

# Canadian Bioinformatics Workshops

[www.bioinformatics.ca](http://www.bioinformatics.ca)

This page is available in the following languages:

Afrikaans বাংলা Català Dansk Deutsch Ελληνικά English (GB) English (US) Esperanto  
 Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)  
 Euskara Suomi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu  
 Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски (latinica) Sotho svenska  
 中文 華語 (台灣) isiZulu



## Attribution-Share Alike 2.5 Canada

### You are free:



**to Share** — to copy, distribute and transmit the work



**to Remix** — to adapt the work



### Under the following conditions:



**Attribution.** You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



**Share Alike.** If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

[Disclaimer](#)

Your fair dealing and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full licence) available in the following languages:  
[English](#) [French](#)

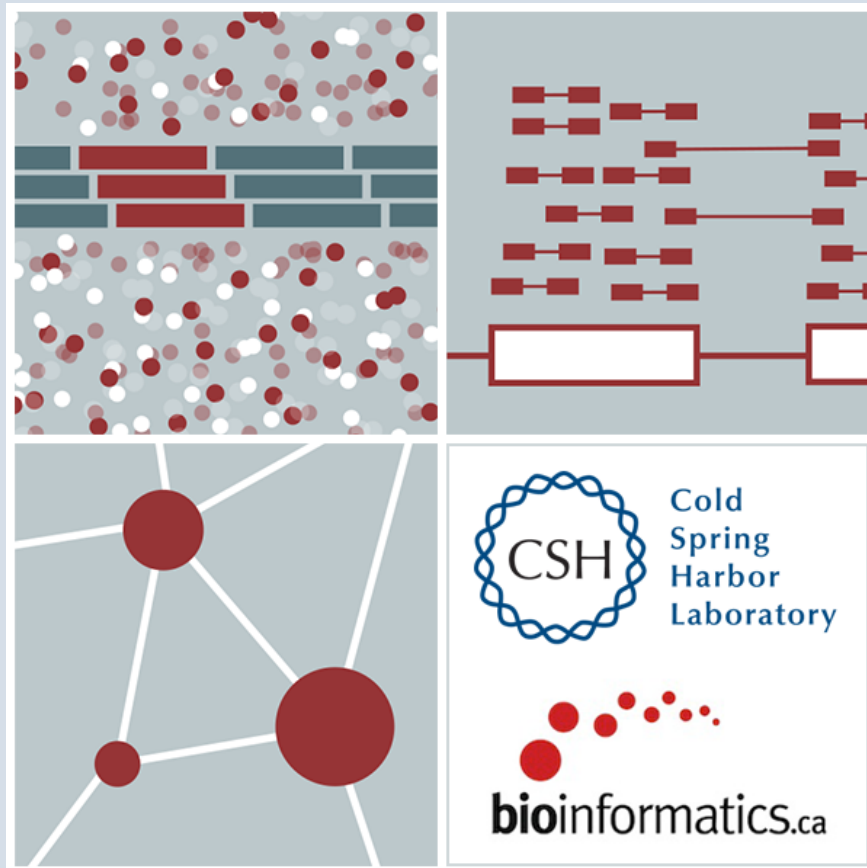
# RNA-Seq Module 4

## Isoform Discovery and Alternative Expression (tutorial)

Malachi Griffith, Obi Griffith, Fouad Yousif

High-Throughput Biology: From Sequence to Networks

March 20-26, 2017



# Learning Objectives of Tutorial

- Learn how to run StringTie in 'reference only', 'reference guided', and 'de novo' modes
- Learn how to use Cuffmerge to combine transcriptomes from multiple Cufflinks runs and compare assembled transcripts to known transcripts
- Learn how to perform differential splicing analysis with Cuffdiff
- Examine junctions counts and Cufflinks differential splicing files at the command line
- Visualize TopHat junction counts and Cufflinks assembled transcripts in IGV

# 5-i,ii. Running cufflinks in 'ref-guided' and 'de-novo' mode

- In Module 3 we ran cufflinks in 'ref-only' mode. This mode gives us an expression estimate for each known gene/transcript
- Now we want to be able to potentially identify novel genes, and novel isoforms of known genes
- To accomplish this we will re-run cufflinks in 'ref-guided' and 'de-novo' modes
  - In 'ref-guided' mode a known transcriptome will be used as a guide
  - In 'de-novo' mode no knowledge of the transcriptome will be used at all

