A computational and experimental study of protein localisation determinants in the mammalian endomembrane system

Tilen Kranjc, MPharm

The thesis is submitted to University College Dublin in fulfilment of the requirements for the degree of Doctor of Philosophy in Bioinformatics and Systems Biology

School of Biology and Environmental Science

Head of School: Prof. Jeremy C. Simpson
Supervisors: Prof. Jeremy C. Simpson, Prof. Denis C. Shields
Doctoral Studies Panel chair: Prof. Emma Teeling

University College Dublin
May 2015
Za babi
# Contents

Abstract v
Acknowledgements vii
Statement of Authorship viii
List of Abbreviations ix
List of Figures xii
List of Tables xv

1 Introduction 1
1.1 Compartmentalisation in eukaryotic cells 2
   1.1.1 Endoplasmic reticulum 2
   1.1.2 Golgi apparatus 7
   1.1.3 Protein targeting to the endomembrane system 9
1.2 Membrane trafficking 11
   1.2.1 Machinery 11
       1.2.1.1 Rab GTPases 12
       1.2.1.2 Transmembrane PRAF and YIPF protein families 17
           1.2.1.2.1 Prenylated Rab acceptors 17
           1.2.1.2.2 YPT interacting proteins 22
       1.2.1.3 Vesicle coat, budding and fusion proteins 24
   1.2.2 Mechanisms of membrane trafficking 29
       1.2.2.1 Trafficking between the ER and Golgi 30
       1.2.2.2 post-Golgi trafficking 31
1.3 Short linear motifs - SLiMs 32
   1.3.1 Properties of SLiMs 34
   1.3.2 Discovery methods 35
   1.3.3 SLiMs facilitating protein localisation 36
1.4 Imaging and image analysis methods in cell biology 36
   1.4.1 Light microscopy 37
CONTENTS

1.4.2 Electron microscopy ................................................. 38
1.4.3 Image analysis ..................................................... 39
1.5 Aims of the study ...................................................... 41

2 Materials and Methods ................................................ 42
2.1 Computational discovery of novel SLiMs ............................ 43
  2.1.1 Localisation dataset .............................................. 43
  2.1.2 SLiMFinder - SLiM discovery tool .............................. 43
  2.1.3 SLiMProb - Search tool for predefined SLiMs ............... 43
2.2 Cloning ............................................................... 43
2.3 Cell culture .......................................................... 45
2.4 DNA transfection ...................................................... 45
2.5 siRNA transfection .................................................... 45
2.6 Quantitative PCR (qPCR) ............................................. 46
2.7 Fixation and immunostaining ....................................... 46
2.8 Co-immunoprecipitation ............................................. 46
2.9 Mass spectrometry and data analysis ............................... 47
2.10 Western blotting ........................................................ 48
2.11 Fluorescence Protease Protection (FPP) assay ..................... 48
2.12 Confocal imaging ..................................................... 49
2.13 Epifluorescence microscopy ....................................... 49
2.14 Image analysis ...................................................... 50
  2.14.1 Co-localisation with RWC ...................................... 50
  2.14.2 Localisation with texture features ........................... 50

3 Discovery of novel SLiMs involved in protein localisation .... 51
3.1 Introduction ........................................................... 52
3.2 Results ............................................................... 53
  3.2.1 Compilation of the protein localisation dataset ................. 53
  3.2.2 Discovery of SLiMs with SLiMFinder .......................... 58
  3.2.3 Experimental validation ......................................... 64
3.3 Discussion ........................................................... 65

4 Characterisation of PRAF proteins .................................. 69
4.1 Introduction ........................................................... 70
4.2 Results ............................................................... 72
  4.2.1 Design of GFP- and HA-tagged constructs ..................... 72
  4.2.2 Localisation ...................................................... 73
  4.2.3 Membrane topology ............................................... 80
  4.2.4 Co-localisation with organelle markers and Rab proteins .... 80
  4.2.5 Interaction network ............................................... 87
  4.2.6 Functional studies ............................................... 88
4.3 Discussion ........................................................... 94
## 5 Characterisation of YIPF proteins

5.1 Introduction ................................................................. 99

5.2 Results ................................................................. 100

5.2.1 Design of GFP-tagged YIPF constructs .............................................. 100

5.2.2 Localisation ................................................................. 100

5.2.3 Membrane topology ................................................................. 106

5.2.4 Co-localisation with organelle markers .............................................. 109

5.2.5 Golgi morphology ................................................................. 111

5.2.6 Interaction network ................................................................. 112

5.3 Discussion ................................................................. 115

## 6 Classification of cell phenotypes using image texture features

6.1 Introduction ................................................................. 120

6.2 Results ................................................................. 122

6.2.1 Construction of localisation dataset .............................................. 122

6.2.2 Development of the image analysis pipeline .............................................. 122

6.2.3 Unsupervised clustering of texture features .............................................. 124

6.3 Discussion ................................................................. 128

## 7 General discussion and future prospects

Appendices

A Sequences and primers ................................................................. 137

B Source code of analysis scripts ................................................................. 141

C Permissions for reproduction of figures ................................................................. 150

Bibliography
Abstract

The subcellular localisation of a protein, together with its sequence and structure, provide the first information about its function. Although several approaches for determining localisation exist, the most widely used experimental methods involve the overexpression of a GFP-tagged construct of the protein in a cultured cell or staining the endogenous protein with specific fluorescently-labelled antibodies. Computational sequence-based localisation predictors have been developed, and are of value, but they are still limited in their predictive power. Considering the now extensive use of imaging approaches, it is therefore unsurprising that much effort has been put into the development of automated image analysis methods to classify localisation. Image classifiers are typically trained on ground truth data, which introduces certain bias. The aim of modern algorithms is to eliminate human interaction and instead perform unsupervised classification of the images. The study presented in this thesis addresses three aspects of protein localisation methodology: sequence-based localisation prediction with short linear motifs (SLiMs), unsupervised image analysis with texture features and experimental determination of protein localisation.

SLiMs are 3-12 amino acid long linear peptides enriched in disordered regions of proteins that interact with domains of other proteins. The aim of this study was to search for novel targeting SLiMs in a dataset of proteins for which their localisation had already been experimentally determined. The software SLiMFinder was used to find and evaluate overrepresented regions in the sequences grouped by localisation, topology or the effect of the fluorescent protein on the localisation. The discovery was performed in the whole sequences, limited to disordered regions or limited to the residues at each termini. None of the above approaches yielded any statistically significant SLiM candidate. However, the SLiM F.M.AE, that was discovered among Golgi-localising proteins, was borderline significant and therefore selected for experimental evaluation. A peptide with this SLiM sequence was synthesised, attached to a GFP and transfected into cells. The localisation observed was similar to that seen of the soluble GFP alone, therefore suggesting that this putative SLiM was unable to target GFP to the Golgi.

The experimental localisation data used to discover novel targeting SLiMs were manually annotated, which may introduce certain bias. Therefore an image analysis pipeline to automatically group localisation profiles based on image texture features was developed. As a test, Haralick, Gabor and Threshold Adjacency Statistics texture features were extracted from images of cells overexpressing GFP-tagged variants of two protein families (PRAF and YIPF) localising across various compartments of the endomembrane system. The dataset was first plotted and visually in-
spected by reducing the dimensionality with principal components analysis. This revealed two clearly separated clusters of localisation phenotypes: one small cluster corresponded to the Golgi and punctate structures; and a large cluster corresponded to ER, Golgi and punctate structures. The large cluster could further be divided into cells containing punctae and cells lacking them. The dataset was then automatically clustered with a Gaussian model-based clustering method. Three clusters were identified, which matched the human interpretation of the localisation phenotypes.

The second part of this study focused on experimental characterisation of the localisation and putative function of the PRAF and YIPF protein families. The PRAF proteins were found to localise to the ER and co-localised with a tubular ER marker; they also co-localised to varying extents with Rab3a, Rab4a, Rab13, Rab8a, Rab1a and Rab1b. PRAF proteins have four transmembrane domains with their N- and C-termini oriented towards the cytosol. Deletion mutants, overexpressed in cultured cells, did not exit the ER, indicating that the transmembrane domains, organised into hairpin-like structures, are sufficient for ER localisation. Co-precipitation and mass spectrometry approaches identified a large number of potential PRAF-interacting partners, including proteins associated with lipid biosynthesis pathways and ER and mitochondrial processes. Functional studies showed that PRAF1 affects the morphology of the mitochondria: depletion of PRAF1 significantly increased the size of mitochondria, whereas overexpression of PRAF1 significantly decreased it.

YIPF proteins were found to localise to the Golgi complex: YIPF1, YIPF2 and YIPF6 predominantly co-localised with the trans-Golgi marker TGN46; whereas YIPF3, YIPF4, YIPF5 and YIPF7 predominantly co-localised with the cis-Golgi marker GM130. The YIPF protein family all possess five transmembrane domains, and fluorescence protease protection assays determined that they were oriented with their N-terminus in the cytosol and their C-terminus in the Golgi lumen. Putative interaction partners identified by co-precipitation and mass spectrometry suggested roles in linkage to the cytoskeleton and a function in cell migration.

This study addresses computational and experimental aspects of the subcellular localisation and function of proteins, with specific focus on proteins of the PRAF and YIPF families. The human 'localisome' is incomplete, and therefore additional methodology and information to further refine this knowledge is vital. Image analysis techniques contribute to these efforts and in the future further advances in cytometry and automated interpretation of image data are expected.
Acknowledgements

The journey I started four years ago is coming to an end. I would like to thank the people that accompanied me on this journey and turned it into an unforgettable experience.

First I would like to thank my supervisors Prof. Jeremy Simpson - Jez and Prof. Denis Shields for the opportunity to work with them and continuous support throughout the studies and writing of this thesis. Jez, thank you for introducing me into the beautiful world of dark rooms, lasers and optics (it’s microscopy, not laser tag). Thank you also for friendly advice and nerdy chats about computers and airplanes. Denis, thank you for helping me discover my love for numbers (and ‘characters’). Thank you both for being an example of good scientists and supervisors and for convincing me that a career in science can be fun and exciting.

Thanks to George, Angela, Linda, Kenan, Maeve, Mariana, Niamh, Badriah and Eugene for being the best team. Thanks also to Elaine, Juan and Vasanth, who became my friends from the very first day in Dublin. Thanks also to Katey and Ciara.

Thanks to Bronagh and Ant for making sure the money paperwork was in place and the SBES school office for making sure the school paperwork was in place. Thanks to IRC for providing a generous grant that allowed several generations of PhD students to evolve into great scientists. Thanks also to Ad Futura for support.

Thanks to Skylab crew for sharing the hours of sweat and cold on the rooftop of SBES. Thanks also to Batlab for friendship, laughs and non-scientific discussions. Thanks to SBES technical team, particularly Gwyneth and Brendan.

Several other people accompanied me on this journey. Thanks to UCD Choral scholars and Enchiriadis for all the music and Detelca for reminding me of my homeland. Special thanks goes to all of my friends, in particular Urban for helping me plan this journey, Mauro and George for giving me an excuse to take a break, Miha for everything I know about reproduction (of cows, of course), Jaka for reminding me when the weekend starts and all the others.

The last paragraph is reserved for the special people that are the closest to me and accompanied me from the beginning to the end and beyond. First thanks to my mom, dad and my two brothers for love and support. Although far away, you were always with me, side by side. And finally, to the most loved Bojana for all the love and patience. Thank you for being brave and taking this journey with me. Thank you also for tolerating my princess-like behaviour in last few weeks (I think you actually enjoyed it) and providing all the treats. And most importantly, thanks for listening and providing friendly advice in the hard times of this journey.
Statement of Authorship

I hereby certify that the submitted work is my own work, was completed while registered as a candidate for the degree stated on the Title Page, and I have not obtained a degree elsewhere on the basis of the research presented in this submitted work.
List of Abbreviations

3D  Three-dimensional
ADP  Adenosine diphosphate
ATP  Adenosine triphosphate
BCA  Bicinchoninic acid
BIC  Bayesian information criterion
BSA  Bovine serum albumin
CCD  Charge-coupled device
CCV  Clathrin coated vesicles
cDNA  Complementary DNA
CFP  Cyan fluorescent protein
CLEM  Correlative light-electron microscopy
CORVET  Class C core vacuole/endosome tethering
DMEM  Dulbecco’s Modified Eagle Medium
DNA  Deoxyribonucleic acid
DTT  Dithiothreitol
ECM  Extracellular matrix
EDTA  Ethylenediaminetetraacetic acid
EGF  Epidermal growth factor
ELM  Eukaryotic Linear Motifs database
EM  Electron microscopy
ER  Endoplasmic reticulum
ERES  Endoplasmic reticulum exit sites
ERGIC  Endoplasmic reticulum-Golgi intermediate compartment
ERMES  ER-mitochondria encounter structure
FCS  Fetal calf serum
FDR  False discovery rate
FPP  Fluorescence protease protection
FRAP  Fluorescence recovery after photobleaching
GAP  GTPase activating protein
GDF  GDI displacement factor
GDI  GDP dissociation inhibitor
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDP</td>
<td>Guanosine diphosphate</td>
</tr>
<tr>
<td>GEF</td>
<td>Guanine nucleotide exchange factor</td>
</tr>
<tr>
<td>GET</td>
<td>Golgi ER trafficking complex</td>
</tr>
<tr>
<td>GFP</td>
<td>Green fluorescent protein</td>
</tr>
<tr>
<td>GGT</td>
<td>Geranylgeranyl transferase</td>
</tr>
<tr>
<td>GPI</td>
<td>Glycosylphosphatidylinositol</td>
</tr>
<tr>
<td>GRASP</td>
<td>Golgi reassembly stacking proteins</td>
</tr>
<tr>
<td>GTP</td>
<td>Guanosine triphosphate</td>
</tr>
<tr>
<td>HA</td>
<td>Hemagglutinin</td>
</tr>
<tr>
<td>HCS</td>
<td>High-content screening</td>
</tr>
<tr>
<td>HEPES</td>
<td>4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid</td>
</tr>
<tr>
<td>HOPS</td>
<td>Homotypic fusion and protein sorting</td>
</tr>
<tr>
<td>IMM</td>
<td>Inner mitochondrial membrane</td>
</tr>
<tr>
<td>LB</td>
<td>Lysogeny broth</td>
</tr>
<tr>
<td>LDL</td>
<td>Low density lipoprotein</td>
</tr>
<tr>
<td>LDLR</td>
<td>Low density lipoprotein receptor</td>
</tr>
<tr>
<td>M6PR</td>
<td>Mannose-6-phosphate receptor</td>
</tr>
<tr>
<td>MAM</td>
<td>Mitochondria associated membrane</td>
</tr>
<tr>
<td>MCS</td>
<td>Multiple cloning site</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger ribonucleic acid</td>
</tr>
<tr>
<td>MTC</td>
<td>Multisubunit tethering complexes</td>
</tr>
<tr>
<td>MVB</td>
<td>Multivesicular body</td>
</tr>
<tr>
<td>NES</td>
<td>Nuclear export signal</td>
</tr>
<tr>
<td>NLS</td>
<td>Nuclear localisation signal</td>
</tr>
<tr>
<td>OMM</td>
<td>Outer mitochondrial membrane</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>PAGE</td>
<td>Polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>PBS</td>
<td>Phosphate buffered saline</td>
</tr>
<tr>
<td>PC</td>
<td>Principal component</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal components analysis</td>
</tr>
<tr>
<td>PCh</td>
<td>Phosphatidylcholine</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PE</td>
<td>Phosphatidylethanolamine</td>
</tr>
<tr>
<td>PFA</td>
<td>Paraformaldehyde</td>
</tr>
<tr>
<td>PI</td>
<td>Phosphatidylinositol</td>
</tr>
<tr>
<td>PI3K</td>
<td>Phosphoinositol 3 kinase</td>
</tr>
<tr>
<td>PI-4,5-P2</td>
<td>Phosphatidylinositol-4,5-bisphosphate</td>
</tr>
<tr>
<td>RAF</td>
<td>Prenylated Rab acceptor family</td>
</tr>
<tr>
<td>PS</td>
<td>Phosphatidylserine</td>
</tr>
<tr>
<td>PVDF</td>
<td>Polyvinylidene fluoride</td>
</tr>
<tr>
<td>qPCR</td>
<td>Quantitative polymerase chain reaction</td>
</tr>
<tr>
<td>REP</td>
<td>Rab escort protein</td>
</tr>
</tbody>
</table>
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>RER</td>
<td>Rough endoplasmic reticulum</td>
</tr>
<tr>
<td>RHD</td>
<td>Reticulon homology domain</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>RNC</td>
<td>Ribosome-nascent protein chain complex</td>
</tr>
<tr>
<td>RT-qPCR</td>
<td>Reverse transcription quantitative polymerase chain reaction</td>
</tr>
<tr>
<td>RWC</td>
<td>Rank weighted coefficient</td>
</tr>
<tr>
<td>SDS-PAGE</td>
<td>Sodium dodecyl sulphate polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>SER</td>
<td>Smooth endoplasmic reticulum</td>
</tr>
<tr>
<td>siRNA</td>
<td>Small interference RNA</td>
</tr>
<tr>
<td>SLIM</td>
<td>Short linear motif</td>
</tr>
<tr>
<td>SMP</td>
<td>Synaptotagmin-like, mitochondrial and lipid-binding proteins</td>
</tr>
<tr>
<td>SNARE</td>
<td>Soluble N-ethylmaleimide-sensitive factor attachment protein receptor</td>
</tr>
<tr>
<td>SP</td>
<td>Signal peptide</td>
</tr>
<tr>
<td>SPT</td>
<td>Single particle tracking</td>
</tr>
<tr>
<td>SRP</td>
<td>Signal recognition particle</td>
</tr>
<tr>
<td>TAC</td>
<td>Tip attachment complex</td>
</tr>
<tr>
<td>TAS</td>
<td>Threshold adjacency statistics</td>
</tr>
<tr>
<td>TBS</td>
<td>Tris-buffered saline</td>
</tr>
<tr>
<td>TEM</td>
<td>Transmission electron microscope</td>
</tr>
<tr>
<td>TFA</td>
<td>Trifluoroacetic acid</td>
</tr>
<tr>
<td>TGN</td>
<td>Trans-Golgi network</td>
</tr>
<tr>
<td>TMD</td>
<td>Transmembrane domain</td>
</tr>
<tr>
<td>TRAPP</td>
<td>Transport protein particle</td>
</tr>
<tr>
<td>TTBS</td>
<td>Tween TBS - TBS supplemented with 0.1% polysorbate 20</td>
</tr>
<tr>
<td>TULIP</td>
<td>Tubular lipid-binding proteins</td>
</tr>
<tr>
<td>UDP</td>
<td>Uridine diphosphate</td>
</tr>
<tr>
<td>UPC</td>
<td>Unrelated protein cluster</td>
</tr>
<tr>
<td>VSV-G</td>
<td>Vesicular stomatitis Indiana virus G-protein</td>
</tr>
<tr>
<td>YFP</td>
<td>Yellow fluorescent protein</td>
</tr>
<tr>
<td>YIPF</td>
<td>Ypt interacting protein family</td>
</tr>
</tbody>
</table>
# List of Figures

1.1 Overview of the organelles and trafficking pathways in endomembrane system of a mammalian cell. ........................................... 3
1.2 Sheets and tubules of ER ............................................. 4
1.3 Four types of ER shaping by cytoskeleton .......................... 5
1.4 Membrane topology of ER membrane shaping proteins ........ 6
1.5 Seven steps of vesicle trafficking ................................... 12
1.6 Localisation map of Rab proteins in mammalian cell .......... 13
1.7 Rab activation cycle .................................................. 14
1.8 Membrane topology of PRAF and YIPF proteins ............... 18
1.9 Localisation and function map of SNARE proteins in a mammalian cell ........................................... 29
3.1 The dataset workflow ................................................. 54
3.2 Proteins that changed localisation when fused with a fluorescent protein ........................................... 55
3.3 Features used to classify the proteins based on their membrane topology ........................................... 57
3.4 Counts of the proteins in the three topology classes ........... 57
3.5 The masking strategies used in the SLiM discovery ............. 58
3.6 The interaction networks of the proteins with SLiM instances ........................................... 61
3.7 Images of live HeLa Kyoto cells expressing the GFP-tagged peptide corresponding to the putative Golgi SLiM ........................................... 65
3.8 Images of fixed HeLa Kyoto cells expressing the GFP-tagged peptide corresponding to putative Golgi SLiM ........................................... 65
4.1 Schematics of the designed PRAF constructs ................. 73
4.2 Agarose gels of the restriction digestion of the wild-type and the mutant PRAF constructs .............................. 74
4.3 Localisation of the GFP-tagged PRAF proteins ................ 75
4.4 Localisation of GFP- and HA-tagged PRAF proteins (part 1) ........................................... 76
4.5 Localisation of GFP- and HA-tagged PRAF proteins (part 2) ........................................... 77
4.6 Localisation of GFP- and HA-tagged PRAF proteins (part 3) ........................................... 78
4.7 Localisation of GFP-tagged PRAF mutants ...................... 79
4.8 Membrane topology of PRAF proteins .......................... 81
4.9 Comparison of PRAF localisation to ER tubule and sheet markers ........................................... 82
4.10 Images of co-localisation of PRAF proteins with the tubular ER marker REEP5 or the sheet ER marker CLIMP-63. 82
4.11 Co-localisation between PRAF proteins and GM130. 83
4.12 Image processing diagram for RWC measurements. 85
4.13 Heatmap showing the RWC co-localisation values between GFP-PRAF and mCherry-Rab proteins. 85
4.14 Images and co-localisation between GFP-PRAF and mCherry-Rab1a and Rab1b proteins. 86
4.15 Venn diagram of counts of ‘hits’ for each GFP-PRAF bait. 87
4.16 Heatmap showing fold-changes of selected identified proteins that co-immunoprecipitated with N-terminally GFP-tagged PRAF proteins. 88
4.17 Overexpression of PRAF-GFP proteins and immunostaining of REEP5 and CLIMP-63. 89
4.18 Relative PRAF mRNA levels after siRNA transfection. 89
4.19 Depletion of PRAF proteins and immunostaining of REEP5 and CLIMP-63. 90
4.20 Areas of tubular and sheet ER in cells depleted for PRAF proteins. 91
4.21 Diagram of the ER-CFP construct. 91
4.22 Intensity of the ER-CFP fluorescence signal in cells depleted for PRAF1. 91
4.23 Image analysis pipeline for measuring mitochondria size. 92
4.24 The ER and mitochondria in live cells following either depletion or overexpression of PRAF1. 93
4.25 Measurements of the width and the length of the mitochondria in cells either overexpressing or depleted for PRAF1. 94
5.1 Agarose gels of the diagnostic digestion reactions of YIPF GFP-tagged constructs. 101
5.2 Localisation phenotypes of N- and C-terminally GFP-tagged YIPF constructs. 102
5.3 N- and C-terminally GFP-tagged YIPF1, YIPF2 and YIPF6 stained with corresponding anti-YIPF antibodies. 103
5.4 N- and C-terminally GFP-tagged YIPF3, YIPF4 and YIPF5 stained with corresponding anti-YIPF antibodies. 104
5.5 N- and C-terminally GFP-tagged YIPF7 stained with anti-YIPF5 antibodies. 105
5.6 Rank weighted coefficient of co-localisation between GFP-tagged YIPF proteins and anti-YIPF antibody staining. 105
5.7 SDS-PAGE and western blotting of lysates of cells expressing GFP-YIPF proteins. 106
5.8 Profiles of GFP digestion in a fluorescence protease protection assay. 107
5.9 Membrane topology model of YIPF proteins. 108
5.10 Quantification of co-localisation of GFP-tagged YIPF proteins with GM130 and TGN46. 108
5.11 Linear profiles and images of co-localisation of GFP-tagged YIPF1, YIPF2 and YIPF6 with the cis-Golgi marker GM130 and the trans-Golgi marker TGN46. 109
5.12 Linear profiles and images of co-localisation of GFP-tagged YIPF3, YIPF4, YIPF5 and YIPF7 with the cis-Golgi marker GM130 and the trans-Golgi marker TGN46. 110
5.13 Area and count of Golgi fragments after depletion of YIPF proteins. 111
5.14 The effect of YIPF depletion on Golgi morphology. 112
LIST OF FIGURES

5.15 Fold-changes of selected identified proteins co-immunoprecipitated with N-terminally GFP-tagged YIPF proteins. ......................... 113
5.16 Fold-changes of identified cytoskeletal, YIPF, PRAF, Rho GTPases, Rab GTPases and related proteins co-immunoprecipitated with N-terminally GFP-tagged YIPF proteins. ........................................ 114
6.1 The threshold adjacency statistics (TAS) texture features. ....................... 121
6.2 The localisation profiles of PRAF and YIPF proteins. ............................... 123
6.3 The image analysis pipeline. ................................................................. 124
6.4 The data analysis pipeline exploiting texture features. ................................ 125
6.5 Plots of the first four principal components of the texture features. ......... 126
6.6 Bayesian Information Criterion (BIC) plot for the determination of the optimal model and the optimal number of clusters. ...................... 127
6.7 Unsupervised clustering of the texture features. ........................................ 127
List of Tables

1.1 Coat complexes and associated trafficking pathways. .......................... 25
1.2 List and description of metacharacters used in SLiM regular expressions. .... 34
2.1 Amounts of reagents used in DNA transfections. ................................. 45
2.2 Amounts of reagents used in siRNA transfections. ............................... 46
2.3 Antibodies used in immunofluorescence staining and corresponding dilutions used. 47
2.4 Antibodies and corresponding dilutions used in western blotting. ............... 48
2.5 List of fluorophores used with corresponding excitation lasers and emission filters on an FV1000 laser scanning confocal microscope. ................. 49
2.6 List of filters on the Leica DMI6000B system. ..................................... 49
3.1 Statistics of the dataset with excluded proteins with N-terminal signal or transit peptides. ................................................................. 56
3.2 Example contingency table for putative Golgi SLiM F.M.AE ..................... 59
3.3 The motifs discovered in the localisation dataset. ................................ 60
3.4 The motifs discovered in the topology groups. ..................................... 62
3.5 The motifs discovered in the terminal regions of the proteins whose localisation was affected by the fluorescent tag. ................................. 62
3.6 The experimentally evaluated SLiM with the Sig-value, the enrichment factor and the enrichment p-values. ............................................. 64
3.7 The experimentally evaluated SLiM with the example protein, the localisation class, the peptide and the DNA sequence. ................................. 64
4.1 A list of the designed PRAF fusion constructs. ..................................... 72
4.2 Potential targeting motifs in PRAF proteins retrieved from the ELM database. .. 75
4.3 Rab proteins used as organelle markers and their localisation. ................. 84
5.1 Summary of YIPF localisation to subdomains of the Golgi. ...................... 109
6.1 List of texture features measured in the images, .................................. 124
6.2 A comparison of the visually interpreted localisation class and the classification based on the unsupervised model-based clustering. .................. 128
Chapter 1

Introduction
CHAPTER 1. INTRODUCTION

1.1 Compartmentalisation in eukaryotic cells

A key feature that evolved in eukaryotic cells is the formation of cellular compartments or organelles. These compartments are aqueous spaces, separated from the cytosol by a phospholipid bilayer, which enclose biological processes that need either spatial separation (such as lysosomes, where molecules are degraded) or need conditions that are different from those found in the cytosol (for example oxidative conditions in the endoplasmic reticulum). Seminal work on characterising the endomembrane system (i.e. the system of internal membranes) was done by George Palade, who was also the first to describe the exchange of cargo between compartments (Palade 1975). Cargo is packaged into membrane bound carriers budding from a donor membrane, translocated by active transport mechanisms involving the cytoskeleton and cellular motors to an acceptor membrane where the carriers then fuse in turn delivering their cargo. Such trafficking is a very selective process and each trafficking interface carefully selects which cargo will be trafficked from the donor compartment and which carriers will be allowed to fuse with the acceptor membrane.

Membrane trafficking is particularly important to deliver proteins from the place where they are synthesised to their place of action - these two locations are often separated from each other. An overview of the endomembrane system is illustrated on Figure 1.1. Proteins of the endomembrane system are synthesised in the endoplasmic reticulum (ER) and then travel through the maze of ER tubules to its trafficking interface termed ER exit sites (ERES). Here they are recognised by cargo receptors, packaged into budding transport carriers and trafficked towards the Golgi complex. Following transit through the Golgi complex, a further series of transport carriers dispatch the cargo to their place of action, which can be an endosome, lysosome, plasma membrane or even extracellular environment. This high state of continuous membrane flux means that the endomembrane system needs to be highly regulated, and therefore requires a large number of auxiliary proteins as part of its regulatory and maintenance mechanisms.

1.1.1 Endoplasmic reticulum

The ER is the largest volume component of the endomembrane system, making up approximately half of the space in the cell. It was first observed in 1902 by Emilio Veratti, but it took another 50 years (and the invention of electron microscope) until Veratti’s discovery was accepted by the scientific community (Veratti 1961). A more detailed overview of ER structure was described by George Palade and Keith Porter (Palade et al. 1954; Palade 1955; Porter et al. 1957). Generally it has two parts: the nuclear envelope and the peripheral ER. The nuclear envelope is a cistern enclosed in two layers of membrane bilayer wrapping the chromatin, whereas the peripheral ER consists of a network of cisternae and tubules spreading throughout the cell, from its centre out to the plasma membrane. Although both parts are connected with each other into a single compartment, they do differ in structure, function and constitution. Several subdomains can be visually distinguished from each other at the level of the electron microscope.

The peripheral ER comes in two forms: sheet-like cisternae and a network of tubules with three-way junctions (Figure 1.2). ER sheets are flat elongated structures with low curvature with
the exception of the edges where the membrane folds upon itself. In contrast, the tubular ER is made of small high curvature tubules spreading from the centre of the cell to the very edge of it. Interestingly, the diameter of a tubule and thickness of a sheet is similar, approximately 50 nm (Bernales et al. 2006; Shibata et al. 2010). Tubules have high surface-to-volume ratio rendering them better suited for surface-dependent processes and sheets more suitable for volume-dependent processes. Ribosomes usually attach to ER sheets, which gives them the characteristic ‘rough’ look, whereas they do not attach to ER tubules which look ‘smooth’ (Voeltz et al. 2002; Shibata et al. 2006). The main role of the rough ER (RER) is the biosynthesis of membrane and transmembrane proteins, their folding and modification. The smooth ER (SER) on the other

Figure 1.1: Overview of the organelles and trafficking pathways in endomembrane system of a mammalian cell. The biosynthetic or secretory pathway is labelled with grey arrows and the endocytic or retrograde pathway is labeled with black arrows. Unconventional pathways are labelled with dashed arrows. Reproduced with permission from De Matteis and Luini 2011, copyright Massachusetts Medical Society.
CHAPTER 1. INTRODUCTION

Figure 1.2: Sheets and tubules of ER. Sheets are flat structures with a high volume-to-surface ratio which is better suited for luminal processes, such as protein synthesis. The distance between two sheets is defined by two CLIMP63 proteins and high-curvature regions at the edges are induced by reticulons. Synthesised proteins then flow to tubules (the flow is marked with white arrows). The high curvature of tubules is actively shaped by ER shaping proteins, such as reticulons. Tubules have a high surface-to-volume ratio and are more suitable for surface processes. Reproduced with permission from Westrate et al. 2015, copyright Annual Reviews.

Sheets and tubules of ER. Sheets are flat structures with a high volume-to-surface ratio which is better suited for luminal processes, such as protein synthesis. The distance between two sheets is defined by two CLIMP63 proteins and high-curvature regions at the edges are induced by reticulons. Synthesised proteins then flow to tubules (the flow is marked with white arrows). The high curvature of tubules is actively shaped by ER shaping proteins, such as reticulons. Tubules have a high surface-to-volume ratio and are more suitable for surface processes. Reproduced with permission from Westrate et al. 2015, copyright Annual Reviews.

hand is responsible for the synthesis of lipids and sterols and it also forms contacts with other organelles. The transitional ER, also known as ERES, is the trafficking interface of the ER. Some specialised cells clearly demonstrate how both structures are related to their function - adrenal cells, that produce steroids, and muscle cells with high metabolism of Ca\(^{2+}\) ions for muscle contraction have proliferated SER; secretory gland cells on the other hand, which produce large amounts of proteins, have proliferated RER (Yamasaki et al. 1997; Terasaki et al. 2013). Interestingly, the RER in salivary secretory gland cells has tightly stacked ER sheets connected with twisted membrane structures allowing them to maximise the ER volume (Terasaki et al. 2013; Guven et al. 2014).

Studies by Vale and Hotani in 1988 and Voeltz et al. in 2006 revealed that shaping of ER is an active process facilitated by cytoskeleton and shaping proteins (Vale et al. 1988; Voeltz et al. 2006). The cytoskeleton shapes ER tubules in three ways: sliding, where a tubule is pulled out of an existing tubule and expanded along a microtubule; directly attached to a growing microtubule through the tip attachment complex (TAC); or ring rearrangements, where rings in the ER are dragged around with microtubules (illustrated on Figure 1.3) (reviewed in Westrate et al. 2015). In the sliding type of dynamics the ER tubule is pulled out of the existing tubule with the help of kinesin-1 and dynactin walking down the microtubule filament (Waterman-Storer et al. 1998) (Figure 1.3a). They preferentially bind to acetylated microtubules with a slightly different structure and high resistance to nocodazole (Friedman et al. 2010). An ER tubule attached to a microtubule through TAC grows and shrinks together with the elongation or degradation of the microtubule (Figure 1.3b). The protein EB1, a member of TAC, binds to the plus end of microtubules and interacts with the integral ER membrane protein STIM1. This mechanism of tubule growth is more specific in regions close to the plasma membrane because of its attachment to the plus end of the microtubule (Grigoriev et al. 2008). In the ring rearrangements mechanism ring structures in
CHAPTER 1. INTRODUCTION

**Figure 1.3**: Four types of ER shaping by cytoskeleton. a) The sliding mechanism, where tubules (green) are pulled out of an existing tubule and expanded along a microtubule (red). b) The tip attachment mechanism. An ER tubule is attached to a microtubule with the tip attachment complex (TAC) at the tip of the microtubule. The images show a TAC protein STIM1 (red), which localises to the tip of the expanding ER tubule (green). c) ER ring rearrangements. A ring between two proximal three-way junctions becomes smaller and smaller until it closes. Rings (green) commonly circumscribe other organelles, such as endosomes (red). d) ER tubules (green) can also expand with a translocating organelle, such as an endosome (red). Reproduced with permission from Weststrate et al. 2015, copyright Annual Reviews.

The tubular ER move around the cell along microtubules with active dynamics (Figure 1.3c). These rings are basically two three-way junctions, where one of them is static and the dynamics of the other junction then determines the size of the ring. Usually it moves closer to the static junction until the ring closes. Many of these rings occur at contact sites between the ER and endosomes or mitochondria. The two organelles get squeezed by the closing ER ring and these are the sites where fission of endosomes and mitochondria occurs (Friedman et al. 2011; Rowland et al. 2014).

A high surface-to-volume ratio of tubular ER is not an intrinsic membrane condition and therefore needs an active mechanism of shaping. The cytoskeleton is important in the creation of the reticular shape of the tubular network, but it does not induce curvature into the membrane. Curvature is induced by integral membrane proteins reticulons and DP1/Yop1p, which create a scaffold by forming extensive homo- and heteromers in an arc like shape (Shibata et al. 2008; Hu et al. 2008). They specifically localise to high-curvature regions of tubules and edges of sheets (Voeltz et al. 2006). Figure 1.4 shows models of some ER shaping proteins. They have characteristic struc-
Reticulons, REEP5, Atlastin and Spastin have one or two short hairpin domains and protrudin has a short hairpin and a full length hairpin domain.

ER sheets are formed differently. They are usually bundled with ribosomes to a much higher extent than tubular ER, which suggests possible stabilisation of sheet structure by ribosomes. Indeed drug treatments that release ribosomes from the ER induce destabilisation of sheets and formation of tubules (Puhka et al. 2007). Sheets have a consistent thickness of 50 nm throughout the ER, which indicates that another mechanism stabilises the two membrane bilayers at this particular distance. The transmembrane protein CLIMP-63 has a luminal coiled-coil domain that dimerises with the coiled-coil domain of CLIMP-63 in the opposite membrane layer, thus setting the distance between them (Schweizer et al. 1993; Klopfenstein et al. 1998). While depletion of CLIMP-63 does not destabilise the sheet it does reduce its thickness from 50 nm to 30 nm (Shibata et al. 2010). Sheet shape is also regulated by reticulons and DP1 which reside in the high-curvature regions at the bends.

The ER network is highly dynamic, continuously expanding and changing. Two ER membrane elements can fuse when they come close to each other, with fusion being mediated by atlastins and Rab GTPases. One molecule of atlastin dimerises with another atlastin on the neighbouring membrane and through the activation of its GTPase domain initiates fusion of both membranes (Orso et al. 2009; Bian et al. 2011; Byrnes et al. 2011). Atlastins localise to three way junctions and
depletion leads to tubules with fewer branches (Hu et al. 2009). The opposite effect is observed after depletion of Rab10 and Rab18, which leads to the proliferation of peripheral sheet ER with loss of tubular ER (English et al. 2013; Gerondopoulos et al. 2014). Lunapark is another ER shaping protein with function similar to atlastins (Chen et al. 2012).

Certain other proteins are also known for ER shaping. Protrudin binds to atlastins and regulates the balance between sheets and tubules (Chang et al. 2013; Hashimoto et al. 2014). Arf-like-6-interacting partner 1 (ARL6IP1) also binds to atlastins and when overexpressed induces extensive proliferation of tubular ER (Yamamoto et al. 2014). YIPF5 interacts with DP1/Yop1 and its disruption causes formation of whorl-like structures in the ER (Dykstra et al. 2010).

Shaping of membranes in the endomembrane system is of high interest because disruptions of shaping mechanisms are linked to neuropathologies, such as hereditary spastic paraplegia, Alzheimer’s disease and Parkinson’s disease. The ER is also the main regulator of Ca\(^{2+}\) metabolism which is disrupted if ER tubulation is lost. This could be fatal for the cell - commonly it ends with apoptosis (Jozsef et al. 2014).

Most resident proteins of the endomembrane system and secreted proteins are synthesised at the ER and then trafficked from there to their final destination. Trafficking starts at ERES that have the characteristic shape of small protrusions, which are nascent COPII trafficking vesicles (Shaywitz et al. 1995; Orci et al. 1991). Transport carriers originating from ERES are trafficked towards the Golgi apparatus.

### 1.1.2 Golgi apparatus

The Golgi apparatus is a distinctive membrane bound organelle that functions as a trafficking hub in the cell. A second important function is modification of cargo with oligosaccharides. Modifications are of critical importance for proper function of proteins and also for their delivery, indeed many trafficking pathways recognise glycosylation profiles of proteins and read this information to help determine their delivery to for example basolateral versus apical membranes (Mellman et al. 1992).

The Golgi is structured as a ribbon of four to seven stacks of disc-shaped cisternae (Ladinsky et al. 1999). These cisternae are not homogenous but remain polarised both in composition and function (Dunphy et al. 1985; Farquhar 1985). Cisternae at the cis- side of the Golgi are involved in trafficking pathways between the ER and Golgi. They also contain enzymes that carry out the first processing steps of the cargo. In medial and trans cisternae cargo gets further processed with other modification enzymes. Finally cargo reaches the trans-Golgi network (TGN) where it is sorted and packaged into clathrin-coated or other vesicles (Bard et al. 2006; De Matteis et al. 2008). Trafficking mechanisms between cisternae remain inconclusive, despite being a topic of much interest, and it is still unclear how compartmentalisation in this organelle is achieved. What is clear though, is that the constituents of each cisterna are actively retained in the correct location.

The ribbon-like organisation of the Golgi is maintained by cytoplasmic and transmembrane proteins surrounding the cisternae, a structure called the Golgi matrix (Slusarewicz et al. 1994). The Golgi matrix forms a scaffold for the assembly of the Golgi and maintains its physical struc-
ture. Two main components of the Golgi matrix are involved in the maintenance of its structure: coiled-coil proteins termed golgins and Golgi reassembly stacking proteins (GRASPs) (Seemann et al. 2000b; Seemann et al. 2000a; Barr et al. 1997; Shorter et al. 1999). GRASP55 and GRASP65 are two key members of the GRASP family and they localise distinctly to medial and cis-Golgi elements, respectively (Shorter et al. 1999; Barr et al. 1997). They both interact with golgins: GRASP65 with GM130 and GRASP55 with golgin-45. Disruption of the GRASP55 - golgin-45 complex results in dispersal of the Golgi and inhibition of trafficking (Short et al. 2001). Interaction between GRASP65 and GM130 is needed for correct localisation of both proteins to the Golgi as well as for trafficking, likely because GRASP65 interacts with cargo receptors of p24 family (Barr et al. 1998; Barr et al. 2001). Golgins and GRASPs are also involved in vesicle tethering. Golgins function as ‘tentacles’ that capture vesicles approaching the Golgi and assist in guiding them through the matrix (Sinka et al. 2008; Hayes et al. 2009).

Another important factor in positioning and structuring the Golgi is the cytoskeleton. Microtubules and actin are involved in structuring the Golgi in the opposing ways: the disruption of microtubules with nocodazole disperses the Golgi into mini-stacks, while depolymerisation of actin (with either cytochalasin, latrunculin or other similar compounds) makes the Golgi small and compact (Ho et al. 1989; Campli et al. 1999; Valderrama et al. 1998). This indicates that the two cytoskeletal systems are kept in balance to maintain normal Golgi structure. The molecular motors cytoplasmic dynein 1 and 2 (CD1 and CD2) link microtubules to the Golgi as well as their regulators dynactin and NUDE-like (NDEL) (Karki et al. 1999; Liang et al. 2004a). Depletion of several Golgi localising actin binding partners, such as Arp2/3, WHAMM and cortactin, mimic the effect of actin depolymerisation (Campellone et al. 2008; Kirkbride et al. 2012). Microtubules are particularly important for positioning the Golgi stacks near the centrosome in mammalian cells. Such proximity of the stacks enables the formation of tubular connections between them, which enhances the efficiency of glycosylation (Puthenveedu et al. 2006). Such positioning is even more important for post-Golgi trafficking - cargo is delivered to the cis-Golgi along microtubules and then trafficked outward from the TGN to the plasma membrane also along microtubules. Placing all Golgi stacks in the vicinity of the centrosome makes transport along microtubules more efficient (Glick et al. 2009; Efimov et al. 2007).

Individual Golgi cisternae effectively operate as distinct compartments with specific set of reactions taking place in each. Cargo has to be transported between these compartments and there are at least three different theories explaining this process: an anterograde vesicular transport model, a cisternal maturation model and a rapid partitioning model. In the anterograde vesicular transport model each cisterna contains a constant set of enzymes. Cargo is trafficked from cis-Golgi to TGN in an anterograde way using vesicular intermediates (Orci et al. 2000; Ostermann et al. 1993). In the cisternal maturation model each cisternal compartment dynamically changes and 'matures' from cis- to trans-Golgi. Cargo remains in the same compartment while enzymes are transported in a retrograde manner, also via vesicular intermediates (Glick et al. 1997; Glick et al. 2011; Glick et al. 2009). This model is supported by strong evidence - intra-Golgi transport vesicles indeed contain higher levels of glycosylation enzymes. Furthermore, some large molecules that pass through the Golgi, such as pro-collagen, are never seen in vesicles but only in cisternae, and still effectively trafficked outwards from the cell (Hoe et al. 1995; Bonfanti et al. 1998). However,
this model assumes that all cargo is trafficked at the same rate, which has been questioned in a recent study in which different transport rates of various cargoes were observed. A new model was proposed instead, whereby Golgi cisternae are interconnected thus forming a single compartment. In this model the transport is regulated by fusion and fission events and cisternal identity is regulated by a lipid gradient (Patterson et al. 2008). This model currently best fits experimental evidence, however these models are not mutually exclusive (Glick et al. 2011).

Endomembrane proteins, after being processed in the Golgi, are sorted and further dispatched at the TGN. The destination of a cargo protein is determined by complex criteria, either by sequence determinants, glycosylation profiles, three-dimensional determinants or a combination of all. Most of these rules are not explained yet, but we do understand some early targeting mechanisms, such as protein synthesis/targeting to the ER and trafficking between the ER and Golgi. In the following section, the early stages of targeting to the endomembrane system will be discussed.

