

The Galaxy workflow

George Magklaras PhD RHCE

Biotechnology Center of Oslo &
The Norwegian Center of Molecular Medicine

University of Oslo, Norway

<http://www.biotek.uio.no>

<http://www.ncmm.uio.no>

<http://www.no.embnet.org>

Talleres Internacionales de Bioinformática - UNAM- Enero 2012

Agenda

- Workflows and credible research
- Galaxy: quick overview of the framework
- Sign up to a server and upload/get data
- Add steps to the history and make workflows
- Publish your histories and workflows
- Demo of workflows on our Galaxy server

What is a workflow?

- “A workflow consists of a sequence of concatenated (connected) steps.”
- “It is a depiction of a sequence of operations...”

Source: Wikipedia



Talleres Internacionales de Bioinformática - UNAM- Enero 2012

Reproducible research

- How do you judge the quality of research work?
- You need: Paper + Data + computing environment + workflow (Jon Claerbout, Stanford)
- Paper + Data = No longer good enough

Bioinformatics Workflow Management System

- A BWMS gives you the data and the toolset.
- The workflow as a series of well-defined computational steps.
- Helps you design your processing pipeline and get your results.
- Helps you convince yourself and others about the credibility of your research.

BWMS Example 1: Taverna

The screenshot displays the Taverna Workbench v1.7.1.0 interface. The top-left pane shows a list of available processors, including various WSDL services from Biomart, Soaplab, and Biomoby. The bottom-left pane, titled 'Advanced model explorer', shows a workflow object named 'Fetch Dragon images from BioMoby' with a table of its components and data links.

Workflow object	Retries	Delay	Backoff	Threads	Critical
Fetch Dragon images from BioMoby					
Workflow inputs					
Workflow outputs					
Processors					
id : cho	0	0	1	1	
namespace : DragonDB:Allele	0	0	1	1	
Decode_base64_to_byte	0	0	1	1	
getJpegFromAnnotatedImage	0	0	1	1	
getDragonSimpleAnnotatedImages	0	0	1	1	
Object	0	0	1	1	
Parse_Moby_Data_JPEGImage	0	0	1	1	
Parse_Moby_Data_SimpleAnnotatedJPEGImage	0	0	1	1	
Data links					
Decode_base64_to_byte:bytes-images					

The main workspace shows a graphical workflow diagram. It starts with an 'Object' node receiving 'id' and 'namespace' inputs. The workflow proceeds through 'getDragonSimpleAnnotatedImages', 'getJpegFromAnnotatedImage', and 'Parse_Moby_Data_JPEGImage'. The final step is 'Parse_Moby_Data_SimpleAnnotatedJPEGImage', which produces two outputs: 'images' and 'annotations'.

```
graph TD; id[id] --> Object[Object]; namespace[namespace] --> Object; Object --> getDragonSimpleAnnotatedImages[getDragonSimpleAnnotatedImages]; getDragonSimpleAnnotatedImages --> getJpegFromAnnotatedImage[getJpegFromAnnotatedImage]; getJpegFromAnnotatedImage --> Parse_Moby_Data_JPEGImage[Parse_Moby_Data_JPEGImage]; Parse_Moby_Data_JPEGImage --> Decode_base64_to_byte[Decode_base64_to_byte]; Parse_Moby_Data_JPEGImage --> Parse_Moby_Data_SimpleAnnotatedJPEGImage[Parse_Moby_Data_SimpleAnnotatedJPEGImage]; Decode_base64_to_byte --> images[images]; Parse_Moby_Data_SimpleAnnotatedJPEGImage --> annotations[annotations];
```

BWMS Example 2: Pipeline Pilot

The screenshot displays the Pipeline Pilot Professional Client interface. The main window shows a workflow diagram with the following components: SD Reader, HTML Table Viewer, Fjernet kolonner (A:=B), Molecule to PNG, Table, nødvendig for å kunne lese..., and PDF Report Writer. The left sidebar shows a tree view of components, with 'Molecule to JPEG' selected. The bottom panel shows the 'Parameters' section for the 'Molecule to JPEG' component.

