RNAseq Data Analysis Template

Glossary of Terms

Insert, fragment and read

DNA of interest is the insert
Add adaptor sequences to the insert and it becomes a fragment
A sequenced fragment is a read.

'+' and '-' strand

'+' strand is also called coding/ forward strand/ sense strand, while '-' strand is called template /reverse strand/ anti-sense strand

The forward strand has to do with the synthesis of the RNA, which proceeds in the 5' → 3' direction. Forward strand in humans is also assigned as the 5' end closest to Centromere.

It is called coding strand because it is reference for mRNA.
i.e, in the coding strand if T is replaced with U , it gives the mRNA sequence

** The orientation of the gene on the DNA determines which strand is the coding strand and which is the template strand.
If the gene is on forward strand, the reverse strand is used as template.
If gene is on the reverse strand, the forward strand is used as template.

The reverse strand is the template strand used to synthesize mRNA.It is complimentary to both dna and mRNA, as it runs in 3' -> 5' direction

The reverse complement of the coding strand gives the sequence of the template strand, read in 5'->3'

cDNA

DNA synthesized from mRNA template is cDNA. It represents only expressed genes and lacks introns. Reverse transcriptase synthesizes the complementary dna strand from mRNA.

Covariates

covariates can include any measurable factors that may influence the results of an experiment or study.

•Demographic Variables: Age, sex, race, ethnicity
•Clinical Variables: Disease status, treatment type, medication use, comorbidities
•Technical Variables: Batch effects, sequencing depth, quality metrics, lab conditions
•Environmental Variables: Diet, lifestyle, exposure to toxins or pathogens

covariates can significantly impact the analysis of gene expression data. Including covariates in the analysis can help in adjusting for confounding factors, improving the accuracy of the differential expression analysis

How to adjust for covariates ?

meta/adjust_for_covariates.

Contigs:

A set of overlapping DNA segments that together represent a consensus region of DNA

Mate:

Mate refers to one of the reads in a paired end.

Data collection

Does the data generated have the potential to answer the biological question of interest?

Experimental Design

- **Library Type**
  - Strand preserving library
  - Use of strand preserving libraries that retain information of the DNA strand that is actually expressed.
  - Paired and sequencing - contains fragments (i.e reads + adaptors)..
- **RNA extraction protocol**
  - rRNA is highly abundant and constitutes 90% of the total RNA in the cell, leaving the 1-2% mRNA (messenger RNA(mRNA))
  - There are two methods for RNA extraction, enrich for mRNA using poly(A) selection or deplete rRNA using ribosomal depletion. The latter is the preferred method in most cases due to limited quantity of biological samples.
- **Number of replicates**
  - Variability in measurements exists in the form of Technical variability and Biological noise. It is recommended to have 3 Biological replicates while performing the experiment.
- **Sequencing Depth**
  - Recommended - 5 million(medium and highly expressed genes) to 100 million mapped reads(low expressed genes)
- **Statistical power of the analysis**
  - For the desired effect size (fold change), how many samples are needed to have the maximum probability of getting significant result.(see section 10)

Yeast as a Model to Study BRCA2 Functions and Cell Death Programs
doi: 10.3389/fonc.2022.908442

In the experiment , the author used yeast as a model for functional genomics studies to identify new genes with a potential role in BRCA2-mediated control of cell death processes. To this aim, they performed an RNA-seq analysis in yeast either growing or undergoing Acetic Acid Induced Regulated Cell Death in the presence or absence of BRCA2 expression.
Here I am taking a subset of samples from their experiment to study the role of BRCA2 in growth phase.

Description of samples:

Condition 1. Yeast cells harboring a recombinant pYES plasmid containing BRCA2 cDNA grown in SC-Gal for 16 hours (Y-BRCA2)

Condition 2. Yeast cells harboring empty pYES plasmid grown in SC-Gal for 16 hours (Y-pYES2)

1. Download samples from GEO/ENA
   - download_fastqs.sh
2. Download ref_genome, gtf,gff files
   - http://ftp.ensemblgenomes.org/pub/release-57/fungi/gff3/saccharomyces_cerevisiae_ensembl_gtf.gz
   - http://sgd-archive.yeastgenome.org/sequence/S288C_reference/genome_releases/

The downloaded files are in fasty format

File Type

FASTA:
•Header Line: Starts with a ">" character followed by a description or identifier for the sequence.
•Sequence Lines: The actual sequence of nucleotides or amino acids, which follows the header line.

