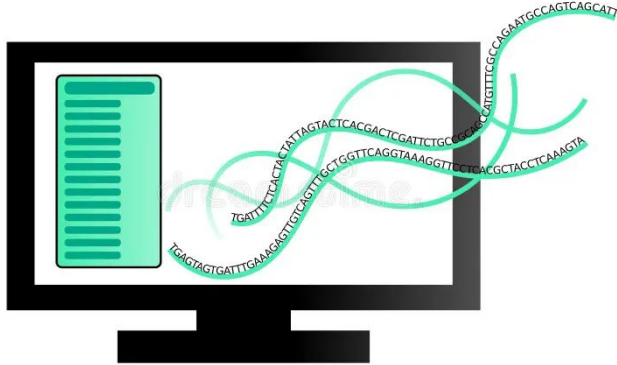


BULK RNA-SEQ: UPSTREAM ANALYSIS



Presenter: Duy Dao

TABLE OF CONTENTS

RNA SEQ: UPSTREAM ANALYSIS

01

INTRODUCTION

02

RAW DATA PROCESSING

03

ALIGNMENT

04

MAPPED READS POST-PROCESSING

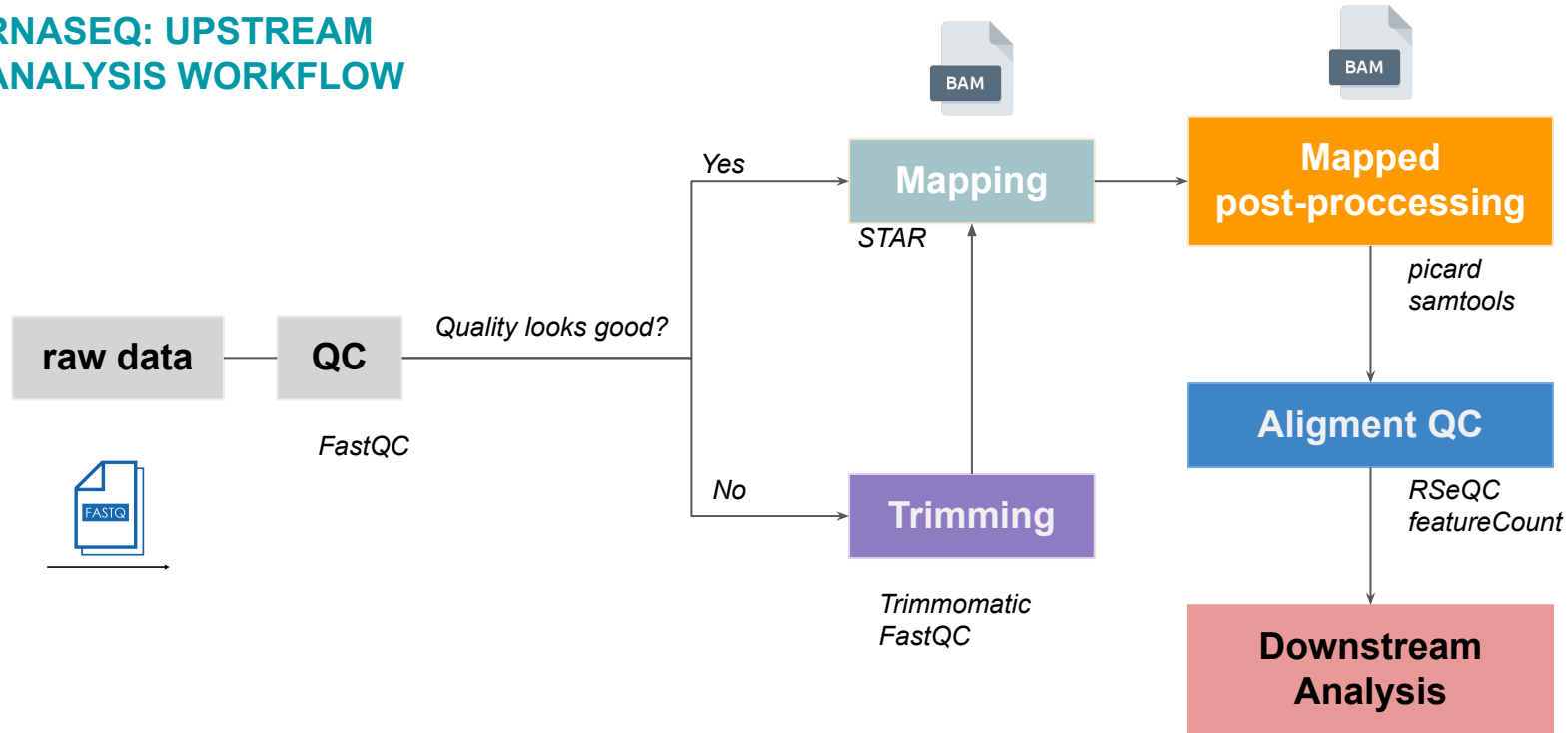
05

ALIGNMENT DATA: QUALITY CONTROL



INTRODUCTION


RNASEQ: UPSTREAM ANALYSIS WORKFLOW














RAW DATA PROCESSING

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 |

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:

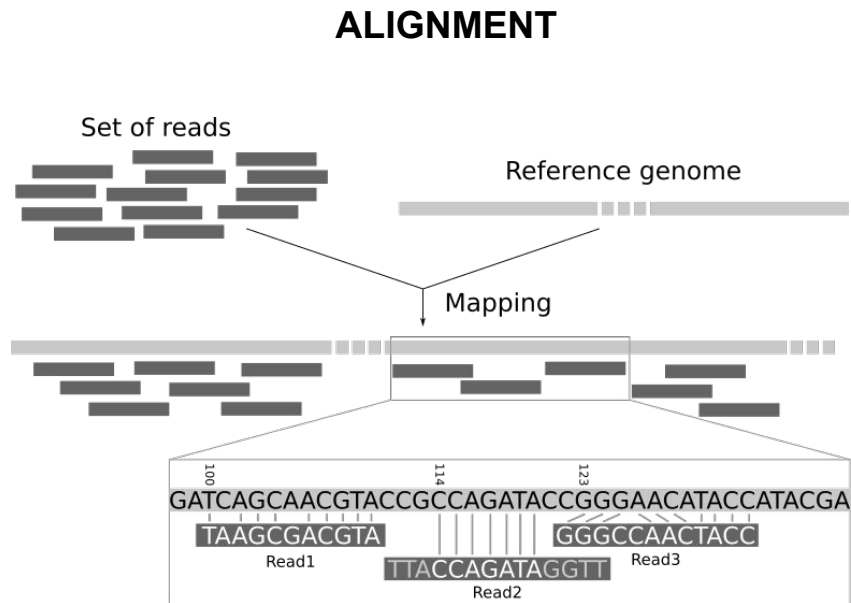
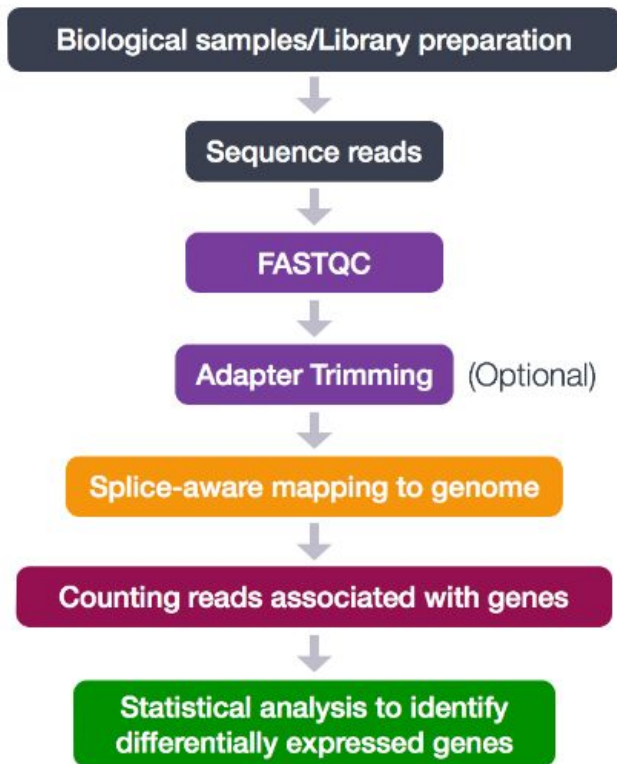
- Quality trimming
- Adapter trimming.

