University of HUDDERSFIELD

Integrated genome-scale discovery of the SFPQ-DNA and -RNA regulatory interactome and its output in cancer biology

Joseph Alexander Cogan
MRes (Biological Sciences)

University of Huddersfield
School of Biological Sciences

January 2022

Acknowledgements

I would first like to thank Dr James Boyne for his support and advice over the past two years, not only through research supervision but in general mentorship throughout my progression from undergraduate to postgraduate research. I also express my appreciation for Dr Chinedu Anthony Anene who guides the development of my skills in bioinformatics. Dr Anene's patience and dedication throughout my MRes has helped lay a foundation of techniques and good practise that will prove extremely useful towards my progression in biomedical research. Finally, I owe gratitude to my friends and family for their support throughout my studies.

Copyright Statement

Abstract

Splicing factor proline- and glutamine- rich (SFPQ) is a multifunctional DNA- and RNA-binding protein, part of the drosophila behaviour human splicing (DBHS) family of proteins. First discovered as a splicing factor, it has since been revealed that SFPQ is involved in a large range of cellular mechanisms, including transcriptional regulation, DNA damage repair, and paraspeckle formation. Misregulation of SFPQ is associated with aetiology of neurodegenerative diseases and a range of cancers, including prostate, renal, and colorectal cancers, and is portrayed with significant potential as a diagnostic or prognostic biomarker in cancer. Studies regarding the regulatory role of SFPQ, through interaction with other transcripts have propelled in recent years, however the complete scale of the SFPQ regulatory network remains unknown.

In this study, we mapped the SFPQ regulome exhaustively, constructing an interaction database of novel interactors for further study. We utilised publicly available datasets from several experimental techniques at NCBI GEO and ENCODE, including SFPQ-knockdown RNA-seq, RIP-seq, ChIP-seq, PAR-CLIP and eCLIP datasets. We interrogated transcripts with altered expression levels in multiple cell lines depleted with SFPQ in order to explore the scope of genes regulated by SFPQ function. We next identified RNA physical interactors of SFPQ through integration RIP-seq data targeting SFPQ across multiple cell lines, as well as revealing the binding locations of target of SFPQ upon target transcripts through analysis of eCLIP and PAR-CLIP datasets. To complete the baseline of our interaction database, we integrated ChIP-seq datasets in order to tease out transcripts regulated by SFPQ at the DNA level, and the extent at which SFPQ is involved in transcriptional regulation.

We observed a wide range of biological pathways associated with SFPQ interactors, and compiled results into the mineable database and R package, idbSFPQ. Additionally, our analyses lay template for application of methodology to investigate regulatory networks of other proteins.

1.0 Introduction

Splicing factor proline- and glutamine-Rich (SFPQ), also known as PTB-associated splicing factor (PSF), is a multifunctional DNA- and RNA-binding protein and is part of the Drosophila behaviour/ human splicing (DBHS) family of proteins. As the alias implies, SFPQ was first discovered as a protein involved in pre-mRNA splicing via interactions with polypyrimidine tract-binding protein (PTB) (Patton et al., 1993). SFPQ has since been revealed as a protein with great complexity, evidenced with importance in regulation of a large range of cellular mechanisms, including transcriptional regulation, DNA damage repair, and paraspeckle formation (Bladen et al., 2005; de Silva et al., 2019; Emili et al., 2002; Rosonina et al., 2005). Existing evidence surrounding the role of SFPQ in regulation of cellular functions, as well as the potential implications of misregulation will be reviewed in detail below.

1.1 RNA-binding proteins (RBPs)

RNA-binding proteins (RBPs) are a largely diverse class of proteins that regulate gene expression by interacting with target RNA substrates to form ribonucleoprotein complexes. RBPs are involved in multiple cellular processes regarding regulation of gene expression, including RNA translation, localisation and stability, as well as mediating pre-mRNA splicing, cleavage and polyadenylation (Osera et al., 2015; Weidensdorfer et al., 2009; Yoon et al., 2014; Zhou et al., 2019). The regulatory roles of RBPs are vital in human physiology and defective RBPs are associated with the development of a broad range of diseases, including cancer, neurodegeneration (de Silva et al., 2019; Giordana et al., 2010; Neumann et al., 2009; Peng et al., 2018; Sakurai et al., 2014).

As part of the Encyclopedia of DNA elements (ENCODE) project phase III, datasets containing large-scale systematic mapping of 356 human RBPs, including SFPQ, were made publicly available (Van Nostrand et al., 2020). Datasets consist of RBP analysis in K562 and HepG2 cells utilising five assay types, including enhanced Crosslinking and immunoprecipitation (eCLIP), RBP-knockdown (kd)

RNA-seq, chromatin immunoprecipitation and sequencing (ChIP-seq), RNA bind-N-seq, and immunofluorescence assays (Van Nostrand et al., 2020).

1.2 Drosophila behaviour/ human splicing (DBHS)

The DBHS family of proteins are a group of complex proteins defined by an evolutionarily conserved DBHS region which consists of three heavily conserved motifs, namely N-terminal RNA recognition motifs (RMM1 and RMM2), a NonA/ paraspeckle domain (NOPS) and a C-terminal coiled-coil (Knott et al., 2016; Y. Lee & Rio, 2015).

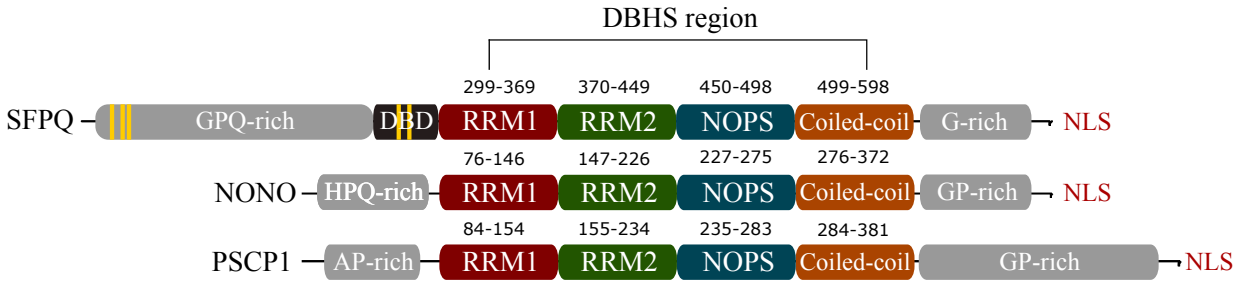


Figure 1: Schematic diagram representing the sequence structure of the three human DBHS proteins: SFPQ, NONO, and PSPC1. DBHS region consists of the RNA recognition motifs (RRMs) RRM1 (red) and RRM2 (green), NonA/ paraspeckle domain (NOPS) (blue) and coiled-coil (orange). Each domain in the DBHS interface is labelled with amino acid position. Schematic includes SFPQ DNA binding domain (DBD) (black), low-complexity regions (black), and C-terminal nuclear localisation signals (NLS). Yellow lines represent RGG motifs and within the schematic. Schematic recreated from Knott, Bond and Fox, 2016.

Within the DBHS region, the human DBHS proteins share more than 70% pairwise sequence identity (Passon et al., 2012). Alongside SFPQ, there are two other members of the DBHS family in humans, non-POU domain-containing octamer-binding protein (NONO), and paraspeckle protein component 1 (PSPC1) and aside from the previously described defining characteristics, these proteins differ significantly. Localisation of DBHS proteins is contextually dependant, although they are frequently deemed as nuclear factors due to their C-terminal nuclear localisation signal (Figure 1) (Shav-Tal & Zipori, 2002). DBHS proteins can also be found within paraspeckles, localised to chromatin, or DNA

damage foci (Bajc Česnik et al., 2019; Major et al., 2019). DBHS proteins are also evidenced to re-localise from nucleoplasm to the cell surface (Ren et al., 2014).

1.3 Interactions between DBHS proteins

Critical to the function of DBHS proteins, dimers are formed through both homodimeric and heterodimeric composition, resulting in the potential for six possible dimer combinations (Passon et al., 2012). DBHS obligate dimerisation occurs through reciprocal interactions between RRM2, NOPS domain and part of the coiled-coil (Knott et al., 2015; M. Lee et al., 2015). Although dimerisation involves interaction across the entire DBHS region, it mainly consists of hydrophobic interactions between the RRM2 and the NOPS domain (Knott et al., 2015). In the absence of RRM1, SFPQ is still able to form an obligate dimer, mutations of residues within RRM2-NOPS domains hindered localisation and function of NONO/PSPC1 dimerisation (M. Lee et al., 2015; Passon et al., 2012). Dimerisation of DBHS proteins provides structural integrity but importantly, it is critical for DBHS function. For instance, DBHS proteins are involved in the formation and localisation of subnuclear paraspeckles and DBHS proteins unable to dimerise, due to mutations at key hydrophobic residues, were unable to form paraspeckles (Passon et al., 2012).

1.4 Structure of SFPQ

DBHS proteins show low levels of similarity in protein architecture aside from the shared DBHS region (Figure 1). SFPQ has two major isoforms with the major isoform (SFPQ-A) consisting of 707 amino acids, and the shorter spliced isoform (SFPQ-F) consisting of 669 amino acids. SFPQ-F differs from SFPQ-A after amino acid 662, with an additional seven amino acids and lack of a C-terminal nuclear localisation signal (Patton et al., 1993).

Oligomerisation of SFPQ dimers is also reported to be critical to its function, the SFPQ C-terminal coiled-coil projects out from the dimer core, and promotes oligomerisation with the coiled-coil of neighbouring dimers (M. Lee et al., 2015). X-ray scattering analysis of SFPQ revealed that SFPQ oligomerisation is dynamic and reversible, additionally to being critical for protein function (M. Lee et al., 2015). Through truncation of the C-terminal coiled-coil, prevention of SFPQ oligomerisation resulted in reduced function in DNA-binding and transcriptional regulation, with implications to the potential involvement of deficient oligomerisation in the misregulation of SFPQ (Huang et al., 2020; M. Lee et al., 2015; Thomas-Jinu et al., 2017).

The specific boundary of the SFPQ DNA-binding domain (DBD) remains uncategorised, however it is reported that the sequence N-terminal to RRM1 is required for DNA binding (Figure 1) (Ha et al., 2011; M. Lee et al., 2015)

1.5 SFPQ is a multifunctional RBP

SFPQ is often described as a "multifunctional" RBP due to its great spectrum of interaction partners deeming involvement in a large range of processes (Knott et al., 2016). We discuss some of the SFPQ-associated functions including paraspeckle formation, RNA processing, transcriptional regulation, alternative splicing, DNA damage repair and apoptosis.

1.5.1 Paraspeckle formation

Paraspeckles are a distinct class of subnuclear bodies, ranging from 0.5 – 1 μm in diameter (Cardinale et al., 2007). These organelles are located within the interchromatin space, between chromatin and larger nuclear speckles. Paraspeckles are formed within the cell nucleus when specific proteins bind to the nascent long non-coding RNA (lncRNA) transcript, Nuclear paraspeckle assembly transcript 1 (NEAT1) (Clemson et al., 2009). Paraspeckle functions, as well as the mechanisms underlying them, are not completely annotated, however they are reported to be involved in gene regulatory processes,

including mRNA retention and cleavage, and protein sequestration (L.-L. Chen & Carmichael, 2009; Hirose et al., 2014). In formation of paraspeckles, SFPQ and its paralogue, NONO, bind NEAT1 to form a minimal ribonucleoprotein (Clemson et al., 2009). Additionally, siRNA knockout of SFPQ in HeLa cells resulted in loss of paraspeckles, suggesting its critical role in the formation and structural integrity of paraspeckles (Sasaki et al., 2009).

1.5.2 RNA processing

As previously mentioned, SFPQ is both nuclear and cytoplasmic localised in a contextually dependent manner. SFPQ is reported with involvement in translation, mediated through internal ribosome entry sites (IRESs) in the cytoplasm (King et al., 2014; Sharathchandra et al., 2012). *In vitro* assays demonstrated that SFPQ can bind directly to IRES element on the p53 gene (Sharathchandra et al., 2012). Additionally, knockdown of SFPQ in H1299 cells resulted in downregulation of IRES- and non-IRES-dependent expression of p53, it is unclear whether SFPQ binds IRES elements in cells but an indirect effect is suggested (Sharathchandra et al., 2012).

SFPQ is involved in biogenesis of mRNA through 3' end processing, reported through identification of SFPQ as a component of the snRNP-free U1A (SF-A) complex via sucrose gradient fractionation and immunoprecipitation (IP) of HeLa cells (O'Connor et al., 1997). The SF-A complex comprises of other splicing factors, including NONO, and immunodepletion experiments affected cleavage and polyadenylation (Hall-Pogar et al., 2007; Liang & Lutz, 2006).

1.5.3 Transcriptional regulation

SFPQ is involved in transcriptional regulation through acting as a corepressor and coactivator. SFPQ/NONO are reported to directly bind RNA Polymerase II, with implications to involvement in transcription initiation, as well as RNA Polymerase II-dependent transcriptional elongation (Emili et al., 2002). Through binding RNA Polymerase II, SFPQ/NONO provide integrity for transcriptional machinery, via acting as a bridge between RNA Polymerase II and promoter-bound transcriptional

activators, coactivating transcription (Emili et al., 2002). One example of SFPQ involvement in positive transcriptional regulation is through binding to the promoter of Phosphodiesterase 3A (PDE3A) gene, and is critical for PDE3A expression (Rhee et al., 2017). SFPQ is also evidenced to bind to a proposed enhancer sequence, which positively regulates expression of ribosomal protein genes, including 60S ribosomal protein L18 (RPL18) (Roepcke et al., 2011).

Inversely, SFPQ is also known to act as a repressor of transcriptional events. SFPQ acts as a transcriptional corepressor when interacting with members of the nuclear hormone receptor (NHR) family. Transcriptional regulation of downstream genes through via NHRs is dependent upon the binding of a specific ligand to the NHR. Interactions between SFPQ and NHRs result in recruitment of the corepressor, SIN3 transcription regulator family member A (SIN3A). Interactions with SIN3A aid in the assembly of histone deacetylases (HDACs), inducing repressive chromatin marks and transcriptional repression of target genes (Mathur et al., 2001). It is notable that transcriptional repression of androgen receptor (AR) is only regulated by SFPQ-A isoform, and not SFPQ-F, suggesting that functional dissimilarities are caused by the lack of nuclear localisation signal (NLS) on SFPQ-F (Dong et al., 2007).

1.5.4 Alternative splicing

SFPQ was initially characterised for its role in early formation of the spliceosome, through association with PTB (Patton et al., 1993). Following its discovery, immunodepletion of nuclear extract with antibodies targeting SFPQ suppressed spliceosome assembly in the initial stages (Patton et al., 1993). The modern understanding of SFPQ involvement in spliceosome assembly is that it can regulate the process in a substrate-dependent manner, through control of alternative splicing. Although, despite being the initial characterising function, there remains a gap in the understanding of the exact mechanisms encompassed by SFPQ in splicing. Alternative splicing is a process involving the rearrangement of intron and exon elements, allowing messenger RNA (mRNA) to synthesise different protein variants. For instance, SFPQ is involved in alternative splicing of cluster of differentiation 45

(CD45) and Tau genes through RNA interaction (Ray et al., 2011). In alternative splicing of the Tau gene, SFPQ interacts with a stem-loop structure at the exon-intron boundary downstream from exon 10, leading to a lack of exon 10 in the final mRNA (Ray et al., 2011). Upregulation of SFPQ leads to exclusion of exon 10, and consistently, downregulation facilitates inclusion of exon 10 (Ray et al., 2011). SFPQ also promotes exon skipping in CD45 through interactions with pyrimidine-rich regions on exon 4 leading to suppressed inclusion of exon 4 (Heyd & Lynch, 2010; Melton et al., 2007).

SFPQ also influences alternative splicing via promotion of exon inclusion. For example, SFPQ is reported to interact with a purine-rich sequence on exon 7 of survival of motor neuron 2 (SMN2), promoting its inclusion in final mRNA (Cho et al., 2014).

In some cases, alternative splicing also occurs through SFPQ interaction with other splicing factors, such as through interaction with heterogeneous nuclear ribonucleoprotein (hnRNP) M, where overexpression of SFPQ induced greater exon inclusion in a splicing minigene reporter, preprotachykinin (PPT), potentially through interaction with hnRNP (Marko et al., 2010). The hnRNP protein family is associated with nuclear splicing processes, predominantly through interactions with pre-mRNA substrates (Llères et al., 2010). Another interactor of SFPQ in splicing is the splicing factor, Fox-3. SFPQ is potentially a coactivator of Fox-3, which, in the mouse central nervous system, enhances the inclusion of exon N30 in non-muscle myosin heavy chain (NMHC) II-B (Kim et al., 2011). Although these mechanisms are not annotated in entirety, they support the claim that SFPQ influences alternative splicing through interaction with other splicing factors.

1.5.5 DNA damage repair

Another heavily annotated function of SFPQ is its role in DNA damage response. The DNA damage response is a complex network of signaling pathways and acts as a cellular defence system in maintaining genome stability, which is critical in maintenance of cellular homeostasis and is associated with the evolution of cancer (Halazonetis et al., 2008). One of the common initiators of

DNA damage response is double-stranded breaks in DNA, and cellular defence against double-stranded breaks consists of robust pathways such as homologous recombination (HR) and error prone pathways such as non-homologous end joining (NHEJ) (Ferguson et al., 2000).

Many RBPs are associated with the DNA damage response through indirect regulation of genes involved in DNA damage response pathway, such as KH-type splicing regulatory protein (KSRP), which does so through regulation of miRNAs (X. Zhang et al., 2011). Whereas it is reported that SFPQ is directly involved in the DNA damage response (Ha et al., 2011; Rajesh et al., 2011; Salton et al., 2010).

Through SFPQ/NONO interaction with proteins associated with DNA damage response, SFPQ has been implicated in both HR and NHEJ. It was reported that SFPQ forms a preligation complex through interaction with the KU70/KU80 heterodimer (Bladen et al., 2005). Additionally, SFPQ interacts with IGFBP-3, with involvement in IGFBP-3-dependent repair of double-strand breaks (de Silva et al., 2019).

Under conditions inducing DNA damage, SFPQ/NONO heterodimer levels increased rapidly, with implications to their role in the initial stages of the DNA damage response (Ha et al., 2011; Salton et al., 2010). Consistent with this observation, the SFPQ/NONO heterodimer interacts with Matrin 3 (MATR3) (Salton et al., 2010). MATR3 is a target of Ataxia Telangiectasia Mutated (ATM), a nuclear kinase in repair of double-stranded breaks (Salton et al., 2010).

In maintenance of genome stability, SFPQ has been recently reported in regulation at telomeres, maintaining telomere integrity (Petti et al., 2019). In absence of SFPQ/NONO heterodimer, telomere recombination events significantly increased, inducing alterations in telomere length in both telomerase-positive and negative cells (Petti et al., 2019). SFPQ/NONO binds telomere repeat-containing lncRNA (TERRA), and in doing so, prevents formation of RNA-DNA hybrids, defective DNA replication, HR and DNA damage at telomeres (Petti et al., 2019).

1.5.6 Apoptosis

An important role of SFPQ, especially in the prevention of cancer progression, is through association with apoptosis. Apoptosis is a process involving programmed cell death, primarily of cells which are damaged beyond repair. SFPQ is reported with a role in regulation of cell death in cancer cell lines, it is known that SPFQ interacts with peroxisome proliferator-activated receptor γ (PPARγ), which is involved in cell proliferation and apoptosis (Tsukahara, Haniu, et al., 2013). Interestingly, genetic perturbance of SFPQ in PPARγ-expressing DLD1 colon cancer cell lines caused a loss of microtubule-associated proteins, 1A/1B light chain 3B (LC3B) and induced apoptosis (Tsukahara, Haniu, et al., 2013; Tsukahara, Matsuda, et al., 2013). Consistent with SFPQ apoptotic function, depletion of SFPQ in supressed proliferation and induced S-phase arrest and apoptosis in melanoma $BRAF^{V600E}$ mutant colorectal cancer cell lines (Klotz-Noack et al., 2020) During apoptotic processes associated with SFPQ, nuclear levels of SFPQ are significantly reduced when observed via monoclonal antibody staining, suggesting subcellular localisation of SFPQ during apoptosis. (Shav-Tal et al., 2001).

1.6 Role of SFPQ in neuronal function and neurodegenerative diseases

SFPQ function is heavily annotated in association with neuronal maintenance and development, such as axon growth (Thomas-Jinu et al., 2017). Therefore, is it unsurprising that abnormal SFPQ function has implications in neurodegenerative disorders such as frontotemporal lobar degeneration (FLTD), Alzheimer's disease (AD) and amyotrophic lateral sclerosis (ALS) (Ishigaki et al., 2020; Lu et al., 2018; Luisier et al., 2018). The role of SFPQ In neuronal function, as well as their implications in development of neurodegenerative diseases are discussed below.

1.6.1 Maintenance of neuronal function

One way in which SFPQ function is integral to neuronal maintenance is through transcriptional regulation. SFPQ is reported to regulate transcriptional elongation of long genes (>100 kb) in the developing mouse brain through binding to long intron-containing pre-mRNA, and this was confirmed in recent bioinformatics analysis of SFPQ in neuronal cells (Iida et al., 2020; Takeuchi et al., 2018). A proposed mechanism of neuronal transcriptional regulation is through recruitment of cyclin-dependent kinase 9 (CDK9) and activation of RNA polymerase II (Takeuchi et al., 2018).

As mentioned in section 1.6.4, SFPQ is involved in alternate splicing of pre-mRNA (Heyd & Lynch, 2010; Kim et al., 2011). Alternative splicing in neuronal cells is mediated by SFPQ through coactivation and recruitment of neuronal nuclei (NeuN) (Kim et al., 2011). Importantly, SFPQ forms a complex with RBP Fused in sarcoma (FUS) , in regulation of alternative splicing of the *Mapt*, which codes for the protein, tau and is important in maintaining tau isoform ratio (Ishigaki et al., 2017). Depletion of either SFPQ or FUS resulted in accumulation of phosphorylated tau and neuronal loss (Ishigaki et al., 2017).

It is described that cytoplasmic localisation of SFPQ may be involved in maintenance of neurons via transport of RNA to dendrites (Furukawa et al., 2015; Kanai et al., 2004). This is suggested through colocalization of SFPQ/NONO heterodimer with RBP, HERMES, and G3BP1 in neuronal differentiation within retinal ganglion cells, which formed cytoplasmic neuronal granules (Furukawa et al., 2015). The cytoplasmic neuronal granules formed are presumed RNA transport granules, which are involved in mRNA dendritic localisation (Furukawa et al., 2015).

1.6.2 Implications of aberrant SFPQ in neurodegenerative diseases

As previously mentioned, Takeuchi *et al*., 2018 downregulated SFPQ in the developing mouse brain and demonstrated that as well is impeding transcriptional elongation, loss of SFPQ promoted neuronal

apoptosis (Iida et al., 2020; Takeuchi et al., 2018). Bioinformatics analysis reported that many of the downregulated genes are associated with axon guidance, neuronal migration and formation of synapses, suggesting that aberrant SFPQ function may repress neuronal development and inhibit survival (Iida et al., 2020).

It is reported that interactions between SFPQ and FUS are critical to neuronal homeostasis through regulation of the isoform ratio of tau, where suppression of SFPQ/FUS interactions were coupled with abnormal ratios of tau within ALS and FTLD-affected neurons (Ishigaki et al., 2017, 2020). Abnormal intron retention of the SFPQ transcript, along with nuclear loss, is recognised as a molecular hallmark of ALS, where irregular intron retention was observed during motor neuron differentiation in induced pluripotent stem-cell (iPSC) lines derived from ALS patients (Luisier et al., 2018). FUS binds to the aberrantly retained intron 9 of the SFPQ transcript, which is the most prominently retained intron significantly associated with ALS mutation (Luisier et al., 2018; Tyzack et al., 2019). Interactions between FUS and SFPQ, and aberrant intron retention of SFPQ is critical to mislocalisation of FUS and SFPQ in ALS cytoplasmic accumulation (Luisier et al., 2018; Tyzack et al., 2019).

1.7 Implications of SFPQ in cancer progression

Although SFPQ is associated in the pathogenesis of several diseases, the primary focus of this study is to explore the regulatory targets of SPFQ and the biological processes involved in the SFPQ regulatory network in several cancer cell lines. Notably, the SFPQ-regulated mechanisms in cancer progression vary in a tissue-specific manner. Below we discuss existing studies regarding the carcinogenic implications of aberrant SFPQ function across several tissue types, namely breast, colorectal, melanoma, and prostate cancer.

1.7.1 Breast cancer

Mitobe *et a*l demonstrated that SFPQ expression is significantly associated with poor prognosis in breast cancer patients (Mitobe et al., 2020). SFPQ is involved in post-transcriptional regulation of oestrogen receptor 1 (ESR1) mRNA expression and nuclear transport (Mitobe et al., 2020). Through interaction with ESR1, SFPQ expression is significantly associated with oestrogen signaling and SFPQ knockdown supressed cell proliferation in oestrogen receptor (ER)-positive MCF7 cells (Mitobe et al., 2020). Oestrogen often plays a large role in proliferation and progression of breast cancer, as well as metastasis, and defective regulation of the oestrogen signaling pathways worsen the progression of ER-positive tumours (C. Zhang et al., 2013).

As previously mentioned, SFPQ/NONO heterodimers are heavily associated with early stage double-strand break repair (Salton et al., 2010). Interactions between SFPQ and IGFBP-3, associated with PARP-dependent double-strand break repair, in triple-negative breast cancer induce chemoresistance (de Silva et al., 2019). Chemoresistance is the ability for cancer cells to evade chemotherapeutic treatment and is major obstacle in cancer treatment (Li & Melton, 2012).

1.7.2 Colorectal cancer

SFPQ regulates expression of Metastasis associated with lung adenocarcinoma transcript 1 (MALAT1), a lncRNA widely documented to be overexpressed in multiple cancers including cervical, melanoma, prostate, and breast (Dai et al., 2019; Luan et al., 2016; Ou et al., 2019; Xia et al., 2018). Decreased expression levels of SFPQ are reported to be essential for the oncogenic role of MALAT1, inducing proliferation and migration in colorectal cancer cells (Ji et al., 2014).

