CHAPTER 1

INTRODUCTION

1.1 Extracellular receptor-ligand interactions

Extracellular protein-protein interactions are central to many biological processes including cell signalling, migration and fusion. Beyond interacting with proteins from the same organism, some membrane associated proteins are also exploited by pathogens to evade host immune recognition and enable successful reproduction in host tissue (Pizarro-Cerdá and Cossart, 2006). Therefore, determining extracellular receptor-ligand pairs essential to such processes is pertinent to gain a more complete understanding of signalling pathways that govern development and disease. In addition, cell surface proteins are attractive targets for drug discovery due to their accessibility to systemically administered therapeutics such as antibodies. The identification of conserved host-pathogen interactions at the cell surface has also helped to inform vaccine development (Ord et al., 2015).

Despite their importance in development and infection, interactions involving plasma membrane proteins are not well characterised due to their amphipathic nature and the specialised environment in which they function. This presents problems when using common biochemical methods for large-scale protein interaction screening, which are more often suited for studying intracellular complexes. As a result, interactions involving plasma membrane proteins tend to be underrepresented in large-scale proteomic interaction datasets (Futschik et al., 2007). Several methods have been developed to address the unique characteristics of extracellular interactions, but can be restricted to investigating certain classes

of membrane proteins or require large investments in terms of cost and equipment. Consequently, the development of more cost-effective and comprehensive approaches for large-scale extracellular interaction screening would facilitate the discovery of novel receptor-ligand pairs.

1.1.1 Challenges of studying extracellular protein-protein interactions

Solubilising membrane proteins with detergents interferes with mass spectrometry analysis

Intracellular proteins are often soluble in aqueous solutions, forming stable globular structures with solvent-exposed hydrophilic amino acid side chains facing outward and hydrophobic moieties hidden within the core of the protein. As a result, intracellular protein complexes can be easily solubilised for further characterisation without disrupting binding between members of the complex. In contrast, integral membrane proteins possess distinct hydrophilic and hydrophobic domains and may be dependent on insertion into the plasma membrane to maintain their 3-dimensional structure. Due to their amphipathic nature, membrane proteins are not as easily solubilised in aqueous solvents as intracellular proteins are. Detergents such as Triton X-100 and sodium dodecyl sulfate (SDS) can be helpful for solubilising proteins with hydrophobic domains, but may at the same time denature secondary or tertiary structures and therefore disrupt native protein complexes.

The use of detergents to solubilise membrane protein complexes is particularly relevant when using affinity purification in tandem with mass spectrometry (AP-MS) to characterise protein complexes. AP-MS involves the enrichment of protein complexes containing a specific protein, called a 'bait' protein, which is typically tagged and overexpressed in a cell line of interest. Enrichment of protein complexes from cell lysate is performed using a matrix with high affinity for the epitope tag. Purified complexes are denatured and size-separated by gel electrophoresis before being digested into peptides with trypsin. Rapid and sensitive identification of peptide sequences is achieved by tandem mass spectrometry and peptide fragments mapped to a database of possible fragments from known proteins in that species to elucidate protein identity and abundance (Huttlin et al., 2015). Although some detergent (0.05%) is generally used during protein extraction and purification to reduce non-specific interactions and loss of protein due to adsorption to surfaces, the higher concentrations of detergent (0.5 - 1%) needed to solubilise many membrane protein complexes are known to interfere with tryptic digest as well as mass spectrometry analysis of the resulting peptides

(Zhang and Li, 2004). In addition, membrane proteins are often glycosylated, which alters the mass of subsequent peptides in unpredictable ways, making them challenging to identify using mass spectrometry. Unsurprisingly, membrane protein complexes are therefore underrepresented in large-scale interactome studies using AP-MS (Huttlin et al., 2015).

Low affinity interactors may be lost during multiple stringent wash steps

Cell surface receptors diffuse within the plasma membrane and are thought to sometimes form local concentrations of receptors together with accessory proteins required for functional signalling. Clusters of proteins also serve to increase the avidity of an interaction occurring by providing multiple binding sites for receptor-ligand interaction. A corollary of this is that extracellular interactions, especially those between two cell surface proteins, can have low monomeric affinities and still be biologically functional. For instance, the T-cell receptor co-regulatory complex CD55-CD97, has a very high equilibrium dissociation constant (K_D) of 86 x 10^{-6} M, indicating a very low affinity. In contrast, most antibodies have K_D values in the nanomolar range (10^{-7} to 10^{-9} M). Blocking the CD55-CD97 interaction with antibodies against the extracellular domains of either CD55 or CD97 results in a significant inhibition of T-cell proliferation, indicating that interactions of such low affinity can still be functionally important (Abbott et al., 2007). Therefore, it is critical that screening technologies aimed at identifying novel extracellular interactions should be sensitive enough to detect such low affinity interactions. Further to the use of detergents reducing the likelihood of detecting low-affinity binding partners, affinity purification requires multiple stringent wash steps to reduce background contaminants due to non-specific binding to the affinity matrix, which may cause further loss of low-affinity interactors. This makes AP-MS less than ideal for identifying extracellular receptor-ligand interactions.

Detecting transient interactions with mass spectrometry is challenging

Low affinities facilitate the formation of extremely transient interactions needed for dynamic processes regulated by cell surface molecules. Leukocyte extravasation is a good example of an interaction that undergoes quick formation and dissociation as leukocytes roll and attach to endothelial cells under flow rates of 5-10 μm/s (van der Merwe and Barclay, 1994). This process is mediated by three lectins present on the endothelium (P-selectin and E-selectin) and on leukocytes (L-selectin), and is thought to be dependent on fast dissociation rates of the lectins from their respective binding partners. In addition to

dissociation rates, highly regulated expression of binding partners at the surface can also give rise to transient interactions. For instance, P and E-selectin are not expressed on the surface of quiescent endothelial cells but upregulated in response to pro-inflammatory factors like tumour necrosis factor (TNF)-α and interleukin (IL)-1. P-selectin is usually present in storage granules called Weibel–Palade bodies and is rapidly transported to the cell surface upon endothelial cell activation. This stimulates transcription of E-selectin from the *SELE* gene which is then trafficked to the plasma membrane. Maximal expression of E-selectin occurs around 6–12 hours after cytokine stimulation, with levels returning to baseline after 24 hours (Leeuwenberg et al., 1992). The short timeframe in which E-selectin is present on the surface highlights how dynamically regulated membrane proteins can form interactions only transiently.

