Widely-used genomic file formats

Luigi Grassi < lg490@medschl.cam.ac.uk >
Romina Petersen < rp520@medschl.cam.ac.uk >

TRAIN MALTA SUMMER SCHOOL 2016
September 2016

General information

The following standard icons are used in the hands-on exercises to help you locating:

i Important Information

📄 General information / notes

👣 Follow the following steps

Q Questions to be answered

! Warning – PLEASE take care and read carefully

✸ Optional Bonus exercise

✸ Optional Bonus exercise for a champion

BED file formats

BED (Browser Extensible Data) format provides a flexible way to define the data lines that are displayed in an annotation track. BED lines have three required fields and nine additional optional fields. The first three required BED fields are:

- chrom - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
- chromStart - The starting position of the feature in the chromosome or scaffold.
- chromEnd - The ending position of the feature in the chromosome or scaffold.

The 9 additional optional BED fields are:

- name - Defines the name of the BED line to be displayed.
- score - A score between 0 and 1000.
- strand - Defines the strand - either '+' or '-'.
- thickStart - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays).

- thickEnd - The ending position at which the feature is drawn thickly (for example, the stop codon in gene displays).
- itemRgb - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line itemRgb attribute is set to "On", this RBG value will determine the display color of the data contained in this BED line.
- blockCount - The number of blocks (exons) in the BED line.
- blockSizes - A comma-separated list of the block sizes. The number of items in this list should correspond to blockCount.
- blockStarts - A comma-separated list of block starts. All of the blockStart positions should be calculated relative to chromStart. The number of items in this list should correspond to blockCount.

Manage and visualize bed files

Open the Terminal.

First, go to the folder where the data are stored.

```
cd /data/day2/QC/practical/FileFormats
```

Questions

1. How many bed file are in the folder? Try to use ls and wc to count.
2. How many fields each one contains? Hint: the fields are tab separated, you can use the awk program to count them

   ```
   awk -F"\t" '{print NF}' Bt_6.bed | sort | uniq
   ```

3. How many chromosomes each one refers to? Hint: Use cat or awk to print only the first field of each line, then use sort and uniq
4. Do you figure out which specie each one refers to? Hint: the name field is a NCBI Reference Sequence ID.

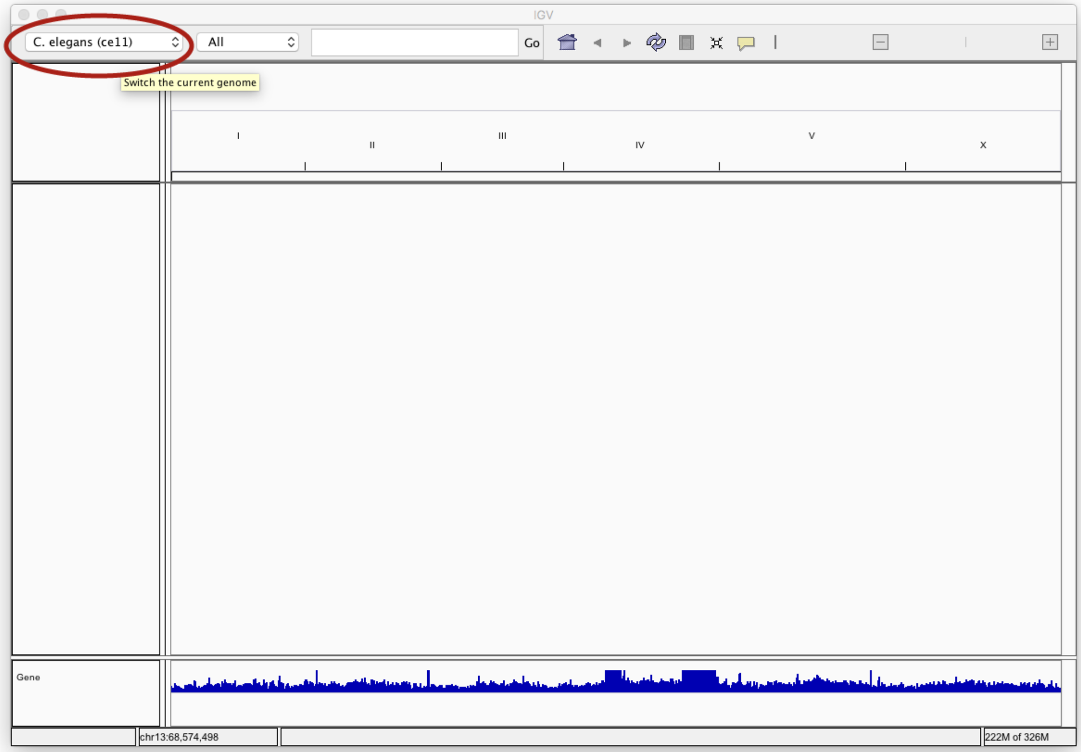
Loading files on IGV

The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations. Because NGS datasets are very large, it is often impossible or inefficient to read them entirely into a computer's memory when searching for a specific piece of data. In order to more quickly retrieve the data we are interested in analyzing or viewing, most programs have a way of treating these data files as databases. Database indexes enable one to rapidly pull specific subsets of the data from them.

Resources used

Integrative Genomics Viewer (IGV) : http://software.broadinstitute.org/software/igv/home

- Launch IGV and switch to the C. elegans genome (release ce11) .



- From the main window of IGV, click on File Load from File Ce_ce11.bed, upon you downloaded it using the Apache server links on the course server.

There are a lot of things you can do in IGV. Here are a few:

- Zoom in using the slider in the upper right.

