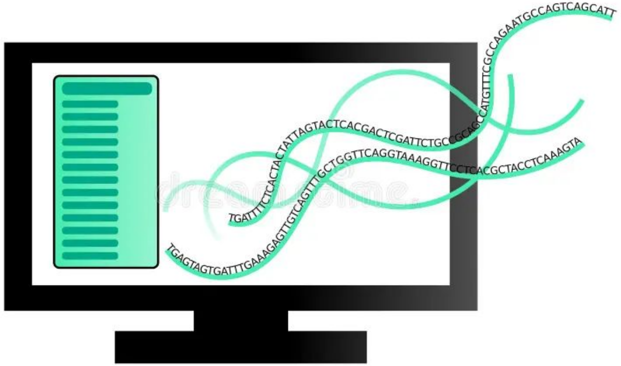
BULK RNA-SEQ: UPSTREAM ANALYSIS


TGATTTTCTCACTACTATTAGTACTCACGACTCGATTCTGCCGCAGCCATGTTTCGCCAGAATGCCAGTCAGCATT
TGAGTAGTGATTTGAAAGAGTTGTCAGTTTGCTGGTTCAGGTAAAGGTTCCTCACGCTACCTCAAAGTA

Presenter: Duy Dao

-03/15/2025-

TABLE OF CONTENTS

RNA SEQ: UPSTREAM ANALYSIS

01 INTRODUCTION

02 RAW DATA PROCESSING

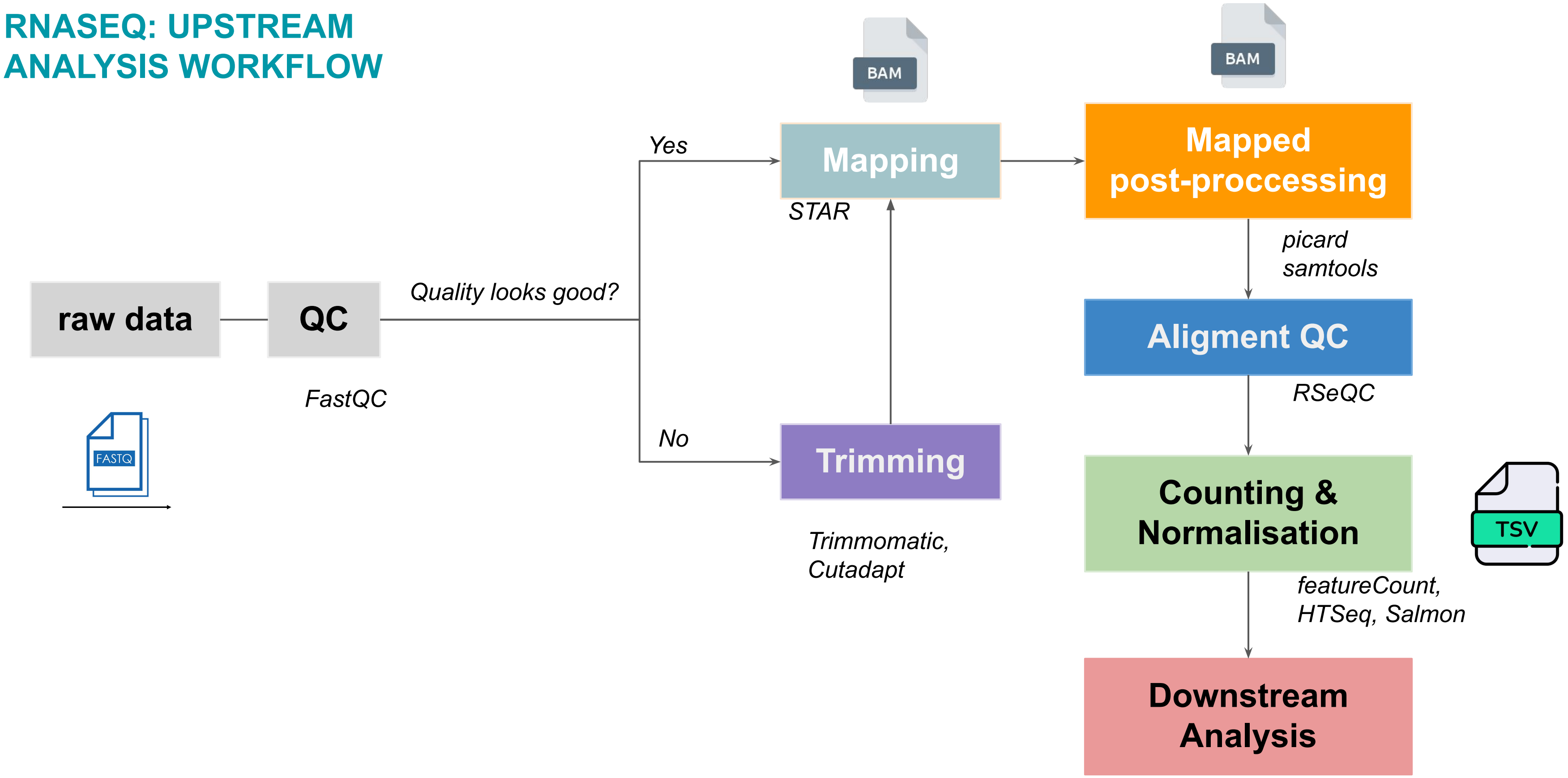
03 ALIGNMENT

04 MAPPED READS POST-PROCESSING

05 ALIGNMENT DATA: QUALITY CONTROL

INTRODUCTION

RNASEQ: UPSTREAM ANALYSIS WORKFLOW



raw data

QC

FastQC

Quality looks good?

Yes

No

Mapping

STAR

Trimming

Trimmomatic, Cutadapt

Mapped post-proccessing

picard samtools

Aligment QC

RSeQC

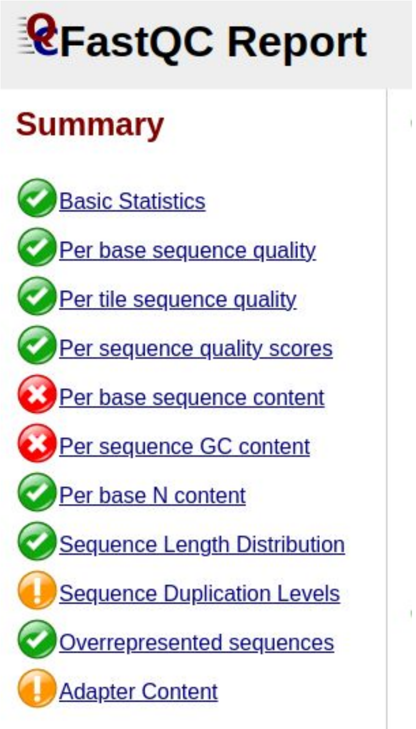
Counting & Normalisation

featureCount, HTSeq, Salmon

Downstream Analysis

SEQUENCE QUALITY CONTROL (FASTQC)

> FASTQC Summary

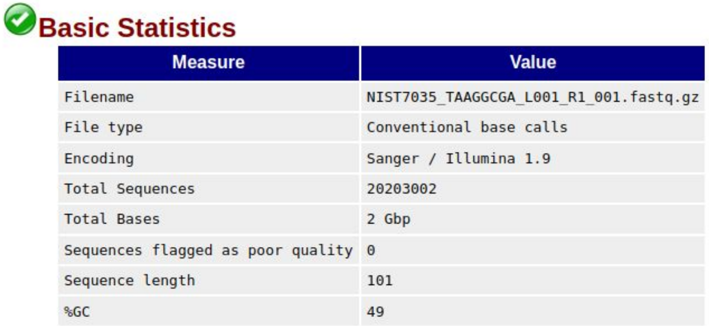


FastQC Report

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ✓ Overrepresented sequences
- ! Adapter Content

"FASTQC is a useful tool to check sequences quality."

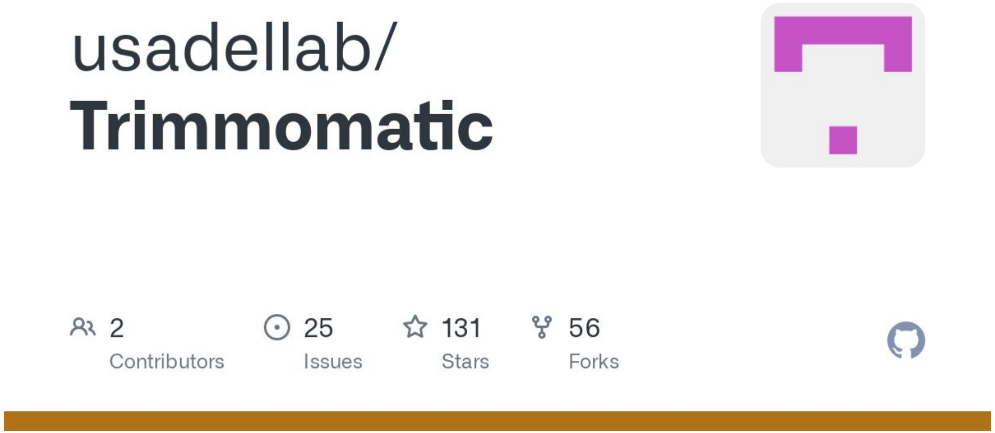


✓ **Basic Statistics**

Measure	Value
Filename	NIST7035_TAAGGCGA_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	20203002
Total Bases	2 Gbp
Sequences flagged as poor quality	0
Sequence length	101
%GC	49

RAW DATA PROCESSING

READ TRIMMING & FILTERING



usadellab/
Trimmomatic

2 Contributors

25 Issues

131 Stars

56 Forks

This program does adaptive quality trimming, head and tail crop, and adaptor removal.

Check QC → Trim → Check QC again.

Trimming:

- Quality trimming
- Adapter trimming.

https://home.cc.umanitoba.ca/~psgendb/doc/TrimmomaticManual.pdf

RAW DATA PROCESSING

Quality problems

- Quality problems typically originate either in the sequencing itself or in the preceding library preparation.

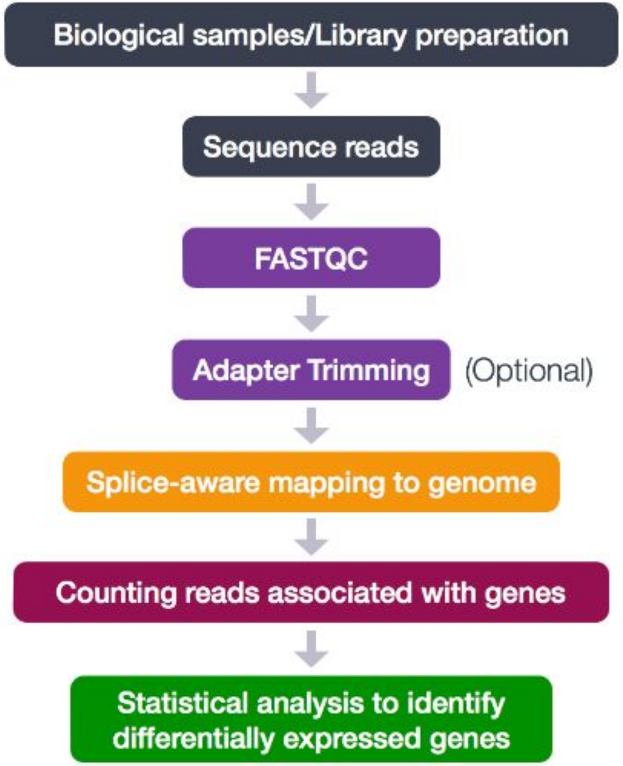
- They include low-confidence bases, sequence-specific bias, 3′/5′ positional bias, polymerase chain reaction (PCR) artifacts, untrimmed adapters, and sequence contamination.

- These problems can seriously affect mapping to reference, assembly, and expression estimates, but luckily many of them can be corrected for by filtering, trimming, error correction, or bias correction.

- Some problems cannot be corrected for, but you should at least be aware of them when interpreting results.

ALIGNMENT

RNA-SEQ: ALIGNMENT / MAPPING



Biological samples/Library preparation

Sequence reads

FASTQC

Adapter Trimming (Optional)

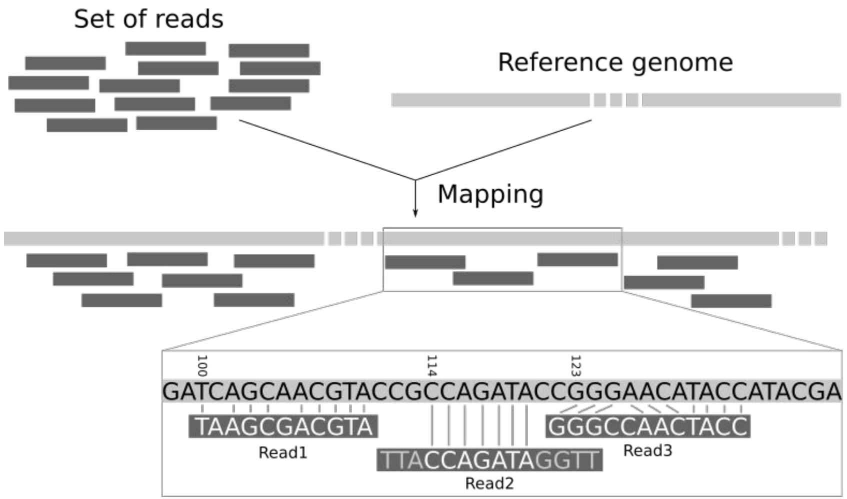
Splice-aware mapping to genome

Counting reads associated with genes

Statistical analysis to identify differentially expressed genes

Workflow for a RNA-seq analysis.

