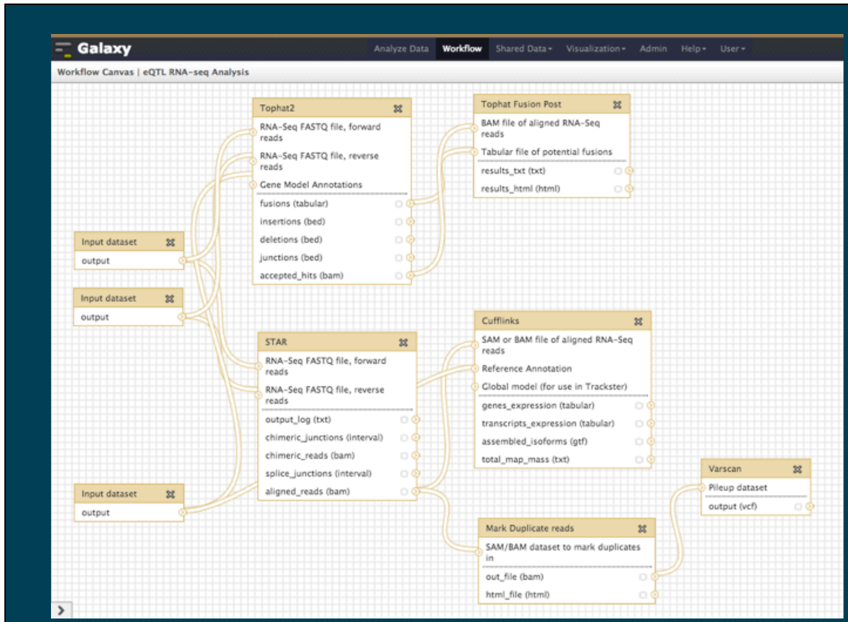
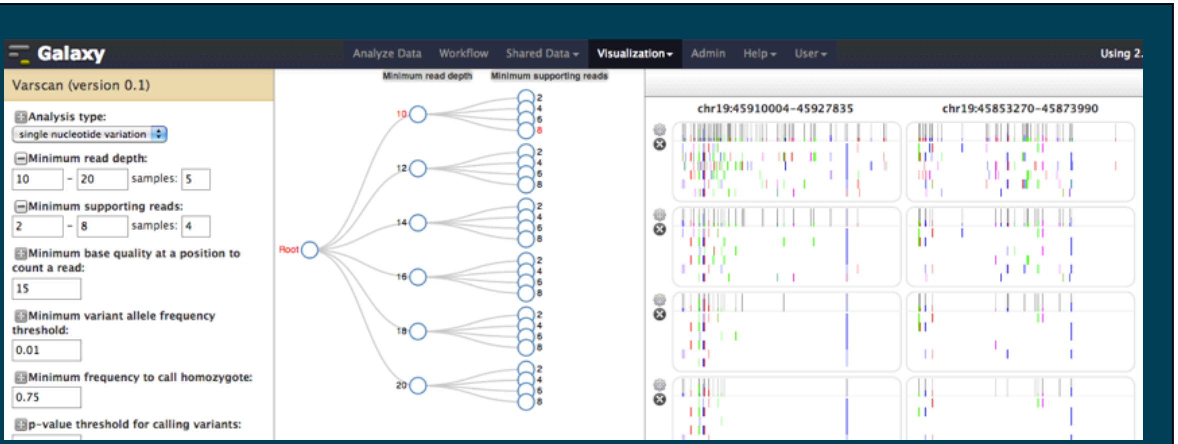




Varscan (version 0.1)

Analysis type: single nucleotide variation

Minimum read depth: 10 - 20 samples: 5

Minimum supporting reads: 2 - 8 samples: 4

Minimum base quality at a position to count a read: 15

Minimum variant allele frequency threshold: 0.01

Minimum frequency to call homozygote: 0.75

p-value threshold for calling variants:

Galaxy as a Platform for High-throughput Genomics

Jeremy Goecks

EMORY
UNIVERSITY

Topics

Galaxy

Analyzing Cancer Genomes & Transcriptomes

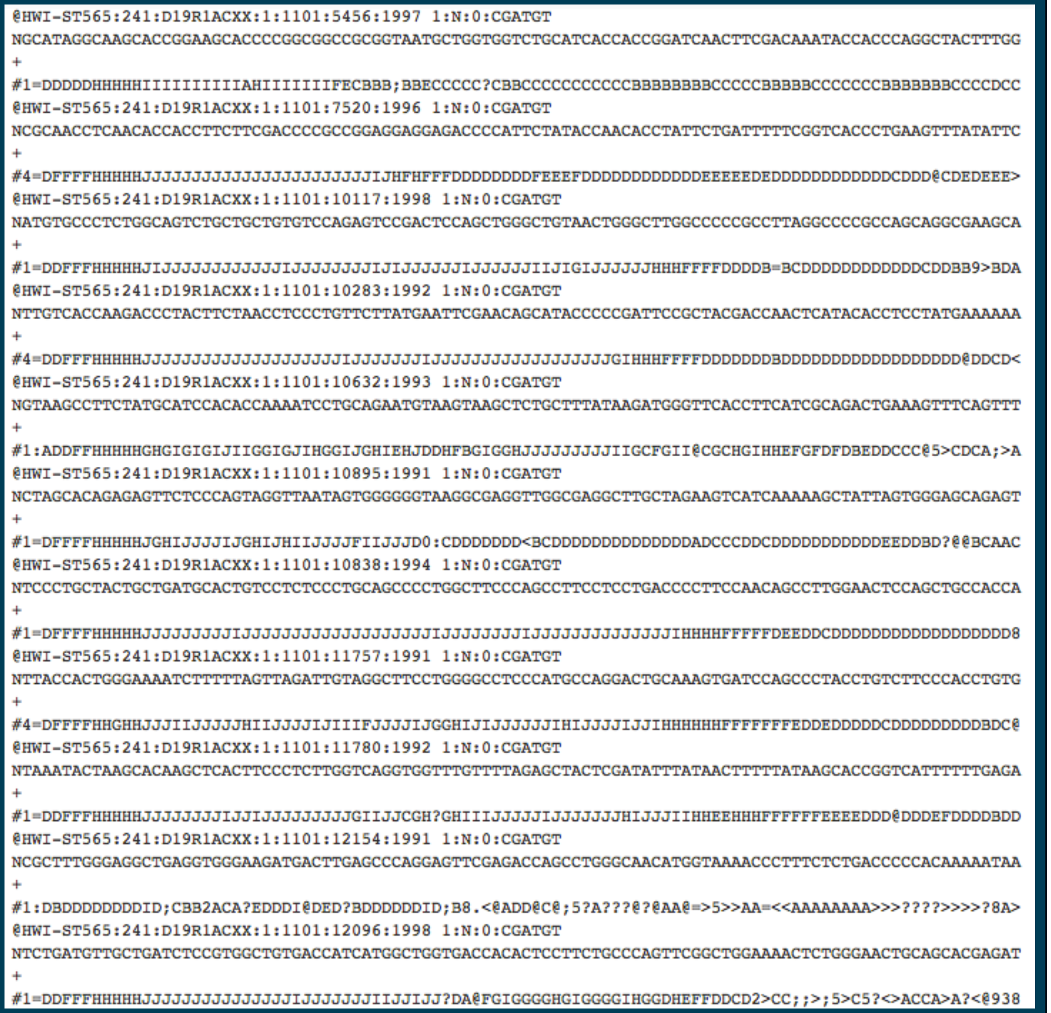
Web-based Visual Analysis

Genomic Analyses are Difficult

Investigators unfamiliar with computation

Creating and reproducing workflows (pipelines) hindered by complexity: systems, scripts, tools, parameters

Collaboration and reuse difficult because current approaches do not support computational artifacts well



```
@HWI-ST565:241:D19R1ACXX:1:1101:5456:1997 1:N:0:CGATGT
NGCATAGGCAAGCACCGGAAGCACCCCGGCGGCCGCGGTAATGCTGGTGGTCTGCATCACCACCGGATCAACTTCGACAAATACCACCCAGGCTACTTTGG
+
#1=DDDDDHHHHHIIIIIIIIIIAHIIIIIIIFECBBB;BBECCCCC?CBBCCCCCCCCCCCBBBBBBBBCCCCCBBBBBCCCCCCCBBBBBBBCCCCDCC
@HWI-ST565:241:D19R1ACXX:1:1101:7520:1996 1:N:0:CGATGT
NCGCAACCTCAACACCACCTTCTTCGACCCCGCCGGAGGAGGAGACCCCATTCTATACCAACACCTATTCTGATTTTTCGGTCACCCTGAAGTTTATATTC
+
#4=DFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJIJHFHFFFDDDDDDDDFEEEFDDDDDDDDDDDDEEEEEDEDDDDDDDDDDDDCDDD@CDEDEEE>
@HWI-ST565:241:D19R1ACXX:1:1101:10117:1998 1:N:0:CGATGT
NATGTGCCCTCTGGCAGTCTGCTGCTGTGTCCAGAGTCCGACTCCAGCTGGGCTGTAACTGGGCTTGGCCCCCGCCTTAGGCCCCGCCAGCAGGCGAAGCA
+
#1=DDFFFHHHHHJIJJJJJJJJJJJJIJJJJJJJJIJIJJJJJJIJJJJJJIIJIGIJJJJJJHHHFFFFDDDDB=BCDDDDDDDDDDDDCDDBB9>BDA
@HWI-ST565:241:D19R1ACXX:1:1101:10283:1992 1:N:0:CGATGT
NTTGTCACCAAGACCCTACTTCTAACCTCCCTGTTCTTATGAATTCGAACAGCATACCCCCGATTCCGCTACGACCAACTCATACACCTCCTATGAAAAAA
+
#4=DDFFFHHHHHJJJJJJJJJJJJJJJJJJJJIJJJJJJJIJJJJJJJJJJJJJJJJJGIHHHFFFFDDDDDDDBDDDDDDDDDDDDDDDDDDD@DDCD<
@HWI-ST565:241:D19R1ACXX:1:1101:10632:1993 1:N:0:CGATGT
NGTAAGCCTTCTATGCATCCACACCAAAATCCTGCAGAATGTAAGTAAGCTCTGCTTTATAAGATGGGTTCACCTTCATCGCAGACTGAAAGTTTCAGTTT
+
#1:ADDFFHHHHHGHGIGIGIJIIGGIGJIHGGIJGHIEHJDDHFBGIGGHJJJJJJJJJIIGCFGII@CGCHGIHHEFGFDFDBEDDCCC@5>CDCA;>A
@HWI-ST565:241:D19R1ACXX:1:1101:10895:1991 1:N:0:CGATGT
NCTAGCACAGAGAGTTCTCCCAGTAGGTTAATAGTGGGGGGTAAGGCGAGGTTGGCGAGGCTTGCTAGAAGTCATCAAAAAGCTATTAGTGGGAGCAGAGT
+
#1=DFFFFHHHHHJGHIJJJJIJGHIJHIIJJJJFIIJJJD0:CDDDDDDD<BCDDDDDDDDDDDDDDDADCCCDDCDDDDDDDDDDDDEEDDBD?@@BCAAC
@HWI-ST565:241:D19R1ACXX:1:1101:10838:1994 1:N:0:CGATGT
NTCCCTGCTACTGCTGATGCACTGTCCTCTCCCTGCAGCCCCTGGCTTCCCAGCCTTCCTCCTGACCCCTTCCAACAGCCTTGGAACTCCAGCTGCCACCA
+
#1=DFFFFHHHHHJJJJJJJJIJJJJJJJJJJJJJJJJJJJIJJJJJJJJIJJJJJJJJJJJJJIHHHHFFFFFDEEDDCDDDDDDDDDDDDDDDDDDD8
@HWI-ST565:241:D19R1ACXX:1:1101:11757:1991 1:N:0:CGATGT
NTTACCACTGGGAAAATCTTTTTAGTTAGATTGTAGGCTTCCTGGGGCCTCCCATGCCAGGACTGCAAAGTGATCCAGCCCTACCTGTCTTCCCACCTGTG
+
#4=DFFFFHHGHHJJJIIJJJJJHIIJJJJIJIIIFJJJJIJGGHIJIJJJJJJIHIJJJJIJJIHHHHHHFFFFFFFEDDEDDDDCDDDDDDDDDBDC@
@HWI-ST565:241:D19R1ACXX:1:1101:11780:1992 1:N:0:CGATGT
NTAAATACTAAGCACAAGCTCACTTCCCTCTTGGTCAGGTGGTTTGTTTTAGAGCTACTCGATATTTATAACTTTTTATAAGCACCGGTCATTTTTTGAGA
+
#1=DDFFFHHHHHJJJJJJJIJJIJJJJJJJJJGIIJJCGH?GHIIIJJJJJIJJJJJJJHIJJJIIHHEEHHHFFFFFFEEEEDDD@DDDEFDDDDBDD
@HWI-ST565:241:D19R1ACXX:1:1101:12154:1991 1:N:0:CGATGT
NCGCTTTGGGAGGCTGAGGTGGGAAGATGACTTGAGCCCAGGAGTTCGAGACCAGCCTGGGCAACATGGTAAAACCCTTTCTCTGACCCCCACAAAAATAA
+
#1:DBDDDDDDDDID;CBB2ACA?EDDDI@DED?BDDDDDDID;B8.<@ADD@C@;5?A???@?@AA@=>5>>AA=<<AAAAAAAA>>>????>>>>?8A>
@HWI-ST565:241:D19R1ACXX:1:1101:12096:1998 1:N:0:CGATGT
NTCTGATGTTGCTGATCTCCGTGGCTGTGACCATCATGGCTGGTGACCACACTCCTTCTGCCCAGTTCGGCTGGAAAACTCTGGGAACTGCAGCACGAGAT
+
#1=DDFFFHHHHHJJJJJJJJJJJJJJJIJJJJJJJIIJJIJJ?DA@FGIGGGGHGIGGGGIHGGDHEFFDDCD2>CC;;>;5>C5?<>ACCA>A?<@938
```

