UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1660

Membrane-bound proteins

Characterization, evolution, and functional analysis

MISTY M. ATTWOOD

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2020

ISSN 1651-6206
ISBN 978-91-513-0926-2
urn:nbn:se:uu:diva-407778

Dissertation presented at Uppsala University to be publicly examined in Room A1:107, Uppsala Biomedicinska Centrum (BMC), Husargatan 3, Uppsala, Monday, 25 May 2020 at 10:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English. Faculty examiner: Professor Arne Elofsson (Stockholm University, Department of Biochemistry and Biophysics).

Abstract
Attwood, M. M. 2020. Membrane-bound proteins. Characterization, evolution, and functional analysis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine* 1660. 45 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-0926-2.

Alpha-helical transmembrane proteins are important components of many essential cell processes including signal transduction, transport of molecules across membranes, protein and membrane trafficking, and structural and adhesion activities, amongst others. Their involvement in critical networks makes them the focus of interest in investigating disease pathways, as candidate drug targets, and in evolutionary analyses to identify homologous protein families and possible functional activities. Transmembrane (TM) proteins can be categorized into major groups based the same gross structure, i.e., the number of transmembrane helices, which are often correlated with specific functional activities, for example as receptors or transporters. The focus of this thesis was to analyze the evolution of the membrane proteome from the last holozoan common ancestor (LHCA) through metazoans to garner insight into the fundamental functional clusters that underlie metazoan diversity and innovation. Twenty-four eukaryotic proteomes were analyzed, with results showing more than 70% of metazoan transmembrane protein families have a pre-metazoan origin. In concert with that, we characterized the previously unstudied groups of human proteins with three, four, and five membrane-spanning regions (3TM, 4TM, and 5TM) and analyzed their functional activities, involvement in disease pathways, and unique characteristics. Combined, we manually curated and classified nearly 11% of the human transmembrane proteome with these three studies. The 3TM data set included 152 proteins, with nearly 45% that localize specifically to the endoplasmic reticulum (ER), and are involved in membrane biosynthesis and lipid biogenesis, proteins trafficking, catabolic processes, and signal transduction due to the large ionotropic glutamate receptor family. The 373 proteins identified in the 4TM data set are predominantly involved in transport activities, as well as cell communication and adhesion, and function as structural elements. The compact 5TM data set includes 58 proteins that engage in localization and transport activities, such as protein targeting, membrane trafficking, and vesicle transport. Notably, ~60% are identified as cancer prognostic markers that are associated with clinical outcomes of different tumour types. This thesis investigates the evolutionary origins of the human transmembrane proteome, characterizes formerly dark areas of the membrane proteome, and extends the fundamental knowledge of transmembrane proteins.

Keywords: Transmembrane protein, alpha-helical membrane protein, evolution membrane proteins, 3TM, 4TM, 5TM, trispanin

Misty M. Attwood, Department of Neuroscience, Schiöth: Functional Pharmacology, Box 593, Uppsala University, SE-751 24 Uppsala, Sweden.

ISSN 1651-6206
ISBN 978-91-513-0926-2
urn:nbn:se:uu:diva-407778 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-407778)

Yes, but under very different circumstances from those expected.

Robert F. Scott
Upon reaching the South Pole, 1912

Misty M. Attwood
Upon finishing doctoral thesis in self-isolation during worldwide pandemic, 2020

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I **Attwood MM**, Krishnan A, Almén MS, Schiöth HB. (2017) Highly diversified expansions shaped the evolution of membrane bound proteins in metazoans. *Scientific Reports,* 7:12387.

II **Attwood MM**, Schiöth HB. (2020) Classification of trispanins: A diverse group of proteins that function in membrane synthesis and transport mechanisms. *Frontiers in Cell and Developmental Biology*, 7:386.

III **Attwood MM**, Krishnan A, Pivotti V, Yazdi S, Almén MS, Schiöth HB. (2016) Topology based identification and comprehensive classification of four-transmembrane helix containing proteins (4TMs) in the human genome. *BMC Genomics*, 17:268.

IV **Attwood MM**, Schiöth HB. (2020) Characterization of five transmembrane proteins: with focus on the Tweety, Sidoreflexin, and YIP1 domain families. *Submitted manuscript*.

Additional Papers

Attwood MM, Jonsson J, Rask-Andersen M, Schiöth HB. (2020) Soluble ligands as drug targets. *Nat Reviews Drug Discovery*. *Submitted manuscript*.

Bondarev AD, **Attwood MM**, Jonsson J, Chubarev VN, Tarasov VV, Schiöth HB. (2020) Opportunities and challenges for drug discovery in modulating Adhesion GPCR functions. *Expert Opinion On Drug Discovery*. *Submitted manuscript.*

Bondarev AD, **Attwood MM**, Jonsson J, Chubarev VN, Tarasov VV, Schiöth HB. (2020) Advances in new HDAC inhibitors development: Current targets, new molecules and targets. *Submitted manuscript.*

Attwood MM, Rask-Andersen M, Schiöth HB. (2018) Orphan Drugs and Their Impact on Pharmaceutical Development. *Trends in Pharmacological Sciences*, 39:525-535.

Crüsemann M, Reher R, Schamari I, Brachmann AO, Ohbayashi T, Kuschak M, Malfacini D, Seidinger A, Pinto-Carbó M, Richarz R, Reuter T, Kehraus S, Hallab A, **Attwood MM**, Schiöth HB, Mergaert P, Kikuchi Y, Schäberle TF, Kostenis E, Wenzel D, Müller CE, Piel J, Carlier A, Eberl L, König GM. (2018) Heterologous Expression, Biosynthetic Studies, and Ecological Function of the Selective Gq-Signaling Inhibitor FR900359. *Angew. Chem. Int. Ed, 57*, 836.

Boström AE, Ciuculete D-M, **Attwood MM**, Krattinger R, Nikontovic L, Titova OE, et al. (2017) A MIR4646 associated methylation locus is hypomethylated in adolescent depression. *Journal of Affective Disorders*, 220:117–28.

Hauser A, **Attwood MM**, Rask-Andersen M, Schiöth HB and Gloriam DE. (2017) Trends in GPCR drug discovery: new agents, targets and indications. *Nat Reviews Drug Discovery*, 16:829-842.

Contents

Abbreviations

EC	Enzyme Commission number
ER	Endoplasmic reticulum
FDA	US Federal Drug Administration
GO	Gene ontology
GPCR	G protein-coupled receptor
LHCA	Last holozoan common ancestor
TCDB	Transporter Classification Database
TM	Transmembrane

Introduction

Membrane proteins function in many essential cell processes and engage in a variety of important activities such as functioning as receptors and in signal transduction, acting as recognition molecules of the immune system, providing structural components, and transporting ions and molecules across impermeable lipid membranes [1]. Membrane proteins are also theorized to be instrumental in the development of multicellularity. Furthermore, they are key factors in the evolution of metazoans and the development of many functions that are unique to complex organisms including cell-cell adhesion, signalling, immune defence, and developmental processes [2].

Membrane-embedded protein structure

Transmembrane proteins (TM) are comprised of two secondary structural elements: either beta-barrels or alpha-helices, which is the focus of this thesis. Transmembrane alpha-helical segments contain largely hydrophobic residues such as leucine, isoleucine, alanine, and methionine and typically extend ~20 amino acids, approximately the width of the phospholipid bilayer. The aromatic residues tryptophan and tyrosine are frequently enriched at the membrane-aqueous interface, which serves to anchor and stabilize the transmembrane orientation [3]. The alpha-helical structure is also stabilized through specific interactions between side chains, hydrogen bonding, isolation of hydrophobic residues, aromatic residue stacking, as well as helix-helix interactions [1]. Polytopic membrane proteins, i.e. those that span the membrane multiple times, contain loops between the transmembrane segments and positively charged residues are found at higher frequencies in the cytoplasmic loops, noted as the 'positive-inside rule' [4]. Transmembrane helices can be short, long, kinked, cross the membrane at oblique angles, or form re-entrant loops where they enter the membrane part-way and then return to the same side [5]. These varying forms of transmembrane structure can make it difficult for transmembrane prediction algorithms to correctly evaluate and predict transmembrane regions. Hence, for this thesis, we attempt to focus on transmembrane helices that completely span the membrane.

Transmembrane topology

In addition to the cytosolic regions that contain more hydrophilic amino acids which interact with the aqueous environment, transmembrane proteins also contain an exoplasmic region if located in the plasma membrane or luminal/non-cytoplasmic region if spanning an organelle membrane. The terminal ends of the protein contain an N-terminal, a positively charged amine group (NH2) which indicates the 'start' of the polypeptide, and a C-terminal, composed of the negatively charged carboxyl group (COOH). The orientation of the N- and C- terminals in relation to the membrane can be determined by the protein assembly and the manner in which the polypeptide chain is inserted into the membrane [5]. The terminal ends of a transmembrane protein can be engaged in a variety of functions, for example an N-terminal of a receptor that is located in the extra-cellular environment can be involved in stabilizing the first transmembrane helix which subsequently can ensure correct receptor structure and facilitate ligand binding and signal transduction [6]. An intracellular C-terminal can, for instance, be involved in the assembly and regulation of channel subunits, as in the case of voltage-gated ion-channels, and further be engaged in electrical signalling [7]. The topology of a membrane protein includes the number of transmembrane segments as well as the orientation of the N- and C-terminals [5] (Figure 1).