# ‘-g’, ‘-G’ woe is me...

- tophat has a ‘-G’ option
  - Used to supply a transcriptome GTF file
  - This will be used to **assist the alignment** step by allowing alignment to both transcriptome and genome sequences
  - Coordinates from alignments to transcriptomes will be converted back to genome coordinates
  - Even though we supply a transcriptome, tophat will not be limited in anyway to known transcripts
- tophat also has a ‘-g’ option
  - Used to specify the maximum number of multiple mappings for a single read
- cufflinks has a ‘-G’ option
  - Used to supply a transcriptome GTF file
  - If specified, cufflinks will quantitate against reference transcript annotations
  - We call this the ‘ref-only’ analysis mode
- cufflinks also has a ‘-g’ option
  - Use to supply a transcriptome GTF file
  - Use reference transcript annotations to **guide assembly**
  - We call this ‘reference-guided’ analysis mode
- Running cufflinks with neither ‘-G’ or ‘-g’
  - We call this ‘de-novo’ analysis mode
- cuffdiff requires a GTF file but it is not specified with a ‘-G’ or ‘-g’ option, but rather is simply supplied as a file path when you run cuffdiff

# The tophat 'junctions.bed' file

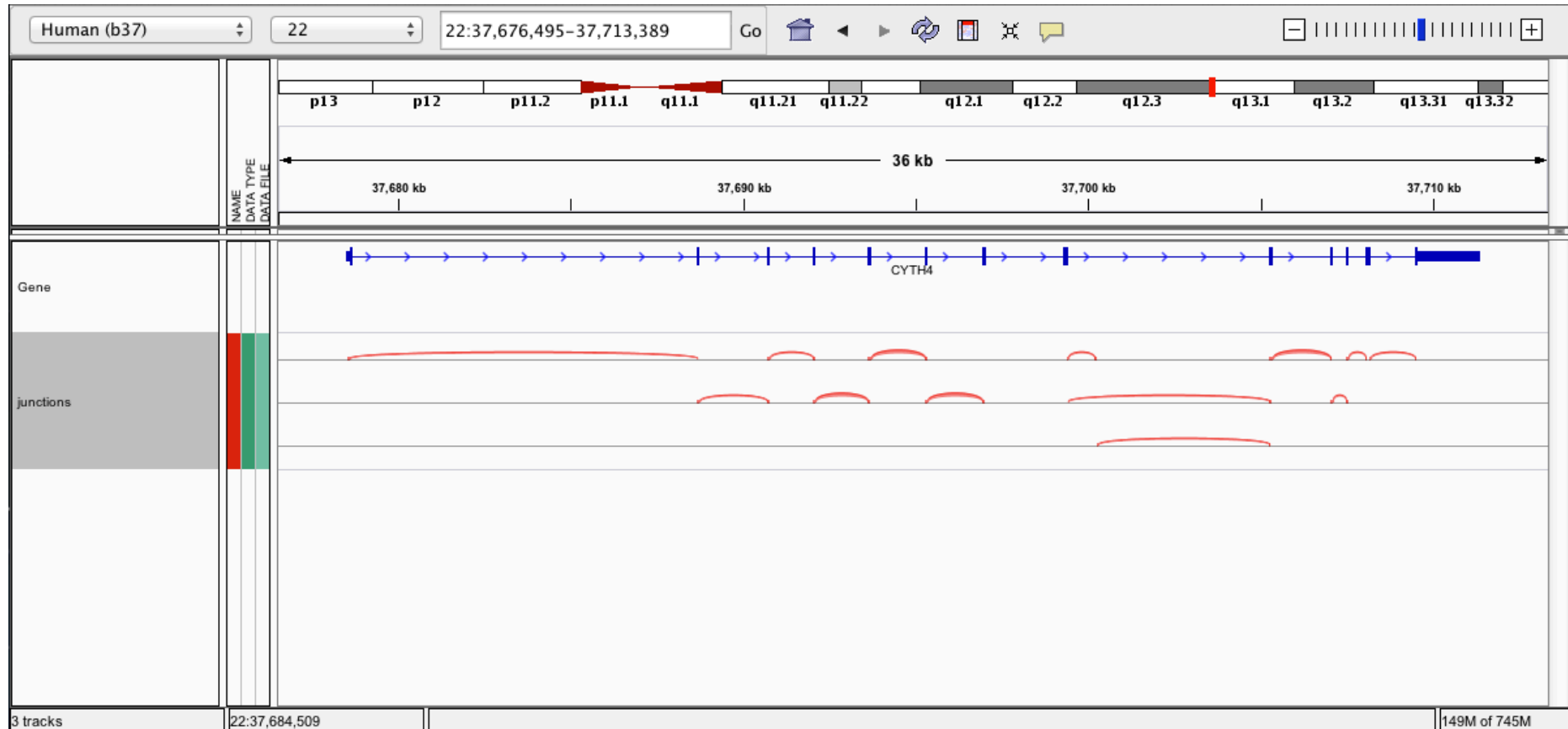
- After alignment, tophat creates a summary of all reads that support exon-exon junctions
  - e.g. exon1-exon2 has 5 reads
  - e.g. exon1-exon3 has 9 reads
- This file reports all of the unique exon-exon junctions observed and the read counts for each
  - In BED format

```
track name=junctions description="TopHat junctions"
22 17062079 17063415 JUNC000000001 3 - 17062079 17063415 255,0,0 2 98,19 0,1317
22 17092740 17095057 JUNC000000002 5 + 17092740 17095057 255,0,0 2 43,91 0,2226
22 17117940 17119543 JUNC000000003 6 + 17117940 17119543 255,0,0 2 40,75 0,1528
22 17152466 17156100 JUNC000000004 3 - 17152466 17156100 255,0,0 2 12,88 0,3546
22 17525819 17528242 JUNC000000005 1 + 17525819 17528242 255,0,0 2 71,29 0,2394
22 17528261 17538007 JUNC000000006 1 + 17528261 17538007 255,0,0 2 55,45 0,9701
22 17566071 17577976 JUNC000000007 10 + 17566071 17577976 255,0,0 2 48,25 0,11880
22 17577951 17578785 JUNC000000008 24 + 17577951 17578785 255,0,0 2 25,99 0,735
22 17578093 17578710 JUNC000000009 1 + 17578093 17578710 255,0,0 2 76,24 0,593
```



