Introduction to Galaxy

Jeremy Goecks The Galaxy Team http://usegalaxy.org

Session Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

- public website
- local instance
- on the cloud
- + tool shed/contributing tools

Galaxy 101 Exercise

Workshop Coverage

Session 1: Galaxy framework and support for analyses

+ Galaxy is tool-agnostic

Session 2: Galaxy for HTS data analysis

Vision

Galaxy is an open, Web-based platform for accessible, reproducible, and transparent computational biomedical research

What is Galaxy?

GUI for genomics

+ for complete analyses: analyze, visualize, share, publish

A free (for everyone) web service integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

Open source software that makes integrating your own tools and data and customizing for your own site simple

Session Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

- public website
- local instance
- on the cloud
- + tool shed/contributing tools

Galaxy 101 Exercise

Galaxy Analysis Workspace



Filter and Sort

- Filter data on any column using simple expressions
- · Sort data in ascending or descending order
- Select lines that match an expression

GFF FILES

- Extract features from GFF file
- Filter GFF file by attribute using simple expressions
- Filter GFF file by feature count using simple expressions

Extract Features Fetch Sequences Fetch Alignments **Get Genomic Scores Operate on Genomic Intervals** Statistics Graph/Display Data **Regional Variation** Multiple regression **Multivariate Analysis** Evolution Metagenomic analyses EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation NGS: Mapping NGS: SAM Tools NGS: Indel Analysis NGS: Peak Calling

list

RGENETICS

SNP/WGA: Data: Filters SNP/WGA: QC; LD; Plots **SNP/WGA: Statistical Models**

Workflows

xy Analysis Workspace



Filter and Sort

- <u>Filter</u> data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an e) Operate on Genomic Intervals
 - Intersect the intervals of two queries
 - E) Subtract the intervals of two queries
 - Fi Merge the overlapping intervals
 - <u>Concatenate</u> two queries into one query
 - <u>Base Coverage</u> of all intervals
 - <u>Coverage</u> of a set of intervals on second set of intervals
 - <u>Complement</u> intervals of a query
 - <u>Cluster</u> the intervals of a query
 - Join the intervals of two queries side-by-side
 - <u>Get flanks</u> returns flanking region/s for every gene
 - Fetch closest feature for every interval
 - Profile Annotations for a set of genomic intervals

xy Analysis Workspace





C Q. Google

History

0-

Genes

E18

Trimmed

Trimmed

2: E18 PE.2 Reads

1: E18 PE.1 Reads

Sample E18

than ref. base

10: Variants from sample

7: Map with Bowtie for

Illumina on data 6 and data 5

imported: SNP Pileup Analysis for

15: Variants from sample @ 0 22

14: UCSC mm9 RefSeq Genes @ 0 22

13: Variants from sample @ 0 22 E18 where consensus base different

9: Generate pileup on data 8 @ 0 %

8: SAM-to-BAM on data 7 @ 0 20

6: E18 PE.2 Reads Groomed, @ 0 12

5: E18 PE.1 Reads Groomed, @ 0 22

4: E18 PE.2 Reads Groomed @ 0 10

3: E18 PE.1 Reads Groomed @ 0 22

E18, consensus different, in RefSeq

Options -

002

002

002

0002

User

Filter and Sort

Filter	data on any colum	nn using	
simple	e expressions	Filter pileup	
<u>Sort</u> d descer	ata in ascending nding order	Select dataset:	ce
<u>Select</u> e) <u>Op</u> G <u> </u> E) <u> </u> E) <u> </u>	lines that match erate on Genon Intersect the inte queries Subtract the inte queries	which contains: Pileup with six columns (simple) See "Types of pileup datasets" below for examples Do not consider read bases with quality lower than: 20 No variants with quality below this value will be reported	ory Options
^{SI} • <u> </u> <u>Fi</u>	<u>Merge</u> the overla of a query	Do not report positions with coverage lower than: 3	Variants from sample O X , consensus different, in RefSeq 105
u - (NGS: SAM Too Filter SAM o	Only report variants?: Yes See "Examples 1 and 2" below for explanation	UCSC mm9 RefSeq Genes @ 0 % Variants from sample @ 0 % where consensus base different n ref. base
- (<u>Convert SAN</u> SAM to BAN	Convert coordinates to intervals?:	Variants from sample 👁 Ø 🕱 enerate pileup on data 8 👁 Ø 🕱
- (format to BA	Print total number of differences?:	AM-to-BAM on data 7
- (format to SA	See "Example 3" below for explanation Print quality and base string?:	18 PE-2 Reads Groomed. (4) 0 (2) nmed
	 Merge BAM files togethe 	Yes See "Example 4" below for explanation	18 PE.1 Reads Groomed. 40 0 22 nmed 18 PE.2 Reads Groomed 40 0 22
- (<u>Generate pil</u> dataset 	Execute	18 PE.1 Reads Groomed ゆ ク 23 18 PE.2 Reads ゆ ク 33
•	 Filter pileup of SNPs 	aligner designed to be ultrafast and memory-efficient. It is developed by Ben annull Please city: Langment P. Trangell C. Peo M. Salzberg St. Ultrafast and	E18 PE1 Reads 🛛 👁 🖉 🕱
- 1	 <u>Pileup-to-Int</u> pileup format 	erval condenses into ranges of	

Filter and Sort Filter data on any column using simple expressions Filter pileup Sort data in ascending History Options ablaSelect dataset: descending order 10: Variants from sample E18 - Select lines that match ()) 📄 which contains: e Operate on Genon Pileup with six columns (simple) Variant Analysis for Sample E18 Intersect the interview See "Types of pileup datasets" below for examples queries Do not consider read bases with quality lower than: 15: Intersect to get Variants OE from sample E18, consensus different, 20 Subtract the inte aueries in RefSeq Genes No variants with quality below this value will be reported FI si Do not report positions with coverage lower than: Merge the overla 14: UCSC mm9 RefSeg Genes @ 0 💥 3 of a query Fi Pileup lines with coverage lower than this value will be skipped u . 13: Filter to get Variants from @ Ø 💥 NGS: SAM Too Only report variants?: sample E18 where consensus base Filter SAM o Yes 🛟 different than ref. base values See "Examples 1 and 2" below for explanation н. Convert coordinates to intervals?: Convert SAN ш. 10: Filter pileup to get • / X No 🛟 Variants from sample E18 SAM-to-BAN See "Output format" below for explanation format to BA . Print total number of differences?: 9: Generate pileup on data 8 No 🛟 • / X BAM-to-SAM See "Example 3" below for explanation format to SA . 8: SAM-to-BAM on data 7 • 1 X Print quality and base string?: Merge BAM . Yes 🛟 files togethe See "Example 4" below for explanation 7: Map with Bowtie for • / X Generate pil Illumina on data 6 and data 5 . Execute) dataset 000X 6: E18 PE.2 Reads Groomed, · Filter pileup on coverage and • Trimmed **SNPs** ligner designed to be ultrafast and memory-efficient. It is develo pnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL, U ent of short DNA sequences to the human genome. Genome Bic Pileup-to-Interval condenses 5: E18 PE.1 Reads Groomed, • () % pileup format into ranges of Trimmed bases 12

ilter an	d Sort		This dataset is	large a	nd only the firs	t meg	abyte is sh	own bel	ow.			
Filter of	data on any co		<u>5110w all</u> <u>5ave</u>									
Simple	expressions	chr10 chr10 chr10	6882036 6882037 14243075 14943079	A 1424307	A 107 6 G	0 G	60 96	32 0	.\$, 60 60	C 35 t.		
descer	ata in ascendir nding order	chr10 chr10	14465082 14465083	1446508	3 T 34 G	к к	173 144	176 144	60 60	35 GG 35	Optic	ons ≂)
Colores.	lines that was	chr10 chr10 chr10	14465084 14465085 14465252	1446508 1446508	5 T 16 G	Т С	117 70 79	0 0	60 60 60	38 38		
	erate on Geno	chr10 chr10	14465258 14465263	1446525	9 A 34 A	Ă A	137 136	0 0	60 60	46 61		<∕∕ ⊟
1.1	Intersect the in	chr10 chr10 chr10	14465366 14465371 14465410	1446536	7 A 12 G	A G C	101 137	0 0	60 60 60	38 g\$ 50 .\$	nalysis for Sample E18	
4	queries	chr10 chr10	14465447 14465456	1446544 1446545	1 G .8 T .7 G	T G	186 193	0	60 60	65 .\$ 70	sect to get Variants	0 %
E	Subtract the ini	chr10 chr10 chr10	14465465 14465485 14465569	1446546 1446548 1446552	6 T 6 C	T T T	177 129 219	0 129 0	60 60 62	63 .\$ 34 t\$ 84	nple E18, consensus diff	erent,
E	queries	chr10 chr10	14465581 14465586	1446558	2 G 17 C	é c	240 248	0	60 60	84 ,\$ 82 .\$	1 Genes	
si 🛛	Merge the over	chr10 chr10 chr10	14465621 14465658 14465660	1446562 1446565 1446566	2 C 19 C	C C T	134 134	0 0	60 60 62	49 ., 49 ,, 55		0
E	of a query	chr10 chr10	14465691 14465778	1446569	12 G 19 C	Ġ	133 128 89	0	60 60	42 .\$ 34 ,\$. mm9 RefSeq Genes @	0 %
u . (NCS SAM TO	chr10 chr10	14465791 14465881 17445089	1446579	2 G 12 G	G G	104 110	0	60 60	33 ,\$ 41	to got Variants from @	$\square \otimes$
(- Eiltor SAM	chr10 chr10	17445000 17445271 17731269	1744508	2 A 20 T	A T	55 113	0	60 60	34 34 42 ,\$	18 where consensus ba	se 🔅
- 1	values	chr10 chr10 chr10	19928287 19928468	1992828	8 G ,9 C	A T	135 132	135 132	60 60 60	36 AA 35 T\$	than ref. base	
	. Convert SA	chr10 chr10	19928494 19928527	1992849 1992852	5 Ĉ 28 A	T A	138 134	138 0	60 60	37 TI 45 ,,		
	- <u>convert se</u>	chr10 chr10 chr10	19928538 19928543 19928741	1992853 1992854 1992874	.9 G .4 A 10 т	G G T	144 147 80	0 147 0	60 60 60	52 ,\$ 40 Gg 30	pileup to get 💿	0 23
	SAM-to-B/ format to I	chr10 chr10	20799826 28750217	2079982 2875021	,7 G .8 C	Ĝ T	117 138	0 138	60 60	37 ,\$ 37 Ťī	from sample E18	
	normat to r	chr10 chr10 chr10	28750397 28750401 28750423	2875039 2875040 2875042	8 A 12 A 24 C	C A T	154 128 113	211 0 113	60 60 60	64 C\$ 47 ,\$ 35 #4	ate nileun on data 8 @	0.52
- (BAM-to-SA format to 1	chr10 chr10	28750438 28750446	2875043 2875044	9 Å 7 Å	Â G	95 165	0 165	60 60	36 .9 46 G9		~ ~ ~
	Normal to .	chr10 chr10 chr10	28750487 28750512 28750548	2875048 2875051 2875054	8 A .3 G 19 G	A G C	80 220 255	0 0 255	60 60 60	31 72 .\$ 97 C\$	o-BAM on data 7 🛛 👁	0 23
	files toget	chr10 chr10	28750574 28750577	2875057 2875057	5 T 18 T	Ť T	237 234	0	60 60	83 . 82 ,		
	- Conceptor	chr10 chr10 chr10	28750593 28750640	2875057 2875059 2875064	9 T 14 G 11 T	т G С	242 220 165	0 165	ь0 60 60	76 ,9 75 .9 46 Co	ith Bowtie for @	0 %
	dataset	chr10 chr10	28750746 28750766	2875074 2875076	7 G	Á	202 205	202	60 60	58 AA 59 G\$	on data 6 and data 5	
- 1	- Eilter pilou	chrit	20/30/69	2873077		C m	175	1/5	50	6: E18 P	E.2 Reads Groomed.	0 52
-	SNPs	2 on c	overage and		oper designed to be a	Itrafact	and memory of	ficient It is	develo	Trimme	d	0.00
	- Pileun-to-	Intenva	l condenses	ap	nell. Please cite: Langi ent of short DNA sequ	mead B, uences to	Trapnell C, Pop o the human ge	M, Salzberg nome. Geno	g SL. U me Bic			
	pileup forn	nat int	o ranges of							<u>5: E18 P</u>	E.1 Reads Groomed, @	0 %
	bases					1-1	.			Trimme	<u>d</u>	