Parameters

Output	PNG_Image
Image Options	
ImageSize	250
WidthToHeightRatio	1.0
Transparent	True
Caption Property	Name
Chemistry Options	

Talleres Internacionales de Bioinformática - UNAM- Enero 2012

The Galaxy BWMS

- User friendly: Only requires an up-to-date web browser and an internet connection.
- Contains already a large number of integrated tools for NGS
<https://bitbucket.org/galaxy/galaxy-central/wiki/NGSLocalSetup>
- Framework for integrating other tools
<https://community.g2.bx.psu.edu>
- Has an active community that develops the base code plus modules

The galaxy web interface

The screenshot displays the Galaxy web interface with three main panels:

- Tools (left):** A sidebar with a search bar and a list of tool categories including Motif Tools, Multiple Alignments, Metagenomic analyses, FASTA manipulation, NCBI BLAST+, NGS: QC and manipulation, ILLUMINA FASTQ, ROCHE-454 DATA, and AB-SOLID DATA. The 'FASTQ Groomer' tool is highlighted under the 'NGS: QC and manipulation' section.
- FASTQ Groomer (version 1.0.4) (center):** A configuration form for the selected tool. It includes fields for 'File to groom' (set to '15: FASTQ Groomer on data 13'), 'Input FASTQ quality scores type' (set to 'Sanger'), and 'Advanced Options' (set to 'Hide Advanced Options'). An 'Execute' button is at the bottom. Below the form is a 'What it does' section with detailed text explaining the tool's function and conversion options.
- History (right):** A list of workflow steps. The current step, '15: FASTQ Groomer on data 13', is highlighted in green. Below it, a previous step '13: SRR000749.fastq' is also highlighted, showing its details: 464.3 Mb, format: fastq, database: 2, and info: uploaded fastq file. A preview of the FASTQ data is shown below the step details.

Tool selection

Selected tool form

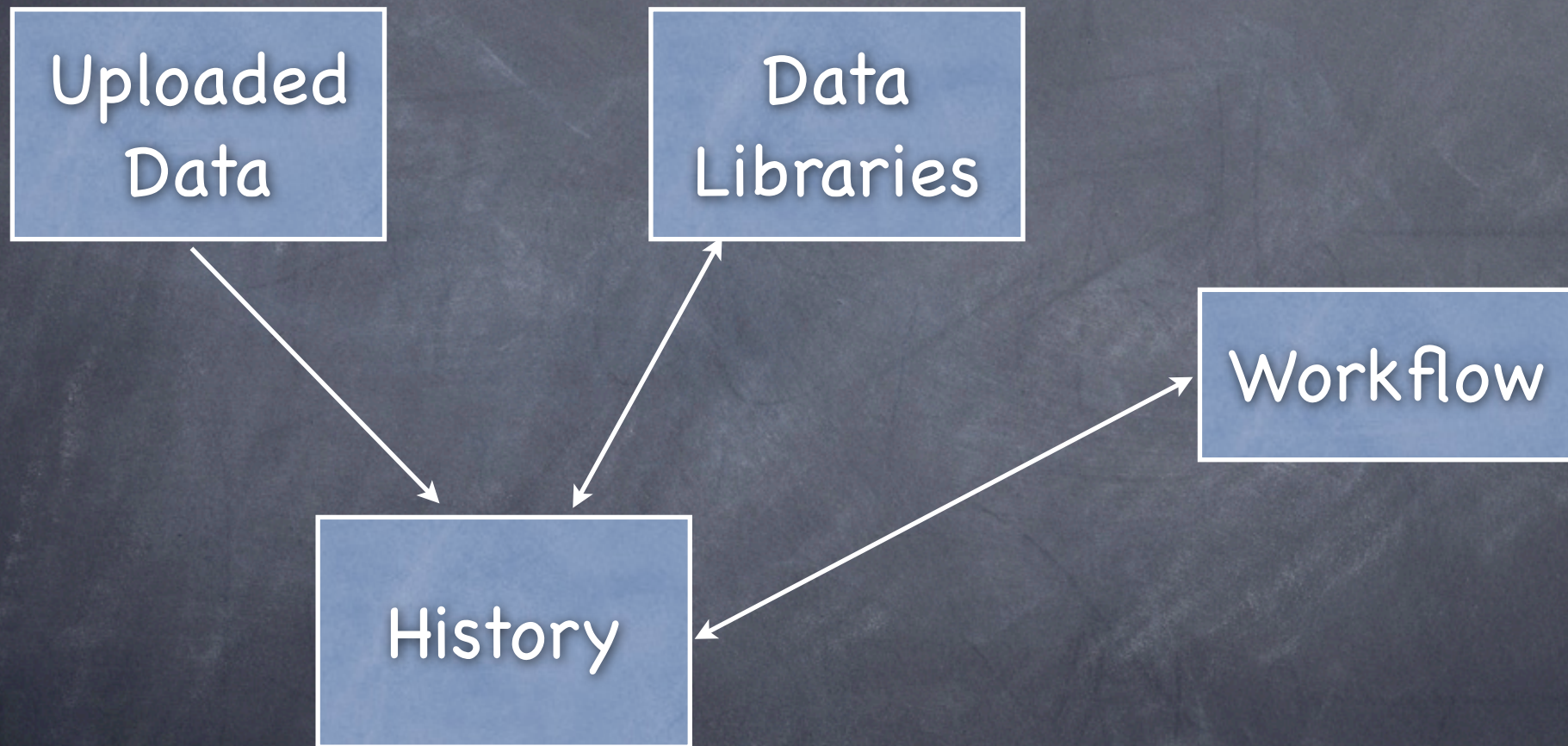
Workflow (history)