SFPQ has been described as a tumour-suppressor protein, through regulation of the oncogenic activity of PTB, where cell proliferation and migration increase in release of PTB from SFPQ/PTB complexes (Ji et al., 2014). SFPQ is reported as a downstream effector of miRNA, miR-1296, in colorectal cancer

cells (Tao et al., 2018). Despite the tumour suppressive role of miR-1296 in several other cancers via targeting and downregulation of oncogenic genes, miR-1296 is overexpressed in colorectal cancer tissues and high expression levels are associated with colon cancer recurrence, with potential indication of poor prognosis (Bobowicz et al., 2016; Majid et al., 2010; Tao et al., 2018). Tao *et al* demonstrate that miR-1296 is involved in cell proliferation and metastasis in colorectal cancer, through directly targeting SFPQ and downregulating its role in tumour suppression (Tao et al., 2018).

1.7.3 Melanoma

SFPQ has been associated with oncogenesis in melanoma through interactions with lncRNA, Llme23, a transcript with oncogenic properties and exclusively detected in human melanoma cell lines (C.-F. Wu et al., 2013). Specific physical binding of recombinant and native SFPQ to Llme23 was determined *in vitro* and *in vivo* via immunoprecipitation and interestingly, Llme23 was only detectable with inclusion of antibodies targeting SFPQ *in vivo* (C.-F. Wu et al., 2013). Although the specific molecular mechanisms surrounding the effect of Llme23 on melanoma malignancy remain unannotated, Llme23 knockdown in melanoma cell line, YUSAC, resulted in suppression of malignant properties and downregulated expression levels of proto-oncogene, Rab23 (C.-F. Wu et al., 2013).

SFPQ has since been explored as an essential gene for melanoma progression. GapmeR knockdown of SFPQ repressed cancer cell phenotypic activities in melanoma cells compared with primary melanocytes (Bi et al., 2021). Depletion of SFPQ had an adverse effect on melanoma viable cell growth and migratory potential, as well as increased number of cells in late stage apoptosis (Bi et al., 2021). Bi *et al*., 2021 also interrogated potential mechanisms by which SFPQ contributes to cancer phenotypes in melanoma through analysis of lncRNA interaction partners, LINC00511 and LINC01234 (Bi et al., 2021). LINC00511 sponges the miRNA, miR-625-5p, in several cancers and

demonstrates glucose consumption in the context of melanoma (Bi et al., 2021; Z. Chen et al., 2019; Xue & Zhang, 2020).

Additionally, in integration of melanoma patient data, SFPQ was shown to be highly expressed in metastatic melanoma and increased SFPQ expression negatively correlated with melanoma patient survival, suggesting its potential as a prognostic biomarker in melanoma (Bi et al., 2021; Muralidhar et al., 2019).

1.7.4 Prostate cancer

Takayama *et al*., 2017 demonstrated that SFPQ is involved in development and progression of prostate cancer through regulation of the spliceosome (K. Takayama et al., 2017). The androgen signaling pathway is important in in castration-resistant prostate cancer (CRPC) (Q. Wang et al., 2009). SFPQ was observed in high expression levels within multiple prostate cancer cells, including DU145, LNCaP and LTAD (K. Takayama et al., 2017). In androgen receptor (AR)-dependent prostate cancer cells, SFPQ was shown to moderate transcriptional mechanisms upon genes associated with cell cycle, such as p53 and SMAD3, and post-transcriptionally regulate AR (K.-I. Takayama et al., 2013).

Takayama et al., 2017 also report the involvement of the regulatory role of SFPQ in the spliceosome development in the progression of aggressive prostate cancer, with most of the SFPQ-targeted spliceosome genes, including U2AF2, DDX23, CHERP, and HNRNPU, upregulated in metastatic prostate cancer (K.-I. Takayama et al., 2013).

1.7.5 Known interactors of SFPQ in cancer

As previously mentioned, SFPQ has a great spectrum of interaction partners and it is likely that the majority remain unannotated. Here we list several notable interaction partners of SFPQ which are already evidenced in potential association with cancer progression through interactions with SFPQ.

Table 1: A list of several known interactors of SFPQ in association with cancer, noting the tissue context in which interaction has been evidenced.

Interactor	Context	Notes	Reference
NEAT1	Hepatocellular carcinoma, liver cancer, chronic myeloid leukaemia		(Ru et al., 2018; Weidensdorfer et al., 2009; Zeng et al., 2018)
MALAT1	Osteosarcoma, colorectal cancer, lung adenocarcinoma	Downregulation of SFPQ/PTB complex.	(Fang et al., 2015; L. Hu et al., 2018; Ji et al., 2014)
TFE3	Renal cell carcinoma	Upregulation of PI3K/AKT/mTOR pathways	(Damayanti et al., 2018)
DHX9	Hepatocellular carcinoma	Oncogenic splicing of BIN1	(Z. Hu et al., 2020)
BRAF	Colorectal cancer		(Klotz-Noack et al., 2020)
GAPLINC	Colorectal cancer	Targeting of SNAI2	(Yang et al., 2016)
SNHG1	Gastric carcinoma		(S. Wang et al., 2021)

1.8 Integrative bioinformatics analysis (Meta-analysis)

In rise of high throughput sequencing (HTS) technologies over the last 2 decades has seen an astounding increase in the availability of gene expression quantification data through public repositories. For instance, National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) freely stores and distributes high throughput functional genomics data from over 100,000 studies, including microarray and next generation sequencing analysis (Barrett et al., 2013).

This development has led to an undoubtable acceleration in studies utilising already existing publicly available datasets and proves a revolutionary method for computational and biomedical research. For example, Bell et al. (2017) employed meta-analysis upon microarray gene expression data from primary and metastatic breast cancer in order to explore potential cancer biomarkers, and identified COX2 and RRM2 as markers for metastasis in breast cancer tumours (Bell et al., 2017). Similarly, Chen et al. (2014) integrated gene expression data from 13 human lung cancer studies and report PTK7 as a highly and specifically expressed genes in lung adenocarcinomas, with potential as a therapeutic target (R. Chen et al., 2014).

Robust bioinformatics analysis begins with method-based data selection, where implementing systematic sampling of some sort is beneficial in reducing bias within data collection. Public repositories such as, NCBI GEO and ArrayExpress allow for automatic searches using keywords (Athar et al., 2019; Barrett et al., 2013). However, manual filtering is necessary to some extent in determining inter-study heterogeneity in the diversity of the study, where high levels of technical or biological diversity between datasets improves the generalisation of results but has a negative effect on statistical power (Jaksik et al., 2015; Waldron & Riester, 2016).

Many methods of data integration through meta-analysis are proposed, each having contextually dependent advantages. For instance, in cases where comparisons are made between data from the same platform or condition, effects size combination can be applied. Although, this method relies on

effect sizes following normal distribution, where a fixed effect model (FEM) or random effect model (REM) can be applied depending on the homogeneity of the data (Hedges, 1982). In the case of gene expression analysis, effect size refers to the differential expression between two groups (Toro-Domínguez et al., 2021).

Whereas, where comparisons involve datasets generated by different platforms or involve measuring different conditions, alternate methods of meta-analysis must be implemented. One example is through rank combination, which involves ranking sets of genes based on their fold change (FC) value in differential gene expression (DGE) analysis using either rank sum or rank product statistical models, depending on the number of studies included. This method reduces the impact of highly significant individual P-values on the rest of the data, proving useful for combination of heterogenous data or data generated from multiple platforms (Nagy et al., 2019). However, rank combination methods are highly sensitive to diversity of gene-specific variance, which can lead to over-optimistic P-values and lead to false positive (Breitling & Herzyk, 2005).

The other major method of meta-analysis for datasets where effect size combinations are unsuitable, P-value combination can be utilised. P-value combination consists of integrating P-values from multiple individual analyses and collapsing them to produce one P-value per gene. P-value combination is useful when combining heterogeneous analyses (Marot et al., 2009). One caveat in P-value combination is that, in direct comparison of P-values, expression pattern and directionality are lost when genes have alternate expression patterns in different analyses (Siangphoe & Archer, 2017). One way to bypass this caveat is through only categorising P-values as significant if expression values follow complementary FC directions in DEG (Rau et al., 2014).

An important consideration in data integration and significance testing is the tolerance of type I and type II error rates, as this varies depending on the aims of the study. Type I errors, otherwise known as false positives, represent cases where the null hypothesis is rejected incorrectly. The method of calculating the false discovery rate (FDR), proposed by Benjamini & Hochberg (1995), attempts to

control type I errors more closely, by offering control of experiment-wide errors in multiple testing (Benjamini & Hochberg, 1995). However, by reducing the likelihood of type I errors we increase type II errors (false negatives) to varying degrees.

1.9 Aims and objectives

SFPQ has been widely studied regarding its association with disease progression, especially in neurodegenerative diseases and cancers, and is portrayed with significant potential as a diagnostic or prognostic biomarker in cancer. Research regarding the regulatory role of SFPQ in interactions with other transcripts, have propelled in recent years, however the complete scale of the SPFQ regulatory network remains unknown.

In this study, we aim to map the SFPQ regulome exhaustively and identify SFPQ interactions that occur at the RNA and DNA level. Through data-mining analysis of GEO and ENCODE, we take advantage of publicly available datasets containing SFPQ analysis via multiple experimental techniques. Differential gene expression analysis of RNA-seq datasets on cell lines depleted with SFPQ and meta-analysis via P-value combination, will allow us to explore genetic interactors with altered expression levels in loss of SFPQ function. Additionally, doing so across multiple cancer cell lines will reveal genes which are regulated consistently across multiple tissues.

Integration of immunoprecipitation datasets combined with high throughput sequencing, such as RIP-seq and ChIP-seq, reveal which of the SFPQ genetic interactors are directly regulated by SFPQ through physical binding. Combination of enriched genes in RIP-seq data and genetic interactors allow us to identify RNA transcripts which are post-transcriptionally regulated by SFPQ through physical interactions. Incorporation of eCLIP and PAR-CLIP datasets allow us to observe specific SFPQ binding locations on RNA transcripts, mapping the transcript biotype binding favoured by SFPQ interactions. Furthermore, integration of ChIP-seq data will unveil DNA transcripts which are transcriptionally regulated by SFPQ. Integration of these experimental techniques will construct a

comprehensive annotation of genes at multi-omics level, identifying SFPQ interactors regulated at the genetic level, RNA-level through physical binding, and transcriptional regulation through DNA binding.

Gene ontology analysis will allow us to understand the biological processes associated with SFPQ interactors, underpinning molecular mechanisms and functions regulated by SFPQ through genetic networks, as well as post-transcriptional and transcriptional regulation. By understanding the spectrum of biological processes regulated by SFPQ in cancer cell lines, we can dissect the mechanistic roles of SFPQ in cancer development and progression.

Ultimately, the analysis performed throughout this study will construct a novel database, idbSFPQ (interaction database of SFPQ). The database will consist of all integrative analysis performed throughout this study, including genetic, physical, and transcriptional interactors of SFPQ, and the biological processes associated. This study portrays broad potential in future research of SFPQ, through identification of novel interactors, as well as classification of the binding locations of known interactors.

2.0 Methods

Unless specified, statistical analysis and functional programming described was completed utilising R (4.0.3-4.1.1) and results were visualised using ggplot2 (Wickham, 2016).

2.1 Data availability and systematic mining of public repositories

RISMed R package was utilised to search for dataset titles containing the gene name as a keyword, at the National Center for Biotechnology Information (NCBI) PubMed and Gene Expression Omnibus (GEO). Corresponding dataset accession IDs were reviewed, and the individual datasets were manually curated with annotation of their relevance to our study (See supplementary: S1), based on experimental design. Additionally, all available datasets targeting SFPQ at Encyclopedia of DNA

Elements (ENCODE) were inspected and reviewed manually alongside PubMed and GEO dataset curation. For literature-based text mining, RISMed R package was also used to identify known interactors of SFPQ in existing literature. We retrieved NCBI PubMed articles with cooccurring instances of SFPQ or aliases in titles and abstracts. Results were compiled into a dataset and annotated for relevant evidence of SFPQ interactors.

2.2 RNA-seq analysis

For GEO and ENCODE RNA-seq datasets, Trimmomatic was used to filter raw reads, removing the adaptors and low-quality reads ($q < 20$) (Bolger, Lohse and Usadel, 2014). Processed reads were aligned to the human reference genome GRCh38/hg38 assembly using HISAT (v2.1.0) with default settings (Pertea *et al.*, 2016). Expression counts across genomic features were generated using HTSeq (v0.11.1) on human GRCh38 reference annotation (GENCODE release 32) (Anders, Pyl and Huber, 2015).

We used the R package, edgeR, for differential gene expression (DGE) analysis between two groups of expression counts, based on the negative binomial distribution (Robinson, McCarthy and Smyth, 2010). EdgeR assumes that variance of gene counts depends on the negative binomial and the quasi-likelihood dispersion parameters (Ren and Kuan, 2020).

Prior to DGE analysis, raw expression counts were normalised for counts per million (CPM) using the trimmed mean of M values (TMM) method (Robinson & Oshlack, 2010). CPM of genes were filtered to remove genes with low expression using the edgeR function, `edgeR::FilterByExpr`, by which genes with an expression count of <10 in x samples per group, where x is equal to the number of samples in the group with the fewest samples, are removed (Robinson et al., 2010). Filtered data were treated with one of two statistical analysis methods in order to complete differential expression comparisons between groups of expression counts. In datasets containing comparison between genes in only one cell line, the exact test was used to compare between two groups of expression counts i.e. control and SFPQ-kd.

In cases where datasets contained biological replicates, DGE analysis was performed by fitting the Generalised Linear Model (GLM) for analysis by quasi-likelihood F-test in order to account for covariate factors, where cell lines were grouped as factors (Lund et al., 2012). In both cases, resultant tables consist of computed P-values, FDR adjusted P-values, and comparative Log_2 fold change (logFC) (See supplementary: S3). Fold change (FC) is expressed as logFC to allow for clear comparison where results are symmetrically expressed and zero centralised, where positive and negative values with a change greater than 0.2 are defined as 'upregulated' and 'downregulated', respectively. In our analysis, we were interested in a large-scale observation of statistically significant DEG, where FC thresholds were used with low stringency to define the direction of change in expression and a large number of DEG are removed in downstream integration of other data i.e. RIP-seq, eCLIP, PARCLIP and ChIP-seq.

DGE analysis workflow was compiled into a reusable R function, `idbSFPQ::deEdgeR`, where statistical thresholds can be altered as necessary.

2.3 Gene ontology analysis

We used gene ontology to assesses the biological pathways associated with SFPQ interactors. Gene ontology (GO) analysis utilises evidence-based networks to describe the biological roles of listed genes (Ashburner et al., 2000). We programmed the function, `clusterProfiler_OR`, utilising the R package clusterProfiler and human Bioconductor annotation database (org.Hs.eg.db) to reveal the functional pathways associated with gene lists (Yu *et al.*, 2012). ClusterProfiler uses the over-representation (OR) test to identify enriched Gene Ontology (GO) terms of input genes. OR tests use statistical modelling to analyse whether specific functional pathways are OR within an experimentally derived gene list.

GO terms are defined by the GO Consortium, with biological processes relating to the specific biological objective enriched by a gene product and molecular function denoting the associated biochemical activity (Ashburner *et al.*, 2000). Returned ontology terms from OR tests were

considered significant with $q < 0.1$. GO terms are organised with acyclic structure of hierarchical parent-child terms, and therefore parent terms may be enriched because a child-term is significantly over-represented (T. Wu et al., 2021). Thus, significant terms were then filtered for redundancies by semantic similarities (Wang *et al.*, 2007). In cases where we observed GO analysis of multiple gene lists, clusterProfiler function `clusterProfiler::compareCluster` was used to analyse and compare ontologies. Results were visualised using integration of R package enrichplot, via dotplots, cnetplots, and emapplots.

2.4 Peak analysis

In order to uncover the binding locations from publicly available immunoprecipitation datasets, pre-processed peaks were analysed using the R package ChIPseeker (Yu, Wang and He, 2015) and UCSC annotation packages (TxDb.Hsapiens.UCSC.hg38.knownGene ; org.Hs.eg.db). For datasets containing peaks aligned to a reference genome other than hg38, called peaks were lifted over to hg38 alignment using the R package, rtracklayer with the `liftOver` function (T. Wu et al., 2021). Annotative analysis relied on the input of browser extensible data (BED) files containing an annotated chromosome and genomic coordinates containing start and end nucleotides of peak sites. For datasets where significant peaks were not defined, additional parameters were utilised in downstream analysis including experiment-specific signal values, and statistical parameters including *p* and *q* values. Genomic coordinates of peak sites were annotated for overlapping genes (Ensembl ID, Entrez ID and HGNC symbol) and transcripts and categorised genomic features (intron, exon, 3' UTR, 5' UTR, promoter (< 1kb proximity), intergenic, distal intergenic) using the R package, ChIPseeker. For each genomic coordinate, set sizes were calculated and visualised via upset plot using the R package, UpSetR (Conway et al., 2017).

2.5 Downstream analysis to annotate SFPQ-RNA binding locations

2.5.1 eCLIP analysis

SFPQ interactions assessed by eCLIP were obtained from ENCODE in narrowPeak BED format. eCLIP assays are used to identify the binding sites of RNA-binding proteins (RBP). The inclusion of barcoded single-stranded DNA adaptors allows for reduced amplification bias at single-nucleotide resolution, with increased efficiency and specificity compared with regular CLIP experiments by modified preparation protocol (Van Nostrand *et al.*, 2016). ENCODE eCLIP datasets were collected for analysis and pre-processed genomic coordinates of significant peaks ($q < 0.05$) were annotated using peak analysis (as described, 2.5).

2.5.2 PAR-CLIP analysis

SFPQ interactions assessed by PAR-CLIP were obtained from GEO in BED format. Photoactivatable Ribonucleoside-enhanced Cross-Linking and Immunoprecipitation (PAR-CLIP) is another derivative of CLIP methodology useful for identifying binding sites of RBP. PAR-CLIP experimental approach consists of incorporating RNA transcripts with photoactivatable ribonucleoside, commonly 4-thiouridine (4SU), by which binding sites can be identified through characteristic T>C mutations during reverse transcription, allowing for high accuracy results with low background signal (Danan, Manickavel and Hafner, 2016). Pre-processed GEO PAR-CLIP datasets were analysed where significant peaks (PARalyzer score > 0.5). PARalyzer score is an output parameter from analysis via the PARalyzer tool where T>C mutations are exploited in order to produce a value determining the binding status of RBP upon transcripts (Corcoran et al., 2011). Significant peaks were annotated using peak analysis (as described, 2.5). For the integration of PAR-CLIP datasets, each analysis was labelled and combined into a single dataset.

2.6 Downstream ChIP-seq analysis to annotate SFPQ-DNA binding locations

Chromatin Immunoprecipitation (ChIP), combined with high throughput sequencing methods, is a useful technique in identifying target DNA binding sites for proteins (Blecher-Gonen et al., 2013). Succinctly, DNA-protein complexes are induced through in vivo treatment with formaldehyde before Immunoprecipitation with antibodies specific to the DNA-binding protein. Sequencing allows for the identification of targets pulled down in DNA-protein complexes. First, we filtered pre-called ChIP-seq peaks for significant peaks ($p < 0.05$). Significant peaks were then annotated via peak analysis (as described, 2.5). Following peak analysis, ChIP-seq datasets were tagged with a unique dataset label and combined into a single dataset.

2.7 Identification of epigenetic mechanisms underlying SFPQ interactions

We reasoned analysis of SFPQ epigenetic mechanisms due to previous evidence of SFPQ involvement in transcriptional regulation (C.-F. Wu et al., 2013). Epigenetic analysis and comparisons in HepG2 cells were performed using histone datasets at ENCODE. Histone 3 tri-methylation of lysine 3 (H3K4me3) (ENCSR575RRX) and lysine 27 (H3K27me3) (ENCSR000AOL), histone 3 mono-methylation of lysine 3 (H3K4me1) (ECNCSR000APV), and histone 3 acetylation of lysine 27 (H3K27ac) (ENCSR000AMO) ChIP-seq datasets were downloaded from ENCODE and compiled into a single BED file annotated by histone modification mark. Active promoters were defined in peaks that overlap H3K4me3 and lack H3K4me1 (Bae & Lesch, 2020; Barski et al., 2007). Peaks overlapping H3K27me3 mark regions of transcriptional repression (Barski et al., 2007). Peaks overlapping both H3K4me1 and H3K27ac mark enhancers, whereas enrichment of repressive mark, H3k27me3, in place of H3K27ac denotes poised enhancers (Heintzman et al., 2007; Zentner et al., 2011). Poised enhancers are described as distal regulatory elements with absent

histone acetylation, and are therefore not transcribed but can be rapidly activated upon cellular differentiation (Creyghton et al., 2010).

First, SFPQ ChIP-seq datasets were filtered for peaks generated from HepG2 datasets, constructing a BED file which only contained analysed SFPQ-DNA interactions in HepG2 cells. Next, we compressed ENCODE ChIP-seq histone modification datasets into a single BED file, labelled by their respective histone mark. We used Bedtools (v2.28.0) `intersect` function to pull down SFPQ ChIP-seq peaks which overlap with histone mark peaks. Based on combinations of overlapping histone mark peaks, SFPQ ChIP-seq peaks were categorised based on combinatorial histone modifications into regions of active promoters, enhancers, poised enhancers, and repressors.

2.8 Data integration

For integration of RNA-seq DGE results we programmed the R function, `DGEintegrateR`. `DGEintegrateR` uses base R functions to load DGE results i.e., Gene ID, P-value and logFC, collapsing them into a single R dataframe, using Ensemble ID as a gene identifier. RNA-seq datasets were integrated based on P-value and logFC, where genes were categorised twice per dataset. The first categorisation classified gene-wise significance in individual studies where genes tested significant with FDR<0.1. Second, genes were categorised based on the direction in which they were Differentially expressed (DE), with a logFC threshold of 0.2 i.e. upregulated with logFC>0.2 and downregulated with -0.2>logFC.

Genetic interactors were then characterised as genes which test significant (FDR<0.1) and are DE in the same direction in two or more datasets. For genetic analysis, the means that genes must be significantly DE in the same direction in 40% of SFPQ-kd RNA-seq datasets. Our justification for this relaxed threshold is to account for the large interstudy differences in sample size due to dataset-specific variability. This level of stringency dampens the rise of type I errors, where genes may be discarded due to a lack of coverage across datasets. Whereas many false positives will be filtered out as we begin to integrate multi-omics analysis as previously mentioned.

Identification of genetic interactors in HepG2 involved selecting genes which were significantly differentially expressed in either of the HepG2 SFPQ-kd RNA-seq datasets. In rare cases where regulatory direction differed between HepG2 datasets (2.7%), the consensus direction from complete integration was selected.

This methodology was slightly altered for RIP-seq analysis where downregulated genes were discarded, and genes were characterised as enriched if significantly upregulated in two or more datasets

In order to integrate peaks datasets, significant peaks were compiled in a row-wise manner into a single dataset per technique.

2.9 idbSFPQ: Database and R package

All results from this study, including analysis of SFPQ-kd RNA-seq, RIP-seq, eCLIP, PAR-CLIP, and ChIP-seq data were compiled into the database, idbSFPQ, and stored in a GitHub repository. In addition, to base analysis of interactors at each omics level, results from GO analysis are also available in idbSFPQ. In doing so, idbSFPQ can be mined for information regarding SFPQ interactions for applications in further study. For example, idbSFPQ can be used to mine information of novel interactors of SFPQ to inspire validation of interactions *in vivo*.

In addition to acting as a mineable database, idbSFPQ is a loadable R package which contains functions programmed for analysis workflow throughout this project, including deEdgeR (as described, 2.2), clusterProfiler_OR (as described 2.3) and DGEintegrateR (as described 2.9). This means that functions loaded from idbSFPQ can be used in contexts outside the SFPQ interactome and can be applied in investigating the regulatory network of other genes.

3.0 Results

Previous studies have shown that SFPQ regulates a range of biological processes through transcriptional, physical and genetic interactions at RNA and DNA level (Knott, Bond and Fox, 2016;

Iida, Hagiwara and Takeuchi, 2020). However, the complete spectrum and scale of the SFPQ regulatory network remains unknown. We aimed to map the SFPQ regulome exhaustively, constructing an interaction database of novel interactors for further study.

In order to map the SFPQ network, we took advantage of publicly available datasets containing high throughput sequencing analyses targeting SFPQ, including genetic, physical and transcriptional data across multiple tissue cell lines (Figure 2). We began by conducting data mining analysis to curate all available datasets containing experiments targeting SFPQ in human cell lines at GEO and ENCODE.

Table 2

Datasets collected through systematic data mining, including SFPQ-kd RNA-seq, RIP-seq, eCLIP, PAR-CLIP, and ChIP-seq datasets. Datasets are annotated by dataset UID, experimental technique, cell line context, and publication reference.

Dataset UID	*Experimental technique*	*Cell line*	*Notes*	*Reference*
ENCSR782MXN	SFPQ-kd RNA-seq	HepG2	shRNA	(Consortium, 2012)
ENCSR535YPK	SFPQ-kd RNA-seq	K562	shRNA	(Consortium, 2012)
GSE157622	SFPQ-kd RNA-seq	HepG2	siRNA	(Stagsted et al., 2021)
GSE157622	SFPQ-kd RNA-seq	HEK293T	siRNA	(Stagsted et al., 2021)
GSE149370	SFPQ-kd RNA-seq	CaCo2	shRNA	(Klotz-Noack et al., 2020)
GSE94243	RIP-seq	22Rv1, LTAD, LNCaP	-	(Takayama *et al.*, 2017)
GSE114394	RIP-seq	22Rv1	-	(Takayama *et al.*, 2017)
GSE133423	RIP-seq	MCF-7	-	(Iino et al., 2020)
ENCSR965DLL	eCLIP	HepG2	-	(Consortium, 2012)
GSE113349	PAR-CLIP	HeLa	-	(Yamazaki et al., 2018)
GSE113349	PAR-CLIP	U2OS	-	(Yamazaki et al., 2018)

ENCSR757HBB	ChIP-seq	HepG2	-	(Consortium, 2012)
ENCSR258SXK	ChIP-seq	HepG2	-	(Consortium, 2012)
ENCSR647PJW	ChIP-seq	K562	-	(Consortium, 2012)
GSE94577	ChIP-seq	LTAD	-	(K. Takayama et al., 2020)

Datasets were manually filtered for relevance to this study based on experimental techniques, distinguishing whether datasets would be useful for initial analysis or validation of results. The initial analysis consisted of five RNA-seq datasets with SFPQ-knockdown (SFPQ-kd), three RIP-seq, one eCLIP, two PAR-CLIP and four ChIP-seq datasets (Figure 2B, Table 2).

