

# Galaxy: A Web-Based Genome Analysis Tool for Experimentalists

UNIT 19.10

Daniel Blankenberg,<sup>1,5</sup> Gregory Von Kuster,<sup>1,5</sup> Nathaniel Coraor,<sup>1,5</sup>  
Guruprasad Ananda,<sup>1,5</sup> Ross Lazarus,<sup>2,5</sup> Mary Mangan,<sup>3</sup>  
Anton Nekrutenko,<sup>1,5</sup> and James Taylor<sup>4,5</sup>

<sup>1</sup>The Huck Institutes for the Life Sciences, Pennsylvania State University,  
University Park, Pennsylvania

<sup>2</sup>Channing Laboratory, Harvard Medical School, Boston, Massachusetts

<sup>3</sup>OpenHelix LLC, Bellevue, Washington

<sup>4</sup>Emory University, Atlanta, Georgia

<sup>5</sup>The Galaxy Team, Pennsylvania State University, University Park, Pennsylvania

## 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 distributed 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 laboratories. Either way, it allows experimentalists without informatics or programming expertise to perform complex large-scale analysis with just a Web browser. *Curr. Protoc. Mol. Biol.* 89:19.10.1-19.10.21. © 2010 by John Wiley & Sons, Inc.

Keywords: 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 laboratory 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 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.

Galaxy (<http://galaxyproject.org>) rectifies this challenge 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 computation 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. It is directed primarily at experimentalists, and makes use only of analysis tools available at the public

Informatics for  
Molecular  
Biologists

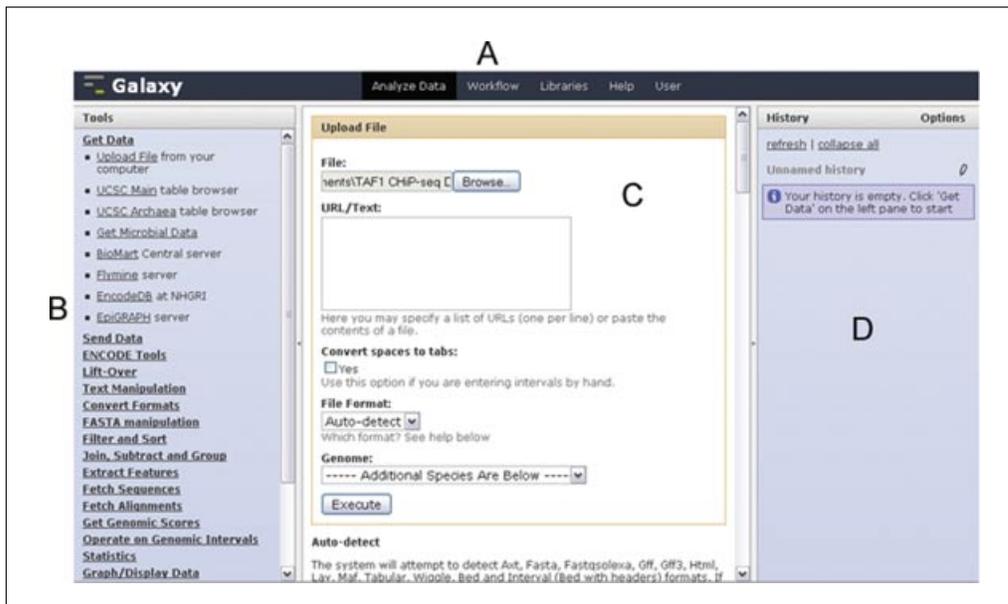
19.10.1

Supplement 89

Galaxy service at <http://usegalaxy.org>. Various components and tools of the public Galaxy server are 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 is an introduction to the Galaxy approach—finding promoters containing TAF1 binding sites identified from a ChIP-seq experiment. Basic Protocol 2 is a bit more data manipulation—finding coding exons with the most SNPs. Support Protocol 1 describes how to save results in Galaxy and share data with others. Basic Protocol 3 describes generating a workflow from a history in Galaxy. Support Protocol 2 describes modifying a parameter of the workflow in Galaxy. Support Protocol 3 describes running workflows with Galaxy. Support Protocol 4 describes sharing workflows with Galaxy. Basic Protocol 4 describes generating workflows from scratch with Galaxy. Basic Protocol 5 describes extracting sequences and alignments with Galaxy—SNPs in exons example.

These protocols cover the basic aspects of the functionality of Galaxy. 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 is beneficial to review 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 is 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 19.10.1 in detail will familiarize the user with the layout of Galaxy’s interface, including a history for the user and the tools menu.



**Figure 19.10.1** 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 (see Basic Protocol 1, step 1); clicking the Execute button will cause the data to be uploaded and appear in the history panel. See the TAF1 screencast (<http://galaxycast.org/cpmb-2009-1>) for more details.