1.1.3 Protein targeting to the endomembrane system

Protein targeting is a fundamental mechanism that delivers proteins to their correct locations inside or outside of the cell. Such processes can insert a protein into a membrane bilayer (transmembrane proteins) or to the lumen of the organelle. They can also direct proteins to be secreted out of the cell or attached (anchored) to the membrane from the cytosolic site. Approximately 30% of proteins in the human proteome are targeted to the endomembrane system based on simple criteria; namely that soluble proteins that either reside in the lumen or are secreted out of the cell have a cleavable signal peptide at their N-terminus; and that transmembrane proteins, that are incorporated into the membrane bilayer, are translocated based on the presence of TMDs. Proteins are translocated through a translocation channel in the ER largely made of a protein termed Sec61. Translocating mechanisms can be separated into two groups: cotranslational, where a protein is translocated while it is synthesised and posttranslational, where a protein is translocated after it is synthesised.

The majority of endomembrane proteins are translocated by a signal recognition particle (SRP) mediated mechanism (Gilmore et al. 1982b; Gilmore et al. 1982a; Walter et al. 1980; Walter et al. 1983; Walter et al. 1984), involving a signal peptide, up to 30 residues long, at the N-terminus of the protein to be translocated. When the N-terminal region of a protein emerges from of a ribosome the signal peptide gets recognised by SRP. Binding of SRP to the ribosome-nascent protein chain complex (RNC) blocks protein synthesis and then, through the interaction with the SRP receptor located at ER membranes, the RNC is brought into close vicinity of the Sec61 translocation channel. The synthesis of a protein then continues and the peptide chain is directly synthesised into the lumen of the ER (Rapoport 2007; Akopian et al. 2013). In case of transmembrane proteins, some regions of a protein do not enter the translocation channel and stay in the cytosol, thus forming the cytosolic regions of a protein (Mothes et al. 1997). Co-translational targeting mechanisms minimise protein aggregation, misfolding and prevent exposure of certain regions to cytosolic proteins thus preventing unwanted interactions and activity.
While most endomembrane proteins follow this mechanism many others are localised after the synthesis completes. In this case the Sec61 translocation channel partners with the tetrameric membrane protein complex Sec62/Sec63 and the luminal chaperone BiP, an ATPase from Hsp70 family (Meyer et al. 2000; Deshaies et al. 1991; Panzner et al. 1995). Translocation starts when a synthesised protein, partially folded and with attached cytosolic chaperones, binds to the translocation complex. At this stage all cytosolic chaperones are removed and the N-terminal signal peptide intercalates in the channel (Plath et al. 2000; Shaw et al. 1988). The peptide chain could move in or out of the channel, but at this stage BiP binds to it and pulls it through the channel into the lumen. BiP then interacts with a 'J-domain' on Sec63, which exchanges ATP for ADP resulting in the closure of the peptide binding pocket on BiP. The peptide chain is therefore anchored in the lumen by BiP (Matlack et al. 1999). After the polypeptide moves forward another BiP molecule binds to it and the whole process repeats. In both co- and posttranslational translocation mechanisms the N-terminal sequence is cleaved off by signal peptidase and further degraded by signal peptide peptidase (Weihofen et al. 2002).

Transmembrane proteins can also start with an N-terminal signal peptide, but their TMDs are sufficient for the initiation of translocation. Translocation starts when the hydrophobic domain, either the signal peptide or the first TMD, emerges from the ribosome. The complex then binds Sec61 as described above. The TMD then exits the channel through its lateral gates - four TMDs that move away from each other forming an opening (Van den Berg et al. 2004). The highly hydrophobic TMD residing in the hydrophilic translocation channel gets access to the membrane and thus transitions into the lipophilic phase (Rapoport et al. 2004). The direction of the first TMD also determines the topology of the protein although there are some exceptions (Wessels et al. 1988; Rapoport et al. 2004). Transmembrane proteins with a signal peptide usually have their N-terminus inside the lumen, while the orientation of proteins without a signal peptide depends on the N-terminal sequence - highly charged residues tend to reside on the cytosolic side thus forming a cytosolic N-terminus (Rapoport et al. 2004).

A particular type of transmembrane protein with a long N-terminal cytosolic domain and a C-terminal hydrophobic region, also termed tail-anchor (TA) proteins, are translocated using a different mechanism. Because their translocation is initiated only after the C-terminal hydrophobic region has been translated, these proteins cannot exploit the typical SRP-dependent translocation mechanism (Yabal et al. 2003). These proteins therefore exploit an alternative mechanism present in the cell, which is mediated through the GET (Golgi to ER trafficking) complex (Schuldiner et al. 2008). In yeast, the soluble protein Get3 probes newly synthesised proteins for the presence of the C-terminal hydrophobic region and forms a complex with them. The Get3-TA complex then binds with complex of ER transmembrane proteins Get1-Get2, which results in insertion of the TA protein into the membrane (Schuldiner et al. 2008).

Some soluble proteins, such as Rab GTPases, attach to the membrane with a lipid tail that is added posttranslationally (reviewed in Nadolski et al. 2007). Cytoplasmic and luminal proteins can be modified with a lipid tail: modifications of cytoplasmic proteins are usually N-myristoylation, S-palmitoylation and prenylation (Resh 1999; Zhang et al. 1996; Smotrys et al. 2004; Greaves et al. 2007); and the best characterised lipid modification of luminal proteins is the attachment of glycosylphosphatidylinositol (GPI) anchors (Orlean et al. 2007). Proteins with a GPI anchor
are translocated to the plasma membrane and are anchored to the extracellular side of it. Some important regulators of membrane trafficking, such as Rab proteins, are soluble cytosolic proteins which attach to membranes with a diprenylated lipid anchor. Modification enzymes, for example geranylgeranyl transferase (GGT) that modifies Rab proteins, recognises a C-terminal Caa. motif, a cysteine (C) followed by two aliphatic amino acids (a) and . which determines whether a protein is farnesylated or geranylgeranylated. Prenylation occurs also on variations of the above motif, such as C-terminal CC or C.C (Zhang et al. 1996).

Attachment of cytosolic membrane trafficking regulators to the membranes with lipid anchors is essential, not only for their proper localisation but also for their regulatory role. In contrast to transmembrane proteins, attachment via lipid tails is transient - such proteins can be extracted from the membrane by proteins such as guanosine diphosphate (GDP) dissociation inhibitors (GDIs) that neatly pack their lipid anchor thus preventing their reattachment to the membrane and keeping them in a soluble state. In the following section, how the transition from membrane-bound to soluble state also determines their activity and makes them a molecular switch will be discussed.

1.2 Membrane trafficking

Exchange of cargo between compartments is facilitated by small membrane bound vesicles that emerge from a donor membrane and deliver the cargo to a acceptor membrane. Several trafficking pathways are known in the cell, each characterised by direction, relevant compartments and underlying machinery. Generally speaking, cargo can travel in two directions. Proteins synthesised at the ER are trafficked in an anterograde or secretory manner from the ER to the plasma membrane. In the reverse direction, material from extracellular environment is taken up by the process of endocytosis and trafficked towards the ER in a retrograde manner. Pathways are also organelle specific - ER/Golgi trafficking is mechanistically different from endocytosis, and even within organelle specific pathways there are variants, depending on the mechanism and machinery that is involved.

1.2.1 Machinery

The formation of a vesicle, its transport and fusion (Figure 1.5) can be summarised in 7 steps: (1) The trafficking process is initiated by the binding of a cargo to a cargo receptor on the membrane. (2) Coat proteins assemble on the membrane reshaping the membrane into a bud. (3) The bud is excised and (4) the newly formed vesicle gets uncoated and (5) translocated along cytoskeletal elements. At the acceptor membrane the vesicle (6) tethers with the tethering proteins and (7) docks with the membrane. In the final step it is (8) fused with the acceptor membrane releasing the cargo into the lumen. In each step specific machinery is recruited: for example coat proteins for membrane budding, tethers and transmembrane Soluble N-ethylmaleimide-sensitive factor Attachment Protein Receptors (SNAREs) for fusion. The whole process is overseen by Rab GT-Pases, lipid anchor proteins cycling between the membrane and the cytosol with the mission of keeping the membrane trafficking system 'under control'.
Small GTPase Membrane-distal coat protein

Membrane-proximal coat protein

Donor compartment

Soluble cargo

Transmembrane cargo

Regulator

Budding (2)

Coat assembly (1)

Scission (3)

Uncoating (4)

Translocation (5)

Tethering (6)

SNARE assembly (7)

Fusion (8)

Tethering factor

v-SNARE

Cytoskeletal element

Small GTPase

Membrane-distal coat protein

Membrane-proximal coat protein

Small GTPase

Cytosol

Figure 1.5: Seven steps of vesicle trafficking. 1) initiation of vesicle trafficking and coat assembly; 2) vesicle budding; 3) vesicle pinches off the donor membrane; 4) uncoating and 5) association with cytoskeleton guides; 6) tethering with the acceptor membrane; 7) docking of the vesicle - formation of the SNARE complex; 8) fusion with the acceptor membrane. Figure reproduced from Bonifacino et al. 2004.

1.2.1.1 Rab GTPases

Rab GTPases are a large family of proteins that function as molecular switches, switching between an activated ‘ON’ guanosine triphosphate (GTP) bound state and an inactivated ‘OFF’ GDP bound state. There are 60 Rab GTPases in the human proteome, localising to distinct membrane compartments (Figure 1.6) (Stenmark 2009; Chavrier et al. 1990; Zerial et al. 2001). Their regulatory role depends on cycling between soluble and membrane attached states - for example when anchored to the membrane and activated they attract trafficking components, such as SNARE proteins, that mediate vesicle fusion (discussed below).

After a Rab protein is synthesised it is captured by Rab escort protein (REP). REP presents it to the geranylgeranyl transferase (GGT), which attaches two geranylgeranyl tails to two cysteine residues arranged as ...CC, C.C or CC.. motifs at the C-terminus (Seabra 1996; Shen et al. 1996). The modified Rab remains in complex with REP, which delivers it to its donor membrane and releases it for the first activation (Figure 1.7) (Shen et al. 1996). The Rab is activated by exchanging GDP for GTP resulting in a major conformational change of the switch region of the protein (Pfeffer 2005). The exchange reaction is catalysed by a guanine nucleotide exchange factor (GEF), which recognises a specific region in the Rab switch region and facilitates release of GDP (Burstein et al. 1992; Burton et al. 1994; Delprato et al. 2004). A molecule of GTP promptly binds into the switch region due to the high concentration of GTP in the cytosol. The activated Rab attracts effectors, mostly proteins involved in membrane trafficking such as coat proteins, tethers, sorting proteins, molecular motors and adapters. The Rab protein ‘deactivates’ through GTPase activity. Hydrolysis of GTP is an intrinsic process catalysed by a GTPase activating protein (GAP) (Trahey et al. 1987; Gibbs et al. 1988; Becker et al. 1991; Barr et al. 2010). Although the inactive GDP bound Rab has a tendency to exchange GDP for GTP due to the high concentration of GTP in the cytosol, this is prevented by the Rab binding to a GDP dissociation inhibitor (GDI) protein (Sasaki et al. 1990;
Matsui et al. 1990; Garrett et al. 1994; Shapiro et al. 1995). GDI also extracts the Rab from the membrane, chaperones its lipid tail while in the cytosol, and delivers it back to the membrane (Ullrich et al. 1994; Ullrich et al. 1993; Soldati et al. 1994; Rak et al. 2003). When a Rab protein is delivered to the membrane the GDI is first detached from it, which allows the Rab to enter the activation cycle. Detachment of a GDI from the Rab protein is mediated by a GDI displacement factor (GDF) (Sivars et al. 2003).

The first link between Rab GTPases and vesicle trafficking was established in seminal studies, starting with a genetic screen identifying yeast mutants with impaired protein secretion (Novick et al. 1980). Subsequent studies linked Rab8 to post-Golgi trafficking and Rab1 to the early se-
cretory pathway (Goud et al. 1988; Salminen et al. 1987; Segev et al. 1988; Bacon et al. 1989). Later it was revealed that Rab1 and Rab5 directly promote membrane fusion by regulating SNAREs through the recruitment of tethering factors to the membranes (Gorvel et al. 1991; Segev 1991; Lian et al. 1994; Søgaard et al. 1994; Sapperstein et al. 1996; Cao et al. 1998; Christoforidis et al. 1999a; McBride et al. 1999; Allan et al. 2000; Shorter et al. 2002). After 20 years of intensive studies it is now clear that Rab GTPases regulate all steps of membrane trafficking events. They recruit cargo adapters that select cargo for trafficking. An example of this is the late endosomal Rab9 which is involved in the recycling of mannose-6-phosphate receptor (M6PR) from late endosomes to the TGN. Rab9 recruits a cargo adapter TIP47 that recognises the cytosolic tail of M6PR and recruits it for trafficking. Recruitment of TIP47 by Rab9 also increases its affinity for M6PR thus further enhancing its role in cargo selection (Carroll et al. 2001). Several other Rab GTPases can directly or indirectly function in cargo selection: Rab5 is involved in the formation of clathrin-coated pits and has a role in clathrin dependent endocytosis of the transferrin receptor (McLauchlan et al. 1998); Rab11 directly interacts with the Ca$^{2+}$ channels TRPV5 and TRPV6 and targets them to the plasma membrane (Graaf et al. 2006); Rab11 also regulates the recycling of β2- and β1-adrenergic receptor through direct interaction (Parent et al. 2009; Gardner et al. 2011). After the vesicle is formed Rab GTPases can initiate the uncoating process. For example, the interaction between the clathrin-coat cargo adapter AP2 and the plasma membrane is stabilised by phosphoinositide phosphatidylinositol (PI)-4,5-bisphosphate (PI-4,5-P$_2$). Rab5 on endosomes recruits PI-3-kinase (PI3K) which induces PI-4,5-P$_2$ turnover initiating the uncoating of a vesicle (Christoforidis et al. 1999b; Shin et al. 2005). The transport of vesicles is mediated by highly specific interactions between vesicles and cellular motors. Rab27a recruits the adapter melanophilin to melanosomes which interacts with the actin motor myosin Va (Wu et al. 2002). Myosin Va (Lindsay et al. 2013; Sun et al. 2014), myosin Vb (Schafer et al. 2014; Thoeni et al. 2014; Roland et al. 2011) and myosin Vc (Bultema et al. 2014; Xu et al. 2009) are common effectors of Rab GTPases. Trafficking along microtubules requires plus-end (towards the cell periphery) motors kinesins, and opposite minus-end motors dyneins (reviewed in Horgan et al. 2011). Kinesins

---

**Figure 1.7:** Rab activation cycle. Rab-GDP protein is stabilised and protected from activation by a GDI protein. In transition to activated state it dissociates from the GDI and exchanges GDP for GTP with the help of a GEF. Rab-GTP is anchored into the membrane by its farnesylated lipid tail. Then GTPase is activated by a GAP, Rab hydrolyses GTP to GDP and is extracted from the membrane and stabilised in its GDP form by a GDI. Freshly synthesised Rab protein first undergoes lipid transformation by the help of a REP which presents an untransformed Rab to a GGT.
can be direct effectors of Rab GTPases, such as rabkinesin-6 recruited by Rab6 in murine cells (Echard et al. 1998). Human rabkinesin-6 localises to the nucleus and plays a crucial role in cytokinesis (Fontijn et al. 2001). An example of indirect interaction is KIF16B, involved in endocytic recycling, recruited to endosomes through the interaction with PI-3-phosphate which is recruited to the endosomal membrane by Rab5 (Hoepfner et al. 2005). Dyneins interact indirectly, for example Rab11-FIP3 interacts with cytoplasmic dynein-1 in the transport from sorting to recycling endosomes (Horgan et al. 2010). During vesicle fusion Rab GTPases function at different levels. First they interact with tethers and regulate vesicle recognition and quality control. This process was firstly described by Cao and Barlowe while investigating the tethering of COPII vesicles in yeast. Ypt1, a yeast homolog of Rab1, recruits the cytosolic protein Uso1p to the Golgi membrane, which tethers the vesicle to the membrane (Cao et al. 1998; Cao et al. 2000). Tethers can recruit SNARE regulators, for example recruitment of VPS45, a SNARE regulator, by rabenosyn-5 (Nielsen et al. 2000). The involvement of Rab GTPases in such a variety of processes is reflected in their localisation map (Figure 1.6) - they localise to all compartments of the endomembrane system. This phenomenon was first described by Chavrier et al. and further developed into an interesting proposal that Rab GTPases define membrane identity (Bourne 1988; Pfeffer 2001; Zerial et al. 2001). This is indeed the case in the early endosomal pathway where Rab5 and Rab7 proteins identify early and late endosomes (Rink et al. 2005), but is less defined across the other compartments.

The mechanism by which Rab GTPases recognise their target membranes is not completely understood, but there are at least two models attempting to explain the process. GDF activity (in the form of the Prenylated Rab acceptor family 1 (PRAF1) protein) was discovered by Sivars and colleagues, when they observed a change in the distribution of Rab9 between a soluble and membrane fraction on modulation of PRAF1 levels (Sivars et al. 2003). This model proposes that individual or groups of Rab proteins would interact with a specific GDF, which would probably be a transmembrane protein specifically localised to the relevant compartment. This model has several downfalls. The only known protein so far confirmed to have GDF activity is PRAF1 (Janoueix-Lerosey et al. 1995; Sivars et al. 2003). Its homologs, PRAF2 and PRAF3, have been proposed to be GDFs, however there is no evidence that they biochemically possess GDF activity. In addition, the depletion of any individual PRAF protein shows only minimal or, in some cases, no effect on Rab localisation (Bucci et al. 1999). The proposed diversity of possible GDF proteins is also too small to cover all the compartments to which Rabs localise.

Recently two independent studies indicated that RabGEF is sufficient for localisation of Rab proteins to the membrane. In these experiments RabGEFs Rabex-5, DrmA, Rabin8 and BLOC-3 of Rab5A, Rab1A, Rab8A and Rab32/38 proteins, respectively, were fused to mitochondrial outer membrane targeting sequences which changed the localisation of RabGEF and relevant Rab protein to the mitochondria (Gerondopoulos et al. 2012; Blümer et al. 2013). Nucleotide exchange is essential for localisation as a mutated GEF, unable to catalyse the GDP/GTP exchange, showed lower levels of localisation change (Blümer et al. 2013). This model could therefore explain the mechanism of Rab localisation, but it raises new questions about membrane recruitment and recognition of GEFs. The first proposed explanation is vesicle-coat dependent GEF targeting. Several examples are known: recruitment of the Rab1 GEF Transport Protein Particle-I (TRAPP-I) by COPII (Cai et al. 2007b) and recruitment of the Rab35 GEF DENND1A or the Rab5 GEF RME-6, both
CHAPTER 1. INTRODUCTION

by the clathrin coat (Allaire et al. 2010; Yoshimura et al. 2010; Sato et al. 2005b; Semerdjieva et al. 2008). A second explanation is Rab coupling. This happens when an activated Rab directly or indirectly recruits another Rab to the membrane. The most basic mechanism is the recruitment of another Rab protein by coupling through an effector of an activated Rab. An example is rabenosyn-5, which has binding sites for both Rab4 and Rab5. Overexpression of rabenosyn-5 enhances the overlap between Rab4 and Rab5 containing domains, indicating a potential role in connecting both domains (Renzis et al. 2002). More advanced examples of Rab coupling are Rab cascades, which can be GDP/GTP exchange, GTPase activation or mixed cascades. A typical example of a GDP/GTP exchange cascade is found with the Rab5 GEF Rabex5-rabaptin complex which is a Rab5 GEF and an effector for Rab4 and Rab5 (Horiuchi et al. 1997). This case is also a demonstration of signal amplification in a Rab network; Rab5 recruits its own GEF thus activating more Rab5 molecules in its vicinity. Many other activating cascades are known, such as the recently discovered Ric1-Rgp1 complex, a Rab6a GEF and effector of Rab33b (Pusapati et al. 2012) and the GEF of Rab32/38, which is an effector of Rab6 (Kloer et al. 2010; Gerondopoulos et al. 2012). Rab effectors can also be GAPs, which inactivate the previously activated Rab. In yeast, Gyp1 is an effector of Ypt32 (Rab32) and acts as a GAP that deactivates Ypt1 (Rab1) (Rivera-Molina et al. 2009). Similarly in mammalian cells, the Rab9 effectors RUTBC1 and RUTBC2 are also GAPs for Rab32 and Rab36, respectively (Nottingham et al. 2011; Nottingham et al. 2012). A more complex form of Rab coupling is Rab conversion - a network loop, where the first Rab, when activated, recruits a GEF for the second Rab, which then deactivates the first Rab. This is in its origins a negative feedback loop, but it turns out that such a mechanism forms a bistable switch. An example is Rab5 recruitment of the Homotypic fusion and protein sorting (HOPS)-tethering complex (discussed below), the VPS39 subunit of which is a GEF for Rab7 (Rink et al. 2005). SAND-1/Mon1 then stops the activation of Rab5 by displacing Rabex5 from the membrane and blocking the recruitment of Rab7 by interaction with the HOPS complex (Conte-Zerial et al. 2008; Poteryaev et al. 2010). A bistable switch is formed when activation of the first Rab recruits the GEF for the second Rab, which then recruits a GAP for the first Rab. This particular example has only been recently proposed, although it is a plausible model, especially with highly interlinked Rab networks, where GEFs and GAPs would compete with each other which would result in unnecessary work (Barr 2013).

The above models cannot however explain how the first Rab in a cascade finds its membrane. One plausible explanation includes specific PIs, which are produced by phosphatases and kinases on the membrane. This suggests that Rab proteins are able to control the local constitution of PIs, which is supported by an example where multiple Rab proteins are able to recruit OCRL, a PI-5-phosphatase (Fukuda et al. 2008). Even more interesting is an example involving Rab5, which recruits PI3K to produce PI-3-phosphate on the membranes of endosomes (Christoforidis et al. 1999b). PI-3-phosphate together with Rab5 then recruits its effectors EEA1, rabenosyn5 and rabankyrin5, which all have a FYVE domain that binds PI-3-phosphate (Simonsen et al. 1998; Nielsen et al. 2000; Schnatwinkel et al. 2004). Rab5 in particular also recruits other PI kinases, such as PI4K and PI5K (Shin et al. 2005).

A second issue that remains open is the role of GDF in the Rab activation cycle. It seems that at least some Rab proteins can localise and activate without GDF, so its function may be limited to only a few Rab proteins. However, PRAF and Ypt-interacting proteins family (YIPF) proteins,
potential GDF proteins, do localise to the endomembrane system and seem to have a role in membrane trafficking. Their potential roles are discussed in the following section.

1.2.1.2 Transmembrane PRAF and YIPF protein families

A typical transmembrane protein has one or more hydrophobic domains, usually in the shape of \( \alpha \)-helices, which span through the phospholipid bilayer. Hydrophobic domains of a protein are connected with soluble loops, pointing either into the cytoplasm or lumen of the membrane-bound compartment (or extracellular environment for plasma membrane proteins). Transmembrane proteins with \( \alpha \) TMDs account for approximately 27% of the human proteome, although this is based on predictions, because solving the three-dimensional (3D) structure of a transmembrane protein is difficult. Their hydrophobic nature makes them hard to express in large quantities and they are difficult to extract, solubilise, purify and crystallise (Carpenter et al. 2008). The first attempts to predict hydrophobic regions in proteins were based on the hydrophobicity of amino acids. Kyte and Doolitle in 1982 proposed a scale of amino acid hydrophobicity values. For example, hydrophobic isoleucine has a value of 4.5, glycine has neutral value of -0.8 and charged arginine has the lowest value of -4.5 (Kyte et al. 1982). On a hydrophobicity plot (Figure 1.8a), amino acids are plotted from N to C terminus with the corresponding hydrophobicity values on the vertical axis. Regions above 0 are considered hydrophobic (Kyte et al. 1982). In 30 years the hydrophobicity scale has been improved and modern predictors rather utilise machine learning methods. The best predictors today are TMHMM and Phobius (Krogh et al. 2001; Käll et al. 2004). Both of these are based on hidden Markov models and distinguish transmembrane regions, signal peptides and even whether soluble loops are cytoplasmic or luminal (Figures 1.8a and 1.8b).

Several families of transmembrane proteins are known for their function in membrane trafficking. Particularly important are SNARE proteins that provide the minimal needed machinery for vesicle fusion (SNARE proteins are discussed below). Here, a group of transmembrane proteins, PRAF and YIPF, which are proposed GDF candidates are discussed.

1.2.1.2.1 Prenylated Rab acceptors

Proteins of the Prenylated Rab acceptor family can be identified by three features. The main criterion is a PRA1 domain, which has been noted to be conserved from yeast to human. The second feature is their characteristic arrangement of four TMDs. Lastly, they interact with Rab GTPases. The prototype of the family, PRAF1 (or PRA1), was discovered as an interactor of Rab6 in mouse (Janoueix-Lerosey et al. 1995). It has a yeast ortholog, Yip3p, and 2 homologs in animals, PRAF2 and PRAF3. A greater variety of PRA1 proteins can be found in plants, for example Arabidopsis has 19 PRA1 proteins (Alvim Kamei et al. 2008). PRAF1 and PRAF2 are ubiquitously expressed in human tissues (Martincic et al. 1997; Bucci et al. 1999; Abdul-Ghani et al. 2001; Schweneker et al. 2005). PRAF3, in contrast, is hardly detectable in human tissues, but its expression can be induced by retinoic acid (Lin et al. 2001b). A recent study showed that PRAF3 is expressed in developing tissues and declines after birth (Maier et al. 2009).
Figure 1.8: Membrane topology of PRAF and YIPF proteins. a) Hydrophobicity plot of PRAF and YIPF proteins using the Kyte and Doolittle hydrophobicity scale (Kyte et al. 1982). Scores above 0 represent high hydrophobicity. Two hydrophobic regions are clearly seen in PRAF proteins and a large hydrophobic region, corresponding to 5 TMDs, in YIPF proteins. b) TMHMM prediction of TMDs (Krogh et al. 2001). The red colour represents TMDs, blue and purple show inside and outside regions, respectively. c) PRAF proteins have 4 transmembrane domains (TMDs) and YIPF proteins have 5 TMDs.

1.2.1.2.1 Structure

The most striking feature of PRAF1 is its TMDs, which resemble those of ER shaping proteins, particularly reticulons (Figure 1.8c) (Voeltz et al. 2006; Shibata et al. 2009). Additionally, the first short hairpin-like domain has a proline between the first and second TMD, which forms a hinge bending the α-helix (Cordes et al. 2002). Unlike reticulons however, PRAF1 is primarily a Golgi protein (Abdul-Ghani et al. 2001; Gougeon et al. 2002; Liang et al. 2004b; Alvim Kamei et al. 2008; Bhagatji et al. 2010; Geng et al. 2005). There is evidence for a role in Golgi structure - mutation of PRAF1 residues N70T and H166A induces Golgi fragmentation and S76A induces Golgi condensation (Gougeon et al. 2002). However, the observed fragmentation phenotypes are more likely a consequence of disrupted interactions, as both Rab3a and VAMP2 bind to the mutated region (Gougeon et al. 2002). Alternatively PRAF1 could shape the ER even when present in low
quantities, as some localisation to the ER has been reported (Abdul-Ghani et al. 2001; Liang et al. 2004b; Alvim Kamei et al. 2008; Geng et al. 2005).

Besides the Golgi and ER, PRAF1 localises to punctate structures that co-localise with the late-endosomal Rab7 (Bhagatji et al. 2010). In Arabidopsis 13 (out of 19) PRA1 proteins localised to punctate structures that belong to the endosomal compartment, but have also been associated with the ER. The other 6 were localised to the ER only (Alvim Kamei et al. 2008). The localisation of PRAF1 could depend on the ability of the protein to form oligomers. The mutation V185A in mouse PRAF1 prevents the oligomerisation of the protein and it re-localises to the ER (Liang et al. 2004b). It is possible that the protein forms multiple oligomeric complexes that localise to different membranes. Evidence of homo- and heterodimers has been also reported in Arabidopsis (Alvim Kamei et al. 2008).

PRAF2 and PRAF3, human homologs of PRAF1, also have their TMDs shaped into hairpin-like structures. Two 20 residue long TMDs are positioned one after another, forming a 40 residue long hydrophobic region. Two pairs of TMDs are linked by a loop of 15–20 residues. The role of transmembrane regions in organelle shaping has not been reported, but PRAF3 indirectly induces the reorganisation of ERES. Overexpressed PRAF3 reduced the number of ERES, which was rescued by the addition of the cytosol saturated with Rab1a. Overexpressed PRAF3 probably binds Rab1a which prevents its access to the ERES and leads to the blockage of COPII transport (Maier et al. 2009).

Both PRAF2 and PRAF3 are seemingly ER proteins (Schweneker et al. 2005; Abdul-Ghani et al. 2001; Liu et al. 2008b). In gliomas and neuroblastomas PRAF2 localises to endosomes and the ER, and in the brain to synaptic vesicles (Borsics et al. 2010; Geerts et al. 2007; Koomoa et al. 2008). A shorter isoform of PRAF2, 16 kDa long, was found in the nuclear envelope fraction (Borsics et al. 2010). Both proteins form homo- and heterodimers, but whether localisation depends on the oligomerisation state is not known. Interestingly, dimeric PRAF2 resides in the cytosol, while the monomeric conformation predominantly eluted in the membrane fraction in biochemical experiments (Borsics et al. 2010). There are examples of proteins being both transmembrane and cytosolic - bacterial pore-forming toxins transition between a soluble monomeric and transmembrane multimeric form (Yamashita et al. 2014); examples of soluble multimers are also known (Slovic et al. 2005).

1.2.1.2.1.2 Interacting partners

Little is known about the membrane shaping ability of the transmembrane regions in PRAF proteins, but through interactions they might be indirectly involved with the reticulon ER shaping proteins. In yeast PRAF1 interacts with reticulon 1 and YIPF5, which are both involved in the shaping of ER (Geng et al. 2005; Dykstra et al. 2010; Voeltz et al. 2006). Similarly, PRAF3 interacts with reticulon 2B in human, both in vitro and in vivo (Liu et al. 2008b).

PRAF1 was initially discovered as an interactor of Rab6, a Golgi protein involved in retrograde trafficking (Janoueix-Lerosey et al. 1995; Girod et al. 1999; Goud et al. 1990). It also interacts with Rab1, an early secretory pathway Rab. In addition, it binds to endosomal Rab4a, Rab4b, Rab5c, Rab7, Rab17, Rab22 (Bucci et al. 1999) and Rab3a, present in synaptic vesicles (Martincic
et al. 1997). The described interactions are consistent with the localisation phenotypes of PRAF1-Golgi, endosomes and ER. Why is one single protein interacting with such a variety of Rab proteins on different organelles? This could be related to its GDF function as was proposed by Sivars and colleagues (Sivars et al. 2003). PRAF1 could help with the dissociation of GDI from the Rab protein (Dirac-Svejstrup et al. 1997). Indeed, PRAF1 does co-precipitate with GDI (Hutt et al. 2000) and a biochemical in vitro assay showed that PRAF1 facilitates dissociation of Rab9, Rab7 and Rab5 but not Rab1 or Rab2 from GDI. Membrane attachment of Rab9 was also reduced in vivo, while the attachment of Rab1 did not change (Sivars et al. 2003). The GDF activity of PRAF1 seems to be specific for endosomal Rabs, which questions the role of PRAF1 in the early secretory pathway.

Another important endosomal protein interacts with PRAF1. Vesicle associated membrane protein 2 or VAMP2 (or Synaptobrevin 2) is a SNARE protein that facilitates fusion of vesicles with the acceptor membranes in the post-Golgi compartment (Jahn et al. 2006). VAMP2 interacts with conserved N- and C-terminal cytoplasmic loops of PRAF1 and the binding is competitive with Rab3, so they both bind to the same region (Martincic et al. 1997; Hutt et al. 2000).

VAMP2 and Rab3 are both involved in synaptic vesicle trafficking. The synaptic vesicle protein α-synuclein also binds to PRAF1. They are both recruited to synaptic exocytic vesicles and overexpression of both resulted in enlarged synaptic vesicles (Lee et al. 2011a). Another study associated PRAF1 with NFκB signalling through the interaction with LMP1, an inducer of NFκB pathway (Liu et al. 2006).

Other PRAF proteins show fewer interactions with Rabs. PRAF3 has been extensively studied in the context of glutamate transport, but the only known Rab interactor is Rab1 (Maier et al. 2009). This corresponds with the ER localisation of PRAF3 (Schweneker et al. 2005; Abdul-Ghani et al. 2001; Maier et al. 2009). PRAF3 has been mostly studied for its role in cell glutamate uptake. It interacts with the EAAC1 receptor but not its homolog EAAC2 (Lin et al. 2001a). Two other interactors are very interesting: Reticulon 2 (Rtn2) and ARL6IP1 (Liu et al. 2008b; Akiduki et al. 2008). They both shape the ER in a similar fashion (Yang et al. 2007; Yamamoto et al. 2014).

Interestingly, PRAF3 forms heterodimers with PRAF2 (Schweneker et al. 2005). PRAF2 is the least studied protein in the PRAF family. It was discovered as a CCR5 interactor and it can reduce the levels of CCR5 on the plasma membrane (Schweneker et al. 2005), but whether it is involved in CCR5 trafficking is not known. Interesting is the interaction with Bcl-xL and multiple members of Bcl-2 family - overexpression of PRAF2 induces apoptosis by blocking the cell cycle in G1 (Vento et al. 2010; Yco et al. 2013).

1.2.1.2.1.3 Function

The only explained and experimentally supported model for PRAF function is for a role in the Rab activation cycle (Sivars et al. 2003). The authors of the above study proposed that GDFs act as membrane anchors. Although the evidence is relatively strong the study is controversial. Indeed there is also evidence directly countering this model - for example the depletion of PRAF1 or other PRAF proteins did not change Rab localisation (Bucci et al. 1999). Two recent studies showed evidence that Rab GEF could in certain circumstances provide the minimal machinery for localisation of Rab proteins (discussed above). The GDF model also conflicts with the order of
events in the Rab activation cycle. PRAF1 mildly interacts with GDI (Hutt et al. 2000) and it binds only activated and prenylated Rabs (Bucci et al. 1999). On the contrary, Rab is activated only after the detachment of GDI (Sivars et al. 2003; Stenmark 2009; Zerial et al. 2001). In addition, the prenylated tail that attaches the Rab to the membrane is packed inside GDI (Soldati et al. 1994). GDF is supposed to break the complex between GDI and Rab, so it should interact with GDI and the inactive Rab. These models therefore need further experimental work and explanation.

Little is known about the function of PRAF2. It is overexpressed in brain, colon, lung and ovary cancer tissues (Fo et al. 2006). When overexpressed, it also decreases the levels of CCR5 protein on the plasma membrane (Schweneker et al. 2005). The subcellular processes that facilitate these changes are yet to be identified.

PRAF3 is, in contrast to other two PRAFs, hardly detectable in the cell (Lin et al. 2001a). Overexpression of PRAF3 reduces the cell uptake of glutamate without changing EAAC1 levels and plasma membrane localisation (Lin et al. 2001a). The effect was rescued by depletion of PRAF3 - the glutamate uptake increased 4-fold (Lin et al. 2001a; Butchbach et al. 2002). This particular receptor is also important for the uptake of glutathione, an antioxidant, into the cell. Overexpression of PRAF3 induces glutathione related oxidative stress, but this could be rescued by depletion of PRAF3 (Watabe et al. 2007). Its effect on glutamate (and glutathione) trafficking is mediated through modulation of the ER exit of the EAAC1 receptor. Upregulated PRAF3 prevented the exit of EAAC1 from the ER to the Golgi. ER membranes proliferated, probably as a consequence of the general block of ER exit (explained in Ward et al. 2001). Blockage of ER exit has also been demonstrated with Vesicular stomatitis Indiana virus antigen VSV-G. Trafficking of VSV-G in mammalian cells can be blocked by incubation at 15°C, therefore it is used as a tool to study the dynamics of anterograde trafficking. The trafficking of VSV-G to the Golgi was significantly reduced in upregulated PRAF3 (Maier et al. 2009). The observed reduced numbers of ERES were a consequence of the competitive interaction between PRAF3 and Rab1a. High levels of PRAF3 depleted the levels of available Rab1a, which blocked the ‘maturation’ of ERES (Maier et al. 2009). Rab1a is also involved in neurite outgrowth, which is reduced with overexpressed PRAF3 (Maier et al. 2009).

The interaction between PRAF3 and ARL6IP1 might indicate a role in ER shaping. Low expression levels in developed cells do not support this hypothesis however, but the ability to induce a shape to the membrane bound organelle could be important in some pathologies. For example, neurological disorders, such as hereditary spastic paraplegia, are caused by misshaped ER, and ARL6IP1 is one of the genes that causes this disease (Hübner et al. 2014).

Clinically PRAF3 has been related to several neurological diseases. It is upregulated after morphine treatment and is required for the acquisition of ethanol resistance (Wu et al. 2011; Li et al. 2008). It also significantly suppresses cell migration, metastases, invasion and ILK pathway-induced angiogenesis, probably through the trafficking of integrin \( \alpha_\text{V}/\beta_3 \) (Bai et al. 2009; Lu et al. 2013). PRAF3 is downregulated in hepatocellular carcinoma and esophageal squamous cell carcinoma (Wu et al. 2014a; Shi et al. 2012) and is a prognostic marker in gastric cancer (Wang et al. 2012). Depletion of PRAF3 also reduces bone density and morphology (Wu et al. 2014b).
YPT interacting proteins

The YPT interacting protein family in yeast has 6 members, but only 4 of them are orthologs of the human YPT interacting protein family or YIPF. Yip1p or YIPF5 in human has been extensively studied for its role in membrane trafficking, together with its homolog Yif1p or YIF1A and YIF1B in human (Shakoori et al. 2003). Both genes are essential in yeast (Yang et al. 1998; Matern et al. 2000). Yip4p is related to YIPF4 and YIPF6 and Yip5p to YIPF1 and YIPF2 (Shakoori et al. 2003). Yip3p is an ortholog of PRAF proteins, discussed in the previous section, while Yip2p (or Yop1p) is related to the REEP5 protein in humans, which has a structural role in the endomembrane system (Voeltz et al. 2006). The human YIP family has two further proteins, YIPF3 and YIPF7, which are less related to yeast orthologs (Shakoori et al. 2003).

All of the YIPF proteins are expressed in a broad range of human tissues. Only a single RNA transcript was found for YIPF3 and YIF1A, while the others have multiple splicing variants. The expression was the highest in testis for all YIPF proteins, suggesting their role in spermatogenesis. High expression was also detected in liver, pancreas and prostate, in general tissues with active exocytic pathways (Shakoori et al. 2003).

All YIPFs localise to the Golgi. YIPF4 and YIPF5 have been reported to co-localise with the cis-Golgi, while others show somehow different Golgi profiles, more resembling medial-Golgi and trans-Golgi profiles. High expression levels disrupt the localisation: YIPF proteins showed ER localisation, YIPF1 and YIPF2 have also been found on punctate structures and on the plasma membrane (Shakoori et al. 2003).

YIPF3, YIPF4, YIPF5 and YIF1A all localise to the cis-Golgi, but differences exist. When separated on native PAGE, YIPF3 resolves in 3 bands. First it is translated into a 40 kDa protein, then N-glycosylated in the ER and O-glycosylated in the Golgi. This produces a band of 46 kDa. Another smaller band of 36 kDa is seen as well, which is probably a truncated protein (Tanimoto et al. 2011). YIPF5 and YIF1A also localise to the ER and punctate structures and particularly co-localise with ERGIC-53, which is a marker for the intermediate complex between the ER and Golgi (ERGIC) (Tang et al. 2001; Yoshida et al. 2008). Treatment with GTP-locked mutant Sar1, which induces relocation of the ERGIC to the ER, relocated both YIPF5 and YIF1A to the ER (Yoshida et al. 2008). These two particular proteins interact and form large complexes of 320 kDa, 160 kDa and 120 kDa (Yoshida et al. 2008).

Structure

Human YIPF proteins are made of two parts: a long (~150 residues) cytosolic region at the N-terminus followed by a hydrophobic region made up of 5 TMDs (Figure 1.8c). These domains are 20 residues long and packed closely together, some of them forming hairpin-like structures. The linkers between them are slightly longer, between 3–10 residues. YIPF4, YIPF5, YIPF6, YIPF7 and YIF1A end with a TMD, whereas YIPF1, YIPF2 and YIPF3 have another luminal C-terminal region of approximately 50 residues. Sterical hindrance between TMDs is less likely. Considering an average diameter of an α-helix is 12 Å and the length of amino acid is approximately 8 Å, the sterical hindrance would be overcome by loops longer than 2 residues. Despite this, YIPF5 seems to have a role in ER shaping, although the experimental data are inconsistent. Kano et al. reported
that depletion of YIPF5 does not have an effect on the structure of the ER (Kano et al. 2009), but one year later, Dykstra et al. observed formation of whorl-like membrane structures in the ER after the depletion of the same protein (Dykstra et al. 2010). The effect was replicated by the E95K mutation and rescued by overexpression of wild-type YIPF5. This mutation is distant to the TMDs, so the reorganisation of the ER might occur because of the disrupted interactions. Indeed, yeast YIPF5 interacts with Yop1p, an ER shaping protein known as REEP5 in human (Calero et al. 2001). However, the mutation in the region that facilitates interaction with REEP5 is not sufficient to induce reorganisation, so there is likely another interaction partner or mechanism involved (Dykstra et al. 2010).

1.2.1.2.2 Interacting partners

The interaction networks of YIPF proteins are much less explored than those of PRAFs. They form oligomers within the family: YIPF1 and YIPF2 strongly interact with YIPF6, which was also seen between yeast orthologs YIP4 and YIP5 (Shakoori et al. 2003; Calero et al. 2002b). Another important interaction occurs between YIPF4, YIPF5 and YIF1A (Shakoori et al. 2003). Interaction between YIPF5 and YIF1A has been extensively studied in the context of membrane trafficking and they seem to complement each other in some manner (Jin et al. 2005; Yoshida et al. 2008; Calero et al. 2001; Barrowman et al. 2003; Chen et al. 2004; Calero et al. 2002a; Dykstra et al. 2010). They localise to similar compartments and co-localise with the same markers. Depletion of YIF1A changes the glycosylation profile of YIPF5, but the opposite deletion does not produce this effect. This suggests that YIPF5 works upstream of YIF1A (Calero et al. 2001).

Few interactions with Rab proteins have been reported. In yeast Yip1p (human YIPF5) and Yif1p (human YIF1A) interact with Ypt1 (Rab1 in human) (Calero et al. 2001; Barrowman et al. 2003). Interactions with other Golgi Rabs have also been reported: Ypt31, Ypt32 and Ypt6 (Chen et al. 2004). In human cells there was a significant decrease in the membrane fraction of Rab6α (52%) and milder decrease of Rab1β (82%) after the depletion of YIPF5 (Kano et al. 2009). YIPF5 interacts with membrane shaping proteins, for example REEP5, although it does not cause reorganisation of the ER after depletion or mutation (Dykstra et al. 2010). Another interesting interaction might explain this observation. Yos1p is a poorly characterised protein with two hydrophobic regions and has a similar localisation profile to YIPF5. Depletion of this protein does induce some reorganisation of ER membranes, however its role in shaping the ER is not fully established. It does co-precipitate with YIPF5 and the interaction can be disrupted by the E70K mutation in yeast Yip1p (Heidtman et al. 2005). Mutation of E95K in human YIPF5, which is equivalent to E76K in yeast, induced a whorl-like reorganisation of the ER. A similar effect was observed by mutating residues around E95K, particularly L92 and L96 (Dykstra et al. 2010). The interaction with Yos1p may therefore be facilitated through this particular region and hence it may be that it is the disruption of this interaction that induces reorganisation of the ER.

