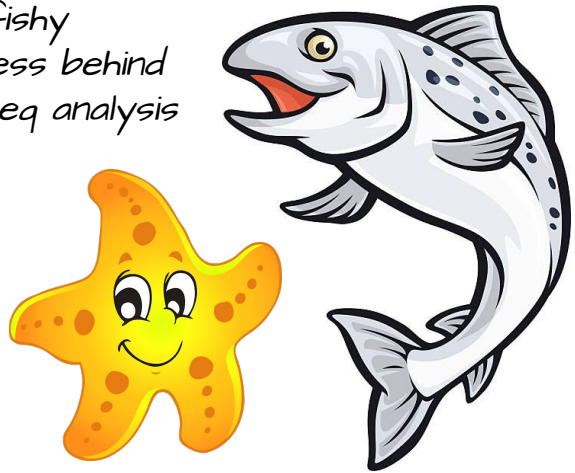
BCB 716 - Sequence Analysis

The fishy business behind RNAseq analysis



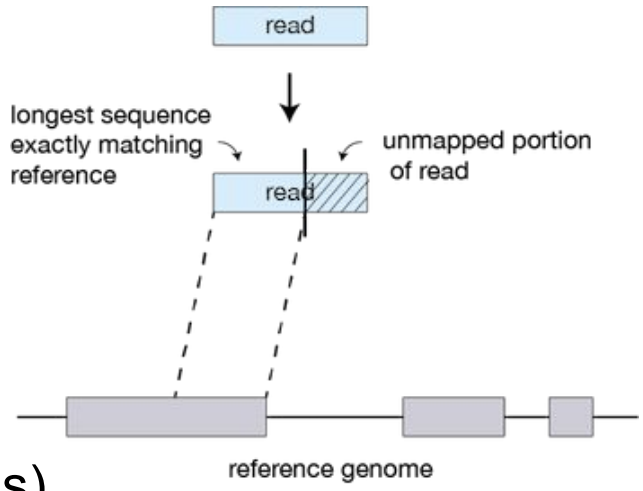
- Problems with problem sets
- In-class exercises
- Final take home?

RNA Analysis Pipelines

Alignment based RNAseq tools

Splice-aware genome aligners (STAR, HiSat2, MapSplice2, GSNAP, etc)

- STAR (Spliced Transcripts Alignment to a Reference)
- Fast two-stage alignment
  - Seed search
  - Clustering, stitching, and scoring
- STAR searches for the longest exact match to one or more genomic locations. These longest matching sequences are called the Maximal Mappable Prefixes (MMPs)



Seed Mapping

Different parts of the read, called ‘seeds’, are mapped separately. The first MMP that is mapped to the genome is called seed1.

STAR next searches for the unmapped portion of the read to find the next MMP, which will be seed2.

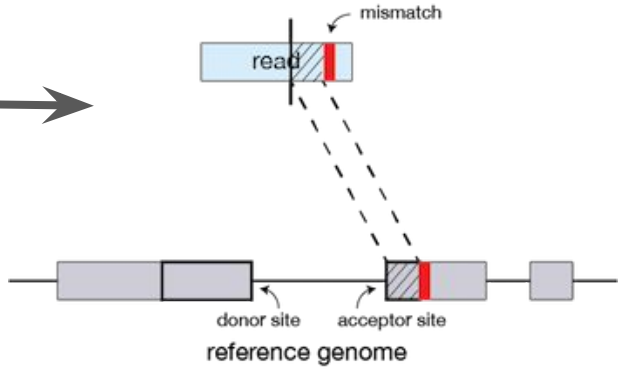
This sequential "greedy" search for unmapped portions of reads underlies the efficiency of the STAR algorithm.

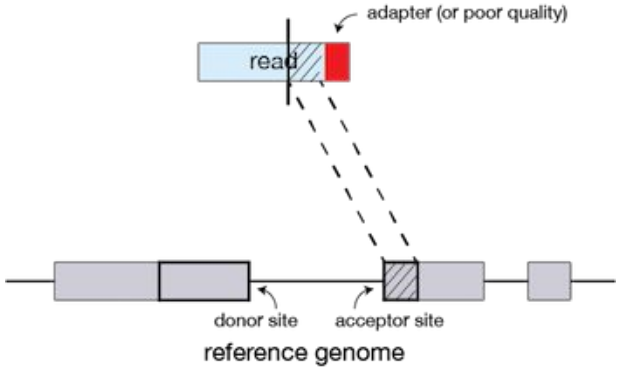
STAR uses an uncompressed suffix array (SA) to search for MMPs.