Galaxy Project: Fundamental Questions

When genomics (or any other biomedical science) becomes dependent on computational methods, how to:

- make tools and workflows **accessible** to scientists?
- ensure that analyses are **reproducible**?
- enable transparent **communication and reuse** of analyses?

Vision

Galaxy is an **open, Web-based platform** for accessible, reproducible, and collaborative computational genomics

Goecks et al. (2010) *Genome Biology*

Galaxy Demo

What is Galaxy?

Platform for high-throughput genomics

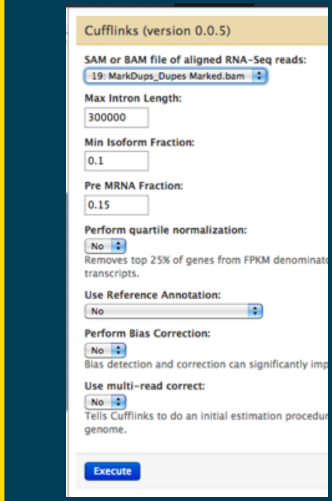
1. get and integrate public, private data
2. analyze data and create workflows
3. visualization, sharing, publication

Customizable open-source software on various HPC resources

- public website — http://usegalaxy.org
- local instance
- on the cloud

Galaxy platform

- run tools, workflows on HPC resources
- create workflows, visualizations, pages
- share *everything*



Cufflinks (version 0.0.5)

SAM or BAM file of aligned RNA-Seq reads: 19: MarkDups_Dupes Marked.bam

Max Intron Length: 300000

Min Isoform Fraction: 0.1

Pre MRNA Fraction: 0.15

Perform quartile normalization: No

Removes top 25% of genes from FPKM denominat transcripts.

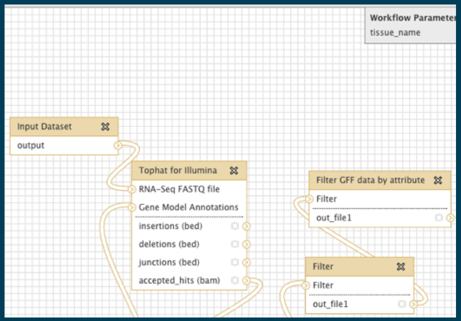
Use Reference Annotation: No

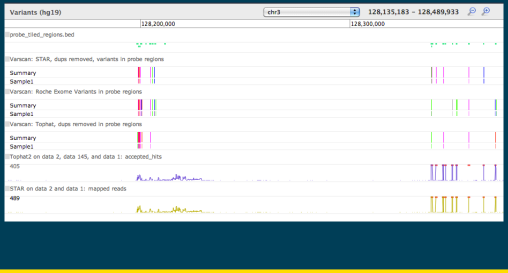
Perform Bias Correction: No

Bias detection and correction can significantly imp

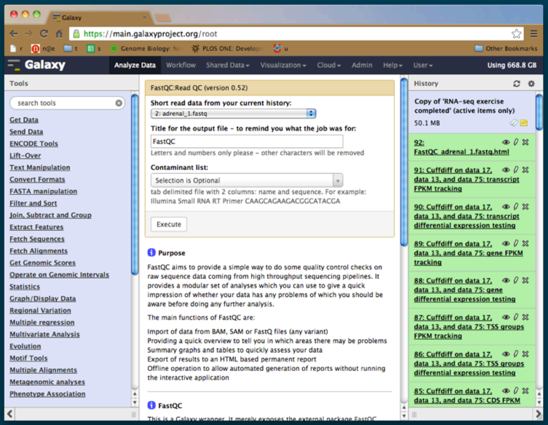
Use multi-read correct: No

Tells Cufflinks to do an initial estimation procedur genome.





Variants (hg19)
chr3
128,135,183 – 128,489,933
128,200,000
128,300,000
probe_tiled_regions.bed
Varscan: STAR, dups removed, variants in probe regions
Summary
Sample1
Varscan: Roche Exome Variants in probe regions
Summary
Sample1
Varscan: Tophat, dups removed in probe regions
Summary
Sample1
Tophat2 on data 2, data 145, and data 1: accepted_hits
405
STAR on data 2 and data 1: mapped reads
489



FastQC:Read QC (version 0.52)

Short read data from your current history: 2: adrenal_1.fastq

Title for the output file – to remind you what the job was for: FastQC

Letters and numbers only please – other characters will be removed

Contaminant list: Selection is Optional

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Purpose

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

Import of data from BAM, SAM or FastQ files (any variant)
Providing a quick overview to tell you in which areas there may be problems
Summary graphs and tables to quickly assess your data
Export of results to an HTML based permanent report
Offline operation to allow automated generation of reports without running the interactive application

FastQC

This is a Galaxy wrapper. It merely exposes the external package FastQC

Copy of 'RNA-seq exercise completed' (active items only)
50.1 MB

92: FastQC adrenal 1.fastq.html

91: Cuffdiff on data 17, data 13, and data 75: transcript FPKM tracking

90: Cuffdiff on data 17, data 13, and data 75: transcript differential expression testing

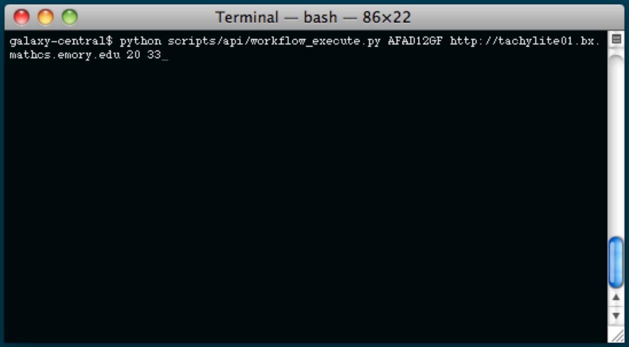
89: Cuffdiff on data 17, data 13, and data 75: gene FPKM tracking

88: Cuffdiff on data 17, data 13, and data 75: gene differential expression testing

87: Cuffdiff on data 17, data 13, and data 75: TSS groups FPKM tracking

86: Cuffdiff on data 17, data 13, and data 75: TSS groups differential expression testing

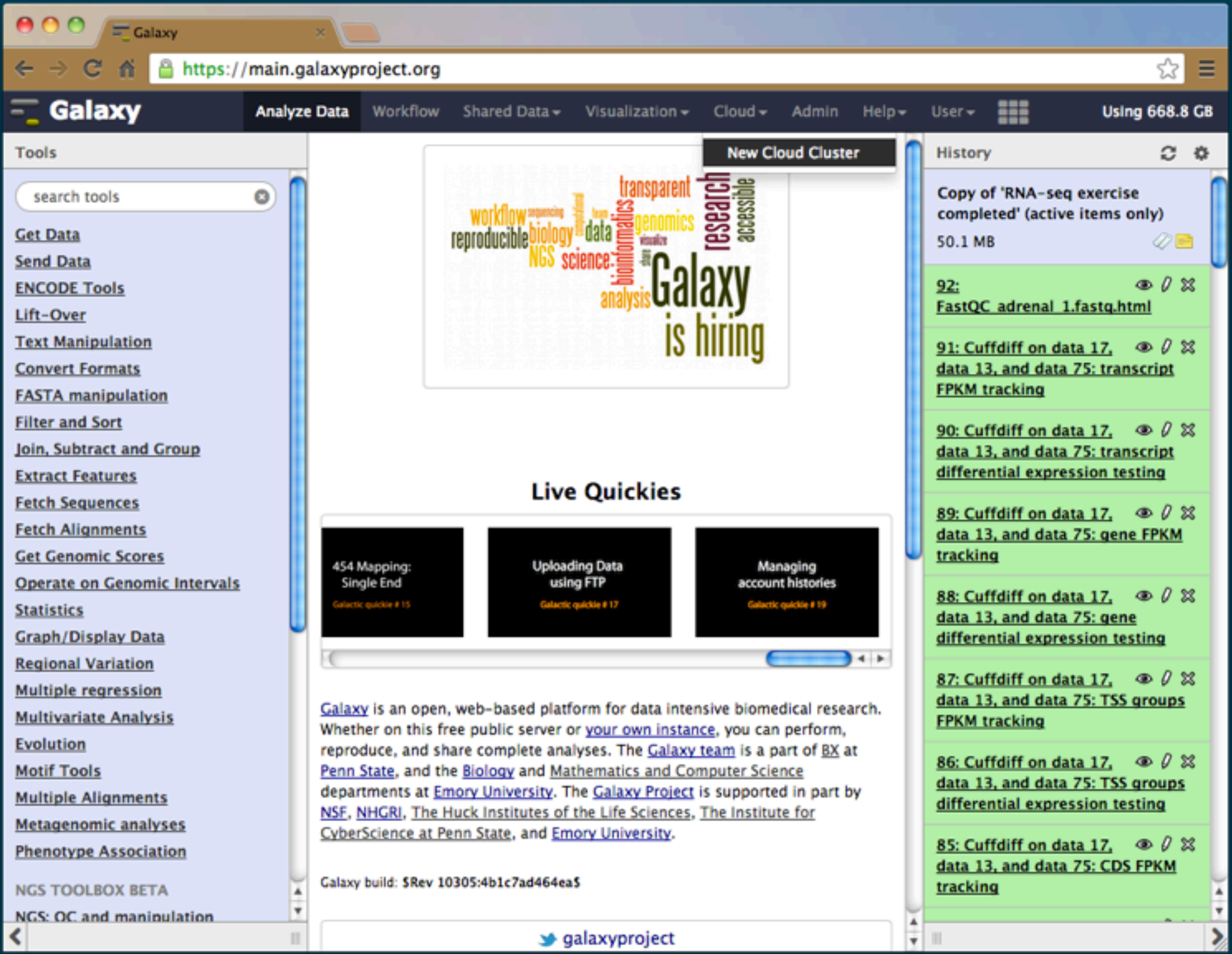
85: Cuffdiff on data 17, data 13, and data 75: CDS FPKM



