Published in final edited form as:

Curr Protoc Mol Biol. 2010 January ; 0 19: Unit–19.1021. doi:10.1002/0471142727.mb1910s89.

Galaxy, a web-based genome analysis tool for experimentalists

Daniel Blankenberg^{1,5}, Gregory Von Kuster^{1,5}, Nathaniel Coraor^{1,5}, Guruprasad Ananda^{1,5},
Ross Lazarus^{2,5}, Mary Mangan³, Anton Nekrutenko^{1,5}, and James Taylor^{4,5}

¹The Huck Institutes for the Life Sciences, Penn State University

²Channing Laboratory, Harvard Medical School

³OpenHelix LLC

⁴Dept. of Biology, Emory University

Abstract

High-throughput data production has revolutionized molecular biology. However, massive increases in data generation capacity require analysis approaches that are more sophisticated, and often very computationally intensive. Thus making sense of high-throughput data requires informatics support. Galaxy (<http://galaxyproject.org>) is a software system that provides this support through a framework that gives experimentalists simple interfaces to powerful tools, while automatically managing the computational details. Galaxy is available both as a publicly available web service, which provides tools for the analysis of genomic, comparative genomic, and functional genomic data, or a downloadable package that can be deployed in individual labs. Either way, it allows experimentalists without informatics or programming expertise to perform complex large-scale analysis with just a web browser.

Key terms

Galaxy; analysis; bioinformatics; workflow; algorithm; pipeline; genomics; SNPs

Introduction

Research in the life sciences continues to become more data intensive. With new high throughput experimental techniques, an individual lab can generate raw data of a scale that was unthinkable only a few years ago. These developments represent an enormous opportunity for basic and applied research. However, they are also creating a crisis for many experimental scientists, since making sense of this wealth of data requires significant analysis infrastructure. Without informatics support, experimental biologists, who possess key biological knowledge and experience, and thus the best potential for making novel discoveries, cannot effectively use the available data.

Correspondence should be addressed to: Anton Nekrutenko, anton@bx.psu.edu or James Taylor, james.taylor@emory.edu.

⁵<http://galaxyproject.org>

Galaxy (<http://galaxyproject.org>) rectifies this situation by providing the needed informatics infrastructure (Taylor et al, 2007). For experimentalists, it provides an analysis environment in which they can perform analysis interactively, while ensuring that the resulting analyses are transparent and reproducible. The Galaxy framework encapsulates high-end computational tools, and gives them intuitive user interfaces while hiding the details of compute and storage management. It thus eliminates the need for specialized informatics expertise when performing many common types of large-scale analysis.

This unit describes the functionality of Galaxy using a series of examples. The material in this section is directed primarily at experimentalists, and makes use only of analysis tools available at the public Galaxy service at <http://usegalaxy.org>. Various components and tools of the public Galaxy server will be explored by following several connected, but independent, protocols. Although the data being investigated in these protocols may not be of personal research interest, the techniques demonstrated are useful in a wide-array of applications. Each of the protocols below is accompanied by a screencast (a real-time movie showing the steps of the protocol as they appear on the screen) available from <http://galaxycast.org/CPMB>. Following along with the screencasts is recommended, and they provide an alternate presentation of details not easily conveyed by text. This unit is divided into the following protocols:

- Basic Protocol 1. An introduction to the Galaxy approach: Finding promoters containing TAF1 binding sites identified from a ChIP-seq experiment
- Basic Protocol 2. A bit more data manipulation: Finding coding exons with most SNPs
- Support Protocol 2.1. Saving results in Galaxy and sharing data with others
- Basic Protocol 3. Generating a workflow from a history in Galaxy
- Support Protocol 3.1. Modify a parameter of the workflow in Galaxy
- Support Protocol 3.2. Running workflows with Galaxy
- Support Protocol 3.3. Sharing workflows with Galaxy
- Basic Protocol 4. Generating workflows from scratch with Galaxy
- Basic Protocol 5. Extracting sequences and alignments with Galaxy: A SNPs in exons example

These protocols cover the basic aspects of Galaxy's functionality. They are sufficient for overcoming the initial learning curve, but Galaxy has much more to offer, including complex analyses of next generation sequencing data such as metagenomic applications or re-sequencing studies. Additionally, the Galaxy project is progressing rapidly with new tools and features added on a monthly basis. The best way to keep up with these enhancements is to regularly check the screencast page at <http://galaxycast.org>.

Before beginning the protocols it will be beneficial to review some terminology and concepts. Many of the formats ("datatypes") used in genomics are composed of rows of tab-delimited columns which contain varied data (known as tabular data and similar in function

to a spreadsheet). One of these, known as interval, in which each row represents the position of a genomic feature in a particular genome. The interval format contains at least three columns: (1) the chromosome, (2) the start position within that chromosome and (3) the end position within that chromosome. Other columns commonly included are name, strand, score, and exon information (when the intervals are gene annotations). Additional formats beyond those composed of tabular columns are used, but the intricacies of their formats can be largely ignored in this introductory text as Galaxy can handle most of the details needed for performing complex analysis. The practice of matching rows between tabular datasets with Galaxy is known as “joining.” Two different Join tools are used here. The first Join tool works on interval datasets (using multiple columns to determine matching) and creates a dataset where rows are matched if their interval on the genome overlaps (by a user specified number of nucleotides) and combined into a single row. The second type of join works on a single column from each dataset and is useful for matching between identifiers. Every time a tool is run, one or more datasets are created in the user’s history. The box surrounding the dataset will change color based upon its state: a query in the queue will be indicated by a gray box, a running query will be yellow and a completed query will have a green box. Although a dataset is only ready to be viewed or used as input after it has turned green, additional analysis steps can be lined-up for non-completed queries by using the desired tools as normal; the tools will wait in the queue for the dataset needed to finish before running. Examining Figure 1 in detail will familiarize the user with the layout of Galaxy’s interface, including a user’s history and the tools menu.

Basic Protocol 1. An introduction to the Galaxy approach: Finding promoters containing TAF1 binding sites identified from a ChIP-seq experiment

Suppose a ChIP-seq experiment was performed and identified a series of genomic regions that bind TAF1-protein. The next task is to identify a list of genes that contain such sites. This can be easily done with Galaxy in just a few steps. This protocol uses a file arranged by tab-delimited columns, where each column contains information about the genomic positions (“intervals”), as well as name and score data, for TAF1-binding sites from a ChIP-seq experiment. Each row in this file represents an individual TAF1-binding site by listing the chromosome and the start and end positions within that chromosome. Here we assume that the ChIP-seq data has already been processed into putative binding regions, since this procedure is currently very experiment and lab specific. However, as best-practices are defined for performing and evaluating the quality these procedures, appropriate tools will be added to Galaxy.

Materials

- a file containing genomic coordinates for TAF1- binding sites from the ChIP-seq experiment (an example file can be downloaded at http://galaxy.psu.edu/CPMB/TAF1_CHiP.txt). (Kim et al, 2005)
- an internet accessible computer with any modern web browser (Firefox, Safari, Opera, IE).

Steps: *(Please note that the items to alter will be stated in the text. If other menus and options are not referenced, leave those settings in their default or existing condition.)*

1. Upload the TAF1 ChIP-seq data. Before beginning the analysis the ChIP-seq data needs to be uploaded into Galaxy's workspace (known as a user's "history" throughout this document).
  - a. Go to the public Galaxy site at <http://usegalaxy.org>.
  - b. Click "Get Data".
  - c. Click "Upload file".
  - d. Middle panel of the interface will change and allow selection of the desired file (use the example file that can be downloaded from http://galaxy.psu.edu/CPMB/TAF1_CHiP.txt). Note: It is possible to skip the downloading step and directly upload the data by entering the URL into the paste box, causing Galaxy to fetch the URL contents automatically.
  - e. Click "Execute". The dataset will be uploaded and will appear as dataset #1 within the right panel.
2. Set properties of the TAF1 dataset. To begin the analysis, a number of properties for the ChIP-seq dataset need to be set.
  - a. Expand the dataset by clicking on the name of the item in the history list (TAF1_CHiP.txt).
  - b. Click "?" next to "database:". A new interface will appear in the middle panel.
  - c. Use the "Database/Build" dropdown to select "Human Mar. 2006 (hg18)". Click the Save button. The dataset is now designated as originating from the human genome.
  - d. Back in the right panel, click the pencil icon. A new interface will appear in the middle.
  - e. Use the "New Type" dropdown within the "Change data type" box to select "interval". The "interval" datatype describes data representing genomic coordinates or "intervals" (chromosomes, start positions and end positions within chromosomes, as well as variable data, for a set of genomic features).
  - f. Click the "Save" button immediately below the box. The upper part of the interface will change.
  - g. Set "Chrom column", "Start column" and "End column" to 2, 3 and 4, respectively. Check the "Name" checkbox and select "5" from the adjacent dropdown.
  - h. Click the "Save" button immediately below.