ALIGNMENT

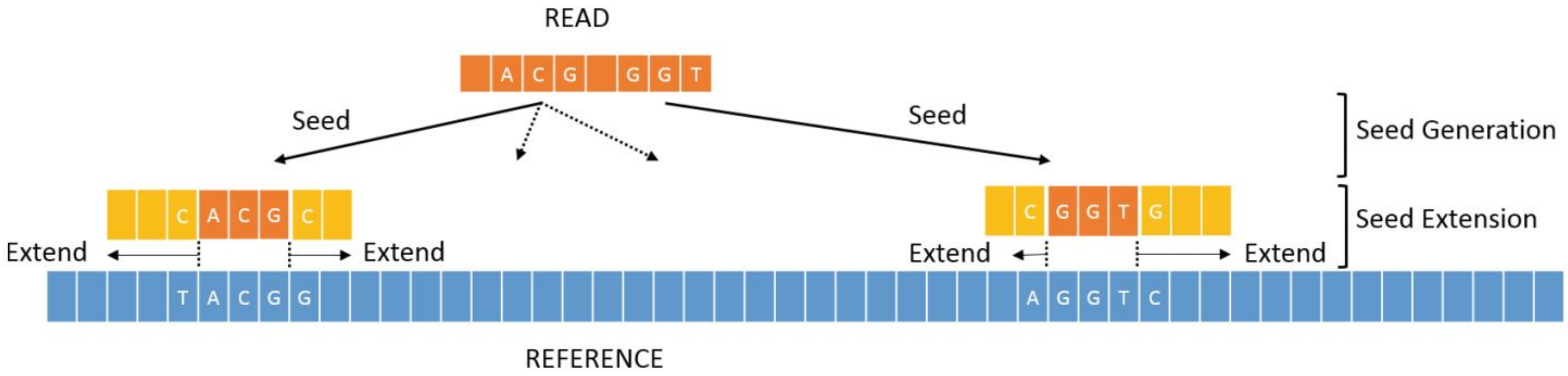


https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

RNA-SEQ: ALIGNMENT / MAPPING

Basic alignment (Contiguous Alignment / Non-spliced Alignment)

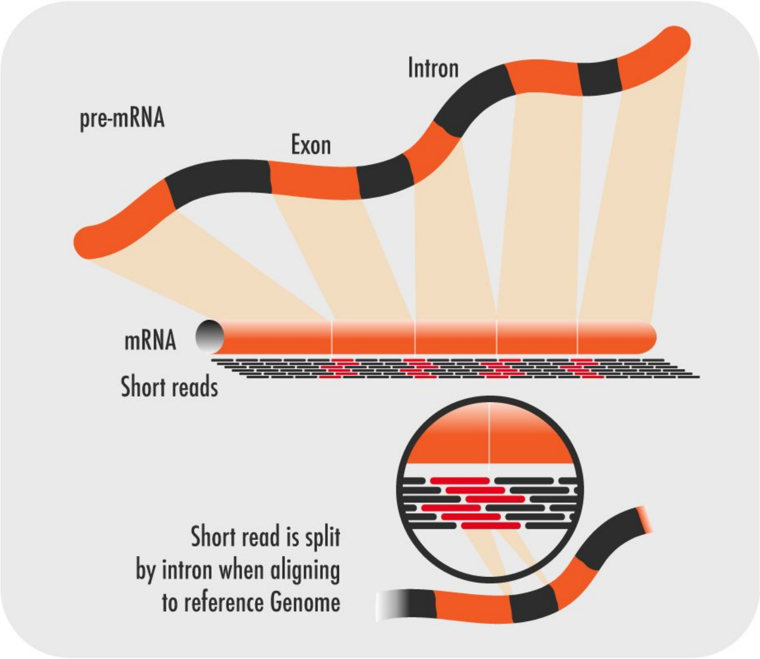
In contiguous alignment, sequences are aligned continuously without any gaps or interruptions.

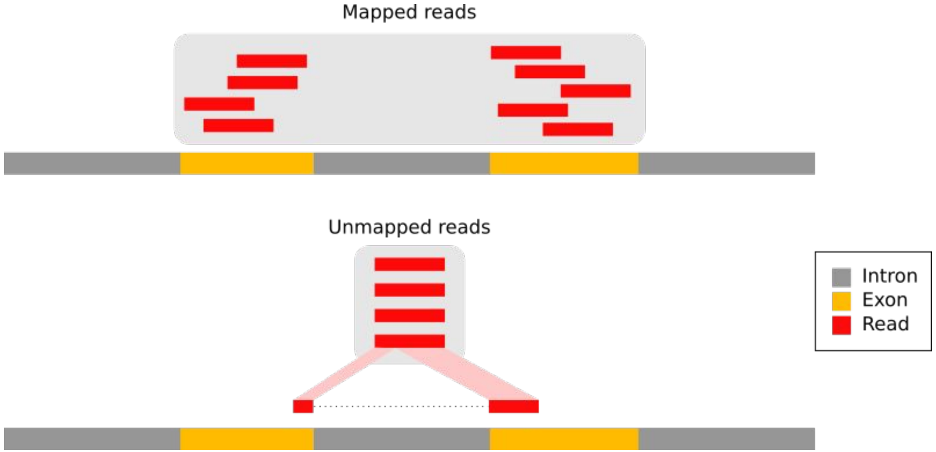


http://ce-publications.et.tudelft.nl/publications/1520_gpuaccelerated_bwamem_genomic_mapping_algorithm_using_ada.pdf

RNA-SEQ: ALIGNMENT / MAPPING

Problem when using basic alignment to map RNA-seq data



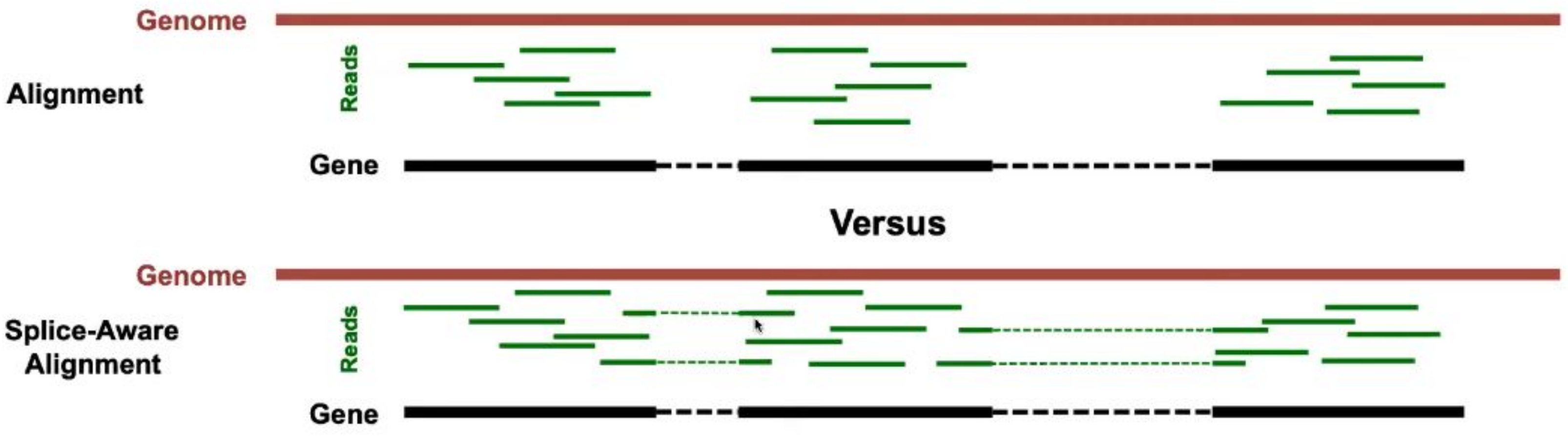


- Unmapped reads due to intron splicing.

https://www.rna-seqblog.com/introduction-and-application-of-transcriptome-sequencing/

RNA-SEQ: ALIGNMENT / MAPPING

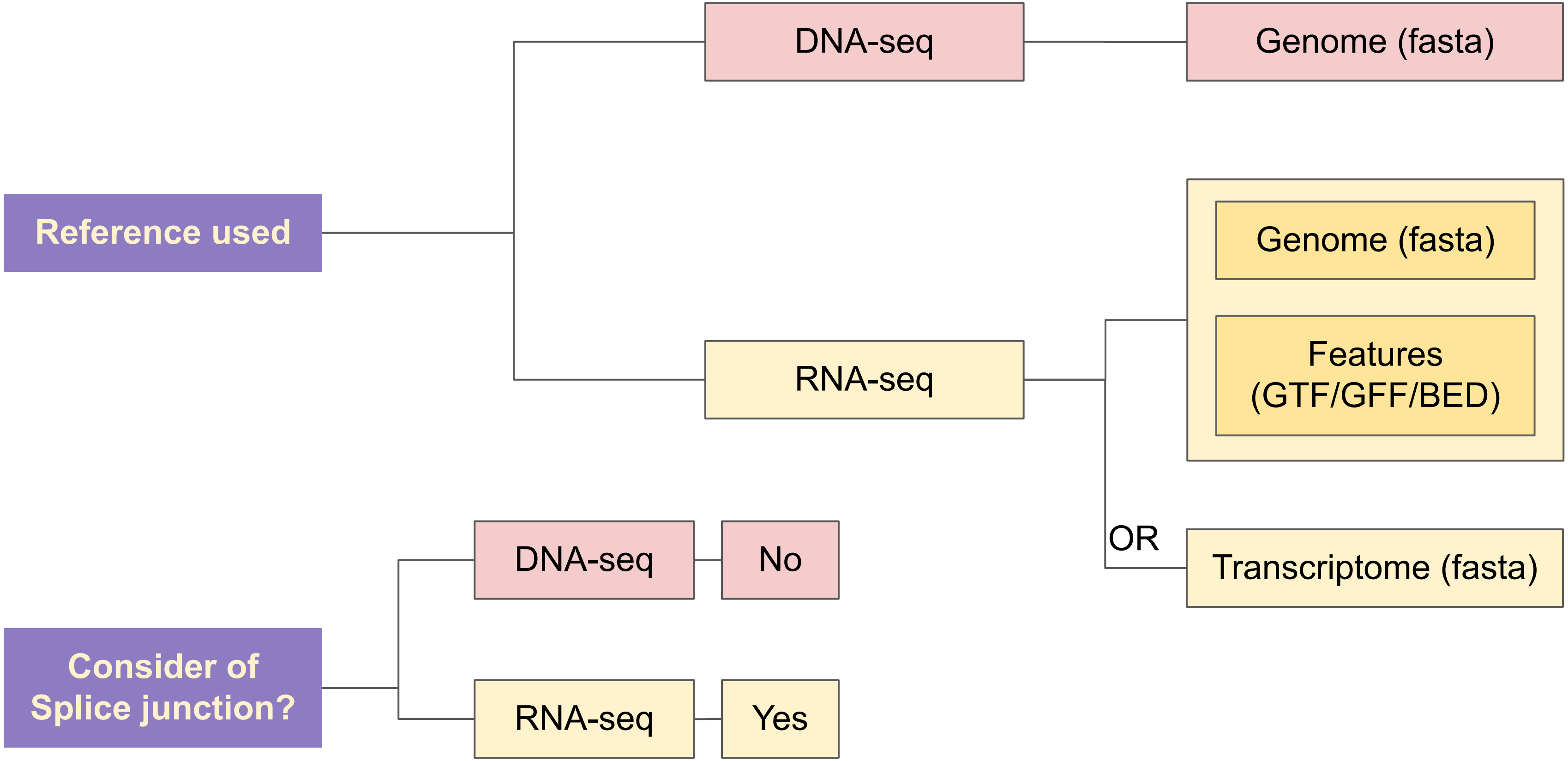
Contiguos Alignment vs Splice-Aware Alignment



- Contiguous Aligners: BWA, Bowtie2,...
- Spliced Aligners: HiSAT2, TopHat, STAR,...

RNA-SEQ: ALIGNMENT / MAPPING

Compare the Alignment of DNA-seq and RNA-seq



Reference used

DNA-seq

Genome (fasta)

RNA-seq

Genome (fasta)

Features (GTF/GFF/BED)

OR

Transcriptome (fasta)

Consider of Splice junction?

