An Introduction to Galaxy

Daniel Blankenberg The Galaxy Team http://UseGalaxy.org

Overview

What is Galaxy?

What you can do in Galaxy

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

Galaxy 101 Exercise

The 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

Overview

What is Galaxy?

What you can do in Galaxy

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

Galaxy 101 Exercise

Galaxy Analysis Workspace

00	Calaxy	
◄ ► + Shttp://main.g2.bx.psu.	edu/ c Q+ G	Google
- Galaxy	Analyze Data Workflow Shared Data Visualization Help User	
Galaxy Tools Options Get Data Options Send Data Options ENCODE Tools Ift-Over Text Manipulation Options Convert Formats FASTA manipulation FILter and Sort Join, Subtract and Group 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	Analyze Data Workflow Shared Data Visualization Help User Map with Bowtie for Illumina	History Options imported: SNP Pileup Analysis for Sample E18 15: Variants from sample (*) % E18, consensus different, in RefSeq Genes 14: UCSC mm9 RefSeq Genes (*) % 13: Variants from sample (*) % E18 where consensus base different than ref. base 10: Variants from sample (*) % E18 9: Generate pileup on data 8 (*) % 8: SAM-to-BAM on data 7 (*) % 7: Map with Bowtie for (*) %
EMBUSS NGS TOOLBOX BETA NGS: QC and manipulation NGS: Mapping NGS: SAM Tools NGS: Indel Analysis NGS: Peak Calling RGENETICS SNP/WGA: Data: Filters SNP/WGA: QC: LD: Plots SNP/WGA: Statistical Models	FR (for Illumina) Bowtie settings to use: Commonly used For most mapping needs use Commonly used settings. If you want full control use Full parameter list Suppress the header in the output SAM file: Image: Solution of the several lines of header information by default Execute What it does	6: E18 PE.2 Reads Groomed, ● ∅ ∅ Trimmed ● ∅ ∅ ∅ 4: E18 PE.2 Reads Groomed ● ∅ ∅ ∅ 3: E18 PE.1 Reads Groomed ● ∅ ∅ ∅ 2: E18 PE.2 Reads ● ∅ ∅ ∅ 1: E18 PE.1 Reads ● ∅ ∅ ∅
Workflows	Bowtie is a short read aligner designed to be ultrafast and memory-efficient. It is developed by Ben Langmead and Cole Trapnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10:R25.	A Y

- 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

RGENETICS

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

Workflows

xy Analysis Workspace

Galaxy				
۵) (۵	• Google	2		
Analyze Data Workflow Shared Data Visualization Help User				
with Bowtie for Illumina		History	Ор	tions 👻
you select a reference genome from your history or use a built-in index?: e a built-in index : t-ins were indexed using default options		imported: SNP Pile Sample E18	up Analysis	🖉 🖻 for
nct a reference genome:		15: Variants from E18, consensus d Genes	<u>sample</u> ifferent, in F	⊕ / % tefSeq
his library mate-paired?:		14: UCSC mm9 Re	fSeq Genes	@ / X
Forward FASTQ file: 1: E18 PE.1 Reads		13: Variants from E18 where conser- than ref. base	<u>sample</u> nsus base di	⊕ Ø % fferent
Must have Sanger-scaled quality values with ASCII offset 33 Reverse FASTQ file:		10: Variants from E18	sample	@0%
1: E18 PE.1 Reads Must have Sanger-scaled quality values with ASCII offset 33		9: Generate pileu	o on data 8	• 0 %
Maximum insert size for valid paired-end alignments (-X):		8: SAM-to-BAM o	n data 7	• 0 x
1000 The upstream/downstream mate orientation for valid paired-end alignment against the forward reference strand (fr/rf/fl):		7: Map with Bowt Illumina on data	i <u>e for</u> 6 and data 5	• / ×
FR (for Illumina)		6: E18 PE.2 Reads	Groomed,	@ / X
Bowtie settings to use: Commonly used For most mapping needs use Commonly used settings. If you want full control use Full parameter		5: E18 PE.1 Reads Trimmed	Groomed,	● / X
Suppress the header in the output SAM file:		4: E18 PE.2 Reads	Groomed	• / %
✓ Bowtie produces SAM with several lines of header information by default		3: E18 PE.1 Reads	Groomed	• / %
Execute		2: E18 PE.2 Reads	L.	• / %
What it does		1: E18 PE.1 Reads		• / %
<u>Bowtie</u> is a short read aligner designed to be ultrafast and memory-efficient. It is developed by Ben Langmead and Cole Trapnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10:R25.	4			

u

- <u>Filter</u> data on any column using simple expressions
- <u>Sort</u> data in ascending or descending order
- Select lines that match an e) Operate on Genomic Intervals
 - Intersect the intervals of two queries
 - EI Subtract the intervals of two queries
 - si <u>Merge</u> the overlapping intervals Fi of a query
 - <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
 - · Cluster 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
 - <u>Profile Annotations</u> for a set of genomic intervals

xy Analysis Workspace

	Galaxy							
				Ċ	Q• Go	ogl	e	
Analyze Data Workflow	Shared Data	Visualization	Help	User				
wtie for Illumina							History Op	tions 👻
ect a reference genome from in index re indexed using default option erence genome: me of interest is not listed - co ry mate-paired?: STQ file: Reads anger-scaled quality values with STQ file: Reads anger-scaled quality values with STQ file: Reads anger-scaled quality values with sert size for valid paired-er im/downstream mate orientate erence strand (fr/rf/1 ina) ngs to use: used ipping needs use Commonly u e header in the output SAM for uces SAM with several lines of	a your history of hs intact Galaxy te th ASCII offset 3 th ASCII offset 3 ind alignments (- ation for valid p ff): sed settings. If y file: header informat	r use a built-in am 3 -X): paired-end align you want full cor ion by default	nment ag	jainst the Full param	eter	t.	Imported: SNP Pileup Analysis Sample E18 IS: Variants from sample E18, consensus different, in R Genes Id: UCSC mm9 RefSeq Genes Id: UCSC mm9 RefSeq Genes Id: Variants from sample E18 where consensus base difthan ref. base IO: Variants from sample E18 where consensus base difthan ref. base IO: Variants from sample E18 where consensus base difthan ref. base IO: Variants from sample E18 where consensus base difthan ref. base IO: Variants from sample E18 where consensus base difthan ref. base IO: Variants from sample E18 where consensus base difthan ref. base IO: Variants from sample E18 where consensus base difthan ref. base S. SAM-to-BAM on data 7 IO: Variants from sample E18 SAM-to-BAM on data 7 IIlumina on data 6 and data 5 S. E18 PE.2 Reads Groomed, Trimmed S. E18 PE.1 Reads Groomed G. E18 PE.2 Reads Groomed I.E18 PE.2 Reads Groomed I.E18 PE.2 Reads Groomed	for for for for for for for for
							1: E18 PE.1 Reads	• / ¤
ort read aligner designed to be I Cole Trapnell. Please cite: La ent alignment of short DNA se	e ultrafast and m ngmead B, Trapr equences to the	nemory-efficient nell C, Pop M, Sa human genome.	. It is dev Izberg SI Genome	eloped by L. Ultrafast Biology 10	Ben and):R25.	Ā		
								16