- i. The appearance of dataset #1 within the right panel will change -- column headers will appear, and the format will change to “interval”.
3. Upload gene annotations from the UCSC Table Browser (Karolchik et al, 2008 ; Karolchik et al, 2004). To identify which genes' promoters contain the TAF1 binding sites, the gene coordinates must first be uploaded.
  - a. Click “Get Data” in the Tools menu list on the left panel.
  - b. Click “UCSC Main”. The UCSC Table Browser interface will be displayed in the middle panel.
  - c. Because the data is of human annotations, make sure that “clade”, “genome” and “assembly” are set to “Mammal”, “Human” and “Mar. 2006”, respectively.
  - d. Set “group” to “Genes and Gene Prediction tracks” and “track” to “RefSeq Genes”.
  - e. Select the radio button “genome”.
  - f. Make sure “output format” is set to “BED - browser extensible data” and the checkbox by “Send output to Galaxy” is checked. The BED format is a specialized version of the interval format discussed earlier.
  - g. Click “get output”. A new interface will appear.
  - h. Make sure the “Whole Gene” radio button is selected.
  - i. Click “Send query to Galaxy”. At this point a new dataset, #2, will appear in Galaxy's history on the right panel. This dataset contains the genomic positions for all RefSeq genes from the March 2006 human genome assembly.
  - j. Rename dataset #2 to “RefSeq” by clicking the pencil icon and typing “RefSeq” in the “Name” field that appears within the center panel. Click the Save button.
4. Transform coordinates of genes into coordinates of putative promoters.
  - a. Click “Operate on genomic intervals” in the Tools menu list on the left panel.
  - b. Click “Get flanks”. A new interface will appear in the center panel.
  - c. Make sure the “Select data” dropdown is set to dataset #2 RefSeq.
  - d. Set “Length of the flanking region/s” to “1000”.
  - e. Click “Execute”. A new dataset, #3, containing the putative promoters, will appear within the right panel.
  - f. Rename dataset #3 to “Promoters” by clicking the pencil icon and typing “Promoters” in the “Name” field that will appear within the center panel. Save.

5. Remove unnecessary columns for dataset #3. Only five columns are needed from this dataset. Galaxy's Cut tool allows the removal of unwanted columns.
  - a. Click "Text Manipulation" in the Tools menu list on the left panel.
  - b. Click "Cut columns from a table". A new interface will appear in the center panel.
  - c. Type "c1,c2,c3,c4,c6" in the "Cut columns" text box.
  - d. Click "Execute". A new dataset, #4, will appear within the right panel, containing only columns one through four and column six, which correspond to the chromosome, start position, end position, name and strand, respectively, for the calculated promoter regions.
  - e. Rename dataset #4 to "Clean Promoters" by clicking the pencil icon and typing "Clean Promoters" in the "Name" field that will appear within the center panel. Save.
  - f. Because the Cut tool breaks column assignment, the pencil icon will need to be clicked. A new interface will appear in the middle.
  - g. Use the "New Type" dropdown within the "Change data type" box to select "interval". The "interval" datatype describes data representing genomic coordinates (intervals).
  - h. Click the "Save" button immediately below the box. The upper part of the interface will change.
  - i. Set "Chrom column", "Start column" and "End column" to 1, 2 and 3, respectively. Check the "Strand" checkbox and select "5" from the adjacent dropdown. Check the "Name" checkbox and select "4" from adjacent dropdown.
  - j. Click the "Save" button immediately below.
  - k. The appearance of dataset #4 within the right panel will change -- column headers will appear, and the format will change to "interval".
6. Identify promoters containing the TAF1 binding sites. Now join the coordinates of TAF1 binding sites from dataset #1 with the coordinates of putative promoters from dataset #4. The Genomic Interval Join Tool matches two separate sets of genomic coordinates (intervals) according to their overlap, creating a single output containing the matched rows.
  - a. Click "Operate on Genomic Intervals" in the Tools menu list on the left panel.
  - b. Click "Join". A new interface will appear in the center panel.
  - c. Select dataset #4 Clean Promoters from the first dropdown called "Join".
  - d. Select dataset #1 TAF1_CHiP.txt from the second dropdown called "with".

- e. Make sure the “Return” dropdown is set to “Only records that are joined”.
  - f. Click “Execute”. A new dataset, #5, will appear in the right panel. This dataset will list coordinates of putative promoters and TAF1 binding sites side by side as shown in Figure 5.
7. Visualize results of this analysis using the UCSC Genome Browser.
  - a. Click “Graph/Display Data” in the Tools menu list on the left panel.
  - b. Click “Build custom track”. A new interface will appear in the center panel.
  - c. Click “Add new Track” and select dataset #1 TAF1_CHiP.txt from the “Dataset” dropdown.
  - d. Type “TAF1” in the “Name” and “Description” boxes. Leave other settings as default.
  - e. Click “Add new Track” to create Track 2, and select dataset #4 Clean Promoters from the “Dataset” dropdown.
  - f. Type “Promoters” in the “Name” and “Description” boxes. Set the color to Blue.
  - g. Click “Add new Track” to create Track 3 and select dataset #5 from the “Dataset” dropdown.
  - h. Type “Overlap” in the “Name” and “Description” boxes. Set the color to Purple.
  - i. Click “Execute”. A new dataset, #6 Build custom track on data 5, data 4, and data 1, will appear within the right panel.
  - j. Once dataset #6 becomes green in the History panel list, click on its name and then click the “display at UCSC main” link. The browser will open a new window or a tab with the UCSC Genome Browser interface displaying datasets 1, 4 and 5 as custom tracks.

Basic Protocol 2. Combining and filtering genome annotations: Finding Exons with the highest number of nucleotide polymorphisms

The objective of this protocol is to demonstrate joining, grouping, sorting, and filtering of genomic annotations in Galaxy. To explore these features using real data an illustrative example will be used: identification of exons containing the largest number of single nucleotide polymorphisms (SNPs).

Materials

- an internet accessible computer with any modern web browser (Firefox, Safari, Opera, IE).

Steps: *(Please note that it is beneficial to clear the current history and start re-numbering from 1 by accessing the History Options and selecting “Create a new empty history”. It simplifies following with the numbered steps.)*

1. Upload exon annotations from the UCSC Table Browser.
  - a. Click “Get Data” in the Tools menu list on the left panel.
  - b. Click “UCSC Main”. The UCSC Table Browser interface will be displayed in the middle panel.
  - c. Because the data of interest are human annotations make sure that “clade”, “genome” and “assembly” are set to “Mammal”, “Human” and “Mar. 2006”, respectively.
  - d. Set “group” to “Genes and Gene Prediction tracks” and “track” to “UCSC Genes”.
  - e. Select the radio button “position” and type “chr22” within the adjacent text box. This will limit the annotations to the entirety of chromosome 22.
  - f. Make sure “output format” is set to “BED - browser extensible data” and the checkbox by “Set output to Galaxy” is checked. The BED format is a specialized version of the interval format discussed earlier; it contains the information required to represent a genomic position.
  - g. Click “get output”. A new interface will appear.
  - h. Make sure the “Coding Exons” radio button is selected in the “Create one BED record per:” area. The genes on chromosome 22 will be divided into coding exons, with each exon having its own set of genomic intervals.
  - i. Click “Send query to Galaxy”. At this point a new dataset, #1, will appear in Galaxy's history on the right panel. A query in the queue will be indicated by a gray box, a running query will be yellow, and a completed query will have a green box.
  - j. When the query has completed, rename dataset #1 to “exons” by clicking the pencil icon and typing “exons” in the “Name” field that will appear within the center panel. Click the Save button.
2. Upload SNP coordinates.
  - a. Click “Get Data” from the Tools menu list on the left panel.
  - b. Click “UCSC Main”. The UCSC Table Browser interface will be displayed in the middle panel.
  - c. Because the data of interest are human annotations, make sure that “clade”, “genome” and “assembly” are set to “Mammal”, “Human” and “Mar. 2006”, respectively.
  - d. Set “group” to “Variation and Repeats” and “track” to “SNPs (129)”.

- e. Select the radio button “position” and type “chr22” within the adjacent text box.
        - f. Make sure “output format” is set to “BED - browser extensible data” and a checkbox by “Set output to Galaxy” is checked.
        - g. Click “get output”. A new interface will appear.
        - h. Make sure the “Whole Gene” radio button is selected in the “Create one BED record per:” area.
        - i. Click “Send query to Galaxy”. At this point, a new dataset, #2, will appear in Galaxy's history in the right panel.
        - j. When the query has completed, rename dataset #2 to “snps” by clicking the pencil icon and typing “snps” in the “Name” field that will appear within the center panel. Save.
3. Join coordinates of exons with coordinates of SNPs to identify those exons that contain SNPs.
  - a. Click “Operate on Genomic Intervals” in the Tools menu list on the left panel.
  - b. Click “Join” in the Operate submenu. A new interface will appear in the center panel.
  - c. Select dataset #1 exons from the first dropdown called “Join”.
  - d. Select dataset #2 snps from the second dropdown called “with”.
  - e. Make sure the “Return” dropdown is set to “Only records that are joined (INNER JOIN)”.
  - f. Click the “Execute” button. A new dataset, #3 Join on data 2 and data 1, will appear in the right panel. This dataset will list coordinates of exons and SNPs side by side as shown in Figure 7. The data can be examined by clicking on the name.
4. Count the number of SNPs per exon using the Group tool. In Figure 7, it is seen that if an exon contains multiple SNPs its name is repeated. It is possible to take advantage of this by using the Group tool. By counting the number of times each exon's name appears within dataset #3, the number of SNPs within that exon will be obtained.
  - a. Click “Join, Subtract, and Group” from the Tools menu list on the left panel.
  - b. Click “Group”. A new interface will appear in the center panel.
  - c. Set the “Select data:” dropdown to dataset #3 Join on data 2 and data 1.
  - d. Set “Group by column” to “c4”, as this column contains exon identifiers.