DNA-seq

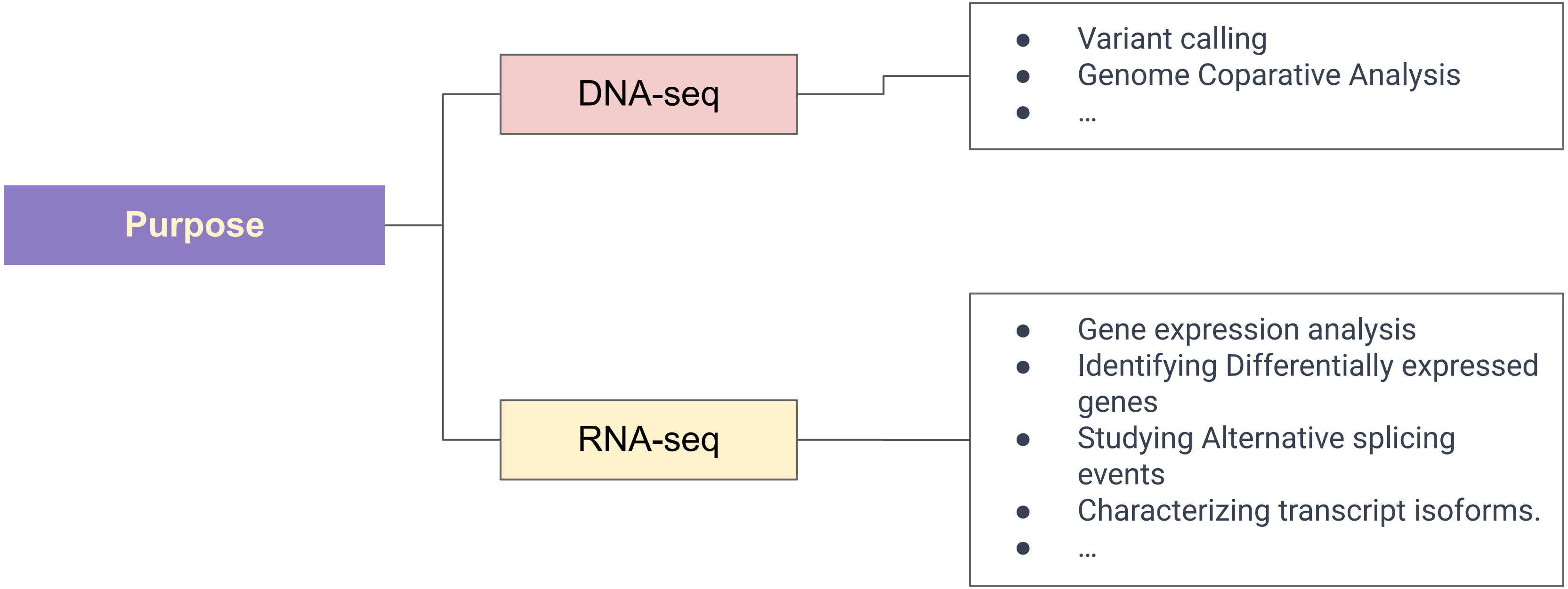
No

RNA-seq

Yes

RNA-SEQ: ALIGNMENT / MAPPING

Compare the Alignment of DNA-seq and RNA-seq



Purpose

DNA-seq

- Variant calling
- Genome Coparative Analysis
- …

RNA-seq

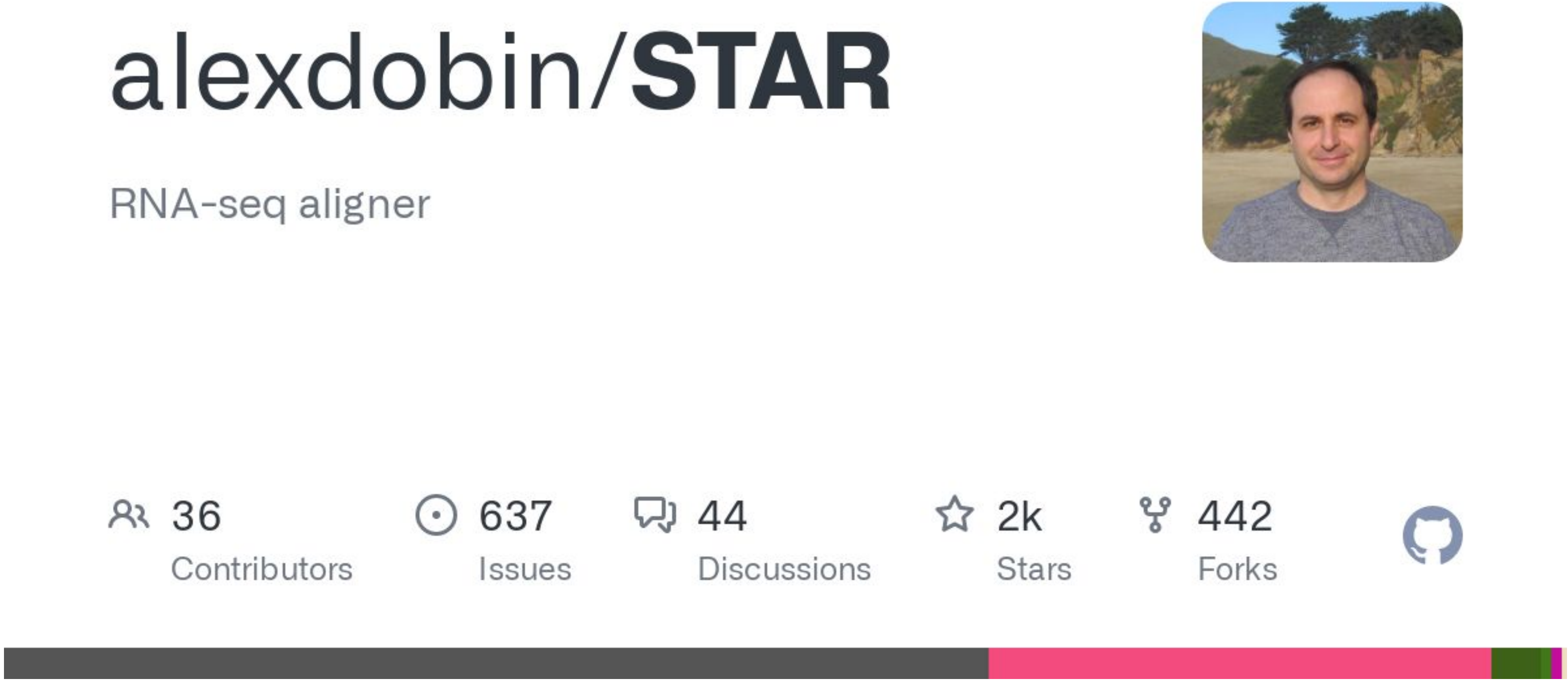
- Gene expression analysis
- Identifying Differentially expressed genes
- Studying Alternative splicing events
- Characterizing transcript isoforms.
- …

RNA-SEQ: ALIGNMENT / MAPPING

Spliced-aware alignment algorithms employ various strategies to handle splice junctions, such as:

- **Split reads:** Allows for precise alignment across the splice junctions.
- **Novel splice junction detection:** Detect previously unknown splicing events, providing insights into alternative splicing patterns and transcriptome complexity.
- **Splice junction annotation:** Aligners may utilize existing splice junction annotations, such as those obtained from databases or previous studies, to guide the alignment process.

RNA-SEQ: ALIGNMENT / MAPPING



alexdobin/**STAR**

RNA-seq aligner

36 Contributors

637 Issues

44 Discussions

2k Stars

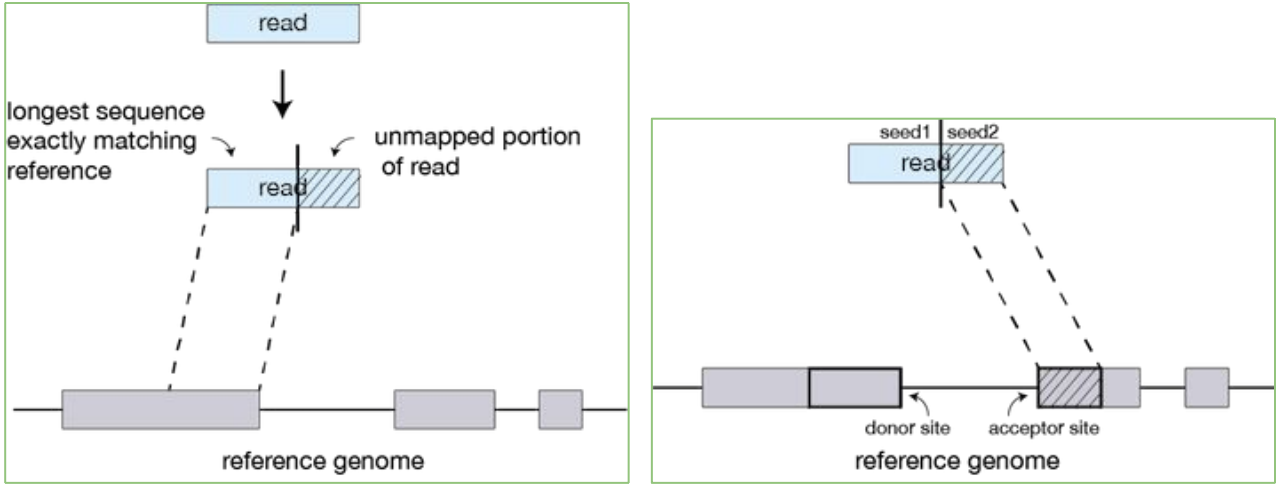
442 Forks

https://github.com/alexdobin/STAR

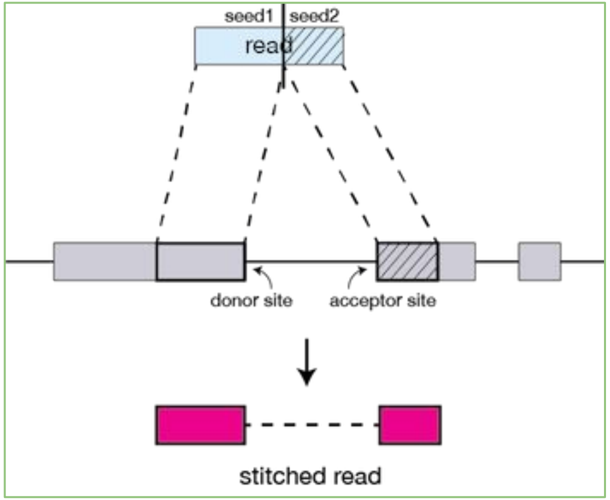
RNA-SEQ: ALIGNMENT / MAPPING

STAR (Spliced Transcripts Alignment to a Reference)

STAR alignment strategy



Seed searching…



Clustering, stitching, and scoring

RNA-SEQ: ALIGNMENT / MAPPING

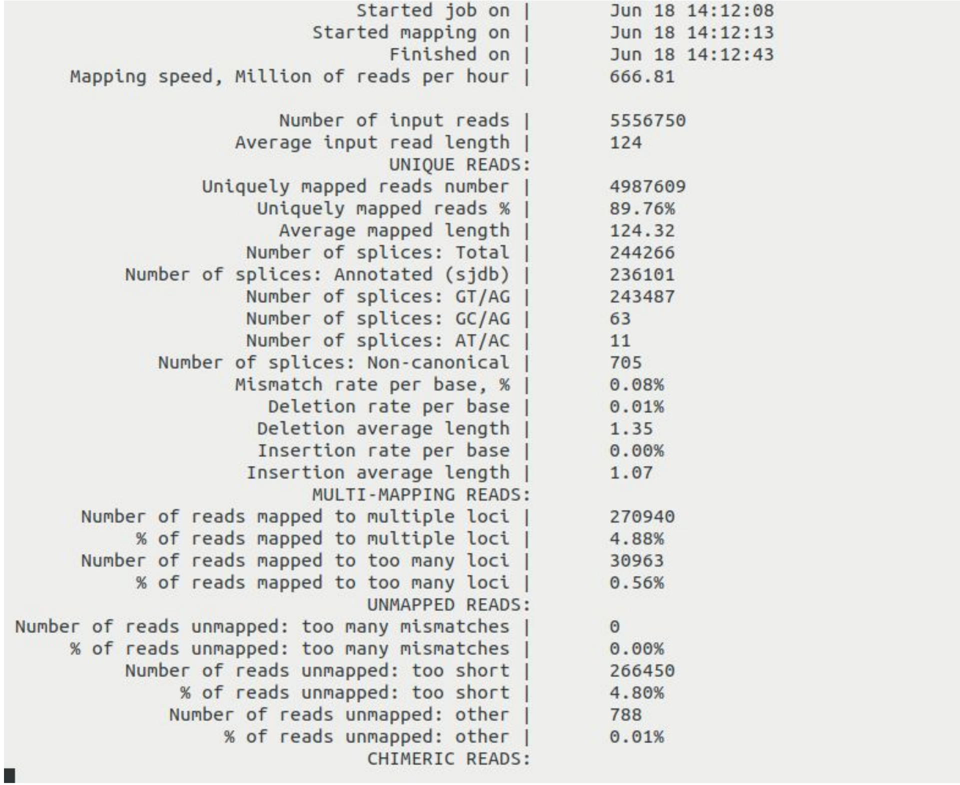
Input → Output

- *R1.fastq*
- *R2.fastq*
- *genome.fa*
- *genome.gtf*

- *SAM/BAM*
- *SJ file*
- *Log file*

RNA-SEQ: ALIGNMENT / MAPPING

Alignment statistics & Utilities for manipulating alignment files.

