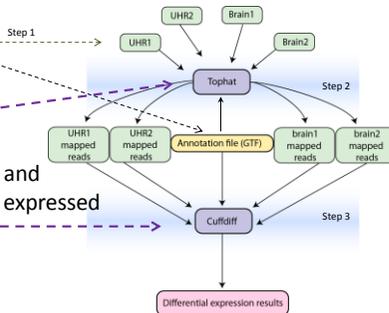


# Hands-on 1

## Quantification and assay for differential expression of reference annotation

1. Get input files
2. Map reads with Tophat
3. Quantify and find genes and transcripts differentially expressed with Cuffdiff



These are our input files

Reads are 35 nt long in fastqsanger format. You will have to map each of the 4 fastqsanger input files with tophat. Reads were downloaded from GEO: <http://www.ncbi.nlm.nih.gov/sra/SRX016368>.

**Note:** when you do your own analysis, if your reads are not in fastqsanger format you have to convert them to fastqsanger using the **NGS: QC and manipulation -> FASTQ Groomer** tool.

## Step 1: Get input Files

- Click on the **Shared Data** Tab and select **Published Histories**

- Select **RNAseqInputs\_1\_DifferentialGeneExpression**

Click on **Import history** then click on **start using this history**

## Step 2: Map each of the four fastqsanger read with Tophat (v1.3.3)

Red arrows show the parameters that you have to set, or settings I want you to notice. Leave all other settings as they are.

Check next slide for the rest of the parameters. This screenshot doesn't show the full interface.

# Step 2\_cont: Map each of the four fastqsanger read with Tophat (v1.3.3)

Number of mismatches allowed in each segment alignment for reads mapped independently: 2

Minimum length of read segments: 17

Use Open Translation: Yes

Use Gene Annotation Model: Yes

Gene Model Annotation: hg19.refGene.gtf

Use Clustal Search: No

Use Coverage Search: Yes

Minimum intron length that may be found during coverage search: 10

Maximum intron length that may be found during coverage search: 20000

Use Microarray Search: No

With this option, the tool will attempt to find alignments incident to microarrays; works only for reads 50bp or longer.

**After selecting all parameters click Execute**

Hint: Once you have sent the first sample you can click on the rerun icon and change only the fastq input file to submit the other samples. All the other settings will be already selected.

# Outputs from step 2 (Tophat)

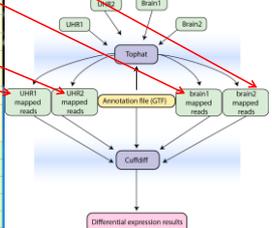
History

Imported: RNAseqtophats\_1\_DifferentialExpression

- 21. Tophat for Illumina on data 5 and data 4: accepted\_hits
- 20. Tophat for Illumina on data 5 and data 4: splice junctions
- 19. Tophat for Illumina on data 5 and data 4: deletions
- 18. Tophat for Illumina on data 5 and data 4: insertions
- 17. Tophat for Illumina on data 5 and data 3: accepted\_hits
- 16. Tophat for Illumina on data 5 and data 3: splice junctions
- 15. Tophat for Illumina on data 5 and data 3: deletions
- 14. Tophat for Illumina on data 5 and data 3: insertions
- 13. Tophat for Illumina on data 5 and data 2: accepted\_hits
- 12. Tophat for Illumina on data 5 and data 2: splice junctions
- 11. Tophat for Illumina on data 5 and data 2: deletions
- 10. Tophat for Illumina on data 5 and data 2: insertions
- 9. Tophat for Illumina on data 5 and data 1: accepted\_hits
- 8. Tophat for Illumina on data 5 and data 1: splice junctions
- 7. Tophat for Illumina on data 5 and data 1: deletions
- 6. Tophat for Illumina on data 5 and data 1: insertions
- 5. hg19.refGene.gtf
- 4. brain\_2.Antassembler
- 3. brain\_1.Antassembler
- 2. UHR\_2.Antassembler
- 1. UHR\_1.Antassembler

This is how your history looks now, after you have sent four tophat jobs. Each tophat job produces 4 output files.

You don't have to wait for those jobs to finish to send Cuffdiff.



# TopHat output example

Info: Tophat v1.3.3

This file contains the mapped reads. We'll use on the next step.

7. Tophat for Illumina on data 5 and data 1: deletions

9. Tophat for Illumina on data 5 and data 1: accepted\_hits

8. Tophat for Illumina on data 5 and data 1: splice junctions

If you want to see the output from TopHat, import history *Hands-on 1 - Output* from Published Histories.

