

#### Using Galaxy to Understand Cancer Genomes

#### Jeremy Goecks

**Computational Biology Institute** 

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC



# The High-throughput Sequencing Era Galaxy

Cancer Genomics with Galaxy

#### **A Revolution in Biology**



#### 50-100 Gb / day

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



# 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



#### OPINION

#### The real cost of sequencing: higher than you think!

Andrea Sboner<sup>1,2</sup>, Xinmeng Jasmine Mu<sup>1</sup>, Dov Greenbaum<sup>1,2,3,4,5</sup>, Raymond K Auerbach<sup>1</sup> and Mark B Gerstein<sup>\*1,2,6</sup>



### **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

@HW1-ST565:241:D19RIACXX:1:1101:5456:1997	I:N:U:CGATGT
NGCATAGGCAAGCACCGGAAGCACCCCGGCGGCCGCGGTAAT	GCTGGTGGTCTGCATCACCACCGGATCAACTTCGACAAATACCACCCAGGCTACTTTGG
+	
#1=DDDDDDHHHHHIIIIIIIIAHIIIIIIFCBBB;BBE	CCCCC?CBBCCCCCCCCBBBBBBBBBCCCCCBBBBBCCCCCC
NCGCAACCTCAACACCACCTTCTTCGACCCCGCCGGAGGAGG	AGACCCCATTCTATACCAACACCTATTCTGATTTTTCGGTCACCCTGAAGTTTATATTC
#4=DFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	FFDDDDDDDDFEEEFDDDDDDDDDDDDDEEEEEDEDDDDDD
@HWI-ST565:241:D19R1ACXX:1:1101:10117:1998	1:N:0:CGATGT
NATGTGCCCTCTGGCAGTCTGCTGCTGTGTCCAGAGTCCGAC	rccagctgggctgtaactgggcttggcccccgccttaggccccgccagcaggcgaagca
#1=DDFFFHHHHHJIJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	JJJIJJJJJJIIJIGIJJJJJJHHHFFFFDDDDB=BCDDDDDDDDDDDDDDDBB9>BDA
@HWI-ST565:241:D19R1ACXX:1:1101:10283:1992	1:N:0:CGATGT
NTTGTCACCAAGACCCTACTTCTAACCTCCCTGTTCTTATGA	ATTCGAACAGCATACCCCCGATTCCGCTACGACCAACTCATACACCTCCTATGAAAAAA
+	
#4=DDFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	JJJJJJJJJJJJJJJJJJJJJJJJGIHHHFFFFDDDDDDDDDDDDDDDDDDDDDDDDDDDCCD<
@HWI-ST565:241:D19R1ACXX:1:1101:10632:1993	1:N:0:CGATGT
NGTAAGCCTTCTATGCATCCACACCAAAATCCTGCAGAATGT	AAGTAAGCTCTGCTTTATAAGATGGGTTCACCTTCATCGCAGACTGAAAGTTTCAGTTT
+	
#1:ADDFFHHHHHGHGIGIGIJIIGGIGJIHGGIJGHIEHJD	DHFBGIGGHJJJJJJJJJIIGCFGII@CGCHGIHHEFGFDFDBEDDCCC@5>CDCA;>A
@HWI-ST565:241:D19R1ACXX:1:1101:10895:1991	1:N:0:CGATGT
NCTAGCACAGAGAGTTCTCCCCAGTAGGTTAATAGTGGGGGGGT	AAGGCGAGGTTGGCGAGGCTTGCTAGAAGTCATCAAAAAGCTATTAGTGGGAGCAGAGT
+	
#1=DFFFFHHHHHJGHIJJJJIJGHIJHIIJJJJFIIJJJD0	CDDDDDDDCBCDDDDDDDDDDDDDDDDDDDDDDDDDDD
eHw1=ST565:241:D19RIACXX:1:1101:10838:1994	1:N:0:CGATGT
TCCCTGCTACTGCTGATGCACTGTCCTCTCCCTGCAGCCCC	IGGCITUCCAGUCITUCTUCTGAUCUCUTTUCAACAGUCITIGGAAUTUCAGUTGUCACUA
+	
#1=DFFFFHHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	1.N.O.COMCT
VERMI-SISOS:241:DISKIACAX:1:1101:11/5/:1991	1:N:0:CGAIG1
	IGGGGCCICCATGCCAGGACIGCAAAGIGATCCAGCCCIACCIGICIICCCACCIGIG
#4=DFFFFFHGHHILITTILITITILITITITITI	
8HWT_ST565+241+D19R1aCXX+1+1101+11780+1992	1.N.O.CGATCT
NTAAATACTAAGCACAAGCTCACTTCCCTCTTGGTCAGGTGG	THE THE TRACE OF T
+	
#1=DDFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	CHTTTTTTTTTTTTTTTTTTTHEEHHHFFFFFFFFFFFFF
0HWT_ST565:241:D1981ACXX:1:1101:12154:1991	1:N:0:CGATGT
NCGCTTTGGGAGGCTGAGGTGGGAAGATGACTTGAGCCCCAGG	AGTTCGAGACCAGCCTGGGCAACATGGTAAAACCCTTTCTCTGACCCCCCACAAAAATAA
+	
#1:DBDDDDDDDDDD:CBB2ACA?EDDDI@DED?BDDDDDDT	D:B8.<@ADD@C@:5?A???@?@AA@=>5>>AA=< <aaaaaaaa>&gt;&gt;????&gt;&gt;&gt;&gt;???</aaaaaaaa>
@HWI-ST565:241:D19R1ACXX:1:1101:12096:1998	1:N:0:CGATGT
NTCTGATGTTGCTGATCTCCGTGGCTGTGACCATCATGGCTG	STGACCACACTCCTTCTGCCCAGTTCGGCTGGAAAACTCTGGGAACTGCAGCACGAGAT
+	na kana manana kana kana kana kana kana
#1=DDFFFFHHHHHITTTTTTTTTTTTTTTTTTTTTTTTTTT	12DA&FGTGGGGHGTGGGGTHGGDHEFFDDCD2>CC++>+5>C52C52<>ACCA>A2<8938