FASTQ:
•Header Line: Starts with a "@" character followed by a description or identifier for the sequence.
•Sequence Line: The actual nucleotide or amino acid sequence.
•Plus Line: A "+" character (often followed by the same identifier as the header line).
•Quality Line: Contains quality scores for each base in the sequence, encoded as ASCII characters.

=Contains both sequence data and associated quality scores.

Pre-alignment QC

2. FastQC

batch_fastqc.sh

Per Base Sequence Content: FAIL because of presence of primers in the first 12 bases of RNA-seq data

Sequence duplication levels FAIL : high deduplication levels after >50 on x axis, but not surprising.

Sequence length distribution: most samples have most sequences of length 76bp

%GC content: around 40%, close to actual value 38%

Per base sequence quality score: Red line: median across every position, Blue line : avg across all nucleotides
Above 31 for all samples

Observations:

per base sequence quality
red line is the median across every position, blue line is the avg across all nucleotides
They are above 31 for all samples

per sequence quality score
Majority of the sequences have mean quality score of 36, with only bumps of lower values on x-axis, i.e they have 1 in 1000 chance of a wrong sequence call

Number of reads
The number of reads are sufficiently high for samples looking BRCA2 genes(for 10m vs 6m for sample containing BRCA2 gene)

Many over represented sequences, that need removal

3. Cutadapt

cut_adapt.sh

From fastqc, data.txt, extract the overrepresented sequences and pass these to cutadapt to remove them from the samples

4. Trimmomatic

batch_trimmomatic2.py

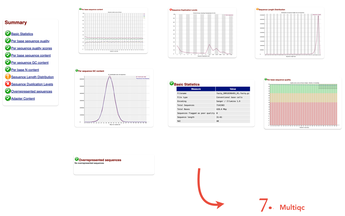
Removes first 12 bases and any low quality reads, hence fixing per base QC content

5. Fastp

matic_to_fastp2.R

Trim 3 bases from the ends of the read

6. FastQC



7. Multiqc

Post-alignment QC

batch_qualimap.sh

9. FastQC

Uncorrected BAMs (Without step 3,4,5)

Corrected BAMs

10. Qualimap

Coverage profile

Showing the overall distribution of reads across the reference genome by depth of coverage

X-Axis: Represents different genes or regions within the genes. Each tick on the axis may correspond to a specific gene or a segment of a gene.
Y-Axis: Represents the coverage depth. i.e., the number of reads mapping to each gene or region.
Peaks in Curve: The curve or bars in the graph show the coverage for each gene or region. Peaks indicate high coverage, while valleys or dips indicate lower coverage.
genes or regions with the highest peaks, represent regions with the most reads mapped, indicating high coverage.

A well-formed histogram with a single peak suggests uniform coverage. Multiple peaks or a wide distribution may indicate uneven coverage, which could be due to sequencing biases or technical issues.

There are 700 transcripts that have no coverage. From the plot we can assume the the desired coverage level is 50. Most transcripts have a coverage of 50.

Genebody coverage

showing how uniformly the reads cover the entire length of genes

Typically gene body coverage should give a relatively flat line across percentiles indicating Uniform Coverage, in this diagram we can spot greater coverage between 0-20 percentile on the 5' end indicating coverage bias.

Mean gene body coverage across the whole transcript

$$= \sum \frac{\text{2depth of all bases in a transcript(no. of reads)}}{\text{Number of bases in the transcript}}$$

3 parameters are required to calculate genebody coverage :-
(1) Mean coverage across the whole transcript
(2) Mean Coverage at the 5' end - mean coverage in first 100 bp
(3) Mean coverage across the 3' end - mean coverage in the last 100 bp region
5' bias is the ratio between mean coverage at the 5' region and the whole transcript.
3' bias is the ratio between mean coverage at the 3' region and the whole transcript. 5'-3' bias is the ratio between both biases.

