

Bioinformatics Lab

Analyses of ChIP-Seq and RNA-Seq data using Galaxy and UCSC Genome Browser

version 1.0, May 28 2009

Aim: The aim of this lab is to introduce Galaxy as a web server tool that enables custom analyses of genome-wide data. In this exercise, this potential will be exemplified by analyses of target genes for pluripotency associated transcription factors, non-coding RNAs and studies of their regulation in embryonic stem cells using available RNA annotations and ChIP-Seq data of transcription factor binding sites and histone modifications.

Outline:

Task 1: Gene regulation in embryonic stem cells and induced pluripotent cells are in part regulated by three important transcription factors, Oct4, Sox2 and Nanog. The binding sites of these three factors in embryonic stem cells was recently investigated by sequencing of regions bound by the factors (ChIP-Seq). The first task consists of identifying the genomic regions bound by **all** three factors and then to find the closest gene to each region.

Task 2: Visualize RNA-Seq data as a custom track in the genome browser

In this part, we will download rna-sequencing data from Cloonan et al. Nature Methods 2008 from embryonic stem cells and look at the expression of a few genes using the UCSC Genome Browser. Thus, you will learn how to make use of latest sequencing data to find out what regions are transcribed in embryonic stem cells and embryonic bodies.

Task 3: A set of custom non-coding RNAs (long intron-containing non coding RNAs; lincRNAs; defined in Guttman et al, Nature 2009) will be uploaded to Galaxy. The total set of lincRNAs will be filtered by only selecting the lincRNAs that are under purifying selection, by using the evolutionary conservation of the regions in other mammals.

Available data:

Data have been converted from supplementary tables (Chen et al. Cell 2008) into bed format genome annotations (for an description of bed format, see this link <http://genome.ucsc.edu/FAQ/FAQformat#format1>).

Data will be accessible from the web site on the laboratory day.

Galaxy URL: <http://main.g2.bx.psu.edu>

Screen cast about Galaxy usage: <http://g2.trac.bx.psu.edu/wiki/ScreenCasts>

Detailed instructions:

Task 1: Genes regulated by Oct4/Sox2/Nanog in mouse embryonic stem cells

Open Galaxy in firefox (<http://main.g2.bx.psu.edu/>), you should see a collection of tools in the left menu, a blank history in right side of the firefox window.

1.1 Upload the custom transcription factor binding sites data

Download the bed files containing the binding regions of the three transcription factors: Oct4_mm8.bed Sox2_mm8.bed Nanog_mm8.bed, save them on your computer, e.g. the Desktop.

Note. The bed files were converted from the Supplemental Table S3 of Chen et al. Cell 2008 [pdf] from the cell homepage:

[http://www.cell.com/supplemental/S0092-8674\(08\)00617-X](http://www.cell.com/supplemental/S0092-8674(08)00617-X)

[Table S3. Coordinates of Loci Bound by Transcription Factors \(XLS 7480 kb\)](#)

Next, upload the bed files into galaxy, by selecting “Get Data” from the left hand tools menu, further select “Upload File from your computer”. As “File” input, enter the path to the Oct4_mm8.bed file you downloaded to your computer above. Select File Format to “bed” and Genome to “Mouse Feb 2006 (mm8)”. Click execute and the file should appear on the right side menu, turning green when fully uploaded. Continue to upload the Sox2 and Nanog file similarly.

1.2 Find the intersection of the regions bound by the factors

1.2.1 Extend the regions bound by Oct4 +/- 100 bp to find closely locate binding regions when later comparing the Oct4 bound regions to Sox2 and Nanog.

First, select “Text Manipulations” and “Compute” from the left tools menu. Specify the expression: “c2-100” which means that the column 2 which holds the regions start genomic coordinates should be subtracted by 100. Specify the “Oct4_mm8.bed” as the file to add this column to. Check the “Round results” to “yes” in order for the computed values to be stored as an integer (needed for later steps). Click “Execute”.

Redo the same procedure on the newly generated file, but here instead selecting the expression: “c3+100” to add 100 to the regions end genomic coordinate. Round the results and click execute.

The newest files contains the old bed information plus the two new columns we added that holds the extended genomic coordinates. Now we need to tell galaxy to update the file attributes to instead treat the new columns as the start and end coordinates. Click the 'pen icon' to the right of the newly generated file. Click change data type and select new type: “bed”. In the next page you modify bed file information and selects the new columns as the start end end region. Change the name to: e.g. Ext_Oct4_mm8 and specify column 4 as start columns and column 5 as end column. Unclick the Name/Identifier checkbox and then click “save”. Done. You may also delete the intermediary file (still called something like “compute on data 1”, by clicking on the 'X' icon to the right of the file name in the right menu.

