# From the Ground to the Cloud in 25 Minutes: Building a Customized Galaxy Analysis Server Using Only a Web Browser

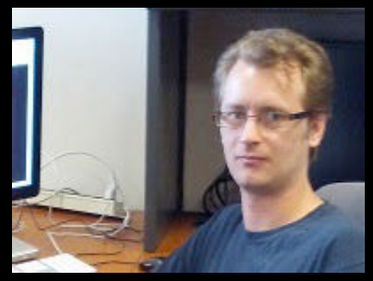
Daniel Blankenberg
Postdoctoral Research Associate
**The Galaxy Team | Nekrutenko Lab@Penn State**
http://UseGalaxy.org

# The Galaxy Team

Enis Afgan

Dannon Baker

Dan Blankenberg



Dave Bouvier

Marten Čech

John Chilton

Dave Clements

Nate Coraor

Carl Eberhard

Jeremy Goecks

Sam Guerler

Jen Jackson

Ross Lazarus

Anton Nekrutenko

Nick Stoler

James Taylor

Greg Von Kuster

http://wiki.galaxyproject.org/GalaxyTeam

# Overview

## What is Galaxy?

## ToolShed

## Data Managers

## Demo

# Overview

## What is Galaxy?

## ToolShed

## Data Managers

## Demo

# Galaxy Project Mission

**Galaxy** is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research.

**Accessible**: Users without programming experience can easily specify parameters and run tools and workflows.

**Reproducible**: Galaxy captures information so that any user can repeat and understand a complete computational analysis.

**Transparent**: Users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis.

# What is Galaxy?

A **data analysis and integration** tool

**A free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

**Open source software** that makes integrating your own tools and data and customizing for your own site simple

There are several **ways to use Galaxy**

# Using Galaxy - 4 ways

- **Public Main** Galaxy web instance: ***usegalaxy.org***
- **Local** instance: ***getgalaxy.org***
- **Cloud** instance: ***usegalaxy.org/cloud***
- **Other Public** Galaxy web instances hosted by various groups: wiki.galaxyproject.org/PublicGalaxyServers

# Galaxy as a *Genomics WorkBench*

**Dataset:**

Any input, output or intermediate set of data + metadata. A record of a specific data or analysis step.

**History:**

A series of inputs, analysis steps, intermediate datasets, and outputs. A record of a group of data and analysis steps.

**Tool:**

An operation within Galaxy that acts upon dataset(s) as an analysis step. May be developed by Galaxy team or a 3rd party program that has been "wrapped" for Galaxy.

**Workflow:**

A series of analysis steps executed as a unit.

