

Reproducible & Transparent Computational Science with Galaxy

Jeremy Goecks
The Galaxy Team

Doing Good Science

Previous talks: performing an analysis

- setting up and scaling Galaxy
- adding tools
- libraries and sample tracking
- visualizations

Next step: using an analysis to do good science

Galaxy Vision

Supporting accessible, **reproducible**, and **transparent** computational science

- *genomic science is computational*

Transparency ~ sharing and communicating experimental outputs in a meaningful way

- facilitate understanding, reproducing, extending, best practices, collaboration, and publication

Challenges

Computational (genomic) science is difficult to reproduce and communicate:

- large data sets
- complex operations
- details matter
- influx of new tools
- data flow among tools

Galaxy Approach

Open, web-based platform

- easy to access, view, and use analysis objects
- leverage web as an “everything” medium

Integrate analysis workspace with viewing & reading workspace

- quick, simple sharing and reuse
- enable interactive reading

Reproducibility and Transparency in Galaxy

Workflows ~ repeating analyses

Display Framework ~ sharing, viewing

Annotations & Tags ~ explanations, context

Pages ~ communicating and publishing

Workflows

Galaxy workflow (“pipeline”) ~ an abstract analysis that can be repeatedly applied to many different datasets

- › choose datasets and Galaxy runs workflow

Can create workflows by example or via interactive, GUI editor

Highly reusable for individuals and community

- › completely repeatable analyses
- › core component for supporting best practices

Workflow by Example

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: Demo Workflow

Tools to include in workflow:

- Upload File
- Cufflinks
- Cuffcompare
- Cuffdiff

History items created:

- 1: Sample 1 - E18.sam
- 2: Sample 1 - P77.sam
- 3: Cufflinks on data 2: gene expression
- 4: Cufflinks on data 2: transcript expression
- 5: Cufflinks on data 2: assembled transcripts
- 6: Cufflinks on data 1: gene expression
- 7: Cufflinks on data 1: transcript expression
- 8: Cufflinks on data 1: assembled transcripts
- 9: Cuffcompare on data 8 and data 5: data 5 tmap file
- 10: Cuffcompare on data 8 and data 5: data 5 refmap file
- 11: Cuffcompare on data 8 and data 5: data 8 tmap file
- 12: Cuffcompare on data 8 and data 5: data 8 refmap file
- 13: Cuffcompare on data 8 and data 5: combined transcripts
- 14: Cuffcompare on data 8 and data 5: transcript tracking
- 15: Cuffcompare on data 8 and data 5: transcript accuracy
- 16: Cuffdiff on data 1, data 2, and data 13: isoform expression
- 17: Cuffdiff on data 1, data 2, and data 13: gene expression
- 18: Cuffdiff on data 1, data 2, and data 13: TSS groups expression
- 19: Cuffdiff on data 1, data 2, and data 13: CDS Expression FPKM Tracking
- 20: Cuffdiff on data 1, data 2, and data 13: isoform FPKM tracking
- 21: Cuffdiff on data 1, data 2, and data 13: gene FPKM tracking
- 22: Cuffdiff on data 1, data 2, and data 13: TSS groups FPKM tracking
- 23: Cuffdiff on data 1, data 2, and data 13: CDS FPKM tracking
- 24: Cuffdiff on data 1, data 2, and data 13: splicing diff
- 25: Cuffdiff on data 1, data 2, and data 13: promoters diff
- 26: Cuffdiff on data 1, data 2, and data 13: CDS diff

Create a workflow from a history

Can include some or all steps

Workflow Editor

The screenshot displays the Galaxy Workflow Editor interface. The main canvas shows a workflow diagram with the following components:

- Input datasets:** Three 'Input dataset' boxes on the left, each with an 'output' port.
- Cufflinks tools:** Two 'Cufflinks' tool boxes. Each takes a 'SAM file of aligned RNA-Seq reads' as input and produces outputs: 'genes_expression (expr)', 'transcripts_expression (expr)', and 'assembled_isoforms (igt)'. The top Cufflinks tool also takes 'genes_expression (expr)' as input.
- Cuffdiff tool:** A 'Cuffdiff' tool box that takes 'Transcripts' and 'SAM file of aligned RNA-Seq reads' as input. Its outputs include: 'isoforms_exp (tabular)', 'genes_exp (tabular)', 'ts_group_exp (tabular)', 'cds_exp_fpkm_tracking (tabular)', 'isoforms_fpkm_tracking (tabular)', 'genes_fpkm_tracking (tabular)', 'ts_group_fpkm_tracking (tabular)', 'cds_fpkm_tracking (tabular)', 'splicing_diff (tabular)', 'promoters_diff (tabular)', and 'cds_diff (tabular)'.
- Cuffcompare tool:** A 'Cuffcompare' tool box that takes 'CTF file produced by Cufflinks' and 'Reference Annotation' as input. Its outputs include: 'input1_rmap (tabular)', 'input1_refmap (tabular)', 'input2_rmap (tabular)', 'input2_refmap (tabular)', 'transcripts_combined (igt)', 'transcripts_tracking (tabular)', and 'transcripts_accuracy (txt)'.

