

# Understanding Cancer Genomes (and Transcriptomes!) using Galaxy

Jeremy Goecks

Department of Biology  
 Department of Math and Computer Science  
 Emory University



EMORY  
 UNIVERSITY

# Advancing Computing and Genomics

## Research model

1. find computing challenge while doing genomics
2. invent new computing technology to address challenge
3. demonstrate usefulness of new technology via genomics investigation

# Advancing Computing and Genomics

## Realizing research model

1. computing challenge: **analyzing cancer genomes**
2. new computing technology: **Galaxy tools, workflows, and visual analysis**
3. genomics investigation: **pancreatic cancer transcriptome**

# Roadmap

## Galaxy

Analyzing Cancer Genomes and  
Transcriptomes

# Vision

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

# A User Perspective of Galaxy

## GUI for high-throughput, high-performance genomics

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

## Customizable open-source software on various HPC resources

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

# A Developer Perspective of Galaxy

## Module (plug-in) architecture

- ✦ {attributes + behaviors} define a module, and module implementations are written as needed

## There are Galaxy module definitions/support for:

- ✦ tools
- ✦ data types (file formats)
- ✦ data sources (e.g., sequencers)
- ✦ data stores (file systems)
- ✦ job scheduling engines (e.g., DRMAA, Condor)
- ✦ visualizations/visual analysis (e.g., genome browser, Circos plot, scatterplot)

# Roadmap

Galaxy

**Analyzing Cancer Genomes and  
Transcriptomes**



# Cancer Genomics

The New York Times

March 26, 2013

## New Prostate Cancer Tests Could Reduce False Alarms

By ANDREW POLLACK

Sophisticated new prostate cancer tests are coming to market that might supplement the unreliable P.S.A. test, potentially saving tens of thousands of men each year from unnecessary biopsies, operations and radiation treatments.

Some of the tests are aimed at reducing the false alarms, and accompanying anxiety, caused by elevated P.S.A. readings. Others, intended for use after a definitive diagnosis, examine the genetic workings of the cancer to distinguish dangerous tumors that need treatment from slow-growing ones that might be left alone.

The New York Times

April 21, 2013

## Cancer Centers Racing to Map Patients' Genes

By ANEMONA HARTOCOLLIS

The promise of whole genome sequencing can be seen in trials like one for bladder cancer at Memorial, where the effects of a drug normally used for breast cancer were disappointing in all but one of about 40 patients, whose tumor went away, Dr. Baselga said. Investigators sequenced the patient's whole genome. "The patient had a mutation in one gene that was right on the same pathway as the therapy," Dr. Baselga said. "And that explained why this worked."

# Using Galaxy for Analysis of Cancer Genomes/Transcriptomes

## New tools

- ✦ e.g. variant calling, fusion detection, variant annotation and filtering, VCF manipulation