Splice junction

Most reads as expected are Exonic

A splice junction is a region where two exons are joined together after the removal of an intervening intron during the process of RNA splicing
Splice junctions occur at the boundaries between exons and introns in pre-mRNA.
There are two main types of splice junctions:
•5' Splice Site (Donor Site): The boundary at the beginning of the intron (and of the upstream exon).
•3' Splice Site (Acceptor Site): The boundary at the end of the intron (start of the downstream exon).
Novel splice junctions are those that do not match any known annotated junctions in the reference database. In this case the reference database is the yeast gtf file from Ensemble

11. Infer experiment

run_infer_experiment.sh

The majority of the reads (78.25%) have the expected orientation for paired-end sequencing.

Alignment

8. STAR

Why align RNAseq reads to genome and not transcriptome?

there are several disadvantages of mapping the RNA sequences directly to the predicted transcripts:
1 We wouldn't know if we are looking at the complete gene as many genes structures are incomplete and also inaccurate.
2 The genes we identify in the RNA samples may not be annotated in the genome yet. In this case, we would completely miss the gene from our analysis.
3 Mapping the sequences to the genome can help us identify the genes that are missing from our annotation or annotated incorrectly.
4. Full-Length Alignment: Aligning to the genome ensures that full-length transcripts, including those that span multiple exons and introns, and also allows for precise mapping of reads to intron-exon boundaries, which is crucial for identifying splicing events.
5. Variant Calling: Genome alignment is essential for identifying RNA variants (e.g., SNPs and indels) and studying allele-specific expression.

All exons are not protein coding
only a fraction of exonic sequences are protein coding (~50% in humans). Both exons and introns are also present in untranslated regions (UTRs) and non-coding RNA.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5012253/

What is the difference between a cds and exon?

Protein coding exons between start and stop codon are CDS

run_star_alignment.sh

The log output file shows the percentage of mapped reads. A good mapping range is between 70-90%.The percentage of mapped reads in this experiment is 89%

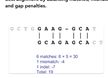
```
/App2Load/snl/Software/bin/cellranger/STAR/bin/MacOSX_x86_64/STAR --genomeDir
"$GENOME" \
     --readFilesIn "$R1" "$R2" \
     --sjdbGTFfile "$GTF_FILE" \
     --runThreadN "$THREADS" \
     --readFilesCommand "zcat" \
     --outTmpDir "$TMP_DIR" \
     --outSAMtype BAM SortedByCoordinate
done
```

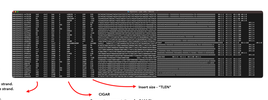
`--runThreadN` Number of threads to use.
`--genomeDir` Directory where the STAR index is located.
`--readFilesIn` Paths to the input RNA-seq read files (can be paired-end or single-end).
`--readFilesCommand zcat` Decompress .gz files.
`--outSAMtype` Output file format (e.g., BAM sorted by coordinate).
`--outTmpDir` Temp dir for output, typically /tmp/ to ensure reads spanning known splice junctions are reprocessed used in this case.

STAR (Spliced Transcripts Alignment to a Reference) is a widely used aligner for RNA sequencing (RNA-seq) data. It is designed to efficiently align short RNA-seq reads to a reference genome, including the ability to detect and align reads across splice junctions.

Gapped mapper

Unlike simple alignment tools that may only align sequences without accounting for gaps, gapped mappers are specifically designed to align sequences that have mismatches, insertions, or deletions relative to a reference sequence.
Gapped mappers use dynamic programming algorithms (e.g., Smith-Waterman or Needleman-Wunsch) to calculate the optimal alignment by considering gaps. These algorithms help in finding the best alignment by balancing matches, mismatches, and gap penalties.