# More Galaxy Terminology

**Share:**

Make something available to someone else

**Publish:**

Make something available to everyone

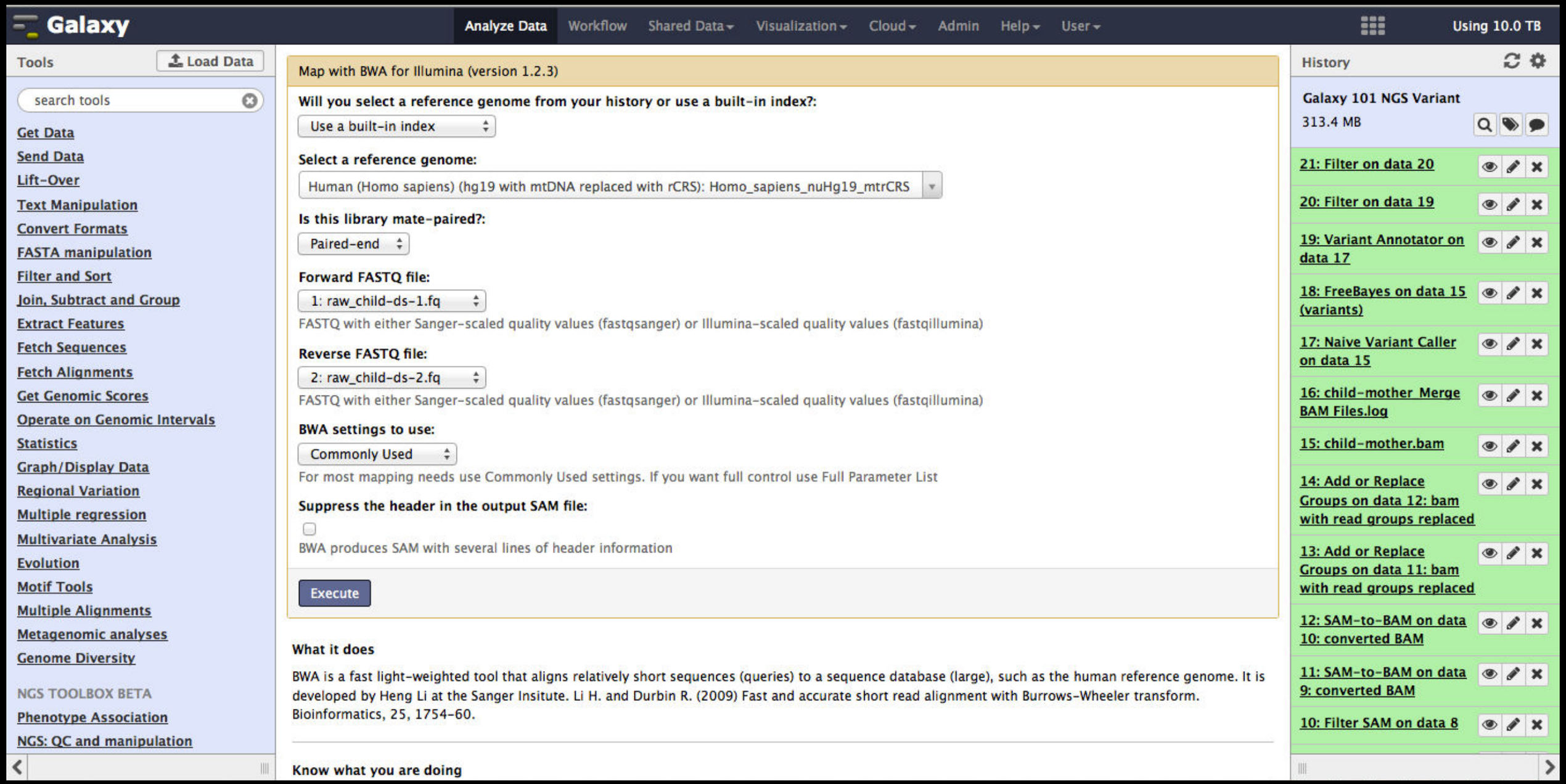
**Galaxy Page:**

Analysis documentation within Galaxy; easy to embed and link to any Galaxy object (histories, datasets, workflows, visualization) or external resource (video, graphics, publications).