The right-hand side of the interface shows the 'Details' panel for the selected 'Cuffdiff' tool, including parameters like 'False Discovery Rate' (0.05), 'Min SAM Mapping Quality' (0), and 'Min Alignment Count' (0). It also includes an 'Edit Step Attributes' section and a 'Cuffdiff Overview' section with a warning icon and text: 'There is no such thing as an automated gearshift in expression analysis. It is all like stick-shift driving in San Francisco. In other words, running this tool with default parameters will probably not give you meaningful results. A way to deal with this is to understand the parameters by carefully reading the documentation and experimenting. Fortunately, Galaxy makes experimenting easy.'

Drag, drop, and connect analysis steps to create or edit a workflow

- Validates data flow
- can set parameters in workflow or during runtime

Any tool can be added to a workflow

Reproducibility and Transparency in Galaxy

Workflows ~ repeating analyses

Display Framework ~ sharing, viewing

Annotations & Tags ~ explanations, context

Pages ~ communicating and publishing

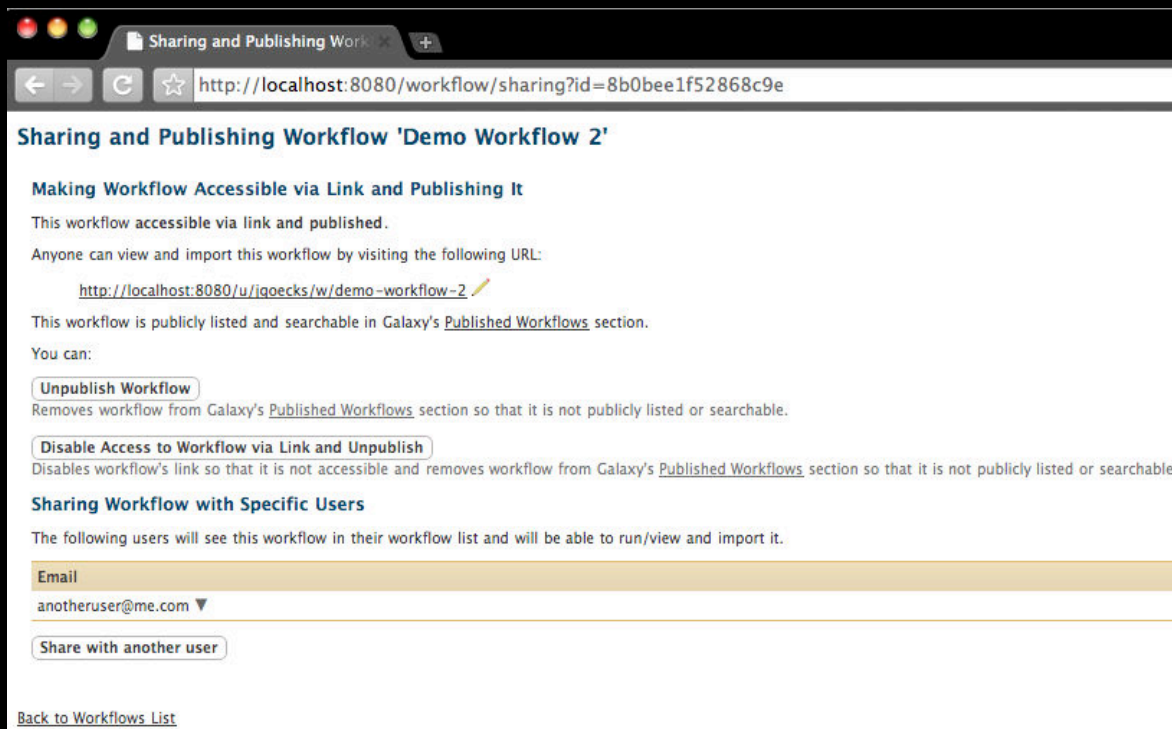
Display Framework

Makes it easy to share or publish items via the web

Shared, published items can be viewed, copied into workspace

Connects viewing & reading with analysis workspace to facilitate reproduction and reuse