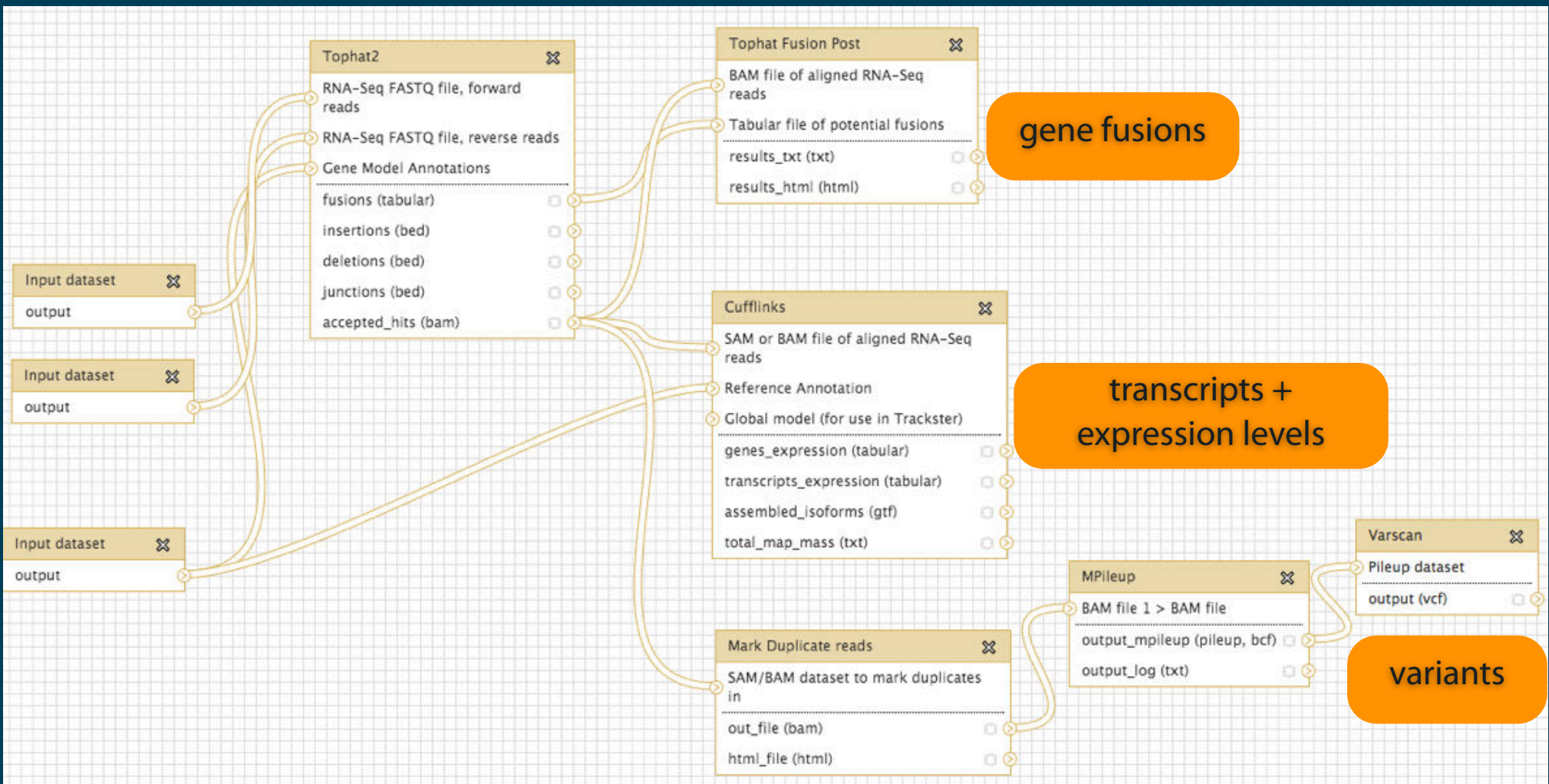
## New workflows

- ✦ workflows are understandable and extendable

## New visual analysis applications

- ✦ visualize and call variants in a Web browser

# Single Sample Transcriptome Analysis



# Advantages of Galaxy Workflows

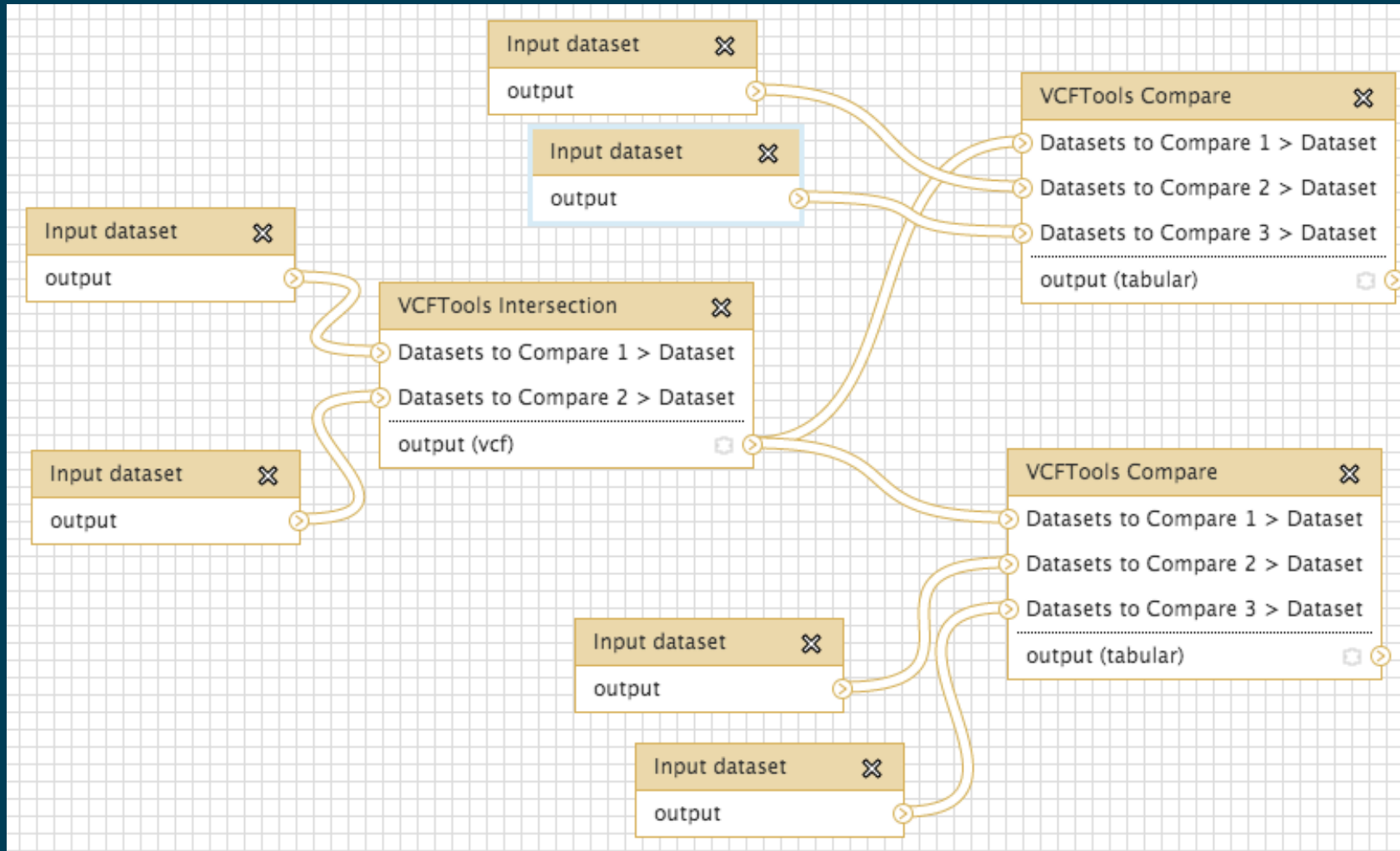
