C3G Analysis Workshop - Visualization and reporting

J. Hector Galvez

2019-01-22

C3G Analysis Workshop - Visualization and reporting

J. Hector Galvez

2019-01-22

Introduction

Visualization and summarization are the last steps of an RNA-seq analysis. But in many respects, they are the most important part of the analysis. These final steps will enable the interpretation of results to answer scientific questions. GenPipes generates an HTML report that facilitates the understanding and exploration of results. Additionally, using track files or alignment files, users can visualize their alignments with a genome browser directly.

Learning objectives

1. Learn how to generate the GenPipes report.
2. Understand the contents of the GenPipes report.
3. Learn how to open alignment and track files using genome browsers.

** Note:** We will be generating the GenPipes report on the server and then move it to our laptop for visualization.

Theoretical Background: Interpreting RNA-Seq results

Each scientific project is unique. The experiments consist of different samples, different designs, and they are ultimately trying to answer different questions.

Therefore, the downstream analysis and interpretation of RNA-seq results can vary greatly between projects. This is why, it is important to understand the outputs of any standard RNA-seq analysis, and to tie those results to the respective reserarch objective. GenPipes tries to facilitate this process by generating a report in which results are grouped by step and summarized using tables and figures. This HTML report is a first step to understanding the output of an analysis, and can help guide the direction of additional downstream analyses by providing convenient links and descriptions.

Exercise 1: Generating the RNA-Seq report using GenPipes

GenPipes does not generate a report every time the pipeline is run, to avoid reporting incomplete or intermediate results. To generate a report, a user must manually run the pipeline script with the `--report` flag, which will instruct GenPipes to search through the output files and produce a report. Aftewards, the report can be downloaded and opened with an internet browser to explore interactively. This exercise will guide users through the process of generating and opening a report using GenPipes.

Step A: Run the pipeline script to generate the report and open it

Reports can be generated for one or more steps using the same command used to launch a GenPipes analysis, with the addition of the `--report` flag.

Question 1

1. Generate a report script for all the steps in the analysis that have been done in this workshop, and then run the script to generate the report.

Solution (click here) *Command:*

```
rnaseq.py -c $MUGQIC_PIPELINES_HOME/pipelines/rnaseq/rnaseq.base.ini $MUGQIC_PIPELINES_HOME/
    -r readset.rnaseq.txt \
    -d design.rnaseq.txt \
    -j slurm \
    -o . \
    --report \
    -s 1-23 > report_steps1-23.sh

bash report_steps1-23.sh
```

Question 2

2. Once you have run the previous step without errors, make sure the report was created by looking into the `report` directory and verifying that `index.html` exists.

Solution (click here) *Command:*

```
cd report

ls
```

Question 3

3. If the report has been generated and `index.html` can be found inside the directory, download the full report folder into your computer so you can open it with your browser.

Solution (click here) Use CyberDuck to download the results to your local computer. Alternative, you also can download it from the server using `Rsync`:

Command:

```
#From the terminal on your laptop type the command AFTER you change <my_cc_account> to your
#rsync -rltv <USERNAME>@<SERVER_ADDRESS>:<PATH/TO/REPORT> LOCAL/PATH
rsync -rltv <my_cc_account>@mp2.ccs.usherbrooke.ca:/home/<my_cc_account>/RNAseq_workshop/RNA
```

Make sure you are downloading the full **report** directory, and not just `index.html`, otherwise the report will not display properly.

Step B: Understand the different parts of the report

Once you have saved the report in your computer, make sure you can open it with your internet browser.

Question 1

1. Using your file explorer tool, go to the place where the report folder has been saved. Double click on the `index.html` file and wait for it to open in your internet browser. Once it has opened, read through the titles of each section and try to understand what information each section contains.