# Step 3: Run Cuffdiff to find genes differentially expressed

To get to this screen you have to click on **Add New Group** under Groups heading and also click on **Add New Replicate**.

Galaxy / WIBR

Cuffdiff (version 0.5.5) Cuffdiff v1.3.0

Transcripts: 5: hg19.refGene.gtf

Performance analysis: Performance analysis by replicates, cuffdiffcomp, or other source.

Group 1 (no spaces or commas): UHR

Group 2 (no spaces or commas): brain

Group name (no spaces or commas): data 5 and data 4: accepted hits

Group name (no spaces or commas): data 5 and data 2: accepted hits

Group name (no spaces or commas): data 5 and data 3: accepted hits

**After selecting all samples click Execute**

data 5 and data 4: accepted hits

data 5 and data 1: accepted hits

data 5 and data 2: accepted hits

data 5 and data 3: accepted hits

Perform pairwise normalization: Yes

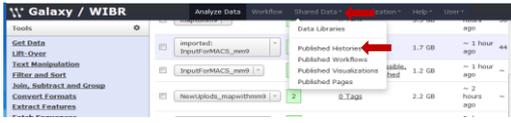
Perform Bias Correction: No

Set Parameters for Paired-end Reads? (not recommended): No



# Step 1: Get input Files

- Click on the **Shared Data** Tab and select **Published Histories**

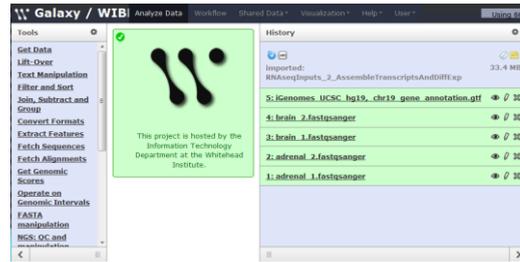


- Select **RNAseqInputs\_2\_AssembleTranscriptsAndDiffExp**



Click on **Import history** then click on **start using this history**

# These are our input files

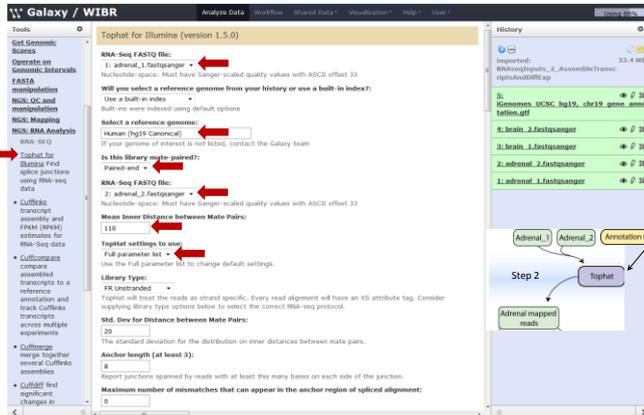


Hg19 Illumina gene annotation for chr19  
Brain paired end reads chr19 only  
Adrenal paired end reads chr19 only

Reads are 50 nt long **paired end reads** in fastqsanger format. Only reads mapping to chr19 are included. You will have to map each of the 2 sets of reads.

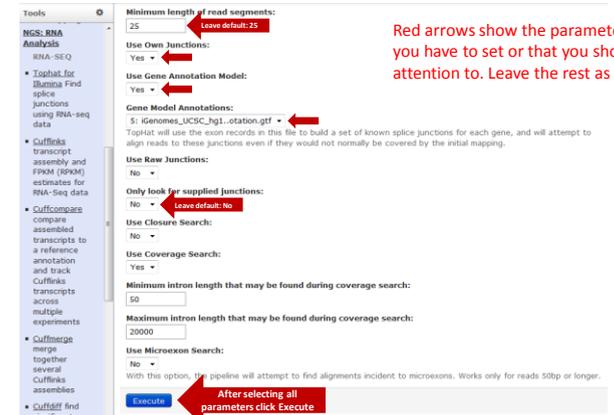
Annotation file for chr 19 is coming from Illumina igenomes <http://cufflinks.cbcb.umd.edu/igenomes.html>  
These input files were taken from the published RNA-seq Analysis Exercise <https://main.g2.bx.psu.edu/u/jeremy/p/galaxy-rna-seq-analysis-exercise>

# Step 2: Map reads with Tophat (v1.3.3)



Check next slide for the rest of the parameters. This screenshot doesn't show the full interface.

