

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



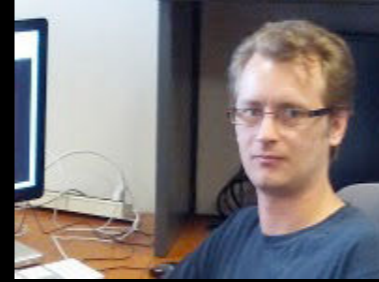
Enis Afgan



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

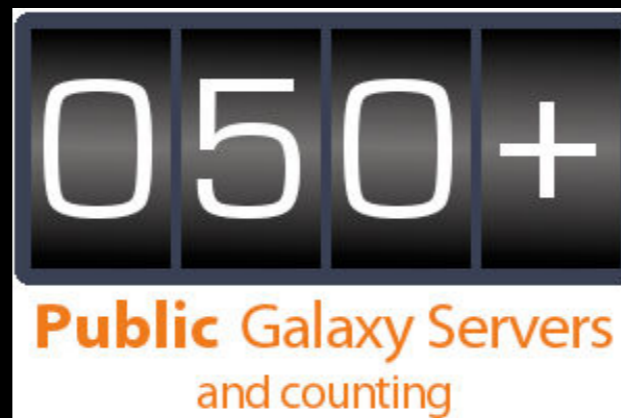
A **free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

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

The screenshot displays the Galaxy Analysis Workspace interface. At the top, the navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Admin', 'Help', and 'User'. The top right corner shows 'Using 10.0 TB'.

The main workspace is divided into three sections:

- Tools Panel (Left):** A sidebar with a search bar and a list of tool categories including 'Get Data', 'Send Data', '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', 'Genome Diversity', 'NGS TOOLBOX BETA', 'Phenotype Association', and 'NGS: QC and manipulation'. A 'Load Data' button is also present.
- Tool Configuration (Center):** The 'Map with BWA for Illumina (version 1.2.3)' tool is active. The configuration includes:
 - Reference Genome:** 'Human (Homo sapiens) (hg19 with mtDNA replaced with rCRS): Homo_sapiens_nuHg19_mtrCRS'
 - Library Type:** 'Paired-end'
 - Forward FASTQ file:** '1: raw_child-ds-1.fq'
 - Reverse FASTQ file:** '2: raw_child-ds-2.fq'
 - BWA settings:** 'Commonly Used'
 - Suppress the header in the output SAM file:** UncheckedAn 'Execute' button is located at the bottom of the configuration area.
- History Panel (Right):** A list of previous jobs with their names and sizes. The jobs are:
 - Galaxy 101 NGS Variant (313.4 MB)
 - 21: Filter on data 20
 - 20: Filter on data 19
 - 19: Variant Annotator on data 17
 - 18: FreeBayes on data 15 (variants)
 - 17: Naive Variant Caller on data 15
 - 16: child-mother Merge BAM Files.log
 - 15: child-mother.bam
 - 14: Add or Replace Groups on data 12: bam with read groups replaced
 - 13: Add or Replace Groups on data 11: bam with read groups replaced
 - 12: SAM-to-BAM on data 10: converted BAM
 - 11: SAM-to-BAM on data 9: converted BAM
 - 10: Filter SAM on data 8

At the bottom of the interface, there is a status bar with the text 'Know what you are doing'.

Galaxy Analysis Workspace

The screenshot displays the Galaxy Analysis Workspace interface. On the left is a sidebar with various tool categories like 'Get Data', 'Text Manipulation', and 'Filter and Sort'. The main area shows a workflow step 'Map with BWA for Illumina (version 1.2.3)' with configuration options for reference genome, library type, and FASTQ files. An 'Execute' button is visible. On the right, a 'History' panel lists previous jobs, including 'Galaxy 101 NGS Variant' and several 'Filter on data' jobs. A central pop-up window shows a list of jobs: '8: A job that will surely fail' (red), '7: top 5 exons' (white), '6: Select first on data 5' (yellow), and '5: Sort on data 4' (green). The '5: Sort on data 4' job is selected, showing it has 15,310 lines in tabular format for the hg19 database. Below this, a table of genomic coordinates is displayed, with the first row highlighted in blue.

