CHAPTER 1

INTRODUCTION

Communication between cells is crucial for the normal development and functioning of a multicellular organism. Membrane-compartmentalised cells receive instructional information from their surroundings by extracellular signalling cues which are often initiated via specific binding events made by plasma-membrane embedded receptors. While some extracellular signals are received via soluble factors (e.g. secreted proteins, hormones, autacoids, and neurotransmitters), others act through direct cell-cell interactions between specific receptors exposed on the outer surface of apposing cells [1]. Understanding of the molecular basis of cellular recognition events has wide implications as a multitude of cellular processes such as differentiation, motility and proliferation of cells depend on inter-cellular communication mediated by membrane receptors. Interactions between our own cells (e.g. neural and immunological recognition) as well as between host cells and pathogens also rely on extracellular recognition and signalling events. This, together with the fact that membrane receptors are directly accessible to systematically delivered biological reagents such as monoclonal antibodies [2] and small-molecule antagonists [3], make the study of membrane receptors therapeutically attractive. In fact, cell surface receptors currently make up the largest group (44 %) of human drug targets [4]. Thus, elucidation of cellular recognition processes is of significant interest for advances in biomedical research.

1.1 Molecules mediating cell surface recognition

A wide range of biochemically diverse molecules located in the plasma membrane of cells such as proteins, carbohydrates, and lipids have the potential to mediate cell surface interactions that are required for a vast range of biological processes [5]. In this section, I will review some aspects of each of these in turn.

1.1.1 Proteins

Nearly a quarter of the human genome encodes for either secreted or membrane-bound proteins [6, 7]. An intregral membrane protein that is embedded within the lipid bilayer of the plasma membrane is intrinsically amphipathic in nature. The domain exposed to the exterior of the cell (ectodomain) is hydrophilic, often glycosylated and is responsible for binding to the ligand, whereas the transmembrane region is hydrophobic to allow the receptor to exist within the plasma membrane. The intracellular domain of membrane receptors is in direct contact with the components of the cellular cytoplasm, which allows some membrane proteins to function as 'signal transducers' to transmit extracellular signals across the membrane to influence cellular behaviour in a context (a cell or signal type) dependent manner [8].

Membrane proteins that behave as signal transducers can be broadly categorised into three groups: Ion channels, G-protein linked receptors, (GPCR) and enzyme-linked receptors. Ion channels are usually multimeric and form an aqueous pore in the plasma membrane that, upon specific perturbation, allow movement of inorganic ions in and out of the cells [9]. Voltage-gated ion channels respond to change in membrane potential whereas transmitter gated channels respond to binding of neurotransmitters [10, 11]. The second class of signal transducers, GPCRs, are one of the most diverse receptors that can transmit diverse extracellular signals ranging from peptides, lipids, neurotransmitters and nucleotides to light, Ca^{+} and odorants [12, 13, 14, 15]. There are approximately 800 identified members in the GPCR superfamily in the human genome. Despite the large diversity of GPCRs, they interact with a relatively small number of heterotrimeric (composed of α, β and γ subunits) G proteins to initiate intracellular signalling cascades [16]. GPCRs are usually characterised by the presence of seven-transmembrane regions and this includes a class of approximately 140 GPCRs called 'orphan GPCRs' whose sequence is known but the endogenous ligands to which they bind are still largely unknown [17]. The third class of signal transducers are enzyme linked receptors which, unlike the GPCRs, are usually single transmembrane proteins that, instead of recruiting a G-protein for signalling activity, usually contain intrinsic enzymatic activity in their cytoplasmic domain or associate directly with enzymes with signalling capabilities. Six different classes of enzyme linked receptors have been described: (1) receptor guanylyl cyclases, (2) receptor tyrosine phosphatases, (3) receptor serine/threonine kinases, (4)

receptor tyrosine kinases, (5) tyrosine kinase-associated receptors, and (6) histidine-kinase-associated receptors [18].

1.1.2 Glycans

The second class of molecules that mediate cellular interactions are complex carbohydrates or glycans. This is one of the most diverse classes of macromolecules found in nature. Mammalian glycomes are predominantly created out of numerous combinations of nine common sugars (Glucose, N-acetylglucosamine, Galactose, N-acetylgalactosamine, Mannone, Fucose, Glucosamine, (or isomer L-Iduronic acid), Xylose, and N-Acetylneuraminic acid). Glycans can be found displayed at cell surfaces, incorporated into the extracellular matrix and covalently attached to secreted glycoproteins. Most eukaryotic cells are surrounded with a dense coat consisting of glycans and glycoconjugates (glycocalyx) that is important not only for providing protective, organisational, and barrier functions to the cell but also for mediating cellular communication [19]. Glycans can act as direct receptors for glycan binding proteins (GBP), which can be broadly classified into two groups: lectins and glycosaminoglycan (GAG) binding proteins. The origins of lectin, a term derived from the Latin word “legere,” meaning “to select”, dates back to 1888 when Herrmann Stillmark first described the animal red blood cell agglutination properties of extracts of castor bean seeds. Subsequently, lectins were identified in almost every plant species and today the stock of known lectins has increased vastly to include those from viruses (hemagglutinins), bacteria (adhesins and toxins), invertebrates and vertebrates. Lectins are usually characterised by the presence of evolutionarily conserved “carbohydrate recognition domains” (CRDs) [20]. This is in contrast to GAG binding proteins that are evolutionarily unrelated to each other and rather than possessing a specific binding domain, rely on basic residues to mediate interaction with the negatively charged sulphated groups of GAGs. An important consideration for the interaction of GBPs with their glycan ligands is the principle of multivalency. The binding affinity of a single glycan to GBP is low; thus to achieve biologically significant interaction, glycans that bind GBPs contain a multiple repeating structure for increased avidity [21].

1.1.3 Lipids

The third class of macromolecules that mediate cellular recognition and signalling processes are lipids. Although lipids have been suggested to be as

diverse as proteins, we have a much poorer understanding of their functions. In the past, lipids were mainly studied in the context of structural components of the plasma membrane or intermediary metabolites. The initial studies delineating the pathophysiological roles of specific lipid molecules (prostaglandins and leukotrienes) first suggested a role for lipids as intracellular signalling molecules [22, 23]. Plasma membrane lipids are mainly composed of phospholipids (phosphatidylcholine, phosphatidylethanolamine, phosphatidylserine, and sphingomyelin) and cholesterol. Since the identification of inositol phospholipids as a plasma membrane lipid crucial for cellular signalling [24], much research has been devoted to understanding the mechanism by which lipid molecules carry out signalling functions in cells [25]. The cellular lipid pool also includes glycans and proteins that are covalently modified by lipids. Both plasma membrane and intracellular organelle membrane glycans are modified by lipids to form glycolipids (e.g, glyceroglycolipids, glycosphingolipids, sulfoglycosphingolipids and gangliosides), which serve both as signalling molecules, recognition sites for cell–cell interactions [26, 27, 28], and receptors of bacteria and bacterial toxins [29, 30, 31].