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

# **Galaxy Demo**

#### Accessibility

# All tools looks the same

# No command line or programming

Easy to chain tools together into larger analyses

#### 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-SEO

- <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.bar

Max Intron Length: 300000

Min Isoform Fraction:

0.1

Pre MRNA Fraction: 0.15

a a construction and a construction of the

Perform quartile normalization:

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:

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

Use multi-read correct:

No C

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 provide precise reproducibility

Users can add tags and annotations for additional context

#### Galaxy | Published Page | pr

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

#### Galaxy

Published Pages | webb | polar-bears

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

Analyze Data Workflow Shared Data • Visualization

Webb Miller, Stephan C. Schuster, Andreanna J. Welch, Aakrosh Ratan, Oscar C. Bedoya-Reina, Fangging 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. 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-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 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 00 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.

-0¢ 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 800 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.

**F** Galaxy Dataset | bear SAPs 00 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 nported 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 vorkflow (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 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

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 58 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

#### Goecks et al., Genome Biology, 2010

Galaxy Workflow | bear sweep table

0 10

00

00



Galaxy

Datasets

-

0

scaffoldl

scaffold1 scaffold1 scaffold1

scaffold1

scaffoldl

scaffold1 scaffold1

scaffold1

scaffold1

scaffold1 scaffold1 scaffold1 scaffold1

scaffoldl

scaffoldl

scaffold1 scaffold1

scaffold1

ow has 11 columns.

-

hese SNPs, and has 31 columns.

17

### 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

				Workflow Parameter tissue_name
ufflinks (version 0.0.5)				
AM or BAM file of aligned RNA-Seq reads:				
19: MarkDups_Dupes Marked.bam	Input Dataset 🛛 🗱			
lax Intron Length:	output			
00000		Tophat for Illumina 22		
	(	RNA-Seg FASTO file	Filter GFF	data by attribute 🕱
in Isoform Fraction:		Cane Model Apportations	Filter	
0.1	(	Gene model Annotations	out_file1	0 0
re MRNA Fraction:		insertions (bed)		
		deletions (bed)		<u></u>
.15	X	junctions (bed)	Filter	×
erform quartile normalization:		accepted_hits (bam)	Filter	
No 🕄			out_file1	o Dette
emoves top 25% of genes from FPKM denominati		N. Contraction of the second s	X	
anscripts.				
se Reference Annotation:			1	
NO	Variants (hg19)	128.200.000	( chr3	128,135,183 - 128,489,933
erform Bias Correction:	probe_tiled_regions.bed			
No 🔹	Diamona \$750 dam monand	and and a market section		
as detection and correction can significantly imp	Summary			111
	Sample1	1.11		
se multi-read correct:	Disease Reales France Madaste	a make and and		
se multi-read correct:	EVarscan: Roche Exome Variants	In probe regions		
se multi-read correct: No : Ils Cufflinks to do an initial estimation procedui	IVarscan: Roche Exome Variants Summary Sample1	In probe regions		
se multi-read correct: so : IIIS Cufflinks to do an initial estimation procedur prome.	IVarscan: Roche Exome Varlants Summary Sample1 IVarscan: Tophat, dups removed Summary	In probe regions		
se multi-read correct: so	IVansan: Roche Dome Varianta Summary Sample1 IVanson: Tophat, dups removed Summary Sample1	In probe regions		
se multi-read correct: so lis Cutflinks to do an initial estimation procedur mome.	IVancen: Roche Exome Varianta Summary Sample1 IIVanson: Tophat, dups removed Summary Sample1 IITophat2 on data 2, data 145, an 405	In poste regiona In poste regione dista 1: accepted_htm		



