

Galaxy 2011



Community Conference

25-26 May Lunteren, The Netherlands

Dank u NBIC!

- ▶ Hailiang Lai
- ▶ Femke Francissen
- ▶ Freek de Bruijn
- ▶ Anita Radstaat
- ▶ Works at De Werelt
- ▶ Marc van Driel
- ▶ Rob Hooft
- ▶ Mons Barend



The 2011 Galaxy Community Conference is generously sponsored by the [US National Science Foundation](#) and the [Netherlands Bioinformatics Centre \(NBIC\)](#). Galaxy is developed by [Galaxy Team](#) and funded by [NSF](#), [NHGRI](#), [Penn State](#), [Emory University](#), [Penn State Institute for CyberScience](#) and the [Huck Institutes of the Life Sciences](#). Galaxy is free for all.



Introduction to **Galaxy** and **GCC**

The Team

Slide Organization

Deeply Meaningful Title

Not so clear content

Person who can clarify slide for beer →



Talk Organization

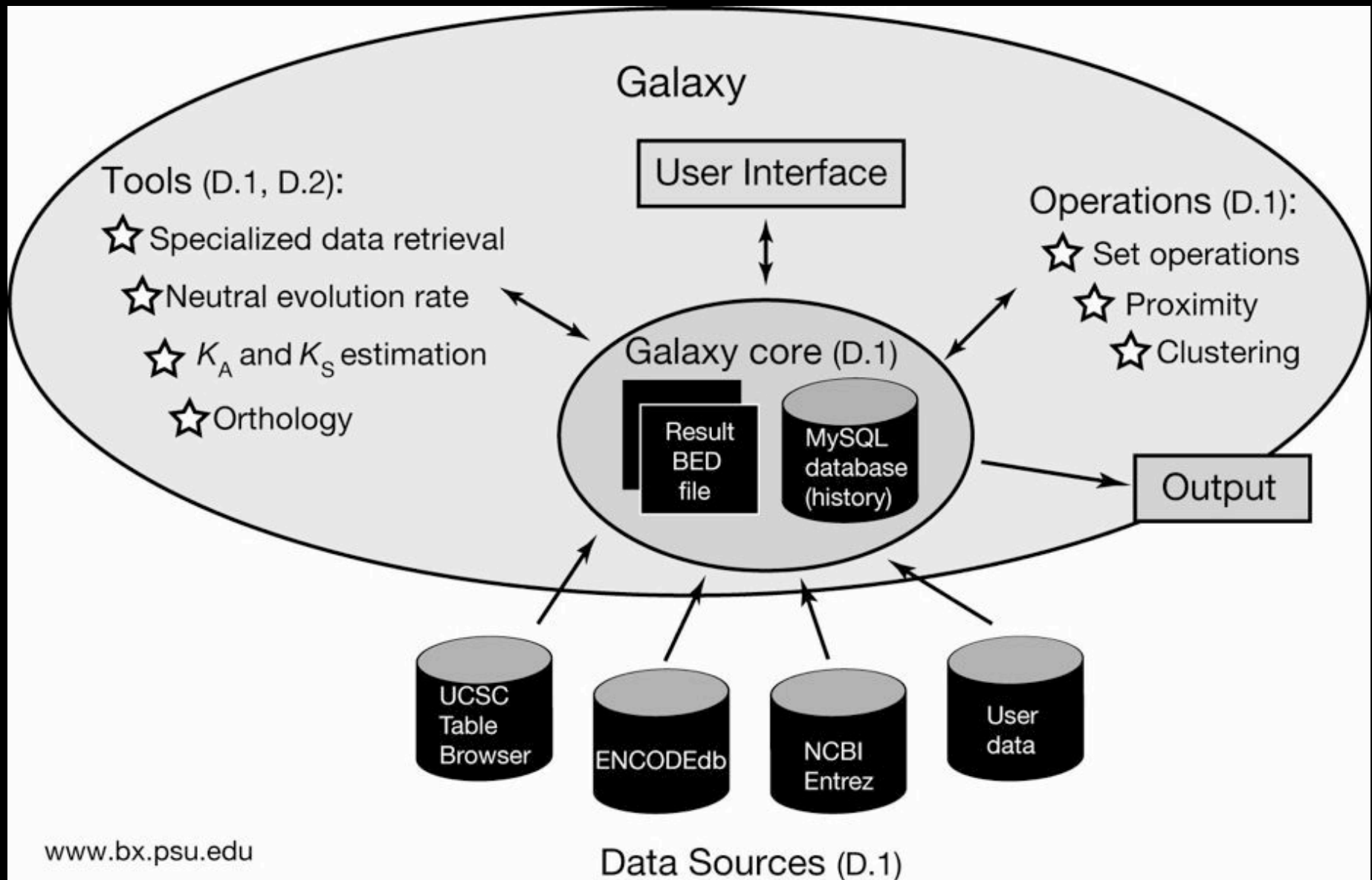
The next 90 minutes

- ▶ History (2005 – 2010)
- ▶ Present (2010 – 2011)
- ▶ The “Vision” (2011 – ∞)
- ▶ The Community
- ▶ Beer

The next 89 min & 59 seconds

- ▶ History (2005 – 2010)
- ▶ Present (2010 – 2011)
- ▶ The “Vision” (2011 – ∞)
- ▶ The Community
- ▶ Beer

Galaxy as a single Perl script (!)



Galaxy as a single Perl script (!)

The image displays four screenshots of the Galaxy web interface, labeled A, B, C, and D, illustrating the workflow for running a query operation.

Panel A: Table Browser
This panel shows the 'Table Browser' interface. It includes a description of the tool and various configuration options:
- **clade:** Vertebrate
- **genome:** Human
- **assembly:** May 2004
- **group:** Genes and Gene Prediction Tracks
- **track:** CCDS
- **table:** ccdsGene
- **region:** genome, ENCODE, position (selected)
- **identifiers (names/accessions):** paste list, upload list
- **filter:** create
- **intersection:** create
- **output format:** query results to Galaxy
- **output file:** (leave blank to keep output in browser)
- **file type returned:** plain text (selected), gzip compressed
Buttons for 'get output' and 'summary/statistics' are at the bottom.

Panel B: Galaxy: History Page
This panel shows the 'History Page' with the following configuration:
- **Genome:** Human
- **Assembly:** hg17: May 2004
- **Your Previous Queries:** 1: ccdsGene cds (limit to chr11:2110531-2116578) [3 regions]
- **Action to Perform:** Get output, Perform operations like intersection, etc., Run analysis tools, Delete selected queries
Buttons for 'Go' and 'Refresh' are at the bottom.

Panel C: Galaxy: Query Operations
This panel shows the 'Query Operations' interface. It includes the following configuration:
- **Assembly:** Human, hg17
- **Selected Queries:** 1: ccdsGene cds (limit to chr11:2110531-2116578) [3 regions], 2: snp (limit to chr11:2110531-2116578) [40 regions]
- **Operation:** Help, Union, Intersection (selected), Subtraction, Complement, Restrict, Proximity, Clusters
- **Intersection details:** return whole regions from query #1, where overlap >= 1 bp
Buttons for 'Go' and 'Refresh' are at the bottom.

Panel D: Galaxy: History Page
This panel shows the 'History Page' with the following configuration:
- **Genome:** Human
- **Assembly:** hg17: May 2004
- **Your Previous Queries:** 1: ccdsGene cds (limit to chr11:2110531-2116578) [3 regions], 2: snp (limit to chr11:2110531-2116578) [40 regions], 3: regions from query 1 that intersect regions from query 2 [2 regions]
- **Action to Perform:** Get output (selected), Perform operations like intersection, etc., Run analysis tools, Delete selected queries
Buttons for 'Go' and 'Refresh' are at the bottom.

Etymology

- ▶ Galaxy = Gala + XL (Bob Harris, author of Lastz)
- ▶ GALA = Genome Alignment and Annotation Database
- ▶ Brainchild of Ross Hardison
- ▶ Taking over the universe was not our original intension

Ross Hardison



Tuesday, May 31, 2011

Pythonic Age (mid 2005)

Basic Statistics:

Histogram: `histogram.tool`
Scatter Plot: `scatter.tool`
Filtering: `filtering.tool`
Correlation: `correlation.tool`
Region Length: `region_length.tool`
Score distribution: `scoreGraph.tool`

Operations:

Complement: `complement.tool`
Restrict: `restrict.tool`
Merge overlapping regions: `merge.tool`
Cluster: `cluster.tool`
Union: `union.tool`
Intersect: `intersect.tool`
Subtract: `subtract.tool`
Proximity: `proximity.tool`
Join Lists: `joinLists.tool`
Vicinity: `vicinity.tool`
Join Same Coordinates Region: `joinSameCoor.tool`

Sequence Tools:

Extract sequences: `fasta-subseq-wrapper.tool`
Extract blastZ alignments: `extractAxt_wrapper.tool`

Data Sources:

UCSC query: `ucsc.tool`
Genbank: `genbank.tool`
Encode DB: `encodedb.tool`
Featured datasets: `import.tool`

Format Converters:

BED and xBED converter: `bed_convert.tool`

The screenshot shows a web browser window titled "History Page". The address bar displays "http://nekrut.bx.psu.edu:9000". The browser has a search bar with "Google" and navigation buttons. Below the browser window, the "History Page" interface is visible. It features a "Data" section with a "Query" input field and a "Genome" dropdown menu. A "refresh" link is present. The "Perform Actions" section lists several categories: "Basic Statistics" (with links to Histogram, Scatter Plot, Filtering, Correlation, Region Length, and Score distribution), "Operations", "Sequence Tools", "Data Sources", "Format Converters", and "Upload Data". Each link in the "Basic Statistics" section is accompanied by a brief description and a "[help]" link.

