Developing Workflows for Functional Genomics Using Galaxy
Whole Genome ChIP-Seq Data Analysis

Many questions in bioinformatics require the use of multiple types of data; but often, different types of data require different programs to analyze them. Further, many programs are limited in the amount of input data that can be uploaded, meaning we ideally would run analyses on local installations of programs where we can tackle our questions with large datasets. Thirdly, doing complex bioinformatics analyses with large datasets usually means you have developed a pipeline that you would like to re-run on multiple datasets, or that others might want to run to answer similar questions in their datasets. Lastly, and most importantly, very few bioinformatic and genomic studies have reproducible results, even though reproducibility is the cornerstone of good scientific inference.

Overview: In this exercise, we'll look at some chromatin immunoprecipitation (ChIP) data to find binding sites for a transcription factor. For these analyses, we'll use Galaxy, a suite of integrated tools for working with complex genomic data that allows one to perform a series of steps (ranging from data cleaning to analysis in a single workflow) that can be saved, shared, and executed with other datasets.[1]

Chromatin immunoprecipitation (ChIP) is a technique to identify where proteins that interact with DNA bind. Chromatin is purified from cells of interest (that is, DNA is extracted under conditions that won't dislodge bound proteins) and the DNA is fragmented. An antibody specific to a protein of interest is then used to precipitate (pull out) that protein along with whatever DNA fragments remain bound to it, then the protein is released and only those DNA fragments are sequenced ***(ChIP-seq)***. This results in a DNA sequence dataset that includes only candidate binding sites for the protein being studied. Alternatively, the DNA fragments that are released after the protein pull-down can then be hybridized to an array ***(ChIP-on-chip)*** if the experiment is conducted in a well-studied organism for which arrays are available.

Transcription factors (TFs) are a class of proteins that bind to DNA and regulate transcription. Because transcription levels often must be finely regulated (that is, certain genes are turned on in certain cells at certain times in certain environments), understanding where and when TFs bind to DNA can reveal a lot about the biology of an organism. Figuring out which genes a given TF regulates can be accomplished by performing a *ChIP-seq* experiment for a given TF, and then aligning the regions of the genome that TF binds to with whole genome sequence and looking to see what promoters or genes are nearby. In this exercise, you will use the results from a TAF1 transcription factor ChIP experiment and the human genome sequence to identify what genes might be regulated by this TF binding to candidate promoter regions nearby.

Part I. Uploading a ChIP dataset into Galaxy

Before working with Galaxy, register (for free!) so that you can save and share your workflows for later. To register, click on the **User** button in the top (black) menu bar, and select **Register.** Enter all the necessary information (email, password and public name) and click **Submit**. After registering, you need to activate your account (via the link sent to your email), and then log in to Galaxy using your account information.

Our data file contains 200 chromosome regions identified in one ChIP experiment with TAF1. Given this list, we could, for example, examine nucleotides 116099071 through 116100373 from chromosome #7 (the first region in the list) using a genome browser, identify any genes or known transcripts in the vicinity, then move on to the next region. However, even with the relatively short list, this would become tedious quickly. Using Galaxy will allow us to analyze all 200 at essentially the same time. The first step is to load our ChIP file into Galaxy.

1. Download the file **TAF1_ChIP.txt** from the workshop website and save it on your desktop.
2. Navigate to Galaxy at **http://main.g2.bx.psu.edu** or use the link on the workshop website.
3. In the left sidebar, click **Get Data** and choose **Upload File**. Click on **File: Browse** in the center panel, locate the file you saved and click **Execute**. Your job will appear in the right sidebar and turn green when it's uploaded. *Everything you do in Galaxy will appear in this right hand sidebar while it is processing and will turn green once it is complete.*

[1] This exercise is modified from a Galaxy tutorial screencast (http://main.g2.bx.psu.edu/screencast) and from Bock *et al.*, Methods in Molecular Biology (2010)

4. Click on the filename link in the right sidebar to see a preview of the first few lines of the file. Notice the chromosome number, and the start (**chromStart**) and end (**chromEnd**) nucleotides identify each sequence that bound to TAF1, but there are also some columns with data that we don't need.

5. Notice the buttons at the upper right of the green box. They allow you to view all of the data (eye icon) or to edit the information for this data file (pencil icon). Galaxy needs to know what kind of information is contained in this file, so click the **pencil icon** and then in the **Datatype tab**, under **New Type:** type "**in**" and then select **interval**; this indicates that this file contains a list of intervals of interest from across the genome. Click **Save**.

6. When the notification "**changed the type of dataset 'TAF1_ChIP.txt' to interval**" appears at the top of the center window, additional options will appear under **Edit Attributes (**in the **Attributes tab).** Galaxy needs to know what organism these data come from, in the **Database/Build** box type **hg18** to use this version of the human genome.