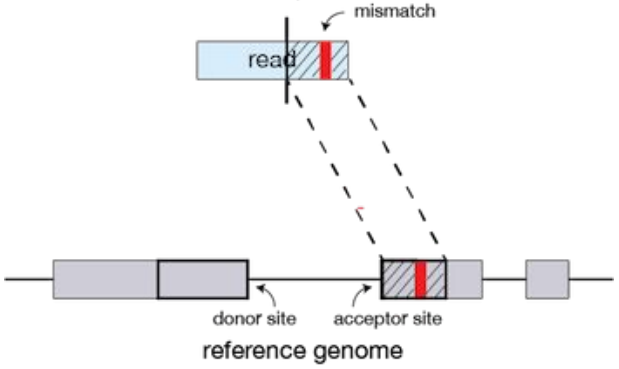
Handling Mismatches

If STAR doesn't find an exact match sequence for each part of the read due to mismatches or indels, the previous MMPs are extended.

If extension does not give a good alignment, the poor quality or adapter sequence (or other contaminating sequence) will be soft clipped.



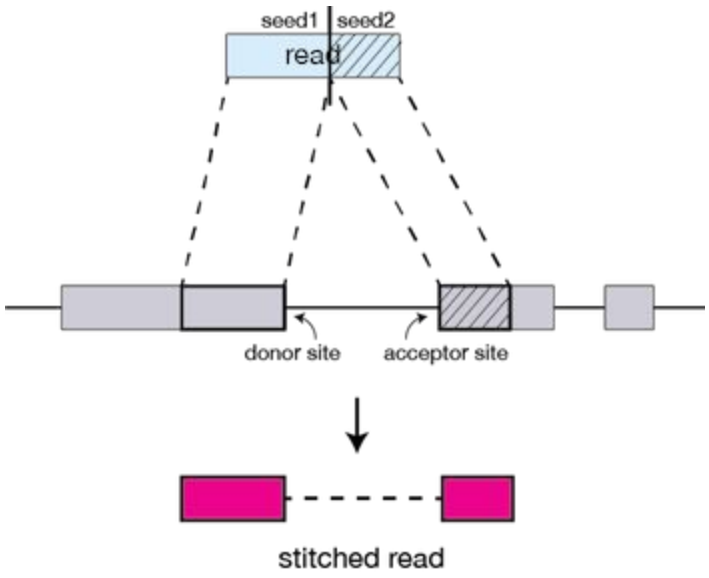




Clustering, Stitching, and Scoring

Separate seeds are stitched together to create a complete read by first clustering the seeds together based on proximity to a set of ‘anchor’ seeds, or seeds that are not multi-mapping.

Then the seeds are stitched together based on the best alignment for the read (scoring based on mismatches, indels, gaps, etc.).



Star Command-line Options

The basic options to **generate genome indices** using STAR are as follows:

- `--genomeDir`: /path/to/store/genome_indices
- `--readFilesIn`: /path/to/FASTQ_file
- `--outFileNamePrefix`: prefix for all output files
- `--runThreadN`: number of threads
- `--outSAMtype`: output filetype (SAM default)
- `--outSAMunmapped`: what to do with unmapped reads
- `--sjdbOverhang`: readlength -1

