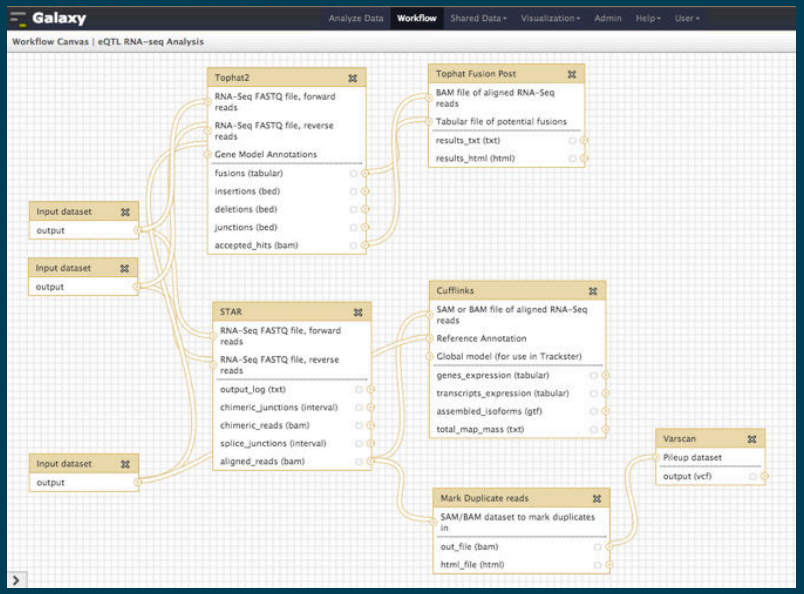
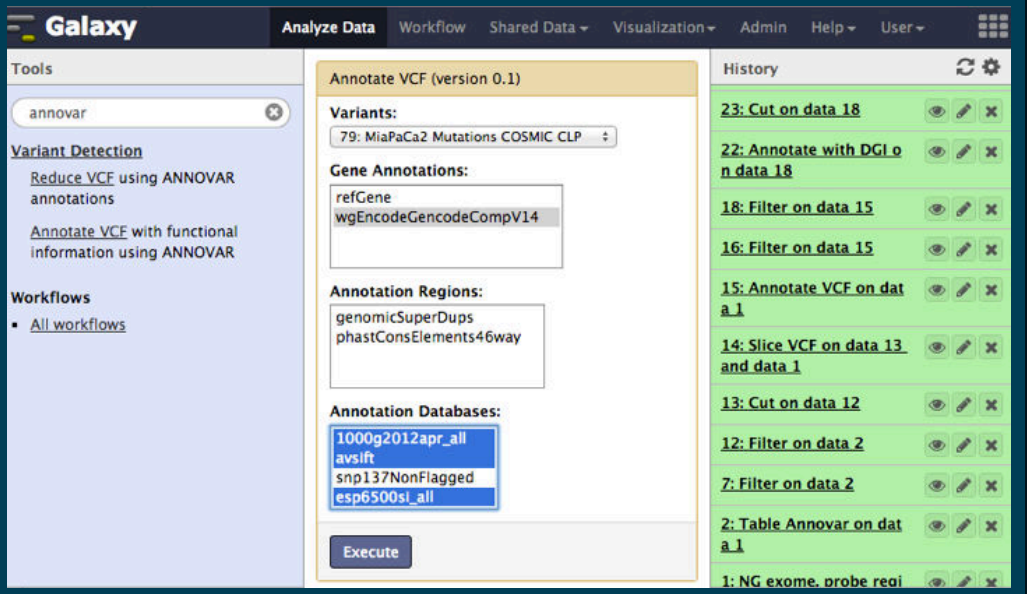




Using Galaxy to Understand Cancer Genomes

Jeremy Goecks

Computational Biology Institute

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Topics

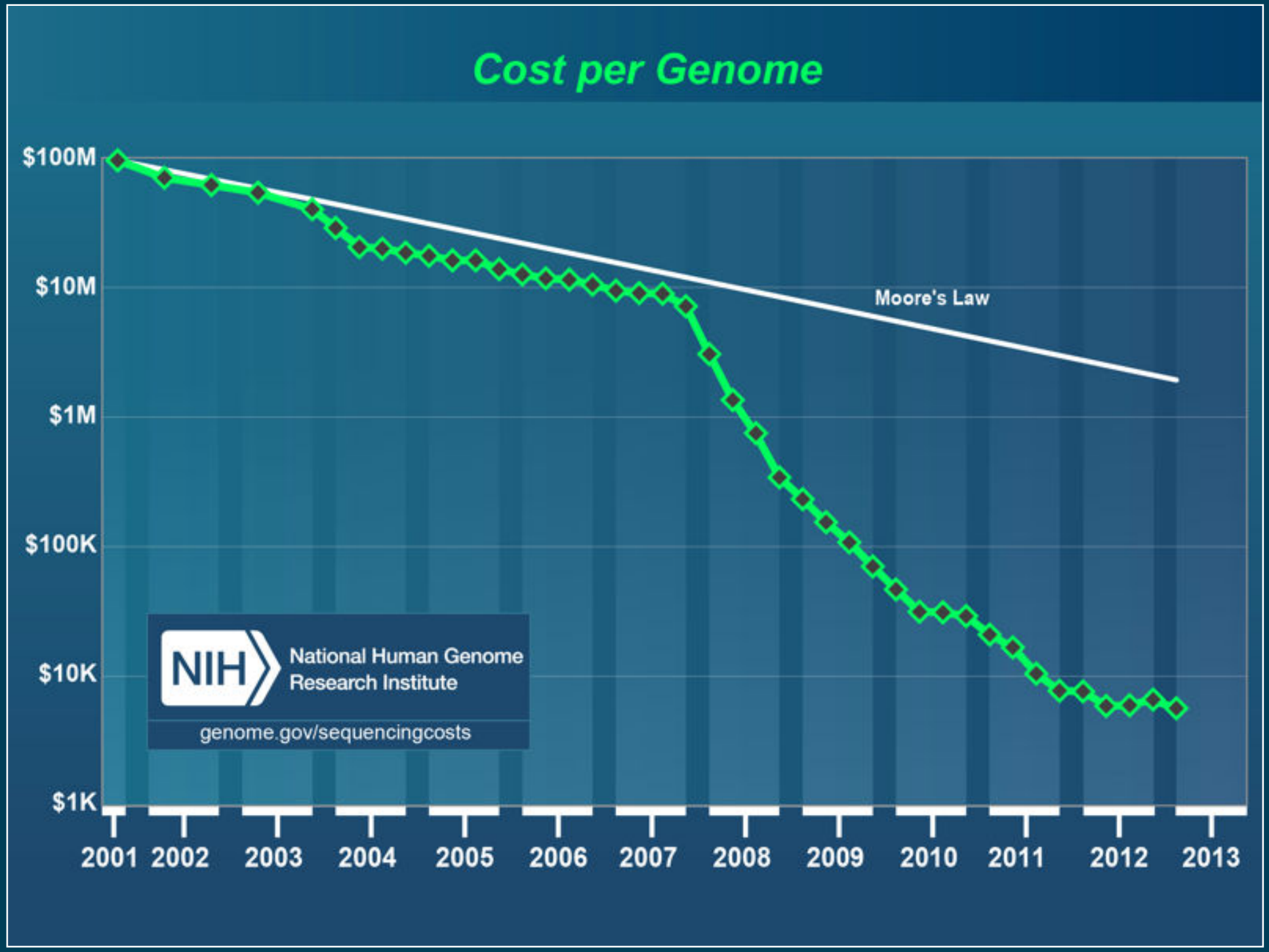
The High-throughput Sequencing Era

Galaxy

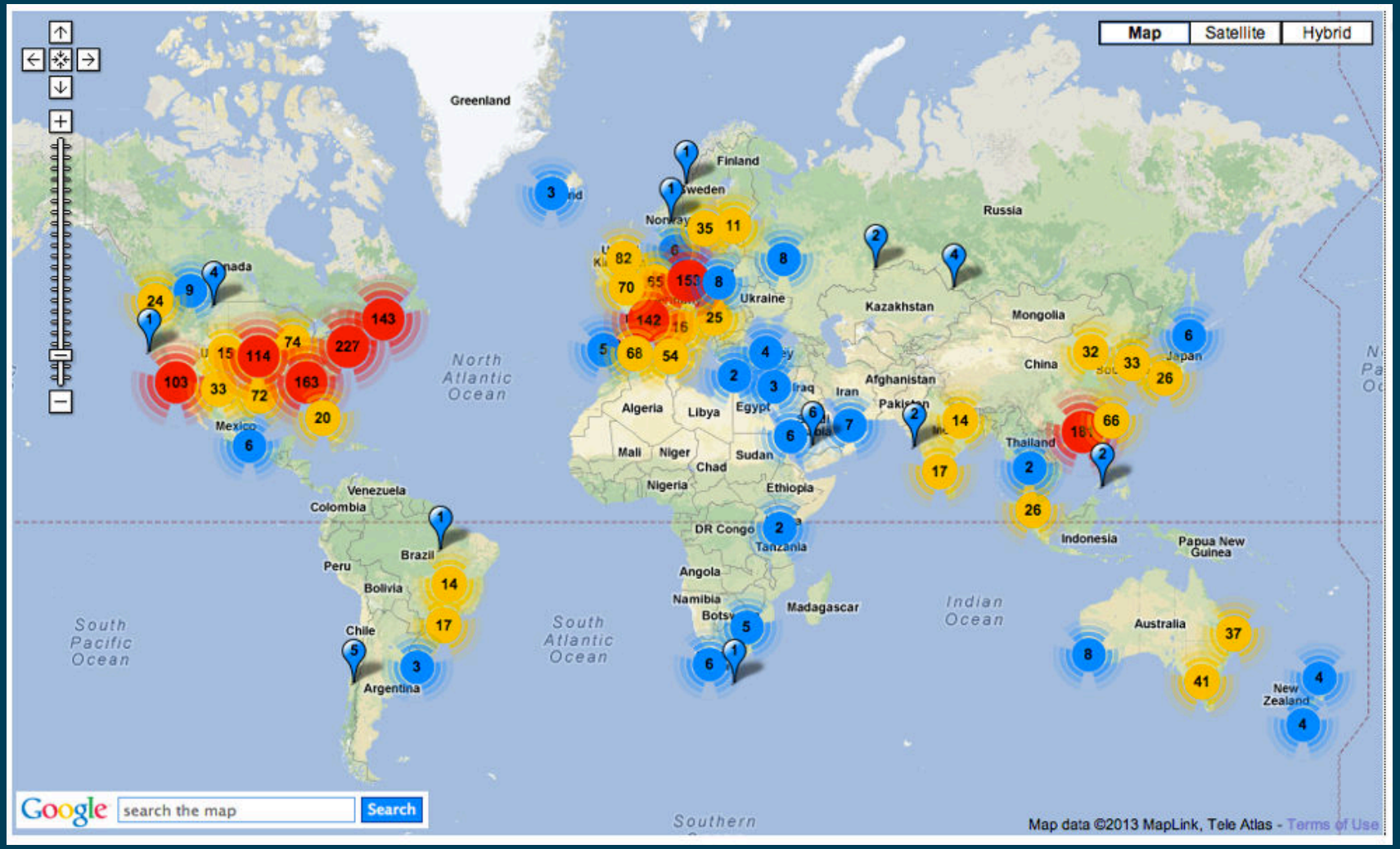
Cancer Genomics with Galaxy

A Revolution in Biology

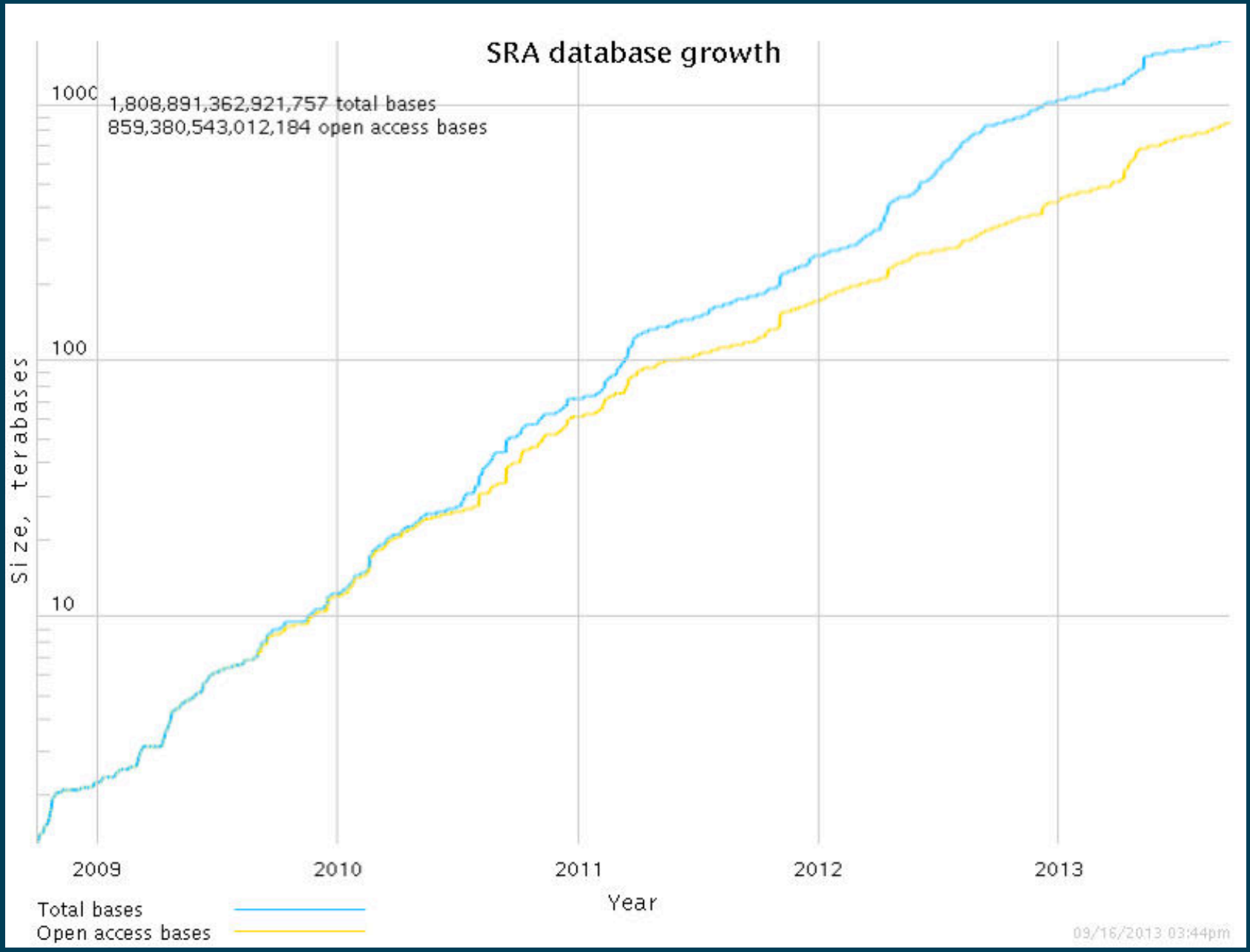
Low-cost, high-throughput sequencing technologies have become widespread


