

Understanding Cancer Genomes (and Transcriptomes!) using Galaxy

Jeremy Goecks

Department of Biology
 Department of Math and Computer Science
 Emory University



EMORY
 UNIVERSITY

Roadmap

Galaxy

Analyzing Cancer Genomes and
Transcriptomes

Vision

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

What is 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

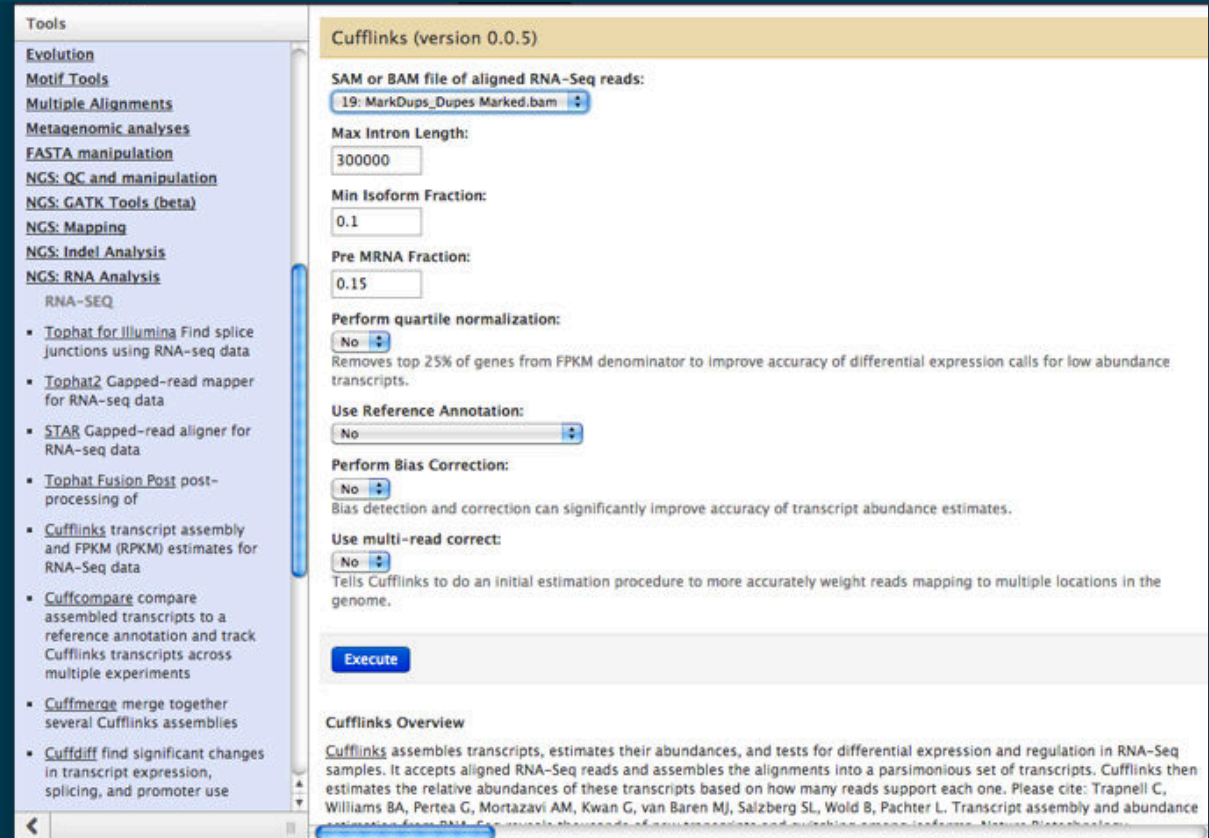
Galaxy Demo

Accessibility

All tools looks the same

No command line or programming

Easy to chain tools together into larger analyses



The screenshot displays a web-based interface for bioinformatics tools. On the left, a sidebar lists various tools under the heading 'Tools', including Evolution, Motif Tools, Multiple Alignments, Metagenomic analyses, FASTA manipulation, NGS: QC and manipulation, NGS: GATK Tools (beta), NGS: Mapping, NGS: Indel Analysis, and NGS: RNA Analysis. Under 'RNA-SEQ', several tools are listed with brief descriptions: Tophat for Illumina, Tophat2, STAR, Tophat Fusion Post, Cufflinks, Cuffcompare, Cuffmerge, and Cuffdiff. The main panel shows the configuration page for 'Cufflinks (version 0.0.5)'. It includes input fields for 'SAM or BAM file of aligned RNA-Seq reads' (set to '19: MarkDups_Dupes Marked.bam'), 'Max Intron Length' (300000), 'Min Isoform Fraction' (0.1), and 'Pre MRNA Fraction' (0.15). There are also dropdown menus for 'Perform quartile normalization' (No), 'Use Reference Annotation' (No), and 'Perform Bias Correction' (No). A 'Use multi-read correct' option is also present. An 'Execute' button is located at the bottom of the configuration area. Below the button, there is a 'Cufflinks Overview' section with a brief description of the tool's function and a citation: Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L. Transcript assembly and abundance estimation from RNA-Seq data using cufflinks. Genome Biology. 2009;10(9):R119.

Reproducibility

The screenshot displays the Galaxy workflow editor interface. The browser address bar shows the URL: <https://main.g2.bx.psu.edu/workflow/editor?id=48f2ede636d64d33#>. The page title is "Galaxy" and the current workflow is titled "Workflow Canvas | BodyMap Mapping and Assembly".

The interface includes a left-hand sidebar with a search bar and a list of tool categories and their descriptions:

- Filter and Sort**
 - Filter data on any column using simple expressions
 - Filter on ambiguities in polymorphism datasets
 - GFF
 - Filter GFF data by attribute using simple expressions
 - Filter GFF data by feature count using simple expressions
 - Filter GTF data by attribute values list
- Fetch Alignments**
 - Filter MAE by specified attributes
- Operate on Genomic Intervals**
 - Intersect the intervals of two datasets
 - Subtract the intervals of two datasets
 - Cluster the intervals of a dataset
- Graph/Display Data**
 - VCF to MAF Custom Track for display at UCSC
- Regional Variation**
 - Filter nucleotides based on quality scores
 - Fetch Indels from 3-way alignments

The main workflow canvas shows a sequence of tools connected by arrows:

- Input Dataset** (output)
- Tophat for Illumina** (RNA-Seq FASTQ file, Gene Model Annotations, insertions (bed), deletions (bed), junctions (bed), accepted_hits (bam))
- Filter GFF data by attribute** (Filter, out_file1)
- Filter** (Filter, out_file1)
- Filter** (Filter, out_file1)
- Cufflinks** (SAM or BAM file of aligned RNA-Seq reads, Reference Annotation, Global model (for use in Trackster), genes_expression (tabular), transcripts_expression (tabular), assembled_isoforms (gtf), total_map_mass (txt))
- Input dataset** (output)

The workflow parameters section on the right indicates the parameter "tissue_name". A small preview window in the bottom right corner shows a heatmap visualization.

