

Omar S. Harb, PhD

RNA Sequencing Analysis Basics

Slides adapted from originals provided by Dr. Kathryn Crouch

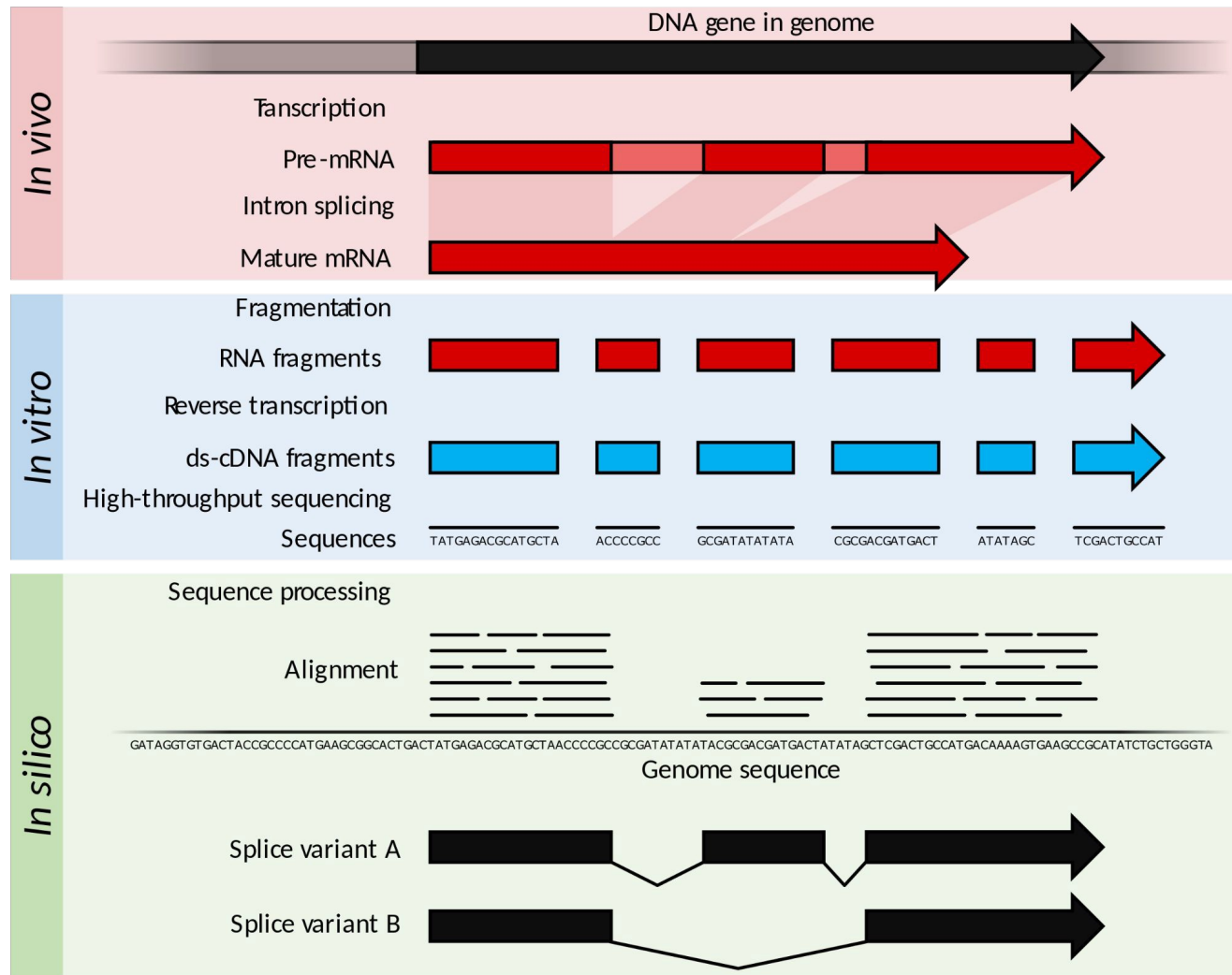
Why do we want to sequence the transcriptome?

Capture full unbiased view of full repertoire of transcripts for gene model prediction

Functional studies for conditions such as stress and drug resistance

Explore alternative splicing and complex patterns of expression and regulation

Transcriptome sequencing



Experimental Considerations



Do you have enough biological replicates?

Do you have enough RNA?

Is the RNA what you want?

Are you using an appropriate selection method?

Are you interested in strand differentiation?

Do you have appropriate controls?



Experimental Considerations

Do you have enough biological replicates?

- At least three are recommended

Do you have enough RNA?

Is the RNA what you want?

Are you using an appropriate selection method?

Are you interested in strand differentiation?

Do you have appropriate controls?

Experimental Considerations

Do you have enough biological replicates?

Do you have enough RNA?

- If you want to capture rare transcripts, you will need to sequence more deeply

Is the RNA what you want?

Are you using an appropriate selection method?

Are you interested in strand differentiation?

Do you have appropriate controls?

Experimental Considerations

Do you have enough biological replicates?

