

Learning Goals

1. Run Tophat/Bowtie alignments of reads to see what are expressed regions
2. Run EST/Transcripts to genome alignments to find genes
3. Run protein to genome alignments to find genes
4. Visualize these results in Browser (IGV)

Running RNASeq Alignments

1. Download sequence for genome, proteins, and RNAs

```
wget http://stajichlab.github.io/GenomeAnnotation/data/locus.tar.gz
```

2. Uncompress the file

```
tar xzf locus.tar.gz # uncompress the small dataset
```

3. Align the raw sequence reads against the genome locus with Bowtie/TopHat

```
bowtie2-build locus.fa locus # index the database
tophat locus RNASeq_locusonly.3H.fq # run the search
samtools index tophat_out/accepted_hits.bam
```

- Let's investigate that alignment file.
- Open IGV.
- Load locus.fa from the Genomes menu
- File - Load the tophat_out/accepted_hits.bam
- File - Load locus.fungidb.gff

Aligning ESTs to the genome

1. Align ESTs to genome with exonerate

```
exonerate -m e2g ESTs.fa locus.fa --showtargetgff > EST.aln.gff
```

- Now load this GFF into IGV to visualize

Aligning Proteins to the genome

5. Align proteins to genome with BLASTX

```
makeblastdb -in mory_proteins.fa -dbtype prot # format the db for BLAST
blastx -query locus.fa -db mory_proteins.fa -outfmt 6 # run BLASTX to find homologs
python blast2gff.py mory.BLASTX.tab BLASTX LGV_locus test > mory_proteins.BLASTX.gff
```

- Now load this GFF into IGV to visualize

6. Align proteins to genome with exonerate

```
exonerate -m p2g mory_proteins.fa locus.fa --showtargetgff > mory_proteins.aln.gff
```

- Now load this GFF into IGV to visualize

Practice with larger datasets

```
wget http://stajichlab.github.io/GenomeAnnotation/data/big.tgz
```

```
tar xzf big.tgz
```

```
wget http://www.fungidb.org/common/downloads/Current_Release/Fgraminearum_PH-1/fasta/data/Fu
```

1. Look in the new folder 'big'
2. there is a whole chromosome file now NcraOR74A_LGV.fa; Index this with bowtie2-build and run tophat
3. Use this file Ncra3H_ChrV_reads.fastq to align to the genome with tophat.
4. Load your new aligned bamfile reads (Step 3) and the genes in Ncra_OR74A_LGV.genes.gff
5. Use this file Nc5H-Trinity.fasta to align transcripts to the chromosome with exonerate
6. Load the chromosome NcraOR74A_LGV.fa into IGV and load its annotations NcraOR74A_LGV.genes.gff
7. Use the downloaded file from another genome FungiDB-27_Fgraminearum_PH-1_AnnotatedProteins.fasta to align proteins to this chromosome with BLASTX
 - You can try to run exonerate but it works better if you already have a subset of proteins that align to this chromosome as exonerate will try to align all proteins in the file (will take a while).
8. Load some of the alignments into IGV if you get it to work