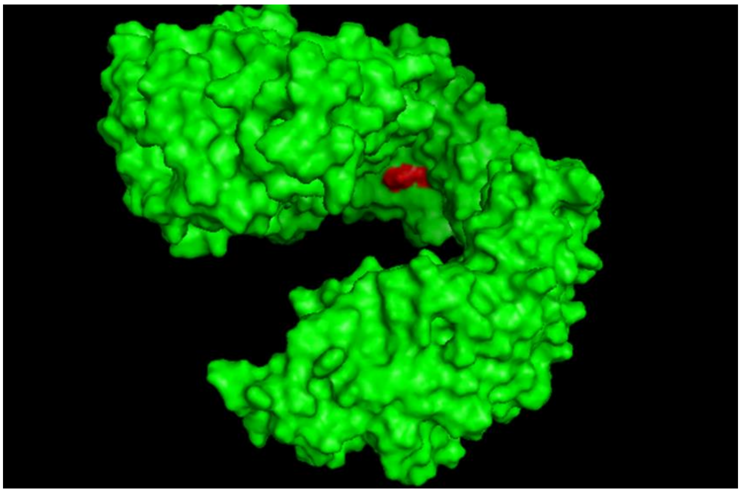
Understanding the role of novel cancer-related gene C1ORF112 in relation to DNA damage and repair



Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

By

Jacob Unekwu-ojo Edogbanya

February 2022

Supervisors: Drs Lesley Iwanejko, Jason Parsons, Pedro de Magalhaes and Simon Tew.

Abstract

The improvement of DNA sequencing technology has increased the amount of genetic information available. This has also increased the types of analysis that can be carried out concerning the information available. Linkage analysis is one of the various tools utilised in understanding genetic diseases especially when they have recognisable phenotypes. Genetic mapping and genotyping have used to correlate neurological diseases with their chromosomal location, in addition, a single-gene or multi-gene involvement in disease state can also be established.

Functionally characterizing the new genes with the amount of sequence information is the new bottleneck in bioinformatic genomics. Many genes are now being revealed and gene association studies are being carried out to functionally annotate novel genes. C1ORF112 is a gene whose functional attributes are currently understudied. C1ORF112 appears to be strongly co-expressed with DNA repair and proto-oncogenes such as RAD51 and CCDC6. C1ORF112 is also co-expressed with many genes in the BRCA-Fanconi anaemia (FA) DNA damage response pathway, including BRCA1, BRCA2, FANCD2, and FANCI.

Using bioinformatic analyses it was determined that C1ORF112 is a well-conserved protein to Opisthokonts, particularly in Primates and other Metazoa and land plants. C1ORF112 is also co-expressed with genes associated with chromosome integrity, segregation, and cell replication. Model structures for C1ORF112 were also generated, to determine that C1ORF112 is an alpha-helical protein. The structural analysis has also determined that C1ORF112 has two possible sites of phosphorylation one at the N-terminus and the other at the C-terminal.

This thesis has also shown that C1ORF112 is a cytoplasmic protein and when C1ORF112 is knocked down by CRISPR in cells, it affects the growth rate for the first 24hr after replating, but this recovers afterwards. C1ORF112 knockdown also increases the sensitivity of the cells to agents of DNA damage, especially x-ray radiation and hydrogen peroxide. This thesis concluded that while there is currently no strong evidence that C1ORF112 has a direct role in DNA damage repair, its absence appears to affect cell growth and replication.

Acknowledgements

I would like to thank my dad for sponsoring this doctorate program and my entire education. I would also like to thank Dr Joao de Magalhaes for offering the project to me and for his patience. I would also like to thank, Dr Jason Parson for allowing me into his lab even for the brief period I was there and for his tutelage, support, and knowledge. I would also like to Dr Leslie Iwanejko and Dr Simon Tew for being there at the most difficult period of this project, for their support and guidance.

To everyone in both labs that helped me during my work and my family for their support.

A very special thanks to my partner Leah and my son Malachy, my north star, for bearing with me through this entire period.

Table of Contents

Abstract 2

Acknowledgements 3

Abbreviations 11

Chapter 1: Introduction 15

1.1 DNA and The Genome 15

1.1.1 DNA Structure 16

1.1.2 Chromosome and Chromosome structure 18

1.1.3 Chromatin modification 19

1.1.4 DNA Damage 20

1.1.5 Oxidation 23

1.1.6 Ultraviolet radiation and Ionizing radiation 24

1.1.7 Repair pathway conservation 25

1.1.8 Cell cycle and Cell cycle checkpoints 25

1.1.9 DNA Damage Response and Repair 30

1.2 Protein Conservation 33

1.3. Protein structure 36

1.3.1 Determination of Protein 3D structure 37

1.3.2 Experimental determination of protein 3D structure 37

1.3.3 Computational modelling of Protein 3D structure 38

1.3.4 Template based methods 39

1.4 Structure analysis 40

1.5 C1ORF112 41

1.6 Aims 43

Chapter 2: Methods 44

2.1 Bioinformatic analysis 44

2.1.1 Sequence analysis 44

2.1.2 Sequence alignment and phylogenetic tree 44

2.1.3 Expression of C1ORF112 across tissues 45

2.1.4 Analysis of genes co-expressed with C1ORF112 and possible physical interactors. 45

2.2 Structural analysis 45

2.2.1 I-TASSER 45

2.2.3 Ramachandran plot analysis 46

2.2.4 QMEAN and Z-score 46

2.3 Experimental analysis 46

2.3.1 Laboratory Materials 46
2.3.1 Mammalian cell line culture 47
2.3.2 Thawing cells 48
2.3.3 Passaging cells 48
2.3.4 Long-term storage of cells 48
2.3.5 Seeding cells 48
2.3.6 Harvesting cells 49
2.3.7 Cell fractionation 49
2.3.8 Measuring protein concentration 49
2.3.9 Antibodies 49
2.3.10 Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) 50
2.3.11 Immunoblotting 51
2.3.12 Agarose gel electrophoresis 52
2.3.13 Induction of oxidative stress 52
2.3.14 Clonogenic assays 52
2.3.15 Neutral comet assay 53
Chapter 3: Conservation, Expression and Co-expression of C1ORF112 54
3.1 Introduction 54
3.2 Sequence conservation and phylogenetic analysis 55
3.3 Expression of C1ORF112 in tissues and co-expression analysis 60
3.4 Analysis of Genes co-expressed of C1ORF112 71
3.5 Possible physical interaction with C1ORF112 79
3.6 Discussion 80
Chapter 4: C1ORF112 is an alpha-helical protein with a possible kinase domain 83
4.1 Introduction 83
4.2 Alpha helical model generation using I-TASSER 83
4.3 Hydrophobic distribution of residues on the C1ORF112 model 101
4.4 Post-translational modification of C1ORF112 102
5.5 Discussion 106
Chapter 5: C1ORF112 knockout cells showed no phenotypic difference with normal cells and microarray analysis show cell cycle association 109
5.1 Introduction 109
5.2 Western blot to ensure knock out of C1ORF112 in cells 110
5.3 C1ORF112 mutant cells grow the same as normal cells 112
5.4 Differential gene expression between C1ORF112 knockout cells and C1ORF112 normal cells115
5.5 Pathways affected by knockdown of C1ORF112 127

6.6 Discussion...137

Chapter 6: Results IV: C1ORF112 knockout cells are sensitive to x-rays and hydrogen peroxide treatments compared to normal cells 139

6.1 Introduction 139

6.2 C1ORF112 is a cytoplasmic protein...140

6.3 C1ORF112 mutant cells are sensitive to x-rays compared to C1ORF112 normal cells............140

6.4 C1ORF112 mutant cells are sensitive to hydrogen peroxide compared to C1ORF112 normal cells 148

6.5 Discussion...153

7: Discussion...155

8: References 163

List of Figures

Figure 1. 1 **Schematic of DNA structure and pairing of double-stranded DNA**. 17
Figure 1. 2 **Chromosome and sister chromatids**. 19
Figure 1. 3 **Schematic of cell cycle**. 26
Figure 1. 4 **Schematic of the function of the Aurora kinases in chromosome segregation**. 29
Figure 1. 5 **Amphitelic attachment of the kinetochore during prometaphase**. 30
Figure 1. 6 **DNA damage and corresponding repair pathways**. 32
Figure 1. 7 **The Fanconi-Anaemia (FA) pathway**. 42

Figure 3. 1 **Multiple sequence alignment of representative organisms.** 58
Figure 3. 2 **C1ORF112 gene tree**. 59
Figure 3. 3 **Relative expressions of C1ORF112 across various cells in the body**. 62
Figure 3. 4 **Age-related expression of C1ORF112 in Vagina, Uterus, Cervix uteri, Ovary and Fallopian tube**... 63
Figure 3. 5 **Age-related expression of C1ORF112 in the Thyroid and the Testis**. 64
Figure 3. 6 **Age-related expression of C1ORF112 in the Stomach, the Spleen, and the Oesophagus**. 65
Figure 3. 7 **Age-related expression of C1ORF112 in the Small intestine and the Skin**. 66
Figure 3. 8 **Age-related expression of C1ORF112 in the Salivary gland and the Prostate**. *Salivary gland showed the largest variance in C1ORF112 expression, and the prostate showed similar average levels of expression across all age groups* 66
Figure 3. 9 **Age-related expression of C1ORF112 in the Pituitary gland and the Pancreas**. 67
Figure 3. 10 **Age-related expression of C1ORF112 in the Muscle and the Brain and Heart**. 68
Figure 3. 11 **Age-related expression of C1ORF112 in the Lung and the Liver, the Kidney, and the Colon**. 69
Figure 3. 12 **Age-related expressions of C1ORF112 in the breast, the Nerve, the Blood vessel, the Blood, and Bladder**. 70
Figure 3. 13 **KEGG pathway over-representation in Anatomical parts, Cancers, and Cell lines**. 75
Figure 3. 14 **Possible physical interactors with C1ORF112**. 79

Figure 4. 1 **Reference Ramachandran plot**. 84
Figure 4. 2 **Secondary structure of Model 1 of C1ORF112 generated using I-TASSER**. 85
Figure 4. 3 **Model 1 analysis using Ramachandran plot and QMEAN analysis**. 86
Figure 4. 4 **Secondary structure of Model 2 of C1ORF112 generated using I-TASSER.** 88
Figure 4. 5 **Model 2 analysis using Ramachandran plot and QMEAN analysis**. 89
Figure 4. 6 **Secondary structure of Model 3 of C1ORF112 generated using I-TASSER.** 90
Figure 4. 7 **Model 3 analysis using Ramachandran plot and QMEAN analysis.** 91
Figure 4. 8 **Secondary structure of Model 4 of C1ORF112 generated using I-TASSER.** 92
Figure 4. 9 **Model 4 analysis using Ramachandran plot and QMEAN analysis.** 93
Figure 4. 10 **Secondary structure of Model 5 of C1ORF112 generated using I-TASSER.** 95
Figure 4. 11 **Model 5, Ramachandran plot and QMEAN analysis.** 96
Figure 4. 12 **Model of C1ORF112 showing the position of the WCF tripeptide position**. 98
Figure 4. 13 **Primary structure of C1ORF112 generated by ALPHAFOLD prediction server**. 99
Figure 4. 14 Ramachandran and QMEAN analysis of AlphaFold model of C1ORF112. 100
Figure 4. 15 **Hydrophobic distribution of C1ORF112**. 102
Figure 4. 16 **Position of post-translational modification along the sequence of C1ORF112**. 105
Figure 4. 17 **Surface area resolution of the AlphaFold model of C1ORF112.** 106

Figure 5. 1 **Western blot analysis between normal C1ORF112 and mutant C1ORF112**. 111
Figure 5. 2 **Comparison of the cell growth between Normal HAP1 and Mutant HAP1**. 113
Figure 5. 3 **Volcano plot showing the differential expression of upregulated genes (red) and down-regulated genes (green) at 24hrs.** 116
Figure 5. 4 **Volcano plot showing the differential expression of upregulated genes (red) and down-regulated genes (green) at 48hrs.** 117

Figure 5. 5 ***Overall expressions of C1ORF112 compared to genes selected from the microarray data and β-actin as a housekeeping gene***.......... 126
Figure 5. 6 ***PI3K-Akt signalling pathway with the DEGs***.......... 131
Figure 5. 7 ***PI3K-Akt signalling pathway with the DEGs***.......... 132
Figure 5. 8 ***PI3K-Akt signalling pathway with the DEGs***.......... 133
Figure 5. 9 ***PI3K-Akt signalling pathway with the DEGs***.......... 134
Figure 5. 10 ***PI3K-Akt signalling pathway with the DEGs***.......... 135
Figure 5. 11 ***Cell cycle pathway***.......... 136

Figure 6. 1 ***Cell fractionation of mutant HAP1 cells indicating cytoplasmic presence of C1ORF112***. 140
Figure 6. 2 ***Representative images of clonogenic seeding densities***.......... 141
Figure 6. 3 ***Representative plates showing a reduction in cell colonies after treatment with x-rays.*** 142
Figure 6. 4 ***Sensitivity of mutant HAP1 cells to x-rays***.......... 142
Figure 6. 5 ***Generation of DSBs after x-ray exposure.*** 144
Figure 6. 6 ***Repair kinetics after generation of DSBs after exposure to x-rays***.......... 147
Figure 6. 7 ***Representative plates showing a reduction in cell colonies after treatment with hydrogen peroxide.***.......... 148
Figure 6. 8 ***Sensitivity of mutant HAP1 cells to hydrogen peroxide***.......... 149
Figure 6. 9 ***Generation of DSBs after hydrogen peroxide exposure.***.......... 150
Figure 6. 10 *R**epair kinetics after exposure to hydrogen peroxide***. 152

List of Tables

Table 1. 1 **Summary of DNA repair mechanisms***.* 31

Table 2. 1 **Primary antibody***.* 50
Table 2. 2 **Secondary antibodies.** 50

Table 3. 1 **List of representative C1ORF112 homologues in various species.** 56
Table 3. 2 **Top 25 genes positively co-expressed with C1ORF112.** 73
Table 3. 3 **Gene ontology (GO) terms for biological processes that are over-represented in genes positively co-expressed with C1ORF112.** 74
Table 3. 4 **Top 25 genes negatively co-expressed with C1ORF112 in Anatomical parts, Cancers, and Cell lines** 77
Table 3. 5 **Gene ontology (GO) terms for biological processes that are over-represented in genes negatively co-expressed with C1ORF112.** 78

Table 4. 1 **QMEAN analysis of Model 1** 87
Table 4. 2 **QMEAN analysis of Model 2.** 90
Table 4. 3 **QMEAN analysis of Model 3.** 92
Table 4. 4 **QMEAN analysis of Model 4.** 94
Table 4. 5 **QMEAN analysis of Model 5.** 96
Table 4. 6 **Comparison of the QMEAN and Z-scores of Models 1, 3 &4.** 98
Table 4. 7 **Possible phosphorylation sites on C1ORF112***.* 104

Table 5. 1 **Top 25 DEGs downregulated in C1ORF112 knockout cells versus wild type at 24 hours post-seeding***.* 118
Table 5. 2 **Top 25 DEGs upregulated in C1ORF112 knockout cells versus wild type at 24 hours post-seeding***.* 119
Table 5. 3 **Gene ontology (GO) terms for biological processes that are over-represented in down-regulated genes differentially expressed at 24hr post-seeding.** 120
Table 5. 4 **Gene ontology (GO) terms for biological processes that are over-represented in up-regulated genes differentially expressed at 24hr post-seeding.** 121
Table 5. 5 **Top 25 DEGs downregulated in C1ORF112 knockout cells versus wild type at 48 hours post-seeding***.* 122
Table 5. 6 **Top 25 DEGs upregulated in C1ORF112 knockout cells versus wild type at 48 hours post-seeding***.* 123
Table 5. 7 **Gene ontology (GO) terms for biological processes that are over-represented in down-regulated genes differentially expressed at 48hr post-seeding.** 124
Table 5. 8 **Gene ontology (GO) terms for biological processes that are over-represented in up-regulated genes differentially expressed at 48hr post-seeding.** 125
Table 5. 9 **Pathways with the most DEGs after C1ORF112 knockout***.* 128
Table 5. 10 **Pathways with the least number of DEGs after C1ORF112 knockout.** 129

Abbreviations

2D – 2 dimensional

3D – 3 dimensional

γH2AX – gamma H2AX

A – Adenine

AP site – Apurinic/apyrimidinic site

ATM – Ataxia–Telangiectasia Mutated

ATR – Ataxia Telangiectasia and Rad3-related

AURKB – Aurora kinase B

BER – Base Excision Repair

BRCA1 – Breast cancer-associated protein 1

BRCA2 – Breast cancer-associated protein 2

BSA – Bovine serum albumin

BUB1 – Budding uninhibited by benzimidazoles 1

C – Cytosine

CDC25A – Cdc25A M-phase inducer phosphatase 1

CDC2 – Cyclin-dependent kinase 1

CDKs – Cyclin-dependent kinases

CDK4 – Cyclin-dependent kinase 4

CENPA – Histone H3-like centromeric protein A

CHK1 – Checkpoint kinase 1

CHK2 – Checkpoint kinase 2

CPD – Cyclobutane pyrimidine dimers

DDR – DNA damage response

DEG – Differentially expressed genes

DMC1 – Meiotic recombination protein

DMEM – Dulbecco's Modified Eagle Medium

DMSO – Dimethyl sulfoxide

DNA – Deoxyribose Nucleic Acid

DNA-PK – Deoxyribose Nucleic Acid protein kinase

DSB – Double-strand break

DSBR – Double-strand break repair

EDTA – Ethylenediaminetetraacetic acid

FBS – Foetal bovine serum

G – Guanine

Gy – Gray

H2AX – Histone H2AX

HEPES – Hydroxyethyl piperazineethanesulphonic acid

HR – Homologous Recombination

HNE – 4-hydroxynonenal

ICL – Inter-strand crosslink

IR – Ionizing radiation

Kb – Kilobase

kDa – Kilo-Dalton

Lig – DNA ligase

MCM10 – Mini chromosome maintenance 10

MDA – malondialdehyde

MDM2 – Mouse double mutant 2 homolog

MDM10 – Mouse double mutant 2 homolog

Me – methylation

MMR – Mismatch repair

mRNA – messenger RNA

MRN complex – Mre11-Rad50-Nbs1

MSA – Multiple Sequence Alignment

MW – Molecular weight

NBS1 – Nibrin

NER – Nucleotide Excision Repair

NHEJ – Non-homologous end-joining

NTH1 – Endonuclease III-like protein 1

O^6MeG - O6-methyl-guanine

p53 – Cellular tumour antigen p53

PARP1 – Poly (ADP-ribose) polymerase 1

PBS – Phosphate buffered saline

PE – Plating efficiency

PIKKs – Phosphoinositide-3-kinase-related kinases

PMSF – Phenylmethylsulphonyl fluoride

POLE2 - DNA Polymerase Epsilon 2

ROS – reactive oxygen species

RF – Replication fork

RNA – Ribonucleic Acid

RNA-Seq – RNA-sequencing

rRNA – Ribosomal RNA

SDS-PAGE – Sodium dodecyl sulphate-polyacrylamide gel electrophoresis

SSB – Single-strand break

SSBR – Single-strand break repair

TEMED – Tetramethyl ethylenediamine

T – Thymine

TPM – Transcripts per Million

Ub – Ubiquitylation

UV – Ultraviolet radiation

Uvr – Uvr - A/B/C endonuclease

WB – Western blot

WT – Wild type

XP – Xeroderma pigmentosum

XRCC1 – X-ray repair cross-complementing protein 1

Chapter 1: Introduction

1.1 DNA and The Genome

The human genome contains the complete set of evolutionary information necessary for survival (Simonti and Capra 2015). The genome consists of protein and non-protein-coding genes, it also contains cis and trans-regulatory elements for both the protein and non-protein-coding regions, transcription, and translation information (The ENCODE Project Consortium 2004, Davidson 2006, The ENCODE Project Consortium 2012). In addition, the genome contains non-coding RNAs, micro RNAs and there are other areas of the genome whose functions are not yet elucidated (The ENCODE Project Consortium 2004, The ENCODE Project Consortium 2012). Improved sequencing technology has enabled the complete sequencing of the human genome, a work which was started in the early 90s (Gyapay, Morissette et al. 1994, Dib, Faure et al. 1996, International Human Genome Sequencing 2004). This revealed that the euchromatic portion of the genome has about 2.8 billion nucleotides. Accessibility of these nucleotides is dependent on cell type and gene expression requiring a switch from heterochromatin to euchromatin (Lorch, LaPointe et al. 1987).

Genome stability is linked to chromosome stability and maintenance. Genomic information is written sequentially using four nucleotides, monomeric subunits, linked by phosphodiester bonds that make up the backbone of the DNA. The DNA contains continuous complementary stands of polynucleotides held together by hydrogen bonds, which are then wound around histone proteins, and packed into supercoiled structures called chromosomes. DNA packed into chromosomes are stable and increase genome stability. This increases control of expression or repression of genes through chromatin remodelling (Lorch, LaPointe et al. 1987). Genes are single independent functional units of DNA. Genes usually encode for proteins, major components that enable numerous biological processes to concurrently and co-ordinately be active in every living cell. Genes are generally thought to be evolutionary conserved especially when they are involved in important functions such as homeostatic maintenance, cell cycle regulation, cell replication and genome stability and maintenance (Tatusov, Koonin et al. 1997, Neuwald, Aravind et al. 1999, Tamames 2001).

The improvement of DNA sequencing technology has increased the amount of genetic information available. This has also increased the types of analysis that can be carried out concerning the information available. Linkage analysis is one of the various tools utilised in understanding genetic diseases especially when they have recognisable phenotypes. Genetic mapping and genotyping have enabled correlation of genetic diseases with their chromosomal location, in addition, a single-gene or multi-gene involvement in disease state can also be established (Pulst 1999). Genome-wide association studies (GWAS) is another approach used to study disease-state, especially in larger populations, understanding the genetic variation with a population and how genetic drift occurs around different populations (Zhao, Gupta et al. 2011). Gene co-expression studies is yet another tool available for analysing groups of genes that may appear to be functionally related. Functionally characterizing new genes is the new bottleneck in bioinformatic genomics, with the amount of sequence information. Many genes are now being revealed and gene association studies are being carried out to functionally annotate novel genes. C1ORF112 is a gene whose functional attributes are currently understudied. C1ORF112 appears to be strongly co-expressed with DNA repair and proto-oncogenes such as RAD51 and CCDC6 (van Dam, Cordeiro et al. 2012). C1ORF112 is also co-expressed with many genes in the BRCA-Fanconi anaemia (FA) DNA damage response pathway, including BRCA1, BRCA2, FANCD2 and FANCI (van Dam, Cordeiro et al. 2012, Nalepa and Clapp 2018).

1.1.1 DNA Structure

Watson and Crick used x-ray diffraction to determine the structure of DNA (Watson and Crick 1953). The commonest DNA structure is the B-DNA, which is the right-handed double helix. DNA is comprised of monomeric nucleotides, when single-stranded, nucleotides are held together by phosphodiester bonds, the binding of the 5’ phosphate of the preceding to the 3’ carbon of the hydroxyl group on the subsequent nucleotide. The phosphodiester bond on the backbone of the DNA creates the directionality of reading the DNA sequence. When double-stranded, nucleotides on both strands interact with nucleotide specificity with Adenine (A) binding to Thymine (T) and Guanine (G) binding to Cytosine (C) (Figure 1.1).

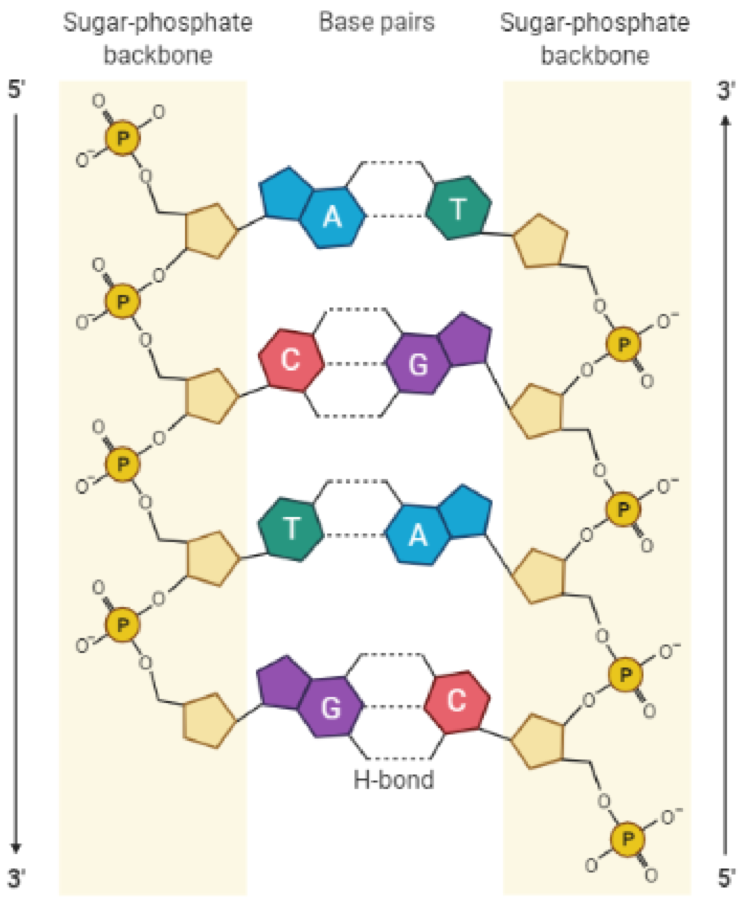


Figure 1. 1 **Schematic of DNA structure and pairing of double-stranded DNA**.

Directionality and base-pairing because of the hydrogen bonding of the Watson-Crick base pairs. The affinity of the nitrogenous bases is also shown guanine (purple) binds to cytosine (red) and adenine (blue) binds to thymine (green). Illustration created on Bio Render.

There are other forms of DNA aside from the B-form DNA, such as, Z-DNA in alternating purine pyrimidine sequences, G-quartets, triplex DNA, and these can be caused due to various phenomena such as mirror repeats, direct repeats homopurine-homopyrimidine tracts (including G-tracts), and inverted repeats (McPherson and Longo 1993, Raghavan and Lieber 2007). Non-B DNA is usually regions of the genome that contain repetitive sequences. These repetitive sequences tend to become areas more susceptible to DNA breaks, chromosomal translocation, deletions, or amplifications leading to genome instability (Wang and Vasquez 2006). These regions are also more associated with genetic diseases such as CTG and CGG triplet repeats being associated with Fragile X syndrome (Fu, Kuhl et al. 1991, Bowater and Wells 2001) and spinocerebellar ataxia type 10 (Matsuura, Yamagata et al. 2000).

The DNA in humans of which there are approximately 6.4 million base pairs are packaged into superstructures in the nucleus called chromosomes. Human somatic cells are diploid, meaning they contain 23 pairs of chromosomes, with one set inherited from either parent. On the other hand, gamete sets are haploid, containing only a single copy of each chromosome, which is passed onto the offspring.

1.1.2 Chromosome and Chromosome structure

DNA is condensed into structures called chromatin to form chromosomes. The interaction of the DNA helical strands, whereby, it wraps around a core complex of eight positively charged proteins called histones. The histone complex is made up of two copies of H2A, H2B, H3 and H4. These make an octamer complex, of which approximately 150 – 200 bases of DNA wraps around 1.65 times called a nucleosome. Histones being positively charged, and DNA being negatively charged creates a strong binding affinity. The nucleosome has a diameter of 11nm due to the tight binding of the DNA to the nucleosomes, histone H1 acts as a linker between nucleosomes by binding to short of DNA on entry and exit of the wrapping, giving beads on a string structure. The nucleosome along with the linking H1 histone is known as the chromatosome.

The way DNA is packaged in cells depends on the stage of the cell cycle at which the cell is. Chromatin is the normal state of packed DNA. Chromatin is the unravelled form of DNA to package into the nucleus. Chromosome, on the other hand, is the highest form of condensed DNA that appears during metaphase. The main function of the chromosome is to enable the proper segregation of the genome into the respective daughter cells (Becker and Horz 2002, Sif 2004). Chromatin is condensed at 50 times the normal helical turn and folded into 30nm fibres generally appear during interphase (Lee 2001, Sif 2004). They are long, thin, singular uncoiled structures that allow for the DNA to be packed into the nucleus while also allowing for the regulation of gene expression and DNA replication. The compaction of chromatin fibres first into 300nm loops and then 250nm fibre bundles leads to chromosomes, condensed 10,000 times than normal DNA (Becker and Horz 2002). As stated earlier, chromosomes appear during the metaphase of the cell cycle and anaphase. They are thick compact chromatin fibres that exist in pairs that are lined up at the cellular equator for segregation during cell division. Chromosome does not have any metabolic activity; it also provides a level of protection to the genome and allows for epigenetic gene control for certain areas of the DNA.

Chromosomes contain thousands of genes packed together sequentially. The locus is the position of a gene on the chromosome. Genes are inherited from each parent and both genes on the same locus are called alleles. If the two alleles are the same, they are genotyped as homozygous and if they are different, they are genotypes as heterozygous. Furthermore, alleles can be dominant or recessive, in addition, the chromosome also consists of telomeres, centromeres, the origin of replication and kinetochore (Lee 2001). The centromere of a chromosome is a region of condensed repetitive alpha satellite sequences, on either side of the centromere lie disc-shaped proteinaceous structures called kinetochore as shown in figure 2 (Chan, Liu et al. 2005, Shen 2019). The centromere alongside the kinetochore allows for microtubule attachment and chromosome motility during segregation (Chan, Liu et al. 2005). The centromeres also serve as connecting points between two sister chromatids. Telomeres are found at the ends of the chromosome and serve as caps to prevent DNA degradation.

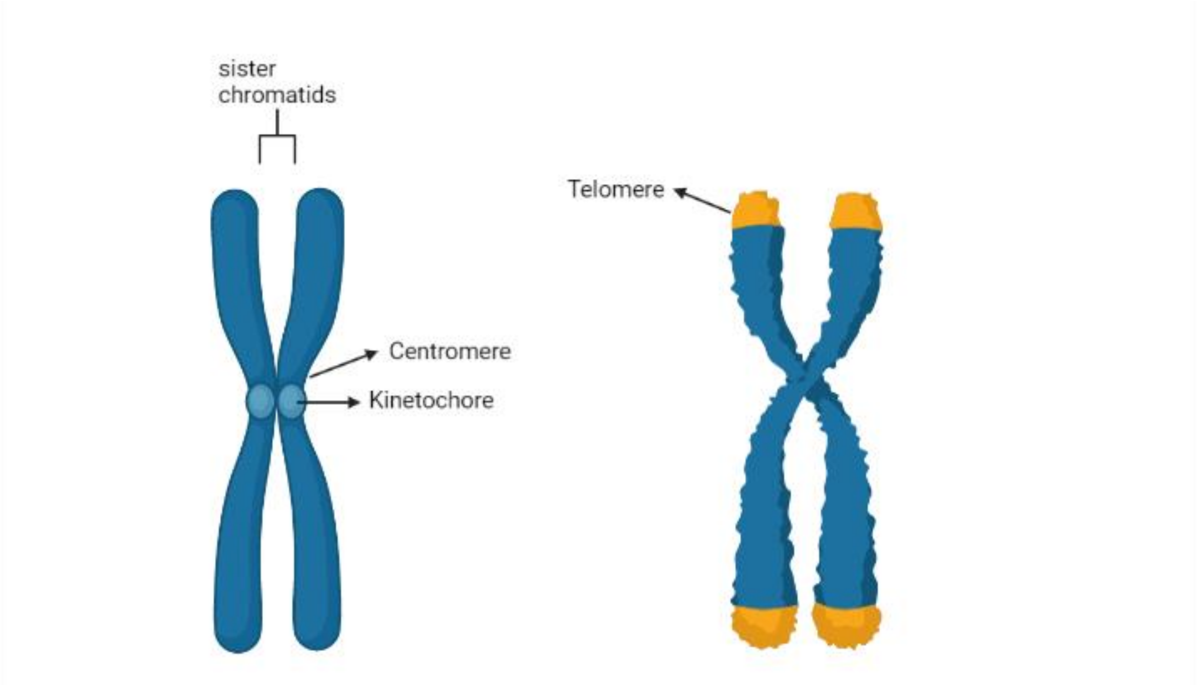


Figure 1. 2 **Chromosome and sister chromatids**.

Image indicating the location of telomere, centromere, and kinetochore. Kinetochore holds the sister chromatids together until the time for separation into two cells. Illustration created on Bio Render.

1.1.3 Chromatin modification

Repression or expression of gene expression is determined by the accessibility of chromatin by DNA replicating proteins. Chromatin structure can either be open (euchromatin) or closed (heterochromatin) (Lee 2001). In addition, the tight packing of the heterochromatin region can have downstream effects, such as preventing gene expression for genes surrounding the region and not just genes present within the heterochromatin region. Interactions with enzymes that remodel chromatin structure help switch from heterochromatin to euchromatin and vice versa. There are two types of enzymes that have been identified as major factors that affect chromatin structure. The first group of enzymes are ATP hydrolysing enzymes that remodel chromatin at the nucleosome level and the second are enzymes that catalyse post-translational modifications of histones (Wolffe 1998, Kouzarides 2002). Both activities act on the nucleosome.

Chromatin modifications usually target the histone proteins and modifications such as methylation of arginine, sumoylation, ADP ribosylation, deimination and proline isomerization regulate gene transcription. Acetylation and phosphorylation regulate transcription, replication, repair, and condensation of chromatin. Methylation of lysine and ubiquitylation regulate transcription and repair of chromatin. DNA-bound transcription factors supply chromatin with modifying enzymes such as SWI/SNF, ISWI, NuRD/Mi-2/CHD, INO80 and SWR1 which are ATP-dependent remodelling complexes (Kouzarides 2007). Chromatin re-modellers tend to be functionally specific and are involved in certain processes such as DNA repair (SWI/SNF), apoptosis or transcription repression (NuRD/Mi-2/CHD) (Kouzarides 2002, Kouzarides 2007, Wang, Allis et al. 2007).

Chromatin remodelling is an essential response step to DNA damage. Evidence has indicated that slackening of the chromatin acts to promote repair protein recruitment and is initiated by PARP1 protein followed by the re-modeller Alc1, which contains the ADP-ribose-binding domain (Ahel, Horejsi et al. 2009). Another process for chromatin relaxation involves γH2AX bound by MDC1, RNF8, and NBS1. This complex is usually formed after exposure irradiation (Stucki, Clapperton et al. 2005, Mailand, Bekker-Jensen et al. 2007, Chapman and Jackson 2008). Chromatin re-modelling is also capable of exerting tumour suppression function through fine-tuning at critical steps such as cell growth and division steps, like cell-cycle progression, DNA repair and chromosome segregation (Hanahan and Weinberg 2011)

1.1.4 DNA Damage

Alteration of the chemical structure of DNA either from strand breaks, missing bases or base modification that disrupts the helical structure of DNA can be referred to as DNA damage. There are other forms of chemical alterations that do not result in the alteration of structure of the DNA but does affect the coding information present in the sequence of the DNA such as uracil and O6meG Maintaining the integrity of the DNA is essential for cell survival, and successfully error-free replication is necessary for survival for cellular progeny. The double-helical structure does offer some level of protection from genotoxicity, DNA is still very much susceptible to various types of damage (Hoeijmakers 2009). The DNA sequence always needs a pre-existing template as it cannot be effectively copied without this. There are different sources of DNA damage, and they can either be endogenous (because of normal cellular processes and by-products) or exogenous (environmental factors such as UV, natural radioactive substances, plant alkaloids) (Friedberg, McDaniel et al. 2004). The persistence of DNA damage and resulting genomic instability and mutations can lead to many diseases such as cancer, ageing, congenital syndromes, and neurodegenerative diseases (Subba Rao 2007, Hoeijmakers 2009). They are various types of DNA damage including but not limited to the site of base loss (apurinic/apyrimidinic or AP sites), base mismatch, base oxidation, single-strand breaks

(SSBs), double-strand breaks (DSBs), DNA-protein and DNA-DNA cross-links. A summary of the different sources of DNA damage and their resulting damage are shown in Table 1.1

Table 1. 1 **Summary of types of endogenous DNA damage and approximate frequency of occurrence**.

Showing the types of DNA under the different sources; endogenous and exogenous.

Endogenous DNA damage	**Damage caused**	**Approx. Frequency**
Replication errors	Base substitutions, DNA base mismatches, single base insertion and deletion errors	10^{-6} to 10^{-8} per base (Lindahl 1993, Fatemi, Pao et al. 2005)
Hydrolysis	Loss of nitrogenous bases. SSB. DSB.	$5x10^3$ - $1x10^3$ times per cell per day (Lindahl 1993)
Oxidation (ROS)	SSB, DSB, 8-oxm m,o-G, 8-oxo-A, 5-hydroxy-C, 5-hydroxy-U, etheno-A, faPy-G, faPy-A, lipid peroxidation.	10,000 per cell per day (Lindahl and Barnes 2000)
Endogenous Alkylation	Methylation of nitrogenous bases. 5-methylcytosine, 7-methylguanine, 1-methyladenine	4000 7-methylguanine, 600 3-methyladenine and 10–30 O^6-methylguanine residues per day (De Bont and van Larebeke 2004)
Spontaneous DNA damage	DNA hydrolysis leading to deamination of adenine, cytosine, and guanine. Depurination, DNA oxidation, Non-enzymatic DNA methylation,	Rarely occurs (Lindahl 1993)

Table 1. 2 **Summary of exogenous DNA damage**.

Showing the damage caused and the sources of damage

Exogenous DNA damage	**Damage caused**	**Sources of damage**
Ionizing radiation	SSB, DSB, Dimerization, Tautomeric isomerism of nitrogenous bases (keto-enol and amino-imino)	Gamma rays, X-rays, photons
Ultra-violet radiation	cyclobutane pyrimidine dimers (CPD), 6-4 photoproducts, SSB, DSB	Sun, UV-producing lamps (Atillasoy, Seykora et al. 1998)
Alkylating agents	Methylation of nitrogenous bases. 5-methylcytosine, 7-methylguanine, 1-methyladenine	haloalkanes (dichloromethane, trichloromethane (chloroform), tetrachloromethane), alkyl sulfonates, nitrosoureas, and others (Preston, Singer et al. 1986)
Cross-linking agents	Inter-strand crosslinks	Psoralens, Mitomycin C, nitrous acids, and DNA-protein crosslinkers (Huang and Li 2013)
Polyaromatic hydrocarbons (PAHs)	DNA adducts, oxidative stress	benz[a]anthracene, benzo[b]fluoranthene, benzo[j]fluoranthene, BaP,dibenz[a,h]anthracene, 7H-dibenzo[c,g]carbazole, dibenzo[a,h]pyrene, dibenzo[a,i]pyrene, indeno[1,2,3-cd]pyrene, benzo[k]fluoranthene, dibenzo[a,e]pyrene, dibenzo[a,l]pyrene, and 5-methylchrysene (Moorthy, Chu et al. 2015)
Intercalating agents	Base deletions, intra-strand, and inter-strand crosslinks	Acridine, Ethidium bromide, actinomycin D, cisplatin

Alterations of the information content of DNA is directly related to the 'sense' of the encoded information. Alterations could occur because of compounding factors. For example, loss of bases because of DNA strand breaks, ineffective repair mechanisms, bases conversion leading to wrong bases being copied. This can lead to a variety of mutations such as missense mutations, nonsense mutations, frameshift mutations, structural alterations in DNA leading to genomic rearrangements (Crick 1966, Fujiwara, Ichihashi et al. 1981, Selby and Sancar 1997, Paiva and Bozza 2014).

1.1.5 Oxidation

Cellular oxidation is carried out by the electron transport chain in cells that perform aerobic respiration. A by-product of cellular respiration, catabolic and anabolic processes and peroxisomal metabolism is reactive oxygen species (ROS) (Henle and Linn 1997). ROS play significant cellular functions in redox signalling as cellular messengers. ROS are also vital in the innate response to pathogenic infection by the immune system especially by phagocytes or lung epithelial cells (Friedberg 2005, Malle, Furtmuller et al. 2007). When there are increased levels of ROS in cells, this can cause a variety of damages such as oxidative base lesions and 2'-deoxyribose modifications (Bjelland and Seeberg 2003, Cadet, Douki et al. 2010, Cadet and Wagner 2014). Under normal conditions, the activities of ROS are regulated in cells by restricting cellular respiration to the mitochondrial, thereby, protecting other organelles. Enzymes such as superoxide dismutase, catalase, and peroxiredoxin are antioxidants that prevent ROS damage to nucleosomes when there is surplus ROS (Riley 1994, Mates, Perez-Gomez et al. 1999, Mates and Sanchez-Jimenez 1999). Regardless, ROS accumulation is associated with many human diseases such as cancer, heart failure, diabetes, Alzheimer's, and Parkinson's (Giacco and Brownlee 2010, Liou and Storz 2010, Mohsenzadegan and Mirshafiey 2012, Hafstad, Nabeebaccus et al. 2013).

There are a few ROS species and the most prominent of these species are superoxide radicals ($O_2{\bullet}^-$), hydrogen peroxide (H_2O_2), and the hydroxyl radical (•OH) (Henle and Linn 1997). The -•OH radicals, produced as a by-product of the Fenton' reaction of H_2O_2 with Fe^{2+}, is also the most reactive of the ROS species and efficient at damaging DNA, proteins, and lipids (Imlay and Linn 1988, Dizdaroglu, Rao et al. 1991). The hydroxyl radical, is an electrophilic radical that interacts with DNA bases in a couple of ways such as removing hydrogen from methyl groups, interacting with the sugar residue (Breen and Murphy 1995, Winterbourn 2008). The mechanism by which •OH interacting with DNA molecules can be highlighted with a few examples. •OH, can attack either G or A to form an imidazole ring, generating a fragmented purine structure formamidopyrimidine (Chetsanga, Lozon et al. 1981, Friedberg, McDaniel et al. 2004, Friedberg 2005). Hydroxylation of C-8 residue of guanine at the saturated imidazole ring produces 7,8 dihydro-8-oxoguanine (8-oxo-G), which is another form of damage that can be caused by •OH, 8-oxo-guanine which pairs incorrectly with A instead of C (Kasai, Hayami et al. 1984, Cadet, Douki et al. 2010, Cadet, Douki et al. 2011, Cadet and Wagner 2014).

In mammalian cells, there is approximately 2300 SSBs due to the action of ROS species, ROS can also generate DSBs (Giloni, Takeshita et al. 1981, Henner, Rodriguez et al. 1983). In addition, lipid peroxidation by hydroxyl radicals can generate aldehyde products such as malondialdehyde (MDA) and 4-hydroxynonenal (HNE), which can react with adenine, guanine, and cytosine to form pro-mutagenic exocyclic DNA adducts, such as N^2-propanodGuo adducts, which can be repaired through base excision repair, nucleotide excision repair, mismatch repair, and AP endonuclease-mediated repair. (Xu, Wu et al. 1999, Marnett 2000, Plastaras, Riggins et al. 2000, Hang 2004).

1.1.6 Ultraviolet radiation and Ionizing radiation

Aside from endogenous damage of DNA from cellular ROS, DNA is capable of sustaining damage from exogenous sources in the form of “physical damage”. The most common sources of this form of damage are ionizing radiation (IR) and ultra-violet radiation (UV). IR releases electrons from atoms generating ions that are capable of breaking covalent bonds. Different types of IR are alpha and beta particles, neutron, gamma rays and X-rays (Borrego-Soto, Ortiz-Lopez et al. 2015). IR and UV can induce double-strands (DSB) and single-strand breaks (SSB) in DNA. Furthermore, IR can also generate ROS species, alongside the high energy electrons to damage DNA via chemical attacks. UV can generate cyclobutane pyrimidine dimers (CPDs) and 6-4 photoproducts (6-4 PPs) (Borrego-Soto, Ortiz-Lopez et al. 2015). Quiescent cells tend to be less radiosensitive compared to highly proliferative cells. Possible due to the quiescent cells having their DNA as heterochromatin and transcriptionally inactive, while proliferative cells have more euchromatin and as such more susceptible. For example, p53 is a 20Kb gene found on the short arm of chromosome 17, which encodes a 53 kDA, phospho-protein primarily involved in transcription and regulation of the cell cycle. Under normal circumstances p53 existing in an inactive state that is comparatively inefficient at binding DNA (Lakin and Jackson 1999). In the presence of UV and IR DNA damage, including radio-mimetic drugs and chemicals such as methane sulfonate (MMS), when the DNA is present in the euchromatin form, (G_0/G_1 phase), the amount of p53 is increase and then p53 peptide is phosphorylated at a serine residue in either the amino or carboxyl terminus and activated (Chen, Ko et al. 1996, Benjamin and Ananthaswamy 2007). Activation of p53 in this cycle phase leads to long term cell cycle arrest to enable the cell repair or remove the damage. Absence or mutation of p53 decreases the likelihood cell cycle arrest and inaccurate DNA damage repair, sensitizing cells to UV and IR.

An effect of radiation damage can be seen in the form of Xeroderma pigmentosum (XP). XP is a congenital condition characterised by extreme sensitivity to UV due to defects inefficient repair of UV-related DNA damage (Lehmann, McGibbon et al. 2011). There are 8 XP gene products, all of which are involved in the removal of DNA damage (Bradford, Goldstein et al. 2011). Although, there is variation in the presentation of the XP, including skin lesions, skin pigmentation and melanoma, the underlying causation is the inability to repair DNA damage (Thielmann, Popanda et al. 1991).

1.1.7 Repair pathway conservation

Due to the necessity of maintaining genome stability, DDR pathways are evolutionarily conserved from bacteria to mammals. DDR is a complex signalling pathway that integrates many different proteins that actively contribute to the cellular response. Most of these processes are interdependent, and sometimes proteins are involved are redundant in functions with different specificities, cells must precisely coordinate them for restoration of the genome (Shimada and Nakanishi 2013). Like other canonical signalling pathways, the DDR is composed of sensors, transducers, and effectors these proteins and their functionalities are conserved along the evolutionary timeline in the form of homologues (Friedberg 1995). For example, OGG1 from eukaryotes and MutM glycosylase of the *Escherichia coli* which are involved in BER are functional homologues despite having about 38% sequence similarity (Radicella, Dherin et al. 1997). The other repair pathways also have evolutionary similarities, such as RAD51 and RecA proteins (Friedberg 1995). XRCC4, DNA-PK, Ligase IV etc. have functional homologues as well in other species but not in bacteria (Taylor and Lehmann 1998). This conservation has enabled the understanding of the repair pathway and the difference in conservation highlights functional relevance and redundant approach to ensure that genome stability is guaranteed throughout the cell cycle.

1.1.8 Cell cycle and Cell cycle checkpoints

The cell cycle is the process a cell undergoes to replicate its genome into new daughter cells. There are 4 stages of the cell cycle where the cell prepares to replicate its genome. G1, S, G2 and M are the 4 phases of the cell cycle. G1 and G2 are the gap phases where the cell grows, incorporate the growth signals, and prepares for either DNA synthesis (S-phase) or Mitosis of the replicated genome (M-phase). The cell cycle is controlled by the serine/threonine kinase family of proteins called cyclin-dependent kinases (CDKS). CDKs and their corresponding cyclins work in unison to control the cell through each phase of the cell cycle, starting with Cyclin E and CDK2 moving the cells from G1 into S-phase followed by Cyclin A-CDK1/CDK2 in S-phase and finally Cyclin B-CDK1 in M-phase. While the cell grows and replicates there are concurrent checks to ensure DNA damage is repaired and organelles are duplicated as well as accurate segregation of chromosomes (Killander and Zetterberg 1965, Fantes 1977).

Cell cycle checkpoints are control mechanisms to prevent early entrance into the replicative cycle when the cell is not capable of sustaining itself through the synthesis phase or replication of genomic errors across to daughter cells while giving cells time to decide if the level of accumulated damage is sufficient for apoptosis. There are 3 major cell cycle checkpoints as seen in Figures 1.3. The G1/S checkpoint prevents early progression into the S-phase, this has been extensively studied in yeast, CDC25 and Wee1 are major proteins that respond to cell size and regulate CDC2-Cyclin B (Nurse 1975, Barnum and O'Connell 2014). The G1/S check point is a size control checkpoint and acts in response to genotoxic stress, CDK4/6-Cyclin D and CDK2-Cyclin E act in concert to inhibit the action of retinoblastoma (Rb), DNA damage, replicative senescence and withdrawal of growth factors are indicators that can inhibit the actions of CDK4/6-Cyclin D and CDK2-Cyclin E (Besson, Dowdy et al. 2008). CHK1 pathway is an evolutionarily conserved pathway in the S/G2 checkpoint (Demidova, Aau et al. 2009, Tapia-Alveal, Calonge et al. 2009). ATR-CHK1 degradation of CDC25A ensures the cell cycle arrest as checkpoint proteins assemble on RPA-coated damaged DNA, co-currently ATM/ATR/DNA-PK are involved in the activation of p53 and protected from E3 ubiquitin ligase MDM2 activates a plethora of genes including cyclin-dependent kinase inhibitor CK1 ensuring cell cycle arrest for DDR (Demidova, Aau et al. 2009, Barnum and O'Connell 2014).

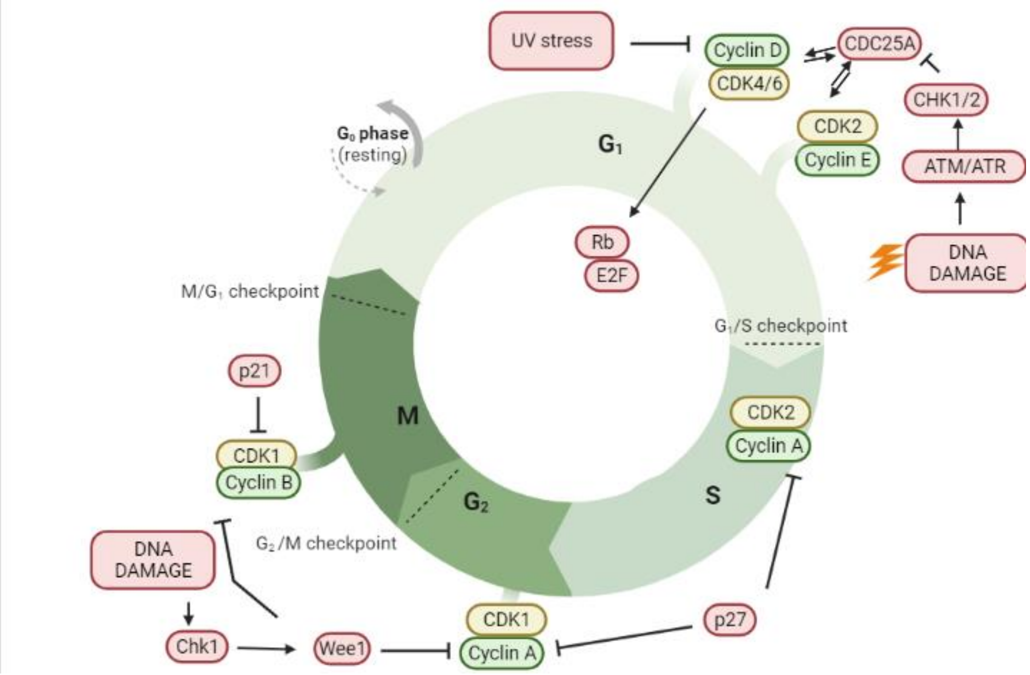


Figure 1. 3 **Schematic of cell cycle**.

Progression of the cell across the different phases of the cell cycle. The 3 cell cycle checkpoints. G1/S checkpoint is required for size control to ensure the cell does not progress into the synthesis phase earlier than required. G2/M checkpoint checks proper genome replication and no damage induced errors are present. M/G1 checkpoint is to ensure appropriate chromosome assembly and segregation. G1, S-phase and G2 are collectively known as Interphase. Illustration adapted on Biorender.

Failure to resolve DNA damage can lead to programmed cell death or senescence as the cell does not progress into M-Phase. Mitosis or M-phase is when fully replicated cellular aggregates including the replicated sister chromatids are polarised and separated into new daughter cells, this would be referred to as meiosis in germ-line cells. M-phase is sub-divided into 5 stages: Prophase, Prometaphase, Metaphase, Anaphase and Telophase or Cytokinesis. Prophase begins with the breakdown of the nuclear membrane, condensation of chromosomes and separation of the centrosome to form the pole where the organelles and chromatids would migrate to (Metz 1925). During prometaphase mitotic spindle formation occurs to allow for migration of the chromosome. Mitotic spindles extend from the centrosomes to the kinetochore. The kinetochore is made up of 80 different proteins constructed at the centromere of each sister chromatid. Cohesin, a ring-shaped protein that holds the replicated chromatids together (Nasmyth and Haering 2009), NDC80 and CENP-A constitute the inner and outer kinetochore plates respectively (Musacchio and Desai 2017). CENP-A, an H3 nucleosome variant recruits CCAN, constitutive centromere associated network a complex of 16 different proteins called the KMN network which consists of the KNL1, Mis12 and Ndc80 sub-complexes (Allshire and Karpen 2008). Ndc80 sub-complex is made up of Ndc80/Hec1, Nuf2, Spc24 and Spc25, they are tasked with the load-bearing attachments of the microtubule to the kinetochore and Ndc80 is responsible for recruiting the RZZ complex consisting of Zw10, Rod and Zwilch which in turn recruit dynein/dynactin minus-end motor complex and the Mad proteins (Mad1 and Mad2) (Chen 2002, Liu, Hittle et al. 2003, Karess 2005). Aside from enabling microtubule attachment, the kinetochore controls the stability of microtubule attachment favouring an orientation to allow for proper segregation of the sister chromatids (Desai, Guha et al. 2009, Musacchio and Desai 2017). These proteins are also well conserved due to the nature of their function. For example, in *S. cerevisiae*, Cse4 Nucleosome is homologous to CENP-A in humans and the homologs of the CCAN complex, which in humans is made up in of the CENP proteins family corresponds to the various proteins for which are Mif2, MCM family sub-complexes Ctf19, MFH family Cnn1 and Wip1 (Lechner 1994, Lechner and Ortiz 1996, Hyland, Kingsbury et al. 1999, Ghosh, Poddar et al. 2001, Musacchio and Desai 2017). Furthermore, the NDC80 subcomplex is also well conserved in both *S. cerevisiae* and humans and the Knl1/Zwint is homologous to Spc105/Ydr532 (Doheny, Sorger et al. 1993, Desai, Guha et al. 2009, Musacchio and Desai 2017).

Regulation of kinetochore assembly is carried out by competitors of contrasting functions Aurora B and Protein phosphatase 1 (PP1). Aurora kinase is a family of Serine/Threonine kinases, of which there are 3 of them with different subcellular localization in mammals. Although all 3 play different roles in chromosome segregation, they act at different mitotic times and different cellular regions. Aurora-A is mostly active at the centrosome during interphase and Aurora-C acts chiefly during the telophase

(Ke, Dou et al. 2003, Meraldi, Honda et al. 2004, Venoux, Basbous et al. 2008). Proteins such as Cdc25B, TPX2, Eg5, Lats2, TACC, Ajuba and BRCA1 etc are substrates phosphorylated by Aurora-A, enabling spindle formation and stabilizing proteins that commit the cell to mitosis (Meraldi, Honda et al. 2004, Venoux, Basbous et al. 2008). Aurora-B acts at the kinetochore during prophase correcting erroneous spindle fibre attachment. When monothelic, syntelic or meterotelic attachments occur phosphorylation of certain proteins by Aurora-B causes destabilization of kinetochore/microtubule attachments leaving the unattached kinetochore to form proper attachments (Ditchfield, Johnson et al. 2003). Monothelic attachments happen when one kinetochore is attached to the spindle fibre microtubules, but the other is not. This tends to occur early in mitosis and is known as normal intermediate, synthelic attachment, on the other hand, happens when both kinetochores of the same chromosome bind to one microtubule on the same spindle pole (Ault and Rieder 1994, Hauf 2003). Both monothelic and synthelic attachments take time to resolve by Aurora-B and trigger the spindle assembly checkpoint (SAC) delaying mitotic progression. Correction of both attachment errors involves Aurora-B forming Chromosome Passenger complex (CPC) with INCENP, survivin and borealin, phosphorylating Ndc80- which inhibits Ndc80 binding to microtubules and release of the kinetochore (Ault and Rieder 1994, Hauf 2003). When a kinetochore is attached to microtubules from both poles merotelic attachment is said to have occurred, it is also resolved by the action of Aurora-B, but it does not trigger SAC and is usually resolved during the onset of anaphase (Cimini, Fioravanti et al. 2002). Prolongation of SAC can lead to M-phase cell cycle arrest (Figure 1.4), this is particularly important in the context of cancer and progression as any of the erroneous attachments could lead to segregation aberration (unequal distribution of chromosomes) and either apoptosis or cancer progression. Over-expression of the Aurora kinases have been reported to cancer-stasis (Gregan, Polakova et al. 2011).

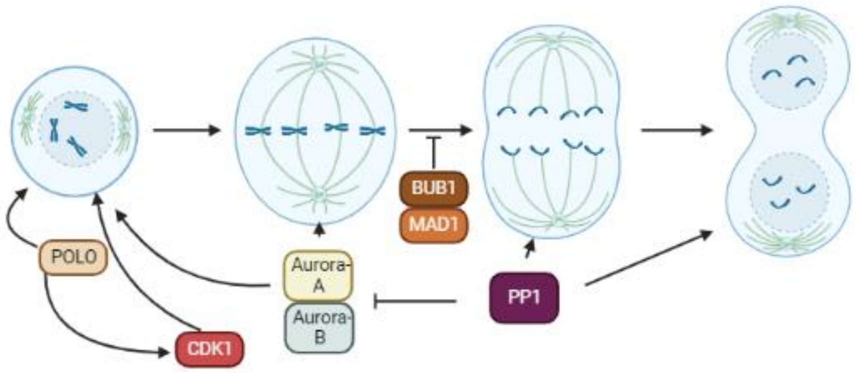


Figure 1. 4 **Schematic of the function of the Aurora kinases in chromosome segregation**.

Interaction of the Aurora kinases, Protein phosphatase 1 and Polo-kinase and how they regulate chromosome segregation through spindle fibre attachments. Illustration created in BioRender.

Amphitelic attachment is the correct and accurate attachment of the sister chromatids to the opposing spindle microtubules (Figure 1.5), this allows the sister chromatids to be segregated properly. When this occurs protein phosphatase 1 (PP1) counteracts the actions of Aurora-B kinase by dephosphorylating Aurora-B substrates and destabilizes the kinetochore machinery at the outer core (Francisco, Wang et al. 1994, Emanuele, Lan et al. 2008, Vanoosthuyse and Hardwick 2009). Although

PP1 is functionally diverse, it is recruited to the kinetochore by Knl1 which is an Aurora-b substrate.

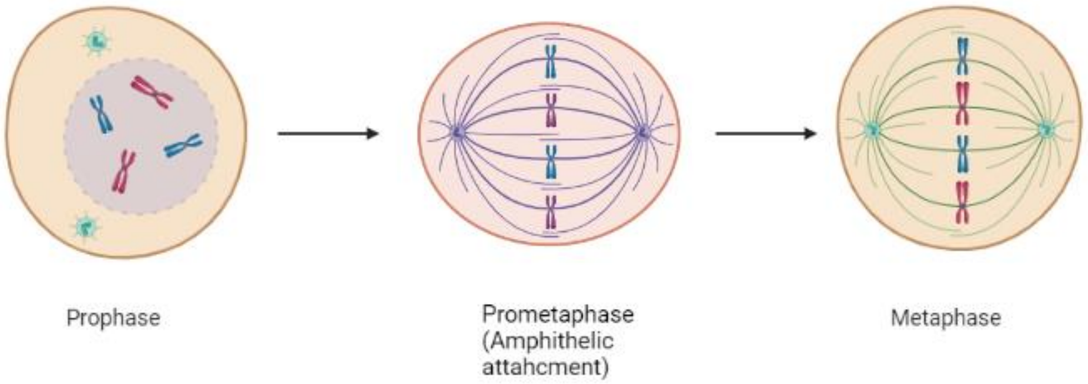


Figure 1. 5 **Amphitelic attachment of the kinetochore during prometaphase**.

Accurate kinetochore-microtubule attachment at the cellular equator leads to mitotic progression to metaphase. Illustration created in Biorender.

Aurora-B and PP1 are highly conserved proteins that work in tandem to ensure proper chromosomal segregation and progression of the cycle and alongside an array of other proteins.

1.1.9 DNA Damage Response and Repair

Depending on the damage induced mammalian cells have developed several repair pathways to efficiently repair the damage induced. However, some pathways are more efficient than others in repairing the damage. Understanding DDR is best understood through DNA-damage checkpoints and response to different types of damage. There are a few response pathways and damage that are resolved which have been summarised in Table 1.4.

Table 1. 1 **Summary of DNA repair mechanisms**.

Types of DNA repair pathways and accuracy of the pathways.

Repair pathway	Lesions
Base excision repair	Corrects DNA damage from oxidation, deamination and alkylation, also single-strand breaks
Nucleotide excision repair	oxidative endogenous lesions such as cyclo-purine, sunlight-induced thymine dimers (cyclo-butane dimers and pyrimidine (6-4) pyrimidine photoproducts and other bulky lesions)
Homology-directed repair	double-strand breaks in the mid-S phase or mid-G2 phase of the cell cycle
Non-homologous end-joining	Double strand breaks if cells are in the G0 phase. the G1 phase or the G2 phase of the cell cycle
DNA mismatch repair	base substitution mismatches and insertion-deletion mismatches generated during DNA replication
Translesion synthesis	DNA damage tolerance process that allows the DNA replication machinery to replicate past DNA lesions
Fanconi anaemia repair	Corrects damage from either intrastrand or inter-strand crosslinks (ICLs) primarily in the S-phase.