## AN INTRODUCTION TO THE GALAXY APPROACH: FINDING PROMOTERS CONTAINING TAF1 BINDING SITES IDENTIFIED FROM A ChIP-Seq EXPERIMENT

### BASIC PROTOCOL 1

This protocol presents an example situation in which a ChIP-Seq experiment 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, it is assumed that the ChIP-Seq data has already been processed into putative binding regions, since this procedure is currently very experiment and laboratory specific. However, as best practices are defined for performing and evaluating the quality of these procedures, appropriate tools will be added to Galaxy.

A screencast of the protocol can be viewed at <http://galaxycast.org/cpmb-2009-1>.

**NOTE:** The items to alter are stated in the protocol. If other menus and options are not referenced, leave those settings in their default or existing condition.

### 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](http://galaxy.psu.edu/CPMB/TAF1_ChIP.txt); Kim et al., 2005)

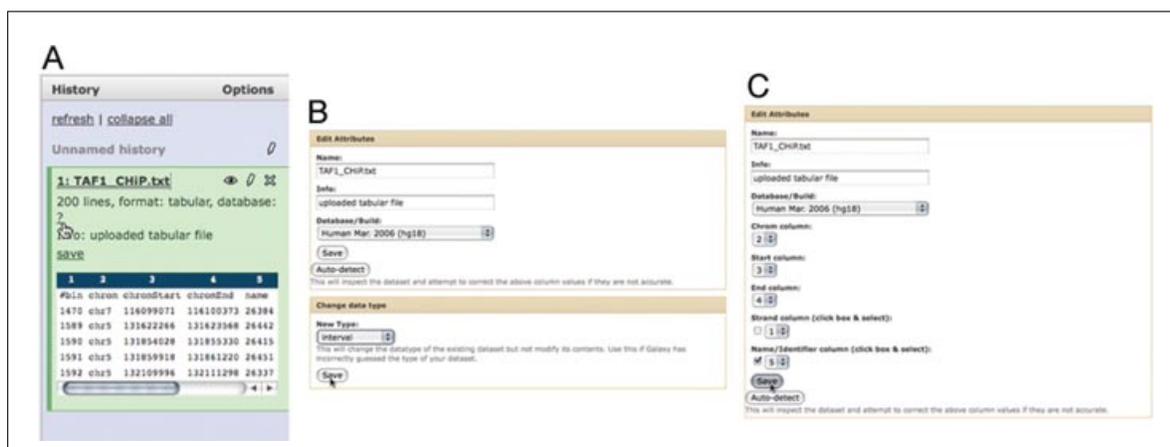
An internet-accessible computer with any modern Web browser (Firefox, Safari, Opera, Internet Explorer)

Informatics for  
Molecular  
Biologists

## 19.10.3

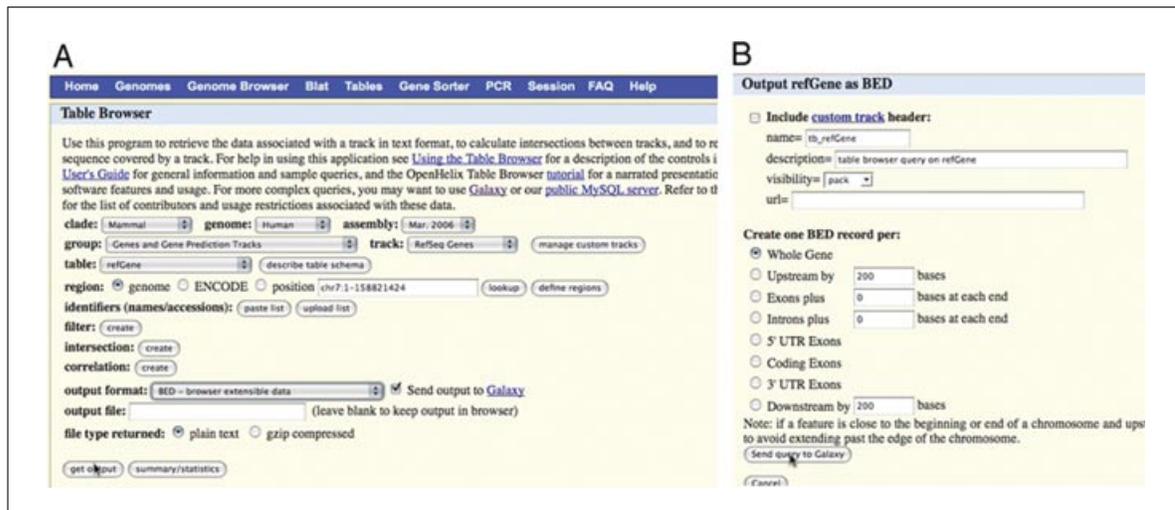
1. Upload the TAF1 ChIP-Seq data. Before beginning the analysis, the ChIP-Seq data needs to be uploaded into Galaxy's workspace (known as a user's "history" throughout this document).
  - a. Go to the public Galaxy site at <http://usegalaxy.org>.
  - b. Click Get Data.
  - c. Click Upload File.
  - d. The 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](http://galaxy.psu.edu/CPMB/TAF1_ChIP.txt)).
 

*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 no. 1 within the right panel.
2. Set properties of the TAF1 dataset (Fig. 19.10.2). 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 box and select 5 from the adjacent dropdown.