Sharing and Publishing



- Simple sharing model: share with an individual, make accessible via link, or publish
- › for histories, workflows, and visualizations
 - › more complex for datasets

Each shared/published item has its own automatically-generated webpage

- › can customize item URL
- › tags and annotations included as well

Viewing a Shared Item

The screenshot displays the Galaxy web interface for a published workflow. The browser address bar shows the URL: `http://localhost:8080/u/jgoecks/w/demo-workflow-using-rna-seq`. The page title is "Galaxy | Published Workflow". The main content area is divided into three steps:

- Step 6: Cuffcompare**
 - Combine assembled transcripts.
 - Parameters: Use Reference Annotation? No; Is this library mate-paired? Single-end.
 - Output: GTF file produced by Cufflinks; Output dataset 'assembled_isoforms' from step 5.
 - Parameters: Use Another GTF file produced by Cufflinks? Yes; Reference Annotation: Yes; Reference Annotation: Output dataset 'output' from step 3; Ignore reference transcripts that are not overlapped by any transcript in input files: True.
- Step 7: Cuffdiff**
 - Use initial read set with assembled transcripts to compute FPKM and differential data.
 - Parameters: Transcripts: Output dataset 'transcripts_combined' from step 6; SAM file of aligned RNA-Seq reads: Output dataset 'output' from step 1; SAM file of aligned RNA-Seq reads: Output dataset 'output' from step 2; False Discovery Rate: 0.05; Min SAM Mapping Quality: 0; Min Alignment Count: 0; Is this library mate-paired? Single-end.
- Step 8: Filter**
 - Filter out isoforms whose confidence interval includes 0.
 - Parameters: Filter: Output dataset 'isoforms_exp' from step 7; With following condition: `c8>0`.

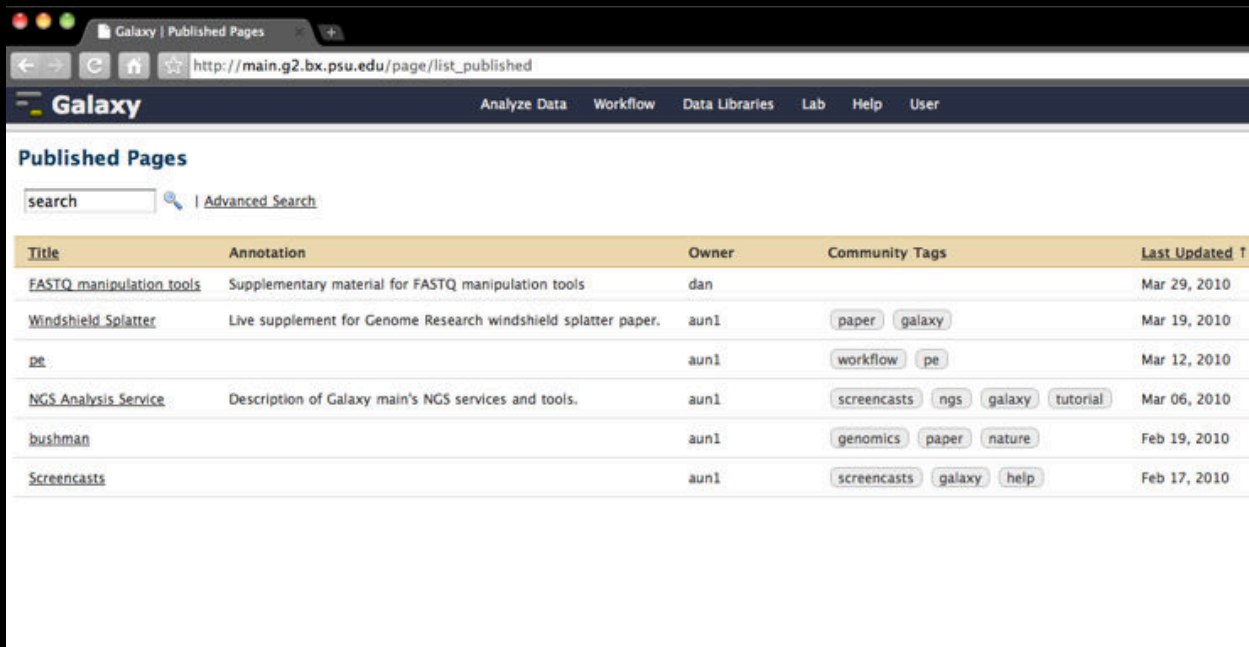
The right sidebar contains "About this Workflow" information, including the author (jgoecks), related workflows, and tags (cufftools, rna-seq).

