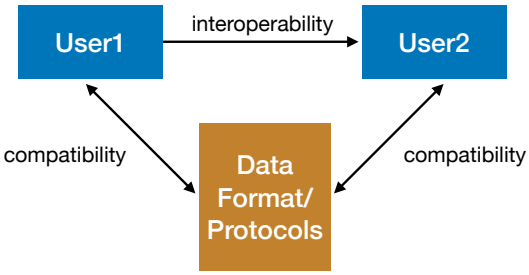
Bioinformatics file formats

Eric Ross
Stowers Institute for Medical Research
&
Jessen Bredeson
DOE Joint Genome Institute

Why are standardized file formats important?

Findable, Accessible, Interoperable, and Reusable (FAIR)



Text vs Binary

Text Formats

- Comma Delimited Text

- Tab-delimited Text
  - BED
  - SAM
  - GFF/GTF
  - VCF
- Multi-line records
  - FASTA
  - FASTQ
  - GENBANK
  - JSON
  - YAML
  - XML
  - HTML

Comma Delimited Text

```
SMED30001362,ACC97187.1,SMEDWI-3
SMED30007406,Q2Q5Y9,SMEDWI-1
SMED30014446,Q2Q5Y8,SMEDWI-2
```

TAB Delimited Text

```
SMED30001362    ACC97187.1    SMEDWI-3
SMED30007406    Q2Q5Y9        SMEDWI-1
SMED30014446    Q2Q5Y8        SMEDWI-2
```

Why I hate CSV

ENSRNOP00000008035 3-oxoacyl-ACP synthase, mitochondrial [Source:RGD Symbol;Acc:1311092]
ENSRNOP00000012271 abhydrolase domain containing 6, acylglycerol lipase [Source:RGD Symbol;Acc:1359323]
ENSRNOP00000074573 abhydrolase domain containing 6, acylglycerol lipase [Source:RGD Symbol;Acc:1359323]
ENSRNOP00000077100 SWI/SNF-related, matrix-associated actin-dependent regulator of chromatin, subfamily a, containing DEAD/H box 1` [Source:RGD Symbol;Acc:1309640]

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
```

WHY?: used for intervals on genomes (Peaks, binding sites, genes etc.)

Tools: bedtools

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
```

WHY?: used for alignment of large numbers of reads (often short reads)

Tools: samtools

GFF3/GTF

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

GFF: Generic Feature Format / **GTF:** Gene Transfer 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
```

WHY?: Gene annotations

Tools: genometools, bedtools

VCF/BCF

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

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
```

WHY?: Variants
Tools: GATK, vcftools, bcftools, picard

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

WHY?: Sequences (nucleotide or peptide)
Tools: BLAST, BLAT . . .