u

- <u>Filter</u> data on any column using simple expressions
- <u>Sort</u> data in ascending or descending order
- Select lines that match an
 e) Operate on Genomic Intervals
 - Intersect the intervals of two queries
 - E = <u>Subtract</u> the intervals of two Fi queries
 - si <u>Merge</u> the overlapping intervals Fi of a query
 - .
 - 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
 - Generate pileup from BAM dataset
 - <u>Filter pileup</u> on coverage and SNPs
 - <u>Pileup-to-Interval</u> condenses pileup format into ranges of bases

xy Analysis Workspace



simple expressions	Filter pileup	
Sort data in ascending descending	Select dataset: 10: Variants from sample E18	ce
Select lines that match e: Operate on Genom G Intersect the inter 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	ory Options -
Fi queries si <u>Merge</u> the overla	No variants with quality below this value will be reported Do not report positions with coverage lower than:	orted: SNP Pileup Analysis for ple E18 <u>Variants from sample</u>
Fi of a query	3 Pileup lines with coverage lower than this value will be skipped Only report variants?:	UCSC mm9 RefSeq Genes (1) (2) (2) (2) (2) (2) (2) (2) (2) (2) (2
Filter SAM o values	Yes See "Examples 1 and 2" below for explanation Convert coordinates to intervals?:	vhere consensus base different n ref. base Variants from sample Variants from sample
SAM-to-BAM format to BA	No See "Output format" below for explanation Print total number of differences?:	ienerate pileup on data 8 👁 0 💥 AM-to-BAM on data 7 🔹 0 2 💥 Aap with Bowtie for 🔹 0 💥
BAM-to-SAM format to SAM	No See "Example 3" below for explanation Print guality and base string?:	mina on data 6 and data 5 18 PE.2 Reads Groomed, @ (/ X nmed
Merge BAM files togethe Generate pil	Yes See "Example 4" below for explanation	18 PE.1 Reads Groomed.
dataset	on coverage and	18 PE.1 Reads Good C X 18 PE.2 Reads Image: Comparent of the second of
SNPs	aligner designed to be ultrafast and memory-efficient. It is developed by Ben apnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and ment of short DNA sequences to the human genome. Genome Biology 10:R25.	
pileup forma bases	t into ranges of	

bases

Filter data on any colu	mn lising
simple expressions	Filter pileup
Sort data in ascending descending order	Select dataset: 10: Variants from sample E18
Select lines that match e) Operate on Genon	which contains: Pileup with six columns (simple)
G Intersect the inter queries	See "Types of pileup datasets" below for examples Do not consider read bases with quality lower than:
E = <u>Subtract</u> the inte Fi queries	20 No variants with quality below this value will be reported
si • <u>Merge</u> the overla Fi of a query	Do not report positions with coverage lower than:
UIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	Pileup lines with coverage lower than this value will be skipped Only report variants?:
■ <u>I</u> values	See "Examples 1 and 2" below for explanation
• <u>Convert SAN</u>	Convert coordinates to intervals?:
 <u>SAM-to-BAR</u> format to B/ 	Print total number of differences?:
BAM-to-SAM format to SAM	No See "Example 3" below for explanation
 J = <u>Merge BAM</u> files togethe 	Print quality and base string?: Yes See "Example 4" below for explanation
• <u>Generate pil</u> dataset	Execute
 Filter pileup SNPs 	on coverage and aligner designed to be ultrafast and memory-efficient. It is dev
 Pileup-to-In pileup forma 	terval condenses at into ranges of

History Ontions =
U 🗆 🧷 🖻
Variant Analysis for Sample E18
<u>15: Intersect to get Variants</u> \textcircled{O}
from sample E18, consensus different, in RefSeq. Cones
14: UCSC mm9 RefSeq Genes @ 0 🛛
13: Filter to get Variants from @ Ø 🖇
sample E18 where consensus base
different than ref. base
10: Filter pileup to get
Variants from sample E18
9: Generate pileup on data 8 👁 🖉 💥
8: SAM-to-BAM on data 7 (10) 1/ 22
7: Map with Bowtie for
lilumina on data o and data 5
6: E18 PE.2 Reads Groomed,
Trimmed
5: E18 PE.1 Reads Groomed, @ Ø 🕸
Trimmed



This dataset is large and only the first megabyte is shown below. Show all I Save