History Page

Query Genome

[refresh](#)

Perform Actions

- [Basic Statistics](#)
 - [Histogram](#) - builds histogram for any numeric column [\[help\]](#)
 - [Scatter Plot](#) - builds scatterplot for any numeric column [\[help\]](#)
 - [Filtering](#) - filters data on any column using simple conditional expressions [\[help\]](#)
 - [Correlation](#) - computes Pearson's correlation between any two numerical columns [\[help\]](#)
 - [Region Length](#) - computes length of bed intervals [\[help\]](#)
 - [Score distribution](#) - display the score distribution of a selected score name [\[help\]](#)
- [Operations](#)
- [Sequence Tools](#)
- [Data Sources](#)
- [Format Converters](#)
- [Upload Data](#)

2006

Galaxy

columns | stack | tools | history

Data

Edit

Filter and Sort

Format

Sequences and Alignments

Scores

Interval Functions

Statistics

Graph

EMBOSS

PHYLIP

Cluster regions of a single Query

Cluster and Merge regions within specified distance of each other

Complement a single Query

Coverage density of the regions of two queries

Difference find non-overlapping segments between two Queries

Intersect return overlapping segments of two Queries

Cluster

Cluster regions of Query: 3: UCSC: knownGene (chr7:12747119)

Regions per cluster: 2

Within distance: 10

Execute

TIP: If your query does not appear in the pulldown menu -> it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns

Syntax

- Regions per cluster sets the minimum number of regions to start a cluster
- Within distance sets the maximum separation (in base pairs) at which two regions are still considered to be within the same cluster

Example

Cluster				
Query1		■		■
Query2	■	■		■
Result	■	■	■	■

refresh | collapse all

3: UCSC: knownGene (chr7:127471196-127495720), running, bed, hg17

2: Maf to concatenated FASTA on data 1, 8 sequences, fasta, hg17

1: UCSC: multiz8way (chr22:23153675-23162878), 511 lines, maf, hg17

History

Done

Tuesday, May 31, 2011

Sometime before today

The screenshot shows the Galaxy web interface in a browser window. The address bar displays `http://main.g2.bx.psu.edu/`. The page title is "Galaxy". The navigation bar includes links for "Info: report bugs | wiki | screencasts" and "Account: create | login".

Get Data

- Upload File from your computer
- UCSC Main table browser
- Get Microbial Data

Get ENCODE Data

ENCODE Tools

Edit Queries

Filter, Sort, Join and Compare

Convert Formats

Fetch Sequences and Alignments

Get Genomic Scores

Operate on Genomic Intervals

Statistics

Graph Data

EMBOSS

HYPHY

for a description of the controls in this form. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data.

clade: Vertebrate **genome:** Human
assembly: Mar. 2006
group: Variation and Repeats **track:** SNPs (126)
table: snp126
[describe table schema](#)
region: ☐ genome ☒ position
chr7:113265469-114910668 [lookup](#) [define regions](#)
identifiers (names/accessions): [paste list](#) [upload list](#)
filter: [create](#)
intersection: [create](#)
correlation: [create](#)
output format: BED - browser extensible data
☒ Send output to [Galaxy](#)
output file: (leave blank to keep output in browser)
file type returned: ☒ plain text ☐ gzip compressed
[get output](#) [summary/statistics](#)
To reset all user cart settings (including custom tracks), [click here](#).

refresh | collapse all

1: UCSC Main on Human: knownGene (chr7:113265469-114910668)

History options...

- You must be [logged in](#) to store or switch histories.
- [share](#) current history
- [delete](#) current history

Done

The First Galaxy Developer Conf



The next 90 – x minutes

- ▶ History (2005 – 2010)
- ▶ Present (2010 – 2011)
- ▶ The “Vision” (2011 – ∞)
- ▶ The Community
- ▶ Beer

Galaxy Today

- **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

Galaxy: accessible analysis system

The screenshot displays the Galaxy web interface in a browser window. The address bar shows <http://main.g2.bx.psu.edu/>. The top navigation bar includes links for **Analyze Data**, **Workflow**, **Data Libraries**, **Admin**, **Help**, and **User**.

Tools Panel (Left): A list of available tools categorized into sections: **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**, **Metagenomic analyses**, **EMBOSS**, **NGS TOOLBOX BETA**, **NGS: QC and manipulation**, **NGS: Mapping**, **NGS: SAM Tools**, **NGS: Peak Calling**, **RGENETICS**, **SNP/WGA: Data; Filters**, and **SNP/WGA: QC; LD; Plots**.

Main Content Area: A central panel titled "Here is what's happening..." features a large box with the text "Mapping Pipeline for Illumina, 454, and SOLiD" and a prominent "USE IT NOW!" button. Below this, a section titled "Live Quickies (more after May 17 ...)" displays three cards: "Basic fastQ manipulation: Galaxy quickie # 13", "Advanced fastQ manipulation: Galaxy quickie # 14", and "454 Mapping: Single End Galaxy quickie # 15". At the bottom of the main area, text states: "The Galaxy team is a part of BX at Penn State. This project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, and The Institute for CyberScience at Penn State. Galaxy build: \$Rev 3885:1ab9d6b0ddfc\$".

History Panel (Right): A list of previous jobs under the heading "History" and "Options". The jobs are numbered 4 through 16, each with a description and icons for viewing, deleting, and refreshing. The jobs include: "4: FASTA-to-Tabular on", "5: Add column on data 4", "6: Tabular-to-FASTA on data 5", "7: Megablast on data 6", "8: Megablast on data 6", "9: Compute sequence length on data 6", "10: Concatenate queries on data 8 and data 7", "11: Join two Queries on data 9 and data 10", "12: Filter on data 11", "13: Fetch taxonomic representation on data 12", "14: Find lowest diagnostic rank on data 13", "15: Summarize taxonomy on data 13", and "16: Draw phylogeny on data 14".

Tools

Integrating existing tools into a uniform framework

The image shows a Galaxy tool interface for a tool named 'Cluster'. On the left, a code editor displays the XML definition of the tool. The XML includes a description, command, inputs, and various parameters like 'format', 'distance', 'minregions', and 'returntype'. On the right, a graphical user interface (GUI) for the tool is shown. It features a dropdown menu for 'Cluster intervals of:' with '1: UCSC Main on Human genome' selected. Below this are input fields for 'max distance between intervals:' (set to 1) and 'min number of intervals per cluster:' (set to 2). A 'Return type:' dropdown is set to 'Merge clusters into single intervals'. An 'Execute' button is at the bottom. A tip box states: 'TIP: If your query does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.' Below the tip are sections for 'Screencasts!' and 'Syntax'. The 'Syntax' section lists: 'Maximum distance is greatest distance in base pairs allowed between intervals that will be merged', 'Minimum intervals per cluster', 'Merge clusters into single intervals', 'Find cluster intervals', and 'Find cluster intervals'.

```
<?xml version="1.0" encoding="UTF-8"?>
<tool id="gops_cluster_1" name="Cluster">
  <description>[[Cluster]] the intervals of a query</description>
  <command interpreter="python">
    gops_cluster.py $input1
    -d $dist
  </command>
  <inputs>
    <param format="interval"
      <label>Cluster interval
    </param>
    <param name="distance"
      <label>max distance between
    </param>
    <param name="minregions"
      <label>min number of intervals
    </param>
    <param name="returntype"
      <option value="1">Merge
      <option value="2">Find
      <option value="3">Find
      <option value="4">Find
      <option value="5">Find
    </param>
  </inputs>
  <help>
    .. class:: infomark
    **TIP:** If your query does
    -----
    **Screencasts!**
    See Galaxy Interval Operation
    .. _Screencasts: http://www.
    -----
    **Syntax**
    - **Maximum distance** is gr
    - **Minimum intervals per cl
    - **Merge clusters into sing
    - **Find cluster intervals;
    - **Find cluster intervals;
  </help>
</tool>
```

- Defined in terms of an abstract interface (inputs and outputs)
- In practice, mostly command line tools, a declarative XML description of the interface, how to generate a command line
- Designed to be as easy as possible for tool authors, while still allowing rigorous reasoning

Dan Blankenberg, Guru Ananda, Kelly Vincent, Ross Lazarus

NGS: QC and manipulation

ILLUMINA DATA

- [FASTQ Groomer](#) convert between various FASTQ quality formats
- [FASTQ splitter](#) on joined paired end reads
- [FASTQ joiner](#) on paired end reads
- [FASTQ Summary Statistics](#) by column

ROCHE-454 DATA

- [Build base quality distribution](#)
- [Select high quality segments](#)
- [Combine FASTA and QUAL](#) into FASTQ

AB-SOLID DATA

- [Convert](#) SOLiD output to fastq
- [Compute quality statistics](#) for SOLiD data
- [Draw quality score boxplot](#) for SOLiD data

GENERIC FASTQ MANIPULATION

- [Filter FASTQ](#) reads by quality score and length
- [FASTQ Trimmer](#) by column
- [FASTQ Quality Trimmer](#) by sliding window

Evolution

Metagenomic analyses

Human Genome Variation

EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation

NGS: Mapping

ILLUMINA

- [Map with Bowtie](#) for Illumina
- [Map with BWA](#) for Illumina

ROCHE-454

- [Lastz](#) map short reads against reference sequence
- [Megablast](#) compare short reads against htgs, nt, and wgs databases

- [Parse blast XML output](#)

AB-SOLID

- [Map with Bowtie](#) for SOLiD

NGS: SAM Tools

NGS: Indel Analysis

NGS: Peak Calling

NGS: RNA Analysis

RGENETICS

SNP/WGA: Data; Filters

SNP/WGA: QC; LD; Plots

SNP/WGA: Statistical Models

Workflows

NGS TOOLBOX BETA

NGS: QC and manipulation

NGS: Mapping

NGS: SAM Tools

- [Filter SAM](#) on bitwise flag values
- [Convert SAM](#) to interval
- [SAM-to-BAM](#) converts SAM format to BAM format
- [BAM-to-SAM](#) converts BAM format to SAM format
- [Merge BAM Files](#) merges BAM files together
- [Generate pileup](#) from BAM dataset
- [Filter pileup](#) on coverage and SNPs
- [Pileup-to-Interval](#) condenses pileup format into ranges of bases
- [flagstat](#) provides simple stats on BAM files

NGS: Indel Analysis

NGS: Peak Calling

NGS: RNA Analysis

RGENETICS

SNP/WGA: Data; Filters

SNP/WGA: QC; LD; Plots

SNP/WGA: Statistical Models

Workflows

NGS: SAM Tools

NGS: Indel Analysis

- [Filter Indels](#) for SAM
- [Extract indels](#) from SAM
- [Indel Analysis](#)

NGS: Peak Calling

- [MACS](#) Model-based Analysis of ChIP-Seq
- [GeneTrack indexer](#) on a BED file
- [Peak predictor](#) on GeneTrack index