- e. Click the “Add new operation” button. A new section of interface named “Operation 1” will appear below.
        - f. Within the new interface section, set “Type” to “Count” and “On column” to “c4”.
        - g. Click the “Execute” button. A new dataset, #4 Group on data 3, will appear in the right panel. It will contain two columns: (1) exon id and (2) SNP count.
5. Sort exon by SNP count. To see the highest possible number of SNPs per exon in this dataset, sort the dataset from the previous step.
  - a. Click “Filter and Sort” from the Tools menu list on the left panel.
  - b. Click “Sort”. A new interface will appear in the center panel.
  - c. Set “Sort Query” to dataset #4 Group on data 3.
  - d. Set “on column” to “c2” (the SNP count calculated above).
  - e. Click “Execute”. Dataset #5 Sort on data 4 will appear in the history panel.
  - f. Click on the eye icon to see which exons have the highest SNP count.
6. Restrict dataset #5 to exons that have ten or more SNPs.
  - a. Click “Filter and Sort” from the Tools menu list on the left panel.
  - b. Click “Filter”.
  - c. Set “Filter” to dataset #5 Sort on data 4.
  - d. Set “With following condition” to “c2 >= 10” (without the quotes). This is because column 2 (c2) contains the count of SNPs per exon; only rows in which the contents of column two is greater-than or equal-to ten will be kept.
  - e. Click “Execute”. Dataset #6 Filter on data 5 will appear in the history panel.
7. Restore genomic location for exons containing ten or more SNPs. The previous step has produced a list of exons containing ten or more SNPs, however, information about their genomic position, strand orientation, etc has been lost. Because dataset #6 contains the exon identifier field, it can be used to restore genomic context information by joining with dataset #1. The Join two Queries tool is different than the Genomic Operations Join which was used earlier; this tool matches two separate datasets by matching column contents between any tab-delimited dataset (including interval datasets).
  - a. Click “Join, Subtract, and Group” from the Tools menu list on the left panel.

Steps

1. Rename the history. All histories are given the default name “unnamed”, which is obviously not very descriptive. This is easily changed by following the steps below.
  - a. Click the pencil icon immediately above the history items (on a lavender background). A text box will appear to the left of the pencil.
  - b. Type “Exons and SNPs” in the textbox and hit the Enter key or return key on the keyboard.
2. Click the “Options” button above the history panel. A list of history actions will appear in the middle panel.
3. Click the “Share current history” link
4. Enter the e-mail address of an existing Galaxy user and click “Submit”. This history is now shared.

Basic Protocol 3. Generating a workflow from a history in Galaxy

Protocols 1 and 2 demonstrate interactive analysis in Galaxy, the result is a “history” which documents each step of an analysis. Galaxy also allows the construction of reusable multi-step analysis “workflows”. In this protocol, the creation of a workflow from an existing analysis history is demonstrated.

Materials

- an internet accessible computer with any modern web browser (Firefox, Safari, Opera, IE).
- completed protocols – the history that was created by following Basic Protocol 2.
- A Galaxy account (created by clicking “Register” in the Galaxy interface). All workflow in Galaxy requires the user to be logged in with an account.

Steps

1. Ensure a non-empty history is loaded (for this example, the history resulting from the completion of Basic Protocol 2 is used).
2. In the header of the History panel (top-right of the Galaxy analysis interface) click “Options”. This will load a menu of options, which apply to the current history, in the center panel.
3. In the center panel, click “Construct workflow from the current history”. This will load a list of the actions (tool runs) which generated each dataset in the current history. A subset of tools can be selected by clicking the checkboxes on this page (e.g. if more than one analysis has been performed in the current history, but a workflow is only to be created from one of them). Certain tools cannot be used in workflows (including most external data sources), in these cases the dataset can be treated as an input to the workflow. Here, a workflow is constructed from the entire history, so do not change any checkboxes.

4. Provide a name for the new workflow by entering a name of choice in the text box underneath the label “Workflow name”.
5. Click the “Create Workflow” button to create the new workflow; a message will be displayed in the center panel confirming that the workflow was created.

Support Protocol 3.1. Modify a parameter of the workflow in Galaxy

After constructing a workflow from an existing analysis, the Workflow Editor can be used to modify tool parameters (or even add and remove steps).

Materials

- an internet accessible computer with any modern web browser (Firefox, Safari, Opera, IE).
- completed protocols – the workflow that was created by following Basic Protocol 3.

Steps

1. Move from the “Analyze Data” view to the “Workflow” view by clicking “Workflow” in the top panel of the Galaxy interface. The workflow view provides access to all workflow management functionality (editing, sharing, etc).
2. Load the workflow in the workflow editor.
  - a. Click the triangle next to the name of the workflow that was just created.
  - b. Select “edit” from the menu that appears.
3. Drag the workflow canvas (center panel) until the box labeled “Filter” is visible. Each box in the canvas represents a step of the workflow. The canvas viewport can be moved by dragging the background or by dragging the blue box in the overview panel (bottom right).
4. Click the box for the “Filter” step in the canvas, it will be outlined in blue, showing that it is the active step, and a form showing the tool parameters will appear in the right panel.
5. Modify this step to filter regions with 50 or more SNPs by entering the text “c2 >= 50” (without the quotes) in the textbox under the label “With the following condition” and clicking “Save” in the right panel.
6. Save the changes to the workflow by clicking “Save” in the header of the center panel. It is important to save both the changes to the step and the changed step to the workflow.

Support Protocol 3.2 Running workflows with Galaxy

Once a workflow has been constructed, it can be run in the analysis view just like any other tool in Galaxy.

Materials

- an internet accessible computer with any modern web browser (Firefox, Safari, Opera, IE).
- completed protocols – the workflow that was saved in Basic Protocol 3.

Steps

1. Return to the Analyze Data view by clicking “Analyze Data” in the top panel.
2. Create a new empty history in which to store the result of running the workflow.
  - a. In the header of the History panel (top-right of the Galaxy analysis interface) click “Options”.
  - b. In the center panel click “Create a new empty history”
3. Get exon and SNP annotations for human chromosome X from the UCSC Table Browser.
  - a. Follow Steps 1 and 2 of Basic Protocol 2; however, enter “chrX” instead of “chr22”.
4. Click “Workflow” at the bottom of the tool menu (left panel), then click “All workflows” in the list of options that appears.
5. In the center panel, click the name of the workflow created above in Basic Protocol 3. This will load the workflow in the center panel with prompts for parameters that need values.
6. Under “Step 1: Input Dataset” select the second item in the history (the SNPs).
7. Under “Step 2: Input Dataset” select the first item in the history (the exons).
8. Click “Run Workflow” at the bottom of the form in the center panel. A message will be displayed confirming that the workflow has been run, and the datasets for each workflow step will be added to the history (in the 'queued' state). At this point, the workflow is running, and each step will execute once the data it requires has been generated by previous steps. The box surrounding the dataset will change color based upon its state as the steps progress: a query in the queue will be indicated by a gray box, a running query will be yellow and a completed query will have a green box.

Support Protocol 3.3 Sharing workflows with Galaxy

Galaxy allows researchers to share workflows with others. Workflows can either be shared with a specific Galaxy user, or made publicly accessible by a special link.

Materials

- an internet accessible computer with any modern web browser (Firefox, Safari, Opera, IE).
- completed protocols – the workflow that was created using Basic Protocol 3

Steps

1. Move from the “Analyze Data” view to the “Workflow” view by clicking “Workflow” in the top panel of the Galaxy interface. This is located in the dark blue navigation banner across the top of the interface.
2. Click the triangle next to the name of the workflow to be shared and select “Sharing”.
3. To allow others to import the workflow by following a link.
  - a. Click “Enable import via link”.
  - b. Copy the link that is displayed.
  - c. Provide the link to anyone wanting to access the workflow (e.g. via email, including in a publication, etc).
4. To share the workflow only with a specific user.
  - a. Click “Share with another user”.
  - b. Enter the email address of a user to share with in the textbox.
  - c. Click “Share”. If the email address corresponds to another Galaxy user, they will now see the workflow in their workflow view.

Basic Protocol 4. Generating workflows from scratch with Galaxy

In addition to creating workflows from existing histories, Galaxy allows the creation of a workflow from scratch. In this protocol a simple workflow that finds the 50 longest intervals from a dataset in a 6 column BED file is constructed. The BED format is a specialized version of the interval format discussed earlier; it contains the information required to represent a genomic position. A 6 column BED file contains the chromosome, start position in the chromosome, end position in the chromosome, name, score, and strand for a set of genomic positions. Insert Figure 9 here.

Materials

- an internet accessible computer with any modern web browser (Firefox, Safari, Opera, IE).
- A Galaxy account (created by clicking “Register” in the Galaxy interface). All workflow in Galaxy requires the user to be logged in with an account.

Steps

1. Move from the “Analyze Data” view to the “Workflow” view by clicking “Workflow” in the top panel of the Galaxy interface.
2. Create a new empty workflow.
  - a. Click the “Add a new workflow” button near the top.
  - b. Enter a name for the workflow in the text box under the label “Workflow name”.