Similarly, proteins are also post-translationally modified with lipids (e.g, myristoylation, palmitoylation, geranylgeranylation, GPI (Glycosylphosphatidylinositol) anchored) [32]. Some of these lipid-modified proteins, specifically those with a GPI anchor, cholesterol-linked and palmitoylated proteins such as hedgehog, are particularly enriched in the specific micro-environment of the plasma membrane; these are termed ‘lipid rafts’. Lipid rafts that contain defined sets of proteins are known to be important for signal transduction processes in cells [33].

Here, I will mainly focus on extracellular protein-protein and protein-glycan interactions including those mediated by glycoproteins, glycolipids, proteoglycans and lipid modified proteins. I will first discuss some of the challenges in studying interactions mediated by membrane receptors. Next I will introduce the major methods that have been designed to address some of these challenges and discuss their applicability and their limitations. Finally, I will introduce the recent genetic loss-of-function screening approaches including those using the CRISPR-Cas9 technology and their potential for studying cell surface interactions.

1.2 The challenges of studying extracellular ligand-receptor interactions

Even though extracellular interactions mediated by membrane receptors have been recognised to be of biological and pharmacological importance, investigating such interactions remains technically challenging. The inherent biochemical properties of extracellular space exposed membrane proteins, such as their relatively low abundance (typically 10^4 to 10^5 copies per cell) and amphipathic nature causing poor solubility, make them difficult to isolate, purify, solubilise, and biochemically manipulate [34]. Advances in approaches such as (Immuno)affinity-based techniques combined with mass-spectometry (MS) [35, 36, 37], yeast two-hybrid-based methods (Y2H) [38], and in-vitro array-based technologies [39, 40], now allow interrogation of protein-protein interactions at a large scale. However, such approaches are generally considered unsuitable to study extracellular interactions as the affinities of binding between two membrane receptor proteins are usually weak (K_D in μM to mM range) with fast dissociation rate constants (half lives of the order of one second). This poses challenges in detecting such interactions in approaches that require stringent wash steps such as affinity purification and MS. Moreover, membrane proteins are usually post-translationally modified with structurally critical glycans and disulfide bonds; this limits the use of prokaryotic heterologous systems or cell-free systems that lack the oxidative environment and the cellular machinary to generate correctly folded and glycosylated recombinant membrane proteins [41, 42].

Historically, the study of interactions between glycans and proteins has also been challenging because of the high diversity of glycans and the difficulties in obtaining glycans in high quantities in homogenous form [43]. Common methods to detect the low-affinity monovalent interactions between a single carbohydrate ligand and a single binding domain of a protein, such as inhibition studies using soluble mono- and oligosaccharides, isothermal calorimetry, surface plasmon resonance (SPR), and enzyme-linked lectin assays (ELLA), require large quantities of purified glycans with precise structures, which are not always readily available [21]. Unlike proteins, glycans cannot be readily cloned as they are secondary gene products and are not encoded directly in the genome. The particularly high diversity of glycans is the result of biosynthetic enzymes (glycosyltransferases, glycosidases, sulfotransferases, etc.) that

generate them, which act together in numerous combinatorial possibilities to generate highly heterogenous glycan chains [44]. These enzymes are dynamic and respond to environmental cues to act in a context specific manner, which makes it difficult to predict the exact nature of the glycan generated by cells. Glycans have also been found to have different roles in cell cultures compared to whole organisms. In the relatively simple environment of a cell culture system, genetic defects in glycans usually do not have severe biological consequences; however, the same defects have been shown to have major phenotypic consequence in a complex multicellular organism, in which glycans are involved in mediating important cell-cell and cell-matrix interactions [45, 46]. This has complicated the interpretation of the role of glycans from genetic studies.

1.3 Methods to study extracellular protein-protein interactions

1.3.1 General overview

Historically, a range of biochemical and genetic methods have been used to identify specific extracellular receptor-ligand interactions. Almost four decades ago, the receptors for platelet-derived growth factor (PDGF), epidermal growth factor (EGF) and insulin were all isolated using similar biochemical methods that involved co-purification of chemically cross-linked receptors from cells that interacted with radiolabelled (Iodine-125) ligand [47, 48, 49, 50]. Upon isolation of the target receptor, common approaches were to either (i) microsequence parts of the protein to generate nucleic acid probes to screen a complementary DNA (cDNA) library generated from the tissue of interest to isolate the cDNA that encoded for the receptor of interest or, (ii) generate antibodies from the purified protein to carry out immunoscreening of cDNA expression libraries (reviewed in [51]). cDNA expression cloning was commonly used especially in the field of cell surface receptor characterisation of lymphocytes, where cDNA clones were transiently transfected into cells and screened with monoclonal antibodies to "pan" for cells that express the receptor [52, 53]. The use of interaction blocking monoclonal antibodies has also led to discovery of extracellular interactions especially in the context of immune cell interactions [54, 55]. Monoclonal antibodies have also contributed majorly to the identification of virus receptors for human immunodeficiency virus [56, 57] and rhinovirus [58, 59].

Early methods also included direct cell-cell binding assays using cell types that either expressed the interacting molecules in their endogenous forms or that had been transfected to express the receptor (or ligand) of interest. This was first demonstrated with T-lymphocytes expressing CD2 cell surface marker that formed “rosettes” with erythrocytes that expressed CD58, the receptor for CD2 [60]. A variation on this when studying interactions mediated by non-erythrocytic cells, was the cell aggregation assay, which has been used to identify both hereophilic [61] and homophilic [62] interactions. Finally, genetic approaches have also been mainly in the context of neuronal cell interactions to identify receptors using inferences from phenotypes of specific gene targeted mutants in model organisms (*Caenorhabditis elegans* and *Drosophila melanogaster*) [63, 64, 65].

The completion of the human genome project was quickly followed by multiple studies that mapped the human membrane proteome to annotate proteins that have the potential to participate in extracellular interactions [66, 67, 68]. While the basic principles of the methods used in the post-genomic era have not changed dramatically compared to the conventional methods, the knowledge of the cellular secretome has allowed these methods to be high-throughput. Here I will highlight the key aspects of some of these post-genomic methods and their utility in understanding molecular mechanisms of cellular interactions. These will exclude the genetic knockdown or knockout screening approaches, which will be discussed separately in section 1.6.

1.3.2 Cell-based binding assays

Cell-cell binding assays

Extracellular receptor-ligand interactions can be studied by investigating direct cell-cell adhesion, in which two cell types that express different sets of membrane proteins are mixed together and the binding is detected with microscopy or by labeling with a radioisotope or fluorochrome [69, 70, 71]. In the recent years, efforts have been made to improve such assays in terms of their sensitivity and quantifiability. For example, in a proof-of-principle demonstration of one of the recent approaches, the low-affinity interaction between the adhesion receptors, JAM-B and JAM-C was detected using a cell–cell mixing experiment with cDNA transfected CHO cells expressing the receptors in a GPI anchored form [71]. It was suggested that the addition of the GPI anchor to the surface receptors aids their lateral diffusion, thereby facilitating

ligand-induced clustering. The cells were differentially labelled with DiI or DiD phospholipid binding dyes, which could then be analysed by flow-cytometry, such that interactions could be readily detected based on the fraction of cells containing signals from both dyes. While such approaches provide platforms to study interactions in the context of the plasma membrane, their results can be confounded by the potential for non-specific cell–cell clustering usually mediated by endogenous adhesion receptors whose expression can be altered during the course of the experiment (for example by transfection or signalling derived from the exogenously expressed genes).