Terminal — bash — 86×22

```
galaxy-central$ python scripts/api/workflow_execute.py AFAD12GF http://tachylite01.bx.
mathcs.emory.edu 20 33_
```

Cloud Launch



Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

Galaxy build: $Rev 10305:4b1c7ad464ea$

Copy of 'RNA-seq exercise completed' (active items only)

50.1 MB

92: FastQC adrenal 1.fastq.html

91: Cuffdiff on data 17, data 13, and data 75: transcript FPKM tracking

90: Cuffdiff on data 17, data 13, and data 75: transcript differential expression testing

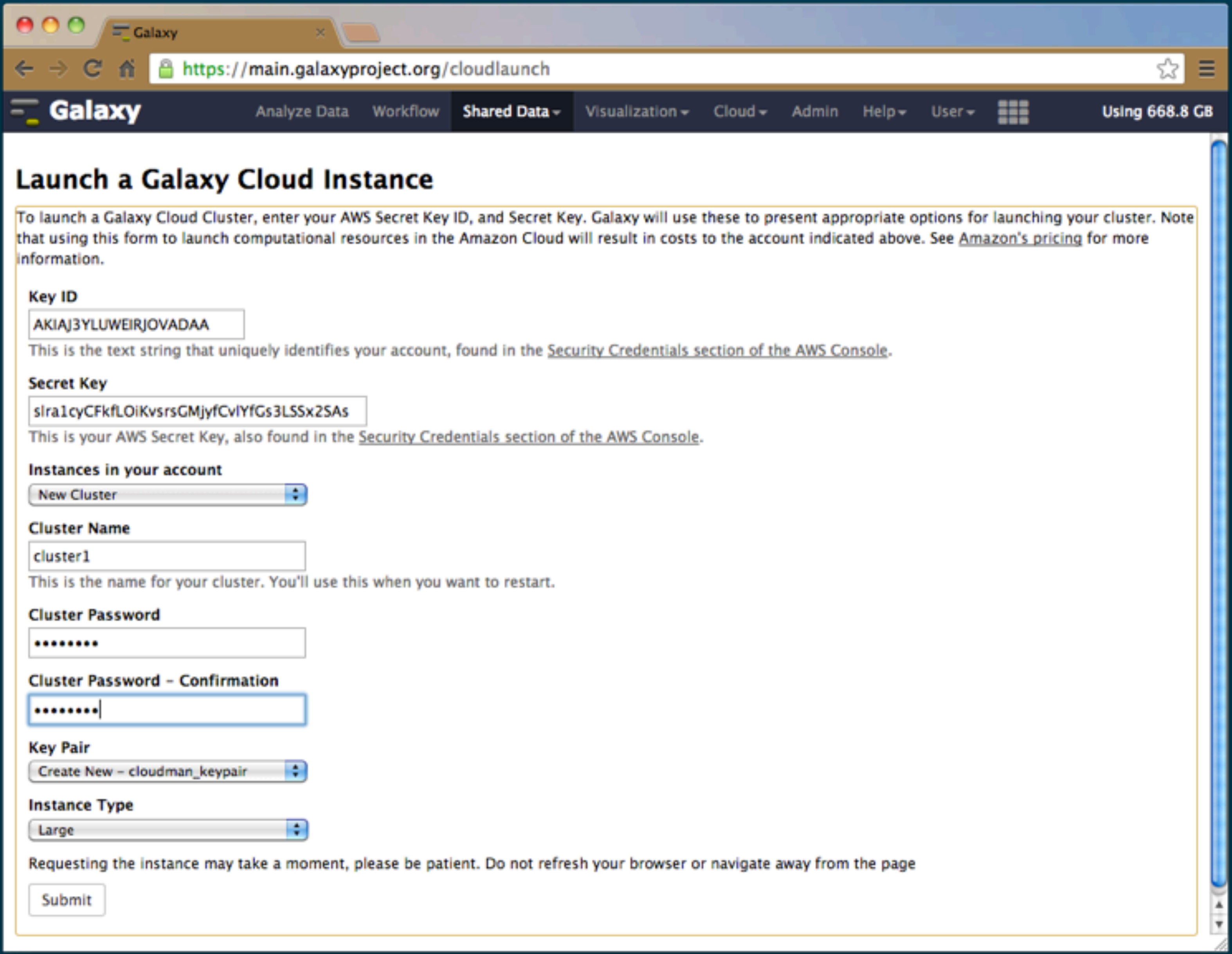
89: Cuffdiff on data 17, data 13, and data 75: gene FPKM tracking

88: Cuffdiff on data 17, data 13, and data 75: gene differential expression testing

87: Cuffdiff on data 17, data 13, and data 75: TSS groups FPKM tracking

86: Cuffdiff on data 17, data 13, and data 75: TSS groups differential expression testing

85: Cuffdiff on data 17, data 13, and data 75: CDS FPKM tracking



Launch a Galaxy Cloud Instance

To launch a Galaxy Cloud Cluster, enter your AWS Secret Key ID, and Secret Key. Galaxy will use these to present appropriate options for launching your cluster. Note that using this form to launch computational resources in the Amazon Cloud will result in costs to the account indicated above. See Amazon's pricing for more information.

Key ID

AKIAJ3YLUWEIRJOVADAA

This is the text string that uniquely identifies your account, found in the Security Credentials section of the AWS Console.

Secret Key

slra1cyCFkfLOiKvsrsGMjyfCvlYfGs3LSSx2SAs

This is your AWS Secret Key, also found in the Security Credentials section of the AWS Console.

Instances in your account

New Cluster

Cluster Name

cluster1

This is the name for your cluster. You'll use this when you want to restart.

Cluster Password

••••••••

Cluster Password - Confirmation

••••••••

Key Pair

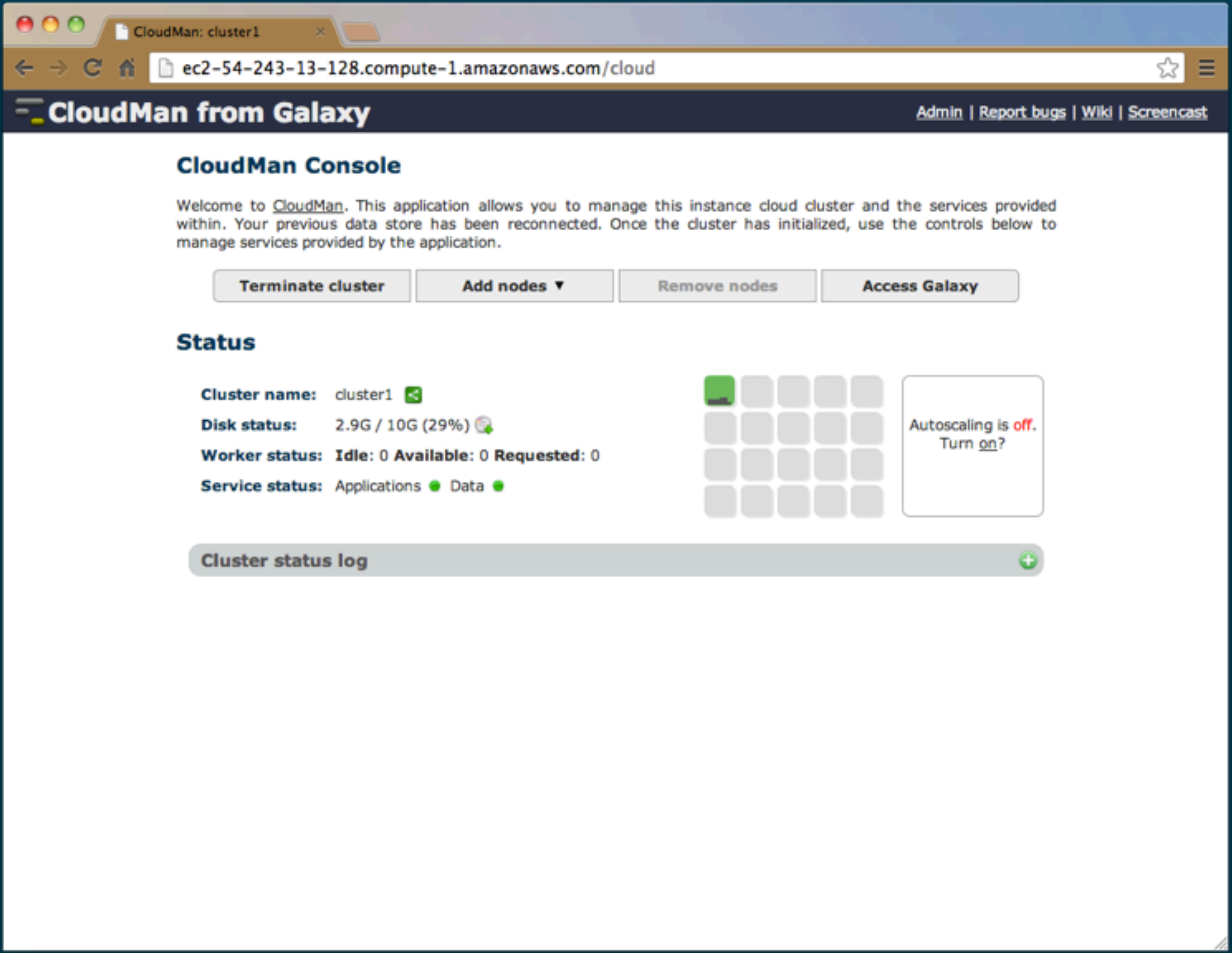
Create New - cloudman_keypair

Instance Type

Large

Requesting the instance may take a moment, please be patient. Do not refresh your browser or navigate away from the page

Submit


CloudMan: cluster1
ec2-54-243-13-128.compute-1.amazonaws.com/cloud
CloudMan from Galaxy
Admin | Report bugs | Wiki | Screencast

CloudMan Console

Welcome to CloudMan. This application allows you to manage this instance cloud cluster and the services provided within. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to manage services provided by the application.