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

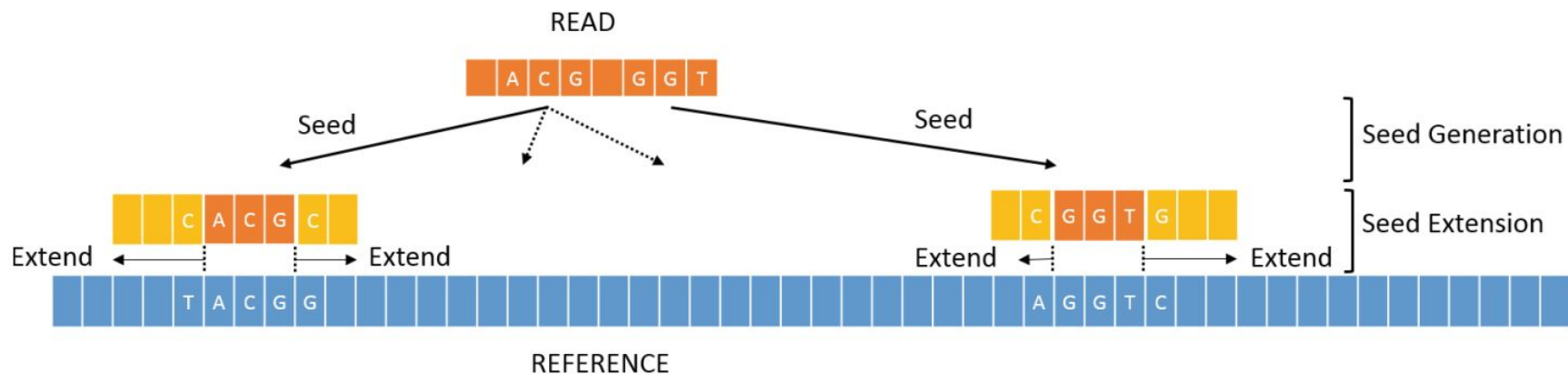


Workflow for a RNA-seq analysis.

RNA-SEQ: ALIGNMENT / MAPPING

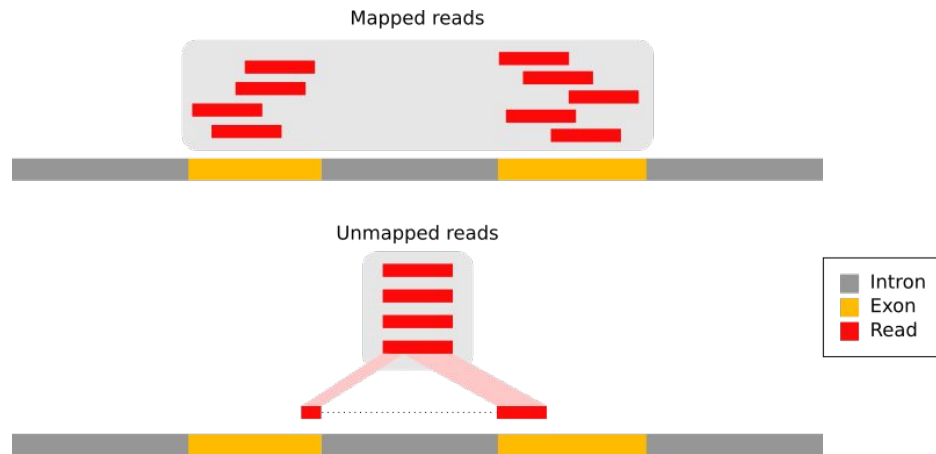
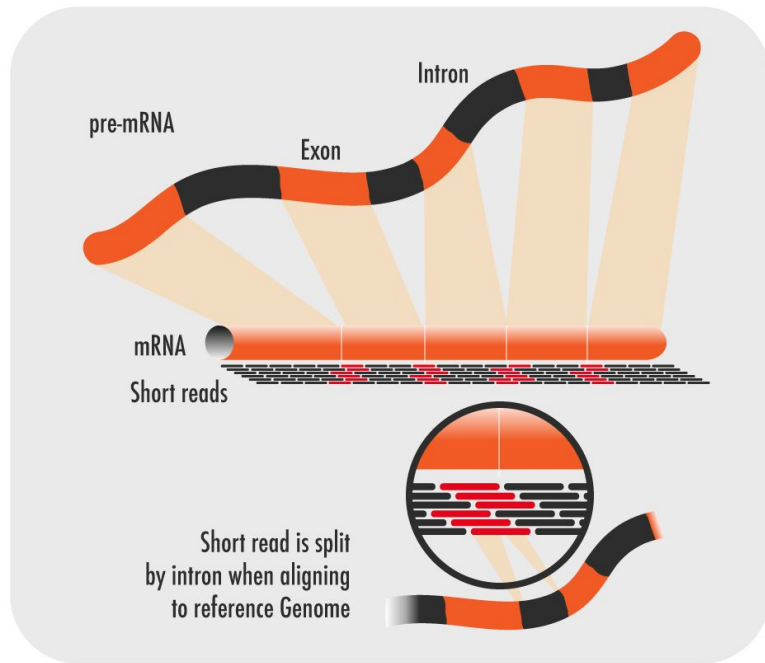
Basic alignment (Contiguous Alignment / Non-spliced Alignment)

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



RNA-SEQ: ALIGNMENT / MAPPING

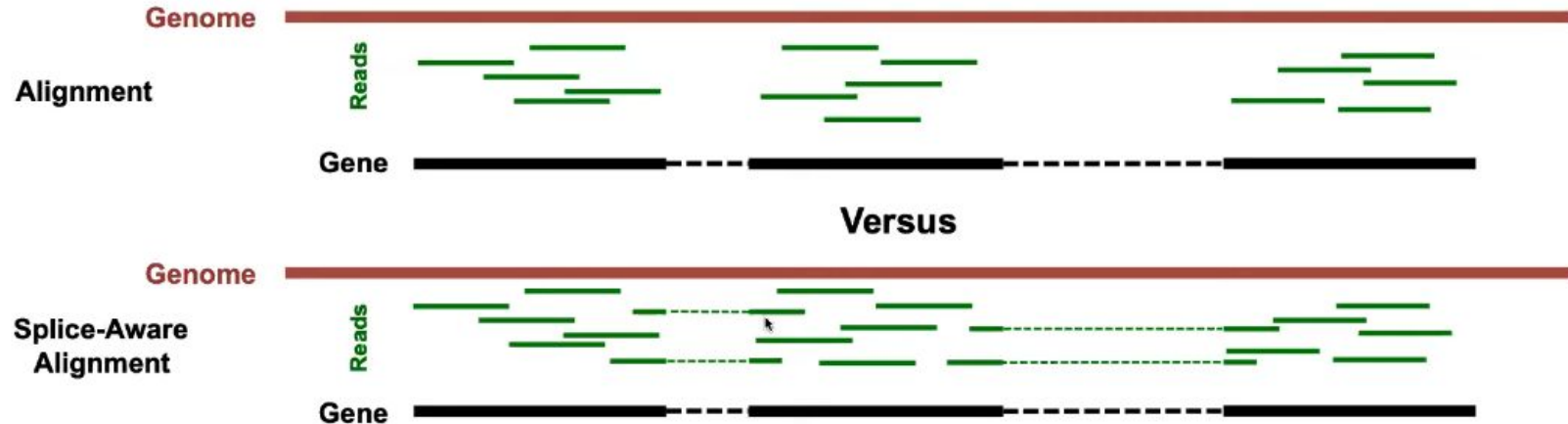
Problem when using basic alignment to map RNA-seq data



- Unmapped reads due to intron splicing.

