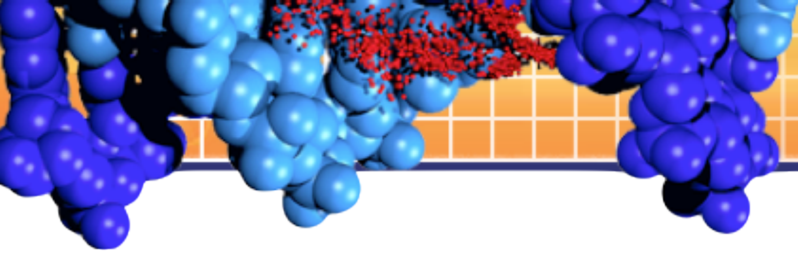


CENTER FOR BIOINFORMATICS
& COMPUTATIONAL BIOLOGY

Bioinformatics Short Course: RNA-Seq Data Analysis

Part IV: Transcriptome Assembly (Exercises)

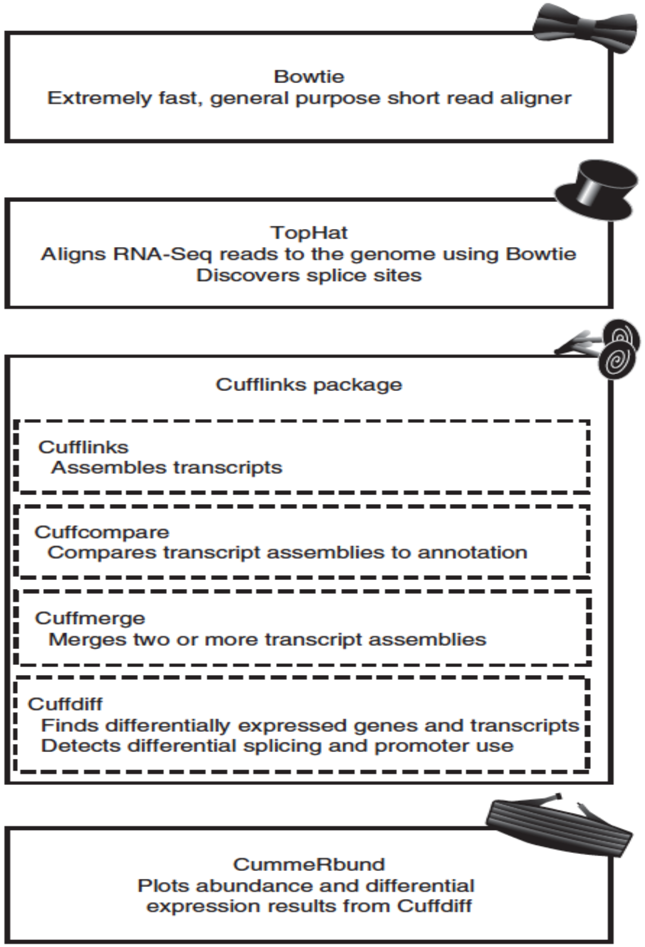
Chuming Chen, Ph.D.
University of Delaware

May 22-23, 2012

Objectives

- Learn RNA-Seq data analysis using open source software packages:
    - Tuxedo suite
        - Bowtie
        - TopHat
        - Cufflinks, Cuffcompare, Cuffmerge
    - SAMtools
    - IGV (Integrative Genomics Viewer)
- Gain hands-on experience in using these tools.

Software Components of Tuxedo Suite Tools



Bowtie
Extremely fast, general purpose short read aligner

TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites

Cufflinks package

Cufflinks
Assembles transcripts

Cuffcompare
Compares transcript assemblies to annotation

Cuffmerge
Merges two or more transcript assemblies

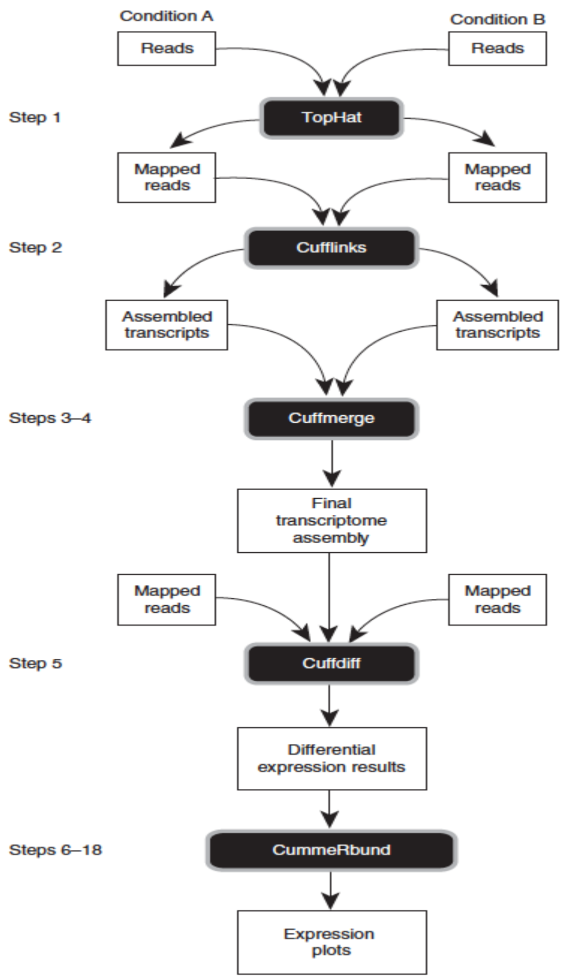
Cuffdiff
Finds differentially expressed genes and transcripts
Detects differential splicing and promoter use

CummeRbund
Plots abundance and differential
expression results from Cuffdiff

- **Bowtie** forms the algorithmic core of TopHat, which align reads to the reference genome.
- **TopHat**'s read alignments are assembled by **Cufflinks** and its associated utility program (**Cuffmerge**, **Cuffcompare**) can produce a transcriptome annotation of the genome.
- **Cuffdiff** quantifies this transcriptome across multiple conditions using the TopHat read alignments.
- **CummeRbund** explores and visualizes the differential expression data (Genes and Transcripts) generated by Cuffdiff.

(Trapnell et al., **Nat Protoc**. 2012 Mar 1;7(3):562-78)

Overview of Analysis Protocol

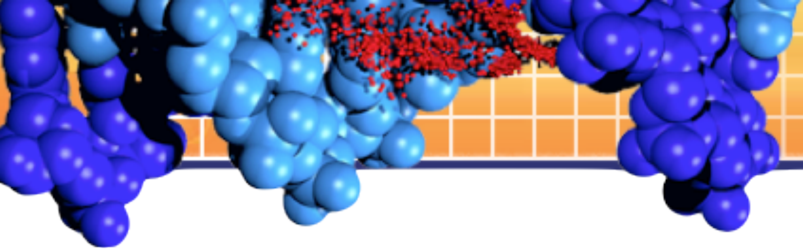


- Two condition experiment.
- Reads are first mapped to the genome with TopHat, biological replicate are mapped separately.
- Cufflinks creates one assembled transfrags file for each replicate.
- The assembled transfrags files are merged with the reference transcriptome annotation to form a unified annotation file.
- Cuffdiff quantifies the merged annotation file in each condition and generates expression data tables.
- These files are indexed and visualized with CummeRbund to facilitate the exploration of genes/transcripts identified by Cuffdiff as differentially expressed, spliced, or transcriptionally regulated genes.