User Metadata

History	Options 👻
- 3	42 🖻
Variant Analysis for Sampl	e E18
Tags:	
snp x pileup x boy	wtie ×
demo × sample:e18 ×	0 /2
Annotation / Notes: Perform a variant analysis w parameters to identify varian E18 that lie in annotated ge	ith default nts in sample nes.

10: Vari sample 26,742 database Info: 	ants fror E18 regions, f e: mm9	<u>n</u> 'ormat: in	ي te	rva <	0 🗙 al, 2 🖻					
Tags:										
pileup	×)(sam	ple:e18 🔅	x)							
snps >	0 🔏 🛛									
Annotat	ion:									
Find v covera quality	Find variants with coverage >= 30 and quality score >= 20.									
display	at UCSC	<u>main</u> v	iev	N.	in					
GeneTra Current	<u>ck</u> displ	lay at Ens	er	nb						
				_						
1.Chrom	2.Start	3.End	4	5	6 1					
chr10	14042035	14042037	а С	а С	107 1					
chr10	14243073	14243076	c	c	106					
chr10	14465082	14465083	Ť	ĸ	173					
chr10	14465083	14465084	G	к	144					
chr10	14465084	14465085	т	т	117					
)	4 🕨					

Datasources

Upload file from your computer

• FTP support for large datasets

UCSC table browser

BioMart

interMine / modMine

EuPathDB server

EncodeDB at NHGRI

EpiGRAPH server

Tool Suites

Text Manipulation Format Converters Filtering and Sorting Join, Subtract, Group Sequence Tools Multi-species Alignment Tools Genomic Interval Operations Summary Statistics Graphing / Plotting Regional Variation EMBOSS Evolution / Phylogeny RNA-seq ChIP-seq GATK Picard RGenetics ...and more

NGS: QC and manipulation

ILLUMINA DATA

- FASTQ Groomer convert between various FASTQ quality formats
- <u>FASTQ splitter</u> on joined paired end reads
- <u>FASTQ joiner</u> on paired end reads
- <u>FASTQ Summary Statistics</u> by column

ROCHE-454 DATA

- Build base quality distribution
- Select high quality segments
- <u>Combine FASTA and QUAL</u> into FASTQ

AB-SOLID DATA

- <u>Convert</u> SOLID output to fastq
- <u>Compute quality statistics</u> for SOLID data
- <u>Draw quality score boxplot</u> for SOLID data

GENERIC FASTQ MANIPULATION

- <u>Filter FASTQ</u> reads by quality score and length
- FASTQ Trimmer by column
- <u>FASTQ Quality Trimmer</u> by sliding window

Evolution Metagenomic analyses

Human Genome Variation EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation NGS: Mapping

ILLUMINA

- Map with Bowtie for Illumina
- <u>Map with BWA</u> for Illumina

ROCHE-454

- <u>Lastz</u> map short reads against reference sequence
- <u>Megablast</u> compare short reads against htgs, nt, and wgs databases
- Parse blast XML output

AB-SOLID

Map with Bowtie for SOLID

NGS: SAM Tools NGS: Indel Analysis NGS: Peak Calling NGS: RNA Analysis

RGENETICS

<u>SNP/WGA: Data; Filters</u> <u>SNP/WGA: QC; LD; Plots</u> <u>SNP/WGA: Statistical Models</u>

NGS TOOLBOX BETA

NGS: QC and manipulation

NGS: Mapping

NGS: SAM Tools

- <u>Filter SAM</u> on bitwise flag values
- Convert SAM to interval
- <u>SAM-to-BAM</u> converts SAM format to BAM format
- <u>BAM-to-SAM</u> converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- <u>Generate pileup</u> from BAM dataset
- <u>Filter pileup</u> on coverage and SNPs
- <u>Pileup-to-Interval</u> condenses pileup format into ranges of bases
- <u>flagstat</u> provides simple stats on BAM files

NGS: Indel Analysis

NGS: Peak Calling NGS: RNA Analysis

RGENETICS

SNP/WGA: Data; Filters SNP/WGA: QC; LD; Plots SNP/WGA: Statistical Models

NGS: SAM Tools

NGS: Indel Analysis

- Filter Indels for SAM
- <u>Extract indels</u> from SAM
- Indel Analysis

NGS: Peak Calling

- MACS Model-based Analysis of ChIP-Seq
- <u>GeneTrack indexer</u> on a BED file
- <u>Peak predictor</u> on GeneTrack index

NGS: RNA Analysis

RNA-SEQ

- <u>Tophat</u> Find splice junctions using RNA-seq data
- <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>Cuffdiff</u> find significant changes in transcript expression, splicing, and promoter use

FILTERING

 Filter Combined Transcripts using tracking file

Dozens of tools for different HTS applications packaged with Galaxy

VCF Tools

- Intersect Generate the intersection of two VCF files
- <u>Annotate</u> a VCF file (dbSNP, hapmap)
- Filter a VCF file
- <u>Extract</u> reads from a specified region

NGS: Picard (beta)

QC/METRICS FOR SAM/BAM

- BAM Index Statistics
- Sam/bam Alignment Summary Metrics
- Sam/bam GC Bias Metrics
- Estimate Library Complexity
- Insertion size metrics for PAIRED data
- <u>Sam/bam Hybrid Selection</u>
 <u>Metrics</u> For (eg exome) targeted data

BAM/SAM CLEANING

- Add or Replace Groups
- Reorder SAM
- Replace Sam Header
- <u>Paired Read Mate Fixer</u> for paired data
- Mark Duplicate reads

FASTQC: FASTQ/SAM/BAM

 Fastqc: Fastqc QC using FastQC from Babraham

NGS: GATK Tools Alpha REALIGNMENT

- <u>Realigner Target Creator</u> for use in local realignment
- Indel Realigner perform local realignment

BASE RECALIBRATION

- Count Covariates on BAM files
- Table Recalibration on BAM files
- <u>Analyze Covariates</u> perform local realignment

GENOTYPING

 <u>Unified Genotyper</u> SNP and indel caller

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

- public website
- local instance
- on the cloud
- + tool shed/contributing tools

Galaxy 101 Exercise

Data Library "Bushman"

Library Actions 🔻

These are the data underlying the analyses reported in the paper "Complete Khoisan and Bantu genomes from southern Africa" by S. C. Schuster et al., published in the journal Nature, February 18, 2010. Each data set can be downloaded and/or imported into a Galaxy history. Data will be updated as the project progresses.

Name	Information	Uploaded By	Date	File Size
□ <u>All SNPs in personal genomes</u> ▼	Summary table of SNPs in all individuals	greg@bx.psu.edu	2010-01-28	676.8 Mb
□ <u>Alu insertions in KB1</u> ▼		greg@bx.psu.edu	2010-02-10	14.9 Kb
		greg@bx.psu.edu	2010-02-10	6.5 Kb
□ <u>KB1 microsatellites.txt</u> ▼		greg@bx.psu.edu	2010-02-15	3.5 Mb
□ <u>NB1 microsatell tos.txt</u> ▼		greg@bx.psu.edu	2010-02-15	828.5 Kb
☐ amino acid differences ₩th functional predictions		greg@bx.psu.edu	2010-02-05	1.1 Mb
📃 gene copy numbers in 175 and raher personal genoral V		greg@bx.psu.edu	2010-02-15	2.1 Mb
indels in AB1 V		greg@bx.psu.edu	2010-02-03	105.3 Kb
□ indels in KB1 ▼		greg@bx.psu.edu	2010-02-03	14.2 Mb
🗌 indels in MDb 🚀		greg@bx.psu.edu	2010-02-03	109.8 Kb
□ indels in NB1 ▼		greg@bx.;mu.c.tu	2010-02-03	21895 KP
□ indels in TK1 ▼		greg@bx.psu.edu	2010-02-03	123.2 Kb
nove <u>/ SNPs in ABT</u>		greg@bx.psu.edu	2010-02-09	9.4 Mb
□ novel SNPs in KB1 ▼		greg@bx.psu.edu	2010-02-09	16.9 Mb
novel SNPs in: MDE		greg@bx.psu.edu	2010-02-09	594.1 Kb
novel <u>SNPs in NBi</u> S'		greg@bx.psu.edu	2010-02-09	4.1 Mb
□ novel SNPs in TK1 ▼		greg@bx.psu.edu	2010-02-09	722.6 Kb
sequenced exon-containing intervals		greg@bx.psu.edu	2010-02-03	3.1 Mb
For selected items: Import into your current history	Go			

http://usegalaxy.org/bushman

Managing Libraries

Loading Data

- Upload a single file
- Import datasets from a Galaxy history
- Upload a directory of files
- Directly from Sequencer using Sample Tracking System

Accessing Data

- Data contents on disk are not copied
- Dataset security: public, Role-based access control (RBAC)

