Galaxy RNA-Seq Analysis: *S. tuberosum*

*Tutorial*
*Research Informatics Support Systems*
*Minnesota Supercomputing Institute*
*University of Minnesota*
*Version 1.1*
*10/24/2012*
<table>
<thead>
<tr>
<th>Appendix I: Workflows</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>9.1</strong> Extract workflow from current history</td>
<td>31</td>
</tr>
<tr>
<td><strong>9.2</strong> Edit the workflow</td>
<td>32</td>
</tr>
<tr>
<td><strong>9.3</strong> Create new history</td>
<td>34</td>
</tr>
<tr>
<td><strong>9.4</strong> Run workflow</td>
<td>35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Appendix II: Visualizing alignments with IGV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>10.1</strong> Load BAM alignment files and GTF into new history</td>
<td>36</td>
</tr>
<tr>
<td><strong>10.2</strong> Load Genome Fasta file into current history</td>
<td>37</td>
</tr>
<tr>
<td><strong>10.3</strong> Download files to your desktop</td>
<td>38</td>
</tr>
<tr>
<td><strong>10.4</strong> Start IGV</td>
<td>39</td>
</tr>
<tr>
<td><strong>10.5</strong> Load the potato genome and bam files</td>
<td>40</td>
</tr>
<tr>
<td><strong>10.6</strong> Look at a housekeeping gene</td>
<td>41</td>
</tr>
<tr>
<td><strong>10.7</strong> Look at a gene with differential expression</td>
<td>42</td>
</tr>
</tbody>
</table>
# Introduction

## 1.1 Scope of this tutorial
This is a practical, hands-on tutorial designed to give participants experience with RNA-Seq data analysis using Tophat and Cufflinks in Galaxy. The analysis in this tutorial is typical of experiments in plant species with adequate quality genomes and genome annotation. Participants are expected to be familiar with next-generation sequence data and basic theory of RNA-Seq. Participants do not need previous experience with Tophat, Cufflinks, or Galaxy.

## 1.2 Reference materials
RNA-Seq tutorial PDFs on MSI website: [www.msi.umn.edu/content/transcriptomics](http://www.msi.umn.edu/content/transcriptomics)
Tophat manual: [tophat.cbcb.umd.edu/manual.html](http://tophat.cbcb.umd.edu/manual.html)
Cufflinks manual: [cufflinks.cbcb.umd.edu/manual.html](http://cufflinks.cbcb.umd.edu/manual.html)
Galaxy screencasts: [galaxycast.org](http://galaxycast.org)

## 1.3 Loading Galaxy
a) Navigate your web browser to [http://galaxy.msi.umn.edu](http://galaxy.msi.umn.edu)
b) Log in using your MSI username and password

## 1.4 Outline of tutorial
1 Introduction
2 Loading Data Libraries
3 Quality control with FastQC
4 Mapping with Tophat
5 Galaxy Workflows
6 Visualizing alignments with Trackster
7 Computing differential expression with cuffdiff
8 Sample clustering with MDS plot
9 Cuffdiff visualization with CummeRbund
10 Appendix I: Workflows
2 Data Libraries

★ Tutorial Dataset (Sect 2.1 on page 5)
This tutorial will identify genes whose expression levels differ between pathogen-infected and mock-infected potato tubers. The sample dataset used in this tutorial was created from a private dataset kindly provided by Dr. James Bradeen that consists of three mock-infected and three samples infected with potato blight. Reads mapping to a 2MB region at the beginning of chromosome 5 were extracted along with some unmapped reads. Each fastq file contains about 75,000 50 base-pair paired-end reads.

NOTE: This dataset was chosen to allow for fast processing and response times in a classroom setting where dozens of people will be submitting jobs at once to the server. It is not ideal due to the small sample sizes (leading to atypical-looking graphs in some cases and poor statistics).

