BINF*6999 - Master Project

Integrative Analyses on RNA-seq and Proteomics Data of Glioblastomas Cell Lines Treated with A Selective Inhibitor of PRMT5 and A Control

Name: Duong, Bang Chi

ID: 0981462

Primary advisor: Dr. Panagiotis Prinos

Co-advisor: Dr. Lewis Lukens

Department of Mathematics and Statistics

Department of Integrative Biology

University of Guelph

Canada

August 4, 2017

Contents

1 Abstract **2**

2 Introduction **3**

3 Methods **6**

3.1 RNA sequencing 6

3.2 Proteomics 7

3.3 Differential Analysis 7

3.4 Enrichment Analysis 8

3.5 Protein-Protein Interaction (PPI) Network Analysis 10

4 Results and Discussion **11**

4.1 RNA sequencing 11

4.2 Proteomics 16

4.3 Integrative Analysis 21

5 Conclusion **26**

References **27**

1 Abstract

Three different Glioblastomas cell lines (G561, G564, G583) were treated with either GSK591, a PRMT5 inhibitor, or SGC2096, a corresponding inactive control. Transcriptomic (RNA-seq) and proteomic (quantitative label-free mass spectrometry, followed by MaxQuant data process procedure) experiments were carried out. Individual and integrative analyses were performed. We hypothesised that there is a difference in gene expression or proteomic levels compared between the two groups. Specifically, genes or proteins associated with cell cycle regulation are expected to be down-regulated in the GSK591 treated samples (i.e. effective treatment), compared to SGC2096 treated samples. In the case of RNA-seq, the library size is 26,485 genes. There were 1,032 differentially expressed genes found (471 down and 561 up-regulated genes) using edgeR package. For the proteomics data, the library size is 46,435 peptides or 4,533 proteins. There were 191 differentially expressed proteins found (98 down and 93 up-regulated proteins) using MSqRob package. A number of enrichment analyses were performed separately on the two datasets, including GO terms, KEGG pathways, InterPro, Pfam, and Gene Set Enrichment Analysis. For the integration analysis, intersection region of each enrichment pathway was investigated. On the gene level, 21 genes was found to be common in both differentially expressed sets. On the pathway level, it was found that the transcriptomic and proteomic data both agree on enrichment in metabolic pathways, cell cycle process, DNA replication, spliceosome, and immune system regulation.

2 Introduction

Brain tumours have a hierarchical cellular organisation which is an indication of a stem cell foundation [1]. Glioblastomas (GBM) are the most aggressive brain tumours arising from astrocytes [2]. They are highly malignant since they proliferate very rapidly and are nourished by a large network of blood vessels. There are two types of GBM: primary (de novo) and secondary (recurrent) tumours. Those tumours that belong to the former type are the most common (90% of GBM) and aggressive since they show up and reproduce quickly. The secondary GBM grow slower, they begin as lower-grade tumours and eventually become higher grade. There is a prominent involvement of epigenetics in the aetiology of these tumours.

Chemical probes are small molecules which can potently and selectively inhibit or antagonise the target protein *in vitro* with a defined mode of action [3]. The compounds help researchers connect selective antagonism or inhibition of a specific protein target with a biological and disease phenotype in cell-based assays, with high confidence. These probes demonstrate potential new drug discovery for oncology and therapeutic medicine, by allowing preclinical target validation in industrial as well as academic laboratories [4]. The intention for these chemical probes are preliminary target validation and phenotypic profiling studies in cell lines as well as primary patient samples cultured *in vitro*; as a result, individual probe has been illustrated to be cell-permeable and stable in cells [5–7]. Also, it has been demonstrated that the inhibition of the proposed target in the cell occurs at low µM concentrations or lower. There are three criteria for the SGC chemical probes: *in vitro* potency of less than 100 nM, more than 30-fold selectivity versus other subfamilies, and demonstration of on-target in cells at 1 µM. Nevertheless, these chemicals may not necessarily have favourable pharmacokinetic properties for *in vivo* studies. The Structural Genomics Consortium (SGC) Epigenetic Chemical Probe Library currently has more than thirty well-characterised chemicals that selectively and potently inhibit or antagonise specific chromatin regulatory proteins or domains involved in epigenetic control, including protein methyltransferases, demethylases, and bromodomains [8]. Most of them are paired with a control compound that is structurally similar to the active probe, but is much less active or inactive. They are essential to control

for potential off-target effects.

Three classes of epigenetic proteins have been discovered by studies of post-translational modification (PTM) of lysine and/or arginine residues in histones and other proteins, through methylation and acetylation. The first class is writers which add a methyl or acetyl group using S-adenosylmethionine and Acetyl-CoA as co-factors, respectively in methylation and acetylation. Up to three methyl groups are transferred to the e-amino group of a lysine residue by histone methyltransferases; and up to two methyl groups are transferred to the guanidine group of an arginine residue [9]. In the case of arginine dimethylation, symmetrical or unsymmetrical orientation can occur. In the case of acetylation, only lysine residues are acetylated and the event only happens once. The second class is erasers which remove a methyl or acetyl group. And the third class is readers which bind histones or proteins containing a particular PTM.

Protein arginine methyltransferases (PRMTs) are importantly involved in a number of biological processes [10]. Studies have implied that various human diseases (including cancer) are associated with overexpression of PRMTs [11–13]. Nine PRMTs have been identified and are categorised into three groups: types I, II, and III. Type II PRMTs (5 and 9) catalyses symmetric dimethylation of arginine residues, of which PRMT5 is the predominant enzyme. Interactions of PRMT5 with a number of binding partners characterise the enzyme's substrate specificity. Notably, PRMT5 methyltransferase activity requires MEP50, which is a member of the WD40 family of proteins.

Studies have reported that PRMT5 plays a role in mantle cell lymphoma (MCL), shown by its up-regulation in patient samples [14, 15]. Therefore, designing a chemical probe for PRMT5 would be very useful in testing therapeutic and biological hypotheses. GlaxoSmithKline and Epizyme co-developed the first compound of PRMT5 [16]. SGC has received an analog of the aforementioned probe as GSK591 [17] for distribution. At IC50 = 11 nM in an *in vitro* biochemical assay, the chemical probe potently inhibits the PRMT5/MEP50 complex from methylating histone H4 [17]. It also inhibits the symmetric arginine methylation of SmD3, with EC50 = 56 nM, in the case of Z-138 cells [17]. Compared to other methyltransferases, GSK591 is selective for PRMT5 up to 50 µM [17]. A control compound for GSK591 is

SGC2096 which is inactive up to 10 µM.

Several research groups, including ours, are capable of deriving and maintaining primary GBM lines derived from individual tumours grown either as neurospheres or adherent monolayers, which have been shown to retain stem-cell characteristics and the capacity to form tumours when serially transplanted in recipient immunocompromised mice. Moreover, these can be propagated *in vitro* and screened using chemical probes in a high-throughput fashion. We are using these patient-derived tumour lines for extensive genomic, epigenomic, transcriptomic, proteomic and metabolomic profiling as well as functional phenotypic analysis in order to gain insights into their biology and find molecular vulnerabilities for therapeutic intervention. Moreover, we are functional screening these cells with a library of epigenetics chemical probes and integrating these data with all the omics data in order to identify potential drug targets and define biomarkers of response to allow personalised medicine.

The primary goal of the project is integrative analyses of different omics data types that are currently being generated for the first cohort of thirty primary GBM lines. Through our functional screening, we have identified chemical probes that show promising effects, i.e. inhibition of cell proliferation of several primary GBM lines. In an effort to validate the observed phenotypes and better understand the mechanism of action, we have performed RNA-seq and proteomic profiling experiments in three GBM lines (G516, G564, and G583) treated with GSK591, a PRMT5 inhibitor, and an inactive negative control which is SGC2096. Individual and integrative differential analyses of the RNA-seq and proteomic datasets were performed. The analyses utilised various algorithms and R packages (limma, edgeR, MSqRob) to identify significant hits, i.e. genes and proteins that are affected by GSK591 treatment. Multiple enrichment analyses were performed including GO terms, KEGG pathways, Pfam, InterPro, and Gene Set Enrichment Analysis (GSEA), using various methods such as STRING database, and GSEA-Cytoscape. We hypothesised that cell cycle regulated genes are down-regulated in the GSK591 treated samples, compared to the control group. The expectation for the integrative analyses is that the two omics levels will agree with each other to some degree, and more insights regarding epigenetics mechanism of action can be unraveled.

3 Methods

3.1 RNA sequencing

3.1.1 Library Preparation

Approximately 5 million cells were plated in each well of a 6-well plate on day 1. Three different cell lines G561, G564, and G583 each was treated with 1 µM of GSK591 or control probe SGC2096 for 3 days, after which mRNA was prepared from cell pellets by using an RNeasy minikit (Qiagen, Valencia, CA) according to manufacturer's protocol including the on-column DNase digestion step. 2 micrograms of RNA was sequenced using the Illumina HiSeq 2500 instrument, on a high throughput flowcell, V4 chemistry. Sequencing was done as paired end, with read length of 126-base. The library was stranded using the NEBNext Ultra Directional RNA library prep kit.

3.1.2 RNA-seq Read Processing

In this study, FastQC [18] was used to examine the quality of the reads. Average library size for each sample is as follows: G561-GSK591 with 406-bp, G561-SGC2096 with 387-bp, G564-GSK591 with 397-bp, G564-SGC2096 with 405-bp, G583-GSK591 with 399-bp, and G583-SGC2096 with 386-bp. In terms of gene library size, there were 26,485 genes.

3.1.3 Alignment

The three softwares needed were STAR [19], SAMtools [20], and HTSeq [21]. STAR is a spliced read alignment software for RNA-seq, SAMtools is a program for interacting and manipulating sequencing data, and HTSeq was used to generate read counts from alignment files. Outputs from HTSeq were used for differential analysis (using edgeR [22]) and subsequent enrichment analyses.

STAR was used to create an index from the reference genome files. Indexing ensured that alignment happened in a reasonable amount of time. It is important to check that the FASTA and the GTF files are compatible. STAR was then used to align the reads. Next,

SAMtools was used to sort by gene name the BAM files generated from STAR, and index the sorted files. Lastly, HTSeq was used to generate raw read counts from genes and it only accounts for non-overlapping features.

3.2 Proteomics

3.2.1 Library Preparation, Mass Spectrometry and MaxQuant Process

Three different cell lines (G561, G564 and G583) were treated with PRMT5 inhibitors (GSK591 or LLY283) or inactive compound (SGC2096) for 7 days. Each condition was represented by three replicates (four replicates for G564 cell line). Cells were lysed in denaturing urea buffer, proteins were digested into peptides with trypsin. After desalting on C18, peptides were analysed by 14-hour DLC runs on Orbitrap Fusion (approximately 1 µg/injection). MaxQuant [23] was used for peptide identification, protein inference and label-free quantification (LFQ). There were 46,435 peptides or 4,533 proteins indentified by MaxQuant. Subsequent data analysis steps were done in R [24] and STRING [25].

3.3 Differential Analysis

For RNA-Seq data, edgeR package was used for the differential analysis. The generalised linear model used was based on the negative binomial distribution. Original library size was 26,485 genes, of which 13,636 genes were retained for further analysis, where only genes that have more than 1 count per million (cpm) in at least 3 samples were kept. Maximum likelihood test was used to find differentially expressed genes. False discovery rate (FDR) threshold was chosen to be 5% for multiple hypothesis testing, and the adjustted p-values would be called q-values. The filtering method was acceptable for both statistical reason, where genes with very low counts across all libraries provide little evidence for differential expression, and biological reason, where a gene must be expressed at some minimal level before it is likely to be translated into a protein or to be biological significant.

