CGEBM Assembly Workshop December 2015 Galaxy Walkthrough

1. Open firefox, enter login1/galaxy in the address. Click user, register and login to Galaxy

= Galaxy	Analyze Data	Workflow	Shared Data 🗸	Visualization	Help -	User 🗸		Using 0 byte
You have been logged out. You can log in again, go back to the	page you were vis	<u>siting</u> or <u>go t</u>	o the home page.			Login Register	r	
Create account								
Email address:								
Password:								
Confirm password:								
Public name:								
Your public name is an identifier that wi length and contain only lower-case lette	II be used to gene ers, numbers, and	erate addres the '-' chara	ses for informatio acter.	n you share pub	olicly. Publ	lic names m	nust be at least f	our characters in
Submit								

=_ Galaxy	Analyze Data	Workflow	Shared Data -	Visualization	Help -	User -	 Using 0 bytes
Login						Login	
Email address:						Register	
Password:							
Forgot password? Reset here							
Login							

2. Click on Shared Data->Data Libraries and select Assembly Workshop 2015

📮 Galaxy	Analyze Data Workf	low Shared Data 🗸	Visualization -	Admin He	elp∓ User∓	=	Using 0%
Data Libraries		Data Libraries Data Libraries	deprecated				
search dataset name, info, message, dbkey	Q						
Advanced Search		Published His	tories				
		Published Wo	rkflows				
Data library name:		Published Vis	ualizations y d	escription			
Assembly Workshop 2015		Published Pa	jes iles	5			

2.b Select sequence_1.fq and sequence_2.fq -> import to current history and click Go

Data Library "Assembly tutorial Dec15"

Simulated SacCer reads

🔽 Name	Message	Data type	Date uploaded	File size
✓ sequence_1.fq		fastqsanger	Thu Nov 26 13:21:06 2015 (UTC)	71.3 MB
✓ sequence_2.fq		fastqsanger	Thu Nov 26 13:21:42 2015 (UTC)	71.3 MB
For selected datasets: Import to cur	rent history ᅌ G	o		

1 TIP: You can download individual library datasets by selecting "Download this dataset" from the context menu (triangle) next to each dataset's name.

3. Select Analyze Data tab, then search and choose FastQC. Select multiple datasets, shift click to selet sequence_1.fq and sequence_2.fq and click Execute.

🔁 Galaxy	Analyze Data Workflow Shared Data			Using 0%
Tools	FastQC Read Quality reports (Galaxy Tool Version 0.63)		History	€‡⊡
fastqc 🙁	Short read data from your current history		search datasets	8
NGS: QC and manipulation	C 2: sequence_2.fq 1: sequence_1.fq		Unnamed history	
FastQC Read Quality reports	Multiple datasets		0 b	S
All workflows	🔛 🚠 This is a batch mode input field. A separate job will be triggered for each dataset.		2: sequence 2.fq	• / ×
	Contaminant list		<u>1: sequence 1.fq</u>	⊛ & ×
	tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA			
	Submodule and Limit specifing file			
	□ ℓ2 □ Nothing selected •			
	a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter	=		
	✓ Execute			

3.a To visualize the results click on the eye icon next to the "FastQC on data #:Webpage"

T Galaxy	Analyze Data Workflow Shared Data			Using 0%
Tools	FastQC Read Quality reports (Galaxy Tool Version 0.63)		History	<i>€</i> ¢ ⊡
fastqc 🙁	Short read data from your current history		search datasets	8
NGS: QC and manipulation FastQC Read Quality reports	2: sequence 2.fq 1: sequence_1.fq		Unnamed history 2 shown	
Workflows			0 b	V > 9
 <u>All workflows</u> 	🚠 This is a batch mode input field. A separate job will be triggered for each dataset.		2: sequence 2.fg	• 🖋 🗙
	Contaminant list		1: sequence 1.fq	• 🖋 🗙
	Image: Constraint of the with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA			
	Submodule and Limit specifing file			
	C 4 D Nothing selected -			
	a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter	=		
	✓ Execute			



4. Search and choose Trim_Galore In the middle panel, select Paired-end and sequence_1.fq, sequence_2.fq. Set the "Discard Reads that became shorter than int" to 50 and click Execute button.