Not a black box, so can swap out tools and modify parameters

- ✦ recomputable

Human readable, especially for non-programmers

“-able”: import, export, share, publish, embed

# Comparing Called Variants with Public Datasets



# Patient Mutations vs.



<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
HP MIA (84)	6	6	5	5	4	4	19
HP PC (1769)	21	29	23	14	29	15	49

**Cell line does not appear very similar to tumors**

OM = OncoMap, HP = hybrid capture with probes

# Using Mutations for Characterizing Tumors

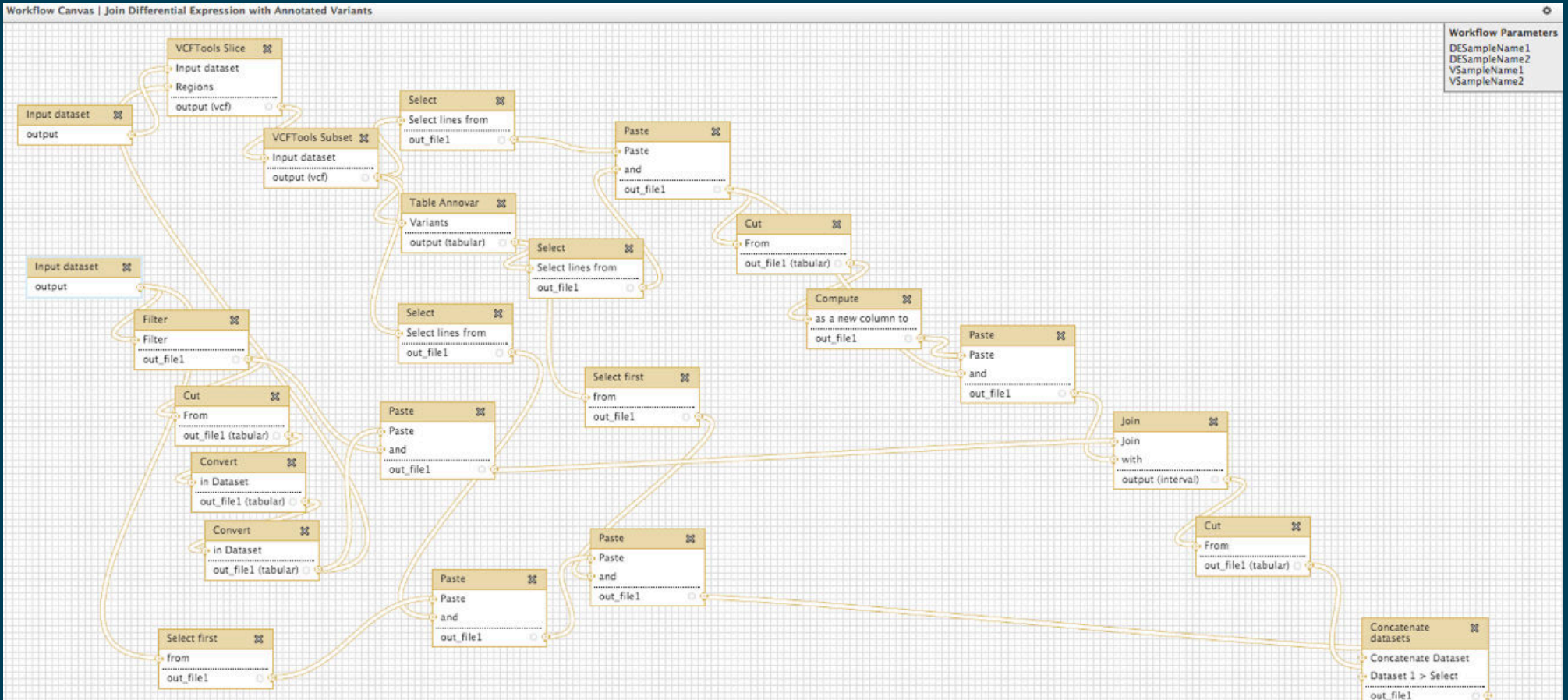
	P1	P2	P3	P4	P5	P6
OM MIA (4)	0	1	1	0	0	0
OM PC (11)	0	1	1	0	0	0
HP MIA (84)	6	6	5	5	4	4
HP PC (1769)	21	29	23	14	29	15
Tumor %	90%	90%	100%	0%?	60%	40%

