

Galaxy 'RNA' workshop

Delphine POTIER, Rekin's JANKY &
Stein AERTS (LCB - KU Leuven)

Test case:

Drosophila retinal determination

Sequencing platform:

HiSeq 2000 (illumina)

Tools

- [Get Data](#)
- [Send Data](#)
- [ENCODE Tools](#)
- [Lift-Over](#)
- [Text Manipulation](#)
- [Convert Formats](#)
- [FASTA manipulation](#)
- [Filter and Sort](#)
- [Join, Subtract and Group](#)
- [Extract Features](#)
- [Fetch Sequences](#)
- [Fetch Alignments](#)
- [Get Genomic Scores](#)
- [Operate on Genomic Intervals](#)
- [Statistics](#)
- [Graph/Display Data](#)
- [Regional Variation](#)
- [Multiple regression](#)
- [Multivariate Analysis](#)
- [Evolution](#)
- [Motif Tools](#)
- [Multiple Alignments](#)
- [Metagenomic analyses](#)
- [Phenotype Association](#)
- [Genome Diversity](#)
- [EMBOSS](#)

- [NGS TOOLBOX BETA](#)
- [NGS: QC and manipulation](#)
- [NGS: Mapping](#)
- [NGS: SAM Tools](#)
- [NGS: GATK Tools \(beta\)](#)
- [NGS: Indel Analysis](#)

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

History of your data pre-processing & analysis

Galaxy available tools

First step will be to
load the data

History



Unnamed history



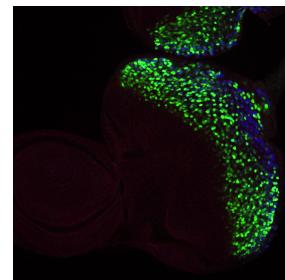
0 bytes

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

RNA-seq data:

Eye development, glass mutant VS wild type eye/antennal imaginal discs

- The transcription factor ‘*glass*’ is required for normal development of photoreceptor cells in eye imaginal disc during *Drosophila* retinal determination (larval stage 3).
- We will study the differential expression in *gl* mutants compared to wild type eye/antennal discs. To this aim will provide you single-end HiSeq2000 RNA-seq data from:
 - Wild type eye/antennal disc
(2 replicates)
 - Glass mutant eye/antennal disc
(2 replicates)



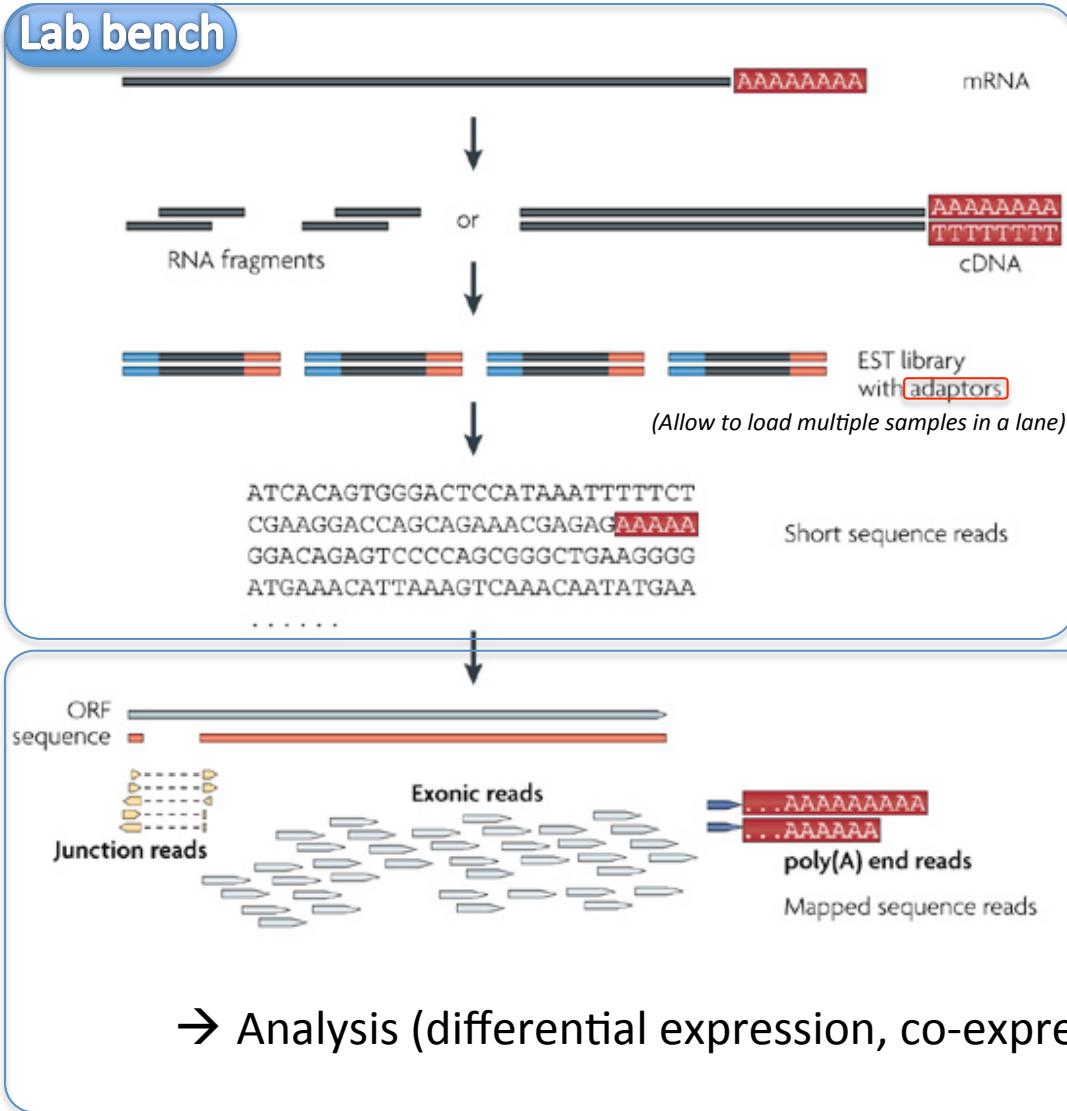
* You will work on a partial dataset to reduce calculation time

Results for the entire dataset are available here:

<http://galaxy.bits.vib.be/u/delphine-potier/h/ngs-workshop>

Bioinformatics treatment linked to lab work

Lab bench



Computer

- Demultiplexing (usually done by the platform)
- Adapter clipping (Fast-x clipper)
- Checking reads quality (FastQC)
- Mapping (tophat)
- Visualization
- Analysis (differential expression, co-expression, cis-regulation etc...)

Modified from Zhong Wang et al., RNA-Seq: a revolutionary tool for transcriptomics, Nat Rev Genet (2009)

Understand .fastq reads

```
@HWI-ST571_69:4:1101:11162:2311:ACAGTG/1  
ATTTATTCTAATTGTTATTATGTTTATTCTT  
+HWI-ST571_69:4:1101:11162:2311:ACAGTG/1  
Gggggfggggegggggggggggggggggggggggggg
```

*Illumina sequence identifier
Sequence*

Encoded sequencing quality score for each base

Illumina sequence identifier:

- **HWI-ST571_69** the unique instrument name
- **4** flowcell lane
- **1101** tile number within the flowcell lane
- **11162** 'x'-coordinate of the cluster within the tile
- **2311** 'y'-coordinate of the cluster within the tile
- **ACAGTG** index number for a multiplexed sample (0 for no indexing)
- **/1** the member of a pair, /1 or /2 (*paired-end or mate-pair reads only*)

From raw data to pre-processed data

- Demultiplexing

→ Gives the .fastq files from which you will start

- 24 different adapters allow to load 24 samples in one lane (~120-150 million base sequenced/lane), the more samples loaded the less base sequenced/sample
- Unused adapter shouldn't be found (quality check)

- Adapter cleaning (Fastx-clipper)

- After demultiplexing, primers and/or adapter can still be part of the reads, you have to clip them to keep only RNA sequence. In our case:

- Multiplex PCR primer

CATTGTTGGCTATTAATTGAACAAATGAGATCGGAAGAGCACACGTCT

- Sample specific True-seq adapters

GATCGGAAGAGCACACGTCTGAACTCCAGTCACXXXXXXATCTCGTATGCCGTCTTGCTTG

```
cat MyFile.fastq /software/fastx_toolkit-0.0.13/src/fastx_clipper/fastx_clipper -Q33 -a  
CATTGTTGGCTATTAATTGAACAAATGAGATCGGAAGAGCACACGTCT -M15 -n -v -l 20 | /software/fastx_toolkit-0.0.13/src/fastx_clipper/  
fastx_clipper GATCGGAAGAGCACACGTCTGAACTCCAGTCACXXXXXXATCTCGTATGCCGTCTTGCTTG -M15 -n -v -l 20 >MyCleanFile.fastq
```



Keep in mind that the options that we are using today fits to our data (e.g.:RNA –seq, single end sequencing...)
You should be careful to all options (it can dramatically change your results)



*Note that all these steps can be done in an automatic pipeline with command lines in a shell.
(Quicker for numerous and big datasets)*

From raw data to pre-processed data

- **Reads quality checking (FastQC)**
 - A report (containing graphs representing basic statistics, sequence quality, GC content, N content, Overrepresented sequences etc...) will help to know if each step went well.
 - Example of FastQC reports:
 - Good illumina dataset: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc/fastqc_report.html
 - Poor illumina dataset: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc/fastqc_report.html

```
/software/FastQC/fastqc MyCleanFile.fastq
```



- **Mapping (TopHat)**
 - TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

```
/software/tophat/bin/tophat -p 20 -o MyMapped_file.tophat /bowtie_index/dmel-all-chromosome-r5.45 MyCleanFile.fastq
```

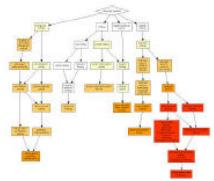
Data analysis

Getting to differential expression

- Cuffdiff
 - find significant changes in transcript/gene expression starting from BAM files and a GTF file using FPKM (To test whether an observed difference in a gene's expression is significant, Cuffdiff compares the log ratio of gene's expression in two conditions against the log of one.)