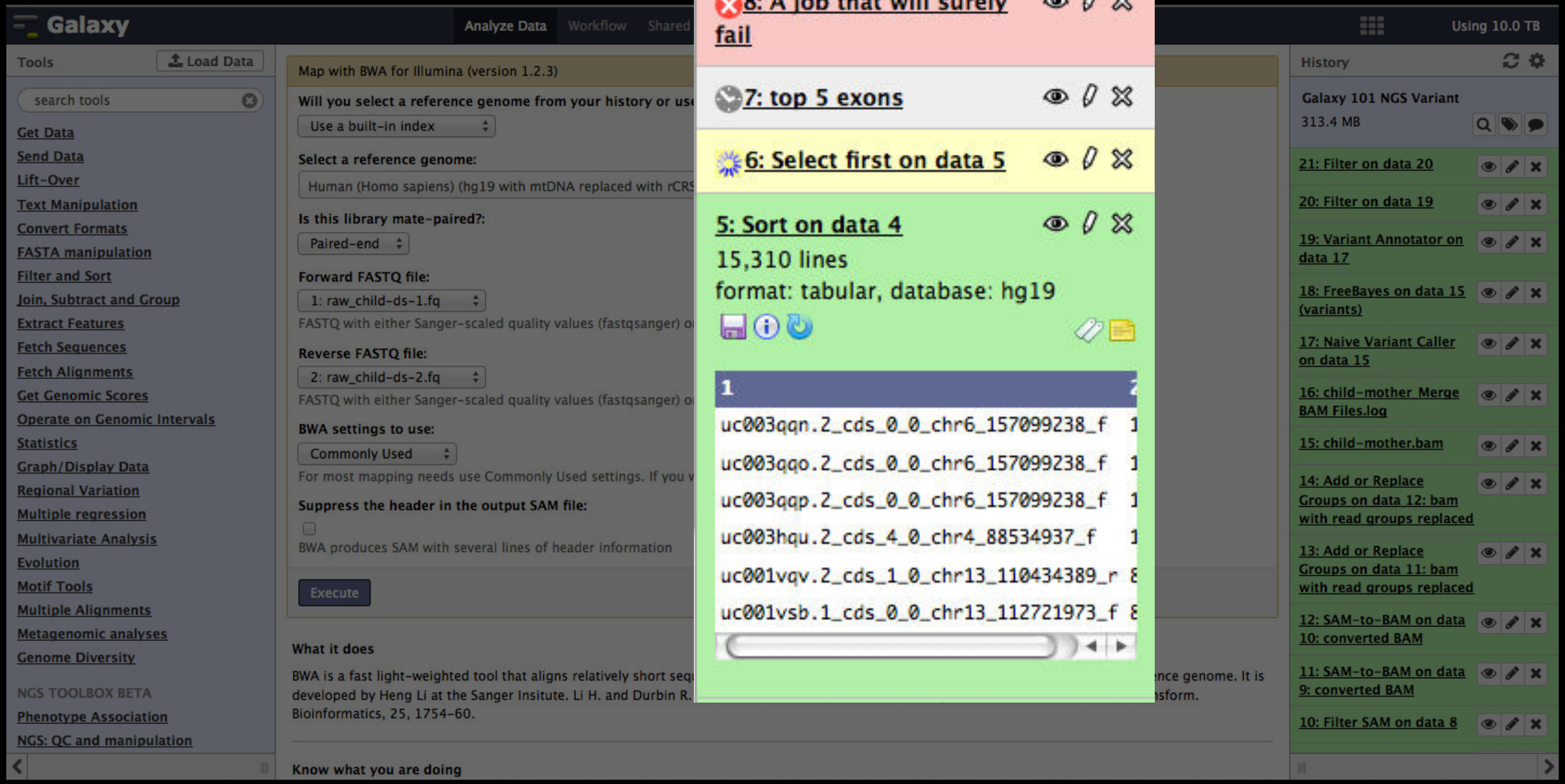
**Visualize:**

External resources. Trackster. Galaxy Charts (D3/NVD3).

