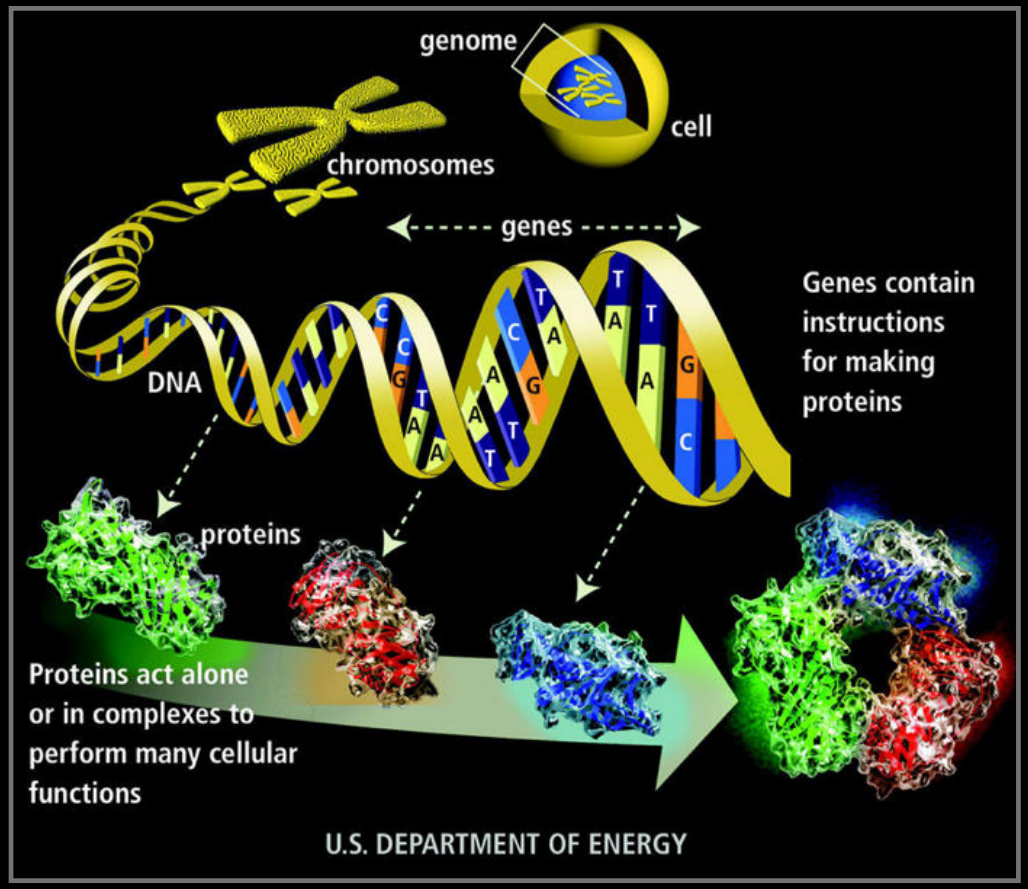
# Galaxy: A Web-based Platform for Accessible, Reproducible, and Transparent High-throughput (Genome) Biology

Jeremy Goecks
Depts. of Biology and Math & Computer Science
Emory University

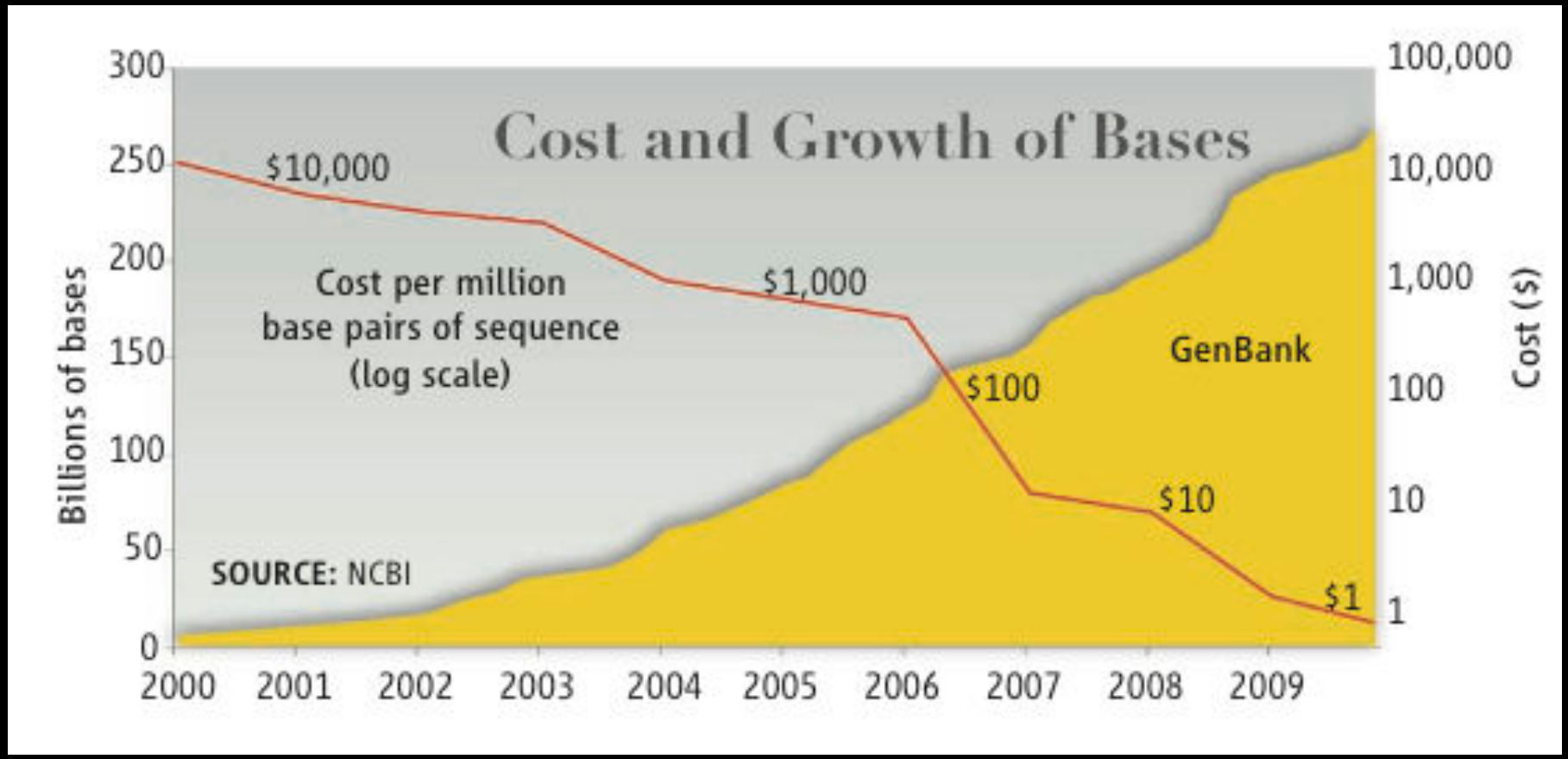
# Genomics




Identify and annotate all functional genomic elements

Understand interactions among elements

Apply genome knowledge to address biomedical challenges


Cost and Growth of Bases
Billions of bases
300
250
200
150
100
50
0
100,000
10,000
1,000
100
10
1
Cost ($)
$10,000
$1,000
$100
$10
$1
Cost per million
base pairs of sequence
(log scale)
GenBank
SOURCE: NCBI
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009

11 FEBRUARY 2011 VOL 331 **SCIENCE** www.sciencemag.org
*Published by AAAS*

# Will Computers Crash Genomics?

**New technologies are making sequencing DNA easier and cheaper than ever, but the ability to analyze and store all that data is lagging**

you-go service, accessible from one's own desktop, that provides rented time on a large cluster of machines that work together in parallel as fast as, or faster than, a single powerful computer. "Surviving the data deluge means computing in parallel," says Michael



"Will Computers Crash Genomics?", Pennisi, E., *Science*, Feb 11, 2011

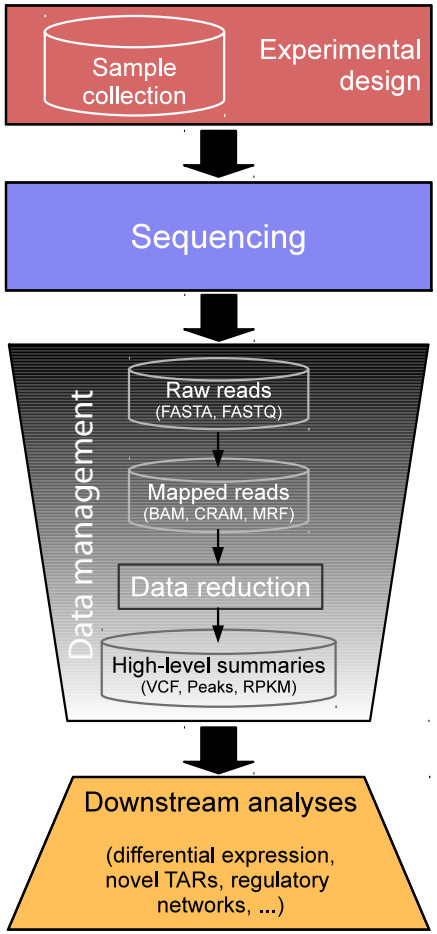
Sboner et al. Genome Biology 2011, 12:125
http://genomebiology.com/2011/12/8/125