1.2.1.2.3 Function

YIPF proteins localise to different regions of the Golgi and they interact with each other (Shakoori et al. 2003). It is not yet clear whether they work in synergy or whether their function is dis-
tributed to different parts of the Golgi - YIPF3 and YIPF4 at the cis-Golgi, YIPF5 and YIF1A at the ERGIC and YIPF1 and YIPF2 at the trans-Golgi (Shakoori et al. 2003). No co-localisation data are available for YIPF6 and YIPF7. They might be involved in structuring the Golgi, as depletion of YIPF3 or YIPF4 induce Golgi fragmentation. Secretion assays do not show any change in Vesicular stomatitis Indiana virus G-protein (VSV-G) trafficking from the ER to the Golgi (Tanimoto et al. 2011). Depletion of YIPF5 or YIF1A also induces fragmentation of the Golgi (Yang et al. 1998; Tang et al. 2001; Dykstra et al. 2010), although several more recent studies show the contrary, namely that depletion of YIPF5 or YIPF7 does not induce Golgi fragmentation (Dykstra et al. 2013; Barone et al. 2015). YIPF5 is also incorporated into COPII vesicles (Tang et al. 2001), and experiments using blocking antibodies have demonstrated that YIPF5 is needed for the fusion of COPII vesicles. Fusion was blocked only when the antibody was added before the budding process, suggesting that YIPF5 might be important for packaging of fusion machinery. YIPF5 interacts with ER SNAREs and with the cytosolic domain of the vSNARE-Bos when expressed alone (Barrowman et al. 2003). In the same assay the budding process was only mildly blocked, but another study has demonstrated efficient blockage of COPII budding with the same blocking antibodies. In contrast, tethering and fusion remained intact. These antibodies prevented the recruitment of the complete COPII machinery on to budding vesicles (Heidtman et al. 2003; Chen et al. 2004). The resulting blockage of COPII transport with these antibodies resulted in proliferated ER membranes (Heidtman et al. 2003; Kaiser et al. 1990).

A study by Kano et al. could not find any effect of YIPF5 on ER structure, COPII trafficking, or Shiga toxin uptake. In contrast they observed that depletion of YIPF5 did block the retrograde transport that operates between the Golgi and ER independent of COPI (Kano et al. 2009). In addition, they also found that YIPF5 interacts with Rab6, which is involved in retrograde transport (Kano et al. 2009; Girod et al. 1999; Goud et al. 1990). YIPF5 therefore does seem to be involved in membrane trafficking, although its precise role remains unknown.

1.2.1.3 Vesicle coat, budding and fusion proteins

Membrane shaping mechanisms are important for organelle morphology as well as for membrane trafficking. The formation of a transport vesicle relies on reshaping of a membrane into a bud that is then pinched off. Reshaping of membranes is a mechanical process in which specialised coat proteins, specific for each trafficking pathway, assemble at the membrane into a cage-like structure and mechanically lift the membrane to form high curvature. At least four different coat complexes are known in eukaryotic cells, each associated with specific trafficking pathways (Table 1.1).

Vesicles coated with COPII proteins mediate ER to Golgi trafficking. Vesicles bud only from specific microdomains of ER elements, free of ribosomes, called ERES. The minimal machinery for COPII coat formation requires only 6 proteins. COPII vesicle formation is initiated by a small GTPase Sar1, which is activated by its GEF, the transmembrane protein Sec12 (Barlowe et al. 1993). Activation of Sar1 exposes an amphipathic α-helix which probably induces initial curvature of the membrane (Lee et al. 2005a; Bielli et al. 2005). Sar1-GTP then recruits the Sec23/24 complex which has a bowtie-like structure with a concave surface facing the membrane thus stabilising
Table 1.1: Coat complexes and associated trafficking pathways.

<table>
<thead>
<tr>
<th>Coat complex</th>
<th>Trafficking pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>COPII</td>
<td>ER to Golgi trafficking</td>
</tr>
<tr>
<td>COPI</td>
<td>Golgi to ER, intra-Golgi trafficking</td>
</tr>
<tr>
<td>Clathrin, caveolin1</td>
<td>Plasma membrane to endosome trafficking</td>
</tr>
<tr>
<td>Retromer</td>
<td>Early endosome to TGN trafficking</td>
</tr>
</tbody>
</table>

the curvature (Matsuoka et al. 1998; Bi et al. 2002). The Sec23/24 complex exhibits multiple functions. First, Sec23 is a GAP for the activation of Sar1 (Yoshihisa et al. 1993; Bi et al. 2002). Sec24 is a cargo binding protein that interacts with ER export motifs on cargo proteins or cargo adapters, such as ERGIC-53 or cornichons (Miller et al. 2002; Mossessova et al. 2003; Mancias et al. 2008; Appenzeller et al. 1999; Herzig et al. 2012; Sauvageau et al. 2014; Pagant et al. 2015). Lastly, Sec23/24 complex recruits the outer layer of the coat, a heterotetramer composed of Sec13 and Sec31. The Sec13/31 complex can self assemble into multiple different geometries further increasing the curvature of the membrane (Stagg et al. 2006; Stagg et al. 2008) and significantly enhances the GTPase activity of Sar1, up to 10-fold (Antonny et al. 2001). A newly formed transport vesicle is then pinched off in a process that involves the N-terminal domain (Lee et al. 2005a) and GTPase activity of Sar1 (Bielli et al. 2005). Rapidly after fission, coat proteins disassemble and leave the vesicle (Barlowe et al. 1994). The release of COPII components from the vesicle is a consequence of the increased hydrolysis of GTP by Sar1 (Antonny et al. 2001), however with a slight delay as both cargo and Sec12 (a Sar1 GEF) can prevent coat disassembly by interacting with the coat itself (Sato et al. 2005a) or keeping the Sar1 activated through GEF activity (Futai et al. 2004; Sato et al. 2005a), respectively. Recent studies show that components of the COPII coat, particularly Sec23, can interact with the Golgi tether TRAPPI and are needed for the recognition of vesicles before docking takes place (Cai et al. 2007b). This suggests that vesicle uncoating occurs only after it is tethered to the Golgi membrane (Lord et al. 2011), although this would not be supported by earlier live observations of the process which seem to show that uncoating occurs rapidly after ER exit (Watson et al. 2005).

Retrograde trafficking, between the Golgi cisternae and from Golgi to ER, is predominantly mediated by COPI vesicles. Cargo in COPI vesicles recycling from the Golgi to ER is recognised by the presence of a C-terminal KDEL motif (luminal proteins) (Lewis et al. 1990), FF.. motif (p24 proteins) (Fiedler et al. 1996) or dilyisine KK..$ and K.K..$ motifs (transmembrane proteins) (Nilsson et al. 1989; Jackson et al. 1990). Assembly of the COPI coat is initiated when p23, a p24 family protein, recruits the GTPase Arf1 to cis-Golgi membranes (Gommel et al. 2001; Donaldson et al. 1992). Arf1 is then activated by its GEFs GBF1, BIG1 or BIG2, all containing a Sec7 domain (reviewed in D’Souza-Schorey et al. 2006). Nucleotide exchange induces a conformational change in Arf1 and exposure of a myristoylated tail which anchors the GTPase into the membrane (Franco et al. 1996; D’Souza-Schorey et al. 2006). Activated GTPases recruit the preassembled heptameric COPI coatomeric complex which forms the inner layer—adapter and trimeric α−, δ−, β−, ζ− COP complex which forms the outer layer—cage (Duden et al. 1991; Harrison-Lavoie et al. 1993; Malhotra et al.
sequential degradation) to lysosomes. This is to promote the export of cargo from endosomes, therefore preventing its trafficking (and concomitantly recruiting Vps/five.pnum/Vps/one.pnum/seven.pnum complex (Seaman et al./one.pnum/nine.pnum/eight.pnum; Haft et al./two.pnum/zero.pnum/zero.pnum/zero.pnum). The function of the retromer is another type of a coat functioning in post-Golgi trafficking. It was first discovered as associated with the trafficking of the TGN sorting receptor Vps10 from endosomes to the TGN (Seaman et al. 1998). It is composed of two subcomplexes: the first is composed of Vps5 and Vps17 sorting nexins that can sense membrane tubules or membrane bending (Carlton et al. 2004; Weering et al. 2012); the second is a trimeric complex composed of Vps26, Vps29 and Vps39 and is important for cargo recognition (commonly named cargo selective complex or CSC). The trimeric complex cannot directly bind to the membrane and thus binds to the previously recruited Vps5/Vps17 complex (Seaman et al. 1998; Haft et al. 2000). The function of the retromer is to promote the export of cargo from endosomes, therefore preventing its trafficking (and consequential degradation) to lysosomes.
CHAPTER 1. INTRODUCTION

All transport carriers approaching the target membrane must be identified before they are accepted for fusion. This is done in the process of tethering, whereby specific tethers recognise different types of coat and cargo (Barlowe 1997; Cao et al. 1998). Tethers are usually peripheral membrane proteins that are recruited to the target membrane as effectors of Rab and Arl GTPases (Mayer et al. 1997; Ungermann et al. 1998; Cai et al. 2007a). They can be divided into two major groups: homodimeric long coiled-coil proteins and multisubunit tethering complexes (MTCs) (Gillingham et al. 2003; Whyte et al. 2002). Coiled-coil tethers are long, around 3000 residues in size, and are peripheral membrane proteins that can capture a vesicle at distances of over 200 nm from the membrane (Gillingham et al. 2003). They are mostly found in the Golgi, therefore they are commonly named ‘golgins’. EEA1 is one exception, which shows an endosomal localisation (Christoforidis et al. 1999a). Golgins are recruited to the membrane in three different ways: as effectors of Rab or Arl proteins by interacting via their carboxy-terminal GRIP domains (golgin-245, GMAP-210) (Panic et al. 2003a; Panic et al. 2003b); or they have a carboxy-terminal TMD (giantin, CASP) (Gillingham et al. 2002); or they bind to specific lipids in the membrane (Simon-ensen et al. 1998). Golgins act like ‘tentacles’ at the Golgi membrane, allowing vesicles to remain in close proximity to the Golgi. They can also prevent the release of homotypic vesicles in intra-Golgi trafficking (Sinka et al. 2008).

MTCs can be also further divided into two groups: early secretory pathway MTCs, also called CATCHR, comprising of Dsl-1, COG, GARP and exocyst; and endosomal/lysosomal MTCs, composed of VpsC, HOPS and class C core vacuole/endosome tethering (CORVET). TRAPP is present on both pathways as well as playing a role in autophagy. The exocyst complex is made of 8 subunits: Sec3, Sec5, Sec6, Sec8, Sec10, Sec15, Exo70 and Exo84 (Novick et al. 1980; He et al. 2009a); this large complex is important for trafficking of Golgi derived vesicles to the plasma membrane. Six subunits are recruited to the Golgi derived vesicle by a yeast Rab Sec4 and the remaining two, Sec3 and Exo70, mark the sites of secretion on plasma membrane by directly interacting with PI-4,5-P$_2$ and several non-Rab GTPases (Boyd et al. 2004; Zhang et al. 2008; Guo et al. 2001; Yamashita et al. 2010). The exocyst complex also affects assembly of SNARE complexes (Sivaram et al. 2005; Grote et al. 2000). The Dsl-1 complex has only three subunits and is involved in retrograde Golgi to ER trafficking (Kamena et al. 2004; Kraynack et al. 2005; Zink et al. 2009). Dsl-1 forms a 20 nm tall tower anchored to the ER through the interaction with ER t-SNAREs. At the tip of the tower is a disordered region forming a lasso loop that interacts with COPI subunits (Ren et al. 2009). The COG complex functions in trafficking pathways between Golgi cisternae. It is made of 2 subcomplexes: COG2-COG3-COG4 and COG5-COG6-COG7, linked with heterodimers of COG1 and COG8 (Ungar et al. 2002; Ungar et al. 2005). The complex interacts with Ypt1/Rab1, Sly1 and the SNARE Sed5, a yeast ortholog of syntaxin-5 (Suvorova et al. 2002; Laufman et al. 2009; Shestakova et al. 2007). Mutations in the COG complex cause defects in glycosylation and lead to various diseases (Richardson et al. 2009). The GARP complex is involved in tethering endosomes to the TGN and has four subunits: Vps51, Vps52, Vps53 and Vps54 (Pérez-Victoria et al. 2009). The complex cooperates with Rab6a and interacts with SNAREs Tlg1, Syntaxin-6, Syntaxin-16 and VAMP4 (Fridmann-Síkis et al. 2006; Pérez-Victoria et al. 2010). HOPS and CORVET function at endosomes/lysosomes. They have four shared subunits: Vps11, Vps16, Vps18 and Vps33 and an additional two specific but homologous subunits (Nickerson et al. 2009). CORVET is required for
fusion of Golgi-derived vesicles with endosomes (Raymond et al. 1990). Its two specific subunits are Vps3 and Vps8 with which it interacts with Vps21, a yeast ortholog of Rab5 (Markgraf et al. 2009). HOPS is required for fusion events at endosomes, autophagosomes, multivesicular bodies (MVBs) and AP3 Golgi-derived vesicles. It tethers the vesicles by interacting with Rab7 on membranes (Hickey et al. 2009).

A unique feature of TRAPP complexes is their GEF activity towards Ypt1/Rab1 (Jones et al. 2000). Three different forms of TRAPP are known: TRAPPI, TRAPPII and TRAPPIII. TRAPPI has seven subunits: Bet5, Trs20, Bet3A, Bet3B, Trs23, Trs31, and Trs33 (Jiang et al. 1998; Kim et al. 2006; Sacher et al. 2000); TRAPPII has an additional three subunits: Trs65, Trs120, and Trs130 (Sacher et al. 2001); and TRAPPIII has one additional subunit, Trs85, making eleven in total (Lynch-Day et al. 2010). TRAPPI is essential for the tethering of COPI vesicles with Golgi membranes - the Bet3 subunit interacts with the COPII subunit Sec23 (Cai et al. 2007b). TRAPPII specifically mediates intra-Golgi transport and interacts with COPI, but not COPII vesicles (Cai et al. 2005; Yamasaki et al. 2009). It is also a GEF for Ypt1/Rab1 (Yamasaki et al. 2009). TRAPPIII is involved in autophagosome formation and localises to the phagophore assembly site (Lynch-Day et al. 2010). It also functions in trafficking from the endosomes to the Golgi, which includes protein Atg9, involved in the formation of the autophagosome (Shirahama-Noda et al. 2013).

Recognition of a vesicle by a tether is followed by fusion. Membrane fusion is mediated by a family of membrane-bound SNARE proteins. SNARE proteins have a coiled-coil domain of approximately 70 residues in heptad repeats. There are 38 SNAPES in humans, which can be divided into v-SNAREs or t-SNAREs, depending on whether they function on vesicle or target membranes, respectively (Figure 1.9). Another nomenclature is also in use, based on the residue that mediates interaction in parallel four-helix: Q-SNAREs have a conserved glutamine and R-SNAREs have a conserved arginine (Fasshauer et al. 1998). In a synaptic SNARE complex, made of syntaxin-1A, synaptobrevin-2 and SNAP-25B, three glutamines, contributed by syntaxin-1A and SNAP-25B, interact with arginine on synaptobrevin-2. Most R-SNAREs act as v-SNAREs and most Q-SNAREs act as t-SNAREs, however there are exceptions: the R-SNARE Ykt6 acts in a t-SNARE complex and Q-SNAREs GS15, Bet1, and Slt1 act as v-SNAREs. Typically SNAREs have one or two SNARE motifs and a C-terminal membrane anchor, usually a TMD. Seven SNAREs do not have TMDs and instead undergo palmitoylation at a C-terminal cysteine motif (SNAP-23, SNAP-25, SNAP-29, SNAP47, Stx9/19, and Stx11) or farnesylation followed by palmitoylation (Ykt6) (Fukasawa et al. 2004). Some SNARE also have an N-terminal peptide before the SNARE motif. These peptides are important for interactions with other proteins, for example interaction with SM proteins (Stx1, Stx5, and Stx16), interaction with the protein’s own SNARE motif which inhibits formation of a SNARE complex (Habc peptide in Stx16 (Misura et al. 2000) and longin peptide in Vamp7 (Schäfer et al. 2012)), or they are special motifs that regulate trafficking (dileucine motif in VAMP4 recycles the protein from endosomes to the TGN (Zeng et al. 2003)). Different combinations of SNAREs form SNARE complexes and the same SNARE can be part of different complexes on different organelles. Each SNARE complex has at least one t-SNARE from the Stx family (Hong et al. 2014).

In a fusion event the v-SNARE interacts with the t-SNARE, forming a trans-SNARE complex, in which four SNARE motifs, one from the R-SNARE and three from Q-SNAREs, assemble into a parallel four-helix bundle bringing the membranes close to each other eventually enabling fusion.
Figure 1.9: Localisation and function map of SNARE proteins in a mammalian cell. Specific SNARE proteins mediate vesicle docking and fusion at each membrane trafficking interface in an eukaryotic cell. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology Jahn et al. 2006, copyright 2006.

(Hong et al. 2014; Poirier et al. 1998). Formation of a SNARE complex is regulated by SM proteins and tethering factors and these specific interactions allow only physiological pairing combinations. Four SM proteins are known: Sly1, VPS45, VPS33 and Munc18. Each of them functions in its own pathway: Sly1 in ER/Golgi trafficking, VPS45 at TGN/endosomes, VPS33 in endocytic/lysosomal trafficking and Munc18 in trafficking to the plasma membrane. SM proteins interact with SNAREs and promote only physiological pairing (Peng et al. 2002). Tethers also interact with SNAREs to influence the assembly of intermediate complexes, stabilise complexes or spatially and temporally organise fusion events. Subunits of GARP, Dsl-1 and COG can stabilise the formation of t-SNARE subcomplexes (Laufman et al. 2013; Ren et al. 2009; Pérez-Victoria et al. 2009). The Cog6 subunit stabilises the complex Stx6-Stx16-Vti1-VAMP4 (Laufman et al. 2011). The HOPS complex stabilises trans-SNARE complexes by inhibiting its disassembly by Sec17p/Sec18p (SNAP/NSF) (Xu et al. 2010). The relation can be also reversed - some tethers specifically recognise SNAREs on certain membranes. For example, SNAREs can enhance binding of HOPS to liposomes (Stroupe et al. 2009) and ER SNAREs recruit the Dsl-1 complex for tethering COPI vesicles (Meiringer et al. 2011).

1.2.2 Mechanisms of membrane trafficking

The Golgi is a central point in membrane trafficking pathways, functioning as a hub where the fate of the cargo is decided. Accordingly, trafficking between the ER and Golgi and post-Golgi trafficking can be seen as distinct entities in terms of their spatial and temporal organisation. Two
membrane trafficking mechanisms will be discussed in this section: trafficking between ER and Golgi; and post-Golgi trafficking.

1.2.2.1 Trafficking between the ER and Golgi

Anterograde trafficking from ER to Golgi is the first trafficking step in a lifetime of a protein. Newly synthesised soluble or transmembrane proteins are packaged into COPII vesicles at long-lived ER subdomains termed ERES. Transport vesicles then attach to microtubules through the interaction between the COPII component Sec23A and the motor subunit dynactin (Watson et al. 2005) and then ultimately fuse with the Golgi membrane - the vesicle is first recognised by a tether, then uncoated and fused with cis-Golgi with the help of SNAREs (Cai et al. 2007b; Lord et al. 2011). Recent studies with high speed imaging showed that the cis-Golgi in yeast might form contacts with ERES and capture the vesicles soon after they are formed (Kurokawa et al. 2014). Metazoan cells have another compartment between the ER and Golgi called the ERGIC (Presley et al. 1997). The ERGIC can be considered as a distinct organelle involved in sorting of the COPI retrograde cargo (Klumperman et al. 1998), concentrating the cargo in anterograde trafficking (Martínez-Menárguez et al. 1999) and performing a secondary quality control of cargo function (Breuza et al. 2004). COPII vesicles travel towards the ERGIC in a process that can occur without functional microtubules. ERGIC or ERGIC derived vesicles likely travel towards the cis-Golgi in a microtubule-dependent way (Scales et al. 1997; Presley et al. 1997; Hammond et al. 2000). Some cells need to export very large cargo, such as procollagen (300–400 nm) (Bonfanti et al. 1998) or lipoprotein particles (100–500 nm), that would not fit into a typical sized COPII vesicle (60–90 nm) (Barlowe et al. 1994). Large cargo still exits the ER using COPII coated carriers, which indicates that the COPII coat is flexible and can adapt its size depending on the cargo (Bonfanti et al. 1998). Cargo can be incorporated into vesicles by default (bulk-flow) (Thor et al. 2009) or selected by cargo receptors (Kuehn et al. 1998; Giraudo et al. 2003).

Proteins that arrive at the Golgi are generally trafficked further along the secretory pathway or they can be recycled back to the ERGIC or ER. In the cisternal maturation model of intra-Golgi trafficking no anterograde carriers have been observed, therefore cargo stays in the same cisterna which slowly matures from cis- to trans-Golgi (Bonifacino et al. 2004). Instead, trafficking machinery and glycosylation enzymes are recycled backwards in COPI vesicles, from trans- to cis-Golgi (Bonifacino et al. 2004). COPI coated vesicles also recycle proteins from the Golgi to the ER. These proteins are mostly components of trafficking machinery, such as SNAREs, cargo receptors and certain glycosylation enzymes. Proteins recycled from the Golgi to the ER have motifs, commonly KDEL or dilyssine KK..$, that mark them for retrograde trafficking (Jackson et al. 1990). In metazoans the ERGIC brings in another level of complexity. It is the first point where ER proteins can be recycled back to the ER in COPI vesicles (Martínez-Menárguez et al. 1999; Ben-Tekaya et al. 2005). This requires disassembly of cargo from its cargo receptor, which indicates that the environment has to be different. Some studies suggest, that release of cargo from ERGIC-53 is pH and Ca$^{2+}$ dependent (Appenzeller-Herzog et al. 2004). A slightly lower pH and decreased Ca$^{2+}$ concentration in the ERGIC compartment would therefore allow release of cargo from receptors. Lower pH has also been shown to increase binding of proteins with a KDEL motif to the KDEL receptor, which
allows its recycling to the ER (Scheel et al. 1996). Other functions of the ERGIC is concentration of cargo for further anterograde transport (Martínez-Menárguez et al. 1999) and it also performs quality control of received proteins. The ERGIC has the highest levels of several chaperones and the glycoprotein folding sensor UDP-glucose:glycoprotein glucosyltransferase (Zuber et al. 2001; Zuber et al. 2004; Breuza et al. 2004).

Some proteins are recycled back to the ER independently of the COPI coat or Arf1 (Girod et al. 1999). This pathway has been shown to traffic some toxins, such as *Shigella dysenteriae* Shiga toxin and *Escherichia coli* Shiga-like toxin (both lacking a KDEL retrieval sequence) and also some Golgi-resident glycosylation enzymes to the ER (Girod et al. 1999). This pathway is regulated by Rab6a, but it remains poorly characterised, mostly due to the involvement of Rab6a with other pathways (Storrie et al. 2012; Heffernan et al. 2014).

### 1.2.2.2 post-Golgi trafficking

Cargo delivered to the cis-Golgi is processed in Golgi cisternae by glycosylation enzymes. Proteins get simple sugars attached in the ER, but these are then further processed in the Golgi forming complex glycosylation profiles (Moremen et al. 2012). When the cargo reaches the TGN its fate is decided. Secreted proteins are packaged into transport carriers that fuse with the plasma membrane. Components of the extracellular matrix are secreted continuously in a process called constitutive exocytosis. On the other hand several proteins, such as α1-antitrypsin, are only secreted following stimulus. This process is called regulated exocytosis. The majority of proteins that reach the TGN never leave the cell and are instead trafficked to other internal cellular organelles or the plasma membrane. These proteins are particularly important in endocytic pathways. Endocytosis is a process of material uptake from the plasma membrane, usually mediated by cell surface receptors. A typical endocytic route starts by binding of a ligand to its receptor, for example low density lipoprotein (LDL) binds to its receptor LDLR. The ligand-receptor complex is then internalised and packaged into a clathrin-coated vesicle. This vesicle is transported to an early endosome, where the acidic environment promotes release of ligand from receptor. The receptor is then recycled back to the plasma membrane via recycling endosomes. Early endosomes containing ligands, proteins and some non-recycled receptors, such as epidermal growth factor (EGF) receptor, provide sorting capacity. They can have vesicular and tubular domains, and elements of the early endosomes can gradually bud away and be transformed into late endosomes. The transformation is induced by trafficking of lysosomal enzymes from the TGN to early endosomes, which changes the internal environment by decreasing pH and activating degradation enzymes. Late endosomes can be transformed into MVBs and then to lysosomes, where ligand and proteins are completely degraded (for a review on endocytosis see Doherty et al. 2009).

Furthermore, receptors with their ligands can be endocytosed by different mechanisms, which can be organised in three different classes: clathrin-dependent, caveolin-1-mediated and clathrin/caveolin-1-independent endocytosis. Clathrin-dependent endocytosis involves recruitment of the clathrin coat to membrane domains enriched in the AP2 adapter protein. Similar mechanisms mediate trafficking of receptors and lysosomal enzymes between the TGN and endosome populations (Robinson 2004). The caveolin-1-mediated mechanism is a non-clathrin
pathway involving the formation of invaginations termed caveolae or ‘small caves’, coated with caveolin-1 (Rothberg et al. 1992; Parton et al. 2007). This pathway is associated with transcytosis, a direct trafficking of vesicles from the apical to the basolateral membrane in some cell types (Heltianu et al. 1989). A further pathway, independent of both clathrin and caveolin-1, is associated with the uptake of certain bacterial toxins, such as cholera toxin, GPI-anchored proteins, growth hormones and many other molecules (Kirkham et al. 2005).

Interlinked endocytic pathways demand tight regulation of their trafficking steps as well as careful cargo selection. Regulation is facilitated by Rab proteins. Rab5 localises to early endosomes and is probably the most widely studied Rab protein. Stimulation of a receptor, such as EGFR, also activates Rab5 which activates trafficking of the receptor to the early endosome (Barbieri et al. 2000). Activated Rab5 also recruits the HOPS complex which contains Rab7 GEF and eventually activates it. This may lead to the conversion of early endosomes to late endosomes, as Rab7 is specific for late endosomes and lysosomes (Rink et al. 2005). Enzymes that are trafficked to late endosomes and lysosomes are selected at the TGN by their specific glycosylation profile - typically decorated with M6PR (Pearse 1985). These proteins bind mannose-6-phosphate receptors at the TGN and are transported to endosomes and lysosomes. The receptor is then recycled back to the TGN by retromer coated carriers. Plasma membrane receptors also have specific motifs that mark them for internalisation. For example, the transferrin receptor, a common tool to study clathrin mediated endocytosis, has a YTRF motif on its cytoplasmic tail, which binds to the AP2 adapter and eventually drives clathrin-coated vesicle production (Collawn et al. 1990). Three consensus motifs are known: Y...φ, [D,E]...L[L,I] and F.NP.Y (Ohno et al. 1995; Bonifacino et al. 2003) with some derivatives, such as the ones discovered in a recent screen for novel endocytic motifs (Kozik et al. 2010).

1.3 Short linear motifs - SLiMs

In the past decade the focus of much cell biology research has changed from describing components of the cell to describing systems. Proteins are the main effectors in the cell and their function arises from their interaction networks. The specificity (and strength) of interactions between proteins is of great importance for the stability of the systems and most of the biological processes in a cell are defined by its specific interaction networks. Proteins have a modular structure with different subunits performing different functions. Protein domains and many important modules in a protein are well known. Globular domains form three dimensional structures, and appear as structured regions in a protein. They are conserved and relatively easy to identify with the aid of computational tools, such as SMART or Pfam (Finn et al. 2006). Proteins also contain short linear motifs (SLiMs) that have regulatory roles. Many examples of post-translational modifications are known, where only a short linear sequence is recognised by a modification enzyme. These linear motifs usually appear in unstructured or disordered regions of a protein. These regions are relatively poorly investigated, mostly because structural studies have been focused on globular domains and these unstructured regions were often omitted from expressed proteins (e.g. physically removed from constructs).
SLiMs are also difficult to study because of their short length (3–12 residues) and because they are abundant in poorly conserved regions of sequence. Their short length also makes them more evolutionarily labile, which could also be an advantage - their regulatory role can be fine-tuned by a single point mutation. Another advantage of their short length is their ability to label certain proteins with a small distinct footprint, such as the KDEL motif used in the recycling of ER proteins (Lewis et al. 1990).

Interactions between proteins can be classified into several groups. Two proteins can interact with their globular domains. Domain-domain interactions involve large interfaces and have high binding affinity in the nanomolar range. They are probably the best characterised and much data have been collected in online databases, such as DIMA (Luo et al. 2011). Mutual fit interactions are characteristic for SNARE proteins. They also involve large interfaces between natively unstructured regions, which form a stable secondary structure after interaction. In SNARE proteins, their SNARE motifs have a linear structure and after binding with three other SNARE motifs form a stable secondary structure (Poirier et al. 1998). Induced fit interactions are interactions between an unstructured region and a structured domain. The unstructured region forms a stable structure only after it binds to a structured domain. An example of a large unstructured region interacting with a domain are the SARA proteins that interact with SMAD transcription factors (Wu et al. 2000).

Linear motifs are a subset of induced fit interactions. In this case a relatively short unstructured region interacts with a domain. As mentioned earlier, this is particularly convenient for modification enzymes, both because of easy to control specificity and simple interaction mechanisms. Proteins that have central roles in regulatory networks have many such motifs and are often called hub proteins. Interactions involving multiple motifs are also more specific and bind with higher affinity.

Known and validated linear motifs are entered into the Eukaryotic Linear Motifs (ELM) database (Dinkel et al. 2014). They are grouped into 6 classes, depending on their function. Post translational modifications motifs (PTMs) were described above. Docking motifs (DOC) are also involved in modifications, but in this case a motif just attracts the modification enzyme and is then modified in a different region (Reményi et al. 2006). In previous sections, several targeting motifs (TRGs), such as KDEL and dilyisine motifs were described. Another very abundant class are cleavage motifs (CLVs), motifs that are recognised by proteases. Functionally related are degron motifs (DEGs), which direct polyubiquitination of a protein and target it for degradation in the proteasome (Schrader et al. 2009). The most abundant class of motifs are ligand motifs (LIG), which are involved in binding to other proteins.

SLiMs are represented as strings of single letter amino acids. Some particular places in a motif are degenerate which means that a residue does not need a specific amino acid at this place for the functionality of a motif. The degenerate nature of SLiMs is represented by regular expressions. Regular expressions are strings of characters and metacharacters that are commonly used to search for patterns in text. Fixed characters in a string are represented by a letter and ambiguous spaces are represented by predefined metacharacters. For example, a regular expression representing a dilyisine motif is K.{0,1}K.{2,3}$ (Dinkel et al. 2014). It starts with a lysine, represented as a single letter symbol K, which is followed by zero or one instance ({0,1}) of any amino
Table 1.2: List and description of metacharacters used in SLiM regular expressions.

<table>
<thead>
<tr>
<th>Regular expression metacharacter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>KDEL</td>
<td>Sequence of amino acids as single letter amino acid codes</td>
</tr>
<tr>
<td>.</td>
<td>Any amino acid</td>
</tr>
<tr>
<td>[A,L,F]</td>
<td>Any amino acid from the list</td>
</tr>
<tr>
<td>{1,2}</td>
<td>Repeat last symbol 1 or 2 times</td>
</tr>
<tr>
<td>^</td>
<td>N-terminus</td>
</tr>
<tr>
<td>$</td>
<td>C-terminus</td>
</tr>
<tr>
<td>[^K,R]</td>
<td>Any amino acid except amino acids from the list</td>
</tr>
</tbody>
</table>

acid, represented as a dot (.). This is then followed by another lysine (K) and 2 or 3 instances of any amino acid ([2,3]). A $ sign marks the C-terminus of a protein. Other metacharacters are listed in Table 1.2.

Discovery methods, nowadays mostly computational, exploit some of the known properties of SLiMs. It is not difficult to characterise the environment where one would expect to find a SLiM, but because of their short length they are, like transcription factor motifs, prone to overprediction. A good predictor will be therefore particularly good at finding true positives, which is at present the main issue of SLiM discovery.

1.3.1 Properties of SLiMs

It is estimated, that about 85% of SLiMs occur in disordered regions (Fuxreiter et al. 2007). Disordered regions are intrinsically unfolded, lacking any three-dimensional structure. The amino acid composition of disordered regions is biased, they are slightly enriched in E, G, K, P, Q and S and depleted in C, F, I, L, V, W and Y (Romero et al. 2001). Composition bias is not sufficient to predict whether a region is disordered because many physicochemical properties and interactions between residues can induce the disorder. Modern predictors therefore use all of the above properties and more sophisticated ones are even based on machine learning algorithms. A large leap in quality of predictions happened when DisProt - a database of experimentally determined disorder - was established (Sickmeier et al. 2007). At present, more than 50 different predictors are available, some of them are described in He et al. 2009b.

SLiMs can appear or disappear due to a small number of mutations, which makes them hard to discover, but at the same time introduces a certain level of flexibility into regulation networks. Interactions involving SLiMs can change much easier and faster than domain interactions (Neduva et al. 2005a). Discovery with conservation methods is difficult as only few sites, sometimes a single one, can be conserved and many SLiMs are specific to only some species (Neduva et al. 2005b). Despite this, SLiMs are better conserved than surrounding unstructured regions (Neduva et al. 2005a).
The binding affinity of SLiMs is orders of magnitude lower than the affinity of domains - SLiMs would bind in the micromolar range, usually between 0.1-150 μM, while domains bind with nanomolar or even picomolar affinity. For example, the affinity of a cyclin binding motif L.C.E, present in the transcription factor E2F-2 and binding with retinoblastoma tumor suppressor, is 0.19 μM (Lee et al. 2002) and the affinity of a 14-3-3 binding motif found in beta-chains of granulocyte-macrophage colony-stimulating factor, interleukin-3 and interleukin-5 receptors is 0.15 μM (Stomski et al. 1999). This is suitable for short lived and quick response interactions. Binding of several motifs in the same interaction significantly increases the affinity, sometimes to the level of domain interactions (Drake et al. 2000).

Only a few of the sites in a SLiM are defined, which does not allow high specificity of interactions. It has been shown that the residues surrounding defined residues (the context of a SLiM) increase the specificity by increasing the binding affinity (Stein et al. 2008). For example, a Pbs2 peptide in yeast with a motif P..P binds only to the Sho1 SH3 domain out of 27 SH3 domains known in yeast (Zarrinpar et al. 2003). This implies that secondary information from the context of a motif defines its specificity.

A SLiM gets its secondary structure after binding to a domain. Induced secondary structure is another property of a SLiM. Some estimates suggest that one third of SLiMs form an α-helix structure in the interaction and one third form beta augmentation, i.e. addition of another strand to an existing β-sheet (Remaut et al. 2006). The last third of SLiMs form various structures (Vacic et al. 2007).

There is a certain bias towards some amino acids in SLiMs. G and L are generally avoided and there is a preference of R and K. Fixed sites in motifs are enriched in P, W, T and C while hydrophobic residues I, M and V are enriched in ambiguous sites demonstrating their interchangeability due to their similar chemical properties (Davey et al. 2012).

The properties of SLiMs described in this section are readily leveraged for discovery of novel SLiMs. While there is a great potential for discovery of novel SLiMs - it is estimated that only 5% of SLiMs have been discovered so far - the discovery process is far from trivial.

### 1.3.2 Discovery methods

The linear nature of SLiMs makes them more suitable for computational prediction. A handful of predictors have been developed and most of them assess a SLiM by its unexpectedness, i.e. how unlikely it is that a discovered SLiM would occur by chance (Kelil et al. 2014). According to benchmarks (Kelil et al. 2014) the most successful predictors are based on text pattern search methods, which look for overrepresented peptide sequences in a set of proteins. The second type is based on pattern models such as hidden Markov models. Some predictors, such as SLiMPred, can also predict regions of proteins with high probability of a SLiM and these are based on machine learning models, such as neural networks.

SLiMFinder (Edwards et al. 2007) is a text pattern type of discovery tool and is at present, according to benchmarks, one of the best known SLiM predictors (Kelil et al. 2014). The software is based on the algorithms SLiMBuild and SLiMChance. The predictor first accounts for redundancy in a set of protein sequences by classifying them into unrelated protein clusters (UPC). Then the
sequences can be masked by different strategies, such as disorder masking, conservation masking, Uniprot feature masking and terminus masking. The SLiMBuild algorithm then finds all possible motifs in masked UPCs. These candidate SLiMs are then statistically evaluated by the SLiMChance algorithm which calculates appropriate significance values based on binomial distribution.

SLiMPred (Mooney et al. 2012) is another predictor that does not search for motif pattern but rather finds regions in a protein that are more likely to have a SLiM. The predictor is based on an artificial neural network, trained with data on protein conservation, disorder prediction and known SLiMs from ELM database. Anchor (Dosztányi et al. 2009) is also a predictor of motif containing regions and is based on physicochemical properties of the region. FIRE-pro (Lieber et al. 2010) is a predictor based on mutual information exploration to find motifs enriched in certain dataset. Another recently developed text pattern predictor is MotifHound (Kelil et al. 2014), which works in a similar way to SLiMFinder.

1.3.3 SLiMs facilitating protein localisation

According to the ELM database (Dinkel et al. 2014) there are 23 recorded SLiMs that are involved in protein targeting. Among these are dilyssine, diarginine and the KDEL motif (discussed previously), which label proteins for recycling in COP1-dependent retrograde transport. Two motifs are present on cargo adapters in the formation of the clathrin coat and are involved in binding accessory proteins, usually cargo adapters, to the AP2, subunit. One motif has been found in ER-Golgi transport cargo receptors and there are 4 SLiMs that are involved in cargo sorting at the TGN. Besides endomembrane-specific SLiMs there are also 4 well known nuclear localisation motifs (NLS) and one nuclear export signal (NES). Two motifs are involved in targeting to the cilium and 5 are involved in targeting to peroxisomes.

The computational discovery of SLiMs is usually based on high-throughput experimental data that established the relations between different proteins. High-throughput studies on protein-protein interactions and post-translational modifications have been used with computational SLiM discovery tools and some of them have yielded interesting results (Ba et al. 2012; Edwards et al. 2011). However, no high-throughput prediction of localisation motifs has been published to date. Some interesting datasets are available, based on expression of GFP-tagged cDNA constructs (Simpson et al. 2000) or antibody staining, such as in the Human Protein Atlas project (Uhlén et al. 2015). Both datasets are manually curated, which introduces some ambiguity into the determination of protein localisation. An interesting approach would be to first classify the phenotypes by automated image analysis and then perform SLiM discovery. The analysis can then be further expanded with rich data about interactions and signalling pathways.

1.4 Imaging and image analysis methods in cell biology

The development of the light microscope was probably the most important advance and prerequisite to discover and study cells. A traditional light microscope is in its principle a stack of lenses that magnify the object and its features so that they are observable by a human eye. The first light microscope, dating back to the 16th century, has since evolved into a sophisticated piece of equip-
ment with many divergent and specialised versions of it. The level of detail that we can observe with a light microscope is limited by the wavelength of light, which proved to be, even in its best case, insufficient to observe greater details of cell organisation or even smaller objects, such as atoms. In 1926 Hans Busch invented electromagnetic lenses, which became the foundation of an electron microscope, prototyped in 1931 by Ernst Ruska and Max Knoll (Siegahn et al. 1993; Mulvey et al. 1996). Electron microscopy overcomes the limitations of resolution in light microscopy - while the light microscope can achieve the highest useful magnification of ~1500x, electron microscopes can go beyond 10,000,000x. More informative is the resolution, which indicates the minimum distance between two objects to be recognised as two separate entities. A light microscope with good lenses would reach the resolution of ~250 nm and an electron microscope can go better than 50 pm (Erni et al. 2009).

The information that is held in an image is easily interpreted by humans with prior knowledge. However, this observation-based interpretation can be dramatically improved with quantification techniques. Image analysis is an emerging field in microscopy, especially in high throughput studies, where manual interpretation is simply not feasible.

1.4.1 Light microscopy

The very basic use of a light microscope is observation of very small objects by trans-illuminating them from the back and capturing the transmitted light through the objective lens. A major advance was the introduction of fluorescence microscopy. In this case fluorophores are attached to biomolecules and then these biomolecules can be observed by exciting the attached fluorophore and capturing the emitted fluorescent light. Biomolecules in fixed specimens can be labelled by permeabilising the plasma membrane and incubating the specimen with fluorescently labelled antibodies, specific for the desired biomolecule. Some cellular compartments can also be tagged with cell-penetrating fluorophores, such as LysoTracker® for lysosomes or MitoTracker® for mitochondria. A major breakthrough in fluorescence microscopy was the discovery of green fluorescent protein (GFP), which can be genetically encoded in the cells and used to fluorescently label proteins (Prasher et al. 1992; Chalfie et al. 1994). Many variations with different excitation and emission properties have been developed since then, allowing tagging of several genes in the same specimen. The major advantage of fluorescent proteins is that they can be visualised in both live cells and fixed specimens, which makes them particularly useful in mechanistic studies. Fluorescent proteins with special features have been developed recently. These include fluorescent proteins that produce oxygen species upon excitation, such as KillerRed (Bulina et al. 2006), or pH and metal ion sensing fluorescent proteins (Miesenböck et al. 1998; Barondeau et al. 2002). The most important are photoswitchable fluorescent proteins, which can be activated by a pulse of light of a particular wavelength (Dickson et al. 1997). This led to the invention of the super-resolution microscope, for which a Nobel prize was awarded in 2014 (Klar et al. 2000; Betzig 1995; Moerner et al. 2003). Super-resolution microscopy allows the capture of light microscopy images with higher resolution than that of a conventional fluorescence light microscope. This is achieved by either overcoming some limitations of typical light microscopy or applying mathematical transformations to the images.
The resolution in light microscopy is defined by the wavelength and numerical aperture of lens. It can be expressed by the following equation:

$$d = \frac{0.61\lambda}{NA}$$

where $\lambda$ represents the wavelength of light and $NA$ the numerical aperture of lens system. Various sources of light are used in fluorescence microscopy and the wavelength can have an effect on the resolution of the image. Therefore it is important to carefully select the most appropriate fluorophores and illumination sources.

Another major development in fluorescence microscopy was the invention of the confocal microscope. The confocal microscope introduces a pinhole into the light path that blocks out-of-focus light and allows signal from a single plane of the specimen to reach the detector. With confocal microscope we can observe greater detail, but this comes at a price. The pinhole blocks part of the emitted light, which significantly decreases the signal. To overcome this we need to increase the illumination of the specimen, which can be achieved by using laser light sources. To prevent bleaching of fluorophores, associated with increased illumination, the images are usually acquired in a scanning mode, i.e. pixel by pixel. This technique illuminates each pixel separately thus reducing the overall bleaching. The signal in a scanning confocal microscope is captured by a photomultiplier, while wide field microscopes typically use charge-coupled device (CCD) type of detector (typically used in digital cameras).

Microscopes have been further enhanced by robotics and automation. These systems require little or no input from the user and can run uninterrupted for days. Such systems are able to load the plate, automatically detect the desired focal planes and acquire multiple images from the same well. They can be used with any optical setup, even immersion objectives and confocal units.

### 1.4.2 Electron microscopy

High resolution in microscopy can be achieved by changing the light illumination system for electrons. Since electrons can have 100,000x shorter wavelengths than photons of visible light, the resolution of an electron microscope can be as low as 50 pm and beyond. The idea of electron microscopy (EM) was developed after the invention of the electromagnetic lens in 1926. Modern microscopes use electromagnetic and electrostatic lenses as an alternative to the optical lenses in light microscopes.

The transmission electron microscope (TEM) is the most common type of electron microscope and follows the model of a light microscope. A beam of electrons is produced by an electron gun and accelerated by an anode towards a cathode. It passes through condensing lenses which focus the beam, and then pass it through the specimen. The specimen can be partly transparent to electrons and partly scatter them. The beam emerging from the specimen carries information about the specimen structure and is magnified through a set of electromagnetic and electrostatic objective lenses. The image can then be visualised on phosphor plates and captured by a CCD camera. TEM is used to investigate biological samples, including protein structure and materials, particularly nanomaterials. The most difficult process in TEM is sample preparation. Samples
need to be very thin (<100 nm), therefore they first need to be fixed, dehydrated, embedded into a polymer and sectioned with a sharp knife, usually made of diamond. The samples are commonly treated with heavy metals to increase the contrast. All these steps can severely damage the ultrastructure of the cell which can be partly improved by cryofixation and freeze dehydration.

In EM, there is no direct alternative for fluorescent probes, an indispensable tool in light microscopy. Alternatively, subcellular structures can be labelled with antibodies conjugated with gold nanoparticles, a technique called immunogold-EM. Nanoparticles can be identified in images as spherical structures and different diameters of nanoparticles can be used to label multiple structures. An EM image can be overlaid by an image of the same area from a light microscope which correlates fluorescence signal with ultrastructural information. This particular technique is called correlative light EM (CLEM) and has also been used in membrane trafficking studies, more recently to investigate the fission of endosomes in neurosecretory cells (Gormal et al. 2015).