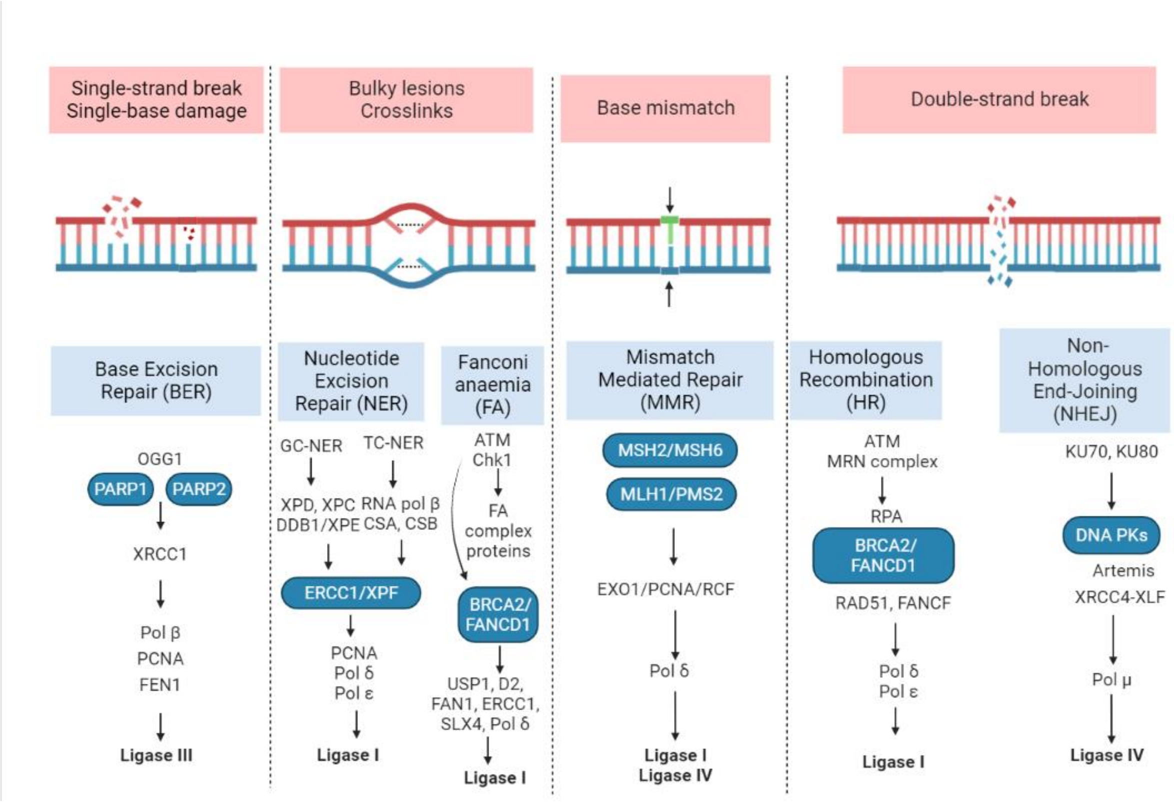


Figure 1. 6 **DNA damage and corresponding repair pathways**.

Summary of the types of DNA damage, the possible source of the damage and the repair response pathway used to resolve the damage. Illustration created in Biorender.

Double strand break repair

DSB is carried out in three ways: non-homologous end joining (NHEJ), conservative homologous recombination (HR) and single-strand alignment, also called non-conservative homologous recombination (SSA) (Langerak and Russell 2011). HR is an error-free repair pathway as it uses undamaged DNA as a template for the repair. However, HR is a slow process and occurs mostly in the S/G2 phases of the cell cycle and if the template is already erroneous based on previous repair errors this is passed on to the new strand. When HR is initiated, MRE11/RAD50/NBS1 (MRN) complex is recruited to the DSB sites, followed by ataxia–telangiectasia mutated (ATM) a member of the family of phosphoinositide-3-kinase-related kinases (PIKKs) (Pan, Penney et al. 2014). This is followed by RAD50-NBS1, ATM-dependent phosphorylation, which in turn phosphorylates the BRCA genes, EXO1 and RPA and then Pol δ and LIG1. An important protein in the cascading event leads to the resolution of DSB through the HR repair pathway is RAD51. Rad51 replaces the replication protein a (RPA) on the single strand overhangs that have been treated by the MRN complex, this enables strand invasion to initiate HR repair (Baumann and West 1998). RAD51 promotes replication restart through break induced replication (BIR) at broken replication forks and plays a role in non-repair functions at stalled replication forks by promoting fork reversal and stabilizing stalled forks by protecting nascent DNA from nucleolytic degradation (Morrison, Shinohara et al. 1999, Petermann, Orta et al. 2010, Hashimoto, Puddu et al. 2011). Upon successful invasion of the RAD51 coated DNA strand, Polymerase activity is carried out by Pol δ or Pol ε and then ligated by ligase 1.

The other two repair pathways, NHEJ and SSA are more error-prone because processing the ends of the DNA during repair can lead to loss or modification of information. NHEJ is the commonest repair pathway for DSB in eukaryotes as it is also the quickest. When NHEJ is initiated, DNA-dependent protein kinase (DNA–PK) and KU70/80 are first recruited to the site of the damage. This is then followed by phosphorylation of WRN, TDP" and Artemis followed by Pol μ and then XRCC4, LIG4 and XLF (Pan, Penney et al. 2014). ATM and ATR are essential for the G1/S, intra-S-phase, and G2/M DNA-damage checkpoints, and are critical for the maintenance of genomic integrity. Mutation in either HR or NHEJ can lead to serious congenital syndromes such as human ataxia-telangiectasia (AT), an autosomal recessive disorder characterized by cerebellar ataxia, progressive mental retardation, impaired immune functions, neurological problems, and malignancies other possible syndromes are ATR-Seckel and LIG4 syndrome (O'Driscoll and Jeggo 2006).

There are also other forms of DNA repair pathways (Figure 1.3), base excision repair (BER), nucleotide excision repair (NER), and mismatch mediated repair (MMR). BER is used to correct base lesions caused by ROS, IR, or chemotherapeutics (alkylating agents) that alter the DNA helix. OGG1, the DNA glycosylase responsible for the excision of 8-oxo-guanine, functions by cleaving the N-glycosyl bond between the sugar and the base to form an abasic site. This is followed by the action of PARP1, PARP2 and XRCC1 to ensure the excision site is suitable for repair and Pol β with its AP lyase activity alongside PCNA synthesize with a short or a long stretch of sequence and then the final step is the action of DNA ligase to ligate the nick in the sequence (Seeberg, Eide et al. 1995, Maynard, Schurman et al. 2009).

1.2 Protein Conservation

The conservation of gene order between different organisms is an informative property of genomes and is currently being used to study the functional relationships between genes and predict the functions of novel genes (Tamames 2001). Orthologous genes between species are exposed to the evolutionary pressures of the species and so essential genes in some organisms can become not so essential in other organisms (Bergmiller, Ackermann et al. 2012). Often genes sequences are either lost completely due to faulty replication mechanism, are duplicated or have sequence mutation and so have their functions lost or transformed in their sister taxa (Bergmiller, Ackermann et al. 2012).

Gene essentiality is currently understood to be both context-dependent and evolvable in all phylogenies (Rancati, Moffat et al. 2018). This understanding of non-absolute gene essentiality drastically changes our approach towards biological processes and gene conservation among the evolutionary phylogenies (Rancati, Moffat et al. 2018). The inventory for housekeeping genes and understanding the differences in the genetic basis of their functionality in the different phylogenetic lineages is critical to understanding life at the level of a single cell (Tatusov, Koonin et al. 1997). Complete sequences are crucial to achieving this goal as they hold the information required to describe the complex relationships between genes from different genomes (Tatusov, Koonin et al. 1997). In addition, complete genomes sequences make it possible to establish proteins implicated in essential functions in some genomes and not in other genomes (Tatusov, Koonin et al. 1997). Nonetheless, there are instances of well-preserved clusters of genes even in divergent species, the best examples are the ribosomal proteins (Nikolaichik and Donachie 2000). There are a few reasons as to the conservation of gene order which includes but are not limited to; firstly, the species have diverged only recently and the gene order is not yet destroyed; secondly, there has been lateral gene transfer of a block of genes, or the integrity of the cluster is important for the fitness of the cell (Tamames 2001). The latter seems to be the case especially when the gene order traverses several species across the phylogenetic tree.

In bacteria, there have been extensive studies into the conservation of gene families which has led to enhanced understanding of the conservation of genes in multicellular organisms. The conservation of such genes across the species gives insight as to the evolution of each phyla group. Genes such as the 16S and 23S ribosomal RNAs (rRNA) have given insights into the relationship of the bacteria phyla *Planctomycetes*, *Chlamydiae*, *Lentisphaerae*, and *Verrucomicrobia* (Pilhofer, Rappl et al. 2008). Genes such as 16S and 23S ribosomal RNAs also have homologues in multicellular organisms. These genes with importance for cellular integrity and fitness are often well conserved among all phylogenetic classes. Another example is the three major eukaryotic cytoskeletal families, actin, tubulin, and intermediate filaments, which are represented across the phylogenetic landscape and in the bacterial phyla as FtsA/MreB/ParM, FtsZ/BtubAB (Gitai 2007, Pilhofer, Rappl et al. 2008). MutT gene class an E. coli protein that is part of the GO system is an error avoidance pathway devoted to enhancing the fidelity of DNA replication and is another example of genes required for cell integrity being conserved across species (Michaels and Miller 1992, Koonin 1993). With gene conservation comes gene order conservation and the question, why are certain neighbouring genes more difficult to separate compared to others during evolution? Recombination events that result in the shuffling of genes in genomes during evolution are more apparent in eukaryotes as compared to bacteria (Davila Lopez, Martinez Guerra et al. 2010). Bacterial genes are organised in operons and as such, shuffling of genes is restricted (Davila Lopez, Martinez Guerra et al. 2010). While eukaryotic genes are not subject to this restriction their gene order is not completely random (Eichler and Sankoff 2003, Michalak 2008, Davila Lopez, Martinez Guerra et al. 2010). Hence, comparison of two eukaryotic genomes that are only distantly related, there is a likelihood that two genes are in the same order between the two species (Davila Lopez, Martinez Guerra et al. 2010).

Studies have investigated the evolution of gene order and recombination in eukaryotic chromosomes and how this affects gene expression. Chromosomes evolve through modification, acquisition, deletion, and re-arrangement of genetic material (Eichler and Sankoff 2003). Understanding chromosomal evolution involves approaching genome evolution from two perspectives; firstly, comparing the number of chromosomes and the order of homologous segments among closely and distantly related species (O'Brien, Menotti-Raymond et al. 1999). The other approach involves analysis of small blocks of DNA sequences, through comparative sequencing among closely related species (Eichler and Sankoff 2003). Homologous recombination could lead to shuffling of genes across chromosomes; however, genes of similar expression tend to cluster more commonly than expected by chance and there is evidence that functionally related genes also tend to cluster together (Pal and Hurst 2003, Davila Lopez, Martinez Guerra et al. 2010). The emergence of large-scale sequencing of eukaryotic genomes, the ability to comparatively analyse complete genomes gives a comprehensive view of physical co-localization of genetic loci on the same chromosome, gene order and regions of non-conservation (Eichler and Sankoff 2003). The biggest advantage of large-scale gene sequencing is the ability to generate gene maps for species whose genomes have been completely sequenced. Mammals tend to contain between 10,000 and 20,000 genes arranged in a linear order in their chromosomes, with chromosome numbers ranging from a minimum of 3 pairs (2N = 6; Indian muntjac, *Muntiacus muntjak*) to a maximum of 67 pairs (2N = 134; Black rhinoceros, *Diceros bicornis*) (O'Brien, Menotti-Raymond et al. 1999). While the level of quality of the gene map would be directly proportional to the sequence quality the most effective gene maps contain and integrate three categories of markers (O'Brien and Graves 1990, O'Brien, Menotti-Raymond et al. 1999). Coding genes through which DNA sequence comparison and comparative mapping are essential for identification of gene orthologs in distantly related species are referred to as Type 1 markers (Wang, Fan et al. 1998); short tandem repeats, hypervariable microsatellites (STRs) that are informative in pedigree, forensic and population assessments as there are approximately 100,000 near-randomly dispersed throughout the mammalian genome are Type 2 markers and less useful for locus recognition between species compared to type 1 markers and because each carries multiple alleles (Cargill, Altshuler et al. 1999). Type 3 markers are bi-allelic single nucleotide polymorphisms (SNPs) within coding regions or non-coding regions such as introns (Lipshutz, Fodor et al. 1999). SNPs like STRs are valuable for pedigree or family screens within species, especially genotyping, automated array-based analysis (O'Brien, Menotti-Raymond et al. 1999).

Understanding the evolution of multicellular organisms involves recognising the core conserved features of “house-keeping genes” in relation to their developmental programs and the genetic modifications that influence their phenotypic features (Schilde, Lawal et al. 2016). The information encoded in the genome has the potential to manifest the physical form of the organism and help adapt to the changing environment in which the cell occupies. The house-keeping genes are expressed constitutively and are necessary for the shape and basic physiology of the organism, while other genes are expressed when required (Dekel, Mangan et al. 2005). Housekeeping genes and other essential genes are imperative for the maintenance of organisms as they are central in many critical cellular functions (Juhas, Eberl et al. 2011, Luo, Gao et al. 2015). The natural selection that acts on essential genes is expected to be stricter compared to other nonessential genes (Luo, Gao et al. 2015). The evolutionary strictness could be applied to functionally related genes but due to genetic recombination events are not deemed as essential. The average eukaryotic gene is thought to be randomly distributed within the genome and independently expressed of their neighbours through promoter sequences and sequence-specific transcription factors, with notable exceptions of gene clusters such as the Hox and β-globin (Michalak 2008). However, there is growing evidence that genes that are functionally related or genes proximally located to each other tend to show certain levels of co-expression (Cho, Campbell et al. 1998, Michalak 2008). Cumulatively, genes of similar expression tend to cluster more commonly than expected by chance, in addition, functionally related genes tend to cluster including genes involved in stable complexes and the same metabolic pathways (Boutanaev, Kalmykova et al. 2002, Lee and Sonnhammer 2003, Davila Lopez, Martinez Guerra et al. 2010). A strong factor associated with non-random gene order is intergenic distance, which in yeast is a strong indicator of gene order conservation (Poyatos and Hurst 2007). Mammalian gene pairs tend to have short intergenic distances, where genes are divergently transcribed (Adachi and Lieber 2002, Trinklein, Aldred et al. 2004). In addition to intergenic regions, eukaryotic genomes possess bidirectional divergent organization (head-to-head), an additional level of complexity allowing genes to be encoded by the same DNA sequences and be acted upon by different promoters and transcription factors (Adachi and Lieber 2002, Yang and Elnitski 2008)

1.3. Protein structure

DNA is transcribed to mRNA and in turn translated into amino acids, which folds into functionally active tertiary structures. This flow of information is the central dogma of cellular biology. The amino acid sequence determines the possible folding pattern and stability of the quaternary structure of a protein. The side chain of the amino acid is the most important part of the amino acid as the chemical properties are important in the folding pattern of the transcribed protein sequences. The folding pattern of amino acids usually follows the thermodynamic favourability, in turn, determining the structural class of a protein. Proteins can be classified into four structural classes; all-a class, which are helices and a small amounts of strands, all-b class, which are strands and only a small number of helices, a/b class, which includes both helices and mostly parallel strands, and a + b class, which includes both helices and mostly antiparallel strands.

Protein structure has usually been determined experimentally, a slow arduous process after the biological function has been well characterized, and the structure used to support its function. The improvement of sequencing technology has increased the access to protein sequences and a better understanding of protein conservation and domain structures have allowed for functional studies of proteins alongside structure prediction while waiting for better structure resolution technologies (Ingolfsson and Yona 2008, Loewenstein, Raimondo et al. 2009).

1.3.1 Determination of Protein 3D structure

Understanding the three-dimensional conformation of amino acids is an ongoing subject of inquiry and different explanations of the underlying mechanism have been proposed, especially the Gibbs free energy for the protein/water system. Peptide bonds are formed through a reaction of the carboxyl group in one amino acid and the amino group of another amino acid and the release of water. This coupling of the amino acids starts the polypeptide from the N-terminus and carries on to the C-terminus (Pauling and Corey 1951). Carbonyl groups in the peptide chain form stable hydrogen bonds with amino groups creating regular and stable secondary structures called α-helices or β-strands (Pauling and Corey 1951). These secondary structures are usually associated with domains that can have their sequences, evolutionarily conserved as unique families of amino acids (Andrade, Petosa et al. 2001). Hence, the sequence of the amino acids in the primary structure plays a role in the tertiary structure (Anfinsen 1961, Anfinsen 1973). In addition, one of the major propelling forces behind protein folding is the hydrophobic effect, which makes the hydrophobic amino acids cluster together in the protein interior, while polar and charged amino acids are on the surface to interact with the surrounding water molecules (Kauzmann 1959). There are two main methods of resolving protein structure experimentally and computational modelling.

1.3.2 Experimental determination of protein 3D structure

The resolution of a protein structure experimentally can be done in 3 ways: X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and electron microscopy (EM). When ascertaining a protein structure via X-ray crystallography, the protein is purified and then crystallized, and the crystal is subjected to an intense synchrotron X-ray beam. The electrons in the protein scatter the X-rays in a specific pattern, which is then used to calculate an electron density map where the amino acids in the protein are already defined to give a structure (McPherson 2004). In NMR spectroscopy, the solution of the purified protein is placed in a strong magnetic field and then probed with radio waves (Marion 2013). The resonance can be observed in a spectrum, and analysis of the bond conformation is enabled through nuclei atom proximity. There are conformational restraints as well, which are then used to build the structure of the protein. Alternatively, EM is used for large macromolecular complexes, which are subjected to beams of electrons to obtain a 3D image (Kuhlbrandt 2013). EM has often been used in combination with X-ray crystallography or NMR spectroscopy to obtain the atomic details of the complexes (Callaway 2015). EM has been reformed to produce high-resolution models quickly, that X-ray crystallography and other approaches have not been able to solve. X-ray crystallography also relies on obtaining a protein crystal, while NMR spectroscopy is limited to low molecular weight proteins. In addition, experimental methods are tedious, time-consuming, and expensive. Therefore, computational modelling of proteins structures has become more frequently used.

1.3.3 Computational modelling of Protein 3D structure

Prediction of 3D protein structures is a relatively novel field of Bioinformatics, and this is because of the better understanding of domain conservation and protein families (Altman and Dugan 2003, Zhang, Arakaki et al. 2005). Protein structure prediction involves transferring sequence information into a 3D structure and prediction functionality of the structure (Chou and Fasman 1974, Creighton 1990). Various algorithms and methods have been established to determine how much sequence information determines the function of the final protein structure. Homology modelling (also called comparative modelling), fold recognition methods, and first principal prediction with and without database information are now used to determine the function of novel genes of clinical relevance (Hunter 2006, Xiang 2006). Prediction of structure without the use of database information is known as *ab initio* modelling, this method uses only the Gibbs free energy or minimum global free energy of the amino acids sequence in determining the possible structure of the protein (Osguthorpe 2000, Bonneau and Baker 2001). The major limitation to *ab initio* modelling is the possible conformations of the peptide sequence, which is restricted to the Gibbs free energy state (Dorn, MB et al. 2014). However, this is not always representative of the protein. An advantage of this method is the prediction of new protein folds without existing templates (Bonneau and Baker 2001, Dorn, MB et al. 2014). There is another method of template-free modelling which uses sub-sequence information from the target protein to scan databases for the folding of similar fragments, the folding orientation of these fragments are then stacked and scored to the lowest energy state (Xiang 2006). There are many challenges to this method like the conformations such fragments can take, and the scoring functions required to determine acceptable conformations. In addition, combining different fragments to develop the whole structure can be difficult but this method is more advantageous compared to ab initio modelling as the fragment conformational search helps narrow the probabilities of structures to be determined (Xiang 2006, Dorn, MB et al. 2014).

1.3.4 Template based methods

The use of existing protein structures (templates) to resolve structures of novels sequences is a widely used approach in structural bioinformatics. Templates are more useful in structure prediction as they are more accurate and can be used for proteins with longer sequences. Homology modelling is one of the commonest template-based modelling. It relies on the evolutionary relationship between proteins to ascertain structural relationships (Mullins 2012). Due to the convergent and divergent nature of evolution, protein relationships may not be as straightforward as being descendent from the same ancestral protein. Homologs can also be divided into orthologs and paralogs. With orthologs, proteins have evolved independently in different species but have the same function. Paralogs, on the other hand, are found in one species, but the proteins have differentiated into having different functions. Homology modelling assumes that similar protein sequences, i.e., homologous proteins, fold into a similar 3D structure. Several steps are required for template-based protein modelling to enable accurate prediction; sequence analysis, structure association and prediction of primary and secondary structure are required.

Sequence and domain searching

There are several databases where protein sequences are stored and annotated. These can be searched either manually or using a script to find homologous sequences. The Basic Local Alignment Search Tool (BLAST) (Altschul, Gish et al. 1990) at NCBI enables searching of a nucleotide (BLASTN) or protein sequence (BLASTP) (Altschul, Gish et al. 1990) against a selected database in search of local similarity, which is then reported as a statistical significance. There are different variations of BLAST, such as the position-specific iterated BLAST (PSI-BLAST) (Altschul, Madden et al. 1997) and context-specific iterated BLAST (CSI-BLAST) (Biegert and Soding 2009). PSI-BLAST uses position-specific scoring matrices (PSSM), which represent multiple sequence alignments with numbers so that each number indicates the probability of a certain amino acid at every position, particularly useful in domain prediction. Use of a target sequence is required for a basic BLAST search before additional search iterations can be carried out. Statistically significant results can be aligned together to generate a PSSM to create a sequence profile that can be used to find distantly related protein sequences. Aside from BLAST, species-specific databases are available, they contain sequences that are present and have been annotated to specific species. These can be accessed to reduce the time of determining paralogs, however, BLAST is still the most advantageous in determining orthologues and domain attributes of proteins.

Prediction of Primary and Secondary Structures

Generation of a PSSM enables prediction of the domain composition of the target sequence and sequence pattern. Domains are compact, semi-independent functional sequence units, which can be evolutionarily conserved depending on function. When domains are not obvious within the target sequence, sequence threading is employed to determine the structure. Threading is a fold recognition-based method on the conclusion that structure is more conserved than sequence and, therefore, two proteins can have the same fold although there is no apparent sequence similarity and evolutionary relationship between them (Levitt and Chothia 1976, Finkelstein and Ptitsyn 1987, Xiang 2006). Additionally, sequence profiles are used to describe and detect larger areas or domains of the sequence, including variable regions (Gribskov, McLachlan et al. 1987). The threading process works by linking the target protein sequence sequentially onto the known 3D structure in an optimal way and, through this, identifying homologous (evolutionary related) or analogous (no direct evolutionary relationship) templates (Dorn, MB et al. 2014). The energy of the target sequence in a certain 3D fold assesses the quality and is used to estimate the likelihood of the query sequence to adopt this fold.

Some databases are dedicated to domain architecture and prediction and can be used to query target sequences. Simple Modular Architecture Research Tool (SMART) (Schultz, Milpetz et al. 1998, Letunic and Bork 2019) and Pfam (Finn, Attwood et al. 2017) are protein domain repositories that hold valuable information relating to structure prediction and the functional association of uncharacterized proteins. There are also online servers used for the generation and prediction of protein structures. These are databases of consensus repeats crucial for the structure or function of domains or protein families (Mulder and Apweiler 2002, Wu, Huang et al. 2003). PROSITE is a major server used for detecting sequence patterns (de Castro, Sigrist et al. 2006). Early methods for secondary structure prediction were based on the probability of a certain amino acid is in a specific secondary structure. Leucine, Isoleucine, and Valine are amino acids usually found in β-strands probably because they are non-polar amino acids (Chou and Fasman 1974). In addition to PROSITE, several other servers such as I-TASSER and PSIPRED are used to predict secondary structures and a general theory that proteins with approximately 30 % or more sequence identity have a similar fold (Chou and Fasman 1974, Pavlopoulou and Michalopoulos 2011, Yang and Zhang 2015).

1.4 Structure analysis

Prediction of protein structure either through the template-free or template-based methods still required further analysis to determine the accuracy of the models. Special attention should be paid to the target template sequence identity, and the higher the overall sequence identity is, the effective it will be, when used with the sequence of interest (Tramontano and Morea 2003). The use of templates in structure prediction can be complicated by the presence or absence of cofactors and ligands, also oligomerization and conformation state can affect structure prediction and analysis (Kopp and Schwede 2004, Kopp, Bordoli et al. 2007). Analysis of protein structures often involve comparisons to one another to emphasize the similarities and differences. Structural comparison can also help infer evolutionary relationships even when the proteins have less than 25 % sequence identity (Katoh and Standley 2013). It can also be used for the classification of proteins and their domains into families (Murzin, Brenner et al. 1995, Orengo, Michie et al. 1997). Protein structure comparison is done by the superimposition of two or more structures and, during the process; one of the molecules is rotated and oriented to fit on top of the other molecule (Maiti, Van Domselaar et al. 2004). During 4superimposition of protein structures for comparison, the question of interest is the level of similarity between the structures, and if it is at a local level around the ligand-binding site or if it is the global fold. The commonly used measure for this is the Root Mean Square Deviation (RMSD), which is calculated by adding together the square of the difference in distance (Ångström [Å]) between equal Cα-atom pairs and dividing the sum by the number of compared atoms. Hence, the lower the RMSD, the more similar are the compared structures.

1.5 C1ORF112.

The advances in the technologies to further the understanding of mammalian genomes has also furthered the understanding of mechanisms that underlie human diseases and basic cellular functions. The growth in genomics and proteomics has enabled rapid generation and analysis of large amounts of genomic and proteomic data to identify novel genes that could be avenues to further understanding cellular networks and pathways and identify novel therapeutic candidates for genetic disorders and age-related ailments such as Alzheimer's, Parkinson's and cancer. Various bioinformatic approaches have been used to identify candidate genes for study, to understand cellular mechanics and one of such approaches is the use of guilty by association method used by van Dam et al. (2012) to identify the mouse gene *BC055324* gene (van Dam, Cordeiro et al. 2012). *C1ORF112* is the human homologue of *BC055324*, C1ORF112 is strongly co-expressed with genes such as RAD51, CCDC6 as well as many genes in the BRCA-Fanconi anaemia (FA) DNA damage response pathway, including BRCA1, BRCA2, FANCD2 and FANCI (van Dam, Cordeiro et al. 2012).

The genes strongly co-expressed with C1ORF112 such as the breast cancer susceptibility proteins BRCA1 (also known as FANCS), BRCA2 (FANCD1) and RAD51 (FANCR) and its paralogs including the XRCC-2 (FANCU), and XRCC3 all function in homologous recombination repair (HRR). HRR is a critical DNA repair process that operates on directly occurring DNA double-strand breaks, but also in the repair of broken and stalled DNA replication forks (Baumann and West 1998;Pan, Penney et al. 2014). FANCD2 and FANCI are critical proteins that are mono-ubiquitylated in the activation of the Fanconi anaemia (FA) pathway required for the repair of inter-strand crosslinks (ICLs) (Sato, Toda et al. 2012). FANCD2 and FANCI are at the core of the FA pathway and are mono-ubiquitylated by the upstream FA core complex (comprising nine FA or FA-associated proteins). Downstream of the monoubiquitylation of FANCD2 and FANCI are the downstream FA proteins that function directly in DNA repair including HRR. ICLs are a specific type of DNA damage that block transcription and DNA replication and require removal by several DNA repair processes including translesion DNA synthesis and HRR that are co-ordinated by the FA pathway (Figure 1.7). Whilst repair by HRR is largely error-free when it, or the FA pathway, is defective DNA double-strand breaks and broken replication forks may be erroneously repaired by non-homologous end joining (NHEJ) (Lesport, Ferster et al. 2018).

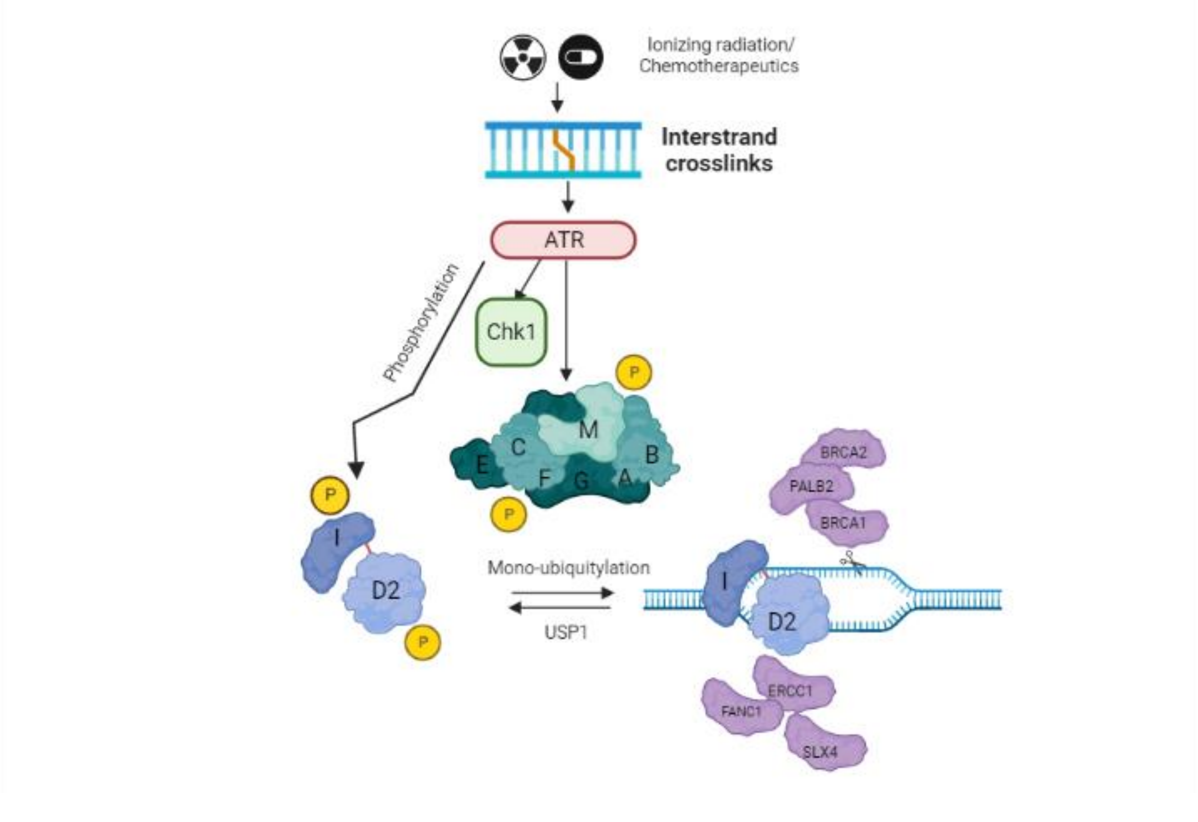


Figure 1. 7 **The Fanconi-Anaemia (FA) pathway**.

The interaction between the genes co-expressed with C1ORF112 and the role they play in DNA damage repair in the FA pathway

C1ORF112 has been highlighted in studies involved in HR DNA damage response processes (Fernandes, Duhamel et al. 2018). There are earlier studies which have also reported the presence of C1ORF112 in cancer tissue (Leo, Wang et al. 2005, Sanchez-Carbayo, Socci et al. 2007). For example, a study of bladder cancer progression, the genomic and proteomic profiles in the association of TP53 showed that *C1ORF112* has a fold change corresponding to an increased expression with tumours having mutant *TP53*, a tumour suppressor gene was, whose mutation is involved in driving various cancer (Sanchez-Carbayo, Socci et al. 2007). Down-regulation of *C1ORF112* expression in response to regulation by progesterone hormone-independent breast cancer cells transfected with progesterone receptor might indicate *C1ORF112* might be a target for progesterone regulation (Leo, Wang et al. 2005).

Chromosome 1 Open Reading Frame 112 (C1ORF112) codes for 9 transcripts. The chromosomal locus for C1ORF112 is 1q24.2. It has also been identified as FLJ10706 (HGNC) or ENSG00000000460 (Ensemble). Of the 9 C1ORF112 transcripts, 5 are translated into proteins, while 4 of them undergo nonsense-mediated decay. The first two transcripts are 853 amino acids in length and are 4355 bps with 24 exons and 4011 bps with 25 exons respectively (Zerbino, Achuthan et al. 2018). There are no domain motifs currently attributed to this protein as it is currently classed under the domain of unknown function DUF4487 (Finn, Attwood et al. 2017). Proteins in this domain family have a conserved WCF tripeptide sequence which may be functionally relevant (Finn, Attwood et al. 2017).

1.6 Aims

C1ORF112 is of interest because it has been found to be co-expressed with genes such as BRCA1, BRCA2, FANCD2 and FANCI. Since there are no current ascribed functions to C1ORF112 and there is gap in the knowledge about the protein itself. The aim of this thesis would be to

- To understand the evolution and conservation of C1ORF112 in species across the eukaryotic phylum and determine how C1ORF112 is level of expression in human tissues using data from the Gtex database. This thesis will also look at the genes co-expressed and associated with C1ORF112, and possible physical interactor with C1ORF112
- This thesis will aim to determine the structure of C1ORF112 and possible functional domain and post-translation modifications.
- This thesis will finally look to characterise phenotypic properties of the C1ORF112 knockdown cells lines. Looking at the differential gene expression between the knockout cell and the normal cells and to determine if the knockout cells are more sensitive to agents of DNA damage, especially x-ray radiation and hydrogen peroxide.

Chapter 2: Methods

2.1 Bioinformatic analysis

2.1.1 Sequence analysis

The sequence of C1ORF112 was obtained from UniProt ID Q9NSG2 isoform Q9NSG2-1 shown below. A BLASTP search was initially carried out on the BLAST server (Altschul, Gish et al. 1990). The default setting was used initially after the sequence was entered to the query box. Databases was set to Standard databases (set at non-redundant protein sequences), no organisms were excluded, or any boxes ticked in the exclude section. The algorithm under program selection was set to blastp (protein-protein BLAST). This sequence was used for the blast to identify possible homologs and curate the phylogenetic tree of C1ORF112.

```
>sp|Q9NSG2|CA112_HUMAN Uncharacterized protein C1orf112 OS=Homo
sapiens OX=9606 GN=C1orf112 PE=1 SV=1
MFLPHMNHLTLEQTFFSQVLPKTVKLFDDMMYELTSQARGLSSQNLEIQTTLRNILQTMV
QLLGALTGCVQHICATQESIILENIQSLPSSVLHIIKSTFVHCKNSESVYSGCLHLVSDL
LQALFKEAYSLQKQLMELLDMVCMDPLVDDNDDILNMVIVIHSLLDICSVISSMDHAFHA
NTWKFIIKQSLKHQSIIKSQLKHKDIITSLCEDILFSFHSCLQLAEQMTQSDAQDNADYR
LFQKTLKLCRFFANSLLHYAKEFLPFLSDSCCTLHQLYLQIHSKFPPSLYATRISKAHQE
EIAGAFLVTLDPLISQLLTFQPFMQVVLDSKLDLPCELQFPQCLLLVVVMDKLPSQPKEV
QTLWCTDSQVSETTTRISLLKAVFYSFEQCSGELSLPVHLQGLKSKGKAEVAVTLYQHVC
VHLCTFITSFHPSLFAELDAALLNAVLSANMITSLLAMDAWCFLARYGTAELCAHHVTIV
AHLIKSCPGECYQLINLSILLKRLFFFMAPPHQLEFIQKFSPKEAENLPLWQHISFQALP
PELREQTVHEVTTVGTAECRKWLSRSRTLGELESLNTVLSALLAVCNSAGEALDTGKQTA
IIEVVSQLWAFLNIKQVADQPYVQQTFSLLLPLLGFFIQTLDPKLILQAVTLQTSLLKLE
LPDYVRLAMLDFVSSLGKLFIPEAIQDRILPNLSCMFALLLADRSWLLEQHTLEAFTQFA
EGTNHEEIVPQCLSSEETKNKVVSFLEKTGFVDETEAAKVERVKQEKGIFWEPFANVTVE
EAKRSSLQPYAKRARQEFPWEEEYRSALHTIAGALEATESLLQKGPAPAWLSMEMEALQE
RMDKLKRYIHTLG
```

In addition, sequence analysis was carried out using https://web.expasy.org/protparam/ (Wilkins, Gasteiger et al. 1999).

2.1.2 Sequence alignment and phylogenetic tree

The evolutionary history of *C1ORF112* using 67 orthologous sequences from representative species of all major groups of Eukaryotes was analysed. The orthologues were sought in through the Orthologous Matrix project (OMA) (Zahn-Zabal, Dessimoz et al. 2020), cross-referenced using the Pfam entry (El-Gebali, Mistry et al. 2019) entry DUF4487, and through local BLAST searches (Altschul, Madden et al. 1997, Boratyn, Schaffer et al. 2012). Protein sequences were aligned using the L-INS-I strategy from MAFFT v747 (Katoh and Standley 2013). The gene tree of *C1ORF112* was inferred using the maximum likelihood program IQ-TREE multicore version 1.6. for Linux (Nguyen, Schmidt et al. 2015). The best model of substitution (JTT+F+G4) was selected using ultrafast bootstrap replicates (Minh, Schmidt et al. 2020). iTOL v5.5.1 (Letunic and Bork 2019) was used for gene tree visualisation and the images were obtained from PhyloPic (http://phylopic.org/).

2.1.3 Expression of C1ORF112 across tissues

Gene expression data from the genome type expression project (GTEx) version 8 (dbGaP Accession phs000424.v8. p2) was analysed using R script to determine the level of expression of C1ORF112 across tissues. The expression data was reviewed as transcripts per minute (TPM) The aim was to determine the baseline level of expression of C1ORF112 in normal tissue looking at the level of expression in each tissue by age, while also examining the tissue specificity of C1ORF112.

2.1.4 Analysis of genes co-expressed with C1ORF112 and possible physical interactors.

This project carries on from the work of van Dam, Cordeiro et al. (2012), who identified that C1ORF112 was co-expressed with genes such as RAD51, CCDC6 as well as many genes in the BRCA-Fanconi anaemia (FA) DNA damage response pathway, including BRCA1, BRCA2, FANCD2 and FANCI (van Dam, Cordeiro et al. 2012). The breast cancer susceptibility proteins BRCA1 (also known as FANCS), BRCA2 (FANCD1) and RAD51 (FANR) and its paralogs including the XRCC-2 (FANCU), and XRCC3 all function in homologous recombination repair (HRR). The database Genevestigator (version 8.0.1 were used to carry out this analysis. For Genevestigator, the Gene search tool was used to search the entire content for genes that are specifically expressed in a chosen set of conditions (a specific tissue type, cell line, cancer type or neoplasm, or perturbation). The Perturbations meta-profile comprises responses to various experimental conditions (drugs, chemicals, hormones, etc.), diseases, and genotypes. Results from the analysis above were cross-referenced with genes in the BIO-GRID database to find the most likely physical interactors with C1ORF112.

Gene ontology (GO) enrichment analysis was carried out on both the co-expressed genes using DAVID (Huang, Sherman et al. 2009), a threshold of 1.0 and then filtered by FDR P-Value and Benjamini correction of < 0.05. The genes and their p-values were then inputted into Revigo (Supek, Bosnjak et al. 2011) and a similarity threshold of 0.5 was set against a background of UniProt genes to give a Log10 p-value to determine the significance of the enrichment analysis.

2.2 Structural analysis

2.2.1 I-TASSER

The absence of a defined functional homology in any eukaryotic species prompted the use of the *de novo* method of structural analysis of C1ORF112. The de novo tool used was the I-TASSER online tool https://zhanglab.dcmb.med.umich.edu/I-TASSER/. I-TASSER uses an iterative threading hierarchical approach to predict protein structure. This is built on the existing platform LOMETS, a local meta threading server, which uses a PDB database to build a multiple sequence alignment profile to create deep sequence template profiles which can be used for full sequence *de novo* modelling. The results were then analysed to determine the quality of the structures provided and the best possible model structure was identified.

2.2.3 Ramachandran plot analysis

Another online tool called RAMPAGE was used to analyse the protein structures generated by I-TASSER. The online tool is called RAMPAGE. RAMPAGE was used to generate Ramachandran profiles for each of the structures. The PDB files were loaded onto the server and the plots were received as pdf files. Ramachandran plots are used the determine the torsion angles of the amino acid residues in a peptide chain known as the phi (ϕ)and psi (ψ) bonds, and how well they fit into the defined regions determined for alpha-helical structures or beta-sheet structures. The server used initially for RAMPAGE is now defunct and can no longer be accessed. The Ramachandran server used for subsequent analysis was https://zlab.umassmed.edu/bu/rama/index.pl and the ERRAT server used for validation analysis was https://www.doe-mbi.ucla.edu/errat/

2.2.4 QMEAN and Z-score

Further analysis was carried out on the structures using the QMEAN and Z-score analysis to determine how close the structure would be to the native structure of the protein, using data analysis for already resolved figures. QMEAN is the qualitative model energy analysis, a composite scoring function describing the major geometrical aspects of protein structure. 6 different structural descriptors (QMEAN6) are used. The local geometry is analysed by a new kind of torsion angle potential over three consecutive amino acids. A secondary structure-specific distance-dependent pairwise residue-level potential is used to assess long-range interactions. The descriptors used for the analysis are C-beta interaction energy, all-atom pairwise energy, solvation energy, torsion angle energy, secondary structure agreement and solvent accessibility agreement. This is then used to generate a z-score a mean average on the other scores.

2.3 Experimental analysis

2.3.1 Laboratory Materials

Reagents for general laboratory use were obtained from Bio-Rad (Hemel Hempstead, UK) Fisher Scientific (Loughborough, UK), and Sigma-Aldrich (St Louis, USA). Cell culture reagents were obtained from Sigma-Aldrich (St Louis, USA) and Thermofisher (Waltham, Massachusetts, USA). C1ORF112 CRISPR knockdown HAP1 cells were obtained from Horizon Discovery (ref HZGHC004317c001). The

HAP1 cell line is a human near-haploid cell line derived from male chronic myelogenous leukaemia (CML) cell line KBM-7. HZGHC004317c001 clone line has a 1bp insertion in exon 5 of C1orf112. This insertion is specifically at the chr1:169,803,224 The sequencing result is

“anactGTGTTTTGTTCATTGCTGTATTTGTAGCACCCAGCATGCTGACTAATACCTTTTCAGTGCACAAAAAATA
TATTCTAAGTGAAATTTCCTTCCTTATTCACAGACAATGGTGCAGCTCTTAGGAGCTCTCACAGGATGTGTTCA
GCATATCTGTGCCACACAAGGAATCCATCATTTTGGAAAATATTCAGAGTCTCCCCTCCTCAGTCCTTCATATAA
TTAAAAGCACATTTGTGCATTGTAAGGTGAGTAAAGGTCTAATTATACTTTGAATGGTATATAATCAATGTGCA
TAGGGGctgnAgtAAA”

Blast Align shows an A insertion highlighted in red

```
Query    102  TTCACAGACAATGGTGCAGCTCTTAGGAGCTCTCACAGGATGTGTTCAGCATATCTGTGC  161
C112     521  TTTACAAACAATGGTGCAGCTCTTAGGAGCTCTCACAGGATGTGTTCAGCATATCTGTGC  580


Query    162  CACACAAGGAATCCATCATTTTGGAAAATATTCAGAGTCTCCCCTCCTCAGTCCTTCATA  221
C112     581  CACAC– AGGAATCCATCATTTTGGAAAATATTCAGAGTCTCCCCTCCTCAGTCCTTCATA  639


Query    222  TAATTAAAAGCACATTTGTGCATTGTAAG  250
C112     640  TAATTAAAAGCACATTTGTGCATTGTAAG  668
```

when translated shows multiple STOP codons.

“LEU THR GLN ASN LYS **STOP** ARG HIS LYS HIS ARG GLY SER TYR ASP **STOP** LEU TRP LYS SER HIS VAL PHE PHE ILE **STOP** ASP SER LEU STOP ARG LYS GLU **STOP** VAL SER VAL THR THR SER ARG ILE LEU GLU SER VAL LEU HIS LYS SER TYR ARG HIS GLY VAL PHE LEU ARG **STOP STOP** ASN LEU LEU **STOP** VAL SER GLU GLY ARG SER GLN GLU VAL TYR **STOP** PHE SER CYS LYS HIS VAL THR PHE HIS SER PHE PRO ASP **STOP** TYR GLU THR TYR HIS ILE LEU VAL THR ARG ILE PRO ASP SER PHE”.

2.3.1 Mammalian cell line culture

All tissue culture work was carried out using an aseptic technique and was performed in a class Il hood with the laminar flow that was cleaned with 70 % ethanol both before and after use. Cells were stored and grown at 37°C and 5 % CO2 in a humidified cell culture incubator and were cultured using tissue culture grade plastics. All cell culture reagents (obtained from Sigma Aldrich, St Louis, USA; listed below) were pre-warmed in a water bath at 37°C before use. Dulbecco's Modified Eagle's Medium (DMEM) - 25 mM HEPES and sodium bicarbonate, 4500 mg/L glucose, sterile filtered, further supplemented with 10 % foetal bovine serum (FBS; sterile-filtered, non-US origin), 1 %

penicillin/streptomycin, 2 mM L-glutamine and 1% non-essential amino acids. 0.25 % Trypsin-EDTA solution was used to detach the cell from the flasks and Dulbecco's Phosphate Buffered Saline (PBS) was used to wash the cells during passaging.

2.3.2 Thawing cells

Cryovials containing cells frozen in DMEM and 10 % Dimethyl sulfoxide (DMSO) were defrosted in a water bath at 37°C for 30 sec. 1 ml DMEM was added to the cells and gently mixed via pipetting. The cell suspension was centrifuged at 1300 rpm for 5 min and the DMSO containing supernatant was removed. The cell pellet was resuspended in 1 ml DMEM, transferred to a T75 culture flask containing 11 ml DMEM and incubated at 37°C and 5 % CO2 in a humidified cell culture incubator.

2.3.3 Passaging cells

Once cells had reached 80-90 % confluency they were ready to be split. The existing DMEM was aspirated, and the monolayer of cells was washed with 5 ml PBS. The PBS was then aspirated, and cells were incubated with 1 ml 0.25 % trypsin-ETDA for 5 min at 37°C and 5 % CO2 in a humidified cell culture incubator. Once all adherent cells had detached from the flask, 9 ml DMEM was added to neutralise the trypsin and cells were mixed via pipetting to create a single-celled suspension. Cells were split 1:10 with 1 ml of cell suspension being added to a new T75 culture flash with 12 ml of DMEM. Cells were incubated at 37°C and 5 % CO2 in a humidified cell culture incubator.

2.3.4 Cryogenic storage of cells

Cells were prepared as described above (see section 2.3.3) to bring the cells into a single-celled suspension. Following trypsinisation and subsequent neutralisation with DMEM, the cell suspension was transferred to a 15 ml tube and centrifuged at 1300 rpm for 3 min. The medium was removed, and the cell pellet was resuspended in a 1ml freezing medium (DMEM with 10 % DMSO) and transferred to a cryovial. The cryovials were placed in a cell freezing container CoolCell from corning and placed into an 80°C freezer for 24 h before being removed and transferred to long-term storage in liquid nitrogen.

2.3.5 Seeding cells

Cells were seeded prior to various experiments and assays, cell confluency of between 70 – 80% confluency were considered optimal for use. A single-celled celled suspension was obtained following trypsinisation as described (see section 2.3.3). The number of cells was counted using a haemocytometer. In general, 1 x 10^6 cells were seeded in 5 ml supplemented DMEM for a T75 flask

2.3.6 Harvesting cells

Tissue culture dishes containing cells that were <90 % confluent was removed from the humidified incubator and the media was aspirated. The dishes were washed with a volume of cold PBS (5 ml for 10 cm dish or 2 ml for 3.5 cm dish) before being aspirated. The second volume of cold PBS was added, adhered cells were carefully Scraped and transferred to a pre-cooled 15 ml tube -this process was repeated a second time. The 15 ml tube was centrifuged at 2000 rpm for 5 min at 4°C, the Supernatant removed, and the pellet resuspended in 1 ml of cold PBS before.

2.3.7 Cell fractionation

Fresh cell pellets were prepared as described in **section 3.3.7** and suspended in 2 PCV of buffer I (10mM Tris-HCl pH 8, 2.5mM $MgCl_2$, 0.5% NP-40, 1ug/ml of each of the following protease inhibitors (pepstatin, aprotinin, chymostatin, and leupeptin), 100mm PMSF, 1mM N-ethylmaleimide (NEM)). The resuspended pellet was incubated in ice for 10 mins and then centrifuged at 10000 rpm and 4^{o}c for and the supernatant containing the soluble (S) fraction was transferred to a fresh 1.5ml tube. The remaining nuclear pellet was resuspended in 2PCV of buffer II (20mM $2NaPO_2$ pH 8.0, 0.5M NaCl, 1mM EDTA, 0.75% Triton X-100, 10% glycerol, 5mM $MgCl_2$, 1ug/ml of each of the following protease inhibitors (pepstatin, aprotinin, chymostatin, and leupeptin), 100mm PMSF, 1mM N-ethylmaleimide (NEM)) and then incubated in ice for 10 mins and then centrifuged at 10000 rpm and 4^{o}c. The supernatant containing the chromatin-bound (CB) fraction was into a fresh 1.5ml tube, the protein concentration was measured using Bradford assay (**see section 3.3.8**)

2.3.8 Measuring protein concentration

Protein concentration was measured using the Bradford protein assay. In a 3 ml plastic cuvette, 960ul of diluted Bradford protein assay dye reagent (1 in 4 using dH2O), 38 l of dH20 and 2ul of protein extract were mixed. A protein standard sample was prepared by mixing 96ul of diluted Bradford protein assay dye reagent and 40ul of 0.2 mg/ml BSA. A blank was also prepared by mixing 960ul of diluted Bradford protein assay dye reagent and 40ul of dH20. Samples were incubated for 5 min at room temperature and then absorbance was measured at an optical density (OD) of 595 nm (A595) using a UV spectrophotometer, following zeroing of the spectrophotometer with the blank sample. The protein standard sample was used as a reference to convert the A595 into mg/ml using the calculation: Sample concentration (mg/ml) =(0.2/A595BsA) x 40 x A595sample

2.3.9 Antibodies

The antibodies employed throughout this project were used to probe for specific proteins after immunodepleting of endogenous proteins during immunoblot analysis. The primary antibodies are condensed in Table 3.1 and secondary antibodies in Table 3.2 below.

Table 2. 1 **Primary antibody**.

List of primary antibodies used throughout this research project. Host organism, clonality, dilution and source.

Antibodies	Host organism	Clonality	Dilution	Source	Code
Anti-C1ORF112	Rabbit	polyclonal	1:500	Sigma-Aldrich	HPA024451-25UL
Anti- Actin	Mouse	Monoclonal	1:20000	Sigma-Aldrich	A5441
Anti-Fibrillarin	Mouse	Monoclonal	1:2000	Abcam	ab4566

Table 2. 2 **Secondary antibodies.**

List of the fluorescently tagged secondary antibodies used throughout this research project to target the primary antibodies. Host organism, target immunoglobulin isotype, dilution and source are displayed.

Antibodies	Host organism	Dilution	Source	Code
Alexa Fluor 680 Anti-Mouse lgG	Goat	1:10000	Invitrogen	A21057
Alexa Fluor 680 Anti-Rabbit lgG	Goat	1:10000	Invitrogen	A21076
R Dye 800 Anti-Rabbit lgG	Goat	1:10000	Li-Cor	926-32211
R Dye 800 Anti-Mouse lgG	Goat	1:10000	Li-Cor	926-32210

2.3.10 Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE)

SDS-PAGE was employed to separate the proteins present in a sample based on their molecular weight. Gels (10 %) were prepared by initially making the separating portion of the gel (377 mM Tris-HCl pH 8.8, 0.1 % SDS, 2 mM EDTA, 10 % acrylamide/bis solution (30:0.8; Bio-Rad, Hemel, Hempstead, UK), 0.1 % ammonium persulphate (APS), and 0.1 % tetramethyl ethylenediamine (TEMED) and pour this into a 1.5 mm gel cassette until the cassette was % full. The separating gel solution was then overlaid with 1 ml of 100 % ethanol and left to set, to create a level separating gel and remove any bubbles. After 30 min the gel had set, the ethanol was poured off and the cassette was washed with dH_2O. To the remaining top % of the gel cassette, the 5% stacking gel solution (125 mM Tris-HCl pH 6.8, 0.1 % SDS, 2 mM EDTA, 5 % acrylamide/bis solution (30:0.8; Bio-Rad, Hemel, Hempstead, UK), 0.1% ammonium persulphate (APS), and 0.1 % tetramethyl ethylenediamine (TEMED) was poured on top of the separating gel. Either a 10-well or 15-well comb was inserted into the gel cassette and the stacking portion of the gel was left to set for -30 min. Protein extracts (typically 40 g) or in vitro ubiquitylation reaction samples were prepared in SDS-PAGE sample buffer (25 mM Tris-HCI pH 6.8, 2.5 % mercapto-ethanol, 1 % SDS, 10 % glycerol, 0.05 mg/ml bromophenol blue, and 1 mM EDTA) and heated for 5 min at 95°C before loading on the 10% polyacrylamide gel. Samples were loaded and electrophoresis was performed in 1x Tris-glycine SDS (TGS) running buffer (25 mM Tris-HCI pH 8.3, 192 glycine, and 0.1 % SDS: Fisher Scientific, Loughborough, UK) at 125 V for 110 min in an SDS-PAGE Mini Gel Tank (Fisher Scientific, Loughborough, UK). The Precision Plus Protein All Blue Pre-Stained Protein Standards (10 KDa- 250 KDa; Bio-Rad, Hemel, Hempstead, UK) were used as standard protein markers. Protein levels following electrophoretic separation via SDS-PAGE were analysed by Immunoblotting.

2.3.11 Immunoblotting

Immunoblot analysis was conducted to probe for and visualise specific proteins of interest by transferring the proteins from the SDS-PAGE gel following electrophoresis (see section 3.2) to an Immobilon-FL polyvinylidene difluoride (PVDF) membrane (Millipore, Watford, UK). Firstly, the PVDF membrane was activated in 100 % methanol for 15 sec, washed in dH2O for 1 min and washed in cold transfer buffer (1x Tris-glycine (TG; 25 mM Tris-HCl pH 8.3, 192 mM glycine; Fisher Scientific, Loughborough, UK), 20 % methanol) for >1 min. Two pieces of filter paper and two sponges were also soaked in cold transfer buffer for>1 min. One sponge was taken and placed in the base of a Mini Blot Module (Fisher Scientific, Loughborough, UK), followed by a piece of filter paper. The SDS-PAGE gel was removed from the gel cassette, rinsed in transfer buffer before being placed on top of the filter paper in the blot module. The PVDF membrane was placed on top of the acrylamide gel, followed by the second piece of filter paper and the second sponge. The top of the blot module was affixed, and the blot module was transferred to the Mini Gel Tank. The interior of the blot module was filled with cold transfer buffer and the tank was filled with dH20. The transfer was conducted at 20 V for 1 h. Following the transfer, the PVDF membrane was washed in 1xPB for 5 min and then blocked in Odyssey blocking buffer (LI-COR Biosciences, Cambridge, UK) for 1 h at room temperature with 25 rpm rocking. The blocking buffer was then removed and the PVDF membrane incubated with the primary antibody (Table 3.1) diluted in Odyssey blocking buffer diluted 1:1 with 1x PBS and containing 0.1% Tween 20 at 4°C overnight at 25 rpm rocking. The PVDF membrane was then rinsed three times with 1x PBS and contained 0.1% Tween 20 at room temperature for 5 min and 25 rpm rocking. The PVDF membrane was incubated with the secondary antibody (Table 3.2) diluted in Odyssey blocking buffer diluted 1:5 with 1x PBS and containing 0.1% Tween 20 for 1h at room temperature with 25 rpm rocking. The PVDF membrane was then rinsed three times with 1x PBS and containing 0.1% Tween 20, followed by one wash with 1x PBS at room temperature for 5 min and 25 rpm rocking. The membrane was imaged and quantified using the L-COR Odyssey Infrared Imaging System (LI-COR Biosciences, Cambridge, UK).

2.3.12 Agarose gel electrophoresis

Agarose gel electrophoresis was utilised to analyse DNA molecules based on their size in kilobases (kb). Agarose gels (1 % or 0.8 %) were prepared by dissolving agarose (broad separation range) in 1x Tris-acetate-EDTA (TAE) buffer (both from Fisher Scientific, Loughborough, UK), and heated using a microwave until the agarose had dissolved. Once cooled to the touch, 5nM SYBR Safe DNA gel stain (Fisher Scientific, Loughborough, UK) was added to the solution and poured into the gel tray of a Mini-Sub Cell GT electrophoresis tank (Bio-Rad, Hemel, Hempstead, UK), comb was added and left to set. Once set the comb was removed from the gel and the gel tray was placed into the electrophoresis tank and submerged in 1x TAE. DNA samples were prepared in 1x DNA loading dye (Fisher Scientific, Loughborough, UK) and loaded onto the gel. The Gene-Ruler 1 kb DNA ladder (Fisher Scientific, Loughborough, UK) was also prepared in 1x DNA loading dye and loaded onto the gel as a reference for DNA size. Electrophoresis was performed at 80 V for 1 hr and the gel was imaged using the L-COR Odyssey Infrared Imaging System (LI-COR Biosciences, Cambridge, UK).

2.3.13 Induction of oxidative stress

DNA damaged HAP1 cells were seeded into 10 cm dishes, grown until 30-50 % confluency. Oxidative stress was induced through treatment with 10 Gy ionising radiation (IR) using the CellRad X-Ray irradiator (Faxitron, Tuscan, USA). Unirradiated controls were treated with media only. Following this, the serum-containing media was removed, cells washed with 1x PBS, and media replaced. Cells were subsequently harvested at various time points post-irradiation (0-4 hr) following further incubation at 37°C.

2.3.14 Clonogenic assays

HAP1 cells were seeded into 3.5 cm dishes, grown until 30-50 % confluent before being treated with hydrogen peroxide or x-rays. Cells were incubated on ice for 5 min to suppress DNA repair activity following radiation (0-4 Gy), before being washed with 1x PBS, trypsinised with 200 ul 0.25 % trypsin-EDTA and neutralised with 800 pl supplemented DMEM. The number of cells was counted using a haemocytometer. A defined number of cells were then seeded in triplicate at two seeding densities in 2 ml supplemented DMEM per six-well plate. The six-well plates and incubated for 7-10 days at 37°C and 5 % CO2 in a humidified cell culture incubator to allow colony growth. Colonies were grown until -well defined, non-overlapping colonies could be visualised under a light microscope. Following this, the media was removed, wells washed in PBS and the colonies were then fixed and stained with 6% glutaraldehyde, 0.5 % crystal violet (Fisher Scientific, Loughborough, UK) for 30 min, washed and left to air dry. Colonies were counted using the GelCount colony counter from Oxford Optronix (Oxford, UK). Relative colony forming units (surviving fractions) were expressed as colonies per treatment relative to colonies observed in the unirradiated control for each treatment, calculated using the calculation shown below. Average surviving fractions were calculated, and values plotted on a log scale against ionising radiation. Statistical analysis was performed by the CFAssay for R package (348).

$$plating\ efficiency = \frac{number\ of\ colonies\ for\ unirradiated}{seeding\ densityof\ unirraddiated\ control}$$

$$surviving\ fraction = \frac{number\ of\ clonies\ for\ selected\ condition}{seeding\ densityof\ selected\ condition * average\ plating\ efficiency}$$

2.3.15 Neutral comet assay

HAP1 cells were seeded into T25 flasks, grown until 70-80 % confluent before being briefly trypsinised, and diluted to 200,000 cells/ml. 250 ul aliquots of the cell suspension were then pipetted into a 24 well plated and treated with either hydrogen peroxide (100uM – 400uM) or x-rays (2Gy -8gy) and embedded on a microscope slide in low melting agarose (Bio-Rad, Hemel Hempstead, UK). For repair studies, the slides were incubated for up to x h in a humidified chamber at 37°C to allow for DNA repair, before lysis containing 2.5 M NaCl, 100 mM EDTA, 10 mM Tris-HCl pH 10.5, 1 % N-lauroylsarcosine, 1 % DMSO and 1 % (v/v) Triton X-100. for 1 h at 4°C. The slides were then incubated in the dark for 30 min in cold electrophoresis containing 1 × TBE, pH 9.5 at 25 V, ~20 mA for 25 min. Finally, slides were washed three times with 1 × PBS before allowing them to air-dry overnight. The slides were rehydrated for 30 min in water (pH 8.0), stained for 30 min with SYBR Gold (Life Technologies, Paisley, UK) diluted 1:10,000 in water (pH 8.0) and again air-dried overnight. Cells (50 per slide, in duplicate) were analysed from the dried slides using the Komet 6.0 image analysis software (Andor Technology, Belfast, Northern Ireland) and % tail DNA values averaged from at least three independent experiments.