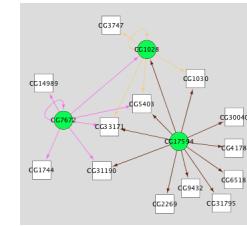
```
cuffdiff --no-update-check -q -p 4 -c 10 --FDR 0.050000 -N --labels wild_types,mutants transcript.gtf  
Dmel_rep1.bam,Dmel_rep2.bam mut_rep1.bam,mut_rep2.bam
```

- DESeq, EDASEQ, EdgeR
 - Other differential expression calculation tools available in galaxy and/or R



Data analysis

complementary analysis



- **Gorilla** - <http://cbl-gorilla.cs.technion.ac.il/>
 - tool for identifying and visualizing enriched GO terms in (ranked) lists of genes Available for *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Homo sapiens*, *Mus musculus* and *Rattus norvegicus*.
- **i-regulon** - cytoscape plugin
 - computational framework uniting *cis*-regulatory sequence analysis with network biology.
 - iRegulon analyzes a *human*, *mouse*, or *fly* gene network or gene signature and identifies enriched *cis*-regulatory motifs.
 - For each candidate TF, iRegulon determines the optimal sub-network as direct targetome of the TF.

Workshop in details

1-load data

(available on <http://ngsworkshop.aertslab.org/> or in the library)

Choose your format, file to download and related genome (here we are working with RNA-seq data from D. melanogaster)
Or select datasets from “shared data → data libraries → Galaxy RNA workshop 20 sep 2012” we prepared for you

The screenshot shows the Galaxy / BITS web interface. The top navigation bar includes 'Galaxy / BITS', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A green arrow points from the explanatory text above to the 'Shared Data' dropdown. Two red arrows point to the 'Get Data' link in the left sidebar and the 'Upload File from your computer and via FTP' section in the main pane.

Tools

- search tools
- Get Data**
- Upload File from your computer and via FTP.
 - UCSC Main table browser
 - UCSC Archaea table browser
 - BioMart Central server
 - CBI Rice Mart rice mart
 - GrameneMart Central server
 - modENCODE fly server
 - Flymine server
 - modENCODE modMine server
 - Ratmine server
 - YeastMine server
 - metabolicMine server
 - modENCODE worm server
 - WormBase server
 - EuPathDB server
 - EncodeDB at NHGRI
 - EpiGRAPH server
- Lift-Over
- Text Manipulation
 - Add column to an existing dataset

Upload File (version 1.1.3)

File Format: fastq

Which format? See help below

File: Choose File no file selected

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://ngsworkshop.aertslab.org/chrX_9000000_9400000_Dmel1.fastq`

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

Files uploaded via FTP:

File	Size	Date
Your FTP upload directory contains no files.		

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at `galaxy.bits.vib.be` using your Galaxy credentials (email address and password).

Convert spaces to tabs: Yes

Use this option if you are entering intervals by hand.

Genome: D. melanogaster Apr. 2006 (BDGP R5)

Execute

History

Unnamed history 0 bytes

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

1-load data

(available on <http://ngsworkshop.aertslab.org/> or in the library)

The loaded data will appear in your history

The screenshot shows the Galaxy / BITS web interface. On the left, the 'Tools' sidebar lists various data sources like UCSC Main table browser, BioMart Central server, and modENCODE fly server. The main area displays the 'Upload File (version 1.1.3)' tool. The 'File Format' dropdown is set to 'fastq'. The 'File' section shows a 'Choose File' button with 'no file selected'. A note below says: '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)'. The 'URL/Text' field contains the URL: http://ngsworkshop.aertslab.org/chrX_9000000_9400000_Dmel1.fastq. Below it is a note: 'Here you may specify a list of URLs (one per line) or paste the contents of a file.' The 'Files uploaded via FTP:' section shows a table with columns 'File', 'Size', and 'Date'. A note says: 'Your FTP upload directory contains no files.' The 'Convert spaces to tabs:' checkbox is unchecked. The 'Genome:' dropdown is set to 'D. melanogaster Apr. 2006 (BDGP R5)'. At the bottom is a blue 'Execute' button. On the right, the 'History' panel shows an 'Unnamed history' entry with a size of 6.4 Mb. The entry details are: '1: http://ngsworkshop.aertslab.org/chrX_9000000_9400000_Dmel1.fastq'. Three red boxes with arrows point to the history entry: one labeled 'Edit data attributes' points to the gear icon; one labeled 'Look at data' points to the eye icon; and one labeled 'Remove data' points to the trash can icon.

Galaxy / BITS

Analyze Data Workflow Shared Data Visualization Help User

Using 64%

Tools

search tools

Get Data

- Upload File from your computer and via FTP.
- UCSC Main table browser
- UCSC Archaea table browser
- BioMart Central server
- CBI Rice Mart rice mart
- GrameneMart Central server
- modENCODE fly server
- Flymine server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- metabolicMine server
- modENCODE worm server
- WormBase server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server

Lift-Over

Text Manipulation

- Add column to an existing dataset

Upload File (version 1.1.3)

File Format: fastq

Which format? See help below

File: Choose File no file selected

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://ngsworkshop.aertslab.org/chrX_9000000_9400000_Dmel1.fastq

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

Files uploaded via FTP:

File	Size	Date
Your FTP upload directory contains no files.		

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at [galaxy.bits.vib.be](#) using your Galaxy credentials (email address and password).

Convert spaces to tabs:

Yes

Use this option if you are entering intervals by hand.

Genome:

D. melanogaster Apr. 2006 (BDGP R5)

Edit data attributes

Look at data

Remove data

History

Using 64%

Unnamed history 6.4 Mb

1: http://ngsworkshop.aertslab.org/chrX_9000000_9400000_Dmel1.fastq

LCB

1-load data

(available on <http://ngsworkshop.aertslab.org/> or in the library)

The loaded data will appear in your history, you should start with 4 datasets 2 wild types and 2 mutants

The screenshot shows the Galaxy / BITS web interface. On the left, the 'Tools' sidebar lists various data sources like UCSC Main table browser, BioMart Central server, and modENCODE fly server. The main area displays the 'Upload File (version 1.1.3)' tool. The 'File Format' dropdown is set to 'fastq'. The 'File' section shows 'no file selected' and a URL input field containing 'http://ngsworkshop.aertslab.org/chrX_9000000_9400000_Dmel1.fastq'. The 'History' panel on the right shows four datasets added from the URL: 'Unnamed history' (36.1 Mb), '4:' (http://ngsworkshop.aertslab.org/chrX_9000000_9400000_glass-mutant_r3.fastq), '3:' (http://ngsworkshop.aertslab.org/chrX_9000000_9400000_glass-mutant_r2.fastq), and '2:' (http://ngsworkshop.aertslab.org/chrX_9000000_9400000_Dmel2.fastq). A red arrow points to the fourth dataset in the history list.

Galaxy / BITS

Analyze Data Workflow Shared Data Visualization Help User Using 64%

Tools

search tools

Get Data

- Upload File from your computer and via FTP.
- UCSC Main table browser
- UCSC Archaea table browser
- BioMart Central server
- CBI Rice Mart rice mart
- GrameneMart Central server
- modENCODE fly server
- Flymine server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- metabolicMine server
- modENCODE worm server
- WormBase server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server

Lift-Over

Text Manipulation

- Add column to an existing dataset

Upload File (version 1.1.3)

File Format: fastq

Which format? See help below

File: Choose File no file selected

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://ngsworkshop.aertslab.org/chrX_9000000_9400000_Dmel1.fastq

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

Files uploaded via FTP:

File	Size	Date
Your FTP upload directory contains no files.		

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at [galaxy.bits.vib.be](#) using your Galaxy credentials (email address and password).

Convert spaces to tabs:

Yes

Use this option if you are entering intervals by hand.

Genome:

D. melanogaster Apr. 2006 (BDGP R5)

Execute

History

Unnamed history 36.1 Mb

4: http://ngsworkshop.aertslab.org/chrX_9000000_9400000_glass-mutant_r3.fastq

3: http://ngsworkshop.aertslab.org/chrX_9000000_9400000_glass-mutant_r2.fastq

2: http://ngsworkshop.aertslab.org/chrX_9000000_9400000_Dmel2.fastq

1: http://ngsworkshop.aertslab.org/chrX_9000000_9400000_Dmel1.fastq

LCB

2- FASTQ Groomer

Use the FASTQ Groomer to get the data in a galaxy-tools readable format

3- Clip adapter sequences

FASTX-clipper

Adapters/primer can still be part of some reads, use FastX-clipper to remove them with the following options.

The screenshot shows the Galaxy / BITS web interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A progress indicator 'Using 64%' is shown in the top right. The left sidebar lists various tools under 'Tools' categories: FASTQ processing (FASTQ de-interlacer, Manipulate FASTQ, FASTQ to FASTA, FASTQ to Tabular, Tabular to FASTQ), FASTX-TOOLKIT FOR FASTQ DATA (Quality format converter, Compute quality statistics, Draw quality score boxplot, Draw nucleotides distribution chart), and sequencing artifacts (FASTQ to FASTA, Filter by quality, Remove sequencing artifacts, Barcode Splitter). A red arrow points to the 'Clip adapter sequences' option. The main panel displays the 'Clip (version 1.0.1)' tool configuration. It includes fields for 'Library to clip' (set to '6: FASTQ Groomer on data 2'), 'Minimum sequence length (after clipping, sequences shorter than this length will be discarded)' (set to '15'), 'Source' (set to 'Enter custom sequence'), 'Enter custom clipping sequence' (containing 'CATTGGTTGGCTATTAATTG'), 'enter non-zero value to keep the adapter sequence and x bases that follow it' (set to '0'), 'Discard sequences with unknown (N) bases' (set to 'No'), and 'Output options' (set to 'Output both clipped and non-clipped sequences'). A large blue 'Execute' button is at the bottom of this panel. Below the main panel, a 'What it does' section describes the tool's function: 'This tool clips adapters from the 3'-end of the sequences in a FASTA/FASTQ file.' A 'Clipping Illustration' section shows a diagram of a sequence being clipped, and a 'Clipping Example' section shows two sequence boxes. The right side of the interface features a 'History' panel listing nine previous runs, each with a preview icon, a delete icon, and a link to the tool and input data used. The history items are: 9: Clip on data 5, 8: FASTQ Groomer on data 4, 7: FASTQ Groomer on data 3, 6: FASTQ Groomer on data 2, 5: FASTQ Groomer on data 1, 4: http://ngsworkshop.aertslab.org/chrX_9000000_9400000_glass-mutant_r3.fastq, 3: http://ngsworkshop.aertslab.org/chrX_9000000_9400000_glass-mutant_r2.fastq, 2: http://ngsworkshop.aertslab.org/chrX_9000000_9400000_Dmel2.fastq, and 1: http://ngsworkshop.aertslab.org/chrX_9000000_9400000_Dmel1.fastq. A red arrow also points to the 'Execute' button.

