

Galaxy

Info: report bugs | wiki | screencasts | blog | logged in as auni@psu.edu: manage | logout

Galaxy Genomics Penn State

Tools

- Get Data
- Get ENCODE Data
- ENCODE Tools
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
  - Intersect the intervals of two queries
  - Subtract the intervals of two queries
  - Merge the overlapping intervals of a query
  - Concatenate two queries into one query
  - Base Coverage of all intervals
  - Coverage of a set of intervals on second set of intervals
  - Complement intervals of a query
  - Cluster the intervals of a query
  - Join the intervals of two queries side-by-side
  - Get flanks returns flanking regions for every gene
- Statistics
- Graph/Display Data
- Evolution: HYPHY
- EMBOSS

Join

Join: 1: RefSeq coding exons

First query

with: 2: SNPs

Second query

with min overlap: 1

(bp)

Return: Only records that are joined (INNER JOIN)

Execute

**TIP:** If your query does not appear in the pulldown menu -> 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 specifies the minimum overlap between intervals that allows them to be joined.
- Return only records that are joined returns only the records of the first query that join to a record in the second query. This is analogous to an INNER JOIN.
- Return all records of first query (fill null with ".") returns all intervals of the first query, and any intervals that do not join an interval from the second query are filled in with a period(.). This is analogous to a LEFT JOIN.
- Return all records of second query (fill null with ".") returns all intervals of the second query, and any intervals that do not join an interval from the first query are filled in with a period(.). Note that this may produce an invalid interval file, since a period(.) is not a valid chr chrom, start, end or strand.
- Return all records of both queries (fill nulls with ".") returns all records from both queries, and fills on either the right or left with periods. Note that this may produce an invalid interval file, since a period(.) is not a valid chrom, start, end or strand.

Example

If First query is:

```
chr1 10 100 Query1.1
chr1 100 1000 Query1.2
chr1 1100 1250 Query1.3
```

and Second query is:

```
chr1 20 50 Query2.1
chr1 50 100 Query2.2
chr1 2500 3000 Query2.3
```

The four return options will generate:

- Return only records that are joined:

History (options)

refresh | collapse all

7: Human/Chimp/Dog alignments for Exons from Step 6

957 sequences, format: fasta, database: hg18

Info: 319 regions were extracted successfully.

save

>hg18\_chr8(+)1:18301793-18302666  
 ATGGACATTTGGACATATTTTTCAGAAATTTGGCTATAGAACTCT7  
 >panTro2  
 ATGGACATTTGGACATATTTTTCAGAAATTTGGCTATAGAACTCT7  
 >canFam2

6: Coordinates of Exons with at least 5 SNPs

5: IDs Exons with at least 5 SNPs

4: Number of SNPs per Exon

3: Join of Exons and SNPs

2: SNPs

1: RefSeq coding exons

259,080 regions, format: bed, database: hg18

Info: Uploaded from UCSC  
 save | display at UCSC mail

1 2 3 4

```
chr1 67052400 67052451 NM_024763_cds_0_0_chr1_259_080 regions, format: bed, database: hg18
chr1 67066631 67067088 NM_024763_cds_1_0_chr1_259_080 regions, format: bed, database: hg18
chr1 67065090 67065317 NM_024763_cds_2_0_chr1_259_080 regions, format: bed, database: hg18
chr1 67066082 67066481 NM_024763_cds_3_0_chr1_259_080 regions, format: bed, database: hg18
chr1 67071895 67071977 NM_024763_cds_4_0_chr1_259_080 regions, format: bed, database: hg18
chr1 67072261 67072419 NM_024763_cds_5_0_chr1_259_080 regions, format: bed, database: hg18
```

## Tools (left)

Tools (currently over 100) are organized in categories. Clicking on a category expands and shows available tools. For example, here the *Operate on Genomic Intervals* category is expanded.

## Interfaces (center)

If you click a tool, its interface will appear here. For example, clicking on *Join* (left pane) will bring up its interface (shown in the center pane). If this case we are joining RefSeq exons with SNPs. The names of datasets in the two dropdowns ("1:RefSeq coding exons" and "2:SNPs") correspond to items listed in the right (History) pane. The tool interface also provides help information.

## History (right)

This is where all your data and analyses results go. The output of every tool creates another history item, so that your original data is never changed. For example, we started with two datasets imported from UCSC: "1:RefSeq coding exons" and "2:SNPs". Joining the two datasets created a new History Item called "3:Join on Exons and SNPs". And so on... Thus the History serves as a complete record of every analyses. Clicking "options" (just above the right pane) allows saving and sharing (!) of histories.