RNA-SEQ: ALIGNMENT / MAPPING

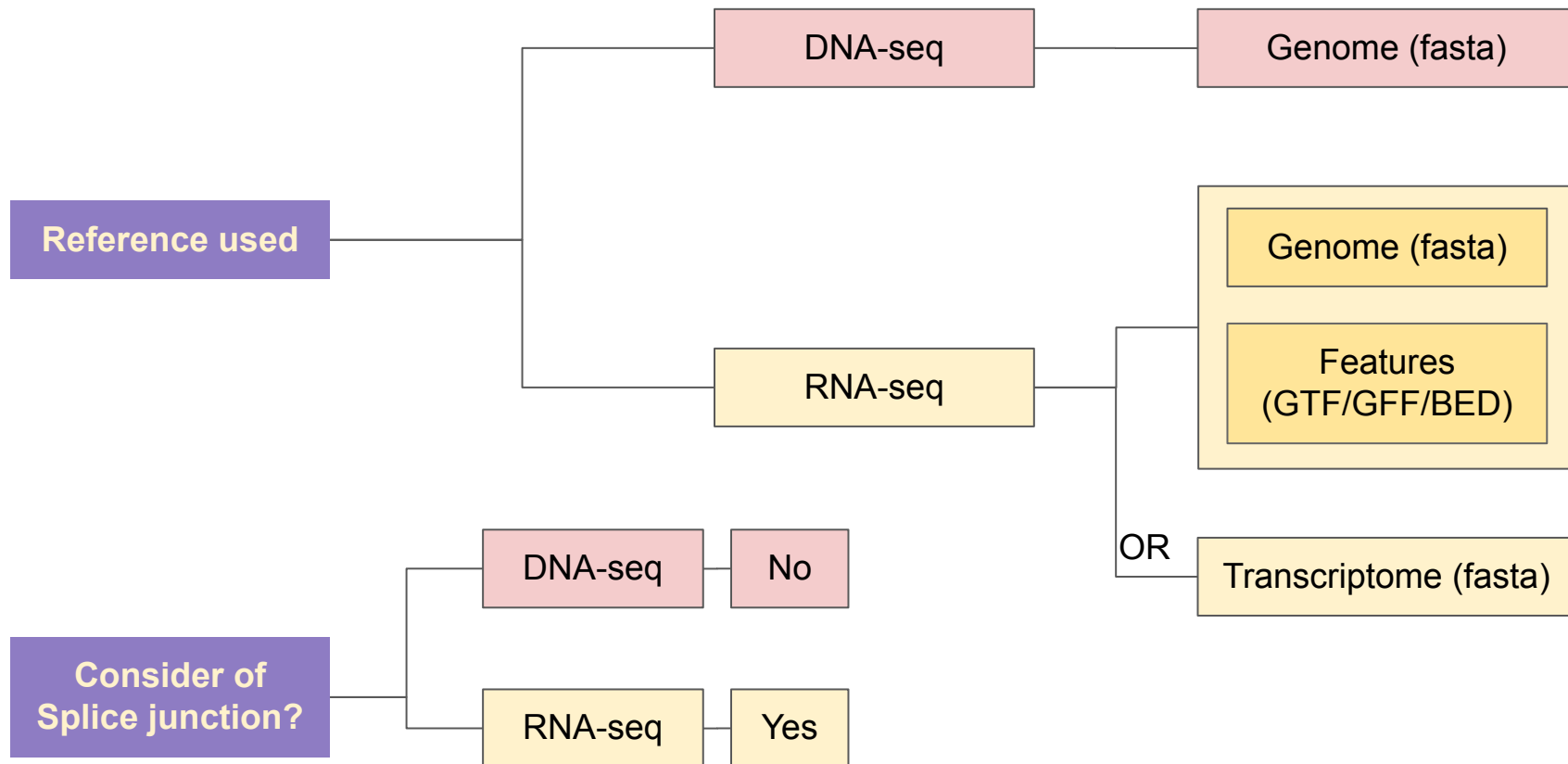
Contiguous Alignment vs Splice-Aware Alignment



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

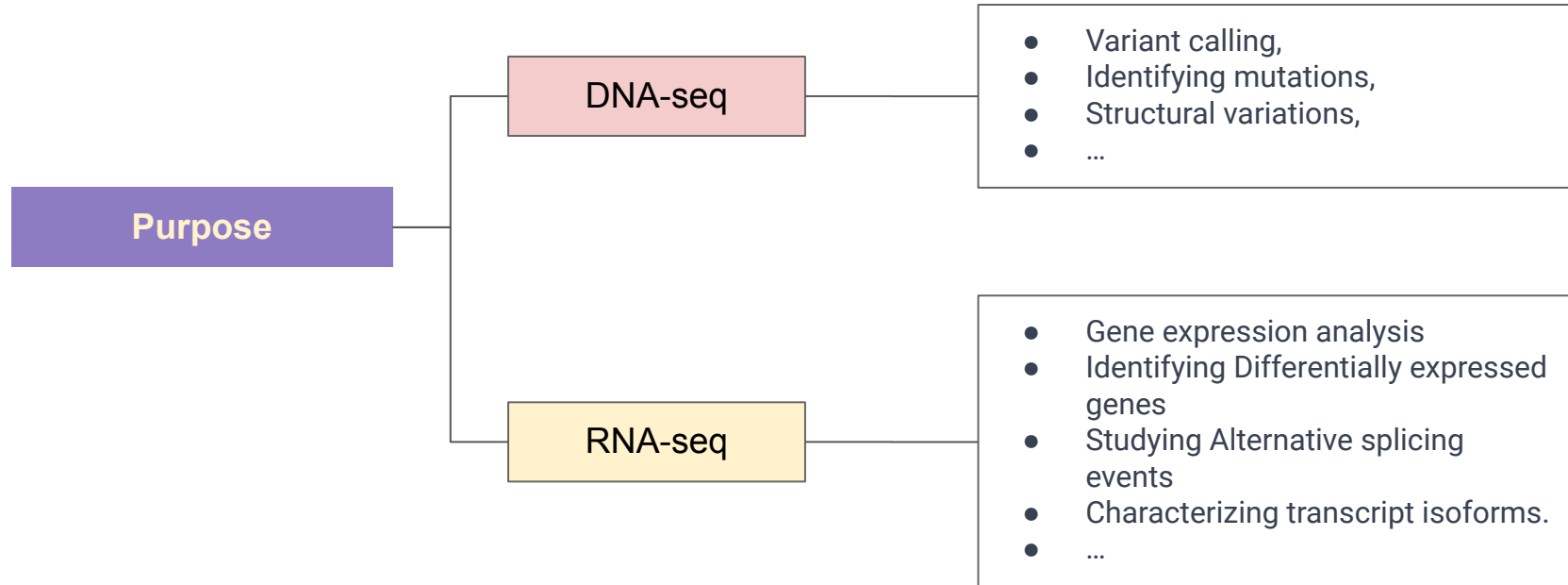
RNA-SEQ: ALIGNMENT / MAPPING

Compare the Alignment of DNA-seq and RNA-seq



RNA-SEQ: ALIGNMENT / MAPPING

Compare the Alignment of DNA-seq and RNA-seq



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.

