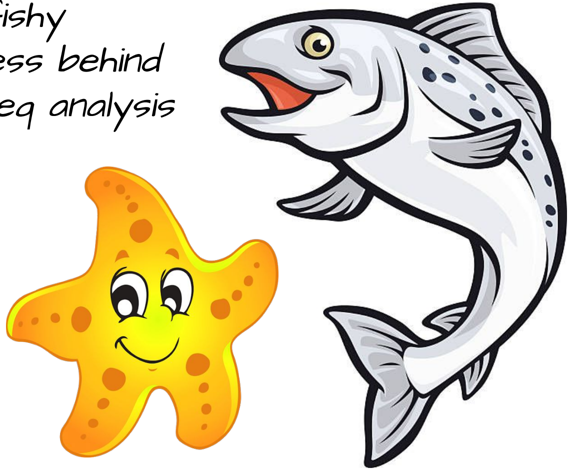
BCB 716 - Sequence Analysis

The fishy business behind RNAseq analysis



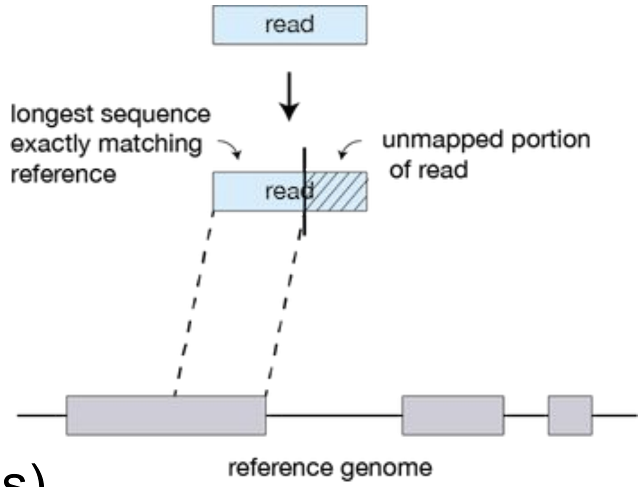
- Regarding the problem sets… at least you are all in the same boat

RNA Analysis Pipelines

Alignment based RNAseq tools

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

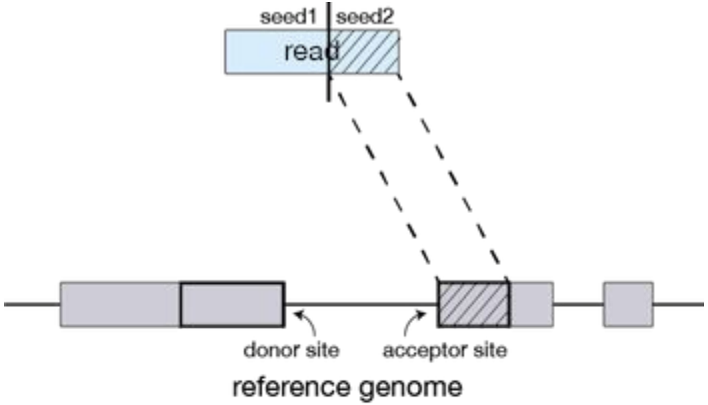
- STAR (Spliced Transcripts Alignment to a Reference)
- Fast two-stage alignment
  - Seed search
  - Clustering, stitching, and scoring
- STAR searches for the longest exact match to one or more genomic locations. These longest matching sequences are called the Maximal Mappable Prefixes (MMPs)



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.

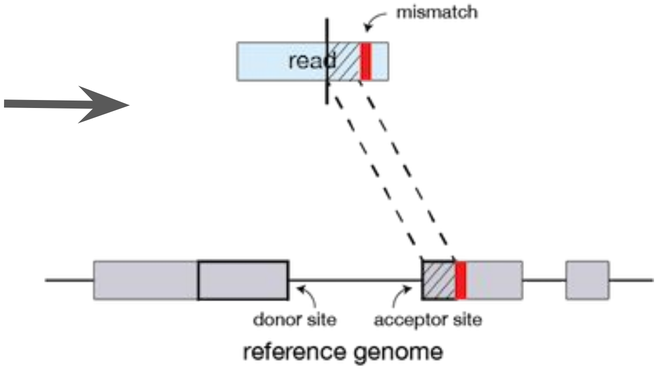


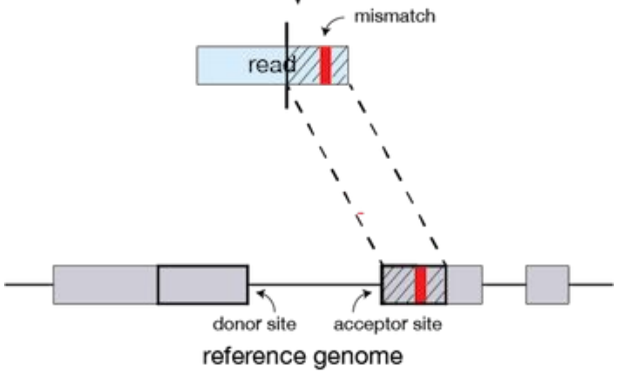
This sequential "greedy" search for unmapped portions of reads underlies the efficiency of the STAR algorithm.

