Series 2

Genomics and bioinformatics - Week 2

September 27, 2011

1 Introduction

1.1 Description

In today's session, you will use publicly available genome sequence and annotation data for a particular species to extract some biological information about that species.

1.2 Before we begin...

If you do not have a working copy of Python and R on your computer please go through last week's tutorial before starting this exercise.

2 Getting data with UCSC

2.1 Visualizing genome data

1. Go to the UCSC Genome Browser and select the *Mus musculus* genome.
2. Visualize the most recent assembly (mm9) of mouse chromosome 18.
3. Scroll down to "Mapping and Sequencing Tracks" and load the GC percent track.

You can obtain more information about the tracks by clicking on them.

2.2 Downloading genome data

Copy the sequence file `chr18.fa` and annotation files `chr18.gtf`, `chr18_mod.txt` from the USB keys provided by us.

3 Manipulating data with Python

1. Load the `.fa` file for chr18 and extract the sequence.
2. Determine the length of the sequence.
3. Calculate the number of As, Gs, Cs and Ts in the sequence.
4. Compute the GC-content of the chromosome.
5. Plot GC content along mouse chr18 using an appropriate window (bin) sizes (use `matplotlib`).
6. Write the start and end coordinates of each bin and it's corresponding GC content to a file, as follows:
   `binStart binEnd GC_content`

4 Manipulating data with R

4.1 Exons

The `.gtf` file is a tab-delimited file, with the following column headers:
`chromosome source feature start end score strand frame attributes`

1. Load the `.gtf` file for chr18 in R.
2. Extract the rows corresponding to exons from the `feature` column to a new table.
3. Compute exon sizes, and attach them to the table in a new column "`exonSize`".
4. Plot the exon size distribution for chr 18.

4.2 Genes

The modified annotation file `chr18_attributes.txt` is also a tab-delimited file, with the following column headers:

```
chromosome source feature start end score strand frame gene_id transcript_id exon_number
gene_name gene_biotype    transcript_name protein_id
```

1. Load the modified `.txt` annotation file for chr18 in R.
2. Find out the ID and name of the gene containing, a) the longest exon, b) the shortest exon and c) most number of exons.
3. List all the intron-less genes in the chromosome.

4.3 GC content

1. Load the file (table) generated by your python script into R.
2. Recreate the GC content plot for chr 18 in R.

5 Reference documentation

For R - `http://cran.r-project.org/doc/manuals/refman.pdf`
For Python - `http://docs.python.org/tutorial/`

If you need help:

1. Go through last week's exercise session for examples
2. Use Google
3. Use the ? or help() with R commands
4. Ask us