

Creating and Using Tumor Genome Profiles with Galaxy

Jeremy Goecks

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC





Cancer Genomics with Galaxy



My Perspective

I do computational biology/bioinformatics

I analyze all sorts of high-throughput sequencing data for many different purposes

- model, non-model
- molecular evolution, systematics, cancer

Analyzing sequence data is hard and difficult to automate

However, anyone can do it with the proper computational systems

What is the Cost of Analyzing NGS Data?

Musings

Highly accessed

The \$1,000 genome, the \$100,000 analysis?

Elaine R Mardis

Correspondence: Elaine R Mardis emardis@wustl.edu

Author Affiliations

The Genome Center at Washington University School of Medicine, 4444 Forest Park Blvd, St Louis, MO 63108, USA

Genome Medicine 2010, 2:84 doi:10.1186/gm205

The electronic version of this article is the complete one and can be found online at: http://genomemedicine.com/content/2/11/84

Published: 26 November 2010

© 2010 BioMed Central Ltd



OPINION

The real cost of sequencing: higher than you think!

Andrea Sboner^{1,2}, Xinmeng Jasmine Mu¹, Dov Greenbaum^{1,2,3,4,5}, Raymond K Auerbach¹ and Mark B Gerstein^{*1,2,6}



Galaxy Project: Fundamental Questions

When genomics (or any 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

Galaxy Demo

\varTheta 🔿 🔿 📑 Galaxy			
$\leftarrow \rightarrow \mathbf{C} \widehat{\mathbf{n}} \underline{\mathbf{n}} \text{tachylite01.bx.}$	mathcs.emory.edu/g/jeremy/roo	t	☆ =
- Galaxy	Analyze Da	ata Workflow Shared Data - Visualization - Admin Help - User -	Using 2.1 TB
Tools	Hello world! It's runnir	ng	History C 🌣
search tools		edit static/welcome.html	PanCan P3 9.3 GB
Get DataSend DataENCODE ToolsLift-OverText ManipulationFilter and SortJoin, Subtract and GroupConvert FormatsExtract FeaturesFetch SequencesFetch AlignmentsGet Genomic ScoresOperate on Genomic Intervals	ł	www.fssmbb. grow noodly appendages www.missioner.eter www.missioner	22: Tophat Fusion Post Image: Organization of the second state of the second sta
Statistics Wavelet Analysis Graph/Display Data Regional Variation Multiple regression		usegalaxy.org	and data 4: assembled transcripts 16: Cufflinks on data 9 and data 4: transcript expression 15: Cufflinks on data 9 ● Ø ※
Multivariate Analysis Evolution Motif Tools Multiple Alignments Metagenomic analyses FASTA manipulation NGS: QC and manipulation	This project is supported in part	by <u>NSF</u> , <u>NHGRI</u> , and <u>the Huck Institutes of the Life Sciences</u> .	and data 4: gene expression 14: Tophat2 on data 3, ● Ø ※ data 4, and data 2: accepted hits 13: Tophat2 on data 3, data 4, and data 2: splice junctions
NGS: GATK Tools (beta) NGS: Mapping NGS: Indel Analysis NGS: RNA Analysis	A T		12: Tophat2 on data 3, Image: Operation of the second
<			

Accessibility

All tools looks the same

No command line or programming

Easy to chain tools together into larger analyses

C Tools Evolution Motif Tools Multiple Alignments Metagenomic analyses FASTA manipulation

NGS: QC and manipulation NGS: GATK Tools (beta)

NGS: Mapping

NGS: Indel Analysis NGS: RNA Analysis

RNA-SEQ

- <u>Tophat for Illumina</u> Find splice junctions using RNA-seq data
- <u>Tophat2</u> Gapped-read mapper for RNA-seq data
- <u>STAR</u> Gapped-read aligner for RNA-seq data
- <u>Tophat Fusion Post</u> postprocessing of
- <u>Cufflinks</u> transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- <u>Cuffcompare</u> compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- <u>Cuffmerge</u> merge together several Cufflinks assemblies
- <u>Cuffdiff</u> find significant changes in transcript expression, splicing, and promoter use

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 denominator to improve accuracy of differential expression calls for low abundance transcripts.