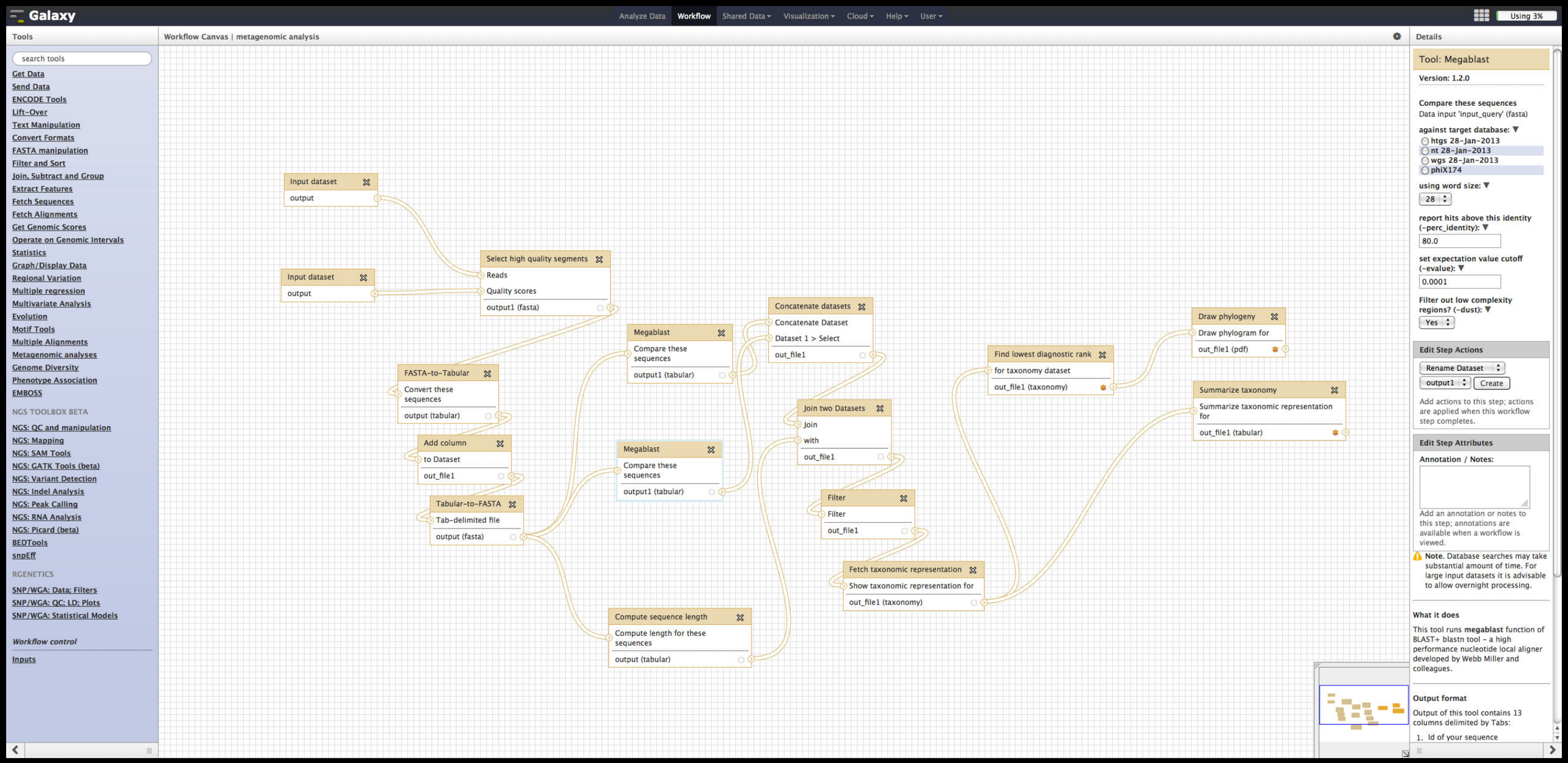
# Galaxy Analysis Workspace



# Galaxy Analysis Workspace



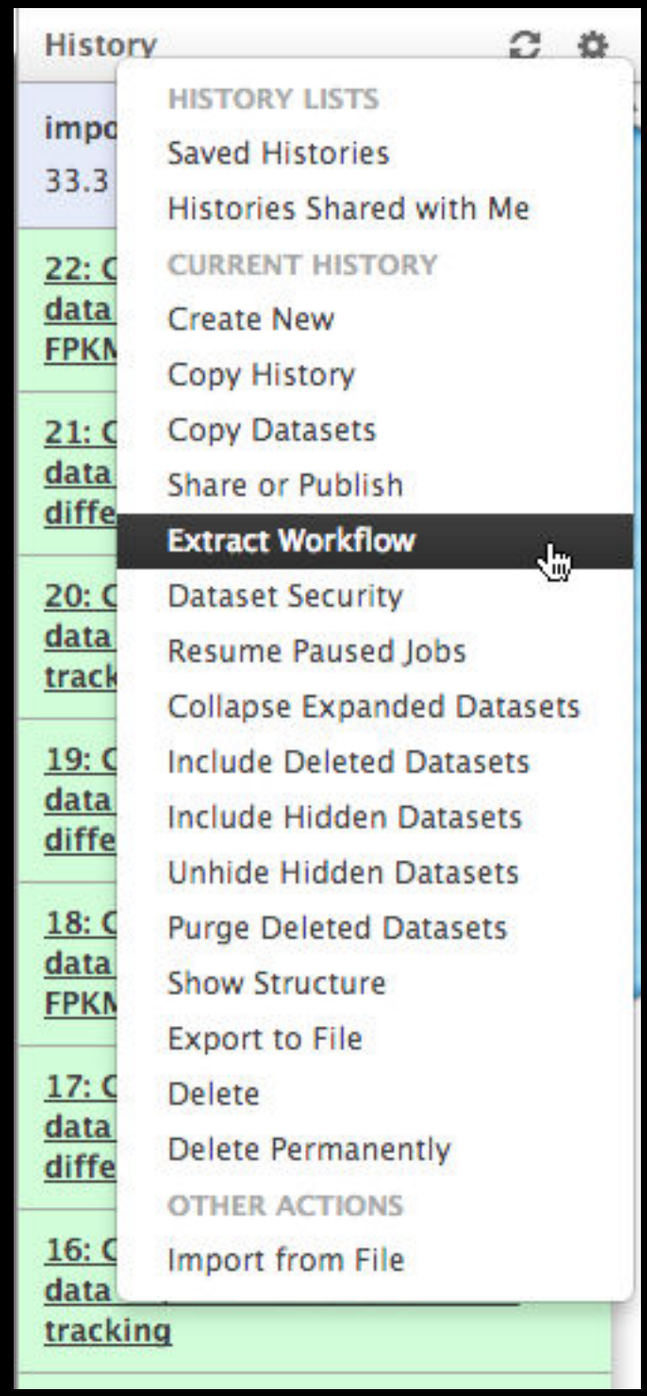
# Workflow Editor



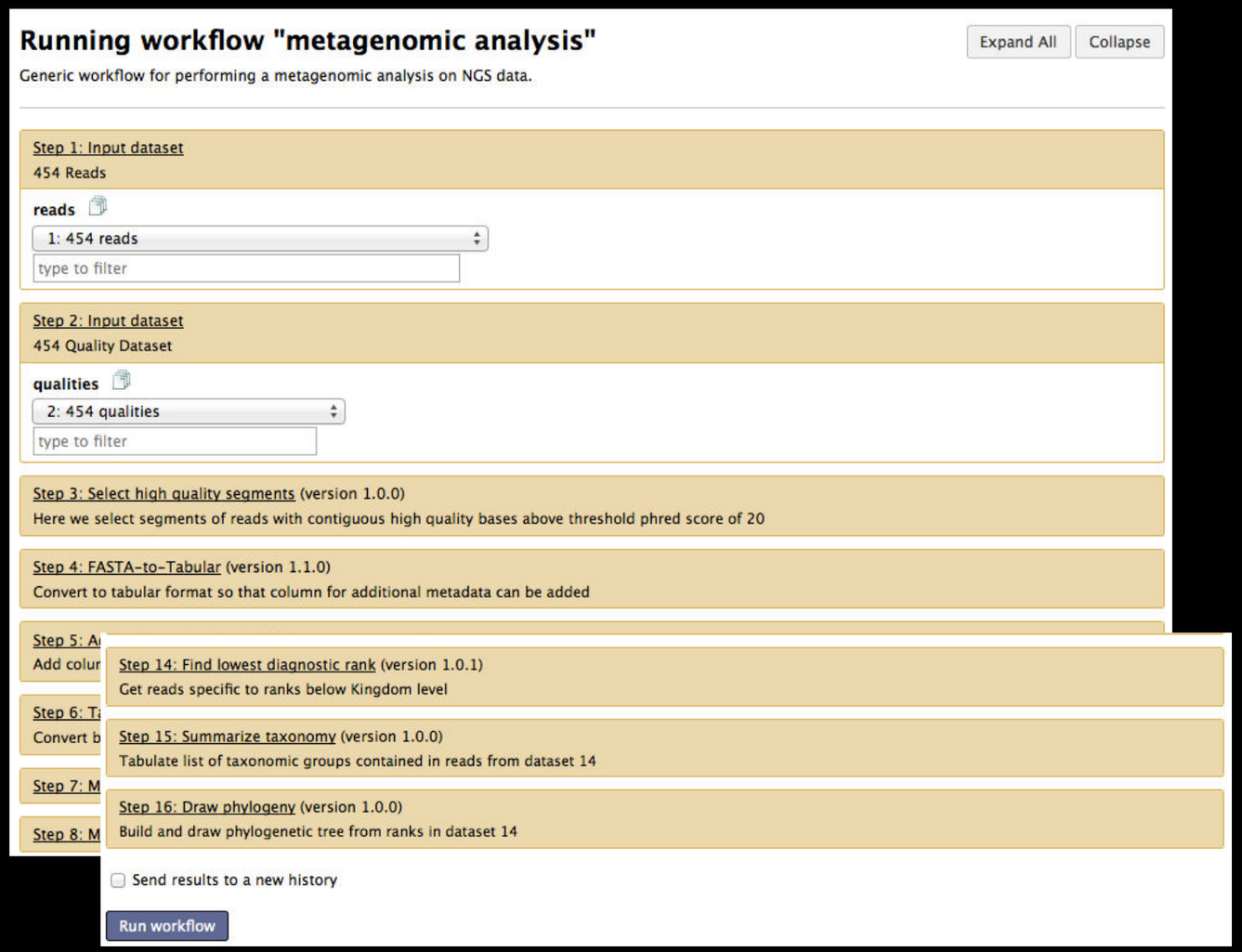