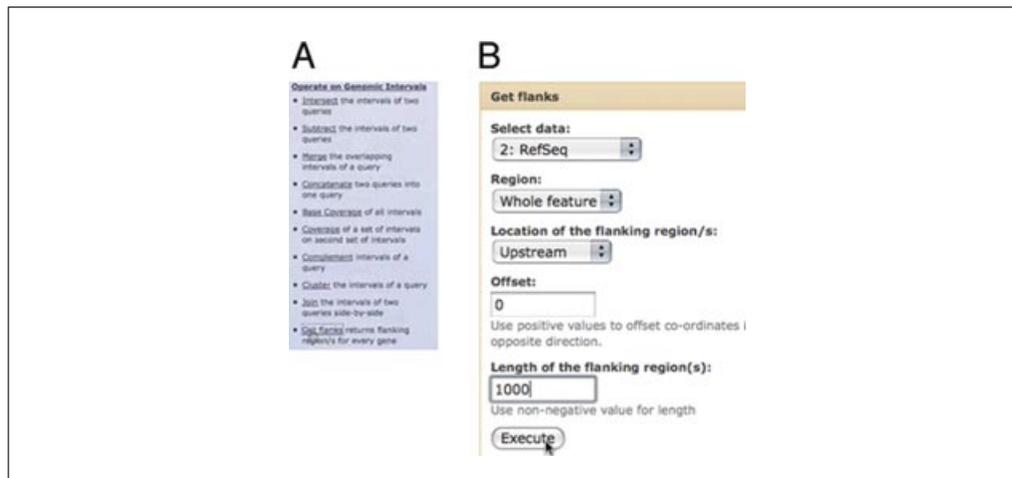
**Figure 19.10.2** To change the properties of a dataset (see Basic Protocol 1, step 2), click on the question mark (or the pencil icon) associated with the dataset in the history panel (A). This causes the Edit Attributes page to appear in the center panel (B) where the datatype has been changed from tabular to interval. Clicking Save causes the page to refresh, allowing additional interval-specific information to be set (C).

## 19.10.4



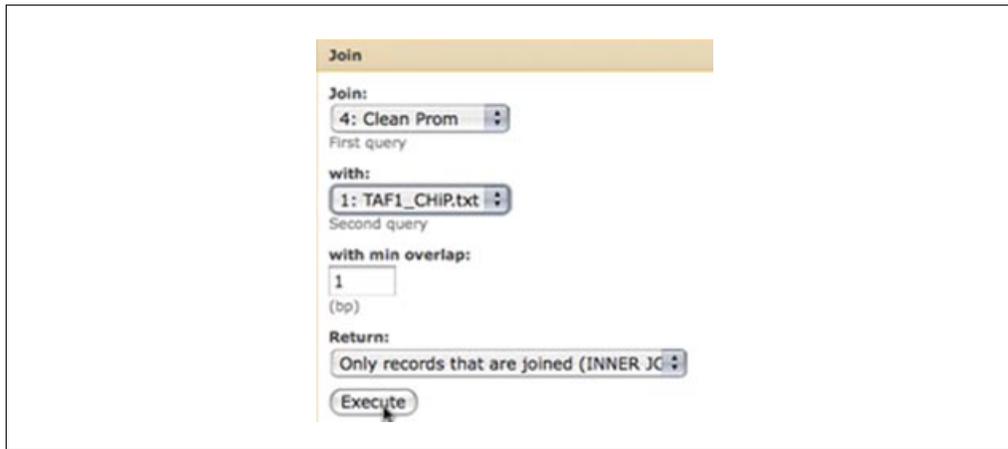
**Figure 19.10.3** The UCSC Table browser tool has been selected and its interface (**A**) appears in the center panel. The refGene table has been selected and the output is marked to be sent to Galaxy (see Basic Protocol 1, step 3). Once output style is specified (**B**), clicking Send query to Galaxy will create a new dataset in the history panel. The history item has been renamed to RefSeq after clicking on the pencil icon next to its name and making the required changes in the Edit Attributes page (see Fig. 19.10.2) which appears.

- h. Click the Save button immediately below.
  - i. The appearance of dataset no. 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 (Fig. 19.10.3; Karolchik et al., 2004, 2008). 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, no. 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 no. 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 (Fig. 19.10.4).
    - 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.