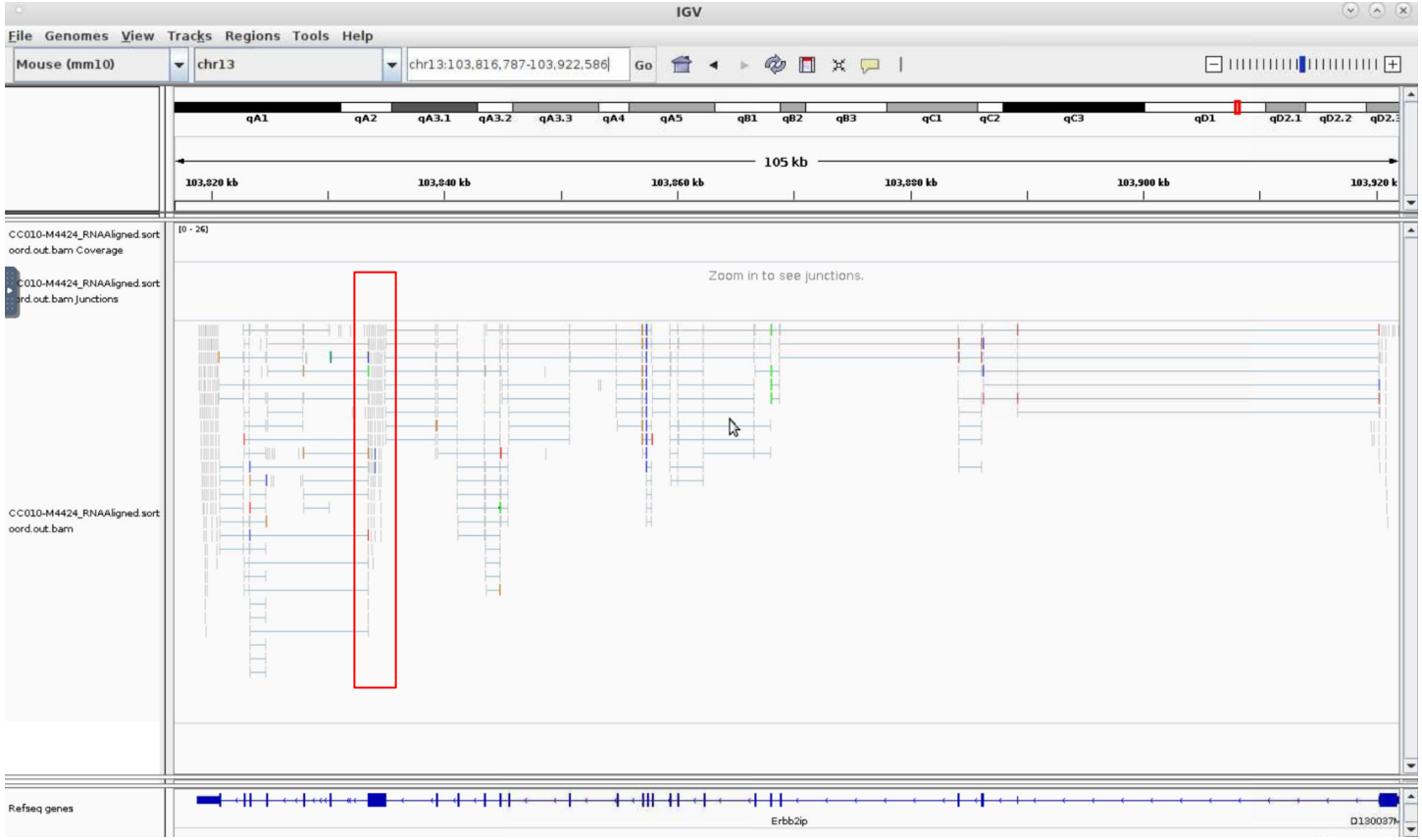
Star Sbatch Script

```
longleaf-login4$ cat bin/runStar
#!/bin/bash
#SBATCH --ntasks=6
#SBATCH --time=4:00:00
#SBATCH --mem=32G
reffile=$1
outfile="${!#}"
fqfiles="${@:2:$#-2}"
echo "ref genome:" $reffile
echo "input files:" $fqfiles
echo "output file:" $outfile
star --genomeDir $reffile --runThreadN 6 --readFilesIn $fqfiles --outFileNamePrefix
$outfile --outSAMtype BAM SortedByCoordinate --outSAMunmapped Within --outSAMattributes
Standard
longleaf-login4$ sbatch runStar /proj/seq/data/STAR_genomes_v277/GRCm38_p6_GENCODE_primary
/proj/mcmillanlab/BCB716F21/RNAseq/CC001_F5694_L001_R1.fastq
/proj/mcmillanlab/BCB716F21/RNAseq/CC001_F5694_L001_R2.fastq
/pine/scr/m/c/mcmillan/alignments/CC001_F5694_RNA
Submitted batch job 33793629
longleaf-login4$ samtools index
/pine/scr/m/c/mcmillan/alignments/CC001_F5694_RNAAligned.sortedByCoord.out.bam
```