(Trapnell et al., **Nat Protoc**. 2012 Mar 1;7(3):562-78)

Requirements

- Software (already installed on biohen cluster)
  - Bowtie (http://bowtie-bio.sourceforge.net/index.shtml/)
  - TopHat (http://tophat.cbcb.umd.edu/)
  - Cufflinks, Cuffcompare, Cuffmerge, Cuffdiff (http://cufflinks.cbcb.umd.edu/)
  - CummeRbund (http://compbio.mit.edu/cummeRbund/)
  - SAM tools (http://samtools.sourceforge.net/)
  - Integrative Genomics Viewer (http://www.broadinstitute.org/igv/home)
- Data
  - Reference
    - Ensembl 64 chicken chromosome 1 sequence in FASTA format (**gallus_chr1.fa**) and its annotations in GTF format (**gallus_chr1.gtf**).
    - http://useast.ensembl.org/info/data/ftp/index.html
  - RNA-Seq reads
    - Adipose Tissues of Fat/Lean Line Chicken. 2 Fat line and 2 Lean line samples were multiplexed in one Illumina HiSeq 2000 lane.
    - Randomly selected 100,000 paired-end reads that can be mapped to chicken chromosome 1 for each sample.
      - **FL1-1.trim.paired.fastq and FL1-2.trim.paired.fastq**
      - **FL2-1.trim.paired.fastq and FL2-2.trim.paired.fastq**
      - **LL1-1.trim.paired.fastq and LL1-2.trim.paired.fastq**
      - **LL2-1.trim.paired.fastq and LL2-2.trim.paired.fastq**



Setup Working Environment

- Commands to be executed from Linux shell are prefixed with a '$' (Don't include it in the command when you type in).
- Blue text follows the command is the output from the commands.

Login to biohen cluster through either a SSH client or command console:

```
$ ssh –X yourlogin@biohen.dbi.udel.edu
```

```
$ ls –l ~/rnaseq-shared
$ ls –l ~/rnaseq-work
```

If directory "~/rnaseq-shared" and "~/rnaseq-work" DO NOT exist, run the following commands , otherwise skip this step:

```
$ ln –s /net/biohen/shared/rna-seq-course ~/rnaseq-shared
$ mkdir ~/rnaseq-work
```

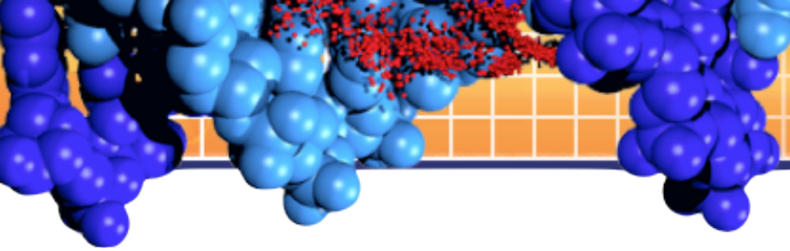
Copy data files:

```
$ cd ~/rnaseq-work
$ cp -r ~/rnaseq-shared/reference .
```

Exercise 1

Index Reference Genome

(Bowtie)



Bowtie

- An ultrafast, memory-efficient short read aligner.
- It uses an extremely economical data structure called the Burrows-Wheeler index to store the reference genome sequence and allows it to be searched rapidly at a rate of tens of millions reads per CPU hour.
- It makes a number of compromises to achieve its high speed:
  - If one or more exact matches exist for a read, it is guaranteed to find one.
  - If the best match is not exact match, then it is not guaranteed in all cases to find the highest quality alignment.
  - It may fail to align reads with multiple mismatches.
- Furthermore, Bowtie does not allow alignments between a read and the genome to contain large gaps; hence, it cannot align reads that span introns. TopHat was created to address this limitation.
- Web Site: http://bowtie-bio.sourceforge.net/index.shtml

Build Bowtie index

Check the “bowtie-build” command options:

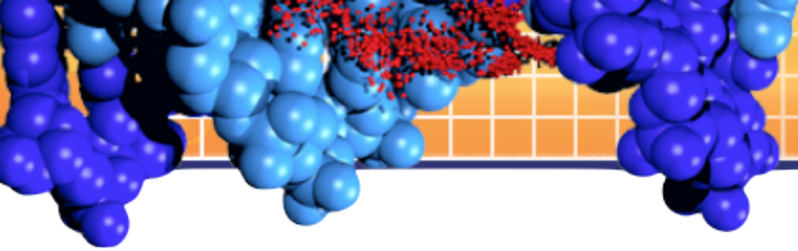
```
$ bowtie-build
```

Build an index for Chicken chromosome 1 using biohen cluster:

```
$ cd ~/rnaseq-work

$ mkdir index

$ cp reference/gallus_chr1.fa index/

$ cat ~/rnaseq-shared/pbs_scripts/bowtie_build.qs
#PBS -N BowtieBuild
#PBS -S /bin/bash
#PBS -V
#PBS -l ncpus=1,walltime=16:00:00,cput=10:00:00,mem=2000mb,nodes=1:ppn=1
#PBS -q rnaseq

cd $PBS_O_WORKDIR
bowtie-build index/gallus_chr1.fa index/gallus_chr1

$ qsub ~/rnaseq-shared/pbs_scripts/bowtie_build.qs
90249.biohen.dbi.local

$ qstat —a
biohen.dbi.local:
                                                                         Req'd  Req'd   Elap
Job ID               Username Queue    Jobname          SessID NDS   TSK Memory Time  S Time
-------------------- -------- -------- ---------------- ------ ----- --- ------ ----- - -----
90249.biohen.dbi     chenc    cbcb     BowtieBuild        8360     1   1 2000mb 600:0 R 00:01
```

Check Bowtie index

```
$ ls -tlr index/
total 412964
-rw-r--r-- 1 chenc cwu  48797843 May 17 11:34 gallus_chr1.4.ebwt
-rw-r--r-- 1 chenc cwu     89909 May 17 11:34 gallus_chr1.3.ebwt
-rw-r--r-- 1 chenc cwu  60083414 May 17 11:36 gallus_chr1.1.ebwt
-rw-r--r-- 1 chenc cwu  24398928 May 17 11:36 gallus_chr1.2.ebwt
-rw-r--r-- 1 chenc cwu  60083414 May 17 11:39 gallus_chr1.rev.1.ebwt
-rw-r--r-- 1 chenc cwu  24398928 May 17 11:39 gallus_chr1.rev.2.ebwt
-rw-r--r-- 1 chenc cwu 205013902 May 17 11:40 gallus_chr1.fa

$ bowtie
No index, query, or output file specified!
Usage:
  bowtie [options]* <ebwt> {-1 <m1> -2 <m2> | --12 <r> | <s>} [<hit>]
...
...

$ bowtie –c index/gallus_chr1 Aagggttt
0     +     chr1  117204729  AAGGGTTT   IIIIIIII   4261
# reads processed: 1
# reads with at least one reported alignment: 1 (100.00%)
# reads that failed to align: 0 (0.00%)
Reported 1 alignments to 1 output stream(s)
```