3- Clip adapter sequences

FASTX-clipper

Data should be clipped 2 times, first with the Multiplex PCR primer and second with the specific True-seq adapter

The screenshot shows the Galaxy / BITS interface with the Clip tool (version 1.0.1) selected. The left sidebar lists various tools under the 'Tools' category, including FASTQ de-interlacer, Manipulate FASTQ reads, FASTQ to FASTA converter, FASTQ to Tabular converter, Tabular to FASTQ converter, FASTX-TOOLKIT FOR FASTQ DATA, Quality format converter, Compute quality statistics, Draw quality score boxplot, Draw nucleotides distribution chart, FASTQ to FASTA converter, Filter by quality, Remove sequencing artifacts, Barcode Splitter, Clip adapter sequences, Collapse sequences, Rename sequences, Reverse-Complement, and Trim sequences.

The Clip tool configuration includes:

- Library to clip:** 6: FASTQ Groomer on data 2
- Minimum sequence length (after clipping, sequences shorter than this length will be discarded):** 15
- Source:** Enter custom sequence
- Enter custom clipping sequence:** CATTGGTTGGCTATTAATTG
- enter non-zero value to keep the adapter sequence and x bases that follow it:** 0
- use this for hairpin barcoding. keep at 0 unless you know what you're doing.**
- Discard sequences with unknown (N) bases:** No
- Output options:** Output both clipped and non-clipped sequences

What it does: This tool clips adapters from the 3'-end of the sequences in a FASTA/FASTQ file.

Clipping Illustration: A diagram showing a sequence being clipped by a red bar representing the adapter sequence.

Clipping Example: A sequence example showing the adapter being removed from the end of a read.

History: The history panel shows the following steps:

- Unnamed history (106.9 Mb)
- 12: Clip on data 8 (5.0 Mb, format: fastqsanger, database: dm3)
 - Info: Clipping Adapter: CATTGGTTGGCTATTAATTGACAAATGAGATCGGAAGAGCACACGTCT Min. Length: 15 Input: 49706 reads. Output: 45058 reads. discarded 4350 too-short reads. discarded 298 adapter-only reads.
- 11: Clip on data 7 (format: fastqsanger, database: dm3)
- @HWI-ST571:83:D06DMACXX:6:1101:1720 GCCGNNNNNNNGGGGGGGGGGNCGGCNNNNNNN + #####
- @HWI-ST571:83:D06DMACXX:6:1101:2235 ATTCACNNNAACGGTCATAACGTTGGCTTNNNA/
- 10: Clip on data 6
- 9: Clip on data 5
- 8: FASTQ Groomer on data 4

A red box highlights the step 12: Clip on data 8, and a red arrow points to the 'Infos about the number of discarded sequences' text.

Infos about
the number
of discarded
sequences

LCB

4- Check reads quality with fastQC

- After clipping you can check the quality of your reads with the fastQC quality control tool (“NGS: QC and manipulation → Fastqc: Fastqc QC”)
- For some examples of good and bad results:
<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- If you think that the mean quality of some base of your reads is too low you can trim these reads with (“NGS: QC and manipulation → FASTQ Trimmer”)

5- Mapping of the reads

Tophat

The screenshot shows the Galaxy / BITS interface with the following details:

- Tools Panel:** NGS TOOLS section expanded, showing:
 - NGS: QC and manipulation**
 - NGS: Picard (beta)**
 - NGS: Mapping**
 - NGS: Indel Analysis**
 - NGS: RNA Analysis** (highlighted with a red arrow)
 - RNA-SEQ**
 - Tophat for Illumina** (highlighted with a red arrow)
 - Tophat for SOLiD**
 - Cufflinks**
 - Cuffcompare**
 - Cuffmerge**
 - Cuffdiff**
 - FILTERING**
 - Filter Combined Transcripts**
 - DE Seq**
- Tool Configuration:** Tophat for Illumina (version 1.5.0) selected.
 - Reference genome: Drosophila melanogaster genome 3
 - RNA-Seq FASTQ file: 13: Clip on data 9
 - Is this library mate-paired?: Single-end
 - TopHat settings to use: Default settings
- Execute Button:** A red arrow points to the "Execute" button.
- History Panel:** Shows a list of previous analyses:
 - Unnamed history (168.1 Mb)
 - 21: FastQC data 17.html
 - 20: FastQC data 16.html
 - 19: FastQC data 14.html
 - 18: FastQC data 13.html
 - 17: Clip on data 12
 - 16: Clip on data 11
 - 14: Clip on data 10
 - 13: Clip on data 9
 - 12: Clip on data 8
 - 11: Clip on data 7
 - 10: Clip on data 6
 - 9: Clip on data 5
 - 8: FASTQ Groomer on data 4
 - 7: FASTQ Groomer on data 3
 - 6: FASTQ Groomer on

5- Mapping of the reads

Tophat

4 files are produced by Tophat, our mapped reads are in the “accepted_hits” one

The screenshot shows the Galaxy / BITS web interface. The top navigation bar includes "Galaxy / BITS", "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", "User", and "Using 64%". The left sidebar under "Tools" lists "NGS TOOLS" with sub-sections: "NGS: QC and manipulation", "NGS: Picard (beta)", "NGS: Mapping", "NGS: Indel Analysis", "NGS: RNA Analysis", "RNA-SEQ", and "FILTERING". Under "RNA-SEQ", there are links for Tophat, Cufflinks, Cuffcompare, Cuffmerge, Cuffdiff, and FastQC. A green success message box in the center states: "The following job has been successfully added to the queue:" followed by four job entries: "23: Tophat for Illumina on data 13: insertions", "24: Tophat for Illumina on data 13: deletions", "25: Tophat for Illumina on data 13: splice junctions", and "26: Tophat for Illumina on data 13: accepted_hits". Below this message, a note says: "You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered." To the right, the "History" pane lists previous runs: "Unnamed history" (168.4 Mb), "26: Tophat for Illumina on data 13: accepted_hits" (highlighted with a red border), "25: Tophat for Illumina on data 13: splice junctions", "24: Tophat for Illumina on data 13: deletions", "23: Tophat for Illumina on data 13: insertions", "21: FastQC data 17.html", "20: FastQC data 16.html", "19: FastQC data 14.html", "18: FastQC data 13.html", "17: Clip on data 12", "16: Clip on data 11", "14: Clip on data 10", "13: Clip on data 9", and "12: Clip on data 8".

6- Get differential expression

cuffdiff

For differential expression calculation we will switch to the full dataset that can be found here:

- * <http://ngsworkshop.aertslab.org/> → load the 4 Tophat* files via URL box in “get data”
→ load the gtf file Flybase2006.gtf

* Or in “shared data → data libraries → Galaxy RNA workshop 20 sep 2012”

Use replicates → add new group : wild_type → add new replicates : Dmel1
→ add new replicates : Dmel2
→ add new group : glass_mutant → add new replicates : mut_r2
→ add new replicates : mut_r3

Other options to change :

- Perform quartile normalization : yes

6- Get differential expression

cuffdiff (DESeq, EDSeq etc... Can also be used)

Galaxy / BITS Using 67%

Analyze Data Workflow Shared Data Visualization Help User

Tools

- Get Data
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- FASTA manipulation
- NGS TOOLS
- NGS: QC and manipulation
- NGS: Picard (beta)
- NGS: Mapping
- NGS: Indel Analysis
- NGS: RNA Analysis
- RNA-SEQ
- Tophat for Illumina
- Find splice junctions using RNA-seq data
- Tophat for SOLID
- Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffmerge merge together several Cufflinks assemblies
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use
- FILTERING
- Filter Combined Transcripts using tracking file
- DE Seq Run Differential Expression analysis from SAM To Count data
- SAM To Counts Produce count data from SAM files
- NGS: SAM Tools
- NGS: GATK Tools (beta)
- NGS: Variant Detection
- NGS: Peak Calling

Transcripts:
44: http://ngsworkshop.aertslab.org/Flybase2006.gtf

A transcript GTF file produced by cufflinks, cuffcompare, or other source.

Perform replicate analysis:
Yes

Perform cuffdiff with replicates in each group.

Groups

- Group 1**
- Group name (no spaces or commas):** wild_type
- Replicates**
- Replicate 1**
- Add file:** 40: http://ngsworkshop.aertslab.org/accepted_hits.bam
- Remove Replicate 1**
- Replicate 2**
- Add file:** 41: http://ngsworkshop.aertslab.org/accepted_hits.bam
- Remove Replicate 2**
- Add new Replicate**
- Remove Group 1**
- Group 2**
- Group name (no spaces or commas):** glass_mutant
- Replicates**
- Replicate 1**
- Add file:** 42: http://ngsworkshop.aertslab.org/accepted_hits.bam
- Remove Replicate 1**
- Replicate 2**
- Add file:** 43: http://ngsworkshop.aertslab.org/accepted_hits.bam
- Remove Replicate 2**
- Add new Replicate**
- Remove Group 2**
- Add new Group**

False Discovery Rate: 0.05

The allowed false discovery rate.

Min Alignment Count: 10

The minimum number of alignments in a locus for needed to conduct significance testing on changes in that locus observed between samples.