•	Filter data on any co		<u>511017 un</u> (5									
	simple expressions					_				_	- 1	
		chr10	6882036 6882	037 A A	107	0	60	32	\$,	C	· · <u>.</u> ·	
	Cont data in accordin	chr10	14243075	14243076	Ğ	C C	96	0	60	30	τ.	
	<u>Sort</u> data in ascendir	chr10	14465082	14465083	Ť	ĸ	173	176	60	35	ĠĠ	
	descending order	chr10	14465083	14465084	Ĝ	ĸ	144	144	60	35		
	descending order	chr10	14465084	14465085	т	Т	117	0	60	38		
		chr10	14465085	14465086	Ģ	Ģ	70	0	60	38		
	Select lines that mate	chr10	14465257	14465258	ç	ç	79	0	60	42		
	Operate on Geno	chr10	14465258	14465259	A.	А Л	137	0	60 40	46		
	e, operate on deno	chr10	14465366	14465367	Ä	Ä	101	ŏ	60	38	ά\$	alveis for Sam
	Intersect the in	chr10	14465371	14465372	Ğ	Ĝ	137	ŏ	60	50		alysis for sam
	G = <u>interseet</u> the in	chr10	14465410	14465411	G	G	184	0	60	69	.\$.	
	queries	chr10	14465447	14465448	T	T	186	0	60	65	.\$	
	-	chr10	14465456	14465457	G	G	193	0	60	70	ن <u>ہ</u> :	sect to get Varia
•	E - Subtract the int	chr10	14465485	14463466	Ċ	T T	1//	129	60	63 34	.⊅ +¢•	pla E18 conco
	 <u>Subtract</u> the initial 	chr10	14465569	14465570	Ť	Ť	219	0	60	84		iple E16, conse
_	ci queries	chr10	14465581	14465582	Ĝ	Ĝ	240	ŏ	ĞŎ	84	.\$	Cenes
-	ri querres	chr10	14465586	14465587	С	С	248	0	60	82	÷\$	denes
	Si na na	chr10	14465621	14465622	ç	ç	134	0	60	49	- -	
	Merge the over	chr10	14465658	14465659	C	C	134	0	60	49	-	
	ci of a query	chr10	14463660	14463661	T	T G	103	0	60 60	22	- '6 L	mm9 RefSeg G
•	FI Dia query	chr10	14465778	14465779	č	č	89	ő	60	34	ŝ	
	11	chr10	14465791	14465792	Ğ	Ğ	104	ŏ	60	33	÷.	
		chr10	14465881	14465882	G	G	110	0	60	41	- 11	
	NGS: SAM TO	chr10	17445088	17445089	A	A	103	0	60	34	· · · [to get Variants
		chr10	17445271	17445272	A	A	55	0	60	34	· .	10 whore course
	Filter SAM	chr10	17731269	17731270	T G	T A	113	135	60	42	áð I	18 where conse
	Notice 1	chr10	19928468	19928469	č	T	132	132	60	35	T\$	than ref base
	- values	chr10	19928488	19928489	Ă	Â	119	0	60	44		than rei. base
		chr10	19928494	19928495	С	т	138	138	60	37	TT'	
	Convert SA	chr10	19928527	19928528	A	A	134	0	60	45	-24	
		chr10	19928538	19928539	G	6	144	0	60	52	<u>چ</u>	pileup to get
		chr10	19928543	19928544	<u>ዳ</u> ጥ	Մ Մ	147	147	60 60	40	Ggi.	pridup to get
	SAM-to-BA	chr10	20799826	20799827	Ġ	Ġ	117	ñ	60	37	ś	from sample E1
	format to I	chr10	28750217	28750218	č	Ť	138	138	60	37	ŤŤ	
	- i format to r	chr10	28750397	28750398	A	C	154	211	60	64	C\$	
	(chr10	28750401	28750402	A	A	128	0	60	47	_\$	to mileum en de
	BAM-to-SA	chr10	28750423	28750424	ç	T	113	113	60	35	T\$	tte pileup on da
		chr10	28730438	28750439	н 2	e G	90 145	165	БÜ 60	36 46	. \$ G¢u	
	I format to .	chr10	28750487	28750488	Ä	A	80	0	60	31	GΦI	
		chr10	28750512	28750513	Ĝ	Ĝ	220	ŏ	60	72	(\$.	-RAM on data
	A La Morgo PAN	chr10	28750548	28750549	Ğ	ē	255	255	60	97	Ċ\$	0-bAM OII data
	- I - Merge BAN	chr10	28750574	28750575	Т	Т	237	0	60	83	.\$	
	files toget	chr10	28750577	28750578	T	T	234	0	60	82	,ş	
	inco to gett	chr10	28750578	28750579	T	T	242	0	60	76	-,\$	ith Bowtie for
		chr10	28750640	28750641	т Т	č	165	165	60	46		in some for
	Generate p	chr10	28750746	28750747	Ĝ	Ă	202	202	60	58	ĂĂ:	on data 6 and d
	I dataset	chr10	28750766	28750767	Ā	G	205	205	60	59	G\$-	
	uataset	chr10	28750769	28750770	Т	C	175	175	60	49	cei	
			00550505	00550500	m	m	000	<u> </u>			A	

- Filter pileup on coverage and **SNPs**
- <u>Pileup-to-Interval</u> condenses pileup format into ranges of bases

aligner designed to be ultrafast and memory-efficient. It is develo apnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. U ment of short DNA sequences to the human genome. Genome Bic

Trimmed

vsis Options 👻 02 🖻 ple E18 ants 👁 🖉 💥 nsus different, ienes 👁 🖉 🕱 from ensus base • 0 X L<u>8</u> ta 8 • / X • / X 7 • / X lata 5 6: E18 PE.2 Reads Groomed, • 1 × Trimmed 5: E18 PE.1 Reads Groomed, • 1 ×

User Metadata

History Options -
0 🗖 🧷 🖻
Variant Analysis for Sample E18
Tags:
snp x pileup x bowtie x
demo 🗙 sample:e18 🗙 🆧
Annotation / Notes: Perform a variant analysis with default parameters to identify variants in sample E18 that lie in annotated genes

10: Variants from ● Ø X sample E18 26,742 regions, format: interval, database: mm9 Info: □ □ ○ □ ○ □ ○ □ ○								
pileup	× sam	ple:e18	c)					
snps	2							
Annota	or o							
Find v covera quality	Find variants with coverage >= 30 and quality score >= 20.							
display <u>GeneTra</u> <u>Current</u>	y at UCSC <u>ack</u> displ	i <u>main</u> v lay at Ens	iev en	v in nbl				
1.Chrom	2.Start	3.End	4	56 1				
chr10	6882036	6882037	A	A 107				
chr10	14243075	14243076	G	G 96 I				
chr10	14243079	14243080	С	C 106				
chr10	14465082	14465083	Т	к 173 :				
chr10	14465083	14465084	G	K 144 :				
chr10	14465084	14465085	Т	T 117 (
				7411				

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

- <u>FASTQ Groomer</u> 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
- Map with BWA 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

<u>NGS: SAM Tools</u> <u>NGS: Indel Analysis</u> <u>NGS: Peak Calling</u> 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
- <u>Convert SAM</u> to interval
- <u>SAM-to-BAM</u> converts SAM format to BAM format
- <u>BAM-to-SAM</u> converts BAM format to SAM format
- <u>Merge BAM Files</u> 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

- <u>Filter Indels</u> 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
- <u>Sam/bam Alignment Summary</u> <u>Metrics</u>
- 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

 <u>Fastqc: Fastqc QC</u> 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

Galaxy 101 Exercise

Data Library "Bushman"