Annotating Library Data: Library Templates

- Build user fillable forms
- Associate at Library, Folder or Dataset level

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

- public website
- local instance
- on the cloud
- tool shed/contributing tools

Galaxy 101 Exercise

00					Calaxy					
🔺 🕨 🕂 🚱 http:/	/main.g2.bx.psu	.edu/						¢	Q. Google	
💳 Galaxy			Analyze Data	Workflow	Shared Data	Visualizatio	on Help	User		
Tools	Options -								П Ні	stor History Lists
search tools			Show all Sav	s large and oi <u>re</u>	nly the first me	gabyte is show	vn below.			Saved Histories
Get Data Send Data ENCODE Tools Lift-Over Text Manipulation Convert Formats FASTA manipulation Filter and Sort Join, Subtract and Gro Extract Features Fetch Sequences Fetch Alignments Get Genomic Scores Operate on Genomic In Statistics Graph/Display Data Regional Variation Multiple regression Multivariate Analysis	up ntervals	chr10 chr10	6002006 600200 14243075 14245062 14455062 14455062 14455063 14455063 14455063 14455063 14455063 14455257 14455257 14455410 14455410 1445545 14455563 14455563 14455563 14455563 14455563 14455563 14455563 14455563 14455563 14455563 14455563 14455563 14455563 14455563 14455563 14455563 14455563 14455563 14455773 14455773 14455783 14455783 14455783 14455783 14455783 14455783 14455783 14455783 14455783 14455883 17445883 17445883 17445273 1445588 17445273 1445588 17445273 1445588 1445588 14455888 17445273 14455888 14455888 14455888 14455888 14455888 14455888 14455888 14455888 14455888 14455888 14454888 14454888 14458888 144548888 144548888 144548888 144548888 144548888 144548888 144548888 144548888 144548888 144548888 1445488888 144548888 144548888 144548888 144548888 144548888 144548888 144548888 1445488888 1445488888 1445488888 144548888 1445488888 1445488888 1445488888 1445488888 1445488888 144548888	7 h h 14243020 14245026 14455023 14455025 14455025 14455025 14455229 14455229 14455229 14455229 14455248 14455248 14455448 14455448 1445546 14455572 14455572 14455282 144558282 144558282 144558282 14455882 14458882 144558882 1	000 th to caa a go to to coordogaate at at	60 99 105 107 144 144 177 70 79 137 137 137 137 137 137 137 137	32 \$ 0 60 0 60 1144 60 0 60 <th>· </th> <th>1. 1379-1 - 1999</th> <th>Baved Histories P B Histories Shared with Me Current History 0: M Create New mp 6,74 Gamp Clone Share or Publish Extract Workflow Dataset Security Show Deleted Datasets Show Hidden Datasets Show Structure http://doi.org/14465083 Tk 173 http://doi.org/14465083 Tk 173 http://doi.org/14465083 Tk 173 http://doi.org/14465084 Tk465085 T 7 117</th>	· 	1. 1379-1 - 1999	Baved Histories P B Histories Shared with Me Current History 0: M Create New mp 6,74 Gamp Clone Share or Publish Extract Workflow Dataset Security Show Deleted Datasets Show Hidden Datasets Show Structure http://doi.org/14465083 Tk 173 http://doi.org/14465083 Tk 173 http://doi.org/14465083 Tk 173 http://doi.org/14465084 Tk465085 T 7 117
Evolution Metagenomic analyses EMBOSS	L	chr10 chr10 chr10 chr10 chr10 chr10 chr10 chr10	19928525 19928538 19928543 19928741 20799826 28750217 28750397 28750397	19928528 19928539 19928544 19928742 20799627 28750218 28750218 28750398 28750398	4648-60-44	134 144 80 117 138 154 128	0 60 0 60 147 60 0 60 0 60 138 60 211 60 0 60	40 30 37 64 47	99 d	Senerate pileup on O X ata 8 SAM-to-BAM on data O X
NGS: OC and manipula NGS: Mapping NGS: SAM Tools NGS: Indel Analysis NGS: Peak Calling RGENETICS SNP/WGA: Data: Filters SNP/WGA: QC: LD: Plot SNP/WGA: Statistical M	tion i s todels	chri0 chri0	28750423 28750426 20756426 28750426 28750426 28750574 28750574 28750574 28750574 28750577 28755077 28755074 28750756 28750756 28750756 28750757 28750757 28750757 28750757 28750757 28750757 28750813 207550813 28750855 28750855	287534924 201750427 201750447 201750448 201750448 20175048 20175049 20175049 20175057 20175057 20175057 20175057 20175070 20175070 20175070 20175070 20175070 201750904 201750904 201750904 201750904 201750904	tagagortetgovagortovaago vaagoortetgetgattovagoo	113 113 145 220 225 227 227 227 227 227 227 227 227 227	$\begin{array}{cccccccccccccccccccccccccccccccccccc$	57843129788276746899490447552882	sie: e: e: 89899 e: e: e: 8989 ▲ ★ ()	Map with Bowtie for lumina on data 6 and data 5 (073,928 lines, format: sam, atabase: mm9 fo: Sequence file aligned. Ø OUMLE 2.FLAG 9.1 VI-EAS269: 3:1:1449:913 99 cha WI-EAS269: 3:1:1449:913 147 cha WI-EAS269: 3:1:1709:832 99 cha WI-EAS269: 3:1:1709:832 147 cha
Workflows		0)	4 10 1	WI-EAS269:3:1:1422:1087 99 cha

00	Tool		History items created	_
🔺 🕨 🕂 🚱 http://main.	Upload File		1: E18 PE.1 Reads)
- Galaxy	This tool cannot be used in workflows		✓ Treat as input dataset	
Tools Opt				History Lists
search tools	Upload File		2: E18 PE.2 Reads	Saved Histories
Get Data	This tool cannot be used in workflows		Treat as input dataset	Histories Shared with Me
ENCODE Tools				Current History
Lift-Over Text Manipulation	FASTQ Groomer			Create New Clone
Convert Formats	☑ Include "FASTQ Groomer" in workflow		3: E18 PE.1 Reads Groomed	Share or Publish
Filter and Sort				Dataset Security
Join, Subtract and Group	EASTO Groomer			Show Deleted Datasets
Fetch Sequences		►	4: E18 PE.2 Reads Groomed	Show Hidden Datasets
Fetch Alignments	✓Include "FASTQ Groomer" in workflow			Show structure
Get Genomic Scores				Delete
Statistics	FASTQ Trimmer		5: E18 DE 1 Paads Croomed	14465082 14465083 T K 173 :
Graph/Display Data Regional Variation	✓ Include "FASTQ Trimmer" in workflow	Þ	Trimmed	14465083 14465084 G K 144 1 14465084 14465085 T T 117 4
Multiple regression			-	- 24161
Multivariate Analysis Evolution				erate pileup on 👁 🖉 🕱
Metagenomic analyses	FASTQ Trimmer		6: E18 PE.2 Reads Groomed,	
EMBOSS	✓Include "FASTQ Trimmer" in workflow		Trimmed	1-to-BAM on data 👁 🖉 🗱
NGS TOOLBOX BETA				
NGS: QC and manipulation NGS: Mapping	Map with Bowtie for Illumina		7. Man with Develop for Illuming on	a with Bowtie for @ 0 % na on data 6 and data 5
NGS: SAM Tools	☑ Include "Map with Bowtie for Illumina"	►	data 6 and data 5	928 lines, format: sam, ise: mm9
NGS: Peak Calling	in workflow			equence file aligned.
RGENETICS				
SNP/WGA: Data: Filters	SAM-to-BAM			E 2.FLA9 3.1 - 13269:3:1:1449:913 99 cba
SNP/WGA: QC; LD; Plots	Include "SAM-to-BAM" in workflow		8: SAM-to-BAM on data 7	i\$269:3:1:1449:913 147 cha - i\$269:3:1:709:832 99 cha
West-Research Models				15269:3:1:709:832 147 cb
worktiows				13269:3111422:1087 99 CM
	Generate pileup		9: Congrate piloup on data 9	
	☑ Include "Generate pileup" in workflow		5. Generate plieup on data 8	









Example: Workflow for differential expression analysis of RNA-seq using Tophat/ Cufflinks tools



Example: Diagnosing low-frequency heterosplasmic sites in two tissues from the same individual

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- + visualization
- sharing
- Pages

Where you can use and build Galaxy

- public website
- local instance
- on the cloud
- + tool shed/contributing tools

Galaxy 101 Exercise

Visualize

Send data results to external genome browsers

Trackster: Galaxy's genome browser

External Genome Browsers

UCSC Ensembl GBrowse

IGV



Integrative Genomics Viewer (IGV)

1: Sample data	• / ×	000			
format: bam, database: mi Info: uploaded bam file	m9 🧷 💼 <u>t</u>		The application "IGV 1 requesting access to y The digital signature could r	5" from "www.broad our computer. not be verified.	institute.org" is
Binary bam alignments file			Allow all applications fro	m "www.broadinstitute.org	" with this signature
		?	Show Details	Deny	Allow
Mouse mm9 🛟 Chr	1 🗘 chr1:9	98,582,224-98,597,370	IGV Co 音 🛷 🔲		
	qA2 qA4 q	18 qC1.1 qC1.3 d	ąC3 qC4 qD qE1.1 qE2.2	qE3 qF qG1 qH1	qH2.3 qH4 qH6
NAM E DATA FILE DATA FILE DATA TYPE	98,584 kb	98,586 kb 	15 kb 98,588 kb 98,590 kb 	98,592 kb 98,594 kb 	98,596 kb
galaxy_f2979acbfb2c63 75.bam Coverage					i
galaxy_f2979acbfb2c63 75.bam		1	I		
chr1:085.80702					112M of 26.9M
CIII 1.30303/35					1131101 20811

Galaxy

- tool integration framework
- heavy focus on usability
- sharing, publication framework

Trackster

Genome Browser

- physical depiction of data
- visually identify correlations
- + find interesting regions, features

Trackster

View your data from within Galaxy

- No data transfers to external site
- Use it locally, even without internet access

Supports common filetypes

+ BAM, BED, GFF/GTF, WIG

Unique features

- custom genomes
- highly interactive


e o o	Galaxy Publ	ished Visualization GCC2(011-1: Viewing and Navigating	~	(Or Coople
Galaxy	Analyze Data	Workflow Shared Data	Visualization Help User		Google
Published Visualizations jeremy GCC2011 [630,000 UCSC Main on Human: knownGene (chr19) ♥ UCSC Main on Human: all_est (chr19) ♥ UCSC Main on Human: phyloP46wayPrimates (chr19) ♥	-1: Viewing and chr19 640,000	650,000	,719 - 682,581 🔎 🖉	670,000	680,000 Auto (Squish) 👻 Dense 😴 Histogram 😴
-1 h1-hESC Tophat Mapped Reads 🛩					Auto (Squish) 👻
630,000	640,000	650,000	660,000	670,000	680,000
Display a menu					10



But really, why another genome browser

From static browsing to visual analysis

Visual feedback and experimentation needed for complex tools with many parameters

Leverage Galaxy strengths: a very sound model for abstracting interfaces to analysis tools and already integrates an enormous number