Use Reference Annotation:

No
Perform Bias Correction:

No 🛟

Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

4

Use multi-read correct:

No :

Tells Cufflinks to do an initial estimation procedure to more accurately weight reads mapping to multiple locations in the genome.

Execute

Cufflinks Overview

<u>Cufflinks</u> assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one. Please cite: Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L. Transcript assembly and abundance

Reproducibility



Workflows enable reuse and precise reproducibility

Users can add tags and annotations for additional context

😝 🔿 🔿 🧷 = Galaxy | Published Page | pr 🔅

G f https://main.g2.bx.psu.edu/u/webb/p/polar-bears

Galaxy

Analyze Data Workflow Shared Data • Visualization • Cloud • Admin Help •

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. Drautz, 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. The 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-y 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 to closular experienced a prolonged and dramatic decline in its effective population size during the last ca. 500,000 years. We demonstrate that brown bears to closular explored adpation 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	
	ontains 13,038,705 putative SNPs that were identified using a de novo assembly of the polar corresponds to a SNP, and has <u>117 columns</u> .	r bear genome (rather than the dog
	Galaxy Dataset bear assembly SNPs	🖬 O 🔗
The "bear mitochondrial SNPs" tal 31 columns.	ble contains 1,698 positions where not all 28 individuals had the same nucleotide. Each row	represents one of these SNPs, and ha
	Galaxy Dataset bear mitochondrial SNPs	🖬 🖯 🞯
The "bear SAPs" table contains 75 columns.	,501 variant position in putative protein-coding regions, both synonymous and non-synony	mous changes. Each row has 11
	Galaxy Dataset bear SAPs	
	a table) uses a streamlined file with the locations of 19,014 dog genes (basically, each one is	the longest of a set of overlapping

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 "eve" for 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 builte re-run button, and clicking on "Execute". You also may need to give Calaxy 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 bears 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 512 of the supplement (which has one chromosome shown in Figure 48 of the main paper). The new figure indicates the 3Mb interval on the left end of each dog chromosome, which are treated as heterchromatin 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
 Galaxy Workflow | bear admixture map
 Ge
 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,
 Le, where an allele having a selective advantage increased in frequency in the population and brought along with it the neighboring alleles. The table appeared

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 <u>Table S8</u> 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.

