Galaxy Exercises

Marc van Driel- June 4th, 2013 – Course (NBIC/WUR)- RNA-seq course: "The Power of RNA-seq"

Contents

EXERCISE 1: length of exons on chromosome 22

Make sure you have a clean history in galaxy (click on the *gear* icon at right hand top next to history, click create new)

1. **First get the exon information**
2. Menu: **Get data**
3. Select **UCSC main**
    - **a.** Settings:
        - **i.** **clade:** mammal
        - **ii.** **genome:** human
        - **iii.** **assembly:** HG19
        - **iv.** **Group:** Gene and gene predictions
        - **v.** **Track:** UCSC known genes
        - **vi.** **region, position:** chr22 (click lookup to get the coordinates)
        - **vii.** **output format:** bed
        - **viii.** **Check :** sent output to Galaxy
        - **ix.** **Click get output.**
    - **b.** New screen:
        - **i.** Select coding exons
        - **ii.** **Click send query to Galaxy**
4. Data set is loading in Galaxy. The history indicates: grey: in the queue; yellow: running; green: ready
5. Explore the functionality of the history:
    - Pencil, rename (renaming to a functional and descriptive name is **highly recommended**!)
    - Rename to "*Coding Exons*"
    - Format options
6. Now compute the exon length

7. Menu: **Text manipulation**
8. Select **Compute an expression on every row**
    a. **Add expression:** c3-c2 (length = end - start position)
    b. Select the dataset produced under 5: "*Coding Exons"*
    c. **Click execute**
9. Explore the new dataset in the history pane. Note the extra column!
10. Rename the dataset to "*Exon Lengths*"
11. Now do some simple statistics on the exon lengths
12. Menu: **Statistics**
13. Select **Summary statistics**
    a. Summary statistics on the data set "*Exon Lengths*"
    b. **Click execute**
14. The result is main length: 158; longest coding exon: 6762bp
15. Now filter "*Exon Lengths*" data set on coding exons larger than 6kb.
    a. Menu: **Filter and Sort**
    b. Select **Filter data on any column using simple expressions**
    c. **Filter**: dataset "*Exon Lengths*"
    d. **With following condition:** c7>=6000
16. The result is 1 region
17. Select in the history of the last dataset **display in UCSC browser**
18. The PJDREJ gene has a huge exon
19. Explore the function of PJDREJ(egg jelly precursor ; human reproduction)
20. Rename the history to *Exercise 1*
21. Explore the Tags options for the histories
22. Select **Extract workflow** in the history and rename the workflow. You can run and edit the workflow. Give it a try.
23. Explore the sharing options (show link, download) in the history menu.

EXERCISE 2: Which RefSeq genes on chr22 have a CpG island in the potential promotor region?

Here we try to find out which RefSeq genes on chromosome 22 have a CpG island in their potential promotor region. [Unlike CpG sites in the coding region of a gene, in most instances the CpG sites in the CpG islands of promoters are unmethylated if the genes are expressed. This observation led to the speculation that methylation of CpG sites in the promoter of a gene may inhibit gene expression. Ref: http://en.wikipedia.org/wiki/CpG_site]

Make sure you have a clean history in galaxy (click on the *gear* icon at right hand top next to history, click create new)

1. **First get the CpG information**
2. Menu: **Get data**
3. Select **UCSC main**
    a. Settings:
        i. **clade:** mammal
        ii. **genome:** human
        iii. **assembly:** HG19
        iv. **Group:** Regulation
        v. **Track:** CpG islands
        vi. **region, position:** chr22 (click lookup to get the coordinates)
        vii. **output format:** bed

viii. **Check :** sent output to Galaxy
ix. **Click get output.**

b. New screen:

i. Select whole gene
ii. **Click send query to Galaxy**

4. The result is a dataset with 719 CpG Islands; Rename the dataset to *CpG islands*

5. Now get the RefSeq genes

6. Menu: **Get data**
7. Select **UCSC main**

a. Settings:

i. **clade:** mammal
ii. **genome:** human
iii. **assembly:** HG19
iv. **Group:** Genes and Gene predications
v. **Track:** RefSeq Genes
vi. **region, position:** chr22 (click lookup to get the coordinates)
vii. **output format:** bed
viii. **Check :** sent output to Galaxy
ix. **Click get output.**

