BIN 508: Next Generation Sequence Analysis & Informatics

Assignment 04

Author: Taha Ahmad

Instructor: Dr Yesim Aydin Son

Galaxy history:

https://usegalaxy.eu/u/taha.ahmad/h/bin-508-assignment04-part1-1

https://usegalaxy.eu/u/taha.ahmad/h/bin-508-assignment-04-part2

Question 1) Briefly describe RPKM, FPKM, and TPM. Which metric is more appropriate for RNA-Seq analysis?

Answer :

- RPKM (Reads per Kilobase Million) normalizes for sequence depth and gene length. It is used for single-end RNA-Seq where each read represents a single fragment

Formulae : $$RPKM = \frac{Reads\ mapped\ to\ gene}{(Gene\ length\ in\ kb)\ *(Total\ Reads\ in\ million)}$$

- TPM (Transcripts per Million) normalizes first by gene length and then by sequencing depth thus ensuring the sum of all TPM values equals 1 million. This preserves the proportionality of expression levels within the sample

Formulae : $$TPM = \frac{\frac{Reads\ mapped\ to\ gene}{Gene\ length\ in\ kb}}{\boldsymbol{\Sigma}\left(\frac{reads}{gene\ length}\right) for\ all\ genes}$$

- FPKM (Fragments per kilobase million) is similar to RPKM but is used for paired-end RNA-Seq. it counts fragments to avoid double-counting overlapping reads from the same fragment

Formulae : $$FPKM = \frac{Fragments\ mapped\ to\ gene}{Gene\ legnth\ in\ kb * Total\ fragments\ in\ million}$$

TPM is more appropriate for RNA-Seq analysis as it reflects the relative abundance of transcripts with a sample and ensures consistent scaling across sample (notice the summation in the equation). It also avoids biases introduced by highly expressed genes which skew the normalization in the other metrics (RPKM/FPKM)

IMPORTANT NOTE !

GSM461177 → Untreated paired
GSM461180 → Treated paired

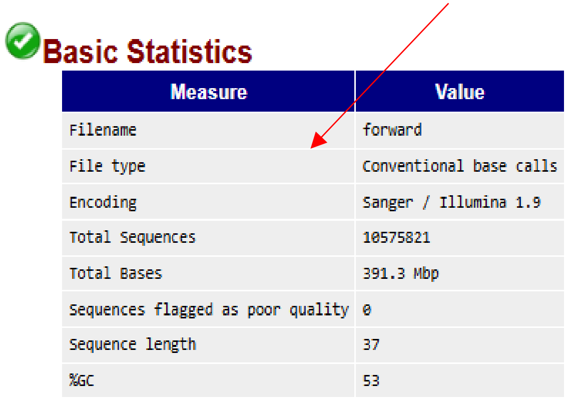
Q2 A)

Before Cutadapt

FastQC

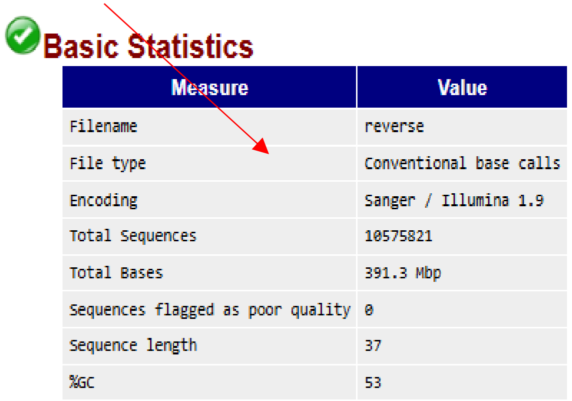
After Cutadapt

GSM 461177



Basic Statistics

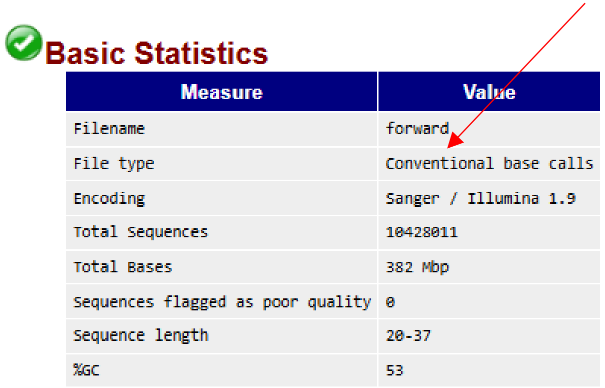
Measure	Value
Filename	forward
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10575821
Total Bases	391.3 Mbp
Sequences flagged as poor quality	0
Sequence length	37
%GC	53



Basic Statistics

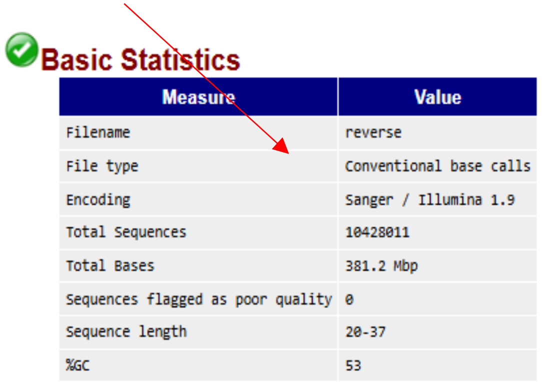
Measure	Value
Filename	reverse
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10575821
Total Bases	391.3 Mbp
Sequences flagged as poor quality	0
Sequence length	37
%GC	53

GSM461177



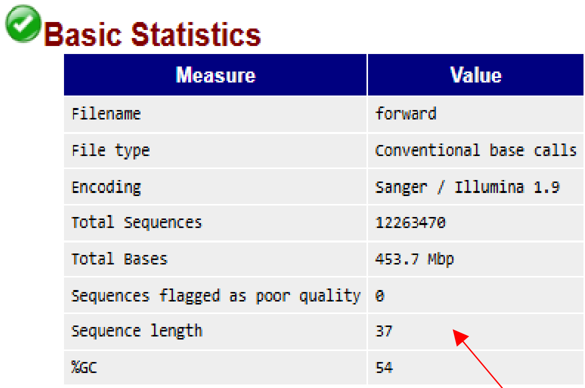
Basic Statistics

Measure	Value
Filename	forward
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10428011
Total Bases	382 Mbp
Sequences flagged as poor quality	0
Sequence length	20-37
%GC	53



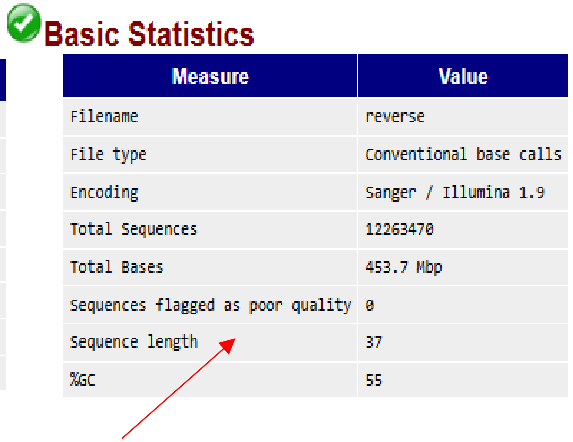
Basic Statistics

Measure	Value
Filename	reverse
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10428011
Total Bases	381.2 Mbp
Sequences flagged as poor quality	0
Sequence length	20-37
%GC	53



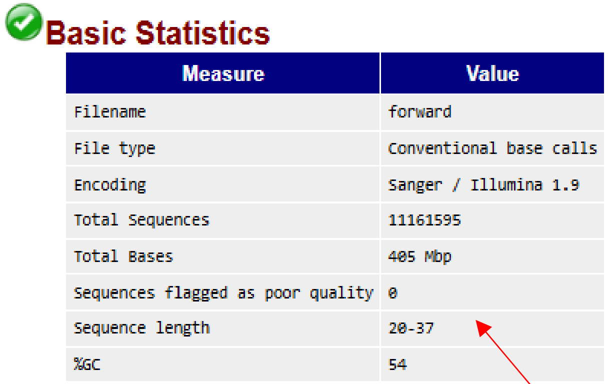
Basic Statistics

Measure	Value
Filename	forward
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	12263470
Total Bases	453.7 Mbp
Sequences flagged as poor quality	0
Sequence length	37
%GC	54



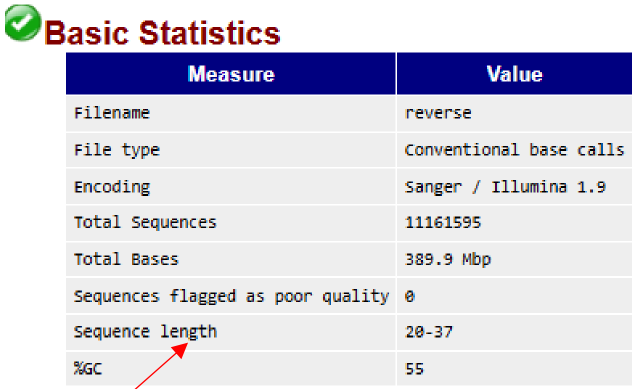
Basic Statistics

Measure	Value
Filename	reverse
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	12263470
Total Bases	453.7 Mbp
Sequences flagged as poor quality	0
Sequence length	37
%GC	55



Basic Statistics

Measure	Value
Filename	forward
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	11161595
Total Bases	405 Mbp
Sequences flagged as poor quality	0
Sequence length	20-37
%GC	54



Basic Statistics

Measure	Value
Filename	reverse
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	11161595
Total Bases	389.9 Mbp
Sequences flagged as poor quality	0
Sequence length	20-37
%GC	55

GSM461180

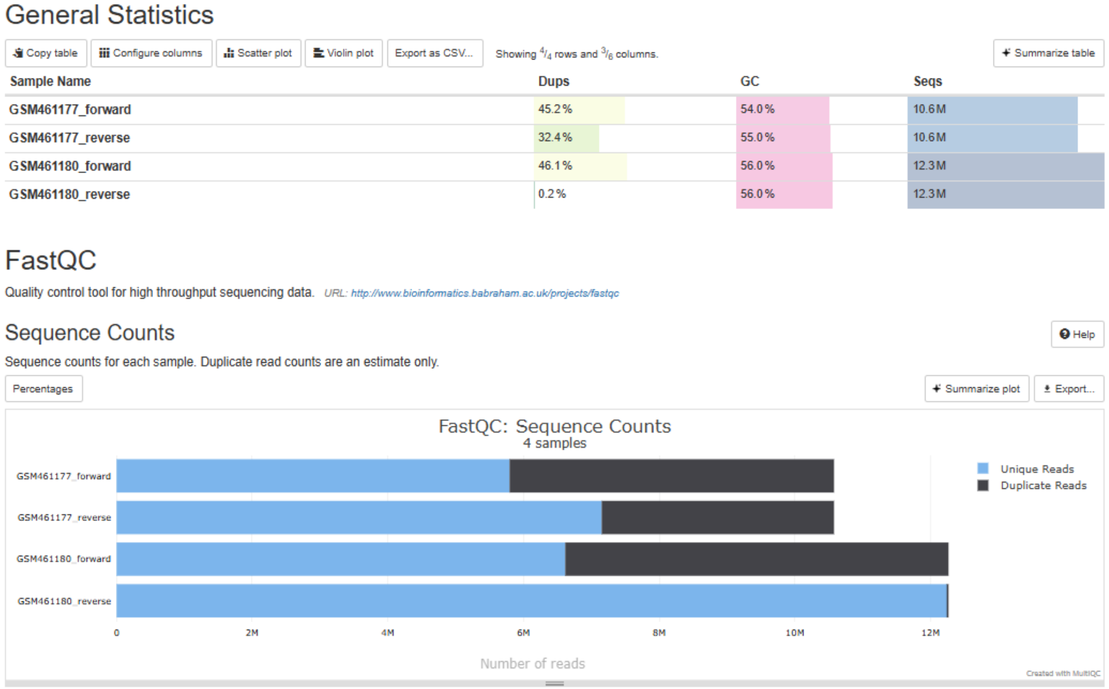
GSM461180

All samples have a sequence length of 37 bp

Removes read with quality < 20 and length < 20
(shown by the decrease in total sequences and bases)

MultiQC

Before cutadapt



General Statistics

Showing 4/4 rows and 3/6 columns.

Sample Name	Dups	GC	Seqs
GSM461177_forward	45.2%	54.0%	10.6M
GSM461177_reverse	32.4%	55.0%	10.6M
GSM461180_forward	46.1%	56.0%	12.3M
GSM461180_reverse	0.2%	56.0%	12.3M

FastQC

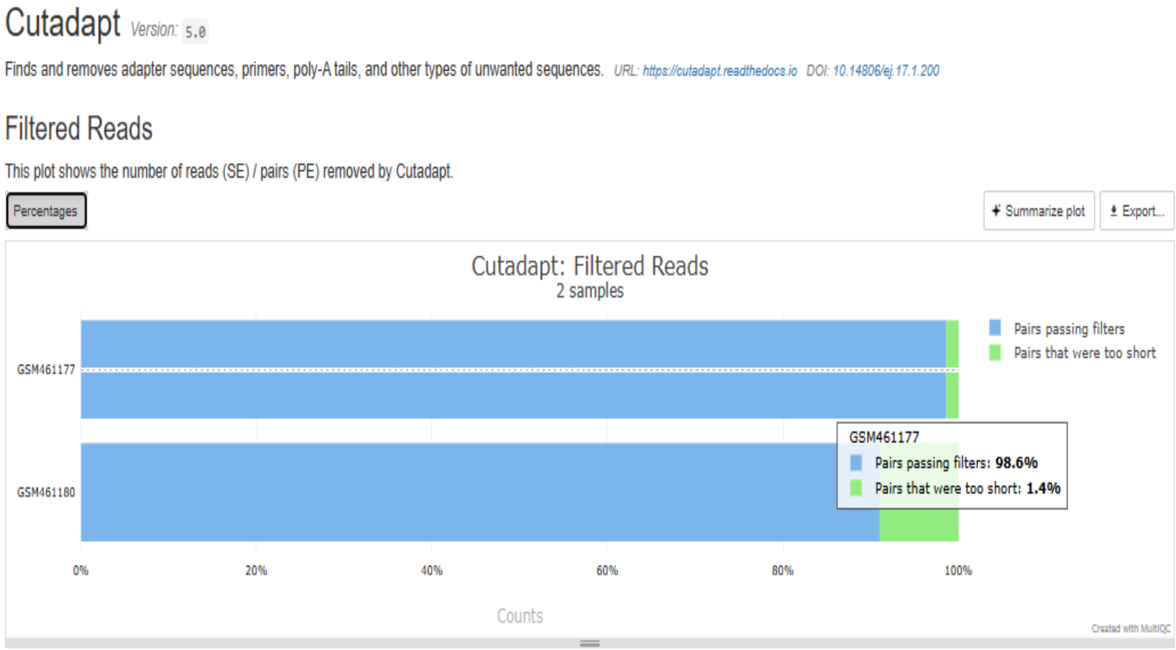
Quality control tool for high throughput sequencing data. *URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc*

Sequence Counts

Sequence counts for each sample. Duplicate read counts are an estimate only.

All reads except GSM461180_reverse shows a high rate of duplicated reads, which is normal for RNA-Seq data.