SAM FLAG
0x1 (1): The read is paired in sequencing.
0x2 (2): The read is mapped in a proper pair.
0x4 (4): The read is unmapped.
0x8 (8): The mate is unmapped.
0x10 (16): The read is mapped to the reverse strand.
0x20 (32): The mate is mapped to the reverse strand.
0x40 (64): The read is the first in the pair.
0x80 (128): The read is the second in the pair.
0x100 (256): The alignment is not primary (a read having multiple alignments).
0x200 (512): The read fails platform/vendor quality checks.
0x400 (1024): The read is a PCR or optical duplicate. 0x800 (2048): The read is a supplementary alignment.

CIGAR
Compact representation of a BAM file.
Compact Idiosyncratic Gapped Alignment Report

File Type

Fields of a SAM file

QNAME : Query Template Name
FLAG : Bitwise flag encodes various properties of the read
RNAME: Reference Sequence Name
POS: Leftmost mapping position on the reference genome
MAPQ: Mapping Quality
CIGAR : Shorthand description of the alignment
RNEXT: Reference name of the mate
PNEXT: Position of the mate read
TLEN : length of insert size (used to confirm physical distance between mates)
SEQ : Sequence segment
QUAL: Quality of each base in the read, encoded in ASCII

File Type

Fields of a gtf(gene transfer format) file

Seqname : Name of the chromosome or scaffold
Source : Indicates the program or database that generated this feature
Feature : type of feature : 'gene', 'transcript', 'exon', 'CDS', 'start_codon', 'stop_codon'
Start : start position of the feature
End : end position of the feature
Score : A floating point value
Strand : '+', '-', or '.'
Frame : '0', '1' or '.'
Attribute : A semicolon-separated list of tag-value pairs providing additional information about the feature.Each tag-value pair is formatted as "tag "value";"

GTF: Gene transfer format

GFF:General Feature Format

GTF is less flexible for complex feature relationships.
GFF supports hierarchal relationships among features using parent and ID attributes.
There is no fixed formatting for attributes unlike GTF

Differential Gene Expression Analysis

12. FeatureCounts

batch_FeatureCounts.sh

check_countdata.R

Check for NAs and other non numeric values in dataset

13. DESeq2

DESeq2.R

Dispersion plot - Let's asses the variability in gene expression
X-axis: Avg. expression level of each gene
Y-axis:the spread of counts for each gene
Each point represents the dispersion estimate for a single gene
We can see that the genes having lower counts have higher variability in expression, and are less reliable.

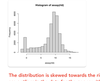
The red line :DESeq2 starts by estimating dispersion for each gene independently and a trend line is fitted to the observations. The observed gene dispersions are shrunk to the fitted line(Empirical Bayes Shrinkage) to improve the statistical reliability of genes with lower counts.

In the plot above, we can see that there are quite a few gene dispersion estimates that are unusual i.e the light blue dots on top of the red line

Analysis of the effect of BRCA2 expression on yeast cells growing for 16 h up to stationary phase (comparison Y-BRCA2 vs Y-pYES).

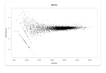
The brighter unmixed colors ('neon green' and 'bright red') are highly similar, and of contrasting conditions while the 'olive drab green' represents dissimilar profiles. In the above plot, samples 'SRR1839651:2' and 'SRR1839649' have similar DE profiles and this is reflected in the PCA plot where they cluster together

Histogram of assayVsd - The vsd transformation is used to stabilize the variance across different expression levels, making the data more suitable for downstream analysis, such as clustering or principal component analysis (PCA)



The distribution is skewed towards the right and there are outliers in the data for the genes with log transformed counts before/between 0 and 5, possibly NAs?

MA plots can help identifying up-regulated(above M =0)&down-regulated genes for two conditions, In this case there are a set of down-regulated genes with <10 base mean expression, and most differentially expressed genes fall within +-2 fold change levels.



MA plot :Visualize the relationship between the mean expression (average signal) and the log fold change of genes.
X-Axis: A (Mean Average): The average log expression level of each gene across the two conditions, plotted on the x-axis.
Y-Axis: M (Log Ratio): The difference in log expression levels between two conditions (log fold change), plotted on the y-axis.