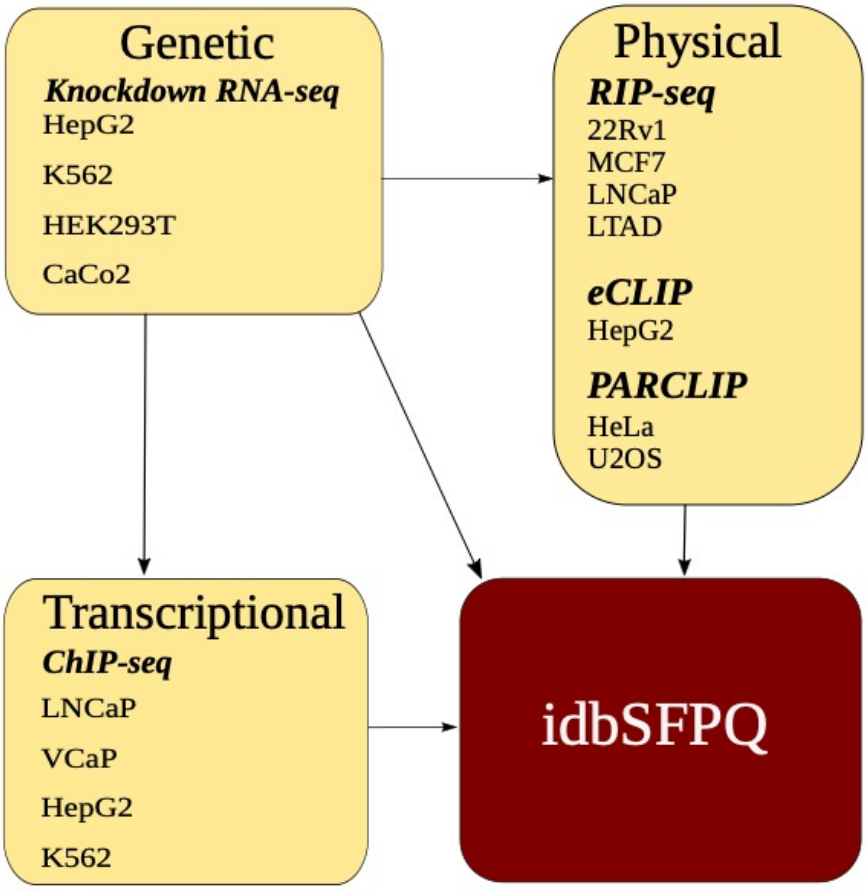


Figure 2:

Workflow depicting integration of multiple sequencing analysis techniques from numerous publicly available datasets to construct interaction database, 'idbSFPQ'.

3.1 Analysis of SFPQ genetic interactors

Initially, we set out to identify genetic interactors of SFPQ, by performing downstream analysis of SFPQ-kd RNA-seq datasets. We define 'genetic interactors' as genes with significantly altered expression levels under conditions where SFPQ is perturbed genetically. Our database search identified five RNA-seq datasets on cell lines depleted with SFPQ across four tissue types, including liver (HepG2; ENCSR782MXN, GSE157622), kidney (HEK293T; GSE157622), colon (CaCo2; GSE149370), and chronic myelogenous leukaemia (CML) (K562; ENCSR535YPK) (Figure 2).

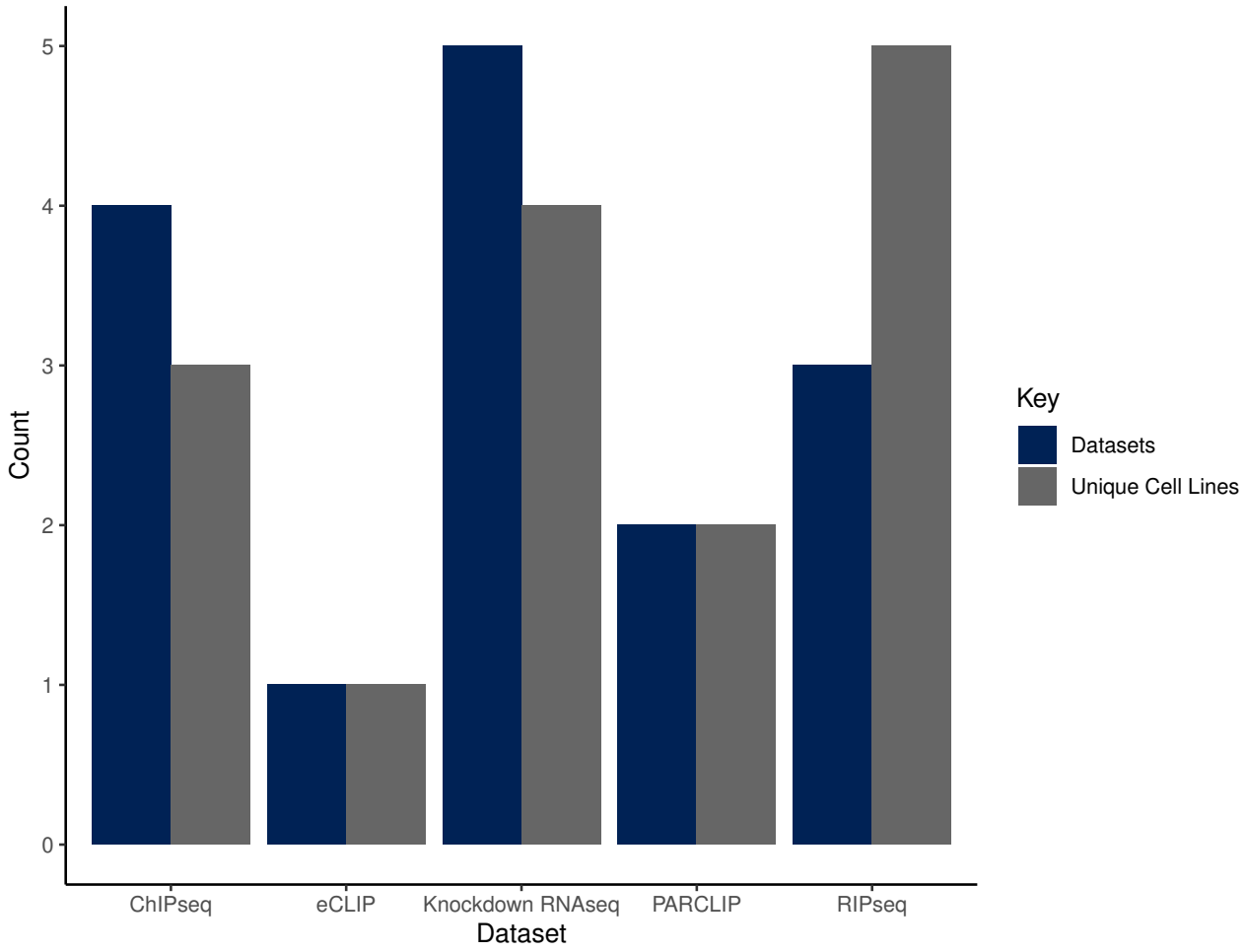


Figure 3: Number of datasets collected per analysis technique.

Differential gene expression (DGE) analysis was performed on SFPQ-kd RNA-seq datasets individually, comparing expression levels of genes between control cell lines and cell lines depleted

of SFPQ. In this instance, where each SFPQ-kd dataset only contains analysis for one cell line, gene-wise exact tests were performed on negative binomially distributed expression counts to observe the effect of SFPQ depletion on gene expression levels. Differentially expressed genes (DEG) were defined as genes with False discovery rate (FDR) < 0.1 and Log fold change (LogFC) > 0.2 within each dataset.

Table 3: Total genes within each SFPQ-kd dataset following normalisation and filtering in RNA-seq analysis. DEG expressed as a percentage of total genes in corresponding dataset. Further categorisation of total upregulated and downregulated DEG and the percentage proportion they occupy within the DEG of that dataset.

Dataset	*Cell line*	*Total genes*	*DEG*	*Upregulated*	*Downregulated*
ENCSR782MXN	HepG2	14, 573	3, 861 **(26.5%)**	1, 976 **(51.2%)**	1, 885 **(48.8%)**
GSE157622	HepG2	18, 693	5, 002 **(26.8%)**	3, 119 **(62.4%)**	1, 883 **(37.6%)**
GSE157622	HEK293T	18, 538	6, 193 **(33.4%)**	3, 116 **(50.3%)**	3, 077 **(49.7%)**
GSE149370	CaCo2	19, 301	2,163 **(11.2%)**	1, 118 **(51.7%)**	1, 045 **(48.3%)**
ENCSR535YPK	K562	14, 682	5, 534 **(37.7%)**	2, 812 **(50.8%)**	2, 722 **(49.2%)**

More than 2000 DEG were identified in each cell line, with variation of absolute values between datasets. The number of DEG within each dataset ranged from 2,163 (CaCo2) and 6,193 (HEK293T). Generally, the percentage of DEG in the total genes per filtered dataset was close to 30% with the exception of CaCo2 which was much lower than the other datasets (11.2%), despite having the largest total genes. Additionally, the proportion of upregulated DEG compared with downregulated DEG were mostly even, excluding HEK293T which contained almost 25% more upregulated genes than downregulated. In all datasets, although contrasts were often minor, more genes were upregulated than downregulated in depleting of SFPQ (Table 3).

3.2 Integration of genes with altered expression under knockdown levels of SFPQ

With aims of mapping the SFPQ regulatory network in a comprehensive manner, we were interested in DEG consistently altered in multiple experiments across multiple tissues. Therefore, we identified DEG which overlap between datasets. First, many observed genes are altered in a dataset-specific way as a large amount of the total DEG are only altered in a single dataset. However, when we observe the individual proportion of DEG, we find that on average, over 58% of the DEG in each dataset are DE in at least one other dataset (Figure 4, Table 3).

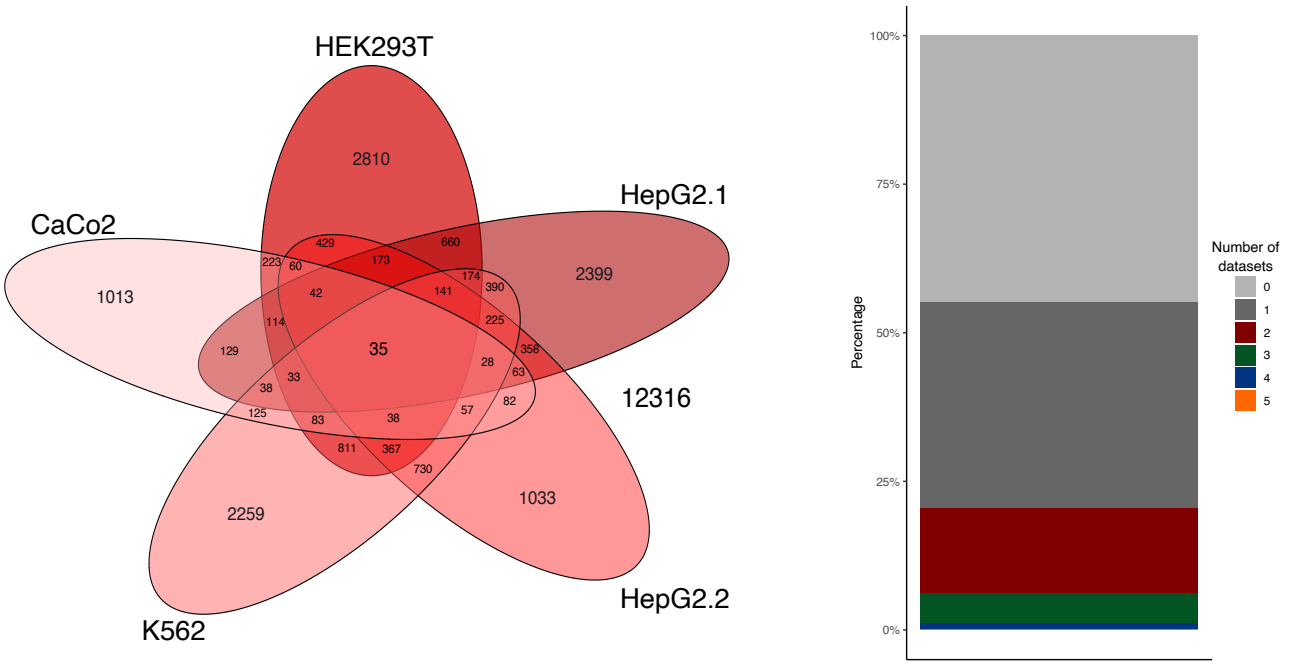


Figure 4

Venn diagram of overlapping DEG between each SFPQ-knockdown RNA-seq dataset (left). Percentages of genes which are significantly DE in a given number of datasets (right).

Amongst the total genes observed in genetic analysis, 44.9% of genes were not DE in any of the datasets, 34.7% of genes were DE in only one dataset, and 16.2% were defined DEG, DE in at least

two datasets. 4.9% of total genes were DE in three datasets, 1% were DE in four datasets, and a total of 35 genes (0.1%) were DE in all five genetic analysis datasets. (Figure 4).

In full integration of SFPQ genetic interaction datasets, genes were further categorised based on the direction in which gene expression was altered under reduced SFPQ expression levels. Genetic interactors were only selected from DEG if significantly DE in the same direction in at least two of the five genetic datasets (as described 2.8). With depleted levels of SFPQ, 2458 genes were upregulated and 1973 downregulated.

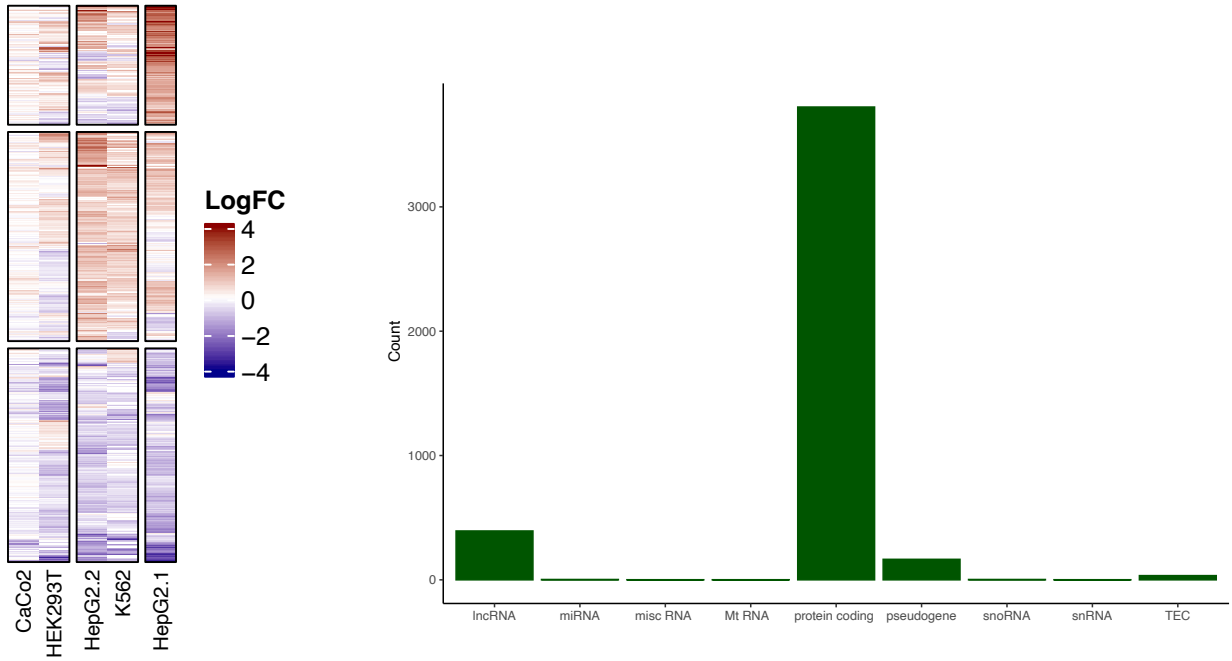


Figure 5

Heatmap displaying regulatory direction of a selection of genes across genetic analysis datasets, were positive LogFC represents genes that are upregulated with depleted levels of SFPQ (left). Number of significant DEG representing different genetic biotypes (right).

By constructing a heat map, based on fold change of DEG which test significance in three or more datasets, we were able to confirm that, in addition to DEG being shared amongst datasets, they are mostly consistent in regulatory direction (Figure 5). Interestingly, integration revealed 19 genes

regulated in opposite when compared between datasets e.g., upregulated in two RNA-seq datasets, whilst also downregulated in two other datasets. These genes include protein-coding genes CDK18, QSOX1, DEPP1. However, DEG with contrasting regulatory directions only occupy 0.4% of the total DEG, whereas the majority are distinctly regulated in a common direction between datasets.

Next, we explored the gene biotypes that SFPQ interacts with at the genetic level, utilising the R package, biomaRt and Ensembl database server. The majority of SFPQ genetic interactors are protein-coding genes (86.4%), with fewer non-coding RNA (13.6%), including lncRNA transcripts (8.9%) and pseudogenes (3.7%) (Figure 6).

3.3 SFPQ is associated with a large range of biological processes

To investigate the biological processes regulated by SFPQ, we applied gene ontology (GO) analysis on DEG (defined in 2.3) of SFPQ. We observe that SFPQ regulates a large range of biological processes, involving ontologies associated with localisation of organelles, cellular response to stress signals, and metabolic processes. Interestingly, many of the ontologies are already evidenced to be associated with SFPQ, including transcription regulation of neuron projection development. (Figure 6, red text). SFPQ regulates many developmental processes, specifically, a large proportion of the top 15 upregulated ontologies (ranked by FDR) are associated with tissue development (Figure 6).

There is a clear distinction between biological processes regulated by upregulated and downregulated genes, where, excluding regulation of small GTPase mediated signal transduction, ontology terms do not seem to be shared between upregulated and downregulated gene clusters (Figure 6). In SFPQ-kd, upregulated biological processes are mostly responsible for regulation of developmental processes, homeostasis, response to stress signals/ damage stimuli and rhythmic processes, and downregulated ontologies majorly regulate metabolic processes, progressions in cell cycle processes, and transport of organelles.

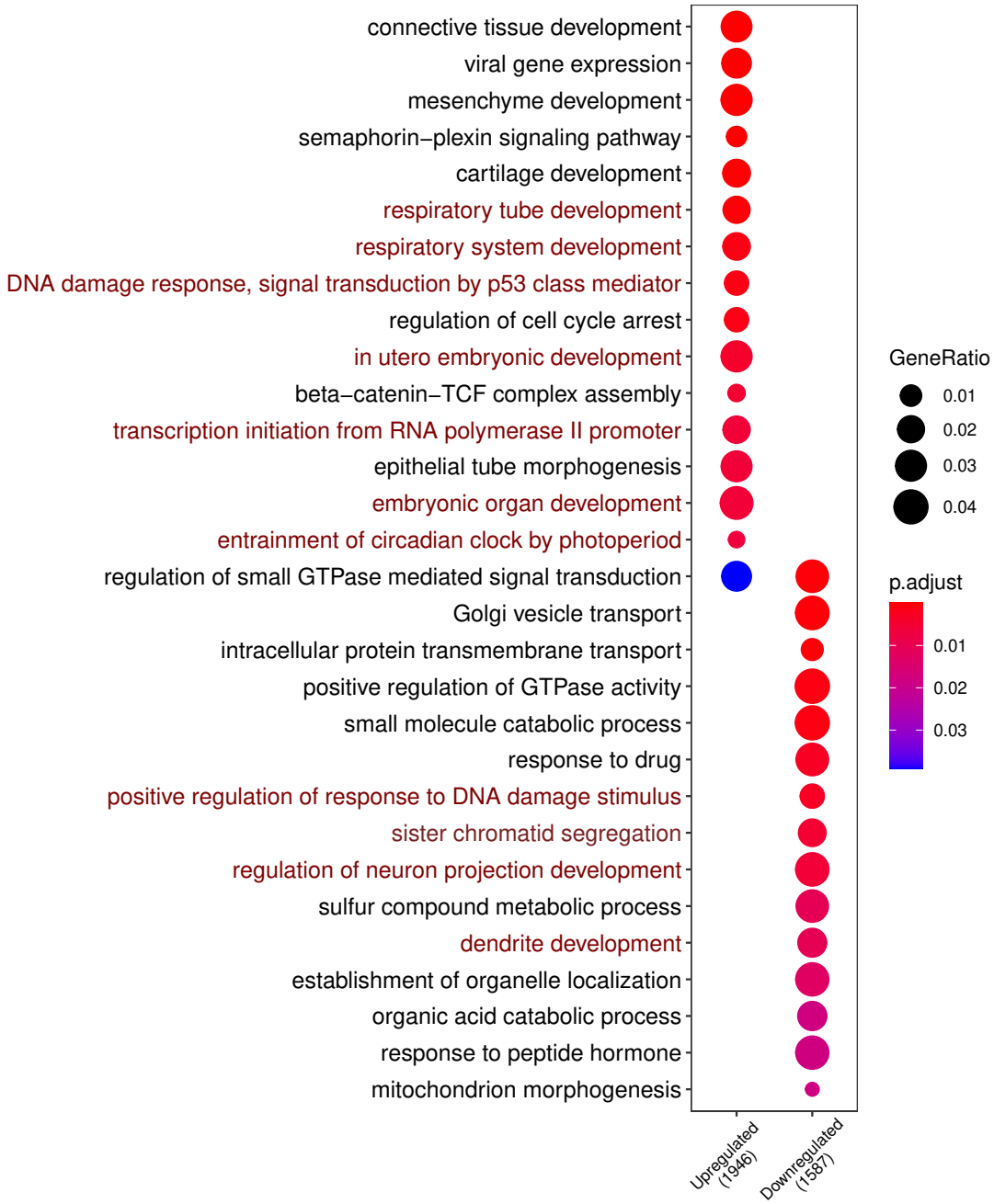


Figure 6

The top 15 ranked GO terms according to adjusted P-Value (FDR), in both upregulated and downregulated clusters of DEG. Terms highlighted red are ontologies already associated with SFPQ. DEG, differentially expressed genes, FDR, False discovery rate.

As well as sub-cellular organisations, depletion of SFPQ appears to upregulate genes involved in extracellular arrangement, including cell-substrate adhesion, tissue migration and extracellular matrix organisation. Further, SFPQ regulates genes involved in numerous cell-signalling pathways, through

regulation of hormone receptor, extrinsic apoptotic and semaphorin-plexin signalling pathways. (See Supplementary: S4).

3.4 Investigating SFPQ physical RNA interactors

Analysis of SFPQ-kd RNA-seq datasets helps us identify genetic interactors but does not indicate whether genes are being regulated directly or indirectly. To address this, we investigated transcripts that physically associate with SFPQ. Assessment of physical interactors began with downstream analysis of RIP-seq data targeting SFPQ in multiple tissue cell lines, which provides a global understanding of SFPQ-RNA interaction. Prostate cancer (22Rv1, LTAD, LNCaP; GSE94243, GSE114394) and breast cancer (MCF-7; GSE133423) datasets were analysed via full RNA-seq analysis to generate gene-wise expression counts and DEG (as described, 2.2) (Table 2). For the integration of RIP-seq analysis, SFPQ physical interactors from RIP-seq data were deemed significant if they were significantly enriched (FDR < 0.1, LogFC > 0.2) in at least two of the three datasets. Following full integration of RIP-seq analysis datasets, we identified 3331 genes which form protein-RNA complexes with SFPQ.

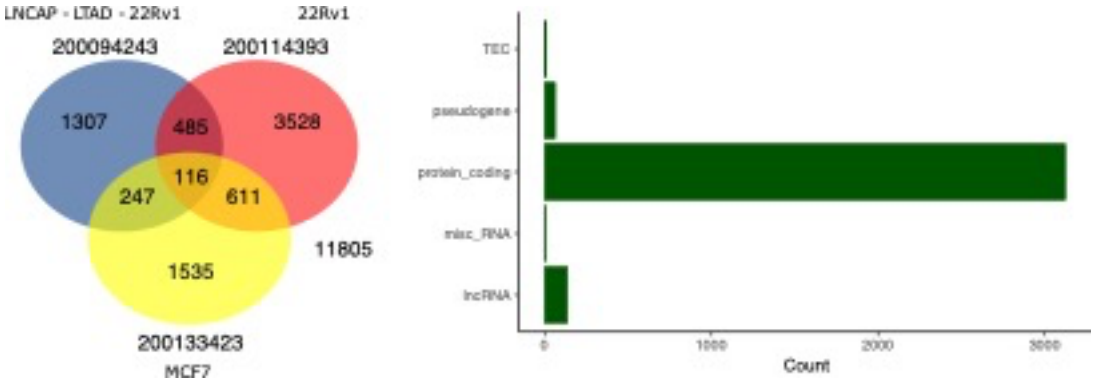


Figure 7

Venn diagram distributing overlapping genes in integration of RIP-seq datasets (left). Number of genetic biotypes physically binding SFPQ from RIP-seq analysis integration

Consistent with the distribution of genetic interactors, the most common physical interactors of SFPQ are protein-coding genes (93.8%), with a lower frequency of non-coding genes, with the majority of non-coding genes classifying as lncRNA (4.1%) and pseudogenes (1.8%) (Figure 7).