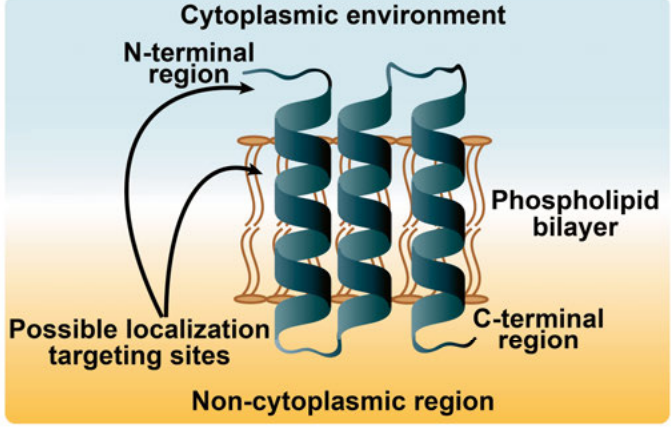

Figure 1. Example of transmembrane protein topology with three alpha-helical segments spanning the phospholipid bilayer.

The membrane environment

The biogenesis, folding, final three-dimensional structure, as well as targeting and transport of alpha-helical membrane proteins are important components in understanding membrane proteins. Transmembrane proteins are assembled

and guided into the membrane and achieve their final three-dimensional structure through the Sec61 translocon complex as well as additional ancillary proteins that aid in assembling and insertion [8,9]. The 'sequential-insertion' scheme where hydrophobic segments are guided into the membrane as they emerge from the ribosome is one hypothesis [10], although newer studies with marginally hydrophobic transmembrane regions indicate that a more dynamic relationship driven by thermodynamically controlled interactions may influence translocon-guided transmembrane protein insertion into the membrane [8,10,11].

Membrane composition

Transmembrane proteins are embedded in different biological membranes including the plasma membrane as well as organelles such as the nucleus, vacuoles, vesicles, mitochondria, the endoplasmic reticulum (ER), and the Golgi apparatus. Most cell membranes are composed of glycerophospholipids, which are molecules made of glycerol, a phosphate group, and two fatty acid chains, and hence the phospholipid bilayer is hydrophobic in nature [12]. This impermeable barrier separates the interior and exterior environments, for example the plasma membrane separates the cell cytosol from the extra cellular environment or intracellular membranes keep the lumen of an organelle segregated from the cytoplasmic region. However, organelles differ both in quantity and variety of lipid content, that is they contain varied distributions of different types of lipids [13]. Furthermore, the amount and types of transmembrane proteins within a membrane vary depending on different factors, for example the cell type and subcellular location. Accordingly, while membranes are comprised of a combination of lipids and proteins, the membrane composition of different organelles is dependent on the distribution of the different types of lipids and proteins [13]. In fact, the variable distribution of lipids and proteins as well as the trafficking and exchange of them between membranes contribute to the dynamic quality of membranes. And importantly, the distribution of different kinds of membrane proteins engender specific functions respective to the that membrane (Figure 2).

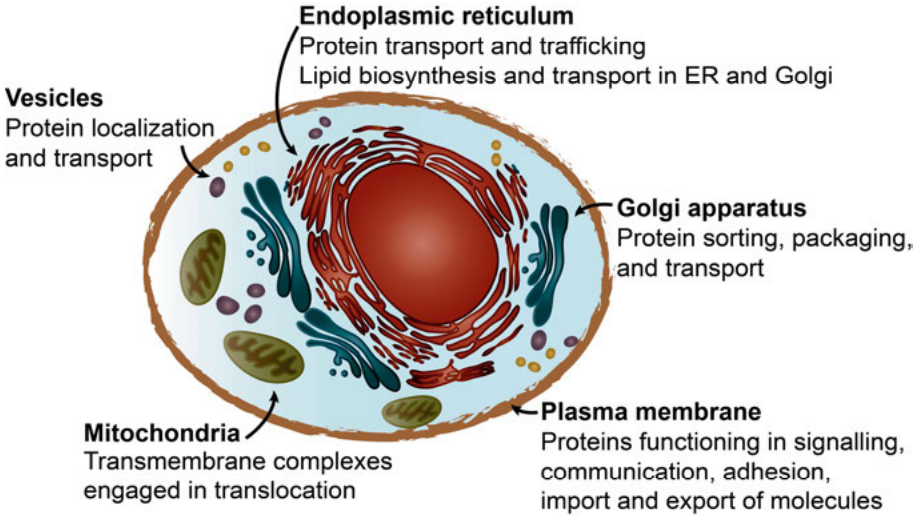

Figure 2. Functions associated with different organelles with emphasis on activities identified in the three, four, and five transmembrane protein datasets.

Membrane and protein trafficking

The development of organelles and formation of subcellular compartments has been facilitated through the evolutionary retargeting of membrane proteins, where the subcellular localization of a protein is altered to shared or different locations in the course of evolution [14]. Indeed, the development of organelles and protein sorting mechanisms most likely co-evolved [15]. Proteins are found to localize to multiple locations, with the Human Protein Atlas concluding that more than half of the genes in their database encode multilocalizing proteins [16,17]. Protein targeting and trafficking can be separated into two primary groups: direct co-translational transport across or into a membrane (as described above) or post-translational transport of premade proteins, where in the latter proteins are either guided by escorting proteins to the organellar surface and are subsequently integrated or translocated across a membrane, or the proteins are moved through vesicle-mediated transport [15,18]. The ability of proteins to switch locations confers several advantages, including quick cellular responses upon a changing environment. Furthermore, multilocalizing proteins can also engage in multifunctional activities that may correlate with different locations [16].

Protein localization

In order for the important functional tasks of membrane proteins to be carried out, they need to be directed or transported to their appropriate destinations. Targeting signals, also called signal peptides or targeting peptides, can help

direct proteins and enable cellular transport machinery to correctly deliver the protein. One type of protein targeting system uses the signal peptide, which can be found at the N-terminal region and is composed of hydrophobic amino acids, to direct proteins for insertion into the endoplasmic reticulum (ER) or embedding into the ER membrane [19]. Another common signalling method lies within the structure of the protein itself, often the first transmembrane domain or other intra-membrane components, and sometimes includes additional residues in the N- or C-terminal [20,21]. Different determinants within transmembrane domains such as the exact length and amino acid composition have been found to be strongly correlated with the intracellular localization of the protein. For example, the transmembrane region of proteins that localize predominantly to the ER were found to be shorter than those of plasma membrane proteins, which were found to have an asymmetric distribution of residues along the alpha-helix [20]. Furthermore, post-translational modifications such as glycosylation and acetylation can also affect subcellular targeting as well as topological orientations of the terminals [22,23].

Evolution of membrane proteins

Investigating the patterns of conservation, expansion, and reduction of membrane protein families throughout metazoans can help elucidate mechanisms which have contributed to developing organismal complexity. Moreover, comparisons of membrane proteomes can provide insight into how different biological systems evolved and how environment and life style can influence the evolution of species. The proteins that are conserved throughout lineages can indicate their functional importance and help in identifying the minimum protein set needed by organisms. The expansion of protein families through the creation of new genes is predominantly achieved through the duplication of present genes followed by divergence of the paralogs [24,25]. This may result in neo- or sub-functionalization, which can facilitate further adaptive qualities or specialized functions, although it often leads to forming non-functional pseudo-genes [26]. Genome reduction has been observed in lineages adapted to specialized environments, where population sizes, mutation rates and recombination rates, amongst other factors, appear to affect genetic reductions [27,28]. Studying the patterns of gene addition or reduction in species can give insights in how novel adaptations developed and evolved.

Comparison of metazoan genomes

Comparisons of important membrane protein families in metazoan species has previously revealed that many crucial families are unique to metazoans, although a substantial portion also has a pre-metazoan origin. For example,

tyrosine kinases originated pre-opisthokont and show multiple radiations in pre-metazoans, although cytoplasmic and receptor tyrosine kinases (RTK) show a difference in divergence patterns with RTKs significantly expanding in metazoan lineages [29,30]. G protein-coupled receptors (GPCRs), the largest human receptor superfamily with 800+ members, show a minimal number of representatives in unicellular holozoans [31,32] although it has dramatically expanded in specific metazoan lineages, as we show in paper I. Extensive comparisons of membrane proteomes across species to include pre-metazoan lineages gives insight into the origins of multicellularity and cell differentiation, which are poorly understood [33].

Involvement in disease pathology

Given the pivotal functional roles that membrane proteins are engaged in, they are often involved in disease pathology and subsequently common targets for pharmaceutical agents. Indeed, it has been estimated that 60% of drug targets are located at the cell surface [34], and GPCRs alone compose ~35% of all drugs approved by the FDA [35]. Transmembrane proteins such as receptors and ion channels have always been primary drug targets, however with increased understanding of disease pathways as well as with continued advancements in biopharmaceutical engineering, other classes of polytopic proteins are being explored. For example, tetraspanins in general, and CD37 in particular, have been explored as therapeutic targets in haematological malignancies [36]. Furthermore, studies are being developed involving receptor mediated transcytosis (RMT) where the membrane bound transferrin receptor or insulin receptor is used to transport drugs into previously inaccessible areas, for example across the blood brain barrier [37]. New bioengineering developments include therapeutic agents such as antibodies, nanobody-photosensitizer conjugates [38] and antibody-drug conjugates to target membrane proteins for treatment of cancers [39]. Hence, polytopic membrane proteins are expected to be of increasing interest as therapeutic antibody targets [40].