Communication and Reuse

Galaxy Workflow | bear sweep table

- 0 d

A O O Calaxy Rublished Page p - x		
G f Attps://main.g2.bx.psu.edu/u/webb/p/polar-bears		
Galaxy Analyze Data Workflow Shared Data - Visualization - Cloud - Admin Help - User -	Using 588.3 G8	
Published Pages webb polar-bears	About this Page	
suring the last ca. 500,000 years. We demonstrate that brown bears and PBs have had sufficiently independent evolutionary histories over the last +-5 million years to leave imprints in the PB nuclear genome that likely are associated with ecological adaptation to the Arctic environment.	Author	
Datasets	webb	
Aany of the analyses reported in the paper were based on the five datasets given here. (You can also find them under Shared Data -> Data Ibraries -> Genome Diversity, then under bear and dog.)	Related Pages	
he first consists of 12,023,192 dog-based "SNPs", i.e., positions in the dog genome where we detected two distinct nucleotides in the	Published pages by webb	
orresponding bear locations (among the our three bear species, polar bear, brown bear, and American black bear). Each row in the table orresponds to a SNP, and has <u>124 entries</u> .	Rating Community de de de de	
Galaxy Dataset bear SNPs Galaxy Dataset bear SNPs Gold A Control of the polar bear genome (rather	(0 ratings, 0.0 average) Yours de te de de de	
than the dog assembly). Each row of the table corresponds to a SNP, and has <u>117 columns</u> .	Tags	
Galaxy Dataset bear assembly SNPs	Community: evolution climate-change	
scaffold1 370 T C 999 36 0 2 135 0 scaffold1 441 A G 89.9 41 0 2 155 0 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 1 224 scaffold1 1464 G T 999 14 6 1 29 scaffold1 1693 C T 999 0 26 0 75	bears Yours:	
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$		
scaffold1 4901 C T 999 0 30 0 87 scaffold1 5591 C A 999 0 36 0 105 4	Contraction of the second seco	
scaffold1 5969 T C 999 0 28 0 81	< 🔿 C 🐔 🔓 https://main.g2.bx.psu.edu/u/webb/p/polar-bears	☆ =
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 <u>31 columns</u> .	- Galaxy Analyze Data Workflow Shared Data - Visualization - Cloud - Admin Help+ User+	Using 588.3 GB
Galaxy Dataset bear mitochondrial SNPs Golaxy Dataset bear mitochondrial SNPs Golaxy Dataset bear SAPs' table contains 79,501 variant position in putative protein-coding regions, both synonymous and non-synonymous changes. Each	Published Pages webb polar-bears	About this Page
row has <u>11 columns</u> .	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 ownands run (which takes a couple of minutes), click on the "yee" 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. RCA on data 6", clicking on the blue re-run button, and clicking on "Execute". You also may need to give Calaxy a minute after the workflow Kinshes to put the output files in the correct places.]	Author webb Related Pages All published pages
	Galaxy Workflow bear PCA Galaxy Book bear bea	Published pages by webb Rating
	treated as heterchromatin 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	Yours Tags Community:
	Step 6: Filter	evolution climate-change bears fours:
	Filter Output dataset 'output' from step 5 With following condition	l I
	c6!='chrX' and c12S>=0.5 Step 7: Admixture	
	SNP dataset Output dataset 'out_file1' from step 6 Ancestral population 1 individuals	
	Output dataset 'output' from step 3 Ancastral nonulation. 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 <u>Tables</u> 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	* = >
		Calaxy C C A A https://main Calaxy Workflow "bear admixture r
		You can start using this wor

Galaxy Visualizations

Many visualizations

- biological: genome browser, Circos plot, phylogenetic tree
- numerical: bar charts, histograms, etc.

Developers can add new visualizations

Completely Web-base: no data or software downloads





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 for 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

				Workflow Parameter tissue_name
Cufflinks (version 0.0.5) SAM or BAM file of aligned RNA-Seq reads: 19: MarkDups_Dupes Marked.bam	Input Dataset 🛛 🗱 output	1 97)		
300000	,	Tophat for Illumina 😫		
	(RNA-Seq FASTQ file		F data by attribute 🔮
Min Isoform Fraction: 0.1		Gene Model Annotations	Filter	
0.1	(insertions (bed)	out_file1	<u>_</u>
Pre MRNA Fraction:	()	deletions (bed)		
0.15	×	junctions (bed) 🛛 💿	Filter	×
Perform quartile normalization:		accepted_hits (bam)	Filter	
No : Removes top 25% of genes from FPKM denominate transcripts.			out_file1	0 0
Use Reference Annotation:				
No	Variants (hg19)	128.200.000	_ chr3	128,135,183 - 128,489,933 P
	Il probe_tiled_regions.bed			
Perform Bias Correction:		19 C 10 C		
No	EVarsean: STER, durs removed	variants in mohe regions		
No	EVarscan: STAR, dups removed, Summary	, variants in probe regions		111
No : Blas detection and correction can significantly imp				111
No 3 Blas detection and correction can significantly imp Use multi-read correct: No 3	Summary Sample1 [Varscan: Roche Exome Variant Summary			
No Contraction and correction can significantly imp Blas detection and correct: No Contraction of the second secon	Summary Sample1 [[Varscan: Roche Exome Variant	s in probe regions		
No Contraction and correction can significantly imp Blas detection and correct: No Contraction of the second secon	Summary Sample1 IIVarscan: Roche Exome Varlant Summary Sample1 IIVarscan: Tophat, dups remove Summary	s in probe regions		
No Contraction and correction can significantly imp Use multi-read correct: No Contraction of the second se	Summary Sample1 IIVarscan: Roche Exome Varlant Summary Sample1 IIVarscan: Tophat, dups remove Summary Sample1 IITophe2 on data 2, data 146, e	s in probe regions		
No Contraction and correction can significantly imp Blas detection and correct: No Contraction of the second secon	Summary Sample1 IIVarsan: Roche Exome Variant Summary Sample1 IIVarsan: Tophat, dups remove Summary Sample1	s in probe regions d is probe regions d is probe regions out data 1: accepted, JNts		
Blas detection and correction can significantly imp Use multi-read correct: No C Tells Cufflinks to do an initial estimation procedur genome.	Summary Sample1 IIVarscan: Roche Exome Varlant Summary Sample1 IIVarscan: Tophat, dups remove Summary Sample1 IITophe2 on data 2, data 146, e	s in probe regions d in probe regions d in probe regions ord data 1: accepted, Jins		