To address the challenges of capturing low-affinity or transient extracellular interactions with mass spectrometry-based approaches, a chemically-defined crosslinking reagent, TRICEPS, has been developed to covalently link a ligand of interest to glycoproteins at the cell surface (Frei et al., 2012). TRICEPS is a trifunctional reagent that contains an N-hydroxysuccinimide (NHS)-ester for coupling to polypeptide ligands, a hydrazine moiety for capturing glycoprotein receptors on the cell surface, and a biotin moiety for affinity purification of the cross-linked receptor-ligand complex. The suggested workflow also involves performing the trypsin digest before rather than after affinity purification, which circumvents issues associated with purifying intact plasma membrane proteins, such as limited solubility and nonspecific interactions through exposed hydrophobic domains. This approach may enable the identification of receptors under near-physiological conditions; however it may result in high background signals due to nonspecific crosslinking to neighbouring glycoproteins as well as more complicated analysis of mass spectrometry data to account for the presence of the cross-linking agent in the sample.

Although AP-MS is a sensitive technique for detecting proteins, the purification of specific protein complexes before MS analysis necessitates a large amount of starting material, typically from cell lines, which can be easily expanded, or from tissue lysates provided the interaction of interest is fairly stable and abundant in the tissue. This poses a challenge for detecting interactions which occur in a short timeframe or between rare cell types as it could be difficult to obtain sufficient amounts of primary material for AP-MS.

Lack of proper post-translational modifications hinders analysis by yeast-two hybrid

Many proteins undergo post-translational modifications crucial for their function. In eukaryotes, the majority of proteins synthesised in the endoplasmic reticulum (ER) are modified by the addition of carbohydrates in a process called glycosylation. Glycosylation results in the addition of a glycosyl group to either asparagine, hydroxylysine, serine, or threonine residues in a polypeptide. Secreted and membrane-bound proteins are synthesised in the ER before being trafficked to the surface. Consequently, most secreted and membrane-bound proteins are glycosylated, and the addition of these large, bulky oligosaccharide chains modulates binding properties as well as receptor recognition (Ulloa-Aguirre et al., 1999). In some cases, the sugar moieties themselves can function as adhesive molecules independent of the core protein they are attached to. This is exemplified by the role of glycosaminoglycans (GAGs) in infectious disease, where they have been shown to act as receptors for initial attachment of a wide variety of microbial pathogens (Jinno and Park, 2015). GAGs are formed by the sulfation of mannose and N-acetylglucosamine (GlcNAc) residues of a carbohydrate side chain in a process catalysed by sulfotransferases using 3'-phosphoadenosine-5'-phosphosulfate (PAPS) as a sulfuryl donor.

Yeast-two hybrid (Y2H) is a genetic system to detect direct binary interactions between proteins in a living cell. Y2H uses the yeast Gal4 transcription factor, which consists of distinct DNA-binding and transcriptional activation domains, to bind to a conserved Upstream Activation Sequence (UAS) to regulate the expression of galactose-induced genes. Through the expression of two hybrid proteins - a bait protein fused to the N-terminal DNA-binding domain of Gal4 and a prey protein fused to the C-terminal transcriptional activator domain - interactions between bait and prey can be detected by measuring transcription levels of a reporter gene with a UAS (Fields and Song, 1989). An adaptation of the Y2H system to study interactions involving membrane proteins (MYTH) utilises a split ubiquitin system where bait proteins are fused to a C-terminal fragment of ubiquitin, as well as a transcription factor, and prey proteins are fused to the N-terminal fragment of ubiquitin. Interaction between bait and prey proteins results in the formation of a full-length 'pseudoubiquitin' molecule, which can be recognised and cleaved by cytosolic deubiquitinising enzymes. Cleavage releases the transcription factor which then enters the nucleus and activates a reporter gene to provide a readout for the presence or absence of interaction (Snider et al., 2010). This method is a cheap, efficient strategy for detecting interactions between intracellular proteins, and has been performed at scale to detect binary interactions at a proteome level (Rolland et al., 2014; Snider and Stagljar, 2016). Crucially, Y2H and MYTH proteins are expressed in yeast cells,

which may not fully recapitulate the range of post-translational modifications needed for human membrane proteins. For instance, *Saccharomyces cerevisiae* proteins undergo both N- and O-linked glycosylation, but do not form complex carbohydrate structures (Tanner and Lehle, 1987).

In addition, Y2H is performed in the reducing environment of the cytoplasm, which does not favour the formation of disulfide bonds needed to provide extracellular proteins with the proper tertiary structure for interaction (Feige and Hendershot, 2011). Disulfide bonds are formed between sulfhydryl groups on cysteine residues, and due to their covalent nature form extremely stable structures crucial for ligand binding and cell adhesion. The presence of hydrophobic transmembrane domains in integral membrane proteins also precludes them from being expressed as both soluble and functional baits/preys for Y2H, unless truncated to express only hydrophilic domains. *In vivo*, secreted and membrane-associated proteins are synthesised in the ER, which maintains an oxidising environment that better resembles the extracellular milieu, along with chaperones for protein folding and insertion of integral membrane proteins into the ER membrane.

In summary, extracellular proteins and their interactions have unique attributes that allow them to function extracellularly in a membrane-embedded context. These attributes render such interactions unsuitable for investigation using standard biochemical techniques like AP-MS and Y2H, and require novel strategies for accurate and sensitive characterisation.