Association between structure and function

Of the ~20,000 protein-coding genes in the human genome, approximately 25-30% (~5500) are predicted to be alpha-helical transmembrane proteins [17,41]. Membrane-spanning proteins function in many crucial cell activities where they catalyse interactions, engage in numerous signalling pathways, facilitate the transport of molecules throughout the cell, and regulate many processes at specific times and locations, amongst other activities. For example, GPCRs function as receptors to extracellular stimuli that bind at

specific active sites and then signals are transduced along distinct pathways. Regulation of activities and coordination of actions are particularly important in transmembrane proteins, for instance through ubiquitination the timing and specificity of downregulating proteins can be controlled, which affects activities such as cell fate specification and neurotransmission [42].

Studies point towards a strong correlation between membrane structure and function, as noted by Sällman Almén and colleagues in their detailed classification of the human membrane proteome that analysed groups of transmembrane proteins based on their number of membrane-spanning regions [43]. They identified that some membrane topologies are more common with certain functional classes, i.e. receptors, transporters and enzymes. Proteins that function as receptors often contain one or seven TM helices while proteins that have a high number of TMs can be an indicator of transport activity, such as the solute carrier families that have 10 – 14 TM regions [44]. Approximately 60% of membrane bound enzymes and nearly 30% of non-GPCR receptors are single transmembrane proteins [41].

Our previous investigation of the membrane proteome suggested that there were several categories of proteins that contained the same gross structure, in particular proteins with fewer numbers of transmembrane regions, that have not been functionally characterized or analysed. In addition, these groups contain many proteins that have unknown functions. Hence the impetus for this thesis was to identify and functionally characterize proteins that contain three-, four-, and five-transmembrane regions and analyse these datasets along with cellular localizations, tissue enrichment patterns, protein-protein interactions, and involvement in disease to produce comprehensive analyses about these proteins. Further, we wanted to examine membrane proteins in an evolutionary context to investigate the innovations, expansions, and reductions of functional protein groups.

Aims

The aim of this thesis was to analyse the evolution of the membrane proteome and characterize membrane proteins. Further, we wanted to investigate previously unstudied groups of transmembrane proteins that contain similar tertiary structures and characterize the predominant functional activities, associations with disease, and identify possible pharmaceutical targets. The specific aims for each paper are listed below.

Paper I
The aim for *paper I* was to investigate what transmembrane proteins were present at the origin of metazoans and give a comparative analysis of the evolution of membrane proteomes in holozoans.

Paper II
The focus of *Paper II* was to investigate the previously unstudied group of proteins that contain three alpha-helical transmembrane regions (3TM or trispanin) The aim was to classify the dataset and then hypothesize on the functions of uncharacterized trispanins that are potentially involved in important pathways or are associated with disease

Paper III
The aim for *Paper III* was to investigate proteins that contain four transmembrane regions (4TM) and provide qualitative functional classifications of this relatively large group of proteins. We wanted to research the role of the topological orientation of the N- and C-terminals with the primary functional activities of 4TMs.

Paper IV
In *paper IV*, the aim was to characterize proteins that contain five alpha-helical transmembrane (5TM) regions and investigate with greater detail the tweety, sidoreflexin, and YIP1 domain families. Studying the role of 5TMs in disease was also a focus of this project.

Materials and Methods

The analyses in the four papers followed similar methodology involving retrieval of proteomes, data processing, *in silico* transmembrane predication followed by manual curation and classification of human transmembrane sequences (Figure 3). Paper I had additional steps including pairwise sequence alignments using BLASTp and clustering proteins by homology and functional descriptions.

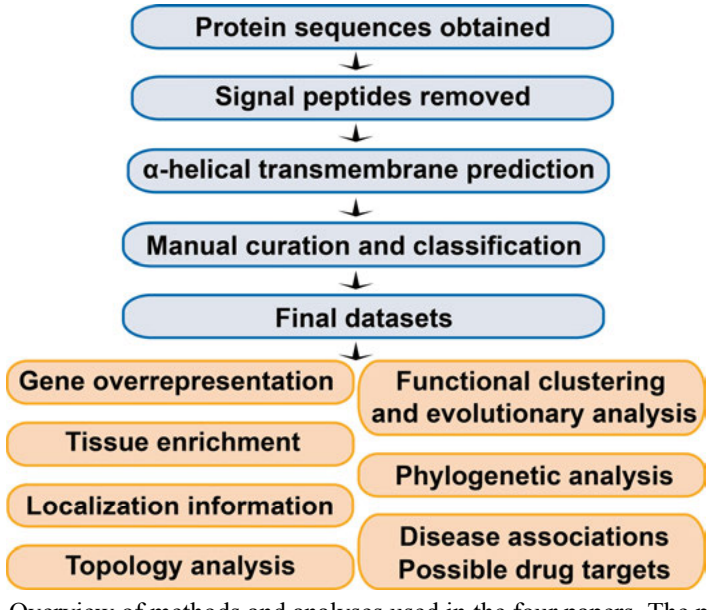

Figure 3. Overview of methods and analyses used in the four papers. The processing of the sequences and creation of the datasets are in blue while the analyses performed are in orange.

Proteome retrieval

The most current assemblies of the human reference genome GRCh38 were downloaded from the National Center for Biotechnology Information (NCBI). The GenCode gene annotations, which were used for paper I, initially includes more transcripts as it contains pseudogenes, long non-coding RNAs and small non-coding RNAs which can be useful in projects investigating different gene

features. The Consensus Coding Sequence (CCDS) annotations, which were used in papers II, III and IV, have more extensive manual curation and is more conservative, leading to high quality, very stable protein-coding annotations. In the evolutionary analysis study, other species proteomes were obtained with consideration to the best available resources, including the Genome Reference Consortium as well as other collaborations with high quality assemblies and gene prediction pipelines. In papers I and III, the sequences were assessed and if a gene produced multiple protein isoforms, the longest sequence for each gene was used. In papers II and IV the canonical sequence as identified in UniProt was retained, which is determined as the most prevalent isoform, the most similar to orthologous sequences, the composition of the residues in the sequence, or typically the longest sequence.

Alpha-helical transmembrane region prediction

Computational membrane prediction methods are often used in studies involving membrane proteins due to the difficulty in obtaining experimentally determined transmembrane protein structures. As alpha-helices are hydrophobic in nature, there can be issues in attempting to unfold and then refold them *in vivo*. Additionally, extracting membrane-embedded proteins often requires a high concentration of detergent which can denature the proteins or impede crystallization, and furthermore membrane proteins are generally present in low concentrations in the cell so they must be over-expressed to obtain enough to work with [45]. The Protein Data Bank (PDB) [46] is the repository for experimentally determined proteins structures, totalling 161,273 structures in the database as of March, 2020; however, the number of alpha-helical transmembrane proteins totals between only 1896 according to the Orientations of Proteins in Membranes (OPM) database [47] and 4387 in the Protein Data Bank of Transmembrane Proteins (PDBTM) [48].

Before using alpha-helical transmembrane prediction methods, the sequences need to be assessed for signal peptides because prediction methods can have difficulty differentiating between N-terminal signal peptides and transmembrane segments. SignalP v4.1 [49] was used with default settings in papers I, II, and III, while SignalP v5 [50], which was released in early 2019, was used in papers II and IV.

There are many different algorithms that have been developed to predict the number of transmembrane helices and the orientation of the N- and C-terminals of the protein in the membrane. Two general approaches are used: one uses sequence statistics that are based on the topologies of known transmembrane proteins while the other uses the position-specific amino acid contributions to the free-energy of the membrane insertion of the protein, also called the sequential-insertion scheme, to predict membrane topology.

Prediction methods use machine learning techniques with different algorithms including hidden Markov models (HMM), neural networks, and support vector machines. As each method uses a different algorithm, there can be considerable discrepancy in the number of alpha-helices predicted between the methods and hence using a consensus method that consists of several underlying predictors has been suggested [51]. The consensus prediction method TOPCONS-single, which is appropriate to use for whole full proteome scans, was used in all four studies to discern alpha-helical transmembrane regions [52]. The default parameters for TOPCONS-single uses four different prediction methods: S-TMHMM [53], HMMTOP [54], MEMSAT [55], and SCAMPI-single [56]. An additional step was added in papers II and IV in which sequences were re-evaluated with TOPCONS2 [57], which is a more recent iteration of the TOPCONS series and can more successfully predict transmembrane regions due to the integration of information from scans of homology-based resources such as Pfam [58] and the Conserved Domain Database (CDD) [59].

Characterization of the dataset

Primary classification of proteins

The primary functional classification categories of proteins in each of the papers were transporters, enzymes, receptors, and other varied activities. The UniProt resource was used to obtain annotations including gene symbols, sequence status, functional descriptions, protein family information, and also cross-references resources to obtain the Transporter Classification Database (TCDB) number, Enzyme Commission (EC) number, and Gene Ontology (GO) terms [60]. The TCDB system incorporates both functional and phylogenetic information into an International Union of Biochemistry and Molecular Biology (IUBMB) approved classification system for membrane proteins [61]. The EC nomenclature scheme, which also has IUBMB approval, is based on the chemical reactions the enzymes catalyse. The International Union of Basic and Clinical Pharmacology (IUPHAR)/British Pharmacological Society (BPS) Guide to Pharmacology [62], which is a curated resource of ligand-target relationships, was used to determine receptors. Proteins that were not classified in one of these three groups were manually categorized using similar Pfam families, functional descriptions, and GO terms.

Gene-disease associations

Several different resources were used in each of the papers to identify proteins and genes that are associated with diseases. The Online Mendelian Inheritance

in Man (OMIM) database [63], which contains a catalogue of genetic traits and disorders with all known Mendelian disorders, was cross-referenced through UniProt and annotations for the proteins in the 4TM dataset were identified. The Jensen Diseases database was used with the 3TM and 4TM studies, which incorporates disease-gene associations from automatic text mining, literature, cancer mutation data, and genome-wide association studies with evidence confidence scores for each association [64]. The DisGeNET gene encyclopaedia [65], which integrates data from expert curated resources to produce homogeneously annotated ontologies of genotype-phenotype relationships, was used in both the 3TM and 5TM analyses. DisGeNET uses evidence metrics to identify relevant associations. The Pathology Atlas [66], which contains protein expression data from different types of human tumours, was used to identify candidate prognostic genes that are associated with the clinical outcome of different tumour types in the 5TM study.