After cutadapt



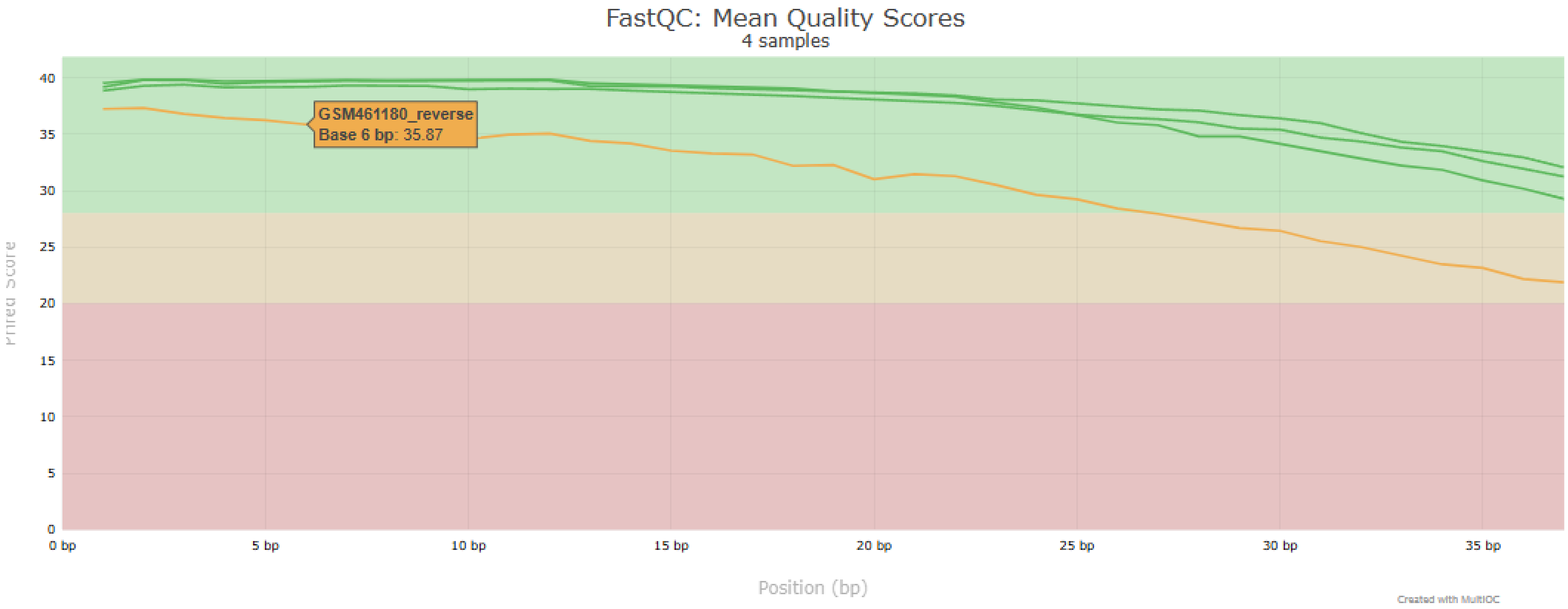
Cutadapt *Version:* 5.0

Finds and removes adapter sequences, primers, poly-A tails, and other types of unwanted sequences. *URL: https://cutadapt.readthedocs.io DOI: 10.14806/ej.17.1.200*

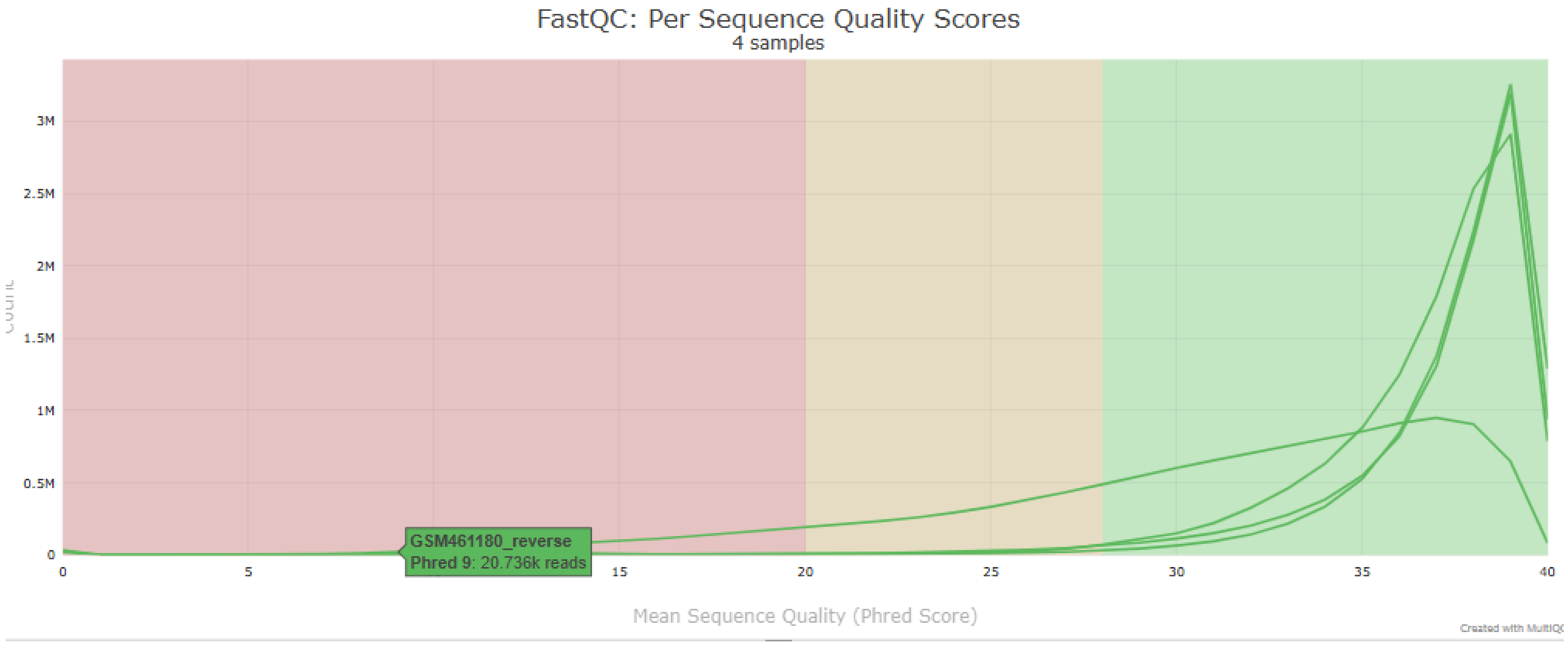
Filtered Reads

This plot shows the number of reads (SE) / pairs (PE) removed by Cutadapt.

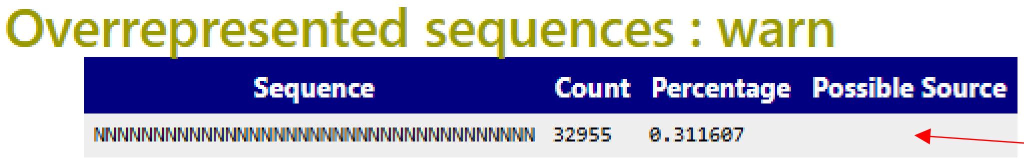
A small portion of reads were too short, with 1.4% in GSM461177 and a higher 9% in GSM461180.



The mean quality score across the reads is generally high, though there's a slight variation in the distribution for GSM461180_reverse.



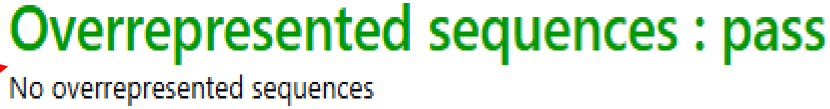
The Per base sequence quality is good overall, though there’s a sharp drop toward the end of most reads, with a noticeably sharper decline in GSM461180_reverse.



Overrepresented sequences : warn

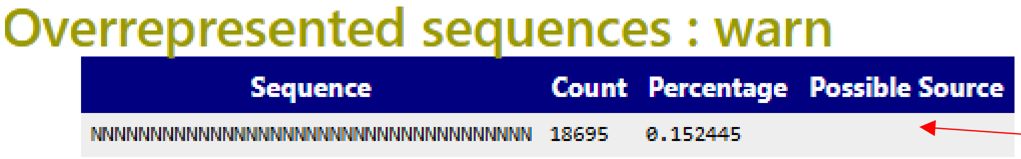
Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	32955	0.311607	