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
☐ All SNPs in personal genomes	Summary table of SNPs in all individuals	greg@bx.psu.edu	2010-01-28	676.8 Mb
☐ Alu insertions in KB1		greg@bx.psu.edu	2010-02-10	14.9 Kb
🔲 Alu insurvious in 182 V		greg@bx.psu.edu	2010-02-10	6.5 Kb
─ KB1 microsatellites.txt ▼		greg@bx.psu.edu	2010-02-15	3.5 Mb
■ <u>NB1</u> microsatellites.txt		greg@bx.psu.edu	2010-02-15	828.5 Kb
☐ amino acid differences with functional predictions		greg@bx.psu.edu	2010-02-05	1.1 Mb
■ gene copy numbershit(NP3 (undirathen beitsbhalt genorne)		greg@bx.psu.edu	2010-02-15	2.1 Mb
indels in ABT		greg@bx.psu.edu	2010-02-03	105.3 Kb
indels in KB1		greg@bx.psu.edu	2010-02-03	14.2 Mb
□ indels ín MD& ¥		greg@bx.psu.edu	2010-02-03	109.8 Kb
🔲 indels <u>in NB1</u> 🔻		greg@bx.j)/a.c/au	2010-92-03	275/13 Kp
☐ indels in TK1		greg@bx.psu.edu	2010-02-03	123.2 Kb
nove' SNPs in ABT		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 MER V		greg@bx.psu.edu	2010-02-09	594.1 Kb
novel SNPs ir NB1 V		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

Galaxy 101 Exercise

00					Calaxy						
+ http://main.g2.b	x.psu.e	du/						Ċ	Q+ Go	ogle	\supset
🚾 Galaxy			Analyze Data	Workflow	Shared Data	Visualization	Help	User			
Tools Options	-									Histor	
search tools			This dataset Show all Sa	is large and or <u>ve</u>	nly the first me	gabyte is shown	below.			Saved Histories	h
Get Data		chr10	6882036 688203	7.4	107 0	60 32		c	- 11	SNP PI Histories Shared with Me	
Send Data		chr10 chr10	14243075 14243079	14243076 14243080	ê ê	96 0 106 0	60 60	35 35	ŧ.	Current History	
ENCODE Tools		chr10 chr10	14465082 14465083	14465083 14465084	T K G K	173 176 144 144	60 60	35 35	66	10: V. Create New	
Lift-Over Text Manipulation		ohr10 ohr10	14465084 14465085	14465085 14465086	- - -	117 0 70 0	60 60	38 38		Samp Clone	
Convert Formats		chr10 chr10	14465257 14465258	14465258 14465259	¢ ¢	79 0 137 0	60 60	42 46		datab Share or Publish	
FASTA manipulation		chr10 chr10	14465263 14465366	14465264 14465367	ż ż	136 0 101 0	60 60	61 38	ġŝ	Info: Extract Workflow	- 11
Filter and Sort		ohr10 ohr10	14465371 14465410	14465372 14465411	6 6 6 6	137 0 184 0	60 60	50 69	- 5	Extract worknow	-11
Join, Subtract and Group		ohr10 ohr10	14465447 14465456	14465448 14465457	0 0	186 0 193 0	60 60	65 70	.\$	disp Dataset Security	
Extract Features		chr10 chr10	14465465 14465485	14465466 14465486	č Ŧ	129 129	60	63 34	÷\$	Gene Show Deleted Datasets	
Fetch Sequences		chr10 chr10	14465569 14465581	14465570 14465582	ē ē	219 0 240 0	60 60	84 84	36	Show Hidden Datasets	
Fetch Alignments		ohr10 ohr10	14465586 14465621	14465587 14465622	c c	248 0 134 0	60 60	82 49	-8	Show structure	
Get Genomic Scores		chr10 chr10	14465658 14465660	14465659 14465661	ç ç	134 0 153 0	60 60	49 55		ohri0 Delete	
Operate on Genomic Intervals		chr10 chr10	14465691 14465778	14465692 14465779	8 8	128 0 89 0	60 60	42 34	.s ,s	cpull 14542012 14542000 C C 100	
Statistics		chr10 chr10	14465791 14465881	14465792 14465882	6 6 6 6	104 0 110 0	60 60	33 41	,\$	chr10 14465082 14465083 T K 173	
Graph/Display Data		ohr10 ohr10	17445088 17445271	17445089 17445272	à à	103 0 55 0	60 60	34 34		chr10 14465083 14465084 G K 144	
Regional Variation		chr10 chr10	17731269 19928287	17731270 19928288	5 1	113 0 135 135	60 5 60	42 36	a l		
Multiple regression		chr10 chr10	19928468 19928488	19928469 19928489	Č I	132 132 119 0	2 60 60	35 44	75		
Multivariate Analysis		chr10 chr10	19928494	19928495		138 138 134 0	3 60 60	37	77		
Evolution		ohr10 ohr10	19928538	19928539	G G	144 0 147 147	60	52 40	1	data 8	
Metagenomic analyses		chr10 chr10	19920741 20799826	19928742		00 0 117 0	60	30 37	14	-	
EMBOSS		chr10 chr10	28750217	28750218	Č Ť	138 138 154 211	3 60 60	37	11	8: SAM-to-BAM on data @ 0 🕸	: U
NGS TOOLBOX BETA		ohr10 ohr10	28750401 28750423	28750402	Å Å	128 0 113 113	60 8 60	47	-8	Z	
NGS: OC and manipulation		chr10 chr10	28750438 28750446	28750439 28750447	à ò	95 0 165 165	60 5 60	36 46	.\$	7. Man with Bowtie for @ 0 %	,
NGS: Mapping		chr10 chr10	28750487 28750512	28750468 28750513	à à	80 0 220 0	60	31 72	16	Illumina on data 6 and data 5	1
NGS: SAM Tools		chr10 chr10	28750548 28750574	28750549 28750575	6 C	255 255 237 0	5 60	97 83	C\$	9,073,928 lines, format: sam,	
NGS: Indel Analysis		ohr10 ohr10	28750577	28750578	1 1	234 0 242 0	60	82 76	,8	database: mm9	
NGS: Peak Calling		chr10 chr10	28750593 28750640	28750594 28750641	ê ê	220 0 165 165	60 5 60	75 46	18	Info: Sequence file aligned.	
DEENETICS		chr10 chr10	28750746 28750766	28750747 28750767	G Á A G	202 202 205 205	2 60 5 60	58 59	AA GS-		
RUENE (IC3		chr10 chr10	28750769 28750787	28750770 28750788	T C	175 175 255 0	5 60 60	49 90	00 ,8	1.QNAME 2.FLAG 3.	[
SNP/WGA: Data; Filters		ohr10 ohr10	28750797 28750813	28750798 28750814	с с	180 0 195 0	60 60	64 67	,8	HWI-EAS269:3:1:1449:913 99 cl	۵
SNP/WGA: QC; LD; Plots	U	chr10 chr10	28750833 28750835	28750834 28750836	ź ź	152 0 139 0	60 60	53 52	1	NVI-EAS269:3:1:1449:913 147 cl	0
SNP/WGA: Statistical Models	Ă	chr10 chr10	28750860 28750873	28750861 28750874	6 6 C C	101 0 83 0	60 60	38 32	\$ ¥	HVI-EA3269:3:1:709:832 147 el	4
Workflows	Ŧ	0)	4 1	HWI-EAS269:3:1:1422:1087 99 ch	ب د