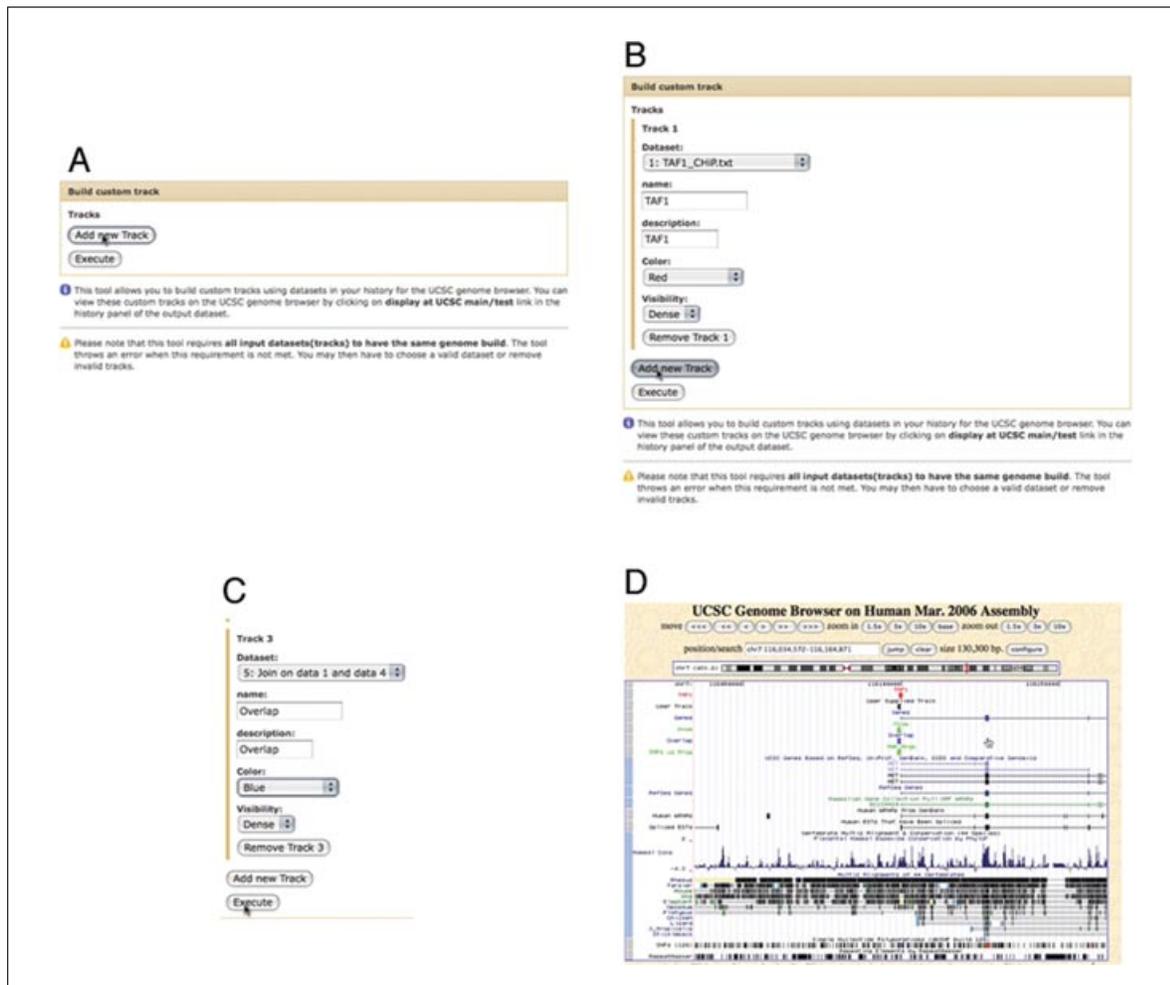
**Figure 19.10.4** Selecting the Get flanks tool (see Basic Protocol 1, step 4) from the Operate on Genomic Intervals Section (A) allows the creation of new data containing the region 1000 nucleotides upstream of our RefSeq genes (B).

- c. Make sure the Select Data dropdown is set to dataset no. 2 RefSeq.
  - d. Set Length of the Flanking Region/s to “1000”.
  - e. Click Execute. A new dataset, no. 3, containing the putative promoters, will appear within the right panel.
  - f. Rename dataset no. 3 to Promoters by clicking the pencil icon and typing “Promoters” in the Name field that will appear within the center panel. Click the Save button.
5. Remove unnecessary columns for dataset no. 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, no. 4, will appear within the right panel, containing only columns one through four and column six, which respectively correspond to the chromosome, start position, end position, name and strand for the calculated promoter regions.
  - e. Rename dataset no. 4 to Clean Promoters by clicking the pencil icon and typing “Clean Promoters” in the Name field that will appear within the center panel. Click the Save button.
  - f. Because the Cut tool breaks column assignment, click the pencil icon. 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 no. 4 within the right panel will change—column headers will appear, and the format will change to Interval.



**Figure 19.10.5** The Join tool is used to create a dataset that contains the coordinates of putative promoters and TAF1 binding sites side by side (see Basic Protocol 1, step 6).

6. Identify promoters containing the TAF1 binding sites. Now join the coordinates of TAF1 binding sites from dataset no. 1 with the coordinates of putative promoters from dataset no. 4. The Genomic Interval Join Tool (Fig. 19.10.5) 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 no. 4 Clean Promoters from the first dropdown called Join.
  - d. Select dataset no. 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, no. 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 19.10.5.
  
