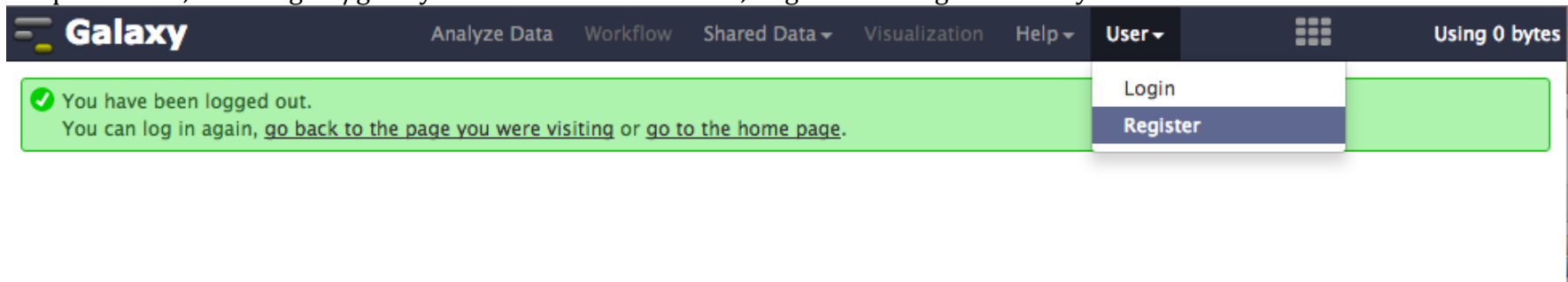


CGEBM Assembly Workshop December 2015 Galaxy Walkthrough

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



Create account

Email address:

Password:

Confirm password:

Public name:

Your public name is an identifier that will be used to generate addresses for information you share publicly. Public names must be at least four characters in length and contain only lower-case letters, numbers, and the '-' character.

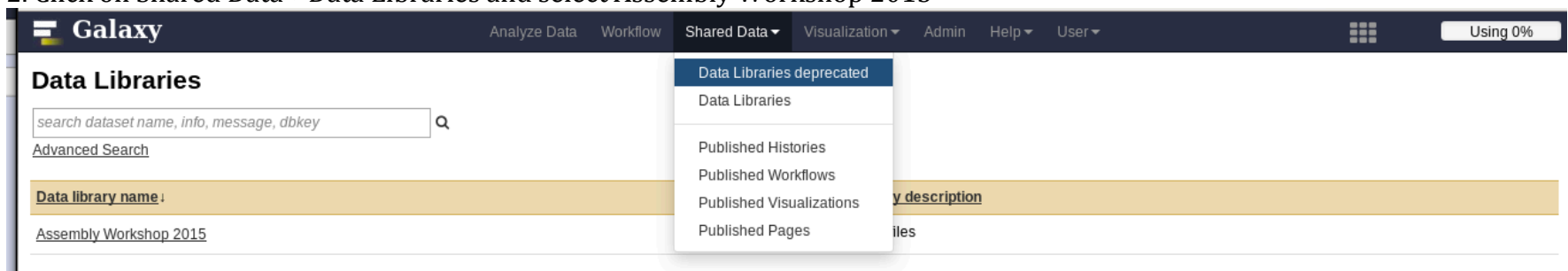
Login

Email address:

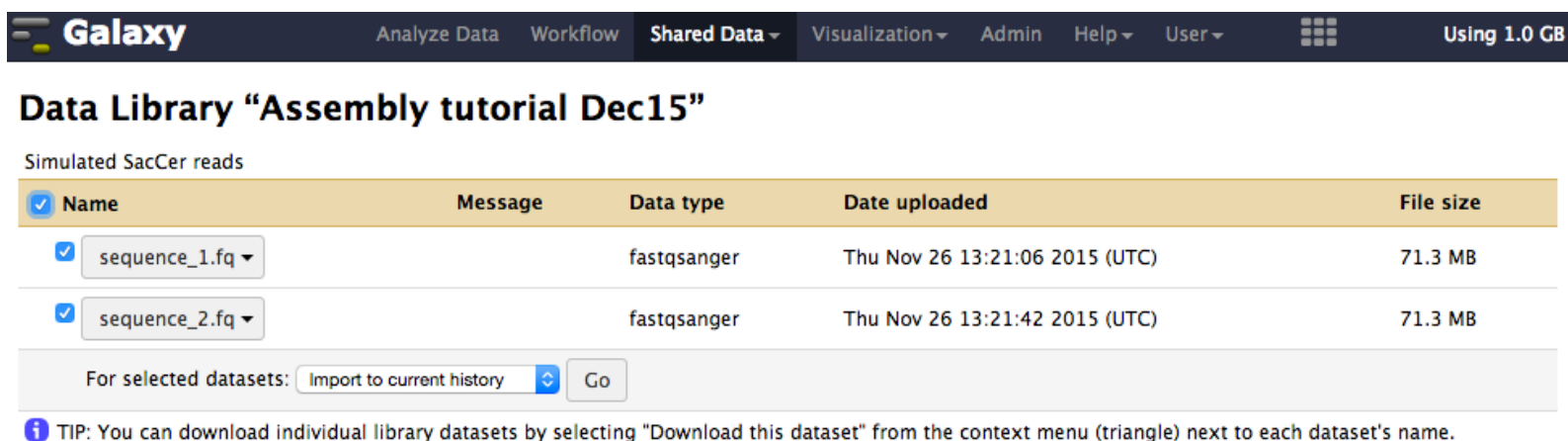
Password:

[Forgot password?](#) [Reset here](#)

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



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



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

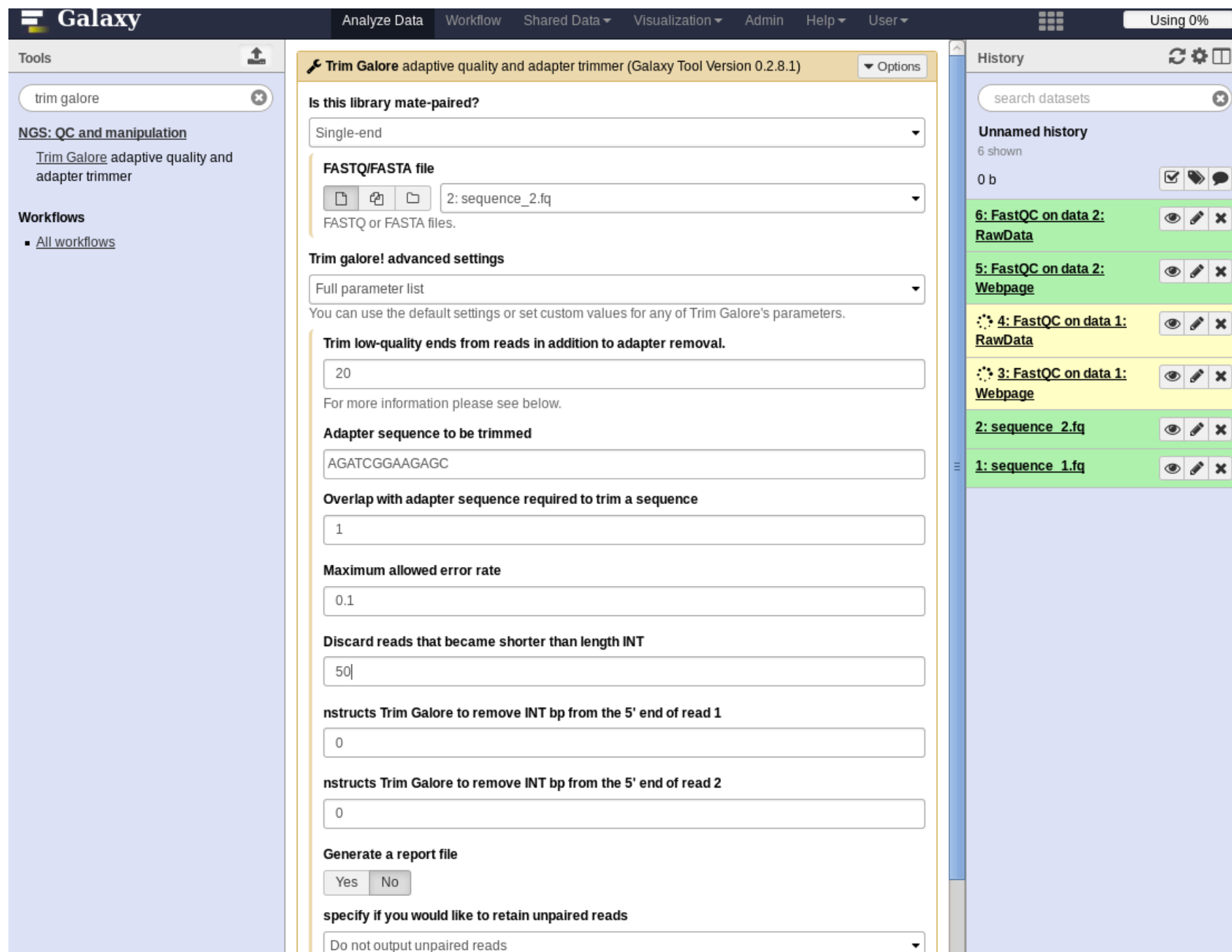
Note! These materials are for the use of University of Aberdeen staff and students only. They are not for wider distribution. Copyright laws apply.
Copyright © 2015 Centre for Genome Enabled Biology and Medicine