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013

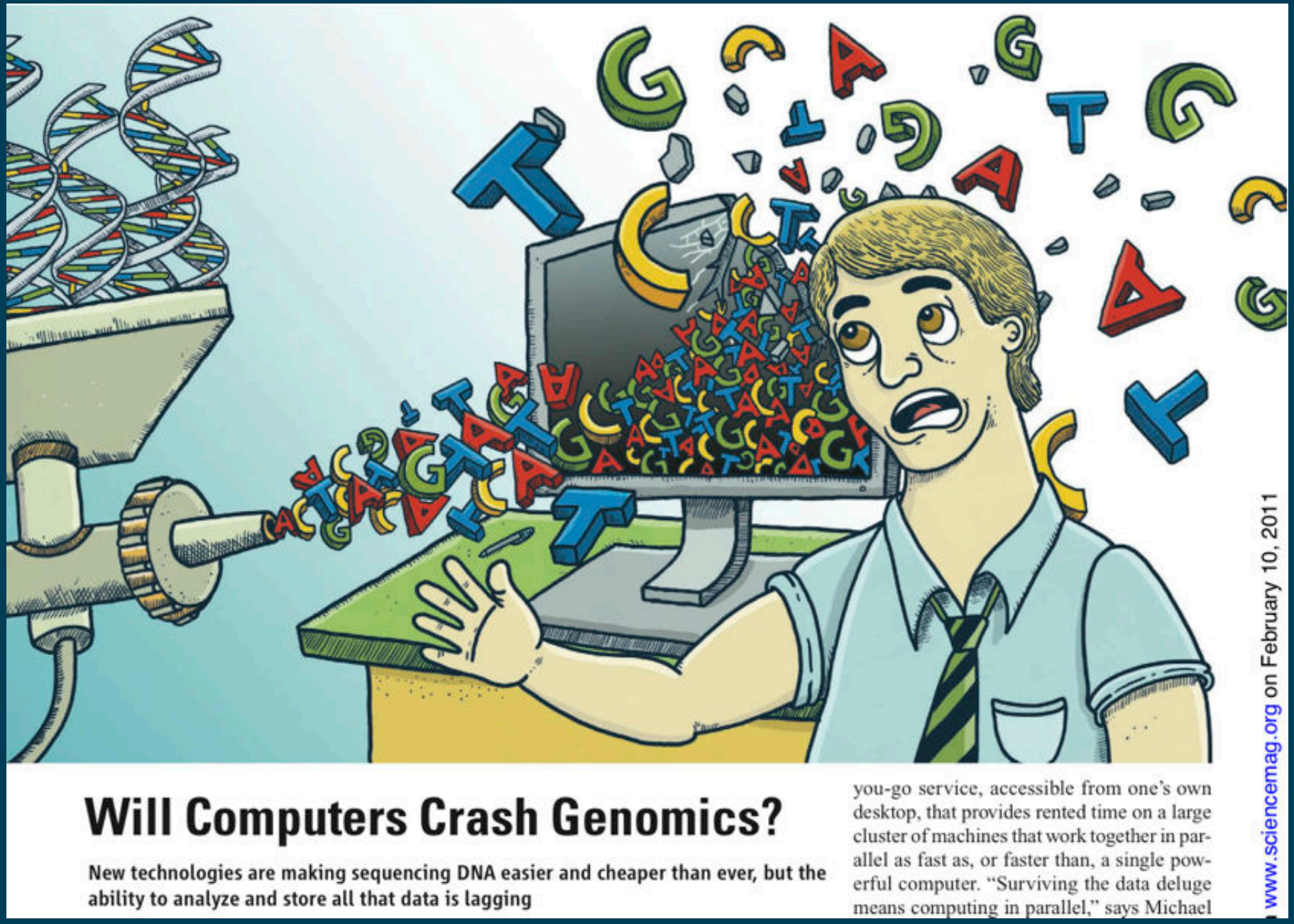
World Sequencing Capacity > 15Pb/year



http://omicsmaps.com



http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi



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

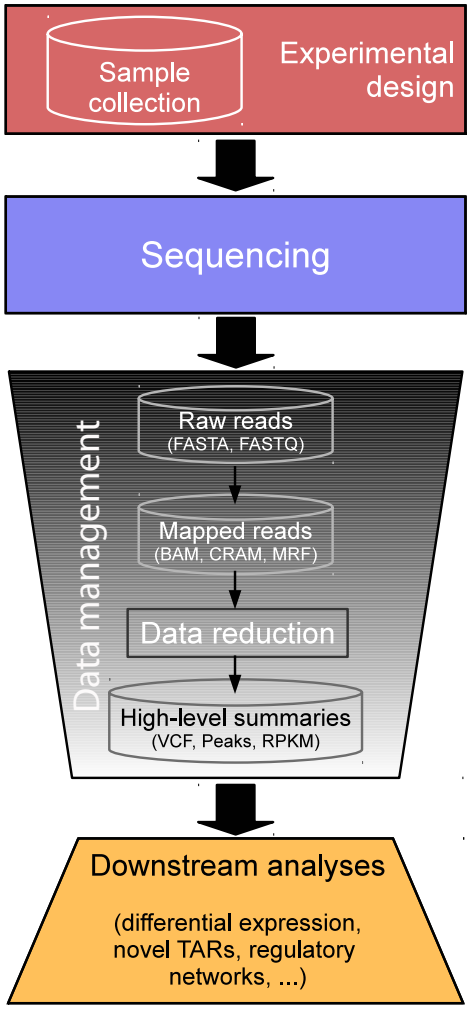
Sboner et al. Genome Biology 2011, 12:125
http://genomebiology.com/2011/12/8/125

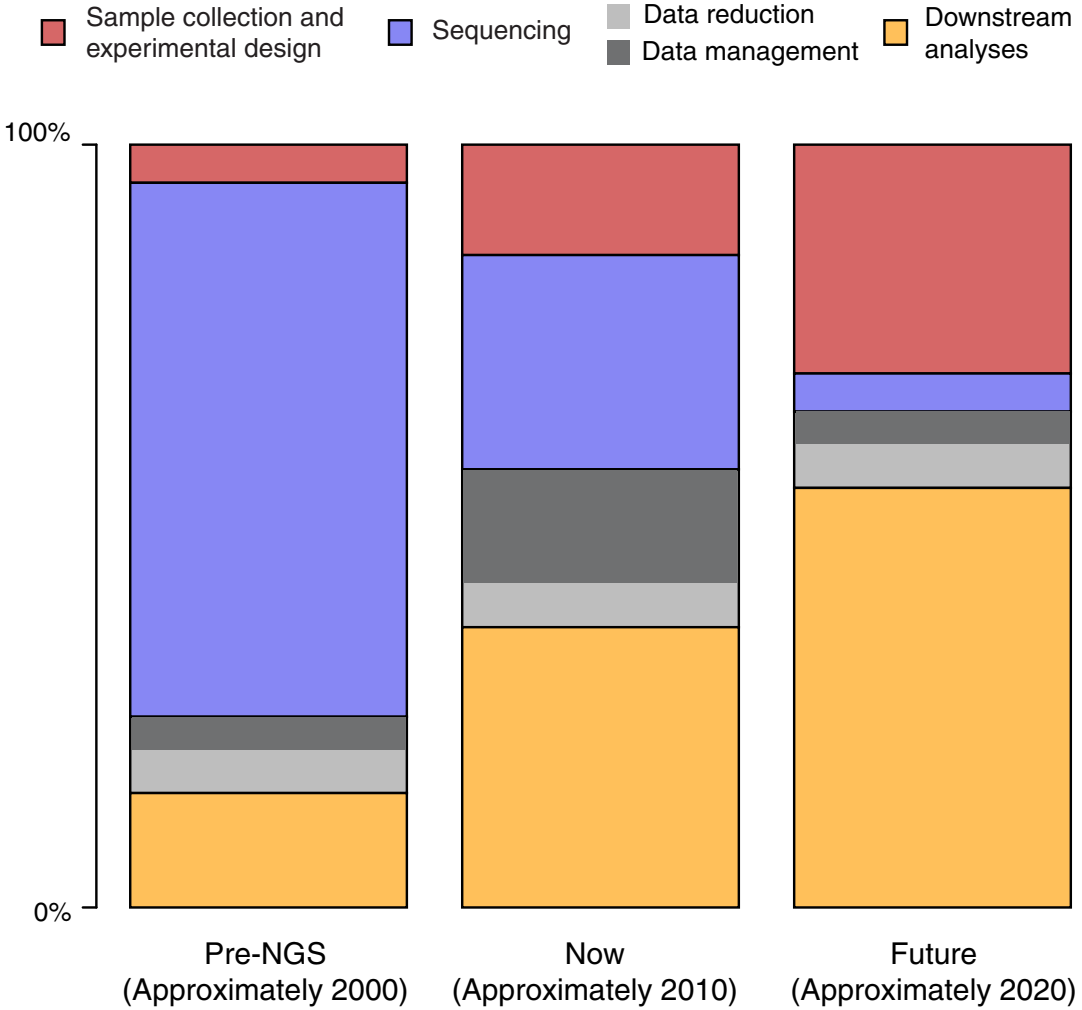
Genome Biology

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]