Perform quartile normalization: Yes

Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Perform Bias Correction: No

Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Set Parameters for Paired-end Reads? (not recommended): No

Execute

History

- Unnamed history 3.1 Gb
- 44: http://ngsworkshop.aertslab.org/Flybase2006.gtf
- 43: http://ngsworkshop.aertslab.org/Tophat_for_Illumina_on_mut-r3_accepted_hits.bam
- 42: http://ngsworkshop.aertslab.org/Tophat_for_Illumina_on_mut-r2_accepted_hits.bam
- 41: http://ngsworkshop.aertslab.org/Tophat_for_Illumina_on_Dmel2_accepted_hits.bam
- 40: http://ngsworkshop.aertslab.org/Tophat_for_Illumina_on_Dmel1_accepted_hits.bam
- 38: Tophat for Illumina on data 17; accepted hits
- 34: Tophat for Illumina on data 16; accepted hits
- 30: Tophat for Illumina on data 14; accepted hits
- 26: Tophat for Illumina on data 13; accepted hits
- 21: FastQC_data 17.html
- 20: FastQC_data 16.html
- 19: FastQC_data 14.html
- 18: FastQC_data 13.html
- 17: Clip on data 12
- 16: Clip on data 11
- 14: Clip on data 10
- 13: Clip on data 9
- 12: Clip on data 8
- 11: Clip on data 7
- 10: Clip on data 6
- 9: Clip on data 5
- 8: FASTQ Groomer on data 4
- 7: FASTQ Groomer on data 3
- 6: FASTQ Groomer

7- Visualization of your mapped reads

While cuffdiff is running you can visualize your mapped data:

Download bam and bai to have a look in local IGV, or use the visualization button (new visualisation).

The screenshot shows the Galaxy / BITS web interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A message in the center says: 'The following job has been successfully added to the queue:' followed by four job entries: 35: Tophat for Illumina on data 17: insertions, 36: Tophat for Illumina on data 17: deletions, 37: Tophat for Illumina on data 17: splice junctions, and 38: Tophat for Illumina on data 17: accepted_hits. Below this, a note says: 'You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.' On the left, a 'Tools' sidebar lists 'NGS TOOLS' (NGS: QC and manipulation, NGS: Picard (beta), NGS: Mapping, NGS: Indel Analysis, NGS: RNA Analysis) and 'RNA-SEQ' (Tophat for Illumina, Cufflinks, Cuffcompare, Cuffmerge, Cuffdiff). On the right, a 'History' pane shows a list of completed jobs: 38: Tophat for Illumina on data 17: accepted_hits (5.2 Mb, format: bam, database: dm3, Info: TopHat v1.4.0), 34: Tophat for Illumina on data 16: accepted_hits, 30: Tophat for Illumina on data 14: accepted_hits, 26: Tophat for Illumina on data 13: accepted_hits, 21: FastQC data 17.html, 20: FastQC data 16.html, 19: FastQC data 14.html, 18: FastQC data 13.html, and 17: Clin on data 12. A context menu is open over the first job in the history list, showing options: 'Download Dataset', 'ADDITIONAL FILES', and 'Download bam_index'.

Laboratory of Computational Biology

LCB

7- Visualization in galaxy

Iz example (only known glass target)



7- Visualization in IGV

using the entire BAM file (File: Load from URL)

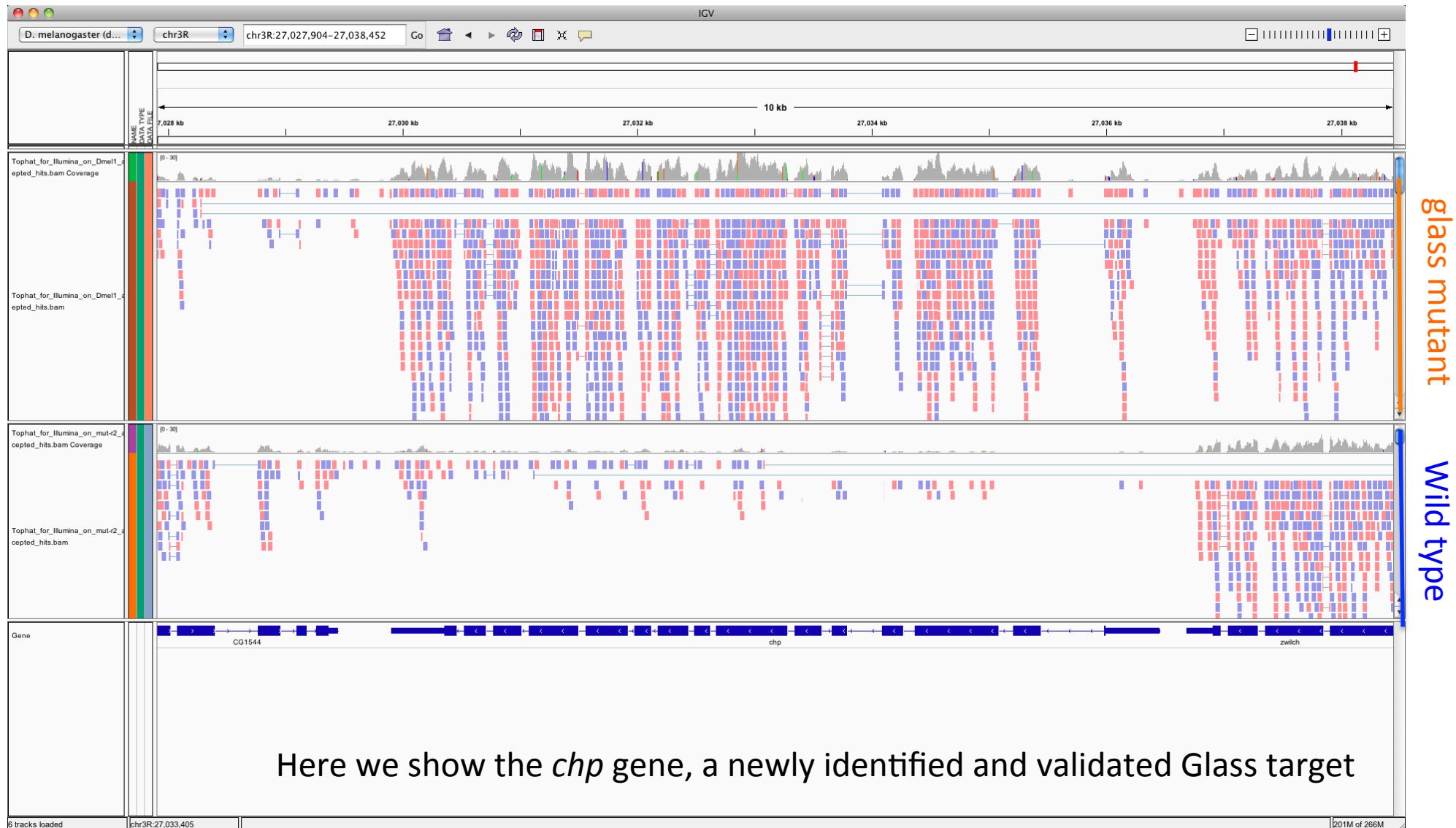
Copy BAM file URLs from <http://ngsworkshop.aertslab.org/>



7- Visualization in IGV

using the entire BAM file (File: Load from URL)

Copy BAM file URLs from <http://ngsworkshop.aertslab.org/>



6- Cuffdiff results

Cuffdiff produce multiple files, we will now work with the gene differential expression testing

The screenshot shows the Galaxy/BITS interface with a search bar and a list of tools. The history panel on the right lists several analysis steps, with one specific entry highlighted by a red box:

- S2: Cuffdiff on data 42, data 43, and others: gene differential expression testing**
- 44: http://ngsworkshop.aertslab.org/Flybase2006.gtf
- 43: http://ngsworkshop.aertslab.org/Tophat for Illumina on mut-r3 accepted hits.bam
- 42: http://ngsworkshop.aertslab.org/Tophat for Illumina on mut-r2 accepted hits.bam
- 41: http://ngsworkshop.aertslab.org/Tophat for Illumina on Dmel2 accepted hits.bam
- 40: http://ngsworkshop.aertslab.org/Tophat for Illumina on Dmel1 accepted hits.bam
- 38: Tophat for Illumina on data 17: accepted hits
- 34: Tophat for Illumina on data 16: accepted hits
- 30: Tophat for Illumina on data 14: accepted hits
- 26: Tophat for Illumina on data 13: accepted hits
- 21: FastQC data 17.html
- 20: FastQC data 16.html
- 19: FastQC data 14.html
- 18: FastQC data 13.html
- 17: Clip on data 12
- 16: Clip on data 11
- 14: Clip on data 10
- 13: Clip on data 9
- 12: Clip on data 8
- 11: Clip on data 7
- 10: Clip on data 6

8- GO analysis

rank the cuffdiff result file

Rank the result file with “Filter and Sort → sort” according to:

- the best Q-value
- For an equal Q-value rank according to the logFC (glass being an activator, in the ascending order)

Galaxy / BITS Analyze Data Workflow Shared Data Visualization Help User Using 67%

Tools

search tools

[Get Data](#)

[Lift-Over](#)

[Text Manipulation](#)

[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](#)
- [Convert Formats](#)
- [Extract Features](#)
- [Fetch Sequences](#)
- [Fetch Alignments](#)
- [Get Genomic Scores](#)
- [Operate on Genomic Intervals](#)
- [Statistics](#)
- [Graph/Display Data](#)
- [Regional Variation](#)
- [Multiple regression](#)
- [Multivariate Analysis](#)
- [Motif Tools](#)

Sort (version 1.0.1)

Sort Query: 52: Cuffdiff on data ..ion testing

on column: c13

with flavor: Numerical sort

everything in: Ascending order

Column selections

Column selection 1

on column: c10

with flavor: Numerical sort

everything in: Ascending order

Add new Column selection

Execute

TIP: If your data is not TAB delimited, use [Text Manipulation->Convert](#)

Syntax

This tool sorts the dataset on any number of columns in either ascending or descending order.

Numerical sort orders numbers by their magnitude, ignores all characters besides numbers, and evaluates a string of numbers to the value they signify. Alphabetical sort is a phonebook type sort based on the conventional order of letters in an alphabet. Each nth letter is compared with the nth letter of other words in the list, starting at the first letter of each word and advancing to the second, third, fourth, and so on, until the order is established. Therefore, in an alphabetical sort, 2 comes after 100 (1 < 2).