Three dimensional images can be acquired with electron tomography. Multiple images of a specimen are taken at different angles and are then used to make a 3D reconstruction of the specimen. EM tomography can be correlated with light microscopy (3D-CLEM) or with immunogold labeling. An impressive 3D reconstruction of the Golgi was designed to study the role of Rab6 in trans-Golgi trafficking (Storrie et al. 2012).

1.4.3 Image analysis

The real power of imaging techniques in cell biology is revealed with the use of image analysis techniques. While the images can be visually interpreted, quantification of image and object features leverages more information, produces measurable outputs and reduces bias. Features can be measured on the entire image or in desired regions of interests or objects. The image is usually segmented so that each object (cell) is analysed separately. Several segmentation algorithms are in use today, but the most accurate ones are based on watershed segmentation (Malpica et al. 1997). Cells are segmented in a two step process. The algorithm first detects seeds of objects by identifying easily segmented objects, such as nuclei. These seeds are then expanded to cover the area of the cell. The boundaries of the cell are identified with intensity profiling. Advanced segmentation algorithms also consider shape and size of the object, resulting in better segmentation (Jones et al. 2005). Cells can be further segmented into subcellular structures. This is reasonably trivial for clearly segmented punctate structures, such as vesicles. However, few methods have been developed to effectively segment complex structures, such as the ER or membrane tubules, which would require advanced image analysis and machine learning approaches.

Basic image analysis quantification techniques are routinely used in microscopy. The simplest among them are intensity measurements, usually measured as a total (integrated) intensity or mean intensity. More advanced is the fluorescence recovery after photobleaching (FRAP) technique, which observes the dynamics of fluorescence signal recovery after photobleaching a small area of the image. A sequence of images is required for proper quantification of dynamics, so the signal needs to be corrected for photobleaching induced by imaging itself. Advanced intensity and FRAP techniques are very useful in studying dynamics of membrane trafficking. Recently FRAP has been used to investigate the role of Annexin A11 in stabilisation of Sec31A on ERES (Shibata et al.
This study showed that ratios between mobile and immobile fractions of Sec31A change after depletion of ALG-2, which mediates physical interaction between Sec31A and Annexin A11. FRAP has been used to describe the dynamics of Golgi proteins. A study by Cole et al. investigated how Golgi proteins retain their localisation amidst the constant flux of cargo through the Golgi. They observed that Golgi resident proteins are not laterally immobilised, but are transported between Golgi stacks, thereby retaining their Golgi localisation (Cole et al. 1996). The Fluorescence Protease Protection (FPP) assay is used to study the topology of proteins (Lorenz et al. 2006). In FPP, cells are expressing either N- or C-terminally GFP-tagged transmembrane proteins. Cells are semi-permeabilised with digitonin and then treated with a protease, such as trypsin or proteinase K. The GFP molecule that is facing the cytoplasm is completely digested by the enzyme, while a molecule in the lumen of the endomembrane system is protected and thus retains the fluorescence. This assay can therefore indicate which terminus of a protein is facing the cytoplasm or lumen.

Time lapse microscopy provides additional dimension to data. Videos, capturing the dynamics of the observed process, can also be quantified. Segmentation is particularly important here, because boundaries of structures have to be precisely defined in order to track them from frame to frame. In a recent study on dynamics of contacts between ER and endosomes the authors captured an endosome fission event in a time-lapse video (Rowland et al. 2014). Well segmented punctate structures can be tracked from frame to frame to profile hundreds of structures. So-called single particle tracking (SPT) analysis methods are particularly useful for tracking the cell uptake of fluorescently labelled nanoparticles. A recent publication demonstrates SPT and object-based quantification of two channels - nanoparticles and lysosomes (Varela et al. 2014).

In the above example authors measured co-localisation, which is a commonly used technique to measure the overlap between two signals. This is particularly useful in determining the localisation of a protein because some markers with precisely defined localisation can be used as localisation reference points. Co-localisation provides information on the amount of the investigated protein localising to the same structure as the marker. Several metrics of co-localisation are in use, either quantifying pixel co-occurrence or pixel intensity correlation. Pixel intensity correlation is calculated as the Pearson’s correlation between two channels in the region of interest. Most commonly used metrics of pixel co-occurrence is the Manders’ coefficient (Manders et al. 1993). A new algorithm, rank weighted coefficient (RWC), has been developed recently, which overcomes some drawbacks of the Pearson’s and Manders’ coefficients (Singan et al. 2011).

Identifying the localisation of a protein by co-localisation methods requires specific organelle markers, which is not feasible in high-throughput studies. Alternatively, texture quantification methods have been used, which extract texture features (or descriptors) of an image or object. Texture features are criteria that transform a multidimensional array of pixels into a numerical value. Commonly used are Haralick features (Haralick et al. 1973), Gabor features and threshold adjacency statistics (TAS) features. TAS features in particular have been used in automatic classification of cellular phenotypes (Hamilton et al. 2007). Object classification accuracy can be further improved by texture quantification across multiple channels (Coelho et al. 2013). Texture features have been also used for the quantification of biomarkers in immunohistological images (Kumar et al. 2014).
Recently, a new approach in bioimage informatics has become popular. There is a significant shift towards the use of unsupervised machine learning methods and structured learning. This is demonstrated in the first attempts of image-derived generative models (Buck et al. 2012) and use of morphological features to identify cell shape regulation networks (Yin et al. 2013; Sailem et al. 2014). This can be further expanded by non-imaging data, such as genomics, proteomics and metabolomics. The above studies do not rely on human input and are therefore the first attempts to programmatically interpret biological images.

1.5 Aims of the study

Protein localisation is, together with its sequence and structure, among the most basic determinants for its function. Imaging techniques can be used to experimentally determine protein localisation, however it is difficult to explain underlying localisation mechanisms. The overall aim of this study was to use computational and experimental methods to study localisation mechanisms of proteins in cultured mammalian cells. More specifically:

- to discover protein sequence determinants in form of SLiMs that are related to specific subcellular localisation and experimentally validate them;

- to experimentally determine localisation, sequence determinants, membrane topology and functional context of PRAF and YIPF family proteins;

- to develop image analysis methods to quantify protein subcellular localisation profiles.
Chapter 2

Materials and Methods
2.1 Computational discovery of novel SLiMs

2.1.1 Localisation dataset

The localisation dataset was obtained from Simpson et al. 2000. In that study the localisation of 3981 cDNA constructs corresponding to 1165 unique proteins fluorescently tagged (GFP, CFP or YFP) on either C- or N-terminus was determined by fluorescence microscopy and manual image interpretation. If different localisations were detected for the same protein, the most abundant or the most probable was selected. The general localisation classes were also arbitrarily selected. The dataset was then compiled as a table of Uniprot identifier, localisation of N-terminally tagged protein, localisation of C-terminally tagged protein, general localisation class and annotations about localisation discrepancies and membrane topology. Uniprot identifiers of proteins for each classification strategy were extracted from this spreadsheet and exported as text files, one per class, using an R script. These files were then used to download protein sequences from Uniprot (20th February 2015) in fasta format leveraging Uniprot REST service with a custom Python script.

2.1.2 SLiMFinder - SLiM discovery tool

SLiMFinder, a module of SLiMSuite, is a computational tool for the discovery of novel short linear motifs in a subset of protein sequences (Edwards et al. 2007). SLiMFinder 4.7 was used in these experiments. The software was running on a cluster utilising 20 nodes. Disorder masking was used to limit the search to disordered regions of a protein; terminal masking (by using case mask) was used to search only in N or C terminal regions. By default all methionine residues at position 1 were masked, as were low complexity regions with 5 repeated residues in a window of 8 residues. Default statistical evaluation by the SLiMCalc algorithm was used unless stated otherwise.

2.1.3 SLiMProb - Search tool for predefined SLiMs

SLiMProb is a module of SLiMSuite used to search for a particular SLiM in a sequence dataset. Novel SLiM candidates discovered by SLiMFinder were searched for in the whole dataset. The enrichment within each localisation class was then calculated and evaluated with the Fisher’s exact test.

2.2 Cloning

Sequences and cDNA clones representing all members of the PRAF and YIPF families were obtained from the I.M.A.G.E. consortium (see Appendix A for identifiers). Restriction sites were introduced into the 3’ and 5’ ends of the sequence using custom primers (Eurofins Genomics) in PCR amplification (see Appendix A for the list and sequences of primers). The components of the PCR reaction mix were: 5 ng of template DNA, 50 pmol of each primer, 1x PCR buffer (Invitrogen), 10 mM of each dNTP (Invitrogen), 1 mM MgSO4 (Invitrogen) and 0.4 µl of Platinum Pfx polymerase (2.5 U/µl, Invitrogen). The PCR was then performed in MasterCycler Pro (Eppendorf) with the following programme: 5 min incubation at 95°C, followed by 25 cycles of 30 s denaturation at 94°C,
30 s annealing at 45°C and 90 s elongation at 68°C. After the last cycle a 7 min elongation step was performed at 68°C. The PCR products were purified with a PCR purification kit (Genomed) following the manufacturer’s protocol.

‘Sticky ends’ were introduced into the amplified DNA molecules and pEGFP-C1 or pEGFP-N1 vectors with XhoI and SacII endonucleases (NEB) by incubating for 1 h at 37°C. The fragments were purified with a PCR purification kit (Genomed) following the manufacturer’s protocol and ligated into linearised plasmids with T4 DNA ligase (NEB) overnight at room temperature. Linearised plasmid without insert fragment was used as a self-ligation control in the ligation reaction. The following day 50 µl of DH5α competent cells (Invitrogen) were transformed with the ligation mix. Competent cells were first incubated with the ligation mix for 20 min on ice, then heat-shocked at 42°C for 45 s and placed on ice for 2–5 min. Next 250 µl of lysogeny broth (LB) medium was added and transformations were incubated in an orbital shaker for 2 h at 37°C. Next the cells were pelleted by centrifugation at 1000×g for 5 min, diluted in 100 µl of LB-broth medium, spread on LB-agar plates containing 50 µg/µl kanamycin at 37°C and incubated overnight. The following day 5 colonies were picked from the plate, inoculated into 4 ml of LB-broth medium containing kanamycin for 8 h and then the DNA was extracted using a Miniprep DNA isolation kit (Qiagen) following the manufacturer’s protocol. Isolated DNA was digested with XhoI and SacII endonucleases as described above followed by separation on a 1% w/v agarose gel made with Tris/Borate/EDTA (TBE) buffer and 1x SYBR Safe DNA Gel stain (Life Technologies) for DNA visualisation. The gel was run at 100 V for 45 h and then visualised on a GeneGenius (Syngene) gel documentation system by transilluminating the gel with ultraviolet light (302 nm). Successful ligation was confirmed by DNA sequencing (Source Bioscience). Sequence-verified colonies were inoculated into 100 ml of LB-broth containing kanamycin and incubated overnight and then the DNA was extracted with a Midiprep kit (Promega) following the manufacturer’s protocol.

Sequences of predicted SLiMs were synthesised as single strand DNA molecules of ~50 nucleotides in length (see Appendix A for sequences). The sequences were reconstituted in annealing buffer (10 mM Tris, pH 7.5–8.0, 50 mM NaCl, 1 mM EDTA). Equal amounts of both strands were mixed in an Eppendorf tube and heated at 95°C for 5 min in a metal block. The block was then allowed to slowly cool to room temperature (approximately 90 min). The annealed DNA was then ligated with linearised plasmid as described above. After the second digestion with endonucleases the restriction reactions were analysed by polyacrylamide gel electrophoresis, as the inserts were too small for detection on agarose gels. 5% polyacrylamide gels were prepared with the following components: 7.2 ml distilled water, 1 ml 10X TBE buffer, 1.66 ml 30% Protogel (37.5:1 acrylamide:bis-acrylamide) (National Diagnostics), 40 µl tetramethylethylenediamine (TEMED) (National Diagnostics) and 200 µl ammonium persulphate (BDH). The electrophoresis was performed for 90 min at 300 V. The DNA was stained by incubating with 1X SYBR Safe DNA gel stain in TBE buffer for 15 min and recording with a GeneGenius (Syngene) gel documentation system by transilluminating the gel with ultraviolet light (302 nm). Sequences of selected colonies were verified by DNA sequencing (Source Bioscience), inoculated into 100 ml of LB-broth containing kanamycin and incubated overnight, followed by DNA extraction using a Midiprep kit (Promega) following the manufacturer’s protocol.
2.3 Cell culture

HeLa Kyoto cells were cultured at 37°C in Dulbecco’s Modified Eagle Medium (DMEM) (Gibco) supplemented with 10% foetal calf serum (FCS) (Gibco) and 1% L-Glutamine (Gibco) in a humidified atmosphere 5% CO₂/95% air. Cells were routinely passaged 1:10 on reaching 70% confluency by detaching with 0.25% trypsin-EDTA (Gibco). After 15 passages the cells were discarded.

2.4 DNA transfection

One day before transfection HeLa Kyoto cells were seeded in a suitable cell culture dish or plate (cell numbers are listed in Table 2.1). On the day of transfection FuGENE6 transfection reagent (Promega) was diluted in OptiMEM (Gibco) medium according to the manufacturer’s protocol (reaction mix volumes are listed in Table 2.1) and incubated for 5 min at room temperature. In transfections with two different DNAs the amount of FuGENE6 reagent was doubled. The diluted reagent was mixed with the appropriate amount of DNA in a fresh tube and incubated for 15 min at room temperature. After incubation the mixture was added drop-wise to the cells, mixed and incubated for 24 h.

Table 2.1: Amounts of reagents used in DNA transfections.

<table>
<thead>
<tr>
<th>Dish type</th>
<th>Vol. of FuGENE6 per transfection</th>
<th>Vol. of OptiMEM per transfection</th>
<th>Amount of DNA per transfection</th>
<th>Number of cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 cm dish</td>
<td>27 µl</td>
<td>373 µl</td>
<td>10 µg</td>
<td>360000</td>
</tr>
<tr>
<td>35 mm dish</td>
<td>3 µl</td>
<td>100 µl</td>
<td>1 µg</td>
<td>60000</td>
</tr>
<tr>
<td>12-well plate</td>
<td>1.5 µl</td>
<td>50 µl</td>
<td>0.5 µg</td>
<td>30000</td>
</tr>
<tr>
<td>24-well plate</td>
<td>0.75 µl</td>
<td>50 µl</td>
<td>0.25 µg</td>
<td>15000</td>
</tr>
</tbody>
</table>

2.5 siRNA transfection

One day before transfection cells were seeded in a suitable cell culture dish or plate (cell numbers are listed in Table 2.2). The following day the cells were transfected with Oligofectamine (Life Technologies) transfection reagent according to the manufacturer’s protocol (reaction mix volumes are listed in Table 2.2). Two siRNAs targeting the same gene at different sites were mixed together and diluted with OptiMEM (Gibco). In a separate tube Oligofectamine was diluted 1:10 in OptiMEM and incubated for 7 min. After that the diluted Oligofectamine was added to the diluted siRNAs and incubated for 20 min. The cell medium was exchanged for FCS-free DMEM and then the transfection mixture was added drop-wise. The plate was gently swirled and incubated for 4 h. After 4 h the FCS-free DMEM medium was exchanged for DMEM medium supplemented with 10% FCS (Gibco) and 1% L-Glutamine (Gibco).
Table 2.2: Amounts of reagents used in siRNA transfections.

<table>
<thead>
<tr>
<th>Dish type</th>
<th>Vol. of Oligofectamine per transfection</th>
<th>Amount of each siRNA</th>
<th>Vol. of OptiMEM for siRNA dilution</th>
<th>Number of cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-well plate or 35 mm dish</td>
<td>2 μl + 18 μl OptiMEM</td>
<td>20 pmol</td>
<td>180 μl</td>
<td>12000</td>
</tr>
<tr>
<td>24-well plate</td>
<td>0.5 μl + 4.5 μl OptiMEM</td>
<td>5 pmol</td>
<td>45 μl</td>
<td>3000</td>
</tr>
</tbody>
</table>

2.6 Quantitative PCR (qPCR)

Efficiency of siRNA-mediated gene depletion was measured using RT-qPCR. Cells were seeded in a 6-well plate (12000 cells for 72 h transfections and 24000 cells for 48 h transfections) and transfected as described above. Total RNA was extracted 48 h or 72 h after transfection using an InviTrap spin universal RNA kit (Stratec) following the manufacturer’s protocol. Next 500 ng of total RNA was transcribed into cDNA using a High-Capacity cDNA Reverse Transcription kit (Applied Biosystems) according to the manufacturer’s protocol. The reaction was carried out in MasterCycler Pro (Eppendorf) with the following cycling program: 25°C, 10 min → 37°C, 120 min → 85°C, 5 min. Complementary DNA was then diluted with RNase-free water to the final volume 100 μl. Gene expression was then quantified in qPCR reactions using a Fast SYBRGreen PCR mastermix (Applied Biosystems). The reaction mix was composed of the PCR mastermix, 5 μl of diluted cDNA and 200 nM of each gene-specific primer. Quantitative PCR was performed in a 7500 FAST real-time PCR system (Applied Biosystems) using the following cycling program: 50°C, 2 min → 95°C, 10 min → (95°C, 15 s → 60°C, 1 min) × 40. A melting curve stage was included in every run in order to assess the synthesis of the specific PCR product and four technical replicates of each sample were analysed. The data were analysed as relative gene expression compared to that seen in cells treated with negative control (non-targeting) siRNAs using the -ΔC_T method.

2.7 Fixation and immunostaining

Cells were fixed with 3% paraformaldehyde (PFA) (Sigma) in PBS (Sigma) at 37°C with 0.2 μg/ml Hoechst33342 (Sigma) for 20 min. Cells were permeabilised with 0.1% Triton-X100 (Sigma) in PBS for 5 min. Then the cells were stained with antibodies at dilutions listed in Table 2.3. Coverslips were incubated with primary antibodies for 30 min, followed by 3 washes with PBS and then incubated with secondary antibodies for 30 min. They were again washed 3-times and mounted on to glass slides with Mowiol.

Cells stained with ER markers REEP5 and CLIMP-63 were fixed with methanol at -20°C for 4 min. In this case additional permeabilisation was not needed and cells were stained and mounted as described above.

2.8 Co-immunoprecipitation

Hela Kyoto cells were cultured in 35 mm or 10 cm cell culture dishes and transfected with GFP-tagged DNA constructs as described above. After 24 h the cells were washed with ice-cold PBS,
CHAPTER 2. MATERIALS AND METHODS

Table 2.3: Antibodies used in immunofluorescence staining and corresponding dilutions used.

<table>
<thead>
<tr>
<th>Antibody</th>
<th>Dilution in PBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rabbit α REEP5 (Proteintech)</td>
<td>1:100</td>
</tr>
<tr>
<td>Mouse α CLIMP-63 (Enzo)</td>
<td>1:1000</td>
</tr>
<tr>
<td>Mouse α GM130 (BD Biosciences)</td>
<td>1:500</td>
</tr>
<tr>
<td>Mouse α HA (Abcam)</td>
<td>1:1000</td>
</tr>
<tr>
<td>Goat α mouse Alexa 488 (Life Technologies)</td>
<td>1:400</td>
</tr>
<tr>
<td>Goat α rabbit Alexa 488 (Life Technologies)</td>
<td>1:400</td>
</tr>
<tr>
<td>Goat α mouse Alexa 568 (Life Technologies)</td>
<td>1:400</td>
</tr>
<tr>
<td>Goat α rabbit Alexa 647 (Life Technologies)</td>
<td>1:400</td>
</tr>
<tr>
<td>Goat α mouse Alexa 568 (Life Technologies)</td>
<td>1:400</td>
</tr>
<tr>
<td>Goat α YIPF antibodies (from Nobuhiro Nakamura)</td>
<td>1:100</td>
</tr>
</tbody>
</table>

detached using cell scrapers, pelleted and frozen at -80°C. Pellets were thawed on ice and lysed with 1 ml of LAP buffer (50 mM HEPES, 5 mM EDTA, 175 mM KCl, 10% glycerol) containing 1% NP-40 and protease inhibitors cocktail (Roche) by incubating on ice with gentle rocking for 30 min. Cell lysates were then centrifuged at 16000 × g, 4°C for 20 min. Supernatants were collected and pre-cleared with 50 µl of BSA (0.1%) blocked Sepharose beads (Sigma) in LAP buffer (1:1) for 4 h. Beads were removed by centrifugation at 10000 × g for 2 min. Pre-cleared cell lysates were incubated with 1 µg of anti-GFP (Roche) antibody overnight at 4°C. The next day 50 µl of BSA blocked beads were added to the lysates and incubated at 4°C for 4 h. Then the beads were pelleted by centrifugation at 10000 × g for 2 min and washed 3-times with lysis buffer, followed by two washes with LAP buffer without detergent. The beads were stored at -80°C prior to analysis by mass spectrometry.

Sepharose beads conjugated with Protein G (Sigma) were prepared by removing the supernatant, followed by 3 washes with LAP buffer. Beads were then incubated with 0.1% BSA (NEB) in LAP buffer for 1 h at 4°C. After that the supernatant was removed and beads were resuspended in LAP buffer at a dilution of 1:1.

2.9 Mass spectrometry and data analysis

Proteins co-precipitating with GFP-tagged baits were identified and quantified using mass spectrometry. Beads prepared by the co-immunoprecipitation protocol described above were thawed on ice and then resuspended in 100 µl of 8 M urea, followed by addition of 100 µl 200 mM ammonium bicarbonate, pH 8.5 and 2 mM dithiothreitol (DTT). The sample was incubated at room temperature for 15 min. Then 500 mM iodoacetamide solution was added to give a final concentration of 4 mM and incubated in the dark for 15 min. The bead suspension was then diluted with 200 mM ammonium bicarbonate, pH 8.5, to dilute the urea to <2 M. The proteins were digested with 1 µg of trypsin at 37°C overnight. The beads were then centrifuged at 10000 × g for 3 min and the supernatant was transferred into a new tube. The beads were washed once with 0.1% trifluoroacetic acid (TFA), centrifuged as before and the supernatant was combined with the sample (first supernatant). The sample was then acidified with 100% TFA to pH ~2.0. The samples were then cleaned using the ‘ZipTip’ technique. Each tip was first wetted with 100% acetonitrile.
and then equilibrated with 0.1% TFA. Then the peptides were loaded into the column by slowly pipetting the sample up and down 20 times. Bound peptides were then washed with 0.1% TFA 3 times. Then the peptides were eluted 3 times into a new tube with 60% acetonitrile in 0.1% TFA. The samples were then analysed on a Q-Exactive Hybrid Quadrupole-Orbitrap mass spectrometer (Thermo Scientific).

Raw spectral data were processed, analysed and statistically evaluated with MaxQuant software (version 1.5.2.8) (Cox et al. 2008). Relative levels of proteins were quantified with the MaxLFQ algorithm, which is based on raw spectral intensities and applies several levels of normalisation (Cox et al. 2014). An Uniprot database of reviewed human proteins was used, acquired on 18th February 2015.

2.10 Western blotting

Protein lysates were prepared as described above. First the amounts of proteins in each sample were quantified using the bicinchoninic acid (BCA) protein assay kit (Thermo Scientific). Then, 25 µg of protein lysates were boiled with loading buffer (100 mM Tris-Cl (pH 6.8) with 4% w/v sodium dodecyl sulphate, 0.2% w/v bromophenol blue and 20% v/v glycerol) and 200 mM DTT reducing reagent for 10 min. Then the proteins were separated on a 12% acrylamide gel (National Diagnostics) and transferred to a polyvinylidene fluoride (PVDF) membrane (Perkin Elmer). The membrane was first activated in 100% ethanol, followed by incubation in distilled water. Then, four sheets of filter paper were soaked in transfer buffer (48 mM Tris, 39 mM glycine, 0.0375% sodium dodecyl sulphate and 20% ethanol) and assembled into a 2x filter paper – PVDF membrane – gel – 2x filter paper stack and a semi-dry transfer was performed at 2.5 mA/cm² for 45 min. The PVDF membranes were then blocked with skimmed milk for 30 min and probed with primary antibodies (see Table 2.4 for dilutions) overnight at 4°C. The next day the membranes were washed in TTBS (Tris-buffered saline (TBS) supplemented with 0.1% polysorbate 20) and incubated with alkaline phosphatase conjugated secondary antibodies (see Table 2.4) for 1 h at room temperature. The membranes were washed 5-times with TTBS, 2-times with TBS and incubated with Attophos reagent (Promega) in the dark for 5 min. The membranes were imaged for 10 s with a LAS 4000 (Fujifilm) gel documenting system.

Table 2.4: Antibodies and corresponding dilutions used in western blotting.

<table>
<thead>
<tr>
<th>Antibody</th>
<th>Dilution</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse α GFP (Roche)</td>
<td>1:1000</td>
</tr>
<tr>
<td>goat α mouse conjugated with alkaline phosphatase (Sigma)</td>
<td>1:10000</td>
</tr>
</tbody>
</table>

2.11 Fluorescence Protease Protection (FPP) assay

HeLa Kyoto cells were transfected with N- or C-terminally tagged GFP constructs in 35 mm live cell culture dishes as described above. After 24 h the medium was exchanged for KHM buffer (110 mM potassium acetate, 20 mM HEPES, 2 mM magnesium chloride; pH 7.4) pre-warmed at
37°C. During the following steps the fluorescence of GFP was imaged with a Leica DMI6000B fluorescence microscope with HCX PL APO 63x/1.40 oil immersion objective in 5 s intervals. The cells were first semi-permeabilised with 20 µM digitonin (Sigma) in KHM buffer for 45 s and then 8 M trypsin (Sigma) in KHM buffer was added. Cells were manually segmented and intensity levels analysed using ImageJ. The decrease in fluorescence after the addition of trypsin was plotted. The fluorescence levels of the final time points were normalised to the fluorescence levels before the addition of trypsin. The protocol was adapted from Lorenz et al. 2006.

2.12 Confocal imaging

Slides with fixed cells transfected with GFP constructs were imaged on a FluoView FV1000 laser scanning microscope (Olympus), equipped with motorised stage, six laser lines and five photomultipliers. A combination of argon gas laser and laser diodes provided excitation wavelengths of 405 nm, 458 nm, 488 nm, 515 nm, 559 nm and 635 nm (see Table 2.5 for the combinations of fluorophores and excitation lasers). Images were acquired with a 60x UPLSAPO 1.35NA (Olympus) oil immersion objective at a resolution of 1024x1024 pixels in sequential scanning model. The pixel dwell time was 12.5 µs and images were processed with the Kalman filter. The images were saved in Original Imaging Format (OIF), which includes greyscale TIFF image files.

Table 2.5: List of fluorophores used with corresponding excitation lasers and emission filters on an FV1000 laser scanning confocal microscope.

<table>
<thead>
<tr>
<th>Fluorophore</th>
<th>Excitation laser</th>
<th>Emission range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hoechst33342</td>
<td>405 nm</td>
<td>425–475 nm</td>
</tr>
<tr>
<td>EGFP, Alexa488</td>
<td>488 nm</td>
<td>500–545 nm</td>
</tr>
<tr>
<td>mCherry, Alexa568</td>
<td>559 nm</td>
<td>575–620 nm</td>
</tr>
<tr>
<td>Alexa647</td>
<td>635 nm</td>
<td>655–755 nm</td>
</tr>
</tbody>
</table>

2.13 Epifluorescence microscopy

Live or fixed cells were imaged on a Leica DMI6000B wide-field microscope. The microscope was equipped with a motorised stage, metal-halide light source, a DFC340FX CCD camera (pixel size 4.4 x 4.4 µm) and five dichroic excitation filters (see Table 2.6). Images were acquired using a HCX PL APO 63x/1.40 oil immersion objective at a resolution of 1600x1200 pixels and exported as 16-bit greyscale TIFF files. The microscope was controlled by Leica Application Suite software.

Table 2.6: List of filters on the Leica DMI6000B system.

<table>
<thead>
<tr>
<th>Dichroic filter name</th>
<th>Excitation filter (nm)</th>
<th>Emission range (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L5</td>
<td>480/40</td>
<td>527/30</td>
</tr>
<tr>
<td>CFP/YFP</td>
<td>CFP: 436/12; YFP: 500/20</td>
<td>CFP: 467/37; YFP: 545/45</td>
</tr>
<tr>
<td>CY5</td>
<td>635/10</td>
<td>720/60</td>
</tr>
<tr>
<td>CY3</td>
<td>560/10</td>
<td>610/65</td>
</tr>
<tr>
<td>DAPI/FITC</td>
<td>DAPI: 410/15; FITC: 490/20</td>
<td>DAPI: 455/20; FITC: 535/45</td>
</tr>
</tbody>
</table>
2.14 Image analysis

2.14.1 Co-localisation with RWC

Co-localisation between two fluorophores was calculated as a Rank Weighted Coefficient (RWC) value (Singan et al. 2011), and analysed using CellProfiler 2.1.1 software (Carpenter et al. 2006). First, the nuclei were identified as primary objects and then the cells were segmented as secondary objects. Only transfected cells were selected by applying a filter based on GFP intensity. A custom CellProfiler plugin was developed to calculate the RWC values between two channels. The RWC values were then evaluated by t-statistics with false discovery rate (FDR) correction for multiple testing (Benjamini et al. 1995). The custom plugin was developed in the Python programming language according to CellProfiler specifications. Statistical analysis scripts were developed in the R programming language.

2.14.2 Localisation with texture features

Images of GFP-tagged constructs were analysed with CellProfiler 2.1.1 (Carpenter et al. 2006). Image segmentation was performed as described above followed by extraction of Haralick and Gabor texture features using built-in analysis algorithms. A custom plugin was developed to extract Threshold Adjacency Statistics (TAS) features (Hamilton et al. 2007).

A custom plugin was developed in the Python programming language according to CellProfiler specifications. Statistical analysis scripts were developed in the R programming language. The source code of the CellProfiler module is provided in Appendix B.
Chapter 3
Discovery of novel SLiMs involved in protein localisation
3.1 Introduction

Protein targeting mechanisms make use of protein-protein interactions to deliver proteins to their site of function. Two proteins can interact with each other through domains, and these interactions are usually strong and specific. Domains also interact with short linear motifs (SLiMs), which are linear peptides, 3–12 amino acids long, usually placed in the unstructured regions of proteins (Fuxreiter et al. 2007; Russell et al. 2008). Only a few residues of a SLIM typically have fixed amino acids; many positions are ambiguous; hence SLiMs are evolutionarily labile and can evolve de novo through a small number of mutations (Neduva et al. 2005a). Known and validated SLiMs are collected in the Eukaryotic Linear Motifs (ELM) (Dinkel et al. 2014) database and are categorised based on the function that they are related to. One such category are targeting SLiMs, which include the well known SLiMs facilitating localisation, such as the COPI retrograde trafficking motifs KDEL (Lewis et al. 1990), K.[0,1]K..$ (Jackson et al. 1990) and RR (Schutze et al. 1994); the nuclear localisation signal KRK (Dingwall et al. 1982); and more complicated targeting SLiMs, such as the nuclear export signals (Fischer et al. 1995; Wen et al. 1995).

SLiMs as localisation signals are convenient because of their low footprint. Despite their low specificity and binding affinity they are able to control the localisation of proteins. When fused with fluorescent proteins they can be used to monitor localisation mechanisms in a cell. Particularly interesting are KDEL (Lewis et al. 1990) and Tat (Frankel et al. 1988; Green et al. 1988) sequences which can be fused with a fluorescent protein and target it to the ER or allow its penetration across the membrane, respectively (Ma et al. 2009). In both cases a short peptide attached to a fluorescent protein is able to target the fusion protein to a certain location.

It is estimated that only 5% of all SLiMs have been discovered. Modern discovery methods are computational and usually search for overrepresented regions in a group of protein sequences. The discovered SLiMs, also called candidate SLiMs, are then evaluated for their unexpectedness (Kelil et al. 2014). This type of a search algorithm is implemented in the SLiMFinder software, which is currently the best performing software for SLiM discovery (Edwards et al. 2007; Kelil et al. 2014). SLiMFinder first groups the proteins into Unrelated Protein Clusters (UPCs) in such a way that proteins in each UPC are not more than 70% identical to any other protein outside of its UPC. Then the sequences are masked to limit the search only to certain parts of the sequences, such as disordered regions, terminal regions or structural features. The masked sequences are then searched for overrepresented motifs by the SLiMBuild algorithm and the resulting candidate SLiMs are statistically evaluated by the SLiMChance algorithm. SLiMChance calculates the probability of finding a certain motif in an UPC based on the frequency of each amino acid in a dataset and the motif length. These probabilities are then corrected for: protein ‘unrelatedness’ within each UPC; search space limitations because of masking and the fact that these motifs are unknown (for details see Edwards et al. 2007). SLiMFinder by default assumes that nothing is known about the discovered SLiMs, however it is possible to provide prior information to restrict the search.

The ELM database currently holds 23 targeting motifs, which is a small number in comparison to the other classes. Novel targeting motifs could be discovered by searching for overrepresented regions in sequences grouped by localisation. Localisation datasets, such as the localisation screen
of GFP-tagged proteins (Simpson et al. 2000), can be used as the ground truth localisation data. In the above localisation screen authors tagged 107 open reading frames on either N- or C-termini with DNA encoding fluorescent proteins CFP or YFP, respectively. The fusion constructs were then transfected into mammalian cells and imaged with a fluorescence microscope. The subcellular localisation was determined for the N- and C-terminally tagged constructs of the proteins, potentially providing information about the effect of the fluorescent protein on the localisation. The fused fluorescent proteins could interrupt interactions with nearby terminal regions of the tagged protein, in particular, they can affect the unstructured nature of the terminal regions (Lotti et al. 2011). Although the initial localisation screen assessed only a small part of human proteome it has been since expanded to over 1000 proteins (Stadler et al. 2013; Mehrle et al. 2006; Bannasch et al. 2004).

The aim of this study was to predict and experimentally validate novel localisation SLiMs by using the experimentally determined localisation data of 1165 proteins. The dataset of GFP-tagged proteins was acquired from Simpson et al. 2000 and subsequent studies (Stadler et al. 2013). The proteins were grouped into the localisation classes and SLIMFinder was used to discover novel SLiMs specific for the localisation classes. The search was performed with different masking strategies to focus on the disordered or terminal regions. Finally, one SLiM candidate was experimentally evaluated.

3.2 Results

3.2.1 Compilation of the protein localisation dataset

The authors of the localisation study (Simpson et al. 2000; Stadler et al. 2013) provided a spreadsheet with the original localisation data of 107 proteins and the additional localisation data acquired in subsequent studies. The localisation of the proteins fused with a fluorescent protein on the N- or C-termini was recorded in replicates and for different source clones, resulting in 3981 localisation records for 1165 unique proteins. A comprehensive database of the proteins with all of the localisation records was first assembled. The consensus localisation for each construct and the final localisation class were determined (Figure 3.1a) and then duplicate records were removed: duplicate records with the same localisation were merged; otherwise a single record was retained depending on the quality and the consistency of the localisation data; and in the case of transmembrane proteins or proteins with signal peptides a record showing localisation to the endomembrane system was retained. This resulted in 14 localisation classes and 3 additional classes of the proteins: those that did not express, those with undetermined localisation and those that formed unknown aggregates. The localisation classes with less than 10 proteins were not considered for further analysis, resulting in 13 localisation classes for subsequent analysis (Figure 3.2).

The localisation of 328 proteins was found to differ depending on whether they were tagged at the N- or C-terminus. This could be a consequence of the interference between the fluorescent tag and the terminal region. The proteins were thus classified into three groups: the N group, where the localisation of proteins changed after they were tagged on the N-terminus; the C group, where the localisation of proteins changed after they were tagged on the C-terminus; and the O
CHAPTER 3. DISCOVERY OF NOVEL SLiMS INVOLVED IN PROTEIN LOCALISATION

Figure 3.1: The dataset workflow. The protein sequences were first classified based on the localisation (a). The localisation classes were then further classified based on whether the localisation was affected by the fluorescent protein (b) or based on topology (c).

group, where the localisations of the N- and C-terminally tagged proteins were the same. (Figures 3.1b and 3.2a). The highest number of proteins in the N and C groups were the endomembrane- and mitochondria-localising proteins (Figure 3.2a). Many endomembrane proteins typically have an N-terminal signal peptide and discovery of SLiMs in this region would be irrelevant, because signal peptides are cleaved off before the protein is fully translated and becomes functional. Mitochondrial proteins also have an N-terminal transit peptide that is cleaved off. Proteins with an N-terminal signal or transit peptide were excluded from groups C, N or 0 using the annotations from Uniprot. 175 proteins were therefore removed from the above groups, leaving 224 proteins that were affected by the fluorescent protein tag (Figure 3.2b). The groups containing 10 or more proteins were considered for further analysis, which resulted in 9 classes (Figure 3.2b).

The attachment of a fluorescent protein had the greatest effect on proteins that localised to the mitochondria (68%), plasma membrane (65%) and nuclear envelope (64%). Approximately 6%
**Figure 3.2:** Proteins that changed localisation when fused with a fluorescent protein. a) Counts of all proteins in the dataset that changed the localisation when fused to a fluorescent protein. Group 0 are the proteins that localised the same regardless on which termini they were fused with a fluorescent protein; group C are the proteins that changed localisation when a fluorescent protein was fused to the C-terminus; group N are the proteins that changed localisation when a fluorescent protein was fused to the N-terminus. The numbers in the bars show the counts of proteins in each group and the numbers at the top of the bars show the total counts in each localisation class. b) Same as a) but without the proteins with an N-terminal signal or transit peptide.
### Table 3.1: Statistics of the dataset with excluded proteins with N-terminal signal or transit peptides.

<table>
<thead>
<tr>
<th>Localisation class</th>
<th>Proportion</th>
<th>Odds ratio</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centrosome</td>
<td>36%</td>
<td>1.90</td>
<td>0.33</td>
</tr>
<tr>
<td>Cytoplasm</td>
<td>6%</td>
<td>0.24</td>
<td>0.00</td>
</tr>
<tr>
<td>Cytoplasm &amp; Nucleus</td>
<td>3%</td>
<td>0.09</td>
<td>0.00</td>
</tr>
<tr>
<td>Cytoskeleton</td>
<td>41%</td>
<td>2.39</td>
<td>0.00</td>
</tr>
<tr>
<td>Endosomes / lysosomes</td>
<td>52%</td>
<td>3.64</td>
<td>0.00</td>
</tr>
<tr>
<td>ER</td>
<td>25%</td>
<td>1.12</td>
<td>0.68</td>
</tr>
<tr>
<td>ER &amp; Golgi</td>
<td>61%</td>
<td>5.36</td>
<td>0.00</td>
</tr>
<tr>
<td>Golgi</td>
<td>33%</td>
<td>1.71</td>
<td>0.20</td>
</tr>
<tr>
<td>Mitochondria</td>
<td>68%</td>
<td>7.08</td>
<td>0.00</td>
</tr>
<tr>
<td>Nuclear envelope</td>
<td>64%</td>
<td>6.14</td>
<td>0.00</td>
</tr>
<tr>
<td>Nucleolus</td>
<td>26%</td>
<td>1.21</td>
<td>0.59</td>
</tr>
<tr>
<td>Nucleus</td>
<td>22%</td>
<td>0.95</td>
<td>0.84</td>
</tr>
<tr>
<td>Peroxisomes</td>
<td>100%</td>
<td>Infinite</td>
<td>0.01</td>
</tr>
<tr>
<td>Plasma membrane</td>
<td>65%</td>
<td>6.42</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Whole dataset</strong></td>
<td><strong>23%</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The proportion is calculated between the number of proteins whose localisation was affected by a fluorescent protein and the number of all proteins in the localisation class. The odds ratio was calculated between the odds of the proteins whose localisation was affected by the tag and the proteins that were not affected by the tag in each localisation class versus the odds of the whole dataset. The odds ratio p-values were calculated using the Fisher’s exact test.

of the cytoplasmic proteins were affected by the attachment of the GFP, while the other classes had at least 20% of the proteins affected (Table 3.1). The counts of proteins with altered localisation between the N- and C-terminal tags were compared with the ratio of the whole dataset, which was 23%. The odds ratios were also compared using the Fisher’s exact test. The localisation classes cytoplasm, cytoplasm and nucleus, and nucleus were under-enriched, whereas the mitochondrial proteins showed a high 7-fold enrichment over the whole dataset. Interestingly, the ER or Golgi proteins individually did not show any significant enrichment of the proteins that were affected by the tag, while 61% of the proteins that localised both to the ER and the Golgi were affected.

Proteins that interact with membranes can have 3 topologies: cytosolic, luminal or transmembrane. Transmembrane proteins can be identified by the presence of transmembrane regions, luminal proteins have a signal peptide (and no transmembrane domains) and cytosolic proteins have neither of these (Figure 3.3). Transmembrane proteins and signal peptides can be predicted by hydrophobicity plots (Kyte et al. 1982) or by using advanced predictors, such as TMHMM (Krogh et al. 2001). They are also annotated in the Uniprot database based on experimental data or predictions. Uniprot annotations were then used to classify the proteins from the dataset into 3 classes: cytosolic, luminal and transmembrane (Figure 3.3c). The groups of proteins, containing 10 or more proteins, and localising to the endomembrane system and mitochondria were considered for further analysis, yielding 13 classes (Figure 3.4).
**CHAPTER 3. DISCOVERY OF NOVEL SLIMS INVOLVED IN PROTEIN LOCALISATION**

**Figure 3.3:** Features used to classify the proteins based on their membrane topology. Grey boxes represent the whole protein sequence. Membrane proteins have transmembrane domains (TMD; orange box) and can have a signal peptide (SP; green box). Luminal proteins have an N-terminal signal peptide (SP; green box) and no transmembrane domains. Cytosolic proteins have neither transmembrane domains nor a signal peptide.

**Figure 3.4:** Counts of the proteins in the three topology classes. The proteins were classified based on their topology. Red bars represent the cytosolic proteins, green bars represent the luminal proteins and blue bars represent the transmembrane proteins. The numbers in the bar represent the counts of proteins in each topology class. The numbers at the top of the bars are the total counts in each localisation class.

The proteins from the dataset were categorised into 14 localisation classes and 3 ambiguous classes (unknown localisation, no expression, aggregates) (Figures 3.2 and 3.4). The localisation of 90 proteins changed when C-terminally tagged, 134 changed when N-terminally tagged and 766 were not affected by the tag (Figure 3.2b). Most affected were the mitochondria localising proteins (27 out of 40), followed by the plasma membrane localising proteins (32 out of 49) and the endosomes/lysosomes localising proteins (16 out of 31). An N-terminal tag affected the local-
CHAPTER 3. DISCOVERY OF NOVEL SLiMS INVOLVED IN PROTEIN LOCALISATION

Figure 3.5: The masking strategies used in the SLiM discovery. Dark grey boxes show the parts of sequences that were not masked and which were used in the SLiM discovery; light grey boxes show the masked regions. The masking of the internal sequence revealed only the terminal regions. The disorder masking revealed the regions of proteins with the predicted level of disorder >0.2.

...isation more than a C-terminal tag. Topology-wise, 849 proteins were predicted as cytosolic; 233 were transmembrane and 83 luminal (Figure 3.4). According to Uniprot annotation, most of the transmembrane proteins were in ER group (89 out of 148), followed by the plasma membrane group (31 out of 62) and the endo-lysosomal group (17 out of 36). As expected, the transmembrane and luminal proteins were mostly found in the proteins seen to localise to the endomembrane system and the mitochondria. Interestingly, several cytoplasmically-localising proteins had predicted transmembrane domains or signal peptides, which could be an artefact of overexpression or a prediction error. Such proteins were retained in the original localisation class.

3.2.2 Discovery of SLiMs with SLiMFinder

The strategy to discover novel localisation determining SLiMs was to first group the sequences into biologically relevant groups based on the localisation, the membrane topology or the interference of the fluorescent protein tag with a localisation signal, as described in the previous section. Next, a search for overrepresented short regions in the sequences from the same group of proteins was performed computationally using the SLiMFinder software. Three different sequence masking strategies were used to restrict the search space and increase the significance of the discovered motifs (Figure 3.5). In the first round the sequences were masked with the default settings, which includes masking of the N-terminal methionine and low complexity regions (regions with 5 identical amino acids within a window of 8 residues). In the next round, non-disordered regions were additionally masked out by using the disorder predictor IUPred (Dosztányi et al. 2005) with the default disorder cut-off (0.2). Alternatively, for the proteins whose localisation was affected by the fluorescent protein, the default masking options with the additional masking of the internal protein sequences were used, exposing only 30 residues at the N- and C-termini. This limited the search space to the terminal regions, which is where the interference with the fluorescent protein would be expected.

