bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

Creative Commons

This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE) Euskara Suomeksi 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 srpski (latinica) Sotho svenska 中文 華語 (台灣) isiZulu

creative commons

Attribution-Share Alike 2.5 Canada

You are free:

to Share — to copy, distribute and transmit the work

to Remix — to adapt the work

Free Cultural Works APPROVED FOR

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

Learn how to distribute your work using this licence

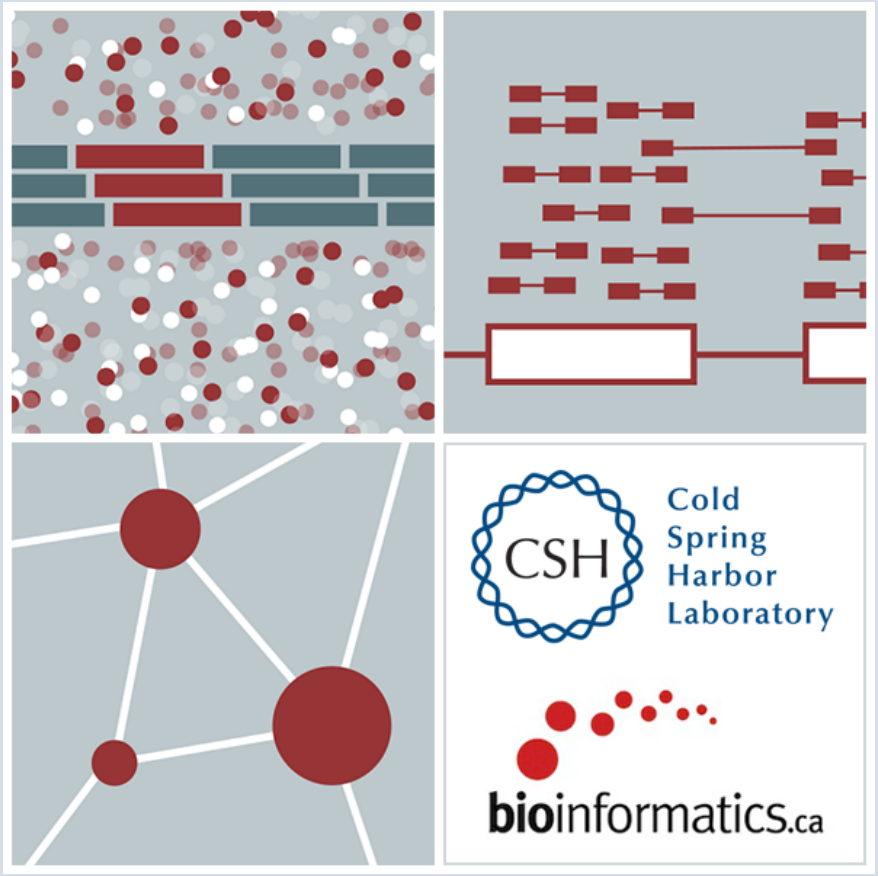
RNA-Seq Module 4
Isoform Discovery and Alternative Expression (tutorial)

Malachi Griffith, Obi Griffith, Fouad Yousif

Informatics for RNA-seq Analysis

July 10-12, 2017

bioinformatics.ca



THE GENOME INSTITUTE at Washington University

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 Ballgown
- Examine junctions counts with RegTools and StringTie alternative transcript files at the command line
- Visualize junction counts and StringTie assembled transcripts in IGV

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

- In Module 3 we ran StringTie 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

Options that govern use of existing transcript information

- During indexing of the genome with hisat2, transcript information is provided
  - A transcriptome GTF file is used to extract splice sites and exons
  - These are supplied during the index step to build a better index
  - These 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 transcriptome info, hisat2 will not be limited in to known transcripts or splice sites
- Stringtie '-G' option
  - Used to supply a transcriptome GTF file
  - If specified, uses the reference annotation file (in GTF or GFF3 format) to guide the assembly process. We call this the '**ref-guided**' analysis mode
- Stringtie '-e' option
  - Limits the processing of read alignments to only estimate and output the assembled transcripts matching the reference transcripts given with the -G option
  - We call this '**reference-only**' analysis mode
- Running StringTie with neither '-G' or '-e'
  - We call this '**de-novo**' analysis mode

