

Canadian Bioinformatics Workshops

www.bioinformatics.ca
bioinformaticsdotca.github.io



CC BY-SA 4.0 DEED

Attribution-ShareAlike 4.0 International

Canonical URL: <https://creativecommons.org/licenses/by-sa/4.0/>

[See the legal code](#)

You are free to:

Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

 **Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

 **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.

No additional restrictions — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

Notices:

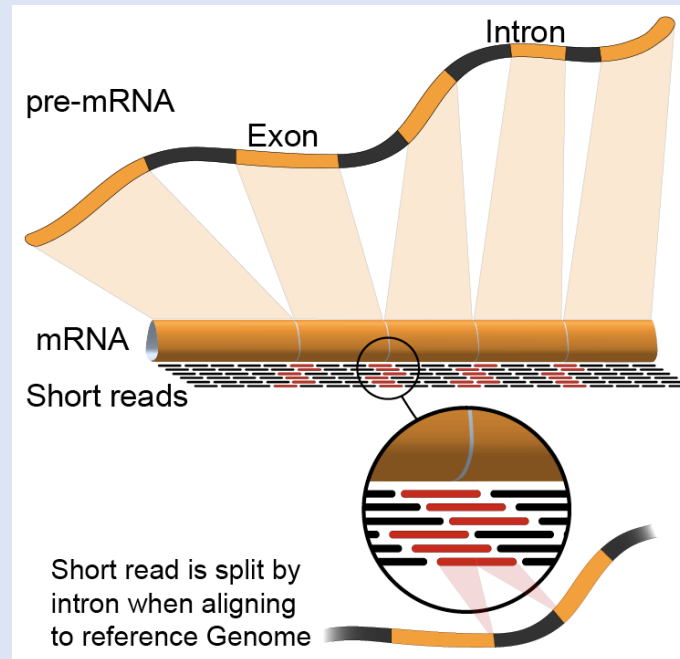
You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable [exception or limitation](#).

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as [publicity, privacy, or moral rights](#) may limit how you use the material.

RNA-Seq Module 1:

FASTA/FASTQ/GTF

Malachi Griffith, Obi Griffith, Isabel Risch, Vida Talebian
RNA-seq Analysis 2024. June 17-19, 2024



Fasta – format for representing nucleic acid or amino acid sequences

```
>AY274119.3 Severe acute respiratory syndrome-related coronavirus isolate  
Tor2, complete genome
```

```
ATATTAGGTTTTTACCTACCCAGGAAAAGCCAACCAACCTCGATCTCTTGTAGATCTGTTCTCTAAACGA  
ACTTTAAAATCTGTGTAGCTGTCGCTCGGCTGCATGCCTAGTGCACCTACGCAGTATAACAATAATAAA  
TTTTACTGTCGTTGACAAGAAACGAGTAACTCGTCCCTCTTCTGCAGACTGCTTACGGTTTCGTCCGTGT  
TGCAGTCGATCATCAGCATACCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTTT  
TTGGTGTCAACGAGAAAACACACGTCCAACCTCAGTTTGCCTGTCCTTCAGGTTAGAGACGTGCTAGTGCG  
TGGCTTCGGGGACTCTGTGGAAGAGGCCCTATCGGAGGCACGTGAACACCTCAAAAATGGCACTTGTGGT  
...
```

```
>FJ882960.1 SARS coronavirus ExoN1 isolate P3pp34, complete genome  
CGATCTCTTGTAGATCTGTTCTCTAAACGAACCTTTAAAATCTGTGTAGCTGTCGCTCGGCTGCATGCCTA  
GTGCACCTACGCAGTATAACAATAATAAATTTTACTGTCGTTGACAAGAAACGAGTAACTCGTCCCTCT  
TCTGCAGACTGCTTACGGTTTCGTCCGTGTTGCAGTCGATCATCAGCATACCTAGGTTTCGTCCGGGTGT  
...
```

First line starts with “>” header or “Comment”; used as a summary/description, often starting with unique accession/identifier

Subsequent lines contain sequence

- Interleaved: sequence broken into multiple lines of characters
- Sequential: entire sequence on a single line

Multiple sequence FASTA obtained by simply concatenating multiple FASTA records together

Fastq – format for representing raw sequence – base calls and quality values

@HWUSI-EAS100R:6:73:941:1973#0/1

CTTTTTTATTTTGTCTGACTGGGTTGATTCAAAA

+

CCCFFFFFFHHHGGJHIIJHIHIIIFHIJJJJIJJGIBBFGE

First line starts with “@” header or “Comment”; followed by sequence identifier and optional description

Sequence line

Spacer line

Quality values

@HWUSI-EAS100R:6:2303:11793:37095#0/1

ATGAATTATAGGGCTGTATTTTAATTTTGCATTTTAA

+

@@??BDDFFF<FHEGFFGGIEBGHIIIIIBEHIIGIH<FHE

Next sequence record

Read naming conventions

@HWUSI-EAS100R:6:73:941:1973#0/1

Instrument ID

Lane

X/Y coords

Index

Pair

Filter
status

@EAS139:136:FC706VJ:2:2104:15343:197393:GATTACT+GTCTTAAC 1:Y:0:ATCACG

Instrument ID

Run

Flowcell

lane

tile

x/y coords

UMI

Pair

Control #

Index

Quality values - Phred scores and ASCII glyphs

Phred Q	Probability (P) of Wrong Base	Base Call Accuracy	Sanger "Q + 33" Shift	Sanger "Q + 33" Shift ASCII glyph
0	1	0	33	!
1	0.794	0.206	34	"
2	0.631	0.369	35	#
10	0.1	0.9	43	+
20	0.01	0.99	53	5
30	0.001	0.999	63	?

