Bioinformatics Coursework 1

Jude Popham K20041606

Multiple Choice Questions

Question 1

1. What is the final score of the alignment below calculated using the running score (blast scoring system)?:

```
Seq.1 CTTCCTTATGAATTGGAAGAAAACACAGACAAAGCCGTTA
      |||  ||||| |||||||||||| |||||  |||| ||||
Seq.2 CTTTTTTATGCATTGGAAGAAAA-ACAGA--AAGCTGTTA
```

A- 5
B- 10
C- 15
D- 20

Answer: C

Via the BLAST scoring system, there is a tallied score between the two sequences of 15 as a nucleotide match (alignment) is cumulative +1, a mismatch is -1 and a gap opening equals a score of -5. Continuing a gap is -2 cumulative. However, if the gap extends, there is an equation: - (opening cost + (length of gap x continuing cost)). There is a gap at the end which is two nucleotides long and thus produces a cost of -9 in this case.

Question 2

2. The alignment procedure that tries to align the entire sequence is

A- Multiple sequence alignment
B- Pair wise alignment
C- Global alignment
D- Local alignment

Answer: C

We are comparing a target and query sequence. This is a global alignment (option C). Global alignment reflects sequence variation as the whole sequence is aligned end-to-end, usually when the two sequences are approximately the same length or quite similar (local alignment only aligns selected parts of the sequence that have partial homology to identify regions with high similarity). Pairwise alignment is the umbrella term for local and global. Multiple sequence alignment is not considered in this case as it aligns 3 or more biological sequences with similarity.

Question 3

3. Sequence alignment helps scientists

A – to trace out evolutionary relationships
B – to infer the functions of newly synthesized genes
C – to predict new members of gene families
D – all the above

Answer: D

Sequence alignment and its *alignment score/Hamming distance* can produce evidence for all the above by comparing a sequence in a species with a known gene to a database of sequences that exist in the species to find sequences above a certain threshold for that gene. This is because biological sequences contain a lot of information that can be used to infer genetic relationships. These relationships can be extrapolated backwards to trace out divergent and convergent evolutionary relationships (A) by identifying homology, testing evolutionary models and building phylogenies. Newly synthesised gene functions can be inferred by functional analyses which identify conserved gene regions and aligning exons, protein coding regions and sequence mutations that produce new alleles (B). Conserved gene regions can also identify protein families (C): a known gene in one species can be searched for in a different species to predict paralogs and orthologs for that gene.

Question 4

4. The human genome is a reverse complementary DNA, given this sequence
ATCGATGCAATTGGCATATAT
Which is the correct reverse complementary sequence?

A– ATCGATGCAATTGGCATTTTA
B– ATATATGCCAATTGCATCGAT
C– ATCAATGCAATTGGCATATAT
D– ATGTATGGCAATTGCATCGAT
E– ATATATGCCAATTGCATCATT

Answer: B

The complement of DNA is 3’ to 5’ and the reverse is written in the 5’ to 3’ direction as DNA is antiparallel. The new sequence is determined by complimentary base pairing. The reverse complement of the provided sequence is B where the sequence is read backwards and swapped to complimentary nucleotides (A for T and C for G). The sequences are thus reversely complementary. The reverse is: TATATACGGTTAACGTAGCTA. The complement of the reversal is: ATATATGCCAATTGCATCGAT.

Question 5

5. The ENSEMBL Genome Browser you find a transcript version classified as non-sense-mediated mRNA decay. This means that the transcript contains

A – a non-coding exon
B – a partially coding exon
C – a pre-mature stop codon
D – a splice variant
E – a fusion transcript

Answer: C

Non-sense-mediated mRNA decay (NMD) is a translational mechanism in eukaryotes that eliminates mRNA containing premature stop codon. This means the original transcript that is classified as NMD in the ENSEMBL Genome Browser was eliminated as it contained a premature stop codon (produced by an error in gene expression) as it could produce a mutation that has the potential to be detrimental to the health and function of the cell.