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

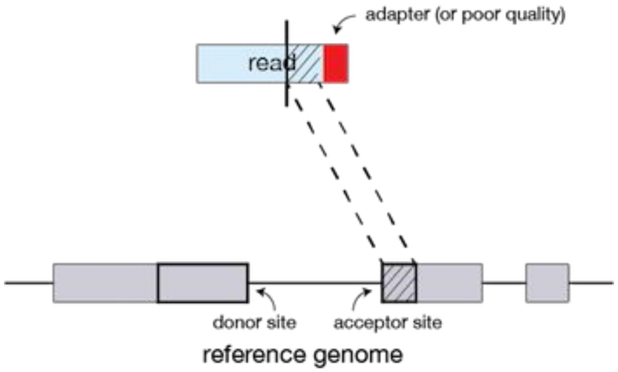
Handling Mismatches

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.





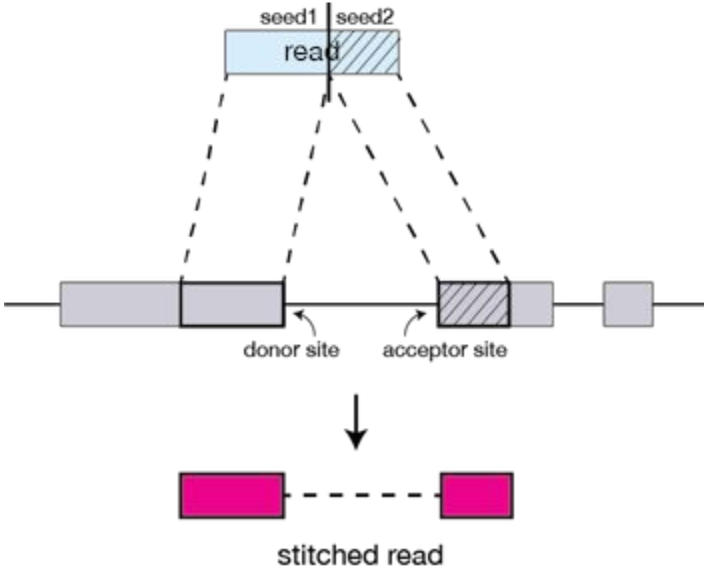
If extension does not give a good alignment, the poor quality or adapter sequence (or other contaminating sequence) will be soft clipped.