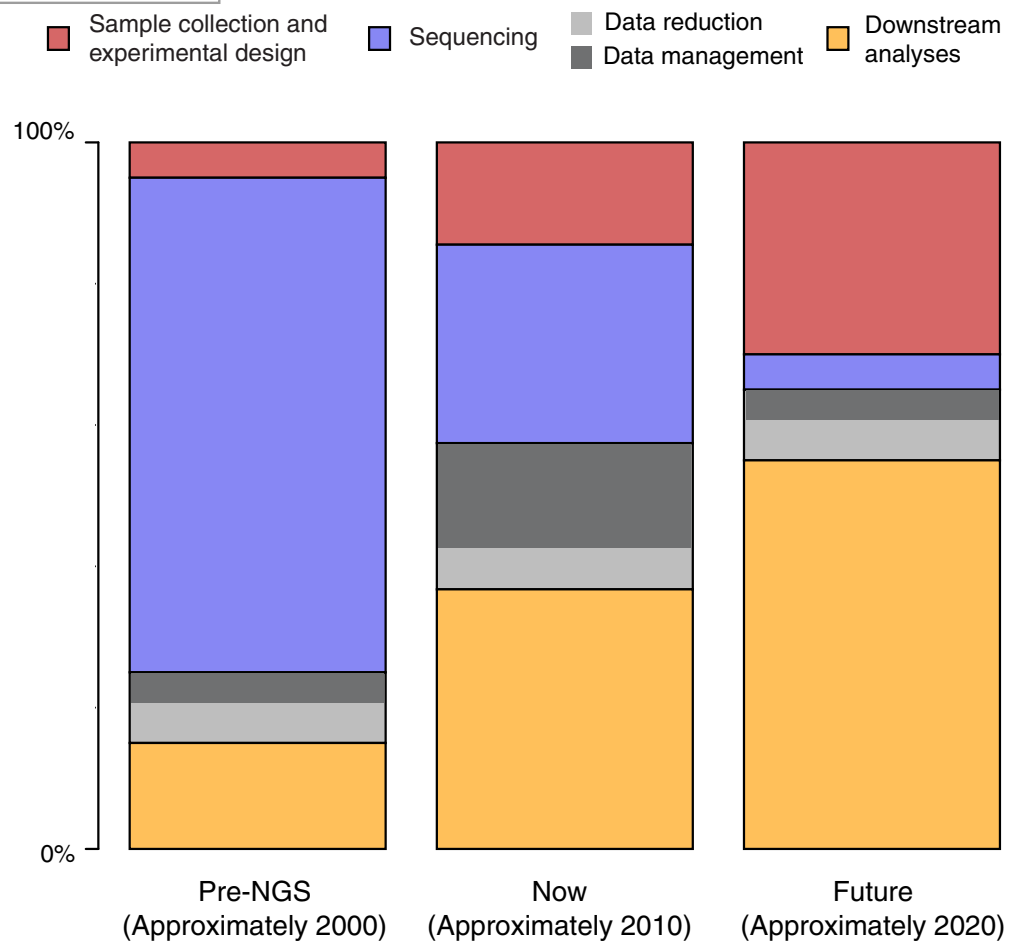
Genome Biology

OPINION

# The real cost of sequencing: higher than you think!

Andrea Sboner[1,2], Xinmeng Jasmine Mu[1], Dov Greenbaum[1,2,3,4,5], Raymond K Auerbach[1] and Mark B Gerstein*[1,2,6]



# A Key Challenge in Genomics

Generating data is easy

- high-throughput/next-generation sequencing (HTS/NGS) technologies improving rapidly
- datasets are many MBs or GBs

Analyzing data is THE bottleneck

- computation is essential due to dataset size

# Computation in Science?

Scientists unfamiliar with computation

Reproducibility hindered by complexity: systems, scripts, tools, parameters

Collaboration and publishing difficult because current media do not support computational artifacts well

# Galaxy Project: Fundamental Questions

When Biology (or any science) becomes dependent on computational methods:

- how best to make methods accessible to scientists?
- how best to ensure that analyses are reproducible?
- how best to enable transparent communication and reuse of analyses?

# Vision

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

# What is Galaxy?

**GUI for genomics**

- for complete analyses: analyze, visualize, share, publish

**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

# Accessibility

# Accessibility

Galaxy

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

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User

Tools | Options

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: Indel Analysis
NGS: Peak Calling

RGENETICS

SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models

Workflows

**Map with Bowtie for Illumina**

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

Built-ins were indexed using default options

Select a reference genome: mm9

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

Is this library mate-paired?: Paired-end

Forward FASTQ file: 1: E18 PE.1 Reads

Must have Sanger-scaled quality values with ASCII offset 33

Reverse FASTQ file: 1: E18 PE.1 Reads

Must have Sanger-scaled quality values with ASCII offset 33

Maximum insert size for valid paired-end alignments (-X): 1000

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

Bowtie settings to use: Commonly used

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

Execute

**What it does**

Bowtie is a short read aligner designed to be ultrafast and memory-efficient. It is developed by Ben Langmead and Cole Trapnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10:R25.

History | Options

Imported: SNP Pileup Analysis for Sample E18

15: Variants from sample E18, consensus different, in RefSeq Genes
14: UCSC mm9 RefSeq Genes
13: Variants from sample E18 where consensus base different than ref. base
10: Variants from sample E18
9: Generate pileup on data 8
8: SAM-to-BAM on data 7
7: Map with Bowtie for Illumina on data 6 and data 5
6: E18 PE.2 Reads Groomed, Trimmed
5: E18 PE.1 Reads Groomed, Trimmed
4: E18 PE.2 Reads Groomed
3: E18 PE.1 Reads Groomed
2: E18 PE.2 Reads
1: E18 PE.1 Reads

# Accessibility

**Filter and Sort**

- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an e

**Operate on Genomic Intervals**

- Intersect the intervals of two queries
- Subtract the intervals of two queries
- Merge the overlapping intervals of a query