For proteomic data, a robust ridged peptide-based model was used [26]. Many studies have demonstrated that the peptide-based model approach is a much more reliable and accurate

method for differential analysis, in contrast to the summarisation-based approach which only looks at the protein level and discards peptide information [27–31]. Proteomic data suffers from non-random missing values, leading to sparsity issue and unstable residual variance estimates. It also suffers from heteroscedastic variance [27]. The proposed model is an extension of a linear (mixed) regression model. It provides an improvement on the estimation of the model parameters via: 1. a ridge regression which shrinks the log2 fold change (LFC) estimates but the LFCs become more stable; 2. empirical Bayes estimation of the variance which further stabilises variance estimators; and 3. M-estimation with Huber weights which reduces the impact of peptide intensity outliers. The degrees of freedom are calculated using the trace of the hat matrix. The response variable is the quantile normalised of log2 of LFQ intensity. MSqRob package in R provides an efficient analysis pipeline for the robust ridge peptide-based model [26]. Standard data filtering was also carried out based on the MaxQuant output files. The filtering criteria were removal of contaminants, removal of reverse strands, and removal of proteins that were only identified by a modification site. This resulted in 26,280 peptides or 3,702 proteins. Again, FDR threshold for differentially expressed proteins was kept at 5% for the comparison between samples treated with GSK591 and those treated with SGC2096. Furthermore, only those proteins that had corresponding genes (identified by MaxQuant) were kept. A protein may have more than one gene associated with it, but subsequent analyses required a list of unique gene symbols, thus only the first gene was retained since MaxQuant put the first gene as the most relevant one. This may result in two unique proteins having the same associated gene, therefore the gene with a more significant q-value was chosen.

3.4 Enrichment Analysis

3.4.1 Gene Set Enrichment Analysis (GSEA)

Extracting biological mechanisms insight from omics analysis information poses many challenges. First of all, a long list of statistically significant genes from the differential analysis very often does not point towards a major biological theme. Cellular processes usually influence sets of genes working in concordance, thus a small change in all genes

participating in a metabolic pathway may significantly affect the pathway mechanism; and sometimes, this observation is more important than a large fold change observed for a single gene. It is shown in [32] that two lists of differentially expressed genes from two different groups studying the same biological system had few overlapping genes.

Gene Set Enrichment Analysis [33] provides a method to find pathways and processes, making interpretation for large-scale experiments easier. If members of a gene set show a strong cross-correlation, GSEA can boost the signal-to-noise ratio. This helps detect moderate changes. There are two main differences between GSEA and single-gene analysis. First, all genes are considered in GSEA, whilst only those below an arbitrary cut-off probability (for example, a 5% FDR threshold) are looked at in single-gene analysis. Second, a more accurate null model is provided because correlations between genes are preserved in the significance assessment.

There are three main steps in GSEA. First, one needs to calculate an Enrichment Score (ES) of a set S based on a ranked list of genes L. The genes can be ranked based on different methods [33], in this study, the genes were ranked based on differential analysis output ("rank" = "LFC" times "-log(q-value)"). ES shows how much a set S is overrepresented of the ranked list L. It is calculated by walking down L, increasing a statistic if genes in S are encountered, and decreasing if genes in S are not encountered. ES is the maximum deviation from zero encountered in the random walk, which is very similar to Kolmogorov–Smirnov statistic [34]. Second, significance level (nominal p-value) of ES is estimated using an empirical class-based permutation test, which retains the gene correlation structure. Lastly, multiple hypothesis testing adjustment is performed by first normalising the ES, producing a normalised ES (NES) for each gene set that takes into account the size of each set. FDR corresponding to each NES is then calculated. GSEA was carried out in R through Java command line for both RNA-seq and proteomic datasets. Threshold for FDR was kept at 5%. The results were then visualised by EnrichmentMap, clusterMaker2 and AutoAnnotate applications in Cytoscape [35–37], where clustering method was performed using Markov Cluster (MCL) algorithm [38].

3.4.2 Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways, Protein Family (Pfam and InterPro) Enrichments

For individual transciptomic and proteomic dataset, the GO terms, KEGG pathways, Pfam and InterPro enrichments analyses were all performed using STRINGdb [25] package downloaded from Bioconductor in R. The reference list for each dataset was chosen to be all genes or proteins available from the RNA-seq or proteomic experiment accordingly. The input was all the differentially expressed genes or proteins that are kept at 5% false discovery rate. Multiple hypothesis testing was also performed on the enrichments and the cut-off FDR was also kept at 5%. Results were shown as word clouds, where the size of each term represents the significance of the term, and accordingly -log(q-value) was used as term frequency. In the case of GO terms and protein family domains enrichments, there were many instances that can be clustered together for a higher hierarchical representative. Hence, unsupervised clustering was performed using MCL algorithm [38] which can be accessed directly from Cytoscape [35]. For the integration analysis on these enrichments, overlapping terms were presented, and clustering was carried out if many generic and/or related terms occured (especially for the GO terms enrichment).

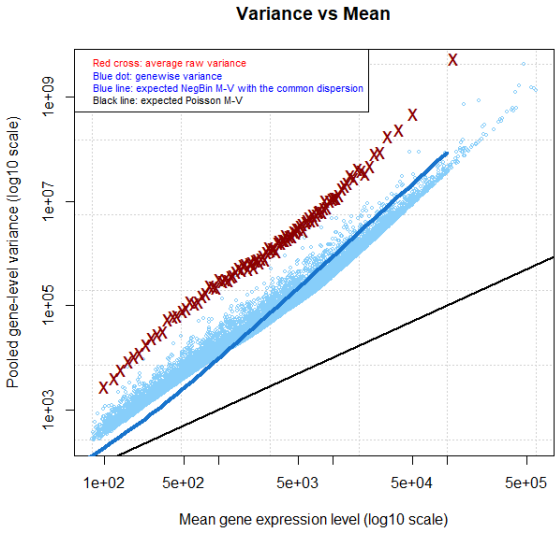
3.5 Protein-Protein Interaction (PPI) Network Analysis

The PPI network is constructed from the STRING database [25]. It shows genome or protein-wide interactions, and is derived from large-scale screens or analyses of multiple data sets. A combined score from both biological evidence and machine learning scores is assigned to each interaction ranging from 0 to 1 representing lowest to highest confidence or reliability of the (functional) interaction between proteins. Only the top 100 differentially expressed genes or proteins were mapped onto the PPI network.