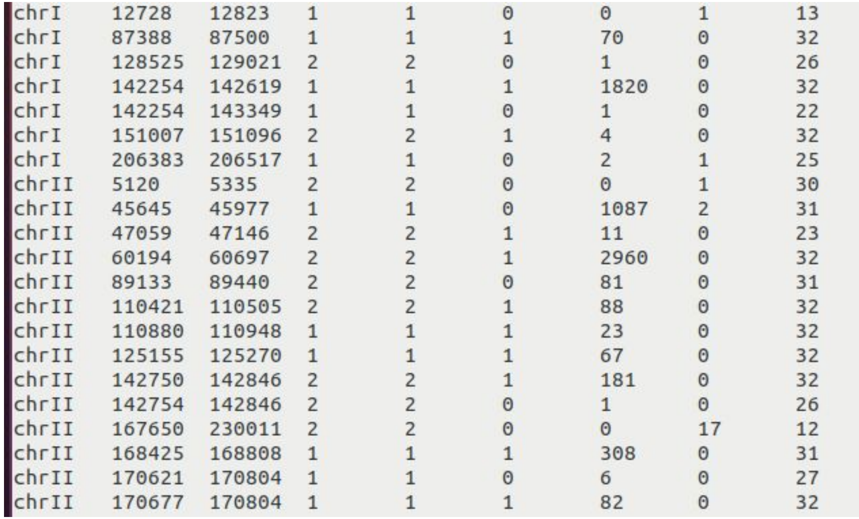


```
                                 Started job on |       Jun 18 14:12:08
                             Started mapping on |       Jun 18 14:12:13
                                    Finished on |       Jun 18 14:12:43
       Mapping speed, Million of reads per hour |       666.81

                          Number of input reads |       5556750
                      Average input read length |       124
                                    UNIQUE READS:
                   Uniquely mapped reads number |       4987609
                        Uniquely mapped reads % |       89.76%
                          Average mapped length |       124.32
                       Number of splices: Total |       244266
            Number of splices: Annotated (sjdb) |       236101
                       Number of splices: GT/AG |       243487
                       Number of splices: GC/AG |       63
                       Number of splices: AT/AC |       11
               Number of splices: Non-canonical |       705
                      Mismatch rate per base, % |       0.08%
                         Deletion rate per base |       0.01%
                        Deletion average length |       1.35
                        Insertion rate per base |       0.00%
                       Insertion average length |       1.07
                             MULTI-MAPPING READS:
        Number of reads mapped to multiple loci |       270940
             % of reads mapped to multiple loci |       4.88%
        Number of reads mapped to too many loci |       30963
             % of reads mapped to too many loci |       0.56%
                                  UNMAPPED READS:
  Number of reads unmapped: too many mismatches |       0
       % of reads unmapped: too many mismatches |       0.00%
            Number of reads unmapped: too short |       266450
                 % of reads unmapped: too short |       4.80%
                Number of reads unmapped: other |       788
                     % of reads unmapped: other |       0.01%
                                  CHIMERIC READS:
```

Alignment statistics (log file)

SJ.out.tab

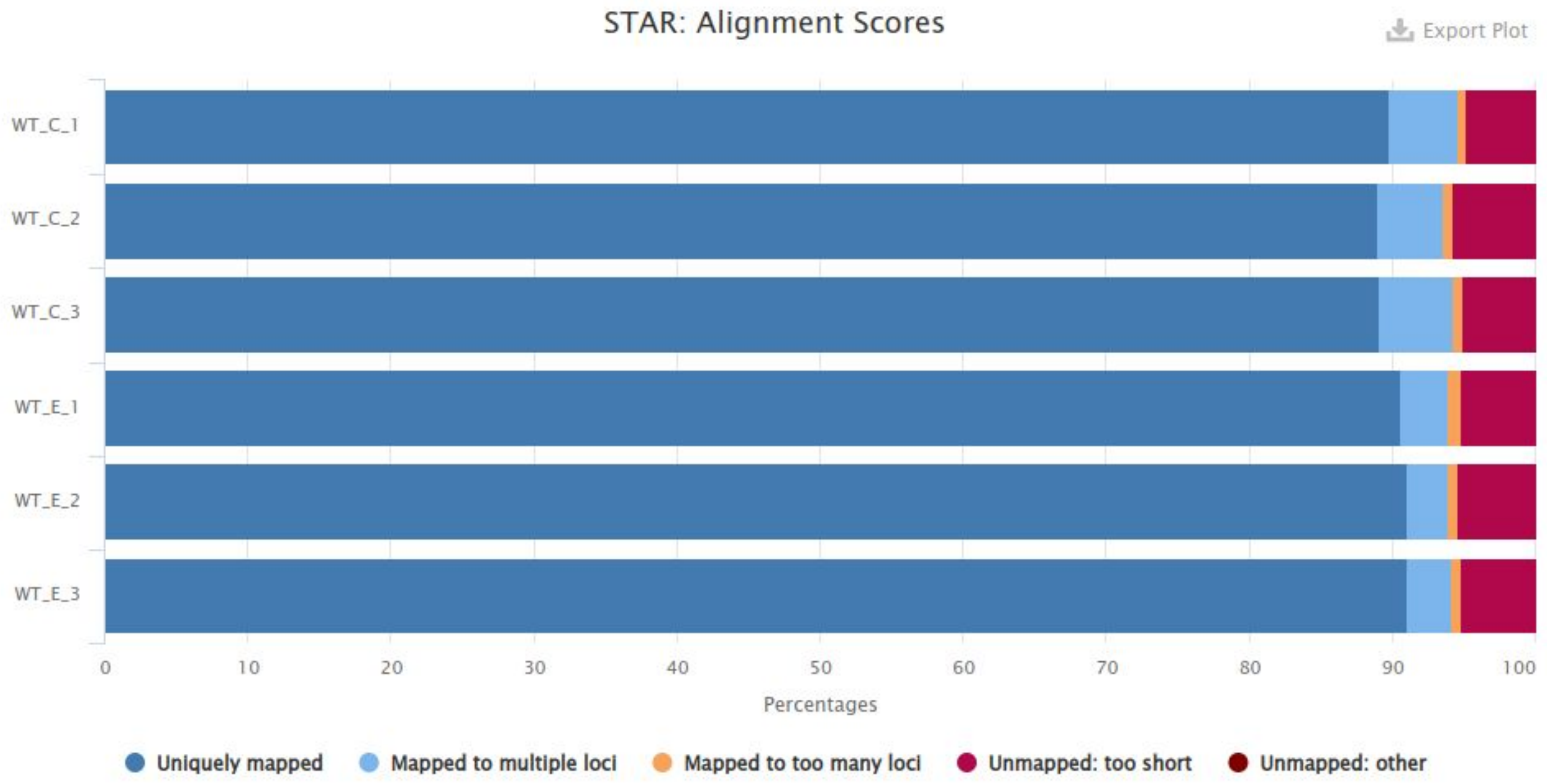


```
chrI    12728   12823   1       1       0       0       1       13
chrI    87388   87500   1       1       1       70      0       32
chrI    128525  129021  2       2       0       1       0       26
chrI    142254  142619  1       1       1       1820    0       32
chrI    142254  143349  1       1       0       1       0       22
chrI    151007  151096  2       2       1       4       0       32
chrI    206383  206517  1       1       0       2       1       25
chrII   5120    5335    2       2       0       0       1       30
chrII   45645   45977   1       1       0       1087    2       31
chrII   47059   47146   2       2       1       11      0       23
chrII   60194   60697   2       2       1       2960    0       32
chrII   89133   89440   2       2       0       81      0       31
chrII   110421  110505  2       2       1       88      0       32
chrII   110880  110948  1       1       1       23      0       32
chrII   125155  125270  1       1       1       67      0       32
chrII   142750  142846  2       2       1       181     0       32
chrII   142754  142846  2       2       0       1       0       26
chrII   167650  230011  2       2       0       0       17      12
chrII   168425  168808  1       1       1       308     0       31
chrII   170621  170804  1       1       0       6       0       27
chrII   170677  170804  1       1       1       82      0       32
```

The SJ.out.tab contains filtered splice junctions detected in the mapping

RNA-SEQ: ALIGNMENT / MAPPING

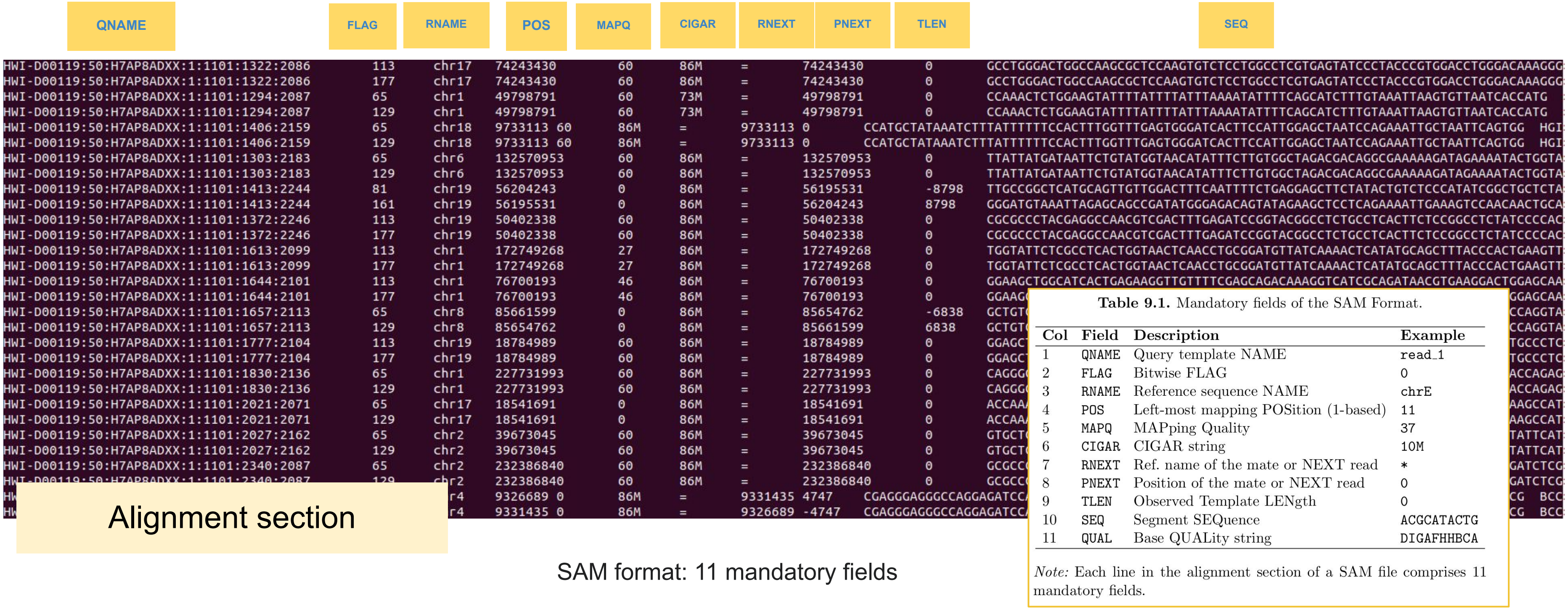
Alignment statistics & Utilities for manipulating alignment files.



RNA-SEQ: ALIGNMENT / MAPPING

Alignment statistics & Utilities for manipulating alignment files.

SAM format



```
HWI-D00119:50:H7AP8ADXX:1:1101:1322:2086      113    chr17   74243430        60      86M     =       74243430        0       GCCTGGGACTGGCCAAGCGCTCCAAGTGTCTCCTGGCCTCGTGAGTATCCCTACCCGTGGACCTGGGACAAAGGG
HWI-D00119:50:H7AP8ADXX:1:1101:1322:2086      177    chr17   74243430        60      86M     =       74243430        0       GCCTGGGACTGGCCAAGCGCTCCAAGTGTCTCCTGGCCTCGTGAGTATCCCTACCCGTGGACCTGGGACAAAGGG
HWI-D00119:50:H7AP8ADXX:1:1101:1294:2087      65     chr1    49798791        60      73M     =       49798791        0       CCAAACTCTGGAAGTATTTTATTTTATTTAAAATATTTTCAGCATCTTTGTAAATTAAGTGTTAATCACCATG
HWI-D00119:50:H7AP8ADXX:1:1101:1294:2087      129    chr1    49798791        60      73M     =       49798791        0       CCAAACTCTGGAAGTATTTTATTTTATTTAAAATATTTTCAGCATCTTTGTAAATTAAGTGTTAATCACCATG
HWI-D00119:50:H7AP8ADXX:1:1101:1406:2159      65     chr18   9733113 60      86M     =       9733113 0       CCATGCTATAAATCTTTATTTTTTCCACTTTGGTTTGAGTGGGATCACTTCCATTGGAGCTAATCCAGAAATTGCTAATTCAGTGG  HGI
HWI-D00119:50:H7AP8ADXX:1:1101:1406:2159      129    chr18   9733113 60      86M     =       9733113 0       CCATGCTATAAATCTTTATTTTTTCCACTTTGGTTTGAGTGGGATCACTTCCATTGGAGCTAATCCAGAAATTGCTAATTCAGTGG  HGI
HWI-D00119:50:H7AP8ADXX:1:1101:1303:2183      65     chr6    132570953       60      86M     =       132570953       0       TTATTATGATAATTCTGTATGGTAACATATTTCTTGTGGCTAGACGACAGGCGAAAAAGATAGAAAATACTGGTA
HWI-D00119:50:H7AP8ADXX:1:1101:1303:2183      129    chr6    132570953       60      86M     =       132570953       0       TTATTATGATAATTCTGTATGGTAACATATTTCTTGTGGCTAGACGACAGGCGAAAAAGATAGAAAATACTGGTA
HWI-D00119:50:H7AP8ADXX:1:1101:1413:2244      81     chr19   56204243        0       86M     =       56195531        -8798   TTGCCGGCTCATGCAGTTGTTGGACTTTCAATTTTCTGAGGAGCTTCTATACTGTCTCCCATATCGGCTGCTCTA
HWI-D00119:50:H7AP8ADXX:1:1101:1413:2244      161    chr19   56195531        0       86M     =       56204243        8798    GGGATGTAAATTAGAGCAGCCGATATGGGAGACAGTATAGAAGCTCCTCAGAAAATTGAAAGTCCAACAACTGCA
HWI-D00119:50:H7AP8ADXX:1:1101:1372:2246      113    chr19   50402338        60      86M     =       50402338        0       CGCGCCCTACGAGGCCAACGTCGACTTTGAGATCCGGTACGGCCTCTGCCTCACTTCTCCGGCCTCTATCCCCAC
HWI-D00119:50:H7AP8ADXX:1:1101:1372:2246      177    chr19   50402338        60      86M     =       50402338        0       CGCGCCCTACGAGGCCAACGTCGACTTTGAGATCCGGTACGGCCTCTGCCTCACTTCTCCGGCCTCTATCCCCAC
HWI-D00119:50:H7AP8ADXX:1:1101:1613:2099      113    chr1    172749268       27      86M     =       172749268       0       TGGTATTCTCGCCTCACTGGTAACTCAACCTGCGGATGTTATCAAAACTCATATGCAGCTTTACCCACTGAAGTT
HWI-D00119:50:H7AP8ADXX:1:1101:1613:2099      177    chr1    172749268       27      86M     =       172749268       0       TGGTATTCTCGCCTCACTGGTAACTCAACCTGCGGATGTTATCAAAACTCATATGCAGCTTTACCCACTGAAGTT
HWI-D00119:50:H7AP8ADXX:1:1101:1644:2101      113    chr1    76700193        46      86M     =       76700193        0       GGAAGCTGGCATCACTGAGAAGGTTGTTTTCGAGCAGACAAAGGTCATCGCAGATAACGTGAAGGACTGGAGCAA
HWI-D00119:50:H7AP8ADXX:1:1101:1644:2101      177    chr1    76700193        46      86M     =       76700193        0       GGAAGC                                                              GGAGCAA
HWI-D00119:50:H7AP8ADXX:1:1101:1657:2113      65     chr8    85661599        0       86M     =       85654762        -6838   GCTGTC                                                              CCAGGTA
HWI-D00119:50:H7AP8ADXX:1:1101:1657:2113      129    chr8    85654762        0       86M     =       85661599        6838    GCTGTC                                                              CCAGGTA
HWI-D00119:50:H7AP8ADXX:1:1101:1777:2104      113    chr19   18784989        60      86M     =       18784989        0       GGAGCT                                                              TGCCCTC
HWI-D00119:50:H7AP8ADXX:1:1101:1777:2104      177    chr19   18784989        60      86M     =       18784989        0       GGAGCT                                                              TGCCCTC
HWI-D00119:50:H7AP8ADXX:1:1101:1830:2136      65     chr1    227731993       60      86M     =       227731993       0       CAGGGC                                                              ACCAGAG
HWI-D00119:50:H7AP8ADXX:1:1101:1830:2136      129    chr1    227731993       60      86M     =       227731993       0       CAGGGC                                                              ACCAGAG
HWI-D00119:50:H7AP8ADXX:1:1101:2021:2071      65     chr17   18541691        0       86M     =       18541691        0       ACCAAA                                                              AAGCCAT
HWI-D00119:50:H7AP8ADXX:1:1101:2021:2071      129    chr17   18541691        0       86M     =       18541691        0       ACCAAA                                                              AAGCCAT
HWI-D00119:50:H7AP8ADXX:1:1101:2027:2162      65     chr2    39673045        60      86M     =       39673045        0       GTGCTC                                                              TATTCAT
HWI-D00119:50:H7AP8ADXX:1:1101:2027:2162      129    chr2    39673045        60      86M     =       39673045        0       GTGCTC                                                              TATTCAT
HWI-D00119:50:H7AP8ADXX:1:1101:2340:2087      65     chr2    232386840       60      86M     =       232386840       0       GCGCCC                                                              GATCTCG
HWI-D00119:50:H7AP8ADXX:1:1101:2340:2087      129    chr2    232386840       60      86M     =       232386840       0       GCGCCC                                                              GATCTCG
HW                                                  r4      9326689 0       86M     =       9331435 4747    CGAGGGAGGGCCAGGAGATCCA                                              CG   BCC
HW                                                  r4      9331435 0       86M     =       9326689 -4747   CGAGGGAGGGCCAGGAGATCCA                                              CG   BCC
```