8: A job that will surely fail

7: top 5 exons

6: Select first on data 5

5: Sort on data 4
15,310 lines
format: tabular, database: hg19

1	2
uc003qqn.2_cds_0_0_chr6_157099238_f	1
uc003qqo.2_cds_0_0_chr6_157099238_f	1
uc003qqp.2_cds_0_0_chr6_157099238_f	1
uc003hqu.2_cds_4_0_chr4_88534937_f	1
uc001vqv.2_cds_1_0_chr13_110434389_r	8
uc001vsb.1_cds_0_0_chr13_112721973_f	8

Workflow Editor

Galaxy Analyze Data **Workflow** Shared Data Visualization Cloud Help User Using 3%

Tools Workflow Canvas | metagenomic analysis Details

search 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
- Genome Diversity
- Phenotype Association
- EMBOSS
- NGS TOOLBOX BETA
- NGS: QC and manipulation
- NGS: Mapping
- NGS: SAM Tools
- NGS: GATK Tools (beta)
- NGS: Variant Detection
- NGS: Indel Analysis
- NGS: Peak Calling
- NGS: RNA Analysis
- NGS: Picard (beta)
- BEDTools
- snpEff
- RGENETICS
- SNP/WGA: Data: Filters
- SNP/WGA: QC: LD: Plots
- SNP/WGA: Statistical Models
- Workflow control
- Inputs

```
graph LR; I1[Input dataset] --> S[Select high quality segments]; I2[Input dataset] --> S; S --> F[FASTA-to-Tabular]; S --> M1[Megablast]; S --> M2[Megablast]; F --> A[Add column]; A --> T[Tabular-to-FASTA]; T --> C[Compute sequence length]; T --> M1; T --> M2; C --> M1; C --> M2; M1 --> J[Join two Datasets]; M2 --> J; J --> F1[Filter]; J --> F2[Filter]; F1 --> T1[Fetch taxonomic representation]; F2 --> T1; T1 --> R[Find lowest diagnostic rank]; R --> S1[Summarize taxonomy]; R --> S2[Summarize taxonomic representation for]; S1 --> P1[Draw phylogeny]; S2 --> P1; P1 --> P2[Draw phylogram for];
```

Tool: Megablast
Version: 1.2.0

Compare these sequences
Data input 'input_query' (fasta)

against target database: ▼
 htgs 28-Jan-2013
 nt 28-Jan-2013
 wgs 28-Jan-2013
 phiX174

using word size: ▼
28

report hits above this identity (-perc_identity): ▼
80.0

set expectation value cutoff (-evalue): ▼
0.0001

Filter out low complexity regions? (-dust): ▼
Yes

Edit Step Actions
Rename Dataset
output1 Create

Add actions to this step; actions are applied when this workflow step completes.

Edit Step Attributes

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

Note: Database searches may take substantial amount of time. For large input datasets it is advisable to allow overnight processing.

What it does
This tool runs megablast function of BLAST+ blastn tool - a high performance nucleotide local aligner developed by Webb Miller and colleagues.

Output format
Output of this tool contains 13 columns delimited by Tabs:
1. Id of your sequence

Create Workflow Automatically

Extract Workflow from History


Create a workflow from a History that you created interactively.

Run it

Running workflow "metagenomic analysis" Expand All Collapse


Generic workflow for performing a metagenomic analysis on NGS data.