NGS: RNA Analysis

RNA-SEQ

- [Tophat](#) 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
- [Cuffdiff](#) find significant changes in transcript expression, splicing, and promoter use

FILTERING

- [Filter Combined Transcripts](#) using tracking file

Dozens of tools for different NGS applications packaged with Galaxy

Galaxy Tool Shed

http://community.g2.bx.psu.edu/ Google

Galaxy Tool Shed / (beta) [Tools](#) [Help](#) [User](#)

Community

Tools

- [Browse by category](#)
- [Browse all tools](#)
- [Login to upload](#)

Categories

search [Advanced Search](#)

Name ↓	Description	Tools
Convert Formats	Tools for converting data formats	5
Data Source	Tools for retrieving data from external data sources	1
Fasta Manipulation	Tools for manipulating fasta data	5
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	7
Ontology Manipulation	Tools for manipulating ontologies	1
SAM	Tools for manipulating alignments in the SAM format	0
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	10
SNP Analysis	Tools for single nucleotide polymorphism data such as WGA	1
Statistics	Tools for generating statistics	1
Text Manipulation	Tools for manipulating data	3
Visualization	Tools for visualizing data	1

Display a menu

The Galaxy Tool Shed allows the community to contribute, share, and evaluate tools

Greg Von Kuster

Galaxy Tool Shed

http://community.g2.bx.psu.edu/

Galaxy Tool Shed / (beta) Tools Help User

Community

Tools

- Browse by category
- Browse all tools
- Login to upload

View Tool

This is the latest approved version of this tool suite Tool Actions ▾

Mothur Metagenomics

Tool Id:
Mothur_toolsuite

Version:
1.15.1

Description:
Mothur metagenomics commands as Galaxy tools

User Description:
Provides galaxy tools for the commands in the Mothur metagenomics package: http://www.mothur.org/wiki/Main_Page


Uploaded by:
jjohnson

Date uploaded:
about 22 hours ago

Categories:

- Sequence Analysis

Tool Contents

 [Mothur toolsuite 1.15.1.tar.gz](#)

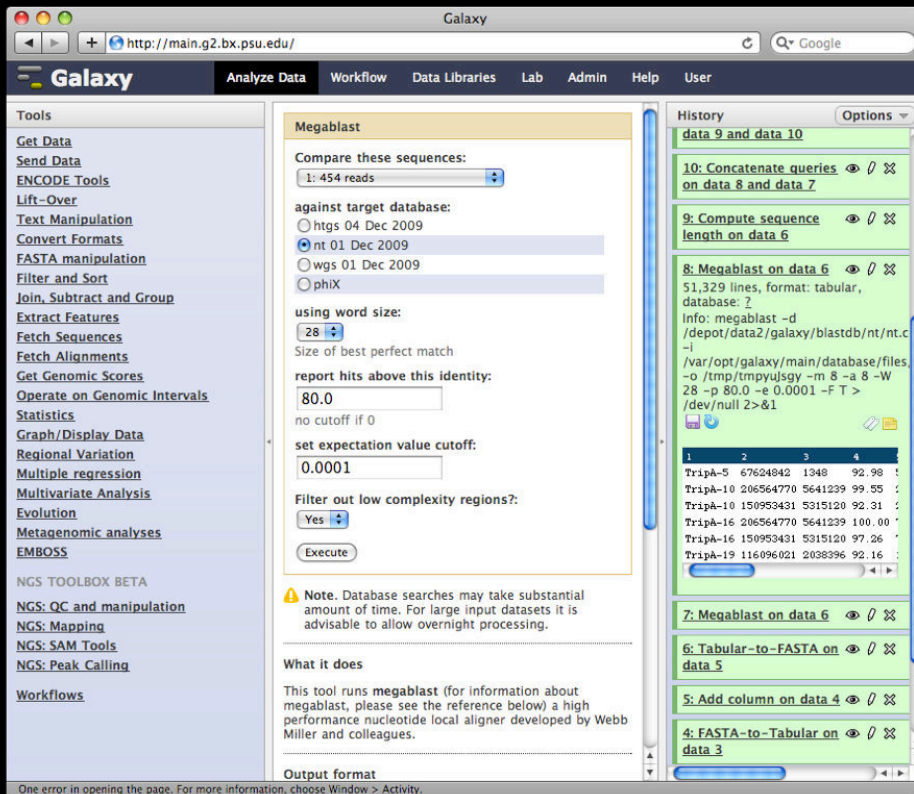
- [mothur/](#)
- [mothur/tools/](#)
- [mothur/tools/mothur/](#)
- [mothur/tools/mothur/split.abund.xml](#)

For example, complete wrappers for the Mothur metagenomics suite
from **Jim Johnson** (UMN)

Jim Johnson, Konrad Paszkiewicz, Peter Cock, Vipin Sreedharan

Analysis environment

Galaxy analysis interface



- Consistent tool user interfaces automatically generated
- History system facilitates and tracks multistep analyses

Automatically tracks every step of every

7: Map with Bowtie for Illumina on data 6 and data 5

9,073,928 lines, format: sam,
database: mm9
Run this job again

1. QNAME	2. FLAG	3. RNAME
HWI-EAS269:3:1:1449:913	99	chr1
HWI-EAS269:3:1:1449:913	147	chr1
HWI-EAS269:3:1:709:832	99	chr1
HWI-EAS269:3:1:709:832	147	chr1
HWI-EAS269:3:1:1422:1087	99	chr1
HWI-EAS269:3:1:1422:1087	147	chr1

Map with Bowtie for Illumina

Will you select a reference genome from your history or use a built-in index?

Built-ins were indexed using default options

Select a reference genome:

if your genome of interest is not listed - contact Galaxy team

Is this library mate-paired?:

Forward FASTQ file:

Must have Sanger-scaled quality values with ASCII offset 33

Reverse FASTQ file:

Must have Sanger-scaled quality values with ASCII offset 33

Maximum insert size for valid paired-end alignments (-X):

The upstream/downstream mate orientation for valid paired-end alignment against the forward reference strand (--fr/--rf/--ff):

Bowtie settings to use:

For most mapping needs use Commonly used settings. If you want full control use Full parameter list





Suppress the header in the output SAM file:
☒

Bowtie produces SAM with several lines of header information by default

Dan Blankenberg

As well as user-generated metadata and annotation...


History Options ▾




Variant Analysis for Sample E18

Tags:

snp x pileup x bowtie x





demo x sample:e18 x 

Annotation / Notes:
Perform a variant analysis with default parameters to identify variants in sample E18 that lie in annotated genes.

10: Variants from sample E18   


26,742 regions, format: interval, database: mm9

Info:

Tags:

pileup x sample:e18 x


snps x 

Annotation:

Find variants with coverage ≥ 30 and quality score ≥ 20 .

| display at UCSC [main](#) | view in [GeneTrack](#) | display at Ensembl [Current](#)

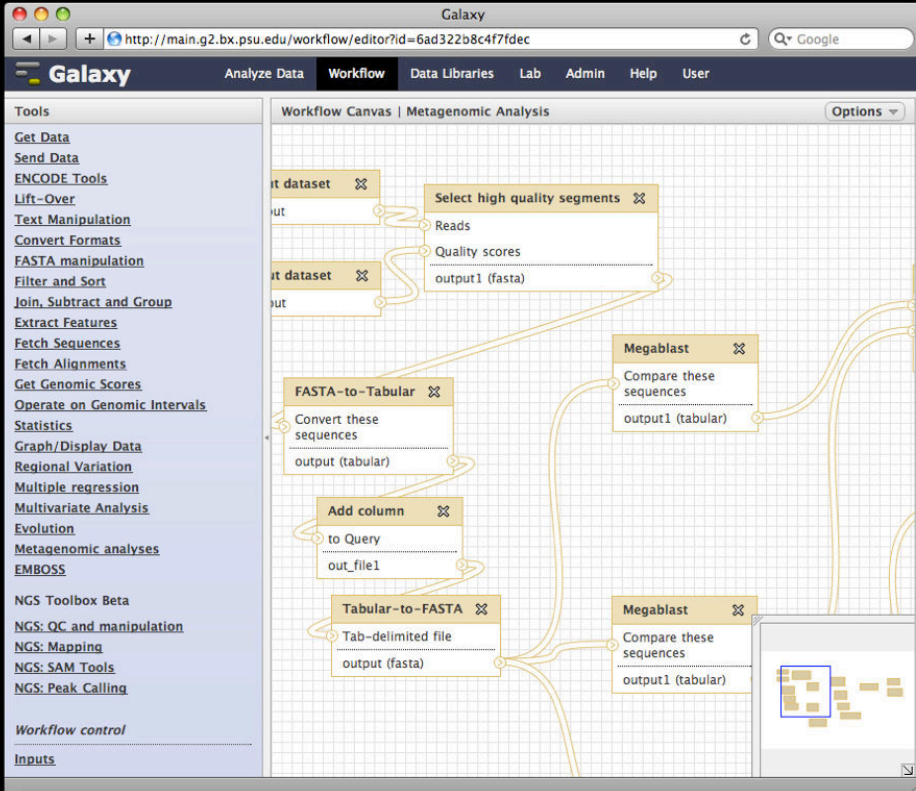
1. Chrom	2. Start	3. End	4	5	6
chr10	6882036	6882037	A	A	107
chr10	14243075	14243076	G	G	96
chr10	14243079	14243080	C	C	106
chr10	14465082	14465083	T	K	173
chr10	14465083	14465084	G	K	144
chr10	14465084	14465085	T	T	117



Jeremy Goecks

Workflows

Galaxy workflow system



- Workflows can be constructed from scratch or extracted from existing analysis histories
- Facilitate reuse, as well as providing precise reproducibility of a complex analysis

Dannon Baker

Galaxy

psu.edu/

Analyze Data Workflow Shared Data Lab Visualization Admin Help User

Upload File
This tool cannot be used in workflows

UCSC Main
This tool cannot be used in workflows

UCSC Main
This tool cannot be used in workflows

Tophat
☒ Include "Tophat" in workflow

Tophat
☒ Include "Tophat" in workflow

Cufflinks
☒ Include "Cufflinks" in workflow

2: imported: GM12878 Sample Dataset
☒ Treat as input dataset

3: imported: UCSC Main on Human: refGene chr19 BED
☒ Treat as input dataset

4: imported: UCSC Main on Human: refGene chr19 GTF
☒ Treat as input dataset

7: Tophat on data 1: splice junctions

8: Tophat on data 1: accepted_hits

9: Tophat on data 2: splice junctions

10: Tophat on data 2: accepted_hits

11: Cufflinks on data 8: gene expression

12: Cufflinks on data 8: transcript expression

13: Cufflinks on data 8:

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

Show Structure

Export to File

Delete

Other Actions

Import from File

27: C data

26: C data diff

25: C data

24: C data track

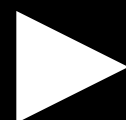
23: C data FPKM

22: C data 10, and data 4: gene FPKM tracking

21: Cuffdiff on data 8, data 10, and data 4: isoform FPKM tracking

20: Cuffdiff on data 8, data 10, and data 4: CDS Expression FPKM Tracking

19: Cuffdiff on data 8, data 10, and data 4: TSS groups



Galaxy

http://main.g2.bx.psu.edu/workflow/editor?id=5928d603676a469a

Analyze Data Workflow Shared Data Lab Visualization Admin Help

Workflow Canvas | Workflow constructed from history 'RNA-seq exercise (full)'

Input dataset output

Input dataset output

Input dataset output

Input dataset output

Tophat RNA-Seq FASTQ file

insertions (bed)

deletions (bed)

junctions (bed)

accepted_hits (bam)

Tophat RNA-Seq FASTQ file

insertions (bed)

deletions (bed)

junctions (bed)

accepted_hits (bam)

Cufflinks SAM or BAM file of aligned reads

genes_expression (tabular)

transcripts_expression (tabular)

assembled_isoforms (gtf)

Cufflinks SAM or BAM file of aligned reads

genes_expression (tabular)

transcripts_expression (tabular)

assembled_isoforms (gtf)

Cuffdiff Transcripts

SAM or BAM file of aligned reads

SAM or BAM file of aligned reads

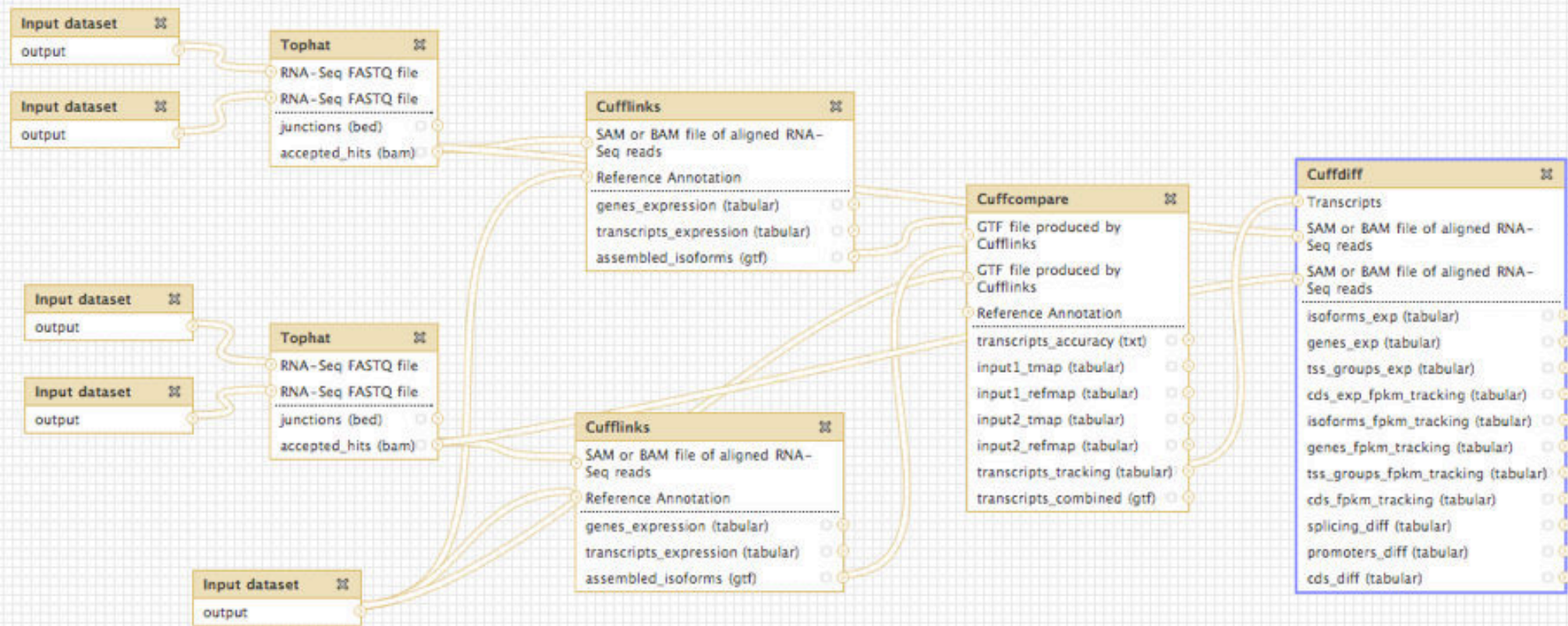
isoforms_exp (tabular)

genes_exp (tabular)

tss_groups_exp (tabular)

cds_exp_fpk_tracking (tabular)

Display a menu



Data Libraries

Galaxy

http://main.g2.bx.psu.edu/library

Galaxy Analyze Data Workflow Shared Data Visualization Help User

Data Library "mtProject"

Name	Message	Uploaded By	Date	File Size
<input type="checkbox"/> F4-bM4C3 ▾	Family 4, child M4C3, blood, PCR (no replicates)	anton@bx.psu.edu	2010-09-01	2.4 Gb
<input type="checkbox"/> F4-bM5G-1 ▾	Family 4, grandmother M5G, blood, PCR1	anton@bx.psu.edu	2010-09-01	2.0 Gb
<input type="checkbox"/> F4-bM5G-2 ▾	Family 4, grandmother M5G, blood, PCR2	anton@bx.psu.edu	2010-09-01	2.4 Gb
<input type="checkbox"/> F4-bM9 ▾	Family 4, sister M9, blood, PCR (no replicates)	anton@bx.psu.edu	2010-09-01	1.4 Gb
<input type="checkbox"/> F4-cM4C3 ▾	Family 4, child M4C3, cheek, PCR (no replicates)	anton@bx.psu.edu	2010-09-01	1.5 Gb
<input type="checkbox"/> F4-cM5G-1 ▾	Family 4, grandmother M5G, cheek, PCR1	anton@bx.psu.edu	2010-09-01	1.6 Gb
<input type="checkbox"/> F4-cM5G-2 ▾	Family 4, grandmother M5G, cheek, PCR2	anton@bx.psu.edu	2010-09-01	1.7 Gb
<input type="checkbox"/> F4-cM9 ▾	Family 4, sister M9, cheek, PCR (no replicates)	anton@bx.psu.edu	2010-09-01	1.6 Gb
<input type="checkbox"/> F4-bM4C2-1 ▾	Family 4, child M4C2, blood, PCR1	anton@bx.psu.edu	2010-01-08	247.1 Mb
<input type="checkbox"/> F4-bM4C2-2 ▾	Family 4, child M4C2, blood, PCR2	anton@bx.psu.edu	2010-01-08	426.8 Mb
<input type="checkbox"/> F4-cM4C2-1 ▾	Family 4, child M4C2, cheek, PCR1	anton@bx.psu.edu	2010-01-08	92.3 Mb
<input type="checkbox"/> F4-cM4C2-2 ▾	Family 4, child M4C2, cheek, PCR2	anton@bx.psu.edu	2010-01-08	157.7 Mb
<input type="checkbox"/> F4-bM4-1 ▾	Family 4, mother M4, blood, PCR1	anton@bx.psu.edu	2010-01-08	85.7 Mb
<input type="checkbox"/> F4-bM4-2 ▾	Family 4, mother M4, blood, PCR2	anton@bx.psu.edu	2010-01-08	132.3 Mb
<input type="checkbox"/> F4-cM4-1 ▾	Family 4, mother M4, cheek, PCR1	anton@bx.psu.edu	2010-01-08	150.2 Mb
<input type="checkbox"/> F4-cM4-2 ▾	Family 4, mother M4, cheek, PCR2	anton@bx.psu.edu	2010-01-08	99.8 Mb
<input type="checkbox"/> F7-bM10C2-1 ▾	Family 7, child M10C2, blood, PCR1	anton@bx.psu.edu	2010-01-08	90.8 Mb
<input type="checkbox"/> F7-bM10C2-2 ▾	Family 7, child M10C2, blood, PCR2	anton@bx.psu.edu	2010-01-08	156.3 Mb
<input type="checkbox"/> F7-cM10C2-1 ▾	Family 7, child M10C2, cheek, PCR1	anton@bx.psu.edu	2010-01-08	162.3 Mb
<input type="checkbox"/> F7-cM10C2-2 ▾	Family 7, child M10C2, cheek, PCR2	anton@bx.psu.edu	2010-01-08	170.0 Mb
<input type="checkbox"/> F7-bM10-1 ▾	Family 7, mother M10, blood, PCR1	anton@bx.psu.edu	2010-01-08	192.9 Mb
<input type="checkbox"/> F7-bM10-2 ▾	Family 7, mother M10, blood, PCR2	anton@bx.psu.edu	2010-01-08	118.1 Mb
<input type="checkbox"/> F7-cM10-1 ▾	Family 7, mother M10, cheek, PCR1	anton@bx.psu.edu	2010-01-08	243.3 Mb

Greg Von Kuster

Tuesday, May 31, 2011

Sharing and publishing


Everything can be shared

Sharing and Publishing History 'Variant Analysis for Sample E18'

Making History Accessible via Link and Publishing It

This history accessible via link and published.

Anyone can view and import this history by visiting the following URL:

<http://main.g2.bx.psu.edu/u/jgoecks/h/variant-analysis-for-sample-e18> 

This history is publicly listed and searchable in Galaxy's [Published Histories](#) section.

You can:

Unpublish History

Removes history from Galaxy's [Published Histories](#) section so that it is not publicly listed or searchable.

Disable Access to History via Link and Unpublish

Disables history's link so that it is not accessible and removes history from Galaxy's [Published Histories](#) section so that it is not publicly listed or searchable.

Sharing History with Specific Users

You have not shared this history with any users.