#### \varTheta 🔿 🔿 Terminal — bash — 86×22

galaxy-central\$ python scripts/api/workflow\_execute.py AFAD126F http://tachylite01.bx. mathcs.emory.edu 20 33\_

19

#### **Cloud Launch**



20

	*									
A D C A A https://	//main galaxyn	roject ora/	cloudlaunch							A
	//manisganasyp.	Tojection gr	cioudiasisci.							M
- Galaxy	Analyze Data	Workflow	Shared Data -	Visualization +	Cloud +	Admin	Help+	User		Using 668.8 G
Louisek a Calavy	Cloud Inc	1								
Launch a Galaxy	Cloud Ins	tance								
To launch a Galaxy Cloud Clus that using this form to launch	iter, enter your AW	/S Secret Key	ID, and Secret Kr	ey. Galaxy will use	these to pr	resent app	ropriate o	ptions for	/ launching yo	our cluster. Note
nformation.	Comparational rec	Unices in and	/ MINDLOIT STUDE	Win result in costs	to the acce	Auto Instance	HEU BASES	h See Chile	LEVIL 2 MILLING	Tor more
Key ID										
AKIAJ3YLUWEIRJOVADAA	1									
This is the text string that un	niquely identifies y	our account	, found in the Ser	curity Credentials	section of the	he AWS Co	insole.			
Secret Key										
slra1cyCFkfLOiKvsrsGMjyfC	vlYfGs3LSSx2SAs	1								
This is your AWS Secret Key,	also found in the f	Security Cred	Jentials section o	of the AWS Console	É.					
Instances in your account										
New Cluster										
Cluster Name	1									
cluster1		an an private states								
This is the name for your clu	ster, You'll use the	s when you v	want to restart.							
Cluster Password										
•••••										
Cluster Password - Confirm	mation									
••••••										
Key Pair										
Create New - cloudman_keyp	air 📫									
Instance Type										
Large										
Requesting the instance may	/ take a moment, p	lease be pat	ient. Do not refre	esh your browser o	ir navigate (	away from	the page	£		
Submit										





#### **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	Attributes Convert Format Datatype Permissions	History C 🌣
Tophat for Illumina Find splice     junctions using RNA-seq data	Edit Attributes	Small Sample/Treatment Differential Expression Analysis 10.7 MB
for RNA-seq data	Name:	15: Differential @ 0 🕱
<ul> <li><u>Tophat Fusion Post</u> post- processing of</li> </ul>	Differential Transcript Expression Info:	Transcript Expression 2,535 lines
<ul> <li><u>Cufflinks</u> transcript assembly and FPKM (RPKM) estimates for RNA–Seq data</li> </ul>		
<ul> <li><u>Cuffcompare</u> compare assembled transcripts to a reference annotation and track</li> </ul>	Annotation / Notes: None	TCONS_00000001 = NM_001005240 XLOC_( TCONS_00000002 = NM_130760 XLOC_(
Cufflinks transcripts across multiple experiments	Add an annotation or notes to a dataset; annotations are available when a history is viewed.	TCONS_00000003 = NM_130762 XLOC_( TCONS_00000004 = NM_033513 XLOC_(
<ul> <li><u>Cuffmerge</u> merge together several Cufflinks assemblies</li> </ul>	Human Feb. 2009 (GRCh37/hg19) (hç	TCONS_00000005 = NM_004359 XLOC_4
<ul> <li><u>Cuffdiff</u> find significant changes in transcript expression, splicing, and promoter use</li> </ul>	Number of comment lines:	(CON2_00000000 = NM_00231/ XCOC_(
FILTERING	Save	14: Cuffdiff on data 1,
Filter Combined Transcripts     using tracking file	This will inspect the dataset and attempt to correct the above column values if they are not accurate.	13: Cuffdiff on data 1, I I X
NGS: SAM Tools NGS: Variant Detection		data 2, and data 3: transcript differential expression testing
NGS: Peak Calling		12: Cuffdiff on data 1,
NGS: Simulation		data 2, and data 3: gene FPKM
		· ···· · · · · · · · · · · · · · · · ·