GSM461177
reverse



Overrepresented sequences : pass

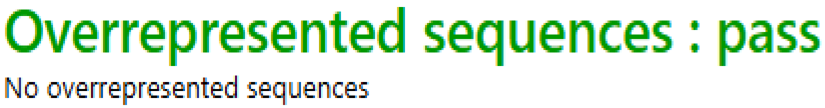
No overrepresented sequences



Overrepresented sequences : warn

Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	18695	0.152445	

GSM461180
reverse



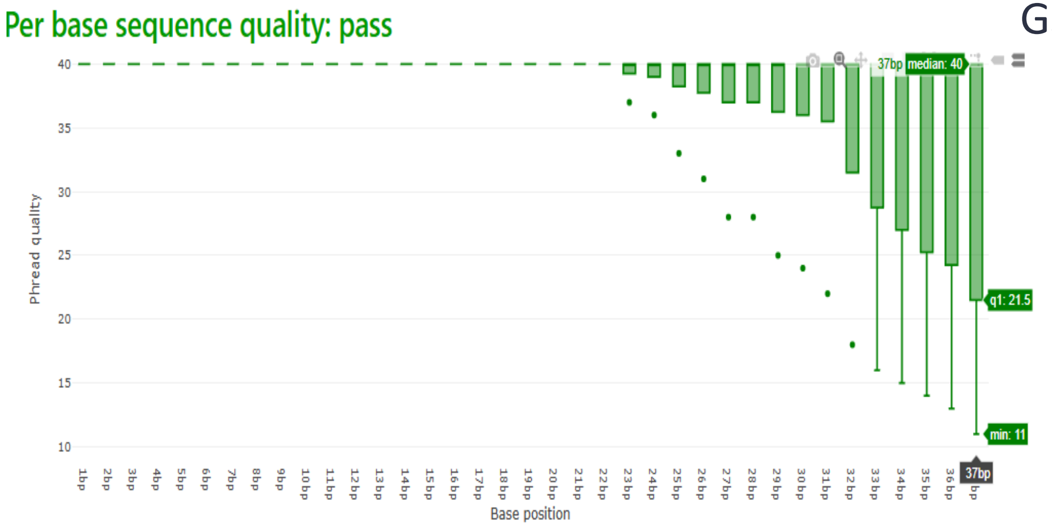
Overrepresented sequences : pass

No overrepresented sequences

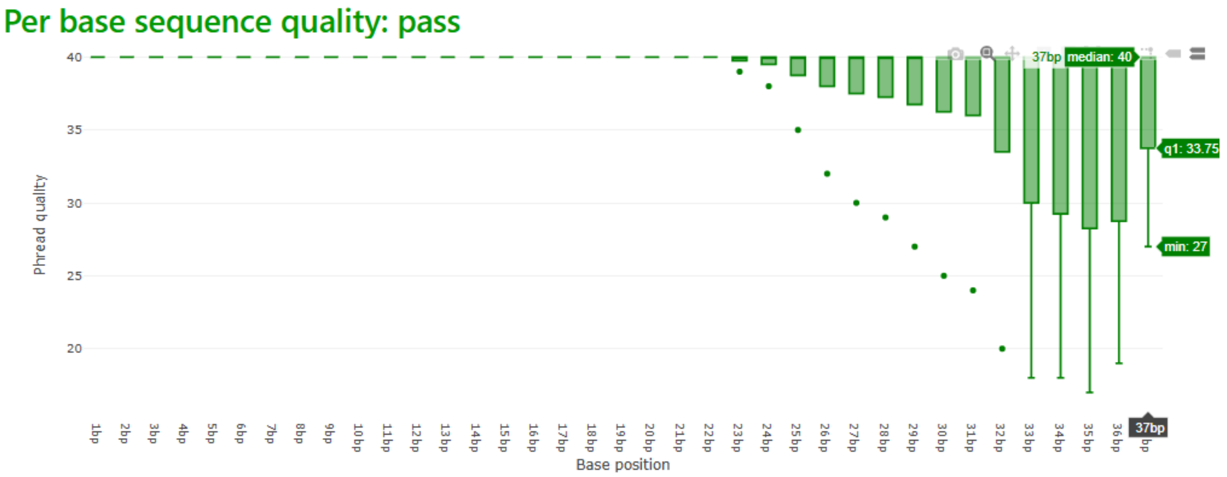
Before Cutadapt

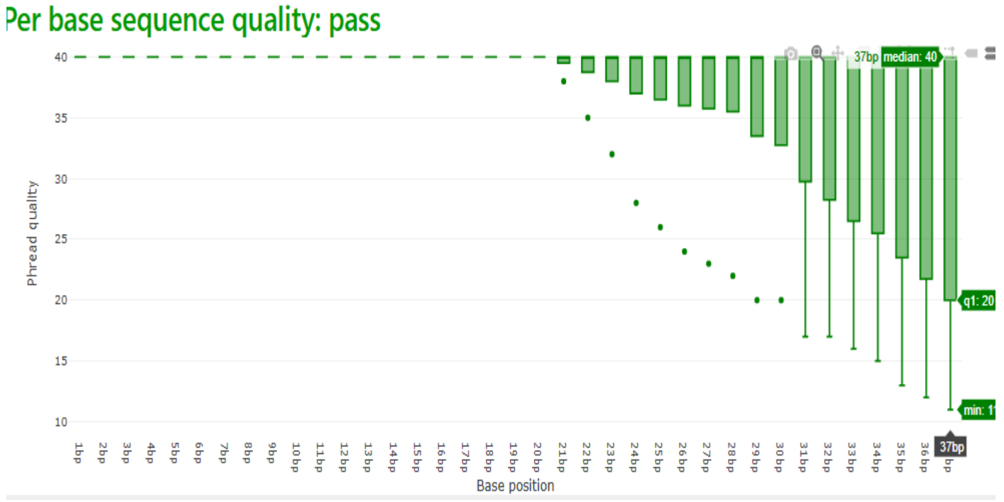
After Cutadapt

GSM461177

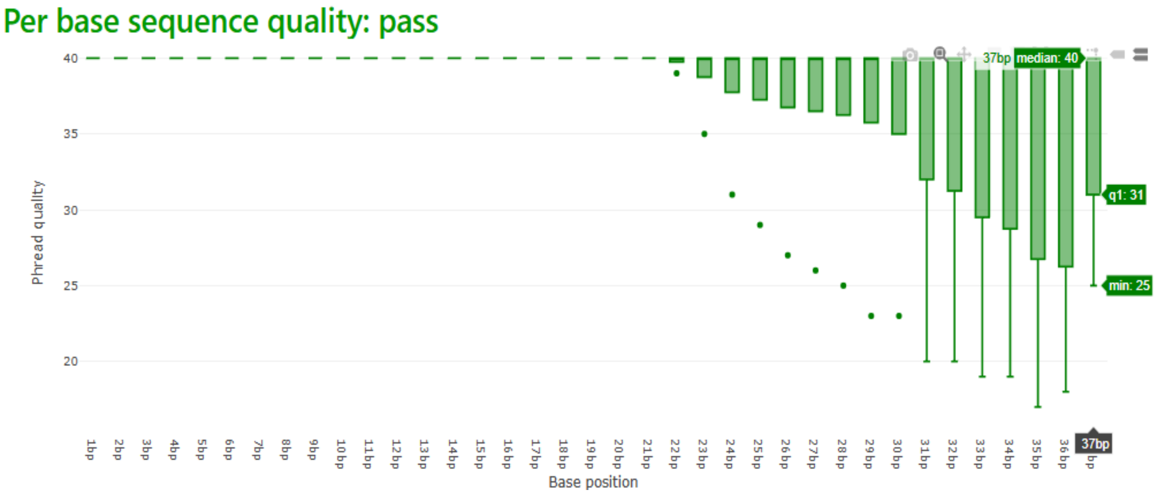


Forward





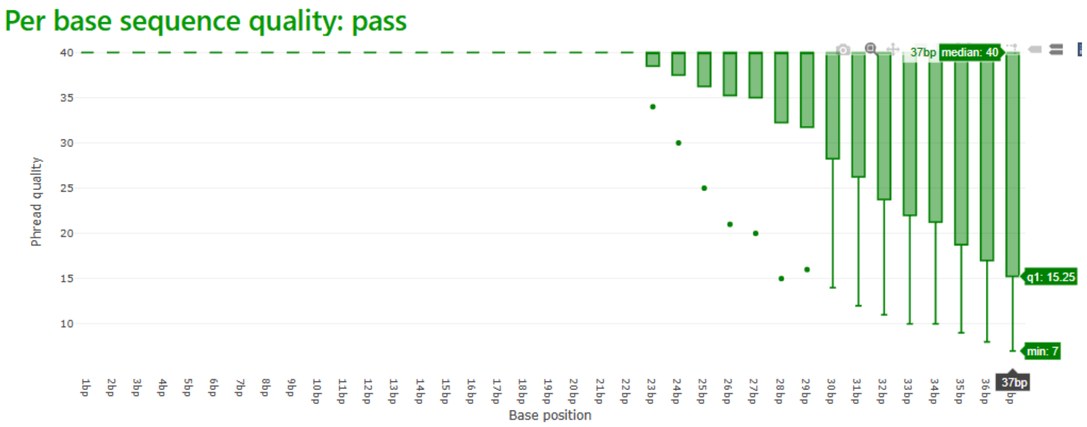
Reverse



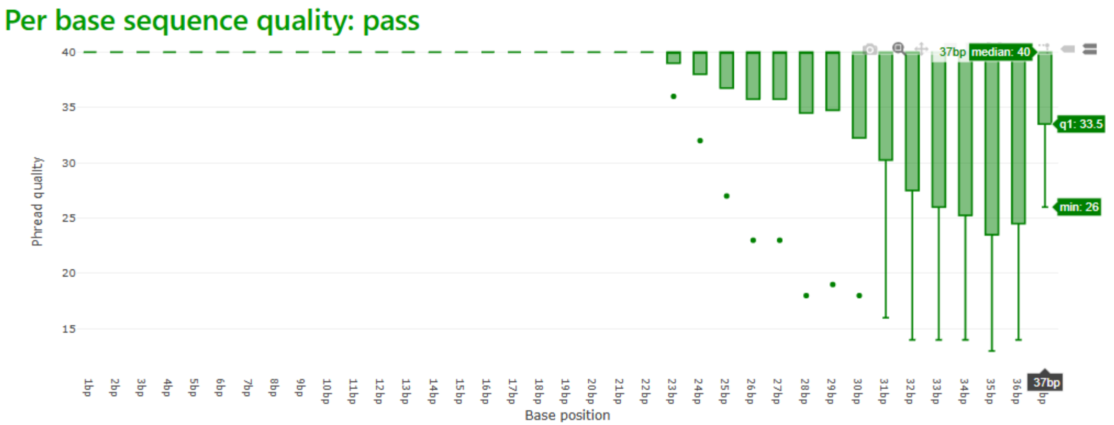
Before cutadapt

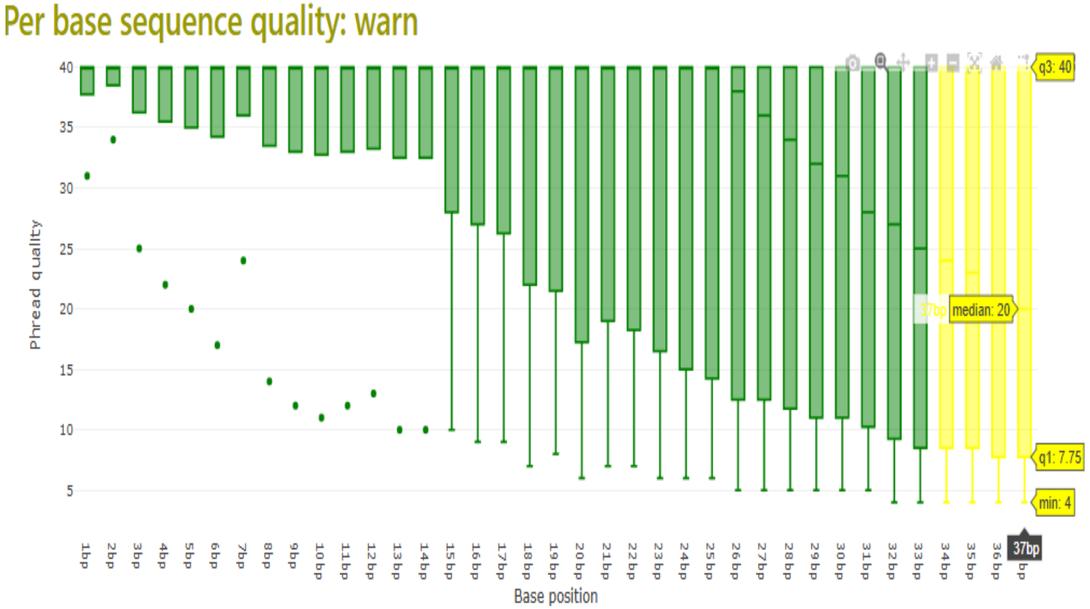
After cutadapt

GSM461180

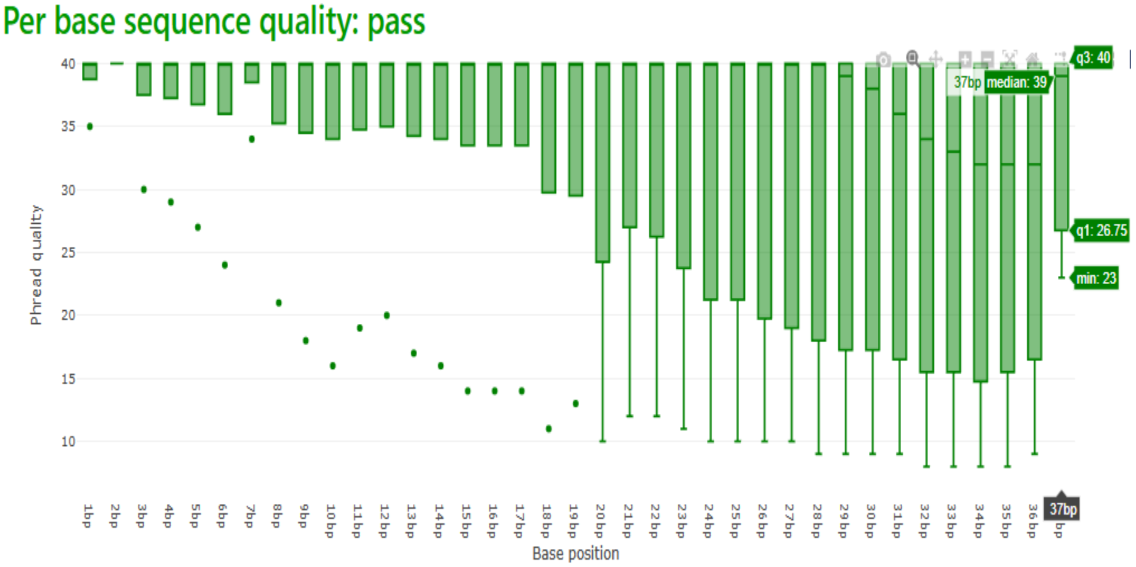


Forward





Reverse

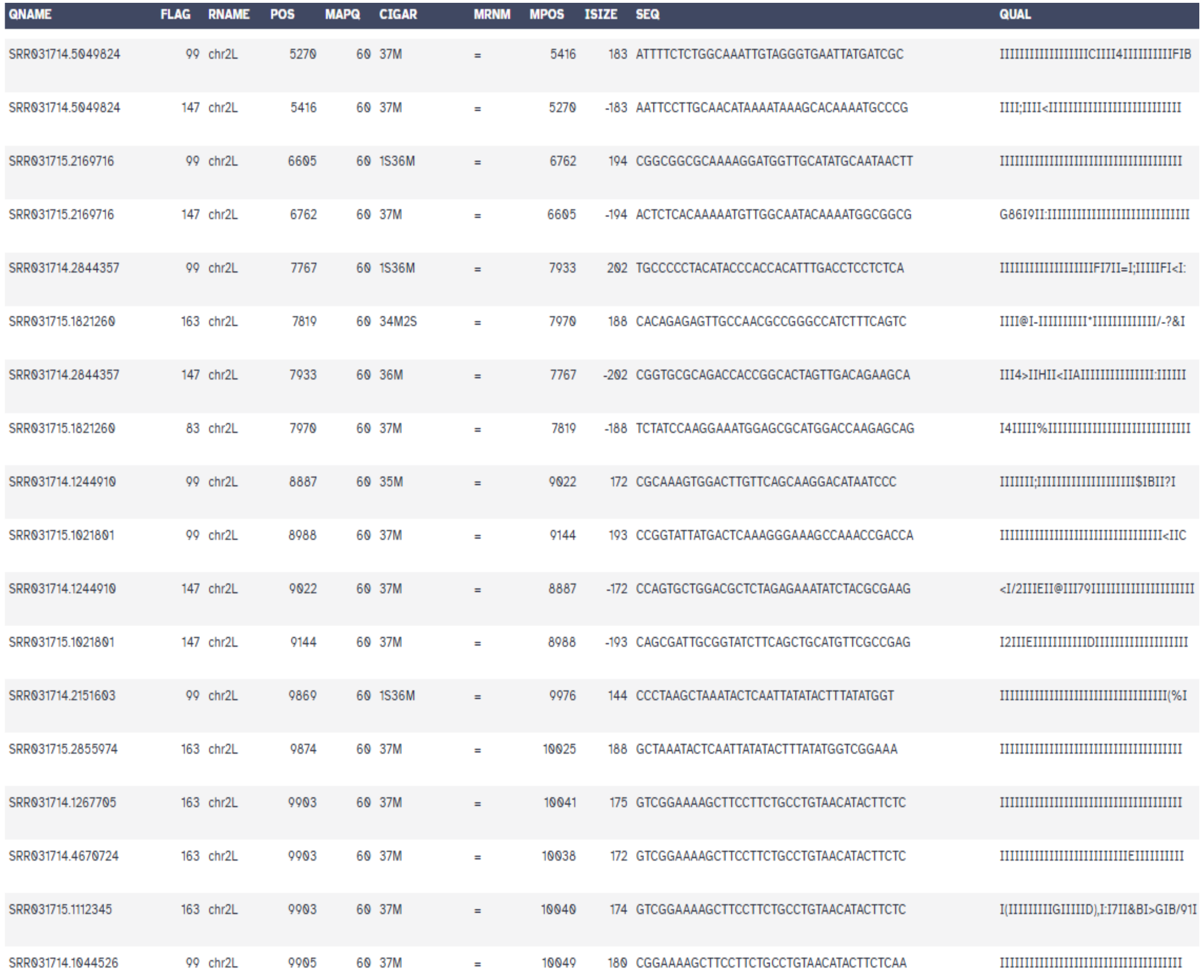


Before Cutadapt

After Cutadapt

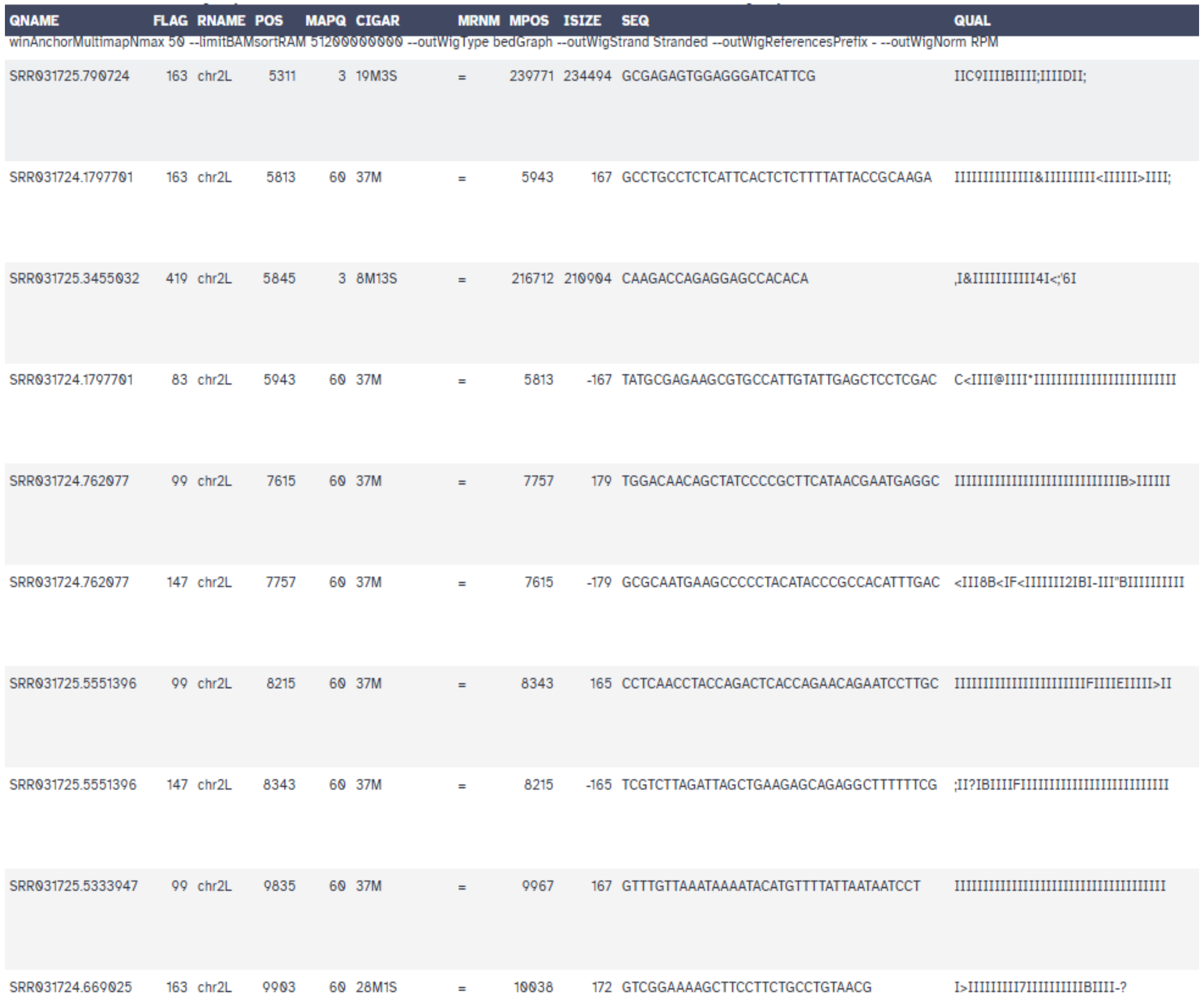
Q2 B)

RNA STAR BAM file



QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	ISIZE	SEQ	QUAL
SRR031714.5049824	99	chr2L	5270	60	37M	=	5416	183	ATTTTCTCTGGCAAATTGTAGGGTGAATTATGATCGC	IIIIIIIIIIIIIIIIICIIII4IIIIIIIIIIFIB
SRR031714.5049824	147	chr2L	5416	60	37M	=	5270	-183	AATTCCTTGCAACATAAAATAAAGCACAAAATGCCCG	IIII;IIII<IIIIIIIIIIIIIIIIIIIIIIIIIII
SRR031715.2169716	99	chr2L	6605	60	1S36M	=	6762	194	CGGCGGCGCAAAAGGATGGTTGCATATGCAATAACTT	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
SRR031715.2169716	147	chr2L	6762	60	37M	=	6605	-194	ACTCTCACAAAAATGTTGGCAATACAAAATGGCGGCG	G86I9II:IIIIIIIIIIIIIIIIIIIIIIIIIIIII
SRR031714.2844357	99	chr2L	7767	60	1S36M	=	7933	202	TGCCCCCTACATACCCACCACATTTGACCTCCTCTCA	IIIIIIIIIIIIIIIIIIFI7II=I;IIIIIFI<I:
SRR031715.1821260	163	chr2L	7819	60	34M2S	=	7970	188	CACAGAGAGTTGCCAACGCCGGGCCATCTTTCAGTC	IIII@I-IIIIIIIII*IIIIIIIIIIII/-?&I
SRR031714.2844357	147	chr2L	7933	60	36M	=	7767	-202	CGGTGCGCAGACCACCGGCACTAGTTGACAGAAGCA	III4>IIHII<IIAIIIIIIIIIIIII:IIIIII
SRR031715.1821260	83	chr2L	7970	60	37M	=	7819	-188	TCTATCCAAGGAAATGGAGCGCATGGACCAAGAGCAG	I4IIIII%IIIIIIIIIIIIIIIIIIIIIIIIIIIII
SRR031714.1244910	99	chr2L	8887	60	35M	=	9022	172	CGCAAAGTGGACTTGTTCAGCAAGGACATAATCCC	IIIIII;IIIIIIIIIIIIIIIIIII$IBII?I
SRR031715.1021801	99	chr2L	8988	60	37M	=	9144	193	CCGGTATTATGACTCAAAGGGAAAGCCAAACCGACCA	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII<IIC
SRR031714.1244910	147	chr2L	9022	60	37M	=	8887	-172	CCAGTGCTGGACGCTCTAGAGAAATATCTACGCGAAG	<I/2IIIEII@III79IIIIIIIIIIIIIIIIIIII
SRR031715.1021801	147	chr2L	9144	60	37M	=	8988	-193	CAGCGATTGCGGTATCTTCAGCTGCATGTTCGCCGAG	I2IIIEIIIIIIIIIIDIIIIIIIIIIIIIIIIIII
SRR031714.2151603	99	chr2L	9869	60	1S36M	=	9976	144	CCCTAAGCTAAATACTCAATTATATACTTTATATGGT	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII(%I
SRR031715.2855974	163	chr2L	9874	60	37M	=	10025	188	GCTAAATACTCAATTATATACTTTATATGGTCGGAAA	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
SRR031714.1267705	163	chr2L	9903	60	37M	=	10041	175	GTCGGAAAAGCTTCCTTCTGCCTGTAACATACTTCTC	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
SRR031714.4670724	163	chr2L	9903	60	37M	=	10038	172	GTCGGAAAAGCTTCCTTCTGCCTGTAACATACTTCTC	IIIIIIIIIIIIIIIIIIIIIIIIIIEIIIIIIIIII
SRR031715.1112345	163	chr2L	9903	60	37M	=	10040	174	GTCGGAAAAGCTTCCTTCTGCCTGTAACATACTTCTC	I(IIIIIIIIGIIIIID),I:I7II&BI>GIB/91I
SRR031714.1044526	99	chr2L	9905	60	37M	=	10049	180	CGGAAAAGCTTCCTTCTGCCTGTAACATACTTCTCAA	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