1.1.2 Interaction screening using soluble recombinant ectodomains

Several studies have addressed the challenge of extracellular interaction screening by performing plate-based screens using libraries of recombinant secreted and plasma membrane proteins from cell types known to interact (Bushell et al., 2008; Özkan et al., 2013; Wojtowicz et al., 2007) (Figure 1.1). In these screens, truncated ectodomains of membrane proteins are fused to various biochemical tags to form soluble probes for directly detecting binding between ectodomains or secreted proteins. Binding is generally indicated by the retention of an ectodomain fused to an enzymatic tag, which is detected by addition of a colourimetric substrate. In order to detect low-affinity interactions, ectodomains are also often oligomerised to increase local avidity during the binding assay.

The exact strategies for affinity capture onto plates, ectodomain oligomerisation, and interaction detection differ between studies. Wojtowicz *et al.* expressed the ectodomains of 92 isoforms of the highly polymorphic Drosophila Down Syndrome cell adhesion molecule

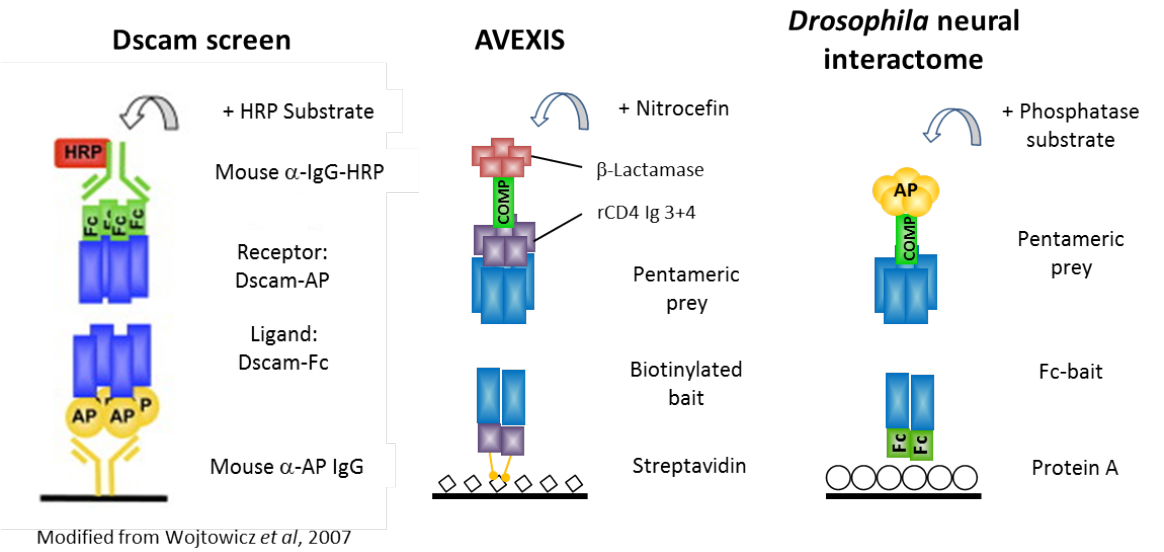

Figure 1.1 Various strategies for plate-based interaction screening. The ectodomains of cell surface receptors are produced as soluble recombinant form fused to biochemical tags for immobilisation, oligomerisation and detection by enzymatic activity. Immunoglobulin Fc-fusions (Fc) and the pentamerising domain of rat cartilage oligomeric matrix protein (COMP) mediate oligomerisation for increased avidity. Alkaline phosphatase (AP), horseradish peroxidase (HRP) and β-lactamase provide enzymatic activity for detection with the relevant substrates. AP and Fc tags also facilitate immobilisation onto α-AP antibody or Protein A-coated plates. Biotinylation enables capture and clustering on streptavidin-coated plates.

(Dscam) gene as alkaline phosphatase (AP)-tagged and immunoglobulin Fc-tagged constructs (Wojtowicz et al., 2007). AP-tagged ectodomains were captured on an anti-AP antibody coated 96-well plate before incubation with Fc-tagged ectodomains. Fc-fusion proteins are generally expressed as homodimers due to the disulfide bond in the hinge region of the Fc domain, increasing its avidity as compared to a monomeric probe (Czajkowsky et al., 2012). After washing to remove unbound Fc-fusion proteins, any remaining Fc-tagged ectodomains were detected with an anti-human IgG antibody conjugated to horseradish peroxidase (HRP), and detected with an appropriate HRP substrate. This study showed that a large number of DSCAM isoforms interact in a homophilic manner, providing a mechanistic explanation for how they mediate self-avoidance during neural circuit formation.

In contrast, Bushell *et al.* expressed the ectodomains of 110 zebrafish immunoglobulin (Ig) proteins as biotinylated 'baits' and pentameric 'preys' (Bushell et al., 2008). Both constructs contained a fusion of the ectodomains to the third and fourth Ig domains of rat Cd4 as a carrier. Carrier domains fold independently of other domains and can boost ex-

pression of a recombinant protein (Brown and Barclay, 1994). Baits contained an additional 17-amino acid biotinylation peptide, which becomes biotinylated when co-expressed with biotin ligase (BirA) from *E. coli*, whilst preys were additionally fused to the pentamerisation domain of rat cartilaginous oligomeric matrix protein (COMP) and the enzyme β-lactamase, separated by a flexible protein linker sequence (-LDRNLPPLAPLGP-). Streptavidin-coated plates were used to capture biotinylated baits. Interactions were detected by incubating with pentameric preys and captured preys were detected using a nitrocefin hydrolysis assay. This strategy, termed Avidity-based extracellular interaction screening (AVEXIS), eliminates an antibody incubation and wash step needed in the Dscam screen, thereby shortening the screen duration and reducing the loss of low affinity interactors during wash steps. The use of pentamers rather than dimers for prey proteins could also facilitate the detection of low-affinity interactions, although no systematic comparison has been conducted.

More recently, Ozkan *et al.* (2013) investigated the interactions between 202 *Drosophila* neuronal proteins by expressing them as Fc-fusion baits and pentameric, AP-tagged preys. The 202 proteins included members of the immunoglobulin superfamily (IgSF), fibronectin type III (FnIII), and leucine-rich repeat (LRR) families. Fc-fusion baits were captured using Protein A-coated plates, whilst interactions were detected by assaying for phosphatase activity using colourigenic BluePhos phosphatase substrate. Again, the benefit of oligomerising both bait and prey proteins was not explicitly demonstrated, although it could ostensibly promote the detection of very low affinity interactions. However, it could be argued that capturing baits on streptavidin also clusters them, providing little benefit over producing biotinylated monomeric baits. This study uncovered a set of previously unknown interactions between a 20 member subfamily of defective-in-proboscis-response IgSF proteins, and showed that they selectively interact with an 11 member subfamily of previously uncharacterized IgSF proteins.