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 (Fig. 19.10.6). A new interface will appear in the center panel.
  - c. Click Add New Track and select dataset no. 1 TAF1\_ChIP.txt from the Dataset dropdown.
  - d. Type “TAF1” in the Name and Description boxes. Leave other settings as defaults.
  - e. Click Add New Track to create Track 2, and select dataset no. 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 no. 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, no. 6 Build custom track on data 5, data 4, and data 1, will appear within the right panel.
  - j. Once dataset no. 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.



**Figure 19.10.6** The Build custom track tool (see 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**).

**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 is presented: identification of exons containing the largest number of single nucleotide polymorphisms (SNPs).

A screencast of the protocol can be viewed at <http://galaxycast.org/cpmb-2009-2>.

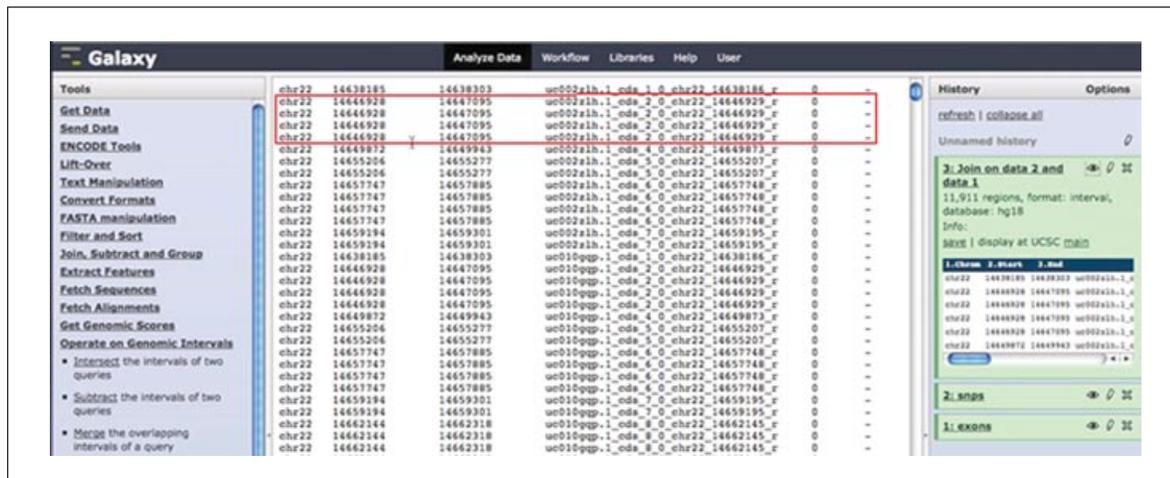
**Materials**

An internet-accessible computer with any modern Web browser (Firefox, Safari, Opera, Internet Explorer)

**NOTE:** It is beneficial to clear the current history and start re-numbering from 1 by accessing the History Options and selecting Create New. It simplifies following 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 send 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, no. 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 no. 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 send 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, no. 2, will appear in Galaxy’s history in the right panel.
  - j. When the query has completed, rename dataset no. 2 to snps by clicking the pencil icon and typing “snps” in the Name field that will appear within the center panel. Click the Save button.
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 no. 1 exons from the first dropdown called Join.



**Figure 19.10.7** A dataset containing exons and overlapping SNPs was created (see Basic Protocol 2, step 4) using the Join tool and 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 four SNPs. See the Exons and SNPs screencast (<http://galaxycast.org/cpmb-2009-2>) for more details.

- d. Select dataset no. 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, no. 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 19.10.7. The data can be examined by clicking on the name.
4. Count the number of SNPs per exon using the Group tool. In Figure 19.10.7, 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 no. 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 no. 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, no. 4 Group on data 3, will appear in the right panel. It will contain two columns: (1) exon identifier 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 no. 4 Group on data 3.
    - d. Set On Column to "c2" (the SNP count calculated above).
    - e. Click Execute. Dataset no. 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 no. 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 no. 5 Sort on data 4.
  - d. Set With Following Condition to  $c2 \geq 10$ . This is because column 2 (c2) contains the count of SNPs per exon; only rows in which the contents of column 2 is  $\geq 10$  will be kept.
  - e. Click Execute. Dataset no. 6 Filter on data 5 will appear in the history panel.
7. Restore genomic location for exons containing ten or more SNPs. Step 6 produced a list of exons containing ten or more SNPs; however, information about their genomic position, strand orientation, etc. has been lost. Because dataset no. 6 contains the exon identifier field, it can be used to restore genomic context information by joining with dataset no. 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.
  - b. Click Join two Queries. A new interface will appear within the center panel.
  - c. Set Join to dataset no. 1 exons.
  - d. Set Using Column to “c4” as this column contains exon identifiers in dataset no. 1.
  - e. Set With to dataset no. 6 Filter on data 5.
  - f. Set And Column to “c1” as this column contains exon identifiers in dataset no. 6.
  - g. Click Execute. Dataset no. 7 Join two Queries on data 6 and data 1 will appear within the history panel on the right. It contains full genomic context information about exons containing ten or more SNPs in chromosome 22.
8. Visualize dataset no. 7 Join two Queries on data 6 and data 1 in UCSC Genome Browser.
  - a. Go to the right (history) panel and expand dataset no. 7 by clicking on the name of the dataset.
  - b. Click the Display at UCSC main link. A new browser tab (or window) will open dataset no. 7 within the UCSC Genome Browser. The query will be displayed as a track called User Supplied Track. Access to menu control of this track is available in the menus area below, and the track will be available in the UCSC Table Browser for further query and manipulation.
9. To save the analysis and share it with colleagues continue on to Support Protocol 1.

