**Assessment Cycle 2**

**Summary**

A single-nucleotide polymorphism (SNP) in the transcription factor, forkhead protein (FKHR) mRNA, is suspected to be the underlying cause of a genetic disease by changing gene expression. Using variant calling on Illumina sequencing data, a substitution at position 1103 was identified as the most likely to be the disease-causing single-nucleotide polymorphism (SNP), changing cytosine to guanine and resulting in a Proline-to-Arginine amino acid change.

Differential analysis of RNA-seq comparing mutant and wild-type transcription factors revealed thousands of differentially expressed genes (DEGs). Gene ontology (GO) enrichment analysis of the DEGs revealed the mutation effects on transcription associated with upregulation of mitotic processes, chromosome segregation, organelle fusion, and mRNA, while downregulating ion homeostasis and autophagy pathways.

**Part A**

Using the Windows command line interface, Ubuntu (Barnes, 2021), Illumina sequencing reads were mapped to the reference sequence to identify the engineered amino acid change in the transcription factor (TF) gene. The reads were aligned using BWA-MEM, an algorithm optimised for high-throughput mapping of short reads with high accuracy (Li, 2013).

SAM files were converted to BAM format using SAMtools for efficient data structuring (Clarke et al., 2012). The haplotype-based variant detector, FreeBayes (Garrison and Marth, 2012), was used for variant calling, inferring SNPs using a Bayesian statistical framework. Variants were validated using Integrative Genomics Viewer (IGV; Robinson et al., 2011; Figure 1). **A screen shot of a computer screen

AI-generated content may be incorrect.**

**Figure 1. Integrative Genomics Viewer (IGV) visualising SNPs identified in the forkhead protein (FKHR) mRNA transcription factor coding sequence.**

The three homozygous SNPs were identified as C1103G, T1142C, and T1878A. C1103G induces a non-conservative proline-to-leucine change, which alters the rigid, helix-breaking residue to a positively charged, flexible one. Such changes can disrupt secondary structure, potentially altering protein folding, transcriptional regulation ability and functional domains in DNA-binding proteins like transcription factors (MacArthur and Thornton, 1991; Carland et al., 2004). In contrast, T1142C and T1878A induce phenylalanine-to-leucine substitution, which both have similar size and hydrophobicity, qualities that are typically well-tolerated substitutions in protein structure and function (Brender and Zhang, 2015).

All candidates had high QUAL/support scores and produced an amino acid change. These amino acid changes are under the assumption that the reference is a clean coding sequence starting at position 1 and that it's in frame.

**Table 1. Summary of homozygous SNPs identified in the forkhead protein (FKHR) mRNA transcription factor coding sequence.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Mutation** | **Codon Change** | **Amino Acid (AA) Change** | **AA Property Change** | **Quality Score** | **Depth** | **Alternative Observations** |
| C1103G | CCC🡺CGC | Proline 🡺Arginine | Non-polar 🡺Positively Charged | 101150 | 6809 | 6672 |
| T1142C | TTC🡺CTC | Phenylalanine 🡺Leucine | Aromatic 🡺 Aliphatic (hydrophobic) | 107039 | 7130 | 7002 |
| T1878A | TTC🡺ATC | Phenylalanine 🡺 Isoleucine | Aromatic 🡺 Aliphatic (hydrophobic) | 106998 | 7214 | 7061 |

However, the translated reference sequence was input into InterPro to assess functional domains. C1103G and T1142C, at translated positions 368 and 381, were not within any functional domains but were at the centre of the FOXO family. Only T1878A, translated to position 626, was identified as within the transactivation domain (TAD; PF16676) of the FOXO protein family (593-634). This domain is known for its promiscuous binding abilities, interacting with CBP/p300’s KIX domain at two distinct binding sites, TAZ1/TAZ2 domains simultaneously, allowing for flexible binding (Figure 2). Changes like an SNP here could affect the protein's ability to recruit coactivators, altering gene expression. Therefore, T1878A has a clear potential to affect the transcriptional capacity of the protein, meaning that this is likely the engineered mutation (Wang et al., 2012).