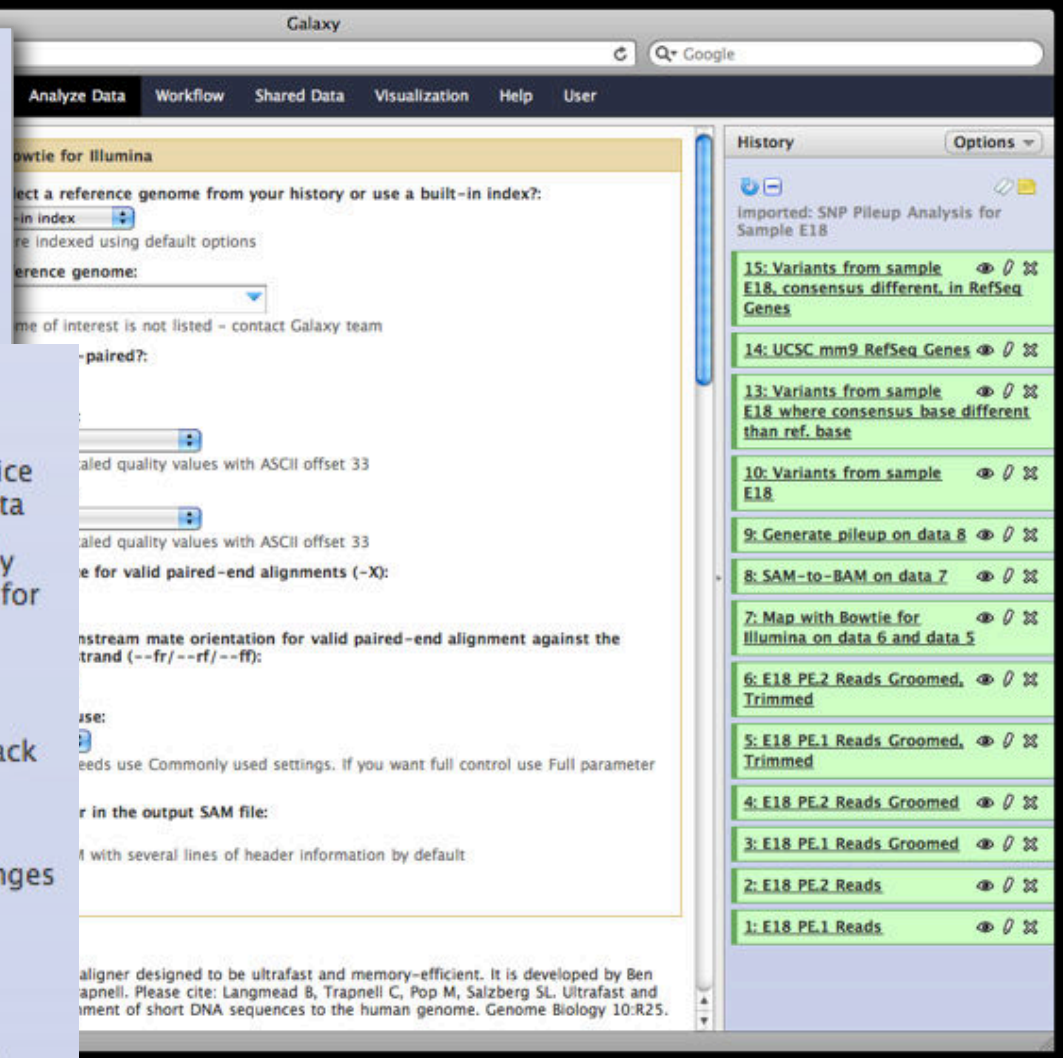
**NGS: RNA Analysis**

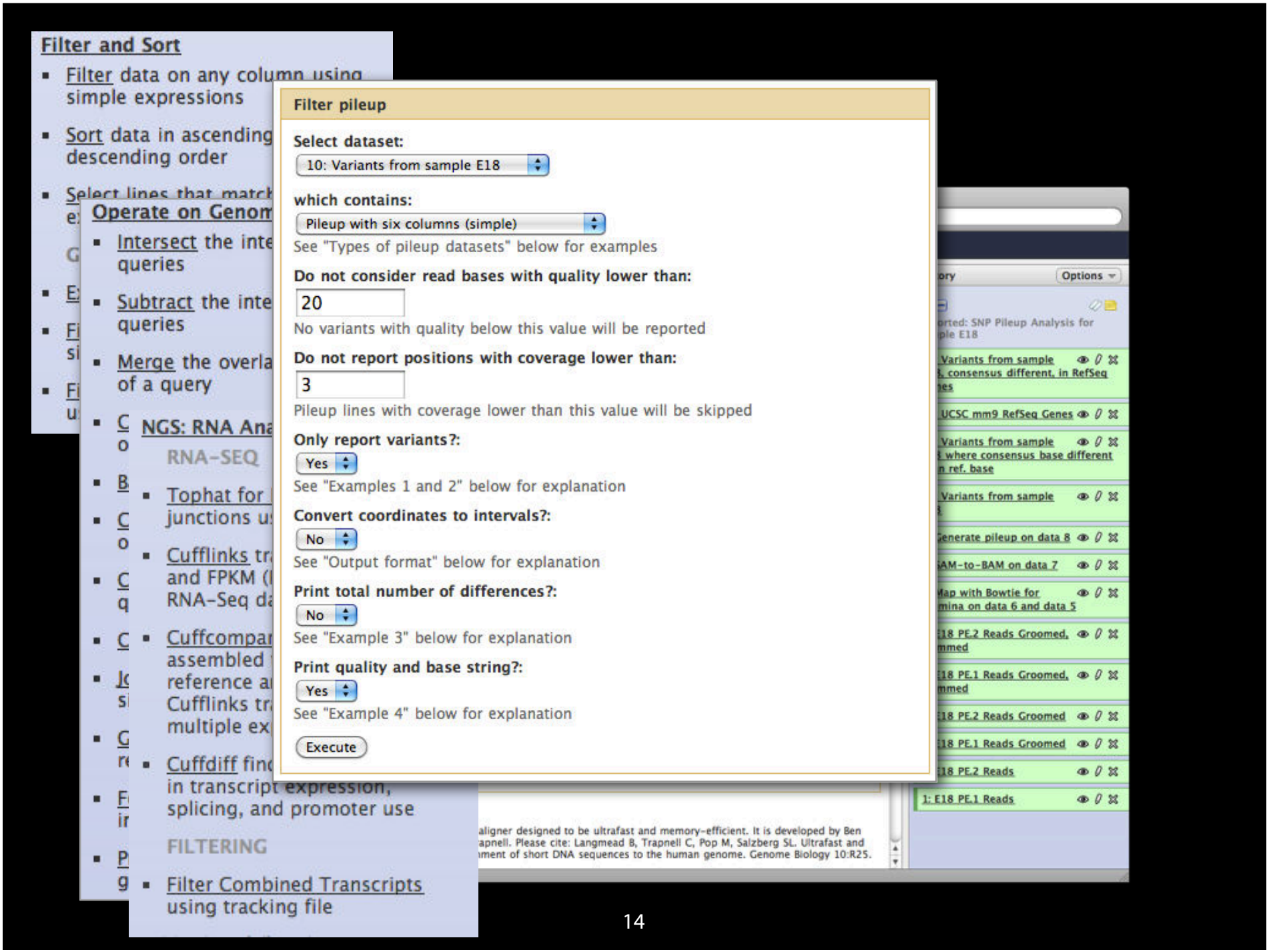
RNA-SEQ

- Tophat for Illumina 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




Filter and Sort
Filter data on any column using simple expressions
Sort data in ascending descending order
Operate on Genom
Intersect the inte queries
Subtract the inte queries
Merge the overla of a query
NGS: RNA Ana
RNA-SEQ
Tophat for junctions u
Cufflinks tr and FPKM ( RNA-Seq da
Cuffcompar assembled reference a Cufflinks tr multiple ex
Cuffdiff find in transcript expression, splicing, and promoter use
FILTERING
Filter Combined Transcripts using tracking file
Filter pileup
Select dataset:
10: Variants from sample E18
which contains:
Pileup with six columns (simple)
See "Types of pileup datasets" below for examples
Do not consider read bases with quality lower than:
20
No variants with quality below this value will be reported
Do not report positions with coverage lower than:
3
Pileup lines with coverage lower than this value will be skipped
Only report variants?:
Yes
See "Examples 1 and 2" below for explanation
Convert coordinates to intervals?:
No
See "Output format" below for explanation
Print total number of differences?:
No
See "Example 3" below for explanation
Print quality and base string?:
Yes
See "Example 4" below for explanation
Execute
Options
1: E18 PE.1 Reads

Filter and Sort

- Filter data on any column using simple expressions
- Sort data in ascending descending order
- Select lines that match

Operate on Genom

- Intersect the inte queries
- Subtract the inte queries
- Merge the overla of a query

NGS: RNA Ana

RNA-SEQ

- Tophat for junctions u
- Cufflinks tr and FPKM ( RNA-Seq da
- Cuffcompar assembled reference a Cufflinks tr multiple ex
- Cuffdiff find in transcript expression, splicing, and promoter use

FILTERING

- Filter Combined Transcripts using tracking file

Filter pileup

Select dataset:

10: Variants from sample E18

which contains:

Pileup with six columns (simple)

See "Types of pileup datasets" below for examples

Do not consider read bases with quality lower than:

20

No variants with quality below this value will be reported

Do not report positions with coverage lower than:

3

Pileup lines with coverage lower than this value will be skipped

Only report variants?:

Yes

See "Examples 1 and 2" below for explanation

Convert coordinates to intervals?:

No

See "Output format" below for explanation

Print total number of differences?:

No

See "Example 3" below for explanation

Print quality and base string?:

Yes

See "Example 4" below for explanation

Execute

aligner designed to be ultrafast and memory-efficient. It is develo
apnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL.
ment of short DNA sequences to the human genome. Genome Bi

History Options

Variant Analysis for Sample E18

15: Intersect to get Variants from sample E18, consensus different, in RefSeq Genes

14: UCSC mm9 RefSeq Genes

13: Filter to get Variants from sample E18 where consensus base different than ref. base

10: Filter pileup to get Variants from sample E18

9: Generate pileup on data 8

8: SAM-to-BAM on data 7

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

6: E18 PE.2 Reads Groomed, Trimmed

5: E18 PE.1 Reads Groomed, Trimmed

Filter and Sort

- Filter data on any co simple expressions
- Sort data in ascendir descending order
- Select lines that mat

Operate on Geno

- Intersect the in queries
- Subtract the int queries
- Merge the over of a query

NGS: RNA Ar

RNA-SEQ

- Tophat fo junctions
- Cufflinks and FPKM RNA-Seq
- Cuffcomp assembled reference Cufflinks multiple e
- Cuffdiff fi in transcript expression, splicing, and promoter use

FILTERING

- Filter Combined Transcripts using tracking file

This dataset is large and only the first megabyte is shown below.
Show all | Save

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| chr10 | 6882036 | 6882037 | A | A | 107 | 0 | 60 | 32 | .$..,.....C..... |
| chr10 | 14243075 | | 14243076 | | G | G | 96 | 0 | 60 | 35 | t. |
| chr10 | 14243079 | | 14243080 | | C | C | 106 | 0 | 60 | 35 | .. |
| chr10 | 14465082 | | 14465083 | | T | K | 173 | 176 | 60 | 35 | GG |
| chr10 | 14465083 | | 14465084 | | G | K | 144 | 144 | 60 | 35 | .. |
| chr10 | 14465084 | | 14465085 | | T | T | 117 | 0 | 60 | 38 | .. |
| chr10 | 14465085 | | 14465086 | | G | G | 70 | 0 | 60 | 38 | .. |
| chr10 | 14465257 | | 14465258 | | C | C | 79 | 0 | 60 | 42 | .. |
| chr10 | 14465258 | | 14465259 | | A | A | 137 | 0 | 60 | 46 | .. |
| chr10 | 14465263 | | 14465264 | | A | A | 136 | 0 | 60 | 61 | .. |
| chr10 | 14465366 | | 14465367 | | A | A | 101 | 0 | 60 | 38 | g$ |
| chr10 | 14465371 | | 14465372 | | G | G | 137 | 0 | 60 | 50 | .$ |
| chr10 | 14465410 | | 14465411 | | G | G | 184 | 0 | 60 | 69 | .$ |
| chr10 | 14465447 | | 14465448 | | T | T | 186 | 0 | 60 | 65 | .$ |
| chr10 | 14465456 | | 14465457 | | G | G | 193 | 0 | 60 | 70 | .. |
| chr10 | 14465465 | | 14465466 | | T | T | 177 | 0 | 60 | 63 | .$ |
| chr10 | 14465485 | | 14465486 | | C | T | 129 | 129 | 60 | 34 | t$ |
| chr10 | 14465569 | | 14465570 | | T | T | 219 | 0 | 60 | 84 | .. |
| chr10 | 14465581 | | 14465582 | | G | G | 240 | 0 | 60 | 84 | ,$ |
| chr10 | 14465586 | | 14465587 | | C | C | 248 | 0 | 60 | 82 | .$ |
| chr10 | 14465621 | | 14465622 | | C | C | 134 | 0 | 60 | 49 | .. |
| chr10 | 14465658 | | 14465659 | | C | C | 134 | 0 | 60 | 49 | ,, |
| chr10 | 14465660 | | 14465661 | | T | T | 153 | 0 | 60 | 55 | ,, |
| chr10 | 14465691 | | 14465692 | | G | G | 128 | 0 | 60 | 42 | .$ |
| chr10 | 14465778 | | 14465779 | | C | C | 89 | 0 | 60 | 34 | ,$ |
| chr10 | 14465791 | | 14465792 | | G | G | 104 | 0 | 60 | 33 | ,$ |
| chr10 | 14465881 | | 14465882 | | G | G | 110 | 0 | 60 | 41 | .. |
| chr10 | 17445088 | | 17445089 | | A | A | 103 | 0 | 60 | 34 | .. |
| chr10 | 17445271 | | 17445272 | | A | A | 55 | 0 | 60 | 34 | .. |
| chr10 | 17731269 | | 17731270 | | T | T | 113 | 0 | 60 | 42 | ,$ |
| chr10 | 19928287 | | 19928288 | | G | A | 135 | 135 | 60 | 36 | AA |
| chr10 | 19928468 | | 19928469 | | C | T | 132 | 132 | 60 | 35 | T$ |
| chr10 | 19928488 | | 19928489 | | A | A | 119 | 0 | 60 | 44 | .. |
| chr10 | 19928494 | | 19928495 | | C | T | 138 | 138 | 60 | 37 | TT |
| chr10 | 19928527 | | 19928528 | | A | A | 134 | 0 | 60 | 45 | ,, |
| chr10 | 19928538 | | 19928539 | | G | G | 144 | 0 | 60 | 52 | ,$ |
| chr10 | 19928543 | | 19928544 | | A | G | 147 | 147 | 60 | 40 | Gg |
| chr10 | 19928741 | | 19928742 | | T | T | 80 | 0 | 60 | 30 | ,, |
| chr10 | 20799826 | | 20799827 | | G | G | 117 | 0 | 60 | 37 | ,$ |
| chr10 | 28750217 | | 28750218 | | C | T | 138 | 138 | 60 | 37 | TT |
| chr10 | 28750397 | | 28750398 | | A | C | 154 | 211 | 60 | 64 | C$ |
| chr10 | 28750401 | | 28750402 | | A | A | 128 | 0 | 60 | 47 | .$ |
| chr10 | 28750423 | | 28750424 | | C | T | 113 | 113 | 60 | 35 | T$ |
| chr10 | 28750438 | | 28750439 | | A | A | 95 | 0 | 60 | 36 | .$ |
| chr10 | 28750446 | | 28750447 | | A | G | 165 | 165 | 60 | 46 | G$ |
| chr10 | 28750487 | | 28750488 | | A | A | 80 | 0 | 60 | 31 | ,, |
| chr10 | 28750512 | | 28750513 | | G | G | 220 | 0 | 60 | 72 | .$ |
| chr10 | 28750548 | | 28750549 | | G | C | 255 | 255 | 60 | 97 | C$ |
| chr10 | 28750574 | | 28750575 | | T | T | 237 | 0 | 60 | 83 | .$ |
| chr10 | 28750577 | | 28750578 | | T | T | 234 | 0 | 60 | 82 | ,$ |
| chr10 | 28750578 | | 28750579 | | T | T | 242 | 0 | 60 | 76 | ,$ |
| chr10 | 28750593 | | 28750594 | | G | G | 220 | 0 | 60 | 75 | .$ |
| chr10 | 28750640 | | 28750641 | | T | C | 165 | 165 | 60 | 46 | Cc |
| chr10 | 28750746 | | 28750747 | | G | A | 202 | 202 | 60 | 58 | AA |
| chr10 | 28750766 | | 28750767 | | A | G | 205 | 205 | 60 | 59 | G$ |
| chr10 | 28750769 | | 28750770 | | T | C | 175 | 175 | 60 | 49 | cc |

aligner designed to be ultrafast and memory-efficient. It is develo
apnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. U
ment of short DNA sequences to the human genome. Genome Bi