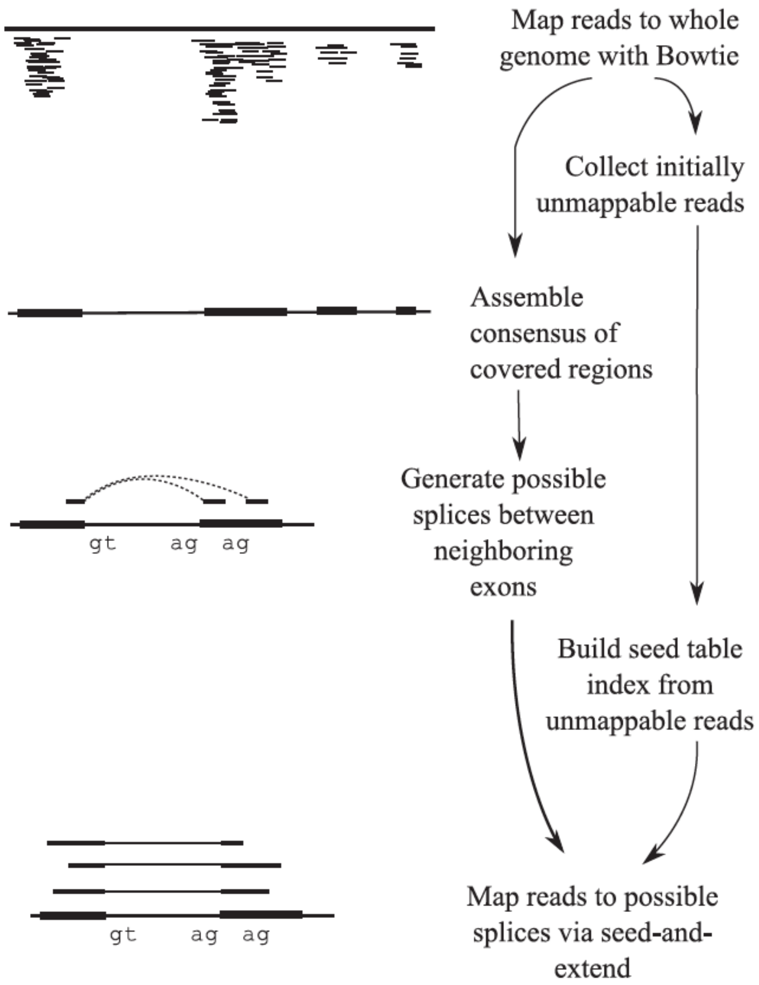


Exercise 2

Align Reads to the Reference Genome

(TopHat)

TopHat



Map reads to whole genome with Bowtie

Collect initially unmappable reads

Assemble consensus of covered regions

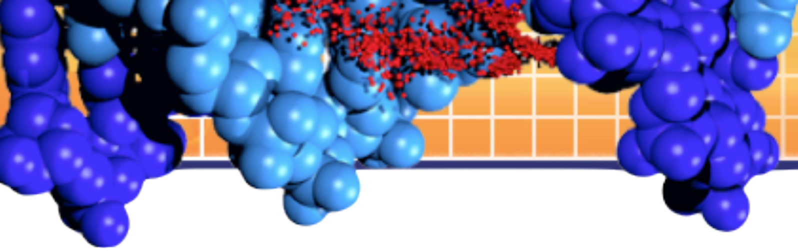
Generate possible splices between neighboring exons

Build seed table index from unmappable reads

Map reads to possible splices via seed-and-extend

- Use Bowtie as alignment engine.
- Break up reads Bowtie cannot align into segments then align them independently.
- When several of a read's segments aligned to the genome far apart, TopHat infers that the read spans a splice junction and estimate the splice site.
- By using the 'initially unmapped' reads, TopHat can build an index of splice sites in the transcriptome on the fly without a prior gene or splice site annotations.
- Web site: http://tophat.cbcb.umd.edu/

(Trapnell et al. **Bioinformatics**. 2009 May 1;25(9):1105-11)



Align the RNA-Seq Reads to the Genome

Create symbolic links to the RNA-Seq reads if they don't exist, otherwise skip this step:

```
$ ls -tlr *trim.paired.fastq
ls: cannot access *.trim.paired.fastq: No such file or directory

$ ln -s trimmed_sequences/* .

$ ls -tlr *trim.paired.fastq
lrwxrwxrwx 1 chenc cwu 41 May 17 14:00 FL1-1.trim.paired.fastq -> trimmed_sequences/FL1-1.trim.paired.fastq
lrwxrwxrwx 1 chenc cwu 41 May 17 14:00 FL1-2.trim.paired.fastq -> trimmed_sequences/FL1-2.trim.paired.fastq
lrwxrwxrwx 1 chenc cwu 41 May 17 14:00 FL2-1.trim.paired.fastq -> trimmed_sequences/FL2-1.trim.paired.fastq
lrwxrwxrwx 1 chenc cwu 41 May 17 14:00 FL2-2.trim.paired.fastq -> trimmed_sequences/FL2-2.trim.paired.fastq
lrwxrwxrwx 1 chenc cwu 41 May 17 14:00 LL1-1.trim.paired.fastq -> trimmed_sequences/LL1-1.trim.paired.fastq
lrwxrwxrwx 1 chenc cwu 41 May 17 14:00 LL1-2.trim.paired.fastq -> trimmed_sequences/LL1-2.trim.paired.fastq
lrwxrwxrwx 1 chenc cwu 41 May 17 14:00 LL2-1.trim.paired.fastq -> trimmed_sequences/LL2-1.trim.paired.fastq
lrwxrwxrwx 1 chenc cwu 41 May 17 14:00 LL2-2.trim.paired.fastq -> trimmed_sequences/LL2-2.trim.paired.fastq
```

Check the "tophat" command options:

```
$ tophat
tophat:
TopHat maps short sequences from spliced transcripts to whole genomes.

Usage:
    tophat [options] <bowtie_index> <reads1[,reads2,...]> [reads1[,reads2,...]] \
                                    [quals1,[quals2,...]] [quals1[,quals2,...]]
...
```

Align the Reads of Sample FL1

```
$ cat ~/rnaseq-shared/pbs_scripts/tophat_FL1.qs
#PBS -N TopHatFL1
#PBS -S /bin/bash
#PBS -V
#PBS -l ncpus=1,walltime=16:00:00,cput=10:00:00,mem=2000mb,nodes=1:ppn=4
#PBS –q rnaseq

cd $PBS_O_WORKDIR
tophat -p 4 -g 1 -G reference/gallus_chr1.gtf -r 300 -o tophat_out_FL1 index/gallus_chr1 FL1-1.trim.paired.fastq FL1-2.trim.paired.fastq

$ qsub ~/rnaseq-shared/pbs_scripts/tophat_FL1.qs
90253.biohen.dbi.loca

$ qstat -a
biohen.dbi.local:
                                                                    Req'd  Req'd   Elap
Job ID               Username Queue    Jobname          SessID NDS   TSK Memory Time  S Time
-------------------- -------- -------- ---------------- ------ ----- --- ------ ----- - -----
90253.biohen.dbi     chenc    cbcb     TopHatFL1          8586     1   4 2000mb 600:0 R 00:01

$ ls -tlr tophat_out_FL1/
total 15656
-rw-r--r-- 1 chenc cwu       65 May 17 11:53 left_kept_reads.info
-rw-r--r-- 1 chenc cwu       65 May 17 11:53 right_kept_reads.info
drwxr-xr-x 2 chenc cwu     4096 May 17 11:54 logs
-rw-r--r-- 1 chenc cwu     7638 May 17 11:55 insertions.bed
-rw-r--r-- 1 chenc cwu     7577 May 17 11:55 deletions.bed
-rw-r--r-- 1 chenc cwu   624585 May 17 11:55 junctions.bed
-rw-r--r-- 1 chenc cwu 15375917 May 17 11:55 accepted_hits.bam
```