Solution (click here)

If the report was properly generated, it will have the following section titles:

- Introduction
- Analysis and Results
- Read Trimming and Clipping of Adapters
- Read Alignment to a Reference Genome
- Trimming and Alignment Metrics per Sample

- Wiggle Graphs/Tracks Generation
- FPKM Analysis
  - FPKM Values Generation
- Metrics and Exploratory Analysis
  - FPKM Metrics
  - Exploratory Analysis
- Transcriptome Assembly with Cufflinks
- Differential Expression Analysis - Methods
  - Differential Analysis Design
  - Differential Gene Analysis Description
  - Differential Transcripts Analysis Description
  - Gene Ontology (GO) Analysis of the Differential Expression Results
- Differential Expression Analysis - Results
- H1ESC_GM12787 Results
- Analysis Configuration Parameters
- References

Read the descriptions of each section to make sure you understand what it contains. Try to download the full tables for each section by clicking on the link labelled: **download full table**. Open the table using a spreadsheet tool (such as Excel) or using a text editor.

Step C: Interpret the plots produced as part of the report

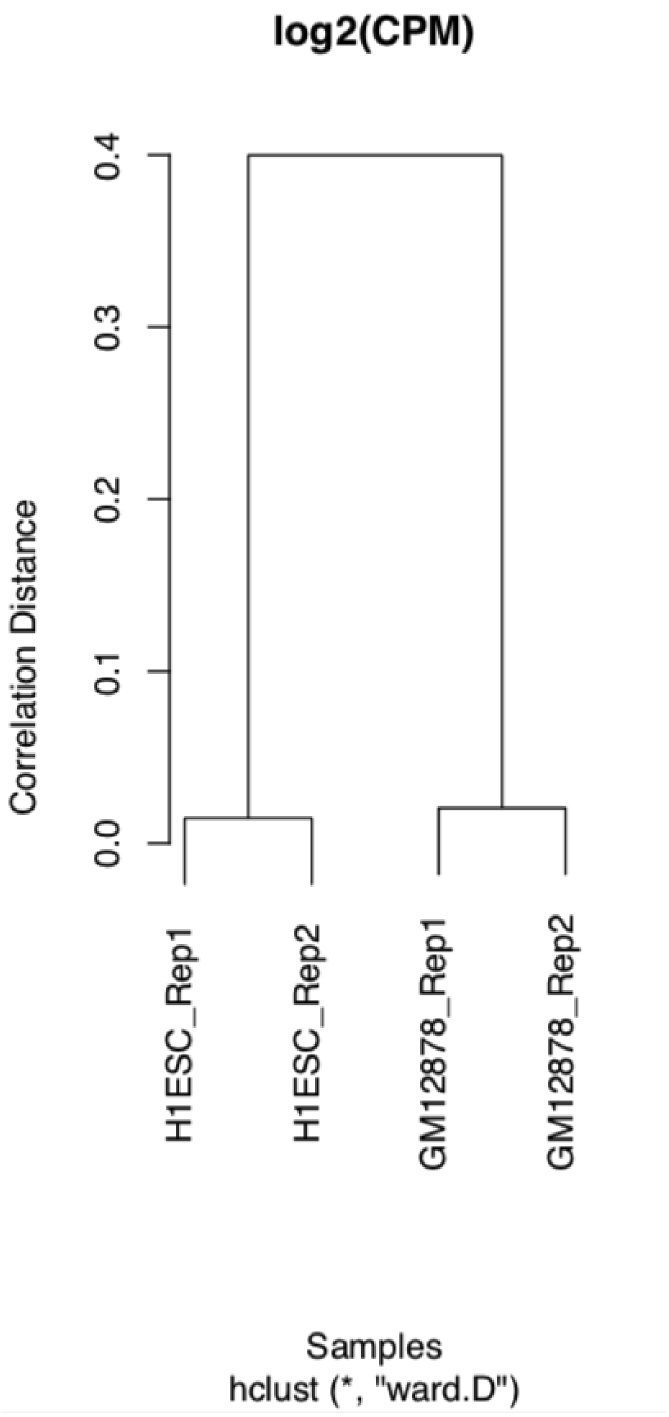
GenPipes generates several plots for data exploration and quality assurance. It is important to understand what these plots contain and what they look like.

