

# Canadian Bioinformatics Workshops

[www.bioinformatics.ca](http://www.bioinformatics.ca)

[bioinformaticsdotca.github.io](https://bioinformaticsdotca.github.io)

Supported by



This page is available in the following languages:

Afrikaans Ελληνικά Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto  
 Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)  
 Euskara Suomi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu  
 Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik срpski srpski (latinica) Sotho svenska  
 中文 華語 (台灣) isiZulu



## Attribution-Share Alike 2.5 Canada

### You are free:



to **Share** — to copy, distribute and transmit the work



to **Remix** — to adapt the work



### Under the following conditions:



**Attribution.** You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



**Share Alike.** If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

[Disclaimer](#)

Your fair dealing and other rights are in no way affected by the above.  
 This is a human-readable summary of the Legal Code (the full licence) available in the following languages:  
[English](#) [French](#)

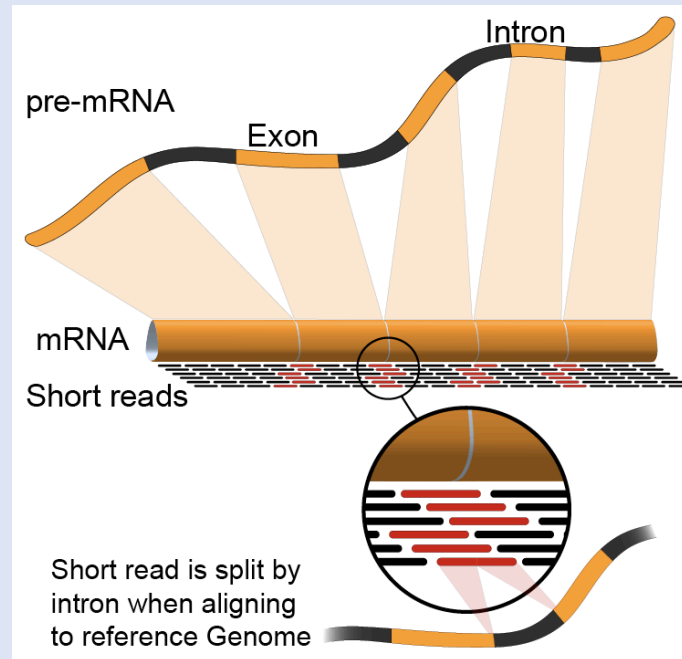
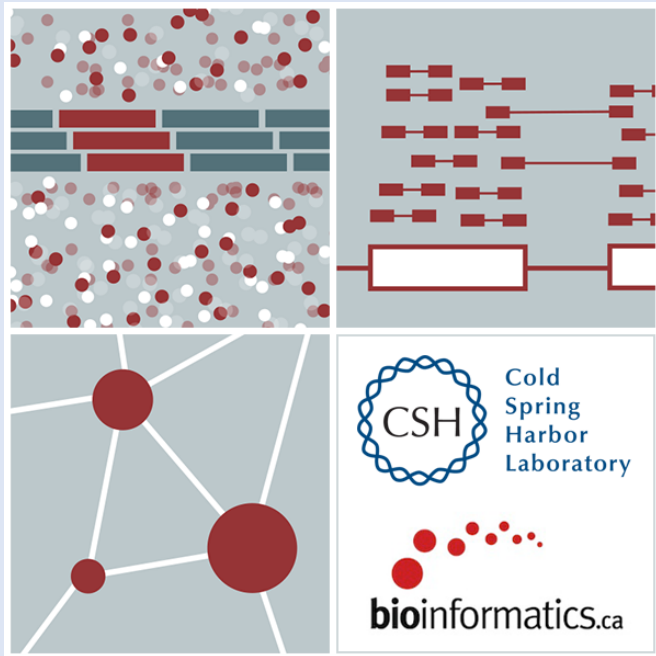
[Learn how to distribute your work using this licence](#)

# FASTA/FASTQ/GTF

Kelsy Cotto, Obi Griffith, Malachi Griffith, Saad Khan, Allegra Petti, Huiming Xia

Informatics for RNA-seq Analysis

June 17-19, 2020



# 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  
CGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTAGCTGTCGCTCGGCTGCATGCCTA  
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