b. New screen:

i. Select whole gene
ii. **Click send query to Galaxy**

8. The result is a dataset with 922 genes; Rename the dataset to *RefSeq Genes*

9. Now we create the a dataset with the upstream regions of the RefSeq genes

10. Menu: **Operate on genomic intervals**
11. Select Get flanks returns flanking region/s for every gene
    a. Explore the manual of the Get flanks tool by scrolling down in the middle pane
    b. **Select data:** *RefSeq Genes* dataset
    c. **Region:** whole feature
    d. **Location of the flanking region/s:** Upstream
    e. **Offset:** 0
    f. **Length of the flanking regions:** 1000 (potential promotor region)
12. Rename the resulting dataset in *Upstream regions*
13. Explore the dataset in the UCSC browser by clicking on display in UCSC browser in the history of dataset *Upstream regions*
    a. *E.g.* go to gene PRODH by clicking the gene
    b. Click genome browser in the gene details page
    c. Zoom out 3X
    d. Show the CpG track in browser
14. Now we have to intersect the upstream regions of the RefSeq genes with the CpG island track:
    a. Menu: **Operate on genomic intervals**
    b. Select Intersect the intervals of two datasets with settings:
        i. **Return:** Overlapping intervals (NOTE: check the manual for the details!)
        ii. **Of:** *Upstream regions* (Intersect of GENES overlapping with CpG islands, order matters!)
        iii. **That intersect:** *CpG islands*
        iv. **For at least:** 1bp
        v. **Click execute**
15. The resulting dataset contains 587 regions of original total of 922.
16. Rename the dataset to *Upstream regions with CpG*

17. Go to the history and display the *Upstream regions with CpG* dataset in the UCSC browser
    a. Switch on the CpG track
    b. Go e.g. to the DGCR2 gene (chr22:19,103,416-19,117,777 ; HG19)
18. Optionally: Join two datasets *Refseq Genes* and *Upstream regions with CpG* to find back the genes with a CpG upstream
19. Menu: **Join, Subtract and Group**
    a. Select: **Join two Datasets side by side on a specified field** with settings:
        i. **Join:** *RefSeq Genes*
        ii. **using column:** c4
        iii. **with:** *Upstream regions with CpG*
        iv. **and column:** c4
        v. **Click execute**
20. The resulting dataset contains the original RefSeq Genes with upstream CpG islands

EXERCISE 3: LiftOver

Since all the genome information in the human browser and Galaxy depends on coordinates using the right genome build version is important. Galaxy uses the LiftOver tool to convert genomic coordinates between build versions.

We convert SNP137 from hg19 to hg18 to illustrate the conversion:

1. **First get the SNP137 information**
2. Menu: **Get data**
3. Select **UCSC main**
    a. Settings:
        i. **clade:** mammal
        ii. **genome:** human
        iii. **assembly:** HG19
        iv. **Group:** Variation and Repeats
        v. **Track:** Common SNPs(137)
        vi. **region, position:** chr22 (click lookup to get the coordinates)
        vii. **output format:** bed
        viii. **Check :** sent output to Galaxy
        ix. **Click get output.**
    b. New screen:
        i. **Click send query to Galaxy**
    c. Rename the dataset to *SNPs137 hg19*
4. Menu:**Lift-Over**
    a. Select: **Convert genome coordinates between assemblies and genomes,** with settings:
        i. **Convert coordinates of**: *SNPs137 hg19*
        ii. **To:** Human hg 18
        iii. **Click Execute**
5. Two histories will be created: Mapped and Unmapped results
6. The Mapped results are converted to the new genomics coordinates, while the unmapped SNPs are still based on the old coordination system.

Exercise 4: Create exercise (1) with the workflow editor

1. Click workflow in the top bar
2. Click Create new workflow
3. The Canvas is empty
4. As you select the tools on the left hand panel these will appear on the canvas
5. Try to reconstruct exercise 1 using this graphical interface.