The candidate SLiMs were evaluated with the built-in SLiMChance algorithm which calculates the probability that each individual SLiM occurs by chance. The probability is expressed as a Sig-value, which is a p-value, corrected for the size of the search space. The number of overrepresented regions is typically very large, even in a small dataset and the SLiMFinder output is limited to 5000 SLiMs per dataset. Therefore the output was restricted by setting the cut-off Sig value to
CHAPTER 3. DISCOVERY OF NOVEL SLIMS INVOLVED IN PROTEIN LOCALISATION

0.10, which yielded a small number of candidate SLiMs for each localisation class. The significance values of most of the discovered SLiMs were above the consensus p-value threshold of 0.05. The consensus threshold p-value was corrected for multiple testing by dividing 0.05 by the number of separate localisation groups tested.

The SLiMChance algorithm evaluates the overrepresentation of a SLiM in a group of sequences, but it cannot consider underrepresentation or exclusiveness of a motif in a certain group. Therefore the enrichment of discovered SLiMs was calculated for each class over the rest of the dataset. For example, the motif F..M.AE was discovered in 5 sequences of the Golgi localising proteins and in 7 sequences from the dataset without Golgi proteins. The enrichment of the motif in the Golgi dataset over non-Golgi dataset was 10-fold. This approach was used to annotate the SLiM candidates discovered in the SLiMFinder trials. The SLiM candidates were searched for in all localisation datasets and negative datasets. The negative dataset of a localisation class is a dataset with all of the proteins from the localisation screen excluding the proteins from that particular localisation class. The counts of UPCs with a candidate motif in the localisation dataset and the negative dataset were used to calculate the enrichment of a motif:

\[
\text{enrichment} = \frac{\text{count in positive dataset} \times \text{total UPCs in negative dataset}}{\text{count in negative dataset} \times \text{total UPCs in positive dataset}}
\]

The counts were also used to statistically evaluate the enrichment score using the Fisher’s exact test. Table 3.2 shows an example of a contingency table. P-values were corrected for multiple testing with the Benjamini and Hochberg method (Benjamini et al. 1995). The enrichment value can be used to determine whether a SLiM is not enriched in the particular localisation class, rather than whether it is enriched in the class, because the enrichment can be biased towards the class in which the SLiM was discovered.

Table 3.2: Example contingency table for putative Golgi SLiM F..M.AE

<table>
<thead>
<tr>
<th>UPCs with SLiM</th>
<th>Positive set</th>
<th>Negative set</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPCs without SLiM</td>
<td>5</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>UPCs without SLiM</td>
<td>36</td>
<td>582</td>
<td>618</td>
</tr>
<tr>
<td>Total</td>
<td>41</td>
<td>589</td>
<td>630</td>
</tr>
</tbody>
</table>

The SLiM discovery with the default masking settings yielded 16 motifs with an uncorrected Sig-value <0.10 (Table 3.3). Variants of C..C and [KR][KR][KR][KR], representing zinc finger motifs and nuclear localisation motifs, were found among the nucleus-localising proteins with the enrichment between 2 and 3. A variant of zinc finger motif was also found among the ER localising proteins, as well as the motif P[DE][DE], which resembles TRG_ER_FFAT_. The motif P[DE][DE] was not enriched in the ER class and is probably a common motif in the whole dataset. The motif F..M.AE had a relatively low Sig-value and was found among the Golgi-localising proteins, enriched 10-fold. Two similar motifs, [IV]..L.R..E and [ILM]..L.R, were discovered in the proteins localising to the nuclear envelope. The motif [ILM]..L.R is similar to a nuclear receptor box motif (LIG_NRBOX) typically found in proteins interacting with nuclear receptors. The motifs discovered in the cytoskeleton-localising proteins were interesting for two reasons: E.[IM]..[ILM].K has two
The motifs discovered in the localisation dataset.

<table>
<thead>
<tr>
<th>Class</th>
<th>Sig</th>
<th>SLiM</th>
<th>Characterisation</th>
<th>Masking</th>
<th>Enrichment</th>
<th>P-value</th>
<th>UPCs with SLiM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleus</td>
<td>3.35E-05</td>
<td>C.CG</td>
<td>Zinc finger motif</td>
<td>Default</td>
<td>3.04</td>
<td>0.04</td>
<td>12</td>
</tr>
<tr>
<td>Nucleus</td>
<td>3.94E-05</td>
<td>KRR</td>
<td>Nuclear localisation motif</td>
<td>Default</td>
<td>3.00</td>
<td>0.0002</td>
<td>31</td>
</tr>
<tr>
<td>Nucleus</td>
<td>8.97E-05</td>
<td>KRK</td>
<td>Nuclear localisation motif</td>
<td>Default</td>
<td>2.16</td>
<td>0.01</td>
<td>27</td>
</tr>
<tr>
<td>Nucleus</td>
<td>8.58E-04</td>
<td>E.LK</td>
<td>Nuclear localisation motif</td>
<td>Default</td>
<td>1.07</td>
<td>1.00</td>
<td>29</td>
</tr>
<tr>
<td>Nucleus</td>
<td>0.001</td>
<td>[KR][KR][KR][KR]</td>
<td>Nuclear localisation motif</td>
<td>Default</td>
<td>2.00</td>
<td>0.01</td>
<td>32</td>
</tr>
<tr>
<td>Nucleus</td>
<td>0.005</td>
<td>C.C[GS]</td>
<td>Zinc finger motif</td>
<td>Default</td>
<td>2.30</td>
<td>0.14</td>
<td>14</td>
</tr>
<tr>
<td>Nucleus</td>
<td>0.023</td>
<td>T.[0,1]C.C</td>
<td>Zinc finger motif</td>
<td>Default</td>
<td>2.21</td>
<td>0.27</td>
<td>12</td>
</tr>
<tr>
<td>ER</td>
<td>0.028</td>
<td>CP.CR</td>
<td>Similar to zinc finger motif</td>
<td>Default</td>
<td>5.67</td>
<td>0.095</td>
<td>6</td>
</tr>
<tr>
<td>ER</td>
<td>0.054</td>
<td>P[DE][DE]</td>
<td>Similar to TRG_ER_FFAT_1</td>
<td>Default</td>
<td>1.02</td>
<td>1.00</td>
<td>44</td>
</tr>
<tr>
<td>Mitochondria</td>
<td>0.042</td>
<td>RI.Y[0,1]F</td>
<td>Nuclear localisation motif</td>
<td>Default</td>
<td>17.0</td>
<td>0.004</td>
<td>6</td>
</tr>
<tr>
<td>Mitochondria</td>
<td>0.066</td>
<td>F.[FY]DT</td>
<td>Nuclear localisation motif</td>
<td>Default</td>
<td>12.8</td>
<td>0.006</td>
<td>6</td>
</tr>
<tr>
<td>Cytoskeleton</td>
<td>0.061</td>
<td>E.[IM],[ILM].K</td>
<td>Nuclear localisation motif</td>
<td>Default</td>
<td>7.58</td>
<td>0.0003</td>
<td>10</td>
</tr>
<tr>
<td>Cytoskeleton</td>
<td>0.088</td>
<td>C.S.LH.Y</td>
<td>Nuclear localisation motif</td>
<td>Default</td>
<td>37.9</td>
<td>0.11</td>
<td>2</td>
</tr>
<tr>
<td>Nuclear envelope</td>
<td>0.052</td>
<td>[IV].[L.R.E]</td>
<td>Nuclear localisation motif</td>
<td>Default</td>
<td>15.6</td>
<td>0.0003</td>
<td>6</td>
</tr>
<tr>
<td>Golgi</td>
<td>0.026</td>
<td>KE.K</td>
<td>Nuclear localisation motif</td>
<td>Default</td>
<td>1.73</td>
<td>0.36</td>
<td>13</td>
</tr>
<tr>
<td>Golgi</td>
<td>0.006</td>
<td>F.M.AE</td>
<td>Nuclear localisation motif</td>
<td>Default</td>
<td>10.26</td>
<td>0.013</td>
<td>5</td>
</tr>
<tr>
<td>Nucleolus</td>
<td>0.01</td>
<td>C.[IM].C</td>
<td>Nuclear localisation motif</td>
<td>Default</td>
<td>9.81</td>
<td>0.0006</td>
<td>8</td>
</tr>
<tr>
<td>Nucleus</td>
<td>0.057</td>
<td>QR.[0,2]R.R</td>
<td>Similar to MOD_PKA_1 and CLV_PCSK motifs</td>
<td>Disorder</td>
<td>5.64</td>
<td>0.001</td>
<td>13</td>
</tr>
<tr>
<td>Nucleus</td>
<td>7.79E-06</td>
<td>KRK</td>
<td>Nuclear localisation motif</td>
<td>Disorder</td>
<td>2.16</td>
<td>0.01</td>
<td>27</td>
</tr>
<tr>
<td>Nucleus</td>
<td>3.79E-06</td>
<td>KRK</td>
<td>Nuclear localisation motif</td>
<td>Disorder</td>
<td>3.00</td>
<td>0.0002</td>
<td>31</td>
</tr>
<tr>
<td>Cytoplasm</td>
<td>0.009</td>
<td>[FLM].QN</td>
<td>Nuclear localisation motif</td>
<td>Disorder</td>
<td>1.92</td>
<td>0.012</td>
<td>14</td>
</tr>
<tr>
<td>Cytoplasm and nucleus</td>
<td>0.08</td>
<td>C.F.K</td>
<td>Similar to zinc finger motif</td>
<td>Disorder</td>
<td>1.48</td>
<td>1.00</td>
<td>7</td>
</tr>
<tr>
<td>Nuclear envelope</td>
<td>0.096</td>
<td>[ILM].L.R</td>
<td>Similar to LIG_NRBOX</td>
<td>Disorder</td>
<td>2.36</td>
<td>0.005</td>
<td>12</td>
</tr>
</tbody>
</table>

The significance values were calculated with the SLiMChance algorithm. The threshold Sig-value after Bonferroni correction was 0.004 (=0.05/13 localisation classes). The enrichment was calculated as a ratio between the proportions of SLiMs in the positive versus the negative dataset. P-value represents enrichment p-value and is calculated with Fisher’s exact test.
The interaction networks of the proteins with SLiM instances. The interaction network of the instances of proteins with a) the cytoskeletal motifs E.[IM]..[ILM].K and C.S.LH..Y. b) the mitochondrial motifs RI..Y.(0,1)F and F..[FY]DT. c) the Golgi luminal motif K[ST].EE.K.

charged residues, which are usually involved in interactions with other proteins and C.S.LH..Y has a cysteine, which forms dimers. The motif C.S.LH..Y was enriched 37-fold, although the enrichment was surprisingly not significant. The disorder masking yielded fewer SLiMs: besides the already mentioned [ILM]..L.R motif found in the proteins localising to nuclear envelope there was also the motif QR.(0,2)R..R found in the proteins localising to the nucleus. This motif has three positively charged arginines and resembles the PKA phosphorylation site (MOD_PKA_1) and the proprotein convertase motifs (MOD_PCSK motifs).

Two proteins from the cytoskeleton group had the C.S.LH..Y motif and 11 proteins had E.[IM]..[ILM].K. The proteins gamma-syntrophin and serine/threonine-protein kinase MARK2 had both motifs, protein cyclin-A1 had only the C.S.LH..Y motif and cyclin-B2 had only the E.[IM]..[ILM].K motif. Interestingly, cyclin-B2 was previously reported to localise to the Golgi (Jacketman et al. 1995). Among the proteins with the E.[IM]..[ILM].K motif were also two poorly characterised proteins: FAM161A and C1orf49. The Golgi motif KE..K was found in 35 proteins localising to the Golgi. Among them was coatamer subunit beta (COPB1), which had two instances of the motif. In total 5 proteins had two instances of the motif, which were >100 residues apart from each other. The second Golgi motif F.M.AE and its variant F.M.E were found in 10 proteins, including COPB1 and Rab1b. The motif is repeated three times in the MON2 homolog protein. The motif [ILM]..L.R and its variants [IV]..L.R.E and [IL]..L.R were found in 13 proteins.

The proteins containing the motifs listed in Table 3.3 were investigated for common interactors with STRING (Szklarczyk et al. 2015). The mitochondrial proteins isocitrate dehydrogenase 3 gamma (IDH3G) and 4-aminobutyrate aminotransferase (ABAT) both interacted with aldehyde dehydrogenase 5 member 1 (ALDH5A1) (Figure 3.6b). Isocitrate dehydrogenases 1, 2 and 3 alpha and beta (IDH1, IDH2, IDH3A, IDH3B) were also annotated as interactors with ALDH5A1, although they do not contain the above motif. The two cytoskeletal proteins containing E.[IM]..[ILM].K -
CHAPTER 3. DISCOVERY OF NOVEL SLIMS INVOLVED IN PROTEIN LOCALISATION

Table 3.4: The motifs discovered in the topology groups.

<table>
<thead>
<tr>
<th>Class</th>
<th>Sig</th>
<th>SLiM</th>
<th>Characterisation</th>
<th>Masking</th>
<th>Enrichment</th>
<th>P-value</th>
<th>UPCs with SLiM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER-TMEM</td>
<td>0.004</td>
<td>CP.CR</td>
<td>Similar to zinc finger motif</td>
<td>Default</td>
<td>5.67</td>
<td>0.095</td>
<td>6</td>
</tr>
<tr>
<td>ER-TMEM</td>
<td>0.012</td>
<td>C.CR</td>
<td>Similar to zinc finger motif</td>
<td>Default</td>
<td>2.60</td>
<td>0.17</td>
<td>11</td>
</tr>
<tr>
<td>ER-TMEM</td>
<td>0.002</td>
<td>K.{0,1}K.$</td>
<td>ER motif TRG_ER_dilyl_1</td>
<td>Default</td>
<td>3.15</td>
<td>0.17</td>
<td>8</td>
</tr>
<tr>
<td>Golgi-luminal</td>
<td>0.009</td>
<td>K{ST}_EE.K</td>
<td>ER motif TRG_ER_diLys_1</td>
<td>Default</td>
<td>23.9</td>
<td>0.003</td>
<td>5</td>
</tr>
<tr>
<td>Golgi-TMEM</td>
<td>0.08</td>
<td>K.F.Q</td>
<td></td>
<td>Default</td>
<td>2.39</td>
<td>0.39</td>
<td>6</td>
</tr>
</tbody>
</table>

The Sig values were calculated with the SLiMChance algorithm. The threshold Sig-value after the Bonferroni correction was $0.004 (=0.05/13 topology classes). The enrichment was calculated as a ratio between the proportion of SLiMs in the positive versus the negative dataset. TMEM = transmembrane.

Table 3.5: The motifs discovered in the terminal regions of the proteins whose localisation was affected by the fluorescent tag.

<table>
<thead>
<tr>
<th>Class</th>
<th>Sig</th>
<th>SLiM</th>
<th>Masking</th>
<th>Enrichment</th>
<th>P-value</th>
<th>UPCs with SLiM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleus - C tag</td>
<td>0.079</td>
<td>L.KIS</td>
<td>C-terminal only</td>
<td>3.03</td>
<td>0.12</td>
<td>3</td>
</tr>
<tr>
<td>Mitochondria - N tag</td>
<td>0.081</td>
<td>G.OK</td>
<td>N-terminal only</td>
<td>1.72</td>
<td>0.03</td>
<td>4</td>
</tr>
<tr>
<td>Mitochondria - N tag</td>
<td>0.094</td>
<td>[HY][FV].E</td>
<td>N-terminal only</td>
<td>2.43</td>
<td>0.11</td>
<td>17</td>
</tr>
</tbody>
</table>

The Sig values were calculated with the SLiMChance algorithm. The threshold Sig-value after the Bonferroni correction was $0.0056 (=0.05/9 classes). The enrichment was calculated as a ratio between the proportion of SLiMs in the positive versus the negative dataset.
MAP/microtubule affinity-regulating kinase 2 (MARK2) and RhoGEF2 (ARHGEF2) - directly interact with each other according to STRING (Figure 3.6a). Cyclin A1 (CCNA1) and B2 (CCNB2) both interacted with two cyclin-dependent kinases (CDK1 and CDK2), although the above motifs are unlikely involved, because they do not share a common motif. In a network of the proteins containing Golgi motifs, proteins ArfGAP2 and COPB1 interacted directly (not shown). No indirect interactions were found.

Further segregation of the endomembrane localisation groups based on membrane topology yielded 5 motifs with Sig-values <0.10 (Table 3.4). Two zinc finger motifs were identified among the transmembrane proteins localising to the ER. The motif K..[ST].K..$ is a C-terminal ER targeting motif specific for transmembrane proteins. Interestingly, the enrichment score was neither high nor significant. Two motifs with positively charged lysines and negatively charged glutamic acids were found in the Golgi dataset. The motif K[ST].EE.K was found in 5 proteins from the group of soluble luminal Golgi proteins, namely osteopontin, clusterin, proprotein convertase subtilisin/kexin type 5, leucine-rich repeat-containing protein 17 and synapse-associated protein 1. The enrichment factor was 23.9, showing high specificity of this motif for this dataset. The motif K..F.Q was found in the group of transmembrane Golgi proteins, among them were surfactant locus 4, YIPF2 and two less characterised proteins FAM3A and FAM198B. The motif discovery in the disordered regions did not yield any SLiMs with Sig-values <0.10.

The interaction network of the proteins containing the Golgi motif K[ST].EE.K showed an interesting network of osteopontin and integrins, which were not among the motif-containing proteins (Figure 3.6c). This network cluster was connected to proteins clusterin and proprotein convertase subtilisin/kexin type 5. No direct nor indirect interactions were found between the proteins containing the motif K..F.Q.

The attachment of a fluorescent protein did alter the localisation of 224 proteins according to the dataset. The dataset was separated into the proteins that changed from original localisation class when a fluorescent protein was attached to their N-terminus (N group) or when it was attached to their C-terminus (C group). The discovery was performed as follows: the first 30 residues at the N-terminus were unmasked in N group and the last 30 residues at the C-terminus were unmasked in C group. The searches were performed separately and yielded 3 SLiMs with Sig<0.10 (Table 3.5). The SLiM LKIS was discovered in the C group of the nuclear proteins and was similar to an actin binding WH2 motif (LIG_Actin_WH2_2_b or LIG_Actin_WH2_2_d), however these are N-terminal motifs and LKIS was only discovered at the C-terminus. Therefore the similarity is likely coincidental. The SLiMs G.CK and [HY][FV]..E were discovered in the N group of the mitochondrial proteins. G.CK was similar to a palmitoylation motif in Wnt proteins (MOD_WntLipid), but it matched only with two sites. [HY][FV]..E partly overlapped with a complex nuclear export signal (TRG_NES_CRM1_1_a or TRG_NES_CRM1_1_b). Four proteins had the SLiM [HY][FV]..E in their sequence, but their localisation changed from mitochondria to cytoplasm. It is therefore unlikely that this is a true nuclear export signal.
3.2.3 Experimental validation

Computational predictions may be prone to high false discovery rates, therefore their validity needs to be confirmed with experimental validation. A SLiM can be experimentally validated by 'removing' it from the protein sequence, i.e. by inducing point mutations to change the residues of the SLiM to an inert amino acid, such as alanine. Alternatively, the sequence of the SLiM can be synthesised as a synthetic DNA or peptide, ligated to a (fluorescent) tag and transfected into cells. The last approach can be used to investigate localisation motifs, as has been shown for the ER retention KDEL motif, which alone can redirect a GFP to the lumen of the ER (given that the GFP is first translocated across the membrane by a signal peptide). Therefore the above approach was used to experimentally validate one predicted SLiM. Two single strands of DNA encoding the peptide corresponding to the discovered motif were synthesised, annealed and ligated into a pEGFP-C1 vector. The construct generated had a short peptide encoding the SLiM attached to the C-terminus of a GFP molecule. The construct was expressed in HeLa Kyoto cells and imaged after 24 hours.

One predicted SLiM was experimentally evaluated (Table 3.6). SLiM F..M.AE (motif F) scored highly in the SLiMChance evaluation and it was highly enriched over the negative dataset. In addition, this motif was found in COPB1, which is a multisubunit complex and a component of the basic machinery in COPI retrograde trafficking. This motif could facilitate the interactions between the subunits. The DNA sequence of the above motif was copied from the motif region of the COPB1 ORF and de novo synthesised as single strands of DNA (Table 3.7).

Table 3.6: The experimentally evaluated SLiM with the Sig-value, the enrichment factor and the enrichment p-values.

<table>
<thead>
<tr>
<th>SLiM</th>
<th>Sig</th>
<th>Enrichment</th>
<th>Enr p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motif F: F..M.AE</td>
<td>0.006</td>
<td>10.3 (Golgi)</td>
<td>0.025</td>
</tr>
</tbody>
</table>

The SLiMChance threshold p-value corrected for multiple testing with the Bonferroni correction was 0.004.

The DNA sequence shown in Table 3.7 was synthesised as single strands and then annealed to form a double stranded DNA. The fragment was then ligated into a pEGFP-C1 vector and the insertion into the vector was confirmed by sequencing.

Table 3.7: The experimentally evaluated SLiM with the example protein, the localisation class, the peptide and the DNA sequence.

<table>
<thead>
<tr>
<th>SLiM</th>
<th>Protein</th>
<th>Localisation</th>
<th>Peptide</th>
<th>cDNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>F..M.AE</td>
<td>Coatomer subunit beta</td>
<td>Golgi</td>
<td>FRQMWAE</td>
<td>TTTCGCCAGATGTGGGCGGAA</td>
</tr>
</tbody>
</table>

The peptide and DNA sequence were extracted from the example protein.

The designed construct was transfected into HeLa Kyoto cells and images of live and fixed cells were then acquired. As shown in Figures 3.7 and 3.8 the localisation of the GFP-SLiM construct was not visibly different from the localisation of the soluble GFP without the attached peptide. This indicates that the SLiM F..M.AE alone was unable to influence the targeting of the soluble GFP molecule.
Figure 3.7: Images of live HeLa Kyoto cells expressing the GFP-tagged peptide corresponding to the putative Golgi SLiM F..M.AE. The images were acquired on a wide-field fluorescence microscope.

Figure 3.8: Images of fixed HeLa Kyoto cells expressing the GFP-tagged peptide corresponding to putative Golgi SLiM F..M.AE. The images were acquired on a wide-field fluorescence microscope. Green colour shows the GFP channel and blue colour shows the nuclei.

3.3 Discussion

The aim of this study was to discover SLiMs that are overrepresented in groups of proteins with similar localisation. The SLiMs were computationally predicted using SLiMFinder software that searches for overrepresented regions in the dataset, but it does not account for exclusiveness of a SLiM to a certain dataset. The exclusiveness was determined by calculating the enrichment factor of a SLiM for a particular localisation class. The proteins from the dataset were classified into the localisation classes and SLiMs were searched for within each class. Some of the proteins visualised in the endomembrane system were probably misclassified, either because they were not yet in their final location or because overexpression caused overflow of the proteins into the neighbouring compartments. An example of this are the proteins that contain the motif K[ST].EE.K in their sequence. Proteins osteopontin, clusterin and proprotein convertase subtilisin/kexin type
5 were classified as Golgi proteins although they are components of the extracellular matrix and are eventually secreted from the cell (Burkey et al. 1992; Horton et al. 2009; Craig et al. 1990).

The localisation data of N- and C-terminally tagged proteins were available in the original dataset, therefore particular attention was paid to the proteins that changed their localisation as a result of the attachment of a fluorescent protein. Fusion of a protein with a fluorescent protein can cause a change in the protein conformation, which also affects disordered regions, typically rich with SLiMs (Lotti et al. 2011). The largest localisation classes were cytoplasm, nucleus and ER, but the proteins localising to the nuclear envelope, Golgi and mitochondria were the most affected by the attachment of the fluorescent protein. A comparison with the proportion of the affected proteins in the whole dataset revealed that most of the localisation classes are affected, albeit by differing amounts, by the attachment of a fluorescent protein. By contrast, the cytoplasmic proteins were affected inversely - in fact the localisation of less than 6% of cytoplasmic proteins was affected. In total, 23% of the proteins in the dataset were affected. This is an interesting observation and reveals the importance of careful interpretation of data acquired with overexpressed fusion proteins. A comparative screen between the localisations of fluorescent protein fusions and immunofluorescence staining of endogenous proteins did report a high correlation between the localisation of the overexpressed GFP-tagged proteins and native proteins. From a dataset of more than 500 proteins analysed, 21% of the proteins showed dissimilarities in the localisation between both techniques and the majority of the discrepancies was observed among the proteins localising to the endomembrane system. The discrepancies were likely often caused by the presence of N-terminal signal or transit sequences, or because immunofluorescence studies require fixation and permeabilisation of cells, which could disrupt small structures within the cell and change the epitopes (Stadler et al. 2013). The same study also compared the localisation of N- versus C-terminally tagged proteins. They reported that 26% of the proteins showed dissimilarities between the N- and C-terminally tagged protein, which is 3% higher than the proportion of dissimilarities reported in the current work (23%).

The classification of the proteins based on their topology was an attempt to focus the SLiM discovery on localisation mechanisms that are specific for proteins of a certain topology. This assumption was supported by the ER retention motifs, which are different for transmembrane and luminal proteins, KK..$ and KDEL, respectively.

The discovery of SLiMs in the localisation classes yielded a large number of motifs, but only the SLiMs with Sig-values <0.10 were retained for further evaluation. In general, Sig-values in all searches were relatively high, with the exception of some well known motifs, such as the nuclear localisation signal and the dilyssine ER retention motif. A typical threshold Sig-value 0.05 was corrected with Bonferroni correction for multiple testing, which excluded the majority of hits as likely false positives. Despite that, certain motifs did show interesting profiles. The motif F..M.AE was present in COPB1, Rab1b and MON2, which are all involved in membrane trafficking (Efe et al. 2005). The motif [ILM]..L.R, which was found in the nuclear envelope-localising proteins, is similar to the motif that binds nuclear receptors (Heery et al. 1997). Interesting were also the cytoskeletal motifs E,[IM]..[ILM],K and C.S.LH..Y, although no reliable associations between the proteins with these motifs could be found. All of the above proteins were also enriched in their localisation classes. Proteins osteopontin, clusterin and proprotein convertase subtilisin/kexin type 5 had mo-
tif K[ST].EE.K in their sequence and were classified as Golgi proteins although they are secreted (see above). This motif may be important for the interaction with cell surface integrins, which were, according to STRING, interacting with these proteins but did not have the motif in their sequence. It could also be a post-translational modification site in secreted proteins. The discovery of SLiMs in the terminal regions of the proteins did not yield any significant SLiMs, although the localisation of 24% of the proteins was disrupted by the GFP tag.

A property of localisation SLiMs is that they are specific for proteins that localise to a certain compartment in a cell. Therefore a strategy to assess the exclusiveness of a SLiM in a particular localisation group was investigated and implemented by calculating the enrichment of a SLiM in the class over the negative dataset. This provides information about how specific a SLiM is for the dataset. This score needs to be interpreted side by side with the SLiMChance Sig value, which estimates the probability that the SLiM was found in a particular dataset just by chance. It is expected, that many motifs would be enriched in their localisation class, because of the biased search space.

The aim of this study was also to experimentally validate the SLiM candidates. A peptide with a putative SLiM, extracted from the sequence of the protein in which the SLiM was discovered, was fused with a GFP and expressed in mammalian cells. The peptide itself did not have an effect on the localisation of the GFP. However, the validation strategy was not optimal, because localisation SLiMs often just fine-tune the localisation and work in synergy with other localisation mechanisms in the protein. In addition, only a single SLiM was tested, making it unclear as to whether this is realistically an efficient approach to validate potential SLiMs. An alternative approach would be to mutate the motif in the original protein and track its localisation. This approach is more likely to show the effect of a SLiM on the localisation, but requires more time and resources and may not be feasible for screening a large number of SLiM candidates. A related problem is the high false positives ratio among predictions, resulting in the high rate of negative experimental validations. This could be avoided by determining the real false discovery rate by performing a pilot screen with a random sample of predicted SLiMs. The Sig values could then be corrected with the real false discovery rate, which would eliminate the SLiM candidates that are likely to be the false positive predictions.

The localisation data in the original localisation dataset from Simpson et al. 2000 was manually interpreted, which does introduce certain bias. Alternatively, the images from the dataset could be classified by an unsupervised classification algorithm and texture features, as is demonstrated in Chapter 6. The SLiM discovery would then be performed within the automatically determined localisation classes. This would provide an unbiased classification and could also find novel localisation relationships within the dataset.

The SLiMFinder performs well because of its completeness. It takes into account the evolutionary relationships between sequences, the size of the search space and also provides a variety of masking options. It could be improved by taking into account the exclusiveness of a SLiM. A promising approach was developed for another SLiM discovery tool, FirePro (Lieber et al. 2010). This algorithm builds a protein profile for each SLiM and determines the mutual information within that profile. It does estimate the exclusiveness as well, but it does not take into
account the evolutionary relationships between the sequences, which would wrongly estimate whether a SLiM is a real instance or overrepresented because of high identity.

Overall, this study aimed to discover novel SLiMs in a dataset of proteins with the experimentally determined localisation. SLiM prediction is a difficult process with high false discovery rates, because SLiM signals are weak. While it remains an interesting and prospective research field there is a need for the improved estimation of true positives.
Chapter 4

Characterisation of PRAF proteins
4.1 Introduction

The endoplasmic reticulum (ER) is the largest organelle in endomembrane system occupying as much as half of the cell volume depending on the cell type. The ER is composed of a network of tubules and cisternae, and is the site of synthesis of proteins that are needed in the endomembrane system or secreted from the cell. Alongside this main function it is also the main storage of Ca\(^{2+}\) ions and the site of lipid biosynthesis. These activities are facilitated by maintaining close contacts with several organelles in the cell, including the Golgi, peroxisomes, mitochondria, plasma membrane and endosomes. Contacts with mitochondria are probably the most studied; at these sites the membranes of the mitochondria and ER come close together but they do not fuse. The gap between these two membranes is 10-30 nm, which means they can be tethered to each other by tethering proteins (Csordás et al. 2006). ER-mitochondria contacts are important for lipid biosynthesis, Ca\(^{2+}\) signalling and maintenance of morphology and dynamics of both organelles (reviewed in Rowland et al. 2012).

The region where the ER and mitochondria form a contact is rich in proteins involved in lipid biosynthesis; the mitochondria-associated membrane (MAM) is enriched in proteins such as phosphatidylerine (PS) synthase, phosphatidylethanolamine (PE) synthase and phosphatidylcholine (PCh) synthase (Stone et al. 2000; Vance 1990; Voelker 2000). The synthesis of PCh and PE starts by the synthesis of PS in the ER, which is translocated to the outer mitochondrial membrane (OMM) and then to the inner mitochondrial membrane (IMM), where it is converted into PE by mitochondrial enzymes. PE is then translocated back to the ER, where it is finally converted into PCh. Another mechanism translocates them again to mitochondria (reviewed in Osman et al. 2011). The mechanism of synthesis and exchange of the above phospholipids between the ER and mitochondria is still largely uncharacterised. It is partly facilitated through a four member complex ER-mitochondria encounter structure (ERMES), which was identified in a screen of yeast mutants where ER-mitochondria contacts could be rescued by an artificial tethering protein. Cells where ERMES components were not rescued had lower PS to PCh conversion rates (Kornmann et al. 2009). Also, four members of ERMES, Mdm10, Mdm34/Mmm2, Mdm12 and Mmm1, have a synaptotagmin-like, mitochondrial and lipid-binding proteins (SMP) domain, which belongs to the tubular lipid-binding proteins (TULIP) superfamily of hydrophobic ligand binding domains. These proteins are thus likely involved in actual exchange of lipids rather than merely tethering two membranes (Kopec et al. 2010). The role of ERMES in lipid exchange has been debated in several studies, for example where no difference in the PS to PE conversion rate was observed after depletion of the ERMES component Gem1 in yeast (Nguyen et al. 2012). Recently a contact site between the ER and vacuole, called vCLAMP, was discovered in yeast, which essentially provides a link between the ER and endocytic compartments (Elbaz-Alon et al. 2014; Hönscher et al. 2014). vCLAMP was able to rescue ERMES mutants and vice versa, while mutants of both ERMES and vCLAMP were lethal (Elbaz-Alon et al. 2014). vCLAMP mutants with gradual depletion of ERMES resulted in a complete loss of lipid exchange between the ER and mitochondria (Elbaz-Alon et al. 2014).

Recent studies have also reported on the role of ER shaping proteins in lipid exchange between the ER and mitochondria (Voss et al. 2012). In particular reticulon 1c was shown to affect the
morphology of mitochondria, which is likely related to the physiology of ER-mitochondria contacts (Reali et al. 2015). Reticulons are a family of proteins that form a scaffold needed for shaping membranes into tubule like structures in the ER. They are four-pass transmembrane proteins with shorter transmembrane domains, which form a short-hairpin structure. The short hairpins wedge into membrane bilayers, but because of their shorter length, they can only penetrate through the first layer of phospholipids, thus increasing the area of one layer by ~10% (Voeltz et al. 2006; Zurek et al. 2011). Other ER shaping proteins, DP1, atlastins and protrudin, also have a short-hairpin domain (Voeltz et al. 2006; Chang et al. 2013; Hashimoto et al. 2014).

Such a short hairpin structure is also a structural feature of the PRAF protein, which has been previously associated with membrane trafficking. It was proposed to be a GDI displacement factor (GDF) in the Rab activation cycle, facilitating removal of GDI from Rab which would enable its transition to GTP-bound active state (Sivars et al. 2003). Recent studies have suggested that the GDP/GTP exchange factor (GEF) is sufficient for the localisation and activation of a Rab protein, therefore the role of GDF is questioned (Blümer et al. 2013; Gerondopoulos et al. 2012). PRAF3 has been widely studied for its role in glutamate trafficking, but is normally present only in developing tissues. Overexpression of PRAF3 blocks the exit of glutamate transporter EAAC1 from ERES because of a competitive interaction with Rab1a (Maier et al. 2009). The least is known about PRAF2. Unlike PRAF3 it is expressed in adult tissues and overexpression reduces the levels of CCR5 at the plasma membrane (Fo et al. 2006; Schweneker et al. 2005).

All PRAF proteins localise to the early secretory pathway, PRAF1 to the Golgi and ER (Abdul-Ghani et al. 2001; Gougeon et al. 2002; Liang et al. 2004b; Alvim Kamei et al. 2008; Bhagatji et al. 2010; Geng et al. 2005); PRAF2 and PRAF3 to the ER (Schweneker et al. 2005; Abdul-Ghani et al. 2001; Liu et al. 2008b). They possess an interesting organisation of transmembrane domains - two long hydrophobic regions have been predicted in PRAF proteins. The length of the two hydrophobic regions in PRAF2 and PRAF3 is 40 residues each, which corresponds with lengths of two typical TMDs. This organisation of TMDs would form a hairpin-like structure, where two TMDs are tightly packed parallel to each other with little or no linking sequence exposed outside of the membrane bilayer. Hydrophobic regions of PRAF1 are slightly shorter, 34 residues long, which would correspond to two short, 17 residues long, TMDs also shaped as a hairpin-like structure. N- and C-terminal regions are soluble as well as the loop between both hydrophobic regions. PRAF1 and PRAF2 are ubiquitously expressed in human tissues (Martinicic et al. 1997; Bucci et al. 1999; Abdul-Ghani et al. 2001; Schweneker et al. 2005), while PRAF3 is mainly expressed in developing tissues and declines after birth (Maier et al. 2009).

Experimental validation of the PRAF1 topology has confirmed the above predictions and also determined the orientation of soluble loops, which are all cytosolic (Lin et al. 2001b). Moreover, the authors reported on sequential membrane integration of the PRAF1 protein, following the co-translational model; they also reported, that single TMDs are not sufficient for membrane integration and hence an intact hydrophobic region is required.

Depletion of PRAF1 resulted in differential expression of proteins involved in lipid metabolism and cell migration. Immunofluorescence showed accumulation of cholesterol and changes in localisation of the regulated proteins. The phenotype was rescued by overexpression of PRAF1 or mimicked with cholesterol transport inhibitor U18666A (Liu et al. 2011a)
Motivated by conflicting and incomplete characterisation of PRAF proteins systematic subcellular localisation, membrane topology and functional characterisation of the PRAF proteins was carried out. They all localised to the ER; PRAF2 and PRAF3 showing a clear reticular phenotype, whereas PRAF1 changed the structure of ER tubules. Profiling of their interaction networks showed involvement in lipid processing events. Depletion of PRAF1 did not show a significant effect on ER structure, but it rather altered the morphology of mitochondria. This suggests that PRAF1 is an ER localising protein that may be involved in lipid metabolism thus having an effect on the morphology of mitochondria.

### 4.2 Results

#### 4.2.1 Design of GFP- and HA-tagged constructs

A library of wild-type and mutated PRAF constructs fused with a GFP or a HA tag was designed to investigate the localisation, function and localisation determinants of PRAF proteins. Three wild-type PRAF proteins were designed: N-terminally GFP-tagged, C-terminally GFP-tagged and N-terminally HA-tagged (Figure 4.1). Four mutated versions of each PRAF were designed, as illustrated on Figure 4.1: PRAFxA and PRAFbc were cloned into vector pEGFP-N1; PRAFxC and PRAFb were cloned into vector pEGFP-C1.

The open reading frames (ORFs) of PRAF proteins were amplified with custom primers, digested and ligated into the multiple cloning site (MCS) of pEGFP-C1, pEGFP-N1 or pCMV-HA vectors (see Materials and Methods for full details). The ligation reaction was validated by a restriction digestion - the plasmid was extracted from the bacterial colonies transformed with the ligated product and digested with the original cloning enzymes. The reaction was then separated on an agarose gel (Figure 4.2). A successful ligation reaction separates into two bands, one representing the linearised vector backbone (~4700 bp) and the other one representing the inserted ORF of the wild-type (~550 bp) or the deletion mutants (~250-500 bp). The ligation was verified by sequencing.

In total 3 N-terminally GFP-tagged, 3 C-terminally GFP-tagged and 3 C-terminally HA-tagged wild type PRAF constructs were designed (Table 4.1). Four mutants of each PRAF were also designed: PRAFxC and PRAFb were GFP-tagged on the N-terminus; and PRAFxA and PRAFbc were GFP-tagged on the C-terminus (Table 4.1).

<table>
<thead>
<tr>
<th>Construct</th>
<th>N-terminal GFP tag</th>
<th>C-terminal GFP tag</th>
<th>N-terminal HA tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRAFwt</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PRAFxA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PRAFxC</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PRAFb</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PRAFbc</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 4.1: A list of the designed PRAF fusion constructs.
4.2.2 Localisation

The localisation of PRAF proteins was determined by imaging HeLa Kyoto cells transfected for 24 h with the GFP-tagged PRAF constructs. Images were first captured from live cells. The N- and C-terminally tagged PRAF proteins localised to the ER, regardless of the position of the tag, indicating that the presence of the fluorescent protein does not interfere with possible localisation signals in the terminal regions (Figure 4.3). In PFA-fixed cells the PRAFs also localised to the ER (Figure 4.4).

A subset of live and fixed cells overexpressing GFP-tagged PRAF1 did not show a typical reticular pattern, but instead a profile resembling soluble GFP with aggregates was observed. Co-staining with the tubular ER marker REEP5 (Figure 4.10) showed that overexpressed PRAF1 was not soluble but rather changed the shape of the tubular ER.

The fusion with the GFP tag could have an effect on the localisation of the protein, therefore the observed localisation patterns were compared with the localisation phenotypes of PRAF proteins with a shorter haemagglutinin (HA) tag, which is only 9 residues long (Figure 4.4). The localisation of co-expressed GFP- and HA-tagged PRAF1 showed the expected phenotype - ER localisation with a poorly defined reticular pattern. C-terminally GFP-tagged PRAF1 showed a weak retic-
Figure 4.2: Agarose gels of the restriction digestion of the wild-type and the mutant PRAF constructs. a) The open reading frames of the wild-type PRAF proteins were ligated into pEGFP-N1, pEGFP-C1 or pCMV-HA vectors. b) The mutated PRAF ORFs were ligated into pEGFP-C1 or pEGFP-N1 vectors. Multiple gels are stitched together. The first gel fragment in each set of the gels shows the molecular weight markers.

ular pattern, and a number of larger aggregates were also present. Some cells also formed small protrusions on expression of the C-terminal fusion, an effect not observed with the N-terminally GFP-tagged PRAF1. None of the proteins showed any localisation to the Golgi.

In the cells overexpressing different PRAF constructs (Figure 4.4) a change in the reticular phenotype could be observed. C-terminally GFP-tagged PRAF2 showed a normal reticular localisation pattern, however the N-terminally GFP-tagged PRAF2 was more similar to the PRAF1 phenotype, specifically showing poorly defined reticular pattern. It also formed aggregates, which were not seen with the overexpression of HA construct.

A reticular pattern was also seen in cells overexpressing the PRAF3 (Figure 4.4) fusion proteins. N-terminally GFP-tagged PRAF3 showed small changes in the ER architecture and a small number of elongated and straight tubules could also be seen.

Several unusual phenotypes were observed when different combinations of GFP- and HA-tagged PRAF proteins were co-transfected into the cells (Figures 4.5 and 4.6). GFP-PRAF1/PRAF2-HA showed poorly defined reticular pattern, similar to the PRAF1 phenotype (Figure 4.5). Similar phenotypes were observed with GFP-PRAF3/PRAF2-HA and PRAF3-GFP/PRAF2-HA (Figure 4.5). The combinations GFP-PRAF1/PRAF3-HA, GFP-PRAF2/PRAF1-HA and PRAF2-GFP/PRAF1-HA showed several spherical compartments around the nucleus with a bright aggregate in the mid-
CHAPTER 4. CHARACTERISATION OF PRAF PROTEINS

**Figure 4.3:** Localisation of the GFP-tagged PRAF proteins expressed in HeLa Kyoto cells for 24 h. Live cells were imaged on a wide-field fluorescence microscope.

**Table 4.2:** Potential targeting motifs in PRAF proteins retrieved from the ELM database.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Motif name</th>
<th>Motif expression</th>
<th>Sequence in protein</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRAF1</td>
<td>TRG_ENDOCYTIC_2</td>
<td>Y.[LMVIF]</td>
<td>YQSNYVFVFL GLILYCVV TSPM</td>
<td>78-81* 88-91*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRAF2</td>
<td>TRG_ER_diArg_1</td>
<td>Y[(LIVMFYWPR)R [&quot;YFWDE][0,1]R (R[&quot;YFWDE][0,1]R [LIVMFYWPR])]</td>
<td>RAAV RRCRR SHPA</td>
<td>91-93 92-95 93-96</td>
</tr>
<tr>
<td>PRAF3</td>
<td>possible TRG_ER_dilys_1</td>
<td>K.[0,1]K.[2,3]S</td>
<td>DYISKVKES</td>
<td>185-188</td>
</tr>
</tbody>
</table>

* The motif is in the transmembrane domain.

dle (Figure 4.6). The ER was visually disrupted in PRAF1-GFP/PRAF3-HA, GFP-PRAF2/PRAF3-HA and PRAF2-GFP/PRAF1-HA (Figures 4.4, 4.5 and 4.6).