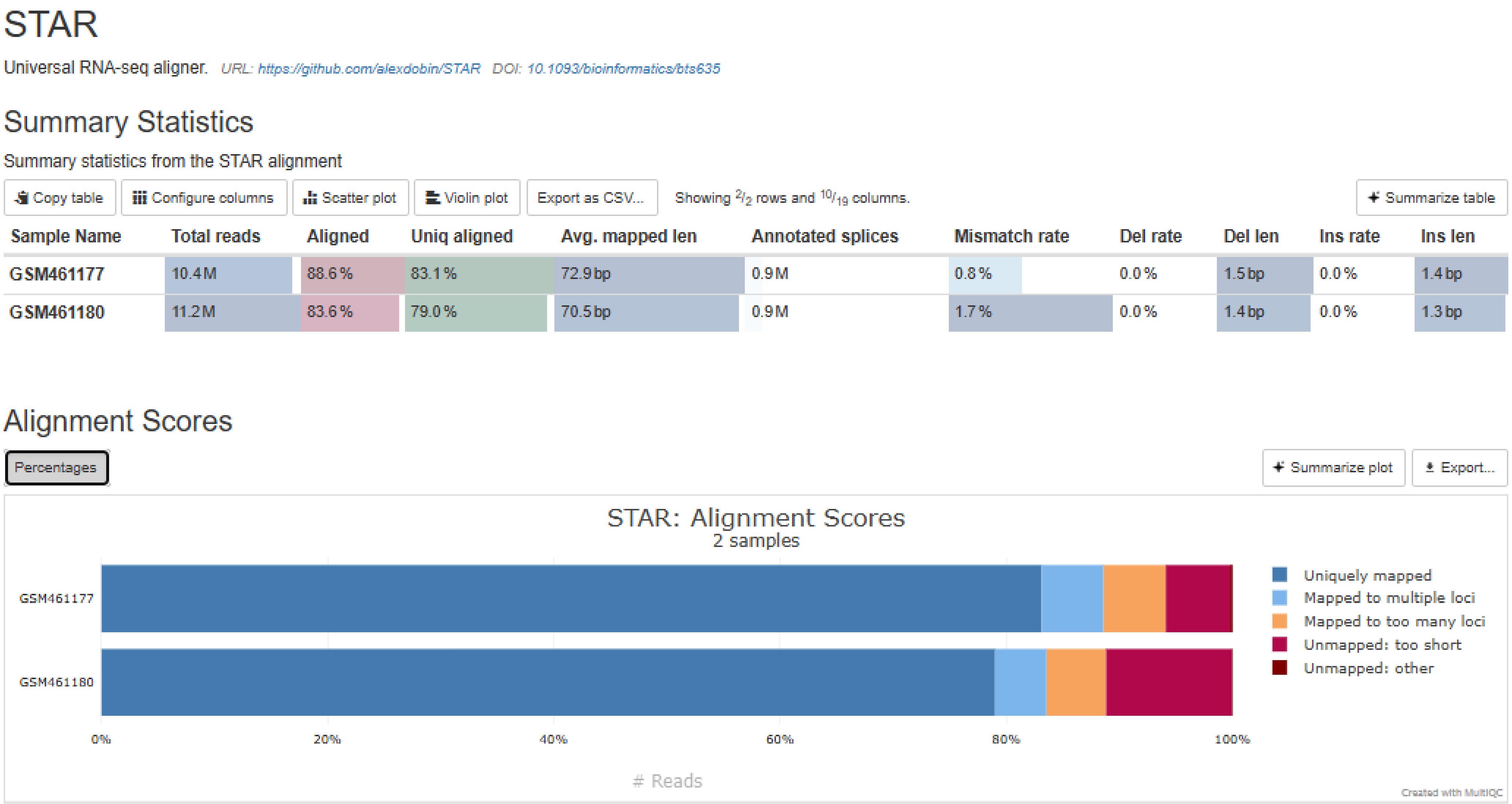
GSM461177
(untreated_paired)



winAnchorMultimapNmax 50 --limitBAMsortRAM 51200000000 --outWigType bedGraph --outWigStrand Stranded --outWigReferencesPrefix - --outWigNorm RPM

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	ISIZE	SEQ	QUAL
SRR031725.790724	163	chr2L	5311	3	19M3S	=	239771	234494	GCGAGAGTGGAGGGATCATTCG	IIC9IIIIBIIII;IIIIDII;
SRR031724.1797701	163	chr2L	5813	60	37M	=	5943	167	GCCTGCCTCTCATTCACTCTCTTTTATTACCGCAAGA	IIIIIIIIIIIII&IIIIIIIII<IIIIII>IIII;
SRR031725.3455032	419	chr2L	5845	3	8M13S	=	216712	210904	CAAGACCAGAGGAGCCACACA	,I&IIIIIIIIIII4I<;'6I
SRR031724.1797701	83	chr2L	5943	60	37M	=	5813	-167	TATGCGAGAAGCGTGCCATTGTATTGAGCTCCTCGAC	C<IIII@IIII*IIIIIIIIIIIIIIIIIIIIIIII
SRR031724.762077	99	chr2L	7615	60	37M	=	7757	179	TGGACAACAGCTATCCCCGCTTCATAACGAATGAGGC	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIB>IIIIII
SRR031724.762077	147	chr2L	7757	60	37M	=	7615	-179	GCGCAATGAAGCCCCCTACATACCCGCCACATTTGAC	<III8B<IF<IIIIIII2IBI-III"BIIIIIIIIII
SRR031725.5551396	99	chr2L	8215	60	37M	=	8343	165	CCTCAACCTACCAGACTCACCAGAACAGAATCCTTGC	IIIIIIIIIIIIIIIIIIIIIIFIIIIEIIIII>II
SRR031725.5551396	147	chr2L	8343	60	37M	=	8215	-165	TCGTCTTAGATTAGCTGAAGAGCAGAGGCTTTTTTCG	;II?IBIIIIFIIIIIIIIIIIIIIIIIIIIIIIIII
SRR031725.5333947	99	chr2L	9835	60	37M	=	9967	167	GTTTGTTAAATAAAATACATGTTTTATTAATAATCCT	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
SRR031724.669025	163	chr2L	9903	60	28M1S	=	10038	172	GTCGGAAAAGCTTCCTTCTGCCTGTAACG	I>IIIIIIII7IIIIIIIIIIBIIII-?

GSM461180
(treated_paired)



STAR

Universal RNA-seq aligner. *URL: https://github.com/alexdobin/STAR DOI: 10.1093/bioinformatics/bts635*

Summary Statistics

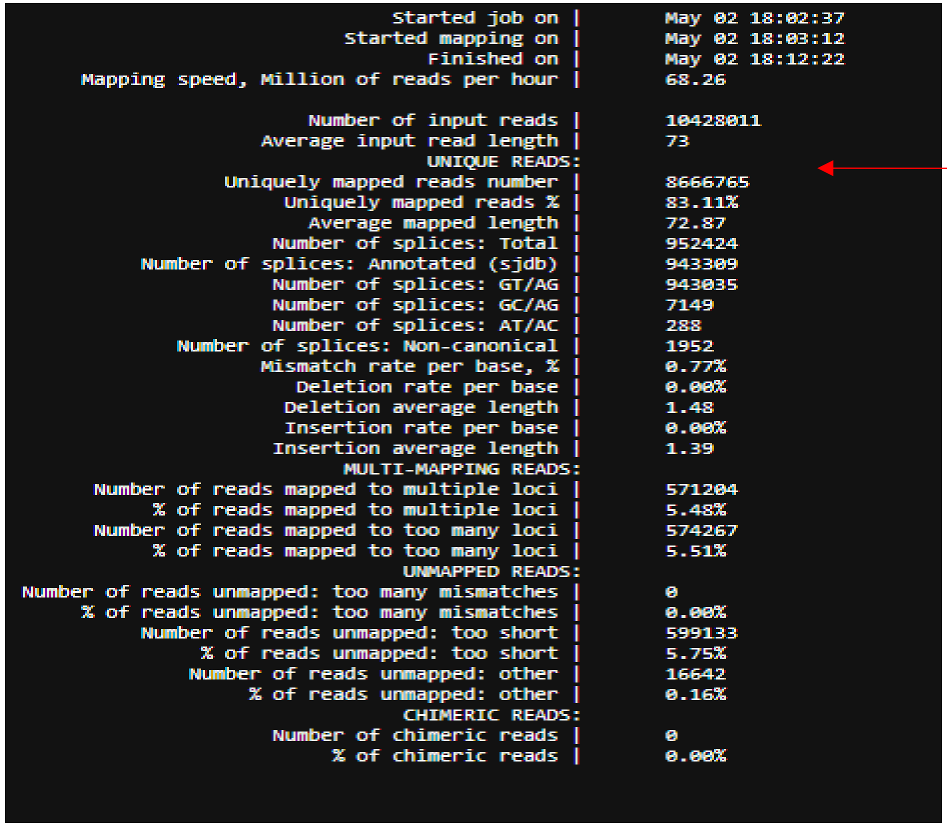
Summary statistics from the STAR alignment

Sample Name	Total reads	Aligned	Uniq aligned	Avg. mapped len	Annotated splices	Mismatch rate	Del rate	Del len	Ins rate	Ins len
GSM461177	10.4M	88.6%	83.1%	72.9bp	0.9M	0.8%	0.0%	1.5bp	0.0%	1.4bp
GSM461180	11.2M	83.6%	79.0%	70.5bp	0.9M	1.7%	0.0%	1.4bp	0.0%	1.3bp

Alignment Scores

The MultiQC report shows that around 80% of reads in both samples map uniquely to the reference genome while Less than 10% of reads are mapped to multiple location. For the treated pair 11.1% reads that were too short compared to the 5.7% of the untreated pair. The number of unmapped reads are less than 0.2% in both treated and untreated pairs

RNA STAR Logs

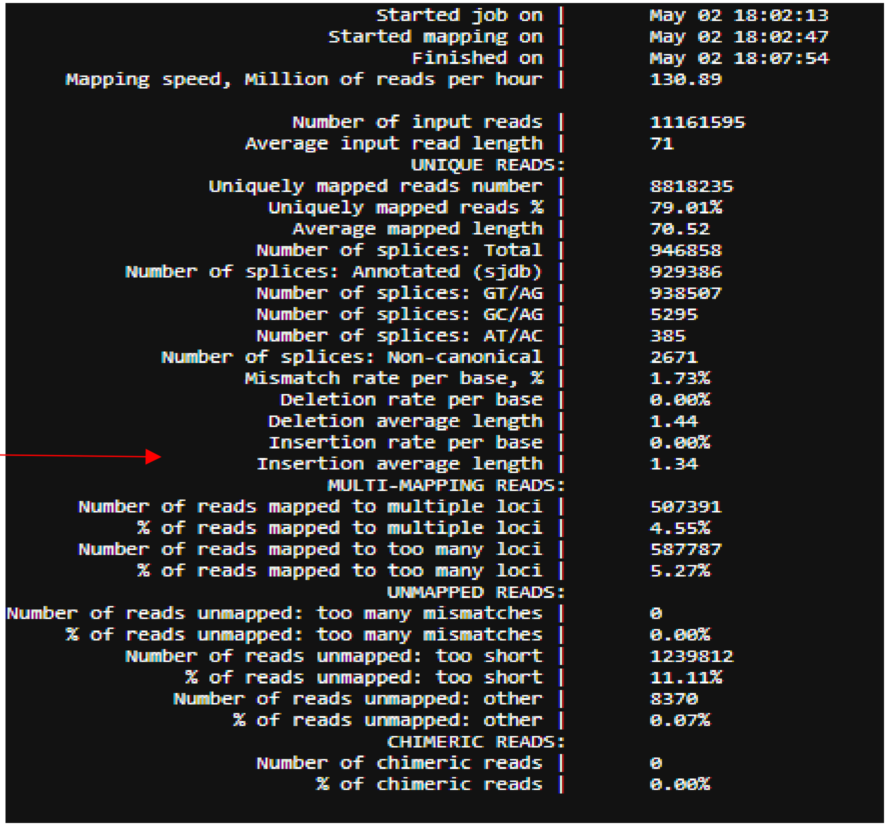


```
                                 Started job on |       May 02 18:02:37
                             Started mapping on |       May 02 18:03:12
                                    Finished on |       May 02 18:12:22
       Mapping speed, Million of reads per hour |       68.26

                          Number of input reads |       10428011
                      Average input read length |       73
                                    UNIQUE READS:
                   Uniquely mapped reads number |       8666765
                        Uniquely mapped reads % |       83.11%
                          Average mapped length |       72.87
                       Number of splices: Total |       952424
            Number of splices: Annotated (sjdb) |       943309
                       Number of splices: GT/AG |       943035
                       Number of splices: GC/AG |       7149
                       Number of splices: AT/AC |       288
               Number of splices: Non-canonical |       1952
                      Mismatch rate per base, % |       0.77%
                         Deletion rate per base |       0.00%
                        Deletion average length |       1.48
                        Insertion rate per base |       0.00%
                       Insertion average length |       1.39
                             MULTI-MAPPING READS:
        Number of reads mapped to multiple loci |       571204
             % of reads mapped to multiple loci |       5.48%
        Number of reads mapped to too many loci |       574267
             % of reads mapped to too many loci |       5.51%
                                  UNMAPPED READS:
  Number of reads unmapped: too many mismatches |       0
       % of reads unmapped: too many mismatches |       0.00%
            Number of reads unmapped: too short |       599133
                 % of reads unmapped: too short |       5.75%
                Number of reads unmapped: other |       16642
                     % of reads unmapped: other |       0.16%
                                  CHIMERIC READS:
                       Number of chimeric reads |       0
                            % of chimeric reads |       0.00%
```

GSM461177

GSM461180



```
                                 Started job on |       May 02 18:02:13
                             Started mapping on |       May 02 18:02:47
                                    Finished on |       May 02 18:07:54
       Mapping speed, Million of reads per hour |       130.89

                          Number of input reads |       11161595
                      Average input read length |       71
                                    UNIQUE READS:
                   Uniquely mapped reads number |       8818235
                        Uniquely mapped reads % |       79.01%
                          Average mapped length |       70.52
                       Number of splices: Total |       946858
            Number of splices: Annotated (sjdb) |       929386
                       Number of splices: GT/AG |       938507
                       Number of splices: GC/AG |       5295
                       Number of splices: AT/AC |       385
               Number of splices: Non-canonical |       2671
                      Mismatch rate per base, % |       1.73%
                         Deletion rate per base |       0.00%
                        Deletion average length |       1.44
                        Insertion rate per base |       0.00%
                       Insertion average length |       1.34
                             MULTI-MAPPING READS:
        Number of reads mapped to multiple loci |       507391
             % of reads mapped to multiple loci |       4.55%
        Number of reads mapped to too many loci |       587787
             % of reads mapped to too many loci |       5.27%
                                  UNMAPPED READS:
  Number of reads unmapped: too many mismatches |       0
       % of reads unmapped: too many mismatches |       0.00%
            Number of reads unmapped: too short |       1239812
                 % of reads unmapped: too short |       11.11%
                Number of reads unmapped: other |       8370
                     % of reads unmapped: other |       0.07%
                                  CHIMERIC READS:
                       Number of chimeric reads |       0
                            % of chimeric reads |       0.00%
```

The GSM461177 untreated sample has 83.11% uniquely mapped reads (8,666,765 out of 10,428,011), while the GSM461180 treated sample has 79.01% (8,818,235 out of 11,161,595), both indicating strong reliability for gene expression analysis. The untreated sample shows slightly better mapping quality, with 5.48% of reads mapping to multiple loci and 5.51% to too many loci, compared to 4.55% and 5.27% in the treated sample. However, the treated sample has a higher percentage of unmapped reads due to shortness (11.11% vs. 5.75%), suggesting potential differences in read quality or length. Both samples exhibit 0% chimeric reads meaning no fusion events were detected.