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?CBBCCCCCCCCCCCBBBBBBBBCCCCCBBBBBCCCCCCBBBBBBBCCCCDCC
@HWI-ST565:241:D19R1ACXX:1:1101:7520:1996 1:N:0:CGATGT
NCGCAACCTCAACACCACCTTCTTCGACCCCGCCGGAGGAGGAGACCCCATTCTATACCAACACCTATTCTGATTTTTCGGTCACCCTGAAGTTTATATTC
+
#4=DFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJIJHFHFFFDDDDDDDDFEEEFDDDDDDDDDDDDEEEEEDEDDDDDDDDDDDDDCDDD@CDEDEEE>
@HWI-ST565:241:D19R1ACXX:1:1101:10117:1998 1:N:0:CGATGT
NATGTGCCCTCTGGCAGTCTGCTGCTGTGTCCAGAGTCCGACTCCAGCTGGGCTGTAACTGGGCTTGGCCCCCGCCTTAGGCCCCGCCAGCAGGCGAAGCA
+
#1=DDFFFHHHHHJIJJJJJJJJJJJIJJJJJJJJIJIJJJJJJIJJJJJJIIJIGIJJJJJJHHHFFFFDDDDB=BCDDDDDDDDDDDCDDBB9>BDA
@HWI-ST565:241:D19R1ACXX:1:1101:10283:1992 1:N:0:CGATGT
NTTGTCACCAAGACCCTACTTCTAACCTCCCTGTTCTTATGAATTCGAACAGCATACCCCCGATTCCGCTACGACCAACTCATACACCTCCTATGAAAAAA
+
#4=DDFFFHHHHHJJJJJJJJJJJJJJJJJJJJIJJJJJJJIJJJJJJJJJJJJJJJJJGIHHHFFFFDDDDDDDBDDDDDDDDDDDDDDDDD@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<BCDDDDDDDDDDDDDDADCCCDDCDDDDDDDDDDDDEEDDBD?@@BCAAC
@HWI-ST565:241:D19R1ACXX:1:1101:10838:1994 1:N:0:CGATGT
NTCCCTGCTACTGCTGATGCACTGTCCTCTCCCTGCAGCCCCTGGCTTCCCAGCCTTCCTCCTGACCCCTTCCAACAGCCTTGGAACTCCAGCTGCCACCA
+
#1=DFFFFHHHHHJJJJJJJJIJJJJJJJJJJJJJJJJJJIJJJJJJJJIJJJJJJJJJJJJJIHHHHFFFFFDEEDDCDDDDDDDDDDDDDDDDDD8
@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=DDFFFHHHHHJJJJJJJJIJJIJJJJJJJJJGIIJJCGH?GHIIIJJJJJIJJJJJJJHIJJJJIIHHEEHHHFFFFFFEEEEDDD@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
```

Topics

The High-throughput Sequencing Era

Galaxy

Cancer Genomics with Galaxy

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*

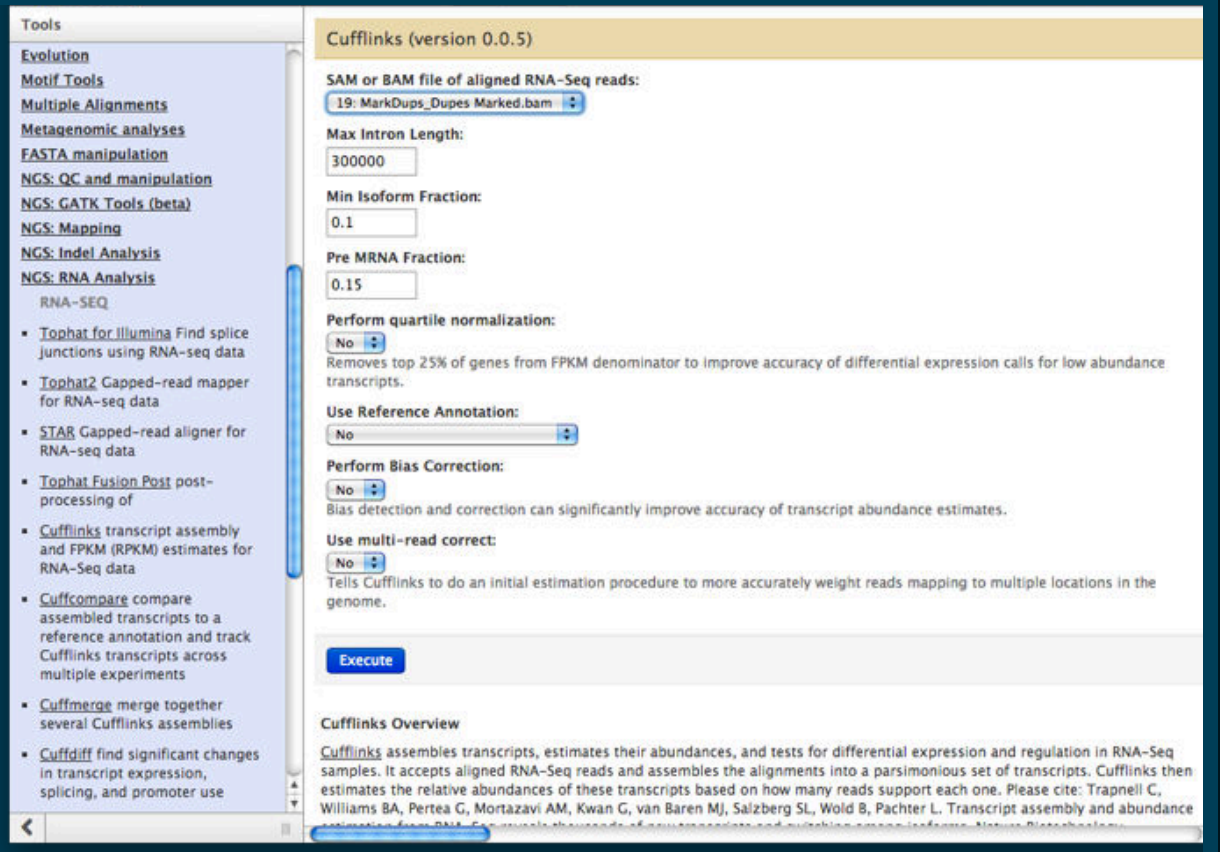
Galaxy Demo

Accessibility

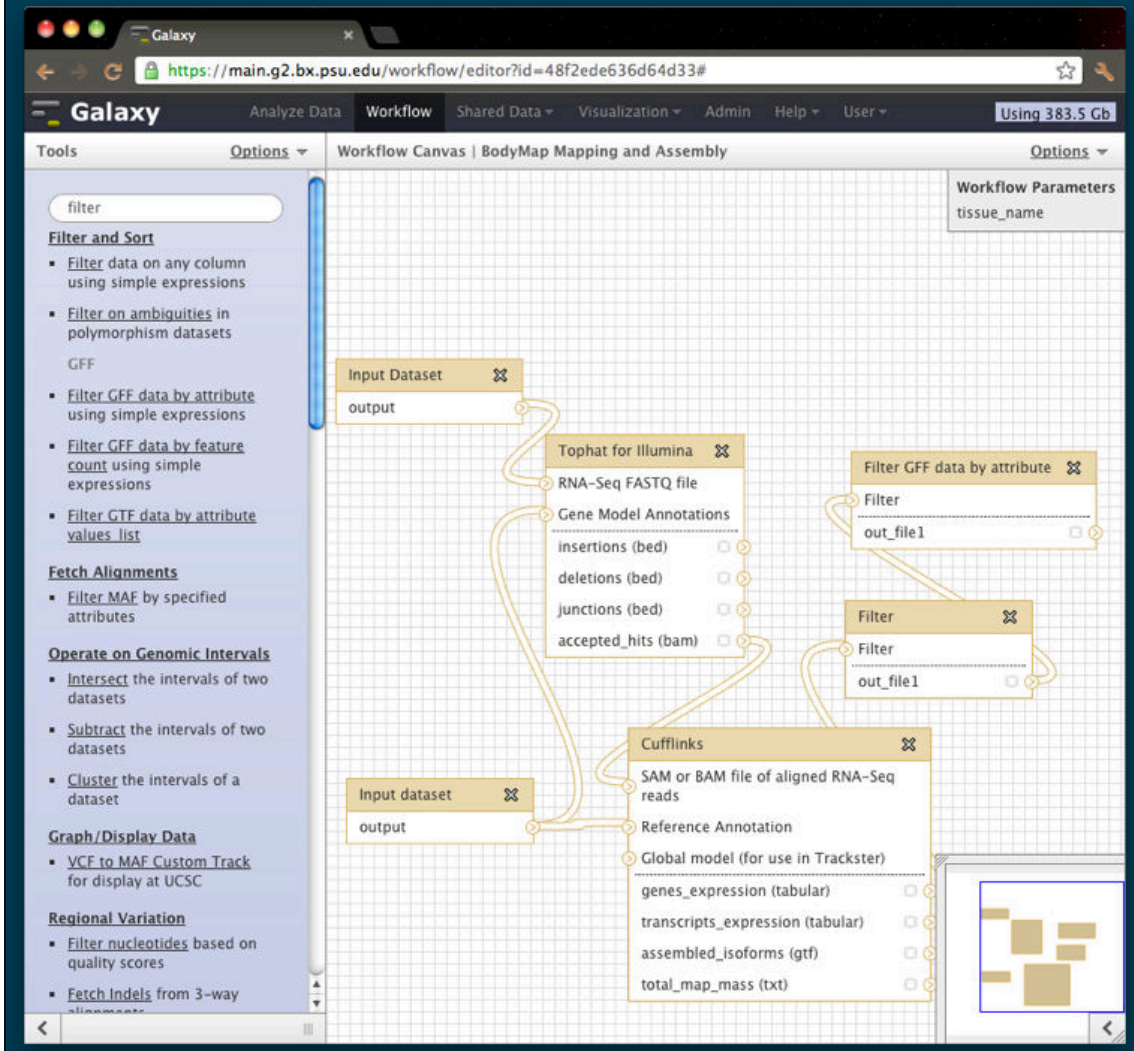
All tools looks the same

No command line or programming

Easy to chain tools together into larger analyses



Reproducibility

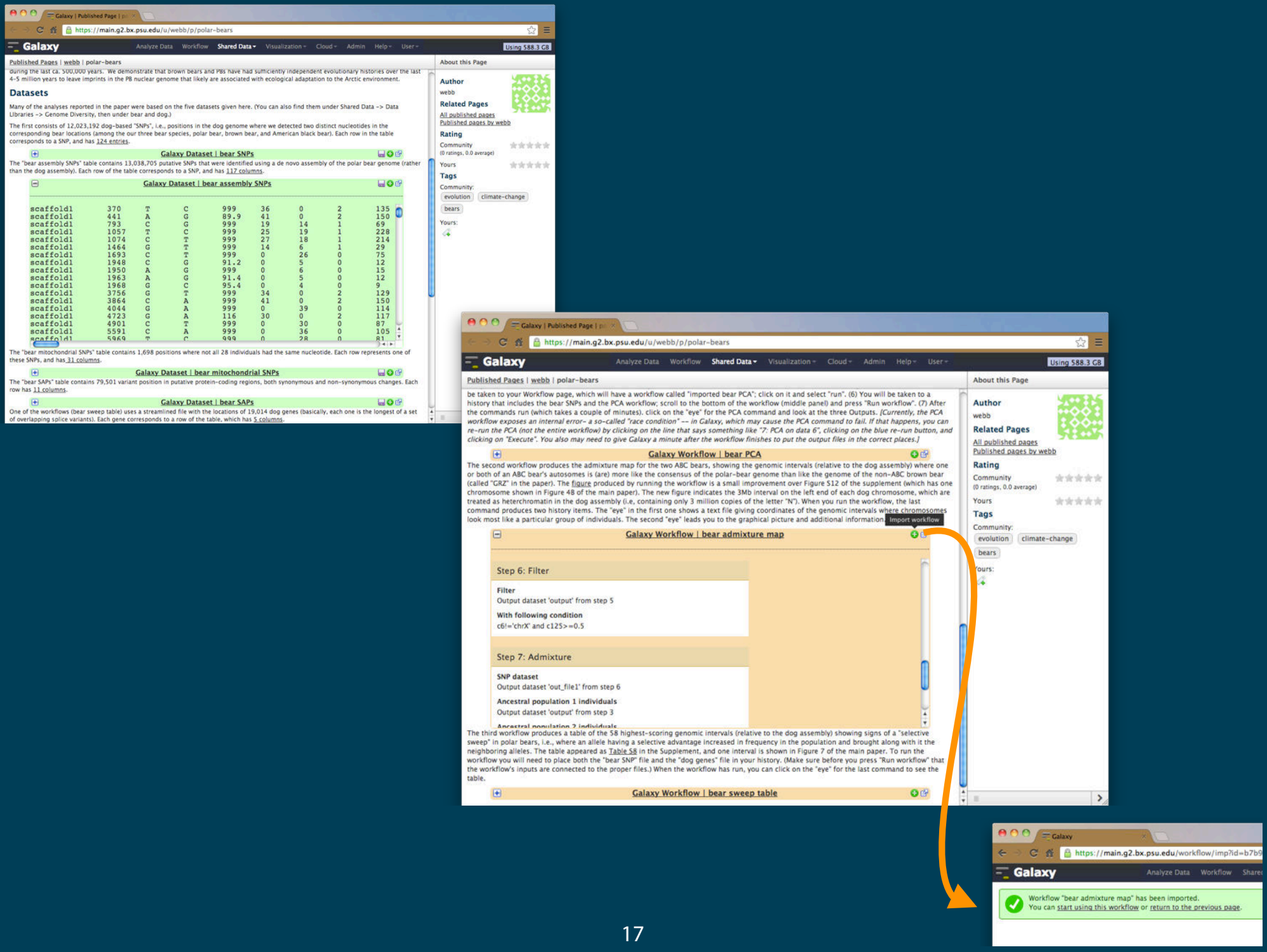


Workflows enable reuse and provide precise reproducibility

Users can add tags and annotations for additional context

Communication and Reuse

Goecks et al., *Genome Biology* , 2010


https://main.g2.bx.psu.edu/u/webb/p/polar-bears
Galaxy
Published Pages | webb | polar-bears
Datasets
Galaxy Dataset | bear SNPs
Galaxy Dataset | bear assembly SNPs
Galaxy Dataset | bear mitochondrial SNPs