Align the Reads of Sample FL2, LL1 and LL2

```
$ qsub ~/rnaseq-shared/pbs_scripts/tophat_FL2.qs
90255.biohen.dbi.local
$ qsub ~/rnaseq-shared/pbs_scripts/tophat_LL1.qs
90256.biohen.dbi.local
$ qsub ~/rnaseq-shared/pbs_scripts/tophat_LL2.qs
90257.biohen.dbi.local

$ qstat -a

biohen.dbi.local:
                                                                          Req'd  Req'd    Elap
Job ID               Username Queue    Jobname          SessID NDS   TSK Memory Time  S Time
-------------------- -------- -------- ---------------- ------ ----- --- ------ ----- - -----
90255.biohen.dbi     chenc    cbcb     TopHatFL2          8848     1   4 2000mb 600:0 C 00:03
90256.biohen.dbi     chenc    cbcb     TopHatLL1          8877     1   4 2000mb 600:0 C 00:03
90257.biohen.dbi     chenc    cbcb     TopHatLL2         62619     1   4 2000mb 600:0 C 00:03
```

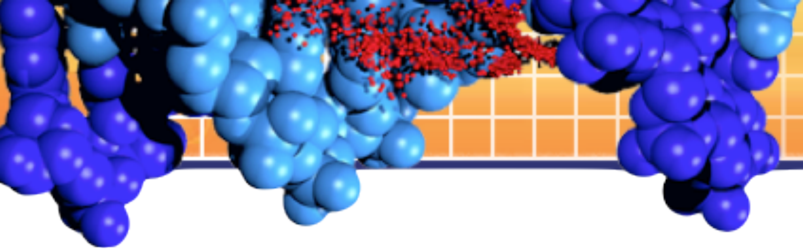
Check Alignment Outputs

```
$ ls -tlr tophat_out_*
tophat_out_FL1:
total 15656
-rw-r--r-- 1 chenc cwu       65 May 17 11:53 left_kept_reads.info
-rw-r--r-- 1 chenc cwu       65 May 17 11:53 right_kept_reads.info
drwxr-xr-x 2 chenc cwu     4096 May 17 11:54 logs
-rw-r--r-- 1 chenc cwu     7638 May 17 11:55 insertions.bed
-rw-r--r-- 1 chenc cwu     7577 May 17 11:55 deletions.bed
-rw-r--r-- 1 chenc cwu   624585 May 17 11:55 junctions.bed
-rw-r--r-- 1 chenc cwu 15375917 May 17 11:55 accepted_hits.bam

tophat_out_FL2:
total 15528
-rw-r--r-- 1 chenc cwu       65 May 17 12:00 left_kept_reads.info
-rw-r--r-- 1 chenc cwu       65 May 17 12:00 right_kept_reads.info
drwxr-xr-x 2 chenc cwu     4096 May 17 12:01 logs
-rw-r--r-- 1 chenc cwu     8734 May 17 12:01 insertions.bed
-rw-r--r-- 1 chenc cwu     7486 May 17 12:01 deletions.bed
-rw-r--r-- 1 chenc cwu   609676 May 17 12:01 junctions.bed
-rw-r--r-- 1 chenc cwu 15256657 May 17 12:02 accepted_hits.bam

tophat_out_LL2:
total 15488
-rw-r--r-- 1 chenc cwu       65 May 17 12:00 left_kept_reads.info
-rw-r--r-- 1 chenc cwu       65 May 17 12:00 right_kept_reads.info
drwxr-xr-x 2 chenc cwu     4096 May 17 12:01 logs
-rw-r--r-- 1 chenc cwu     8849 May 17 12:02 insertions.bed
-rw-r--r-- 1 chenc cwu     6557 May 17 12:02 deletions.bed
-rw-r--r-- 1 chenc cwu   603884 May 17 12:02 junctions.bed
-rw-r--r-- 1 chenc cwu 15218475 May 17 12:02 accepted_hits.bam

tophat_out_LL1:
total 15812
-rw-r--r-- 1 chenc cwu       65 May 17 12:00 left_kept_reads.info
-rw-r--r-- 1 chenc cwu       65 May 17 12:00 right_kept_reads.info
drwxr-xr-x 2 chenc cwu     4096 May 17 12:02 logs
-rw-r--r-- 1 chenc cwu     7648 May 17 12:02 insertions.bed
-rw-r--r-- 1 chenc cwu     6713 May 17 12:02 deletions.bed
-rw-r--r-- 1 chenc cwu   621429 May 17 12:02 junctions.bed
-rw-r--r-- 1 chenc cwu 15536842 May 17 12:02 accepted_hits.bam
```

Check Mapping Status for Each Alignment

```
$ samtools flagstat tophat_out_FL1/accepted_hits.bam
168306 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
168306 + 0 mapped (100.00%:-nan%)
168306 + 0 paired in sequencing
84906 + 0 read1
83400 + 0 read2
153700 + 0 properly paired (91.32%:-nan%)
158160 + 0 with itself and mate mapped
10146 + 0 singletons (6.03%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)

$ samtools flagstat tophat_out_FL2/accepted_hits.bam
168562 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
168562 + 0 mapped (100.00%:-nan%)
168562 + 0 paired in sequencing
85081 + 0 read1
83481 + 0 read2
154460 + 0 properly paired (91.63%:-nan%)
158288 + 0 with itself and mate mapped
10274 + 0 singletons (6.10%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

```
$ samtools flagstat tophat_out_LL1/accepted_hits.bam
170172 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
170172 + 0 mapped (100.00%:-nan%)
170172 + 0 paired in sequencing
86030 + 0 read1
84142 + 0 read2
157496 + 0 properly paired (92.55%:-nan%)
160390 + 0 with itself and mate mapped
9782 + 0 singletons (5.75%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)

$ samtools flagstat tophat_out_LL2/accepted_hits.bam
168022 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
168022 + 0 mapped (100.00%:-nan%)
168022 + 0 paired in sequencing
84757 + 0 read1
83265 + 0 read2
154414 + 0 properly paired (91.90%:-nan%)
157656 + 0 with itself and mate mapped
10366 + 0 singletons (6.17%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

