Some Phenotype Association Tools in Galaxy: Looking for Disease SNPs in a Full Genome

Belinda M. Giardine,[1] Cathy Riemer,[1] Richard Burhans,[1] Aakrosh Ratan,[1] and Webb Miller[1]

[1] Pennsylvania State University, University Park, Pennsylvania

ABSTRACT

This unit focuses on some of the tools available on the public Galaxy server that are useful for exploring possible associations between human genetic variants and phenotypes. We trace step-by-step through an example illustrating several methods for examining a single full-coverage genome to look for single-nucleotide polymorphisms (SNPs) that are either known to be associated with disease or suspected to have impact for other reasons. It makes use of public genomic data, tools designed specifically for working with variants, and also some general tools for text manipulation and operations on genomic coordinates. *Curr. Protoc. Bioinform.* 39:15.2.1-15.2.27. © 2012 by John Wiley & Sons, Inc.

Keywords: disease • SNP • genome variation • coding • non-coding • gene-based analysis • Web application

INTRODUCTION

Galaxy (*http://galaxyproject.org*; Giardine et al., 2005; Blankenberg et al., 2010; Goecks et al., 2010) is a software framework that provides Web-based tools for bioinformatics, including tasks useful in the analysis of human variation. The developers maintain a public server at Penn State, and the software is also freely available for local installation and customization. Galaxy is highly extensible, so as new tools become available (not necessarily written specifically for Galaxy), they can be added to increase the power and flexibility of the system.

This unit focuses on some of the tools available on the public Galaxy server that are useful for exploring possible associations between human genetic variants and phenotypes. We trace step-by-step through an example illustrating several methods for examining a single full-coverage genome to look for single-nucleotide polymorphisms (SNPs) that are either known to be associated with disease or suspected to have impact for other reasons. It makes use of public genomic data, tools designed specifically for working with variants, and also some general tools for text manipulation and operations on genomic coordinates. A version of this tutorial is also available in HTML format (see Internet Resources, below). For a more basic introduction to using Galaxy in general, please see UNIT 10.5, "Using Galaxy to Perform Large-Scale Interactive Data Analyses."

For this example, we will use an artificial dataset consisting of the SNP calls from the Complete Genomics genome GS12880 (Drmanac et al., 2009) with a few known disease variants added. This will provide a realistic background to search for the disease SNPs, but not necessarily a realistic collection of disease SNPs for a single individual to have. We chose an assortment of six SNPs from the PhenCode database (Giardine et al., 2007), representing different genes and different parts of the gene (Table 15.2.1). There are two coding SNPs (heterozygous) and four non-coding (one heterozygous and three homozygous). The four non-coding SNPs are located in a promoter region, a UTR, and two introns.

Current Protocols in Bioinformatics 15.2.1-15.2.27, September 2012
Published online September 2012 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/0471250953.bi1502s39

Table 15.2.1 Disease SNPs Planted in the Sample Dataset[a]

chr11	5248153	5248154	G	Intron	HbVar:thalassemia
chr11	5255743	5255744	C	UTR	HbVar:thalassemia
chr11	5275879	5275880	C/T	Coding	HbVar:Hb E
chr16	222915	222916	T/C	Coding	HbVar:Hb Lyon-Bron
chrX	31279779	31279780	T/C	Intron	LMDp:muscular dystrophy
chrX	100641248	100641249	G	Promoter	BTKbase:Agammaglobulinemia

[a] The position of the SNP is given here using the same coordinate conventions as in BED format, to match what is seen in the results. The last column shows the database the SNP is from and the associated disease.

This example builds a single sequential history, but it is organized into several protocols according to the type of analysis being performed. These illustrate the following skills.

Basic Protocol 1: Preparing input data

- Uploading files
- Using Galaxy libraries
- Basic filtering

Basic Protocol 2: Selecting known coding SNPs predicted to be damaging, then finding their genes and associated pathways

- PolyPhen-2
- Gene-based analysis

Basic Protocol 3: Running new predictions of coding SNPs likely to be detrimental

- SIFT
- Using published workflows

Basic Protocol 4: Finding SNPs that fall in suspected functional regions

- Predicted regulatory regions
- ENCODE functional data
- PhyloP conserved positions

Conventions Used in Figures (Screen Shots) for Tutorial Purposes

Red arrows or boxes on the screen shots indicate settings or things you will need to do. Green arrows indicate the "go" buttons once the settings or parameters are selected. Blue arrows or boxes point out additional information that you should note, but that do not require any action.

BASIC PROTOCOL 1

USING GALAXY TO LOOK FOR DISEASE SNPs IN A FULL GENOME: PREPARING INPUT DATA

Overview

- Start with a full set of SNP calls from a particular individual, in the masterVar format used by Complete Genomics.
- Upload the file to a Galaxy history via URL or FTP.

- Convert to pgSnp format.
- Obtain a set of SNPs found in healthy individuals from the Galaxy library.
- Subtract the SNPs found in healthy individuals to narrow the search.

Necessary Resources

Hardware

An Internet-connected computer

Software

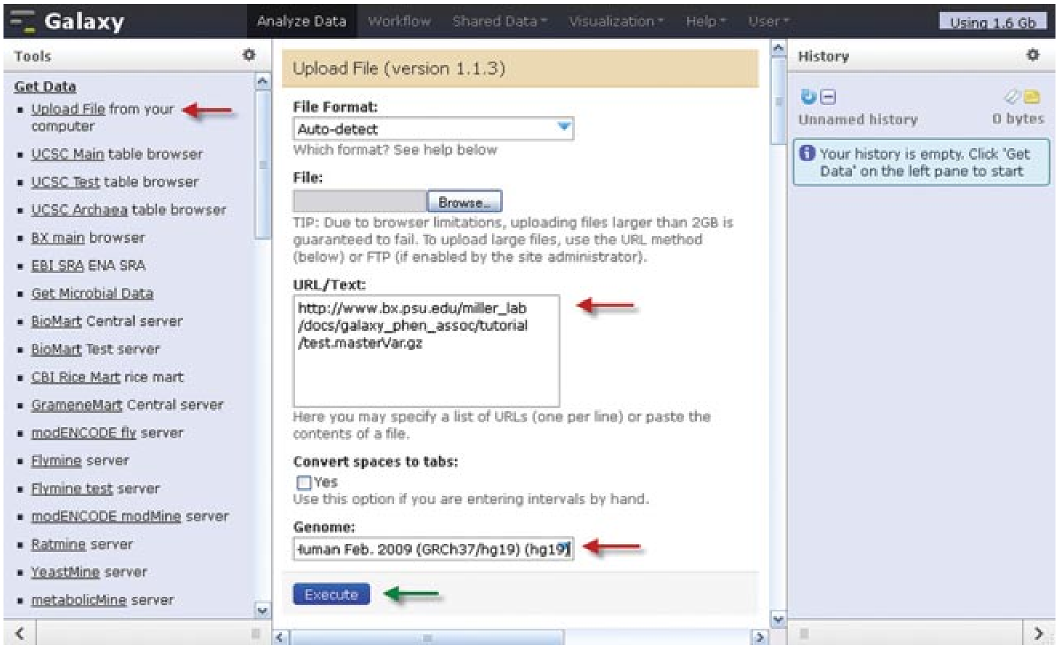
Internet browser that supports JavaScript (e.g., most current browsers such as Mozilla Firefox, Safari, Opera, Chrome, or Microsoft Internet Explorer)

Files

Supplementary file: `test.masterVar.gz` (found at *http://www.bx.psu.edu/miller_lab/docs/galaxy_phen_assoc/tutorial/test.masterVar.gz*). See Supplementary File at the end of this unit for an alternate URL to access the supplementary file.

Uploading a data file

1. There are two ways to upload large files to Galaxy: one is to use FTP, and the other is to place the file in a Web-accessible location and give Galaxy the URL. For this example, we will use the URL method. Open your Web browser, go to the Galaxy portal at *galaxyproject.org*, and choose "Use the free public server." In the tool panel on the left, click on the section called Get Data to open it, then click on the Upload File tool (indicated by a red arrow in Fig. 15.2.1). The input form for the selected tool will appear in the center panel, along with other information about the tool.



Upload File (version 1.1.3)

File Format: Auto-detect

Which format? See help below

File:

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

http://www.bx.psu.edu/miller_lab
/docs/galaxy_phen_assoc/tutorial
/test.masterVar.gz

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Convert spaces to tabs:

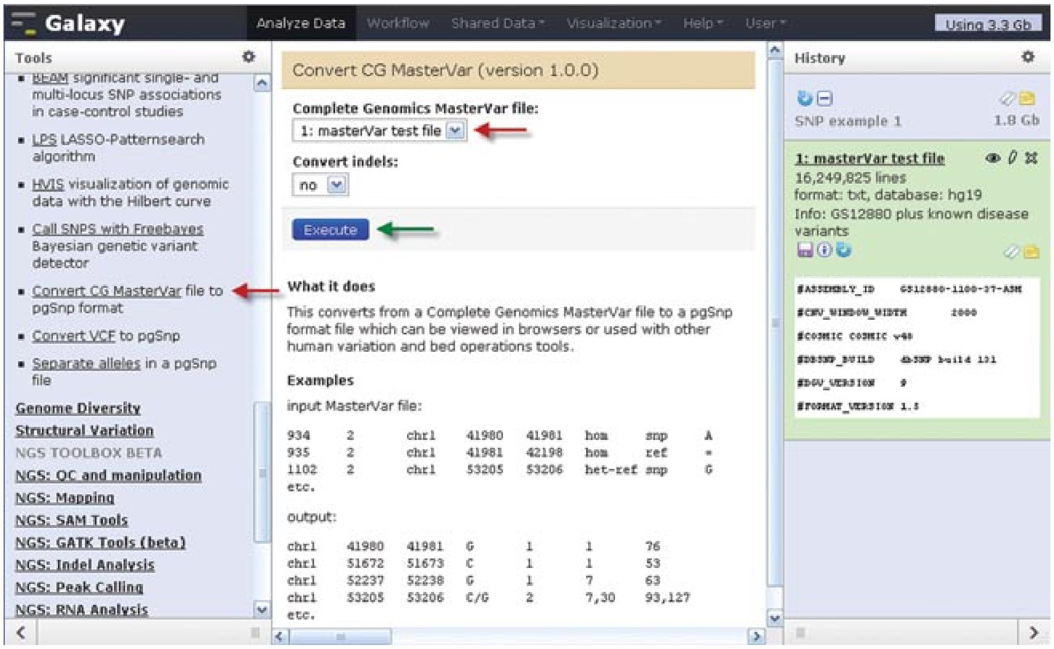
□ Yes

Use this option if you are entering intervals by hand.

Genome: Human Feb. 2009 (GRCh37/hg19) (hg19)

Your history is empty. Click 'Get Data' on the left pane to start

Figure 15.2.1 Uploading a data file. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.



- BEAM significant single- and multi-locus SNP associations in case-control studies
- LPS LASSO-Patternsearch algorithm
- HVIS visualization of genomic data with the Hilbert curve
- Call SNPS with Freebayes Bayesian genetic variant detector
- Convert CG MasterVar file to pgSnp format
- Convert VCF to pgSnp
- Separate alleles in a pgSnp file

Convert CG MasterVar (version 1.0.0)

Complete Genomics MasterVar file: 1: masterVar test file

Convert indels: no

Execute

What it does

This converts from a Complete Genomics MasterVar file to a pgSnp format file which can be viewed in browsers or used with other human variation and bed operations tools.

Examples

input MasterVar file:

```
934     2       chr1    41980   41981   hom     snp     A
935     2       chr1    41981   42198   hom     ref     =
1102    2       chr1    53205   53206   het-ref snp     G
etc.
```

output:

```
chr1    41980   41981   G       1       1       76
chr1    51672   51673   C       1       1       53
chr1    52237   52238   G       1       7       63
chr1    53205   53206   C/G     2       7,30    93,127
etc.
```

1: masterVar test file
16,249,825 lines
format: txt, database: hg19
Info: GS12880 plus known disease variants

```
#ASSEMBLY_ID    GS12880-1100-37-ASM
#CNV_WINDOW_WIDTH       2000
#COSMIC COSMIC v48
#DBSNP_BUILD    dbSNP build 131
#DGV_VERSION    9
#FORMAT_VERSION 1.5
```

Figure 15.2.2 Converting to pgSnp format. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

2. Type or paste the file's URL in the large text box: *http://www.bx.psu.edu/miller_lab/docs/galaxy_phen_assoc/tutorial/test.masterVar.gz*.

3. Also be sure to choose the corresponding genome assembly in the Genome box. Typing a search term in the select box (e.g., `hg19`) will limit the list of choices to items containing the specified text. Then, click Execute (green arrow) to upload the dataset into your history.

   Compressed files are unpacked automatically.

Converting to pgSnp format

4. Our uploaded dataset is in the masterVar file format used by Complete Genomics, but that doesn't work with very many of the Galaxy tools. Therefore we will convert it to pgSnp, which is a specialized BED format for SNPs. In the tool panel, open the Phenotype Association section and click on the tool called Convert CG MasterVar (red arrow in Fig. 15.2.2). Then, in the center panel, select the dataset with the masterVar file (red arrow), and click the Execute button (green arrow).

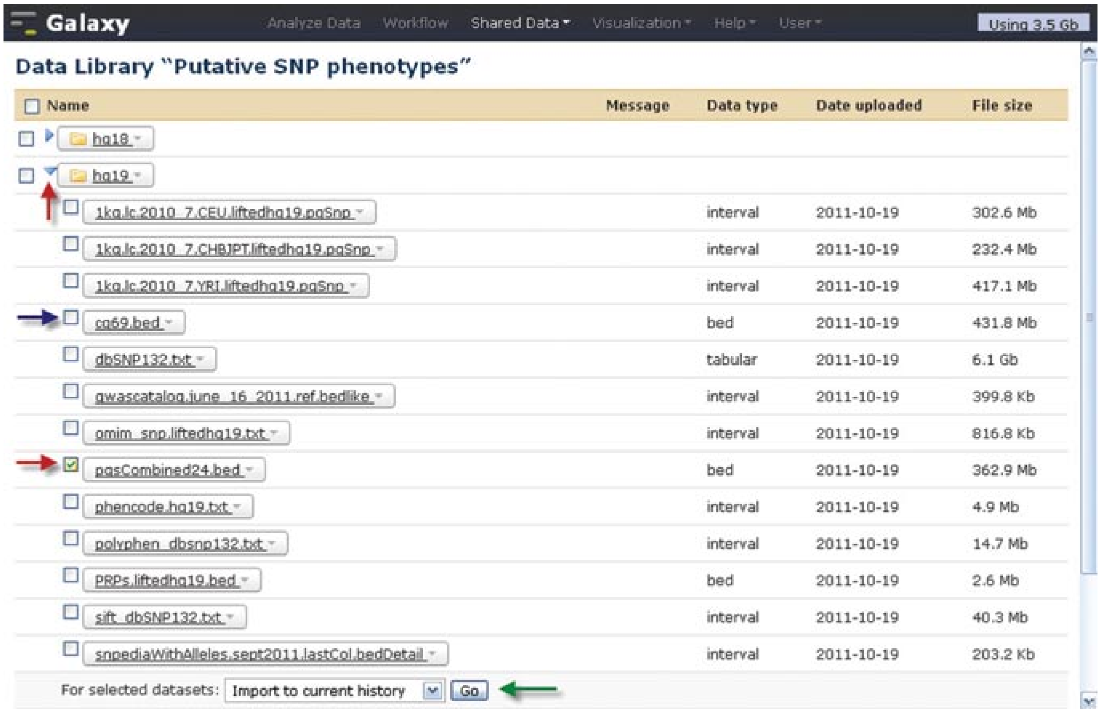
   It is good practice to rename the major datasets in your history to remind yourself later what each one represents. Here we have renamed the uploaded dataset, and also the history itself (see UNIT 10.5).

Galaxy libraries

5. To look for dominantly inherited disease SNPs, it is helpful to remove SNPs found in healthy individuals. To do this we will pull in some data from a Galaxy library. Click on Shared Data in the menu bar and select Data Libraries. Then, look for a library called Putative SNP Phenotypes and click on it.

Putative SNP Phenotypes library

6. First open the folder with the hg19 datasets by clicking on the blue triangle next to the folder name (red arrow pointing up in Fig. 15.2.3). To import datasets into your history, check the ones you want (red arrow) and then click the Go button (green arrow).



Data Library "Putative SNP phenotypes"

□ Name	Message	Data type	Date uploaded	File size
□ hg18				
□ hg19				
□ 1kg.lc.2010_7.CEU.liftedhg19.pgSnp		interval	2011-10-19	302.6 Mb
□ 1kg.lc.2010_7.CHBJPT.liftedhg19.pgSnp		interval	2011-10-19	232.4 Mb
□ 1kg.lc.2010_7.YRI.liftedhg19.pgSnp		interval	2011-10-19	417.1 Mb
□ cg69.bed		bed	2011-10-19	431.8 Mb
□ dbSNP132.txt		tabular	2011-10-19	6.1 Gb
□ gwascatalog.june_16_2011.ref.bedlike		interval	2011-10-19	399.8 Kb
□ omim_snp.liftedhg19.txt		interval	2011-10-19	816.8 Kb
☑ pgsCombined24.bed		bed	2011-10-19	362.9 Mb
□ phencode.hg19.txt		interval	2011-10-19	4.9 Mb
□ polyphen_dbsnp132.txt		interval	2011-10-19	14.7 Mb
□ PRPs.liftedhg19.bed		bed	2011-10-19	2.6 Mb
□ sift_dbSNP132.txt		interval	2011-10-19	40.3 Mb
□ snpediaWithAlleles.sept2011.lastCol.bedDetail		interval	2011-10-19	203.2 Kb

For selected datasets: Import to current history [Go]

Figure 15.2.3 Putative SNP Phenotypes library. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

In this library, there are currently two sets of full-coverage SNPs from healthy individuals. One has 24 public genomes from a variety of sources, populations, and sequencing technologies (red arrow). The other is a group of 69 from Complete Genomics (blue arrow). In most cases, you will want to use both of these for filtering; however since our example input dataset is based on one of the CG genomes, we will skip that set here.

Removing SNPs found in healthy individuals

7. Clicking on the Analyze Data link in the menu bar returns us to the page with the history and tool panels.

8. Now we are ready to remove the "benign" SNPs from our input. In the tool panel, open the section called Operate on Genomic Intervals. Click on Subtract (red arrow in Fig. 15.2.4) to see the options for that tool in the center panel. We want to remove the SNPs that were found in the 24 genomes from our converted set (red arrows). Next, click Execute (green arrow).