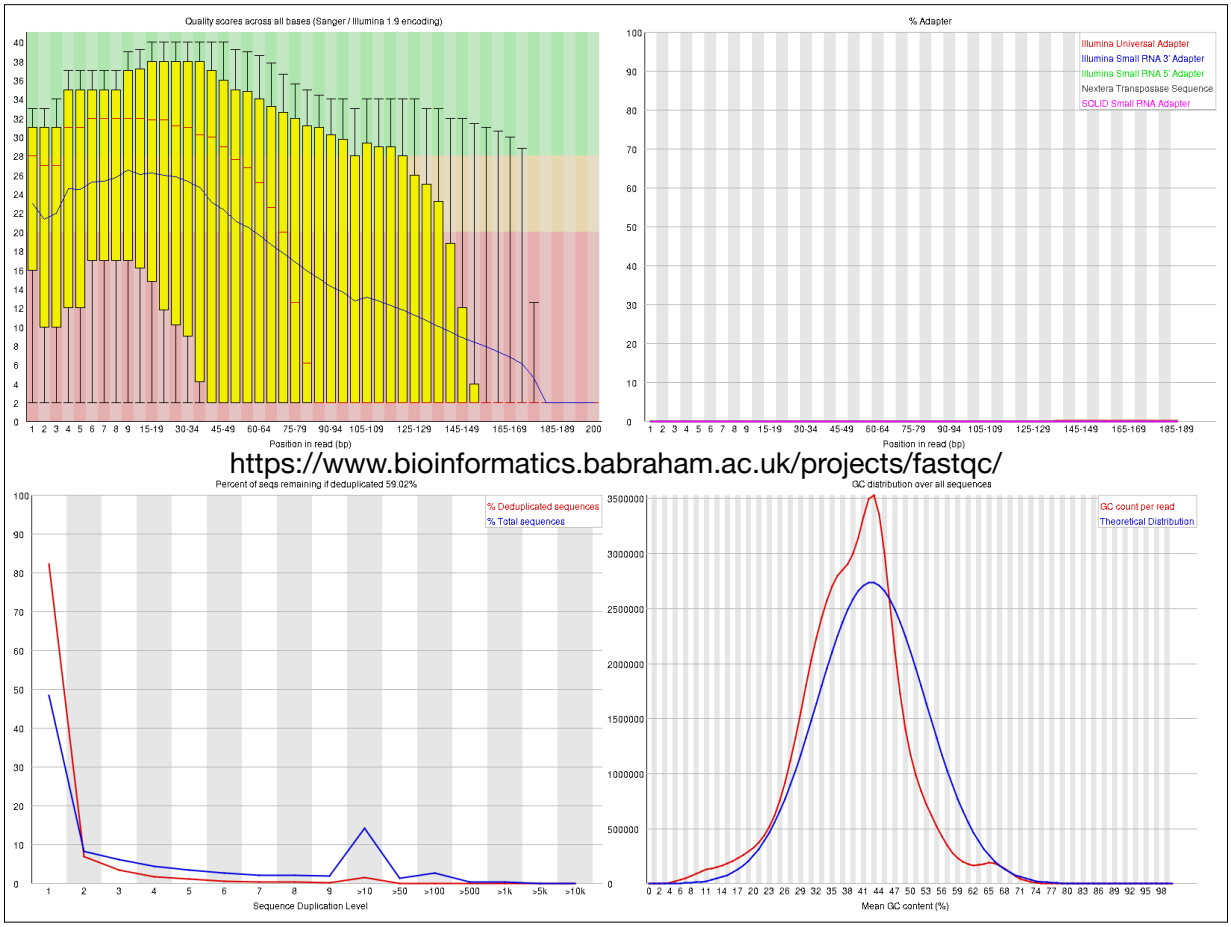
FASTQ

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:
ATATATTAATTAATAATTAATTTATAATAAATATATGATATTAATTAATATATATATATATAATATATTTAATAA
+
AAFFFJFJJJJJJ<FJJJAFJJ<FFAJFJ--FJJJJ-FFF-<FFJFA-FJJ-AJ-<<-FFFAFJJJJJAJ-7---
```

WHY?: Nucleotide sequences, typically output of NextGen sequencing

Tools: FASTQC



Genbank

https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

Genbank: GenBank format

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
```

WHY?: NCBI

Tools: AntiSMASH

JSON

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

JSON: JavaScript Object Notation

```
{
  "first_name": "John",
  "last_name": "Smith"
  "address": {
    "street_address": "21 2nd Street",
    "city": "New York",
    "state": "NY"
  },
  "phone_numbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": [
    "Catherine",
    "Thomas"
  ]
}
```

WHY?: Complex data - kind of human readable (dictionary - list format)

YAML

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

YAML: Yet Another Markup Language

```yaml
receipt:     Oz-Ware Purchase Invoice
date:        2012-08-06
customer:
    first_name:   Dorothy
    family_name:  Gale

items:
    - part_no:   A4786
      descrip:   Water Bucket (Filled)
      price:     1.47
      quantity:  4

    - part_no:   E1628
      descrip:   High Heeled "Ruby" Slippers
      size:      8
      price:     133.7
      quantity:  1
```

WHY?: Complex data - kind of human readable

HTML

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

HTML: HyperText Markup Language

```html
<!DOCTYPE html>
<html>
<body>

<h1>My First Heading</h1>
<p>My first paragraph.</p>

</body>
</html>
```

WHY?: Webpages

Tools: wget, curl, web browsers

XML

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

XML: Extensible Markup Language

```
<note>
<to>Tove</to>
<from>Jani</from>
<heading>Reminder</heading>
<body>Don't forget me this weekend!</body>
</note>
```

WHY?: Complex data - NOT human readable

Tools: BLAST XML output . . .