Potential drug targets

Two sources were used in the 3TM, 4TM, and 5TM studies to identify possible transmembrane proteins that are targeted by pharmaceutical agents. The DrugBank database was cross-referenced through UniProt and contains information on proteins that are targeted drugs that have been approved by the US Food and Drug Administration (FDA) as well as investigative agents in clinical trials [67]. Additionally, drug-target information was obtained through the manually curated and updated dataset on clinically established as well as novel drug targets from Rask-Andersen and colleagues [68,69].

Enrichment analyses

Gene overrepresentation

More nuanced insights into the predominant functional activities and regions in the cell where proteins were localized were achieved through enrichment analyses. The PANTHER Classification System, which uses GO annotations to identify functional descriptions and classifications of gene products, was used with the Statistical Overrepresentation Test in the 3TM and 5TM studies to analyse if or how the functional activities of these proteins were overrepresented in these datasets in comparison to the entire human membrane proteome [70]. In the Statistical Overrepresentation Test, the parameters Fischer's Exact test was chosen as it assumes a hypergeometric distribution that is more accurate for smaller gene lists and the Benjamini-Hochberg False Discovery Rate (FDR) correction ($p<0.05$) was used to control the false positive rate in the statistical test results [70]. The FDR adjusted p-value results in fewer false positives as it implies that 5% of the significant results

will be false positives. The annotation sets used in the tests included GO molecular functions and biological processes, which aided in determining group functional activities, and also cellular components, which used GO terms to identify cell locales the proteins were found to be overrepresented in comparison to all human transmembrane proteins. The reference list for the homo sapiens membrane proteome was obtained from [71].

Tissue enrichment

The recently released resource TissueEnrich: Tool for tissue-specific gene enrichment in human and mouse was utilized to analyse tissue-specific gene enrichment in the 3TM and 5TM datasets in comparison to the human membrane proteome. TissueEnrich uses RNA-Seq data to describe tissue-specific genes [72]. The Human Protein Atlas dataset, which consists of RNA-Seq data from 35 human tissues, was the dataset chosen in the studies. TissueEnrich uses the hypergeometric test to calculate the enrichment of tissue-specific genes in the 3TM and 5TM dataset and the Benjamini-Hochberg correction to correct for multiple hypothesis [72].

Creating functional clusters

In paper I, a two-step classification process was implemented to identify groups of proteins with similar functional activities in the 24 eukaryotic proteomes (Figure 4). The first level of clustering was based on sequence similarities using the BLAST+ package where all-versus-all pairwise alignments of all transmembrane sequences from the eukaryotic proteomes were generated. A homology network was constructed from the BLAST results, which were processed as an undirected network with the edges weighted by the BLAST e-values using the Markov Clustering (MCL) algorithm.

The MCL algorithm establishes clusters by a mathematical bootstrapping procedure and is based on simulation of stochastic flow in graphs [73]. The MCL algorithm uses Markov matrices and iteratively performs two operations called expansion and inflation. The expansion step corresponds to computing random walks of higher length, where it associates new probabilities with all pairs of nodes. And since higher length paths are more common within clusters than between different clusters, intra-cluster node pairs will generally have higher probabilities as there are many ways of going from one node to the other [73]. The inflation step changes the probabilities associated with the collection of random walks departing from one particular node through favouring more probable walks over less probable ones and hence boosting the probabilities of the intra-cluster walks. The expansion and inflation steps are iterated until there is no longer any significant changes and the algorithm

converges, resulting in the final set of clusters. This algorithm is hailed as one of the more successful approaches to clustering proteins based on sequence similarities [74].

The second level of clustering grouped the first level BLAST clusters based on annotations from the Pfam families or other functional activity descriptions. The membrane proteins of all twenty-four species in this dataset were evaluated against the Pfam database using a local instalment and the associated Pfam families for each protein were obtained. In the second MCL clustering, two clusters that contained proteins of the same Pfam family were connected in the network with an edge weighted by the fraction of proteins that contained that Pfam family.

The first level BLAST clusters were classified according to the characterization of the human membrane proteome (described previously) and a consensus first level description was built for each cluster. The second level clusters were then manually assessed to control for conflicting classifications between the first level descriptions and Pfam annotations to reach an appropriate annotation for each second level cluster. There did exist some clusters that were annotated as ambiguous as consensus classification failed.

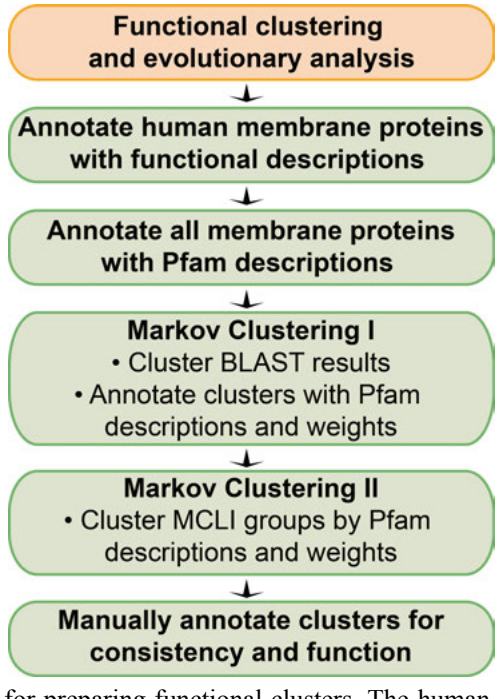

Figure 4. Pipeline for preparing functional clusters. The human membrane proteins were annotated with functional descriptions and all pre-processed sequences from the 24 eukaryotic membrane proteomes were annotated with Pfam descriptions. Two-step Markov clustering was performed on first the BLAST results and then on those clusters using Pfam annotation. The first and second level clusters were manually annotated for consistency and function using Pfam and functional descriptions.

Phylogenetic analysis

Phylogenetic reconstruction was used in paper IV to evaluate the evolutionary relationships of the genes in the sidoreflexin family. The first step was to create a multiple sequence alignment (MSA) in which the variation of residues at each position of the alignment could be determined. The MSA was created in MAFFT using the E-INS-i iterative refinement method, which aligns sequences that have several conserved regions embedded with unalignable regions [75]. The MSA was then manually curated in Jalview [76]. Two different methods were used for phylogenetic tree reconstruction to ensure reliability over the tree topology: Bayesian inference was used to estimate the posterior probabilities of the trees and Maximum Likelihood (ML) to verify the topology with bootstrap analysis. Both Bayesian inference and ML methods are character-based approaches that simultaneously compare all sequences in the alignment and consider one site in the alignment (or character) at a time to compute the tree score [77]. The ML method uses complex substitution models and has many adjustable parameters to estimate the most credible tree, however it involves heavy computation time. Bayesian analysis also has realistic substitution models and the parameters are considered random variables with statistical distributions, whereas in ML they are unknown fixed constants. Before Bayesian analysis, the parameters are assigned a prior distribution which is then combined with the data to produce the posterior distribution [77]. The sidoreflexin phylogenetic tree was constructed using MrBayes version 3.2.7a [78] to generate tree support using posterior probabilities and RAxML version 8.2.12 to cross verify with bootstrap analysis [79].

Results

Paper I: Evolution of the membrane proteome

Twenty-four eukaryotic proteomes were investigated that originally included nearly 500,000 sequences, with 123,014 transmembrane sequences predicted. More than 85% – 105,757 membrane proteins – were functionally characterized into 2181 groups after first and second level clustering (Figure 5). The characterized proteins were classified into several main categories: receptors consist of the largest proportion of proteins (24%) in comparison with the other main functional classes including enzymes (17%) and transporters (21%). In all five vertebrate species the receptor class contained the most number of proteins within these three categories, ranging from 20% to 36%. In 60% of the invertebrate species, transporters were the largest functional class of these three, ranging from 16% to 21% of their classified proteins. In most of the investigated species, however, the group of proteins that have varied functional activities contained the most number of proteins, and comprised 33% of the characterized proteins. Furthermore, 3% of the characterized proteins contained a conserved, but unknown, Pfam domain.

To determine the functional clusters that were present at the origin of metazoan evolution from the last holozoan common ancestor (LHCA), clusters were evaluated in which at least one protein was identified in either unicellular holozoan outgroup and also present in metazoan species. We identified 604 conserved functional clusters, which across all species totalled 90,906 proteins. The possible LHCA membrane proteome was created by averaging together the number of proteins (rounding up) of the two unicellular eukaryotic outgroups *M. brevicollis* and *C. owczarzaki* in each of the 604 clusters, which totalled 2283 proteins. Presumably, these are the possible proteins that comprise the functional core of the LHCA, although the proposed proteome does not include proteins that have been lost from LHCA and hence we identified proteins that are likely to be present buy may not cover all of them.