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

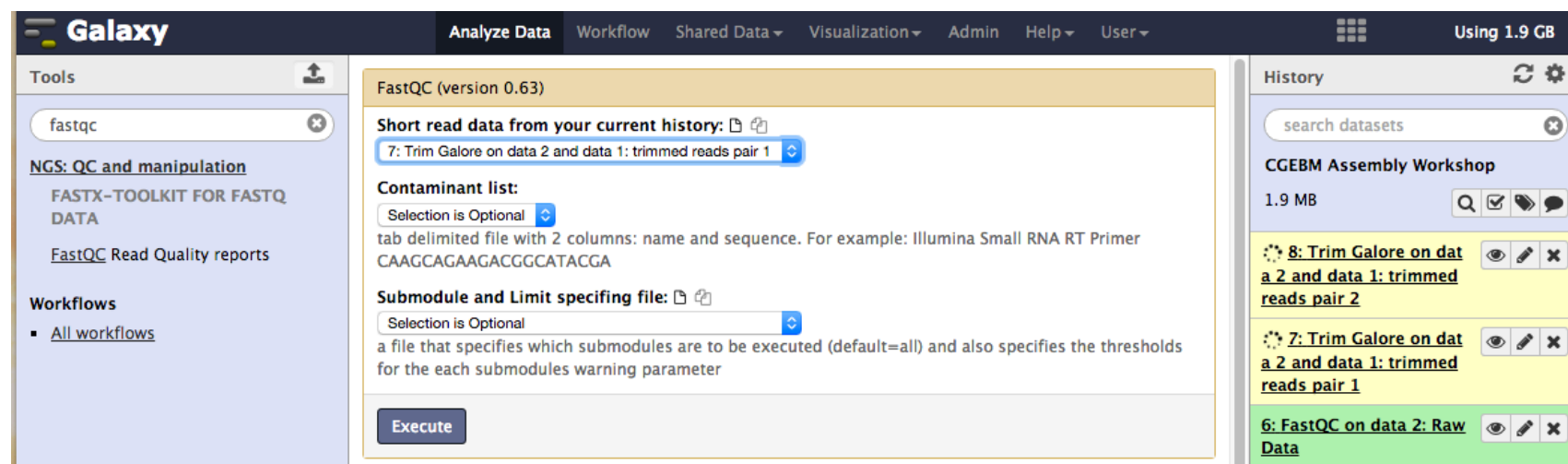
Measure	Value
Filename	sequence_1.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	333909
Sequences flagged as poor quality	0
Sequence length	100
%GC	39

Note! These materials are for the use of University of Aberdeen staff and students only. They are not for wider distribution. Copyright laws apply. Copyright © 2015 Centre for Genome Enabled Biology and Medicine

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.



5. Repeat FastQC using Trim_Galore results



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”

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

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**

Assembly
 Augustus gene prediction for eukaryotic genomes

Workflows
 All workflows

Augustus (version 3.1.0)

Genome Sequence:
 13: velvetg on data 11: Contigs

Don't report transcripts with in-frame stop codons (--noInFrameStop):
 Otherwise, intron-spanning stop codons could occur. (--noInFrameStop)

Predict genes independently on each strand:
 This allows overlapping genes on opposite strands. (--singlestrand)

Predict the untranslated regions in addition to the coding sequence:
 This currently works only for human, galdieria, toxoplasma and caenorhabditis. (--UTR)

Model Organism:
 Saccharomyces cerevisiae
 Choose a specialised trainingset.