# Create Workflow Automatically

## Extract Workflow from History

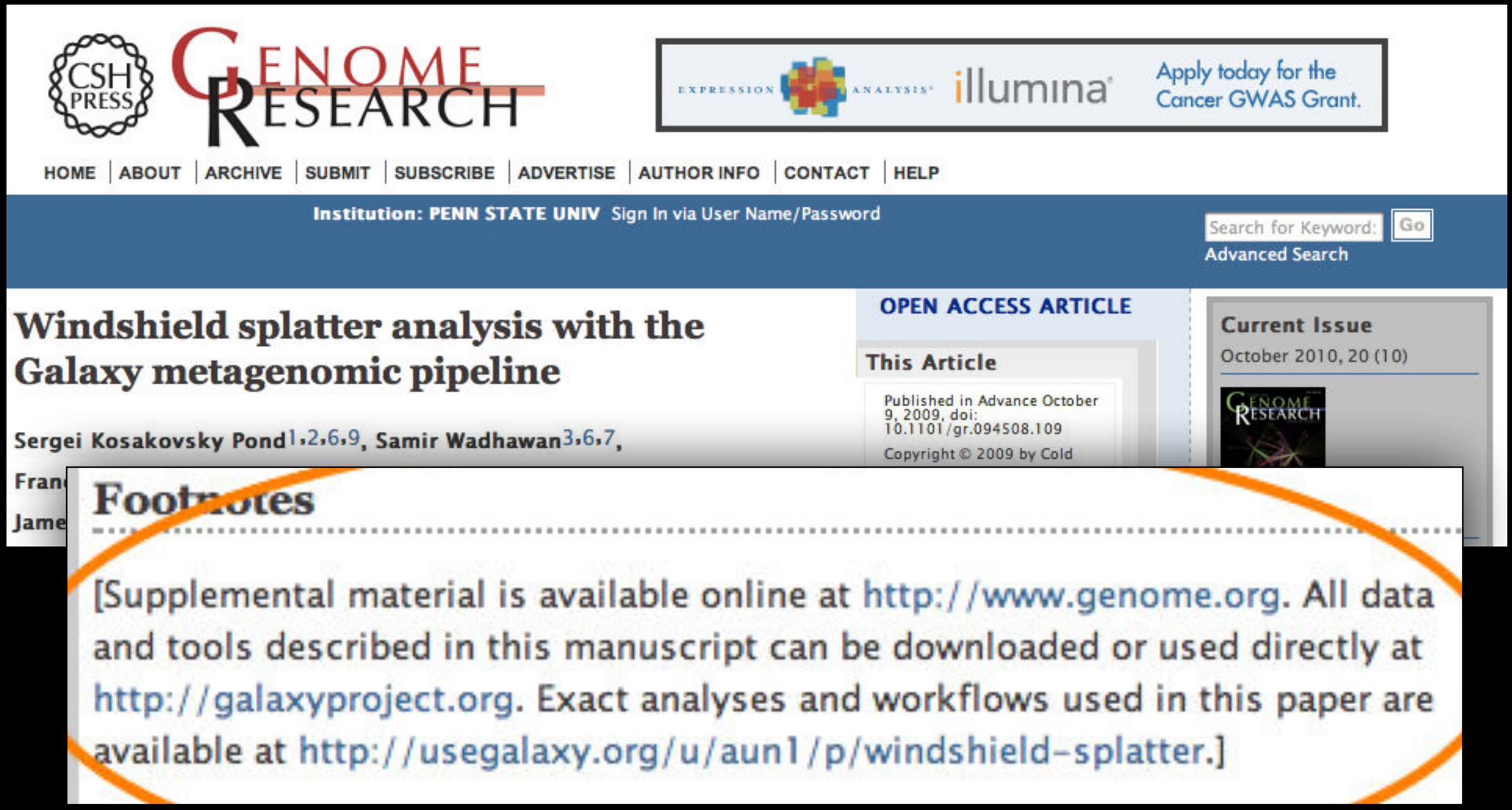
Create a workflow from a History that you created interactively.




