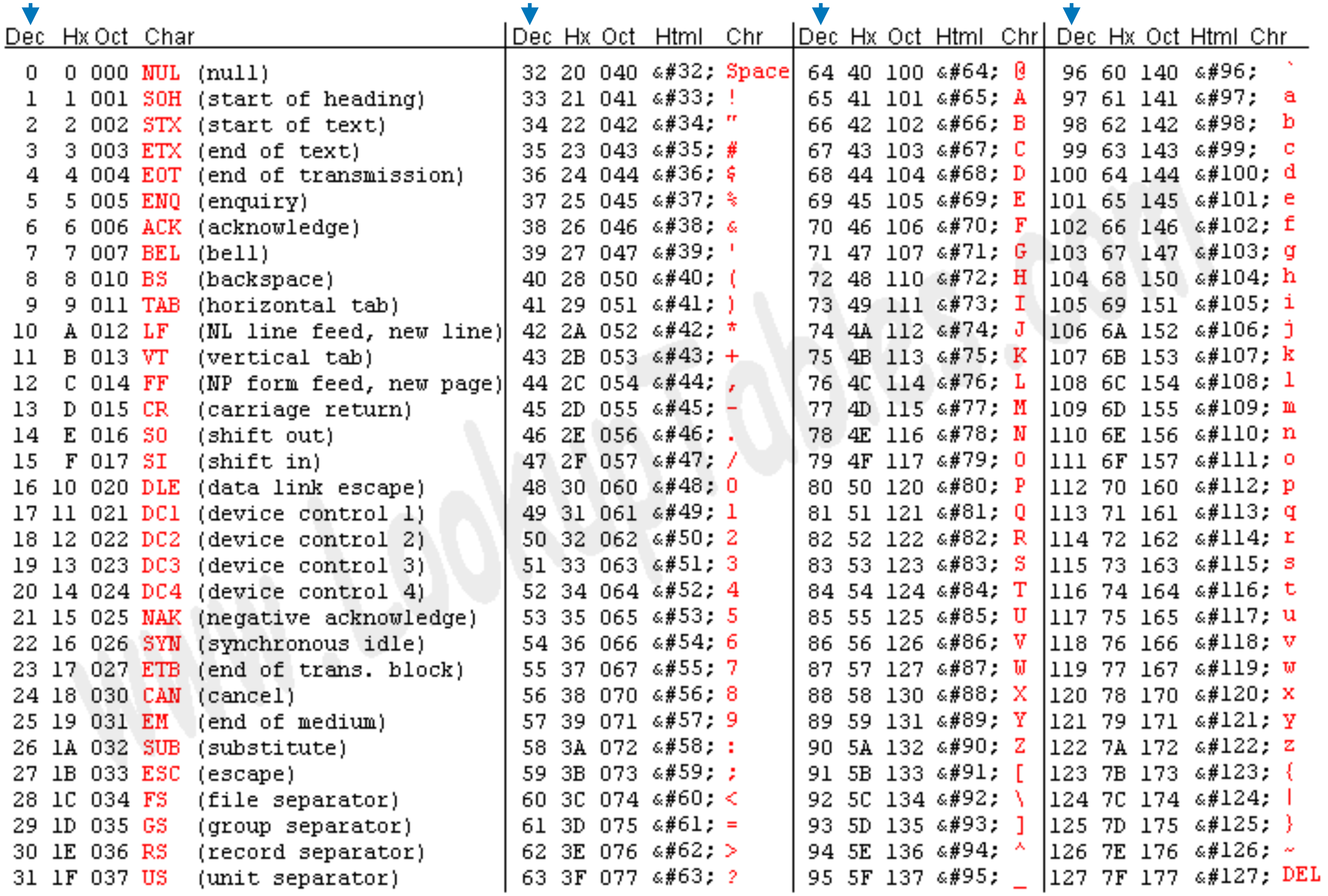
Should *NOT* be wrapped flush

PHRED encoding

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

P = fractional probability that the base call is wrong

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

Source: www.LookupTables.com

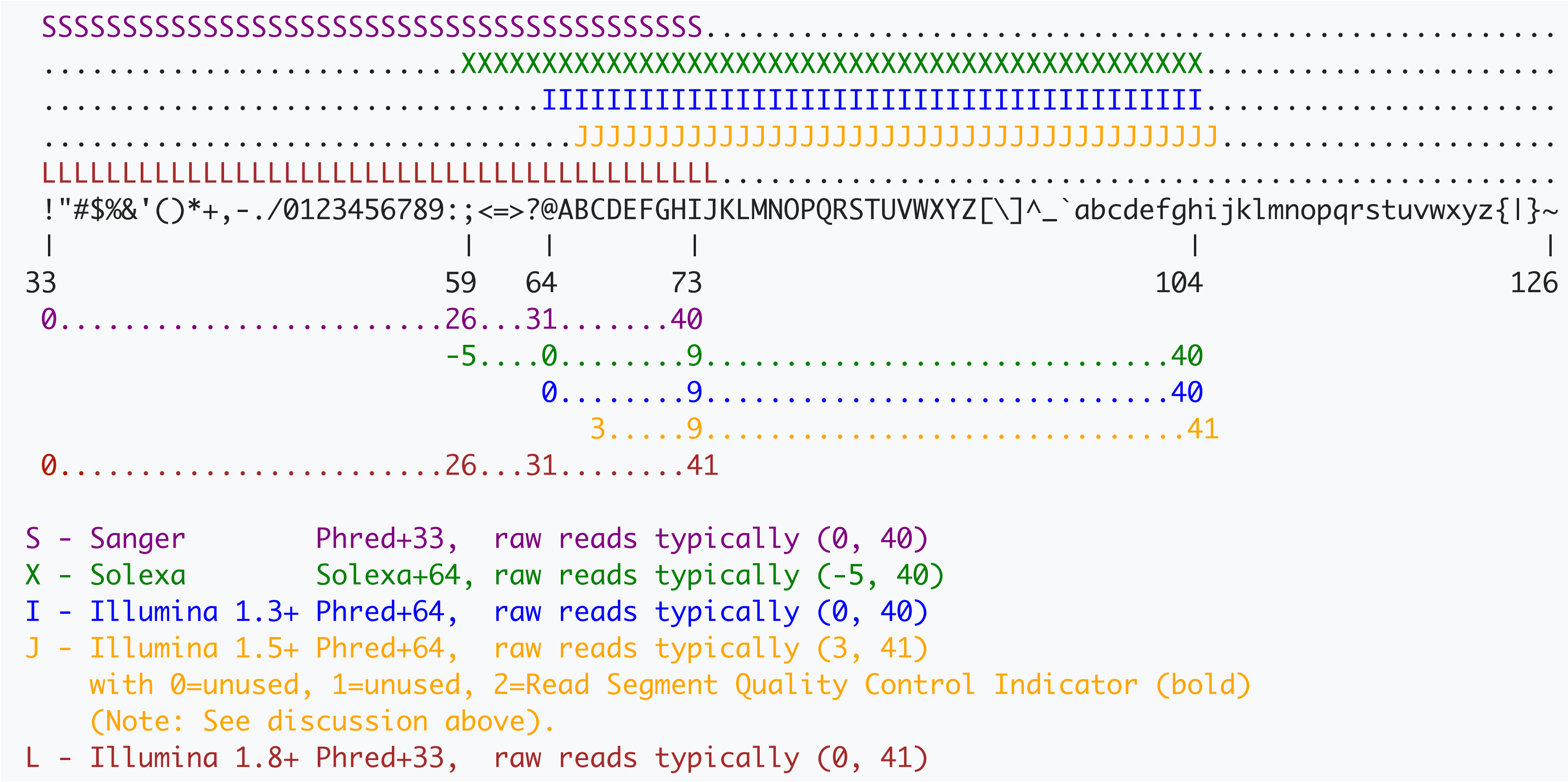
P	Phred
1×10^{0}	0
1×10^{-1}	10
1×10^{-2}	20
1×10^{-3}	30
1×10^{-4}	40
1×10^{-5}	50
1×10^{-6}	60

PHRED encoding

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

P = fractional probability that the base call is wrong

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



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9..............................41
 0........................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