Alignment section

SAM format: 11 mandatory fields

Table 9.1. Mandatory fields of the SAM Format.

Col	Field	Description	Example
1	QNAME	Query template NAME	read_1
2	FLAG	Bitwise FLAG	0
3	RNAME	Reference sequence NAME	chrE
4	POS	Left-most mapping POSition (1-based)	11
5	MAPQ	MAPping Quality	37
6	CIGAR	CIGAR string	10M
7	RNEXT	Ref. name of the mate or NEXT read	*
8	PNEXT	Position of the mate or NEXT read	0
9	TLEN	Observed Template LENgth	0
10	SEQ	Segment SEQuence	ACGCATACTG
11	QUAL	Base QUALity string	DIGAFHHBCA

Note: Each line in the alignment section of a SAM file comprises 11 mandatory fields.

RNA-SEQ: ALIGNMENT / MAPPING

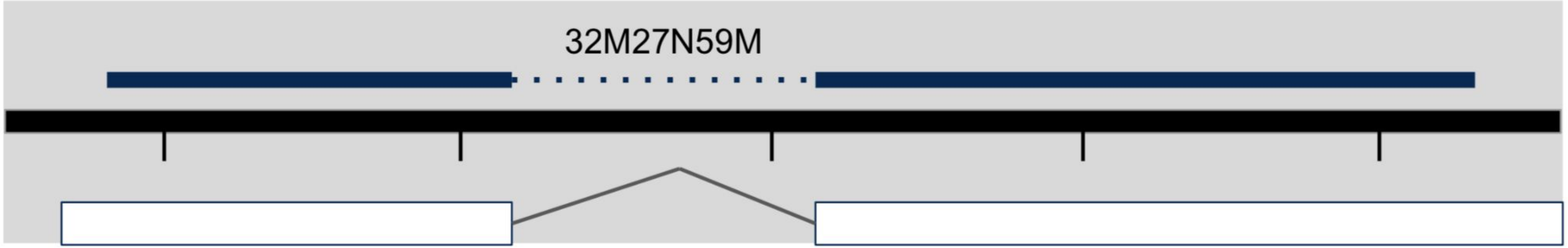
Alignment statistics & Utilities for manipulating alignment files.

CIGAR string with “N”

The "N" in the CIGAR string represents a stretch of skipped reference bases (also known as introns or gaps) in a sequence alignment.

It indicates that the read aligns to the reference genome, but there is a region of the reference sequence that is not covered by the read.

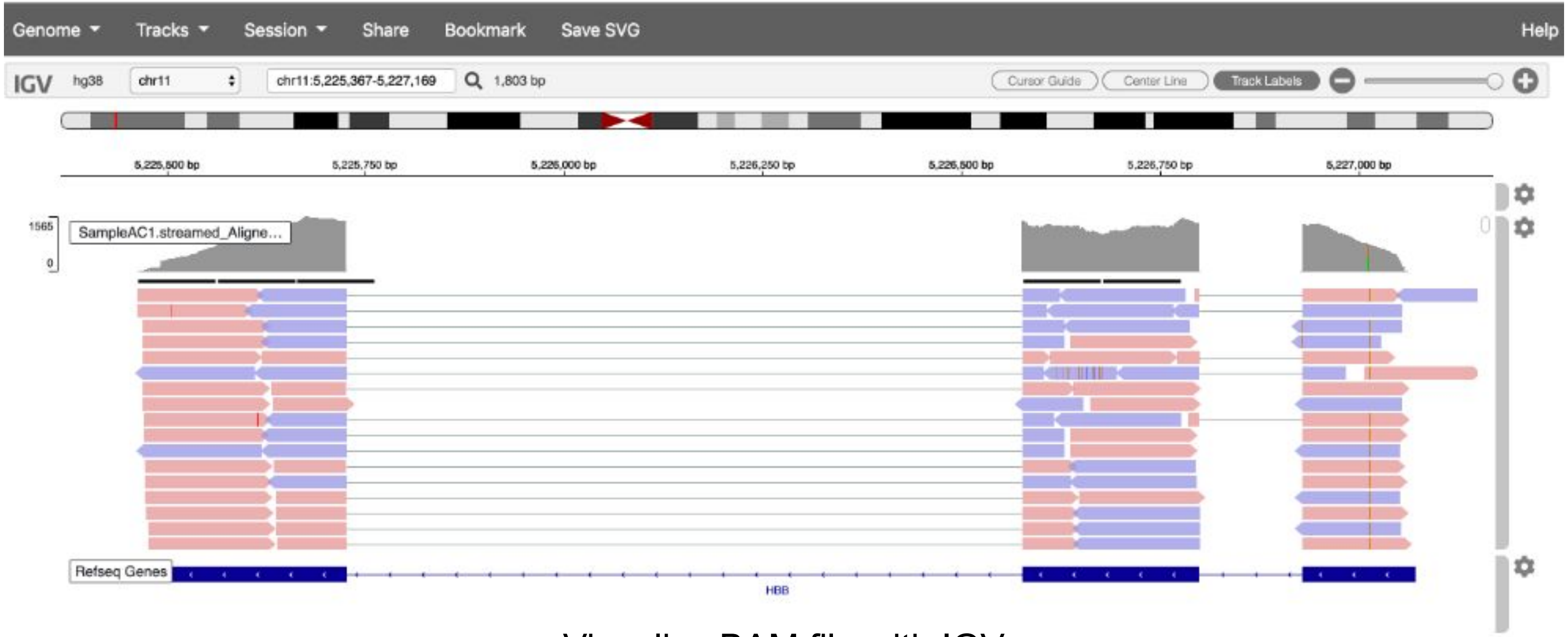
Splicing:



https://ucdavis-bioinformatics-training.github.io/2019_March_UCSF_mRNAseq_Workshop/data_reduction/alignment.html

RNA-SEQ: ALIGNMENT / MAPPING

Alignment statistics & Utilities for manipulating alignment files.



Visualize BAM file with IGV

ALIGNMENT QUALITY CONTROL

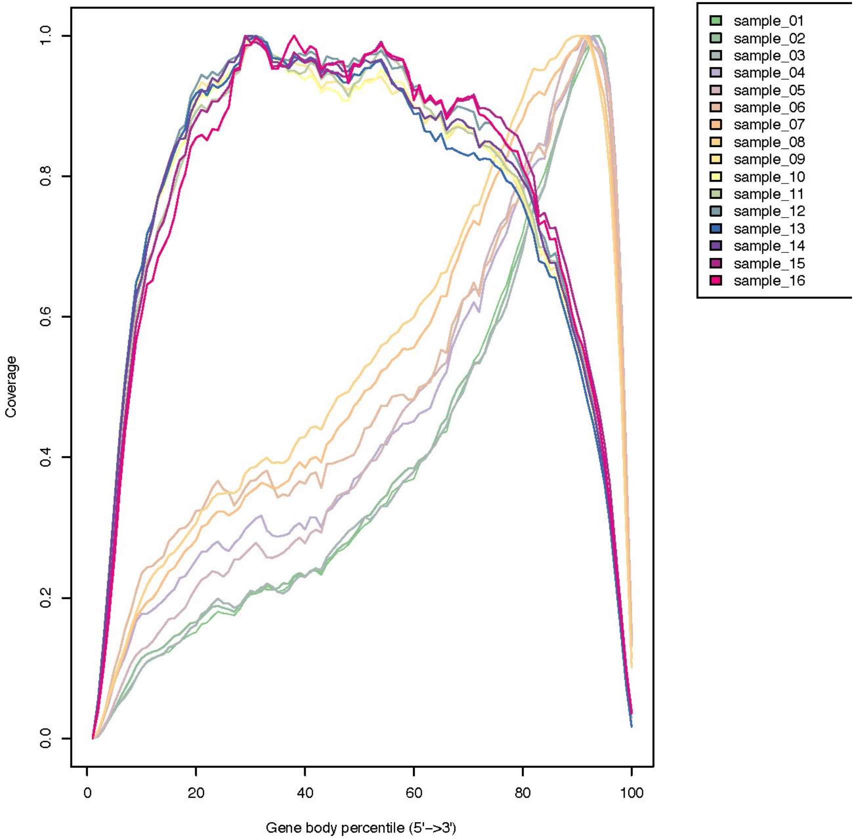
RNA-SEQ: ALIGNMENT QUALITY CONTROL

Alignment quality metrics:

- Coverage uniformity along transcripts
- Saturation of sequencing depth
- Ribosomal RNA content (rRNA)
- Read distribution between exons, introns & intergenic regions.
- …

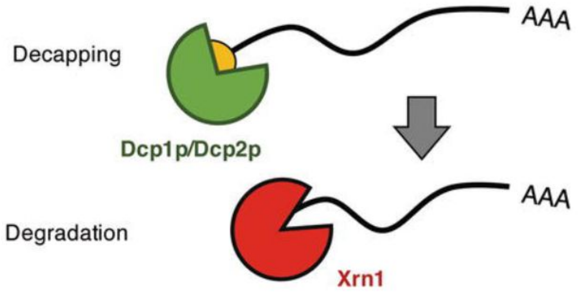
RNA-SEQ: ALIGNMENT QUALITY CONTROL

RSeQC: Genebody Coverage



→ Used to assess the sequencing depth and coverage across the entire length of genes

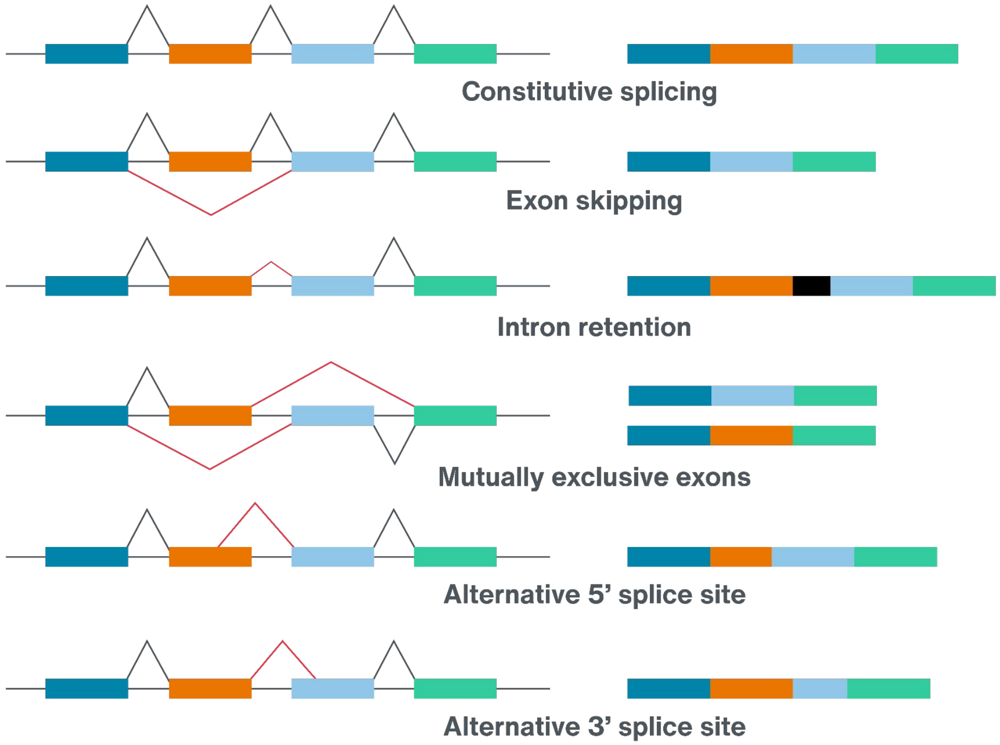
- Gene expression quantification
- Transcript isoform analysis
- Detection of gene expression biases
- Assessing RNA integrity and sample quality



Which samples may have been degraded?