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 1500 publications

Galaxy is Very Popular

60+ local installations all over the world

香港中文大學 - 華大基因跨組學創新研究院 华大县因。 CUHK-BGI Innovation Institute of Trans-Omics GGG



http://bit.ly/gxyservers



Galaxy Cancer Genomics with Galaxy

Using Galaxy for Cancer Genomics

New tools

complement existing high-throughput sequencing analysis tools

New workflows

workflows are understandable, extendable, sharable

New visualizations and visual analysis applications

 e.g. interactively visualize and call variants in a Web browser

Motivating Applications

Tumor-cell line matching

- primary pancreatic adenocarcinoma tumors
 vs. pancreatic cancer cell lines
- variants + gene expression

Deep targeted tumor sequencing

- small, inexpensive panel of ~26 genes sequenced at 5-10k coverage
- focus on clinically actionable genomic markers





http://usegalaxy.org/cancer



Videos to Get Started



expression:

Galaxy Accessible Page X							
$\leftarrow \rightarrow \mathbf{C} \bigcirc \text{https://usegalaxy.org/u/jeremy/p/cancer-analyses}$							
Galaxy Analyze Data Workflow Shared Data - Visualization - Cloud - Help - User -	Using 805.1 GB						
Accessible Page Cancer Analyses	About this Page						
(1) This is the basic tumor exome analysis workflow that calls variants from targeted exome resequencing data:	About this rage						
	Author						
Galaxy Workflow Exome Basics Analysis Image: Comparison of the second secon	jeremy						
(2) This is the RNA-seq analysis workflow. This workflow analyzes tumor RNA-seq data to find small variants, gene fusions, and quantify gene expression:	Related Pages All published pages						
Galaxy Workflow Tumor RNA-seq Analysis	Published pages by jeremy						
	Rating						
(3) This is the integrated variant analysis workflow. To use this workflow, two datasets in the same history are needed: (a) a variants dataset from either the exome or transcriptome analysis workflow) (b) Cufflinks Gene Expression dataset. This workflow then identifies:	Community (0 ratings, 0.0 average)						
deleterious variants	Yours ****						
deleterious and druggable variants	Tags						
deleterious variants in highly-expressed genes deleterious and druggable variants in highly-expressed genes	Community: none						
Galaxy Workflow Integrated Variant Analysis: Expression/Functional/Drug Image: Comparison of the second seco	Yours:						
(4) This is an extended workflow for use when only a tumor exome is available. Starting with tumor exome sequencing data, it identifies deleterious variants and druggable variants:							
Galaxy Workflow Tumor Exome Analysis Geographies							
(5) VCF Variant recovery. Use this workflow to obtain a list of variants in VCF format from a ANNOVAR table of variants. Variants in VCF format are useful for visualization.							
Galaxy Workflow Recover Variants from ANNOVAR George							
(6) Workflow to convert Tophat-fusion-post results to chrint format, which can be used to visualization fusion in Circster:							
Galaxy Workflow Tophat fusion post output to chrint							
To use these workflows on a Galaxy instance other than this one, take these steps:							
1. As an admin user, download the workflows that you want to use and follow the prompts to install needed tools. Here is more							
explanation on installing tools needed for workflows. 2. Download and install ANNOVAR (no automatic installation is possible due to ANNOVAR's licensing):							