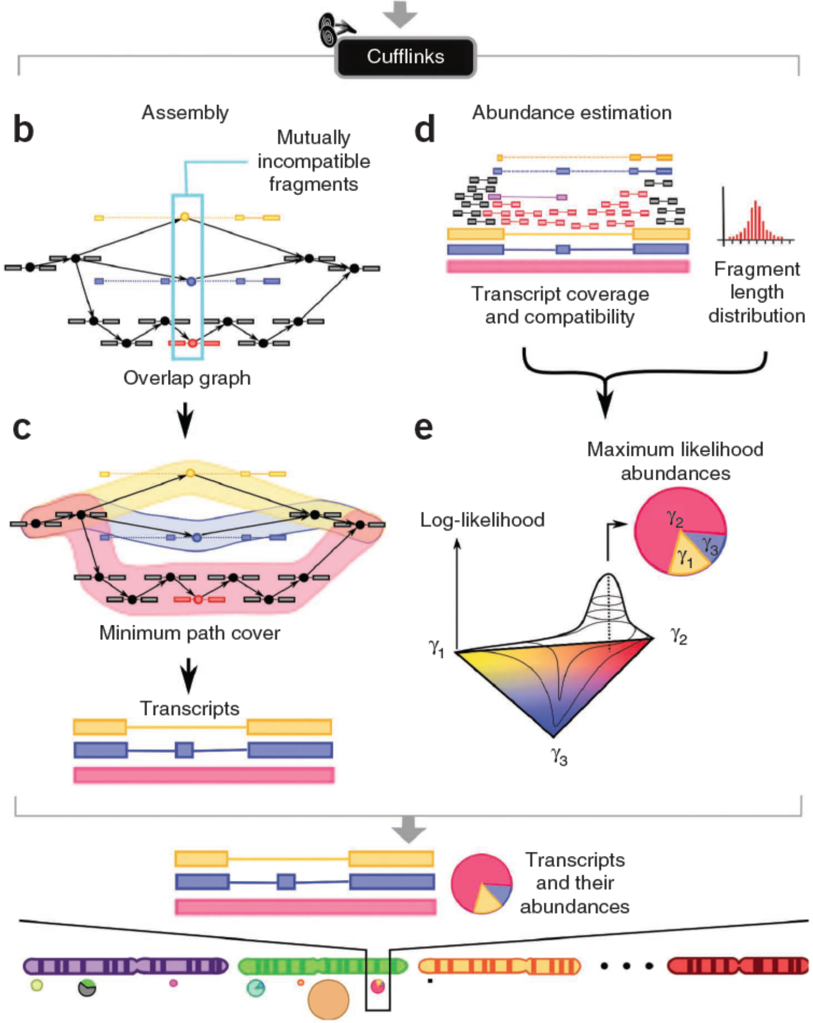


Exercise 3

Assemble Transcriptome

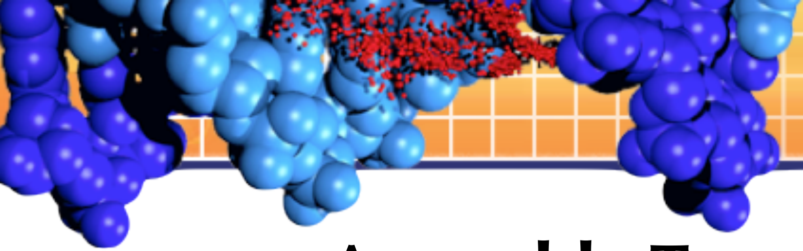
(Cufflinks, Cuffcompare, Cuffmerge)

Cufflinks


Cufflinks
b
Assembly
Mutually incompatible fragments
Overlap graph
c
Minimum path cover
Transcripts
d
Abundance estimation
Transcript coverage and compatibility
Fragment length distribution
e
Maximum likelihood abundances
Log-likelihood
γ_1
γ_2
γ_3
Transcripts and their abundances

- Assembles individual transcripts from RNA-Seq reads that have been aligned to the genome.
- Reports as few full-length transcript fragments or 'transfrags' as are needed to 'explain' all the splicing events in the input data.
- Quantifies the expression level of each transfrag in the sample using a rigorous statistical model of RNA-Seq to filter out background or artifactual transfrags such as immature primary transcripts.
- Quantifies transcript abundance using a reference annotation.
- Web site: http://cufflinks.cbcb.umd.edu/

(Trapnell et al. **Nat Biotechnol**. 2010 May;28(5):511-5)



Assemble Transcriptome of Sample FL1

Check the “cufflinks” command options:

```
$ cufflinks
cufflinks v1.3.0
linked against Boost version 104000
-----------------------------
Usage:   cufflinks [options] <hits.sam>
General Options:
  -o/--output-dir              write all output files to this directory              [ default:     ./ ]
  -p/--num-threads             number of threads used during analysis                [ default:      1 ]
  --seed                       value of random number generator seed                 [ default:      0 ]
  -G/--GTF                     quantitate against reference transcript annotations
  -g/--GTF-guide               use reference transcript annotation to guide
...
```

Assemble transcriptome for sample FL1 using biohen cluster:

```
$ cat ~/rnaseq-shared/pbs_scripts/cufflinks_FL1.qs
#PBS -N CufflinksFL1
#PBS -S /bin/bash
#PBS -V
#PBS -l ncpus=1,walltime=16:00:00,cput=10:00:00,mem=2000mb,nodes=1:ppn=4
#PBS –q rnaseq

cd $PBS_O_WORKDIR
cufflinks -p 4 -g reference/gallus_chr1.gtf -o cufflinks_out_FL1 tophat_out_FL1/accepted_hits.bam

$ qsub ~/rnaseq-shared/pbs_scripts/cufflinks_FL1.qs
90258.biohen.dbi.local

$ ls -tlr cufflinks_out_FL1/
total 9900
-rw-r--r-- 1 chenc cwu       0 May 17 12:17 skipped.gtf
-rw-r--r-- 1 chenc cwu 9520019 May 17 12:17 transcripts.gtf
-rw-r--r-- 1 chenc cwu  380896 May 17 12:17 isoforms.fpkm_tracking
-rw-r--r-- 1 chenc cwu  232907 May 17 12:17 genes.fpkm_tracking
```

Assemble Transcriptomes of Sample FL2, LL1 and LL2

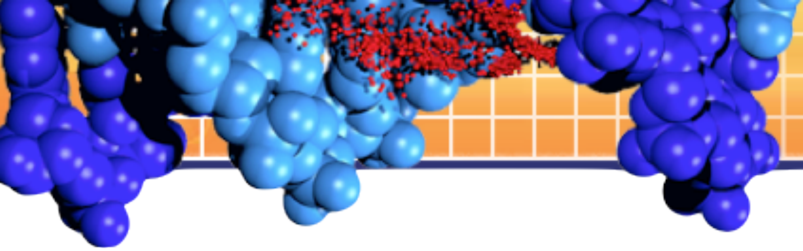
```
$ qsub ~/rnaseq-shared/pbs_scripts/cufflinks_FL2.qs
90259.biohen.dbi.local
$ qsub ~/rnaseq-shared/pbs_scripts/cufflinks_LL1.qs
90260.biohen.dbi.local
$ qsub ~/rnaseq-shared/pbs_scripts/cufflinks_LL2.qs
90261.biohen.dbi.local

$ qstat -a

biohen.dbi.local:
                                                                         Req'd  Req'd   Elap
Job ID               Username Queue    Jobname          SessID NDS   TSK Memory Time  S Time
-------------------- -------- -------- ---------------- ------ ----- --- ------ ----- - -----
90259.biohen.dbi     chenc    cbcb     CufflinksFL2      15430     1   4 2000mb 600:0 C 00:00
90260.biohen.dbi     chenc    cbcb     CufflinksLL1      15816     1   4 2000mb 600:0 C 00:00
90261.biohen.dbi     chenc    cbcb     CufflinksLL2      15826     1   4 2000mb 600:0 C 00:00
```

Check Assembly Outputs

```
$ ls -ltr cufflinks_out_*
cufflinks_out_FL1:
total 9900
-rw-r--r-- 1 chenc cwu       0 May 17 12:17 skipped.gtf
-rw-r--r-- 1 chenc cwu 9520019 May 17 12:17 transcripts.gtf
-rw-r--r-- 1 chenc cwu  380896 May 17 12:17 isoforms.fpkm_tracking
-rw-r--r-- 1 chenc cwu  232907 May 17 12:17 genes.fpkm_tracking
cufflinks_out_FL2:
total 9928
-rw-r--r-- 1 chenc cwu       0 May 17 12:25 skipped.gtf
-rw-r--r-- 1 chenc cwu 9549020 May 17 12:25 transcripts.gtf
-rw-r--r-- 1 chenc cwu  380683 May 17 12:25 isoforms.fpkm_tracking
-rw-r--r-- 1 chenc cwu  231673 May 17 12:25 genes.fpkm_tracking
cufflinks_out_LL1:
total 10048
-rw-r--r-- 1 chenc cwu       0 May 17 12:25 skipped.gtf
-rw-r--r-- 1 chenc cwu 9663479 May 17 12:25 transcripts.gtf
-rw-r--r-- 1 chenc cwu  383853 May 17 12:25 isoforms.fpkm_tracking
-rw-r--r-- 1 chenc cwu  234230 May 17 12:25 genes.fpkm_tracking
cufflinks_out_LL2:
total 9916
-rw-r--r-- 1 chenc cwu       0 May 17 12:25 skipped.gtf
-rw-r--r-- 1 chenc cwu 9536165 May 17 12:25 transcripts.gtf
-rw-r--r-- 1 chenc cwu  378921 May 17 12:25 isoforms.fpkm_tracking
-rw-r--r-- 1 chenc cwu  231467 May 17 12:25 genes.fpkm_tracking
```