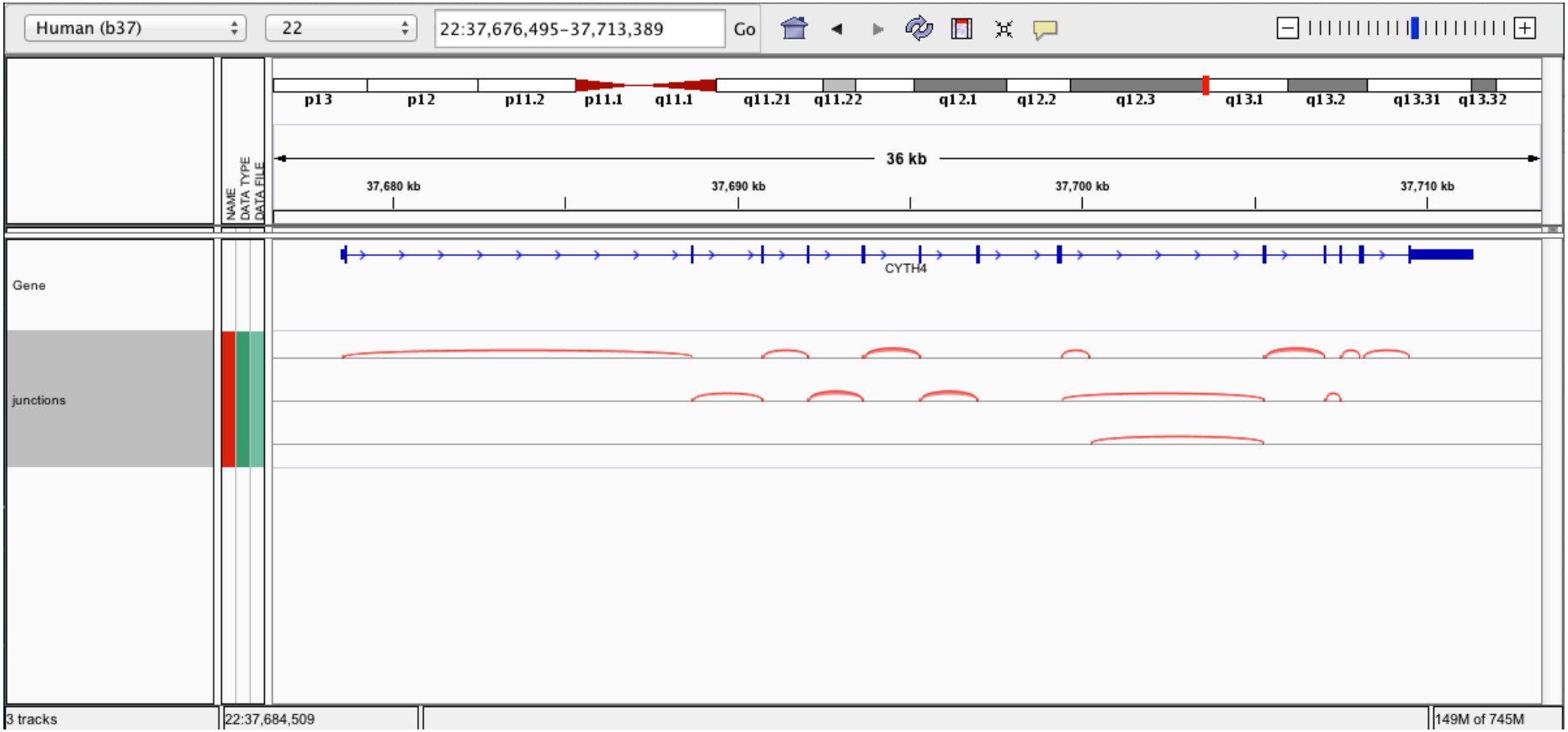
A ‘junctions.bed’ file

- After alignment, we can create 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        JUNC00000001    3       -       17062079        17063415        255,0,0 2       98,19   0,1317
22      17092740        17095057        JUNC00000002    5       +       17092740        17095057        255,0,0 2       43,91   0,2226
22      17117940        17119543        JUNC00000003    6       +       17117940        17119543        255,0,0 2       40,75   0,1528