A slightly different approach in this context has been the Baculovirus (BV) display system, in which a library of BV particles expressing the membrane protein of interest is used to screen for interactions (reviewed in [72]). For this, insect cells are transduced with BV encoding for a given membrane receptor and as the viral particle buds off from the cell membrane it will incorporate the overexpressed membrane protein. The virions can be used as 'nanoparticles' to screen for interaction with whole cells that express the interacting receptor [73]. This approach has been validated for low-affinity interaction study by detecting the interaction between BV particles displaying CD58, CD40 and glucocorticoid-induced tumor necrosis factor (TNF) receptor family-related protein (GITR) and cells expressing the respective ligands, CD2, CD40 ligand (CD40L) and GITR-ligand (GITRL), using a flow-cytometry based binding assay.

Cell-recombinant protein interaction assays

Recombinant proteins are common tools used in many methods designed to study cell surface interactions. A common theme in any approach that uses recombinant proteins is the purposeful oligomerisation of proteins, which generates avid probes, allowing detection of low-affinity interactions. Common approaches for this include tagging proteins with coiled-coil sequence from rat cartilage oligomeric matrix protein (COMP)–which causes spontanous protein pentamerisation or with human immunoglobulin-γ (Fc)-fusion or alkaline phosphatase (AP)-fusion proteins– which allows for protein dimerisation. Further multimerization approaches, for example, conjugating Fc-fusion tagged protein to protein-A microbeads and conjugating AP to anti-AP antibody to generate tetramers are also used to achieve even higher avidity [42]. Avid probes generated in this way can be used to screen for interactions with endogenous or

over-expressed receptors (for example by cDNA transfection) on the intact cell surface.

Libraries of cDNAs encoding a broad range of receptors have been widely used to study extracellular receptor-ligand interactions in the context of immune regulation, axon guidance mechanisms, 'de-orphanising' GPCRs, egg-sperm interaction, and host-pathogen interactions [74, 75, 76, 77, 78, 79]. In this method, cDNA expression vectors, with each vector expressing one receptor protein, are used to transfect cells to generate an 'expression library', which is then screened with avid probes to identify the receptor-ligand interaction. The clones of interest are then recovered for further analysis. In one of the comparatively earlier studies, the interaction between a transmembrane protein netrin-G1 ligand (NGL-1) and netrin-G1 was identified by screening 300 secreted or transmembrane (TM) human proteins expressed as Fc-fusion proteins with 400 putative cell-surface human proteins that were transiently expressed in COS7 cells [74]. A focused screen investigating interactions between TNF ligands and TNF receptors in mouse and human has also used a similar approach, in which cDNA clones expressing TNF receptors linked to a C-terminal GPI anchor sequence were individually transfected into HEK293T cells and and binding with soluble recombinant Fc and FLAG tagged TNF ligands was investigated using a 96-well format flow cytometry [75]. A slightly different iterative cDNA cloning approach was used to identify Juno as the oocyte receptor for Izumo1. In this method, a normalised mouse oocyte cDNA library was transfected into HEK293T cells in a pooled format and probed with pentamerised Izumo1. Transfected cells that bound the probe were selected and cDNAs from the cells were extracted and re-transfected into the cells in an iterative manner until the single cDNA clone that expressed Juno was identified [79].

A commercial group (Retrogenix) has designed a cell microarray technology using a library of cDNA encoding approximately 4500 membrane proteins. The approach used here is to spot cDNA together with the transfection reagent in glass slides and overlay mammalian cells such that cells are 'reverse-transfected' to overexpress a wide range of membrane proteins. Such libraries have been probed with avid reagents to identify low-affinity interactions [77, 78].

1.3.3 Cell-free protein interaction approaches

In the recent years, methods that use cell-free binding assays have been commonly used to study extracellular protein interactions [80, 81, 82, 83]. The

ability to generate large libraries of recombinant proteins has allowed for the development of methods that can map interaction networks in a high-throughput manner. In these approaches, soluble recombinant proteins that recapitulate the ectodomain of the membrane proteins and retain their binding properties are used. Recombinant proteins can be produced either with cellular systems or with In Vitro Transcription and Translation (IVTT) system. Here I will be predominantly focusing on techniques that use mammalian and insect cell expression systems because these systems are better suited to studying extracellular proteins as they are better at achieving correct folding, glycosylation and secretion of recombinant proteins compared to IVTT systems.

AVEXIS

The Wright lab has developed a method for detecting of low affinity protein-protein interactions, termed 'avidity-based extracellular interaction screen' (AVEXIS). In AVEXIS, a mammalian expression system (with HEK293 cells) is used to produce libraries of soluble ectodomains of membrane proteins that can be used for interaction screening. The proteins are tagged either with biotin ('bait' protein) or with COMP alongside beta-lactamase tags ('prey' protein). Bait proteins are captured on streptavidin-coated plates and the binding between bait and prey is quantified using the colorimetric beta-lactamase substrate nitrocefin.The method has been used extensively to characterise interactions in zebrafish neural system interactions. It was first used to identify receptor–ligand pairs within the zebrafish immunoglobulin superfamily (IgSF) [84] and later expanded to screen for interactions mediated by leucine-rich repeat (LRR) proteins, secreted factors and proteins from floor-plate microenvironment in zebrafish [85, 86, 87].

AVEXIS has also been used to identify interactions between malaria causing parasite *P. falciparum* and the human red blood cell. In two studies, recombinant parasite proteins were used to probe a library of soluble ectodomains that represented the red blood cell surface protein repertoire to identify host-pathogen interactions between basigin (BSG)- reticulocyte-binding protein homolog 5 (RH5) and Semaphorin7A-merozoite thrombospondin-related anonymous protein (MTRAP) [88, 89]. The interaction between BSG and RH5 was found to be an essential interaction required by all tested strains of *P. falciparum* in mediating invasion of red-blood cells. Recently, a variation of AVEXIS has been used to characterise interdependencies of the other parasite proteins that interact with RH5 and BSG to make up the RH5 invasion complex [90].

Glass-slide based microarray techniques

Several studies have compiled lists of extracellular proteins to generate recombinant protein libraries using mammalian and insect based cell expression systems to carry out high throughput screens in glass-slide based microarray format, in which recombinant proteins are captured on a glass slide and interactions are detected using multivalent probes. In this regard, one of the first demonstrations of protein-microarray approach was the successful detection of the low-affinity protein interaction between CD200 coupled to multivalent microbeads and its receptor CD200R [91] immobilized on epoxysilane-coated glass slides [92, 93]. Subsequently, this approach was used to screen an array of 1,334 proteins with 89 immunoglobulin superfamily (IgSF) receptor baits using bait-Fc fusion proteins bound on Protein A coated microbeads [83]. Recently, this method was also used to describe a virus-host extracellular interaction map using an array of 1,500 human proteins and human adenovirus (HAdV) encoded immunomodulatory protein baits [94].