3.5 Identification of SFPQ-RNA binding sites

A disadvantage of RIP analysis is that it pulls down whole transcripts, without indicating specific SFPQ-transcript binding regions. To discover SFPQ physical interaction sites, we took advantage of enhanced cross-linking (eCLIP) and photoactivatable-ribonucleoside-enhanced cross-linking (PAR-CLIP) immunoprecipitation datasets. We analysed three datasets (Figure 2) including SFPQ eCLIP in liver cancer cells (HepG2; ENCSR965DLL), and SFPQ PAR-CLIP in cervical cancer (HeLa**;** GSE113349) and osteosarcoma epithelial (U2OS; GSE113349) cells.

Pre-processed signal peaks (eCLIP: P-value < 0.05 or PAR-CLIP: PARalyzer score > 0.5) were annotated, using the R package, ChIPseeker, and UCSC annotation databases, to identify overlapping genes and further annotate transcripts, exons, and biotypes (promoter < 1kd, intron, exon, 5’ UTR, 3’ UTR, etc.)

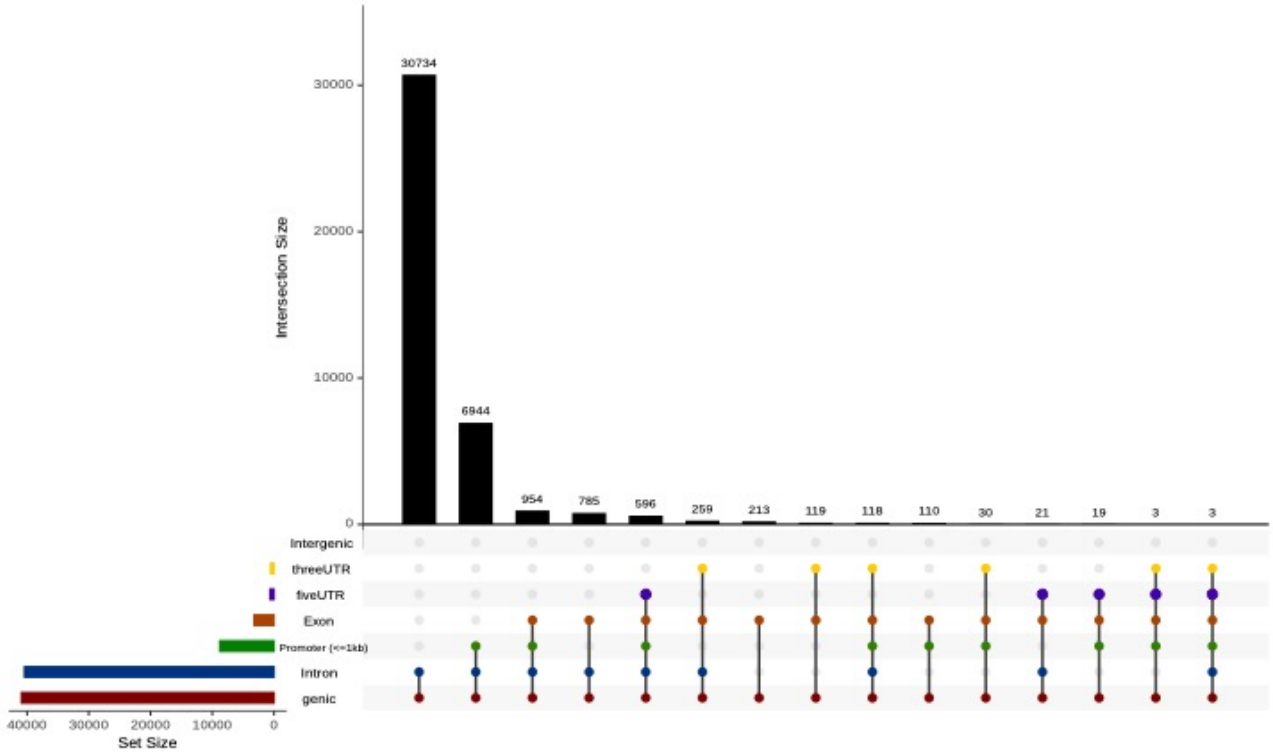


Figure 7

Upset plot distributing gene biotypes of SFPQ eCLIP interactions.

In eCLIP analysis, most SFPQ-RNA binding regions overlap introns, with a considerable number of peaks binding proximal to promoters (Figure 8). Similarly, SFPQ-RNA interactions in PAR-CLIP analysis occur majorly at introns, alongside some binding near promoter regions (See Supplementary: S6). In the majority of cases, RNA interactions do not occur at 5' UTR and 3' UTR regions, simultaneously. In cases where SFPQ-RNA interactions overlap both 5' UTR and 3' UTR, SFPQ lacks bias in binding specific biotypes, and appears to bind many regions across target transcripts.

3.6 Integrative analysis of SFPQ physical interactors

Under the assumption that integrating datasets from multiple experimental analyses will enhance the robustness of our interaction network, we combined transcript-level resolution RIP datasets with nucleotide-resolution methods (eCLIP and PAR-CLIP). Significant interactors from RIP-seq datasets (as described previously) were extracted and overlapped with genes containing at least one significant peak (see methods) in eCLIP and PAR-CLIP datasets. We identified 892 genes that significantly

interact with SFPQ across the three analysis methods. Ontology analysis of these genes revealed significant enrichment of cell morphological processes, including positive regulation of cell projection organisation and regulation of cell morphogenesis (Figure 9).

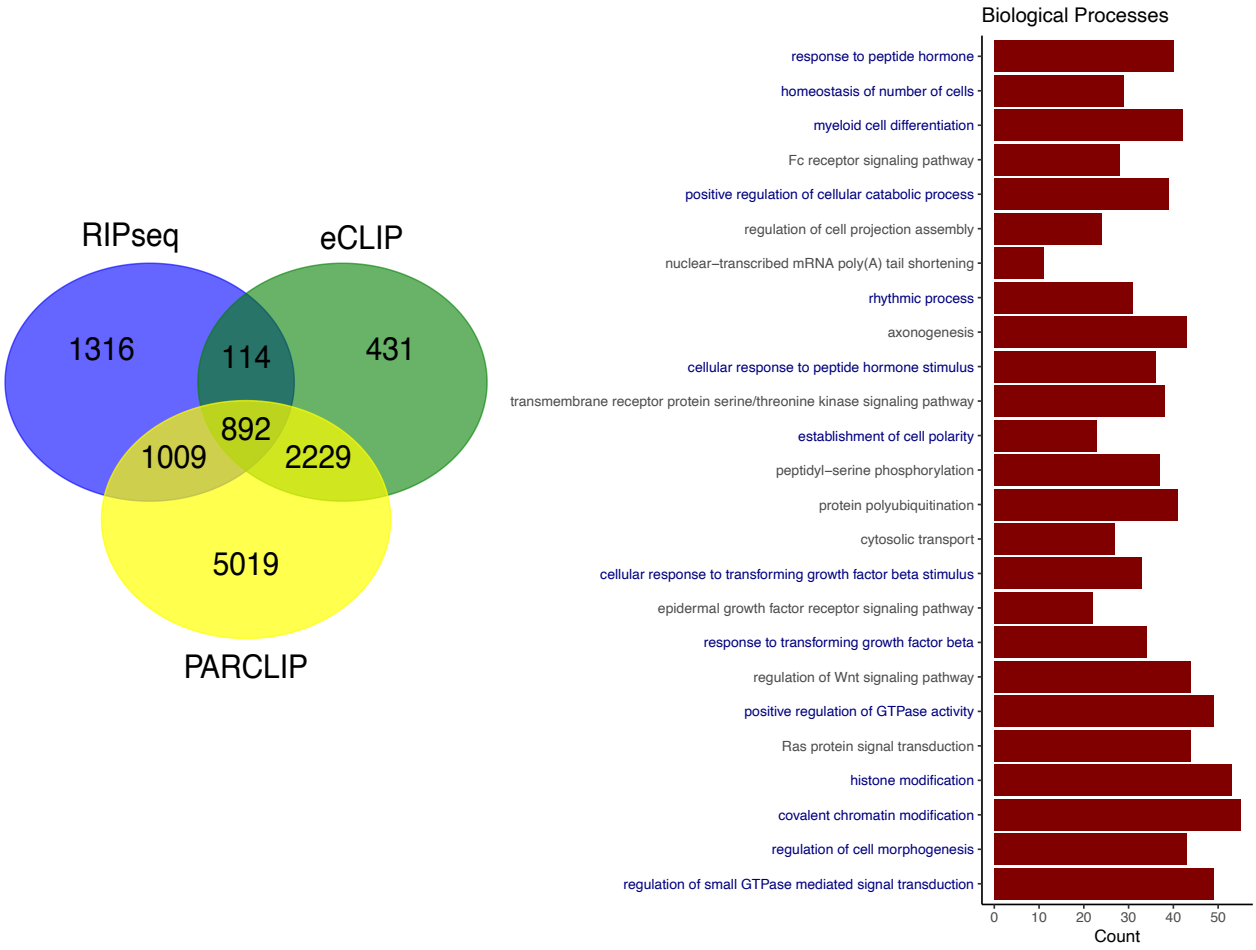


Figure 8

Venn diagram of overlapping significant SFPQ interactors in physical analysis data (left). Top 25 gene ontology terms implicated by integrated physical interactors of SFPQ, ranked by FDR. Blue terms are ontologies consistent with GO analysis in genetic interactors (right).

The remaining terms among the top 25 ontologies include a broad range of processes such as involvement in physiological rhythmic processes, epigenetic modifications, growth factor/ hormone response and Wnt signalling. As one might expect, several ontologies are shared with those associated with genetic interactors (Figure 6), namely regulation of GTPase activity, response to peptide hormones and dendrite development (Figure 9).

3.7 Investigating SFPQ transcriptional interactors

In the third level of multi-omics analysis, we sought to identify genes that interact with SFPQ at DNA level, integrating genes regulated at the transcriptional level into the interaction network. To explore SFPQ-DNA interactions, we interrogated Chromatin immunoprecipitation sequencing (ChIP-seq) datasets. We explored DNA-binding of SFPQ in three tissues, including liver cells (HepG2; ENCSR757HBB, ENCSR258SXK), CML (K562; ENCSR647PJW) and prostate cells (LTAD; GSE94577). To identify interactors pulled down in ChIP-seq datasets, pre-processed signal peaks were analysed using the R package ChIPseeker (as previously described), by which binding regions are assessed for overlapping genes. Through further annotation of overlapping transcripts and exons, with characterisation of biotypes, it is evident that SFPQ binds regions overlapping introns (Figure 10).

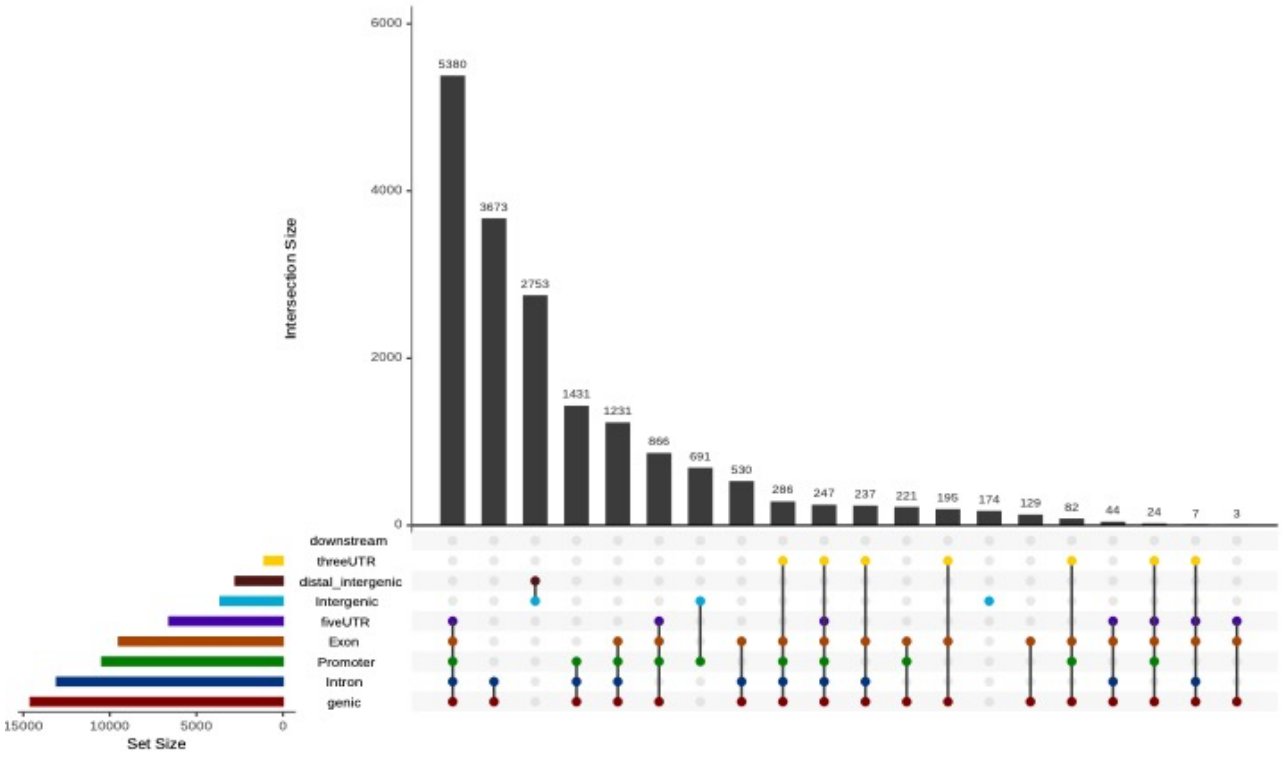


Figure 9

UpSet plot distributing the biotypes overlapped by SFPQ interaction sites from ChIP-seq experiments.

However, in comparison with eCLIP analysis of SFPQ, ChIP-seq results imply copious binding sites overlapping other transcript biotypes. Also, in contrast to RNA interactors, there are a great number of occasions where SFPQ binds intergenic and distal-intergenic regions. There is a general contradiction between interaction sites overlapping 5' UTR and 3' UTR on the same gene, and in the cases where 5' UTR and 3' UTR are overlapped simultaneously, SFPQ also binds a wide array of other transcript biotypes. The most common case of genic binding at the DNA level contains interaction regions which overlap intronic, promoter, exonic and 5' UTR regions.

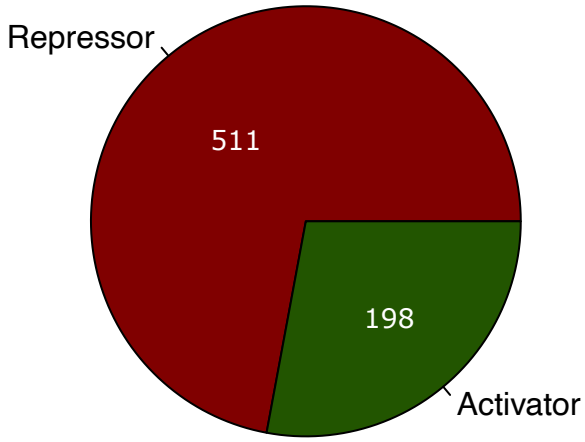


Figure 10

Proportion of genes which interact with promoters in ChIP-seq data and are downregulated in SFPQ-kd (SFPQ acts as a transcriptional activator) and upregulated in SFPQ-kd (SFPQ acts as a transcriptional repressor).

3.8 SFPQ status as transcriptional activator/ repressor

In attempt to estimate the mechanisms by which SFPQ regulates genes at the transcriptional level, we explored the genes which interaction locations overlapping promoter regions in ChIP-seq data, which are also DE under SFPQ-kd conditions (709 genes). We then categorised these genes based on whether they were upregulated or downregulated following depletion of SFPQ, assuming that if SFPQ interacts, with the promoter of a gene which is upregulated in SFPQ-kd, it is likely that SFPQ regulates that gene by acting as a transcriptional repressor, It is clear that of the 709 genes analysed, a large amount of these are regulated via transcriptional repression (511), whereas a smaller number of genes appear to be transcriptionally activated by SFPQ binding (198) (Figure 11).

3.9 Discovery of the SFPQ-DNA regulatory network

After identifying genes which SFPQ bind at the DNA level, we compared results with genetic interactors to explore which genes are regulated by transcription. Data across the full study was integrated in order to subset genes which were DE in SFPQ-kd and interactors in ChIP-seq data but are not physical interactors of SFPQ at the RNA level. We identified 876 genes which SFPQ interacts with exclusively at the DNA level. To observe the biological processes controlled by SFPQ via transcriptional regulation, GO analysis was performed on this subset (Figure 12).

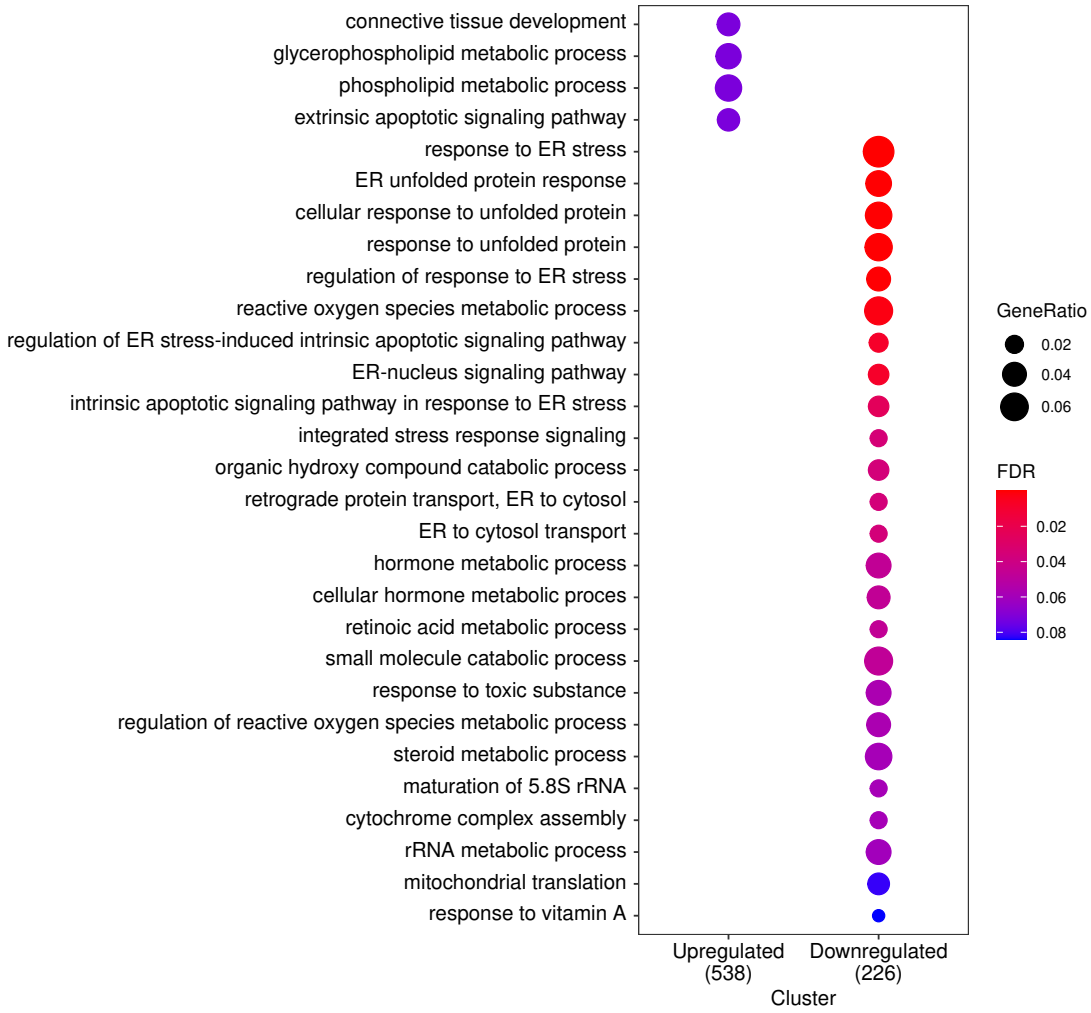


Figure 11

Biological processes enriched by genes DE in SFPQ-kd and which SFPQ binds at DNA level.

Despite the genetic and transcriptional interactors being predominantly upregulated in SFPQ-kd (538 upregulated; 226 downregulated), there are only four associated ontologies, whereas previous GO analysis of SFPQ generated a large range of ontologies associated with both upregulated and downregulated genes. The ontologies that are upregulated in SFPQ-kd include connective tissue development, phospholipid/ glycerophospholipid metabolic processes, and the extrinsic apoptotic signalling pathway. Interestingly, many SFPQ-DNA interactors are associated with processes involving the endoplasmic reticulum (ER), primarily through response to unfolded proteins and stress signals.

3.10 Analysis of SFPQ-regulated biological processes in HepG2

There was only one cell line that was consistently integrated throughout each level of omics analysis, HepG2, consisting of two SPFQ-kd RNA-seq datasets (ENCSR782MXN, GSE157622), one eCLIP dataset (ENCSR965DLL) and two ChIP-seq datasets (ENCSR757HBB, ENCSR258SXK). ENCODE phase III repository contained eCLIP analysis of RBP in HepG2 and K562, however as far as we are aware, eCLIP data targeting SFPQ in K562 cells is unavailable, otherwise this level of analysis would also be possible for K562-specfic SFPQ analysis (Van Nostrand et al., 2020).

First, in selection of SFPQ genetic interactors in HepG2 cells we took the union of significant DEG, where genes were characterised as SFPQ genetic interactors if they were DE in at least one of the two HepG2 RNA-seq datasets, defining 3727 genes as genetic interactors in HepG2 cells. In cases where genetic interactors contrasted in regulatory direction with SFPQ depletion (2.7%), genetic interactors were categorised as upregulated or downregulated based on the consensus direction from previous integration of genetic interactors. SFPQ interactors at post-transcriptional level consist of genetic interactors which were enriched in HepG2 eCLIP data. Similarly, transcriptional interactors include genetic interactors which were enriched in HepG2 ChIP-seq data.

In order to explore the biological processes associated with SFPQ interactors at transcriptional and post-transcriptional level, GO analysis was performed. First, there are several ontologies which are upregulated in post-transcriptional interactors and upregulated in transcriptional interactors.

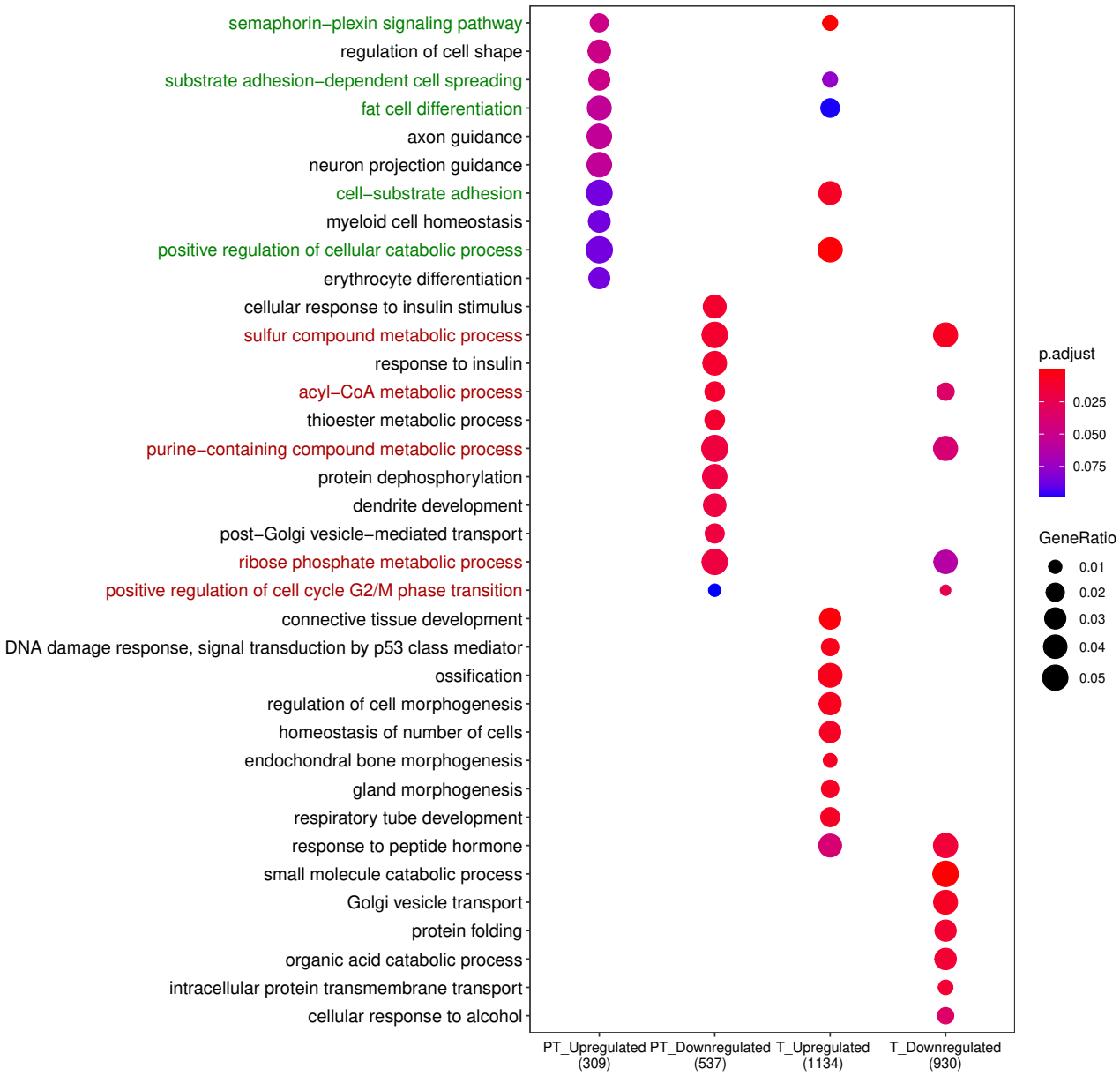


Figure 12

Biological processes of SFPQ interactors, clustered by post-transcriptional interactors (P_T) and transcriptional interactors (T). Transcriptional and post-transcriptional interactors are further clustered into a status defining whether they were upregulated or downregulated in SFPQ-kd RNA-seq. Overlapping ontologies between upregulated transcriptional and post-transcriptional interactors (green). Overlapping ontologies between downregulated transcriptional and post-transcriptional interactors (red).