CTTTTTTATTTTTGTCTGACTGGGTTGATTCAAAA

+

CCCFHHHHGJHIJHIHIIIFHIJJJIJGIBBFGE

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

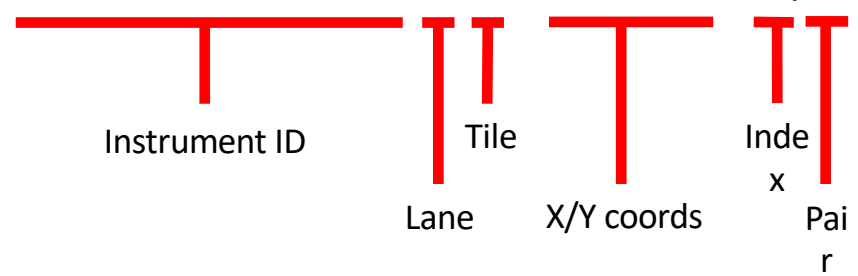
+

@@??BDDFFF<FHEGFFGGIEBGHIIIIIBEHIIGIH<FHE

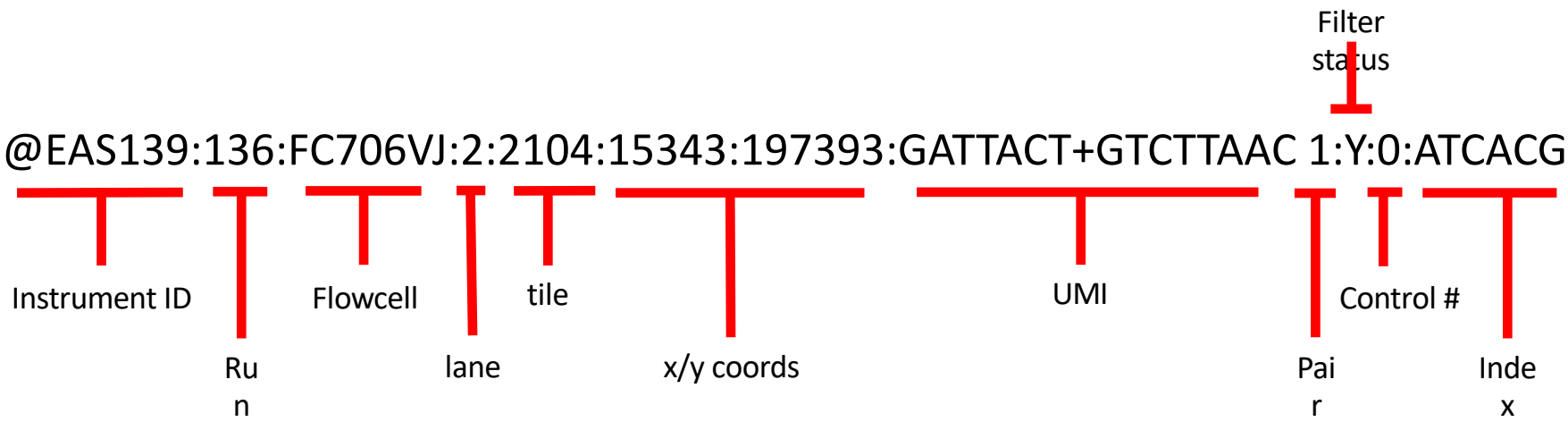
Next sequence record

# Read naming conventions

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



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



# 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 9<sup>th</sup> column



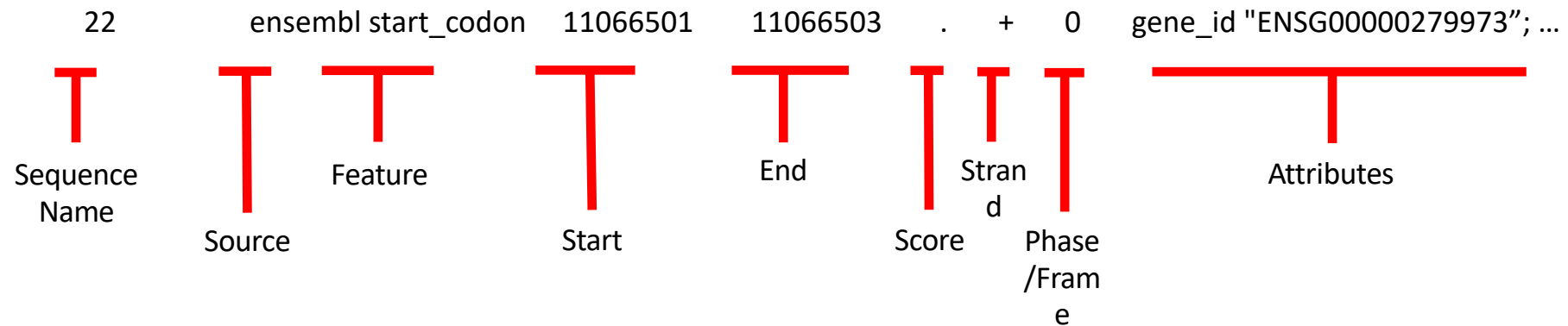
# GFF/GTF – general structure

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. <a href="#">Augustus</a> or <a href="#">RepeatMasker</a> ) or an organization (like <a href="#">TAIR</a> ).
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 <a href="#">standards released by the Sequence Ontology Project</a> .
4	start	Genomic start of the feature, with a <b>1-base offset</b> . This is in contrast with other 0-offset half-open sequence formats, like <a href="#">BED files</a> .
5	end	Genomic end of the feature, with a <b>1-base offset</b> . This is the same end coordinate as it is in 0-offset half-open sequence formats, like <a href="#">BED files</a> . <sup>[<i>citation needed</i>]</sup>
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 <a href="#">Sense (molecular biology) strand</a> 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](https://en.wikipedia.org/wiki/General_feature_format)

# Ensembl GTF example record



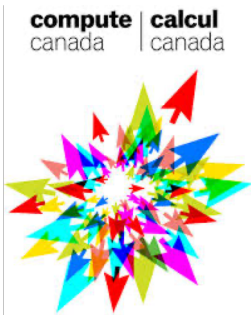
## Example of attributes string:

```
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