Chapter 3: Conservation, Expression and Co-expression of C1ORF112

3.1 Introduction

Examining the evolutionary conservation of proteins is a valuable tool for understanding their function. At the sequence level, the conservation of each residue can be used to infer the importance of the regions of proteins such as the domains necessary for activities or interaction with other proteins (Wong 2019). Sequence conservation analysis of protein complexes has been widely applied to identify various protein homologues to detect residues required for functionality (Choi, Yang et al. 2009). In evolutionary history, all amino acids have been under selective evolutionary pressure from various factors such as folding, recombination rate, and protein-protein interaction (Elcock and McCammon 2001, Sim and Creamer 2004). Understanding the conservation of a protein among the various Phyla could give insight into its functional importance and relevance. In most cases, full sequence conservation is rare and instead particular sets of sequences are conserved, this is the basis for domain conservation and is intricately linked to protein functionality.

In addition, control of gene expression which involves transcription, translation and the turnover of mRNA and protein is linked to the role of the subsequent protein product and its turnover. Multi-protein complexes are involved in essentially all cellular processes. A protein's function can be defined as a combination of its properties, cellular localization, stoichiometry and its interacting partners (Kerrigan, Xie et al. 2011). Relevant context is key to studying and understanding protein function to give a better understanding of the innate role of the protein in question. Characterizing protein function would naturally involve various experiments such as high-throughput screening (HTS) an example being the yeast two-hybrid (Y2H) system, and tandem affinity purification (TAP) followed by mass spectrometry (MS) to characterize interacting partners. Cellular experiments such as sensitivity assays, knockout down and over-expression analysis and colocalization analysis reveal the innate role of the protein.

In this chapter, the conservation, expression levels and co-expression of C1ORF112 will be examined using bioinformatics modelling and analysis of databases. Preceding studies showed that C1ORF112 was co-expressed with genes such as RAD51, CCDC6 as well as many genes in the BRCA-Fanconi anaemia (FA) DNA damage response pathway, including BRCA1, BRCA2, FANCD2 and FANCI (van Dam, Cordeiro et al. 2012). The breast cancer susceptibility proteins BRCA1 (also known as FANCS), BRCA2 (FANCD1) and RAD51 (FANR) and its paralogs including the XRCC-2 (FANCU), and XRCC3 all function in homologous recombination repair (HRR) (see Figure 1.3). Genevestigator (version 8.0.1) was the primary tool used to carry out the co-expression analysis. The Gene Search tool was used to search the entire content for genes that are specifically expressed in a chosen set of conditions (a specific tissue type, cell line, cancer type or neoplasm, or perturbation). The Perturbations meta-profile comprises responses to various experimental conditions (drugs, chemicals, hormones, etc.), diseases, and genotypes. The sequence of C1ORF112 obtained from the UniProt database is illustrated alongside various homologues used for the conservation study. Multiple sequence alignment (MSA) for chosen model organisms is also shown and analysed, in addition to, tissues expression profile at various age categories. Finally, co-expression analysis, gene ontology (GO) analysis, KEGG pathway analysis and hypothesized C1ORF112 functional pathway are discussed.

3.2 Sequence conservation and phylogenetic analysis

The sequence of C1ORF112 was obtained from UniProt ID Q9NSG2 isoform Q9NSG2-1. This sequence was used for the blast to identify possible homologs and curate the phylogenetic tree of C1ORF112. The amino acid sequence for C1ORF112 was obtained from the UniProt database https://www.uniprot.org/uniprot/Q9NSG2#sequences the total number of amino acids is 853, the molecular weight of the protein product is 96.6kDa (96554.36Da). The theoretical isoelectric point for C1ORF112 is 5.64. BLAST analysis was initially carried out using the default setting to maximize the number of sequence orthologues captured, this was then followed by cross-referencing the captured sequences with available data across several databases such as PFAM and OMA, to identify consistent orthologues. The final list of orthologues was used to generate the phylogenetic association of C1ORF112.

Table 3. 1 **List of representative C1ORF112 homologues in various species.**

Sequences used for the phylogenetic analysis showing the orthologous relationship, protein ID and per cent similarity. Ranking the protein similarity from highest to lowest with light grey shading indicating mammals, orange shading indicating reptiles, green shading indicating birds, yellow indicating fish and no shading indicates Branchiostoma floridae, the earliest of Chordates.

Taxon	**Protein ID**	**% Sequence similarity with Human C1ORF112**	**Phylogenetic Order**
Pan paniscus	A0A2R9AG57	99.1	Primate
Pan troglodytes	H2R2Z2	99.1	Primate
Gorilla gorilla	ENSGGOG00000016160	98.8	Primate
Sus scrofa	A0A286ZX35	86.5	Artiodactyla
Equus caballus	ENSECAG00000008931	86.5	Perissodactyla
Mus musculus	CA112_MOUSE	72.4	Rodentia
Ornithorhynchus anatinus	F6WQR8	67.9	Monotremata
Chrysemys picta bellii	ENSCPBG00000013015.1	63.8	Testudines
Chelonoidis abingdonii	ENSCABG00000018673.1	62.8	Testudines
Melopsittacus undulatus	ENSMUNG00000013129.1	59.8	Psittaciformes
Gallus gallus	ENSGALG00000003368	58.7	Galliformes
Ficedula albicollis	U3JZM8	58.4	Passeriformes
Latimeria chalumnae	H2ZSH9	52.4	Actinistia
Lepisosteus oculatus	W5LZV5	50.7	Lepisosteiformes
Danio rerio	CA112_DANRE	45.9	Cypriniformes
Branchiostoma floridae	C3Z504	32.4	Leptocardii

Table 3.1 shows the ranking of the protein sequences from the most similar, to the least similar. Pan paniscus (Bonobo) and Pan troglodytes (Chimpanzee) are primates and they both have the highest level of similarity of 99.1% with the human C1ORF112 protein. In the higher order of phylogenetic superfamily, Hominidae humans and chimpanzees (Pan genus) deviate from the same ancestor. C1ORF112 similarity between the genus Homo and Pan shows the same pattern as it may have been present in the ancestor of both humans and Chimpanzees, as shown in Figure 3.1. The level of protein similarity declines the more removed from humans, the organism as shown in Table 3.1. The following organisms on the table with high similarity are vertebrates and there is a range of the level sequence similarity amongst them from 80% to 30%, which is the Branchiostoma floridae, a lancelet of the Branchiostoma genus. Although the Branchiostoma is morphologically related to tunicates, sequence analysis suggests they are more closely related to the vertebrates (Boore, Daehler et al. 1999).

To further understand the sequence variation between species, 3 representative sequences from different vertebrates were picked, usually, this is carried out with species that are scientific models such as C. elegans and Drosophila, C1ORF112 does not appear to be present in these organisms but is present in other organisms in their phylogenetic clades. The 3 representative organisms picked were mouse (mammal), chimpanzee (primate) and frog (amphibian)

3.2.2 Multiple sequence alignment

Multiple sequence alignment of the various homologues shows high levels of conservation in the gene sequences in the models above. The total sequence length for each model is above 800 amino acids, using human C1ORF112 as a reference, as it is 830 amino acids in length, only the chimpanzee has the same number of nucleotides. The other model species have at least 50 additional nucleotides, and if these nucleotides have an impact on the function of the protein, is currently unknown. These proteins were grouped under the domain of unknown function DUF4487 by Pfam (El-Gebali, Mistry et al. 2019). The family of proteins were said to contain a conserved WCF tripeptide in most of the sequences of the species, which is usually preceded by the amino acids LAMDA, followed by the amino acids LARY predominantly in the vertebrate sequences. However, if these amino acids are functionally relevant is not evident yet. However, sequence alignment of model organisms for which the gene is present shows the WCF tripeptide consistently being present.

The evolutionary history of *C1ORF112* using 67 orthologous sequences from representative species of all major groups of Eukaryotes was investigated. The orthologues were sought in the Orthologous Matrix project (OMA) (Zahn-Zabal, Dessimoz et al. 2020), using the Pfam entry (El-Gebali, Mistry et al. 2019) entry DUF4487, and through local BLAST searches (Altschul, Madden et al. 1997, Boratyn, Schaffer et al. 2012). Protein sequences were aligned using the L-INS-I strategy from MAFFT v747

(Katoh and Standley 2013). The gene tree of *C1ORF112* was inferred using the maximum likelihood program IQ-TREE multicore version 1.6. for Linux (Nguyen, Schmidt et al. 2015). The best model of substitution (JTT+F+G4) was selected using ultrafast bootstrap replicates (Minh, Schmidt et al. 2020). iTOL v5.5.1 (Letunic and Bork 2019) was used for gene tree visualisation and the images were obtained from PhyloPic (http://phylopic.org/).

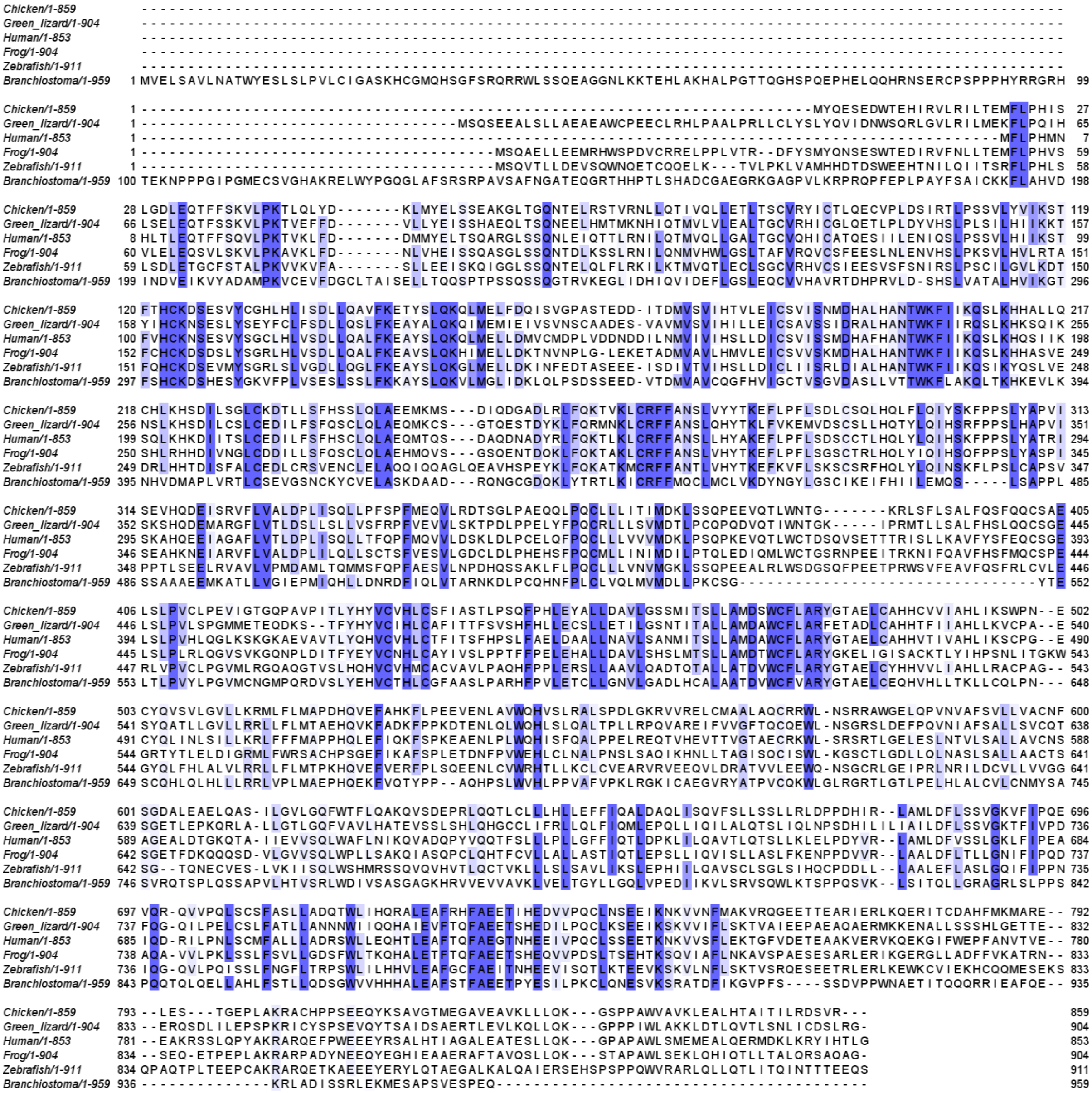


Figure 3. 1 **Multiple sequence alignment of representative organisms.**

Representative organisms from different Classes of the Phylum Chordate to understand sequence conservation of C1ORf112 within the Phylum. The threshold of 80% was set (conserved residues highlighted in blue), showing approximately 30% conservation across the different models chosen Image obtained using Jalview.

These organisms were chosen to understand the conservation of C1ORF112 in vertebrates, since the sequence similarity varies across the different classes in the phylum. In addition, conservation of DNA repair pathways such as HR pathway and FA pathways are well conserved in vertebrates (Yuan, Song et al. 2010, Kawale and Sung 2020).

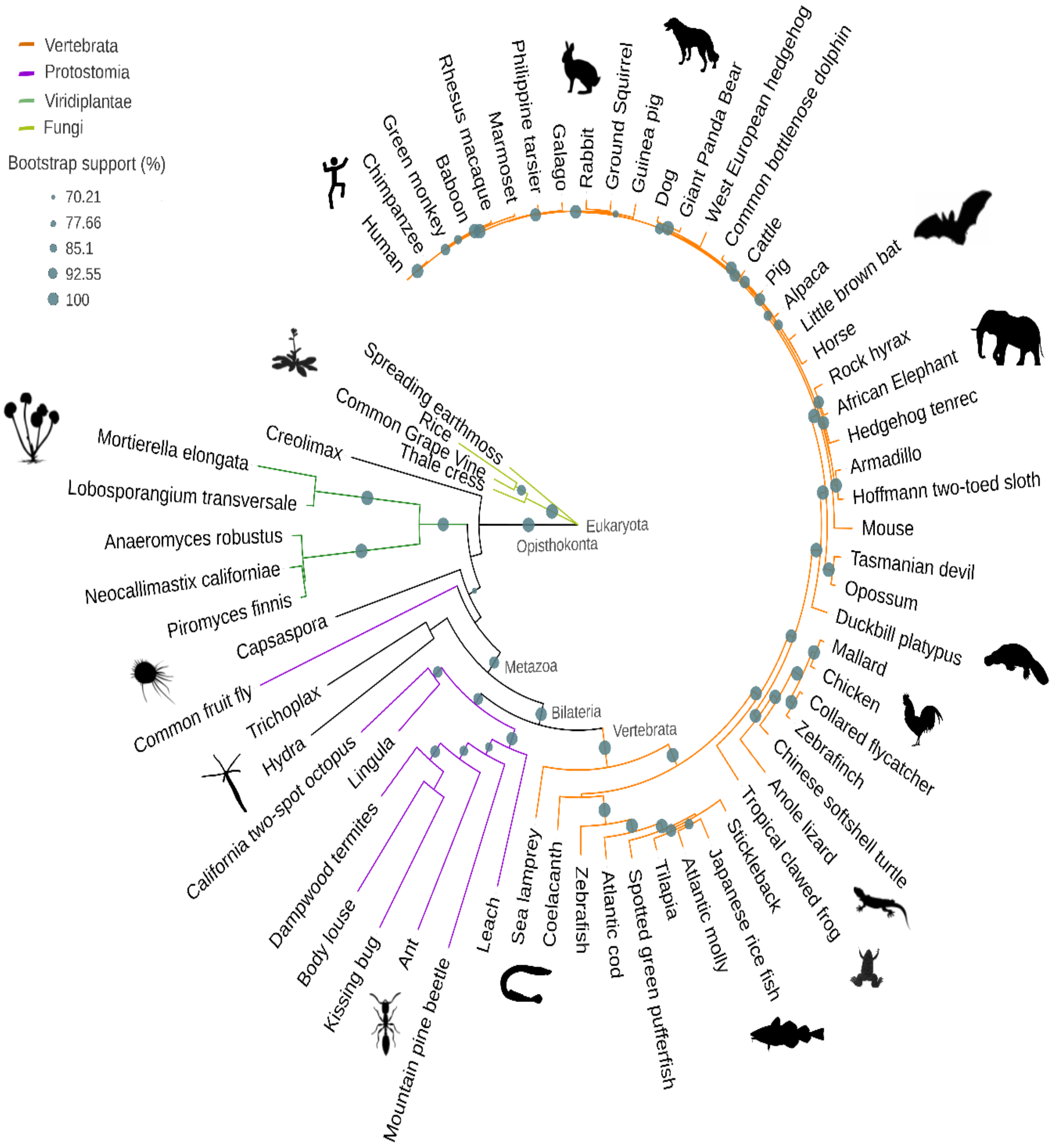


Figure 3. 2 **C1ORF112 gene tree**.

The phylogenetic relationship between the main groups of Eukaryota and between species are in general, well conserved from plants to humans, with bootstrap support >70%. Orange indicates phylum Vertebrata, light green for fungi and dark green for plants and purple for protostome. Grey lines are for unicellular organisms.

Higher-order conservation of C1ORF112 is reflected in Figure 3.2 as with Table 3.1 showing the level of sequence similarity of human C1ORF112 with the homologues in other species. All the primates used in the phylogenetic conservation are clustered together at the top of the figure alongside other species in the phylum Vertebrata (orange lines), all the way down to coelacanth and sea lamprey, which are also classed furthest on the vertebrate phylum and closer to cephalochordates such as Branchiostoma. This level of conservation shows that C1ORF112 does retain the natural phylogenetic classification and could be a hereditary protein from ancestral species, that is C1ORF112 is orthologous across species rather than paralogues. Hence the possibility of it being lost in certain species such as *C. elegans and Drosophila* is highly likely. Considering this, a possible homologue was found in Drosophila, but it lacked the WCF tripeptide and had a 12% sequence similarity with the human C1ORF112 and as seen in Figure 3.1 (purple lines) does not conserve well with other Protostomia which consists of nematodes, arthropods, flatworms, annelids, and molluscs. C1ORF also appears to be present in plants and fungi (green lines). The presence in the plants does indicate that C1ORF112 could be a well-conserved historical protein, while its function is currently not well studied, it may play a vital role in the cell.

3.3 Expression of C1ORF112 in tissues and co-expression analysis

C1ORF112 was established to be conserved across the Metazoa phyla, it is also present in some fungi species and some plants and worms. It was necessary to determine the level of expression of C1ORF112 across human tissues. Data from the genome type expression project (GTEx) version 8 (dbGaP Accession phs000424.v8. p2) was analysed to determine the level of expression of C1ORF112 across tissues. The aim was to determine the baseline level of expression of C1ORF112 in normal tissue looking at the level of expression of C1ORF112 in broad tissues and expression in each tissue by age, while also examining the tissue specificity of C1ORF112.

Expression for C1ORF112 was observed to be higher transcripts per kilobase million (TPM) in testis (light purple) and cells transformed by Epstein-Barr virus (orange) shown in Figure 3.2. the expression was shown to have a mean of above 10TPM (white bar) but ranging up to 20TPM, the rest of the tissues had less than 5TPM, with the brain (yellow), vagina and uterus (light purple, far-right) having the next highest levels of expression. Cultured fibroblasts (purple middle) did show a range of expression; however, the mean expression level was below 5TPM. This shows that C1ORF112 is constitutively expressed in all tissues, but it is expressed more frequently in testis and EBV-transformed lymphocytes. These sets of cells are particularly known to be rapidly proliferating cells could be a likely explanation as to why C1ORF112 has a higher read count in these cells. The testis is tasked with sperm production and androgen and the level of sperm production is controlled by the Follicle stimulation hormone, there is the presence of active rapidly dividing cells (Wang and Swerdloff 1992, Tiwana and Leslie 2022), this is like EBV transformed lymphocytes which, become actively proliferating lymphoblastoid cell lines through the activation of the NF-kB pathway and are associated with several cancers as a result (Cahir McFarland, Izumi et al. 1999).

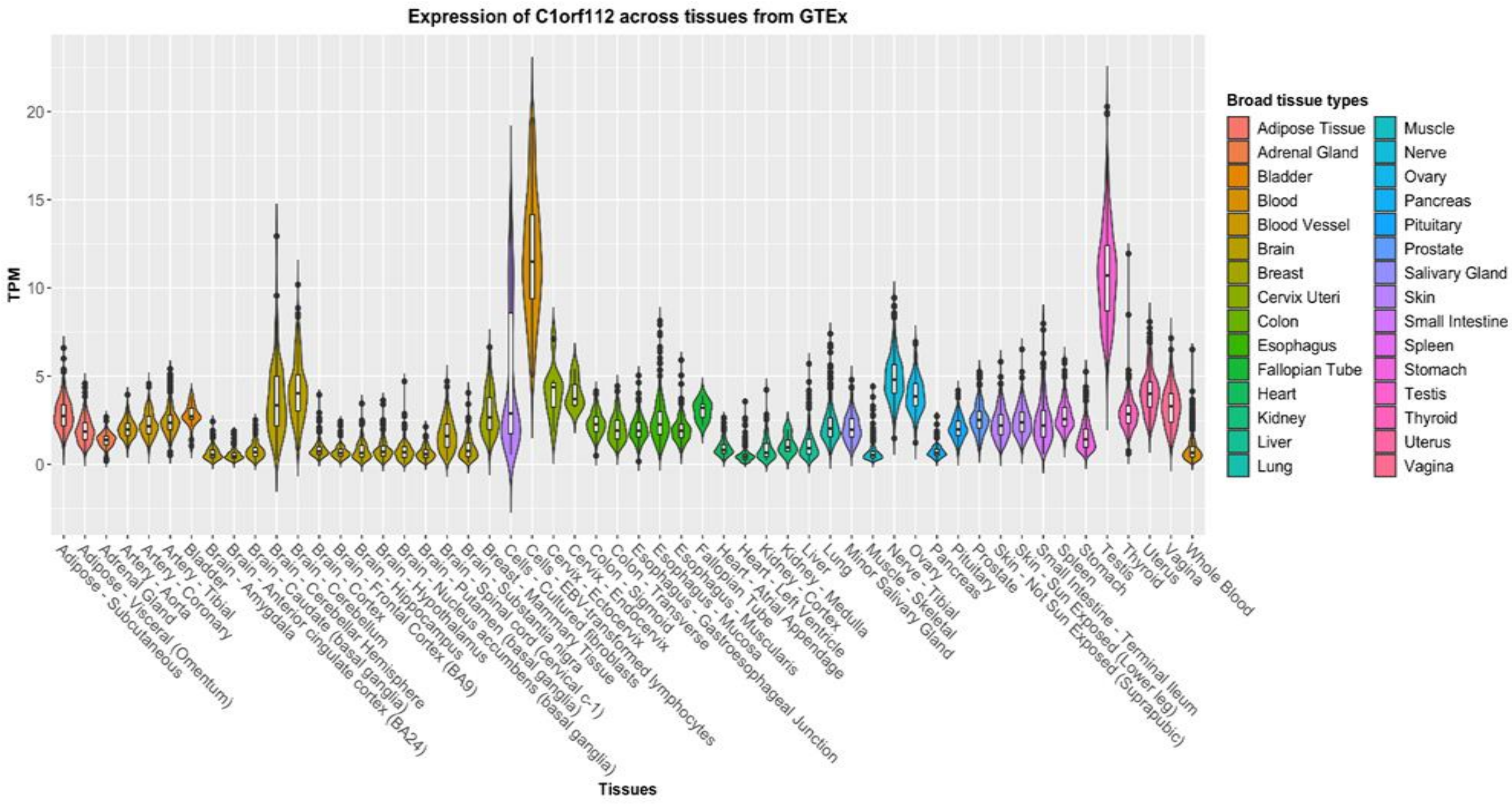


Figure 3. 3 **Relative expressions of C1ORF112 across various cells in the body**.

Looking at the expression level in terms of transcripts per minute, C1ORF112 shows higher levels of expression in the testis and Epstein-Barr virus-transformed lymphocytes cells compared to other cells in the body. The image was generated after analysis of samples from the Gtex portal https://gtexportal.org/home/gene/C1ORF112

To understand if ageing had a role in C1ORF112 expression, data from the GTex database was downloaded and tissue samples were grouped into age brackets and a comparison of the expression levels in each tissue was analysed using R (Figure 3.3 – Figure 3.11).

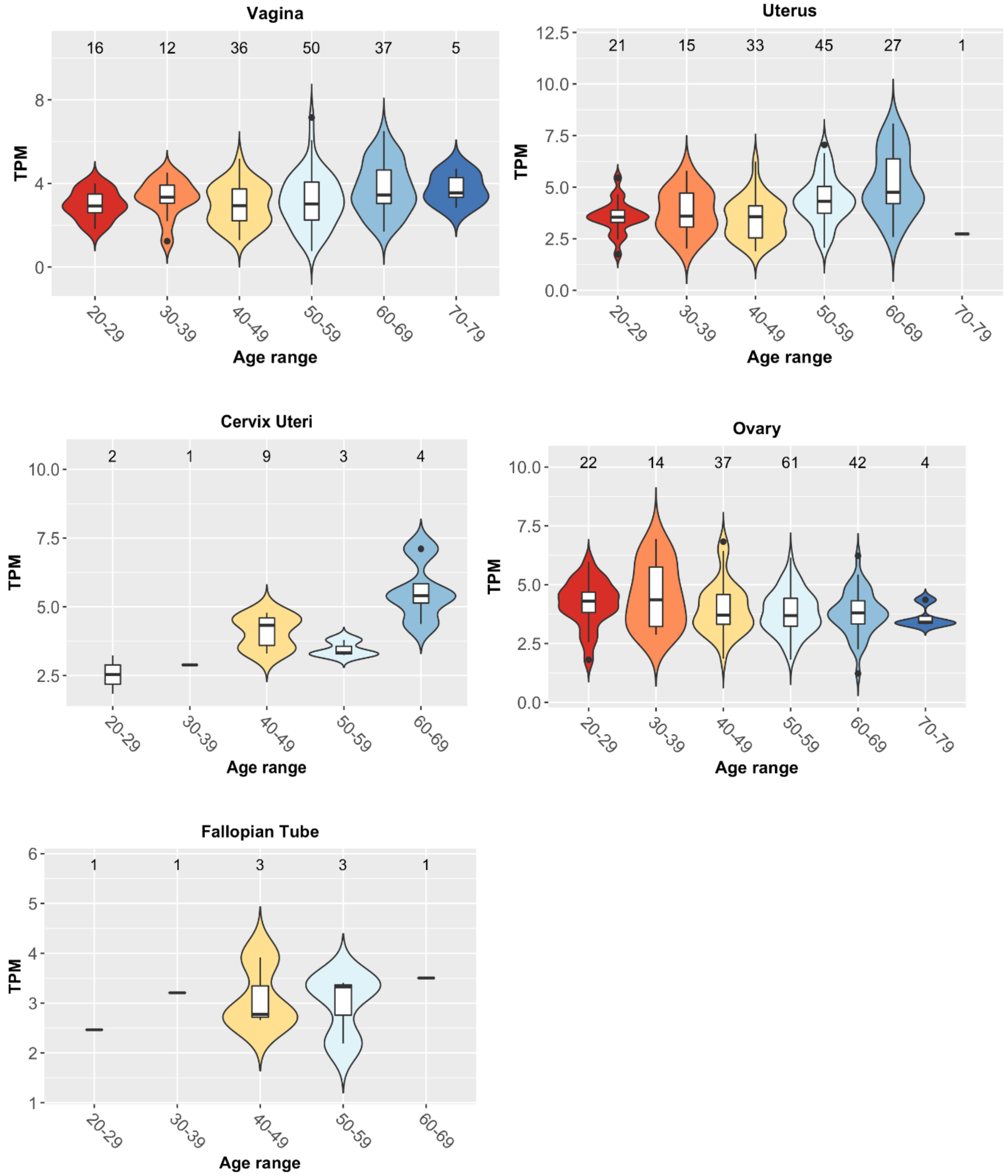


Figure 3. 4 **Age-related expression of C1ORF112 in Vagina, Uterus, Cervix uteri, Ovary and Fallopian tube**.

The number at the top of each age category signifies the number of samples available in each category.

Vagina shows the average level of expression around or below 4TPM. The level of expression of C1ORF112 in the Uterus is similar with the average just above 5TPM. The Cervix and Fallopian tube did not have enough sample sizes for adequate analysis. The Ovary had a mean expression level below 5TPM across all samples including the 70-79 age group. C1ORF112 expression in the Vagina showed consistency across the age categories, with the biggest variability in expression around 50 -59 and 60-69, this variation can also be seen in the Uterus as the average TPM is slightly increased compared to the other age categories. The 70-79 age group had a smaller number of samples for adequate comparison, with 1 sample available for comparison in the Uterus and 5 in the Vagina. The Cervix and Fallopian tubes did not have enough samples for analysis across the age groups and the 70-79 age group had no samples and so the expression of C1ORf112 could not be accurately analysed. The ovary followed the uterus and the vagina with having a mean TPM of less than 5 across all age groups.

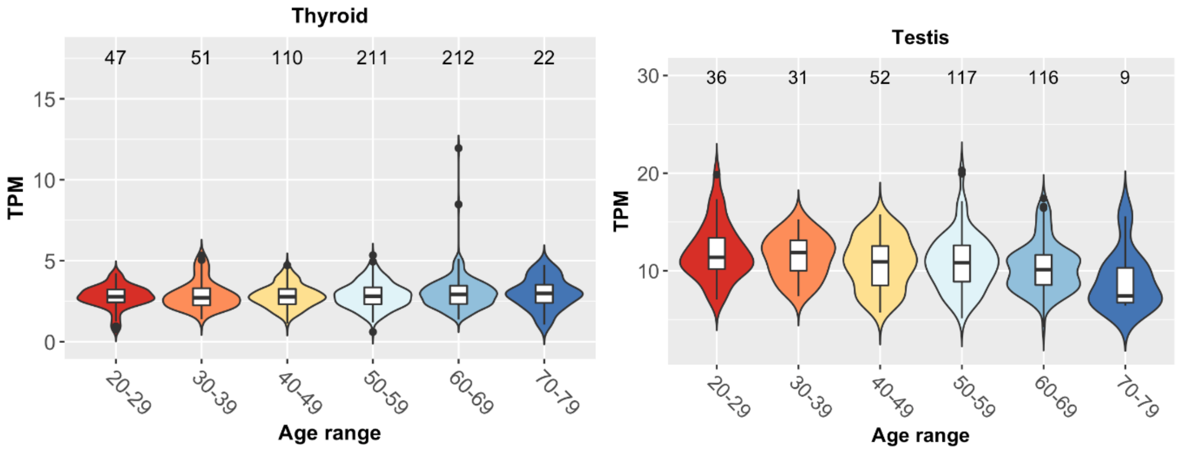


Figure 3. 5 **Age-related expression of C1ORF112 in the Thyroid and the Testis**.

The number at the top of each age category signifies the number of samples available in each category. C1ORF112 expression in the Thyroid showed no significant changes across the categories.

The thyroid showed no significant difference in C1ORF112 expression across the age groups, although the 50-59 and 60-69 categories had the largest sample size, the 60-69 age group did show a possibility of outliers in expression compared to the other age groups. Looking at all the tissues, the testis has the largest Average TPM as seen in Figure 3.2 and again is reflected in Figure 3.4 when comparing the age categories. The average TPM is 10 and this decreases slightly in the 70-79 age group to just below then. This could be because of a decrease in sperm production in the testis in old age, also the sample number in the age group is 9 and may skew the results of the analysis.

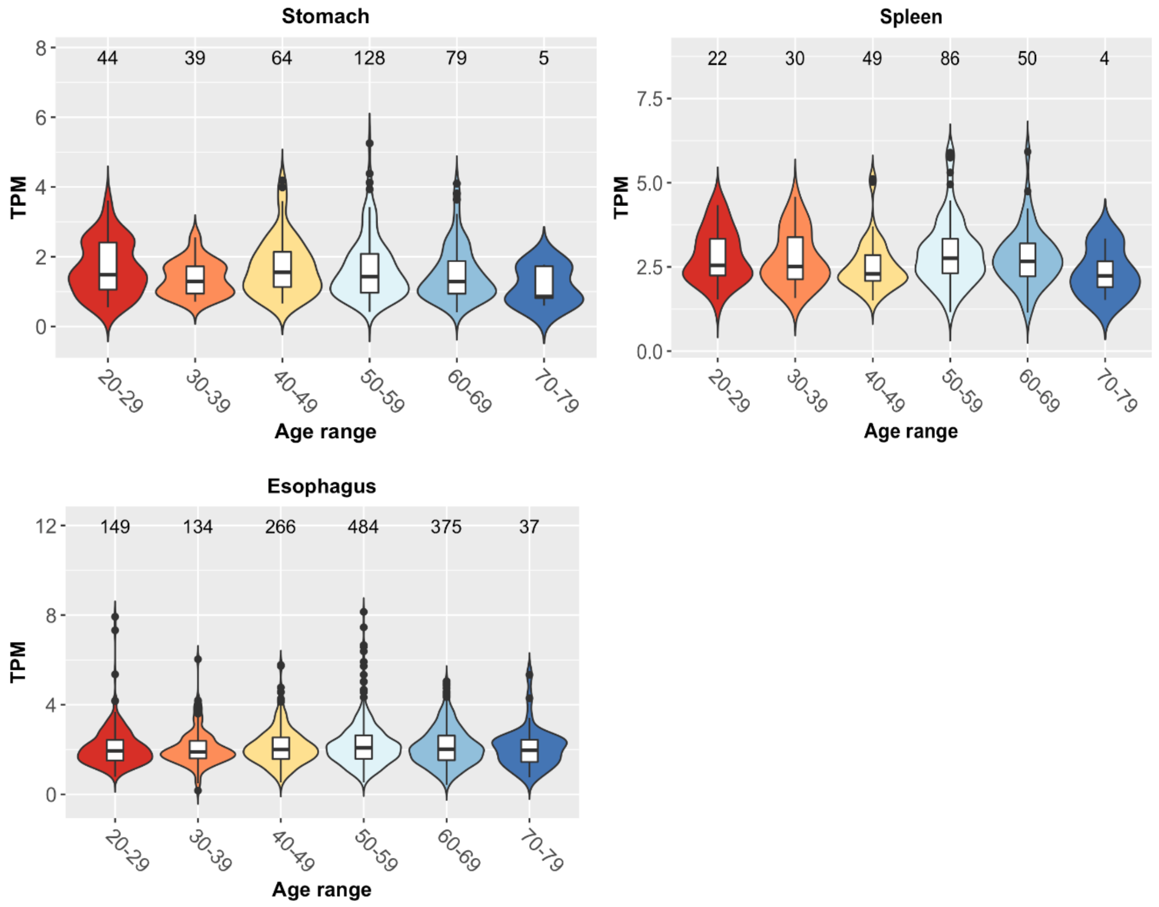


Figure 3. 6 **Age-related expression of C1ORF112 in the Stomach, the Spleen, and the Oesophagus**.

The number at the top of each age category signifies the number of samples available in each category. Expression of C1ORF112 in the stomach and spleen average about 2.5TPM. the oesophagus also showed stable expression levels of C1ORF112 across all time points.

The expression of C1ORF112 in the stomach, the spleen, and the oesophagus is about 2.5TPM and below, the number of samples available for the 70-79 age group for stomach and spleen was and 5 and 4 respectively and might affect the results as their average TPM seemed slightly lower compared to the age groups in both tissues. The oesophagus, on the other hand, had a larger sample size compared to the stomach and the spleen for the 70-79 age group, however, this is less when compared to the other age groups and this reflects the results as the mean TPM is like the other age groups.

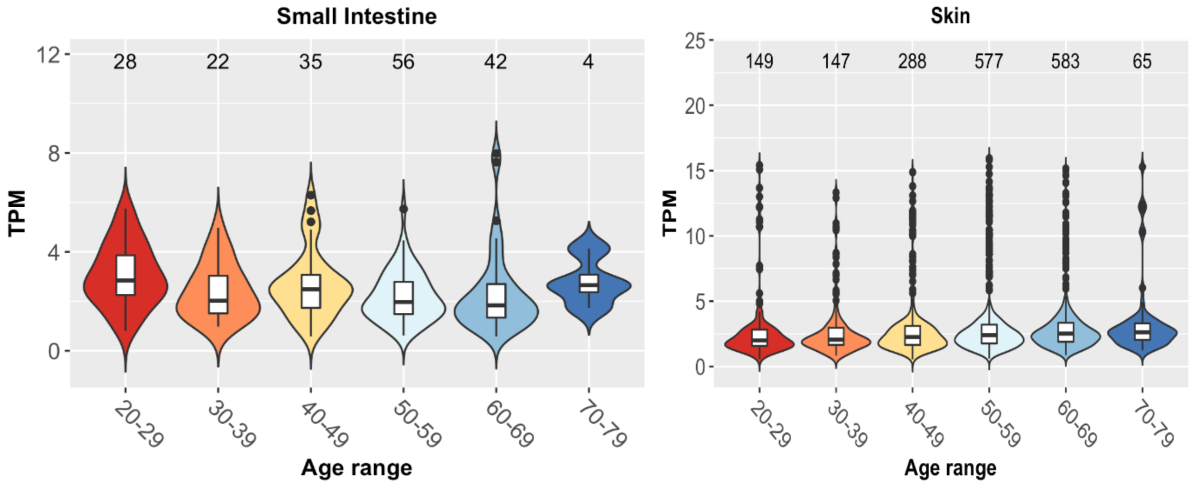


Figure 3. 7 **Age-related expression of C1ORF112 in the Small intestine and the Skin**.

The number at the top of each age category signifies the number of samples available in each category. The expression of C1ORF112 in the small intestine is below 4TPM across all age groups, the skin also had a mean expression below 5TPm, however, several outliers are present across the age groups.

The small intestine had the mean expression across all age groups less than 4TPM, as with other tissues analysed the 70-79 age group had fewer samples available for analysis with only 4 present. The skin had the most samples available of all the tissues analysed and as a result showed more outliers across all age groups, even in the 70-79 age group even though it had the least number of samples available. However, it did not affect the mean TPM expression, which is in line with the other age categories.

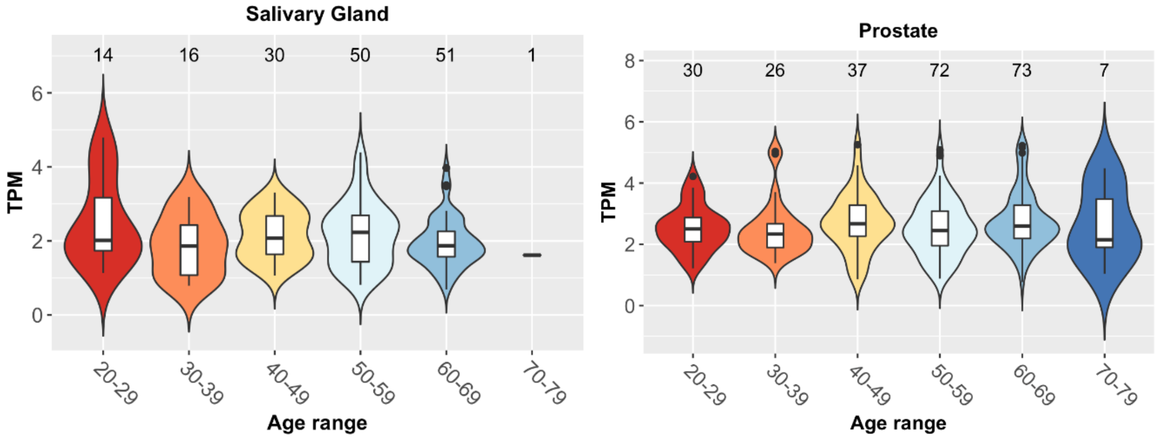


Figure 3. 8 **Age-related expression of C1ORF112 in the Salivary gland and the Prostate**. Salivary gland showed the largest variance in C1ORF112 expression, and the prostate showed similar average levels of expression across all age groups

Despite having the second least sample size the 20-29 age group showed the biggest variance in gene expression of C1ORF112 in the salivary gland, the 70-79 age group had only one sample and so could not be used in the analysis like the uterus (Figure 3.3). The prostrate on the other hand had the biggest variance of expression of C1ORF112 in the 70-79 age group with only 7 samples available for analysis, however, the average TPM is still in line with the other age groups less than 4TPM.

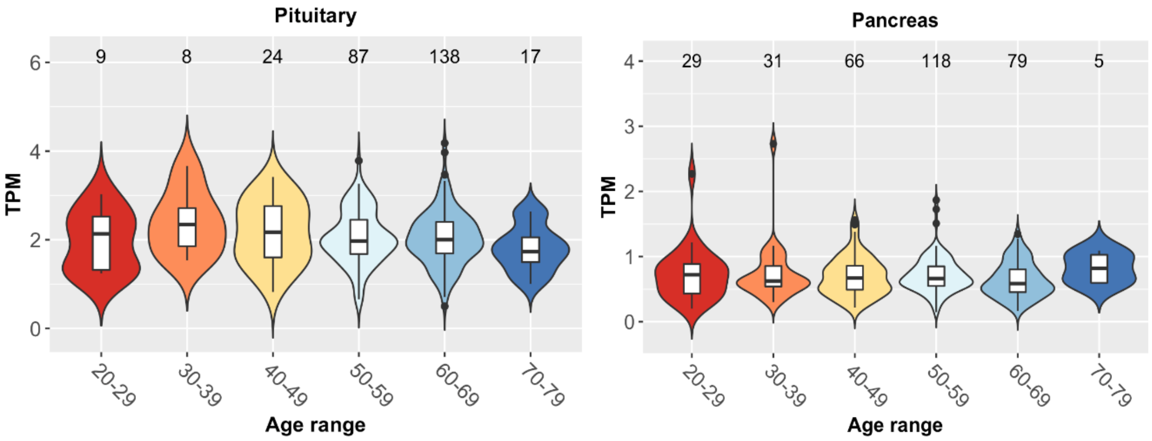


Figure 3. 9 **Age-related expression of C1ORF112 in the Pituitary gland and the Pancreas**.

The number at the top of each age category signifies the number of samples available in each category. The 20-29 and 30-39 age groups had fewer samples compared to the 70-79 age group for analysis in the pituitary, the mean expression stayed similar across all age groups. The pancreas showed the least expression level of C1ORF112, at less than 1TPM compared to other tissues.

The pituitary showed similar levels of expression of C1ORF112 across all the age groups although the 20-29 and the 30-39 age groups had smaller sample sizes compared to the 70-79. The pancreas also had the same mean expression of C1ORF112 across all age groups, the 70-79 age group also had a smaller sample size like the rest of the tissues analysed.

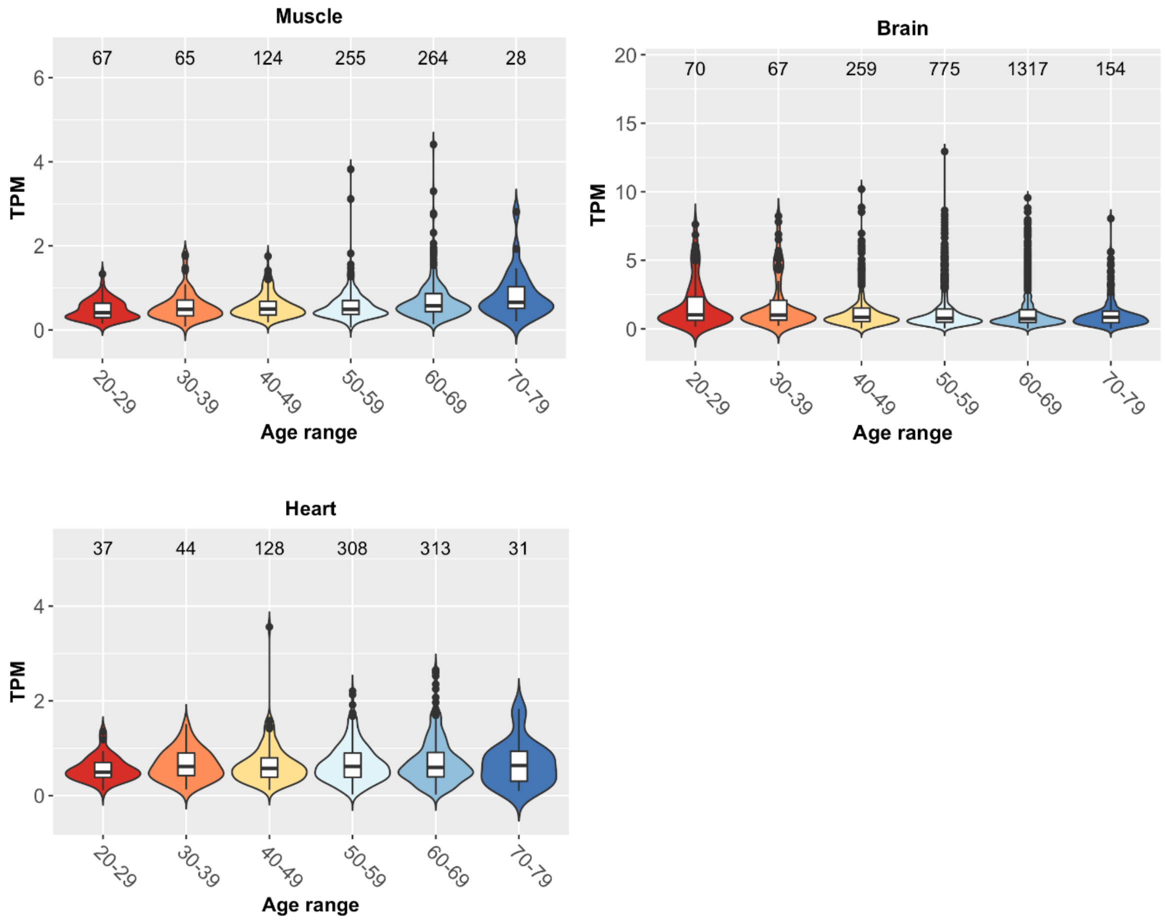


Figure 3. 10 **Age-related expression of C1ORF112 in the Muscle and the Brain and Heart**.

The number at the top of each age category signifies the number of samples available in each category. The muscle, brain and heart showed less than 2TPM expression of C1ORF112 across all age groups with an adequate sample size across each tissue and age group.

These 3 tissues show comparatively lower levels of C1ORF112 expression compared to the other tissues analysed like tissues such as the pancreas (Figure 3.8), it could be as because these tissues do not undergo active cellular division, the pancreas does have cells in the islets of Langerhans that consist of younger cells capable of dividing.

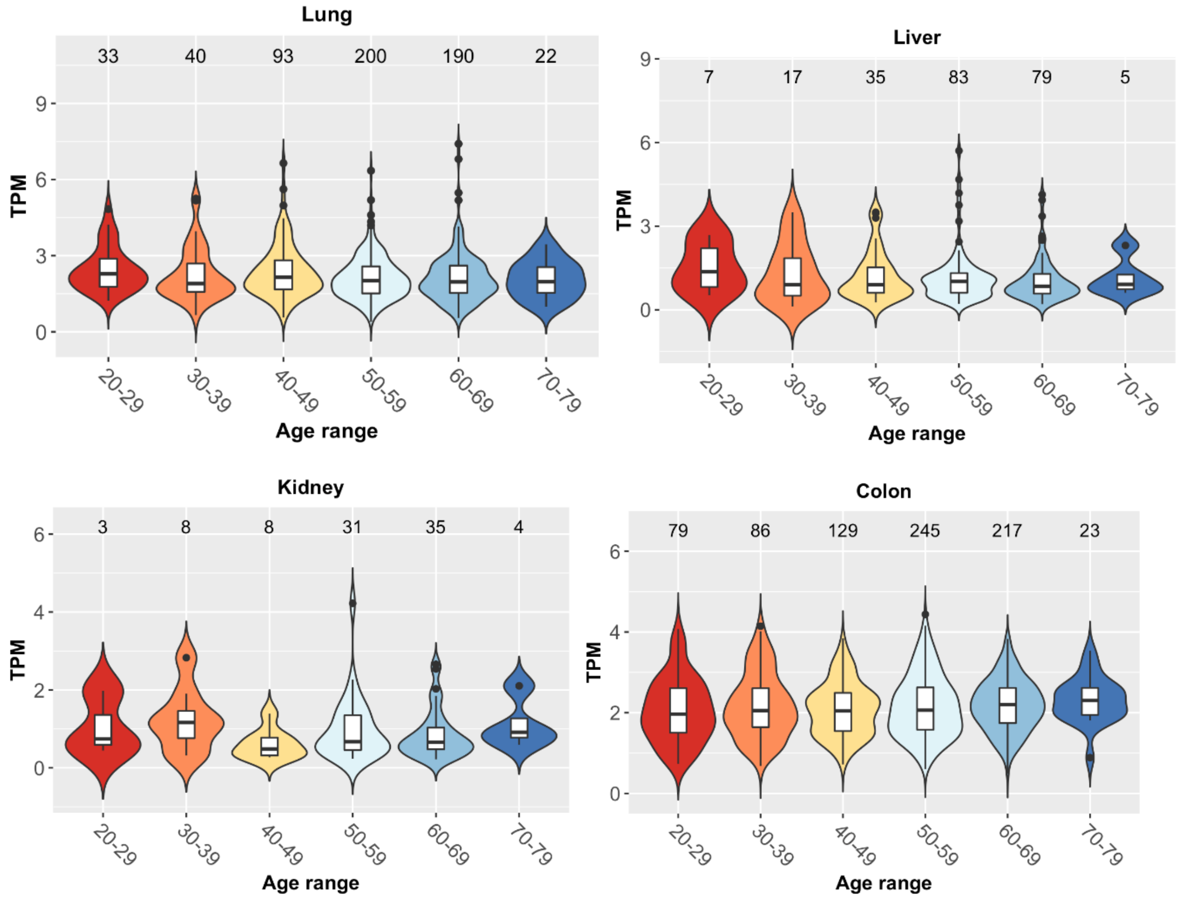


Figure 3. 11 **Age-related expression of C1ORF112 in the Lung and the Liver, the Kidney, and the Colon**.

The number at the top of each age category signifies the number of samples available in each category. The kidney and the liver have fewer samples for analysis in the 70-79 age group, although it does not seem to affect the average TPM expression levels of C1ORF112. The Lung and the Colon have stable expression levels of C1ORf112 across each age group.

C1ORF112 is stable in the lung, liver, kidney, and colon across all age groups. Although the kidney and the liver had fewer samples in the 70-79 age group, this did not affect the mean TPM.

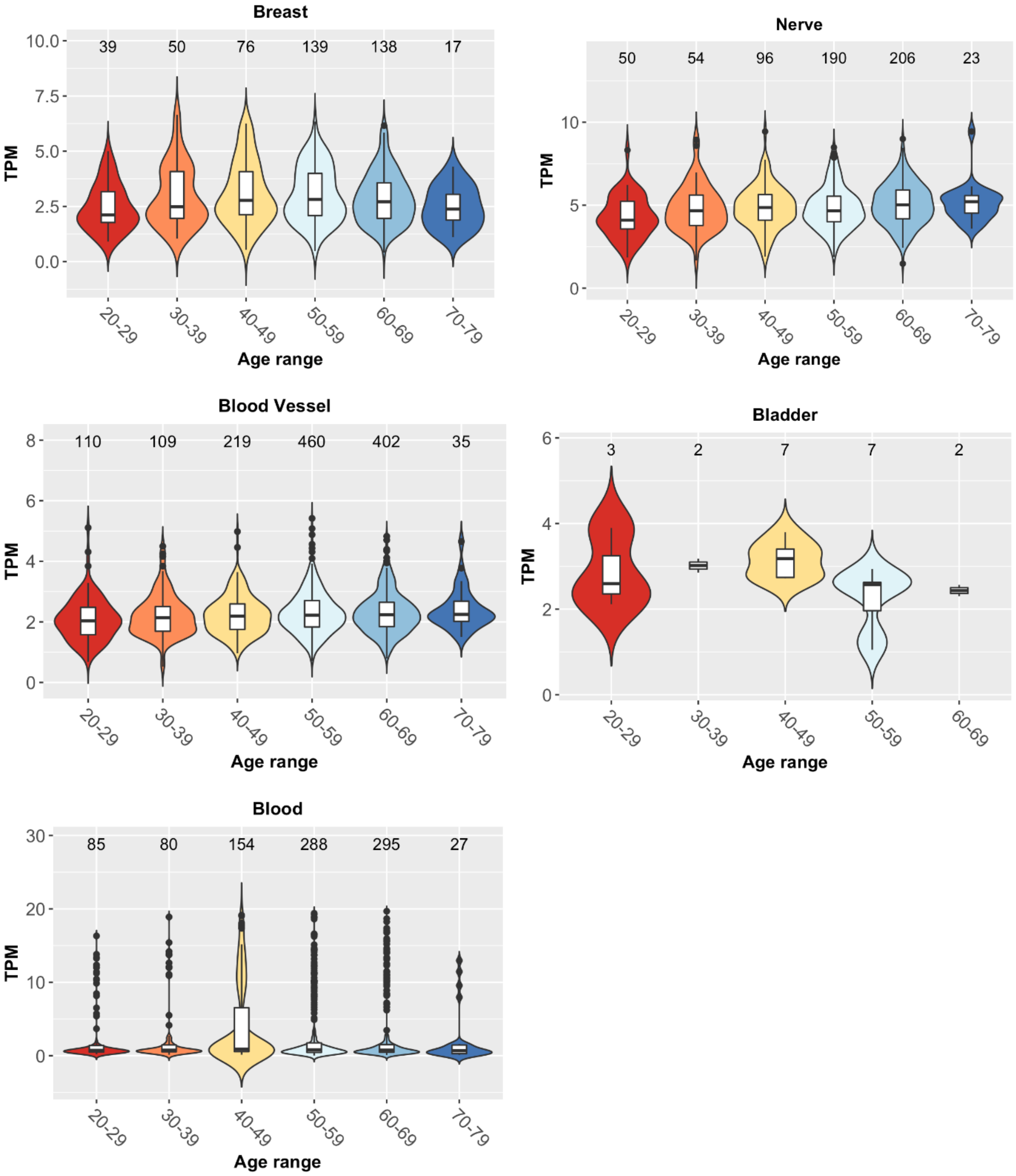


Figure 3. 12 **Age-related expressions of C1ORF112 in the breast, the Nerve, the Blood vessel, the Blood, and Bladder**.

Expression of C1ORF112 was analysed in whole tissues, while overall there is no age-dependent level of expression, there is variability in the expression levels due to the number of samples present for each tissue at different age groups. The Bladder has few samples for analysis and so the results are skewed like other tissues, the expression across the breast, nerve, blood vessels and blood were relatively stable across all age groups.

C1ORF112 expression in blood is quite low compared to the other tissues, the is a lot of outliers as well probably because of the number of samples available, the mean TPM for C1ORF112 in the blood is just above zero, similarly, the blood vessel had outliers in the expression of C1ORF112 across all age groups again possibly because of the number of samples available, identical to the brain and heart (Figure 3.9). The expression of C1ORF112 in breast and nerve is stable across all age categories, though the nerve does show a few outliers in expression.

Generally, there was no difference in expression of C1ORF112 level among the age groups in most tissues with enough samples for robust analysis. In tissues that comprised the female reproduction such as uterus, cervix, and fallopian tube there were not enough samples for analysis, and this could be attributed to the relative importance of the tissues as they are prime reproductive years, especially for the fallopian tube where no samples were present for the 20-29 and 30-39 age groups. The only other tissue to have a low sample size across all age categories was the bladder. The analysis did not consider gender as a factor as the C1ORF112 was not determined to be gender-specific, although, the testis did show higher levels of expression (TPM rate) compared to other tissue.

Furthermore, tissue-specific expression of C1ORF112 was evaluated using tissue specificity analysis as described in Palmer, Fabris et al. (2021). Tissue specificity is an indicator of the level of expression of any gene, either broad expression across all tissues or specifically in certain tissues. τ is the index used to express the specificity of the gene, is calculated using the equation below

$$\tau = \frac{\sum_{i=1}^{N} (1 - x_i)}{N - 1}$$

Where N is the number of tissues studied and x_i is the expression profile component for a given tissue, normalized by the expression value of the tissue in which the gene of interest has the highest expression. Taking that into account, the tissue-specific expression of C1ORF112 was analysed, as C1ORF112 showed a higher average TPM in testis (Figure 3.2, 3.4). As such, C1ORF112 is not expressed in a tissue-specific manner, the tau score for C1ORF112 is 0.68, with a threshold cut-off of 0.8 for tissue-specific genes (Palmer, Fabris et al. 2021).

3.4 Analysis of Genes co-expressed of C1ORF112

After establishing that C1ORF112 was not expressed in a tissue-specific manner, it became important to understand what underlying cellular processes C1ORF112 would be involved with. Initial co-expression analysis was carried out by van Dam, Cordeiro et al. (2012), using GeneFriends. To ascertain a consensus of proteins co-expressed with C1ORF112, Genevestigator was used as described in the

Methods section (2.1.4). Genevestigator is a single cell transcriptomic repository that allows for analysis of curated bulk tissues and single cells. Genevestigators' expansive database allows for more co-expression studies to be carried out in different sample sets. The genes co-expressed with C1ORF112 was analysed across different tissues and cellular conditions, the co-expression was looked at across Anatomical parts (tissues), Cell lines, Neoplasms (Cancers), Perturbations (comparison between different cellular states i.e., diseased vs healthy state). The top 400 genes positively co-expressed, and the top 400 genes negatively co-expressed with C1ORF112 in each condition was determined in each condition was analysed and narrowed down to the top 25 genes. These genes were most frequently co-expressed with C1ORF112 in Anatomical parts, Cancers and Cell lines were then pooled together to determine the genes most commonly co-expressed across all 3 conditions and the top 25 genes were determined, alongside their co-expression scores (see Table 3.2).

Gene Ontology over-representation analysis of genes whose expression correlated positively with C1ORF112 in all categories from Genevestigator was conducted (Hruz, Laule et al. 2008). The size of the dot on the dot plot showed the number of genes overrepresented in each pathway and the p-value cut-off was set at 0.05.

Top 25 Genes positively correlated with the expression of C1ORF112 in Anatomical parts, Cancers, and Cell lines

The top 3 genes positively co-expressed with C1ORF112 are FANCI, NCAPG2 and NUF2 have been reported to be involved with cell cycle progression, DNA repair and chromosome segregation (DeLuca, Moree et al. 2002, Smogorzewska, Matsuoka et al. 2007, Liu, Tanasa et al. 2010). To ascertain the biological processes over-represented by the genes co-expressed with C1ORF112 and gain some functional insight. The genes were enriched using DAVID annotation (Huang da, Sherman et al. 2009) as described in the methods section 2.1.4, this method was robust enough to focus on enriched functional categories and pathways and employing Revigo further filtered out redundant categories. This analysis resulted in several GO terms and the top 20 shown in Table 3.3, showing that the genes co-expressed with C1ORF112 are involved in processes such as cell division (GO:0051301), sister chromatid cohesion (GO:0007062), mitotic nuclear division (GO:0007067).

Table 3. 2 **Top 25 genes positively co-expressed with C1ORF112.**

The genes most frequently co-expressed with C1ORF112 in Anatomical parts, Cancers, and Cell lines

	Score	**Gene**	**Gene name**
1	0.87	FANCI	FA Complementation Group I
2	0.87	NCAPG2	Non-SMC Condensin II Complex Subunit G2
3	0.87	NUF2	NUF2 Component Of NDC80 Kinetochore Complex
4	0.87	WDHD1	WD Repeat And HMG-Box DNA Binding Protein 1
5	0.87	BUB1	BUB1 Mitotic Checkpoint Serine/Threonine Kinase
6	0.87	KIF11	Kinesin Family Member 11
7	0.87	CCNA2	Cyclin A2
8	0.87	SPC25	SPC25 Component Of NDC80 Kinetochore Complex
9	0.86	NCAPG	Non-SMC Condensin I Complex Subunit G
10	0.86	KIF4A	Kinesin Family Member 4A
11	0.86	BIRC5	Baculoviral IAP Repeat Containing 5
12	0.86	AURKB	Aurora Kinase B
13	0.86	RFC3	Replication Factor C Subunit 3
14	0.86	DONSON	DNA Replication Fork Stabilization Factor DONSON
15	0.86	MND1	Meiotic Nuclear Divisions 1
16	0.85	CENPA	Centromere Protein A
17	0.85	HJURP	Holliday Junction Recognition Protein
18	0.85	GINS1	GINS Complex Subunit 1
19	0.85	RFC4	Replication Factor C Subunit 4
20	0.85	UBE2T	Ubiquitin Conjugating Enzyme E2 T
21	0.85	MIS18A	MIS18 Kinetochore Protein A
22	0.85	AURKA	Aurora Kinase A
23	0.85	RFC2	Replication Factor C Subunit 2
24	0.85	KIF18A	Kinesin Family Member 18A
25	0.85	CDC25C	Cell Division Cycle 25C

Table 3. 3 **Gene ontology (GO) terms for biological processes that are over-represented in genes positively co-expressed with C1ORF112.**

The P-values obtained from DAVID, for significantly overrepresented GO terms for biological processes genes co-expressed with C1ORF112 in descending order.