Status

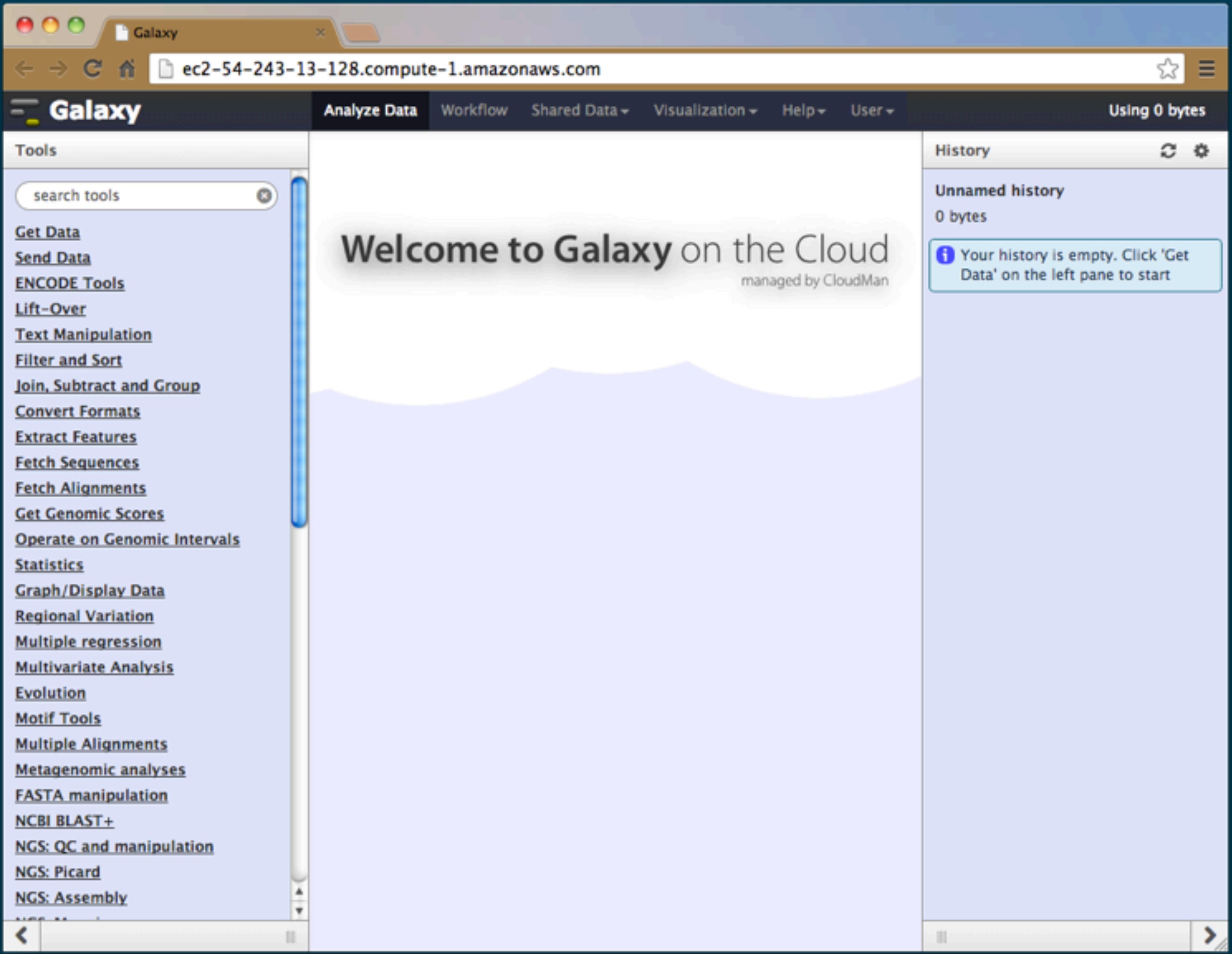
Cluster name: cluster1

Disk status: 2.9G / 10G (29%)

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

Service status: Applications ● Data ●

Autoscaling is off.
Turn on?



Welcome to Galaxy on the Cloud

managed by CloudMan

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

Cloud Features

Resource configuration

Autoscaling

Snapshotting

Galaxy is Very Popular

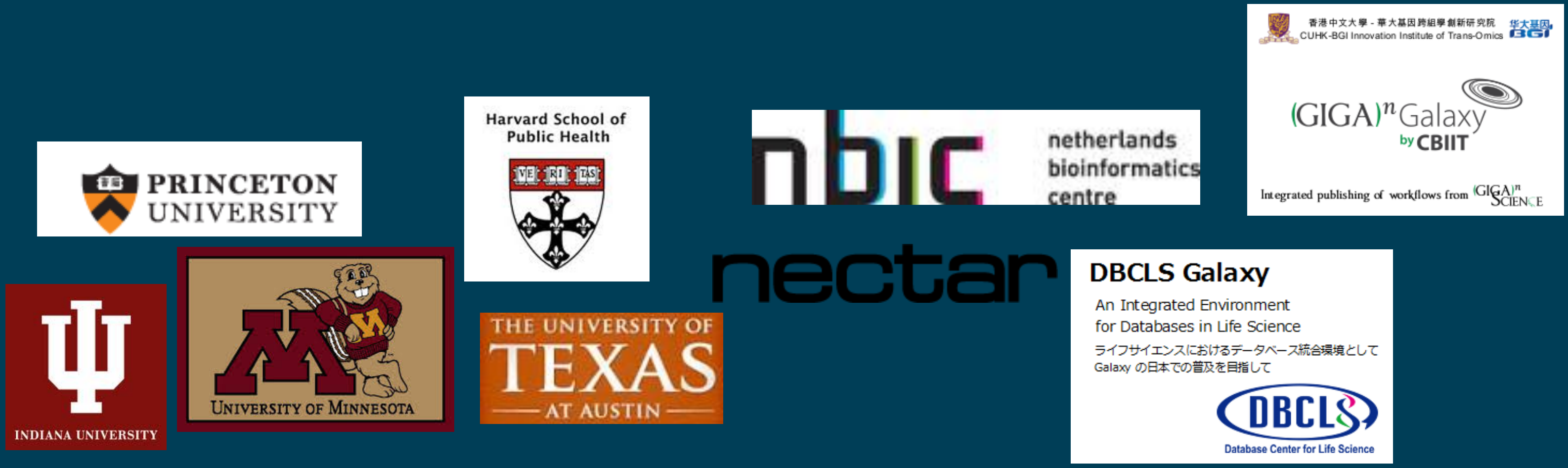
Public Website (http://usegalaxy.org), anybody can use:

- ~500 new users per month, ~200 TB of user data, ~130,000 analysis jobs per month

Used and cited in more than 1000 publications

Galaxy is Very Popular

Local installations all over the world


PRINCETON UNIVERSITY
Harvard School of Public Health
VE RI TAS
nbic netherlands bioinformatics centre
香港中文大學 - 華大基因跨組學創新研究院
CUHK-BGI Innovation Institute of Trans-Omics
华大基因 BGI
(GIGA)^n Galaxy by CBIIT
Integrated publishing of workflows from (GIGA)^n SCIENCE
INDIANA UNIVERSITY
UNIVERSITY OF MINNESOTA
THE UNIVERSITY OF TEXAS AT AUSTIN
nectar
DBCLS Galaxy
An Integrated Environment for Databases in Life Science
ライフサイエンスにおけるデータベース統合環境として Galaxy の日本での普及を目指して
DBCLS
Database Center for Life Science

http://bioteam.net/slipstream/galaxy-edition/

Topics

Galaxy

Analyzing Cancer Genomes & Transcriptomes

Web-based Visual Analysis

Preliminary Data

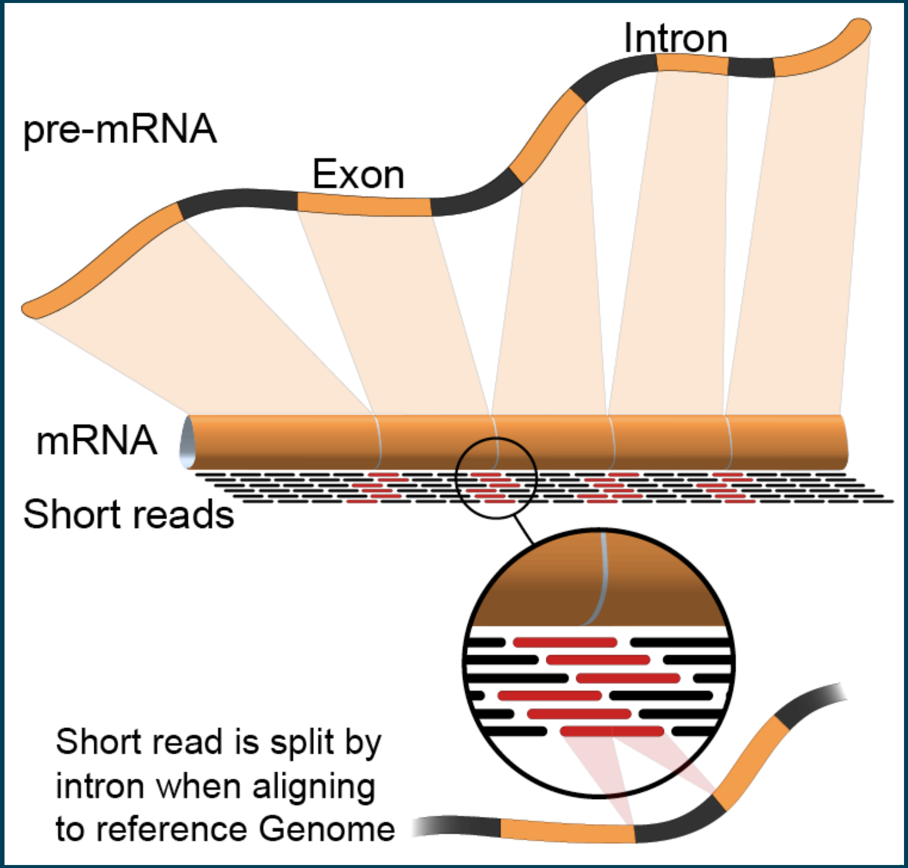
6 patients, whole transcriptome sequencing (RNA-seq) of primary tumor

- mixed populations!
- 3 +ERCC, 3 -ERCC (via IHC)

