



**University of
Reading**

Defining the Proximal Interactome of Death-Associated Protein Kinase 1

James E. Tomkins

Thesis submitted for the degree of Doctor of Philosophy

School of Pharmacy

November 2019

DECLARATION

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

J E Tomkins

ACKNOWLEDGEMENTS

There are many people who have supported me during this project and I am sincerely grateful to all those involved. It has been a fantastic experience.

First and foremost, I would like to thank Dr Patrick A. Lewis for the opportunity of working on this project and for his excellent project supervision. Patrick has been an inspirational role model to work alongside and I am eternally grateful for his guidance and support throughout the past four years.

Second, I would like to thank Dr Claudia Manzoni for her involvement in the supervision of this project. A considerable part of this work was facilitated by training in bioinformatics that I received from Claudia; her patience, guidance and encouragement throughout this journey are very much appreciated.

In addition, I would like to thank Dr Eva Kevei for the opportunity of collaboration and for introducing me to the world of *C. elegans* science. Her infectious enthusiasm for *C. elegans* and expertise in this area have assisted with guiding the direction of the project.

I am also thankful to all of the Lewis/Kevei lab members and other researchers in the Hopkins and Knight Buildings, past and present, who have been a pleasure to work alongside.

This project was supported by generous funding from the Biotechnology and Biological Sciences Research Council (BBSRC).

My special thanks to Caitlan Onabanjo for her tolerance, support and encouragement, particularly during the process of writing this thesis.

Finally, I would like to express my appreciation to my family and friends, and in particular my parents, Keith and Cari Tomkins, for their continued support and encouragement.

ABSTRACT

Death-associated protein kinase 1 (DAPK1) is a multidomain cell signalling macromolecule which has been implicated in a plethora of biological processes. The domain topology of this protein comprises a calcium/calmodulin dependent serine/threonine kinase, a Ras of complex proteins (ROC) GTPase and a number of further protein-protein interaction (PPI) interfaces. The precise role and regulation of DAPK1 is unclear although it appears to be complex. Due to the potential for targeting this protein for therapeutic intervention, most notably in relation to cancer and neurodegeneration, understanding the physiological function of DAPK1 is important. Developing our understanding of protein function in the wider cellular context can be achieved by defining its proximal interactome.

In this research a PPI network analysis of the human ROCO proteins was performed utilising literature-derived PPI data and novel experimental data to shed light on the commonalities and distinctions within the interaction and functional profiles of these structurally related proteins. This was facilitated by the development of a PPI query resource, termed Protein Interaction Network Online Tool (PINOT), for extracting and processing PPI data from a number of major molecular interaction data repositories. The pursuit of defining the DAPK1 interactome was then translated into the context of the *Caenorhabditis elegans* proteome for predictive and evidence based mapping of the DAPK-1 interactome. This revealed intriguing novel DAPK-1 interactors, MEP-1, SYD-9 and UNC-14, for further investigation. In addition, a number of mutant *dapk-1 C. elegans* strains, FLAG-*dapk-1*, *dapk-1 K57W* and *dapk-1 T715N*, were engineered to further assess the role and regulation of DAPK-1 *in vivo*. Initial phenotypic analysis provided insight into novel DAPK-1 related functions for future examination.

Collectively, the analyses performed and resources developed throughout the course of this research project contribute to our understanding of the DAPK1/DAPK-1 interactomes and will guide future investigation into the complex functionality of this protein.

LIST OF CONTENTS

| | |
|--|-----------|
| Declaration | 2 |
| Acknowledgements | 3 |
| Abstract | 4 |
| List of Contents | 5 |
| List of Figures | 10 |
| List of Tables | 12 |
| Abbreviations | 13 |
| Chapter One - Introduction | 14 |
| 1.1 Investigating protein-protein interactions: the rationale, resources and rewards | 15 |
| 1.2 ROCO proteins | 17 |
| 1.2.1 Human ROCO proteins: structure, regulation and function | 17 |
| 1.3 Death-associated protein kinase 1 (DAPK1) | 19 |
| 1.3.1 Identification, transcription and splicing of DAPK1 | 19 |
| 1.3.2 Structure and regulation of DAPK1 | 20 |
| 1.3.3 DAPK1 function | 22 |
| 1.3.3.1 Cell death and autophagic pathways | 22 |
| 1.3.4 DAPK1 in disease and as a therapeutic target | 23 |
| 1.3.5 Death-associated protein kinase (DAPK) family | 23 |
| 1.4 Caenorhabditis elegans | 24 |
| 1.5 Overall aims | 25 |
| Chapter Two - Evaluating the Human ROCO Protein Interaction Network | 27 |
| 2.1 Introduction | 28 |
| 2.1.1 Project aims | 29 |
| 2.2 Materials and Methods | 30 |
| 2.2.1 Constructing the literature-derived network | 30 |
| 2.2.1.1 Data acquisition | 30 |
| 2.2.1.2 Data formatting and quality control | 31 |
| 2.2.1.3 Interaction detection method reassignment | 32 |
| 2.2.1.4 Interaction confidence scoring | 33 |
| 2.2.1.5 Network output | 34 |
| 2.2.2 Generating the experimental network | 36 |

| | |
|--|-----------|
| 2.2.2.1 Protein production and purification | 36 |
| 2.2.2.2 Protein microarray | 36 |
| 2.2.2.3 Network construction | 37 |
| 2.2.3 Creating the common core network | 37 |
| 2.2.3.1 Coexpression analysis | 38 |
| 2.2.4 Functional enrichment analysis (FEA) | 38 |
| 2.2.4.1 g:Profiler and functional annotation reassignment | 38 |
| 2.2.4.2 Panther and WebGestalt cross-validation | 39 |
| 2.3 Results | 40 |
| 2.3.1 Literature-derived ROCO protein interaction network | 40 |
| 2.3.2 Experimental ROCO protein interaction network | 44 |
| 2.3.3 Common core ROCO protein interaction network | 47 |
| 2.3.4 Enriched functions within the ROCO protein interaction networks | 51 |
| 2.4 Discussion | 54 |
| 2.4.1 The utility of WPPINA for analysis of the human ROCO proteins | 54 |
| 2.4.2 Adding novel insight from ROCO protein microarray data | 55 |
| 2.4.3 Data integration strategies and the common core network | 56 |
| 2.4.4 Gaining functional insight | 57 |
| 2.4.5 Considerations for literature-derived PPI network analysis approaches | 59 |
| 2.4.6 Challenges and ongoing developments: WPPINA to PINOT | 60 |
| 2.4.7 Conclusions | 64 |
| Chapter Three - Assessing the DAPK-1 Protein Interaction Profile in <i>C. elegans</i> | 65 |
| 3.1 Introduction | 66 |
| 3.1.1 Project aims | 68 |
| 3.2 Materials and Methods | 69 |
| 3.2.1 General laboratory materials and media/buffer compositions | 69 |
| 3.2.2 Developing a <i>C. elegans</i> query option for PINOT | 71 |
| 3.2.2.1 Data acquisition and pre-processing | 71 |
| 3.2.2.2 Query submission | 74 |
| 3.2.2.3 Interaction confidence scoring | 74 |
| 3.2.2.4 Network output | 75 |
| 3.2.3 Predicting DAPK-1 interactors | 75 |
| 3.2.4 DAPK-1 yeast two-hybrid screen | 76 |
| 3.2.5 Generating a FLAG- <i>dapk-1</i> <i>C. elegans</i> strain | 77 |
| 3.2.5.1 FLAG- <i>dapk-1</i> sequence design | 77 |
| 3.2.6 FLAG- <i>dapk-1</i> sequence validation | 77 |

| | |
|---|------------|
| 3.2.6.1 <i>C. elegans</i> maintenance | 78 |
| 3.2.6.2 Freezing and recovering <i>C. elegans</i> | 78 |
| 3.2.6.3 Obtaining synchronised life-stage <i>C. elegans</i> populations | 78 |
| 3.2.6.4 DNA extraction, PCR and restriction digestion | 79 |
| 3.2.7 FLAG-DAPK1 protein expression analysis | 81 |
| 3.2.7.1 Protein extraction | 81 |
| 3.2.7.2 Western blot | 81 |
| 3.2.8 FLAG- <i>dapk-1</i> transcript expression analysis | 82 |
| 3.2.8.1 RNA extraction | 82 |
| 3.2.8.2 cDNA synthesis | 83 |
| 3.2.8.3 qPCR | 83 |
| 3.3 Results | 85 |
| 3.3.1 Implementation of <i>C. elegans</i> query option in PINOT | 85 |
| 3.3.1.1 PINOT performance for <i>C. elegans</i> data | 85 |
| 3.3.2 Predicted DAPK-1 protein interaction networks | 87 |
| 3.3.2.1 Predicted DAPK-1 interactors using MIST | 88 |
| 3.3.2.2 Predicted DAPK-1 interactors using Ortholist 2.0 | 90 |
| 3.3.2.3 Incorporation of reported interactors within the predicted DAPK-1 network | 90 |
| 3.3.3 Hybrigenics DAPK-1 yeast two-hybrid screen | 92 |
| 3.3.3.1 <i>dapk-1</i> sequence validation and compatibility | 92 |
| 3.3.3.2 Identification of novel DAPK-1 interactors | 92 |
| 3.3.4 Generation of a FLAG- <i>dapk-1 C. elegans</i> strain | 94 |
| 3.3.4.1 FLAG- <i>dapk-1</i> sequence validation | 94 |
| 3.3.4.2 FLAG-DAPK-1 protein validation | 95 |
| 3.3.4.3 FLAG- <i>dapk-1</i> transcript analysis | 98 |
| 3.4 Discussion | 100 |
| 3.4.1 Capturing <i>C. elegans</i> PPI data in PINOT | 100 |
| 3.4.2 Creating predicted DAPK-1 PPI networks from orthologous inference | 101 |
| 3.4.3 Novel DAPK-1 binding partners | 103 |
| 3.4.4 Validation challenges with the FLAG- <i>dapk-1 C. elegans</i> strain | 105 |
| 3.4.5 Conclusions | 107 |
| Chapter Four - Developing Novel Mutant <i>dapk-1 C. elegans</i> Models | 109 |
| 4.1 Introduction | 110 |
| 4.1.1 Project aims | 113 |
| 4.2 Materials and Methods | 114 |
| 4.2.1 Generating novel <i>dapk-1</i> mutant strains | 114 |

| | |
|--|------------|
| 4.2.1.1 Genomic modifications of the <i>dapk-1</i> K57W and T715N mutant strains | 114 |
| 4.2.2 Validation of <i>dapk-1</i> mutant strains | 115 |
| 4.2.2.1 DNA extraction | 115 |
| 4.2.2.2 PCR and restriction digestion | 115 |
| 4.2.3 Outcrossing novel <i>dapk-1</i> mutant strains | 118 |
| 4.2.3.1 Generating male <i>C. elegans</i> | 118 |
| 4.2.3.2 Outcrossing | 118 |
| 4.2.4 Phenotyping <i>dapk-1</i> mutant strains | 119 |
| 4.2.4.1 Scoring head morphology defects | 119 |
| 4.2.4.2 Lifespan assay | 119 |
| 4.2.4.3 Progeny counts | 120 |
| 4.3 Results | 121 |
| 4.3.1 Validation of mutant <i>dapk-1 C. elegans</i> strains | 121 |
| 4.3.1.1 <i>dapk-1 gk219, ju4, knu450 T715N</i> and <i>knu458 T715N</i> genotype validation | 121 |
| 4.3.1.2 Optimising K57W genotype validation | 124 |
| 4.3.2 Outcrossing the <i>dapk-1 (knu458 T715N)</i> strain | 127 |
| 4.3.3 Phenotypic analysis | 129 |
| 4.3.3.1 Head morphology defect | 129 |
| 4.3.3.2 Lifespan assessment | 131 |
| 4.3.3.3 Viable progeny counts | 133 |
| 4.4 Discussion | 135 |
| 4.4.1 Further insight into mutant <i>dapk-1</i> associated cuticle abnormalities | 135 |
| 4.4.2 Exploring new <i>dapk-1</i> phenotype avenues | 136 |
| 4.4.3 Genotype validation challenges | 138 |
| 4.4.4 Conclusions | 138 |
| Chapter Five - General Discussion and Suggested Future Work | 140 |
| 5.1 General Discussion | 141 |
| 5.1.1 Research overview | 141 |
| 5.1.2 Outlook | 143 |
| 5.1.2.1 The human DAPK1 and <i>C. elegans</i> DAPK-1 interactomes | 143 |
| 5.1.2.2 PINOT and PPI data curation | 144 |
| 5.2 Future Work | 146 |
| References | 148 |
| Appendix - Supporting Files, Conferences and Publications | 165 |
| 7.1 Supporting files | 166 |

7.2 Conference Presentations

167

7.3 Publications

168

LIST OF FIGURES

| | | |
|------|---|-----|
| 1.1 | Domain organisation and ROC-COR sequence similarity of the human ROCO proteins | 18 |
| 1.2 | <i>Caenorhabditis elegans</i> life cycle | 25 |
| 2.1 | Distribution of molecular interaction data entries for the human ROCO proteins across multiple primary databases | 31 |
| 2.2 | Weighted Protein-Protein Interaction Network Analysis (WPPINA) pipeline | 35 |
| 2.3 | Purification of 3xFLAG-tagged proteins | 37 |
| 2.4 | Literature-derived human ROCO protein interaction network | 42 |
| 2.5 | Experimental human ROCO protein interaction network | 45 |
| 2.6 | Common core human ROCO protein interaction network | 48 |
| 2.7 | mRNA coexpression profiling of ROCO proteins in relation to nodes within the common core network | 50 |
| 2.8 | Human ROCO protein interactome functional enrichment map | 52 |
| 3.1 | Primary structure comparisons between human DAPK1 and <i>C. elegans</i> DAPK-1 | 67 |
| 3.2 | Data processing pipeline underlying PINOT for <i>C. elegans</i> data | 73 |
| 3.3 | Screenshot of an example PINOT <i>C. elegans</i> data output | 75 |
| 3.4 | Inserted 3xFLAG and linker sequence at the <i>dapk-1</i> locus | 77 |
| 3.5 | Experimental design and predicted result for FLAG- <i>dapk-1</i> genotype validation | 80 |
| 3.6 | Performance comparison of PINOT and MIST for <i>C. elegans</i> PPI data | 87 |
| 3.7 | DAPK1 protein interaction network and orthologous <i>C. elegans</i> interactors | 89 |
| 3.8 | Predicted DAPK-1 protein interaction network based on orthologous interactor predictions and reported interactions of these nodes | 91 |
| 3.9 | FLAG- <i>dapk-1</i> tag validation at the genomic level | 95 |
| 3.10 | FLAG-DAPK-1 tag validation at the protein level | 97 |
| 3.11 | FLAG- <i>dapk-1</i> transcript analysis by qPCR | 98 |
| 3.12 | <i>dapk-1</i> and <i>act-2</i> mRNA expression through the developmental stages of <i>C. elegans</i> | 106 |
| 4.1 | Structural inference of <i>ju4</i> S179L mutation site based on human DAPK1 kinase structure | 111 |
| 4.2 | Mutant <i>dapk-1</i> genotyping experimental design using PCR and restriction digestion | 117 |
| 4.3 | <i>dapk-1</i> mutant genotype validation for <i>gk219</i> , <i>ju4</i> and T715N strains | 123 |
| 4.4 | Optimising mutation validation of <i>dapk-1</i> K57W strain | 126 |
| 4.5 | Representative genotype verification upon outcrossing the <i>dapk-1</i> (<i>knu458</i> T715N) strain onto the N2 background by PCR and restriction digestion | 128 |

| | | |
|-----|--|-----|
| 4.6 | Head morphology defect in mutant <i>dapk-1 C. elegans</i> strains | 130 |
| 4.7 | Analysis of lifespan in mutant <i>dapk-1 C. elegans</i> strains | 132 |
| 4.8 | Viable progeny assessment in mutant <i>dapk-1 C. elegans</i> strains | 134 |

LIST OF TABLES

| | | |
|-----|---|-----|
| 1.1 | Substrates of DAPK1 kinase activity | 23 |
| 2.1 | Snapshot of method annotation reassignment | 33 |
| 2.2 | Overview of confidence score assignment | 34 |
| 2.3 | Snapshot of functional term reannotation | 39 |
| 2.4 | Functional overview of common interactors within literature-derived network | 43 |
| 2.5 | MASL1-interacting kinases | 46 |
| 2.6 | LRRK2 interactors identified via protein microarray that have been previously reported with the literature | 49 |
| 2.7 | Top three significantly enriched GO BP terms in various ROCO protein interactome functional enrichment analyses | 53 |
| 2.8 | Advantages and limitations of PPI detection methods | 61 |
| 2.9 | Comparison of features across PPI query resources | 63 |
| 3.1 | Primer sequences for FLAG- <i>dapk-1</i> genotype validation by PCR | 79 |
| 3.2 | PCR thermocycler programme for FLAG- <i>dapk-1</i> genotype validation | 80 |
| 3.3 | List of antibodies used for Western blot | 82 |
| 3.4 | Primer sequences for FLAG- <i>dapk-1</i> transcript validation by qPCR | 84 |
| 3.5 | Novel DAPK-1 interactors identified by yeast two-hybrid (Y2H) | 93 |
| 3.6 | Functional and phenotypic insight into novel DAPK-1 interactors identified by Y2H | 104 |
| 4.1 | Mutant <i>dapk-1</i> <i>C. elegans</i> strains with reported cuticle defect phenotype | 111 |
| 4.2 | Primer sequences for mutant <i>dapk-1</i> validation by PCR | 116 |
| 4.3 | PCR thermocycler programme for <i>gk219</i> , <i>ju4</i> , T715N and K57W genotyping | 116 |
| 4.4 | Restriction endonucleases used for genotype validation | 116 |

ABBREVIATIONS

| | |
|-----------------|---|
| aa | amino acid |
| Amp | ampicillin |
| AP-MS | affinity purification mass spectrometry |
| BP | biological process |
| bp | base pairs |
| COR | C-terminal of ROC |
| DAPK1 | death-associated protein kinase 1 |
| GDP | guanosine diphosphate |
| GO | gene ontology |
| GTP | guanosine triphosphate |
| HIPPIE | Human Integrated Protein–Protein Interaction Reference |
| ID | identifier |
| kb | kilobases |
| kDa | kilodalton |
| LRRK1 | leucine-rich repeat kinase 1 |
| LRRK2 | leucine-rich repeat kinase 2 |
| MASL1 or MFHAS1 | malignant fibrous histiocytoma amplified sequence 1 |
| MIST | Molecular Interaction Search Tool |
| NGM | nematode growth media |
| NTC | no template control |
| PCR | polymerase chain reaction |
| PI3KC3 | class III phosphatidylinositol 3-kinase |
| PINOT | Protein Interaction Network Online Tool |
| PPI | protein-protein interaction |
| ROC | Ras of complex proteins |
| STRING | Search Tool for Recurring Instances of Neighbouring Genes |
| ULK1 | unc-51-like kinase 1 |
| UTR | untranslated region |
| WLB | worm lysis buffer |
| WPPINA | weighted protein-protein interaction network analysis |
| Y2H | yeast two-hybrid |

CHAPTER ONE

Introduction

1.1 Investigating protein-protein interactions: the rationale, resources and rewards

The subcellular environment is a dynamic interconnected landscape of intricate signalling events that regulate cell homeostasis. Much of this signal transduction is driven by physical protein-protein interactions (PPIs) which convey molecular messages through cells at a systems-level to coordinate biological processes. Impairment or alterations to these pathways can lead to aberrant signalling and disease states. Therefore, investigating PPI networks is a valued strategy for developing a holistic view of the mechanisms which underlie cell function and dysfunction.

Motivation to identify PPIs within the proteome is longstanding and increasingly, high-throughput approaches, such as yeast two-hybrid (Y2H) and protein microarray, are used to detect thousands of PPIs in a single screen [1]. Furthermore, the BioPlex project, led by the Gygi and Harper labs at Harvard Medical School, has utilised a traditionally low-throughput technique, affinity purification coupled with mass spectrometry (AP-MS), in a high-throughput manner to identify PPIs at an almost proteome-wide scale [2,3]. Of note, each PPI detection method has inherent limitations and certain types of PPIs, such as transient low-affinity PPIs, are challenging to detect, which has led to the development of a diverse toolbox of techniques available for detecting PPIs.

To enhance the accessibility of this wealth of PPI data generated, extensive efforts in primary database curation are ongoing. Data repositories, such as BioGRID [4] and IntAct [5], have been manually cataloguing PPI data entries from published studies for over a decade and increasingly encourage data deposits from researchers performing high-throughput experiments. Typically, PPI data entries contain gene/protein identifiers (IDs) for both pairwise interactors, species IDs, method detection information and related publication records. This curation initiative is supported by the International Molecular Exchange (IMEx) consortium [6] which includes numerous PPI data providers and promotes standardisation of high quality data curation across the major molecular interaction data resources.

These databases facilitate the construction of PPI networks, an analysis approach based on graph theory, which provides a representation of the connectivity between proteins. In a graph (or alternatively termed a network), nodes are connected by edges which depict proteins and interactions, respectively. In addition, nodes which represent proteins of interest, for example proteins queried against a PPI database, are termed seeds. Where nodes and edges form chains of connection within a network, potential signal transduction cascades can be inferred.

This network analysis approach is often complemented with functional annotation analysis to gain biological insight into proteins and PPI networks. Numerous ongoing functional annotation projects, for example Gene Ontology (GO) [7], Reactome [8] and WikiPathways [9], assign and maintain brief functional descriptions of genes and pathways, which form the basis of analyses such as functional enrichment. Furthermore, utilising the guilt-by-association principle [10], which hypothesises that proteins which interact are likely involved in the same or similar functional pathways, functional inferences can be deduced, a strategy particular useful for poorly characterised proteins.

Utilising a PPI network analysis approach is particularly useful for exploring the interactions and, in turn, functional profiles of related proteins. For example, structurally related proteins, such as those with common domain topologies [11,12] or disease related proteins, for instance those which have disease association through nominated genetic loci [13–16]. In these example cases, a protein network approach facilitates dissection of structure-function relationships between proteins and enables functional insight to be gathered into potential pathological processes influencing disease states, respectively. Since these network analysis approaches often utilise many data points from multiple seed proteins collectively, it enables the identification of common interactors between seed proteins, such observations would be missed if analysing individual interactomes. These areas of overlap between interaction profiles suggest functional convergence and hence the concept of shared pathways between seed proteins can be explored.

To date, the study of PPIs to better understand cellular processes and shed light on the complexity of biological systems has been fruitful. For example, the detailed characterisation of molecular assembly, mTOR illustrated how the recruitment of different proteins within the complex leads to different functional outcomes in relation to downstream signalling events [17]. Upon scaling PPI analysis up to a network level, the utility of this approach broadens, including attempts at mapping the entire human interactome [1,3], exploring PPI networks related to human disease [13,14,16,18] and utilising this approach for tracing protein evolution [19]. The future direction of this field is likely to involve data integration across a multitude of omic data types to construct *in silico* models of living systems.

1.2 ROCO proteins

The ROCO proteins are a diverse family of large multidomain proteins characterised by a Ras of complex proteins (ROC) – C-terminal of ROC (COR) supra-domain [20]. These two domains always exist in tandem and in the same orientation (ROC-COR), suggesting they act as a single functional unit [21]. The ROC region of this conserved unit resembles a GTPase, similar but phylogenetically distinct from small single-domain GTPases such as Ras, Rab, and Rho proteins [22]. This protein family was first identified in the amoeba slime mould *Dictyostelium discoideum* in 2003 and has since been mapped in a range of species from prokaryotes to mammals [22]. Interestingly from an evolutionary perspective, the distribution and domain topologies of ROCO proteins throughout different species is highly variable [21,23].

1.2.1 Human ROCO proteins: structure, regulation and function

The human ROCO protein family comprises of death-associated protein kinase 1 (DAPK1), leucine-rich repeat kinase 1 (LRRK1), leucine-rich repeat kinase 2 (LRRK2) and malignant fibrous histiocytoma amplified sequence 1 (MASL1 or MFHAS1) [20]. The ROC-COR tandem domain in these proteins is flanked by a diverse range of protein-protein interaction (PPI) interfaces and in DAPK1, LRRK1 and LRRK2, an active kinase domain (Figure 1.1A). This arrangement of intrinsic dual catalytic activities (kinase and GTPase) is exclusive to these three proteins in the human proteome. Of note, the kinase domain in DAPK1 is structurally distinct from the LRRK1 and LRRK2 kinase domains. The DAPK1 kinase domain belong to the calcium/calmodulin-dependent serine/threonine kinase family with a unique basic loop (residues 45-56) which is a defining feature of death-associated protein kinases (DAPKs) [24,25], whereas the LRRK1 and LRRK2 kinase domains resemble tyrosine kinase-like (TKL) origin and specifically mixed-lineage kinase (MLK) homology, yet possess serine/threonine kinase activity [26,27]. The conserved ROC-COR region is most similar between the LRRK paralogs, although considerably similarity is apparent across all human ROCO proteins (Figure 1.1B and 1.1C).

The ROC domain of these proteins contains five conserved G-domain sequence motifs which are characteristic of GTPases [28]. The ability of these proteins to bind and hydrolyse GTP has been tested, demonstrating that all four human ROCO proteins are active GTPases [29–33]. Furthermore, intramolecular modulation of the kinase activity by GTP binding (in the ROC domain) has been reported for DAPK1, LRRK1 and LRRK2. In GTP binding deficient forms of DAPK1, the kinase activity is enhanced, suggesting GTPase activity of the ROC domain negatively regulates DAPK1 kinase activity [29,30], whereas in LRRK1 and LRRK2, impairment of GTP

binding properties of the ROC domain results in kinase inactivity [31,34]. Therefore, it appears that the GTPase activity within the ROC domain of these proteins is acting as a molecular switch for kinase activity, and for MASL1 this may be an unknown extrinsic kinase domain.

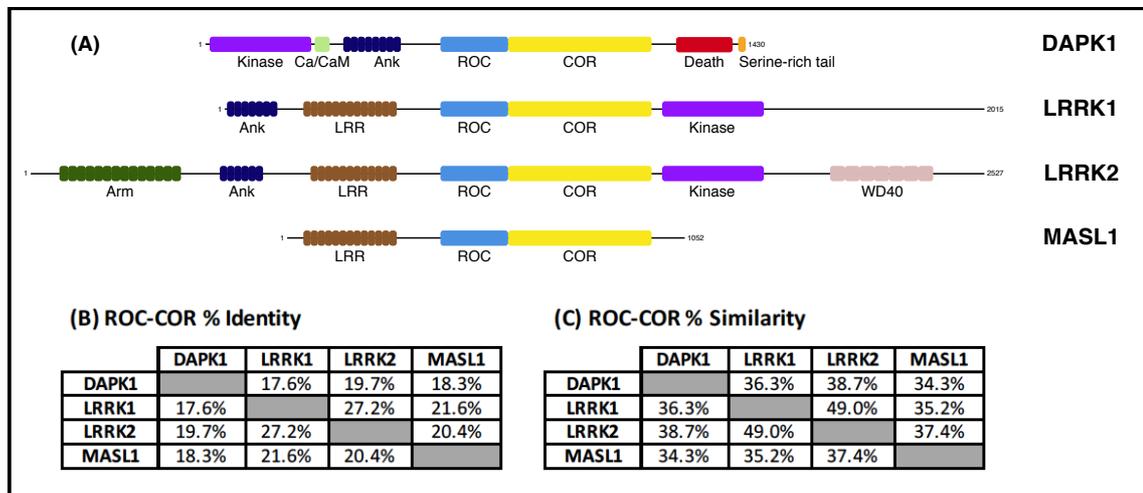


Figure 1.1 – Domain organisation and ROC-COR sequence similarity of the human ROCO proteins. The human ROCO proteins are characterised by a conserved ROC-COR tandem domain. (A) Domain topologies of DAPK1, LRRK1, LRRK2 and MASL1. Amino acid sequence identity (B) and similarity (C) of the conserved ROC-COR region. ROC-COR region defined as residues 612–1225 (DAPK1), 574–1143 (LRRK1), 1271–1790 (LRRK2) and 345–972 (MASL1). Abbreviations: Ank, ankyrin repeats; Arm, armadillo repeats; Ca/CaM, calcium/calmodulin regulatory domain; COR, C-terminal of ROC; LRR, leucine-rich repeats; ROC, Ras of complex proteins

One area of uncertainty regarding the catalytic activity within the ROC domain of these proteins is the GDP/GTP cycle underlying GTPase activity. The conventional nucleotide-state cycle of small GTPases, whereby guanine nucleotide exchange factors (GEFs) facilitate the dissociation of GDP resulting in the GTPase adopting its GTP-bound active form and GTPase activating proteins (GAPs) induce GTP hydrolysis, resulting in the GTPase returning to its inactive GDP-bound form, may not be the case for ROCO proteins. A recently proposed mechanism, assessed in the context of LRRK2 and bacterial ROCO proteins, is that this protein family do not require GEFs to induce GTPase activation due to the low affinity and high dissociation rate of GDP/GTP in relation to these proteins. Instead, dimerisation of ROCO proteins and consequential conformational changes in protein structure results in GDP dissociation [28]. This is an intriguing hypothesis, which is supported by evidence of ROCO protein dimerisation [35].