## Run it



# Sharing and Publishing Your Work




**Histories, workflows, visualizations** and ***pages*** can be shared with others or published to the world.

http://usegalaxy.org/u/aun1/p/windshield-splatter

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

SERGEI KOSAKOVSKY POND[1,2,*], SAMIR WADHAWAN[3,6*], FRANCESCA CHIAROMONTE[4], 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 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:

**Galaxy History | Galaxy vs MEGAN**
**Comparison of Galaxy vs. MEGAN pipeline.**

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**):

**Galaxy History | metagenomic analysis**

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**):

**Galaxy Workflow | metagenomic analysis**
**Generic workflow for performing a metagenomic analysis on NGS data.**

## Accessing the Data

Windshield Splatter datasets analyzed in this manuscript can be accessed through this Galaxy Library. From



### Author

aun1

### Related Pages

All published pages
Published pages by aun1

### Rating

Community
(6 ratings, 5.0 average)

### Tags

Community:
paper galaxy megan

http://usegalaxy.org/u/aun1/p/windshield-splatter

# Overview

## What is Galaxy?

## ToolShed

## Data Managers

## Demo


Enables sharing of Galaxy Utilities:

- tools
- proprietary datatypes
- exported Galaxy workflows

Automatically install tools and tool suites, and their dependencies, into a Galaxy instance

Galaxy Utilities can be created and shared by any member of the community

https://wiki.galaxyproject.org/ToolShed

Galaxies on private clouds

Galaxies on public clouds

http://usegalaxy.org

**Galaxy Tool Shed**

http://usegalaxy.org/toolshed

...

private Tool Sheds

1

2

3

...