🔁 Galaxy		Analyze Data Workflow Shared Data → Visualization → Admin Help → User →		Using 0%
Tools	1	Trim Galore adaptive quality and adapter trimmer (Galaxy Tool Version 0.2.8.1)	History	€\$□
trim galore	8	Is this library mate-paired?	search dataset	s 8
NGS: QC and manipulation		Single-end -	Unnamed history	/
<u>Trim Galore</u> adaptive quality an adapter trimmer	d	FASTQ/FASTA file	0 b	S
Workflows		FASTO or FASTA files.	6: FastQC on data	a 2: 💿 🖋 🗙
 <u>All workflows</u> 		Trim galore! advanced settings	RawData	-
		Full parameter list	5: FastQC on data Webpage	<u>12:</u> (1) (1) (1) (1) (1) (1) (1) (1) (1) (1)
		You can use the default settings or set custom values for any of Trim Galore's parameters.	A: FastQC on	data 1: 💿 🖋 🗙
		Trim low-quality ends from reads in addition to adapter removal.	: 3: FastOC on	data 1: 🔿 🖉 🛩
		For more information please see below.	<u>Webpage</u>	
		Adapter sequence to be trimmed	2: sequence 2.fq	●
		AGATCGGAAGAGC	≡ <u>1: sequence 1.fq</u>	● & ×
		Overlap with adapter sequence required to trim a sequence		
		1		
		Maximum allowed error rate		
		0.1		
		Discard reads that became shorter than length INT		
		50		
		nstructs Trim Galore to remove INT bp from the 5' end of read 1		
		0		
		nstructs Trim Galore to remove INT bp from the 5' end of read 2		
		0		
		Generate a report file		
		Yes No		
		specify if you would like to retain unpaired reads		
		Do not output unpaired reads		

5. Repeat FastQC using Trim_Galore results

💳 Galaxy	Analyze Data Workflow Shared Data - Visualization - Admin Help - User -		Using 1.9 GB
Calaxy Tools fastqc fastqc NGS: QC and manipulation FASTX-TOOLKIT FOR FASTQ DATA FastQC Read Quality reports Workflows • All workflows	Analyze Data Workflow Shared Data + Visualization + Admin Help + User + FastQC (version 0.63) Short read data from your current history: 1 <th>History Search datasets CGEBM Assembly Wo 1.9 MB C 8: Trim Galore on a 2 and data 1: trimm reads pair 2 C 7: Trim Galore on a 2 and data 1: trimm</th> <th>Using 1.9 GB</th>	History Search datasets CGEBM Assembly Wo 1.9 MB C 8: Trim Galore on a 2 and data 1: trimm reads pair 2 C 7: Trim Galore on a 2 and data 1: trimm	Using 1.9 GB
	for the each submodules warning parameter Execute	a 2 and data 1: trimm reads pair 1 6: FastQC on data 2: F Data	ea Raw 🕑 🖋 🗙

6. Search and select velveth on the Tools menu. "Add new input file", set file to "fastq" and read type to "shortPaired reads" and Dataset "trimmed reads pair 1". Click on "Add new input file" and file to "fastq" and read type to "shortPaired reads" and Dataset "trimmed reads pair 2"

=_ Galaxy	Analyze Data Workflow Shared Data - Visualization - Admin Help - User -	Us	ing 2.0 GB
Tools	velveth (version 1.2.10.0)	History	C 🌣
velvet	Hash Length:	search datasets	8
Assembly	21 😳 k-mer length in base pairs of the words being hashed.	CGEBM Assembly Worksh	ор
velvetg Velvet sequence assembler for very short reads	Use strand specific transcriptome sequencing:	145.2 MB	. 🗹 📎 🗩
velveth Prepare a dataset for the Velvet velvetg Assembler	□ If you are using a strand specific transcriptome sequencing protocol, you may wish to use this option for better results.	<u>10: FastQC on data 7: Ra</u> wData	• / ×
	Input Files	9: FastOC on data 7: Web	@ # ¥
All workflows	Input Files 1	page	
<u>In Hornions</u>	file format: fastq 📀	8: Trim Galore on data 2 and data 1: trimmed rea	• / ×
	read type: shortPaired reads	7: Trim Galore on data 2	
	Dataset: 🗅 🕲	and data 1: trimmed rea ds pair 1	
	Remove Input Files 1	<u>6: FastQC on data 2: Raw</u> Data	• / ×
	Input Files 2	5: FastQC on data 2: Web page	• / ×
	file format: fastq 📀	<u>4: FastQC on data 1: Raw</u> <u>Data</u>	• / ×
	read type: shortPaired reads	3: FastQC on data 1: Web	• / ×
	Dataset: 🗅 🖄 B: Trim Galore on data 2 and data 1: trimmed reads pair 2 😜	<u>2: sequence 2.fq</u>	• / ×
	Remove Input Files 2	<u>1: sequence 1.fq</u>	• / ×
	Add new Input Files		
	Execute		