## SAVING RESULTS IN GALAXY AND SHARING DATA WITH OTHERS

How can researchers ensure that the analyses they have just conducted are safely stored and that they are able to go back to them at anytime? They will need to create a free account within Galaxy. This is the only requirement to save analyses. The protocol below explains how to store results and introduces sharing analyses with colleagues. A screencast can be viewed at <http://galaxycast.org/cpmb-2009-3> to walk the user through the process.

## *SUPPORT PROTOCOL 1*

**Informatics for  
Molecular  
Biologists**

**19.10.11**

### **Materials**

An internet-accessible computer with any modern Web browser (Firefox, Safari, Opera, Internet Explorer)

Results from Basic Protocol 2

A Galaxy account (created by clicking Register in the Galaxy interface); histories must be linked to a user to be stored and shared

1. Rename the history. All histories are given the default name Unnamed, which is not very descriptive. Change name 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 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 link.
4. Enter the e-mail address of an existing Galaxy user and click Submit. This history is now shared.

### **GENERATING A WORKFLOW FROM A HISTORY IN GALAXY**

Basic Protocols 1 and 2 demonstrate interactive analysis in Galaxy, the result is a history that documents each step of an analysis. Galaxy also allows the construction of reusable multi-step analysis “workflows” (Fig. 19.10.8). In this protocol, the creation of a workflow from an existing analysis history is demonstrated.

A screencast of the protocols can be viewed at <http://galaxycast.org/cpmb-2009-4>.

### **Materials**

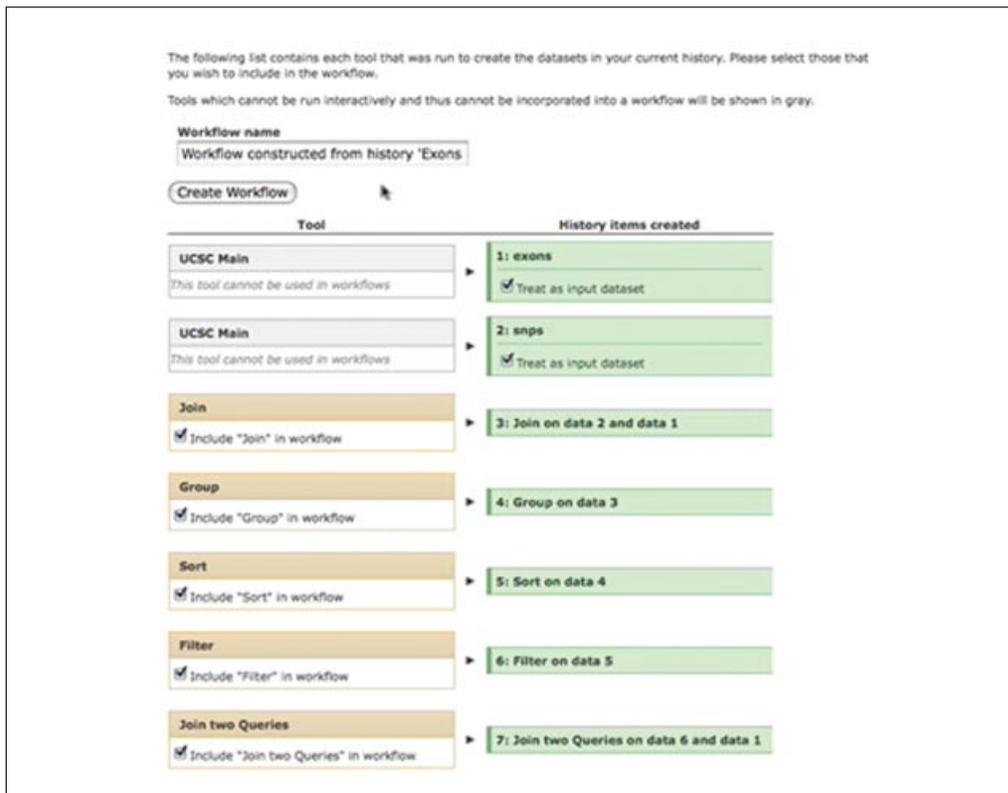
An internet-accessible computer with any modern Web browser (Firefox, Safari, Opera, Internet Explorer)

History created from Basic Protocol 2

A Galaxy account (created by clicking Register in the Galaxy interface); all workflow manipulation in Galaxy requires the user to be logged in with an account

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 that apply to the current history.
3. Click Extract Workflow. This will load a list of the actions (tool runs) that 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.