RNA-SEQ: ALIGNMENT QUALITY CONTROL

RSeQC: Junction Annotation



Splice variation

The junction-annotation command:

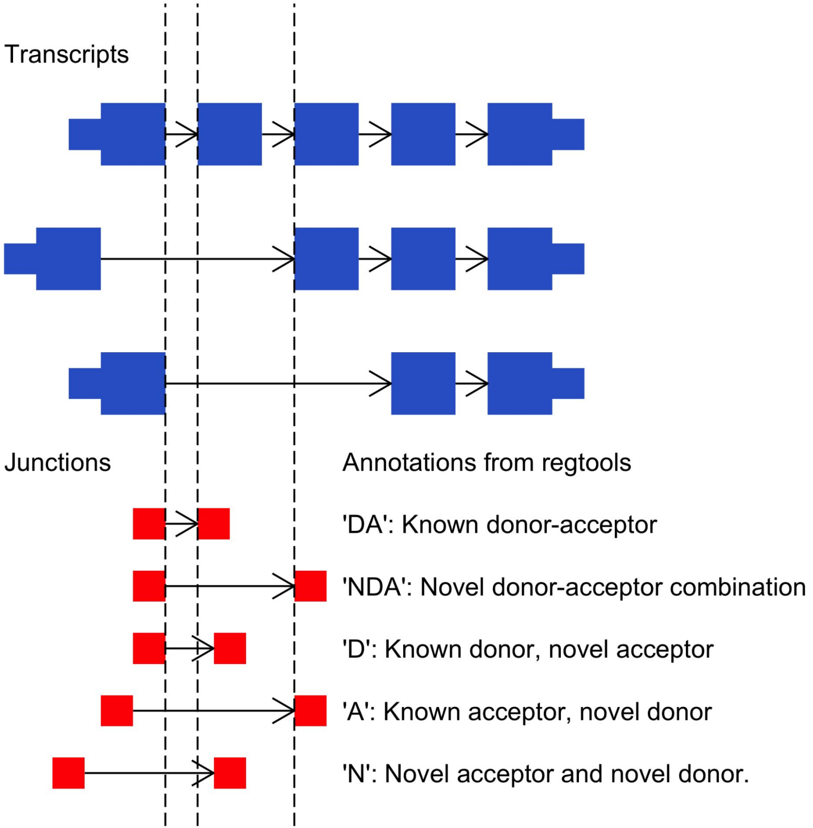
1. Search an RNA-Seq bam file for splice junctions.

2. Compare them to a gene model.

3. Output whether the found junctions are novel, partially novel, or already annotated in a gene model.

RNA-SEQ: ALIGNMENT QUALITY CONTROL

RSeQC: Junction Annotation



Transcripts

Junctions

Annotations from regtools

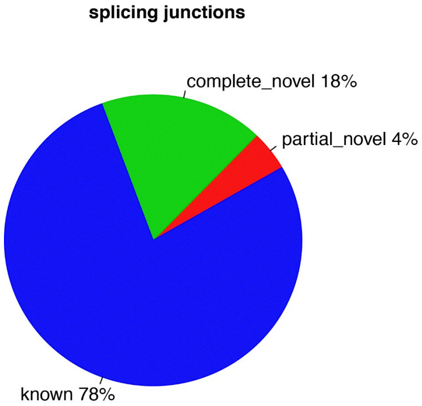
'DA': Known donor-acceptor

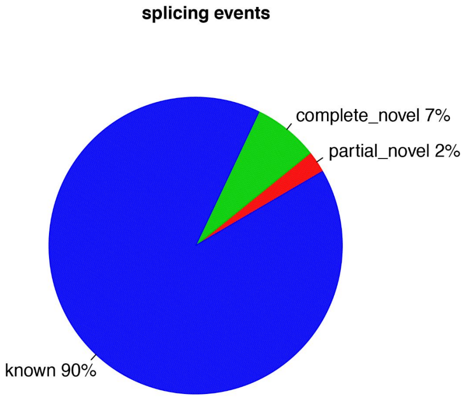
'NDA': Novel donor-acceptor combination

'D': Known donor, novel acceptor

'A': Known acceptor, novel donor

'N': Novel acceptor and novel donor.

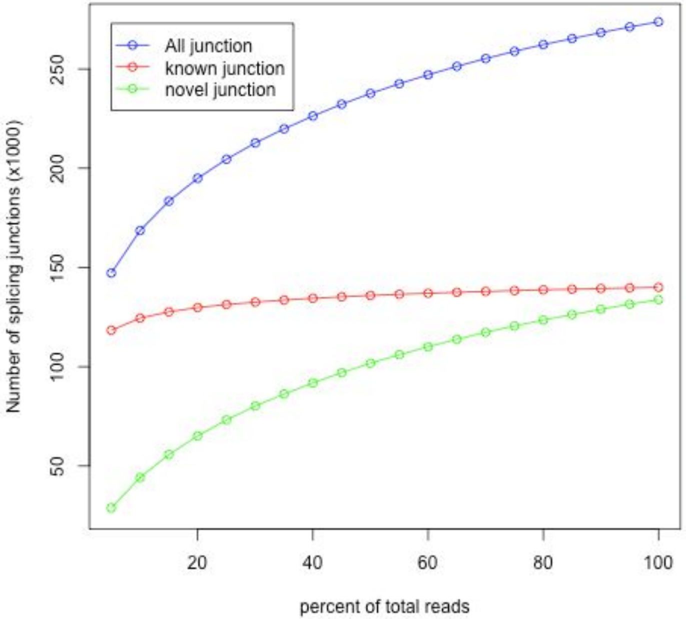




https://regtools.readthedocs.io/en/latest/commands/junctions-annotate/

RNA-SEQ: ALIGNMENT QUALITY CONTROL

RSeQC: Junction Saturation



A sample that reaches a plateau before getting to 100% data indicates that all junctions in the library have been detected, and that further sequencing will not yield more observations.

Junction Saturation Analysis

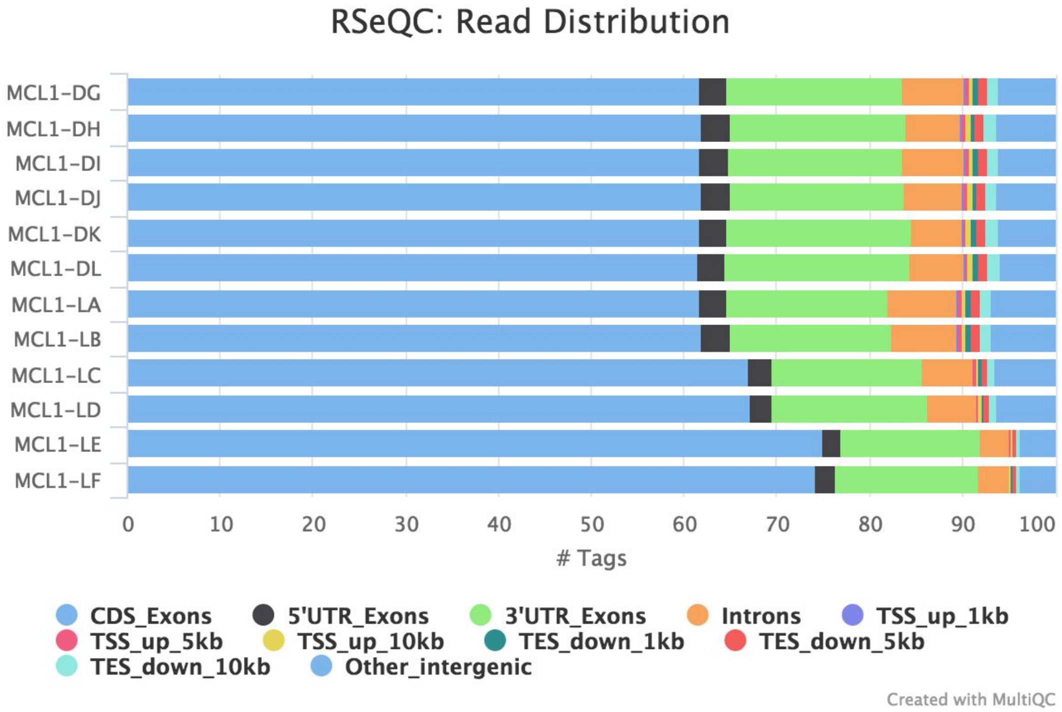
- Evaluates the depth of sequencing coverage at splice junctions.
- It helps determine if sufficient sequencing depth has been achieved to capture the full diversity of splice junctions.

→ Guides decisions on whether additional sequencing is needed to achieve more comprehensive coverage.

→ Ensures confidence in downstream analyses (alternative splicing analysis, isoform discovery).

RNA-SEQ: ALIGNMENT QUALITY CONTROL

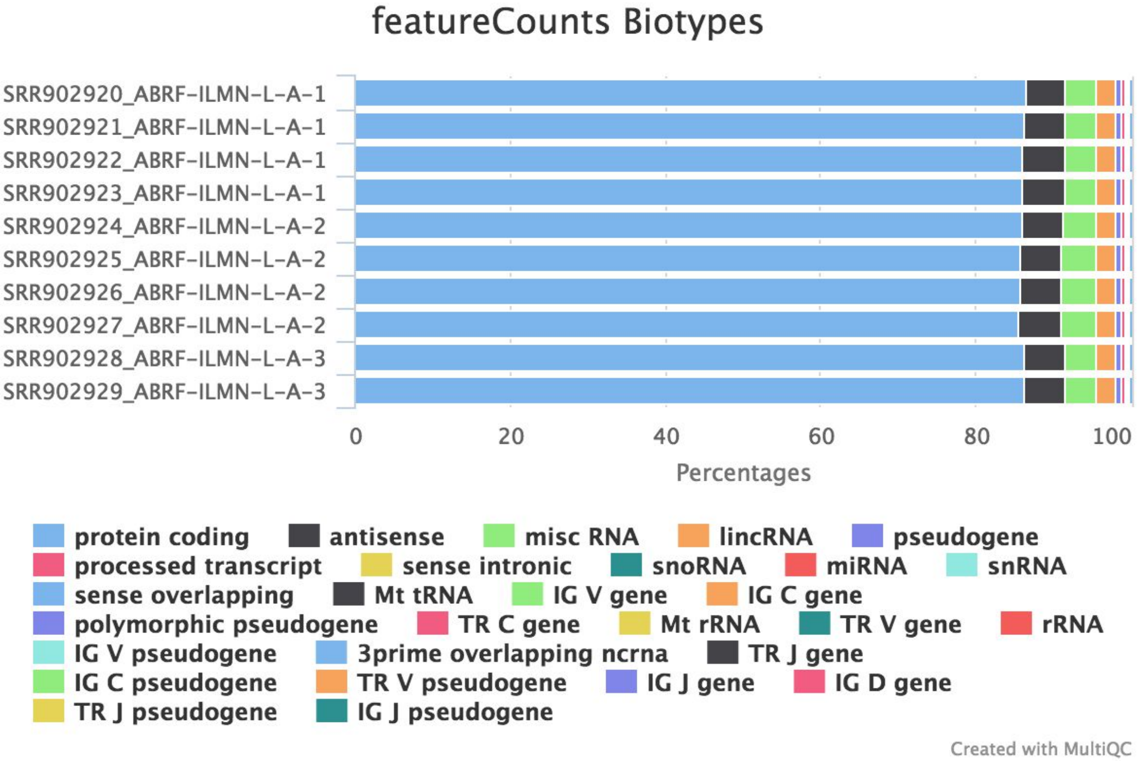
RSeQC: Read Distribution



Calculate how mapped reads were distributed over genome feature (like CDS exon, 5'UTR exon, 3' UTR exon, Intron, Intergenic regions).

RNA-SEQ: ALIGNMENT QUALITY CONTROL

Biotypes Count

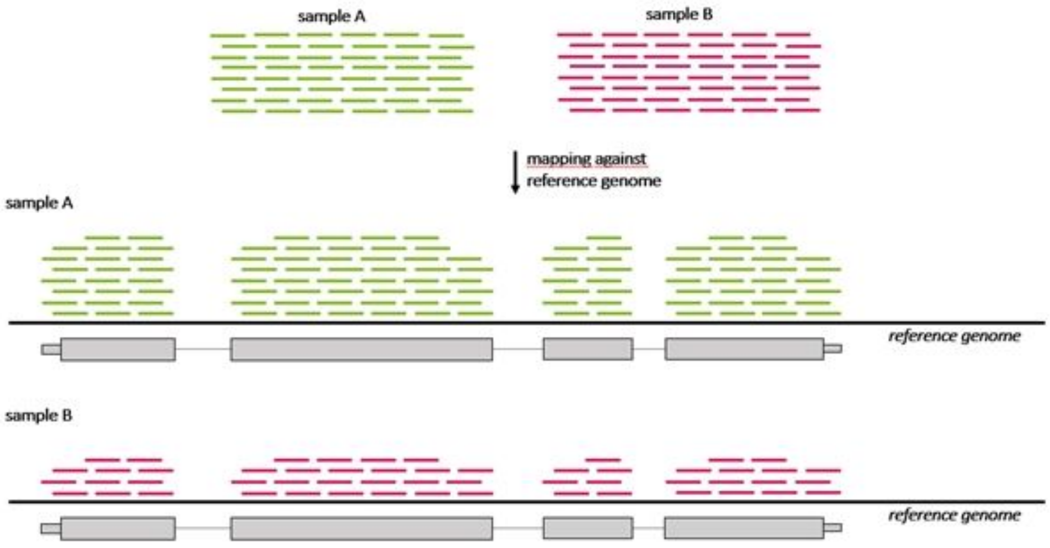


- A good RNAseq sample should have a large portion of the reads coming from protein coding genes.
- This plot can help you spot problems with your library such as incomplete rRNA depletion.