Talleres Internacionales de Bioinformática - UNAM- Enero 2012

Basic Galaxy terminology

- Analysis step: A tool which accepts input data and generates output data.
- History: All ordered analysis steps plus the data plus the settings on each step.
- Workflow: Ordered processing steps without the data ("blueprint" of a history)
- Datasets: The input and output data of analysis steps.
- Data Libraries: Specific datasets organized for reference.

Building blocks and information flow



User registration

- You should register in order to get the most out of the Galaxy environment. It allows you to:
 - build and access workflows
 - have access to non-public data files and workflows.
 - Save your datasets and workflow histories.

User registration (2)

Galaxy - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Galaxy New Tab

biotin.uio.no:8080/user/login?webapp=galaxy&use_panels=True Google

Most Visited BOINCstats | User s... MAXMIND GeolIP CPAN ΣΚΑΪ Player TV LIVE... UOP Intranet - -

ant.com Search the Search Download Player Browse Rank: 22,748 Help About Preferences

Galaxy Analyze Data Workflow Shared Data Help **User**

Login Register

Login

Email address:

Password:

Forgot password? [Reset here](#)

Login

http://biotin.uio.no:8080/user/create?

User registration (3)

Create account

Email address:

Password:

Confirm password:

Public name:

Your public name is an identifier that will be used to generate addresses for information you share publicly. Public names must be at least four characters in length and contain only lower-case letters, numbers, and the '-' character.

Galaxy "history"

- Your workflow's scratchpad
- You record and annotate your steps
- You can generate, report, export data to it from data libraries
- You can save, publish your history

History annotation

History Options ▾

Unnamed history 1.8 Gb

Tags:

Annotation / Notes:

Simple workflow to demonstrate FASTQ file processing plus additional steps.

25: transeq on data 16 👁️ ✎ ✕

4,680,118 sequences
format: fasta, database: ?

📄 ⓘ ↻ 📎 📁

```
>SRR000749.1_1 USI-EAS21_60_6387:4:1:380:871
AYVEKYGAX
>SRR000749.2_1 USI-EAS21_60_6387:4:1:466:121
GKISY*KFX
>SRR000749.3_1 USI-EAS21_60_6387:4:1:551:138
DIFLVYKMX
```

16: FASTQ to FASTA on data 15 👁️ ✎ ✕

15: FASTQ Groomer on data 13 👁️ ✎ ✕

13: SRR000749.fastq 👁️ ✎ ✕

Importing data to history

Data Library "Escherichia coli"

Containing the E. Coli reference genome

<input type="checkbox"/> Name	Message	Uploaded By
<input checked="" type="checkbox"/> AP012306.fasta ▼		gmagklaras@gmail.com

For selected datasets:

i TIP: You can download individual library datasets by selecting "Download this dataset" from the context menu (triangle) next to each dataset's name

i TIP: Several compression options are available for downloading multiple library datasets simultaneously:

- gzip: Recommended for fast network connections
- bzip2: Recommended for slower network connections (smaller size but takes longer to compress)
- zip: Not recommended but is provided as an option for those who cannot open the above formats

Exporting data from history

Analyze Data

Workflow

Shared Data

Admin

Help

User

Upload files to a data library

Active datasets in your current history (E.coli mapping)