**Figure 19.10.8** To create a workflow from an existing history (see Basic Protocol 3), the user needs to make sure that they are logged in and then select History Options and click Extract Workflow. A new workflow will be populated from the current history as shown; the workflow can now be renamed and created. See the Workflow screencast (<http://galaxycast.org/cpmb-2009-4>) for more details.

- 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.

## MODIFY A PARAMETER IN 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, Internet Explorer)

Workflow created by Basic Protocol 3

- 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.).*

- Load the workflow in the workflow editor.
  - Click the triangle next to the name of the workflow that was just created.
  - Select Edit from the menu that appears.
- 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

## SUPPORT PROTOCOL 2

Informatics for  
Molecular  
Biologists

### 19.10.13

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” in the textbox under the label With the Following Condition.
6. Save the changes to the workflow by clicking Save in the header of the center panel.

## **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, Internet Explorer)  
Workflow saved in Basic Protocol 3

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. Click Create New.
3. Get exon and SNP annotations for human chromosome X from the UCSC Table Browser.
  - a. Follow Basic Protocol 2, steps 1 and 2; however, enter “chrX” instead of “chr22”.
4. Click Workflows 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 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.

## **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, Internet Explorer)

Workflow created by Basic Protocol 3

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, perform:
  - 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, perform:
  - 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.

### **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.

A screencast of this protocol can be viewed at <http://galaxycast.org/cpmb-2009-5>.

### **Materials**

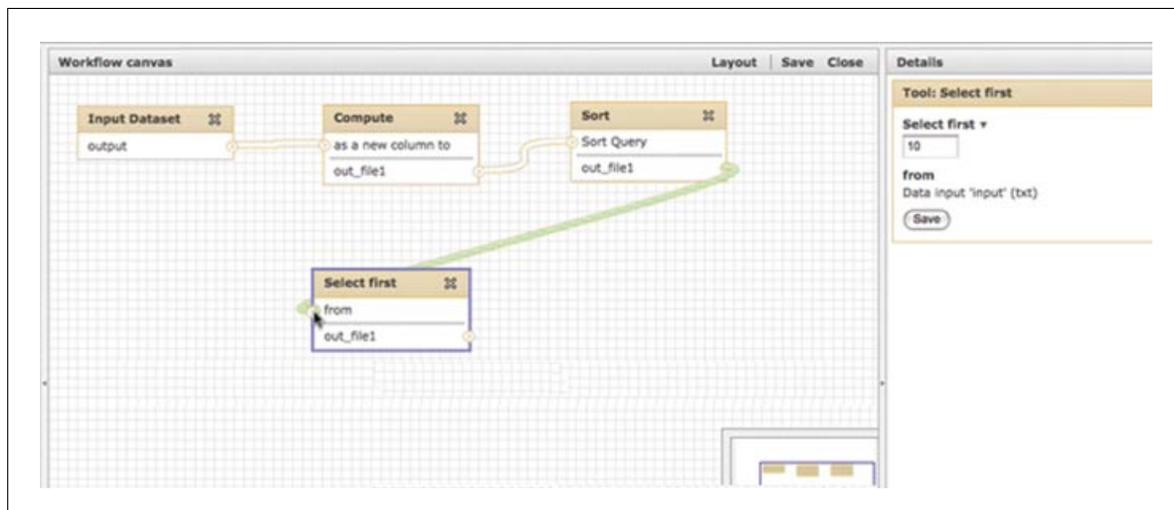
An internet-accessible computer with any modern Web browser (Firefox, Safari, Opera, Internet Explorer)

A Galaxy account (created by clicking Register in the Galaxy interface); all workflow manipulation in Galaxy requires the user to be logged in with an account

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 create 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.

### **BASIC PROTOCOL 4**

4. Add an input dataset to the workflow.
  - a. In the left 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). 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 (Fig. 19.10.9). 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.



**Figure 19.10.9** 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. Here, the output of the Sort tool is being connected to the Select first tool (see 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.

- 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 Add Expression, enter “c3-c2” to subtract column 2 (start position) from column 3 (end position). Note that this may already be the default value for this field.
  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.
  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.
  12. Click the Save button in the title bar header of the center workflow canvas panel to save the workflow.
  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.

### **EXTRACTING SEQUENCES AND ALIGNMENTS WITH GALAXY: AN 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 provides a brief illustration of how easily Galaxy can be used to handle the often tricky manipulation of these files.

A screencast of this protocol can be viewed at <http://galaxycast.org/cpmb-2009-6>.