Galaxy Dataset | bear SAPs
About this Page
Author
webb
Related Pages
All published pages
Published pages by webb
Rating
Tags
Community:
evolution
climate-change
bears
Yours:
Galaxy Workflow | bear PCA
Galaxy Workflow | bear admixture map
Import workflow
Step 6: Filter
Filter
Output dataset 'output' from step 5
With following condition
c6!='chrX' and c125>=0.5
Step 7: Admixture
SNP dataset
Output dataset 'out_file1' from step 6
Ancestral population 1 individuals
Output dataset 'output' from step 3
Galaxy Workflow | bear sweep table
Workflow "bear admixture map" has been imported.
You can start using this workflow or return to the previous page.

What is Galaxy?

Platform for high-throughput genomics

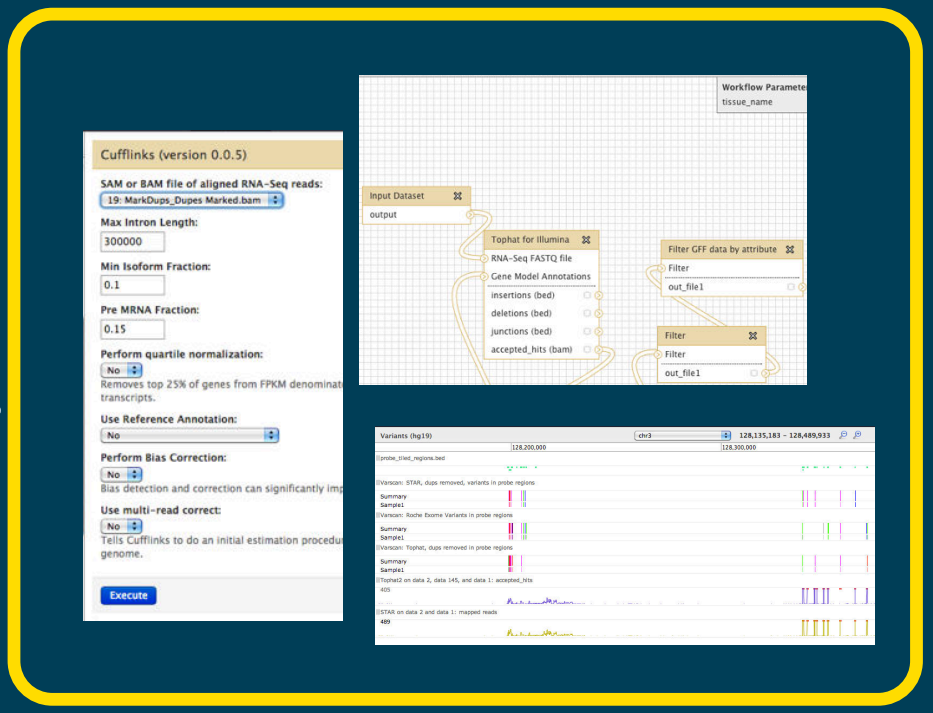
1. get and integrate public, private data
2. analyze data and create workflows
3. visualization, sharing, publication

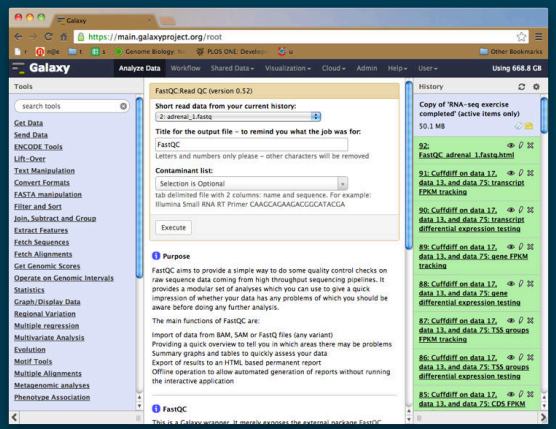
Customizable open-source software on various HPC resources

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

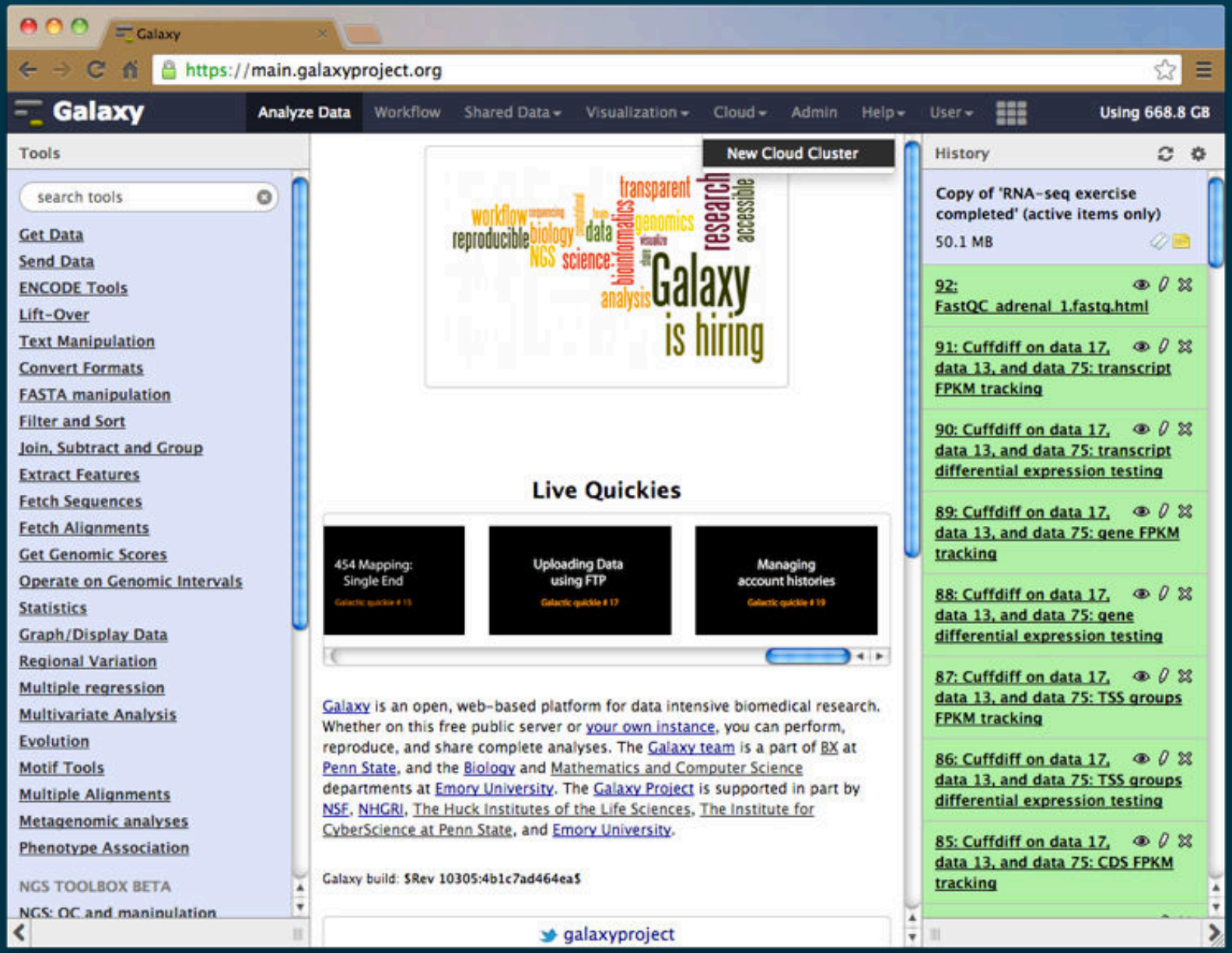
Galaxy platform

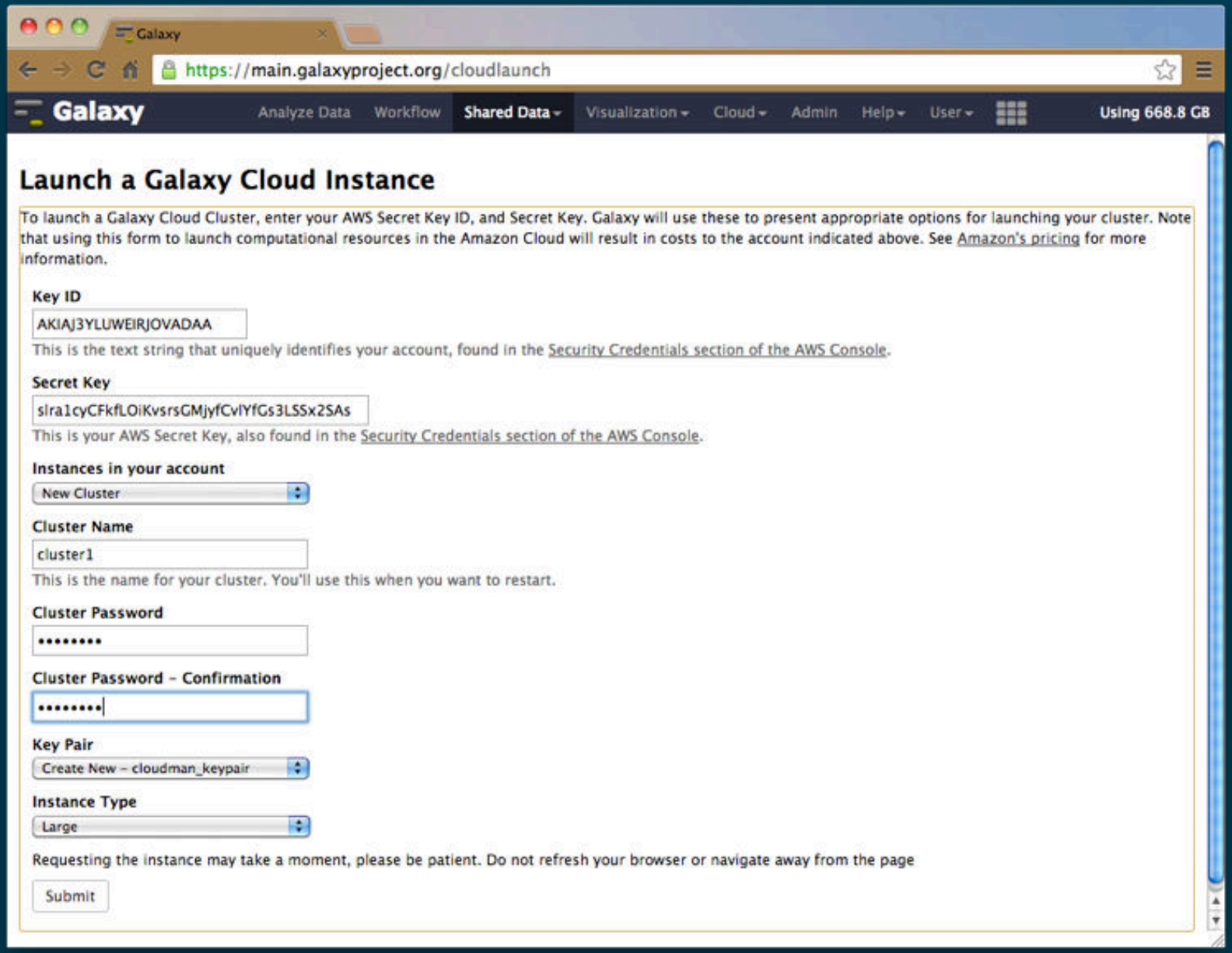
- run tools, workflows on HPC resources
- minimizes data movement
- create workflows, visualizations, pages
- share *everything*

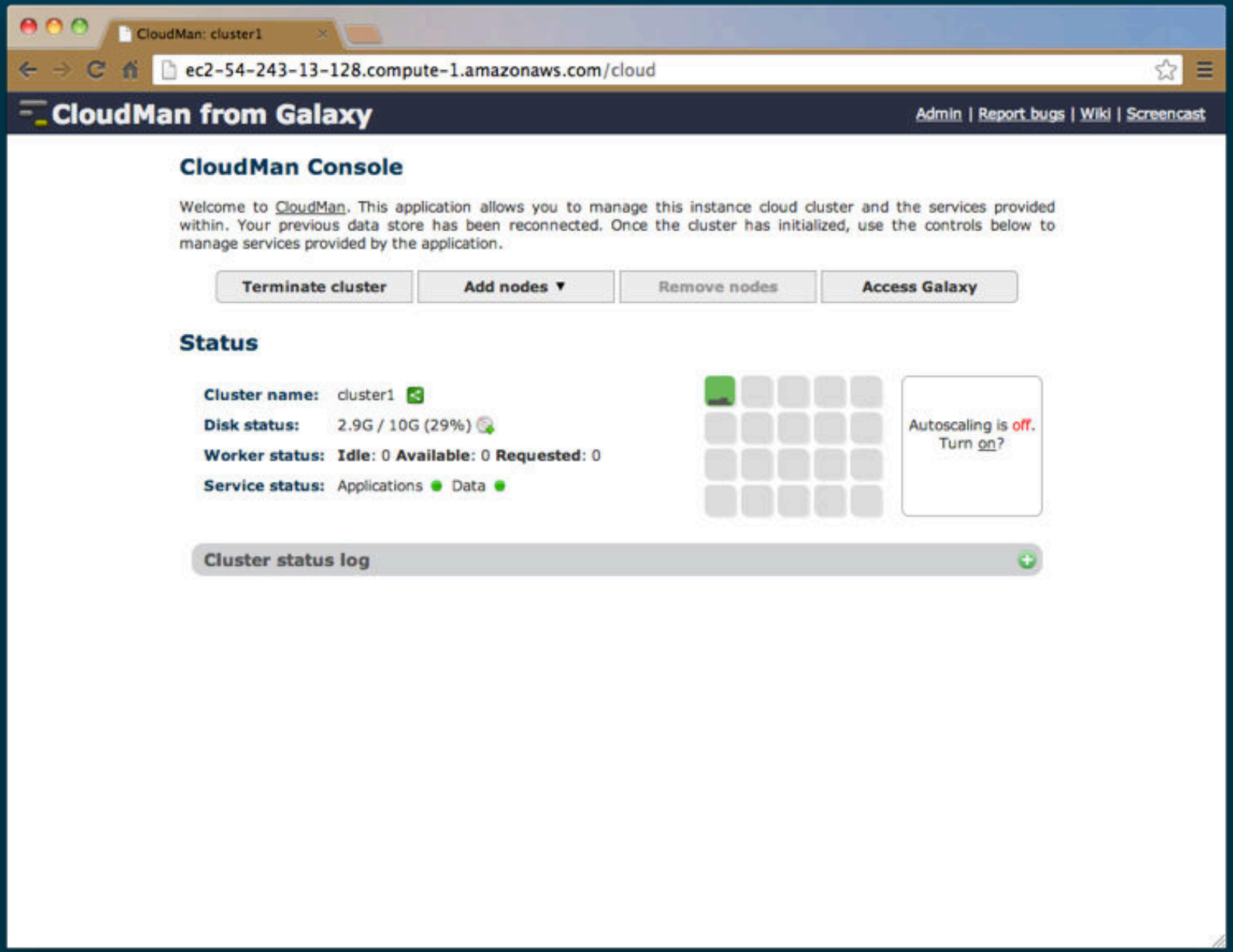


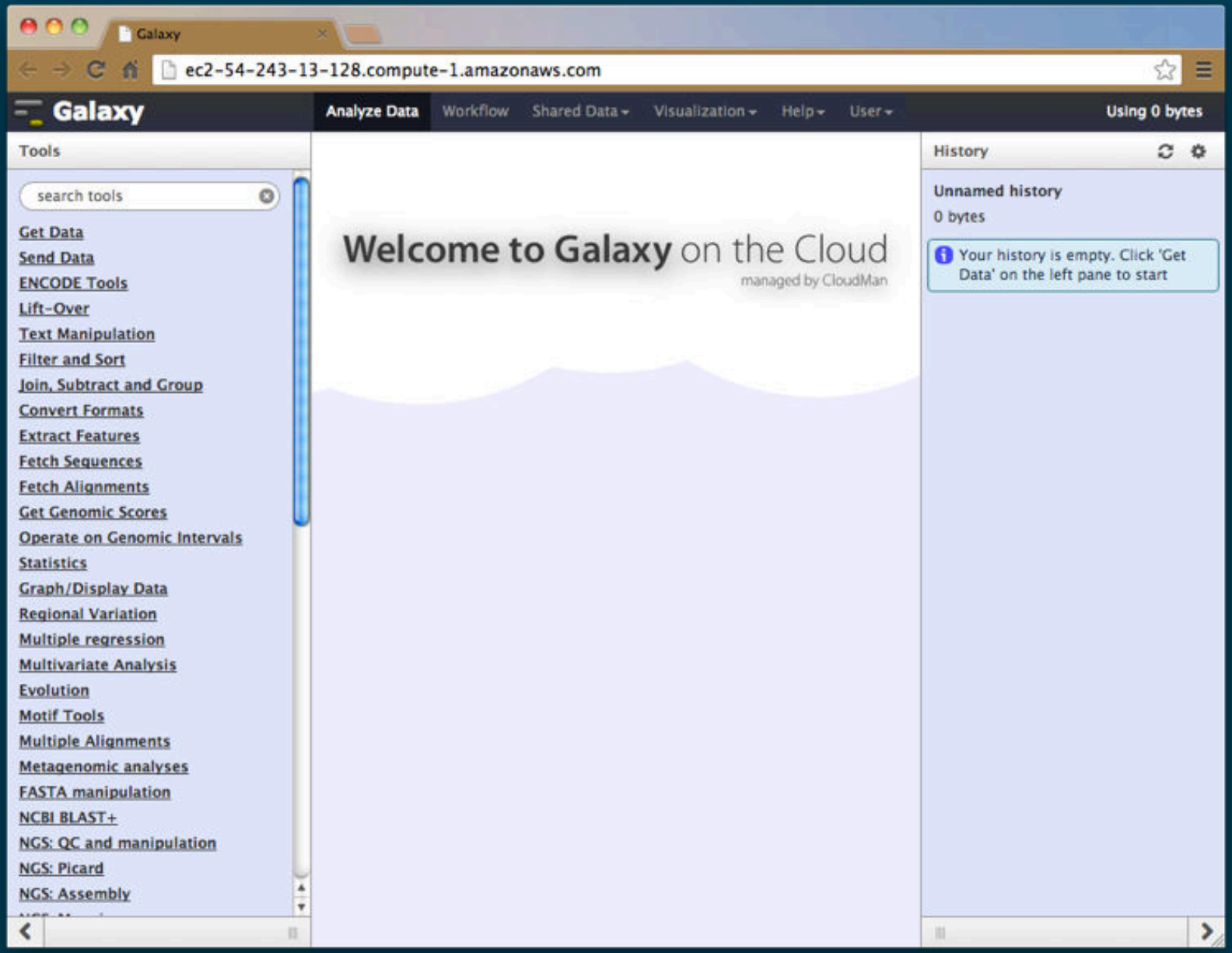


Cloud Launch




Galaxy
https://main.galaxyproject.org/cloudlaunch
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Admin
Help
User
Using 668.8 GB
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.
Terminate cluster
Add nodes
Remove nodes
Access Galaxy
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?
Cluster status log


Galaxy
ec2-54-243-13-128.compute-1.amazonaws.com
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 0 bytes
Tools
search tools
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
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
FASTA manipulation
NCBI BLAST+
NGS: QC and manipulation
NGS: Picard
NGS: Assembly
Welcome to Galaxy on the Cloud
managed by CloudMan
History
Unnamed history
0 bytes
Your history is empty. Click 'Get Data' on the left pane to start

Cloud Features

Resource configuration

- CPUs (read mapping) vs. Memory (assembly)

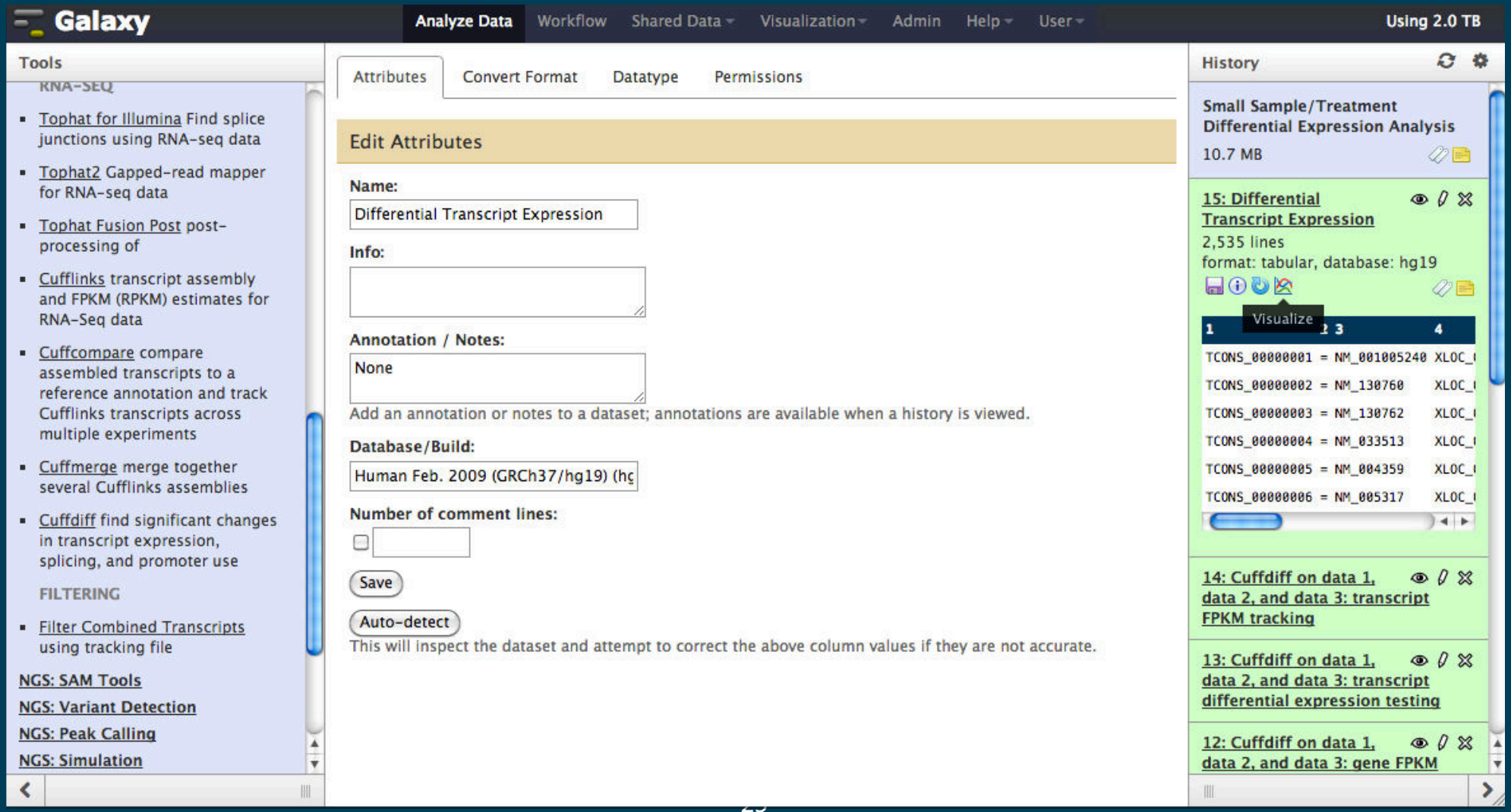
Autoscaling