The histogram visualizes the frequency distribution of the log-transformed count values from the rld object.From the graph, we can get insights into variance stabilization and data distribution

The shape of the histogram should ideally be a normal distribution with the center representing the most commonly log-transformed count values. The narrower the spread, indicates reduced variance. Any bars far from the main peak can indicate outliers or genes with extreme expression levels.

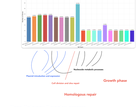
A PCA (Principal Component Analysis) plot visualizes the major sources of variation in a dataset by reducing its dimensionality while preserving the key patterns in the data, allowing for the identification of patterns, clusters, and outliers. In RNA-seq data analysis, PCA plots are used to visualize the overall gene expression differences between different samples or conditions.

In the plot, most of the variation in data =60% is explained by the presence/ absence of the gene of interest(PC1), and the two control samples and one sample 'SRR1839649' may be an outlier as it does not cluster near it's 'type'. However due to the small sample size of the set, this cannot be confirmed.

Hence BRCA2 expression does not affect yeast cell growth.

The analysis did not reveal any difference of gene expression between the two conditions. Although the expression profiles for the conditions is different as established by pca plot and heatmap, there are no significantly differentially expressed genes between the two conditions as seen on the volcano plot (absence of green dots)
This is not expected to change with the addition of more samples as demonstrated by Contanza et al.

14. Gene Ontology Biological process



Gene Ontology Biological process -
Top 20 biological processes can be roughly divided across 3 categories- 'Nucleoside metabolic processes', 'Mitosis and DNA Repair',
•Plasmid introduction and expression-
Nucleoside metabolic processes- There are 10 nucleoside synthesis processes that are up-regulated. These can be explained by the experimental condition of the yeast cells where they were cultured until stationary phase, which would mean during the log phase where yeast cells in S-phase, there would be increase in nucleoside synthesis in preparation for the intensive energy requirement for growth. Once growth phase begins these nucleoside metabolic processes would be responsible for nucleoside recycling and energy regulation, by conversion to nucleotides i.e ATP
Mitosis and DNA Repair-
The most frequent GO term in the dataset is dna recombination . There are two circumstances that involve dna recomination in this case(1)Mitosis and (2)DNA repair.
As the cells undergoing growth are exponentially dividing, dna recombination during mitosis is what is reflected on the graph. In the case of DNA repair, BRCA2's role in repairing double stranded breaks through Homologous Recombination is well-established.In the absence of an efficient HR pathway, double stranded breaks accumulate resulting in genome instability, thus supporting tumorigenesis. Hence it is possible that 'dna recombination' also includes homologous recombination that would have occurred in cells containing the Gene of Interest.
•Plasmid introduction and expression-
The yeast cells may have been transformed through transfection and this can explain the occurrence of GO terms such as 'viral gene expression' , 'dna integration' and 'retrotransposition'

Pathway Enrichment

15. Power Analysis

Power_analysis_real data.R

Power_analysis_simulated_data.R

Power analysis - The statistical power of this study

The power of any statistical test is the probability that it will yield a statistically significant outcome.
What would a statistically significant result be in this case?
If we assume the null hypothesis to be true, i.e there is no difference between the expression profiles of plasmids containing BRCA2 and those without, the probability of getting a small difference between the two groups or more by chance is 0.05 (alpha). This means there is no actual difference, hence we would get p-value of greater than 0.05.
The power, which is the probability of getting a statistically significant outcome is expected to be extremely low.

The power is affected by three factors-
The size of the bias
The sample size
Measurement error

Since the sample size of this experiment is extremely small and the difference we are trying to detect if it exists is negligible, a power analysis on real data showed a 0 percent chance of getting a statistically significant outcome(i.e with 1.5 fold change as effect size at alpha as 0.05).If the sample size of this experiment is increased by 100 times, the "Estimated power is 0.00019202674337028)", which 0.02% chance of getting a statistically significant p-value.

alpha - level of significance, p - the value of significance