7. Search for velvetg, select the velveth results change "coverage cutoff" and "expected coverage" to "automatically determined". Set minimum contig length to 600 and using paired reads to "Yes" with Insert Length 600. Click Execute

=_ Galaxy	Analyze Data Workflow Shared Data - Visualization - Admin Help - User -		Us	ing 2.0 GB
Tools	velvetg (version 1.2.10.0)		History	C 🌣
velvet	Velvet Dataset:		search datasets	8
Assembly	11: velveth on data 7 and data 8 Prepared by velveth.		CGEBM Assembly Worksh	юр
<u>velvetg</u> Velvet sequence assembler for very short reads	Generate a AMOS.afg file:		145.2 MB	2 > >
velveth Prepare a dataset for the		1	11: velveth on data 7 an	• / ×
Velvet velvetg Assembler		<u>c</u>	10: EastOC on data 7: Ra	
Workflows All workflows	Generate velvet LastGraph file:	Y	wData	• / *
	No C	9	9: FastQC on data 7: Web	• / ×
		1	8: Trim Galore on data 2	• / ×
	Coverage cutoff:	<u>a</u>	and data 1: trimmed rea ds pair 2	
	Automatically Determined	1	7: Trim Galore on data 2	• / ×
	Expected Coverage of Unique Regions:	a (and data 1: trimmed rea ds pair 1	
	Set minimum contig length:	6	6: FastQC on data 2: Raw	• / ×
	Yes 📀 minimum contig length exported to contigs.fa file (default: hash length * 2).		<u>Data</u>	
	minimum contig length:	1	page	• / X
	600	4	4: FastQC on data 1: Raw	• / ×
	Using Paired Reads:		3: FastOC on data 1: Web	@ # ¥
	Yes 📀	I	page	
	Insert Length in Paired-End Read dataset (ignored when -1):	2	2: sequence_2.fq	• / ×
	Expected distance between two paired end reads	1	<u>1: sequence 1.fq</u>	• / ×
	Velvet Advanced Options:			
	Everyte			

8. Search for Augustus in the Tools pane and select velvetg Contigs as input and Model Organism "Saccharomyces cerevisiae"

💳 Galaxy	Analyze Data Workflow Shared Data - Visualization - Admin Help - User -	Using 2.1 GB
Tools	Augustus (version 3.1.0)	History C 🌣
augustus	Genome Sequence:	search datasets
Assembly	13: velvetg on data 11: Contigs 😳	CGEBM Assembly Workshop
<u>Augustus</u> gene prediction for eukaryotic genomes	Otherwise, intron-spanning stop codons could occur. (noInFrameStop)	270.9 MB
Workflows	Predict genes independently on each strand:	13: velvetg on data 11: C O X
<u>All workflows</u>	This allows overlapping genes on opposite strands. (singlestrand)	12: velvetg on data 11: S () S X
	Predict the untranslated regions in addition to the coding sequence:	11: velveth on data 7 an 💿 🖋 🗙
	This currently works only for human, galdieria, toxoplasma and caenorhabditis. (UTR)	d data 8
	Model Organism:	10: FastQC on data 7: Ra wData
	Choose a specialised trainingset.	9: FastQC on data 7: Web 💿 🖋 🗙
	Predict genes on specific strands:	8: Trim Galore on data 2 💿 🖋 🗙
	(strand)	<u>and data 1: trimmed rea</u> <u>ds pair 2</u>
	complete ᅌ	7: Trim Galore on data 2
	Gene Model to predict, for more information please refere to the help. (genemodel)	ds pair 1
	Standard output is GTF. (gff3)	6: FastQC on data 2: Raw 💿 🥒 🗙 Data
	Output options:	5: FastQC on data 2: Web 💿 🖋 🗙
	Select All Unselect All	4: EastOC on data 1: Paw on A w
	 coding sequences (protein) coding sequence as comment in the output file (codingseq) 	Data
	predicted intron sequences (introns) predicted start codons (start) predicted start codons (start)	3: FastQC on data 1: Web 🕑 🖋 🗙
	CDS region (cds)	2: sequence 2.fq 💿 🖋 🗙
	Execute	1: sequence 1.fq

9. Search for blastp in the Tools pane, select Augustus protein sequence as query, and Swiss-Prot as database. Change the type of blast to blastp-fast, expectation cut-off to 0.00001 and click on tabular (select which columns) and scroll down to select extra columns as per the next slide