These ontologies include semaphorin-plexin signaling pathway, substrate adhesion-dependent cell spreading, fat cell differentiation, cell-substrate adhesion, and positive regulation of cellular catabolic processes, although it is notable that post-transcriptional upregulated clustered ontologies have low levels of confidence in FDR value (Figure 13). Similarly, there are several ontologies associated with genes that are both downregulated in SFPQ interactors at transcriptional and post-transcriptional level, including many metabolic processes, and G2/M phase transition in the cell cycle (Figure 13). At post-transcriptional level, SFPQ also interacts with genes associated with response to insulin stimulus, which are downregulated in depletion of SFPQ. Whereas SFPQ interactors at transcriptional level associated with cell and tissue morphogenesis, including gland and endochondral bone, were upregulated in SFPQ depletion (Figure 13). Additionally, transcriptionally downregulated genes in absence of SFPQ are associated with catabolic processes, and intracellular transport (Figure 13).

3.11 Epigenetic analysis of SFPQ interactors at transcriptional level in HepG2 cells

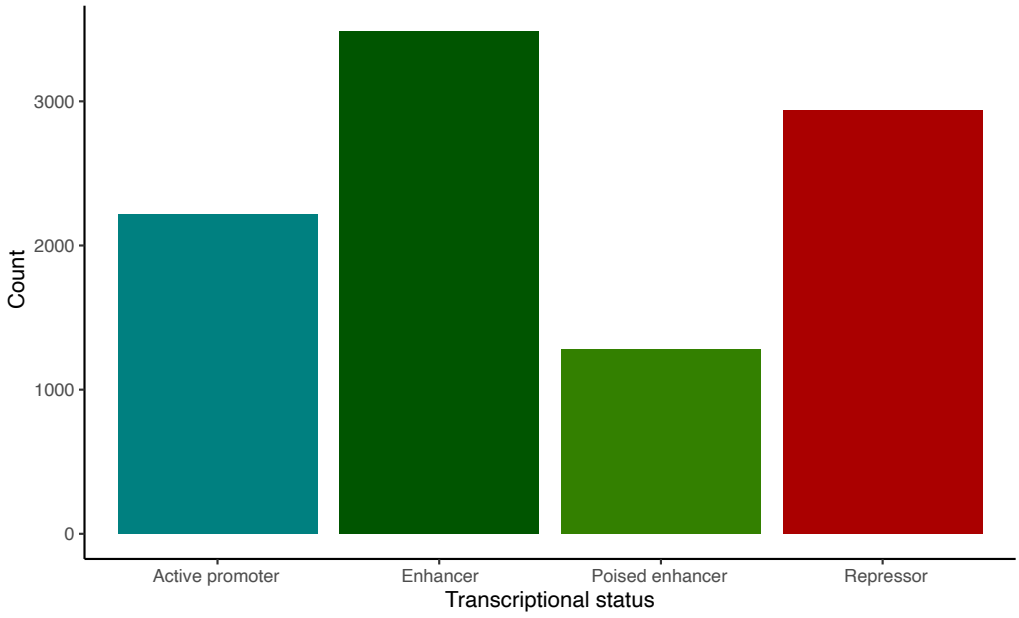


Figure 13

Distributed number of genes which overlap combinations histone mark combinations denoting different transcriptional statuses.

In order to estimate the modes by which SFPQ regulates genes at the transcriptional level, we integrated pre-processed ChIP-seq data containing peaks for genomic DNA methylated and acetylated regions, focusing only on HepG2 ChIP-seq peaks which overlap with transcriptional histone marks. SFPQ-targeted ChIP-seq data from HepG2 cells were compared with histone modification data at ENCODE. Combinatorial histone modifications were annotated in attempt to estimate frequency by which SFPQ binds regions of potential active promoters, enhancers, poised enhancers, and repressors. Peaks overlapping with H3K4me3 and lack of H3K4me1 mark active promoters (Bae & Lesch, 2020; Barski et al., 2007). Peaks which overlap regions annotated enriched with H3K27me3 modifications, indicate transcriptional repression (Barski et al., 2007). SFPQ appears enriched at DNA regions overlapping active enhancer marks most frequently (35.8%) compared with repression marks (29.6%). SFPQ enriched enhancer and repressor regions more frequently than active promoter regions (22.3%) and showed enrichment with histone marks implicating poised enhancer regions (12.9%) (Figure 14).

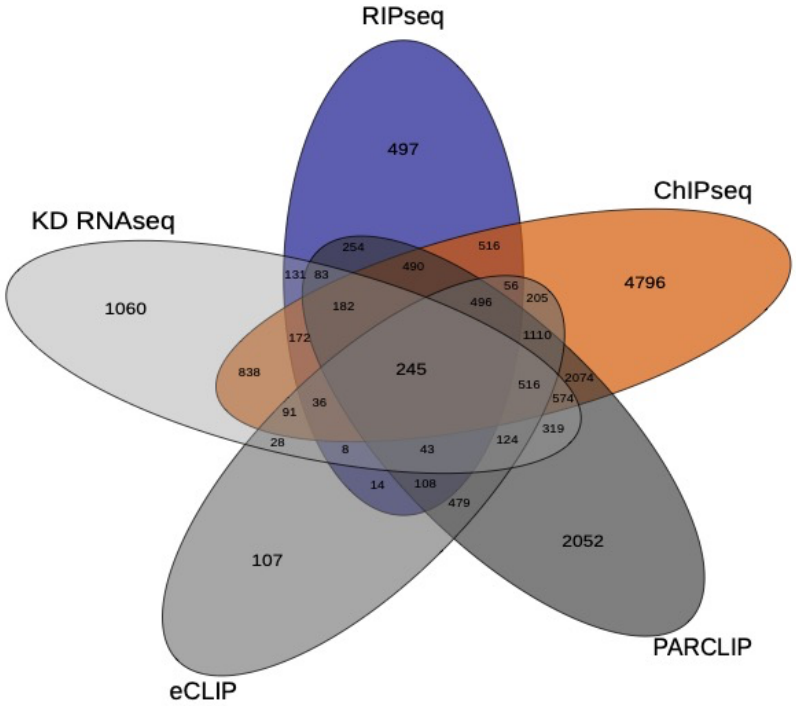


Figure 14

Venn diagram displaying overlapping genes between analysis results from each dataset and analysis technique.

3.12 Complete integrative analysis of SFPQ regulatory network

To characterise robust interactors of SFPQ across multiple scales of regulatory interaction, we integrated results from all levels of analysis to identify genes that are defined significant interactors in all observed datasets (SFPQ-kd RNA-seq, RIP-seq, eCLIP, PAR-CLIP, ChIP-seq). We found 245 genes that matched these criteria and are consistently interacting with SFPQ across multiple cell lines (Figure 15).

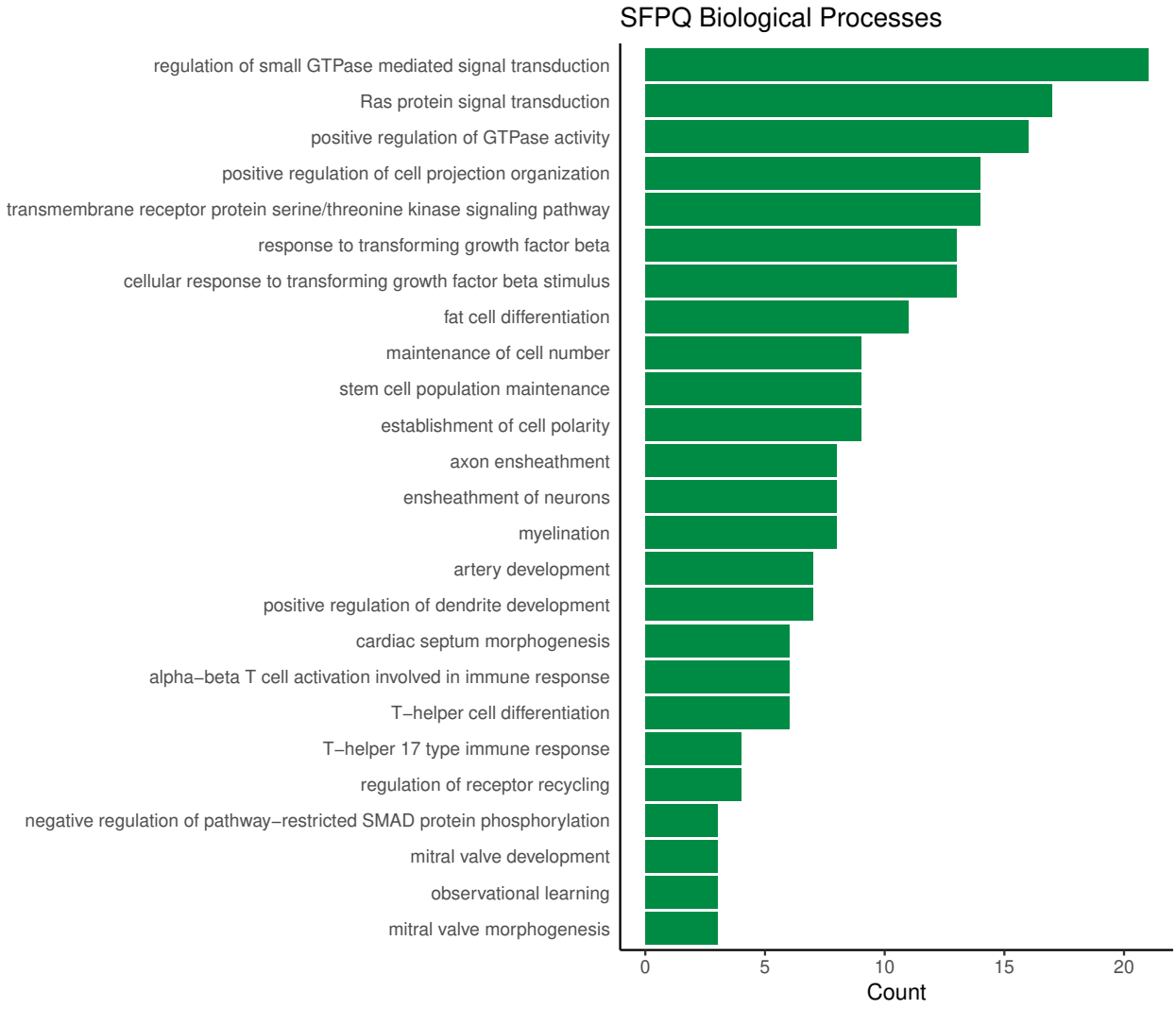


Figure 15

Top 25 biological processes (ranked by FDR) associated with interactors at the genetic, physical, and transcriptional level.

To explore the biological processes associated with fully interactors of SFPQ, GO analysis was conducted on the 245 common interactors (Figure 16). GO analysis uncovered many ontologies

encompassing neurological processes, including neuron/axon ensheathment, myelination and dendrite development. Additionally, a notable number of ontologies implicate regulation of processes involving the immune response, namely T-helper cell differentiation and activation.

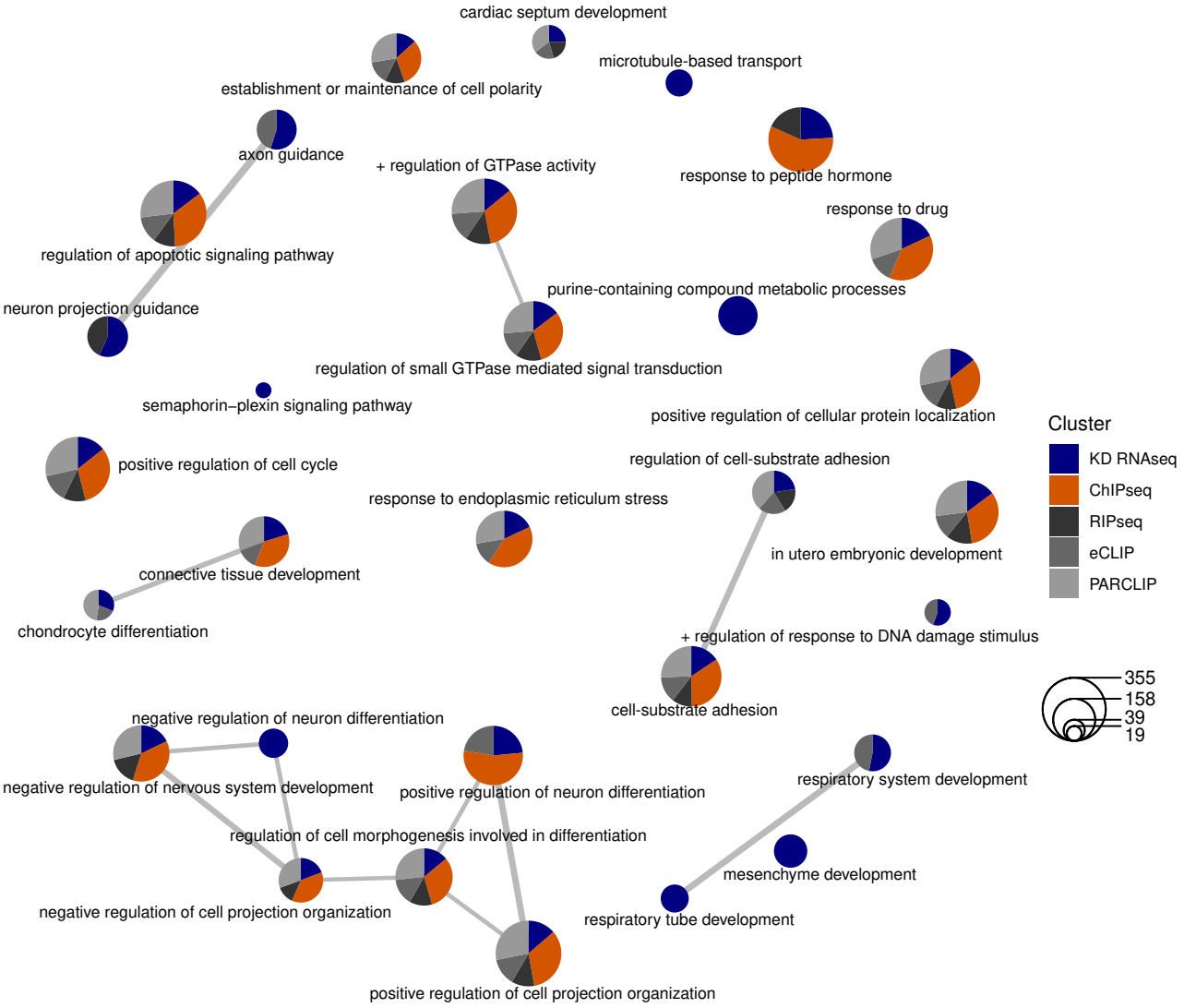


Figure 16

Network map of top ontologies, with pie charts describing which datasets associate with each ontology. Nodes are sized by gene count.

In assessment of shared ontologies across datasets, GO analysis was performed upon fully integrated gene clusters for each mode of analysis. A network plot was constructed for top 30 enriched ontologies (ranked by FDR) across all analyses (Figure 17). We found many biological processes

enriched across all three layers of analysis, including embryonic development, positive regulation of cell cycle, and cell-substrate adhesion. Interestingly, there are several top ontologies which are only enriched in SFPQ-kd and RNA physical binding analysis, for instance response to DNA damage stimulus, respiratory system/cardiac septum development, and neuron projection guidance.

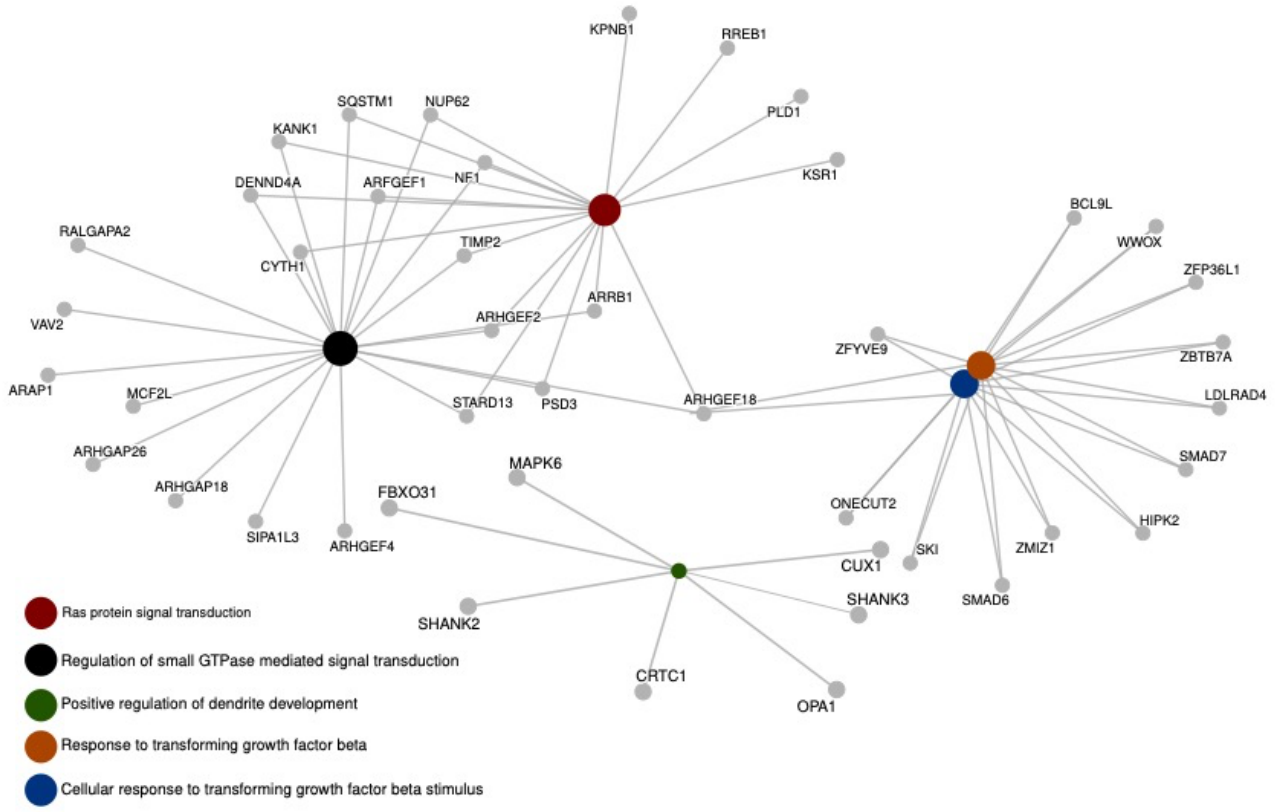


Figure 17

Top 5 ontologies (ranked by FDR) associated with interactors at the genetic, physical and transcriptional level, indicating the genes involved in each ontology. Nodes are sized by gene count.

We interrogated the genes associated with the top five FDR ranked ontologies enriched across our complete analysis (Figure 18). These top five ontologies included Ras protein signal transduction, regulation of small GTPase transduction, Positive regulation of dendrite development, and two related ontologies involving response to transforming growth factor beta. There are many genes which are involved in multiple biological processes, with a great number of genes overlapping between Ras protein signal transduction and small GTPase mediated signal transduction. Protein coding gene, ARHGEF18, is involved in four out of the top five ontologies. As suspected, many associated genes

are shared between the two most related ontologies. Genes involved in regulation of dendrite development are not associated with any of the other top ontologies.

4.0 Discussion

SFPQ is a multifunctional DBHS protein with a large spectrum of interaction partners both through indirect gene regulation, and through physical binding at the RNA and DNA level (Ha et al., 2011; Rosonina et al., 2005). In normal function, SFPQ is involved in a broad range of cellular processes including DNA damage repair, transcriptional regulation, and alternative splicing (de Silva et al., 2019; Ha et al., 2011; Kim et al., 2011; Rosonina et al., 2005).

Aberrant SFPQ is associated with disease aetiology and is known to interact with specific genes such as NEAT1 and MALAT1 (Hassan et al., 2019; Imamura et al., 2014; Ji et al., 2014). Additionally, SFPQ has been studied regarding its role in neurological pathways, including maintenance of neuronal cells through promoting neuronal transcription dendritic transport of RNA and large-scale multilateral analysis revealed SFPQ targets in neuronal cells (Furukawa et al., 2015; Iida et al., 2020; Kanai et al., 2004). However, as far as we are aware the regulatory role of SFPQ has not yet been studied on a large-scale across different cancer types and so we sought to integrate datasets targeting SFPQ across various cell lines in order to investigate the scope of SFPQ interactions in various cancer types.

Thus, we aimed to construct an interaction database deciphering the SFPQ-DNA and -RNA regulatory network, which could then be utilised to further study SFPQ and its regulatory output in disease pathogenesis. We integrated data from SFPQ-kd RNA-seq, RIP-seq, ChIP-seq, eCLIP, and PAR-CLIP analysis across multiple cancer tissues including liver, kidney, prostate, breast, and CML. These datasets contain information on both known and novel interactors of SFPQ and integrative analysis will prove useful for future research, in understanding SFPQ interactions and the mechanisms

regulate by SFPQ function. In turn, by understanding the biological processes associated with SFPQ interactors, we can explore potential mechanisms by which SFPQ misregulation promotes carcinogenesis.

The SFPQ regulatory network and its implications in disease pathogenesis, according to our analyses combined with existing reports, will be discussed in detail below.

4.1 Genetic interactors of SFPQ

First, we defined genetic interactors of SFPQ as genes by which expression levels are modified by the expression of SFPQ, this definition allows us to capture both direct and indirect interactors of SFPQ in order to explore the SFPQ regulatory network on a large scale. We incorporated a meta-analysis approach, consisting of P-value combination to identify transcripts with significantly altered expression levels in RNA-seq datasets where SFPQ was depleted in kidney, liver, CML and colon cancer cell lines. Through incorporating meta-analysis into our study, we enhance statistical power through increasing sample size from multiple cohorts.

Prior to integration, each dataset contained over 1000 genes which were only DE in that dataset (Figure 4). It is possible that many SFPQ interactions are cell line- or dataset-specific. However, following DGE analysis, the HepG2 dataset published by Stagsted et al. (2021) surfaced 1141 more significant DE genes than the HepG2 dataset published at ENCODE (ENCSR782MXN), suggesting that alongside potential cell line specificities, there are contrasts in DEG between datasets (Table 3). We addressed contrasts between datasets by incorporating integrative analysis methods, revealing genetic interactors across multiple cell lines and datasets (Figure 2).

Using GO analysis, we discovered that genetic interactors of SFPQ are involved in a large range of biological processes and it is clear that SFPQ is involved in the regulation of a wide array of functions (Figure 6). The breadth of processes which SFPQ regulates is unsurprising, as SFPQ and other

members of the DBHS family are frequently described as multifunctional proteins (Rhee et al., 2017; Shav-Tal & Zipori, 2002). Interestingly, many of these ontologies are already evidenced to be associated with SFPQ including sister chromatid segregation, cellular response to DNA damage stimulus and regulation of neuron projection development, which were all downregulated in absence of SFPQ (Ha et al., 2011; Lowery et al., 2007; Rajesh et al., 2011) (Figure 6).

Aberrant expression of SFPQ is often associated with disease aetiology and GO analysis of SFPQ-kd data reports processes that are misregulated with impaired SFPQ function (Duhoux et al., 2011; Jiang et al., 2013; Klotz-Noack et al., 2020). For instance, genes associated with extra-cellular matrix organisation and cell-substrate adhesion including the cell surface adhesion protein, ITGA3, are upregulated with loss of SFPQ and are often associated with metastatic activity in cancers (Jiao et al., 2019; Kawataki et al., 2007) (Figure 6).

Loss of SFPQ also resulted in downregulation of genes involved in G2/M phase transition of the cell cycle, inducing cell cycle arrest (Figure 6). Induction of G2 phase cell cycle arrest in the absence of SFPQ has been reported by Rajesh et al. (2011) with relation to the vital role of SFPQ in homologous repair of double-stranded DNA breaks, which in disturbance, results in progression of cells through S phase and accumulation at the G2 phase of the cell cycle (Rajesh et al., 2011). Furthermore, genome stability is hindered through downregulation of genes associated with sister chromatid separation, where inability for chromatid cohesion impedes processes in repair of radiation-induced double-stranded breaks (Ström et al., 2007). Inversely, SFPQ depletion also upregulated genes involved in positive regulation of cell cycle progression, including TP53 and FOXA1 (Figure 6; See Supplementary: S7). Interestingly, we also observed that loss of SFPQ downregulated RAD51ADP1 and RAD51 paralog, RAD51B. RAD51B depletion is reported to increase sensitivity to DNA damage, whilst RAD51 paralogues are also known to trigger the p53 checkpoint response (Kuznetsov et al., 2009; P. S. Lee et al., 2014; Rajesh et al., 2011) (Figure 6; See Supplementary: S7). These

results combined with our observations suggests that our genetic interaction network captured cognate genetic interactions of SFPQ.

Via DGE analysis of SFPQ-kd RNA-seq, we observed that depleting SFPQ altered expression levels of protein coding genes more frequently than other biotypes (Figure 5). A possible explanation would be that many of the protein coding genes are being co-regulated via SFPQ-lncRNA interactions, which represent a much smaller proportion of the SFPQ genetic interactors. LncRNA manipulate gene expression through a range of processes including chromatin modification and transcriptional regulation (Colognori et al., 2019; Dueva et al., 2019; Rinn et al., 2007; Stojic et al., 2016). Similarly, SFPQ also genetically interacts with a small proportion of pseudogenes. Pseudogenes are reported to regulate gene expression through ceRNA networks and are often processed as siRNA (An et al., 2017; Chan & Chang, 2014).

However, the large proportion of protein coding interactors in comparison with non-coding genes is likely due to experimental methods in RNA-seq, for instance, SFPQ-kd RNA-seq from ENCODE (ENCSR782MXN and ENCSR535YPK) prepared RNA-seq libraries through enrichment of polyadenylated mRNA transcripts, and therefore the data only contains transcripts which are polyadenylated. Many non-coding RNA transcripts, including miRNAs, snoRNAs, and some lncRNAs, are not polyadenylated and have alternative 3' end processing pathways (Wilusz et al., 2012). Although it would require greater sequencing depth in experiments, it would be useful to generate more SFPQ-kd RNA-seq data by which RNA-seq libraries have been prepared through physical rRNA depletion in order to include more information on non-polyadenylated transcripts.