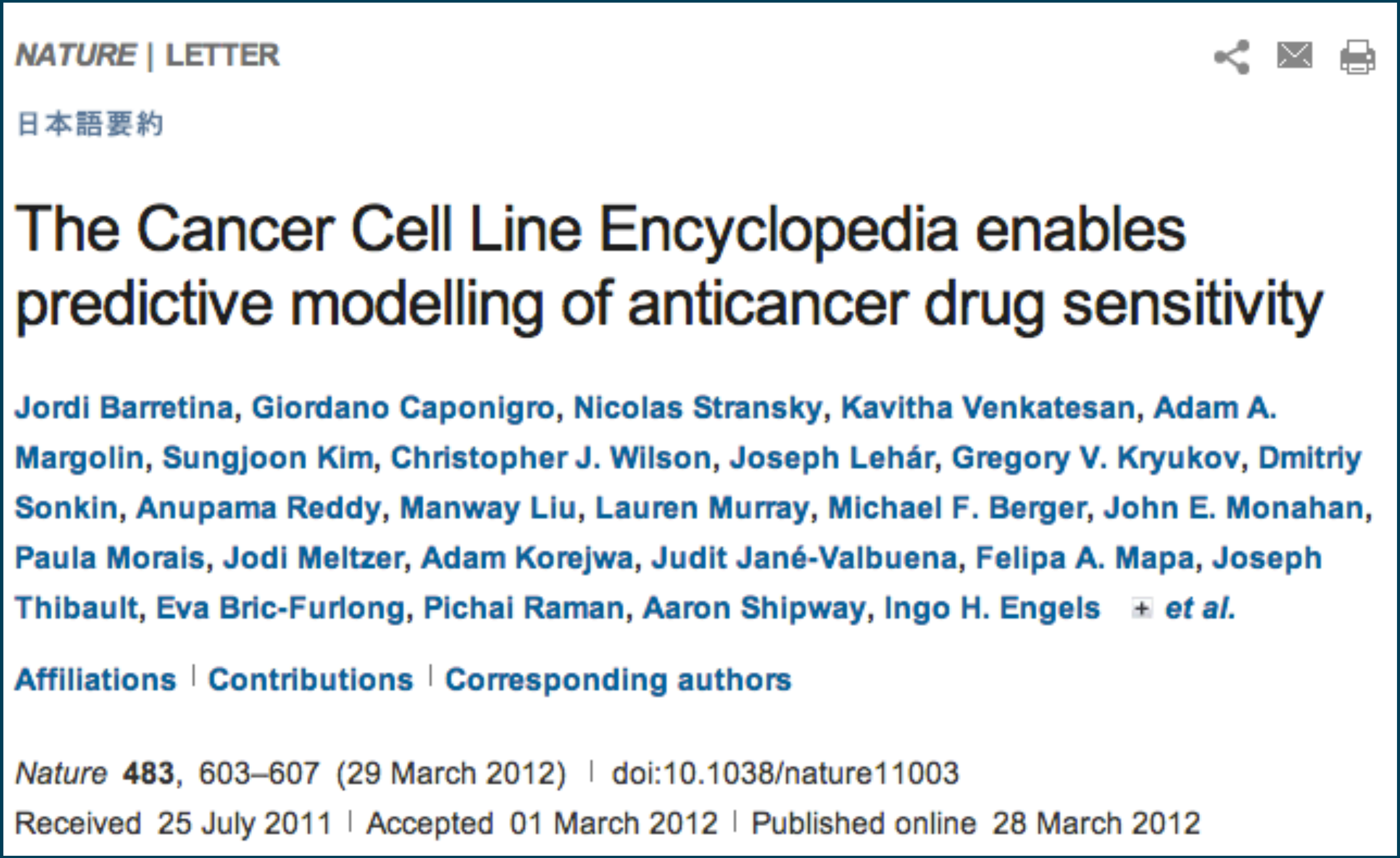
MiaPaCa2 cell line

- whole transcriptome
- targeted exome

Total sequencing data: ~70 GB


Intron
pre-mRNA
Exon
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

http://en.wikipedia.org/wiki/RNA-Seq



NATURE | LETTER

日本語要約

The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity

Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, Joseph Lehár, Gregory V. Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F. Berger, John E. Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jané-Valbuena, Felipa A. Mapa, Joseph Thibault, Eva Bric-Furlong, Pichai Raman, Aaron Shipway, Ingo H. Engels + *et al.*

Affiliations | Contributions | Corresponding authors

Nature **483**, 603–607 (29 March 2012) | doi:10.1038/nature11003

Received 25 July 2011 | Accepted 01 March 2012 | Published online 28 March 2012

Big Questions

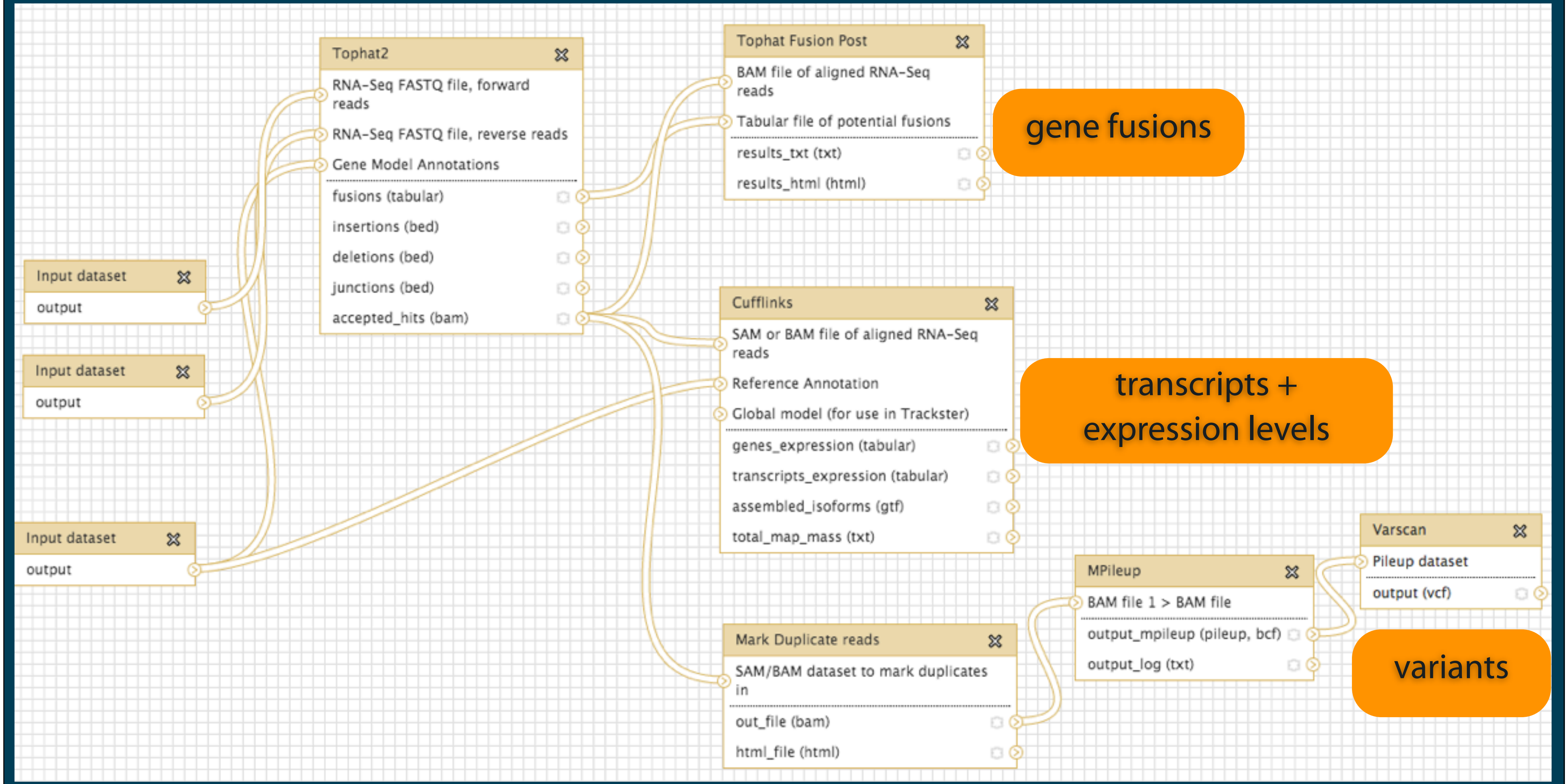
How closely does MiaPaCa2 match to primary pancreatic tumors?

How to match patient to “best” CCLE cell line(s)?

Using Galaxy for Analysis of Cancer Transcriptomes

- New tools
  - complement existing transcriptome analysis tools
- New workflows
  - workflows are understandable, extendable, sharable
- New visual analysis applications
  - visualize and call variants in a Web browser

Single Sample Transcriptome Analysis

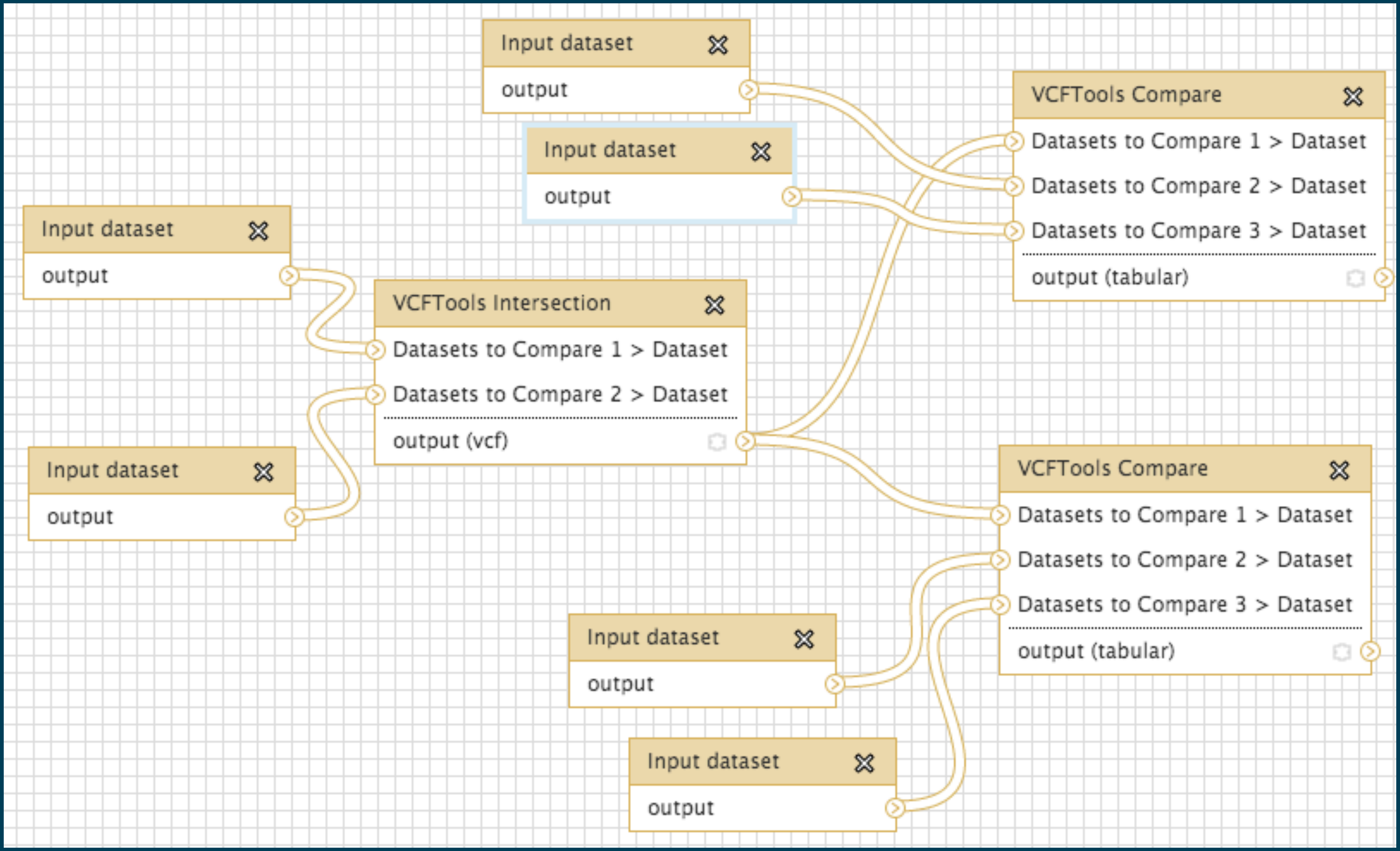


gene fusions

transcripts + expression levels

variants

Comparing Called Variants with Public Datasets


Input dataset
output
Input dataset
output
VCFTools Compare
Datasets to Compare 1 > Dataset
Datasets to Compare 2 > Dataset
Datasets to Compare 3 > Dataset
output (tabular)
Input dataset
output
VCFTools Intersection
Datasets to Compare 1 > Dataset
Datasets to Compare 2 > Dataset
output (vcf)
Input dataset
output
VCFTools Compare
Datasets to Compare 1 > Dataset
Datasets to Compare 2 > Dataset
Datasets to Compare 3 > Dataset
output (tabular)
Input dataset
output
Input dataset
output

Patient Mutations vs. CCLE Cancer Cell Line Encyclopedia

http://www.broadinstitute.org/ccle/home

	P1	P2	P3	P4	P5	P6	CL
OM MIA (4)	0	1	1	0	0	0	4
OM PC (11)	0	1	1	0	0	0	4
OM ALL (114)	0	2	1	1	1	1	4
HP MIA (84)	3	6	4	5	4	3	15
HP PC (1769)	16	23	19	11	23	8	39
HP ALL (64,669)	110	180	143	97	136	65	87

OM = OncoMap, HP = hybrid capture with probes

Patient Mutations vs.