Question 1

1. Open the `index.html` file with your internet browser and go to the section labelled `Exploratory analysis`. What figures are contained in this section? Open each figure and understand what they are depicting.

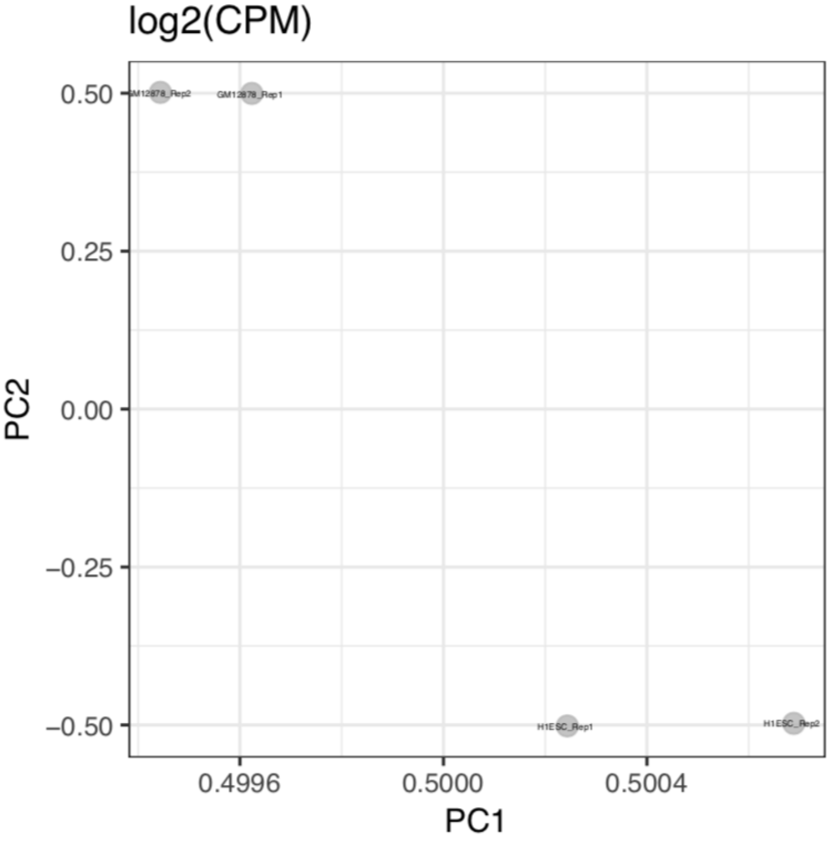
Hierachical clustering based on the correlation distance, log2(CPM) (click here)

This figure shows how the samples cluster based on their gene expression levels (in log2CPM). We expect to see samples from the same experimental group clustering next to each other, instead of clustering with samples from different groups.


log2(CPM)
Correlation Distance
0.0
0.1
0.2
0.3
0.4
H1ESC_Rep1
H1ESC_Rep2
GM12878_Rep1
GM12878_Rep2
Samples
hclust (*, "ward.D")