22      17152466        17156100        JUNC00000004    3       -       17152466        17156100        255,0,0 2       12,88   0,3546
22      17525819        17528242        JUNC00000005    1       +       17525819        17528242        255,0,0 2       71,29   0,2394
22      17528261        17538007        JUNC00000006    1       +       17528261        17538007        255,0,0 2       55,45   0,9701
22      17566071        17577976        JUNC00000007    10      +       17566071        17577976        255,0,0 2       48,25   0,11880
22      17577951        17578785        JUNC00000008    24      +       17577951        17578785        255,0,0 2       25,99   0,735
22      17578093        17578710        JUNC00000009    1       +       17578093        17578710        255,0,0 2       76,24   0,593
```

Junction read count

Viewing the junctions.bed in IGV

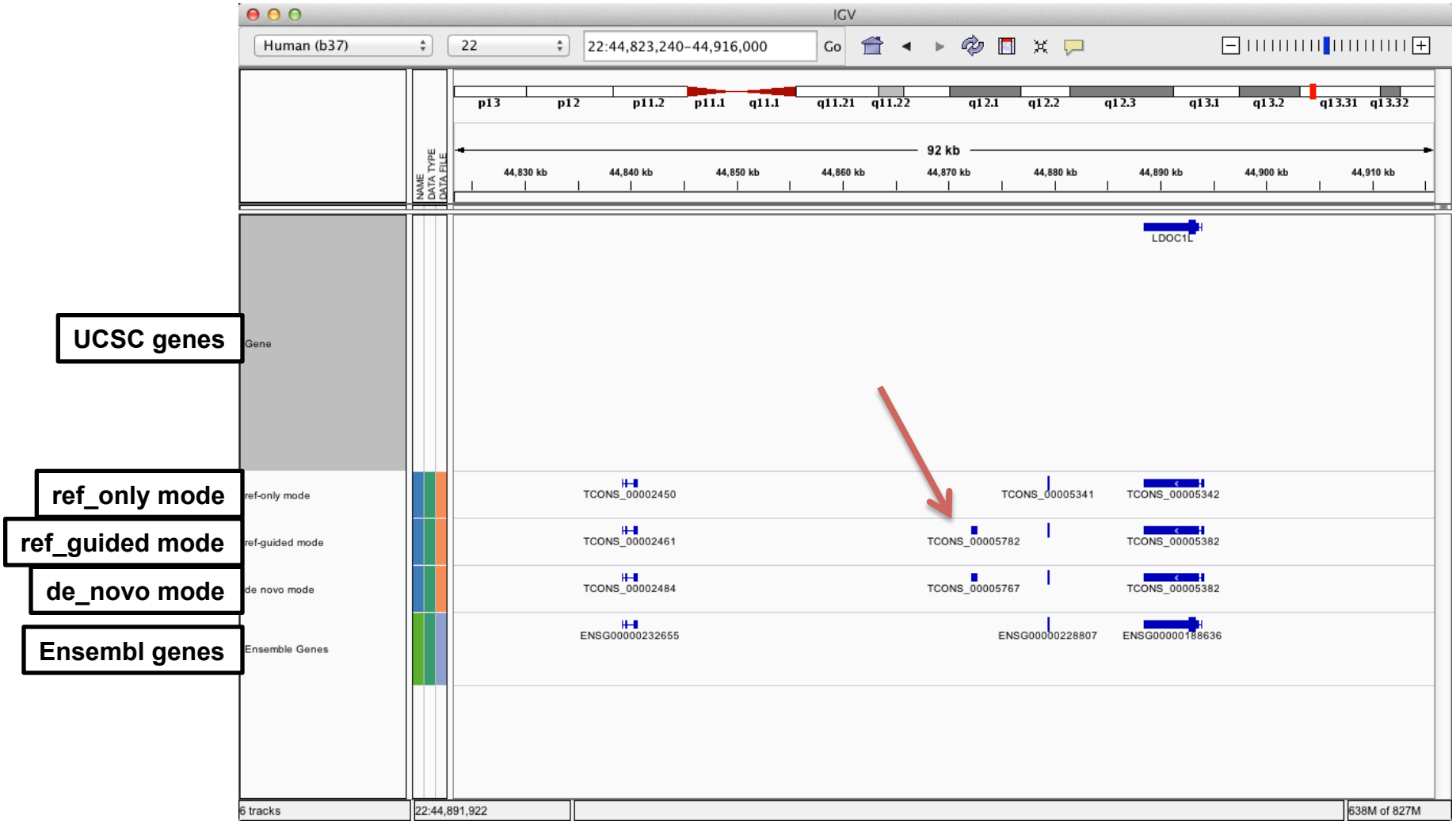


5-iii,iv. Cuffmerge

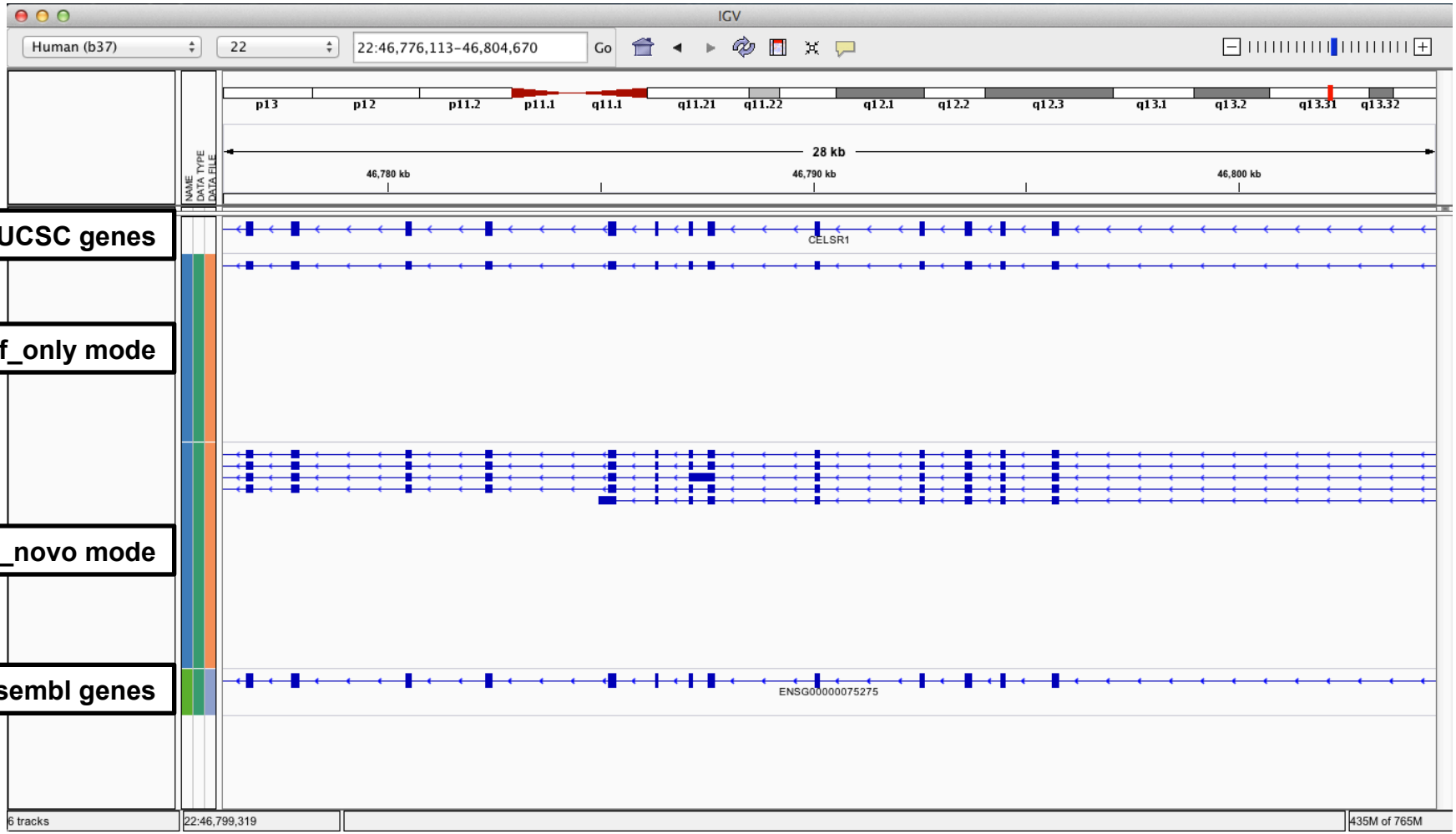
- http://cufflinks.cbcb.umd.edu/manual.html#cuffmerge

- 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

5-v. Comparison of merged GTFs from each StringTie mode



Comparison of merged GTFs from each StringTie mode



We are on a Coffee Break & Networking Session