AVEXIS itself has also been miniaturised to a microarray format [95], in which biotinylated bait proteins are directionally arrayed on streptavidin coated glass slides and probed with pentamerised probes tagged with FLAG tag ('DYKDDDDK' epitope) for detection using anti-FLAG antibody. This approach has been utilised to identify FcϵR1α as a receptor for PEAR1 from a ectodomain library representing the secretome of the human platelet [96].

1.3.4 Mass-spectrometry-based methods

Mass spectrometry (MS) based proteomics is a powerful tool used for the identification and quantification of peptides, but its application in the analysis of cell surface receptors has been challenging mainly because of the difficulty in obtaining homogenous plasma membrane protein regions, low abundance of membrane proteins, and technical difficulties in identifying hydrophobic regions of membrane proteins [97, 98]. This has been significantly improved in the past decade as several approaches have been described, which has allowed the use of MS in the identification of cell surface proteins [97, 98, 99, 100, 101]. That said, there are still very few mass spectrometry based methods that not only identify the cell surface molecules but also directly investigate low affinity interactions mediated at the surface of cells.

One such approach that was recently described is the ligand-based, receptor-capture (LRC) technology using TRICEPS [102]. TRICEPS is a chemoproteomic reagent that consists of three distinct sites: an amine-reactive site

for non-specific amine conjugation of purified ligand of interest; a protected aldehydereactive site for covalent conjugation with carbohydrate groups on glycoproteins under oxidising conditions; and a biotin site purifying the receptor peptides for identification by quantitative mass spectrometry. In this approach, ligands conjugated to TRICEPS are added to cells that have been previously exposed to oxidising conditions. Upon a stable ligand-receptor interaction, TRICEPS is covalently captured with nearby carbohydrates. After the reaction, the cells are lysed, trypsinized and the captured glycopeptides are enriched using the biotin tag and identified using MS. This approach was first validated by detection of known and novel interaction mediated by extracellular ligands of diverse nature such as peptides, glycoprotein, therapeutic antibodies and intact viruses. It has subsequently been commercialised (LRC-TriCEPS; Dualsystems Biotech AG) and used in other studies to identify cell surface interactions between secreted proteins and membrane receptors [103].

1.4 Methods to study extracellular protein-glycan interactions

1.4.1 General overview

Studies involving protein-glycan interactions have utilised techniques from genetics, structural biology, biochemistry, organic, and analytical chemistry. In the past, plant lectins were crucial tools in the field of glycobiology mainly because of their high abundance, easy sourceability and ability to mediate high affinity interactions (K_D of nM range) with monosaccarides. Early studies utilised the agglutination properties of lectins to characterise blood group antigens, antigenic determinants of which are specified by terminal sugar residues that are recognised by specific lectins. Plant lectins such as concanavalin A, L-phytohemagglutinin lentil lectin (LCA), and *Maackia amurensis* leukoagglutinin (MAL) have been used extensively in affinity chromatography to isolate major glycan structures present in animal cells [104]. Over the years, a wide range of methods including monoclonal antibody blocking, expression of glycosyltransferases by transfection with cDNA encoding specific enzymes, chemical and enzymatic manipulation of carbohydrates and direct binding with immobilized (glycan arrays and SPR based methods) or soluble (NMR based methods) glycans have been used to study interactions between glycans and proteins [105].

1.4.2 Binding inhibition approaches

In the past, the main method to identify interactions mediated by glycans with the protein of interest was through binding inhibition studies that used soluble probes such as monoclonal antibodies and purified mono-(oligo-)saccharides. One of the best characterised functional evidence for protein-glycan interactions is the low affinity interactions that allow for leukocytes to roll along the vascular surface. The first indications of such interactions were studied using the monoclonal antibody (MEL14) that specifically blocked the binding of leukocytes to the high endothelial venules (HEV). Similarly, monoclonal antibodies were also used to identify ELAM 1 (endothelial-leukocyte adhesion molecule), which was shown to be expressed on stimulated but not unstimulated human endothelial cells and gp140, which was expressed in activated platelets. These molecules were termed as selectins and are known today as L-selectin (MEL 14 antigen), E-selectin (ELAM 1), and P-selectin (gp140). Subsequent binding inhibition studies with monoclonal antibodies identified sialyl-Lewis^x (siaLe^x) as the glycan necessary for the interaction of P- and E-selectins with both fucose and sialic acid residues required for binding (reviewed in [106, 107]).

Chemical or enzymatic manipulation of glycans presented on the cell surface has also been a valuable tool for identification of interactions mediated by glycans [108]. Tunicamycin, a compound that inhibits N-glycosylation, has been used extensively for studying the role of N-glycans in mediating interactions. One common approach when studying interactions mediated by sulfated receptors is the treatment of cells with chlorate, which has been shown to inhibit the production of the high-energy sulfate precursor 3'-phosphoadenosine 5' -phosphosulfate (PAPS). This leads to the generation of undersulfated glycoproteins and proteoglycans in cells [109]. The ligand for L-selectin was determined to be a sulfated version of siaLe^x based on the decrease in binding to undersulfated glycoproteins.

Enzymatic deglycosylation is also a common method to study the role of glycans in mediating interactions. Common enzymes include those that selectively remove N-linked oligosaccarides (PNGaseF), sialic acid residues in an O-glycan (sialidase), and core-O-glycans (O-glycosidase) [110]. The first prediction of the nature of endogenous ligands for L-selectin was made on the basis of loss of binding of lymphocytes to cells treated with sialidase [107]. Similarly, neuraminidase (a type of sialidase) treatment of red-blood cells, which abolished the binding of the malaria causing parasite *P. falciparum*

protein EBA175 to these cells led to the discovery of the first host receptor, glycophorin A, for this parasite [111, 112].

1.4.3 cDNA expression methods

Another approach to studying glycan-protein interactions is to express the cDNA encoding a glycotransferase in a cell culture system and to detect the binding mediated by the new glycans (neoglycans) on the cell surface. This method has been used in studying interactions mediated by selectins and siglecs (sialic acid-recognizing Ig-superfamily lectins). In the example of P-selectin, transfecting cDNA encoding a human α-1,3-fucosyltransferase together with cDNA encoding P-selectin glycoprotein ligand (PSGL-1) into nonmyeloid cell line (CHO cell line) was shown to confer high affinity binding of P-selectin to these cells [113]. In another example, one of the early studies in identifying a receptor for the B cell adhesion molecule CD22 used a cDNA expression cloning approach, which identified a sialyltransferase required for adhesion [114]. Although the original study suggested the sialyltransferase enzyme to be the direct receptor, further biochemical experiments soon demonstrated that CD22 bound to a sialic acid residues through its extracellular domain, the production of which required the enzyme. From there on, homology based studies led to recognition of a new family protein called siglecs, which mediate binding in a sialic acid dependent manner [115].