An important similarity between interaction screening studies described above is the investigation of protein families with single, contiguous ectodomains containing distinct structural domains. To produce soluble ectodomain fusions that retain the ability to bind their native ligands, these ectodomains should consist of a single polypeptide chain that is able to fold independently of the transmembrane region. This presents difficulties for investigating multimeric receptors or integral membrane proteins with more than one transmembrane domain, as ligand binding interfaces may be formed from non-contiguous ectodomains, even within the same protein. Integrins are a prime example of heterodimeric receptors, as they are composed of an α and β chain. At least 18 α subunits and eight β subunits have been identified in humans, which are able to generate 24 different integrins with distinct binding

specificities (Stupack and Cheresh, 2002). Although another study has tried to address this by constructing different vectors to express tagged ectodomains of heterodimeric receptors and proteins of different membrane topologies (Sun et al., 2012), identifying extracellular interactions involving multimeric or structurally unstable binding sites on a large scale remains a challenge.

1.1.3 Interaction screening using cDNA overexpression

Ectopic surface expression of functional multipass membrane proteins is routinely achieved by transient transfection of mammalian cells with full-length cDNA constructs. The expression of transmembrane proteins in their native context increases the likelihood of having endogenous glycosylation patterns as well as the proper extracellular domain conformation. Expression cloning is a method of identifying genes that code for proteins with a specific attribute, and has been used to identify the receptors of numerous hormones and other secreted ligands (Simonsen and Lodish, 1994). This strategy involves creating pools of cDNA clones from cell or tissue lysate, transfection into mammalian cells, and selection of pools based on binding to a ligand of interest. cDNA clones from that pool are expanded and split into further pools which undergo the same process until the binding phenotype can be attributed to one or a few transcripts which are identified by sequencing. Instead of pooling, cells with binding phenotype can also be selected by cell sorting, separation with magnetic beads, or 'panning' approaches. However, the construction of good cDNA libraries from cells is technically challenging and can be heavily biased towards shorter or incomplete inserts. In addition, cell surface receptors are generally not highly expressed in cells, and as a result large fractions of the cDNA library would be occupied by transcripts coding for cytosolic housekeeping proteins that were not of interest. Although strategies have been developed to deplete expression cloning libraries of unwanted transcripts, this is challenging and not always fully effective. This means not only that a lot of primary material would be needed to capture transcripts of interest, but also that a large number of cells have to be screened in order to detect one or two positive events (Simonsen and Lodish, 1994).

After the sequencing of the human genome, large collections of individual clones containing defined cDNA transcripts became available, avoiding the need to create cDNA libraries from transcripts expressed in particular cell lines or tissues. Transfected cell microarrays were developed to utilise this resource for interrogating the function of genes in a systematic, unbiased, and high-throughput fashion (Ziauddin and Sabatini, 2001). For extracellular interaction screening, clusters of cells expressing different cell surface receptors

are formed by printing nanolitre volumes of plasmids and transfection reagent on glass slides in a known layout, then covering the slides with mammalian cells in medium. The glass slides are then incubated with a fluorescently-labelled ligand of interest and examined for clusters of cells which show increased ligand binding. Since each cluster overexpresses a single receptor, the positions of ligand-binding cell clusters would indicate the identity of the receptor. This approach circumvents technical challenges associated with constructing cDNA libraries from RNA, including overrepresentation of cytosolic housekeeping proteins and bias towards shorter inserts. In addition, sensitivity is increased from concentrating the signal in a localised region as compared to pooled expression cloning selection where only few cells in a well might show increased ligand binding. Known multi-subunit receptors can also be screened using this method by co-transfection of plasmids encoding individual subunits. The use of transfected cell microarrays for identifying important receptors involved in infection and metabolism (Mullican et al., 2017; Turner et al., 2013) highlights the utility of cell-based overexpression assays for more comprehensive screening of proteins with complex membrane architectures.

Nonetheless, transfected cell microarrays have yet to be widely adopted for extracellular interaction screening in individual laboratories due to high set-up costs needed in terms of procuring a comprehensive set of plasma membrane protein cDNA clones and hardware needed for arraying reagents. Instead, transfected cell microarray screening is typically provided as a service by biotechnology or pharmaceutical companies. Thus, an approach for large-scale extracellular interaction screening that involves low set-up costs and little specialised equipment could provide more flexibility for researchers to fine-tune their assays for studying any ligand of interest. Such a screening platform should be able to test for interactions against ideally all cell surface receptors encoded in the human genome, and be sensitive enough to detect low-affinity interactions with micromolar K_{D}.

1.2 CRISPR/Cas9-based transcriptional activation

Recent developments in CRISPR/Cas9 technology have allowed for specific transcriptional activation of target genes on a genome-wide scale. Cas9 is an RNA-guided DNA endonuclease functioning alongside Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) in bacteria to recognise and cleave foreign DNA elements (Sapranauskas et al., 2011). Since its discovery, this system has been adapted to manipulate eukaryotic genomes and transcriptomes by delivering wildtype or modified versions of Cas9, along with

a single guide RNA (gRNA), into eukaryotic cells. Each gRNA contains a 20 nucleotide (nt) long guide sequence complementary to a target genomic locus upstream of a Protospacer Adjacent Motif (PAM), and also encodes a scaffold for forming a complex with Cas9. Nuclease-deficient Cas9 (dCas9) fused to transcriptional activator domains can also complex with gRNA, and result in transcriptional upregulation when directed to the promoter regions of genes in a process termed CRISPR activation (CRISPRa).