Single Sample Transcriptome Analysis



Exome: From Sequence to Drugs

Workflow Canvas | Exome Variant Calling + Annotation + Drugs

\$





Integrates private and public data

Workflow Canvas | Exome Variant Calling + Annotation + Drugs





Understandable, editable, sharable

⊖ ⊙ ⊙ ⊊Galaxy Accessible Page ×	м ^л
← → C	☆ 💿 🖓 🛠 🄌 🗖 🔍 🚍
Galaxy Analyze Data Workflow Shared Data - Visualization - Cloud - Help - User -	Using 805.1 GB
Accessible Page Cancer Analyses	About this Page
Analysis Histories for Cell Line Data Using the first three workflows above, here are the analysis histories for the three pancreatic cancer cell lines, Mia PaCa2, HPAC, and PANC-1.	Author jeremy Related Pages
Mia PaCa2 Exome:	All published pages Published pages by jeremy
• Galaxy History Mia PaCa2 Exome Analysis • □	Rating
Mia PaCa2 Transcriptome:	Community
Galaxy History Mia PaCa2 RNA-seq Analysis	(O ratings, 0.0 average)
Mia PaCa2 Integrated Variant Analysis:	Yours the test
Galaxy History Mia PaCa2 Integrated Variant Analysis Galaxy History Mia PaCa2 Integrated Variant Analysis	Tags Community: none
HPAC Exome:	Yours:
Galaxy History HPAC Exome Analysis Image: Comparison of the second secon	G.
HPAC Transcriptome:	
Galaxy History HPAC RNA-seq Analysis Image: Comparison of the seq Analysis	
HPAC Integrated Variant Analysis:	
Galaxy History HPAC Integrated Variant Analysis	
PANC-1 Exome:	
Galaxy History PANC-1 Exome Analysis C	
PANC-1 Transcriptome:	
Galaxy History PANC-1 RNA-seq Analysis	
PANC-1 Integrated Variant Analysis:	
Galaxy History PANC-1 Integrated Variant Analysis	

🗧 🕤 🗿 💐 Galaxy Accessit	ole Page ×					R _M
← → C A https://use	galaxy.org/u/jeremy/p/	/cancer-analyses			☆ 💿 🖬 🔮	e 🕫 🖾 🔌
🕎 Galaxy	Analyze Data	Workflow Shared Dat	a→ Visualization→ (Cloud + Help + User +		Using 805.1 GB
Accessible Page Cancer Analy	yses				About this Page	
Analysis Histories Using the first three workflows PANC-1. Mia PaCa2 Exome: Mia PaCa2 Transcriptome: Mia PaCa2 Integrated Variant Analysis HPAC Exome: HPAC Transcriptome: HPAC Integrated Variant Analysis	above, here are the analysis <u>Galaxy History </u> <u>Galaxy History Malysis:</u> <u>Galaxy History Mia P</u> <u>Galaxy History</u>	s histories for the three p Mia PaCa2 Exome Mia PaCa2 RNA-see	Analysis Analysis I Analysis Iriant Analysis alysis	s, Mia PaCa2, HPAC, and View history C C C C C C C C C C C C C	Author jeremy Related Pages All published pages Published pages by j Rating Community (0 ratings, 0.0 average) Yours Tags Community: none Yours:	eremy *****
+	Galaxy History HP	AC Integrated Vari	ant Analysis	⊕ 🖗		
PANC-1 Exome:	Salaxy History HP	Ac integrated valle	ant Analysis			
+ PANC-1 Transcriptome:	Galaxy History	PANC-1 Exome A	<u>nalysis</u>	O A		
PANC-1 Integrated Variant Analy	ysis:	PANC-1 RNA-seq		• 7		
usegalaxy.org/u/jeremy/h/mia-pace	Galaxy History PAN a2-exome-analysis-1	IC-1 Integrated Var	iant Analysis	€		>