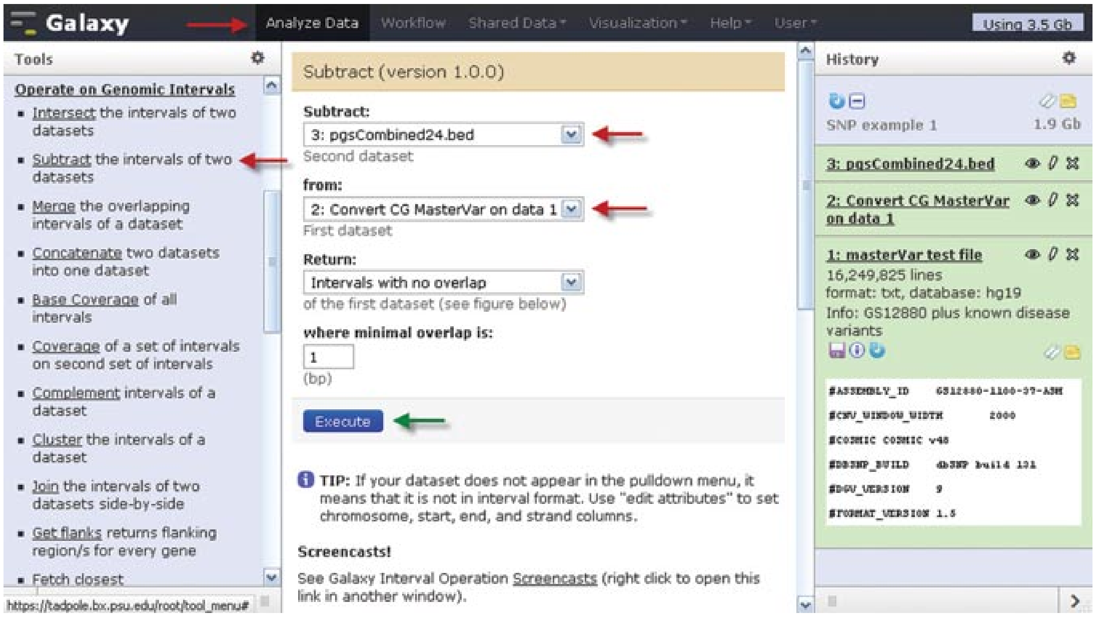
   The tools in the Operate on Genomic Intervals section are examples of tools that cannot be run on a masterVar file; they only work with "interval" formats, which include pgSnp.

   If we had not used a Complete Genomics dataset as our starting point, the next step would be to repeat the subtraction using the result from the first subtraction and the CG dataset from the library (blue arrow in Fig. 15.2.3).

Completed input dataset

9. Open the results in the history panel by clicking on the dataset name. We see that the number of SNPs is greatly reduced, from around 3.4 million to 96 thousand (blue arrows in Fig. 15.2.5).

   We have renamed the results of this step, moving the automatically generated name to the info field (red box). This dataset will be the starting point in the subsequent protocols for this example. The filtering we have done is relevant when looking for both coding and non-coding SNPs.



Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 3.5 Gb

Tools

Operate on Genomic Intervals

- Intersect the intervals of two datasets
- Subtract the intervals of two datasets
- Merge the overlapping intervals of a dataset
- Concatenate two datasets into one dataset
- Base Coverage of all intervals
- Coverage of a set of intervals on second set of intervals
- Complement intervals of a dataset
- Cluster the intervals of a dataset
- Join the intervals of two datasets side-by-side
- Get flanks returns flanking region/s for every gene
- Fetch closest

https://tadpole.bx.psu.edu/root/tool_menu#

Subtract (version 1.0.0)

Subtract: 3: pgsCombined24.bed
Second dataset

from: 2: Convert CG MasterVar on data 1
First dataset

Return: Intervals with no overlap
of the first dataset (see figure below)

where minimal overlap is: 1
(bp)

Execute

TIP: If your dataset does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.

Screencasts!

See Galaxy Interval Operation Screencasts (right click to open this link in another window).

History

SNP example 1 1.9 Gb

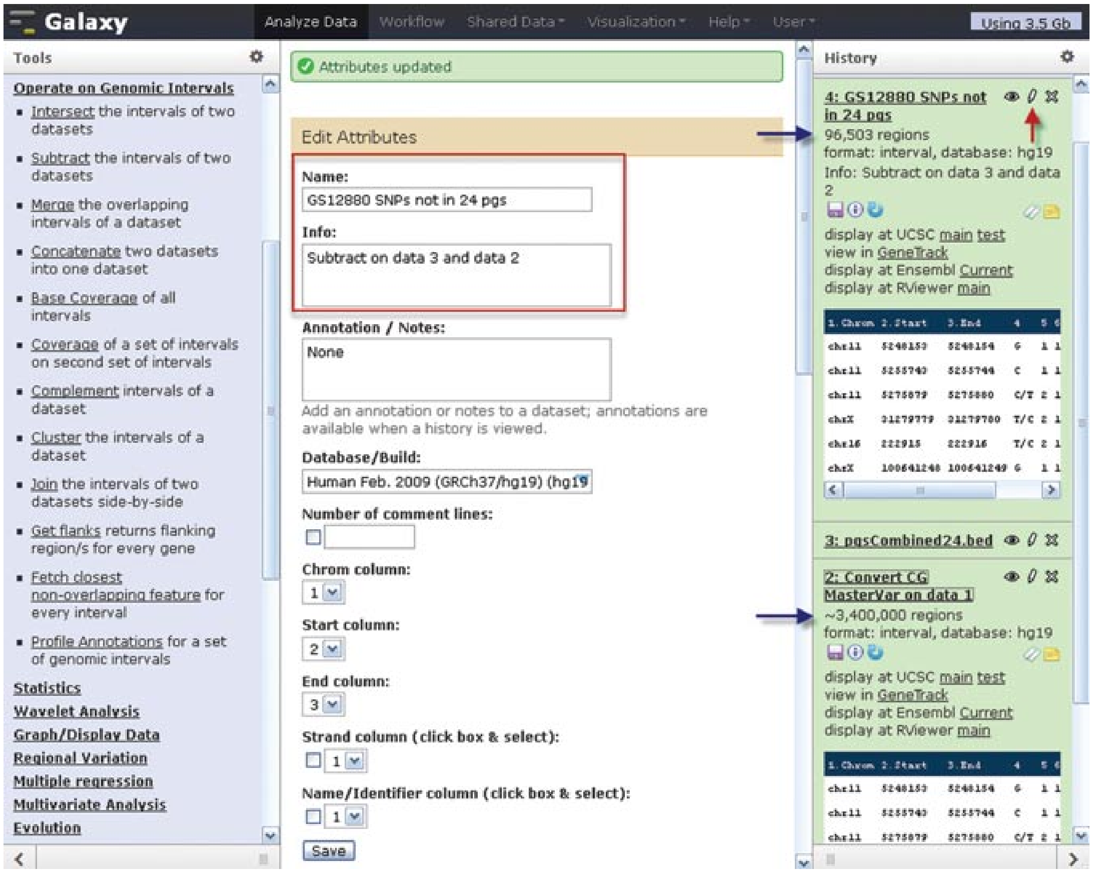
3: pgsCombined24.bed

2: Convert CG MasterVar on data 1

1: masterVar test file
16,249,825 lines
format: txt, database: hg19
Info: GS12880 plus known disease variants

```
#ASSEMBLY_ID     GS12880-1100-37-ASM
#CNV_WINDOW_WIDTH        2000
#COSMIC COSMIC v48
#DBSNP_BUILD     dbSNP build 131
#DGV_VERSION     9
#FORMAT_VERSION 1.5
```

Figure 15.2.4 Removing SNPs found in healthy individuals. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.



Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 3.5 Gb

Tools

Operate on Genomic Intervals

- Intersect the intervals of two datasets
- Subtract the intervals of two datasets
- Merge the overlapping intervals of a dataset
- Concatenate two datasets into one dataset
- Base Coverage of all intervals
- Coverage of a set of intervals on second set of intervals
- Complement intervals of a dataset
- Cluster the intervals of a dataset
- Join the intervals of two datasets side-by-side
- Get flanks returns flanking region/s for every gene
- Fetch closest non-overlapping feature for every interval
- Profile Annotations for a set of genomic intervals

Statistics
Wavelet Analysis
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution

Attributes updated

Edit Attributes

Name: GS12880 SNPs not in 24 pgs

Info: Subtract on data 3 and data 2

Annotation / Notes: None

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build: Human Feb. 2009 (GRCh37/hg19) (hg19

Number of comment lines: □

Chrom column: 1

Start column: 2

End column: 3

Strand column (click box & select): □ 1

Name/Identifier column (click box & select): □ 1

Save

History

4: GS12880 SNPs not in 24 pgs
96,503 regions
format: interval, database: hg19
Info: Subtract on data 3 and data 2
display at UCSC main test
view in GeneTrack
display at Ensembl Current
display at RViewer main

```
1.Chrom 2.Start   3.End     4   5 6
chr11   5248153   5248154   G   1 1
chr11   5255743   5255744   C   1 1
chr11   5275879   5275880   C/T 2 1
chrX    31279779  31279780  T/C 2 1
chr16   222915    222916    T/C 2 1
chrX    100641248 100641249 G   1 1
```

3: pgsCombined24.bed

2: Convert CG MasterVar on data 1
~3,400,000 regions
format: interval, database: hg19
display at UCSC main test
view in GeneTrack
display at Ensembl Current
display at RViewer main

```
1.Chrom 2.Start   3.End     4   5 6
chr11   5248153   5248154   G   1 1
chr11   5255743   5255744   C   1 1
chr11   5275879   5275880   C/T 2 1
```

Figure 15.2.5 Completed input dataset. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

SELECTING KNOWN CODING SNPs PREDICTED TO BE DAMAGING, THEN FINDING THEIR GENES AND ASSOCIATED PATHWAYS

BASIC PROTOCOL 2

Overview

• Import a public library file containing pre-computed results from running the PolyPhen-2 program (Adzhubei et al., 2010) on the dbSNP database.

• Join the input dataset with the PolyPhen-2 results row-by-row, based on interval overlap. This adds new columns to the input set, including the UniProt protein accession ID and the predicted effect of the SNP.

• Filter the results to select rows containing the word “damaging.”

• Translate the UniProt IDs to HUGO gene symbols by joining with an identifier table imported from UCSC.

• Run the CTD tool to extract curated pathway associations for these genes from the Comparative Toxicogenomics Database.

Necessary Resources

Hardware

An Internet-connected computer

Software

Internet browser that supports JavaScript (e.g., most current browsers such as Mozilla Firefox, Safari, Opera, Chrome, or Microsoft Internet Explorer)

PolyPhen-2 predictions

1. Go to Shared Data > Data Libraries > Putative SNP Phenotypes > hg19 as before. This time look for the dataset `polyphen_dbsnp132.txt`. To learn more about this dataset, click on its name instead of marking the box.

Details about the PolyPhen-2 dataset

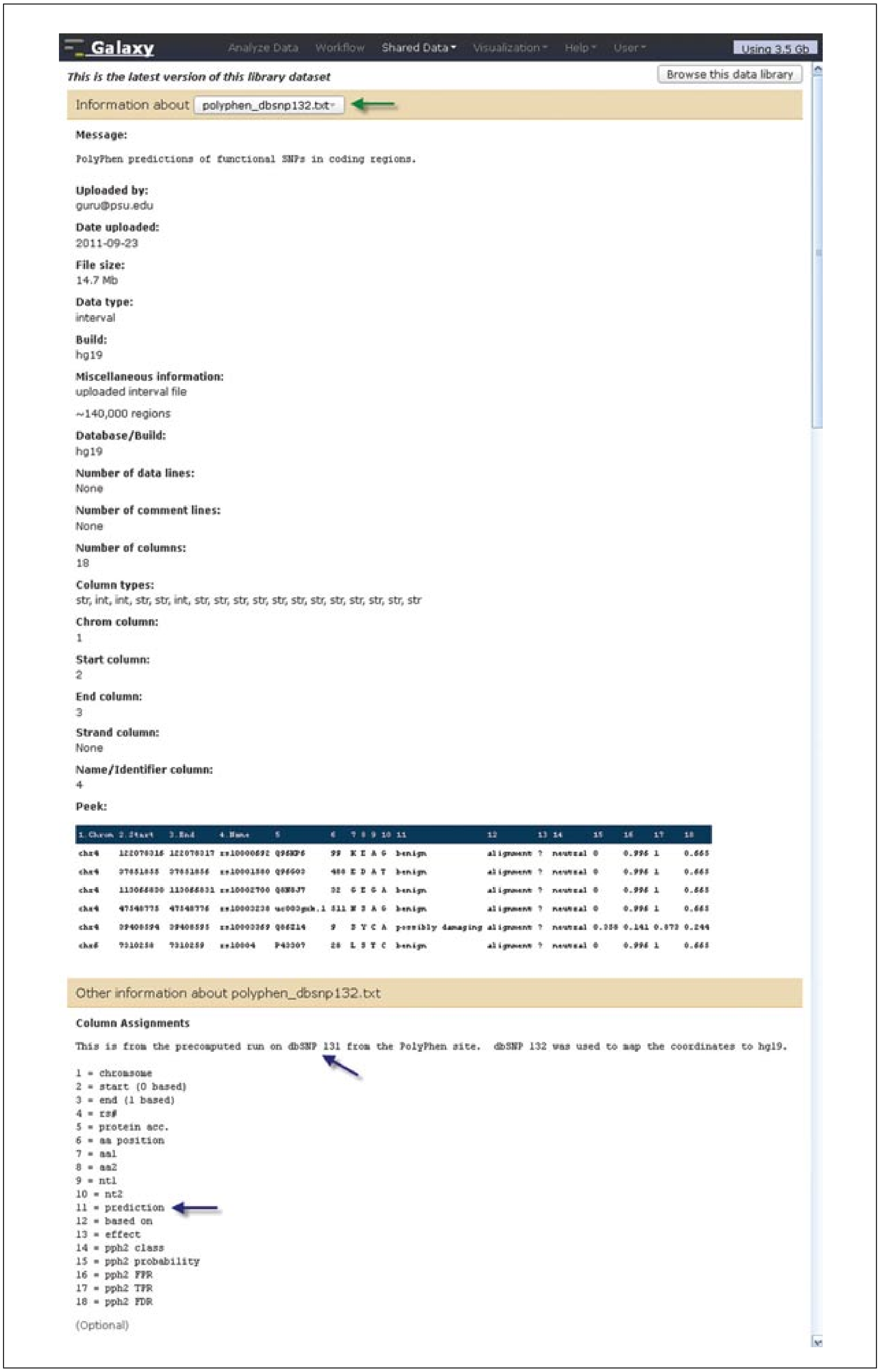
2. Most of the first section is information automatically computed about the dataset, including its size, date uploaded, and a peek at the actual data (Fig. 15.2.6). The Message at the top, as well as the second section, Other Information, is provided by the person who added the file to the library. Here we find labels for the data columns, and learn that dbSNP version 131 was used as input to compute the predictions, which are in column 11.

3. To add this dataset to your history from here, click on its name (green arrow) and select “Import this dataset into selected histories.” On the next page that appears, your current history should already be selected as the default destination, so just click the “Import library datasets” button. Then, click Analyze Data in the menu bar to return to the page with the history and tool panels.

Joining on genomic intervals

4. We will do a join between the two datasets to get the predictions associated with our SNPs. We do not have a shared identifier to join on, so instead we will join together rows of the two datasets whenever their positions on the genome overlap. The result will have all of the information from both datasets, for any position on the genome where there are data in both datasets.

   In the Operate on Genomic Intervals section of the tool panel, click on Join (Fig. 15.2.7). Then, in the center panel, select the filtered SNPs from Basic Protocol 1 as the first dataset and the PolyPhen-2 predictions as the second. We



This is the latest version of this library dataset

Information about polyphen_dbsnp132.txt

Message:

PolyPhen predictions of functional SNPs in coding regions.

Uploaded by:
guru@psu.edu

Date uploaded:
2011-09-23

File size:
14.7 Mb

Data type:
interval

Build:
hg19

Miscellaneous information:
uploaded interval file

~140,000 regions

Database/Build:
hg19

Number of data lines:
None

Number of comment lines:
None

Number of columns:
18

Column types:
str, int, int, str, str, int, str, str, str, str, str, str, str, str, str, str, str, str

Chrom column:
1

Start column:
2

End column:
3

Strand column:
None

Name/Identifier column:
4

Peek:

1.Chrom	2.Start	3.End	4.Name	5	6	7	8	9	10	11	12	13	14	15	16	17	18
chr4	122078316	122078317	rs10000692	Q96RP6	99	K	E	A	G	benign	alignment	?	neutral	0	0.996	1	0.665
chr4	37851855	37851856	rs10001580	Q96G03	488	E	D	A	T	benign	alignment	?	neutral	0	0.996	1	0.665
chr4	113068830	113068831	rs10002700	Q8NBJ7	32	G	E	G	A	benign	alignment	?	neutral	0	0.996	1	0.665
chr4	47540775	47540776	rs10003238	uc003gxh.1	311	N	S	A	G	benign	alignment	?	neutral	0	0.996	1	0.665
chr4	39408594	39408595	rs10003369	Q86Z14	9	S	Y	C	A	possibly damaging	alignment	?	neutral	0.358	0.141	0.872	0.244
chr8	7310258	7310259	rs10004	P43307	26	L	S	T	C	benign	alignment	?	neutral	0	0.996	1	0.665

Other information about polyphen_dbsnp132.txt

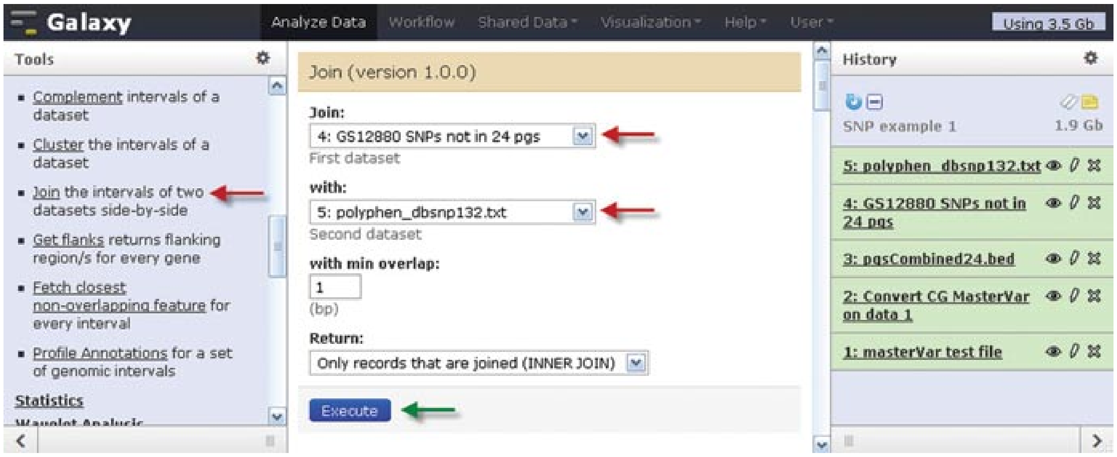
Column Assignments

This is from the precomputed run on dbSNP 131 from the PolyPhen site. dbSNP 132 was used to map the coordinates to hg19.

```
1 = chromsome
2 = start (0 based)
3 = end (1 based)
4 = rs#
5 = protein acc.
6 = aa position
7 = aa1
8 = aa2
9 = nt1
10 = nt2
11 = prediction
12 = based on
13 = effect
14 = pph2 class
15 = pph2 probability
16 = pph2 FPR
17 = pph2 TPR
18 = pph2 FDR
```