alexdobin/**STAR**

RNA-seq aligner



36

Contributors

637

Issues

44

Discussions

2k

Stars

442

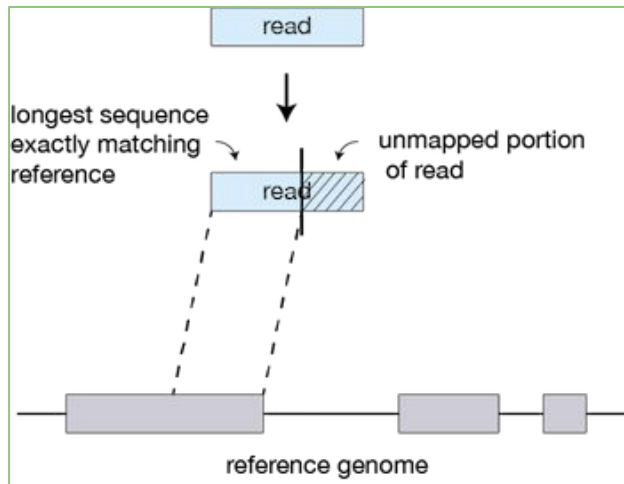
Forks



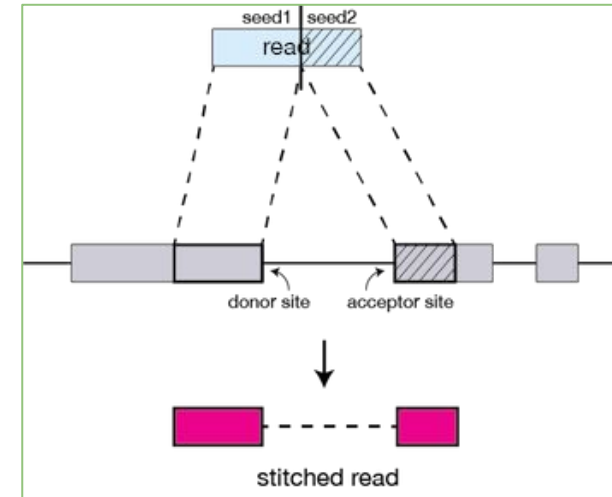
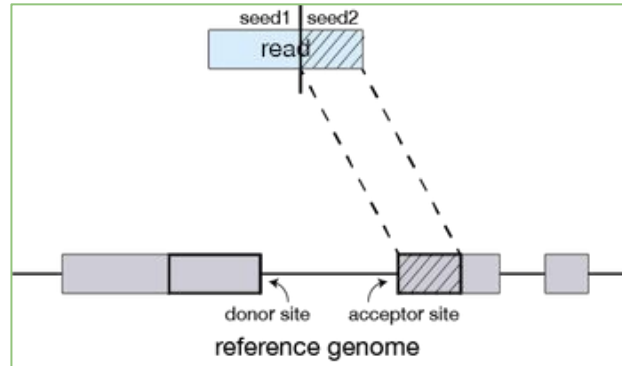
RNA-SEQ: ALIGNMENT / MAPPING

STAR (Spliced Transcripts Alignment to a Reference)

STAR alignment strategy

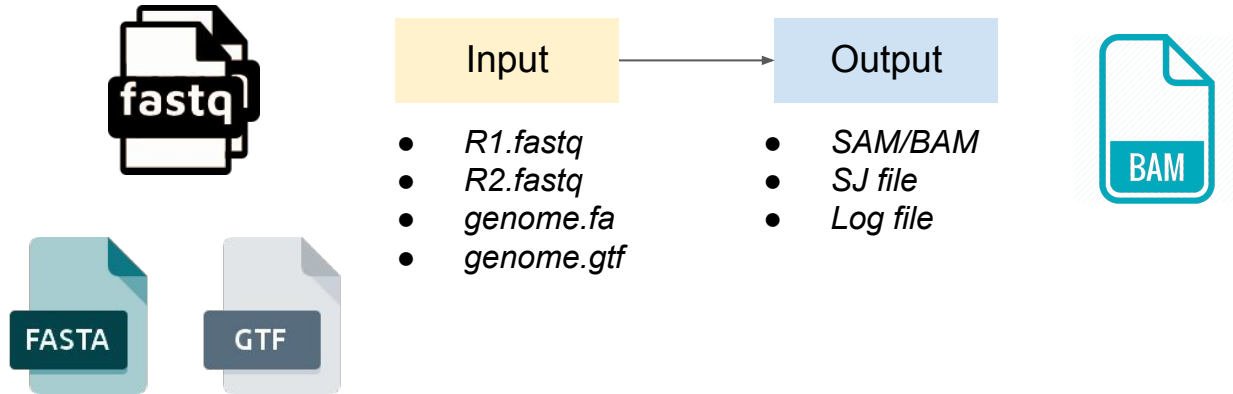


Seed searching...