Exercises

Exercise 1

1. Order the following steps to make an appropriate workflow for detecting differential expression using RNA-seq (30%):

1. Library preparation
2. Sequencing
3. Differential expression analysis
4. Read mapping
5. Quantification

Answers:

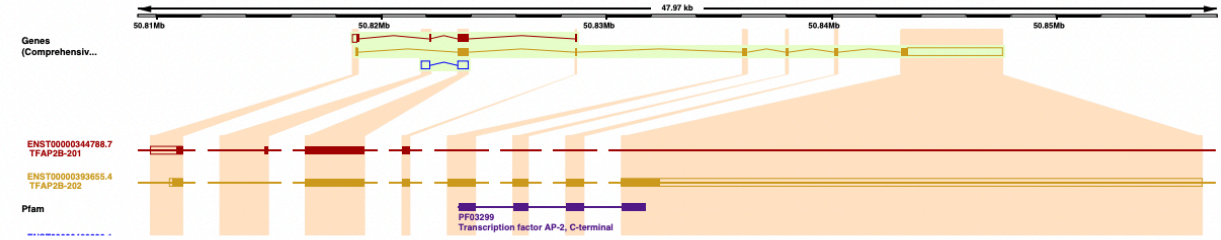
1.1) Library Preparation, Sequencing, Read Mapping, Quantification, Differential Expression Analysis

The first step is to make the library. Sequencing and then read mapping is used to match sequences and the data is quantified to count the number of reads that map with programmes like HTSeq-count. Finally, DSEQ is run with bioinformatics software to get results.

Exercise 2

2. The transcription factor TFAP2B has two isoforms, one of which is missing the DNA binding domain (identified with the id PF03299).

The two isoforms are shown in the image below, the shorter one in red and the longer one in yellow. Each box is an exon, and the lines are the introns. The domain is depicted in purple.



Perform a pairwise sequence alignment using the two sequences below, describe the result reporting:

- The % of identity
- The % of similarity
- A screenshot of the alignment where you highlight the part of the missing domain

>ENST00000344788.7|198
MLWKLVENVKYEDIYEMLVHTYSSMDRHDGVPSHSSRLSQLGSVSQGPYSSAPPL
SHTPSSDFQPPYFPPPYQPLPYHQSQDPYSHVNDPYSLNPLHQPQQHPWGQRQR
QEVGSEAGSLLPQPRAALPQLSGLDPRRDYHSVRRPDVLLHSAHHGLDAGMGDS
LSLHGLGHPGMEDVQSVEDANNSGMNLLDQSVIKK

>ENST00000393655.4|460
MHSPPRDQAAIMLWKLVENVKYEDIYEDRHDGVPSHSSRLSQLGSVSQGPYSSAP
PLSHTPSSDFQPPYFPPPYQPLPYHQSQDPYSHVNDPYSLNPLHQPQQHPWGQR
QRQEVGSEAGSLLPQPRAALPQLSGLDPRRDYHSVRRPDVLLHSAHHGLDAGMG
DSLSLHGLGHPGMEDVQSVEDANNSGMNLLDQSVIKKVPVPPKSVTSLMMNKDGF
LGGMSVNTGEVFCSVPGRLSLLSSTSKYKVTVGEVQRRLSPPECLNASLLGGVLR
RAKSKNGGRSLRERLEKIGLNLPAGRRKAANVTLLTSLVEGEAVHLARDFGYICETE
FPAKAVSEYLNRQHTDPSDLHSRKNMLLATKQLCKEFTDLLAQDRTPIGNSRPSPIL
EPGIQSCLTHFSLITHGFGAPAICAALTALQNYLTEALKGMDKMFLNNTTTNRHTSG
EGPGSKTGDKEEKHRK

Answers:

2.1) The % of identity is 40.3%

2.2) The % of similarity is 40.3%

2.3) The image of the missing part of the domain in the alignment on the right. **(A)**

A

Explanation:

EMBOSS *Needle* Global alignment was used first because it includes gaps (EMBOSS *Water* local alignment does not). This way we can see that the sequences are actually very similar, apart from the amino acids of the missing domain at the end (highlighted in blue). The gap is 280/469(59.7%). The final gap is the missing domain: an isoform of TFAPB missing the DNA binding domain.

The identity of 40.3% suggests a 100% correspondence and precise nucleotide matching of 189 identical nucleotides of the 469 total.

A similarity of 40.3% infers the resemblance between the isoform sequences and thus quantifies the similarity between the proteins. It includes perfect matches and close mismatches. If identity and similarity are the same %, there are no close mismatches between the two sequences, only perfect matches or gaps. To refine our results, we could run a local alignment with water afterwards to gauge the similarity between the region without big gaps as 95.5%.

Question 3

3) For this exercise we will use the count files from a study (Trapnell et al 2012) where they studied the loss of the developmental transcription factor HOXA1 in lung fibroblasts.