★ GTF/GFF Files (Sect 2.1 on page 5)
GTF and GFF files identify the genomic locations of genes and their exons. If a GTF or GFF file for your organism is not available in Galaxy send a request to MSI, or find one online at sites such as www.ensembl.org/info/data/ftp/index.html, genome.ucsc.edu/cgi-bin/hgTables?command=start, or NCBI. The GTF files provided in the Illumina iGenomes collection (cufflinks.cbcb.umd.edu/igenomes.html) have been specially modified for maximum compatibility with the Cufflinks program. The GFF file used in this tutorial was downloaded from the potato genome sequencing consortium website: solanaceae.plantbiology.msu.edu/data/PGSC_DM_v3_2.1.11_pseudomolecule_annotation.gff.zip

★ Reference Genomes (Sect 2.2 on page 6)
It is important that the genome database you select is generated from the same reference genome as the GTF/GFF you are using because the chromosome names and coordinates used in the GTF file must be the same as those used in the database. If the reference genome for your organism is not listed email a request to MSI to have it added.
2.1 Import Fastq files for one sample and annotation file into current history

- **Tutorial Dataset**
- **GTF/GFF Files**

  a) At the top of the screen click “Shared Data”
  b) Select “Data Libraries”
  c) Select “RISS-tutorial-Stuberosum” from the list of data libraries
  d) Expand the “Fastq” folder and check the boxes next to the first two files
  e) Check the box next to the “Stuberosum_DM_v3_chr05.gff” file
  f) Near the bottom of the page click the “Go” button to import the selected datasets to the current history
  g) Click on “Analyze Data” to return to the current analysis
2.2 Set file attributes: database/build and datatype

- **Reference Genomes**
  a) In the history pane click on the pencil icon next to the mock-1_R1.fastq file
  b) Enter “potato” in the “Database/Build” box. A list of available databases will appear as you type. Select the second (chr05) potato database
  c) Click save

d) Click the Datatype tab
  e) Enter “fastqsanger” in the “New Type” box. A list of available data types will appear as you type.
  f) Click save
3 Quality control with FastQC

★ Quality Control (Sect 3.1 on page 8)
It is important to always verify the integrity of a dataset before starting to analyze it. Quantifying dataset quality may uncover problems that might otherwise go undetected. Data quality problems such as sequencing adaptor contamination or low read quality require remedial processing not covered in this tutorial.

★ Galaxy Job Queue (Sect 3.1 on page 8)
Executing a program in Galaxy is called a job. Each job you start is submitted to a job queue and is run on a remote server together with other users’ jobs. After submitting a job you may continue to submit other jobs, or close your browser and return later to view the results of the job. Submitted jobs that haven’t run yet are shown in the history pane as grey with a clock icon. Running jobs are yellow, failed jobs are red, and successfully completed jobs are green.

★ FastQC Output (Sect 3.2 on page 8)
The graphs generated in this tutorial are not entirely typical due to the small sample datasets used. See the typical output from a good Illumina dataset: www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc/fastqc_report.html, and a poor Illumina dataset: www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc/fastqc_report.html. For more information about interpreting FastQC output refer to the RISS galaxy 101 handout https://www.msi.umn.edu/sites/default/files/Galaxy-101.pdf
3.1 Run FastQC

★ Quality Control

a) Load the FastQC tool from the tool pane: “NGS: QC and manipulation -> FastQC: Read QC”
b) Set the input file: select “mock-1_R1.fastq” from the dropdown menu under “Short read data from your current history”
c) Click “Execute”

★ Galaxy Job Queue

3.2 Review FastQC output

a) Click the eye icon to display the file in the center pane

★ FastQC Output

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
4 Mapping with Tophat

★ Mean Inner Distance – Part I (Sect 4.1 on page 10)
This is the expected (mean) inner distance between mate pairs. For example, the BMGC’s default fragment selection size is 200, so 200 – (2 * read length) is a good value to use for this parameter. We will determine the exact fragment length in the next section.