Predict genes on specific strands:
 both
 (--strand)

Gene Model:
 complete
 Gene Model to predict, for more information please refer to the help. (--genemodel)

GFF formatted output:
 Standard output is GTF. (--gff3)

Output options:
 Select All Unselect All

- predicted protein sequences (--protein)
- coding sequence as comment in the output file (--codingseq)
- predicted intron sequences (--introns)
- predicted start codons (--start)
- predicted stop codons (--stop)
- CDS region (--cds)

Execute

History

search datasets

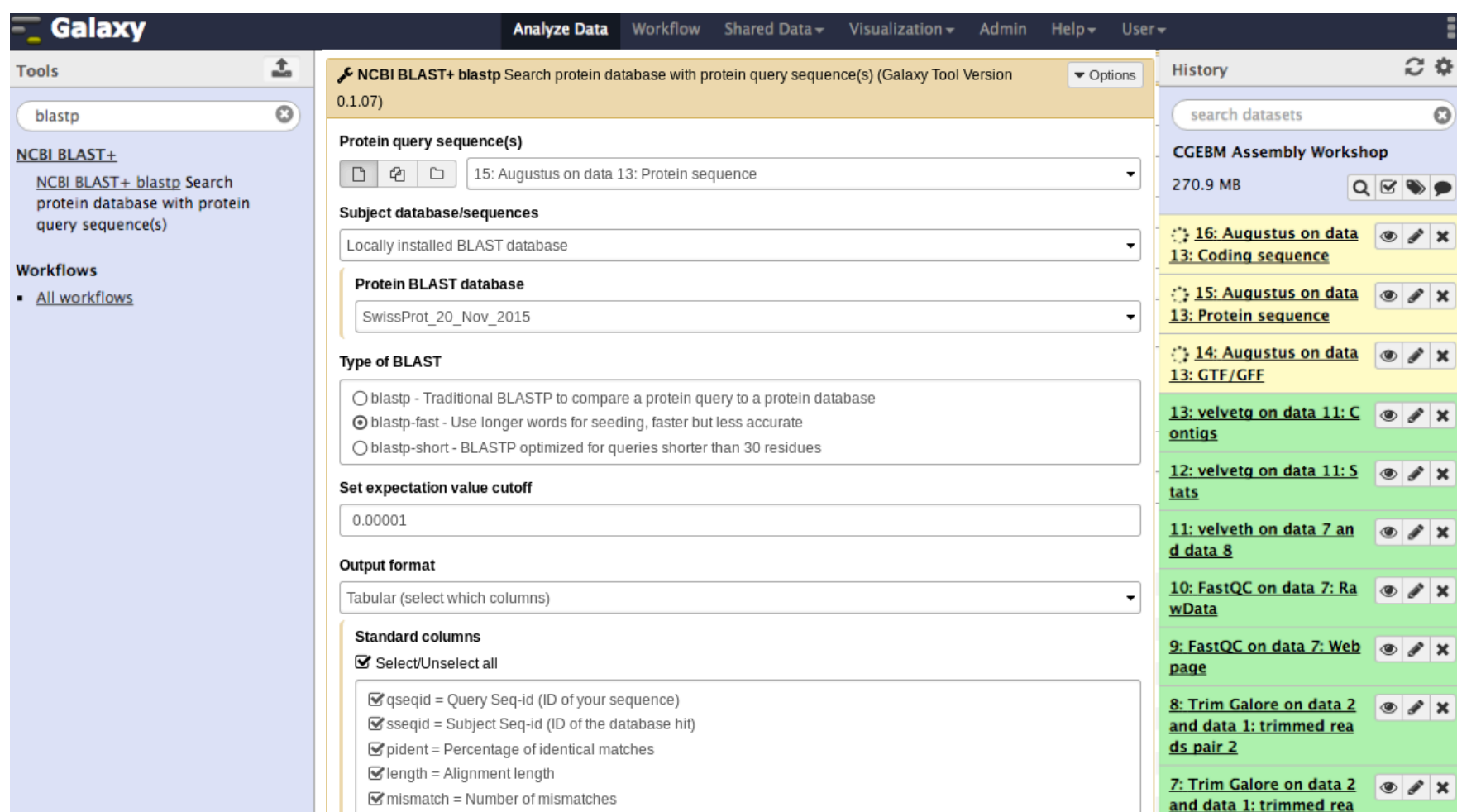
CGEBM Assembly Workshop
 270.9 MB

- 13: velvetg on data 11: Contigs
- 12: velvetg on data 11: Stats
- 11: velvetg on data 7 and data 8
- 10: FastQC on data 7: Raw Data
- 9: FastQC on data 7: Web page
- 8: Trim Galore on data 2 and data 1: trimmed reads pair 2
- 7: Trim Galore on data 2 and data 1: trimmed reads pair 1
- 6: FastQC on data 2: Raw Data
- 5: FastQC on data 2: Web page
- 4: FastQC on data 1: Raw Data
- 3: FastQC on data 1: Web page
- 2: sequence 2.fq
- 1: sequence 1.fq