History

52: Cuffdiff on data 42, data 43, and others: gene differential expression testing
14,050 lines
format: tabular, database: dm3
Info: cuffdiff v1.3.0 (3022)
cuffdiff --no-update-check -q -p 4
-c 10 --FDR 0.050000 -N --labels
wild_type,glass_mutant
/mnt/galaxydb/files/001/dataset_1
441.dat
/mnt/galaxydb/files/001/dataset_1
437.dat,/mnt/galaxydb/files/001/d
ataset_1438.dat /mnt/galaxydb/file
1 2 3 4
test_id gene_id gene locus
CG00000 CG00000 - chrX:23877-367:
CG10000 CG10000 - chr3R:24574104:
CG10001 CG10001 - chr3R:24562830:
CG10002 CG10002 - chr3R:24406804:
CG10005 CG10005 - chr3R:7800876-
44: http://ngsworkshop.aertslab.org/
Flybase2006.gtf
43: http://ngsworkshop.aertslab.org/
Tophat for Illumina on mut-
r3 accepted_hits.bam
42:

8- GO analysis

rank the cuffdiff result file

Rank the result file with “Filter and Sort → sort” according to:

- the best Q-value
- For an equal Q-value rank according to the logFC (glass being an activator, in the ascending order)

Keep only the first column with “Text manipulation → cut”

The screenshot shows the Galaxy web interface with the following details:

- Tools Panel:** Shows various tools under "Text Manipulation". The "Cut" tool (version 1.0.1) is selected.
- Cut Tool Configuration:**
 - Cut columns:** c1
 - Delimited by:** Tab
 - From:** 58: Sort on data 52
 - Execute:** A blue button with a red arrow pointing to it.
- Tool Help:**
 - WARNING:** This tool breaks column assignments. To re-establish column assignments run the tools and click on the pencil icon in the latest history item.
 - The output of this tool is always in tabular format (e.g., if your original delimiters are commas, they will be replaced with tabs). For example:

apple,is,good
windows,is,bad

will give:

apple good
windows bad
- What it does:** This tool selects (cuts out) specified columns from the dataset.
- Example:** Input dataset (six columns: c1, c2, c3, c4, c5, and c6):
chr1 10 1000 gene1 0 +
chr2 100 1500 gene2 0 +
cut on columns "c1,c4,c6" will return:

History Panel: Shows the following items:

- 58: Sort on data 52
- 52: Cuffdiff on data 42, data 43, and others: gene differential expression testing
- 44: http://ngsworkshop.aertslab.org/Flybase2006.gtf
- 43: http://ngsworkshop.aertslab.org/Tophat for Illumina on mut-r3 accepted hits.bam

8- GO analysis

GORILLA

Save your ranked file to use it in Gorilla: <http://cbl-gorilla.cs.technion.ac.il/>

- Option:
- Single ranked list of genes
 - All ontologies

Galaxy / BITS Analyze Data Workflow Shared Data Visualization Help User

Tools

search tools

Get Data

Lift-Over

Text Manipulation

- Add column to an existing dataset
- Compute an expression on every row
- Concatenate datasets tail-to-head
- Cut columns from a table
- Merge Columns together
- Convert delimiters to TAB
- Create single interval as a new dataset
- Change Case of selected columns
- Paste two files side by side
- Remove beginning of a file
- Select random lines from a file
- Select first lines from a dataset
- Select last lines from a dataset
- Trim leading or trailing characters
- Line/Word/Character count of a dataset
- Secure Hash / Message Digest on a dataset
- Filter on ambiguities in polymorphism datasets
- Arithmetic Operations on tables

Filter and Sort

- Filter data on any column using

CG6518
CG5653
CG4178
CG9935
CG6188
CG6821
CG31313
CG17975
CG18331
CG2559
CG2044
CG6910
CG34076
CG4927
CG1744
CG7105
CG8799
CG14994
CG1090
CG11650
CG15825
CG13360
CG31775
CG11163
CG1028
CG1112
CG3747
CG9432
CG11064
CG3389
CG34377
CG11051
CG14989
CG31619
CG9887
CG30040
CG3625
CG31795
CG32458
CG2297

History

Drosophila melanogaster

Unnamed history 3.1 Gb

59: Cut on data 58 14,050 lines format: tabular, database: dm3

58: Sort on data 52

52: Cuffdiff on data 42, data 43, and others: gene differential expression testing

44: http://ngsworkshop.aertslab.org/Flybase2006.gtf

43: http://ngsworkshop.aertslab.org/Tophat for Illumina on mut-r3 accepted hits.bam

42: http://ngsworkshop.aertslab.org/Tophat for Illumina on mut-r2 accepted hits.bam

41: http://ngsworkshop.aertslab.org/Tophat for Illumina on Dmel2_accepted_hits.bam

40:

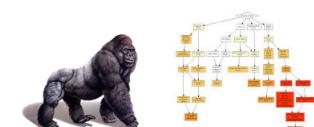
Laboratory of Computational Biology

Reset form



Step 1: Choose organism

Drosophila melanogaster



Step 2: Choose running mode

- Single ranked list of genes Two unranked lists of genes (target and background)

Step 3: Paste a ranked list of gene/protein names

Names should be separated by an <ENTER>. The preferred format is gene symbol. Other supported formats are: gene and protein RefSeq, Uniprot, Unigene and Ensembl. Use [WebGestalt](#) for conversion from other identifier formats.

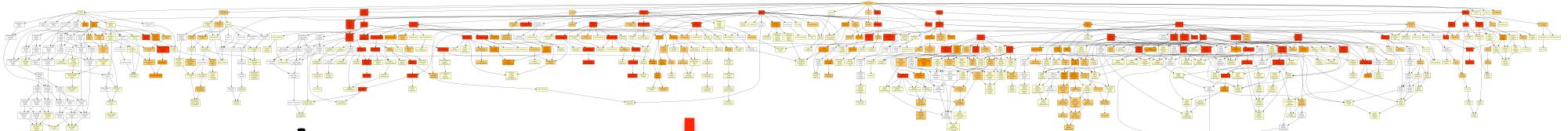
Or upload a file: Galaxy59-[C...abular.rdp]

Choose File Galaxy59-[C...abular.rdp]

Step 4: Choose an ontology

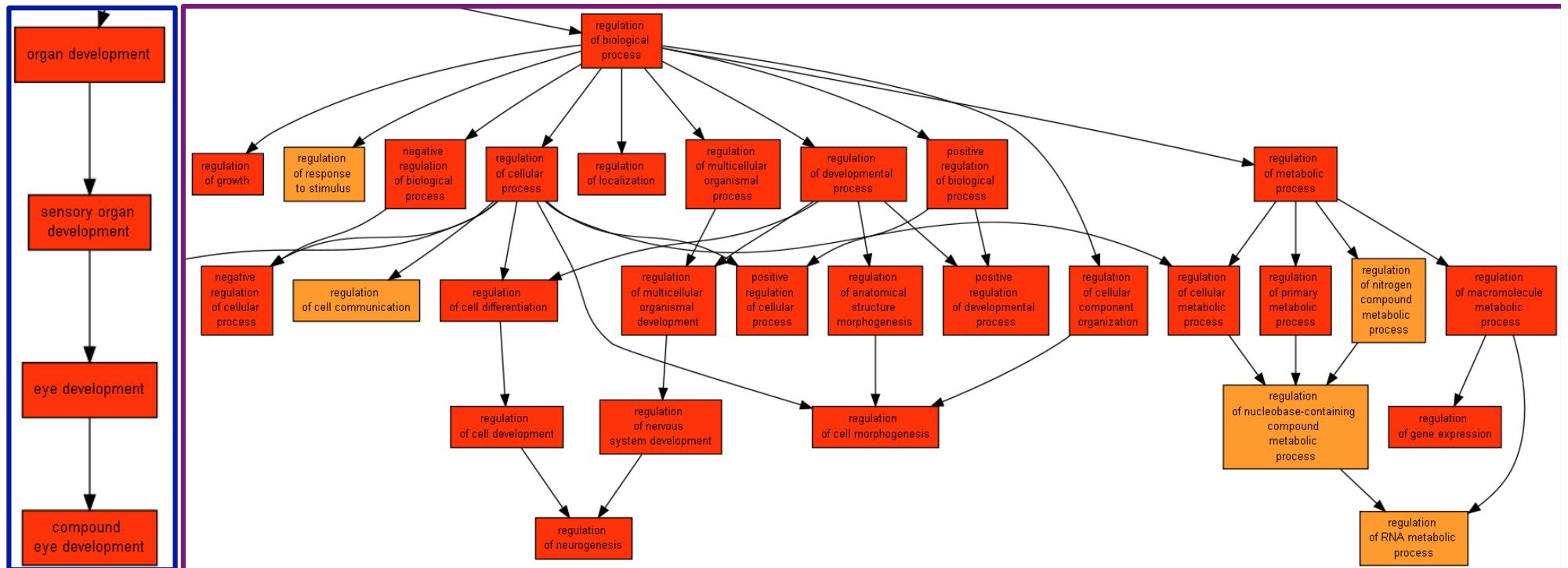
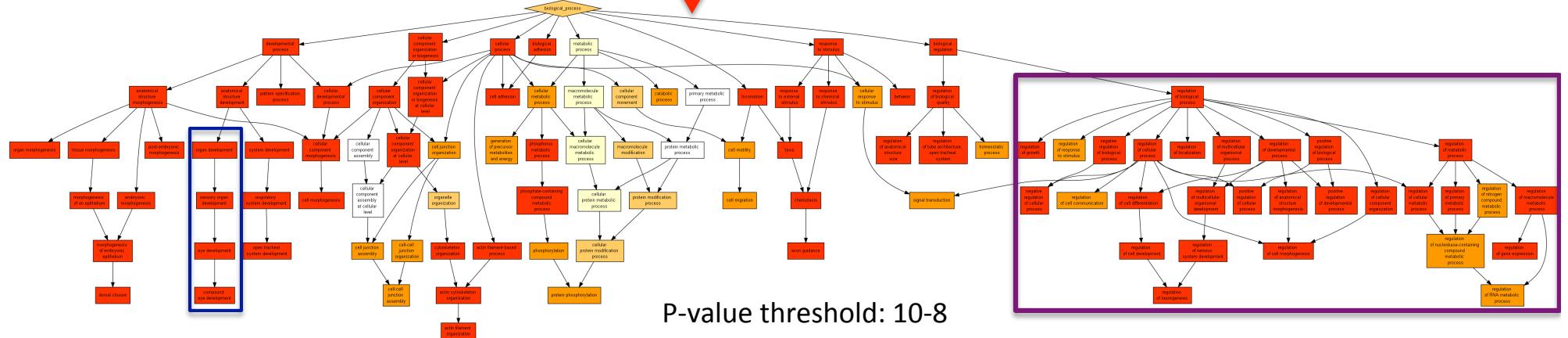
- Process Function Component All

Search Enriched GO terms!