1.2.2 To compare the binding regions between the factors, select “Operate on Genomic Intervals” in the tools menu. Click “Intersect”. Look at the bottom figure to see what overlap will be reported when using the “Overlapping intervals” selection (default). Define Ext_Oct4_mm8.bed as the source (“of”) to be intersected with “that intersect” Sox2_mm8.bed. Require at least on bp overlap (as suggested as default). Click execute. The new set should appear in the right menu. Change name of the created file by selecting the 'pen icon' to the right of the newly created file name. Change the name to 'Oct4_Sox2_mm8' and click “save”. Make sure the name changed in the right side menu. Repeat the process of defining the intersection, but now instead selecting the “Oct4_Sox2_mm8” set and the last Nanog_mm8.bed file. Change the name of the resulting file into

“Oct4_Sox2_Nanog_mm8”. The file should contain 656 regions.

1.3 Find the closest genes to the regions bound by all three factors

Download gene information (NCBI Refseq database) for mouse (assembly mm8) by using the Galaxy to UCSC Genome Browser interface. Select “Get Data” under tools menu, and further select “UCSC Main table browser”. The UCSC Genome browser webpage should appear in the middle section of firefox window. Select genome: “mouse”, assembly: “Feb 2006” (which corresponds to mm8), group: “Genes and Gene Prediction Tracks”, track: “Refseq Genes”. Make sure the checkbox “Send output to Galaxy” is checked and that the output format is “XXX”. Click “get output” and “send query to galaxy” to fetch the data into Galaxy without changing any of the fields. When completed, the new file should appear in green in the right side history panel. We now have to convert it into bed format by selecting the 'pen icon' to the right of the file name. Click change data type and select new type: “bed”. Click save. Next is a page where you define what features will be converted into the columns of the bed file you are creating. Change the name to “Refseq Genes” and update chrom column to 3, start column to 5 and end column to 6. Specify name column as 13.

Now, finding the Refseq Genes closest to our bound regions, we use the “Operate on Genomic Intervals” - “Fetch closest features”. For every feature in the “Ext_Oct4_Sox2_Nanog_mm8” file fetch the closest feature in the downloaded Refseq file (called something like UCSC Main on Mouse: refGene (genome)”. Click Execute.

1.4 Examine the results

Click on the 'eye icon' to the right of the file name in the right hand menu to visually inspect the created file. Do you understand the information in the field? Click the “save” link next to the file name in the right menu and save the file to e.g. Desktop. The file can now be opened in e.g. excel for further analysis, e.g. can you determine if Sox2 is among the putative regulated genes?

1.5 (optional) microRNA genes may be closest

Redo the mapping to closest genes but use instead the database of microRNAs to find potential microRNAs to be regulated by the the triad of transcription factors. Get Data from UCSC Main (as for Refseq) but select track: “miRNA”. Find the closest miRNA genes using the “Operate on Genomic Intervals” - “Fetch closest features”.

1.6 (optional) Oct4 was not found to be a regulated gene in the analysis above, lets look at the raw data in the Oct4 locus to figure out why it was not included among the results.

Task 2: Visualize RNA-Seq data as a custom track in the genome browser

In this part, we will download rna-sequencing data from Cloonan et al. Nature Methods 2008 from embryonic stem cells and look at the expression of a few genes using the UCSC Genome Browser. A web page for the publication is maintained from the Grimmond labs homepage:

<http://grimmond.imb.uq.edu.au/mESEB.html>

there you can find files in the wiggle format (for a description of the wiggle format look here: <http://genome.ucsc.edu/goldenPath/help/wiggle.html>)

We are going to use two files,

<http://grimmond.imb.uq.edu.au/files/nicole/mm9/ESEB/ES.all.combined.positive.wig.rounded.condsd.20plus.gz>

<http://grimmond.imb.uq.edu.au/files/nicole/mm9/ESEB/ES.all.combined.negative.wig.rounded.condsd.20plus.gz>

keep the firefox window open – you are going to copy the link to the datafiles soon.

In a new firefox window, open the custom track page in UCSC Genome Browser, click on “add custom track”, insert the two links above into the “Paste URLs or data” field. Make sure the genome is set to Mouse and assembly is July 2007 (mm9). Press “submit”. Press the “go to genome browser” button. When in the genome browser, select Sox2, press the link on the search results page. Now you should see the Sox2 region (chr3:34,548,927-34,551,382). What are your impression of the short read coverage of the region? Zoom out to see if you can find evidence for other regions being transcribed in embryonic stem cells that are not annotated Refseq regions.

Task 3. Purifying selection of non-coding RNAs