Clustering, stitching, and scoring

RNA-SEQ: ALIGNMENT / MAPPING



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:
```

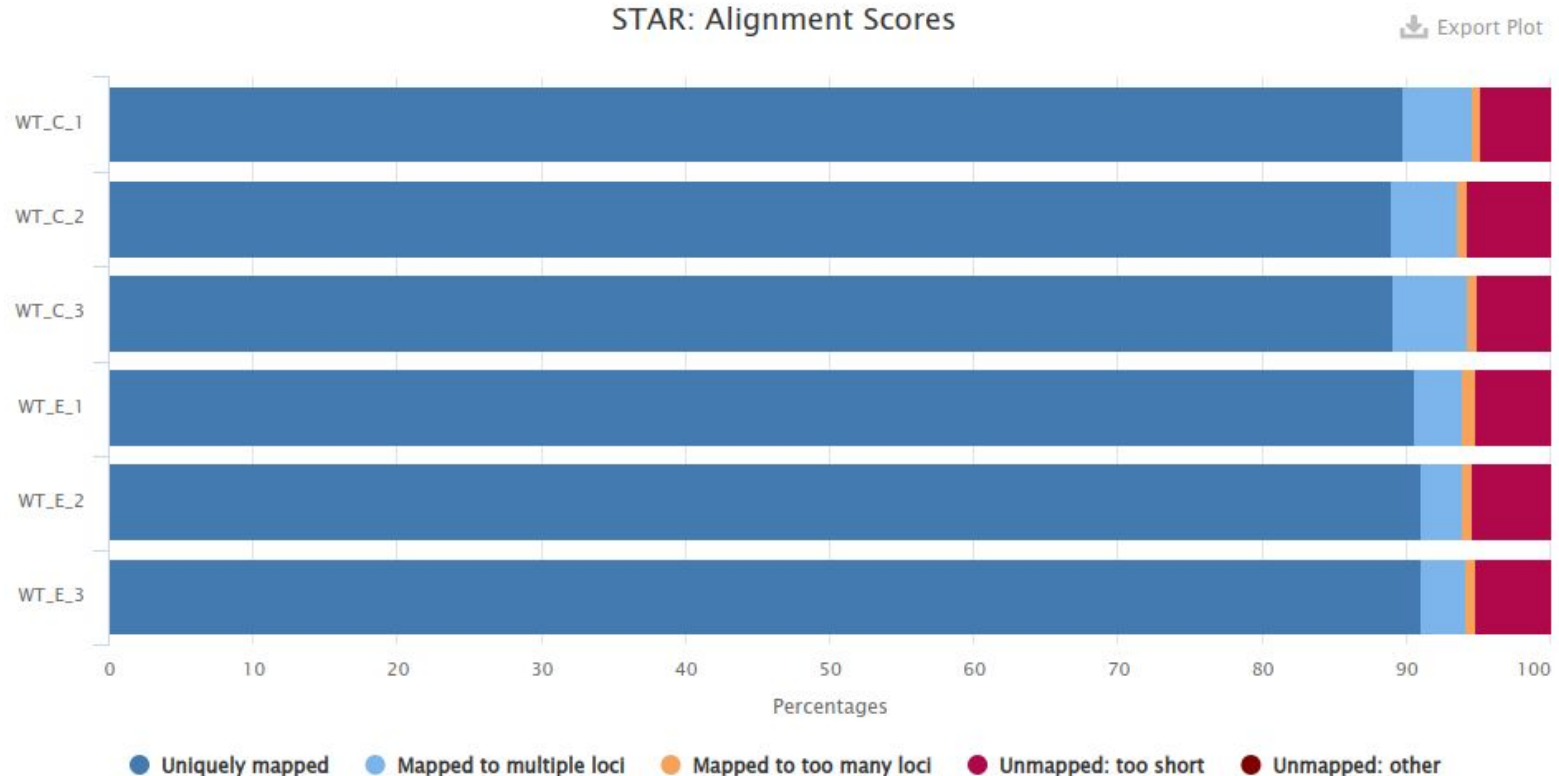
SJ.out.tab

| | | | | | | | | |
|-------|--------|--------|---|---|---|------|----|----|
| 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

| QNAME | FLAG | RNAME | POS | MAPQ | CIGAR | RNEXT | PNEXT | TLEN | SEQ |
|--|------|-------|------------|------|-------|---------------|--|-------|---|
| HWI-D00119:50:H7AP8ADXX:1:1101:1322:2086 | 113 | chr17 | 74243430 | 60 | 86M | = | 74243430 | 0 | GCCTGGGACTGCCAAGCGCTCCAAGTGTCTCTGGCCTCGTGAGTATCCCTACCCGTGGACCTGGGACAAGGG |
| HWI-D00119:50:H7AP8ADXX:1:1101:1322:2086 | 177 | chr17 | 74243430 | 60 | 86M | = | 74243430 | 0 | GCCTGGGACTGCCAAGCGCTCCAAGTGTCTCTGGCCTCGTGAGTATCCCTACCCGTGGACCTGGGACAAGGG |
| HWI-D00119:50:H7AP8ADXX:1:1101:1294:2087 | 65 | chr1 | 49798791 | 60 | 73M | = | 49798791 | 0 | CCAAACTCTGGAAGTTTATTTATTTTAAAAATATTTTCAGCATCTTTGTAATTAAGTGTTATCCACCTG |
| HWI-D00119:50:H7AP8ADXX:1:1101:1294:2087 | 129 | chr1 | 49798791 | 60 | 73M | = | 49798791 | 0 | CCAAACTCTGGAAGTTTATTTATTTTAAAAATATTTTCAGCATCTTTGTAATTAAGTGTTATCCACCTG |
| HWI-D00119:50:H7AP8ADXX:1:1101:1406:2159 | 65 | chr18 | 9733113 60 | 86M | = | 9733113 0 | CCATGCTATAAATCTTTATTTTTTCCACTTTGGTTTGAGTGGGATCACTTCATTGGAGCTAATCCAGAAATGCTAATTCAGTGG | HGI | CGGACCA |
| HWI-D00119:50:H7AP8ADXX:1:1101:1406:2159 | 129 | chr18 | 9733113 60 | 86M | = | 9733113 0 | CCATGCTATAAATCTTTATTTTTTCCACTTTGGTTTGAGTGGGATCACTTCATTGGAGCTAATCCAGAAATGCTAATTCAGTGG | HGI | CCAGGTA |
| HWI-D00119:50:H7AP8ADXX:1:1101:1303:2183 | 65 | chr6 | 132570953 | 60 | 86M | = | 132570953 | 0 | TTATTATGATAATTCGTATGGTAAACATATTTCTGTGGCTAGACGACAGCGGCAAAAGATAGAAAATACTGGTA |
| HWI-D00119:50:H7AP8ADXX:1:1101:1303:2183 | 129 | chr6 | 132570953 | 60 | 86M | = | 132570953 | 0 | TTATTATGATAATTCGTATGGTAAACATATTTCTGTGGCTAGACGACAGCGGCAAAAGATAGAAAATACTGGTA |
| HWI-D00119:50:H7AP8ADXX:1:1101:1413:2244 | 81 | chr19 | 56204243 | 0 | 86M | = | 56195531 | -8798 | TTCGCGCCCTACGAGTGTGTGGACTTTCAAATTTCTGAGGAGCTCTCTATCACTGTCTCCCATATCCGGCTGCTCTA |
| HWI-D00119:50:H7AP8ADXX:1:1101:1413:2244 | 161 | chr19 | 56195531 | 0 | 86M | = | 56204243 | 8798 | GGGATGTAATAATAGAGCAGCGGATATGGGACAGACTATAGAAAGTCTCCAGAAAATGAAAGTCCACAACTGCA |
| HWI-D00119:50:H7AP8ADXX:1:1101:1372:2246 | 113 | chr19 | 50402338 | 60 | 86M | = | 50402338 | 0 | CGCGCCCTACGAGGCCAAGCTCGACTTTGAGATCCGGTACGGCTCTCGCTCACTTCTCCGGCTCTATCCCCAC |