Do you have enough RNA?

Is the RNA what you want?

- If you are sequencing clinical or environmental samples, consider the abundance of the target organism

Are you using an appropriate selection method?

Are you interested in strand differentiation?

Do you have appropriate controls?

Experimental Considerations

Do you have enough biological replicates?

Do you have enough RNA?

Is the RNA what you want?

Are you using an appropriate selection method?

- PolyA selection is common - but not all RNA species are poly-adenylated

Are you interested in strand differentiation?

Do you have appropriate controls?

Experimental Considerations

Do you have enough biological replicates?

Do you have enough RNA?

Is the RNA what you want?

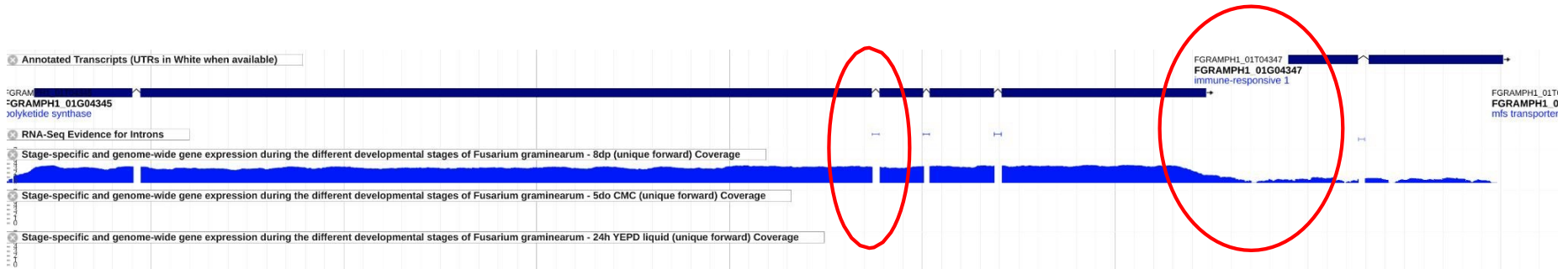
Are you using an appropriate selection method?

Are you interested in strand differentiation?

- Stranded library kits allow you to distinguish the strand from which the sequencing read originated.

Do you have appropriate controls?

What Can We Learn from RNA-seq?

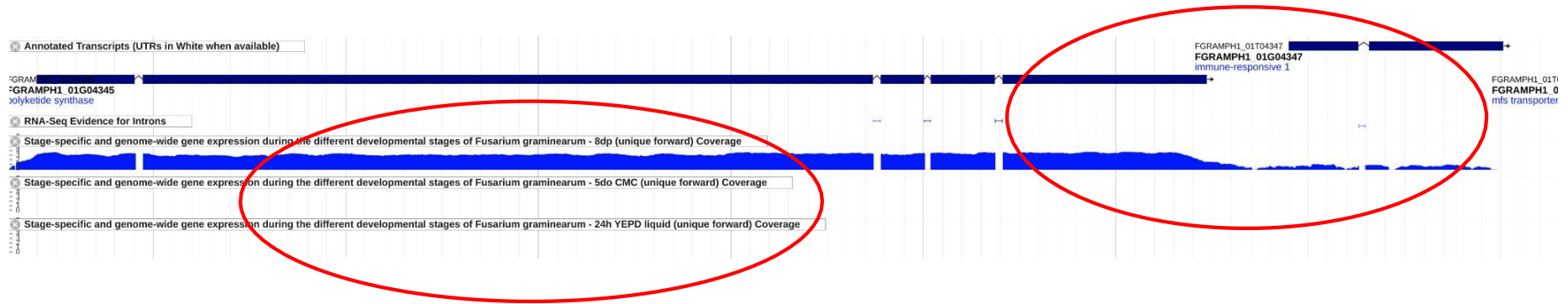


Gene Model Prediction

Alignment of RNA-seq reads to a genomic reference can help us to predict and confirm gene model structure

- Introns can be predicted based on coverage and on individual reads that cross splice junctions
- UTRs can be predicted based on coverage
- Differential splicing can also be predicted from coverage

What Can We Learn from RNA-seq?