CCLE Cancer Cell Line Encyclopedia

http://www.broadinstitute.org/ccle/home

	P1	P2	P3	P4	P5	P6	CL
OM MIA (4)	0	1	1	0	0	0	4
OM PC (11)	0	1	1	0	0	0	4
OM ALL (114)	0	2	1	1	1	1	4
HP MIA (84)	3	6	4	5	4	3	15
HP PC (1769)	16	23	19	11	23	8	39
HP ALL (64,669)	110	180	143	97	136	65	87

Cell line does not appear very similar to tumors

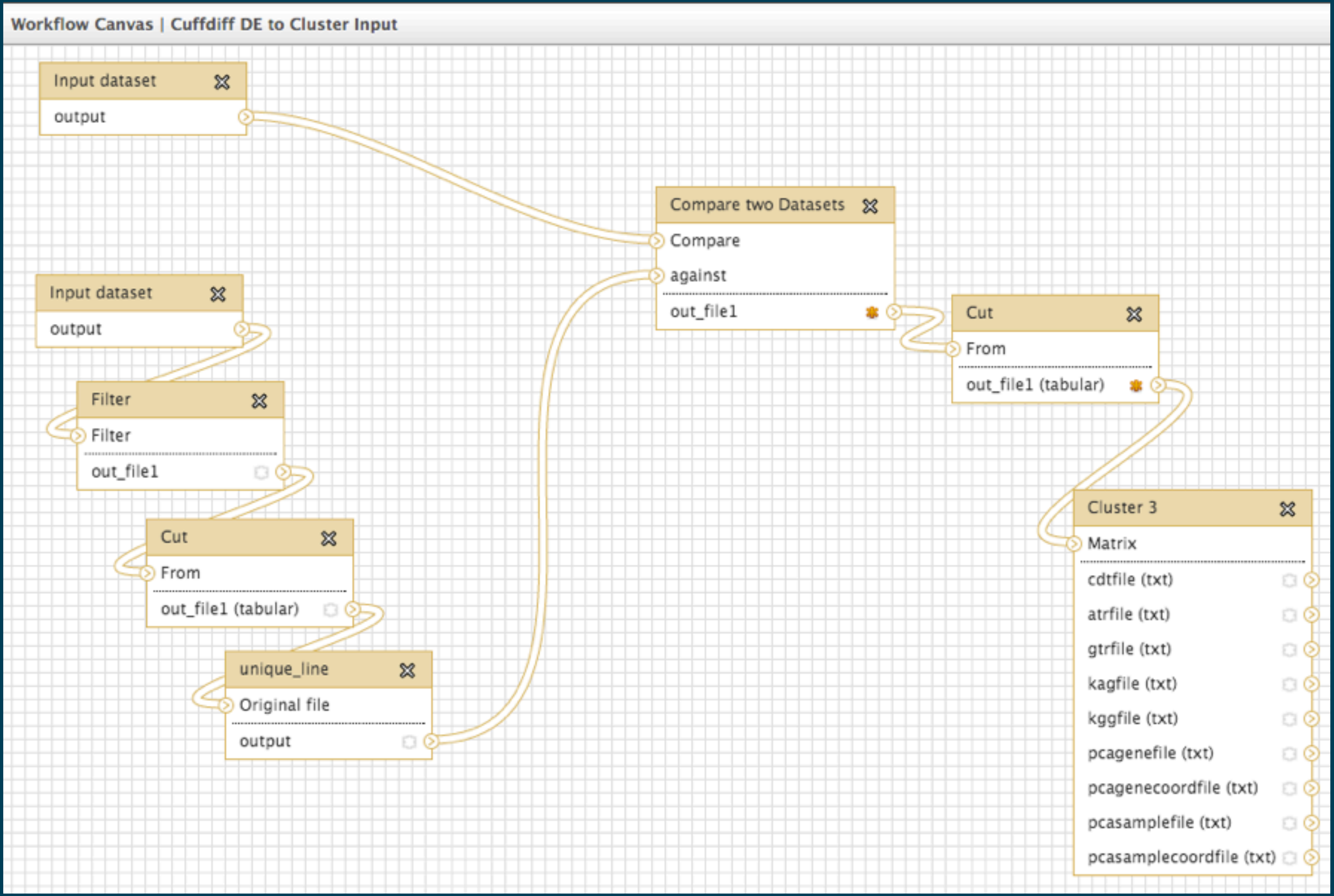
OM = OncoMap, HP = hybrid capture with probes

Patient Mutations to Predict Tumor Attributes

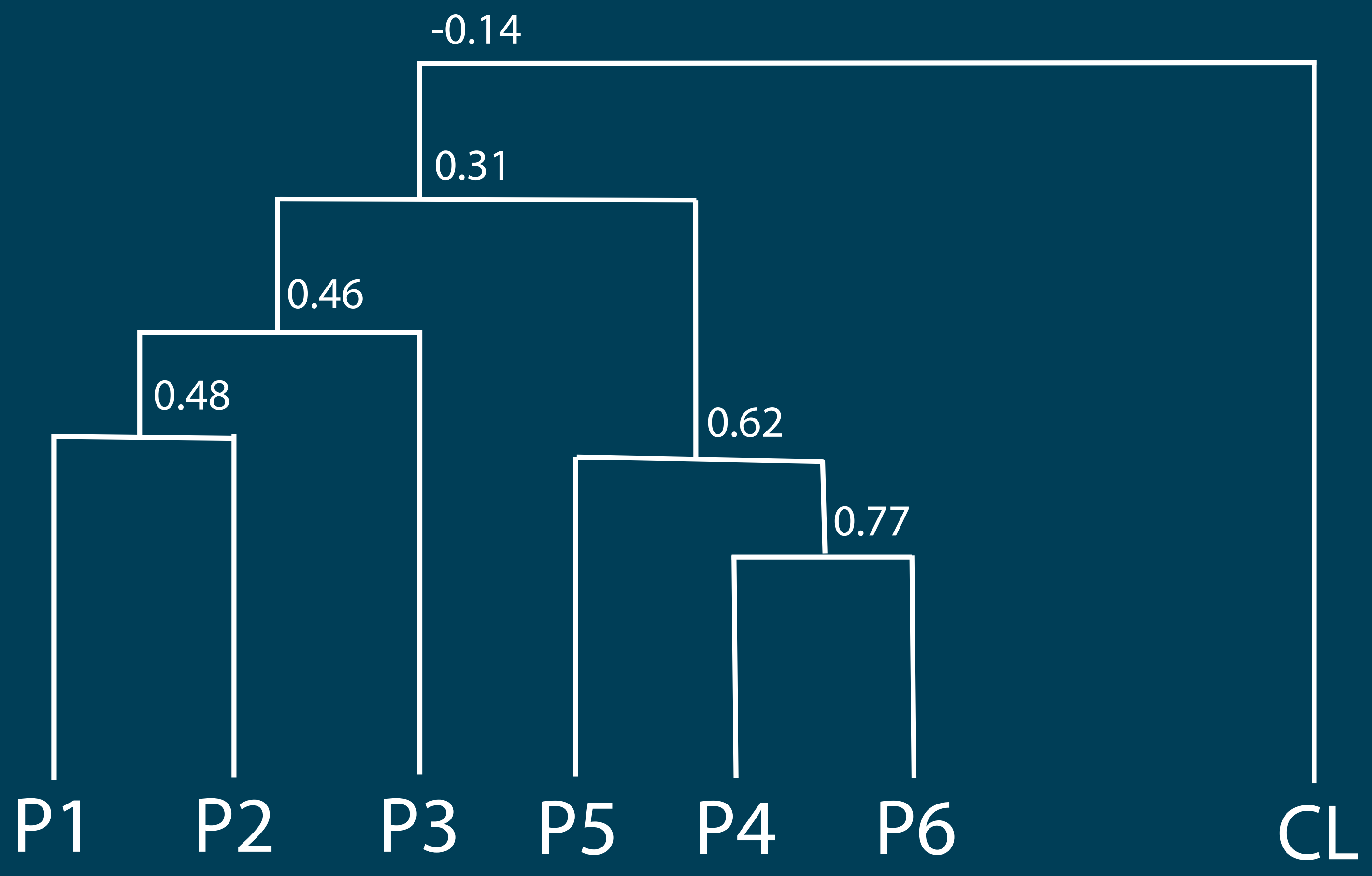
	P1	P2	P3	P4	P5	P6
OM MIA (4)	0	1	1	0	0	0
OM PC (11)	0	1	1	0	0	0
OM ALL (114)	0	2	1	1	1	1
HP MIA (84)	3	6	4	5	4	3
HP PC (1769)	16	23	19	11	23	8
HP ALL (64,669)	110	180	143	97	136	65
Tumor %	90%	90%	100%	0%?	60%	40%