(Optional)

Figure 15.2.6 Details about the PolyPhen-2 dataset. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 3.5 Gb
Tools
Complement intervals of a dataset
Cluster the intervals of a dataset
Join the intervals of two datasets side-by-side
Get flanks returns flanking region/s for every gene
Fetch closest non-overlapping feature for every interval
Profile Annotations for a set of genomic intervals
Statistics
Wavelet Analysis
History
SNP example 1
1.9 Gb
5: polyphen_dbsnp132.txt
4: GS12880 SNPs not in 24 pgs
3: pgsCombined24.bed
2: Convert CG MasterVar on data 1
1: masterVar test file
Execute

Join (version 1.0.0)

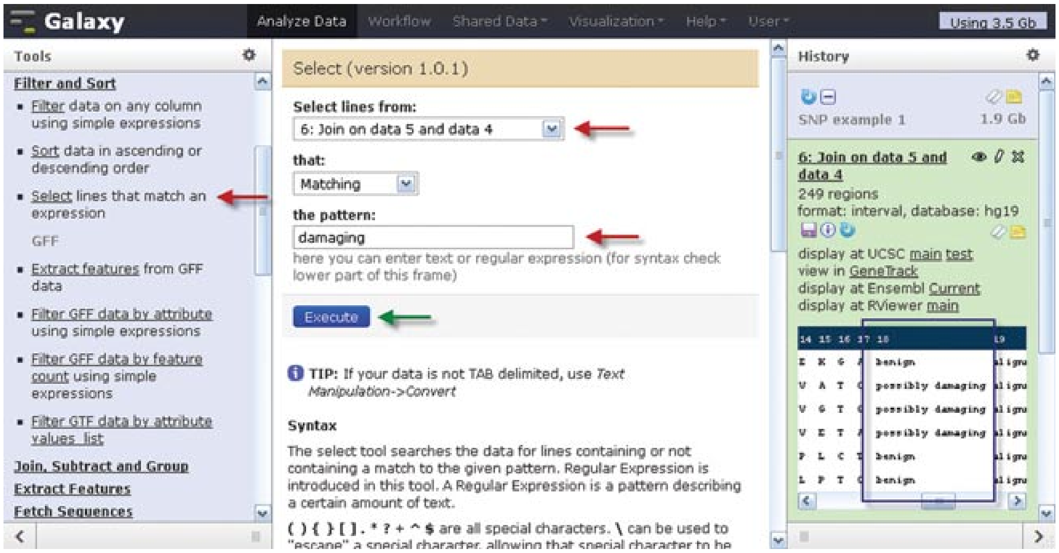
Join: 4: GS12880 SNPs not in 24 pgs
First dataset

with: 5: polyphen_dbsnp132.txt
Second dataset

with min overlap: 1
(bp)

Return: Only records that are joined (INNER JOIN)

Figure 15.2.7 Joining on genomic intervals. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 3.5 Gb
Tools
Filter and Sort
Filter data on any column using simple expressions
Sort data in ascending or descending order
Select lines that match an expression
GFF
Extract features from GFF data
Filter GFF data by attribute using simple expressions
Filter GFF data by feature count using simple expressions
Filter GTF data by attribute values_list
Join, Subtract and Group
Extract Features
Fetch Sequences
History
SNP example 1
1.9 Gb
6: Join on data 5 and data 4
display at UCSC main test
view in GeneTrack
display at Ensembl Current
display at RViewer main
Execute

Select (version 1.0.1)

Select lines from: 6: Join on data 5 and data 4

that: Matching

the pattern: damaging
here you can enter text or regular expression (for syntax check lower part of this frame)

TIP: If your data is not TAB delimited, use *Text Manipulation->Convert*

Syntax

The select tool searches the data for lines containing or not containing a match to the given pattern. Regular Expression is introduced in this tool. A Regular Expression is a pattern describing a certain amount of text.

( ) { } [ ] . * ? + ^ $ are all special characters. \ can be used to "escape" a special character, allowing that special character to be

249 regions
format: interval, database: hg19

```
14 15 16 17 18                  19
E  K  G  A  benign              aligu
V  A  T  C  possibly damaging   aligu
V  G  T  C  possibly damaging   aligu
V  E  T  A  possibly damaging   aligu
P  L  C  T  benign              aligu
L  P  T  C  benign              aligu
```

Figure 15.2.8 Selecting damaging results. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

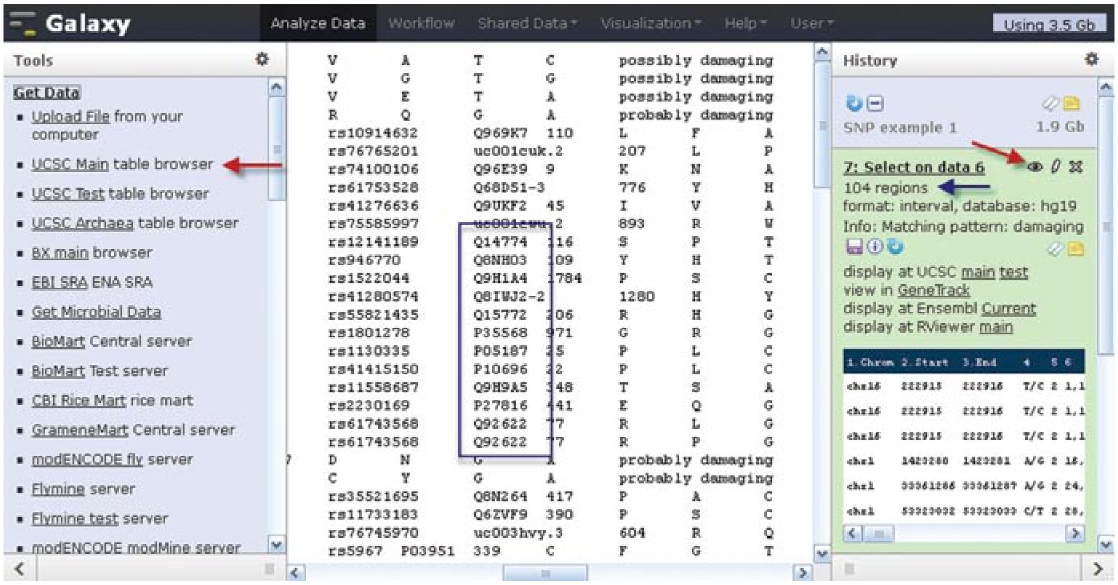
are only interested in SNPs that appear in both sets, so leave the default setting of doing an Inner Join. Then click Execute.

Selecting damaging results

5. There are 249 predictions (Fig. 15.2.8), but some of those say that the SNP is benign. In the history panel, a few rows of data are shown; use the scrollbar to scroll over to the predictions (blue box). The ones we are interested in use the word "damaging."

6. Open the Filter and Sort section and choose the Select tool, which selects lines from a dataset that match (or do not match) the given pattern. To run it, choose the dataset with the join results and type in `damaging` for the pattern. The search is case-sensitive, so type it exactly as it appears in the dataset. Then click Execute.

   The predictions are no longer in column 11 because we requested the columns from the SNP dataset to be listed first in the join results.



```
V         A         T        C         possibly damaging
V         G         T        G         possibly damaging
V         E         T        A         possibly damaging
R         Q         G        A         probably damaging
rs10914632          Q969K7   110       L         F        A
rs76765201          uc001cuk.2         207       L        P
rs74100106          Q96E39   9         K         N        A
rs61753528          Q68D51-3           776       Y        H
rs41276636          Q9UKF2   45        I         V        A
rs75585997          uc001cwu.2         893       R        W
rs12141189          Q14774   116       S         P        T
rs946770            Q8NH03   109       Y         H        T
rs1522044           Q9H1A4   1784      P         S        C
rs41280574          Q8IWJ2-2           1280      H        Y
rs55821435          Q15772   206       R         H        G
rs1801278           P35568   971       G         R        G
rs1130335           P05187   25        P         L        C
rs41415150          P10696   22        P         L        C
rs11558687          Q9H9A5   348       T         S        A
rs2230169           P27816   441       E         Q        G
rs61743568          Q92622   77        R         L        G
rs61743568          Q92622   77        R         P        G
D         N         G        A         probably damaging
C         Y         G        A         probably damaging
rs35521695          Q8N264   417       P         A        C
rs11733183          Q6ZVF9   390       P         S        C
rs76745970          uc003hvy.3         604       R        Q
rs5967    P03951    339      C         F         G        T
```

7: Select on data 6
104 regions
format: interval, database: hg19
Info: Matching pattern: damaging
display at UCSC main test
view in GeneTrack
display at Ensembl Current
display at RViewer main

1.Chrom	2.Start	3.End	4	5	6
chr16	222915	222916	T/C	2	1,1
chr16	222915	222916	T/C	2	1,1
chr16	222915	222916	T/C	2	1,1
chr1	1423280	1423281	A/G	2	16,
chr1	33361286	33361287	A/G	2	24,
chr1	53323032	53323033	C/T	2	28,

Figure 15.2.9 PolyPhen-2 results. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

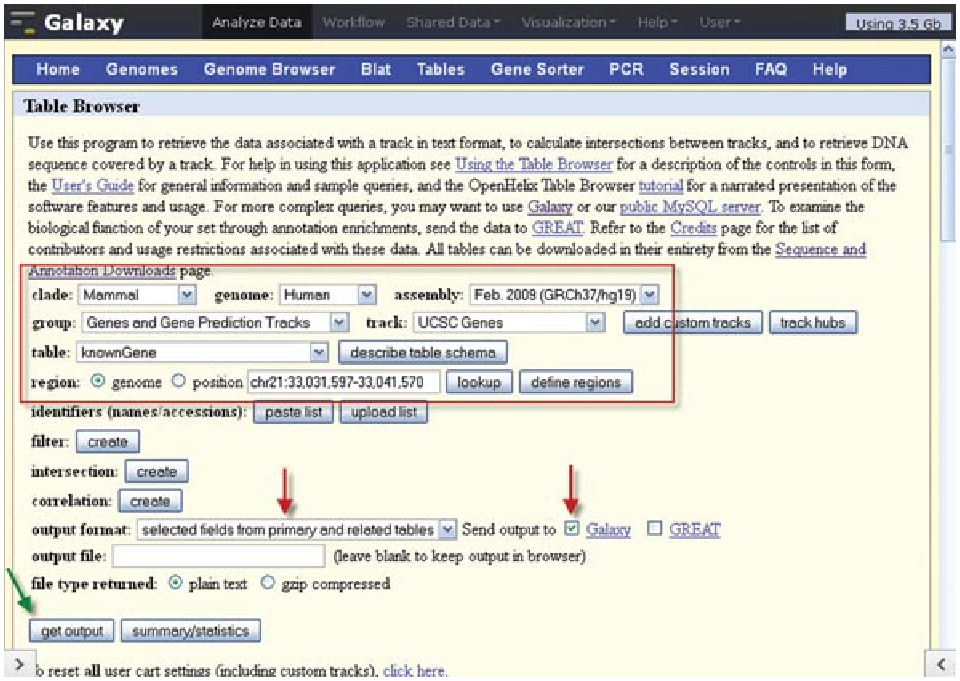


Table Browser

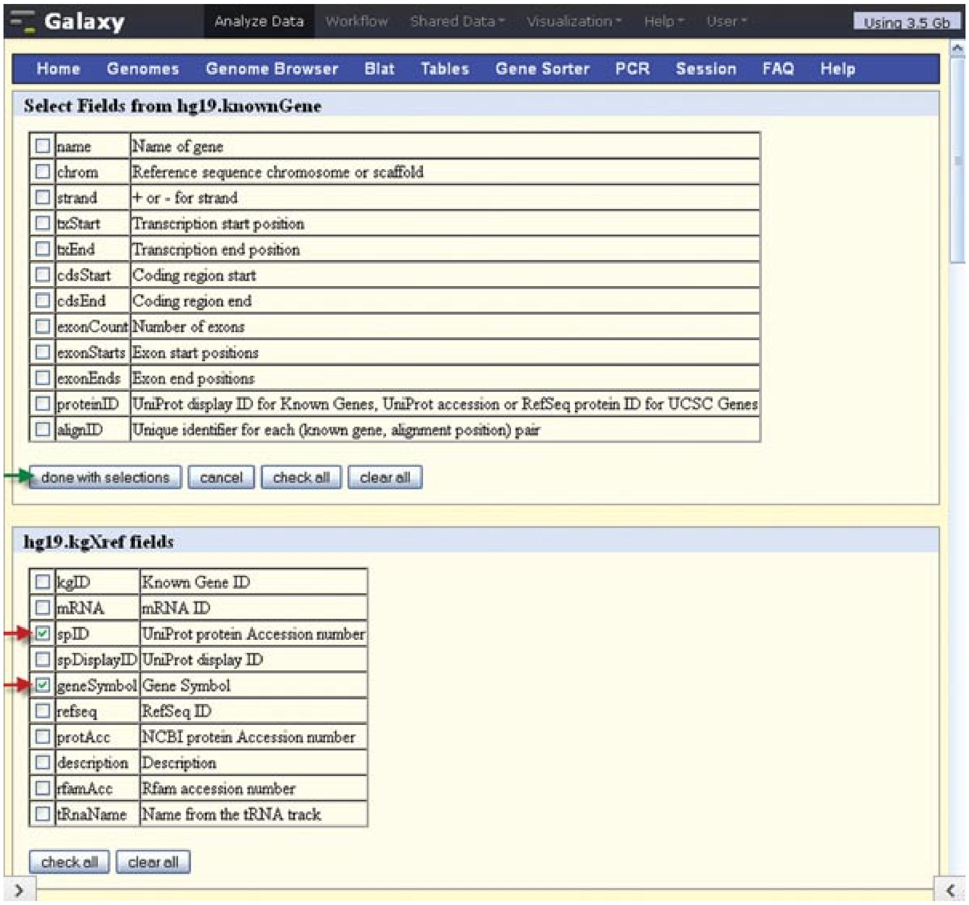
Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal genome: Human assembly: Feb. 2009 (GRCh37/hg19)
group: Genes and Gene Prediction Tracks track: UCSC Genes add custom tracks track hubs
table: knownGene describe table schema
region: ● genome ○ position chr21:33,031,597-33,041,570 lookup define regions
identifiers (names/accessions): paste list upload list
filter: create
intersection: create
correlation: create
output format: selected fields from primary and related tables Send output to ☑ Galaxy □ GREAT
output file: (leave blank to keep output in browser)
file type returned: ● plain text ○ gzip compressed
get output summary/statistics
o reset all user cart settings (including custom tracks), click here.

Figure 15.2.10 Mapping between identifiers. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

PolyPhen-2 results

7. We find 104 SNPs that are predicted to be damaging by PolyPhen-2 (blue arrow in Fig. 15.2.9), including one of our two coding disease SNPs (not shown). For small datasets like this, you can display the entire contents by clicking on the eye icon in the history panel.



Select Fields from hg19.knownGene

□	name	Name of gene
□	chrom	Reference sequence chromosome or scaffold
□	strand	+ or - for strand
□	txStart	Transcription start position
□	txEnd	Transcription end position
□	cdsStart	Coding region start
□	cdsEnd	Coding region end
□	exonCount	Number of exons
□	exonStarts	Exon start positions
□	exonEnds	Exon end positions
□	proteinID	UniProt display ID for Known Genes, UniProt accession or RefSeq protein ID for UCSC Genes
□	alignID	Unique identifier for each (known gene, alignment position) pair

hg19.kgXref fields

□	kgID	Known Gene ID
□	mRNA	mRNA ID
☑	spID	UniProt protein Accession number
□	spDisplayID	UniProt display ID
☑	geneSymbol	Gene Symbol
□	refseq	RefSeq ID
□	protAcc	NCBI protein Accession number
□	description	Description
□	rfamAcc	Rfam accession number
□	tRnaName	Name from the tRNA track

Figure 15.2.11 Choosing the identifier fields. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

In the blue box you can see that we now have UniProt IDs for many of the genes associated with these variants. We can look for pathways associated with the genes by using the CTD tool; however, that tool requires HUGO/NCBI identifiers (Seal et al., 2011) rather than UniProt IDs for its input, so first we will download a file from the UCSC Table Browser (Karolchik et al., 2004) to map between the identifiers (red arrow in the tool panel).

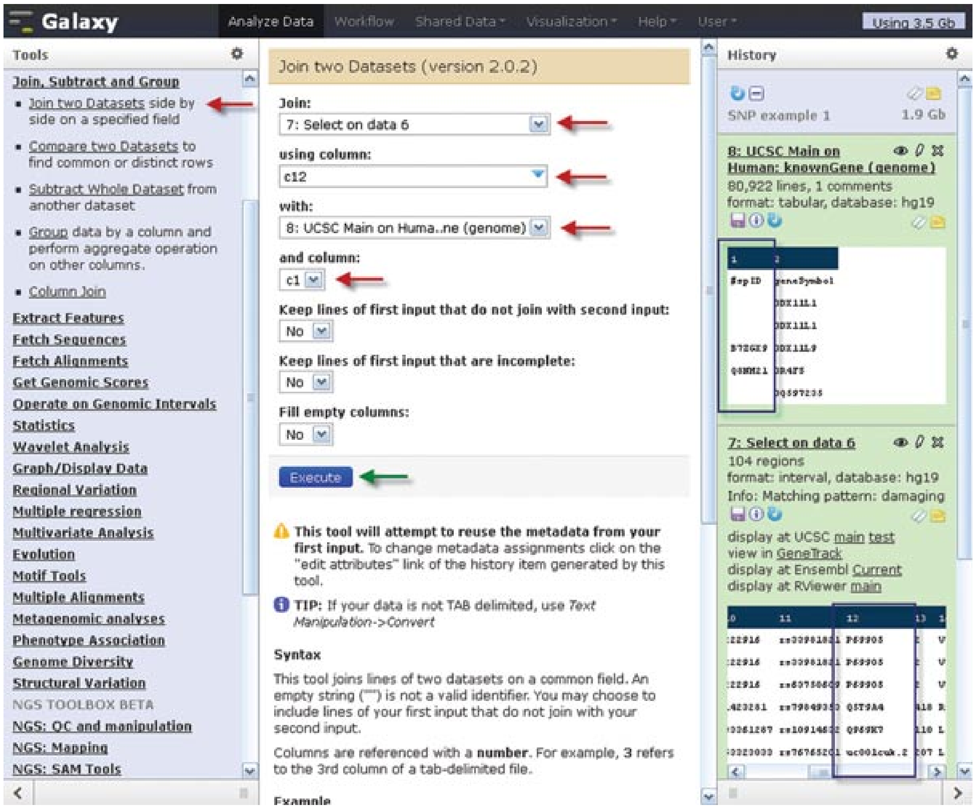
The columns do not line up in this view because the data are tab-separated and the values in each column are not all of the same length.