# Step 2: Map reads with Tophat cont.

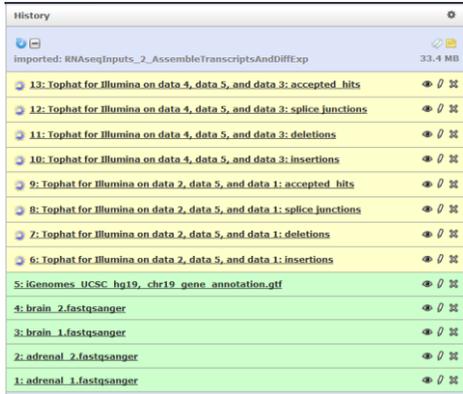


Red arrows show the parameters that you have to set or that you should pay attention to. Leave the rest as default.

After selecting all parameters click **Execute**

# Tophat output

This is how your history should look after sending the 2 tophat jobs



# Step 3: Assemble and quantify transcripts for each sample with Cufflinks

After selecting all parameters click Execute

You have to send 2 Cufflinks jobs. The 2 input files are the "accepted hits" files marked with green arrows.

19

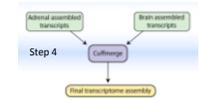
# Cufflinks Output

Assembled transcripts file. Will be the input file for Cuffmerge and Cuffcompare

Transcript expression: FPKM per transcript

Gene expression: FPKM per gene

# Step 4: Merge assemblies with Cuffmerge (v1.0.0)

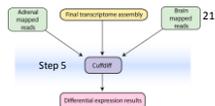


After selecting all parameters click Execute

The input files are the assembled transcripts files generated by Cufflinks (green arrows).

20

## Step 5: Find genes with Cuffdiff



## Step 5 output: Cuffdiff output Gene and transcript files

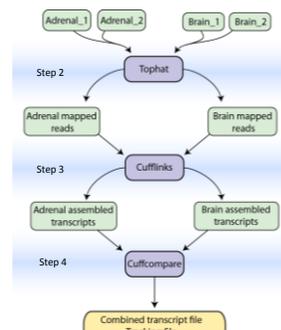
To access the outputs from Hands-on 2, import history *Hands-on 2 and 3\_ Output* from Published Histories. Note that because we are working with a small sample size none of the differences found are significant.

## Step 5 output: Cuffdiff output TSS, promoters and splicing files

Note that because we are working with a small sample size none of the differences found are significant.

## Hands-on 3 Assemble transcripts and compare assemblies

1. Get input files
2. Map reads with tophat
3. Assemble and quantify transcripts for each sample using Cufflinks
4. Compare assemblies with Cuffcompare



We have already done steps 1-3 in Hands-on 2. We only need to run Cuffcompare.

For simplicity the "annotation file" input has been removed "Annotation file" is an optional input

## Step 4: Compare assemblies with Cuffcompare

The screenshot shows the Galaxy web interface for the Cuffcompare v1.3.0 tool. The left-hand navigation menu has 'Cuffcompare' highlighted. The main configuration area is titled 'Cuffcompare (version 0.0.5) Cuffcompare v1.3.0'. It contains several sections: 'CTF file produced by Cufflinks:' with a dropdown menu; 'Additional CTF Input Files:' with a list of files and a 'Remove Additional CTF Input Files 1' button; 'Add new Additional CTF Input Files:' button; 'Use Reference Annotation:' with a 'Yes' radio button selected; 'Reference Annotation:' with a text input field containing 'Genome\_LUCSC\_Tg1\_stations.gtf'; 'Ignore reference transcripts that are not overlapped by any transcript in input files:' with a 'Yes' radio button selected; 'Use Sequence Data:' with a 'Yes' radio button selected; and 'Choose the source for the reference list:' with a dropdown menu set to 'Locally cached'. At the bottom, there is an 'Execute' button. A red callout box with the text 'After selecting all parameters click Execute' points to the Execute button. The right-hand panel shows a list of historical runs with their respective parameters and dates.

## Cuffcompare output files

- Cuffcompare gives a gtf file of combined transcripts that you can use in other analysis
- Cuffcompare also gives tracking, reformat and tmap files that point to transcripts that fully or partially match between samples
- For more detailed description of output files see [http://cufflinks.cbcb.umd.edu/manual.html#cuffcomp\\_output](http://cufflinks.cbcb.umd.edu/manual.html#cuffcomp_output)
- To access the outputs from Hands on 3, import history *Hands on 2 and 3 \_ Output* from Published Histories