- c. Click “Create”.
3. Load the (empty) workflow in the workflow editor.
  - a. Click the triangle next to the name of the workflow that was just created.
  - b. Select “edit” from the menu that appears.
4. Add an input dataset to the workflow.
  - a. In the right panel, click “Inputs” from the Tools menu on the left, and then select “Input Dataset”. A box (called a “node”) will appear in the middle of the center panel (called the “Workflow canvas”).
  - b. Move the node representing the input dataset to the top-left of the editor canvas by dragging it by the title.
5. Add a step to the workflow to compute the length of each interval in the input dataset.
  - a. Click “Text Manipulation” from the Tools menu on the left and then select “Compute” in the left panel.
  - b. Drag the newly created node up and place it to the right of “Input Dataset” (leaving some space). Note: it is also possible to click “Layout” in the top of the center panel to automatically organize nodes in the canvas.
6. Create a connection between the input dataset and the Compute node. Outputs of a node are represented by circled arrowheads overlapping the right edge of a node, while data inputs are circled arrowheads overlapping the left edge of a node. Connections are made by dragging.
  - a. Click and hold the arrowhead next to the label “output” in the “Input Dataset” node.
  - b. Drag the mouse, a curve should follow the mouse pointer.
  - c. Drag over the arrowhead next to the label “as a new column to” in the “Compute” node. The curve should turn green indicating that the connection is valid (datatypes are compatible).
  - d. Release the mouse to make the connection.
7. Add a step to the workflow to sort the intervals by length.
  - a. Click “Filter and Sort” from the Tools menu on the left and then choose “Sort” in the left panel.
  - b. Position the new node to the right of the previously created nodes by dragging.
  - c. Create a connection between the output labels “out_file1” of “Compute” with the input “Sort query” of “Sort”.
8. Add a step to the workflow to select the longest intervals.

- a. Click “Text Manipulation” from the Tools menu on the left and then “Select first lines from a Query in the left panel.
  - b. Position the new node to the right of the previously created nodes by dragging.
  - c. Create a connection between the output labels “out_file1” of “Sort” with the input “from” of “Select first”.
9. Edit the parameters of the “Compute” step to calculate interval length.
  - a. Click the “Compute” node in the canvas. In the right panel a new form will appear. The parameters for this workflow action can be edited using the text boxes and menu choices.
  - b. In the text box under “on column” enter “c3 - c2” (without the quotes) to subtract column 2 (start position) from column 3 (end position). Note that this may already be the default value for this field.
  - c. Click the “Save” button in the right panel to validate and save the parameters.
10. Edit the parameters of the “Sort” step to sort on the correct column.
  - a. Click the “Sort” node in the canvas. Its action options will appear in the right panel form.
  - b. In the text box under “on column” enter “7”. Since the data is a 6 column BED file, the length computed in the “Compute” step will have been stored in column 7.
  - c. Click “Save” in the right panel to validate and save the parameters.
11. Edit the parameters of the “Select first” step to select the first 50 intervals.
  - a. Click the “Select first” node in the canvas. Its action options will appear in the right panel.
  - b. In the text box under “Select first” enter “50”.
  - c. Click “Save” in the right panel to validate and save the parameters.
12. Click the “Save” button in the title bar header of the center workflow canvas panel to save the workflow as a whole.
13. Click “Close” in the header of the workflow canvas panel to return to the workflow list. This workflow can now be run in the same fashion as described in Support protocol 3.2.

Basic Protocol 5. Extracting Sequences and alignments with Galaxy: A SNPs in exons example

This protocol demonstrates how Galaxy is used to extract genomic sequences and multiple species alignments corresponding to regions of interest. It starts with the data that was

generated in Basic Protocol 2, where human coding exons with high SNP counts were found. Two types of data will be extracted for these regions: the genomic sequence of each region, and pieces of a whole genome alignment between human and other species overlapping these regions. Because the whole genome alignment used here (produced by Multiz, a local aligner) is fragmented into pieces, these pieces will then be projected back onto the regions of interest (exons) to facilitate per-exon analysis of the alignments (the result is sometimes called a “pseudo-global” alignment. This protocol is a brief illustration of how easily the often tricky manipulation of these files is with Galaxy.

Materials

- an internet accessible computer with any modern web browser (Firefox, Safari, Opera, IE).
- completed protocols – the completed and saved the history that was created by following Basic Protocol 2 and Support Protocol 2.1.

Steps

1. Return to the main Galaxy interface by going to <http://usegalaxy.org>.
2. Load the history created in Basic Protocol 2:
  - a. Click “Options”, located at the top right of the History Panel; History Options will appear in the middle panel.
  - b. Click “List previously stored histories”; all histories associated with the current user will be displayed.
  - c. Click on the history that was saved earlier as “Exons and SNPs”.
3. The history panel will refresh with the selected history, which should contain 7 steps. Dataset #7 Join two Queries on data 6 and data 1 contains the genomic coordinates of the exons of interest.
4. Extract Genomic DNA corresponding to each of the exons.
  - a. In the Tools panel, click to expand the “Fetch Sequences” menu.
  - b. Click the item titled “Extract Genomic DNA”; the tool’s interface will appear in the center.
  - c. Select dataset #7 Join two Queries on data 6 and data 1 as the input query.
  - d. Set output datatype to “FASTA”.
  - e. Click “Execute”.
5. When the query finishes, a dataset containing one human sequence for each of the exons (total of 109 sequences) called “Extract Genomic DNA on data 7” is created. This dataset contains the human genomic DNA corresponding to each of the 109 exons in FASTA format, a very common format for storing multiple named sequences.
6. Extract multiple species alignment blocks for each of the human exon locations.

- a. In the Tools panel, click to expand the “Fetch Alignments” menu.
  - b. Click the item titled “Extract MAF blocks”; the tool’s interface will appear in the center. MAF stands for multiple alignment format, a standard format for large alignments of multiple genomes. A genome-wide local alignment of multiple genomes consists of many regions of high scoring alignment called “blocks”.
  - c. Select dataset #7 Join two Queries on data 6 and data 1 as the interval source.
  - d. Set MAF Source to “Locally Cached Alignments”.
  - e. Under Choose alignments, set “5-way multiZ” using the pulldown menu. This alignment contains 5 different mammals including human. Most of the locally cached multiple-species whole-genome alignments available in Galaxy were generated using the UCSC/Penn State Bioinformatics comparative genomic alignment pipeline; these original source alignments are available from the UCSC download site.
  - f. Click the “Select All” button to choose to extract all species.
  - g. Click the “Execute” button.
7. A new history item is created, #9 Extract MAF blocks on data 7, which contains the portions of the source alignment which overlap with the exon regions. For the 109 regions, 387 alignment blocks were retrieved, which is due to multiple local alignment blocks overlapping individual exons. Thus, the resulting dataset contains every local alignment block overlapping an exon, trimmed to just include the portion of the alignment that overlapped. This dataset is useful for examining the conservation of exons in aggregate, however the relationship between exons and alignments has been lost.
8. Create one projected alignment per human exon.
  - a. In the Tools panel, ensure that the “Fetch Alignments” menu is still expanded.
  - b. Click the item titled “Stitch MAF blocks”; the tool’s interface will appear in the center.
  - c. Select dataset #7 Join two Queries on data 6 and data 1 as the input intervals.
  - d. Set MAF Source to “Alignments in Your History”.
  - e. Select dataset #9 Extract MAF blocks on data 7 for MAF File.
  - f. Click the “Select All” button to choose to extract all species.
  - g. Click “Execute”.
9. A new history item is created, #10 Stitch MAF blocks on data 7 and data 9, which contains one alignment block for each of the human exons, with regions where no

alignment was found represented as gaps (-). Click the eye icon to examine the data in the center panel. The projected alignment is in FASTA format, suitable for downstream analysis in most phylogenetic software packages, including those available in Galaxy. For 109 regions, 545 FASTA sequences (109 regions each with sequences for 5 species) were generated in 109 alignment blocks.

Commentary

Galaxy successfully bridges the gap between data collection and analysis. The public Galaxy server allows researchers across the globe to perform computationally intensive, large-scale analyses with the only equipment requirement consisting of an internet-connected web browser. Users are not required to delve into the intricacies of how to execute a large collection of unrelated programs, but instead have access to a unified point and click interface. Galaxy provides both experimental biologists and their computational brethren with a framework to facilitate truly reproducible cutting edge science.

The protocols contained within this unit offer only a glimpse of possible analyses and tool functionality. The text contained here should only be considered an introduction to performing complex analysis with Galaxy. New datasets, tools, and features will be added regularly. Some new menu choices may arise or move. In addition to the screencasts that accompany these protocols; many more screencasts that demonstrate additional functionality are available at <http://galaxycast.org> and others will be added over time.

Transparency and reproducibility

Open and transparent research is essential to the process of science. Research papers cannot be published without making the protocols and generated experimental data publically available. Unfortunately, the same standards are often not applied to computational analysis. When analysis is performed within Galaxy, every detail is preserved in the “history” and can be inspected later. These histories can be shared or published, and can be reproduced (with or without modification) through the workflow system. Thus, without additional effort on the part of the user, Galaxy facilitates greater transparency and reproducibility of computational analyses.