For high-throughput gain-of-function screening, complex pools of different 20 nt guide sequences can be easily synthesised to form gRNA libraries targeting multiple promoter regions. This overcomes some of the technical challenges associated with using cDNA libraries for overexpression screening, such as restrictions on insert size. Moreover, each 20 nt guide sequence doubles up as a unique barcode that can be detected by next generation sequencing (NGS), and multiple gRNAs targeting the same promoter provide degeneracy in the library as well as a greater ability to distinguish between biological effects and gRNA-specific artefacts. Thus, the ability to overexpress virtually all receptors in the genome regardless of transcript length and assay for binding in the context of the plasma membrane makes CRISPRa an attractive potential platform for extracellular interaction screening.

1.2.1 Origins as a bacterial adaptive immune system

CRISPR loci were first discovered in bacteria and gained interest due to the unusually regular repetitions of DNA sequences separated by short, regularly-sized 'spacer' sequences, that occur throughout the loci. The importance of CRISPR arrays, along with flanking CRISPR-associated (Cas) genes, to the survival of prokaryotic cells is underscored by their presence in approximately 40% and 90% of sequenced bacterial and archaeal genomes respectively (Sorek et al., 2008). Spacer sequences were found to originate from bacteriophages or plasmids, and to mediate sequence-specific cleavage of these foreign genetic elements by Cas effector proteins (Mojica et al., 2005). Several types of CRISPR systems have been discovered and classified based on their effector modules and signature Cas proteins. Of these, Cas9 is a single effector protein characteristic of Class II, Type II CRISPR systems, which means that it performs gRNA binding, genomic DNA binding, and nuclease activity with a single protein. In contrast, Class I CRISPR systems are characterised by multi-subunit effector complexes where nucleic acid binding and nuclease activity are performed by separate Cas proteins (Makarova et al., 2015). Although the CRISPR/Cas9

system was not the first CRISPR system discovered, the simplicity of having a single effector like Cas9 has made it the system of choice for heterologous use.

For genome editing in mammalian cells, both Cas9 and at least one guide RNA (gRNA) must be introduced into the cell. The gRNA provides targeting specificity through complementary base-pairing between the target genomic loci and a 20 nt region within the gRNA scaffold. The rest of the scaffold forms a complex with Cas9 and samples PAM-containing sites in the genome by random diffusion/collision (Sternberg et al., 2014). PAMs differ between species, although the most commonly used Cas9 proteins come from *Streptococcus pyogenes* and recognise a 5'-NGG-3' motif. Upon reaching a site which is complementary to the guide sequence, Cas9 initiates a double strand break just upstream of the PAM (Figure 1.2A). In most cells, this double strand break is repaired by non-homologous end joining (NHEJ), resulting in small deletions or insertions ranging from 1-15 nt at the double strand break (Brandsma and Gent, 2012). Thus, targeting the Cas9-gRNA complex to coding exons of a gene often results in a frameshift mutation and functional knock-out of the encoded protein.

1.2.2 Transcriptional activation with CRISPR/Cas9

Cas9 contains two conserved nuclease sites, HNH and RuvC, for cleaving both strands of genomic DNA. The HNH domain is responsible for cleaving the target DNA strand (complementary to gRNA) whilst the RuvC domain cleaves the non-target DNA strand (Jinek et al., 2012). Inactivating mutations to either one of the nuclease domains create Cas9 nickases, which can only catalyse single-stranded DNA breaks and can be useful for promoting homologous recombination with a supplied donor template. A fully inactive Cas9 enzyme (dCas9) with mutations to both nuclease domains is devoid of nuclease activity but retains its RNA-guided DNA binding activity, transforming it into a nucleic acid-programmable DNA-binding protein. The *S. pyogenes* dCas9 (D10A/H840A) mutant is most commonly used. When recruited near a transcriptional start site (TSS), dCas9 blocks transcription and elongation, resulting in a reduction in gene expression. However, the addition of tethered activation domains like VP16 to dCas9 enabled transcriptional upregulation when targeted to regions upstream of the TSS (Maeder et al., 2013) (Figure 1.2B). VP16 is a transcription factor from the herpes simplex virus involved in the expression of viral immediate-early genes. *In vivo*, VP16 binds DNA indirectly through host factors Oct-1 and human factor C1 (HCF), and initiates transcription by recruiting appropriate factors through a C-terminal transcriptional activation domain. The VP16 activation domain

interacts with numerous host factors including basal transcription factors and the Mediator complex, which then recruits RNA polymerase II. VP16 also binds histone modifiers such as the SAGA and NuA complexes to promoters, causing chromatin decondensation detectable by fluorescence microscopy (Tumbar et al., 1999). In dCas9 fusion proteins, the VP16 activation domain is commonly present in four copies for enhanced activity, and referred to as VP64.

An issue with simple dCas9-VP64 CRISPRa systems was that many gRNAs did not induce activation on their own, but only when combined with other gRNAs targeting the same locus. Furthermore, targeting of endogenous genes gave modest levels of activation, with most displaying under ten-fold difference at mRNA level (Gilbert et al., 2013). An early attempt to boost gene induction by adding an N-terminal VP64 fusion, creating a dCas9 enzyme flanked by two VP64 domains, showed modest increase in efficacy of about four-fold compared to using single C-terminal dCas9-VP64 fusions for a single gene, Myod1 (Chakraborty et al., 2014). The need for multiple gRNAs targeting the same loci for significant mRNA upregulation severely limited the use of CRISPRa, especially for high-throughput pooled screening applications where each cell receives only one gRNA and large effect sizes are desirable for better signal-to-noise ratios.