| HWI-D00119:50:H7AP8ADXX:1:1101:1372:2246 | 177 | chr19 | 50402338 | 60 | 86M | = | 50402338 | 0 | CGCGCCCTACGAGGCCAAGCTCGACTTTGAGATCCGGTACGGCTCTCGCTCACTTCTCCGGCTCTATCCCCAC |
| HWI-D00119:50:H7AP8ADXX:1:1101:1613:2099 | 113 | chr1 | 172749268 | 27 | 86M | = | 172749268 | 0 | TGCTATTCTCGCCCTCACTGGTAACCTCAACCTCGGGAATGTATCAAAACCTATAGCAGCTTTACCCCACTGAAGTT |
| HWI-D00119:50:H7AP8ADXX:1:1101:1613:2099 | 177 | chr1 | 172749268 | 27 | 86M | = | 172749268 | 0 | TGCTATTCTCGCCCTCACTGGTAACCTCAACCTCGGGAATGTATCAAAACCTATAGCAGCTTTACCCCACTGAAGTT |
| HWI-D00119:50:H7AP8ADXX:1:1101:1644:2101 | 113 | chr1 | 76700193 | 46 | 86M | = | 76700193 | 0 | GGAAGCTGGCATCATGGAAGGTTGTTTTCGACGACAGCAAAAGGTCATCGCAGATAACGCTGAAGGACTGGACAA |
| HWI-D00119:50:H7AP8ADXX:1:1101:1644:2101 | 177 | chr1 | 76700193 | 46 | 86M | = | 76700193 | 0 | GGAAGCTGGCATCATGGAAGGTTGTTTTCGACGACAGCAAAAGGTCATCGCAGATAACGCTGAAGGACTGGACAA |
| HWI-D00119:50:H7AP8ADXX:1:1101:1657:2113 | 65 | chr8 | 85651599 | 0 | 86M | = | 85654762 | -6838 | GCTGTG |
| HWI-D00119:50:H7AP8ADXX:1:1101:1657:2113 | 129 | chr8 | 85654762 | 0 | 86M | = | 85661599 | 6838 | GCTGTG |
| HWI-D00119:50:H7AP8ADXX:1:1101:1777:2104 | 113 | chr19 | 18784989 | 60 | 86M | = | 18784989 | 0 | GGAGG |
| HWI-D00119:50:H7AP8ADXX:1:1101:1777:2104 | 177 | chr19 | 18784989 | 60 | 86M | = | 18784989 | 0 | GGAGG |
| HWI-D00119:50:H7AP8ADXX:1:1101:1830:2136 | 65 | chr1 | 227731993 | 60 | 86M | = | 227731993 | 0 | CAGGG |
| HWI-D00119:50:H7AP8ADXX:1:1101:1830:2136 | 129 | chr1 | 227731993 | 60 | 86M | = | 227731993 | 0 | CAGGG |
| HWI-D00119:50:H7AP8ADXX:1:1101:2021:2071 | 65 | chr17 | 18541691 | 0 | 86M | = | 18541691 | 0 | ACCAAA |
| HWI-D00119:50:H7AP8ADXX:1:1101:2021:2071 | 129 | chr17 | 18541691 | 0 | 86M | = | 18541691 | 0 | ACCAAA |
| HWI-D00119:50:H7AP8ADXX:1:1101:2027:2162 | 65 | chr2 | 39673045 | 60 | 86M | = | 39673045 | 0 | GTGCTG |
| HWI-D00119:50:H7AP8ADXX:1:1101:2027:2162 | 129 | chr2 | 39673045 | 60 | 86M | = | 39673045 | 0 | GTGCTG |
| HWI-D00119:50:H7AP8ADXX:1:1101:2340:2087 | 65 | chr2 | 232386840 | 60 | 86M | = | 232386840 | 0 | CGCCCT |
| HWI-D00119:50:H7AP8ADXX:1:1101:2340:2087 | 129 | chr2 | 232386840 | 60 | 86M | = | 232386840 | 0 | CGCCCT |
| HWI-D00119:50:H7AP8ADXX:1:1101:2340:2087 | 129 | chr4 | 9326689 0 | 86M | = | 9331435 4747 | CGAGGGAGGGCCAGGAGATCC | 0 | CG BCC |