**BASIC  
PROTOCOL 5**

**Informatics for  
Molecular  
Biologists**

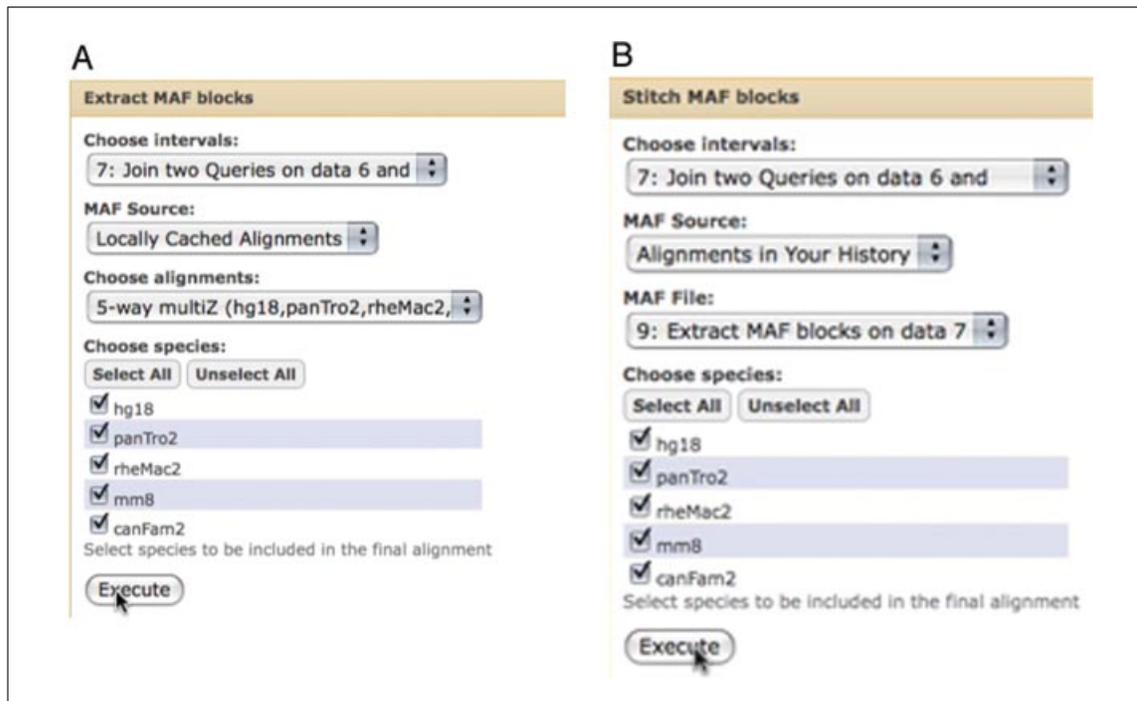
**19.10.17**

## Materials

An internet-accessible computer with any modern Web browser (Firefox, Safari, Opera, Internet Explorer)

Completed and saved history created by Basic Protocol 2 and Support Protocol 1

1. Go to the main Galaxy interface at <http://usegalaxy.org>.
2. Load the history created in Basic Protocol 2:
  - a. Click Options, located at the top right of the History panel.
  - b. Click saved 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 seven steps. Dataset no. 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 no. 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 entitled Extract MAF Blocks; the tool's interface will appear in the center (Fig. 19.10.10A). MAF is defined as the 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 no. 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 dropdown menu. This alignment contains five 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, no. 9 Extract MAF blocks on data 7, which contains the portions of the source alignment that overlap with the exon regions. For the



**Figure 19.10.10** Several options exist for obtaining multi-species alignments (see Basic Protocol 5). The Extract MAF blocks tool (**A**) creates a MAF dataset, which contains only the trimmed alignment blocks that 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. See the SeqAlign screencast (<http://galaxycast.org/cpmb-2009-6>) for more details.

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 (Fig. 19.10.10B).
  - c. Select dataset no. 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 no. 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, no. 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 five 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 colleagues 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 herein should only be considered as an introduction to performing complex analysis with Galaxy. New datasets, tools, and features are 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 analyses. In particular, Galaxy’s library system provides for sharing of datasets within research groups, complete 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 laboratory 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>.

### ***Acknowledgements***

A vision for Galaxy was originally articulated by Ross Hardison, who is also the major source of support and critical feedback. The authors 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 the Galaxy design. The following individuals contributed to the Galaxy project at different stages: Richard Burhans, Laura Elnitski, Belinda Giardiane, Bob Harris, Jianbin He, Webb Miller, Cathy Riemer, and Yi Zhang. The authors 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 Pennsylvania State University, NSF BD&I grant 0543285, NIH-NHGRI grant R01 HG004909 as well as funds from the Pennsylvania Department of Public Health.

### Literature Cited

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32:D493-D496.

Karolchik, D., Kuhn, R.M., Baertsch, R., Barber, G.P., Clawson, H., Diekhans, M., Giardine, B., Harte, R.A., Hinrichs, A.S., Hsu, F., Miller, W., Pedersen, J.S., Pohl, A., Raney, B.J., Rhead, B., Rosenbloom, K.R., Smith, K.E., Stanke, M., Thakkapallayil, A., Trumbower, H., Wang, T., Zweig, A.S., Haussler, D., and Kent, W.J. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.* 36:D773-D779.

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D., and Ren, B. 2005. A high-resolution map of active promoters in the human genome. *Nature* 436:876-880.

Taylor, J., Schenck, I., Blankenberg, D., and Nekrutenko, A. 2007. Using galaxy to perform large-scale interactive data analyses. *Curr. Protoc. Bioinformatics* 19:10.5.1-10.5.25.