OM = OncoMap, HP = hybrid capture with probes

(See ISMB talk for discussion of clustering patients via gene expression data from RNA-seq.)

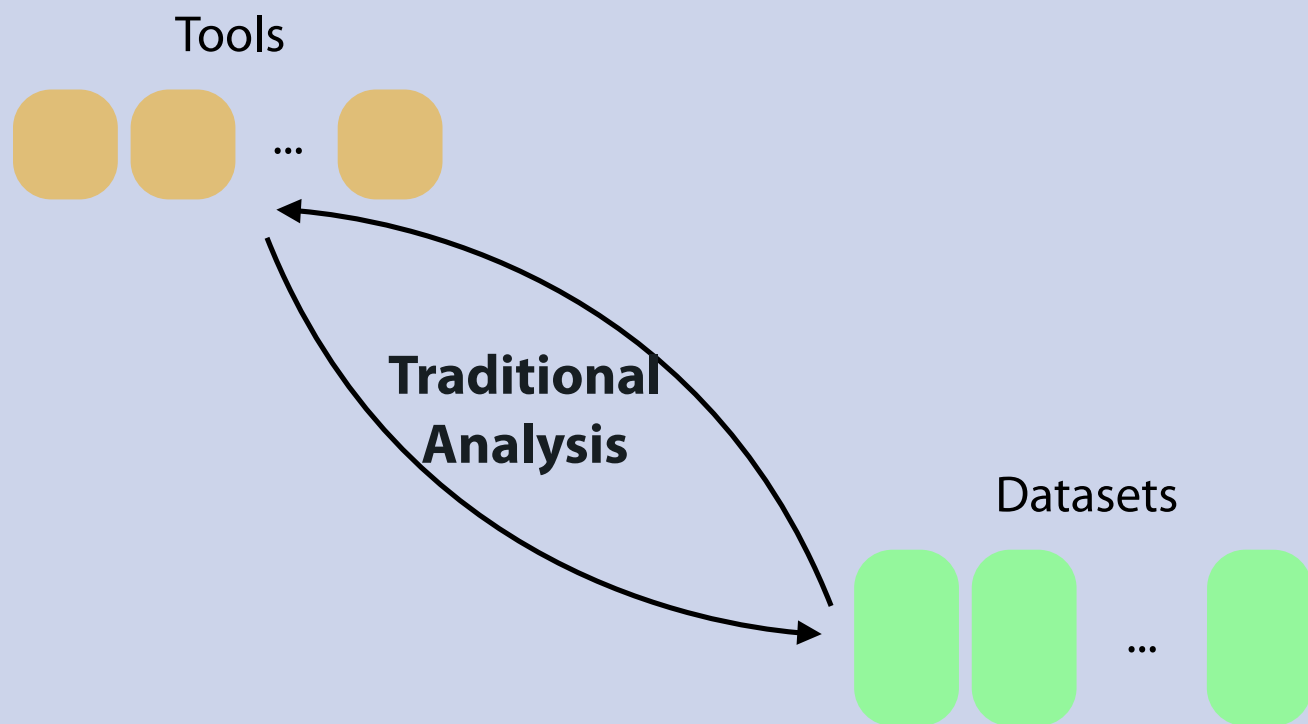


# Variants + Gene Expression + Annotation: Targeted eQTL Analysis

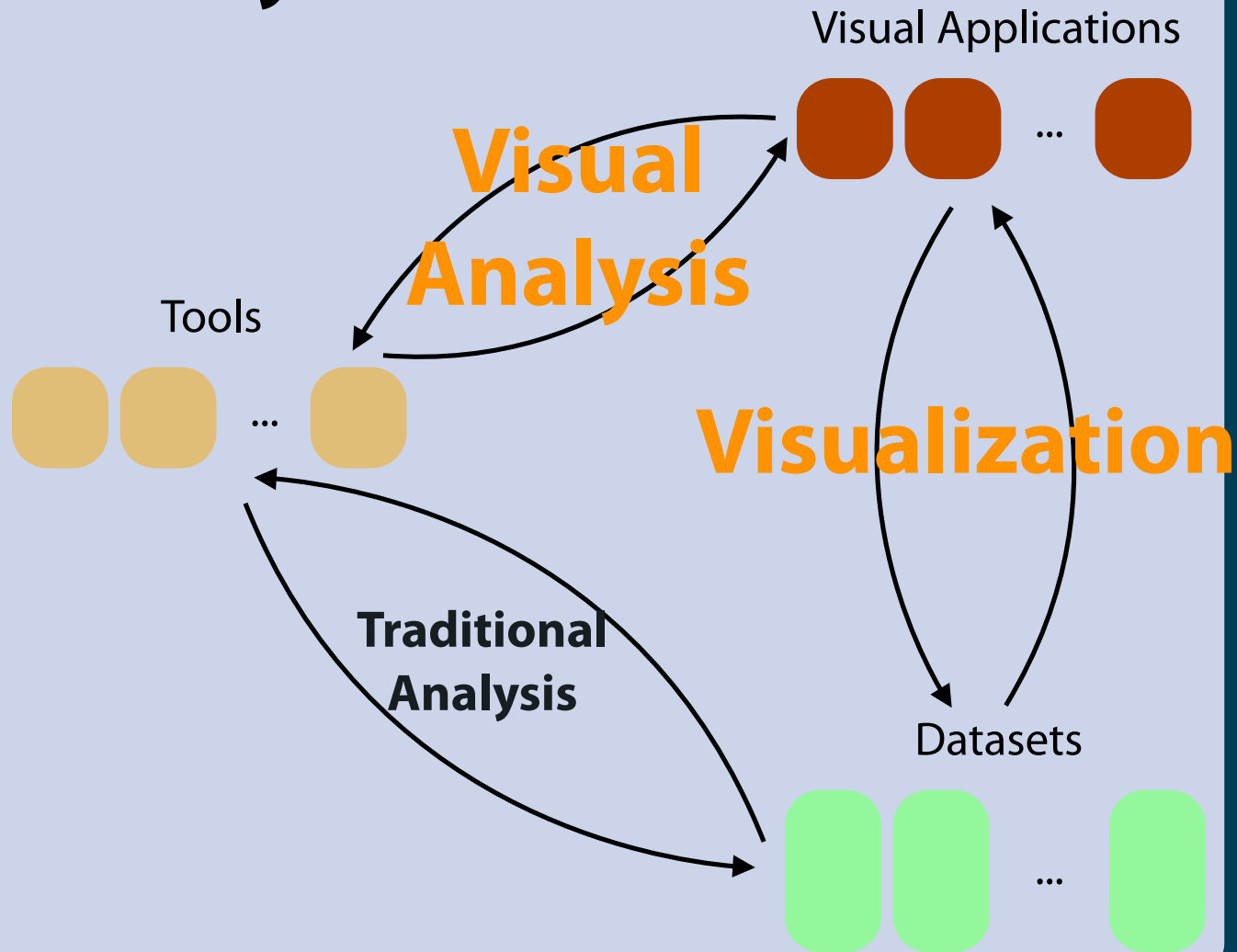


Finds and annotates variants in differentially-expressed genes or isoforms

# Galaxy



# Galaxy



# Web-based Visualization for High-throughput Genomic Datasets

## State-of-the-art data management

- ✦ automatic indexing for aggregate data and individual data points
- ✦ data on demand + multi-level caching