Results

To keep only the most significant results, a filter on the P-value can be applied



LCB

8- GO analysis

GORILLA

Selected parts of the Biological Process Gorilla result table:

GO:0048729	tissue morphogenesis	2.78E-15	8.09E-13	3.67 (10774,178,875,53)
GO:0040008	regulation of growth	9.09E-14	1.73E-11	5.18 (10774,181,402,35)
GO:0060284	regulation of cell development	1.07E-13	1.97E-11	4.57 (10774,191,444,36)
GO:0051960	regulation of nervous system development	3.1E-13	4.8E-11	5.10 (10774,151,434,31)
GO:0007423	sensory organ development	1.37E-12	1.62E-10	2.74 (10774,217,1087,60)
GO:0007411	axon guidance	2.84E-11	2.7E-9	2.61 (10774,183,1288,57)
GO:0048749	compound eye development	2.05E-10	1.61E-8	4.02 (10774,111,748,31)-----
GO:0090066	regulation of anatomical structure size	2.93E-10	2.2E-8	3.25 (10774,95,1291,37)

- 10774 annotated with a Biological Process
- 111 genes annotated “compound eye development”
- 31 genes are annotated “compound eye development” in the first 748 genes of our ranking

9- Find glass direct target i-regulon analysis

Glass being an activator, select only the down-regulated genes with a filter on the sorted data (c10<0).

The screenshot shows the Galaxy / BITS interface. The top navigation bar includes Analyze Data, Workflow, Shared Data, Visualization, Help, and User. The main area is titled "Filter (version 1.1.0)". A dropdown menu "Filter:" is set to "65: Sort on data 52". Below it, "With following condition:" contains the expression "c10<0". To the right, the "History" panel lists several workflow steps, including "66: Cut on data 65", "65: Sort on data 52", and "52: Cuffdiff on data 42, data 43, and others: gene differential expression testing". The left sidebar under "Tools" has a red arrow pointing to the "Filter and Sort" section, which includes options like "Filter data on any column using simple expressions" and "Select lines that match an expression". Another red arrow points to the "Execute" button at the bottom of the filter form. The bottom of the page contains syntax and TIP sections.

Galaxy / BITS

Analyze Data Workflow Shared Data Visualization Help User Using 67%

Tools

- Get Data
- Lift-Over
- Text Manipulation
- 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
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments

Filter (version 1.1.0)

Filter:
65: Sort on data 52

Dataset missing? See TIP below.

With following condition:
c10<0

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')

ℹ 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".

ℹ 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 statements.

Columns are referenced with **c** and a **number**. For example, **c1** refers to the first column of a tab-delimited file

Make sure that multi-character operators contain no white space (e.g., <= is valid while < = is not valid)

When using 'equal-to' operator **double equal sign '==' must be used** (e.g., **c1=='chr1'**)

Non-numerical values must be included in single or double quotes (e.g., **c6=='+'**)

Filtering condition can include logical operators, but make sure operators are all lower case (e.g., **(c1!='chrX' and c1!='chrY') or not c6=='+'**)

History

- Unnamed history 3.1 Gb
- 66: Cut on data 65
- 65: Sort on data 52
- 52: Cuffdiff on data 42, data 43, and others: gene differential expression testing
- 44: http://ngsworkshop.aertslab.org/Flybase2006.gtf
- 43: http://ngsworkshop.aertslab.org/Tophat for Illumina on mut-r3 accepted hits.bam
- 42: http://ngsworkshop.aertslab.org/Tophat for Illumina on mut-r2 accepted hits.bam
- 41: http://ngsworkshop.aertslab.org/Tophat for Illumina on Dmel2 accepted hits.bam
- 40: http://ngsworkshop.aertslab.org/

9- Find glass direct target i-regulon analysis

Select the top 100 down regulated genes

The screenshot shows the Galaxy / BITS interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The 'Tools' panel on the left lists various options under 'Text Manipulation', with 'Select first' highlighted. A red arrow points to the 'Select first' link. Another red arrow points to the 'Execute' button, which is highlighted with a blue box and a red arrow. The main workspace displays the 'Select first (version 1.0.0)' tool configuration. It shows '100' lines selected from '69: Filter on data 65'. The 'History' panel on the right lists several previous steps, including filtering, cutting, sorting, and cuffdiff analysis, along with their corresponding URLs and descriptions.

Galaxy / BITS

Analyze Data Workflow Shared Data Visualization Help User Using 67%

Tools

Get Data Lift-Over **Text Manipulation**

- Add column to an existing dataset
- Compute an expression on every row
- Concatenate datasets tail-to-head
- Cut columns from a table
- Merge Columns together
- Convert delimiters to TAB
- Create single interval as a new dataset
- Change Case of selected columns
- Paste two files side by side
- Remove beginning of a file
- Select random lines from a file
- Select first** lines from a dataset
- Select last lines from a dataset
- Trim leading or trailing

Select first (version 1.0.0)

Select first:

100 lines

from:

69: Filter on data 65

Execute

What it does

This tool outputs specified number of lines from the beginning of a dataset

Example

Selecting 2 lines from this:

```
chr7 56632 56652 D17003_CTCF_R6 310 +
chr7 56736 56756 D17003_CTCF_R7 354 +
chr7 56761 56781 D17003_CTCF_R4 220 +
chr7 56772 56792 D17003_CTCF_R7 372 +
chr7 56775 56795 D17003_CTCF_R4 207 +
```

will produce:

```
chr7 56632 56652 D17003_CTCF_R6 310 +
chr7 56736 56756 D17003_CTCF_R7 354 +
```

History

- Unnamed history 3.1 Gb
- 69: Filter on data 65
- 66: Cut on data 65
- 65: Sort on data 52
- 52: Cuffdiff on data 42, data 43, and others: gene differential expression testing
- 44: http://ngsworkshop.aertslab.org/Flybase2006.gtf
- 43: http://ngsworkshop.aertslab.org/Tophat for Illumina on mутr3 accepted_hits.bam
- 42: http://ngsworkshop.aertslab.org/Tophat for Illumina on мутr2 accepted_hits.bam
- 41: http://ngsworkshop.aertslab.org/Tophat for Illumina on Dmel2 accepted_hits.bam

9- Find glass direct target i-regulon analysis

Cut the 1st column and convert CG identifier to symbol in flybase

The screenshot shows the Galaxy / BITS interface. On the left, the 'Tools' panel is open, displaying a list of 'Text Manipulation' tools. Two specific tools are highlighted with red arrows: 'Cut columns from a table' and 'Cut first lines from a dataset'. The main area shows a history of operations:

- 71: Cut on data 70
- 70: Select first on data 69
- 69: Filter on data 65
- 66: Cut on data 62
- 65: Sort on data 52
- 52: Cuffdiff on data 42, data 43, and others: gene differential expression testing
- 44: http://ngsworkshop.aertslab.org/Flybase2006.gtf
- 43: http://ngsworkshop.aertslab.org/Tophat for Illumina on mtr3 accepted hits.bam
- 42: http://ngsworkshop.aertslab.org/Tophat for Illumina on mtr2 accepted hits.bam

On the right, a list of CG identifiers is displayed:

- CG6518
- CG5653
- CG4178
- CG9935
- CG6188
- CG6821
- CG31313
- CG17975
- CG18331
- CG2559
- CG2044
- CG6910
- CG34076
- CG4927
- CG1744
- CG7105
- CG8799
- CG14994
- CG1090
- CG11650
- CG15825
- CG13360
- CG31775
- CG11163
- CG1028
- CG1112
- CG3747
- CG9432
- CG11064

http://flybase.org/static_pages/downloads/IDConv.html

The screenshot shows the FlyBase ID Converter tool. The URL in the address bar is http://flybase.org/static_pages/downloads/IDConv.html. The page header includes the FlyBase logo and navigation links: Home, Tools, Files, Species, Documents, Resources, News, Help, Archives, Jump to Gene, and Go.

The main form is titled 'ID Converter' and contains the following fields:

- Validate Only (Update to Current IDs)
- Validate and Convert into: Genes
- Enter IDs or Symbols: (text input field containing: CG17975, CG18331, CG2559, CG2044, CG6910, CG34076)
- or Upload File of IDs: (file upload button labeled 'Choose File')
- Go button (highlighted with a red arrow)
- Reset button

A note at the bottom states: You may enter FlyBase IDs or Symbols, including Annotation Symbols and Clone Names.

9- Select genes for an i-regulon analysis

1-click on Flybase HitList

Export converted IDs to: FlyBase HitList file, uniq IDs only file, conversion table

Conversion report			
Submitted ID	Current ID	Converted ID	Related record
CG6518	FBgn0004784	FBgn0004784	inaC
CG5653	FBgn0035943	FBgn0035943	CG5653
CG4178	FBgn0002563	FBgn0002563	Lsp1beta
CG9935	FBgn0039916	FBgn0039916	CG9935
CG6188	FBgn0038074	FBgn0038074	CG6188
CG6821	FBgn0002564	FBgn0002564	Lsp1gamma
CG31313	FBgn0051313	FBgn0051313	CG31313
CG17975	FBgn0028562	FBgn0028562	sut2
CG18331	FBgn0036181	FBgn0036181	Muc68Ca
CG2559	FBgn0002562	FBgn0002562	Lsp1alpha
CG2044	FBgn0002535	FBgn0002535	Lcp4

2-Removed other species corresponding genes
Click on "HitList Conversion Tools → Batch download"

Batch Download

Output Format: FASTA Sequence, Database Format: Full Data Only, Field Data: Selected Fields Only

Output Options: As tab-separated file

Send results to: File, Enter IDs, Symbols or Sequence Coordinates: FBgn0012034, FBgn0260642, FBgn0263111, FBgn0038247, FBgn0037238, FBgn0033484

Allow synonyms: Select fields:

General Information: Check Section (Symbol, Name, Feature Type, Gene Model Status), Uncheck Section