Step 1: Input dataset
454 Reads

reads 

1: 454 reads

Step 2: Input dataset
454 Quality Dataset

qualities 

2: 454 qualities

Step 3: Select high quality segments (version 1.0.0)
Here we select segments of reads with contiguous high quality bases above threshold phred score of 20

Step 4: FASTA-to-Tabular (version 1.1.0)
Convert to tabular format so that column for additional metadata can be added

Step 5: Add color

Step 6: Tabulate
Convert to tabular format

Step 7: Merge

Step 8: Merge



Step 14: Find lowest diagnostic rank (version 1.0.1)
Get reads specific to ranks below Kingdom level

Step 15: Summarize taxonomy (version 1.0.0)
Tabulate list of taxonomic groups contained in reads from dataset 14

Step 16: Draw phylogeny (version 1.0.0)
Build and draw phylogenetic tree from ranks in dataset 14

Send results to a new history


Run workflow

History  

HISTORY LISTS

- Saved Histories
- Histories Shared with Me

CURRENT HISTORY

- Create New
- Copy History
- Copy Datasets
- Share or Publish
- Extract Workflow** 
- Dataset Security
- Resume Paused Jobs
- Collapse Expanded Datasets
- Include Deleted Datasets
- Include Hidden Datasets
- Unhide Hidden Datasets
- Purge Deleted Datasets
- Show Structure
- Export to File
- Delete
- Delete Permanently

OTHER ACTIONS

- Import from File

Sharing and Publishing Your Work

The screenshot shows the top of a Genome Research article page. At the top left is the CSH PRESS logo and the 'GENOME RESEARCH' title. To the right is an advertisement for Illumina's Cancer GWAS Grant. Below the header is a navigation menu with links: HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTACT | HELP. A blue bar below the menu contains the text 'Institution: PENN STATE UNIV Sign In via User Name/Password' and a search box with 'Search for Keyword: Go' and 'Advanced Search' options. The main article title is 'Windshield splatter analysis with the Galaxy metagenomic pipeline' by Sergei Kosakovsky Pond and Samir Wadhawan. A 'Footnotes' section is highlighted with an orange oval, containing a paragraph about supplemental material availability. To the right of the article, there is a 'Current Issue' section for October 2010 and an 'OPEN ACCESS ARTICLE' section with publication details.

CSH PRESS GENOME RESEARCH

EXPRESSION ANALYSIS[®] illumina[®] Apply today for the Cancer GWAS Grant.

HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTACT | HELP

Institution: PENN STATE UNIV Sign In via User Name/Password

Search for Keyword: Go
Advanced Search

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond^{1,2,6,9}, Samir Wadhawan^{3,6,7},
Fran
Jame

Footnotes

[Supplemental material is available online at <http://www.genome.org>. All data and tools described in this manuscript can be downloaded or used directly at <http://galaxyproject.org>. Exact analyses and workflows used in this paper are available at <http://usegalaxy.org/u/aun1/p/windshield-splatter>.]

OPEN ACCESS ARTICLE

This Article

Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109
Copyright © 2009 by Cold

Current Issue
October 2010, 20 (10)

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 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:

[+](#) **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** [+](#) [↗](#)

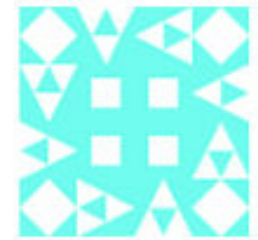
This is the Galaxy workflow for generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3B**):

[+](#) **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 there, they can be re-analyzed through Galaxy using the above workflows or downloaded.

About this Page



Author

aun1

Related Pages

[All published pages](#)
[Published pages by aun1](#)

Rating

Community
(6 ratings, 5.0 average)



Tags

Community:

- paper
- galaxy
- megan

Overview

What is Galaxy?