# DON'T DO GENOMICS ALONE!

Introducing GALAXY, a revolutionary tool for collaborative genomic research!

<http://g2.bx.psu.edu>

(since you cannot click the paper - flip it!)

See what GALAXY can do for you:

### Biologists

Use our public site (<http://g2.bx.psu.edu>) to access popular sources of data like the UCSC Table Browser. Run analyses right on the spot using a variety of integrated tools. Your results are always available and can be easily shared with others.

### Software Developers

Galaxy is an easy-to-use, open-source, scalable framework for tool and data integration. Stop wasting time writing interfaces and get your tools used by biologists! GALAXY includes everything you need to get started, so download and start integrating!



Plecion fulvicollis



## Why?

You are an experimental biologist. You keep watching databases fill with more and more data. You keep thinking: *"even if I knew how to use Excel as a pro, it would probably not load 12,435,654 SNPs"*. So how do you perform analyses without calling somebody on the Computer Science side of campus? Suppose you want to find human exons with the highest SNP density. There is no straightforward way of doing it without learning programming first. And this is why:

### Databases are not analyses tools

Databases are where you get the data. Browsers are where you visualize the results. For a bench biologist there is not much in between besides spreadsheets or Perl scripting. GALAXY fills this gap by providing a tool-rich analysis medium.

### No tools for essential datatypes

Some datatypes generated by high throughput genomics are so new that there are no tools to analyze them. For example, how do you extract sequences of coding exons from the latest 28-way alignments of vertebrate genomes or analyze genome-wide disease association data? With GALAXY!

### Is genomics reproducible?

The Methods section of too many papers sound like *"the data were analyzed using a collection of in-house scripts"*. How do you repeat such an analysis? GALAXY saves every step of your analysis and allows you to share these workflows with others.

### Too many tools

*Bioinformatics* has published 166 Application Notes since January 2007. How does one know which tool to use? GALAXY integrates a multitude of different tools by giving them the same "look and feel" and linking them to data warehouses.



## How?

So is it all hot air or can you actually do something with GALAXY that you cannot do otherwise? Let's try this:

*find all human exons that contains at least 5 non-synonymous (amino acid changing) SNPs and extract their alignments with chimp and dog*

This analysis requires 7 steps outlined below but can be completed easily and quickly with GALAXY (in fact, 90% of that time is spent downloading SNP data from the UCSC Table Browser). These steps correspond exactly to the contents of GALAXY's left pane on the opposite side of this brochure. You can watch a QuickTime screencast detailing this analysis at <http://g2.bx.psu.edu>.

### Download exon and SNP coordinates

Here we just grab coordinates of all coding exons and SNPs from the UCSC Table Browser (the efforts of the UCSC Team were crucial for this integration). This creates History items #1 and #2 (right pane of Galaxy's interface on the opposite side)

### Join exons with SNPs

This will find all exons that contain SNPs. This operation is performed using the *Join* function and creates History item #3. *Join* is a part of our large suite of tools for manipulation of genomic intervals that include intersection, coverage, and other functions.

### Compute number of SNP per Exon

Here we first compute the number of SNPs per exon with the *Count* function (History item #4) and then use the *Filter* tool to find exons with at least 5 SNPs (History item #5).

### Extract human/chimp/dog alignments

Finally, we extract human/chimp/dog alignments corresponding to the exons from the previous step. These alignments are trimmed per exon boundaries. If multiple alignment blocks overlap an exon - they are stitched together (History item #6).



## See it!

Still not convinced? Go to <http://g2.bx.psu.edu>. There you will see a link to screencasts (you will need a free QuickTime player available from Apple website to view them). Screencasts highlight all aspects of GALAXY's functionality and also show examples of very complex analyses.



## Share!

Best of all, GALAXY's history system provides a complete analyses record that can be shared. Every history is an analysis workflow, which can be used to reproduce the entire analyses.

### History is an analysis record

Look at GALAXY's interface on the opposite side. Every step of the analyses is recorded. You can have any number of histories saved. This way you can go back to your analyses anytime.

### Share your analyses

Alice works at Penn State, while Bob suffers from the terrible San Diego climate. Alice wants Bob to see her analyses. Alice clicks the *"share"* link and enters Bob's e-mail address. Now Alice's history is visible to Bob (see *"Sharing history"* screencast).

### Now your results are reproducible!

When publishing results, replace *"the data were analyzed using a collection of in-house scripts"* with a URL pointing to Galaxy's history. Your reviewers will have no further questions. That's reproducible genomics!