At a functional level, the human ROCO proteins have been associated with a multitude of diverse biological processes, positioning them as important signalling modulators. First, DAPK1 has been extensively linked to cell death pathways (discussed in detail in '1.3.3.1 Cell death and

autophagic pathways'). LRRK1 and LRRK2 although structurally very similar (Figure 1.1A), appear to be involved in different functions [36]. LRRK1 has been implicated in epidermal growth factor receptor (EGFR) trafficking [37,38], spindle orientation during mitosis [39], humoral immunity [40] and osteoclast function [41]. Furthermore, mutations in *LRRK1* have been reported in patients with osteosclerotic metaphyseal dysplasia [42,43]. LRRK2 has functional links to many cellular process, including autophagy, inflammation and apoptosis [44]. Moreover, LRRK2 is the focus of extensive research efforts in relation to Parkinson's disease (PD), since mutations in *LRRK2* are a known to cause familial PD and variants in this gene also contribute to increased risk of sporadic PD [45]. Finally, *MASL1* is the least studied human ROCO protein and hence little is known about its cellular role [46]. Current functional associations relate to macrophage polarisation [47] and erythropoiesis [48], and from a clinical perspective, *MASL1* is a reported oncogene in relation to various cancer types [46].

1.3 Death-associated protein kinase 1 (DAPK1)

DAPK1 is a 160kDa calcium/calmodulin regulated kinase with extensive functional connections to cell death pathways and tumour suppression [49] in addition to a plethora of other proposed functions. As described in '1.2.1 Human ROCO proteins', DAPK1 belongs to the ROCO protein family due to structural homology in the ROC-COR supra-domain, a GTPase functional unit within the centre of the protein (Figure 1.1A) and is also a member of the death-associated protein kinase (DAPK) family, based on kinase homology. This introductory section will provide an overview of DAPK1 from its identification to targeting this complex protein for therapeutic intervention.

1.3.1 Identification, transcription and splicing of *DAPK1*

DAPK1 was discovered through an unbiased screen for genes upregulated upon interferon- γ stimulation, using an antisense cDNA library and growth of HeLa cells as a positive readout, approximately 25 years ago [50]. This cytokine insult induces apoptosis and hence from the discovery of *DAPK1*, it was hypothesised that this gene positively regulates programmed cell death. The *DAPK1* gene was subsequently mapped to chromosome 9 and spans 211.4kb [51]. Its transcription is tightly regulated by a number of epigenetic modifications and transcription factors. The promoter region of *DAPK1* is densely distributed with CpG islands which are prone to methylation [52], a modification which represses gene transcription. In addition, several transcriptional regulatory elements have been mapped upstream of and within the *DAPK1* gene, these include SP1, AP1, AP2, E2F and NF κ B binding sites, a CAAT box and an E box [53].

Functional characterisation of *DAPK1* transcription has identified two transcription factors, CP2 and HNF3B/FOXA2, which initiate transcription of *DAPK1* from two distinct promoter regions. Interestingly, *DAPK1* transcription from these two promoter regions result in distinct transcripts, yet a single protein [53]. Additional transcriptional regulation of *DAPK1* has been reported, whereby p53, C/EBP- β and SMAD induce transcription, whereas STAT3 and p52NF- κ B suppress transcription [52]. Post-transcriptionally, miRNAs (specifically miRNA-103 and miRNA-107) target the 3' UTR of *DAPK1* mRNA and suppress protein expression [54]. Furthermore, multiple splice variant forms of *DAPK1* have been discovered: DAPK1- α which is considered the full-length transcript results in translation of a 1430 amino acid (aa) protein; DAPK1- β which results in the translation of a slightly larger 1440 aa protein which is particularly interesting from a functional perspective, since the murine ortholog appears to have cytoprotective properties [55,56]; and a much smaller splice variant, s-DAPK1, which encodes a 337 aa protein which lacks the kinase and death domains, and has a novel C-terminus [57].

1.3.2 Structure and regulation of DAPK1

The primary structure of DAPK1 is well defined into a number of domains. From N-terminus to C-terminus, the protein consists of a kinase domain, a calcium/calmodulin regulatory domain, a series of ankyrin repeats, ROC-COR domains, a death domain and a serine rich tail (Figure 1.1A). This primary structural organisation will influence the three-dimensional (3D) conformations that DAPK1 is likely to adopt and ultimately the structure-function relationship of the protein. Currently, crystal structures for the kinase and calcium/calmodulin regulatory domains have been determined [24][58], however a full-length DAPK1 3D structure is yet to be solved.

The kinase domain of DAPK1 harbours serine/threonine kinase activity which is dependent on residue K42 within the ATP binding site [29,30]. The 1.5 Å structural resolution of this domain [24] has shed light on key motifs within this region. These include a basic loop which is a characteristic motif of the DAPK family, these 12 residues are thought to be involved in homodimerisation of DAPK1 and potential heterodimerisation with DAPK3 (or ZIPK) which is a reported substrate of DAPK1 [59,60]. Further elucidation from structural analysis revealed that the kinase domain maintains a closed active conformation and therefore kinase activation *via* phosphorylation of an activation loop, which is common to many other kinases, appears not to be necessary [24,25]. Moreover, the DAPK1 kinase domain interacts with HSP90, a chaperone protein critical for the stability of DAPK1 [61,62], hence this interaction is likely to be long-lived and DAPK1 could be considered to exist as a protein complex.

The calcium/calmodulin regulatory domain of DAPK1 is central to its kinase activity; activation of this region by calcium-activated calmodulin results in activation of the kinase activity [63]. Furthermore, when inactive this domain acts as a pseudo-substrate, physically negatively regulating kinase activity by blocking entry of exogenous substrates to the kinase catalytic cleft [25]. Genetic deletion of this domain results in a constitutively active DAPK1 kinase [63]. This region is further negatively regulated by an autophosphorylation event on residue S308 (in the calcium/calmodulin regulatory domain), which reduces the affinity of calmodulin binding [64]. This residue is dephosphorylated by PP2A phosphatase upon stimuli such as ceramide, TNF α , ER stress and ischaemia, which consequently activates the DAPK1 kinase domain [65–67].

The latest domains to be mapped within DAPK1 are the ROC-COR tandem domains. As described in the context of ROCO proteins, the ROC domain of this functional unit is able to bind and hydrolyse GTP. The role of this catalytic activity in DAPK1 is largely unexplored, however it is evident that in GTP binding deficient forms of DAPK1, kinase activity is enhanced [29,30] suggesting an important intramolecular regulatory role for this region. Furthermore, it has been proposed that the ROC domain is a homodimerisation interface in addition to the kinase domain [29]. A region that overlaps the ROC domain and flanks the region towards the N-terminus is thought to be involved in the localisation of DAPK1 to the cytoskeleton, since deletion mutant forms of DAPK1 lacking residues 641-835 resulted in mislocalisation of DAPK1 to the cytosol [63,68].

The death domain at the C-terminus of DAPK1 is a protein-protein interacting region which is involved in a diverse range of functions spanning both interdomain regulation and activation of downstream molecular cascades. This death domain is required for the docking of extracellular signal-regulating kinase (ERK), which subsequently phosphorylates Ser735 in the ROC domain and promotes kinase activity at the N-terminus of the protein [69]. Regulatory interactions such as this, illustrate the complexity and significance of interdomain cross-talk for the functioning of multidomain proteins. In addition, the death domain has a role in the stability of the protein, since this region binds adaptor proteins which recruit a ubiquitin E3 ligase complex to initiate polyubiquitination and ultimately protein degradation *via* the proteasome [70].

Following the death domain is a 17 amino acid sequence known as the serine rich tail, which denotes the C-terminus of the protein. It has been established that the serine rich tail has no direct influence on kinase activity but it is thought to negatively regulate apoptosis by modulating PPIs elsewhere in the protein [71].

1.3.3 DAPK1 function

DAPK1 has been implicated in a diverse range of biological processes. Most notably, this cell signalling molecule has functional links to cell death and autophagic pathways [49], however further evidence has also suggested roles for DAPK1 in inflammation [72], cell migration [73], synaptic function [74], and beyond [75]. At a molecular level, a number of substrates of DAPK1 kinase activity have been identified (Table 1.1). Current understanding of how DAPK1 is involved in cell death and autophagic pathways is outlined below in further detail.

1.3.3.1 Cell death and autophagic pathways

Cell death signalling is traditionally classified into three pathways based on mechanistic and cellular morphological features prior to cell death (i.e. apoptosis, autophagy, necrosis) [76]. Although these classifications are widely used, a recent proposal for reclassification of cell death signalling highlights the complexity and diversity of these pathways [77]. This is reflected in the involvement of DAPK1 in cell death, which is functionally linked to all three of the traditionally classified pathways. Since its identification, it has been reported that the activation of DAPK1 harbours pro-apoptotic properties upon stimulation from a variety of signals: interferon- γ , TNF α , TGF β , ceramide, ischaemia, Fas signalling, p53 signalling [50,78–82]. At a cellular level, pro-apoptotic DAPK1 signalling results in membrane blebbing and the formation of autophagic vesicles [83].

DAPK1 has also been associated with promoting autophagic cell death *via* ER stress signals and through DAPK3 [60,65]. Moreover, the molecular basis of the role of DAPK1 in autophagy has been mapped from two angles. First, DAPK1 is known to phosphorylate beclin 1, which results in the dissociation of beclin 1 from Bcl-2 [84]. This signalling event enables beclin 1 to become active with respect to assembly of class III phosphatidylinositol 3-kinase (PI3KC3) complex I for autophagy induction. Through an independent mechanism, DAPK1 activates PI3KC3 complex I *via* phosphorylating PKD [85] which in turn, phosphorylates Vps-34 (a component of PI3KC3 complex I) to activate the complex for autophagy induction. It has also been shown that the activation of PKD by DAPK1 results in JNK signalling and cellular necrosis [85]. A further mechanism whereby DAPK1 induces cell death is by anoikis (programmed cell death due to detachment of cells from the cellular matrix), whereby DAPK1 is thought to interfere with integrin function [86].

Table 1.1 – Substrates of DAPK1 kinase activity. Reported DAPK1 substrates, mapped phosphorylation sites and proposed functional influence.

| Substrate | Phospho sites | Related function | References |
|---------------------|------------------------------------|---|------------|
| BECN1 (Beclin1) | T119 | Autophagy | [84] |
| DAPK3 (or ZIPK) | T299, S309, S311, S312, S318, S326 | Autophagic cell death | [60] |
| MCM3 | S160 | DNA replication, cell cycle | [87] |
| MLC | S19 | Apoptotic membrane blebbing | [68] |
| NR2B | S1303 | Ischaemic neuronal death | [88] |
| p53 | S23 | Apoptosis and necrosis | [89] |
| PKD | Undefined | Necrosis | [85] |
| STX1A (Syntaxin 1A) | S188 | Synaptic vesicle activity | [74] |
| Tropomyosin | S283 | Stress fibre formation, cytoskeletal dynamics | [90] |

1.3.4 DAPK1 in disease and as a therapeutic target

The dysregulation of DAPK1 has been linked to a number of human diseases. A widespread epigenetic feature of many tumour and cancer types is hypermethylation of the DAPK1 promoter region and hence suppression of *DAPK1* transcription [91]. Due to the pro-apoptotic properties of DAPK1, targeting this region with demethylating agents has been a therapeutic avenue explored in oncology. Evidence from preclinical tests supports this as a potential therapeutic strategy [92,93].

Furthermore, DAPK1 has been associated with several neurological disorders. This protein has been proposed as a therapeutic target for ischaemic stroke and brain injury due to evidence that suggests DAPK1 contributes to neuronal death in the brain following ischaemia or trauma [88,94]. DAPK1 has also been linked to Alzheimer’s disease (AD) *via* data showing increased *DAPK1* expression in brains of AD patients and suggested roles of DAPK1 in tau and amyloid precursor protein regulation [95,96], although this require further exploration. *DAPK1* expression is also elevated in the brain of epilepsy patients [97].

1.3.5 Death-associated protein kinase (DAPK) family

In addition to the ROCO protein family, DAPK1 also belongs to the DAPK family based on kinase homology and cell death related functions. The other two DAPKs, DAPK2 (or DRP1) and DAPK3 (or ZIPK) are much smaller than DAPK1 at just 42kDa and 55kDa, respectively, with just the kinase domain in common throughout all three. Unlike the ROCO proteins, whereby the DAPK1 kinase domain is more distantly related to the other kinase domains of the protein family, DAPKs share >80% sequence homology within the kinase domains [25].

1.4 *Caenorhabditis elegans*

Caenorhabditis elegans is a free-living nematode that naturally inhabits soil and feeds on bacteria. However, for over fifty years *C. elegans* has been used as a model organism in laboratories for a multitude of biological research from understanding the foundations of genetics to preclinical drug screening for potential therapeutics. The widely used wild-type strain, termed N2 (Bristol), was isolated from compost near Bristol, UK.

C. elegans has a short life cycle of approximately 3 days from fertilisation to a reproductively active adult nematode, although this timescale is variable with temperature [98]. Through this life cycle (Figure 1.2), embryos undergo a short *in utero* development period, followed by *ex utero* embryonic development and hatching of an L1 larvae. In favourable condition, which includes a bacterial food source, L1 larvae undergo a series of larval moults (L1-L4) prior to development into a young adult and then finally a reproductively active adult. The *C. elegans* lifespan is also relatively short, approximately 17 days at 20°C [99], hence this model organism is frequently used in ageing research.

In addition, *C. elegans* exists in two sexes, self-fertilising hermaphrodites and males. However, males are present in a very low frequency (<0.2%) in laboratory populations [100] and hence the majority of progeny arise from self-fertilisation. For reproduction, hermaphrodite nematodes produce and store sperm prior to the production of oocytes and therefore can produce a finite number of fertilised embryos (typically 250-300), whereas upon cross-fertilisation whereby sperm is provided by male nematodes, hermaphrodites can produce a brood size of approximately 1000 progeny. Adult hermaphrodite nematodes produce progeny for approximately 5 days. A short generation time, together with the potential to produce hundreds of progeny *via* either self- or cross-fertilisation positions *C. elegans* an attractive model organisms for genetic studies.

Furthermore, the cell fate of each cell in *C. elegans* has been mapped [101]. Adult hermaphrodite nematodes possess 959 somatic cells which form simple organ systems, with a nervous system composed of 302 neurons. Moreover, the body of *C. elegans* is transparent and therefore fluorescent tracking of proteins and cells is achievable *in vivo*. In addition, there is a considerable degree of conservation between the *C. elegans* and human genomes, with a report that over 80% of *C. elegans* genes have a human ortholog [102]. This highlights the genetic tractability of this species in the context of the human genome/proteome. This is supported by a comprehensive resource for *C. elegans* research, WormBase [103].

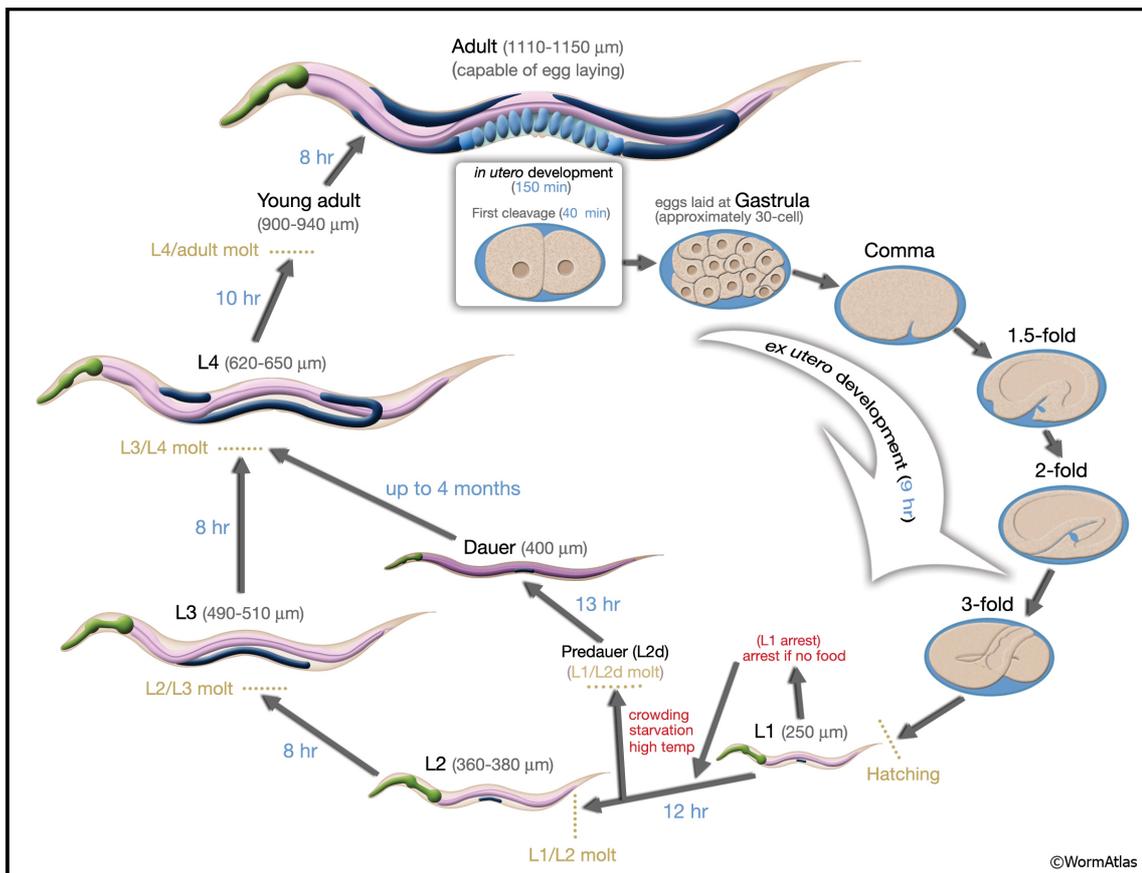


Figure 1.2 – *Caenorhabditis elegans* life cycle. A stepwise visualisation of the *C. elegans* life cycle when nematodes are maintained at 22°C. Timespan for each stage is annotated in blue text, 0 mins corresponds to time of fertilisation. Length of nematodes at each stage is annotated in grey and in brackets. Image from wormatlas.org (available at <https://www.wormatlas.org/hermaphrodite/introduction/IMAGES/introfig6.jpg> [accessed November 2019])

1.5 Overall aims

The primary aim of the research presented in this thesis is to progress our current understanding of the physiological role of DAPK1 by defining its proximal interactome. This aim will be achieved through a range of bioinformatic and functional modelling approaches. First my goal is to determine the current knowledgebase on the human DAPK1 interaction profile, and more broadly the ROCO protein family as a whole, by mapping PPI networks based on existing data from published studies and use this interactome analysis as the basis for gathering functional insight. To achieve this goal, I aim to develop a data mining and processing pipeline to maximise the utility of freely available PPI data. The ultimate aim with this pipeline is to release it as a resource for the research community. I then aim to integrate this literature-derived ROCO protein interaction network with novel ROCO PPI findings as a strategy to prioritise interactors identified *via* a high-throughput screen for further investigation.

Next, I intend to translate the human DAPK1 interactome data to *C. elegans* in order to predict DAPK-1 interactors in this model organism. I then aim to gain evidence based insight into the *C. elegans* DAPK-1 interactome in order to assess the validity of predictions and understand where DAPK-1 signalling may fit within the wider cellular context. To support this goal, I plan to develop a genome edited *C. elegans* FLAG tagged DAPK-1 strain. Furthermore, I aim to enhance the *C. elegans* toolbox available for dissecting the role of DAPK-1 *in vivo* by engineering novel mutant strains whereby key residues in relation to the DAPK-1 catalytic activities are targeted. Finally, I aim to phenotypically characterise these new strains to obtain an idea of the role of DAPK-1 in *C. elegans* physiology and to suggest routes for future examination.

CHAPTER TWO

Evaluating the Human ROCO Protein Interaction Network

2.1 Introduction

The resources for mapping protein interactomes are becoming well established, both in terms of curated data repositories and available technical approaches [104]. These tools facilitate the construction of PPI networks for interpreting PPI data collectively and in the wider cellular context. As described in Chapter One – ‘Investigating protein-protein interactions: the rationale, resources and rewards’, these network analysis approaches are often coupled with functional annotation analyses to add a functional perspective to PPI networks. Since interacting proteins are likely involved in the same or similar functional pathways [10], the utility of functional annotations is particularly beneficial for inferring biological meaning to functionally uncharacterised proteins within PPI networks. Additionally, these are valuable approaches for assessing the interaction and functional profiles of related proteins.

The human ROCO proteins are an attractive protein family for PPI network analysis due to the diverse, yet partially similar, multidomain topologies and dual enzymatic activities of these proteins. These structurally related proteins consist of a conserved tandem supra-domain, termed ROC-COR, flanked by a plethora of PPI regions (Figure 1.1A). A number of these PPI interfaces and repeat motifs are conserved between two or more ROCO proteins, such as the leucine-rich repeats. In addition, several functional domains are unique to specific ROCO proteins. This poses an interesting route of investigation to dissect commonalities and distinctions in the ROCO protein interaction profiles based on protein structure.

Furthermore, three of the four human ROCO proteins (DAPK1, LRRK1 and LRRK2) possess both kinase and GTPase catalytic activities, a phenomenon exclusive to these three proteins in the human proteome. Since kinases and GTPases are key molecules in signal transduction cascades [105,106] and epigenetic/genetic alterations to the ROCO proteins, in particular DAPK1 and LRRK2, lead to disease phenotypes, exploring the ROCO protein interactome will likely shed light on their potentially critical contribution to cell signalling. Consequently, defining these PPI events will facilitate the targeting of ROCO proteins and associated functional pathways in existing and novel therapeutic development programmes.

Prior studies to gather insight into the interaction profiles of the human ROCO proteins have taken an individualistic approach. The DAPK1 interactome has been assessed by using a traditional literature review strategy, which shed light on a multitude of pathways potentially influenced by DAPK1 [75]. Furthermore, the LRRK2 interactome has been mapped in two cases using PPI data repositories, which provided insight into the complex functional landscape of

LRRK2 biology, but also showcased the utility of PPI databases for PPI network analyses [107,108]. However, knowledge gaps persist in our understanding of human ROCO protein biology, including why such structurally similar proteins appear to be differentially involved in cellular function and dysfunction. In addition, where these complex multidomain enzymes fit within the landscape of subcellular signalling is underexplored. With this in mind and since the interaction profiles of this protein family as a whole had not been performed, a collective PPI network analysis approach was adopted in this project.

2.1.1 Project aims

Within this section of my investigation the aims of the research were two-fold, first from the perspective of exploring the ROCO protein interactome utilising both literature-derived and novel experimental data, then second, using the initial part of this analysis as a case study for trialling and further developing an in-house bioinformatic pipeline for maximising the utility of high quality PPI data available in the public domain, for PPI network analysis.

More specifically, my aims were to systematically assess the extent of the PPI literature for the human ROCO proteins in order to construct a confidence-weighted ROCO PPI network, using this network as the foundation for functional enrichment analysis (FEA) to gain functional insight into this protein family. These two approaches in combination were then utilised to identify commonality and distinction within the interaction and functional profiles of the ROCO proteins. Furthermore, I aimed to integrate novel PPI data derived from protein microarray experiments (performed by collaborators) with existing PPI data to prioritise interactors from the microarray screen for further validation.

Finally, in relation to my methodological development aim, my aim was to use the ROCO PPI network analysis based on literature-derived data as a prototype analysis to facilitate ongoing development of a structured bioinformatic pipeline for extracting, processing and confidence scoring PPI data. The rationale for this stems from our vision of releasing this pipeline as a freely-available user-friendly online resource and for this to materialise, extensive trialling of the pipeline with a broad range of protein query sets was crucial to identify and address potential malfunctions within the coding scripts of the pipeline. Hence the ROCO PPI network analysis contributed to this trialling procedure.

2.2 Materials and Methods

2.2.1 Constructing the literature-derived network

The literature-derived ROCO protein interaction network was generated by utilising a structured data mining approach coupled to an in-house data processing pipeline. This enabled extraction of PPI data which was readily available within the public domain to be processed, using a transparent filtering and quality control procedure, and subsequently assigned a confidence score based on the method detection and publication records. The pipeline (Figure 2.2), termed weighted protein-protein interaction network analysis (WPPINA) [13], consisted of a series of coding scripts developed in R (version 3.2.2). Each step of the pipeline is described below.

2.2.1.1 Data acquisition

PPI data was sourced by querying the Proteomics Standard Initiative Common Query Interface [109] (PSICQUIC [available at <http://www.ebi.ac.uk/Tools/webservices/psicquic/view/main.xhtml>]) for each ROCO protein (DAPK1, LRRK1, LRRK2 and MASL1), independently. To ensure a wide capture of reported PPIs, data was downloaded, *via* the PSICQUIC, from six primary databases: BioGRID [4], InnateDB, InnateDB-All, InnateDB-IMEx [110], IntAct [5] and MINT [111] (data downloaded on 12th Jan 2017 in a PSI-MITAB 2.5 format). This resulted in the generation of 17 text-files, each file was specific to a query protein and a source database (e.g. DAPK1_IntAct or LRRK2_BioGRID), and contained multiple entries of molecular interaction data (Figure 2.1).

The data entries consist of identifiers (IDs) for each interactor, the interactor species ID, a detection method annotation (this is derived from the EBI interaction detection method annotation ontology, annotation prefix “MI:”) and the PubMed ID relating to the publication that reports the interaction. In most cases, further information is stored in each data entry, such as database ID, however, this is not utilised for downstream data processing.

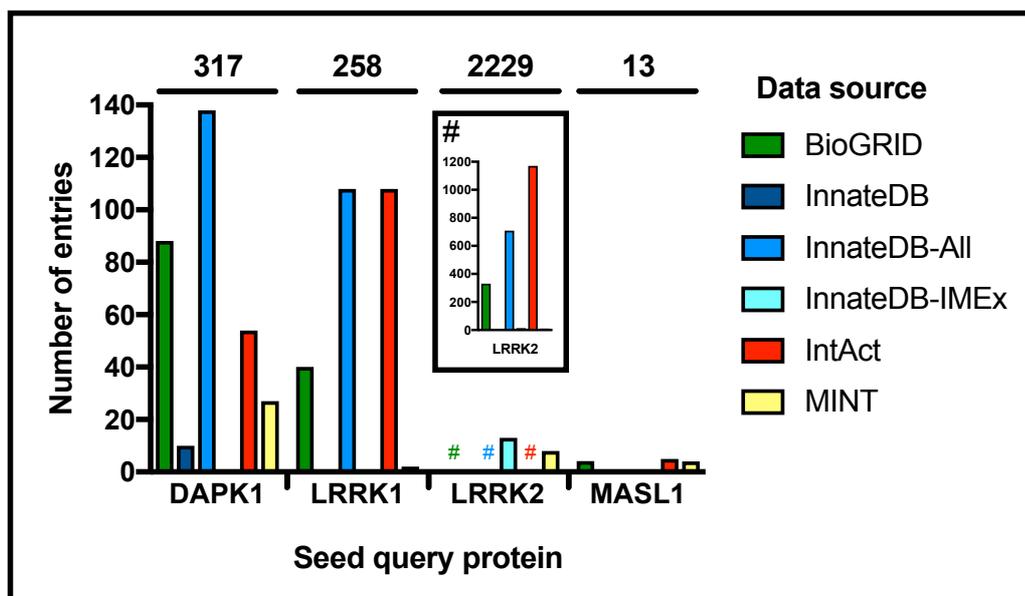


Figure 2.1 – Distribution of molecular interaction data entries for the human ROCO proteins across multiple primary databases. Number of PPI data entries available (on 12th January 2017) in BioGRID, InnateDB, InnateDB-All, InnateDB-IMEx, IntAct and MINT for each of the human ROCO proteins that were extracted for data processing within the WPPINA pipeline.

2.2.1.2 Data formatting and quality control

Upon downloading the data, the first challenge in data processing was faced. The downloaded data files originate from six different database sources, which although essentially curate the same type of data, the file formats differ slightly. These format inconsistencies hamper the ability to straightforwardly merge the files for further data processing. In particular, the inconsistencies of importance were where the gene/protein ID is located within the file and which gene/protein ID type is used to report a PPI. Achieving consistency in these two elements of the datasets early in the pipeline was essential to ameliorate potential data processing challenges downstream.

Three coding scripts were developed to overcome these format inconsistency issues. First, for data downloaded from Innate, a script to extract the Uniprot ID and store it in a more prominent position which aligns with other dataset formats (e.g. data downloaded from IntAct), since entries in Innate databases list the Uniprot ID nested within a string of alternative ID types. A further two scripts were developed to achieve a consistent ID type for each entry; BioGRID reports Entrez gene IDs, whereas the other data sources (Innate, IntAct, MINT) reports Uniprot protein IDs. These ID conversion scripts transposed the reported ID (Entrez or Uniprot) into the

alternative ID type and respective HGNC approved symbol (gene name) ID, based on a dictionary file of ID synonyms (developed from Uniprot-derived data for all human proteins annotated with a Swiss-Prot ID on 13th Jan 2017). The output files from these two latter scripts listed the PPI data entries with three ID types (Entrez, Swiss-Prot and gene name), which were consistently formatted.