Item is displayed in webpage

Community tags

Links to related items, public repositories

Public Repositories



The screenshot shows a web browser window with the URL http://main.g2.bx.psu.edu/page/list_published. The page title is "Galaxy" and the navigation menu includes "Analyze Data", "Workflow", "Data Libraries", "Lab", "Help", and "User". The main content area is titled "Published Pages" and features a search bar and a link to "Advanced Search". Below this is a table with the following data:

Title	Annotation	Owner	Community Tags	Last Updated ↑
FASTQ manipulation tools	Supplementary material for FASTQ manipulation tools	dan		Mar 29, 2010
Windshield Splatter	Live supplement for Genome Research windshield splatter paper.	aun1	paper galaxy	Mar 19, 2010
pe		aun1	workflow pe	Mar 12, 2010
NGS Analysis Service	Description of Galaxy main's NGS services and tools.	aun1	screencasts ngs galaxy tutorial	Mar 06, 2010
bushman		aun1	genomics paper nature	Feb 19, 2010
Screencasts		aun1	screencasts galaxy help	Feb 17, 2010

Where published items live

- searchable

Local to a Galaxy instance

For histories, workflows, visualizations, and Pages

Reproducibility and Transparency in Galaxy

Workflows ~ repeating analyses

Display Framework ~ sharing, viewing

Annotations & Tags ~ explanations, context

Pages ~ communicating and publishing

Annotations

32: Filter on data 21 👁 🗑 🔗
31 lines, format: tabular,
database: mm9
Info: Filtering with $c8 > 0$,
kept 68.89% of 46 lines.
Skipped 1 invalid lines starting at
line #1: "test_id gene locus status
value_1 value_2 log(fold_change)
test_stat p_value significant"

Annotation:
Isoforms with non-zero
FPKM for Sample 2.

1	2	3
TCONS_00000001	- chr1:5083438-50835	
TCONS_00000002	- chr1:5085696-50857	
TCONS_00000003	- chr1:5114537-51152	
TCONS_00000005	- chr1:6251091-62511	
TCONS_00000006	- chr1:6251951-62651	
TCONS_00000008	- chr1:6251951-62651	

Tool: Filter

Filter
Data input 'input' (tabular)
With following condition ▼
c8>0

Edit Step Attributes

Annotation / Notes:
Filter out isoforms whose
confidence interval includes 0.

Add an annotation or notes to this step;
annotations are available when a workflow
is viewed.

Notes about an item or step

- text and HTML

Useful for making analyses
easier to understand

- provides context
- explains details

For histories, history steps,
workflows, workflow steps,
visualizations, and Pages

Tags

32: Filter on data 21 👁 0 ✕

31 lines, format: tabular,
database: mm9
Info: Filtering with $c8 > 0$,
kept 68.89% of 46 lines.
Skipped 1 invalid lines starting at
line #1: "test_id gene locus status
value_1 value_2 log(fold_change)
test_stat p_value significant"

Tags:

sample_no:2 ✕ mm9 ✕
cufftools ✕

demo

1	2	3
TCONS_00000001	- chr1:5083438-50835	
TCONS_00000002	- chr1:5085696-50857	
TCONS_00000003	- chr1:5114537-51528	
TCONS_00000005	- chr1:6251091-62511	
TCONS_00000006	- chr1:6251951-62657	
TCONS_00000008	- chr1:6251951-62657	

Edit Workflow Attributes

Name:
Demo Workflow 2

Tags:

cufftools ✕
rna_seq

Apply tags to make it easy to search for and
find items with the same tag.

Annotation / Notes:
None
Add an annotation or notes to a workflow;
annotations are available when a workflow is
viewed.