Workflows enable reuse and provide precise reproducibility

Users can add tags and annotations for additional context

Communication and Reuse

Galaxy | Published Page | p...
https://main.g2.bx.psu.edu/u/webb/p/polar-bears

Galaxy
Analyze Data Workflow Shared Data Visualization Cloud Admin Help User

Published Pages | webb | polar-bears

Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change

Webb Miller, Stephan C. Schuster, Andreanna J. Welch, Aakrosh Ratan, Oscar C. Bedoya-Reina, Fangqing Zhao, Hie Lim Kim, Richard C. Burhans, Daniela I. Drutz, Nicola E. Wittekindt, Lynn P. Tomsho, Enrique Ibarra-Laclette, Luis Herrera-Estrella, Elizabeth Peacock, Sean Farley, George K. Sage, Karyn Rode, Martyn Obbard, Rafael Montiel, Lutz Bachmann, Ólafur Ingólfsson, Jon Aars, Thomas Mailund, Øystein Wiig, Sandra L. Talbot, and Charlotte Lindqvist

Summary of the paper

Polar bears (PBs) are superbly adapted to the extreme Arctic environment and have become emblematic of the threat to biodiversity from global climate change. Their divergence from the lower-latitude brown bear provides a textbook example of rapid evolution of distinct phenotypes. However, limited mitochondrial and nuclear DNA evidence conflicts in the timing of PB origin as well as placement of the species within versus sister to the brown bear lineage. We gathered extensive genomic sequence data from contemporary polar, brown, and American black bear samples, in addition to a 130,000- to 110,000-year old PB, to examine this problem from a genome-wide perspective. Nuclear DNA markers reflect a species tree consistent with expectation, showing polar and brown bears to be sister species. However, for the enigmatic brown bears native to Alaska's Alexander Archipelago, we estimate that not only their mitochondrial genome, but also 5-10% of their nuclear genome, is most closely related to PBs, indicating ancient admixture between the two species. Explicit admixture analyses are consistent with ancient splits among PBs, brown bears and black bears that were later followed by occasional admixture. We also provide paleodemographic estimates that suggest bear evolution has tracked key climate events, and that PB in particular experienced a prolonged and dramatic decline in its effective population size during the last ca. 500,000 years. We demonstrate that brown bears and PBs have had sufficiently independent evolutionary histories over the last 4-5 million years to leave imprints in the PB nuclear genome that likely are associated with ecological adaptation to the Arctic environment.

Datasets

Many of the analyses reported in the paper were based on the five datasets given here. (You can also find them under Shared Data -> Data Libraries -> Genome Diversity, then under bear and dog.)

The first consists of 12,023,192 dog-based "SNPs", i.e., positions in the dog genome where we detected two distinct nucleotides in the corresponding bear locations (among the our three bear species, polar bear, brown bear, and American black bear). Each row in the table corresponds to a SNP, and has [124 entries](#).

[Galaxy Dataset | bear SNPs](#)

The "bear assembly SNPs" table contains 13,038,705 putative SNPs that were identified using a de novo assembly of the polar bear genome (rather than the dog assembly). Each row of the table corresponds to a SNP, and has [117 columns](#).

[Galaxy Dataset | bear assembly SNPs](#)

The "bear mitochondrial SNPs" table contains 1,698 positions where not all 28 individuals had the same nucleotide. Each row represents one of these SNPs, and has [31 columns](#).

[Galaxy Dataset | bear mitochondrial SNPs](#)

The "bear SAPs" table contains 79,501 variant position in putative protein-coding regions, both synonymous and non-synonymous changes. Each row has [11 columns](#).

[Galaxy Dataset | bear SAPs](#)

One of the workflows (bear sweep table) uses a streamlined file with the locations of 19,014 dog genes (basically, each one is the longest of a set of overlapping splice variants). Each gene corresponds to a row of the table, which has [5 columns](#).

[Galaxy Dataset | dog genes](#)

Workflows

This page presents three "workflows" that produce results presented in the polar-bear paper. Almost all of the commands that they use are from the "Genome Diversity" tool set. (See the left panel under "Analyze Data".)

The first workflow generates the data for [Figure 4A](#) of the paper. (Those data were used to produce a more attractive PCA plot that includes other information.) The workflow needs to be applied to the "bear SNPs" data set as follows: (1) Under "Analyze Data" (in the black bar) create an empty history. (2) Under "Shared Data" -> "Published Pages", view this page. (3) Import the "bear SNPs" data set ("+" in the green circle near the right of the green bar), then click on "return to the previous page". (4) Import the "Bear PCA" workflow, and click on "start using this workflow". (5) You will be taken to your Workflow page, which will have a workflow called "Imported bear PCA"; click on it and select "run". (6) You will be taken to a history that includes the bear SNPs and the PCA workflow; scroll to the bottom of the workflow (middle panel) and press "Run workflow". (7) After the commands run (which takes a couple of minutes), click on the "eye" for the PCA command and look at the three Outputs. [Currently, the PCA workflow exposes an internal error - a so-called "race condition" -- in Galaxy, which may cause the PCA command to fail. If that happens, you can re-run the PCA (not the entire workflow) by clicking on the line that says something like "7: PCA on data 6", clicking on the blue re-run button, and clicking on "Execute". You also may need to give Galaxy a minute after the workflow finishes to put the output files in the correct places.]

[Galaxy Workflow | bear PCA](#)

The second workflow produces the admixture map for the two ABC bears, showing the genomic intervals (relative to the dog assembly) where one or both of an ABC bear's autosomes is (are) more like the consensus of the polar-bear genome than like the genome of the non-ABC brown bear (called "GRZ" in the paper). The [figure](#) produced by running the workflow is a small improvement over [Figure S12](#) of the supplement (which has one chromosome shown in [Figure 4B](#) of the main paper). The new figure indicates the 3Mb interval on the left end of each dog chromosome, which are treated as heterochromatin in the dog assembly (i.e., containing only 3 million copies of the letter "N"). When you run the workflow, the last command produces two history items. The "eye" in the first one shows a text file giving coordinates of the genomic intervals where chromosomes look most like a particular group of individuals. The second "eye" leads you to the graphical picture and additional information.

[Galaxy Workflow | bear admixture map](#)

The third workflow produces a table of the 58 highest-scoring genomic intervals (relative to the dog assembly) showing signs of a "selective sweep" in polar bears, i.e., where an allele having a selective advantage increased in frequency in the population and brought along with it the neighboring alleles. The table appeared as [Table S8](#) in the Supplement, and one interval is shown in [Figure 7](#) of the main paper. To run the workflow you will need to place both the "bear SNP" file and the "dog genes" file in your history. (Make sure before you press "Run workflow" that the workflow's inputs are connected to the proper files.) When the workflow has run, you can click on the "eye" for the last command to see the table.

[Galaxy Workflow | bear sweep table](#)

Galaxy | Published Page | p...
 https://main.g2.bx.psu.edu/u/webb/p/polar-bears

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Admin Help User Using 588.3 GB

Published Pages | webb | polar-bears

During the last ca. 300,000 years, we demonstrate that brown bears and PBS have had sufficiently independent evolutionary histories over the last 4-5 million years to leave imprints in the P8 nuclear genome that likely are associated with ecological adaptation to the Arctic environment.

Datasets

Many of the analyses reported in the paper were based on the five datasets given here. (You can also find them under Shared Data -> Data Libraries -> Genome Diversity, then under 'bear and dog'.)

The first consists of 12,023,192 dog-based "SNPs", i.e., positions in the dog genome where we detected two distinct nucleotides in the corresponding bear locations (among the our three bear species, polar bear, brown bear, and American black bear). Each row in the table corresponds to a SNP, and has 124 entries.

Galaxy Dataset | bear SNPs

The "bear assembly SNPs" table contains 13,038,705 putative SNPs that were identified using a de novo assembly of the polar bear genome (rather than the dog assembly). Each row of the table corresponds to a SNP, and has 117 columns.

Galaxy Dataset bear assembly SNPs										
scaffold1	370	T	C	999	36	0	2	135		
scaffold1	441	A	G	89.9	41	0	2	150		
scaffold1	793	C	G	999	19	14	1	69		
scaffold1	1057	T	C	999	25	19	1	228		
scaffold1	1074	C	T	999	27	18	1	214		
scaffold1	1464	G	T	999	14	6	1	29		
scaffold1	1693	C	T	999	0	26	0	75		
scaffold1	1948	C	G	91.2	0	5	0	12		
scaffold1	1950	A	G	999	0	6	0	15		
scaffold1	1963	A	G	91.4	0	5	0	12		
scaffold1	1968	G	C	95.4	0	4	0	9		
scaffold1	3756	G	T	999	34	0	2	129		
scaffold1	3864	C	A	999	41	0	2	150		
scaffold1	4044	G	A	999	0	39	0	114		
scaffold1	4723	G	A	116	30	0	2	117		
scaffold1	4901	C	T	999	0	30	0	87		
scaffold1	5591	C	A	999	0	36	0	105		
scaffold1	5969	T	C	999	0	28	0	81		

The "bear mitochondrial SNPs" table contains 1,698 positions where not all 28 individuals had the same nucleotide. Each row represents one of these SNPs, and has 31 columns.