Of note, Uniprot protein IDs are subcategorised into manually reviewed Swiss-Prot IDs or unreviewed automatically assigned TrEMBL IDs. Since the dictionary file used in the ID conversion scripts is curated using human Swiss-Prot IDs only, PPIs defined by TrEMBL or obsolete IDs, non-protein data entries (e.g. miRNAs, small molecules) and non-human PPIs were discarded from further data processing. In addition, incomplete data entries (e.g. a data entry lacking a PubMed ID) were removed. Once the datasets were consistently formatted and filtered, the data were merged into one file per seed protein. These files were a collation of human PPI data from each of the six source data repositories.

2.2.1.3 Interaction detection method reassignment

Each of the merged data files was then subjected to method annotation reassignment based on categorisation of the EBI molecular interaction detection method annotation ontology. A coding script was developed which grouped specific EBI annotations into a custom ontology, curated in-house, based on technical similarity of the methodology (an insight into this grouping is provided in Table 2.1 and the complete list is provided in Supporting File S1). This method reassignment step enables accurate confidence scoring, which was the subsequent step in the pipeline, whereby data entries were scored based on distinct methodologies used to detect an interaction. Hence, technically similar methodological approaches which have unique EBI annotations needed to be considered collectively.

Through this process each data entry was assigned a new method detection annotation (Table 2.1). Data entries that were reassigned the annotation ID, UNSPM (unspecified), were discarded from further data processing. In cases such as this, the initial EBI annotation lacked detail and hence the specific data entry curated into the source database was considered poor quality.

Table 2.1 – Snapshot of method annotation reassignment. EBI molecular interaction detection method annotations are grouped and reassigned to new annotations based on technical similarity.

| EBI Annotation | EBI Annotation ID | Reassigned Annotation (ID) |
|------------------------------------|-------------------|------------------------------|
| coimmunoprecipitation | MI:0019 | Coimmunoprecipitation (COIP) |
| anti bait coimmunoprecipitation | MI:0006 | |
| anti tag coimmunoprecipitation | MI:0007 | |
| two hybrid | MI:0018 | Two Hybrid (TWO) |
| two hybrid array | MI:0397 | |
| barcode fusion genetics two hybrid | MI:2215 | |
| validated two hybrid | MI:1356 | |
| experimental interaction detection | MI:0045 | Unspecified (UNSPM) |
| docking | MI:0035 | |
| biophysical | MI:0013 | |

2.2.1.4 Interaction confidence scoring

The PPI data was then further processed to allow for confidence scoring. A coding script was developed to first remove repeated equivalent data entries, arising when a binary interaction originating from a single methodology and publication were replicated in the datasets undergoing processing. Replicated data entries occurred due to the same entries being curated into multiple primary databases and since this approach collated data from numerous of these data repositories, these cases would have persisted as replicate identifications of a PPI if disregarded, when in fact these entries are a single PPI identification report.

Next, each binary interaction was confidence scored based on three parameters (Table 2.2): the number of distinct methods used to detect an interaction (MS, method score), the number of publications that report the interaction (PS, publication score) and the likelihood the interactor is an affinity purification mass spectrometry (AP-MS) contaminant (CS, CRAPome score). For the method score (MS), each binary interaction scored a value of 1 or 2 based on whether a single or multiple reassigned method annotations were associated with that interaction, respectively. Likewise, for the publication score (PS), a value of 1 or 2 was assigned to the interaction based on whether one or multiple PubMed IDs were annotated to that interaction, respectively.

The third scoring parameter, the CRAPome score (CS), utilised the CRAPome repository (version 1.1) [112], a database of known contaminants of AP-MS experiments. Contaminant likelihood was negatively scored by querying each interactor detected *via* AP-MS against the CRAPome

repository to identify how common it was within the 411 CRAPome datasets (scored on January 18th 2017). If the interactor was present within >50% of the datasets and had not been detected by an alternative method, it was scored -1; if the interactor was present within >50% of the datasets but had been detected by an alternative distinct method or if it was present within 30-50% of the CRAPome datasets, it was scored -0.5; and if the interactor was present in <50% of the datasets it was assigned a CS of 0 (interactors not detected *via* AP-MS were also assigned a CS of 0). Refer to Table 2.2 for further detail regarding the CS assignment. These three scores (MS, PS, CS) were then added to generate a final confidence score.

Table 2.2 – Overview of confidence score assignment. Conditions for scoring each PPI data entry, based on three scoring parameters: Method Score (MS), Publication Score (PS) and CRAPome Score (CS). The final score equates to $MS + PS + CS$

| Score Type | Condition | Assigned Score |
|------------------------|---|----------------|
| Method Score (MS) | Single method used | 1 |
| | Multiple distinct methods used | 2 |
| Publication Score (PS) | Single publication record | 1 |
| | Multiple publication records | 2 |
| CRAPome Score (CS) | Interactor reported in >50% of CRAPome datasets and not detected by an alternative (non-AP-MS) method | -1 |
| | Interactor reported in >50% of CRAPome datasets and detected by an alternative (non-AP-MS) method | -0.5 |
| | Interactor reported in 30-50% of CRAPome datasets and not detected by an alternative (non-AP-MS) method | |
| | Interactor reported in <50% of CRAPome datasets and detected by an alternative (non-AP-MS) method | 0 |
| | Interactor reported in <30% of CRAPome datasets and not detected by an alternative (non-AP-MS) method | |
| | Interactor not reported by AP-MS | |

2.2.1.5 Network output

For generating the literature-derived network, a confidence threshold was set on the scored PPI datasets, only interactors that exceeded a confidence score of 2 were mapped as nodes within the network. Hence, the network represents a subset of reported PPIs from literature-derived data sources which had elevated confidence-weighting. Network visualisations were generated using Cytoscape [113] (version 3.3.0).

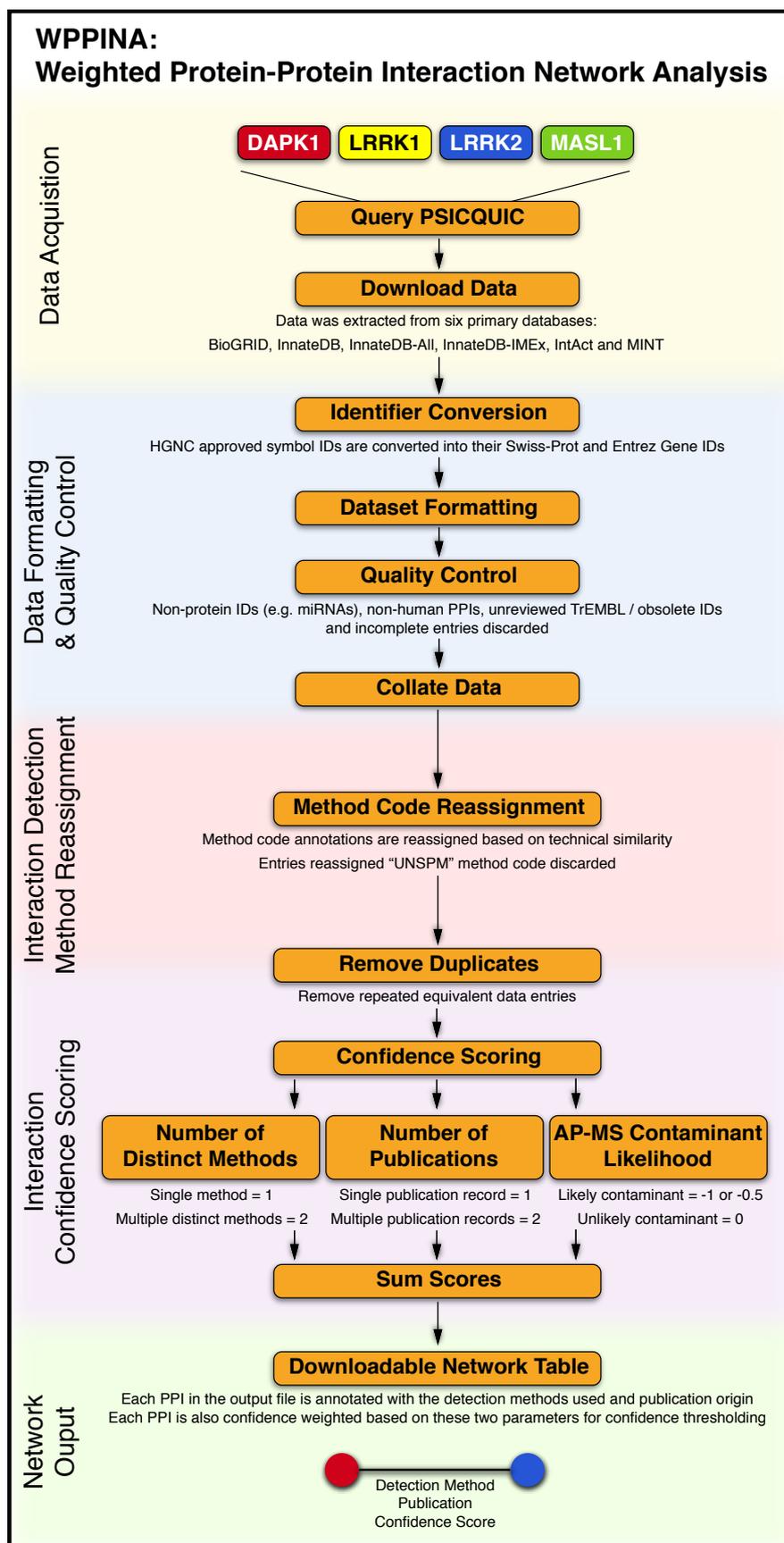


Figure 2.2 – Weighted Protein-Protein Interaction Network Analysis (WPPINA) pipeline. Outline of the data mining, processing and confidence scoring procedure underlying WPPINA, see Chapter Two – ‘2.2.1 - Constructing the literature-derived network’ for further detail.

2.2.2 Generating the experimental network

An experimental ROCO protein PPI network was generated based on novel protein microarray derived data. The PPI data which underlies this experimental network were generated by Dr Sybille Dihanich, Dr Alexandra Beilina and Dr Mark Cookson at the National Institutes of Health (NIH), USA. Materials and methods for this process are summarised below to provide context for these results. I then utilised this post-processed data for PPI network analysis in this project.

2.2.2.1 Protein production and purification

As described in reference [12], 3xFLAG-tagged DAPK1, LRRK1, LRRK2, MASL1 or GFP were transfected into HEK293T cells using polyethylenimine (PEI). Cells were harvested 24 hrs after transfection and lysed using lysis buffer: 20mM Tris (pH 7.5), 150mM NaCl, 1mM EDTA, 1% Triton X-100, 10% glycerol, 1x protease inhibitor cocktail (Roche) and 1x phosphatase inhibitor cocktail (Thermo Scientific). Lysates were precleared by centrifugation at 20 000 x g for 10 mins at 4°C, then incubated with EZview Red Protein G beads (Sigma-Aldrich) for 1 hr at 4°C to remove proteins which bind non-specifically to agarose. Following preclear with Protein G beads, lysates were incubated with EZview Red ANTI-FLAG M2 beads (Sigma-Aldrich) for 1 hr at 4°C to pull-down FLAG fusion proteins of interest. Beads were then washed six times using wash buffer: 20mM Tris (pH 7.5), 400mM NaCl and 1% Triton X-100, and protein eluted using elution buffer: 25mM Tris (pH 7.5), 150mM NaCl and 100µg/ml 3xFLAG peptide (Sigma-Aldrich). Protein yields and purity were estimated by gel staining with Coomassie brilliant blue (Thermo Scientific) (Figure 2.3).

2.2.2.2 Protein microarray

Once purified, 6µg of each either 3xFLAG-tagged full-length DAPK1, LRRK1, LRRK2, MASL1 or GFP were used to probe protein microarrays (Protoarray, version 4.1; Invitrogen) in accordance with manufacturer's instructions with the modification that following 3xFLAG-tagged protein probing, arrays were probed with monoclonal ANTI-FLAG BioM2 (Biotin Clone M2) antibody (Sigma-Aldrich) then with Alexa Fluoro 647 streptavidin (Invitrogen). Microarrays were imaged using an Axon GenePix 4000B fluorescence scanner and images were acquired using GenePix Pro software. Significant hits were determined using ProtoArray Prospector software and Z-scores, signal difference in standard deviations above background fluorescence, were used as estimates for binding strength. Proteins on the array were immobilised in duplicate and reported values were averaged for both data points.

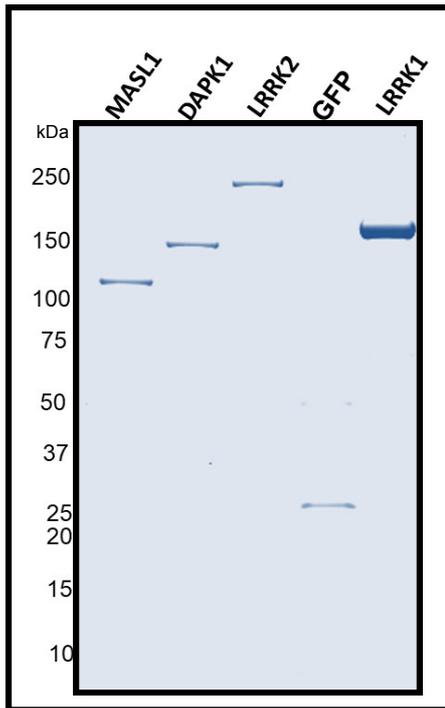


Figure 2.3 – Purification of 3xFLAG-tagged proteins. Purified 3xFLAG-tagged MASL1, DAPK1, LRRK2, GFP and LRRK1 subjected to SDS-PAGE and subsequently stained with Coomassie brilliant blue. Experiment performed by Dr Sybille Dihanich and Dr Alexandra Beilina at the NIH.

2.2.2.3 Network construction

A PPI network was constructed to represent potential ROCO protein interactors based on the protein microarray data. The array data was first filtered against the GFP positive hits, used as a negative control, to identify hits specific to DAPK1, LRRK1, LRRK2 and/or MASL1. Furthermore, a confidence threshold of $Z < 3$ was assigned to the datasets and only proteins associated with signals that exceeded this threshold were considered potential interactors. Each positive hit was then mapped to its corresponding Uniprot ID, hits corresponding to hypothetical proteins were discarded and the Uniprot IDs were converted to its corresponding HGNC approved symbol (gene name), using a coding script previously described (see ‘data formatting and quality control’). These gene names were then utilised for constructing the experimental network, using Cytoscape (version 3.3.0).

2.2.3 Creating the common core network

Data from both literature-derived and protein microarray network analysis approaches were considered in parallel to identify nodes common to both data sources. These nodes were mapped onto a network (the common core network) with particular detail annotating which data source the interactor originated, i.e. derived from published literature or our protein microarray experiments. In contrast to the literature-derived network, the common core network was inclusive of literature-derived data regardless of confidence score.

2.2.3.1 Coexpression analysis

In addition to using this data integration strategy to prioritise nodes of interest, the nodes of the common core network were assessed in terms of their tissue-specific mRNA expression pattern in relation to their associated seed protein mRNA expression. This co-expression profiling, performed by Dr Raffaele Ferrari (UCL Institute of Neurology, UK), utilised median gene expression values from the Genotype-Tissue Expression (GTEx) project [114] (data obtained 5th February 2018; GTEx V7 data release). This data was derived from RNA sequencing a wide variety of non-diseased tissue samples from 635 donor individuals, at a transcriptome-wide scale. The results are catalogued in the GTEx portal, which is a freely-accessible online resource.

For this analysis a stringent gene expression threshold of three reads per kilobase of transcript per million mapped reads (RPKM) was assigned to identify specific gene expression signal from background noise. Pairwise interactors were considered to be coexpressed when median mRNA expression levels for both interactors exceeded this 3 RPKM threshold. Expression levels were assessed across 13 tissue types: bladder, brain, heart, kidney (cortex), liver, lung, skeletal muscle, skin, intestine, spleen, thyroid, reproductive apparatus and whole blood. Data analysed for brain, heart, skin, intestine and reproductive apparatus tissue types were mean averages of the subcategorised tissue types available at the time of query.

2.2.4 Functional enrichment analysis (FEA)

Enrichment analysis of gene ontology (GO) biological process (BP) annotations was performed to gather functional insight into specific aspects of the ROCO protein networks. This approach analyses GO BP term annotations associated with a sample protein input list in comparison to a reference set, in this case, the entire genome, to determine functional terms which are enriched within a query list.

2.2.4.1 g:Profiler and functional annotation reassignment

FEA was predominantly performed using g:Profiler g:GOST [115] (on 23rd June 2017), an online web resource for functional profiling of gene/protein lists. Despite many resources available for this application [116,117], g:Profiler was chosen because it is actively maintained and functional annotations are regularly updated, unlike numerous of the alternative available tools. The analysis parameters were as follows: statistical significance was determined by the Fisher's exact

test with the g:SCS algorithm for multiple testing correction; $p < 0.05$ was assigned as the significance threshold; and the output data was not subjected to hierarchical filtering.

To aid interpretation of the FEA results, significantly enriched GO BP terms were classified *via* an in-house ontology which grouped functionally similar annotations into functional blocks and further subcategories, semantic classes. The coding script which underlies this process relies on semantic classification of GO terms, a dictionary list of GO BP terms was mapped to corresponding reassigned functional annotations that were manually curated (Table 2.3). This approach enables a clearer overview of the functional landscape enriched within a query protein set, yet retains GO term-specificity. Terms grouped into the ‘general’, ‘enzyme’, ‘physiology’ and ‘metabolism’ functional blocks provided very limited functional insight and were discarded from functional enriched map visualisations.

2.2.4.2 Panther and WebGestalt cross-validation

Two alternative FEA tools, Panther [118] and WebGestalt [119], were also utilised (on 22nd November 2017) to cross-validate the initial enrichment results *via* distinct approaches. These resources implement different, more conservative statistical methodology to correct for multiple testing to the one used in g:Profiler, which is a custom g:Profiler algorithm [120]. The analyses *via* both Panther and WebGestalt, utilised the over-representation enrichment test with Bonferroni correction. The entire genome was used as the reference set for comparison in all cases. This replication strategy enabled identification of major differences in enrichment results which may be due to FEA tool-specific biases.

Table 2.3 – Snapshot of functional term reannotation. *GO terms are grouped into functional blocks and semantic classes based on functional similarity.*

| GO term | GO ID | Reassigned functional block – semantic class |
|---------------------------------------|------------|--|
| regulation of apoptotic process | GO:0042981 | cell death - apoptosis |
| execution phase of apoptosis | GO:0097194 | |
| inflammatory cell apoptotic process | GO:0006925 | |
| autophagosome assembly | GO:0000045 | waste disposal - autophagy |
| positive regulation of macroautophagy | GO:0016239 | |
| autophagosome organization | GO:1905037 | |
| chaperone-mediated autophagy | GO:0061684 | |
| regulation of multi-organism process | GO:0043900 | general |
| cellular process | GO:0009987 | |
| regulation of molecular function | GO:0065009 | |

2.3 Results

Using a multi-layered methodological approach, a collection of ROCO protein interaction network analyses were performed and are presented hereafter. These provide a comprehensive overview of the existing identified ROCO protein interaction landscape, incorporating a novel contribution of potential PPIs derived from protein microarray analysis. These network analyses provide the foundation for gathering functional insight into the ROCO protein family. Furthermore, integrating data from these different approaches allows for prioritisation of protein interactors for further investigation.

2.3.1 Literature-derived ROCO protein interaction network

The literature-derived protein interaction network (Figure 2.4D) provides a confidence-weighted overview of the interactors that have been identified and reported in published literature for the human ROCO proteins. This network was constructed using a custom designed bioinformatic pipeline, Weighted Protein-Protein Interaction Network Analysis (WPPINA) [12,13] which has recently been developed into a freely available online PPI query resource, Protein Interaction Network Online Tool (PINOT) [121]. Each ROCO protein, DAPK1, LRRK1, LRRK2 and MASL1 were designated seed proteins and therefore these protein identifiers served as query inputs for this literature-mining procedure.

The WPPINA pipeline pools PPI data from six IMEx consortium associated databases, processes these data entries through quality control and filtering steps which enables a confidence score to be assigned to each interaction. This confidence score is based on the number of distinct methods used for interaction detection, the number of publications that report a specific interaction and the likelihood that the interaction detected is a false positive from an AP-MS experiment based on data within the CRAPome repository [112]. Only interactions that were replicated by method and/or publication record, and were unlikely to be AP-MS contaminants exceeded the confidence threshold set for generating the network.

The literature-derived network (Figure 2.4D) displays a strong bias for interactors within the LRRK2 interactome, in comparison to the other three ROCO protein interactomes. 113 interactors exceeded the confidence threshold and were mapped into the network for LRRK2, whereas 38, 14 and 4 interactors were mapped into the DAPK1, LRRK1 and MASL1 interactomes, respectively (Figure 2.4C). This skewed network topology is likely driven by literature bias, for example, a protein of great interest (especially if associated with human disease) will be

extensively studied, including investigations into its interacting molecules, and hence findings of this nature are regularly reported in the published literature, giving rise to an abundance of PPI data for this protein of interest in data repositories. The relationship between the number of PubMed entries and the number of reported interactors for each ROCO protein supports this hypothesis (Figure 2.4B). In this case, LRRK2 is a widely studied protein in relation to understanding and potentially developing therapeutics for Parkinson's Disease [122], whereas MASL1 is comparably neglected [46]. Alternatively, LRRK2 may have a broader interaction profile merely due to its greater molecular mass, in comparison to the other three ROCO proteins.

Interestingly, this topological trend evident in the network differs when considering the seed protein interactomes prior to applying the confidence threshold, i.e. all reported interactors regardless of confidence score. For example, the LRRK1 interactome is comparably larger than the DAPK1 interactome prior to confidence thresholding, 85 interactors versus 57 interactors, respectively (Figure 2.4A and 2.4C). Whereas upon confidence thresholding, the DAPK1 interactome consists of a greater number of interactors (Figure 2.4C and 2.4D). The retention of interactors following confidence thresholding reflects the experimental efforts into interactor discovery and characterisation. This retention of interactors is much greater for DAPK1 and MASL1, whereby 66.7% and 57.1%, respectively, of the reported interactors exceed the confidence threshold. In comparison, only 16.5% and 23.5% of the LRRK1 and LRRK2 interactors, respectively, have been replicated and hence persist into mapping the network.

Furthermore, this literature mining and subsequent PPI network mapping approach allows for straightforward identification of common interactors between proteins of interest. Here, in the case of the human ROCO proteins, five common interactors were extracted from the literature and persisted through the pipeline to mapping the network (Figure 2.4D). Of these five interactors, three were common to DAPK1 and LRRK2 (FADD, MYO1B and MYO1D) and two were common to LRRK1 and LRRK2 (BAG5 and HSPA8). Functional detail of these common interactors is summarised in Table 2.4, providing insight into potential overlapping cellular functions amongst the human ROCO proteins.

Figure 2.4 – Figure legend continued (A) Network visualisation of human ROCO protein interaction data extracted and processed using the Weighted Protein-Protein Interaction Network Analysis (WPPINA) pipeline prior to confidence thresholding. Data extracted 12th January 2017. (B) Relationship between number of PubMed entries and number of interactors in primary databases (determined using WPPINA pipeline) for each ROCO protein. The two variables are date matched to the data extraction date listed above. $r = 0.977$, $p = 0.023$ upon Pearson’s correlation testing. (C) Quantification of ROCO protein interactors before and after applying the replication confidence threshold, confidence scoring assigned as part of the WPPINA pipeline. (D) Network visualisation of human ROCO protein interaction partners following data processing via the WPPINA pipeline, confidence thresholded to include only interactions that have been replicated.

Table 2.4 – Functional overview of common interactors within literature-derived network.

| Seeds | Common Interactor (UniProt Swiss-Prot ID) | Functional Overview | References |
|---------------|---|---|------------|
| DAPK1 & LRRK2 | FADD (Q13158), FAS-associated death domain protein | Apoptosis, necrosis, autophagy, inflammation, innate immunity, cell proliferation, cell cycle progression, lipid metabolism | [123,124] |
| | MYO1B (O43795), Unconventional myosin-Ib | Intracellular transport, cell motility, cytoskeletal organisation | [125] |
| | MYO1D (O94832), Unconventional myosin-I d | Intracellular transport, cell motility, cytoskeletal organisation | [125] |
| LRRK1 & LRRK2 | BAG5 (Q9UL15), Bcl-2 associated athanogene 5 | Protein chaperone | [126,127] |
| | HSPA8 (P11142), Heat shock cognate 71kDa protein | Protein chaperone, transcriptional repression, autophagy, endocytosis | [128] |

2.3.2 Experimental ROCO protein interaction network

The experimental protein interaction network (Figure 2.5A) represents novel data derived from protein microarray screens, whereby each of the four human ROCO proteins were purified and introduced to an array chip of 9480 immobilised proteins, independently. These screens were conducted through a collaboration with Dr Mark Cookson (National Institutes for Health, USA) and the data generated was utilised within this project. Positive hits (i.e. potential ROCO protein interaction partners) were defined as signals which exceeded the Z-score threshold ($Z < 3$), to separate out reliable signal from background noise, and had been filtered against GFP positive hits, which was used as a negative control to identify non-specific binding events. These data processing steps reduced the likelihood of false-positive hits persisting in the network.

Through this hypothesis-free approach 303 interactions were detected for the human ROCO proteins: 87 DAPK1 interactions, 51 LRRK1 interactions, 78 LRRK2 interactions and 87 MASL1 interactions (Figure 2.5B). From mapping the network (Figure 2.5A), it was evident that the distribution of nodes (226 in total) surrounding the seed proteins was relatively even in comparison to the network topology of the literature-derived network (Figure 2.4D) which was heavily biased in the direction of the LRRK2 interactome. This finding opposes the hypothesis that LRRK2 has a greater number of interactors due to its greater molecular mass, in relation to the other ROCO proteins. A further notable observation of the experimental network is the extent of common interactors between multiple ROCO proteins. These mutual nodes account for 23.5% of the entire network, suggesting a higher degree of overlap between interaction profiles than depicted in the literature-derived network. Furthermore, 8.4% of the nodes within the network are common to three or more seed protein and five nodes (2.2% of the nodes in the network) are mutual nodes connecting all four ROCO proteins (Figure 2.5A and 2.5C).

Moreover, due to the lack of existing PPI data for MASL1, all potential interactors identified within this screen for this seed protein are novel contributions to understanding the MASL1 PPI landscape. In contrast, for the other seed proteins, several of the potential interactors detected by protein microarray exist in the published literature (1 DAPK1 interactor, 3 LRRK1 interactors and 11 LRRK2 interactors), increasing confidence in these interactions based on this replication. Interestingly, a number of kinases were identified as potential MASL1 interactors within the experimental network (Table 2.5). In particular, six of these kinases, CLK1, LIMK1, MAP3K4, NEK11, ROR1 and STK25 were detected to be specific to MASL1.

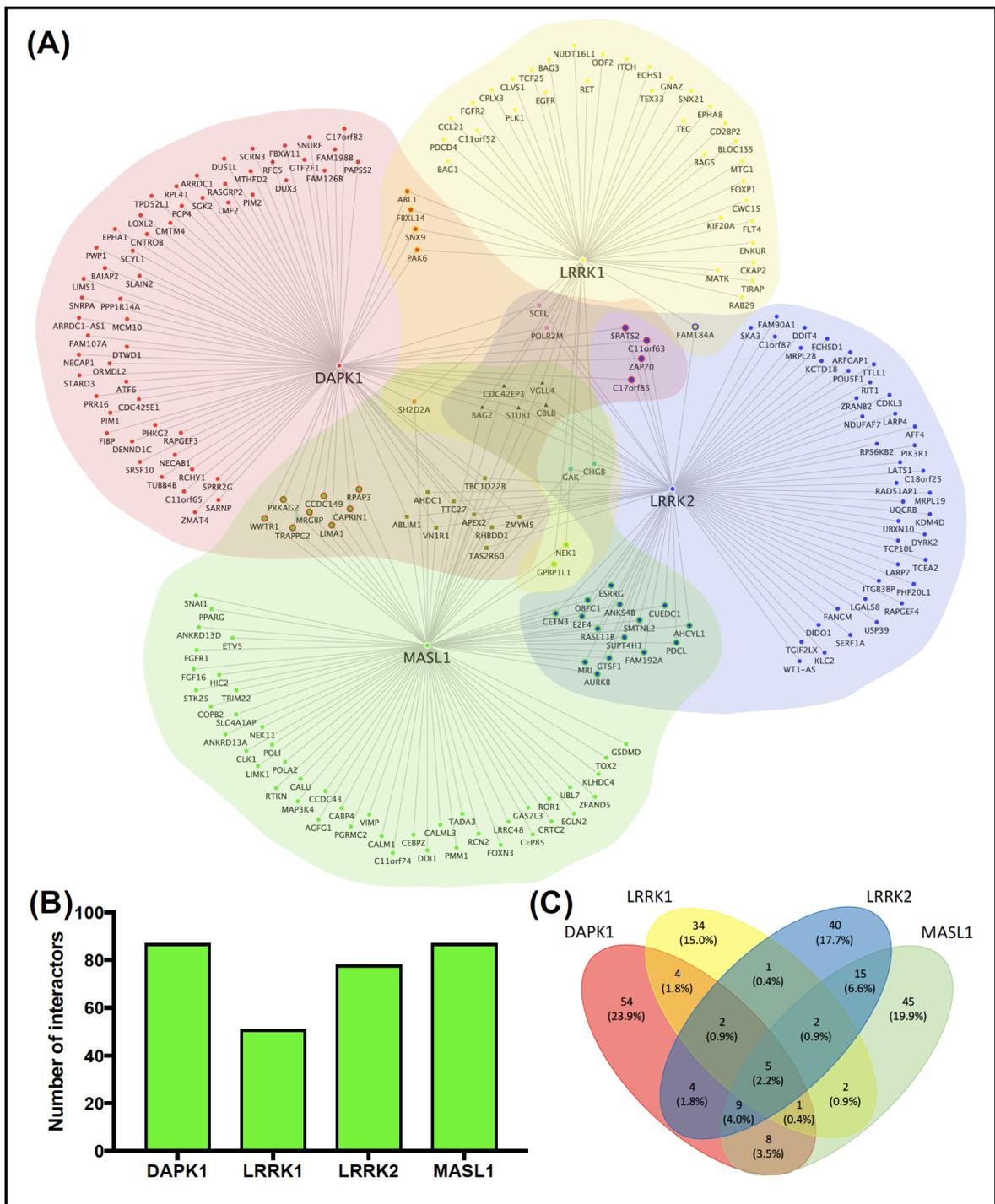


Figure 2.5 – Experimental human ROCO protein interaction network. (A) Network visualisation of the positive interactors identified for the human ROCO proteins using protein microarray experimentation (experimental network), utilising the four proteins of interest as prey proteins, independently. (B) Quantification of the number of positive interactors identified by protein microarray for each ROCO protein. 303 positive interactions identified across 226 nodes. (C) Venn diagram indicating the extent of overlap between ROCO protein interactomes in the experimental network. Number of interactors and percentage of entire network (in terms of nodes) presented.