Share with a user

[Back to Histories List](#)

Galaxy | Published Pages

http://main.g2.bx.psu.edu/page/list_published

Galaxy Analyze Data Workflow **Shared Data** Lab Visualization Admin Help User

Published Pages

search Advanced Search

Title	Annotation	Owner	Community Rating	Community Tags	Last Updated ↓
ChrY 1000 Genomes	A demo workshop project during CSHL course on Computational Genomics Nov 2010	ericy	★★★★★		2 days ago
Galaxy Exercises	Various exercises for learning about Galaxy	james	★★★★★		5 days ago
Galaxy 101: The first thing you need to try	An elementary guide to Galaxy	aun1	★★★★★	exons snps tutorial	Nov 03, 2010
Windshield Splatter	Live supplement for Genome Research windshield splatter paper.	aun1	★★★★★	megan paper galaxy	Oct 27, 2010
Galaxy RNA-seq Analysis Exercise	An exercise that illustrates how to use Galaxy for RNA-seq analyses.	jeremy	★★★★★		Oct 27, 2010
heteroplasmy		aun1	★★★★★	heteroplasmy bwa resequencing illumina	Oct 26, 2010

Pervasive search allows others to find published items of interest

The screenshot shows a web browser window with the address bar displaying `http://main.g2.bx.psu.edu/u/aun1/p/heteroplasmy`. The browser's title bar reads "Galaxy | Published Page | heteroplasmy". The page header features the "Galaxy" logo and a navigation menu with links: "Analyze Data", "Workflow", "Shared Data", "Lab", "Visualization", "Admin", "Help", and "User". Below the header, a breadcrumb trail shows "Published Pages | aun1 | heteroplasmy".

Dynamics of mitochondrial heteroplasmy in three families: A fully reproducible re-sequencing study

Hiroki Goto¹, Benjamin Dickins², Enis Afgan^{3,5}, Ian M. Paul⁴, James Taylor^{3,5}, Kateryna D. Makova¹, and Anton Nekrutenko^{2,5}

Correspondence should be addressed to [KDM](#), [JT](#), or [AN](#).

1. How to use this document

This document is a live copy of supplementary materials for the manuscript. It provides access to all the data as well as to 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 sequencing data. 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. To make this even easier, we created several screencasts (very short movies) to help you:

- [access our datasets](#)
- [re-use workflows listed on this page](#)
- [view and import histories listed on this page](#)

In addition, we created two longer screenacasts:

- [Watch the analysis of one family \(F7\) from start \(Illumina reads\) to finish \(a list of variable position\):](#)
- [Watch how the complete analysis can be performed on the Amazon Cloud.](#)

If you experience any problems while using this page, please e-mail our [bug report list](#) and we will get back to you.

2. Accessing the Data

All datasets discussed in the paper can be found in two places:

- [A Galaxy Library called mtProject;](#)
- [An S3 bucket on the Amazon Cloud.](#)

Galaxy Page for a recent study on mitochondrial heteroplasmy

Galaxy | Published Page | heteroplasmy

http://main.g2.bx.psu.edu/u/aun1/p/heteroplasmy

Galaxy Analyze Data Workflow Shared Data Lab Visualization Admin Help User

Published Pages | aun1 | heteroplasmy

M10, M10C2, M15, and M15C2;

- the workflow 'mt analysis 0.01 strand-specific (*fastq single*)' was run four times on datasets that lacked PCR replicates: M9 and M4C3;

for this we created three separate histories: one for each family. Each history (F4 = Family 4, F7 = Family 7, F11 = Family 11) can be examined in detail and imported below ([see a Screencast explaining how to do this](#)):

+

Galaxy History | F4

+

+

Galaxy History | F7

+

+

Galaxy History | F11

+

Each of the histories contain original Illumina datasets and outputs of workflows.

3.3 Generating initial summary datasets

In the previous step we identified variable sites in all samples. Now we need to merge the results by generating reports for each family. To do this we first copied results workflow executions into a new history called "F4-F7-F11 final report" ([for explanation on how to copy datasets between histories see this Screencast](#)):

+

Galaxy History | F4-F7-F11 final report

+

Within this history individual datasets are merged into summaries generated for each family. To be more specific, datasets 1 through 10 were merged into dataset 19 called "F4 summary", datasets 11 - 14 were joined into history item 22 called "F7 summary", and, finally, datasets 15 - 18 were used to generate #24 called "F11 summary". Merging of datasets was performed with "Join, Subtract, and Group -> Column Join" tool. Let's look at datasets "F7 summary" to understand what this means:

+

Galaxy Dataset | F7 summary

+

Results of heteroplasmy workflow for all individuals of family 7 joined together. You can click in "rerun" button above to see the parameters.

Actual histories and datasets directly accessible from the text

Galaxy | Published Page | Heteroplasmy pilot

http://184.73.9.52/u/jtxt/p/heteroplasmy-pilot

AWS Management Console Galaxy | Published Page | Heteroplasmy...

Galaxy

Analyze Data Workflow Data Libraries Help User

Published Pages | jtxt | Heteroplasmy pilot

We analyzed the mitochondrial genome from three mother/child pairs. For each mother and child pair the DNA was collected from cheek swab specimen and from blood at Penn State Medical School. mtDNA was amplified with PCR using two primer sets L2815 and H11571; L10796 and H3370. These primers are originally described in Tanaka et al. (1996). To control for possible PCR-induced errors, each amplification was performed twice. In total we generated 24 Illumina datasets (eight for each mother and child pair – two mtDNA amplification for each cheek swab and blood samples

[Galaxy History | mt datasets](#)

Reads were mapped against hg19 version of the human genome using bwa. Only those reads aligning exactly once to the mitochondrial genome and having no hits to the nuclear genome were retained. This procedure eliminated potential contamination of our data with reads associated with numts (our PCR strategy enriched mt DNA but did not eliminate nuclear DNA from the sample: approximately 10–20% of the reads mapped to the nuclear genome and were subsequently eliminated from the analysis). Using PCRs replicates for each sample, the following workflow estimates the methodological error rate by comparing mapping results between two amplifications. To do so we identified all sites where in one replicate there were no deviant reads (all reads contained the same nucleotide; i.e. 1000 'A' bases) but the other contained such sites (e.g., 1000 As and 12 Cs). Dividing the number of deviant reads (12 in this case) by the total read coverage (1012) at such positions gave us error the rate of 1.18% (12/1012) at this position.

Galaxy Workflow | Determine threshold from PCR replicates

`c1==chrM and c10 >= 200`

Step 16: Filter

Replicate 2: Keep only positions that map to chrM and have quality adjusted coverage greater than 200

Filter
Output dataset 'out_file1' from step 14
With following condition
`c1=='chrM' and c10 >= 200`

Step 17: Join

Create a joined file containing the pileup information for all positions that have sufficient quality to consider in both replicates


Join
Output dataset 'out_file1' from step 15
with
Output dataset 'out_file1' from step 16

Histories resulting from first workflow on each pair: [History 'mt replicates pair 1'](#), [History 'mt replicates pair 2'](#), [History 'mt](#)

Display a menu

About this Page

Author

jtxt 

Related Pages

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

Tags

Community:
[cloud](#) [heteroplasmy](#) [ngs](#)

Yours:
[heteroplasmy](#) [cloud](#)
[ngs](#)

Workflows and other entities can also be embedded

The screenshot displays the Galaxy web interface for a workflow titled "Determine threshold from PCR replicates". The interface is divided into several sections:

- Top Bar:** Shows the Galaxy logo and navigation links: Analyze Data, Workflow, Data Libraries, Help, User.
- Left Sidebar:** Contains a list of tool categories: Get Data, Text Manipulation, Filter and Sort, Statistics, Join, Subtract and Group, Operate on Genomic Intervals, Graph/Display Data, NGS ToolBox Beta, NGS: QC and manipulation, NGS: Mapping, and NGS: SAM Tools.
- Central Canvas:** Displays the workflow steps:
 - Generate pileup:** Takes a BAM file and a file for the pileup as input.
 - Filter pileup:** Selects a dataset (out_file1 (tabular)).
 - Filter:** Filters the output (out_file1).
- Right Sidebar:** Contains details for the selected step (Filter):
 - Details:** Shows a dropdown menu for "lower than" with a value of 30. It also includes checkboxes for "Do not report positions with coverage lower than", "Only report variants?", "Convert coordinates to intervals?", "Print total number of differences?", and "Print quality and base string?".
 - Edit Step Attributes:** Includes an "Annotation / Notes" section with the text: "Replicate 2: Filter pileup for positions with high coverage (over 200 reads that map with quality of at least 30)".
- Bottom Section:** Shows the workflow steps in a list:
 - Step 16: Filter:** Output dataset 'out_file1' from step 14. With following condition: c1=='chrM' and c10 >= 200.
 - Step 17: Join:** Join. Output dataset 'out_file1' from step 15 with Output dataset 'out_file1' from step 16.

And imported for inspection, verification, and reuse

The power of Galaxy publishing and

- Galaxy's publishing features facilitate access and reproducibility without any extra leg work
- One click grants access to the actual analysis you performed to generate your original results
 - Not just data access: the full pipeline
 - Annotate each step

Jeremy Goecks



Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond^{1,2,6,9}, Samir Wadhawan^{3,6,7},
Francesca Chiaromonte⁴, Guruprasad Ananda^{1,3}, Wen-Yu Chung^{1,3,8},
James Taylor^{1,5,9}, Anton Nekrutenko^{1,3,9} and The Galaxy Team¹

[+ Author Affiliations](#)

Abstract

How many species inhabit our immediate surroundings? A straightforward collection technique suitable for answering this question is known to anyone who has ever driven a car at highway speeds. The windshield of a moving vehicle is subjected to numerous insect strikes and can be used as a collection device for representative sampling. Unfortunately the analysis of biological material collected in that manner, as with most metagenomic studies, proves to be rather demanding due to the large number of required tools and considerable computational infrastructure. In this study, we use organic matter collected by a

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>.]

Article published online before print. Article and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.094508.109>.