★ Junctions (Sect 4.1 on page 10)
Tophat can attempt to identify exon-exon splice junctions solely using your dataset, or you may supply a set of gene model annotations as a GTF or GFF file. In this tutorial we will provide a GFF annotation file because the potato genome has decent annotation.

★ Advanced Tophat Parameters (Sect 4.1 on page 10)
See https://www.msi.umn.edu/sites/default/files/RNA_seq_Mod3_v1.pdf for more details on setting parameters properly for other organisms.

★ Mean Inner Distance – Part II (Sect 4.2 on page 12)
It is important that the mean inner distance Tophat parameter is set correctly in order to get the best mapping results. The actual average fragment size for each sample can be determined by running Tophat with an estimated inner distance and then calculating the true value from the mapped reads. Rerunning Tophat with the true value will give improved results.

★ Insert Size Histogram (Sect 4.2 on page 12)
The insert size histogram generated from this sample dataset is noisier than a typical histogram, shown here:

★ Mapping Statistics (Sect 4.4 on page 13)
It is important to determine how well the RNA-Seq reads align to the reference genome. Low mapping rates require further investigation to determine the cause.
4.1 Initial Tophat run

- Mean Inner Distance – Part I
- Junctions
- Advanced Tophat Parameters

a) Load the Tophat2 tool from the tool pane: NGS: RNA Analysis -> Tophat2
b) Set the following options:
   - "Is this library mate-paired" -> ‘Paired-end’
   - “RNA-Seq FASTQ file, forward reads” -> ‘mock-1_R1.fastq’
   - “RNA-Seq FASTQ file, reverse reads” -> ‘mock-1_R2.fastq’
   - “Mean Inner Distance between Mate Pairs” -> ‘100’
   - “Select a reference genome” -> ‘Potato Stuberosum_DM_v3-chr05’
   - “TopHat settings to use” -> ‘Full parameter list’
Mapping with Tophat

c) Set these additional options:
- “Minimum intron length” -> ‘45’
- “Maximum intron length” -> ‘5000’
- “Minimum intron length found during split-segment (default) search” -> ‘45’
- “Maximum intron length found during split-segment (default search)” -> ‘5000’
- “Use Own Junctions” -> ‘Yes’
- “Use Gene Annotation Model” -> ‘Yes’

d) Click “Execute” to submit the job
4.2 Determine insert size

- Mean Inner Distance – Part II
- Insert Size Histogram
  a) Open the “NGS: Picard (beta)” section and choose the “Insertion size metrics” tool
  b) Click Execute
  
  ![Insertion size metrics](image)

  a
  b

c) Click on the “eye” icon next to the output file in the history pane to view the output in the central pane

d) Identify the mode (highest frequency) insert size from the program output

![Output view](image)

c d
4.3 Rerun Tophat with correct insert size

a) Click on the name of any one of the Tophat2 output files in the history pane to expand it
b) Click on the circular blue arrow icon to display the Tophat2 tool in the central pane with the parameters with which it was run

c) Change the “Mean Inner Distance between Mate Pairs” to the correct value: Picard value –
   \[2 \times \text{read length}\] = 165 – (2 \times 50) = 65

d) Click “Execute” to submit the job

4.4 Viewing General Tophat2 Statistics

★ Mapping Statistics

a) The Tophat2 run produced an “align_summary” file, click the ‘eye’ to view it.
5 Galaxy Workflows

In this version of the RNA-seq tutorial the section on workflows has been moved to Appendix I. The details of creating a workflow are also covered in the Galaxy 101 tutorial (Galaxy101).

**Galaxy Workflows (Sect 9.1 on page 31)**
All of the steps that have been performed on the mock-1 sample would need to be repeated for the two other mock-treated samples and the three pathogen-treated samples. Galaxy workflows provide an easy method to automate an analysis pipeline.

**Workflow Parameters (Sect 9.2 on page 33)**
The workflow we set up in this section will run FastQC, Tophat2 and Insertion size metrics. Tophat2 will be run just once using the inner mate distance calculated from the first sample. Samples that were sequenced together in the same batch often have very similar average insert sizes and the same inner mate distance can be used for all samples. Check the Insertion size metrics results after running the workflow to verify that is the case.