Synergistic activation domains facilitate transactivation

Second generation CRISPRa systems sought to address this issue by using synergistic combinations of activator domains or chromatin modifiers. To mimic the coordinated recruitment of transcriptional machinery *in vivo*, Chavez *et al.* screened a series of candidate effectors for activation of a fluorescent reporter in HEK293 cells and selected the three most active activation domains (VP64, Rta, and p65) to create a tripartite activator fused directly to dCas9. Along with the respective gRNAs, this dCas9-VPR fusion was capable of inducing mRNA levels of more than 100-fold greater than that with dCas9-VP64 alone (Chavez et al., 2015). Rta (Replication and Transcription Activator) is an immediate-early gene product found in murine gammaherpesviruses. The endogenous function of Rta is to activate viral lytic genes by binding to a 27 bp RTA-responsive element (RRE) in the promoter regions of target genes. The transactivation domain of Rta has been compared to VP16 in competition assays and shown to activate genes by distinct mechanisms than VP16, suggesting that these domains can synergise (Hardwick et al., 1992). The third activator domain comes from p65 (also known as RelA), a human transcription factor and member of the NF-κB family. Accordingly, p65 is involved in innate and adaptive immune responses, and can be

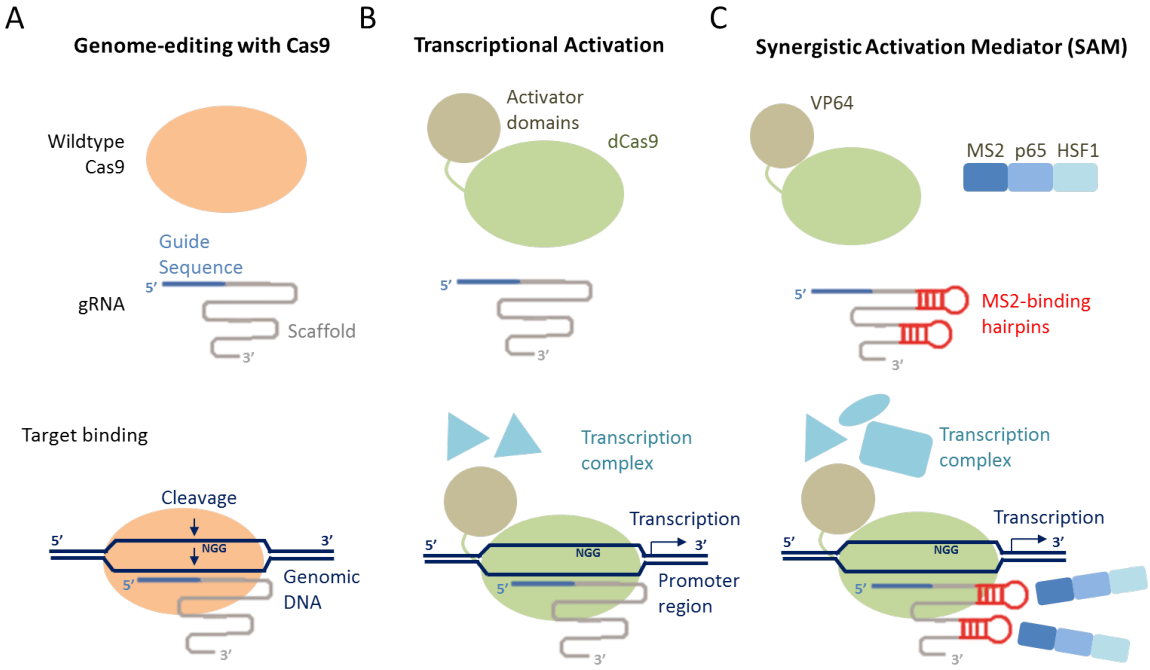

Figure 1.2 Original and modified CRISPR/Cas9 systems. A) Nuclease-active Cas9 complexes with gRNA and is targeted to genomic loci complementary to a 20 nt guide sequence at the 3' end of the gRNA and upstream of a PAM sequence (NGG). Upon binding, Cas9 catalyses a double strand break upstream of the PAM sequence. B) Nuclease-inactive Cas9 (dCas9) is fused to an activator domain like VP16 and retains gRNA-mediated targeting to genomic loci upstream of a coding gene. Binding of dCas9-activator results in recruitment of transcription initiation complex and subsequently transcriptional activation of downstream gene. C) Second generation dCas9-activator system SAM utilises synergistic activation domains from VP64, p65 and HSF1 to achieve high levels of transactivation from a single gRNA. MS2-fused p65 and HSF1 transactivator domains are recruited to MS2-binding hairpin loops engineered into the gRNA scaffold.

activated by proinflammatory factors TNF-α and IL-1. The transactivation domain of p65 interacts with members of the basal transcription complex such as TATA-binding protein and Transcription factor II B. Notably, p65 recruits transcription factors distinct from VP16, such as Activator protein 1 and the cAMP response element binding protein (CREB) family (van Essen et al., 2009).

In contrast, Konermann *et al.* opted for an engineered gRNA scaffold containing MS2-binding hairpin loops to recruit additional transactivator domains from p65 and HSF1 to the dCas9-VP64:gRNA complex (Figure 1.2C). The resulting system, called Synergistic Activation Mediator (SAM), demonstrated higher levels of transcriptional activation using a single guide compared to using multiple guides without the engineered scaffold (Konermann et al., 2015). HSF1 is a human transcription factor and the major mediator of the heat shock response. HSF1 induces transcription of heat shock proteins in response to environmental stress through a 150 amino acid long C-terminal region containing two transcriptional activator domains TAD1 and TAD2. TAD1 interacts with TATA-box binding protein-associated factor TAF9, as well as members of the histone remodelling SWI/SNF complex, whilst little is known about the binding partners of TAD2 (Dayalan Naidu and Dinkova-Kostova, 2017).

Rather than using different activator domains, Tanenbaum *et al.* sought to boost CRISPRa efficiency by recruiting multiple copies of VP64 to a repeating peptide array using single chain variable fragment (svFc) antibodies. A svFC consists of the epitope binding regions of the light and heavy chains of the antibody, fused to form a single polypeptide. Unlike conventional antibodies which generally do not fold properly in the cytoplasm, svFCs have been successfully expressed in soluble form in cells. dCas9 was fused to a peptide array containing 10 copies of a short peptide epitope, and expressed along with the corresponding svFC fused to VP64. This strategy was able to increase transcription of endogenous *CXCR4* by 10 - 50 fold of that inducd by dCas9-VP64 alone (Tanenbaum et al., 2014). Importantly, all three second generation CRISPRa sytems demonstrated the ability to achieve robust upregulation of mRNA abundance using a single gRNA. A systematic comparison of the three second generation CRISPR activation systems confirmed the increased efficiency of second generation systems over dCas9-VP64, and found that the SAM system from Konermann *et al.* was the most consistent in delivering high levels of upregulation across multiple genes in HEK293T cells (Chavez et al., 2016).