Options

nalysis for Sample E18

sect to get Variants
nple E18, consensus different,
Genes

mm9 RefSeq Genes

to get Variants from
E18 where consensus base
than ref. base

pileup to get
from sample E18

ate pileup on data 8

o-BAM on data 7

ith Bowtie for
on data 6 and data 5

6: E18 PE.2 Reads Groomed, Trimmed

5: E18 PE.1 Reads Groomed, Trimmed

# Galaxy Accessibility

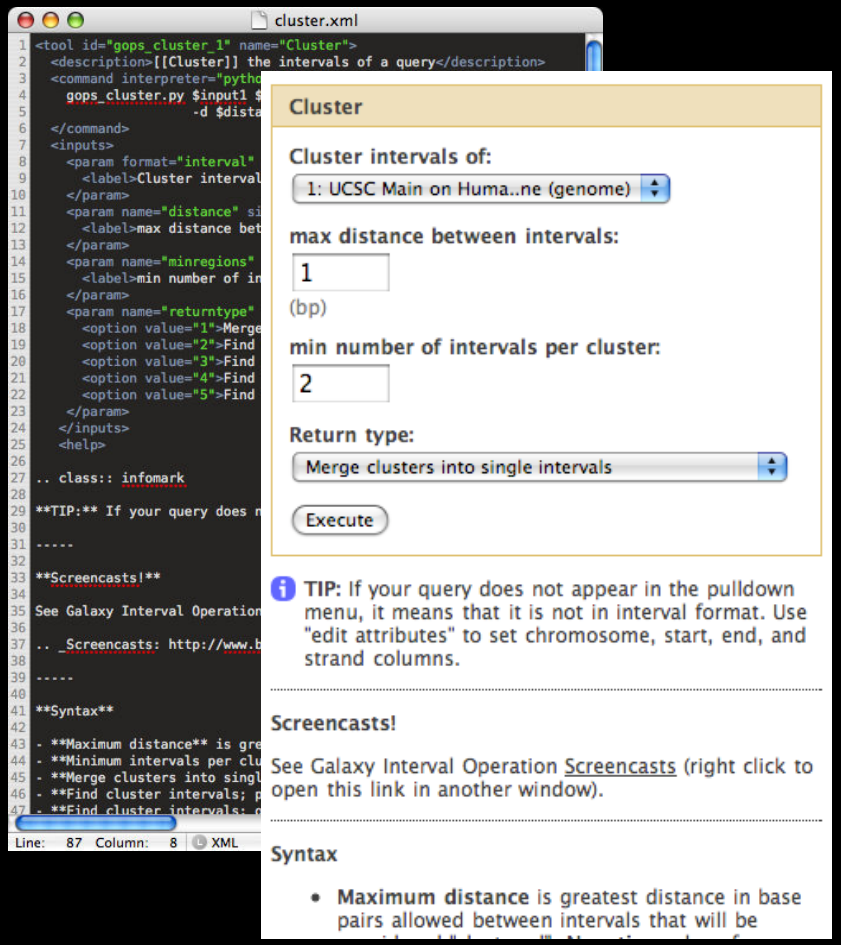
Only a Web browser is required

Standardization

- tools, parameters, outputs all look the same

Easy to use output of one tool as input for another tool

# Accessibility for Tool Developers




Defined via abstract interface:

- inputs & outputs
- parameters
- how to generate command line

As simple as possible but allows for rigorous

**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
- FASTQ Masker by quality score

FASTQC: FASTQ/SAM/BAM

- Fastqc: Fastqc QC using FastQC from Babraham

**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: Picard (beta)**

QC/METRICS FOR SAM/BAM

- BAM Index Statistics
- Sam/bam Alignment Summary Metrics
- Sam/bam GC Bias Metrics
- Estimate Library Complexity
- Insertion size metrics for PAIRED data
- Sam/bam Hybrid Selection Metrics For (eg exome) targeted data

BAM/SAM CLEANING

- Add or Replace Groups
- Reorder SAM
- Replace Sam Header
- Paired Read Mate Fixer for paired data
- Mark Duplicate reads

**NGS: GATK Tools**

REALIGNMENT

- Realigner Target Creator for use in local realignment
- Indel Realigner - perform local realignment

BASE RECALIBRATION

- Count Covariates on BAM files
- Table Recalibration on BAM files
- Analyze Covariates - perform local realignment

GENOTYPING

- Unified Genotyper SNP and indel caller

**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

RGENETICS

Dozens of tools for different HTS applications packaged with Galaxy

# Amplification

Many tools available in a single place means that tools can be combined in novel ways

Users, developers, community benefit

# Reproducibility

# Reproducibility in Genomics

18 *Nat. Genetics* experiments in microarray gene expression

<50% of reproducible

Problems

- missing data (38%)
- missing software, hardware details (50%)
- missing method, processing details (66%)

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

# Reproducibility in Genomics

18 *Nat. Genetics* experiments in microarray gene expression

<50% of reproducible

Problems

- missing data (38%)
- missing software, hardware details (50%)
- missing method, processing details (66%)

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

14 re-sequencing experiments in *Nat. Genetics*, *Nature*, *Science*

0% reproducible?

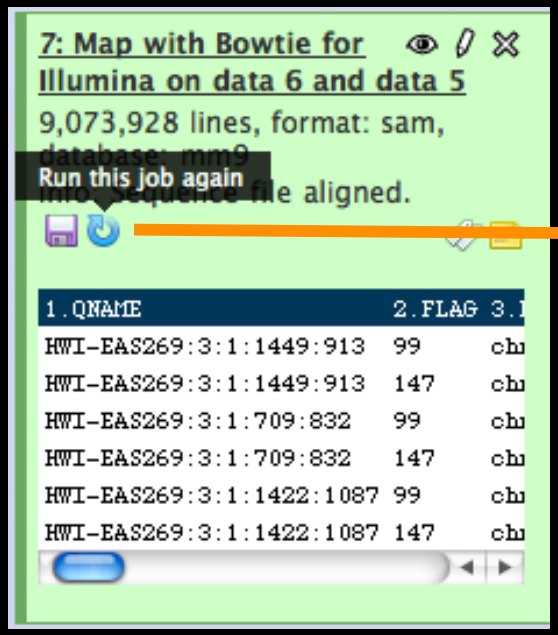
Problems

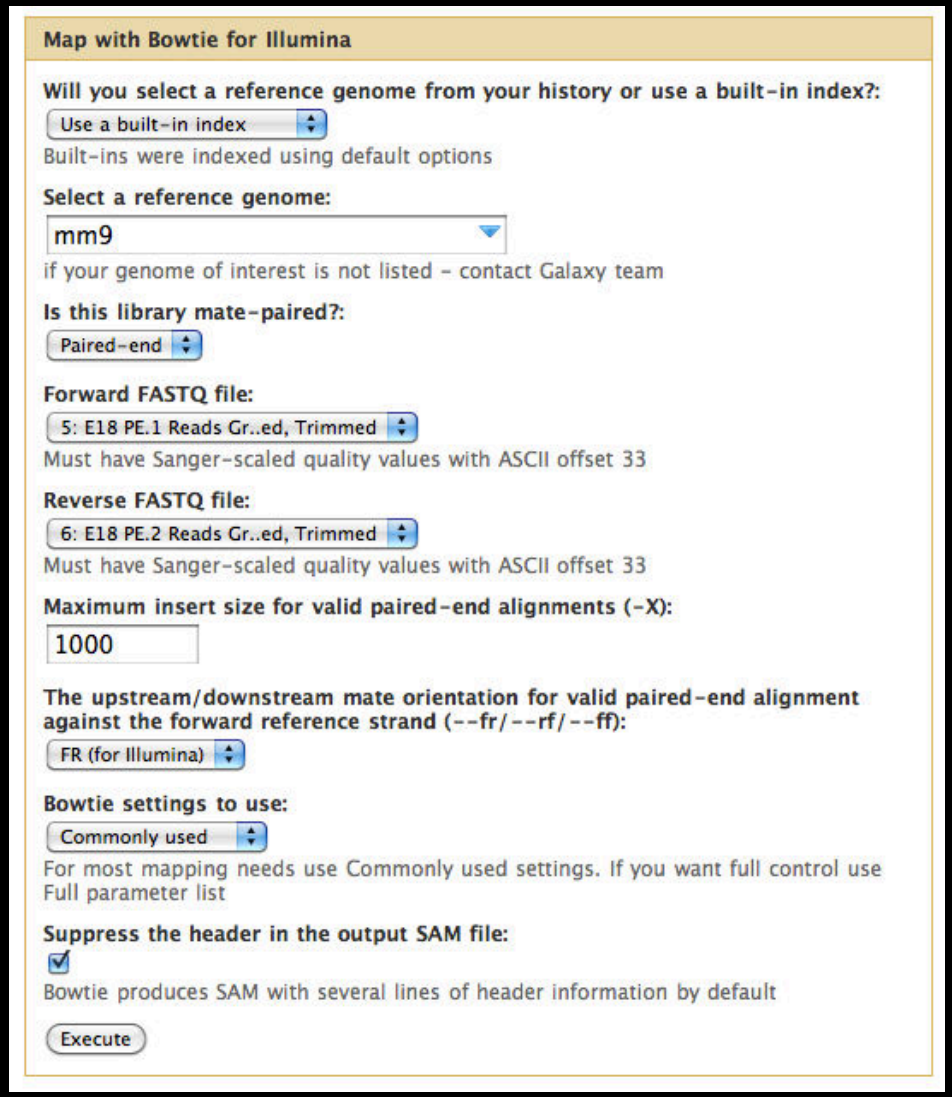
- missing primary data (50%)
- tools unavailable (50%)
- missing parameter setting, tool versions (100%)