6 Visualizing alignments with Trackster

**Visualization (Sect 6.6 page 29)**
Visualizing alignments is a quick and easy way to check for major problems with the data. You may wish to verify that housekeeping genes are indeed roughly evenly covered with reads, or documented differentially-expressed genes indeed have differential coverage between samples of different groups.

**Galaxy Visualization Options (Sect 6.4 page 27)**
Galaxy supports three genome browsers for visualizing data:
- Trackster is a genome browser built into Galaxy. Any data file that can be viewed in Trackster will have a Trackster icon next to it in the history pane.
- The Integrative Genomics Viewer (IGV) is a well-regarded genome browser because it is fast, powerful, and easy to use for common genomes. However, it can be difficult to set up for less common genomes.
- The Integrated Genome Browser (IGB) is similar to IGV, but most users prefer to use IGV.

**Sample Dataset (Sect 6.1 page 24)**
In this section we start with Bam alignment files that have already been generated for all six mock- and pathogen-infected samples. These Bam files were generated using the workflow previously described in this tutorial.

**Using IGV (Appendix II: Visualizing alignments with IGV on page 36)**
Appendix II covers using IGV to visualize your sequence alignment. Currently IGV has a few useful features that are not available in Trackster. In particular IGV provides the ability to jump to chromosomal loci by searching for their 'common names'.
6.1 Load BAM alignment files and GTF into new history

★ Sample Dataset

a) Create a new history by clicking on the gear icon at the top of the history window and selecting “Create New” from the drop-down menu

b) Click on “Shared Data” at the top of the window

c) Choose “Data Libraries”

d) Click on the “RISS-tutorial-Stuberosum” data library

e) Expand the “Bam” folder

f) check the boxes next to each bam file

g) Check the box next to the “Stuberosum_DM_v3_chr05.gff” file

h) Click “Go” near the bottom of the center pane
6.2 View ‘bam’ and ‘gtf’ files in Trackster

a) Click “Analyze Data” to return to the history view

b) Click on “mock-1_accepted_hits.bam” in the history panel to expand it

c) Click the ‘Chart’ icon under the “mock-1_accepted_hits.bam” entry

d) Choose “Trackster” from the menu

e) Choose “View in new visualization”

f) Make sure the Potato genome (chr5) is selected

g) Click “Create”
h) Using the chromosome dropdown menu select “chr05”
i) Click the “Add Tracks” button (green square with a ‘+’)

j) Select all entries except “mock-1_accepted_hits.bam”
k) Click “add”

l) Click the “Save” Button (Floppy Disk)
m) Click “Analyze Data” to return to the history view
n) Expand the “Stuberosum_DM_v3_chr05.gff” entry in the history bar
o) Click the chart icon
p) Select “Trackster” from the drop down menu

q) Choose “View in save visualization”

r) Click the check box next to your saved visualization
s) Click “Add to Visualization”
6.3 Group Samples

a) Click the “Add Group” button twice (colored blocks with a ‘+’)

b) For both new “New Group”s use the “Edit Settings” menu to change its name. Set one to “Mock” the other to “Patho”

c) Use the areas of close set vertical lines on the far left to grab tracks and drag them to their proper groups. Put “Mock” sets in the “Mock” group and “Patho” datasets in the “Patho” group. Do not add the ‘gff’ file to a group.

d) Change the “Max value” for all tracks to ‘25’ by click on the number

---

Copyright 2013 Regents of the University of Minnesota

RISS Minnesota Supercomputing Institute, University of Minnesota
e) Use the “Show composite track” option on each group to provide a combined view of the datasets

f) Use the “Set display mode option” on each group to set the view to “filled”
6.4 Look at a housekeeping gene