Dynamic Filtering



Integrating Tools and Visualization

GCC3: Running Tools (hg19) chr19 1,523,098 - 1,545,232 P P 1,530,000 1,540,000 1,540,000 III UCSC Main on Human: knownGene マ P P P 21 ti_2 P P P P III USSC Main on Human: knownGene マ P P P P III USSC Tophat mapped reads マ P P P P III h-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] マ Cuffinks P Max Intron Length 150000 0.5 P Min Isoform Fraction 0.5 P P P F. 138.1 Cuff I 41, 1055 Cuff I 41, 1055 Cuff I 41, 1055 T. 138.1 Cuff I 41, 1055 Cuff I 41, 1055 Cuff I 41, 1055	GCC3: Running Tools (hg19) chr19 1,523,098 - 1,545,232 IUCSC Main on Human: knownGene	Galaxy	Analyze Data	Workflow	Shared Data	Visualization	Admin	Help	User	
1,530,000 1,540,000 III UCSC Main on Human: knownGene ▼ 22[t1].2 1 10 1 11 1 11 1 12[t1].2 1 12[t1].2 1 12[t1].2 1 11 1 11 1 11 1 12[t1].2 1 11 1 11 1 11 1 11 1 11 1 11 1 11 1 11 1 11 1 11 1 11 1 11 1 11 1 12	1,530,000 1,540,000 III UCSC Main on Human: knownGene ▼ 2215,2 1 2216,2 1 III h1-hESC Tophat mapped reads ▼ III h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] ▼ Cufflinks Max Intron Length 150000 Min Isoform Fraction 0.05 Per MRNA Fraction 0.05 Perform quartile normalization No € F. 138.1 CUFF.139.1 CUFF.148.1 000 CUFF.148.1 000	GCC3: Running Tools (hg19)	chr19		• 1	,523,098 - 1,54	5,232	₽₽		
UCSC Main on Human: knownGen ▼ 221tJ.2 22	UCSC Main on Human: knownGene ▼ 221tj.2 221tj.2 21th.1 221tj.2 21th.2 21th.2 </td <td></td> <td>1,530,000</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>1,540,000</td>		1,530,000							1,540,000
221tj.2 2 2 2 2 2 2 2 2 2 2 2 2 2	21 tj.2 21	UCSC Main on Human: knownGene 👻								
III h1-hESC Tophat mapped reads ▼ III h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] ▼ Cufflinks Max Intron Length 150000 Min Isoform Fraction 0.5 Per MRNA Fraction 0.05 Perform quartile normalization No ♀ F.138.1 CUFF.148.1 CUFF.148.1 CUFF.148.1 CUFF.148.1 CUFF.148.1 CUFF.148.1 CUFF.148.1	III h1-hESC Tophat mapped reads ▼ III h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] ▼ Cufflinks Max Intron Length 150000 Min Isoform Fraction 0.5 Per MRNA Fraction 0.5 Perform quartile normalization No € Fr. 138.1 CUFF.148.1 CUFF.149.1 CUFF.149.1 CUFF.149.1 CUFF.142.1	221tj.2 221tl.1 221tk.2	······		***************************************		· · · · · · · · · · · · · · · · · · ·	· · · · · · · · · · · · · · · · · · ·	·····	
III h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] ▼ Cufflinks Max Intron Length IS0000 Min Isoform Fraction 0.5 Per MRNA Fraction 0.05 Perform quartile normalization No \$ ** <tr< td=""><td>III h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] → Cufflinks Max Intron Length 150000 Min Isoform Fraction 0.5 Perform quartile normalization Run on complete dataset Run on visible region</td><td> h1-hESC Tophat mapped reads 👻</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr<>	III h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] → Cufflinks Max Intron Length 150000 Min Isoform Fraction 0.5 Perform quartile normalization Run on complete dataset Run on visible region	h1-hESC Tophat mapped reads 👻								
CUFF.139.1 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	CUFF.139.1 S>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	IIII n1-nESC assembled transcripts - region=[all], param Cufflinks Max Intron Length Min Isoform Fraction 0.5 Pre MRNA Fraction Run on complete dataset FE 138.1	eters=[150000, 0.5, 0.05, M	vol 🗢						
		>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	•••••••••••••••••••••••••••••••••••••••		CUFF.139.1		CUFF.140 C	0.1 >> UFF.141.1 CUFF.1	•••• 42.1 <mark>>></mark>	•

h1-hESC assembled transcripts - regi	on=[all], parameters=[150	0000, 0.5, 0.05, No] 🔻		
h1-hESC assembled transcripts - regi Jfflinks Max Intron Length	on=[all], parameters=[150	0000, 0.5, 0.05, №] 🛩		
h1-hESC assembled transcripts - regi ufflinks Max Intron Length Min Isoform Fraction	on=[all], parameters=[15(150000 0.05	0000, 0.5, 0.05, No] 👻		
h1-hESC assembled transcripts - regi ufflinks Max Intron Length Min Isoform Fraction Pre MRNA Fraction	on=[all], parameters=[150	0000, 0.5, 0.05, No] 🔻		
h1-hESC assembled transcripts - regi ufflinks Max Intron Length Min Isoform Fraction Pre MRNA Fraction Perform quartile normalization Run on complete dataset Run o	on=[all], parameters=[150 150000 0.05 0.05 No 1 visible region	0000, 0.5, 0.05, No] 👻		

CUFF.3.1						***********************************	***************************************
>>>>>	203	and and the constructions that is an a	e voneaue von	the concerne concerne la	CUFF	.4.1 >>>>>>:	CUFF.6.1 >>
CUFF.3.2						CUFF.5.1 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	CUFF.7.1 >>>>
CUFF.3.3				******************		***********************************	***************************************
	Contraction of the second	5 (L)		54 D.S.			CUFF.8.1 >>>



Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- + sharing
- Pages

Where you can use and build Galaxy

- public website
- local instance
- on the cloud
- + tool shed/contributing tools

Galaxy 101 Exercise

Sharing and Publishing

Sharing and Publishing History 'Variant Analysis for Sample E18'

Making History Accessible via Link and Publishing It

This history is currently restricted so that only you and the users listed below can access it. You can:

Make History Accessible via Link

Generates a web link that you can share with other people so that they can view and import the history.

Make History Accessible and Publish

Makes the history accessible via link (see above) and publishes the history to Galaxy's <u>Published Histories</u> section, where it is publicly listed and searchable.

Sharing History with Specific Users

You have not shared this history with any users.

Share with a user

Back to Histories List

Sharing and Publishing

Sharing and Publishing History 'Variant Analysis for Sample E18'

Making History Accessible via Link and Publishing It

This history accessible via link and published.

Anyone can view and import this history by visiting the following URL:

http://main.g2.bx.psu.edu/u/jgoecks/h/variant-analysis-for-sample-e18 /

This history is publicly listed and searchable in Galaxy's Published Histories section.

You can:

Unpublish History

Removes history from Galaxy's Published Histories section so that it is not publicly listed or searchable.

Disable Access to History via Link and Unpublish

Disables history's link so that it is not accessible and removes history from Galaxy's <u>Published Histories</u> section so that it is not publicly listed or searchable.

Sharing History with Specific Users

You have not shared this history with any users.

Share with a user

Back to Histories List

O O Galaxy Pul	blished Hi	story Variant Analysis for Sample E18	
+ Ohttp://main.g2.bx.psu.edu/u/jgoecks/h/variant-analysis-fo	r-sample-r	e18 C Q+ Google	e ::
- Galaxy Analyze Data W	lorkflow	Shared Data Visualization Help User	
Published Histories jgoecks Variant Analysis for Sample E18			About this History
Galaxy History ' Variant Analysis for Sample E18' Annotation: Perform a pileup analysis with default parameters to identify varia	ants in sam	Import history sple E18.	Author Jaoecks
Dataset		Annotation	Related Histories
1: E18 PE.1 Reads	۹	Forward reads from sample E18.	All aubliched histories
2: E18 PE.2 Reads	æ	Reverse reads from sample E18.	All published histories Published histories by igoecks
3: E18 PE.1 Reads Groomed	۹	Groom reads to convert quality scores from Solexa 1.0 to Solexa 1.3	Rating Community
4: E18 PE2 Reads Groomed	۲	Groom reads to convert quality scores from Solexa 1.0 to Solexa 1.3	(1 rating, 4.0 average) Yours de de de de de d
S: E18 PE.1 Reads Groomed, Trimmed	۹	Trim reads from 3° end to remove low-quality nts.	Tags
6: E18 PE-2 Reads Groomed, Trimmed	Ø	Trim reads from 3' to remove low-quality nts.	Community: snp pileup bowtie demo
7: Map with Bowtie for Illumina on data 6 and data 5	æ	Map paired-end reads with default parameters.	* sample
8: SAM-to-BAM on data 7	۹	Need to convert Bowtie SAM to BAM so that pileup analysis can be performed.	Yours:
9: Generate pileup on data 8	æ	Plieup analysis with default parameters	demo 🗙 sample:e18 🗙 🚑
10: Filter pileup to get Variants from sample E18	Ð	Find variants with coverage >= 30.	
13: Filter to get Variants from sample E18 where consensus base different than ref. base	•	Filter pileup to find variants where the consensus base is different than the reference base.	
14: UCSC mm9 RefSeq Genes	40	UCSC mm9 RefSeq genes.	
15: Intersect to get Variants from sample E18, consensus different, in RefSeq Genes	æ	Variants with consensus different that occur in RefSeq genes.	

000	Galaxy Published	Workflow	SNP variant	detection from	n paired	-end reads	0.0		_
Galaxy	Analyze Data	Workflow	Shared Data	Visualization	Help	User	Q* 60	ogie	
Published Workflows jgoecks SNP variant detection fro	om paired-end reads					- 10 A.N.	1	About this Workflow	
Published Workflows jaoecks SNP variant detection fro Step 6: FASTQ Trimmer FASTQ File Output dataset 'output_file' from step 4 Define Base Offsets as Absolute Values Offset from 3' end 9 Keep reads with zero length False Step 7: Map with Bowtle for Illumina Will you select a reference genome from your history or Use a built-in index Select a reference genome /galaxy/data/apiMel3/bowtie_index/apiMel3 Is this library mate-paired? Paired-end Forward FASTQ file Output dataset 'output_file' from step 5 Maximum insert size for valid paired-end alignments 1 1000 The upstream/downstream mate orientation for valid y Perfor Illumina) Bowtie settings to use Commonly used Suppress the header in the output SAM file	r use a built-in index? (-X) paired-end alignment a	gainst	Trim reads to	remove low-qual	ity bases. eter values			About this Workflow Author jgoecks Related Workflows All published workflows Published w	s *****
Step 8: SAM-to-BAM Choose the source for the reference list Locally cached			Convert Bowti can be run.	e SAM output to E	BAM forma	t so that pileu	p		

 + http://www.elements/alemanner/a	//main.g2.bx.psu.edu/history/list_published	Gala	xy Published	Histories		¢ Q+ Ge	ogle	
Galaxy	Analyze Data	Workflow	Shared Data	Visualization	Help	User		
Published Hist	tories							
search	L Advanced Search							
Name	Annotation		<u>Owner</u>	Communit Rating 1	ΩX.	Community Tags	Last Updated	
<u>Galaxy vs MEGAN</u>	Comparison of Galaxy vs. MEGAN pipeline.		aunl	****	r#	metagenomics megan galaxy	Mar 19, 2010	
<u>metagenomic</u> analysis			aunl	****	r#	(metagenomics) (galaxy)	Mar 19, 2010	
<u>SM 1186088</u>	Datasets correspond to our paper published in Peleg et al. entitled : Altered histone acetylation associated with age-dependent memory impairs Experiment layout: This history contains 4 datas form of BED files of uniquely mapped reads pro chip-seq for histone modifications H4K12ac an mouse hippocampus of 3 months (young) and 1 (old) mice after fear conditioning. For detailed in please refer to supplementary materials and me respective work by peleg et al.	Science by 1 is ment. sets in the duced after d H3K9ac in 16 months nformation thods of the	fischerlab	****	r#r		Apr 19, 2010	
<u>Variant Analysis</u> for Sample E18	Perform a pileup analysis with default paramete variants in sample E18.	ers to identify	jgoecks	****	r it	snp pileup bowtie demo sample	2 minutes ago	
<u>get longest exon</u>			henri	***	nk	chr22 longest marc exon human workshop	Sep 02, 2010	
FASTA to Tabular Test			J	skakaka	rŵ:		Aug 26, 2010	
EKLE			yzc109	***	nk.		Aug 24, 2010	

Sharing Trackster Visualizations

"A picture is worth a 1000 words."