1.4.4 Glycan arrays

Many past methods investigated the glycan binding ability of proteins in material-intensive assays (e.g. inhibition assays using purified glycans, lectins and monoclonal antibodies). This was changed when technological advances in isolating 'natural' glycans from sources such as cells, tissues, and pathogens, and generating synthetic glycans from chemical and enzymatic methods, allowed for the generation of glycan arrays [43]. The two largest mammalian glycan libraries compiled from natural and synthetic sources are currently from Consortium for Functional Glycomics (CFG) ($\sim$ 609 glycans) and Feizi and coworkers ($\sim$ 830 glycans). These arrays have been used to probe for interactions with plant and microbial lectins, glycan-binding proteins involved in the innate and adaptive immune system, glycan-specific antibodies, virus particles, and whole cells (reviewed in [116]).

1.5 Limitations of the existing methods

While the methods described above have been successful in the past, they have several limitations:

1. One of the main drawbacks of methods that require large libraries of cDNA or recombinant proteins is that the initial compilation of libraries containing hundreds of molecules is resource intensive, and most researchers' interests are usually focussed on a single or small number of proteins rather than the networks of interactions within a larger family.

2. The cDNA transfection approach of membrane proteins requires all expressed proteins to be transported on the surface of cells; however cell surface targeting of transmembrane proteins is strictly regulated and often requires cellular accessory factors such as transporters, chaperones and correct oligomeric assembly. Thus, transfecting a single cDNA might not be enough to achieve cell surface expression in many cases.

3. Recombinant protein based methods require that the receptor binding function is retained when expressed by heterologous cells out of the context of the plasma membrane as a soluble recombinant protein. Whilst this is generally the case for proteins that span the membrane once (single-pass (type I, type II) or GPI-anchored), it is not so for receptor complexes and membrane proteins that span the membrane multiple times. Therefore, interactions made by these latter classes of complexes and proteins are usually underrepresented. In addition, although care can be taken to produce recombinant proteins in heterologous systems that increase the likelihood of structurally critical post-translational being added properly, this is not always achieved. An example of this is the under-glycosylation of HEK-293 expressed recombinant glycophorin A, which does not retain its property to interact with its known *P. falciparum* interaction partner EBA175 [117].

4. While the use of recombinant proteins makes the manipulation of membrane receptors biochemically tractable, it is challenging to investigate other cellular components required for cellular recognition at the surface of the cell when interactions are studied in isolation between purified probes rather than in the context of a cellular system. In a cell, interactions occur in a complex environment, which includes contributions from

a charged glycocalyx of carbohydrates and lipids displayed on a dynamic membrane [118, 119, 120]. As a consequence of taking the reductionist approach of studying cell surface molecules as soluble recombinant proteins, the contribution from cellular factors other than direct receptors that ultimately contribute in cellular recognition is usually ignored.

5. Mass-spectrometry based methods such as LRC do allow for the interrogation of cell surface interaction mediated by endogenous receptors in the context of the plasma membrane but they still require chemical manipulation of the cell surface (e.g, oxidation), which can alter the biochemical nature of the molecules present on the surface of the cells. In addition, this method relies on the biological properties of the ligand being preserved during the process of non-specific amine conjugation. This is difficult to achieve for proteins that rely on lysine residues (residues where amine conjugation is normally done) to mediate interactions. Moreover, the method requires the receptor to be glycosylated, which is not always the case.

6. The success of glycan arrays depends on the representation of the glycans they contain. While there exist multiple efforts to define the 'glycome' of an organism using mass-spectrometry, lectin and antibody array based approaches, due to the very high diversity of glycans, the size of mammalian glycan array libraries available today still pales in comparison to the DNA libraries; such mammalian glycan array libraries may contain a fraction of those glycans present in nature.

7. Methods that tether glycans to a surface such as SPR, glycan array or ELISA-based methods limit the number of ways in which glycans can be presented, which can change the binding properties of the isolated glycans. On the surface of cells, glycans are presented on a glycoprotein or a glycolipid scaffold in diverse conformations, which can be critical for ligand recognition.

8. Methods that use isolated glycans require the glycan receptors to independently mediate binding with the ligand. This is not always the case as many glycoprotein backbones also participate in the interaction. The best known example of this is the interaction mediated by P-selectin, whose interaction depends both on the glycan and the adjacent sulfotyrosine residues in the glycoprotein receptor PSGL1 [121].

1.6 Loss-of-function genetic approaches to study cellular recognition process

1.6.1 General overview

Genetic mutations can alter cellular processes, therefore a method to study a function of the gene is to investigate the effect of its absence. Genetic mechanisms of multiple biological processes can be studied by analysing loss-of-function (LOF) mutants, in which the altered gene product lacks the molecular function of the wild-type gene. The underlying principle of a LOF approach is to ablate the function of a gene by targeting DNA, RNA or protein. LOF approaches range from non-targeted chemical mutagenesis approaches to generate mutants, to systematic large-scale generation of mutant libraries for genetic screens using genome-editing technologies such as short interfering (si) or short hairpin (sh) RNA, and more recently the CRISPR (clustered regularly interspaced short palindromic repeats)-Cas9 (CRISPR-associated protein 9) system (reviewed in [122]). These approaches can be applied both in a large-scale manner for the identification of novel factors in different cellular contexts and in a small-scale manner to explore the roles of few candidate genes usually involved in disease processes that are identified from large-scale screens. The pioneering works of genome-wide LOF screens carried out in *Caenorhabditis elegans* [123] and *Drosophila melanogaster* [124, 125] demonstrated how large-scale mutagenesis could be applied to assign gene function. These approaches have already been applied to many areas of biological and biomedical research including the investigation into cellular recognition events, which I will discuss in this section.

1.6.2 Study of naturally occurring mutants

Traditional loss-of-function analysis involved characterising naturally occurring mutants. For example, in the context of host-pathogen interactions, blood-group polymorphisms that occur in the human population were used to identify receptors required for invasion of host cells by malaria causing parasites *P. falciparum* and *P. vivax*. One of the best described examples of this is the identification of the mutation that abolishes the expression of the Duffy antigen receptor of chemokines (DARC) on the surface of red blood cells (the so-called 'Duffy-negative phenotype') in western and central Africa, which was shown to confer almost complete protection from *P. vivax* infection [126, 127]. This

led to the identification of DARC as a critical receptor for *P. vivax* invasion into red-blood cells and the subsequent identification of its parasite binding partner, Duffy binding protein (DBP). Untill today this receptor-ligand interaction remains the only known host-pathogen interaction in the context of *P. vivax* infection. Similarly, the studies characterising the host receptors for *P. falciparum* merozoite surface proteins have also utilised naturally occuring erythrocytes with mutations in blood group antigens. For example, the parasite-host interactions between EBA175 with glycophorin A (GYPA), EBA140 with glycophorin C (GYPC) and EBL1 with glycophorin B (GYPB) were studied using blood groups that lacked the respective receptors; En(a-) (absence of GYPA), Leach phenotype (absence of GYPC) and S-s-U-phenotype (absence of GYPB) (reviewed in [128]). Another well-known example of this is the study in individuals resistant to HIV infection, who were found to carry homozygous mutation in the cell surface receptor protein CCR5, which acts as the viral co-receptor [129]. While the study of natural mutants using biochemical or population genetics approach has been useful in the past, the difficulty in acquisition of mutants, the presence of rare alleles in a population, and the limited shelf life of biological materials makes this approach generally unsuitable for systematic investigation of cellular recognition processes.