GO Term	Description	Log10 P-Value
GO:0051301	cell division	-59.75
GO:0007062	sister chromatid cohesion	-49.21
GO:0007067	mitotic nuclear division	-48.43
GO:0000777	condensed chromosome kinetochore	-33.05
GO:0000082	G1/S transition of mitotic cell cycle	-29.74
GO:0000776	kinetochore	-25.32
GO:0006270	DNA replication initiation	-23.63
GO:0006281	DNA repair	-23.21
GO:0005524	ATP binding	-19.96
GO:0006271	DNA strand elongation involved in DNA replication	-15.40
GO:0031145	anaphase-promoting complex-dependent catabolic process	-14.61
GO:0000722	telomere maintenance via recombination	-13.50
GO:0000722	telomere maintenance via recombination	-13.50
GO:0007077	mitotic nuclear envelope disassembly	-11.42
GO:0016925	protein sumoylation	-11.33
GO:0051436	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	-10.77
GO:0051437	positive regulation of ubiquitin-protein ligase activity involved in the regulation of mitotic cell cycle transition	-10.32
GO:0051437	positive regulation of ubiquitin-protein ligase activity involved in the regulation of mitotic cell cycle transition	-10.32
GO:0000731	DNA synthesis involved in DNA repair	-10.08
GO:0000732	strand displacement	-8.77

Results from GO enrichment analysis was like the KEGG pathway analysis (Figure 3.12), the top KEGG pathways over-represented from the analysis include cell cycle, DNA replication and Fanconi anaemia pathway. Amyotrophic lateral sclerosis was also top of the list of pathways and how this is related to C1ORF112 is currently unknown.

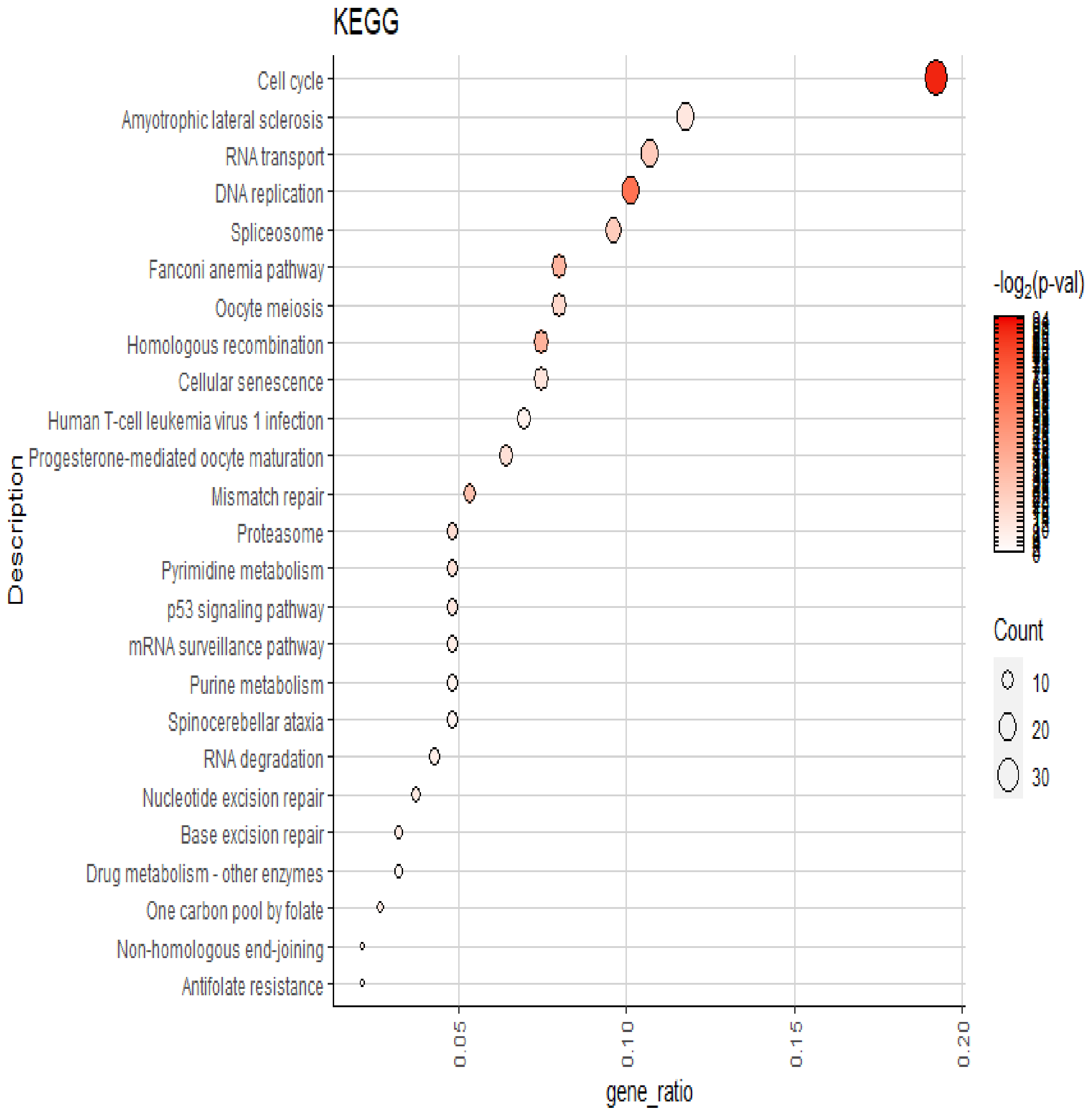


Figure 3. 13 **KEGG pathway over-representation in Anatomical parts, Cancers, and Cell lines**.

Gene Ontology and KEGG pathway analysis for the gene co-expressed with C1ORF112

This analytical process was repeated for genes negatively co-expressed with C1ORF112 and the top 25 genes derived from the analysis are shown in Table 3.4. and CES4A, ADHFE1, and PIK3IP1 are the top genes negatively co-expressed with C1ORF112. CES4A is part of the carboxylesterase large family responsible for the hydrolysis of or transesterification of various xenobiotics, they also are involved in fatty acyl and cholesterol ester metabolism (Hosokawa, Furihata et al. 2007), ADHFE1 is responsible for the oxidation of 4-hydroxybutyrate in mammalian tissues (Kardon, Noel et al. 2006) and PIK3IP1 is predicted to enable phosphatidylinositol 3-kinase catalytic subunit binding activity and involved in negative regulation of phosphatidylinositol 3-kinase (Joshi, Wei et al. 2016).

To better under the enriched GO terms for the gene negatively co-expressed with C1ORF112, the same analysis pipeline was employed results shown in Table 3.5. The most enriched GO terms were plasma membrane (GO:0005886), An integral component of membrane (GO:0016021), and bicarbonate transport (GO:0015701), indicating the genes may be involved in the homeostatic balance of the cell and membrane integrity.

Genes negatively correlated with the expression of C1ORF112 in Anatomical parts, Cancers, and Cell lines

Cell cycle was the top over-represented pathway for gene co-expressed with C1ORF112 in all categories individually and combined. This implies that C1ORF112 would be expressed in replicative cells. Alongside are co-current pathways such as DNA replication and pathways involved in DNA damage and repair. Top GO terms over-represented for the co-expressed genes are ATPase activity, catalytic activity on DNA, DNA dependent activity, Serine/Threonine activity and other activities that either directly impact DNA/RNA or indirectly but are linked to control and maintenance of DNA replication, maintenance of DNA integrity, progression of the cell cycle. When the uniquely co-expressed genes from all categories are obtained, it is very clear as genes such as FANCI, CCNA2, BIRC5, AURKB, CENPA are found to be consistently co-expressed with C1ORF112. This suggests that C1ORF112 transcription occurs co-currently with these genes. However, its function is not clear as co-expression does not necessarily mean functional interaction. But it does increase the likelihood of C1ORF112 playing a role in one of several pathways involved in the cell cycle and DNA replication.

Table 3. 4 **Top 25 genes negatively co-expressed with C1ORF112 in Anatomical parts, Cancers, and Cell lines**

	Score	Gene	Gene name
1	-0.66	CES4A	Carboxylesterase 4A
2	-0.66	ADHFE1	Alcohol Dehydrogenase Iron Containing 1
3	-0.65	PIK3IP1	Phosphoinositide-3-Kinase Interacting Protein 1
4	-0.65	HBB	Haemoglobin Subunit Beta
5	-0.64	GGTA1P	Glycoprotein Alpha-Galactosyltransferase 1
6	-0.63	RNU6-1083P	RNA, U6 Small Nuclear 1083
7	-0.63	CLDN5	Claudin 5
8	-0.63	GRASP	Trafficking Regulator And Scaffold Protein Tamalin
9	-0.63	GPIHBP1	Glycosylphosphatidylinositol Anchored High Density Lipoprotein Binding Protein 1
10	-0.62	RAMP3	Receptor Activity Modifying Protein 3
11	-0.62	VWF	Von Willebrand Factor
12	-0.62	CSF1R	Colony Stimulating Factor 1 Receptor
13	-0.62	HBA2, HBA1	Haemoglobin Subunit Alpha 2/1
14	-0.62	NKAPL	NFKB Activating Protein Like
15	-0.62	TMEM204	Transmembrane Protein 204
16	-0.62	PRORY	PRORY Y-Linked LncRNA
17	-0.62	ADCY4	Adenylate Cyclase 4
18	-0.61	MRVI1	Inositol 1,4,5-Triphosphate Receptor Associated 1
19	-0.61	COX7A1	Cytochrome C Oxidase Subunit 7A1
20	-0.61	CACNA1C	Calcium Voltage-Gated Channel Subunit Alpha1 C
21	-0.61	LGI4	Leucine Rich Repeat LGI Family Member 4
22	-0.61	FMO2	Flavin Containing Dimethylaniline Monoxygenase 2
23	-0.61	MYH11	Myosin Heavy Chain 11
24	-0.61	CLEC14A	C-Type Lectin Domain Containing 14A
25	-0.61	PTGDS	Prostaglandin D2 Synthase

Table 3. 5 **Gene ontology (GO) terms for biological processes that are over-represented in genes negatively co-expressed with C1ORF112.**

The P-values obtained from DAVID, for significantly overrepresented GO terms for biological processes genes negatively co-expressed with C1ORF112 in descending order.

GO Term	Description	Log10 P-Value
GO:0005886	plasma membrane	-6.82
GO:0016021	An integral component of membrane	-5.39
GO:0015701	bicarbonate transport	-3.19
GO:0031012	extracellular matrix	-3.17
GO:0031720	haptoglobin binding	-3.17
GO:0031838	haptoglobin-haemoglobin complex	-2.86
GO:0072562	blood microparticle	-2.62
GO:0007190	activation of adenylate cyclase activity	-2.44
GO:0042542	response to hydrogen peroxide	-2.06
GO:0043202	lysosomal lumen	-2.01
GO:0015254	glycerol channel activity	-1.86
GO:0005833	haemoglobin complex	-1.86
GO:0016056	rhodopsin mediated signalling pathway	-1.82
GO:0007601	visual perception	-1.82
GO:0009992	cellular water homeostasis	-1.76
GO:0015793	glycerol transport	-1.76
GO:0036159	inner dynein arm assembly	-1.76
GO:0005344	oxygen transporter activity	-1.73
GO:0015671	oxygen transport	-1.64
GO:0015250	water channel activity	-1.62

3.5 Possible physical interaction with C1ORF112

Possible protein-protein interactors for C1ORF112 were extracted from BioGRID and Gene mania. The results from the BioGRID data are from high throughput affinity mass spectrophotometry (Hein, Hubner et al. 2015) or yeast hybrid system (Fernandes, Duhamel et al. 2018) to determine the possibility of physical interaction. Based on this available information C1ORF112 has 8 direct physical interactors (Figure 4.10). When cross-referenced with co-expressed genes FIGNL1, and SPATA5L1 are the 2 genes that positively co-expressed with C1ORF112 and whose protein products interact physically.

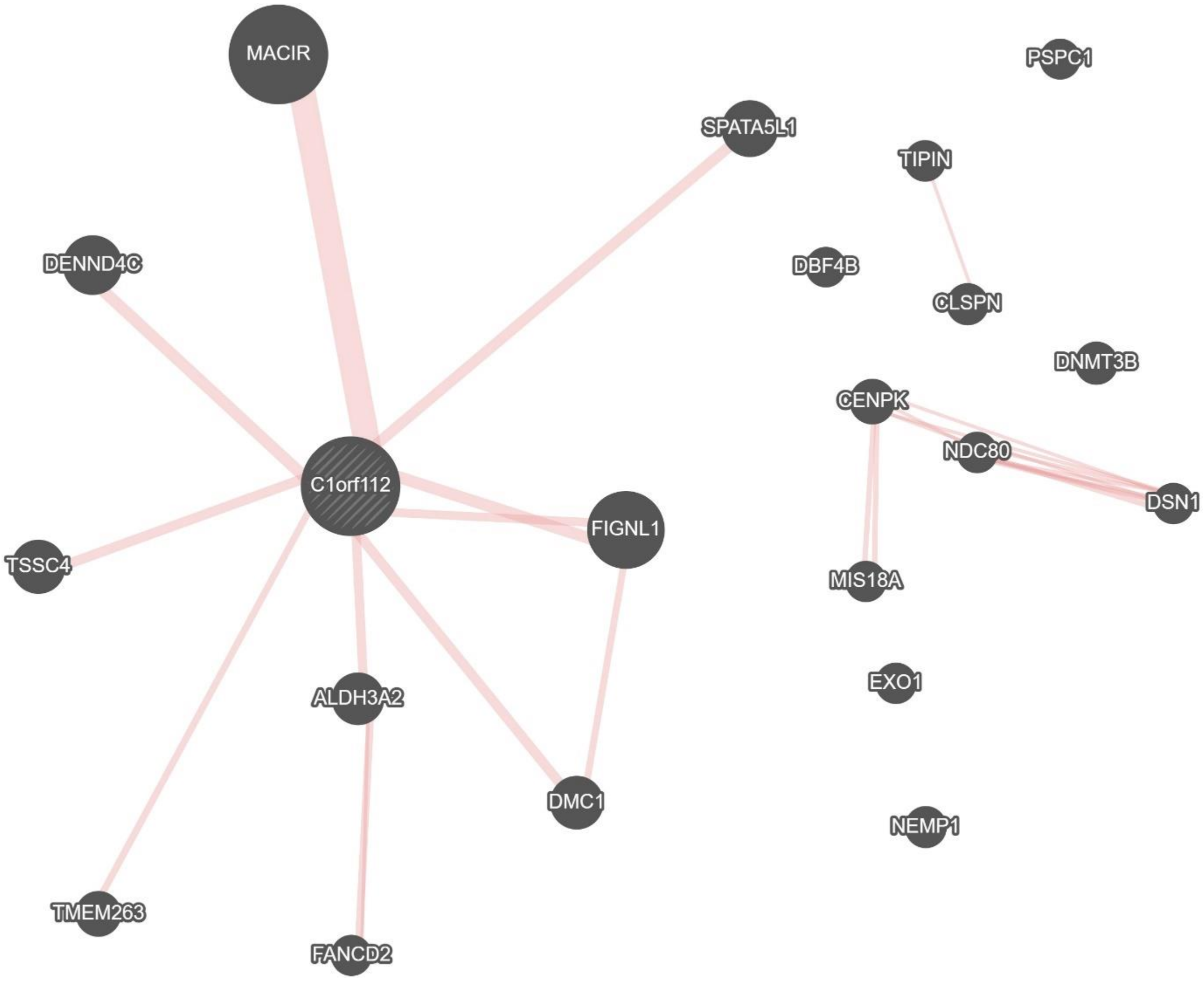


Figure 3. 14 **Possible physical interactors with C1ORF112**.

Proteins that have been highlighted to have possible physical interactions with C1ORF112 based on affinity capture of yeast two-hybrid system. Protein on the right do not have and evidence of direct physical interaction with C1ORF112, but the pink lines indicate interaction with each other.

3.6 Discussion

C1ORF112 is a highly conserved gene from mammals to choanoflagellates. It is also present in insects and plants. The sequence similarity shows high sequence conservation in vertebrates, especially in mammals, reptiles, and birds. This shows that C1ORF112 serves a relatively important function or different functions. It is not quite clear yet as to the role or roles C1ORF112 performs within the cell. Although C1ORF112 orthologues are present in many metazoans, it appears to be absent in *Caenorhabditis elegans*, and *Drosophila melanogaster*, however, in Drosophila it appears there is a sequence present, but it lacks the canonical WCF tripeptide coupled with low sequence similarity of about approx. <12% so it's not quite clear if it is a true homologue. The reason C1ORF112 is absent in *C. elegans*, and *Drosophila* is not particularly clear since, it is present in other worms particularly Platyhelminthes, worms with lifecycles in several vertebrates. This is also consistent with a recently published article by Fernandes, Duhamel et al. (2018), where they described C1ORF112 as FLIP as a historically conserved protein alongside FIGNL1 and DMC1 which are both conserved in *Caenorhabditis elegans*, and *Drosophila melanogaster,* interestingly, they also did not identify any FLIP homologues in Fungi. FIGNL1 is a member of the AAA ATPase family recruited to sites of DNA damage playing a role in DSB via the HR through the binding interaction with RAD51 (Yuan and Chen 2013). FIGNL1 is recruited to sites of DSBs in a BRCA2 independent manner and its depletion does not affect RAD51 loading unto ssDNA (Yuan and Chen 2013). Fernandes, Duhamel et al. (2018) also suggested that C1ORF112 is homologous to a protein in Arabidopsis called MEICA which has been indicated to play a role in chromosome crossovers in homologous recombination (Hu, Li et al. 2017). Considering that, C1ORF112 appears to be functionally important in the cell and possibly a historically conserved protein as well and is co-expressed with FA related genes such as FANCI and FAND2.

C1ORF112 aside from being well conserved is also co-expressed with genes that are responsible for the control and maintenance of cell cycle, DNA replication and cell replication. With a co-expression score of approx. 70% across the different categories in Genevestigator, the number of unique co-expressed genes include but are not limited to MCM10, MAD2L1, CENPA, AURKB, BUB1, POLE2, and CDC6 Kinetochore complexing proteins. These genes are involved in the control of chromosomal segregation and the progression of the cell through cellular division. For example, MCM10 is part of the mini-chromosome maintenance family of proteins, which are highly conserved and are involved in the initiation of genome replication (Miotto, Chibi et al. 2014). POLE2, DNA Polymerase Epsilon 2 participates in DNA repair and in DNA replication through its direct interaction with DNA and dimerization with POLE1 (Li, Pursell et al. 2000). CENPA, Centromere protein A and AURKB, Aurora kinase B are involved in kinetochore assembly, chromosome alignment and segregation (Sullivan, Hechenberger et al. 1994, Thoresen, Campsteijn et al. 2014). BUB1 is a serine/threonine-protein kinase mitotic checkpoint protein bound to kinetochores and is involved in spindle fibre assembly checkpoint (Roberts, Farr et al. 1994). The co-expression of these genes indicate that C1ORF112 may have a role in DNA replication, chromosome stability and accurate segregation.

However, C1ORF112 is not overtly expressed in all cells or tissues, looking at the relative expression of C1ORF112 in tissues (Figure 3.2). Aside from the testis, and Epstein-Barr virus-transformed lymphocytes cells, which have relatively higher expression levels when compared to the other tissues, this however does not make it tissues specific. In addition, age-related expression of C1ORF112 is relatively stable across tissues as well (Figure 3.4 -3.11), that is, the expression of C1ORF112 does not change dramatically with the increase in age. As stated earlier while there were disparities in sample size due to the availability of some samples in tissues, such as those comprising of the female reproduction such as the uterus, cervix, fallopian tube, and other tissues in the older age category as well. In addition, C1ORF112 has been suggested to be a potential biomarker for several tumours by Chen, Mai et al. (2021). In the correlation study of C1ORF112 expression and patient survival across several tumours, lower expression of C1ORF112 was shown to increase the likelihood of better survival for several tumours including but not limited to bladder urothelial carcinoma(BLCA), breast invasive carcinoma (BRCA), cholangiocarcinoma(CHOL), colon adenocarcinoma (COAD), oesophageal carcinoma(ESCA), glioblastoma multiforme (GBM), HNSC, kidney renal papillary cell carcinoma (KIRP), LIHC, lung adenocarcinoma(LUAD), lung squamous cell carcinoma (LUSC), rectum adenocarcinoma (READ) (Chen, Mai et al. 2021). This also follows along with the work of Zhang, Tan et al. (2021), also suggesting that the presence of C1ORF112 predict poor outcomes in patients with low-grade glioma. In their study, high expression of C1ORF112 positively correlated with immune cells such as B cells, CD8+T cells, CD4+T cells, macrophages, neutrophils, and dendritic cells infiltrating low-grade gliomas and was an independent factor in overall survival. Nonetheless, analysis of available samples does show that expression of C1ORF112 does not change over time in normal cells. Its expression is stable and consistent across all tissues over time.

Gene set enrichment analysis of gene co-expressed with C1ORF112 with showed that top GO terms over-represented for the co-expressed genes that possess ATPase activity, catalytic activity on DNA, DNA dependent activity, Serine/Threonine activity and other activities and are involved in the mitotic nuclear division, condensed chromosome kinetochore, G1/S transition of the mitotic cell cycle, DNA replication initiation (Tables 3.3 & 3.5). In addition, cell cycle RNA transport, Fanconi anaemia, homologous recombination and cellular senescence are top pathways overrepresented in all categories. C1ORF112 also has 7 possible physical interactors (Figure 4.10), including FIGL1 and DMC1 stated to be associated via tandem affinity purification coupled to mass spectrometry (TAP-MS) using overexpressed FIGL1 as a bait (Fernandes, Duhamel et al. 2018). Finally, results show that C1ORF112 is co-expressed with genes involved with DNA replication, kinetochore assembly, cell cycle progression, and cell replication.

Chapter 4: C1ORF112 is an alpha-helical protein with a possible kinase domain

4.1 Introduction

Functional analysis of proteins is typically linked to the cumulative domain architecture present within its entire sequence. Each domain typically has a specific function and to decipher the function of a protein it is necessary first to determine its domains and characterize their functions (Galzitskaya and Melnik 2003). Since domains are recurring patterns, assigning a function to a domain family can shed light on the function of the many proteins that contain this domain, which makes the task of automated function prediction feasible. Considering the massive protein sequence data that is generated, this is an important goal (Ingolfsson and Yona 2008). A protein domain hypothesis domain can be defined as:

- A domain is a protein unit that can fold independently.
- a specific cluster in three-dimensional (3D) space
- It performs a specific task/function.
- It is a movable unit that was formed early during evolution

Regardless of the definition knowledge of the protein domain and the binding site would yield more insight into the mechanism by which protein carry out their functions *in vitro*. To predict possible domains with protein, DNA, or ligand binding regions on the sequence of C1ORF112, the knowledge-based approach was used. Based on the functional annotation of highly similar genes an *ab-initio* model structure was generated.

There are various mechanisms for protein modelling, from template-based modelling to *ab initio* / computational modelling (Ingolfsson and Yona 2008). I-TASSER generates an *ab initio* by stringing fragments of Multiple Sequence Alignment (MSA) of proteins of similar structure to generate a model without the use of a template described in the method section 2.2.1. *ab initio* protein modelling was used in this instance as there were no available functional homologs to C1ORF112 to be used as templates.

4.2 Alpha helical model generation using I-TASSER

After the models were generated, they were analysed using the Ramachandran plot to determine how well the residues in the models fitted the phi (ϕ)and psi (ψ) bonds torsion angles in the models.

to understand the Ramachandran plot, see figure 4.1, which highlights the favoured regions for α-helices and β-sheets and their accompanying favoured regions.

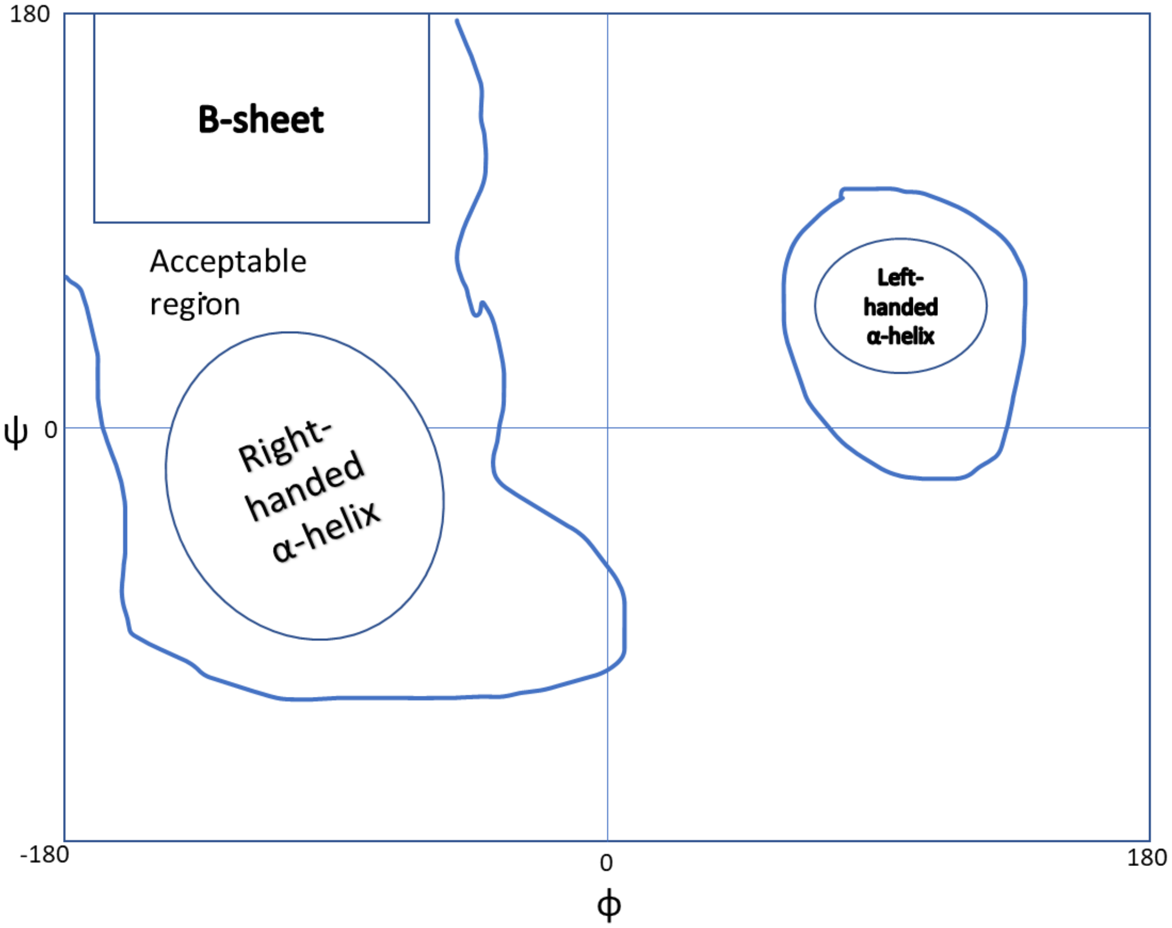


Figure 4. 1 **Reference Ramachandran plot**.

Oval shapes are the α-helical acceptable regions, and the box shape is the favoured region for β-sheets. The blue outline is the acceptable regions residues to account for misalignment, conformational changes, or poor modelling.

Model 1 of C1ORF112 generated by I-TASSER

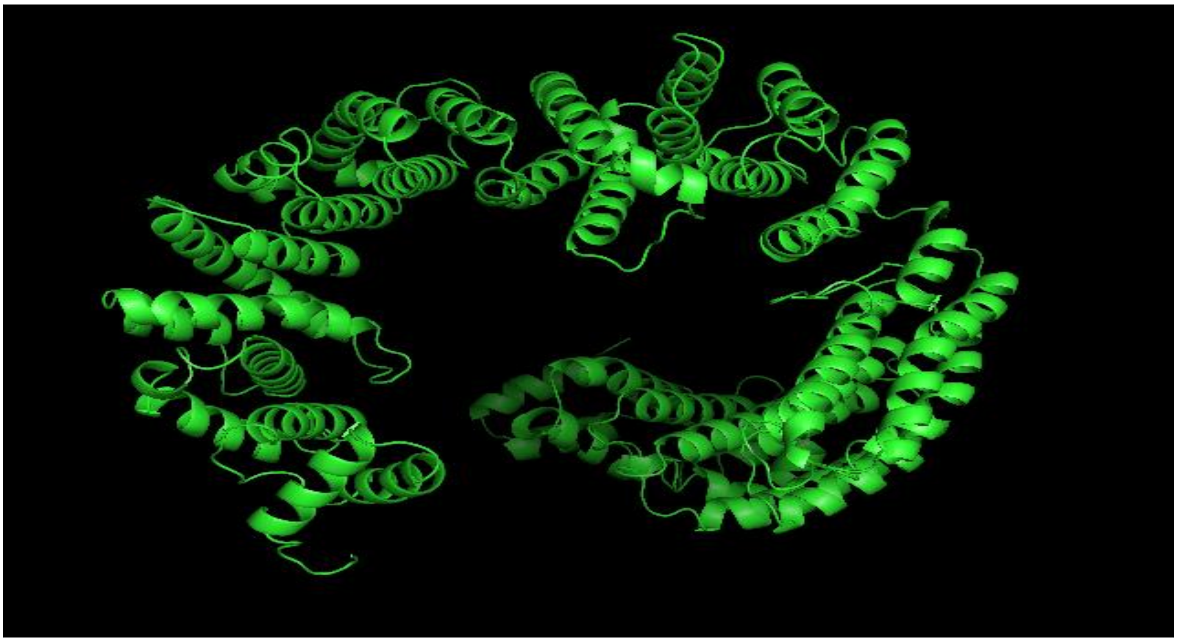


Figure 4. 2 **Secondary structure of Model 1 of C1ORF112 generated using I-TASSER**.

Model 1 generated by I-TASSER reveals an alpha-helical superstructure.

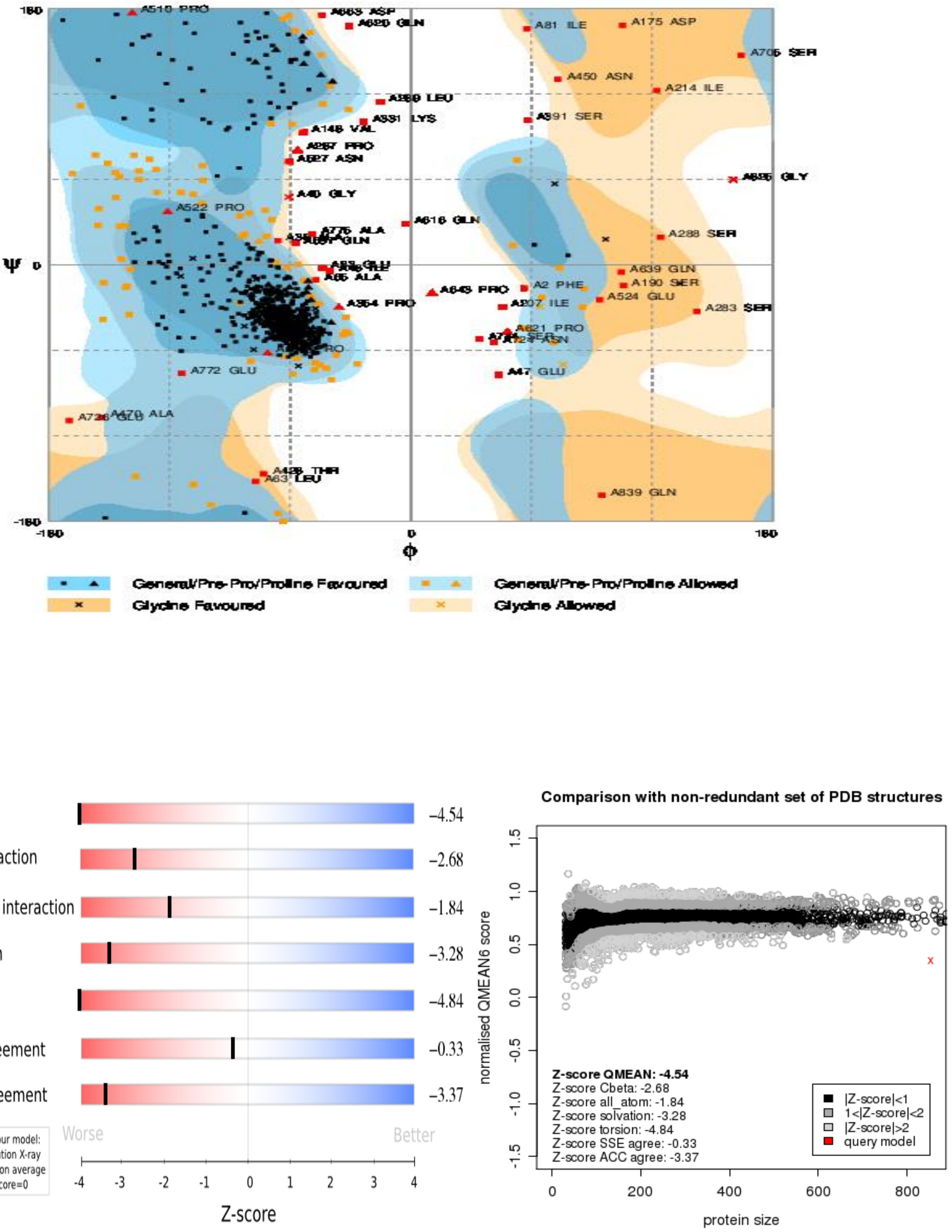


Figure 4. 3 **Model 1 analysis using Ramachandran plot and QMEAN analysis**.

Ramachandran plot of model 1 shows how well residue conforms to the expected regions. The blue area is the region generally favoured by alpha-helices and the brown region is generally favoured by beta-sheets. The normalised QMEAN score (red indicating query model) shows the model to be out of range with similar PDB structures, also presenting low z-scores with an overall z score of -4.54.

Number of residues in favoured region (~98.0% expected): 719 (84.5%)

Number of residues in allowed region (~2.0% expected): 88 (10.3%)

Number of residues in outlier region: 44 (5.2%)

The number of residues in the favoured region is 84.5%, 719 residues out of 853, 88 residues in the allowed region which accounts for 10.3%, and 44 residues in the outlier region making the final 5.2%

Table 4. 1 **QMEAN analysis of Model 1**.

Scoring function term	Raw score	Z-score
C-beta interaction energy	75.29	-2.68
All-atom pairwise energy	-9733.34	-1.84
Solvation energy	8.84	-3.28
Torsion angle energy	28.47	-4.84
Secondary structure agreement	77.1%	-0.33
Solvent accessibility agreement	62.6%	-3.37
QMEAN6 score	0.352	-4.54

__Model 2 of C1ORF112 generated by I-TASSER__

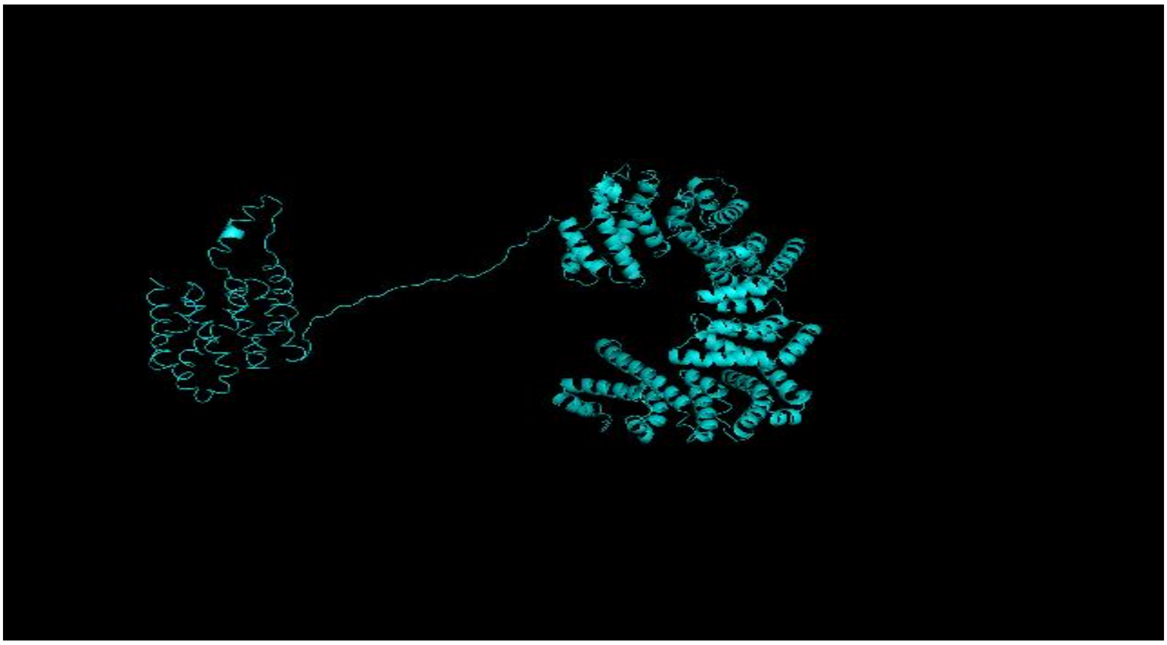


Figure 4. 4 **Secondary structure of Model 2 of C1ORF112 generated using I-TASSER.**

Starting from the left, model 2 is a more disordered helical shape, a few linking sequences and then shows a more compact alpha-helical structure on the right.

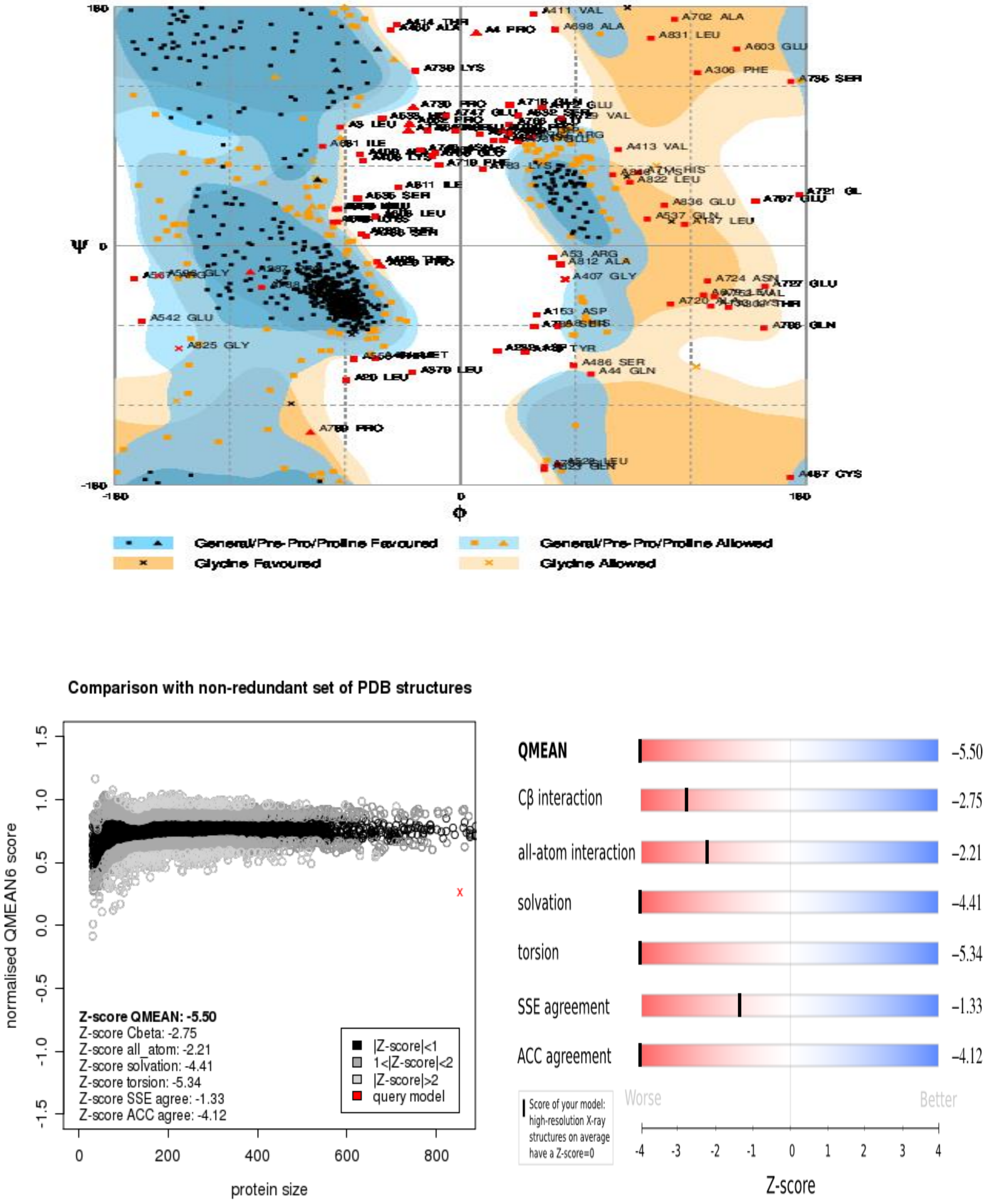


Figure 4. 5 **Model 2 analysis using Ramachandran plot and QMEAN analysis**.

Ramachandran plot of Model 2 shows more residues present in the disordered region than acceptable for alpha-helices. The normalised QMEAN score (red indicating query model) shows the model to be out of range with similar PDB structures, also presenting low z-scores with an overall z score of -5.50.

Number of residues in the favoured region (~98.0% expected): 601 (70.6%)

Number of residues in allowed region (~2.0% expected): 158 (18.6%)

Number of residues in outlier region: 92 (10.8%)

Table 4. 2 **QMEAN analysis of Model 2.**

Scoring function term	**Raw score**	**Z-score**
C-beta interaction energy	77.12	-2.75
All-atom pairwise energy	-6892.45	-2.21
Solvation energy	38.23	-4.41
Torsion angle energy	55.35	-5.34
Secondary structure agreement	72.1%	-1.33
Solvent accessibility agreement	58.5%	-4.12
QMEAN6 score	0.266	-5.50

Model 3 of C1ORF112 generated by I-TASSER

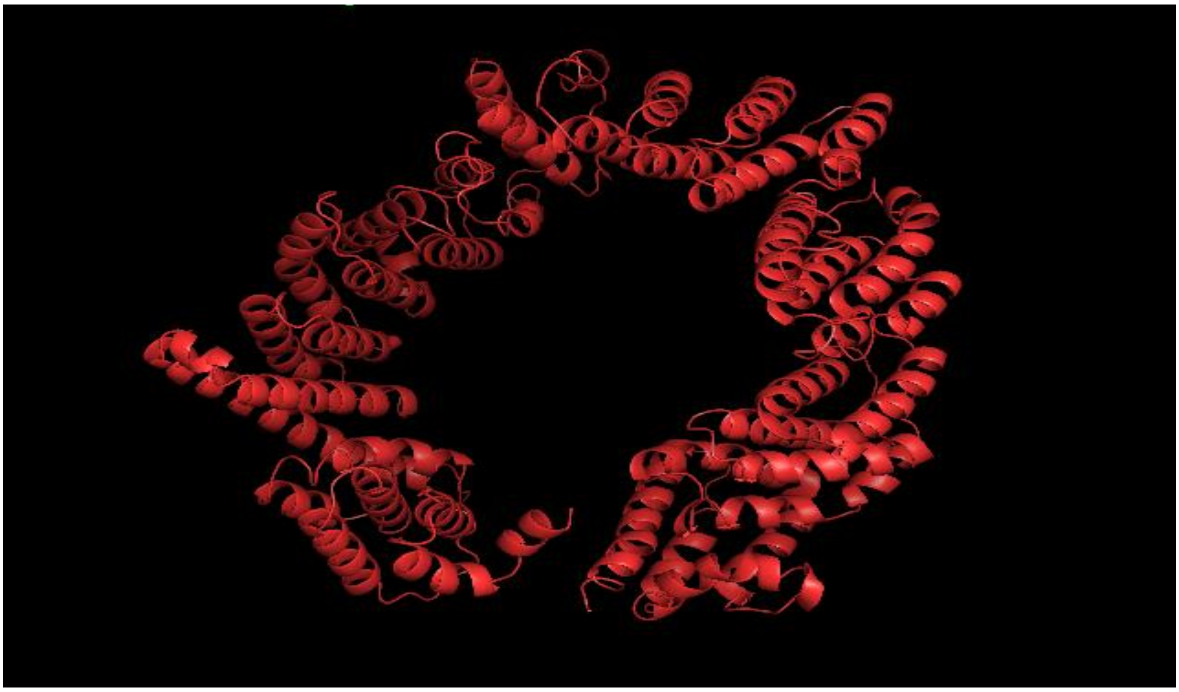


Figure 4. 6 **Secondary structure of Model 3 of C1ORF112 generated using I-TASSER.**

Model 3 is like Model 1 in terms of helical organisation and full helical assembly through the entire sequence.

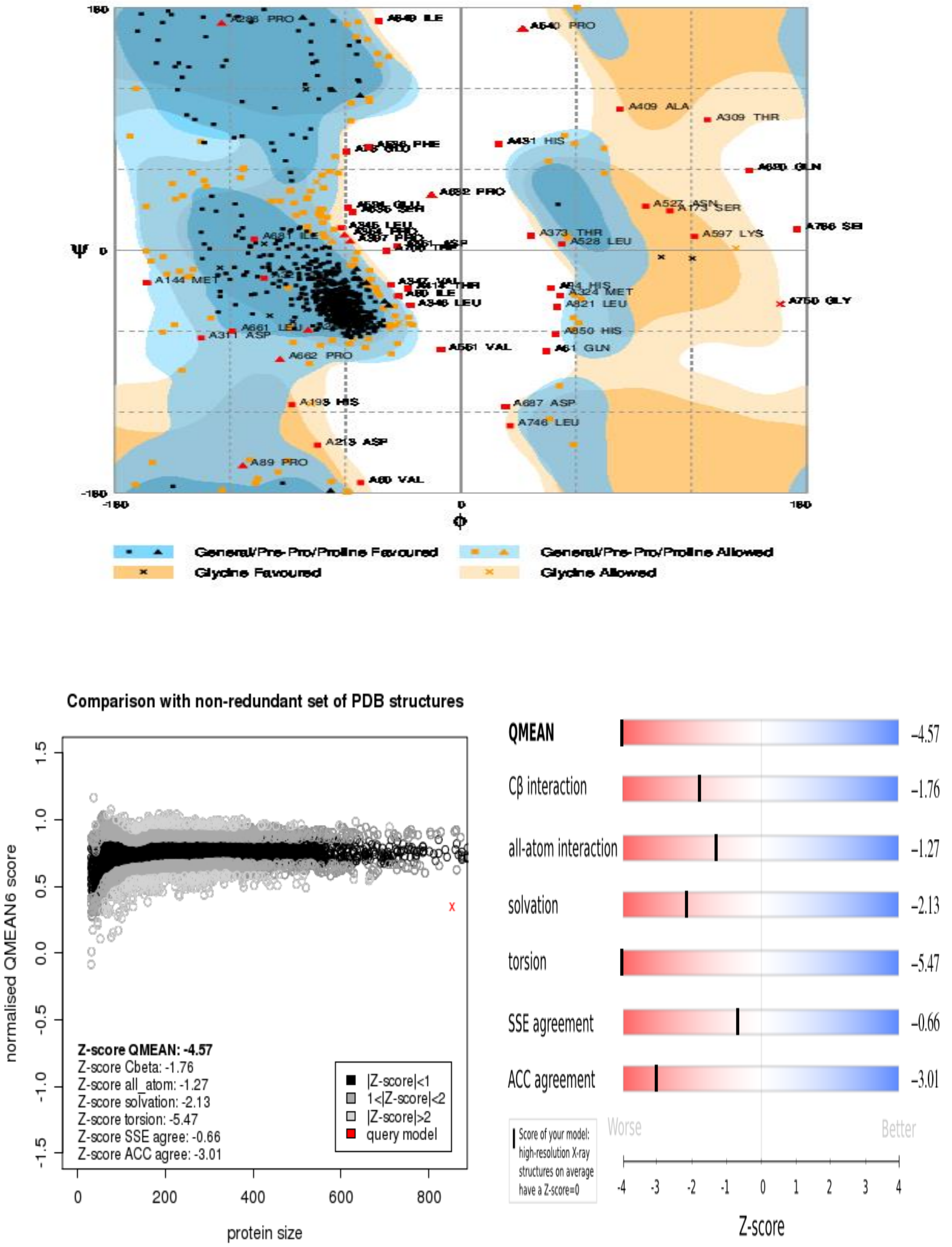


Figure 4. 7 **Model 3 analysis using Ramachandran plot and QMEAN analysis.**

Ramachandran plot of model 3 shows more of the amino acid residues within the expected regions. The blue area is the region generally favoured by alpha-helices and the brown region is generally favoured by beta-sheets. The normalised QMEAN score (red indicating query model) shows the model to be out of range with similar PDB structures, also presenting low z-scores with an overall z score of -4.57

Number of residues in the favoured region (~98.0% expected): 678 (79.7%)

Number of residues in allowed region (~2.0% expected): 126 (14.8%)

Number of residues in outlier region: 47 (5.5%)

Table 4. 3 **QMEAN analysis of Model 3.**

Scoring function term	Raw score	Z-score
C-beta interaction energy	-46.94	-1.76
All-atom pairwise energy	-13147.44	-1.27
Solvation energy	-21.22	-2.13
Torsion angle energy	62.93	-5.47
Secondary structure agreement	75.5%	-0.66
Solvent accessibility agreement	64.6%	-3.01
QMEAN6 score	0.350	-4.57

Model 4 of C1ORF112 generated by I-TASSER

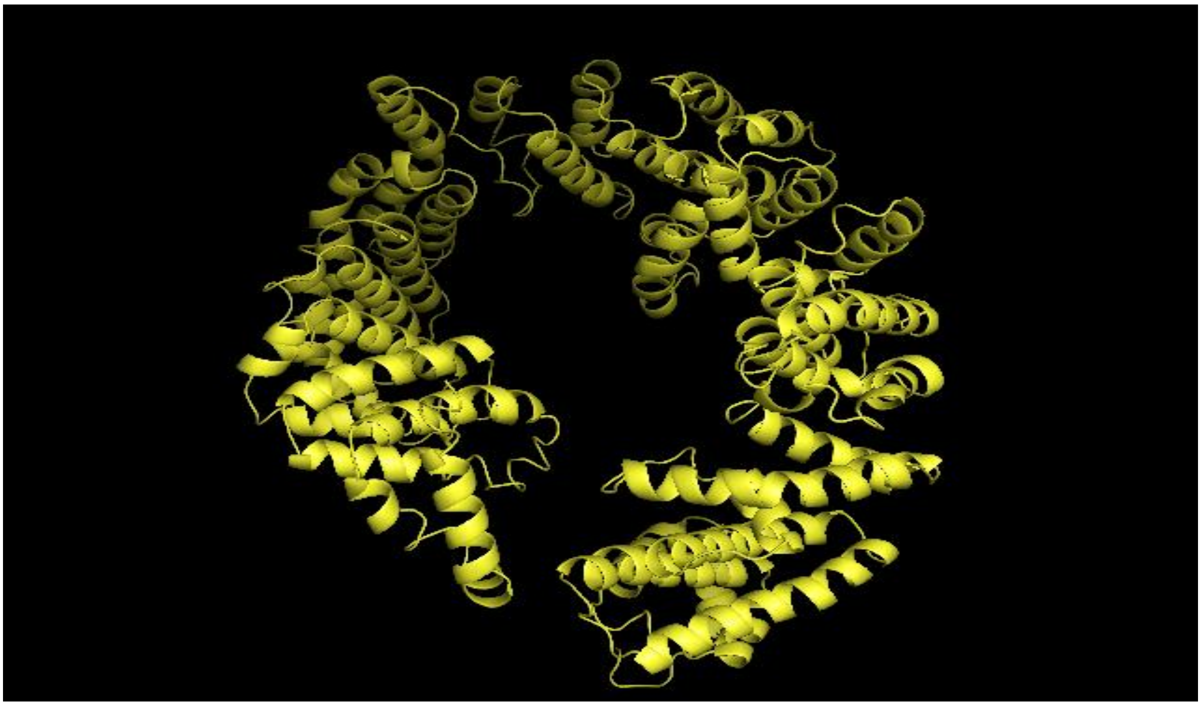


Figure 4. 8 **Secondary structure of Model 4 of C1ORF112 generated using I-TASSER.**

Model 4 is like Model 1 in terms of helical organisation and full helical assembly through the entire sequence

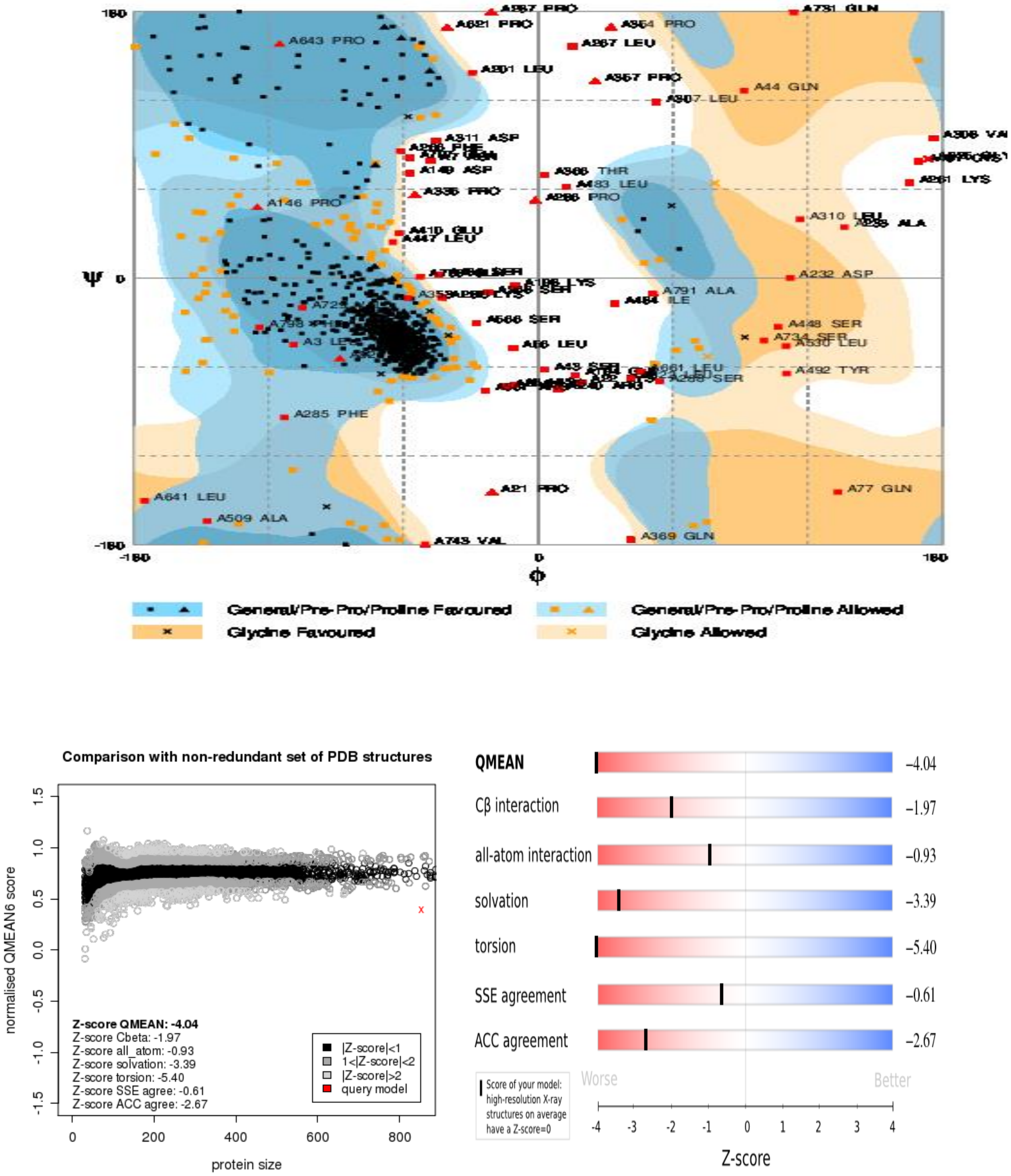


Figure 4. 9 **Model 4 analysis using Ramachandran plot and QMEAN analysis.**

Ramachandran plot of model 4 shows more of the amino acid residues within the expected regions like models 1 and model 3. The blue area is the region generally favoured by alpha-helices and the brown region is generally favoured by beta-sheets. The normalised QMEAN score (red indicating query model) shows the model to be out of range with similar PDB structures, also presenting low z-scores with an overall z score of -4.04

Number of residues in favoured region (~98.0% expected): 681 (80.0%)

Number of residues in allowed region (~2.0% expected): 106 (12.5%)

Number of residues in outlier region: 64 (7.5%)

Table 4. 4 **QMEAN analysis of Model 4.**

Scoring function term	Raw score	Z-score
C-beta interaction energy	-18.38	-1.97
All-atom pairwise energy	-14180.44	-0.93
Solvation energy	11.76	-3.39
Torsion angle energy	58.75	-5.40
Secondary structure agreement	75.7%	-0.61
Solvent accessibility agreement	66.5%	-2.67
QMEAN6 score	0.398	-4.04

Model 5 of C1ORF112 generated by I-TASSER

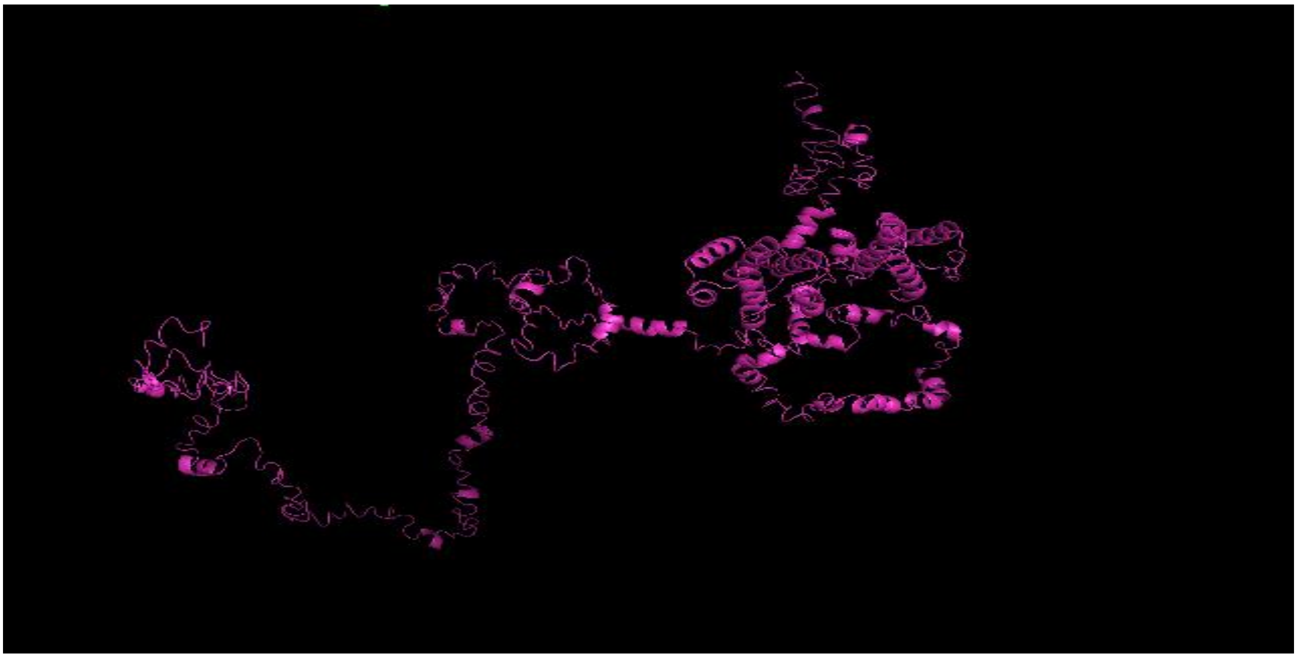
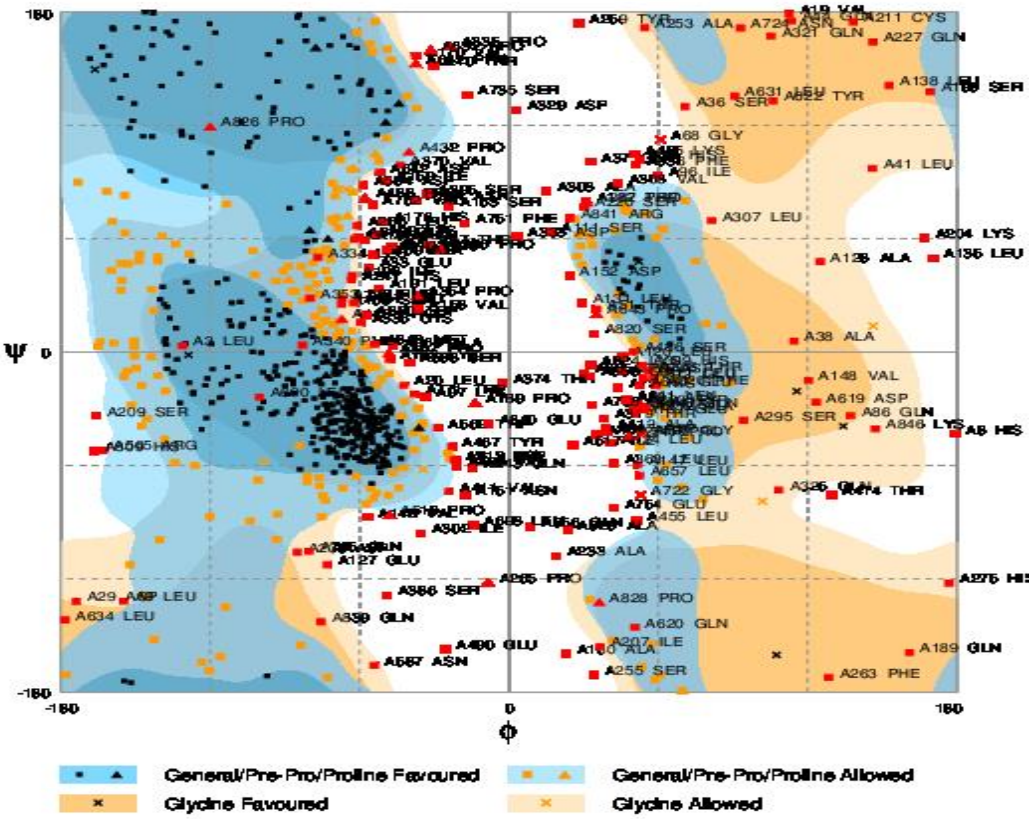


Figure 4. 10 **Secondary structure of Model 5 of C1ORF112 generated using I-TASSER.**

Model 5 presents as highly disordered with incomplete helical structures.