*"Devil in the details," Nature, vol. 470, 305-306 (2011).*

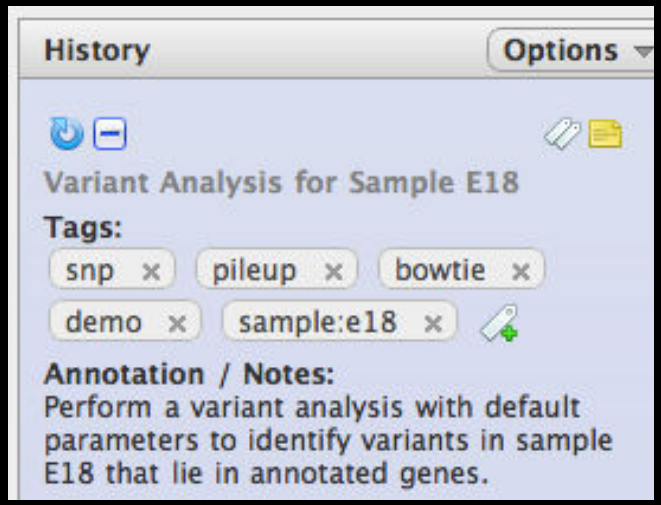
Metadata = Reproducibility

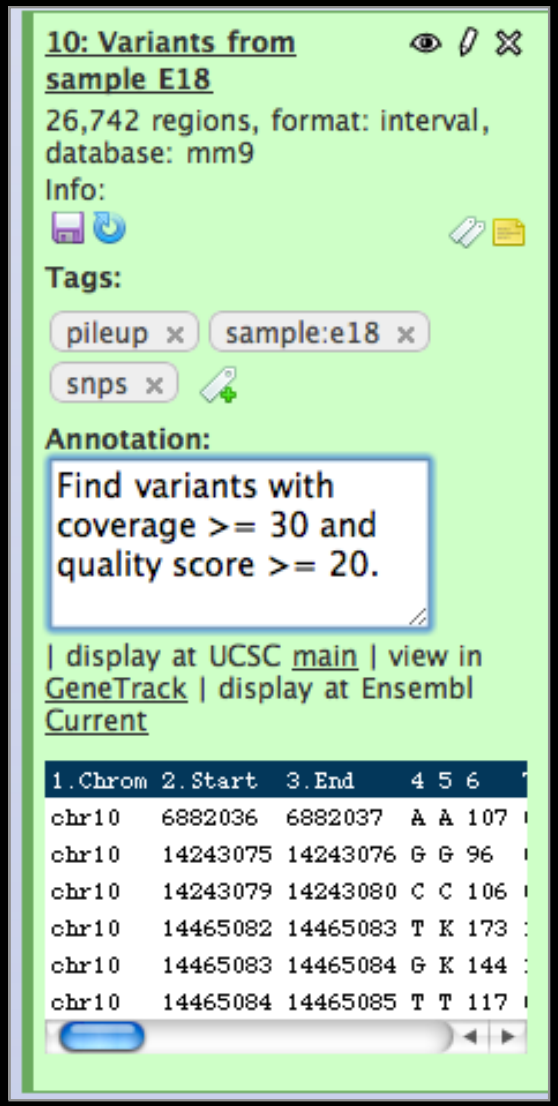
# Automatic Metadata



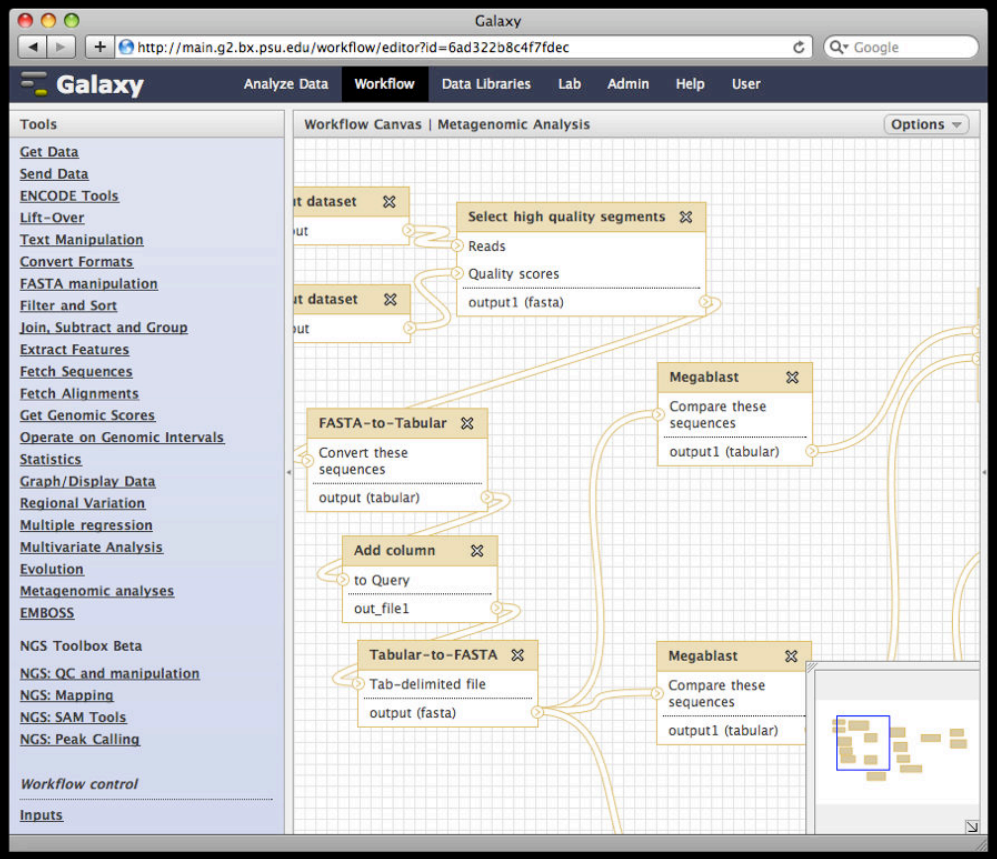



# User Metadata



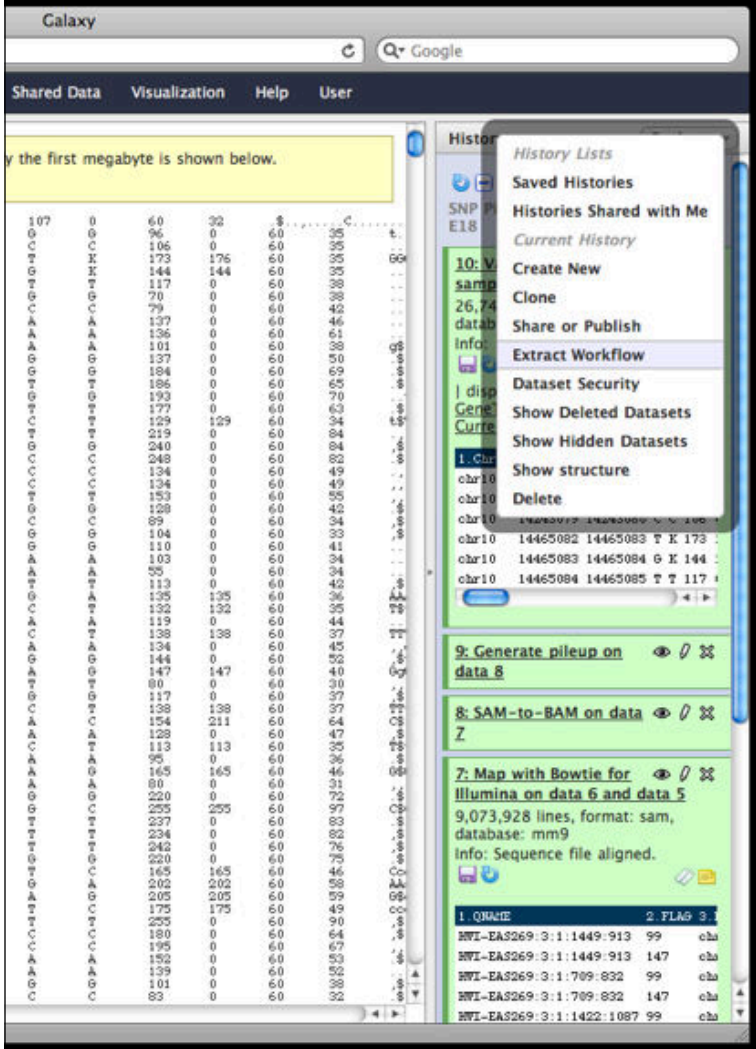
# Galaxy Workflow System

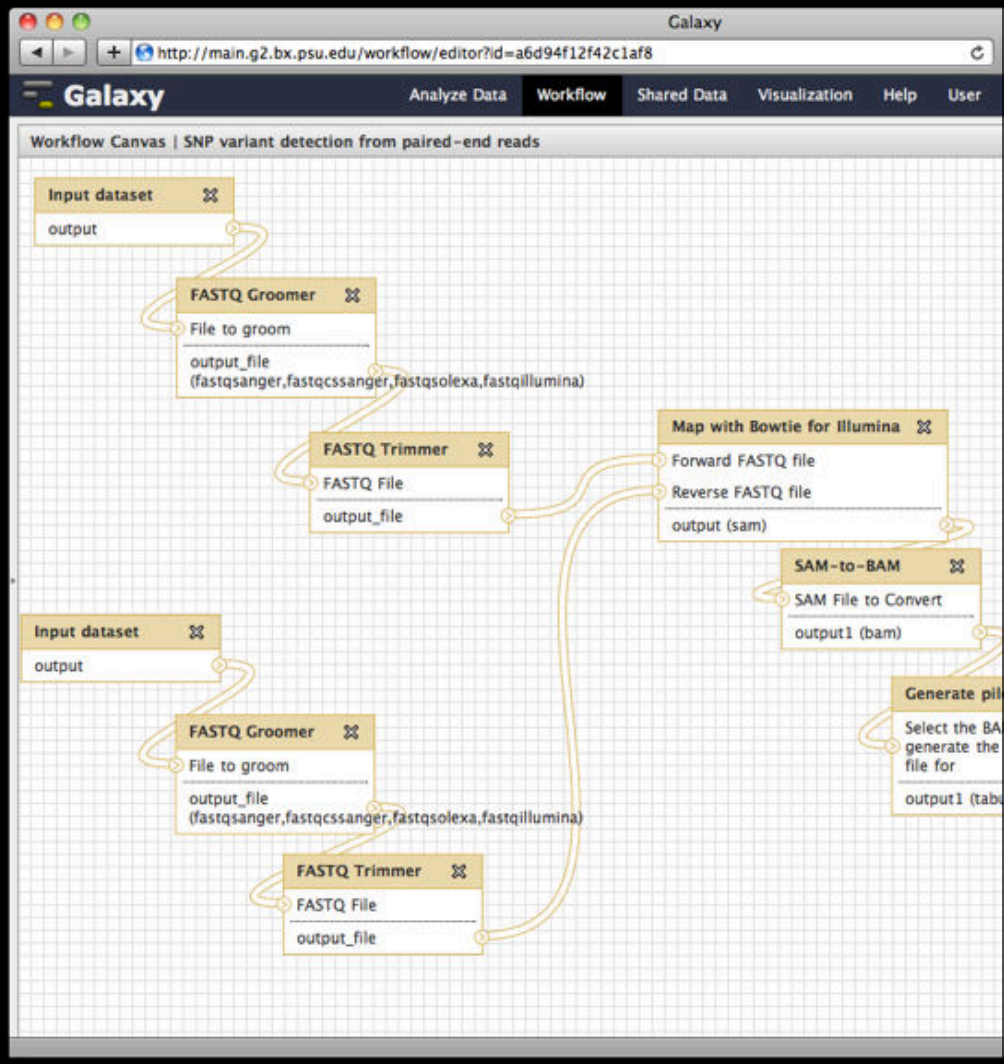


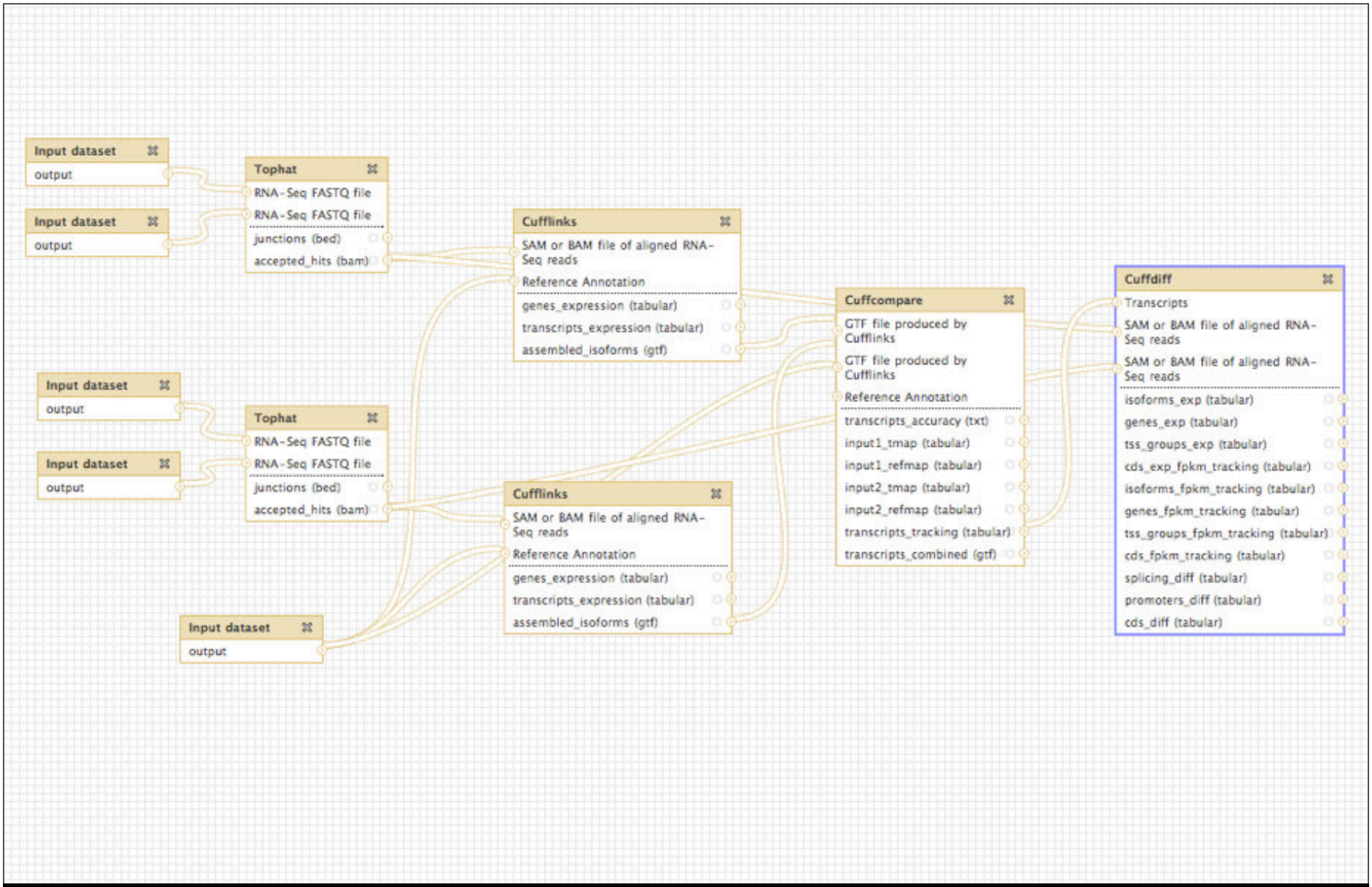

Workflows can be constructed from scratch *or* extracted from existing analysis histories

Facilitate reuse and provide precise reproducibility of a complex analysis

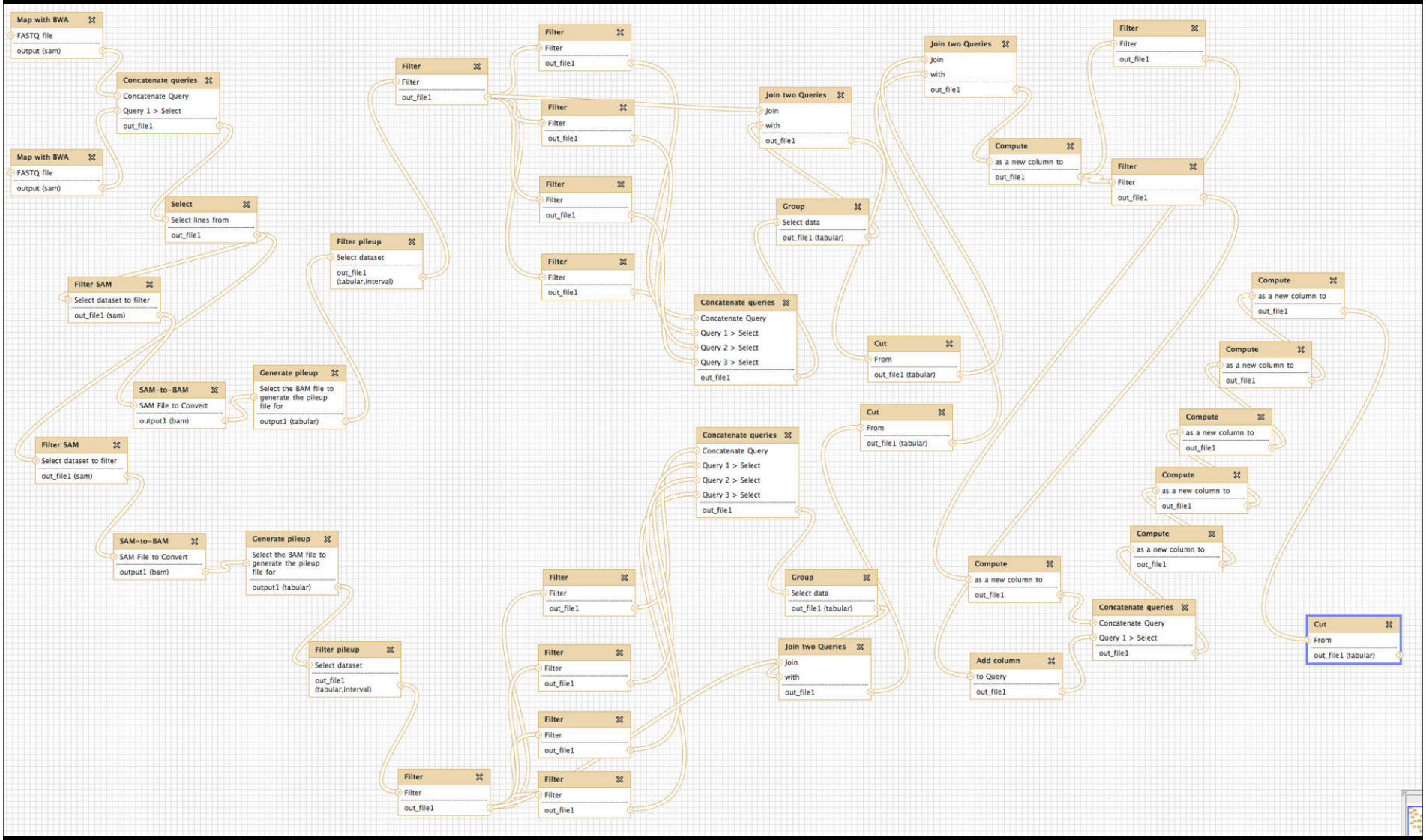
# Galaxy Workflows






**Example:** Workflow for differential expression analysis of RNA-seq using Tophat/Cufflinks tools



**Example**: Diagnosing low-frequency heterosplasmic sites in two tissues from the same individual

Transparency ~ Sharing, Publishing, Reusing

# *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 History | Variant Analysis for Sample E18

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

Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Help | User

Published Histories | jgoecks | Variant Analysis for Sample E18

## Galaxy History ' Variant Analysis for Sample E18'

Import history

Annotation: Perform a pileup analysis with default parameters to identify variants in sample E18.

| Dataset | Annotation |
|---|---|
| 1: E18 PE.1 Reads | Forward reads from sample E18. |
| 2: E18 PE.2 Reads | Reverse reads from sample E18. |
| 3: E18 PE.1 Reads Groomed | Groom reads to convert quality scores from Solexa 1.0 to Solexa 1.3 |
| 4: E18 PE.2 Reads Groomed | Groom reads to convert quality scores from Solexa 1.0 to Solexa 1.3 |
| 5: E18 PE.1 Reads Groomed, Trimmed | Trim reads from 3' end to remove low-quality nts. |
| 6: E18 PE.2 Reads Groomed, Trimmed | Trim reads from 3' to remove low-quality nts. |
| 7: Map with Bowtie for Illumina on data 6 and data 5 | Map paired-end reads with default parameters. |
| 8: SAM-to-BAM on data 7 | Need to convert Bowtie SAM to BAM so that pileup analysis can be performed. |
| 9: Generate pileup on data 8 | Pileup analysis with default parameters |
| 10: Filter pileup to get Variants from sample E18 | Find variants with coverage >= 30. |