4.2 SFPQ physical interactors at RNA level

Although identifying genetic interactors provides a wide investigation of genes affected by SFPQ function, it does not tell us whether the genetic networks are indirect or through direct physical

interactions. Thus, we analysed RIP-seq analysis datasets targeting SFPQ into our database, in order to uncover genes that physically interact with SFPQ at RNA level. As expected, RIP-seq datasets showed variation of enriched genes between cell lines and datasets.

After full integration of RIP-seq datasets we identified 3331 genes that form RNA complexes with SFPQ in at least two of the three datasets. 900 of the 3331 enriched genes overlap with SFPQ genetic interactors, confirming 900 genes to be directly regulated by SFPQ through physical interactions. Of the 900 physical interactors, (interferon regulatory factor 1) IRF1 is shown to physically enrich SFPQ in RIP-seq and was upregulated in SFPQ depletion (See Supplementary: S3, S4). A recent study supports this observation as SFPQ was reported responsible for the alternative splicing of IRF1, producing the isoform IRF1Δ7 in T helper 1 cells and impeding antitumour response (Bernard et al., 2021). Another known interactor of SFPQ is NEAT1, NEAT1 was enriched through RIP-seq and downregulated with depleted levels of SFPQ.

However, this does not necessarily mean that there are only 900 genetic interactors directly regulated by SFPQ. Partial overlap between physical and genetic interactors could be explained by certain transcript expression levels only being regulated in specific cancer cell lines, where there are dissimilarities in observed cell lines between SFPQ-kd RNA-seq and RIP-seq analysis. SFPQ-kd analysis consisted of liver, kidney, erythroleukemia and colorectal cancer cell lines, whereas the RIP-seq data available consisted of breast cancer and prostate cancer cell lines. Hopefully, with the spiking interest of SFPQ, more RIP-seq data targeting SFPQ will become publicly available, by which further integration of datasets will make physical analysis more robust. Additionally, certain transcripts may interact with SFPQ solely at DNA level, this will be explained in later sub-chapters.

We found that at RNA level, SFPQ binding favours intronic regions of target genes in both eCLIP and PAR-CLIP analysis (Figure 8). Through integration of RNA-seq and CLIP-seq data, SFPQ has already been reported to bind long intron containing genes (Iida et al., 2020). It has been evidenced

that SFPQ binds long introns in RNA processing through intron retention, where abnormal intron retention of SFPQ has been associated with motor neuron differentiation in ALS patients (Luisier *et al.*, 2018). RNA interactors showed some binding across RNA regions overlapping 3' UTR of transcripts, SFPQ binding of motifs within 3' UTR on RNA transcripts has been reported potentially necessary for intracellular trafficking, including neurotrophin-dependent axonal localization (Cosker et al., 2016) (Figure 8). Alternatively, miRNAs regulate gene expression through binding to 3' UTR of mRNA, therefore it is possible that in 3' UTR binding of SFPQ, miRNAs are sequestered from binding target mRNA (L. Wu et al., 2006).

We integrated RIP-seq data with nucleotide resolution immunoprecipitation datasets (eCLIP and PAR-CLIP) and found significant but partial overlap between the list of genes enriching SFPQ in RIP-seq, eCLIP, and PAR-CLIP datasets, where only partial overlap between results of different experimental techniques was to be expected. We performed GO analysis on clusters of genes enriched in all three analysis techniques in order to assess the biological processes regulated by interactions between SFPQ and genes at RNA level (Figure 9).

At first inspection of GO analysis, there are many top ontology terms which are also associated with genetic interactors of SFPQ, such as response to peptide hormone, myeloid cell differentiation, and histone modification (Figure 9). SFPQ physical binding of genes associated with cell morphogenesis has been shown to regulate neuronal cell projection organisation and axon development (Thomas-Jinu et al., 2017).

SFPQ physically binds RNA transcripts involved in processes associated with regulation of a range of signaling pathways, including the Fc receptor signaling pathway, the Wnt pathway, and the Transforming growth factor (TGF-β). SFPQ physically interacts with 28 genes involved in the Fc receptor signaling pathway, the response of a cell-surface receptor of inflammatory immune response (Vogelpoel et al., 2015). Interestingly, SFPQ binds genes involved in the Wnt pathway, a cascade

closely associated with carcinogenic processes in melanoma and leukaemia, as well as breast and gastrointestinal cancers (Biechele et al., 2012; Christie et al., 2013; Khramtsov et al., 2010; Lane et al., 2011; Xu et al., 2015). Additionally, SFPQ binds 34 RNA transcripts involved in response to Transforming growth factor (TGF)-β, an important response pathway in immune homeostasis, but mutations that eliminate the TGF-β pathway aid tumour progression through creating an immune suppressive tumour microenvironment. In SFPQ-kd RNA-seq analysis, regulation of the TGF-β pathway was upregulated, suggesting that SFPQ may potentially dampen the immune response regulated by the TGF-β signaling pathway and enhance tumour progression in cases of high SFPQ expression levels (Figure 6).

Interestingly, genes associated with histone modification and covalent chromatin modification are amongst the top ranked physical interactors at RNA level, suggesting the importance of SFPQ in regulating post-translation modifications of RNA target transcripts (Figure 9).

4.3 SFPQ physical interactors at DNA level

Consistent with RNA physical interactors, SFPQ most frequently binds regions of DNA which overlap introns, across analysis of liver, CML, and prostate cancer cell lines via ChIP-seq (Figure 10). SFPQ physically binds regions of DNA proximal to promoters of genes, with implications of potentially high levels of transcriptional regulation through DNA binding (Figure 10). SFPQ has been studied regarding its role in transcriptional regulation. In some instances, SFPQ acts as a coactivator through promoting the binding of RNA polymerase II to promoter-bound transcriptional activators (Emili et al., 2002; Rosonina et al., 2005). Inversely, SFPQ is also involved in transcriptional repression through interactions with DNA-binding domains of Nuclear hormone receptors (NHR) (Dong et al., 2005, 2007). Through binding multiple biotype regions per gene at DNA level, including intronic, promoter, exonic and 5’ UTR regions, it is possible that SFPQ does not have one specific

mode of regulation through genes and thereby regulates many genes through various mechanisms of regulation (Figure 10).

We performed epigenetic analysis on SFPQ by comparing ChIP-seq peaks in HepG2 with regions of popular epigenetic histone marks, namely H3K4me3, H3K4me1, H3K27ac, and H3K27me3. Of active promoters, enhancers, poised enhancers, and repressors, SFPQ binds DNA regions of transcriptional enhancers and repression most frequently (Figure 14). SFPQ and NONO directly bind to an enhancer region which upregulates transcription of RPL18 (Roepcke et al., 2011). SFPQ has been evidenced in transcriptional repression, where SFPQ is reported to induce tumour suppression through binding to the promoter of proto-oncogene, *Rab23* (C.-F. Wu et al., 2013). It is implicated that SFPQ induces transcriptional repression in this way in the context of many other DNA transcripts, given the large scale of SFPQ-DNA interactors with binding at 'repressive' chromatin marks. These observations, combined with existing evidence, suggest a broad role for SFPQ in transcriptional regulation. As far as we are aware, ChIP-seq data for H3K27me1 in HepG2 is not available at ENCODE, otherwise we could employ additional combinatorial epigenomic conditions and assess the extent at which SFPQ acts as a transcriptional activator.

GO analysis was performed on SFPQ targets enriched at both genetic and DNA level but not at RNA level, revealing only four biological processes associated with genes bound by SFPQ at DNA level that were upregulated in integrated analysis of SFPQ-kd datasets (Figure 12). These processes include connective tissue development, phospho- and glycerophospholipid metabolic processes, and extrinsic apoptotic signaling pathways (Figure 12). Enrichment of extrinsic apoptotic signaling pathway is consistent with observations in melanoma cell, whereby depletion of SPFQ resulted in significant induction of apoptosis (Bi et al., 2021). Similarly, depletion of SFPQ increased apoptosis in $BRAF^{V600E}$ mutated colorectal cancer cell (Klotz-Noack et al., 2020). At transcriptional level, many SFPQ-DNA targets are associated with processes involving endoplasmic reticulum (ER) and were downregulated in SFPQ-d (Figure 12). Frequent ontologies relate to response to ER stress, with

implications to the role of SFPQ in transcriptional regulation of protein folding, as well as transport to the cytosol.

Amidst integration of SFPQ interactors at multi-omics level, we integrated results generated only from HepG2 cells in pursuit to observe the SFPQ regulatory network in a single cell line across genetic, RNA and DNA interactions. Firstly, there are clear overlaps between pathway-associated interactors regulated at both transcriptional and post-transcriptional level. In depletion of SFPQ, interactors associated with regulation of metabolic processes were upregulated at both transcriptional and post-transcriptional level. Hosokawa *et al*. (2019) demonstrated that loss of SFPQ downregulated the expression of a cluster of metabolic pathways (Hosokawa et al., 2019). Many of the GO results obtained through analysis of HepG2 are consistent with our previous observations, including post-transcriptional interaction with genes associated with response to insulin and transcriptional regulation of neuronal function, cell morphogenesis and protein folding. It could be argued that high presence of HepG2 datasets throughout our analyses could be creating bias, however our observations are consistent with existing SFPQ study.

4.4 Complete integration of the SFPQ regulatory interactome at genetic, physical and transcriptional level

Our aim was to integrate the different omics datasets to construct a compressive and robust analysis of the SFPQ interactome. To this end we integrated clusters of genetic, physical and transcriptional interactors in order to analyse the biological processes associated with genes overlapping all observed SFPQ omics datasets.

We observed SFPQ interactors associated with the top five ontologies (ranked by FDR) overlapping at each level of analysis based on FDR. Strikingly, the top ranked term interrogated genes involved in Ras protein signal transduction (Figure 18). Consistently through all levels of analysis, SFPQ

enriches 18 genes associated with this pathway, including transcription factor, Ras-responsive element-binding protein 1 (RREB1) (See Supplementary: S7). It is reported that RREB1 is a downstream effector of the Mitogen-activated protein kinase (MAPK) signaling pathway (Thiagalingam et al., 1996). RREB1 has been studied with involvement in many different pathways in the development of numerous cancers. RREB1 is overexpressed in prostate cancer, where high levels of RREB1 decrease zinc levels, providing a microenvironment for enhanced growth and survival of prostate cancer cells (Zou et al., 2011).

Additionally, through acting as a downstream effector of the MAPK signaling pathway, RREB1 is involved in melanoma development through inhibition of several tumour suppressors, including p53, p16INK4a and miR-143/145 (Turri-Zanoni et al., 2013). Many of the SFPQ interactors involved in Ras signal transduction are also associated with regulation small GTPase mediated signal transduction. Small GTPase mediated signal transduction is a broad term and is present in many aspects of cell biology; it is likely that a large number of cellular processes regulated by SFPQ involve regulation of small GTPase signal transduction (Figure 18).

Our analyses show recurring enrichment of genes involved in neuronal developmental processes, including axon ensheathment and myelination (Figure 16). SFPQ is described with association with neurological function including promoting axon viability and neuronal development, therefore it is not surprising to discover that SFPQ heavily interacts with genes associated with neurological processes consistently across multiple levels of analysis, as well as through multiple cell lines (Cosker et al., 2016; Iida et al., 2020; Ishigaki et al., 2017; Luisier et al., 2018; Thomas-Jinu et al., 2017).

Interestingly, many integrated SFPQ interactors are associated with immune response pathways, through differentiation and activation of T-cells (Figure 4B). Each ontology containing T-cell associated process appear based upon SFPQ interactions with the gene, RORA, implying the

possibility that in this case SFPQ activity in immune response involves regulation of CD4+ T-cell response (Haim-Vilmovsky et al., 2020).

4.5 idbSFPQ R interface

We created a flexible R interface for querying this database, herein called interaction database of SFPQ (idbSFPQ), available at the repository github.com/jcogan1/idbSFPQ. idbSFPQ contains results from genetic, physical, and transcriptional interaction analysis of SFPQ, as well as GO analysis at each level. Throughout this study, we programmed functions with consideration that analysis could be expanded to discovering regulatory networks for other genes. In compiling idbSFPQ functions into an R package, we lay the foundations for an analysis workflow to investigate regulatory networks for other genes of interest.

The prospects of this database are broad, but opportunistic. The results are currently stored in a Github repository, however there are aims to build a relational database using SQL to allow ease of online datamining for users unfamiliar with R. It is in our aims to build an interactive web app using R shiny. Shiny allows for the transformation of existing R code into a live application which can provide output based on user input with a minimum work.

4.6 Limitations of this study

We focused on commonalities amongst datasets in integration of RNA-seq data, in order to maintain statistical accuracy in comparison of analyses generated through different platforms. It is important to highlight that our genetic interactions network could have expelled important genes in selection of overlapping DEG. When selecting genetic interactors based on common genes between datasets, it is often the case that important genes can be lost due to genes not being shared between datasets (Bobak

et al., 2020). Our relaxed FDR threshold (0.1) reflects this problem, where we compensate for type II errors by implementing a tolerance for a large amount of type I errors.

Notably, our database has been constructed solely using cell line data, and therefore it is possible that SFPQ genetic interactions may differ when analysed *in vivo*. To address this in future study, we could utilise experiments which validate novel SFPQ interactions *in vivo*, such as through RIP targeting SPFQ, followed by RNA extraction and RT-PCR analysis, similarly described by Wu et al. (2013) (C.-F. Wu et al., 2013).

Additionally, we explore the changes in gene expression introduced by perturbed levels of SFPQ and although this may give indication of SFPQ regulatory interactions, changes in gene expression are not always indicative of consequences in cellular activity. For instance, a change in the mRNA expression of a protein coding gene does not necessarily determine a change in levels of the phosphorylated protein, where phosphorylation indicates an active form of the protein. In further study of protein coding genetic interactors of SFPQ, this can be challenged through western blotting using antibodies against phosphorylated forms of the protein of interest, thus giving a more detailed insight to the downstream consequences of SFPQ regulation.

By employing a meta-analysis approach to SFPQ investigation, our study directs focus towards commonalities of SFPQ interactions across multiple cancer cell lines and through multiple modes of analysis. Therefore, it is possible that our study fails to highlight specific perturbations of SFPQ in individual cancer types. Moreover, the broad spectrum of SPFQ has been stressed throughout this report, and by limiting our attention to only enriched biological functions which overlap across several datasets, we potentially overlook important functions which are driving progression in individual cancers. Our database contains RNA-seq DGE analysis results prior to integration adj therefore data for individual cell lines is mineable, however incorporation of other analyses (ChIP-seq, PARCLIP, eCLIP) is reliant upon enhanced data availability from future studies.

Although we map a large database of SFPQ interactors, many SFPQ interactions occur through dynamic structural changes formed by heterodimerisation with other RBP, such as NONO (de Silva et al., 2019; Huang et al., 2018; Salton et al., 2010). Although we are able to identify genetic interactors of SFPQ heterodimers in SFPQ-kd RNA-seq data, this isn't the case for genes that aren't present in genetic analysis and it is likely that the ultimate scale of the SFPQ regulatory network is even greater than our database suggests, through obligate interactions with other members of the DBHS family.

4.7 Concluding remarks

In summary, we constructed interaction database, idbSFPQ, depicting the regulatory interactome of SFPQ at genetic, RNA, and DNA level, taking advantage of publicly available experimental datasets targeting SFPQ in multiple techniques, across multiple cell lines. We also show extensive analysis of the biological processes associated with transcripts regulated or enriched by SFPQ, emphasising that SFPQ is involved in a wide range of biological pathways. Novel SFPQ interactors discovered by integrative analysis offer the opportunity for further study of SFPQ and the mode of SFPQ regulation upon specific transcripts of interest.

Through implementation of meta-analysis, we identified a range of SFPQ interactors at RNA and DNA level in a range of cancer cell lines. Although the cost of high throughput sequencing has improved greatly, it remains an expensive technique and therefore individual studies often studies are often burdened by a low number of replicates, thus by combining multiple studies we were able to increase the number of samples observing the knockdown of SFPQ. Additionally, by analysing each RNA-seq dataset with consistent normalisation and statistical testing, we enable the comparison of individual datasets to observe effects which may be specific to a particular cell line or dataset. Ultimately, our database offers an indication of genes regulated by SFPQ, many of which will be interesting for further study regarding their functional relationship with SFPQ in cancer.

Integrative analysis was limited by the amount of unique cell line coverage and although we offer an exhaustive exploration of the SFPQ regulome, we are restricted by data availability. However, Interest in SFPQ function has propelled over recent years and as new datasets are published, our analysis pipeline will allow us to integrate newly available datasets, strengthening the SFPQ interaction database.

SFPQ is an extensively complex RBP with potential as a prognostic and diagnostic biomarker in cancer progression, displaying a colossal range of interactors and enriched biological pathways. It will be interesting to develop a greater understanding of SFPQ regulatory mechanisms via further study focusing specific target transcripts.

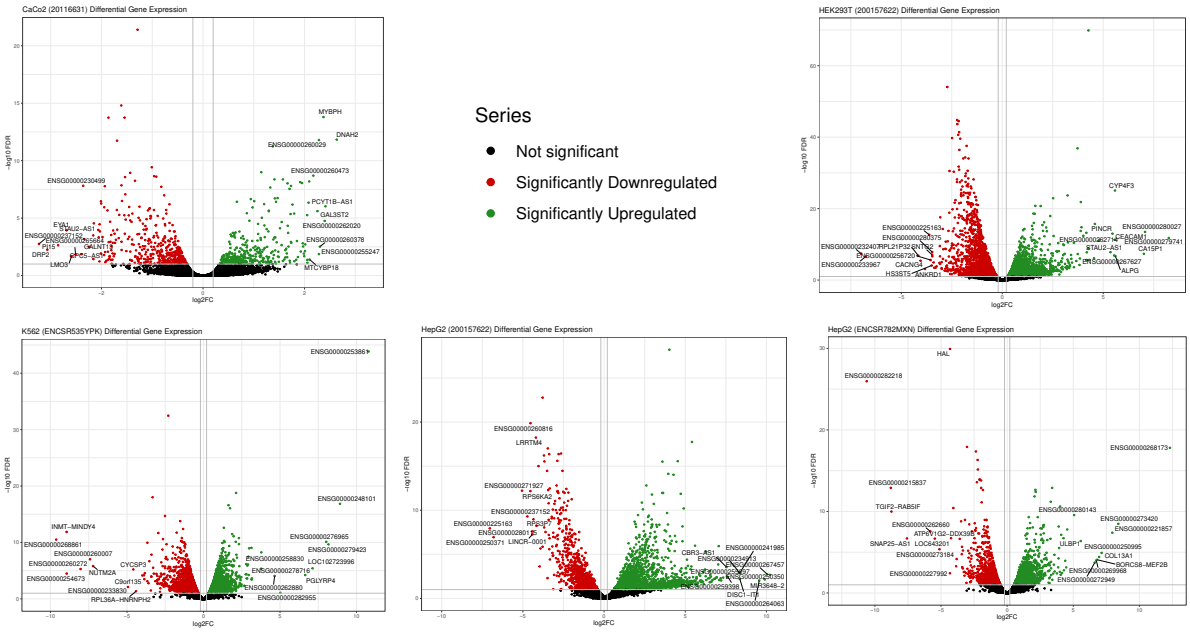
Supplementary

S1

Results table from data mining analysis. Publicly available datasets were reviewed and inclusion in this study was dependent on technique and data types available. Data used in this study are coloured green.

UID	Technique	Context	Used (T/F)
200157622	SFPQ-kd RNA-seq	HepG2, HEK293T	T
200149371	SFPQ-kd RNA-seq	CaCo2	T
200131539	SFPQ-kd Microarray	MDAH	F
200095057	TREND-seq	BE(2)-C	F
200113349	PAR-CLIP	HeLa, U2OS	T
200114394	RIP-seq	22Rv1	T
200133423	RIP-seq	MCF-7	T
200100239	SFPQ-kd Microarray	LNCaP	F
200094243	RIP-seq	22Rv1, LNCaP, LTAD	T
ENCSR782MXN	SFPQ-kd RNA-seq	HepG2	T
ENCSR535YPK	SFPQ-kd RNA-seq	K562	T
ENCSR757HBB	ChIP-seq	HepG2	T
ENCSR258SXK	ChIP-seq	HepG2	T
ENCSR647PJW	ChIP-seq	K562	T
ENCSR965DLL	eCLIP	HepG2	T
ENCSR951YCV	RNA bind-n-seq	Cell-free sample	F

S2



Volcano plots of individual SFPQ-kd RNA-seq DGE results. Vertical lines represent logFC threshold (0.2) and horizontal lines represent FDR threshold (FDR < 0.1). Significantly upregulated genes are coloured green. Significantly downregulated genes are coloured red. Non-significant genes coloured black. Top 10 upregulated and downregulated genes are annotated in each dataset.

S3

Comma separated variable (CSV) files containing DGE analysis results of individual SFPQ-kd RNA-seq analysis. 'gene_id' contains Ensembl IDs of genes. 'logFC' defines logFC where genes with positive or negative logFC values are upregulated or downregulated, respectively. Datasets also include logCPM, P-Value, and FDR

Joseph_Cogan_Supplementary/S3

/ENCSR782MXN.csv ; / GSE157622_hepg2.csv ; /GSE157622_hek293t.csv ; /GSE149370.csv ; /ENCSR535YPK.csv

S4

CSV file containing list of genetic interactors identified by Ensembl ID ('gene_id'). 'genetic' column denotes the regulatory status in SFPQ-kd, where '1_sig' defines significantly upregulated genes, '-

1_sig' defines significantly downregulated genes, '2_sig' defines genes with significantly altered expression but in an uncategorised direction, and any data containing 'notsig' is not significant.

Joseph_Cogan_Supplementary/S4.csv

S5

CSV file containing list of RIP-seq interactors identified by Ensembl ID ('gene_id'). 'physical' column denotes enrichment status, where '1_sig' defines significantly enriched genes.

Joseph_Cogan_Supplementary/S5

S6

CSV file containing annotated and integrated peak datasets for eCLIP, PAR-CLIP and ChIPseq. Dataset contains chromosome name (chr), peak start and end coordinates (start, end), nucleotide width of peak (width), dataset label, score, strand, signal value, log10 P-value and log10 q-value (-1 where unavailable), peak value, gene start, gene end, gene length, gene strand, distance to transcription start site (distanceToTSS), entrez id, gene symbol, Ensembl gene ID (gene_id), Ensembl transcript ID (transcriptID), biotype annotation, and gene name.

PAR-CLIP also contains PARalyzer score.

Joseph_Cogan_Supplementary/S6

/eclip.csv ; /parclip.csv ; /chip.csv

S7

CSV files containing GO analysis data. Datasets contain cluster, GO ID, ontology description, gene ratio, Bg ratio, P-value, FDR (p.adjust), q value and associated genes (geneID).

Joseph_Cogan_Supplementary/S7

/GO_genetic.csv ; /GO_physical.csv ; /GO_transcriptional.csv ; /GO_fully_integrated ; /GO_hepg2.csv

References

An, Y., Furber, K. L., & Ji, S. (2017). Pseudogenes regulate parental gene expression via ceRNA network. *Journal of Cellular and Molecular Medicine*, *21*(1), 185–192. https://doi.org/10.1111/jcmm.12952

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., Snow, C., Fonseca, N. A., Petryszak, R., Papatheodorou, I., Sarkans, U., & Brazma, A. (2019). ArrayExpress update – from bulk to single-cell expression data. *Nucleic Acids Research*, *47*(Database issue), D711–D715. https://doi.org/10.1093/nar/gky964

Bae, S., & Lesch, B. J. (2020). H3K4me1 Distribution Predicts Transcription State and Poising at Promoters. *Frontiers in Cell and Developmental Biology*, *8*, 289. https://doi.org/10.3389/fcell.2020.00289

Bajc Česnik, A., Darovic, S., Prpar Mihevc, S., Štalekar, M., Malnar, M., Motaln, H., Lee, Y.-B., Mazej, J., Pohleven, J., Grosch, M., Modic, M., Fonovič, M., Turk, B., Drukker, M., Shaw, C. E., & Rogelj, B. (2019). Nuclear RNA foci from C9ORF72 expansion mutation form paraspeckle-like bodies. *Journal of Cell Science*, *132*(5), jcs224303. https://doi.org/10.1242/jcs.224303

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets--update. *Nucleic Acids Research*, *41*(Database issue), D991-995. https://doi.org/10.1093/nar/gks1193

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., & Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, *129*(4), 823–837. https://doi.org/10.1016/j.cell.2007.05.009

Bell, R., Barraclough, R., & Vasieva, O. (2017). Gene Expression Meta-Analysis of Potential Metastatic Breast Cancer Markers. *Current Molecular Medicine*, *17*(3), 200–210. https://doi.org/10.2174/1566524017666170807144946

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Bernard, A., Hibos, C., Richard, C., Viltard, E., Chevrier, S., Lemoine, S., Melin, J., Humblin, E., Mary, R., Accogli, T., Chalmin, F., Bruchard, M., Peixoto, P., Hervouet, E., Apetoh, L., Ghiringhelli, F., Végran, F., & Boidot, R. (2021). The Tumor Microenvironment Impairs Th1 IFNγ Secretion through Alternative Splicing Modifications of Irf1 Pre-mRNA. *Cancer Immunology Research*, *9*(3), 324–336. https://doi.org/10.1158/2326-6066.CIR-19-0679

Bi, O., Anene, C. A., Nsengimana, J., Shelton, M., Roberts, W., Newton-Bishop, J., & Boyne, J. R. (2021). SFPQ promotes an oncogenic transcriptomic state in melanoma. *Oncogene*, *40*(33), 5192–5203. https://doi.org/10.1038/s41388-021-01912-4

Biechele, T. L., Kulikauskas, R. M., Toroni, R. A., Lucero, O. M., Swift, R. D., James, R. G., Robin, N. C., Dawson, D. W., Moon, R. T., & Chien, A. J. (2012). Wnt/β-Catenin Signaling and AXIN1 Regulate Apoptosis Triggered by Inhibition of the Mutant Kinase BRAFV600E in Human Melanoma. *Science Signaling*, *5*(206), ra3–ra3. https://doi.org/10.1126/scisignal.2002274

Bladen, C. L., Udayakumar, D., Takeda, Y., & Dynan, W. S. (2005). Identification of the polypyrimidine tract binding protein-associated splicing factor.p54(nrb) complex as a candidate DNA double-strand break rejoining factor. *The Journal of Biological Chemistry*, *280*(7), 5205–5210. https://doi.org/10.1074/jbc.M412758200

Blecher-Gonen, R., Barnett-Itzhaki, Z., Jaitin, D., Amann-Zalcenstein, D., Lara-Astiaso, D., & Amit, I. (2013). High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo protein-DNA interactions and epigenomic states. *Nature Protocols*, *8*(3), 539–554. https://doi.org/10.1038/nprot.2013.023

Bobak, C. A., McDonnell, L., Nemesure, M. D., Lin, J., & Hill, J. E. (2020). Assessment of Imputation Methods for Missing Gene Expression Data in Meta-Analysis of Distinct Cohorts of Tuberculosis Patients. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, *25*, 307–318.