Results Analysis/Refinement: Show related Genes

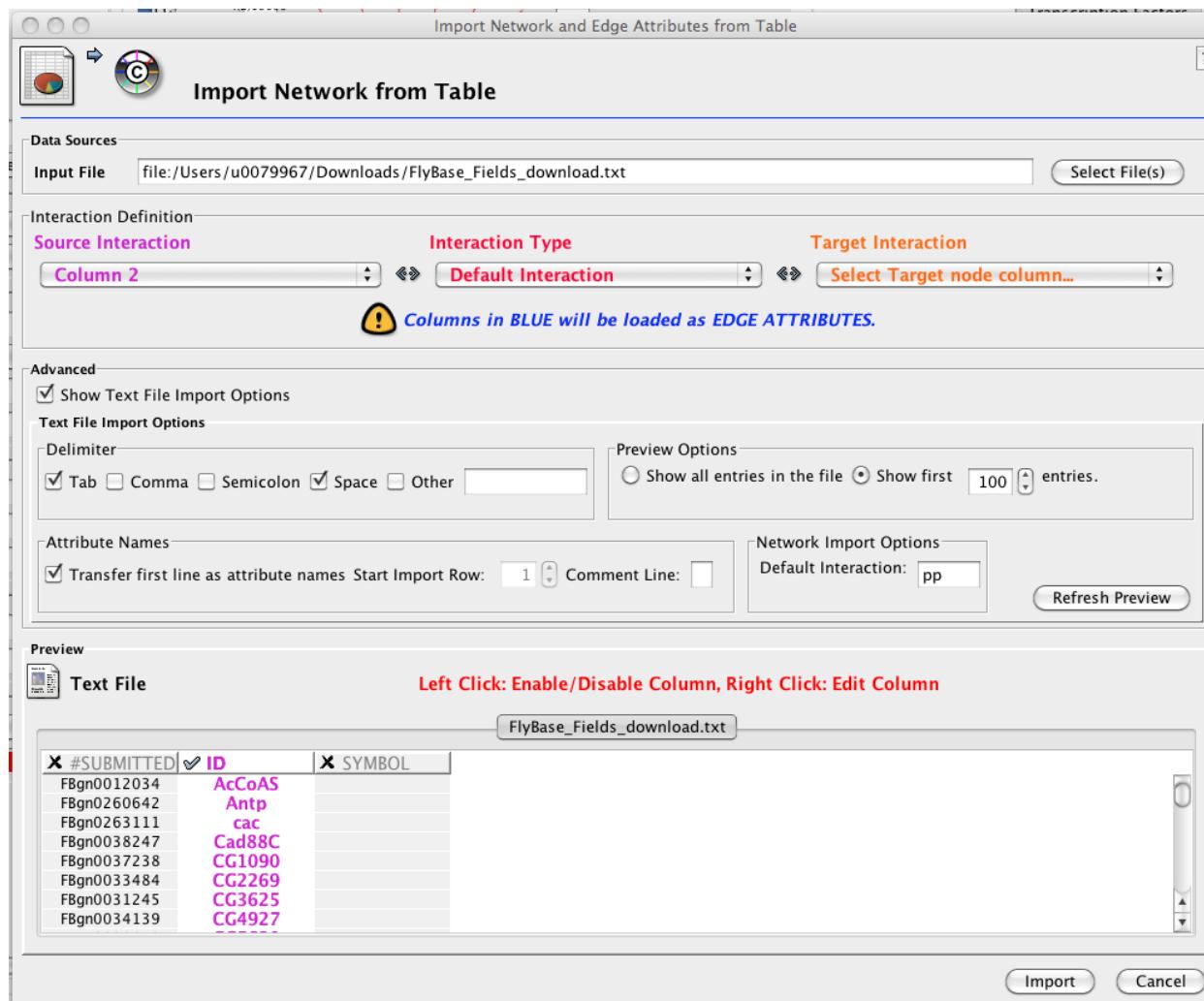
HitList Conversion Tools: 61 matches

#	Symbol	Name	Annotation ID	Cytology	Alleles #	Stocks #	Clones #
1	AcCoAS	Acetyl Coenzyme A synthase	CG9390	78C3-78C4	12	10	150
2	Antp	Antennapedia	CG1028	84A6-84B2	174	106	39
3	cac	cacophony	CG43368	10F7-11A1	57	13	65
4	Cad88C	Cadherin 88C	CG3389	88C10-88C10	9	5	3
5	CG1090	-	CG1090	82A4-82A4	12	10	47
6	CG2269	-	CG2269	46E4-46E4	5	2	73
7	CG3625	-	CG3625	21B7-21B7	11	7	49
8	CG4927	-	CG4927	53C8-53C8	3	3	9
9	CG5630	-	CG5630	93A3-93A4	2	3	36
10	CG5653	-	CG5653	66E5-66E5	2	3	3
11	CG6188	-	CG6188	87C5-87C5	4	3	76
12	CG6910	-	CG6910	69A1-69A1	4	5	59
13	CG9935	-	CG9935	102D1-102D1	6	4	28
14	CG11163	-	CG11163	41F10-41F11	16	13	40
15	CG13360	-	CG13360	1C4-C4	2	2	15
16	CG14989	-	CG14989	64A5-64A5	3	3	141
17	CG31313	-	CG31313	88C6-88C6	4	2	15
18	CG31619	-	CG31619	39F1-39F3	15	12	74
19	CG31775	-	CG31775	35B5-35B5	-	-	105
20	CG32082	-	CG32082	68A8-68A9	6	3	35
21	CG34377	-	CG34377	94A5-94A5	14	13	1
22	CG42351	-	CG42351	55F1-55F1	9	7	35
23	chp	chaoptic	CG1744	100B5-100B6	16	5	78
24	Dmir CG2269	-	-	-	-	-	-
25	Dmir CG4821	-	-	-	-	-	-
26	Decam3	Down syndrome cell adhesion molecule 3	CG31190	90B1-90B1	12	8	20
27	Dsim AcCoAS	Acetyl Coenzyme A synthase	-	-	-	-	-
28	Dyak GE16568	-	GE16568	-	-	-	-
29	Dyak GE21891	-	GE21891	-	-	-	-
30	Dyak pst	-	GE20418	-	-	-	-
31	Eaat1	Excitatory amino acid transporter 1	CG3747	30A8-30A8	15	4	236
32	fon	fondue	CG15825	37D3-37D3	5	1	161
33	futsch	futsch	CG34387	2A3-2A3	22	7	36
34	Gad1	Glutamic acid decarboxylase 1	CG14994	64A5-64A5	20	11	100
35	IA-2	IA-2 ortholog	CG31795	21E3-21E3	9	7	70
36	inaC	inactivation no afterpotential C	CG6518	53E1-53E1	21	11	53
37	jeb	jelly belly	CG30040	48E1-48E2	13	8	17
38	I(2)01289	lethal (2) 01289	CG9432	42C6-42C7	23	9	138
39	I(2)03659	lethal (2) 03659	CG8799	45D1-45D1	6	6	3
40	Lcp1	Larval cuticle protein 1	CG11650	44C6-44C6	4	3	30
41	Lcp4	Larval cuticle protein 4	CG2044	44D1-44D1	5	5	22
42	Lsp1a	Larval serum protein 1 a	CG2559	11A12-11A12	7	3	36
43	Lsp1b	Larval serum protein 1 b	CG4178	21E2-21E2	8	5	155
44	Lsp1y	Larval serum protein 1 y	CG6821	61A6-61A6	8	2	277
45	lz	lozenge	CG1689	80D5-80D6	160	62	5
46	mp	multiplexin	CG42543	65E1-65E2	18	14	48
47	mtND3	mitochondrial NADH-ubiquinone oxidoreductase chain 3	CG34076	-	1	-	-
48	Muc68Ca	Mucin 68Ca	CG18331	68C15-68C15	1	2	1
49	Nlp2	Neuropeptide-like precursor 2	CG11051	70A6-70A6	2	3	322
50	nrm	neuromusculin	CG43079	80C3-80C3	21	17	120
51	Obp44a	Odorant-binding protein 44a	CG2297	44B3-44B3	3	4	103
52	Proct	Proctolin	CG7105	28D2-28D2	5	5	211
53	pst	pastrel	CG8588	65F6-65F7	7	3	276
54	retn	retained	CG5403	59F5-59F5	29	8	102
55	Rfabg	Retinoid- and fatty acid-binding glycoprotein	CG11064	102F8-102F8	16	8	354
56	Scr	Sex combs reduced	CG1030	84A5-84A5	116	62	15
57	scro	scarecrow	CG17594	-	3	3	6

3- Get field data

i-regulon analysis

Use the file from Flybase to run an i-regulon analysis



- Run cytoscape
- File
- import
→import network from table
- Select your flybase file
- Select the column2 as source interaction
- Tick "show text file import options"
→ Transfert first line as attri...

i-regulon analysis

Load i-regulon plugin (“plugins → i-regulon → add sidepanel”)

iRegulon

Predict regulators and targets Query metatargetome

Name for analysis: selected

Species and gene nomenclature: Drosophila melanogaster, FlyBase names

Database

Region- or gene-based analysis? Region Based

Database: 136K regions (11 species)

Region-based specific parameters

Overlap fraction: 0.4

5kb upstream and full transcript

Upstream region: 5000

Downstream region: 5000

Motif prediction

Enrichment score threshold: 3.18

ROC threshold for AUC calculation: 0.01

Rank threshold: 5000

TF prediction

Minimum orthologous identity: 0.0

Maximum motif similarity FDR: 0.001

Node information

Node attribute that corresponds to genID: canonicalName

Number of valid genes (nodes): 1

Submit

Select your putative co-regulated genes

Change default parameter to:

- 5kb upstream and full transcript
- ROC thresold for AUC calculation: 0.01

Submit!

FlyBase_Fields_download-1.txt.1											
CG34377	alpha-Est2	CG3175	alpha-Est7	CG1163	CG13360	CG1090	scro	Scr	sut2		
CG4927	SPS55	Tequila	CG691D	Syn	Vmat	VGlut	ppk13	Obp4a	CG3133		
pst	Proct	CG6188	rdgC	CG9935	Rbp6	Rfabg	CG5653	retn	mp		
mt:ND3	lz	CG15293	Mhc	CG12239	nrm	nrv3	CC42492	Muc68Ca	Nplp2		
Lcp1	Lcp4	CG5630	Il201289	Il203659	CG32082	Lsp1gamma	CG42351	Lsp2	Lsp1alpha		
Lsp1beta	gfA	Gad1	futsch	CC2269	fon	jeb	inaC	IA-2	Hsc1		
Dscam2	dpr18	CG3625	DopEcR	Cyt-b5-r	Est-6	CG31619	Eaat1	CG14989	Dscam4		
Dscam8	CG13023	CG8665	CG11538	CG16712	CG13830	CG34166	CG32198	CG9150	CG12116		
CG4577	CG5210	CG4019	CG7255	CG10186	CG6701	CG10006	CG33509	CG3818	Antp		
Adh	AcCoAS	abd-A	CG42613	CG14042	Cad88C	cac	Cyp6g1	Cyp4d1	cn		
chp											

LCB

i-regulon analysis

Result panel: gl appears at the 2nd position

FlyBase_Fields_download-1.txt

Transcription Factor: gl

motif

Transcription Factors Motifs

...	...	TF	NES	#Targets	#Motifs
✓	1	ara	4.268	37	7
✓	2	gl	3.677	35	2
✓	3	lola	3.668	27	2
✓	4	MTF-1	3.526	12	2

Enriched Motif ID NES AUC ... #T... #TF Filter Tr... #Motifs Orthol... Motif Si... R... #... Target...