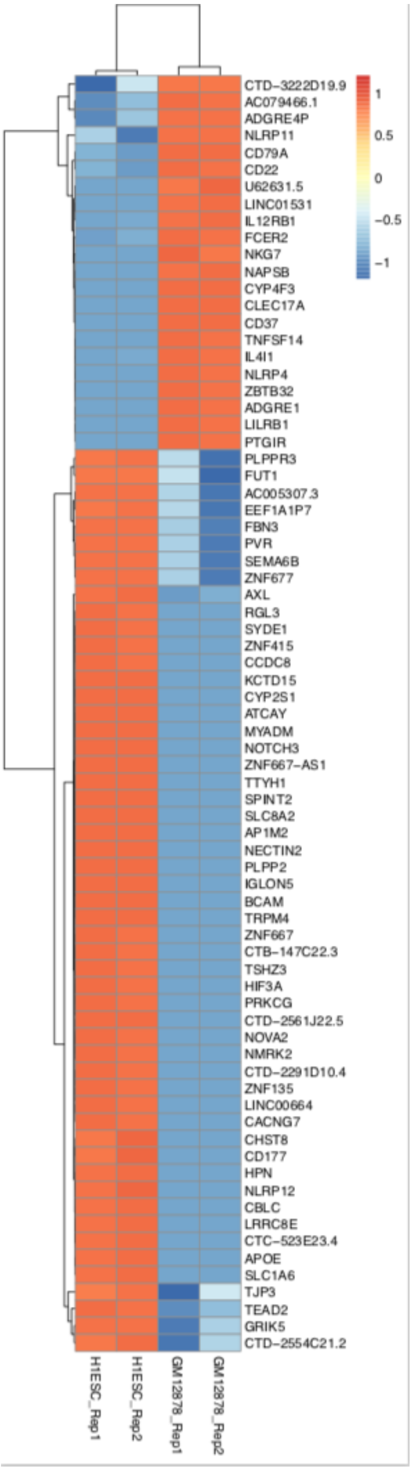
PCA (first two components) of the gene log2CPM values (click here)

A PCA plot shows the projection of the samples into a two-dimensional space, based on their first and second principal components. The principal components are usually enough to differentiate groups of samples, based on the experimental design.



Heat map of most varying genes and most varying transcripts (click here)

To get a rough overview of the genes and transcripts that are contributing the most to the samples' clustering pattern, heatmaps of the most variable genes and transcripts are created. Please note that these genes have not been selected to reflect any particular gene group or pathway, and therefore the biological relevance of these hetamaps is limited. They are mostly used to make sure that there are no biases underlying the final results.


CTD−3222D19.9
AC079466.1
ADGRE4P
NLRP11
CD79A
CD22
U62631.5
LINC01531
IL12RB1
FCER2
NKG7
NAPSB
CYP4F3
CLEC17A
CD37
TNFSF14
IL4I1
NLRP4
ZBTB32
ADGRE1
LILRB1
PTGIR
PLPPR3
FUT1
AC005307.3
EEF1A1P7
FBN3
PVR
SEMA6B
ZNF677
AXL
RGL3
SYDE1
ZNF415
CCDC8
KCTD15
CYP2S1
ATCAY
MYADM
NOTCH3
ZNF667−AS1
TTYH1
SPINT2
SLC8A2
AP1M2
NECTIN2
PLPP2
IGLON5
BCAM
TRPM4
ZNF667
CTB−147C22.3
TSHZ3
HIF3A
PRKCG
CTD−2561J22.5
NOVA2
NMRK2
CTD−2291D10.4
ZNF135
LINC00664
CACNG7
CHST8
CD177
HPN
NLRP12
CBLC
LRRC8E
CTC−523E23.4
APOE
SLC1A6
TJP3
TEAD2
GRIK5
CTD−2554C21.2
1
0.5
0
−0.5
−1
H1ESC_Rep1
H1ESC_Rep2
GM12878_Rep1
GM12878_Rep2

All figures (click here)

This zip file contains all the figures discussed above, as well as additional figures that are mostly variations of the ones previously mentioned. They are not always useful, but if a surprising result is detected in the previous 4 plots, it is useful to open all the additional figures to gain additional insight.

Exercise 2: Explore the data interactivelly using the IGV Genome Browser

For this exercise, you will require the IGV Genome Browser software. If you have not done so, please download and install it now before continuing.

Step A: Open the BAM files using IGV

The Integrative Genomics Viewer, IGV is a desktop genome browser that has many powerful tools that help visualize and analyze genomics data. It has a lot of pre-loaded data, including several reference genomes for model organisms as well as pre-defined parameters for common SAM/BAM flags. Because of its graphical user interface, the best way to learn IGV is actually to explore data interactivelly and search through the menus and options.