1.6.3 Genetic screening approaches

Chemical and insertional mutagenesis

A different approach to study a gene function is to generate loss-of-function mutants using genome editing tools. The classical approach of mutant generation was by random mutagenesis through chemical mutagens (e.g, ethyl methanesulfonate (EMS)), that can introduce a variety of genetic lesions that are expressed as complete or partial loss of function of the gene product [130]. Such approaches were commonly carried out in *Saccharomyces cerevisiae*, which served as an ideal model organism because of its rapid generation time and its haploid genome, which permitted efficient generation of homozygous mutants for the study of recessive phenotypes. This approach served as a fast and effective method to generate a large number of mutants so that they could be screened for the phenotype of interest. Genetic screens in yeast using temperature sensitive mutants served as powerful tools to study cell essential mechanisms. The early works in yeast temperature sensitive mutants, which identified components of the SEC genes [131, 132], has been extremely valu-

able in the field of membrane protein biology as they have shaped our current understanding of the general secretory pathway that is required for the transport of the majority of membrane and secreted proteins. Similarly, chemical mutagenesis has also been used to mutate the genome of a hermaphrodite model organism *Caenorhabditis elegans*, which has allowed study into, among other cellular processes, the genetic control of neuronal development. For example, the molecular basis of UNC-6/Netrin signalling in axon guidance was first identified by characterisation of mutants generated through random mutagenesis in this organism (reviewed in [133]).

A different approach of random mutagenesis proceeds through insertional mutagenesis, which employs the strategy by which exogenous retroviral and transposable DNA can be inserted randomly into the genome such that if the insertions is in the coding or regulatory region of the gene, the gene product will be rendered non-functional. This approach has an advantage over chemical mutagenesis as it facilitates the identification of the mutated gene as the inserted DNA, whose sequence is known, serves as a molecular tag. While powerful, the mutagenesis techniques had limited use for the generation of homozygous mutants in mammalian cell culture systems mainly because of the diploid nature of mammalian cells, the lack of strategies to set up genetic crosses, and the low rates of homologous recombination. One of the strategies employed to address this limitation was to utilise mouse embryonic stem cell (ESC) lines that are deficient for Bloom's syndrome protein (BLM), in which cells exhibit a high rate of mitotic recombination to generate a genome-wide library of homozygous mutant cells from heterozygous mutations induced with insertional mutagenesis approach [134]. Furthermore, the discovery of haploid human cell lines, KBM7 [135, 136] and HAP1 (a derivative of KBM7) [137], and haploid mouse embryonic stem cells [138] further facilitated the use of insertional mutagenesis appraoch in mammalian cell culture system. The application of random mutations techniques in haploid cells can directly cause loss-of-function phenotypes, which can be studied in a high-throughput manner. Such approaches have been used in the context of host-pathogen interactions mediated by bacteria and viruses (reviewed in [139]). Examples with the gene trap[1] approach in haploid human cell lines include identification of host factors for bacterial toxins [140, 141, 142], intracellular receptor for the

[1] A type of insertional mutagenesis using a vector that contains a strong splice acceptor site, an efficient polyadenylation signal and a marker gene. The insertion of the vector into the intronic (or exonic) region leads to inactivation of the target gene.

Ebola virus [137], deciphering the glycosylation of the Lassa virus receptor α-dystroglycan (α-DG) [143], and receptor switching mediated by the Lassa virus upon virus internalisation [144]. More recently, the same approach has been used for the identification of the host receptor for adeno-associated virus (AAV) serotype 2 [145].

RNAi mediated approaches

During the time when the creation of homozygous mutant libraries for cells and organisms with diploid genomes was time- and capital-intensive and low throughput, RNA interference (RNAi) technology provided an alternative approach to allow high throughput gene silencing through sequence-specific targeting of mRNAs [146]. In this approach, RNAi reagents (such as synthetic siRNAs, siRNA precursors (short hairpin RNAs (shRNAs)), or long double-stranded RNAs (dsRNAs)) are introduced to cells or organisms via methods such as transduction, transfection, microinjection or, in case of *C.elegans*, by simply feeding organisms with *E.coli* expressing dsRNAs [147]. Once in the cells, siRNAs are incorporated directly into the RNA-induced silencing complex (RISC), whereas dsRNAs are processed by DICER to first generate siRNA. The siRNA-RISC complex can mediate gene silencing by cleaving the complementary mRNA (when the sequences are perfectly complementary) or by interfering with translation (when the sequences are partially complementary). The approach has allowed for the implementation of both small scale and genome-wide loss-of-function screens in human and Drosophila cell lines for the identification of genes and gene networks involved in signal transduction processes [148], identification of regulators of cell adhesion, [149] and extensively in host-pathogen interactions (e.g. colonisation of Drosophila cells by sinbad virus [150], dengue virus [151]; mammalian host factors required by hepatitis C virus (HCV) [152, 153], west nile virus [154], vaccinia virus [155] and multiple screens for host factors required by HIV (reviewed in [156])). Similar screens have also been carried out in hematopoietic progenitor cells that can be differentiated into erythroblasts, thereby allowing forward genetic screens to be carried out in the otherwise genetically intractable anucleate red-blood cells. A shRNA mediated genetic screen in such a system has identified CD55 as a crucial host receptor for *P. falciparum* invasion into red-blood cells, although the ligand on the parasite end is still unknown [157].

One of the biggest challenges of using RNAi as a tool to study gene function involves the sequence-specific off-target effects of siRNA. RNAi uses

an existing cellular pathway governed by endogenous microRNAs (miRNAs) that regulate cellular gene expression levels using incomplete complementarity between the miRNA and its target. Exogenously supplied siRNA can function as an endogenous miRNA and mediate target recognition by partial sequence complementarity, which can lead to translational repression and/or degradation of non targeted mRNAs. Such off target effects can confound the interpretation of the phenotypic effects and potentially create cellular toxicities [158]. In addition, the genetic perturbation using RNAi frequently results in incomplete silencing, which, combined with the off-target effects can lead to a decrease in sensitivity and inconsistent results. One such example of inconsistent results is the lack of overlap between three large-scale RNAi-mediated knockdown approaches conducted by independent laboratories for the identification of factors responsible for HIV infection, which have identified 842 putative factors out of which only 37 genes were identified in more than one screen and merely three genes identified in all three studies [159, 160, 161].