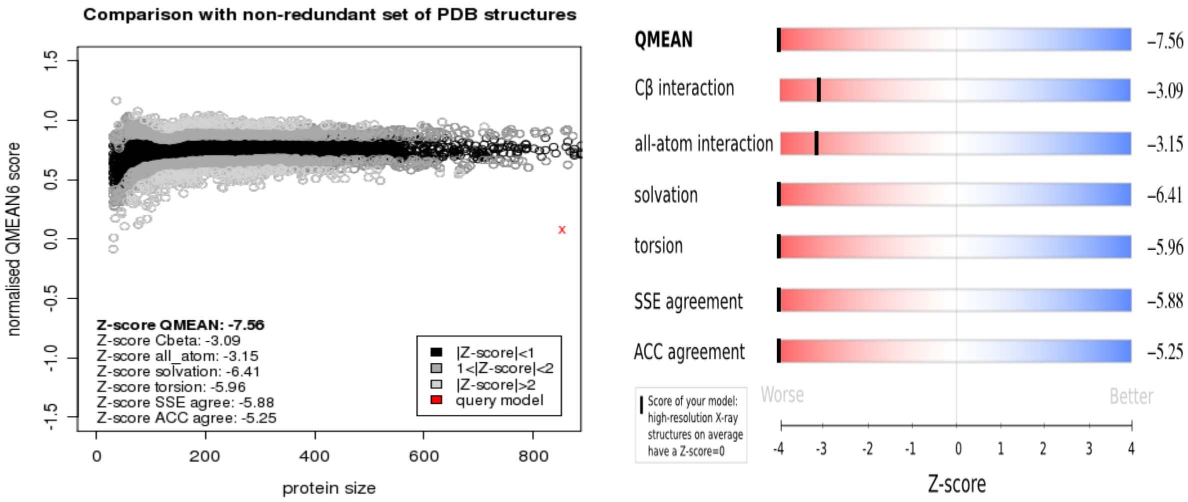


Figure 4. 11 **Model 5, Ramachandran plot and QMEAN analysis.**

Ramachandran's analysis of model 5 shows a highly disordered protein with only half the amino acid residues in the alpha-helical region. Normalized QMEAN analysis shows the query model is closer to 0 and farther from resolved PDB structures, and a QMEAN z-score of -7.56.

Number of residues in the favoured region (~98.0% expected): 458 (53.8%)

Number of residues in allowed region (~2.0% expected): 216 (25.4%)

Number of residues in outlier region: 177 (20.8%)

Table 4. 5 **QMEAN analysis of Model 5.**

Scoring function term	Raw score	Z-score
C-beta interaction energy	87.75	-3.09
All-atom pairwise energy	-1326.01	-3.15
Solvation energy	90.41	-6.41
Torsion angle energy	89.47	-5.96
Secondary structure agreement	49.4%	-5.88
Solvent accessibility agreement	52.3%	-5.25
QMEAN6 score	0.079	-7.56

Comparison of Model structures generated by I-TASSER

The structural validation and reliability of the predicted models of the 5 C1ORF112 models were carried out by RAMPAGE, ERRAT, and Verify-3D. Each of these looks at different aspects of the model to determine how close to the native structure of the model is. RAMPAGE was used to build the Ramachandran plot of the 5 models. Ramachandran plot looks at the torsional angles - phi (ϕ)and psi (ψ) of the amino acids in the peptide sequences and placed them in regions favoured for right-handed helices, beta sheets and left-handed helices and models 1,3, and 4 they had a minimum of approx. 80% of the peptide residues in the favoured region, which is the right-handed helical region. There are amino acids in the unfavoured region more than generally allowed for precise structural prediction. However, it should be noted that that modelling technique was ab initio modelling (non-template) and this was done using I-TASSER which was picked as it was the only program that used the full sequence for modelling. QMEAN analysis was also carried out for all the predicted models. QMEAN, which stands for Qualitative Model Energy Analysis, is s mathematical scoring function that describes the geometrical aspects of protein structures describing the C-beta interaction energy, All-atom pairwise energy, Solvation energy, Torsion angle energy, Secondary structure agreement, Solvent accessibility agreement and the then QMEAN6 (6 criteria that are being looked at) which is the weighted linear combination of the 6 scoring functions. The Z-score is a degree of nativeness to the actual structure and as a rule of thumb Z-scores of around 0.0 reflect a true native structure and less than -4.0 are low-quality structures. As shown from the results all the models are less than -4.0. Models 1,3, and 4 are within -0.4 and -4.5 (Table 5.6), although are low-quality models are reasonably better than the other 2 models. Although the models do not fall in the high-quality region, there is no resolved functional homologue that can be used as a template so they can only be used with caution as a prototype to understanding the function of C1ORF112. In addition, proteins undergo post-translation modifications that can modify their structures and these modifications are not always captured by non-template modelling and mathematical analysis of peptide torsional angles and these modifications can cause transitional states for example proteins such as CHK2, Checkpoint kinase 2, a serine/threonine kinase is activated through phosphorylation by ATM at the T68 priming site and on other residues at the SCD site leading to a conformational change of CHK2 monomer and allows the protein to dimerize (Zannini, Delia et al. 2014). these changes caused by post-translational modifications are not easily caught by *ab initio* modelling.

The modelling information generated, lead to narrowing down the models with the highest probability of being closely related to the native structure for further analysis. Models 1, 3 and 4 show very close similarities in structure, QMEAN and Z-scores. Although structure 4 has a slightly better score than the others indicating it might be a more likely native structure for the C1ORF112 protein.

Table 4. 6 **Comparison of the QMEAN and Z-scores of Models 1, 3 &4.**

Scoring function term	Raw score	Z-score
QMEAN6 score Model 1	0.352	-4.54
QMEAN6 score Model 3	0.350	-4.57
QMEAN6 score Model 4	0.398	-4.04

Further analysis was carried out on models 1, 3 and 4 on ERRAT and Verify-3D to establish the nature of the quality of the models. ERRAT scores of 94.9, 92.8 and 91 were scored respectively by each model. All models failed the Verify-3D analysis as they all had less than 80% of amino acids ≥ 0.2 in the 3D/1D profile. Although the results suggest that the models are not high-quality models, it does reveal that C1ORF112 is an alpha-helical protein with no discernible domains yet, it is also possible that I-TASSER may not be the best possible platform for modelling C1ORF112. The position of the WCF tripeptide as shown in Figure 4.11 does not provide any information as to its functional relevance and it is currently unclear if it has any role in the function of C1ORF112

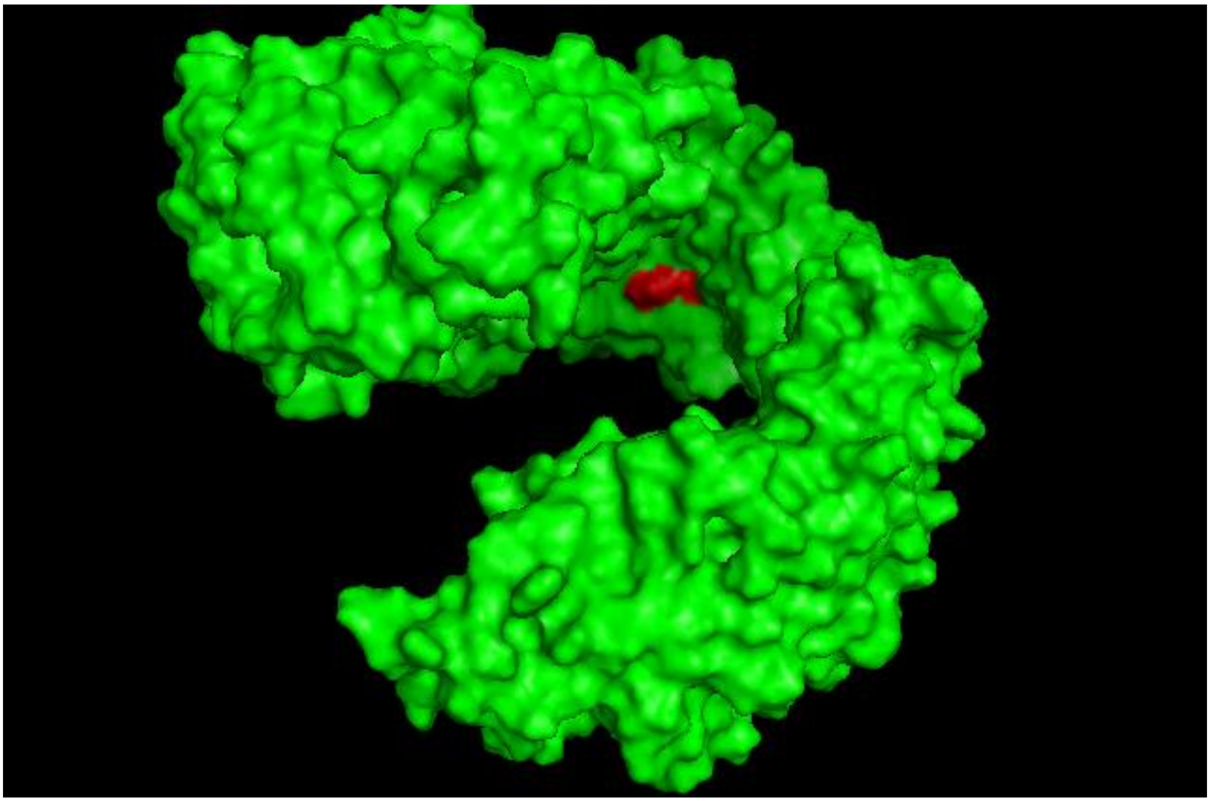


Figure 4. 12 **Model of C1ORF112 showing the position of the WCF tripeptide position**.

Surface area resolution of C1ORF112 modelled structure showing the position of the WCF tripeptide in the inner region in the region of the protein.

4.3 Comparison of I-TASSER generated models with another Protein structure prediction server AlphaFold*.*

AlphaFold is a protein prediction tool that predicts the 3D coordinates of heavy atoms using primary amino acid sequence, aligned with sequences of homologues as inputs. It is used MSA to pre-determine templates and increases the likelihood of better model resolution (Jumper, Evans et al. 2021). AlphaFold uses AI technology developed by Deepmind in collaboration with The European Bioinformatics Institute (EBI), models generated by AlphaFold are rated using pLDDT (per residue confidence scores based on RMSD of the MSA template (Varadi et al, 2021). AlphaFold is a more recent protein prediction software compared to I-TASSER and as such was used later as a comparative model with the ad initio models generated earlier by I-TASSER. The pre-determined structure of C1ORF112 was accessed from the server as a .pdb file and comparative analysis was carried out using the Ramachandran plot and QMEAN analysis, to validate the quality of the model. Figure 4.13 shows the primary structure of the predicted model of C1ORF112.

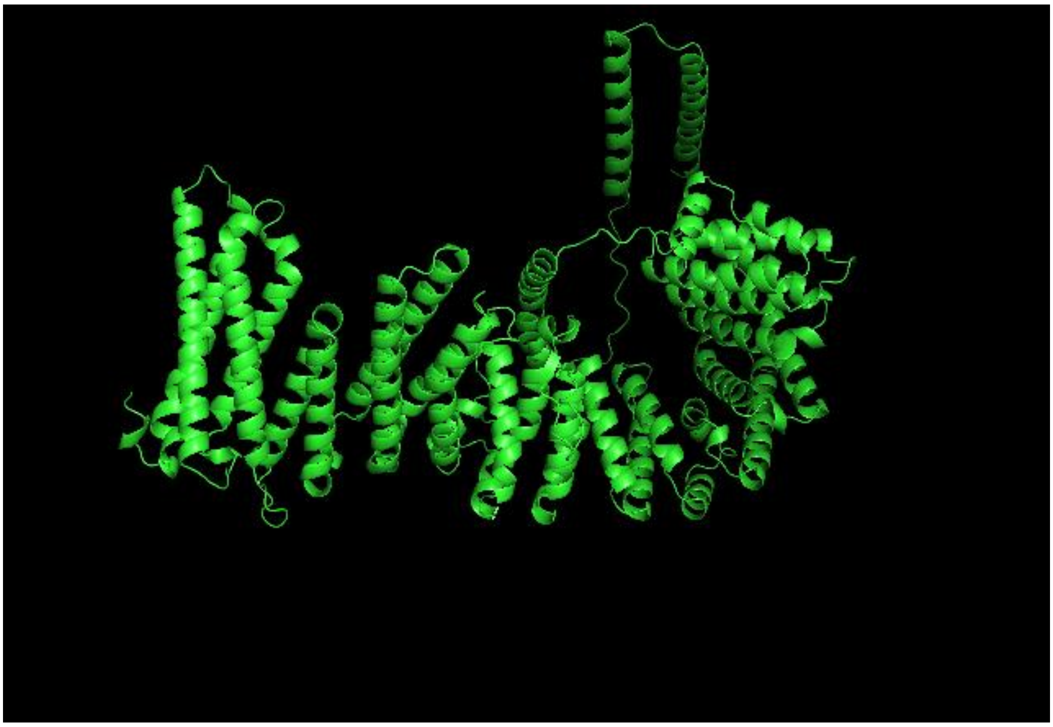
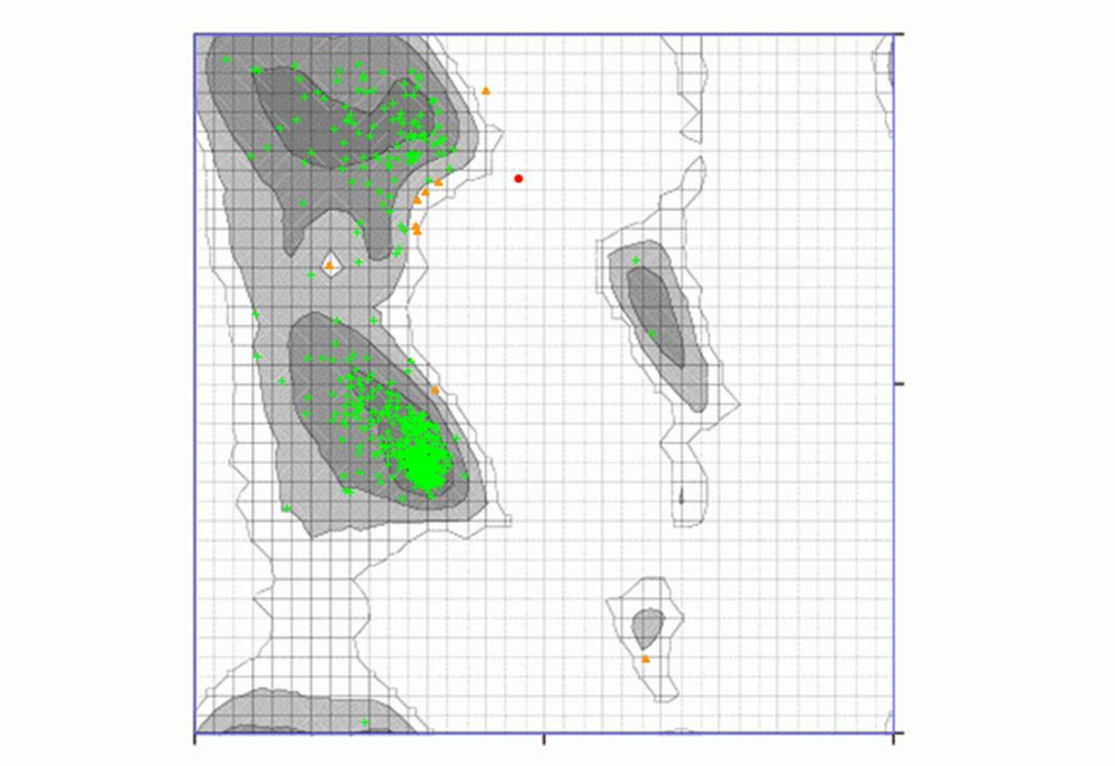


Figure 4. 13 **Primary structure of C1ORF112 generated by ALPHAFOLD prediction server**.

ALPHAFOLD modelled structure of C1ORF112 shows a more compact structure of the alpha-helix.



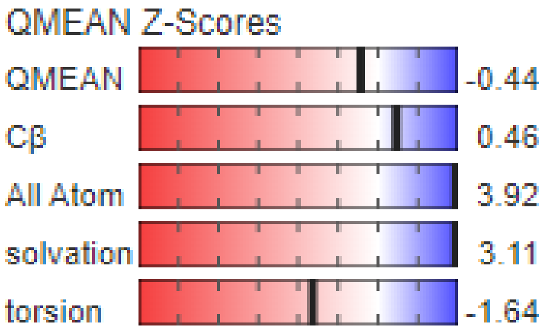


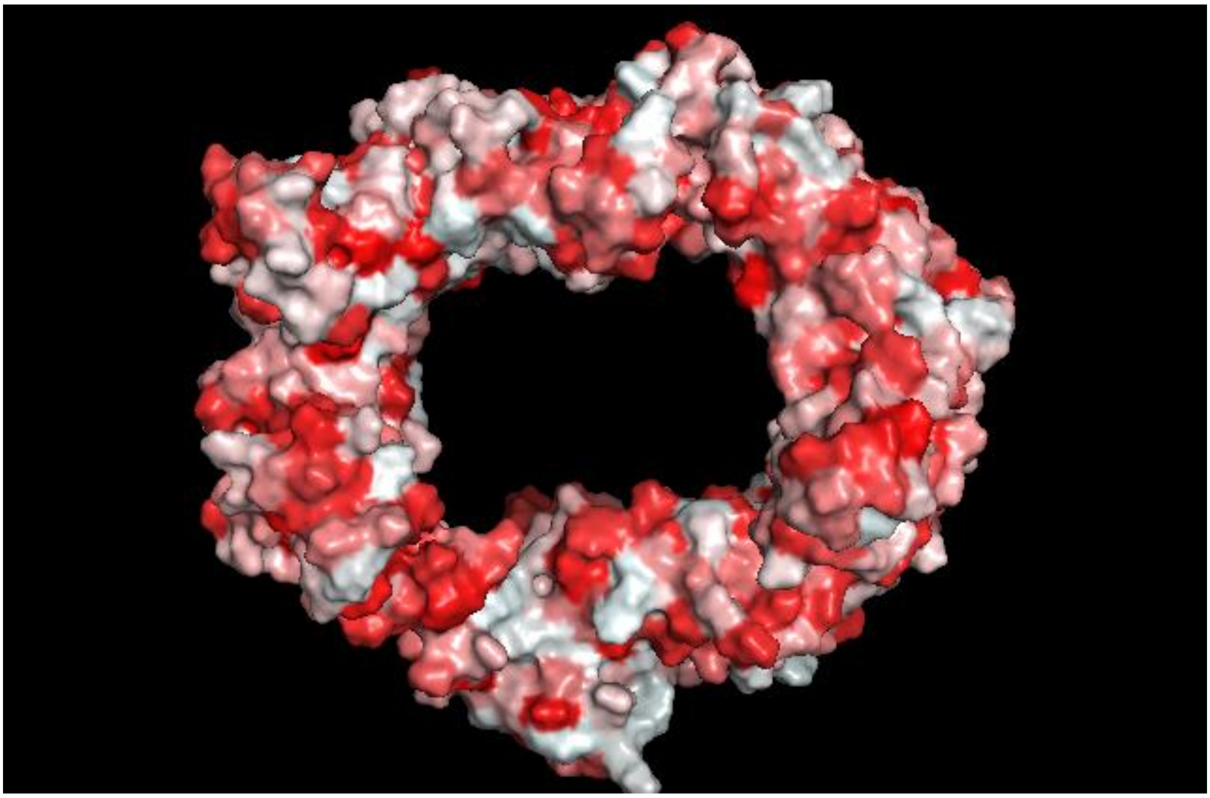
Figure 4. 14 Ramachandran and QMEAN analysis of AlphaFold model of C1ORF112.

The Ramachandran plot shows that all the peptide residues (green dots) do fall into the expected regions for alpha helices and the QMEAN evaluation of -0.44.

The Ramachandran analysis shows that most of the residues from the AlphaFold models fall in the favoured region for right-handed helices (see figure 4.14), and few residues are in the expected region for β-sheets. A Ramachandran score of 98% was achieved for the residues in the favoured region (green) and residues in the allowed region (brown) was 1.29% and residues in the outlier region (red) was less than 1%. The results are significantly better compared to the I-TASSER models, especially for whole sequence modelling. In addition, a QMEAN z-score of -0.44 suggests that the AlphaFold model is a lot closer to the native structure of C1ORF112. Furthermore, the ERRAT score for the overall quality factor was 98% and a VERIFY-3D score of 83.94% of the residues with averaged 3D-1D score >= 0.2 a passing score indicating a good quality model. These results show that the AlphaFold model of C1ORF112 is better compared to the I-TASSER generated models and as such closer to the native structure of C1ORF112, however, it does validate the prediction of the I-TASSER that C1ORF112 is an α-superhelix.

4.3 Hydrophobic distribution of residues on the C1ORF112 model

To further understand the structural relevance of C1ORF112 in cells, as no domains had been identified based on sequence similarity and homology modelling. The hydrophobic distribution of the amino acids in C1ORF112 using the models generated is shown in Figure 4.15 below.



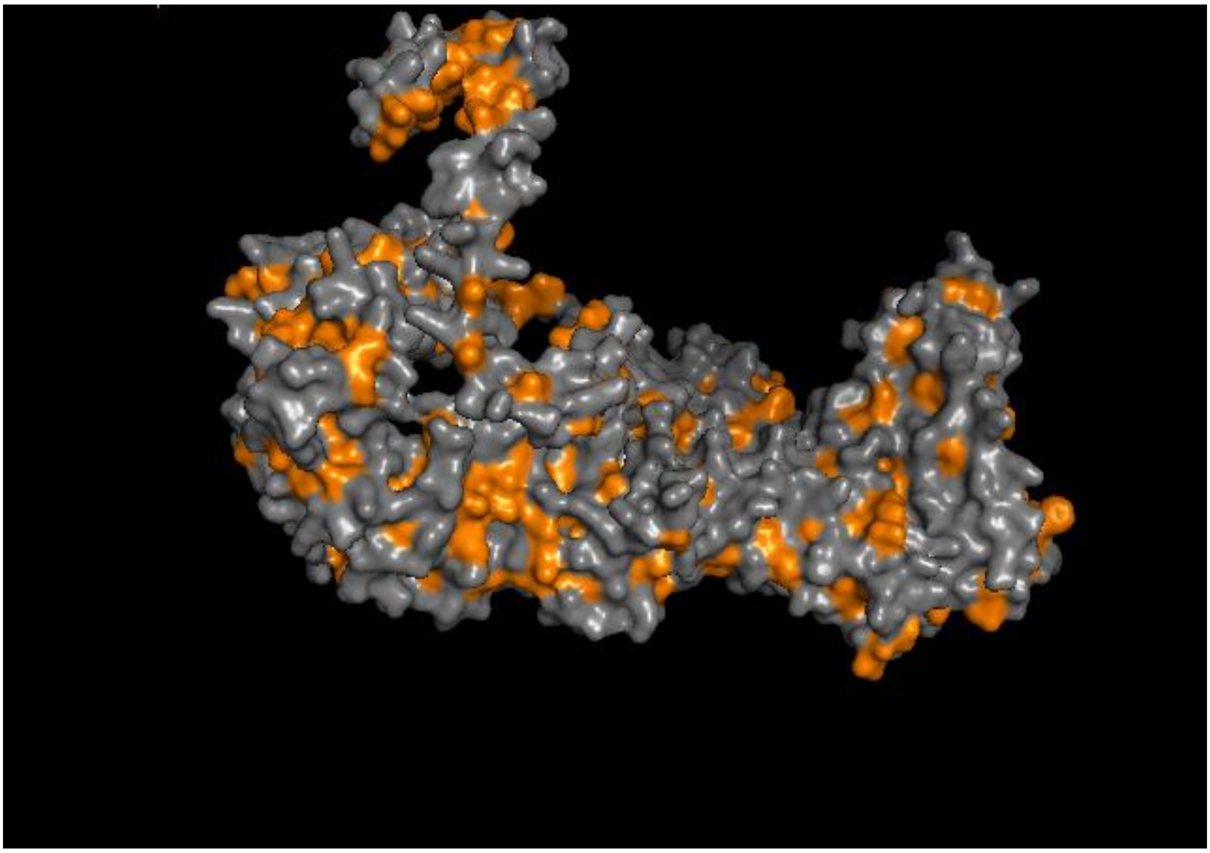


Figure 4. 15 **Hydrophobic distribution of C1ORF112**.

(Top) Hydrophobic residues shown in red are evenly distributed across the model generated from I-TASSER, (below) also even distribution of the hydrophobic residues in orange for the model generated from ALPHAFOLD.

The hydrophobic distribution of the amino acids does show that the hydrophobic residues that comprise C1ORF112 are evenly spread across the structure of the protein. Determining the hydrophobicity of C1ORF112 was essential to predict would present as a membrane interacting protein, as hydrophobic residues than to be lipophilic and as such interact with either the cellular or nuclear membrane (Kyte and Doolittle 1982). The hydrophobic coverage of the ALPHAFOLD model does suggest that C1ORF112 would be a cytosolic protein as the hydrophobic residues seem to be evenly spread and cover less surface area compared to the hydrophilic residues. This result does not suggest if C1ORF112 would fit as an integral or peripheral membrane protein. Inner membrane proteins tend to follow the positive inside rule, where the cytoplasmic loops have a higher number of positively charged residues. It also does not show the accumulation of the hydrophobic residues towards the interior of the model structure (Virkki, Peters et al. 2014).

4.4 Post-translational modification of C1ORF112

To further understand the structure of C1ORF112 and its amino acid residue, the post-translational modification it undergoes was estimated and analysed using the Phosphosite database https://www.phosphosite.org/homeAction.action. the results showed that C1ORF112 has 6 possible sites of phosphorylation and 11 possible sites of ubiquitylation as shown in Table 5.7.

Table 4. 7 **Possible phosphorylation sites on C1ORF112**.

6 possible phosphorylation sites on C1ORF112. Small alphabets denote phosphorylated residue (supplementary data 1 for full list)

Site	Peptide sequence from the alignment	Position in alignment	Position in the target protein	No. of species analysed	%Conservation out of total no. of proteomes	%Conservation in aligned orthologues	No. of species aligned	Phosphorylation likelihood in PSP	Observations in PSP	Phosphorylation likelihood in PA
S	TsQARGLssQNLEIQ	147	42	100	49	67.1	73	-	-	Low
S	sQARGLssQNLEIQT	148	43	100	59	80.8	73	Low	1	Medium
S	IHsKFPPsLYATRIs	414	288	100	66	90.4	73	Low	1	-
Y	SKFPPSLyATRISKA	416	290	100	58	79.5	73	Low	1	-
T	FPPSLYAtRISKAHQ	418	292	100	21	28.8	73	Low	1	-
S	ETKNKVVsFLEKTGF	961	744	100	45	61.6	73	Medium	2	-

The serine amino acids at the N-terminal ETKNKVVsFLEKTGF showed a higher possibility of being phosphorylated by interacting kinases, to enable the function of C1ORF112. 73 orthologues were aligned and the conservation of the sequence at the phosphorylating site was calculated to show there was >60% conservation in most of the sites for the orthologues. Only the Threonine residue in FPPSLYAtRISKAHQ Phosphosite showed less than 60% conservation. This suggests a higher probability of having the sites phosphorylated. As of the writing of this thesis a study by Xu, Ali et al. (2021), highlighted the N-terminal of C1ORF112 as an interacting domain for Polo-like kinase (PLK) and may mediate its function cell cycle regulation and in control of cell proliferation. Further analysis, as shown in Figure 4.16 shows that C1ORF112 is also ubiquitylated at various points across its sequences indicating an increased likelihood of tight control of expression.

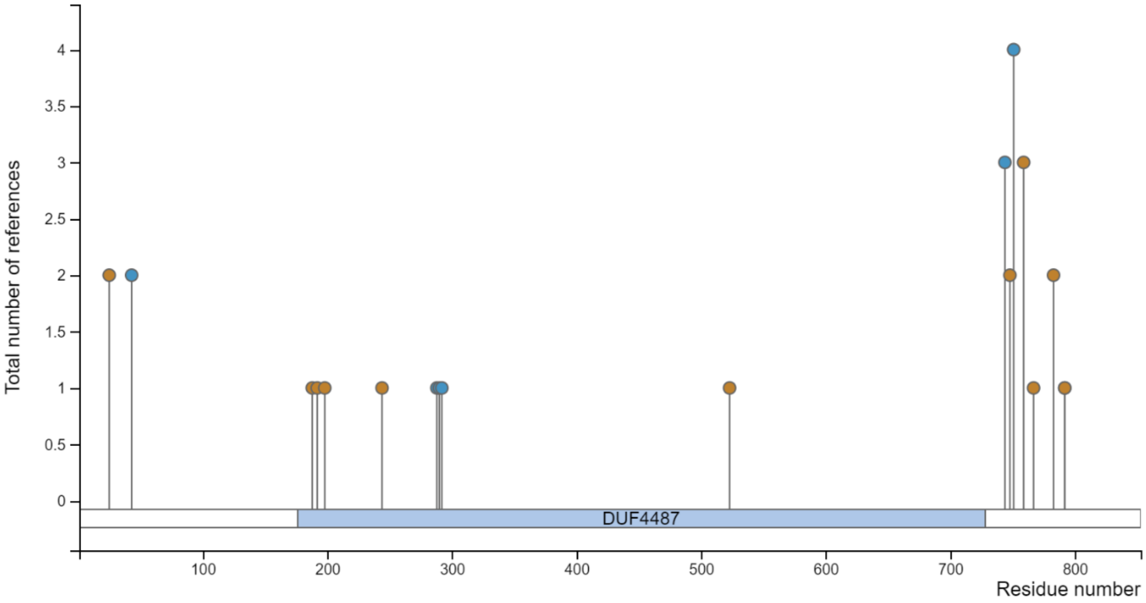


Figure 4. 16 **Position of post-translational modification along the sequence of C1ORF112**.

Brown colour denotes sites of ubiquitylation, and blue spots are the possible phosphorylation sites.

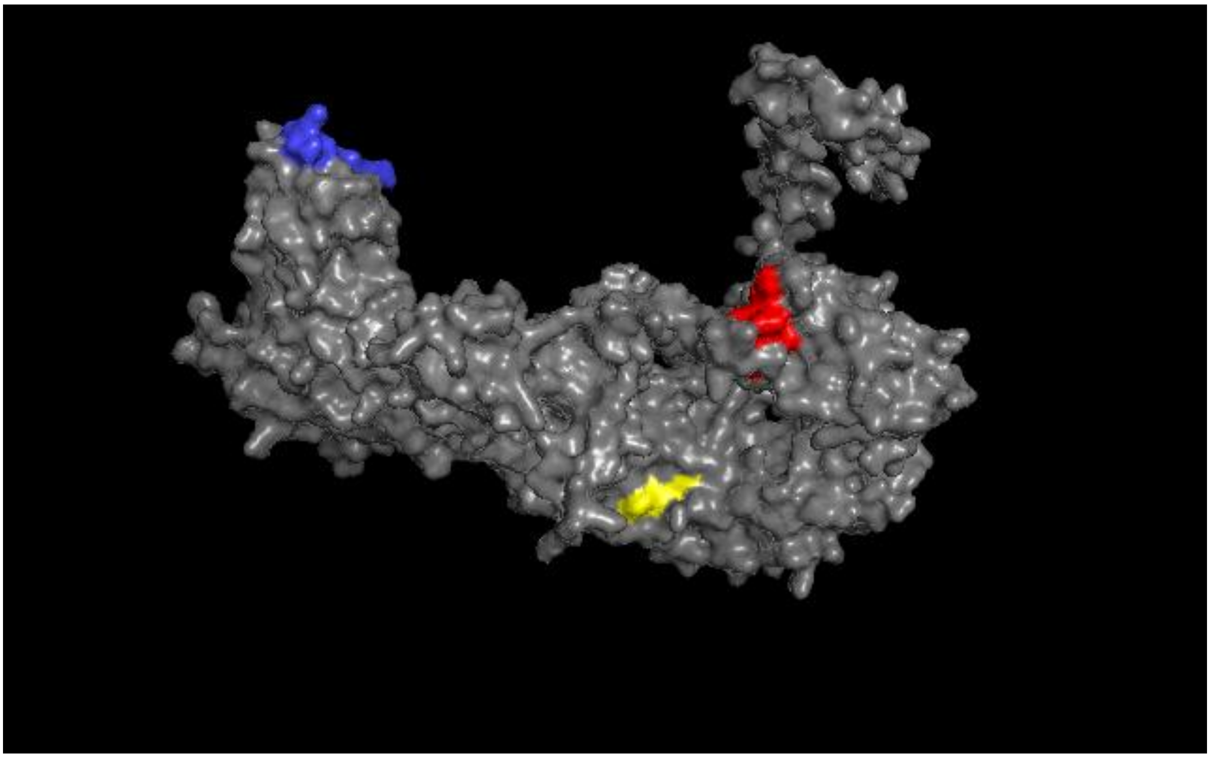


Figure 4. 17 **Surface area resolution of the AlphaFold model of C1ORF112.**

Yellow indicates the position of the WCF tripeptide, blue and red are the positions of the proposed phosphorylation sites discussed in section 4.4.

5.5 Discussion

Generation of 3-D protein models from amino acid sequences is generally the aim of protein model prediction, however, the efficacy of the prediction is dependent on several factors, one of which is the prediction tools used. I-TASSER hosted by the Zhang server has been used several times since its inception in 2006 for protein structure prediction with over 41 publications. In this case, the models created via I-TASSER has shown that C1ORF112 is an alpha superhelix. However, the models were not the best fit when further analysed using QMEAN, RAMPAGE, ERRAT, and Verify-3D. The results of the I-TASSER models did predict an alpha-helical structure, however, it did have limitations on the super fold structure, which could either be an algorithm issue or the sequence threading capacity of the server is limited, and this could be because of the sequence models used for its model prediction. It does seem, it could benefit from an improved model selection process. AlphaFold on the other hand was better at the structural prediction of C1ORF112. Further analysis using QMEAN, RAMPAGE, ERRAT, and Verify-3D showed that AlphaFold prediction of C1ORF112 is closer to its native protein compared to the I-TASSER selection. Other structural prediction software and servers such as the Swiss-Prot, were not utilized as they did not give full sequence model prediction for C1ORf112 and as such did not full the criteria for model prediction selection. So C1ORF112 *ab initio* model is α-super helical ARM repeats, no homology is present with a characterized functional protein. Co-expression analysis suggests involvement in DNA replication, repair, and recombination processes. MSA of helical proteins involved in the same processes and proteins with ARM repeats would indicate possible interacting regions on the sequence of C1ORF112

ARM repeats are imperfect tandem helical repeats of approx. 42 amino acids, with the motif composing three helical turns (Coates 2003, Tewari, Bailes et al. 2010). ARM repeat-containing proteins are called that because it was first characterised in the Drosophila polarizing protein Armadillo (Peifer, Berg et al. 1994). Proteins in this structural class appear not to share sequence similarity and play multi-functional roles in cells. Due to the sequence fluidity of ARM proteins, determining homologues and orthologues becomes difficult (Tewari, Bailes et al. 2010). C1ORF112 does maintain its sequence across species as discussed in Section 4, however, defining functional homologues within a species has proven difficult and would require more studies. In addition, C1ORF112 and other ARM proteins with the tandem superhelix as shown with the AlphaFold model forms a versatile platform for interacting with other proteins, this enables them to carry out a few functions based on their post-translational modification and domain architecture. As such several ARM proteins serve key functions in cells. For example, B-Catenin (Armadillo homologue) is typically a cell adhesion and a signalling protein, involved in the Wnt signalling pathway, transducing extracellular signals to modify gene expression in the nucleus (Cadigan and Peifer 2009, MacDonald, Tamai et al. 2009). In addition, B-Catenin is a hub protein for other signalling networks. Interestingly, B-Catenin has been shown to localise to the centrosome to interact with microtubules regulating their regrowth, cohesion, and separation (Huang, Senga et al. 2007, Bahmanyar, Kaplan et al. 2008). Wnt signals facilitate the splitting centrosomes and as such show both structural and transcriptional roles dependent on cellular localization and post-translational modification (Huang, Senga et al. 2007, Bahmanyar, Kaplan et al. 2008, Hadjihannas, Bruckner et al. 2010).

ARM proteins are not limited to structural and transcriptional roles, as another example of ARM proteins with different functional roles are the importin and exportin proteins. Importins are transport cargo proteins that shuttle proteins to and from the nucleus and consist of 10 ARM repeats (Mason, Stage et al. 2009). There are 3 sub-classes of importin-α in animals which are also evolutionally conserved in plants, fungi, amoeba and choanoflagellates (Mason, Stage et al. 2009). This level of conservation indicates the important role this class of ARM proteins plays across the Phylum, especially in spermatogenesis and gametogenesis, where their absence elicit sterility in Drosophila (Holt, Ly-Huynh et al. 2007, Ratan, Mason et al. 2008). These few examples highlight the range of roles ARM proteins play in the cell and how determining the function of C1ORF112 is a complex process.

The study by Xu, Ali et al. (2021), looking into Polo-like Kinase 1 (PLK1), C1ORF112, termed Apolo 1 was identified as aiding feedback control for chromosome segregation. This interaction between PLK1 and C1ORF112 was enabled mechanistically by binding the Polo-box domain (PBD) on PLK1 with its N-terminal region which corresponds to the serine amino acids at the N-terminal ETKNKVVsFLEKTGF. This suggests that the N-terminal region of C1ORF112 is a PBD binding domain. They also state that the C-terminal of C1ORF112 contains a Protein phosphatase 1γ (PP1γ) docking motif, hereby, interacting with both PLK1 and PP1γ. This alongside other studies further strengthen the idea that C1ORF112 could play a role in cell cycle regulation and control of cell proliferation. To further understand the role of C1ORF112, HAP1 cells modified to have the C1ORF112 knocked out were studied to understand its role in cells and the gene expression profile arising for silencing. Considering these limitations, both prediction tools do confirm that C1ORF112 is an alpha superhelix, which seems to be Armadillo repeats (ARM). In addition, the possible ubiquitylation of C1ORF112 at several points across its sequence indicate possible tight control of its expression, this could explain the tissue expression studies carried out in Chapter 4, showing the relatively low level of expression across tissues (Figure 4.2) and age groups (Figure 4.3).

Chapter 5: C1ORF112 knockout cells showed no phenotypic difference with normal cells and microarray analysis show cell cycle association

5.1 Introduction

Gene expression is the process where genomic information, in the form of a gene, is translated from DNA to a protein. Gene expression is the measure of the state of the cellular system at the point in time. Gene expression allows the cell to carry out certain functionalities over a period. It is interesting to note that differential gene expression is a snapshot of the cellular activities caused by changes in gene expression due to either internal or external conditions. Stochastic expression of genes can lead to random variation within the cell, because of fluctuations in gene expression (McAdams and Arkin 1999, Elowitz, Levine et al. 2002). Processes that alter gene expression can cause various downstream effects which can be described as cellular noise. Expression of genes can be affected by factors, exogenous sources such as drug treatments, exposure to either endogenous or exogenous sources of DNA damaging agents etc (discussed in section 1.1.4).

Gene expression regulation begins at the level of the DNA from chromosome remodelling largely through post-translational modification of histone proteins. When accessibility of the target genes is ensured, regulation of gene expression carries on at the level of transcription of genes and subsequent processes such as differential translation and post-translational modification of mRNA species (Elowitz, Levine et al. 2002). Insight into gene function can be extrapolated based on co-expressed genes and processed where gene expression is up or downregulated (discussed in Chapter 4). Further insights into gene function, can be obtained via over-expression of knockout studies of target genes. It is usually helpful if gene function or functional attributes can be narrowed down based on co-expression studies, or domain architecture, as it is generally known that genes with similar domains tend to have similar functions with different levels of specificity or downstream targets. There are new and different ways to understand changes to gene expression when a target gene has been either over-expressed or knocked down. Cell phenotype can be characterised and micro-array or RNA-seq studies can be carried out to elucidate the function of a gene through dysregulation of expression.

Microarray technology is an efficient way of addressing genetic questions by revealing patterns of gene expression and classifying sample datasets based on the patterns of expression (Kerr, Martin et al. 2000). This is very similar to RNA-sequencing (RNA-Seq), each with its pros and cons. One of the key differences between both technologies is that micro-array is a probe hybridization-based method used to capture genes while RNA-seq uses short nucleotide reads to map genes thereby increasing its capacity to sequence the whole transcriptome while micro-array can only profile pre-defined targets. This makes RNA-seq a more attractive technique compared to micro-array as it is more information based on its ability to capture a higher dynamic range at better resolution, however, it is more expensive and now, analysis of its data is not yet standardized compared to the micro-array which is cheaper and more standardized in analysis (Rao, Van Vleet et al. 2018). Nevertheless, both technologies are currently employed in studies to characterize the differentially expressed genes (DEG) across various perturbations, cell types and disease states to identify genes of interest.

At the start of this study, there was a significant lack of information on the roles of C1ORF12 except for a few studies such as Sanchez-Carbayo, Socci et al. (2007) and Leo, Wang et al. (2005) where C1ORF112 was shown to be co-expressed with cancer propagating genes. This study was therefore carried out to understand the role of C1ORF112 in cells, through phenotypic characterizations and microarray analysis. C1ORF112 mutant chronic myeloid leukaemia (HAP1) transformed cells were obtained commercially from Horizon Discovery (discussed in Chapter 2). To ensure the mutation generated no protein, western blot was performed on the mutant and normal cell lines. This was then followed by phenotypic characterizations to determine if there were any phenotypic differences between the normal cell line and the mutant cell lines. Microarray analysis was then followed to determine DEGs between the mutant and normal cells lines. Based on the co-expression results (Chapter 4), the hypothesis of this study was to determine if knockdown of C1ORF112 would cause any changes in cell growth between the normal and knockout cell lines and would give a gene expression profile complimenting those of the co-expression studies. An expectation is that C1ORF112 facilitates or compliments the functions of other cell cycle-related genes.

HAP1 cells were chosen as they are haploid cells, meaning a gene mutation in a single allele will not be masked by the presence of the second allele. This would allow for easier observation of gene knockdown effects and prevent double knockdown mutations from being carried out, reducing cost genetic manipulation.

5.2 Western blot to ensure knock out of C1ORF112 in cells

To generate the haploid cell line, which is HAP1, the Horizon performed as a series of subcloning the heterogeneous human leukaemia cell line, KBM-7 described in Kotecki, Reddy et al. (1999). KBM-7 cells have approximately half the human diploid DNA content and a haploid karyotype except for the disomic chromosome 8 and a portion of chromosome 15 (Kotecki, Reddy et al. 1999). From the KBM-7 cells, subclone P1-55 was isolated with higher stability compared to other near-haploid clones (Kotecki, Reddy et al. 1999). Further subcloning resulted in the isolation of the HAP1 subclones, the major difference of HAP1 to other near-haploid subclones is that it contains a single copy of all chromosomes except for a heterozygous 30-megabase fragment from chromosome 15, this fragment is integrated on chromosome 19 and encompasses 330 genes (Essletzbichler, Konopka et al. 2014). The HAP1 clone is a very stable subclone of the near-haploid subclones from the KBM-7 cell line (Essletzbichler, Konopka et al. 2014), Horizon then used the CRISPR/Cas9 technology on the HAP1 cells to delete the disomic copy of the Chromosome 15 present in the HAP1 cell line (Essletzbichler, Konopka et al. 2014). To generate the C1ORF112 knockout cell lines, Horizon used the same CRISPR/Cas9 technology on the HAP1 cells to induce a 1bp insertion to the exon 5 of the C1ORF11, to determine the nature of the mutation caused by the 1bp insertion in exon 5 of the C1ORF112 sequence in the knockout cell line. The western blot procedure was carried out as described in Chapter 3 (sections 2.1.8 – 2.1.12), Figure 5.1 shows that mutation produced no protein. As stated in section 3.2, the MW of C1ORF112 is 96.6kDa, this can be seen from the western blot analysis on both lanes where the normal C1ORF112 is present.

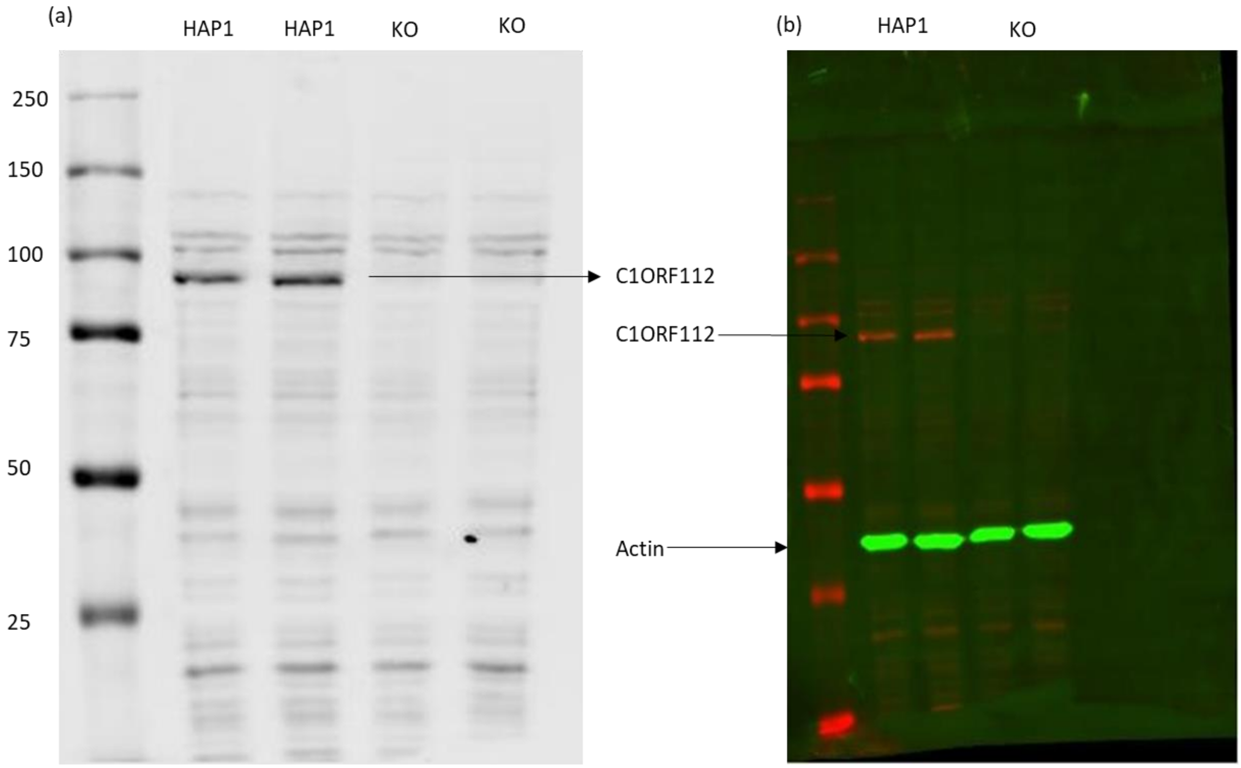


Figure 5. 1 **Western blot analysis between normal C1ORF112 and mutant C1ORF112**.

(a) blot shows C1ORF112 just below the 100kDa as C1ORF112 is 96kDA on the two lanes just after the protein marker lane. Knock out (KO) lanes showing no presence of the C1ORF112 protein. (b) the actin control is in green and C1ORF112 is in red in the same position.

Image (a) on figure 5.1 shows the thick bands of C1ORF112 on the two lanes for normal HAP1 cells, just below the 100 kDa marker on the protein ladder, which is absent on the two lanes for the C1ORF112 knockout cells. Across the blot, there is the similarity of non-specific binding of the antibodies which the exception of the C1ORF112 bands in the KO cells lanes. The image (b) also shows the C1ORF112 protein in red (Alexa flour rabbit 680 see methods 2.3.9) just below the 100 kDa marker and actin green (Alexa flour mouse 800) below the 50 kDa marker, as actin has a molecular weight of 42. Since the mutant cells produce no C1ORF112 it was necessary to determine if the knockdown of C1ORF112 directly affect the cell growth in the mutant cell line compared to the normal cells.

5.3 C1ORF112 mutant cells grow the same as normal cells

To understand if there were any phenotypic differences between the mutant cell line and the normal cell lines, a cell counting analysis was performed shown in Figure 6.2, this was carried out to determine if the knockdown of C1ORF112 would show any immediate effect on the cell's capacity to proliferate. 1x10^6 cells were seeded in T75 flasks for each cell line (Normal and Knockout), for each time point (24hr, 48hr and 72hr). Each flask was then trypsinised at each time point and the live cell count was taken using the Bio-Rad T20 automatic cell counter, to determine if there was any significant difference in cell numbers at each time point.

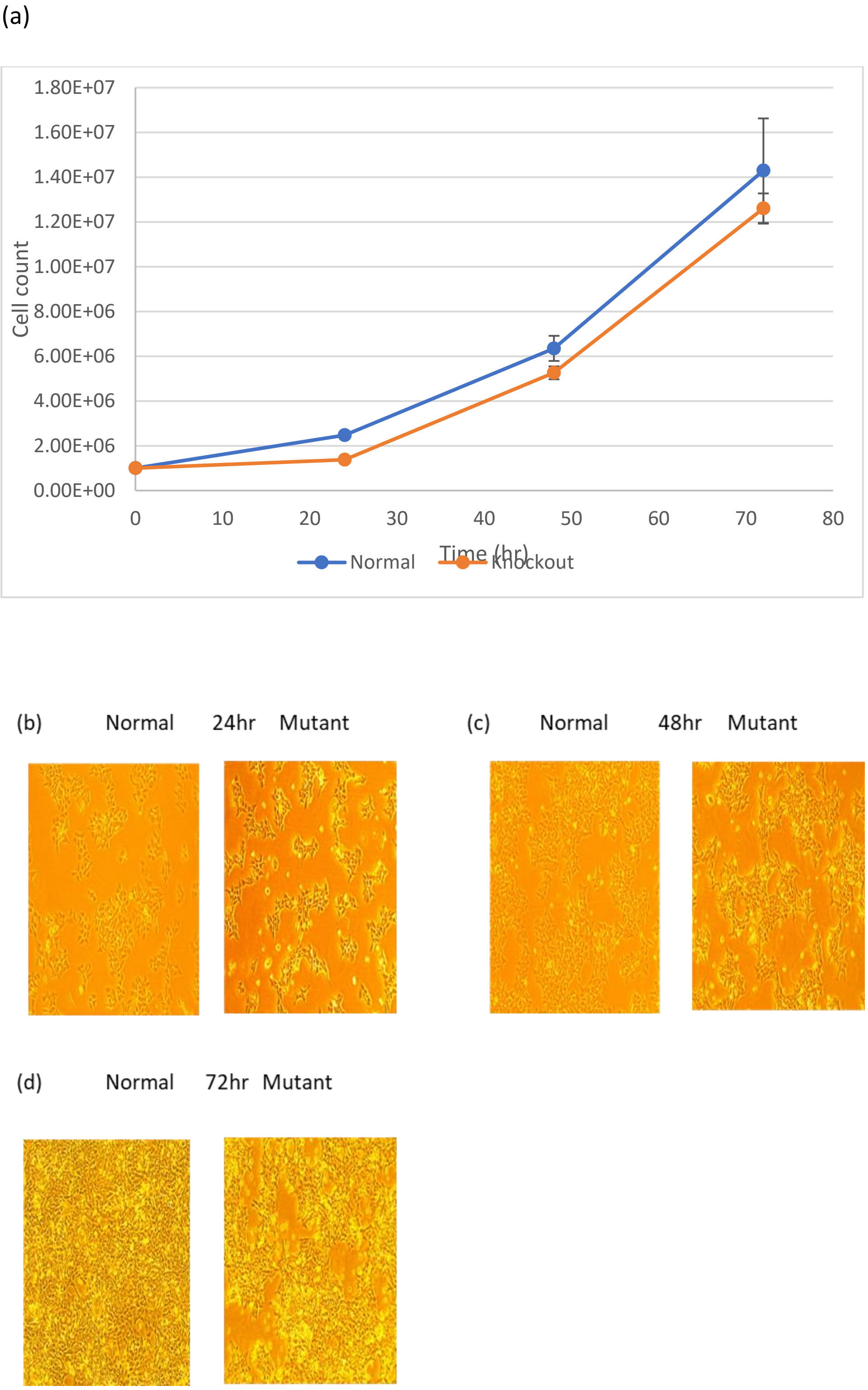


Figure 5. 2 **Comparison of the cell growth between Normal HAP1 and Mutant HAP1**.

(a) The figure indicates the increase in cell numbers of both normal (blue) and knockout (orange) error bars indicate standard deviation in cell numbers. (b-d) representative image of cell confluency between normal and knockout cells at each time point before counting.

At the 24hr time point post-seeding, the confluency of the cells was determined to be about 30-40% for both normal and mutant cells. After counting the cells for both cell lines, the average live cell count was ascertained and used to determine the growth rate and doubling time between the normal and knockout cells. To calculate the cell doubling time and the growth rate the following calculation was used.

$$\text{Growth rate} = \frac{\text{In(Final concentration of cells/ initial concentration of cells)}}{\text{Duration (time)}}$$

$$\text{Doubling time} = \text{Duration (time) X} \quad \frac{\text{In(2)}}{\text{In(Final concentration of cells/ initial concentration of cells)}}$$

At 24hr the doubling time for the normal HAP1 cells was determined to be 18hrs 22mins and the growth rate of the cells was 2 cells per min, for the knockout HAP1 cells, the doubling time was 52hrs 2min and the growth rate was 0.8 cells per min. Statistical analysis using the student’s T.Test to ascertain the difference in cell number between the normal HAP1 cells and the C1ORF112 knock out cells had a p-value of 0.02, indicating the was a significant difference in the cell numbers between both cells line after 24hrs. This analysis was carried out subsequently for the 48hr and 72hr, after 48hr post-seeding, the cell doubling time for the normal cells was determined to 17hr 59 mins and the cell growth rate 2.31 cells per min, and at 72hrs the doubling was 18hr 45mins and cell growth rate at 2.2 cells per min. for the C1ORF112 knockout cells, at 48hrs the doubling time had reduced to 20hr and 2mins and the cell growth rate was 2 cells per min, and at 72hr the doubling time was 19hrs 41 mins, with the cell growth rate of 2.1 cells per min. This is reflected in the growth curve (figure 5.2a) as the normal cells have a higher cell count at 24hrs and then steady growth at 48 and 72hrs. C1ORF112 knockout cells, on the other hand, have a lower cell count after 24hrs and then the growth rate picked up and tracked with the normal cells at the 48hr and 72hr time points. Statistical analysis at the 48 and 72hr time points between the normal mutant cells showed no significant differences with p-values of 0.24 and 0.15 respectively. The confluence of the cells (figure 5.2 b-d) also reflects the results as in figure 5.2b, the 24hr time point the normal cells appear to be more confluent compared to knockout cells and this carries on into the 48hr and 72hr time points, where the confluence for both cell lines is about 50-60% and that of the 72hr time point is >80%.

Following this, and once it was established that mutant HAP1 cells did not produce any form of C1ORF112 (truncated or functionally inactive) because of the modification and there was no significant difference between both cell lines (mutant and normal) at the 48hr and 72hr time points but the significant difference at the 24hr time point, I proceeded to analyse the mutant and normal HPA1 cells for microarray analysis to establish the differential gene expression profile between both cell lines and to ascertain if cell cycle-related genes would be significantly downregulated because of the C1ORF112 protein not being produced.

5.4 Differential gene expression between C1ORF112 knockout cells and C1ORF112 normal cells

The microarray analysis was carried out to determine the differential expression of protein-coding genes between normal HAP1 and mutant HAP1 at the two-time points post-seeding (24hr and 48hr). The 24hr time point had 3 biological sample replicates for both the normal cells and the mutant cells, the 48hr Cut-off FDR P-Val (false discovery rate) was set at 0.05 and gene fold change (FC) was set at +/- 1.5 (+ = indicative of upregulation; - indicative of down-regulation).

Differential expression between Mutant HAP1 vs Normal HAP1 at 24hrs

After filtering all protein-coding genes differentially expressed based on the criteria, the total number of differentially expressed protein-coding genes was 730. The total number of protein-coding genes that were down-regulated were 289, and the total number of protein-coding genes that were up-regulated was 441, 24hr post-seeding (Figure 5.3)

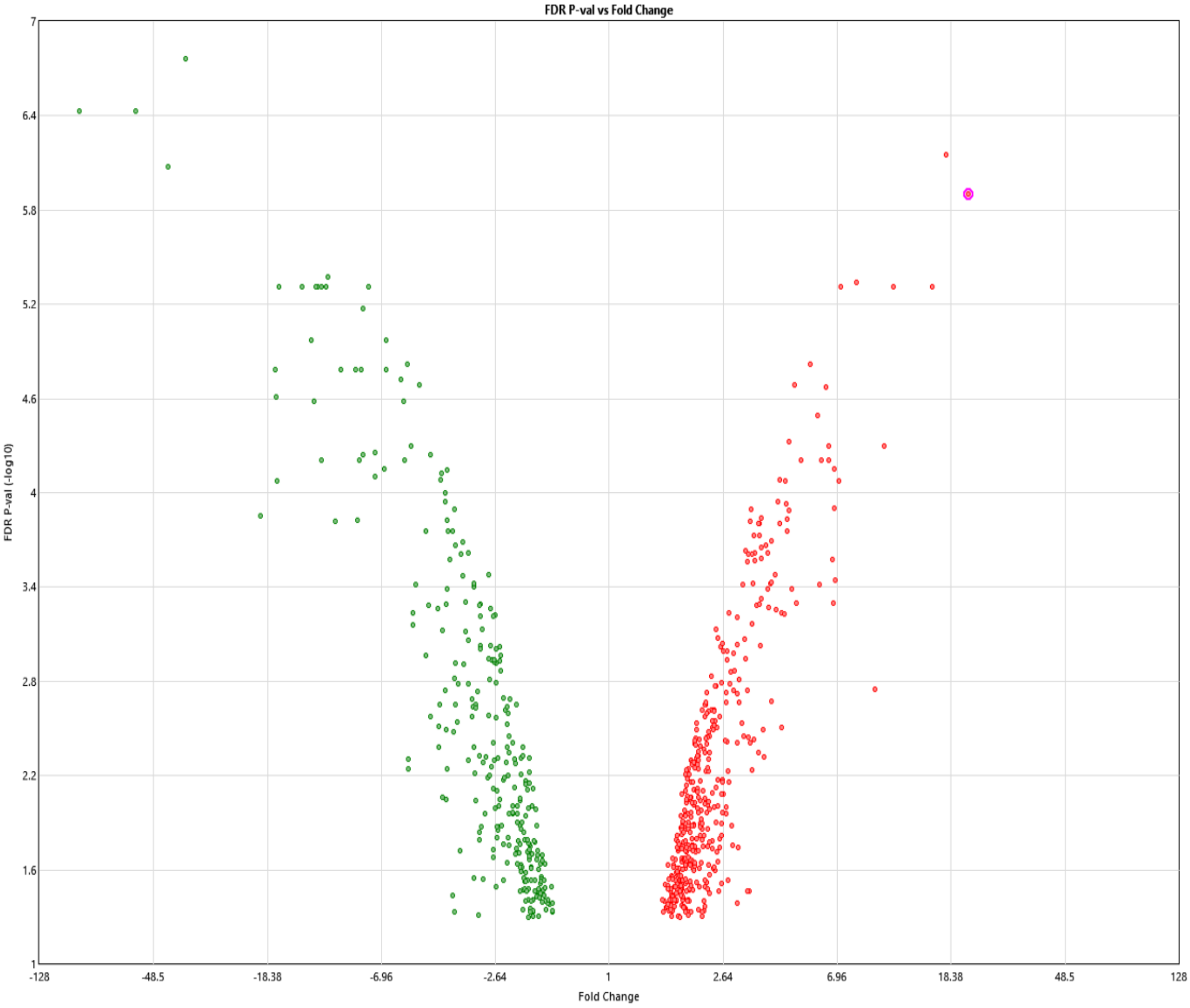


Figure 5. 3 **Volcano plot showing the differential expression of upregulated genes (red) and down-regulated genes (green) at 24hrs.**

Filtered for a cut-off of +/- 1.5 FC and 0.05 FDR. Obtained from the Transcription Analysis Control (TAC) a software for micro-array analysis (Thermofisher, UK).