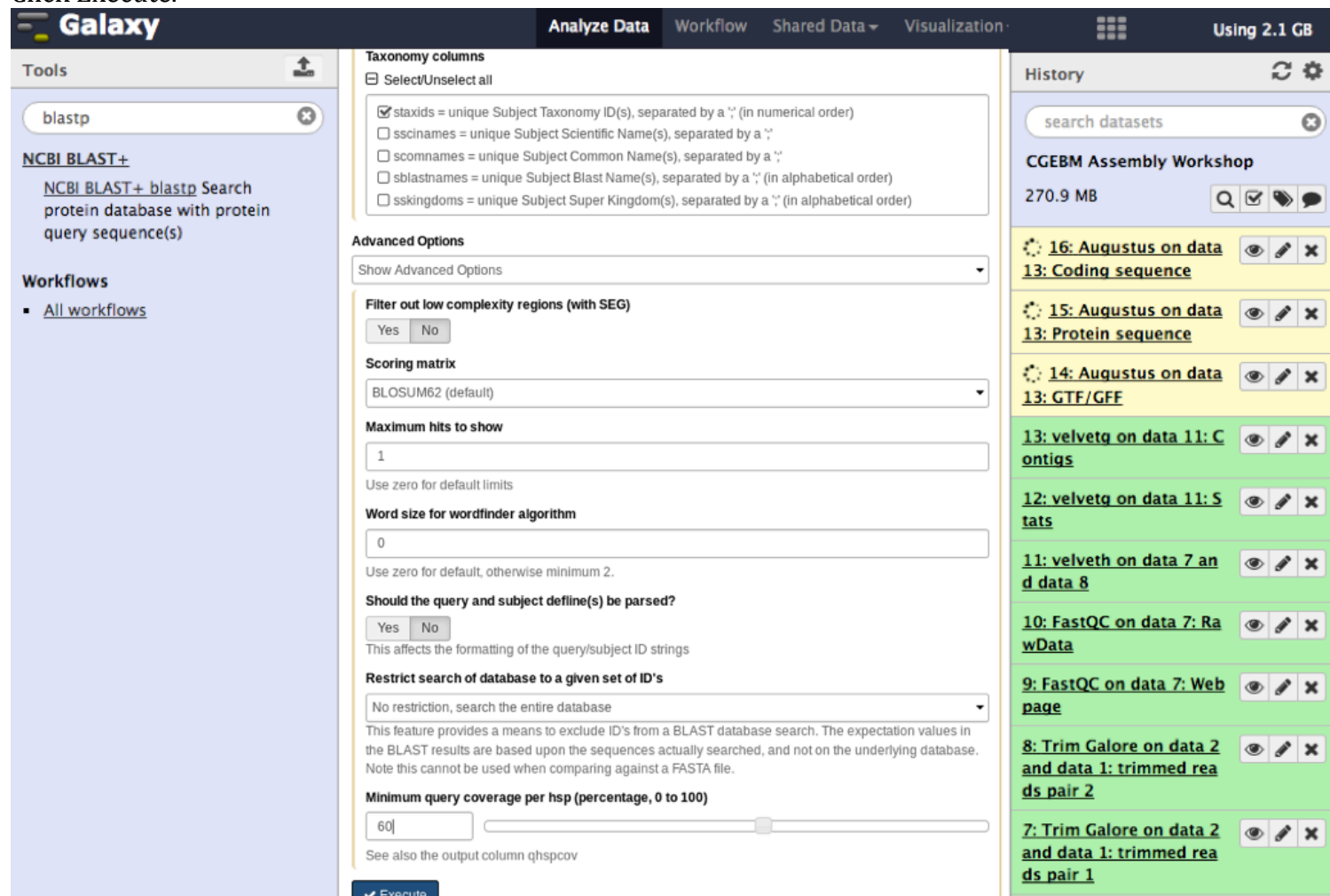
Note! These materials are for the use of University of Aberdeen staff and students only. They are not for wider distribution. Copyright laws apply.