7. Galaxy requires you to specify which columns in the file contain the chromosome number **(Chrom column: select 2)**, the start positions **(Start column: select 3)**, and the end positions **(End column: 4)**. Be sure these were detected correctly and correct any that were not (no other data in the file are necessary to this experiment). You can also change the name for these data to something friendlier (e.g., TAF1 ChIP data) if you like. Click **Save**.

Part II. Downloading all the genes in the human genome

In order to analyze where TAF1 binds relative to known genes, we need a list of known genes. Galaxy can pull data directly from major genetic databases, so we'll start by retrieving a list of annotated genes from the **RefSeq database**. RefSeq is one of the many databases housed at NCBI, but is unique in that it contains a "reference" version of each well-annotated gene region (DNA, RNA, and protein data) currently available (without the redundancy of Genbank, which contains many versions of each region).

8. Under **Get Data** in the left sidebar, choose **UCSC Main table browser**.

9. A form will appear in the middle panel, check and/or adjust *ALL* the following parameters:
genome: Human
assembly: Mar. 2006 (NCBI36/hg18)
track: RefSeq Genes
region: select genome button
output format: BED – browser extensible data
Send output to Galaxy: select/check this box

10. Then, click **get output.**

11. When the next screen appears, just click **Send query to Galaxy** button on the bottom. A new box will appear in the sidebar representing this job, which should turn green when the gene list is downloaded.

12. Click the title link [**UCSC Main on Human: refGene (genome)]** located in the right-hand window for a quick preview of the file, and to see how many annotated regions were downloaded. Use the pencil icon to change the name of this job to something more memorable, like **RefSeq genes**, and **Save** the new name (Remember click the eye icon to view data).

Part III. Extracting putative promoter regions upstream of each gene

Indeed, we have some biological information about TAF1—it is known that it binds near transcriptional start sites. Let's assume that for the moment that most transcriptional start sites are *upstream* of genes, thus we *actually* want to extract the region upstream of every gene, not the gene itself.

13. To generate a file with the start and end points for 1000 bp regions upstream of each gene, in the left sidebar under **Operate on Genomic Intervals,** choose **Get flanks**. Check and/or adjust the parameters bulleted below:
Select data: RefSeq genes
Location of the flanking region/s: upstream
Length of the flanking region/s: 1000

14. Click **Execute**. You should now have a new job with a new table of upstream regions for every gene. (Wasn't that easy?) Rename this job **Upstream Flanks** by clicking on the pencil icon next to it (remember to **Save**). *NOTE: this dataset will be so large, that even when you click on the eye icon, only a part of the dataset will appear in the central window, you will have to download the dataset to see it all.*

15. This table contains more information than we are currently interested in (i.e., locations of the introns and exons of original genes). Since we are interested in promoter regions, let's clean this data set up. Under **Text Manipulation** in the left sidebar, choose **Cut columns from a table**. "Cutting" columns actually means keeping them. Fill in **c1, c2, c3, c4, c6**. Click **Execute.**

16. Now, you have a new table with just the desired information for each promoter region (name this **Upstream flanks clean**), but because we have modified the table, Galaxy did not retain the column titles. To fix this, click on the pencil icon and make sure Galaxy auto-detected which columns contained data. (If you go back and view the table preview in your **Upstream flanks** job, you can see that the columns names are: 1 [chromosome number], 2 [start position], 3 [end position], 4 [name], and 6 [strand] to make sure they match in your clean dataset – *NOTE that column 6 in "Upstream flanks" got shifted to column 5 in the "Upstream flanks clean").*

Part IV. Joining the datasets and displaying results

17. Now we're ready for the interesting part: joining the ChIP results with the **Upstream flanks clean** list to identify which genes TAF1 might actually bind to and regulate. Click on the **Operate on Genomic Intervals** link (in the menu on the left) and choose **Join the intervals of two datasets side-by-side**.

18. From the **Join** drop-down menu, specify that we want to **join: Upstream flanks clean** (First dataset) **with: TAF1 ChIP data** (Second dataset). Click **Execute**.

19. A new job will appear listing all TAF1 binding sequences that overlap with one of the candidate promoter regions. Name this new job **candidate TAF1-binding promoters.** Click on the job title to see how many regions were returned by joining these 2 lists.

How many "regions" appear to have a TAF1 binding site upstream of them? ___________

20. Click the eye icon to view the **candidate TAF1-binding promoters** data: candidate promoter regions (from the Upstream Flanks dataset) are shown in the *left* columns, and the TAF1 binding sites from your ChIP data that overlap them are shown to the *right*. Scroll through your results.

Are there any binding sites upstream of genes located on chromosome 3? ________

Notice there are apparent repeats in the "Upstream flanks clean" columns on the left (in terms of chr, start, and stop position). But if you look at the column with "name" data (Col 4), each row is actually unique. Based on what you know about databases and copy number variation in genes think about why this might occur. If multiple RefSeq sequences have been annotated as reference variants for the same specific locus, it is not a surprise that the upstream flank of each variant would be the same and would therefore match the same TAF1 binding site. In fact, if you look at the set up RefSeq genes that you downloaded, you should notice there are way more than the ~21,500 genes estimated in the human genome.