GenBank

https://www.ncbi.nlm.nih.gov/genbank/samplerecord

GenBank: GenBank format (file suffix: .gb)

```
LOCUS       NM_001349598   32 bp    mRNA    linear   PLN 20-OCT-2022
DEFINITION  Arabidopsis thaliana uncharacterized protein (AT4G12485),
            partial mRNA.
ACCESSION   NM_001349598
VERSION     NM_001349598.1
DBLINK      BioProject: PRJNA116
            BioSample: SAMN03081427
KEYWORDS    RefSeq.
SOURCE      Arabidopsis thaliana (thale cress)
  ORGANISM  Arabidopsis thaliana
            Eukaryota; ... Arabidopsis.
REFERENCE   1  (bases 1 to 32)
  AUTHORS   Mayer,K., Schuller,C., ... and McCombie,W.R.
  TITLE     Sequence and analysis of chromosome 4 of the plant
            Arabidopsis thaliana
  JOURNAL   Nature 402 (6763), 769-777 (1999)
  REMARK    Protein update by submitter
COMMENT     REVIEWED REFSEQ: This record has been curated by TAIR ...
            COMPLETENESS: incomplete on the 3' end.
```

(Continues on right)

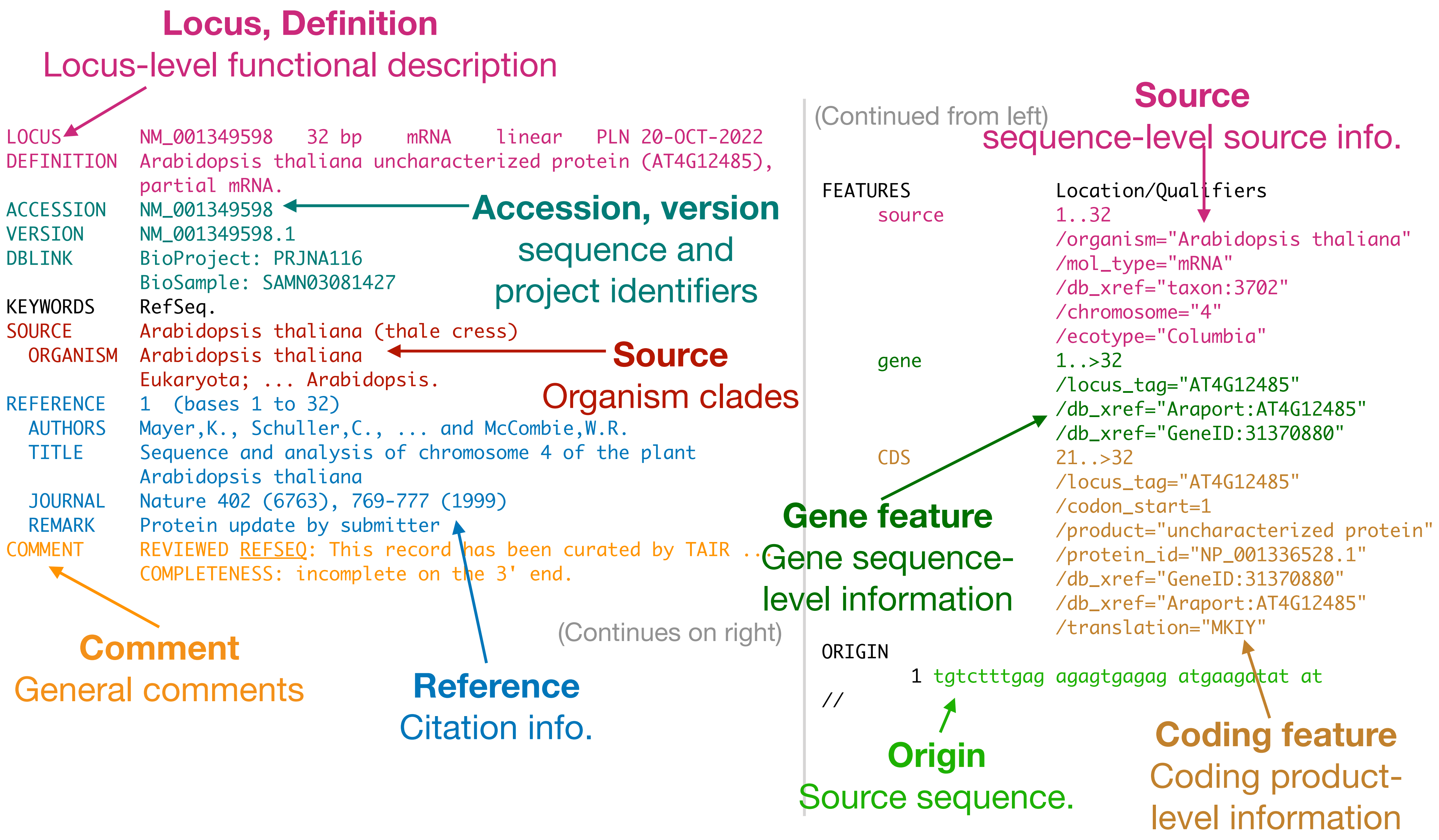
(Continued from left)

```
FEATURES             Location/Qualifiers
     source          1..32
                     /organism="Arabidopsis thaliana"
                     /mol_type="mRNA"
                     /db_xref="taxon:3702"
                     /chromosome="4"
                     /ecotype="Columbia"
     gene            1..>32
                     /locus_tag="AT4G12485"
                     /db_xref="Araport:AT4G12485"
                     /db_xref="GeneID:31370880"
     CDS             21..>32
                     /locus_tag="AT4G12485"
                     /codon_start=1
                     /product="uncharacterized protein"
                     /protein_id="NP_001336528.1"
                     /db_xref="GeneID:31370880"
                     /db_xref="Araport:AT4G12485"
                     /translation="MKIY"
ORIGIN
        1 tgtctttgag agagtgagag atgaagatat at
//
```