- Navigate by clicking and dragging in the window. This is how you move left and right along the genome.

- Navigate more quickly. Use page-up page-down, home, end.

- Jump right to a gene. Type its name into the search box. Try "mib-1".

- Change the appearance of genes. Right click on the gene track and try "expanded". Experiment with the other options.

Questions

1. How many transcripts are reported for the gene "cul-6" ?
2. How many are protein-coding ?

GTF file formats

GTF (Gene Transfer Format) is a refinement to the GFF (General Feature Format).

Fields are tab-separated. Also, all but the final field in each feature line must contain a value; "empty" columns should be denoted with a ’.’

- seqname - name of the chromosome or scaffold; chromosome names can be given with or without the ’chr’ prefix.

- source - name of the program that generated this feature, or the data source (database or project name)

- feature - feature type name, e.g. Gene, Variation, Similarity

- start - Start position of the feature.

- end - End position of the feature.

- score - A floating point value.

- strand - defined as + (forward) or - (reverse).

- frame - One of ’0’, ’1’ or ’2’. ’0’ indicates that the first base of the feature is the first base of a codon, ’1’ that the second base is the first base of a codon, and so on..

- attribute - A semicolon-separated list of tag-value pairs, providing additional information about each feature.

Manage and visualize gtf files

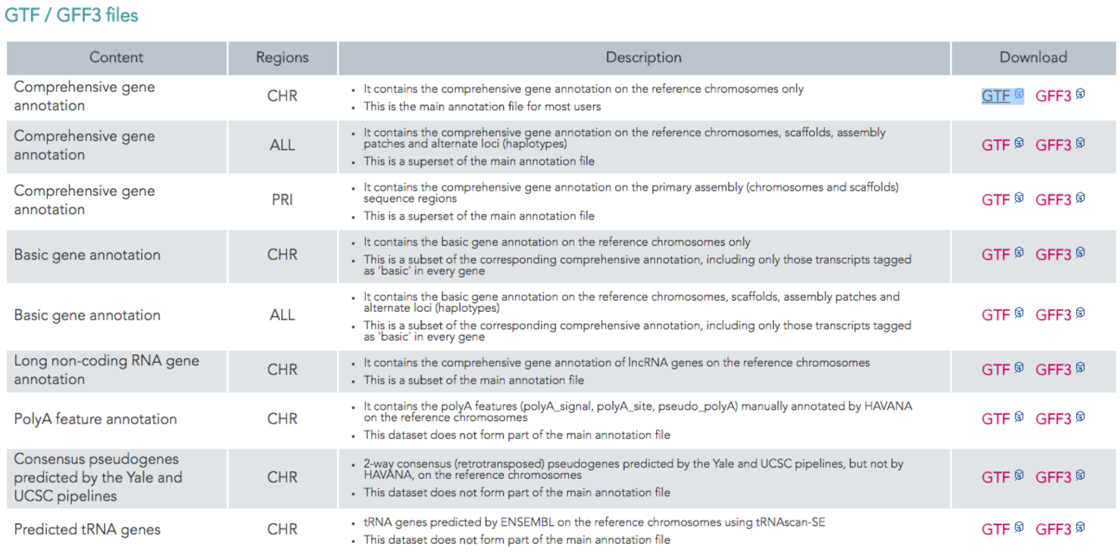
Open the Terminal.

First, go to the folder where you can download the data.

```
cd ~/scratch/day2/QC/practical/FileFormats
```

Questions

1. Download the current release of the gencode gtf for the chromosome only Try to use wget. Hint: http://www.gencodegenes.org/releases/current.html



GTF / GFF3 files

Content	Regions	Description	Download
Comprehensive gene annotation	CHR	• It contains the comprehensive gene annotation on the reference chromosomes only • This is the main annotation file for most users	GTF GFF3
Comprehensive gene annotation	ALL	• It contains the comprehensive gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes) • This is a superset of the main annotation file	GTF GFF3
Comprehensive gene annotation	PRI	• It contains the comprehensive gene annotation on the primary assembly (chromosomes and scaffolds) sequence regions • This is a superset of the main annotation file	GTF GFF3
Basic gene annotation	CHR	• It contains the basic gene annotation on the reference chromosomes only • This is a subset of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene	GTF GFF3
Basic gene annotation	ALL	• It contains the basic gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes) • This is a subset of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene	GTF GFF3
Long non-coding RNA gene annotation	CHR	• It contains the comprehensive gene annotation of lncRNA genes on the reference chromosomes • This is a subset of the main annotation file	GTF GFF3
PolyA feature annotation	CHR	• It contains the polyA features (polyA_signal, polyA_site, pseudo_polyA) manually annotated by HAVANA on the reference chromosomes • This dataset does not form part of the main annotation file	GTF GFF3
Consensus pseudogenes predicted by the Yale and UCSC pipelines	CHR	• 2-way consensus (retrotransposed) pseudogenes predicted by the Yale and UCSC pipelines, but not by HAVANA, on the reference chromosomes • This dataset does not form part of the main annotation file	GTF GFF3
Predicted tRNA genes	CHR	• tRNA genes predicted by ENSEMBL on the reference chromosomes using tRNAscan-SE • This dataset does not form part of the main annotation file	GTF GFF3

2. Let?s see what the file we just downloaded looks like. Hint: use less or head to visualise it.
3. How many chromosomes each it refers to? Hint: Use cat or awk to print only the first field of each line, then use sort and uniq
4. How many genes are reported in the current realese? Hint: look at the third field of the file.
5. Load it in IGV Hint: take into account the reference genome release.