Q2 C) In Figure 1 the lines connecting the reads represent spliced alignments or junctions between the reads. These lines indicate where the RNA sequencing reads span intronic regions of the pre-mRNA that are typically removed during splicing to form mature mRNA. This is a key difference from figure 2 where such lines are absent, suggesting those reads were either unspliced or aligned without highlighting splicing events. The presence of these lines shows the use of RNA-Seq data, where STAR alignment maps reads across exon-exon junctions.

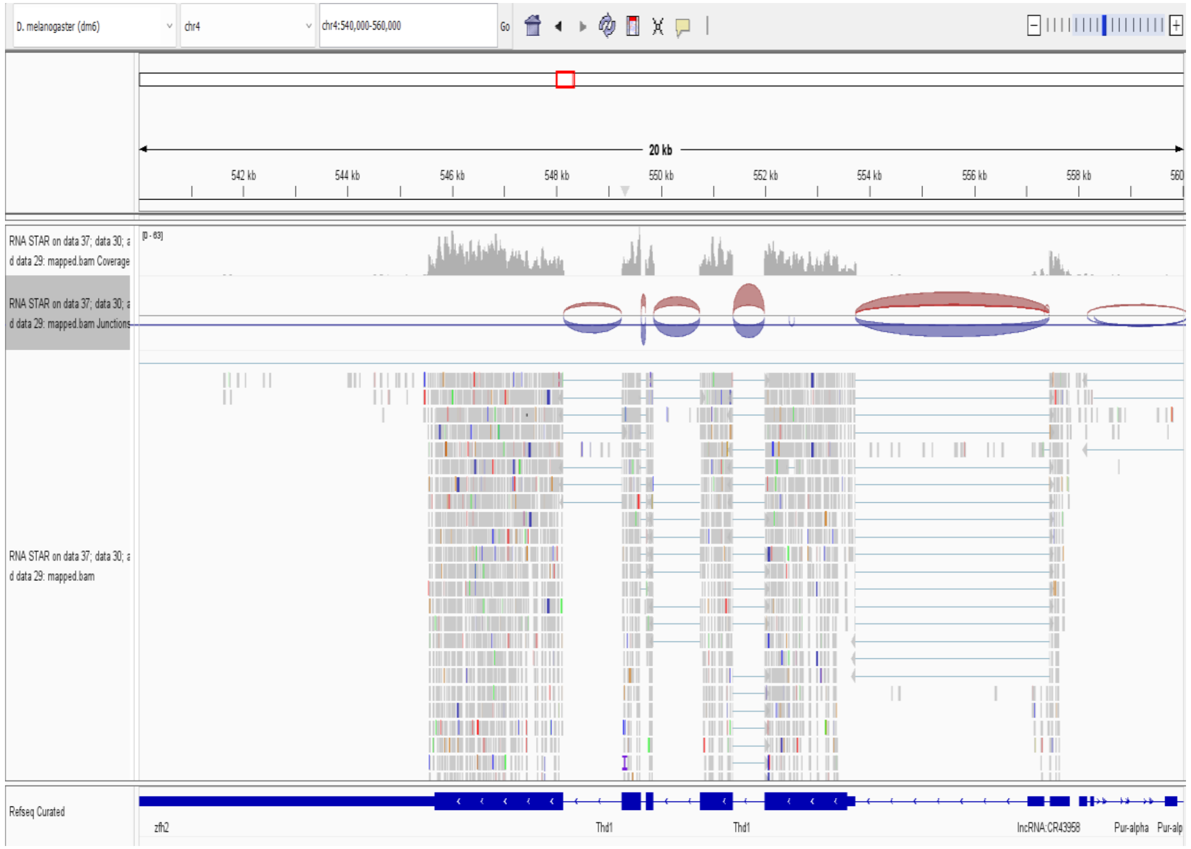

D. melanogaster (dm6)
chr4
chr4:540,000-560,000
Go
20 kb
542 kb
544 kb
546 kb
548 kb
550 kb
552 kb
554 kb
556 kb
558 kb
560
RNA STAR on data 37; data 30; and data 29: mapped.bam Coverage
[0 - 63]
RNA STAR on data 37; data 30; and data 29: mapped.bam Junctions
RNA STAR on data 37; data 30; and data 29: mapped.bam
Refseq Curated
zfh2
Thd1
Thd1
lncRNA:CR43958
Pur-alpha
Pur-alp

Figure 1 (current assignment)

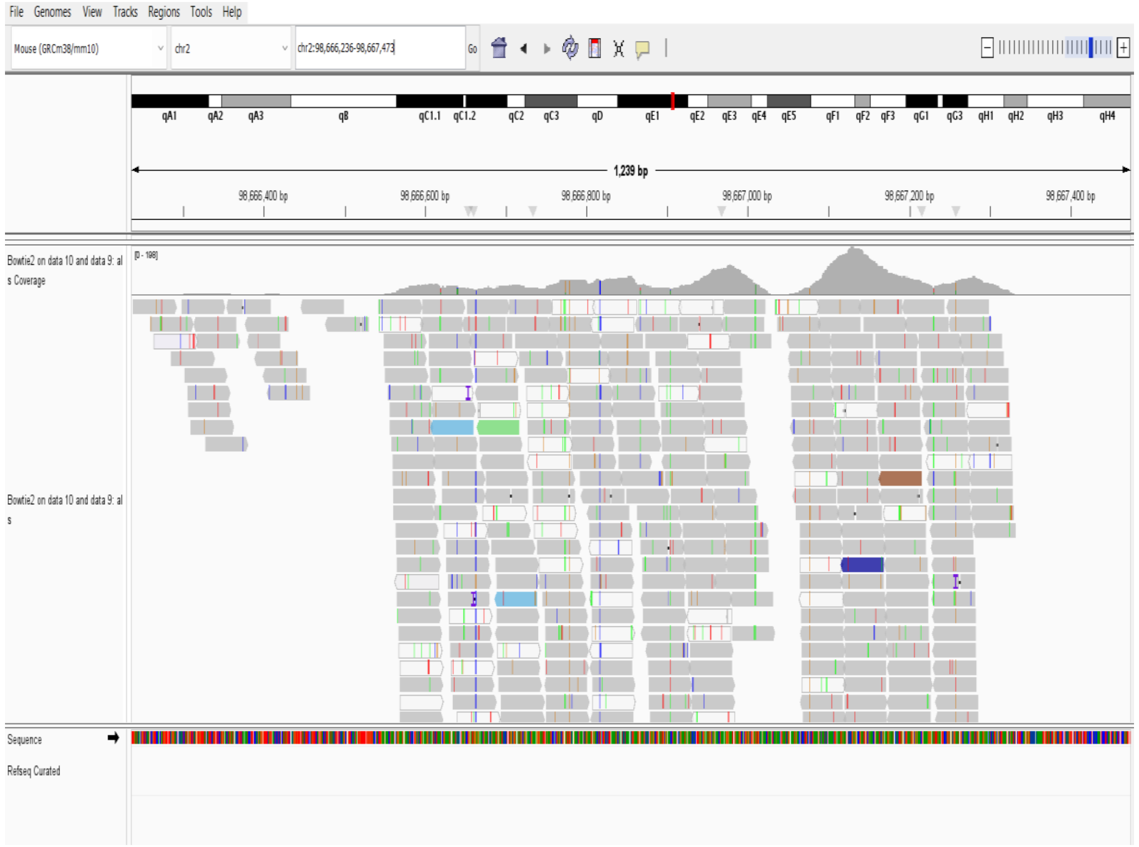
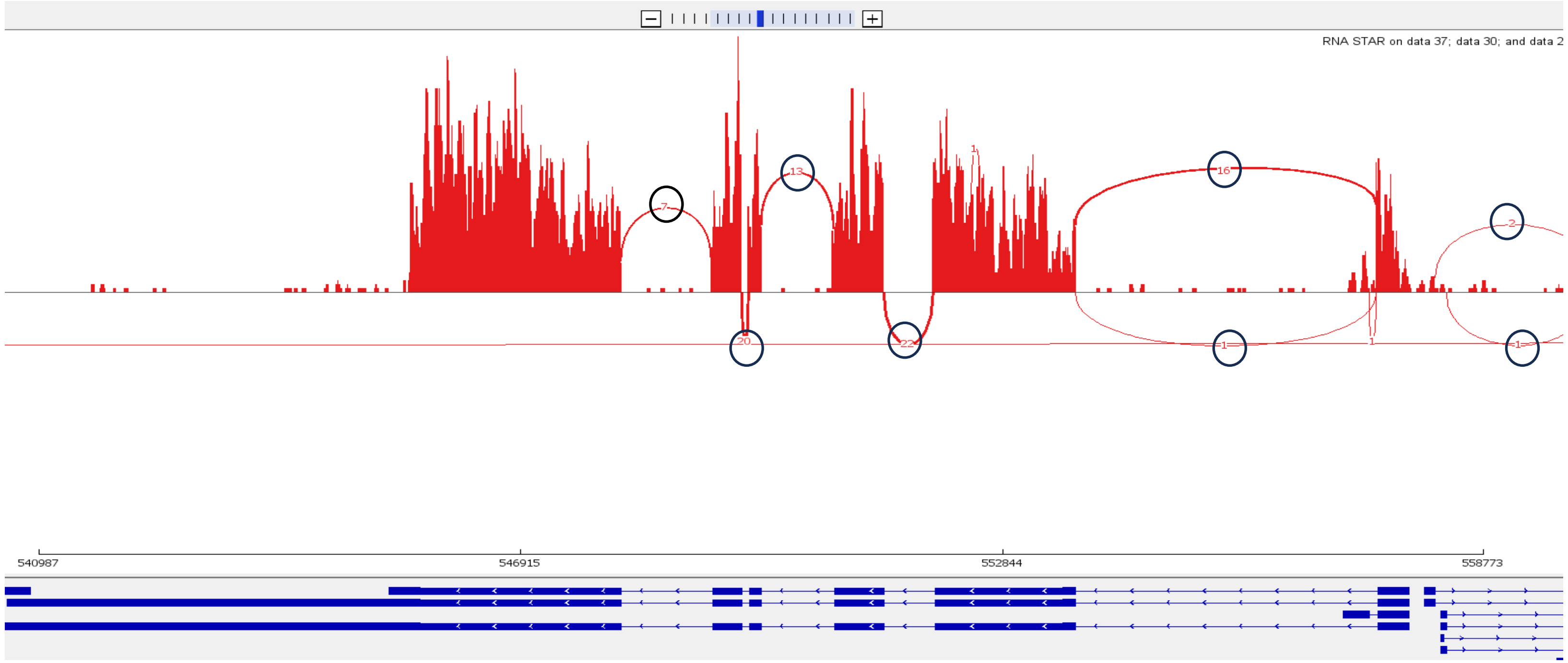


Figure 2 (old assignment)

Q2 D)

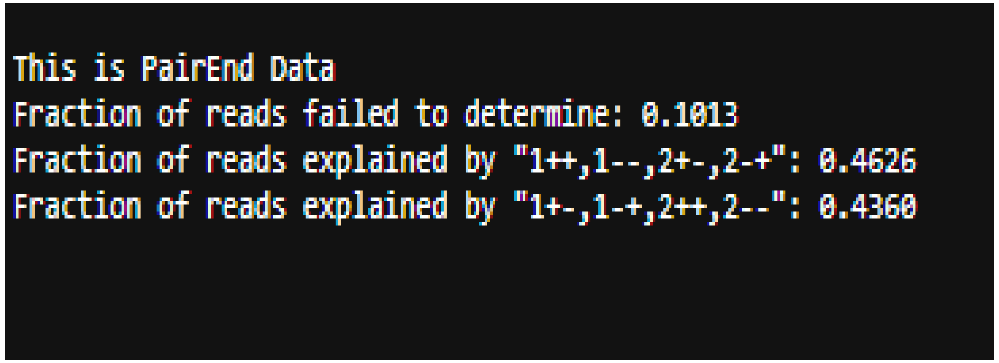
Sashimi Plot



The Sashimi plot shows splicing events across a genomic region. The lines represent splice junctions, connecting exons where reads span introns, indicating splicing events during RNA processing. The numbers on the lines [eg 7,13,16 (labeled)] denote the number of observed junction reads. Red peaks show coverage above the junctions, reflecting the depth of reads mapped to each genomic position, with gaps corresponding to introns.

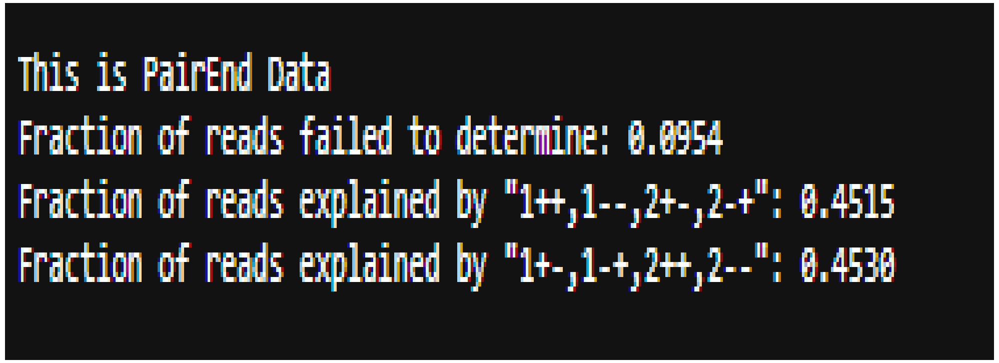
Q2 E)

Infer Experiment



```
This is PairEnd Data
Fraction of reads failed to determine: 0.1013
Fraction of reads explained by "1++,1--,2+-,2-+": 0.4626
Fraction of reads explained by "1+-,1-+,2++,2--": 0.4360
```

GSM461177 (untreated)



```
This is PairEnd Data
Fraction of reads failed to determine: 0.0954
Fraction of reads explained by "1++,1--,2+-,2-+": 0.4515
Fraction of reads explained by "1+-,1-+,2++,2--": 0.4530
```

GSM461180 (treated)

Strandedness: Strandedness in RNA-Seq indicates whether the sequencing protocol preserves the strand origin (sense or antisense) of RNA transcripts, aiding in gene expression analysis for overlapping genes, unlike unstranded protocols.