★ Visualization

a) Click on the position range next to the chromosome dropdown. Enter “chr05:1419937-1429990” into the box. This will bring up the “ASC1” housekeeping loci.
b) Note the similar expression profiles within a group and between groups

6.5 Look at a gene with differential expression

a) Expand the tracks by clicking “Show individual tracks”
b) Set the max value for each track to ‘150’
c) Collapse the tracks by clicking “Show composite Track”
d) Enter “chr05:237646-239446” in the position box to view a gene that is expected to be differentially expressed

e) Note the differential expression
Computing differential expression with cuffdiff

★ Cuffdiff Output (Sect 0 on page 25)
Cuffdiff produces many output files. In this tutorial we look at the gene differential expression testing file which shows which genes are differentially expressed. The other output files also contain important data, including the results of differential expression testing for spliced transcripts, primary transcripts, and coding sequences. See the cufflinks manual for detailed information about what information is in each file: cufflinks.cbcb.umd.edu/manual.html - cuffdiff_output

★ Differential Gene Expression (Sect 0 on page 25)
The gene differential expression testing output file is a tab-delimited text file with one row for each gene. Our sample dataset only covers a small portion of chr05 so most genes will have too few aligned reads for a differential expression test. These genes are indicated with “NOTEST” or “LOWDATA” in column 7.

★ De novo gene/transcript discovery (Sect 7.1 on page 24)
The analysis pipeline used in this tutorial will quantify the expression of known genes in a reference annotation. If you are interested in discovering novel genes or spliceforms more steps need to be added to the pipeline. Refer to the Nature Protocols paper “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks” for more information: www.ncbi.nlm.nih.gov/pubmed/22383036

★ Column Summary (Sect 7.2 on page 25)
Clicking on files in the history view will expand the box to include additional information. This can be particularly useful for remembering column numbers when setting the parameters for a tool such as the “Filter” tool.
Computing differential expression with cuffdiff

7.1 Run cuffdiff

★ De novo gene/transcript discovery

a) Open the “NGS: RNA Analysis” section of the ‘Tools’ panel
b) Open the “Cuffdiff” program
c) Set “Perform replicate analysis” -> Yes
d) Click twice on “Add new Group”
e) Name:
   - Group1 -> “mock”
   - Group 2 -> “patho”

f) Add 3 replicates to each group (by clicking “Add new Replicates” x3 per group)
g) Set the three ‘mock’ bam files as the three replicates in Group 1, and the three ‘patho’ bam files as the three replicates in Group 2
h) Under the “Select outputs for history datasets,” click “Select All”

i) Click “Execute” to submit the job

7.2 View and filter cuffdiff output

★ Cuffdiff Output
★ Differential Gene Expression
★ Column Summary

a) View the Cuffdiff output file “gene differential expression testing” by clicking on the “eye” icon next to the filename in the history pane
b) Open the “Filter and Sort” section of the toolbar

c) Open the “Filter” tool

d) Set the Cuffdiff output file “gene differential expression testing” as the file to filter

e) Filter out genes with significant change in expression with a log fold-change of at least 1 by entering “c14 == ‘yes’ and abs(c10)>1” in the “with following condition” text box

f) Click “Execute” to submit the job

g) Click on the “eye” icon next to the filter output filename to view the results in the center pane
8 Cuffdiff visualization with CummeRbund

★ CummeRbund (Sect 8.1 on page 28)
CummeRbund is an easy to use R package that takes the output files from a cuffdiff run and creates a SQLite database of the results. This allows the user to explore data for genes, transcripts, transcription start sites, and CDS regions across multiple samples or conditions. CummeRbund implements numerous plotting functions for commonly used visualizations. The CummeRbund wrapper in Galaxy allows easy access to much of CummeRbund’s functionality. For more details about available plots refer to the CummeRbund website: compbio.mit.edu/cummeRbund/

★ Density Plots (Sect 8.2 on page 29)
A Kernel density plot is interpreted the same as a histogram. The density plot shows the distribution of gene expression levels across different samples. All samples should have reasonably similar distributions. A log10(FPKM) of 0 = 1 FPKM, which is very low expression.