Annotation tools: GenBank, ANTISMASH

GenBank

https://www.ncbi.nlm.nih.gov/genbank/samplerecord



Locus, Definition
Locus-level functional description

```
LOCUS       NM_001349598    32 bp    mRNA    linear   PLN 20-OCT-2022
DEFINITION  Arabidopsis thaliana uncharacterized protein (AT4G12485),
            partial mRNA.
ACCESSION   NM_001349598
VERSION     NM_001349598.1
DBLINK      BioProject: PRJNA116
            BioSample: SAMN03081427
KEYWORDS    RefSeq.
SOURCE      Arabidopsis thaliana (thale cress)
  ORGANISM  Arabidopsis thaliana
            Eukaryota; ... Arabidopsis.
REFERENCE   1  (bases 1 to 32)
  AUTHORS   Mayer,K., Schuller,C., ... and McCombie,W.R.
  TITLE     Sequence and analysis of chromosome 4 of the plant
            Arabidopsis thaliana
  JOURNAL   Nature 402 (6763), 769-777 (1999)
  REMARK    Protein update by submitter
COMMENT     REVIEWED REFSEQ: This record has been curated by TAIR ...
            COMPLETENESS: incomplete on the 3' end.
```

Accession, version
sequence and project identifiers

Source
Organism clades

Comment
General comments

Reference
Citation info.

(Continues on right)

(Continued from left)

Source
sequence-level source info.

```
FEATURES             Location/Qualifiers
     source          1..32
                     /organism="Arabidopsis thaliana"
                     /mol_type="mRNA"
                     /db_xref="taxon:3702"
                     /chromosome="4"
                     /ecotype="Columbia"
     gene            1..>32
                     /locus_tag="AT4G12485"
                     /db_xref="Araport:AT4G12485"
                     /db_xref="GeneID:31370880"
     CDS             21..>32
                     /locus_tag="AT4G12485"
                     /codon_start=1
                     /product="uncharacterized protein"
                     /protein_id="NP_001336528.1"
                     /db_xref="GeneID:31370880"
                     /db_xref="Araport:AT4G12485"
                     /translation="MKIY"
ORIGIN
        1 tgtctttgag agagtgagag atgaagatat at
//
```