- Download the count files from this link and save them on your computer: https://www.dropbox.com/sh/en8ezqqk43e3ys1/AADJnPufUZ4F3PyRQZ6Kse5sa?dl=0
- Login in Galaxy
- Upload the 6 files by using the "upload data" button on the left and click on "Choose local file". Select the 6 files and click on Start. The 6 files should show up in your history.

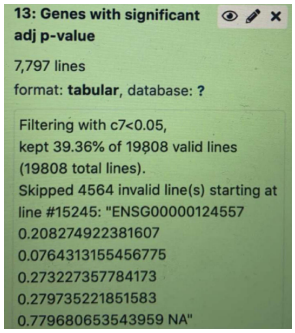
Perform a differential expression analysis using Deseq2 and answer the following questions:

3.1) how many genes are differentially expressed between control and Hoxa-depleted samples using a p-value cutoff of 0.05?

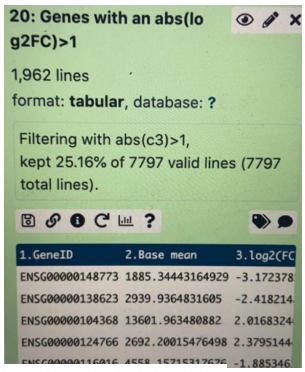
3.2) how many genes have a log fold change greater than 1?

3.3) attach and describe the plots resulting from the analysis

A



B



Answers:

3.1) There are 7797 genes **(A)**

3.2) There are 1962 genes **(B)**

3.3) Images (x5) on the right

1

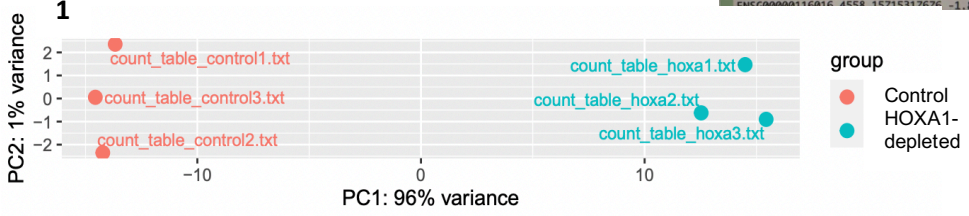


Figure 1: This is a Principal Component Analysis (PCA) plot which plots the count files as 6 individual data points which visualises covariates and batch effects. It shows clusters of samples based on their similarity: the first dimension of DESeq is PC1 which separates the control from HOXA1 developmental transcription factor (depleted) group with 96% variance and the second-dimension, PC2, which separates the 3 count files within each of the two groups from each other which results in 1% variance. The plot visualises the samples from the control group and the group with a loss of HOXA1 in a 2D plane across their first two principal components.

2

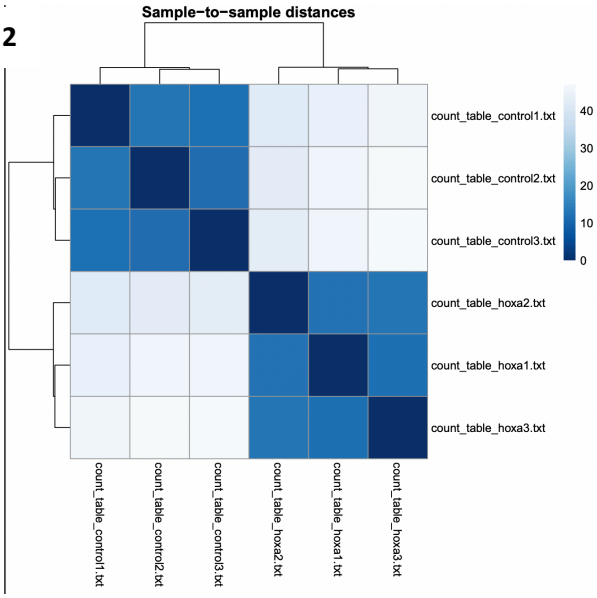


Figure 2: Heat map distance matrix of similarities and dissimilarities between the 3 HOXA1-depleted datasets and the 3 controls. It is a matrix calculated based on clustering (sample similarities): dark blue represents a shorter distance which means closer samples given the normalized counts. The range in distance is 0 to 40 (lighter as the distance is greater). The HOXA1-depleted samples and controls are the furthest from each other and show the lightest blue. When the samples are the same, a dark shade of blue representing 0 difference is used.