Question 1

1. Download an alignment file and its index to your local computer from the pipeline results.

Solution (click here)

From the `alignment` folder choose a sample. Use CyberDuck to download the alignment to your local computer desktop. Rembember to download both the alignment (`sorted.mdup.bam`) and index (`.bai`) files for that sample. Repeat for as many samples as you want to visualize.

You can also use `Rsync` to do the same thing, using as an example the following command (from your local computer):

```
cd  ~/Desktop
```

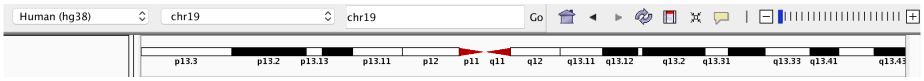
```
rsync -rltv <USERNAME>@<SERVER_ADDRESS>:<PATH/TO/PROJECT>/alignment/GM12878_Rep1/GM12878_Rep
```

Question 2

2. Open the *IGV* program using the icon in your application launcher. Change the reference genome to `GRCh38` and open the alignment file you just downloaded to your desktop.

Solution (click here)

To change the reference genome, select **"Human (hg38)"** from the drop-down menu in the top left corner of the toolbar. Then, from the second menu to the right, select **chr19** to visualize only chromosome 19 (see the figure below).



IGV Toolbar

Once the appropriate reference has been selected, open the alignment file by clicking `File > Load from File...` and then selecting `GM12878_Rep1.sorted.mdup.bam` from the desktop.

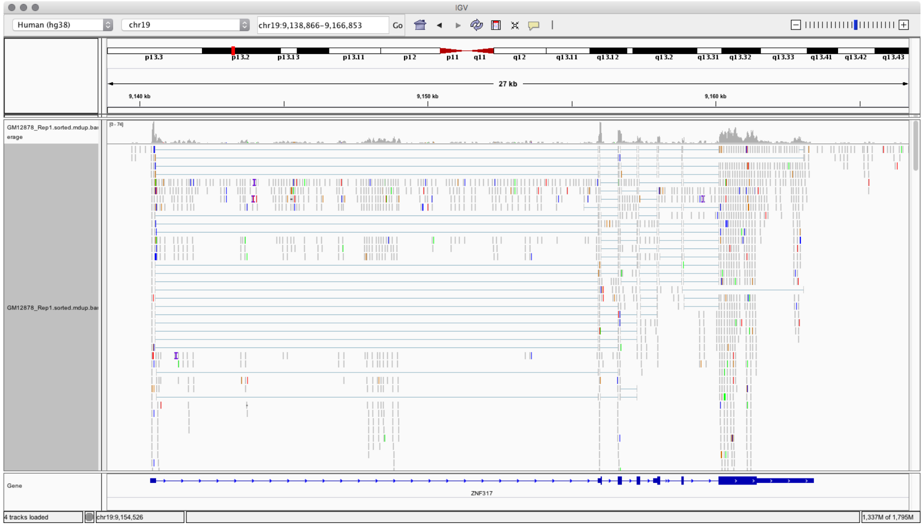
Remember that in order to open an alignment file on IGV, an index file (`.bai`) must always be available in the same directory

Question 3

3. Explore the data from the alignment using the genome browser and familiarize yourself with the interface.

Solution (click here)

The figure below shows an example of how the data is displayed once the alignment files are open:



Alignment Data on IGV

The following are useful exercises to practice using IGV to visualize alignment data:

- **Zoom in and out using the slider in the top right hand corner of the interface**
  - It is also possible to zoom to a particular locus in the chromosome by selecting it in the cartoon representation using your mouse.
- **Use the gene annotation track at the bottom to find annotated genes**
  - Observe how the reads align to all the components of the gene (exons, introns, UTRs, etc.).
  - Do these alignments make sense?
- **Use the search bar in the *top centre* of the toolbar to find a specific gene**
  - Try typing the name of a gene that was found to have differential expression between the two cell lines.
  - The following options are available to search for a gene:
    - Gene symbol
    - Gene ID (e.g. ENSEMBL)
    - Gene coordinates
- **Right-click on the *track* panel and see the options displayed in the menu**
  - Try changing the range of the coverage track.
  - Notice how that modifies the view of the data without actually changing any results.
- **Once you feel more confident, try downloading and opening other alignment files**
  - How does opening more than one alignment affect the visualization of the data?
  - Are the ranges of all tracks the same?
  - How can the data ranges be changed so that all tracks have the same scale?