Galaxy Dataset | bear mitochondrial SNPs

The "bear SAPs" table contains 79,501 variant position in putative protein-coding regions, both synonymous and non-synonymous changes. Each row has 11 columns.

Galaxy Dataset | bear SAPs

One of the workflows (bear sweep table) uses a streamlined file with the locations of 19,014 dog genes (basically, each one is the longest of a set of overlapping splice variants). Each gene corresponds to a row of the table, which has 5 columns.

Galaxy | Published Page | p...
 https://main.g2.bx.psu.edu/u/webb/p/polar-bears

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Admin Help User Using 588.3 GB

Published Pages | webb | polar-bears

be taken to your Workflow page, which will have a workflow called "Imported bear PCA"; click on it and select "run". (6) You will be taken to a history that includes the bear SNPs and the PCA workflow; scroll to the bottom of the workflow (middle panel) and press "Run workflow". (7) After the commands run (which takes a couple of minutes), click on the "eye" for the PCA command and look at the three Outputs. (Currently, the PCA workflow exposes an internal error—a so-called "race condition"—in Galaxy, which may cause the PCA command to fail. If that happens, you can re-run the PCA (not the entire workflow) by clicking on the line that says something like "?: PCA on data 6", clicking on the blue re-run button, and clicking on "Execute". You also may need to give Galaxy a minute after the workflow finishes to put the output files in the correct places.)

Galaxy Workflow | bear PCA

The second workflow produces the admixture map for the two ABC bears, showing the genomic intervals (relative to the dog assembly) where one or both of an ABC bear's autosomes is (are) more like the consensus of the polar-bear genome than like the genome of the non-ABC brown bear (called "GRZ" in the paper). The figure produced by running the workflow is a small improvement over Figure S12 of the supplement (which has one chromosome shown in Figure 4B of the main paper). The new figure indicates the 3Mb interval on the left end of each dog chromosome, which are treated as heterochromatin in the dog assembly (i.e. containing only 3 million copies of the letter "N"). When you run the workflow, the last command produces two history items. The "eye" in the first one shows a text file giving coordinates of the genomic intervals where chromosomes look most like a particular group of individuals. The second "eye" leads you to the graphical picture and additional information. [Import workflow](#)

Galaxy Workflow | bear admixture map

Step 6: Filter

Filter

Output dataset 'output' from step 5

With following condition

c6!="chrX" and c12>=0.5

Step 7: Admixture

SNP dataset

Output dataset 'out_file1' from step 6

Ancestral population 1 individuals

Output dataset 'output' from step 3

Ancestral population 2 individuals

The third workflow produces a table of the 58 highest-scoring genomic intervals (relative to the dog assembly) showing signs of a "selective sweep" in polar bears, i.e., where an allele having a selective advantage increased in frequency in the population and brought along with it the neighboring alleles. The table appeared as Table S8 in the Supplement, and one interval is shown in Figure 7 of the main paper. To run the workflow you will need to place both the "bear SNP" file and the "dog genes" file in your history. (Make sure before you press "Run workflow" that the workflow's inputs are connected to the proper files.) When the workflow has run, you can click on the "eye" for the last command to see the table.

Galaxy Workflow | bear sweep table

Galaxy

https://main.g2.bx.psu.edu/workflow/imp?id=b7b9

Galaxy Analyze Data Workflow Shares

Workflow "bear admixture map" has been imported.
 You can start using this workflow or return to the previous page.

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

- ✦ complement existing transcriptome analysis tools

New workflows

- ✦ workflows are understandable and extendable

New visual analysis applications

- ✦ visualize and call variants in a Web browser

Varscan (version 0.1)

Pileup dataset:

21: MPileup on data 3, data 6, and others

Analysis type:

single nucleotide variation

Minimum read depth:

8

Minimum depth at a position to make a call

Minimum supporting reads:

2

Minimum supporting reads at a position to make a call

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

sample_names:

Separate sample names by comma; leave blank to use default s

Execute

Tophat Fusion Post (version 0.1)

BAM file of aligned RNA-Seq reads:

7: CL Tophat2 on data 2, data 145, and data 1: accepted_hits

Tabular file of potential fusions:

182: Filter on data 17

Num Fusion Reads:

3

Fusions with at least this many supporting reads will be reported.

Num Fusion Pairs:

2

Fusions with at least this many supporting pairs will be reported.

Num Fusion Reads + Pairs:

0

The sum of supporting reads and pairs is at least this number for

Fusion Read Mismatches:

2

Reads support fusions if they map across fusion with at most this

Reads that map to more than this many places will be ignored

2

Is your data from humans?:

Yes

Execute

http://tophat.cbcb.umd.edu/fusion_index.html

<http://varscan.sourceforge.net/>

VCFTools Subset (version 0.1)

Input dataset:

168: VCFTools Slice on data 166 and

Columns:

P1,P3,P5

Remove alternate alleles if not found:

No

Exclude rows not containing variants:

No

Execute

VCFTools Slice (version 0.1)

Input dataset:

168: VCFTools Slice on data 166 and data 23

Regions:

166: Select last on data 165

Execute

VCFTools Compare (version 0.1)

Datasets to Compare

Datasets to Compare 1

Dataset name:

All

Dataset:

23: Varscan on data 21

Datasets to Compare 2

Dataset name:

P1

Dataset:

168: VCFTools Slice on data 166 and data 23

Add new Datasets to Compare

Comparison window:

0

Execute

Table Annovar (version 0.1)

Variants:

2: Varscan on data 21

Gene Annotations:

refGene
wgEncodeGencodeCompV14

Annotation Regions:

genomicSuperDups
phastConsElements46way

Annotation Databases:

snp137NonFlagged
esp6500si_all
snp137
cosmic64

Execute

Annovar Variants Reduction (version 0.1)

Variants:

301: Select on data 300

Gene annotation:

wgEncodeGencodeCompV14

Filtering by Regions

Filtering by Region 1

Regions to use for filtering:

phastConsElements46way

Keep or remove variants in regions:

keep

Remove Filtering by Region 1

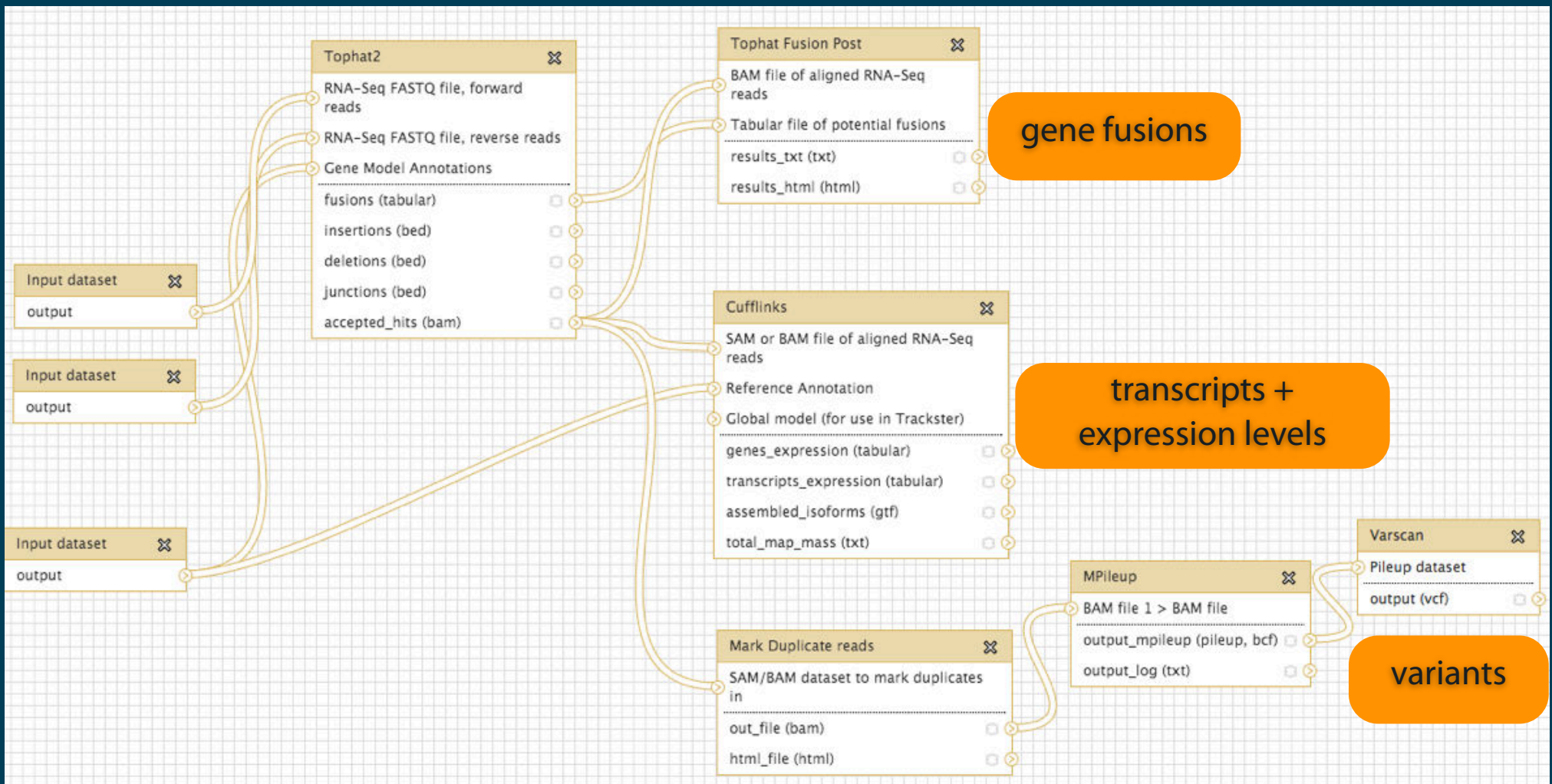
Add new Filtering by Region

Databases to use for filtering:

1000g2012apr_all
avsift
snp137NonFlagged
esp6500si_all

Execute

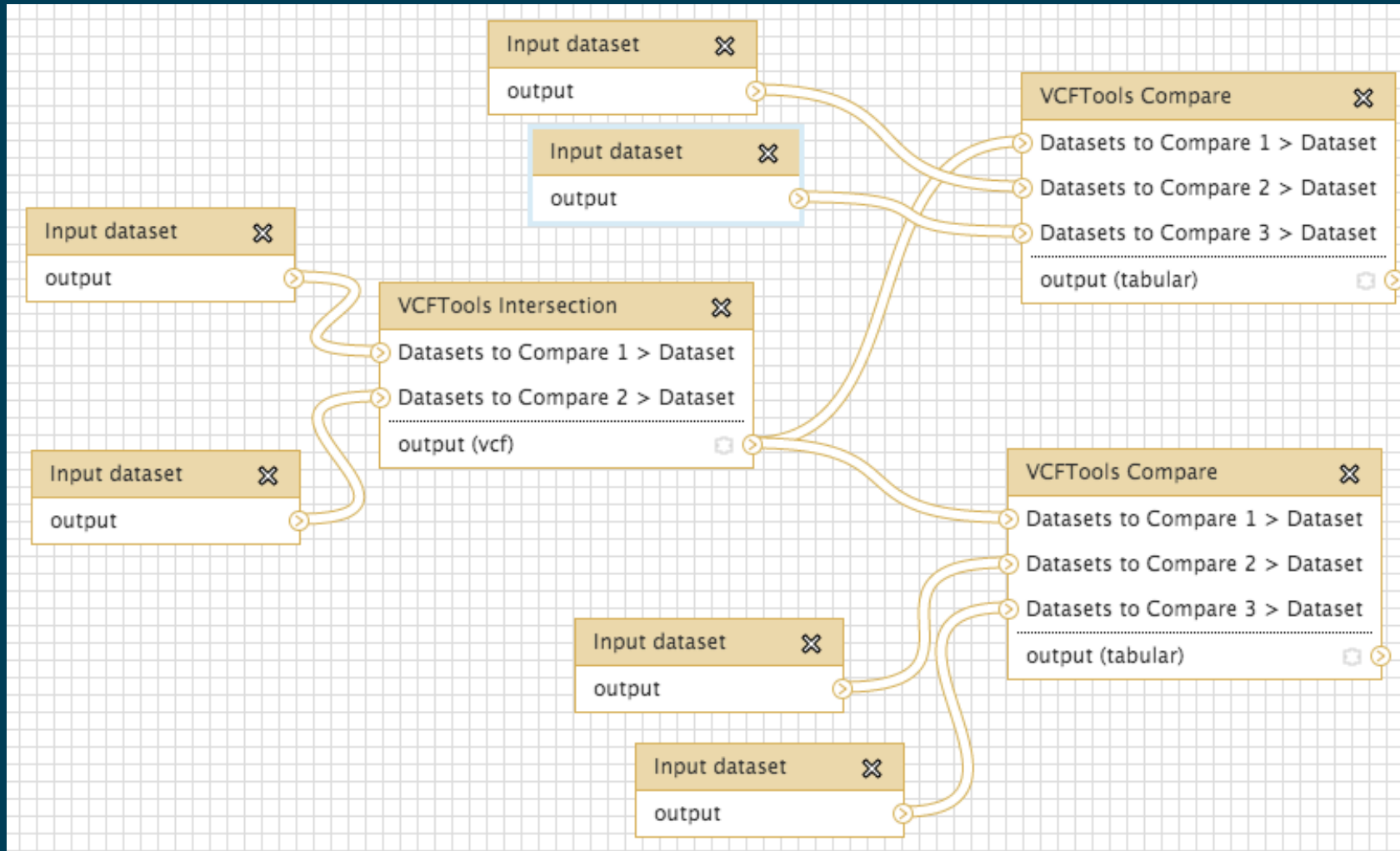
Single Sample Transcriptome Analysis



Aggregate Analysis

1. Differential expression with Cuffdiff
2. Join variants with MPileup

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	3	2	1	2	1	4
HP MIA (84)	6	6	5	5	4	4	19
HP PC (1769)	21	29	23	14	29	15	49
HP ALL (64,669)	247	357	332	214	280	233	263

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	3	2	1	2	1	4
HP MIA (84)	6	6	5	5	4	4	19
HP PC (1769)	21	29	23	14	29	15	49
HP ALL (64,669)	247	357	332	214	280	233	263

Cell line does not appear very similar to tumors

OM = OncoMap, HP = hybrid capture with probes

Patient Mutations vs.