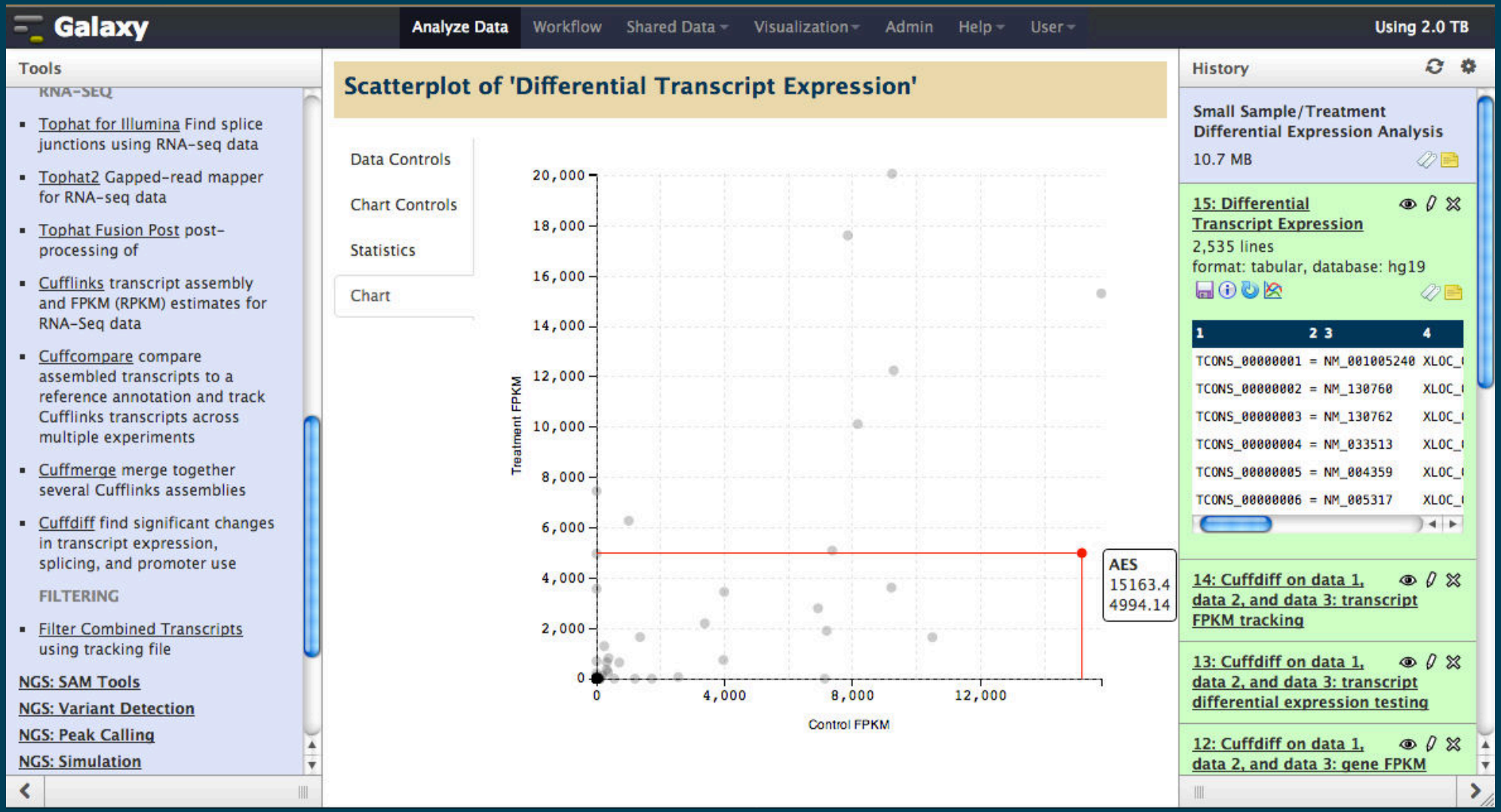
- automatically scale cluster as needed

Snapshotting

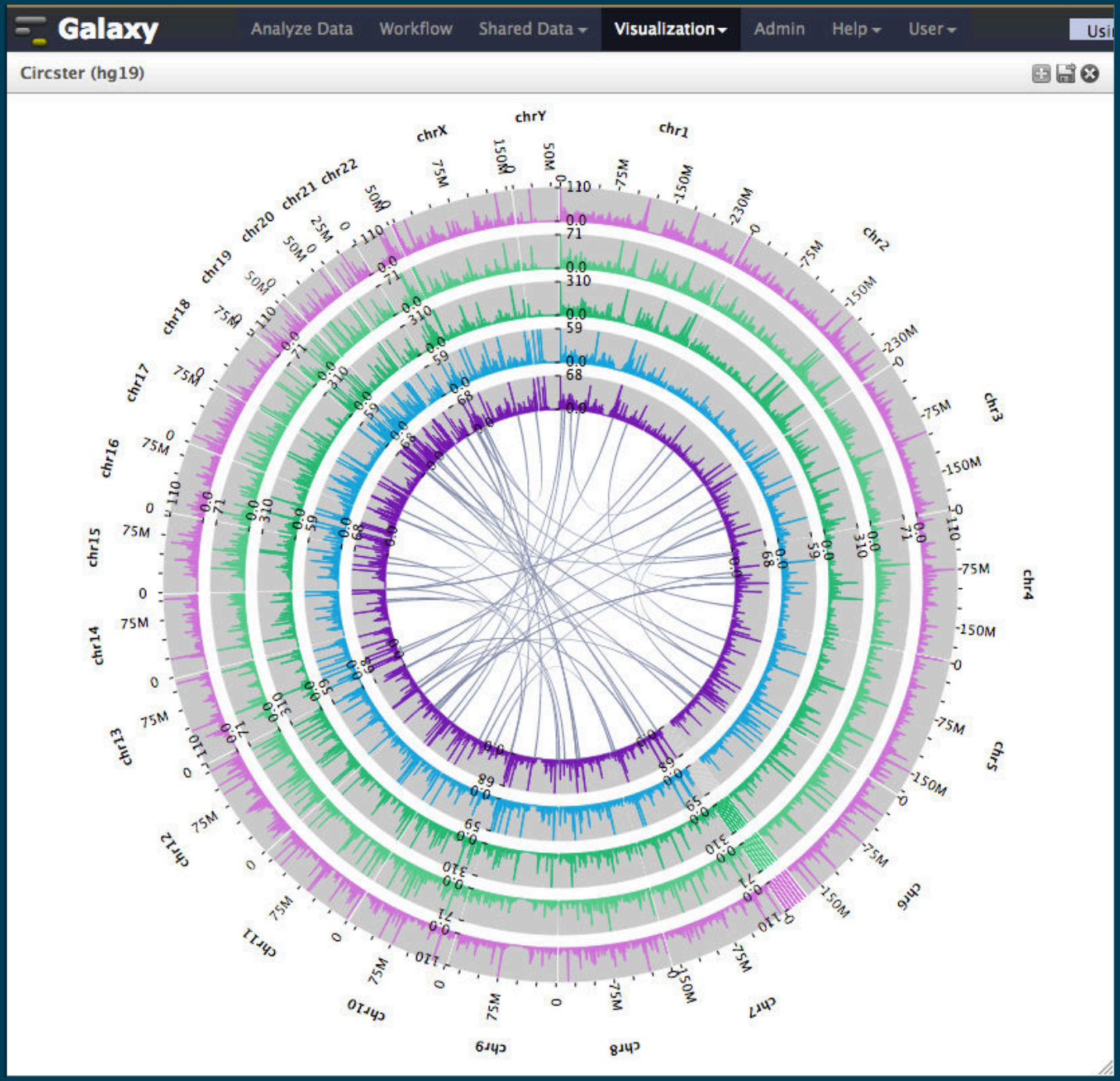
- share a complete Galaxy that others can copy and use

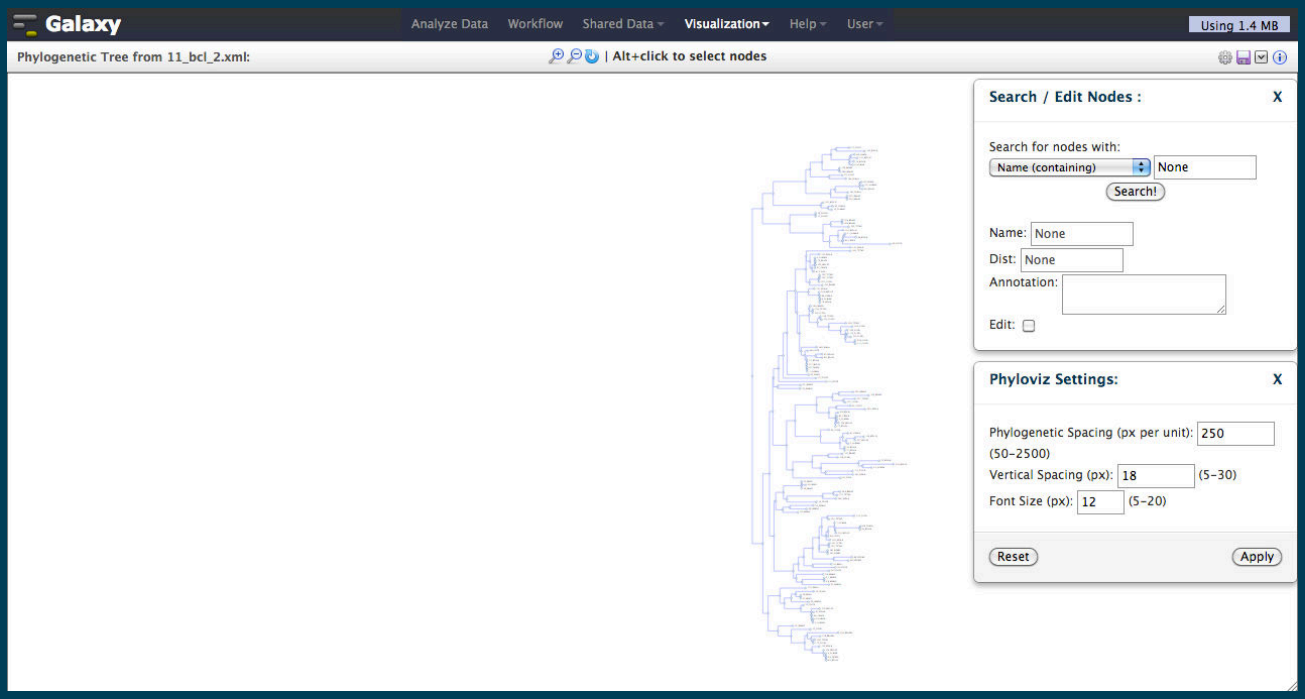
Visualization Framework

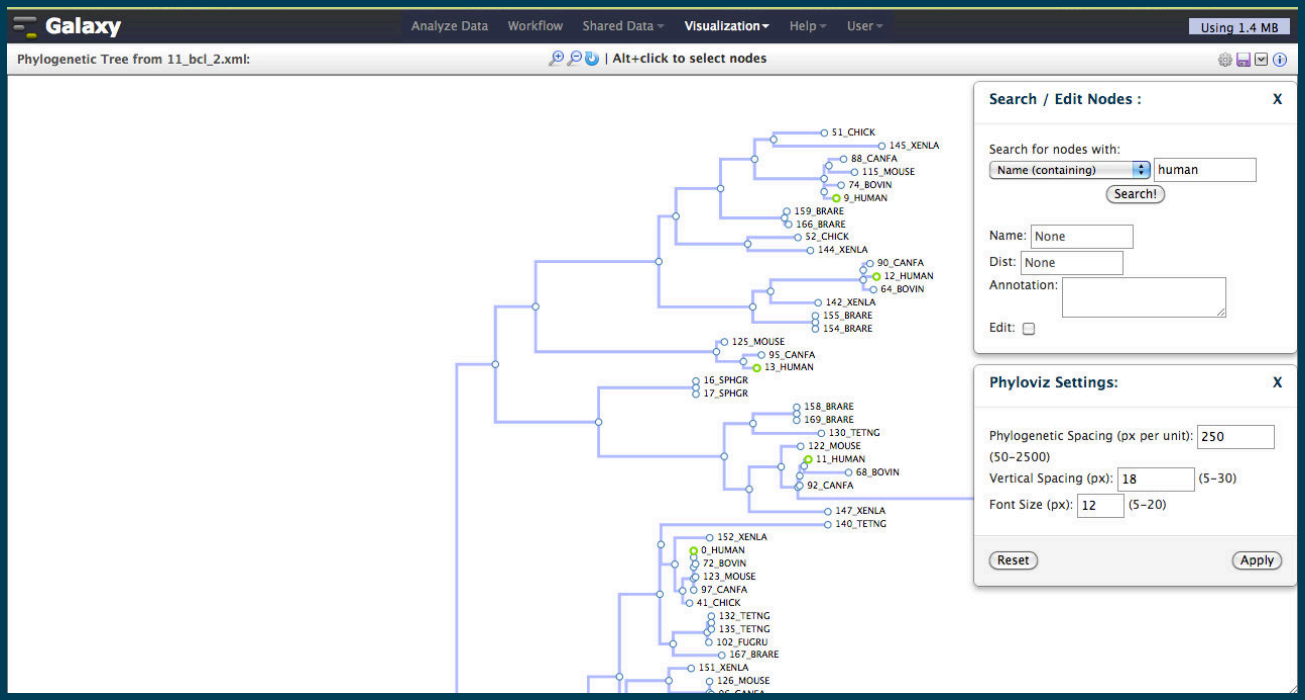



Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 2.0 TB
Tools
RNA-SEQ
Tophat for Illumina Find splice junctions using RNA-seq data
Tophat2 Gapped-read mapper for RNA-seq data
Tophat Fusion Post post-processing of
Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
Cuffmerge merge together several Cufflinks assemblies
Cuffdiff find significant changes in transcript expression, splicing, and promoter use
FILTERING
Filter Combined Transcripts using tracking file
NGS: SAM Tools
NGS: Variant Detection
NGS: Peak Calling
NGS: Simulation
Scatterplot of 'Differential Transcript Expression'
Data Controls
Chart Controls
Statistics
Chart
Treatment FPKM
Control FPKM
20,000
18,000
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
0
4,000
8,000
12,000
AES
15163.4
4994.14
History
Small Sample/Treatment Differential Expression Analysis
10.7 MB
15: Differential Transcript Expression
2,535 lines
format: tabular, database: hg19
1 2 3 4
TCONS_00000001 = NM_001005240 XLOC_
TCONS_00000002 = NM_130760 XLOC_
TCONS_00000003 = NM_130762 XLOC_
TCONS_00000004 = NM_033513 XLOC_
TCONS_00000005 = NM_004359 XLOC_
TCONS_00000006 = NM_005317 XLOC_
14: Cuffdiff on data 1, data 2, and data 3: transcript FPKM tracking
13: Cuffdiff on data 1, data 2, and data 3: transcript differential expression testing
12: Cuffdiff on data 1, data 2, and data 3: gene FPKM


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Figure 1 (hg19)
chr22
29,189,815 - 29,197,337
29,190,000
29,191,000
29,192,000
29,193,000
29,194,000
29,195,000
29,196,000
UCSC: Gencode v7
Custom Composite Track Group
Four UCSC Conservation Tracks
UCSC Main on Human: all_est (genome)
Skeletal muscle mapped RNA-seq reads
Cufflinks
Cufflinks on data 149: assembled transcripts
Cufflinks
Max Intron Length
300000
Min Isoform Fraction
0.001
Pre MRNA Fraction
0.01
Perform quartile normalization
Yes
Run on complete dataset
Run on visible region
Cufflinks - region=[chr22:29189782-29197304], parameters=[300000, 0.1, 0.05, Yes]