Junction read count

# Viewing the junctions.bed in IGV

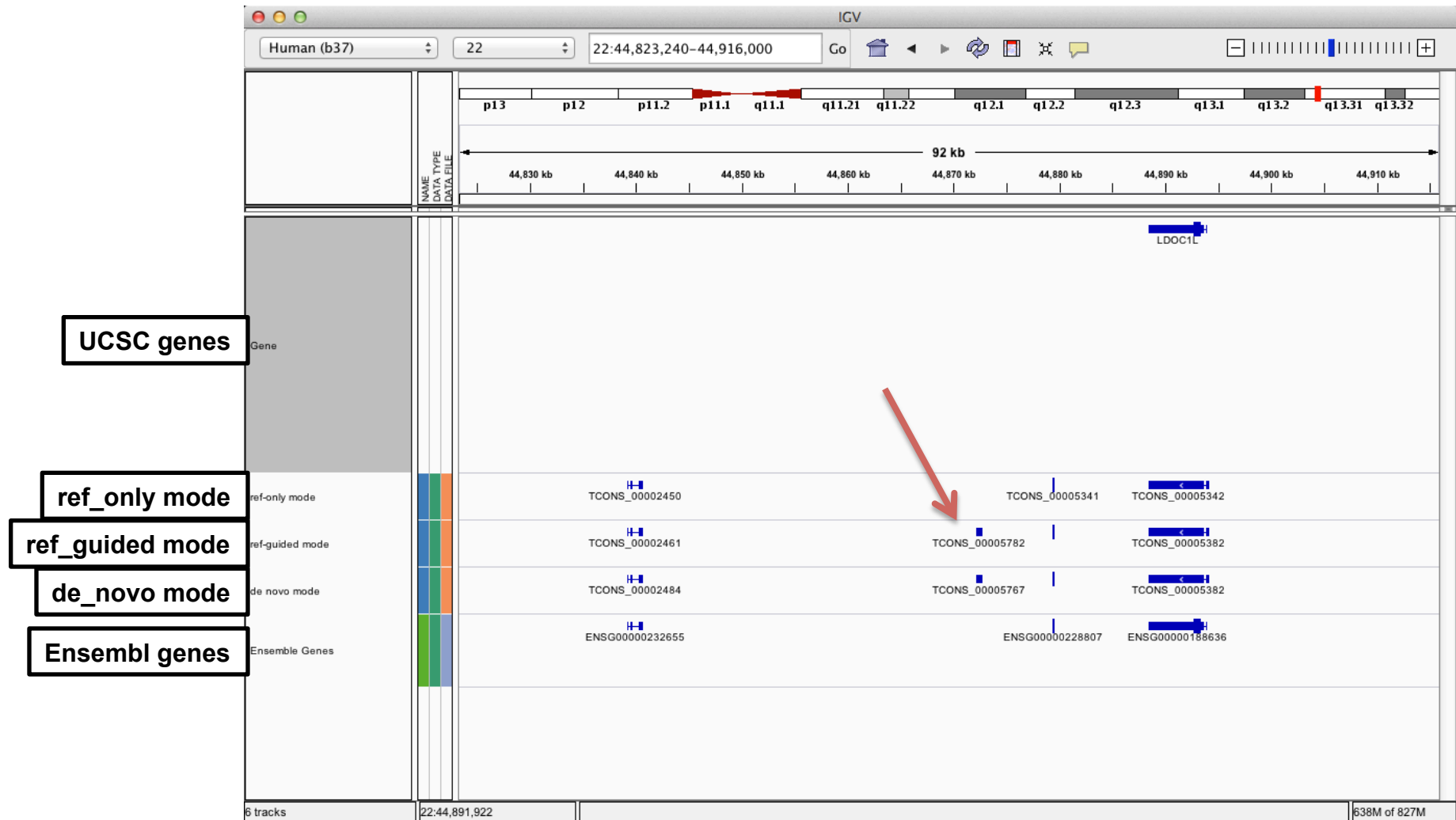




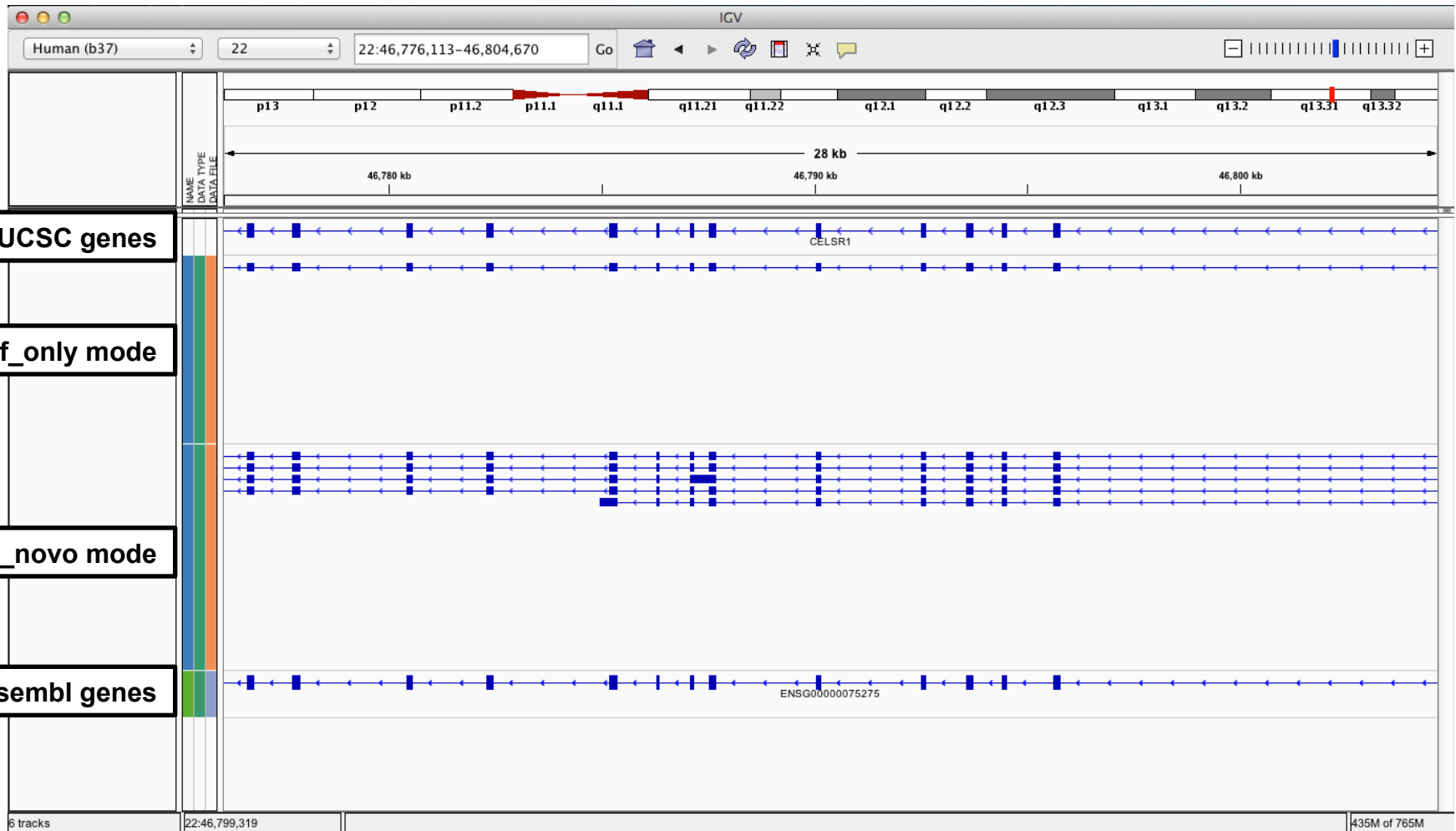
## 5-iii,iv. Cuffmerge

- <http://cufflinks.cbcb.umd.edu/manual.html#cuffmerge>
- Cuffmerge combines transcripts predicted from multiple RNA-seq data sets into one view of the transcriptome
  - Do this before running cuffdiff to compare between multiple conditions
- Cuffmerge can also simultaneously compare transcripts to the known transcripts GTF file from Ensembl, etc.
  - [http://cufflinks.cbcb.umd.edu/manual.html#class\\_codes](http://cufflinks.cbcb.umd.edu/manual.html#class_codes)

# 5-v. Comparison of merged GTFs from each cufflinks mode



# Comparison of merged GTFs from each cufflinks mode



# What if I return to my lab and can not get this to work on my own data?

- Refer to the materials provided with this course for clues
- Refer to the Nature Protocols tutorial (Trapnell et al. 2012)
  - In particular refer to the troubleshooting table (next slide)
- Search BioStars, SeqAnswers, and Google
  - <http://www.biostars.org/>
  - <http://www.seqanswers.com>
- If your question is not already answered on BioStars...
  - Ask it! Then follow up so that others that have the same problem in the future know whether this solution worked