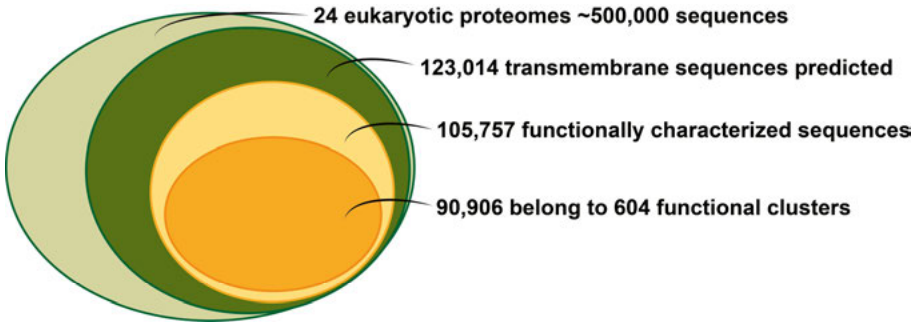

Figure 5. Overview of the number of sequences involved in this analysis.

In comparison to the proposed membrane proteome of the LHCA, metazoan proteome sizes have increased – nearly doubling in size and even tripling in several species, with the exception of the simplified orthonectid *I. linei*. While many essential protein families were present before metazoan expansion and innovation, there are major differences in how the enzyme, transporter, and receptors classes evolved throughout metazoan species. The enzymes in the membrane proteome of the LHCA was the largest classified group with 517 proteins (23%) divided into 36 functional clusters, while in metazoan species enzymes exhibit distinct increases in both vertebrates and lophotrochozoans. The transporter class in the LHCA included 370 proteins (16%) in 95 clusters, and has many pronounced expansions in the lophotrochozoans species as well as *S. purpuratus*, while in vertebrates this group does not appear to increase as significantly. The receptor class in the LHCA membrane proteome is the smallest with 309 proteins (14%) divided among 31 clusters, while many of the receptor families in metazoan lineages have doubled or expanded more than eight times in comparison.

Paper II: Trispanins

The 152 proteins identified with three transmembrane regions included 35 proteins with an EC identifier and classified as enzymes; 26 proteins with a TCDB number and labelled transporters; 21 receptors including the 18 members of the ionotropic glutamate receptor family; 43 proteins with varied functional activities; and 27 trispanins that are uncharacterized (Table 1). The orientation of the N- and C-terminals in relation to the membrane can indicate functional activities. However, the relationships between membrane topology and functional activities were difficult to assess with trispanins as 72 had the N-terminal oriented in the cytoplasmic region while 80 proteins had the N-terminal in the non-cytoplasmic region.

Table 1. Summary of functional classes identified in the trispanins dataset.

Functional classes	Total	Topology		Disease associations	Drug targets
		In	Out		
Enzymes	35	18	17	9	5
Transporters	26	11	15	8	1
Receptors	21	1	20	14	18
Varied functional activities	43	25	18	10	2
Uncharacterized	27	17	10	-	-

Trispanins function in several essential cell processes (Table 2). More than 20% of trispanins engage in forming membrane complexes that localize primarily to the plasma membrane, mitochondria, and the endomembrane system including the endoplasmic reticulum (ER) and Golgi apparatus. Trispanins are members of translocase complexes, the mitochondrial respiratory complex II, the mitochondrial respiratory electron transport chain, and ER-associated degradation pathways. Transport and trafficking are also major functional activities with 40% of the proteins in the dataset engaged in transport with 20% involved in intracellular transport and protein targeting as well as membrane and vesicle trafficking. Nearly 20% of trispanins are involved in lipid biogenesis and metabolic processes while ~16% are engaged in catabolic processes that result in the breakdown of substances. The 18 members of the ionotropic glutamate receptor family are characterized as trispanins and are described with many different annotations, including signal transduction activity, cell communication, regulation of membrane potential, and neurotransmitter receptors.

Using our bioinformatic protocols, we were able to analyse the 27 proteins in the dataset that are uncharacterized and identify those proteins that are potentially involved in significant activities or disease pathways. Many were identified that localize to the ER and mitochondria and several were predicted to interact with proteins engaged in functional activities already common in trispanins such as protein targeting, lipid metabolic processes, the ER-associated degradation process and transport, amongst others.

Nearly one-third of trispanins have disease-gene associations or are targeted by an agent approved by the US Federal Drug Administration (FDA) or an investigative drug in clinical trials. Disease-gene associations with neurological disorders are prevalent, including neurodegenerative diseases, intellectual disabilities, addiction disorders as well as other diseases such as diabetes, cancer, and viral infections.

Table 2. Trispanins are engaged in varied types of functional activities and multilocalize to different organelles in the cell. The activities in green involve different classes of proteins, although primarily those labelled with varied functional activities. The transport and trafficking activities in orange are chiefly transporters; the enzymatic functions in blue are principally enzymes; while descriptions in maroon include receptors as well as other classes of membrane proteins.

Functional activity	Total	ER	PM	Golgi apparatus	Mito-chondria	Vesicles	Nucleus
Membrane protein complex	34	20	21	5	7	10	2
Cell junction	27	12	26	5	1	12	2
Structural molecule	7	0	6	1	0	2	2
ER-associated degradation	9	8	1	3	3	5	0
Transport	61	32	37	13	8	24	11
Intracellular transport	18	12	7	7	5	7	5
Vesicle mediated transport	19	10	13	9	1	12	4
Protein transport	21	12	3	5	6	6	4
Membrane trafficking	9	8	4	6	1	4	2
Lipid metabolic processes	31	25	12	5	3	3	3
Catabolic processes	24	17	9	3	4	7	2
Heme binding	9	6	2	0	3	0	0
Cell communication	35	20	26	7	4	15	5
Signalling receptor activity	26	15	20	3	0	9	3
Regulation of membrane potential	24	11	24	3	0	9	2
Neurotransmitter receptor complex	20	11	20	3	0	9	1
Synaptic adhesion molecule	8	6	8	1	0	5	0

Paper III: 4TMs

The 373 transmembrane proteins identified as containing four alpha-helical membrane regions were classified into five main categories (Table 3). The enzyme category contains 45 proteins, with predominantly transferases (25 proteins) and oxidoreductases (12 proteins). The 66 transporters are primarily *alpha-type channels* which include 26 proteins and *auxiliary transport proteins* with 18 proteins identified. The 46 neurotransmitter gated ion channel (NGIC) family of proteins engage in both receptor and transporter activities and are thus classified as a dual function group. One other 4TM is identified as dual function as it has both enzymatic and transport identifiers and activities. Only two receptors are classified in the dataset. Proteins that have varied functional activities (also called miscellaneous) contain 213

proteins that are subsequently classified into subgroups including receptor-like activity, cell adhesion, and gap junctions, amongst others. The proteins with varied functional activities also contain 32 uncharacterized proteins, with 24 having various conserved Pfam domains.

Table 3. Summary of functional classes identified in the 4TM dataset.

Functional classes	Totals	Topology		Disease associations	Cancer	Drug targets
		In	Out			
Enzymes	45	40	5	26	8	1
Transporters	66	65	1	41	17	2
Dual function proteins	47	1	46	41	8	35
Receptors	2	2	-	2	-	1
Varied functional activities	181	177	4	99	39	2
Uncharacterized proteins	32	31	1	5	-	-

Nearly 65% (237 proteins) of the 4TMs are found within eight major Pfam domain families or clans (Figure 6). Pfam clans consist of homologous domain families while domain families are formed through the collection of proteins with the same conserved Pfam domain, for example the NGIC family does not belong to any clan and does not have any homologous sister families. The largest clan is the transport superfamily clan with 78 proteins from six different conserved families, including claudin and connexin domains, that function in cell adhesion, transporter activity, regulation, and cell communication via gap junctions. The tetraspanin-like clan contains 53 proteins from two different families that function in protein scaffolding as well as cell regulation and differentiation activities. The NGIC family contains 46 members while the marvel-like clan includes 24 proteins. The zinc-beta ribbon clan is comprised of 18 enzymes that act as transferases while the reticulon family has seven proteins involved in transport and apoptotic pathways, although one member is completely uncharacterized. The last families include the six member L6-membrane family that act in transport and regulation, although three members are uncharacterized, and the Got/Sft2-like family with five members that function in vesicle mediated transport.

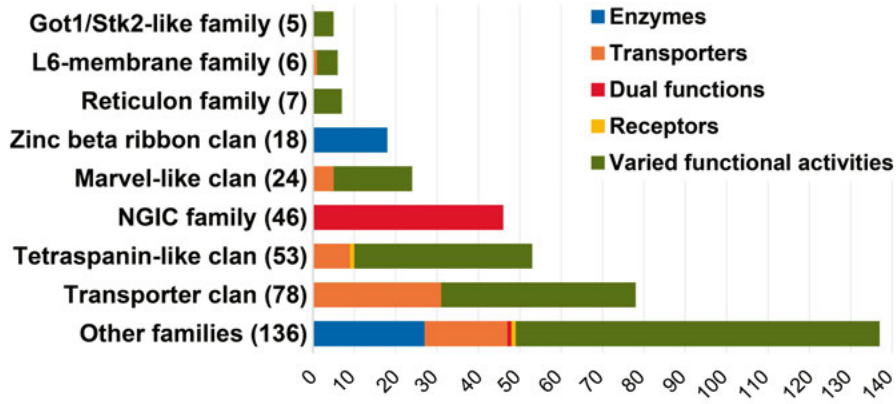

Figure 6. Functional classification of major clans and families in the 4TM dataset.

Almost 60% (215 genes) in the 4TM dataset are associated with disease conditions with 72 recognized in different types of cancers. Cancers such as lung, colon, breast, and liver were the most common types identified in the dataset. Many proteins were associated with more than one type of disease. Neurological disorders were common, with genes associated with schizophrenia, autism, Alzheimer's disease, and epilepsy as well as nicotine and alcohol dependence.