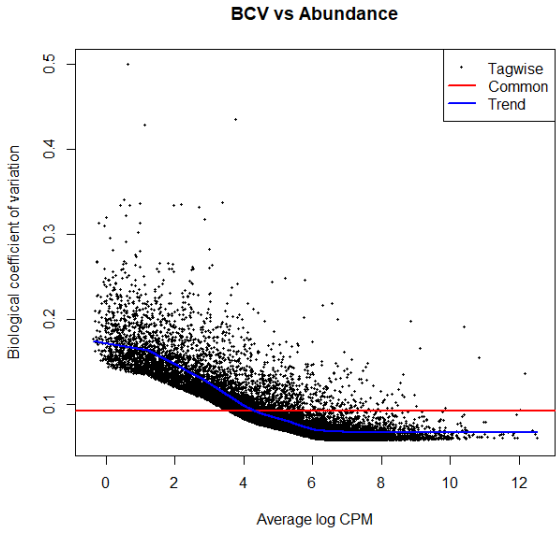
4 Results and Discussion

4.1 RNA sequencing

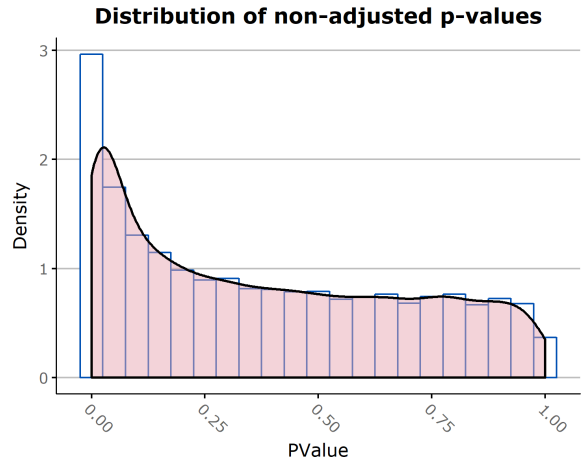
4.1.1 Quality Control



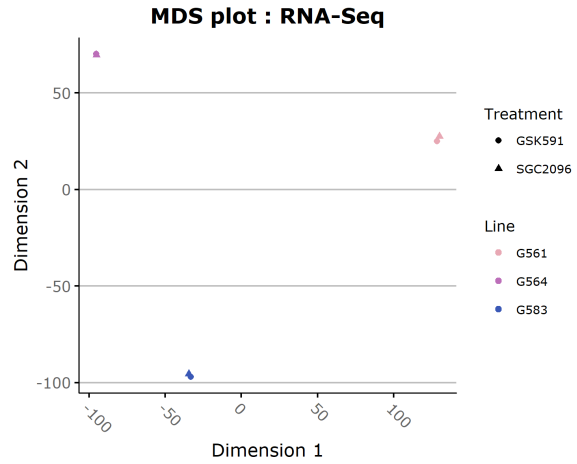
(a) Variance-Mean plot



(b) Biological Coefficient of Variation plot



(c) Histogram of non-adjusted p-values



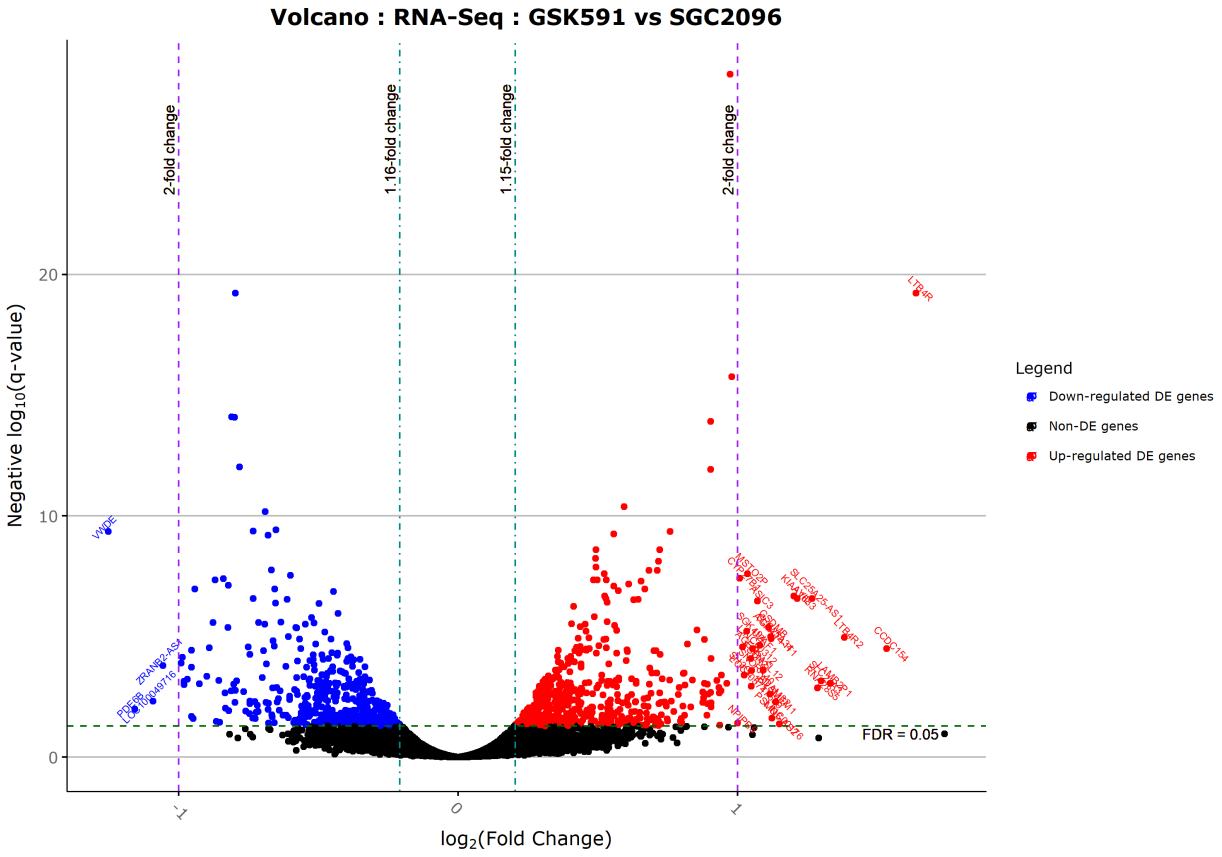
(d) Multidimensional scaling plot

Figure 1: **Quality control plots for RNA-seq data**

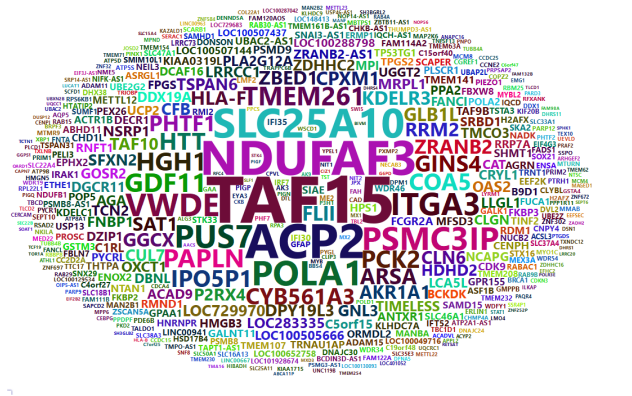
Figures 1a and 1b show that the read counts followed a negative binomial distribution, with a common dispersion parameter of 0.008562743, and hence 0.09253509 was the common biological coefficient of variation. Figure 1c shows a right-skew histogram, indicating that the analysis was done quite successfully with many differentially expressed genes detected

(p-value near zero). Figure 1d suggests that samples from the same cell line behaved more similarly compared to those from different cell lines (dimension 1 separation).

4.1.2 Differentially Expressed Genes



(a) Volcano plot where blue represents down-regulation, and red represents up-regulation



(b) Word cloud for down-regulated genes (fold change > 1.16 and FDR < 0.05)

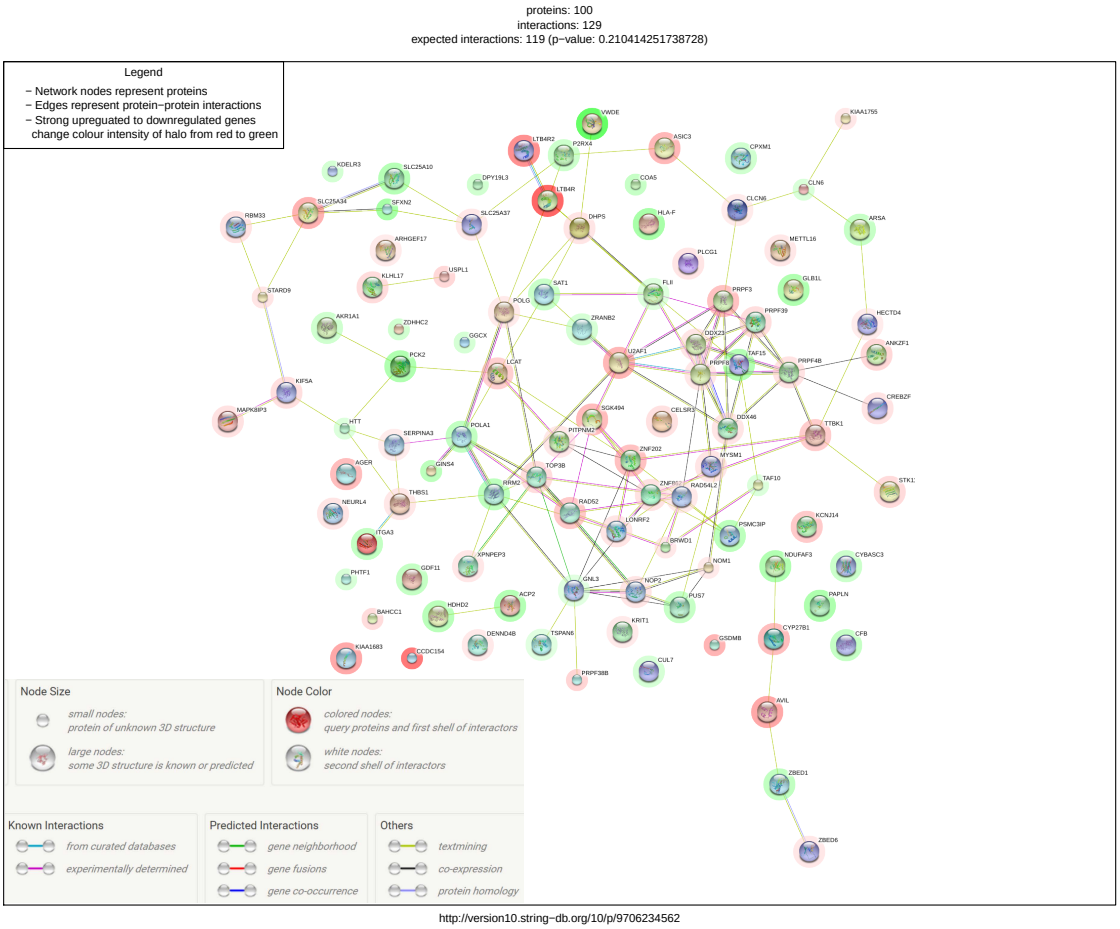
(c) Word cloud for up-regulated genes (fold change > 1.15 and FDR < 0.05)

Figure 2: **Differentially expressed genes detected in RNA-seq data using edgeR**

There were 1,032 differentially expressed genes found in RNA-seq data, of which 471 were down regulated and 561 were up regulated (fold change > 1.15 and FDR < 0.05). Figures 2a, b, and c show the volcano plot, and word clouds for down and up-regulated genes, respectively. Many genes were detected to change more than 2-fold in GSK591 treated compared to SGC2096 treated samples; for instance, LTB4R and LTB4R2 (involved in immune response) were up-regulated, and ZRANB2-AS1, LOC100049716 (both are affiliated with non-coding RNA class), VWDE (involved in calcium ion binding), and PDE6B (involved in metabolism of fat-soluble vitamins) were down-regulated. Also, TAF15 (TATA-Box Binding Protein Associated Factor 15) and PRPF3 (Pre-MRNA Processing Factor 3) were among the top significant hits. As aforementioned in Section 3.4, a large fold change in a single gene may not reveal much about biological significances since very often multiple genes work together in harmony for a biological pathway. Thus interactions and pathways should be investigated.

4.1.3 Protein-Protein Interaction Network

Proteins interact with each other to display their funtional roles. Thus constructing a PPI network for the DEGs was carried out. The analysis used the STRING database. 1,116 of the 13,636 retained genes were not identified by STRING, of these 119 were differentially expressed (FDR < 0.05). Therefore the number of DEGs reduced to 913 from 1,032 genes. Figure 3 shows PPI network for the top 100 DEGs. The interaction scores displayed in Figure 3b represent combination of probabilities from different evidence channels and then corrected for the probability of randomly observing an interaction. Proteins with top interactions score were NOP2, GNL3, PRPF8, PRPF3, DDX23, POLA1, GINS4, U2AF1, RRM2, ITGA3, THBS1, LTB4R, and LTB4. The associated annotations were shown in Figure 3b. Some of the proteins do not have well-characterised 3D structures (small nodes) such as STARD9 and GINS4. Another observation was that the interaction network looks sparse compared to PPI network obtained from the proteomics data (Figure 7a). Further studies into interaction-dense region may discover biological significance, which may give ideas on therapeutic treatment improvement.



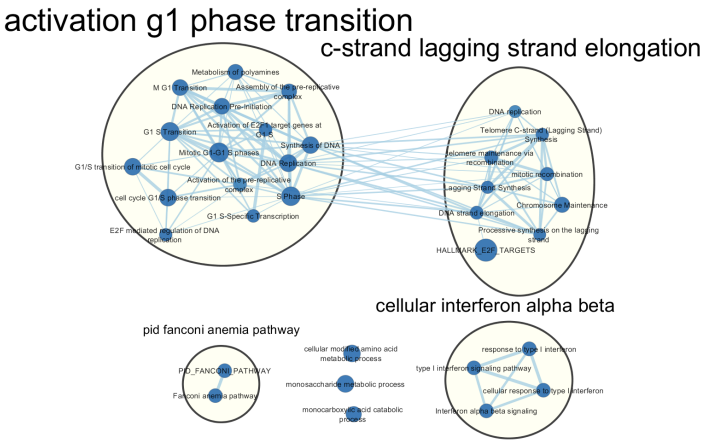
(a) PPI Network for the top 100 DEGs. A link to the network stored online is provided.