Cufflinks - region=[chr22:29189782-29197304], parameters=[300000, 0.001, 0.001, Yes]
Score [449-1000]
frac [0-0.64]
cov [8-253]
conf_lo [7862517-313270351]
FPKM [10511944-326534916]
conf_hi [13161372-339799481]
Run on complete dataset


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Circster (hg19)
chrX
chrY
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 1.4 MB
Phylogenetic Tree from 11_bcl_2.xml:
| Alt+click to select nodes
Search / Edit Nodes :
X
Search for nodes with:
Name (containing)
None
Search!
Name: None
Dist: None
Annotation:
Edit:
Phyloviz Settings:
X
Phylogenetic Spacing (px per unit): 250
(50-2500)
Vertical Spacing (px): 18 (5-30)
Font Size (px): 12 (5-20)
Reset
Apply


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 1.4 MB
Phylogenetic Tree from 11_bcl_2.xml:
| Alt+click to select nodes
Search / Edit Nodes :
X
Search for nodes with:
Name (containing)
human
Search!
Name: None
Dist: None
Annotation:
Edit:
Phyloviz Settings:
X
Phylogenetic Spacing (px per unit): 250
(50-2500)
Vertical Spacing (px): 18 (5-30)
Font Size (px): 12 (5-20)
Reset
Apply
51_CHICK
145_XENLA
88_CANFA
115_MOUSE
74_BOVIN
9_HUMAN
159_BRARE
166_BRARE
52_CHICK
144_XENLA
90_CANFA
12_HUMAN
64_BOVIN
142_XENLA
155_BRARE
154_BRARE
125_MOUSE
95_CANFA
13_HUMAN
16_SPHGR
17_SPHGR
158_BRARE
169_BRARE
130_TETNG
122_MOUSE
11_HUMAN
68_BOVIN
92_CANFA
147_XENLA
140_TETNG
152_XENLA
0_HUMAN
72_BOVIN
123_MOUSE
97_CANFA
41_CHICK
132_TETNG
135_TETNG
102_FUGRU
167_BRARE
151_XENLA
126_MOUSE

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

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

Topics

The High-throughput Sequencing Era

Galaxy

Cancer Genomics with Galaxy

Personalized Oncology

EMORY | WINSHIP CANCER INSTITUTE
A Cancer Center Designated by the National Cancer Institute

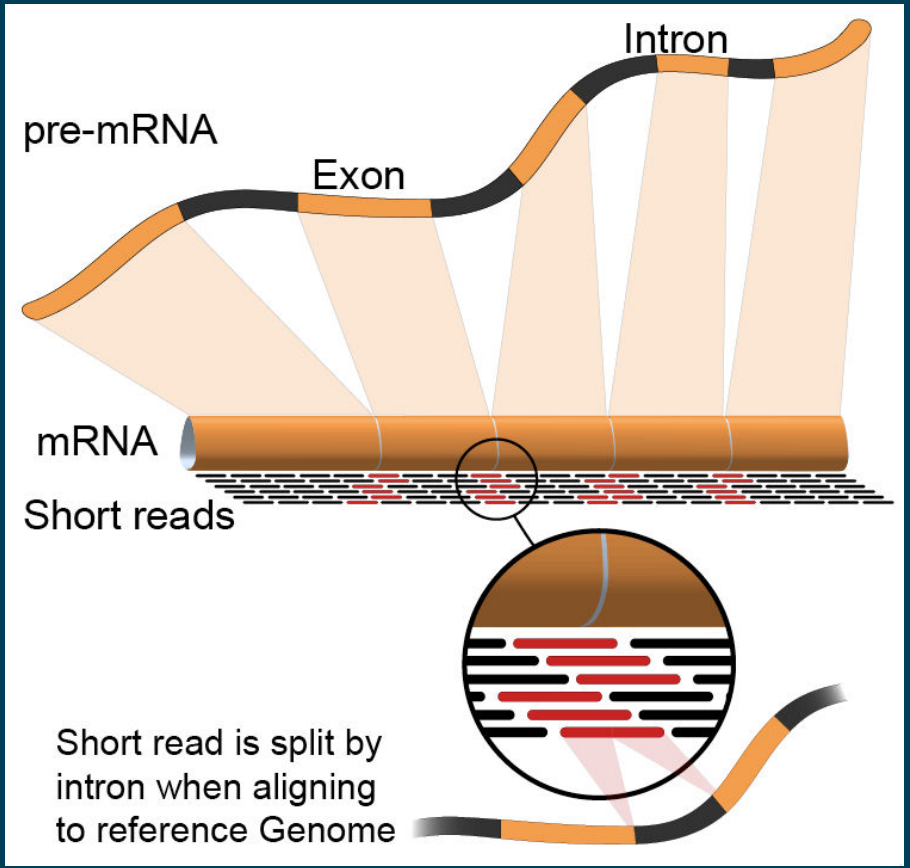
6 patients, whole transcriptome sequencing (RNA-seq) of primary tumor

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

3 pancreatic cancer cell lines

- whole transcriptome
- targeted exome (cancer genes)

Total sequencing data: ~70 GB



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

Big Questions

What drugs is a patient likely to respond to given his/her genomic profile?

- genomic profile = mutations, gene expression, copy number, ...