| 13: Filter to get Variants from sample E18 where consensus base different than ref. base | Filter pileup to find variants where the consensus base is different than the reference base. |
| 14: UCSC mm9 RefSeq Genes | UCSC mm9 RefSeq genes. |
| 15: Intersect to get Variants from sample E18, consensus different, in RefSeq Genes | Variants with consensus different that occur in RefSeq genes. |

### About this History

**Author**

jgoecks

**Related Histories**

All published histories
Published histories by jgoecks

**Rating**

Community
(1 rating, 4.0 average)

Yours

**Tags**

Community:
snp | pileup | bowtie | demo | sample

Yours:
snp × | pileup × | bowtie × | demo × | sample:e18 ×

Galaxy | Published Histories

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

Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Help | User

## Published Histories

search | Advanced Search

| Name | Annotation | Owner | Community Rating ↑ | Community Tags | Last Updated |
|---|---|---|---|---|---|
| Galaxy vs MEGAN | Comparison of Galaxy vs. MEGAN pipeline. | aun1 | ★★★★★ | metagenomics, megan, galaxy | Mar 19, 2010 |
| metagenomic analysis | | aun1 | ★★★★★ | metagenomics, galaxy | Mar 19, 2010 |
| SM_1186088 | Datasets correspond to our paper published in Science by Peleg et al. entitled : Altered histone acetylation is associated with age-dependent memory impairment. Experiment layout: This history contains 4 datasets in the form of BED files of uniquely mapped reads produced after chip-seq for histone modifications H4K12ac and H3K9ac in mouse hippocampus of 3 months (young) and 16 months (old) mice after fear conditioning. For detailed information please refer to supplementary materials and methods of the respective work by peleg et al. | fischerlab | ★★★★★ | | Apr 19, 2010 |
| Variant Analysis for Sample E18 | Perform a pileup analysis with default parameters to identify variants in sample E18. | jgoecks | ★★★★☆ | snp, pileup, bowtie, demo, sample | 2 minutes ago |
| get longest exon | | henri | ☆☆☆☆☆ | chr22, longest, marc, exon, human, workshop | Sep 02, 2010 |
| FASTA to Tabular Test | | jj | ☆☆☆☆☆ | | Aug 26, 2010 |
| EKLF | | yzc109 | ☆☆☆☆☆ | | Aug 24, 2010 |