A fully-interactive visualization is worth many more words

📤 🌑 🌒 🗮 Galaxy Published Visualizar 🛪 💽				
← → C fi ③ main.g2.bx.psu.edu/u/jeremy/v/gcc201	1-1-viewing-and-navigating			☆ 🔍
🔁 Galaxy	Analyze Data Workflow Shared Data	Visualization Admin Help User		
Published Visualizations jeremy GCC2011-1: Viewin chr19 0 [1,000,000 UCSC Main on Human: knownGene (chr19) * UCSC Main on Human: all_est (chr19) * 1431 UCSC Main on Human: phyloP46wayPrimates (chr19) * 1 -1 h1-hESC Tophat Mapped Reads * 8732 h1-hESC Cufflinks assembled transcripts *	1,290 - 4,168,475 2,000,000		Auto (Squish) + Auto (coverage histogram) + Auto (coverage histogram) + Histogram + Auto (coverage histogram) + Auto (coverage histogram) + Auto (Squish) + Auto (Squish) +	hor my attention and a second
0 1,000,000	2,000,000	3,000,000	4,000,00	

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

- public website
- local instance
- on the cloud
- tool shed/contributing tools

Galaxy 101 Exercise

A web-based, interactive medium for presenting all aspects of an analysis: data, methods, and results

Conception Conception Conception Conception Conception Conception Conception Conception Conception Conception Conception Conception Conception Published Pages goocks Variant Analysis for sample E18 Defense Conception Author Jarrent Analysis of Embryonic Mouse Brain Tissue Author Jarrent Conception Author Joecks Related Pages Author Joecks Results Related Pages To demonstrate how Galaxy can support accessible, reproducible, and transparent NGS re-sequencing studies, we performed a simple variant analysis experiment. This experiment identifies variants from a set of 4,536,964 RNA-seq reads obtained from sequencing a sample of mm9 brain tissue from day 18 of embryonic development. Author sequencing a sample of mm9 brain tissue form day 18 of embryonic development. Author sequencing a sample of mm9 brain (Orminanity) (Orminanity (Orminanity) (Orminanity (Orminanity (Orminanity) (Orminanity (Orminanity (Orminanity (Orminanity) (Orminanity (Ormin		Galaxy Published Page Variant Analy	rsis for sample E18		
Published Pages jgoecks Variant Analysis for sample E18 About this Page Variant Analysis of Embryonic Mouse Brain Tissue Author Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team Jgoecks Results Related Pages To demonstrate how Galaxy can support accessible, reproducible, and transparent NGS re-sequencing studies, we performed a simple variant analysis experiment. This experiment identifies variants from a set of 4,536,964 RNA-seq reads obtained from sequencing a sample of mm9 brain tissue from day 18 of embryonic development. All published pages Published pages Published pages Published pages by jgoecks The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus baseas Community (0 ratinos, 0.0 average)	- Galaxy	Analyze Data Workflow Shared Data Vi	sualization Help User	12	
Variant Analysis of Embryonic Mouse Brain Tissue Author Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team jgoecks Results Related Pages To demonstrate how Galaxy can support accessible, reproducible, and transparent NGS re-sequencing studies, we performed a simple variant issue from day 18 of embryonic development. Author The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus baseas Author Community (or ratings, 0.0 average) Community (or ratings, 0.0 average) Author	Published Pages jgoecks Variant Analysis fo	r sample E18		About this Page	
To demonstrate how Galaxy can support accessible, reproducible, and transparent NGS re-sequencing studies, we performed a simple variant analysis experiment. This experiment identifies variants from a set of 4,536,964 RNA-seq reads obtained from sequencing a sample of mm9 brain tissue from day 18 of embryonic development. All published pages Published pages Published pages by igoecks The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus baseas Community (0 ratings, 0.0 average)	Variant Analysis of Emb Jeremy Goecks, Anton Nekrutenko, James Taylo	ryonic Mouse Brain Tissue		Author jgoecks Related Pages	ð.
descended have been and a second of the seco	To demonstrate how Galaxy can support accessibl analysis experiment. This experiment identifies va tissue from day 18 of embryonic development. The initial analysis produced support for 27,742 ;	e, reproducible, and transparent NGS re-sequencing studio riants from a set of 4,536,964 RNA-seq reads obtained fro possible variants. Of these possible variants, there are 5,62	es, we performed a simple variant om sequencing a sample of mm9 brain 25 where (a) the consensus baseas	All published pages Published pages by jor Rating Community	pecks *****
Calaxy Dataset Intersect to get Variants from sample E18, consensus different, in RefSeq Genes Variants with consensus different that occur in RefSeq genes.	determined by the MAQ modeldiffers from the r 2796 occur in known RefSeq Genes. These potenti Galaxy Dataset Intersect to Variants w	eference base and (b) read coverage at the base is 30x or al variants are: get Variants from sample E18, consensus different, in ith consensus different that occur in RefSeg genes.	greater. Of these potential variants,	Yours Tags	****
Method In the first step of this analysis, the reads were groomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqsanger. Next, the reads were groomed to convert their quality scores; see [1] for this step's rationale and parameter choices. After grooming and trimming, the reads were mapped using the short-read mapper Bowtie [2]. A pileup analysis using SAMtools [3] was then performed and was filtered to identify variants supported by 30+ reads. The complete analysis is contained in this history:	Method In the first step of this analysis, the reads were gri were trimmed from 36bp to 27bp to exclude base grooming and trimming, the reads were mapped is and was filtered to identify variants supported by	pomed to convert their quality scores from Solexa 1.0 to Si pairs with low quality scores; see [1] for this step's ration ising the short-read mapper Bowtie [2]. A pileup analysis i 30+ reads. The complete analysis is contained in this histo	olexa 1.3/Fastqsanger. Next, the reads lale and parameter choices. After using SAMtools [3] was then performed ory:	Community: none Yours:	
Galaxy History Variant Analysis for Sample E18 Get Perform a pileup analysis with default parameters to identify variants in sample E18. Get	Perform a pileup ana	axy <u>History Variant Analysis for Sample E18</u> lysis with default parameters to identify variants in sar	mple E18.		
Here is a workflow for performing this analysis:	Here is a workflow for performing this analysis:				
Galaxy Workflow Variant identification within annotated genes from NGS PE Data Identify variants in annotated genes from NGS paired-end data.	Galaxy Workflow V Identify va	ariant identification within annotated genes from NGS riants in annotated genes from NGS paired-end data.	PE Data		
References [1] Han, X. et al. Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing. Proceedings of the National Academy of Sciences 106, 12741-12746 (2009).	References [1] Han, X. et al. Transcriptome of embryonic and of Sciences 106, 12741-12746 (2009).	neonatal mouse cortex by high-throughput RNA sequenci	ng. Proceedings of the National Academy		
[2] Langmead, B., Trapnell, C., Yop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009). The formation of CAMPACIENT Control of CAMPACIENT 25, 2020. 2020. 2020.	[2] Langmead, B., Trapnell, C., Pop, M. & Salzberg. Genome Biol 10, R25 (2009).	S.L. Ultrarast and memory-efficient alignment of short D	ve sequences to the human genome.		

)
Galaxy Analyze Data Workflow Shared Data Visualization Help User Published Pages jgoecks Variant Analysis for sample E18 About this About this About this The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus baseas About this	
Published Pages jgoecks Variant Analysis for sample E18 About the The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus baseas Image: Construction of the consensus baseas	
determined by the MAQ modeldiffers from the reference base and (b) read coverage at the base is 30x or greater. Of these potential variants, 2796 occur in known RefSeq Genes. These potential variants are:	nis Page
Galaxy Dataset Intersect to get Variants from sample E18, consensus different, in RefSeg Genes igoecks jgoecks Variants with consensus different that occur in RefSeg genes. Related F	Pages
Method In the first step of this analysis, the reads were groomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqsanger. Next, the reads were trimmed from 36bp to 27bp to exclude base pairs with low quality scores; see [1] for this step's rationale and parameter choices. After grooming and trimming, the reads were mapped using the short-read mapper Bowtie [2]. A pileup analysis using SAMtools [3] was then performed and was filtered to identify variants supported by 30+ reads. The complete analysis is contained in this history:	shed pages d pages by igoecks ity 0.0 average)
Galaxy History Variant Analysis for Sample E18 O (*) Yours Perform a pileup analysis with default parameters to identify variants in sample E18. Tags	索索索索索
8: SAM-to-BAM on data 7 Image: SAM on data 7	ity: none
10: Filter pileup to get Variants from Find variants with coverage >= 30. sample E18 Find variants with coverage >= 30.	
13: Filter to get Variants from sample E18 where consensus base different than reference base. ref. base	
14: UCSC mm9 RefSeq Genes UCSC mm9 RefSeq genes.	
15: Intersect to get Variants from sample E18, consensus different, in RefSeq Variants with consensus different that occur in RefSeq genes. Genes Variants with consensus different that occur in RefSeq genes.	
Here is a workflow for performing this analysis:	
Galaxy Workflow Variant identification within annotated genes from NGS PE Data Identify variants in annotated genes from NGS paired-end data.	
References	