Bobowicz, M., Skrzypski, M., Czapiewski, P., Marczyk, M., Maciejewska, A., Jankowski, M., Szulgo-Paczkowska, A., Zegarski, W., Pawłowski, R., Polańska, J., Biernat, W., Jaśkiewicz, J., & Jassem, J. (2016). Prognostic value of 5-microRNA based signature in T2-T3N0 colon cancer. *Clinical & Experimental Metastasis*, *33*(8), 765–773. https://doi.org/10.1007/s10585-016-9810-1

Breitling, R., & Herzyk, P. (2005). Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *Journal of Bioinformatics and Computational Biology*, *3*(5), 1171–1189. https://doi.org/10.1142/s0219720005001442

Cardinale, S., Cisterna, B., Bonetti, P., Aringhieri, C., Biggiogera, M., & Barabino, S. M. L. (2007). Subnuclear Localization and Dynamics of the Pre-mRNA 3′ End Processing Factor Mammalian Cleavage Factor I 68-kDa Subunit. *Molecular Biology of the Cell*, *18*(4), 1282–1292. https://doi.org/10.1091/mbc.E06-09-0846

Chan, W.-L., & Chang, J.-G. (2014). Pseudogene-derived endogenous siRNAs and their function. *Methods in Molecular Biology (Clifton, N.J.)*, *1167*, 227–239. https://doi.org/10.1007/978-1-4939-0835-6_15

Chen, L.-L., & Carmichael, G. G. (2009). Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: Functional role of a nuclear noncoding RNA. *Molecular Cell*, *35*(4), 467–478. https://doi.org/10.1016/j.molcel.2009.06.027

Chen, R., Khatri, P., Mazur, P. K., Polin, M., Zheng, Y., Vaka, D., Hoang, C. D., Shrager, J., Xu, Y., Vicent, S., Butte, A. J., & Sweet-Cordero, E. A. (2014). A Meta-analysis of Lung Cancer Gene Expression Identifies PTK7 as a Survival Gene in Lung Adenocarcinoma. *Cancer Research*, *74*(10), 2892–2902. https://doi.org/10.1158/0008-5472.CAN-13-2775

Chen, Z., Wu, H., Zhang, Z., Li, G., & Liu, B. (2019). LINC00511 accelerated the process of gastric cancer by targeting miR-625-5p/NFIX axis. *Cancer Cell International*, *19*, 351. https://doi.org/10.1186/s12935-019-1070-0

Cho, S., Moon, H., Loh, T. J., Oh, H. K., Williams, D. R., Liao, D. J., Zhou, J., Green, M. R., Zheng, X., & Shen, H. (2014). PSF contacts exon 7 of SMN2 pre-mRNA to promote exon 7 inclusion. *Biochimica et Biophysica Acta*, *1839*(6), 517–525. https://doi.org/10.1016/j.bbagrm.2014.03.003

Christie, M., Jorissen, R. N., Mouradov, D., Sakthianandeswaren, A., Li, S., Day, F., Tsui, C., Lipton, L., Desai, J., Jones, I. T., McLaughlin, S., Ward, R. L., Hawkins, N. J., Ruszkiewicz, A. R., Moore, J., Burgess, A. W., Busam, D., Zhao, Q., Strausberg, R. L., … Sieber, O. M. (2013). Different APC genotypes in proximal and distal sporadic colorectal cancers suggest distinct WNT/β-catenin signalling thresholds for tumourigenesis. *Oncogene*, *32*(39), 4675–4682. https://doi.org/10.1038/onc.2012.486

Clemson, C. M., Hutchinson, J. N., Sara, S. A., Ensminger, A. W., Fox, A. H., Chess, A., & Lawrence, J. B. (2009). An Architectural Role for a Nuclear Non-coding RNA: NEAT1 RNA is Essential for the Structure of Paraspeckles. *Molecular Cell*, *33*(6), 717–726. https://doi.org/10.1016/j.molcel.2009.01.026

Colognori, D., Sunwoo, H., Kriz, A. J., Wang, C.-Y., & Lee, J. T. (2019). Xist Deletional Analysis Reveals an Interdependency between Xist RNA and Polycomb Complexes for Spreading along the Inactive X. *Molecular Cell*, *74*(1), 101-117.e10. https://doi.org/10.1016/j.molcel.2019.01.015

Consortium, T. E. P. (2012). An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature*, *489*(7414), 57. https://doi.org/10.1038/nature11247

Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics*, *33*(18), 2938–2940. https://doi.org/10.1093/bioinformatics/btx364

Corcoran, D. L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R. L., Keene, J. D., & Ohler, U. (2011). PARalyzer: Definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biology*, *12*(8), R79. https://doi.org/10.1186/gb-2011-12-8-r79

Cosker, K. E., Fenstermacher, S. J., Pazyra-Murphy, M. F., Elliott, H. L., & Segal, R. A. (2016). The RNA-binding protein SFPQ orchestrates an RNA regulon to promote axon viability. *Nature Neuroscience*, *19*(5), 690–696. https://doi.org/10.1038/nn.4280

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., & Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(50), 21931–21936. https://doi.org/10.1073/pnas.1016071107

Dai, X., Liu, L., Liang, Z., Guo, K., Xu, S., & Wang, H. (2019). Silencing of lncRNA MALAT1 inhibits cell cycle progression via androgen receptor signaling in prostate cancer cells. *Pathology - Research and Practice*, *215*(4), 712–721. https://doi.org/10.1016/j.prp.2019.01.011

Damayanti, N. P., Budka, J. A., Khella, H. W. Z., Ferris, M. W., Ku, S. Y., Kauffman, E., Wood, A. C., Ahmed, K., Chintala, V. N., Adelaiye-Ogala, R., Elbanna, M., Orillion, A., Chintala, S., Kao, C., Linehan, W. M., Yousef, G. M., Hollenhorst, P. C., & Pili, R. (2018). Therapeutic targeting of TFE3/IRS-1/PI3K/mTOR axis in translocation renal cell carcinoma. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *24*(23), 5977–5989. https://doi.org/10.1158/1078-0432.CCR-18-0269

de Silva, H. C., Lin, M. Z., Phillips, L., Martin, J. L., & Baxter, R. C. (2019). IGFBP-3 interacts with NONO and SFPQ in PARP-dependent DNA damage repair in triple-negative breast cancer. *Cellular and Molecular Life Sciences: CMLS*, *76*(10), 2015–2030. https://doi.org/10.1007/s00018-019-03033-4

Dong, X., Shylnova, O., Challis, J. R. G., & Lye, S. J. (2005). Identification and characterization of the protein-associated splicing factor as a negative co-regulator of the progesterone receptor. *The Journal of Biological Chemistry*, *280*(14), 13329–13340. https://doi.org/10.1074/jbc.M409187200

Dong, X., Sweet, J., Challis, J. R. G., Brown, T., & Lye, S. J. (2007). Transcriptional Activity of Androgen Receptor Is Modulated by Two RNA Splicing Factors, PSF and p54nrb. *Molecular and Cellular Biology*, *27*(13), 4863–4875. https://doi.org/10.1128/MCB.02144-06

Dueva, R., Akopyan, K., Pederiva, C., Trevisan, D., Dhanjal, S., Lindqvist, A., & Farnebo, M. (2019). Neutralization of the Positive Charges on Histone Tails by RNA Promotes an Open Chromatin Structure. *Cell Chemical Biology*, *26*(10), 1436-1449.e5. https://doi.org/10.1016/j.chembiol.2019.08.002

Duhoux, F. P., Auger, N., De Wilde, S., Wittnebel, S., Ameye, G., Bahloula, K., Van den Berg, C., Libouton, J.-M., Saussoy, P., Grand, F. H., Demoulin, J.-B., & Poirel, H. A. (2011). The t(1;9)(p34;q34) fusing ABL1 with SFPQ, a pre-mRNA processing gene, is recurrent in acute lymphoblastic leukemias. *Leukemia Research*, *35*(7), e114-117. https://doi.org/10.1016/j.leukres.2011.02.011

Emili, A., Shales, M., McCracken, S., Xie, W., Tucker, P. W., Kobayashi, R., Blencowe, B. J., & Ingles, C. J. (2002). Splicing and transcription-associated proteins PSF and p54nrb/nonO bind to the RNA polymerase II CTD. *RNA*, *8*(9), 1102–1111.

Fang, D., Yang, H., Lin, J., Teng, Y., Jiang, Y., Chen, J., & Li, Y. (2015). 17β-Estradiol regulates cell proliferation, colony formation, migration, invasion and promotes apoptosis by upregulating miR-9 and thus degrades MALAT-1 in osteosarcoma cell MG-63 in an estrogen

receptor-independent manner. *Biochemical and Biophysical Research Communications*, *457*(4), 500–506. https://doi.org/10.1016/j.bbrc.2014.12.114

Ferguson, D. O., Sekiguchi, J. M., Chang, S., Frank, K. M., Gao, Y., DePinho, R. A., & Alt, F. W. (2000). The nonhomologous end-joining pathway of DNA repair is required for genomic stability and the suppression of translocations. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(12), 6630–6633.

Furukawa, M. T., Sakamoto, H., & Inoue, K. (2015). Interaction and colocalization of HERMES/RBPMS with NonO, PSF, and G3BP1 in neuronal cytoplasmic RNP granules in mouse retinal line cells. *Genes to Cells*, *20*(4), 257–266. https://doi.org/10.1111/gtc.12224

Giordana, M. T., Piccinini, M., Grifoni, S., De Marco, G., Vercellino, M., Magistrello, M., Pellerino, A., Buccinnà, B., Lupino, E., & Rinaudo, M. T. (2010). TDP-43 Redistribution is an Early Event in Sporadic Amyotrophic Lateral Sclerosis. *Brain Pathology*, *20*, 351–360. https://doi.org/10.1111/j.1750-3639.2009.00284.x

Ha, K., Takeda, Y., & Dynan, W. S. (2011). Sequences in PSF/SFPQ mediate radioresistance and recruitment of PSF/SFPQ-containing complexes to DNA damage sites in human cells. *DNA Repair*, *10*(3), 252–259. https://doi.org/10.1016/j.dnarep.2010.11.009

Haim-Vilmovsky, L., Henriksson, J., Walker, J. A., Miao, Z., Natan, E., Kar, G., Clare, S., Barlow, J. L., Charidemou, E., Mamanova, L., Chen, X., Proserpio, V., Pramanik, J., Woodhouse, S., Protasio, A. V., Efremova, M., Griffin, J. L., Berriman, M., Dougan, G., … Teichmann, S. A. (2020). *Rora regulates activated T helper cells during inflammation* (p. 709998). https://doi.org/10.1101/709998

Halazonetis, T. D., Gorgoulis, V. G., & Bartek, J. (2008). An oncogene-induced DNA damage model for cancer development. *Science (New York, N.Y.)*, *319*(5868), 1352–1355. https://doi.org/10.1126/science.1140735

Hall-Pogar, T., Liang, S., Hague, L. K., & Lutz, C. S. (2007). Specific trans-acting proteins interact with auxiliary RNA polyadenylation elements in the COX-2 3′-UTR. *RNA*, *13*(7), 1103–1115. https://doi.org/10.1261/rna.577707

Hassan, N., Zhao, J. T., Glover, A., Robinson, B. G., & Sidhu, S. B. (2019). Reciprocal interplay of miR-497 and MALAT1 promotes tumourigenesis of adrenocortical cancer. *Endocrine-Related Cancer*, *26*(7), 677–688. https://doi.org/10.1530/ERC-19-0036

Hedges, L. V. (1982). Fitting Categorical Models to Effect Sizes from a Series of Experiments. *Journal of Educational Statistics*, *7*(2), 119–137. https://doi.org/10.3102/10769986007002119

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., & Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, *39*(3), 311–318. https://doi.org/10.1038/ng1966

Heyd, F., & Lynch, K. W. (2010). Phosphorylation-Dependent Regulation of PSF by GSK3 controls CD45 Alternative Splicing. *Molecular Cell*, *40*(1), 126–137. https://doi.org/10.1016/j.molcel.2010.09.013

Hirose, T., Virnicchi, G., Tanigawa, A., Naganuma, T., Li, R., Kimura, H., Yokoi, T., Nakagawa, S., Bénard, M., Fox, A. H., & Pierron, G. (2014). NEAT1 long noncoding RNA regulates transcription via protein sequestration within subnuclear bodies. *Molecular Biology of the Cell*, *25*(1), 169–183. https://doi.org/10.1091/mbc.E13-09-0558

Hosokawa, M., Takeuchi, A., Tanihata, J., Iida, K., Takeda, S., & Hagiwara, M. (2019). Loss of RNA-Binding Protein Sfpq Causes Long-Gene Transcriptopathy in Skeletal Muscle and Severe Muscle Mass Reduction with Metabolic Myopathy. *IScience*, *13*, 229–242. https://doi.org/10.1016/j.isci.2019.02.023

Hu, L., Tang, J., Huang, X., Zhang, T., & Feng, X. (2018). Hypoxia exposure upregulates MALAT-1 and regulates the transcriptional activity of PTB-associated splicing factor in A549 lung adenocarcinoma cells. *Oncology Letters*, *16*(1), 294–300. https://doi.org/10.3892/ol.2018.8637

Hu, Z., Dong, L., Li, S., Li, Z., Qiao, Y., Li, Y., Ding, J., Chen, Z., Wu, Y., Wang, Z., Huang, S., Gao, Q., Zhao, Y., & He, X. (2020). Splicing Regulator p54nrb /Non-POU Domain-Containing Octamer-Binding Protein Enhances Carcinogenesis Through Oncogenic Isoform Switch of MYC Box-Dependent Interacting Protein 1 in Hepatocellular Carcinoma. *Hepatology (Baltimore, Md.)*, *72*(2), 548–568. https://doi.org/10.1002/hep.31062

Huang, J., Casas Garcia, G. P., Perugini, M. A., Fox, A. H., Bond, C. S., & Lee, M. (2018). Crystal structure of a SFPQ/PSPC1 heterodimer provides insights into preferential heterodimerization of human DBHS family proteins. *The Journal of Biological Chemistry*, *293*(17), 6593–6602. https://doi.org/10.1074/jbc.RA117.001451

Huang, J., Ringuet, M., Whitten, A. E., Caria, S., Lim, Y. W., Badhan, R., Anggono, V., & Lee, M. (2020). Structural basis of the zinc-induced cytoplasmic aggregation of the RNA-binding protein SFPQ. *Nucleic Acids Research*, *48*(6), 3356–3365. https://doi.org/10.1093/nar/gkaa076

Iida, K., Hagiwara, M., & Takeuchi, A. (2020). Multilateral Bioinformatics Analyses Reveal the Function-Oriented Target Specificities and Recognition of the RNA-Binding Protein SFPQ. *IScience*, *23*(7), 101325. https://doi.org/10.1016/j.isci.2020.101325

Iino, K., Mitobe, Y., Ikeda, K., Takayama, K.-I., Suzuki, T., Kawabata, H., Suzuki, Y., Horie-Inoue, K., & Inoue, S. (2020). RNA-binding protein NONO promotes breast cancer proliferation by post-transcriptional regulation of SKP2 and E2F8. *Cancer Science*, *111*(1), 148–159. https://doi.org/10.1111/cas.14240

Imamura, K., Imamachi, N., Akizuki, G., Kumakura, M., Kawaguchi, A., Nagata, K., Kato, A., Kawaguchi, Y., Sato, H., Yoneda, M., Kai, C., Yada, T., Suzuki, Y., Yamada, T., Ozawa, T.,

Kaneki, K., Inoue, T., Kobayashi, M., Kodama, T., … Akimitsu, N. (2014). Long noncoding RNA NEAT1-dependent SFPQ relocation from promoter region to paraspeckle mediates IL8 expression upon immune stimuli. *Molecular Cell*, *53*(3), 393–406. https://doi.org/10.1016/j.molcel.2014.01.009

Ishigaki, S., Fujioka, Y., Okada, Y., Riku, Y., Udagawa, T., Honda, D., Yokoi, S., Endo, K., Ikenaka, K., Takagi, S., Iguchi, Y., Sahara, N., Takashima, A., Okano, H., Yoshida, M., Warita, H., Aoki, M., Watanabe, H., Okado, H., … Sobue, G. (2017). Altered Tau Isoform Ratio Caused by Loss of FUS and SFPQ Function Leads to FTLD-like Phenotypes. *Cell Reports*, *18*(5), 1118–1131. https://doi.org/10.1016/j.celrep.2017.01.013

Ishigaki, S., Riku, Y., Fujioka, Y., Endo, K., Iwade, N., Kawai, K., Ishibashi, M., Yokoi, S., Katsuno, M., Watanabe, H., Mori, K., Akagi, A., Yokota, O., Terada, S., Kawakami, I., Suzuki, N., Warita, H., Aoki, M., Yoshida, M., & Sobue, G. (2020). Aberrant interaction between FUS and SFPQ in neurons in a wide range of FTLD spectrum diseases. *Brain: A Journal of Neurology*, *143*(8), 2398–2405. https://doi.org/10.1093/brain/awaa196

Jaksik, R., Iwanaszko, M., Rzeszowska-Wolny, J., & Kimmel, M. (2015). Microarray experiments and factors which affect their reliability. *Biology Direct*, *10*(1), 46. https://doi.org/10.1186/s13062-015-0077-2

Ji, Q., Zhang, L., Liu, X., Zhou, L., Wang, W., Han, Z., Sui, H., Tang, Y., Wang, Y., Liu, N., Ren, J., Hou, F., & Li, Q. (2014). Long non-coding RNA MALAT1 promotes tumour growth and metastasis in colorectal cancer through binding to SFPQ and releasing oncogene PTBP2 from SFPQ/PTBP2 complex. *British Journal of Cancer*, *111*(4), 736–748. https://doi.org/10.1038/bjc.2014.383

Jiang, F., He, H., Zhang, Y., Yang, D.-L., Huang, J.-H., Zhu, Y., Mo, R., Chen, G., Yang, S., Chen, Y., Zhong, W., & Zhou, W.-L. (2013). An Integrative Proteomics and Interaction Network-Based Classifier for Prostate Cancer Diagnosis. *PLoS ONE*, *8*(5), e63941. https://doi.org/10.1371/journal.pone.0063941

Jiao, Y., Li, Y., Liu, S., Chen, Q., & Liu, Y. (2019). ITGA3 serves as a diagnostic and prognostic biomarker for pancreatic cancer. *OncoTargets and Therapy*, *12*, 4141–4152. https://doi.org/10.2147/OTT.S201675

Kanai, Y., Dohmae, N., & Hirokawa, N. (2004). Kinesin Transports RNA: Isolation and Characterization of an RNA-Transporting Granule. *Neuron*, *43*(4), 513–525. https://doi.org/10.1016/j.neuron.2004.07.022

Kawataki, T., Yamane, T., Naganuma, H., Rousselle, P., Andurén, I., Tryggvason, K., & Patarroyo, M. (2007). Laminin isoforms and their integrin receptors in glioma cell migration and invasiveness: Evidence for a role of alpha5-laminin(s) and alpha3beta1 integrin. *Experimental Cell Research*, *313*(18), 3819–3831. https://doi.org/10.1016/j.yexcr.2007.07.038

Khramtsov, A. I., Khramtsova, G. F., Tretiakova, M., Huo, D., Olopade, O. I., & Goss, K. H. (2010). Wnt/β-Catenin Pathway Activation Is Enriched in Basal-Like Breast Cancers and Predicts Poor Outcome. *The American Journal of Pathology*, *176*(6), 2911–2920. https://doi.org/10.2353/ajpath.2010.091125

Kim, K. K., Kim, Y. C., Adelstein, R. S., & Kawamoto, S. (2011). Fox-3 and PSF interact to activate neural cell-specific alternative splicing. *Nucleic Acids Research*, *39*(8), 3064–3078. https://doi.org/10.1093/nar/gkq1221

King, H. A., Cobbold, L. C., Pichon, X., Pöyry, T., Wilson, L. A., Booden, H., Jukes-Jones, R., Cain, K., Lilley, K. S., Bushell, M., & Willis, A. E. (2014). Remodelling of a polypyrimidine tract-binding protein complex during apoptosis activates cellular IRESs. *Cell Death and Differentiation*, *21*(1), 161–171. https://doi.org/10.1038/cdd.2013.135

Klotz-Noack, K., Klinger, B., Rivera, M., Bublitz, N., Uhlitz, F., Riemer, P., Lüthen, M., Sell, T., Kasack, K., Gastl, B., Ispasanie, S. S. S., Simon, T., Janssen, N., Schwab, M., Zuber, J., Horst, D., Blüthgen, N., Schäfer, R., Morkel, M., & Sers, C. (2020). SFPQ Depletion Is Synthetically Lethal with BRAFV600E in Colorectal Cancer Cells. *Cell Reports*, *32*(12), 108184. https://doi.org/10.1016/j.celrep.2020.108184

Knott, G. J., Bond, C. S., & Fox, A. H. (2016). The DBHS proteins SFPQ, NONO and PSPC1: A multipurpose molecular scaffold. *Nucleic Acids Research*, *44*(9), 3989–4004. https://doi.org/10.1093/nar/gkw271

Knott, G. J., Lee, M., Passon, D. M., Fox, A. H., & Bond, C. S. (2015). Caenorhabditis elegans NONO-1: Insights into DBHS protein structure, architecture, and function. *Protein Science : A Publication of the Protein Society*, *24*(12), 2033–2043. https://doi.org/10.1002/pro.2816

Kuznetsov, S. G., Haines, D. C., Martin, B. K., & Sharan, S. K. (2009). Loss of Rad51c leads to embryonic lethality and modulation of Trp53-dependent tumorigenesis in mice. *Cancer Research*, *69*(3), 863–872. https://doi.org/10.1158/0008-5472.CAN-08-3057

Lane, S. W., Wang, Y. J., Lo Celso, C., Ragu, C., Bullinger, L., Sykes, S. M., Ferraro, F., Shterental, S., Lin, C. P., Gilliland, D. G., Scadden, D. T., Armstrong, S. A., & Williams, D. A. (2011). Differential niche and Wnt requirements during acute myeloid leukemia progression. *Blood*, *118*(10), 2849–2856. https://doi.org/10.1182/blood-2011-03-345165

Lee, M., Sadowska, A., Bekere, I., Ho, D., Gully, B. S., Lu, Y., Iyer, K. S., Trewhella, J., Fox, A. H., & Bond, C. S. (2015). The structure of human SFPQ reveals a coiled-coil mediated polymer essential for functional aggregation in gene regulation. *Nucleic Acids Research*, *43*(7), 3826–3840. https://doi.org/10.1093/nar/gkv156

Lee, P. S., Fang, J., Jessop, L., Myers, T., Raj, P., Hu, N., Wang, C., Taylor, P. R., Wang, J., Khan, J., Jasin, M., & Chanock, S. J. (2014). RAD51B Activity and Cell Cycle Regulation in Response to DNA Damage in Breast Cancer Cell Lines. *Breast Cancer : Basic and Clinical Research*, *8*, 135–144. https://doi.org/10.4137/BCBCR.S17766

Lee, Y., & Rio, D. C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual Review of Biochemistry*, *84*, 291–323. https://doi.org/10.1146/annurev-biochem-060614-034316

Li, W., & Melton, D. W. (2012). Cisplatin regulates the MAPK kinase pathway to induce increased expression of DNA repair gene ERCC1 and increase melanoma chemoresistance. *Oncogene*, *31*(19), 2412–2422. https://doi.org/10.1038/onc.2011.426

Liang, S., & Lutz, C. S. (2006). P54nrb is a component of the snRNP-free U1A (SF-A) complex that promotes pre-mRNA cleavage during polyadenylation. *RNA*, *12*(1), 111–121. https://doi.org/10.1261/rna.2213506

Llères, D., Denegri, M., Biggiogera, M., Ajuh, P., & Lamond, A. I. (2010). Direct interaction between hnRNP-M and CDC5L/PLRG1 proteins affects alternative splice site choice. *EMBO Reports*, *11*(6), 445–451. https://doi.org/10.1038/embor.2010.64

Lowery, L. A., Rubin, J., & Sive, H. (2007). Whitesnake/sfpq is required for cell survival and neuronal development in the zebrafish. *Developmental Dynamics*, *236*(5), 1347–1357. https://doi.org/10.1002/dvdy.21132

Lu, J., Shu, R., & Zhu, Y. (2018). Dysregulation and Dislocation of SFPQ Disturbed DNA Organization in Alzheimer's Disease and Frontotemporal Dementia. *Journal of Alzheimer's Disease: JAD*, *61*(4), 1311–1321. https://doi.org/10.3233/JAD-170659

Luan, W., Li, L., Shi, Y., Bu, X., Xia, Y., Wang, J., Djangmah, H. S., Liu, X., You, Y., & Xu, B. (2016). Long non-coding RNA MALAT1 acts as a competing endogenous RNA to promote malignant melanoma growth and metastasis by sponging miR-22. *Oncotarget*, *7*(39), 63901–63912. https://doi.org/10.18632/oncotarget.11564