Cuffcompare

- In addition to differential expression analysis, people are often interested in discovering new genes and transcripts.
- Gaps in sequencing coverage will cause breaks in transcript reconstruction and make it difficult to distinguish full-length novel transcripts from partial fragments.
- Cuffcompare can compare the Cufflinks assemblies to reference annotation files and help sort out new genes from known ones.
- Web site: http://cufflinks.cbcb.umd.edu/manual.html#cuffcompare



Compare Transriptome Assemblies to the Reference

Create a file “gtf_out_list.txt” that list all of the GTF files created by Cufflinks.

```
$ find . -name transcripts.gtf > gtf_out_list.txt

$ cat gtf_out_list.txt
./cufflinks_out_FL1/transcripts.gtf
./cufflinks_out_FL2/transcripts.gtf
./cufflinks_out_LL1/transcripts.gtf
./cufflinks_out_LL2/transcripts.gtf
```

Compare each assembly GTF in the list to the reference annotation file “reference/gallus_chr1.gtf”

```
$ cat ~/rnaseq-shared/pbs_scripts/cuffcompare.qs
#PBS -N Cuffcompare1
#PBS -S /bin/bash
#PBS -V
#PBS -l ncpus=1,walltime=16:00:00,cput=10:00:00,mem=2000mb,nodes=1:ppn=1
#PBS -q rnaseq

cd $PBS_O_WORKDIR
cuffcompare -i gtf_out_list.txt -r reference/gallus_chr1.gtf

$ qsub ~/rnaseq-shared/pbs_scripts/cuffcompare.qs
90293.biohen.dbi.local
```

Print Summary Reports

```
$ ls -tlr cufflinks_out_*/*map
-rw-r--r-- 1 chenc cwu 481209 May 17 14:35 cufflinks_out_FL1/cuffcmp.transcripts.gtf.tmap
-rw-r--r-- 1 chenc cwu 196073 May 17 14:35 cufflinks_out_FL1/cuffcmp.transcripts.gtf.refmap
-rw-r--r-- 1 chenc cwu 479849 May 17 14:35 cufflinks_out_FL2/cuffcmp.transcripts.gtf.tmap
-rw-r--r-- 1 chenc cwu   6144 May 17 14:35 cufflinks_out_FL2/cuffcmp.transcripts.gtf.refmap
-rw-r--r-- 1 chenc cwu 482416 May 17 14:35 cufflinks_out_LL1/cuffcmp.transcripts.gtf.tmap
-rw-r--r-- 1 chenc cwu   4466 May 17 14:35 cufflinks_out_LL1/cuffcmp.transcripts.gtf.refmap
-rw-r--r-- 1 chenc cwu 476989 May 17 14:35 cufflinks_out_LL2/cuffcmp.transcripts.gtf.tmap
-rw-r--r-- 1 chenc cwu   3880 May 17 14:35 cufflinks_out_LL2/cuffcmp.transcripts.gtf.refmap
```

Prints a simple table for each assembly that lists how many transcripts in each assembly are complete matches to the know transcripts, how many are partial matches etc.

```
$ find . -name *.tmap | while read file; do echo $file;  awk 'NR > 1 { s[$3]++ } END { for (j in s) { print
j, s[j] }} ' $file; done
```

Cuffcompare Summary Reports

```
./cufflinks_out_FL1/cuffcmp.transcripts.gtf.tmap
u 116
i 20
j 471
x 1
o 6
c 42
p 87
= 2973
e 25
s 2
./cufflinks_out_FL2/cuffcmp.transcripts.gtf.tmap
u 110
i 21
j 479
x 1
o 6
c 42
p 76
= 2971
e 24
s 1
./cufflinks_out_LL1/cuffcmp.transcripts.gtf.tmap
u 129
i 23
j 481
x 1
o 6
c 43
p 89
= 2963
e 24
s 1
./cufflinks_out_LL2/cuffcmp.transcripts.gtf.tmap
u 111
i 17
j 468
x 1
o 9
c 42
p 71
= 2968
e 21
```

Code	Description
=	Complete match of intron chain
c	Contained
j	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript
e	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment
i	A transfrag falling entirely within a reference transcript
o	Generic exonic overlap with a reference transcript
P	Possible polymerase run-on fragment (within 2Kbases of a reference transcript)
r	Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case
u	Unknown, intergenic transcript
x	Exonic overlap with reference on the opposite strand
s	An intron of the transfrag overlaps a reference intro on the opposite strand (likely due to read mapping errors)
-	.tracking file only, indicates multiple classification

(http://cufflinks.cbcb.umd.edu/manual.html#cuffcompare)

Cuffmerge

- In multi-sample RNA-Seq experiment, sometime it is necessary to pool the data and assemble them into a comprehensive set of transcripts before differential analysis.
- Pool aligned reads from all samples and run Cufflinks once on them is not recommended:
  - Assembly becomes more computationally expensive as read depth increases.
  - Complex mixture of splice isoforms for many genes may lead to the incorrectly assembled transcripts.
- As a ‘meta-assembler’, Cuffmerge parsimoniously merges the individually assemblies by Cufflinks by treating the assembled transfrags the way Cufflinks treats the reads.
- It can also performs a reference annotation-based transcript (RABT) (Roberts et al. 2011) assembly to merge reference transcripts with assembled sample transfrags to produce a single annotation file for downstream differential analysis.
- Web site: http://cufflinks.cbcb.umd.edu/manual.html#cuffmerge

Merge Transcriptome Annotations

Check the “cuffmerge” command options:

```
$ cuffmerge
cuffmerge:
cuffmerge takes two or more Cufflinks GTF files and merges them into a
single unified transcript catalog.  Optionally, you can provide the script
with a reference GTF, and the script will use it to attach gene names and other
metadata to the merged catalog.

Usage:
    cuffmerge [Options] <assembly_GTF_list.txt>

Options:
    -h/--help                               Prints the help message and exits
    -o                     <output_dir>     Directory where merged assembly will be written  [ default: ./merged_asm  ]
    -g/--ref-gtf                            An optional "reference" annotation GTF.
    -s/--ref-sequence      <seq_dir>/<seq_fasta> Genomic DNA sequences for the reference.
    --min-isoform-fraction <0-1.0>          Discard isoforms with abundance below this       [ default:          0.05 ]
    -p/--num-threads       <int>            Use this many threads to merge assemblies.       [ default:            1  ]
    --keep-tmp                              Keep all intermediate files during merge
```