OM = OncoMap, HP = hybrid capture with probes

Clustering via Differential Expression

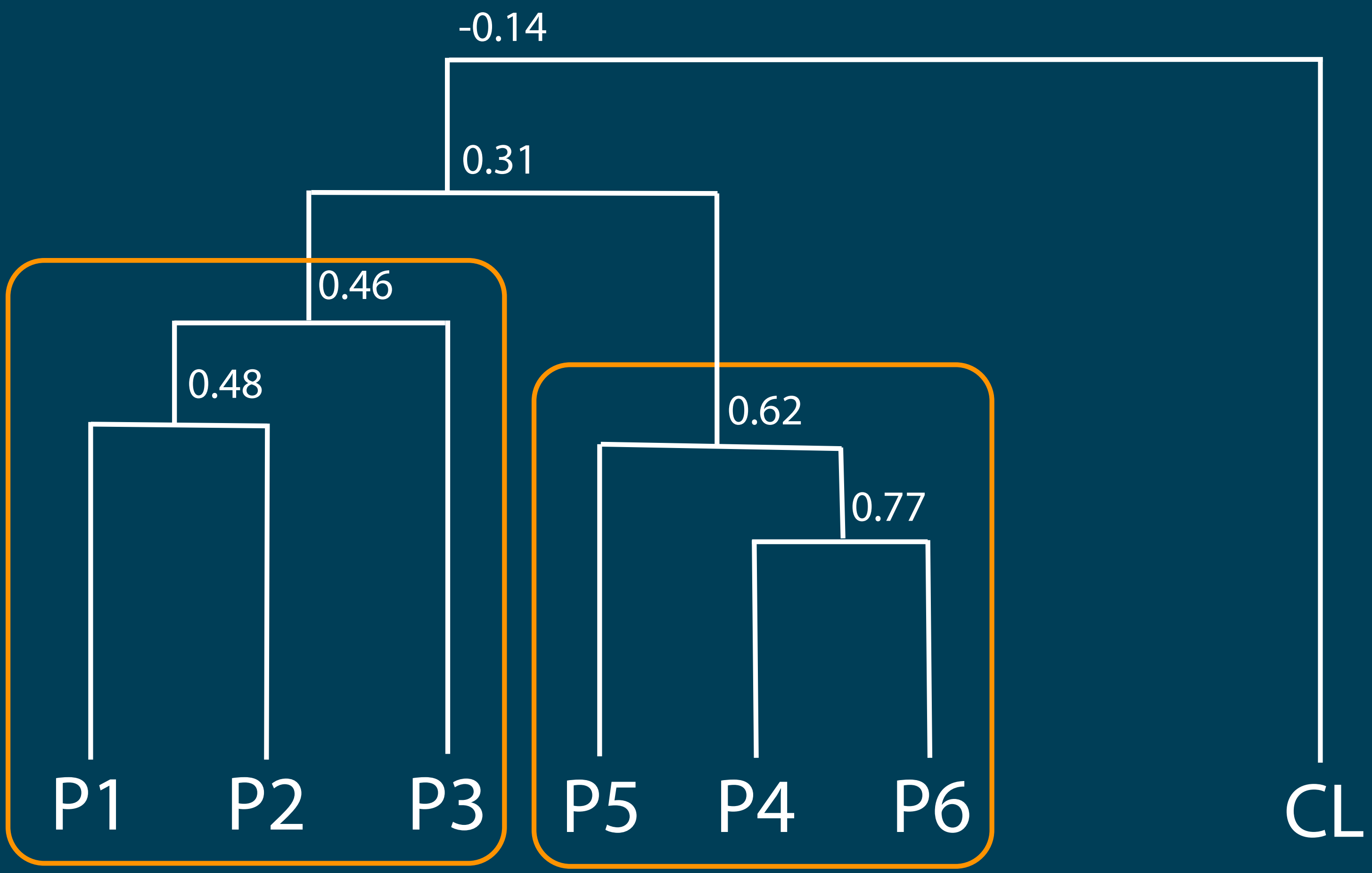

Workflow Canvas | Cuffdiff DE to Cluster Input

Gene Expression Clustering



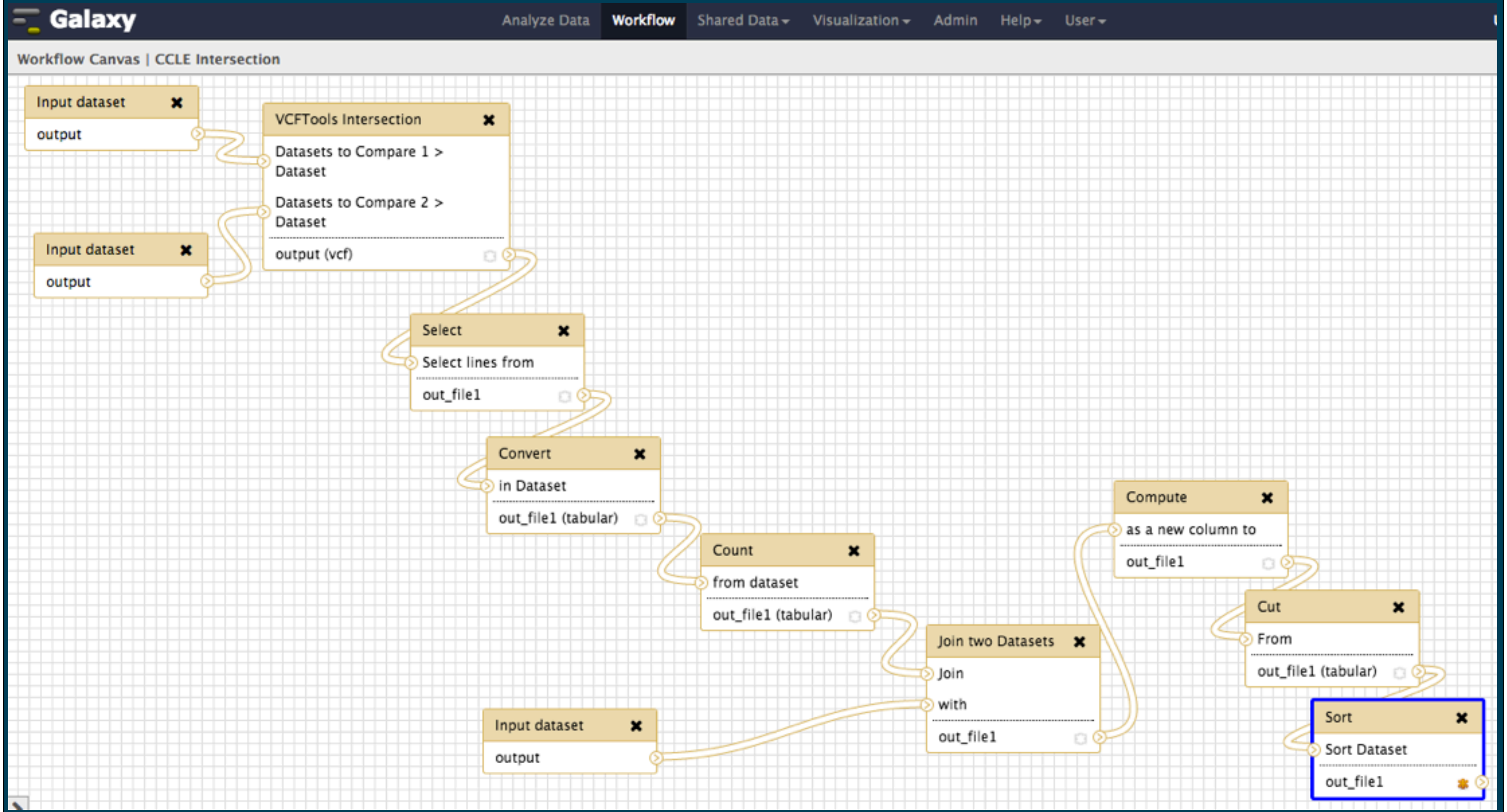
Spearman Correlation

Gene Expression Clustering



Spearman Correlation

Matching Patients to Cell Lines



Matching Patients to Cell Lines

	Best Match (# muts)	Best Match (#, % muts)
P1	KP3, KP2, KP4, PANC0327, PANC1005, QGP1 (4)	KP3 (4, 4.8%)
P2	PANC0327 (8)	CAPAN2 (7, 9.1%)
P3	SNU410, QGP1 (6)	KP3 (5, 6.0%)
P4	CAPAN2, PANC0403, MIAPACA2, PANC0327 (5)	CAPAN (5, 6.5%)
P5	T3M4 (8)	T3M4 (8, 8.7%)
P6	CAPAN2, MIAPACA2 (3)	CAPAN2 (3, 4.8%)

*38 pancreatic cell lines in CCLE

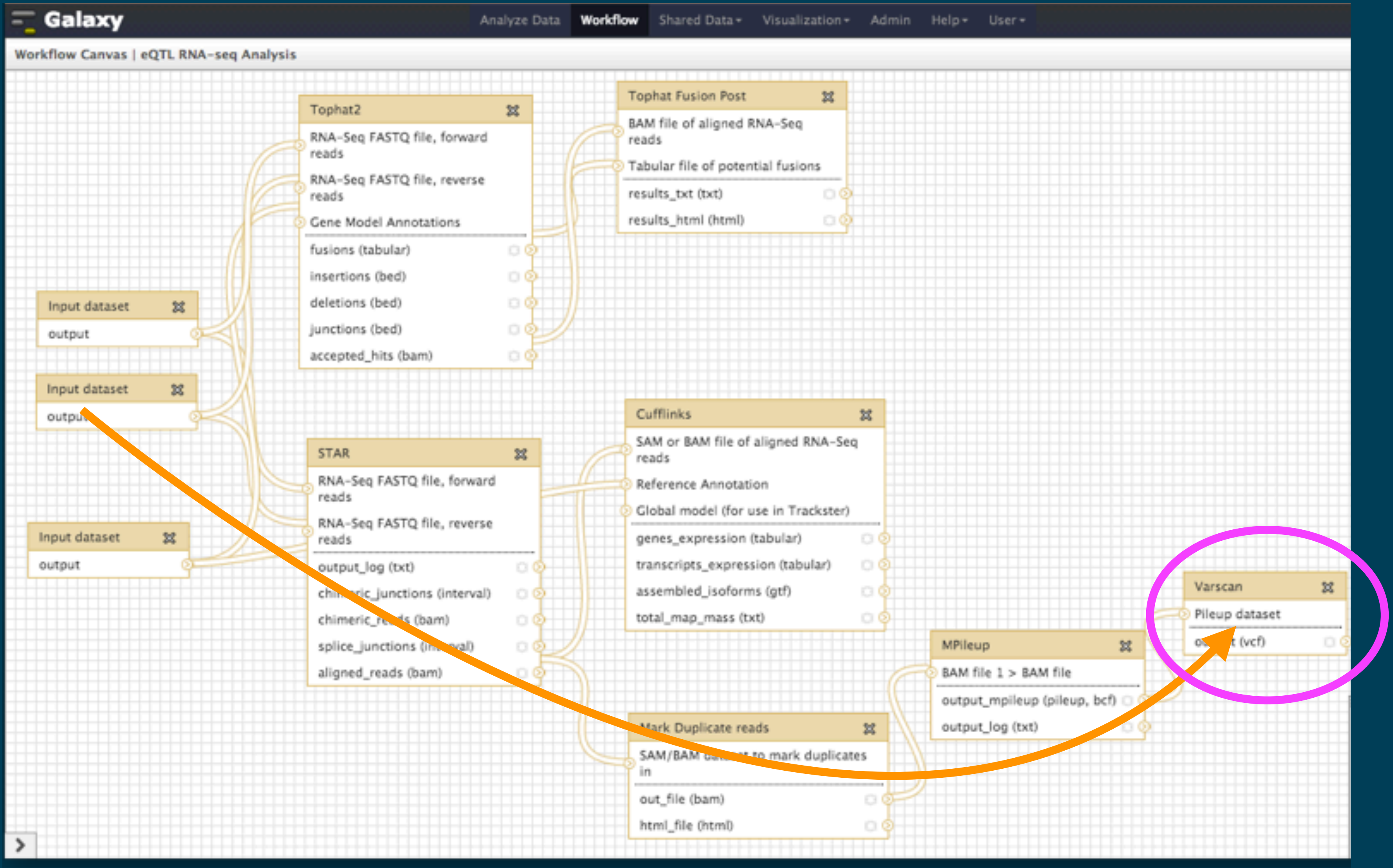
Topics

Galaxy

Analyzing Cancer Genomes & Transcriptomes

Web-based Visual Analysis

Mutation Calling from RNA-seq



Variant calling from 6 patient, 700GB pileup file requires 48 hours to complete

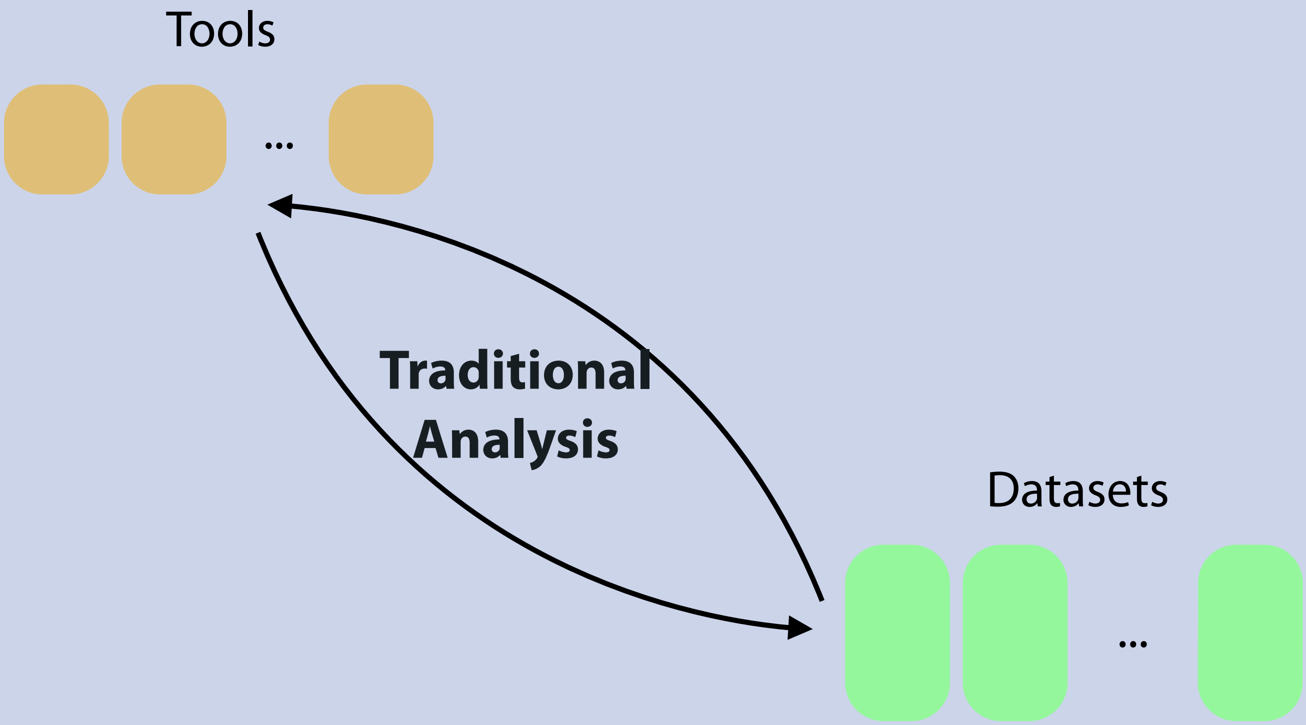
Why Visual Analysis?