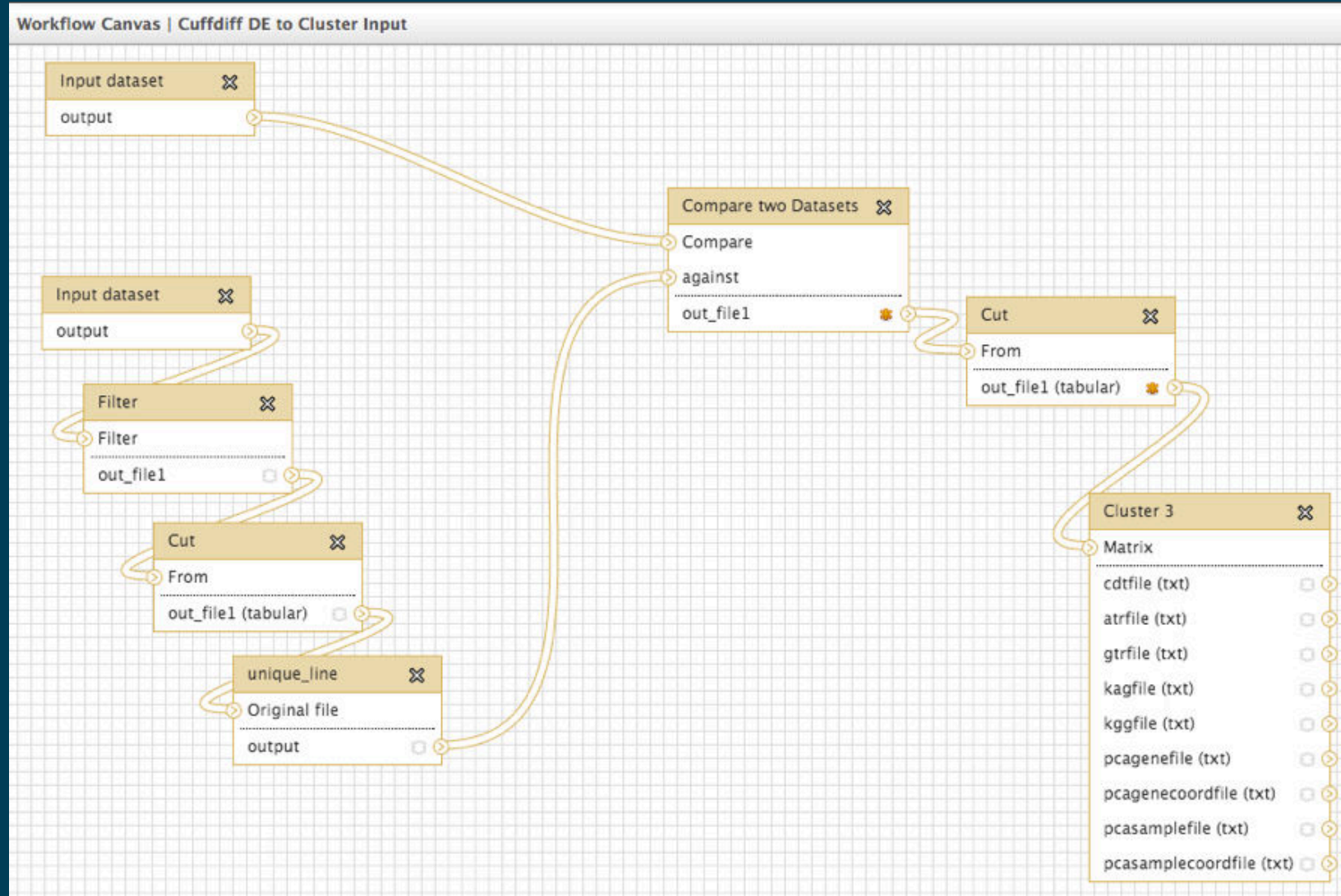


<http://www.broadinstitute.org/ccle/home>

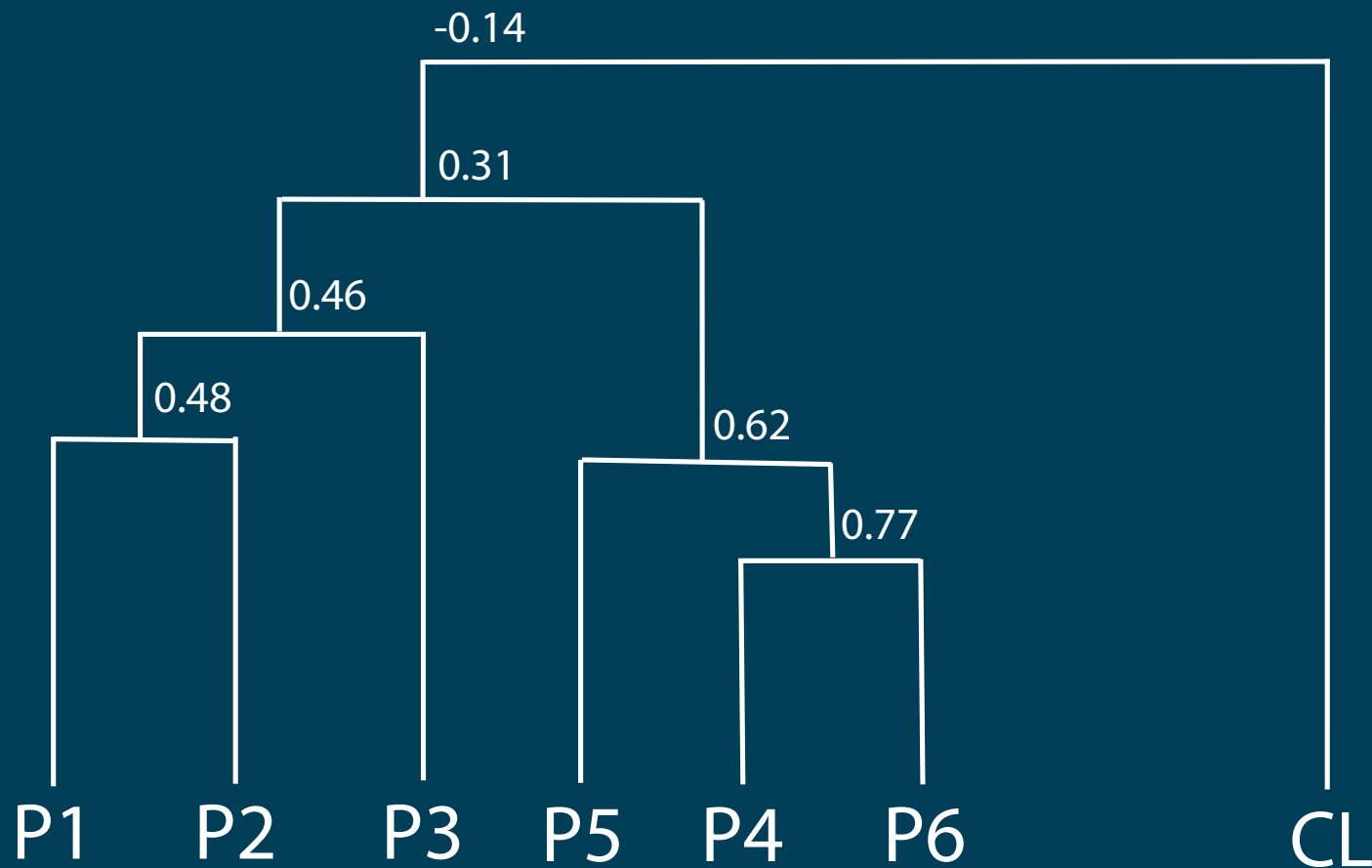
	P1	P2	P3	P4	P5	P6	CL
OM MIA (4)	0	1	1	0	0	0	4
OM PC (11)	0	1	1	0	0	0	4
OM ALL (114)	0	3	2	1	2	1	4
HP MIA (84)	6	6	5	5	4	4	19
HP PC (1769)	21	29	23	14	29	15	49
HP ALL (64,669)	247	357	332	214	280	233	263
Tumor %	90%	90%	100%	0%?	60%	40%	