Mapping between identifiers

8. First make sure the genome and assembly are correct; we want Human build hg19 in order to match our history datasets. We will be using the UCSC Known Genes table because it has the most additional information, including connections between various identifiers, and we want them for the full genome. All of these options are shown in the red box (Fig. 15.2.10). We also want the output format to be "selected fields from primary and related tables," and to have the results sent back to Galaxy (red arrows). Then click the "get output" button (green arrow).

The box for chromosome coordinates is irrelevant when the radio button for the whole genome is selected.

Sometimes having the window divided into three panels makes the center one too narrow to work in easily. In these cases you can click on the large "<" and ">" arrow buttons in the lower corners to hide the tool and/or history panels. When you are done with the



Join two Datasets (version 2.0.2)

Join:
7: Select on data 6

using column:
c12

with:
8: UCSC Main on Huma..ne (genome)

and column:
c1

Keep lines of first input that do not join with second input:
No

Keep lines of first input that are incomplete:
No

Fill empty columns:
No

Execute

This tool will attempt to reuse the metadata from your first input. To change metadata assignments click on the "edit attributes" link of the history item generated by this tool.

TIP: If your data is not TAB delimited, use *Text Manipulation->Convert*

Syntax

This tool joins lines of two datasets on a common field. An empty string ("") is not a valid identifier. You may choose to include lines of your first input that do not join with your second input.

Columns are referenced with a **number**. For example, 3 refers to the 3rd column of a tab-delimited file.

Example

Figure 15.2.12 Joining on identifiers. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

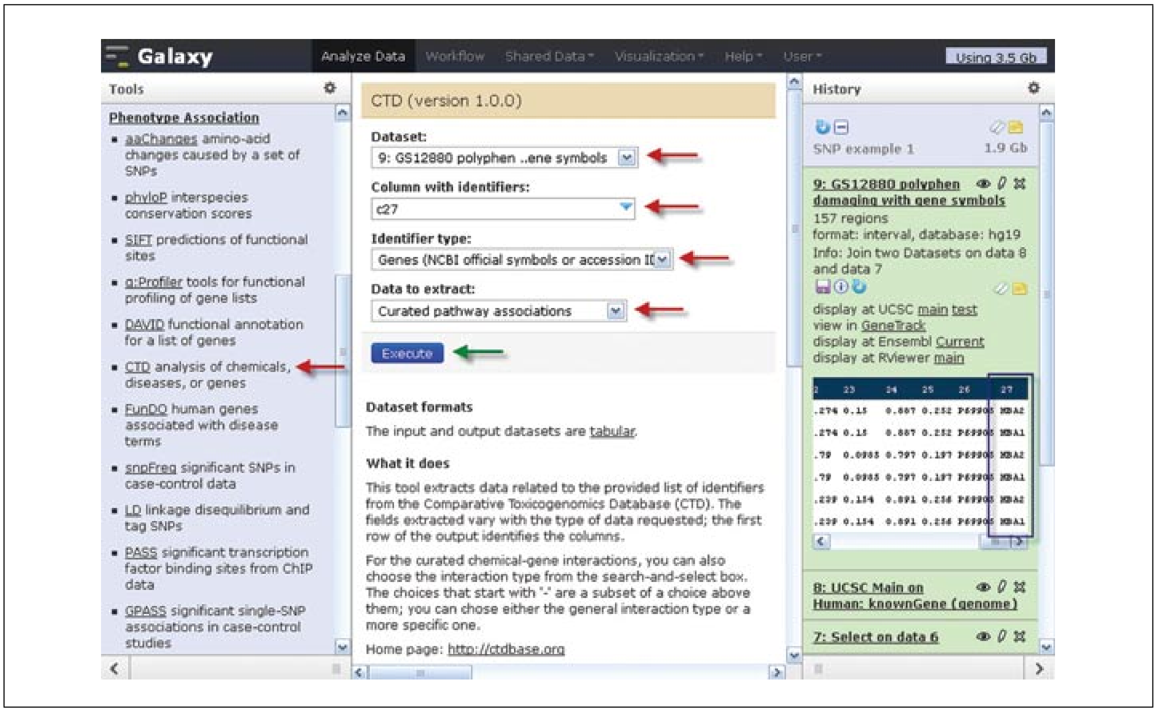
wider center panel, or need to access something in the other panels, just click the arrows again to reopen them. Here we have closed both panels to more easily see the UCSC Table Browser interface.

Choosing the identifier fields

9. Now we select the fields we want (Fig. 15.2.11). The main table does not have what we need, but the cross-reference table below it has both the UniProt ID and the official HUGO gene symbol (red arrows). After marking these boxes, click the "done with selections" button (green arrow). This will bring up one more page with just two buttons; click the one that says "Send query to Galaxy." The Table Browser interface will then close, and the requested data will be sent to your Galaxy history.

Joining on identifiers

10. Back at Galaxy, we are ready to do a join to get the gene symbols associated with the damaging SNPs. This time we do not have genomic positions in both sets, but they do have a field in common: the UniProt ID. We will do the join by matching up the values in those columns. In the section Join, Subtract, and Group there is a tool called Join Two Datasets (Fig. 15.2.12). The parameters for this tool are the two datasets, the columns to match up, and what to do with unmatched rows. By opening the two datasets in the history you can find the column numbers containing the UniProt IDs (blue boxes). Make the selections shown, putting the SNP dataset first so its fields will be first in the result rows. We are not interested in any of the rows that do not join, so leave the rest of the options set to No. Click Execute.



Phenotype Association

- aaChanges amino-acid changes caused by a set of SNPs
- phyloP interspecies conservation scores
- SIFT predictions of functional sites
- g:Profiler tools for functional profiling of gene lists
- DAVID functional annotation for a list of genes
- CTD analysis of chemicals, diseases, or genes
- FunDO human genes associated with disease terms
- snpFreq significant SNPs in case-control data
- LD linkage disequilibrium and tag SNPs
- PASS significant transcription factor binding sites from ChIP data
- GPASS significant single-SNP associations in case-control studies

CTD (version 1.0.0)

Dataset: 9: GS12880 polyphen ..ene symbols

Column with identifiers: c27

Identifier type: Genes (NCBI official symbols or accession ID

Data to extract: Curated pathway associations

Execute

Dataset formats

The input and output datasets are tabular.

What it does

This tool extracts data related to the provided list of identifiers from the Comparative Toxicogenomics Database (CTD). The fields extracted vary with the type of data requested; the first row of the output identifies the columns.

For the curated chemical-gene interactions, you can also choose the interaction type from the search-and-select box. The choices that start with '-' are a subset of a choice above them; you can chose either the general interaction type or a more specific one.

Home page: http://ctdbase.org

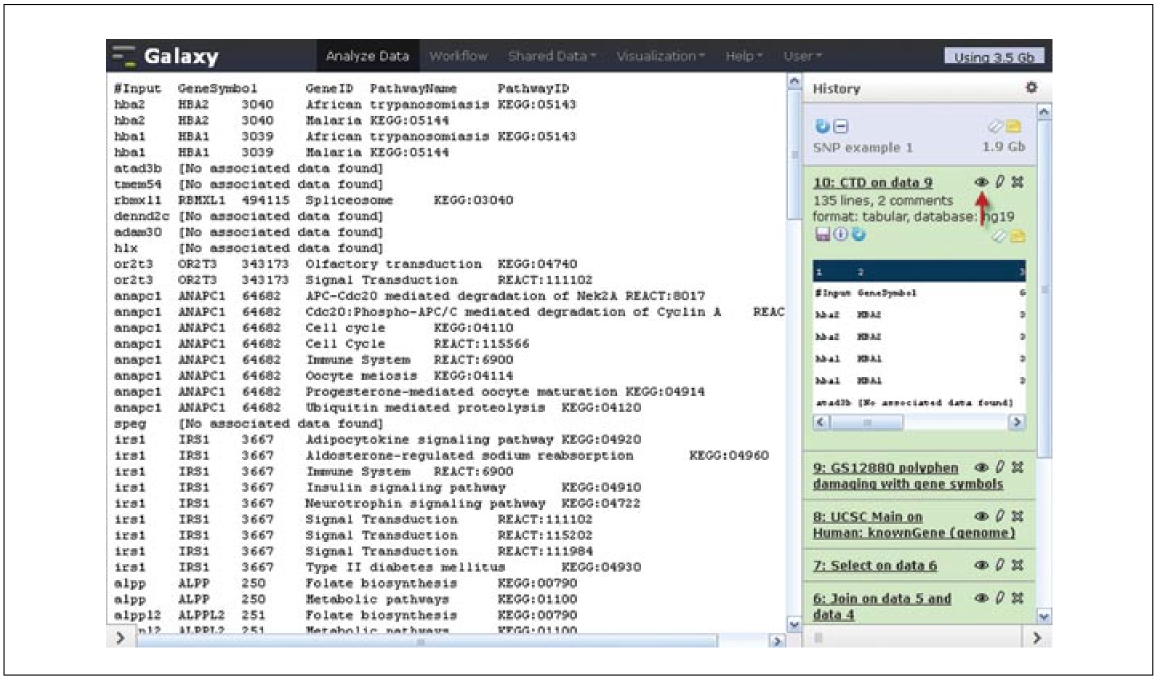
SNP example 1 1.9 Gb

9: GS12880 polyphen damaging with gene symbols
157 regions
format: interval, database: hg19
Info: Join two Datasets on data 8 and data 7
display at UCSC main test
view in GeneTrack
display at Ensembl Current
display at RViewer main

8: UCSC Main on Human: knownGene (genome)

7: Select on data 6

Figure 15.2.13 CTD. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.



```
#Input  GeneSymbol      GeneID  PathwayName     PathwayID
hba2    HBA2    3040    African trypanosomiasis KEGG:05143
hba2    HBA2    3040    Malaria KEGG:05144
hba1    HBA1    3039    African trypanosomiasis KEGG:05143
hba1    HBA1    3039    Malaria KEGG:05144
atad3b  [No associated data found]
tmem54  [No associated data found]
rbmxl1  RBMXL1  494115  Spliceosome     KEGG:03040
dennd2c [No associated data found]
adam30  [No associated data found]
hlx     [No associated data found]
or2t3   OR2T3   343173  Olfactory transduction  KEGG:04740
or2t3   OR2T3   343173  Signal Transduction     REACT:111102
anapc1  ANAPC1  64682   APC-Cdc20 mediated degradation of Nek2A REACT:8017
anapc1  ANAPC1  64682   Cdc20:Phospho-APC/C mediated degradation of Cyclin A    REAC
anapc1  ANAPC1  64682   Cell cycle      KEGG:04110
anapc1  ANAPC1  64682   Cell Cycle      REACT:115566
anapc1  ANAPC1  64682   Immune System   REACT:6900
anapc1  ANAPC1  64682   Oocyte meiosis  KEGG:04114
anapc1  ANAPC1  64682   Progesterone-mediated oocyte maturation KEGG:04914
anapc1  ANAPC1  64682   Ubiquitin mediated proteolysis  KEGG:04120
speg    [No associated data found]
irs1    IRS1    3667    Adipocytokine signaling pathway KEGG:04920
irs1    IRS1    3667    Aldosterone-regulated sodium reabsorption       KEGG:04960
irs1    IRS1    3667    Immune System   REACT:6900
irs1    IRS1    3667    Insulin signaling pathway       KEGG:04910
irs1    IRS1    3667    Neurotrophin signaling pathway  KEGG:04722
irs1    IRS1    3667    Signal Transduction     REACT:111102
irs1    IRS1    3667    Signal Transduction     REACT:115202
irs1    IRS1    3667    Signal Transduction     REACT:111984
irs1    IRS1    3667    Type II diabetes mellitus       KEGG:04930
alpp    ALPP    250     Folate biosynthesis     KEGG:00790
alpp    ALPP    250     Metabolic pathways      KEGG:01100
alppl2  ALPPL2  251     Folate biosynthesis     KEGG:00790
```

SNP example 1 1.9 Gb

10: CTD on data 9
135 lines, 2 comments
format: tabular, database: hg19

9: GS12880 polyphen damaging with gene symbols

8: UCSC Main on Human: knownGene (genome)

7: Select on data 6

6: Join on data 5 and data 4

Figure 15.2.14 CTD results. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

CTD

11. Now that we have a list of gene symbols for the putatively damaging SNPs, we can look for relationships among these genes. There are several tools useful for this, including CTD (Davis et al., 2009), g:Profiler (Reimand et al., 2007), and DAVID (Huang et al., 2009). For this example we will use the CTD tool to extract pathway information from the Comparative Toxicogenomics Database (Fig. 15.2.13).

Select the join results from the previous step, and column 27 which has the gene symbols (blue box). The identifier type we are using is NCBI gene symbols (which are the same as HUGO), and for the output we want the pathways. Click Execute.

CTD results

12. The results are a table of the genes and pathways (Fig. 15.2.14). From here, knowledge of the disease and/or the biology could be used to further filter the SNPs.

BASIC PROTOCOL 3

RUNNING NEW PREDICTIONS OF CODING SNPs LIKELY TO BE DETRIMENTAL

Overview

• Augment the input dataset (from Basic Protocol 1) to add a new column containing both the variant and reference alleles, by running the published workflow "Prep pgSnp file to run SIFT."

• Run the SIFT tool, keeping the original allele column as a comment, and requesting the gene name, OMIM disease, etc., in the output.

• Filter the results to select rows containing the word "DAMAGING."

Necessary Resources

Hardware

An Internet-connected computer

Software

Internet browser that supports JavaScript (e.g., most current browsers such as Mozilla Firefox, Safari, Opera, Chrome, or Microsoft Internet Explorer)

Input for SIFT

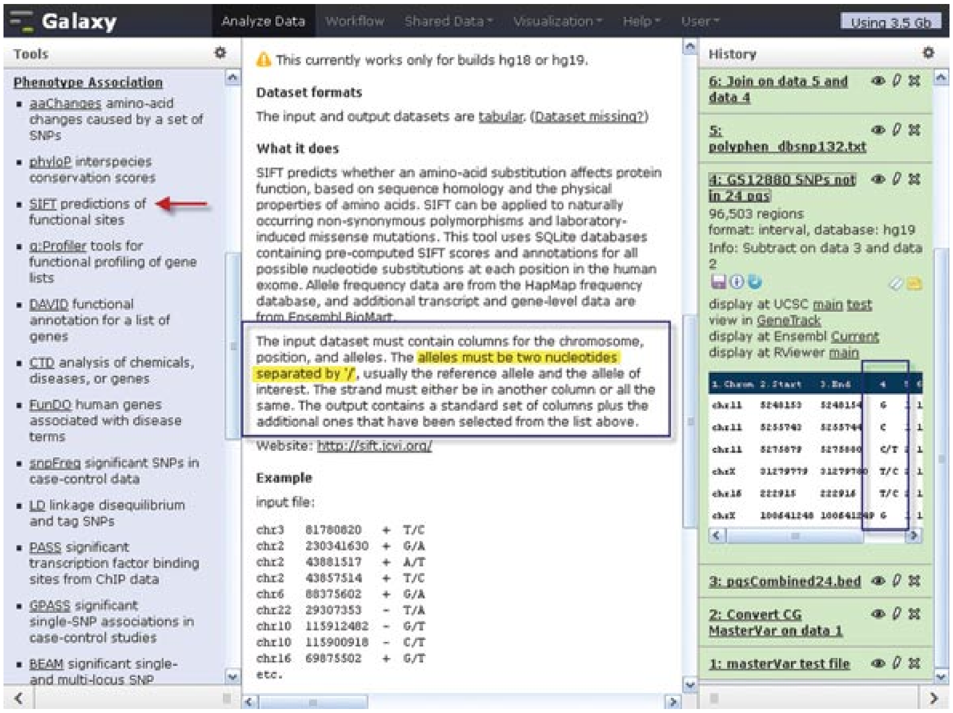
1. PolyPhen-2 is not the only way to predict functional coding changes with Galaxy; the SIFT tool (Kumar et al., 2009) in the Phenotype Association section performs a similar assessment. However, unlike the PolyPhen-2 library dataset which is precomputed only for known SNPs in the dbSNP database, the SIFT tool can make predictions for all possible nucleotide substitutions in the human exome.

   Each Galaxy tool typically has a description and other helpful information located below its input form, so to learn more about SIFT, open it and scroll down in the center panel as shown in Figure 15.2.15. Here, we find that the format needed (blue box) is a little different than that of our pgSnp dataset; SIFT requires two alleles at each SNP position, but pgSnp format lists only a single nucleotide for homozygous genotypes (blue box in history panel).

2. The SIFT instructions suggest adding the reference nucleotide in such cases; this can be done within Galaxy, but involves quite a few steps. For any complex task that you are likely to repeat, it is helpful to have a *workflow*, which is like an automatic recipe. Happily, there is a published workflow to convert from the pgSnp format to what SIFT needs, so we can do this in one step by using the workflow.

Finding the workflow

3. Click on Shared Data in the menu bar and select Published Workflows. In the search box, type `SIFT` and click on the magnifying glass. Find the workflow "Prep pgSnp file to run SIFT" in the resulting list and click on its name.



Phenotype Association

- aaChanges amino-acid changes caused by a set of SNPs
- phyloP interspecies conservation scores
- SIFT predictions of functional sites
- g:Profiler tools for functional profiling of gene lists
- DAVID functional annotation for a list of genes
- CTD analysis of chemicals, diseases, or genes
- FunDO human genes associated with disease terms
- snpFreq significant SNPs in case-control data
- LD linkage disequilibrium and tag SNPs
- PASS significant transcription factor binding sites from ChIP data
- GPASS significant single-SNP associations in case-control studies
- BEAM significant single- and multi-locus SNP

⚠ This currently works only for builds hg18 or hg19.

Dataset formats

The input and output datasets are tabular. (Dataset missing?)