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

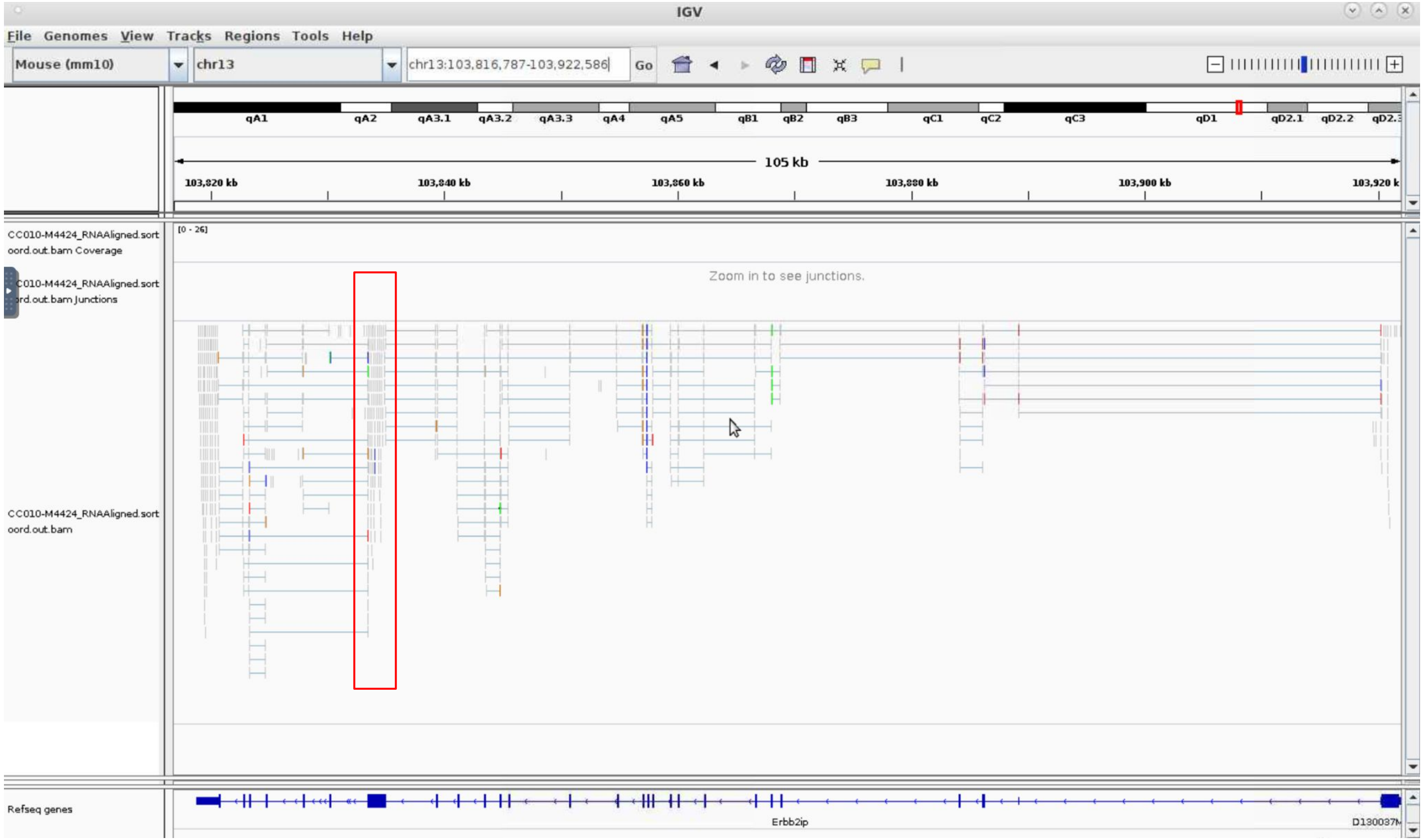
Star Sbatch Script

```
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/BCB716F22/RNAseq/CC001_F5694_L001_R1.fastq
/proj/mcmillanlab/BCB716F22/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



Mouse (mm10)

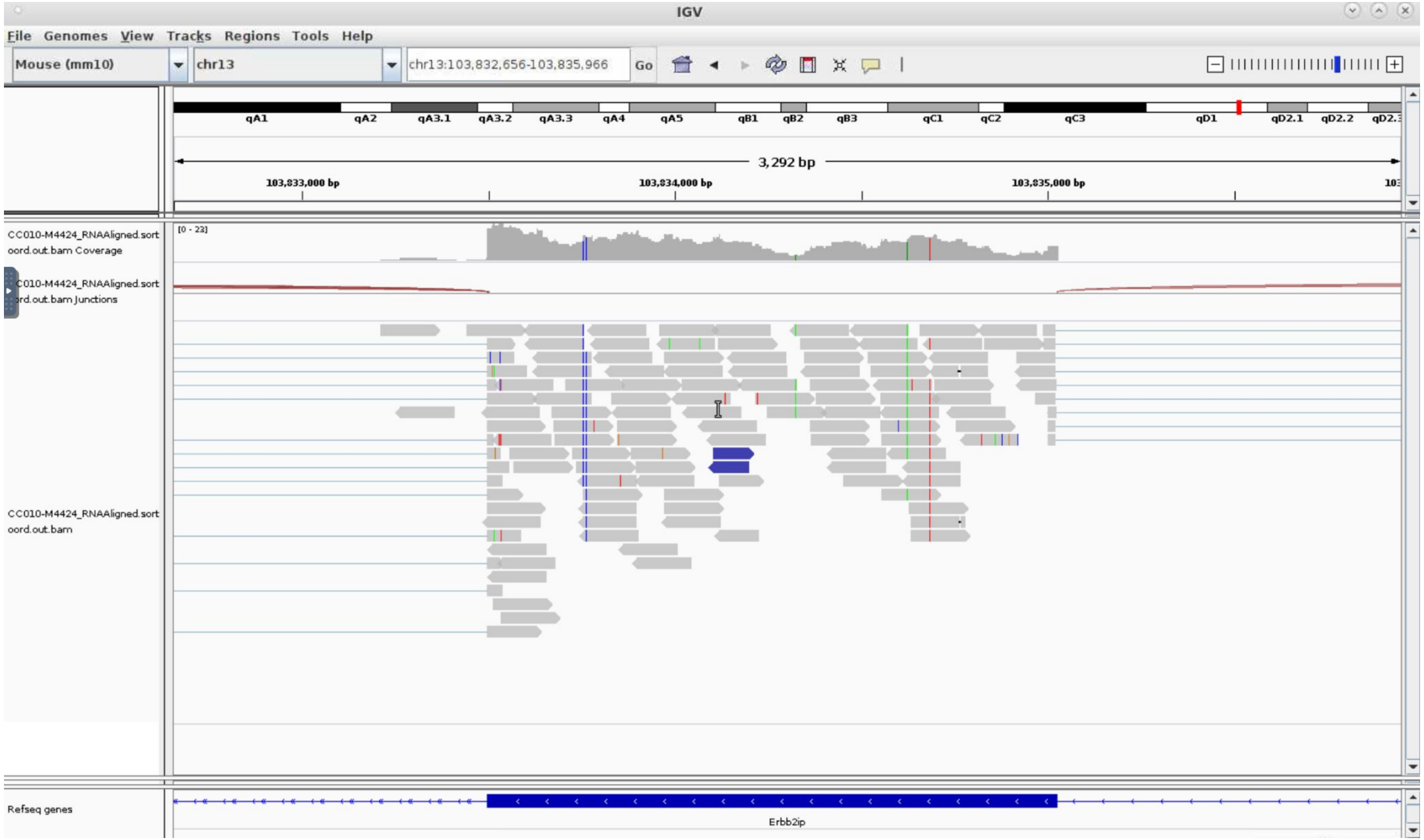
chr13

chr13:103,816,787-103,922,586

IGV
File Genomes View Tracks Regions Tools Help
Go
qA1
qA2
qA3.1
qA3.2
qA3.3
qA4
qA5
qB1
qB2
qB3
qC1
qC2
qC3
qD1
qD2.1
qD2.2
qD2.3
105 kb
103,820 kb
103,840 kb
103,860 kb
103,880 kb
103,900 kb
103,920 kb
CC010-M4424_RNAAligned.sortoord.out.bam Coverage
[0 - 26]
C010-M4424_RNAAligned.sortrd.out.bam Junctions
Zoom in to see junctions.
CC010-M4424_RNAAligned.sortoord.out.bam
Refseq genes
Erbb2ip
D130037M

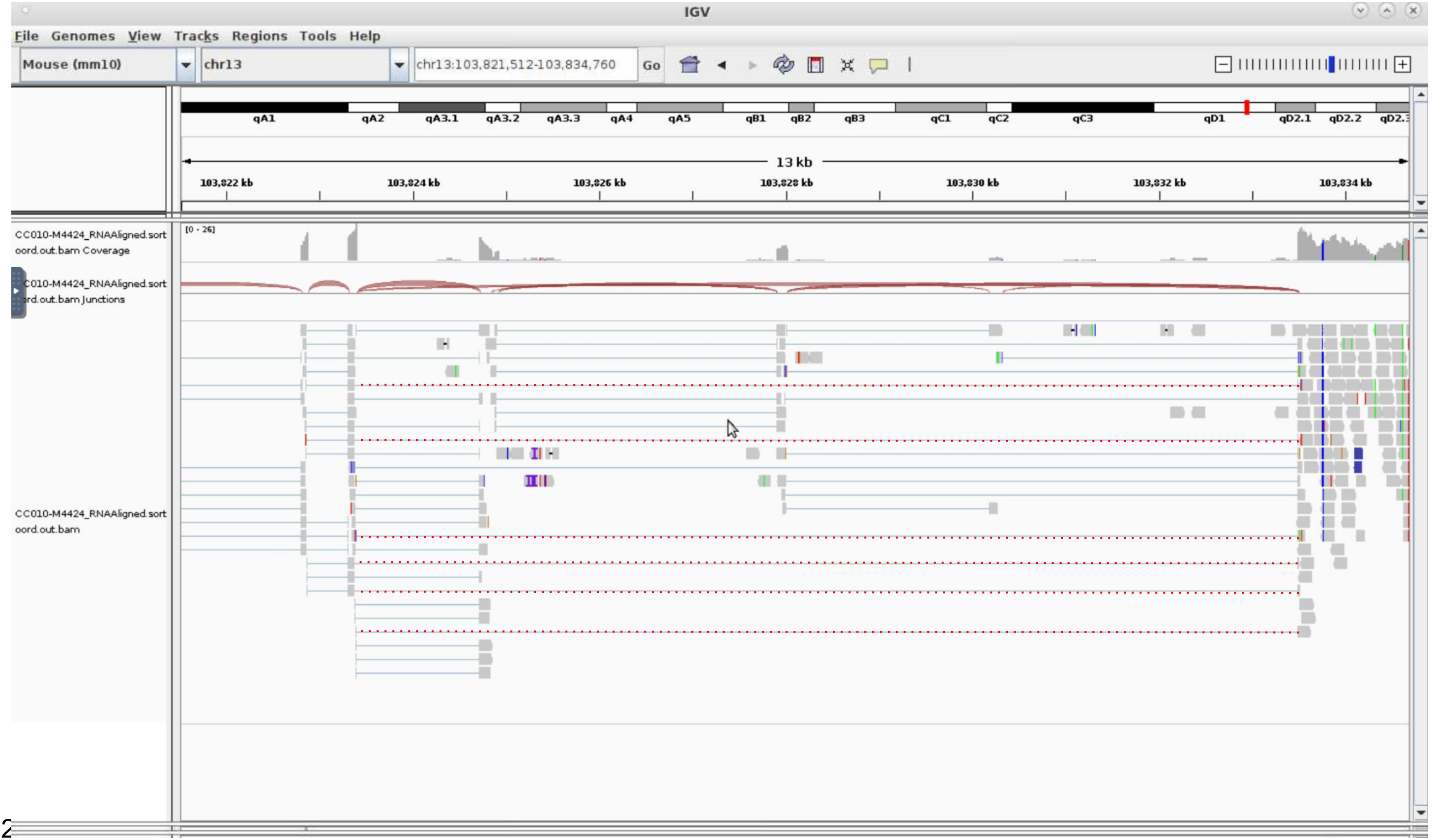