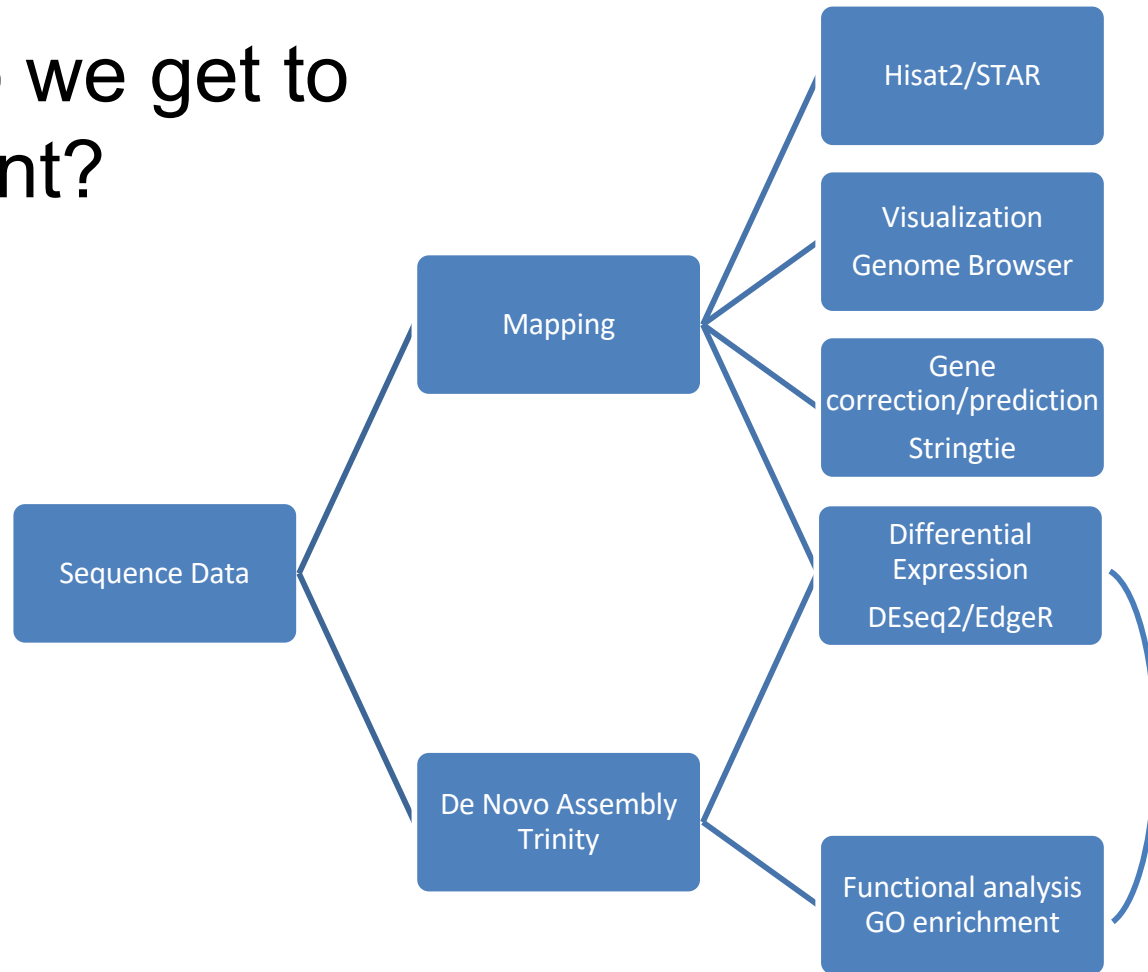


Differential Expression

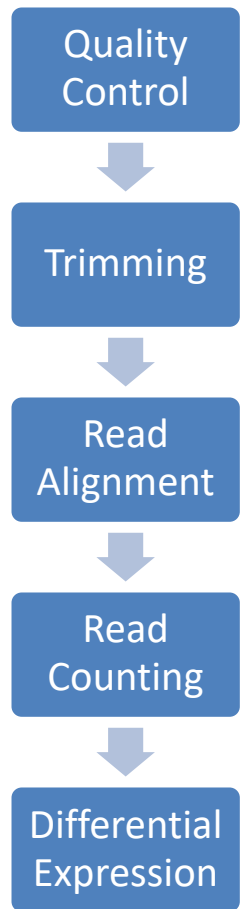
Depth of coverage can help us learn about transcript abundance

- Differential transcript abundance can be observed both within and between samples

How do we get to that point?