Several combinations also showed different intensities in different regions of the ER. The HA-tagged constructs preferentially localised in the periphery of the cell, while the GFP-tagged constructs localised uniformly throughout the cell (for example PRAF1-HA/GFP-PRAF3 on Figure 4.5). The combinations GFP-PRAF1/PRAF2-HA (Figure 4.5), PRAF1-GFP/PRAF1-HA (Figure 4.4) and PRAF3-GFP/PRAF2-HA (Figure 4.5) induced the formation of long cell protrusions, while shorter protrusions were seen in the majority of the other combinations.
Figure 4.4: Localisation of GFP- and HA-tagged PRAF proteins (part 1). Constructs of GFP- and HA-tagged PRAF proteins were expressed in HeLa Kyoto cells for 24 h, stained with anti-HA antibody and imaged with a laser-scanning confocal microscope. The combinations of constructs in each row are indicated on the left side. The first and the second column of the images show the localisation of GFP- and HA-tagged construct, respectively. The third column of the images is a merged image of both channels: green colour represents GFP channel, red colour represents HA channel and blue colour represents nuclear stain Hoechst 33342.
Figure 4.5: Localisation of GFP- and HA-tagged PRAF proteins (part 2). Constructs of GFP- and HA-tagged PRAF proteins were expressed in HeLa Kyoto cells for 24 h, stained with anti-HA antibody and imaged with a laser-scanning confocal microscope. The combinations of constructs in each row are indicated on the left side. The first and the second column of the images show the localisation of GFP- and HA-tagged construct, respectively. The third column of the images is a merged image of both channels: green colour represents GFP channel, red colour represents HA channel and blue colour represents nuclear stain Hoechst 33342.
Figure 4.6: Localisation of GFP- and HA-tagged PRAF proteins (part 3). Constructs of GFP- and HA-tagged PRAF proteins were expressed in HeLa Kyoto cells for 24 h, stained with anti-HA antibody and imaged with a laser-scanning confocal microscope. The combinations of constructs in each row are indicated on the left side. The first and the second column of the images show the localisation of GFP- and HA-tagged construct, respectively. The third column of the images is a merged image of both channels: green colour represents GFP channel, red colour represents HA channel and blue colour represents nuclear stain Hoechst 33342.
Transmembrane ER proteins are actively retained in the ER by a COPI coat-dependent mechanism, which recognises the ER retention signal at their C-terminus. The putative targeting motifs within PRAF proteins were therefore predicted by searching the ELM database. Three putative instances of a diarginine motif were found in the cytosolic loop of PRAF between residues 91 and 96 (Table 4.2). PRAF3 has a K.K.$ motif at the C-terminus which does resemble a dilysine retention motif, but it does not match the consensus dilysine motif. Interestingly, PRAF1 possesses a tyrosine based motif that was found in endocytic proteins and facilitates the interaction with the adapter complex of the clathrin coat (Opresko et al. 1995). In PRAF1 it is located within a predicted transmembrane domain, whereas in all other proteins it would be located in a cytosolic region of the protein. In this configuration therefore, it would be incapable of binding the clathrin coat.

The sequence determinants of the localisation and the predicted targeting motifs were next experimentally investigated by removing the regions of PRAF proteins and then assessing their localisation (Figure 4.1). PRAF mutant constructs were transfected into HeLa Kyoto cells, fixed and
imaged. The localisation profiles are illustrated in Figure 4.7. All mutants retained their ER localisation, even when the predicted ER retention motifs were removed. However, overexpressed PRAF1 truncation mutants did not induce ER reorganisation to the same extent as the wild-type PRAF1. A difference was observed with PRAF1bc, which showed a slightly denser ER network. In some cells it also localised to large punctate structures. These could also be aggregates, however their intensity was not very high, as might be expected for aggregates. PRAF1ab appeared less intense in the peripheral region. PRAF2bc showed some disruptions to the ER structure, which appeared as membrane clusters. Some disruption to the regular ER pattern was also observed in PRAF2ab-expressing cells. A number of high-intensity large punctate structures were observed in cells expressing all PRAF3 mutants. PRAF3ab and PRAF3bc overexpression resulted in the production of punctate structures, which looked like membrane aggregates.

In general, GFP-tagged PRAF proteins localise to the ER regardless of the terminus that the fluorescent protein is attached to. The predicted ER localisation motifs do not seem to be the drivers for ER localisation and all truncation mutants tested showed ER localisation. Certain deformations of the ER structure were observed, likely as a consequence of overexpression, although interestingly these deformations were more pronounced on overexpression of wild-type proteins rather than the truncation mutants.

### 4.2.3 Membrane topology

The predictions and previous reports on the membrane topology of the PRAF proteins were validated using the FPP assay. N- and C-terminally GFP-tagged PRAF proteins were transfected into HeLa Kyoto cells and incubated for 24 h. The cells were then semi-permeabilised with digitonin for 45 s, followed by the addition of trypsin. The fluorescence was recorded with a fluorescence microscope at 5 s intervals. The positive control was a transmembrane protein, stress-related ER protein (SERP), which is a 66 residue long protein with a single predicted TMD (Uniprot annotation based on manual sequence analysis). The fluorescence profile after the treatment with trypsin is shown in Figure 4.8a. The GFP molecule in the cytosol was rapidly digested by the trypsin resulting in a rapid decrease of the fluorescence signal; whereas the luminal GFP molecule was protected from the trypsin and more of the fluorescence signal was retained (Figure 4.8b). As demonstrated in Figure 4.8a, PRAF proteins with either N- and C-terminal GFP tags were rapidly digested after the treatment with trypsin, corresponding to a cytosolic orientation. These results suggest that N- and C-terminal soluble loops are both in the cytosol in the case of PRAF proteins (Figure 4.8c).

### 4.2.4 Co-localisation with organelle markers and Rab proteins

The subcellular localisation of PRAF proteins was further characterised by co-staining the cells overexpressing the GFP-tagged PRAF proteins with various organelle markers. The localisation to the ER subdomains was tested by measuring co-localisation between the tubular ER marker REEP5 and the sheet ER marker CLIMP-63. The RWC values, presented in Figure 4.9, indicate that PRAF proteins co-localised ~2.5 fold higher with REEP5. The images of the individual channels, shown in Figure 4.10, also show reorganisation of the ER in cells overexpressing PRAF1. This reorganisation
Figure 4.8: Membrane topology of PRAF proteins. a) The plots of the FPP profiles of the positive control SERP1 and PRAF proteins after the addition of trypsin. The error bars represent the standard errors of means (N=6-37). b) Example images of the positive control before and after the treatment with trypsin. c) Proposed membrane topology model of PRAF proteins. The transmembrane domains (TMD) are represented with green cylinders and the soluble regions are represented with black curves. The N- and C-termini are marked with letters.
CHAPTER 4. CHARACTERISATION OF PRAF PROTEINS

Figure 4.9: Co-localisation of PRAF proteins with the tubular ER marker REEP5 or the sheet ER marker CLIMP-63. The RWC values are presented in plot a, and the ratios between both RWC values are presented in plot b. The error bars correspond to the standard errors of means (N=29-48).

Figure 4.10: Images of co-localisation of PRAF proteins with the tubular ER marker REEP5 or the sheet ER marker CLIMP-63. PRAF-GFP proteins are shown in green colour; the ER markers are shown in red colour; and the nuclear stain Hoechst 33342 is shown in blue colour. The red channels of the first and the second columns of images show the tubular ER marker REEP5 and the red channels of the third column of the images show the sheet ER marker CLIMP-63. The regions marked with white rectangles are zoomed in the second and the third column. The regions corresponding to the linear profiles in the fourth column are marked with white lines.

was specific to REEP5-labelled ER tubules, indicating that the overexpression of PRAF1 induces reorganisation of ER tubular elements. Confocal linear profile plots of the intensities in the ER region show the co-localisation with both organelle markers. All PRAF proteins showed similar linear profiles, mostly localising with the REEP5 marker, although occasional co-localisation peaks with CLIMP-63 were observed (Figure 4.10).
Figure 4.11: Co-localisation between PRAF proteins and GM130. The first column of the images shows GFP channels, the second column shows GM130 channels and the third column shows the merged images. Green colour in the merged images shows the GFP channel, red colour shows the GM130 immunostain and blue colour shows the nuclear stain Hoechst 33342.

The above results indicate that more than 50% of each PRAF protein consistently co-localises with ER tubules. PRAF2 and PRAF3 were previously described as ER proteins (Schwenekker et al. 2005; Abdul-Ghani et al. 2001; Liu et al. 2008b), although PRAF1 in several studies has been associated with Golgi localisation (Abdul-Ghani et al. 2001; Gougeon et al. 2002; Liang et al. 2004b; Alvim Kamei et al. 2008; Bhagatiji et al. 2010; Geng et al. 2005) and ER localisation (Abdul-Ghani et al. 2001; Liang et al. 2004b; Alvim Kamei et al. 2008; Geng et al. 2005). Co-localisation assays were therefore repeated with the Golgi marker GM130. Although PRAF1 co-localised to a greater extent with GM130 than PRAF2 and PRAF3, the co-localisation values were all less than 0.10, and as can be seen in the images, very little PRAF signal was found in the juxtanuclear region of the cells (Figure 4.11).

Co-localisation with other cellular compartments was systematically determined by quantifying the co-localisation between the GFP-tagged PRAF proteins and overexpressed mCherry-tagged Rab proteins. Rab proteins are a relatively well described family of proteins that localise to different cellular compartments and are commonly used to mark and identify organelles in a cell (Chavrier et al. 1990). Rab proteins selected for this assay with the compartments they localise to are listed in Table 4.3. GFP-PRAF and Rab-mCherry constructs were co-expressed in HeLa Kyoto cells, imaged after 24 h and then the co-localisation was quantified as RWC values. Rab proteins, when expressed in a cell, show two fractions: a membrane attached fraction and a soluble cytosolic fraction. The presence of the strong soluble cytosolic signal interfered with the standard quantification pipeline, therefore all pixels with relative intensity values below 0.15 were excluded.
Table 4.3: Rab proteins used as organelle markers and their localisation.

<table>
<thead>
<tr>
<th>Rab protein</th>
<th>Compartment</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rab1a</td>
<td>Golgi, ER to Golgi trafficking</td>
<td>Saraste et al. 1995; Tisdale et al. 1992</td>
</tr>
<tr>
<td>Rab1b</td>
<td>ER to Golgi trafficking, COPI, COPII</td>
<td>Plutner et al. 1991; Dugan et al. 1995; Peter et al. 1994; Alvarez et al. 2003</td>
</tr>
<tr>
<td>Rab2a</td>
<td>ER/Golgi</td>
<td>Chavrier et al. 1990; Tisdale et al. 1992; Fischer von Mollard et al. 1991</td>
</tr>
<tr>
<td>Rab3a</td>
<td>Synaptic vesicles, synaptic exocytosis</td>
<td>Matteoli et al. 1991</td>
</tr>
<tr>
<td>Rab4a</td>
<td>Early endosomes</td>
<td>Van Der Sluijs et al. 1991</td>
</tr>
<tr>
<td>Rab5a</td>
<td>Early endosomes</td>
<td>Chavrier et al. 1990; Bucci et al. 1992; Alvarez-Dominguez et al. 1996</td>
</tr>
<tr>
<td>Rab6a</td>
<td>Golgi</td>
<td>Goud et al. 1990</td>
</tr>
<tr>
<td>Rab7a</td>
<td>Late endosomes, lysosomes, melanosomes, phagosomes</td>
<td>Chavrier et al. 1990; Bucci et al. 1995; Gomez et al. 2001; Vieira et al. 2003</td>
</tr>
<tr>
<td>Rab8a</td>
<td>Trans-Golgi/plasma membrane/constitutive secretory vesicles, melanosomes, GLUT4 vesicles</td>
<td>Chen et al. 1993; Huber et al. 1993; Chakraborty et al. 2003; Miinea et al. 2005</td>
</tr>
<tr>
<td>Rab9a</td>
<td>Late endosomes, TGN, recycling endosomes</td>
<td>Lombardi et al. 1993; Riederer et al. 1994</td>
</tr>
<tr>
<td>Rab11a</td>
<td>Recycling endosomes, GLUT4</td>
<td>Ullrich et al. 1996; Larance et al. 2005; Marzesco et al. 2002; Morimoto et al. 2005; Nokes et al. 2008; Sun et al. 2010</td>
</tr>
<tr>
<td>Rab13</td>
<td>Tight junctions, recycling endosomes to plasma membrane and TGN, GLUT4</td>
<td>Ozeki et al. 2005; Dejgaard et al. 2008; Gerondopoulos et al. 2014</td>
</tr>
<tr>
<td>Rab18</td>
<td>ER, lipid droplets</td>
<td>Chen et al. 2002; Imai et al. 2004</td>
</tr>
<tr>
<td>Rab27b</td>
<td>Melanosomes, Early secretory vesicles</td>
<td>Sinka et al. 2008; De Leeuw et al. 1998</td>
</tr>
<tr>
<td>Rab30</td>
<td>Golgi, intra-Golgi trafficking</td>
<td>Zheng et al. 1998</td>
</tr>
<tr>
<td>Rab33b</td>
<td>Medial Golgi</td>
<td>Sun et al. 2003</td>
</tr>
<tr>
<td>Rab34</td>
<td>Macropinosomes</td>
<td>Rodriguez-Gabin et al. 2004</td>
</tr>
<tr>
<td>Rab40c</td>
<td>Peri-nuclear endocytosis, recycling endosomes</td>
<td></td>
</tr>
<tr>
<td>Rab43</td>
<td>ER to Golgi trafficking, Golgi biogenesis</td>
<td>Dejgaard et al. 2008; Haas et al. 2007</td>
</tr>
</tbody>
</table>
CHAPTER 4. CHARACTERISATION OF PRAF PROTEINS

Figure 4.12: Image processing diagram for RWC measurements. The pixels with relative intensity values below 0.15 were set to 0. The left column shows example original images and right column shows the images after processing.

Figure 4.13: Heatmap showing the RWC co-localisation values between GFP-PRAF and mCherry-Rab proteins. The RWC values are represented both by the colour code and the number in each tile. The colour scale is beneath the plot.

from the quantification. This thresholding excluded most of the cytosolic fraction but retained membrane structures containing significant fluorescence signal, as shown in Figure 4.12.

The quantitative co-localisation of each PRAF protein with the various Rab proteins is presented in Figure 4.13. As expected, ER and early secretory pathway-localising Rabs showed strong co-localisations with the PRAFs, particularly PRAF2 and PRAF3. For example, Rab18 with PRAF2 (0.68); and Rab1b with PRAF2 (0.75) and PRAF3 (0.75). Less co-localisation was seen with Rabs associated with the late secretory pathway. PRAF1 also co-localised more strongly with the early secretory pathway Rab1a. PRAF1 and PRAF2 highly co-localised with Rab13 (0.76).

Example images of the co-localisation with Rab1a and Rab1b are shown in Figure 4.14. Greatest co-localisation was seen with the reticular membrane structures and the punctate structures in Rab1a- and Rab1b-overexpressing cells, but less with the juxtanuclear Golgi membranes (Figure 4.14).

PRAF2 and PRAF3 co-localised more with Rab1b than Rab1a, which was also seen in the linear profile plots. The punctate Rab1a structures did not co-localise with PRAF2 or PRAF3, while in Rab1b there was particularly high co-localisation in the ER and partly in the Golgi region.
Figure 4.14: Images and co-localisation between GFP-PRAF and mCherry-Rab1a and Rab1b proteins. Green colour in the merged images represents GFP, red colour represents mCherry and blue colour represents nuclear stain Hoechst 33342.
4.2.5 Interaction network

In an attempt to address the controversy about the function of PRAF proteins their interaction partners were probed. The GFP-tagged PRAF proteins were expressed in HeLa Kyoto cells, immunoprecipitated and the co-precipitated proteins were identified and quantified by mass spectrometry. The immunoprecipitation conditions were chosen to preserve the membranes by using mild non-ionic detergent NP-40. These conditions have been applied in a similar study (Poser et al. 2008). In total, 641 potential binding partners were identified for the three PRAF proteins. The relative protein levels were quantified with the MaxLFQ algorithm, which is based on raw spectral intensities and applies multiple levels of normalisation to ensure that relative levels of proteins are comparable between multiple samples (Cox et al. 2014). The fold increase of LFQ intensity over the negative control (soluble GFP) was calculated and an arbitrary cut-off value of 2-fold increase was used to select the ‘hits’, which yielded 347 proteins. Then the list was filtered by removing the components of the translational machinery, by filtering the keyword "ribosom" and not "mitochon" in the Fasta headers; and components of the proteasome by filtering the keyword “proteasome”, also in the Fasta headers, yielding 323 filtered proteins. The counts of the proteins that were selected as the hits are presented in Figure 4.15.

Figure 4.15: Venn diagram of counts of ‘hits’ for each GFP-PRAF bait.

PRAF1 co-precipitated with 188 unique proteins, PRAF2 with 174 and PRAF3 with 133. The proteins within each of these groups were first annotated by using the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Huang et al. 2009a; Huang et al. 2009b) and clustered with the “high” stringency options. The interactors of all PRAF proteins were enriched in mitochondrial terms (10-, 13- and 5-fold for PRAF1, PRAF2 and PRAF3, respectively). Highly enriched (~10-fold) were also the nucleotide binding terms, which included the mitochondrial RNA-binding proteins and the mitochondrial ATPases. A cluster with actin related terms was enriched 7.5-fold in PRAF1 and 5-fold in PRAF3, which included several subunits of the actin related protein 2/3 complex, myosin motors and tropomyosin-2β. Particularly interesting was a cluster of small GTPases, enriched 4-fold in PRAF1 and PRAF3, which included 9 Rab GTPases (see also Figure 4.16).

A selection of proteins that co-precipitated with PRAFs is shown on Figure 4.16. Among them were 8 Rab proteins from Rab1/Sec4 subfamily which have high sequence similarity: Rab1a, Rab1b and Rab1c, related to trafficking between ER and Golgi; Rab8a, Rab8b, Rab10 and Rab13 involved
in GLUT4 trafficking; and Rab15, associated with early endosomes (Zuk et al. 2000; Zuk et al. 1999). Rab34 does not belong to the above subfamily and co-precipitated only with PRAF2, while Rab1/Sec4 Rabs co-precipitated only with PRAF1 and PRAF3. Few other co-precipitated proteins followed this pattern - PRAF1 and PRAF3 co-precipitated with ITPR3, which stimulates release of intracellular Ca\(^{2+}\), while PRAF2 co-precipitated with two Ca\(^{2+}\) ATPases AT2A2 and TERA. Lipid processing proteins were highly represented among co-precipitated proteins, but did not show any particular pattern. Two mitochondrial translocases TIM44 and TOM20 co-precipitated only with PRAF2, while TIM50 co-precipitated with all three PRAF proteins. A mitochondrial pore protein VDAC1 co-precipitated with PRAF1.

### 4.2.6 Functional studies

PRAF proteins have been previously described to have a role in membrane trafficking. PRAF1 has been proposed to act in the Rab activation cycle (Sivars et al. 2003), but PRAF2 and PRAF3 have not to date been demonstrated to have a similar role. Several functional tests were therefore performed to investigate the function of PRAF proteins in the light of their localisation, co-localisation and possible interaction networks described in the previous sections.

When PRAF1 was overexpressed in cells it altered the structure of the tubular ER network. The transmembrane domains of PRAF proteins form short-hairpin structures, which are similar to those found in reticulon proteins and which play a role in physically structuring ER membranes. PRAF1 overexpression altered the structure of the REEP5-labelled membranes, and there was a clear difference between the transfected and non-transfected cells (Figure 4.17). Overexpression of the other PRAF proteins did not have an observable effect on the structure of the ER tubules.
**Figure 4.17:** Overexpression of PRAF-GFP proteins and immunostaining of REEP5 and CLIMP-63. The first column shows the merged images. Green colour represents GFP channel, red colour represents the REEP5 staining and blue colour represents the nuclear stain Hoechst 33342. Yellow dots mark the cells overexpressing PRAF1. White squares mark the zoomed regions. The columns 2, 3 and 4, show the zoomed region of each individual channel.

**Figure 4.18:** Relative PRAF mRNA levels after siRNA transfection. HeLa Kyoto cells were transfected with two pooled siRNA molecules for 48 or 72 h against each cognate gene. The mRNA levels were quantified by RT-qPCR. The relative expression was calculated as a ratio of the gene expression of each PRAF between negative control (siNEG)-treated cells and cells treated with each siPRAF. The error bars represent the standard error of means (N=2). P-values were calculated with ANOVA and Tukey post-hoc tests. **p<0.01, ***p<0.001.
The overexpression of PRAF2 did show a change in the intensity of the CLIMP-63 labelled structures, although the effect was not consistent.

The effect of depletion of PRAF proteins on the structure of the ER was therefore also investigated. PRAF proteins were depleted by the transfection of two pooled siRNAs for 48 and 72 h, which reduced the transcript levels by >90% after 48 h and ~70% after 72 h (Figure 4.18). Due to the latency between RNA depletion and subsequent protein reduction, 72 h siRNA transfections were used in all subsequent experiments.

The effect of the PRAF depletion on cell morphology and ER structure is shown in Figure 4.19. The cells immunostained with the tubular ER marker REEP5 and the sheet ER marker CLIMP-63 were imaged on a laser scanning confocal microscope and then analysed (Figure 4.19). The tubular ER in negative control cells was ~15% larger than the sheet ER. The sheet ER area was significantly (~15%) increased after depletion of PRAF1 or PRAF3, while the other treatments did not show significant changes (Figure 4.20).

Chemical fixation of the cells and staining of the ER tubules with antibodies did not preserve the structure of the ER sufficiently well to describe the morphology of the ER (see Figure 4.17).
Figure 4.20: Areas of tubular and sheet ER in cells depleted for PRAF proteins. Areas were measured in $\mu m^2$. The error bars represent the standard errors of means (N=151-187). P-values between the corresponding negative control (siNEG) and each treatment were calculated by ANOVA and Tukey post-hoc tests. ** p<0.01.

Figure 4.21: Diagram of the ER-CFP construct. A signal peptide (SP) is at the N-termini, followed by a CFP and a C-terminal KDEL ER retention motif.

Figure 4.22: Intensity of the ER-CFP fluorescence signal in cells depleted for PRAF1. The intensity was measured with ImageJ software by manually segmenting the cells, subtracting the background from the measurements and calculating the mean intensity per $\mu m^2$. The error bars represent the standard errors of means (N=8). The p-value was calculated by t-test. *** p<0.001.

The morphology of the ER was therefore inspected in live cells by expressing the construct with N-terminal signal peptide fused with CFP and KDEL C-terminal ER retention motif (Figure 4.21). The ER-CFP construct was expressed for 24 h and imaged using a fluorescence microscope. The intensity of ER fluorescence in cells depleted for PRAF1 was visually lower than the intensity in negative control cells, which was confirmed by intensity measurements made in ImageJ (Figure 4.22). The morphology of the peripheral tubules appeared normal after PRAF1 depletion (Figure 4.24). There was no observable alteration of the structure, size or morphology of the tubules. Also, the tripartite junctions appeared normal, although they could only be seen in the peripheral areas of the cell.
The co-precipitation and mass spectrometry experiments described earlier identified a significant number of mitochondrial proteins co-precipitating with PRAF - the cluster of mitochondrial proteins was enriched 10-fold. The ER forms inter-organelle contacts with the mitochondria, which are required for the exchange of lipids, proteins, ions and other molecules. Many biocatalytic pathways, such as lipid synthesis, are also coupled between the ER and mitochondria. Indeed, many proteins that co-precipitated with PRAFs, were related to lipid synthesis. The effect of PRAF depletion on the morphology of the mitochondria was therefore investigated. A custom image analysis pipeline was developed, as shown in Figure 4.23, to measure the width and the length of the mitochondria. The wide-field images of mitochondria in live cells were filtered with a rolling ball algorithm, which is used to remove the background by creating a smoothed version of the image and subtracting it from the original image. In this image analysis pipeline a ball with the radius 1 pixel (the lowest limit for this algorithm) was used, which essentially worked as an edge detector. The processed image was then thresholded with the Otsu method (Otsu 1979), which resulted in a binary mask of the mitochondria. The binary mask was used to detect and measure the particles. The width was measured as the minimal ferret radius. To measure the length, the binary mask was further processed with the skeletonise algorithm, which 'digests' the particle from all sides until only 1 pixel wide line in the middle of the particle is left, spanning be-
Figure 4.24: The ER and mitochondria in live cells following either depletion or overexpression of PRAF1. a) PRAF1 was depleted in HeLa Kyoto cells by siRNA treatment for 72h and then ER-CFP was overexpressed for 24h. The mitochondria were labelled with MitoTracker®. The cells were imaged live on a wide-field fluorescence microscope. The merged image shows both channels, ER-CFP in green colour and MitoTracker® in red colour. b) PRAF1-GFP was overexpressed in HeLa Kyoto cells for 24h and the mitochondria were labelled with MitoTracker®. The cells were imaged live on a wide-field fluorescence microscope. The merged image shows PRAF1-GFP in green colour and MitoTracker® in red colour. Yellow dots mark the cells overexpressing PRAF1.

between both termini of the particle and thus represents the skeleton of the particle. The length of the skeleton of each particle was measured as the length of the mitochondria. The data points in the top or bottom 5th percentile were removed as outliers.

The overexpression of PRAF1 visually disrupted the morphology of the mitochondria (Figure 4.24b). In the non-transfected cells they appeared as ~5.5 µm thick fibres, slightly curved and unbranched, with a mean length of 17.6 µm (Figure 4.25a and 4.25b). In the PRAF1-overexpressing cells the mitochondria appeared shorter, curved and localised more towards the cell periphery (Figure 4.24b). They were significantly narrower, 5.3 µm wide and significantly shorter than in control cells, the mean length was 16.1 µm (Figure 4.25a and 4.25b).

Next, the effect of PRAF1 depletion on the morphology of the mitochondria was tested. The morphology appeared less disrupted than if PRAF1 was overexpressed. The main difference to the
negative control cells was swelling at the foci of the mitochondria. Visually they looked elongated, thicker and less branched than in negative control cells. Some cells showed more disrupted phenotypes, containing large spherical structures rather than fibres (Figure 4.24a). The thickness was significantly increased, from 5.3 µm in the negative control cells to 5.9 µm in the PRAF1-depleted cells (Figure 4.25c and 4.25d). The length was also significantly increased from 17.3 µm in the negative control cells to 18.7 µm in the PRAF1-depleted cells (Figure 4.25c and 4.25d).

The overexpression of PRAF1 reorganised the ER tubules; they appeared thinner and denser. The depletion of PRAF1 did not show any visual difference in the ER structure, but rather increased the area of the sheet ER by 15%. Dramatic effects were seen in the morphology of the mitochondria: the overexpression of PRAF1 reduced the length and the thickness of the mitochondria, while the depletion resulted in the opposite phenotype: the mitochondria were longer and thicker.

4.3 Discussion

In this study, PRAF proteins were systematically characterised for their localisation, membrane topology and function. First, N- or C-terminally GFP-tagged constructs of three PRAF proteins were designed. The constructs were then expressed in human cells and their localisation was determined by interpreting the images. All three PRAF proteins clearly localised to the ER. While PRAF2 and PRAF3 have been previously reported to localise to the ER (Schweneker et al. 2005; Abdul-Ghani et al. 2001; Liu et al. 2008b), this is not the case for PRAF1, which was reported to localise primarily to the Golgi (Abdul-Ghani et al. 2001; Gougeon et al. 2002; Liang et al. 2004b; Alvim Kamei et al. 2008; Bhagatji et al. 2010; Geng et al. 2005). Although the line plots in Figures 4.10 and 4.11 show localisation to the Golgi region, this is not obvious from the images. Co-localisation was also quantified by RWC and it showed low levels of co-localisation between the Golgi marker GM130 and PRAF proteins. The localisation was investigated in HeLa Kyoto cells, which are not ideal for study of ER proteins. A flatter cell type, such as COS-7, might provide a better insight into the morphology of the ER following modulation of the levels of the PRAF proteins.
ER localisation of transmembrane proteins is maintained by ER retention motifs. ER retention motifs were predicted in the protein sequences of PRAF2 and PRAF3, but not in PRAF1. The role of the predicted ER retention motifs on the localisation and presence of other localisation sequence determinants were investigated by designing truncated versions of the PRAF proteins. All mutants were retained in the ER, suggesting that there is probably another mechanism of ER localisation for these proteins. As all soluble regions were removed in the different truncation combinations, ER localisation is therefore probably maintained by the transmembrane domains themselves, structured into a hairpin. Short hairpin organisation of transmembrane domains can be found in reticulons and indeed these domains alone are sufficient for ER localisation of reticulons (Shibata et al. 2008). Although this evidence would support the retention of hairpin-like proteins in the ER, the deletion mutants could also be retained in the ER because of misfolding. The folding accuracy would need to be experimentally tested, for example by the induction of point mutations.

The ER has multiple structural subdomains, therefore co-localisation with subdomain markers was investigated. Quantification of co-localisation with RWC algorithm revealed that all PRAF proteins co-localise ~2.5-fold more with tubular ER structures than with the sheet ER. Individual image interpretation and analysis demonstrated that PRAF proteins primarily localise to the tubular ER structures and also in some regions to the sheet ER.

Co-localisation with other compartments was determined by co-expressing PRAF and mCherry-tagged Rab proteins, commonly used as organelle markers. The highest co-localisation value was seen with Rab13, which displays a broad localisation profile, including a significant localisation to the plasma membrane (Ioannou et al. 2015). Both the ER and plasma membrane are distributed throughout the cell, but due to resolution limits of the light microscope they would appear to co-localise, and therefore the significance of Rab13 with respect to PRAF function is unclear. Perhaps more significantly, the PRAFs were seen to strongly co-localise with early secretory Rab proteins, in particular Rab1a and Rab1b. Specifically, PRAF2 and PRAF3 co-localised more strongly with Rab1b, whereas PRAF1 showed preference for Rab1a. Inspection of the images revealed that Rab1 proteins concentrate in the Golgi area and faintly in the ER. PRAF proteins have a stronger signal in the ER and do not specifically localise to the Golgi. The high co-localisation coefficient measured is likely a consequence of strong Rab signal in Golgi area. Both Rabs have an overlapping function in the ER/Golgi compartment and similar localisation profiles. Low co-localisation was observed between PRAF1 and endocytic Rabs Rab5a, Rab7a and Rab9a. This was surprising, as PRAF1 was originally identified as a GDF for the above Rabs and the original report also claimed its localisation to endosomes (Sivars et al. 2003). Furthermore, none of the above Rabs were found among the proteins co-precipitating with the PRAFs.

Overexpression of GFP- and HA-tagged PRAF proteins revealed unusual phenotypes at the ER. The most obvious effect seen was reorganisation of the tubular ER in the presence of overexpressed PRAF1. The reorganisation was specific for REEP5- and not for CLIMP-63-labelled structures. The ER appeared disrupted after the fixation and permeabilisation, therefore an alternative approach with milder fixation and permeabilisation conditions should be used, such as PFA/GA combination for fixation and saponin instead of Triton-X100 for permeabilisation. Also, a flatter cell line, such as COS-7, might be better suited for the inspection of ER morphology.
The overexpression of different combinations of PRAF proteins also induced formation of undefined compartments of different sizes close to the nucleus. With PRAF1 they appeared as large structures, while in PRAF2 overexpressing cells there were many smaller ones, resembling lipid droplets (see Figure 1f in Arruda et al. 2014 or Figure 4b4 in Beller et al. 2006). Aggregates were formed in the presence of several combinations of GFP- and HA-PRAFs and also in mutants PRAF1bc, PRAF3ab and PRAF3bc.

Further insight of PRAF protein function was gained determining interaction networks of the proteins. Although PRAF proteins were previously reported to play a role in membrane trafficking, few components of membrane trafficking machinery were found among the interacting proteins. Among the identified Rab proteins were Rab1a and Rab1b, which are involved in trafficking between ER and Golgi, although they also individually function in other compartments. For example, a small population of Rab1b was identified in MAM during virus infection (Horner et al. 2015).

Many co-precipitated proteins were related to mitochondria, or more precisely, transport of metabolites to and from mitochondria. Among them were two translocases across inner membrane TIM44 and TIM50. They both facilitate translocation of proteins across the inner membrane by recognising a presequence in the proteins (Schulz et al. 2014; Kronidou et al. 1994). It has also been shown that the C-terminus of TIM44 interacts with acidic phospholipids, particularly ones containing cardiolipin (Marom et al. 2009). TIM50 interacts with 3-β-HSD II enzyme to promote synthesis of progesterone and androstenedione (Thomas et al. 2014). A mitochondrial channel VDAC1 was found among the proteins that co-precipitated with PRAF1. VDAC1 is a permeability pore protein present at contact sites between the ER and mitochondria and facilitates transport of Ca2+ between the ER and mitochondria (Szabadkai et al. 2006). Transport of Ca2+ depends on two other proteins, SERCA and ITPR3. SERCA is a membrane bound Ca2+ activated ATPase and transports Ca2+ between the ER and cytosol (MacLennan et al. 1985). ITPR3 is a Ca2+ channel in the ER membrane (Miura et al. 2007; Bernhardt et al. 1996; Clapp et al. 2001). Opening of ITPR3 activates VDAC1 and allows entry of Ca2+ ions (Gincel et al. 2001; Rapizzi et al. 2002). Besides Ca2+ ions, VDAC1 also mediates translocation of other metabolites through the outer mitochondrial membrane by forming an ATP microdomain close to SERCA in the ER (Vendelin et al. 2004). Interestingly, VDAC1 has an alternative isoform that localises to the ER (Buettner et al. 2000). Creatine kinase (CKMT1) interacts with VDAC1 and co-localises with both VDAC1 and SERCA (Schlattner et al. 2001). Creatine kinase facilitates transport of lipids across membranes. Together with NDPK it facilitates direct exchange of phospholipids between membranes mimicking ER-mitochondria contacts sites (these membranes are ~30% enriched in cardiolipin) (Epand et al. 2007). Carnitine-O-palmitoyltransferase 1 (CPT1A), located at the outer mitochondrial membrane, also interacts with VDAC1 and facilitates the primary regulatory step in fatty acid oxidation. It also interacts with ACSL3, a protein involved in synthesis of lipids (discussed below) (Lee et al. 2011b).

The second group of interactors was related to lipid metabolism and is closely related to mitochondria, because lipid intermediates are exchanged back and forth between ER and mitochondria during synthesis, degradation or conversion. Fatty-acid CoA-ligase ACSL3 is localised to lipid droplets and has transmembrane domains organised into hairpins. It is probably required for local synthesis of lipids (Poppelreuther et al. 2012). ACSL3 was also required for membrane syn-
thesis after infection with polio virus. In non-infected cells the lipids were transported to lipid droplets (Nchoutmboube et al. 2013). Other proteins involved in lipid metabolism were: acetyl-CoA acetyltransferase (ACAT), involved in cholesterol metabolism (Lei et al. 2010; Kamikawa et al. 2014); lysophosphatidylcholine acyltransferase 1 (LPCAT1) that synthesises phosphatidylcholine (Moessinger et al. 2011); and fatty acid binding protein 5 (FABP5) (Baumeier et al. 2015).

A group of interactors was also associated with oxidation reactions and thus involved in oxidation of lipids. Increased activity of a mitochondrial protein leucine rich PPR-motif containing protein (LRPPRC) induced oxidative phosphorylation of $\beta$-fatty acids and changed the morphology of mitochondria (Liu et al. 2011b). Cytochrome B and cytochrome C are peroxidases catalysing oxidation reactions, including oxidation of lipids (Tolkatchev et al. 1996; Tappel 1953). They require Cu$^{2+}$ ions, which are supplied by SCO1 homolog protein, localised to the inner mitochondrial membrane (Glerum et al. 1996; Buchwald et al. 1991; Stiburek et al. 2006).

PRAFs co-precipitated with several proteins from the outer and inner mitochondrial membrane, suggesting that they are in contact with mitochondria. Indeed, depletion of PRAF1 (clearly an ER-localising protein) increased the size of mitochondria, whereas PRAF1 overexpression reduced it. Overall, a significant majority of the proteins that co-precipitated with PRAFs seem to be related to ER-mitochondria contacts and lipid metabolism. They could therefore play a role in the recruitment of lipid synthesis machinery and also the machinery needed for translocation of nascent lipids between ER and mitochondria. The effect of PRAF proteins on the morphology of mitochondria could be a consequence of disruptions in lipid transport between ER and mitochondria, which would also affect the levels of reactive oxygen species in the mitochondria. Further functional experiments are therefore needed to validate a role for the PRAFs at the ER-mitochondria interface.
Chapter 5

Characterisation of YIPF proteins
5.1 Introduction

The YIPF family of proteins has 7 members with several structural similarities: they have a hydrophobic region in the C-terminal part of the protein and consists of 5 closely stacked TMDs; they also have a 100–150 residue long soluble N-terminal domain. The TMDs are linked with 3–10 residue long soluble linkers, which likely bring all the TMDs close together. The hydrophobic region of YIPF1, YIPF2 and YIPF3 is followed by a shorter, <50 residue long, soluble C-terminal domain, while the other YIPF proteins end with the last TMD.

Previous studies have reported localisation of these proteins to the Golgi, which would be consistent with a role in membrane trafficking. Within the Golgi they apparently localise to different subdomains: overexpression of FLAG- and HA-tagged YIPF constructs showed that YIPF3 and YIPF4 co-localised with the cis-Golgi marker GM130, while the others were distinct from GM130, therefore probably localising to other regions of the Golgi stack (Shakoori et al. 2003). The localisation of YIPF3 and YIPF4 to the cis-Golgi has also been confirmed by antibody staining of the endogenous proteins (Tanimoto et al. 2011). The overexpressed FLAG- and HA-tagged YIPF1, YIPF2, YIPF5 and YIPF6 also have been observed to localise to the ER, but the reticular pattern was faint (Shakoori et al. 2003). The ER localisation was persistent after cycloheximide treatment, therefore it probably reflected a steady state localisation and was not a transit state (Shakoori et al. 2003). YIPF5 also co-localised with ERGIC-53, as shown with antibody staining (Tang et al. 2001; Yoshida et al. 2008); and overexpressed FLAG- and HA-tagged YIPF1 and YIPF2 have been localised to punctate structures and the plasma membrane (Shakoori et al. 2003).

YIPF proteins are ubiquitously expressed in human tissues, although their expression is the highest in testis, which indicates they could be involved in spermatogenesis. They are highly expressed in tissues with high exocytic activity, such as liver, pancreas and prostate (Shakoori et al. 2003).

The most attention has been paid to YIPF5, particularly for its role in structuring the ER. Two conflicting reports have been published on this theme: Dykstra et al. reported on whorling of ER after depletion of YIPF5, supported with electron microscopy images (Dykstra et al. 2010); Kano et al. on the other hand did not observe any change in ER structure after YIPF5 depletion, but rather reported disruption of Rab6a-dependent retrograde trafficking (Kano et al. 2009). The link between YIPF5 and ER structure could be associated to the ER structuring protein REEP5, which was found to interact with YIPF5 (Dykstra et al. 2010). YIPF proteins also have been suggested to play a role in maintaining the structure of the Golgi. Cells depleted of YIPF3, YIPF4 or YIPF5 had fragmented Golgi, although this was observed in only 5% of the cell population (Yang et al. 1998; Tanimoto et al. 2011; Tang et al. 2001; Dykstra et al. 2010), again therefore making interpretation difficult.

Yeast orthologs of YIPF proteins interact with Rab proteins. Yip1p, a yeast ortholog of YIPF5, interacts with Ypt1, which itself is an ortholog of Rab1 (Calero et al. 2001; Barrowman et al. 2003). Interactions with Golgi Rabs Ypt31, Ypt32 and Ypt6 have also been reported (Chen et al. 2004), and clearly interactions with Rab proteins would further indicate involvement of the family in membrane trafficking processes. The depletion of YIPF5 disrupted trafficking of VSV-G and various other secreted proteins (Tang et al. 2001). YIPF5 was seen to be recruited to COPII vesicles (Tang
et al. 2001); and blocking its activity with antibodies prevented fusion of COPII vesicles with their target membrane. The fusion was disrupted only in COPII vesicles that bud after the antibody had been added, indicating that YIPF5 is important for the recruitment of the fusion machinery to COPII vesicles during the budding process (Barrowman et al. 2003).

In this chapter a systematic and detailed localisation analysis of all YIPF proteins, by overexpression GFP-tagged YIPF proteins in HeLa Kyoto cells, was carried out. The localisation of YIPFs to Golgi sub-domains was determined by quantitatively measuring co-localisation with Golgi markers TGN46 and GM130, labelling cis- and trans-Golgi elements, respectively. This systematic study also reports on the likely membrane topology and interaction network of all 7 YIPF proteins.

5.2 Results

5.2.1 Design of GFP-tagged YIPF constructs

In order to investigate the localisation and function of YIPF proteins, a library of GFP-tagged YIPF constructs was designed. Source clones (see Appendix A) were amplified by PCR by using custom primers with overhanging restriction sites. The amplification products were ligated into GFP vectors pEGFP-C1 and pEGFP-N1 to create a set of both N- and C-terminally tagged fusion constructs. Competent cells were transformed with the ligation reactions and grown in selective media. Typically five colonies were tested for presence of the ligation product by diagnostic digestion reactions: the plasmids were isolated from the bacterial cells, digested with the restriction enzymes used in the cloning reactions and separated on agarose gels. The diagnostic digestion reaction mix separated into two DNA fragments: a long fragment of ~4700 bp corresponding to the vector backbone; and a short fragment of 700–1000 bp corresponding to the insert. The presence of both fragments indicated a successful ligation reaction. The sequence of one colony of each YIPF fusion was verified by sequencing and the verified clone was used in further experiments.

In total 7 N-terminally GFP-tagged and 7 C-terminally GFP-tagged wild type YIPF constructs were generated. A source clone for human YIPF7 was not available, therefore a source clone for the mouse (Mus musculus) ortholog of YIPF7 was used. Mouse YIPF7 is 254 residues long and is 74% identical to the 280 residues long human YIPF7. All further experiments were performed with the mouse ortholog of YIPF7. Figure 5.1 shows the agarose gels of the diagnostic digestion reactions of the 14 constructs.

5.2.2 Localisation

The localisation of YIPF proteins was determined by expressing the GFP-tagged YIPF constructs in HeLa Kyoto cells for 48 h and imaging with a fluorescence microscope. All YIPF constructs primarily localised to the Golgi (Figure 5.2), in addition to a number of other cellular structures. YIPF1, YIPF2, YIPF5 and YIPF7 localised to punctate structures. The punctate structures were more concentrated around the Golgi, although they were found in peripheral regions as well. They were particularly abundant with the YIPF5 construct. Some small punctate structures were observed in YIPF3 expressing cells as well. Unlike other YIPF proteins, YIPF4 and YIPF6 did not localise to punctate structures. All constructs also produced a faint reticular pattern in the background,
which resembled an ER localisation. The reticular pattern was strong in YIPF5, YIPF6 and YIPF7 expressing cells. N-terminally tagged YIPF7 had a smaller number of punctate structures than the C-terminally tagged version of the protein. Otherwise, no major differences between the N- and C-terminally tagged versions of the YIPF constructs could be seen.

Live cells over-expressing the GFP-tagged YIPF proteins were then fixed with PFA, permeabilised and stained with YIPF specific antibodies. An antibody specific for YIPF7 was not available, therefore YIPF7-GFP expressing cells were stained with anti-YIPF5 antibody, as YIPF7 is a known homolog of YIPF5 (Tang et al. 2001; Barone et al. 2015). The antibodies stained all or part of the GFP-tagged YIPF proteins as well as the endogenous YIPF proteins (Figure 5.3, 5.4 and 5.5). A significant amount of background staining was seen in all images, likely representing some non-specific binding of the antibodies. This was particularly evident with YIPF5 and YIPF6 antibodies (Figure 5.4 and 5.3), which made it difficult to identify the localisation. Overall, the antibodies did recognise the overexpressed and endogenous YIPF proteins and therefore confirmed the Golgi localisation. YIPF5 showed a Golgi profile that was fainter than in the other YIPF proteins and this phenotype was consistent between the GFP channel and antibody staining. N-terminally tagged
YIPF1 and C-terminally tagged YIPF2 localised to punctate structures and larger regions of the Golgi that were not recognised by the antibodies, and likely represent GFP-YIPF aggregates or cleavage products of the above proteins (Figure 5.3). YIPF1 and YIPF2 also localised to the plasma membrane and cell protrusions and these structures were recognised by the antibodies (Figure 5.3, marked with red arrows; zoomed regions in the insets in Figure 5.3). This phenotype was not observed with the other YIPF proteins. The antibodies recognised both Golgi- and ER-localising YIPF1 and YIPF2 proteins. High levels of ER localisation were observed with N-terminally tagged YIPF1 and YIPF7 (Figure 5.3 and 5.5); C-terminally tagged YIPF2, YIPF3, YIPF4 and YIPF5 (Figure 5.3 and 5.4); and both N- and C-terminally tagged YIPF6 (Figure 5.3). Non-transfected cells present in
Figure 5.3: N- and C-terminally GFP-tagged YIPF1, YIPF2 and YIPF6 stained with corresponding anti-YIPF antibodies. GFP-tagged YIPF proteins were expressed in HeLa Kyoto cells for 48 h, fixed with PFA, stained with antibodies and imaged on a laser-scanning confocal microscope. Images of GFP and antibody channels are shown separately. Red arrows mark the localisation of the protein to protrusions. Insets show zoomed regions marked with white boxes. Merged images have three channels: Hoechst33342 nuclear staining in blue, GFP-tagged YIPF protein in green and antibody staining in red. Co-localising regions can be seen as a yellow colour.
Figure 5.4: N- and C-terminally GFP-tagged YIPF3, YIPF4 and YIPF5 stained with corresponding anti-YIPF antibodies. GFP-tagged YIPF proteins were expressed in HeLa Kyoto cells for 48 h, fixed with PFA, stained with antibodies and imaged on a laser-scanning confocal microscope. Images of GFP and antibody channels are shown separately. Merged images have three channels: Hoechst33342 nuclear staining in blue, GFP-tagged YIPF protein in green and antibody staining in red. Co-localising regions can be seen as a yellow colour.
CHAPTER 5. CHARACTERISATION OF YIPF PROTEINS

**Figure 5.5:** N- and C-terminally GFP-tagged YIPF7 stained with anti-YIPF5 antibodies. GFP-tagged YIPF5 was expressed in HeLa Kyoto cells for 48 h, fixed with PFA, stained with antibodies and imaged on a laser-scanning confocal microscope. Images of GFP and antibody channels are shown separately. Merged images have three channels: Hoechst33342 nuclear staining in blue, GFP-tagged YIPF protein in green and antibody staining in red.