– Galaxy	Analyze Data Workflow Shared Data - Visualization - Admin Help - User	*	
Tools	NCBI BLAST+ blastp Search protein database with protein query sequence(s) (Galaxy Tool Version Options 0.1.07)	History	2*
NCBI BLAST+ NCBI BLAST+ blastp Search protein database with protein query sequence(s)	Protein query sequence(s)	CGEBM Assembly Worksl 270.9 MB	
Workflows All workflows	Protein BLAST database SwissProt_20_Nov_2015	13: Coding sequence	
	Type of BLAST	14: Augustus on data 13: GTF/GFF	• / ×
	 blastp - Traditional BLASTP to compare a protein query to a protein database blastp-fast - Use longer words for seeding, faster but less accurate blastp-short - BLASTP optimized for queries shorter than 30 residues 	<u>13: velvetg on data 11: C</u> ontigs	• / × !
	Set expectation value cutoff	- <u>12: velvetg on data 11: S</u> tats	• # ×
	0.00001	<u>11: velveth on data 7 an</u> <u>d data 8</u>	• / ×
	Tabular (select which columns)	<u>10: FastQC on data 7: Ra</u> <u>wData</u>	• / ×
	Standard columns Select/Unselect all	<u>9: FastQC on data 7: Web</u> page	• / ×
	 ✓ qseqid = Query Seq-id (ID of your sequence) ✓ sseqid = Subject Seq-id (ID of the database hit) ✓ pident = Percentage of identical matches 	<u>8: Trim Galore on data 2</u> and data 1: trimmed rea ds pair 2	● / × !
	✓ Iength = Alignment length ✓ mismatch = Number of mismatches	7: Trim Galore on data 2 and data 1: trimmed rea	• # ×

9.a Scroll and select staxids. Click on show advanced options and set maximum hits to 1 and minimum query coverage to 60. Click Execute.

- Galaxy		Analyze Data	Workflow	Shared Data -	Visualization		Using 2.1 GB
Tools	Taxonomy	columns Inselect all				History	C 🕈
NCBI BLAST+ NCBI BLAST+ blastp Search protein database with protein query sequence(s) Workflows All workflows	Select/U Select/Selec	Inselect all s = unique Subject Taxonomy ID(s), sep mes = unique Subject Scientific Name(s ames = unique Subject Common Name names = unique Subject Blast Name(s), doms = unique Subject Super Kingdom stions ced Options w complexity regions (with SEG)	urated by a ',' (in (), separated by a (s), separated by separated by a ',' s), separated by a ','	numerical order) k '' a '' ' (in alphabetical order) a ''' (in alphabetical ord	ler)	History search datasets CGEBM Assembly Wo 270.9 MB <u>13: Coding sequence</u> 15: Augustus on d	rkshop QVD ata
• <u>All workflows</u>	Yes N Scoring m BLOSUM Maximum 1 Use zero fi	o atrix 32 (default) hits to show r default limits			•	Isragestes on a Isragestes on a	
	Word size 0 Use zero fo Should the	Word size for wordfinder algorithm 0 Use zero for default, otherwise minimum 2. Should the query and subject defline(s) be parsed?			12: velvetg on data 1: tats 11: velveth on data 7 d data 8 10: FastOC on data 7:	<u>an</u>	
	Restrict search of database to a No restriction, search the entire of This feature provides a means to of	o the formatting of the query/subject ID st arch of database to a given set of ID's ion, search the entire database provides a means to exclude ID's from	ings a BLAST databa:	se search. The expecta	• ectation values in	wData 9: FastQC on data 7: V page	Veb @ # X
	the BLAST Note this co Minimum o 60 See also th ✓ Execute	results are based upon the sequences a nnot be used when comparing against uery coverage per hsp (percentage, 0	ctually searched a FASTA file. to 100)	i, and not on the underl	ying database.	and data 1: trimmed r ds pair 2 7: Trim Galore on data and data 1: trimmed r ds pair 1	$\frac{a 2}{ea} \otimes \mathscr{P} \times$