Differential expression between Mutant HAP1 vs Normal HAP1 at 48hrs

The same analysis pipeline was carried out for the cells at 48hr post-seeding, the total number of differentially expressed protein-coding genes was 368. The total number of protein-coding genes that were down-regulated were 144, and the total number of protein-coding genes that were up-regulated was 224 (Figure 5.4)

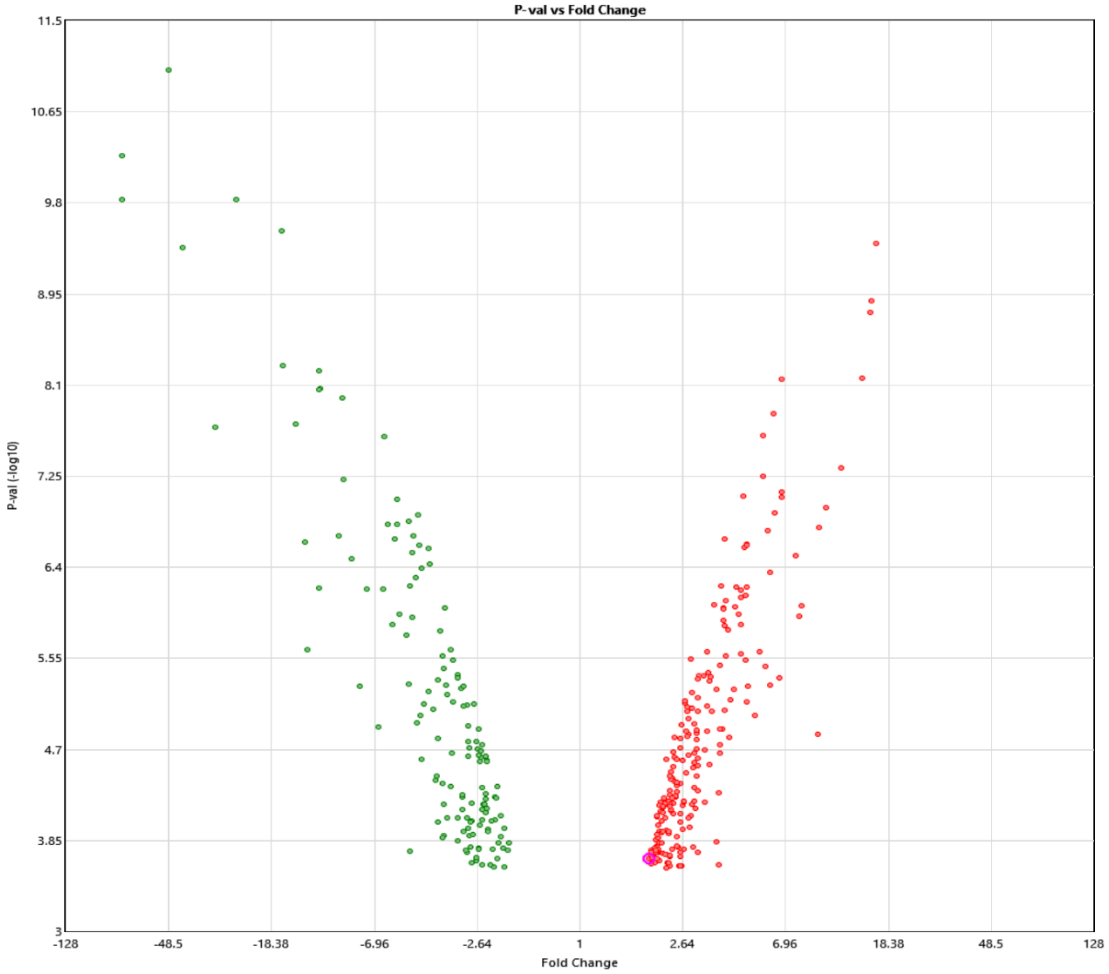


Figure 5. 4 **Volcano plot showing the differential expression of upregulated genes (red) and down-regulated genes (green) at 48hrs.**

Filtered for a cut-off of +/- 1.5 FC and 0.05 FDR. Obtained from the Transcription Analysis Control (TAC) a software for micro-array analysis (Thermofisher, UK).

The top 25 genes differentially expressed at each time point are listed in Table 5.1 -5.4. The top genes upregulated at both 24- and 48- hour time points were GALNT5 (polypeptide N-acetylgalactosaminyltransferase 5) and EYA1 (EYA transcriptional coactivator and phosphatase 1) and the top genes down-regulated in both time points were PRRX1 (paired related homeobox 1) and TFAP2D (transcription factor AP-2 delta (activating enhancer-binding protein 2 delta.

Top 25 DEGs at 24hr

Table 5. 1 **Top 25 DEGs downregulated in *C1ORF112* knockout cells versus wild type at 24 hours post-seeding**.

Top 25 genes differentially expressed based on significance (log10 p-value). The greyed boxes are the genes most significantly regulated based on Log10 p-value and Fold change.

Gene Symbol	Mutant24 Avg (log2)	Wildtype24 Avg (log2)	Fold Change	P-value	Log10 FDR P-value
BZW1	10.2	15.4	-36.83	2.32E-12	-6.76
PRRX1	5.05	11.55	-91.01	1.36E-11	-6.43
TFAP2D	4.36	10.18	-56.63	1.3E-11	-6.43
KLF4	4.87	10.29	-42.87	4.33E-11	-6.08
CEBPZOS	5.71	9.17	-11.01	2.81E-10	-5.37
DPYSL3	5.6	9.66	-16.67	4.97E-10	-5.32
POF1B	7.27	11.04	-13.65	5.91E-10	-5.32
DLX5	12.19	15.8	-12.2	6.86E-10	-5.32
TSPAN5	7.55	11.13	-11.96	5.69E-10	-5.32
ITGB8	6.15	9.68	-11.6	7.85E-10	-5.32
TUBB3; MC1R	8.95	12.43	-11.19	5.42E-10	-5.32
NELL1	7.02	10.04	-8.12	1.14E-09	-5.17
MXRA8	5.16	8.82	-12.62	2.15E-09	-4.97
DLX6	8.84	12.13	-9.83	4.48E-09	-4.79
SCGB3A2	6.28	9.33	-8.29	4.7E-09	-4.79
PREX2	5.22	9.32	-17.23	5.24E-09	-4.78
DOC2B	5.9	9.02	-8.69	5.29E-09	-4.78
C9orf135	5.23	9.32	-17.04	9.39E-09	-4.61
PGM5	5.24	8.87	-12.4	1.04E-08	-4.58
NPR3	4.89	8.43	-11.62	3.37E-08	-4.21
ASCL1	5.24	8.31	-8.41	3.24E-08	-4.21
RCOR2	4.26	8.34	-16.92	5.35E-08	-4.08
BZW1P1	3.49	7.78	-19.58	1.06E-07	-4.00
DTNA	3.69	6.78	-8.52	1.18E-07	-4.00
CRLF1	5.01	8.37	-10.28	1.25E-07	-3.70

Table 5. 2 **Top 25 DEGs upregulated in *C1ORF112* knockout cells versus wild type at 24 hours post-seeding**.

Top 25 genes differentially expressed based on significance (log10 p-value). The greyed boxes are the genes most significantly regulated based on Log10 p-value and Fold change.

Gene Symbol	**Mutant24 Avg (log2)**	**Wildtype24 Avg (log2)**	**Fold Change**	**Log10 FDR P-value**
EYA1	10.27	6.13	17.56	-6.15
GALNT5	12.72	8.31	21.19	-5.90
VSNL1	13.84	10.8	8.2	-5.34
PRAME	10.23	6.26	15.68	-5.32
SYT1	14.2	10.71	11.25	-5.32
NRG4	12.2	9.35	7.19	-5.32
MAGI2-AS3	12.49	10.02	5.53	-4.82
CLDN1	10.26	7.99	4.85	-4.69
PROM1	13.57	10.9	6.36	-4.67
IFI16	11.02	8.47	5.88	-4.49
CSMD1	9.26	5.89	10.36	-4.30
LPL	8.68	5.98	6.48	-4.30
EPHA3	10.82	8.12	6.5	-4.21
TFPI2	8.74	6.13	6.1	-4.21
GNG11	7.86	5.5	5.11	-4.21
ZNF385D	10.41	7.64	6.79	-4.15
PGR	9.09	6.27	7.06	-4.08
GCNT2	8.49	5.72	6.81	-4.00
CROT	10.28	7.53	6.69	-3.52
LRP1B	8	5.22	6.88	-3.40
HOXC9	9.59	7.01	5.99	-3.40
CNPY1	9.6	7.36	4.72	-3.40
PLPPR4	8.22	5.46	6.75	-3.30
PPP2R2C	10.82	8.52	4.93	-3.30
EDIL3	7.93	4.66	9.61	-2.74

Gene ontology (GO) terms for biological processes that are over-represented in the DEGs genes at 24hr post-seeding.

Functional enrichment analysis was carried out for both up-regulated and down-regulated genes to find out the possible collective function of these genes. A functional enrichment cut-off (E. Score) of 1.3 was used as Enrichment scores above 1.3 (which corresponds to P = 0.05) are widely accepted as relevant (Huang, Sherman et al. 2009). A Benjamini correction was applied for correcting for multiple hypothesis testing (Fernandes, Wan et al. 2016).

Table 5. 3 **Gene ontology (GO) terms for biological processes that are over-represented in down-regulated genes differentially expressed at 24hr post-seeding.**

GO	Term	log10 p-value
GO:0048015	phosphatidylinositol-mediated signalling	-3.98
GO:0045668	negative regulation of osteoblast differentiation	-3.74
GO:0030326	embryonic limb morphogenesis	-3.69
GO:0000052	citrulline metabolic process	-3.26
GO:0042472	inner ear morphogenesis	-3.15
GO:0071773	cellular response to BMP stimulus	-3.15
GO:0030501	positive regulation of bone mineralization	-2.89
GO:0045669	positive regulation of osteoblast differentiation	-2.86
GO:0001657	ureteric bud development	-2.76
GO:0060687	regulation of branching involved in prostate gland morphogenesis	-2.43
GO:0008201	Heparin-binding	-2.24
GO:0042391	regulation of membrane potential	-1.71
GO:0060021	palate development	-1.69
GO:0030509	BMP signalling pathway	-1.69
GO:0030509	BMP signalling pathway	-1.69
GO:0045429	positive regulation of nitric oxide biosynthetic process	-1.68
GO:0016597	amino acid binding	-1.46
GO:0043433	negative regulation of sequence-specific DNA binding transcription factor activity	-1.31
GO:0005509	calcium ion binding	-1.00
GO:0010862	positive regulation of pathway restricted SMAD protein phosphorylation	-0.85
GO:0046983	protein dimerization activity	-0.49

Table 5. 4 **Gene ontology (GO) terms for biological processes that are over-represented in up-regulated genes differentially expressed at 24hr post-seeding.**

GO	Term	log10 P-value
GO:0052851	ferric-chelate reductase (NADPH) activity	-4.03
GO:0008823	cupric reductase activity	-4.03
GO:0097461	ferric iron import into the cell	-4.01
GO:0015677	copper ion import	-3.48
GO:0042178	xenobiotic catabolic process	-3.48
GO:1901687	glutathione derivative biosynthetic process	-2.94
GO:0018916	nitrobenzene metabolic process	-2.57
GO:0004364	glutathione transferase activity	-2.20
GO:0051056	regulation of small GTPase mediated signal transduction	-2.06
GO:0050839	cell adhesion molecule binding	-1.99
GO:0008152	metabolic process	-1.96
GO:0097105	presynaptic membrane assembly	-1.82
GO:0043295	glutathione binding	-1.66
GO:0006749	glutathione metabolic process	-1.49
GO:0007158	neuron cell-cell adhesion	-1.34
GO:0003707	steroid hormone receptor activity	-0.93
GO:0055072	iron ion homeostasis	-0.86
GO:0006367	transcription initiation from RNA polymerase II promoter	-0.63
GO:0043547	positive regulation of GTPase activity	-0.47
GO:0043401	steroid hormone-mediated signalling pathway	-0.46
GO:0005096	GTPase activator activity	-0.42

Top 25 DEGs at 48hr

Table 5. 5 **Top 25 DEGs downregulated in *C1ORF112* knockout cells versus wild type at 48 hours post-seeding**.

Top 25 genes differentially expressed based on significance (log10 p-value). The greyed boxes are the genes most significantly regulated based on Log10 p-value and Fold change.

Gene Symbol	Mutant48 Avg (log2)	Wildtype48 Avg (log2)	Fold Change	FDR P-value
BZW1	9.89	15.49	-48.49	-5.90
TFAP2D	4.47	10.71	-75.61	-5.71
PRRX1	4.88	11.12	-75.58	-5.48
CEBPZOS	5.11	9.79	-25.62	-5.48
CRABP1	6.49	10.56	-16.79	-5.26
KLF4	4.96	10.38	-42.66	-5.21
POF1B	7.07	11.11	-16.49	-4.30
DLX5	12.1	15.66	-11.77	-4.27
SCGB3A2	5.69	9.25	-11.81	-4.22
TSPAN5	7.7	11.23	-11.62	-4.22
TUBB3; MC1R	9.14	12.38	-9.45	-4.17
MXRA8	5.77	9.65	-14.73	-4.06
C9orf135	5.18	10.15	-31.35	-4.04
PDGFRA	8.19	10.85	-6.33	-4.00
DOC2B	5.79	9.02	-9.38	-3.70
NPR3	4.7	8.44	-13.42	-3.40
DTNA	3.77	7.05	-9.74	-3.40
ITGB8	6.54	9.15	-6.12	-3.40
PREX2	5.37	8.49	-8.66	-3.30
CRLF1	5.19	8.75	-11.75	-3.10
PGM5	5.65	8.55	-7.5	-3.10
DLX6	9.01	11.69	-6.42	-3.10
BZW1P1	3.99	7.71	-13.15	-2.68
RCOR2	5.04	8.04	-8.04	-2.44
SLC7A8	5.59	8.33	-6.68	-2.19

Table 5. 6 **Top 25 DEGs upregulated in *C1ORF112* knockout cells versus wild type at 48 hours post-seeding**.

Top 25 genes differentially expressed based on significance (log10 p-value). The greyed boxes are the genes most significantly regulated based on Log10 p-value and Fold change.

Gene Symbol	Mutant48 Avg (log2)	Wildtype48 Avg (log2)	Fold Change	Log10P-Val
EYA1	10	5.96	16.38	-5.21
SYT1	14	10.03	15.65	-4.80
GALNT5	12.76	8.81	15.48	-4.72
PRAME	10.14	6.29	14.41	-4.26
VSNL1	14.34	11.6	6.71	-4.26
NRG4	12.06	9.43	6.18	-4.12
CLDN1	10.65	8.15	5.63	-4.00
PGR	9.04	5.49	11.71	-3.70
LPL	8.82	6.07	6.72	-3.70
IFI16	11.58	8.83	6.7	-3.70
DAZL	11.76	9.28	5.6	-3.70
CSMD1	9.85	6.49	10.23	-3.52
PROM1	13.46	10.81	6.26	-3.52
ZNF385D	11.47	8.22	9.54	-3.40
CSMD3	9.47	6.91	5.91	-3.40
TFPI2	8.97	6.03	7.68	-3.30
EPHA3	10.71	8.12	5.99	-3.15
CROT	10.53	7.52	8.06	-2.96
PLPPR4	9.05	6.07	7.9	-2.92
GCNT2	8.55	6.1	5.46	-2.66
RARB	6.98	4.46	5.75	-2.57
HOXC9	9.72	7.01	6.57	-2.49
LRP1B	7.66	5.07	6.02	-2.44
RENBP	8.9	6.52	5.22	-2.27
EDIL3	8.48	5.24	9.41	-2.15

Gene ontology (GO) terms for biological processes that are over-represented in the DEGs genes at 48hr post-seeding

Table 5. 7 **Gene ontology (GO) terms for biological processes that are over-represented in down-regulated genes differentially expressed at 48hr post-seeding.**

GO	Term	log10 p-value
GO:0071682	endocytic vesicle lumen	-3.87
GO:0001501	skeletal system development	-3.51
GO:0042472	inner ear morphogenesis	-3.40
GO:0045669	positive regulation of osteoblast differentiation	-3.16
GO:0050679	positive regulation of epithelial cell proliferation	-3.16
GO:0071773	cellular response to BMP stimulus	-2.99
GO:0030855	epithelial cell differentiation	-2.91
GO:0030501	positive regulation of bone mineralization	-2.79
GO:0030326	embryonic limb morphogenesis	-2.62
GO:0001958	endochondral ossification	-1.89
GO:0042391	regulation of membrane potential	-1.86
GO:0005089	Rho guanyl-nucleotide exchange factor activity	-1.85
GO:0030509	BMP signalling pathway	-1.85
GO:0060021	palate development	-1.85
GO:0035023	regulation of Rho protein signal transduction	-1.77
GO:0045892	negative regulation of transcription, DNA-templated	-1.73
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	-1.66
GO:0045668	negative regulation of osteoblast differentiation	-1.56
GO:0001649	osteoblast differentiation	-1.49
GO:0006898	receptor-mediated endocytosis	-1.44
GO:0010862	positive regulation of pathway restricted SMAD protein phosphorylation	-1.39
GO:0007275	multicellular organism development	-1.22
GO:0098869	cellular oxidant detoxification	-1.10
GO:0072562	blood microparticle	-0.59
GO:0046983	protein dimerization activity	-0.59
GO:0043547	positive regulation of GTPase activity	-0.49
GO:0006351	transcription, DNA-templated	-0.26

Table 5. 8 **Gene ontology (GO) terms for biological processes that are over-represented in up-regulated genes differentially expressed at 48hr post-seeding.**

GO	Term	log10 p-value
GO:0001046	core promoter sequence-specific DNA binding	-2.95
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	-2.60
GO:0003707	steroid hormone receptor activity	-2.53
GO:0043401	steroid hormone-mediated signalling pathway	-1.61
GO:0006367	transcription initiation from RNA polymerase II promoter	-1.58
GO:0045893	positive regulation of transcription, DNA-templated	-1.57
GO:0044212	transcription regulatory region DNA binding	-1.57
GO:0043565	sequence-specific DNA binding	-0.99
GO:0019899	enzyme binding	-0.55
GO:0003700	transcription factor activity, sequence-specific DNA binding	-0.48
GO:0003677	DNA binding	-0.21

Based on the fold change and log10 p-value the top 2 genes for both time points (24hr and 48hr) were picked. EYA1 is a tyrosine phosphatase and a transcriptional coactivator for the proteins in the SIX gene family. SIX genes are required for normal development and before birth is necessary for important issues in the second branchial arch such as the front and sides of the neck, eyes, ears, and kidneys. It also dephosphorylates 'Tyr-142' of histone H2AX and promotes efficient DNA repair via the recruitment of DNA repair complexes containing MDC1 (ref). if there’s an interaction between C1ORF112 it might be an inhibitory one as C1ORF112 is downregulated EYA1 is upregulated however, the direction of interaction is still unclear (Cook, Ju et al. 2009). GALNT5 catalyses the initial reaction in O-linked oligosaccharide biosynthesis, as a transferase, it enables the transfer of N-acetyl-D-galactosamine residue to a serine or threonine residue on the protein receptor (Basu, Wang et al. 2015).

One of the top 2 down-regulated differentially genes is PRRX1, PRRX1 is another transcriptional coactivator from the homeobox family, that colocalise to the nucleus. PRRX1 is the regulator for MCK a kinase whose role is the establishment of diverse mesodermal muscle types by binding to an A/T-rich element in the muscle creatine enhancer. Aside from GALNT5, the other 3 DEGs from the micro-array study appear to be transcriptional factors, two of which are not co-expressed with C1ORF112 (PRRX1 and TFAP2D). Currently, it is still unclear if the DEGs and C1ORF112 are functionally connected but it is interesting that downregulation of C1ORF112 seems to influence transcriptional factors and enhancers. To understand the general levels of expression of these differentially expressed genes. The genes together with C1ORF112, TP53 and β-actin were analysed to ascertain their relative expression levels in normal tissue using the GTex data.

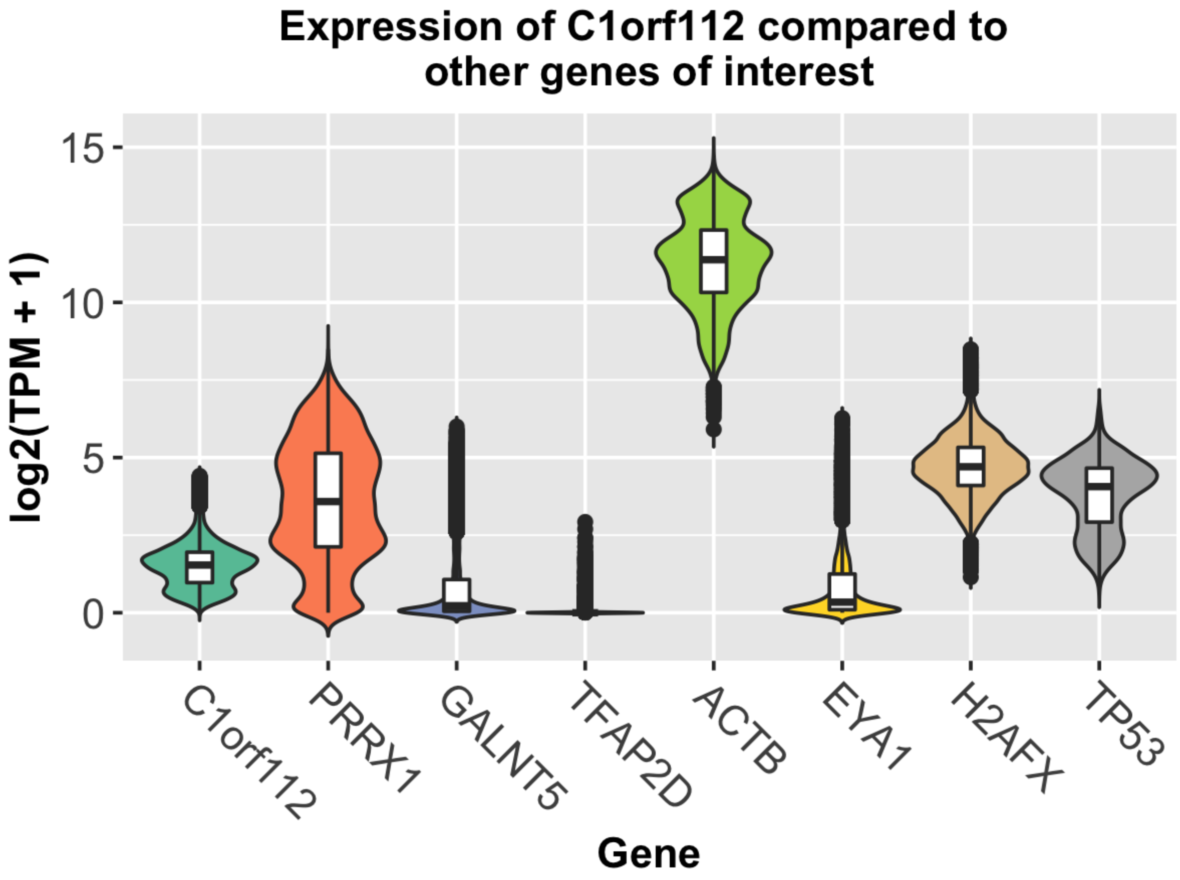


Figure 5. 5 **Overall expressions of C1ORF112 compared to genes selected from the microarray data and β-actin as a housekeeping gene**.

Showing β-actin having a higher level of expression compared to all the other genes. GALNT5 and EYA1 have lower expression levels compared to C1ORF112 as they as well as TFAP2D.

The overall expression level from the selected group of DEGs compared to C1ORF112 is shown in figure 5.5, GALNT5, TFAP2D, and EYA1 all exhibited similar levels of overall expression across all tissues and were consistently lower than the overall expression of C1ORF112. PRRX1 was the only gene with a higher overall expression level compared to C1ORF112. This result suggests that when C1ORF112 is down-regulated, PRRX1 and TFAP2D are also down-regulated and GALNT5 and EYA1 are up-regulated, however, this is just a hypothesis and further analysis would need to be carried out to determine if this is so. Actin was picked as it was expected to have a higher level of expression compared to C1ORF112 and used a high overall expression control. H2AFX and tp53 proteins whose role in cell cycle and DDR are well established and they were used as a secondary comparison to see how the expression of C1ORF112 and the other selected DEGs compared to them (Paull, Rogakou et al. 2000, Liu, Lu et al. 2007). None of actin, H2AFX and tp53 was differentially expressed from the micro-array analysis at both time points (24hr and 48hr). Although, there seems to be a change in the mRNA levels of these genes, there is a possibility that the actual level of proteins transcribed would be different even though it would be logical to think higher mRNA expression or lower mRNA expression would be reflected in the protein expression.

Gene ontology (GO) analysis of the DEGs did not reveal over-representation of genes involved in cell cycle, or DDR or chromosome stability gene at both time points for either the upregulated genes or downregulated genes. The most significantly over-represented biological process at 24hr for down-regulated genes involved in phosphatidylinositol-mediated signalling, and for up-regulated genes at the same time point were genes involved in ferric-chelate reductase (NADPH) activity. At the 48hr time point, the top GO biological process for downregulated genes over-represented was endocytic vesicle lumen and for upregulated genes, the top GO biological process over-represented was core promoter sequence-specific DNA binding. The GO terms for the down-regulated genes were mostly related to metabolic processes such as BMP signalling pathway and positive regulation of pathway restricted SMAD protein phosphorylation, DNA-templated transcription was also an over-represented GO term, however, it was the least significant based on log10 p-value (-0.26) at the 48hr time point.

To further understand the knock-down effect of C1ORF112 pathway analysis was carried out to determine the pathway most affected by C1ORF112

5.5 Pathways affected by knockdown of C1ORF112

DEGs were (shown in Figures 5.9 and 5.10), and how these affected their respective pathways was then analysed looking at the top and bottom affected pathways with the most and least DEGs respectively shown below in Tables 5.1 and 5.2. PI3K-Akt signalling pathway had the most DEGs with 26 in total, the same as Malignant pleural mesothelioma, the third pathway with the most DEG was the PI3K-Akt signalling pathway with mTOR, with 23 DEGs. PI3K-Akt signalling was also the most over-represented biological process for down-regulated DEGs at 24hrs. The PI3K-Akt signalling pathway is important in cell growth metabolism and survival, its activation is a multi-step process involving phosphoinositide-3-kinase (PI3K) (Korkolopoulou, Levidou et al. 2012, Spoerke, O'Brien et al. 2012). It is also a highly conserved and tightly controlled pathway that feeds into several other cell regulatory pathways, the differentially expressed genes because of C1ORF112 knockdown in the PI3K-Akt signalling pathway can be seen in Figures 5.6 – 5.10.

Table 5. 9 **Pathways with the most DEGs after C1ORF112 knockout**.

Pathways with the most DEGs shown in descending order

Pathway	**No of the Genes upregulated**	**No of the Genes downregulated**	**Total number of DEGs**
PI3K-Akt signalling	13	13	26
Malignant pleural mesothelioma	11	15	26
PI3K-Akt-mTOR signalling	12	11	23
Podnet: protein-protein interaction in podocytes	9	14	23
VEGFA-VEGFR2 signalling	14	9	23
Nuclear receptors meta-pathway	17	1	18
Mesodermal commitment	7	10	17
Circadian rhythm genes	12	5	17
Epithelial to mesenchymal transition in colorectal cancer	6	10	16
Ras signalling	10	5	15
Neural crest differentiation	4	10	14
Regulation of actin cytoskeleton	8	6	14
Gastrin signalling	6	7	13
Adipogenesis	10	3	13
Ectoderm differentiation	8	5	13
Male infertility	10	3	13
Hippo-Merlin dysregulation	5	8	13
Sudden infant death syndrome (SIDS) susceptibility	11	2	13
Orexin receptor	5	8	13
MAPK signalling	8	4	12

Table 5. 10 **Pathways with the least number of DEGs after C1ORF112 knockout.**

Pathways with the least DEGs shown in descending order

Pathway	**No of the Genes upregulated**	**No of the Genes downregulated**	**Total number of DEGs**
TCF dependent signalling in response to WNT	-	2	2
Fc epsilon receptor (FCERI) signalling	1	1	2
Processing of capped intron-containing pre-mRNA	1	1	2
Immunoregulatory interaction between lymphoid and non-lymphoid cells	1	1	2
RHO GTPases activate formins	-	1	1
Signalling by FGFR1,2,3	1	-	1
Hedgehog state	-	1	1
Complement system	-	1	1
Degradation of extracellular matrix	-	1	1
Signalling by type 1 insulin growth factor receptor	-	1	1
DDX58/IFIH1 mediated induction of interferon α/β	-	1	1
Protein folding	-	1	1
Signalling by NTRK1	1	-	1
Interleukin-1 family signalling	1	-	1
Factors involved in megakaryocyte development and platelet production	1	-	1
Circadian clock	1		1
Cell cycle	-	1	1
Asparagine n-linked glycosylation	-	1	1

Other notably affected pathways are Ras signalling, MAPK signalling, male infertility and SIDS and similar to the PI3K-Akt signalling pathway, the mechanism behind these are currently unclear but it does give insight into the possible role C1ORF112 might play with the presence of the PP1 docking site at the C-terminal region and possible kinase domain at the N-terminal region. The pathways with the lowest number of DEGs yielded some surprising results such as Cell cycle, Signalling by FGFR1,2,3, and TCF dependent signalling in response to WNT. These pathways are important, especially during early development. These pathways have at least 1 gene up or downregulated, for example, the cell cycle pathway shown in Figure 5.11, had the GADD45G downregulated. GADD45G is a protein that has several functions which included tumour suppression, cellular stress response and human-specific brain development. Lower levels of expression of GADD45G have been associated with a few cancer phenotypes (Zhang, Yang et al. 2014). These associations give insight into the potentially vital role C1ORF112 plays in the maintenance of cell proliferation.

This is in line with structural assessment as explained in Section 3, that, C1ORF112 could have several roles in cells as it is an alpha-helical protein.

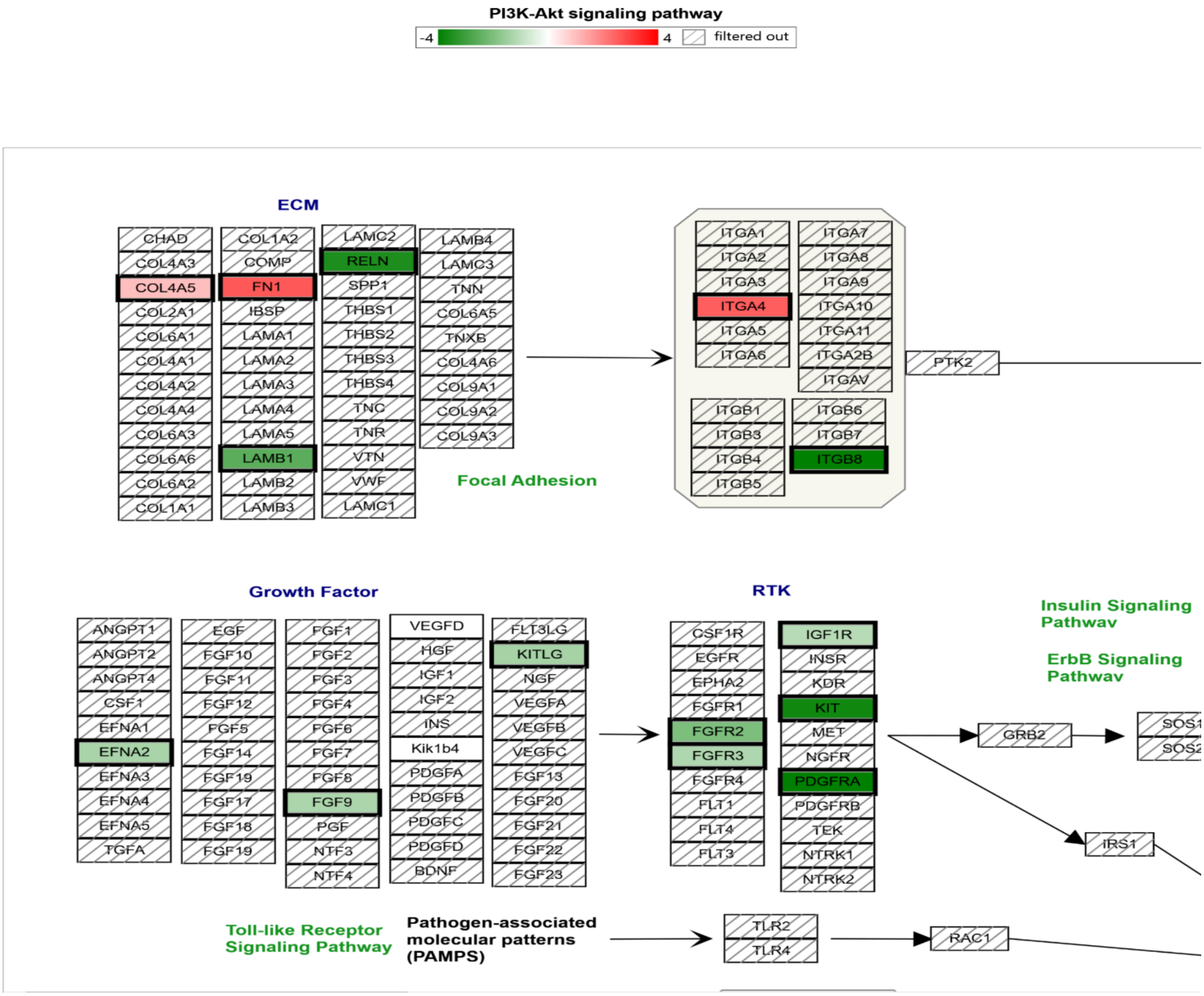


Figure 5. 6 **PI3K-Akt signalling pathway with the DEGs**.

Red indicates upregulated genes, and green indicates those that are downregulated. The extracellular matrix, growth factor, and receptor kinase (RTK) section of the PI3K-AKT signalling pathway.

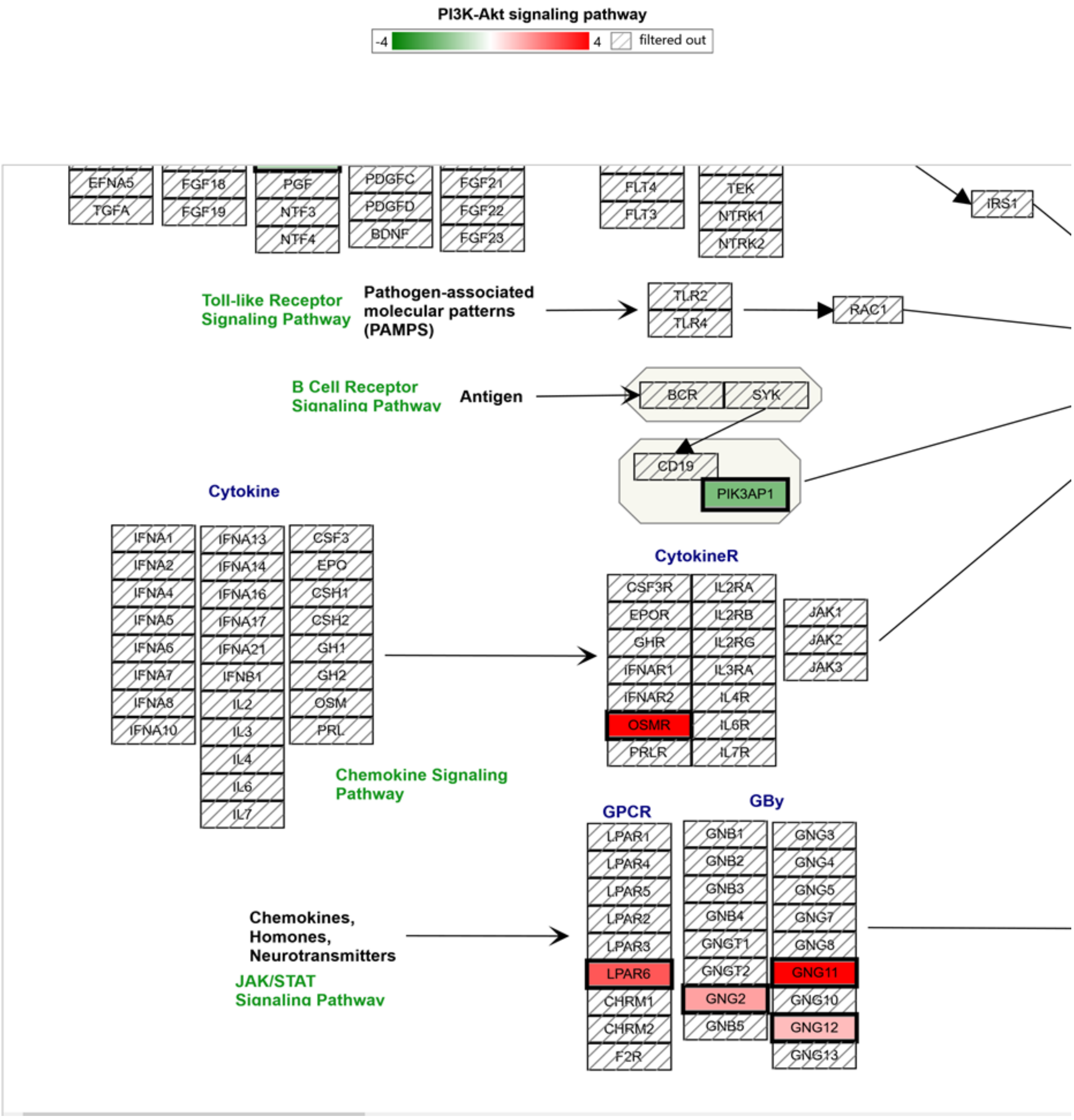


Figure 5. 7 **PI3K-Akt signalling pathway with the DEGs**.

Red indicates upregulated genes, and green indicates those that are downregulated. The cytokine section of the PI3K-AKT signalling pathway.

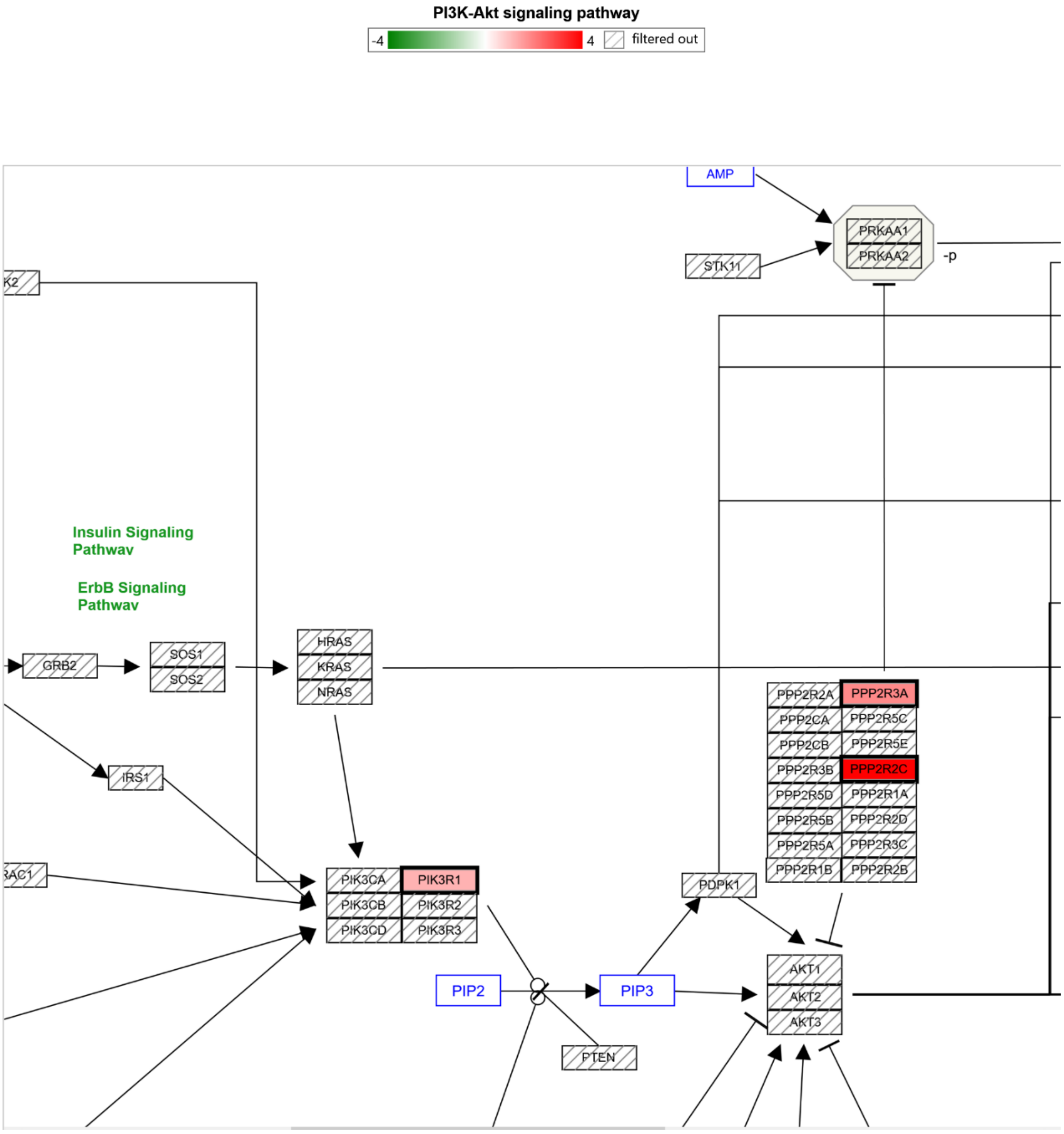


Figure 5. 8 **PI3K-Akt signalling pathway with the DEGs**.

Red indicates upregulated genes, and green indicates those that are downregulated. Signal transduction section of the PI3K signalling pathway, where the various sensors interact (inhibit or activate) with PTEN

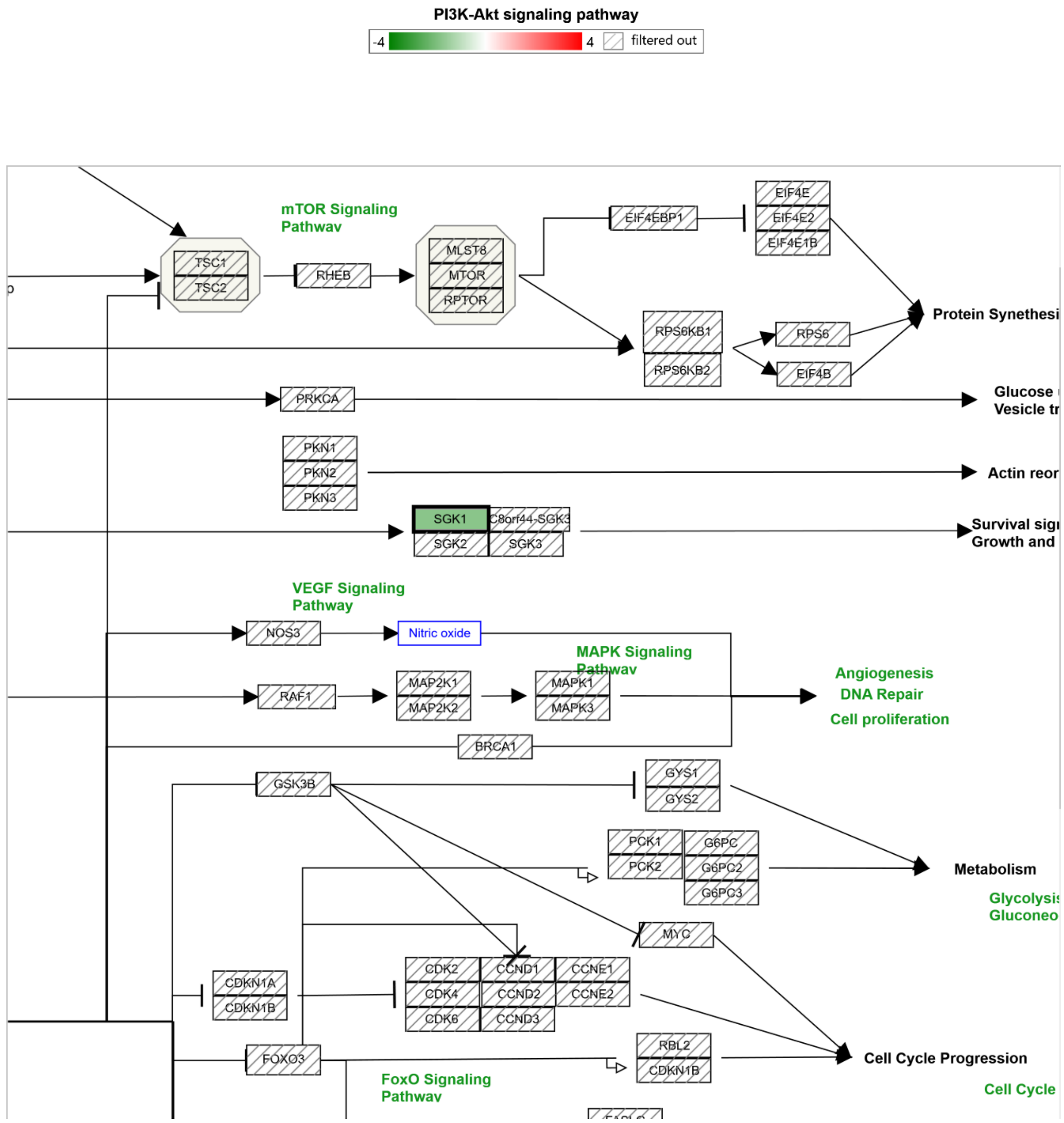


Figure 5. 9 **PI3K-Akt signalling pathway with the DEGs**.

Red indicates upregulated genes, and green indicates those that are downregulated. Effector end of the PI3K pathway feeding into various other pathways such as mTOR, VEGF and FoxO signalling pathways.

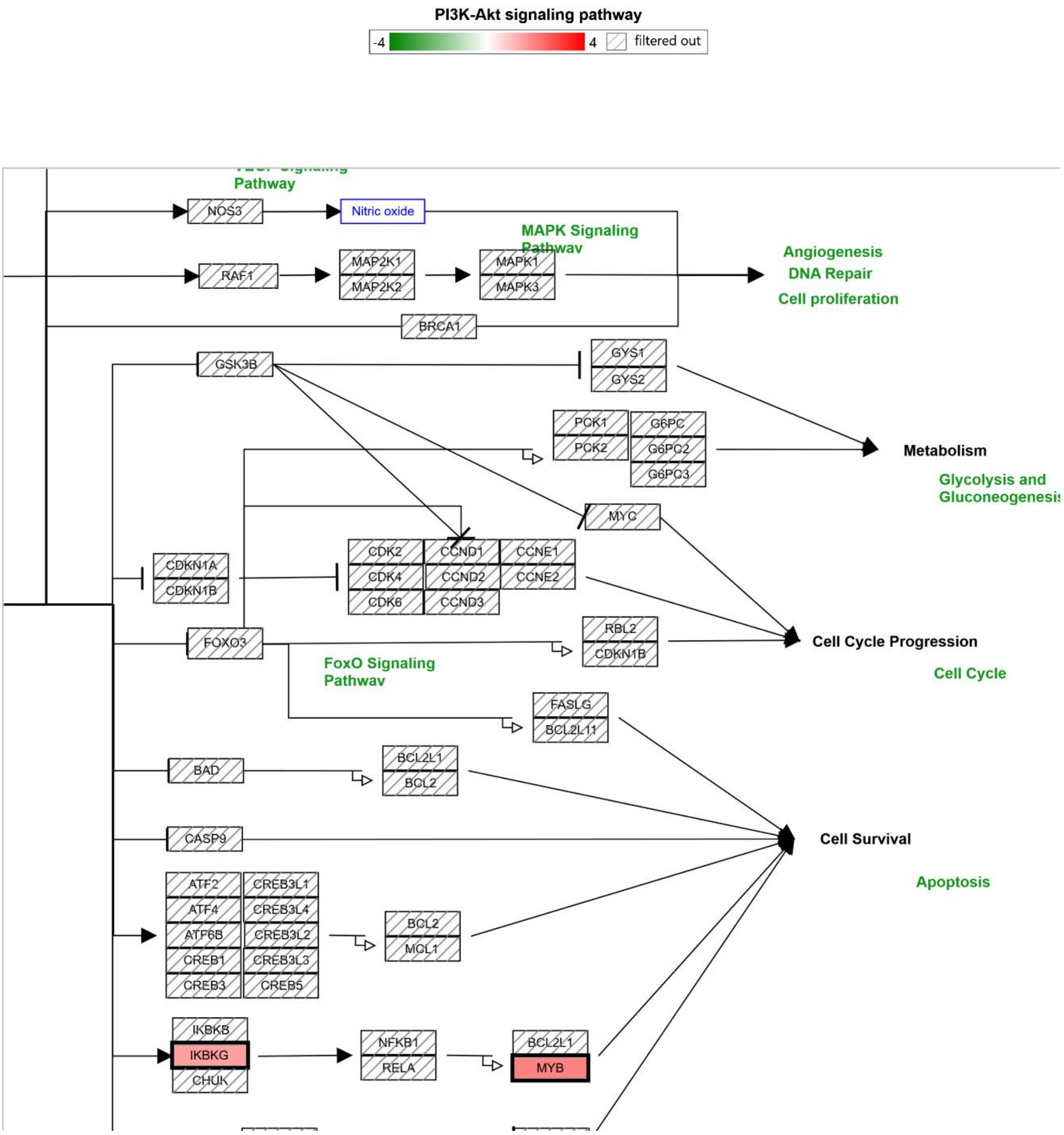


Figure 5. 10 **PI3K-Akt signalling pathway with the DEGs**.

Red indicates upregulated genes, and green indicates those that are downregulated. Effector end of the PI3K pathway leading to control of cell cycle progression, cell survival and metabolism

A total of 26 genes in the PI3K-AKT pathway were differentially expressed, genes such as lamb1 and RELN (figure 5.6) were downregulated. lamb1 plays a role in the attachment and migration of cells into tissues during embryonic development (Radmanesh, Caglayan et al. 2013). RELN, Reelin acts as a ligand for lipoprotein receptors and plays a role in the layering of brain neurons in the cerebral cortex, regulating microtubule formation and neuronal migration (D'Arcangelo, Homayouni et al. 1999).

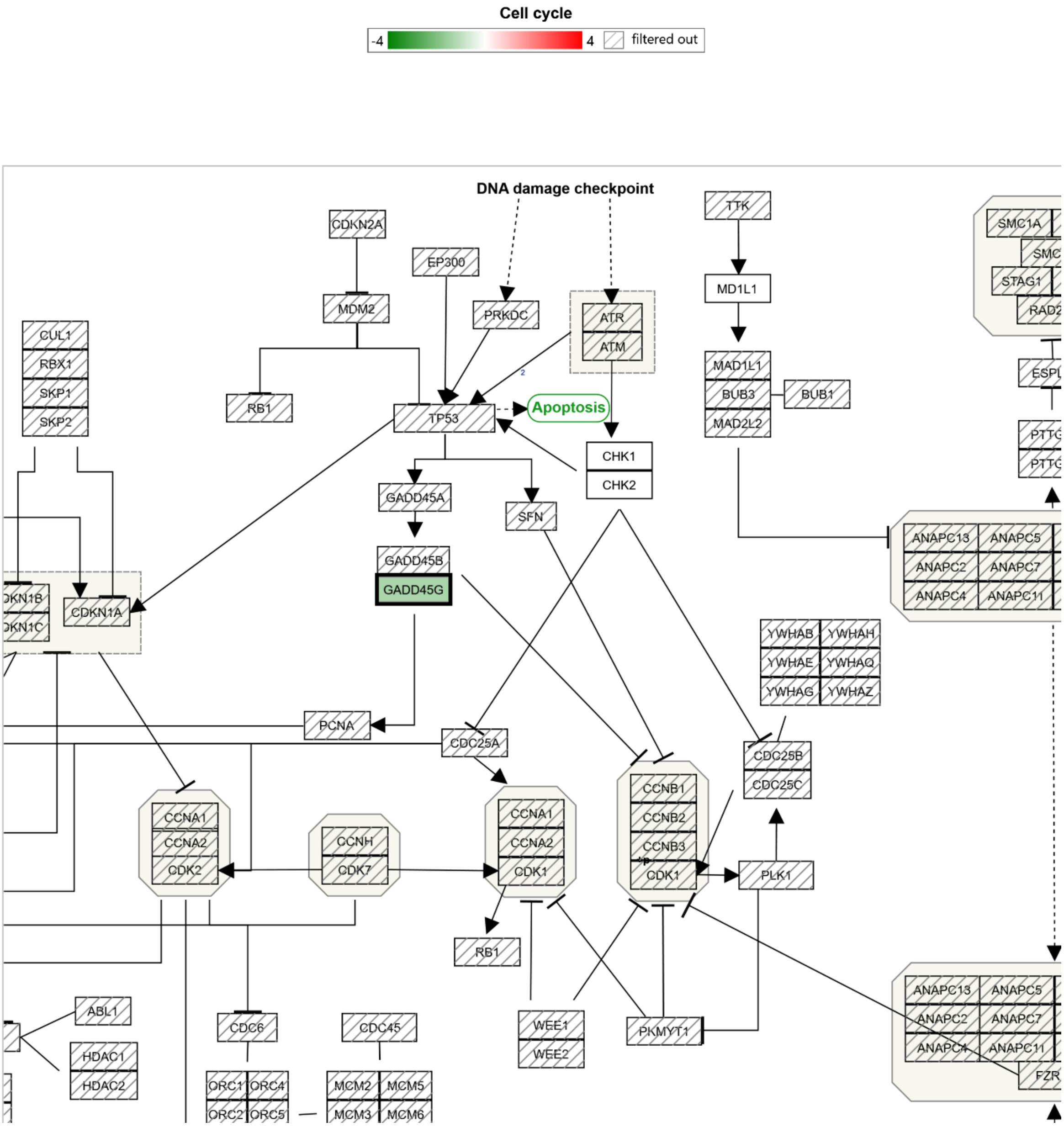


Figure 5. 11 **Cell cycle pathway**.

GADD45G is the only protein whose expression is downregulated in the pathway

The cell cycle pathway is one of the pathways with the least number of differentially expressed genes, and the only gene differentially expressed is GADD45G and it is downregulated (figure 5.11) mentioned earlier in section 5.5, GADD45G directly interacts with PCNA. PCNA proliferation cell nuclear antigen is an adjunct protein of DNA polymerase delta, it plays its role in the control of DNA replication by increasing the processing ability of the DNA polymerase during elongation, it also acts as a loading platform for recruiting DDR proteins following DNA damage repair to complete DNA replication (Cazzalini, Sommatis et al. 2014, Nicolae, Aho et al. 2014).

6.6 Discussion

C1ORF112 HAP1 knockdown cells were obtained commercially from the Horizon Discovery and used to determine the nature and effect of the C1ORF112 protein. Immunoblotting revealed that the mutant cell line produced no C1ORF112 protein product as shown in Figure 5.1, with a clear distinct band present just below the 100 kDa protein marker within the normal cells and no band present in the mutant cell line. This was then followed by cell growth analysis to determine if there was any difference between the normal HAP1 cells and the mutant C1ORF112 cells, of which there was no statistical difference between the normal and mutant cells at the 48hr and 72hr time points but there was a significant difference in the growth rate at the 24hr time between the normal HAP1 cells and the mutant C1ORF112 cells. The results then showed steady growth across both cell lines continuously over then time course. It is currently unclear the reason for the significant difference at the 24hr time point.

Microarray results showed that C1ORF112 was downregulated across both time points post-seeding, this was then followed by DEGs expression across both time points to evaluate the effect of the knockdown of C1ORF112 in the mutant cell line compared to the normal cell line. At the 24hr time point, there was a total of 730 DEGs of which 289 were downregulated genes and 441 were upregulated genes. The 48hr time point had fewer genes deferentially expressed, with a total of 470 DEGs, 189 downregulated and 281 upregulated genes. Functional enrichment for downregulated genes showed that they were involved in PI3K-Akt signalling, upregulated genes were more toward cell homeostasis. In addition, pathway analysis also supported the functional enrichment analysis, showing that in PI3K-Akt signalling (including mTOR) was the pathway with the most genes differentially expressed. PI3K-Akt signalling is a very important and highly conserved pathway in cells. Canonically, cell surface receptors such as growth factors, when activated can directly stimulate class 1A PI3Ks through the Ras protein, bound via their regulatory subunit or adapter molecules such as the insulin receptor substrate (IRS) proteins (Hemmings and Restuccia 2012). This results in a downstream cascade where PI3K is activated by conversion of its catalytic domain of phosphatidylinositol (3,4)-bisphosphate (PIP_2) lipids to phosphatidylinositol (3,4,5)-trisphosphate (PIP_3). PKB/Akt binds to PIP_3 at the plasma membrane, allowing PDK1 to access and phosphorylate T308 in the “activation loop,” leading to partial PKB/Akt activation (Alessi, James et al. 1997). However, the genes with changes in expression are not a part of the canonical pathway but do have a role in the efficacy of the pathway. In the PKB/Akt signal transduction pathway, the knockdown of C1ORF112 resulted in the downregulation of 11 genes in the ECM and Growth factor sections (figure 5.6) these genes play roles in sensing external chemokines that induce cell cycle activation. Downregulation in genes such as FGF2,3 and 9, PDGFRA could reduce the ability of the cell to induce G1 to S phase, as can be seen with the growth rate at 24hr for the mutant C1ORF112 cells, which had a growth rate of 0.8 cells per min compared to the normal HAP1 cells, that had a growth rate of 2 cells per min.

The Ras signalling pathway and MAPK signalling pathways also have genes differentially expressed because of the C1ORF112 down-regulation. These 3 pathways are known to have intersectional activities among them, and they control and maintain the cellular homeostasis and can switch activate or repress gene activation and determine the progress of cells in cycle and proliferation. Furthermore, the pathway with the least number of DEGs also appears to have a role in the cell cycle and proliferation. For example, cell cycle, hedgehog state and complement system are required in early development and although a singular gene might be affected in the pathway, the efficacy of the pathway is necessary for proper cell proliferation and development of the foetus. The mechanism behind C1ORF112 knockdown and the effect it has on this pathway is unclear and although it causes DEGs in the pathway, it does not show any disrupt the signalling pathway such that cell proliferation is inhibited significantly when compared to the normal cells Section 5.3**.**

To understand the mechanistic role of C1ORF112 both the normal and mutant cell lines were exposed to two different types of stress. Chemical stress in the form of hydrogen peroxide (H_2O_2) and exogenous stress in the form of X-rays. The aim was to determine if there was the sensitivity of the mutant cells compared to the normal cells as both forms of stress elicit DNA damage. This was also to ascertain if C1ORF112 [played a role in DNA damage repair or response.

Chapter 6: Results IV: C1ORF112 knockout cells are sensitive to x-rays and hydrogen peroxide treatments compared to normal cells

6.1 Introduction

The role of C1ORF112 in response to DNA damaging events from x-rays and hydrogen peroxide (H_2O_2) is currently unknown, most studies involving C1ORF112 have been correlative studies and its mechanism of action is currently unclear. X-rays and hydrogen peroxide (H_2O_2) are two forms of agents that generate DNA damage, specifically double strand breaks (DSBs) (discussed in sections 1.1.4 – 1.1.6). These agents were chosen as the aim was determine the effect of C1ORF112 loss, on double strand DNA repair, although there is a stronger connection to genes acting in the Fanconi anaemia pathway, it is also worth noting that these agents also induce single strand breaks to a greater frequency compared to DSBs. X-rays are electromagnetic radiation that consists of photons, with a similar spectrum to gamma rays the major difference being the source of the radiation, with gamma rays from radio nuclear radiation and x-rays from the photoelectric effect. Although initial exposure of cells to x-rays leads to direct DNA damage (e.g., SSBs and DSBs), x-rays travelling through tissues majorly lead to secondary ionizations due to subsequent generation of electrons after the initial photons interact with atoms. Therefore, the biological effect of x-rays comprises not only the initial radiation event, but mostly secondary electrons produced (Borrego-Soto, Ortiz-Lopez et al. 2015). This is important due to the clinical use of x-ray in radiotherapy and cancer treatment and the use of lower energy radiation for imaging in dentistry and other medical practices.

Hydrogen peroxide, on the other hand, induces chemical oxidation events in cells and this can either be endogenously or exogenously, that is, through the action of metabolic processes within the cells or external exposure to the compound. Medically, hydrogen peroxide is used as a disinfectant in wound cleaning, sterilization of medical equipment which can lead to exposure, which in cells can lead to various disorders such as Stress-induced premature senescence (SIPS), base oxidation, SSBs and DSBs, through the generation of ROS species like superoxide radicals ($O_2•^-$) and the hydroxyl radical (•OH). H_2O_2 alongside radiation exposure were used to determine the sensitivity of C1ORF112 mutant cells (KO) compared to the normal cells.