- 13: SRR000749.fastq
- 28: AP012306.fasta
- 29: SRR001666_1.fastq
- 30: FASTQ Groomer on data 29

Import to library

Exporting/publishing a history

Open Ant preferences window

Using 1.8 Gb

History

Unnamed history

25: transeq on data 16
4,680,118 sequences
format: fasta, database: 2

```
>SRR000749.1_1 USI-EAS21_60_6387:4:1:380:871  
AYVEKYGAX  
>SRR000749.2_1 USI-EAS21_60_6387:4:1:466:121  
GKISY*KFX  
>SRR000749.3_1 USI-EAS21_60_6387:4:1:551:138  
DIFLVYKMX
```

16: FASTQ to FASTA on data 15

15: FASTQ Groomer on data 13

13: SRR000749.fastq

- History Lists
- Saved Histories
- Histories Shared with Me
- Current History
- Create New
- Clone
- Copy Datasets
- Share or Publish
- Extract Workflow
- Dataset Security
- Show Deleted Datasets
- Show Hidden Datasets
- Purge Deleted Datasets
- Show Structure
- Export to File
- Delete
- Delete Permanently
- Other Actions
- Import from File

How to make a workflow (1)

The screenshot displays the Galaxy web interface for creating a workflow. The main workspace is a grid-based canvas titled "Workflow Canvas | Simple sort of exons". It contains four workflow steps connected by arrows:

- Input dataset**: An orange box with "output" as its name.
- Compute**: An orange box with "as a new column to" and "out_file1" as its name.
- Sort**: An orange box with "Sort Query" and "out_file1" as its name.
- Select first**: An orange box with "from" and "out_file1" as its name. This box is highlighted with a blue border.

The right-hand sidebar shows the configuration for the selected "Select first" tool:

- Tool: Select first**
- Select first:** A dropdown menu set to "20".
- from:** A text field containing "Data input 'input' (txt)".
- Edit Step Actions:** Includes a "Rename Dataset" dropdown set to "out_file1" and a "Create" button.
- Edit Step Attributes:** A section for configuring step attributes.
- Annotation / Notes:** A text area for adding notes to the step.

The left-hand sidebar lists various tool categories under "Tools":

- Get Data
- Send Data
- ENCODE Tools
- 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
- Wavelet Analysis
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- FASTA manipulation
- NCBI BLAST+
- NGS: QC and manipulation
- NGS: Picard (beta)
- NGS: Mapping

Talleres Internacionales de Bioinformática - UNAM- Enero 2012

How to make a workflow (2)

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

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

Workflow name
Workflow constructed from history 'E.coli mapping'

Create Workflow | Check all | Uncheck all

Tool	History items created
Upload File <i>This tool cannot be used in workflows</i>	13: SRR000749.fastq <input checked="" type="checkbox"/> Treat as input dataset
Unknown <i>This tool cannot be used in workflows</i>	28: AP012306.fasta <input checked="" type="checkbox"/> Treat as input dataset
Unknown <i>This tool cannot be used in workflows</i>	29: SRR001666_1.fastq <input checked="" type="checkbox"/> Treat as input dataset
FASTQ Groomer <input checked="" type="checkbox"/> Include "FASTQ Groomer" in workflow	32: FASTQ Groomer on data 29
Map with BWA for Illumina <input checked="" type="checkbox"/> Include "Map with BWA for Illumina" in workflow	34: Map with BWA for Illumina on data 32 and data 28: mapped reads
Convert Genomic Intervals To Strict BED6	37: UCSC Main on Human: knownGene

History Lists

- Saved Histories
- Histories Shared with Me
- Current History
- Create New
- Clone
- Copy Datasets
- Share or Publish
- Extract Workflow
- Dataset Security
- Show Deleted Datasets
- Show Hidden Datasets
- Purge Deleted Datasets
- Show Structure
- Export to File
- Delete
- Delete Permanently
- Other Actions
- Import from File

```
>viral1 seq
ftgaatggatgtcaatccgactctacttttcttaaaaa
caccacattcccttatactggagatcctccatacagcc
catggacacagtaaacagaaacacccaactcagaaa
agagactgggtgcaccccagctcaaccgattgatggac
aagtgggatgcaaacacagactgtgttctagaggcta
```