Luisier, R., Tyzack, G. E., Hall, C. E., Mitchell, J. S., Devine, H., Taha, D. M., Malik, B., Meyer, I., Greensmith, L., Newcombe, J., Ule, J., Luscombe, N. M., & Patani, R. (2018). Intron retention and nuclear loss of SFPQ are molecular hallmarks of ALS. *Nature Communications*, *9*, 2010. https://doi.org/10.1038/s41467-018-04373-8

Lund, S. P., Nettleton, D., McCarthy, D. J., & Smyth, G. K. (2012). Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical*

Applications in Genetics and Molecular Biology, *11*(5), /j/sagmb.2012.11.issue-5/1544-6115.1826/1544-6115.1826.xml. https://doi.org/10.1515/1544-6115.1826

Majid, S., Dar, A. A., Saini, S., Chen, Y., Shahryari, V., Liu, J., Zaman, M. S., Hirata, H., Yamamura, S., Ueno, K., Tanaka, Y., & Dahiya, R. (2010). Regulation of minichromosome maintenance gene family by microRNA-1296 and genistein in prostate cancer. *Cancer Research*, *70*(7), 2809–2818. https://doi.org/10.1158/0008-5472.CAN-09-4176

Major, A. T., Hogarth, C. A., Young, J. C., Kurihara, Y., Jans, D. A., & Loveland, K. L. (2019). Dynamic paraspeckle component localisation during spermatogenesis. *Reproduction*, *158*(3), 267–280. https://doi.org/10.1530/REP-19-0139

Marko, M., Leichter, M., Patrinou-Georgoula, M., & Guialis, A. (2010). HnRNP M interacts with PSF and p54nrb and co-localizes within defined nuclear structures. *Experimental Cell Research*, *316*(3), 390–400. https://doi.org/10.1016/j.yexcr.2009.10.021

Marot, G., Foulley, J.-L., Mayer, C.-D., & Jaffrézic, F. (2009). Moderated effect size and P-value combinations for microarray meta-analyses. *Bioinformatics (Oxford, England)*, *25*(20), 2692–2699. https://doi.org/10.1093/bioinformatics/btp444

Mathur, M., Tucker, P. W., & Samuels, H. H. (2001). PSF Is a Novel Corepressor That Mediates Its Effect through Sin3A and the DNA Binding Domain of Nuclear Hormone Receptors. *Molecular and Cellular Biology*, *21*(7), 2298–2311. https://doi.org/10.1128/MCB.21.7.2298-2311.2001

Melton, A. A., Jackson, J., Wang, J., & Lynch, K. W. (2007). Combinatorial Control of Signal-Induced Exon Repression by hnRNP L and PSF. *Molecular and Cellular Biology*, *27*(19), 6972–6984. https://doi.org/10.1128/MCB.00419-07

Mitobe, Y., Iino, K., Takayama, K., Ikeda, K., Suzuki, T., Aogi, K., Kawabata, H., Suzuki, Y., Horie-Inoue, K., & Inoue, S. (2020). PSF Promotes ER-Positive Breast Cancer Progression via Posttranscriptional Regulation of ESR1 and SCFD2. *Cancer Research*, *80*(11), 2230–2242. https://doi.org/10.1158/0008-5472.CAN-19-3095

Muralidhar, S., Filia, A., Nsengimana, J., Poźniak, J., O'Shea, S. J., Diaz, J. M., Harland, M., Randerson-Moor, J. A., Reichrath, J., Laye, J. P., Weyden, L. van der, Adams, D. J., Bishop, D. T., & Newton-Bishop, J. (2019). Vitamin D–VDR Signaling Inhibits Wnt/β-Catenin–Mediated Melanoma Progression and Promotes Antitumor Immunity. *Cancer Research*, *79*(23), 5986–5998. https://doi.org/10.1158/0008-5472.CAN-18-3927

Nagy, Z. B., Barták, B. K., Kalmár, A., Galamb, O., Wichmann, B., Dank, M., Igaz, P., Tulassay, Z., & Molnár, B. (2019). Comparison of Circulating miRNAs Expression Alterations in Matched Tissue and Plasma Samples During Colorectal Cancer Progression. *Pathology Oncology Research: POR*, *25*(1), 97–105. https://doi.org/10.1007/s12253-017-0308-1

Neumann, M., Rademakers, R., Roeber, S., Baker, M., Kretzschmar, H. A., & Mackenzie, I. R. A. (2009). A new subtype of frontotemporal lobar degeneration with FUS pathology. *Brain*, *132*(11), 2922–2931. https://doi.org/10.1093/brain/awp214

O'Connor, J. P., Alwine, J. C., & Lutz, C. S. (1997). Identification of a novel, non-snRNP protein complex containing U1A protein. *RNA*, *3*(12), 1444–1455.

Osera, C., Martindale, J. L., Amadio, M., Kim, J., Yang, X., Moad, C. A., Indig, F. E., Govoni, S., Abdelmohsen, K., Gorospe, M., & Pascale, A. (2015). Induction of VEGFA mRNA translation by CoCl2 mediated by HuR. *RNA Biology*, *12*(10), 1121–1130. https://doi.org/10.1080/15476286.2015.1085276

Ou, X., Gao, G., Bazhabayi, M., Zhang, K., Liu, F., & Xiao, X. (2019). MALAT1 and BACH1 are prognostic biomarkers for triple-negative breast cancer. *Journal of Cancer Research and Therapeutics*, *15*(7), 1597–1602. https://doi.org/10.4103/jcrt.JCRT_282_19

Passon, D. M., Lee, M., Rackham, O., Stanley, W. A., Sadowska, A., Filipovska, A., Fox, A. H., & Bond, C. S. (2012). Structure of the heterodimer of human NONO and paraspeckle protein component 1 and analysis of its role in subnuclear body formation. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(13), 4846–4850. https://doi.org/10.1073/pnas.1120792109

Patton, J. G., Porro, E. B., Galceran, J., Tempst, P., & Nadal-Ginard, B. (1993). Cloning and characterization of PSF, a novel pre-mRNA splicing factor. *Genes & Development*, *7*(3), 393–406. https://doi.org/10.1101/gad.7.3.393

Peng, W., Furuuchi, N., Aslanukova, L., Huang, Y.-H., Brown, S. Z., Jiang, W., Addya, S., Vishwakarma, V., Peters, E., Brody, J. R., Dixon, D. A., & Sawicki, J. A. (2018). Elevated HuR in Pancreas Promotes a Pancreatitis-Like Inflammatory Microenvironment That Facilitates Tumor Development. *Molecular and Cellular Biology*, *38*(3), e00427-17. https://doi.org/10.1128/MCB.00427-17

Petti, E., Buemi, V., Zappone, A., Schillaci, O., Broccia, P. V., Dinami, R., Matteoni, S., Benetti, R., & Schoeftner, S. (2019). SFPQ and NONO suppress RNA:DNA-hybrid-related telomere instability. *Nature Communications*, *10*, 1001. https://doi.org/10.1038/s41467-019-08863-1

Rajesh, C., Baker, D. K., Pierce, A. J., & Pittman, D. L. (2011). The splicing-factor related protein SFPQ/PSF interacts with RAD51D and is necessary for homology-directed repair and sister chromatid cohesion. *Nucleic Acids Research*, *39*(1), 132–145. https://doi.org/10.1093/nar/gkq738

Rau, A., Marot, G., & Jaffrézic, F. (2014). Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinformatics*, *15*(1), 91. https://doi.org/10.1186/1471-2105-15-91

Ray, P., Kar, A., Fushimi, K., Havlioglu, N., Chen, X., & Wu, J. Y. (2011). PSF Suppresses Tau Exon 10 Inclusion by Interacting with a Stem-Loop Structure Downstream of Exon 10. *Journal of Molecular Neuroscience : MN*, *45*(3), 453–466. https://doi.org/10.1007/s12031-011-9634-z

Ren, S., She, M., Li, M., Zhou, Q., Liu, R., Lu, H., Yang, C., & Xiong, D. (2014). The RNA/DNA-binding protein PSF relocates to cell membrane and contributes cells' sensitivity to antitumor drug, doxorubicin. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology*, *85*(3), 231–241. https://doi.org/10.1002/cyto.a.22423

Rhee, D. K., Hockman, S. C., Choi, S.-K., Kim, Y.-E., Park, C., Manganiello, V. C., & Kim, K. K. (2017). SFPQ, a multifunctional nuclear protein, regulates the transcription of PDE3A. *Bioscience Reports*, *37*(4), BSR20170975. https://doi.org/10.1042/BSR20170975

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E., & Chang, H. Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, *129*(7), 1311–1323. https://doi.org/10.1016/j.cell.2007.05.022

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25

Roepcke, S., Stahlberg, S., Klein, H., Schulz, M. H., Theobald, L., Gohlke, S., Vingron, M., & Walther, D. J. (2011). A tandem sequence motif acts as a distance-dependent enhancer in a set of genes involved in translation by binding the proteins NonO and SFPQ. *BMC Genomics*, *12*, 624. https://doi.org/10.1186/1471-2164-12-624

Rosonina, E., Ip, J. Y. Y., Calarco, J. A., Bakowski, M. A., Emili, A., McCracken, S., Tucker, P., Ingles, C. J., & Blencowe, B. J. (2005). Role for PSF in Mediating Transcriptional Activator-Dependent Stimulation of Pre-mRNA Processing In Vivo. *Molecular and Cellular Biology*, *25*(15), 6734–6746. https://doi.org/10.1128/MCB.25.15.6734-6746.2005

Ru, Y., Chen, X.-J., Guo, W.-Z., Gao, S.-G., Qi, Y.-J., Chen, P., Feng, X.-S., & Zhang, S.-J. (2018). NEAT1_2–SFPQ axis mediates cisplatin resistance in liver cancer cells in vitro. *OncoTargets and Therapy*, *11*, 5695–5702. https://doi.org/10.2147/OTT.S163774

Sakurai, T., Kashida, H., Watanabe, T., Hagiwara, S., Mizushima, T., Iijima, H., Nishida, N., Higashitsuji, H., Fujita, J., & Kudo, M. (2014). Stress Response Protein Cirp Links

Inflammation and Tumorigenesis in Colitis-Associated Cancer. *Cancer Research*, *74*(21), 6119–6128. https://doi.org/10.1158/0008-5472.CAN-14-0471

Salton, M., Lerenthal, Y., Wang, S.-Y., Chen, D. J., & Shiloh, Y. (2010). Involvement of Matrin 3 and SFPQ/NONO in the DNA damage response. *Cell Cycle (Georgetown, Tex.)*, *9*(8), 1568–1576. https://doi.org/10.4161/cc.9.8.11298

Sasaki, Y. T. F., Ideue, T., Sano, M., Mituyama, T., & Hirose, T. (2009). MENε/β noncoding RNAs are essential for structural integrity of nuclear paraspeckles. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(8), 2525–2530. https://doi.org/10.1073/pnas.0807899106

Sharathchandra, A., Lal, R., Khan, D., & Das, S. (2012). Annexin A2 and PSF proteins interact with p53 IRES and regulate translation of p53 mRNA. *RNA Biology*, *9*(12), 1429–1439. https://doi.org/10.4161/rna.22707

Shav-Tal, Y., Cohen, M., Lapter, S., Dye, B., Patton, J. G., Vandekerckhove, J., & Zipori, D. (2001). Nuclear Relocalization of the Pre-mRNA Splicing Factor PSF during Apoptosis Involves Hyperphosphorylation, Masking of Antigenic Epitopes, and Changes in Protein Interactions. *Molecular Biology of the Cell*, *12*(8), 2328–2340.

Shav-Tal, Y., & Zipori, D. (2002). PSF and p54(nrb)/NonO--multi-functional nuclear proteins. *FEBS Letters*, *531*(2), 109–114. https://doi.org/10.1016/s0014-5793(02)03447-6

Siangphoe, U., & Archer, K. J. (2017). Estimation of random effects and identifying heterogeneous genes in meta-analysis of gene expression studies. *Briefings in Bioinformatics*, *18*(4), 602–618. https://doi.org/10.1093/bib/bbw050

Stagsted, L. V. W., O'Leary, E. T., Ebbesen, K. K., & Hansen, T. B. (2021). The RNA-binding protein SFPQ preserves long-intron splicing and regulates circRNA biogenesis in mammals. *ELife*, *10*, e63088. https://doi.org/10.7554/eLife.63088

Stojic, L., Niemczyk, M., Orjalo, A., Ito, Y., Ruijter, A. E. M., Uribe-Lewis, S., Joseph, N., Weston, S., Menon, S., Odom, D. T., Rinn, J., Gergely, F., & Murrell, A. (2016). Transcriptional

silencing of long noncoding RNA GNG12-AS1 uncouples its transcriptional and product-related functions. *Nature Communications*, *7*, 10406. https://doi.org/10.1038/ncomms10406

Ström, L., Karlsson, C., Lindroos, H. B., Wedahl, S., Katou, Y., Shirahige, K., & Sjögren, C. (2007). Postreplicative formation of cohesion is required for repair and induced by a single DNA break. *Science (New York, N.Y.)*, *317*(5835), 242–245. https://doi.org/10.1126/science.1140649

Takayama, K., Fujimura, T., Suzuki, Y., & Inoue, S. (2020). Identification of long non-coding RNAs in advanced prostate cancer associated with androgen receptor splicing factors. *Communications Biology*, *3*, 393. https://doi.org/10.1038/s42003-020-01120-y

Takayama, K., Suzuki, T., Fujimura, T., Yamada, Y., Takahashi, S., Homma, Y., Suzuki, Y., & Inoue, S. (2017). Dysregulation of spliceosome gene expression in advanced prostate cancer by RNA-binding protein PSF. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(39), 10461–10466. https://doi.org/10.1073/pnas.1706076114

Takayama, K.-I., Horie-Inoue, K., Katayama, S., Suzuki, T., Tsutsumi, S., Ikeda, K., Urano, T., Fujimura, T., Takagi, K., Takahashi, S., Homma, Y., Ouchi, Y., Aburatani, H., Hayashizaki, Y., & Inoue, S. (2013). Androgen-responsive long noncoding RNA CTBP1-AS promotes prostate cancer. *The EMBO Journal*, *32*(12), 1665–1680. https://doi.org/10.1038/emboj.2013.99

Takayama, K.-I., Suzuki, T., Fujimura, T., Yamada, Y., Takahashi, S., Homma, Y., Suzuki, Y., & Inoue, S. (2017). Dysregulation of spliceosome gene expression in advanced prostate cancer by RNA-binding protein PSF. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(39), 10461–10466. https://doi.org/10.1073/pnas.1706076114

Takeuchi, A., Iida, K., Tsubota, T., Hosokawa, M., Denawa, M., Brown, J. B., Ninomiya, K., Ito, M., Kimura, H., Abe, T., Kiyonari, H., Ohno, K., & Hagiwara, M. (2018). Loss of Sfpq Causes Long-Gene Transcriptopathy in the Brain. *Cell Reports*, *23*(5), 1326–1341. https://doi.org/10.1016/j.celrep.2018.03.141

Tao, Y., Ma, C., Fan, Q., Wang, Y., Han, T., & Sun, C. (2018). MicroRNA-1296 Facilitates Proliferation, Migration And Invasion Of Colorectal Cancer Cells By Targeting SFPQ. *Journal of Cancer*, *9*(13), 2317–2326. https://doi.org/10.7150/jca.25427

Thiagalingam, A., De Bustros, A., Borges, M., Jasti, R., Compton, D., Diamond, L., Mabry, M., Ball, D. W., Baylin, S. B., & Nelkin, B. D. (1996). RREB-1, a novel zinc finger protein, is involved in the differentiation response to Ras in human medullary thyroid carcinomas. *Molecular and Cellular Biology*, *16*(10), 5335–5345.

Thomas-Jinu, S., Gordon, P. M., Fielding, T., Taylor, R., Smith, B. N., Snowden, V., Blanc, E., Vance, C., Topp, S., Wong, C.-H., Bielen, H., Williams, K. L., McCann, E. P., Nicholson, G. A., Pan-Vazquez, A., Fox, A. H., Bond, C. S., Talbot, W. S., Blair, I. P., … Houart, C. (2017). Non-nuclear Pool of Splicing Factor SFPQ Regulates Axonal Transcripts Required for Normal Motor Development. *Neuron*, *94*(2), 322-336.e5. https://doi.org/10.1016/j.neuron.2017.03.026

Toro-Domínguez, D., Villatoro-García, J. A., Martorell-Marugán, J., Román-Montoya, Y., Alarcón-Riquelme, M. E., & Carmona-Sáez, P. (2021). A survey of gene expression meta-analysis: Methods and applications. *Briefings in Bioinformatics*, *22*(2), 1694–1705. https://doi.org/10.1093/bib/bbaa019

Tsukahara, T., Haniu, H., & Matsuda, Y. (2013). PTB-Associated Splicing Factor (PSF) Is a PPARγ-Binding Protein and Growth Regulator of Colon Cancer Cells. *PLoS ONE*, *8*(3), e58749. https://doi.org/10.1371/journal.pone.0058749

Tsukahara, T., Matsuda, Y., & Haniu, H. (2013). PSF Knockdown Enhances Apoptosis via Downregulation of LC3B in Human Colon Cancer Cells. *BioMed Research International*, *2013*, 204973. https://doi.org/10.1155/2013/204973

Turri-Zanoni, M., Medicina, D., Lombardi, D., Ungari, M., Balzarini, P., Rossini, C., Pellegrini, W., Battaglia, P., Capella, C., Castelnuovo, P., Palmedo, G., Facchetti, F., Kutzner, H., Nicolai, P., & Vermi, W. (2013). Sinonasal mucosal melanoma: Molecular profile and therapeutic

implications from a series of 32 cases. *Head & Neck*, *35*(8), 1066–1077. https://doi.org/10.1002/hed.23079

Tyzack, G. E., Luisier, R., Taha, D. M., Neeves, J., Modic, M., Mitchell, J. S., Meyer, I., Greensmith, L., Newcombe, J., Ule, J., Luscombe, N. M., & Patani, R. (2019). Widespread FUS mislocalization is a molecular hallmark of amyotrophic lateral sclerosis. *Brain*, *142*(9), 2572–2580. https://doi.org/10.1093/brain/awz217

Van Nostrand, E. L., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., Blue, S. M., Chen, J.-Y., Cody, N. A. L., Dominguez, D., Olson, S., Sundararaman, B., Zhan, L., Bazile, C., Bouvrette, L. P. B., Bergalet, J., Duff, M. O., Garcia, K. E., Gelboin-Burkhart, C., … Yeo, G. W. (2020). A large-scale binding and functional map of human RNA-binding proteins. *Nature*, *583*(7818), 711–719. https://doi.org/10.1038/s41586-020-2077-3

Vogelpoel, L. T. C., Baeten, D. L. P., de Jong, E. C., & den Dunnen, J. (2015). Control of Cytokine Production by Human Fc Gamma Receptors: Implications for Pathogen Defense and Autoimmunity. *Frontiers in Immunology*, *6*, 79. https://doi.org/10.3389/fimmu.2015.00079

Waldron, L., & Riester, M. (2016). Meta-Analysis in Gene Expression Studies. In E. Mathé & S. Davis (Eds.), *Statistical Genomics* (Vol. 1418, pp. 161–176). Springer New York. https://doi.org/10.1007/978-1-4939-3578-9_8

Wang, Q., Li, W., Zhang, Y., Yuan, X., Xu, K., Yu, J., Chen, Z., Beroukhim, R., Wang, H., Lupien, M., Wu, T., Regan, M. M., Meyer, C. A., Carroll, J. S., Manrai, A. K., Jänne, O. A., Balk, S. P., Mehra, R., Han, B., … Brown, M. (2009). Androgen Receptor Regulates a Distinct Transcription Program in Androgen-Independent Prostate Cancer. *Cell*, *138*(2), 245–256. https://doi.org/10.1016/j.cell.2009.04.056

Wang, S., Han, H., Meng, J., Yang, W., Lv, Y., & Wen, X. (2021). Long non-coding RNA SNHG1 suppresses cell migration and invasion and upregulates SOCS2 in human gastric carcinoma. *Biochemistry and Biophysics Reports*, *27*, 101052. https://doi.org/10.1016/j.bbrep.2021.101052

Weidensdorfer, D., Stöhr, N., Baude, A., Lederer, M., Köhn, M., Schierhorn, A., Buchmeier, S., Wahle, E., & Hüttelmaier, S. (2009). Control of c-myc mRNA stability by IGF2BP1-associated cytoplasmic RNPs. *RNA*, *15*(1), 104–115. https://doi.org/10.1261/rna.1175909

Wickham, H. (2016). *ggplot2: Elegent Graphics for Data Analysis*. Springer Verlag New York. https://ggplot2.tidyverse.org/

Wilusz, J. E., JnBaptiste, C. K., Lu, L. Y., Kuhn, C.-D., Joshua-Tor, L., & Sharp, P. A. (2012). A triple helix stabilizes the 3′ ends of long noncoding RNAs that lack poly(A) tails. *Genes & Development*, *26*(21), 2392–2407. https://doi.org/10.1101/gad.204438.112

Wu, C.-F., Tan, G.-H., Ma, C.-C., & Li, L. (2013). The Non-Coding RNA Llme23 Drives the Malignant Property of Human Melanoma Cells. *Journal of Genetics and Genomics*, *40*(4), 179–188. https://doi.org/10.1016/j.jgg.2013.03.001

Wu, L., Fan, J., & Belasco, J. G. (2006). MicroRNAs direct rapid deadenylation of mRNA. *Proceedings of the National Academy of Sciences*, *103*(11), 4034–4039. https://doi.org/10.1073/pnas.0510928103

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, *2*(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

Xia, C., Liang, S., He, Z., Zhu, X., Chen, R., & Chen, J. (2018). Metformin, a first-line drug for type 2 diabetes mellitus, disrupts the MALAT1/miR-142-3p sponge to decrease invasion and migration in cervical cancer cells. *European Journal of Pharmacology*, *830*, 59–67. https://doi.org/10.1016/j.ejphar.2018.04.027

Xu, J., Prosperi, J. R., Choudhury, N., Olopade, O. I., & Goss, K. H. (2015). β-Catenin Is Required for the Tumorigenic Behavior of Triple-Negative Breast Cancer Cells. *PLOS ONE*, *10*(2), e0117097. https://doi.org/10.1371/journal.pone.0117097

Xue, J., & Zhang, F. (2020). LncRNA LINC00511 plays an oncogenic role in lung adenocarcinoma by regulating PKM2 expression via sponging miR-625-5p. *Thoracic Cancer*, *11*(9), 2570–2579. https://doi.org/10.1111/1759-7714.13576

Yamazaki, T., Souquere, S., Chujo, T., Kobelke, S., Chong, Y. S., Fox, A. H., Bond, C. S., Nakagawa, S., Pierron, G., & Hirose, T. (2018). Functional Domains of NEAT1 Architectural lncRNA Induce Paraspeckle Assembly through Phase Separation. *Molecular Cell*, *70*(6), 1038-1053.e7. https://doi.org/10.1016/j.molcel.2018.05.019

Yang, P., Chen, T., Xu, Z., Zhu, H., Wang, J., & He, Z. (2016). Long noncoding RNA GAPLINC promotes invasion in colorectal cancer by targeting SNAI2 through binding with PSF and NONO. *Oncotarget*, *7*(27), 42183–42194. https://doi.org/10.18632/oncotarget.9741

Yoon, J.-H., De, S., Srikantan, S., Abdelmohsen, K., Grammatikakis, I., Kim, J., Kim, K. M., Noh, J. H., White, E. J. F., Martindale, J. L., Yang, X., Kang, M.-J., Wood, W. H., Hooten, N. N., Evans, M. K., Becker, K. G., Tripathi, V., Prasanth, K. V., Wilson, G. M., … Gorospe, M. (2014). PAR-CLIP analysis uncovers AUF1 impact on target RNA fate and genome integrity. *Nature Communications*, *5*, 5248. https://doi.org/10.1038/ncomms6248

Zeng, C., Liu, S., Lu, S., Yu, X., Lai, J., Wu, Y., Chen, S., Wang, L., Yu, Z., Luo, G., & Li, Y. (2018). The c-Myc-regulated lncRNA NEAT1 and paraspeckles modulate imatinib-induced apoptosis in CML cells. *Molecular Cancer*, *17*, 130. https://doi.org/10.1186/s12943-018-0884-z

Zentner, G. E., Tesar, P. J., & Scacheri, P. C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Research*, *21*(8), 1273–1283. https://doi.org/10.1101/gr.122382.111

Zhang, C., Luo, X., Liu, L., Guo, S., Zhao, W., Mu, A., Liu, Z., Wang, N., Zhou, H., & Zhang, T. (2013). Myocardin-related transcription factor A is up-regulated by 17β-estradiol and promotes migration of MCF-7 breast cancer cells via transactivation of MYL9 and CYR61. *Acta Biochimica et Biophysica Sinica*, *45*(11), 921–927. https://doi.org/10.1093/abbs/gmt104

Zhang, X., Wan, G., Berger, F. G., He, X., & Lu, X. (2011). The ATM Kinase Induces MicroRNA Biogenesis in the DNA Damage Response. *Molecular Cell*, *41*(4), 371–383. https://doi.org/10.1016/j.molcel.2011.01.020

Zhou, X., Li, X., Yu, L., Wang, R., Hua, D., Shi, C., Sun, C., Luo, W., Rao, C., Jiang, Z., Wang, Q., & Yu, S. (2019). The RNA-binding protein SRSF1 is a key cell cycle regulator via stabilizing NEAT1 in glioma. *The International Journal of Biochemistry & Cell Biology*, *113*, 75–86. https://doi.org/10.1016/j.biocel.2019.06.003

Zou, J., Milon, B. C., Desouki, M. M., Costello, L. C., & Franklin, R. B. (2011). HZIP1 Zinc Transporter Down-Regulation in Prostate Cancer Involves the Over Expression of Ras Responsive Element Binding Protein-1 (RREB-1). *The Prostate*, *71*(14), 1518–1524. https://doi.org/10.1002/pros.21368