00	Tool		History items created	
	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 Send Data	This tool cannot be used in workflows		✓ Treat as input dataset	Histories Shared with Me
ENCODE Tools				Current History
Lift-Over	EASTO Croomer			Create New
Text Manipulation	FASTQ Groomer		3: E18 PE.1 Reads Groomed	Clone
Convert Formats FASTA manipulation	✓Include "FASTQ Groomer" in workflow	-		Share or Publish Extract Workflow
Filter and Sort				Dataset Security
Join, Subtract and Group Extract Features	FASTQ Groomer			Show Deleted Datasets
Fetch Sequences	✓ Include "FASTQ Groomer" in workflow		4: E18 PE.2 Reads Groomed	Show Hidden Datasets
Get Genomic Scores	_			Show structure
Operate on Genomic Intervals				142450177 14245000 C C 100 F
Statistics	FASTQ Trimmer		5: E18 PE.1 Reads Groomed.	14465082 14465083 T K 173 :
Graph/Display Data Regional Variation	Sinclude "FASTQ Trimmer" in workflow		Trimmed	14465083 14465085 T T 117
Multiple regression				(A) (P)
Multivariate Analysis				erate pileup on 🔹 🖉 🕱
Metagenomic analyses	FASTQ Trimmer		6: E18 PE.2 Reads Groomed,	
EMBOSS	☑ Include "FASTQ Trimmer" in workflow		Trimmed	I-to-BAM on data 👁 🖉 🗶
NGS TOOLBOX BETA				
NGS: QC and manipulation NGS: Mapping	Map with Bowtie for Illumina			with Bowtie for
NGS: SAM Tools	Include "Map with Rowtie for Illumina"		7: Map with Bowtie for Illumina on	928 lines, format: sam,
NGS: Indel Analysis	in workflow		data 6 and data 5	se: mm9 equence file aligned.
NGS: Peak Calling				Ø 🖻
RGENETICS	SAM to DAM			E 2.FLAG 3.1
SNP/WGA: QC: LD: Plots	SAM-LO-BAM		8: SAM-to-BAM on data 7	S269:3:1:1449:913 99 cha S269:3:1:1449:913 147 cha
SNP/WGA: Statistical Models	✓Include "SAM-to-BAM" in workflow	-		\$269:3:1:709:832 99 cha
Workflows				sze9:3:1:709:832 147 chu * \$269:3:1:1422:1087 99 chu *
	Cenerate nileun			11
	Generate pneup		9: Generate pileup on data 8	
	Include "Generate pileup" in workflow			



00)	Tool		History items created				
	000			Galaxy				
- <u> </u>		p://main.g2.bx.psu.edu/workflow/edito	r?id=a6d94f12f42c1	ars		C Q+ Google		
Tools search	Galaxy	Analyze C	Data Workflow	Shared Data Visualization H	lelp	Tool: SAM-to-BAM		
Get Dat Send Di ENCOD	Input dataset	×				Choose the source for the reference list		
Lift-Ov Text Ma Convert	output	EASTO Groomer 92				SAM File to Convert Data input 'input1' (sam)		
FASTA Filter an Join, Su	Ģ	File to groom				Edit Step Actions		
Extract Fetch S Fetch A		(fastqsanger,fastqcssanger,fastqsolexa	,fastqillumina)	Map with Bowtie for Illumin	a 🐹	Assign Columns 🗘		
Get Ger Operate Statistic		FASTQ Trimmer		Forward FASTQ file Reverse FASTQ file		Add actions to this step; actions are applied when this workflow step completes.		
Graph/I		output_me	/	output (sam)	9	completes.		
Multiple, Multiva				SAM-to-BAM	M Convert	Edit Step Attributes		
Evolutio	Input dataset	22		output1 (barr	n)	Annotation / Notes:		
Metage EMBOSS NGS TO	output	EASTO Granmar 22			Gener	Convert Bowtie SAM output to BAM format so that pileup can be run.		
NGS: 08 NGS: Ma NGS: SA	Ģ	File to groom	formilluming)	<i>Y</i>	genera file fo outpu	Add an annotation or notes to this step; annotations are available when a workflow is viewed.		
NGS: Pe RGENET SNP/WC SNP/WC SNP/WC Workflo		FASTQ File output_file	,rastqiiumina)					

00	Tool	History items created	
	00	Calaxy	
G	+ http://main.g2.bx.	Edit Workflow Attributes	C Q- Google
Tools search Get Dat Send Di ENCOD Lift-OV Text Mi Convert FASTA Filter a Join, Su Extract Fetch A Get Ger Operator	Galaxy Workflow Canvas SNP variant de Input dataset & output FASTQ Groom File to groom output_file (fastqsanger,fa	 Name: SNP identification within annotated genes from NGS PE Data Tags: snp × ngs × pileup × bowtie × Apply tags to make it easy to search for and find items with the same tag. Annotation / Notes: 	Tool: SAM-to-BAM Choose the source for the reference list Locally cached + SAM File to Convert Data input 'input1' (sam) Edit Step Actions Assign Columns + output1 + Create Add actions to this step; actions are
Statistic Graph/I Regiona Multiple Multiva		Identify variants in annotated genes from NGS paired-end data.	applied when this workflow step completes.
Evolutie Metage EMBOSS NGS TO NGS: OC NGS: OC NGS: Mi NGS: SA NGS: In NGS: Pe RGENET SNP/WC SNP/WC SNP/WC	Input dataset 😒 output FASTQ Groom File to groom output_file (fastqsanger,fa	Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.	Annotation / Notes: Convert Bowtie SAM output to BAM format so that pileup can be run. Add an annotation or notes to this step; annotations are available when a workflow is viewed.