1.6.4 CRISPR-Cas9 approach

Introduction to CRISPR-Cas9

CRISPR-Cas (clustered regularly interspaced short palindromic repeats-CRISPR associated) systems are adaptive immunity strategies developed by bacteria and archaea to protect themselves against foreign genetic elements such as viruses and plasmids [162, 163]. The system involves three key steps: (i) spacer acquisition, during which small DNA fragments from foreign DNA elements are inserted into the CRISPR locus; (ii) crRNA (CRISPR RNA) expression, during which the CRISPR locus is transcribed to generate a long primary CRISPR transcript (the pre-crRNA); (iii) target interference, during which the target is detected and degraded by the crRNA and Cas9 protein(s). Depending on the architecture of the interference molecules, the system can be divided to two main classes, consisting of six types (type I-VI) and 16 subtypes. The CRISPR-Cas9 technology is derived from type II CRISPR-Cas system, which utilises crRNA, tracrRNA (trans-activating crRNA[2]), and a single large multi-domain effector protein (Cas9) to mediate target recognition and cleavage. Under this system, the Cas9 protein, which is an endonuclease, is loaded with RNA duplex (tracrRNA:crRNA) and this riboprotein complex is directed to the target site (i.e., the site complementary to the guide sequence

[2]These are short sequences that are complementary to the corresponding crRNA. They are required for Cas9 to correctly recognise the target DNA.

from crRNA), enabling the Cas9 protein to introduce double stranded breaks (DSBs) in the DNA. Target recognition by Cas9 also requires the protospacer adjacent motif (PAM), which is a species-dependent short (2-6 nucleotide sequences (nts)) DNA sequence immediately following the target DNA sequence (reviewed in [164, 165]). The majority of the current CRISPR-Cas9 systems use the Cas9 protein from *Streptococcus pyogenes* (henceforth referred to as SpCas9), which uses 5′-NGG and 5′-NAG (although at less efficient rates) as the PAM sequence (figure 1.1).

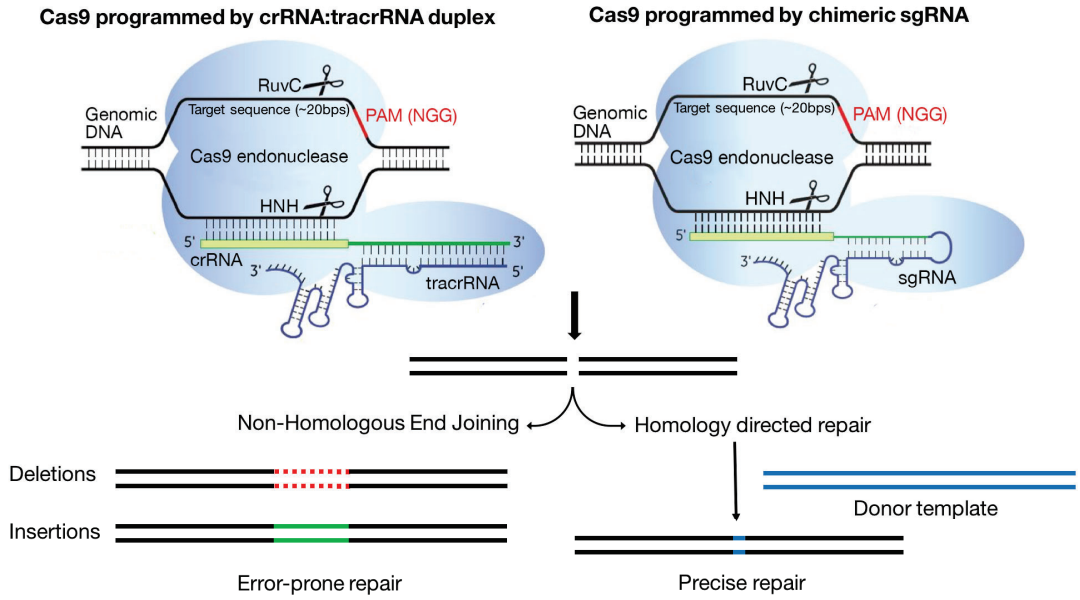

Fig. 1.1 Overview of type II CRISPR-Cas9 mediated gene editing. Cas9 can be programmed either with RNA duplex generated from crRNA and tracrRNA or with a single chimeric RNA, which mimics the RNA duplex. Once loaded with the RNA duplex or the chimeric RNA, the Cas9 protein is targeted to the region of the genome that is complementary to the 20 nts sequence of the crRNA (or the gRNA) where it mediates double stranded DNA cleavage. The Cas9 protein contains two endonuclease domains, the HNH and RuvC domains that either cleave the strand complementary to the guide sequence, or the strand matching the guide sequence. Once a double stranded break is generated the cellular DNA machinery repairs it either with the NHEJ pathway or the HDR pathway (if homologous donor template is present). NHEJ pathway is error-prone and can lead to generation of indel mutations, leading to inactivation of the gene [165]. Figure adapted from [166].

Gene targeting using the CRISPR-Cas9 system

One of the major findings in the field of CRISPR-mediated gene editing has been the observation that the tracrRNA:crRNA duplex can be engineered as a single piece of chimeric RNA (termed as single guide RNA (sgRNA) or gRNA) [167]. Immediately following this observation, CRISPR/Cas9 system

was adapted for genome engineering in human cells, where by changing the 20 nucleotide guide sequence of the gRNA, the DNA sequence of interest could be efficiently targeted by the Cas9 endonuclease to generate double stranded breaks (DSBs) [168, 169]. DSBs in cells are repaired using two major pathways: (i) the error-prone Non-Homologous End Joining (NHEJ) pathway; and (ii) the high-fidelity Homology Directed Repair (HDR) pathway. In higher organisms, the NHEJ pathway, which does not require specific sequence for ligation of DNA, is the major method for repairing DSBs and unlike HDR, this pathway is active during all stages of the cell cycle [170]. DNA repair using the NHEJ pathway is prone to insertion and/or deletion (indel) mutations at the junctional site, which in coding exons can introduce premature stop codons or frameshift mutations, leading to disruption of the targeted gene. This system therefore provides a convenient way to generate loss-of-function mutations in the mammalian genome.

Overview and basic principles of a knockout screen using the CRISPR-Cas9 system

Given the ease of generating libraries containing thousands of gRNAs that could be used to create large knockout collections, CRISPR-Cas9 technology quickly became the method of choice for genome-wide loss-of-function screening approaches. The first genome-wide loss-of-function screens were carried out in human [171, 172] and mouse [173] cells using cell growth as a phenotype and showed successful application with both positive and negative selection results. All the initial screens and the majority of the genome-scale screens that have been described up to now have used a pooled screening approach, the basic principles of which are summarised below.