Differential Expression Analysis



File Formats: FASTQ

```
@NS500205:127:HW272BGXX:1:11101:7788:1040 1:N:0:TCTCGCGC+AGGCTATA
CCTTGNTCCGTCATATTTTTTAGCATTGCAATGACGCTAAGTCCCGATTGACGCGCACGTGCTCACCCGGTTTCC
+
AAAAA#EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

For paired-end reads you will have two files

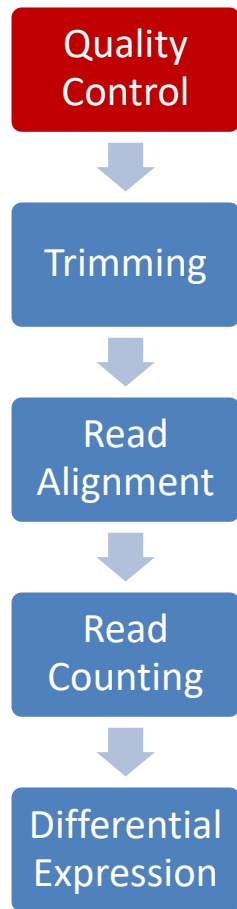
4 lines per read

- Line 1 is a unique header (this will be shared between the pairs)
- Line 2 is the sequence of the read
- Line 4 is the quality for each base
 - Quality is encoded using ASCII
 - <http://www.asciitable.com/>

Phred quality scores are logarithmically linked to error probabilities

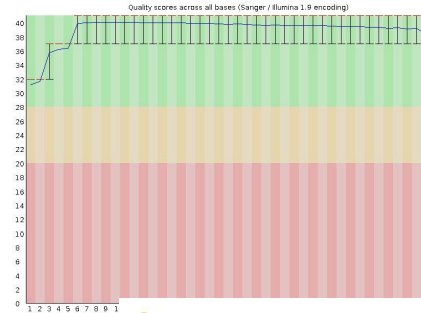
| Phred Quality Score | Probability of incorrect base call | Base call accuracy |
|---------------------|------------------------------------|--------------------|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 99.999% |
| 60 | 1 in 1,000,000 | 99.9999% |

Differential Expression Analysis

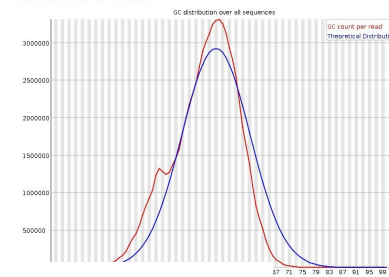


- FastQC <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
 - Overall sequencing quality
 - GC content
 - N content
 - Read length distribution
 - Over-represented sequences
 - Adaptor content
- Output is an html file that can be opened in a web browser

✓ Per base sequence quality



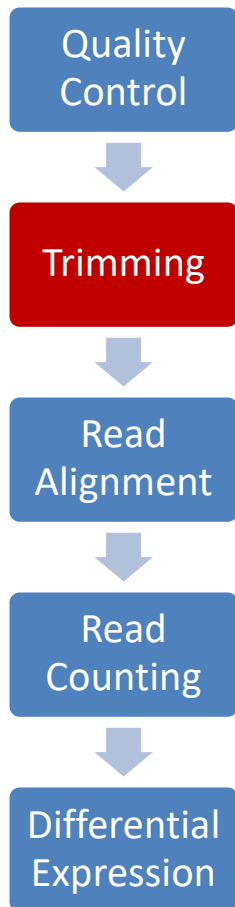
ⓘ Per sequence GC content



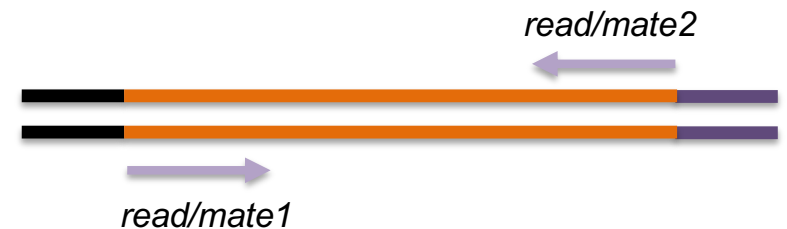
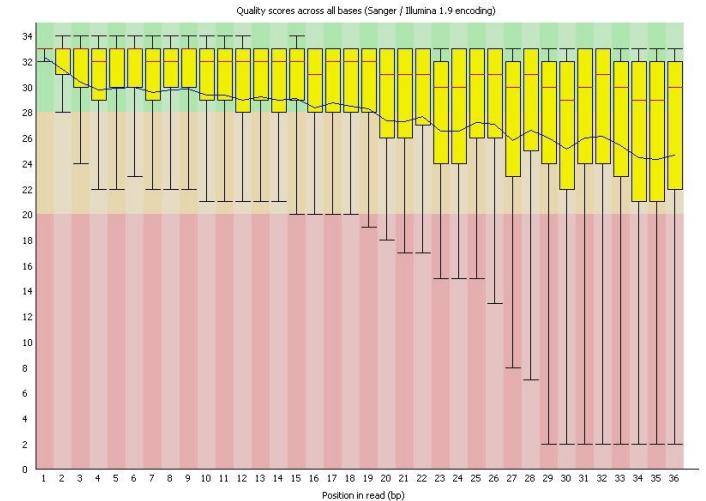
ⓘ Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|--|-------|---------------------|-----------------|
| ACAAGTGTGAACATTAATTGCAAGTTTGCAACGCTGTTCTTTAGTGTT | 78896 | 0.12562741276052788 | No Hit |

Differential Expression Analysis



- Trimmomatic, Sickle and others
- Sequence quality tends to decrease towards the 3' end of reads - these can affect mapping
- Trimmomatic will:
 - Remove poor quality bases from the 3' end of each read
 - Check for reads that are too short and discard them
 - Check that all reads still have a pair and discard those that don't
- Some trimming tools can also remove adaptors



Trimmomatic: <http://www.usadellab.org/cms/index.php?page=trimmomatic>
Sickle: <https://github.com/najoshi/sickle>

Differential Expression Analysis



What is an alignment?

Two sequences:

```
ATTGAAAGCTA
GAAATGAAAAGG
```

How would you align them?