Exercise 5: Basic Protocol 2 - Load and prepare MPromDB data. MPromDB is a curated database that strives to annotate gene promoters identified from ChIP-Seq experiment

Source: https://main.g2.bx.psu.edu/u/galaxyproject/p/using-galaxy-2012

Data: http://wiki.galaxyproject.org/Datafiles/Mouse%20ChIP-Seq%20Data

Screencast: http://screencast.g2.bx.psu.edu/CPB2012_UsingGalaxy_P2/flow.html

Make sure you have a clean history in galaxy (click on the *gear* icon at right hand top next to history, click create new)

1. New history
2. Click the Shared Data – Data Libraries link in the top bar
3. Search mouse
4. Select "*ChIP-Seq Mouse Example*"
5. Check the boxes for the 2 datasets
6. Choose Import to current history
7. Click Go
8. At the top a green bar confirms the import
9. Click Analyse data to return to the history
10. Rename the datasets to *Control Chr19 ungroomed* and *Tags Chr19 ungroomed*
11. Second data set import Annotated Promotor dataset (MM9, chr19): http://wiki.galaxyproject.org/Datafiles/Mouse%20ChIP-Seq%20Data
12. Upload the file 'Get Data' – Upload file
13. **Alternative**: FTP upload:
    a. Host: main.g2.bx.psu.edu
    b. Username: email
    c. Password: galaxy passwd
    d. Connect
    e. Transfer the file to the Galaxy server
    f. Get data, upload
        i. Files for FTP
14. Rename the data set: *MPromDB Promotors chr19*
15. The imported file has a tabular format, which needs to be converted to an interval format to make it useful.Convert the dataset for use with intervals:
16. Tool search function: cut
    a. **Settings:** c2, tab delimited from dataset 3
    b. The result is a dataset with: chr:start..stop promotor data
17. The format with the colon (":") needs to be converted to TAB
18. Tool search: convert delimiters
    a. **Settings**: colons; dataset 4
    b. The result is a dataset with chr TAB start..stop
19. Again the format with the dots ("..") needs to be converted to TAB
    a. Tool search: convert delimiters
    b. **Settings**: dots; dataset 5
20. The result is an interval dataset, but there is no extra data from the original dataset
21. The order of the data set did not change, so we can glue datasets together using Paste tool
22. Tool search: paste
    a. **Settings:** convert on 5 (interval data set) and MpromDB dataset, delimited tab

23. The result is an interval dataset with original data
24. Rename the dataset: *MPromDB Promotors chr19 interval*
    a. Save the changes
    b. Change data type to **interval**
    c. Meta data is now auto detected (this takes time: reload the page/history)
    d. Change meta data: strand and name: (c10 & c8)
    e. View column names: strand and name are set. This is a complete interval data set.
25. Bed format is used often, to generate:
    a. Tool search: cut
    b. Settings: c1,c2,c3,c8,c13,c10 ; tab delimited (c13 = the score)
    c. Change data set to BED
    d. Meta data is now auto detected (this takes time: reload the page/history)
    e. Edit: visualisation score: c5
    f. Rename: *MPromDB Promotors chr19 BED*
26. Third data source is needed: RefSeq dataset (UCSC tables browser)
    a. Get data - UCSC table
    b. Mouse, MM9, Genes and gene predictions, RefSeq, position chr19 lookup, BED, check GALAXY, get data
    c. New page: whole genes, send to galaxy
    d. Rename the data set: *Refseq genes chr19*; change score to c5

Exercise 6: Basic Protocol 3 - Calling Peaks for ChIP-seq Data; Zinc finger CTCF ENCODE experiment

Source: https://main.g2.bx.psu.edu/u/galaxyproject/p/using-galaxy-2012

Data: Basic_Protocol_2 or from the above link. (Direct link: https://main.g2.bx.psu.edu/history/imp?id=8c52d322f1c41705)

Read http://en.wikipedia.org/wiki/CTCF for some background information on CTCF.

Make sure you have a clean history in galaxy (click on the *gear* icon at right hand top next to history, click create new)

Try the **Screencast** yourself: http://screencast.g2.bx.psu.edu/CPB2012_UsingGalaxy_P3/flow.html

There are plenty of screencasts available at http://wiki.galaxyproject.org/Learn/Screencasts

Exercise 7: Finding Heteroplasmic Sites: Two tissues in a single individual

This study is published as a Galaxy paper: http://main.g2.bx.psu.edu/u/aun1/p/ismb2010-demo

Galaxy Papers are interactive datasets which can be reanalysed. More studies can be found at: https://main.g2.bx.psu.edu/page/list_published