Paper IV: 5TMs

The 5TMs are composed of a smaller group of proteins consisting of 58 members. The dataset contains 17 enzymes with an EC number and transferases and hydrolases are the most common enzymatic activities with seven proteins in each class (Table 4). The transporters consist of 16 proteins including five electrochemical potential-driven transporters as well as seven proteins that are identified in incompletely characterized transport systems. Two proteins contain both EC and TCDB identifiers; however, both of them are labelled as enzymes to prevent redundancies. Additionally, one transporter has also been identified as having receptor functions, although its primary function appears to be involved in cholesterol transport. Proteins that have varied functional activities include 25 proteins.

Table 4. Summary of functional classes identified in the 5TM dataset.

Functional classes	Totals	Topology		Disease associations	Prognostic markers
		In	Out		
Enzymes	17	6	11	2	10
Transporters	16	11	5	3	11
Varied functional activities	25	13	12	2	14

More than half of the dataset belongs to twelve families (Table 5), with ten complete families that do not have any other homologous human proteins identified, while the other two families have homologues identified but with a different number of TM segments predicted.

Table 5. The twelve families that contain the 5TM architecture in the dataset. The green colour indicates families classified as engaged in varied functional activities, the orange colour signifies families involved in transport activities with at least member having an TCDB number, and the blue hue indicates families acting in enzymatic activities where at least one member contains an EC identifier.

Protein family	Functional activity	Localization
Dual oxidase maturation factor family (2)	Transport of DUOX1/2 from ER to PM	ER, PM
Prominin family (2)	Cholesterol binding	PM, Vesicle, Nucleoplasm
OXA1/ALB3/YidC family (2)	Insertases: translocation of COX2 and integral membrane proteins	Mitochondria IM
TspO/BZRP family (2)	Transmembrane signalling Mitochondrial respiration Cholesterol transport	ER, Vesicle, Vacuole Mitochondria OM
Tweety family (3)	Swelling-dependent volume-regulated anion channel in astrocytes	PM
Sidoreflexin family (5)	Amino acid transport	Mitochondria IM
YIF1/YIP1 family (9)	COPII-coated ER to Golgi transport vesicle-mediated transport	Golgi apparatus, ER
AB hydrolase superfamily (2)	Hydrolase activity	PM
Emopamil-binding protein family (2)	Cholesterol biosynthesis Lipoprotein internalization	ER, PM, Vesicle, Vacuole
Dual specificity phosphatase catalytic domain (2)	Phosphatase activity	Golgi apparatus, ER
Metallophosphoesterase domain (2)	Hydrolase activity	Nucleoplasm, Mitochondria
TLC domain (2)	Possibly involved in lipid trafficking, metabolism, or sensing	ER, Nucleus

Three major families that contain more than two members and comprise nearly 30% of the dataset were identified: the tweety family, the sidoreflexin family, and the YIP1 domain family (YIPF). The tweety family contains three

members that localize to the plasma membrane and are subunits of the swelling-dependent volume-regulated anion channel in astrocytes. The sidoreflexin family includes five members that localize specifically to the inner or outer mitochondrial membrane and function in amino acid transport. Our phylogenetic analysis of the sidoreflexin family includes members of holomycota (fungi) as well as previously not yet investigated species of archaeplastida, which appear to resolve between the SFXN1/SFXN3/SFXN2 and SFXN5/SFXN4 clades with the fungi species forming more distant phylogenetic groups. The YIPF family includes nine proteins that localize to the Golgi apparatus and ER and function in COPII-coated ER to Golgi transport.

Approximately 60% of 5TMs are identified as cancer prognostic markers that are associated the clinical outcome of different tumour types. Renal, gynaecological, and liver cancers were the most common types of cancers associated with 5TMs.

Discussion and Conclusions

Evolution of the membrane proteome

In our extensive comparative analysis of the evolution of twenty-four eukaryotic membrane proteomes, 86% of the dataset – more than 105,000 proteins – are functionally characterized. We identify 604 conserved functional clusters which contain 90,906 conserved membrane proteins that are found across all investigated species. Thus nearly 75% of the predicted membrane proteins are described in functional clusters that extend from the LHCA, indicating that a substantial proportion of the diversity within the eukaryotic membrane proteome was already present at the origin of holozoans.

Distinct innovations, expansions as well as losses are found in membrane protein families that correlate with the development of specific biological systems. Proteins involved in neuronal and brain development are present in unicellular holozoans and undergo expansions as well as lineage-specific losses in the studied metazoan lineages. Enzymes that are involved in neuronal development and synaptic plasticity show increases in many metazoans and more than doubles in most vertebrate species. Voltage-gated ion channels, which are involved many activities in the nervous system, show both expansions and losses in different lineages. These developments are consistent with the fact that the nervous system was present early in animal evolution and has developed in complexity through the expansion and reduction of neuronal proteins.

Protein expansions throughout metazoans are also associated with developing biological systems. Proteins involved with cell adhesion, which facilitates movement, multicellularity, protein-protein interactions, communication, and signalling were shown to have increased in metazoans. In particular, components of intra- and extra-cellular communication are noted to be critical in the development of complex systems, and expansions in signalling pathways are seen in vertebrates. Protein families involved in innate immunity are found throughout metazoans but with distinct lineage-specific expansions particularly in invertebrates, such as in lophotrochozoans and echinoderms. Unique expansions of adaptive immune response components that are involved in crucial immunological roles are shown in vertebrates. This is consistent with other studies that suggest expansion of immune system components occurred in vertebrates and that the adaptive immune systems

emerged at the origin of vertebrates in conjunction with coevolution of innate defences [80].

This study highlights the patterns of protein family expansions, losses, and conservations across lineages. The potential effects of the environment and lifestyle of a species can be seen in the development of certain protein families, but also in reductions in the size of a species' genome and subsequently the subset of membrane proteins. The reasons behind gene loss are not fully understood, although a study on insects determined that gene loss correlates with the species rates of molecular evolution and radiation times, indicating the gene loss was due to higher evolutionary rate [81]. Another study surmised that in environments where the function of the gene is not necessarily needed, the gene is expendable and subsequently lost [82].

In conclusion, this analysis identifies the functional protein clusters that were present at the origin of metazoan evolution and defines the possible membrane proteome of the LHCA. The development of proteins involved in complex biological pathways such as the nervous system, immune response, cell communication, and multicellularity were facilitated through the expansion of protein families. Further, environmental challenges and species lifestyle contribute to shaping the patterns of gene expansions and reductions.

Trispanins functioning in transport mechanisms

The 152 trispanins include many evolutionarily conserved proteins that are primarily involved in membrane synthesis and lipid genesis, protein trafficking, catabolic processes, and in particular signal transduction as a result of the large ionotropic glutamate receptor family. Their functional activities are supported by where the majority of them are found: they are predominantly localized to intracellular organelles such as the endomembrane system including the ER and mitochondria, where they are over-represented in comparison to the human membrane proteome. More specifically, nearly 45% are localized to the ER and almost 15% localize to the Golgi apparatus, which are major sites of membrane lipid biosynthesis and also centres for directing protein trafficking.

Although ~15% of the identified trispanins are uncharacterized, we were able to reasonably speculate on the activities of these proteins using bioinformatic analyses and literature research. Surprisingly, several of these uncharacterized proteins interact with other trispanins in the dataset and some are even members of important membrane complexes. Our analysis shows many are localized to specific subcellular locations, including the ER and mitochondria. Moreover, many were expressed in specific tissues, including the cerebral cortex, testis, and spleen.

4TMs comprised of evolutionarily conserved groups

The characterization of the 373 4TM proteins represents 7% of the human membrane proteome. In total, proteins that function as transporters or engage in transport activities comprise 37% of the dataset and are the largest class of 4TMs; 17% contain an TCDB identifier, an additional 12% are in the dual function category, while 8% more of proteins with varied functional activities are included. Thus, the 4TM dataset contains more than double proportionally the amount of transporters as the human transmembrane proteome, which is ~15%. Moreover, the 4TM receptor class (23%) and enzyme class (13%) are roughly comparable to the human membrane proteome with 25% and 10%, respectively. This might suggest that the 4TM architecture is more favourable for transport activities than other functions.

An interesting aspect of the 4TM dataset resulting from the functional classification is that more than 60% of the dataset can be described by 8 large Pfam families and clans. Furthermore, up to 97% of the 4TM proteins contain homologous conserved regions that are found in other proteins, indicating very few unique singlet proteins in the dataset.

5TMs engaging in protein targeting and trafficking

Two of the primary functional activities of proteins identified with the 58 proteins in the dataset are the *establishment of localization* and *transport activity*. More than half of the proteins in the 5TM dataset are described as establishing the locales of proteins through engaging in movement, tethering, or selective degradation processes. And one-third of the dataset is described through various annotation sources as involved in transport activities which includes proteins that participate in importing, exporting or trafficking within or between cells. In conjunction with these activities, proteins were found to localize to specific organelles to carry out these activities in disproportionate amounts compared to the entire *Homo sapiens* membrane proteome. For example, proteins localized to the inner and outer membranes of the mitochondria and involved in mitochondrial transport were over-represented. And nearly 30% of the dataset localize to vesicles, including an over-representation of proteins that function in COPII-coated ER to Golgi transport vesicles. Furthermore, more than 40% localize to the nuclear outer membrane-endoplasmic reticulum membrane network and are also over-represented in comparison to the human membrane proteome.

It is interesting to note the types of transport activities the 5TM architecture facilitates, such as membrane trafficking, vesicle-mediated transport, and intracellular protein transport, for example. In contrast, it appears only two families form homo- or heteromeric subunits to create an actual channel or pore for solutes to cross membranes. Hence it appears the 5TM proteins

facilitate transport mechanisms, for example vesicle budding and trafficking, rather than forming oligomeric complexes that create a channel for the transport of molecules.