Table 2.5 – MASL1-interacting kinases. Functional insight into kinases identified as potential MASL1 interactors based on protein microarray data. Shaded rows indicate kinases that are specific to MASL1.

| Kinase Interactor (UniProt Swiss Prot ID) | Additional Seed Interaction | Functional Overview | References |
|--|--|---|-------------------|
| AURKB (Q96GD4), Aurora kinase B | LRRK2 | Interacts with CLK1 (listed below) Mitosis and meiosis Catalytic core of the chromosomal passenger complex (CPC) | [129–131] |
| CLK1 (P49759), Dual specificity protein kinase CLK1 | - | Phosphorylates AURKB (listed above) RNA splicing | [129][132] |
| GAK (O14976), Cyclin-G-associated kinase | LRRK1 & LRRK2 | Catherin-mediated membrane trafficking Centrosome maturation and chromosome congression during mitosis | [133][134] |
| LIMK1 (P53667), LIM domain kinase 1 | - | Cytoskeleton dynamics in mitosis and meiosis | [135][136] |
| MAP3K4 (Q9Y6R4), Mitogen-activated protein kinase kinase kinase 4 | - | p38/MAPK and JNK pathways Apoptosis Development | [137][138] |
| NEK1 (Q96PY6), Serine/threonine- protein kinase Nek1 | LRRK1 | DNA damage response Meiosis Cilium assembly | [139–142] |
| NEK11 (Q8NG66), Serine/threonine- protein kinase Nek11 | - | DNA damage response Meiosis | [140][143] |
| ROR1 (Q01973), Inactive tyrosine- protein kinase transmembrane receptor ROR1 | - | Pseudokinase Transmembrane receptor in Wnt signalling pathway Auditory hair cell development | [144][145] |
| STK25 (O00506), Serine/threonine- protein kinase 25 | - | Lipid metabolism Cell migration Cell polarity | [146][147] |

2.3.3 Common core ROCO protein interaction network

In order to maximise the potential of the protein microarray data and to reduce the burden of approach-specific limitations, the two independent network analysis approaches were considered collectively. Overlaying the literature-derived data onto the experimental dataset enabled cross-validation of positive hits from the protein microarray screen, which lacked independent replication, with existing PPI data and identification of nodes that were common to both network analysis approaches but linked to different seed proteins in each. Presented in Figure 2.6A is the common core network, which represents common nodes across the literature-derived data prior to confidence thresholding and the experimental network.

These common nodes within the common core network can be categorised as: (i) interactors of the same seed protein which are present in both network approaches (these are indicated by the dotted and dashed green edges in Figure 2.6A, e.g. LRRK2 interactors: ARFGAP1, CHGB and GAK); (ii) interactors which are common to the two network analysis approaches but within the interactomes of different seed proteins (when considering unthresholded literature-derived data, 44 interactors [Figure 2.4A]; when considering confidence thresholded literature-derived data, 14 interactors [Figure 2.4D and 2.6B]); (iii) interactors that would exceed the confidence threshold (final score of >2 within the WPPINA pipeline) if the protein microarray data was integrated into literature-derived data and hence representing an independent replication (this was the case for seven LRRK2 interactors [Table 2.6]). It is of note that since these analyses were carried out, the protein microarray data has been curated into the IntAct data repository.

To provide further insight into the likelihood of the binary interactions within the common core network, pairwise coexpression profiling was performed based on tissue-specific gene expression data. Utilising mRNA expression data from the GTEx project [114] and an expression threshold of 3 RPKM, this analysis identified distinct tissues where specific interactor mRNA was coexpressed with specific seed protein mRNA. Of the 44 nodes within the common core network, only one, DUX3, was not recognised as a GTEx query input. This resulted in analysis of 99 pairwise interactions across 13 tissue types.

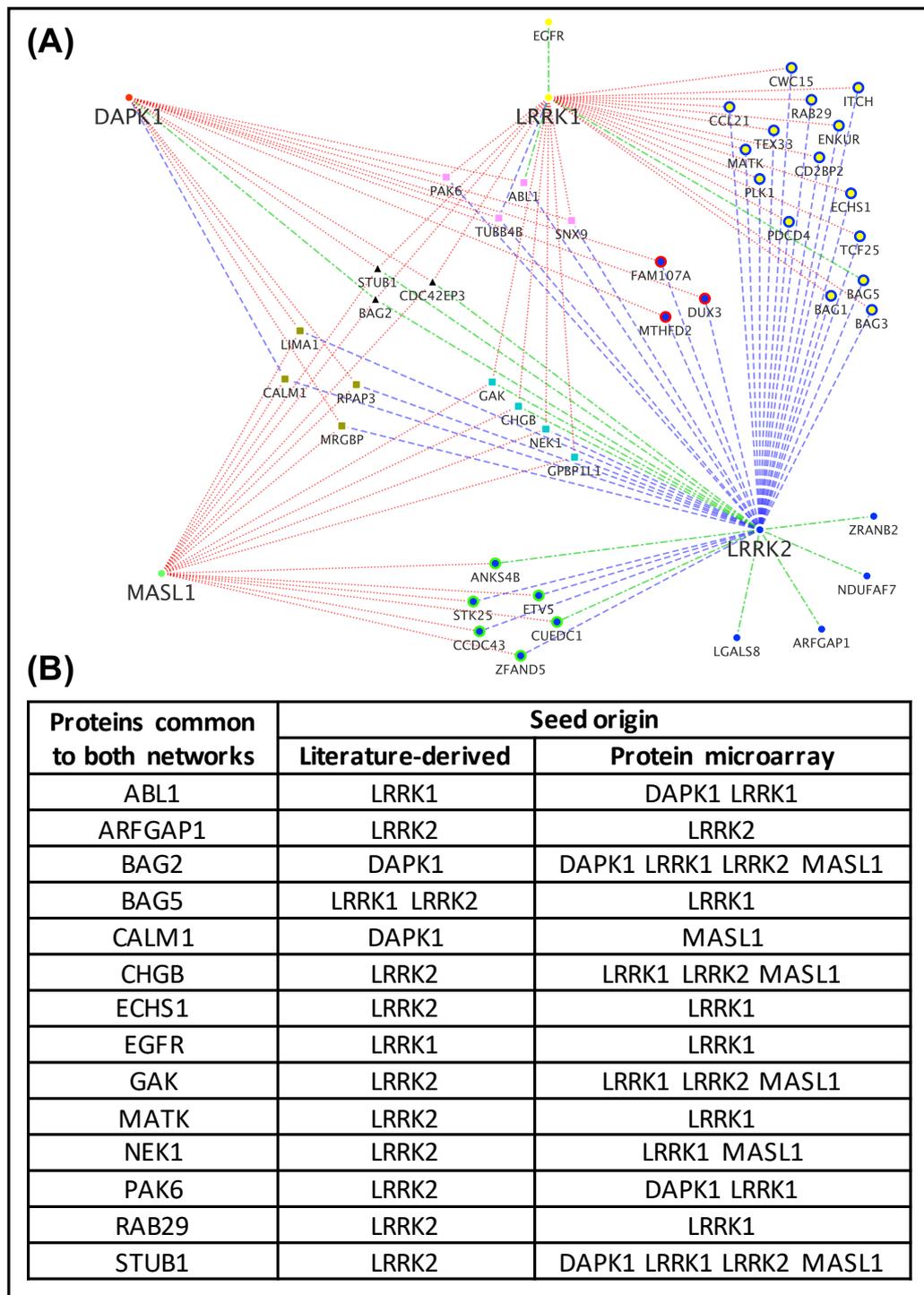


Figure 2.6 – Common core human ROCO protein interaction network. (A) Common core network consisting of common nodes across unthresholded literature-derived and experimental protein microarray data. Dotted red edges indicate protein microarray-derived interactions; dashed blue edges indicate literature-derived interactions; dotted and dashed green edges indicate replicated interactions derived from both protein microarray and literature datasets. Nodes are represented with a circular node if common to one seed, double circular node if common to two seeds, square node if common to three seeds and triangular node if common to all four seeds. (B) Common nodes across the thresholded literature-derived and experimental networks.

The mRNA expression of all four seed proteins exceeded the 3 RPKM threshold in all tissue types analysed, with the exception of skeletal muscle, hence the lack of coexpression signal across all pairwise interactions (Figure 2.7). On average, coexpression of pairwise interactors was evident in 9 (out of 13) tissue types analysed. For 9 pairwise interactors, predominantly DAPK1 interactors, coexpression was evident across 12 tissue types analysed. Whereas 5 pairwise interactors were limited to coexpression in just one or two tissue types analysed, potentially suggesting a specialised role for these proteins in relation to specific tissue types. The highest level of coexpression, in terms of proportion of pairwise interactors analysed, was in the reproductive apparatus (97.0%), brain (93.9%) and intestine (91.9%).

Table 2.6 – LRRK2 interactors identified via protein microarray that have been previously reported with the literature. These interactors would exceed the confidence threshold (final score of >2) assigned within the WPPINA pipeline if the protein microarray data was integrated into the literature based dataset. ⁺unthresholded literature data; *interaction in literature-derived network

| LRRK2 Interactor (UniProt Swiss Prot ID) | Additional Seed Interaction | |
|---|-----------------------------|----------------------|
| | Literature ⁺ | Protein Microarray |
| ANKS4B (Q8N8V4), Ankyrin repeat and SAM domain-containing protein 4B | - | MASL1 |
| BAG2 (O95816), BAG family molecular chaperone regulator 2 | DAPK1* LRRK1 | DAPK1 LRRK1 MASL1 |
| CDC42EP3 (Q9UKI2), Cdc42 effector protein 3 | LRRK1 | DAPK1 LRRK1 MASL1 |
| CUEDC1 (Q9NWM3), CUE domain-containing protein 1 | - | MASL1 |
| LGALS8 (O00214), Galectin-8 | - | - |
| NDUFAF7 (Q7L592), Protein arginine methyltransferase NDUFAF7, mitochondrial | - | - |
| ZRANB2 (O95218), Zinc finger Ran-binding domain-containing protein 2 | - | - |

2.3.4 Enriched functions within the ROCO protein interaction networks

To gain functional insight into the ROCO protein interactomes presented in the PPI networks, functional enrichment analysis (FEA) was performed based on GO BP annotations. This analysis was initially undertaken utilising g:Profiler [115] with subsequent cross-validation using Panther [118] and WebGestalt [119] (see '2.2.4 Functional enrichment analysis (FEA)' for further detail). The data presented corresponds to the g:Profiler derived FEA. Significantly enriched GO BP terms were grouped into functional blocks and then further categorised into semantic classes based on semantic similarity amongst GO BP terms (Table 2.3), enabling an overview of the ROCO protein functional landscape (Figure 2.8A).

FEA of the literature-derived network resulted in 516 significantly enriched GO BP terms which span a diverse range of biological processes (Figure 2.8A). The most significantly enriched terms associate with intracellular organisation and transport related functions (Table 2.7; complete functional enrichment result available in Supporting File S2). In addition, each seed protein interactome (DAPK1, LRRK1, LRRK2 and MASL1) was dissected from the literature-derived network for independent FEA. This seed-specific functional insight highlights cell death and development as significantly enriched functions within the DAPK1 and LRRK1 interactomes, respectively. Intracellular organisation and transport related annotations were most significantly enriched within the LRRK2 interactome (Table 2.7; complete functional enrichment results available in Supporting Files S3 [DAPK1], S4 [LRRK1], and S5 [LRRK2]). As expected, due to the size of the MASL1 interactome (5 nodes in total), no functional annotations were significantly enriched within this sample set. For annotations to reach statistical significance in this enrichment analysis the sample protein set annotations were required to be over-represented in relation to the entire genome (reference set) annotations and withstand multiple testing correction, hence this method has limited power in small sample set cases [148].

Finally, the common core network was subjected to FEA. Since this network was comprised of common nodes between the ROCO proteins, this analysis revealed enriched functions that appear to be common across the four ROCO protein interactomes, which may shed light on the relationship between common structural domains and shared functions. A collection of 63 GO BP terms were significantly enriched. These terms cover a more limited functional landscape than was the case for the literature-derived network FEA (Figure 2.8B) and provide specific insight into a subset of potentially convergent functions associated with ROCO proteins. The top hits, in terms of enrichment significance, were functional terms associated with stress response (Table 2.7; complete functional enrichment result available in Supporting File S6).

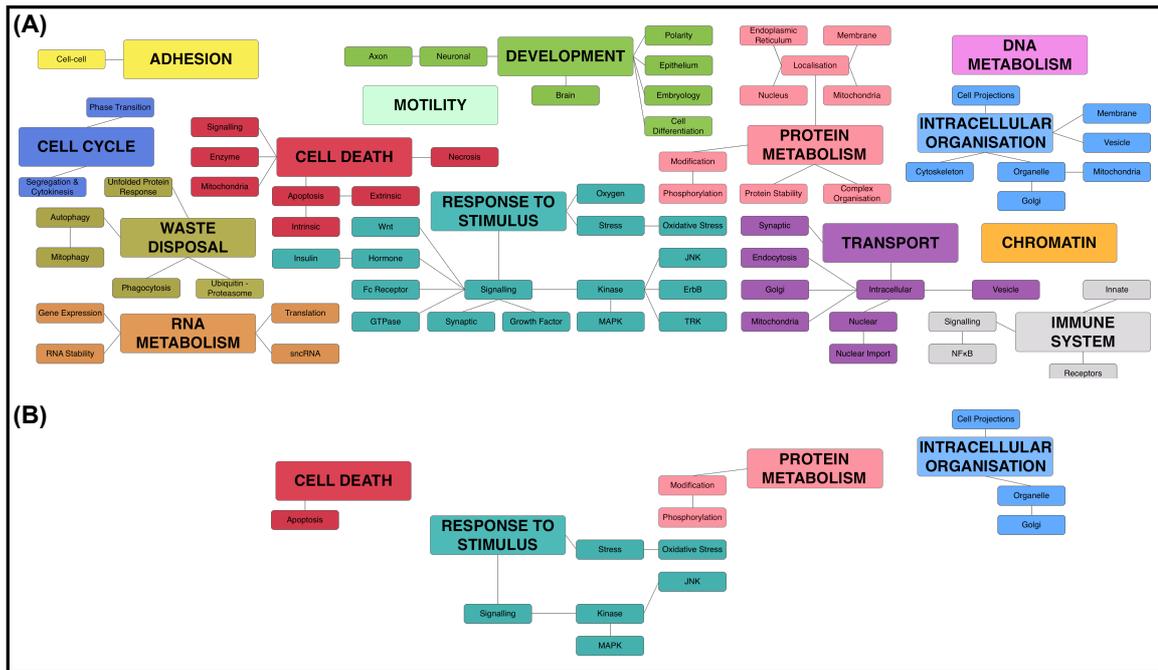


Figure 2.8 – Human ROCO protein interactome functional enrichment map. (A) An overview of the enriched functions associated with the literature-derived network. (B) Enriched functions of the common ROCO proteins interactors of the common core network; insight into the functional overlap amongst the human ROCO proteins. Terms presented represent the functional blocks (uppercase terms) and semantic classes (lowercase terms) that the significantly enriched GO BP terms have been grouped into, utilising g:Profiler coupled to an in-house functional grouping process.

Table 2.7 – Top three significantly enriched GO BP terms in various ROCO protein interactome functional enrichment analyses. A breakdown of the most significantly enriched terms and corresponding functional block – semantic class grouping for the literature-derived network, subset interactomes within this network and the common core network, analysed using g:Profiler. ⁺interactomes based on the literature-derived network.

| Dataset | GO BP term | p-value | Functional block - semantic class |
|--------------------------------|--|------------------------|-----------------------------------|
| Literature-derived network | cellular component organization or biogenesis | 4.60×10^{-36} | Intracellular organisation |
| | intracellular transport | 2.44×10^{-30} | Transport - intracellular |
| | cellular component organization | 4.31×10^{-30} | Intracellular organisation |
| DAPK1 interactome ⁺ | cell death | 3.85×10^{-7} | Cell death |
| | apoptotic process | 4.39×10^{-7} | Cell death - apoptosis |
| | programmed cell death | 1.36×10^{-6} | Cell death |
| LRRK1 interactome ⁺ | neuron projection development | 3.49×10^{-5} | Development - neuronal - axon |
| | cell development | 3.82×10^{-5} | Development |
| | neurogenesis | 9.11×10^{-5} | Development - neuronal |
| LRRK2 interactome ⁺ | cellular component organization or biogenesis | 2.75×10^{-29} | Intracellular organisation |
| | intracellular transport | 4.36×10^{-29} | Transport - intracellular |
| | establishment of localization in cell | 1.61×10^{-26} | Protein metabolism - localisation |
| Common core network | cellular response to stress | 1.14×10^{-5} | Response to stimulus - stress |
| | regulation of cellular response to stress | 1.65×10^{-4} | Response to stimulus - stress |
| | stress-activated protein kinase signalling cascade | 1.15×10^{-3} | Response to stimulus - stress |

2.4 Discussion

The PPI network analyses performed in this study broadened our understanding of the human ROCO protein interactome. By exploring the interaction profiles of these structurally related proteins collectively, new insights were gathered into the commonalities and distinctions present in their proximal interactomes, which was further used to determine shared and exclusive functional associations. In addition, the overall strategy adopted, whereby existing data was utilised to complement novel high-throughput experimental data represents an example of how data integration from previous experimentation reported within the published literature can maximise the potential of newly generated data.

2.4.1 The utility of WPPINA for analysis of the human ROCO proteins

The literature-derived ROCO protein interaction network (Figure 2.4D) represents the extent of DAPK1, LRRK1, LRRK2 and MASL1 interactors from published studies which have been retained following data processing and confidence thresholding *via* the WPPINA pipeline. Three standout aspects of this bioinformatic pipeline are the retention of high quality data entries only, wide coverage of data collated from numerous primary databases, and a confidence score threshold based on interaction replication. Hence the presented network illustrates a robust reflection of the current knowledge for the ROCO protein interaction profiles.

A particular benefit of mapping protein interaction networks of multiple seed proteins using WPPINA is the straightforward identification of common interactors, which would likely be overlooked if surveying literature-derived PPI data using alternative methods. Since the primary structure of the ROCO proteins feature a number of common domains, for example the ROC-COR region which defines the protein family and the leucine-rich repeats (LRR) in LRRK1, LRRK2 and MASL1, common interactors may shed light on potential domain-specific interaction interfaces and shared functions amongst the ROCO proteins (Table 2.4). Interestingly, the potential for heterodimerisation of DAPK1, LRRK1 and LRRK2 (in all three combinations) was evident from this network, suggesting a role for the ROC-COR tandem domain in this conformation. Within the published studies which report these interactions, further characterisation into the interaction interfaces indicated that the ROC-COR region is critical for dimerisation [35].

Two of the DAPK1 and LRRK2 common interactors were unconventional myosin proteins, MYO1B and MYO1D, which are involved in vesicle trafficking (Table 2.4). Of note, Rab proteins

contribute to the regulation of myosin motor function and recent evidence suggested LRRK2 phosphorylates a subset of Rab proteins [149]. This, together with myosin proteins as LRRK2 interactors, supports previous suggestions of LRRK2 playing an important role in intracellular vesicular trafficking [150] and from this PPI network analysis highlights DAPK1 as a potential contributor to the same or related function. In addition, since Rab proteins are increasingly considered *bona fide* LRRK2 substrates [151], an open question persists as to whether LRRK1, which is paralogous to LRRK2 including a highly conserved kinase domain, also impacts the phosphorylation state of Rab proteins. A recent study supports this notion, reporting Rab7a as a LRRK1 substrate [38].

2.4.2 Adding novel insight from ROCO protein microarray data

The generation and analysis of an experimental network (Figure 2.5A) provided novel insight into potential interactors of the ROCO proteins. This hypothesis-free approach evaluates the interaction profile of each seed protein based on an equivalent experimental setup, diluting the effect of ascertainment bias present in the literature-derived network. In particular, insight into the potential MASL1 interactome was massively enhanced through this analysis, since the literature-derived data is extremely limited. In addition, the increased proportion of commonality evident between seed protein interaction profiles comparative to the literature-derived network was intriguing and likely to be a more accurate representation of the capability for ROCO proteins to interact with common nodes, based on the structural organisation of these seed proteins.

In relation to the potential interactors within the MASL1 interactome, numerous kinases were identified and a subset of these kinases are specific to MASL1. These are of particular interest since MASL1 is the only ROCO protein which lacks an intrinsic kinase domain and since the other ROCO proteins possess intra-regulatory cross-talk between the GTPase activity of the ROC domain and the kinase activity [30,31,152], the MASL1 GTPase activity may harbour an equivalent reciprocal relationship with an extrinsic kinase. Further validation to test this hypothesis would determine whether any of these kinases are downstream effector molecules of MASL1.

Although this experimental network revealed a fresh perspective on the ROCO protein interactome, the literature-derived network provided a more robust representation of the ROCO protein interaction profiles to date. This is because the nodes within the experimental network lack independent replication and are sensitive to inherent limitations of the protein

microarray technique, therefore need to be considered as potential interactors that require further validation by alternative PPI detection methods. If reproducible utilising a distinct PPI detection method, this would increase confidence in these initial positive hits, similarly to the concept underlying the scoring process within the WPPINA pipeline. Example intrinsic biases faced during protein microarray experimentation include the choice of protein baits on the array chip, potential alterations to physiological protein conformations and the reliance on antibody detection methodology.

2.4.3 Data integration strategies and the common core network

Since both the literature-derived and experimental network analysis approaches sustain inherent biases, data from the two approaches were considered collectively to reduce the impact of these biases on network analysis. This resulted in the construction of the common core network (Figure 2.6A) which consisted of interactors common to both datasets. The utility of this strategy is two-fold: first to cross validate novel experimental data with the existing PPI knowledgebase, and second to prioritise and nominate interactors for further investigation. From mapping this network, 44 nodes were identified and the majority (89%) linked to at least two seed proteins. This indicated a subset of proteins from the experimental network which are potentially novel common interactors for multiple ROCO proteins based on supporting data within the literature, highlighting avenues for further investigation.

Within the common core network (Figure 2.6A), three proteins (BAG2, CDC42EP3 and STUB1) were mutual nodes for all four ROCO proteins, suggesting potential interaction profile convergence across the whole protein family. For BAG2 (in relation to the DAPK1 interaction [153]) and STUB1 (in relation to the LRRK2 interaction [154,155]), these interactors have also been identified by alternative PPI detection techniques other than protein microarray, which reduces the likelihood that these proteins were false positive hits in the experimental network. In addition, this strengthens the rationale for pursuing further investigation into the potential interaction of these proteins with the other ROCO proteins.

A further subset of proteins of particular interest were the 15 nodes common to both LRRK1 and LRRK2. These proteins were largely novel potential interactors for LRRK1 from the experimental network and previously reported interactors of LRRK2 based on published literature (Figure 2.6A), including interactors which exceeded the WPPINA pipeline confidence threshold (Figure 2.6B). Since the two LRRK proteins are structurally very similar, investigating the validity of these LRRK1 interactions using distinct methodological approaches highlights a further route for

further study. One of these common LRRK1 / LRRK2 interactors is RAB29, which supports a previous discussion point as to whether Rab proteins are phosphorylated by LRRK1, based on the catalytic interplay of LRRK2 and Rab proteins [151].

Since the spatial pattern of protein expression will influence the likelihood of proteins interacting, coexpression profiling of pairwise interactors within the common core network was performed. This analysis revealed similar and distinct coexpression trends for interactors in relation to different seed proteins (Figure 2.7). In particular, a relatively uniform coexpression pattern was evident for the BAG proteins analysed (BAG1, BAG2, BAG3 and BAG5) in relation to the ROCO proteins, which aligns with the role of BAG protein as molecular chaperones and protein complex adaptor elements [126]. Whereas CD2BP2 for example, displayed a distinct tissue-specific coexpression profile for LRRK1 in comparison to LRRK2, suggesting that despite overlapping interaction profiles, this degree of convergence is likely to be altered in different tissue types which may shed light on potential divergent functional profiles. This builds on a more physiological representation of interactome analysis.

Although this expression analysis represents a crude approach for assessing coexpression, it provided insight into the probability of pairwise interactions occurring based on coexistence of corresponding transcripts at a tissue-level. This analysis utilised transcript expression data due to the availability of expression profiles across the majority of the transcriptome at a tissue-specific level. However, protein coexpression profiling would have increased physiological relevance for this analysis, but a comprehensive catalogue of this data-type is currently not readily available.

2.4.4 Gaining functional insight

The functional landscapes of the literature-derived and common core networks were evaluated utilising GO BP annotations for FEA. These analyses provided insight into common and distinct functional pathways which are influenced by the proximal interactomes of the ROCO proteins. First, FEA of the literature-derived network which included reported interactors for all four ROCO proteins, revealed a diverse range of functional associations (Figure 2.8A) suggesting that the ROCO proteins (or due to the persistent LRRK2 bias within this network, LRRK2) may act as signalling hubs for a multitude of molecular cascades. As previously discussed, this network harbours ascertainment bias in relation to the LRRK2 interactome, which persists in the FEA of this network, hence the FEA results of the literature-derived network are largely representative of the LRRK2 interactome. This is supported when comparing the literature-derived network and

LRRK2 network FEA results (Table 2.7), whereby the two most significantly enriched annotations are identical.

Furthermore, single interactome FEA for DAPK1 and LRRK1 networks, subset from the literature-derived network, shed light on functional associations for those seed proteins specifically. These analyses highlighted cell death related and development related functions as enriched in the DAPK1 and LRRK1 interactomes, respectively (Table 2.7). These findings are particularly interesting in relation to LRRK1, since the functional characterisation of this protein is in its infancy. Upon subjecting the common core network to FEA, the most significantly enriched annotations related to the stress response within the cell (Table 2.7), uncovering potential convergence in relation to ROCO protein function; a concept which is plausible based on the structural overlap between these proteins. Further investigation is required to establish the extent of this potential shared functionality of the ROCO proteins.

The use of FEA to interpret gene lists in a functional context escalated in the mid-2000s, to a point whereby it is now often a routine analysis alongside high-throughput genomic, transcriptomic and proteomic analyses. This, together with ongoing functional annotation efforts which increases the usefulness of this approach, has led to wide choice of FEA tools available [116,117]. Each tool differs in relation to the level of maintenance, underlying algorithm and additional features, which can make it difficult to know which tool to rely on. g:Profiler [115] was chosen as the predominant FEA tool for this project, largely because it is actively maintained in relation to up-to-date GO annotations. This is particularly important for interpreting functional enrichment findings in the context of the current functional knowledgebase [156]. Due to the dynamic nature of functional annotation, enrichment results can quickly become outdated.

Furthermore in support of using g:Profiler for FEA, this tool utilises a custom multiple correction algorithm for assigning significance thresholds, termed g:SCS (set counts and sizes). This statistical method is optimised in relation to the discovery of false-positives and outperforms the widely used Benjamini-Hochberg false discovery rate and Bonferroni correction algorithms [120]. In addition and as previously described, to ensure FEA results were not significantly impacted by tool-specific biases, cross-validation of results was achieved using distinct resources.

The post-processing of the FEA output performed following these analyses, which involved grouping GO annotations into functional blocks and semantic classes, addressed two aspects of

the challenge of interpreting lists of significantly enriched GO terms. First, it provided a clearer overview of the types of functions enriched (Figure 2.8), including the extent to which a specific functional block contributed to the enrichment e.g. the proportion of cell death related functional annotations enriched in the literature-derived network FEA was 8.1%. Second, this approach grouped annotations with negligible specific functional meaning into a 'general' category, enabling these terms to be discarded prior to downstream interpretation of FEA results. Examples of non-specific functional annotations which provide very limited insight into the biological roles of proteins within a query sample set are 'negative regulation of cellular process', 'multicellular organismal process' and 'rhythmic process'.