Short words or phrases that describe an item

- hierarchical
- key-value
- individual, community

Useful for metadata, search, reuse

For histories, datasets, visualizations, workflows, and Pages

Reproducibility and Transparency in Galaxy

Workflows ~ repeating analyses

Display Framework ~ sharing, viewing

Annotations & Tags ~ explanations, context

Pages ~ communicating and publishing

Pages

The screenshot displays a Galaxy web interface. At the top, the browser address bar shows the URL: <http://main.g2.bx.psu.edu/u/aun1/p/windshield-splatter>. The page title is "Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement". The author is listed as "aun1". The page includes a list of authors: SERGEI KOSAKOVSKY POND^{1,2,*}, SAMIR WADHAWAN^{3,6*}, FRANCESCA CHIAROMONTE⁴, GURUPRASAD ANANDA^{1,3}, WEN-YU CHUNG^{1,3,7}, JAMES TAYLOR^{1,5}, ANTON NEKRUTENKO^{1,3} and THE GALAXY TEAM^{1*}. A section titled "How to use this document" explains that it is a live copy of supplementary materials for a manuscript, providing access to exact analyses and workflows. Below this, a "Galaxy History" section titled "Galaxy vs MEGAN" shows a comparison of Galaxy vs. MEGAN pipeline. It lists seven datasets with their annotations: 1: s1 (Results of comparison of Dataset 1 from Huson et al. 2007 against the nt database), 2: s234 (Results of comparison of Datasets 2, 3, and 4 from Huson et al. 2007 against the nt database), 3: s1_max_bit_score (Results of comparison of Dataset 1 from Huson et al. 2007 against the nt database), 4: s234_max_bit_score (Blast hits for Datasets 2, 3, and 4 grouped by maximum bitscore reported by megablast), 5: Join two Queries on data 3 and data 1 (Here blast results for dataset s1 are joined with maximum bitscore for each read), 6: Join two Queries on data 4 and data 2 (Here blast results for datasets s1, 2, and 3 are joined with maximum bitscore for each read), and 7: s1 within 5% of max (This dataset contains blast hits +/- 5% of the maximum bitscore). Below the history, a "Galaxy Workflow" section titled "metagenomic analysis" shows a generic workflow for performing a metagenomic analysis on NGS data. It lists two steps: Step 1: Input dataset (454 Quality Dataset) and Step 2: Input dataset (454 Reads).

Web-based documents that communicate a complete analysis

- multiple levels of detail

Support viewing, reproduction, and component reuse

Perfect for online supplement

Page Editor

Galaxy
http://main.g2.bx.psu.edu/page/edit_content?id=faf52b7a99ec14da

Galaxy
Analyze Data Workflow Data Libraries Lab Admin Help User

Page Editor | Title: Windshield Splatter Save Close

Paragraph type Insert Link to Galaxy Object Embed Galaxy Object

Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement

SERGEI KOSAKOVSKY POND^{1,2,*}, SAMIR WADHAWAN^{3,6*}, FRANCESCA CHIAROMONTE⁴, GURUPRASAD ANANDA^{1,3}, WEN-YU CHUNG^{1,3,7}, JAMES TAYLOR^{1,5}, ANTON NEKRUTENKO^{1,3} and THE GALAXY TEAM^{1*}

Correspondence should be addressed to SKP, JT, or AN.

How to use this document

This document is a live copy of supplementary materials for [the manuscript](#). It provides access to the **exact** analyses and workflows discussed in the paper, so you can play with them by re-running, changing parameters, or even applying them to your own data. Specifically, we provide the two histories and one workflow found below. You can view these items by clicking on their name to expand them. You can also import these items into your Galaxy workspace and start using them; click on the green plus to import an item. To import workflows you must [create a Galaxy account](#) (unless you already have one) - a hassle-free procedure where you are only asked for a username and password.

This is the Galaxy history detailing the comparison of our pipeline to MEGAN:

Embedded Galaxy History 'Galaxy vs MEGAN'

[Do not edit this block; Galaxy will fill it in with the annotated history when it is displayed.]

This is the Galaxy history showing a generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and Figure 3A):

Embedded Galaxy History 'metagenomic analysis'

[Do not edit this block; Galaxy will fill it in with the annotated history when it is displayed.]

This is the Galaxy workflow for generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and Figure 3B):

Embedded Galaxy Workflow 'metagenomic analysis'

[Do not edit this block; Galaxy will fill it in with the annotated workflow when it is displayed.]