Running (using) the workflow

The screenshot shows the Galaxy web interface. At the top, there is a navigation bar with 'Galaxy' logo and menu items: 'Analyze Data', 'Workflow', 'Shared Data', 'Admin', 'Help', 'User', and 'Using 3.9 Gb'. On the left, a 'Tools' sidebar lists various categories like 'Get Data', 'Send Data', 'ENCODE Tools', etc. The main workspace displays a workflow titled 'Running workflow "Simple sort of exons"'. The workflow description is 'Get coding exon data from human chr1 and sort them'. It consists of four steps: 'Step 1: Input dataset', 'Step 2: Compute', 'Step 3: Sort', and 'Step 4: Select first'. Step 1 is expanded, showing an 'Input Dataset' field with a dropdown menu set to '37: UCSC Main on Huma...-249250621)' and a 'type to filter' input field. Below the steps is a checkbox for 'Send results to a new history' and a 'Run workflow' button. On the right, a 'History' panel shows a list of jobs. The top job is '37: UCSC Main on Human: knownGene (chr1:1-249250621)' with 58,428 regions. Below it is a table of genomic coordinates and gene names.

1. Chrom	2. Start	3. End	4. Name
chr1	12189	12227	uc010nxq.1_cds
chr1	12594	12721	uc010nxq.1_cds
chr1	13402	13639	uc010nxq.1_cds
chr1	69090	70008	uc001aal.1_cds
chr1	324342	324345	uc009vjk.2_cds
chr1	324438	325605	uc009vjk.2_cds

Publishing a workflow

Share or Publish Workflow 'Simple sort of exons'

Making Workflow Accessible via Link and Publishing It

This workflow is currently restricted so that only you and the users listed below can access it. You can:

Make Workflow Accessible via Link

Generates a web link that you can share with other people so that they can view and import the workflow.

Make Workflow Accessible and Publish

Makes the workflow accessible via link (see above) and publishes the workflow to Galaxy's [Published Workflows](#) section, where it is publicly listed and searchable.

Sharing Workflow with Specific Users

You have not shared this workflow with any users.

Share with a user

[Back to Workflows List](#)

From the top menu bar: "Shared Data" ->
"Published Workflows"

DEMO 1: Get coding exons on Human Chr1 and post-process on start and end positions

Step

Chr1 coding exons
UCSC data



Calculate column 7 as
column 3 - column 2



Sort in descending
order of column7



Select the first 20
records

Galaxy tool

Get Data -> UCSC Main
table browser



Text Manipulation ->
Compute



Filter and Sort -> Sort



Text Manipulation ->
Select First

DEMO 2: Characterize a sequence fragment by BLAST search and sort the hits with by alignment length

Step

Fasta file with nucleotide sequence

Nucleotide to protein translation

Homology search of translated sequence

Sort the results by alignment length

Galaxy tool

Get Data -> Upload File (or history import)

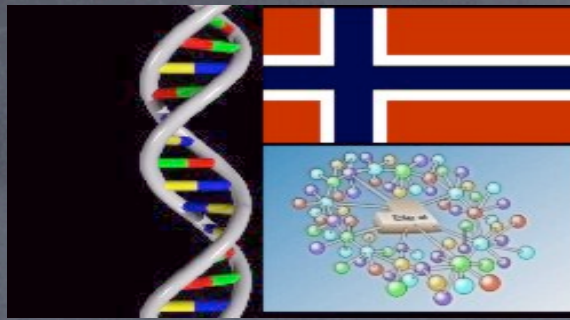
EMBOSS -> Transeq

NCBI BLAST+ -> Blastp

Filter and Sort -> Sort

Exercise

- **Step 1:** User register and login to the Galaxy server (<http://biotin.uio.no:8080>)
- **Step 2:** Locate the Published Workflow called "NGS example". What does it do?
- **Step 3:** Locate the Published Data Library called "Escherichia Coli" which contains an Illumina experiment and a reference Genome.
- **Step 4:** Import both files from this Data Library into your current history.
- **Step 5:** Now run the "NGS example" workflow with the imported files.
- **Step 6:** Can you add steps to your history, make a workflow out of it and publish it for other users to use?



Questions?

admin@embnet.uio.no

and

<http://www.embnet.org/join/ContactRegistration>

Talleres Internacionales de Bioinformática - UNAM- Enero 2012