000)	Galaxy	Dataset I	Filter to get	Variants fro	m samp	le E18 v	where consens	us base di	ifferent than ref. ba	se
	+ 😁 http://n	nain.g2.bx.psu.edu/u/jg	oecks/d/fdo	dbcfd1207f	bd2					C Q+ Got	ogle
G	alaxy		Ana	ilyze Data	Workflow	Shared	Data	Visualization	Help	User	
Dataset	Filter to get Va	riants from sample E1	8 where cor	nsensus ba	se different	than ref.	base				About this Dataset
Galaxy	Dataset ' Filte	r to get Variants fro	om sample	e E18 whe	re consen	sus bas	e diffe	rent than re	f, base'	HO	Author
		a più la mana a sua più									
Annotati	on: Filter pileup to	o find variants where the	e consensus	base is diffe	erent than th	e referen	ce base	2			Jgoecks
₽	This dataset is la	rge and only the first m	egabyte is s	hown below	I Show all						
aba10	14465000	14465003			172	176	60	25	00000 0	0000000 00	
chr10	14465083	14465083	G	×	144	144	60	35	GGGGG.G	ggggggggg.gg	and the second second second
chr10	14465485	14465486	c	T	129	129	60	34	tSTTTTT	TTTTTTTTTTTTTTTTT	
chr10	19928287	19928288	G	A	135	135	60	36	AAAAAA	алалалалалалал	
chr10	19928468	19928469	c	Ξ	132	132	60	35	TSttttt	TTEELELEETEETTT	AND A DECIDENT
chr10	19928494	19928495	C	T	138	147	60	37	CaCaCaa	ttttttttttttttttt	And a state of the
chr10	28750217	28750218	c	T	138	138	60	37	TTtttttt	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	
chr10	28750397	28750398	A	c	154	211	60	64	C\$.\$C\$C	\$C\$C\$C\$C\$C\$CCCCeC	Tags
chr10	28750423	28750424	C	T	113	113	60	35	TSELTTT	LELETTATTTTTTTTT	Community and
chr10	28750446	28750447	A	G	165	165	60	46	G\$GGGgg	GGggGGggGGGGGGgg	Community: none
chr10	28750548	28750549	G	C	255	255	60	97	C\$C\$C\$C	\$C\$CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	Yours
chr10	28750746	28750747	G	2	202	202	60	58	AAaaaaa	222222222222222222222222	Tours.
chr10	28750766	28750767	A	G	205	205	60	59	GSqSGSq	SCSqqqqqqqqqqqG	S.4
chr10	28750769	28750770	T	C	175	175	60	49	ccccCcc	CCCCCeCececeCe	*
chr10	28750924	28750925	c	T	182	217	60	64	TSTSttT	tttTtTtTtTtTtTTg	
chr10	28751092	28751093	a	A	147	0	60	123		**********************	
chr10	28/51096	28/5109/	a	2	222	235	60	25	9999,,,	*************	
chr10	28751117	28751118	T	A	191	198	60	79	asasasa	SaSaSaSaSaSaS, S	
chr10	28918972	28918973	C	м	114	114	60	75	,\$,\$,\$,	\$,,,,,,,,,,,,,,	
chr10	28918975	28918976	a	A	177	0	60	63	,\$,\$,\$.	,,C,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	
chr10	28918995	28918996	C	H	154	154	60	48	aşaaaaa	aaaaaaaaaaaaaaa	
chr10	36721501	36721502	G	×	129	129	60	43	. 585858	ŞAAAAAAAAAAAAAAAA	
chr10	36721507	36721508	c	Ŷ	51	51	60	54			
chr10	36721695	36721696	T	A	120	120	60	31	a\$a\$a\$a	SaSaSaSaSaSaSaS	
chr10	36805412	36805413	A	G	126	126	60	33	GEGEGEGE	000000000000000000000000000000000000000	
chr10	36805605	36805606	A	G	120	120	60	31	dddddd	aaaaaaaaaaaaaaa	
chr10	36853176	368531//	D D	P	138	138	60	35			
chr10	36854675	36854676	T	K	99	99	60	48			
chr10	36854678	36854679	A	м	94	94	60	48			
chr10	36855346	36855347	a	R	159	159	60	50			
chr10	36855350	36855351	a	R	156	156	60	56			
chr10	36855366	36855357	a	G	52	0	60	40		**************	
chr10	36855370	36855371	g	G	25	0	60	32		s.s.s	
chr10	36855665	36855666	A	M	11	11	60	39	c\$c\$c\$c	\$c\$,,,	e 1
chr10	39144591	39144592	G	C	126	126	60	33	ccccccc	ccccccccccccc	
chr10	42722938	42722939	c	Y	63	63	60	34	tststst	\$t\$	
0							_			1. 1	

	9	Galaxy	Published	Page Variant /	Analysis for sam	ple E18				
4 >	+ Http://main.g2.bx.psu.edu/u/jgoecks/p/	variant-analysis	-for-sample-	-e18			Ċ	Q+ Google	8	
- 6	alaxy	Analyze Data	Workflow	Shared Data	Visualization	Help U	lser			
Publish The init determi 2796 oc	ed Pages jgoecks Variant Analysis for sample ial analysis produced support for 27,742 possible v ned by the MAQ modeldiffers from the reference ccur in known RefSeq Genes. These potential varian	E18 ariants. Of thes base and (b) re- ts are:	e possible va ad coverage	iriants, there are at the base is 30	5,625 where (a))x or greater. Of t	the consens these potent	us base- tial variar	-as its,	About this Page	1
	Variants with conse	ensus different	that occur i	n RefSeq genes	it, in Kerseg Gen	<u>es</u>			Related Pages	
Meth in the fi were tri groomir and was	od irst step of this analysis, the reads were groomed to mmed from 36bp to 27bp to exclude base pairs wi ng and trimming, the reads were mapped using the s filtered to identify variants supported by 30+ read	convert their q th low quality so short-read map s. The complete	uality scores cores; see [1] oper Bowtie [2 analysis is c	from Solexa 1.0 for this step's r 2]. A pileup ana contained in this) to Solexa 1.3/Fa ationale and para lysis using SAMto history:	stqsanger. M imeter choic ols [3] was t ir	Vext, the es. After then perf	reads ormed	All published pages Published pages by jg Rating Community (0 ratings, 0.0 average) Your	oecks 来来来来来
	Galaxy Hist Perform a pileup analysis wit	ory Variant A h default paran	nalysis for S neters to ide	Sample E18 entify variants i	n sample E18.		Ö Ö		Tags	NNNNN
	8: SAM-to-BAM on data 7 @	Need to con performed. Pileup analy	vert Bowtie S sis with defa	AM to BAM so t ult parameters	hat pileup analysi	is can be	ĺ		Community: none Yours:	
	10: Filter pileup to get Variants from sample E18	Find variant:	s with covera	ge >= 30.						
	13: Filter to get Variants from sample E18 where consensus base different than ref. base	Filter pileup reference ba	to find varia ise.	nts where the co	onsensus base is o	different tha	n the			
	14: UCSC mm9 RefSeq Genes @	UCSC mm9	RefSeq genes	i.						
	15: Intersect to get Variants from sample E18, consensus different, in RefSeg	Variants with	h consensus	different that oc	cur in RefSeq ger	nes.				
Here is	a workflow for performing this analysis:									
	Galaxy Workflow Variant is Identify variants in	lentification wi annotated ger	thin annotat	ed genes from S paired-end d	NGS PE Data ata.		00			
Refer	rences							4		
Open "ht	tp://main.g2.bx.psu.edu/history/imp?id=e0b8bd5d661b10c2	" in a new tab								1





90	Galaxy	
+ http://main.g2.bx.ps	u.edu/page/edit_content?id=d2523e005e1ec427 Rtcader C Q+ Google	
Galaxy	Analyze Data Workflow Shared Data Visualization Help User	
e Editor Title : Variant Analysis fo	r sample E18	(Save) Clo
I × ⁱ × _i ≡ ≡ i ≡ i ≥ ↔ C	🎭 🎭 🖾 🔲 🛛 Paragraph type 👻 Insert Link to Galaxy Object 👻 Embed Galaxy Object 👻	
Variant Analysis	of Embryonic Mouse Brain Tissue	
Jeremy Goecks, Anton Nekrutenk	o, James Taylor, and The Galaxy Team	
Results		
To demonstrate how Galaxy can su identifies variants from a set of 4,5	pport accessible, reproducible, and transparent NGS re-sequencing studies, we performed a simple variant analysis exp 36,964 RNA-seq reads obtained from sequencing a sample of mm9 brain tissue from day 18 of embryonic development	eriment. This experiment t.
The initial analysis produced suppo from the reference base and (b) rea	Int for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus baseas determine Ind coverage at the base is 30x or greater. Of these potential variants, 2796 occur in known RefSeq Genes. These potenti	d by the MAQ modeldiffers al variants are:
Method		
In the first step of this analysis, the exclude base pairs with low quality <u>Bowtie</u> [2]. A pileup analysis using	reads were groomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqsanger. Next, the reads were tri scores; see [1] for this step's rationale and parameter choices. After grooming and trimming, the reads were mapped u SAMtools [3] was then performed and was filtered to identify variants supported by 30+ reads. The complete analysis is	mmed from 36bp to 27bp to sing the short-read mapper contained in this history:
Here is a workflow for performing	this analysis:	
References		
[1] Han, X. et al. Transcriptome of (2009).	embryonic and neonatal mouse cortex by high-throughput RNA sequencing. Proceedings of the National Academy of Sci	iences 106, 12741-12746
[2] Langmend R. Trannell C. Pon	12 A Pulabase P1 Illingford and another affective affective of short PALA	N-1 10 035 (3000)

00	100		Galaxy	8		
Galaxy	_content?id	d=d2523e005e1e	c427	Visualiz	ation Help User	C Q+ Google
age Editor Title : Variant Analysis for sample E18						(Save) (Cla
■ 1 × × 目目目目日 2 个 8 8 回 □	Parag	aph type 👻 🗌 Ins	ert Link to Galaxy Ol	bject 👻 🗌 Er	nbed Galaxy Object 🔻	٦
Variant Analysis of Emb	search		Advanced Search			
Jeremy Goecks, Anton Nekrutenko, James Taylor,		Name		Tags	Last Updated 1	
Results	۷	Variant Analysis	for Sample E18	5 Tags	15 minutes ago	
		Pileup analysis,	sample E18	4 Tags	2 days ago	
To demonstrate how Galaxy can support accessible identifies variants from a set of 4,536,964 RNA-sec	0	Unnamed histor	у	0 Tags	Sep 07, 2010	variant analysis experiment. This experiment bryonic development.
The initial analysis produced support for 27,742 po	0	Unnamed histor	Y.	0 Tags	Dec 17, 2009	baseas determined by the MAQ modeldiffers
from the reference base and (b) read coverage at th		imported: Hsito	ry with ~100 items	5 Tags	Dec 10, 2009	lenes. These potential variants are:
		imported: Galax	y vs MEGAN	0 Tags	Dec 04, 2009	
Method		imported: Galax	y vs MEGAN	2 Tags	Oct 06, 2009	
		imported: Galax	y vs MEGAN	0 Tags	Oct 06, 2009	
In the first step of this analysis, the reads were gro exclude base pairs with low quality scores; see [1]		imported: metag	genomic analysis	0 Tags	Sep 30, 2009	It, the reads were trimmed from 36bp to 27bp to tads were mapped using the short-read mapper
Rowile [2]. A prieup analysis using SAMtools [3] was		imported: Galax	y vs MEGAN	0 Tags	Sep 30, 2009	complete analysis is contained in this history:
		Page: 1 2	Show all histories of	on one page		
Here is a workflow for performing this analysis:		For 1 selected h	istories:			
	Make Make	the selected hist	ories accessible so	that they ca	in viewed by everyone.	
References					(Embed) (Cancel	
[1] Han, X. et al. Transcriptome of embryonic and n (2009).	eonatal m	ouse cortex by h	igh-throughput RN/	A sequencin	g. Proceedings of the N	lational Academy of Sciences 106, 12741-12746
[2] Langmead, B., Trapnell, C., Pop, M. & Salzberg, :	S.L. Ultrafa	ist and memory-	efficient alignment (of short DN	A sequences to the hun	nan genome. Genome Biol 10, R25 (2009).