```
- - ATTGAAA -GCTA
  |  ||| | | |
GAAATGAAAAGG- -
ATTGAAA -GCTA - - -
  ||| | | |
- - - GAAATGAAAAGG
```

Which is better?

Alignment scoring:

- 1 for match
- -1 for mismatch
- -2 for a gap

Now alignment 1 scores -4 and alignment 2 scores -10 so we would choose alignment 1

Differential Expression Analysis

Quality
Control



Trimming



Read
Alignment



Read
Counting

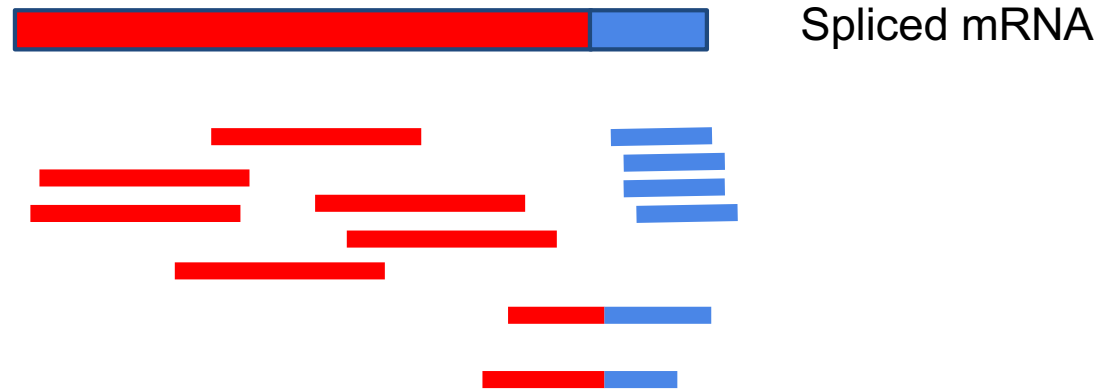
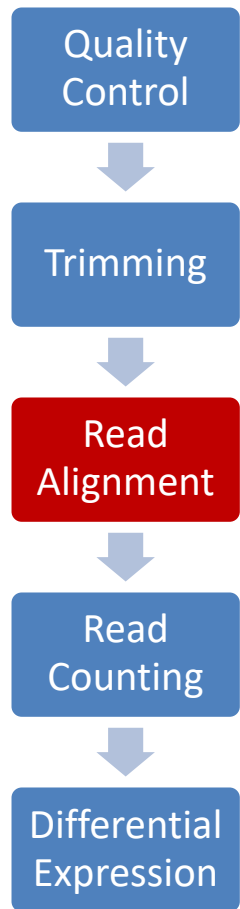


Differential
Expression



Spliced mRNA

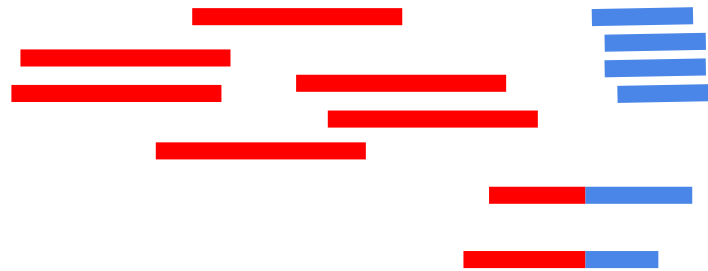
Differential Expression Analysis



Differential Expression Analysis



Spliced mRNA



Reference genome (genomic sequence with introns)

Differential Expression Analysis



 Spliced mRNA

A diagram of a single spliced mRNA molecule. It consists of a long red bar representing the exon and a shorter blue bar representing the intron, joined together.



 Reference genome (genomic sequence with introns)

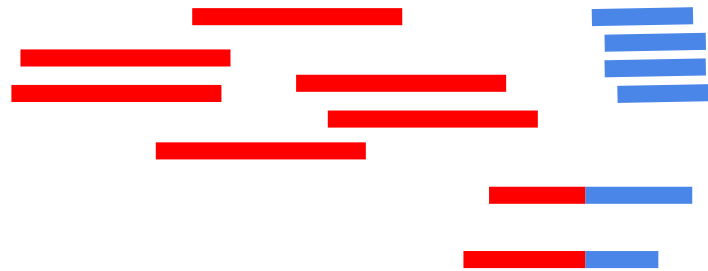
A diagram of a reference genome segment. It shows a long red bar (exon) followed by a thin line (intron) and then a blue bar (exon). Below this, there are several red bars of varying lengths, representing reads that map to the exon. To the right, there are several blue bars of varying lengths, representing reads that map to the intron or the second exon.



Differential Expression Analysis



Spliced mRNA



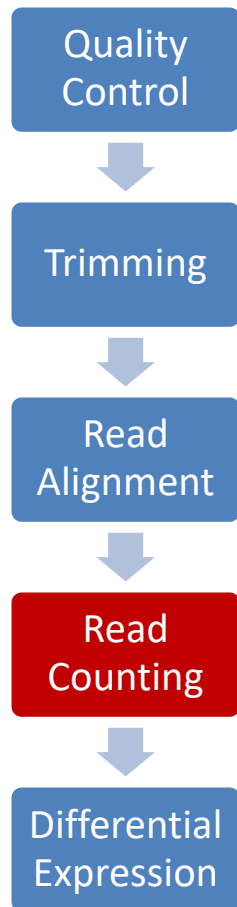
Reference genome (genomic sequence with introns)



Intron spanning reads



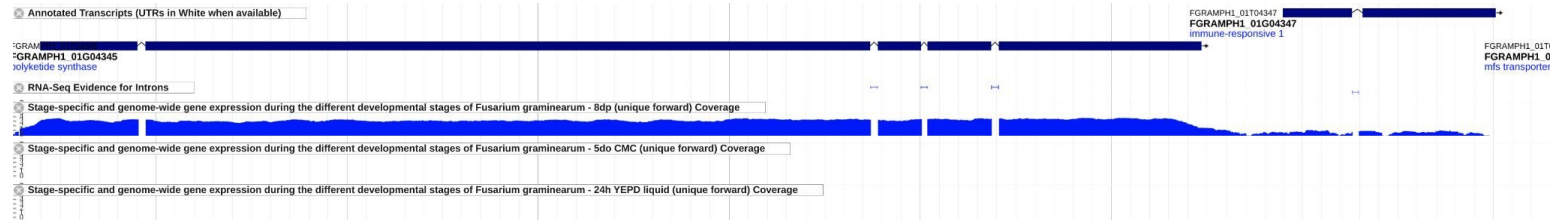
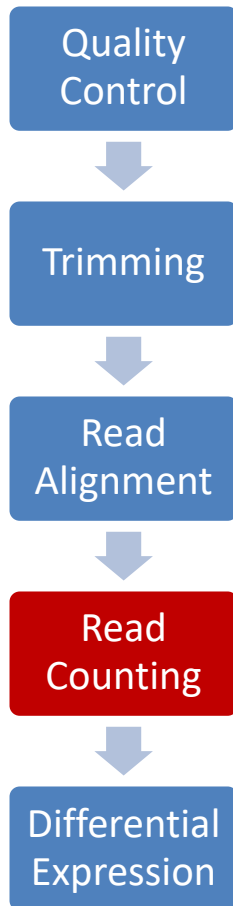
Differential Expression Analysis



Mapping tools for RNA-Seq data

- Capable of aligning millions of reads to a genome
 - In a reasonable amount of time
 - Using a reasonable amount of memory