OM = OncoMap, HP = hybrid capture with probes

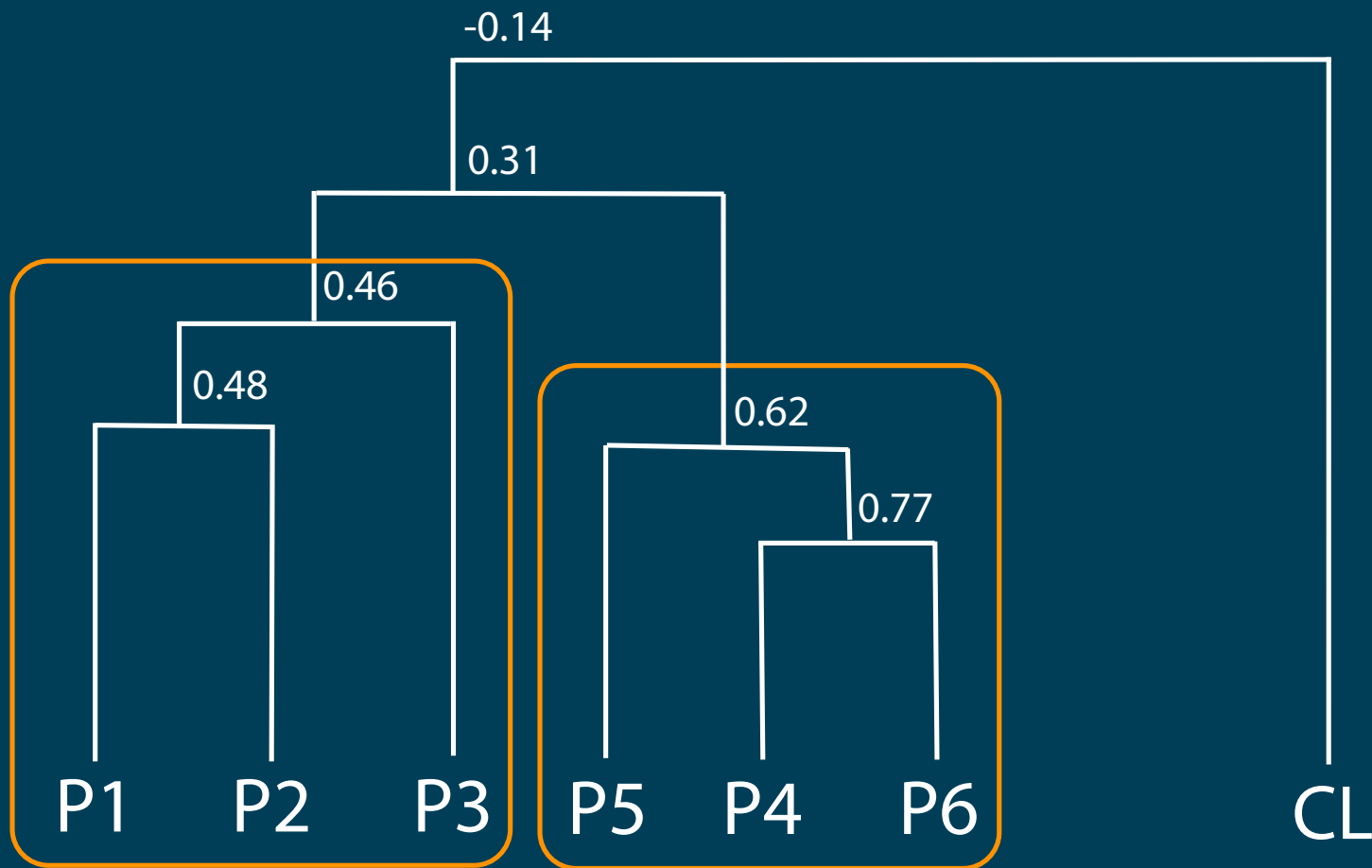
Clustering via Differential Expression



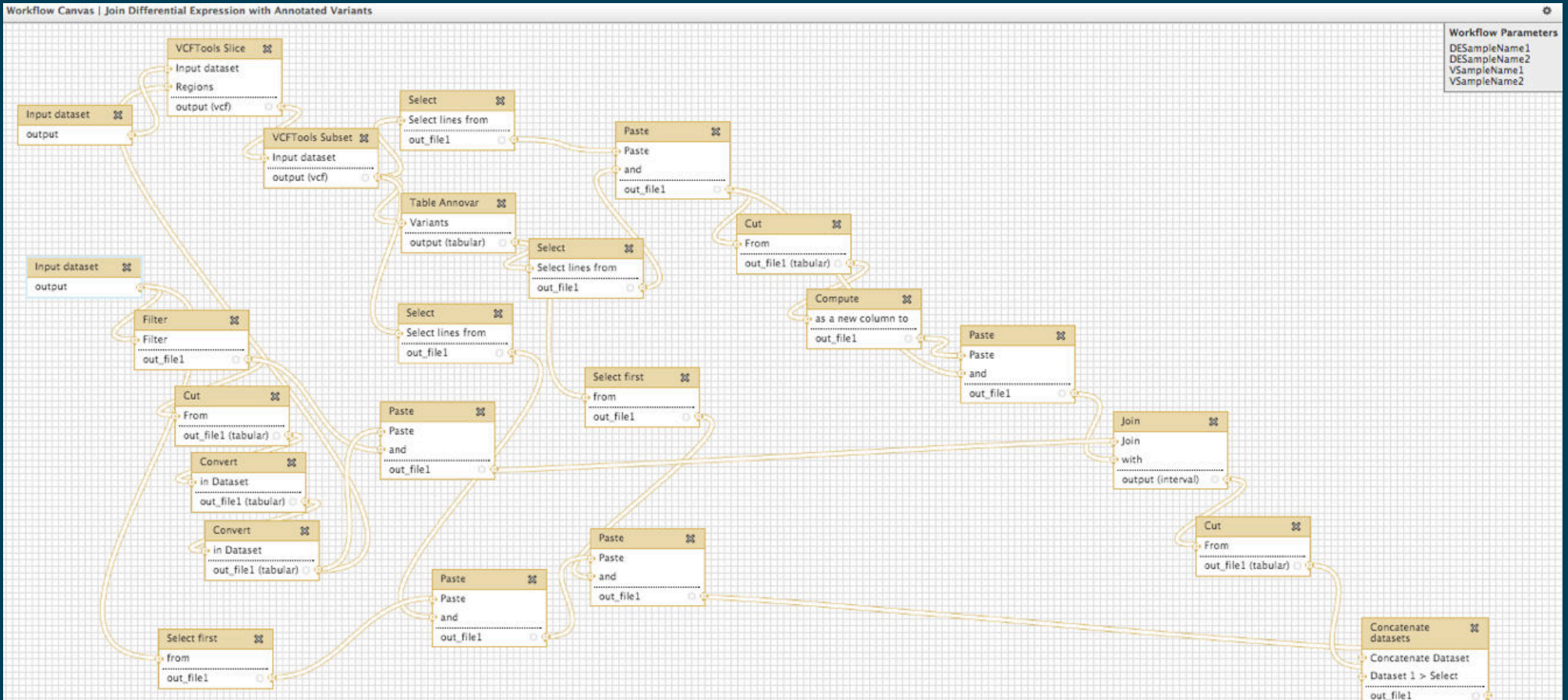
Gene Expression Clustering



Gene Expression Clustering



Variants + Gene Expression + Annotation



Finds and annotates variants in differentially-expressed genes or isoforms

Annotated Variants

1852 genes display differential expression (DE), ~5200 deleterious variants in DE genes

Have identified DE + mutations in known pancreatic cancer genes (e.g. REG1A)

- ✦ step toward eQTL and drug targets

Galaxy Visualization and Visual Analysis

Visualization/visual analysis in Web browser

- ✦ no software or data downloads

Scales to very large datasets

Can share and publish fully-functional visualizations

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

Attributes Convert Format Datatype Permissions

Edit Attributes

Name:

Info:

Annotation / Notes:

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database / Build:

Number of comment lines:

This will inspect the dataset and attempt to correct the above column values if they are not accurate.