**A screenshot of a computer

AI-generated content may be incorrect.**

**Figure 2. Functional domain analysis of the forkhead protein (FKHR) mRNA transcription factor using InterPro.** The functional domains of the transcription factor were used to identify the placement of three SNPs as C1103G (position 368), T1142C (position 381), and T1878A (position 626).

**Part B**

Sample metadata and a matrix containing RNA-seq read counts were loaded into R and processed with the DESeq2 library (Love, Huber and Anders, 2014). This package performs median ratios normalisation to account for RNA composition bias and sequencing depth. Statistical noise was reduced by removing genes with less than 10 total counts. The mutation (mut) was used in the design, with the control as the reference (ctl). The Wald test was used to identify differentially expressed genes (DEGs), adjusting p-values with the Benjamini-Hochberg method.

**Reproducibility among replicates**

The PCA plot of the rlog-transformed counts depicts a clear separation between control and mutant samples along PC1 (99% variance), clustering replicates of each group (Figure 3). PC2 accounted for no variance, but only the mutant replicates clustered.

A graph with numbers and dots

AI-generated content may be incorrect.

**Figure 3. PCA of rlog-transformed data across control and mutant samples.**

The sample distance matrix confirmed the high intra-group similarity of the mutant samples, less similarity of the control samples, and distinct inter-group differences (Figure 4).

A diagram of a graph

AI-generated content may be incorrect.

**Figure 4. Sample distance matrix across control and mutant samples.**

**Differential Expression**

The MA plot presents log2 fold change against the mean normalised counts, highlighting the 3419 upregulated and 3543 downregulated genes in blue (Figure 5).

A graph of a number of dots

AI-generated content may be incorrect.

**Figure 5. MA plot comparing mutant and control conditions.**

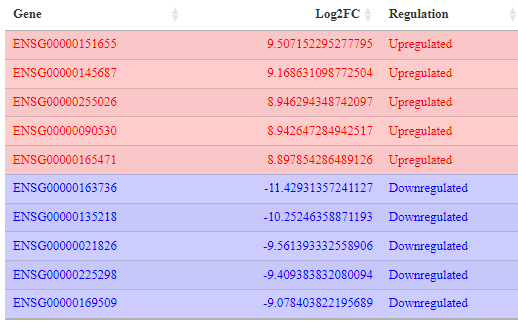
The most significant DEG, *ENSG00000025039*, was strongly downregulated in the mutant condition, with normalised counts reducing from over 2000 in the control to less than 200 in the mutant samples (Figure 6). Table 2 presents the top upregulated and downregulated DEGS.

A white rectangular graph with black text

AI-generated content may be incorrect.

**Figure 6. ENSG00000025039 expression across control and mutant samples.**

**Table 2. Top upregulated and downregulated differentially expressed genes across mutant and control samples.**



**Pathway analysis**

GO enrichment analysis was performed using the clusterProfiler library to retrieve biological process (BP) terms (Yu et al., 2012). The mutant samples upregulated pathways associated with mRNA processing, chromosome segregation, organelle fusion, and nuclear mitotic division (Figure 6). These processes are intricately associated with FOXO transcription factor activity, playing a pivitol role in regulating genes linked to mitotic fidelity and cell cycle progression (Wang, 2021). In contrast, downregulated pathways included monoatomic ion/cation homeostasis, macroautophagy, lipid catabolic process, and protein localisation to the cell periphery (Figure 7)

A graph with red and blue dots

AI-generated content may be incorrect.

Figure 7. GO enrichment dot plot for upregulated genes in the mutant samples.

A graph with dots and lines

AI-generated content may be incorrect.

Figure 8. GO enrichment dot plot for downregulated genes in the mutant samples.