Supplemental Analysis

Comparison between Galaxy pipeline and Megan

(Use [this link](#) to see Galaxy history representing this analysis. Individual elements of this history are referred to as History Item 1, 2 and so on using bold typeface)

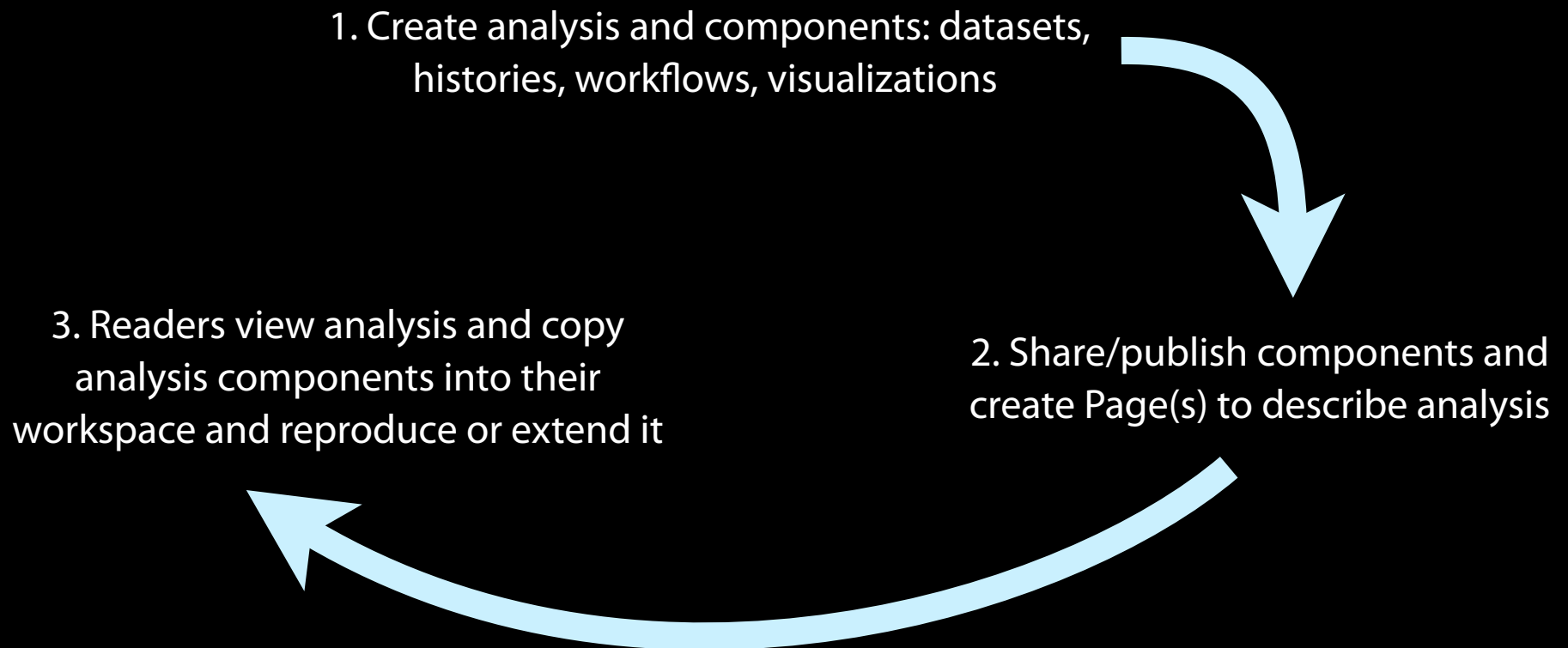
The first step of a homology-based metagenomic analysis is to contrast a collection of sequencing reads against a database whose entries are assigned to taxonomic ranks. Following the procedure of (Huson et al. 2007) we used the non-redundant protein database (NR) from the National Center for Biotechnology Information. There are several avenues for importing large sets of alignments into Galaxy. First, alignments can be generated directly within Galaxy (see the following section). Alternatively, alignments generated elsewhere (e.g., using local BLAST installations of web-based resources such as CAMERA (Seshadri et al. 2007); see below) can be uploaded in either tab-delimited or XML format. To demonstrate this functionality, we generated alignments in BLAST XML format outside of Galaxy using the BLASTx program of the BLAST package (Altschul et al. 1990) and then uploaded them into Galaxy's history. Galaxy includes a parser for XML generated by BLAST programs that produces a tab-delimited format that can be easily used in downstream analyses. Only 243 (or ~2% from 3,812,372 alignments) and 1,192 (or ~11% from 3,581,932 alignments) reads from samples 1 and 2-4, respectively (History Items 1 and 2), did not produce matches against the NR database. These counts were slightly higher than those reported in Huson et al. because we set the BLAST E value flag (-e) to 0.01 instead of the default value of 10 (used in (Huson et al. 2007)) removing many weakly supported alignments and significantly decreasing the size of the resultant file. Similarly to Huson and colleagues we further filtered BLAST alignments by retaining only those hits that were within 5% of the best score for every read using a combination of Galaxy tools (History Items 3 - 8. Here we first selected lines with the highest bit score per read (History Items 3 and 4). Next we joined these lines with the original files using the join tool (History Items 5 and 6). Finally, we selected those lines from datasets 5 and 6 where the bit score was within 5% of the maximum (History Items 7 and 8)). This significantly reduced number of hits to 54,458 and 62,647 in samples 1 and 2-4, respectively, although the number of reads producing these hits did not change (9,757 and 8,808 reads, respectively).