Runs typically take less than 1 hour

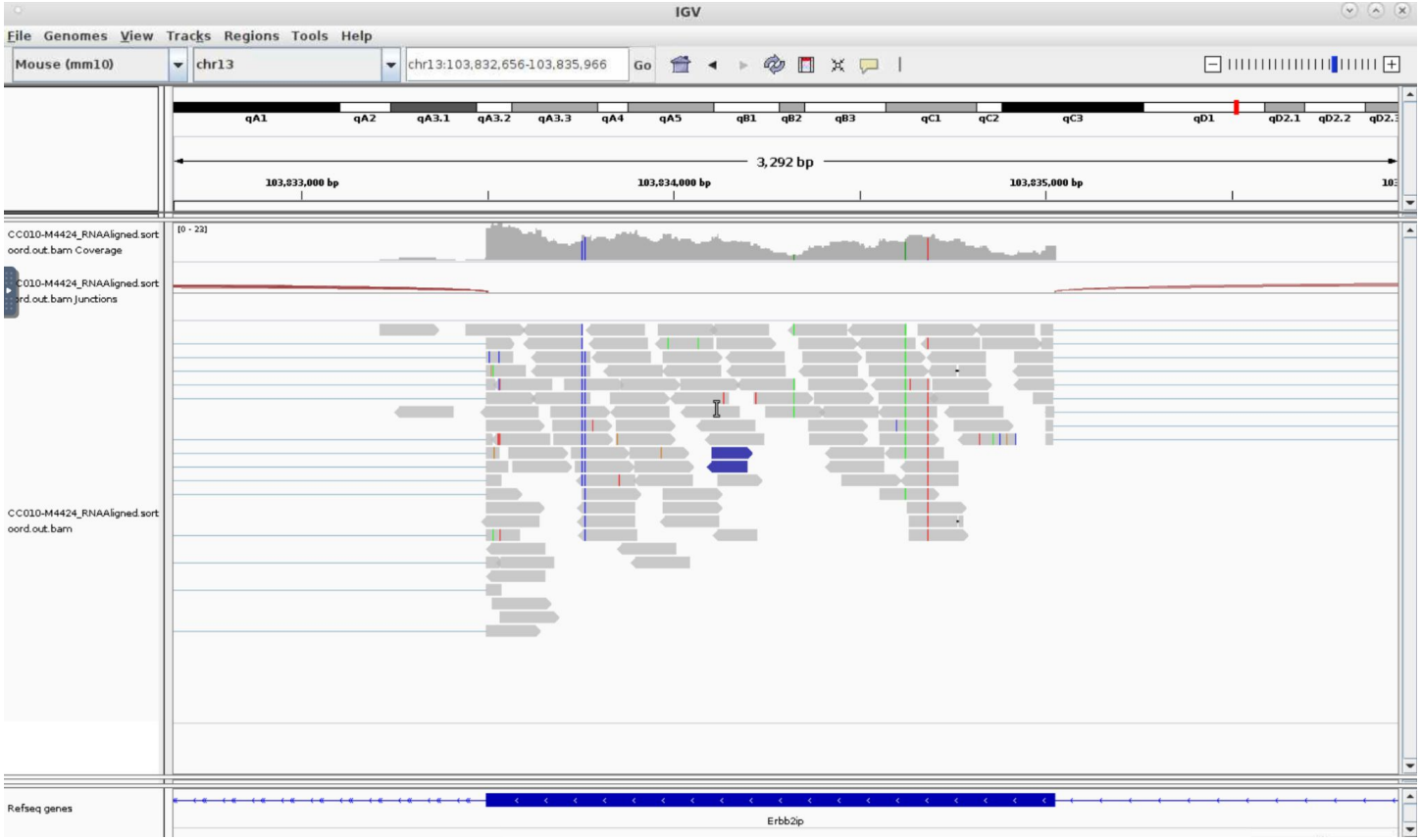
IGV an RNAseq dataset

An expressed gene



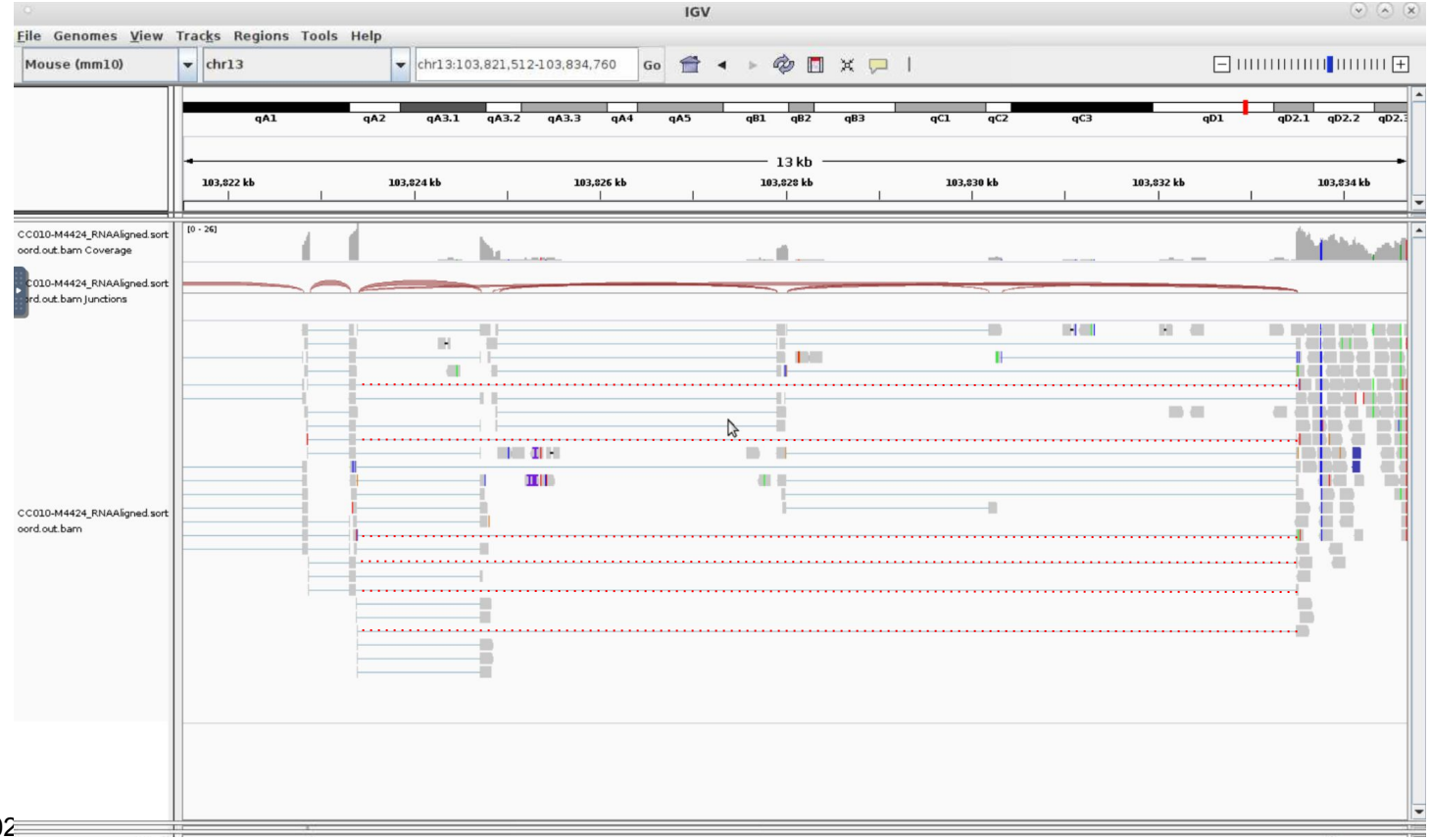
IGV an RNAseq dataset

4 variants in an Exon



IGV an RNAseq dataset

Alternate splice versions



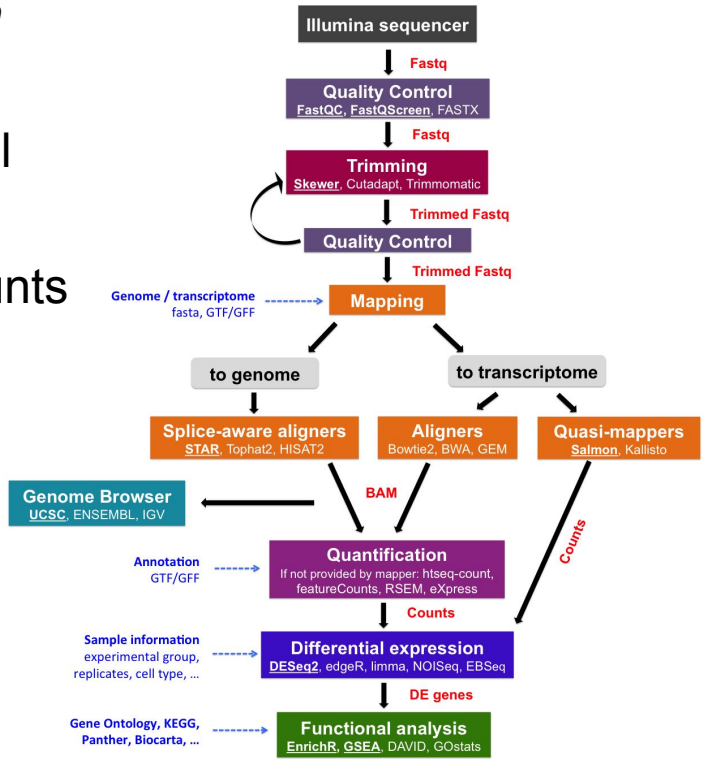
RNAseq quantification

RNA-Seq quantification estimates *gene expression levels* from RNAseq data.