PDB

https://en.wikipedia.org/wiki/Protein_Data_Bank_(file_format)

PDB: Protein Data Bank

```
MODEL        1
ATOM      1  N   MET A   1     -29.546  -3.540  28.854  1.00 63.37
ATOM      2  CA  MET A   1     -28.922  -2.540  27.955  1.00 63.37
ATOM      3  C   MET A   1     -27.392  -2.399  28.080  1.00 63.37
ATOM      4  CB  MET A   1     -29.594  -1.164  28.132  1.00 63.37
ATOM      5  O   MET A   1     -26.861  -1.476  27.493  1.00 63.37
ATOM      6  CG  MET A   1     -31.092  -1.170  27.815  1.00 63.37
ATOM      7  SD  MET A   1     -31.481  -1.761  26.152  1.00 63.37
ATOM      8  CE  MET A   1     -33.128  -1.019  25.952  1.00 63.37
ATOM      9  N   ARG A   2     -26.638  -3.267  28.784  1.00 75.70
ATOM     10  CA  ARG A   2     -25.156  -3.157  28.855  1.00 75.70
ATOM     11  C   ARG A   2     -24.427  -4.012  27.811  1.00 75.70
ATOM     12  CB  ARG A   2     -24.658  -3.460  30.280  1.00 75.70
ATOM     13  O   ARG A   2     -23.435  -3.579  27.246  1.00 75.70
```

WHY?: Protein Structures

Tools: AlphaFold, ChimeraX, PyMol, foldseek

Binary Formats

- Compressed formats
  - gzip (.gz, tar.gz)
  - others (.bz2, .7z, .xz)
- Bioinformatic Specific
  - Binary SAM (.bam)
  - Binary VCF (.bcf)

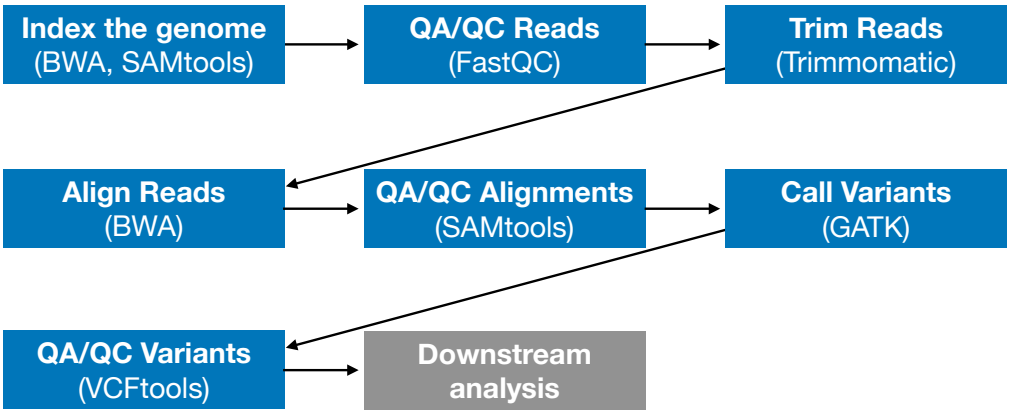
Common file issues

- Non-printable characters
- Non-ASCII encoded characters
- Incorrect formatting (spaces instead of tabs)
- Truncated files

How to find Special Characters

- VI
  - `:set list`
- Unix
  - `od -c filename`
  - `cat -etv filename`

Variant-calling workflow



Reference Slides

Tools / Modules

File manipulation/filtering

pysam	FASTA/Q, BED, B/CR/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	Many	https://biopython.org
pyFaidx	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	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
bamtools	B/SAM	https://github.com/pezmaster31/bamtools
vcftools	B/VCF	https://vcftools.github.io/man_latest.html
Picard	FASTA/Q, BED, B/CR/SAM, B/VCF	https://broadinstitute.github.io/picard/

Tools / Modules

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
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