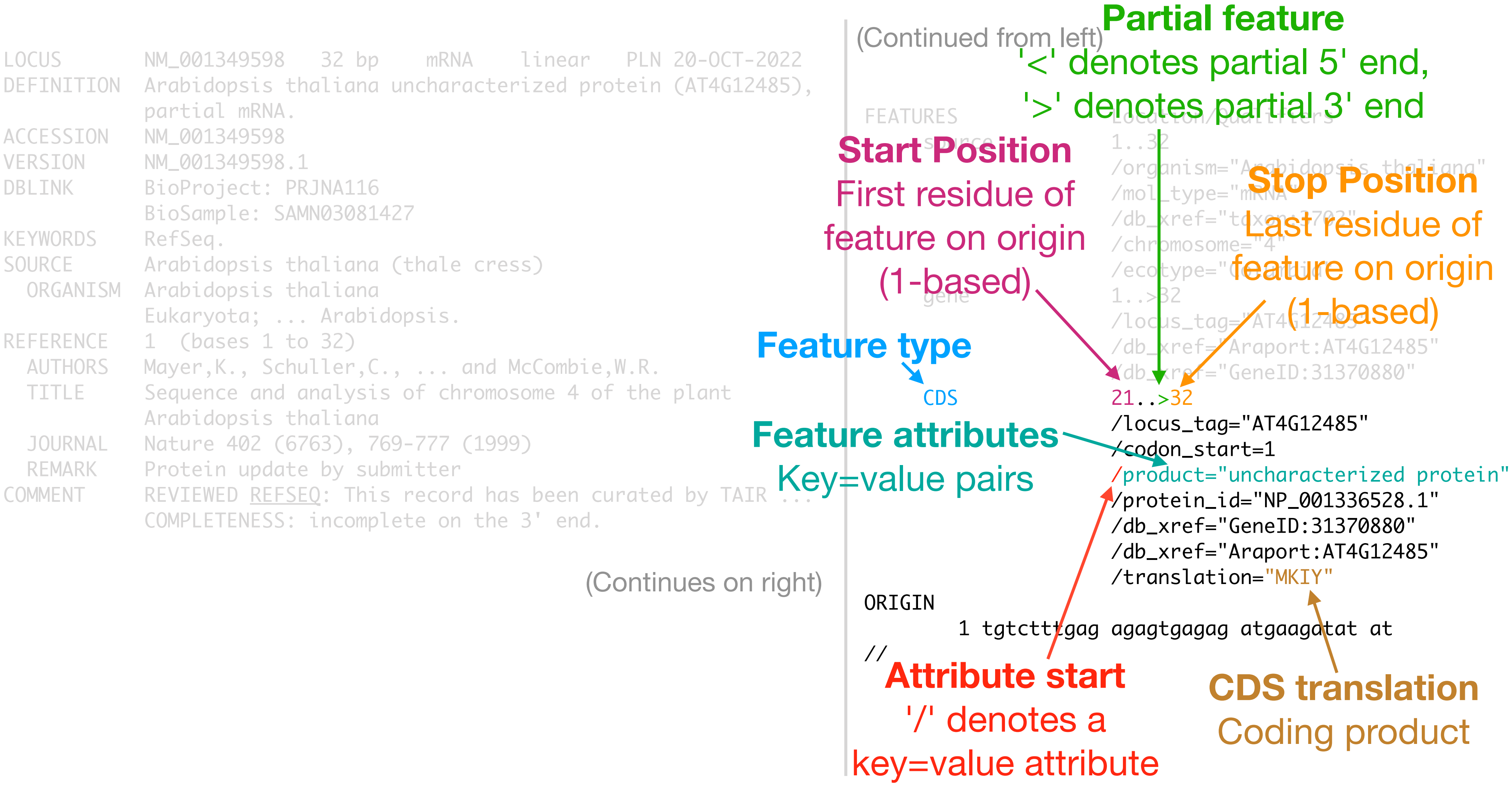
Gene feature
Gene sequence-level information

Origin
Source sequence.

Coding feature
Coding product-level information

GenBank

https://www.ncbi.nlm.nih.gov/genbank/samplerecord



```
LOCUS       NM_001349598    32 bp    mRNA    linear   PLN 20-OCT-2022
DEFINITION  Arabidopsis thaliana uncharacterized protein (AT4G12485),
            partial mRNA.
ACCESSION   NM_001349598
VERSION     NM_001349598.1
DBLINK      BioProject: PRJNA116
            BioSample: SAMN03081427
KEYWORDS    RefSeq.
SOURCE      Arabidopsis thaliana (thale cress)
  ORGANISM  Arabidopsis thaliana
            Eukaryota; ... Arabidopsis.
REFERENCE   1  (bases 1 to 32)
  AUTHORS   Mayer,K., Schuller,C., ... and McCombie,W.R.
  TITLE     Sequence and analysis of chromosome 4 of the plant
            Arabidopsis thaliana
  JOURNAL   Nature 402 (6763), 769-777 (1999)
  REMARK    Protein update by submitter
COMMENT     REVIEWED REFSEQ: This record has been curated by TAIR ...
            COMPLETENESS: incomplete on the 3' end.
```