Open "http://main.g2.bx.psu.edu/history/list_published?sort=rating&f-tags=All" in a new tab

# Galaxy Pages

A web-based, interactive medium for presenting all aspects of an analysis: data, methods, visualization, and results

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

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

Hiroki Goto[1], Benjamin Dickins[2], Enis Afgan[3,5], Ian M. Paul[4], James Taylor[3,5], Kateryna D. Makova[1], 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):

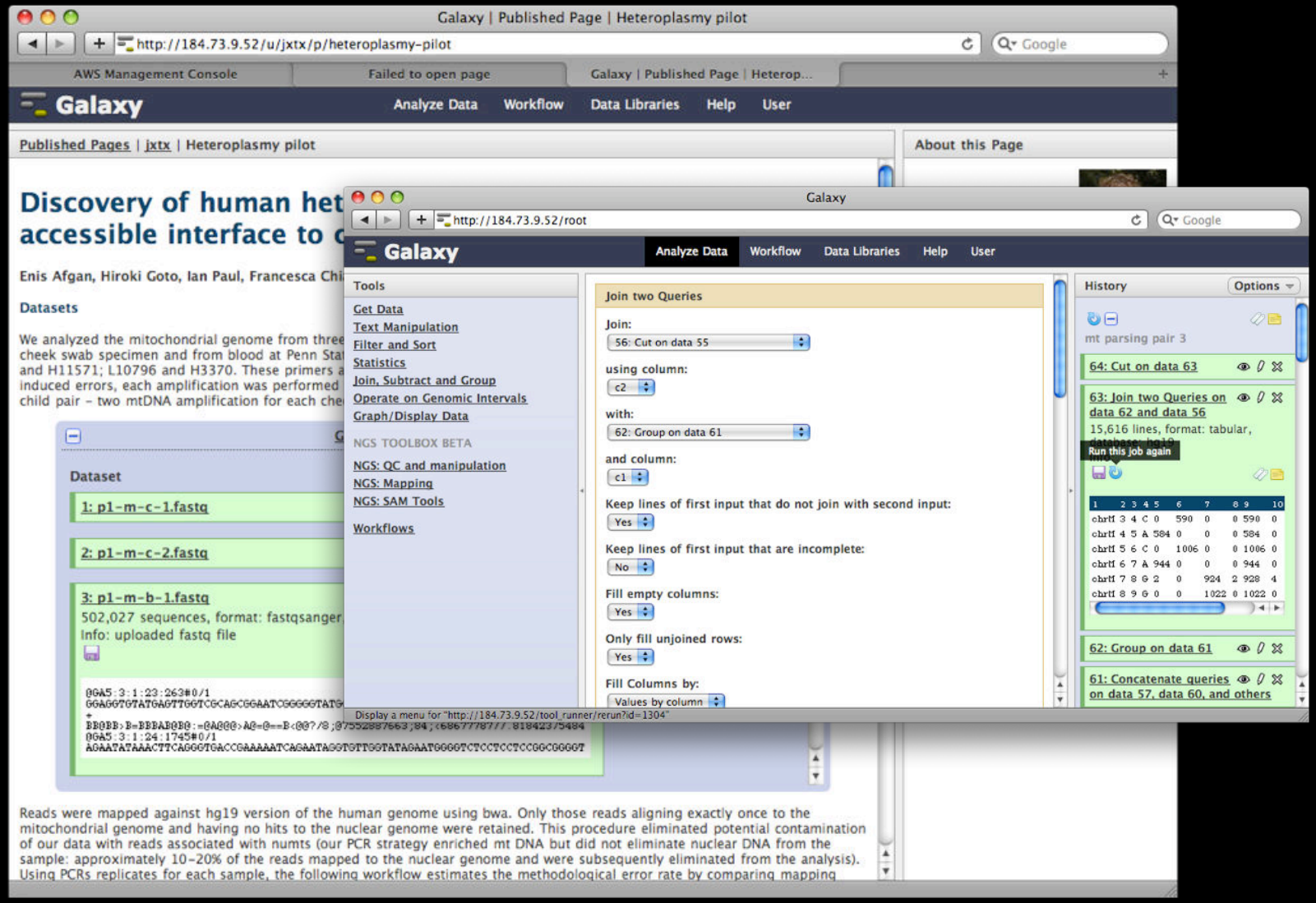
Galaxy History | F4-F7-F11 final report

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




Histories can be imported and the exact parameters inspected

Galaxy | Published Page | Heteroplasmy pilot

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

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

Galaxy — Analyze Data | Workflow | Data Libraries | Help | User

Published Pages | jxtx | 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 where 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**

**Filter**
Output dataset 'out_file1' from step 14

**With following condition**
c1=='chrM' and c10 >= 200

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

**Step 17: Join**

**Join**
Output dataset 'out_file1' from step 15

**with**
Output dataset 'out_file1' from step 16

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

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**

jxtx

**Related Pages**

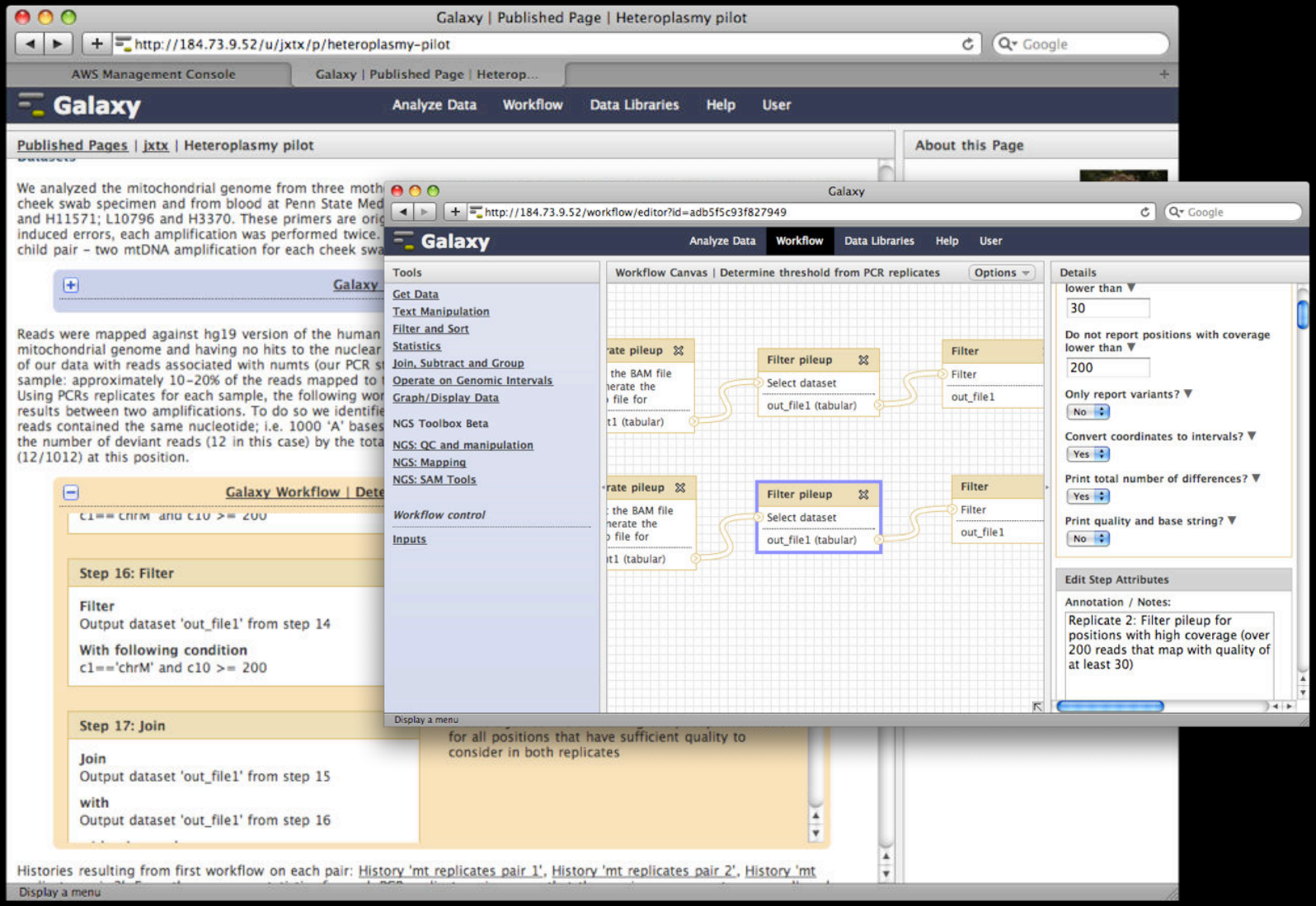
All published pages
Published pages by jxtx

**Tags**

Community:
cloud | heteroplasmy | ngs

Yours:
heteroplasmy × | cloud × | ngs ×

Workflows and other entities can also be embedded




And imported for inspection, verification, and reuse

# The Power of Galaxy Publishing

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 + annotations
- anyone can import your work and immediately reproduce or build on it

Windshield splatter analysis with the Galaxy metagenomic pipeline — Genome Research





# Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond[1,2,6,9], Samir Wadhawan[3,6,7], Francesca Chiaromonte[4], Guruprasad Ananda[1,3], Wen-Yu Chung[1,3,8], James Taylor[1,5,9], Anton Nekrutenko[1,3,9] and The Galaxy Team[1]

[+] 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.



# Three Ways to Use Galaxy

1. Public Website (http://usegalaxy.org)

2. Download and run locally

3. Run on the cloud

# Galaxy main site (http://usegalaxy.org)

Public Website, anybody can use

~500 new users per month, ~100 TB of user data, ~130,000 analysis jobs per month, every month is our busiest month ever...