00					Galaxy				
Galaxy	main.g2.bx.psu.e	edu/page/edit_co	Analyze Data	Workflow	7 Shared Data	Visualization	n Help	Reader C Qr Google	
je Editor Title : Var	iant Analysis for	sample E18						Save)(Clo
$I = s^2 = s_2 \mid \equiv \equiv \equiv$	÷ € ≣ ₽ €	8. 8. 🖬 🖿	Paragraph type	• Insert L	Link to Galaxy Ot	iject 👻 Embed	Galaxy Ob	ject v	
To demonstrate how identifies variants fr	Galaxy can supp om a set of 4,53f	port accessible, r 6,964 RNA-seq r	reproducible, and reads obtained fro	transparent om sequencir	NGS re-sequenting a sample of r	cing studies, we mm9 brain tissu	performed e from day	d a simple variant analysis experiment. This experiment 18 of embryonic development.	
The initial analysis p from the reference b	roduced support lase and (b) read	for 27,742 poss coverage at the	sible variants. Of t base is 30x or gr	these possibl reater. Of the	le variants, ther ese potential var	e are 5,625 whe riants, 2796 occ	ere (a) the ur in know	consensus baseas determined by the MAQ modeldifference of the second s	ers
		Embedde	d Galaxy Datase	t Variants f	from sample E1	8, consensus d	lifferent, i	n RefSeq Genes'	
		[Do no	t edit this block;	Galaxy will f	fill it in with the	annotated data:	set when it	is displayed.]	
In the first step of th exclude base pairs w Bowtie [2]. A pileup	is analysis, the re ith low quality se analysis using SA	eads were groom cores; see [1] for Mtools [3] was t [Do m	ed to convert the r this step's ration then performed ar Embedded G ot edit this block;	ir quality sco iale and para nd was filtere alaxy Histor Galaxy will	ores from Solex. ameter choices. ed to identify va ry 'Variant Piles fill it in with the	a 1.0 to Solexa 1 After grooming riants supported up Analysis for annotated histo	1.3/Fastqs and trimm d by 30+ n Sample E bry when it	anger. Next, the reads were trimmed from 36bp to 27bp ing, the reads were mapped using the short-read mappe eads. The complete analysis is contained in this history: 18" is displayed.]	r
Here is a workflow f	or performing th	is analysis:							
		Embedd	led Galaxy Workf	low 'SNP ide	entification wit	hin annotated	genes fror	n NGS PE Data'	
		[Do not	edit this block; (Galaxy will fi	ill it in with the i	annotated workf	low when	it is displayed.]	
References									
[1] Han, X. et al. Tra (2009).	inscriptome of en							of the National Academy of Sciences 106, 12741-12746	
		nbryonic and neu	onatal mouse con	tex by high-	throughput RNA	sequencing. Pr	roceedings		
[2] Langmead, B., Tr	apnell, C., Pop, N	mbryonic and ne 1. & Salzberg, S.L	onatal mouse cor	tex by high- emory-effici	throughput RN/	A sequencing. Pl	roceedings quences to	the human genome. Genome Biol 10, R25 (2009).	

The power of Galaxy publishing

Galaxy's publishing features facilitate access and reproducibility without any extra leg work

One click grants access to the *actual analysis* you performed to generate your original results

- Not just data access: the full pipeline
- Annotate each step
- Anyone can import your work and immediately reproduce or build on it

Overview

What is Galaxy?

What you can do in Galaxy

- + analysis interface, tools and datasources
- + data libraries
- workflows
- visualization
- sharing
- + Pages

Where you can use and build Galaxy

- public website
- local instance
- on the cloud
- + tool shed/contributing tools

Galaxy 101 Exercise

How to get into Galaxy

Galaxy main site (http://usegalaxy.org)

Public web site, anybody can use

~500 new users per month, ~100 TB of user data, ~130,000 analysis jobs per month, every month is our busiest month ever...

Will continue to be maintained and enhanced, but with limits and quotas

Centralized solution cannot scale to meet data analysis demands

Overview

What is Galaxy?

What you can do in Galaxy

- + analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

- public website
- local instance
- on the cloud
- + tool shed/contributing tools

Galaxy 101 Exercise

Local Galaxy instances (http://getgalaxy.org)

Galaxy is designed for local installation and customization

- Just download and run, completely self-contained
- Easily integrate new tools
- Easy to deploy and manage on nearly any (unix) system
- Run jobs on existing compute clusters

Especially useful for sensitive data

can secure data and abide by regulations

Scale up on existing resources

Move intensive processing (tool execution) to other hosts

Frees up the application server to serve requests and manage jobs

Utilize existing resources

Supports any scheduler that supports DRMAA (most of them)









Running a Production Server

Use a real database server: PostgreSQL, MySQL Run on compute cluster resources External Authentication: LDAP, Kerberos, OpenID Load balancing; proxy support

https://bitbucket.org/galaxy/galaxy-central/wiki/Config/ProductionServer
Lack IT knowledge or resources?

No problem, just use the Cloud

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- + Pages

Where you can use and build Galaxy

- public website
- local instance
- on the cloud
- + tool shed/contributing tools

Galaxy 101 Exercise

Cloud Computing

network accessible compute resources that can be rapidly acquired, configured, and released on demand

Infrastructure as a service

Compute resources provided and configured on demand (compute nodes, storage, network)

Public commercial: Amazon Web Services, Rackspace, ... Build your own: Eucalyptus, Nimbus, OpenStack, ...

When to use the cloud?

Limited informatics expertise or infrastructure Extended or particular resource needs Cannot upload data to a shared resource Need for customization Have oscillating data volume Deploying Galaxy on the AWS Cloud http://usegalaxy.org/cloud

- 1. Open an AWS account (only once)
- 2. Use the AWS Management Console to start a master EC2 instance
- 3. Use the Galaxy CloudMan web interface on the master instance to manage the cluster

2. Start an EC2 Instance



3. Configure Your Cluster



000		Galaxy	Cloud		
< > + S htt	p://ec2-174-129-103-83.compute-1.an	nazonaws.com/cloud		C Qr Google	
- Galaxy				Info: <u>report bu</u>	igs <u>wiki</u> <u>screencasts</u>
	Welcome to Galaxy Cloudman. This a this is your first time running this clu configured, default services will start which jobs are run.	sole pplication will allow you ster, you will need to se and you will be add and	to manage this cloud and t elect an initial data volume s I remove additional services	he services provided within. If size. Once the data store is as well as 'worker' nodes on	
	Status			Process Guidary	
	Disk status: 0 / 0 (0%) 🚱			Starting	
	Worker status: Idle: 0 Availab	ble: 0 Requested: 0		Ready	
	Service status: Applications	Data 鱼		Error	

•

Service status: Applications · Data ·





Can use like any other Galaxy instance, with additional compute nodes acquired and released (*automatically*) in response to usage

+ Thttp://ec2-184-73-135-47.compute-1.amazonaws.com/cloud/ © @ Google AWS Management Console Galaxy Cloud Info: report bugs wiki scree Galaxy Cloudman Console Info: report bugs wiki scree Galaxy Cloudman Console Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove 'worker' nodes for running jobs. Terminate cluster Add nodes • Remove nodes Access Galaxy Status 181M / 100G (1%) @ Autoscaling is off. Turn on? Worker status: Idle: 0 Available: 0 Requested: 0 Idle: 0 Available: 0 Requested: 0
AWS Management Console Galaxy Cloud Galaxy Cloudman Info: report bugs wiki screents Galaxy Cloudman Console Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove 'worker' nodes for running jobs. Terminate cluster Add nodes ▼ Remove nodes Access Galaxy Status Isinstatus: 181M / 100G (1%) ③ Autoscaling is off. Turn on?
Galaxy Cloudman Info: report bugs wiki screent Galaxy Cloudman Console Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove 'worker' nodes for running jobs. Terminate cluster Add nodes ▼ Remove nodes Access Galaxy Status Cluster name: james-cm-31march € Implication (1%) € Autoscaling is off. Turn on? Worker status: Idle: 0 Available: 0 Reguested: 0 Implication (1%) € Implication (1%) €
Galaxy Cloudman Console Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove 'worker' nodes for running jobs. Terminate cluster Add nodes ▼ Remove nodes Access Galaxy Status Image: james-cm-31march S Image: james-cm-31march S Image: james-cm-31march S Autoscaling is off. Turn on? Worker status: Idle: 0 Available: 0 Requested: 0 Image: james-cm-31march S Image: james-cm-31march S Image: james-cm-31march S
Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove 'worker' nodes for running jobs. Terminate cluster Add nodes ▼ Remove nodes Access Galaxy Status Cluster name: james-cm-31march ≤ Image: manage this instance of Galaxy (Mutoscaling is off. Turn on? Worker status: Idle: 0 Available: 0 Requested: 0 Image: manage this instance of Galaxy (Mutoscaling is off. Turn on?
Terminate cluster Add nodes ▼ Remove nodes Access Galaxy Status Cluster name: james-cm-31march ≤ Image: Cluster name: Autoscaling is off. Turn on? Turn o
Cluster name: james-cm-31march S Disk status: 181M / 100G (1%) S Worker status: Idle: 0 Available: 0 Requested: 0
Cluster name: james-cm-31march ≤ Image: Cluster name: Autoscaling is off. Disk status: 181M / 100G (1%) ♀ Image: Cluster name: Autoscaling is off. Worker status: Idle: 0 Available: 0 Requested: 0 Image: Cluster name: Turn on?
Disk status: 181M / 100G (1%) (2 Autoscaling is off. Worker status: Idle: 0 Available: 0 Requested: 0 Turn on?
Worker status: Idle: 0 Available: 0 Requested: 0
Service status: Applications 💿 Data 💿
External Logs: Galaxy Log
Cluster status log



Automation

Cloud instances include all tools available in main Galaxy and more

Tool installation and configuration, image creation, etc, all completely automated and extensible

Same automation approach can be used for configuring tool dependencies for a local Galaxy

VM image with tools (not data) also available, currently at http://s3.amazonaws.com/usegalaxy/UseGalaxy.ova

Overview

What is Galaxy?