What it does

SIFT predicts whether an amino-acid substitution affects protein function, based on sequence homology and the physical properties of amino acids. SIFT can be applied to naturally occurring non-synonymous polymorphisms and laboratory-induced missense mutations. This tool uses SQLite databases containing pre-computed SIFT scores and annotations for all possible nucleotide substitutions at each position in the human exome. Allele frequency data are from the HapMap frequency database, and additional transcript and gene-level data are from Ensembl BioMart.

The input dataset must contain columns for the chromosome, position, and alleles. The alleles must be two nucleotides separated by '/', usually the reference allele and the allele of interest. The strand must either be in another column or all the same. The output contains a standard set of columns plus the additional ones that have been selected from the list above.

Website: http://sift.jcvi.org/

Example

input file:

```
chr3    81780820   +  T/C
chr2    230341630  +  G/A
chr2    43881517   +  A/T
chr2    43857514   +  T/C
chr6    88375602   +  G/A
chr22   29307353   -  T/A
chr10   115912482  -  G/T
chr10   115900918  -  C/T
chr16   69875502   +  G/T
etc.
```

6: Join on data 5 and data 4

5: polyphen_dbsnp132.txt

4: GS12880 SNPs not in 24 pgs
96,503 regions
format: interval, database: hg19
Info: Subtract on data 3 and data 2
display at UCSC main test
view in GeneTrack
display at Ensembl Current
display at RViewer main

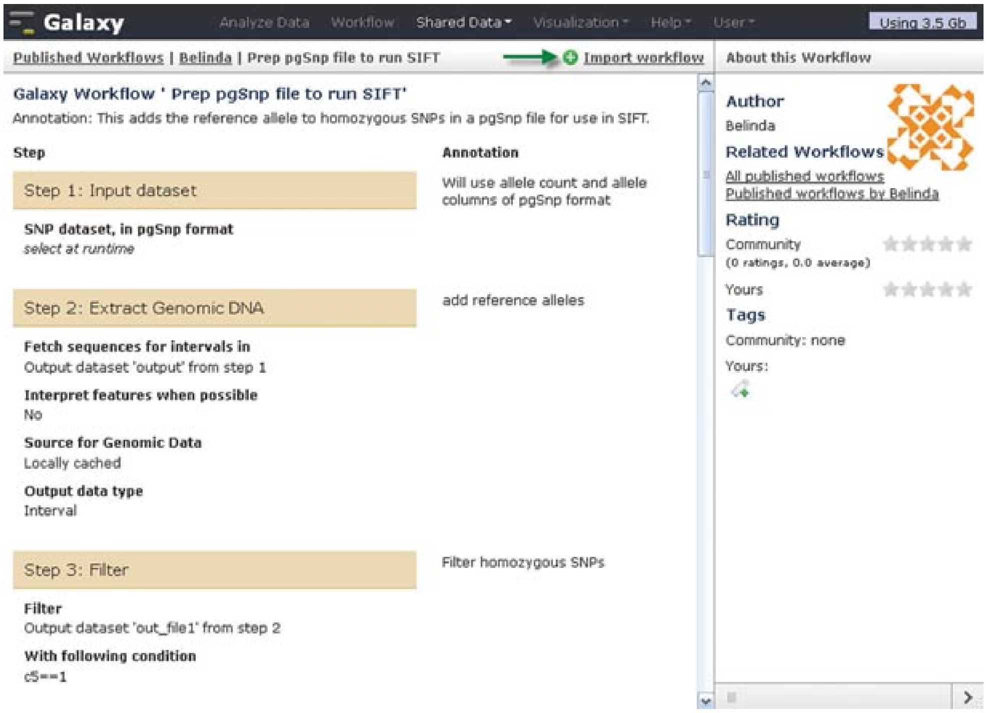
1. Chrom	2. Start	3. End	4	5	6
chr11	5248153	5248154	G	[illegible]	1
chr11	5255743	5255744	C	[illegible]	1
chr11	5275879	5275880	C/T	[illegible]	1
chrX	31279779	31279780	T/C	[illegible]	1
chr16	222915	222916	T/C	[illegible]	1
chrX	100641248	100641249	G	[illegible]	1

3: pgsCombined24.bed

2: Convert CG MasterVar on data 1

1: masterVar test file

Figure 15.2.15 Input for SIFT. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.



Published Workflows | Belinda | Prep pgSnp file to run SIFT

Galaxy Workflow ' Prep pgSnp file to run SIFT'

Annotation: This adds the reference allele to homozygous SNPs in a pgSnp file for use in SIFT.

Step	Annotation
Step 1: Input dataset	Will use allele count and allele columns of pgSnp format
SNP dataset, in pgSnp format *select at runtime*	
Step 2: Extract Genomic DNA	add reference alleles
Fetch sequences for intervals in Output dataset 'output' from step 1 **Interpret features when possible** No **Source for Genomic Data** Locally cached **Output data type** Interval	
Step 3: Filter	Filter homozygous SNPs
Filter Output dataset 'out_file1' from step 2 **With following condition** c5==1	

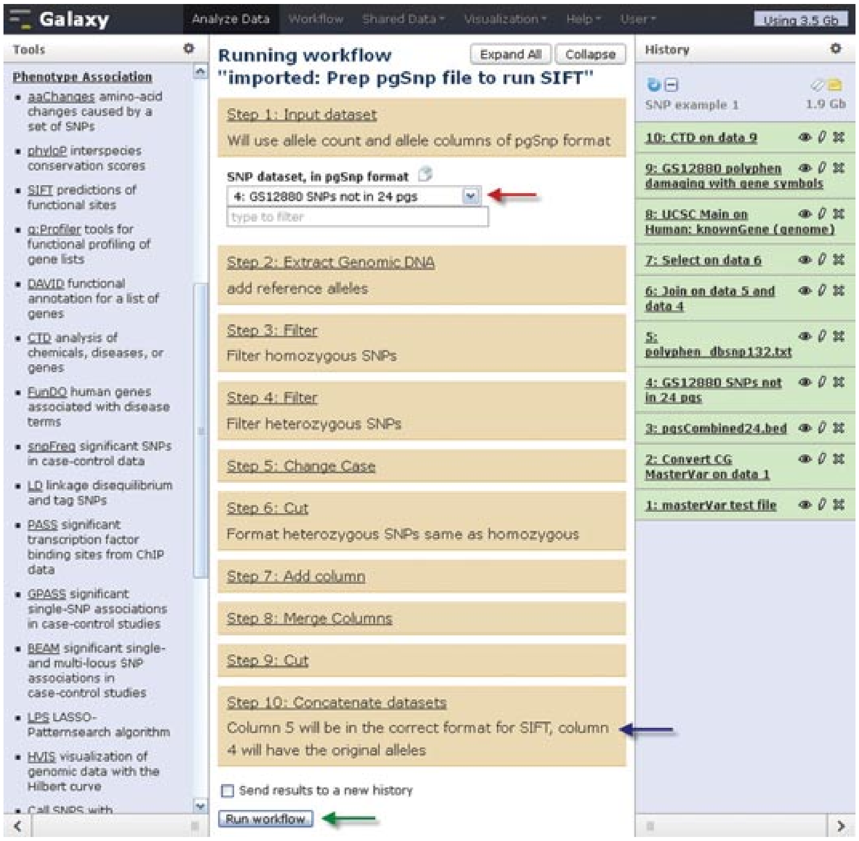
Author
Belinda

Related Workflows
All published workflows
Published workflows by Belinda

Rating
Community ☆☆☆☆☆
(0 ratings, 0.0 average)
Yours ☆☆☆☆☆

Tags
Community: none
Yours:

Figure 15.2.16 Viewing the workflow. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.



Phenotype Association

- aaChanges amino-acid changes caused by a set of SNPs
- phyloP interspecies conservation scores
- SIFT predictions of functional sites
- g:Profiler tools for functional profiling of gene lists
- DAVID functional annotation for a list of genes
- CTD analysis of chemicals, diseases, or genes
- FunDO human genes associated with disease terms
- snpFreq significant SNPs in case-control data
- LD linkage disequilibrium and tag SNPs
- PASS significant transcription factor binding sites from ChIP data
- GPASS significant single-SNP associations in case-control studies
- BEAM significant single- and multi-locus SNP associations in case-control studies
- LPS LASSO-Patternsearch algorithm
- HVIS visualization of genomic data with the Hilbert curve
- Call SNPS with

Running workflow "imported: Prep pgSnp file to run SIFT"

Step 1: Input dataset
Will use allele count and allele columns of pgSnp format

SNP dataset, in pgSnp format: 4: GS12880 SNPs not in 24 pgs

Step 2: Extract Genomic DNA
add reference alleles

Step 3: Filter
Filter homozygous SNPs

Step 4: Filter
Filter heterozygous SNPs

Step 5: Change Case

Step 6: Cut
Format heterozygous SNPs same as homozygous

Step 7: Add column

Step 8: Merge Columns

Step 9: Cut

Step 10: Concatenate datasets
Column 5 will be in the correct format for SIFT, column 4 will have the original alleles

SNP example 1 1.9 Gb

10: CTD on data 9
9: GS12880 polyphen damaging with gene symbols
8: UCSC Main on Human: knownGene (genome)
7: Select on data 6
6: Join on data 5 and data 4
5: polyphen_dbsnp132.txt
4: GS12880 SNPs not in 24 pgs
3: pgsCombined24.bed
2: Convert CG MasterVar on data 1
1: masterVar test file

Figure 15.2.17 Running the workflow. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

Viewing the workflow

4. The full workflow is visible, along with annotations (Fig. 15.2.16). There is a description of what the workflow does, and some notes for individual steps. The right panel shows the author, user rating, and tags for this workflow. Click on "Import workflow" (green arrow) to bring this into your personal set of workflows. Then, on the confirmation screen, click "start using this workflow."

Your workflows

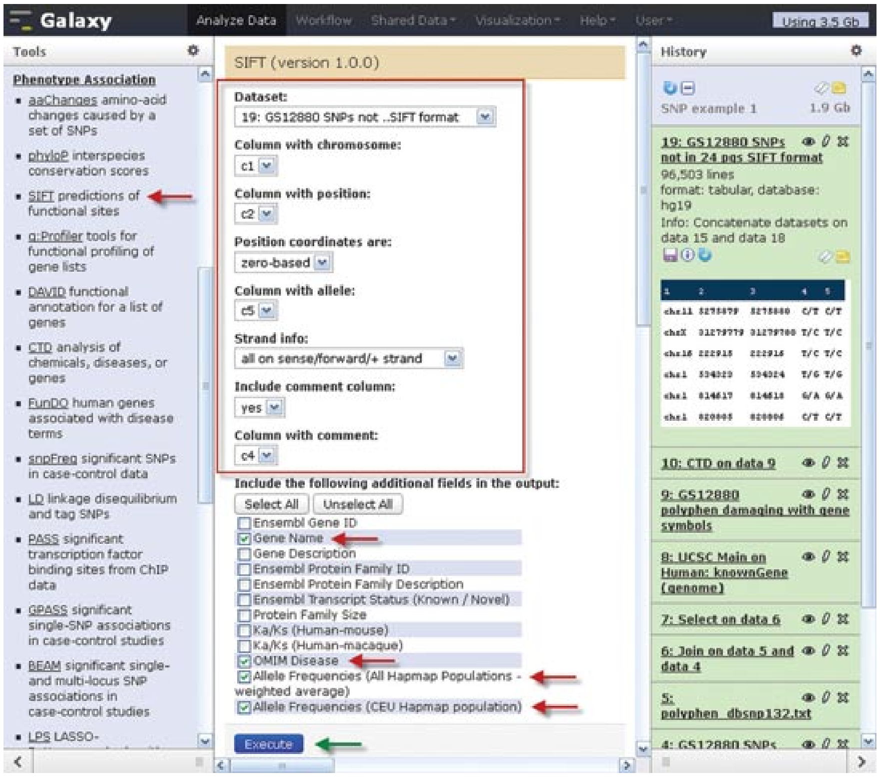
5. You will be presented with the list of workflows that you have acquired. Click on the name of the one you just imported, and select Run from the menu that pops up.

Running the workflow

6. The workflow is in the center panel (Fig. 15.2.17); for its input select the same results from Basic Protocol 1 that we used in Basic Protocol 2 for the join with the PolyPhen-2 predictions. Several of the steps have annotations explaining what the workflow will do. The annotation for step 10 (blue arrow) tells us that column 5 of the output will be what we want to submit to SIFT, and that column 4 will have the original allele data. There are ten steps in the workflow, all of which will be run when you click the "Run workflow" button.

SIFT

7. When the workflow has completed, open the SIFT tool again and fill in the parameters as shown in the red box in Figure 15.2.18. In pgSnp format, the alleles are always on the forward strand, so select "all on sense/forward/+ strand." By choosing "yes" for



Phenotype Association

- aaChanges amino-acid changes caused by a set of SNPs
- phyloP interspecies conservation scores
- SIFT predictions of functional sites
- g:Profiler tools for functional profiling of gene lists
- DAVID functional annotation for a list of genes
- CTD analysis of chemicals, diseases, or genes
- FunDO human genes associated with disease terms
- snpFreq significant SNPs in case-control data
- LD linkage disequilibrium and tag SNPs
- PASS significant transcription factor binding sites from ChIP data
- GPASS significant single-SNP associations in case-control studies
- BEAM significant single- and multi-locus SNP associations in case-control studies
- LPS LASSO-

SIFT (version 1.0.0)

Dataset: 19: GS12880 SNPs not ..SIFT format

Column with chromosome: c1

Column with position: c2

Position coordinates are: zero-based

Column with allele: c5

Strand info: all on sense/forward/+ strand

Include comment column: yes

Column with comment: c4

Include the following additional fields in the output:

Select All | Unselect All

- □ Ensembl Gene ID
- ☑ Gene Name
- □ Gene Description
- □ Ensembl Protein Family ID
- □ Ensembl Protein Family Description
- □ Ensembl Transcript Status (Known / Novel)
- □ Protein Family Size
- □ Ka/Ks (Human-mouse)
- □ Ka/Ks (Human-macaque)
- ☑ OMIM Disease
- ☑ Allele Frequencies (All Hapmap Populations - weighted average)
- ☑ Allele Frequencies (CEU Hapmap population)

SNP example 1 1.9 Gb

19: GS12880 SNPs not in 24 pgs SIFT format
96,503 lines
format: tabular, database: hg19
Info: Concatenate datasets on data 15 and data 18

1	2	3	4	5
chr11	5275879	5275880	C/T	C/T
chrX	31279779	31279780	T/C	T/C
chr16	222915	222916	T/C	T/C
chr1	534023	534024	T/G	T/G
chr1	814617	814618	G/A	G/A
chr1	820005	820006	C/T	C/T

10: CTD on data 9

9: GS12880 polyphen damaging with gene symbols

8: UCSC Main on Human: knownGene (genome)

7: Select on data 6

6: Join on data 5 and data 4

5: polyphen_dbsnp132.txt

4: GS12880 SNPs

Figure 15.2.18 SIFT. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

"Include comment column" we can keep column 4, which has the original alleles, in the output. We have also selected some of the optional fields to be included in our output (red arrows). Click Execute. This tool may take some time to complete, so please be patient.

The workflow's intermediate steps are hidden from view as they complete, to make the history less cluttered. If you wish to see all of the steps in your history, click on the gear icon at the top of the history panel. A menu will open up, and one of the choices is Show Hidden Datasets.

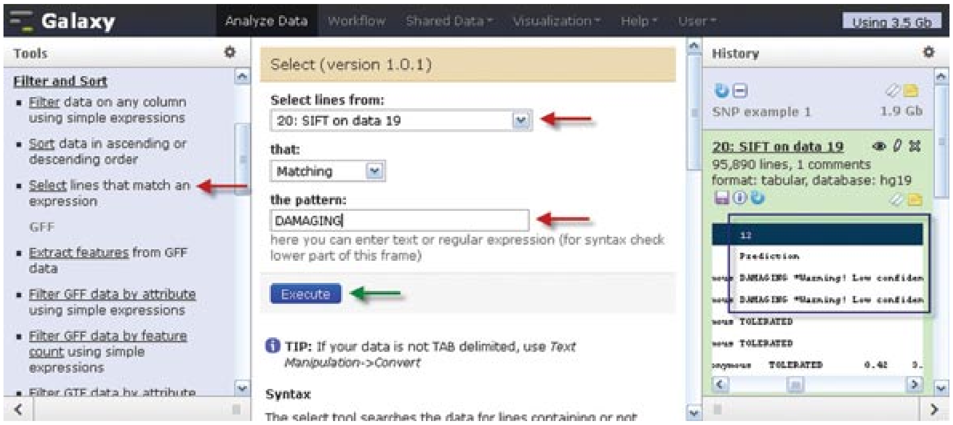
Selecting damaging SNPs

8. As before with the PolyPhen-2 predictions, we now need to filter the SNPs to select the damaging ones. The SIFT output in the history panel (blue box in Fig. 15.2.19) shows that it uses `DAMAGING` in upper case. Once the parameters are set, click Execute.

SIFT results

9. Clicking on the eye icon displays the results in the center panel (Fig. 15.2.20). To see them better, we have temporarily hidden the tool panel by clicking on the arrow button in the bottom left corner. We also renamed this dataset. Within the 209 predicted damaging SNPs, we see that both of our coding disease SNPs are found (blue box). Because we requested the OMIM output field when running SIFT, we have the diseases from OMIM listed here.

Other tools such as CTD and DAVID could now be run on these genes if desired, but for this example we will move on to hunting for disease SNPs in non-coding regions.