#### Galaxy Analyze Data Workflow Shared Data Visualization Admin Help User Using 2.0 TB 0 4 Tools History Scatterplot of 'Differential Transcript Expression' KNA-SEQ Small Sample/Treatment Tophat for Illumina Find splice **Differential Expression Analysis** junctions using RNA-seq data Data Controls 10.7 MB 12 📄 Tophat2 Gapped-read mapper 20,000for RNA-seq data · 1 × 15: Differential Chart Controls Transcript Expression 18,000 -Tophat Fusion Post post-2.535 lines processing of Statistics format: tabular, database: hg19 16,000 <u>Cufflinks</u> transcript assembly 01 Chart 0 and FPKM (RPKM) estimates for **RNA-Seg data** 14,000-1 23 4 <u>Cuffcompare</u> compare TCONS\_0000001 = NM\_001005240 XLOC\_( assembled transcripts to a 12,000 FPKM TCONS 00000002 = NM 130760 XLOC I reference annotation and track Cufflinks transcripts across TCONS\_0000003 = NM\_130762 XLOC\_I time 10,000 multiple experiments TCONS\_00000004 = NM\_033513 XLOC\_I Cuffmerge merge together TCONS\_00000005 = NM\_004359 XLOC I 8,000 several Cufflinks assemblies TCONS\_0000006 = NM\_005317 XLOC\_I Cuffdiff find significant changes 14 + 6,000 in transcript expression, AES splicing, and promoter use 4,000. 14: Cuffdiff on data 1, 00% 15163.4 FILTERING data 2, and data 3: transcript 4994.14 0 **FPKM tracking** Filter Combined Transcripts 2,000 using tracking file 00% 13: Cuffdiff on data 1, NGS: SAM Tools data 2, and data 3: transcript 4,000 8,000 12,000 differential expression testing NGS: Variant Detection Control FPKM NGS: Peak Calling 00% 12: Cuffdiff on data 1, NGS: Simulation data 2, and data 3: gene FPKM < >







### **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/





# The High-throughput Sequencing Era Galaxy

**Cancer Genomics with Galaxy** 

### Personalized Oncology

A Cancer Center Designated by the National Cancer Institute

emory

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 × Varscan × Annotate VCF × Forward FASTQ file Pileup dataset Variants Reverse FASTQ file output (vcf) output (tabular) output (sam) Annotate × Filter × SAM-to-BAM × Input Filter SAM File to Convert output (vcf) out\_file1 output1 (bam) Filter × Annotate with DGI X Mark Duplicate reads × Filter Input SAM/BAM dataset to mark duplicates out\_file1 in output (tabular) out\_file (bam) html\_file (html) 00 Cut × Slice VCF × From Input dataset MPileup × out\_file1 (tabular) Regions BAM file 1 > BAM file output (vcf) output\_mpileup (pileup, bcf) 🗇 📀 output\_log (txt)

#### Workflow Canvas | Exome Variant Calling + Annotation + Drugs ÷ Workflow Parameters sample\_name allele\_freq NHLBI Grand Opportunity Exome Sequencin **Project (ESP)** Map with BWA for Illumina 🗙 000 Genomes Varscan × Annotate VCF Forward FASTQ file Pileup dataset Variants **Reverse FASTQ file** output (vcf) output (tabular) output (sam) Annotate × Filter × SAM-to-BAM × Input Filter SAM File to Convert output (vcf) out file1 output1 (bam) Filter × Annotate with DGI X Mark Duplicate reads × E DRUG GENE INTERACTION DATABASE Filter Input SAM/BAM dataset to mark duplicates out file1 in output (tabular) out\_file (bam) html\_file (html) Cut × Slice VCF × From Input dataset MPileup × out\_file1 (tabular) Regions BAM file 1 > BAM file output (vcf) output\_mpileup (pileup, bcf) 🕤 🤇 output\_log (txt)

#### Integrates private and public data

#### Workflow Canvas | Exome Variant Calling + Annotation + Drugs ÷ Workflow Parameters sample\_name allele\_freq Map with BWA for Illumina × Varscan × Annotate VCF × Forward FASTQ file Pileup dataset Variants **Reverse FASTQ file** output (vcf) output (tabular) output (sam) Annotate × Filter × SAM-to-BAM × Input Filter SAM File to Convert output (vcf) out file1 output1 (bam) Filter × Annotate with DGI X Mark Duplicate reads × Filter Input SAM/BAM dataset to mark duplicates out file1 in output (tabular) out\_file (bam) html\_file (html) Cut × Slice VCF × From Input dataset MPileup × out\_file1 (tabular) Regions BAM file 1 > BAM file output (vcf) output\_mpileup (pileup, bcf) 🕤 🤇 output\_log (txt)

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 highlyexpressed genes:

Found <del>16</del> 6 druggable genes, leading to <del>98</del> 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**



### **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

RIOTATION



### **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



# Thanks! Questions?

**Galaxy** http://galaxyproject.org

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

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