Collaboration

While the scope of this unit is limited to introducing a user to performing data analysis with the public Galaxy server, Galaxy is also an excellent resource for collaborative analysis. Because it is web-based, collaborators at different locations can easily and rapidly share data and analysis. In particular, Galaxy’s library system provides for sharing of datasets within research groups, with access controls and version histories.

Research groups that have their own collections of analysis scripts and binaries will find it worthwhile to download the open source framework, integrate their unique tools, and maintain a private server (a “Galaxy instance”) for lab members to work on their projects. A local Galaxy server makes collaborations between computational and experimental researchers more efficient, since new analysis tools can be effortlessly made available to colleagues, allowing programmers to focus on method development. Although beyond the

scope of this introduction to the user interface, documentation and assistance for programmers is also available on the Galaxy site.

The Galaxy Framework is easily downloaded, quickly configured and effortlessly deployed. Although written in Python, no knowledge of the Python programming language is required to deploy or maintain a personal Galaxy instance. This facilitates local development of new tools, the creation of new Galaxy instances with custom toolsets, and secure private Galaxy instances for analyzing protected data (e.g., genotype data obtained in clinical setting). To download the Galaxy Framework and view detailed installation documentation visit <http://getgalaxy.org>.

Help and feedback

Galaxy is under constant development and is improved based upon user suggestions. Extensive help is available in the form of screencasts as well as active public mailing lists, where both experimentalists and computationalists can request and receive advice. Discussion of feature requests is also encouraged. For links to these resources and to use Galaxy, visit <http://galaxyproject.org>.

Acknowledgments

A vision for Galaxy was originally articulated by Ross Hardison, who is also the major source of support and critical feedback. We thank Jim Kent and David Haussler for their continuing support and making UCSC Genome Browser uplink and connection possible. Istvan Albert pioneered initial aspects of Galaxy design. The following individuals contributed to Galaxy project at different stages: Richard Burhans, Laura Elnitski, Belinda Giardine, Bob Harris, Jianbin He, Webb Miller, Cathy Riemer, and Yi Zhang. We thank Warren Lathe of OpenHelix for critical reading of the manuscript. Galaxy hardware is maintained by Nate Coraor. Robert Castelo, France Denoeud, Roderic Guigo, Erika Kvikstad, Julien Lagarde, and Kateryna Makova provided critical comments during software testing. This work is supported by funds provided by the Eberly College of Science, Huck Institutes of the Life Sciences at Penn State University, NSF BD&I grant 0543285 to AN and JT, NIH-NHGRI grant R01 HG004909 to AN and JT as well as funds from the Pennsylvania Department of Public Health.

Literature cited

- Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, Miller W, Pedersen JS, Pohl A, Raney BJ, Rhead B, Rosenbloom KR, Smith KE, Stanke M, Thakapallayil A, Trumbower H, Wang T, Zweig AS, Haussler D, Kent WJ. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.* 2008 Jan;36:D773–9. [PubMed: 18086701]
- Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 2004 Jan 1; 32(Database issue):D493–6. [PubMed: 14681465]
- Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B. A high-resolution map of active promoters in the human genome. *Nature.* 2005 Aug 11;436:876–80. [PubMed: 15988478]
- Taylor J, Schenck I, Blankenberg D, Nekrutenko A. Using galaxy to perform large-scale interactive data analyses. *Curr Protoc Bioinformatics.* 2007 Sep;Chapter 10(Unit 10.5)

Screencasts

1. Finding promoters containing TAF1 binding sites identified from a ChIP-seq experiment

<http://galaxycast.org/cpmb-2009-1>

Suppose you performed a ChIP-seq experiment and identified a series of genomic regions that bind TAF1-protein. You now need to identify a list of genes that contain such sites. This can be easily done with Galaxy in just a few steps. To run this example yourself you will need a file containing genomic coordinates for TAF1-binding sites from the ChIP-seq experiment (an example file can be downloaded from http://galaxy.psu.edu/CPMB/TAF1_CHiP.txt).

2. Finding Exons with the highest number of nucleotide polymorphisms

<http://galaxycast.org/cpmb-2009-2>

In this protocol we will work with two sets of genomic features: exon coordinates and single nucleotide polymorphism (SNP) coordinates. The objective will be to identify exons that contain the highest number of SNPs. Even if you are not particularly interested in exons and SNPs, we encourage you to follow along as the real goal of this protocol is to introduce you to joining, grouping, sorting, and filtering in Galaxy.

3. Saving your results and sharing data with others

<http://galaxycast.org/cpmb-2009-3>

How can you ensure that the analyses you have just conducted are safely stored and that you are able to go back to them at anytime? You will need to create a free account within Galaxy. This is the only requirement to save your analyses. The protocol below explains this and also introduces sharing analyses with colleagues.

4. Generating a workflow from a history

<http://galaxycast.org/cpmb-2009-4>

Screencasts 1 and 2 within this series demonstrate interactive analysis in Galaxy, the result is a **history** which documents each step of an analysis. Galaxy also allows you to construct reusable multi-step analysis **workflows**. In this protocol, we will demonstrate the creation of a workflow from an existing analysis history.

5. Generating workflows from scratch

<http://galaxycast.org/cpmb-2009-5>

In addition to creating workflows from existing histories, Galaxy allows us to create a workflow from scratch. In this protocol we will construct a simple workflow that finds the 50 longest intervals from a dataset in a 6 column BED file.

6. Extracting Sequences and alignments: A SNPs in exons example

<http://galaxycast.org/cpmb-2009-6>

Here we demonstrate how Galaxy is used to extract genomic sequences and multiple species alignments corresponding to regions of interest. In this protocol, we start with the data that we generated in Screencast 2, where we found human coding exons with high SNP counts. We will extract genomic sequences for our regions in FASTA format as well as obtain corresponding alignment blocks from multiple species whole genome alignments in their native format and also create a pseudo-global alignment in FASTA format, which is suitable for use in existing analysis software.

Figure 1.

Galaxy's Analyze Data interface consists of four regions: the masthead (a) at the top, the tool menu (b) on the left-hand side, the work area (c) in the middle and the history panel (d) on the right. The Get Data section has been expanded in the tool menu and the Upload File tool has been selected. In the work area, a local file containing TAF1 ChIP-seq data has been chosen (Basic Protocol 1, step 1); clicking the "Execute" button will cause the data to be uploaded and appear in the history panel.

Figure 2.

To change the properties of a dataset (Basic Protocol 1, step 2), click on the question mark (or the pencil icon) associated with our dataset in the history panel (a). This causes the Edit Attributes page to appear in the center panel (b) where the datatype has been changed from tabular to interval; clicking “Save” causes the page to refresh allowing additional interval-specific information to be set (c).

a

Home Genomes Genome Browser Blast Tables Gene Sorter PCR Session FAQ Help

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to resequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in the User's Guide for general information and sample queries, and the [OpenTable Tutorial](#) for a narrated presentation of software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). Refer to it for the list of contributors and usage restrictions associated with these data.

clade: ☐ Human ☐ Mouse ☐ Rat ☐ Dog ☐ Pig ☐ Chicken ☐ Zebrafish ☐ Fruit fly ☐ Worm ☐ Yeast ☐ Other

genome: ☐ Human ☐ Mouse ☐ Rat ☐ Dog ☐ Pig ☐ Chicken ☐ Zebrafish ☐ Fruit fly ☐ Worm ☐ Yeast ☐ Other