2.4.5 Considerations for literature-derived PPI network analysis approaches

Despite the utility of WPPINA for constructing literature-derived networks, there are a number of caveats to this type of analysis. A key issue is the partial nature of the data underlying the analysis which strongly implies the PPI network is incomplete, since the analysis relies on literature-derived data and often many interactors within an interaction profile are yet to be discovered. For example, the physiological MASL1 interactome is likely to be greater than 4 interactors, however due to the lack of investigation into the MASL1 interaction profile, MASL1 PPI data is very limited. In contrast, proteins at the centre of extensive investigation, such as LRRK2 in this case, suffer from ascertainment bias in this analysis due to the skew in available data.

This ascertainment bias is further highlighted by the rate of growth in PPI identification reports for LRRK2: an analysis of the LRRK2 interactome using an earlier version of the same pipeline performed in July 2014, yielded 62 interactors [107], in comparison to the 113 interactors reported here (January 2017), hence an almost two-fold increase in 30 months. A more recent analysis using PINOT with the equivalent confidence threshold (July 2019) extracted 220 interactors (over 1400 interactors prior to confidence thresholding) for LRRK2. Furthermore, ongoing high-throughput projects, such as BioPlex [2,3], continue to capture increasing amounts of the human protein interactome. The latest unpublished version of the BioPlex network reported over 70 000 interactions from approximately 7500 independent AP-MS experiments.

Further considerations for interpreting PPI networks based on literature-derived data sources relate back to the quality of the database curation and the technical procedure which underlies the PPI detection. The process of database curation either operates automatically using computation or manually by curators. Both approaches are susceptible to poor quality data

curation, however manual curation involves a more thorough review and verification process. In addition, each PPI detection method has inherent limitations (Table 2.8) and PPIs detected under experimental conditions are to a certain degree artefactual. Therefore, how an experimental setup relates to the physiological setting is important to consider. To address these issues, the WPPINA pipeline comprises a rigorous quality control process and each PPI data entry is fully traceable back to the detection method used and the source publication.

2.4.6 Challenges and ongoing developments: WPPINA to PINOT

The development of a novel bioinformatic approach to mine and process literature-derived PPI data resulted in the WPPINA pipeline. This team effort was predominantly developed and trialled by Dr Claudia Manzoni and myself throughout the course of this project. As previously described (Figure 2.2), this data management and analysis strategy extracts PPI data from numerous primary databases, integrates this expanse of information, performs clear filtering and quality control checks and confidence scores each PPI based on detection method and publication records. This procedure then outputs a network file that can be easily uploaded into network visualisation software, such as Cytoscape [113].

One of the purposes of developing the WPPINA pipeline was to overcome challenges faced when attempting to incorporate and interpret PPI data from multiple repositories. A notable issue that was addressed during the development of the pipeline was the format inconsistencies in curation of different databases which hampered collating data from multiple sources. For example, the majority of the data repositories utilised in this analysis (Innate, IntAct and MINT) report the UniProt protein ID for each interactor, whereas BioGRID reports the NCBI Entrez gene ID. Since BioGRID was a considerable contributory resource towards the total PPI data available (Figure 2.1) and the curation effort is independent of the other data repositories, the absence of integrating this data would have reduced potential data coverage. Hence, the implementation of a ID conversion procedure within the pipeline (described in the '2.1 Material and Methods' section) to enable integration of data from the six data repositories for downstream data processing. The vision of the IMEx consortium includes standardisation across curation efforts, however BioGRID is currently a consortium observer as oppose to an active member.

Table 2.8 – Advantages and limitations of PPI detection methods. Examples of techniques for detecting PPIs with considerations regarding inherent limitations of each approach, expanded from reference [104]. Cross-validation of PPI detection across multiple methods reduces the impact of method-specific limitations and increases the confidence in a PPI event.

| PPI detection method | Advantages | Limitations |
|---|---|---|
| Coimmunoprecipitation (CoIP) | <ul style="list-style-type: none"> Obtain protein complexes in an almost physiological state, in relation to PTMs and conformation Flexibility in downstream detection method, e.g. Western blot or mass spectrometry Protein tag not required Relatively low cost | <ul style="list-style-type: none"> Requires high quality antibodies to be available for proteins of interest Relies on prediction of potential interacting partners Low affinity PPIs not detected <i>Ex vitro</i> / <i>Ex vivo</i> detection from lysate Undetected components of complex may persist |
| Protein Microarray | <ul style="list-style-type: none"> High-throughput, thousands of proteins assessed in parallel Hypothesis-free Able to detect low affinity PPIs Customisable arrays | <ul style="list-style-type: none"> Limited to arrays commercially available or proteins that can be immobilised Non-physiological setting Typically requires protein of interest to be tagged with epitope / fluorophore |
| Yeast Two-Hybrid (Y2H) | <ul style="list-style-type: none"> <i>In vivo</i>, in the context of yeast Well established technique Scalable to a high-throughput format Relatively low cost, initially just requiring yeast, two-hybrid/cDNA constructs and growth media | <ul style="list-style-type: none"> Proteins of interest must be capable of being expressed in yeast and are typically overexpressed PPI detection occurs in nucleus only System may lack the necessary PTM capability and accessory molecules Protein fragments may need to be used in the case of large or membrane-bound proteins |
| Affinity Purification – Mass Spectrometry (AP-MS) | <ul style="list-style-type: none"> Purification step can be done as in CoIP using native antibodies avoiding the need to tag proteins or using an epitope tag where antibodies are of poor quality or unavailable Protein complexes derived from physiological setting Detect multiple components of a complex | <ul style="list-style-type: none"> May require tagging protein of interest Low affinity PPIs not detected <i>Ex vitro</i> / <i>Ex vivo</i> detection from lysate Spatial expression profiles disrupted which may lead to artefactual PPIs occurring in lysate Require specialist expertise for interpreting mass spectra |
| Surface Plasmon Resonance (SPR) | <ul style="list-style-type: none"> Highly sensitive Able to detect low affinity and transient PPIs Binding tracked in real-time Label-free | <ul style="list-style-type: none"> Proteins of interest must be highly purified Artificial conditions setup in flow cell to mimic physiological setting Specialist equipment and expertise required |
| BRET / FRET / FLIM | <ul style="list-style-type: none"> Live cell imaging Provides a spatial and temporal representation of PPIs Highly sensitive Provides insight into proximity of two potential interaction partners | <ul style="list-style-type: none"> Proteins of interest must be fused to bioluminescent / fluorescent molecule Detection molecules at risk of photobleaching Complex experimental design in relation to choice of donor/acceptor molecules and controls |
| Proximity Ligation Assay (PLA) | <ul style="list-style-type: none"> <i>In situ</i> detection, physiological setting Able to detect low affinity and transient PPIs Single molecule resolution | <ul style="list-style-type: none"> Requires proximity probes to be of high quality to specifically detect proteins of interest Limited to probes commercially available or in-house development Dependence on enzymes for ligation and polymerisation of oligos |

Throughout the later stages of the development process the WPPINA pipeline was applied as a tool for PPI network analysis within numerous projects. For example, PPI network analysis was performed for frontotemporal dementia (FTD) associated genes [13], for the human ROCO proteins presented here [12], for Parkinson's disease (PD) associated genes [14] and for exploring the tau interactome [157]. As more studies utilised WPPINA as a strategy for gaining insight into PPI landscapes, the pipeline became more exposed to the breadth of PPI data available and in turn, the related challenges, such as poor data curation hampering data processing. Subsequently, we were able to develop the pipeline further to increase its robustness.

At this stage, the pipeline was limited to a semi-automated format that required the user to run the coding scripts in R. As this analysis strategy gained more interest, we wanted to increase the user-friendliness of this approach, enabling the wider research community to make use of the WPPINA pipeline. This led to the development of PINOT, a fully-automated freely-available online PPI query resource to optimise the utility of PPI data within the public domain (available at http://www.reading.ac.uk/bioinf/PINOT/PINOT_form.html) [121]. The most recent development of the WPPINA pipeline underlies this resource which is coupled to an easy-to-use web-interface which was implemented by Dr Liam McGuffin (University of Reading). PINOT is not targeted at a specific user background, we envisage that the straightforward to use input form and easy to interpret output file will be of benefit to a broad range of user types. In addition, the R scripts which underlie the tool are available from the website to enable transparency of the pipeline and to enable modification if desired.

The utility of PINOT was compared to three alternative PPI query resources, human integrated protein-protein interaction reference (HIPPIE) [158], molecular interaction search tool (MIST) [159] and search tool for recurring instances of neighbouring genes (STRING; Table 2.9) [160]. These other resources also utilise evidence based PPI data to enable construction of PPI networks, however each resource (including PINOT) has distinguishing features. Notably, the added value of PINOT is that the data is extracted at the time of query, for human data, hence the user obtains the most up-to-date representation of the available PPI data. Furthermore, the data output provided by PINOT is of high quality, largely due to the quality control procedures within the data processing pipeline, and therefore every PPI is annotated with its corresponding method detection annotation and publication identifier (PubMed ID).

In terms of performance, i.e. the number of PPIs extracted for a given query, the performance of PINOT is generally comparable to other resources [121]. In test cases, the performance of

PINOT is sometimes slightly lower than in HIPPIE and MIST, and this is due to the quality of the data curation within source databases. For example, if a curated data entry is missing its corresponding PubMed ID, this entry would be discarded from data processing in PINOT, whereas this is not the case in other resources. In other examples, the performance of PINOT hugely exceeds that of alternative resources, these cases usually arise when a high-throughput PPI screen has recently been curated into the source repositories but since alternative resources (HIPPIE and MIST) bank data as oppose to extracting it at the time of query, this data has not been captured, this has been observed in relation to LRRK2 [121].

The most recent developments of PINOT include a stringent or lenient filtering option, required when completing the input form. This stringency filter relates to the method detection annotations and upon selecting the lenient option, method detection annotation corresponding to non-specific method detection annotations which are grouped into the ‘unspecified’ category as part of the method reassignment process, which provide limited detail such as ‘experimental interaction detection’ would be retained. This increases data coverage at the expense of data quality. Furthermore, an additional query species option was recently incorporated, for *Caenorhabditis elegans* PPIs (see ‘3.2.2 Developing a *C. elegans* query option for PINOT’).

Table 2.9 – Comparison of features across PPI query resources. PINOT compared against its closest related PPI query tools, human integrated protein-protein interaction reference (HIPPIE), molecular interaction search tool (MIST) and search tool for recurring instances of neighbouring genes (STRING). ✓ indicates presence and ✗ indicates absence of feature.

| Feature | PINOT | HIPPIE | MIST | STRING |
|--|-------|--------|------|--------|
| Data extracted at the time of query | ✓ | ✗ | ✗ | ✗ |
| Can accommodate large input lists (100s of seeds) | ✓ | ✓ | ✗ | ✗ |
| Parsable output file | ✓ | ✓ | ✗ | ✓ |
| Direct interactions for seeds only | ✓ | ✓ | ✓ | ✗ |
| Multiple species coverage | ✓ | ✗ | ✓ | ✓ |
| Assigned confidence score | ✓ | ✓ | ✓ | ✓ |
| PubMed ID reported | ✓ | ✓ | ✓ | ✗ |
| Detection method reported | ✓ | ✓ | ✓ | ✗ |
| Quality control publication and method annotations | ✓ | ✗ | ✗ | NA |
| NCBI Entrez ID | ✓ | ✓ | ✓ | ✗ |
| UniProt SwissProt ID | ✓ | ✗ | ✗ | ✗ |
| Programming codes available | ✓ | ✗ | ✗ | ✗ |
| Network visualisation | ✗ | ✓ | ✓ | ✓ |
| Orthologous inference | ✗ | ✗ | ✓ | ✗ |

2.4.7 Conclusions

These ROCO protein interaction network analyses and associated investigations provides novel insight into the interaction and functional profiles of the human ROCO proteins. The extent of commonality and distinction between these profiles was evaluated. Evidence supporting potential shared interactors and functional pathways for multiple ROCO proteins, although it appears that despite the structural conservation between these proteins, the ROCO proteins have evolved largely divergent interactomes. Assessing the ROCO protein interaction profiles from multiple angles enabled new experimental data to be considered in the context of existing confidence-weighted PPI literature-derived data, which in turn prioritised specific proteins for further follow-up validation studies. In addition, the construction of the literature-derived network represented a sample analysis for the ongoing development of the WPPINA pipeline, which facilitated improvements to this procedure and the development of PINOT.

Future direction for better understanding the ROCO protein interaction profile would benefit from further integration of novel experimental data, derived from distinct method detection approaches, such as yeast two-hybrid or stable isotope labelling with amino acids in cell culture (SILAC) screens. Furthermore, as cell type specific protein expression data becomes readily available, incorporating this data would also strengthen our understanding of ROCO PPI events in the cell. Finally, this study highlights the value of data integration strategies for accelerating biological understanding, which is particularly relevant in this ever-increasing data-rich era.

CHAPTER THREE

Assessing the DAPK-1 Protein Interaction Profile in *Caenorhabditis elegans*

3.1 Introduction

Human DAPK1 is the product of an ancient evolutionary lineage [21]. As described in Chapter One, mammalian DAPK1 belongs to both the DAPK and ROCO protein families, due to structural homology of the kinase and ROC-COR domains, respectively. The DAPK family diversified within the mammalian clade, with 3 DAPKs in the human proteome, whereas lower organisms only possess a single DAPK protein [161]. This offers a reductionist approach for studying DAPKs in invertebrates. In contrast, the evolutionary history of the ROCO proteins is much more complex, numerous diversification, duplication and loss events occurred throughout the evolutionary timeline [23,162]. DAPKs, and the closely related DRAKs, in invertebrate model organisms, include Drak, a DRAK1/DRAK2 ortholog in *Drosophila melanogaster* and DAPK-1, a DAPK1 ortholog in *Caenorhabditis elegans*. Interestingly, amongst a diverse collection of invertebrates, phylogenetic analysis of DAPK1 and its orthologs identified that the closest relative to human DAPK1 is within the sea urchin *Stongylocentrotus purpuratus* proteome [161]. Furthermore, DAPK1 contains structural homology in the proximity of the LRRs and ROC-COR regions to a plant protein, Tornado1 in *Arabidopsis thaliana* [163], although this protein lacks kinase and death domains.

The DAPK1 ortholog in *C. elegans*, DAPK-1, shares considerable sequence similarity to its human counterpart (Figure 3.1A), especially with regard to the kinase domain. These proteins have equivalent domain topologies and differ in length by only five amino acids. The basic loop motif within the kinase domain, which is a characteristic feature of DAPKs [24], and the P-loop of the ROC domain responsible for GTP binding are largely homologous. In addition, the two amino acid triads of the kinase domain ATP-binding pocket (Lys 42, Glu 64, Glu 100, Glu 143, Asn 144 and Asp 161 in DAPK1 [24]), which are critical for catalysis, are conserved (Figure 3.1B and 3.1C). However, Ser308 (in DAPK1), an autophosphorylation site within in Ca²⁺/CaM regulatory domain which negatively regulates kinase activity upon phosphorylation, is not conserved within *C. elegans* DAPK-1.

Functional characterisation of *C. elegans* DAPK-1 is in its infancy, nevertheless, a number of roles have been proposed. Genetic manipulation approaches, specifically *dapk-1* mutagenesis, RNA interference (RNAi) and transgenesis, have been utilised to assess phenotypic consequences which relate to DAPK1 function. Of note, *dapk-1* loss-of-function results in viable and fertile progeny [161]. Furthermore, studies have provided indications that DAPK-1 may be involved in stress-induced autophagy, epidermal homeostasis, innate immunity, wound healing and neuronal excitotoxicity [164–167]. Despite extensive evidence of human DAPK1 functioning

within cell death pathways, in particular apoptosis, this functional link requires further investigation in relation to *C. elegans* DAPK-1. However, orthologous functional connections between human DAPK1 and *C. elegans* DAPK-1 are evident, such as the roles of DAPK1/DAPK-1 in autophagy [49,164] and the emerging evidence implicating these proteins in excitotoxic neurodegeneration [167,168].

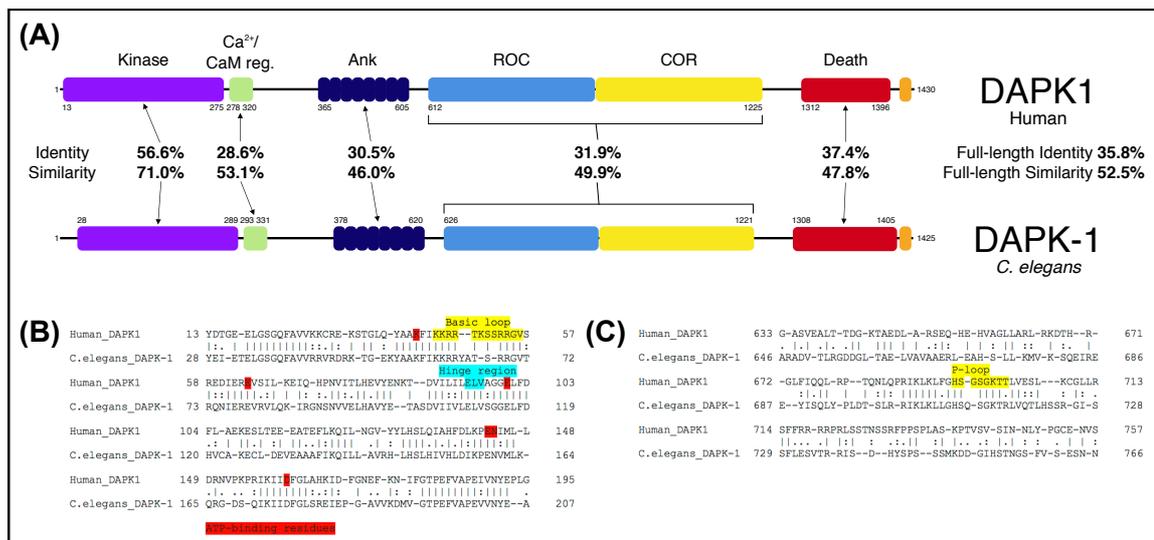


Figure 3.1 – Primary structure comparisons between human DAPK1 and *C. elegans* DAPK-1. (A) Domain topology and sequence similarity of DAPK1 and DAPK-1. Snapshots of annotated protein sequence alignments within the kinase and ROC domains (B and C, respectively). Pairwise sequence alignment analysis performed using the EBI EMBOSS Stretcher tool for full-length and domain protein sequences. Highlighted sequence corresponds to ATP binding residues (red), basic / P loop (yellow) and hinge region (blue).

The role of DAPK-1 at a mechanistic level, in the context of signalling cascades, is poorly understood. The reported PPI profile is very limited, there are no PPI data entries for DAPK-1 in WormBase nor when querying the major molecular interaction repositories via the PSCIQUIC. Insights from functional studies have reported that DAPK-1 physically interacts with PTRN-1 (Patronin), a microtubule-binding protein, to elicit an inhibitory effect on PTRN-1 function [166]. This binding event plays a role in epidermal integrity and wound healing in *C. elegans*. In addition, a DAPK-1 PINN-1 interaction has been hypothesised based on a reported orthologous interaction in human [169] and convergent functional findings in relation to these two proteins in *C. elegans* [167], in the context of excitotoxicity.

Model organisms are powerful tools for exploring biology at a subcellular level to influence inferences to more complex organisms based on genetic tractability, especially in cases where clear orthologs exist between species. However, in the case of *C. elegans* and DAPK1/DAPK-1, human DAPK1 research is more advanced in relation to characterising the interaction and functional landscape of this protein. Nevertheless, there are many benefits to using *C. elegans* as a model system for understanding protein function [100,170]. Most notably, protein function can be assessed at a whole-organism level (*in vivo*) and genetic manipulation is a routinely used procedure to support this *in vivo* investigation into protein function. With this in mind, DAPK1/DAPK-1 remained the centre of investigation, but in the context of *C. elegans*.

3.1.1 Project aims

The aims of this section of research complement and build on those outlined in Chapter Two, from both a method development and an exploratory protein interactome perspective. My first aim was to further develop PINOT to incorporate a *C. elegans* PPI query option. In parallel to this focus, I aimed to gather insight into the protein interaction network of *C. elegans* DAPK-1, which was previously undefined, with the exception of one reported DAPK-1 PPI.

With regard to the further development of PINOT, my aim was to expand the query capacity of our resource to encompass a *C. elegans* option. To address this aim, I planned to modify the coding scripts underlying PINOT, to extract WormBase-curated PPI data for data processing, in relation to method code reassignment and confidence scoring. This modified script would then be implemented onto the server hosting PINOT and upon the user selecting *C. elegans* as a query option, this new script would be utilised. The vision with this additional PINOT development was to broaden the utility of PINOT to the *C. elegans* community, but also to facilitate the assessment of the potential DAPK-1 interactome within the wider cellular context.

In relation to investigating the DAPK-1 interactome, my aim was to utilise various distinct approaches to predict and identify potential binding partners. First, using predominantly human DAPK1 literature-derived PPIs as a foundation for predicting *C. elegans* DAPK-1 interactors based on identifying orthologs of these reported interactors. Second, I aimed to assess the DAPK-1 interaction network experimentally, using the yeast two-hybrid (Y2H) system and by affinity purification - mass spectrometry (AP-MS), to identify novel DAPK-1 interactors. I planned to outsource the Y2H screen to a specialist company. Taken in combination, I aimed to use these approaches as a strategy to test the validity of the DAPK-1 interactor predictions and identify novel interactors.

3.2 Materials and Methods

3.2.1 General laboratory materials and media/buffer compositions

| <u>Reagent</u> | <u>Supplier</u> |
|--|-------------------|
| 1-bromo-3-chloro-propane | Sigma-Aldrich |
| Agar | Fisher Scientific |
| Agarose | Fisher Scientific |
| Ampicillin | Sigma-Aldrich |
| Bromophenol blue | Sigma-Aldrich |
| Calcium chloride | Fisher Scientific |
| Cholesterol | Fisher Scientific |
| Dibasic potassium phosphate | Fisher Scientific |
| Dibasic sodium phosphate | Fisher Scientific |
| Ethanol | Fisher Scientific |
| Ethylenediaminetetraacetic acid (EDTA) | Sigma-Aldrich |
| Fluorodeoxyuridine (FUDR) | Fisher Scientific |
| Gelatin | Fisher Scientific |
| Gibco PBS Tablets | ThermoFisher |
| Glycerol | Fisher Scientific |
| HALT™ protease inhibitor single-use cocktail | ThermoFisher |
| IGEPAL CA-630 | Fisher Scientific |
| Immobilon-P PVDF membrane | Sigma-Aldrich |
| Invitrogen TRI Reagent solution | ThermoFisher |
| Magnesium chloride | Sigma-Aldrich |
| Magnesium sulfate | Fisher Scientific |
| Monobasic potassium phosphate | Fisher Scientific |
| NuPAGE reagents | ThermoFisher |
| Nystatin | Fisher Scientific |
| Peptone | Fisher Scientific |
| Phenylmethylsulfonyl fluoride (PMSF) | Sigma-Aldrich |
| Potassium chloride | Fisher Scientific |
| Propanol | Fisher Scientific |
| Proteinase K | Fisher Scientific |
| Skimmed Milk Powder | Asda |
| Sodium azide | Fisher Scientific |

| | |
|------------------------------|-------------------|
| Sodium chloride | Fisher Scientific |
| Sodium dodecyl sulfate (SDS) | Sigma-Aldrich |
| Sodium hydroxide | Fisher Scientific |
| Sodium hypochlorite | Fisher Scientific |
| 50x TAE buffer | Sigma-Aldrich |
| Tris | Fisher Scientific |
| Triton X-100 | Fisher Scientific |
| Tryptone | Sigma-Aldrich |
| Tween-20 | Sigma-Aldrich |
| Yeast extract | Sigma-Aldrich |
| β -mercaptoethanol | Sigma-Aldrich |

Nematode growth media (NGM)

1.7% (w/v) agar, 0.25% (w/v) peptone, 50mM NaCl, 5 μ g/ml cholesterol [dissolved in ethanol], 25mM KPO₄, 1mM CaCl₂, 1mM MgSO₄, 25 units/ml nystatin, in ddH₂O.

Enriched peptone (EP) media

2.5% (w/v) agar, 2% (w/v) peptone, 20mM NaCl, 5 μ g/ml cholesterol [dissolved in ethanol], 25mM KPO₄, 1mM MgSO₄, 25 units/ml nystatin, in ddH₂O.

M9 buffer

86mM NaCl, 42mM Na₂HPO₄, 22mM KH₂PO₄, 1mM MgSO₄, in ddH₂O.

PBS-T

PBS supplemented with 0.05% Tween-20

Freezing buffer

100mM NaCl, 50mM KH₂PO₄, 5.6mM NaOH, 30% (v/v) glycerol, 300 μ M MgSO₄, in ddH₂O.

Luria Broth (LB) / Luria Agar (LA)

171mM NaCl, 1% (w/v) tryptone, 0.5% (w/v) yeast extract, in ddH₂O. LA was also supplemented with 1.8% (w/v) agar.

Bleaching solution

2% (v/v) NaOCl, 250mM NaOH, in ddH₂O.

Worm lysis buffer 1 (WLB1)

50mM KCl, 10mM Tris (pH 8.3), 2.5mM MgCl₂, 0.45% (v/v) IGEPAL CA-630, 0.45% (v/v) Tween-20, 0.01% (v/v) gelatin, in ddH₂O.

Worm lysis buffer 2 (WLB2)

200mM NaCl, 100mM Tris HCl (pH 8), 1mM ethylenediaminetetraacetic acid (EDTA), 8% (v/v) glycerol, 2% Triton X-100, 1mM phenylmethylsulfonyl fluoride (PMSF [dissolved in propanol]), 2x HALT™ protease inhibitor single-use cocktail, in ddH₂O.

RIPA buffer

Ready-to-use premixed RIPA buffer (Sigma-Aldrich [150 mM NaCl, 50 mM Tris (pH 8.0), 1% IGEPAL CA-630, 0.5% sodium deoxycholate, 0.1% sodium dodecyl sulfate (SDS)]) supplemented with 2x HALT™ protease inhibitor single-use cocktail.

Denaturation buffer

125mM Tris (pH 6.8), 20% (v/v) glycerol, 4% (w/v) sodium dodecyl sulfate (SDS), 5% (v/v) β-mercaptoethanol, 0.03% (w/v) bromophenol blue, in ddH₂O.

3.2.2 Developing a *C. elegans* query option for PINOT

Since *C. elegans* is a widely used model organism for basic biology and translational biomedical research, ongoing efforts to dissect molecular signalling events have resulted in the detection of many PPIs, most of which have been detected in a high-throughput manner. Similarly to human PPI data, *C. elegans* PPI data is actively curated into databases and hence these repositories are a valuable resources for constructing PPI networks. The development of a *C. elegans* query option within PINOT was to utilise this PPI data in relation to a query input and to facilitate the interpretation of this data collectively. A number of modifications were made to the coding scripts which underlie PINOT in order to develop *C. elegans* PPI query capacity, these are described below and the pipeline for *C. elegans* PPI data querying is outlined in Figure 3.2.

3.2.2.1 Data acquisition and pre-processing

Although *C. elegans* PPI data is curated into databases accessed through the PSICQUIC, such as BioGRID and IntAct, which are utilised for human PPI querying in PINOT, there is an independent curation effort for *C. elegans* PPIs carried out by WormBase which encompasses a wider coverage of the PPI data generated for *C. elegans* [171]. This WormBase-curated data is available

via the Alliance of Genome Resources Portal (available at <https://alliancegenome.org>) in a MI TAB format, which is the data format that the pipeline underlying PINOT was designed to process in relation to human PPIs. Hence, much of the already developed pipeline would be compatible with this dataset. Therefore, this dataset (specifically the alliance molecular interactions dataset version 2.1), which contained approximately 760 000 molecular interaction data entries across multiple species, was downloaded for implementation into PINOT.

In contrast to how the human PPI data is acquired in PINOT, the entire WormBase-curated *C. elegans* PPI dataset was downloaded for storage on the server hosting PINOT and therefore upon querying PINOT for *C. elegans* PPI data, the data is extracted from this local data pool. A further difference from the human PPI query pipeline was that the downloaded dataset was pre-processed prior to implementation into PINOT. This pre-processing involved the dataset formatting, quality control and method code reassignment steps of the pipeline underlying PINOT prior to the storage of the dataset on the server. The key reason for this was to reduce the computational load on PINOT for querying *C. elegans* PPI data. Since this complete dataset was static in nature (i.e. was downloaded, stored and not subject to change), undertaking the bulk of the data processing once as oppose to every time a query is submitted was a much more efficient approach, computationally.

More specifically, these pre-processing steps first involved retaining only *C. elegans* data entries from the dataset, reducing the size of the dataset to approximately 37 000 molecular interactions. Then, the dataset was manually reformatted in terms of discarding undesired data columns. The only data columns retained correspond to the WormBase gene ID for each interactor, an alias ID for each interactor (either the common name or locus ID), the interaction detection method annotation and the associated PubMed ID, since this is the relevant information in relation to the data processing and output data generation in the pipeline. Further reformatting reduced information within each data entry by discarding any non-essential characters for data processing, for example non-specific ID prefixes: 'wormbase:WBGene00004945' reduced to 'WBGene00004945'. Next, in line with the quality control step of the pipeline, non-PPI entries, such as protein-DNA interaction entries, were manually removed. This resulted in a dataset of approximately 30 000 PPI data entries. The quality of the data curation was notably high, in comparison to PPI data in BioGRID for example, incomplete data entries were absent in this WormBase-curated data.