Optional Step: **Open the track files using the UCSC Genome Browser**

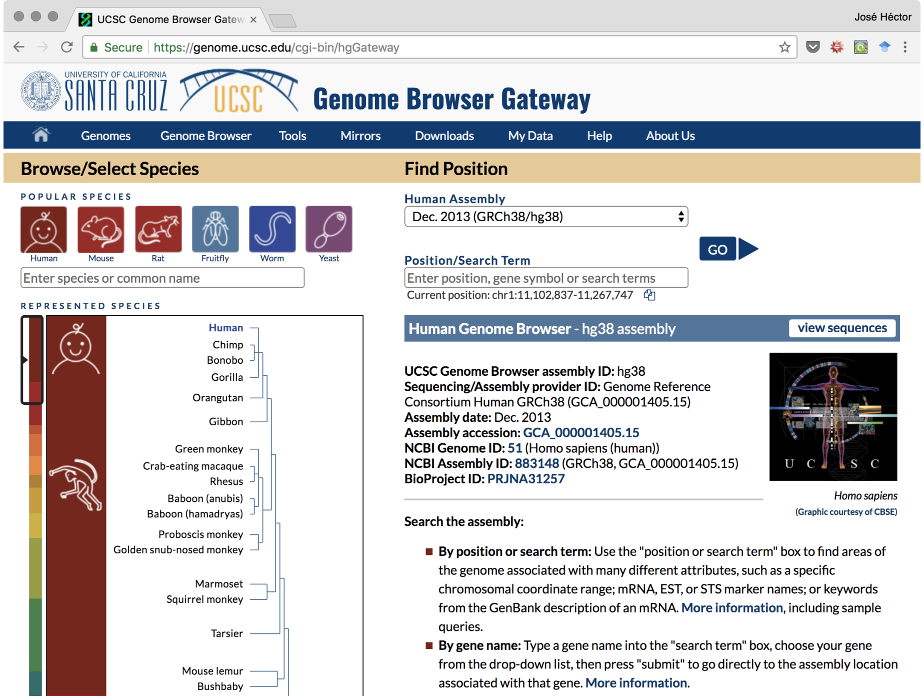
The UCSC Genome Browser is a popular web-based genome browser that is useful for exploring tracks of data and comparing them with one or several curated annotations. Its main strength lies in the fact that it comes pre-loaded with many tracks that contain information ranging from genes and transcripts, to epigenomic marks, making it easy to see overlaps with the transcriptomic data. Unfortunately, it also requires you to host your own files in order to load them into the browser. This falls outside the scope of this workshop, but should you have hosting space available and you are intersted in using this browser, the instructions on how to load GenPipes results can be found below:

Instructions to open BigWig files with UCSC Genome Browser (click here)

1. Open the genome browser using the appropriate version of the human

genome (hg38).

Follow the following link to the Genome Browser Gateway: https://genome.ucsc.edu/cgi-bin/hgGateway. The page that opens should look as follows:



Current position: chr1:11,102,837-11,267,747

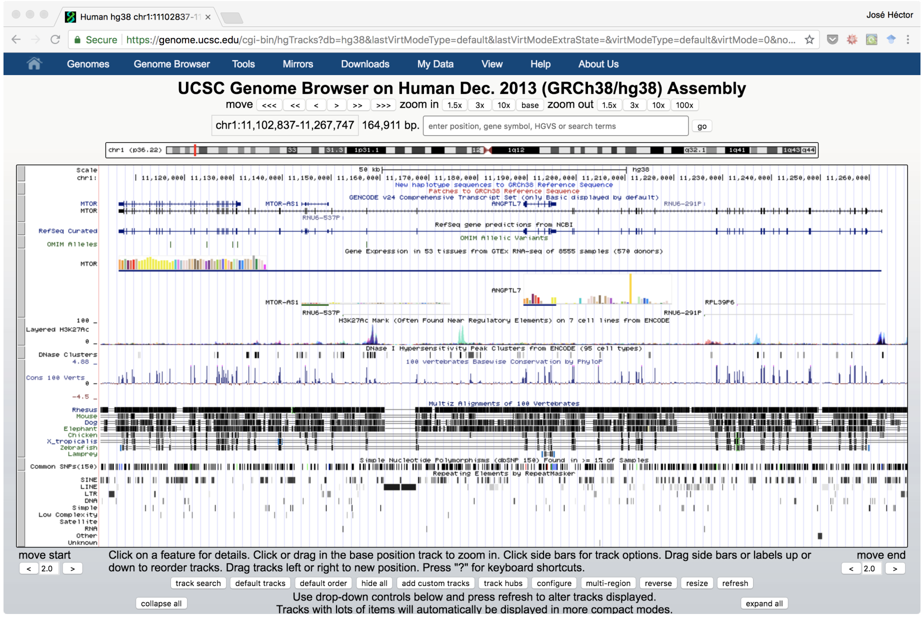
UCSC Genome Browser assembly ID: hg38
Sequencing/Assembly provider ID: Genome Reference Consortium Human GRCh38 (GCA_000001405.15)
Assembly date: Dec. 2013
Assembly accession: GCA_000001405.15
NCBI Genome ID: 51 (Homo sapiens (human))
NCBI Assembly ID: 883148 (GRCh38, GCA_000001405.15)
BioProject ID: PRJNA31257

Search the assembly:

- **By position or search term:** Use the "position or search term" box to find areas of the genome associated with many different attributes, such as a specific chromosomal coordinate range; mRNA, EST, or STS marker names; or keywords from the GenBank description of an mRNA. More information, including sample queries.
- **By gene name:** Type a gene name into the "search term" box, choose your gene from the drop-down list, then press "submit" to go directly to the assembly location associated with that gene. More information.

Genome Browser Gateway

From the drop down menu in the top, select the human assembly labeled *Dec.2013(GRCh38/hg38)* then press **Go**. The Genome Browser should open and produce a view similar to this one:


Human hg38 chr1:11102837-1
José Héctor
Secure https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&no...
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr1:11,102,837-11,267,747 164,911 bp.
enter position, gene symbol, HGVS or search terms
go
Scale
chr1:
New haplotype sequences to GRCh38 Reference Sequence
Patches to GRCh38 Reference Sequence
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
MTOR
MTOR-AS1
ANGPTL7
RNU6-291P
RNU6-537P
RefSeq Curated
RefSeq gene predictions from NCBI
OMIM Alleles
OMIM Allelic Variants
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors)
RPL39P6
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
Cons 100 Verts
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Common SNPs(150)
Simple Nucleotide Polymorphisms (dbSNP 150) Found in >= 1% of Samples
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
move start
< 2.0 >
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.
move end
< 2.0 >
track search
default tracks
default order
hide all
add custom tracks
track hubs
configure
multi-region
reverse
resize
refresh
collapse all
Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.
expand all

Main view of the UCSC Genome Browser

2. Download the GenPipes track files from the server and save them to your local computer.

Use CyberDuck to download tracks from the main results directory as a zipped file called `tracks.zip`.