(Continues on right)

(Continued from left)

```
FEATURES             Location/Qualifiers
     source          1..32
                     /organism="Arabidopsis thaliana"
                     /mol_type="mRNA"
                     /db_xref="taxon:3702"
                     /chromosome="4"
                     /ecotype="Columbia"
     gene            1..>32
                     /locus_tag="AT4G12485"
                     /db_xref="Araport:AT4G12485"
                     /db_xref="GeneID:31370880"
     CDS             21..>32
                     /locus_tag="AT4G12485"
                     /codon_start=1
                     /product="uncharacterized protein"
                     /protein_id="NP_001336528.1"
                     /db_xref="GeneID:31370880"
                     /db_xref="Araport:AT4G12485"
                     /translation="MKIY"
ORIGIN
        1 tgtctttgag agagtgagag atgaagatat at
//
```

Partial feature
'<' denotes partial 5' end,
'>' denotes partial 3' end

Start Position
First residue of feature on origin (1-based)

Stop Position
Last residue of feature on origin (1-based)

Feature type

Feature attributes
Key=value pairs

Attribute start
'/' denotes a key=value attribute

CDS translation
Coding product

JSON

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

https://ecma-international.org/publications-and-standards/standards/ecma-404

JSON: JavaScript Object Notation (file extension: .json)

```
{
  "summary": "json format great for fast machine consumption; highly suitable for
configuration data.",
  "arguments": [
    {
      "name": "--my-first-flag",
      "summary": "",
      "options": [ ]
    },
    {
      "name": "--my-second-flag",
      "summary": "",
      "options": [ ]
    }
  ]
}
```

Used by genome browsers, configuration

JSON

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

https://ecma-international.org/publications-and-standards/standards/ecma-404

JSON: JavaScript Object Notation (file extension: .json)

Top-level dict

Essentially just a dict

```
{
  "summary": "json format great for fast machine consumption; highly suitable for configuration data.",
  "arguments": [
    {
      "name": "--my-first-flag",
      "summary": "",
      "options": [ ]
    },
    {
      "name": "--my-second-flag",
      "summary": "",
      "options": [ ]
    }
  ]
}
```

Nested list

containing two dict items

Key-value pairs

Value can be any type; strings are quoted

Nested dicts

XML

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

https://www.w3.org/TR/xml

XML: eXtensible Markup Language (file extension: .xml)

```
<?xml version="1.0"?>
<catalog>
   <book id="bk101">
      <author type="string">Gambardella, Matthew</author>
      <title type="string">XML Developer's Guide</title>
      <genre type="string">Computer</genre>
      <price type="float">44.95</price>
      <publish_date type="date">2000-10-01</publish_date>
      <description type="string">An in-depth look at creating applications
      with XML.</description>
   </book>
   <book id="bk102">
      <author type="string">Ralls, Kim</author>
      <title type="string">Midnight Rain</title>
      <genre type="string">Fantasy</genre>
      <price type="float">5.95</price>
      <publish_date type="date">2000-12-16</publish_date>
      <description type="string">A former architect battles corporate zombies,
      an evil sorceress, and her own childhood to become queen of the
      world.</description>
   </book>
   <!--book>Commented book</book-->
</catalog>
```

Output by BLAST and InterProScan; used for web sites

XML

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

https://www.w3.org/TR/xml

XML: eXtensible Markup Language (file extension: .xml)

```xml
<?xml version="1.0"?>
<catalog>
   <book id="bk101">
      <author type="string">Gambardella, Matthew</author>
      <title type="string">XML Developer's Guide</title>
      <genre type="string">Computer</genre>
      <price type="float">44.95</price>
      <publish_date type="date">2000-10-01</publish_date>
      <description type="string">An in-depth look at creating applications
      with XML.</description>
   </book>
   <book id="bk102">
      <author type="string">Ralls, Kim</author>
      <title type="string">Midnight Rain</title>
      <genre type="string">Fantasy</genre>
      <price type="float">5.95</price>
      <publish_date type="date">2000-12-16</publish_date>
      <description type="string">A former architect battles corporate zombies,
      an evil sorceress, and her own childhood to become queen of the
      world.</description>
   </book>
   <!--book>Commented book</book-->
</catalog>
```