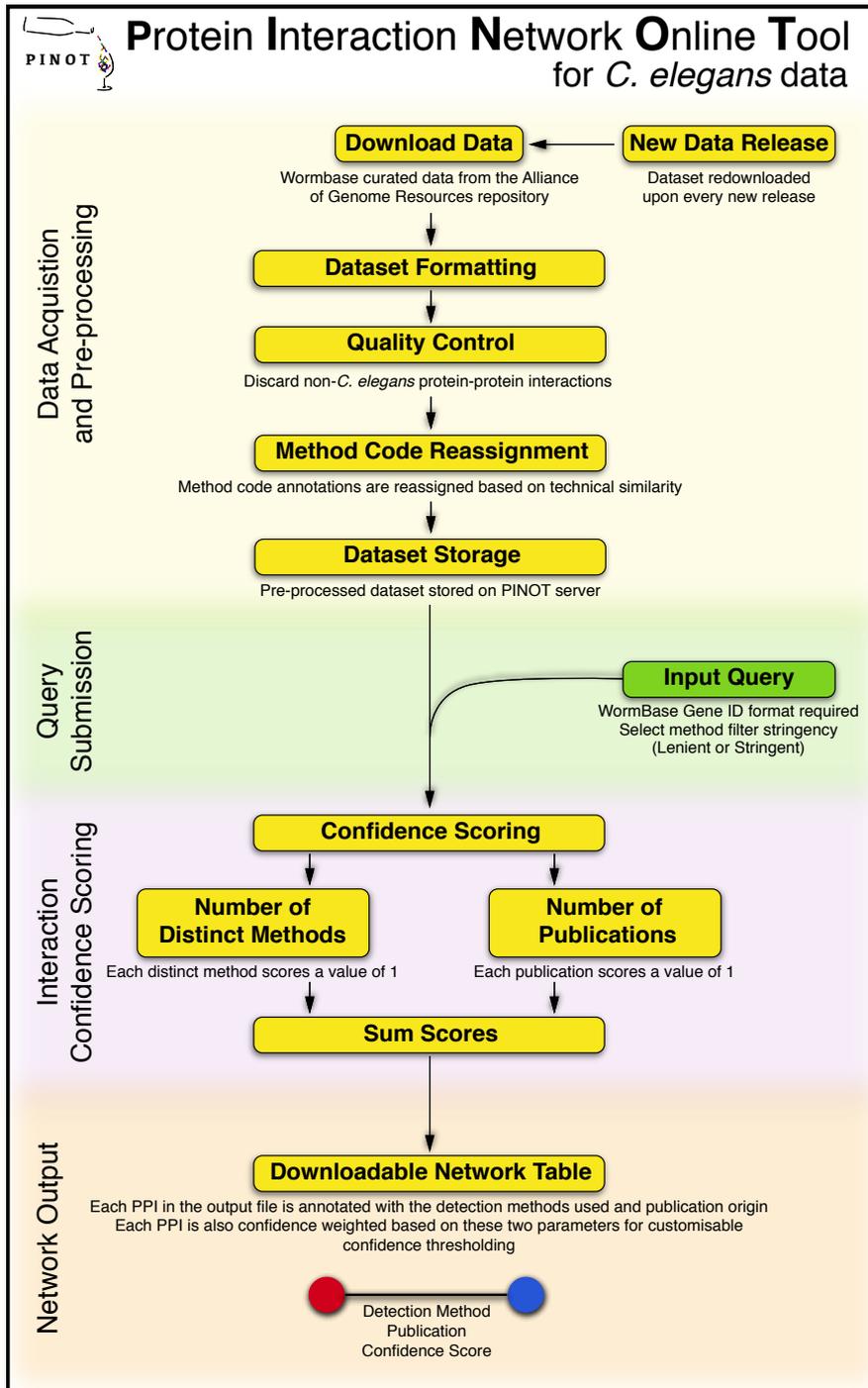


Figure 3.2 – Data processing pipeline underlying PINOT for *C. elegans* data. A stepwise insight into the pipeline. A notable difference to the human PPI data pipeline is that the entire *C. elegans* dataset is downloaded, pre-processed and stored, the query is submitted to extract data from this stored dataset, the data is then scored and the fully processed dataset (which corresponds to the query input) is provided. The downloaded data is updated upon every new WormBase molecular interaction data release.

This dataset was then subjected to the previously developed method code reassignment coding script which reassigns each PPI data entry with a new detection method annotation based on the technical similarity of the technique. Similarly to the human PPI data pipeline, a stringent or lenient filter option was implemented into PINOT for discarding or retaining *C. elegans* data entries which were reassigned the UNSP (unspecified) detection method annotation, respectively. In the case of the stringently filtered dataset, these UNSP-associated data entries were manually removed. These pre-processed datasets, corresponding to the stringent and lenient filtered versions of the *C. elegans* PPI data, were then stored on the server hosting PINOT for further data processing upon query submission.

3.2.2.2 Query submission

The PINOT user interface was modified to accommodate querying *C. elegans* PPI data and this included the addition of a drop-down menu to select this species. The interface development was carried out by Dr Liam McGuffin at the University of Reading. The required input for querying *C. elegans* PPI data and running the pipeline is a list of WormBase gene IDs (example format: WBGene00000063), either entered into the text box within the PINOT interface or uploaded in a .txt file format. This ID type was decided since it is a universal ID for all *C. elegans* genes and is curated alongside each WormBase PPI data entry. Although the gene name ID (example format: act-1) is commonly used to identify *C. elegans* genes and proteins, not all genes have been assigned this ID type.

Upon submitting a query, each data entry row which matches the queried ID is extracted from the downloaded and pre-processed *C. elegans* PPI dataset. A file is then created for each queried ID containing the rows of data entries which match the specific ID of interest. The queried protein (seed protein) is positioned into the first two columns of these files. These files are then temporarily stored on the server.

3.2.2.3 Interaction confidence scoring

The concept underlying the confidence scoring step of the pipeline for *C. elegans* PPI data is identical to the process developed for human PPI data (see '2.2.1.4 Interaction confidence scoring'). Each file which was created in the previous step is assessed for repeated equivalent interaction entries, i.e. the same interaction detected and reported in by a single method/publication curated multiple times. These types of replicate entries are discarded, so only one is retained. Each data entry is then confidence scored based on the number of distinct

methods used to detect an interaction (using the reassigned method codes) and the number of publications which report the interactions (based on PubMed ID). Each data entry is scored a value of 1 for each distinct method used and a value of 1 for each unique publication associated to the interaction. These two scores are added together to calculate the final score which is provided within the output file. Following this confidence scoring of each file, these files are merged to create a single final network file.

3.2.2.4 Network output

Similarly to the output file provided when querying PINOT for human PPI data, the *C. elegans* output file, which is either emailed to the user or can be downloaded from the browser, is straightforward to interpret (Figure 3.3). Each row corresponds to a binary interaction and contains the WormBase gene IDs for each interactor, the common gene name ID or locus ID (for genes lacking a common gene name) for each interactor, the interaction detection method score and annotation, the publication score and PubMed ID, and the final confidence score. This file is compatible for importing into network visualisation software, such as Cytoscape [113].

| WormBaseID.A | Alias.A | WormBaseID.B | Alias.B | Method.Score | Method | Publication.Score | PMIDs | Final.Score |
|----------------|--------------------|----------------|----------------------|--------------|-----------------------------|-------------------|---------------------------------|-------------|
| WBGene00001578 | ges-1(public_name) | WBGene00004798 | sip-1(public_name) | 1 | Chromatography | 1 | pubmed:26009280 | 2 |
| WBGene00004927 | snx-1(public_name) | WBGene00000415 | ced-1(public_name) | 1 | Chromatography | 1 | pubmed:20133524 | 2 |
| WBGene00004927 | snx-1(public_name) | WBGene00000843 | cup-2(public_name) | 3 | Imaging;Chromatography;2Hyb | 2 | pubmed:21722281;pubmed:19123269 | 5 |
| WBGene00004927 | snx-1(public_name) | WBGene00003949 | pbs-3(public_name) | 1 | 2Hyb | 1 | pubmed:12679813 | 2 |
| WBGene00004927 | snx-1(public_name) | WBGene00004336 | ret-1(public_name) | 1 | 2Hyb | 1 | pubmed:14704431 | 2 |
| WBGene00004927 | snx-1(public_name) | WBGene00004378 | rme-8(public_name) | 2 | Chromatography;2Hyb | 1 | pubmed:19763082 | 3 |
| WBGene00004927 | snx-1(public_name) | WBGene00015518 | C06E1.1(public_name) | 1 | 2Hyb | 1 | pubmed:19123269 | 2 |
| WBGene00004927 | snx-1(public_name) | WBGene00017542 | F17E9.5(public_name) | 1 | 2Hyb | 2 | pubmed:14704431;pubmed:19123269 | 3 |

Figure 3.3 – Screenshot of an example PINOT *C. elegans* data output.

3.2.3 Predicting DAPK-1 interactors

Insight into the DAPK-1 protein interaction profile was limited to one reported protein interactor, PTRN-1 [166], and hence predicted interactomes were constructed based on orthologous inference from other species. First, the human DAPK1 PPI network was generated by querying PINOT [121] (available at http://www.reading.ac.uk/bioinf/PINOT/PINOT_form.html) with DAPK1 on 25th June 2019 (*Homo sapiens* and stringent filter query options selected). The output data was confidence thresholded so that only interactors that had been replicated were retained for mapping the network. This corresponds to interactions that were annotated with a final score greater than 2 in the PINOT output file.

Two approaches were utilised for predicting the DAPK-1 interactome by orthologous inference: using the interolog feature in MIST [159] (the Molecular Interaction Search Tool, available at <https://fgertools.hms.harvard.edu/MIST/>), a feature which enables PPI predictions based on orthologs of reported PPIs in other species; and by pinpointing the orthologs of human DAPK1 interactors using Ortholist 2.0 [172] (available at <http://ortholist.shaye-lab.org>), a meta-analysis compiled resource for human – *C. elegans* ortholog conversions. MIST was queried with DAPK-1 on 27th September 2019, with the interolog feature and the filtering option of filtering out low rank interactions selected. The *C. elegans* orthologs of the reported (and replicated) human DAPK1 interactors derived from PINOT were identified using Ortholist 2.0 on 25th June 2019. The HGNC approved symbols were used as the input ID type, no minimum number of programs and no partial matches were selected as criteria for the analysis. These orthologous PPIs were mapped onto two separate predicted DAPK-1 interaction networks.

The predicted DAPK-1 interactors from the latter predicted network (derived from Ortholist-converted orthologs) then formed the basis of a further PINOT query to identify reported *C. elegans* protein binding partners for these predicted DAPK-1 interactors. This PINOT query was performed on 28th August 2019 and the corresponding WormBase gene IDs (obtained from the Ortholist 2.0 output) were used as the query input. *C. elegans* was the selected query species. As before, the output file was confidence thresholded to retain only interactions that had been replicated (final score greater than 2), for mapping onto the network. Networks were visualised in Cytoscape version 3.7.0.

3.2.4 DAPK-1 yeast two-hybrid screen

In order to identify novel putative DAPK-1 interactors, a yeast two-hybrid (Y2H) screen was performed. This experiment was carried out by Hybrigenics Services (France), a specialist Y2H service provider. Full-length *C. elegans dapk-1* (NCBI sequence reference: NM_058439.4) was cloned into their ULTimate Y2H™ pB27 vector with a *LexA* DNA-binding domain (DBD), which binds the promoter region of the *HIS3* reporter gene, fused to the 5' end of the gene. *dapk-1* was screened against a mixed life stage *C. elegans* prey-*Gal4* activation domain construct library (Hybrigenics reference: CEMS). Positive selection of PPI events was determined by mating yeast, bait transformants (α) and prey transformants (α), on media lacking histidine. The DNA of positive clones was sequenced to identify the prey gene underlying the positive hit.

3.2.5 Generating a FLAG-*dapk-1* *C. elegans* strain

With the goal of conducting further experimental interactomic analysis of DAPK-1 adopting distinct interaction detection approaches, such as AP-MS, an epitope tag was engineered onto DAPK-1 using clustered regularly interspaced short palindromic repeats (CRISPR) – Cas9 knock-in technology. This genomic modification was carried out by SunyBiotech (China), a company which specialises in *C. elegans* precision gene editing. The service provided by SunyBiotech covered single guide RNA (sgRNA) design, plasmid construction, plasmid microinjection, screening for and isolating homozygotes harbouring the modification, and sequence validation.

3.2.5.1 FLAG-*dapk-1* sequence design

A 3xFLAG tag and linker sequence was inserted at the endogenous *dapk-1* locus, at the 5' end of the gene (Figure 3.4), on the N2 (wild-type) genetic background. The precise 3xFLAG sequence was suggested by SunyBiotech based on previous successful FLAG tag knock-in projects. The linker sequence was designed to create a short flexible bridge of amino acids between the epitope tag and DAPK-1 in order to connect the tag to the protein of interest yet limit potential structural interference that may impact protein functionality. As is commonly used [173], a short chain of two glycine residues flanking both sides of a serine residue was used as the peptide linker.

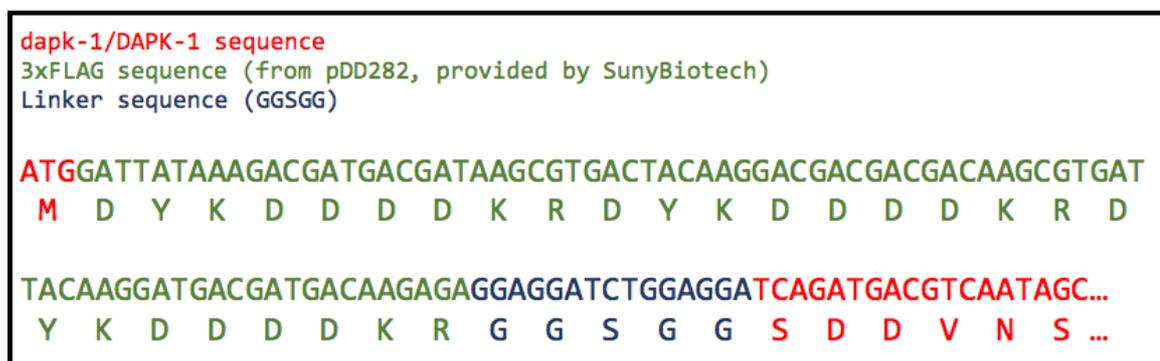


Figure 3.4 – Inserted 3xFLAG and linker sequence at the *dapk-1* locus. DNA sequence with corresponding translated amino acid sequence below. 3xFLAG sequence (green) and flexible linker sequence (blue) inserted at the 5' end of *dapk-1* (red).

3.2.6 FLAG-*dapk-1* sequence validation

In addition to the knock-in verification performed by SunyBiotech, validation of the FLAG-*dapk-1* genotype was conducted in-house. Nematodes were lysed and genomic DNA was crudely

extracted for PCR. Restriction digestion was performed for reactions with primer combinations forward 1 (F₁) and reverse 1 (R₁; see Table 3.1), based on a restriction site in the inserted region, in order to validate the insertion of the tag and linker. The experimental design is illustrated in Figure 3.5. Refer to '3.2.1 General laboratory materials and media/buffer compositions' for media and buffer compositions used in this section and hereafter.

3.2.6.1 *C. elegans* maintenance

C. elegans were routinely maintained on nematode growth media (NGM) plates seeded with *Escherichia coli* OP50 (referred hereafter as OP50) at 20°C. A new OP50 stock culture was streaked on Luria agar (LA) monthly. From this plate, Luria broth (LB) was inoculated with OP50 and incubated at 37°C for 16-24hrs, cultures were stored at 4°C. This liquid culture was inoculated and spread on NGM plates (100µl, 300µl and 1ml on 30mm, 60mm and 100mm plates, respectively) and incubated at 37°C for 16-24hrs, OP50-seeded NGM plates were stored at 4°C. These plates were then used for maintaining *C. elegans*. Nematodes were replated 2 or 3 times per week to prevent starvation. Replating *C. elegans* either involved cutting a cube of NGM from a populated plate and placing it upside down on a fresh NGM plate or by picking individual nematodes with a platinum wire onto fresh plates.

3.2.6.2 Freezing and recovering *C. elegans*

C. elegans stocks were stored at -80°C. For freezing, predominantly starved L1-L2 nematodes were washed off plates with M9 buffer and collated into a 15ml or 50 ml tube. An equal amount of freezing buffer was then added to the M9/nematode suspension, the suspension was mixed and aliquoted into cryopreservation tubes (1ml per tube). These tubes were then placed in polystyrene racks at -80°C. For nematode recovery from -80°C, samples were thawed at room temperature and the nematode suspension was pipetted onto NGM plates, 200-300µl per 60mm plate, and incubated under standard maintenance conditions (20°C).

3.2.6.3 Obtaining synchronised life-stage *C. elegans* populations

C. elegans populations were bleached in preparation for experimentation, this enabled the isolation of embryos for establishing synchronised life-stage populations and also sterilised samples from bacterial/fungal contamination. Bleaching was performed by collecting gravid *C. elegans* off plates using M9 buffer, nematodes were then left to settle to the bottom of a tube, the supernatant was removed and bleaching solution was added (typically 1ml in a 1.5ml tube).

Samples were then shaken vigorously and monitored under the stereo microscope to detect when the nematode bodies had started to disintegrate. Once this had occurred, the samples were centrifuged at 2500 x g for 2 minutes, the supernatant was discarded, 1ml M9 buffer was added, the sample was mixed and then centrifuged as before. This M9 buffer wash step was repeated four further times to wash out the bleaching solution from the sample leaving an M9-embryo solution. This solution was then incubated at 20°C for 16hrs to allow the embryos to hatch into L1 larvae, these larvae were then plated onto NGM.

For ageing synchronised populations of *C. elegans*, to adult day 5 or day 10 for example, late-stage L4s/young adults were transferred onto NGM supplemented with 50µM fluorodeoxyuridine (FUDR) which inhibits DNA synthesis and therefore suppresses the development of embryos [174]. This enabled synchronised ageing of *C. elegans* populations without the technical challenge of managing progeny contamination. NGM plates supplemented with FUDR were seeded with 10x concentrated OP50 broth.

3.2.6.4 DNA extraction, PCR and restriction digestion

Ten adult (day 1) *C. elegans* were picked into 20µl worm lysis buffer 1 (WLB1) supplemented with 1mg/ml proteinase K, per strain (FLAG-*dapk-1*; N2). Four replicates were processed for each strain. Samples were lysed and DNA crudely extracted by incubation at 65°C for 1 hour followed by 95°C for 15 minutes.

Two different PCRs were performed on these DNA extracts based on two different primer combinations (F₁ + R₁ and F₂ + R₁ [see Figure 3.5]). Primer sequences are listed in Table 3.1. PCRs were performed in a total volume of 15µl: 7.5µl 2x PCR BIO HS Taq Mix (PCR Biosystems), 0.6µl 10µM primer mix (forward and reverse), 3µl DNA extract, 3.9µl nuclease-free water. The thermocycler parameters for PCR are outlined in Table 3.2. Samples were incubated in a BioRad T100 Thermal Cycler.

Table 3.1 – Primer sequences for FLAG-*dapk-1* genotype validation by PCR. Primers supplied by Eurofins.

| Primer | Primer sequence (5' to 3') |
|---|----------------------------|
| FLAG- <i>dapk-1</i> forward 1 (F ₁) | TAAAGACGTGCTGAGTGACG |
| FLAG- <i>dapk-1</i> forward 2 (F ₂) | TAAGCGTGACTACAAGGACG |
| FLAG- <i>dapk-1</i> reverse 1 (R ₁) | TTTCCTGAGAGAAGAACGGG |

Table 3.2 – PCR thermocycler programme for FLAG-dapk-1 genotype validation

| Temperature | Time |
|-------------|---------------|
| 95°C | 1 min 30 secs |
| 95°C | 15 secs* |
| 55°C | 15 secs* |
| 72°C | 15 secs* |
| 72°C | 2 mins |

*these three steps were cycled 40 times

PCR products derived from the reactions with primer combinations $F_1 + R_1$ were subjected to restriction digestion with *BseRI* (New England Biolabs). Digest reactions were performed in a total volume of 20µl: 0.3µl *BseRI*, 2µl 10x CutSmart reaction buffer, 7µl PCR amplified product, 10.7µl nuclease-free water. For digestion, samples were incubated at 37°C for 2 hours. DNA fragment size was visualised by 2% (w/v) agarose gel electrophoresis (80V for 1 hour) in TAE buffer, using 0.5x SYBR safe stain for DNA detection and a 100bp DNA ladder. Gels were imaged on a U:Genius 3 (Syngene).

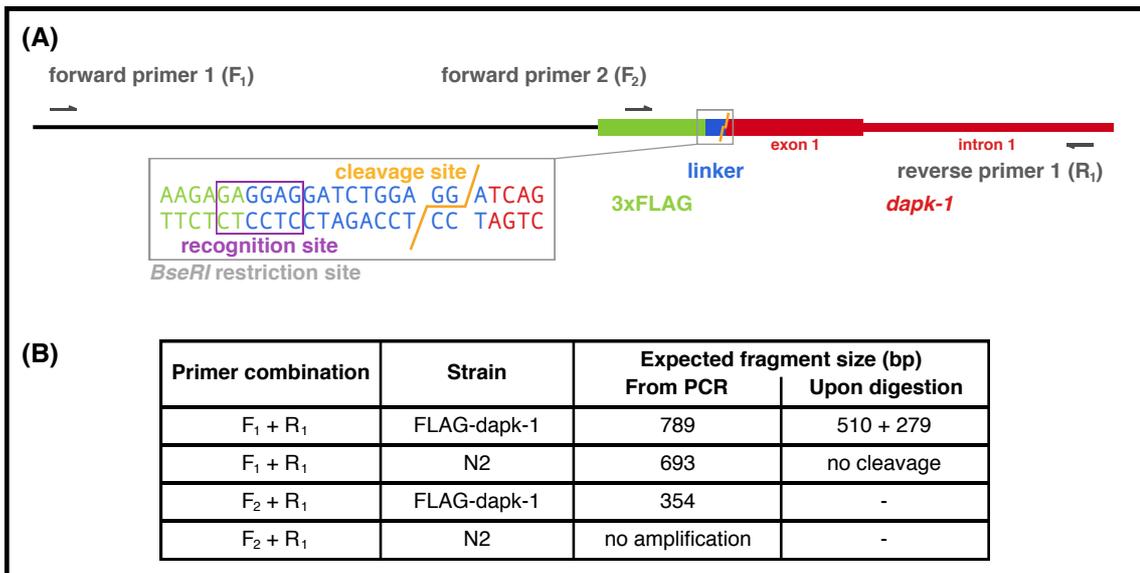


Figure 3.5 – Experimental design and predicted result for FLAG-dapk-1 genotype validation.

(A) Site of 3xFLAG tag and linker knock-in at the 5' end of the *dapk-1* gene, with annotations of where the primers were designed to bind the template. The *BseRI* restriction site is annotated onto the template and a detailed sequence view of the enzyme recognition and cleavage sites are highlighted. (B) Expected DNA fragment sizes following PCR and subsequent restriction digestion with *BseRI* upon amplification with two different primer combinations. Restriction digestion was not performed on amplicons derived from primer combination $F_2 + R_1$.

3.2.7 FLAG-DAPK1 protein expression analysis

3.2.7.1 Protein extraction

C. elegans protein extracts were prepared in various lysis buffers in an attempt to optimise protein detection *via* Western blot. First, synchronised life stage nematodes at various life stages were collected off either NGM or enriched peptone (EP) media plates with M9 buffer. Once collated into a tube, the nematodes were washed at least two times in M9 buffer, by allowing the worms to settle at the bottom of the tube, removing the supernatant and adding fresh M9 buffer. After these washes and removing the supernatant, one of three buffers was added to the sample for nematode lysis and protein extraction, either worm lysis buffer 2 (WLB2), RIPA buffer or denaturation buffer. The volume of this buffer added was 1x the estimated volume of nematodes. Samples were then sonicated (using a probe sonicator) for 5 x 30 seconds, with 10 second intervals, at 55% amplitude. After sonication, the samples were centrifuged at 18 800 x g for 15 minutes at 4°C, the supernatant was removed and stored at -20°C, or immediately processed, for Western blot analysis.

In the cases where the lysis and protein extraction were performed in WLB2 and RIPA buffer, the protein concentrations were estimated using a Pierce™ bicinchoninic acid (BCA) assay kit (ThermoFisher Scientific), as per manufacturer's guidelines. These samples were then aliquoted, diluted in the corresponding lysis buffer to achieve equal protein concentrations across samples and supplemented with 5% (v/v) β-mercaptoethanol and 1x NuPAGE LDS sample buffer. In all cases of sample preparation, the samples were then heated at 95°C for 10 minutes prior to loading for Western blot analysis.

3.2.7.2 Western blot

Protein samples were loaded into NuPAGE™ 4-12% Bis-Tris precast gels assembled in NuPAGE™ MOPS SDS running buffer and subjected to 140V for 90 minutes. PageRuler™ Plus prestained protein ladder was loaded alongside samples. Following protein separation by gel electrophoresis, samples were transferred to polyvinylidene difluoride (PVDF) membrane by wet electro-transfer in NuPAGE™ transfer buffer supplemented with 20% methanol, at 50V for 4 hours. Protein transfer efficiency was assessed by Ponceau S staining. Membranes were blocked with 5% milk in PBS-T (referred to as blocking solution) for 1 hour at room temperature.

For immunoblotting, membranes were cut and probed with primary antibody diluted in blocking solution for 16 hours at 4°C, washed in PBS-T (5 x 5 minute 10ml washes), probed with secondary antibody diluted in blocking solution for 1 hour at room temperature and washed again in PBS-T (5 x 5 minute 10ml washes). Refer to Table 3.3 for specific antibody details. Membranes were then incubated with an enhanced chemiluminescence (ECL) substrate, SuperSignal™ West Pico PLUS chemiluminescent substrate (ThermoFisher Scientific), for 5 minutes at room temperature, then imaged on an ImageQuant LAS 4000 mini (GE Healthcare Life Sciences).

Table 3.3 – List of antibodies used for Western blot

| Antibody | Dilution | Host Species | Supplier | Catalogue Number |
|--|-------------------------------------|--------------|---------------|------------------|
| Anti-β-actin monoclonal | 1:5000 (approx. 0.4µg/ml) | Mouse | Sigma-Aldrich | A1978 |
| Anti-FLAG M2 monoclonal | 1:1000 (approx. 4µg/ml) | Mouse | Sigma-Aldrich | F3165 |
| Anti-mouse IgG polyclonal peroxidase conjugate | 1:2000 (FLAG) or 1:10000 (actin) | Goat | Sigma-Aldrich | A3682 |

3.2.8 FLAG-dapk-1 transcript expression analysis

Quantitative PCR (qPCR) was performed to assess the presence of the FLAG-*dapk-1* transcript within this newly generated strain. In parallel, this technique was also utilised to provide hints at any potential alterations of *dapk-1* transcript expression upon FLAG tag insertion. Two *C. elegans* strains were processed, in triplicate, for this experiment: FLAG-*dapk-1* and N2.

3.2.8.1 RNA extraction

For *C. elegans* lysis, synchronised L4 stage *C. elegans* were washed off NGM with 2ml M9 buffer. Three densely-populated but non-starved 60mm plates were used per sample. Nematodes were washed twice in 1ml M9 buffer to remove bacteria from sample suspension. 1ml TRI Reagent™ solution was added to each sample, then samples were frozen to -80°C. Following thawing, samples were subjected to bead beating with 0.5mm glass beads at 6m/sec for 2 x 20 seconds (with a 1 minute pause interval) in a tissue homogeniser. Samples were left to stand at room temperature for 5 minutes to promote nucleoprotein complex dissociation. The supernatant was then transferred to a new tube.

The RNA isolation process first involved adding 100µl 1-bromo-3-chloro-propane to the lysate samples, which were subsequently shaken vigorously for 15 seconds then left to stand for 3

minutes at room temperature. Samples were then centrifuged at 12 000 x g for 15 minutes at 4°C. After centrifugation the upper colourless phase of the samples were transferred to a new tube, 1 volume of 70% ethanol was added and the samples mixed by vortexing. To immobilise the RNA, the samples were added onto Qiagen RNeasy mini spin columns and centrifuged at 8000 x g for 15 seconds.

The RNA samples were purified through a number of steps using the Qiagen RNeasy mini kit in combination with the Qiagen RNase-free DNase set. 350µl RW1 buffer was added onto the columns, which were then centrifuged at 8000 x g for 15 seconds. 80µl DNase I mix (10µl DNase I [1500 Kunitz units [175]/550µl] + 70µl RDD buffer) was added to the columns, which were then left to stand at room temperature for 15 minutes. Columns were then washed with 350µl RW1 buffer, centrifuged at 8000 x g for 15 seconds, 500µl RPE buffer (diluted in ethanol at a ratio of 1:4) was added and again, columns were centrifuged at 8000 x g for 15 seconds. Again, 500µl RPE buffer (diluted as before) was added to the column, which was then centrifuged at 8000 x g for 2 minutes. Columns were then centrifuged at 8000 x g for 1 minute to remove any residual liquid. New collection tubes were then attached to the columns and RNA was eluted with 30-50µl nuclease-free water by centrifugation at 8000 x g for 1 minute. RNA concentrations were determined using the Nanodrop 2000 spectrophotometer (Thermo Scientific).