In chapter 5, I showed that the knockdown of C1ORF112 did not affect the ability of the HAP1 cells to grow after 48hr time point, meaning C1ORF112 seemed to delay onset cell growth and proliferation, but the micro-array analysis did not reveal significant downregulation of cell cycle-related proteins. To further understand the role of C1ORF112, based on the co-expression results, I hypothesized that the KO cells might be sensitive to DNA damaging agents and particularly through induction of DSBs.

These experiments aimed to determine the sensitivity and cell survival capabilities of both normal and C1ORF112 knockout HAP1 cells to various concentrations of H_2O_2 and x-ray and to determine if there was any significant difference between the cells. I also looked at the repair kinetics after induction of DSBs using both H_2O_2 and x-rays, to determine if C1ORF112 was in the cytoplasm or the nucleus after radiation exposure.

6.2 C1ORF112 is a cytoplasmic protein

To understand the cellular localisation of C1ORF112 before and after stress, both mutant HAP1 cells were irradiated at 10Gy x-rays to establish whether C1ORF112 was a nuclear protein, a cytoplasmic protein, or a cytoplasmic protein that was recruited to the nucleus after stress. The dose 10gy was chosen to generate a high level of DNA damage, which would force the cells to recruit most of the DDR proteins to the nucleus. The C1ORF112 protein was then monitored over 4hr to establish protein levels and possible recruitment kinetics to the nucleus.

(a)

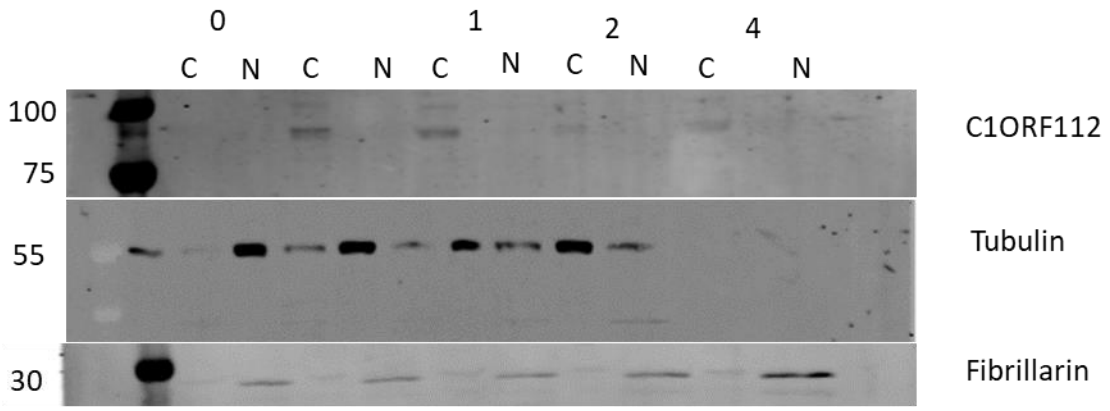


Figure 6. 1 **Cell fractionation of mutant HAP1 cells indicating cytoplasmic presence of C1ORF112**.

(a) showing the C1ORF112 did not move across the nuclear membrane to assist with DDR fibrillarin used as nuclear control and tubulin used as cytoplasmic control.

Although the bands on the immunoblot of the C1ORF112 protein are faint, the cell fractionation assay showed that C1ORF112 did not cross the nuclear membrane to be recruited as part of the DSB DNA repair complex. This indicates that C1ORF112 might not have a significant direct role in DNA repair as it is unable to cross the nuclear membrane after the induction of high levels of DNA damage. I then looked at the sensitivity of the mutant cells to normal cells in response to both x-ray and H_2O_2.

6.3 C1ORF112 mutant cells are sensitive to x-rays compared to C1ORF112 normal cells

Seeding densities for both normal and mutant cells into 6 wells plates for 0Gy (no treatment) were at 500 and 1000 (figure 6.2) and these seeding densities were doubled with an increase in radiation dosage (1,2 and 4Gy). The plates were then incubated for 7 days and then surviving cells were counted and normalised to adjust for each cell line. Representative images for seeded plates are shown in figure 6.2, indicating that with increasing dosage of x-ray the were fewer surviving cells in both normal and mutant cell lines.

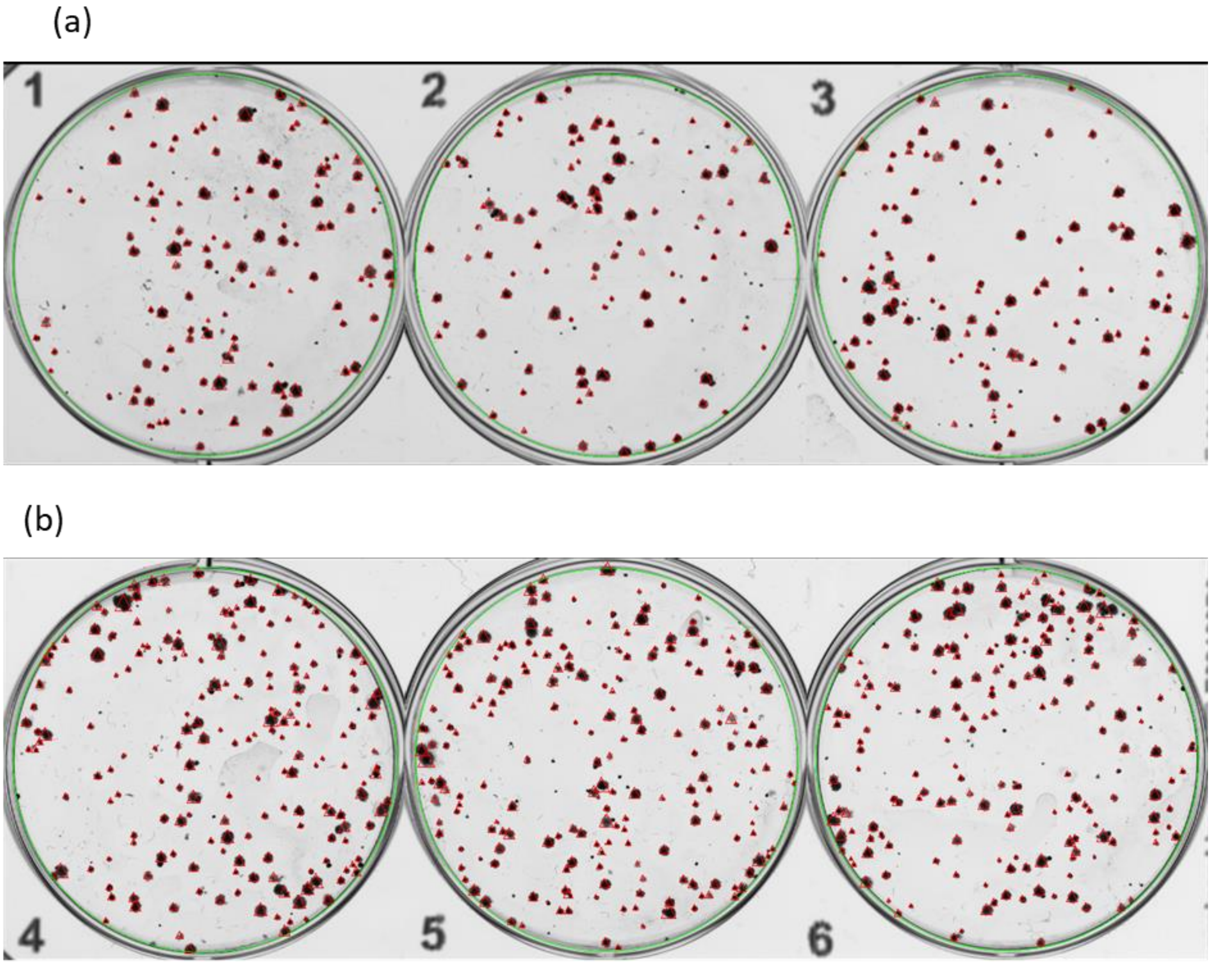


Figure 6. 2 **Representative images of clonogenic seeding densities**.

(a) seeding density for normal cells at 500 cells per well (b) seeding density of normal HAP1 cells at 1000 cells.

The seeding density test performed for the normal C1ORF112 cells and left for 7 days to form colonies showed a plating efficiency of>70% at each seeding density with the treatment of either x-ray or H_2O_2. After establishing the optimal seeding density and plating efficiency, the normal and C1ORF112 mutant cells were then treated with increasing radiation doses (1,2 and 4Gy) before seeding and incubation for 7 days. The plating efficiency for normal C1ORF112 cells treated with 1Gy x-ray was between 20 – 50%, with 2Gy was <20% plating efficiency and for 4Gy, the plating efficiency was less than 1%. The plating efficiency for the C1ORF112 mutant cells was the same across the x-ray dosage.

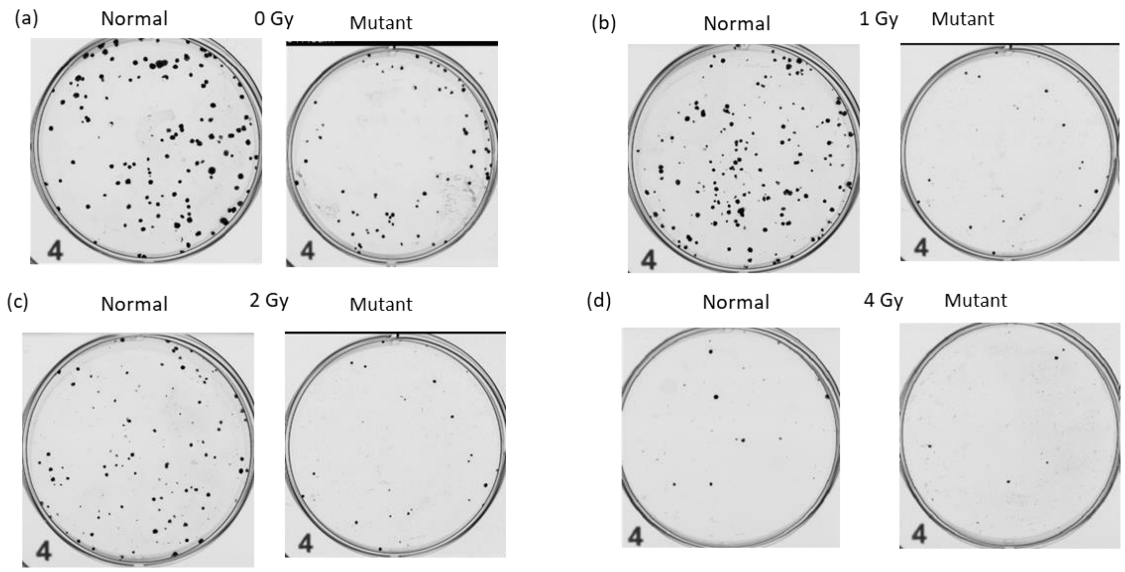


Figure 6. 3 **Representative plates showing a reduction in cell colonies after treatment with x-rays.**

(a) no treatment (b) 1Gy x-ray (c) 2Gy treatment (d) 4Gy treatment. All experiments were carried out in 3 biological triplicates across different days and 3 technical replicates across different plates

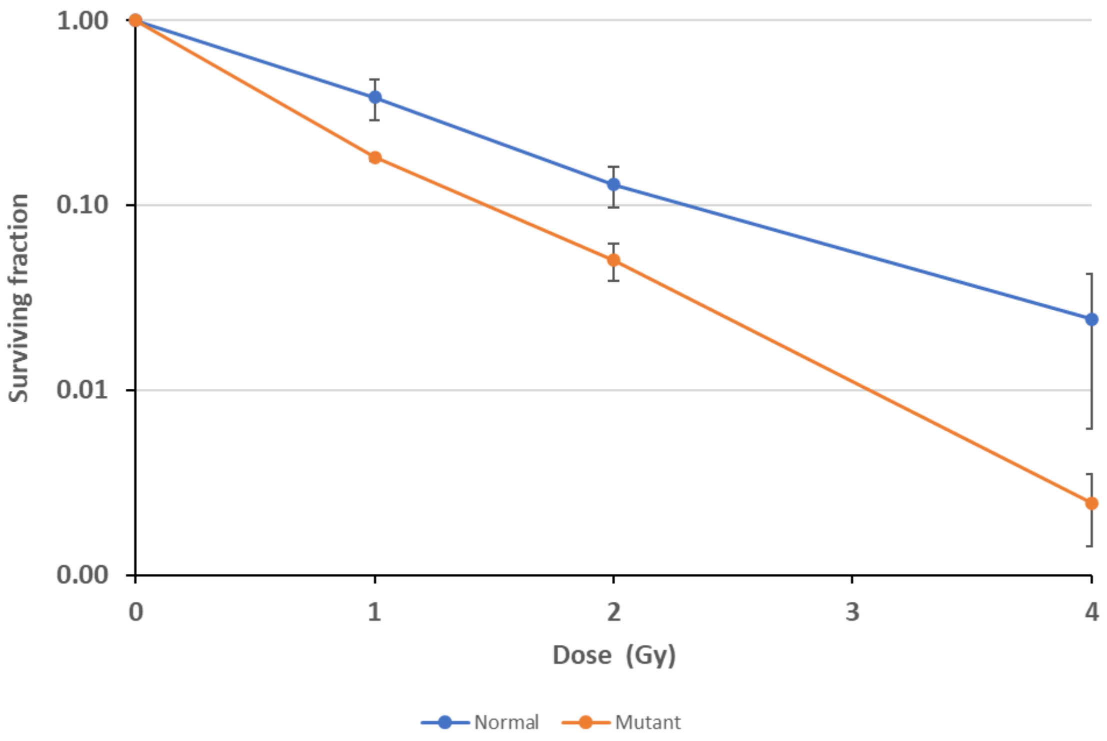


Figure 6. 4 **Sensitivity of mutant HAP1 cells to x-rays**.

Linear regression analysis shows the sensitivity of mutant C1ORF112 cells to x-ray treatment compared to normal C1ORF112 cells.

C1ORF112 mutant HAP1 cells are more sensitive to x-rays compared to normal HAP1 cells as shown in Figure 6.3, with increasing levels of x-ray exposure, normal cells fared better compared to mutant cells. The normal cells showed a steady decline in cell survival, at the 1Gy dose, the cell survival fraction was about 0.5 and at 2Gy the survival fraction was still about 0.1. in contrast, the survival fraction for the mutant HAP1 cells after exposure to 1Gy x-ray was close to 0.10, which is much steeper compared to the normal HAP1 cells. After 2Gy exposure the survival fraction of the mutant C1ORF112 cells is below 0.1, indicating the mutant cells are more sensitive to x-ray treatment compared to the normal C1ORF112. Regression analysis showed there was a significant difference between normal and mutant cells with a p-value of .0002553 (Braselmann, Michna et al. 2015). This was then followed by a comet assay looking at the generation of DSBs when exposed to x-ray radiation to determine if mutant cells induced different levels of DSBs compared to normal cells and their capacity to repair the damage.

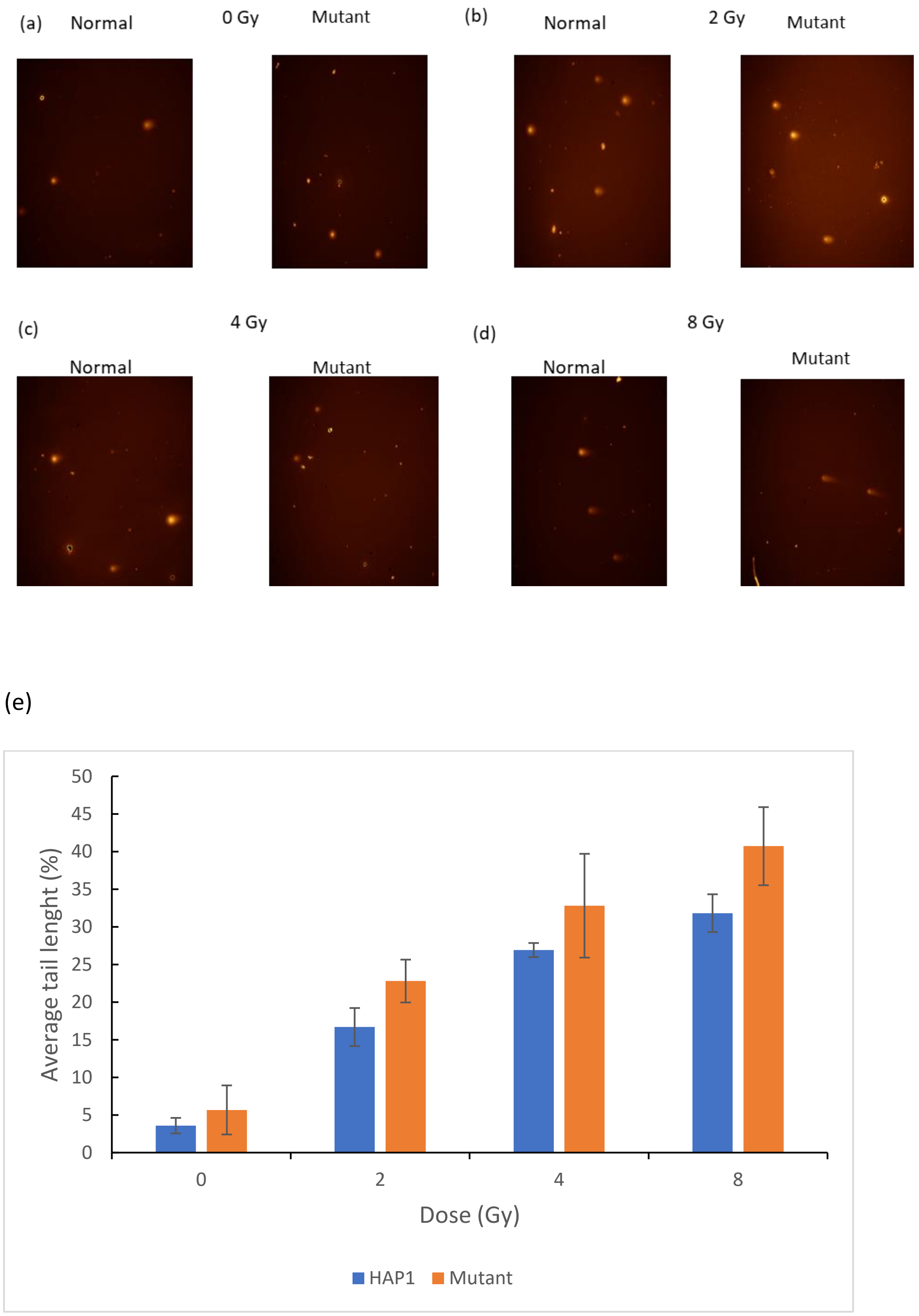
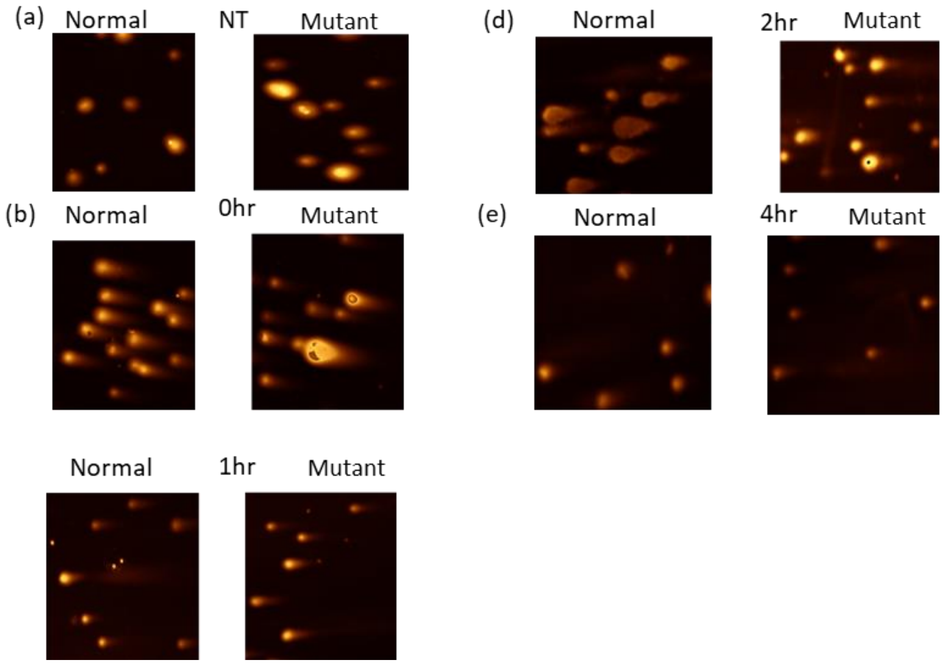


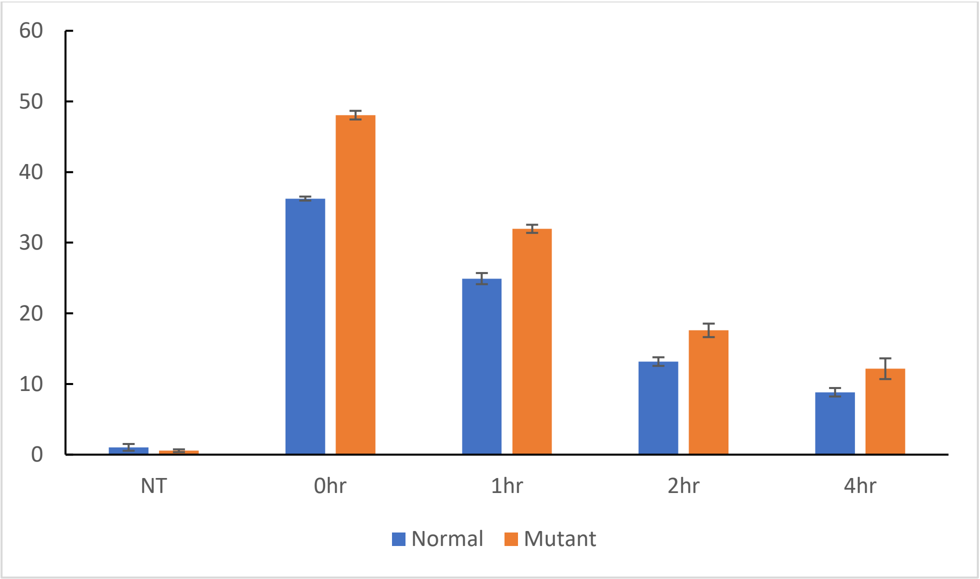
Figure 6. 5 **Generation of DSBs after x-ray exposure.**

(a-d) Generation of DSBs with increasing levels of x-rays (0-8Gy). (e) The amount of DSBs inducted is measured as tail length.

The neutral comet was performed to ascertain the level of induction of DSB after exposure to x-rays, after establishing that mutant cells were more sensitive to the x-ray treatment compared to the normal cells. Both cells were then exposed to increasing doses of x-ray (2-8gy) the cells were embedded on the agarose slides as (described 2.3.15) and ran to determine the level of induction of DSBs generated depending on the dose of x-ray the cells were exposed to. As seen in figure 6.5 (a-d) the amount of DSBs generated increased with an increase in the dosage of x-ray Figure 6.5e, although 0Gy (i.e., not treatment) did see some level of DSB, this could be attributed to the mixing of the cells with agarose before embedding them on the glass slides. The average tail length showed that there were increasing levels of DNA damage with an increase of exposure, however, there was no significant difference between the amount of DSBs induced up until 8Gy, The mutant C1ORF112 cells did show higher induction of DSBs across all dosages for x-rays, compared the normal HAP1 cells (shown as average tail length figure 6.5e). The survival fraction results alongside the dose titration of x-ray showed that 4Gy x-ray could induce a large amount of DSBs without being completely lethal to the cells, and therefore used as the dosage for the repair kinetic studies.


(a)
Normal
NT
Mutant
(d)
Normal
2hr
Mutant
(b)
Normal
0hr
Mutant
(e)
Normal
4hr
Mutant
Normal
1hr
Mutant

(f)


60
50
40
30
20
10
0
NT
0hr
1hr
2hr
4hr
■ Normal
■ Mutant

(g)

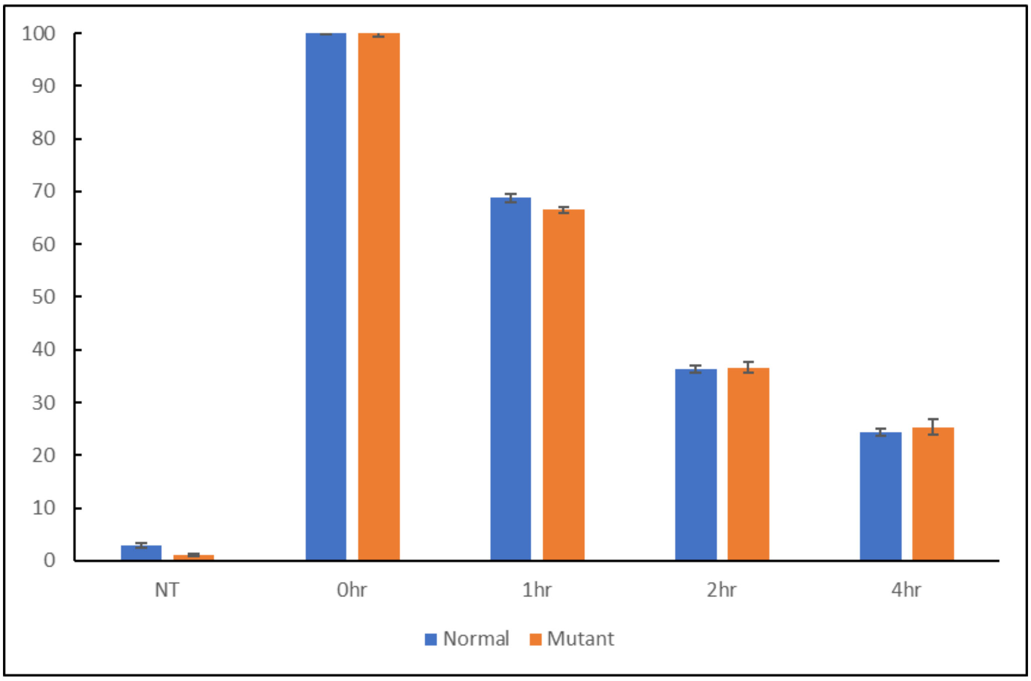


Figure 6. 6 **Repair kinetics after generation of DSBs after exposure to x-rays**.

(a-e) representative slides of the repair kinetics, showing a decrease in DSBs post-exposure (f) Repair kinetics for normal HAP1 and mutant C1ORF112 after exposure to 4Gy x-ray dose. (g) Normalised repair kinetics

The repair kinetics was carried out, after exposing both the normal HAP1 cells and the mutant C1ORF112 cells to 4Gy x-ray. The cells were then given time to repair at various time points (1hr – 4hr, Figure 6.6). Figure 6.6f shows the non-normalised result of the repair kinetics with the mutant C1ORF112 having higher levels of DSBs immediately after exposure to 4Gy x-ray compared to the normal HAP1 cells. Given time for the cells to repair, there was a steady decline in the number of DBSs observed. To determine if there was any significant difference in the repair kinetics, the average tail length was normalised (figure 6.6f), to the 0hr time point (no repair mechanism initiated), the p-value across all time points showed no significant difference in the repair kinetics between the normal HAP1 cells and mutant C1ORF112 cells.

C1ORF112 mutant cells might be more sensitive to x-rays compared to the normal cells, however, when the repair kinetics is considered. There is no significant difference between normal and C1ORF112 cells. The was then repeated using hydrogen peroxide, a different form DNA damaging agent.

6.4 C1ORF112 mutant cells are sensitive to hydrogen peroxide compared to C1ORF112 normal cells

Similar, to the x-ray assay plates, were seeded from 500/1000 cells per well, doubling per increase in peroxide dose.

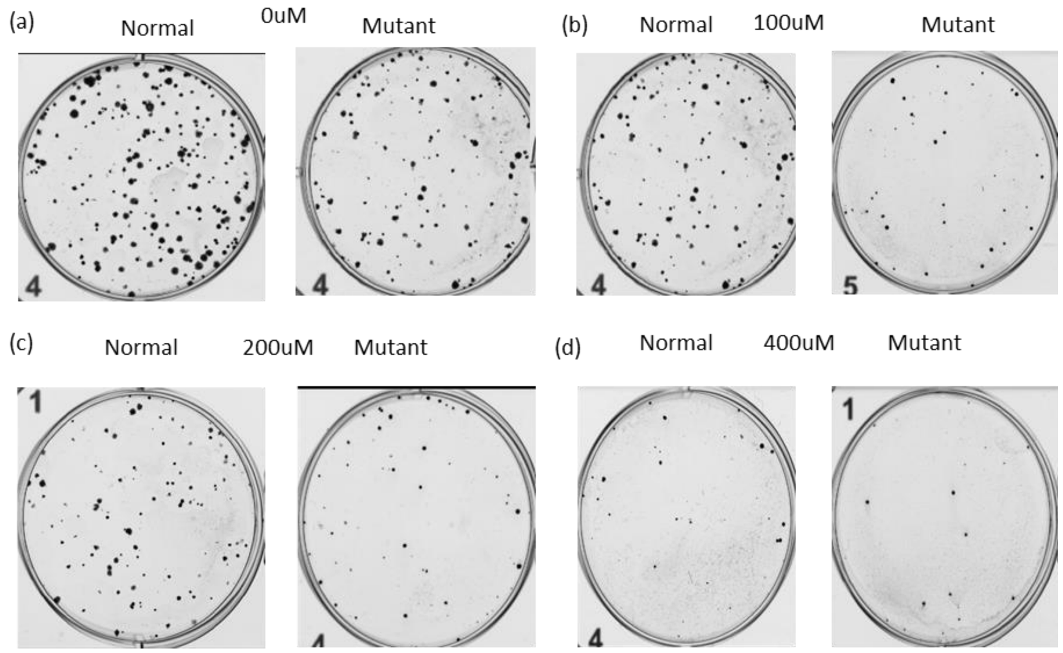


Figure 6. 7 **Representative plates showing a reduction in cell colonies after treatment with hydrogen peroxide.**

(a) no treatment (b) 100uM (c) 200uM (d) 4 00uM. All experiments were carried out in 3 biological triplicates across different days and 3 technical replicates across different plates.

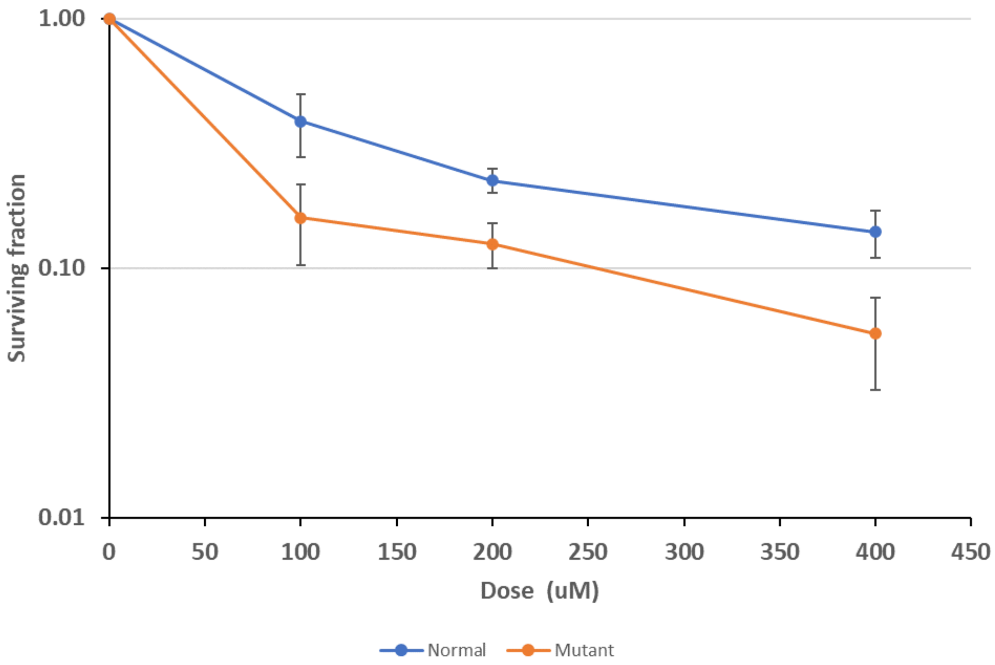


Figure 6. 8 **Sensitivity of mutant HAP1 cells to hydrogen peroxide**.

(Top) representative plates showing a reduction in cell colonies and the surviving fractions after normalisation (bottom).

The mutant cells also showed they were more sensitive to hydrogen peroxide compared to the normal. The mutant C1ORF112, in this case, had a survival fraction above 0.1 for both the 100uM and the 200uM concentrations compared to the normal HAP1 cell, whose survival fraction was also higher than 0.1 at both dosages but also higher than the survival fraction for the C1ORF112 mutant cells. Regression analysis showed there was a significant difference between normal and mutant cells with a p-value of p-value 5.31e-05. Again, the comet assay was performed to ascertain, the level of DSBs induced and if the was any significant difference between the mutant cell and the normal cells and the repair kinetics between them.

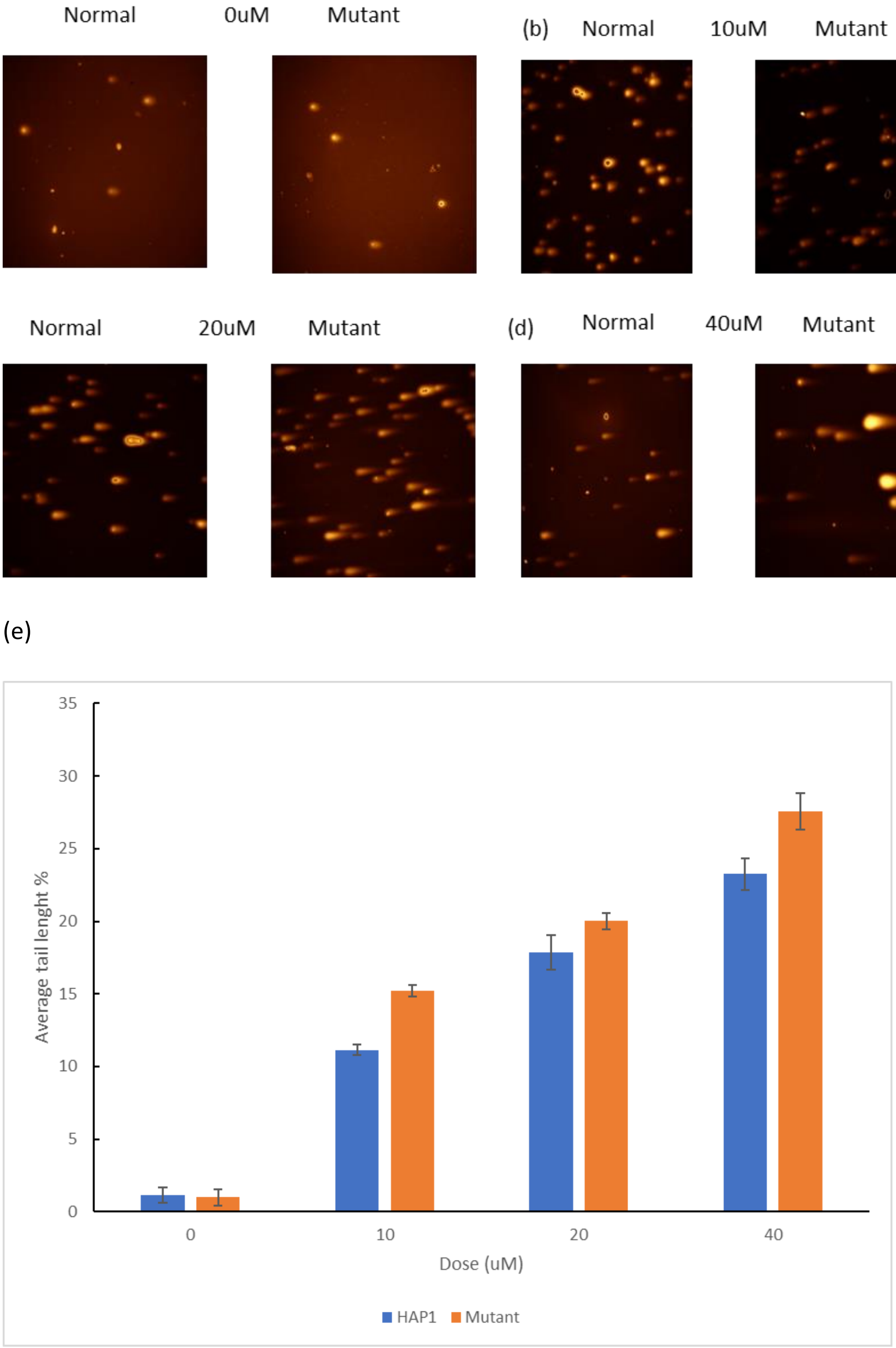
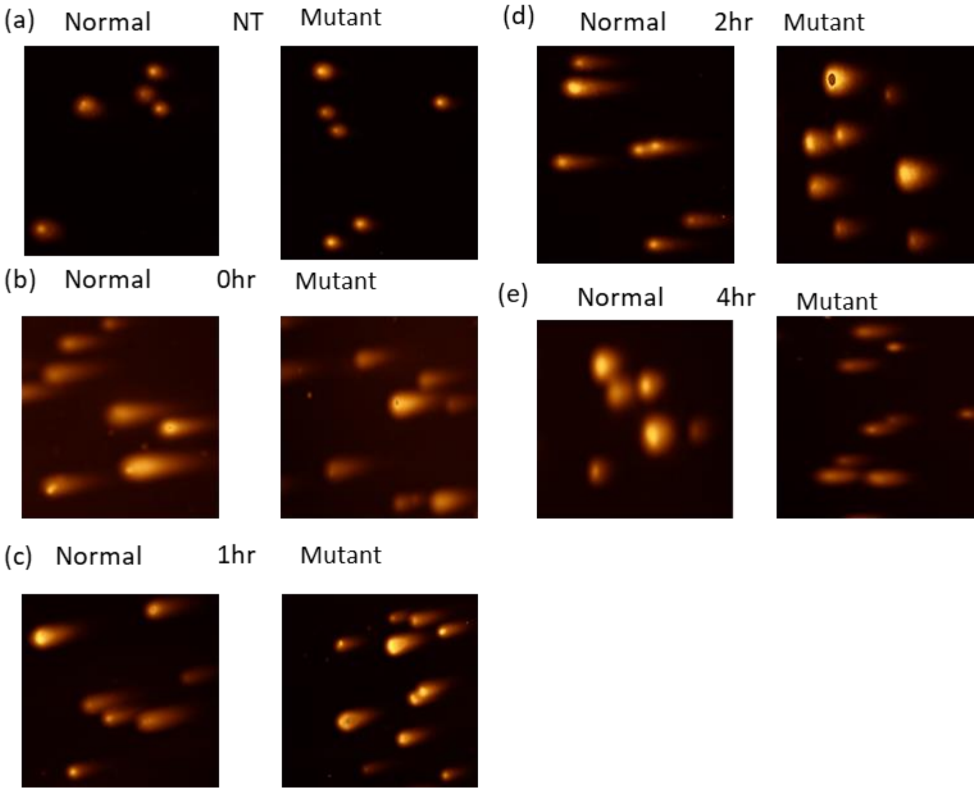


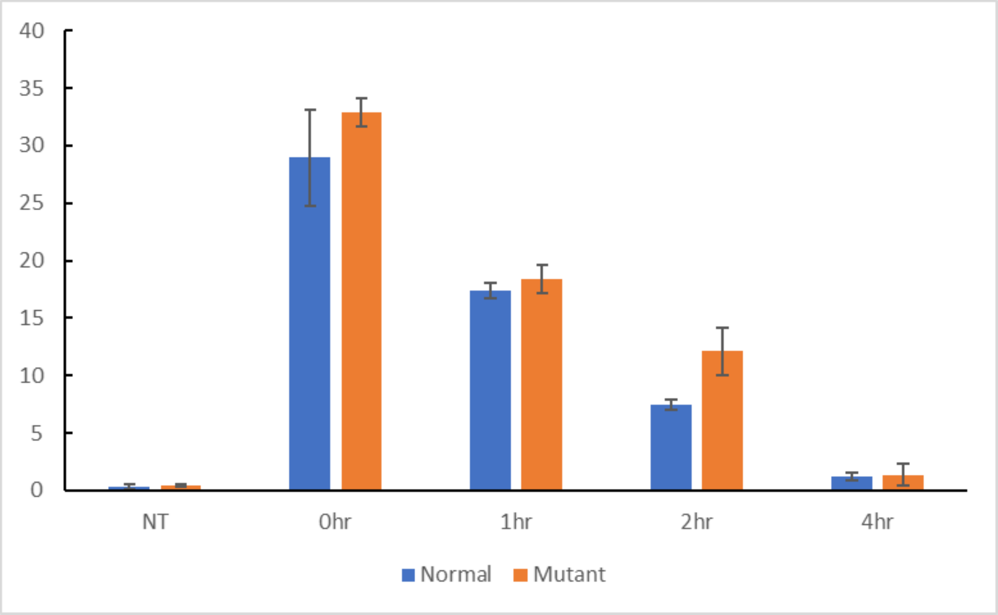
Figure 6. 9 **Generation of DSBs after hydrogen peroxide exposure.**

(a-d) Generation of DSBs with increasing levels of peroxide (10-40uM). (e) The amount of DSBs inducted is measured as tail length. All experiments were carried out in 3 biological triplicates across different days and 3 technical replicates across different plates

After establishing that C1ORF112 mutant cells were again sensitive to exposure to hydrogen peroxide, a neutral comet assay pipeline was performed to determine the level of induction of DSBs at various concentrations like the x-ray treatment. The results showed that C1ORF112 mutant cells induced higher levels of DSBs compared to the normal cells. The 20uM concentration was then determined to be suitable to be used for the repair kinetics as it induced enough DSBs to be well observed (figure 6.10c), without being lethal to the cells.



(f)



(g)

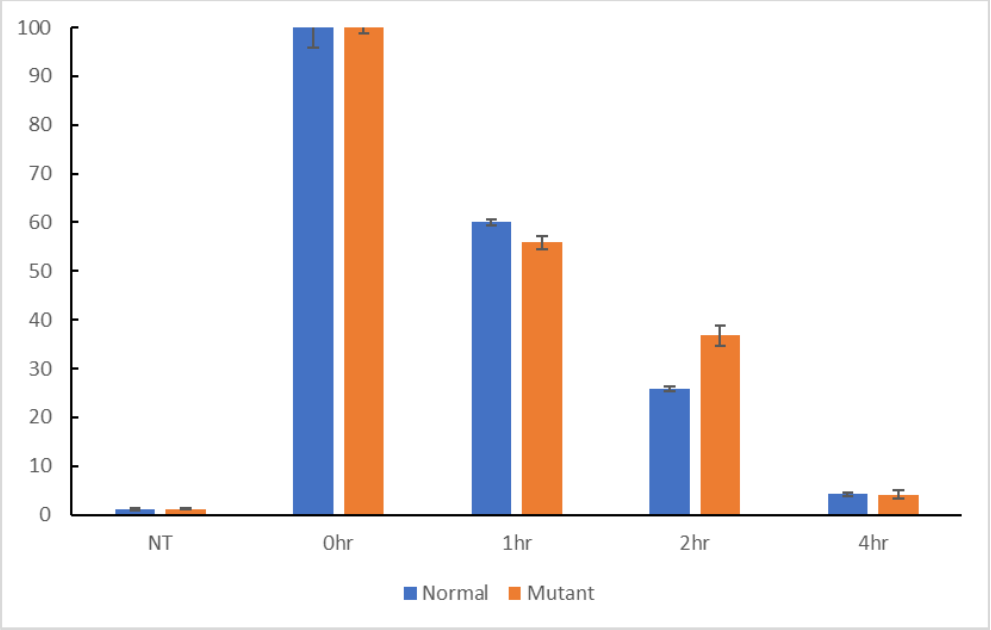


Figure 6. 10 R**epair kinetics after exposure to hydrogen peroxide**.

(a-e) representative slides of the repair kinetics, showing a decrease in DSBs post-exposure. (f) the average tail length for non-normalised repair kinetics (g) the normalised average tail length for the repair kinetics. All experiments were carried out in 3 biological triplicates across different days and 3 technical replicates across different plates

The results of the repair kinetics after exposure to 20uM of hydrogen peroxide showed that at 0hr before the induction of the repair mechanism, C1ORF112 mutant cells induced higher levels of the DSBs like the dose titration, and as the DSB repair mechanism was given time to repair the DSBs the observed DSBs reduced according to for both the normal and mutant cells (figure 6.11f). to determine if the was any significant difference in the repair mechanisms between the mutant C1ORF112 cells and the normal cells, the data were normalised (figure 6.11g) and the was no significant difference at the 1hr and the 4hr time points (p-value >0.05) but at the 2hr mark, the was significance in the repairability between the mutant and the normal cell lines, with the normal cells repairing better compared to the mutant cells (p-value =0.04). it is still unclear as to why the significance in repair presents itself at the 2hr time point.

6.5 Discussion

C1ORF112 protein in the normal HAP1 cells, after exposure to high levels of x-ray radiation, did not show the ability to cross the nuclear membrane in response to induction of DSBs (figure 6.1), this suggests that C1ORF112 may play an indirect role in DDR. On exposure to x-ray radiation, the mutant C1ORF112 cells appear to be more sensitive compared to the normal cells (figure 6.3 and 6.4), as the surviving fraction for the normal cells were significant compared to the surviving fraction for the C1ORF112 mutant cells. Similarly, exposure to increasing dosage of x-ray showed induction of higher levels of DSB in the mutant C1ORF112 cells (figure 6.5) compared to the normal cells, however, the repair kinetics does not show any significant difference in repair mechanism between the mutant C1ORF112 and the normal HAP1 cells (figure 6.6). This is also the case when the C1ORF112 cells are exposed to hydrogen peroxide, the number of surviving colonies of normal HAP1 cells as compared to C1ORF112 knockout cells showed significant differences across time points and levels of concentration figures 6.7 & 6.8.

In addition, the induction of DSBs in mutant C1ORF112 cells when compared to the normal cells for peroxide was largely like the results of the x-ray treatment. The main difference was in the repair kinetics and the 2hr time point for the peroxide treatment showed a significant difference with the normal cells repairing better compared to the mutant cells (p-value 0.04). Aside from that time point, all the other time points showed a significant difference in the repair kinetics. These results alongside the cell fractionation assay indicate that C1ORF112 may not play a direct role in DDR itself but might have an indirect role.

Overall, C1ORF112 is a cytoplasmic protein whose down-regulation does sensitize the cells to different forms of DNA damage, specifically x-ray and hydrogen peroxide. At present it does not show any direct role in DDR, however, it may play an indirect role in DDR by activating other proteins that are directly involved in the DDR, based on the co-expression results (discussed in section 3.6), with proteins with a role in cell cycle regulation, proliferation, and DDR. The sensitivity could also be because of the significant difference observed at the 24hr time point for the cell growth rate (see section 5.3), however, the mechanism is still unclear and would require further investigation.

7: Discussion

C1ORF112 is an evolutionary conserved gene and is well expressed in tissues

C1ORF112 is a well-conserved evolutionary gene from mammals to choanoflagellates very close relative of the Metazoa, present also in plants and certain species of worms and fungi. C1ORF112 is not present in scientific model organisms and the reason this is so, is currently unclear. Although it is not present in Drosophila melanogaster, it is present in insects such as ants, beetles, and weevils. This is similar in Nematodes as it is not present in Caenorhabditis elegans, but it is present in platyhelminths such as tapeworms and liver flukes. The conservation of C1ORF112 has shown that it is possible that the gene could have been lost in these model species while being retained in those of the platyhelminths. However, the sequence conservation is quite low with about <30 sequence similarity, and the presence of the WCF tripeptide used to categorize the sequences in the domain of unknown function DUF4487 is also absent. This makes it difficult to determine if they are true homologs without any established functional data to infer from. In addition, the lack of a functional homolog within a species, with an established function also makes it difficult to ascertain if these sequences are true homologs or not.

Recent evidence into the function of C1ORF112 has started to shape the possible roles C1ORF112 plays within cells and tissues. One such study by Hu, Li et al. (2017), stated that C1ORF112 could be a functional homolog to a plant gene Arabidopsis called MEICA which has been indicated to play a role in chromosome crossovers in homologous recombination. MEICA1 regulates meiotic recombination in rice through its interaction with TOP3α. They also stated that in rice, MEICA1 interacted with MSH7, suggesting its role in preventing nonallelic recombination and anti-crossover activity suppressing the defects of crossover formation in msh5. They showed that MEICA1 aberrant pollens had aberrant chromosome interactions which were visible at the metaphase stage when compared to the wild-type pollen. They also showed that MEICA1 was not required for homologous bivalents formations and functioned in a DMC dependent pathway. DMC encodes for the meiotic recombination protein, which plays a role in homologous strand assimilation required for the resolution of DSBs (Dray, Dunlop et al. 2011). This study was corroborated by Fernandes, et al. (2018), whose study was looking into DMC1 and FIGL1, both of which are also evolutionarily conserved proteins.

DMC1 is DNA meiotic recombinase and FIGNL1 is Fidgetin-like protein 1, both proteins act to resolve double-strand breaks (DSBs) independent of the BRCA2 pathway. DMC1 is important in homologous recombination (HR) and genetic variance during meiosis (Dray, Dunlop et al. 2011). FIGNL1 is required for efficient HR repair and is recruited to sites of DSBs (Yuan and Chen 2013). FIGNL1 interacts with RAD51 through its conserved binding domain and other scaffolding proteins such as KIAA0146/SPIDR in HR repair. These studies highlight, the possible role of C1ORF112 based on its interaction with FIGNL1, and its possible homology with MEICA1. The results of these studies infer that role C1ORF112 might play a role in cell cycle or proliferation, around the S/G2 point of the cell cycle, as FIGNL1 is recruited to sites of DNA damage through its N-terminal domain (Yuan and Chen 2013), and if C1ORF112 is a interacts with FIGNL1 (Fernandes, et al.,2018), it is possible C1ORF112 could be recruited to sites of DNA damage as well during HRR.

The Hu, Li et al. (2017) study does corroborate my sequence conservation study verifying that C1ORF112 is evolutionary conserved even in plants and that the conservation appears to lie in its function rather than the exact sequence. They did not make mention of the WCF tripeptide, indicating it may not be functionally relevant to C1ORF112. C1ORF112 aside from being well conserved is also co-expressed with genes that are responsible for the control and maintenance of cell cycle, DNA replication, and cell replication. With a co-expression score of approx. 70% across the different categories in Genevestigator, the number of unique co-expressed genes include but are not limited to MCM10, MAD2L1, CENPA, AURKB, BUB1, POLE2, and CDC6 Kinetochore complexing proteins. These control chromosome segregation and the progression of the cell through cellular division. However, C1ORF112 is not expressed in a cells or tissues-specific manner, looking at the relative expression of C1ORF112 in tissues (Figure 3.2), with testis having a higher level of expression compared to other tissues. The function of the testis is spermatogenesis and as such meiosis is highly prevalent for the generation of gamete cells, and if C1ORF112 does play a role in the efficacy of chromosomal segregation, kinetochore assembly and DDR, it would explain why it is highly expressed in the testis.

Aside from the testis, Epstein-Barr virus-transformed lymphocytes cells, also have relatively higher expression levels in transcripts per million (TPM) of C1ORF112 when compared to the other tissues, and like the testis, these cells are transformed into a state of uncontrolled cell division and as such would have higher expression of C1ORF112. Furthermore, age-related expression of C1ORF112 is relatively stable across tissues as well (Figure 3.3 - 3.11), analysis of GTex data revealed that when tissues were grouped by age and the expression of C1ORF112 determined, there were no significant changes in the levels of expression of C1ORF112. There were disparities in sample size due to the availability of some tissues like tissues that comprised the female reproduction such as the uterus, cervix, and fallopian tube due to their importance and tissues in the older age 70-79 category as well. This disparity did affect analysis in some tissues such as the Cervix, fallopian tube (Figure 3.3) as there was not enough information to categorically determine the effect of age on the expression of C1ORF112 in these samples. Other tissues such as the uterus and ovary had smaller samples sizes in the 70-79 age group (Figure 3.3) and this affected the results of the analysis with lower TPM levels shown. The uterus did show an increasing range of TPM expression of C1ORF112 with an increase in age, but the biological effect of this is unclear. The testis in contrast showed a decrease in the TPM expression of C1ORF112 at the 70-79 age group but the sample size was 9 for the category and this may have affected the results.

To determine the enriched biological process and pathways C1ORF112 could be involved in, co-expression analysis followed by gene enrichment analysis was carried. Genevestigator revealed that FANCI, NCAPG2 and NUF2 were identified to be the top 3 genes positively co-expressed with C1ORF112 across anatomical parts, cancers, and cell lines. These genes have been reported to be involved in cell cycle progression, DNA repair and chromosome segregation (DeLuca, Moree et al. 2002, Smogorzewska, Matsuoka et al. 2007, Liu, Tanasa et al. 2010). To understand the biological processes over-represented in the co-expression analysis, the results determined cell division (GO:0051301), sister chromatid cohesion (GO:0007062), and mitotic nuclear division (GO:0007067) to be the top biological processes associated with genes positively co-expressed with C1ORF112. These results indicate a strong association of C1ORF112 with these processes. Following on, gene set enrichment analysis of gene co-expressed with C1ORF112 with showed that top GO terms over-represented for the co-expressed genes are ATPase activity, catalytic activity on DNA, DNA dependent activity, serine/threonine activity, and other activities that directly impact DNA/RNA (Figures 3.4-3.9). In addition, cell cycle RNA transport, Fanconi anaemia, HR repair, and and cellular senescence are top pathways overrepresented in all categories. C1ORF112 also has 7 possible physical interactors (Figure 4.10), including FIGL1 and DMC1 stated to be associated via tandem affinity purification coupled to mass spectrometry (TAP-MS) using overexpressed FIGL1 as a bait (J. B. Fernandes et al., 2018). Finally, results show that C1ORF112 is co-expressed with genes involved with DNA replication, kinetochore assembly, cell cycle progression, and cell replication. On the other hand, the genes negatively co-expressed with C1ORF112 are CES4A, ADHFE1, and PIK3IP1. The standout gene from the top 3 genes negatively co-expressed with C1ORF112 is, PIK3IP1 which is predicted to enable phosphatidylinositol 3-kinase catalytic subunit binding activity and is involved in the negative regulation of phosphatidylinositol 3-kinase (Joshi, Wei et al. 2016). Although co-expression does not translate to physical interaction with the proteins, there is the possibility of down-regulation of PIK3IP1 by C1ORF112 could elucidate the role of C1ORF112 in cell growth and proliferation through the PI3k pathway. CES4A encodes Carboxylesterase 4A and ADHFE1 encodes Hydroxyacid-oxoacid transhydrogenase.

C1ORF112 has been suggested to be a potential biomarker for several tumours by J. Chen et al. (2021). The study of C1ORF112 expression across several patient tumours have shown that lower expression of C1ORF112 was shown to increase the likelihood of better survival for several tumours including, but not limited to, bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), cholangiocarcinoma (CHOL), colon adenocarcinoma (COAD), oesophageal carcinoma(ESCA), glioblastoma multiforme (GBM), HNSC, kidney renal papillary cell carcinoma (KIRP), LIHC, lung adenocarcinoma(LUAD), lung squamous cell carcinoma (LUSC), rectum adenocarcinoma (READ) (J. Chen et al., 2021). This alongside the work of Z. Zhang et al. (2021), also suggests that the presence of C1ORF112 predicts poor outcomes in patients with low-grade glioma. In their study, high expression of C1ORF112 positively correlated with immune cells such as B cells, CD8+T cells, CD4+T cells, macrophages, neutrophils, and dendritic cells infiltrating low-grade gliomas and was an independent factor in overall survival. Nonetheless, analysis of available samples does show that expression of C1ORF112 does not change over time in normal cells. Its expression is stable and consistent across all tissues over time.

C1ORF112 is an alpha helical protein with a possible kinase domain and a possible kinase domain at the N-terminal region and PP1γ docking site at the C-terminal domain

Generation of 3D protein models from amino acid sequences is generally the aim of protein model prediction, however, the efficacy of the prediction is dependent on several factors, one of which is the prediction tools used. I-TASSER hosted by the Zhang server has been used several times since its inception in 2006 for protein structure prediction with over 41 publications. In this case, the models created via I-TASSER has shown that C1ORF112 is an alpha superhelix. However, the models were not the best fit when further analysed using QMEAN, RAMPAGE, ERRAT, and Verify-3D. The results of the I-TASSER models did predict an alpha-helical structure, however, it did have limitations on the super fold structure. This could either be an algorithm issue or the sequence threading capacity of the server is limited, and this could be because of the sequence models used for its model prediction. It does seem, it could benefit from an improved model selection process. AlphaFold on the other hand appeared to be better at the structural prediction of C1ORF112. Further analysis using QMEAN, RAMPAGE, ERRAT, and Verify-3D showed that AlphaFold prediction of C1ORF112 is closer to its native protein compared to the I-TASSER selection. Other structural prediction software and servers such as the Swiss-Prot, were not utilized as they did not give full sequence model prediction for C1ORF112 and as such did not full the criteria for model prediction selection. Considering these limitations, both prediction tools do confirm that C1ORF112 is an alpha superhelix, which seems to be Armadillo repeats (ARM).

ARM repeats are imperfect tandem helical repeats of approximately 42 amino acids motif composing three helical turns (Coates, 2003; Tewari, Bailes, Bunting, & Coates, 2010). ARM repeats containing proteins are called that because it was first characterised in the Drosophila polarizing protein Armadillo (Peifer, Berg, & Reynolds, 1994). Proteins in this structural class appear not to share sequence similarity and play multi-functional roles in cells. Due to the sequence fluidity of ARM proteins, determining homologues and orthologues becomes difficult (Tewari et al., 2010). C1ORF112 does maintain its sequence across species as discussed in Section 4, however, defining functional homologues within a species has proven difficult and would require more studies. In addition, C1ORF112 and other ARM proteins with the tandem superhelix as shown with the AlphaFold model forms a versatile platform for interacting with other proteins, this enables them to carry out several functions based on their post-translational modification and domain architecture. As such there are several ARM proteins that serve several functions in cells. For example, Β-Catenin (Armadillo homologue) is typically a cell adhesion and a signalling protein involved in the Wnt signalling pathway, transducing extracellular signals to modify gene expression in the nucleus (Cadigan & Peifer, 2009; MacDonald, Tamai, & He, 2009). In addition, Β-Catenin is a hub protein for other signalling networks. Interestingly, Β-Catenin has been shown to localise to the centrosome to interact with microtubules regulating their regrowth, cohesion, and separation (Bahmanyar et al., 2008; P. Huang, Senga, & Hamaguchi, 2007). Wnt signals facilitate the splitting centrosomes and as such show both structural and transcriptional roles dependent on cellular localization and post-translational modification (Bahmanyar et al., 2008; Hadjihannas, Bruckner, & Behrens, 2010; P. Huang et al., 2007).

ARM proteins are not limited to structural and transcriptional roles, other examples of ARM proteins with different functional roles are the importin and exportin proteins. Importins are transport cargo proteins that shuttle proteins to and from the nucleus and consist of 10 ARM repeats (Mason, Stage, & Goldfarb, 2009). There are 3 sub-classes of importin-α in animals which are also evolutionary conserved in plants, fungi, amoeba and choanoflagellates (Mason et al., 2009). This level of conservation indicates the important role this class of ARM proteins plays across the Phylum, especially in spermatogenesis and gametogenesis, where their absence elicit sterility in Drosophila (Holt et al., 2007; Ratan, Mason, Sinnot, Goldfarb, & Fleming, 2008). These few examples highlight the range of roles ARM proteins play in the cell and how determining the function of C1ORF112 is a complex process.

A recent study by L. Xu et al. (2021), looking into Polo-like Kinase 1 (PLK1), C1ORF112, termed Apolo 1 was identified as aiding feedback control for chromosome segregation. This interaction between PLK1 and C1ORF112 was enabled mechanistically by binding the Polo-box domain (PBD) on PLK1 with its N-terminal region which corresponds to the serine amino acids at the N-terminal sequence ETKNKVVsFLEKTGF. This suggests that the N-terminal region is a PBD binding domain. They also state that the C-terminal of C1ORF112 contains a Protein phosphatase 1γ (PP1γ) docking motif, hereby, interacting with both PLK1 and PP1γ. This, alongside other studies, further strengthens the idea of C1ORF112 playing a role in cell cycle regulation and control of cell proliferation.

***C1ORF112 mutant cells are sensitive to agents of DNA damage especially x-ray radiation and hydrogen peroxide*.**

To further understand the role of C1ORF112, HAP1 cells were genetically modified to delete the C1ORF112 gene, and the cells were studied to further understand its role in cells and the associated gene expression profile. I established that the mutant C1ORF112 produced no protein through immunoblotting (Figure 5.1). this was the followed by a cell count analysis to determine if the knockout cells had any phenotypic difference with the normal HAP1 cells. C1ORF112 mutant cells were appeared to have a slower growth rate at 24hr after passaging, the normal HAP1 has a doubling time of 18hrs 22mins and a growth rate of 2 cells per min, the knockout HAP1 cells had a doubling time of 52hrs 2min and the growth rate was 0.8 cells per min, possibly due to trypsinization. The doubling time for the C1ORF112 knockout cells was halved if the cells were left to grow for the 48hr and 72hr and became comparable to the normal cells at the 48hr and 72 hr time points. To further ascertain if C1ORF112 knockdown caused a change in the expression from file of the cells, a microarray assay was carried out. This revealed that 730 protein-coding genes were differentially expressed. The total number of protein-coding genes that were down-regulated were 289, and the total number of protein-coding genes that were up-regulated was 441, 24hr post-seeding (Figure 5.3). At 48hr post-seeding, the total number of differentially expressed protein-coding genes was 368. The total number of protein-coding genes that were down-regulated were 144, and the total number of protein-coding genes that were up-regulated was 224 (Figure 5.4). in addition, the pathways with the most DEGs were the PI3K-AKT pathway and Malignant pleural mesothelioma pathway. The PI3K-AKT pathway had 11 growth factor and ECM related genes downregulated, this could explain why the cell doubling time was longer at 24hr. it is probable that the knockdown of C1ORF112, caused downregulation of growth factor which in turn caused the cells to inefficiently transduce cell growth signalling to induce cell growth.