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 3.5 Gb
Tools
Filter and Sort
Filter data on any column using simple expressions
Sort data in ascending or descending order
Select lines that match an expression
GFF
Extract features from GFF data
Filter GFF data by attribute using simple expressions
Filter GFF data by feature count using simple expressions
Filter GTF data by attribute
History
SNP example 1
1.9 Gb

Select (version 1.0.1)

Select lines from: 20: SIFT on data 19

that: Matching

the pattern: DAMAGING

here you can enter text or regular expression (for syntax check lower part of this frame)

Execute

TIP: If your data is not TAB delimited, use *Text Manipulation->Convert*

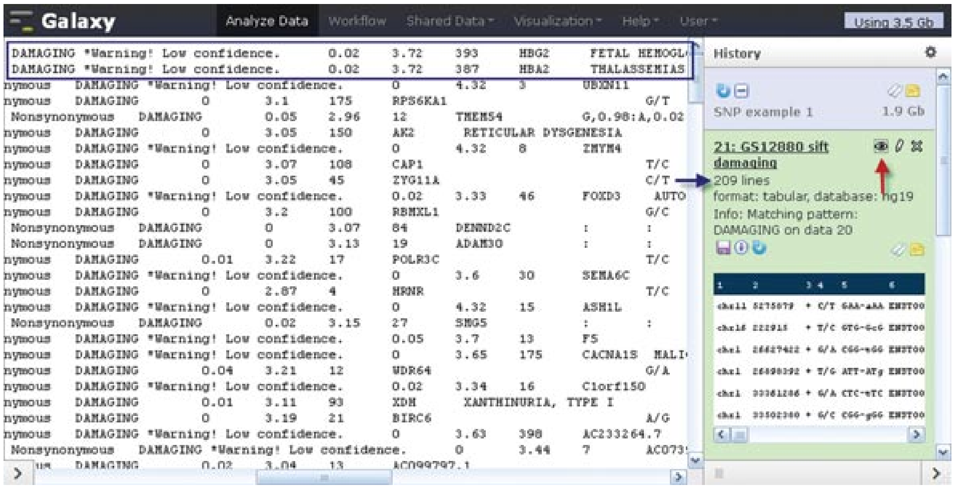
Syntax

The select tool searches the data for lines containing or not

20: SIFT on data 19
95,890 lines, 1 comments
format: tabular, database: hg19

```
     12
     Prediction
nous DAMAGING *Warning! Low confiden
nous DAMAGING *Warning! Low confiden
nous TOLERATED
nous TOLERATED
onymous   TOLERATED        0.42    3.
```

Figure 15.2.19 Selecting damaging SNPs. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.



```
DAMAGING *Warning! Low confidence.      0.02     3.72     393      HBG2      FETAL HEMOGL
DAMAGING *Warning! Low confidence.      0.02     3.72     387      HBA2      THALASSEMIAS
nymous   DAMAGING *Warning! Low confidence.      0        4.32     3         UBXN11
nymous   DAMAGING        0       3.1      175      RPS6KA1                          G/T
 Nonsynonymous   DAMAGING        0.05     2.96     12       THEM54           G,0.98:A,0.02
nymous   DAMAGING        0       3.05     150      AK2       RETICULAR DYSGENESIA
nymous   DAMAGING *Warning! Low confidence.      0        4.32     8         ZMYM4
nymous   DAMAGING        0       3.07     108      CAP1                             T/C
nymous   DAMAGING        0       3.05     45       ZYG11A                           C/T
nymous   DAMAGING *Warning! Low confidence.      0.02     3.33     46        FOXD3     AUTO
nymous   DAMAGING        0       3.2      100      RBMXL1                           G/C
 Nonsynonymous   DAMAGING        0        3.07     84       DENND2C          :         :
 Nonsynonymous   DAMAGING        0        3.13     19       ADAM30           :         :
nymous   DAMAGING        0.01    3.22     17       POLR3C                           T/C
nymous   DAMAGING *Warning! Low confidence.      0        3.6      30        SEMA6C
nymous   DAMAGING        0       2.87     4        HRNR                             T/C
nymous   DAMAGING *Warning! Low confidence.      0        4.32     15        ASH1L
 Nonsynonymous   DAMAGING        0.02     3.15     27       SMG5             :         :
nymous   DAMAGING *Warning! Low confidence.      0.05     3.7      13        F5
nymous   DAMAGING *Warning! Low confidence.      0        3.65     175       CACNA1S   MALI
nymous   DAMAGING        0.04    3.21     12       WDR64                            G/A
nymous   DAMAGING *Warning! Low confidence.      0.02     3.34     16        C1orf150
nymous   DAMAGING        0.01    3.11     93       XDH       XANTHINURIA, TYPE I
nymous   DAMAGING        0       3.19     21       BIRC6                            A/G
nymous   DAMAGING *Warning! Low confidence.      0        3.63     398       AC233264.7
 Nonsynonymous   DAMAGING *Warning! Low confidence.      0        3.44     7         AC073
us   DAMAGING        0.02    3.04     13       AC099797.1
```

21: GS12880 sift damaging
209 lines
format: tabular, database: hg19
Info: Matching pattern: DAMAGING on data 20

```
1      2          3 4    5          6
chr11  5275079    + C/T  GAA-aAA    ENST00
chr16  222915     + T/C  GTG-GcG    ENST00
chr1   26627422   + G/A  CGG-aGG    ENST00
chr1   26698392   + T/G  ATT-ATg    ENST00
chr1   33361286   + G/A  CTC-aTC    ENST00
chr1   33502360   + G/C  CGG-gGG    ENST00
```

Figure 15.2.20 SIFT results. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

BASIC PROTOCOL 4

FINDING SNPs THAT FALL IN SUSPECTED FUNCTIONAL REGIONS

Overview

- Filter the input dataset (from Basic Protocol 1) to keep only rows whose intervals overlap those in a library dataset of predicted regulatory regions.

- In a similar fashion, find rows in the same input dataset that overlap with those in an ENCODE regulatory dataset (DNase clusters) obtained from UCSC.

- Run the PhyloP tool on the same input dataset to add a column of interspecies conservation scores. Then, use the Histogram tool to help choose a suitable score threshold, and filter the SNPs on the score column to keep only those at highly conserved positions.

Necessary Resources

Hardware

An Internet-connected computer

Software

Internet browser that supports JavaScript (e.g., most current browsers such as Mozilla Firefox, Safari, Opera, Chrome, or Microsoft Internet Explorer)

Predicted regulatory regions (PRPs)

1. First we will import another dataset from the Putative SNP Phenotypes library, this time `PRPs.liftedhg19.bed`, which contains computationally predicted regulatory regions (PRPs). Do not forget to click on Analyze Data after importing the dataset to get back to your history and tool page.

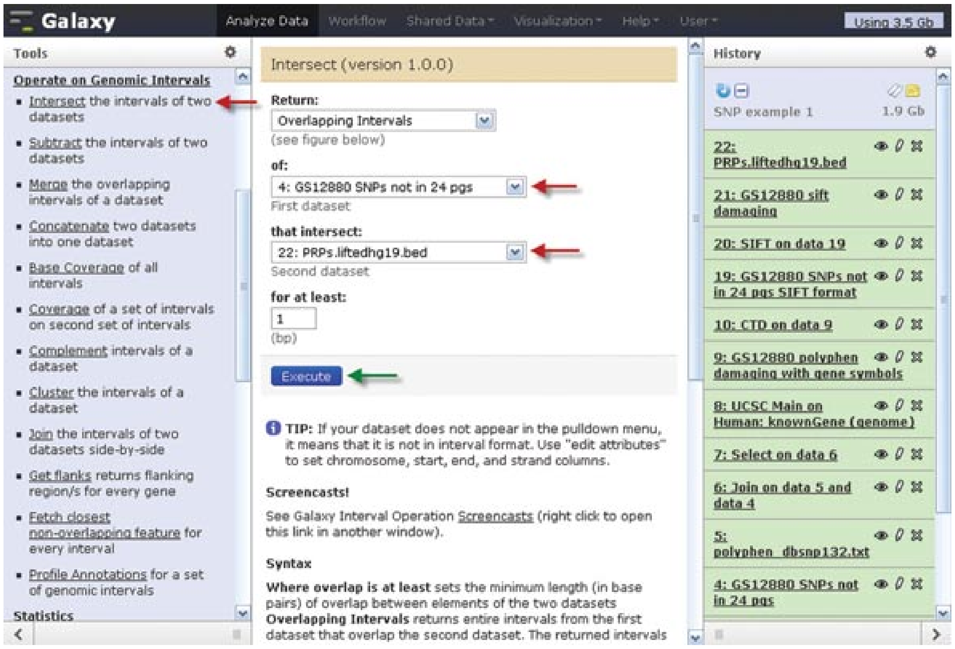
   This is the intersection of the PreMods (Ferretti et al., 2007) and the regions with high Regulatory Potential scores (Taylor et al. 2006). Both of these sets were computed on earlier genome assemblies and then lifted (i.e., mapped) to the current assembly.

Intersecting with the PRPs

2. Exclusion of the SNPs from the 24 healthy individuals is useful for finding both coding and non-coding disease variants, so we will again start with the dataset we prepared in Basic Protocol 1. In the Operate on Genomic Intervals section of the tool panel, select and run the Intersect tool as shown in Figure 15.2.21. An intersection between our input dataset and the PRPs will find the SNPs within the predicted regulatory regions. The output will have the columns from the first dataset, so be sure to specify the SNPs as the first dataset and the PRPs as the second one.

SNPs in PRPs

3. There are 837 SNPs in the predicted regulatory regions (Fig. 15.2.22), including two of our non-coding disease SNPs (blue box).



Operate on Genomic Intervals

- Intersect the intervals of two datasets
- Subtract the intervals of two datasets
- Merge the overlapping intervals of a dataset
- Concatenate two datasets into one dataset
- Base Coverage of all intervals
- Coverage of a set of intervals on second set of intervals
- Complement intervals of a dataset
- Cluster the intervals of a dataset
- Join the intervals of two datasets side-by-side
- Get flanks returns flanking region/s for every gene
- Fetch closest non-overlapping feature for every interval
- Profile Annotations for a set of genomic intervals

Statistics

Intersect (version 1.0.0)

Return:
Overlapping Intervals
(see figure below)

of:
4: GS12880 SNPs not in 24 pgs
First dataset

that intersect:
22: PRPs.liftedhg19.bed
Second dataset

for at least:
1
(bp)

Execute

TIP: If your dataset does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.

Screencasts!

See Galaxy Interval Operation Screencasts (right click to open this link in another window).

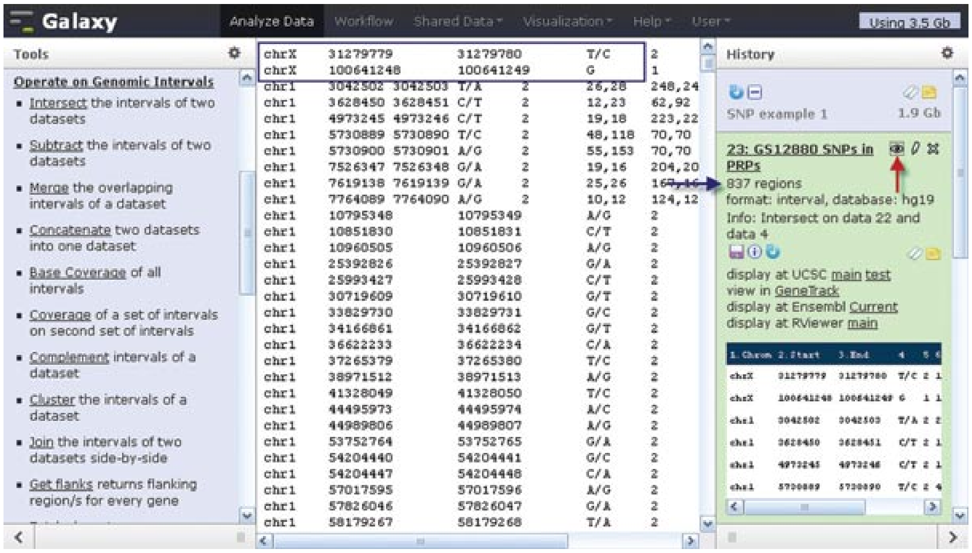
Syntax

Where overlap is at least sets the minimum length (in base pairs) of overlap between elements of the two datasets
Overlapping Intervals returns entire intervals from the first dataset that overlap the second dataset. The returned intervals

SNP example 1 1.9 Gb

22: PRPs.liftedhg19.bed
21: GS12880 sift damaging
20: SIFT on data 19
19: GS12880 SNPs not in 24 pgs SIFT format
10: CTD on data 9
9: GS12880 polyphen damaging with gene symbols
8: UCSC Main on Human: knownGene (genome)
7: Select on data 6
6: Join on data 5 and data 4
5: polyphen_dbsnp132.txt
4: GS12880 SNPs not in 24 pgs

Figure 15.2.21 Intersecting with the PRPs. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.



Operate on Genomic Intervals

- Intersect the intervals of two datasets
- Subtract the intervals of two datasets
- Merge the overlapping intervals of a dataset
- Concatenate two datasets into one dataset
- Base Coverage of all intervals
- Coverage of a set of intervals on second set of intervals
- Complement intervals of a dataset
- Cluster the intervals of a dataset
- Join the intervals of two datasets side-by-side
- Get flanks returns flanking region/s for every gene

```
chrX    31279779            31279780        T/C        2
chrX    100641248           100641249       G          1
chr1    3042502  3042503  T/A      2        26,28      248,24
chr1    3628450  3628451  C/T      2        12,23      62,92
chr1    4973245  4973246  C/T      2        19,18      223,22
chr1    5730889  5730890  T/C      2        48,118     70,70
chr1    5730900  5730901  A/G      2        55,153     70,70
chr1    7526347  7526348  G/A      2        19,16      204,20
chr1    7619138  7619139  G/A      2        25,26      167,16
chr1    7764089  7764090  A/G      2        10,12      124,12
chr1    10795348            10795349        A/G        2
chr1    10851830            10851831        C/T        2
chr1    10960505            10960506        A/G        2
chr1    25392826            25392827        G/A        2
chr1    25993427            25993428        C/T        2
chr1    30719609            30719610        G/T        2
chr1    33829730            33829731        G/C        2
chr1    34166861            34166862        G/T        2
chr1    36622233            36622234        C/A        2
chr1    37265379            37265380        T/C        2
chr1    38971512            38971513        A/G        2
chr1    41328049            41328050        T/C        2
chr1    44495973            44495974        A/C        2
chr1    44989806            44989807        A/G        2
chr1    53752764            53752765        G/A        2
chr1    54204440            54204441        G/C        2
chr1    54204447            54204448        C/A        2
chr1    57017595            57017596        A/G        2
chr1    57826046            57826047        G/A        2
chr1    58179267            58179268        T/A        2
```

SNP example 1 1.9 Gb

23: GS12880 SNPs in PRPs
837 regions
format: interval, database: hg19
Info: Intersect on data 22 and data 4
display at UCSC main test
view in GeneTrack
display at Ensembl Current
display at RViewer main

1. Chrom	2. Start	3. End	4	5	6
chrX	31279779	31279780	T/C	2	1
chrX	100641248	100641249	G	1	1
chr1	3042502	3042503	T/A	2	2
chr1	3628450	3628451	C/T	2	1
chr1	4973245	4973246	C/T	2	1
chr1	5730889	5730890	T/C	2	4

Figure 15.2.22 SNPs in PRPs. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

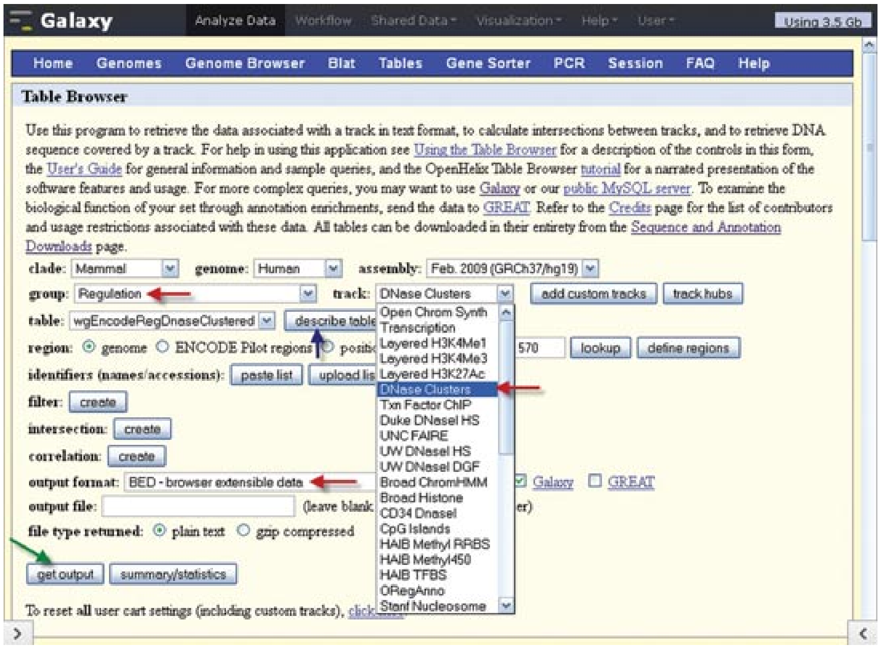


Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

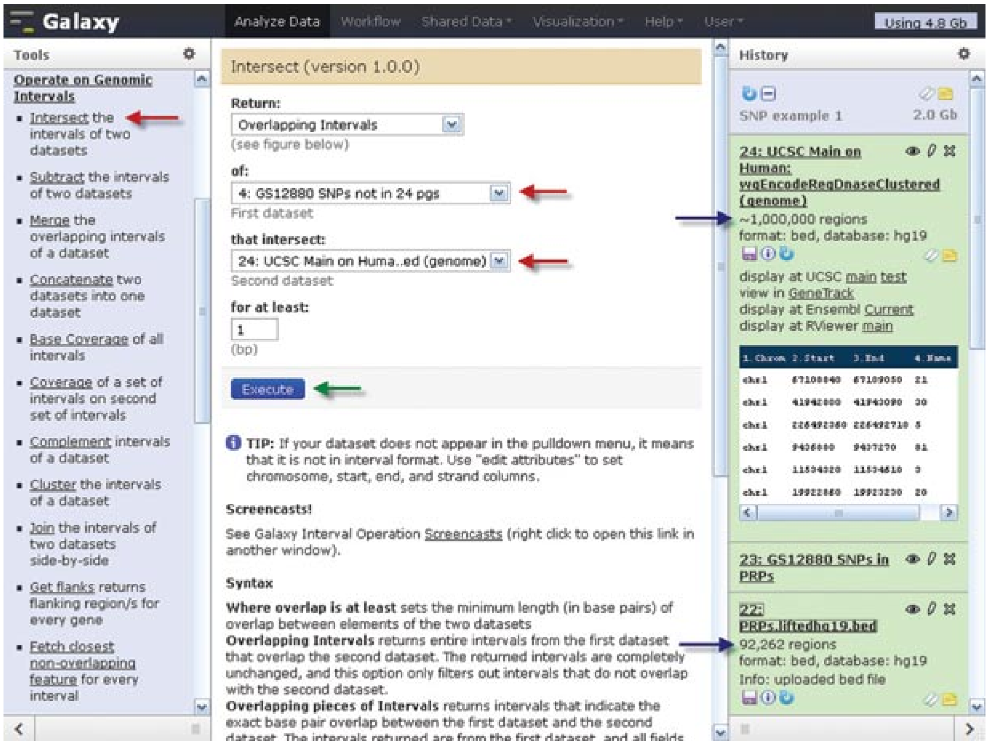
clade: Mammal genome: Human assembly: Feb. 2009 (GRCh37/hg19)
group: Regulation track: DNase Clusters add custom tracks track hubs
table: wgEncodeRegDnaseClustered describe table
region: ● genome ○ ENCODE Pilot regions ○ positi 570 lookup define regions
identifiers (names/accessions): paste list upload lis
filter: create
intersection: create
correlation: create
output format: BED - browser extensible data ☑ Galaxy □ GREAT
output file: (leave blank er)
file type returned: ● plain text ○ gzip compressed
get output summary/statistics

To reset all user cart settings (including custom tracks), click

Figure 15.2.23 DNase hypersensitive sites (HSSs) from ENCODE. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

DNase hypersensitive sites (HSSs) from ENCODE data

4. To obtain data from the ENCODE project, we return to the UCSC Main Table Browser (located in the Get Data section of the tool panel). As before, in order to see the full width of the table browser interface without scrolling, we have temporarily hidden the tool and history panels via the bottom corner arrow buttons (Fig. 15.2.23).