QUANTIFICATION

RNA-SEQ: QUANTIFICATION

Quantification - Read Count



Count how many reads have mapped to each gene.

→Using the **featureCounts** tool to get the gene counts

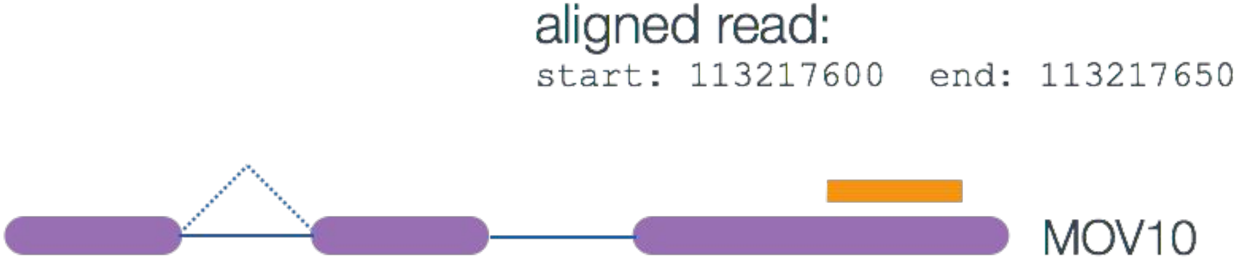
Input: BAM + GTF

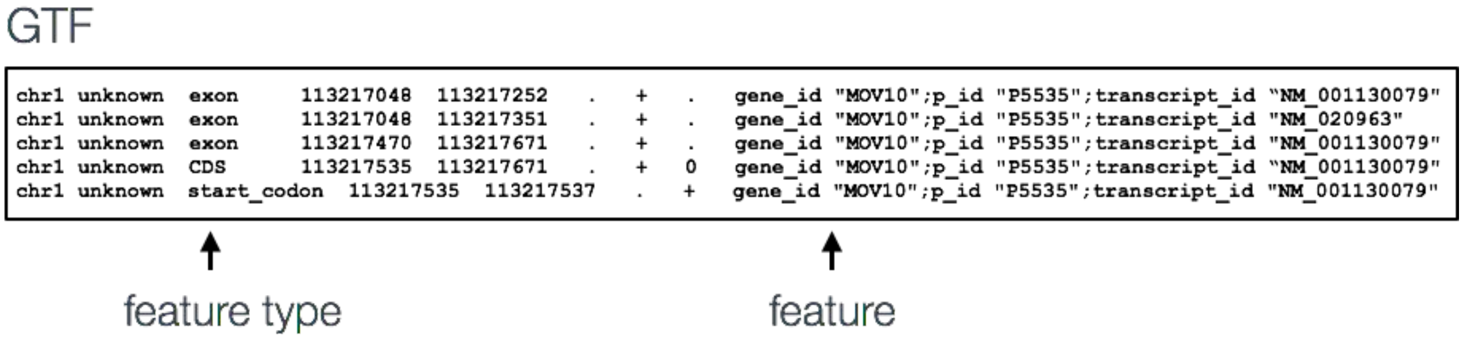
Output: Number of reads (counts) associated with each feature of interest (genes, exons, transcript, etc.).

RNA-SEQ: QUANTIFICATION

Counting reads with featureCounts

- Accurate, fast and is relatively easy to use
- Counts reads that map to a single location (uniquely mapping) and follows the scheme in the figure below for assigning reads to a gene/exon.

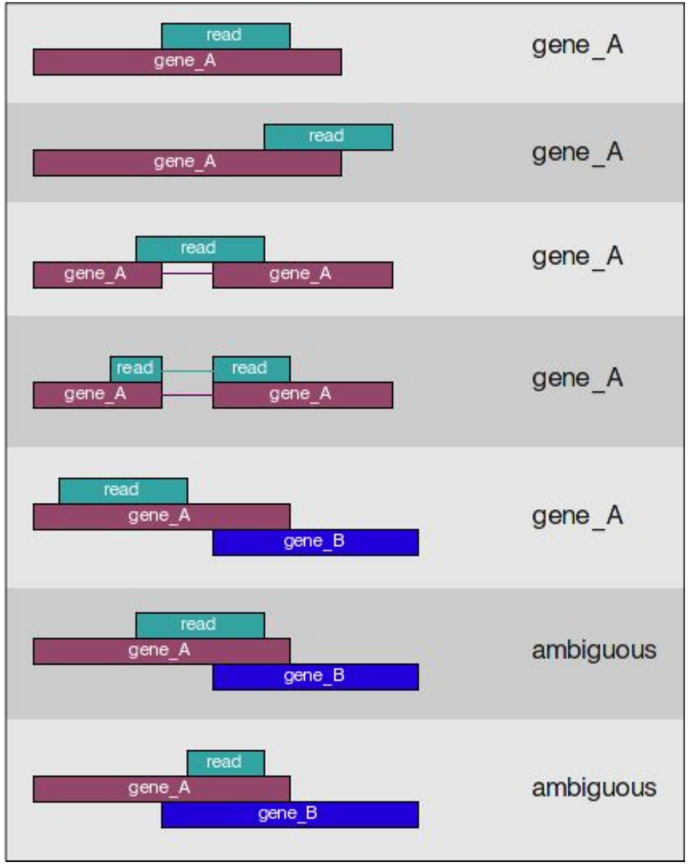



GTF
feature type
feature

```
chr1 unknown  exon        113217048  113217252  .  +  .  gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  exon        113217048  113217351  .  +  .  gene_id "MOV10";p_id "P5535";transcript_id "NM_020963"
chr1 unknown  exon        113217470  113217671  .  +  .  gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  CDS         113217535  113217671  .  +  0  gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  start_codon  113217535  113217537  .  +  gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
```

RNA-SEQ: QUANTIFICATION

Counting reads using featureCounts

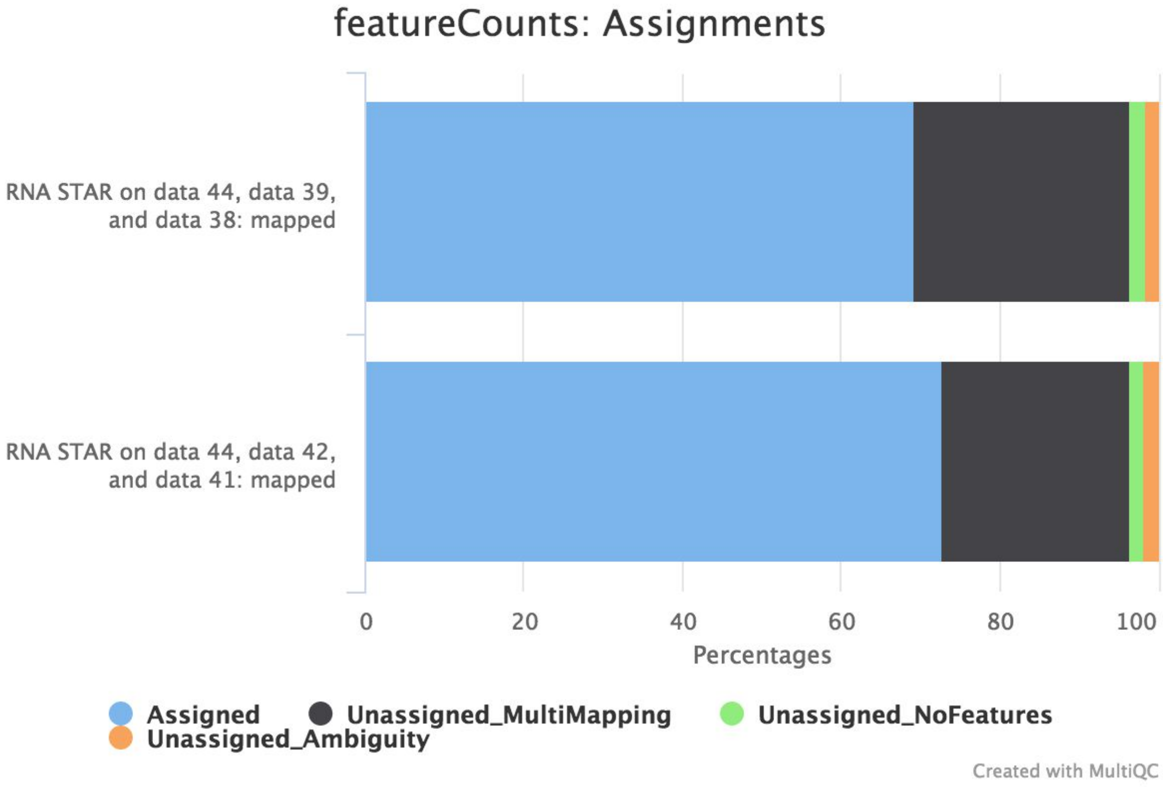


- A read is said to overlap a feature if at least one read base is found to overlap the feature.

- For paired-end data, a fragment (or template) is said to overlap a feature if any of the two reads from that fragment is found to overlap the feature.

- If strandedness is specified, then in addition to considering the genomic coordinates it will also take the strand into account for counting.

RNA-SEQ: QUANTIFICATION

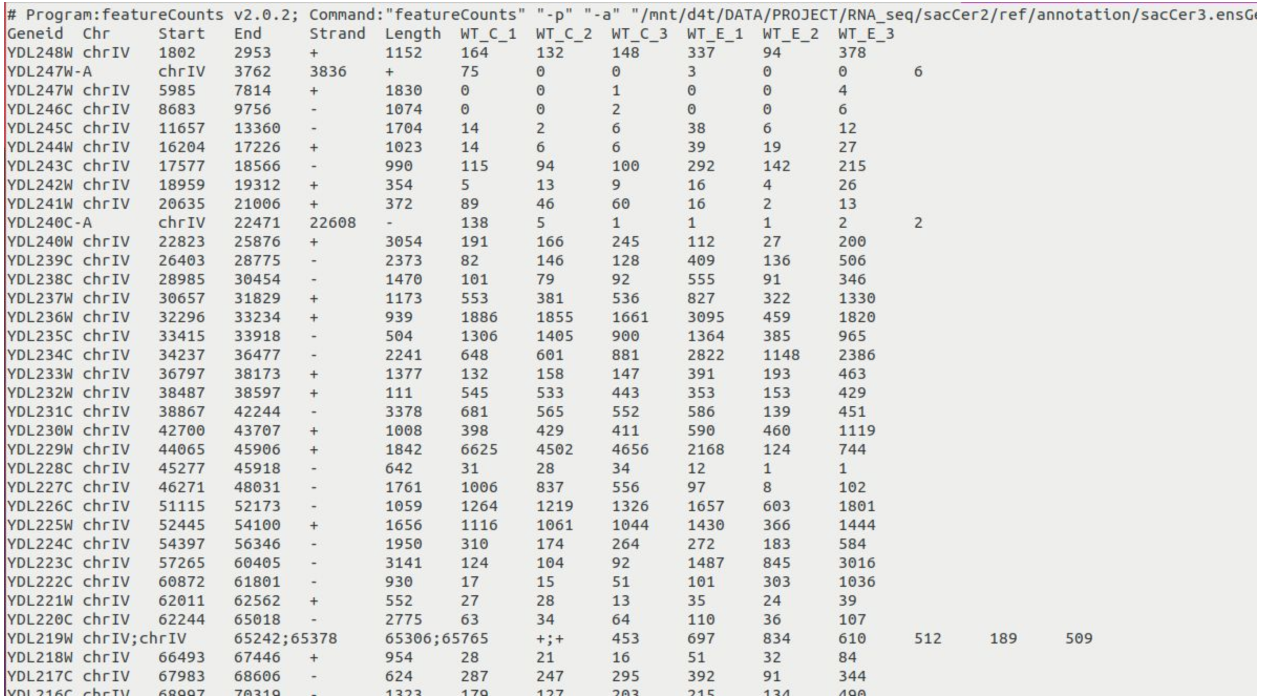


Example of a featureCounts Assignment.