The fields of view generally showed a poor signal for YIPF1 and YIPF2 antibody staining; the YIPF2 antibody faintly stained tubular structures resembling the ER (Figure 5.5).

The co-localisation between the GFP-tagged YIPF proteins and antibody staining was next determined and quantified as RWC (Figure 5.6). Both GFP fusion orientations of proteins YIPF1, YIPF2, YIPF3, YIPF4 and N-terminally tagged YIPF6 co-localised to a high level with the antibody staining, with RWC values between 0.73–0.86. Low RWC values observed for YIPF5 and YIPF7 were probably a consequence of high background staining of the YIPF5 antibody and low specificity of the YIPF5 antibody for YIPF7 (see Figure 5.5). The low RWC of C-terminally tagged YIPF6 was probably related to the low expression levels of the GFP-tagged construct and also high background (see Figure 5.3).

A number of the overexpressed YIPF proteins showed the presence of punctate structures, possibly representing protein aggregates, thereby, prompting their biochemical analysis. To investigate this, lysates of cells overexpressing N-terminally GFP-tagged YIPF proteins were separated with SDS-PAGE, subjected to western blotting and probed with anti-GFP antibodies (Fig-
Figure 5.7: SDS-PAGE and western blotting of lysates of cells expressing GFP-YIPF proteins. Predicted molecular weights of full length unmodified fusion proteins are marked with red stars. Reference molecular weights in kDa are shown on the right side of each membrane.

Bands corresponding to the expected sizes of the fusion proteins were identified on the membranes in each case (Figure 5.7, marked with red stars). Fragments corresponding to larger and smaller proteins were also identified, suggesting that the proteins may have been post-translationally modified, or partially degraded, respectively.

N-terminally tagged YIPF1 and C-terminally tagged YIPF2 clearly localised to compartments that were not stained with the corresponding YIPF-specific antibodies. One explanation is that these structures seen in the imaging experiments were truncated fusion proteins, and specifically were truncations having lost the region where the antibody should bind. Indeed, analysis of these cell lysates by SDS-PAGE and western blotting with anti-GFP antibodies showed several bands of different sizes for both fusion proteins (Figure 5.7). The predicted molecular weights of GFP-YIPF1 and GFP-YIPF2 fusion proteins are 62 kDa and 63 kDa respectively (Figure 5.7, bands marked with red stars). Two shorter bands were identified at ~45 kDa and ~55 kDa, which are likely cleavage products of the two proteins. Larger and intense bands with molecular weights ~70 kDa and ~80 kDa were also identified. Larger bands seen in YIPF1, YIPF2 and YIPF3 may represent glycosylation modifications to the ~50 residue long C-terminal loop predicted to reside in the lumen of the endomembrane system. Bands with similar molecular weights were previously described for YIPF3 and YIPF4 proteins (Tanimoto et al. 2011). Other YIPF proteins do not have this extended C-terminal tail. These proteins (YIPF4, YIPF5, YIPF6 and YIPF7) were mostly present as single bands at the expected size, although a small number of smaller faint bands, probably cleavage products, were also seen (Figure 5.7). YIPF6 lysates also showed a faint band at ~120 kDa, which could be a multimer, and YIPF7 had a faint band corresponding to an unmodified protein (~54 kDa) and a stronger band at ~58 kDa.

5.2.3 Membrane topology

YIPF proteins have two regions: a long soluble N-terminal domain and hydrophobic C-terminal region. TMDs in the hydrophobic region anchor the protein into the membrane and the topology of the protein in the membrane defines whether the N-terminal region is in the cytosol or in the lumen. However, to date there is no experimental data clearly defining the orientation of this family of proteins. The membrane topology of YIPF proteins was therefore experimentally determined using a FPP assay. GFP-tagged YIPF proteins were expressed in HeLa Kyoto cells for
Figure 5.8: Profiles of GFP digestion in a fluorescence protease protection assay. Images were acquired at 5 s intervals following trypsin addition. Cells in the images were manually segmented and integrated fluorescence intensity was measured in each frame. Relative intensity is a ratio between background-corrected cell intensity and background-corrected intensity of the first frame. The red line represents C-terminally tagged YIPF or SERP1 and the green line represents N-terminally tagged YIPF or SERP1. Error bars indicate standard errors of means (N=5–17).
Figure 5.9: Membrane topology model of YIPF proteins. The long N-terminal region is in the cytosol and the C-terminal region, present only in YIPF1, YIPF2 and YIPF3, is in the lumen of endomembrane system. If the C-terminal region is not present the protein ends after the last transmembrane domain.

Figure 5.10: Quantification of co-localisation of GFP-tagged YIPF proteins with GM130 and TGN46. a) RWC values of co-localisation between GFP-tagged YIPF proteins and the cis-Golgi marker GM130 or trans-Golgi marker TGN46. b) Ratio between RWC values of co-localisation of GFP-tagged YIPF proteins with the cis-Golgi marker and the trans-Golgi marker. A ratio <1.0 indicates higher co-localisation with the trans-Golgi marker; a ratio >1.0 indicates higher co-localisation with the cis-Golgi marker; a ratio of 1.0 indicates the same levels of co-localisation between both markers. The error bars represent standard errors of means (N=9–24).

48 h, the cells were semi-permeabilised with digitonin for 45 s and treated with trypsin. The cells were then imaged with a wide-field fluorescence microscope at 5 s intervals. The positive control was a transmembrane protein, stress-related ER protein 1 (SERP1), which is a 66 residue long protein with a single predicted TMD (Uniprot annotation based on manual sequence analysis). All 7 YIPF proteins responded similarly in the assay: the GFP molecule on the N-terminus was rapidly digested after the addition of trypsin, while the C-terminally located GFP molecule was partially protected, similarly to the positive control (Figure 5.8). This indicates that the N-terminus of YIPF proteins resides in the cytosol, while the C-terminus resides in the lumen of the endomembrane system, predominantly the Golgi. The proposed model of membrane topology of YIPF proteins is shown in Figure 5.9.
Figure 5.11: Linear profiles and images of co-localisation of GFP-tagged YIPF1, YIPF2 and YIPF6 with the cis-Golgi marker GM130 and the trans-Golgi marker TGN46. Nuclear stain Hoechst33342 is shown in blue colour, N-terminally GFP-tagged YIPF proteins in green colour and either GM130 or TGN46 in red colour. White boxes in the first column indicate zoomed regions shown in the second and the third column. White lines mark the regions of the linear profiles.

Table 5.1: Summary of YIPF localisation to subdomains of the Golgi.

<table>
<thead>
<tr>
<th>YIPF protein</th>
<th>Localisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>YIPF1</td>
<td>trans-Golgi</td>
</tr>
<tr>
<td>YIPF2</td>
<td>trans-Golgi and unlabelled area of Golgi</td>
</tr>
<tr>
<td>YIPF6</td>
<td>trans-Golgi</td>
</tr>
<tr>
<td>YIPF3</td>
<td>cis-Golgi</td>
</tr>
<tr>
<td>YIPF4</td>
<td>cis-Golgi</td>
</tr>
<tr>
<td>YIPF5</td>
<td>cis-Golgi and unlabelled area of Golgi</td>
</tr>
<tr>
<td>YIPF7</td>
<td>cis-Golgi</td>
</tr>
</tbody>
</table>

5.2.4 Co-localisation with organelle markers

In an attempt to further refine the localisation of the YIPF proteins within the Golgi, a series of experiments were designed to measure the co-localisation between overexpressed GFP-YIPF proteins and the Golgi markers GM130, labelling the cis-Golgi, and TGN46, labelling the trans-Golgi. The GFP-YIPF proteins were all found to co-localise with both markers, likely because they overlap with each other (all YIPF proteins show extensive Golgi localisation), and because of the limits of resolution of this organelle by light microscopy (Figure 5.10). YIPF1 and YIPF2 co-localised to a greater extent with the trans-Golgi marker than with the cis-Golgi marker (Figure 5.10). Although the difference between the two RWC values was small (the ratio between two RWC values was 0.8
and 0.9 for YIPF1 and YIPF2, respectively) (Figure 5.10b). The co-localisation with the trans-Golgi marker could be better seen on a linear plot (Figure 5.11). YIPF3, YIPF4 and YIPF5 co-localised to a greater extent with the cis-Golgi marker (1.4-fold more) (Figures 5.10b and 5.12). RWC analysis of YIPF6 co-localisation suggested similar levels of co-localisation with both markers (Figure 5.10), but the linear profile (Figure 5.11) revealed a better match with the profile of the trans-Golgi marker. YIPF7 co-localised more with the cis-Golgi marker (Figure 5.10), which was also evident from the linear profile (Figure 5.12). The linear plot of YIPF2 also revealed localisation to unlabelled areas of the Golgi (Figure 5.11). Similarly, YIPF5 partly localised to unlabelled areas, although it mostly co-localised with the cis-Golgi marker (Figure 5.12). YIPF7 co-localised with the cis-Golgi marker, although some smaller peaks on the linear plot matched the peaks of the trans-Golgi marker (Figure 5.12). The localisation of YIPF proteins to subdomains of the Golgi is summarised in Table 5.1.

Figure 5.12: Linear profiles and images of co-localisation of GFP-tagged YIPF3, YIPF4, YIPF5 and YIPF7 with the cis-Golgi marker GM130 and the trans-Golgi marker TGN46. Nuclear stain Hoechst33342 is shown in blue colour, N-terminally GFP-tagged YIPF proteins in green colour and either GM130 or TGN46 in red colour. White boxes in the first column indicate zoomed regions shown in the second and the third column. White lines mark the regions of the linear profiles.
**5.2.5 Golgi morphology**

The membrane topology of YIPF proteins resembles hairpin-like structures present in ER shaping proteins. YIPF5 has been previously associated with ER shaping (Dykstra et al. 2010) and depletion of YIPF3, YIPF4 and YIPF5 altered Golgi morphology (Yang et al. 1998; Tanimoto et al. 2011; Tang et al. 2001; Dykstra et al. 2010). In this study, the role of all YIPF proteins on Golgi morphology was systematically investigated using image analysis techniques. YIPF proteins were depleted in HeLa Kyoto cells using two pooled siRNA sequences and then the Golgi was stained with anti-GM130 antibodies. Images were analysed by automatic thresholding with the Otsu method (Otsu 1979) and area of Golgi fragments was measured.

Depletion of YIPF1, YIPF2 and YIPF6 significantly increased the area of Golgi fragments (Figure 5.13a), although the actual number of fragments detected was no different from that seen in the negative control (Figure 5.13b). Depletion of YIPF3 resulted in an increased number of larger fragments. The area of Golgi fragments after depletion of YIPF4, YIPF5 and YIPF7 did not change or was slightly decreased, whereas the number of fragments was greater.

In cells depleted of YIPF1 and YIPF2 the area of Golgi fragments was significantly larger than in negative control (Figure 5.13a), which was also observed in the individual images. The Golgi in cells depleted of YIPF1 or YIPF2 visually appeared larger with greater spacing between the Golgi elements, although it did not seem fragmented (Figure 5.14). Depletion of YIPF3 caused different levels of fragmentation and in some cells it appeared dispersed. Golgi elements could be seen in some cells, although they had small fragments associated with them. Depletion of YIPF4 had the most dramatic effect: the Golgi showed major fragmentation and in some cells was dispersed around the nucleus. The fragments appeared to be similar in size. The Golgi was less fragmented in YIPF5 depleted cells, although it did display an irregular gross morphology. They seemed thinner, had a 'roughe' shape and some branching could be seen. A similar effect was seen with YIPF7.
depletion. In both treatments the Golgi was shaped as a network of short tubules. Depletion of YIPF6 also deformed the Golgi, in this case it became shaped into rough tubule-like structures which also had some curvature (Figure 5.14). Depletion of YIPF3, YIPF5 and YIPF6 also disrupted the shape of nucleus.

5.2.6 Interaction network

Although localisation analysis can be helpful with respect to predicting where the YIPF proteins function in the cell it does not provide specific information on their interaction network. This was therefore investigated by co-immunoprecipitation of overexpressed N-terminally GFP-tagged YIPF proteins and subsequent identification of co-precipitated proteins using mass spectrometry. The identified proteins were quantified label-free with the MaxLFQ algorithm (Cox et al. 2014) and
Figure 5.15: Fold-changes of selected identified proteins co-immunoprecipitated with N-terminally GFP-tagged YIPF proteins. The colour scale shows the fold-change of an identified protein between levels in the GFP-YIPF bait and the soluble GFP bait. The order of YIPF baits shows clustering of interactomes.
Figure 5.16: Fold-changes of identified cytoskeletal, YIPF, PRAF, Rho GTPases, Rab GTPases and related proteins co-immunoprecipitated with N-terminally GFP-tagged YIPF proteins. The colour scale shows the fold-change of an identified protein between levels in the GFP-YIPF bait and the soluble GFP bait. a) YIPF, Rho proteins and GTPases; and b) cytoskeleton associated proteins that co-immunoprecipitated with YIPF baits with colour-coded fold-change levels. Heatmaps in panels a) and b) have the same colour scale at the bottom of the figure.

the fold-change was calculated for the levels of each identified protein found in the YIPF bait over levels found with soluble GFP bait. A 2-fold change was used as an arbitrary cut-off value to select ‘hits’. In total, 379 unique hits were identified for the 7 YIPF proteins. The hits were then further investigated with the DAVID functional annotation suite. A selection of the identified proteins is shown on a clustered heatmap in Figure 5.15. YIPF3, YIPF4 and YIPF5 showed the most similar interactome, followed by YIPF2 and YIPF6. GFP-tagged YIPF proteins - the baits - were identified in all samples except YIPF7 (Figure 5.16a). Half of the identified proteins were less than 10-fold enriched over the negative control and approximately 10% of the proteins were enriched more than 36-fold.

Functional analysis with DAVID revealed that several functional annotation terms are common to all YIPF proteins. High enrichment was observed for nucleotide and ATP binding terms, which included Ca\(^{2+}\) and Na\(^{+}/K\(^{+}\) transporting ATPases, Rab proteins and myosins. Actin related terms were also enriched with all YIPF proteins and consisted of myosins (MYO), actin related proteins (ARP), tropomyosin (TPM2), dystrophin (DMD), nexilin (NEXN), adducin 1 and 3 (ADDA, ADDG), F-actin capping protein (CAZA2), supervillin (SVIL) and coronin 1B (COR1B) (Figure 5.16b). Another abundant cluster of terms was related to mitochondrial proteins, made mostly of mitochondrial
CHAPTER 5. CHARACTERISATION OF YIPF PROTEINS

transporters, such as transporter inner mitochondrial membrane (TIM50 and TIM44) and voltage dependent anion channel 2 (VDAC2) (Figure 5.15).

Several YIPF proteins co-precipitated with other members of YIPF family (Figure 5.16a). A pattern could be seen between YIPF3, YIPF5 and YIPF7, which co-precipitated together. These three proteins also co-precipitated the protein PRAF3. YIPF4 co-precipitated with YIPF3, which did show similar phenotypes in other assays. Overall, cis-Golgi localising YIPF proteins (cis-Golgi YIPFs) seem to co-precipitate together. Trans-Golgi localising YIPF proteins (trans-Golgi YIPFs) did not co-precipitate together in this experiment, however interactions between YIPF1, YIPF2 and YIPF6 have been previously described.

YIPF1, YIPF5 and YIPF7 co-precipitated with Rab proteins from Rab1/Sec4 subfamily of Rab proteins, which show high sequence similarity between each other. YIPF6 interacted only with Rab34 (Figure 5.16a).

All YIPF proteins co-precipitated with components of the cytoskeleton, such as actin-related proteins (ARPs), tubulin subunits (TBA4A), tropomyosin (TPM2), supervillin (SVIL) and several myosins (MYO). Only YIPF7 co-precipitated with dynactins (DCTNs), RHOG and ROCK2. Kinesin 15 (KIF15) co-precipitated only with YIPF6 (Figure 5.16b). Proteins related to endocytosis, such as adaptor complexes (APs), clathrin interactor 1 (CLINT1 or EPN4), epidermal growth factor receptor pathway substrate 15 (EPS15) and FCH domain only 2 (FCHO2) did not show any preference towards cis- or trans-Golgi YIPFs (Figure 5.15).

Cis-Golgi YIPFs co-precipitated with proteins that are involved in the reorganisation of the cytoskeleton: coflin 1 (COF1), dynactin 2 (DCTN2), angiomotin-like 2 (AMOL2), MTA2, calponin (CNN2) and MICAL3 (Figure 5.15). The second group of highly represented proteins was related to focal adhesions: testin (TES), LIM domain containing protein 1 (LIMD) and laminin beta 3 (LAMB3) (Figure 5.15).

Trans-Golgi YIPFs co-precipitated more with trafficking proteins and cargo. Among the trafficking components were SCAMP3, myeloid leukemia factor 2 (MLF2), motor KIF15 and two cholesterol trafficking proteins apolipoprotein A1 (APOA1) and GLUT1 transporter. Among co-precipitated cargo proteins were α1-antitrypsin and collagen (Figure 5.15).

5.3 Discussion

In this systematic study the localisation of YIPF proteins has been determined by overexpressing GFP-tagged YIPF proteins in HeLa Kyoto cells. YIPF proteins localised to the ER, different regions of the Golgi, and punctate structures in live and fixed cells. YIPF1, YIPF2 and YIPF6 localise more towards the trans-Golgi (trans-Golgi YIPFs) and YIPF3, YIPF4, YIPF5 and YIPF7 localise more towards the cis-Golgi (cis-Golgi YIPFs). YIPF1 and YIPF2 were previously reported as Golgi localising proteins and with cytoplasmic punctate structures seen only in high expressing cells (Shakoori et al. 2003). Figures 5.2, 5.3, 5.4 and 5.5 show that they localise to punctate structures even in low expressing cells. Several of these punctate structures were not stained with antibodies directed against YIPF1 and YIPF2, which suggests that the epitope could be post-translationally modified or cleaved off. GFP-tagged YIPF proteins separated into several bands that were larger than native proteins, which is likely due to post-translational modifications. Also several shorter bands were
visible, which are likely cleavage products. Therefore it is possible that the antigen was modified or cleaved off in a subpopulation of proteins and these structures would not be recognised by the antibodies. Post-translational modifications and cleavage products in YIPF3 and YIPF4 have been previously described (Tanimoto et al. 2011). Staining with YIPF2 antibodies also revealed a faint tubular pattern in the cytoplasm, which could be ER or transport tubules. Linear plots of co-localisation with Golgi markers revealed localisation of YIPF2 to a region of the Golgi that was not enriched in either GM130 or TGN46.

It was previously reported that YIPF3 and YIPF4 localise to the cis-Golgi (Tanimoto et al. 2011; Shakoori et al. 2003) and YIPF5 to the cis-Golgi and ERGIC (Yoshida et al. 2008), so the results presented here would be consistent with this. Co-localisation of YIPF5 with both cis- and trans-Golgi markers was proportionally lower than that seen for YIPF3 and YIPF4, indicating that although it still localises to the Golgi, this is not restricted to just the parts of the Golgi tested here. Linear profile plots also support this. Regions that do not co-localise with any marker can be also seen in the related protein YIPF7. These structures did not appear as a peak on the linear profiles, but were rather smeared regions of higher intensity, which might correspond to ER elements. The RWC value of YIPF6 co-localisation gave inconclusive idea about its localisation, but linear plots clearly show good co-localisation with the trans-Golgi marker tested.

Endogenous YIPF1 and YIPF2 were not detected by the antibodies in non-transfected cells, which indicates either that their expression levels in HeLa Kyoto cells are low or the antibody is poor for immunofluorescence work. The other YIPF proteins were expressed in the cells and the endogenous protein could be clearly visualised with the antibodies. All YIPF proteins showed partial localisation to the ER, which would also be consistent with the observation that the RWC values of co-localisation with Golgi markers were all below 0.60. Other studies also support that a proportion of the YIPF proteins localise to the ER (Shakoori et al. 2003).

YIPF proteins are transmembrane proteins with 5 predicted TMDs, a soluble cytosolic N-terminal loop and a soluble luminal C-terminal loop (in YIPF1, YIPF2 and YIPF3). The orientation of the soluble loops therefore largely determines the nature of interactions as they will be the regions that can interact with cytosolic or luminal proteins. This is the first experimental validation of the membrane topology of YIPF proteins. Previously, assumptions about the membrane topology were based on yeast two-hybrid assays, where activation and DNA binding domains were functional only when they were fused with the N-terminus of the YIPF proteins (Shakoori et al. 2003).

Cellular depletion of the YIPF proteins resulted in various Golgi fragmentation or deformation profiles. It appears that depletion of trans-Golgi YIPF proteins increases the area of the Golgi, but does not have a significant effect on fragmentation. On the other hand, depletion of cis-Golgi YIPFs causes fragmentation of the Golgi into small fragments. YIPF3 was the single exception - its depletion resulted in the Golgi being fragmented into larger elements.

A functional characterisation of the YIPF proteins was initiated through a series of co-precipitation assays to identify interacting proteins. Several cytoskeletal proteins were co-precipitated with all YIPF proteins: actin related proteins (ARPs), tubulin subunits (TBA4A), tropomyosin (TPM2), supervillin (SVIL) and several myosins (MYO). Such ubiquitous association
with these cytoskeleton components could indicate that all YIPF proteins are involved in processes linked to the cytoskeleton.

Overall the interactors were related to cytoskeleton organisation, extracellular matrix (ECM) remodelling, tight junctions, endocytosis and exocytosis. These are very broad descriptions, but a common denominator of these processes is that they are involved in the formation of cellular protrusions in healthy cells or invadopodia in cancer cells. Formation of protrusions involves reorganisation of the cytoskeleton to mechanically extend the plasma membrane. This requires delivery and integration of additional lipids and remodelling of the ECM to allow space for the growing protrusion. This is done by controlling the cell-ECM contacts and secretion of ECM degradation enzymes.

Perhaps more informative for functional characterisation are interaction profiles that are specific for individual YIPF protein or groups of similarly localising YIPF proteins. Trans-Golgi YIPFs were previously found to strongly interact with each other through their TMDs (Shakoori et al. 2003). The proteins that co-precipitated with trans-Golgi YIPF proteins were mostly related to trafficking. Protein SCAMP3 regulates degradation of certain cell surface receptors, such as early growth factor receptor (EGFR) (Falguières et al. 2012), but it is also involved in trafficking of metalloproteases to the forming invadopodia (Eisenach et al. 2012). Invadopodia are cholesterol rich domains (Caldieri et al. 2009) which could explain why a cholesterol transporter apolipoprotein A1 (APOA1) and related glucose transporter GLUT1 (Caliceti et al. 2012) were co-precipitated. This indicates that trans-Golgi YIPF proteins could be involved in trafficking of metabolites to the forming protrusions or recycling of surface proteins from the plasma membrane. Surface plasma membrane proteins, such as integrins, provide bridges between cells and the ECM. Two of the co-precipitated proteins, myeloid leukemia factor 2 (MLF2) and KIF15, are indeed involved in endocytosis or trafficking of integrins (Dave et al. 2014). In the initial steps of protrusion formation the recycling of integrins would be upregulated to loosen the contacts of the cell with the ECM and allow space for the degradation of the ECM. Then, the integrins would be trafficked back to the plasma membrane, together with lipids, which would expand the plasma membrane and attach it to the ECM. Trans-Golgi YIPF proteins localise also to the TGN, which is the major sorting site in the cell, therefore trans-Golgi YIPF proteins could provide the machinery to recognise the cargo that is being trafficked to the forming protrusion. The C-terminal luminal loop would be particularly important for this, as it is the only (long enough) luminal region of YIPF protein sequences and therefore able to interact with the cargo. Several possible cargo proteins have been identified among the co-precipitated proteins, such as α1-antitrypsin, which was related to the formation of invadopodia (Chang et al. 2012) and collagen, a component of ECM.

Cis-Golgi YIPFs had a higher representation of cytoskeleton reorganisation machinery among the co-precipitated proteins. Proteins cofillin 1 (COF1), dynactin 2 (DCTN2), angiomiton-like 2 (AMOL2), MTA2, calponin and MICAL3, amongst others, are all involved in the organisation and stabilisation of the cytoskeleton (Lappalainen et al. 1997; Wang et al. 2011; Barone et al. 2011; Winder et al. 1990; Giridharan et al. 2012). The cytoskeleton shapes the protrusion by mechanically expanding the plasma membrane. The cytoskeleton is also directly linked to the ECM through focal adhesions. Co-precipitated proteins testin (TES), LIM domain containing protein 1 (LIMD) and laminin beta 3 (LAMB3) are either components of focal adhesions or provide links between
CHAPTER 5. CHARACTERISATION OF YIPF PROTEINS

the cytoskeleton and focal adhesions (Coutts et al. 2003; Sharp et al. 2004; Huggins et al. 2008; Nishiuchi et al. 2006). These proteins therefore seem to be likely present at the forming protrusion. This suggests that cis-Golgi YIPF proteins are more likely involved in the organisation of the cytoskeleton. Cis-Golgi YIPF proteins (except YIPF3) lack a C-terminal soluble loop, therefore it is less likely that they would interact with cargo and more likely with the cytoskeleton in the cytosol.

Additional indicators that YIPF proteins could be involved in membrane trafficking are co-purified Rab proteins: Rab8a, Rab8b, Rab13 and Rab15 are linked to apical-basolateral trafficking. Polarised apical-like sorting and trafficking was shown to regulate the formation of invadopodia and degradation of the ECM (Caldieri et al. 2012). In parallel, Rab8a has been implicated for its role in exocytosis of metalloproteases that are required for the degradation of the ECM (Bravo-Cordero et al. 2007). Similarly, Rab8a and Rab8b are involved in the formation of cilia (Yoshimura et al. 2007). Rab34, which is not directly involved in apical-basolateral trafficking, controls the morphology and positioning of lysosomes (Wang et al. 2002), which contain enzymes that are necessary for the degradation of ECM, such as cathepsins. Components of lysosomes have been found to localise to forming podosomes (Tu et al. 2008).

A mass spectrometry quantification of differentially expressed proteins in membrane fractions of superinvasive, paclitaxel selected, clonal populations of the human cancer cell line MDA-MB-435S identified 16 proteins showing statistically significant upregulation (Dowling et al. 2007). Among these were 10 proteins that also co-purified with the YIPF proteins: galectin-3, coflin, ATP synthase beta subunit, voltage-dependent anion channel 2, prohibitin, guanine nucleotide binding protein, ATP synthase beta subunit, disulfide-isomerase A3, enolase 1 and heat shock 70 kDa protein. The evidence from this study suggests that YIPF proteins interact with several proteins related to the formation of cell protrusions or invadopodia in cancer cells.

What might therefore be the relevance of YIPF proteins interacting with proteins that are related to the formation of cell protrusions and are related to cell migration and metastasis? These processes require polymerisation of the actin cytoskeleton and trafficking (and fusion) of membrane vesicles to expand the plasma membrane. Newly formed protrusions also have to adhere to the substrate (i.e. ECM). Although this study shows that YIPF proteins primarily interact with proteins that are required for the above processes, they have been previously related to membrane trafficking (Dykstra et al. 2010; Kano et al. 2009; Tanimoto et al. 2011; Barone et al. 2015). Together, this implies a role in trafficking of cell protrusion components, although this would now need specific investigation. One alternative function could be maintenance of Golgi morphology or positioning, particularly during cell migration. In migrating cells the Golgi is relocated to the front of the nucleus (Chen et al. 2013), which also directs the exocytic pathway towards the leading edge of the migration (Schmoranzer et al. 2003), thus optimising the efficiency of secretory pathway. Golgi morphology was affected after the depletion of several YIPF members, which would provide a link here, but as the experiments were carried out in HeLa Kyoto cells that have been selected for being particularly non-motile an alternative cell system would be needed to explore this. Nevertheless, the detailed localisation and interactome information presented here provides a new resource for exploring the function of the YIPF protein family.
Chapter 6
Classification of cell phenotypes using image texture features
CHAPTER 6. CLASSIFICATION OF CELL PHENOTYPES USING IMAGE TEXTURE FEATURES

6.1 Introduction

Determination of protein localisation mostly relies on manual interpretation of images, which is not feasible in high content screening assays. A straightforward approach to determine the localisation is to use organelle markers and quantify the co-localisation. Such markers could be Rab proteins, which have been previously used in localisation, co-localisation and image analysis studies (Chapter 4, Singan et al. 2012). Alternatively, localisation phenotypes may be classified by image analysis algorithms. The latter approach can be stand-alone (Hamilton et al. 2007; Boland et al. 1998; Boland et al. 1997) or combined with co-localisation data (Singan et al. 2012).

Texture features are a powerful tool for the quantification of phenotypes and have been used previously to describe subcellular localisation. They are abstract features, therefore it is difficult to determine which cellular features are actually being measured. Most commonly used texture features in image classification are the Haralick features, which are based on a grey co-occurrence matrix of an image. An image is a two-dimensional matrix of intensity values. A co-occurrence matrix is calculated by transforming the image matrix into a square matrix with dimensions N×N, with N being the number of grey levels in the image. An element $[i,j]$ of the co-occurrence matrix is the count of how many times a pixel with the intensity $i$ is adjacent to a pixel with the intensity $j$. A pixel can have adjacent pixels in four directions: horizontal, vertical and two diagonals. The Haralick features are represented as three statistics calculated from the co-occurrence matrix (Haralick 1979).

The Gabor texture features are calculated from convolutions of the original image with various wavelet functions (Gabor 1946; Daugman 1985). The feature vector is calculated as the statistics of the convoluted images. Image analysis using the Gabor features is thought to be similar to the image processing in the human visual system and indeed the cells in the visual cortex can be modelled with the Gabor wavelet functions (Marčelja 1980; Daugman 1985).

The threshold adjacency statistics (TAS) features are calculated by thresholding an image. First the mean intensity of the pixels with relative intensity above 0.12 (30 in 8-bit images) is calculated. The images are then thresholded with three different upper and lower limits: mean-0.12 to mean+0.12; mean-0.12 to 1; and mean to 1 (Figure 6.1). This results in a binary mask with the pixels having intensities 0 (background) or 1 (foreground). For each foreground pixel the number of neighbouring foreground pixels is counted, which results in a feature vector of length 9 (Figure 6.1). The TAS features have been successfully used for the classification of cell phenotypes with a support vector machine with 94% accuracy when used alone, and 98% accuracy in combination with the Haralick features (Hamilton et al. 2007).

An interesting set of texture features was developed by Perkin Elmer and are designed to recognise particular cellular features, such as spots, valleys or ridges. The SER (Spots, Edges and Ridges) features are based on the statistics of the second differentials of images. They can be extracted with the Acapella suite (Perkin Elmer), but no open source script is available yet.

The classification of phenotypes is typically based on prior information about the phenotypes, but recently several studies have attempted to use unsupervised classification of the phenotypes. This can be done by extracting the cellular features and generating a model that describes the phenotype with a number. The model can be also used to generate random images of the phenotypes (Murphy 2014; Zhao et al. 2007). Texture features were previously used for unsupervised
Calculate the mean intensity value of pixels with intensity >0.12

Thresholded image:
  lower limit: mean-0.12
  upper limit: mean+0.12

Thresholded image:
  lower limit: mean-0.12
  upper limit: 1.00

Thresholded image:
  lower limit: mean
  upper limit: 1.00

Calculate Threshold Adjacency Statistics
Count the number of neighbours of each pixel and normalise to the total number of pixels

Figure 6.1: The threshold adjacency statistics (TAS) texture features. The mean of the image pixels with relative intensity above 0.12 is calculated and the image is then thresholded with three strategies: mean-0.12 to mean+0.12; mean-0.12 to 1; and mean to 1. The number of foreground neighbours of each foreground (white) pixel is counted and the mean count for each neighbour profile is calculated. The resulting feature vector has 9 values.

classification of the localisations of selected Rab proteins associated with the Golgi (Singan et al. 2012).

The aim of this study was to explore the unsupervised classification of localisation phenotypes in the early secretory pathway, in particular the ER, Golgi and punctate structures, and in the context of the localisation profiles of the PRAF and YIPF protein families.
6.2 Results

6.2.1 Construction of localisation dataset

The dataset used in this study was composed of the images of cells expressing 10 proteins from the PRAF and YIPF families, localising to different compartments within the cell. PRAF proteins predominantly localise to the ER and YIPF proteins to the Golgi. PRAF proteins show 2 different phenotypes: PRAF2 and PRAF3 localise to ER tubules, which results in a clear reticular pattern; PRAF1, when overexpressed, induces reorganisation of the tubular ER and therefore shows a less defined reticular pattern (see Chapter 4). Overexpressed YIPF proteins localise to Golgi after 48 hours, but at shorter expression times (24 hours) they also localise to the ER and punctate structures. Both protein families are interesting because they localise to a single compartment or they show mixed localisation phenotypes in the same population of cells. For example, PRAF2 and PRAF3 localise only to the ER, whereas YIPF4 localises to the ER, Golgi or both (see Figure 6.2).

The N-terminally GFP-tagged PRAF and YIPF proteins were expressed in HeLa Kyoto cells for 24 hours, fixed, stained with Hoechst 33342 and imaged on a laser-scanning confocal microscope with a 12.5 μs/pixel dwell time, no zoom and at a resolution 1024 by 1024 pixels. Twelve fields-of-view were acquired and at least 50 cells were imaged with 106.7 cells on average per class. Each image had two channels, one for the nuclear stain and one for the GFP channel. 120 dual-channel images were acquired in total.

The cells exhibited different localisation phenotypes (Figure 6.2 and Table 6.2). PRAF2 and PRAF3 localised to the ER with a clearly defined reticular pattern. PRAF1 also localised to the ER, but had a less defined reticulum and sometimes formed punctate structures. YIPF1 localised to punctate structures in the Golgi area. YIPF2 also localised to punctate structures, but they were dispersed throughout the cell and an ER background could also be seen. YIPF3 had a localisation similar to YIPF2 but with an even stronger ER background. YIPF4 and YIPF5 localised to the ER, Golgi and punctate structures. YIPF6 and YIPF7 localised to the ER with a clear reticular pattern and also a small number of punctate structures could be seen.

6.2.2 Development of the image analysis pipeline

The images were processed with CellProfiler software, which is designed for batch processing and is easily adaptable for use with custom plugins developed in the Python programming language. CellProfiler has built-in analysis modules for segmentation of an image, intensity measurements and measurements of Haralick and Gabor texture features, amongst others. A module to extract TAS features was not part of the package and was therefore developed based on the specifications from Hamilton et al. 2007. The module code is given in Appendix B.

The images were processed in multiple steps (Figure 6.3). First, the images were loaded into CellProfiler and the two channels were annotated as the DAPI channel for the nuclear stain and the GFP channel for the GFP-tagged protein. Then the images were segmented by first detecting the nuclei in the DAPI channel. The nuclei were then used as seeds to detect the cell area in the GFP channel. Automatic segmentation of the cytoplasm was good for PRAF1, PRAF2, PRAF3, YIPF4, YIPF5, YIPF6 and YIPF7, but not for YIPF1, YIPF2 and YIPF3, because of the low intensity of
CHAPTER 6. CLASSIFICATION OF CELL PHENOTYPES USING IMAGE TEXTURE FEATURES

**Figure 6.2:** The localisation profiles of PRAF and YIPF proteins. The N-terminally GFP-tagged PRAF and YIPF proteins were expressed in HeLa Kyoto cells for 24 hours, fixed and imaged on a laser-scanning confocal microscope. PRAF1, PRAF2 and PRAF3 localised to the ER; YIPF1, YIPF2 and YIPF3 localised to punctate structures, with varying intensities of ER also visible; YIPF4 and YIPF5 had a mixed phenotype composed of ER background, strong Golgi and punctate structures; YIPF6 and YIPF7 also localised to the ER and some punctate structures.

the ER background and punctate phenotype. For these proteins the cytoplasmic area was determined by expanding the nuclei by 50 pixels, which sufficiently covered the area of the cytoplasm. The segmented cells were then filtered according to mean intensity value and area. These filters excluded cells containing too many saturated pixel values, which would result in inaccurate measurements, or cells that were very large, likely as a result of poor segmentation. The filtered populations of cells were then used to measure 80 texture features. Among the texture features were 52 Haralick features, 1 Gabor feature and 27 TAS features (Table 6.1). Haralick and Gabor features were measured at a scale of 5 pixels. Haralick features were measured in 4 angles: 0, 45, 90 and 135 degrees. The features were represented as a statistics of each feature in the cell and resulted in a single numerical value per cell.
Segment cells
1) Nuclei detection
2) Cytoplasm detection
Automatic for: PRAF1, PRAF2, PRAF3, YIPF4, YIPF5, YIPF6, YIPF7
or expanded nuclei by 50 px for: YIPF1, YIPF2, YIPF3

Filter cells
1) Remove border objects
2) Area
   between 5000-50000 px²
3) Intensity
   mean intensity between 0.05 and 0.7

Measurements
1) Haralick texture fts
2) TAS texture fts
3) Gabor fts

Figure 6.3: The image analysis pipeline. The cells were first segmented by detecting the nuclei in the DAPI channel and then using the nuclei as seeds for the detection of the cytoplasm, either automatically or by expanding the nuclei by 50 pixels. The objects on the borders of the image were removed, as well as the cells that were very large or contained too many saturated pixels. Three types of texture features (fts) were then extracted from the filtered populations of the cells.

<table>
<thead>
<tr>
<th>TAS features</th>
<th>Haralick features</th>
<th>Gabor features</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAS - Low to High (9 fts)</td>
<td>Angular Second Moment (4 fts)</td>
<td>Gabor Wavelets (1 ft)</td>
</tr>
<tr>
<td>TAS - Low to Max (9 fts)</td>
<td>Contrast (4 fts)</td>
<td></td>
</tr>
<tr>
<td>TAS - Mean to Max (9 fts)</td>
<td>Correlation (4 fts)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Difference Entropy (4 fts)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Difference Variance (4 fts)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Entropy (4 fts)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Info Meas 1 (4 fts)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Info Meas 2 (4 fts)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inverse Difference Moment (4 fts)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sum Average (4 fts)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sum Entropy (4 fts)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sum Variance (4 fts)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Variance (4 fts)</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.1: List of texture features (fts) measured in the images. Three types of TAS features were extracted (see Figure 6.1) and each type had 9 features. 13 types of Haralick features were extracted, each in 4 directions: horizontal, vertical and 2 diagonals. A single Gabor feature was calculated.

6.2.3 Unsupervised clustering of texture features

The texture features were investigated for the ability to cluster the cellular phenotypes by similarity. Principal components analysis was first used to reduce the dimensionality and visually explore the data (Figure 6.4). The data were first scaled to zero mean and unit variance. This was an important step, because unequal variance can result in biased clustering. Next the principal components analysis was performed and the first four principal components were plotted in different combinations (Figure 6.5).

The plot of principal components 1 and 2 divided the cells into two major clusters. The cluster on the right corresponded to YIPF1, YIPF2 and YIPF3, which all exhibited a localisation of a faint ER background and punctate structures. The second cluster contained the remaining cells. It could be further divided into the cells exhibiting an ER localisation: PRAF2, PRAF3, YIPF6, PRAF1 and YIPF7; and the cells exhibiting ER, Golgi and punctate structure localisations: YIPF4 and YIPF5. There appeared to be a linear correlation within the larger cluster in the following order: PRAF2 - PRAF3.
CHAPTER 6. CLASSIFICATION OF CELL PHENOTYPES USING IMAGE TEXTURE FEATURES

- YIPF6 - PRAF1 - YIPF7 - YIPF4 - YIPF5. PRAF2 and PRAF3 had a very well defined ER pattern. YIPF6 and YIPF7 had a well defined ER pattern and some punctate structures. The last two classes in the large cluster, YIPF4 and YIPF5, had the least defined ER phenotype. The first principal component had the highest loading of the feature SumEntropy at an angle of 45 degrees (-0.14); and the second principal component had the highest loading for the feature TAS - Low to Max 2 (-0.24). In general, TAS features had higher loadings in the second principal component.

The plot of principal component 1 and principal component 4 showed a linear correlation within the cluster of YIPF1, YIPF2 and YIPF3. A similar linear relationship within the above cluster was also seen in the plot of principal component 2 and principal component 4. The other principal components did not show any meaningful patterns.

An unsupervised cluster analysis was then performed on the scaled dataset to examine the ability of the algorithm to group the phenotypes by similarity. Clustering based on Gaussian models was used in this case and the analysis was done in the R programming language with the module Mclust (Fraley et al. 2012). Model-based clustering has several advantages over the traditional clustering methods, such as k-means clustering: model-based clustering also calculates the probability that a particular member is a part of the cluster, while the traditional clustering methods simply classify the dataset into the clusters; model-based clustering provides Bayesian Information Criterion (BIC) statistics to determine the optimal number of clusters; and model-based clustering works on non-normalised data (Magidson et al. 2002). The optimal model and the number of clusters was determined by testing 10 different Gaussian models pre-defined in the Mclust package and calculating the Bayesian Information Criterion (BIC) for each model and for maximum 9 clusters (Figure 6.6). The optimal model for the clustering of the dataset was VVV - ellipsoidal, with varying volume, shape, and orientation - for 3 clusters (Figure 6.6). The dataset was then clustered with these parameters and the classification was compared to the original protein labels. The proportions of cluster assignments were then calculated for each protein (Figure 6.7). The model based clustering resembled the clustering done by the visual interpretation (the plot of PC1 and PC2 in Figure 6.5). The proteins showing clear ER localisation (PRAF1, PRAF2, PRAF3, YIPF6 and YIPF7) were strongly represented in cluster 2, the proteins showing punctate structures and low ER background (YIPF1, YIPF2 and YIPF3) were strongly represented in cluster 3, and the proteins with mixed localisations of ER, Golgi and punctate structures (YIPF4 and YIPF5) were strongly represented in cluster 1. Cluster 1 thus represented the punctate structures phe-
Figure 6.5: Plots of the first four principal components of the texture features. The plots show the combinations of the first four principal components (PC). The points were plotted and then labelled with different colours to represent each GFP-tagged protein. The circular points represent the individual cells and the diamonds represent the mean cell.
Figure 6.6: Bayesian Information Criterion (BIC) plot for the determination of the optimal model and the optimal number of clusters. The BIC of 10 models are plotted in different colours for up to 9 clusters. The optimal model was VVV, which had the highest BIC for 3 clusters.