Operate on Genomic Intervals

- Intersect the intervals of two datasets
- Subtract the intervals of two datasets
- Merge the overlapping intervals of a dataset
- Concatenate two datasets into one dataset
- Base Coverage of all intervals
- Coverage of a set of intervals on second set of intervals
- Complement intervals of a dataset
- Cluster the intervals of a dataset
- Join the intervals of two datasets side-by-side
- Get flanks returns flanking region/s for every gene
- Fetch closest non-overlapping feature for every interval

Intersect (version 1.0.0)

Return:
Overlapping Intervals
(see figure below)

of:
4: GS12880 SNPs not in 24 pgs
First dataset

that intersect:
24: UCSC Main on Huma..ed (genome)
Second dataset

for at least:
1
(bp)

Execute

TIP: If your dataset does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.

Screencasts!

See Galaxy Interval Operation Screencasts (right click to open this link in another window).

Syntax

Where overlap is at least sets the minimum length (in base pairs) of overlap between elements of the two datasets
Overlapping Intervals returns entire intervals from the first dataset that overlap the second dataset. The returned intervals are completely unchanged, and this option only filters out intervals that do not overlap with the second dataset.
Overlapping pieces of Intervals returns intervals that indicate the exact base pair overlap between the first dataset and the second dataset. The intervals returned are from the first dataset, and all fields

SNP example 1 2.0 Gb

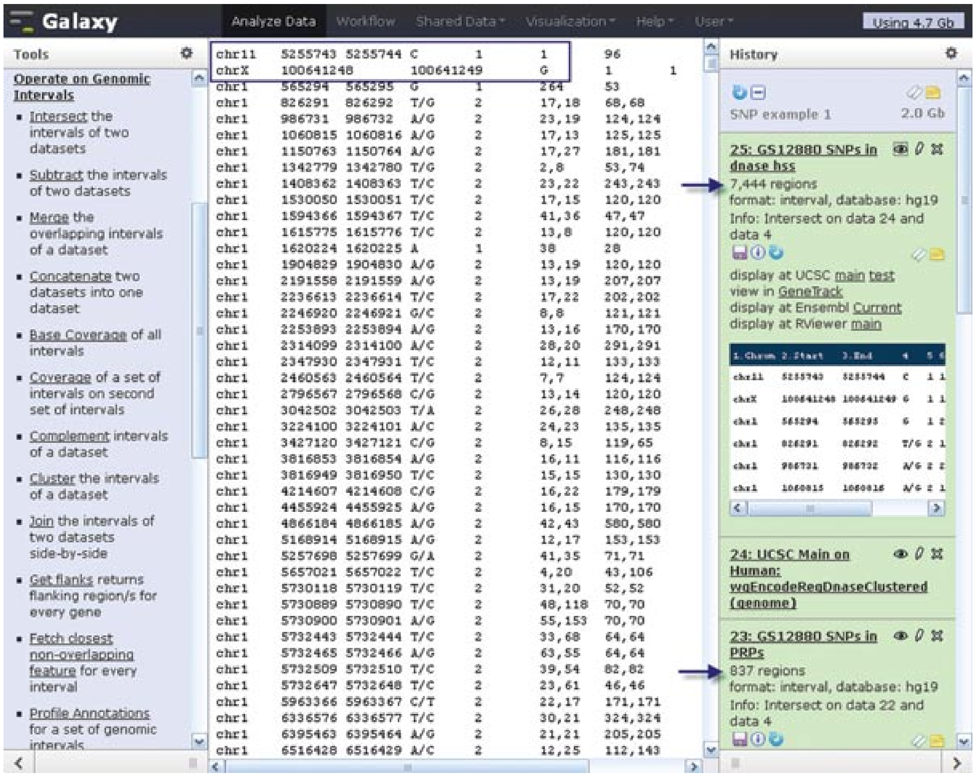
24: UCSC Main on Human: wgEncodeRegDnaseClustered (genome)
~1,000,000 regions
format: bed, database: hg19
display at UCSC main test
view in GeneTrack
display at Ensembl Current
display at RViewer main

1.Chrom	2.Start	3.End	4.Name
chr1	67100840	67109050	21
chr1	41942000	41943090	30
chr1	226492360	226492710	5
chr1	9436000	9437270	81
chr1	11534320	11534810	3
chr1	19922060	19923230	20

23: GS12880 SNPs in PRPs

22: PRPs.liftedhg19.bed
92,262 regions
format: bed, database: hg19
Info: uploaded bed file

Figure 15.2.24 Intersecting with the HSSs. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.



Operate on Genomic Intervals

- Intersect the intervals of two datasets
- Subtract the intervals of two datasets
- Merge the overlapping intervals of a dataset
- Concatenate two datasets into one dataset
- Base Coverage of all intervals
- Coverage of a set of intervals on second set of intervals
- Complement intervals of a dataset
- Cluster the intervals of a dataset
- Join the intervals of two datasets side-by-side
- Get flanks returns flanking region/s for every gene
- Fetch closest non-overlapping feature for every interval
- Profile Annotations for a set of genomic intervals

```
chr11   5255743 5255744 C       1       1       96
chrX    100641248       100641249       G       1       1
chr1    565294  565295  G       1       264     53
chr1    826291  826292  T/G     2       17,18   68,68
chr1    986731  986732  A/G     2       23,19   124,124
chr1    1060815 1060816 A/G     2       17,13   125,125
chr1    1150763 1150764 A/G     2       17,27   181,181
chr1    1342779 1342780 T/G     2       2,8     53,74
chr1    1408362 1408363 T/C     2       23,22   243,243
chr1    1530050 1530051 T/C     2       17,15   120,120
chr1    1594366 1594367 T/C     2       41,36   47,47
chr1    1615775 1615776 T/C     2       13,8    120,120
chr1    1620224 1620225 A       1       38      28
chr1    1904829 1904830 A/G     2       13,19   120,120
chr1    2191558 2191559 A/G     2       13,19   207,207
chr1    2236613 2236614 T/C     2       17,22   202,202
chr1    2246920 2246921 G/C     2       8,8     121,121
chr1    2253893 2253894 A/G     2       13,16   170,170
chr1    2314099 2314100 A/C     2       28,20   291,291
chr1    2347930 2347931 T/C     2       12,11   133,133
chr1    2460563 2460564 T/C     2       7,7     124,124
chr1    2796567 2796568 C/G     2       13,14   120,120
chr1    3042502 3042503 T/A     2       26,28   248,248
chr1    3224100 3224101 A/C     2       24,23   135,135
chr1    3427120 3427121 C/G     2       8,15    119,65
chr1    3816853 3816854 A/G     2       16,11   116,116
chr1    3816949 3816950 T/C     2       15,15   130,130
chr1    4214607 4214608 C/G     2       16,22   179,179
chr1    4455924 4455925 A/G     2       16,15   170,170
chr1    4866184 4866185 A/G     2       42,43   580,580
chr1    5168914 5168915 A/G     2       12,17   153,153
chr1    5257698 5257699 G/A     2       41,35   71,71
chr1    5657021 5657022 T/C     2       4,20    43,106
chr1    5730118 5730119 T/C     2       31,20   52,52
chr1    5730889 5730890 T/C     2       48,118  70,70
chr1    5730900 5730901 A/G     2       55,153  70,70
chr1    5732443 5732444 T/C     2       33,68   64,64
chr1    5732465 5732466 A/G     2       63,55   64,64
chr1    5732509 5732510 T/C     2       39,54   82,82
chr1    5732647 5732648 T/C     2       23,61   46,46
chr1    5963366 5963367 C/T     2       22,17   171,171
chr1    6336576 6336577 T/C     2       30,21   324,324
chr1    6395463 6395464 A/G     2       21,21   205,205
chr1    6516428 6516429 A/C     2       12,25   112,143
```

SNP example 1 2.0 Gb

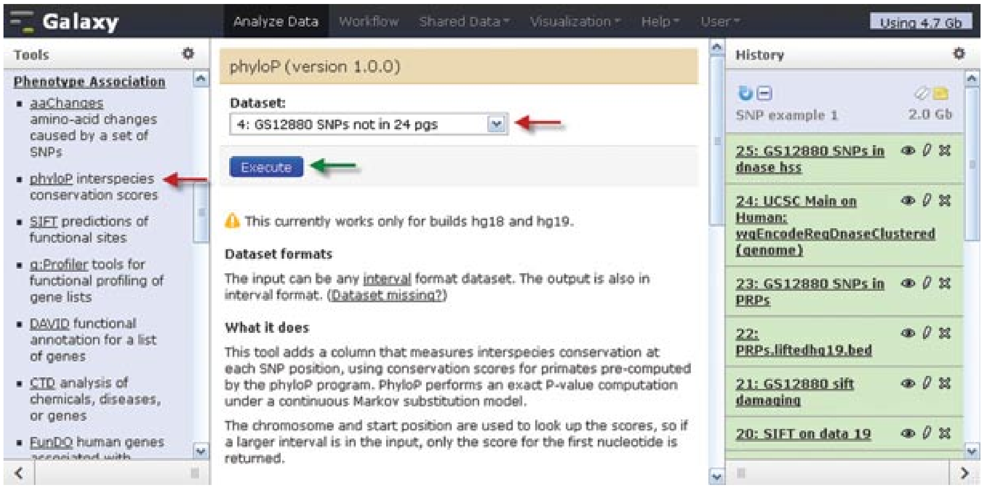
25: GS12880 SNPs in dnase hss
7,444 regions
format: interval, database: hg19
Info: Intersect on data 24 and data 4
display at UCSC main test
view in GeneTrack
display at Ensembl Current
display at RViewer main

1.Chrom	2.Start	3.End	4	5	6
chr11	5255743	5255744	C	1	1
chrX	100641248	100641249	G	1	1
chr1	565294	565295	G	1	2
chr1	826291	826292	T/G	2	1
chr1	986731	986732	A/G	2	2
chr1	1060815	1060816	A/G	2	1

24: UCSC Main on Human: wgEncodeRegDnaseClustered (genome)

23: GS12880 SNPs in PRPs
837 regions
format: interval, database: hg19
Info: Intersect on data 22 and data 4

Figure 15.2.25 SNPs in HSSs. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 4.7 Gb
Tools
Phenotype Association
aaChanges amino-acid changes caused by a set of SNPs
phyloP interspecies conservation scores
SIFT predictions of functional sites
g:Profiler tools for functional profiling of gene lists
DAVID functional annotation for a list of genes
CTD analysis of chemicals, diseases, or genes
FunDO human genes associated with
History
SNP example 1
2.0 Gb
25: GS12880 SNPs in dnase hss
24: UCSC Main on Human: wgEncodeRegDnaseClustered (genome)
23: GS12880 SNPs in PRPs
22: PRPs.liftedhg19.bed
21: GS12880 sift damaging
20: SIFT on data 19

phyloP (version 1.0.0)

Dataset:
4: GS12880 SNPs not in 24 pgs

Execute

This currently works only for builds hg18 and hg19.

Dataset formats

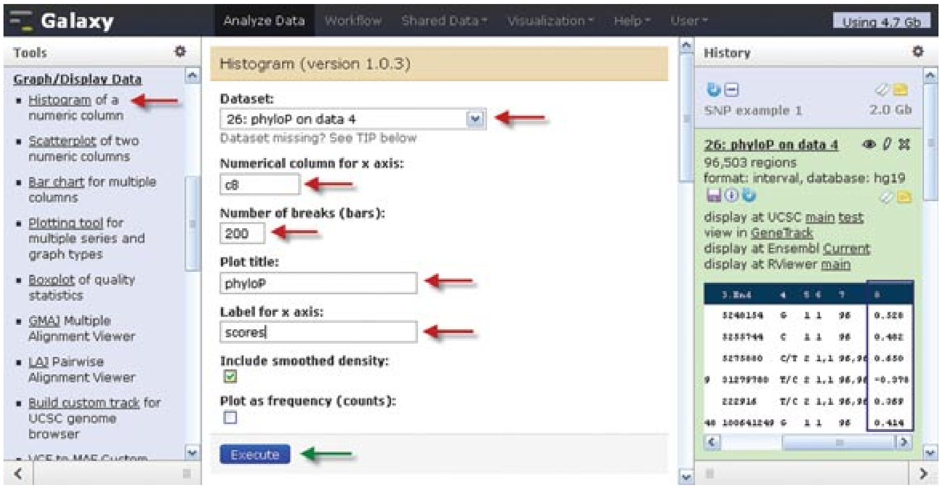
The input can be any interval format dataset. The output is also in interval format. (Dataset missing?)

What it does

This tool adds a column that measures interspecies conservation at each SNP position, using conservation scores for primates pre-computed by the phyloP program. PhyloP performs an exact P-value computation under a continuous Markov substitution model.

The chromosome and start position are used to look up the scores, so if a larger interval is in the input, only the score for the first nucleotide is returned.

Figure 15.2.26 PhyloP. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 4.7 Gb
Tools
Graph/Display Data
Histogram of a numeric column
Scatterplot of two numeric columns
Bar chart for multiple columns
Plotting tool for multiple series and graph types
Boxplot of quality statistics
GMAJ Multiple Alignment Viewer
LAJ Pairwise Alignment Viewer
Build custom track for UCSC genome browser
VCF to MAF Custom
History
SNP example 1
2.0 Gb

Histogram (version 1.0.3)

Dataset:
26: phyloP on data 4
Dataset missing? See TIP below

Numerical column for x axis:
c8

Number of breaks (bars):
200

Plot title:
phyloP

Label for x axis:
scores

Include smoothed density:
☑

Plot as frequency (counts):
☐

Execute

26: phyloP on data 4
96,503 regions
format: interval, database: hg19
display at UCSC main test
view in GeneTrack
display at Ensembl Current
display at RViewer main

```
   3.End     4   5 6   7     8
   5240154   G   1 1   96    0.520
   5255744   C   1 1   96    0.402
   5275880   C/T 2 1,1 96,96 0.650
9  31279780  T/C 2 1,1 96,96 -0.378
   222916    T/C 2 1,1 96,96 0.369
48 100641249 G   1 1   96    0.414
```

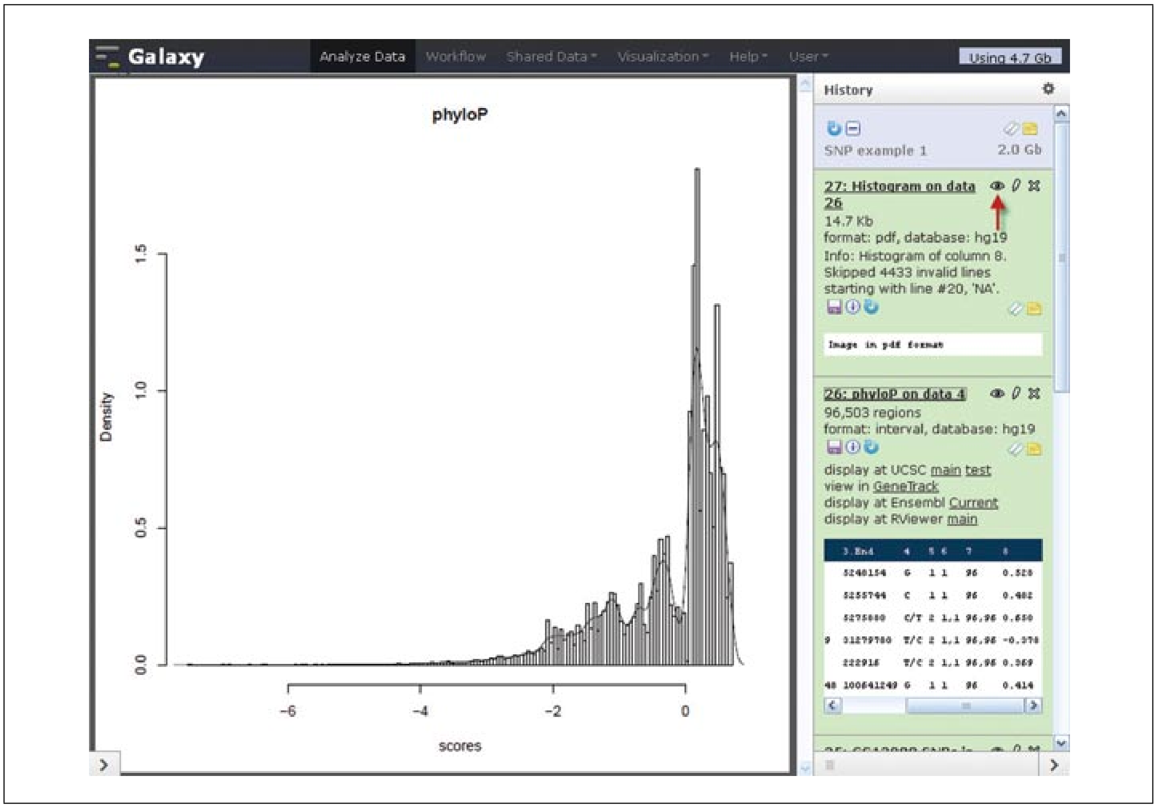
Figure 15.2.27 Distribution of phyloP scores. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

5. Within the group Regulation there is a large list of tracks (drop-down box), many from ENCODE. For this example we have chosen DNase Clusters, which has only one table available: "wgEncodeRegDnaseClustered." The default format for sending data to Galaxy is BED; this is typically the best format for tables with genomic positions, since it allows doing many genomic operations in Galaxy. Once the options are selected, click the "get output" button. For BED output, an intermediate page then appears with options for how to return the intervals. Just leave these at their defaults and click the "Send query to Galaxy" button.

   In general, when choosing a table, clicking on the "describe table schema" button (blue arrow) will tell you more about the table you are selecting, as well as provide a description of what data and methods were used for the track as a whole.

Intersecting with the HSSs

6. The searches we are conducting now are more or less independent, looking for SNPs that affect a variety of region types and functional mechanisms, so rather than



SNP example 1 2.0 Gb

27: Histogram on data 26
14.7 Kb
format: pdf, database: hg19
Info: Histogram of column 8.
Skipped 4433 invalid lines
starting with line #20, 'NA'.