node1	node2	node1 accession	node2 accession	node1 annotation	node2 annotation	score
NOP2	GNL3	ENSP00000382392	ENSP00000395772	*NOP2 nucleolar protein homolog (yeast);...*	*guanine nucleotide binding protein-like 3...*	0.999
GNL3	NOP2	ENSP00000395772	ENSP00000382392	*guanine nucleotide binding protein-like 3...*	*NOP2 nucleolar protein homolog (yeast);...*	0.999
PRPF8	DDX23	ENSP00000304350	ENSP00000310723	*PRP8 pre-mRNA processing factor 8 ho...*	*DEAD (Asp-Glu-Ala-Asp) box polypeptide...*	0.995
DDX23	PRPF8	ENSP00000310723	ENSP00000304350	*DEAD (Asp-Glu-Ala-Asp) box polypeptide...*	*PRP8 pre-mRNA processing factor 8 ho...*	0.995
PRPF8	PRPF3	ENSP00000304350	ENSP00000315379	*PRP8 pre-mRNA processing factor 8 ho...*	*PRP3 pre-mRNA processing factor 3 ho...*	0.992
PRPF3	PRPF8	ENSP00000315379	ENSP00000304350	*PRP3 pre-mRNA processing factor 3 ho...*	*PRP8 pre-mRNA processing factor 8 ho...*	0.992
POLA1	GINS4	ENSP00000368349	ENSP00000276533	*polymerase (DNA directed), alpha 1, catal...*	*GINS complex subunit 4 (Sld5 homolog);...*	0.986
GINS4	POLA1	ENSP00000276533	ENSP00000368349	*GINS complex subunit 4 (Sld5 homolog);...*	*polymerase (DNA directed), alpha 1, catal...*	0.986
U2AF1	PRPF8	ENSP00000291552	ENSP00000304350	*U2 small nuclear RNA auxiliary factor 1;...*	*PRP8 pre-mRNA processing factor 8 ho...*	0.969
PRPF8	U2AF1	ENSP00000304350	ENSP00000291552	*PRP8 pre-mRNA processing factor 8 ho...*	*U2 small nuclear RNA auxiliary factor 1;...*	0.969
RRM2	POLA1	ENSP00000353770	ENSP00000368349	*ribonucleotide reductase M2; Provides th...*	*polymerase (DNA directed), alpha 1, catal...*	0.964
POLA1	RRM2	ENSP00000368349	ENSP00000353770	*polymerase (DNA directed), alpha 1, catal...*	*ribonucleotide reductase M2; Provides th...*	0.964
PRPF8	PRPF4B	ENSP00000304350	ENSP00000337194	*PRP8 pre-mRNA processing factor 8 ho...*	*PRP4 pre-mRNA processing factor 4 ho...*	0.959
PRPF4B	PRPF8	ENSP00000337194	ENSP00000304350	*PRP4 pre-mRNA processing factor 4 ho...*	*PRP8 pre-mRNA processing factor 8 ho...*	0.959
U2AF1	DDX23	ENSP00000291552	ENSP00000310723	*U2 small nuclear RNA auxiliary factor 1;...*	*DEAD (Asp-Glu-Ala-Asp) box polypeptide...*	0.927
DDX23	U2AF1	ENSP00000310723	ENSP00000291552	*DEAD (Asp-Glu-Ala-Asp) box polypeptide...*	*U2 small nuclear RNA auxiliary factor 1;...*	0.927
THBS1	ITGA3	ENSP00000260356	ENSP00000007722	*thrombospondin 1; Adhesive glycoprotei...*	*integrin, alpha 3 (antigen CD49C, alpha 3...*	0.919
ITGA3	THBS1	ENSP00000007722	ENSP00000260356	*integrin, alpha 3 (antigen CD49C, alpha 3...*	*thrombospondin 1; Adhesive glycoprotei...*	0.919
LTB4R2	LTB4R	ENSP00000433290	ENSP00000307445	*leukotriene B4 receptor 2; Low-affinity re...*	*leukotriene B4 receptor; Receptor for extr...*	0.914
LTB4R	LTB4R2	ENSP00000307445	ENSP00000433290	*leukotriene B4 receptor; Receptor for extr...*	*leukotriene B4 receptor 2; Low-affinity re...*	0.914

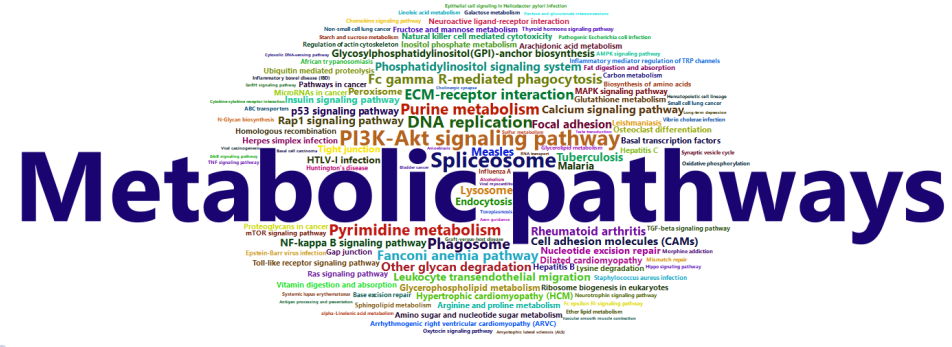
(b) PPI Score for the top 100 DEGs

Figure 3: **Protein-Protein interaction network for the top 100 DEGs**

4.1.4 Enrichment Analyses



(a) A gene set network plot where each node represents a gene set, whose colour is blue if down-regulated or red if up-regulated. The size of each node represents the gene set's size.



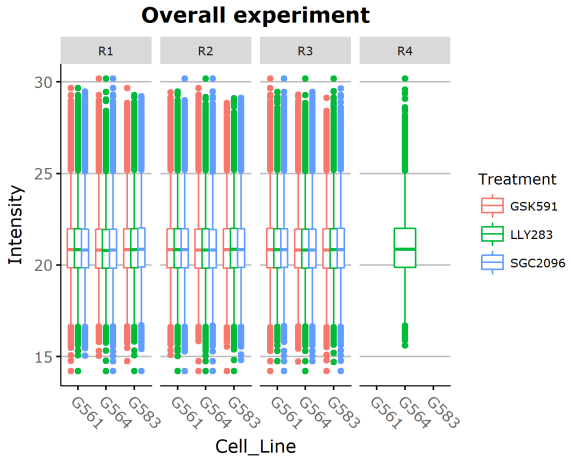
(b) KEGG pathways enrichment. The size of each term represents the significance of the term.

Figure 4: **GSEA and KEGG pathways enrichment for RNA-seq data**

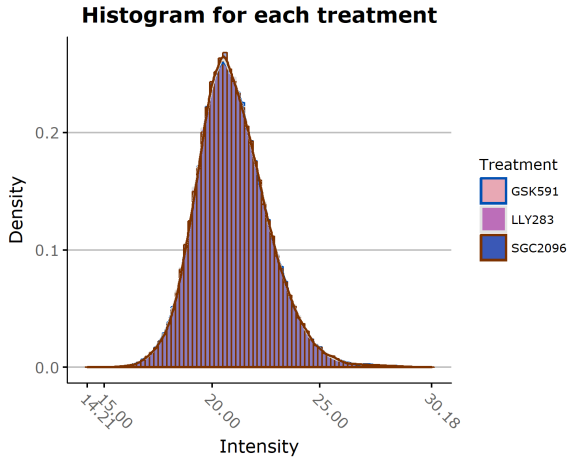
Figure 4a shows the result from gene set enrichment analysis performed on the RNA-seq data. Many gene sets involved in cell cycle regulation were found such as DNA replication, metabolism of polyamines, S/G1 phase; also there were gene sets involved in fanconi pathway, metabolic processes of cellular modified amino acid and monosaccharide, and those involved with interferon signaling pathways (immune response regulation). Figure 4b gives the KEGG pathways that were significant at 5% FDR threshold. A term's size represents how significant the term is. Those enriched were metabolic pathways, PI3K-Akt signaling pathway which is important in regulating cell cycle, spliceosome, phagosome and NF-kappa B signaling pathway which are involved in immune system regulation. The enrichments are very sensible since it was expected that the PRMT5 inhibitor (GSK591) would interfere with the cell proliferation process and block the spliceosome assembly [39].

4.2 Proteomics

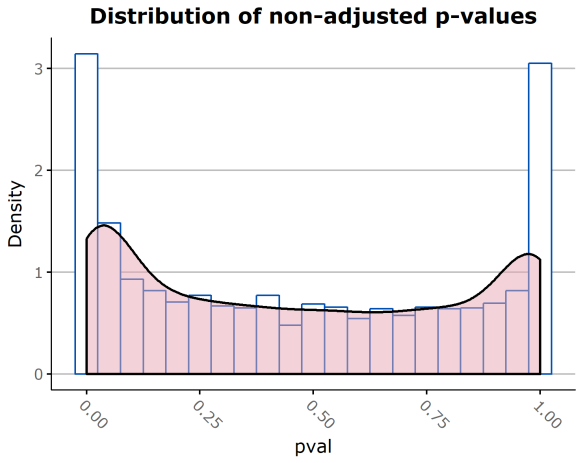
4.2.1 Quality Control



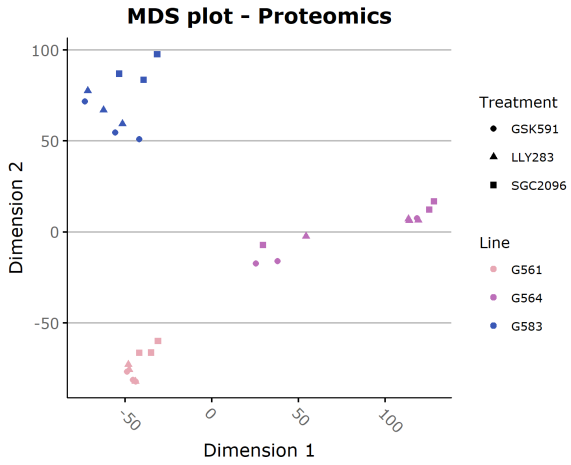
(a) Boxplot of peptide log2(intensity) after quantile normalisation