Limitations

One limitation that affects the studies in this thesis is the reliability of transmembrane prediction software. As mentioned previously, they can have difficulty discriminating different types of transmembrane regions. Also, as they are often trained on experimentally derived structures, and there is still a paucity of 3D membrane structures available, this can affect the reliability of the transmembrane predictions.

In paper I, additional reservations may apply. The inclusion of different species into the evolutionary comparison may affect the different patterns of conservation, expansion, and reduction of membrane proteins. For example, species that develop in different aquatic environments such as brackish, salt water, or fresh water may evolve different manner of adaptations. Another reservation may be in the choice of outgroups, which in turn could affect the extrapolation of the LHCA membrane proteome. There are large differences in genome sizes in unicellular eukaryotes, which would affect perhaps the number and types of conserved functional clusters as well as the estimation of the membrane proteome of the LHCA.

Perspectives

Our comprehensive analysis of three, four, and five membrane-spanning proteins gives further perspective on the correlation of structure and function. Based on similar gross structure, each group of proteins do seem to share over-representations in both functional activities and localization areas and can be briefly summarized in relation to other transmembrane groups in Figure 7.

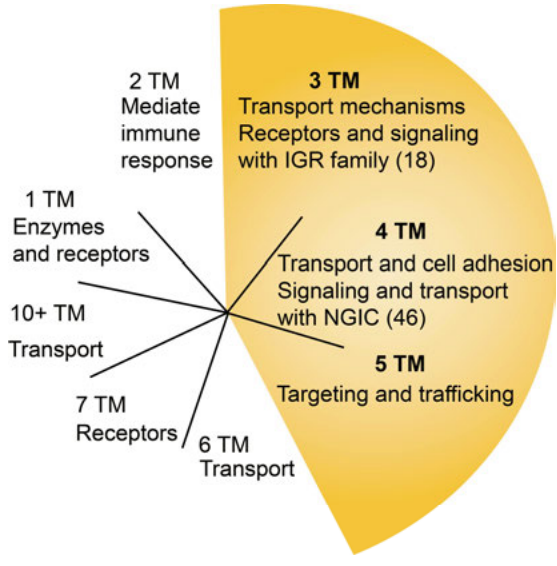

Figure 7. Common correlations between structure and function. The three, four, and five transmembrane proteins coloured in yellow are the focus in this thesis. NGIC: neurotransmitter gated ion channel, IGR: ionotropic glutamate receptors.

Furthermore, in investigating the evolution of membrane proteins, we can see that these three groups of proteins contain many evolutionarily conserved functional activities that are proposed to already exist at the origin of metazoan development, for example lipid synthesis, protein targeting mechanisms, and membrane trafficking. Indeed, the main receptor families identified in 3TMs, the ionotropic glutamate receptors, and in 4TMs, the neurotransmitter gated ion channels, show complex developments with significant lineage-specific expansions. Additionally, our analyses highlight uncharacterized transmembrane proteins that might be involved in possibly novel and important functional activities that are species specific, with special emphasis

on *Homo sapiens*. Thus, our results presents opportunities for further research and development by characterizing previously unstudied transmembrane proteins that may be involved in important cell pathways or engage in disease pathology.

In summary, this thesis examines membrane-bound proteins and presents comprehensive analyses of more than 10% of the human membrane proteome. We have identified and characterized the overarching functional activities, the primary localization areas, and the involvement in disease and possible drug targets of the previously unstudied three-, four-, and five-membrane spanning groups of proteins. Additionally, we have analysed the evolution of the membrane proteome from the LHCA throughout metazoan lineages to garner understanding of the conserved functional clusters that underlie metazoan innovation and diversity. In conclusion, we have investigated the origins and evolutionary patterns of development of human transmembrane proteins, characterized previously unstudied aspects of membrane proteins, and have extended the fundamental knowledge of transmembrane proteins.

Acknowledgements

Thank you to Helgi B. Schiöth for choosing me as a PhD student and creating an independent and autonomous learning environment, yet always ready to provide insight and guidance if asked, particularly in analysing and writing. Thank you to my invaluable colleague Mathias Rask-Andersen – you introduced me to pharmacoinformatics and trusted me with the dataset. Researching and writing projects tied to it have been some of the most fulfilling aspects of my PhD work.

Thank you to the wonderful friends and colleagues I have met through the Functional Pharmacology lab: Mike, Christian, Jessica, Gaia, Gull, Ahmed, Aniruddha, Christina, Arunkumar, Madeleine, Jörgen, and especially Frida, Olga, Diana-Maria, Helen, and Lieve.

Särskilt tack till den fantastiska personal som verkligen gjorde mitt liv mycket lättare. Var och en av er lyssnade på min svenska med min amerikanska accent och hade alltid tålmodig och var mycket proaktiv i att hjälpa mig: Mariana Hooli, Eva-Lena Wahlberg, Berit Hård-Wallenqvist, Maria Lindermann, Cecilia Yates, och Anders Nilsson.

My deepest gratitude goes to my family. My Mom – Nan J Attwood. What can I say that you do not already know; I am grateful you are my Mom. Thank you for teaching me curiosity about the world, for stubbornness and perseverance, for long-distance hugs no matter the country or continent and reminding me to eat the elephant(s) one bite at a time. Most importantly, thank you for your love and belief in me. My sisters and brothers: Kasey – you know how much you mean to me. I am so glad and grateful you came and visited here. I cannot wait until I am closer and get to bring you lattes and we can laugh together! Joni and Cade – I have missed you both - I am so proud of the young man Cade is growing into! Mike and Davy – thank you for being my big brothers! And Layla – I am very glad I have gotten to know you – welcome to the family! A special thank you to my Gramma. At times in hindsight you can the domino of decisions that led you to a certain place, and for me I can see a direct line from our conversations in October, 2002 to here. Thank you.

Jimmy. My husband. Min älskling. Thank you for more things than I can enumerate. From Antarctica to here. To continued adventures elsewhere. It could be worse.

References

1. Müller DJ, Wu N, Palczewski K. Vertebrate Membrane Proteins: Structure, Function, and Insights from Biophysical Approaches. Pharmacol Rev. 2008 Jan 3;60(1):43–78.
2. Nam H-J, Kim I, Bowie JU, Kim S. Metazoans evolved by taking domains from soluble proteins to expand intercellular communication network. Sci Rep. 2015 Apr 29;5:9576.
3. Elofsson A, von Heijne G. Membrane Protein Structure: Prediction versus Reality. Annu Rev Biochem. 2007;76(1):125–40.
4. von Heijne G. Membrane protein structure prediction: Hydrophobicity analysis and the positive-inside rule. J Mol Biol. 1992 May 20;225(2):487–94.
5. von Heijne G. Membrane-protein topology. Nat Rev Mol Cell Biol. 2006 Dec;7(12):909–18.
6. Lindner D, Walther C, Tennemann A, Beck-Sickinger AG. Functional role of the extracellular N-terminal domain of neuropeptide Y subfamily receptors in membrane integration and agonist-stimulated internalization. Cell Signal. 2009 Jan;21(1):61–8.
7. Bagnéris C, DeCaen PG, Hall BA, Naylor CE, Clapham DE, Kay CWM, et al. Role of the C-terminal domain in the structure and function of tetrameric sodium channels. Nat Commun [Internet]. 2013 Sep 19 [cited 2015 Feb 17];4. Available from: http://www.nature.com/ncomms/2013/130919/ncomms3465/full/ncomms3465.html
8. Cymer F, von Heijne G, White SH. Mechanisms of Integral Membrane Protein Insertion and Folding. J Mol Biol. 2015 Mar 13;427(5):999–1022.
9. Rapoport TA. Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. Nature. 2007 Nov 29;450(7170):663–9.
10. White SH. The messy process of guiding proteins into membranes. eLife. 2015 Nov 6;4:e12100.
11. De Marothy MT, Elofsson A. Marginally hydrophobic transmembrane α-helices shaping membrane protein folding. Protein Sci. 2015 Jul 1;24(7):1057–74.
12. Nature Topic Page. Cell Membranes | Learn Science at Scitable [Internet]. 2020 [cited 2020 Mar 22]. Available from: https://www.nature.com/scitable/topicpage/cell-membranes-14052567/
13. van Meer G, Voelker DR, Feigenson GW. Membrane lipids: where they are and how they behave. Nat Rev Mol Cell Biol. 2008 Feb;9(2):112–24.
14. Gabaldón T, Pittis AA. Origin and evolution of metabolic sub-cellular compartmentalization in eukaryotes. Biochimie. 2015 Dec 1;119:262–8.
15. Sommer MS, Schleiff E. Protein Targeting and Transport as a Necessary Consequence of Increased Cellular Complexity. Cold Spring Harb Perspect Biol. 2014 Jan 8;6(8):a016055.
16. Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Blal HA, et al. A subcellular map of the human proteome. Science. 2017 May 26;356(6340):eaal3321.

17. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015 Jan 23;347(6220):1260419.
18. Chen Y, Shanmugam SK, Dalbey RE. The Principles of Protein Targeting and Transport Across Cell Membranes. Protein J. 2019 Jun 1;38(3):236–48.
19. Bauer NC, Doetsch PW, Corbett AH. Mechanisms Regulating Protein Localization. Traffic. 2015;16(10):1039–61.
20. Cosson P, Perrin J, Bonifacino JS. Anchors Aweigh: Protein Traffic Mediated by Transmembrane Domains. Trends Cell Biol. 2013 Oct;23(10):511–7.
21. Shaik S, Pandey H, Thirumalasetti SK, Nakamura N. Characteristics and Functions of the Yip1 Domain Family (YIPF), Multi-Span Transmembrane Proteins Mainly Localized to the Golgi Apparatus. Front Cell Dev Biol [Internet]. 2019 [cited 2020 Mar 26];7. Available from: https://www.frontiersin.org/articles/10.3389/fcell.2019.00130/full
22. Goder V, Bieri C, Spiess M. Glycosylation Can Influence Topogenesis of Membrane Proteins and Reveals Dynamic Reorientation of Nascent Polypeptides within the Translocon. J Cell Biol. 1999 Oct 18;147(2):257–66.
23. Ree R, Varland S, Arnesen T. Spotlight on protein N-terminal acetylation. Exp Mol Med. 2018 Jul;50(7):1–13.
24. Ohno S, Wolf U, Atkin NB. Evolution from Fish to Mammals by Gene Duplication. Hereditas. 1968;59(1):169–87.
25. Long M, Betrán E, Thornton K, Wang W. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 2003 Nov;4(11):865–75.
26. Andersson DI, Jerlström-Hultqvist J, Näsvall J. Evolution of New Functions De Novo and from Preexisting Genes. Cold Spring Harb Perspect Biol [Internet]. 2015 Jun [cited 2020 Mar 22];7(6). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4448608/
27. Lee M-C, Marx CJ. Repeated, Selection-Driven Genome Reduction of Accessory Genes in Experimental Populations. PLOS Genet. 2012 May 10;8(5):e1002651.
28. Lynch M. The Origins of Eukaryotic Gene Structure. Mol Biol Evol. 2006 Feb 1;23(2):450–68.
29. Suga H, Dacre M, de Mendoza A, Shalchian-Tabrizi K, Manning G, Ruiz-Trillo I. Genomic survey of premetazoans shows deep conservation of cytoplasmic tyrosine kinases and multiple radiations of receptor tyrosine kinases. Sci Signal. 2012 May 1;5(222):ra35.
30. Suga H, Torruella G, Burger G, Brown MW, Ruiz-Trillo I. Earliest Holozoan expansion of phosphotyrosine signaling. Mol Biol Evol. 2014 Mar;31(3):517–28.
31. Krishnan A, Almén MS, Fredriksson R, Schiöth HB. The Origin of GPCRs: Identification of Mammalian like Rhodopsin , Adhesion , Glutamate and Frizzled GPCRs in Fungi. PLoS One. 2012 Jan 4;7(1):e29817.
32. de Mendoza A, Sebé-Pedrós A, Ruiz-Trillo I. The Evolution of the GPCR Signaling System in Eukaryotes: Modularity, Conservation, and the Transition to Metazoan Multicellularity. Genome Biol Evol. 2014 Feb 23;6(3):606–19.
33. Brunet T, King N. The Origin of Animal Multicellularity and Cell Differentiation. Dev Cell. 2017 Oct 23;43(2):124–40.
34. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? Nat Rev Drug Discov. 2006 Dec;5(12):993–6.
35. Hauser AS, Attwood MM, Rask-Andersen M, Schiöth HB, Gloriam DE. Trends in GPCR drug discovery: new agents, targets and indications. Nat Rev Drug Discov. 2017 Dec;16(12):829–42.

36. Beckwith KA, Byrd JC, Muthusamy N. Tetraspanins as therapeutic targets in hematological malignancy: a concise review. Front Physiol [Internet]. 2015 Mar 23 [cited 2020 Mar 24];6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4369647/
37. Pulgar VM. Transcytosis to Cross the Blood Brain Barrier, New Advancements and Challenges. Front Neurosci [Internet]. 2019 [cited 2020 Mar 24];12. Available from: https://www.frontiersin.org/articles/10.3389/fnins.2018.01019/full
38. Heukers R, van Bergen en Henegouwen PMP, Oliveira S. Nanobody–photosensitizer conjugates for targeted photodynamic therapy. Nanomedicine Nanotechnol Biol Med. 2014 Oct 1;10(7):1441–51.
39. Chau CH, Steeg PS, Figg WD. Antibody–drug conjugates for cancer. The Lancet. 2019 Aug 31;394(10200):793–804.
40. Carter PJ, Lazar GA. Next generation antibody drugs: pursuit of the 'high-hanging fruit'. Nat Rev Drug Discov. 2018 Mar;17(3):197–223.
41. Almén MS, Nordström KJ, Fredriksson R, Schiöth HB. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. BMC Biol. 2009;7:50.
42. Hicke L, Dunn R. Regulation of Membrane Protein Transport by Ubiquitin and Ubiquitin-Binding Proteins. Annu Rev Cell Dev Biol. 2003;19(1):141–72.
43. Almén MS, Nordström KJ, Fredriksson R, Schiöth HB. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. BMC Biol. 2009 Aug 13;7(1):50.
44. Schlessinger A, Yee SW, Sali A, Giacomini KM. SLC Classification: An Update. Clin Pharmacol Ther. 2013 Jul;94(1):19–23.
45. Arinaminpathy Y, Khurana E, Engelman DM, Gerstein MB. Computational analysis of membrane proteins: the largest class of drug targets. Drug Discov Today. 2009 Dec;14(23–24):1130–5.
46. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Costanzo LD, et al. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 2019 Jan 8;47(D1):D520–8.
47. Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI. OPM: Orientations of Proteins in Membranes database. Bioinformatics. 2006 Mar 1;22(5):623–5.
48. Kozma D, Simon I, Tusnády GE. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Res. 2013 Jan;41(Database issue):D524–9.
49. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011 Oct;8(10):785–6.
50. Armenteros JJA, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 2019 Apr;37(4):420.
51. Nilsson J, Persson B, von Heijne G. Consensus predictions of membrane protein topology. FEBS Lett. 2000 Dec 15;486(3):267–9.
52. Hennerdal A, Elofsson A. Rapid membrane protein topology prediction. Bioinformatics. 2011 May 1;27(9):1322–3.
53. Viklund H, Elofsson A. Best α-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. Protein Sci Publ Protein Soc. 2004 Jul;13(7):1908–17.
54. Tusnády GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics. 2001 Jan 9;17(9):849–50.

55. Jones DT, Taylor WR, Thornton JM. A Model Recognition Approach to the Prediction of All-Helical Membrane Protein Structure and Topology. Biochemistry. 1994;33(10):3038–49.
56. Bernsel A, Viklund H, Falk J, Lindahl E, Heijne G von, Elofsson A. Prediction of membrane-protein topology from first principles. Proc Natl Acad Sci. 2008 May 20;105(20):7177–81.
57. Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res. 2015 Jul 1;43(Web Server issue):W401–7.
58. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D427–32.
59. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 2017 Jan 4;45(Database issue):D200–3.
60. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506–15.
61. Saier MH, Reddy VS, Tsu BV, Ahmed MS, Li C, Moreno-Hagelsieb G. The Transporter Classification Database (TCDB): recent advances. Nucleic Acids Res. 2016 Jan 4;44(Database issue):D372–9.
62. IUPHAR/BPS Guide to PHARMACOLOGY [Internet]. [cited 2020 Mar 7]. Available from: https://www.guidetopharmacology.org/
63. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype–gene relationships. Nucleic Acids Res. 2019 Jan 8;47(Database issue):D1038–43.
64. Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: Text mining and data integration of disease–gene associations. Methods. 2015 Mar 1;74:83–9.
65. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017 Jan 4;45(Database issue):D833–9.
66. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science [Internet]. 2017 Aug 18 [cited 2020 Jan 29];357(6352). Available from: https://science.sciencemag.org/content/357/6352/eaan2507
67. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018 04;46(D1):D1074–82.
68. Rask-Andersen M, Masuram S, Schiöth HB. The Druggable Genome: Evaluation of Drug Targets in Clinical Trials Suggests Major Shifts in Molecular Class and Indication. Annu Rev Pharmacol Toxicol. 2014;54(1):9–26.
69. Attwood MM, Rask-Andersen M, Schiöth HB. Orphan Drugs and Their Impact on Pharmaceutical Development. Trends Pharmacol Sci. 2018 Jun 1;39(6):525–35.
70. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 2019 Jan 8;47(D1):D419–26.
71. Attwood MM, Krishnan A, Almén MS, Schiöth HB. Highly diversified expansions shaped the evolution of membrane bound proteins in metazoans. Sci Rep. 2017 Sep 28;7(1):12387.

72. Jain A, Tuteja G. TissueEnrich: Tissue-specific gene enrichment analysis. Bioinformatics. 2019 Jun 1;35(11):1966–77.
73. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002 Apr 1;30(7):1575–84.
74. Azad A, Pavlopoulos GA, Ouzounis CA, Kyrpides NC, Buluç A. HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks. Nucleic Acids Res. 2018 Apr 6;46(6):e33–e33.
75. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 2019 Jul 19;20(4):1160–6.
76. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009 May 1;25(9):1189–91.
77. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat Rev Genet. 2012 May;13(5):303–14.
78. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001 Aug 1;17(8):754–5.
79. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May 1;30(9):1312–3.
80. Boehm T, Iwanami N, Hess I. Evolution of the Immune System in the Lower Vertebrates. Annu Rev Genomics Hum Genet. 2012;13(1):127–49.
81. Wyder S, Kriventseva EV, Schröder R, Kadowaki T, Zdobnov EM. Quantification of ortholog losses in insects and vertebrates. Genome Biol. 2007;8(11):R242.
82. Albalat R, Cañestro C. Evolution by gene loss. Nat Rev Genet. 2016 Jul;17(7):379–91.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1660

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-407778

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2020