★ MDS Plots (Sect 8.2 on page 29)
MDS plots are similar to Principle Component Analysis (PCA) plots. They are useful for identifying clusters of samples with similar expression patterns. Ideally samples from the same experimental group will be clustered together in the plot indicating that experimental condition is the major source of variation. Samples might also cluster by age, batch, date, technician, or other technical aspect of the experiment.

★ Dendrogram (Sect 8.2 on page 29)
A dendrogram is a tree diagram showing how sample cluster by similarity. Ideally samples from the same experimental group are clustered together.
8.1 Run CummeRbund tool

- CummeRbund
  
  a) Open the “NGS: RNA Analysis” section of the ‘Tool’ panel
  
  b) Open the “cummeRbund”
  
  c) Add three new plots (click three times to generate three plots)

  ![Image of CummeRbund tool interface]

  d) For Plot 1 set “Plot Type” -> “Density” and check the “Replicates?” box
  
  e) For Plot 1 set “Plot Type” -> “MDS” and check the “Replicates?” box
  
  f) For Plot 1 set “Plot Type” -> “Dendrogram” and check the “Replicates?” box
  
  g) Click “Execute” to submit the job

  ![Image of CummeRbund interface with plots displayed]

  Have patience when setting the CummeRbund parameters. After changing each setting it takes several seconds for the center pane to reload. This is common when working with large histories.
8.2 Review CummeRbund plots

★ Density plots, MDS plots, and Dendrograms

a) Click the “eye” icon next to the cummerbund output file to view the three plots

b) Verify that:
- The samples have similar density distributions
- The samples cluster by experimental condition in the MDS plot
- The sample cluster by experimental condition in the dendogram
8.3 Additional CummeRbund plots:

a) Volcano, Heatmap, Expression Plot, and Cluster.

8.4 Troubleshooting

If you experience problems using Galaxy send an email to help@msi.umn.edu with a subject beginning “RISS” and a report of the problem.
9 Appendix I: Workflows

9.1 Extract workflow from current history

Galaxy Workflows

a) At the top of the history pane click on the small gear icon and select “Extract Workflow” from the pop-up menu

b) In the “Workflow name” box enter “QC and Tophat”

c) Uncheck the second (closest to the bottom) Tophat2 run

d) Click “Create Workflow”
9.2 Edit the workflow

a) Click on “Workflow” at the top of the Galaxy window
b) Click on the workflow that was just created and select “Edit” from the drop-down menu
c) Move the elements of the workflow around to make it easier to see how they are connected.
d) Click on the first Input dataset box and set the Name field to ‘forward read’. Repeat for second input dataset (‘reverse read’).

continued on next page...
Appendix I: Workflows

e) Click on the Tophat2 box to display the Tophat2 options in the “Details” pane on the right side.

f) Set the “Mean Inner Distance between Mate Pairs” to 65.

h) Connect the Tophat2 output “accepted_hits (bam)” to the Tophat Stats PE input “BAM file produced by Tophat” by clicking and dragging

j) Return to your history by clicking on “Analyze Data” at the top of the screen

★ Workflow parameters

- **Input dataset output**
  - FastQ/Read QC
    - Short read data from your current history
    - Contaminant list
    - html file (html)

- **Input dataset output**
  - Tophat2
    - RNA-Seq FASTQ file, forward reads
    - RNA-Seq FASTQ file, reverse reads
    - Gene Model Annotations
    - fusions (tabular)
    - insertions (bed)
    - deletions (bed)
    - junctions (bed)
    - accepted_hits (bam)

- **Tophat Stats PE**
  - bam file produced by Tophat
  - Fastq file used to produce the BAM file
  - output (txt)

- **Select a reference genome**
  - Potato Stuberorum DM_v3.0-1
9.3  Create new history

a) Rename the current history: at the top of the history pane click on “Unnamed history” and rename it “mock-1”. (NOTE: you must hit ‘Enter’ after typing the new name, rather than clicking outside the box.)