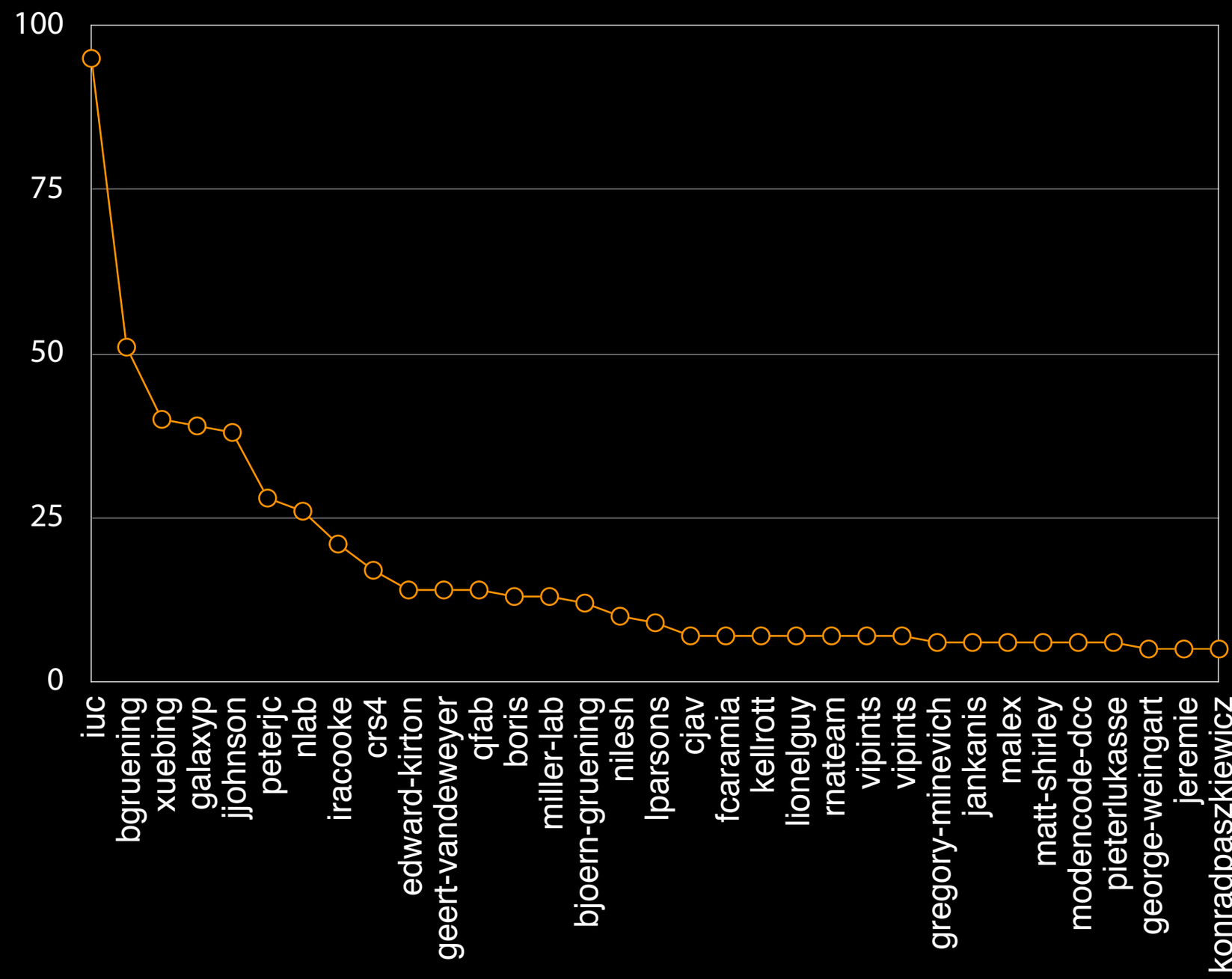
∞

**private Galaxy installations**

# toolshed statistics

- 897 repositories
- 222 unique owners
- 176 Tool dependency package installation recipes
- 2,330 valid tools
- 3,420 valid versions of tools
- 54 exported Galaxy workflows
- 455 custom datatypes
- 62,021 total repository installations

# toolshed contributions



# Overview

What is Galaxy?