OPEN ACCESS ARTICLE

This Article

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

- » Abstract **Free**
- » Full Text (PDF) **Free**
- » Supplemental Material

- All Versions of this Article:
 - gr.094508.109v1
 - 19/11/2144 **most recent**

Article Category

Resource

+ Services

+ Citing Articles

+ Google Scholar

+ PubMed

+ Social Bookmarking

+ Recent Updates

[Follow us on twitter](#)

+ Most Read Articles

[View all ...](#)

Current Issue

October 2010, 20 (10)



+ From the Cover

Alert me to new issues of
Genome Research

- [Advance Online Articles](#)
- [Submit a Manuscript](#)
- [GR in the News](#)
- [Editorial Board](#)
- [E-mail Alerts & RSS Feeds](#)
- [Recommend to Your Library](#)
- [Job Opportunities](#)

Do you know
what your
current research
approach is
missing?

Galaxy deployment models

Galaxy Main (usegalaxy.org)

- ▶ ~130,000 jobs a month
- ▶ Every month is “best ever”
- ▶ Approximately 1Tb in user uploads per week
- ▶ Unsustainable in the long term
Community!

Building local Galaxy instances

- Galaxy is designed for local installation and customization
 - Just download and run, completely self-contained
 - Easily integrate new tools
 - Easy to deploy and manage on nearly any (unix) system
 - Run jobs on existing compute clusters

Scale up on existing resources

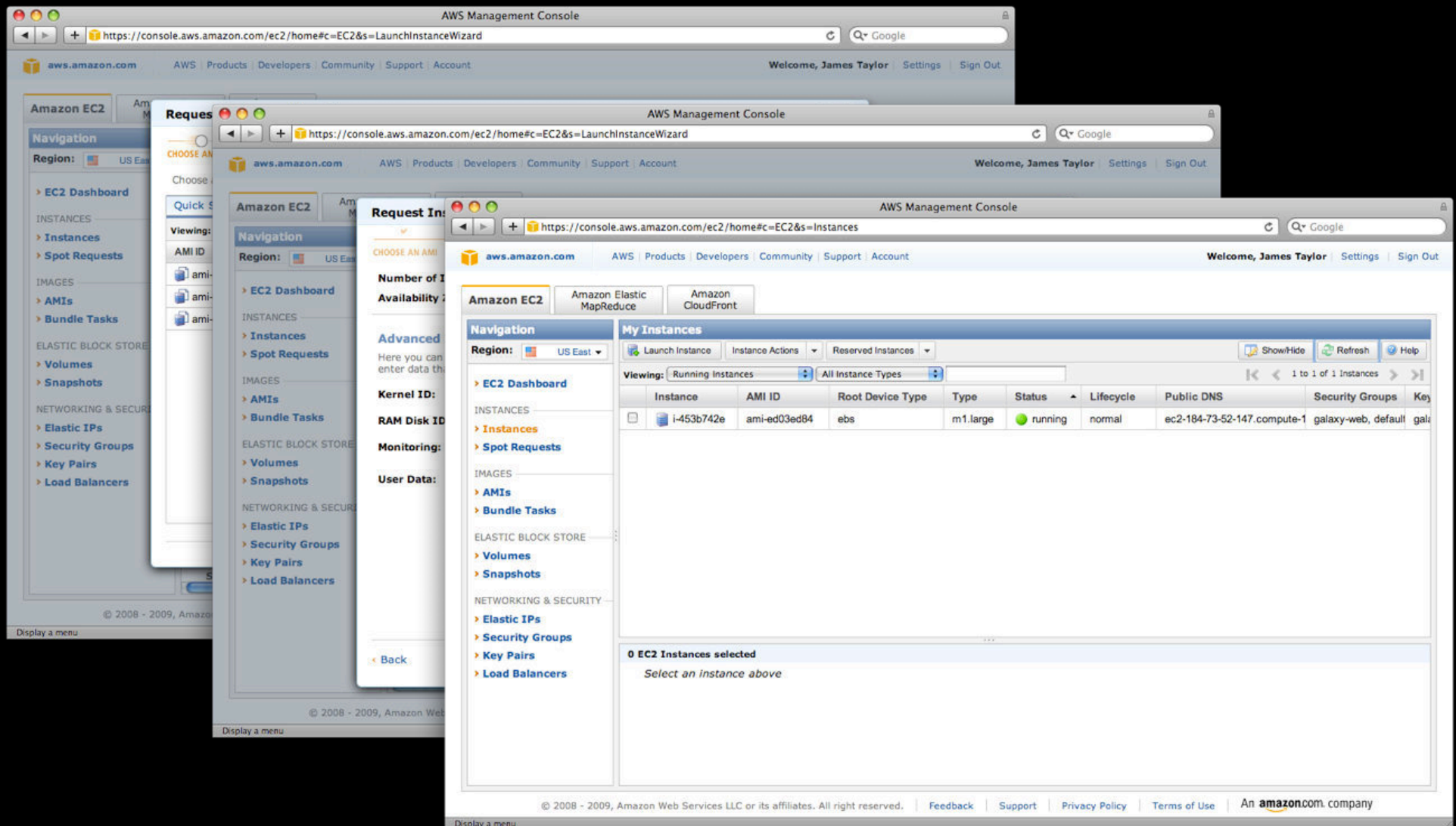
- Move intensive processing (tool execution) to other hosts
- Frees up the application server to serve requests and manage jobs
- Utilize existing resources
- Supports any scheduler that supports DRMAA (most of them)



Cloud computing

- On-demand resource acquisition fits well with the irregular resource needs of many labs working with sequence data
- Our goal is to approach the ease of use of a “software as a service” solution while maintaining the flexibility and control of an infrastructure based solution

Using Amazon EC2: Startup in 3 steps



Galaxy Cloud

+

http://ec2-174-129-103-83.compute-1.amazonaws.com/cloud

Q

Google

Galaxy

Info: [report bugs](#) | [wiki](#) | [screencasts](#)

Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as 'worker' nodes on which jobs are run.

Terminate cluster

Add nodes ▼

Remove nodes

Access Galaxy

Status

Cluster name: ttt

Disk status: 0 / 0 (0%)

Worker status: Idle: 0 Available: 0 Requested: 0

Service status: Applications Data

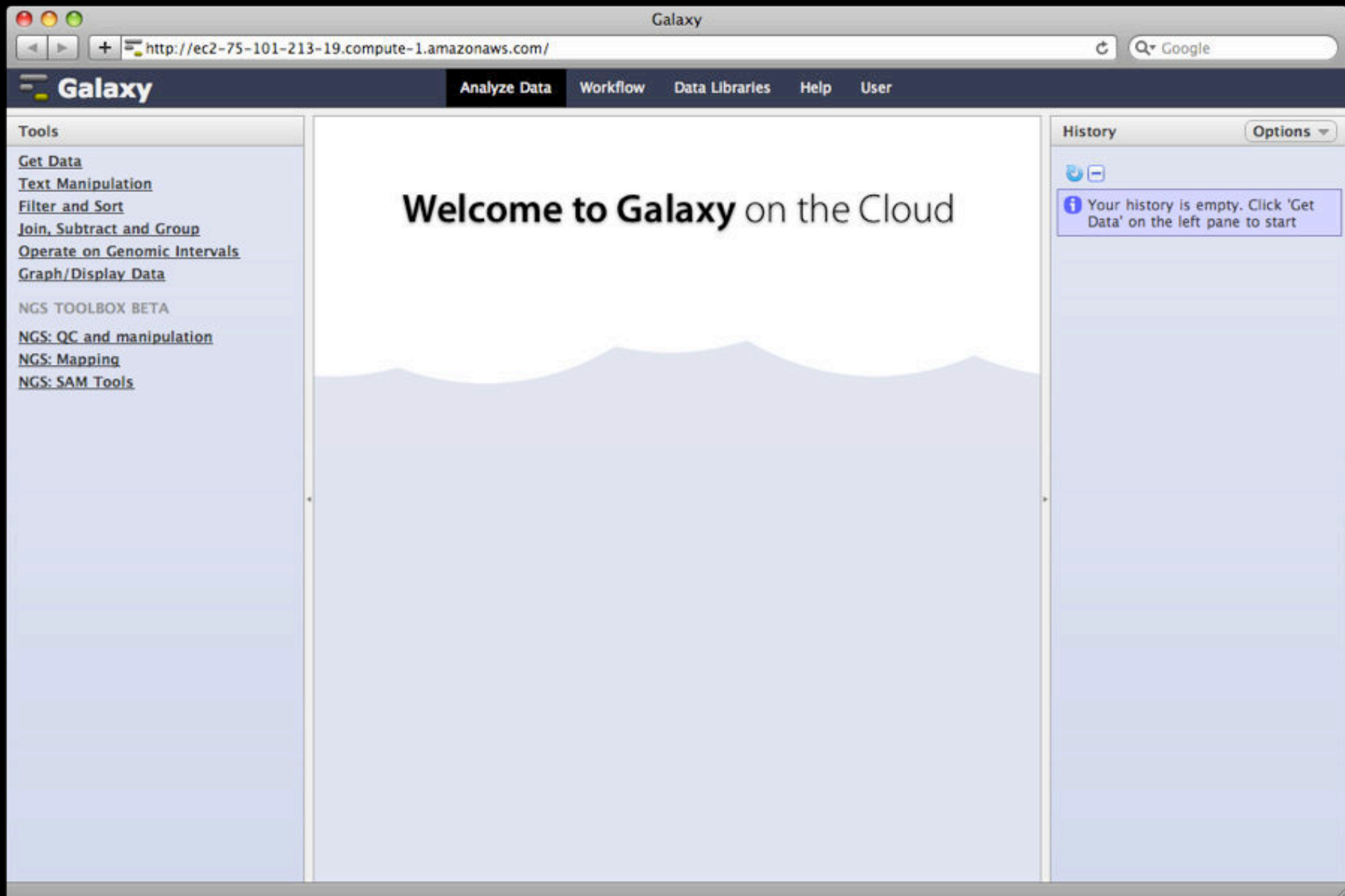
Pending

Starting

Ready

Error

Cluster status log



The image displays two overlapping screenshots of the Galaxy web interface. The left screenshot shows the 'Saved Histories' section, which includes a table of saved workflow histories and a sidebar with navigation links. The right screenshot shows the 'Galaxy Cloud Console', which provides management options for the Galaxy instance, including scaling and monitoring the cluster status log.

Saved Histories Table:

Name	Datasets (by state)	Tags	Sharing	Created	Last Used	
mt.replicates.pair.2	8	96	0 Tags	about 1 hour ago	2 m ago	
mt.replicates.pair.2	8	96	0 Tags	about 1 hour ago	15 min ago	
mt.replicates.pair.1.testing	35	3	66	0 Tags	about 2 hours ago	21 min ago
mt.datasets	24		0 Tags	about 2 hours ago	abo...	