Used for cross-sample comparisons and differential expression analysis

Results are reported in two forms raw mapping counts (reads that map to a gene) and normalized TPM (Transcripts per Million mapped reads) values.

Quantification can be done using alignment-based methods to either the genome or transcriptome but the trend is to quantify using the much faster k-mer based methods.



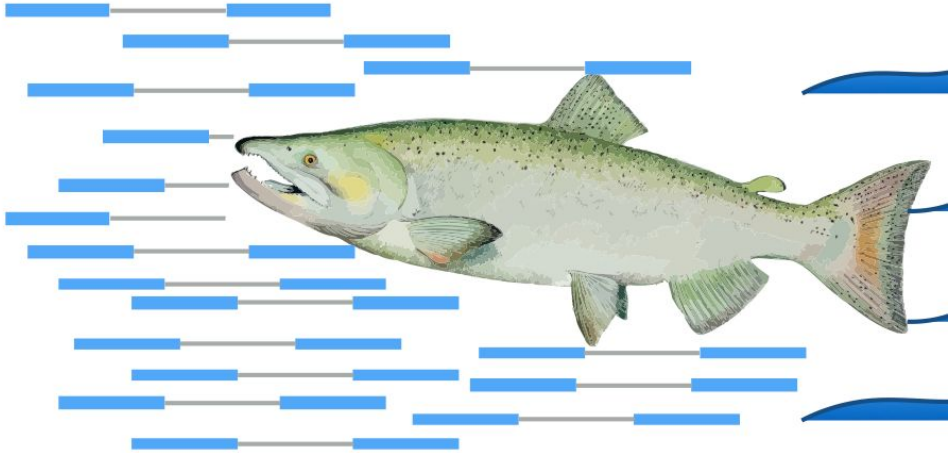
Salmon

Salmon is a tool for quantifying the expression of transcripts using RNA-seq data.

It is a quasi-mapper as it doesn't produce the read alignments (and doesn't output BAM/SAM files).

Salmon "quasi-maps" reads to the *transcriptome* rather than to the *genome* as STAR does.

Salmon can make use of genome alignments (in the form of a SAM/BAM file) to the transcripts rather than using raw reads in the FASTQ format, this is often a good sanity check.