3.2.8.2 cDNA synthesis

cDNA synthesis was performed using reagents from the ProtoScript II First Strand cDNA Synthesis Kit (New England Biolabs). First, a RNA-primer mix containing 1µg RNA, 2µl 60µM random primer mix in a total volume of 8µl (made up with nuclease-free water) was incubated at 65°C for 5 minutes. These RNA-primer mixes were then supplemented with 10µl 2x ProtoScript II Reaction Mix and 2µl 10x ProtoScript II Enzyme Mix, followed by incubation at 25°C for 5 minutes, 42°C for 1 hour and 80°C for 5 minutes. Samples were then stored at -20°C.

3.2.8.3 qPCR

qPCR was performed with primers targeting a region of the FLAG sequence, the *dapk-1* exon 1-2 junction, the *dapk-1* exon 12-13 junction and the *tba-1* exon 1-2 junction (α -tubulin; to act as a housekeeping gene). cDNA used for qPCR was diluted 1:4 in nuclease-free water and 2µl of this diluted template was loaded into sample wells. 8µl of PCR master mix was then added to each well, the final concentrations of these master mix components were 1x FastStart Universal SYBR Green Master (Rox) (Sigma-Aldrich) and 0.3µM primer mix (forward and reverse) in

nuclease-free water (see Table 3.4 for primer sequences). Three experimental replicates, each in duplicate, were setup in one plate. A no template control (NTC) reaction was setup for each master mix (i.e. primer combination).

The samples were incubated at 95°C for 10 minutes then cycled between 95°C for 15 seconds and 62°C for 1 minute for 40 cycles in a StepOnePlus™ Real-Time PCR System thermocycler (Applied Biosystems). The melt curve was determined using a 0.5°C incremental increase in temperature from 62°C to 95°C. Quantitative data was visualised and analysed to determine cycle threshold (C_t) values using StepOne software version 2.3. Data was normalised against the housekeeping gene, *tba-1*, to determine relative expression changes. Representative qPCR products were loaded into a 2.5% (w/v) agarose gel for electrophoresis (80V for 1 hour) to assess for non-specific amplification.

Table 3.4 – Primer sequences for FLAG-dapk-1 transcript validation by qPCR. Primers supplied by Sigma-Aldrich

| Primer | Primer sequence (5' to 3') |
|-----------------------------------|----------------------------|
| FLAG forward | TGACGATGACAAGAGAGGAGG |
| FLAG reverse | TTCGAACGGCGTATCGTCAA |
| <i>dapk-1</i> exons 1-2 forward | AACTTGGGAAGCGGACAATTCG |
| <i>dapk-1</i> exons 1-2 reverse | TTGCGTATCGCCGCTTTTTG |
| <i>dapk-1</i> exons 12-13 forward | TAGACGTGGGTGTTGCTGAT |
| <i>dapk-1</i> exons 12-13 reverse | CGGTTGAATCGACATCTGGC |
| <i>tba-1</i> exons 1-2 forward | AGACCAACAAGCCGATGGAG |
| <i>tba-1</i> exons 1-2 reverse | TCCAGTGCGGATCTCATCAAC |

3.3 Results

3.3.1 Implementation of *C. elegans* query option in PINOT

A *C. elegans* query option was added to PINOT to enable *C. elegans* PPI data to be extracted from the published records in relation to an input query list. As described, the pipeline underlying this aspect of PINOT differs slightly in comparison to using the resource for querying human PPI data. The key difference is that the *C. elegans* PPI data is downloaded, pre-processed and banked within the server hosting PINOT, whereas for human data is extracted and processed at the time of query. The *C. elegans* dataset utilised was curated by WormBase and includes the widest coverage of *C. elegans* PPI data, over four-fold more entries than BioGRID and approximately 70% more than IMEx member databases [171]. In addition, approximately 60% of the data entries curated by WormBase were unique to this curation effort.

Examining this WormBase dataset in further detail, specifically the alliance molecular interactions version 2.1 dataset (downloaded from the Alliance of Genome Resources download portal, available at <https://www.alliancegenome.org/downloads>), it includes 29 774 binary PPI data entries for *C. elegans*, curated from a total of 1308 publications. The majority of these interactions were detected by the Y2H technique (63%) and molecular sieving approaches (33%), with the remaining 4% detected by other methodologies. Hence, the distribution of distinct method detection strategies in this *C. elegans* PPI dataset is biased towards two technical approaches.

3.3.1.1 PINOT performance for *C. elegans* data

The performance of PINOT for querying *C. elegans* PPIs was compared to MIST [159], an alternative PPI query resource which also has *C. elegans* query capacity, the results of this comparison are reported in Figure 3.6. The data acquisition process which underlies these two resources differs slightly, as described, PINOT extracts data from the latest release of WormBase molecular interaction data, whereas MIST utilises data from several repositories, including WormBase, BioGRID and IMEx associated databases. In addition, there are distinct features to these resources, notably the interolog feature in MIST which suggests PPIs based on inference across species, however just evidence-based PPIs for *C. elegans* were used for this comparison.

Overall, the performance of PINOT, in relation to the number of data entries extracted, is comparable to MIST. More specifically, PINOT extracts slightly fewer data entries across the

query cases reported in Figure 3.6. However, upon assessing the completeness of the output data provided by these two resources, in relation to interaction detection method and/or PubMed ID annotations within the data entries, there was a striking difference. Since the PINOT pipeline focusses particular emphasis on the quality of data provided to the user (through a stringent quality control (QC) procedure), all data entries provided in the output file were complete, whereas in MIST, incomplete data entries persisted in the output dataset. In the more abundant data pools, for example when querying the ATP and CED protein (Figure 3.6), incomplete data entries accounted for the majority of the output dataset in MIST.

A further observation from this comparison is that the difference between the number of data entries extracted using PINOT with the stringent or lenient filter applied is very slight (and in most cases none existent; Figure 3.6). This reflects the WormBase curation effort (for the data utilised in PINOT) and in fact, only 381 data entries, which accounts for 1.3% of the entire WormBase dataset, are curated with an unspecified detection method annotation which are discarded in the stringent filter but retained in the lenient filter. Therefore, a difference in the PINOT output, with stringent or lenient filters applied, will only be apparent if the query proteins are present within that very small proportion of the dataset (1.3%).

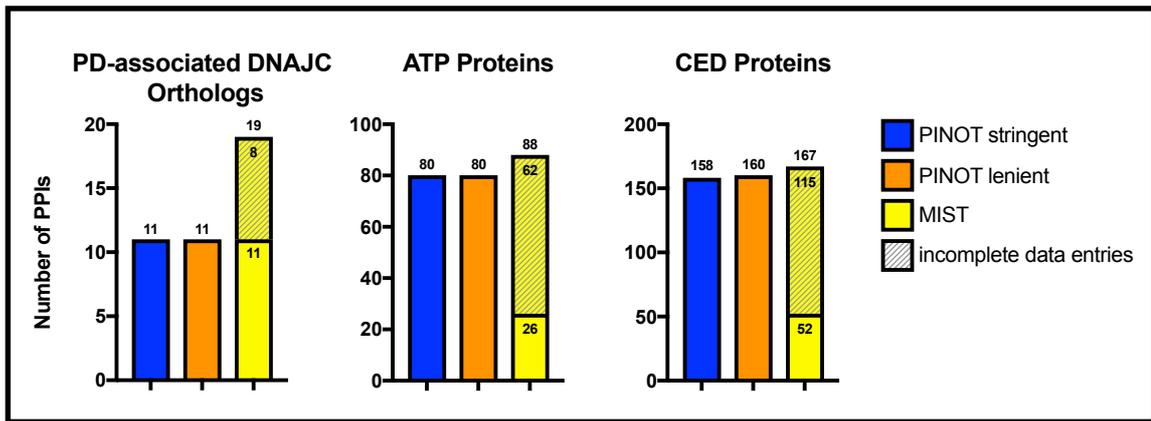


Figure 3.6 – Performance comparison of PINOT and MIST for *C. elegans* PPI data. The number of PPI data entries extracted when using PINOT (with stringent and lenient filters applied, independently) and MIST were assessed with three query lists: PD-associated DNAJC orthologs: DNJ-14, DNJ-25, DNJ-27, Y73B6BL.12, K07F5.16, RME-8 and GAKH-1; ATP proteins: ATP-1, ATP-2, ATP-3, ATP-4, ATP-5 and ATP-6; and CED proteins: CED-1, CED-2, CED-3, CED-4, CED-5, CED-6, CED-7, CED-8, CED-9, CED-10, CED-11, CED-12 and CED-13. The input format used for PINOT was the WormBase gene ID, the common gene name (as listed here) was used for MIST querying and no filter by rank parameter was set. The output data was evaluated in relation to the number of complete and incomplete (lacking detection method and/or PubMed ID annotations) data entries extracted. Analysis performed on 24th September 2019.

3.3.2 Predicted DAPK-1 protein interaction networks

First, an updated human DAPK1 interaction network (in relation to the DAPK1 interactome reported in Chapter Two) was generated using PINOT. Only interactors assigned a final score >2 within the PINOT pipeline and hence were experimentally replicated interactors, were retained for constructing the network. As a result, 34 interactors (including the DAPK1-DAPK1 interaction), from a total of 138 interactors extracted, were mapped onto the DAPK1 interaction network (Figure 3.7A).

Since the *C. elegans* DAPK-1 interaction network was limited to one reported interactor, which was absent from PPI database curation efforts, numerous approaches were adopted to gather insight into the potential DAPK-1 interaction landscape. The first approaches utilised orthologous inference, exploiting the MIST and Ortholist tools independently, to generate predicted DAPK-1 PPI networks (Figure 3.7B and 3.7C).

Surprisingly, there were fewer nodes within the more recent DAPK1 interaction network than were mapped onto the DAPK1 PPI network over two years previously, using an earlier version of the same method (from 38 in 2017 to 34 in 2019). The four nodes absent in the more recent analysis: MIB1, UBC, UNC5A and UNC5C, were all present in the raw output dataset with a final score of 2, whereas previously these interactors were assigned a final score >2. This was investigated to understand the reason for this and it was traced back to a modification of DAPK1 interactor data entries in the BioGRID repository during this time period, for example, one of the data entries listing the DAPK1-MIB1 interaction from a second publication was absent from the more recently acquired BioGRID data. This highlights that the curated data repositories are dynamic datasets and supports the need for providing a date (of data acquisition) alongside this type of analysis. Furthermore, it showed that advances in DAPK1 interactor data curation (and potentially detection), at least for replicated interactors, was stagnant during this time period.

3.3.2.1 Predicted DAPK-1 interactors using MIST

The interolog feature was used in MIST [159] to deduce DAPK-1 interactor predictions based on reported interactors in other species. These ortholog predictions are derived from the integration of the DRSC Integrative Ortholog Prediction Tool (DIOPT) [176] within MIST. Upon querying DAPK-1 in MIST, 25 proteins that ranked moderate within their custom confidence ranking scale were identified (Figure 3.7B), therefore these interactors have been inferred from multiple species, or a single species but identified by multiple methods, or both multiple species and multiple methods. In this analysis, the majority of orthologous inferences were derived from reported human DAPK1 interactions (57%), however inference from mouse and rat Dapk1 interactions also contributed, 31% and 12% respectively.

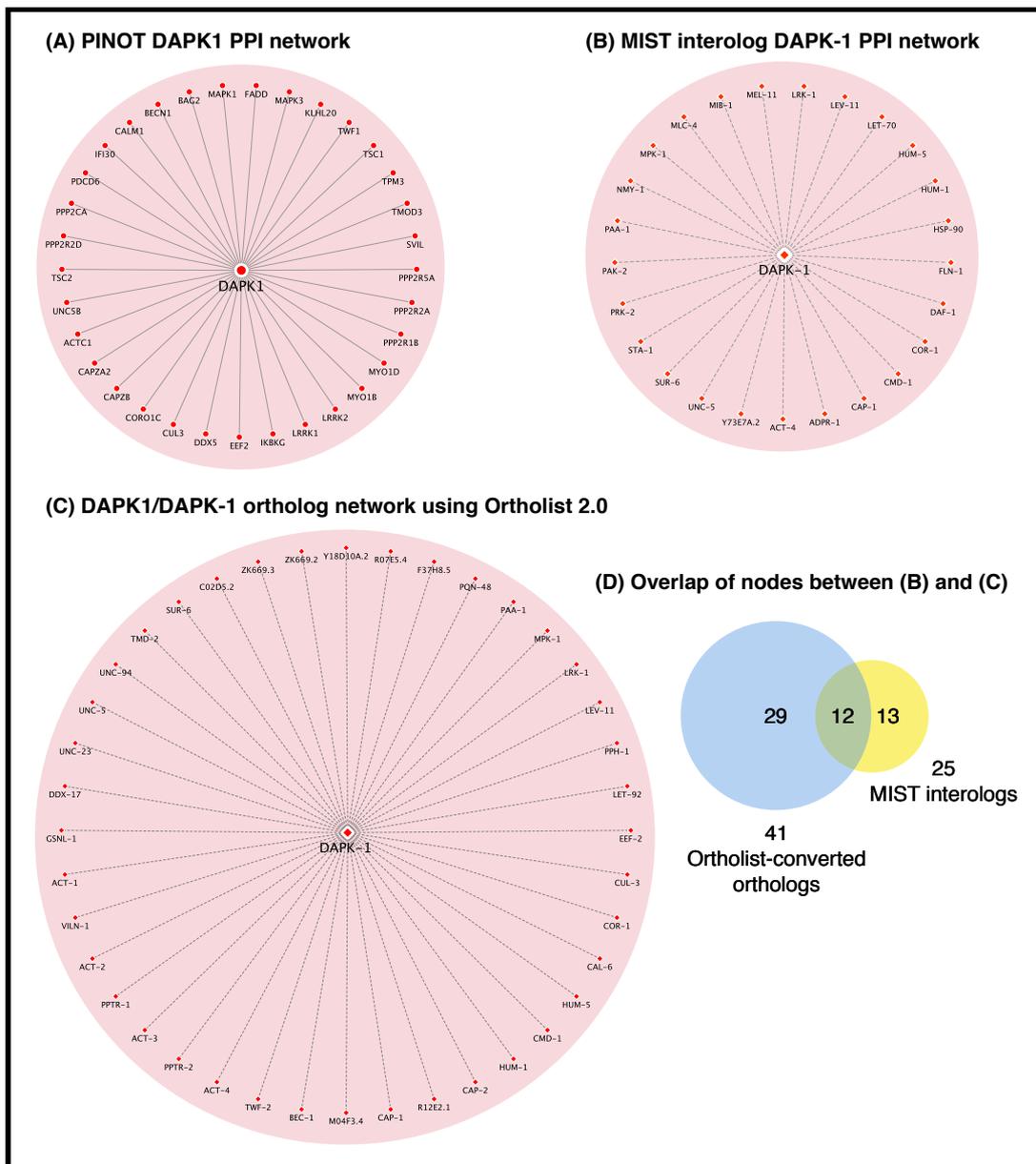


Figure 3.7 – DAPK1 protein interaction network and orthologous *C. elegans* interactors. (A) Human DAPK1 protein interaction network derived from PINOT on 25th June 2019. This network only displays interactions which have been replicated by interaction detection method and/or publication (i.e. PINOT final score >2). (B) Predicted *C. elegans* DAPK-1 protein interaction network derived from the Molecular Interaction Search Tool (MIST) implementing the interolog query option, i.e. these nodes represent potential interactors inferred from reported interaction within a different species. (C) Predicted *C. elegans* DAPK-1 protein interaction network based on the ortholog conversion of interactors within the human DAPK1 PPI network. Orthologs identified using Ortholist 2.0. (D) Number of overlapping nodes between the MIST-derived interolog PPI network and the Ortholist-converted ortholog PPI network.

3.3.2.2 Predicted DAPK-1 interactors using Ortholist 2.0

A second predicted DAPK-1 protein interaction network based on orthologous inference was constructed utilising the confidence-weighted human DAPK1 PPI network presented in Figure 3.7A as a foundation. For this approach, each node in the human network was submitted to Ortholist 2.0 to identify its *C. elegans* orthologs, these orthologs were then mapped onto a PPI network (Figure 3.7C). This network comprised of 41 predicted DAPK-1 interactors. Of note, in some cases multiple orthologs originated from a single human protein and all were retained.

Since the two processes underlying the predicted PPI networks were distinct, but both possessed a level of increased confidence weighting, moderate rank in MIST and derived from replicated human DAPK1 interactors (PINOT final score >2) in the Ortholist network, the networks were overlaid to assess their commonality. As a result, 12 proteins were identified as common between the two predicted networks (Figure 3.7D): ACT-4, CAP-1, CMD-1, COR-1, HUM-1, HUM-5, LEV-11, LRK-1, MPK-1, PAA-1, SUR-6 and UNC-5. This highlights a strategy for prioritisation of predicted interactors for physical and functional validation.

3.3.2.3 Incorporation of reported interactors within the predicted DAPK-1 network

The 41 predicted DAPK-1 interactors from the Ortholist-derived ortholog network (Figure 3.7C) formed the basis of a further analysis of the reported interactors of these predicted interactors. This analysis, performed using the *C. elegans* query option in PINOT, extracted 400 binary interactions that involve the 41 predicted DAPK-1 interactors. Of these reported interactions, 45 were assigned a final score >2 within the PINOT pipeline and these were mapped onto the Ortholist-derived predicted network to add a second layer of interactors to the potential DAPK-1 interactome (Figure 3.8). Of note, this retention of only 11% of interactors from reported to replicated, based on the PINOT scoring procedure, supports the observation that the majority of the PPI data are derived from Y2H screens and it seems only a small proportion are replicated in distinct studies.

This hybrid, prediction and evidence-based, PPI network (Figure 3.8) provides wider and strengthened insight into the potential interactome of DAPK-1, than with predictions alone. Interestingly, a number of the predicted DAPK-1 interactors share common reported interactors, such as LEV-11 and PAA-1 (both of which were within the overlapping subset of nodes from the MIST and Ortholist network comparisons) which are both reported to interact with SMO-1. Furthermore, two of the predicted DAPK-1 interactors, LET-92 and SUR-6, are reported to

interact together and with a common interactor, SAS-6, potentially positioning DAPK-1 in a signalling cascade and/or complex with these three proteins. Potential connectivity of proteins, such as these examples, were apparent in numerous cases (Figure 3.8). This network analysis approach highlights a second strategy for prioritisation of predicted interactors for follow-up studies.

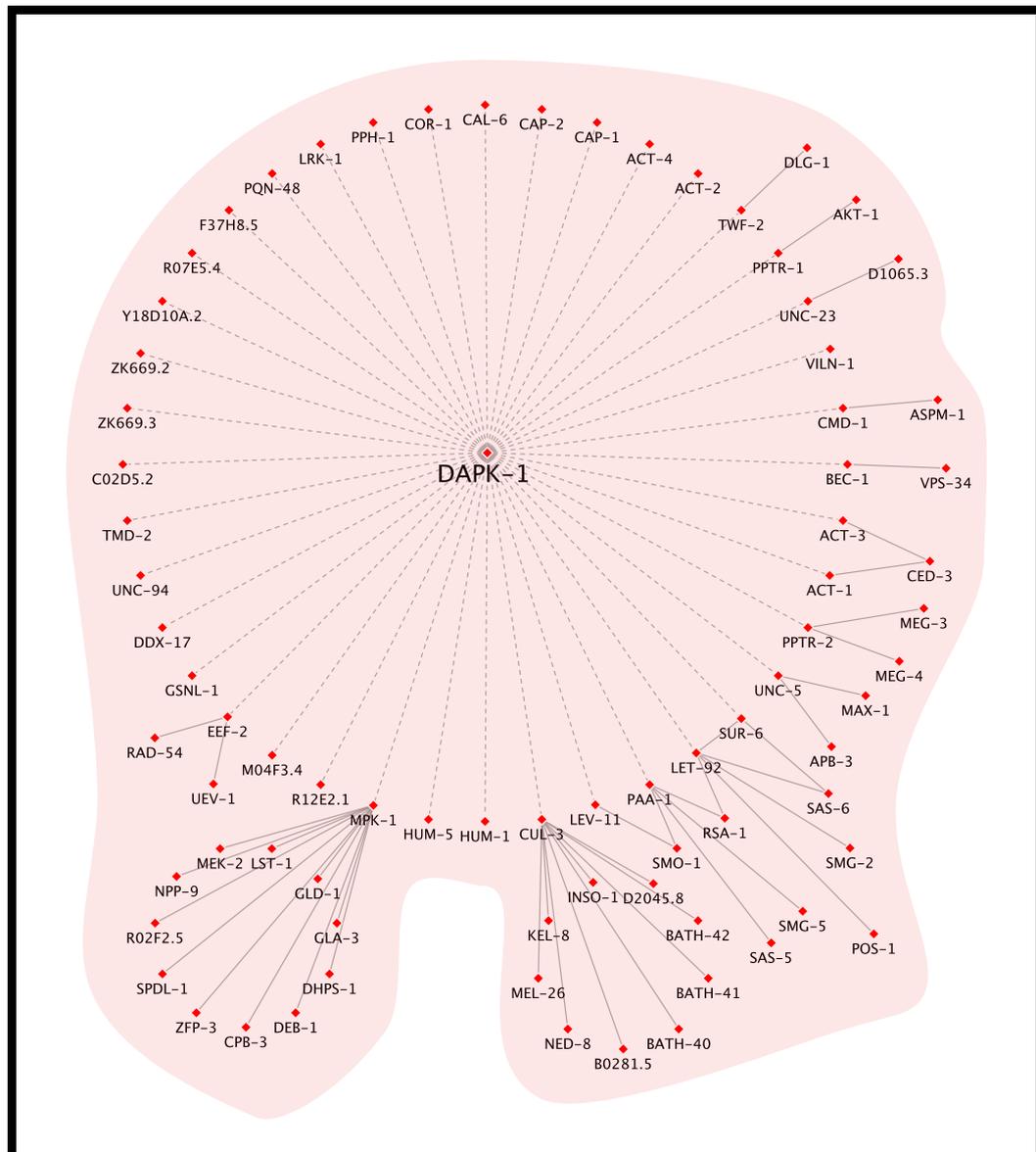


Figure 3.8 – Predicted DAPK-1 protein interaction network based on orthologous interactor predictions and reported interactions of these nodes. Dashed edges indicate predicted interactions based on *C. elegans* ortholog conversions of nodes within the human DAPK1 PPI network. Orthologs identified using Ortholist 2.0. Solid edges indicate reported interactions derived from PINOT, with a confidence threshold set to only retain replicated PPIs.

3.3.3 Hybrigenics DAPK-1 yeast two-hybrid screen

The DAPK-1 Y2H results were obtained by Hybrigenics (France), this experiment was outsourced to this company based on their expertise in this area.

3.3.3.1 *dapk-1* sequence validation and compatibility

Cloning of the full-length *dapk-1* bait sequence into a DBD-fusion vector was performed and validated by Hybrigenics. Initial attempts to assemble this sequence into their vector by ligation of three fragments proved unsuccessful in terms of reconstituting and validating the correct sequence. Therefore, gap repair cloning (directly in yeast) was performed. The sequence validation results from this cloning approach showed positive validation of the *dapk-1* sequence at the protein level, however two synonymous mutations at the DNA level were present, C1350T (I450I) and G2922A (K974K). As these variants did not alter the coding sequence of DAPK-1, this sequence was considered valid for progressing with the screen.

Prior to the full library screen, the *dapk-1* bait was screened on a small-scale to assess its compatibility with the Y2H system. This assessment showed that *dapk-1* was neither toxic nor autoactivating within the experimental setup and therefore this bait was subjected to the full library screen.

3.3.3.2 Identification of novel DAPK-1 interactors

The Y2H screening of the full-length *dapk-1* bait against the *C. elegans* mixed stage prey library resulted in 268 positive clones, from approximately 157 000 000 potential interactions tested. These 268 positive clones were processed by sequencing which identified nine protein binding partners of DAPK-1, three of which were proven technical artefacts of the experimental system. Hence, six novel DAPK-1 interactors were identified (Table 3.5): CMD-1, MEP-1, SYD-9, UNC-14, C39E9.12 and F13H8.5. CMD-1 (or calmodulin) was the only interator identified with very high confidence and this hit was overrepresented within the positive clones (78% of all positive clones). The other five interactors detected were of moderate confidence and were detected from a low number of positive clones (Table 3.5), in some cases just one positive clone corresponded to these hits.

The confidence ratings assigned to these positive hits are derived from a 'predicted biological score' that Hybrigenics calculate based on technical parameters, such as the number of prey

fragments that contribute towards the overall pool of positive clones, and using data generated from previous Hybrigenics screens of the same species. A further feature provided by Hybrigenics in the result output was the identification of a selected interaction domain (SID), which corresponds to the minimum region of protein sequence interacting with the bait protein. The SID boundaries for each interactor identified are listed in Table 3.5. This is calculated based on the prey fragments which interact with the bait and provides an insight into the interaction interface of the prey protein. Of note, if only one prey fragment corresponds to a positive hit, the SID will be the full length of that fragment, whereas if multiple different prey fragments correspond to a positive hit, the SID is likely to be more specific and may represent the interaction interface with increased accuracy.

Table 3.5 – Novel DAPK-1 interactors identified by yeast two-hybrid (Y2H). A summary detailing the positive hit results identified from the outsourced Y2H screen (Hybrigenics, France). Each interactor was assigned a confidence rating and a selected interaction domain (SID) based on the analysis of the sequenced prey fragments. aa = amino acid

| Protein Y2H Positive Hit | Hybrigenics Confidence Rating | Number of Sequenced Clones that Match Hit | Protein Length (aa) | Selected Interaction Domain (SID) (aa) | Human Orthologs |
|-------------------------------------|-------------------------------------|--|---------------------------|---|--|
| CMD-1 T21H3.3 WBGene00000552 | Very high confidence | 209 | 149 | 80-145 | CALM1, CALM2, CALM3 (Calmodulin) |
| MEP-1 M04B2.1 WBGene00003218 | Moderate confidence | 1 | 870 | 400-854 | ZFAT, ZFX, ZFY, ZNF711 |
| SYD-9 ZK867.1 WBGene00044068 | Moderate confidence | 1 | 452 | 54-297 | ZFP57 |
| UNC-14 K10D3.2 WBGene00006753 | Moderate confidence | 6 | 665 | 51-335 | - |
| - C39E9.12 WBGene00008035 | Moderate confidence | 2 | 409 | 2-381 | - |
| - F13H8.5 WBGene00017438 | Moderate confidence | 1 | 513 | 250-507 | - |

3.3.4 Generation of a FLAG-dapk-1 *C. elegans* strain

A 3xFLAG tag was engineered into the *C. elegans* genome at the endogenous 5' *dapk-1* locus to facilitate further experimentation centred around the DAPK-1 interaction profile. This genomic modification was outsourced to SunyBiotech (China).

3.3.4.1 FLAG-*dapk-1* sequence validation

SunyBiotech provided sequencing results supporting the validation of the 3xFLAG tag knock-in at the desired locus. This involved sequencing a 705bp region which included the introduced sequence. The 3xFLAG and linker sequence was validated in line with the expected sequence insertion, however two synonymous mutations were identified which correspond to exon 1 of *dapk-1*, C72T (F24F) and A105T (G35G). In addition, the introduction of the 3xFLAG and linker sequence was validated in-house by PCR and restriction digestion (Figure 3.9; see Figure 3.5 for experimental design and expected results). This analysis also supports the successful insertion of the 3xFLAG and linker sequence at the desired locus.

In brief, the expected amplified fragment size when using primers F₁ and R₁ was 789bp and 693bp on the FLAG-*dapk-1* and N2 genetic backgrounds, respectively. Upon digestion with *BseRI*, which targets a region within the 3xFLAG and linker sequence, the fragments derived from the FLAG-*dapk-1* were expected to be cleaved into fragments of length 510bp and 279bp, whereas the N2-derived fragments were expected to remain intact upon incubation with the *BseRI* endonuclease. This expectation matches the result across all replicates of the FLAG-*dapk-1* and N2 samples (Figure 3.9A), with some undigested fragments persisting in the FLAG-*dapk-1* digested samples. For PCR using primers F₂ and R₁, whereby F₂ was designed to anneal to the 3xFLAG sequence, the expected amplified fragment size was 354bp on the FLAG-*dapk-1* template background whilst no amplification was expected from the N2-derived samples. The gel image corresponding to this PCR (Figure 3.9B) showed expected amplification with regard the FLAG-*dapk-1*-derived samples. However, unexpected weak signal was evident in the N2-derived samples at approximately 200bp and 800bp, although due to the fragment size and signal intensity this is likely as a result of off-target binding of the primers.