XML declaration
(optional)

Opening tag
Before content

Content
Text, space doesn't matter

Closing tag
After content

Attribute
(Optional) name=value

Element
Open tag + content + close tag

Commented element
Uses exclamation mark and double-dashes

YAML

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

https://yaml.org/spec

YAML: Yet Another Markup Language (file suffix: .yaml)

```
pi: 3.14159
xmas: true
french-hens: 3
calling-birds:
  - huey
  - dewey
  - louie
  - fred
plumbers: [Mario, Luigi]
xmas-fifth-day:
  calling-birds: four
  french-hens: 3
  golden-rings: 5
  partridges:
    count: 1
    location: "a pear tree"
  turtle-doves: two
```

Used for configuration and metadata files

YAML

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

https://yaml.org/spec

YAML: Yet Another Markup Language (file suffix: .yaml)

Key-value pair
Top-level, with float value

Key-value pair
Top-level, with boolean value

```
pi: 3.14159
xmas: true
french-hens: 3
calling-birds:
  - huey
  - dewey
  - louie
  - fred
plumbers: [Mario, Luigi]
xmas-fifth-day:
  calling-birds: four
  french-hens: 3
  golden-rings: 5
  partridges:
    count: 1
    location: "a pear tree"
  turtle-doves: two
```

List elements
Nested under 'calling-birds' key

List elements
Nested under 'plumbers' key

key-value pairs
nested under 'partridges' key

Binary file types

- SAM => BAM & CRAM
- VCF => BCF
- Compression:
    - Reasonable compression ratio-to-speed, most common:
        - gzip & bgzip (.gz)
    - Better compression ratio, slow:
        - bzip2 (.bz2)
        - xz/lzma (.xz)
    - Archiving files and folders:
        - zip, tar (.zip and .tar, respectively)

Common file issues

- Non-printable characters
- Non-ASCII (e.g., unicode) encoded characters
- Incorrect formatting (e.g., spaces instead of tabs)
- Truncated files

Check file completeness and find special/hidden characters

- Verify file completeness:
  - `md5sum` - used to verify file completeness
- vi/vim:
  - `set list`
- Unix/Linux:
  - `od –c <filename>`
  - `cat –etv <filename>`

Resources

File manipulation

pysam	FASTA/Q, BED, BAM/CRAM/SAM, B/VCF	https://pysam.readthedocs.io/en/latest/api.html#sam-bam-cram-files
pybedtools	BED/GFF/VCF	https://daler.github.io/pybedtools
BioPython	GenBank, NEWICK, more.	https://biopython.org
pyFaidx	FASTA	https://doi.org/10.7287/peerj.preprints.970v1
json	JSON	https://docs.python.org/3/library/json.html
xml.etree.ElementTree	XML	https://docs.python.org/3/library/xml.etree.elementtree.html
PyYAML	YAML	https://pyyaml.org/wiki/PyYAMLDocumentation
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	B/VCF	https://samtools.github.io/bcftools
genometools	FASTA/Q, GFF, GTF	http://genometools.org
gffread & gffcompare	GFF, GTF	https://github.com/gpertea/gffread https://github.com/gpertea/gffcompare
samtools	FASTA/Q, B/SAM	https://github.com/samtools/samtools
vcftools	B/VCF	https://vcftools.github.io/man_latest.html
Picard	FASTA/Q, BED, B/CR/SAM, B/VCF	https://broadinstitute.github.io/picard/

Python module **Command-line tool**

Resources

Alignment

minimap2	FASTA/Q	https://github.com/lh3/minimap2
miniprot	FASTA	https://github.com/lh3/miniprot
BWA	FASTA/Q	https://github.com/lh3/bwa
hisat2	FASTA/Q	https://daehwankimlab.github.io/hisat2/
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

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