Copyright © 2015 Centre for Genome Enabled Biology and Medicine

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

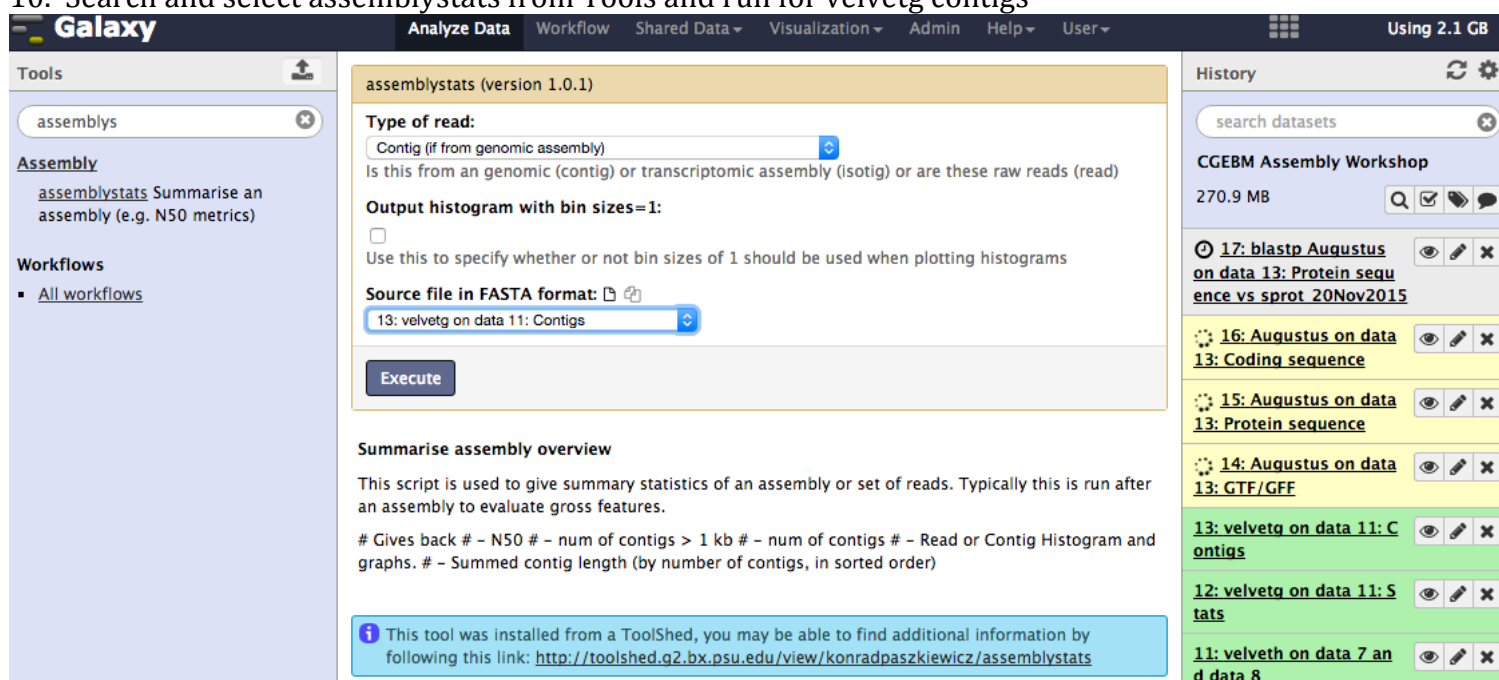


9.a Scroll and select staxids. Click on show advanced options and set maximum hits to 1 and minimum query coverage to 60. Click Execute.