Merge Assembled Transcripts

```
$ cat ~/rnaseq-shared/pbs_scripts/cuffmerge.qs
#PBS -N Cuffmerge
#PBS -S /bin/bash
#PBS -V
#PBS -l ncpus=1,walltime=16:00:00,cput=10:00:00,mem=2000mb,nodes=1:ppn=4
#PBS –q rnaseq

cd $PBS_O_WORKDIR
cuffmerge -p 4 --keep-tmp -g reference/gallus_chr1.gtf -s reference/gallus_chr1.fa -o  cuffmerge_out gtf_out_list.txt

$ qsub ~/rnaseq-shared/pbs_scripts/cuffmerge.qs
90294.biohen.dbi.local

$ qstat -a

biohen.dbi.local:
                                                                   Req'd  Req'd   Elap
Job ID               Username Queue    Jobname          SessID NDS   TSK Memory Time  S Time
-------------------- -------- -------- ---------------- ------ ----- --- ------ ----- - -----
90294.biohen.dbi     chenc    cbcb     Cuffmerge         53740     1   4 2000mb 600:0 C 00:00

$ ls –tlr cuffmerge_out
total 28532
drwxr-xr-x 2 chenc cwu       20 May 17 14:50 logs
drwxr-xr-x 2 chenc cwu     4096 May 17 14:50 tmp
-rw-r--r-- 1 chenc cwu        0 May 17 14:50 skipped.gtf
-rw-r--r-- 1 chenc cwu 14093045 May 17 14:50 transcripts.gtf
-rw-r--r-- 1 chenc cwu   427880 May 17 14:50 isoforms.fpkm_tracking
-rw-r--r-- 1 chenc cwu   213996 May 17 14:50 genes.fpkm_tracking
-rw-r--r-- 1 chenc cwu 14467073 May 17 14:50 merged.gtf
```

Exercise 4

View Alignment, Coverage, and Isoforms

(SAMtools, IGV)

Index Alignment Files for IGV

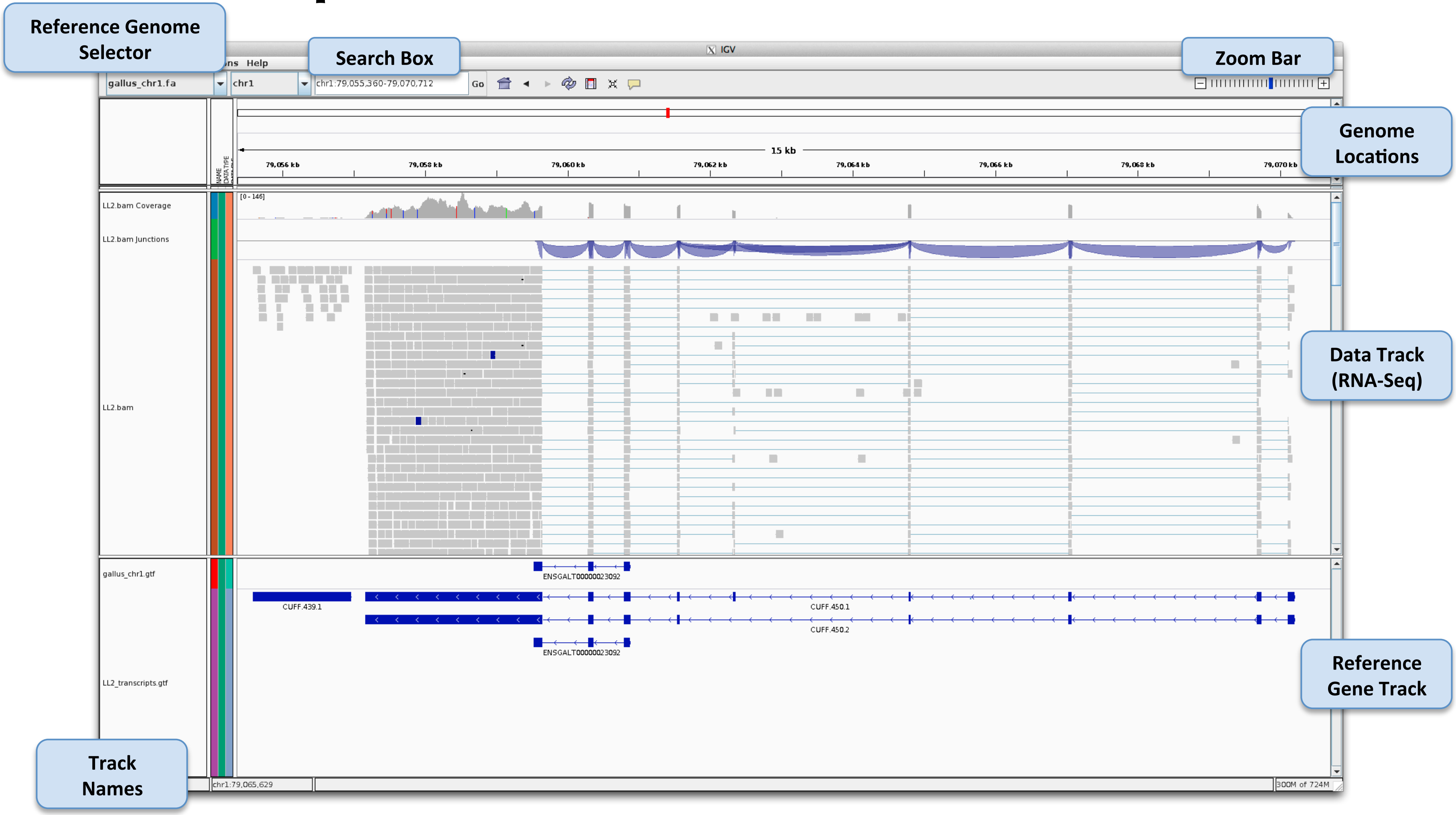
```
$ cat ~/rnaseq-shared/pbs_scripts/samtools_index.qs
#PBS -N SamtoolsIndex
#PBS -S /bin/bash
#PBS -V
#PBS -l ncpus=1,walltime=16:00:00,cput=10:00:00,mem=2000mb,nodes=1:ppn=4
#PBS -q rnaseq

cd $PBS_O_WORKDIR
samtools faidx index/gallus_chr1.fa
ln -s accepted_hits.bam tophat_out_FL1/FL1.bam
ln -s accepted_hits.bam tophat_out_FL2/FL2.bam
ln -s accepted_hits.bam tophat_out_LL1/LL1.bam
ln -s accepted_hits.bam tophat_out_LL2/LL2.bam

samtools index tophat_out_FL1/FL1.bam
samtools index tophat_out_FL2/FL2.bam
samtools index tophat_out_LL1/LL1.bam
samtools index tophat_out_LL2/LL2.bam

ln -s transcripts.gtf cufflinks_out_FL1/FL1_transcripts.gtf
ln -s transcripts.gtf cufflinks_out_FL2/FL2_transcripts.gtf
ln -s transcripts.gtf cufflinks_out_LL1/LL1_transcripts.gtf
ln -s transcripts.gtf cufflinks_out_LL2/LL2_transcripts.gtf

$ qsub ~/rnaseq-shared/pbs_scripts/samtools_index.qs
90297.biohen.dbi.local
$ ls –ltr tophat_out_*/
$ ls -tlr index/
total 412968
-rw-r--r-- 1 chenc cwu  48797843 May 17 14:03 gallus_chr1.4.ebwt
-rw-r--r-- 1 chenc cwu     89909 May 17 14:03 gallus_chr1.3.ebwt
-rw-r--r-- 1 chenc cwu  60083414 May 17 14:06 gallus_chr1.1.ebwt
-rw-r--r-- 1 chenc cwu  24398928 May 17 14:06 gallus_chr1.2.ebwt
-rw-r--r-- 1 chenc cwu  60083414 May 17 14:08 gallus_chr1.rev.1.ebwt
-rw-r--r-- 1 chenc cwu  24398928 May 17 14:08 gallus_chr1.rev.2.ebwt
-rw-r--r-- 1 chenc cwu 205013902 May 17 15:07 gallus_chr1.fa
-rw-r--r-- 1 chenc cwu        23 May 17 15:07 gallus_chr1.fa.fai
```