Galaxy Cloud Console:

Scale

Status

Cluster name: james-galaxy-cluster-9May2010-1
 Cluster status: Ready
 Instance status: Idle: 0 Available: 4 Requested: 4

Cluster status log

```

14:54:40 - Instance i-a3e7b2c8 ready
14:54:40 - Setting up Galaxy
14:54:40 - Starting Galaxy...
14:54:45 - Instance i-a1e7b2c4 ready
14:54:49 - Instance i-a3e7b2c8 ready
14:54:56 - Instance i-a3e7b2c8 reported alive
14:54:56 - Sent master public key to worker instance i-a3e7b2c8.
14:55:00 - Adding instance i-a3e7b2c8 to SGE Execution Host list
14:55:01 - Successfully added instance i-a3e7b2c8 to SGE
14:55:01 - Waiting on worker instance i-a3e7b2c8 to configure itself...
14:55:09 - Instance i-a3e7b2c8 ready
14:55:16 - Galaxy started successfully!
14:55:16 - Ready for use
  
```

Can use like any other Galaxy instance, with additional compute nodes acquired and released (automatically) in response to usage

Galaxy Cloud

http://ec2-184-73-135-47.compute-1.amazonaws.com/cloud/

AWS Management Console Galaxy Cloud

Galaxy Cloudman


Info: [report bugs](#) | [wiki](#) | [screencast](#)


Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove 'worker' nodes for running jobs.



[Terminate cluster](#) [Add nodes ▼](#) [Remove nodes](#) [Access Galaxy](#)

Status


Cluster name: james-cm-31march 

Disk status: 181M / 100G (1%) 

Worker status: Idle: 0 Available: 0 Requested: 0

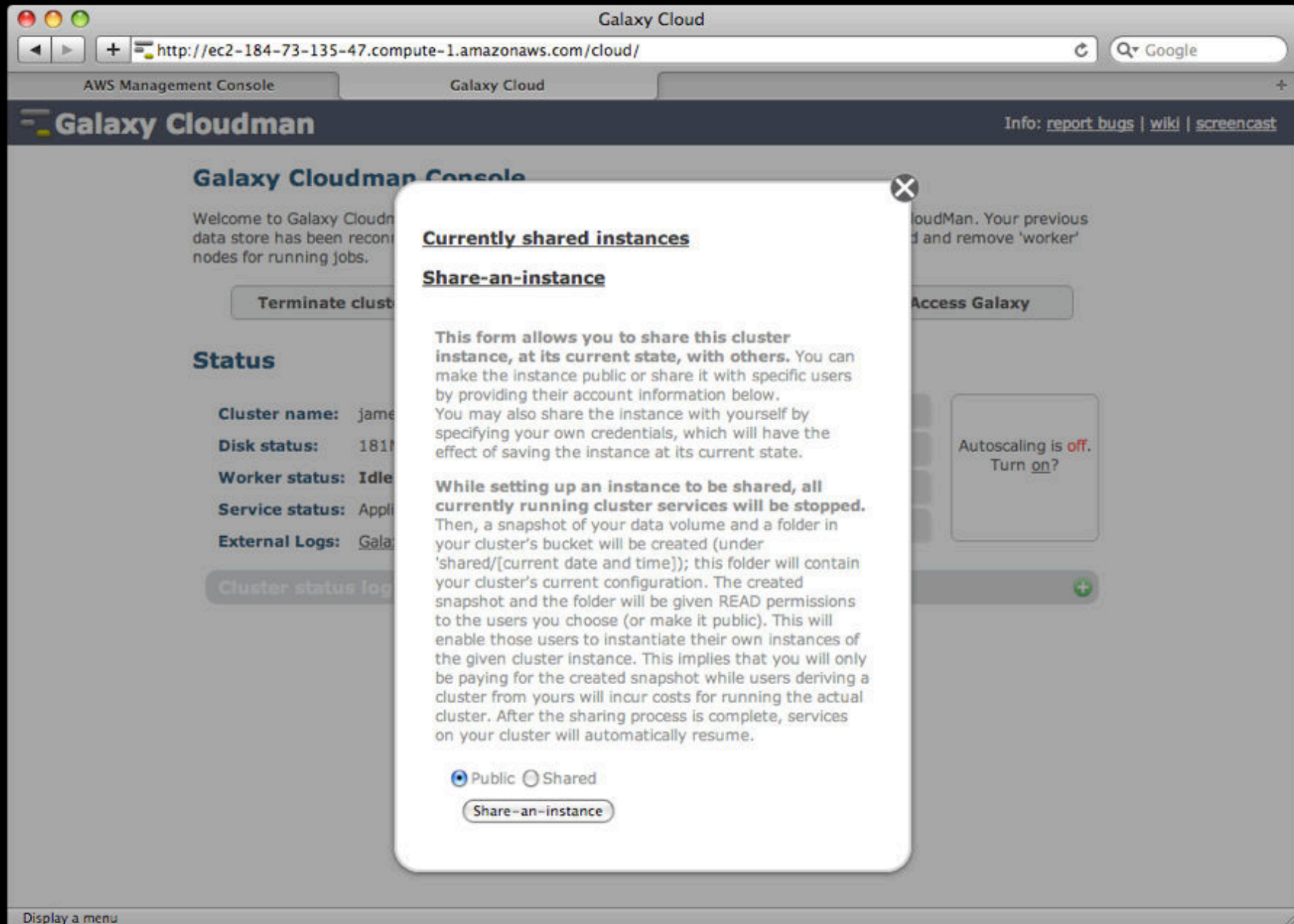
Service status: Applications  Data 

External Logs: [Galaxy Log](#)

Cluster status log 

Autoscaling is **off**. Turn **on**?

Share a snapshot of this instance



The next 90 – x minutes

- ▶ History (2005 – 2010)
- ▶ Present (2010 – 2011)
- ▶ The “Vision” (2011 – ∞)
- ▶ The Community
- ▶ Beer

Where do we go from here

- Why do we want to change the World?
- How are we going to do this?

Why Changing the World?

The workhorse: Illumina



HiSeq 2000

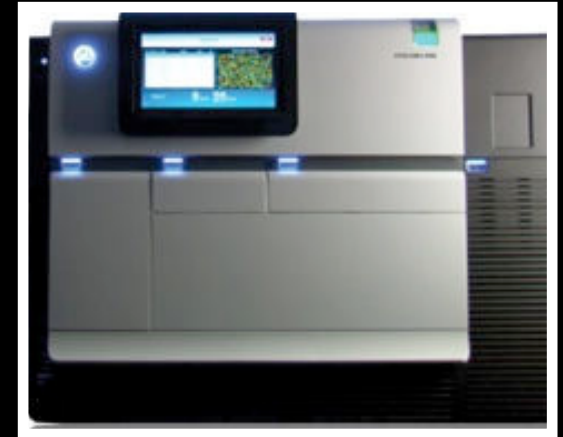
- ~25 GB per day
- \$16k–\$20k per run
- > 1Mb per dollar
- Can multiplex 192 samples per run
- as little as \$100 per sample!



454 GS / Junior:
40–400Mb runs, but
read lengths pushing
1kb



Ion Torrent PGM:
10Mb–1Gb runs,
200–400bp reads, 2
hour runtime, \$500!



PacBio RS: Direct
single molecule
sequencing, only 35k
reads, but long read
lengths, 30 minute
runs!

(plus nanopore and other single molecule
techniques on the horizon)

**Sequence data production capability
is
widely distributed**

Sequencing applications

Genome sequencing

- Direct sequencing of genomic DNA to resolve new genome sequences
- Direct deep sequencing + de novo assembly for novel genomes
- Re-sequencing to identify variations with respect to a reference
 - Single-end resequencing for SNP, copy number variation

RNA-seq for transcriptomics

- The diversity of (known) functional RNAs is enormous
- Even the best understood units (protein coding transcripts) are processed in myriad ways including alternative splicing
- In RNA-seq, capture a class of RNA, sequence (directly or through a cDNA clone)
- Reconstruct (possibly overlapping) RNA sequences and quantify the level at which they

Sequencing for functional annotation

- We can turn many functional annotation problems into sequencing problems (this is only a sample)
- The genome is relatively static within an individual, sequence it once and you are done
- Transcript levels, epigenomic modifications, and chromatin structure vary based on cell type, time, condition, ...
 - Enormous potential for data generation

“Democratization of sequencing”

- Because of the diverse utility of sequencing based assays, investigators across all of biology seek to take advantage of these techniques
- Large community data production projects have become relatively rare, data production is increasingly investigator driven
- Democratization of sequencing has not yet been matched by democratization of analysis infrastructure, burden is largely on the investigator
- Use of these techniques requires sophisticated and computationally intensive approaches

Most biologists don't write code

- High throughput data is very new to Biology, programming is not part of the training (this is not Astronomy...)
- Efforts like Bioperl and Bioconductor (R) have enabled some to pick it up, too often just enough to be dangerous

Much bioinformatics software is “research quality”

- Most software is written for a specific publication
- Poor performance and scalability, not designed with reuse in mind
- The rate with which underlying technologies and methods change makes it pointless to invest in improving
- Difficult to publish purely software papers in good journals, publish updates or improvements to existing software
- Pressure to use only software that is published

Commercial Bioinformatics Software is sustained by ignorance



Key Reproducibility Problems

- **Datasets:** not all available, difficult to access
- **Tools:** inaccessible, hard to record details
- **Publication:** results, data, methods separate

Microarray Experiment Reproducibility

- 18 Nat. Genetics microarray gene expression experiments
- Less than 50% reproducible
- Problems
 - missing data (38%)
 - missing software, hardware details (50%)
 - missing method, processing details