(b) Overlaid histogram of peptide log2(intensity) after quantile normalisation



(c) Histogram of non-adjusted p-values



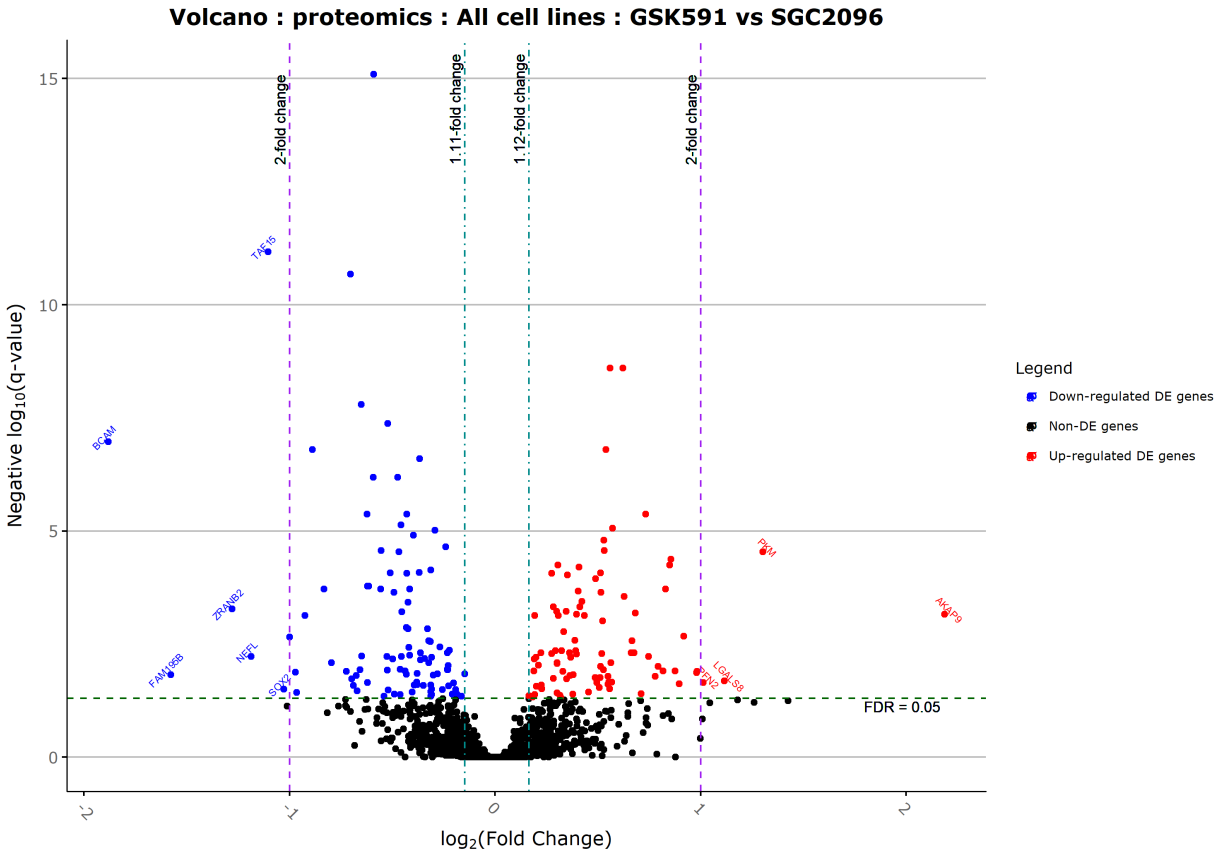
(d) Multidimensional scaling plot

Figure 5: **Quality control plots for the proteomic data**

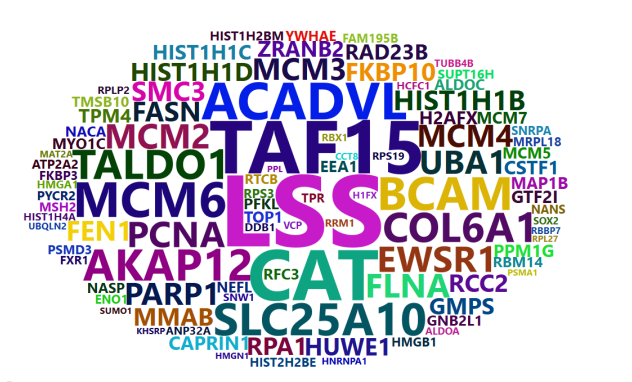
Figures 1a and 1b show that there were many extreme values (as expected from a proteomic data) and that the quantile normalisation successfully transformed the data. Hence the use of a linear robust ridge model (Section 3.3) was appropriate. Figure 1c shows a U-shape histogram on the non-adjusted p-value, suggesting that the analysis was performed successfully. Figure 1d suggests that samples from the same cell line behaved more similarly compared to those from different cell lines (dimension 1 separation). Notably, G564 cell line samples showed high variations, where 4 samples clustered around 50 and the other

6 samples clustered around 125 on dimension 1. Dimension 2 shows treatment separation, where PRMT5 inhibitor treated samples (LLY283 and GSK591) clustered together for each cell line and so did the control treated samples (SGC2096).

4.2.2 Differentially Expressed Proteins



(a) Volcano plot where blue represents down-regulation, and red represents up-regulation



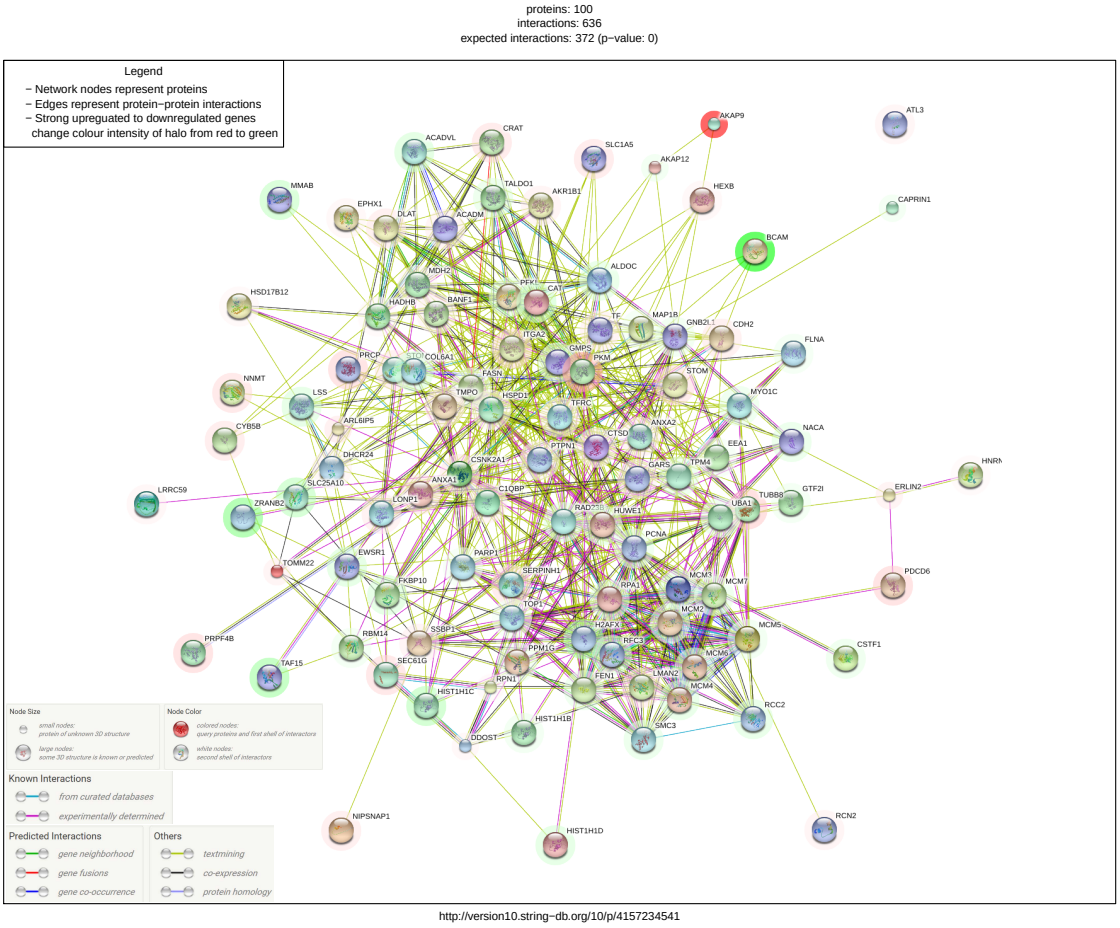
(b) Word cloud for down-regulated genes (fold change > 1.11 and FDR < 0.05)

(c) Word cloud for up-regulated genes (fold change > 1.12 and FDR < 0.05)

Figure 6: **Differentially expressed proteins detected in the proteome using MSqRob package**

There were 191 differentially expressed proteins or genes detected (filtering on the differential result was applied as described in Section 3.3), with 471 down and 561 up-regulated genes. Figures 6a, b, and c show the volcano plot, and word clouds for down and up-regulated genes, respectively. It was observed that TAF15 showed strong evidence (fold change = 2.150 and q-value = 6.740e-12), which was similar to the RNA-seq differential result for TAF15. Section 4.3 demontrates an in-depth comparative analysis for the two omics data.

4.2.3 Protein-Protein Interaction Network



(a) PPI Network for the top 100 DEGs. A link to the PPI network stored online is provided where users can customise their choice of scoring system, as well as looking at different enrichment analyses including GO terms, KEGG pathways, and protein families and domains with InterPro and Pfam databases.

node1	node2	node1 accession	node2 accession	node1 annotation	node2 annotation	score
TFRC	TF	ENSP00000353224	ENSP00000385834	transferrin receptor (p90, CD71); Cellular u...	transferrin; Transferrins are iron binding tr...	0.999
TF	TFRC	ENSP00000385834	ENSP00000353224	transferrin; Transferrins are iron binding tr...	transferrin receptor (p90, CD71); Cellular u...	0.999
RPN1	DDOST	ENSP00000296255	ENSP00000364188	ribophorin I; Essential subunit of the N-olig...	dolichyl-diphosphooligosaccharide--protei...	0.999
RPA1	PCNA	ENSP00000254719	ENSP00000368438	replication protein A1, 70kDa; Plays an ess...	proliferating cell nuclear antigen; Auxiliary...	0.999
RFC3	PCNA	ENSP00000369411	ENSP00000368438	replication factor C (activator 1) 3, 38kDa;...	proliferating cell nuclear antigen; Auxiliary...	0.999
PCNA	RPA1	ENSP00000368438	ENSP00000254719	proliferating cell nuclear antigen; Auxiliary...	replication protein A1, 70kDa; Plays an ess...	0.999
PCNA	FEN1	ENSP00000368438	ENSP00000305480	proliferating cell nuclear antigen; Auxiliary...	flap structure-specific endonuclease 1; Str...	0.999
PCNA	RFC3	ENSP00000368438	ENSP00000369411	proliferating cell nuclear antigen; Auxiliary...	replication factor C (activator 1) 3, 38kDa;...	0.999
MCM7	MCM5	ENSP00000307288	ENSP00000216122	minichromosome maintenance complex c...	minichromosome maintenance complex c...	0.999
MCM7	MCM3	ENSP00000307288	ENSP00000229854	minichromosome maintenance complex c...	minichromosome maintenance complex c...	0.999
MCM7	MCM6	ENSP00000307288	ENSP00000264156	minichromosome maintenance complex c...	minichromosome maintenance complex c...	0.999
MCM7	MCM4	ENSP00000307288	ENSP00000262105	minichromosome maintenance complex c...	minichromosome maintenance complex c...	0.999
MCM7	MCM2	ENSP00000307288	ENSP00000265056	minichromosome maintenance complex c...	minichromosome maintenance complex c...	0.999
MCM6	MCM7	ENSP00000264156	ENSP00000307288	minichromosome maintenance complex c...	minichromosome maintenance complex c...	0.999
MCM6	MCM3	ENSP00000264156	ENSP00000229854	minichromosome maintenance complex c...	minichromosome maintenance complex c...	0.999
MCM6	MCM4	ENSP00000264156	ENSP00000262105	minichromosome maintenance complex c...	minichromosome maintenance complex c...	0.999
MCM6	MCM2	ENSP00000264156	ENSP00000265056	minichromosome maintenance complex c...	minichromosome maintenance complex c...	0.999
MCM6	MCM5	ENSP00000264156	ENSP00000216122	minichromosome maintenance complex c...	minichromosome maintenance complex c...	0.999
MCM5	MCM3	ENSP00000216122	ENSP00000229854	minichromosome maintenance complex c...	minichromosome maintenance complex c...	0.999
MCM5	MCM7	ENSP00000216122	ENSP00000307288	minichromosome maintenance complex c...	minichromosome maintenance complex c...	0.999