ToolShed

Data Managers

Demo

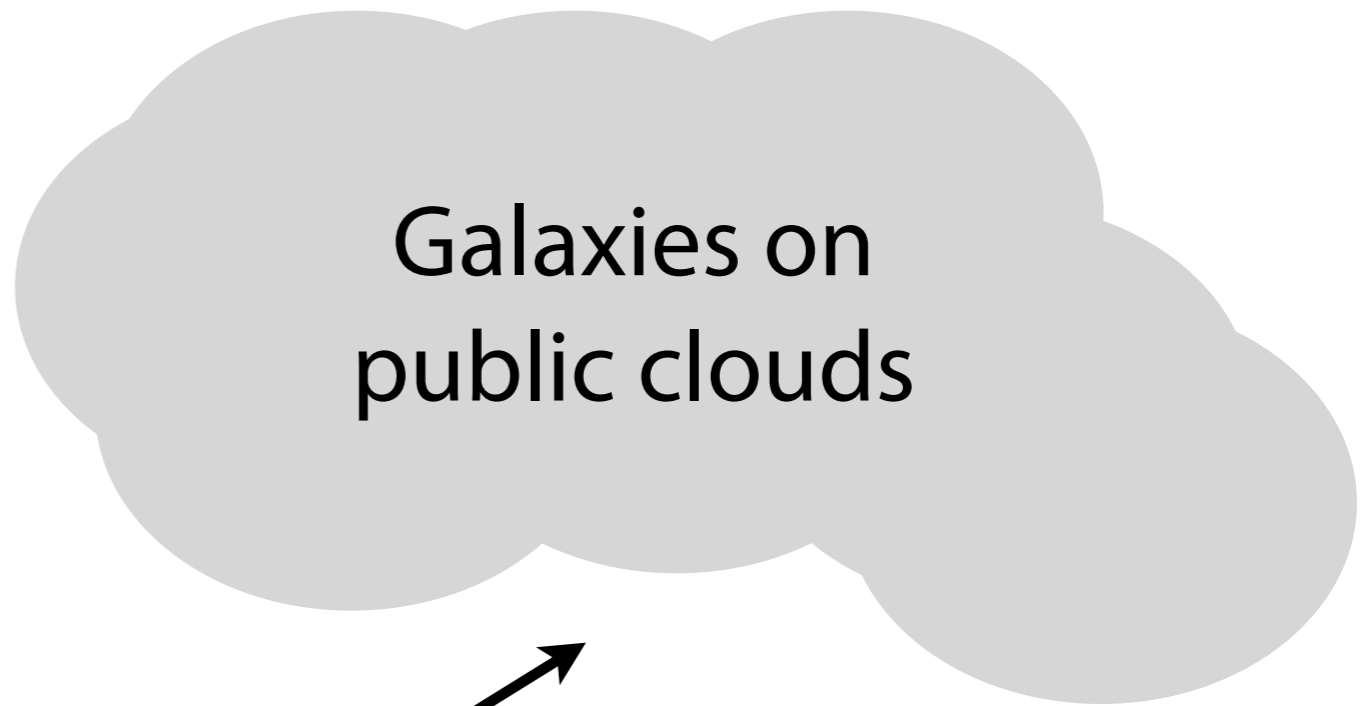


Enables sharing of Galaxy **Utilities**:
tools
proprietary datatypes
exported Galaxy workflows

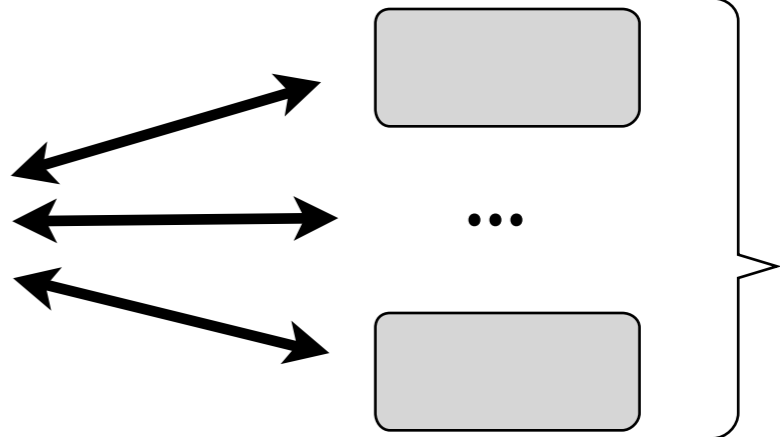
Automatically install tools and **tool suites**, and their **dependencies**, into a Galaxy instance