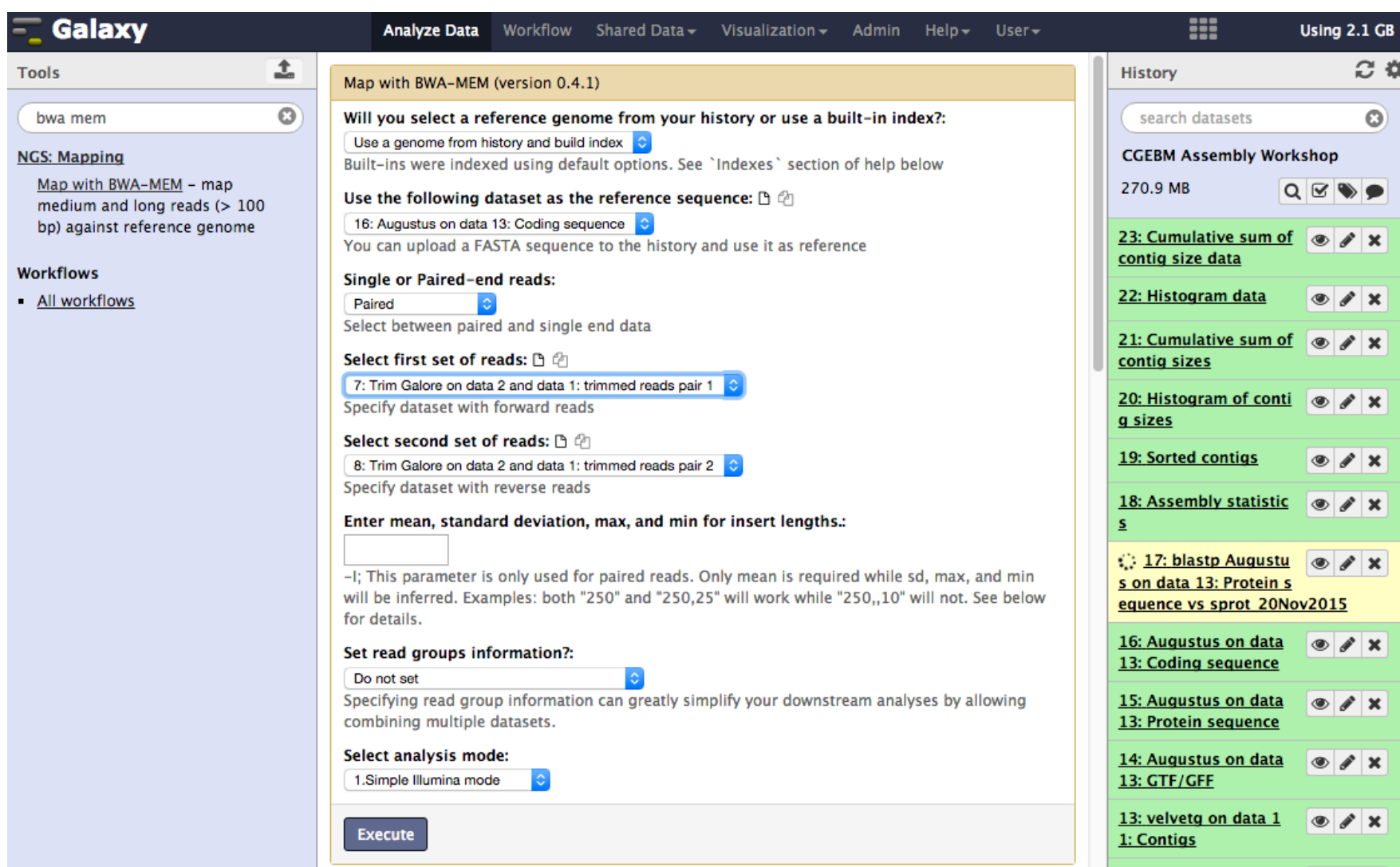


Note! These materials are for the use of University of Aberdeen staff and students only. They are not for wider distribution. Copyright laws apply. Copyright © 2015 Centre for Genome Enabled Biology and Medicine

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



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



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

Click to go forward, hold to see history Tools

blobplot

CGEBM
 Blobplot Creates a plot of GC vs coverage for sequences colored by taxonomy

Workflows
 All workflows

Blobplot (version 0.0.1)

Blast Results:
 17: blastp Augustus on data 13: Protein sequence vs sprot_20Nov2015

Predicted Nucleotide Sequences:
 16: Augustus on data 13: Coding sequence

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:

Taxon level:

Execute

This tool creates a GC vs Coverage plot coloured by taxonomic id for a set of sequences

History

search datasets

CGEBM Assembly Workshop
 270.9 MB

24: Map with BWA-MEM on data 16, data 8, and data 7 (mapped reads in BAM format)

23: Cumulative sum of contig size data

22: Histogram data

21: Cumulative sum of contig sizes

20: Histogram of contig sizes