IGV an RNAseq dataset

4 variants in an Exon


IGV
File Genomes View Tracks Regions Tools Help
Mouse (mm10)
chr13
chr13:103,832,656-103,835,966
Go
qA1
qA2
qA3.1
qA3.2
qA3.3
qA4
qA5
qB1
qB2
qB3
qC1
qC2
qC3
qD1
qD2.1
qD2.2
qD2.3
3,292 bp
103,833,000 bp
103,834,000 bp
103,835,000 bp
[0 - 23]
CC010-M4424_RNAAligned.sortedByCoord.out.bam Coverage
CC010-M4424_RNAAligned.sortedByCoord.out.bam Junctions
CC010-M4424_RNAAligned.sortedByCoord.out.bam
Refseq genes
Erbb2ip

IGV an RNAseq dataset

Alternate splice versions



Mouse (mm10)
chr13
chr13:103,821,512-103,834,760

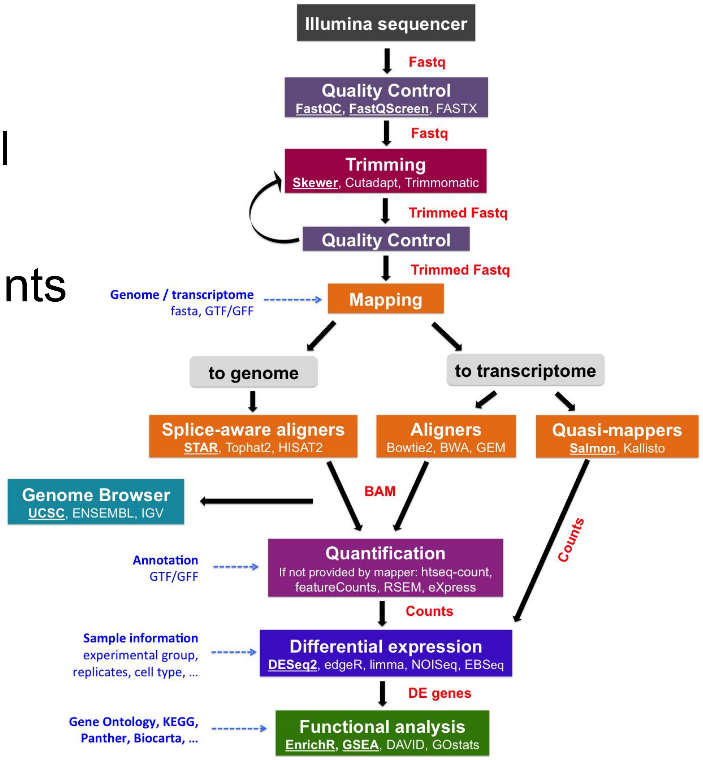
RNAseq quantification

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.



Illumina sequencer

Fastq

Quality Control
FastQC, FastQScreen, FASTX

Fastq

Trimming
Skewer, Cutadapt, Trimmomatic

Trimmed Fastq

Quality Control