Epigenetic modifiers promote transcription and increase deposition of active chromatin marks

Besides transactivating domains, CRISPRa has also been achieved by directly fusing chromatin modifiers to dCas9. Although transcriptional activators like VP16 and Rta recruit chromatin modifiers as part of their transactivating mechanism, fusion of a single histone acetyltransferase domain from p300 (p300core) to dCas9 was shown to induce a significant increase (between 7 to 200 fold) in gene expression of *MYOD*, *OCT4* and *HBD* from promoter regions, as well as both proximal and distal enhancers up to a distance of 46 kb from the TSS (Hilton et al., 2015). No increase in expression was observed when dCas9-VP64 was used and targeted to enhancer regions. Strikingly, the use of dCas9-p300core, but not dCas9-VP64, resulted in increased levels of histone acetylation, specifically of the 27^{th} lysine residue of histone 3 (H3K27ac), at promoter regions of the human β-globin locus when an associated enhancer was targeted. p300 is a transcriptional coactivator first described through its interactions with adenoviral protein E1A (Chan and La Thangue, 2001). This interaction causes a loss of cell cycle control, and inactivating mutations in p300 have been described in several types of cancer, indicating a tumour suppressor role. p300 contains a core histone acetyltransferase domain flanked by two transactivation domains (Chan and La Thangue, 2001). A bromodomain within the acetyltransferase domain recognises acetylated residues and facilitates the transfer of an acetyl group from acetyl-CoA to the ε-amino group of a lysine residue (Dancy and Cole, 2015). Histone acetylation is a key mechanism in regulating transcription and is generally associated with euchromatin and actively transcribed promoters.

1.2.3 Genome-scale gain-of-function screening using CRISPR activation

Second generation CRISPRa systems have been used for genome-scale gain-of-function screening to identify genetic factors influencing cell growth and sensitivity to Cholera-diphtheria Toxin (CTx-DTA) in human myeloid leukemia K562 cells, as well as resistance of the A375 melanoma cell line to BRAF inhibitor vemurafenib treatment (Gilbert et al., 2014; Konermann et al., 2015). In the cell growth screen, the fraction of cells expressing gRNAs and the CRISPRa system was stable over the course of the experiment, indicating that there was no general toxicity associated with the CRISPRa platform used, and that overexpression with CRISPRa was specific. In addition, many of the genes which inhibited growth in the screen had previously known functions in regulating cell cycle and differentiation, including tumour suppressor genes, transcription factor families

involved in differentiation, and genes involved in the negative regulation of mitosis. Similarly, BRAF inhibitor resistance screening identified a number of gene candidates that confirmed known resistance pathways from previous knockout and knockdown screens. For instance, reactivation of the ERK pathway is one of the more well-studied resistance mechanisms, and components of this pathway were enriched in the gain-of-function screen. Taken together, these results highlight the feasibility and utility of CRISPRa for genome-scale gain-of-function screening.

Guide RNA library design

A key component of pooled CRISPRa screening is the gRNA library used. Since the gRNA contains a 20 nt guide sequence responsible for targeting the dCas9-activator complex, the library of gRNAs used defines the search space of the screen and can be designed to target all known promoters in the genome or focused on a subset of genes. Typically, five to ten guides are designed to target an individual locus, as guides vary in their efficiency and off-target profile (Haeussler et al., 2016). This degeneracy provides a buffer for ineffective guides that might have inadvertently and unknowingly been included in the library, and facilitates the differentiation of biological signal from technical artefacts generated by off-target effects of an individual guide. However, a trade-off between degeneracy and practicality exists because the scale of the screen increases with library complexity to maintain sufficient coverage of the library in a pooled screen. A meta-analysis of CRISPR knockout libraries suggests that six guides per gene represents the best trade-off as the amount of information provided by additional guides past six drops rapidly (Ong et al., 2017). As of yet, no similar analyses have been performed with CRISPRa libraries to determine if this can be generalised across different CRISPR screening modalities.

Clearly, any gRNA library would benefit from having a small number of highly active guides per promoter region. Therefore, it is essential to select the most efficient and specific guides to include in a library. The rules for gRNA design have been largely determined for CRISPR knockout systems, but research is underway to determine if these rules apply to CRISPRa as well. A small survey of off-target modifications using nuclease-active Cas9 and 26 gRNAs found that whilst the ratio of off-target modification to on-target modification frequency was generally low (<1%), sites that accounted for 88.3% of off-target modification contained up to four mismatches to the guide sequence (Haeussler et al., 2016). This suggests that when counting the number of potential off-target sites for a guide, sequences with up to four mismatches should be included. However, as dCas9 does not cause

double strand breaks and therefore toxicity, the number of potential off-target sites might be less relevant to CRISPRa as compared to whether they occur within the promoter region of another gene.

On-target gRNA efficiency is affected by gRNA expression levels and other criteria specific to the CRISPR modality used. gRNAs are transcribed by RNA polymerase III (RNAPIII), and in most expression vectors are driven by a U6 or H1 promoter. The murine U6 promoter favours transcription initiation on purine nucleotides (A and G) (Ma et al., 2014), so a common strategy for enhanced gRNA expression is to add a guanine nucleotide to the 5' end of 19 nt guide sequences. In addition, RNAPIII recognises a poly-T termination signal that causes catalytic inactivation and subsequent release of the polymerase from the transcript (Nielsen et al., 2013), suggesting that poly-T stretches of more than three thymine nucleotides within the guide sequence should be avoided.