RNA-SEQ: QUANTIFICATION

Counting reads using featureCounts

gene Location Strand Length Count



```
# Program:featureCounts v2.0.2; Command:"featureCounts" "-p" "-a" "/mnt/d4t/DATA/PROJECT/RNA_seq/sacCer2/ref/annotation/sacCer3.ensGe
Geneid  Chr     Start   End     Strand  Length  WT_C_1  WT_C_2  WT_C_3  WT_E_1  WT_E_2  WT_E_3
YDL248W chrIV   1802    2953    +       1152    164     132     148     337     94      378
YDL247W-A       chrIV   3762    3836    +       75      0       0       3       0       0       6
YDL247W chrIV   5985    7814    +       1830    0       0       1       0       0       4
YDL246C chrIV   8683    9756    -       1074    0       0       2       0       0       6
YDL245C chrIV   11657   13360   -       1704    14      2       6       38      6       12
YDL244W chrIV   16204   17226   +       1023    14      6       6       39      19      27
YDL243C chrIV   17577   18566   -       990     115     94      100     292     142     215
YDL242W chrIV   18959   19312   +       354     5       13      9       16      4       26
YDL241W chrIV   20635   21006   +       372     89      46      60      16      2       13
YDL240C-A       chrIV   22471   22608   -       138     5       1       1       1       2       2
YDL240W chrIV   22823   25876   +       3054    191     166     245     112     27      200
YDL239C chrIV   26403   28775   -       2373    82      146     128     409     136     506
YDL238C chrIV   28985   30454   -       1470    101     79      92      555     91      346
YDL237W chrIV   30657   31829   +       1173    553     381     536     827     322     1330
YDL236W chrIV   32296   33234   +       939     1886    1855    1661    3095    459     1820
YDL235C chrIV   33415   33918   -       504     1306    1405    900     1364    385     965
YDL234C chrIV   34237   36477   -       2241    648     601     881     2822    1148    2386
YDL233W chrIV   36797   38173   +       1377    132     158     147     391     193     463
YDL232W chrIV   38487   38597   +       111     545     533     443     353     153     429
YDL231C chrIV   38867   42244   -       3378    681     565     552     586     139     451
YDL230W chrIV   42700   43707   +       1008    398     429     411     590     460     1119
YDL229W chrIV   44065   45906   +       1842    6625    4502    4656    2168    124     744
YDL228C chrIV   45277   45918   -       642     31      28      34      12      1       1
YDL227C chrIV   46271   48031   -       1761    1006    837     556     97      8       102
YDL226C chrIV   51115   52173   -       1059    1264    1219    1326    1657    603     1801
YDL225W chrIV   52445   54100   +       1656    1116    1061    1044    1430    366     1444
YDL224C chrIV   54397   56346   -       1950    310     174     264     272     183     584
YDL223C chrIV   57265   60405   -       3141    124     104     92      1487    845     3016
YDL222C chrIV   60872   61801   -       930     17      15      51      101     303     1036
YDL221W chrIV   62011   62562   +       552     27      28      13      35      24      39
YDL220C chrIV   62244   65018   -       2775    63      34      64      110     36      107
YDL219W chrIV;chrIV     65242;65378     65306;65765     +;+     453     697     834     610     512     189     509
YDL218W chrIV   66493   67446   +       954     28      21      16      51      32      84
YDL217C chrIV   67983   68606   -       624     287     247     295     392     91      344
YDL216C chrIV   68997   70319   -       1323    179     127     203     215     134     490
```

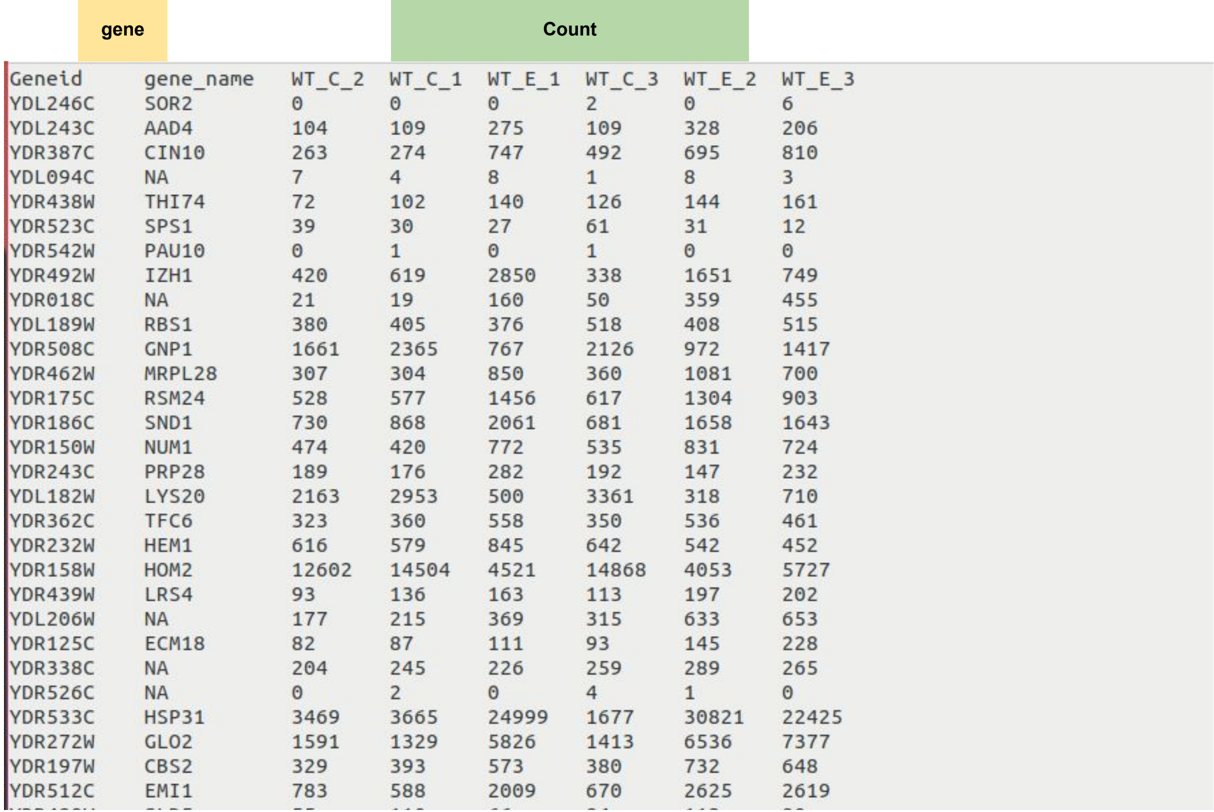
Output: Raw counts

These are the “raw” counts will be used in statistical programs downstream for differential gene expression.

featureCounts output

RNA-SEQ: QUANTIFICATION

Counting reads using featureCounts



Geneid	gene_name	WT_C_2	WT_C_1	WT_E_1	WT_C_3	WT_E_2	WT_E_3
YDL246C	SOR2	0	0	0	2	0	6
YDL243C	AAD4	104	109	275	109	328	206
YDR387C	CIN10	263	274	747	492	695	810
YDL094C	NA	7	4	8	1	8	3
YDR438W	THI74	72	102	140	126	144	161
YDR523C	SPS1	39	30	27	61	31	12
YDR542W	PAU10	0	1	0	1	0	0
YDR492W	IZH1	420	619	2850	338	1651	749
YDR018C	NA	21	19	160	50	359	455
YDL189W	RBS1	380	405	376	518	408	515
YDR508C	GNP1	1661	2365	767	2126	972	1417
YDR462W	MRPL28	307	304	850	360	1081	700
YDR175C	RSM24	528	577	1456	617	1304	903
YDR186C	SND1	730	868	2061	681	1658	1643
YDR150W	NUM1	474	420	772	535	831	724
YDR243C	PRP28	189	176	282	192	147	232
YDL182W	LYS20	2163	2953	500	3361	318	710
YDR362C	TFC6	323	360	558	350	536	461
YDR232W	HEM1	616	579	845	642	542	452
YDR158W	HOM2	12602	14504	4521	14868	4053	5727
YDR439W	LRS4	93	136	163	113	197	202
YDL206W	NA	177	215	369	315	633	653
YDR125C	ECM18	82	87	111	93	145	228
YDR338C	NA	204	245	226	259	289	265
YDR526C	NA	0	2	0	4	1	0
YDR533C	HSP31	3469	3665	24999	1677	30821	22425
YDR272W	GLO2	1591	1329	5826	1413	6536	7377
YDR197W	CBS2	329	393	573	380	732	648
YDR512C	EMI1	783	588	2009	670	2625	2619

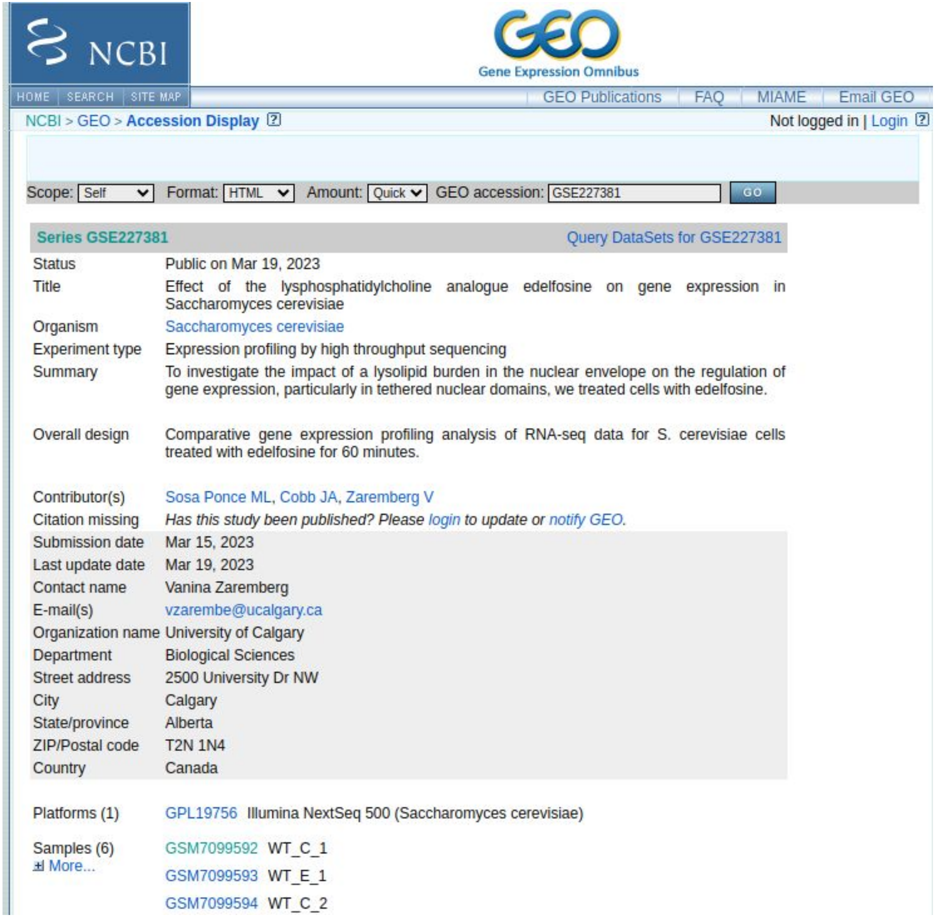
A table of counts

Don’t need information about the genomic coordinates, length

→ Cleaning up the featureCounts matrix

Final output:
A count matrix, with genes as rows and samples are columns

EXAMPLE DATA



Scope: Self Format: HTML Amount: Quick GEO accession: GSE227381

Series GSE227381 Query DataSets for GSE227381

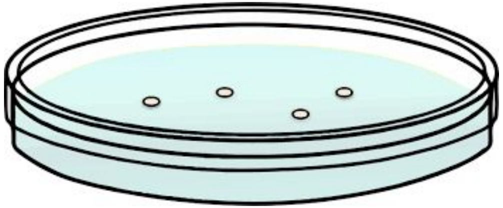
Status	Public on Mar 19, 2023
Title	Effect of the lysphosphatidylcholine analogue edelfosine on gene expression in Saccharomyces cerevisiae
Organism	Saccharomyces cerevisiae
Experiment type	Expression profiling by high throughput sequencing
Summary	To investigate the impact of a lysolipid burden in the nuclear envelope on the regulation of gene expression, particularly in tethered nuclear domains, we treated cells with edelfosine.
Overall design	Comparative gene expression profiling analysis of RNA-seq data for S. cerevisiae cells treated with edelfosine for 60 minutes.
Contributor(s)	Sosa Ponce ML, Cobb JA, Zaremberg V
Citation missing	*Has this study been published? Please login to update or notify GEO.*
Submission date	Mar 15, 2023
Last update date	Mar 19, 2023
Contact name	Vanina Zaremberg
E-mail(s)	vzarembe@ucalgary.ca
Organization name	University of Calgary
Department	Biological Sciences
Street address	2500 University Dr NW
City	Calgary
State/province	Alberta
ZIP/Postal code	T2N 1N4
Country	Canada
Platforms (1)	GPL19756 Illumina NextSeq 500 (Saccharomyces cerevisiae)
Samples (6) ⊞ More...	GSM7099592 WT_C_1 GSM7099593 WT_E_1 GSM7099594 WT_C_2

“Effect of the lysphosphatidylcholine analogue edelfosine on gene expression in *Saccharomyces cerevisiae*”

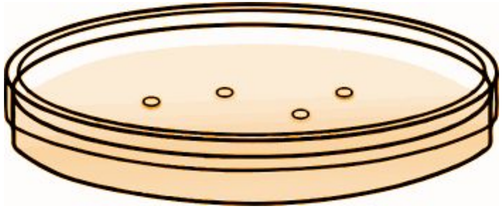
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE227381

EXAMPLE DATA

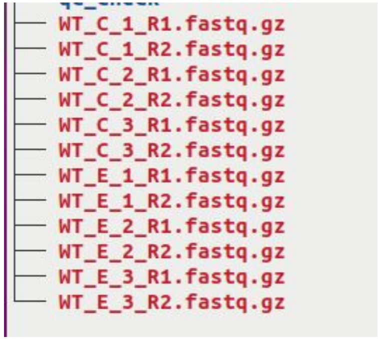
Comparative gene expression profiling analysis of RNA-seq data for S. cerevisiae cells treated with edelfosine for 60 minutes.



Control (X3)



Edelfosine treatment (X3)



```
├── WT_C_1_R1.fastq.gz
├── WT_C_1_R2.fastq.gz
├── WT_C_2_R1.fastq.gz
├── WT_C_2_R2.fastq.gz
├── WT_C_3_R1.fastq.gz
├── WT_C_3_R2.fastq.gz
├── WT_E_1_R1.fastq.gz
├── WT_E_1_R2.fastq.gz
├── WT_E_2_R1.fastq.gz
├── WT_E_2_R2.fastq.gz
├── WT_E_3_R1.fastq.gz
└── WT_E_3_R2.fastq.gz
```

Bulk RNA-seq Analysis

THANK YOU