How best to combine private (patient) data and public data?

- match profile to potential drugs?
- match profile to cell lines, then drugs and/or high-throughput screening?

Using Galaxy for Cancer Genomics

New tools

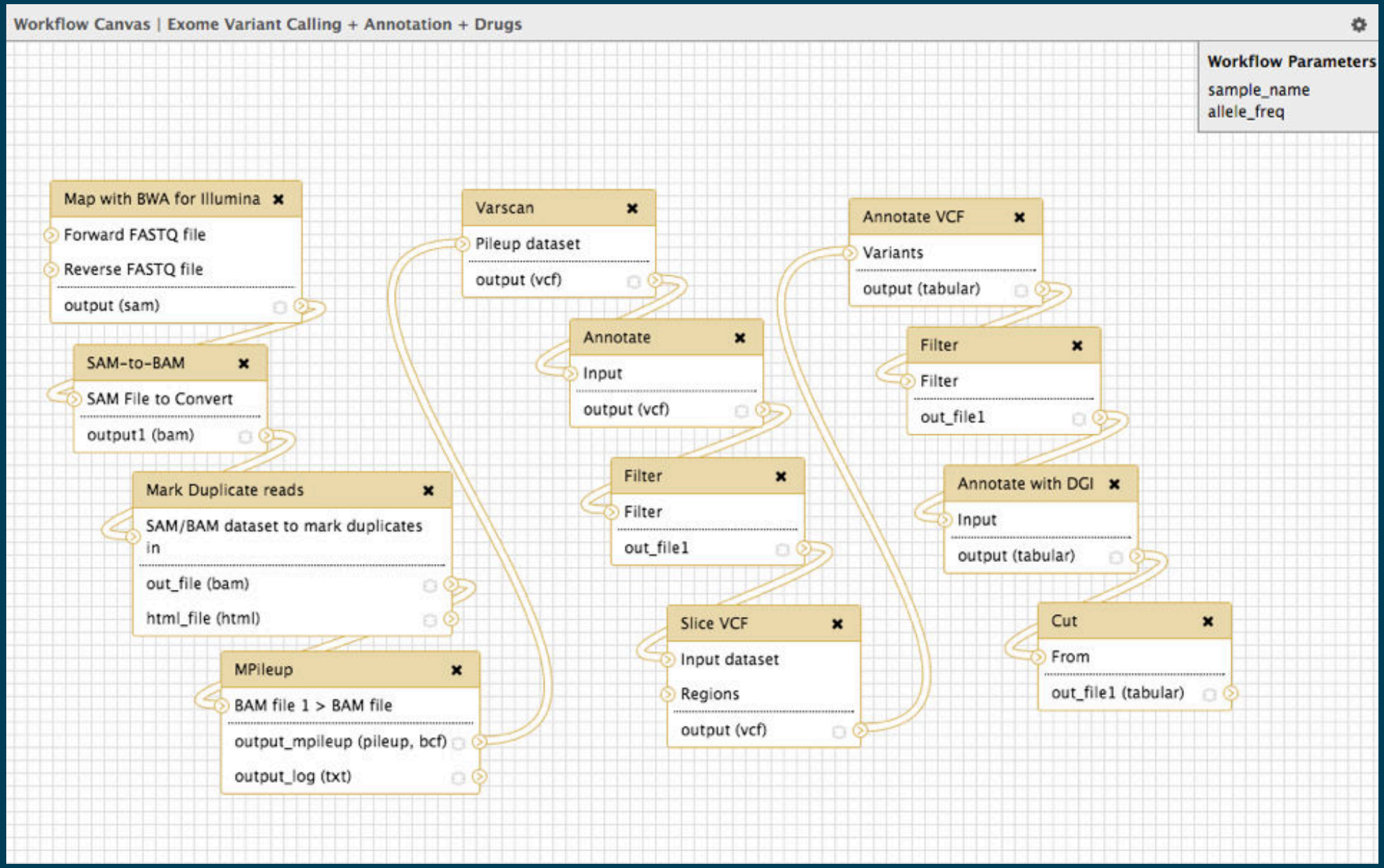
- complement existing high-throughput sequencing analysis tools

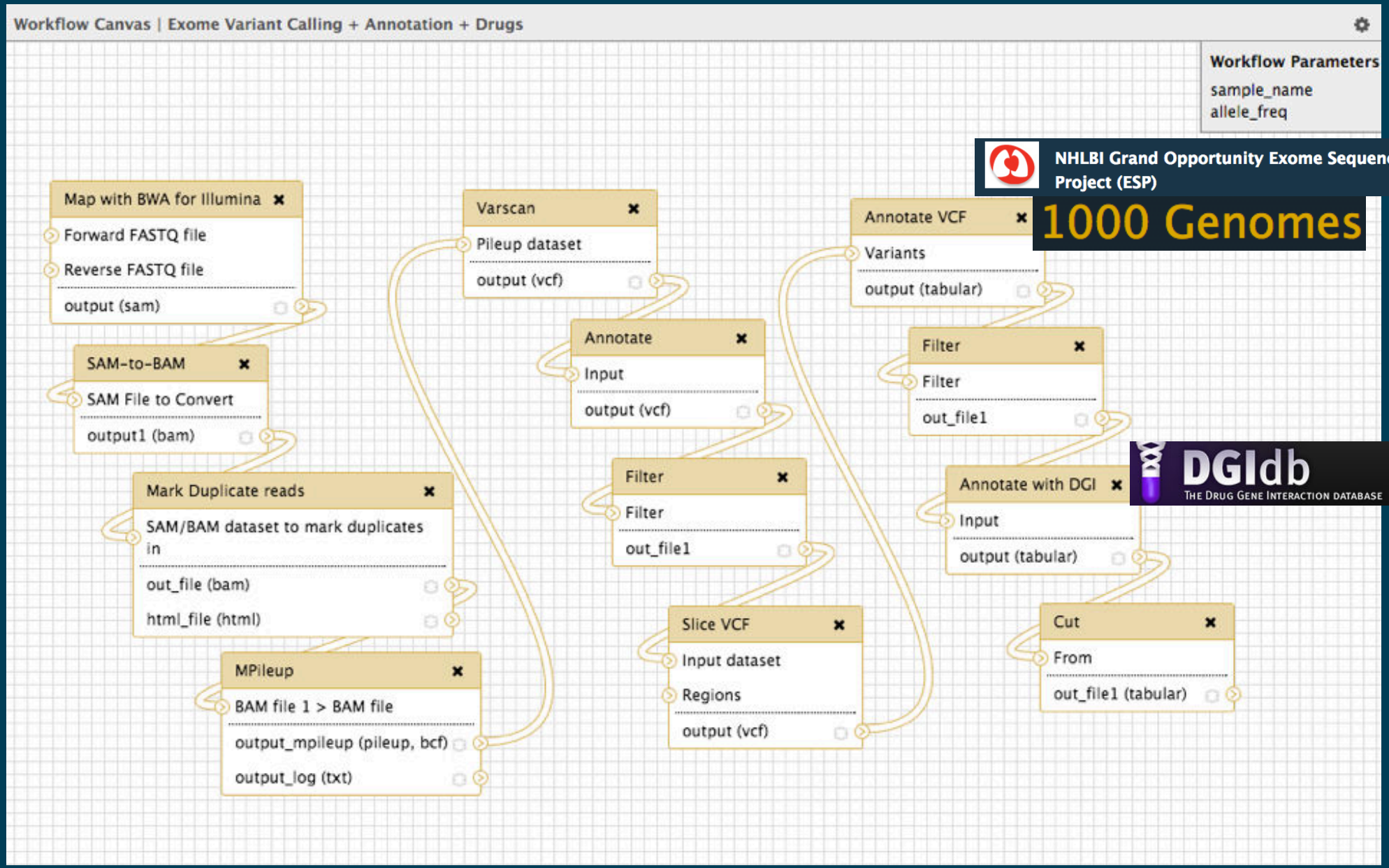
New workflows

- workflows are understandable, extendable, sharable

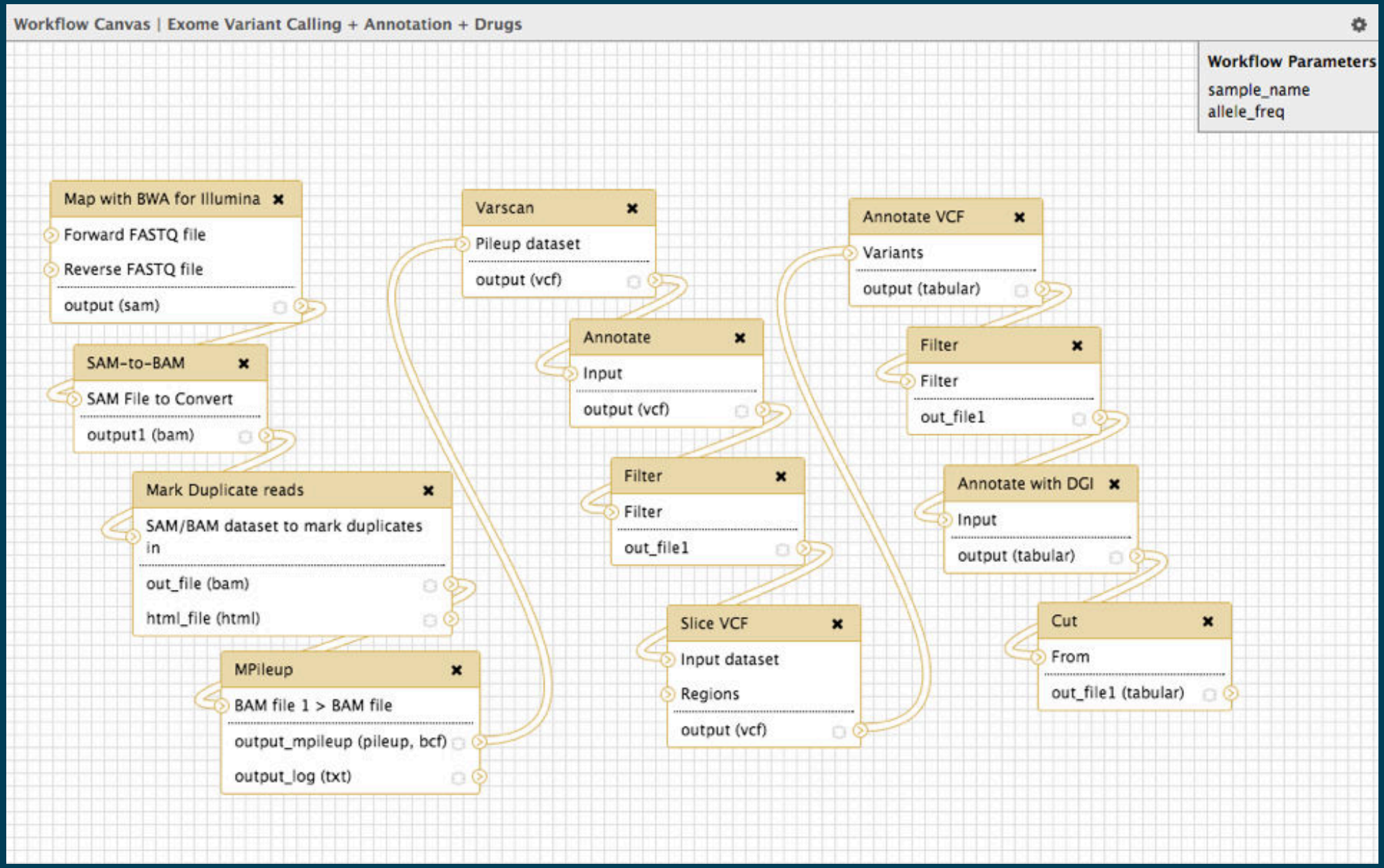
New visual analysis applications

- visualize and call variants in a Web browser


Workflow Canvas | Exome Variant Calling + Annotation + Drugs
Workflow Parameters
sample_name
allele_freq
Map with BWA for Illumina
Forward FASTQ file
Reverse FASTQ file
output (sam)
SAM-to-BAM
SAM File to Convert
output1 (bam)
Mark Duplicate reads
SAM/BAM dataset to mark duplicates in
out_file (bam)
html_file (html)
MPileup
BAM file 1 > BAM file
output_mpileup (pileup, bcf)
output_log (txt)
Varscan
Pileup dataset
output (vcf)
Annotate
Input
output (vcf)
Filter
Filter
out_file1
Slice VCF
Input dataset
Regions
output (vcf)
Annotate VCF
Variants
output (tabular)
Filter
Filter
out_file1
Annotate with DGI
Input
output (tabular)
Cut
From
out_file1 (tabular)



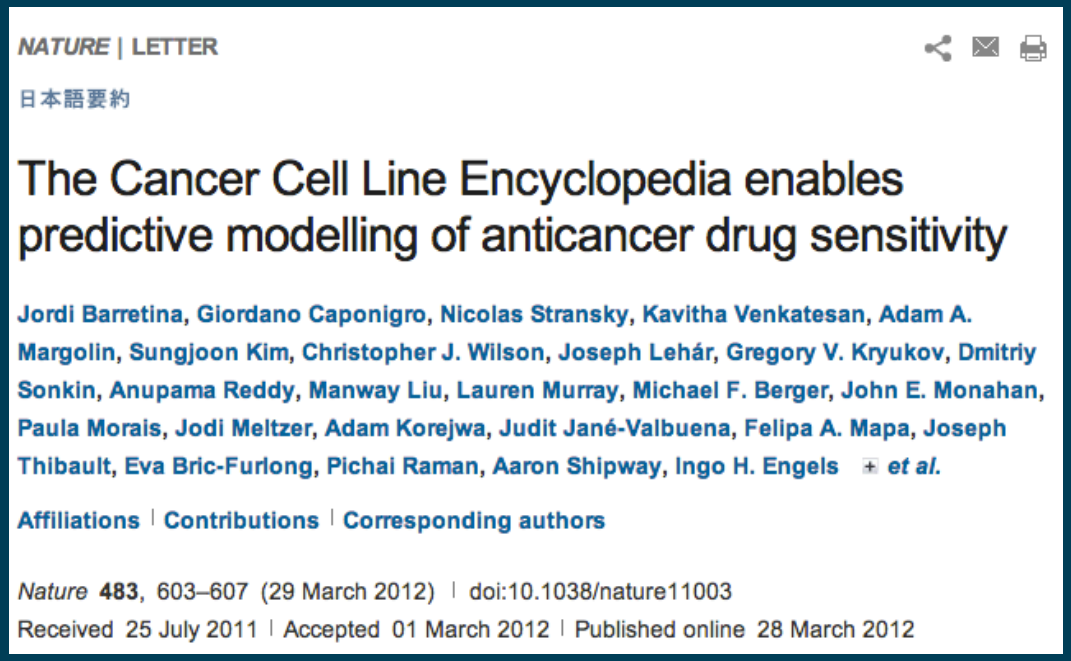
Integrates private and public data



Understandable, editable, sharable

Validation Using Public Data

- Highly targeted exome sequencing of 500+ cancer cell lines
- Drug response curves



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

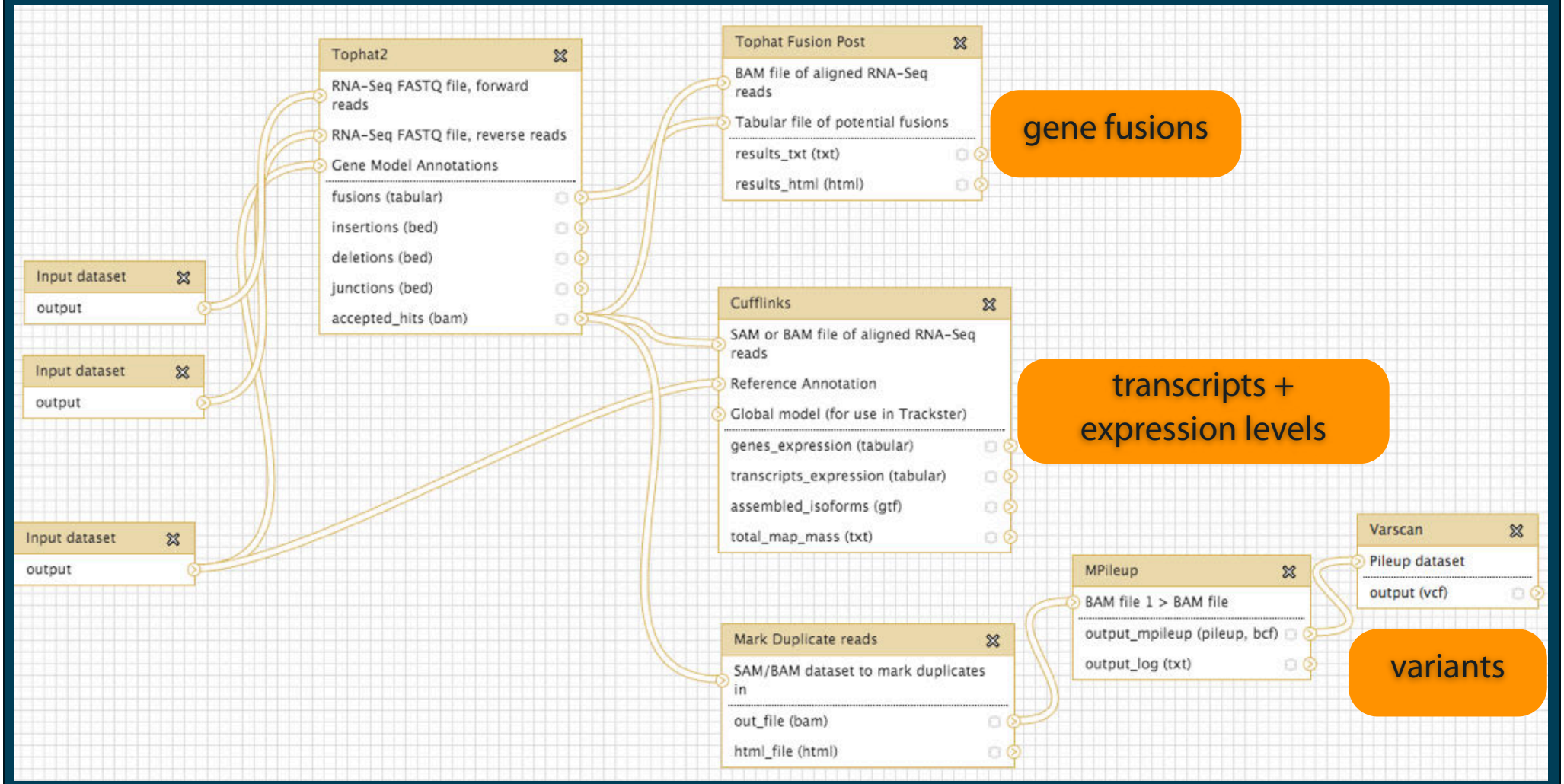
Validation Results Using MiaPaCa2 Cell Line

Able to recover all 23 mutations, includes short insertions and deletions in CCLE

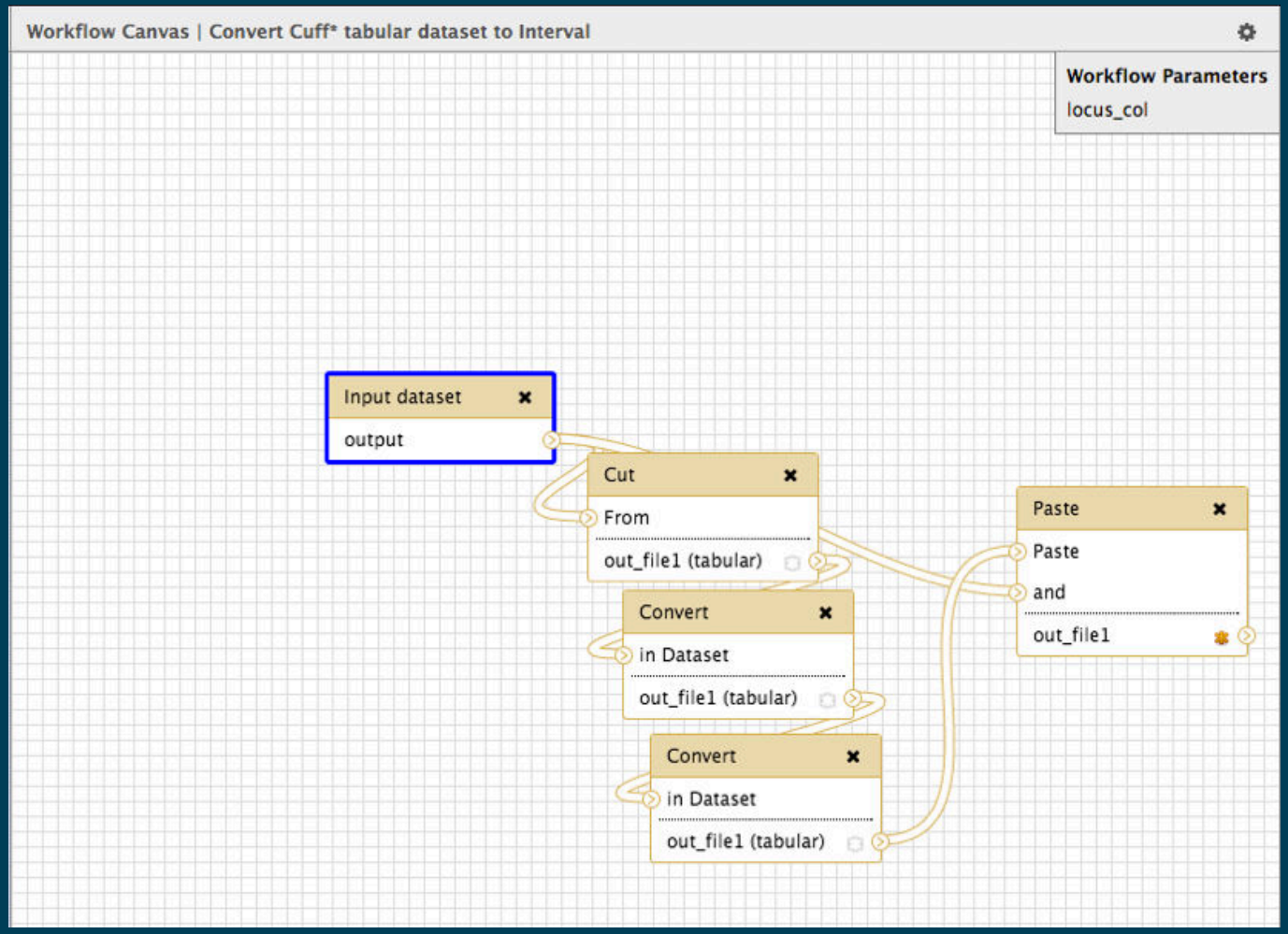
Found 16 druggable genes, leading to 98 potential drugs

Most cancer drugs are inhibitors, so gene expression is important

Single Sample Transcriptome Analysis



Cufflinks DE data to Gene Regions



Validation Results Using MiaPaCa2 Cell Line

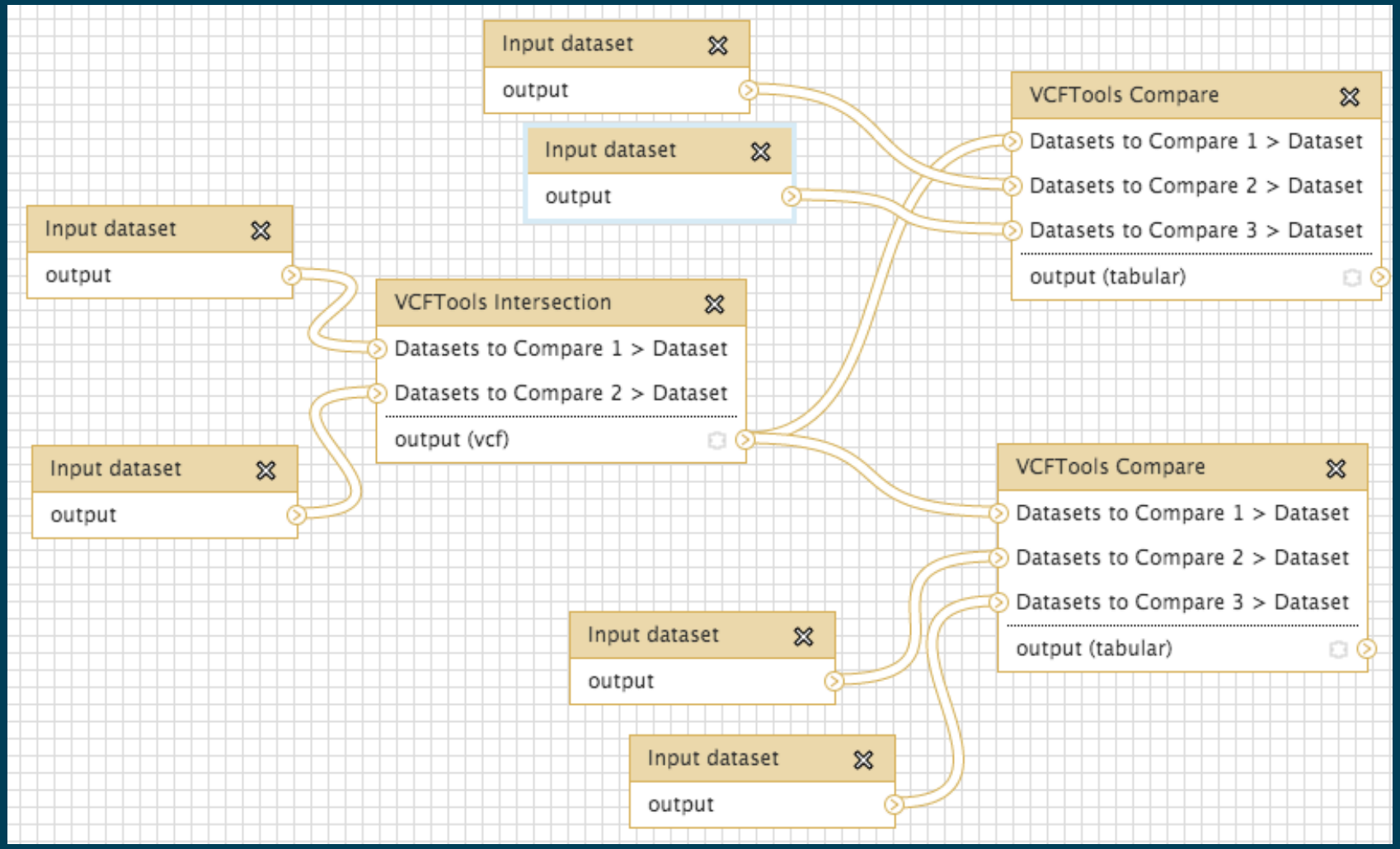
Able to recover all 23 mutations, includes short insertions and deletions in CCLE

After filtering for only mutations in highly-expressed genes:

Found ~~16~~ 6 druggable genes, leading to ~~98~~ 62 potential drugs

Matching Patients and Cell Lines