In general, such factors affecting gRNA expression should be applicable to both CRISPR knockout and CRISPRa as both systems use similar gRNA expression vectors. However, other rules governing gRNA activity relate specifically to each CRISPR modality. For instance, targeting nuclease-active Cas9-gRNA complexes to early exons increases the likelihood of fully disrupting protein function and is thus favourable when designing CRISPR knockout gRNA, but is not as relevant for designing CRISPRa gRNA libraries. Instead, a machine learning algorithm trained on data from nine CRISPRa screens revealed that distance from the TSS and nucleosome positioning were major determinants of CRISPRa gRNA activity (Horlbeck et al., 2016). Fitting the positions of gRNA relative to TSSs using support vector regression showed a broad window of moderate activity between -500 to -50 bp, with a narrow peak between -100 and -250 bp where gRNAs are more likely to be highly active. Periodic patterns of highly active guides that were anti-correlated with nucleosome positioning also indicated an inhibitory effect of nucleosomes on the formation of CRISPRa complexes.

Consequently, designing gRNA libraries for genome-scale screening requires careful selection of parameters including the number of gene targets, number of guides per promoter, positional window for guide selection, and criteria for ranking or filtering guide sequences. Several genome-wide CRISPRa libraries are available and a quick review might provide some insight into the 'best practices' of gRNA library design (Table 1.1). First generation libraries were generally designed based on a combination of low numbers of off-target sites and proximity to the TSS (Gilbert et al., 2014; Konermann et al., 2015). Target gene lists used RefSeq predictions or APPRIS annotation to pick canonical isoforms. Better

Library Name	No. genes targeted	No. guides per gene	Targeting window (bp relative to TSS)	No. of guides (Total)
SAM v1	23,430	3	-200 to 0	70,290
CRISPRa	15,977	10	-400 to -50	198,810
CRISPRa-v2	18,916	5 or 10	-550 to -25	209,080
Calabrese	>18,000	3 or 6	-150 to -75	113,238

Table 1.1 Comparison of CRISPRa gRNA library specifications. There are currently four genome-wide CRISPRa gRNA libraries available. SAM v1 and CRISPRa libraries were first generation libraries designed predominantly based on distance to TSS, whilst second generation libraries like CRISPRa-v2 and Calabrese utilised more complex algorithms taking into account nucleosome positioning and improved TSS prediction. SAM v1 targets all human RefSeq coding isoforms whilst CRISPRa targets a more restricted set of genes that are expressed in human K562 cell line.

predictions of TSSs using FANTOM5 datasets were used in subsequent libraries, along with more sophisticated algorithms which take into account nucleosome positioning (Horlbeck et al., 2016). The most recent genome-wide CRISPRa library was designed following the same principles but selected guides from a much narrower window of 75-150 bp upstream of the TSS (Sanson et al., 2018).

In summary, CRISPRa is a promising alternative to cDNA overexpression for gain-of-function screening and has already been applied to investigate genetic factors underlying cell growth as well as resistance to toxins and drugs. Initial screens show no evidence of inherent toxicity associated with CRISPRa in human cancer cell lines and the identification of known factors suggest that the CRISPRa platform is active and specific. CRISPRa enables pooled screening without the need for barcoding individual constructs as short 20 nt guide sequences double up as barcodes which can be easily retrieved and counted using next generation sequencing (NGS). In addition, cDNA libraries are often limited by a maximum insert size, resulting in the omission of genes with long transcripts (> 3-4 kb) and isoform specification may be an issue for proteins that are not well studied. CRISPRa circumvents these issues by activating transcription from endogenous promoters, enabling upregulation regardless of transcript length and allowing cellular machinery to capture the full diversity of isoform expression. However, potential drawbacks of CRISPRa include limits on the levels of endogenous overexpression compared to cDNA overexpression, which is usually driven by a highly active CMV promoter. In addition, promoter regions of transcriptionally inactive genes may be packed away in chromatin and thus difficult for the CRISPRa complex to access.

Nonetheless, the advantages of CRISPRa screening merit its use alongside cDNA libraries for large-scale gain-of-function screening. This is supported by a side-by-side comparison of CRISPRa and cDNA overexpression screening that showed that both platforms yielded many common hits as well as distinct and complementary hits (Sanson et al., 2018).

1.3 Aims and objectives

As detailed above, extracellular interactions are challenging to identify and as a result are underrepresented in large protein interaction datasets. Nonetheless, the identification of key interactions governing important biological processes is of scientific and clinical interest. Current methods for large-scale screening of cell surface receptors to one or a few defined ligands are restricted to investigating certain classes of membrane proteins or require large investments in terms of cost and equipment. In particular, none of these methods enables genome-scale interrogation of all membrane proteins encoded in the human genome. For instance, the largest plate-based recombinant protein screen tested pairwise interactions of 249 proteins (Martin et al., 2010), whilst the largest available membrane protein cDNA library contains clones encoding 4,493 membrane proteins (Mullican et al., 2017), or an estimated 75% of the human surfaceome. Consequently, the development of more cost-effective and comprehensive approaches for large-scale extracellular interaction screening would facilitate the discovery of novel receptor-ligand pairs.

Recently, CRISPR/Cas9 technologies have provided a highly adaptable platform for genome-scale forward genetic screening. In fact, whole genome CRISPR knockout screening has been successfully applied to elucidate pathways required for cell surface signalling and detect novel extracellular interactions (Sharma et al., 2018). This strategy necessitates first screening of ligands against a panel of cell lines to identify a cell line which exhibits ligand-binding properties. In addition, although CRISPR knockout screening is designed to target all genes encoded in the human genome, it is in practice restricted to genes that are expressed in the particular cell line being used. CRISPR activation (CRISPRa), at least in principle, provides an attractive approach for screening against virtually all cell surface proteins in the human genome with a single cell line.

Therefore, the aim of this thesis was to adapt CRISPRa screening for extracellular receptor-ligand detection by establishing the best parameters to upregulate cell surface receptors using CRISPRa and constructing a CRISPRa gRNA library targeting membrane

proteins. To validate this approach I also apply it to screen for known antibody targets and endogenous interaction partners. Finally, to demonstrate the utility of the CRISPRa approach, I screened several members of the adhesion GPCR family and identified a set of novel interactions between Brain angiogenesis inhibitor 1 (ADGRB1) and three closely related myelin-associated inhibitory proteins (RTN4R, RTN4RL1 and RTN4RL2).