| HWI-D00119:50:H7AP8ADXX:1:1101:2340:2087 | 129 | chr4 | 9331435 0 | 86M | = | 9326689 -4747 | CGAGGGAGGGCCAGGAGATCC | 0 | CG BCC |

Alignment section

| 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 | ACGCATACGT |
| 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

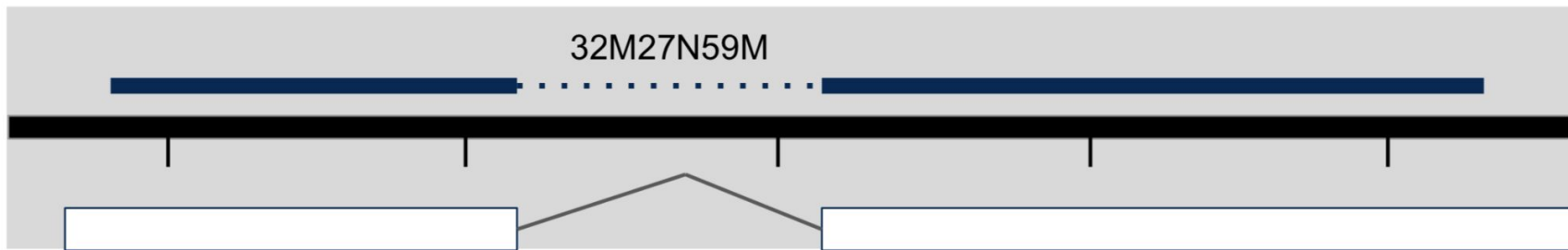
Alignment statistics & Utilities for manipulating alignment files.

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.

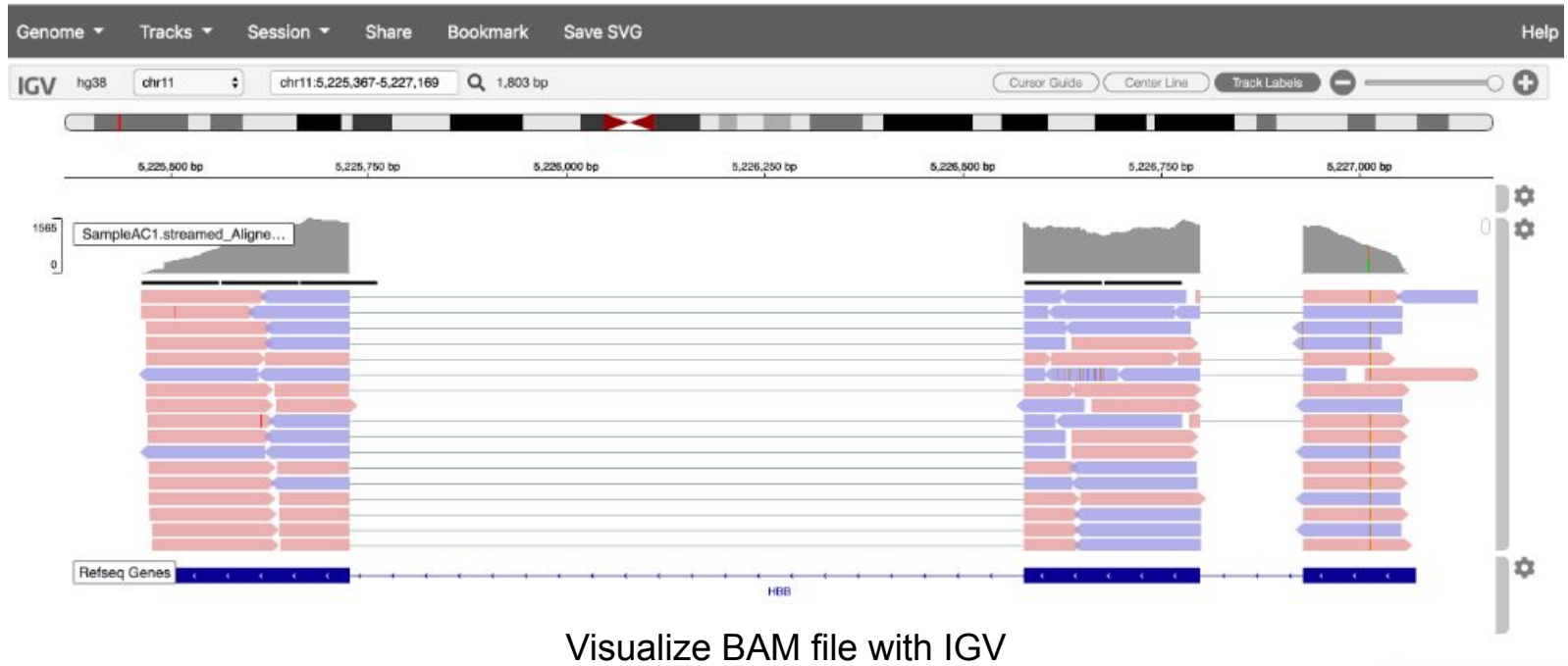
It indicates that the read aligns to the reference genome, but there is a region of the reference sequence that is not covered by the read.

Splicing:



RNA-SEQ: ALIGNMENT / MAPPING

Alignment statistics & Utilities for manipulating alignment files.



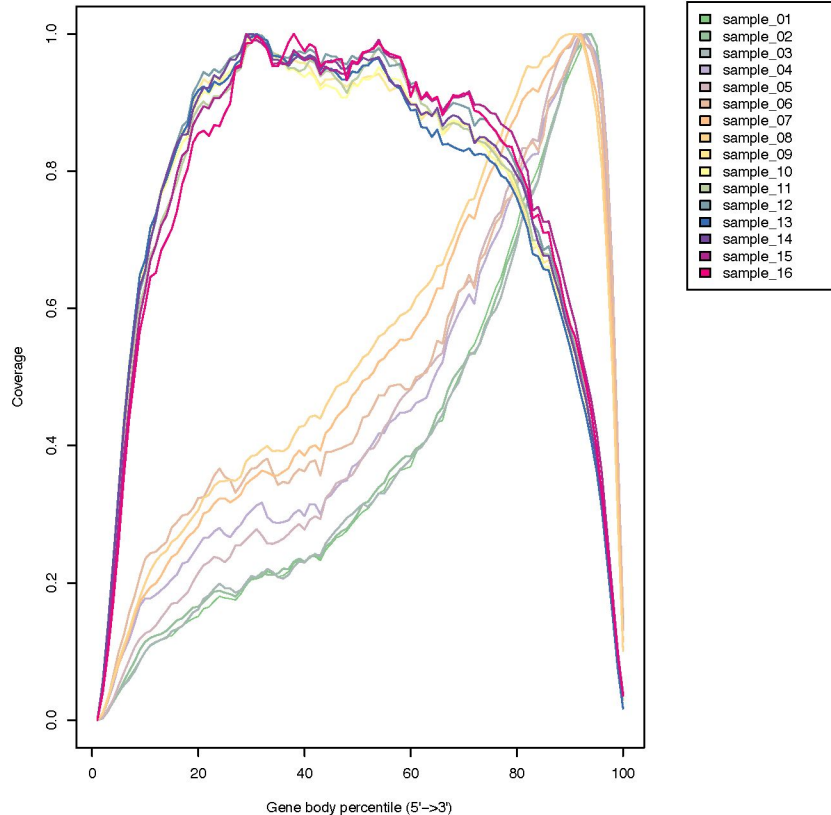
ALIGNMENT QUALITY CONTROL

Alignment quality metrics:

- coverage uniformity along transcripts
- saturation of sequencing depth
- ribosomal RNA content
- read distribution between exons, introns & intergenic regions.
- ...

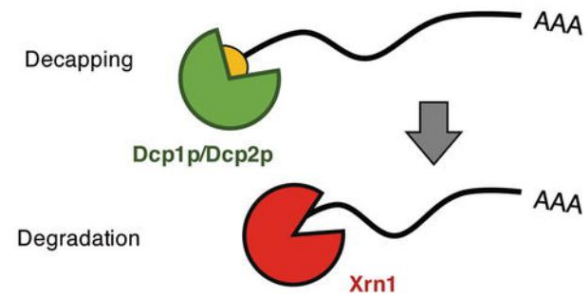
RNA-SEQ: ALIGNMENT QUALITY CONTROL

RSeQC: Genebody Coverage



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

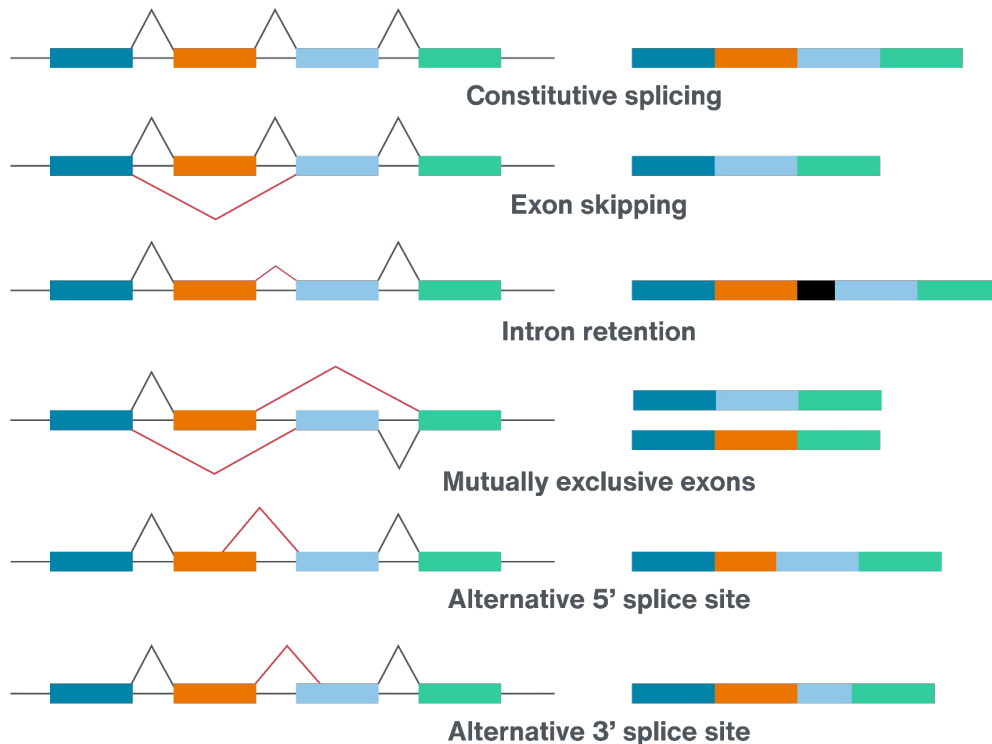
- Gene expression quantification
- Transcript isoform analysis
- Detection of gene expression biases
- Assessing RNA integrity and sample quality



Which samples may have been degraded?

RNA-SEQ: ALIGNMENT QUALITY CONTROL

RSeQC: Junction Annotation



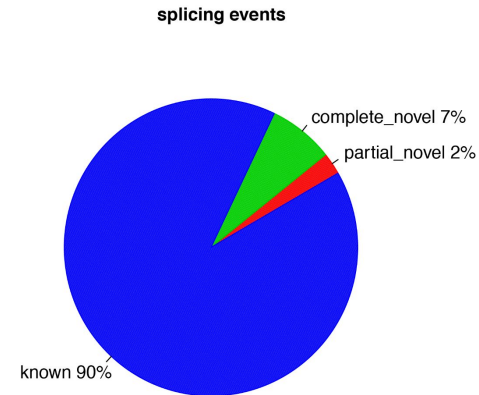
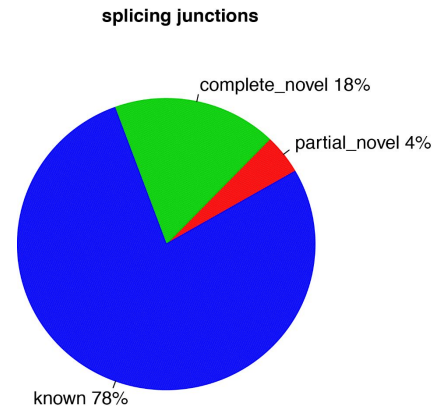
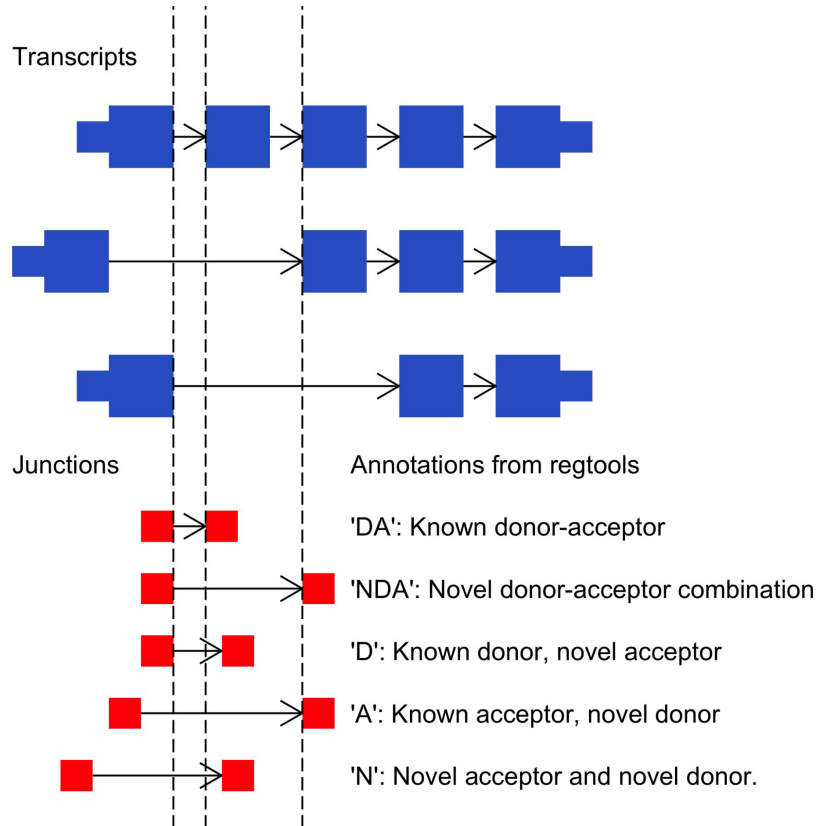
Splice variation

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.

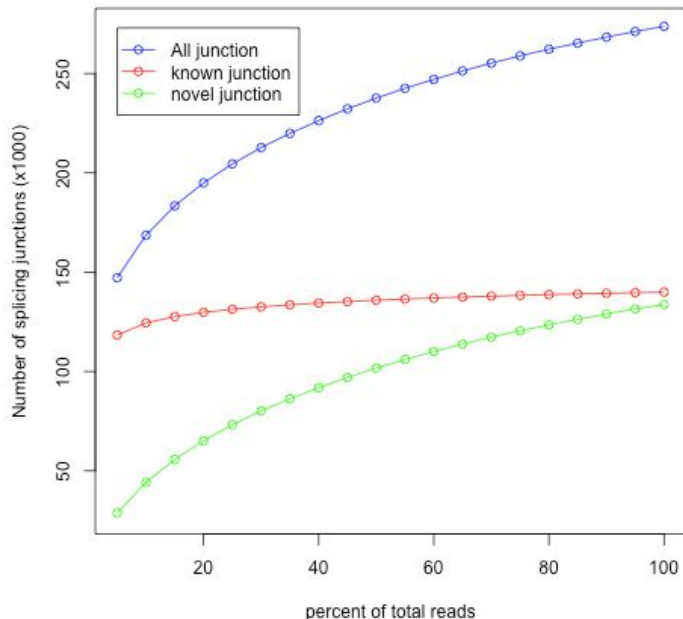
RNA-SEQ: ALIGNMENT QUALITY CONTROL

RSeQC: Junction Annotation



RNA-SEQ: ALIGNMENT QUALITY CONTROL

RSeQC: Junction Saturation



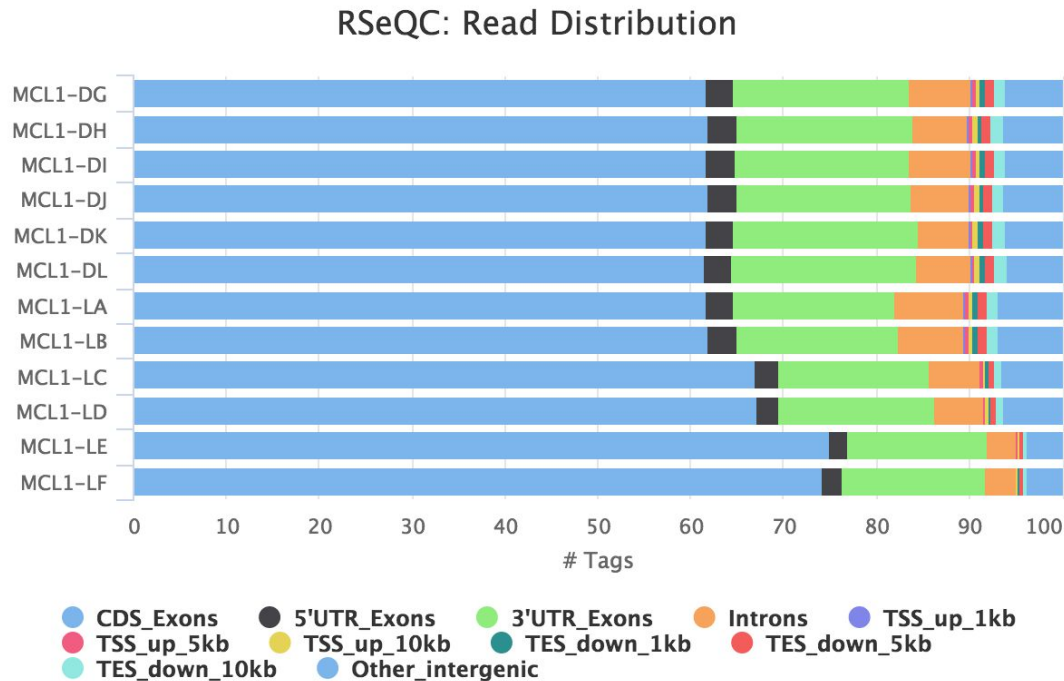
A sample that reaches a plateau before getting to 100% data indicates that all junctions in the library have been detected, and that further sequencing will not yield more observations.

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



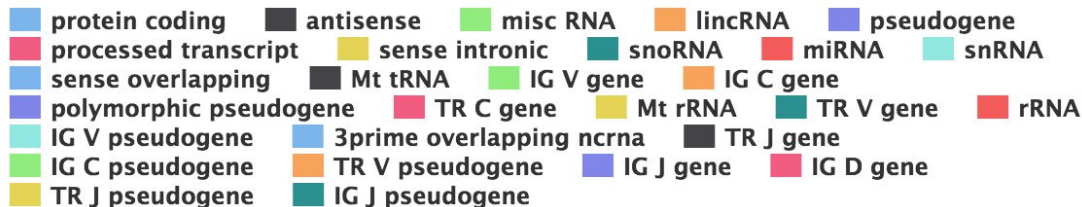
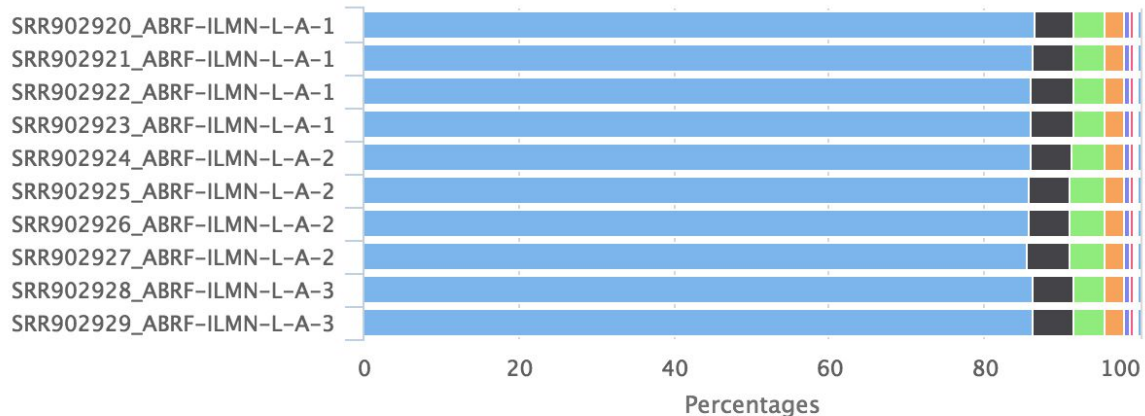
Calculate how mapped reads were distributed over genome feature (like CDS exon, 5'UTR exon, 3' UTR exon, Intron, Intergenic regions).

Created with MultiQC

RNA-SEQ: ALIGNMENT QUALITY CONTROL

Biotypes Count

featureCounts Biotypes