ToolShed

Data Managers

Demo

# Built-in Datasets

## BWA example

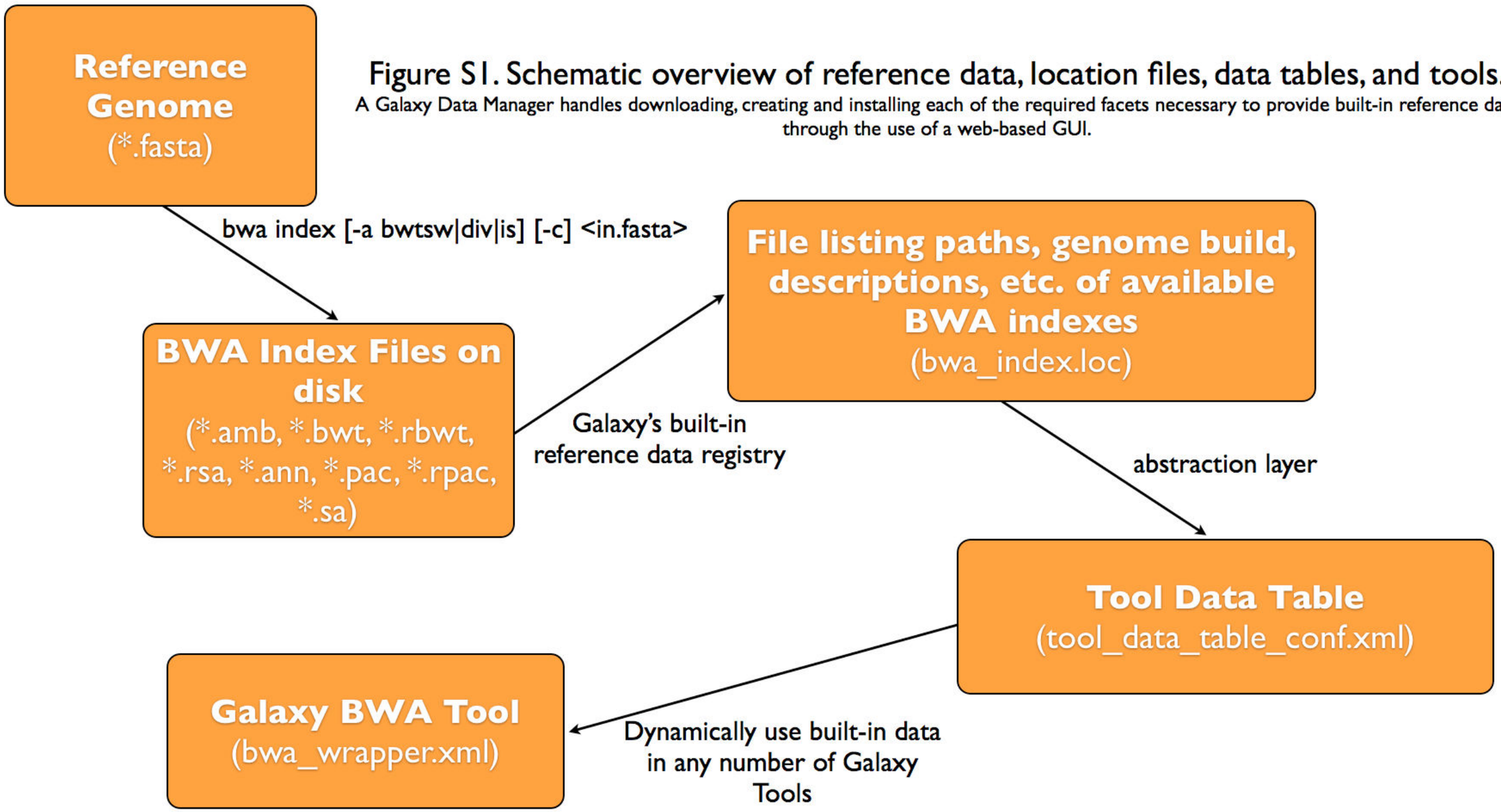



Figure S1. Schematic overview of reference data, location files, data tables, and tools.
A Galaxy Data Manager handles downloading, creating and installing each of the required facets necessary to provide built-in reference data through the use of a web-based GUI.

Blankenberg, et al. Wrangling Galaxy's reference data. Bioinformatics. 2014 Jul 1.;30(13):1917-9.

# Data Managers

Allows for the creation of built-in (reference) data

underlying data

data tables

*.loc files

Specialized Galaxy tools that can only be accessed by an admin

Defined locally or installed from ToolShed

# Data Managers

Flexible Framework

- not just Genomic data
- Interactively Run Data Managers through UI
- Workflow compatible
- API

Examples:

- Adding New genome builds (dbkeys)
- Fetching Genome (FASTA) sequences
- Building short read mapper indexes for genomes

# Overview

## What is Galaxy?

## ToolShed

## Data Managers

## Demo

# Demo

- Fetch the Genome Sequence for sacCer2
  - UCSC as the source
  - Install fetching tool from ToolShed
  - define new Genome build / dbkey
  - all_fasta & __dbkeys__ tables are populated automatically
- Build BWA indexes for sacCer2
  - Install indexing tool from ToolShed
  - Build indexes
  - bwa_index table is populated automatically
- Align some reads to the newly added reference genome
- Call Variants with FreeBayes
  - Install Samtools Indexer from ToolShed
  - Build Samtools dictionary
  - Install and Run FreeBayes

# Data Manager Demo: Full Disclosure

Fresh Install of galaxy-dist stable within CloudMan on AWS

- Removed Many pre-installed tools and data
- The sequencing reads are a small subset from SRR507778, originally downloaded from EBI SRA.
- Already added myself as an admin
- Connected AWS instance to an elastic IP with an A record on my domain

http://ismb2014.dblankenberg.org/

# Demo: Getting Started

- Register User and make admin
- Upload FASTQ files:
  - http://dblankenberg.org/examples/fastq/SRR507778-10k_1.fastqsanger
  - http://dblankenberg.org/examples/fastq/SRR507778-10k_2.fastqsanger
- Install and Run FastQC

# Demo: Adding a new Genome Build

- Look at "Manage Local Data" in Admin interface
- Install Genome Build Data Manager
- Add a new dbkey using sacCer2 from UCSC

# Demo: Mapping Reads with BWA

- Install BWA mapping tool and Index builder
- Build BWA index for sacCer2
- Map reads using BWA

# Demo: Calling Variants with Freebayes

Freebayes requires samtools fasta index

- Install Samtools indexing tool
- Build Samtools index for sacCer2
- Install and Run Freebayes

# GCC2015

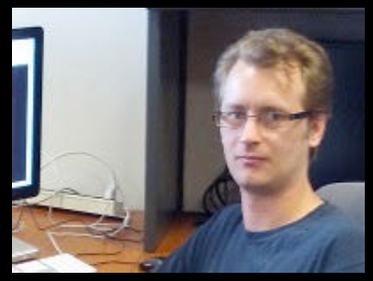
2015 Galaxy Community Conference will be held in Norwich, United Kingdom, at The Sainsbury Lab, 6-8th July

# The Galaxy Team

Enis Afgan

Dannon Baker

Dan Blankenberg



Dave Bouvier

Marten Čech

John Chilton

Dave Clements

Nate Coraor

Carl Eberhard

Jeremy Goecks

Sam Guerler

Jen Jackson

Ross Lazarus

Anton Nekrutenko

Nick Stoler

James Taylor

Greg Von Kuster

http://wiki.galaxyproject.org/GalaxyTeam