Challenges for bioinformatics tools

- time and compute intensive
- many parameters
- parameter sensitive and unpredictable

Approach: repeatedly run tools and visualize outputs to compare and find best parameters/approach

Galaxy



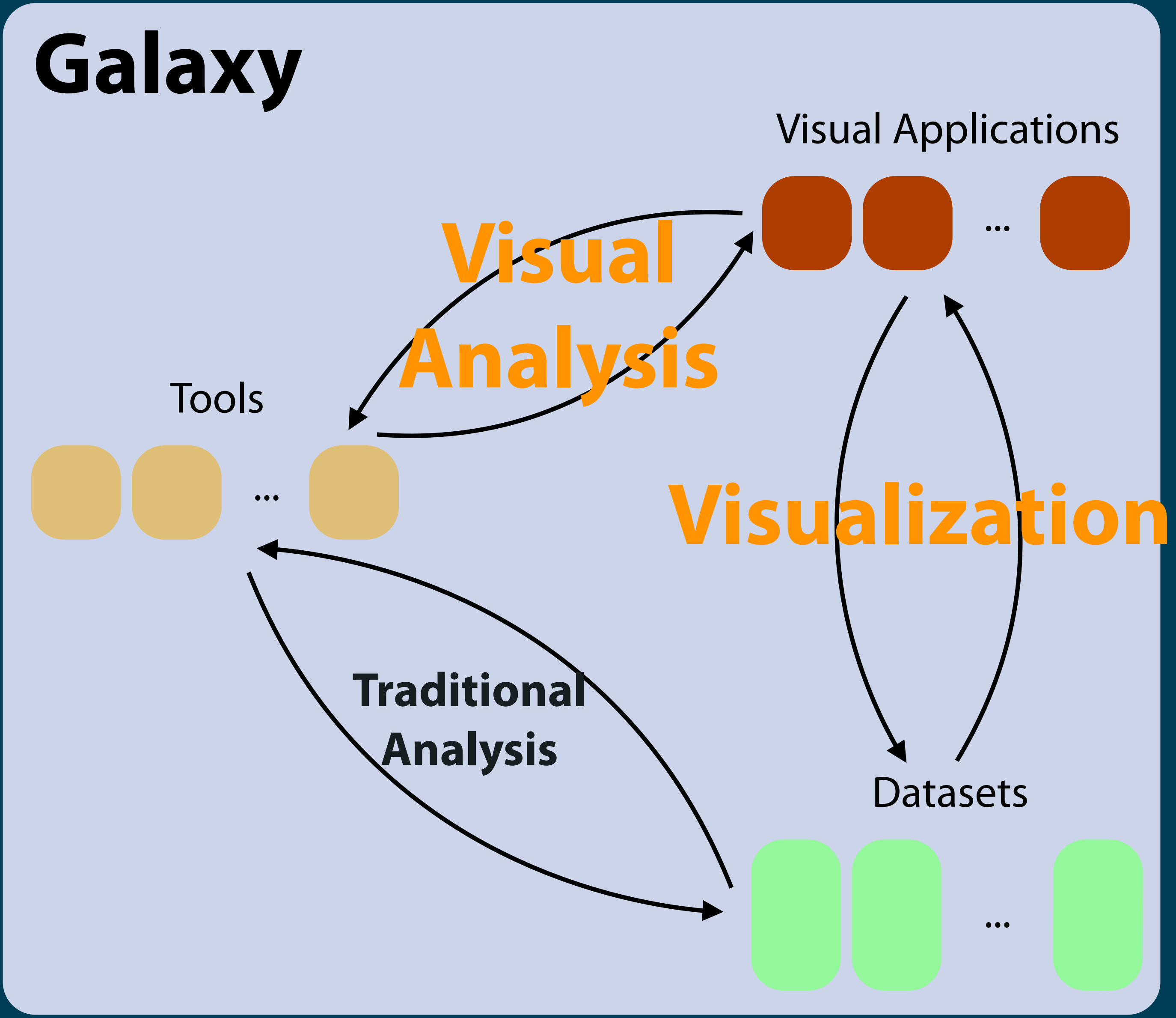
Tools

...

Traditional Analysis

Datasets

...



Galaxy

Visual Applications

Tools

Datasets

Visual Analysis

Visualization

Traditional Analysis

Goecks et al. (2013) *BMC Genomics*

Web-based Visualization for High-throughput Genomic Datasets

State-of-the-art data management

- automatic indexing for aggregate data and individual data points
- retrieve data on demand and cache

Render in browser for speed and flexibility

Can share and publish fully-functional visualizations

Goecks et al. (2011) *IEEE BioVis*

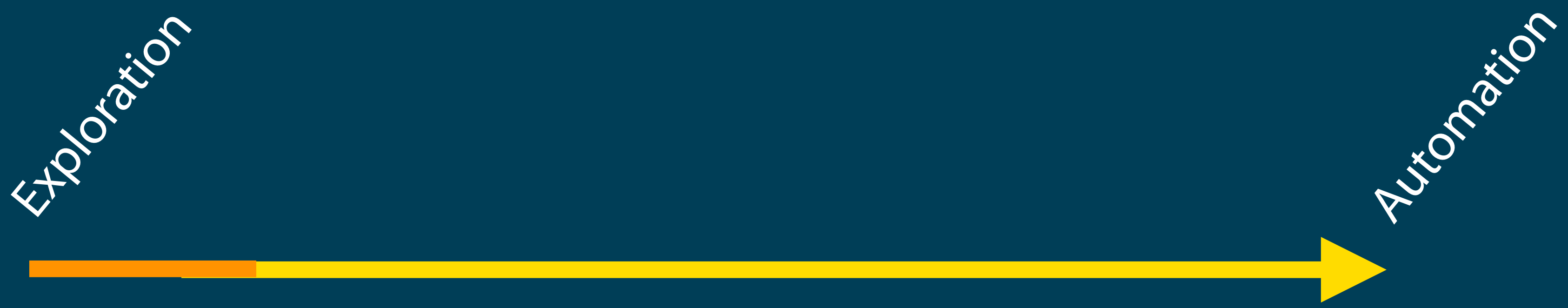
Demo: Visual Analysis

Real-time Visual Analysis

Interactive use of production tool to call and visualize variants for multiple patients using parameter sweeps

A general approach for interactive visual analysis on very large genomics datasets

- any Galaxy visual application, many tools (original application: transcript assembly)
- can decide what data to analyze on the fly



Exploration

Automation

Goecks et al. (2012) *Nature Biotechnology*

Concluding Thoughts

Galaxy is a very useful platform for high-throughout genomics

- accessible, reproducible, collaborative
- public, local, cloud

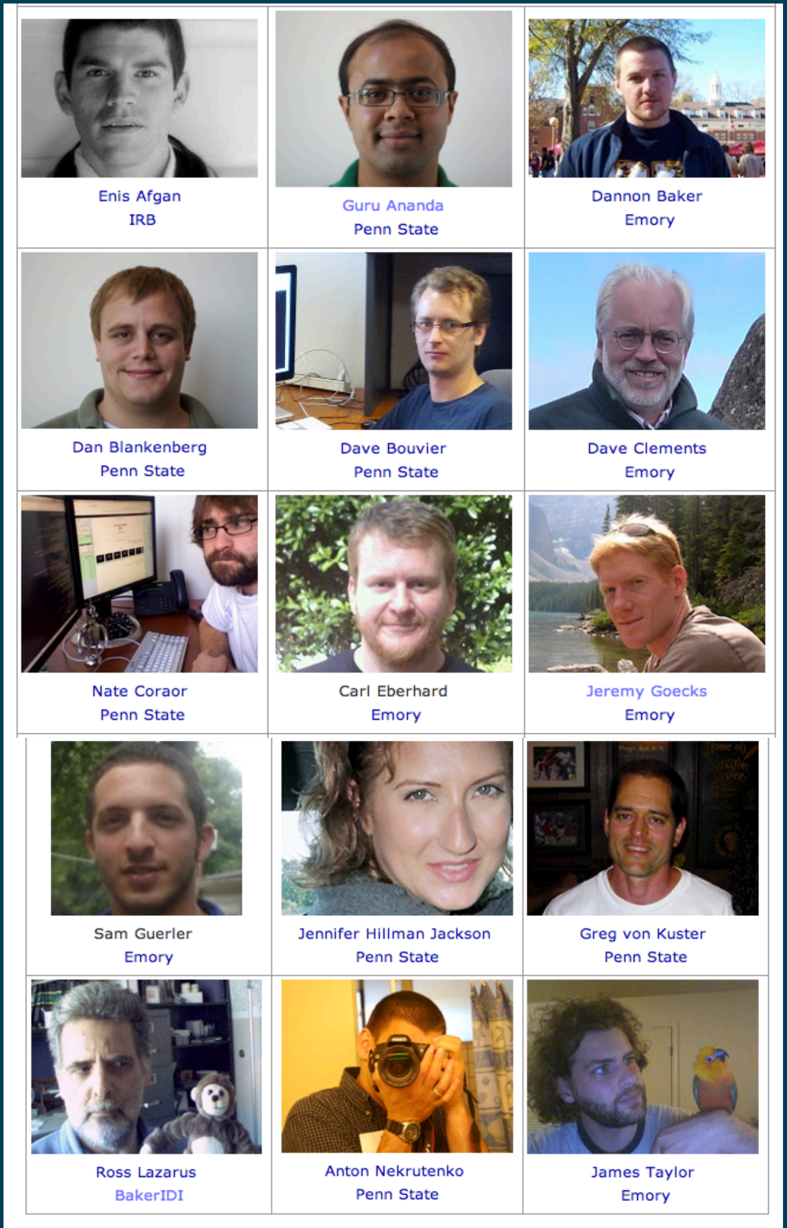
New tools, workflows, and visual analysis tools for analyzing high-throughput cancer sequencing data

- match patients to drug-profiled cancer cell lines via variants
- and soon variants + gene expression

Visualization/visual analysis are first-class objects in Galaxy

- visual analysis affords rapid experimentation with tool parameters

Galaxy



Enis Afgan IRB	Guru Ananda Penn State	Dannon Baker Emory
Dan Blankenberg Penn State	Dave Bouvier Penn State	Dave Clements Emory
Nate Coraor Penn State	Carl Eberhard Emory	Jeremy Goecks Emory
Sam Guerler Emory	Jennifer Hillman Jackson Penn State	Greg von Kuster Penn State
Ross Lazarus BakerIDI	Anton Nekrutenko Penn State	James Taylor Emory

EMORY | WINSHIP CANCER INSTITUTE

A Cancer Center Designated by the National Cancer Institute

Mike Rossi

EMORY UNIVERSITY

PENNSTATE 1855

genome.gov
National Human Genome Research Institute
National Institutes of Health

NSF National Science Foundation
WHERE DISCOVERIES BEGIN

Thanks!
Questions?

Galaxy

http://galaxyproject.org

http://jeremygoecks.com
jeremy.goecks@emory.edu