assembly: ☐ Mar-2006 ☐ Apr-2004 ☐ May-2004 ☐ Jun-2004 ☐ Jul-2004 ☐ Aug-2004 ☐ Sep-2004 ☐ Oct-2004 ☐ Nov-2004 ☐ Dec-2004 ☐ Jan-2005 ☐ Feb-2005 ☐ Mar-2005 ☐ Apr-2005 ☐ May-2005 ☐ Jun-2005 ☐ Jul-2005 ☐ Aug-2005 ☐ Sep-2005 ☐ Oct-2005 ☐ Nov-2005 ☐ Dec-2005 ☐ Jan-2006 ☐ Feb-2006 ☐ Mar-2006 ☐ Apr-2006 ☐ May-2006 ☐ Jun-2006 ☐ Jul-2006 ☐ Aug-2006 ☐ Sep-2006 ☐ Oct-2006 ☐ Nov-2006 ☐ Dec-2006 ☐ Jan-2007 ☐ Feb-2007 ☐ Mar-2007 ☐ Apr-2007 ☐ May-2007 ☐ Jun-2007 ☐ Jul-2007 ☐ Aug-2007 ☐ Sep-2007 ☐ Oct-2007 ☐ Nov-2007 ☐ Dec-2007 ☐ Jan-2008 ☐ Feb-2008 ☐ Mar-2008 ☐ Apr-2008 ☐ May-2008 ☐ Jun-2008 ☐ Jul-2008 ☐ Aug-2008 ☐ Sep-2008 ☐ Oct-2008 ☐ Nov-2008 ☐ Dec-2008 ☐ Jan-2009 ☐ Feb-2009 ☐ Mar-2009 ☐ Apr-2009 ☐ May-2009 ☐ Jun-2009 ☐ Jul-2009 ☐ Aug-2009 ☐ Sep-2009 ☐ Oct-2009 ☐ Nov-2009 ☐ Dec-2009 ☐ Jan-2010 ☐ Feb-2010 ☐ Mar-2010 ☐ Apr-2010 ☐ May-2010 ☐ Jun-2010 ☐ Jul-2010 ☐ Aug-2010 ☐ Sep-2010 ☐ Oct-2010 ☐ Nov-2010 ☐ Dec-2010 ☐ Jan-2011 ☐ Feb-2011 ☐ Mar-2011 ☐ Apr-2011 ☐ May-2011 ☐ Jun-2011 ☐ Jul-2011 ☐ Aug-2011 ☐ Sep-2011 ☐ Oct-2011 ☐ Nov-2011 ☐ Dec-2011 ☐ Jan-2012 ☐ Feb-2012 ☐ Mar-2012 ☐ Apr-2012 ☐ May-2012 ☐ Jun-2012 ☐ Jul-2012 ☐ Aug-2012 ☐ Sep-2012 ☐ Oct-2012 ☐ Nov-2012 ☐ Dec-2012 ☐ Jan-2013 ☐ Feb-2013 ☐ Mar-2013 ☐ Apr-2013 ☐ May-2013 ☐ Jun-2013 ☐ Jul-2013 ☐ Aug-2013 ☐ Sep-2013 ☐ Oct-2013 ☐ Nov-2013 ☐ Dec-2013 ☐ Jan-2014 ☐ Feb-2014 ☐ Mar-2014 ☐ Apr-2014 ☐ May-2014 ☐ Jun-2014 ☐ Jul-2014 ☐ Aug-2014 ☐ Sep-2014 ☐ Oct-2014 ☐ Nov-2014 ☐ Dec-2014 ☐ Jan-2015 ☐ Feb-2015 ☐ Mar-2015 ☐ Apr-2015 ☐ May-2015 ☐ Jun-2015 ☐ Jul-2015 ☐ Aug-2015 ☐ Sep-2015 ☐ Oct-2015 ☐ Nov-2015 ☐ Dec-2015 ☐ Jan-2016 ☐ Feb-2016 ☐ Mar-2016 ☐ Apr-2016 ☐ May-2016 ☐ Jun-2016 ☐ Jul-2016 ☐ Aug-2016 ☐ Sep-2016 ☐ Oct-2016 ☐ Nov-2016 ☐ Dec-2016 ☐ Jan-2017 ☐ Feb-2017 ☐ Mar-2017 ☐ Apr-2017 ☐ May-2017 ☐ Jun-2017 ☐ Jul-2017 ☐ Aug-2017 ☐ Sep-2017 ☐ Oct-2017 ☐ Nov-2017 ☐ Dec-2017 ☐ Jan-2018 ☐ Feb-2018 ☐ Mar-2018 ☐ Apr-2018 ☐ May-2018 ☐ Jun-2018 ☐ Jul-2018 ☐ Aug-2018 ☐ Sep-2018 ☐ Oct-2018 ☐ Nov-2018 ☐ Dec-2018 ☐ Jan-2019 ☐ Feb-2019 ☐ Mar-2019 ☐ Apr-2019 ☐ May-2019 ☐ Jun-2019 ☐ Jul-2019 ☐ Aug-2019 ☐ Sep-2019 ☐ Oct-2019 ☐ Nov-2019 ☐ Dec-2019 ☐ Jan-2020 ☐ Feb-2020 ☐ Mar-2020 ☐ Apr-2020 ☐ May-2020 ☐ Jun-2020 ☐ Jul-2020 ☐ Aug-2020 ☐ Sep-2020 ☐ Oct-2020 ☐ Nov-2020 ☐ Dec-2020 ☐ Jan-2021 ☐ Feb-2021 ☐ Mar-2021 ☐ Apr-2021 ☐ May-2021 ☐ Jun-2021 ☐ Jul-2021 ☐ Aug-2021 ☐ Sep-2021 ☐ Oct-2021 ☐ Nov-2021 ☐ Dec-2021 ☐ Jan-2022 ☐ Feb-2022 ☐ Mar-2022 ☐ Apr-2022 ☐ May-2022 ☐ Jun-2022 ☐ Jul-2022 ☐ Aug-2022 ☐ Sep-2022 ☐ Oct-2022 ☐ Nov-2022 ☐ Dec-2022 ☐ Jan-2023 ☐ Feb-2023 ☐ Mar-2023 ☐ Apr-2023 ☐ May-2023 ☐ Jun-2023 ☐ Jul-2023 ☐ Aug-2023 ☐ Sep-2023 ☐ Oct-2023 ☐ Nov-2023 ☐ Dec-2023 ☐ Jan-2024 ☐ Feb-2024 ☐ Mar-2024 ☐ Apr-2024 ☐ May-2024 ☐ Jun-2024 ☐ Jul-2024 ☐ Aug-2024 ☐ Sep-2024 ☐ Oct-2024 ☐ Nov-2024 ☐ Dec-2024 ☐ Jan-2025 ☐ Feb-2025 ☐ Mar-2025 ☐ Apr-2025 ☐ May-2025 ☐ Jun-2025 ☐ Jul-2025 ☐ Aug-2025 ☐ Sep-2025 ☐ Oct-2025 ☐ Nov-2025 ☐ Dec-2025 ☐ Jan-2026 ☐ Feb-2026 ☐ Mar-2026 ☐ Apr-2026 ☐ May-2026 ☐ Jun-2026 ☐ Jul-2026 ☐ Aug-2026 ☐ Sep-2026 ☐ Oct-2026 ☐ Nov-2026 ☐ Dec-2026 ☐ Jan-2027 ☐ Feb-2027 ☐ Mar-2027 ☐ Apr-2027 ☐ May-2027 ☐ Jun-2027 ☐ Jul-2027 ☐ Aug-2027 ☐ Sep-2027 ☐ Oct-2027 ☐ Nov-2027 ☐ Dec-2027 ☐ Jan-2028 ☐ Feb-2028 ☐ Mar-2028 ☐ Apr-2028 ☐ May-2028 ☐ Jun-2028 ☐ Jul-2028 ☐ Aug-2028 ☐ Sep-2028 ☐ Oct-2028 ☐ Nov-2028 ☐ Dec-2028 ☐ Jan-2029 ☐ Feb-2029 ☐ Mar-2029 ☐ Apr-2029 ☐ May-2029 ☐ Jun-2029 ☐ Jul-2029 ☐ Aug-2029 ☐ Sep-2029 ☐ Oct-2029 ☐ Nov-2029 ☐ Dec-2029 ☐ Jan-2030 ☐ Feb-2030 ☐ Mar-2030 ☐ Apr-2030 ☐ May-2030 ☐ Jun-2030 ☐ Jul-2030 ☐ Aug-2030 ☐ Sep-2030 ☐ Oct-2030 ☐ Nov-2030 ☐ Dec-2030 ☐ Jan-2031 ☐ Feb-2031 ☐ Mar-2031 ☐ Apr-2031 ☐ May-2031 ☐ Jun-2031 ☐ Jul-2031 ☐ Aug-2031 ☐ Sep-2031 ☐ Oct-2031 ☐ Nov-2031 ☐ Dec-2031 ☐ Jan-2032 ☐ Feb-2032 ☐ Mar-2032 ☐ Apr-2032 ☐ May-2032 ☐ Jun-2032 ☐ Jul-2032 ☐ Aug-2032 ☐ Sep-2032 ☐ Oct-2032 ☐ Nov-2032 ☐ Dec-2032 ☐ Jan-2033 ☐ Feb-2033 ☐ Mar-2033 ☐ Apr-2033 ☐ May-2033 ☐ Jun-2033 ☐ Jul-2033 ☐ Aug-2033 ☐ Sep-2033 ☐ Oct-2033 ☐ Nov-2033 ☐ Dec-2033 ☐ Jan-2034 ☐ Feb-2034 ☐ Mar-2034 ☐ Apr-2034 ☐ May-2034 ☐ Jun-2034 ☐ Jul-2034 ☐ Aug-2034 ☐ Sep-2034 ☐ Oct-2034 ☐ Nov-2034 ☐ Dec-2034 ☐ Jan-2035 ☐ Feb-2035 ☐ Mar-2035 ☐ Apr-2035 ☐ May-2035 ☐ Jun-2035 ☐ Jul-2035 ☐ Aug-2035 ☐ Sep-2035 ☐ Oct-2035 ☐ Nov-2035 ☐ Dec-2035 ☐ Jan-2036 ☐ Feb-2036 ☐ Mar-2036 ☐ Apr-2036 ☐ May-2036 ☐ Jun-2036 ☐ Jul-2036 ☐ Aug-2036 ☐ Sep-2036 ☐ Oct-2036 ☐ Nov-2036 ☐ Dec-2036 ☐ Jan-2037 ☐ Feb-2037 ☐ Mar-2037 ☐ Apr-2037 ☐ May-2037 ☐ Jun-2037 ☐ Jul-2037 ☐ Aug-2037 ☐ Sep-2037 ☐ Oct-2037 ☐ Nov-2037 ☐ Dec-2037 ☐ Jan-2038 ☐ Feb-2038 ☐ Mar-2038 ☐ Apr-2038 ☐ May-2038 ☐ Jun-2038 ☐ Jul-2038 ☐ Aug-2038 ☐ Sep-2038 ☐ Oct-2038 ☐ Nov-2038 ☐ Dec-2038 ☐ Jan-2039 ☐ Feb-2039 ☐ Mar-2039 ☐ Apr-2039 ☐ May-2039 ☐ Jun-2039 ☐ Jul-2039 ☐ Aug-2039 ☐ Sep-2039 ☐ Oct-2039 ☐ Nov-2039 ☐ Dec-2039 ☐ Jan-2040 ☐ Feb-2040 ☐ Mar-2040 ☐ Apr-2040 ☐ May-2040 ☐ Jun-2040 ☐ Jul-2040 ☐ Aug-2040 ☐ Sep-2040 ☐ Oct-2040 ☐ Nov-2040 ☐ Dec-2040 ☐ Jan-2041 ☐ Feb-2041 ☐ Mar-2041 ☐ Apr-2041 ☐ May-2041 ☐ Jun-2041 ☐ Jul-2041 ☐ Aug-2041 ☐ Sep-2041 ☐ Oct-2041 ☐ Nov-2041 ☐ Dec-2041 ☐ Jan-2042 ☐ Feb-2042 ☐ Mar-2042 ☐ Apr-2042 ☐ May-2042 ☐ Jun-2042 ☐ Jul-2042 ☐ Aug-2042 ☐ Sep-2042 ☐ Oct-2042 ☐ Nov-2042 ☐ Dec-2042 ☐ Jan-2043 ☐ Feb-2043 ☐ Mar-2043 ☐ Apr-2043 ☐ May-2043 ☐ Jun-2043 ☐ Jul-2043 ☐ Aug-2043 ☐ Sep-2043 ☐ Oct-2043 ☐ Nov-2043 ☐ Dec-2043 ☐ Jan-2044 ☐ Feb-2044 ☐ Mar-2044 ☐ Apr-2044 ☐ May-2044 ☐ Jun-2044 ☐ Jul-2044 ☐ Aug-2044 ☐ Sep-2044 ☐ Oct-2044 ☐ Nov-2044 ☐ Dec-2044 ☐ Jan-2045 ☐ Feb-2045 ☐ Mar-2045 ☐ Apr-2045 ☐ May-2045 ☐ Jun-2045 ☐ Jul-2045 ☐ Aug-2045 ☐ Sep-2045 ☐ Oct-2045 ☐ Nov-2045 ☐ Dec-2045 ☐ Jan-2046 ☐ Feb-2046 ☐ Mar-2046 ☐ Apr-2046 ☐ May-2046 ☐ Jun-2046 ☐ Jul-2046 ☐ Aug-2046 ☐ Sep-2046 ☐ Oct-2046 ☐ Nov-2046 ☐ Dec-2046 ☐ Jan-2047 ☐ Feb-2047 ☐ Mar-2047 ☐ Apr-2047 ☐ May-2047 ☐ Jun-2047 ☐ Jul-2047 ☐ Aug-2047 ☐ Sep-2047 ☐ Oct-2047 ☐ Nov-2047 ☐ Dec-2047 ☐ Jan-2048 ☐ Feb-2048 ☐ Mar-2048 ☐ Apr-2048 ☐ May-2048 ☐ Jun-2048 ☐ Jul-2048 ☐ Aug-2048 ☐ Sep-2048 ☐ Oct-2048 ☐ Nov-2048 ☐ Dec-2048 ☐ Jan-2049 ☐ Feb-2049 ☐ Mar-2049 ☐ Apr-2049 ☐ May-2049 ☐ Jun-2049 ☐ Jul-2049 ☐ Aug-2049 ☐ Sep-2049 ☐ Oct-2049 ☐ Nov-2049 ☐ Dec-2049 ☐ Jan-2050 ☐ Feb-2050 ☐ Mar-2050 ☐ Apr-2050 ☐ May-2050 ☐ Jun-2050 ☐ Jul-2050 ☐ Aug-2050 ☐ Sep-2050 ☐ Oct-2050 ☐ Nov-2050 ☐ Dec-2050 ☐ Jan-2051 ☐ Feb-2051 ☐ Mar-2051 ☐ Apr-2051 ☐ May-2051 ☐ Jun-2051 ☐ Jul-2051 ☐ Aug-2051 ☐ Sep-2051 ☐ Oct-2051 ☐ Nov-2051 ☐ Dec-2051 ☐ Jan-2052 ☐ Feb-2052 ☐ Mar-2052 ☐ Apr-2052 ☐ May-2052 ☐ Jun-2052 ☐ Jul-2052 ☐ Aug-2052 ☐ Sep-2052 ☐ Oct-2052 ☐ Nov-2052 ☐ Dec-2052 ☐ Jan-2053 ☐ Feb-2053 ☐ Mar-2053 ☐ Apr-2053 ☐ May-2053 ☐ Jun-2053 ☐ Jul-2053 ☐ Aug-2053 ☐ Sep-2053 ☐ Oct-2053 ☐ Nov-2053 ☐ Dec-2053 ☐ Jan-2054 ☐ Feb-2054 ☐ Mar-2054 ☐ Apr-2054 ☐ May-2054 ☐ Jun-2054 ☐ Jul-2054 ☐ Aug-2054 ☐ Sep-2054 ☐ Oct-2054 ☐ Nov-2054 ☐ Dec-2054 ☐ Jan-2055 ☐ Feb-2055 ☐ Mar-2055 ☐ Apr-2055 ☐ May-2055 ☐ Jun-2055 ☐ Jul-2055 ☐ Aug-2055 ☐ Sep-2055 ☐ Oct-2055 ☐ Nov-2055 ☐ Dec-2055 ☐ Jan-2056 ☐ Feb-2056 ☐ Mar-2056 ☐ Apr-2056 ☐ May-2056 ☐ Jun-2056 ☐ Jul-2056 ☐ Aug-2056 ☐ Sep-2056 ☐ Oct-2056 ☐ Nov-2056 ☐ Dec-2056 ☐ Jan-2057 ☐ Feb-2057 ☐ Mar-2057 ☐ Apr-2057 ☐ May-2057 ☐ Jun-2057 ☐ Jul-2057 ☐ Aug-2057 ☐ Sep-2057 ☐ Oct-2057 ☐ Nov-2057 ☐ Dec-2057 ☐ Jan-2058 ☐ Feb-2058 ☐ Mar-2058 ☐ Apr-2058 ☐ May-2058 ☐ Jun-2058 ☐ Jul-2058 ☐ Aug-2058 ☐ Sep-2058 ☐ Oct-2058 ☐ Nov-2058 ☐ Dec-2058 ☐ Jan-2059 ☐ Feb-2059 ☐ Mar-2059 ☐ Apr-2059 ☐ May-2059 ☐ Jun-2059 ☐ Jul-2059 ☐ Aug-2059 ☐ Sep-2059 ☐ Oct-2059 ☐ Nov-2059 ☐ Dec-2059 ☐ Jan-2060 ☐ Feb-2060 ☐ Mar-2060 ☐ Apr-2060 ☐ May-2060 ☐ Jun-2060 ☐ Jul-2060 ☐ Aug-2060 ☐ Sep-2060 ☐ Oct-2060 ☐ Nov-2060 ☐ Dec-2060 ☐ Jan-2061 ☐ Feb-2061 ☐ Mar-2061 ☐ Apr-2061 ☐ May-2061 ☐ Jun-2061 ☐ Jul-2061 ☐ Aug-2061 ☐ Sep-2061 ☐ Oct-2061 ☐ Nov-2061 ☐ Dec-2061 ☐ Jan-2062 ☐ Feb-2062 ☐ Mar-2062 ☐ Apr-2062 ☐ May-2062 ☐ Jun-2062 ☐ Jul-2062 ☐ Aug-2062 ☐ Sep-2062 ☐ Oct-2062 ☐ Nov-2062 ☐ Dec-2062 ☐ Jan-2063 ☐ Feb-2063 ☐ Mar-2063 ☐ Apr-2063 ☐ May-2063 ☐ Jun-2063 ☐ Jul-2063 ☐ Aug-2063 ☐ Sep-2063 ☐ Oct-2063 ☐ Nov-2063 ☐ Dec-2063 ☐ Jan-2064 ☐ Feb-2064 ☐ Mar-2064 ☐ Apr-2064 ☐ May-2064 ☐ Jun-2064 ☐ Jul-2064 ☐ Aug-2064 ☐ Sep-2064 ☐ Oct-2064 ☐ Nov-2064 ☐ Dec-2064 ☐ Jan-2065 ☐ Feb-2065 ☐ Mar-2065 ☐ Apr-2065 ☐ May-2065 ☐ Jun-2065 ☐ Jul-2065 ☐ Aug-2065 ☐ Sep-2065 ☐ Oct-2065 ☐ Nov-2065 ☐ Dec-2065 ☐ Jan-2066 ☐ Feb-2066 ☐ Mar-2066 ☐ Apr-2066 ☐ May-2066 ☐ Jun-2066 ☐ Jul-2066 ☐ Aug-2066 ☐ Sep-2066 ☐ Oct-2066 ☐ Nov-2066 ☐ Dec-2066 ☐ Jan-2067 ☐ Feb-2067 ☐ Mar-2067 ☐ Apr-2067 ☐ May-2067 ☐ Jun-2067 ☐ Jul-2067 ☐ Aug-2067 ☐ Sep-2067 ☐ Oct-2067 ☐ Nov-2067 ☐ Dec-2067 ☐ Jan-2068 ☐ Feb-2068 ☐ Mar-2068 ☐ Apr-2068 ☐ May-2068 ☐ Jun-2068 ☐ Jul-2068 ☐ Aug-2068 ☐ Sep-2068 ☐ Oct-2068 ☐ Nov-2068 ☐ Dec-2068 ☐ Jan-2069 ☐ Feb-2069 ☐ Mar-2069 ☐ Apr-2069 ☐ May-2069 ☐ Jun-2069 ☐ Jul-2069 ☐ Aug-2069 ☐ Sep-2069 ☐ Oct-2069 ☐ Nov-2069 ☐ Dec-2069 ☐ Jan-2070 ☐ Feb-2070 ☐ Mar-2070 ☐ Apr-2070 ☐ May-2070 ☐ Jun-2070 ☐ Jul-2070 ☐ Aug-2070 ☐ Sep-2070 ☐ Oct-2070 ☐ Nov-2070 ☐ Dec-2070 ☐ Jan-2071 ☐ Feb-2071 ☐ Mar-2071 ☐ Apr-2071 ☐ May-2071 ☐ Jun-2071 ☐ Jul-2071 ☐ Aug-2071 ☐ Sep-2071 ☐ Oct-2071 ☐ Nov-2071 ☐ Dec-2071 ☐ Jan-2072 ☐ Feb-2072 ☐ Mar-2072 ☐ Apr-2072 ☐ May-2072 ☐ Jun-2072 ☐ Jul-2072 ☐ Aug-2072 ☐ Sep-2072 ☐ Oct-2072 ☐ Nov-2072 ☐ Dec-2072 ☐ Jan-2073 ☐ Feb-2073 ☐ Mar-2073 ☐ Apr-2073 ☐ May-2073 ☐ Jun-2073 ☐ Jul-2073 ☐ Aug-2073 ☐ Sep-2073 ☐ Oct-2073 ☐ Nov-2073 ☐ Dec-2073 ☐ Jan-2074 ☐ Feb-2074 ☐ Mar-2074 ☐ Apr-2074 ☐ May-2074 ☐ Jun-2074 ☐ Jul-2074 ☐ Aug-2074 ☐ Sep-2074 ☐ Oct-2074 ☐ Nov-2074 ☐ Dec-2074 ☐ Jan-2075 ☐ Feb-2075 ☐ Mar-2075 ☐ Apr-2075 ☐ May-2075 ☐ Jun-2075 ☐ Jul-2075 ☐ Aug-2075 ☐ Sep-2075 ☐ Oct-2075 ☐ Nov-2075 ☐ Dec-2075 ☐ Jan-2076 ☐ Feb-2076 ☐ Mar-2076 ☐ Apr-2076 ☐ May-2076 ☐ Jun-2076 ☐ Jul-2076 ☐ Aug-2076 ☐ Sep-2076 ☐ Oct-2076 ☐ Nov-2076 ☐ Dec-2076 ☐ Jan-2077 ☐ Feb-2077 ☐ Mar-2077 ☐ Apr-2077 ☐ May-2077 ☐ Jun-2077 ☐ Jul-2077 ☐ Aug-2077 ☐ Sep-2077 ☐ Oct-2077 ☐ Nov-2077 ☐ Dec-2077 ☐ Jan-2078 ☐ Feb-2078 ☐ Mar-2078 ☐ Apr-2078 ☐ May-2078 ☐ Jun-2078 ☐ Jul-2078 ☐ Aug-2078 ☐ Sep-2078 ☐ Oct-2078 ☐ Nov-2078 ☐ Dec-2078 ☐ Jan-2079 ☐ Feb-2079 ☐ Mar-2079 ☐ Apr-2079 ☐ May-2079 ☐ Jun-2079 ☐ Jul-2079 ☐ Aug-2079 ☐ Sep-2079 ☐ Oct-2079 ☐ Nov-2079 ☐ Dec-2079 ☐ Jan-2080 ☐ Feb-2080 ☐ Mar-2080 ☐ Apr-2080 ☐ May-2080 ☐ Jun-2080 ☐ Jul-2080 ☐ Aug-2080 ☐ Sep-2080 ☐ Oct-2080 ☐ Nov-2080 ☐ Dec-2080 ☐ Jan-2081 ☐ Feb-2081 ☐ Mar-2081 ☐ Apr-2081 ☐ May-2081 ☐ Jun-2081 ☐ Jul-2081 ☐ Aug-2081 ☐ Sep-2081 ☐ Oct-2081 ☐ Nov-2081 ☐ Dec-2081 ☐ Jan-2082 ☐ Feb-2082 ☐ Mar-2082 ☐ Apr-2082 ☐ May-2082 ☐ Jun-2082 ☐ Jul-2082 ☐ Aug-2082 ☐ Sep-2082 ☐ Oct-2082 ☐ Nov-2082 ☐ Dec-2082 ☐ Jan-2083 ☐ Feb-2083 ☐ Mar-2083 ☐ Apr-2083 ☐ May-2083 ☐ Jun-2083 ☐ Jul-2083 ☐ Aug-2083 ☐ Sep-2083 ☐ Oct-2083 ☐ Nov-2083 ☐ Dec-2083 ☐ Jan-2084 ☐ Feb-2084 ☐ Mar-2084 ☐ Apr-2084 ☐ May-2084 ☐ Jun-2084 ☐ Jul-2084 ☐ Aug-2084 ☐ Sep-2084 ☐ Oct-2084 ☐ Nov-2084 ☐ Dec-2084 ☐ Jan-2085 ☐ Feb-2085 ☐ Mar-2085 ☐ Apr-2085 ☐ May-2085 ☐ Jun-2085 ☐ Jul-2085 ☐ Aug-2085 ☐ Sep-2085 ☐ Oct-2085 ☐ Nov-2085 ☐ Dec-2085 ☐ Jan-2086 ☐ Feb