Launch IGV

```
$ qsub -I -X -V -q rnaseq

$ cd ~/rnaseq-work

$ igv.sh
```

- Click “File”, then click “Load Genome from File …”, then select **gallus_chr1.fa** from the directory called **reference**.
- Click “File”, then click “Load from File …”, then select **gallus_chr1.gtf** from the directory called **reference**.
- Click “File”, then click “Load from File …”, then select **LL2.bam** from the directory called **tophat_out_LL2**.
- Click “File”, then click “Load from File …”, then select **LL2_transcripts.gtf** from the directory called **cufflinks_out_LL2**.
- Type in **chr1:79055360-79070712** in the search box and click “Go”.
- Right click **LL2_transcripts.gtf** track and select “Expanded”.

View Splice Junctions and Isoforms of MFAP5



Reference Genome Selector

Search Box

Zoom Bar

Genome Locations

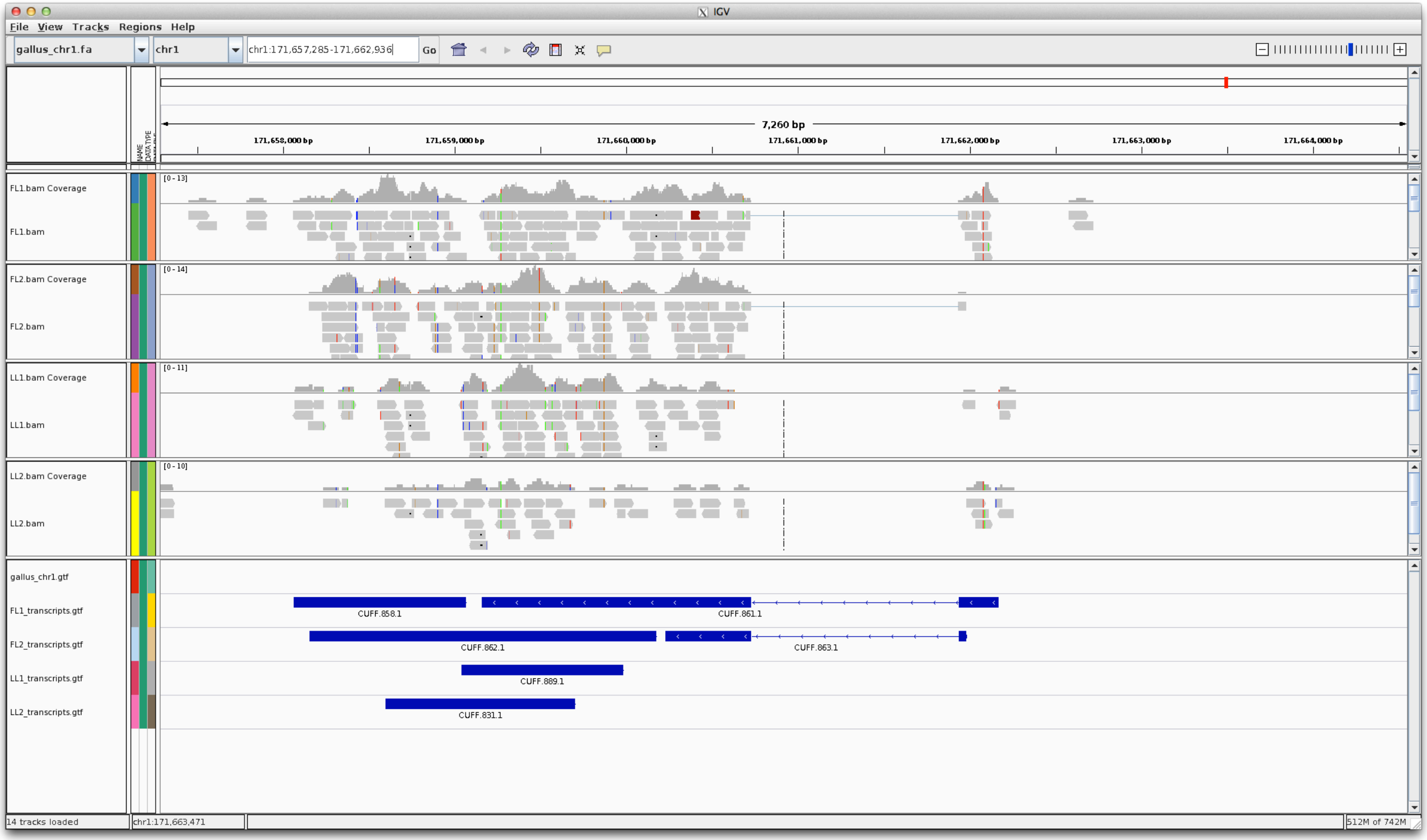
Data Track (RNA-Seq)

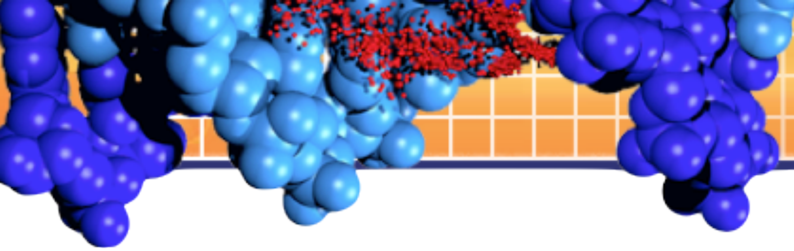
Reference Gene Track

Track Names

View Differentially Expressed Novel Gene

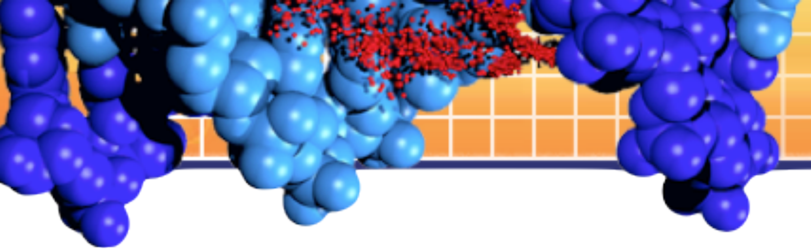
- Type in **chr1:171657285-171662936** in the search box and click “Go”.





Summary

- Create Bowtie index of the reference genome.
- Use TopHat to align paired-end RNA-Seq reads to the reference genome.
- Use Cufflinks to assembly and quantify the transcriptome.
- Use Cuffcompare to find the differences between each individual assembly and the reference genome.
- Use Cuffmerge to merge the individual assemblies generated by Cufflinks.
- Use IGV to view the read alignment and coverage.



References

- Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform first published online April 19, 2012 doi:10.1093/bib/bbs017.
- Langmead B et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
- Li H et al. The Sequence Alignment/Map format and SAMTools. Bioinformatics. 2009 Aug 15;25(16):2078-9.
- Roberts, A et al. Identification of novel transcripts in annotated genomes using RNA-seq. Bioinformatics 27, 2325–2329 (2011).
- Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009 May 1;25(9):1105-11.
- Trapnell C et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5):511-5.
- Trapnell C et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012 Mar 1;7(3):562-78.