## Can share and publish fully-functional visualizations

## Framework for adding new visualizations

- ✦ similar to tool config in Galaxy

# Trackster

Galaxy

localhost:8080/visualization/trackster?id=e56924ee0dd89f8c#chr19:45909678-45927509

Galaxy Analyze Data Workflow Shared Data Visualization Admin Help User Using 2.3 GB

PanCan Patient Mutations (hg19) chr19 45,909,678 - 45,927,509 45,910,000 45,920,000

Varscan

Varscan - region=[all], parameters=[mpileup2snp, 16, 4, 15, 0.01, 0.75, 0.99, no]

Varscan

Analysis type: single nucleotide variation

Minimum read depth: 10

Minimum supporting reads: 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: 0.99

Ignore variants with >90% support on one strand: no

Run on complete dataset Run on visible region

Summary

Sample1

Sample2

Sample3

Sample4

Sample5

Sample6

Varscan - region=[chr19:45909678-45927509], parameters=[mpileup2snp, 32, 4, 15, 0.01, 0.75, 0.99, no]

Summary

Sample1

Sample2

Sample3

Sample4

Sample5

Sample6

Genomes gene annotation

NM\_001983

001166049

NM\_202001

45,910,000 45,920,000

# Sweepster

The screenshot displays the Galaxy web interface for the Sweepster tool. The browser address bar shows the URL: `tachylite01.bx.mathcs.emory.edu/g/jeremy/visualization/sweepster?dataset_id=9784cc0967daee50&hda_ldda=hda&regions=%5B%7B%20chr19%2C%20start%3A45...`

**Varscan (version 0.1) Parameters:**

- 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: 0.99
- Ignore variants with >90% support on one strand: no

**Visualization:**

The central visualization shows a tree diagram with a root node and six intermediate nodes labeled 10, 12, 14, 16, 18, and 20. Each node is connected to a vertical column of nodes representing minimum supporting reads (2, 4, 6, 8). The tree is labeled "Minimum read depth" and "Minimum supporting reads".

**Genomic Tracks:**

The right side of the interface displays a grid of genomic tracks for two regions: `chr19:45910004-45927835` and `chr19:45853270-45873990`. Each track shows read alignments with colored bars representing different alleles. A mouse cursor is visible over one of the tracks.





# 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
- ✦ **workflows soon!**



# Galaxy



EMORY

WINSHIP  
CANCER  
INSTITUTE

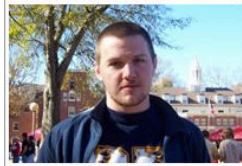
A Cancer Center Designated by  
the National Cancer Institute



Enis Afgan  
IRB



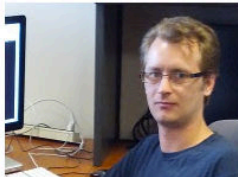
Guru Ananda  
Penn State



Dannon Baker  
Emory



Dan Blankenberg  
Penn State



Dave Bouvier  
Penn State



Dave Clements  
Emory



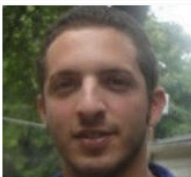
Nate Coraor  
Penn State



Carl Eberhard  
Emory



Jeremy Goecks  
Emory



Sam Guerler  
Emory



Jennifer Hillman Jackson  
Penn State



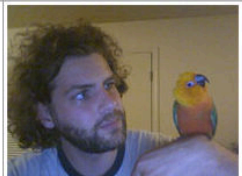
Greg von Kuster  
Penn State



Ross Lazarus  
BakerIDI



Anton Nekrutenko  
Penn State



James Taylor  
Emory



Mike Rossi



EMORY  
UNIVERSITY



genome.gov

National Human Genome Research Institute

National Institutes of Health



National Science Foundation

WHERE DISCOVERIES BEGIN

# Thanks! And More:



<http://galaxyproject.org>

<http://usegalaxy.org>

<http://bitbucket.org/galaxy/galaxy-central>

<http://wiki.galaxyproject.org>



## 1 Longer Talk, more Biology

Sunday, 14:10-14:35

## 2 Integrated Visualization & Computing Workshop

Tuesday, 14:10-16:05