History

Small Sample/Treatment Differential Expression Analysis
 10.7 MB

15: Differential Transcript Expression

2,535 lines
 format: tabular, database: hg19

1	3	4
TCONS_00000001 = NM_001005240	XLLOC_1	
TCONS_00000002 = NM_130760	XLLOC_1	
TCONS_00000003 = NM_130762	XLLOC_1	
TCONS_00000004 = NM_033513	XLLOC_1	
TCONS_00000005 = NM_004359	XLLOC_1	
TCONS_00000006 = NM_005317	XLLOC_1	

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

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

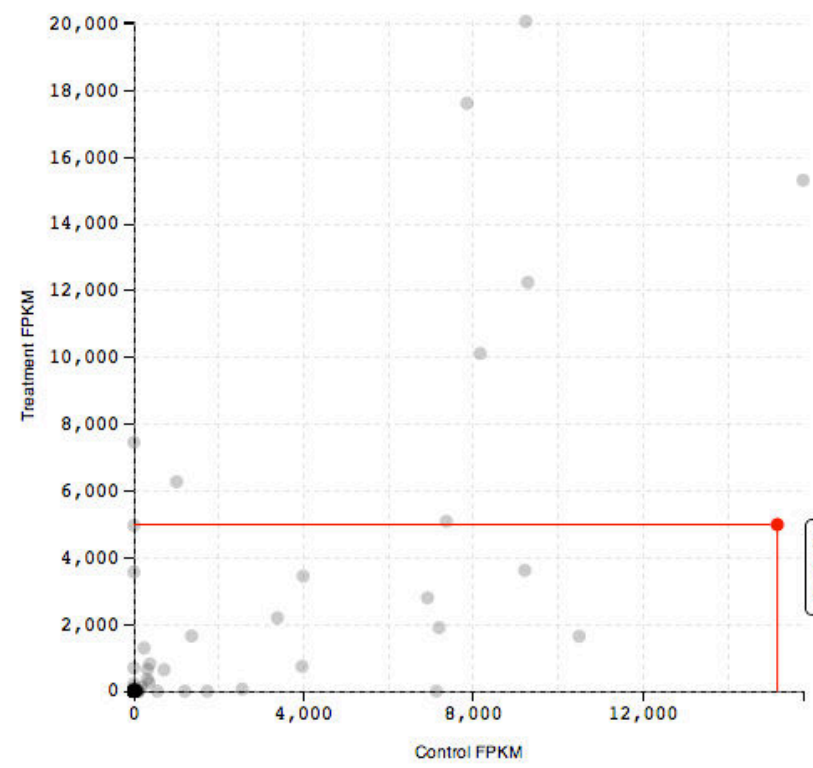
Scatterplot of 'Differential Transcript Expression'

Data Controls

Chart Controls

Statistics

Chart



History

Small Sample/Treatment Differential Expression Analysis
10.7 MB

15: Differential Transcript Expression

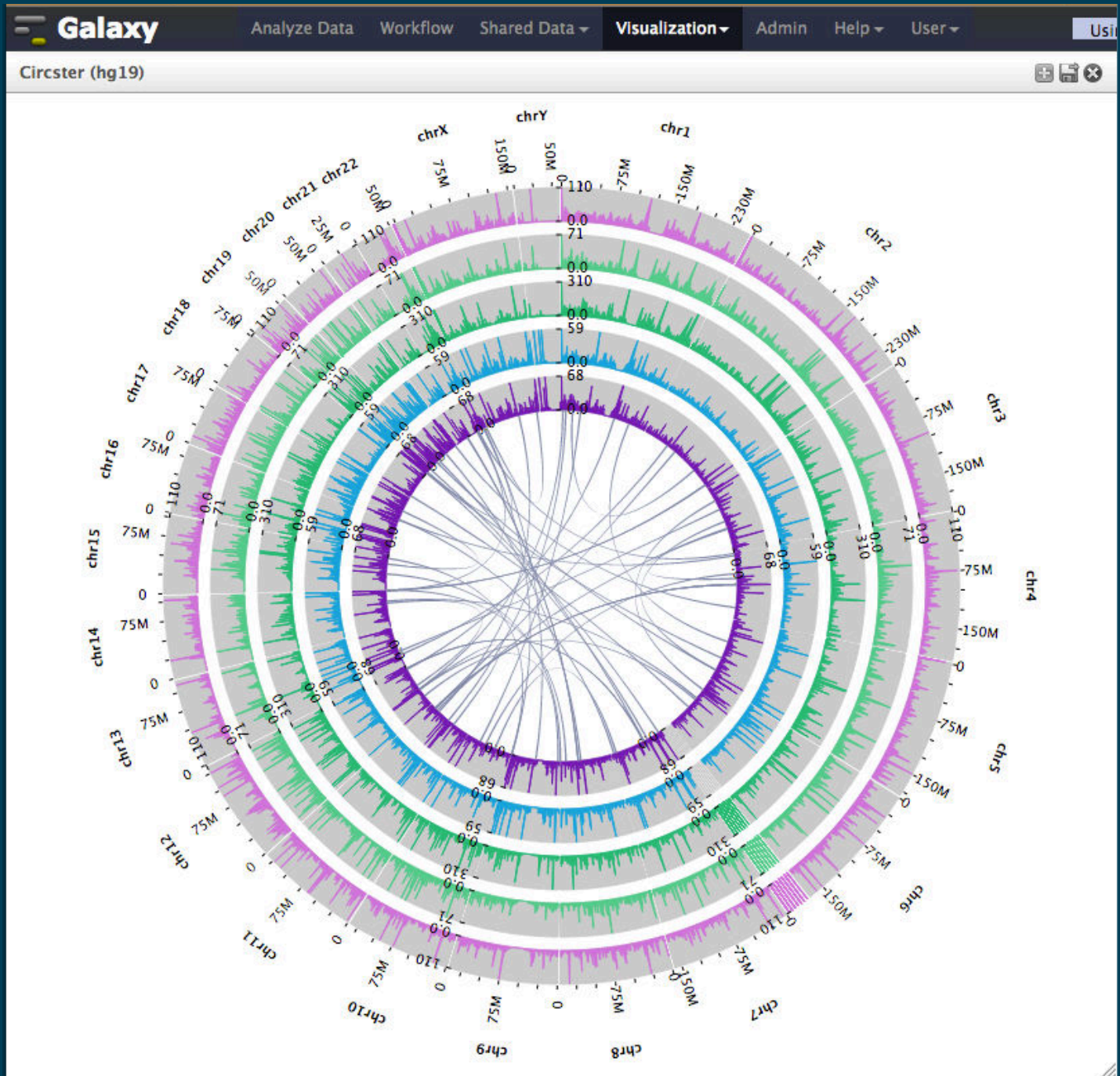
2,535 lines
format: tabular, database: hg19

1	2	3	4
TCONS_00000001 = NM_001005240	XL	LOC_1	
TCONS_00000002 = NM_130760	XL	LOC_1	
TCONS_00000003 = NM_130762	XL	LOC_1	
TCONS_00000004 = NM_033513	XL	LOC_1	
TCONS_00000005 = NM_004359	XL	LOC_1	
TCONS_00000006 = NM_005317	XL	LOC_1	

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 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)

Name:

Dist:

Annotation:

Edit:

Phyloviz Settings: X

Phylogenetic Spacing (px per unit): (50-2500)

Vertical Spacing (px): (5-30)

Font Size (px): (5-20)

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)

Name:

Dist:

Annotation:

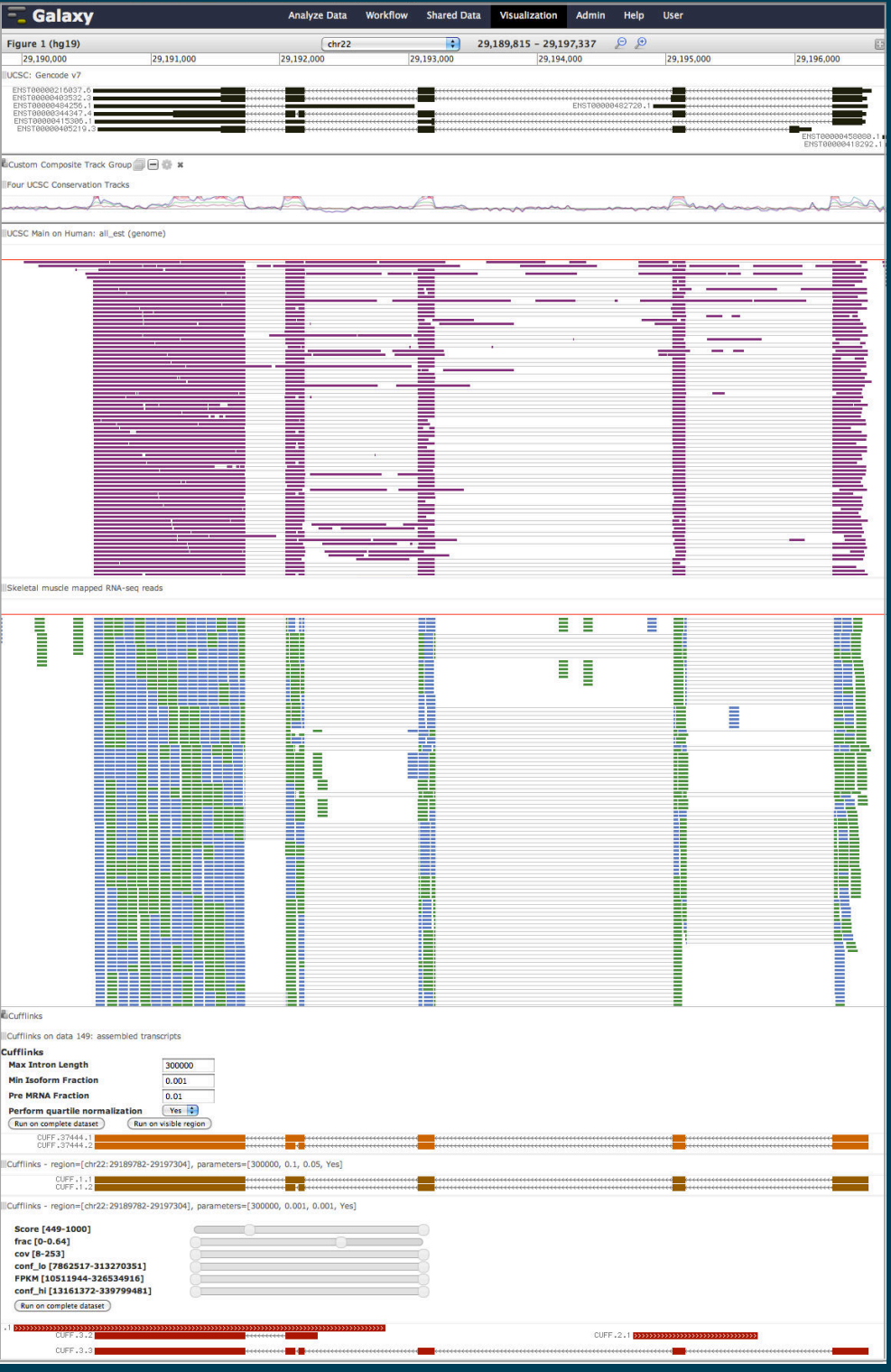
Edit:

Phyloviz Settings: X

Phylogenetic Spacing (px per unit): (50-2500)

Vertical Spacing (px): (5-30)

Font Size (px): (5-20)



Visual Analysis Demo

Real-time Visual Analysis

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

Conclusions

New tools, workflows, and visual analysis tools for analyzing sequence data from cancer

ETA is October 2013, but can be set up locally now if motivated



Galaxy



EMORY

WINSHIP
CANCER
INSTITUTE

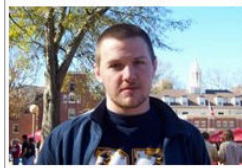
A Cancer Center Designated by
the National Cancer Institute



Enis Afgan
IRB



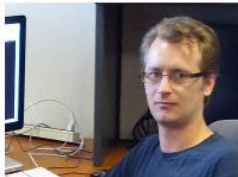
Guru Ananda
Penn State



Dannon Baker
Emory



Dan Blankenberg
Penn State



Dave Bouvier
Penn State



Dave Clements
Emory



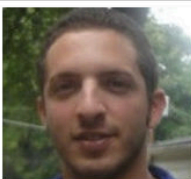
Nate Coraor
Penn State



Carl Eberhard
Emory



Jeremy Goecks
Emory



Sam Guerler
Emory



Jennifer Hillman Jackson
Penn State



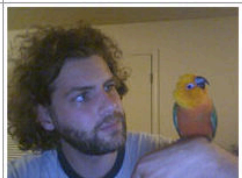
Greg von Kuster
Penn State



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>

jeremy.goecks@emory.edu



1 Integrated Visualization & Computing Workshop

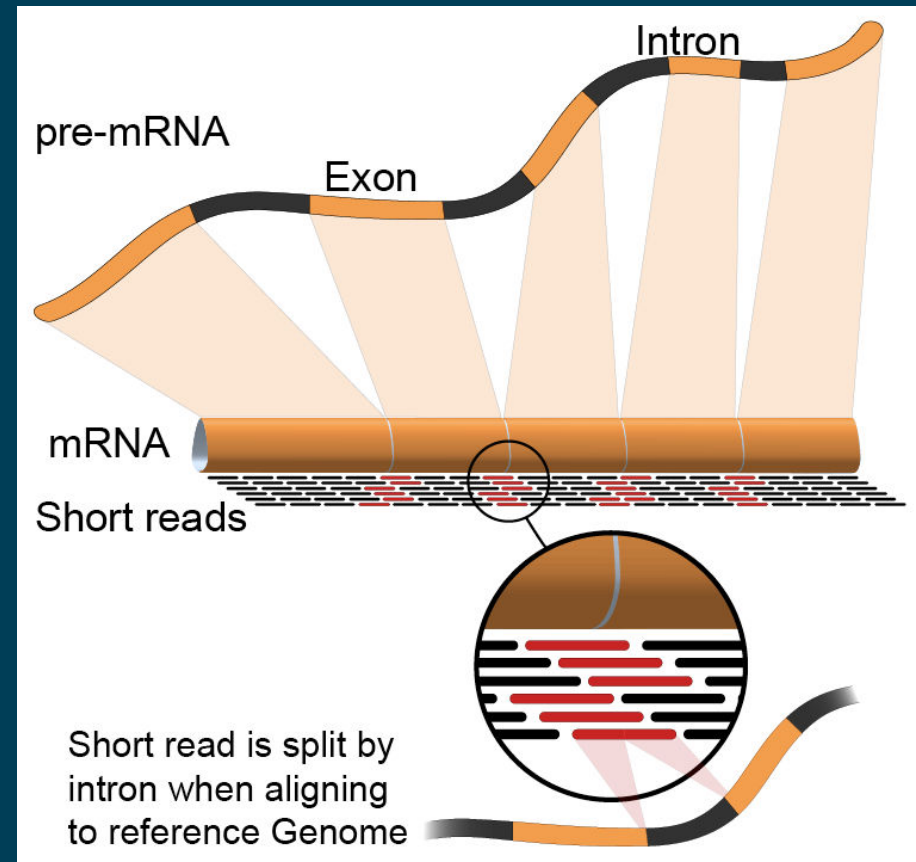
Tuesday, 14:10-16:05

Preliminary Data

6 patients, whole transcriptome sequencing (RNA-seq) of primary tumor
♦ mixed populations!

MiaPaCa2 cell line, whole transcriptome sequencing

Total sequencing data: 50 GB



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