Trimmed Fastq

Genome / transcriptome
fasta, GTF/GFF

Mapping

to genome

to transcriptome

Splice-aware aligners
STAR, Tophat2, HISAT2

Aligners
Bowtie2, BWA, GEM

Quasi-mappers
Salmon, Kallisto

Genome Browser
UCSC, ENSEMBL, IGV

BAM

Counts

Annotation
GTF/GFF

Quantification
If not provided by mapper: htseq-count, featureCounts, RSEM, eXpress

Counts

Sample information
experimental group, replicates, cell type, …

Differential expression
DESeq2, edgeR, limma, NOISeq, EBSeq

DE genes

Gene Ontology, KEGG, Panther, Biocarta, …

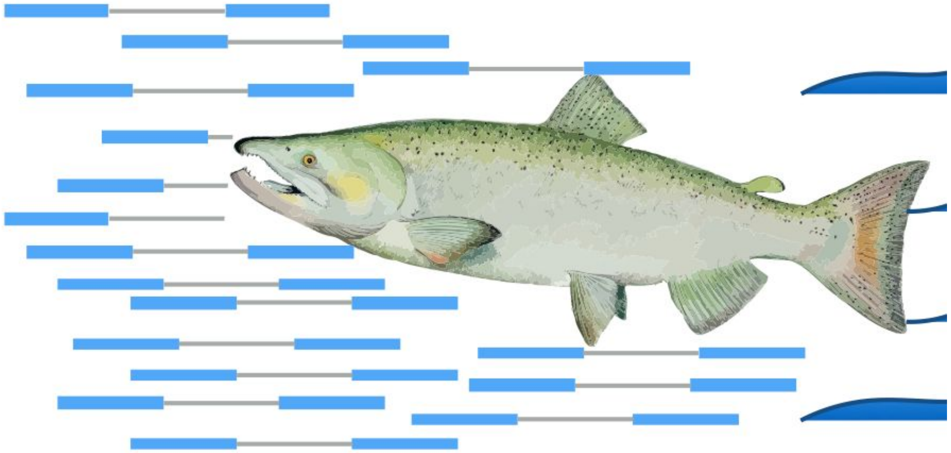
Functional analysis
EnrichR, GSEA, DAVID, GOstats

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:

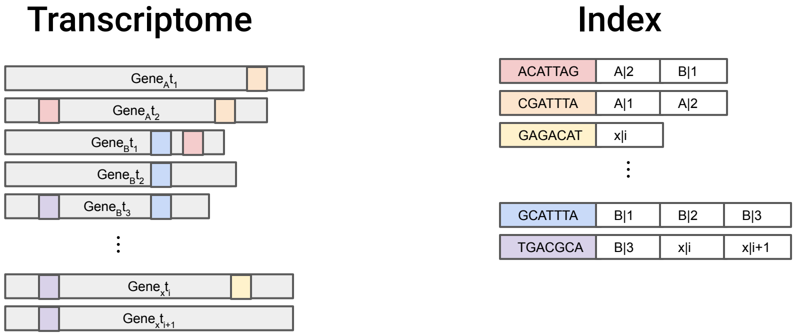
```
$ salmon index -t annotations/gencode.v29.transcripts.fa.gz \
         -i indexes/transcripts --gencode