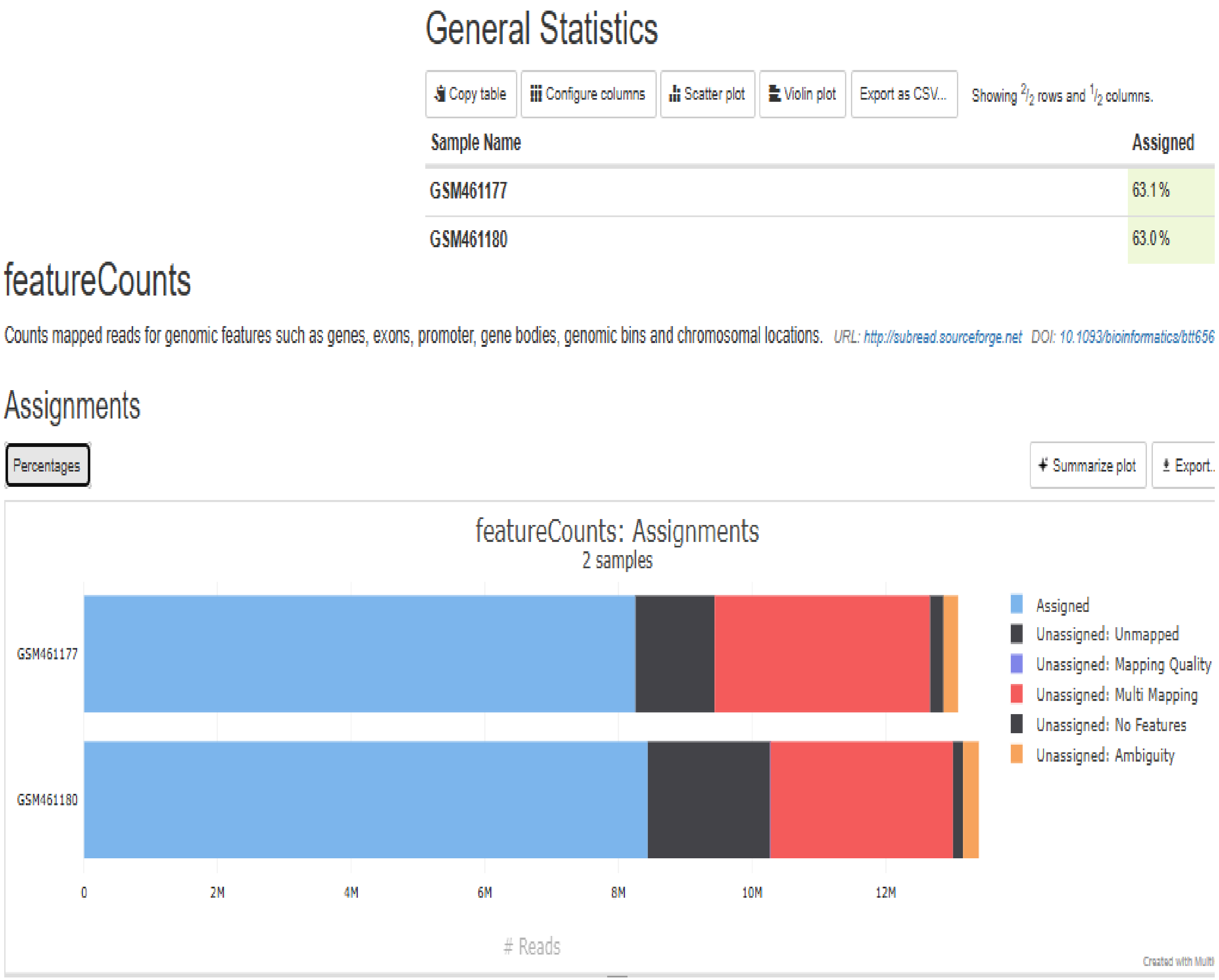
Infer Experiment Output Analysis: For GSM461177_untreated, 46.26% of reads follow "1++,1--,2+-,2-+" and 43.60% follow "1+-,1-+,2++,2--", with 10.13% undetermined. For GSM461180_treated, 45.15% follow "1++,1--,2+-,2-+" and 45.30% follow "1+-,1-+,2++,2--", with 9.54% undetermined. The balanced distribution suggests the data is likely unstranded, as a stranded library would show a strong bias toward one pattern.

Q2 F)

MultiQC on Feature Counts

The featureCounts MultiQC report shows GSM461177 (untreated) with 63.1% assigned reads (~8M) and GSM461180 (treated) with 63.0% (~8.6M), indicating consistent mapping efficiency. Assigned reads (blue) dominate, while unassigned reads (~40% including no features, unmapped, multi-mapping and ambiguity categories) suggest some reads fall outside annotated genomic regions

In short data is good (>50% mapping)



General Statistics

Showing $^2/_2$ rows and $^1/_2$ columns.

Sample Name	Assigned
GSM461177	63.1%
GSM461180	63.0%

featureCounts

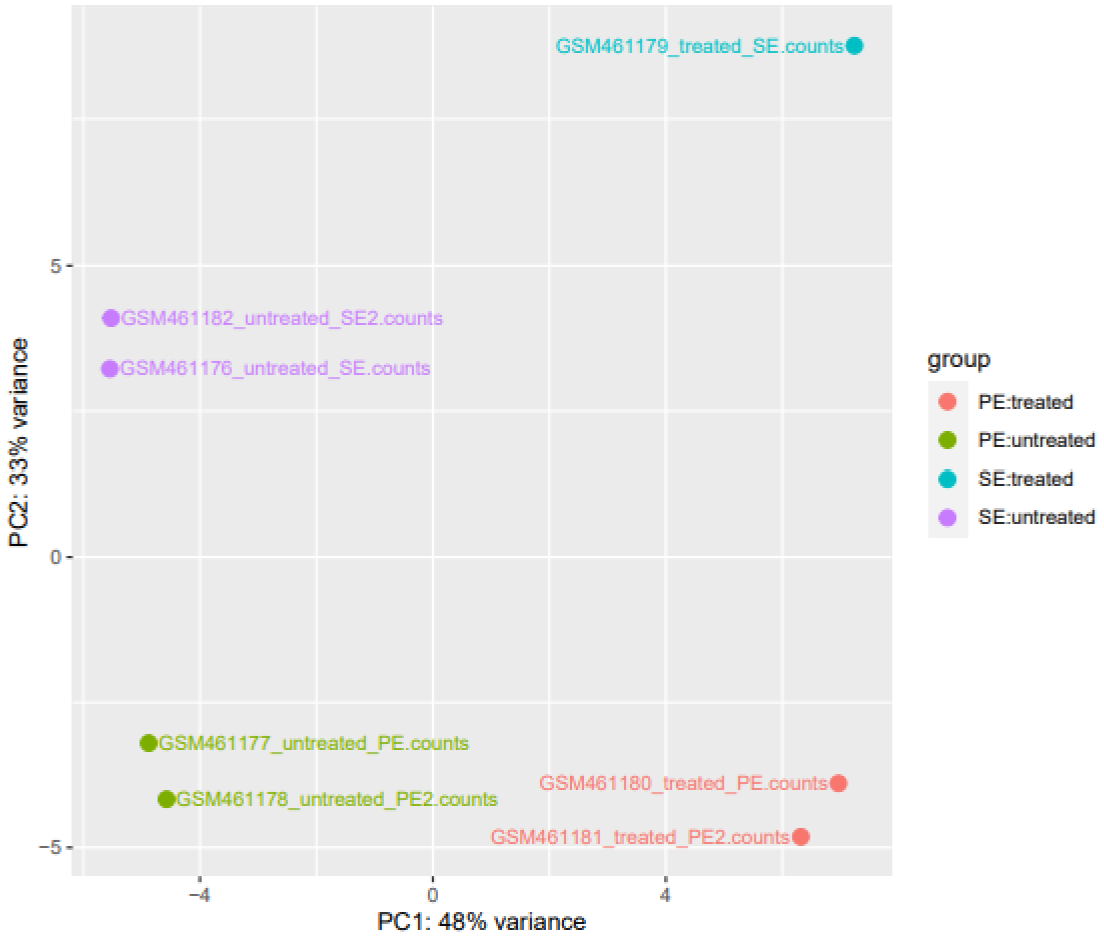
Counts mapped reads for genomic features such as genes, exons, promoter, gene bodies, genomic bins and chromosomal locations. URL: http://subread.sourceforge.net DOI: 10.1093/bioinformatics/btt656

Assignments

Q3 A)

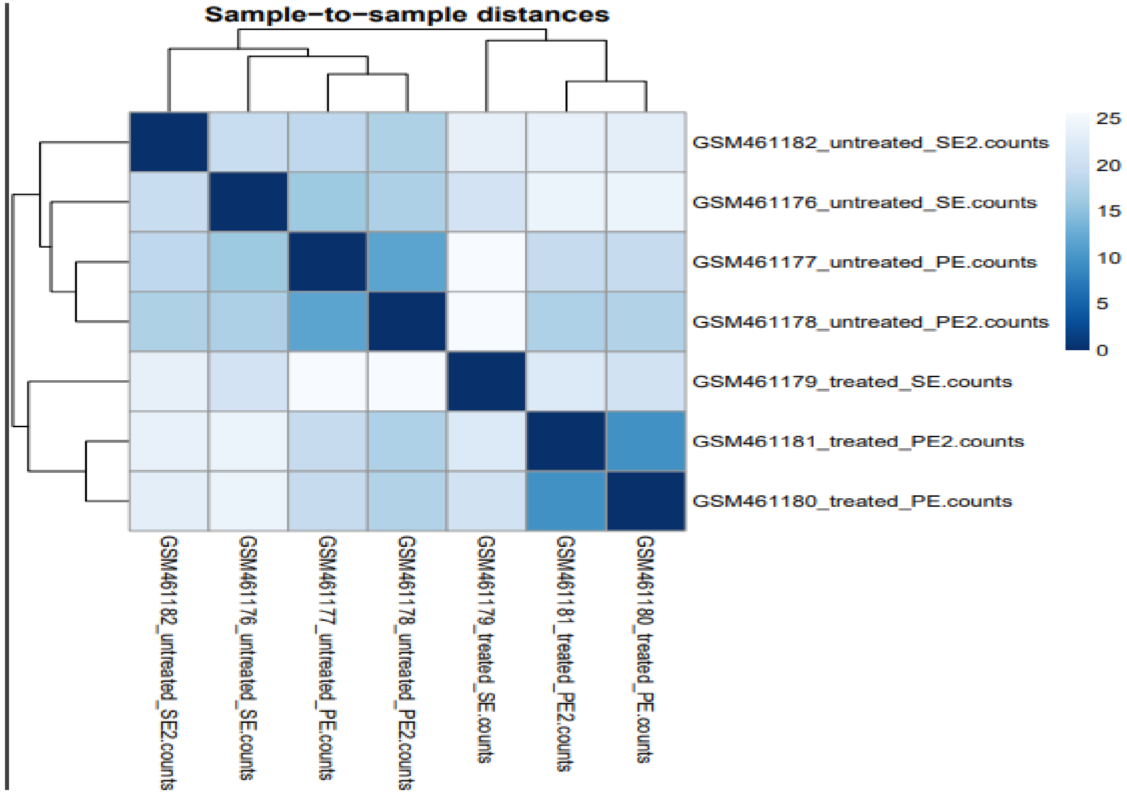
The PCA plot from DESeq2 analysis of GSM461177, GSM461178 (untreated, PE/SE), GSM461180, GSM461181 (treated, PE), and GSM461179 (treated, SE, blue dot) shows PC1 (48% variance) separating untreated (green/purple) from treated (red/blue) samples, capturing biological differences due to treatment. PC2 (33% variance) separates single-end (SE) from paired-end (PE) datasets, reflecting technical variation. Samples cluster tightly by condition and sequencing type, indicating no hidden batch effects

PCA Plot



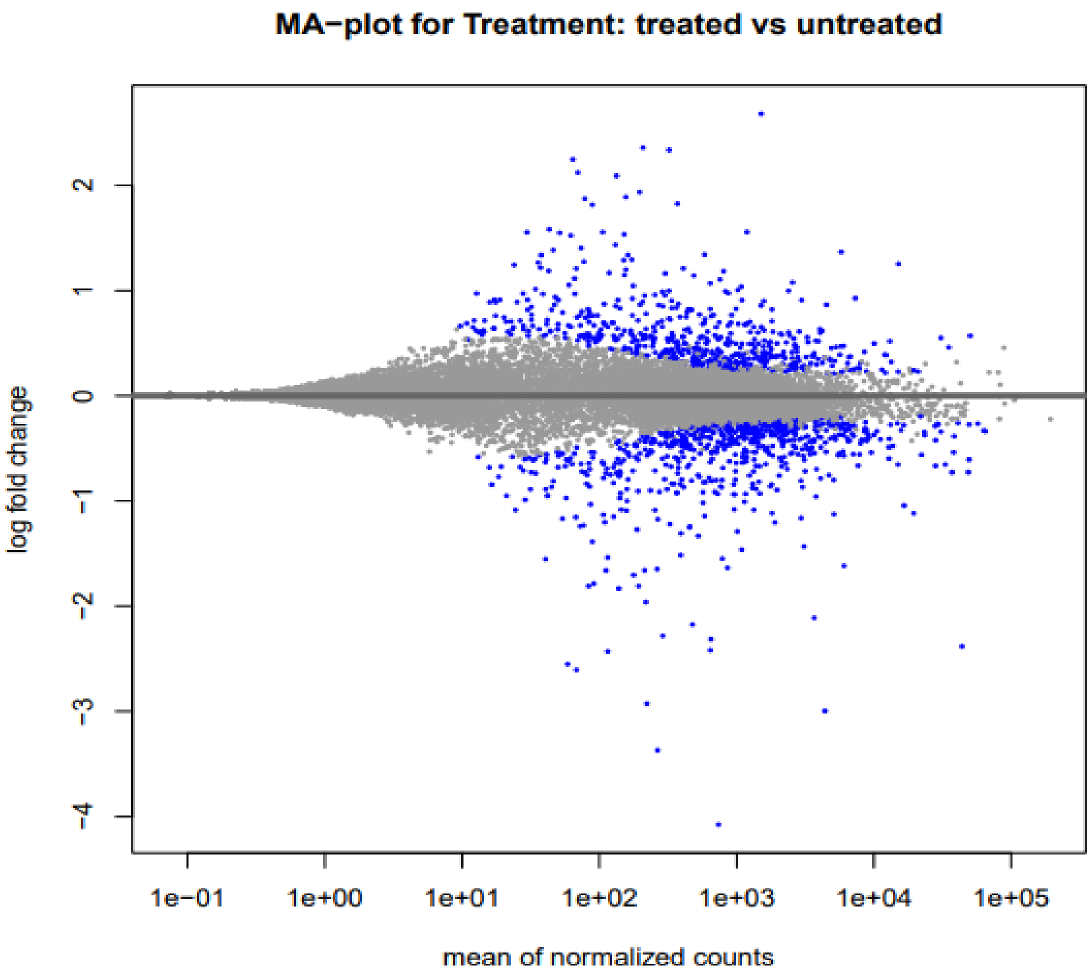
Distance Matrix Plot

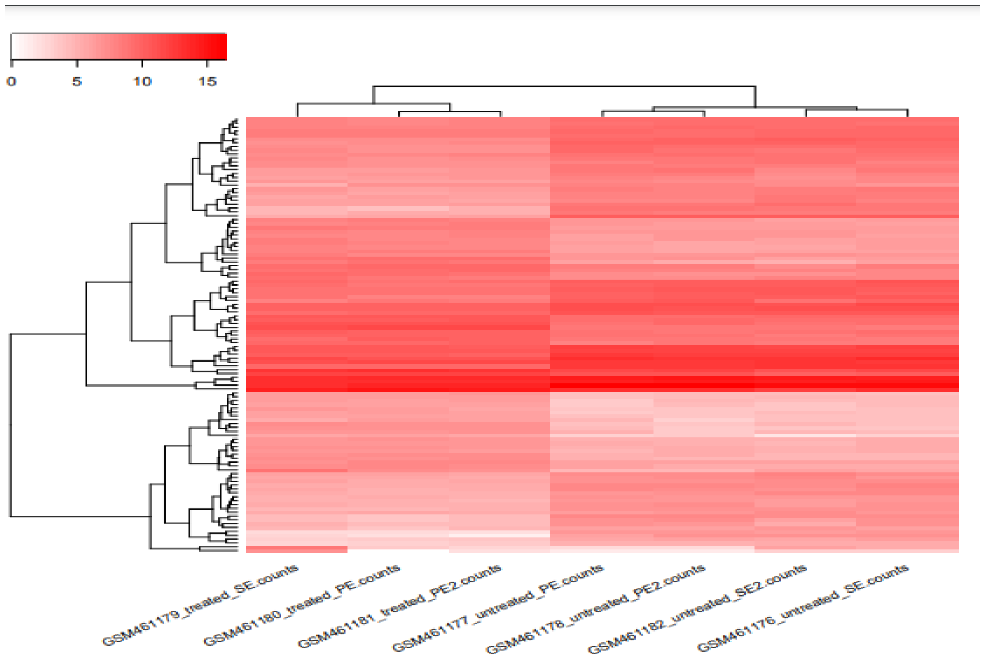
The sample-to-sample distance heatmap, based on normalized counts for GSM461177, GSM461178 (untreated, PE/SE), GSM461180, GSM461181 (treated, PE), and GSM461179 (treated, SE), shows samples grouped first by treatment (untreated vs. treated) and then by sequencing type (PE vs. SE). Darker blue blocks along the diagonal (distances near 0) indicate high similarity within groups, such as between GSM461177 and GSM461178 (untreated) or GSM461180 and GSM461181 (treated, PE), while lighter shades (distances up to 25) reflect greater differences between groups, like untreated vs. treated samples. This clear clustering confirms the separation by treatment and sequencing type, with no evident hidden batch effects.