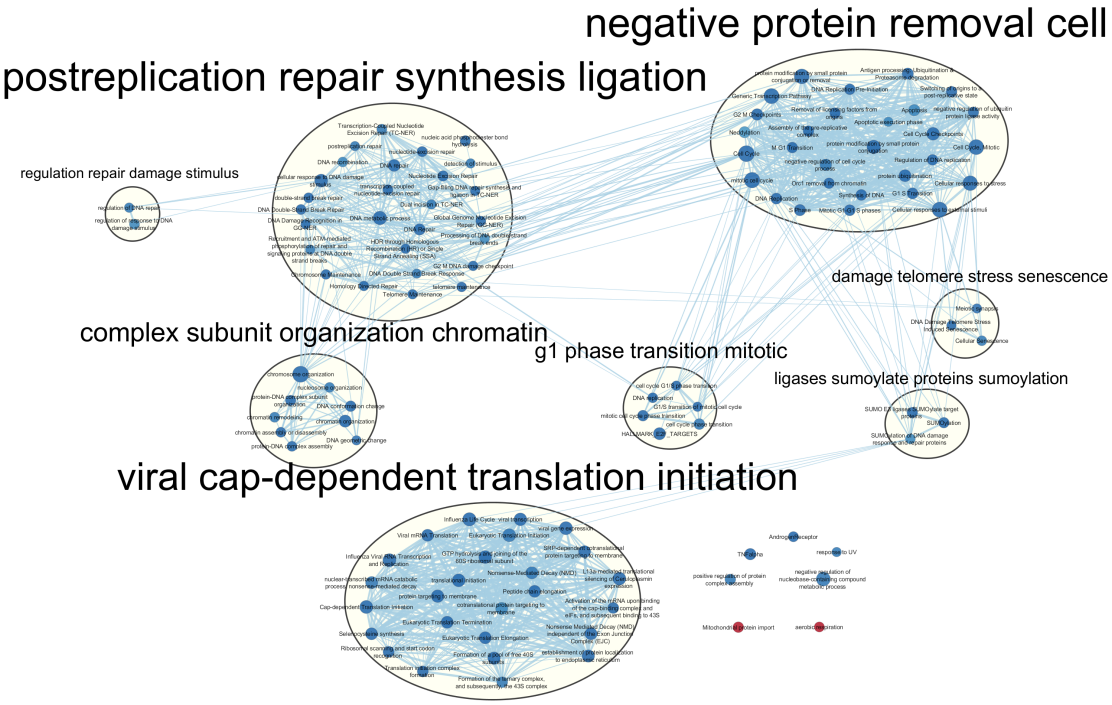
(b) PPI Score for the top 100 DEGs

Figure 7: **Protein-Protein interaction network for the top 100 DEGs**

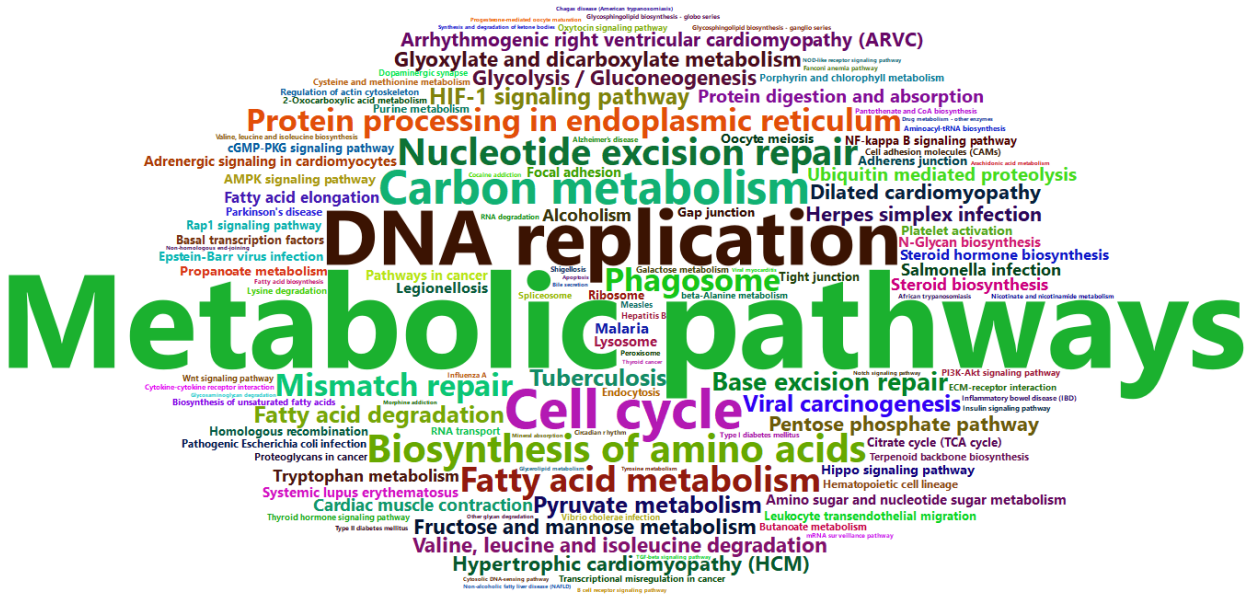
16 genes were not identified by STRING database, of these 1 gene (NEFL, Neurofilament light polypeptide) was differentially expressed (fold change = 2.274, and q-value = 0.00588 with the cut-off FDR = 0.05). Therefore the number of DEGs reduced from 191 to 190 genes. Figure 3 shows PPI network for the top 100 DEGs. The interaction scores displayed in Figure 3b represent combination of probabilities from different evidence channels and then corrected for the probability of randomly observing an interaction. Proteins with top interactions score were the minichromosome maintainance complex proteins (MCM2, MCM3, MCM4, MCM5, MCM6, MCM7), TFRC, TF, RPN1, DDOST, RPA1, PCNA, RFC3, and FEN1.

4.2.4 Enrichment Analyses

Figure 8a shows the result from gene set enrichment analysis performed on the proteomics data. The result was very similar to the transciptomic data. Namely, the key gene sets were cell cycle regulation, DNA replication, S/G1 phase, chromatin related regulation, sumoylation, tumour necrosis factor (TNF) alpha which is involved in the regulation of immune cells, negative regulation of nucleobase-containing compound metabolic process, and cap-dependent translation initiation which affects RNA splicing mechanism. Dysregulation of TNF production has been implicated in human cancer [40]. Figure 4b gives the KEGG pathways that were significant at 5% FDR threshold, and the size of a term represents its significance. The enriched terms agree with the RNA-seq result and Figure 8a.



(a) A gene set network plot where each node represents a gene set, whose colour is blue if the node is down-regulated or red if the node is up-regulated. Naming of each cluster was done based on a word frequency algorithm where the most frequent words and adjacent words inside the cluster would be chosen to a maximum of 4 words. The naming system was done the same for every GSEA network.



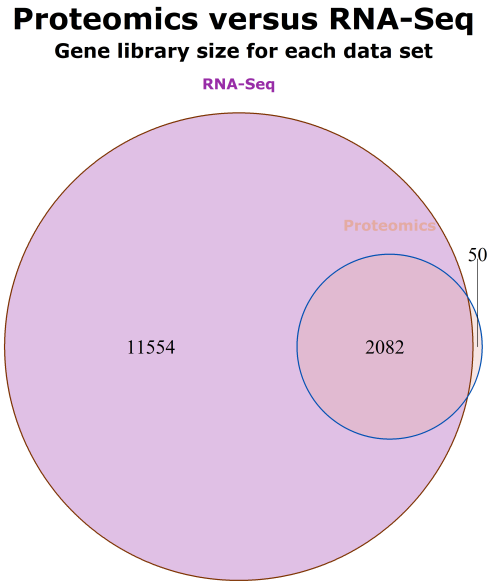
(b) KEGG pathways enrichment

Figure 8: **GSEA and KEGG pathways enrichment for proteomic data**

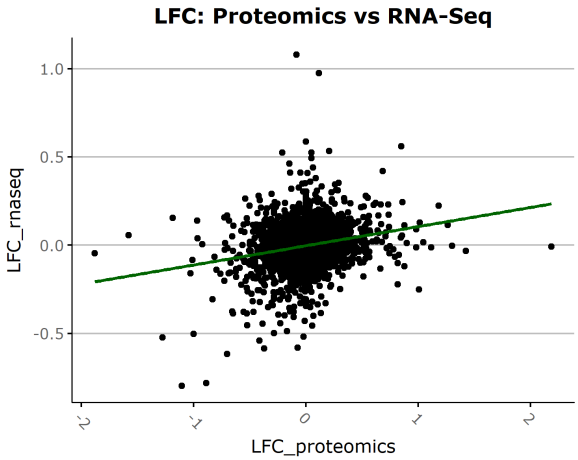
4.3 Integrative Analysis

4.3.1 Differentially Expressed Genes

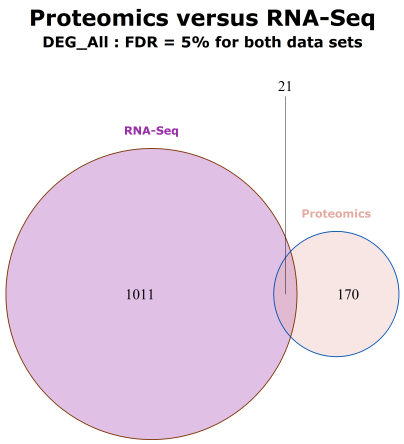
Figure 9a shows that 2,082 genes were common in both RNA-seq and proteomic data.



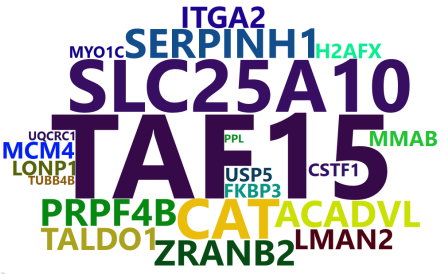
(a) Library sizes comparison between transcriptome and proteome



(b) Correlation of LFC between transciptome and proteome for the 2,082 genes in common (r = 0.2124993 and p-value < 2.2e-16)



(c) Venn diagram showing all overlapping differentially expressed genes



(d) Word cloud for all 21 overlapping differentially expressed genes

(e) Word cloud for 12 down-regulated genes

(f) Word cloud for 6 up-regulated genes

Gene.names	Proteins	Protein.names	LFC_proteomics	LFC_rnaseq	qvalue_proteomics	qvalue_rnaseq
TAF15	Q92804-2;Q92804	TATA-binding protein-associated factor 2N	-1.104656677	-0.797039796	6.74E-12	5.90E-20
SLC25A10	Q9UBX3;Q9UBX3-2	Mitochondrial dicarboxylate carrier	-0.888540111	-0.781487863	1.60E-07	9.60E-13
CAT	P04040	Catalase	-0.703986621	-0.61646429	2.11E-11	0.000170863
SERPINH1	P50454	Serpin H1	0.621461098	0.240163983	2.50E-09	0.015056874
PRPF4B	Q13523	Serine/threonine-protein kinase PRP4 homolog	0.848657825	0.561031762	5.61E-05	3.47E-06
ACADVL	P49748-2;P49748	Very long-chain specific acyl-CoA dehydrogenase, mitochondrial	-0.650714646	-0.386885911	1.61E-08	0.022125318
ZRANB2	O95218-2;O95218	Zinc finger Ran-binding domain-containing protein 2	-1.279317349	-0.523351643	0.000532119	1.62E-06
TALDO1	P37837	Transaldolase	-0.473343841	-0.317240042	6.45E-07	0.015482241
ITGA2	P17301	Integrin alpha-2	0.683014775	0.4189277	0.000655617	3.82E-05
LMAN2	Q12907	Vesicular integral-membrane protein VIP36	0.531582245	0.278358766	2.69E-05	0.0021432
MCM4	P33991	DNA replication licensing factor MCM4	-0.553964463	-0.37874751	2.69E-05	0.039877056
MMAB	Q96EY8	Cob(I)yrinic acid a,c-diamide adenosyltransferase, mitochondrial	-0.832555289	-0.306938073	0.000193495	0.018774191
LONP1	P36776	Lon protease homolog, mitochondrial	0.489698919	0.215439885	0.000114076	0.039877056
H2AFX	P16104	Histone H2AX	-0.999649734	-0.502901741	0.00222161	0.002164979
USP5	P45974-2;P45974	Ubiquitin carboxyl-terminal hydrolase 5	0.283944277	0.311432039	0.018245112	0.000591808
FKBP3	Q00688	Peptidyl-prolyl cis-trans isomerase FKBP3	-0.657387253	-0.376570122	0.011746563	0.000929771
CSTF1	**Q05048**	**Cleavage stimulation factor subunit 1**	**-0.422283253**	**0.279779556**	**0.001451467**	**0.018688299**
MYO1C	O00159-3;O00159	Unconventional myosin-Ic	-0.41456769	-0.255609104	0.005535365	0.015095447
UQCRC1	**P31930**	**Cytochrome b-c1 complex subunit 1, mitochondrial**	**0.224799507**	**-0.226335971**	**0.025518314**	**0.027591862**
TUBB4B	P68371	Tubulin beta-4B chain	-0.518979285	-0.261627469	0.032544841	0.022158316
PPL	**O60437**	**Periplakin**	**-0.540407885**	**0.263682957**	**0.045004324**	**0.027372572**