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.



ACATTAG	A|2	B|1	
CGATTTA	A|1	A|2	
GAGACAT	x|i		
	⋮		
GCATTTA	B|1	B|2	B|3
TGACGCA	B|3	x|i	x|i+1

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

Salmon processing

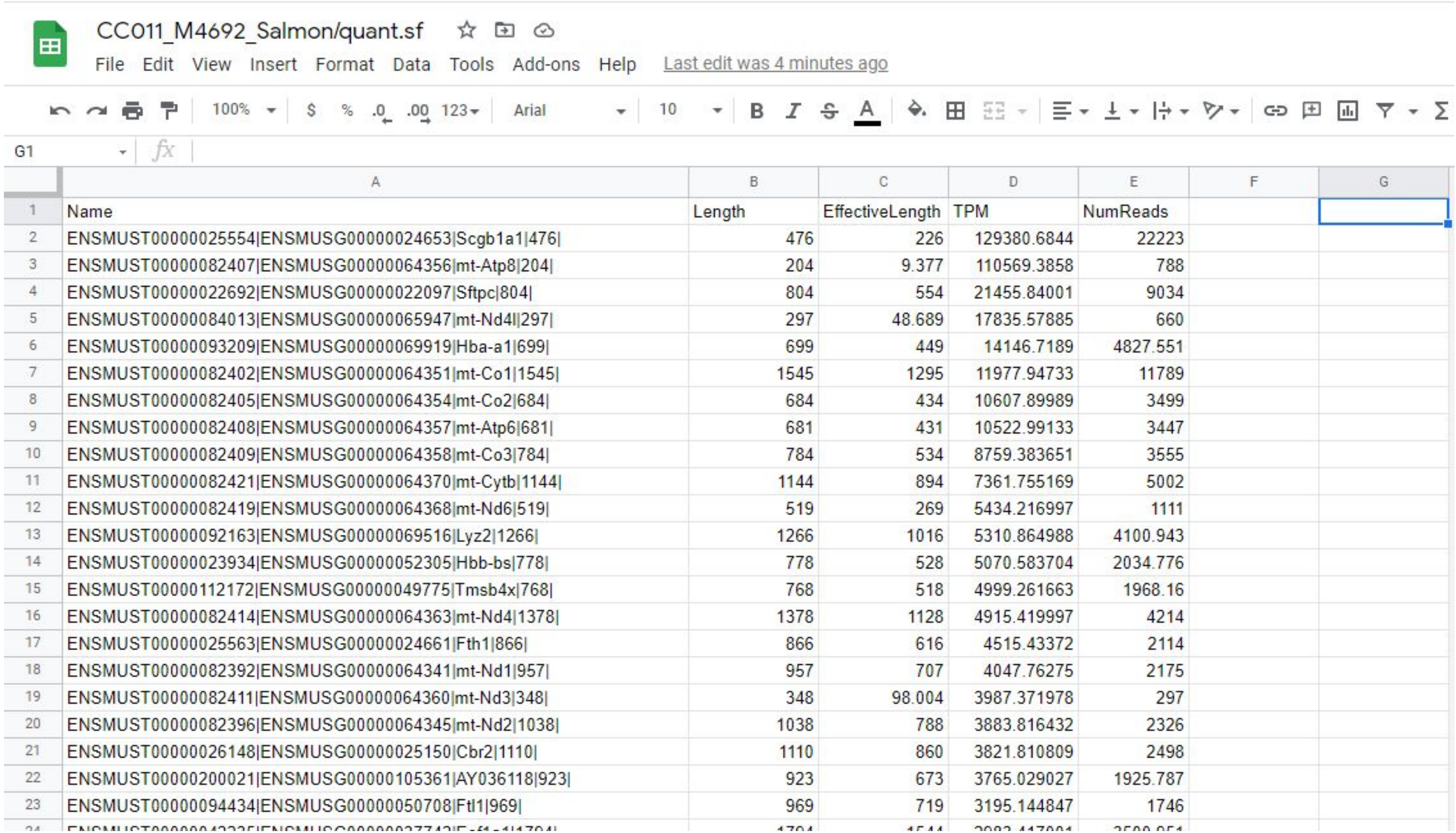
1. Reads are scanned from left to right until a k-mer appears in the hash table
2. The k-mer is looked up in the hash table and the involved transcripts are added to a list
3. Scanning continues after the found k-mer
4. This process is repeated until the end of the read.
5. The final set of mappings 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