To further understand if C1ORF112 has a role in DDR, C1ORF112 knockout cells and normal HAP1 cells were treated with increasing doses of x-ray radiation and hydrogen peroxide. Overall, there was significant difference in the surviving fractions between the normal HAP1 cells and the C1ORF112 knockdown in the clonogenic assay. In addition, the C1ORF112 knockdown cells showed higher levels of DSBs induction compared to the normal C1ORF112 across all dosages for both hydrogen peroxide and x-ray, however, there was no significant difference in the repair kinetics in the x-ray and at the 1hr and 4hr time points of the hydrogen peroxide. There was significant difference at the 2hr time point for the peroxide, but the mechanism behind this is currently unclear. Furthermore, the cell fraction study showed that inducing high levels of DSBs using x-rays did not cause C1ORF112 to move into the nucleus. This is interesting because previous studies by Xu, Ali et al. (2021) and Fernandes, Duhamel et al. (2018) have shown that C1ORF112 is usually involved in DNA associating activity after the breakdown of the nuclear membrane. The is also corroborated by the study in plants by Hu, Li et al. (2017), were MEICA1 is involved in meiotic chromosome segregating during pachytene point of cell division, after the breakdown of the nuclear membrane. This leads to more questions as to how C1ORF112 may be involved in cell growth and proliferation, and possibly DDR.

Future points of study for C1ORF112

Understanding the post-translation modifications of C1ORF112

C1ORF112 has become an interesting protein to study in the last couple of years, it has since been deduced that the N-terminal and C-terminal of C1ORF112 contain phosphorylation sites. It would be interesting to understand the function of phospho-C1ORF112, if there is a possibility of dimerization, possible kinase activity. Of maybe it acts as a bridge between a kinase and a phosphatase as suggested by (Xu, Ali et al. 2021), sub-cellular localization, and the possible downstream interactors. Other post-translational modification of C1ORF112 would also be useful for functional characterizing such as sumoylation, ubiquitylation and acetylation especially if it plays a cell proliferation and chromosome segregation.

The role on C1ORF112 in chromosome segregation and kinetochore complexing

Based on the co-expression analysis, C1ORF112 is co-expressed with several genes that are involved cell proliferation and kinetochore complex genes. It however, unclear if C1ORF112 plays a direct role in the cell proliferation pathways.

Possible control of growth factors

The results for the microarray analysis showed that 11 genes associated with growth factor and Extracellular matrix (ECM) were down regulated in the PI3K-AHt pathway, however the is currently on evidence of direct interaction between C1ORF112 and any of these genes. An interest avenue into the function of C1ORF112 would be to determine if directly interacts with any growth factors and if that could explain why the growth rate was less for the knockout cells compared the to the normal cells.

Conclusion

In conclusion, the results of the thesis have determined that

- C1ORF112 is a well-conserved protein all the way to choanoflagellates, close relative of the Metazoa and particularly in primates and mammals. C1ORF112 is also co-expressed with genes associated with chromosome integrity, segregation, and cell replication.

- C1ORF112 is alpha-helical protein The structural analysis has also determined that C1ORF112 has two possible sites of phosphorylation one at the N-terminal and the other at the C-terminal and could be associated with N-terminal acting as a kinase domain.
- This thesis has also shown that C1ORF112 is cytoplasmic protein and when C1ORF112 is knock out in cells, it affects the growth rate for the first 24hr after trypsinization and replating, but this recovers afterwards. C1ORF112 loss also increases the sensitivity of the cells to agents of DNA damage, especially x-ray radiation and hydrogen peroxide.C1ORF112 does not appear to have a direct role in DDR.

8: References

Adachi, N. and M. R. Lieber (2002). "Bidirectional gene organization: a common architectural feature of the human genome." Cell **109**(7): 807-809.
Ahel, D., Z. Horejsi, N. Wiechens, S. E. Polo, E. Garcia-Wilson, I. Ahel, H. Flynn, M. Skehel, S. C. West, S. P. Jackson, T. Owen-Hughes and S. J. Boulton (2009). "Poly(ADP-ribose)-dependent regulation of DNA repair by the chromatin remodeling enzyme ALC1." Science 325(5945): 1240-1243
Alessi, D. R., S. R. James, C. P. Downes, A. B. Holmes, P. R. Gaffney, C. B. Reese and P. Cohen (1997). "Characterization of a 3-phosphoinositide-dependent protein kinase which phosphorylates and activates protein kinase Balpha." Curr Biol **7**(4): 261-269.
Allshire, R. C. and G. H. Karpen (2008). "Epigenetic regulation of centromeric chromatin: old dogs, new tricks?" Nat Rev Genet **9**(12): 923-937.
Altman, R. B. and J. M. Dugan (2003). "Defining bioinformatics and structural bioinformatics." Methods Biochem Anal **44**: 3-14.
Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.
Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.
Andrade, M. A., C. Petosa, S. I. O'Donoghue, C. W. Muller and P. Bork (2001). "Comparison of ARM and HEAT protein repeats." J Mol Biol **309**(1): 1-18.
Anfinsen, C. B. (1961). "Points of current interest in protein chemistry." Lab Invest **10**: 987-991.
Anfinsen, C. B. (1973). "Principles that govern the folding of protein chains." Science **181**(4096): 223-230.
Atillasoy, E. S., J. T. Seykora, P. W. Soballe, R. Elenitsas, M. Nesbit, D. E. Elder, K. T. Montone, E. Sauter and M. Herlyn (1998). "UVB induces atypical melanocytic lesions and melanoma in human skin." Am J Pathol **152**(5): 1179-1186.
Ault, J. G. and C. L. Rieder (1994). "Centrosome and kinetochore movement during mitosis." Curr Opin Cell Biol **6**(1): 41-49.
Bahmanyar, S., D. D. Kaplan, J. G. Deluca, T. H. Giddings, Jr., E. T. O'Toole, M. Winey, E. D. Salmon, P. J. Casey, W. J. Nelson and A. I. Barth (2008). "beta-Catenin is a Nek2 substrate involved in centrosome separation." Genes Dev **22**(1): 91-105.
Barnum, K. J. and M. J. O'Connell (2014). "Cell cycle regulation by checkpoints." Methods Mol Biol **1170**: 29-40.
Basu, D., W. Wang, S. Ma, T. DeBrosse, E. Poirier, K. Emch, E. Soukup, L. Tian and A. M. Showalter (2015). "Two Hydroxyproline Galactosyltransferases, GALT5 and GALT2, Function in Arabinogalactan-Protein Glycosylation, Growth and Development in Arabidopsis." PLoS One **10**(5): e0125624.
Baumann, P. and S. C. West (1998). "Role of the human RAD51 protein in homologous recombination and double-stranded-break repair." Trends Biochem Sci **23**(7): 247-251.
Becker, P. B. and W. Horz (2002). "ATP-dependent nucleosome remodeling." Annu Rev Biochem **71**: 247-273.
Benjamin, C. L. and H. N. Ananthaswamy (2007). "p53 and the pathogenesis of skin cancer." Toxicol Appl Pharmacol **224**(3): 241-248.
Bergmiller, T., M. Ackermann and O. K. Silander (2012). "Patterns of evolutionary conservation of essential genes correlate with their compensability." PLoS Genet **8**(6): e1002803.
Besson, A., S. F. Dowdy and J. M. Roberts (2008). “CDK inhibitors: cell cycle regulators and beyond.” Dev Cell 14(2): 159-169.
Biegert, A. and J. Soding (2009). "Sequence context-specific profiles for homology searching." Proc Natl Acad Sci U S A **106**(10): 3770-3775.

Bjelland, S. and E. Seeberg (2003). "Mutagenicity, toxicity and repair of DNA base damage induced by oxidation." Mutat Res **531**(1-2): 37-80.

Bonneau, R. and D. Baker (2001). "Ab initio protein structure prediction: progress and prospects." Annu Rev Biophys Biomol Struct **30**: 173-189.

Boore, J. L., L. L. Daehler and W. M. Brown (1999). "Complete sequence, gene arrangement, and genetic code of mitochondrial DNA of the cephalochordate Branchiostoma floridae (Amphioxus)." Mol Biol Evol **16**(3): 410-418.

Boratyn, G. M., A. A. Schaffer, R. Agarwala, S. F. Altschul, D. J. Lipman and T. L. Madden (2012). "Domain enhanced lookup time accelerated BLAST." Biol Direct **7**: 12.

Borrego-Soto, G., R. Ortiz-Lopez and A. Rojas-Martinez (2015). "Ionizing radiation-induced DNA injury and damage detection in patients with breast cancer." Genet Mol Biol **38**(4): 420-432.

Boutanaev, A. M., A. I. Kalmykova, Y. Y. Shevelyov and D. I. Nurminsky (2002). "Large clusters of co-expressed genes in the Drosophila genome." Nature **420**(6916): 666-669.

Bowater, R. P. and R. D. Wells (2001). "The intrinsically unstable life of DNA triplet repeats associated with human hereditary disorders." Prog Nucleic Acid Res Mol Biol **66**: 159-202.

Bradford, P. T., A. M. Goldstein, D. Tamura, S. G. Khan, T. Ueda, J. Boyle, K. S. Oh, K. Imoto, H. Inui, S. Moriwaki, S. Emmert, K. M. Pike, A. Raziuddin, T. M. Plona, J. J. DiGiovanna, M. A. Tucker and K. H. Kraemer (2011). "Cancer and neurologic degeneration in xeroderma pigmentosum: long term follow-up characterises the role of DNA repair." J Med Genet **48**(3): 168-176

Braselmann, H., A. Michna, J. Hess and K. Unger (2015). "CFAssay: statistical analysis of the colony formation assay." Radiat Oncol **10**: 223.

Breen, A. P. and J. A. Murphy (1995). "Reactions of oxyl radicals with DNA." Free Radic Biol Med **18**(6): 1033-1077.

Cadet, J., T. Douki and J. L. Ravanat (2010). "Oxidatively generated base damage to cellular DNA." Free Radic Biol Med **49**(1): 9-21.

Cadet, J., T. Douki and J. L. Ravanat (2011). "Measurement of oxidatively generated base damage in cellular DNA." Mutat Res **711**(1-2): 3-12.

Cadet, J. and J. R. Wagner (2014). "Oxidatively generated base damage to cellular DNA by hydroxyl radical and one-electron oxidants: similarities and differences." Arch Biochem Biophys **557**: 47-54.

Cadigan, K. M. and M. Peifer (2009). "Wnt signaling from development to disease: insights from model systems." Cold Spring Harb Perspect Biol **1**(2): a002881.

Cahir McFarland, E. D., K. M. Izumi and G. Mosialos (1999). "Epstein-barr virus transformation: involvement of latent membrane protein 1-mediated activation of NF-kappaB." Oncogene **18**(49): 6959-6964.

Callaway, E. (2015). "The revolution will not be crystallized: a new method sweeps through structural biology." Nature **525**(7568): 172-174.

Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, N. Shaw, C. R. Lane, E. P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G. Q. Daley and E. S. Lander (1999). "Characterization of single-nucleotide polymorphisms in coding regions of human genes." Nat Genet **22**(3): 231-238.

Cazzalini, O., S. Sommatis, M. Tillhon, I. Dutto, A. Bachi, A. Rapp, T. Nardo, A. I. Scovassi, D. Necchi, M. C. Cardoso, L. A. Stivala and E. Prosperi (2014). "CBP and p300 acetylate PCNA to link its degradation with nucleotide excision repair synthesis." Nucleic Acids Res **42**(13): 8433-8448.

Chan, G. K., S. T. Liu and T. J. Yen (2005). "Kinetochore structure and function." Trends Cell Biol **15**(11): 589-598.

Chapman, J. R. and S. P. Jackson (2008). "Phospho-dependent interactions between NBS1 and MDC1 mediate chromatin retention of the MRN complex at sites of DNA damage." EMBO Rep **9**(8): 795-801.

Chen, J., H. Mai, H. Chen, B. Zhou, J. Hou and D. K. Jiang (2021). "Pan-Cancer Analysis Identified C1ORF112 as a Potential Biomarker for Multiple Tumor Types." Front Mol Biosci **8**: 693651.

Chen, R. H. (2002). "BubR1 is essential for kinetochore localization of other spindle checkpoint proteins and its phosphorylation requires Mad1." J Cell Biol **158**(3): 487-496.

Chen, X., L. J. Ko, L. Jayaraman and C. Prives (1996). "p53 levels, functional domains, and DNA damage determine the extent of the apoptotic response of tumor cells." Genes Dev 10(19): 2438-2451.

Chetsanga, C. J., M. Lozon, C. Makaroff and L. Savage (1981). "Purification and characterization of Escherichia coli formamidopyrimidine-DNA glycosylase that excises damaged 7-methylguanine from deoxyribonucleic acid." Biochemistry **20**(18): 5201-5207.

Cho, R. J., M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart and R. W. Davis (1998). "A genome-wide transcriptional analysis of the mitotic cell cycle." Mol Cell **2**(1): 65-73.

Choi, Y. S., J. S. Yang, Y. Choi, S. H. Ryu and S. Kim (2009). "Evolutionary conservation in multiple faces of protein interaction." Proteins **77**(1): 14-25.

Chou, P. Y. and G. D. Fasman (1974). "Prediction of protein conformation." Biochemistry **13**(2): 222-245.

Cimini, D., D. Fioravanti, E. D. Salmon and F. Degrassi (2002). "Merotelic kinetochore orientation versus chromosome mono-orientation in the origin of lagging chromosomes in human primary cells." J Cell Sci **115**(Pt 3): 507-515.

Coates, J. C. (2003). "Armadillo repeat proteins: beyond the animal kingdom." Trends Cell Biol **13**(9): 463-471.

The ENCODE Project Consortium, E. P. (2004). "The ENCODE (ENCyclopedia Of DNA Elements) Project." Science **306**(5696): 636-640.

The ENCODE Project Consortium, E. P. (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

Cook, P. J., B. G. Ju, F. Telese, X. Wang, C. K. Glass and M. G. Rosenfeld (2009). "Tyrosine dephosphorylation of H2AX modulates apoptosis and survival decisions." Nature **458**(7238): 591-596.

Creighton, T. E. (1990). "Protein folding." Biochem J **270**(1): 1-16.

Crick, F. H. (1966). "Codon--anticodon pairing: the wobble hypothesis." J Mol Biol **19**(2): 548-555.

D'Arcangelo, G., R. Homayouni, L. Keshvara, D. S. Rice, M. Sheldon and T. Curran (1999). "Reelin is a ligand for lipoprotein receptors." Neuron **24**(2): 471-479.

Davidson, E. H. (2006). The Regulatory Genome: Gene Regulatory Networks in Development and Evolution.

Davila Lopez, M., J. J. Martinez Guerra and T. Samuelsson (2010). "Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes." PLoS One **5**(5): e10654.

De Bont, R. and N. van Larebeke (2004). "Endogenous DNA damage in humans: a review of quantitative data." Mutagenesis **19**(3): 169-185.

de Castro, E., C. J. Sigrist, A. Gattiker, V. Bulliard, P. S. Langendijk-Genevaux, E. Gasteiger, A. Bairoch and N. Hulo (2006). "ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins." Nucleic Acids Res **34**(Web Server issue): W362-365.

Dekel, E., S. Mangan and U. Alon (2005). "Environmental selection of the feed-forward loop circuit in gene-regulation networks." Phys Biol **2**(2): 81-88.

DeLuca, J. G., B. Moree, J. M. Hickey, J. V. Kilmartin and E. D. Salmon (2002). "hNuf2 inhibition blocks stable kinetochore-microtubule attachment and induces mitotic cell death in HeLa cells." J Cell Biol **159**(4): 549-555.

Demidova, A. R., M. Y. Aau, L. Zhuang and Q. Yu (2009). "Dual regulation of Cdc25A by Chk1 and p53-ATF3 in DNA replication checkpoint control." J Biol Chem **284**(7): 4132-4139.

Desai, P., N. Guha, L. Galdieri, S. Hadi and A. Vancura (2009). "Plc1p is required for proper chromatin structure and activity of the kinetochore in Saccharomyces cerevisiae by facilitating recruitment of the RSC complex." Mol Genet Genomics **281**(5): 511-523.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot, A. Vignal, P. Millasseau, S. Marc, J. Hazan, E. Seboun, M. Lathrop, G. Gyapay, J. Morissette and J. Weissenbach (1996). "A comprehensive genetic map of the human genome based on 5,264 microsatellites." Nature **380**(6570): 152-154.

Ditchfield, C., V. L. Johnson, A. Tighe, R. Ellston, C. Haworth, T. Johnson, A. Mortlock, N. Keen and S. S. Taylor (2003). "Aurora B couples chromosome alignment with anaphase by targeting BubR1, Mad2, and Cenp-E to kinetochores." J Cell Biol **161**(2): 267-280.

Dizdaroglu, M., G. Rao, B. Halliwell and E. Gajewski (1991). "Damage to the DNA bases in mammalian chromatin by hydrogen peroxide in the presence of ferric and cupric ions." Arch Biochem Biophys **285**(2): 317-324.

Doheny, K. F., P. K. Sorger, A. A. Hyman, S. Tugendreich, F. Spencer and P. Hieter (1993). "Identification of essential components of the S. cerevisiae kinetochore." Cell **73**(4): 761-774.

Dorn, M., E. S. MB, L. S. Buriol and L. C. Lamb (2014). "Three-dimensional protein structure prediction: Methods and computational strategies." Comput Biol Chem **53PB**: 251-276.

Dray, E., M. H. Dunlop, L. Kauppi, J. San Filippo, C. Wiese, M. S. Tsai, S. Begovic, D. Schild, M. Jasin, S. Keeney and P. Sung (2011). "Molecular basis for enhancement of the meiotic DMC1 recombinase by RAD51 associated protein 1 (RAD51AP1)." Proc Natl Acad Sci U S A **108**(9): 3560-3565.

Eichler, E. E. and D. Sankoff (2003). "Structural dynamics of eukaryotic chromosome evolution." Science **301**(5634): 793-797.

El-Gebali, S., J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. E. Tosatto and R. D. Finn (2019). "The Pfam protein families database in 2019." Nucleic Acids Res **47**(D1): D427-D432.

Elcock, A. H. and J. A. McCammon (2001). "Identification of protein oligomerization states by analysis of interface conservation." Proc Natl Acad Sci U S A **98**(6): 2990-2994.

Elowitz, M. B., A. J. Levine, E. D. Siggia and P. S. Swain (2002). "Stochastic gene expression in a single cell." Science **297**(5584): 1183-1186.

Emanuele, M. J., W. Lan, M. Jwa, S. A. Miller, C. S. Chan and P. T. Stukenberg (2008). "Aurora B kinase and protein phosphatase 1 have opposing roles in modulating kinetochore assembly." J Cell Biol **181**(2): 241-254.

Essletzbichler, P., T. Konopka, F. Santoro, D. Chen, B. V. Gapp, R. Kralovics, T. R. Brummelkamp, S. M. Nijman and T. Burckstummer (2014). "Megabase-scale deletion using CRISPR/Cas9 to generate a fully haploid human cell line." Genome Res **24**(12): 2059-2065.

Fantes, P. A. (1977). "Control of cell size and cycle time in Schizosaccharomyces pombe." J Cell Sci **24**: 51-67.

Fatemi, M., M. M. Pao, S. Jeong, E. N. Gal-Yam, G. Egger, D. J. Weisenberger and P. A. Jones (2005). "Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level." Nucleic Acids Res **33**(20): e176.

Fernandes, J. B., M. Duhamel, M. Seguela-Arnaud, N. Froger, C. Girard, S. Choinard, V. Solier, N. De Winne, G. De Jaeger, K. Gevaert, P. Andrey, M. Grelon, R. Guerois, R. Kumar and R. Mercier (2018). "FIGL1 and its novel partner FLIP form a conserved complex that regulates homologous recombination." PLoS Genet **14**(4): e1007317.

Fernandes, M., C. Wan, R. Tacutu, D. Barardo, A. Rajput, J. W. Wang, H. Thoppil, D. Thornton, C. H. Yang, A. Freitas and J. P. de Magalhaes (2016). "Systematic analysis of the gerontome reveals links between aging and age-related diseasesSystematic analysis of the gerontome reveals links between aging and age-related diseases." Human Molecular Genetics **25**(21): 4804-4818.

Finkelstein, A. V. and O. B. Ptitsyn (1987). "Why do globular proteins fit the limited set of folding patterns?" Prog Biophys Mol Biol **50**(3): 171-190.

Finn, R. D., T. K. Attwood, P. C. Babbitt, A. Bateman, P. Bork, A. J. Bridge, H. Y. Chang, Z. Dosztanyi, S. El-Gebali, M. Fraser, J. Gough, D. Haft, G. L. Holliday, H. Huang, X. Huang, I. Letunic, R. Lopez, S. Lu, A. Marchler-Bauer, H. Mi, J. Mistry, D. A. Natale, M. Necci, G. Nuka, C. A. Orengo, Y. Park, S. Pesseat, D. Piovesan, S. C. Potter, N. D. Rawlings, N. Redaschi, L. Richardson, C. Rivoire, A. Sangrador-Vegas, C. Sigrist, I. Sillitoe, B. Smithers, S. Squizzato, G. Sutton, N. Thanki, P. D. Thomas, S. C. Tosatto, C. H. Wu, I. Xenarios, L. S. Yeh, S. Y. Young and A. L. Mitchell (2017). "InterPro in 2017-beyond protein family and domain annotations." Nucleic Acids Res **45**(D1): D190-D199.

Francisco, L., W. Wang and C. S. Chan (1994). "Type 1 protein phosphatase acts in opposition to IpL1 protein kinase in regulating yeast chromosome segregation." Mol Cell Biol **14**(7): 4731-4740.

Friedberg, E. C. (2005). "Suffering in silence: the tolerance of DNA damage." Nat Rev Mol Cell Biol **6**(12): 943-953.

Friedberg, E. C., L. D. McDaniel and R. A. Schultz (2004). "The role of endogenous and exogenous DNA damage and mutagenesis." Curr Opin Genet Dev **14**(1): 5-10.

Friedberg, E. C., Walker, G. C. and Siede, W., (1995). DNA Repair and Mutagenesis, Washington: ASM Press.

Fu, Y. H., D. P. Kuhl, A. Pizzuti, M. Pieretti, J. S. Sutcliffe, S. Richards, A. J. Verkerk, J. J. Holden, R. G. Fenwick, Jr., S. T. Warren and et al. (1991). "Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox." Cell **67**(6): 1047-1058.

Fujiwara, Y., M. Ichihashi, Y. Kano, K. Goto and K. Shimizu (1981). "A new human photosensitive subject with a defect in the recovery of DNA synthesis after ultraviolet-light irradiation." J Invest Dermatol **77**(3): 256-263.

Galzitskaya, O. V. and B. S. Melnik (2003). "Prediction of protein domain boundaries from sequence alone." Protein Sci **12**(4): 696-701.

Ghosh, S. K., A. Poddar, S. Hajra, K. Sanyal and P. Sinha (2001). "The IML3/MCM19 gene of Saccharomyces cerevisiae is required for a kinetochore-related process during chromosome segregation." Mol Genet Genomics **265**(2): 249-257.

Giacco, F. and M. Brownlee (2010). "Oxidative stress and diabetic complications." Circ Res **107**(9): 1058-1070.

Giloni, L., M. Takeshita, F. Johnson, C. Iden and A. P. Grollman (1981). "Bleomycin-induced strand-scission of DNA. Mechanism of deoxyribose cleavage." J Biol Chem **256**(16): 8608-8615.

Gitai, Z. (2007). "Diversification and specialization of the bacterial cytoskeleton." Curr Opin Cell Biol **19**(1): 5-12.

Gregan, J., S. Polakova, L. Zhang, I. M. Tolic-Norrelykke and D. Cimini (2011). "Merotelic kinetochore attachment: causes and effects." Trends Cell Biol **21**(6): 374-381.

Gribskov, M., A. D. McLachlan and D. Eisenberg (1987). "Profile analysis: detection of distantly related proteins." Proc Natl Acad Sci U S A **84**(13): 4355-4358.

Gyapay, G., J. Morissette, A. Vignal, C. Dib, C. Fizames, P. Millasseau, S. Marc, G. Bernardi, M. Lathrop and J. Weissenbach (1994). "The 1993-94 Genethon human genetic linkage map." Nat Genet **7**(2 Spec No): 246-339.

Hadjihannas, M. V., M. Bruckner and J. Behrens (2010). "Conductin/axin2 and Wnt signalling regulates centrosome cohesion." EMBO Rep **11**(4): 317-324.

Hafstad, A. D., A. A. Nabeebaccus and A. M. Shah (2013). "Novel aspects of ROS signalling in heart failure." Basic Res Cardiol **108**(4): 359.

Hanahan, D. and R. A. Weinberg (2011). "Hallmarks of cancer: the next generation." Cell **144**(5): 646-674.

Hang, B. (2004). "Repair of exocyclic DNA adducts: rings of complexity." Bioessays 26(11): 1195-1208

Hashimoto, Y., F. Puddu and V. Costanzo (2011). "RAD51- and MRE11-dependent reassembly of uncoupled CMG helicase complex at collapsed replication forks." Nat Struct Mol Biol **19**(1): 17-24.

Hauf, S. (2003). "Fine tuning of kinetochore function by phosphorylation." Cell Cycle **2**(3): 228-229.

Hein, M. Y., N. C. Hubner, I. Poser, J. Cox, N. Nagaraj, Y. Toyoda, I. A. Gak, I. Weisswange, J. Mansfeld, F. Buchholz, A. A. Hyman and M. Mann (2015). "A human interactome in three quantitative dimensions organized by stoichiometries and abundances." Cell **163**(3): 712-723.

Hemmings, B. A. and D. F. Restuccia (2012). "PI3K-PKB/Akt pathway." Cold Spring Harb Perspect Biol **4**(9): a011189.

Henle, E. S. and S. Linn (1997). "Formation, prevention, and repair of DNA damage by iron/hydrogen peroxide." J Biol Chem **272**(31): 19095-19098.

Henner, W. D., L. O. Rodriguez, S. M. Hecht and W. A. Haseltine (1983). "gamma Ray induced deoxyribonucleic acid strand breaks. 3' Glycolate termini." J Biol Chem **258**(2): 711-713.

Hoeijmakers, J. H. (2009). "DNA damage, aging, and cancer." N Engl J Med **361**(15): 1475-1485.

Holt, J. E., J. D. Ly-Huynh, A. Efthymiadis, G. R. Hime, K. L. Loveland and D. A. Jans (2007). "Regulation of Nuclear Import During Differentiation; The IMP alpha Gene Family and Spermatogenesis." Curr Genomics **8**(5): 323-334.

Hosokawa, M., T. Furihata, Y. Yaginuma, N. Yamamoto, N. Koyano, A. Fujii, Y. Nagahara, T. Satoh and K. Chiba (2007). "Genomic structure and transcriptional regulation of the rat, mouse, and human carboxylesterase genes." Drug Metab Rev **39**(1): 1-15.

Hruz, T., O. Laule, G. Szabo, F. Wessendorp, S. Bleuler, L. Oertle, P. Widmayer, W. Gruissem and P. Zimmermann (2008). "Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes." Adv Bioinformatics **2008**: 420747.

Hu, Q., Y. Li, H. Wang, Y. Shen, C. Zhang, G. Du, D. Tang and Z. Cheng (2017). "Meiotic Chromosome Association 1 Interacts with TOP3alpha and Regulates Meiotic Recombination in Rice." Plant Cell **29**(7): 1697-1708.

Huang da, W., B. T. Sherman and R. A. Lempicki (2009). "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." Nucleic Acids Res **37**(1): 1-13.

Huang, D. W., B. T. Sherman and R. A. Lempicki (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nature Protocols **4**(1): 44-57.

Huang, P., T. Senga and M. Hamaguchi (2007). "A novel role of phospho-beta-catenin in microtubule regrowth at centrosome." Oncogene **26**(30): 4357-4371.

Huang, Y. and L. Li (2013). "DNA crosslinking damage and cancer - a tale of friend and foe." Transl Cancer Res **2**(3): 144-154.

Hunter, P. (2006). "Into the fold. Advances in technology and algorithms facilitate great strides in protein structure prediction." EMBO Rep **7**(3): 249-252.

Hyland, K. M., J. Kingsbury, D. Koshland and P. Hieter (1999). "Ctf19p: A novel kinetochore protein in Saccharomyces cerevisiae and a potential link between the kinetochore and mitotic spindle." J Cell Biol **145**(1): 15-28.

Imlay, J. A. and S. Linn (1988). "DNA damage and oxygen radical toxicity." Science **240**(4857): 1302-1309.

Ingolfsson, H. and G. Yona (2008). "Protein domain prediction." Methods Mol Biol **426**: 117-143.

International Human Genome Sequencing, C. (2004). "Finishing the euchromatic sequence of the human genome." Nature **431**(7011): 931-945.

Joshi, S., J. Wei and N. H. Bishopric (2016). "A cardiac myocyte-restricted Lin28/let-7 regulatory axis promotes hypoxia-mediated apoptosis by inducing the AKT signaling suppressor PIK3IP1." Biochim Biophys Acta **1862**(2): 240-251.

Juhas, M., L. Eberl and J. I. Glass (2011). "Essence of life: essential genes of minimal genomes." Trends Cell Biol **21**(10): 562-568.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis (2021). "Highly accurate protein structure prediction with AlphaFold." Nature **596**(7873): 583-589.

Kardon, T., G. Noel, D. Vertommen and E. V. Schaftingen (2006). "Identification of the gene encoding hydroxyacid-oxoacid transhydrogenase, an enzyme that metabolizes 4-hydroxybutyrate." FEBS Lett **580**(9): 2347-2350.

Karess, R. (2005). "Rod-Zw10-Zwilch: a key player in the spindle checkpoint." Trends Cell Biol **15**(7): 386-392.

Kasai, H., H. Hayami, Z. Yamaizumi, SaitoH and S. Nishimura (1984). "Detection and identification of mutagens and carcinogens as their adducts with guanosine derivatives." Nucleic Acids Res **12**(4): 2127-2136.

Katoh, K. and D. M. Standley (2013). "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." Mol Biol Evol **30**(4): 772-780.

Kauzmann, W. (1959). "Some factors in the interpretation of protein denaturation." Adv Protein Chem **14**: 1-63.

Kawale, A. S. and P. Sung (2020). "Mechanism and significance of chromosome damage repair by homologous recombination." Essays Biochem **64**(5): 779-790.

Ke, Y. W., Z. Dou, J. Zhang and X. B. Yao (2003). "Function and regulation of Aurora/Ipl1p kinase family in cell division." Cell Res **13**(2): 69-81.

Kerr, M. K., M. Martin and G. A. Churchill (2000). "Analysis of variance for gene expression microarray data." J Comput Biol **7**(6): 819-837.

Kerrigan, J. J., Q. Xie, R. S. Ames and Q. Lu (2011). "Production of protein complexes via co-expression." Protein Expr Purif **75**(1): 1-14.

Killander, D. and A. Zetterberg (1965). "A quantitative cytochemical investigation of the relationship between cell mass and initiation of DNA synthesis in mouse fibroblasts in vitro." Exp Cell Res **40**(1): 12-20.

Koonin, E. V. (1993). "A highly conserved sequence motif defining the family of MutT-related proteins from eubacteria, eukaryotes and viruses." Nucleic Acids Res **21**(20): 4847.

Kopp, J., L. Bordoli, J. N. Battey, F. Kiefer and T. Schwede (2007). "Assessment of CASP7 predictions for template-based modeling targets." Proteins **69 Suppl 8**: 38-56.

Kopp, J. and T. Schwede (2004). "Automated protein structure homology modeling: a progress report." Pharmacogenomics **5**(4): 405-416.

Korkolopoulou, P., G. Levidou, E. A. Trigka, N. Prekete, M. Karlou, I. Thymara, S. Sakellariou, P. Fragkou, D. Isaiadis, P. Pavlopoulos, E. Patsouris and A. A. Saetta (2012). "A comprehensive immunohistochemical and molecular approach to the PI3K/AKT/mTOR (phosphoinositide 3-kinase/v-akt murine thymoma viral oncogene/mammalian target of rapamycin) pathway in bladder urothelial carcinoma." BJU Int **110**(11 Pt C): E1237-1248.

Kotecki, M., P. S. Reddy and B. H. Cochran (1999). "Isolation and characterization of a near-haploid human cell line." Exp Cell Res **252**(2): 273-280.

Kouzarides, T. (2002). "Histone methylation in transcriptional control." Curr Opin Genet Dev **12**(2): 198-209.

Kouzarides, T. (2007). "Chromatin modifications and their function." Cell **128**(4): 693-705.

Kuhlbrandt, W. (2013). "Introduction to electron crystallography." Methods Mol Biol **955**: 1-16.

Kyte, J. and R. F. Doolittle (1982). "A simple method for displaying the hydropathic character of a protein." J Mol Biol **157**(1): 105-132.

Lakin, N. D. and S. P. Jackson (1999). "Regulation of p53 in response to DNA damage." Oncogene **18**(53): 7644-7655.

Langerak, P. and P. Russell (2011). "Regulatory networks integrating cell cycle control with DNA damage checkpoints and double-strand break repair." Philos Trans R Soc Lond B Biol Sci **366**(1584): 3562-3571.

Lechner, J. (1994). "A zinc finger protein, essential for chromosome segregation, constitutes a putative DNA binding subunit of the Saccharomyces cerevisiae kinetochore complex, Cbf3." EMBO J **13**(21): 5203-5211.

Lechner, J. and J. Ortiz (1996). "The Saccharomyces cerevisiae kinetochore." FEBS Lett **389**(1): 70-74.

Lee, J. M. and E. L. Sonnhammer (2003). "Genomic gene clustering analysis of pathways in eukaryotes." Genome Res **13**(5): 875-882.

Lee, J. Y. a. O.-W., T.L. (2001). Chromatin. Encyclopedia of Genetics, Academic Press.

Lehmann, A. R., D. McGibbon and M. Stefanini (2011). "Xeroderma pigmentosum." Orphanet J Rare Dis **6**: 70.

Leo, J. C., S. M. Wang, C. H. Guo, S. E. Aw, Y. Zhao, J. M. Li, K. M. Hui and V. C. Lin (2005). "Gene regulation profile reveals consistent anticancer properties of progesterone in hormone-independent breast cancer cells transfected with progesterone receptor." Int J Cancer **117**(4): 561-568.

Lesport, E., A. Ferster, A. Biver, B. Roch, N. Vasquez, N. Jabado, F. L. Vives, P. Revy, J. Soulier and J. P. de Villartay (2018). "Reduced recruitment of 53BP1 during interstrand crosslink repair is associated with genetically inherited attenuation of mitomycin C sensitivity in a family with Fanconi anemia." Oncotarget **9**(3): 3779-3793.

Letunic, I. and P. Bork (2019). "Interactive Tree Of Life (iTOL) v4: recent updates and new developments." Nucleic Acids Res **47**(W1): W256-W259.

Levitt, M. and C. Chothia (1976). "Structural patterns in globular proteins." Nature **261**(5561): 552-558.

Li, Y., Z. F. Pursell and S. Linn (2000). "Identification and cloning of two histone fold motif-containing subunits of HeLa DNA polymerase epsilon." J Biol Chem **275**(40): 31554.

Lindahl, T. (1993). "Instability and decay of the primary structure of DNA." Nature **362**(6422): 709-715.

Lindahl, T. and D. E. Barnes (2000). "Repair of endogenous DNA damage." Cold Spring Harb Symp Quant Biol **65**: 127-133.

Liou, G. Y. and P. Storz (2010). "Reactive oxygen species in cancer." Free Radic Res **44**(5): 479-496.

Lipshutz, R. J., S. P. Fodor, T. R. Gingeras and D. J. Lockhart (1999). "High density synthetic oligonucleotide arrays." Nat Genet **21**(1 Suppl): 20-24.

Liu, H., Z. G. Lu, Y. Miki and K. Yoshida (2007). "Protein kinase C delta induces transcription of the TP53 tumor suppressor gene by controlling death-promoting factor Btf in the apoptotic response to DNA damage." Mol Cell Biol **27**(24): 8480-8491.

Liu, S. T., J. C. Hittle, S. A. Jablonski, M. S. Campbell, K. Yoda and T. J. Yen (2003). "Human CENP-I specifies localization of CENP-F, MAD1 and MAD2 to kinetochores and is essential for mitosis." Nat Cell Biol **5**(4): 341-345.

Liu, W., B. Tanasa, O. V. Tyurina, T. Y. Zhou, R. Gassmann, W. T. Liu, K. A. Ohgi, C. Benner, I. Garcia-Bassets, A. K. Aggarwal, A. Desai, P. C. Dorrestein, C. K. Glass and M. G. Rosenfeld (2010). "PHF8 mediates histone H4 lysine 20 demethylation events involved in cell cycle progression." Nature **466**(7305): 508-512.

Loewenstein, Y., D. Raimondo, O. C. Redfern, J. Watson, D. Frishman, M. Linial, C. Orengo, J. Thornton and A. Tramontano (2009). "Protein function annotation by homology-based inference." Genome Biol **10**(2): 207.

Lorch, Y., J. W. LaPointe and R. D. Kornberg (1987). "Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones." Cell **49**(2): 203-210.

Luo, H., F. Gao and Y. Lin (2015). "Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes." Sci Rep **5**: 13210.

MacDonald, B. T., K. Tamai and X. He (2009). "Wnt/beta-catenin signaling: components, mechanisms, and diseases." Dev Cell **17**(1): 9-26.

Mailand, N., S. Bekker-Jensen, H. Faustrup, F. Melander, J. Bartek, C. Lukas and J. Lukas (2007). "RNF8 ubiquitylates histones at DNA double-strand breaks and promotes assembly of repair proteins." Cell **131**(5): 887-900.

Maiti, R., G. H. Van Domselaar, H. Zhang and D. S. Wishart (2004). "SuperPose: a simple server for sophisticated structural superposition." Nucleic Acids Res **32**(Web Server issue): W590-594.

Malle, E., P. G. Furtmuller, W. Sattler and C. Obinger (2007). "Myeloperoxidase: a target for new drug development?" Br J Pharmacol **152**(6): 838-854.

Marion, D. (2013). "An introduction to biological NMR spectroscopy." Mol Cell Proteomics **12**(11): 3006-3025.

Marnett, L. J. (2000). "Oxyradicals and DNA damage." Carcinogenesis **21**(3): 361-370.

Mason, D. A., D. E. Stage and D. S. Goldfarb (2009). "Evolution of the metazoan-specific importin alpha gene family." J Mol Evol **68**(4): 351-365.

Mates, J. M., C. Perez-Gomez and I. Nunez de Castro (1999). "Antioxidant enzymes and human diseases." Clin Biochem **32**(8): 595-603.

Mates, J. M. and F. Sanchez-Jimenez (1999). "Antioxidant enzymes and their implications in pathophysiologic processes." Front Biosci **4**: D339-345.

Matsuura, T., T. Yamagata, D. L. Burgess, A. Rasmussen, R. P. Grewal, K. Watase, M. Khajavi, A. E. McCall, C. F. Davis, L. Zu, M. Achari, S. M. Pulst, E. Alonso, J. L. Noebels, D. L. Nelson, H. Y. Zoghbi and T. Ashizawa (2000). "Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10." Nat Genet **26**(2): 191-194.

Maynard, S., S. H. Schurman, C. Harboe, N. C. de Souza-Pinto and V. A. Bohr (2009). "Base excision repair of oxidative DNA damage and association with cancer and aging." Carcinogenesis **30**(1): 2-10.

McAdams, H. H. and A. Arkin (1999). "It's a noisy business! Genetic regulation at the nanomolar scale." Trends Genet **15**(2): 65-69.

McPherson, A. (2004). "Introduction to protein crystallization." Methods **34**(3): 254-265.

McPherson, S. and F. J. Longo (1993). "Chromatin structure-function alterations during mammalian spermatogenesis: DNA nicking and repair in elongating spermatids." Eur J Histochem **37**(2): 109-128.

Meraldi, P., R. Honda and E. A. Nigg (2004). "Aurora kinases link chromosome segregation and cell division to cancer susceptibility." Curr Opin Genet Dev **14**(1): 29-36.

Metz, C. W. (1925). "Prophase Chromosome Behavior in Triploid Individuals of DROSOPHILA MELANOGASTER." Genetics **10**(4): 345-350.

Michaels, M. L. and J. H. Miller (1992). "The GO system protects organisms from the mutagenic effect of the spontaneous lesion 8-hydroxyguanine (7,8-dihydro-8-oxoguanine)." J Bacteriol **174**(20): 6321-6325.

Michalak, P. (2008). "Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes." Genomics **91**(3): 243-248.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler and R. Lanfear (2020). "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era." Mol Biol Evol **37**(5): 1530-1534.

Miotto, B., M. Chibi, P. Xie, S. Koundrioukoff, H. Moolman-Smook, D. Pugh, M. Debatisse, F. He, L. Zhang and P. A. Defossez (2014). "The RBBP6/ZBTB38/MCM10 axis regulates DNA replication and common fragile site stability." Cell Rep **7**(2): 575-587.

Mohsenzadegan, M. and A. Mirshafiey (2012). "The immunopathogenic role of reactive oxygen species in Alzheimer disease." Iran J Allergy Asthma Immunol **11**(3): 203-216.

Moorthy, B., C. Chu and D. J. Carlin (2015). "Polycyclic aromatic hydrocarbons: from metabolism to lung cancer." Toxicol Sci 145(1): 5-15.

Morrison, C., A. Shinohara, E. Sonoda, Y. Yamaguchi-Iwai, M. Takata, R. R. Weichselbaum and S. Takeda (1999). "The essential functions of human Rad51 are independent of ATP hydrolysis." Mol Cell Biol **19**(10): 6891-6897.

Mulder, N. J. and R. Apweiler (2002). "Tools and resources for identifying protein families, domains and motifs." Genome Biol **3**(1): REVIEWS2001.

Mullins, J. G. (2012). "Structural modelling pipelines in next generation sequencing projects." Adv Protein Chem Struct Biol **89**: 117-167.

Murzin, A. G., S. E. Brenner, T. Hubbard and C. Chothia (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." J Mol Biol **247**(4): 536-540.

Musacchio, A. and A. Desai (2017). "A Molecular View of Kinetochore Assembly and Function." Biology (Basel) **6**(1).

Nalepa, G. and D. W. Clapp (2018). "Fanconi anaemia and cancer: an intricate relationship." Nat Rev Cancer **18**(3): 168-185.

Nasmyth, K. and C. H. Haering (2009). "Cohesin: its roles and mechanisms." Annu Rev Genet **43**: 525-558.

Neuwald, A. F., L. Aravind, J. L. Spouge and E. V. Koonin (1999). "AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes." Genome Res **9**(1): 27-43.

Nguyen, L. T., H. A. Schmidt, A. von Haeseler and B. Q. Minh (2015). "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies." Mol Biol Evol **32**(1): 268-274.

Nicolae, C. M., E. R. Aho, A. H. Vlahos, K. N. Choe, S. De, G. I. Karras and G. L. Moldovan (2014). "The ADP-ribosyltransferase PARP10/ARTD10 interacts with proliferating cell nuclear antigen (PCNA) and is required for DNA damage tolerance." J Biol Chem **289**(19): 13627-13637.

Nikolaichik, Y. A. and W. D. Donachie (2000). "Conservation of gene order amongst cell wall and cell division genes in Eubacteria, and ribosomal genes in Eubacteria and Eukaryotic organelles." Genetica **108**(1): 1-7.

Nurse, P. (1975). "Genetic control of cell size at cell division in yeast." Nature **256**(5518): 547-551.

O'Brien, S. J. and J. A. Graves (1990). "Report of the committee on comparative gene mapping." Cytogenet Cell Genet **55**(1-4): 406-433.

O'Brien, S. J., M. Menotti-Raymond, W. J. Murphy, W. G. Nash, J. Wienberg, R. Stanyon, N. G. Copeland, N. A. Jenkins, J. E. Womack and J. A. Marshall Graves (1999). "The promise of comparative genomics in mammals." Science **286**(5439): 458-462, 479-481.

O'Driscoll, M. and P. A. Jeggo (2006). "The role of double-strand break repair - insights from human genetics." Nat Rev Genet **7**(1): 45-54.

Orengo, C. A., A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells and J. M. Thornton (1997). "CATH--a hierarchic classification of protein domain structures." Structure **5**(8): 1093-1108.

Osguthorpe, D. J. (2000). "Ab initio protein folding." Curr Opin Struct Biol **10**(2): 146-152.

Paiva, C. N. and M. T. Bozza (2014). "Are reactive oxygen species always detrimental to pathogens?" Antioxid Redox Signal **20**(6): 1000-1037.

Pal, C. and L. D. Hurst (2003). "Evidence for co-evolution of gene order and recombination rate." Nat Genet **33**(3): 392-395.

Palmer, D., F. Fabris, A. Doherty, A. A. Freitas and J. P. de Magalhaes (2021). "Ageing transcriptome meta-analysis reveals similarities and differences between key mammalian tissues." Aging (Albany NY) **13**(3): 3313-3341.

Pan, L., J. Penney and L. H. Tsai (2014). "Chromatin regulation of DNA damage repair and genome integrity in the central nervous system." J Mol Biol **426**(20): 3376-3388.

Pauling, L. and R. B. Corey (1951). "The polypeptide-chain configuration in hemoglobin and other globular proteins." Proc Natl Acad Sci U S A **37**(5): 282-285.

Paull, T. T., E. P. Rogakou, V. Yamazaki, C. U. Kirchgessner, M. Gellert and W. M. Bonner (2000). "A critical role for histone H2AX in recruitment of repair factors to nuclear foci after DNA damage." Curr Biol **10**(15): 886-895.

Pavlopoulou, A. and I. Michalopoulos (2011). "State-of-the-art bioinformatics protein structure prediction tools (Review)." Int J Mol Med **28**(3): 295-310.

Peifer, M., S. Berg and A. B. Reynolds (1994). "A repeating amino acid motif shared by proteins with diverse cellular roles." Cell **76**(5): 789-791.

Petermann, E., M. L. Orta, N. Issaeva, N. Schultz and T. Helleday (2010). "Hydroxyurea-stalled replication forks become progressively inactivated and require two different RAD51-mediated pathways for restart and repair." Mol Cell **37**(4): 492-502

Pilhofer, M., K. Rappl, C. Eckl, A. P. Bauer, W. Ludwig, K. H. Schleifer and G. Petroni (2008). "Characterization and evolution of cell division and cell wall synthesis genes in the bacterial phyla Verrucomicrobia, Lentisphaerae, Chlamydiae, and Planctomycetes and phylogenetic comparison with rRNA genes." J Bacteriol **190**(9): 3192-3202.

Plastaras, J. P., J. N. Riggins, M. Otteneder and L. J. Marnett (2000). "Reactivity and mutagenicity of endogenous DNA oxopropenylating agents: base propenals, malondialdehyde, and N(epsilon)-oxopropenyllysine." Chem Res Toxicol **13**(12): 1235-1242.

Poyatos, J. F. and L. D. Hurst (2007). "The determinants of gene order conservation in yeasts." Genome Biol **8**(11): R233.

Preston, B. D., B. Singer and L. A. Loeb (1986). "Mutagenic potential of O4-methylthymine in vivo determined by an enzymatic approach to site-specific mutagenesis." Proc Natl Acad Sci U S A **83**(22): 8501-8505.

Pulst, S. M. (1999). "Genetic linkage analysis." Arch Neurol **56**(6): 667-672.

Radicella, J. P., C. Dherin, C. Desmaze, M. S. Fox and S. Boiteux (1997). "Cloning and characterization of hOGG1, a human homolog of the OGG1 gene of Saccharomyces cerevisiae." Proc Natl Acad Sci U S A **94**(15): 8010-8015.

Radmanesh, F., A. O. Caglayan, J. L. Silhavy, C. Yilmaz, V. Cantagrel, T. Omar, B. Rosti, H. Kaymakcalan, S. Gabriel, M. Li, N. Sestan, K. Bilguvar, W. B. Dobyns, M. S. Zaki, M. Gunel and J. G. Gleeson (2013). "Mutations in LAMB1 cause cobblestone brain malformation without muscular or ocular abnormalities." Am J Hum Genet **92**(3): 468-474.

Raghavan, S. C. and M. R. Lieber (2007). "DNA structure and human diseases." Front Biosci **12**: 4402-4408.

Rancati, G., J. Moffat, A. Typas and N. Pavelka (2018). "Emerging and evolving concepts in gene essentiality." Nat Rev Genet **19**(1): 34-49.

Rao, M. S., T. R. Van Vleet, R. Ciurlionis, W. R. Buck, S. W. Mittelstadt, E. A. G. Blomme and M. J. Liguori (2018). "Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies." Front Genet **9**: 636.

Ratan, R., D. A. Mason, B. Sinnot, D. S. Goldfarb and R. J. Fleming (2008). "Drosophila importin alpha1 performs paralog-specific functions essential for gametogenesis." Genetics **178**(2): 839-850.

Riley, P. A. (1994). "Free radicals in biology: oxidative stress and the effects of ionizing radiation." Int J Radiat Biol **65**(1): 27-33.

Roberts, B. T., K. A. Farr and M. A. Hoyt (1994). "The Saccharomyces cerevisiae checkpoint gene BUB1 encodes a novel protein kinase." Mol Cell Biol **14**(12): 8282-8291.

Rogakou, E. P., D. R. Pilch, A. H. Orr, V. S. Ivanova and W. M. Bonner (1998). "DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139." J Biol Chem **273**(10): 5858-5868.

Sanchez-Carbayo, M., N. D. Socci, L. Richstone, M. Corton, N. Behrendt, J. Wulkfuhle, B. Bochner, E. Petricoin and C. Cordon-Cardo (2007). "Genomic and proteomic profiles reveal the association of gelsolin to TP53 status and bladder cancer progression." Am J Pathol **171**(5): 1650-1658.

Schilde, C., H. M. Lawal, A. A. Noegel, L. Eichinger, P. Schaap and G. Glockner (2016). "A set of genes conserved in sequence and expression traces back the establishment of multicellularity in social amoebae." BMC Genomics **17**(1): 871.

Schultz, J., F. Milpetz, P. Bork and C. P. Ponting (1998). "SMART, a simple modular architecture research tool: identification of signaling domains." Proc Natl Acad Sci U S A **95**(11): 5857-5864.

Seeberg, E., L. Eide and M. Bjoras (1995). "The base excision repair pathway." Trends Biochem Sci **20**(10): 391-397.

Selby, C. P. and A. Sancar (1997). "Human transcription-repair coupling factor CSB/ERCC6 is a DNA-stimulated ATPase but is not a helicase and does not disrupt the ternary transcription complex of stalled RNA polymerase II." J Biol Chem **272**(3): 1885-1890.

Shen, C. H. (2019). Molecular Diagnosis of Chromosomal Disorders, Diagnostic Molecular Biology.

Shimada, M. and M. Nakanishi (2013). "Response to DNA damage: why do we need to focus on protein phosphatases?" Front Oncol **3**: 8.

Sif, S. (2004). "ATP-dependent nucleosome remodeling complexes: enzymes tailored to deal with chromatin." J Cell Biochem **91**(6): 1087-1098.

Sim, K. L. and T. P. Creamer (2004). "Protein simple sequence conservation." Proteins **54**(4): 629-638.

Simonti, C. N. and J. A. Capra (2015). "The evolution of the human genome." Curr Opin Genet Dev **35**: 9-15.

Smogorzewska, A., S. Matsuoka, P. Vinciguerra, E. R. McDonald, 3rd, K. E. Hurov, J. Luo, B. A. Ballif, S. P. Gygi, K. Hofmann, A. D. D'Andrea and S. J. Elledge (2007). "Identification of the FANCI protein, a monoubiquitinated FANCD2 paralog required for DNA repair." Cell **129**(2): 289-301.

Spoerke, J. M., C. O'Brien, L. Huw, H. Koeppen, J. Fridlyand, R. K. Brachmann, P. M. Haverty, A. Pandita, S. Mohan, D. Sampath, L. S. Friedman, L. Ross, G. M. Hampton, L. C. Amler, D. S. Shames and M. R. Lackner (2012). "Phosphoinositide 3-kinase (PI3K) pathway alterations are associated with histologic subtypes and are predictive of sensitivity to PI3K inhibitors in lung cancer preclinical models." Clin Cancer Res **18**(24): 6771-6783.

Stucki, M., J. A. Clapperton, D. Mohammad, M. B. Yaffe, S. J. Smerdon and S. P. Jackson (2005). "MDC1 directly binds phosphorylated histone H2AX to regulate cellular responses to DNA double-strand breaks." Cell **123**(7): 1213-1226.

Subba Rao, K. (2007). "Mechanisms of disease: DNA repair defects and neurological disease." Nat Clin Pract Neurol **3**(3): 162-172.

Sullivan, K. F., M. Hechenberger and K. Masri (1994). "Human CENP-A contains a histone H3 related histone fold domain that is required for targeting to the centromere." J Cell Biol **127**(3): 581-592.

Supek, F., M. Bosnjak, N. Skunca and T. Smuc (2011). "REVIGO summarizes and visualizes long lists of gene ontology terms." PLoS One **6**(7): e21800.

Tamames, J. (2001). "Evolution of gene order conservation in prokaryotes." Genome Biol **2**(6): RESEARCH0020.

Tapia-Alveal, C., T. M. Calonge and M. J. O'Connell (2009). "Regulation of chk1." Cell Div **4**: 8.

Tatusov, R. L., E. V. Koonin and D. J. Lipman (1997). "A genomic perspective on protein families." Science **278**(5338): 631-637.

Taylor, E. M. and A. R. Lehmann (1998). "Conservation of eukaryotic DNA repair mechanisms." Int J Radiat Biol **74**(3): 277-286.

Tewari, R., E. Bailes, K. A. Bunting and J. C. Coates (2010). "Armadillo-repeat protein functions: questions for little creatures." Trends Cell Biol **20**(8): 470-481.

Thielmann, H. W., O. Popanda, L. Edler and E. G. Jung (1991). "Clinical symptoms and DNA repair characteristics of xeroderma pigmentosum patients from Germany." Cancer Res **51**(13): 3456-3470.

Thoresen, S. B., C. Campsteijn, M. Vietri, K. O. Schink, K. Liestol, J. S. Andersen, C. Raiborg and H. Stenmark (2014). "ANCHR mediates Aurora-B-dependent abscission checkpoint control through retention of VPS4." Nat Cell Biol **16**(6): 550-560.

Tiwana, M. S. and S. W. Leslie (2022). Anatomy, Abdomen and Pelvis, Testicle. StatPearls. Treasure Island (FL).

Tramontano, A. and V. Morea (2003). "Assessment of homology-based predictions in CASP5." Proteins **53 Suppl 6**: 352-368.

Trinklein, N. D., S. F. Aldred, S. J. Hartman, D. I. Schroeder, R. P. Otillar and R. M. Myers (2004). "An abundance of bidirectional promoters in the human genome." Genome Res **14**(1): 62-66.

van Dam, S., R. Cordeiro, T. Craig, J. van Dam, S. H. Wood and J. P. de Magalhaes (2012). "GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases." BMC Genomics **13**: 535.

Vanoosthuyse, V. and K. G. Hardwick (2009). "A novel protein phosphatase 1-dependent spindle checkpoint silencing mechanism." Curr Biol **19**(14): 1176-1181.

Venoux, M., J. Basbous, C. Berthenet, C. Prigent, A. Fernandez, N. J. Lamb and S. Rouquier (2008). "ASAP is a novel substrate of the oncogenic mitotic kinase Aurora-A: phosphorylation on Ser625 is essential to spindle formation and mitosis." Hum Mol Genet **17**(2): 215-224.

Virkki, M. T., C. Peters, D. Nilsson, T. Sorensen, S. Cristobal, B. Wallner and A. Elofsson (2014). "The positive inside rule is stronger when followed by a transmembrane helix." J Mol Biol **426**(16): 2982-2991.

Wang, C. and R. S. Swerdloff (1992). "Evaluation of testicular function." Baillieres Clin Endocrinol Metab **6**(2): 405-434.

Wang, D. G., J. B. Fan, C. J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M. S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, R. Lipshutz, M. Chee and E. S. Lander (1998). "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome." Science **280**(5366): 1077-1082.

Wang, G. and K. M. Vasquez (2006). "Non-B DNA structure-induced genetic instability." Mutat Res **598**(1-2): 103-119.

Wang, G. G., C. D. Allis and P. Chi (2007). "Chromatin remodeling and cancer, Part II: ATP-dependent chromatin remodeling." Trends Mol Med **13**(9): 373-380.

Watson, J. D. and F. H. Crick (1953). "The structure of DNA." Cold Spring Harb Symp Quant Biol **18**: 123-131.

Wilkins, M. R., E. Gasteiger, A. Bairoch, J. C. Sanchez, K. L. Williams, R. D. Appel and D. F. Hochstrasser (1999). "Protein identification and analysis tools in the ExPASy server." Methods Mol Biol **112**: 531-552.

Winterbourn, C. C. (2008). "Reconciling the chemistry and biology of reactive oxygen species." Nat Chem Biol **4**(5): 278-286.

Wolffe, A. P. (1998). "Packaging principle: how DNA methylation and histone acetylation control the transcriptional activity of chromatin." J Exp Zool **282**(1-2): 239-244.

Wong, J. W. H. (2019). "Assessing the Evolutionary Conservation of Protein Disulphide Bonds." Methods Mol Biol **1967**: 9-19.

Wu, C. H., H. Huang, L. S. Yeh and W. C. Barker (2003). "Protein family classification and functional annotation." Comput Biol Chem **27**(1): 37-47.

Xiang, Z. (2006). "Advances in homology protein structure modeling." Curr Protein Pept Sci **7**(3): 217-227.

Xu, A., L. J. Wu, R. M. Santella and T. K. Hei (1999). "Role of oxyradicals in mutagenicity and DNA damage induced by crocidolite asbestos in mammalian cells." Cancer Res **59**(23): 5922-5926.

Xu, L., M. Ali, W. Duan, X. Yuan, F. Garba, M. Mullen, B. Sun, I. Poser, H. Duan, J. Lu, R. Tian, Y. Ge, L. Chu, W. Pan, D. Wang, A. Hyman, H. Green, L. Li, Z. Dou, D. Liu, X. Liu and X. Yao (2021). "Feedback control of PLK1 by Apolo1 ensures accurate chromosome segregation." Cell Rep **36**(2): 109343.

Yang, J. and Y. Zhang (2015). "I-TASSER server: new development for protein structure and function predictions." Nucleic Acids Res **43**(W1): W174-181.

Yang, M. Q. and L. L. Elnitski (2008). "Prediction-based approaches to characterize bidirectional promoters in the mammalian genome." BMC Genomics **9 Suppl 1**: S2.

Yuan, J. and J. Chen (2013). "FIGNL1-containing protein complex is required for efficient homologous recombination repair." Proc Natl Acad Sci U S A **110**(26): 10640-10645.

Yuan, F., L. Song, L. Qian, J. J. Hu and Y. Zhang (2010). "Assembling an orchestra: Fanconi anemia pathway of DNA repair." Front Biosci (Landmark Ed) **15**(3): 1131-1149.

Zahn-Zabal, M., C. Dessimoz and N. M. Glover (2020). "Identifying orthologs with OMA: A primer." F1000Res **9**: 27.

Zannini, L., D. Delia and G. Buscemi (2014). "CHK2 kinase in the DNA damage response and beyond." J Mol Cell Biol **6**(6): 442-457.

Zerbino, D. R., P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Giron, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates and P. Flicek (2018). "Ensembl 2018." Nucleic Acids Res **46**(D1): D754-D761.

Zhang, L., Z. Yang, A. Ma, Y. Qu, S. Xia, D. Xu, C. Ge, B. Qiu, Q. Xia, J. Li and Y. Liu (2014). "Growth arrest and DNA damage 45G down-regulation contributes to Janus kinase/signal transducer and activator of transcription 3 activation and cellular senescence evasion in hepatocellular carcinoma." Hepatology **59**(1): 178-189.

Zhang, Y., A. K. Arakaki and J. Skolnick (2005). "TASSER: an automated method for the prediction of protein tertiary structures in CASP6." Proteins **61 Suppl 7**: 91-98.

Zhang, Z., Z. Tan, Q. Lv, L. Wang, K. Yu, H. Yang, H. Liang, T. Lu, Y. Ji, J. Chen, W. He, Z. Chen, S. Chen and X. Shen (2021). "High Expression of C1ORF112 Predicts a Poor Outcome: A Potential Target for the Treatment of Low-Grade Gliomas." Front Genet **12**: 710944.

Zhao, J., S. Gupta, M. Seielstad, J. Liu and A. Thalamuthu (2011). "Pathway-based analysis using reduced gene subsets in genome-wide association studies." BMC Bioinformatics **12**: 17.