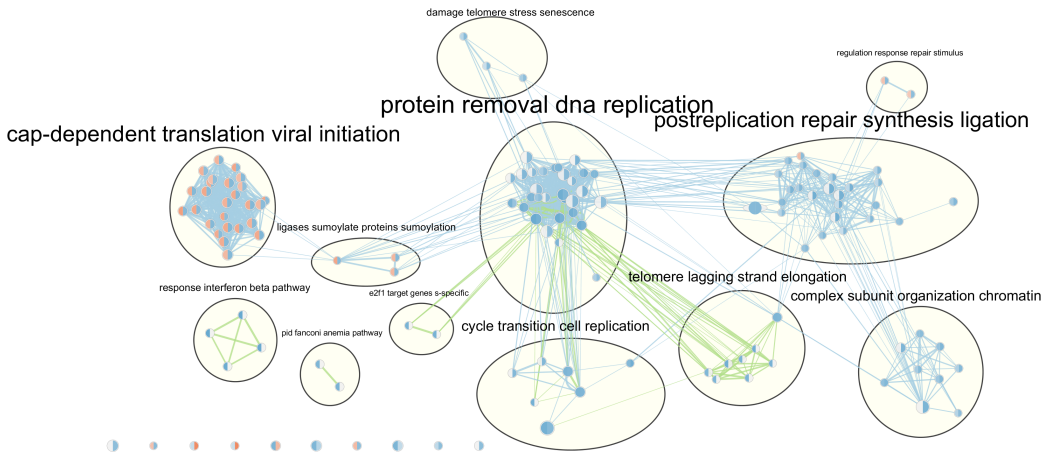
(g) Summary of the 21 overlapping differentially expressed genes

Figure 9: **Differentially expressed genes comparison between RNA-seq and proteomics**

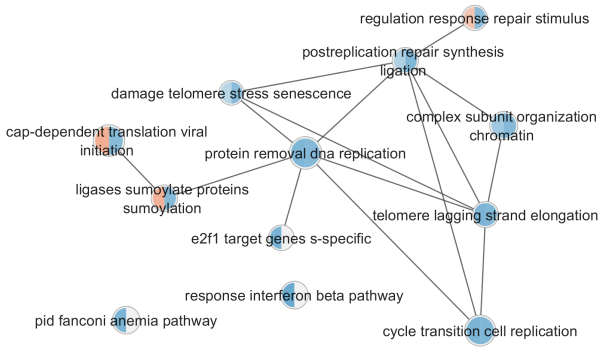
A direct correlation of transciptomic expression and generated protein abundance was always expected based on the central dogma. Nonetheless, many studies have shown that the correlation can be minimal because of post transciption machinery and half lives differences [41–45]. Therefore, in this study, it was expected that low correlation (r = 0.212), and few differentially expressed genes would be detected as common in both transciptomic and proteomic profiles (21 genes), which are shown in Figures 9b and 9c. The 21 overlapping differentially expressed genes are shown in Figure 9d, where 12 of them were both detected as down-regulated and 6 as up-regulated (Figures 9e and 9f). The sign of LFC (positive or negative) of three genes UQCRC1, PPL, and CSTF1 were not the same, as shown in Figure 9g. Genes involved in DNA replication, cell cycle, histone related, and immune system regulation were detected. The enrichment analyses in Section 4.3.2 investigate the omic data on pathways level.

4.3.2 Enrichment Analyses

Figure 10 shows the result from GSEA on both the transcriptomic and proteomic datasets visually (a and b), and the actual numerical result (10c). As aforementioned in Section 3.4.1, GSEA takes into account all genes and their ranks based on the sign of LFC (positive or negative) and the adjusted p-value. The network only shows gene sets that pass the thresholds of 5% FDR for nodes, and similarity constant of 0.375 for edges. Each node in the compact network form (Figure 10b) show the average NES of all gene sets inside the cluster corresponding from Figure 10a. The information for clusters and each singleton (a node that does not belong to any cluster) is displayed in Figure 10c.



(a) Gene set enrichment analysis. Each node contains two sides representing RNA-seq (left side) and proteomics (right side). Blue and red spectra indicate down and up-regulated gene set respectively, and white indicates a missing value. Nodes that do not have clusters are called singletons. Blue and green edges indicate similarity coefficient between nodes for proteomics and RNA-seq respectively.

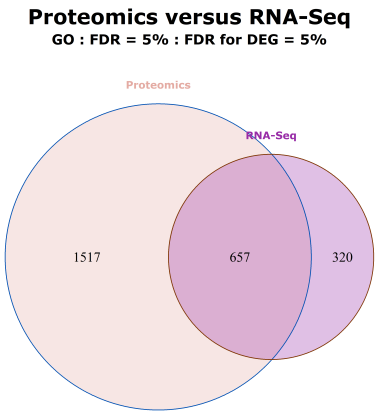


(b) Gene set enrichment network in cluster compact form (from Fig. 10a) without singletons

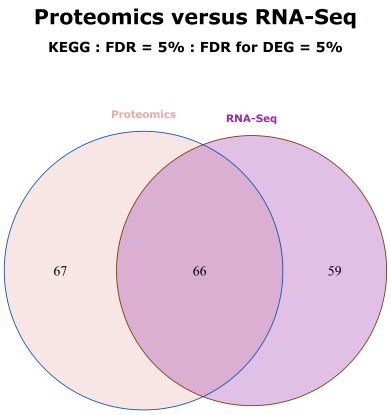
ClusterNumber	ClusterLabel	Nodes	Genes	Average_NES_rnaseq	Average_NES_proteomics
1	protein removal dna replication	31	555	-1.850088889	-1.80016
2	cap-dependent translation viral initiation	26	186	1.253546154	-1.7447
3	postreplication repair synthesis ligation	26	169	-1.111859091	-1.825746154
4	complex subunit organization chromatin	9	174	-1.3405	-1.7887
5	telomere lagging strand elongation	8	107	-1.952375	-1.8526
6	cycle transition cell replication	7	318	-1.93342	-1.863166667
7	response interferon beta pathway	4	55	-2.221725	
8	damage telomere stress senescence	3	33	-0.961233333	-1.702833333
9	ligases sumoylate proteins sumoylation	3	41	1.198333333	-1.758566667
10	regulation response repair stimulus	2	30	0.8561	-1.6856
11	e2f1 target genes s-specific	2	27	-1.9703	
12	pid fanconi anemia pathway	2	47	-1.9225	
Singleton 1	negative regulation of nucleobase-containing compound metabolic process	1	122		-1.6478
Singleton 2	response to UV	1	16	0.9317	-1.7088
Singleton 3	Mitochondrial protein import	1	44	-0.7497	1.8616
Singleton 4	aerobic respiration	1	29	-0.7043	1.8916
Singleton 5	monocarboxylic acid catabolic process	1	73	-1.9024	1.065
Singleton 6	monosaccharide metabolic process	1	102	-1.875	-1.0211
Singleton 7	positive regulation of protein complex assembly	1	44	1.0592	-1.6392
Singleton 8	cellular modified amino acid metabolic process	1	99	-1.9016	-0.8343
Singleton 9	AndrogenReceptor	1	33	-0.8549	-1.6496
Singleton 10	TNFalpha	1	63		-1.6538

(c) Gene set enrichment analysis summary. A missing value indicates that all gene sets inside a cluster are not present (for a transcriptomic or proteomic domain).

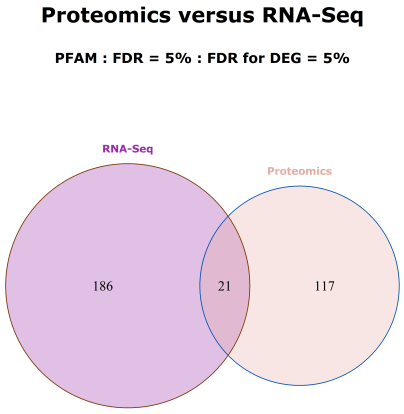
Figure 10: **Integrative gene set enrichment analysis**



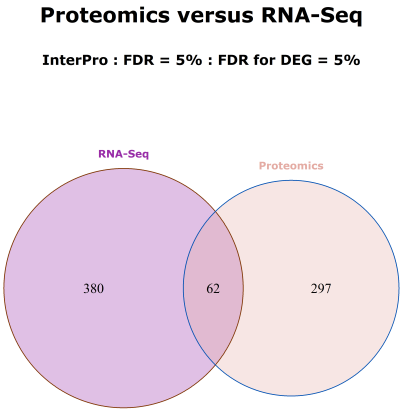
(a) Venn diagram on GO terms enrichment



(b) Venn diagram on KEGG pathways enrichment

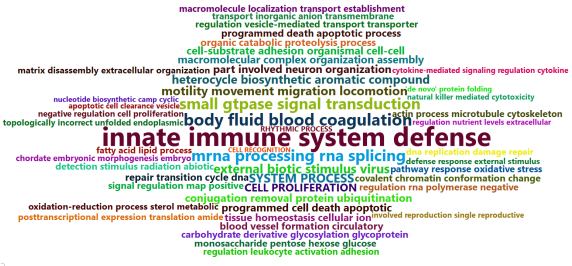


(c) Venn diagram on Pfam enrichment

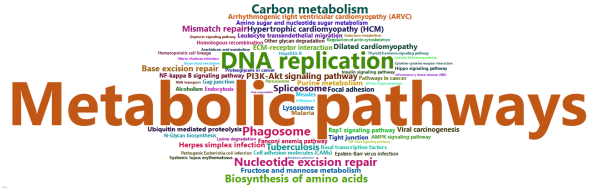


(d) Venn diagram on InterPro enrichment

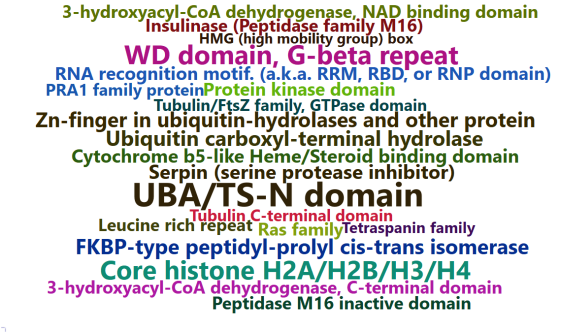
Figure 11: **Overlapping pathways and protein domains enrichments**



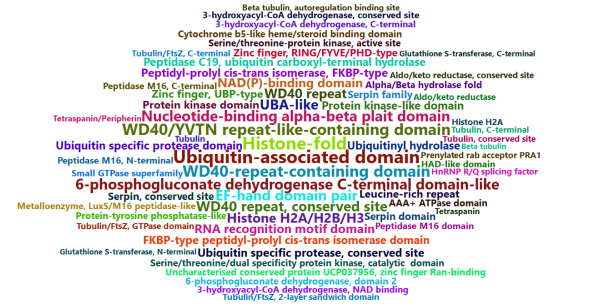
(a) GO terms enrichment with MCL



(b) KEGG pathways enrichment



(c) Pfam terms enrichment with MCL



(d) InterPro terms enrichment with MCL

Figure 12: **GO, KEGG pathways, Pfam and InterPro terms enrichment analyses**

Figure 11 shows the common pathways and protein families or domains compared between transcriptomic and proteomic datasets, with two thresholds: 5% FDR on the list of differentially expressed genes and 5% FDR on the enrichment list. Since the transcriptomic and proteomic domains are different, no merging was done on the reference set. It was observed that 657 GO terms, 66 KEGG pathways, 21 Pfam terms, and 62 InterPro terms were overlapped between the two domains. The common functional contexts were shown as word clouds in Figure 12. Agreement was observed for all analyses on the enrichments of metabolic pathways, DNA replication, spliceosomal machinery, cell cycle, and immune system regulation. Importantly, WD40 domains showed up in both Pfam and InterPro enrichments (Figures 12c and 12d) which confirmed the effectiveness of the PRMT5 inhibitor (GSK591).