How many "regions" did your RefSeq download include? __________________

21. A graphical display would make it much easier to visualize what we've found. We can see the results displayed in the UCSC Genome Browser and even create our own tracks to show our data! To set this up, choose **Graph/Display Data** from the list in the left sidebar then **Build custom track.**

22. In the new window click **Add new Track**, next set **Dataset** to the **candidate TAF1-binding promoters**. Change the track **name** to something short like **TAF1 promoters**, and change **Color** to **Red**. Click **Execute**.

23. When the job is complete, expand it by clicking on the title in the right side bar. Then click the **display at UCSC main** link just above the table preview. This will open the UCSC Genome Browser in a new window to display the data.

24. The graphical display at the top shows where we are relative to the human chromosome set. The first region of interest is on chromosome #1, so the browser is focused there by default. **Zoom out** by 3× in order to see the candidate **TAF1** binding region along the DNA.

The TAF1 binding site should not be very large, but we are seeing 1000 bp blocks. Why? ____________

What gene is the putative promoter at this site associated with? Look for the nearest gene to answer this question. You can click on its abbreviation to find out the real name. Remember, genes can be coded on the + or – strand, so upstream or downstream might be to the left or to the right of a gene *(there are little arrows inside your TAF1 binding site and in the gene nearby that indicate orientation).* ______________

What type of "gene" is this? ____

25. Given that we are looking at putative promoters, we can examine a track showing promoter-specific histone modifications to see if they coincide with our predicted TAF1 binding sites as well. Navigate to **chrX:152,800,000-153,103,900** and look at the ~300,000 bps of data on the long arm of this chromosome more closely. Scroll over the labels on the left of the tracks and click on the **Layered H3K4Me3** track to drag it up next to your custom track and see if it corroborates some of these putative promoter regions. H3K4Me3 (histone H3 Lys4 trimethylation) is a promoter-specific histone modification associated with active transcription and when detected suggests gene activation.

Does this track corroborate these regions as possible TAF1-binding promoters? Why or why not?
____

Using this same pipeline, you could download ChiP-Seq data for every transcription factor for which there are data, and find gene regulation sites throughout the genome.

On your own….

Two of the major findings of the human genome project was that 1) there are more promoters than previously thought, and 2) those promoters can be both upstream *and* downstream of a gene and still be functional. Test the hypothesis that TAF1 is just as likely to bind to promoters on either side of human genes using your skills. (Hint: Go back to step #13 and make a new set of promoters by getting 1000 bp of downstream flank –remember to save this dataset with a new name so you don't get confused. *Continue with the analysis to see how many TAF1 binding sites there are in downstream promoters, and compare this number to the number you wrote to answer to the question at step 3 on this page.)*

Technical Note on Workflows

If you actually worked with transcription factors, it's likely that you'd perform analyses such as these routinely. For example, you would probably run multiple ChIP experiments with different cell types or conditions, and each time you'd want to obtain clean promoter data from the RefSeq gene set and join those data with your ChIP data. It's obviously a lot easier to do this with Galaxy than by hand, but it would be even easier to store your set of steps and recycle them. This stored series of operations is a workflow.

1. From the **Options button:** ⚙ located above the right sidebar, choose **Extract workflow**.
2. The screen that appears in the middle panel will show all the steps you have taken to produce the job list that's currently in the sidebar. Some steps, such as uploading a file, can't be included in the workflow and will be shown in grey. Steps with a tan title bar can become part of your workflow; un-check the ones you don't want (e.g., you may not wish to do the CpG island or SNP searches every time). Name your workflow and click **Create Workflow**.
3. Now, under **Workflows** in the left-sidebar menu select **All workflows**
4. Click on the workflow you've just created. You should see the same steps you completed manually, with drop-down lists allowing you to choose input datasets. If you wanted to analyze another set of ChIP data, you'd just upload the data, get the gene set you want to compare with, select those jobs as input and run the workflow.
5. Back up one step to the list of workflows, and click the button labeled **Switch to workflow management view**. Click the name of your workflow and then choose **Edit**. This view shows how your workflow uses the input data files; you can add, delete or rearrange steps, or edit relationships here.

This exercise was developed for a 2013 workshop at Reed College funded by the National Science Foundation. The authors give permission for the non-commercial educational use of this material.

Further Reading:

1) Nekrutenko, A & Taylor, J. 2012. Next-generation sequencing data interpretation: enhancing reproducibility. Nature Reviews Genetics 13, 667-672. Web-based visual analysis for high-throughput genomics. 2) Goecks J, et al. 2013. Web-based visual analysis for high-throughput genomics. BMC Genomics. 14:397. 3) Xiao S, et al. 2012. Comparative epigenomic annotation of regulatory DNA. Cell 149(6):1381-92.
Tsai, Pei-Fang. 2009. TAF1 Regulation of Gene Expression: Genome-Wide Localization and Transcription Profiling. UC Riverside Dissertation.