0 0 Calaxy Accessible History X							× ⁿ
← → C Attps://usegalaxy.org/u/jeremy/h/mia-paca2-exome-analysis-1						☆ 🎱 🖬 💲 🖌	▶ 🖾 🍳 🗏
Galaxy Analyze Data	Norkflow	Shared Data -	Visualization +	Cloud -	Help+ User+		Using 805.1 GB
Accessible History Mia PaCa2 Exome Analysis				Sw	itch to this history	About this History	
Mia PaCa2 Exome Analysis 90.2 GB search datasets					0	Author jeremy Related Historie All published historie	<u>s</u>
Dataset		Annotation				Published histories by	<u>y jeremy</u>
<u>1: miapaca2 ng exome R1.fastq</u> <u>2: miapaca2 ng exome R2.fastq</u>	•					Rating Community (0 ratings, 0.0 average)	*****
3: tiled probe regions	•					Yours Tags	*****
<u>4: Map with BWA for Illumina on data 2 and data 1: mapp</u> ed reads	۲					Community: none Yours:	
5: MiaPaCa2: mapped exome reads	۲					4	
<u>6: MarkDups MiaPaCa2: mapped exome reads, no dups.b</u> <u>am</u>	۲						
<u>7: MarkDups_MiaPaCa2: mapped exome reads, no dups.h</u> tml	۲					S.	
8: MiaPaCa2: alignment summary	۲						
<u>9: MPileup on data 6</u>	۲						
10: MPileup on data 6 (log)	۲						
11: MiaPaCa2: Varscan variants	۲						
12: MiaPaCa2: Varscan Variants, probe regions	۲					8	
							>
3							

>

Preliminary Validation using Cell Lines

3 cell lines (MiaPaCa2, HPAC, PANC-1): exome + transcriptome sequencing

CCLE

- 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

Preliminary Validation Results

All CCLE variants (84 total) validated in cell lines

MEK inhibitor validation

- pipelines: predicted cell lines would be responsive
- CCLE: drug response data shows cell line are responsive

PANC-1: fewer rare and deleterious mutations in cell line leads to fewer applicable drugs

Preliminary Results: Cell Lines vs. PAAC tumors



Trajectory

Standardize workflows for use by research oncologists

Improve workflows using new tools

Generate clinician-friendly reports from workflow

Common Galaxy Questions

How does Galaxy scale?

to whatever computing resources are available

Can I modify workflows/pipelines?

yes, either via Web or programmatically

Can I add my own tools and visualizations?

yes, there are frameworks for both

Concluding Thoughts

Galaxy is a very useful platform for high-throughout genomics

- accessible, reproducible, collaborative
- tools, workflows, visualizations
- public, local, cloud

Use Galaxy to go from tumor sequence reads to:

- rare, deleterious (RD) variants (exome)
- gene fusions and gene expression (RNA-seq)
- drugs associated with genes that have RD variants (integrative)

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

genome.gov National Human Genome Research Institute



PENN<u>STATE</u>



Galaxy





Nate Coraor Penn State



Sam Guerler Emory



Ross Lazarus BakerIDI



Anton Nekrutenko Penn State





Dannon Baker

Emory

Dave Clements





James Taylor

Emory

Greg von Kuster Penn State





National Institutes of Health





A Cancer Center Designated by the National Cancer Institute



Mike Rossi

EMORY

Dave Bouvier

Carl Eberhard

Emory

Thanks! Questions? Galaxy

http://galaxyproject.org http://usegalaxy.org

Postdoc and software engineer positions available in Interactive Genomics Lab @ THE GEORGE WASHINGTON UNIVERSITY

http://jeremygoecks.com

@jgoecks jgoecks@gwu.edu

WASHINGTON. DC