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

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)



000							IGV							
Mouse mm9	\$	Chr1	¢ ch	r1:98,58	82,224-98,5	97,370	Go	🖆 🖗				<u> </u>		
		qA2	qA4	qB	qC1.1 qC1	.3 qC3	qC4	qD qE1.1	qE2.2	qE3	qF q	61 qH1	qH2.3 q	H4 qH6
	NAME DATA FILE DATA TYPE		98,584 kb 	I	98,586 kb 	98,5 	88 kb 	— 15 kb - 98,590 k 	b 	98,592 kb 	9	8,594 kb 	98,5: 	96 kb
galaxy_f2979acbfb2c63 75.bam Coverage galaxy_f2979acbfb2c63 75.bam		[0 - 10] 	<u>1</u> <u>1</u> <u>1</u>	1				1 - I 						
												-		6.0001
chr1:98589793													113M c	of 268M

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



Http://main.g2.bx.psu.edu/u/ju Galaxy Published Visualizations jeremy GCC2011- 630,000 CSC Main on Human: knownGene (chr19) + CSC Main on Human: all_est (chr19) + CSC Main on Human: phyloP46wayPrimates (chr19) +	jeremy/v/gcc2011-1-view Analyze Dat 1: Viewing and chr19 640,000	ing-and-navigating ta Workflow Share 650,000	d Data Visualization Help Use 625,719 - 682,581 Degree 20 660,000	¢	Ger Google
CSC Main on Human: phyloP46wayPrimates (chr19) →	Analyze Dai	ta Workflow Share	d Data Visualization Help Use 625,719 - 682,581 🔎 🔎 660,000	er 670,000	680,000 Auto (Squish) 🚽 Dense 🚽 Histogram 🚽
Published Visualizations jeremy GCC2011- 630,000 CSC Main on Human: knownGene (chr19) - CSC Main on Human: all_est (chr19) - CSC Main on Human: phyloP46wayPrimates (chr19) -	1: Viewing and chr19 640,000	650,000	625,719 - 682,581 660,000	670,000	680,000 Auto (Squish) 👻 Dense 🛩 Histogram 😴
630,000 CSC Main on Human: knownGene (chr19) → CSC Main on Human: all_est (chr19) → CSC Main on Human: phyloP46wayPrimates (chr19) →	640,000 = = =	650,000		670,000	680,000 Auto (Squish) 👻 Dense 👻 Histogram 👻
CSC Main on Human: all_est (chr19) \scillenge CSC Main on Human: phyloP46wayPrimates (chr19) \scillenge	= :: - · · ·		•		Dense 🛩
CSC Main on Human: all_est (chr19) 🚽					Dense 🛩 Histogram 👻
CSC Main on Human: phyloP46wayPrimates (chr19) 🗢					Histogram 👻
CSC Main on Human: phyloP46wayPrimates (chr19) 🗢					Histogram 👻
1					
1					Auto (Squish) -
					Hata (Silaisi) 4
630,000	640,000	650,000	660,000	670,000	680,000
Display a menu					



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

Galaxy	Analyze Data	Workflow	Shared Data	Visualization	Admin	Help	User	
GCC3: Running Tools (hg19)	chr19		•	1,523,098 - 1,545	,232	ÐÐ		
	1,530,000							1,540,000
UCSC Main on Human: knownGene 👻								
221tj.2 221tl.1 221tk.2	·····				·····	••••••••••••••••••••••••••••••••••••••	••••••••••••••••••••••••••••••••••••••	
h1-hESC Tophat mapped reads 🗢								
h1-hESC assembled transcripts - region=[all], parameters=[1	50000, 0.5, 0.05, N	lo] 🔻						
Cufflinks								
Max Intron Length 150000								
Min Isoform Fraction 0.5								
Pre MRNA Fraction 0.05								
Perform quartile normalization								
(Run on complete dataset) (Run on visible region)								
FF.138.1; ; ;;;;;;;;;;;;;;;;;;;;;;;;;;;;	******	·····	CUFF.139		CUFF.140 CUFF.140	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	**************************************	4





Overview

What is Galaxy?

What you can do in Galaxy

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

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

Galaxy i	Published	History Variant Analysis for Sa	ample E18	e C	Coogle		
- Galaxy Analyze Data	Workflow	v Shared Data Visualization	Help	User	, ooogie		
Published Histories jgoecks Variant Analysis for Sample E18						About this History	
Galaxy History ' Variant Analysis for Sample E18'	riants in s	sample F18.		G Import	history	Author	
Dataset		Annotation				jgoecks	
1: E18 PE.1 Reads	۲	Forward reads from sam	ple E18.			Related Histories	
2: E18 PE.2 Reads	æ	Reverse reads from sam	ple E18.			All published histories Published histories by jgoecks	
3: E18 PE.1 Reads Groomed	Ð	Groom reads to convert	quality score	es from Solex	a 1.0	Rating	
4: E18 PE.2 Reads Groomed	۹	Groom reads to convert	quality score	es from Solex	a 1.0	Community (1 rating, 4.0 average) Yours	
5: E18 PE.1 Reads Groomed, Trimmed	æ	Trim reads from 3' end	to remove lo	w-quality nts	s.	Tags	
6: E18 PE.2 Reads Groomed, Trimmed	æ	Trim reads from 3' to re	move low-q	uality nts.		Community: snp pileup bowtie demo	
7: Map with Bowtie for Illumina on data 6 and data 5	Ð	Map paired-end reads w	with default p	arameters.		 sample 	
8: SAM-to-BAM on data 7	۲	Need to convert Bowtie analysis can be perform	SAM to BAM led.	so that pileu	p	Yours: snp x pileup x bowtie x	
9: Generate pileup on data 8	Ð	Pileup analysis with defa	ault paramete	ers		demo 🗙 sample:e18 🗙 🗸	
10: Filter pileup to get Variants from sample E18	æ	Find variants with covera	age >= 30.				
13: Filter to get Variants from sample E18 where consensus base different than ref. base	۹	Filter pileup to find varia is different than the refe	ants where th erence base.	he consensus	base		
14: UCSC mm9 RefSeq Genes	Ð	UCSC mm9 RefSeq gene	25.				
15: Intersect to get Variants from sample E18, consensus different, in RefSeq Genes	æ	Variants with consensus genes.	different that	at occur in Re	efSeq		

Calaxy Published Workflow	SNP variant detection from paired-end reads		
Http://main.g2.bx.psu.edu/u/jgoecks/w/snp-variant-detection	n-from-paired-end-reads C	Q. Go	ogle
Galaxy Analyze Data Workflow	Shared Data Visualization Help User		
Published Workflows jgoecks SNP variant detection from paired-end reads			About this Workflow
Step 6: FASTQ Trimmer FASTQ File Output dataset 'output_file' from step 4 Define Base Offsets as Absolute Values	Trim reads to remove low-quality bases.		Author Jgoecks Related Workflows All published workflows
Offset from 5' end 0 Offset from 3' end 9 Keep reads with zero length False			Published workflows by igoecks Rating Community (0 ratings, 0.0 average) Yours Tags
Step 7: Map with Bowtie for Illumina Will you select a reference genome from your history or use a built-in index? 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 6 Reverse FASTQ file Output dataset 'output_file' from step 5 Maximum insert size for valid paired-end alignments (-X) 1000	Map reads using default parameter values.		Community: snp bowtie Yours: snp × bowtie ×
The upstream/downstream mate orientation for valid paired-end alignment against the forward reference strand (fr/rf/) FR (for Illumina) Bowtie settings to use Commonly used Suppress the header in the output SAM file True Step 8: SAM-to-BAM Choose the source for the reference list Locally cached	Convert Bowtie SAM output to BAM format so that pileup can be run.		