Building a Salmon index

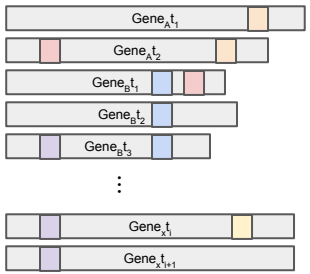
To make an index for Salmon, we need transcript sequences in the FASTA format. Salmon does not need any decompression of the input, so we can index by using this command:

```
$ salmon index -t annotations/gencode.v29.transcripts.fa.gz \
         -i indexes/transcripts --gencode

Version Info: This is the most recent version of salmon.
index ["transcripts"] did not previously exist  . . . creating
it
[2019-04-30 18:12:59.272] [jLog] [info] building index
[2019-04-30 18:12:59.275] [jointLog] [info] [Step 1 of 4] :
counting k-mers

[....]
[2019-04-30 18:18:07.251] [jLog] [info] done building index
```

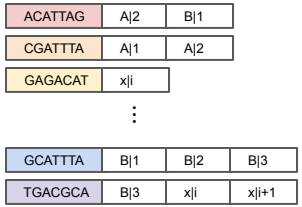
Salmon's Index

Salmon requires an index for quazi-mapping that is derived from a FASTA file of annotated transcripts. This index tracks the association of *informative* k-mers to transcripts.

Transcriptome



Index



The index is used to associate reads with transcripts. Coordinates/offsets aren't important, thus "quasi-mapping"

Salmon processing

1. Reads are scanned from left to right until a k-mer appears in the hash table
2. The k-mer is looked up in the hash table and the involved transcripts are added to a list
3. Scanning continues after the found k-mer
4. This process is repeated until the end of the read.
5. The final set of mappings is determined by a consensus mechanism, i.e. which of the gene/transcript pairs have the most support considering the transcript's orientation
6. A weighted count is added to a table for all potential transcripts consistent with the read.
7. Salmon corrects abundance estimates for any sample-specific biases/factors.
    a. Fragment CG bias
    b. Transcript-length corrections
    c. Fragment-length distributions
    d. Sequence specific biases
8. The weights are then redistributed using an Expectation Maximization (EM) approach.
   Recall if k-mer appeared in multiple transcripts it's count was distributed across all possibilities. However, if other k-mers are one of those transcripts are not seen, we can redistribute the partial counts to transcripts that are supported. This is done iteratively.

Running Salmon

A slurm script for running Salmon, although it typically takes less than 10 mins

```
longleaf-login4$ cat bin/runSalmon
#!/bin/bash
#SBATCH --ntasks=6
#SBATCH --time=4:00:00
#SBATCH --mem=32G
reffile=$1
outfile="${!#}"
fqfiles="${@:2:$#-2}"
echo "ref genome:" $reffile
echo "input files:" $fqfiles
echo "output file:" $outfile
salmon quant -i $reffile --libType A -r $fqfiles -p 16 --validateMappings -o $outfile
longleaf-login4$ sbatch runSalmon /proj/mcmillanlab/BCB716F21/genome/Mouse/salmon
/proj/mcmillanlab/BCB716F21/RNAseq/CC011_M4692_L001_R*
/pine/scr/m/c/mcmillan/alignments/CC011_M4692_Salmon
Submitted batch job 33828974
```