3

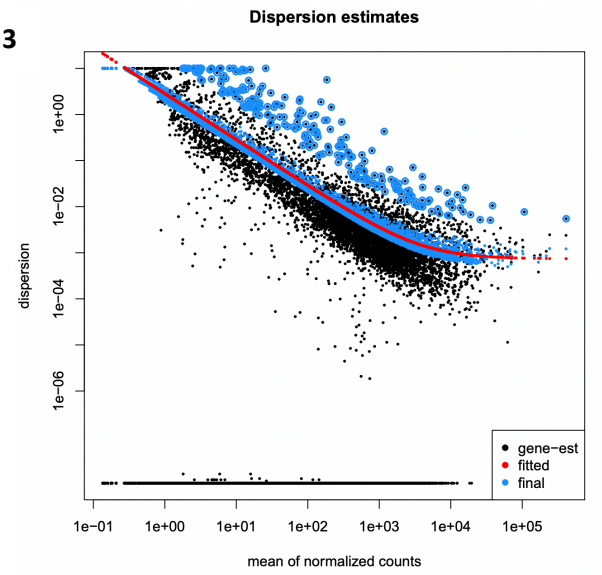


Figure 3: This is a dispersion estimate where dispersion is directly related to variance; it reflects the variance in gene expression for a given mean value. There is an inverse relationship between dispersion estimates and the mean – dispersion is higher for smaller mean counts. It shows gene-wise estimates in black, fitted values in red and the final maximum a posteriori estimates used in testing in blue. Values are plotted as the mean of normalised counts. The final estimates are shrunk from black gene-wise estimates to fitted estimates; the gene-wise estimates that are flagged as outliers are not shrunk towards the fitted estimate values.

4

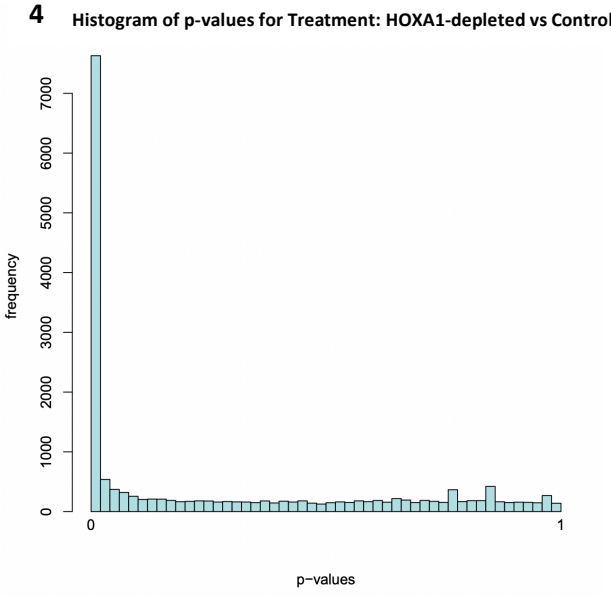


Figure 4: A histogram of p-values for the genes in comparison between HOXA1-depleted group and Control of the Treatment. The histogram frequency shows that over 7000 genes have a low p value near 0 (the cut-off being 0.05 for the differential expression analysis).

5

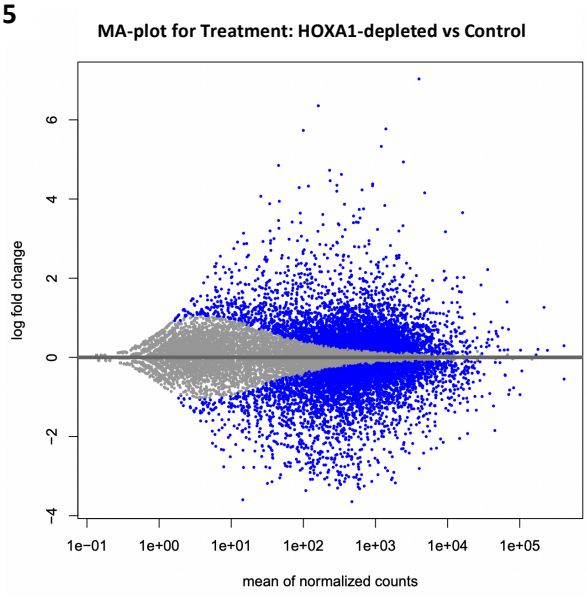


Figure 5: An MA plot which displays a global view of the relationship between the log2 fold changes: the expression change of conditions (log ratios, M), the mean expression strength of the genes (average mean, A) under normalised counts, and the algorithm's ability to detect differential gene expression. Genes that passed (are below) our significance threshold of $p < 0.05$ are in blue so genes with a specific log fold change can be visualised.