Tools / Modules

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

FASTQ

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

FASTQ Sequence
Nucleotide, amino acid, IUPAC codes

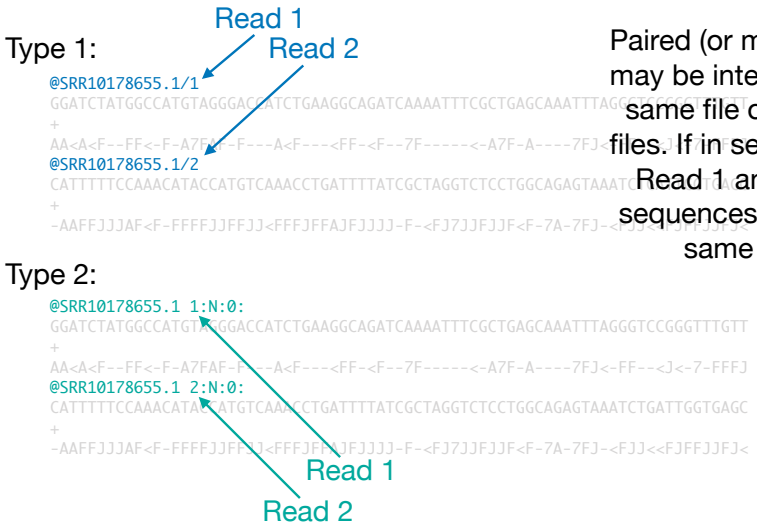
Should *not* be wrapped flush

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

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

FASTQ

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



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: 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.

Should *not* be wrapped flush

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

FASTQ

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

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

FASTQ

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

P = fractional probability that the base call is wrong

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

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 .........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./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

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

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

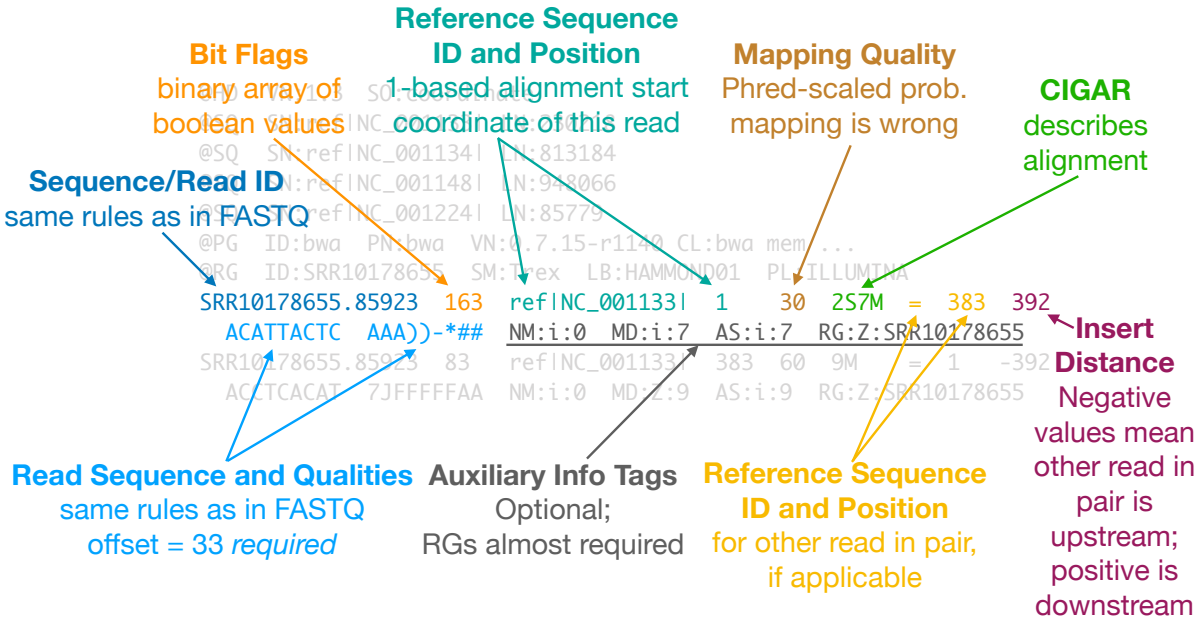
Program processing history (with commands)

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/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.



SAM/BAM/CRAM

CIGAR AND Bitwise flag field details

Useful with `samtools flags` and `samtools view -f -F`

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-GAGG
   ||| |||| ||||| ||
R: ATG-CAGGCCAGATTGATA
```

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

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`		

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: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       513.189  .        DP=36  GT:GQ:AD  2|1:43:0,28,8
Chr1    21[illegible]  GG   GGTATTTTTAG  253.597  .        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)

```
#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       513.189  .        DP=36  GT:GQ:AD  2|1:43:0,28,8
Chr1    21[illegible]  GG   GGTATTTTTAG  253.597  .        DP=64  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">
##INFO=<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=247249719>
##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

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  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  601500  BREAK5  0  +  599500  601500  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

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  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

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

```
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
```

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