In a pooled screening approach, gRNA oligonucleotides are synthesised as a pool and cloned to create plasmid library that is used for virus production. The viral library is then used to transduce Cas9-expressing cells[3] at a low multiplicity of infection (MOI) to generate a library of knockout cells. Ensuring a low MOI (usually MOI of 0.3) is a crucial step in pooled screens to reduce the probability of more than one gRNA being transduced and stably integrated into one cell. The mutant library thus generated is subjected to positive (e.g,

[3] In a single vector approach such as the LentiCRISPR used in [171], both Cas9 and gRNA are encoded from a single plasmid; thus the cells do not express Cas9 prior to transduction with the virus generated from this plasmid. In a dual vector approach, stable cell lines expressing Cas9 are first generated and then transduced with the gRNA library to generate knockout libraries.

drug, toxin resistance), negative (e.g. proliferation) or marker gene selections (expression of surface markers) and the cells of interest are recovered (will be discussed in detail below). The integrated virus serves as a molecular tag for each mutated gene and this can be read-out by isolating genomic DNA from the cell population, sequencing (using next-generation sequencing (NGS)) across the gRNA-encoding regions, and then mapping each sequencing read to a pre-compiled list of designed gRNA library. Computational analysis is then carried out to determine the differences in the abundance of gRNAs between the control and the phenotyped sample to identify the gene product involved [174].

A pooled screen can be carried out with negative, positive or marker gene selection. The goal of a negative selection screen is to identify perturbations that affect the survival or proliferation of cells, which cause the perturbed cells to be depleted during selection. The approach here is to transduce two sets of cell populations and subject one set to the selection while the other serves as a non-selected control. The gRNA abundance in both populations is then analysed to identify gRNAs that have been depleted because of the selection. One of the simplest form of a negative selection screen is the continued growth of cells for an extended amount of time to identify genes that are required for the proliferation of cells. Such screens have been used to identify essential genes for several cell lines [171, 175, 176, 177]. In a positive selection screen, a strong selective pressure is introduced such that the probability of cells being selected without the genetic perturbation is low. These screens have been used to identify perturbations that confer resistance to drug [171, 173], toxin [173, 178, 179], hypoxia stress [180], and pathogen infection [181, 182, 183]. Unlike negative selection screens, the signal for a positive selection is usually strong, as the abundance of relevant gRNAs in such screens increases relative to the rest of the gRNAs, which allows for easy detection of resistant cells [184]. A third type of selection is the marker gene selection in which the phenotype is not based on lethality of the cells but rather on mutations that change marker gene protein expression. In this type of screen, the marker gene is either endogenously-tagged with fluorescent proteins [185] or labelled with highly specific antibodies [186, 187] and cells with gRNAs that target genes whose perturbations contribute to the expression of marker gene are captured using fluorescence-activity cell sorting (FACS)-based approaches.

1.6.5 The scope of CRISPR-Cas9 knockout screening system in the context of cellular interactions

Over the last decade, technological advancements in gene annotation and gene synthesis have facilitated the generation of libraries of recombinant proteins to carry out numerous cell-based or cell-free assays, as discussed here (figure 1.2). The recent development of the genetic-screening method using the CRISPR-Cas9 system has the potential to utilise these libraries of recombinant proteins that could potentially be involved in cellular interaction within different biological contexts (e.g. neural and immunological recognition, and host-pathogen interactions). By using the 'binding' of recombinant proteins to cell lines as a measurable phenotype, a positive selection genome-scale knockout screen can be designed, in which cells within the knockout library that can no longer bind the protein they previously bound can be selected for the identification of the perturbations, which led to the loss in binding. Within those perturbations should, in principle, lie the receptor of the recombinant protein and any components that alter the expression of that receptor. This approach could be used for the study of cellular recognition events and has the following potential advantages over the previously discussed pre-existing methods:

1. Screening on cells that express a wide range of endogenous proteins in the context of a plasma membrane, which contain appropriate post-translational modifications, avoids the necessity of recombinantly generating a large number of proteins. This facilitates the investigation of proteins that are normally difficult to study with biochemical approaches (e.g, large proteins, multi-pass membrane receptors and protein complexes).
2. Genome-scale screens provide a unique platform to study interactions mediated by glycans as the identity of a glycan receptor can be readily inferred by identifying enzymes and intracellular transporters involved in their biosynthesis.
3. The use of an unbiased genome-wide approach to study receptor-ligand interactions on the cell surface facilitates the identification of cellular factors that are not necessarily direct receptors but are involved in their expression on the surface of cells.

4. One important advantage of this approach is that it should be able to identify all gene products required for extracellular interactions without the need to make any prior assumptions regarding the biochemical nature of the receptor.

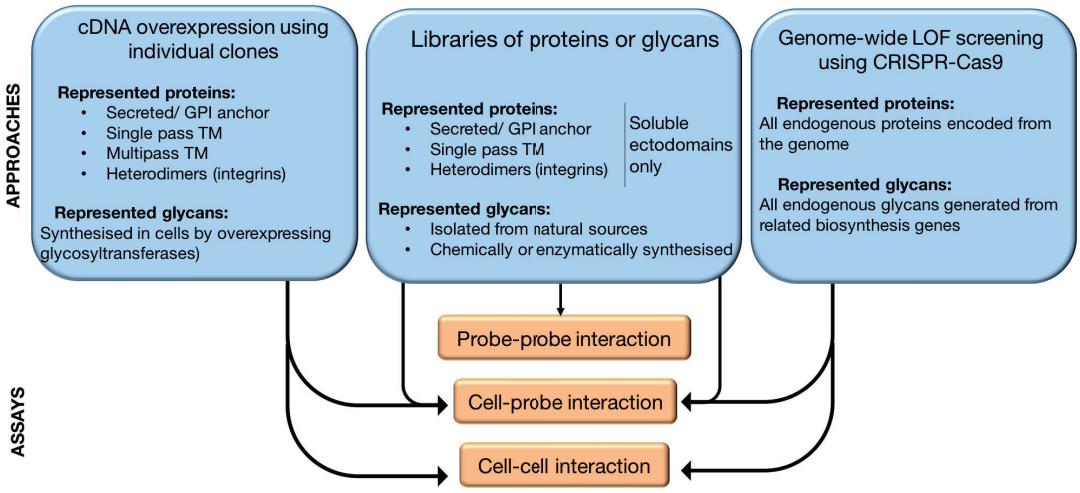

Fig. 1.2 Outline of the major approaches that are utilised for the study of extracellular interactions. Current high throughput approaches utilise compiled libraries of cDNA clones or protein/glycan libraries to investigate protein interactions. Gain-of-function approaches such as cDNA-mediated over-expression strategies have been used in cell-cell binding assays or cell-probe binding assays for the identification of novel extracellular interactions ('probe' in this context refers to avid recombinant proteins or glycans). A genome-scale loss-of-function screening approach has the potential to utilise the avid recombinant proteins or glycans in a cell-binding assay to complement the pre-existing methods to identify direct receptor-ligand interactions. In addition, a genome-scale knockout approach would, in principle, also allow for the identification of intracellular pathways contributing to the biology of the receptor.

1.7 Thesis aims

In this work, I explore whether genome-scale cell-based CRISPR-Cas9 knockout screens can be used to determine the molecular basis of cell surface recognition events. The aim is to develop an approach that utilises the commonly used tools in receptor discovery, monoclonal antibodies and recombinant proteins, to reveal direct receptor-ligand interactions, and assess the feasibility of such an approach to identify other important cellular factors required for the interaction, such as posttranslational modifications, co-receptors, and cytoplasmic proteins involved in receptor trafficking.