a

Get flanks

Select data:
2: RefSeq

Region:
Whole feature

Location of the flanking region/s:
Upstream

Offset:
0
Use positive values to offset co-ordinates in opposite direction.

Length of the flanking region(s):
1000
Use non-negative value for length

Execute

b

Figure 4. Selecting the Get flanks tool (Basic Protocol 1, step 4) from the Operate on Genomic Intervals Section (a) allows the creation of new data containing the region 1000 nucleotides upstream of our RefSeq genes (b).

Join

Join:
4: Clean Prom
First query

with:
1: TAF1_CHIP.txt
Second query

with min overlap:
1
(bp)

Return:
Only records that are joined (INNER JOIN)

Execute

Figure 5.

The Join tool is used to create a dataset which contains the coordinates of putative promoters and TAF1 binding sites side by side (Basic Protocol 1, step 6).

Figure 6.

The Build custom track tool (Basic Protocol 1, step 7) allows the user to design a custom track suitable for display at the UCSC Genome Browser (d) by progressively adding new tracks containing varying datasets (a-c).

Figure 7. A dataset containing exons and overlapping SNPs has been created (Basic Protocol 2, step 4) using the Join tool and has been displayed in the middle panel by clicking on the eye icon next to dataset 3. A red rectangle has been drawn around an exon which overlaps with 4 SNPs.

The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

Workflow name
Workflow constructed from history 'Exons'

Create Workflow

Tool	History items created
UCSC Main <i>This tool cannot be used in workflows</i>	1: exons ☒ Treat as input dataset
UCSC Main <i>This tool cannot be used in workflows</i>	2: snps ☒ Treat as input dataset
Join ☒ Include "Join" in workflow	3: Join on data 2 and data 1
Group ☒ Include "Group" in workflow	4: Group on data 3
Sort ☒ Include "Sort" in workflow	5: Sort on data 4
Filter ☒ Include "Filter" in workflow	6: Filter on data 5
Join two Queries ☒ Include "Join two Queries" in workflow	7: Join two Queries on data 6 and data 1

Figure 8.

In order to create a workflow from an existing history (Basic Protocol 3), the user needs to make sure that they are logged in and then select "History Options" and click "Construct workflow from the current history". A new workflow will be populated from the current history as shown; the workflow can now be renamed and created.

Figure 9.

The Workflow Editor allows users to click to add new tools and connect the output of one tool to the input of another by simple clicking and dragging. The output of the Sort tool is being connected to the Select first tool (Basic Protocol 4, step 9), as is shown by the green rope; when the mouse button is released, the connection will be created and the rope will become white.

Extract MAF blocks

Choose intervals:
7: Join two Queries on data 6 and

MAF Source:
Locally Cached Alignments

Choose alignments:
5-way multiZ (hg18,panTro2,rheMac2)

Choose species:
Select All Unselect All

☒ hg18
☒ panTro2
☒ rheMac2
☒ mm8
☒ canFam2

Select species to be included in the final alignment

Execute

Stitch MAF blocks

Choose intervals:
7: Join two Queries on data 6 and

MAF Source:
Alignments in Your History

MAF File:
9: Extract MAF blocks on data 7

Choose species:
Select All Unselect All

☒ hg18
☒ panTro2
☒ rheMac2
☒ mm8
☒ canFam2

Select species to be included in the final alignment

Execute

a b

Figure 10.

Several options exist for obtaining multi-species alignments (Basic Protocol 5). The Extract MAF blocks tool (a) creates a MAF dataset which contains only the trimmed alignment blocks which overlap a specified set of intervals. The Stitch MAF blocks tool (b) creates a FASTA file which contains a single alignment block per provided interval.