Will continue to be maintained and enhanced, but with limits and quotas

Centralized solution cannot scale to meet data analysis demands

# Download and Run Locally

No configuration needed but everything can be configured

- tools
- computing cluster integration
- proxy server and authentication

Prominent local installations at:

- Cold Spring Harbor Lab
- Dept. of Energy's Joint Genome Institute
- Harvard School of Public Health
- U of Texas System
- Netherlands Bioinformatics Center
- Oxford College

Requires computing resources, technical expertise, and maintenance

# Galaxy on the Cloud

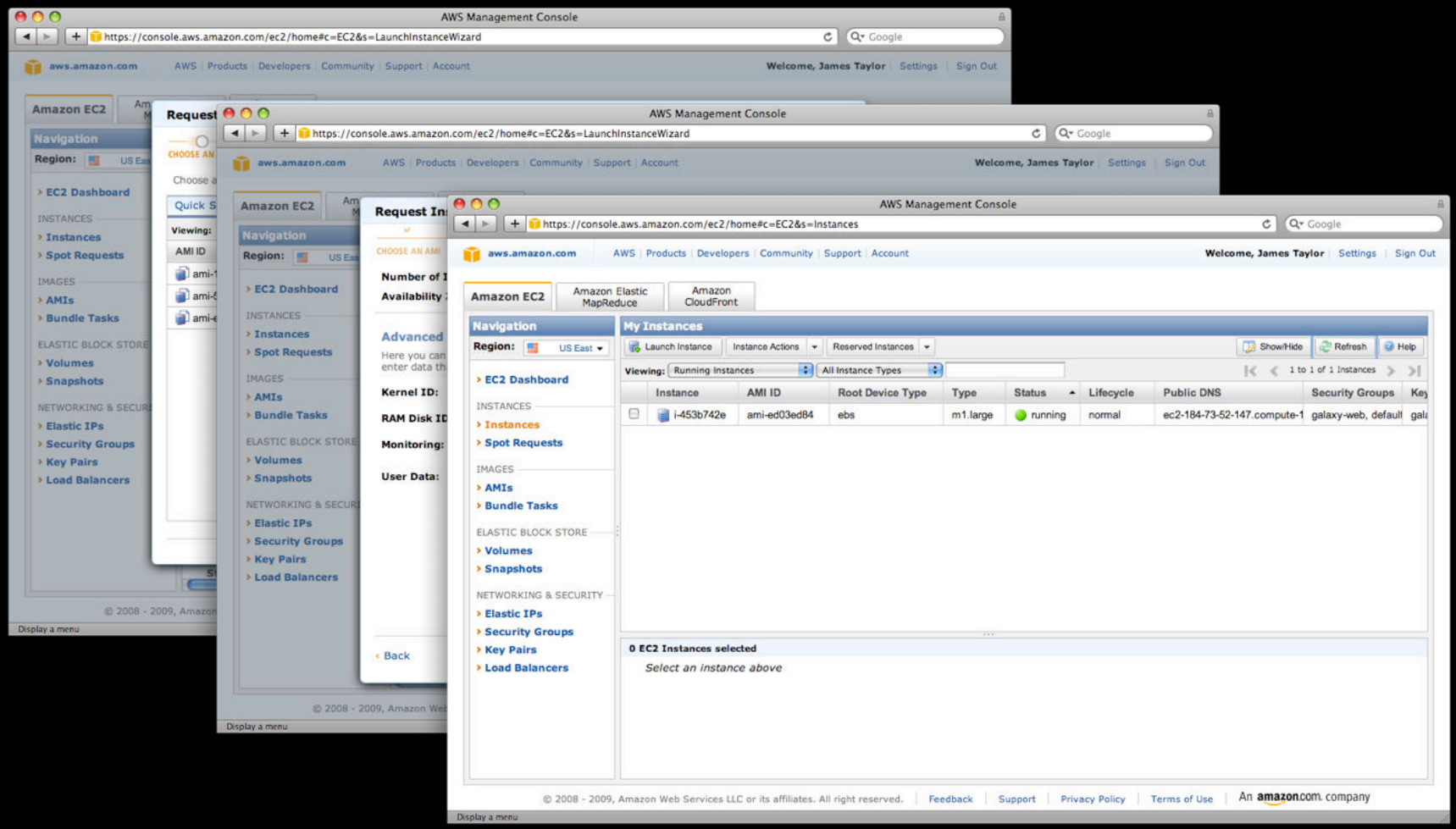
For extended or particular resource needs

- customization necessary
- oscillating data volume

For when informatics expertise or infrastructure is limited

*Work done by Enis Afgan*

# Using Amazon EC2: Startup in 3 steps



ec2-50-17-119-106.compute-1.amazonaws.com/root

Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Help | User

Tools | Options

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Wavelet Analysis
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Metagenomic analyses
- FASTA manipulation

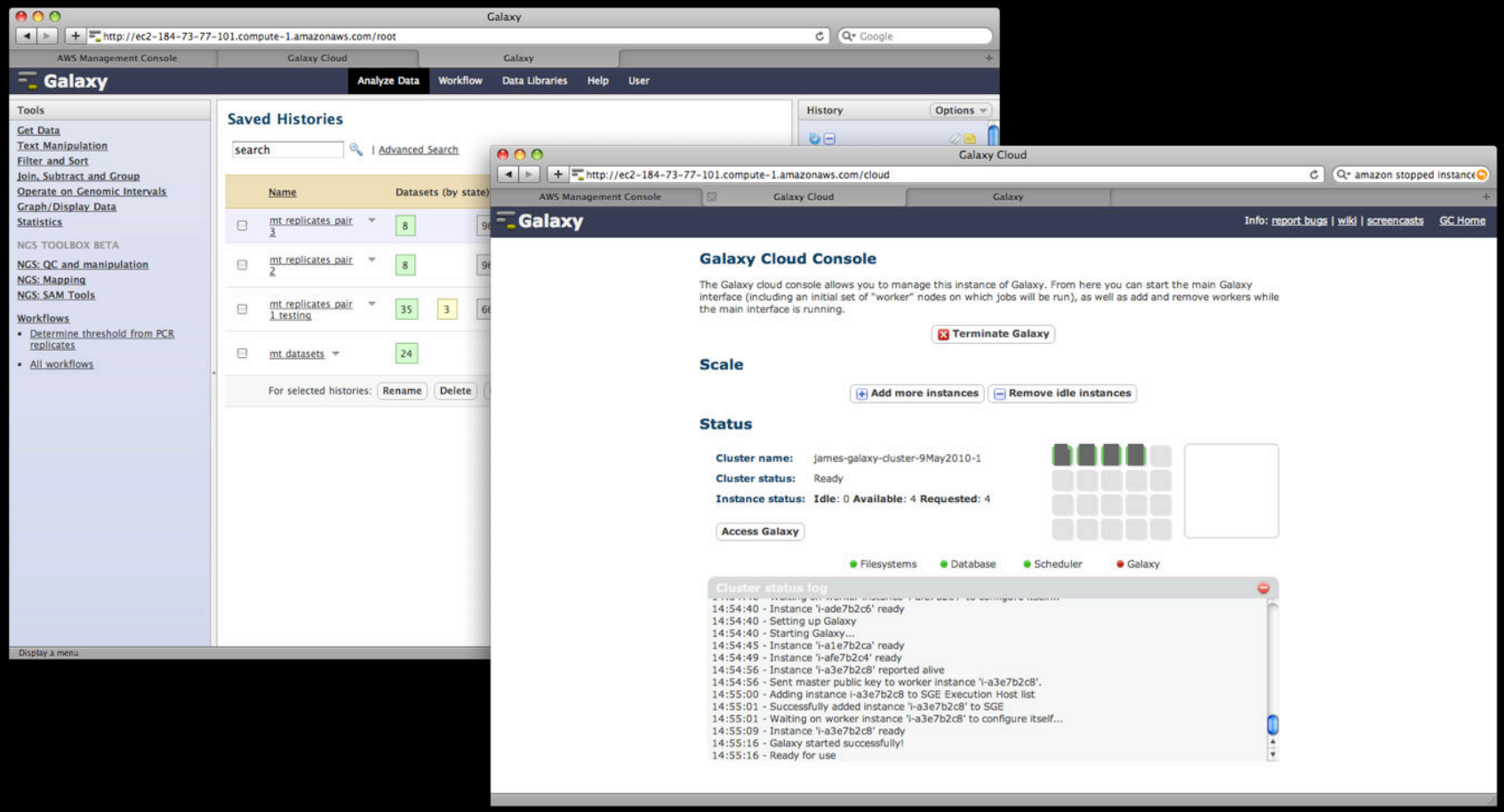
NGS TOOLBOX BETA

- NGS: QC and manipulation
- NGS: Assembly
- NGS: Mapping
- NGS: Indel Analysis
- NGS: Expression Analysis
- NGS: SAM Tools
- NGS: Peak Calling
- Human Genome Variation
- EMBOSS

**Welcome to Galaxy** on the Cloud

History | Options

Your history is empty. Click 'Get Data' on the left pane to start




Like any other Galaxy instance plus:

- additional compute nodes acquired and released *automatically* as needed
- can share or publish an instance

# Challenges Going Forward

Promoting community involvement

- tools, assays, analyses growing too fast for us alone
- facilitate community contributions and usage of contributions

Scaling to many, many Galaxies

- moving objects between Galaxies while ensuring reproducibility
- enabling users to find useful "stuff"

Novel application areas

- genomics ideal application area -- what next?

EMORY

Enis Afgan

Dannon Baker

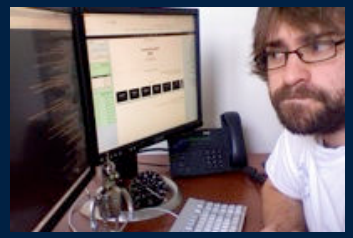
Dave Clements

Jeremy Goecks

James Taylor

PENNSTATE

Dan Blankenberg



Nate Coraor

Jennifer Jackson

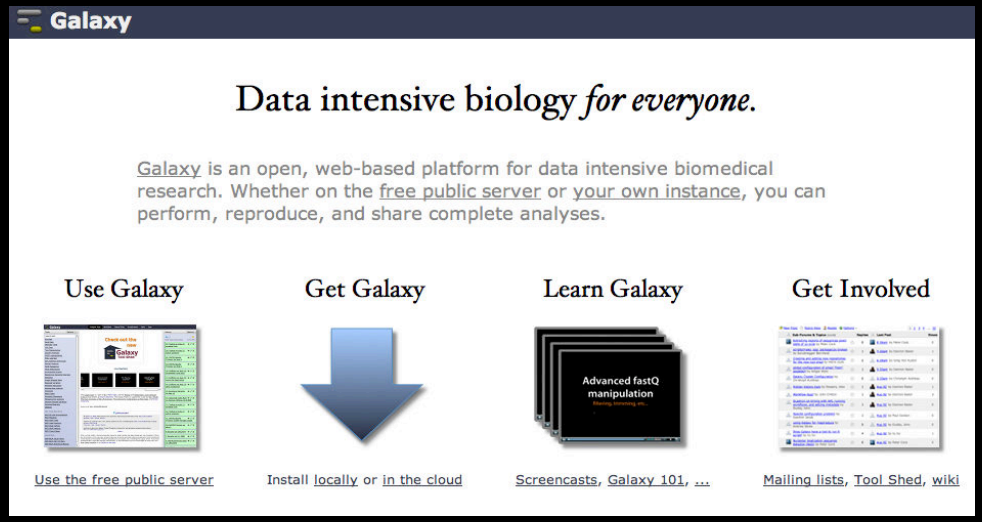
Greg von Kuster

Anton Nekrutenko

Supported by the **NHGRI** (HG005542, HG004909, HG005133), **NSF** (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health

# Thanks! Questions?

http://galaxyproject.org




Galaxy publications: http://galaxyproject.org/wiki/Citing

Galaxy is hiring! http://galaxyproject.org/wiki/GalaxyisHiring

jeremy.goecks@emory.edu