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
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.

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

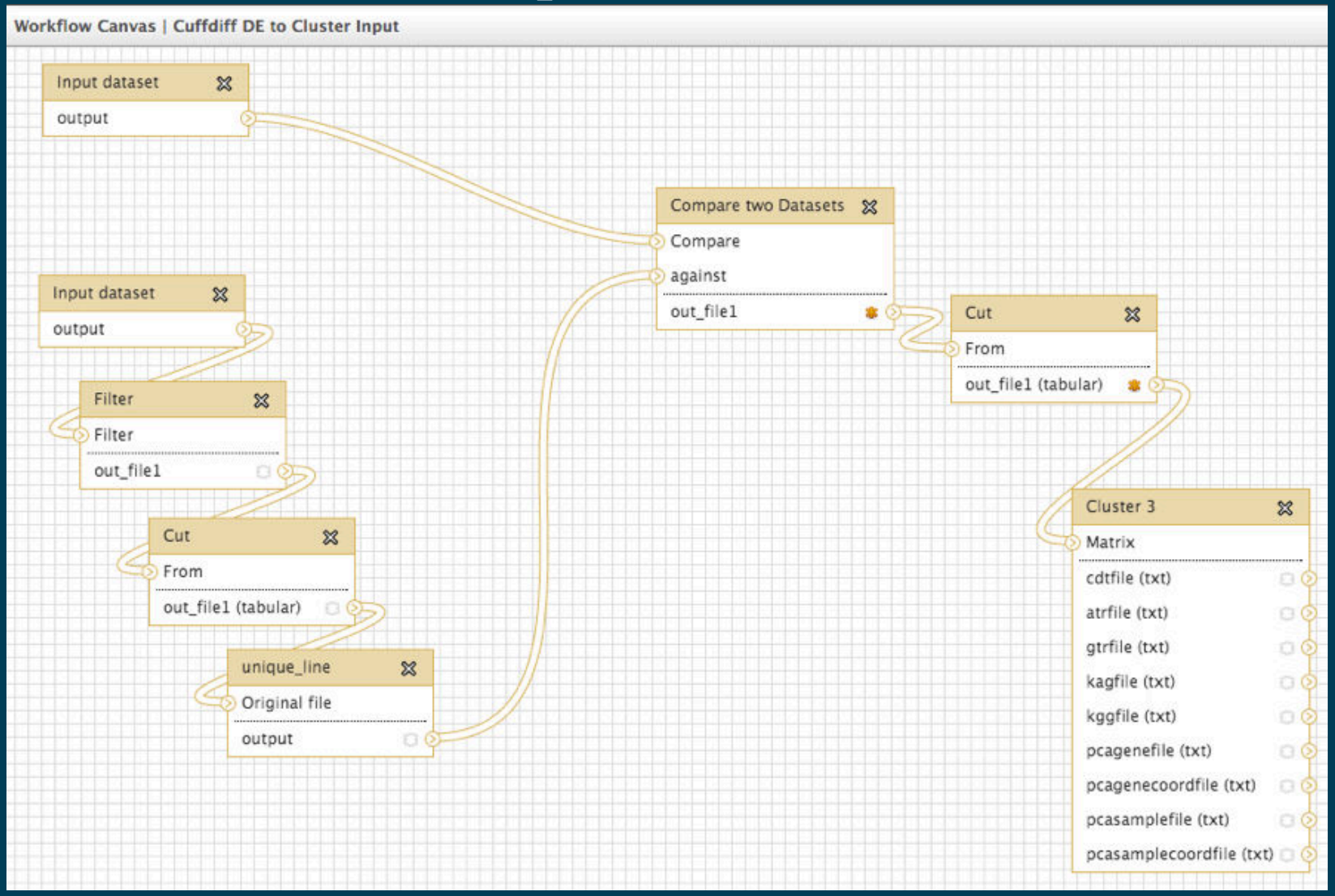
OM = OncoMap, HP = hybrid capture with probes

Patient Mutations to Predict Tumor Attributes

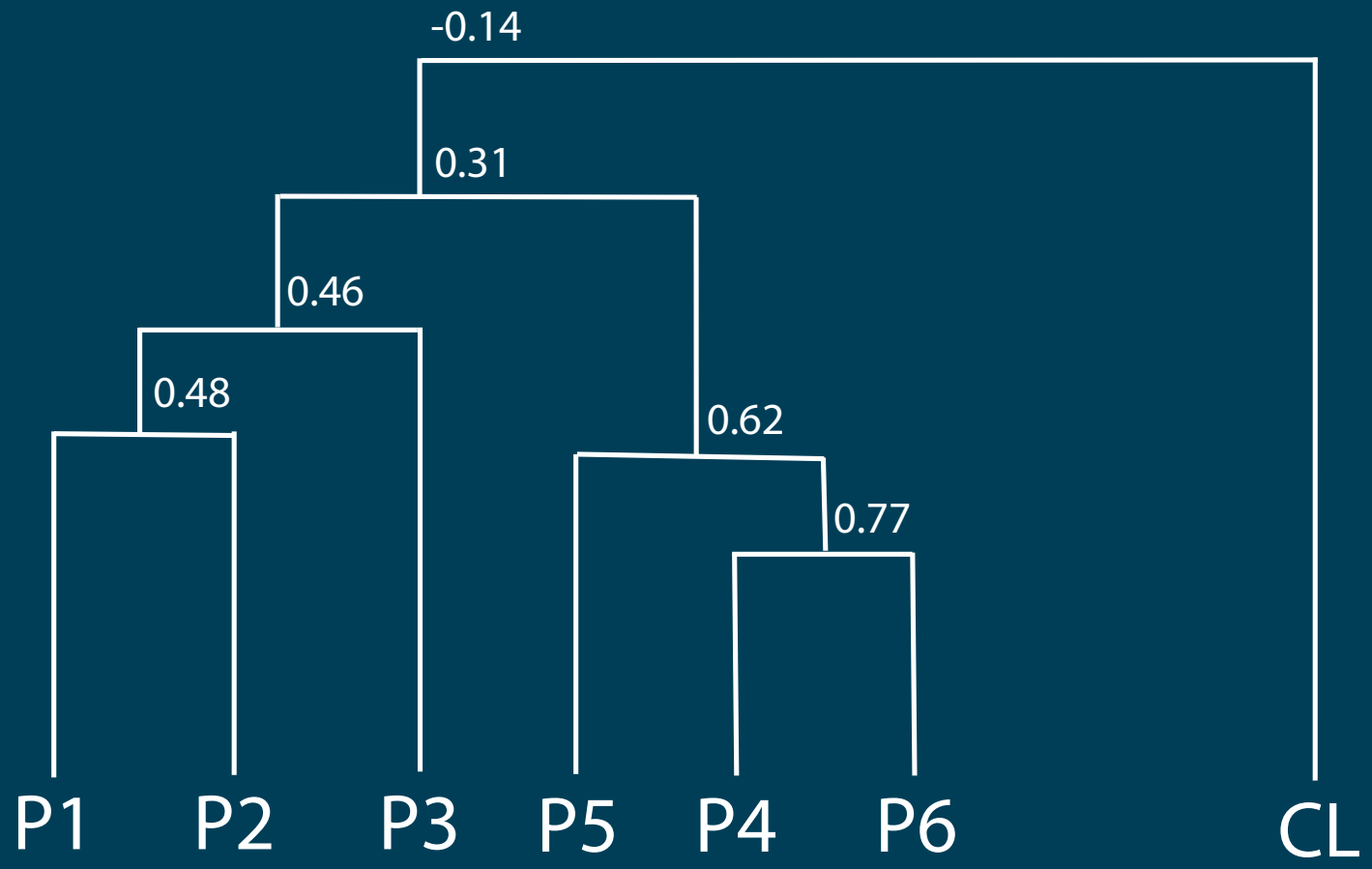
	P1	P2	P3	P4	P5	P6
OM MIA (4)	0	1	1	0	0	0
OM PC (11)	0	1	1	0	0	0
OM ALL (114)	0	2	1	1	1	1
HP MIA (84)	3	6	4	5	4	3
HP PC (1769)	16	23	19	11	23	8
HP ALL (64,669)	110	180	143	97	136	65
Tumor %	90%	90%	100%	0%?	60%	40%

OM = OncoMap, HP = hybrid capture with probes

Clustering via Differential Expression

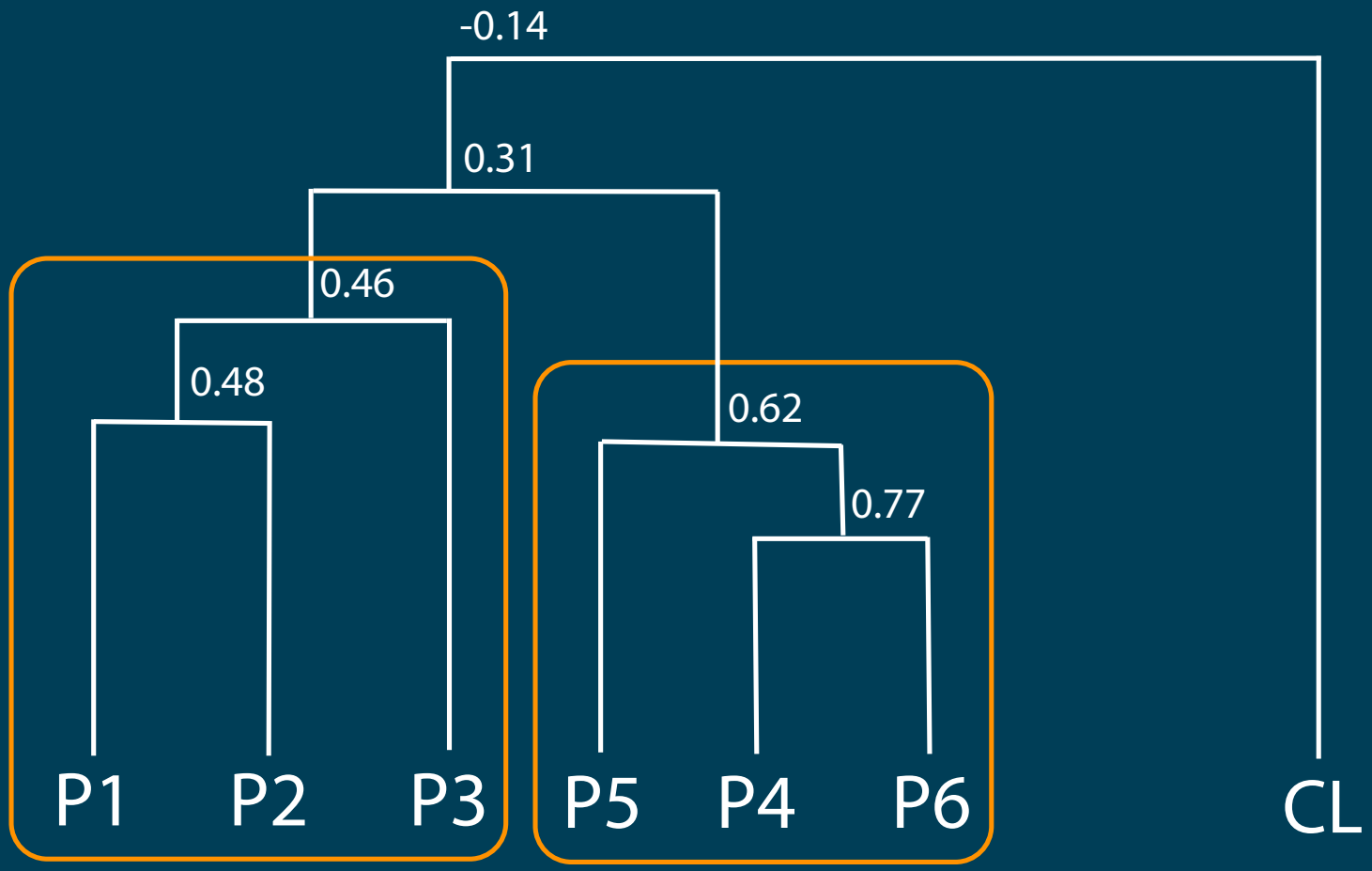


Gene Expression Clustering



Spearman Correlation

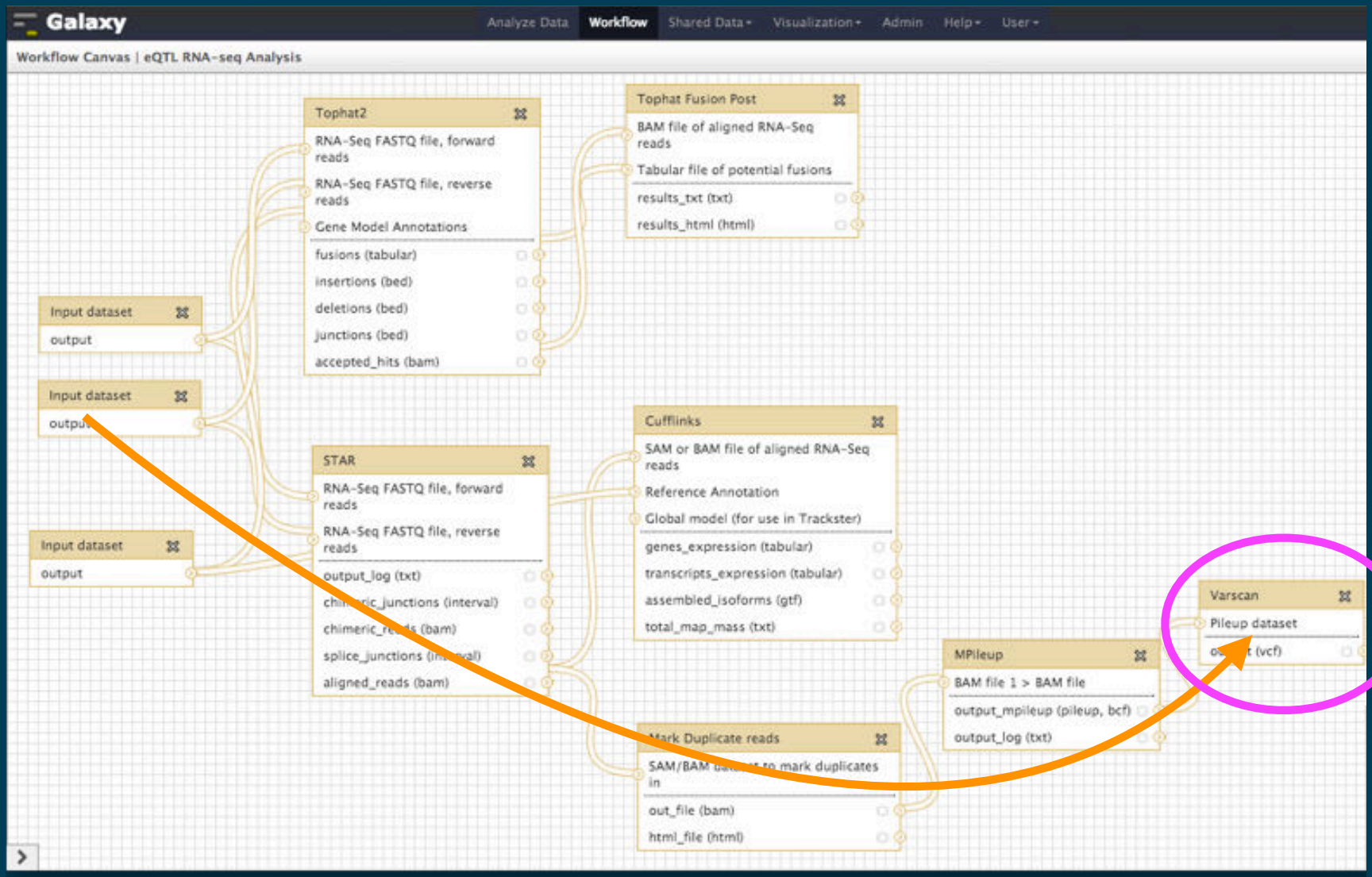
Gene Expression Clustering



Spearman Correlation

Interactive Visual Analysis

Mutation Calling from RNA-seq



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

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

Goecks et al. (2012) *Nature Biotechnology*

Concluding Thoughts

Galaxy is a very useful platform for high-throughout genomics

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

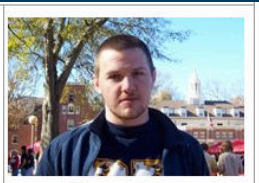
New tools, workflows, and visual analysis tools for analyzing high-throughput cancer sequencing data

- driven by personalized oncology project
- integration is key: genomic profile with mutations + gene expression, private + public data

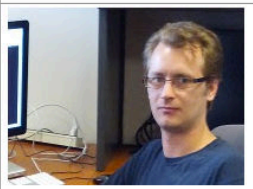
Galaxy


Enis Afgan
IRB

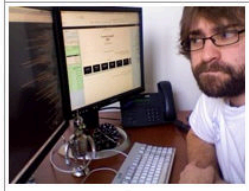

Guru Ananda
Penn State


Dannon Baker
Emory


Dan Blankenberg
Penn State

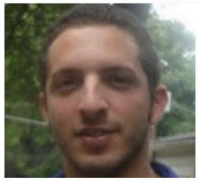

Dave Bouvier
Penn State

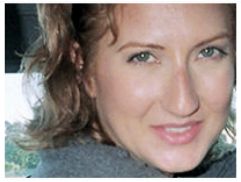

Dave Clements
Emory


Nate Coraor
Penn State


Carl Eberhard
Emory


Jeremy Goecks
Emory

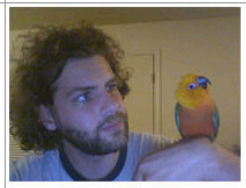

Sam Guerler
Emory


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

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

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

National Science Foundation
WHERE DISCOVERIES BEGIN

Thanks!
Questions?

Galaxy

http://galaxyproject.org

Postdoc and software engineers positions available in
Interactive Genomics Lab @ GW

http://jeremygoecks.com
jgoecks@gwu.edu