Because every entry within the NR database is assigned a taxonomy id, it is straightforward to create a phylogenetic profile of every read that aligns against a database sequence. Galaxy features the Fetch Taxonomic Ranks tool that quickly parses NCBI taxonomy and writes out a taxonomic string consisting of 21 taxonomic ranks from superkingdom to subspecies. Application of this tool to filtered BLAST hits produced 54,458 and 62,647 taxonomic strings for samples 1 and 2-4, respectively (History Items 9 and 10). Note that because the number of taxonomic strings greatly surpasses the number of sequencing reads (9,757 and 8,808, respectively), each read is likely represented by multiple phylogenetic profiles. As a

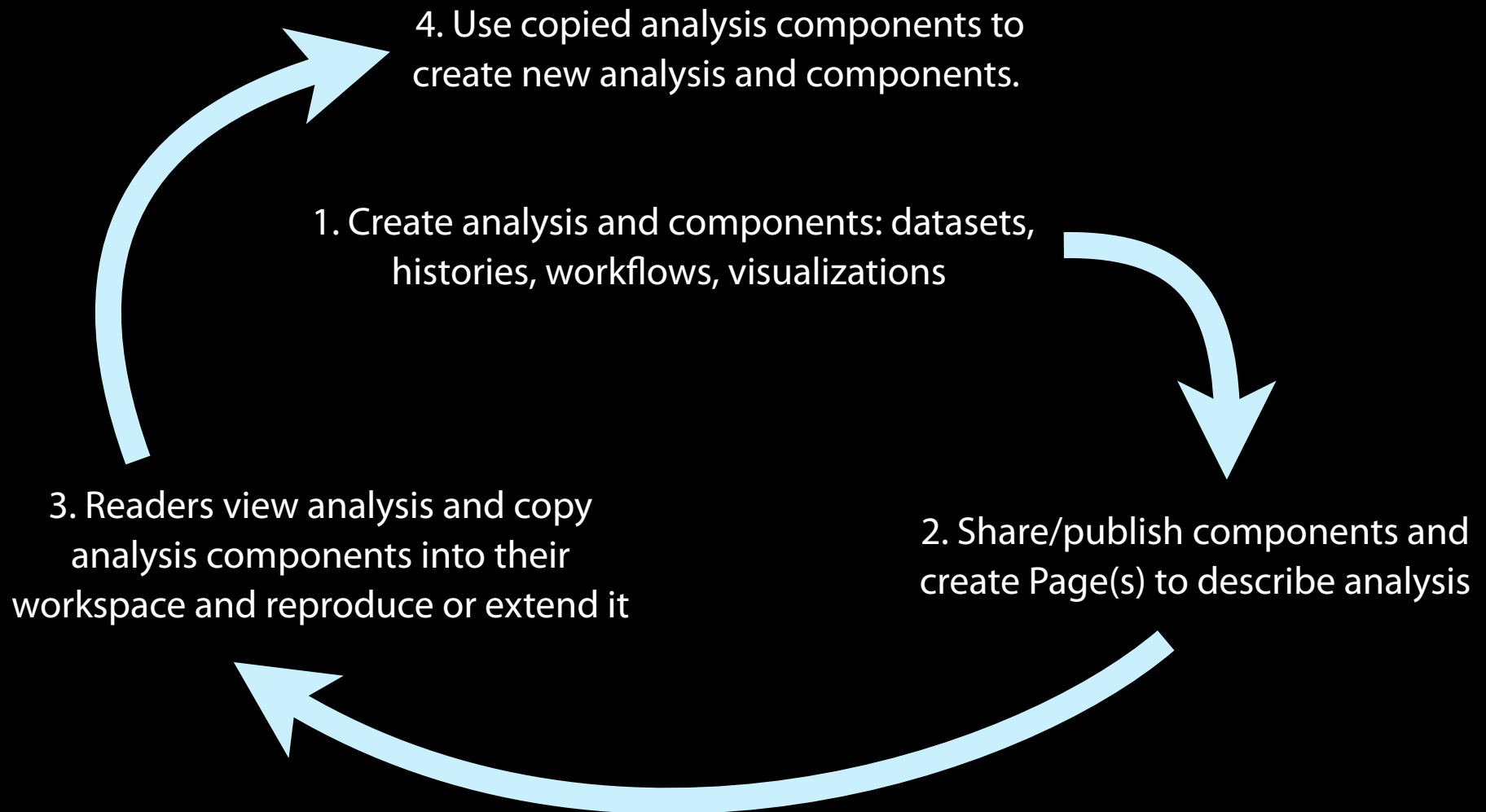
WYSIWYG editor for
HTML + Galaxy objects