- 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.

EXAMPLE DATA

The screenshot shows the NCBI GEO Accession Display page for GSE227381. The page header includes the NCBI logo and the GEO logo (Gene Expression Omnibus). The navigation bar contains links for HOME, SEARCH, SITE MAP, GEO Publications, FAQ, MIAME, and Email GEO. The breadcrumb trail is NCBI > GEO > Accession Display. The user is not logged in, with a login link provided. The search filters are set to Scope: Self, Format: HTML, Amount: Quick, and GEO accession: GSE227381. The series title is "Effect of the lysphosphatidylcholine analogue edelfosine on gene expression in Saccharomyces cerevisiae". The organism is Saccharomyces cerevisiae. The experiment type is Expression profiling by high throughput sequencing. The summary states: "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." The overall design is "Comparative gene expression profiling analysis of RNA-seq data for S. cerevisiae cells treated with edelfosine for 60 minutes." The contributor is Sosa Ponce ML, Cobb JA, Zarembeg V. The citation missing message is "Has this study been published? Please login to update or notify GEO." The submission date is Mar 15, 2023, and the last update date is Mar 19, 2023. The contact name is Vanina Zarembeg, and the email is vzarembeg@ucalgary.ca. The organization is the University of Calgary, Biological Sciences department, located at 2500 University Dr NW, Calgary, Alberta, T2N 1N4, Canada. The platform is GPL19756 Illumina NextSeq 500 (Saccharomyces cerevisiae). The samples are GSM7099592 WT_C_1, GSM7099593 WT_E_1, and GSM7099594 WT_C_2.

NCBI > GEO > Accession Display Not logged in | Login

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

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, Zarembeg 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 Zarembeg

E-mail(s) vzarembeg@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)

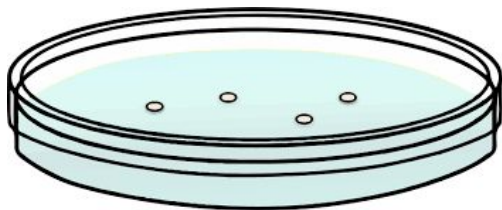
Samples (6) [GSM7099592](#) WT_C_1
[GSM7099593](#) WT_E_1
[GSM7099594](#) WT_C_2

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

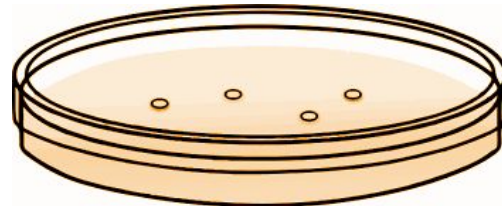
<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE227381>

EXAMPLE DATA

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



Control (X3)



Treated with edelfosine for 60 minutes (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