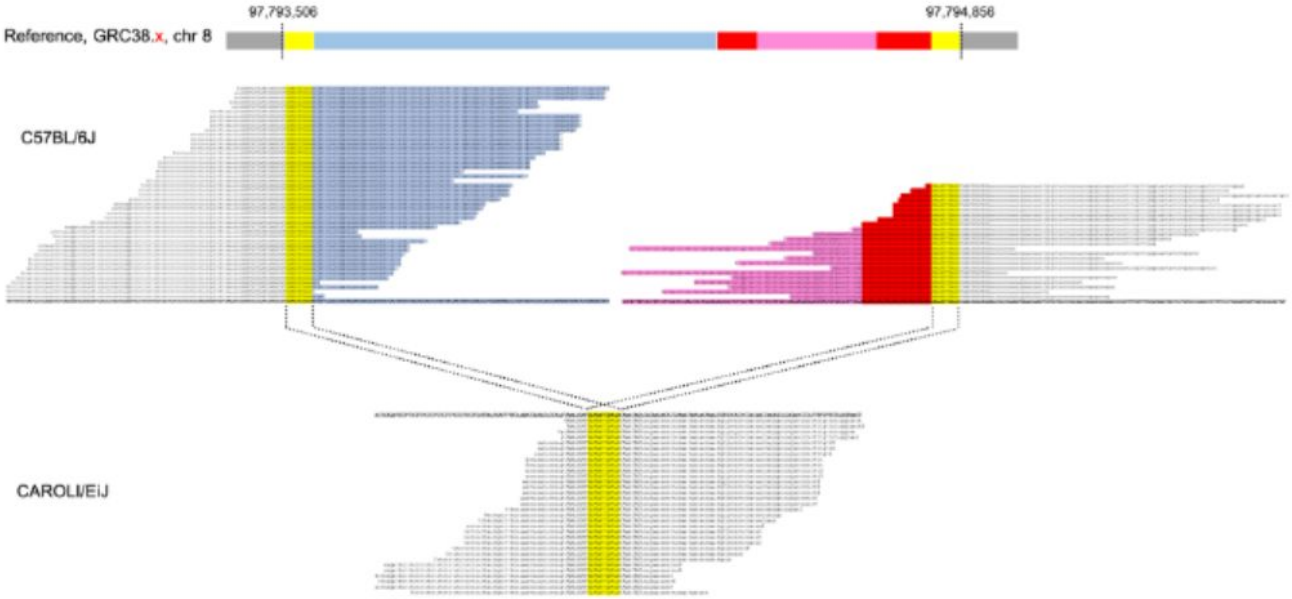
Salmon Results

CC011_M4692_Salmon/quant.sf

File Edit View Insert Format Data Tools Add-ons Help Last edit was 4 minutes ago

Name	Length	EffectiveLength	TPM	NumReads
ENSMUST00000025554|ENSMUSG00000024653|Scgb1a1|476|	476	226	129380.6844	22223
ENSMUST00000082407|ENSMUSG00000064356|mt-Atp8|204|	204	9.377	110569.3858	788
ENSMUST00000022692|ENSMUSG00000022097|Sftpc|804|	804	554	21455.84001	9034
ENSMUST00000084013|ENSMUSG00000065947|mt-Nd4l|297|	297	48.689	17835.57885	660
ENSMUST00000093209|ENSMUSG00000069919|Hba-a1|699|	699	449	14146.7189	4827.551
ENSMUST00000082402|ENSMUSG00000064351|mt-Co1|1545|	1545	1295	11977.94733	11789
ENSMUST00000082405|ENSMUSG00000064354|mt-Co2|684|	684	434	10607.89989	3499
ENSMUST00000082408|ENSMUSG00000064357|mt-Atp6|681|	681	431	10522.99133	3447
ENSMUST00000082409|ENSMUSG00000064358|mt-Co3|784|	784	534	8759.383651	3555
ENSMUST00000082421|ENSMUSG00000064370|mt-Cytb|1144|	1144	894	7361.755169	5002
ENSMUST00000082419|ENSMUSG00000064368|mt-Nd6|519|	519	269	5434.216997	1111
ENSMUST00000092163|ENSMUSG00000069516|Lyz2|1266|	1266	1016	5310.864988	4100.943
ENSMUST00000023934|ENSMUSG00000052305|Hbb-bs|778|	778	528	5070.583704	2034.776
ENSMUST00000112172|ENSMUSG00000049775|Tmsb4x|768|	768	518	4999.261663	1968.16
ENSMUST00000082414|ENSMUSG00000064363|mt-Nd4|1378|	1378	1128	4915.419997	4214
ENSMUST00000025563|ENSMUSG00000024661|Fth1|866|	866	616	4515.43372	2114
ENSMUST00000082392|ENSMUSG00000064341|mt-Nd1|957|	957	707	4047.76275	2175
ENSMUST00000082411|ENSMUSG00000064360|mt-Nd3|348|	348	98.004	3987.371978	297
ENSMUST00000082396|ENSMUSG00000064345|mt-Nd2|1038|	1038	788	3883.816432	2326
ENSMUST00000026148|ENSMUSG00000025150|Cbr2|1110|	1110	860	3821.810809	2498
ENSMUST00000200021|ENSMUSG00000105361|AY036118|923|	923	673	3765.029027	1925.787
ENSMUST00000094434|ENSMUSG00000050708|Ftl1|969|	969	719	3195.144847	1746

Salmon Verification

Next Time

Reference-free Sequence Analysis Approaches