Can embed or link to
datasets, histories,
workflows, and (almost)
visualizations

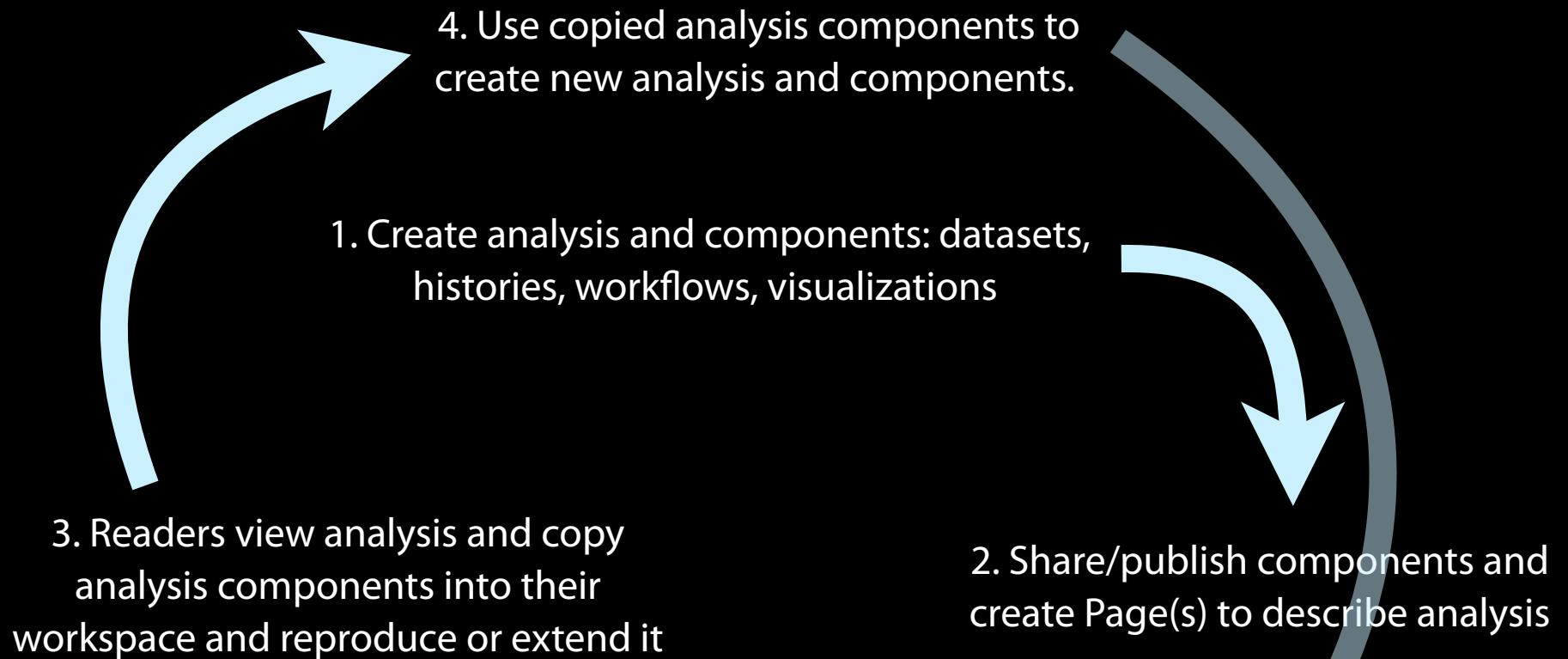
Revisiting Reproducibility and Transparency: The Analysis Lifecycle



Revisiting Reproducibility and Transparency: The Analysis Lifecycle



Revisiting Reproducibility and Transparency: The Analysis Lifecycle



Next Steps

Make published items independent of server

- community space for workflows, histories, pages
- long-term archival, e.g. Dryad

Developing best practices

- usage, ratings, reviews, and comments
- provenance (attribution) for all objects

Thanks! Questions?

<http://usegalaxy.org/>