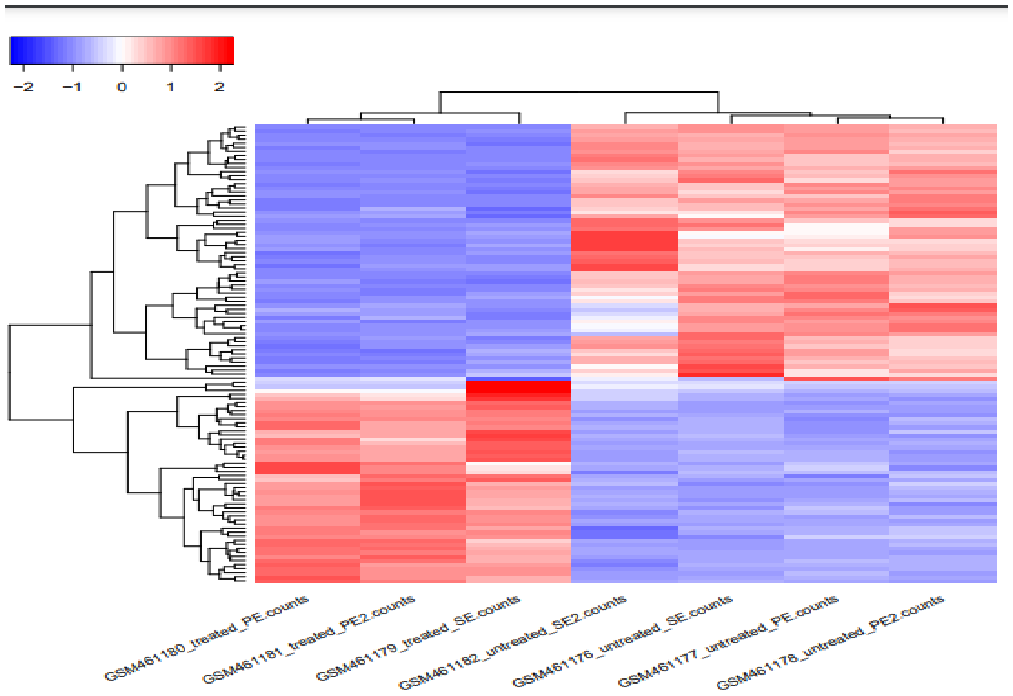
MA Plot

The y-axis (log2 fold change, -4 to 4) shows that while most genes cluster near zero (indicating no significant change), distinct groups of significantly upregulated (positive values) and downregulated (negative values) genes are visible. The x-axis (mean normalized counts, 1e-01 to 1e+05) demonstrates greater variability in fold changes among low-abundance genes, a characteristic feature of RNA-seq data. Notably, the presence of genes with strong fold changes (approaching ±4) suggests the treatment had substantial effects on specific targets.





heatmap of the normalized counts



Z-Score Visualization

A heatmap is a graphical representation of data where values are depicted as colors, making patterns in large datasets easily interpretable. In RNA-seq analysis, heatmaps display normalized counts or expression values for genes (rows) across samples (columns), with clustering revealing groups of genes with similar expression profiles.

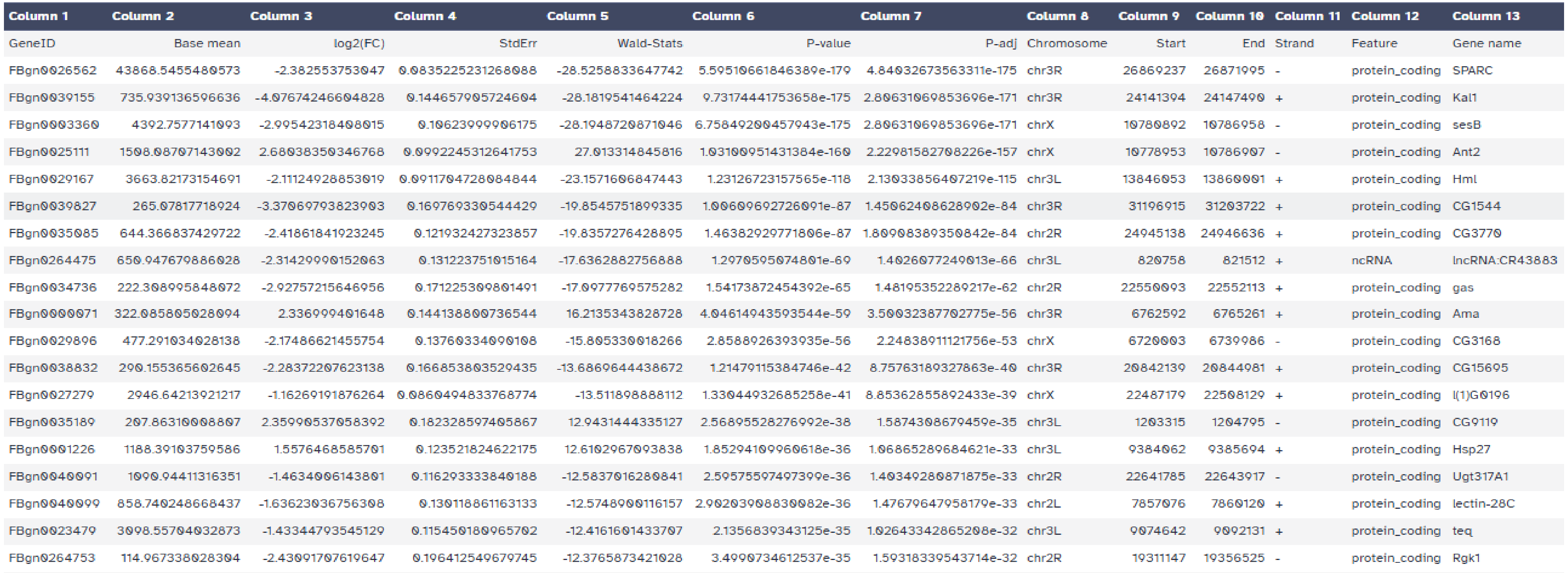
The Z-score standardizes expression by measuring how far each value deviates from the genes mean

Formula: Z = (X – mean) / standard deviation

Red = above mean (upregulated), blue = below (downregulated). This highlights relative changes, making DEG patterns clearer.

Q3 D)

Genes with significant adj p-value & abs(log2(FC)) > 1

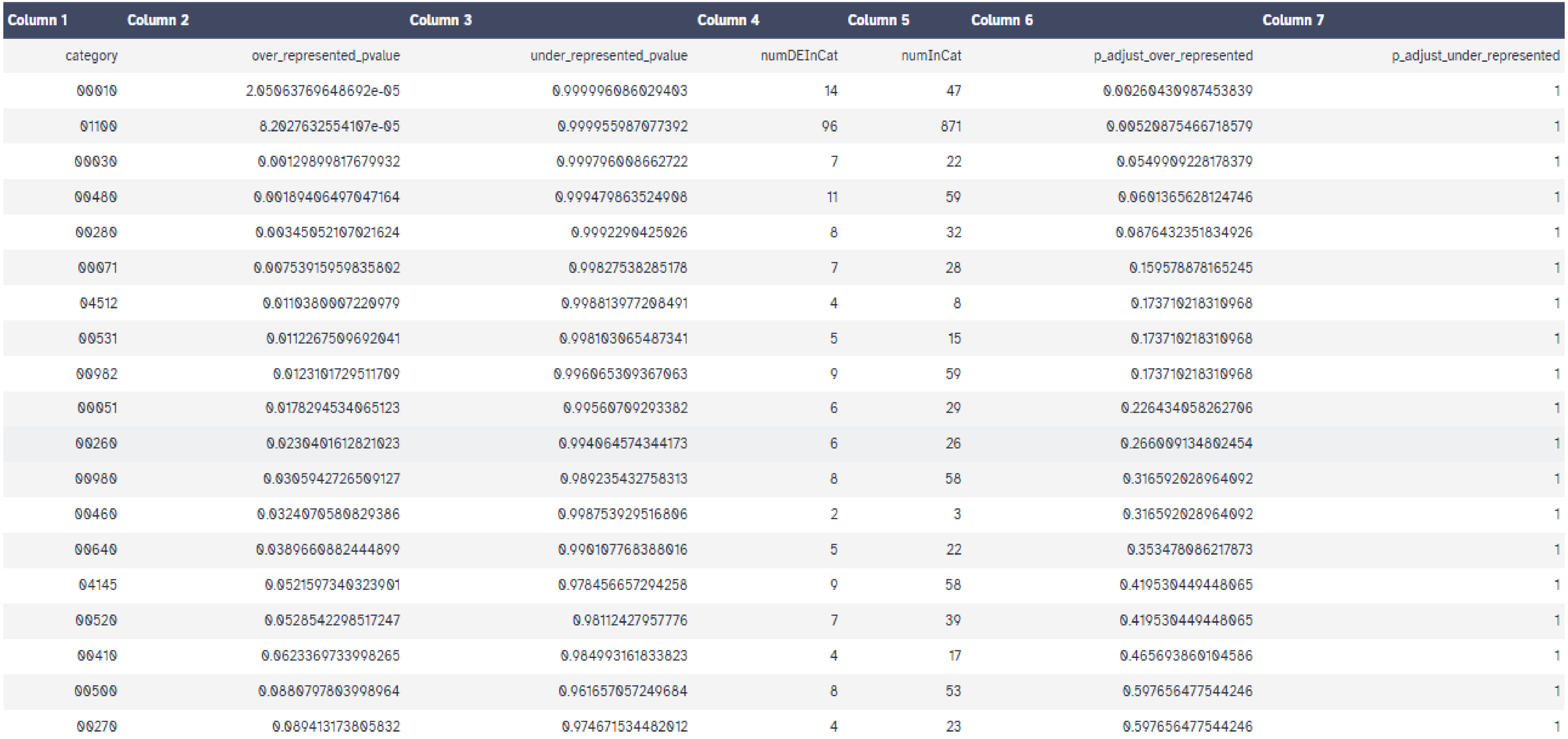


Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Column 11	Column 12	Column 13
GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value	P-adj	Chromosome	Start	End	Strand	Feature	Gene name
FBgn0026562	43868.5455480573	-2.382553753047	0.0835225231268088	-28.5258833647742	5.59510661846389e-179	4.84032673563311e-175	chr3R	26869237	26871995	-	protein_coding	SPARC
FBgn0039155	735.939136596636	-4.07674246604828	0.144657905724604	-28.1819541464224	9.73174441753658e-175	2.80631069853696e-171	chr3R	24141394	24147490	+	protein_coding	Kal1
FBgn0003360	4392.7577141093	-2.99542318408015	0.10623999906175	-28.1948720871046	6.75849200457943e-175	2.80631069853696e-171	chrX	10780892	10786958	-	protein_coding	sesB
FBgn0025111	1508.08707143002	2.68038350346768	0.0992245312641753	27.013314845816	1.03100951431384e-160	2.22981582708226e-157	chrX	10778953	10786907	-	protein_coding	Ant2
FBgn0029167	3663.82173154691	-2.11124928853019	0.0911704728084844	-23.1571606847443	1.23126723157565e-118	2.13033856407219e-115	chr3L	13846053	13860001	+	protein_coding	Hml
FBgn0039827	265.07817718924	-3.37069793823903	0.169769330544429	-19.8545751899335	1.00609692726091e-87	1.45062408628902e-84	chr3R	31196915	31203722	+	protein_coding	CG1544
FBgn0035085	644.366837429722	-2.41861841923245	0.121932427323857	-19.8357276428895	1.46382929771806e-87	1.80908389350842e-84	chr2R	24945138	24946636	+	protein_coding	CG3770
FBgn0264475	650.947679886028	-2.31429990152063	0.131223751015164	-17.6362882756888	1.2970595074801e-69	1.4026077249013e-66	chr3L	820758	821512	+	ncRNA	lncRNA:CR43883
FBgn0034736	222.308995848072	-2.92757215646956	0.171225309801491	-17.0977769575282	1.54173872454392e-65	1.48195352289217e-62	chr2R	22550093	22552113	+	protein_coding	gas
FBgn0000071	322.085805028094	2.336999401648	0.144138800736544	16.2135343828728	4.04614943593544e-59	3.50032387702775e-56	chr3R	6762592	6765261	+	protein_coding	Ama
FBgn0029896	477.291034028138	-2.17486621455754	0.13760334090108	-15.805330018266	2.8588926393935e-56	2.24838911121756e-53	chrX	6720003	6739986	-	protein_coding	CG3168
FBgn0038832	290.155365602645	-2.28372207623138	0.166853803529435	-13.6869644438672	1.21479115384746e-42	8.75763189327863e-40	chr3R	20842139	20844981	+	protein_coding	CG15695
FBgn0027279	2946.64213921217	-1.16269191876264	0.0860494833768774	-13.511898888112	1.33044932685258e-41	8.85362855892433e-39	chrX	22487179	22508129	+	protein_coding	l(1)G0196
FBgn0035189	207.86310008807	2.35990537058392	0.182328597405867	12.9431444335127	2.56895528276992e-38	1.5874308679459e-35	chr3L	1203315	1204795	-	protein_coding	CG9119
FBgn0001226	1188.39103759586	1.5576468585701	0.123521824622175	12.6102967093838	1.85294109960618e-36	1.06865289684621e-33	chr3L	9384062	9385694	+	protein_coding	Hsp27
FBgn0040091	1090.94411316351	-1.4634006143801	0.116293333840188	-12.5837016280841	2.59575597497399e-36	1.40349280871875e-33	chr2R	22641785	22643917	-	protein_coding	Ugt317A1
FBgn0040099	858.740248668437	-1.63623036756308	0.130118861163133	-12.5748900116157	2.90203908830082e-36	1.47679647958179e-33	chr2L	7857076	7860120	+	protein_coding	lectin-28C
FBgn0023479	3098.55704032873	-1.43344793545129	0.115450180965702	-12.4161601433707	2.1356839343125e-35	1.02643342865208e-32	chr3L	9074642	9092131	+	protein_coding	teq
FBgn0264753	114.967338028304	-2.43091707619647	0.196412549679745	-12.3765873421028	3.4990734612537e-35	1.59318339543714e-32	chr2R	19311147	19356525	-	protein_coding	Rgk1

The filtered table identifies high-confidence differentially expressed genes (DEGs) having the constriants of Genes with significant adj p-value & abs(log2(FC)) > 1
Protein-coding genes dominate the significant DEGs, with several showing extreme fold-changes and statistical significance eg SPARC

Q3 D)