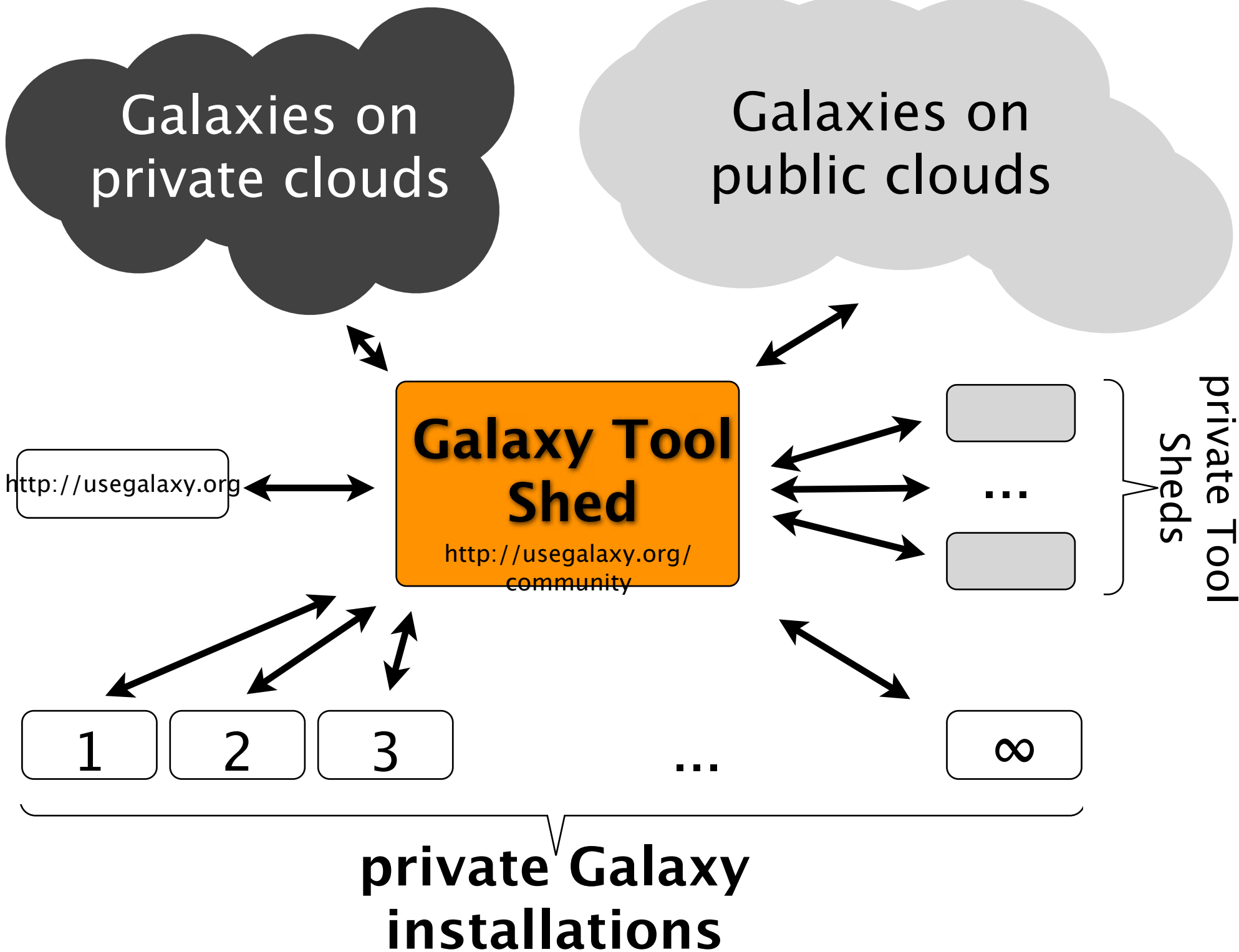
Ioannidis, J.P.A. et al. Repeatability of published microarray gene expression analyses. Nat Genet 41, 149–155 (2009)

NGS Re-sequencing Experiment Reproducibility

- 14 re-sequencing experiments in Nat. Genetics, Nature, and Science (2010)
- 0% reproducible?
- Problems
 - limited access to primary data (50%)
 - some or all tools unavailable (50%)
 - settings & versions not provided (100%)

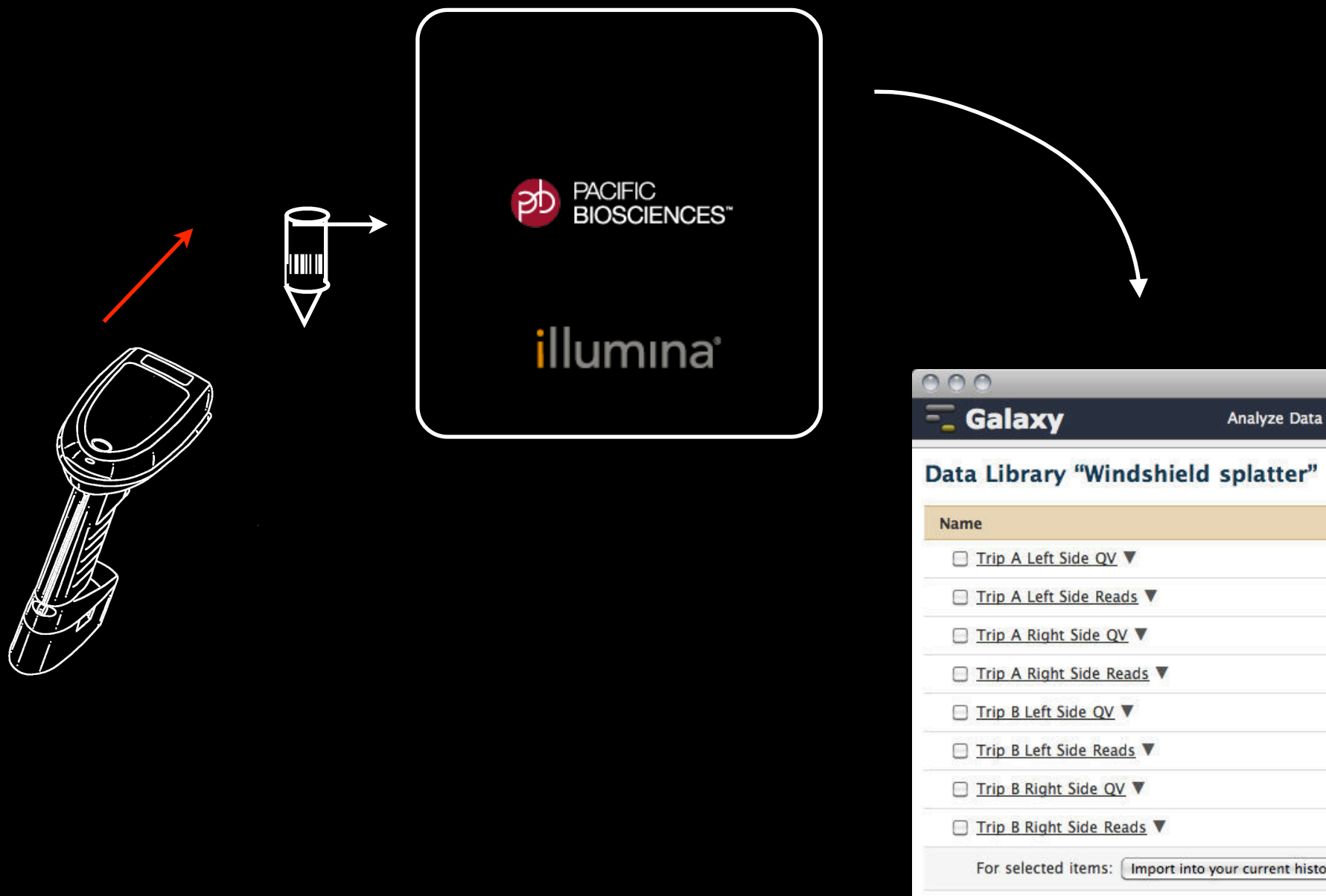
Going Forward

- Development of best practices
- Simplifying deployment of Galaxy and dependencies
- Tool Shed as the HUB
- Transparent publishing of analyses for reproducible research
- Tight Integration with NGS instruments



Sample tracking and instrument integration

(work in progress)



Greg Von Kuster, Dannon Baker

Galaxy

http://localhost:8080/requests

Galaxy Analyze Data Workflow Shared Data Lab Visualization Admin Help User

Request Actions ▾

Add Samples to Sequencing Request "Snail transcriptome"

Name	State	Data Library	Folder	History	Workflow
Sample 1 (required)		Select one ▾	Select one ▾	Lambda History ▾	lambda ▾ Input BAM: Coverage (gff) ▾ Reference Genome: lambda_ref.fasta ▾

For each sample, select the data library and folder in which you would like the run datasets deposited. To automatically run a workflow on run datasets, select a history first and then the desired workflow.

▶ **Additional information**

Copy samples from sample ▾

Select the sample from which the new sample should be copied or leave selection as None to add a new "generic" sample.

Click the Add sample button for each new sample and click the Save button when you have finished adding samples.

▶ **Import samples from csv file**

Display a menu

Consumer creating sequencing requests in Galaxy interface, and

Galaxy Administration

Galaxy

Analyze Data Workflow Shared Data Lab Admin Help User

Administration

- Security
 - Manage users
 - Manage groups
 - Manage roles
- Data
 - Manage data libraries
- Server
 - Reload a tool's configuration
 - Profile memory usage
 - Manage jobs
- Forms
 - Manage forms
- Sample Tracking
 - Sequencer configurations
 - Sequencing requests
 - Find samples

Sequencer configuration "Core Facility 454" Browse this request

Select files for transfer

Sample:
Sample 1

Select the sample with which you want to associate the datasets

Folder path on the sequencer:
/data/run1/ List contents Up

run1.fa
run1.qv

Select & show datasets Select more

Simplest scenario, lab manager manually imports run data and

Galaxy

http://localhost:8080/requests

Galaxy

Analyze Data Workflow Shared Data Lab Visualization Admin Help User

Request Actions ▾

Sequencing request "Snail transcriptome"

Current state:
[Complete](#)

Description:

User:
james.taylor@emory.edu

Request type:
Pacific Biosciences

▶ More

Samples

Name	Barcode	State	Data Library	Folder	History	Workflow	Run Datasets
Sample_1		Done	Pacific Biosciences	Pacific Biosciences	Lambda History	lambda	6

▶ Additional information

Display a menu

Within instrument integration plugins, data acquired automatically and

Sample tracking is completely

- Track manually, with barcodes, or integrate with an existing LIMS
- Everything is configuration driven, capture whatever data and support whatever workflow you want
- Interaction with sequence instruments and secondary analysis is completely pluggable
 - For services that provide a web / REST

**ONLY AS A COMMUNITY WE CAN
ACHIEVE THESE GOALS!**