# TopHat/Cufflinks/Cuffdiff troubleshooting table

**TABLE 2** | Troubleshooting table.

Step	Problem	Possible reason	Solution
1	TopHat cannot find Bowtie or the SAM tools	Bowtie and/or SAM tools binary executables are not in a directory listed in the PATH shell environment variable	Add the directories containing these executables to the PATH environment variable. See the man page of your UNIX shell for more details
2	Cufflinks crashes with a 'bad_alloc' error Cufflinks takes excessively long to finish	Machine is running out of memory trying to assemble highly expressed genes	Pass the <code>-max-bundle-frags</code> option to Cufflinks with a value of <code>&lt;1,000,000</code> (the default). Try 500,000 at first, and lower values if the error is still thrown
5	Cuffdiff crashes with a 'bad_alloc' error Cuffdiff takes excessively long to finish	Machine is running out of memory trying to quantify highly expressed genes	Pass the <code>-max-bundle-frags</code> option to Cuffdiff with a value of <code>&lt;1,000,000</code> (the default). Try 500,000 at first, and lower values if the error is still thrown
	Cuffdiff reports FPKM = 0 for all genes and transcripts	Chromosome names in GTF file do not match the names in the BAM alignment files	Use a GTF file and alignments that has matching chromosome names (e.g., the GTF included with an iGenome index)

We are on a Coffee Break &  
Networking Session