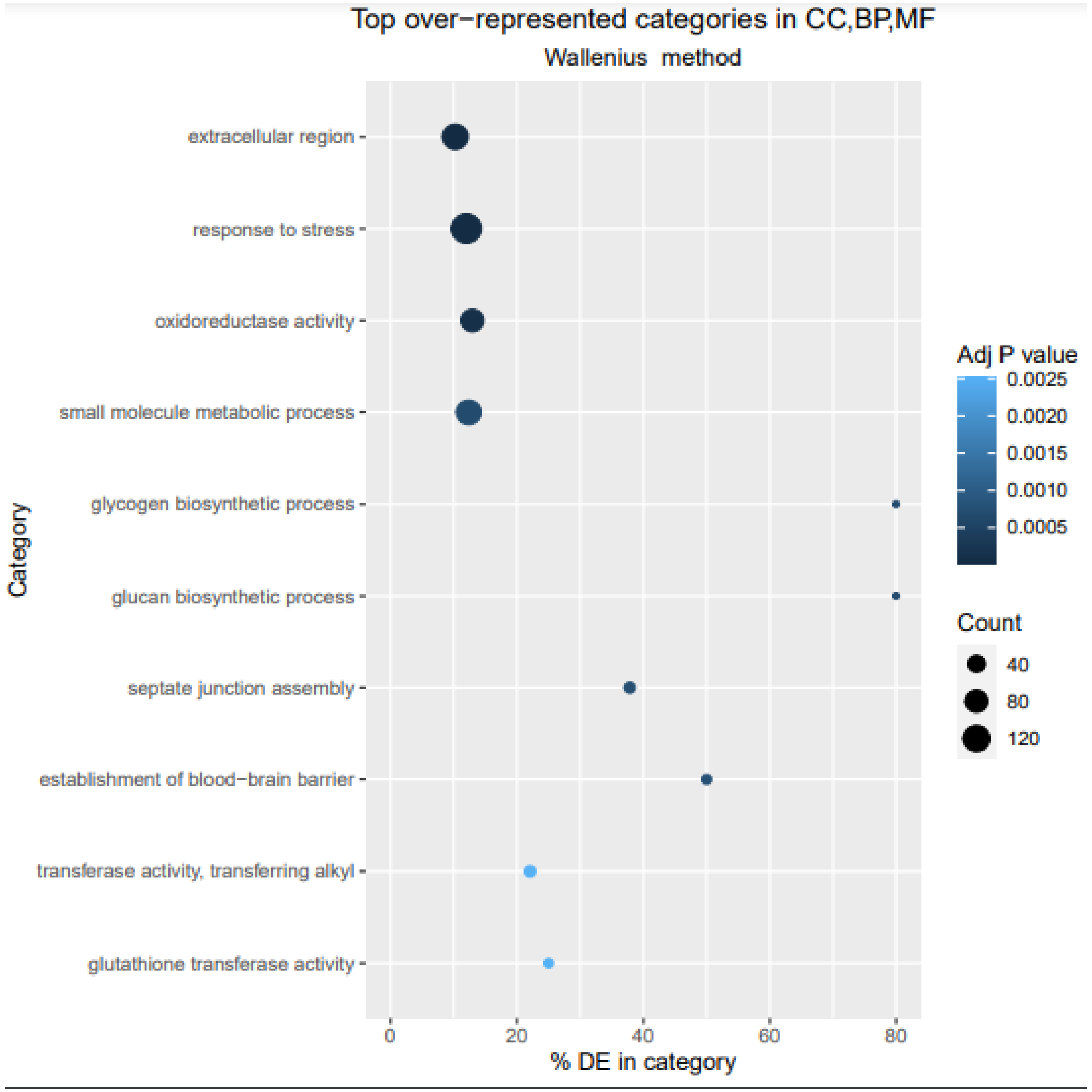
Goseq Ranked Category List



Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
category	over_represented_pvalue	under_represented_pvalue	numDEInCat	numInCat	p_adjust_over_represented	p_adjust_under_represented
00010	2.05063769648692e-05	0.999996086029403	14	47	0.00260430987453839	1
01100	8.2027632554107e-05	0.999955987077392	96	871	0.00520875466718579	1
00030	0.00129899817679932	0.999796008662722	7	22	0.0549909228178379	1
00480	0.00189406497047164	0.999479863524908	11	59	0.0601365628124746	1
00280	0.00345052107021624	0.9992290425026	8	32	0.0876432351834926	1
00071	0.00753915959835802	0.99827538285178	7	28	0.159578878165245	1
04512	0.0110380007220979	0.998813977208491	4	8	0.173710218310968	1
00531	0.0112267509692041	0.998103065487341	5	15	0.173710218310968	1
00982	0.0123101729511709	0.996065309367063	9	59	0.173710218310968	1
00051	0.0178294534065123	0.99560709293382	6	29	0.226434058262706	1
00260	0.0230401612821023	0.994064574344173	6	26	0.266009134802454	1
00980	0.0305942726509127	0.989235432758313	8	58	0.316592028964092	1
00460	0.0324070580829386	0.998753929516806	2	3	0.316592028964092	1
00640	0.0389660882444899	0.990107768388016	5	22	0.353478086217873	1
04145	0.0521597340323901	0.978456657294258	9	58	0.419530449448065	1
00520	0.0528542298517247	0.98112427957776	7	39	0.419530449448065	1
00410	0.0623369733998265	0.984993161833823	4	17	0.465693860104586	1
00500	0.0880797803998964	0.961657057249684	8	53	0.597656477544246	1
00270	0.089413173805832	0.974671534482012	4	23	0.597656477544246	1

The GOSeq analysis output visualizes the top over-represented categories in Gene Ontology (GO) terms across Cellular Component (CC), Biological Process (BP), and Molecular Function (MF) using the Wallenius method. Each dot represents a GO category, with the x-axis showing the percentage of differentially expressed (DE) genes in that category (ranging from 0 to 80%) and the y-axis listing the categories, such as "extracellular region" and "response to stress." Dot size indicates the count of DE genes (40 to 120), and color reflects the adjusted p-value (darker blue for more significant, ranging from 0.0025 to 0.025).

We can observe that "glycogen biosynthetic process" and "glucan biosynthetic process" exhibit the highest % DE, both exceeding 80%, though their gene counts are relatively small (smaller dots). In contrast, categories like "extracellular region" and "response to stress" show moderate % DE (around 10–20%) but have larger gene counts and significant p-values (darker blue). This suggests that while glycogen and glucan biosynthetic processes are highly enriched in terms of % DE, pathways like stress response involve more genes



Kegg Pathway Analysis

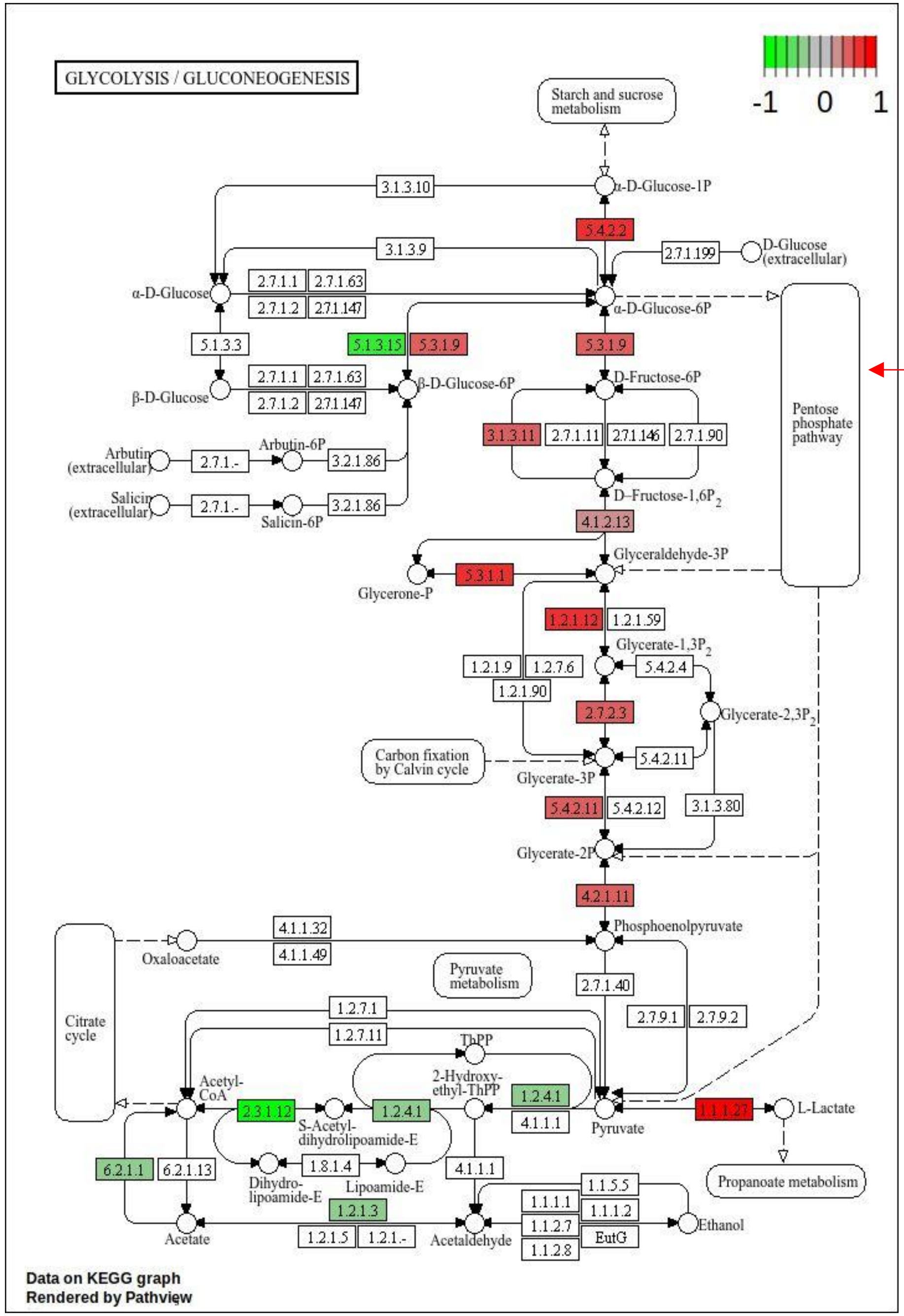


Figure 1

Figure 2

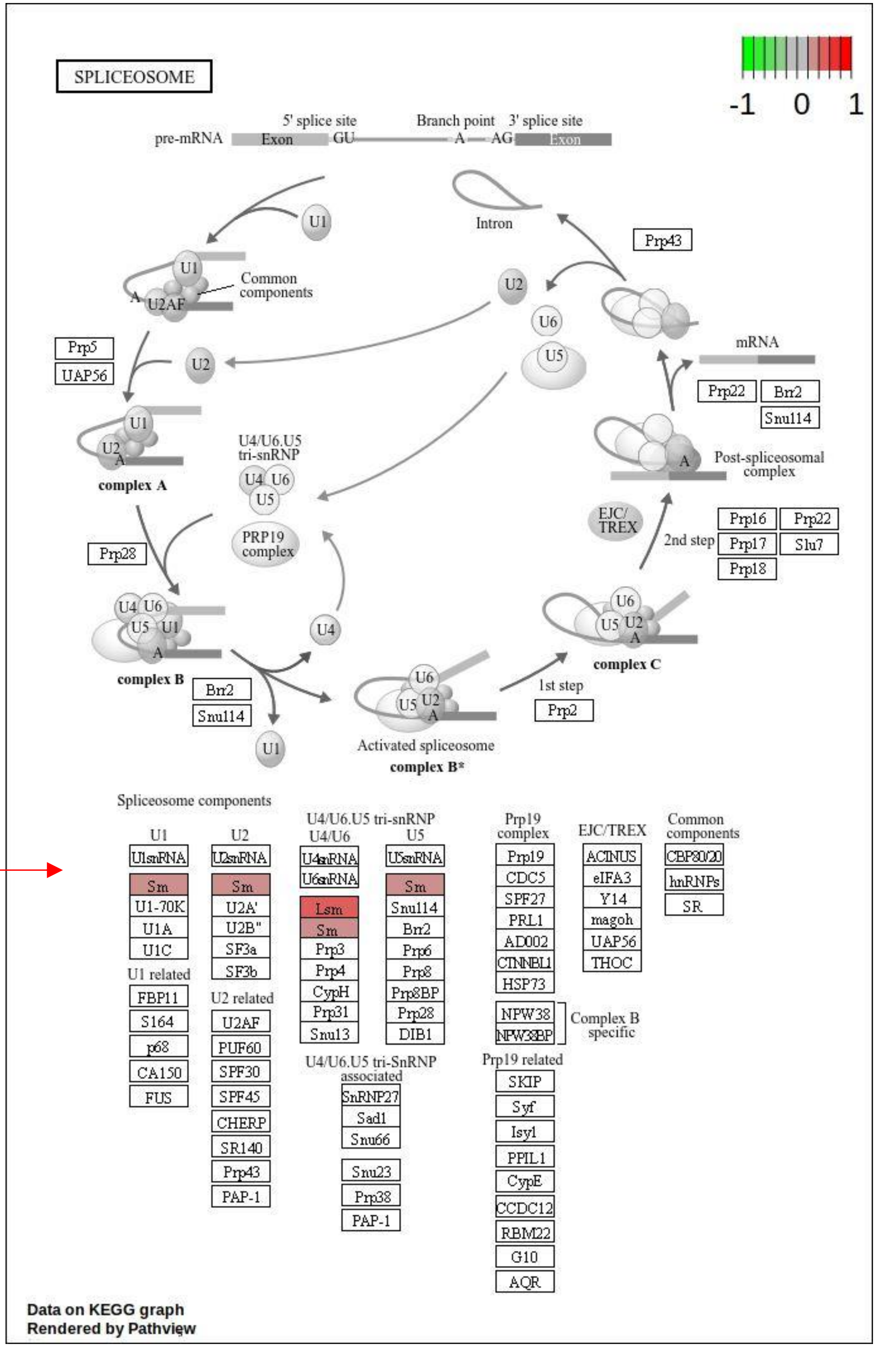


Figure 1 illustrates glycolysis and gluconeogenesis pathways. This metabolic pathway is crucial for glucose metabolism, showing the conversion of glucose to pyruvate (glycolysis) and the reverse process (gluconeogenesis). The red and green colors indicates genes that are differentially expressed, with red representing upregulation and green representing downregulation.
Several key enzymes in both glycolysis and gluconeogenesis appear to be significantly altered

Figure 2 represents the spliceosome pathway. This diagram shows the complex process of RNA splicing, where introns are removed from pre-mRNA and exons are joined to form mature mRNA. The pathway shows various spliceosome components and complexes (A, B, C) involved in this process. Similar to the first diagram, red and green boxes indicate differentially expressed genes within this pathway.

The Kegg Pathway and Goseq Analysis suggest RNA processing (spliceosome), energy metabolism (glycolysis/gluconeogenesis), and stress response pathways are enriched.

Q 4) Briefly answer the following questions

A) Why is normalisation important in RNASeq data analysis?

Answer: Normalization in RNA-Seq is important because it corrects for technical biases that mask true biological differences. Raw read counts are affected by sequencing depth, gene length, and library preparation variations, making direct comparisons misleading. Methods like RPKM,FPKM, TPM, or DESeq2's normalization adjust for these factors, ensuring that expression differences reflect actual biological variation rather than technical artifacts. Without normalization, downstream analyses would likely identify false positives and miss true expression changes.

B) What are “differential expression analysis” and “functional analysis” in RNASeq data analysis? What data is given as the inputs and taken as the outputs of those steps?

Answer: Differential expression analysis identifies genes with statistically significant expression changes between conditions. It takes normalized read counts and experimental metadata as inputs and uses statistical frameworks like DESeq2 or edgeR to produce lists of differentially expressed genes with associated statistics (eg log fold changes, p-values).
Functional analysis interprets the biological significance of these expression changes. It takes differentially expressed gene lists as input and uses databases like Gene Onotlogy or KEGG to identify enriched biological processes or pathways. The output includes significantly enriched functional categories and visualizations.

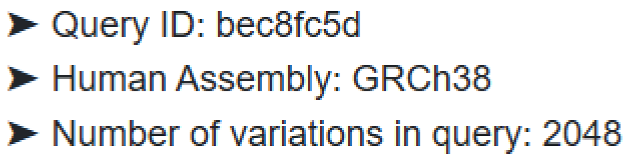
C) Is it a good practice to keep the overrepresented sequences, and not remove them, in RNASeq? Why?

Answer: Keeping overrepresented sequences in RNA-Seq data is generally not good practice. These sequences typically represent technical artifacts like adapter contamination, PCR duplicates, or incompletely depleted rRNA rather than biological signals. Keeping them can skew analysis results, waste computational resources, and reduce effective coverage of genes of interest

Q 5) SNPnexus is a web-based variant annotation tool designed to simplify and assist in selecting and prioritising known and novel genomic alterations. Visit their website here. Check the video tutorial. Using the variant file given to you (sample.vcf), run your analysis to answer the following questions:

A) How many variants are listed in the .vcf from the Ensembl database?

Answer : 2048



- Query ID: bec8fc5d
- Human Assembly: GRCh38
- Number of variations in query: 2048

B) How many exonic variants are present in the list?

Answer:

- 42 → Coding Non-Synonymous
- 14 → Coding Synonymous

C) Compare the number of deleterious/damaging variants annotated with SIFT vs. Polyphen.

Answer:

- 99 → SIFT
- 10 → Polyphen (3 out of 10 benign)