- Use heuristic algorithms
 - Good enough – not perfect!
- Must be capable of aligning intron-spanning reads
- Hisat2
 - Fast, sacrifices sensitivity
 - <https://daehwankimlab.github.io/hisat2/>
- STAR
 - Very sensitive, but slow
 - <https://github.com/alexdobin/STAR>

Differential Expression Analysis



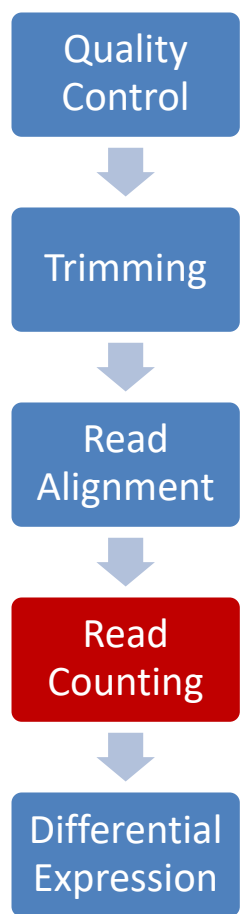
Quantifying Expression

At this point, we can load our alignment into a genome browser and look at expression

Looking at plots like this is great for one gene, but it is too much to look at every gene individually and is not statistically robust

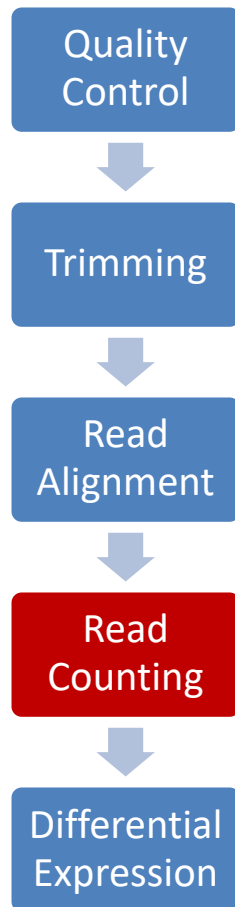
To examine transcript expression globally and perform robust statistics, we must count how many reads map to each gene

Differential Expression Analysis



| | union | intersection_strict | intersection_nonempty |
|--|---|---------------------|-----------------------|
| | gene_A | gene_A | gene_A |
| | gene_A | no_feature | gene_A |
| | gene_A | no_feature | gene_A |
| | gene_A | gene_A | gene_A |
| | gene_A | gene_A | gene_A |
| | ambiguous (both genes with --nonunique all) | gene_A | gene_A |
| | ambiguous (both genes with --nonunique all) | | |
| | alignment_not_unique (both genes with --nonunique all) | | |

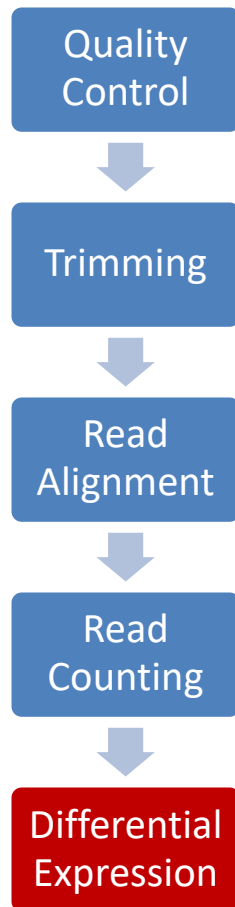
Differential Expression Analysis



Counting Reads Mapped to Genes

- Counting can be strand-specific with a stranded library
- Decide whether to count:
 - Only reads that align uniquely to one feature (stringent, most common)
 - Reads that can be aligned to multiple features (might be useful for multigene families)
- Decide at what level to count reads:
 - Per gene - robust and easy to analyse but lacks information about differential isoform expression