00			Gala	xy Published H	listories				
	+ 🕙 http:	//main.g2.bx.psu.edu/history/list_published	d				୯ ଦ୍ର	ogle	
🗧 Ga	laxy	Analyz	e Data Workflow	Shared Data	Visualization	Help	User		
Publish search	ned Hist	Cories							
Name		Annotation		Owner	Communit Rating 1	¥	Community Tags	Last Updated	
<u>Galaxy v</u>	s MEGAN	Comparison of Galaxy vs. MEGAN pipeline	ŀ.	aunl	****	r#r	metagenomics megan galaxy	Mar 19, 2010	
metageno analysis	omic			aunl	****	r#r	metagenomics galaxy	Mar 19, 2010	
<u>SM 1186</u>	088	Datasets correspond to our paper publish Peleg et al. entitled : Altered histone acety associated with age-dependent memory in Experiment layout: This history contains 4 form of BED files of uniquely mapped read chip-seq for histone modifications H4K12 mouse hippocampus of 3 months (young) (old) mice after fear conditioning. For deta please refer to supplementary materials as respective work by peleg et al.	ed in Science by vlation is mpairment. datasets in the ds produced after eac and H3K9ac in and 16 months ailed information nd methods of the	fischerlab	****	rik		Apr 19, 2010	
<u>Variant A</u> for Samp	<u>nalvsis</u> le E18	Perform a pileup analysis with default par variants in sample E18.	ameters to identify	jgoecks	****	nk	snp pileup bowtie demo sample	2 minutes ago	
get longe	tst exon			henri	statata	nk	chr22 longest marc exon human workshop	Sep 02, 2010	
FASTA to Test	Tabular			'n	****	nk		Aug 26, 2010	
EKLF	(main of here	er adu distan dist auklish finan min filmen di	P is a source	yzc109	****	nk		Aug 24, 2010	

Sharing Trackster Visualizations

"A picture is worth a 1000 words."

A fully-interactive visualization is worth many more words



Overview

What is Galaxy?

What you can do in Galaxy

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

Galaxy 101 Exercise

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

😝 🔿 🧑 🖉 Galaxy Publ	lished Page Variant Analysis for sample	e E18		
+ Ohttp://main.g2.bx.psu.edu/u/jgoecks/p/variant-analysis-for-s	ample-e18	¢ Q• Goo	gle	
Galaxy Analyze Data Work	kflow Shared Data Visualization	Help User		
Published Pages jgoecks Variant Analysis for sample E18			About this Page	
Variant Analysis of Embryonic Mouse	Brain Tissue	Í	Author jgoecks	
Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team			Related Pages	
Results To demonstrate how Galaxy can support accessible, reproducible, and transparent	nt NGS re-sequencing studies, we perform	ied a simple variant	All published pages Published pages by jo	ioecks
analysis experiment. This experiment identifies variants from a set of 4,536,964 tissue from day 18 of embryonic development.	RNA-seq reads obtained from sequencing	g a sample of mm9 brain	Rating	
The initial analysis produced support for 27,742 possible variants. Of these poss determined by the MAQ modeldiffers from the reference base and (b) read cov	e consensus baseas ese potential variants.	Community (0 ratings, 0.0 average)	*****	
2796 occur in known RefSeq Genes. These potential variants are:			Yours	*****
+ Galaxy Dataset Intersect to get Variants from sample EI Variants with consensus different that of	8, consensus different, in RefSeq Genes occur in RefSeq genes.	. 	Tags	
			Community: none	
Method			» 🖧	
In the first step of this analysis, the reads were groomed to convert their quality were trimmed from 36bp to 27bp to exclude base pairs with low quality scores; grooming and trimming, the reads were mapped using the short-read mapper Ba and was filtered to identify variants supported by 30+ reads. The complete analy	scores from Solexa 1.0 to Solexa 1.3/Fast see [1] for this step's rationale and param owtie [2]. A pileup analysis using SAMtool: sis is contained in this history:	qsanger. Next, the reads eter choices. After s [3] was then performed		
Galaxy History Variant Analysi Perform a pileup analysis with default parameters	s for Sample E18 to identify variants in sample E18.	• •		
Here is a workflow for performing this analysis:				
Galaxy Workflow Variant identification within a Identify variants in annotated genes fro	nnotated genes from NGS PE Data om NGS paired-end data.	• •		
References				
 Han, X. et al. Transcriptome of embryonic and neonatal mouse cortex by high of Sciences 106, 12741–12746 (2009). 	h-throughput RNA sequencing. Proceeding	gs of the National Academy		
[2] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-eff Genome Biol 10, R25 (2009).	licient alignment of short DNA sequences	to the human genome.		

0.0.760



References





References

Open "http://main.g2.bx.psu.edu/history/imp?id=e0b8bd5d661b10c2" in a new tab



the pileup dataset was produced by the samtools pileup command (although you

Open "http://main.g2.bx.psu.edu/tool_runner/rerun?id=1703758" in a new tab

Galaxy Published Page Variant Analysis for sample E18	
Image: state stat	Google
Galaxy Analyze Data Workflow Shared Data Visualization Help User	
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	igoecks
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 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	Rating
2796 occur in known RefSeq Genes. These potential variants are:	Community (0 ratings, 0.0 average)
Galaxy Dataset Intersect to get Variants from sample E18, consensus different, in RefSeq Genes Variants with consensus different that occur in RefSeq genes.	Yours ++++
	Tags
Method	Community: none
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:	Yours:
Galaxy History Variant Analysis for Sample E18 Perform a pileup analysis with default parameters to identify variants in sample E18.	
Here is a workflow for performing this analysis: Import workflow	
Galaxy Workflow Variant identification within annotated genes from NGS PE Data Identify variants in annotated genes from NGS paired-end 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).	Y
[2] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009).	
[3] Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078 -2079 (2009).	÷