Image in pdf format

26: phyloP on data 4
96,503 regions
format: interval, database: hg19
display at UCSC main test
view in GeneTrack
display at Ensembl Current
display at RViewer main

```
   3.End      4    5 6  7      8
   5240154    G    1 1  96     0.520
   5255744    C    1 1  96     0.482
   5275080    C/T  2 1,1 96,96 0.650
9  31279780   T/C  2 1,1 96,96 -0.378
   222916     T/C  2 1,1 96,96 0.369
48 100641249  G    1 1  96     0.414
```

Figure 15.2.28 Histogram. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

continuing from the PRP results, we will go back and start yet again with our input dataset from Part A. Once again we are using the Intersect tool, this time to find the SNPs within the DNase hypersensitive sites (Fig. 15.2.24). Note that there are a lot more regions in this set than there were in the PRPs (blue arrows).

SNPs in HSSs

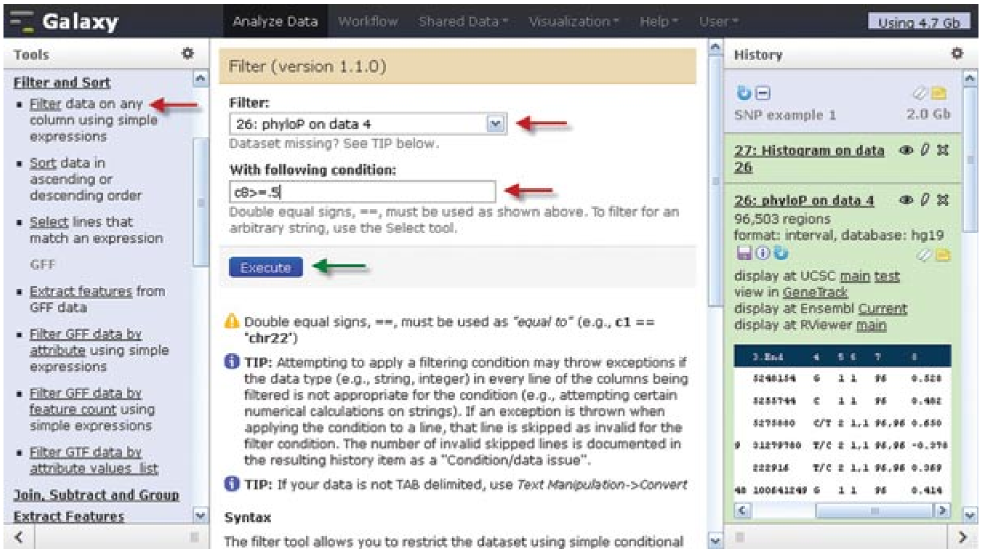
7. As expected, there are many more results from this intersection (blue arrows in Fig. 15.2.25). Again, two of our disease SNPs are found. One was also found via the PRPs; the other is a non-coding SNP missed by that method.

PhyloP

8. Next, we will look at conservation between species using the PhyloP tool, which looks up a precomputed phyloP score (Siepel et al., 2006) for each SNP position. This tool is located in the Phenotype Association section (Fig. 15.2.26). Select the same input dataset from Basic Protocol 1 that we have been using, and then click Execute.

Distribution of phyloP scores

9. To use the phyloP scores, it is helpful to know the distribution of those scores. For this we go to the Graph/Display Data section and select the Histogram tool (Fig. 15.2.27). In the history panel, we can see what column the scores are in, and a few of the values (blue box). The PhyloP tool appends a column to its input dataset, and as expected the last column is new and contains numbers. Select the dataset with the phyloP scores and indicate the column that has the score. Entering a large number of bars gives a finer resolution in the output. Adding a descriptive title and label are not required, but are helpful when coming back to this history later. Click Execute to generate the graph.



- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an expression
- Extract features from GFF data
- Filter GFF data by attribute using simple expressions
- Filter GFF data by feature count using simple expressions
- Filter GTF data by attribute values_list

Filter (version 1.1.0)

Filter:
26: phyloP on data 4
Dataset missing? See TIP below.

With following condition:
c8>=.5
Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Execute

Double equal signs, ==, must be used as *"equal to"* (e.g., **c1 == 'chr22'**)

TIP: Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a "Condition/data issue".

TIP: If your data is not TAB delimited, use *Text Manipulation->Convert*

Syntax

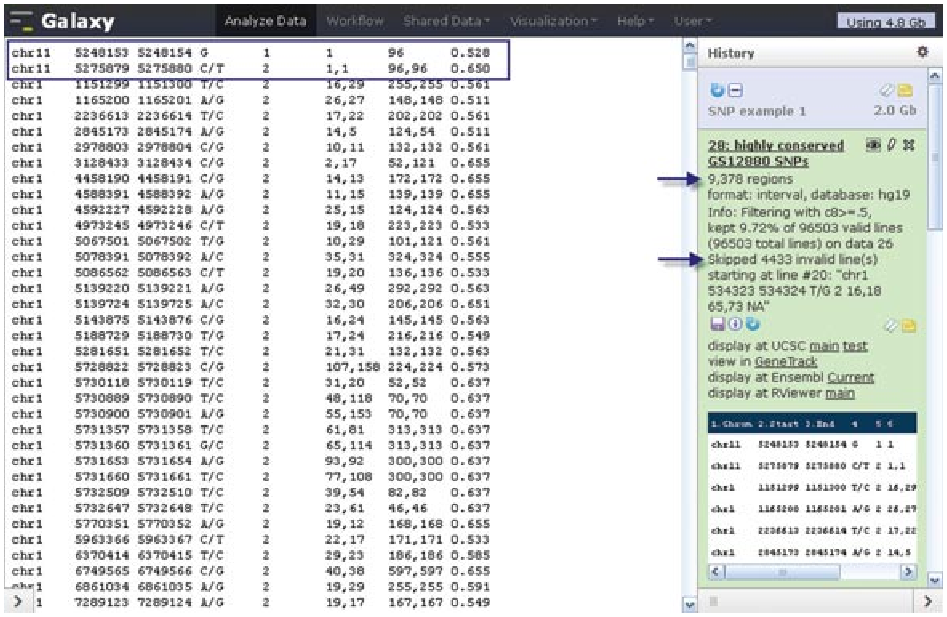
The filter tool allows you to restrict the dataset using simple conditional

27: Histogram on data 26

26: phyloP on data 4
96,503 regions
format: interval, database: hg19
display at UCSC main test
view in GeneTrack
display at Ensembl Current
display at RViewer main

```
   3.End      4   5 6   7      8
   5248154    G   1 1   96     0.528
   5255744    C   1 1   96     0.482
   5275880    C/T 2 1,1 96,96  0.650
9  31279780   T/C 2 1,1 96,96  -0.378
   222916     T/C 2 1,1 96,96  0.369
48 100841249  G   1 1   96     0.414
```

Figure 15.2.29 Filtering the SNPs based on phyloP score. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.



```
chr11   5248153 5248154 G     1       1        96      0.528
chr11   5275879 5275880 C/T   2       1,1      96,96   0.650
chr1    1151299 1151300 T/C   2       16,29    255,255 0.561
chr1    1165200 1165201 A/G   2       26,27    148,148 0.511
chr1    2236613 2236614 T/C   2       17,22    202,202 0.561
chr1    2845173 2845174 A/G   2       14,5     124,54  0.511
chr1    2978803 2978804 C/G   2       10,11    132,132 0.561
chr1    3128433 3128434 C/G   2       2,17     52,121  0.655
chr1    4458190 4458191 C/G   2       14,13    172,172 0.655
chr1    4588391 4588392 A/G   2       11,15    139,139 0.655
chr1    4592227 4592228 A/G   2       25,15    124,124 0.563
chr1    4973245 4973246 C/T   2       19,18    223,223 0.533
chr1    5067501 5067502 T/G   2       10,29    101,121 0.561
chr1    5078391 5078392 A/C   2       35,31    324,324 0.555
chr1    5086562 5086563 C/T   2       19,20    136,136 0.533
chr1    5139220 5139221 A/G   2       26,49    292,292 0.563
chr1    5139724 5139725 A/C   2       32,30    206,206 0.651
chr1    5143875 5143876 C/G   2       16,24    145,145 0.563
chr1    5188729 5188730 T/G   2       17,24    216,216 0.549
chr1    5281651 5281652 T/C   2       21,31    132,132 0.563
chr1    5728822 5728823 C/G   2       107,158  224,224 0.573
chr1    5730118 5730119 T/C   2       31,20    52,52   0.637
chr1    5730889 5730890 T/C   2       48,118   70,70   0.637
chr1    5730900 5730901 A/G   2       55,153   70,70   0.637
chr1    5731357 5731358 T/C   2       61,81    313,313 0.637
chr1    5731360 5731361 G/C   2       65,114   313,313 0.637
chr1    5731653 5731654 A/G   2       93,92    300,300 0.637
chr1    5731660 5731661 T/C   2       77,108   300,300 0.637
chr1    5732509 5732510 T/C   2       39,54    82,82   0.637
chr1    5732647 5732648 T/C   2       23,61    46,46   0.637
chr1    5770351 5770352 A/G   2       19,12    168,168 0.655
chr1    5963366 5963367 C/T   2       22,17    171,171 0.533
chr1    6370414 6370415 T/C   2       29,23    186,186 0.585
chr1    6749565 6749566 C/G   2       40,38    597,597 0.655
chr1    6861034 6861035 A/G   2       19,29    255,255 0.591
   1    7289123 7289124 A/G   2       19,17    167,167 0.549
```

28: highly conserved GS12880 SNPs
9,378 regions
format: interval, database: hg19
Info: Filtering with c8>=.5,
kept 9.72% of 96503 valid lines
(96503 total lines) on data 26
Skipped 4433 invalid line(s)
starting at line #20: "chr1
534323 534324 T/G 2 16,18
65,73 NA"
display at UCSC main test
view in GeneTrack
display at Ensembl Current
display at RViewer main

```
1.Chrom 2.Start 3.End   4   5 6
chr11   5248153 5248154 G   1 1
chr11   5275879 5275880 C/T 2 1,1
chr1    1151299 1151300 T/C 2 16,29
chr1    1165200 1165201 A/G 2 26,27
chr1    2236613 2236614 T/C 2 17,22
chr1    2845173 2845174 A/G 2 14,5
```

Figure 15.2.30 Highly conserved SNPs. See text for details. For color version of figure go to *http://www.currentprotocols.com/protocol/bi1502*.

Histogram

10. Clicking on the eye icon in the history displays the histogram in the center panel (Fig. 15.2.28). Here we have hidden the tool panel to make the graph larger. Any positive phyloP score indicates conservation, but there are many SNPs with scores between 0 and 1. To select the most highly conserved regions, a cutoff of 0.5 seems reasonable.

Filtering the SNPs based on phyloP score

11. In the Filter and Sort section, choose the Filter tool (Fig. 15.2.29). Select the dataset with the phyloP scores, and enter the condition `c8>=.5`, which means we want the rows where the score in column 8 is greater than or equal to 0.5.

Highly conserved SNPs

12. This approach has the highest number of results yet (first blue arrow in Fig. 15.2.30), and finds two of our disease SNPs, including one that has not been found up to this point (blue box). The message about skipped lines (second blue arrow) is not a problem; `NA` in the PhyloP tool's output means that the score was not available, and we want to skip those SNPs anyway.

    This is the end of the protocol. To narrow the results further, one could intersect the output from the various searches, either pairwise or all together.

GUIDELINES FOR UNDERSTANDING RESULTS

It is important to use caution when working with SNP data, as there are always some errors in the sequencing and SNP-calling processes. Even an error rate as low as 0.001% (as reported by Complete Genomics) can produce twenty to thirty errors among the two to three million SNPs typically found in an individual, and in practice error rates often seem much higher (e.g., ~1%), especially for some technologies.

Furthermore, genotype-to-phenotype associations are rarely straightforward. Phenotype association does not necessarily mean phenotype causation; other factors also influence the phenotype, such as other genes, lifestyle, and the environment. For complex diseases, associations determined using genome-wide association studies often reflect just an increase or decrease in risk. The SNP listed may not even be the true phenotype-affecting SNP, but just one that happens to be in linkage disequilibrium with the true SNP. Also, keep in mind that the predictions made for coding SNPs by programs such as PolyPhen-2 and SIFT are only predicting the effect of the SNP on the protein, not on the individual as a whole.

Conversely, many of these caveats apply not only to the SNPs found, but also to the so-called "benign" ones we eliminated. Just because an allele is reported in an apparently healthy individual does not guarantee that it never contributes to problems in other people, or in the same person later on, or even that the person tested actually has that allele.

In summary, results from these protocols should be considered only as rough hints or clues to suggest candidate SNPs potentially worthy of further investigation.

COMMENTARY

Background Information

Data on human genetic variation (e.g., from genotyping, whole-genome sequencing, genome-wide association studies, and mapping of epigenetic features) have the potential to advance medical applications at an unprecedented rate, but only if bioinformatics solutions are provided to bench scientists and clinicians to enable them to find biological insights in the flood of data.

Currently, most software tools, databases, and other resources for these analyses are either not publicly available, or are scattered among many sites and require considerable effort and expertise to install and operate. In several cases, projects to produce open-source resources have published papers, but the resulting tools have not yet been made available.

Galaxy serves as an effective platform for assembling a wide variety of tools into an integrated, easily accessible system that offers one-stop convenience, no need for users to install specialized software, and most importantly, a consistent interface that does not require programming expertise to use. The Phenotype Association section provides access to many tools that are useful for analyzing biomedical applications of genetic variation data, only a few of which are discussed here.

Critical Parameters and Troubleshooting

Be sure to read the description below each tool's input form, especially with regard to the input and output formats.

For each dataset, Galaxy keeps track of the data format and the genome assembly (for datasets containing genomic positions). If a tool takes multiple datasets as input, Galaxy will generally require that they be based on the same assembly (except for tools that are specifically designed to work across species or builds). It will also only allow you to select input datasets that are in a format compatible with the tool. Therefore, if the dataset you want to use as input for a tool is in your history but is not appearing in the list to select from, check the assembly (also known as "database") and format of the dataset to make sure they are correct. There are tools for converting between formats when needed.

Note that if Galaxy ever guesses the type of a dataset incorrectly, you can fix that by clicking on its pencil icon in the history and using the "Change data type" form at the bottom of the page that appears. This is different from actually converting the data from one format to another.

For further assistance, please consult the Critical Parameters and Troubleshooting section of *UNIT 10.5*.

Acknowledgments

We would like to thank the Galaxy Team for developing this innovative and powerful community resource. Funding for our work on assembling and documenting the Phenotype Association tools was provided by NIH grant UL1 RR033184-01 to the Penn State Clinical and Translational Science Institute. This project is funded, in part, under a grant with the Pennsylvania Department of Health using Tobacco CURE Funds. The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions.

Literature Cited

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. 2010. A method and server for predicting damaging missense mutations. *Nat. Methods* 7:248-249.

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., and Taylor, J. 2010. Galaxy: A web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.* 89:19.10.1-19.10.21.

Davis, A.P., Murphy, C.G., Saraceni-Richards, C.A., Rosenstein, M.C., Wiegers, T.C., and Mattingly, C.J. 2009. Comparative Toxicogenomics Database: A knowledgebase and discovery tool for chemical.gene.disease networks. *Nucleic Acids Res.* 37:D786-D792.

Drmanac, R., Sparks, A.B., Callow, J.M., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G, Dahl, F., Fernandez, A., Staker, B., Pant, K.P., Baccash, J., Borcherding, A.P., Brownley, A., Cedeno, R., Chen, L., Chernikoff, D., Cheung, A., Chirita, R., Curson, B., Ebert, J.C., Hacker, C.R., Hartlage, R., Hauser, B., Huang, S., Jiang, Y., Karpinchyk, V., Koenig, M., Kong, C., Landers, T., Le, C., Liu, J., McBride, C.E., Morenzoni, M., Morey, R.E., Mutch, K., Perazich, H., Perry, K., Peters, B.A., Peterson, J., Pethiyagoda, C.L., Pothuraju, K., Richter, C., Rosenbaum, A.M., Roy, S., Shafto, J., Sharanhovich, U., Shannon, K.W., Sheppy, C.G., Sun, M., Thakuria, J.V., Tran, A., Vu, D., Zaranek, A.W., Wu, X., Drmanac, S., Oliphant, A.R., Banyai, W.C., Martin, B., Ballinger, D.G., Church, G.M., and Reid, C.A. 2009. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327:78-81.

Ferretti, V., Poitras, C., Bergeron, D., Coulombe, B., Robert, F., and Blanchette, M. 2007. PReMod: A database of genome-wide mammalian *cis*-regulatory module predictions. *Nucleic Acids Res.* 35:D122-D126.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W.J., and Nekrutenko, A. 2005. Galaxy: A platform for interactive large-scale genome analysis. *Genome Res.* 15:1451-1455.

Giardine, B., Riemer, C., Hefferon, T., Thomas, D., Hsu, F., Zielenski, J., Sang, Y., Elnitski, L., Cutting, G., Trumbower, H., Kern, A., Kuhn, R., Patrinos, G.P., Hughes, J., Higgs, D., Chui, D., Scriver, C., Phommarinh, M., Patnaik, S.K., Blumenfeld, O., Gottlieb, B., Vihinen, M., Väliaho, J., Kent, J., Miller, W., and Hardison, R.C. 2007. PhenCode: Connecting ENCODE data with mutations and phenotype. *Hum. Mutat.* 28:554-562.

Goecks, J., Nekrutenko, A., Taylor, J.; Galaxy Team. 2010. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11:R86.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4:44-57.

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.

Kumar, P., Henikoff, S., and Ng, P.C. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* 4:1073-1081.

Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. 2007. g:Profiler: A web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* 35:W193-W200.

Seal, R.L., Gordon, S.M., Lush, M.J., Wright, M.W., and Bruford, E.A. 2011. genenames.org: The HGNC resources in 2011. *Nucleic Acids Res.* 39:D514-519.

Siepel, A., Pollard, K.S., and Haussler, D. 2006. New methods for detecting lineage-specific selection. *In* Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006), pp. 190-205, Venice, Italy.

Taylor, J., Tyekucheva, S., King, D.C., Hardison, R.C., Miller, W., and Chiaromonte, F. 2006. ESPERR: Learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res.* 16:1596-1604.

Internet Resources

http://galaxyproject.org

The main public instance of Galaxy.

http://phencode.bx.psu.edu

A collection of human phenotype-associated SNPs from Locus-Specific Databases.

http://www.bx.psu.edu/miller_lab/docs/galaxy_phen_assoc/tutorial/

A version of this tutorial in HTML format.

http://genome.ucsc.edu/FAQ/FAQformat.html

Descriptions of file formats used by the UCSC Table Browser.

Supplementary File

http://www.currentprotocols.com/protocol/bi1502

This is an alternate URL to access the file "test.masterVar.gz" cited in Basic Protocol 1, Necessary Resources, Files on page 15.2.3.