Encoding History:

- Sanger Format (shown above): Q of 0 to 93 using ASCII 33 to 126
 - Sanger data, SAM format, Illumina 1.8+
- Solexa/Illumina 1.0: Q of -5 to 62 using ASCII 59 to 126
- Illumina 1.3 to 1.8: Q of 0 to 62 using ASCII 64 to 126
- Illumina 1.5 to 1.7: Phred scores 0 to 2 have a slightly different meaning
- Illumina 1.8+ -> Sanger Format

GFF/GTF - representing sequence features

- GFF – General/Generic Feature Format; Gene Finding Format
 - Two versions in wide use
 - GFF2 (see also GTF)
 - GFF3
 - Added formal support for multiple levels (and direction) of hierarchy (e.g., gene -> transcript -> exon)
- GTF – Gene Transfer Format
 - An extension of GFF2
- GFF2, GFF3 and GTF are all tab-separated files with 9 fields
 - Differing content in 9th column

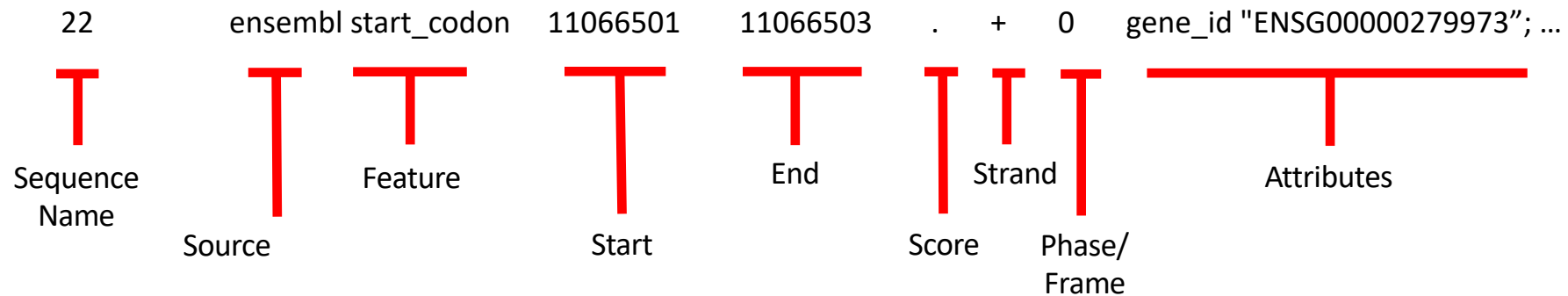
GFF/GTF – general structure

General GFF structure

Position index	Position name	Description
1	sequence	The name of the sequence where the feature is located.
2	source	Keyword identifying the source of the feature, like a program (e.g. Augustus or RepeatMasker) or an organization (like TAIR).
3	feature	The feature type name, like "gene" or "exon". In a well structured GFF file, all the children features always follow their parents in a single block (so all exons of a transcript are put after their parent "transcript" feature line and before any other parent transcript line). In GFF3, all features and their relationships should be compatible with the standards released by the Sequence Ontology Project .
4	start	Genomic start of the feature, with a 1-base offset . This is in contrast with other 0-offset half-open sequence formats, like BED files .
5	end	Genomic end of the feature, with a 1-base offset . This is the same end coordinate as it is in 0-offset half-open sequence formats, like BED files . ^{citation needed}
6	score	Numeric value that generally indicates the confidence of the source on the annotated feature. A value of "." (a dot) is used to define a null value.
7	strand	Single character that indicates the Sense (molecular biology) strand of the feature; it can assume the values of "+" (positive, or 5'→3'), "-", (negative, or 3'→5'), "." (undetermined).
8	phase	phase of CDS features; it can be either one of 0, 1, 2 (for CDS features) or "." (for everything else). See the section below for a detailed explanation.
9	Attributes.	All the other information pertaining to this feature. The format, structure and content of this field is the one which varies the most between the three competing file formats.

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

Ensembl GTF example record



Example of attributes string:

```
gene_id "ENSG00000279973"; gene_version "1"; transcript_id "ENST00000624155"; transcript_version "1";  
exon_number "1"; gene_name "BAGE5"; gene_source "ensembl"; gene_biotype "protein_coding"; transcript_name  
"BAGE5-201"; transcript_source "ensembl"; transcript_biotype "protein_coding"; tag "basic"; transcript_support_level  
"1";
```

Note: there will be many GTF records/rows per transcript per gene (UTRs, start_codon, exons, etc)

We are on a Coffee Break & Networking Session

Workshop Sponsors:



Canadian Centre for
Computational
Genomics



HPC4Health



OICR
Ontario Institute
for Cancer Research



GenomeCanada