10. Search and select assemblystats from Tools and run for velvetg contigs

- Galaxy	Analyze Data worknow Shared Data Visualization Admin Help User	===	Using 2.1 GB
Tools	assemblystats (version 1.0.1)	History	C \$
Assemblys Assemblystats Summarise an assembly (e.g. N50 metrics) Workflows All workflows	assemblystats (version 1.0.1) Type of read: Contig (if from genomic assembly) Is this from an genomic (contig) or transcriptomic assembly (isotig) or are these raw reads (read) Output histogram with bin sizes=1: Use this to specify whether or not bin sizes of 1 should be used when plotting histograms Source file in FASTA format: O: 13: velvetg on data 11: Contigs O: Execute Summarise assembly overview This script is used to give summary statistics of an assembly or set of reads. Typically this is run after an assembly to evaluate gross features. # Gives back # - N50 # - num of contigs > 1 kb # - num of contigs # - Read or Contig Histogram and graphs. # - Summed contig length (by number of contigs, in sorted order) It his tool was installed from a ToolShed, you may be able to find additional information by following this link: http://toolshed.g2.bx.psu.edu/view/konradpaszkiewicz/assemblystats	Search datasets CGEBM Assembly W 270.9 MB © 17: blastp Augus on data 13: Protein s ence vs sprot 20Nov) 16: Augustus on 13: Coding sequence) 16: Augustus on 13: Protein sequence) 14: Augustus on 13: GTF/GFF 13: velvetg on data 2 ontigs 12: velvetg on data 2 itats 11: velveth on data 2 d data 8	C C C C C C C C C C C C C C C C C C C

11. Select BWA mem from Tools. "Use a genome from history" and select Augutus Coding Sequence. Select Paired reads and "trimmed reads pair 1" and "trimmed reads pair 2". Click on Execute.

💳 Galaxy	Analyze Data Workflow Shared Data - Visualization - Admin Help - User -		Using 2.1 GB
Tools	Map with BWA-MEM (version 0.4.1)	History	C \$
Tools bwa mem bwa mem media Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome Workflows • All workflows	Map with BWA-MEM (version 0.4.1) Will you select a reference genome from your history or use a built-in index?: Use a genome from history and build index Built-ins were indexed using default options. See `Indexes` section of help below Use the following dataset as the reference sequence: C 16: Augustus on data 13: Coding sequence You can upload a FASTA sequence to the history and use it as reference Single or Paired-end reads: Paired Select first set of reads: ? 7: Trim Galore on data 2 and data 1: trimmed reads pair 1 Specify dataset with forward reads Select second set of reads: Select second set of reads:	History search datasets CGEBM Assembly We 270.9 MB 23: Cumulative sum contig size data 22: Histogram data 21: Cumulative sum contig sizes 20: Histogram of corr g sizes	C C C C C C C C C C C C C C
	B: Trim Galore on data 2 and data 1: trimmed reads pair 2 Specify dataset with reverse reads Enter mean, standard deviation, max, and min for insert lengths.:	19: Sorted contigs 18: Assembly statist 5	• • ×
	-I; This parameter is only used for paired reads. Only mean is required while sd, max, and min will be inferred. Examples: both "250" and "250,25" will work while "250,,10" will not. See below for details.	17: blastp August s on data 13: Protein equence vs sprot 20	tu 💿 🖋 🗙 s Nov2015
	Set read groups information?: Do not set	16: Augustus on data 13: Coding sequence	<u>a</u> (e) / ×
	Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.	<u>15: Augustus on data</u> <u>13: Protein sequence</u>	a 💿 🖋 🗙
	Select analysis mode: 1.Simple Illumina mode	14: Augustus on data 13: GTF/GFF	a 💿 🖋 🗙
	Execute	<u>13: velvetg on data 1</u> <u>1: Contigs</u>	• / ×

12. Select blobplot from Tools pane. Select blastp results: blastp "Augustus on Protein sequence vs sprot", predicted Nucleotide "Augustus Coding Sequence", Aligned reads "Map with BWA_MEM", cut-off to 0.001 and Taxon level to Order. Click Execute.

- Galaxy	Analyze Data Workflow Shared Data - Visualization - Admin Help - User -		Using 2.1 GB
Tools	Blobplot (version 0.0.1)	History	C \$
blobplot ③ CCEBM Blobplot Creates a plot of GC vs coverage for sequences colored by taxonomy Workflows • All workflows	Blast Results: 17: blastp Augustus on data 13: Protein sequence vs sprot_20Nov2015 Predicted Nucleotide Sequences: 12: 12: 12: 12: 12: 12: 12: 12: 12: 12:	search datasets CGEBM Assembly W 270.9 MB	orkshop
	Aligned reads against Predicted Sequences: 24: Map with BWA-MEM on data 16, data 8, and data 7 (mapped reads in BAM format) Cut off for plot: 0.00 Taxon level: Order	 24: Map with BW, MEM on data 16, dat 8, and data 7 (mapp M format) 23: Cumulative sum contig size data 22: Histogram data 	A- OF C A X
	This tool creates a GC vs Coverage plot coloured by taxonomic id for a set of sequences	<u>contig sizes</u> 20: Histogram of con <u>g sizes</u>	nti 🕑 🖋 🗙