There are multiple integrative approaches on transcriptomic and proteomic data, which can be categorised into 8 different methods [46]. They are: 1. Union of transcriptomic and proteomic data, 2. Extraction of common functional context of transcriptomic and proteomic features, 3. Topological networks approach, 4. Merging datasets in individual domains, 5. Missing value estimation by non-linear optimisation, 6. Multiple regression analysis

to predict contribution of sequence features in mRNA-protein correlation, 7. Clustering approaches, and 8. Dynamic modeling. In this study, types 2 were used on GO, KEGG pathways, Pfam and InterPro enrichment analyses, and types 2, 3, and 7 were used on GSEA. All of the approaches used have a common flaw which is the inability of capturing the dynamic interactions between transcriptomic and proteomic domains. They can only utilise the end-point data or single time snapshot within two different universes of transcriptome and proteome, and find the common features. This is one of the reasons that leads to low correlation observed between two domains. To gain a better insight into the complex biological system through the interactions of the transcriptome and proteome, time series data of mRNAs and proteins (for each cell) as well as the application of dynamic modeling integrative method (type 8) are desired.

5 Conclusion

Comprehensive datasets of transcript expression and protein abundance changes of 3 different GBM cell lines treated either with a PRMT5 inhibitor (GSK591) or its negative control (SGC2096) were conducted. It was demonstrated that there was a low correlation in terms of abundance levels alone (i.e. single-gene analysis), but many agreements were observed on the functional pathways and domains levels (i.e GSEA, GO, KEGG pathways, Pfam, and InterPro enrichment analyses). Specifically, metabolic pathways, immune system regulation, spliceosomal machinery, and cell cycle were the significantly enriched biological processes in the GSK591 treated samples compared to the GSC2096 treated ones. Thus the results confirmed our proposed hypothesis. However, it is still a challenge to develop a uniform and efficient pipeline for the integrative analysis. For future direction, it is proposed that a time series data and dynamic modelling should be used to capture the dynamic interactions of transciptomic and proteomic domains. qPCR should be conducted to confirm the omic results on single-gene analysis. Splicing mechanism can be further studied by transcriptomic analysis on the exon level. Investigation of metabolomic data should also be performed to better understand the metabolite network, which has an impact on the protein network level. Thus new potential drug targets may be discorvered.

References

1. Dirks, P. Brain tumor stem cells: The cancer stem cell hypothesis writ large. *Molecular Oncology* **4,** 420–430 (2010).
2. American Brain Tumor Association. *Glioblastoma (GBM)* http://www.abta.org/brain-tumor-information/types-of-tumors/glioblastoma.html (2017).
3. Frye, S. The art of the chemical probe. *Nat. Chem. Biol.* **6,** 159–161 (2010).
4. Horvath, P. Screening out irrelevant cell-based models of disease. *Nature Reviews Drug Discovery* **15,** 751–769 (2016).
5. Wu, B., Li, L., Huang, Y., Ma, J. & Min, J. Readers, writers and erasers of N(6)-methylated adenosine modification. *Curr. Opin. Struct. Biol.* **47,** 67–76 (2017).
6. Zhang, W., Sartori, M., Makhnevych, T. & et al. Generation and validation of intracellular ubiquitin variant inhibitors for USP7 and USP10. *J. Mol. Biol.* doi:10.1016/j.jmb.2017.05.025 (2017).
7. Denny, R., Flick, A., Coe, J. & et al. Structure-Based Design of Highly Selective Inhibitors of the CREB Binding Protein Bromodomain. *Journal of Medicinal Chemistry.* doi:10.1021/acs.jmedchem.6b01839 (2017).
8. Structural Genomics Consortium. *Epigenetics Probes Collection* http://www.thesgc.org/chemical-probes/epigenetics (2017).
9. Bannister, A. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21,** 381–395 (2011).
10. Bedford, M. & Clarke, S. Protein Arginine Methylation in Mammals: Who, What, and Why. *Mol. Cell* **33,** 1–13 (2009).
11. Wei, H., Mundade, R., Lange, K. & et al. Protein arginine methylation of non-histone proteins and its role in diseases. *Cell Cycle* **13,** 32–41 (2014).
12. Kaniskan, H., Konze, K. & Jin, J. Selective Inhibitors of Protein Methyltransferases. *J. Med. Chem.* **58,** 1596–1629 (2015).
13. Yang, Y. & Bedford, M. Protein arginine methyltransferases and cancer. *Nat. Rev. Cancer* **13,** 37–50 (2013).
14. Chung, J., Karkhanis, V., Tae, S. & et al. Protein arginine methyltransferase 5 (PRMT5) inhibition induces lymphoma cell death through reactivation of the retinoblastoma tumor suppressor pathway and polycomb repressor complex 2 (PRC2) silencing. *J. Biol. Chem.* **288,** 35534–35547 (2013).
15. Pal, S., Baiocchi, R., Byrd, J. & et al. Low levels of miR-92b/96 induce PRMT5 translation and H3R8/H4R3 methylation in mantle cell lymphoma. *EMBO J.* **26,** 3558–3569 (2007).
16. Chan-Penebre, E., Kuplast, K., Majer, C. & et al. A selective inhibitor of PRMT with in vivo and in vitro potency in MCL models. *Nature Chem. Biol.* **11,** 432–441 (2015).

17. Duncan, K., Rioux, N., Boriack-Sjodin, P. & et al. Structure and Property Guided Design in the Identification of PRMT5 Tool Compound EPZ015666. *ACS Med. Chem. Lett.* doi:10.1021/acsmedchemlett.5b00380.

18. Andrew, S. *Andrew S: FASTQC. A quality control tool for high throughput sequence data.* http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/ (2017).

19. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

20. Li, H. *et al.* The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

21. Simons, A. *HTSeq: Analysing high-throughput sequencing data with Python.* http://www-huber.embl.de/users/anders/HTSeq/ (2017).

22. Robinson, M., McCarthy, D. & Smyth, G. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).

23. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372 (2008).

24. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2017). https://www.R-project.org/.

25. Franceschini, A. & et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research (Database issue)* **41** (2013).

26. Ludger, J., Kris, G. & Lieven, C. Peptide-level Robust Ridge Regression Improves Estimation, Sensitivity, and Specificity in Datadependent Quantitative Label-free Shotgun Proteomics. *Molecular & Cellular Proteomics* **15,** 657–668 (2016).

27. Ludger, J., Andrea, A., Lennart, M. & Lieven, C. Summarization vs Peptide-Based Models in Label-Free Quantitative Proteomics: Performance, Pitfalls, and Data Analysis Guidelines. *J. Proteome Res.* **14,** 2457–2465 (2015).

28. Dowle, A., Wilson, J. & Thomas, J. Comparing the Diagnostic Classification Accuracy of iTRAQ, Peak-Area, Spectral-Counting, and emPAI Methods for Relative Quantification in Expression Proteomics. *J. Proteome Res.* **15,** 3550–3562 (2016).

29. Ahrné, E., Glatter, T., Viganò, C. & et al. Evaluation and Improvement of Quantification Accuracy in Isobaric Mass Tag-Based Protein Quantification Experiments. *J. Proteome Res.* **15,** 2537–2547 (2016).

30. Ning, Z., Zhang, X., Mayne, J. & Figeys, D. Peptide-Centric Approaches Provide an Alternative Perspective To Re-Examine Quantitative Proteomic Data. *Analytical Chemistry* **88,** 1973–1978 (2016).

31. Suomi, T., Corthals, G., Nevalainen, O. & Elo, L. Using Peptide-Level Proteomics Data for Detecting Differentially Expressed Proteins. *J. Proteome Res.* **14,** 4564–4570 (2015).

32. Fortunel, N., Otu, H., Ng, H., Chen, J. & et al. Comment on " 'Stemness': Transcriptional Profiling of Embryonic and Adult Stem Cells" and "A Stem Cell Molecular Signature" (II). *Science* **302,** 393 (2003).

33. Subramaniana, A., Tamayoa, P. & et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **102,** 15545–15550 (2005).

34. Hollander, M. & Wolfe, D. *Nonparametric Statistical Methods* (Wiley, New York, 1999).

35. Shannon, P., Markiel, A. & et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).

36. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *PLoS One* **5(11): e13984** (2010).

37. Morris, J. *et al.* clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics* **12,** 436 (2011).

38. Van Dongen, S. *Graph Clustering by Flow Simulation* PhD thesis (University of Utrecht, 2000).

39. Bezzi, M., Teo, S., Muller, J. & et al. Regulation of constitutive and alternative splicing by PRMT5 reveals a role for Mdm4 pre-mRNA in sensing defects in the spliceosomal machinery. *Genes & Dev.* **27,** 1903–1916 (2013).

40. Locksley, R., Killeen, N. & Lenardo, M. The TNF and TNF receptor superfamilies: integrating mammalian biology. *Cell* **104,** 487–501 (2001).

41. Chen, G., Gharib, T., Huang, C. & et al. Discordant protein and mrna expression in lung adenocarcinomas. *Molecular and Cellular Proteomics* **1,** 304–313 (2002).

42. Pascal, L., True, L., Campbell, D. & et al. Correlation of mrna and protein levels: Cell type-specific gene expression of cluster designation antigens in the prostate. *BMC Genomics* **9,** 246 (2008).

43. Gygi, S., Rochon, Y., Franza, B. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell Biol.* **19,** 1720–1730 (1999).

44. Yeung, E. Genome-wide correlation between mrna and protein in a single cell. *Angewandte Chemie International Edition* **50,** 583–585 (2011).

45. Ghazalpour, A., Bennett, B., Petyuk, V. & et al. Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet* **7,** e1001393 (2011).

46. Haider, S. & Pal, R. Integrated Analysis of Transcriptomic and Proteomic Data. *Current Genomics* **14,** 91–110 (2013).