Figure 6.7: Unsupervised clustering of the texture features. The representation of the clusters for each protein is represented as a pie chart with the three clusters plotted in different colours.
CHAPTER 6. CLASSIFICATION OF CELL PHENOTYPES USING IMAGE TEXTURE FEATURES

<table>
<thead>
<tr>
<th>Protein</th>
<th>Localisation</th>
<th>Majority cluster representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRAF1</td>
<td>ER</td>
<td>Cluster 2</td>
</tr>
<tr>
<td>PRAF2</td>
<td>ER</td>
<td>Cluster 2</td>
</tr>
<tr>
<td>PRAF3</td>
<td>ER</td>
<td>Cluster 2</td>
</tr>
<tr>
<td>YIPF1</td>
<td>Punctate structures</td>
<td>Cluster 3</td>
</tr>
<tr>
<td>YIPF2</td>
<td>Punctate structures, faint ER</td>
<td>Cluster 3</td>
</tr>
<tr>
<td>YIPF3</td>
<td>Punctate structures, faint ER</td>
<td>Cluster 3</td>
</tr>
<tr>
<td>YIPF4</td>
<td>ER, Golgi, punctate structures</td>
<td>Cluster 1</td>
</tr>
<tr>
<td>YIPF5</td>
<td>ER, Golgi, punctate structures</td>
<td>Cluster 1</td>
</tr>
<tr>
<td>YIPF6</td>
<td>ER, some punctate structures</td>
<td>Cluster 2</td>
</tr>
<tr>
<td>YIPF7</td>
<td>ER, some punctate structures</td>
<td>Cluster 2</td>
</tr>
</tbody>
</table>

Table 6.2: A comparison of the visually interpreted localisation class and the classification based on the unsupervised model-based clustering.

notype, cluster 2 represented the phenotype of ER, Golgi and punctate structures, and cluster 3 represented the ER phenotype (Table 6.2).

The clustering also informs about the numbers of different localisation phenotypes in the cells. For example, PRAF2 had the cleanest reticular phenotype and had a very high representation in cluster 2. YIPF4 and YIPF5 have a mixed localisation phenotype, represented in cluster 1. YIPF6 localised to punctate structures and the ER, therefore it was classified into either cluster 1 with YIPF4 and YIPF5 or cluster 2 with PRAF2. A similar pattern could be seen with cluster 3: YIPF1 showed the strongest punctate phenotype and no ER background, while YIPF2 and YIPF3 had an increasing amount of ER background, resulting in higher representations in clusters 1 and 2 compared to YIPF1.

6.3 Discussion

In this study texture features were used to describe localisation phenotypes and perform an unsupervised classification of images of cells. The image dataset used comprised the PRAF and YIPF protein families, showing similar localisation phenotypes: ER, punctate structures, Golgi or different combinations of each. Texture features can be relatively easily extracted with the open source image analysis software CellProfiler. The software already includes a module for Haralick and Gabor features, although it does not provide a module for TAS features. Therefore, a custom module to extract TAS features was developed based on previously reported specifications (Hamilton et al. 2007).

The images were segmented automatically, which worked well for the cells with a strong ER background, but not for the cells with punctate structures and faint background. Instead, these cells were segmented by expanding the area of the nucleus by 50 pixels, which covered most of the punctate structures and worked better than the automatic segmentation for these localisation phenotypes.
The texture features were then used for clustering of the phenotypes. First, the whole dataset was scaled by subtracting the mean value of each feature and dividing each feature by the standard deviation. Scaling is desirable in both supervised and unsupervised clustering approaches, because the differences in the ranges of features could introduce bias towards the features with higher scales. It is important for principal components analysis and traditional clustering methods, such as k-means, but it does not affect model based clustering, which is an advantage over the traditional clustering methods (Magidson et al. 2002). To keep the analysis consistent the data were always scaled in the same way. The principal components analysis was then performed and the first four principal components were plotted. The most meaningful was the plot of principal components 1 and 2, which clearly separated the dataset into two major clusters. The larger cluster containing the ER localising proteins also showed a linear relationship, which approximately followed the strength of ER phenotype in the cell. The final two localisation phenotypes were shown by YIPF4 and YIPF5, which have a mixed phenotype with ER, Golgi and punctate structures visible. This part of the cluster could be visually separated from the rest of the cluster, but there was still some overlap between the both parts.

The original scaled texture features were then clustered based on the Gaussian model VVV, which has three variable parameters: volume, shape and orientation (Fraley et al. 2012). The localisation phenotypes were clustered into three clusters, which resembled the visual clustering described above. The localisations showing punctate structures represented the majority of cluster 3, the ER localisation phenotypes were the majority in cluster 2, and the mixed localisation phenotypes were the majority in cluster 1. PRAF proteins, YIPF6 and YIPF7 were strongly represented in the ER cluster 2, but some cells expressing YIPF6 or YIPF7 were also classified as having the mixed localisation phenotype. A similar pattern was observed among YIPF2 and YIPF3, which showed a stronger ER background than YIPF1. In particular YIPF proteins exhibited a variety of phenotypes: well-defined ER, mixed phenotypes, Golgi or punctate structures. Interestingly, the proportion of cells with the particular phenotypes correlated with the abundance of the individual localisation phenotype in the cell, although there is unlikely to be any causative relationship between these observations. For example, in a single cell YIPF6 localises to the ER and punctate structures and the population of YIPF6 expressing cells was classified into cluster 2 (ER) and cluster 1 (punctate structures). The model itself cannot discriminate individual phenotypes in a single cell, therefore this correlation cannot have a causative link.

This short study demonstrated that texture features can be used to cluster phenotypes without any prior knowledge. Although the texture features itself are abstract and it is difficult to determine what exactly is being quantified, they can be combined with the other cellular metrics to improve the description of the phenotypes. Recently much effort has been put into unsupervised cytometry. While this study demonstrates the potential of unsupervised classification of cellular phenotypes based on texture features, they are probably more useful on a subcellular scale, for example to identify individual organelles. Texture features have been used to describe the texture in generative models in other recent studies (Buck et al. 2012; Zhao et al. 2007). The generative models are an interesting approach to analyse phenotypes. Numerous high content screening datasets could be used to build models of the phenotypes and link them with non-imaging resources. This would provide a new insight into old data, providing phenotype information and
visualisation to other researchers as a resource, or even identifying previously unknown links be-
 tween phenotypes, as was recently demonstrated in a study that described novel regulators of
cell morphology by modelling the descriptors of cell shape (Yin et al. 2013; Sailem et al. 2014).
Chapter 7

General discussion and future prospects
After a protein has been synthesised it can only function properly after it is translocated to its site of function. As such, understanding the subcellular localisation of a protein is the initial and among the most valuable descriptor of its basic function (Simpson et al. 2003). Several attempts have been made to systematically (and experimentally) investigate the localisation of the entire human proteome or ‘localisome’ (Simpson et al. 2000; Uhlén et al. 2015); and these approaches have also been invaluable in terms of providing new tools, such as antibodies or fluorescently-tagged constructs. However, the task of mapping the human localisome is still incomplete, but is required for ultimate in depth functional characterisation of the human proteome. Alternative routes to achieve this, for example using computational predictors of subcellular localisation, are helpful, although they still lag behind and are limited in their predictive power (Meinken et al. 2012). Another problem with large scale localisation studies is how image analysis and interpretation is carried out. Although this field is ahead of sequence-based localisation prediction, it still heavily relies on ground truth data, which are typically provided by a human expert. This study therefore has addressed computational, experimental and image analysis aspects of protein localisation and implemented new image and data analysis pipelines that may be of value in future large-scale studies.

Sequence-based subcellular localisation predictors typically apply machine learning techniques to model the localisation. While this is currently the most successful approach, it has difficulties with finding the exact sequence determinants driving the localisation. As such, an alternative approach was used in this study. The initial aim was to search for SLiMs that are over-represented in a localisation class and would therefore be potential sequence determinants for localisation. The discovery method was based on an experimental localisation dataset of 1165 proteins. Software SLiMFinder implements the algorithm SLiMChance that addresses the evolutionary lability of a SLiM candidate, the size of the search space and the expectedness of the SLiM candidate. It returns a Sig value, which is the probability that the SLiM occurs by chance (Edwards et al. 2007). The discovery of novel SLiMs in the localisation groups returned several candidate motifs, however the Sig values showed with high probability that they are false positives. The search was repeated within disordered regions of the proteins, but with the same outcome. The search was also repeated in the topological groups, because localisation mechanisms also depend on the topology of a protein. SLiMs were also specifically searched for in the terminal regions of proteins, particularly in the proteins where the localisation was affected by the position (N- versus C-terminal) of the GFP tag. Neither of these approaches returned strong and significant SLiM candidates.

SLiMs are short and are thus weak localisation signals. They often work in synergy with other SLiMs or stronger localisation mechanisms, such as domain interactions. It would be expected however, that a sequence determinant of localisation would be enriched in a certain localisation class. Several SLiM candidates did show borderline Sig values and were therefore further evaluated by calculating the enrichment factor between the number of instances in the relevant localisation class versus the rest of the dataset. The enrichment factors were low for certain known localisation motifs, such as the nuclear localisation signal, but several candidate motifs were indeed enriched in their localisation class. One convincing novel SLiM candidate was F..M.AE, which had a relatively low Sig value (0.006) and a significant 10-fold enrichment among Golgi-localising
proteins. It was found in coatamer subunit beta and Rab1b, two important components of COP1-mediated trafficking. This SLiM was therefore experimentally investigated by synthesising DNA encoding a peptide with this SLiM and fusing it with GFP. When the construct was transfected into cells it showed a cytoplasmic localisation profile, suggesting that this motif alone could not be a SLiM important for localisation. Although this general experimental approach does work for certain motifs, for example C-terminal ER retention motifs, a better approach may be to mutate the SLiM sequence in the original protein and then investigate the localisation, i.e. look for loss of localisation. SLiMs also rely on other mechanisms and may in fact be needed for fine-tuning the localisation, in which case other localisation determinants would need to be present, besides the SLiM alone, to successfully localise the protein. While mutating a SLiM in the original sequence may be a better approach, it is still not a feasible technique when a large number of validations needs to be performed. Therefore it is important to identify which SLiM predictions are true positives and perform validations only on promising candidates. This can be done by estimating the false discovery rate by validating a certain number of candidates (10 or 20) and determining the number of true positives. The Sig values could be then corrected for false discovery rate, using the Benjamini & Hochberg method (Benjamini et al. 1995), and only strong predictions would be experimentally validated. Discovery of novel SLiMs is a difficult research area, because SLiMs are weak signals and are often lost ‘in the noise’. Therefore a major improvement would be the development of computational techniques that would better separate ‘the signal from the noise’. One such approach could be the use of randomisation techniques. There is also a need for an optimised validation pipeline. High throughput mutagenesis has been used before, although it still incurs high costs (Gajula et al. 2014; Bill et al. 2014). Currently, ‘hit’ selection based on predictions might still not be sufficiently accurate to be worthwhile investing in experimental approaches.

Another consideration is that the localisation profiles in the original dataset used here were interpreted manually, which gives a description of a phenotype that is based on previous knowledge and experience, but it also introduces certain bias and human error. Any method that would automatically cluster localisation phenotypes may be less biased therefore, and one such approach is reported in Chapter 6. The analysis pipeline is based on texture features only, which are abstract descriptors of the image and are difficult, if not impossible, to interpret in terms of biological meaning. Three types of features were measured - Haralick, Gabor and TAS - measuring a total of 80 features. The features were first visualised by reducing the dimensionality with PCA and plotting the first 4 principal components in all combinations. Obvious clustering of the phenotypes was observed in the plot of PC1 and PC2, with punctate phenotypes clearly being separated from ER phenotypes. The phenotypes were then automatically clustered with a model-based clustering algorithm. This effectively clustered the phenotypes into 3 clusters, which corresponded with manually interpreted phenotypes. This demonstrates that texture features can be used to automatically cluster the phenotypes into localisation classes, although no explicit information about the localisation phenotype is previously provided (or could be extracted after). Such clusters could be then used in applications where grouping of proteins is desired, such as SLiM discovery, as described in Chapter 3. This is the first report describing unsupervised model-based clustering of cellular phenotypes based on texture features. Previously, texture features were used for supervised clustering of localisation phenotypes and with a relatively high accuracy (>90%) (Hamilton
et al. 2007) or for unsupervised clustering in combination with co-localisation data (Singan et al. 2012). The trends in image analysis are indeed turning towards unsupervised data analysis and the development of new cytometric approaches. Exciting work has been done on generative models of cellular phenotypes, where an arsenal of features is extracted from the cell and then used to build statistical models (Zhao et al. 2007). The models can be then used to generate (draw) similar phenotypes. This approach could be used to analyse cellular phenotypes in HCS. The resulting resource would provide new insight into the old data, and would also enable other researchers to visually explore the differences between the phenotypes on a case-by-case basis. Such a resource would also pave the way towards a general pipeline to describe cellular phenotypes and the integration of data acquired from different experiments.

The localisation of a protein describes the context of its function and provides a general framework for the processes that the protein could be involved in. Chapters 4 and 5 report on the experimental characterisation of the PRAF and YIPF protein families. The characterisation covered their localisation, co-localisation with organelle markers, membrane topology and functional tests. A series of GFP-tagged constructs was designed and used as a basic tool in the characterisation studies. Tagging of proteins with GFP is a commonly used approach, however GFP is a relatively large molecule, which could affect the conformation of the protein to which it is fused and thus affect its interactions. Therefore a set of HA-tagged PRAF proteins was designed to reduce the size of the tag and thus decrease its impact on the function of the protein. All PRAF proteins were found to localise to the ER, localisations which were previously described for PRAF2 and PRAF3 (Schweneker et al. 2005; Abdul-Ghani et al. 2001; Liu et al. 2008a), however PRAF1 was previously determined to primarily localise to the Golgi (Abdul-Ghani et al. 2001; Gougeon et al. 2002; Liang et al. 2004b; Alvim Kamei et al. 2008; Bhagatji et al. 2010; Geng et al. 2005). Such localisation phenotypes were observed repeatedly with GFP- and HA-tagged PRAF proteins in low and high expressing cells. All PRAF proteins localised to ER tubular elements, which could be clearly seen in PRAF2 and PRAF3 expressing cells, but the localisation was less clear for PRAF1. Co-localisation experiments with an ER tubule marker revealed deformation of tubular ER in the cells overexpressing PRAF1. A series of truncation mutants were designed and expressed in cells in order to identify the localisation sequence determinants. Interestingly, none of the mutants changed localisation, which suggests that the transmembrane domains alone are sufficient for ER localisation. PRAF proteins have 4 transmembrane domains organised into pairs, which form hairpin-like structures. This particular organisation is also present in reticulons, which are proteins that organise ER tubules and indeed reticulon transmembrane domains are sufficient to determine their ER localisation (Shibata et al. 2008). Further experiments suggested that PRAF proteins may be involved in lipid metabolism in cells. This is supported by the profile of the proteins that co-precipitated with PRAFs, which are largely related to lipid synthesis and lipid transport between organelles. In addition, overexpression of PRAF1 induced the proliferation of ER tubules, which is one possible consequence of increased lipid synthesis (Young et al. 1971). Lipid biosynthesis is crucial for cell proliferation and is also related to cancer and infectious diseases (Baenke et al. 2013; Carette et al. 2000). Biosynthetic pathways involve both the ER and mitochondria, with lipid precursors being translocated back and forth during the process (Osman et al. 2011). Lipids are directly translocated between ER and mitochondrial membranes and are not trafficked, which indicates the importance of di-
rec contacts between the organelles. Although PRAF proteins localise to the ER, there were a large number of mitochondrial proteins among the proteins that co-precipitated with the PRAF proteins. In addition, increased levels of PRAF1 reduced the size of mitochondria whereas PRAF1 depletion increased it. One possible hypothesis is that PRAF proteins help localise the lipid biosynthesis machinery at the interface between the ER and mitochondria. Lipid biosynthesis is of high importance for cell proliferation and this would provide a mechanism to efficiently increase the supply of lipids when desired. The effect of depletion of PRAF proteins on cell proliferation could be determined in proliferation assays, but more advanced assays, such as incorporation of heavy C isotopes or lipid imaging, would provide better insight into the underlying mechanisms.

A similar strategy was applied to characterise YIPF proteins. They are all primarily Golgi proteins, although several also show some ER localisation. Two subdomain localisation groups were discovered by quantifying the co-localisation between GFP-tagged YIPF proteins and markers of Golgi subdomains. Proteins YIPF1, YIPF2 and YIPF6 co-localised to a greater extent with a trans-Golgi marker, whereas YIPF3, YIPF4, YIPF5 and YIPF7 co-localised more with a cis-Golgi marker.

The majority of proteins that co-precipitated with YIPFs are involved in cytoskeleton reorganisation, which is a process involved in cell motility, particularly relevant for the invasion of cancer cells. Indeed, 10 out of 16 proteins that were upregulated in super-invasive clone of the MDA-MB-435S cell line also co-precipitated with YIPF proteins (Dowling et al. 2007). The formation of cellular protrusions and cell invasion requires reorganisation of the cytoskeleton and also expansion of cellular membrane, which activates lipid biosynthesis and trafficking of the synthesised lipids towards the forming protrusion at the plasma membrane. A possible hypothesis is that YIPF proteins mediate the above trafficking processes and are selectively activated when a cell needs to deliver metabolites to the protrusions. Indeed, YIPF proteins have been previously associated with membrane trafficking (Dykstra et al. 2010; Kano et al. 2009; Tanimoto et al. 2011; Barone et al. 2015), so this hypothesis seems feasible. However, cell migration assays would be needed to validate this. Several YIPF proteins were present in punctate structures. While such structures could be an artefact of overexpression, they might as well be related to the function of the protein. Co-localisation with organelle markers, such as Rab proteins, would provide more information about their origin.

Curiously, despite the YIPF and PRAF families not being related, there is a possible overlapping function, namely that migrating cells have greater demand for lipids. PRAF proteins could therefore mediate the synthesis of lipids, while the YIPF proteins would assist in their trafficking to the cell surface. This is a curious connection, further endorsed by the fact that the only known overlapping binding partners (in yeast at least) are Rab proteins.

The evidence provided in this study also does not add further support to a role for PRAF1 as a GDI displacement factor (Sivars et al. 2003). According to Sivars et al. PRAF1 should interact with Rab9a, Rab5a and Rab7a, but in this study none of these Rabs were found among PRAF1-co-precipitated proteins. In addition, the study by Sivars also reported that PRAF1 does not interact with Rab1a, which contradicts several other studies (Bucci et al. 1999; Martincic et al. 1997), including this. This study also reports that PRAF1 only poorly co-localises with Rab9a or Rab7a, and to a much lower extent than other proteins, including Rab1a. It would be appropriate to revisit
the role of PRAF1 in the Rab activation cycle with Rabs associated with the endosomal/lysosomal system.

In summary, this study addressed computational and experimental aspects of the subcellular localisation of proteins. The computational aspect included attempts to discover novel targeting SLiMs, and potential problems associated with it, particularly low signal-to-noise ratio. Although the outcome was unsuccessful this still may be a valid approach in terms of prediction of subcellular localisation. Up until now only a small number of sequence determinants for localisation are known and with improved methodologies the identification of novel determinants should be possible. A detection methodology with a better signal-to-noise ratio would be a major improvement in this field.

The second computational aspect was based on texture features and their power to cluster the proteins based on their localisation profiles in an unsupervised way. This study demonstrated that texture features are a powerful tool for the classification of localisation phenotypes. The trends in image analysis in cell biology are heading towards cytometry, development of new features with biological significance and modelling of phenotypes based on this. With these new features it will be easier to integrate different datasets, which will also provide a useful resource for the other scientists.

The experimental part of this study focused on the characterisation of two protein families that previously have not been investigated thoroughly. The primary aim was to define the localisation and to investigate the function of the proteins. The findings of this study provide new insight into these protein families and suggest potential functions. PRAF and YIPF family proteins may play a role in lipid biosynthesis and cell migration, respectively, and ultimately may be relevant to cancer and other pathologies associated with cell proliferation.

Overall, subcellular localisation does provide important information linked to the function of a protein. As we head towards the milestone of obtaining a complete human localisome, exciting developments will undoubtedly be seen in imaging and computational technologies and methodologies. In particular, cytometry based on image analysis together with machine learning techniques will boost the amount of information that we can extract from images. This will also enable us to form phenotypical networks, similar to interaction networks, and integrate other data sources to get an even more complete overview over what is happening inside the cell.
Appendix A: Sequences and primers

PRAF and YIPF source clones

Table A.1: List of genes and source DNA sequences (GenomeCube, Source Bioscience) used in cloning.

<table>
<thead>
<tr>
<th>HGNC gene name</th>
<th>Clone identifier</th>
<th>GenBank accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRAF1</td>
<td>IRAUp969G1130D</td>
<td>BC008950</td>
</tr>
<tr>
<td>PRAF2</td>
<td>IRAUp969H0651D</td>
<td>BC021213</td>
</tr>
<tr>
<td>PRAF3</td>
<td>IRAUp969E0615D</td>
<td>BC005143</td>
</tr>
<tr>
<td>YIPF1</td>
<td>DKFZp761P7010</td>
<td>AL713668</td>
</tr>
<tr>
<td>YIPF2</td>
<td>IRAUp969A1249D</td>
<td>BC013014</td>
</tr>
<tr>
<td>YIPF3</td>
<td>IRAUp969F0848D</td>
<td>BC019297</td>
</tr>
<tr>
<td>YIPF4</td>
<td>IRAUp969E1124D</td>
<td>BC004875</td>
</tr>
<tr>
<td>YIPF5</td>
<td>IRAUp970A1142D</td>
<td>BC024737</td>
</tr>
<tr>
<td>YIPF6</td>
<td>IRAUp970F1221D</td>
<td>BC012469</td>
</tr>
<tr>
<td>YIPF7 (Mus Musculus *)</td>
<td>IRCp5011E051D</td>
<td>BC089576</td>
</tr>
</tbody>
</table>

*No Homo sapiens cDNA was available for this gene.
### Sequences of PRAF and YIPF constructs primers

Table A.2: List of primers used for design of PRAF and YIPF constructs and corresponding sequences.

#### pEGFP-N1 - C-terminally tagged GFP constructs

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>5Xho1PRAF1</td>
<td>GCACTACGAGGTATGGCAGCGCAGAAGGAC</td>
</tr>
<tr>
<td>5Xho1PRAF2</td>
<td>GCACTACGAGGTATGGCGGCTG</td>
</tr>
<tr>
<td>5Xho1PRAF3</td>
<td>GCACTACGAGGTATGGACGTTAATATCGCC</td>
</tr>
<tr>
<td>5Xho1YIPF1</td>
<td>GCACTACGAGGTATGGCAGCAGTAGATGAC</td>
</tr>
<tr>
<td>5Xho1YIPF2</td>
<td>GCACTACGAGGTATGGCCATCGGCCGACGAG</td>
</tr>
<tr>
<td>5Xho1YIPF3</td>
<td>GCACTACGAGGTATGGCAACTACAGCGGCG</td>
</tr>
<tr>
<td>5Xho1YIPF4</td>
<td>GCACTACGAGGTATGGCTAGGCTTTGAAAAC</td>
</tr>
<tr>
<td>5Xho1YIPF5</td>
<td>GCACTACGAGGTATGGCGGAAGCGGAGGAG</td>
</tr>
<tr>
<td>5Xho1YIPF6</td>
<td>GCACTACGAGGTATGGCAACTACAGCGGCG</td>
</tr>
<tr>
<td>5Xho1YIPF7</td>
<td>GCACTACGAGGTATGGCAACTACAGCGGCG</td>
</tr>
<tr>
<td>3Sac2PRAF1N</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2PRAF2N</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2PRAF3N</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2YIPF1N</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2YIPF2N</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2YIPF3N</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2YIPF4N</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2YIPF5N</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2YIPF6N</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2YIPF7N</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
</tbody>
</table>

#### pEGFP-C1 - N-terminally tagged GFP constructs

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>5Xho1PRAF1C</td>
<td>GCACTACGAGGTATGGCAGCGCAGAAGGAC</td>
</tr>
<tr>
<td>5Xho1PRAF2C</td>
<td>GCACTACGAGGTATGGCAGCGCAGAAGGAC</td>
</tr>
<tr>
<td>5Xho1PRAF3C</td>
<td>GCACTACGAGGTATGGCAGCGCAGAAGGAC</td>
</tr>
<tr>
<td>5Xho1YIPF1C</td>
<td>GCACTACGAGGTATGGCAGCGCAGAAGGAC</td>
</tr>
<tr>
<td>5Xho1YIPF2C</td>
<td>GCACTACGAGGTATGGCAGCGCAGAAGGAC</td>
</tr>
<tr>
<td>5Xho1YIPF3C</td>
<td>GCACTACGAGGTATGGCAGCGCAGAAGGAC</td>
</tr>
<tr>
<td>5Xho1YIPF4C</td>
<td>GCACTACGAGGTATGGCAGCGCAGAAGGAC</td>
</tr>
<tr>
<td>5Xho1YIPF5C</td>
<td>GCACTACGAGGTATGGCAGCGCAGAAGGAC</td>
</tr>
<tr>
<td>5Xho1YIPF6C</td>
<td>GCACTACGAGGTATGGCAGCGCAGAAGGAC</td>
</tr>
<tr>
<td>5Xho1YIPF7C</td>
<td>GCACTACGAGGTATGGCAGCGCAGAAGGAC</td>
</tr>
<tr>
<td>3Sac2PRAF1C</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2PRAF2C</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2PRAF3C</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2YIPF1C</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2YIPF2C</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2YIPF3C</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2YIPF4C</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2YIPF5C</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2YIPF6C</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
<tr>
<td>3Sac2YIPF7C</td>
<td>CTCTCTCCGCCGACGCGGCAGTTCTAAGCAG</td>
</tr>
</tbody>
</table>
### APPENDIX A. SEQUENCES AND PRIMERS

#### pCMV-HA - HA-tagged constructs

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>5EcoR1PRAF1</td>
<td>GCACTAGAATTCGTATGGCAGCGCAGAAG</td>
</tr>
<tr>
<td>5EcoR1PRAF2</td>
<td>GCACTAGAATTCGTATGGCAGCGCAGAAG</td>
</tr>
<tr>
<td>5EcoR1PRAF3</td>
<td>GCACTAGAATTCGTATGGCAGCGCAGAAG</td>
</tr>
<tr>
<td>3Xho1PRAF1</td>
<td>CTCTCTCCCGGATTTAGGAGCTCGCTCCAT</td>
</tr>
<tr>
<td>3Xho1PRAF2</td>
<td>CTCTCTCCCGGATTTAGGAGCTCGCTCCAT</td>
</tr>
<tr>
<td>3Xho1PRAF3</td>
<td>CTCTCTCCCGGATTTAGGAGCTCGCTCCAT</td>
</tr>
</tbody>
</table>

#### PRAF deletion mutants

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>5Xho1PRAF1xA</td>
<td>GCACTACTCGAGATGGTGTTCGTGTTCCTG</td>
</tr>
<tr>
<td>5Xho1PRAF2xA</td>
<td>GCACTACTCGAGATGCTCTACTACCAAACC</td>
</tr>
<tr>
<td>5Xho1PRAF3xA</td>
<td>GCACTACTCGAGATGGTAGTGAGCAACCTG</td>
</tr>
<tr>
<td>3Sac2PRAF1AB</td>
<td>CTCTCCCTCCCGGACCCACACCCACACAC</td>
</tr>
<tr>
<td>3Sac2PRAF2AB</td>
<td>CTCTCCCTCCCGGACCCACACCCACACAC</td>
</tr>
<tr>
<td>3Sac2PRAF3AB</td>
<td>CTCTCCCTCCCGGACCCACACCCACACAC</td>
</tr>
<tr>
<td>5Xho1PRAF1BC</td>
<td>GCACTACTCGAGATGCGCACCTTGGAGTCC</td>
</tr>
<tr>
<td>5Xho1PRAF2BC</td>
<td>GCACTACTCGAGATGGAGACCCGCGCAGCT</td>
</tr>
<tr>
<td>5Xho1PRAF3BC</td>
<td>GCACTACTCGAGATGACAGGGTTTGTGG</td>
</tr>
<tr>
<td>3Sac2PRAF1dB</td>
<td>CTCTCCCTCCCGGTCATTAGATAACATCAACAG</td>
</tr>
<tr>
<td>3Sac2PRAF2dB</td>
<td>CTCTCCCTCCCGGTCATTAGATAACATCAACAG</td>
</tr>
<tr>
<td>3Sac2PRAF3dB</td>
<td>CTCTCCCTCCCGGTCATTAGATAACATCAACAG</td>
</tr>
</tbody>
</table>
## DNA sequences of SLiM construct

**Table A.3:** List of sequences used for design of GFP-tagged SLiM construct.

<table>
<thead>
<tr>
<th>Sequence name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-FRQMWAEFE</td>
<td>TCGAGGTTTCGTCAGATGTCGGGCCAATTGAAATAATGACCGC</td>
</tr>
<tr>
<td>3-FRQMWAEFE</td>
<td>GGTCAATTCAAATTCCCACATCTGACGAAACC</td>
</tr>
</tbody>
</table>
Appendix B: Source code of analysis scripts

TAS module - MeasurementTAS.py

'''<b>MeasurementTAS</b> — Measure Threshold adjacency statistics

This module measures the Threshold Adjacency Statistics of an image, object or both.

Please cite the following paper:
Hamilton NA, Pantelic RS, Hanson K, Teasdale RD.: Fast automated cell phenotype image classification.
BMC Bioinformatics 2007, 8:110.

```python
# Import useful Python libraries
import numpy as np
import scipy.ndimage as scind

# Imports from CellProfiler
import cellprofiler.cpimage as cpi
import cellprofiler.cpmodule as cpm
import cellprofiler.measurements as cpmeas
import cellprofiler.objects as cpo
import cellprofiler.settings as cps

from cellprofiler.cpmath.cpmorphology import fixup_scipy_ndimage_result as fix
```

```
# Constants

# I put constants that are used more than once here.

# TAS_LIST are threshold limits
# TAS_WEIGHTS are feature numbers
#
# #################################################################

TAS_LIST = ((('Low', -0.118), ('High', 0.118)), (('Low', -0.118), ('Inf', 1)),
            (('Mean', 0), ('Inf', 0.118)))
TAS_WEIGHTS = range(9)

C_MEASUREMENT_TEMPLATE = "MT"
#
# #################################################################

class MeasurementTAS(cpm.CPModule):
    #################################################################
    #
    # The module class
    #
    # Your module should "inherit" from cellprofiler.cpmodule.CPModule.
    # This means that your module will use the methods from CPModule unless
    # you re-implement them. You can let CPModule do most of the work and
    # implement only what you need.
    #
    #################################################################

    module_name = "MeasurementTAS"
    category = "Measurement"
    variable_revision_number = 1

    #################################################################
    #
    # create_settings is where you declare the user interface elements
    # (the "settings") which the user will use to customize your module
    #
    # You can look at other modules and in cellprofiler.settings for
# settings you can use.
#
###############################################

def create_settings(self):
    # The ImageNameSubscriber "subscribes" to all
    # ImageNameProviders in
    # prior modules. Modules before yours will put images into
    # CellProfiler.
    # The ImageSubscriber gives your user a list of these images
    # which can then be used as inputs in your module.
    #
    self.manual_threshold = cps.Float("Enter threshold:", .118)
    self.input_image_name = cps imageNameSubscriber(
        # The text to the left of the edit box
        "Input image name:",
        # HTML help that gets displayed when the user presses the
        # help button to the right of the edit box
        doc = """This is the image that the module operates on. You
        # can
        # choose any image that is made available by a prior module.
        """
    )
    #
    # The ObjectNameSubscriber is similar – it will ask the user
    # which object to pick from the list of objects provided by
    # upstream modules.
    #
    self.input_object_name = cps.ObjectNameSubscriber(
        "Input object name",
        doc = """These are the objects that the module operates on.
        """
    )
    #
    # The "settings" method tells CellProfiler about the settings you
    # have in your module. CellProfiler uses the list for saving
    # and restoring values for your module when it saves or loads a
    # pipeline file.
    #
    # This module does not have a "visible_settings" method.
    # CellProfiler
    # will use "settings" to make the list of user–interface elements
    # that let the user configure the module. See imagetemplate.py for
    # a template for visible_settings that you can cut and paste here.
    #
    def settings(self):
return [self.input_image_name, self.input_object_name]

# CellProfiler calls "run" on each image set in your pipeline.
# This is where you do the real work.
# def run(self, workspace):
  # Get the measurements object - we put the measurements we
  # make in here
  # meas = workspace.measurements
  
  # Get the input image and object. You need to get the .value
  # because otherwise you'll get the setting object instead of
  # the string name.
  # input_image_name = self.input_image_name.value
  # input_object_name = self.input_object_name.value
  # GETTING AN IMAGE FROM THE IMAGE SET
  # Get the image set. The image set has all of the images in it.
  # The assert statement makes sure that it really is an image
  # set, but, more importantly, it lets my editor do context-sensitive
  # completion for the image set.
  # image_set = workspace.image_set
  
  # Get the input image object. We want a grayscale image here.
  # The image set will convert a color image to a grayscale one
  # and warn the user.
  # input_image = image_set.get_image(input_image_name,
                      must_be_grayscale = True)
  
  # Get the pixels - these are a 2-d Numpy array.
  # pixels = input_image.pixel_data
  
# GETTING THE LABELS MATRIX FROM THE OBJECT SET
# The object set has all of the objects in it.
#
object_set = workspace.object_set
#
# Get objects from the object set. The most useful array in
# the objects is "objects.segmented" which is a labels matrix
# in which each pixel has an integer value.
#
# The value, "0", is reserved for "background" - a pixel with
# a zero value is not in an object. Each object has an object
# number, starting at "1" and each pixel in the object is
# labeled with that number.
#
# The other useful array is "objects.small_removed_segmented"
# which
# is another labels matrix. There are objects that touch the
# edge of
# the image and get filtered out and there are large objects
# that
# get filtered out. Modules like "IdentifySecondaryObjects" may
# want to associate pixels near the objects in the labels
# matrix to
# those objects - the large and touching objects should compete
# with
# the real ones, so you should use "objects.small_removed_
# segmented"
# for those cases.
#
objects = object_set.get_objects(input_object_name)
labels = objects.segmented
#
# The module computes a measurement based on the image
# intensity
# inside an object times a Zernike polynomial inscribed in the
# minimum enclosing circle around the object. The details are
# in the "measure_zernike" function. We call into the function
# with
# an N and M which describe the polynomial.
#
for n, m in TAS_LIST:
    # Compute the TAS for each object, returned in an array
    tas = self.measure_TAS(pixels, labels, n[1], m[1])
    # Add a measurement for this kind of object
for i, t in enumerate(tas):
    feature = self.get_measurement_name(n[0], m[0], i)
    meas.add_measurement(input_object_name, feature, t)

# DISPLAY
# We define is_interactive to be False to tell CellProfiler
# that it should execute "run" in a background thread and then
# execute "display" in a foreground thread.

def is_interactive(self):
    return False

# measure_TAS measures 9 TAS features per object/image

def measure_TAS(self, pixels, labels, n, m):
    # First, get an array that lists the whole range of indexes in
    # the labels matrix.
    #
    if len(labels) == 0:
        n_objects = 0
    else:
        n_objects = np.max(labels)

    if n_objects == 0:
        result = np.zeros((0,))
    else:
        indexes = np.arange(1, np.max(labels)+1, dtype=np.int32)

        # Calculate mean of pixels above int 30
        mean = np.mean(pixels[pixels > 0.118])

    # Set ranges
    rangelow = mean + n
    rangehigh = mean + m

    # Threshold image, create mask
    mask = np.logical_and(pixels < rangehigh, pixels > rangelow)
    thresholded_image = np.zeros(np.shape(pixels))
thresholded_image[mask]=1

# Apply convolution to get the sum of surrounding pixels
# First define a weight array
w=np.array([[1,1,1],[1,0,1],[1,1,1]])
# Create a new array of sums
sums=scind.convolve(thresholded_image,w,mode='constant')
# remove 0 pixels from sums
sums[~mask]=9

# Get the histogram
result = fix(scind.histogram(sums.astype(int),0,9,10,labels=labels, index=indexes))
result = np.vstack(result).T.astype(np.float64)
result = result/np.sum(result,axis=0)
#
return result[0:9]

#########################################################################
# # Here, we go about naming the measurements.
# # Measurement names have parts to them, traditionally separated
# by underbars. There's always a category and a feature name
# and sometimes there are modifiers such as the image that
# was measured or the scale at which it was measured.
# # We have functions that build the names so that we can
# use the same functions in different places.
# def get_feature_name(self, n, m, i=None):
  # ''Return a measurement feature name for the given Zernike''
  # # Something nice and simple for a name...
  # # return "TextureTAS%s%s%s" % (n, m, i)

def get_measurement_name(self, n, m, i=None):
  # ''Return the whole measurement name''
  input_image_name = self.input_image_name.value
  return '_'.join([C_MEASUREMENT_TEMPLATE,
                   self.get_feature_name(n,m,i),
                   input_image_name])
#
# We have to tell CellProfiler about the measurements we produce.
# There are two parts: one that is for database-type modules and one


# that is for the UI. The first part gives a comprehensive list
# of measurement columns produced. The second is more informal and
# tells CellProfiler how to categorize its measurements.
#
#
# get_measurement_columns gets the measurements for use in the
database
# or in a spreadsheet. Some modules need the pipeline because they
# might make measurements of measurements and need those names.
#
def get_measurement_columns(self, pipeline):
    
    # We use a list comprehension here.
    # See http://docs.python.org/tutorial/datastructures.html#list-comprehensions
    # for how this works.
    #
    # The first thing in the list is the object being measured. If
    # it's
    # the whole image, use cpmeas.IMAGE as the name.
    #
    # The second thing is the measurement name.
    #
    # The third thing is the column type. See the COLTYPE constants
    # in measurements.py for what you can use
    #
    input_object_name = self.input_object_name.value
    return [(input_object_name,
             self.get_measurement_name(n[0], m[0], i),
             cpmeas.COLTYPE_FLOAT)
             #for n, m in TAS_LIST for i in TAS_WEIGHTS]
             for n, m in TAS_LIST for i in TAS_WEIGHTS]

#
# get_categories returns a list of the measurement categories
# produced
# by this module. It takes an object name – only return categories
# if the name matches.
#
def get_categories(self, pipeline, object_name):
    if object_name == self.input_object_name:
        return [C_MEASUREMENT_TEMPLATE ]
    else:
        # Don’t forget to return SOMETHING! I do this all the time
        # and CP mysteriously bombs when you use ImageMath
        return []


# Return the feature names if the object_name and category match

```python
def get_measurements(self, pipeline, object_name, category):
    if (object_name == self.input_object_name and
        category == C_MEASUREMENT_TEMPLATE):
        #
        # Use another list comprehension. See docs in get_measurements_columns.
        return [self.get_feature_name(n[0], m[0], i)
                for n, m in TAS_LIST for i in TAS_WEIGHTS]
    else:
        return []
```

# This module makes per-image measurements. That means we need
# get_measurement_images to distinguish measurements made on two
# different images by this module

```python
def get_measurement_images(self, pipeline, object_name, category, measurement):
    if measurement in self.get_measurements(pipeline, object_name, category):
        return [self.input_image_name.value]
    else:
        return []
```
Appendix C: Permissions for reproduction of figures

This is a License Agreement between Tilen Kranjc ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number: 3566531294396
License date: Feb 12, 2015
Licensed content publisher: Nature Publishing Group
Licensed content publication: Nature Reviews Molecular Cell Biology
Licensed content title: Rab GTPases as coordinators of vesicle traffic
Licensed content author: Harald Stenmark
Licensed content date: Aug 1, 2009
Volume number: 10
Issue number: 8
Type of Use: reuse in a dissertation / thesis
Requestor type: academic/educational
Format: print and electronic
Portion: figures/tables/illustrations
Number of figures/tables/illustrations: 1
High-res required: no
Figures: Figure 2 | Localization and function of Rab GTPases.
Author of this NPG article: no
Your reference number: None
Title of your thesis / dissertation: Computational and experimental study of proteins in the mammalian endomembrane system
Expected completion date: May 2015
Estimated size (number of pages): 150
Total: 0.00 EUR
This is a License Agreement between Tilen Kranjc ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

<table>
<thead>
<tr>
<th>License Number</th>
<th>3611001315053</th>
</tr>
</thead>
<tbody>
<tr>
<td>License date</td>
<td>Apr 16, 2015</td>
</tr>
<tr>
<td>Licensed content publisher</td>
<td>Nature Publishing Group</td>
</tr>
<tr>
<td>Licensed content publication</td>
<td>Nature Reviews Molecular Cell Biology</td>
</tr>
<tr>
<td>Licensed content title</td>
<td>SNAREs [mdash] engines for membrane fusion</td>
</tr>
<tr>
<td>Licensed content author</td>
<td>Reinhard Jahn and Richard H. Scheller</td>
</tr>
<tr>
<td>Licensed content date</td>
<td>Aug 16, 2006</td>
</tr>
<tr>
<td>Volume number</td>
<td>7</td>
</tr>
<tr>
<td>Issue number</td>
<td>9</td>
</tr>
<tr>
<td>Type of Use</td>
<td>reuse in a dissertation / thesis</td>
</tr>
<tr>
<td>Requestor type</td>
<td>academic/educational</td>
</tr>
<tr>
<td>Format</td>
<td>print and electronic</td>
</tr>
<tr>
<td>Portion</td>
<td>figures/tables/illustrations</td>
</tr>
<tr>
<td>Number of figures/tables/illustrations</td>
<td>1</td>
</tr>
<tr>
<td>High-res required</td>
<td>no</td>
</tr>
<tr>
<td>Figures</td>
<td>Figure 5</td>
</tr>
<tr>
<td>Author of this NPG article</td>
<td>no</td>
</tr>
<tr>
<td>Your reference number</td>
<td>None</td>
</tr>
<tr>
<td>Title of your thesis / dissertation</td>
<td>Computational and experimental study of proteins in the mammalian endomembrane system</td>
</tr>
<tr>
<td>Expected completion date</td>
<td>May 2015</td>
</tr>
<tr>
<td>Estimated size (number of pages)</td>
<td>150</td>
</tr>
<tr>
<td>Total</td>
<td>0.00 EUR</td>
</tr>
</tbody>
</table>
Bibliography


Caldieri, G., M. Capestrano, K. Bicanova, G. Beznoussenko, M. Baldassarre, and R. Bucci. “Polarised apical-like intracellular sorting and trafficking regulates invadopodia formation and...


BIBLIOGRAPHY


BIBLIOGRAPHY


Fridmann-Sirkis, Y., H. M. Kent, M. J. Lewis, P. R. Evans, and H. R. B. Pelham. “Structural analysis of the interaction between the SNARE Tlg1 and Vps51”. In: Traffic 7.2 (Feb. 2006), pp. 182–190.


Haralick, R. M. “Statistical and structural approaches to texture”. In: *Proc. IEEE* 67.5 (May 1979), pp. 786–804.


Kamikawa, M., X. Lei, Y. Fujiwara, K. Nishitsuji, H. Mizuta, M. Takeya, and N. Sakahita. “ACAT associated Late Endosomes/Lysosomes Significantly Improve Impaired Intracellular Cholesterol Metabolism and the Survival of Niemann-Pick Type C Mice”. In: Acta Histochem. Cytochem. 47.2 (May 2014), pp. 35–43.


BIBLIOGRAPHY


Lee, K., J. Kerner, and C. L. Hoppel. "Mitochondrial carnitine palmitoyltransferase 1a (CPT1a) is part of an outer membrane fatty acid transfer complex". In: J. Biol. Chem. 286.29 (July 2011), pp. 25655–25662.


Pearse, B. M. “Assembly of the mannose-6-phosphate receptor into reconstituted clathrin coats”. In: EMBO J. 4.10 (Oct. 1985), pp. 2457–2460.


Peter, F., C. Nuoffer, S. N. Pind, and W. E. Balch. “Guanine nucleotide dissociation inhibitor is essential for Rab1 function in budding from the endoplasmic reticulum and transport through the Golgi stack”. In: J. Cell Biol. 126.6 (Sept. 1994), pp. 1393–1406.


enhances the transfer of Ca²⁺ microdomains to mitochondria”. In: J. Cell Biol. 159.4 (Nov. 2002), pp. 613–624.


Rapoport, T. A. “Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes”. In: Nature 450.7170 (Nov. 2007), pp. 663–669.


Rink, J., E. Ghigo, Y. Kalaidzidis, and M. Zerial. “Rab conversion as a mechanism of progression from early to late endosomes”. In: Cell 122.5 (Sept. 2005), pp. 735–749.

BIBLIOGRAPHY


Segev, N. “Mediation of the attachment or fusion step in vesicular transport by the GTP-binding Ypt1 protein”. In: Science 252.5012 (June 1991), pp. 1553–1556.

Segev, N., J. Mulholland, and D. Botstein. “The yeast GTP-binding YPT1 protein and a mammalian counterpart are associated with the secretion machinery”. In: Cell 52.6 (Mar. 1988), pp. 915–924.


BIBLIOGRAPHY


BIBLIOGRAPHY


Notes
Notes
Notes