5	yefasco-1330	3.6...	0.0...	2	24	0	✓	gl	1	N/A	Direct	21	1	Mhc
9	flyfactorsurvey-g...	3.4...	0.0...	2	23	1						31	15	Rbp6

Gl motif sequence: GAA CC T GAA

Predicted targets

21	1	Mhc
31	15	Rbp6
34	6	Antp
43	1	VGlut
51	2	CG4019
53	2	CG3814
56	2	mp
66	3	abd-A
1231		CG14989
1621		DopEcR
2013		retn
2072		Dscam3
2171		I(2)012...
2352		CG42492
2641		CG31619
3702		chp
4911		CG13830
4981		CG32082
5482		Syn
5922		futsch
6643		Scr
7261		CG2269
7411		HisCl1
7621		ccoonce

i-regulon analysis

Result panel: gl appears at the 2nd position

FlyBase_Fields_download-1.txt

Transcription Factor: gl

motif

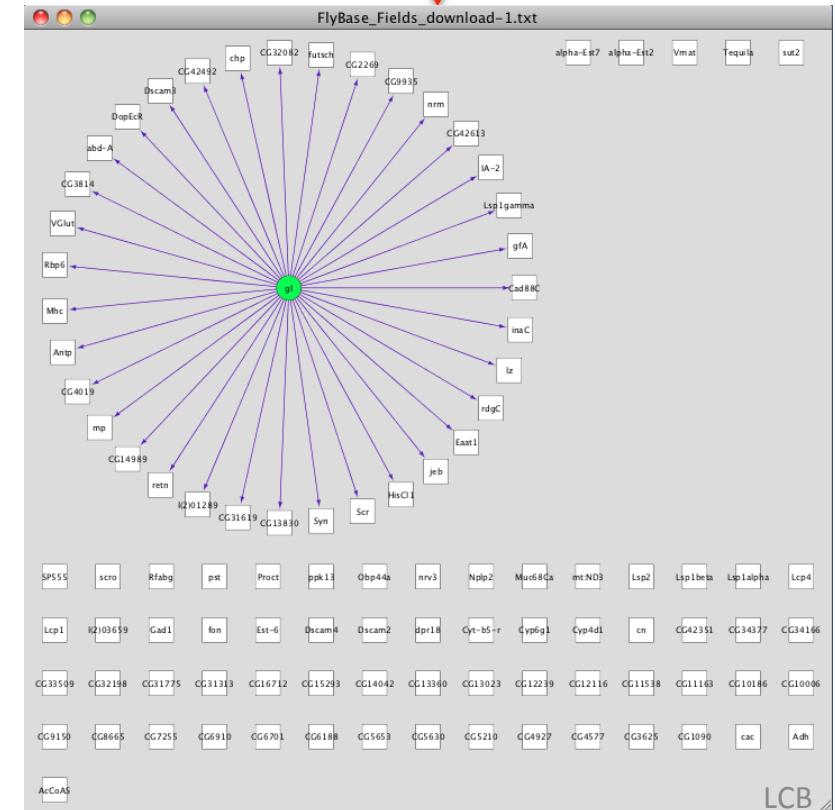
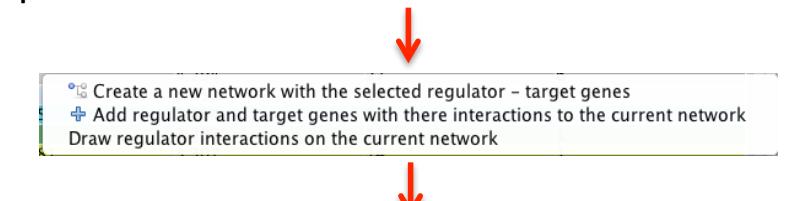
...	...	TF	NES	#Targets	#Motifs
✓	1	ara	4.268	37	7
✓	2	gl	3.677	35	2
✓	3	lola	3.668	27	2
✓	4	MTF-1	3.526	12	2

...	Enriched Motif ID	NES	AUC	#T...	#TF	Filter	Tr...	#Motifs	Orthol...	Motif Si...	R...	#...	Target...	
5	yeftasco-1330	3.6...	0.0...	2	24	0	✓	gl	1	N/A	Direct	21	1	Mhc
9	flyfactorsurvey-g...	3.4...	0.0...	2	23	1						31	15	Rbp6
												34	6	Antp
												43	1	VGlut
												51	2	CG4019
												53	2	CG3814
												56	2	mp
												66	3	abd-A
												1231		CG14989
												1621		DopEcR
												2013		retn
												2072		Dscam3
												2171		I(2)012...
												2352		CG42492
												2641		CG31619
												3702		chp
												4911		CG13830
												4981		CG32082
												5482		Syn
												5922		futsch
												6643		Scr
												7261		CG2269
												7411		HisCl1
												7621		ccoonce

Sequence logo:

GAA CC T GAA

From the TF you choose (right click on it) and its predicted targets, cytoscape can draw the predicted interactions to create a network



i-regulon analysis

Result panel: gl appears at the 2nd position

FlyBase_Fields_download-1.txt

Transcription Factor: gl

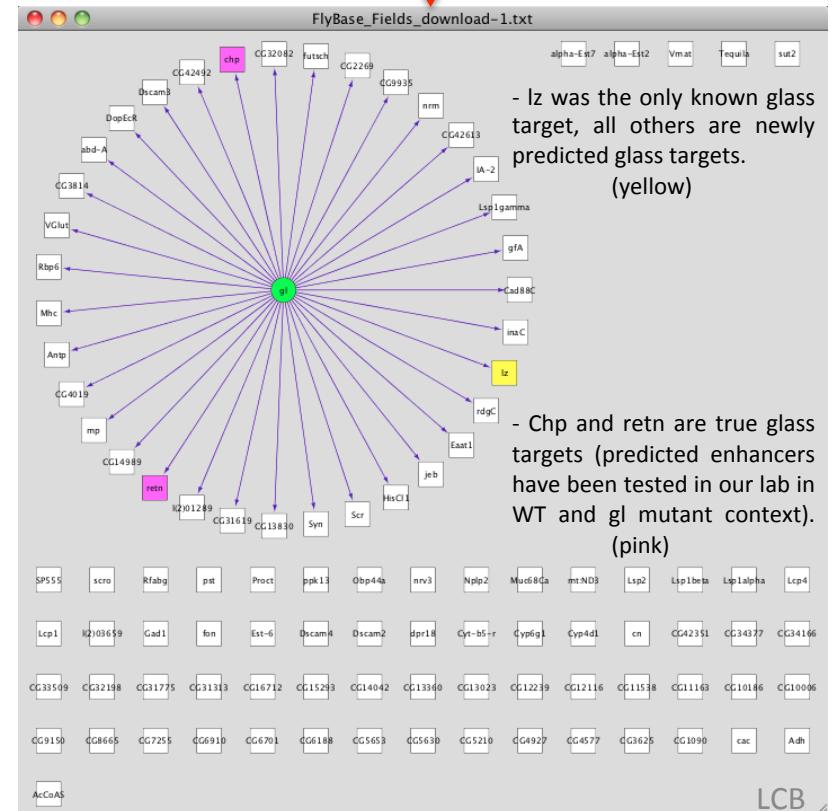
motif

...	...	TF	NES	#Targets	#Motifs
✓	1	ara	4.268	37	7
✓	2	gl	3.677	35	2
✓	3	lola	3.668	27	2
✓	4	MTF-1	3.526	12	2

...	Enriched Motif ID	NES	AUC	#T...	#TF	Filter	Tr...	#Motifs	Orthol...	Motif Si...	R...	#...	Target...	
5	yefascho-1330	3.6...	0.0...	2	24	0	✓	gl	1	N/A	Direct	21	1	Mhc
9	flyfactorsurvey-g...	3.4...	0.0...	2	23	1						31	15	Rbp6
												34	6	Antp
												43	1	VGlut
												51	2	CG4019
												53	2	CG3814
												56	2	mp
												66	3	abd-A
												1231		CG14989
												1621		DopEcR
												2013		retn
												2072		Dscam3
												2171		I(2)012...
												2352		CG42492
												2641		CG31619
												3702		chp
												4911		CG13830
												4981		CG32082
												5482		Syn
												5922		futsch
												6643		Scr
												7261		CG2269
												7411		HisCl1
												7621		cccop8c

From the TF you choose (right click on it) and its predicted targets, cytoscape can draw the predicted interactions to create a network

- >Create a new network with the selected regulator – target genes
- Add regulator and target genes with their interactions to the current network
- Draw regulator interactions on the current network



More galaxy training ...

- Tutorial using human adrenal and brain tissues
 - <https://main.g2.bx.psu.edu/u/jeremy/p/galaxy-rna-seq-analysis-exercise>
- Galaxy Screencasts and Demos
 - <http://wiki.g2.bx.psu.edu/Learn/Screencasts>

More advanced RNA-Seq analysis

- Use of Linux and R/Bioconductor highly recommended
- Read aggregation: HT-Seq (command-line), BEDTools/CoverageBed (command-line), sam2count (also in Galaxy) [need to map on the transcriptome]
- Normalization: EDA-Seq (GC content) (R/Bioconductor), DESeq, edgeR, ...
- Differential expression: DESeq, NOISeq, edgeR, ...
- Discovery of novel transcripts, update existing annotation (Cufflinks, ...)
- Paired-End sequencing:
 - fusion genes (deFuse)
 - de novo transcriptome assembly (Trinity, Velvet, ...)
 - Differential isoforms (DEX-Seq)
- SNP & indel detection (Samtools, ...) [high coverage is needed]