What you can do in Galaxy

- + analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- + Pages

Where you can use and build Galaxy

- public website
- local instance
- on the cloud
- tool shed/contributing tools

Galaxy 101 Exercise

The Problem

You have written a Python script to analyze genomic data and you want to share it with command-line averse colleagues

The Galaxy Solution

Solution: Integrate the script as a new Tool into your own Galaxy server

Steps:

- Obtain and install Galaxy source code (GetGalaxy.org)
- Write an XML file describing the inputs and outputs and how to execute the script
- Instruct Galaxy to load the tool

Adding your Own

Write or download a command-line executable

Determine number and kind of

- Input and Output Datasets
- Input Parameters

Construct a descriptive tool configuration XML file

Write a wrapper script, only if required

Tool Configuration

Tool Action - Default tool action should be adequate (Upload tool uses custom tool action)

Tool Command

Inputs

- Action Used by datasource tools
- Parameters

Outputs

Help

Tests

A Basic Tool

```
<tool id="fa gc content 1" name="Compute GC content">
 <description>for each sequence in a file</description>
  <command interpreter="perl">toolExample.pl $input $output</command>
 <inputs>
   <param format="fasta" name="input" type="data" label="Source file"</pre>
 </inputs>
 <outputs>
   <data format="tabular" name="output" />
 </outputs>
 <tests>
   <test>
     <param name="input" value="fa gc content input.fa"/>
     <output name="out_file1" file="fa_gc_content_output.txt"/>
   </test>
 </tests>
 <help>
This tool computes GC content from a FASTA file.
 </help>
            Compute GC content
</tool>
            Source file:
             1: Uploaded FASTA File
                                          +
              Execute
          This tool computes GC content from a FASTA file.
```

Tools

Get Data

MyTools

 <u>Compute GC content</u> for each sequence in a file

Send Data

<section name="MyTools" id="mTools"> <tool file="myTools/toolExample.xml" /> </section>

tool_conf.xml

Cluster			🗋 cluster.xml					
Cluster intervals of: max distance between intervals: min number of intervals per cluster: Return type: Execute Cluster intervals of: (bp) 2 Merge clusters into single intervals Execute		<pre>1 <tool id="gops_cluster_1" name="Cluster"> 2 <description>[[Cluster]] the intervals of a query</description> 3 <command interpreter="python2.4"/> 4 gops_cluster.py \$input1 \$output -1 \$input1_chromCol,\$input1_startC 5 -d \$distance -m \$minregions -o \$returntype 6 7 <inputs> 8 <param format="interval" name="input1" type="data"/> 9 <label>Cluster intervals of</label> 10 11 <param columns<="" help="(bp 12 <label>max distance between intervals</label></pre></td></tr><tr><td> TIP: If your query doe interval format. Use " li="" name="distance" o="" size="5" strand="" type="integer" value="1"/> Screencasts! See Galaxy Interval Oper another window). </inputs></tool></pre>				es not appear in the pulldown menu -> it is not in edit attributes" to set chromosome, start, end, and ration <u>Screencasts</u> (right click to open this link in	1314 <param name<="" td=""/> 15 <label>mi1617<param name<="" td=""/>18<option td="" v<="">19<option td="" v<="">20<option td="" v<="">21<option td="" v<="">22<option td="" v<=""></option></option></option></option></option></label>	e="minregions" size="5" type="integer" value="2"> in number of intervals per cluster e="returntype" type="select" label="Return type"> value="1">Merge clusters into single intervals value="2">Find cluster intervals; preserve comments and value="3">Find cluster intervals; output grouped by clus value="4">Find the smallest interval in each cluster value="5">Find the smallest interval in each cluster
 Maximum distant intervals that will b distance are allowd Minimum interva minimum number less than this mini Merge clusters in entire cluster. Find cluster interva comments in the fi Find cluster interva grouped together. 	the is greatest distance in base pairs allowed between be considered "clustered". Negative values for ed, and are useful for clustering intervals that overlap. Is per cluster allow a threshold to be set on the of intervals to be considered a cluster. Any area with mum will not be included in the ouput. to single intervals outputs intervals that span the vals; preserve comments and order filters out als while maintaining the original ordering and le. vals; output grouped by clusters filters out als, but outputs the cluster intervals so that they are Comments and original ordering in the file are lost.	24 25 26 27 class:: info 28 29 **TIP:** If you 30 31 32 33 **Screencasts!* 34 35 See Galaxy Inte 36 37Screencasts 38 39	omark ur query does not appear in the pulldown menu -> it is n ** erval Operation Screencasts_ (right click to open this l s: http://www.bx.psu.edu/cgi-bin/trac.cgi/wiki/GopsDesc					
Example	Query Find clusters Merge clusters	40 41 **Syntax** 42 43 - **Maximum dis 44 - **Minimum int 45 - **Merge clust 46 - **Find cluste 47 - **Find cluste Line: 87 Column:	stance** is greatest distance in base pairs allowed betw tervals per cluster** allow a threshold to be set on the ters into single intervals** outputs intervals that span er intervals; preserve comments and order** filters out ar intervals: output grouped by clusters** filters out n 8 C XML Soft Tabs: 2 A A					

Input Parameter types

Basic

- Text
- Integer
- Float
- Select
 - Static
 - Dynamic
- Boolean

- Genome build
- Data column
- Data
- Hidden
- Base URL
- File
- Drill down

- Grouping
 - Conditional
 - Repeat
- Config Files

Datasets and Datatypes

All datasets are associated with a Datatype

- + File format
- Type of Data: genomic intervals, sequence, alignment
- Hierarchical structure useful for inputs
- Automatic conversion possible
- Metadata

datatypes_conf.xml and lib/galaxy/datatypes

Adding your Own Display Application

Define An XML configuration which describes how and where to present the data to the External Web Application

- Static
- Dynamic display options can be loaded from a file

Inform Galaxy about the new display by adding to the appropriate datatype in datatypes_conf.xml

https://bitbucket.org/galaxy/galaxy-central/wiki/ExternalDisplayApplications/Tutorial

Static External Display Application

<datatype extension="bam" type="galaxy.datatypes.binary:Bam"
 mimetype="application/octet-stream" display_in_upload="true">
 <display file="ucsc/bam.xml" />
</datatype>

2: SAM-to-BAM on data 1 @ 0 🖄
660.5 Mb, format: bam, database: mm9
Info:
🖬 🕗 🖉 🖻
display at UCSC <u>main</u>
Binary bam alignments file

BAM at UCSC



Dynamic External Display Application

<display id="ucsc_bam" version="1.0.0" name="display at UCSC">
 <!-- Load links from file: one line to one link -->
 <dynamic_links from_file="tool-data/shared/ucsc/ucsc_build_sites.txt" skip_startswith="#" id="0" name="0">

<!-- Define parameters by column from file, allow splitting on builds -->
<dynamic_param name="site_id" value="0"/>
<dynamic_param name="ucsc_link" value="1"/>
<dynamic_param name="builds" value="2" split="True" separator="," />

<!-- Filter out some of the links based upon matching site_id to a Galaxy application configuration parameter and b
<filter>\${site_id in \$APP.config.ucsc_display_sites}</filter>
<filter>\${dataset.dbkey in \$builds}</filter>

<!-- We define url and params as normal, but values defined in dynamic_param are available by specified name --> <url>\${ucsc_link}db=\${qp(\$bam_file.dbkey)}&hgt.customText=\${qp(\$track.url)}</url> <param type="data" name="bam_file" url="galaxy_\${DATASET_HASH}.bam" strip_https="True" /> <param type="data" name="bai_file" url="galaxy_\${DATASET_HASH}.bam.bai" metadata="bam_index" strip_https="True" /> <param type="template" name="track" viewable="True" strip_https="True"> <param type="template" name="track" viewable="True" strip_https="True">

</param>

</dynamic_links> </display>

	2: SAM-to-BAM on data 1 @ 0 🛛
#Harvested from http://genome.ucsc.edu/cgi_bip/das/dap	660.5 Mb, format: bam, database:
<pre>#harvested from http://genome.ucsc.edu/cgi=bin/das/dsh mainbttp://genome.ucsc.edu/cgi=bin/baTracks2apo[ar1_cef_ce1_ce2_ce3_l</pre>	mm9
#Harvested from http://archaea.ucsc.edu/cgi-bin/das/dsn	Info:
archaea http://archaea.ucsc.edu/cgi-bin/hgTracks? therSibi1,symbTher_IAM148	🔲 🔂 🛛 🖉 📄 🖉
<pre>#Harvested from http://main.genome-browser.bx.psu.edu/cgi-bin/das/dsn by_main_bttp://main_genome-browser_by_psu_edu/cgi-bin/baTracks2_oviAri1_eriEu</pre>	display at UCSC main bx-main
bx-main http://main.genome-browser.bx.psu.edu/tgi-bin/ngrhacks/ oviArii,erieu	
	Binary bam alignments file

You added a tool, now what?

Share it with the community!

Galaxy Tool Shed

- Upload and Download contributed tools
- Rate and provide comments and feedback



Get and Contribute Tools

- Galaxy Tool Shed	/ (beta)	Tools Help User	
Community Tools Browse by category	Categories	dvanced Search	
Browse all tools	<u>Name</u> ↓	Description	Tools
Login to upload	Convert Formats	Tools for converting data formats	4
	Data Source	Tools for retrieving data from external data sources	1
	Fasta Manipulation	Tools for manipulating fasta data	5
	Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	5
	Ontology Manipulation	Tools for manipulating ontologies	1
	SAM	Tools for manipulating alignments in the SAM format	0
	Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	7
	SNP Analysis	Tools for single nucleotide polymorphism data such as WGA	1
	Statistics	Tools for generating statistics	1
	Text Manipulation	Tools for manipulating data	3
	Visualization	Tools for visualizing data	1

http://usegalaxy.org/community

Using Galaxy

Use public Galaxy server: UseGalaxy.org Download Galaxy source: GetGalaxy.org Galaxy Wiki: GalaxyProject.org Screencasts: GalaxyCast.org Public Mailing Lists

- galaxy-bugs@bx.psu.edu
- galaxy-user@bx.psu.edu
- galaxy-dev@bx.psu.edu







Enis Afgan



Dave Clements



Dannon Baker



Jeremy Goecks



Kanwei Li



James Taylor



Dan Blankenberg



Jennifer Jackson



Kelly Vincent



Nate Coraor



Greg von Kuster



Anton Nekrutenko

Supported by the **NHGRI** (HG005542, HG004909, HG005133), **NSF** (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health

Overview

What is Galaxy?

What you can do in Galaxy

- + analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

- public website
- local instance
- on the cloud
- + tool shed/contributing tools

Galaxy 101 Exercise

Galaxy 101 http://usegalaxy.org/galaxy101

http://usegalaxy.org/u/jeremy/p/usc-exercise-clusters

A simple question...

 Which coding exons have highest number of single nucleotide polymorphisms?

Galaxy 101 http://usegalaxy.org/galaxy101

Overview

- Interactively Analyze Data
- Create reusable generic Workflow
- Share analysis Results, History, Workflow

Required Data

Genomic Coordinates of coding exons and SNPs

Genomic Coordinates

Genome		
Chromosome		
Gene	Gene Gene	_

http://library.kiwix.org:4201/A/Human_genome.html

>chr1

taaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta

chrom	start	end	name	score strand
chr1	0	10	first_ten_bases	0 +

see also: https://bitbucket.org/galaxy/galaxy-central/wiki/GopsDesc https://bitbucket.org/galaxy/galaxy-central/wiki/zero_based_coordinates.pdf

Galaxy 101: Basic Steps http://usegalaxy.org/galaxy101

Get Genomic data from UCSC Table Browser

Determine each SNP that overlaps with a specific coding exon

Calculate count of overlapping SNPs for each exon

Sort and select exons by greatest SNP counts