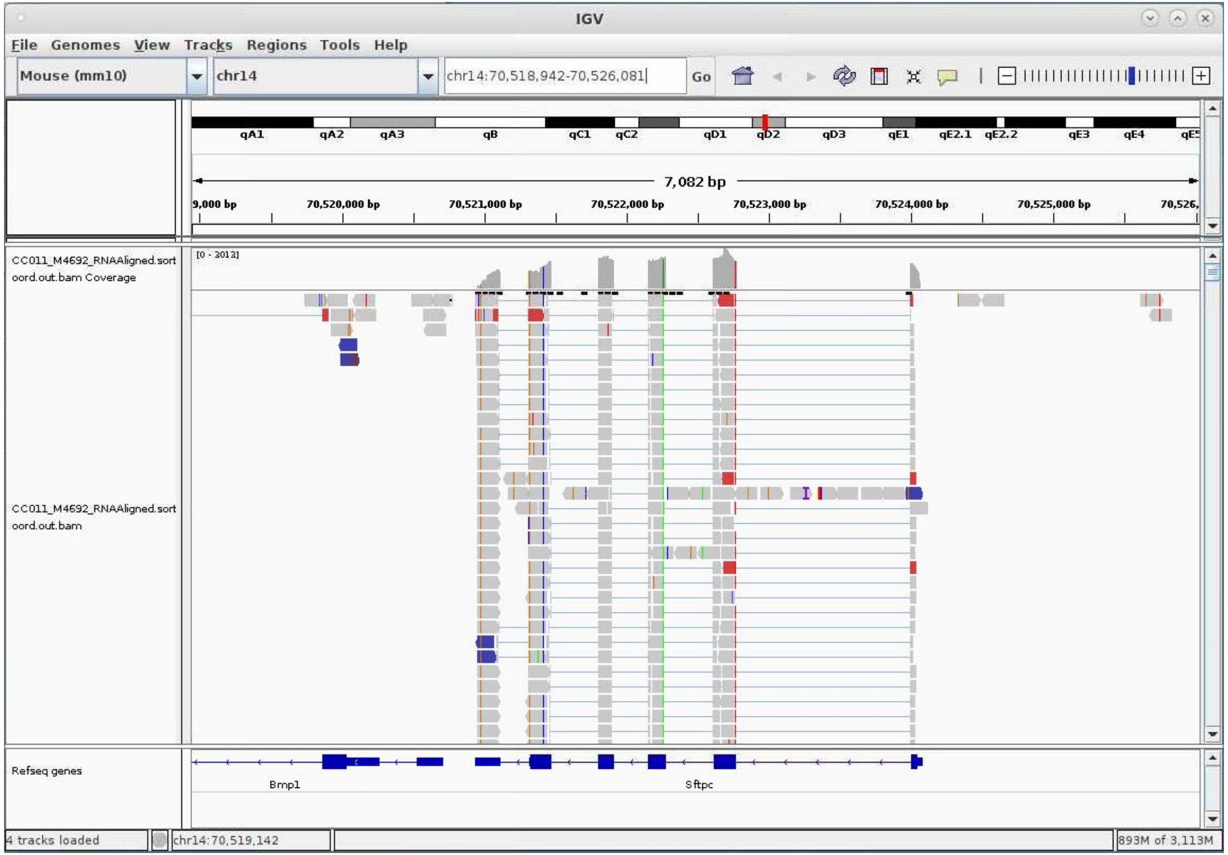
```
longleaf-login4$ cat bin/runSalmon
#!/bin/bash
#SBATCH --ntasks=6
#SBATCH --time=4:00:00
#SBATCH --mem=32G
reffile=$1
outfile="${!#}"
fqfiles="${@:2:$#-2}"
echo "ref genome:" $reffile
echo "input files:" $fqfiles
echo "output file:" $outfile
salmon quant -i $reffile --libType A -r $fqfiles -p 16 --validateMappings -o $outfile
longleaf-login4$ sbatch runSalmon /proj/mcmillanlab/BCB716F21/genome/Mouse/salmon
/proj/mcmillanlab/BCB716F22/RNAseq/CC011_M4692_L001_R*
/pine/scr/m/c/mcmillan/alignments/CC011_M4692_Salmon
Submitted batch job 33828974
```