 - Per transcript - may allow identification of differential expression of known isoforms. Harder to interpret.
 - Per exon - may allow identification of differential expression of novel isoforms. Harder to interpret
- Htseq-count https://htseq.readthedocs.io/en/release_0.11.1/count.html
- featureCounts <http://subread.sourceforge.net/>

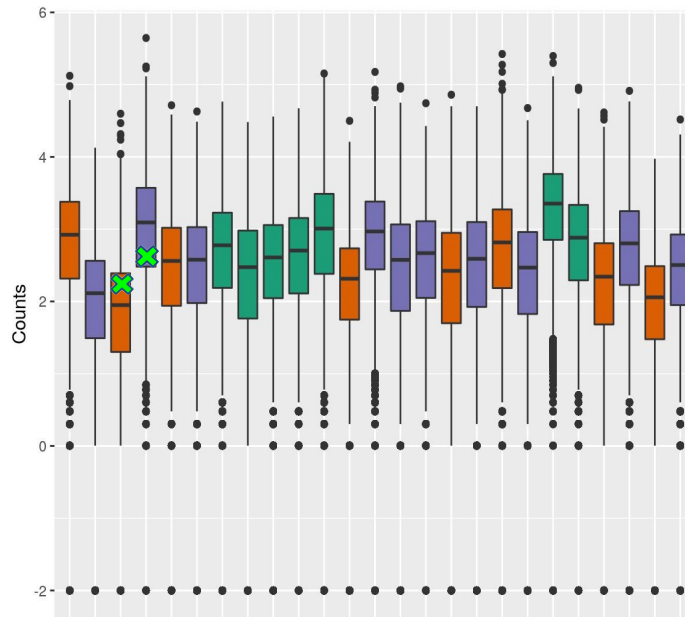
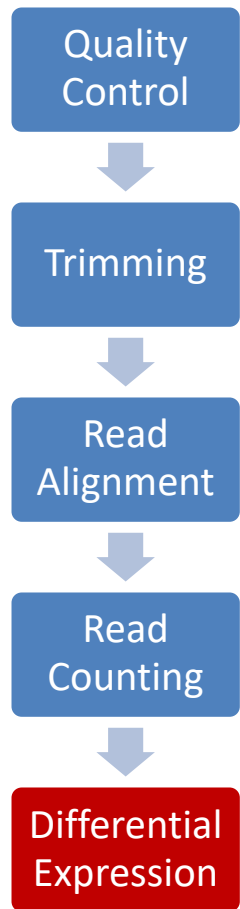
Differential Expression Analysis



Differential Expression

- Compare conditions to see which transcripts differ in abundance
 - Knockout/knockdown/mutant vs wild-type
 - Different lifecycle or cell cycle stages
 - Different nutrient sources
 - Virulent vs avirulent strains
- Sequence coverage depth is used as a proxy for transcript abundance
 - We've already seen how we can visualise this in the alignment
 - Counting the reads mapping to each gene gives us a means to quantify this
- Normalisation for sequencing depth must be carried out
 - Need to account for the total number of reads sequenced
 - Sequencing more reads from one sample will increase the number of reads mapped to each gene even if relative transcript abundance has not changed

Differential Expression Analysis

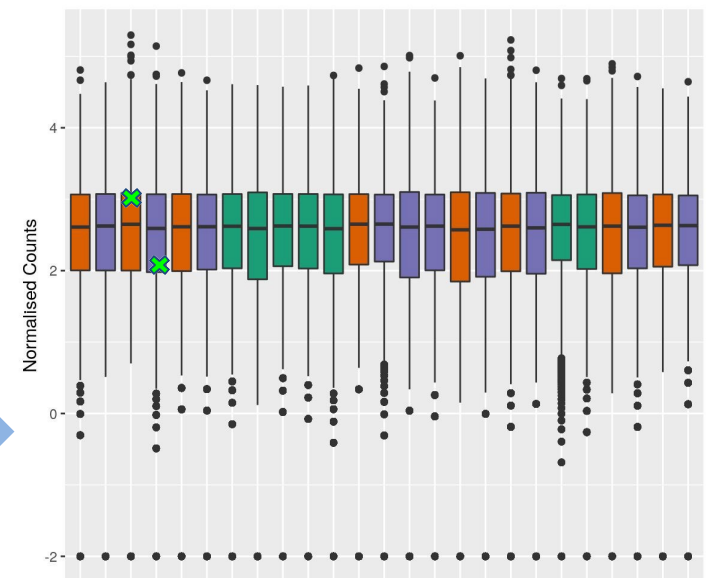


Normalised count data

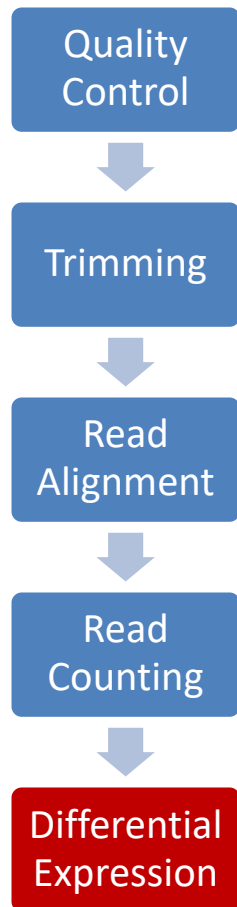
After normalising, count distributions are aligned so individual genes can be compared

Raw count data

Each box represents is one sample and shows the distribution of read counts for each gene



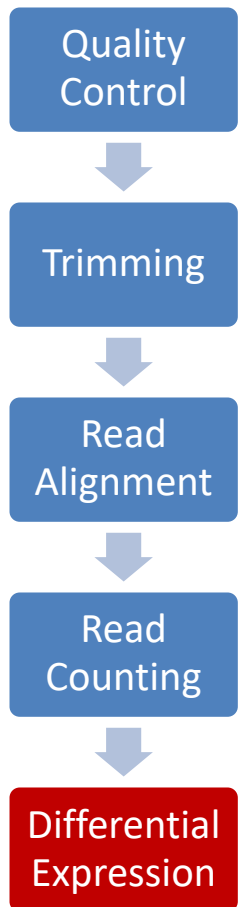
Differential Expression Analysis



Differential Expression - Statistical Analysis

- Robust statistical comparison of quantitative differences in transcript abundance
 - Single Factor experiments comparing two conditions
 - Multi Factor experiments investigating multiple experimental conditions