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

<https://wiki.galaxyproject.org/ToolShed>



<http://usegalaxy.org>

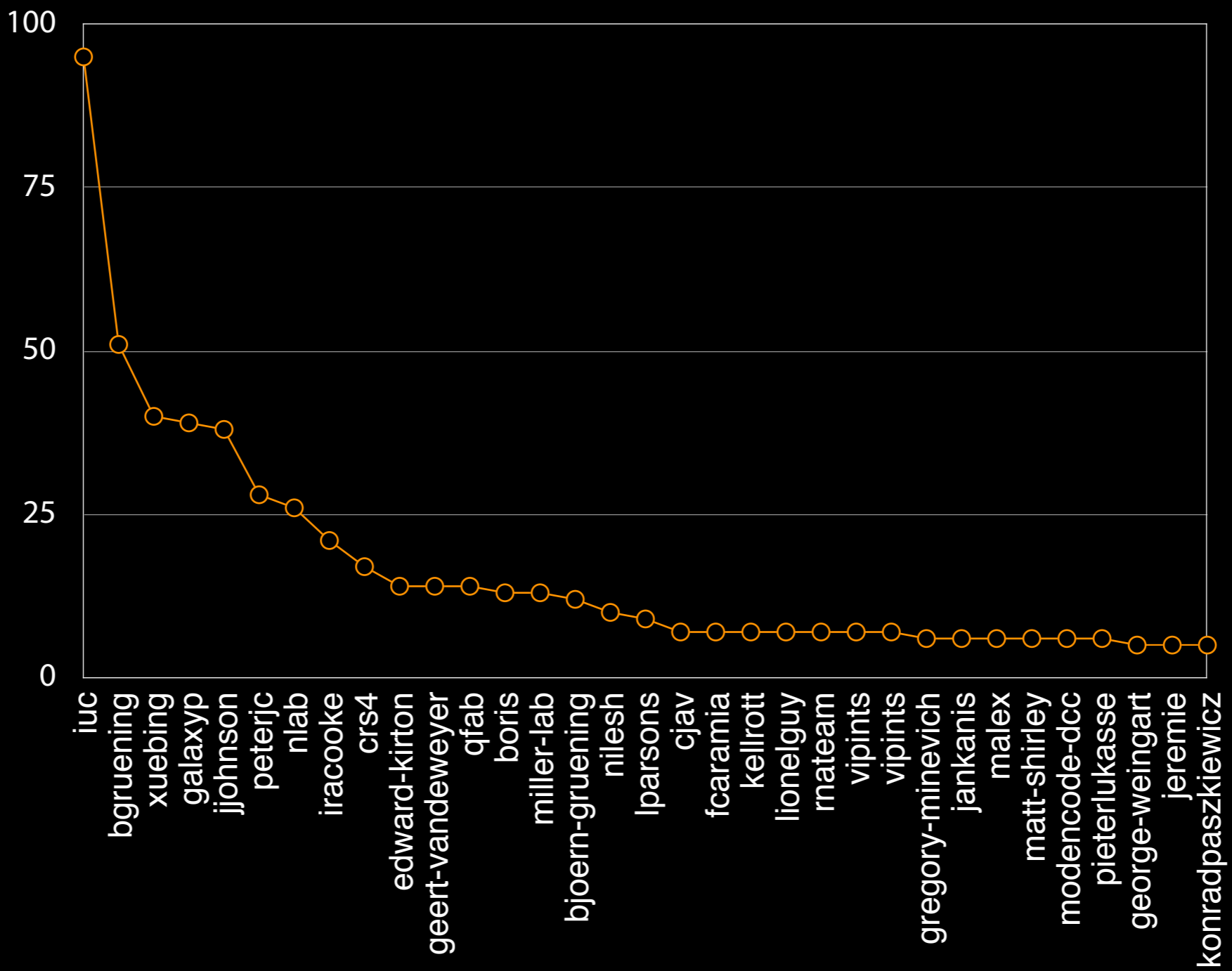


private Galaxy installations

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?