Salmon Results



	A	B	C	D	E
1	Name	Length	EffectiveLength	TPM	NumReads
2	ENSMUST00000025554|ENSMUSG00000024653|Scgb1a1|476|	476	226	129380.6844	22223
3	ENSMUST00000082407|ENSMUSG00000064356|mt-Atp8|204|	204	9.377	110569.3858	788
4	ENSMUST00000022692|ENSMUSG00000022097|Sftpc|804|	804	554	21455.84001	9034
5	ENSMUST00000084013|ENSMUSG00000065947|mt-Nd4l|297|	297	48.689	17835.57885	660
6	ENSMUST00000093209|ENSMUSG00000069919|Hba-a1|699|	699	449	14146.7189	4827.551
7	ENSMUST00000082402|ENSMUSG00000064351|mt-Co1|1545|	1545	1295	11977.94733	11789
8	ENSMUST00000082405|ENSMUSG00000064354|mt-Co2|684|	684	434	10607.89989	3499
9	ENSMUST00000082408|ENSMUSG00000064357|mt-Atp6|681|	681	431	10522.99133	3447
10	ENSMUST00000082409|ENSMUSG00000064358|mt-Co3|784|	784	534	8759.383651	3555
11	ENSMUST00000082421|ENSMUSG00000064370|mt-Cytb|1144|	1144	894	7361.755169	5002
12	ENSMUST00000082419|ENSMUSG00000064368|mt-Nd6|519|	519	269	5434.216997	1111
13	ENSMUST00000092163|ENSMUSG00000069516|Lyz2|1266|	1266	1016	5310.864988	4100.943
14	ENSMUST00000023934|ENSMUSG00000052305|Hbb-bs|778|	778	528	5070.583704	2034.776
15	ENSMUST00000112172|ENSMUSG00000049775|Tmsb4x|768|	768	518	4999.261663	1968.16
16	ENSMUST00000082414|ENSMUSG00000064363|mt-Nd4|1378|	1378	1128	4915.419997	4214
17	ENSMUST00000025563|ENSMUSG00000024661|Fth1|866|	866	616	4515.43372	2114
18	ENSMUST00000082392|ENSMUSG00000064341|mt-Nd1|957|	957	707	4047.76275	2175
19	ENSMUST00000082411|ENSMUSG00000064360|mt-Nd3|348|	348	98.004	3987.371978	297
20	ENSMUST00000082396|ENSMUSG00000064345|mt-Nd2|1038|	1038	788	3883.816432	2326
21	ENSMUST00000026148|ENSMUSG00000025150|Cbr2|1110|	1110	860	3821.810809	2498
22	ENSMUST00000200021|ENSMUSG00000105361|AY036118|923|	923	673	3765.029027	1925.787
23	ENSMUST00000094434|ENSMUSG00000050708|Ftl1|969|	969	719	3195.144847	1746
24	ENSMUST00000043235|ENSMUSG00000037742|Eef1a1|1794|	1794	1544	2983.417001	2500.951

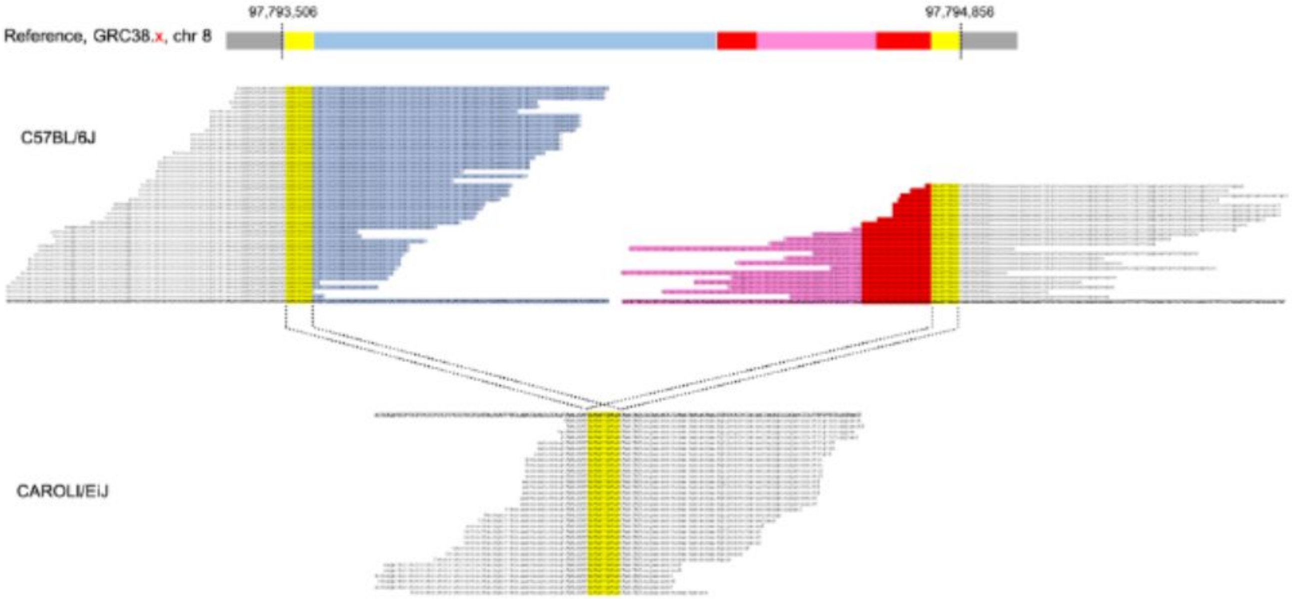
Salmon Verification



IGV
File Genomes View Tracks Regions Tools Help
Mouse (mm10)
chr14
chr14:70,518,942-70,526,081
Go
qA1 qA2 qA3 qB qC1 qC2 qD1 qD2 qD3 qE1 qE2.1 qE2.2 qE3 qE4 qE5
7,082 bp
9,000 bp 70,520,000 bp 70,521,000 bp 70,522,000 bp 70,523,000 bp 70,524,000 bp 70,525,000 bp 70,526,
CC011_M4692_RNAAligned.sortoord.out.bam Coverage
[0 - 3013]
CC011_M4692_RNAAligned.sortoord.out.bam
Refseq genes
Bmp1
Sftpc
4 tracks loaded
chr14:70,519,142
893M of 3,113M

Next Time

Reference-free Sequence Analysis Approaches


97,793,506
97,794,856
Reference, GRC38.x, chr 8
C57BL/6J
CAROLI/EiJ