- RNA-seq data is not normally distributed.
 - Most packages assume a negative binomial distribution
 - Some packages may transform the data to fit it better to this model
 - Common tests used are Wald test (DESeq2) and F-test (EdgeR)
 - Log Ratio test can be useful for time courses
- Output is a table, which reports for each gene
 - Log2 fold-change - a measure of the magnitude and direction of differential expression
 - P-value
 - Q-value (FDR-adjusted P-value)

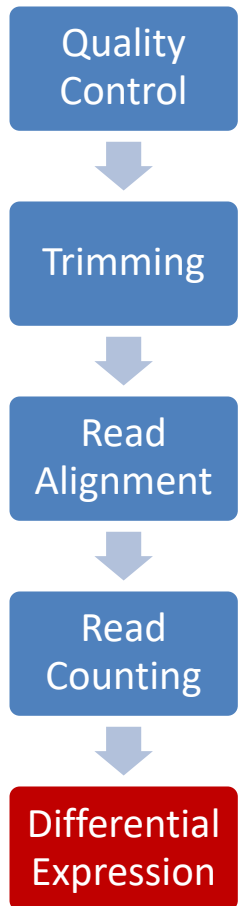
Differential Expression Analysis



Differential Expression - Error Types

- Type II error
 - “False negative”, “acquitting a criminal”
 - Falsely accepting a null hypothesis
- Type I error
 - “False positive”, “convicting an innocent person”
 - Falsely rejecting a null hypothesis
 - Potentially dangerous in science as it results in false discovery
 - Inherently linked to significance level - if you set a p-value cutoff at 0.01 you accept a positive result has a 1 in 100 chance of being a false positive
- If you test 10,000 genes using a p-value cutoff of 0.01, you expect to find 100 false positives
 - Q-value is an adjustment of the P-values from individual tests to reflect this
 - **ALWAYS USE THE Q-VALUE**

Differential Expression Analysis



Differential Expression - Statistical Analysis

- Tools for differential expression analysis usually combine data normalisation, data transformation and statistical testing into a single package
 - For this reason, it is best to use one package for the whole workflow if you aren't sure
- Some common tools include:
 - DESeq2:
<https://bioconductor.org/packages/release/bioc/html/DESeq2.html>
 - EdgeR:
<https://bioconductor.org/packages/release/bioc/html/edgeR.html>
 - Limma:
<https://bioconductor.org/packages/release/bioc/html/limma.html>