b) Create a new history by clicking on the gear icon at the top of the history pane and selecting “Create New” from the pop-up menu

c) Name the new history “mock-2”

d) Import the mock-2 fastq files and the GFF files by clicking on “Shared Data -> Data Libraries” at the top of the screen and selecting the “mock-2_R1.fastq”, “mock-2_R2.fastq”, and “Stuberosum_DM_v3_chr05.gff” files from the “RISS-tutorial-Stuberosum” data library

e) Return to your history by clicking on “Analyze Data” at the top of the screen
9.4 Run workflow

a) Load a workflow by clicking on “Workflow” at the top of the screen
b) Click on the workflow that was just created and select “Run” from the dropdown menu
c) Select the “mock-2_R1.fastq” file in the first drop-down menu and the “mock-2_R2.fastq” file in the second drop-down menu
d) Verify the GFF file is selected in the third drop-down menu
e) Click on “Run workflow” to submit the FastQC, Tophat2, Insertion size metrics, and TophatstatsPE jobs.
10 Appendix II: Visualizing alignments with IGV

10.1 Load BAM alignment files and GTF into new history

a) Create a new history by clicking on the gear icon at the top of the history window and selecting “Create New” from the drop-down menu
b) Click on “Shared Data -> Data Libraries” at the top of the window
c) Click on the “RISS-tutorial-Stuberosum” data library
d) Expand the “Bam” folder and check the box next to each bam file
e) Check the box next to the “Stuberosum_DM_v3_chr05.gff” file
f) Click “Import to current history” near the bottom of the center pane
10.2 Load Genome Fasta file into current history

a) Click on “Shared Data -> Data Libraries” at the top of the window
b) Click on the “Genomes” data library
c) Expand the “Solanum_tuberosum” folder and check the box next to the “Stuberosum DN_v3-chr05.fa” file
d) Click “Import to current history” near the bottom of the center pane
e) Return to your history by clicking on “Analyze Data” at the top of the screen
10.3 Download files to your desktop

a) Click on the “mock-1_accepted_hits.bam” file in the history pane to expand it and click on the floppy disk icon to download the dataset and the bam_index

b) Click on the “patho-1_accepted_hits.bam” file in the history pane to expand it and click on the floppy disk icon to download the dataset and the bam_index

c) Click on the “Stuberosum_DM_v3_chr05.gff” file in the history pane to expand it and click on the floppy disk icon to download the reference gene annotation file

d) Click on the “Stuberosum_DM_v3_chr05.fa” file in the history pane to expand it and click on the floppy disk icon to download the reference genome sequence
10.4 Start IGV

a) In a new browser window load the IGV homepage: [www.broadinstitute.org/igv/](http://www.broadinstitute.org/igv/)
b) In the “Download” section click the “register” link and complete the registration form
c) Click the “Launch with 750 MB” button
d) Open the downloaded file “igv.jnlp”
10.5 Load the potato genome and bam files

a) In IGV select the “File -> Import Genome...” menu option

b) Enter a name and description for the potato genome, and select the FASTA file and Gene file

c) Select the “File -> Load from File...” option and select the two downloaded bam files
10.6 Look at a housekeeping gene

a) Right-click on the Gene track and select “expanded”

b) Enter “ASC1” (with quotes) in the search box to view the reads aligning to a gene that is expected to have similar expression levels in both experimental groups

c) Right-click on the mock-1 coverage track and select “Autoscale” from the pop-up menu to deactivate it

d) Right-click on the mock-1 coverage track and select “Set Data Range”

e) Set the “Max” value to 25

f) Repeat for the patho-1 coverage track
10.7 Look at a gene with differential expression

a) Enter “Zeatin O-glucosyltransferase” in the search box to view the reads aligning to a gene that is expected to be differentially expressed

b) Set the coverage track data range maximums to 150