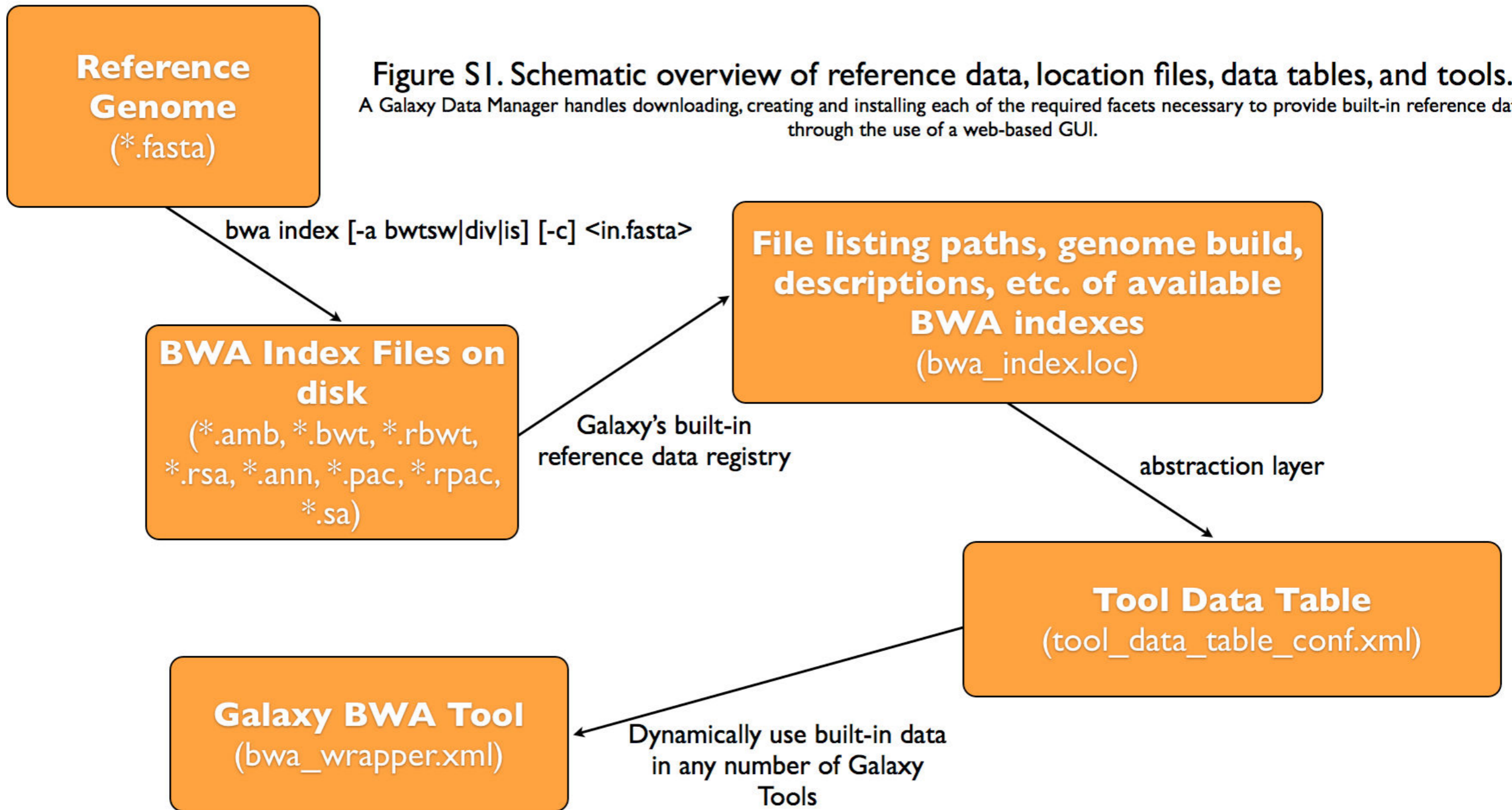
ToolShed

Data Managers

Demo

Built-in Datasets

BWA example



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



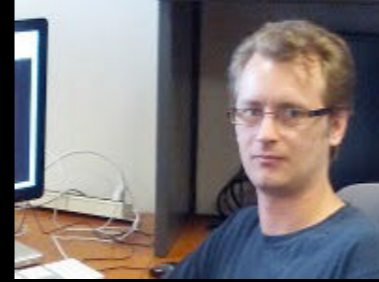
Enis Afgan



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>