Question 4

4. In this exercise, you will be demonstrating how you learned to navigate the Ensembl genome browser. We would like to explore the gene HORMAD1 in human, mouse and alpaca.

4.1) Can you find the chromosomal coordinates for this gene in each species?
4.2) How many amino acids does the longest protein have in each species
4.3) What is the paralogue for this gene – and are the paralogues the same for all 3 species?
4.4) Only in 1 of the three species the gene is on the forward strand - which species is this?

Answers:

4.1) The chromosomal coordinates for the HORMAD1 gene in each species

- **Human**: Chromosome 1: 150,698,060-150,720,895 reverse strand. GRCh38:CM000663.2
- **Mouse**: Chromosome 3: 95,466,988-95,494,982 forward strand. GRCm39:CM000996.3
- **Alpaca**: GeneScaffold_2986: 589,209-607,603 reverse strand.

4.2) How many amino acids does the longest protein have in each species?

- **Human**: 394 aa
- **Mouse**: 392 aa
- **Alpaca**: 392 aa

4.3) What is the paralogue for this gene- and are the paralogues the same for all 3 species

- The paralogues are the **same** for all 3 species
- **Human**: ENSG00000176635, HORMAD2, HORMA domain containing 2 [Source:HGNC Symbol;Acc:HGNC:28383]
- **Mouse**: ENSMUSG00000020419 , Hormad2, HORMA domain containing 2 [Source:MGI Symbol;Acc:MGI:1923078]
- **Alpaca**: ENSVPAG00000009202 , HORMAD2, HORMA domain containing 2 [Source:HGNC Symbol;Acc:HGNC:28383]

4.4) Only in 1 of the three species the gene is on the forward strand - which species is this?

- The **mouse** is the one and only of the species which has the gene in its forward strand

Question 5

5) The gene EGFR is often found mutated in lung cancer. Use the cosmic database:
5.1) find the location within the canonical EGFR gene with highest mutation rate
5.2) in cosmic v 95 what is the mutation count on this position
5.3) mutations in EGFR are associated with altered sensitivities to 8 drugs in cosmicv95– can you name them?
5.4) if you explore the tissue in which point mutations of EGFR were reported, after lung, what is the 2nd most frequently affected tissue in which EGFR mutations have been found?
5.5) the p.T790M is a mutation often reported in lung carcinomas – what is the mutation in the CDS

Answers:

5.1)

- **Answer**: EGFR (COSG150): Amino acid 858 missense substitution.
- **Explanation:** In the canonical EGFR gene the highest mutation rate is at p.L858R (Substitution - Missense, position 858, L➞R)
-

5.2)

- **Answer**: 7965
- **Explanation**: The highest mutation count is 7965 at 858 aa which has an unknown CDS, c? , is 7965.
- **Consideration**: However, the highest mutation of with a known CDS location at c.2573T>G is 2677. Although the 7965 mutation's location may be unknown, it is still proven to be there so remains the highest mutation count.

5.3)

- **Answer**:
    1. AZD3759
    2. Erlotinib
    3. Osimertinib
    4. Afatinib (GDSC1)
    5. Gefitinib
    6. Sapitinib
    7. Bortezomib
    8. Afatinib (GDSC2)
- **Consideration**: Afatinib is listed twice on the database with each drug coming from a different dataset (GDSC1 and GDSC2)

5.4)

- **Answer**: CNS – Central Nervous System
- **Explanation**: If you look at the percentage point mutations, the second most is lung which is 26.39% affected (mutated) and the CNS is 2nd (after lung) but 3rd overall at 9.38% affected.
- **Consideration**: For non-standardised tested point mutations of EFGR, the most is lung with 104646. The second most is the breast with 11011 point-mutations. This however, does not give a standardised 'affected' interpretation between the tissues like % mutated values.

5.5)

- **Answer:** The known CDS mutation is c.2369C>T.
- **Explanation:** p.T790M is a substitution (Missense, position 790, T➞M). c.2369C>T is a substitution at position 2369 from C➞T.