59

Open "http://main.g2.bx.psu.edu/workflow/imp?id=58d16d45527990b7" in a new tab

0	O Galaxy	
Ľ	Exader C Q* Google	
=	Galaxy Analyze Data Workflow Shared Data Visualization Help User	
Pa	ge Editor Title : Variant Analysis for sample E18	
в	i I 😒 🛬 🗄 🗄 i i i i i i i i i i i i i i i i i	
	Variant Analysis of Embryonic Mouse Brain Tissue	
	Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team	
	Results	
	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.	
	The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) 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:	
	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 Bowtig [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:	
	Here is a workflow for performing this analysis:	
	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).	
	[2] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009).	4 4
_	× *	1

00			Galaxy							
+ http://main.g2.bx.psu.edu/page/edit_c	ontent?id=d2	23e005e1ec427	,			Reader 🖒	Q• Google			
🔁 Galaxy	Analyze Dat	a Workflow	Shared Data	Visualiza	tion Help	User				
Page Editor Title : Variant Analysis for sample E18 Save Clos										
в <i>I</i> × ¹ ×₂ ⊟ ⊟ ⊕ ⊕ ⊋ ♂ ♥ № № ⊠ ⊒	I x ² x ₂ IE IE + E + E + C S S S I Paragraph type T Insert Link to Galaxy Object T Embed Galaxy Object T Embed Histories									
Variant Analysis of Embr	search	< I A	Ivanced Search							
Jeremy Goecks, Anton Nekrutenko, James Taylor,	Nam	2		Tags	Last Updat	ted †				
Results	🗹 Varia	nt Analysis for S	ample E18	5 Tags	15 minutes	ago				
	Pileu	o analysis, samp	le E18	4 Tags	2 days ago					
To demonstrate how Galaxy can support accessible, identifies variants from a set of 4,536,964 RNA-seq	🗆 Unna	med history		0 Tags	Sep 07, 20	10 b	ariant analysis experiment. This experiment ryonic development.			
The initial analysis produced support for 27,742 pos	🖯 Unna	med history		<u>0 Tags</u>	Dec 17, 20	09	aseas determined by the MAQ modeldiffers			
from the reference base and (b) read coverage at the	🗆 impo	rted: Hsitory wit	h ~100 items	5 Tags	Dec 10, 20	09	nes. These potential variants are:			
	🗆 impo	rted: Galaxy vs	MEGAN	0 Tags	Dec 04, 20	09				
Method	🗆 Impo	rted: Galaxy vs	MEGAN	2 Tags	Oct 06, 20	09				
	🗆 impo	rted: Galaxy vs	MEGAN	0 Tags	Oct 06, 20	09				
In the first step of this analysis, the reads were groo exclude base pairs with low quality scores; see [1] ft Pender [2] A silver analysis (Strange Strange) [2] we	🗆 impo	rted: metagenor	nic analysis	0 Tags	Sep 30, 20	09	the reads were trimmed from 36bp to 27bp to ds were mapped using the short-read mapper			
powile [2]. A plieup analysis using SAMtools [3] was	🗆 impo	rted: Galaxy vs	MEGAN	0 Tags	Sep 30, 20	09	omplete analysis is contained in this history.			
	Page	1 <u>2</u> Show	v all histories on	one page						
Here is a workflow for performing this analysis:	For 1	selected histori	es:							
	Make the s	elected histories	accessible so th	hat they car	n viewed by	everyone.				
					Embed	Cancel				
References							,			
[1] Han, X. et al. Transcriptome of embryonic and net	onatal mouse	ortex by high-t	hroughput RNA	sequencing	. Proceeding	s of the Natio	onal Academy of Sciences 106, 12741-12746			

[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).

A V

[2] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009).

0	Galaxy Galaxy								
	H http://main.g2.bx.psu.edu/page/edit_content?id=d2523e005e1ec427 Reader C Q* Google Coogle Coogle								
l,	Galaxy Analyze Data Workflow Shared Data Visualization Help User								
Page	Page Editor Title : Variant Analysis for sample E18 Save Close								
в	I x² x₂ 🗄 🗄 👬 🍀 🖉 🖉 🍓 🛄 🛛 Paragraph type 🔻 Insert Link to Galaxy Object 🔻								
\	Variant Analysis of Embryonic Mouse Brain Tissue	Ĩ							
Je	ieremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team								
R	Results								
To	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.								
Tł	The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) 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:								
N	Method								
in eo <u>R</u>	n 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 sowtie [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:								
	Embedded Galaxy History 'Variant Analysis for Sample E18'								
	[Do not edit this block; Galaxy will fill it in with the annotated history when it is displayed.]								
н	Here is a workflow for performing this analysis:								

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).

A V

	00	0			Galaxy							
	•	Http://main.g2.bx.psu.edu/page/edit_c	ontent?id=d2523e	2005e1ec427	,			Reader	C Google			1
	=_	Galaxy	Analyze Data	Workflow	Shared Data	Visualization	Help	User				
	Page Editor Title : Variant Analysis for sample E18											
	B I x ² x ₂ 🗄 🗄 + 🗄 + Ə (* S S S M) Paragraph type 🔻 Insert Link to Galaxy Object 🔻 Embed Galaxy Object 🔻											
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.						ſ	ĺ					
The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus baseas determined by the MAQ modeldifference 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:								ers				
		Embedde	d Galaxy Dataset	Variants fr	om sample F18	consensus diff	ferent, ir	n RefSea (Genes'			Ļ

[Do not edit this block; Galaxy will fill it in with the annotated dataset when it is displayed.]

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:

Embedded Galaxy History 'Variant Pileup Analysis for Sample E18'

[Do not edit this block; Galaxy will fill it in with the annotated history when it is displayed.]

Here is a workflow for performing this analysis:

Embedded Galaxy Workflow 'SNP identification within annotated genes from NGS PE Data'

[Do not edit this block; Galaxy will fill it in with the annotated workflow when it is displayed.]

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).

[2] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009).

[3] Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078 –2079 (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

Galaxy 101 Exercise





Enis Afgan



Dave Clements



Dannon Baker



Jeremy Goecks



Kanwei Li



James Taylor



PENNSTATE.

Dan Blankenberg



Jennifer Jackson



Guru Ananda



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

Galaxy 101 http://usegalaxy.org/galaxy101

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



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

>chr1

taaccctaaccctaaccctaaccctaaccctaaccctaacccta accctaaccctaaccctaaccctaaccctaaccctaac

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