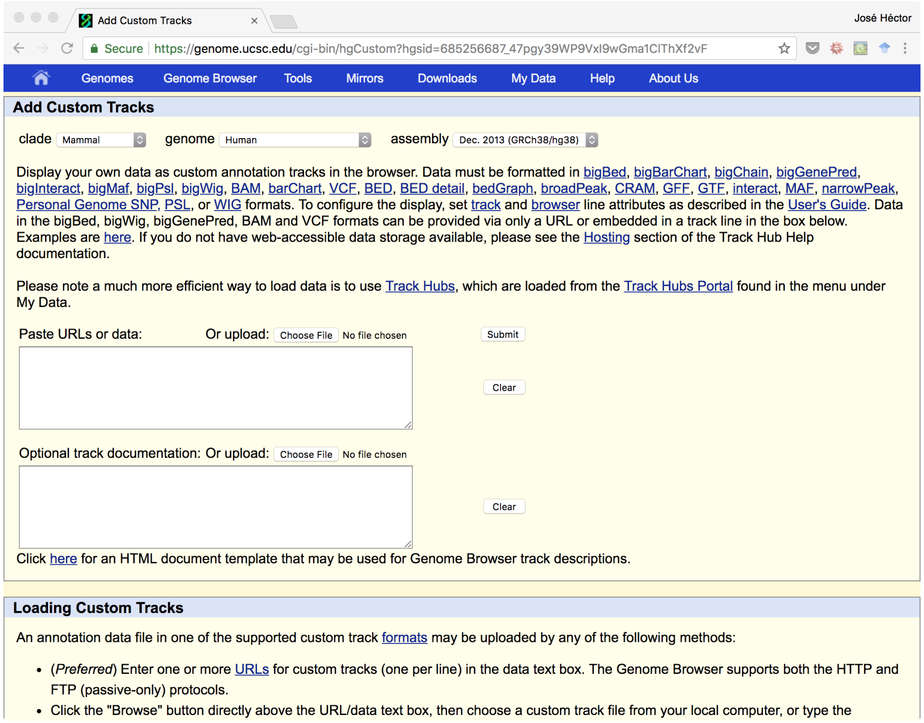
Decompress the `tracks.zip` file and open the directory inside called `BigWig`. Make sure you have two `.bw` files (one `forward` and one `reverse`) for each sample.

3. Upload your BigWig files to your preferred hosting service

To use `.bw` files with the UCSC Genome Browser you need to host them somewhere that is accessible via URL. This is not under the scope of this workshop, so follow the instructions provided by your hosting service.

4. Load the track files to the genome browser

From the Genome Main View, click on the **add custom tracks** button. This will open a window that looks like this:



Add Custom Tracks

clade Mammal genome Human assembly Dec. 2013 (GRCh38/hg38)

Display your own data as custom annotation tracks in the browser. Data must be formatted in bigBed, bigBarChart, bigChain, bigGenePred, bigInteract, bigMaf, bigPsl, bigWig, BAM, barChart, VCF, BED, BED detail, bedGraph, broadPeak, CRAM, GFF, GTF, interact, MAF, narrowPeak, Personal Genome SNP, PSL, or WIG formats. To configure the display, set track and browser line attributes as described in the User's Guide. Data in the bigBed, bigWig, bigGenePred, BAM and VCF formats can be provided via only a URL or embedded in a track line in the box below. Examples are here. If you do not have web-accessible data storage available, please see the Hosting section of the Track Hub Help documentation.

Please note a much more efficient way to load data is to use Track Hubs, which are loaded from the Track Hubs Portal found in the menu under My Data.

Paste URLs or data: Or upload:

Optional track documentation: Or upload:

Click here for an HTML document template that may be used for Genome Browser track descriptions.

Loading Custom Tracks

An annotation data file in one of the supported custom track formats may be uploaded by any of the following methods:

- (*Preferred*) Enter one or more URLs for custom tracks (one per line) in the data text box. The Genome Browser supports both the HTTP and FTP (passive-only) protocols.
- Click the "Browse" button directly above the URL/data text box, then choose a custom track file from your local computer, or type the

Track Upload View

In that window, paste the URL links to the `.bw` files you just uploaded. Make sure to include all the files and then click **Submit**. Wait until the files finish loading and then, return to the Genome Browser to visualize your data.