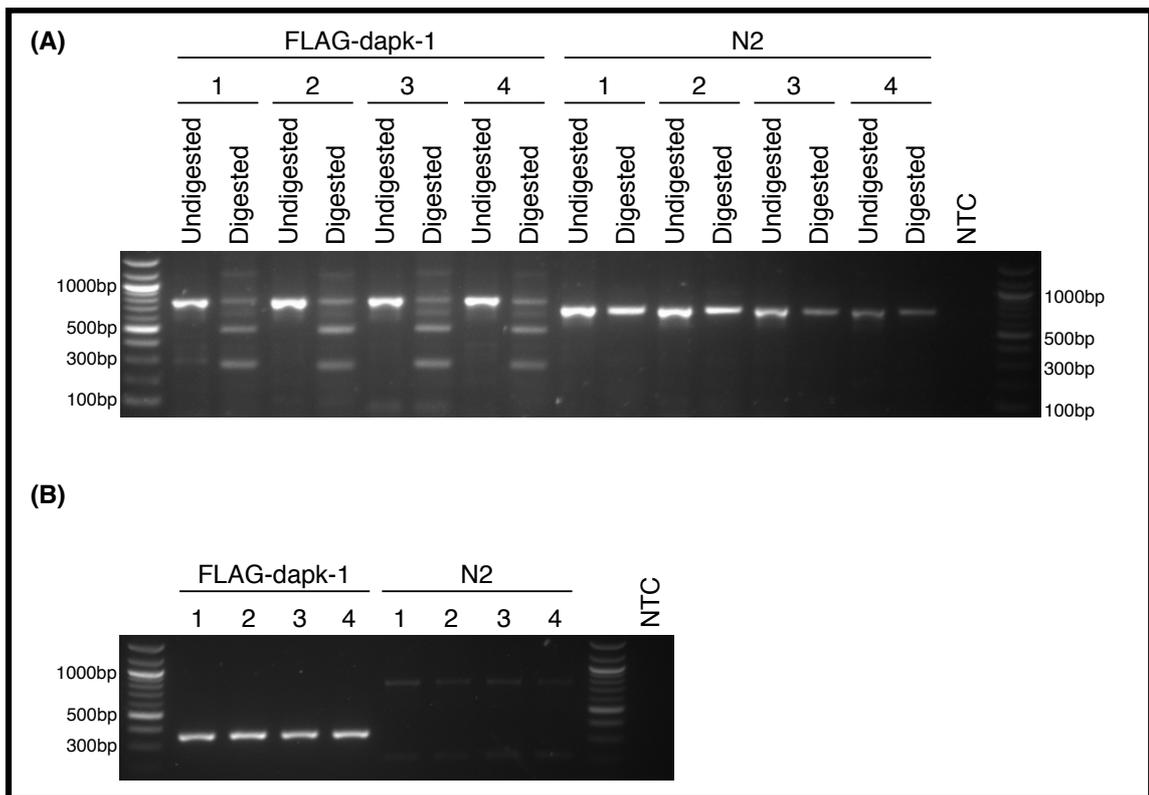


Figure 3.9 – FLAG-dapk-1 tag validation at the genomic level. PCR of two DNA fragments at the 5' *dapk-1* locus to validate insertion of a 3xFLAG tag. (A) Agarose gel image of DNA fragments derived from PCR of a region spanning the 5' untranslated region, *dapk-1* exon 1 and a region of *dapk-1* intron 1, followed by restriction digestion with *BseRI* which targets a recognition site in the inserted 3xFLAG-linker region. (B) Agarose gel image of DNA fragments derived from PCR, whereby the forward primer was designed to anneal to the inserted 3xFLAG sequence. Lane numbering corresponds to sample replicates; NTC, no template control.

3.3.4.2 FLAG-DAPK-1 protein validation

Synchronised life stage *C. elegans* populations were lysed for protein extraction using various lysis buffer compositions and processed for Western blot analysis in an attempt to detect FLAG-DAPK1. Following gel electrophoresis and electro-transfer of the samples, the membranes were cut and probed for either FLAG or β -actin. Samples were collected at various *C. elegans* life stages since DAPK-1 expression is likely to be dynamic through the developmental stages. The rationale underlying this was to identify an appropriate life stage whereby DAPK-1 is expressed at a detectable level (and at its highest) to inform future experimentation with regard FLAG-DAPK1 AP-MS. Each lysis method was replicated at least twice and the resulting blots showed similar outcomes. Representative blot images from these analyses are displayed in Figure 3.10.

From this analysis, it is evident that detecting the FLAG epitope in the FLAG-*dapk-1* *C. elegans* strain proved challenging. The expected result for positive validation of the tag at the protein level would be specific signal in the FLAG-*dapk-1* samples at approximately 160kDa (which would correspond to full-length DAPK-1), which was absent in the N2 samples. *C. elegans* lysis in WLB2 was the initial lysis method used and no specific FLAG-DAPK-1 signal was observed across any of the samples (Figure 3.10A), even following exposure times exceeding 30 minutes during imaging. Specific detection of a β -actin ortholog, likely to be ACT-2, was observed when probing with the anti- β -actin antibody. This verified that this antibody was suitable for *C. elegans* samples, reactivity in this species was not previously reported by the supplier. Furthermore, this signal acted as a positive control for protein extraction.

To identify whether proteolytic cleavage of DAPK-1 may be occurring which may be a reason for the lack of specific signal at the expected molecular weight of full-length FLAG-tagged DAPK-1, samples lysed by the same method (lysis in WLB2) were subjected to Western blot analysis, but for this experiment, the entire membrane was probed for the FLAG epitope (Figure 2.10B). From this blot, high intensity off-target signal was evident across all samples (including negative controls) at an approximate molecular weight of 60kDa. FLAG-*dapk-1* strain specific signal was evident in samples derived from L2 and L4 stage nematodes between 15-35kDa and 15-25kDa, respectively.

Due to unsuccessful validation of FLAG-DAPK-1 using the WLB2 lysis buffer, the lysis and protein preparation methods were then altered to use more stringent lysis buffers, first RIPA buffer and then lysing the samples direct in denaturation buffer. Following nematode lysis in RIPA buffer, again, no specific signal in the FLAG-*dapk-1* samples was observed (Figure 3.10C). Western blot was then trialled on samples derived from lysis directly in denaturation buffer. Estimates of protein quantity were not performed on these samples and this resulted in uneven loading, in relation to the amount of protein loaded per well. This is reflected in the blot image (Figure 3.10D), whereby the lanes corresponding to day 10 samples were overloaded in relation to other samples. This image was overexposed in an attempt to identify any subtle but specific FLAG-DAPK-1 related signal. However, this was not observed.

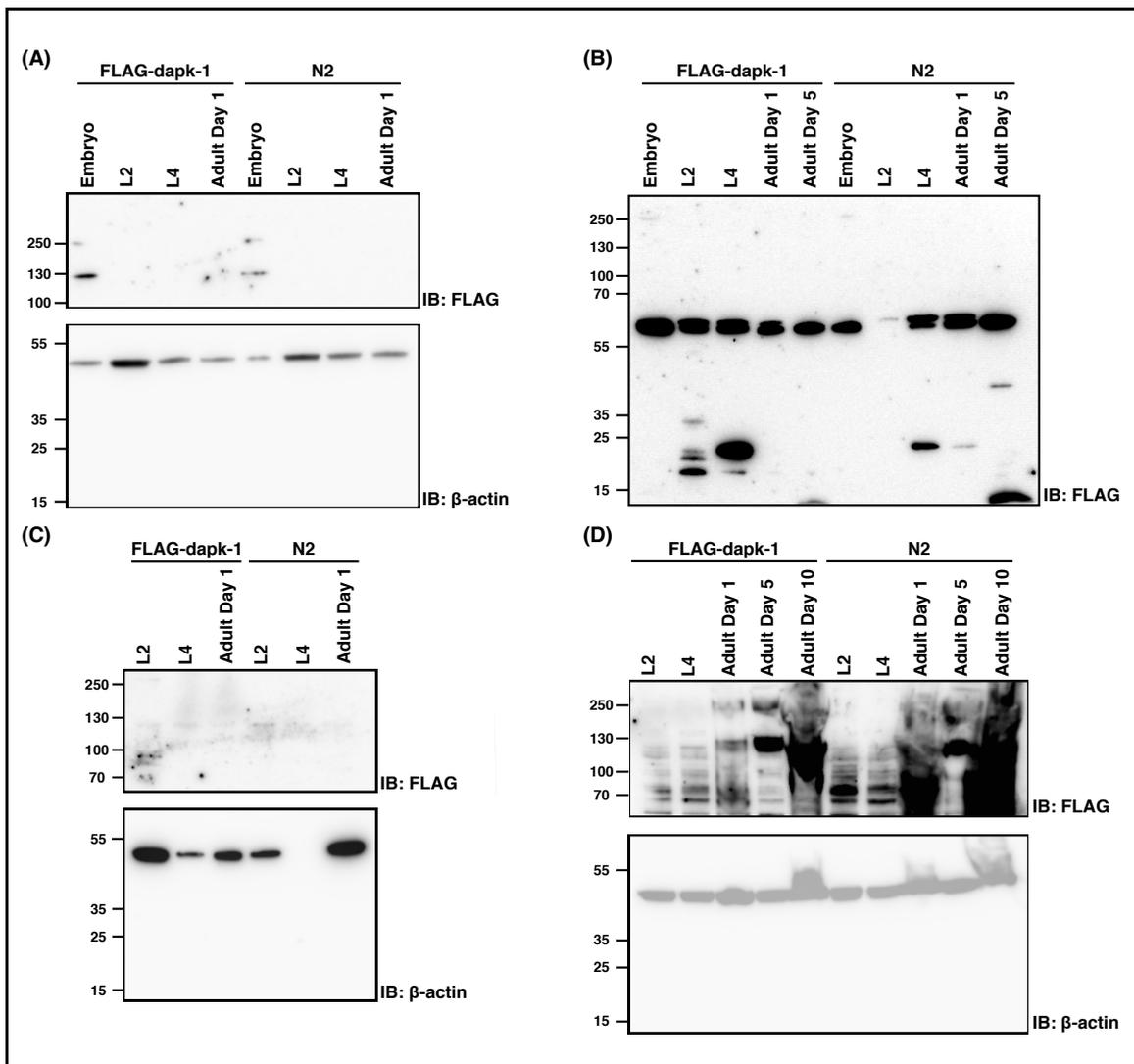


Figure 3.10 – FLAG-DAPK-1 tag validation at the protein level. Representative blot images of various protein preparation methods used in an attempt to detect FLAG-DAPK-1 within lysates from *C. elegans* at different life stages. (A) Samples were lysed in WLB2. (B) Samples were lysed in WLB2 and the whole blot was probed for the FLAG epitope. (C) Samples were lysed in RIPA buffer. (D) Samples were lysed in denaturation buffer. FLAG-associated images were overexposed in attempt to detect weak but specific signal. β -actin was used as positive control for protein extraction and loading control across equivalent life stages.

3.3.4.3 FLAG-*dapk-1* transcript analysis

Since convincing evidence to support the expression of FLAG-DAPK-1 at the protein level was lacking, qPCR was performed to assess whether the FLAG tag was detectable at the transcript level. The amplification reactions targeted two regions of the *dapk-1* sequence, spanning exon junctions at the 5' and 3' end of the gene, the FLAG sequence and exons 1-2 of *tba-1* (housekeeping gene). The expected PCR product sizes were 99bp, 147bp, 103bp and 124bp for *dapk-1* exons 1-2, *dapk-1* exons 12-13, FLAG and *tba-1* primer combinations, respectively. From agarose gel electrophoresis of the qPCR products, it was evident that specific amplification was achieved across the four primer combination reactions and hence *dapk-1* and *tba-1* mRNA was expressed (Figure 3.11A). Moreover, amplification targeting the FLAG sequence was detected, specific to the FLAG-*dapk-1* samples, suggesting transcription of FLAG-*dapk-1*.

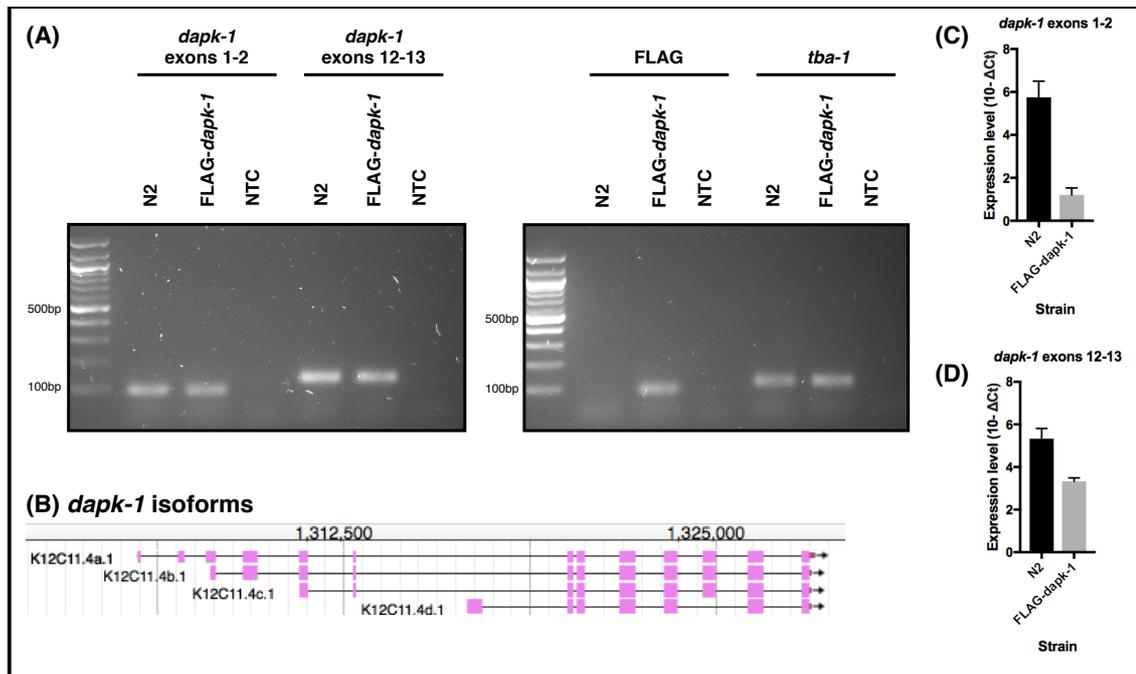


Figure 3.11 – FLAG-*dapk-1* transcript analysis by qPCR. (A) Agarose gel electrophoresis of qPCR products derived from cDNA amplification of four loci: exons 1-2 and exons 12-13 of *dapk-1*, FLAG and *tba-1*, in FLAG-*dapk-1* and N2 nematodes. (B) Exon intron topology of *dapk-1* splice variants, schematic obtained from WormBase. (C) Relative expression level of *dapk-1* exons 1-2 in FLAG-*dapk-1* and N2 nematodes. (D) Relative expression level of *dapk-1* exons 12-13 in FLAG-*dapk-1* and N2 nematodes. Relative expression levels presented as delta- C_t values subtracted from an arbitrary value ($n=3$).

The preliminary quantitative data from this experiment indicates a decrease in *dapk-1* mRNA expression in the FLAG-*dapk-1* strain when compared to the N2 strain, measured at two exon junctions of the gene (Figure 3.11C and 3.11D). The relative expression levels of FLAG-*dapk-1* strain derived transcripts were reduced to approximately 20% and 60% of N2 strain derived transcript expression, at exons 1-2 and exons 12-13 regions of *dapk-1*, respectively (exon numbers corresponds to full-length gene). This is based on fixed cycle threshold (C_t) values of the *dapk-1* targeted amplifications (exons 1-2 or 12-13) normalised against the C_t values for the housekeeping gene measured from the same experimental sample. Of note, shorter splice variants of *dapk-1* lack the 5' portion of the gene (Figure 3.11B).

3.4 Discussion

The approaches developed and used in this body of research harnessed predictive and evidence-based PPI mapping to provide novel insight into the interaction profile of *C. elegans* DAPK-1. The orthologous transition of PPI network analysis from human DAPK1 to *C. elegans* DAPK-1 highlights a strategy for PPI inference for poorly defined interactomes, especially for popular model organisms, such as *C. elegans*, whereby PPI query and ortholog conversions tools are becoming well-established [121,159,172,177]. In addition to PPI predictions, novel DAPK-1 interactors were identified using Y2H technology and an epitope tag was engineered into the *C. elegans* genome at the *dapk-1* locus to facilitate further DAPK-1 PPI screens.

3.4.1 Capturing *C. elegans* PPI data in PINOT

The vision of incorporating a *C. elegans* query option into PINOT led to the successful development of a data processing pipeline which was trialled and integrated into the online version of the resource [121]. This pipeline functions similarly to the pipeline underlying PINOT for human PPI data, however a number of steps are performed prior to the query submission, i.e. the *C. elegans* PPI dataset is pre-processed (as described). This additional species capability broadens the use of PINOT which will, as a result, hopefully be of valuable resource for the *C. elegans* research community.

The added value of PINOT for *C. elegans* PPI data query above alternative resources is multi-layered. First and in relation to data coverage, PINOT utilises data from WormBase [103], which is the primary repository for *C. elegans* information and hence curates the widest coverage of *C. elegans* PPI data [171]. Updated versions of this WormBase dataset are released approximately every 4 months and therefore to maintain an up-to-date resource for *C. elegans* PPI querying, the *C. elegans* dataset within PINOT will be updated in conjunction with these releases.

MIST [159], an alternative tool for *C. elegans* PPI querying, collates data from numerous sources, which includes WormBase and a number of the IMEx member repositories, and hence a wider coverage of data would be expected in comparison to PINOT. This was the case, although the difference was marginal. However, the data quality provided by MIST was compromised, in relation to completeness, in comparison to PINOT (Figure 3.6). In addition and similarly to PINOT for *C. elegans* data, MIST utilises banked PPI data from a bulk download event direct from the source repositories, however it is unclear how regularly this banked dataset is updated, although

the date or version of the latest downloaded dataset is available. Therefore the user may not be provided with an up-to-date representation of the interaction profiles for proteins of interest if there is a considerable time gap between MIST downloading the latest PPI dataset and the time of query.

An alternative approach to access this WormBase curated PPI data is *via* specific gene pages on the WormBase website. Upon querying a gene and then selecting the 'Interactions' tab, a network visualisation of reported PPIs is displayed. However, this can only be done on a protein-by-protein basis and multiple proteins cannot be queried simultaneously. In contrast, there is no limit to the number of proteins that can be queried during a PINOT query submission, therefore this WormBase data can be extracted, collated and considered collectively, in relation to a query input list. This highlights a further advantage of using PINOT for *C. elegans* PPI querying.

As described, the composition of this WormBase curated data is skewed towards PPIs detected by the Y2H approach. Due to the inherent limitations of this technique (briefly outlined in Table 2.8), such as PPI detection only occurring within the nucleus; bait and prey proteins typically overexpressed and in fragments; system potentially lacks the physiological post-translational modification (PTM) and accessory molecule profile for proteins of interest which could impact protein folding; *etc*, these considerations should be accounted for when interpreting the *C. elegans* PPI data from the PINOT output. The final score assigned to each PPI within the PINOT output acts as a confidence score and will facilitate this interpretation. However, it is likely that when using PINOT for *C. elegans* data, the majority of PPIs will score a final score of 2 (which was the case when mapping the DAPK-1 double layered network [Figure 3.8]), for example, a reported PPI based on a single detection from a high-throughput Y2H screen (such as the *C. elegans* interactome study from the Vidal laboratory [178]), and these PPIs should be interpreted with caution.

3.4.2 Creating predicted DAPK-1 PPI networks from orthologous inference

Numerous roles have been proposed for *C. elegans* DAPK-1 (as described in the introduction of this chapter), however mechanistic characterisation of this protein in the context of signalling events remains an area for future investigation. In addition, the interaction profile of *C. elegans* DAPK-1 was largely a blank canvas, with only one reported interactor [166]. Therefore, two strategies were used to map predictive PPI networks for DAPK-1, using MIST and Ortholist 2.0 for orthologous inference from DAPK1 PPIs in other species.

Despite a similar strategy underlying these two approaches, the predicted PPI networks differed in terms of the number of nodes mapped (Figures 3.7B, 3.7C and 3.7D), this reflects the differences in the methodology underlying the ortholog assignment. The key differences are that MIST utilises data from multiple species (human, mouse and rat in this case), whereas Ortholist performs *C. elegans* ortholog conversions from human only and the underlying human DAPK1 PPI network was derived from PINOT. Furthermore, the orthology analysis tools integrated into MIST and Ortholist differ: MIST encompasses 15 orthology algorithms for *C. elegans* via the integration of the DIOPT [176], while Ortholist utilises 6 orthology algorithms [172]. Of note, 5 of these programs incorporated into Ortholist are also used in the MIST interolog feature.

Based on these differences, MIST incorporates a wider coverage of both, species ortholog conversions and orthology algorithms, and would be expected to output a wider array of orthology-predicted PPIs than the Ortholist approach. However, the confidence parameters are unmatched in these two approaches, which contributes to the differences in network size. The ortholog selection in MIST utilises the highest stringency filter in the DIOPT for orthology analysis [159] and hence these interactors are high confidence orthologs. In contrast, when using PINOT coupled with Ortholist, the original reported PPIs are of increased confidence (only interactors with a final score <2 in PINOT) but the ortholog mapping does not filter based on confidence, unlike MIST, hence a wider collection of orthologs are present. Therefore, these two distinct approaches taken together provide a useful strategy for not only predicting PPIs based on orthology, but prioritising these predicted interactors by identifying nodes common to both approaches (i.e. high confidence orthologs derived from replicated reported PPIs [Figure 3.7D]) for validation.

Evidence-based PPI mapping was then incorporated into the Ortholist-derived predicted network to add an element of confidence-weighted experimental data to the potential DAPK-1 interaction landscape. This facilitates a vision of where DAPK-1 may fit within the functional landscape of *C. elegans* biology. Despite a lack of reported interactors for DAPK-1, DAPK-1 predicted interactors possessed many reported interactors within *C. elegans*, with a degree of confidence (Figure 3.8). Therefore, this positions DAPK-1 in the context of potential proximal signalling events, providing routes for DAPK-1 related pathway investigation, in the context of computational pathway inference modelling [179] and functional 'wet-lab' studies.

Furthermore, several nodes of this predicted layer of DAPK-1 interactors possess connectivity

through reported interactions with other predicted DAPK-1 interactors and/or *via* common nodes within the second layer of evidence-based interactors. These cases are of particular interest, because they position DAPK-1 in an already established interconnected network which may suggest the involvement of DAPK-1 in a series of signalling events and/or protein complex. An example of this interconnectivity is between LET-92, SUR-6, SAS-6, RSA-1, PAA-1, SMO-1 and LEV-11 (see lower right side of the network presented in Figure 3.8), which includes a number of predicted DAPK-1 interactors.

LET-92, SUR-6, RSA-1 and PAA-1 are subunits of protein phosphatase 2A (PP2A) [180], which may act on DAPK-1, similarly to human DAPK1 and PP2A [66,67] although the target phosphorylation site (Ser308) is not conserved in *C. elegans* DAPK-1. Alternatively, PP2A may act in concert with DAPK-1 to regulate the phosphorylation status of substrates. Of note, PP2A in *C. elegans* (as a complex and in subunits) influences numerous signalling cascades and hence is functionally linked to a diverse range of biological processes. In addition, several of these interconnected nodes aforementioned are functionally associated with the cell cycle [181–183], and therefore this potentially positions DAPK-1 in the same functional sphere.

3.4.3 Novel DAPK-1 binding partners

Further investigation into the DAPK-1 interaction network was achieved by outsourcing a Y2H screen whereby full-length DAPK-1 was used as the bait protein against a mixed life-stage *C. elegans* prey library. By using this approach experimental PPI data was generated in a hypothesis-free and high-throughput manner. From this screen, six DAPK-1 interactors were identified (Table 3.5). These results uncover novel insight into the DAPK-1 interaction profile and since this was largely undefined previously, provides routes for perturbing potential DAPK-1 signalling events for understanding the role of this protein at molecular and mechanistic levels.

Unsurprisingly, although previously unreported, CMD-1 (or calmodulin) was identified as very high confidence DAPK-1 interactor. This was an expected result since calcium/calmodulin regulation of the kinase activity in human DAPK1 is well characterised [63] and based on structural conservation is likely to also be the case for the *C. elegans* ortholog. In addition, CMD-1 shares a high degree of homology with human calmodulin at a structural level (98% identity), differing by only three residues. Furthermore, CMD-1 was a node present in both predicted DAPK-1 PPI networks (Figures 3.7B and 3.7C). The CMD-1 interaction interface mapped to the C-terminal portion of the protein, within residues 80-145 (Table 3.5). The DAPK-1-CMD-1 interaction is unlikely to elicit a specific DAPK-1 functional pathway, but rather play a critical role

in kinase activation and in turn, influence a multitude of potential signalling cascades. Calmodulin is a regulatory molecule for a whole class of kinases, plus a plethora of other enzymes and structural proteins, within the *C. elegans* proteome [184]. Hence, CMD-1 RNAi results in disruption to numerous fundamental cellular process, such as apoptosis, cell migration and cell proliferation (Table 3.6).

The other DAPK-1 interactors identified by the Y2H screen: MEP-1, SYD-9, UNC-14, C39E9.12 and F13H8.5, were rated with moderate confidence and therefore these results, in particular, would benefit from replication *via* alternative method detection strategies to assess their validity. Nevertheless, these positive hits provide novel insight into the potential DAPK-1 interactome and a foundation for hypothesis-driven DAPK-1 PPI studies. Despite a lack of functional and phenotypic data for two of these potential interactors, the possible role of DAPK-1 can be gleaned from the current understanding of MEP-1, SYD-9 and UNC-14 function (Table 3.6). The human orthologs of these proteins (listed in Table 3.5) were not reported human DAPK1 interactors.

Table 3.6 – Functional and phenotypic insight into novel DAPK-1 interactors identified by Y2H. A list of the six DAPK-1 protein binding partners identified by the yeast-two hybrid approach alongside the current understanding of the protein function and the associated *C. elegans* phenotypes upon genetically manipulating the corresponding gene.

| Protein Y2H Positive Hit | Functional insight | Associated <i>C. elegans</i> phenotype upon gene knock-out ⁻ , RNAi ⁺ , point mutation [*] | Refs |
|-------------------------------------|--|---|-----------|
| CMD-1 T21H3.3 WBGene00000552 | Regulatory calcium-binding molecule for many proteins | Embryonic lethal ⁺ (RNAi microinjected). Apoptosis ⁺ , cell cycle ⁺ and cell migration ⁺ defects, increased cell proliferation ⁺ (RNAi fed). | [185–187] |
| MEP-1 M04B2.1 WBGene00003218 | Larval, oocyte and gonad development, cell differentiation, chromatin regulation | Lethal ⁺ , L1 development arrest ⁺ , slowed growth ⁺ , vulval abnormalities ⁺ , epithelial and intestinal morphology variant ⁺ | [188–191] |
| SYD-9 ZK867.1 WBGene00044068 | Synaptic function, endocytosis | Axon morphology variant ⁻ , synaptic defects ⁺ , developmental delay ⁺ , egg retention ⁺ , enlarged intestine ⁺ , reduced number of progeny ⁺ | [192] |
| UNC-14 K10D3.2 WBGene00006753 | Synaptic function, axon guidance, membrane trafficking | Axon morphology variant ⁺ , locomotion variant ⁺ | [193–196] |
| C39E9.12 WBGene00008035 | Unknown | - | - |
| F13H8.5 WBGene00017438 | Unknown | - | - |

Interestingly, there are common phenotypic traits upon genetic manipulations of these interactors, such as development delay in mutant *mep-1* and *syd-9 C. elegans* [189,192]. Furthermore, MEP-1 and SYD-9 are reported to physically interact [197], potentially positions DAPK-1, MEP-1 and SYD-9 in a functionally relevant interaction triad. Of further interest is the functional commonality between these hits, for example SYD-9 and UNC-14 are both involved in synaptic function and vesicular trafficking. In addition, upon genetic manipulation of the corresponding genes, disruption to axon morphology is evident [192,195,198]. With DAPK-1 possessing functional links to autophagy and cytoskeletal dynamics [161], validating and characterising the interplay between these three proteins appears a logical route for further investigation. Furthermore, UNC-14 contains a RUN domain, a PPI interface which is known to interact with GTPases [199].

3.4.4 Validation challenges with the FLAG-dapk-1 *C. elegans* strain

The generation of a 3xFLAG-tagged *dapk-1 C. elegans* strain was to enable in-house experimentation into the *in vivo* DAPK-1 interaction profile, with the aim of performing AP-MS on FLAG-DAPK-1 isolated from nematode lysates. The reason for inserting an epitope tag at the endogenous *dapk-1* locus was two-fold. First, to enable detection and isolation of the protein since DAPK1 antibodies (developed for the human ortholog) are of low quality. Second, for this detection to be representative of the endogenous DAPK-1 expression profile, avoiding overexpression systems which may result in artefactual PPI events. The FLAG tag was chosen since it is small and hydrophilic [200,201], therefore the likelihood of it interfering with the biological activity and folding of DAPK-1 is low. Moreover, the triplication of the FLAG sequence hugely increases the detection efficiency [202] and since DAPK-1 is likely to be expressed at a low level (Figure 3.12), sensitivity of the detection system was a key consideration.

Upon receiving the FLAG-*dapk-1* strain, a number of validation steps were performed to ensure the tag had been inserted as expected, prior to further experimentation. Sequencing data provided by SunyBiotech and PCR followed by restriction digestion of genomic DNA performed in-house validated the insertion of the tag at the genomic level. However, attempting to detect the FLAG tag within protein extracts derived from this strain was challenging and convincing validation at the protein level is still lacking. The detection of FLAG-DAPK-1 was expected to be difficult, since mRNA expression data for the corresponding transcript showed relatively low expression throughout the *C. elegans* life stages (Figure 3.12). Nevertheless, detecting the 3xFLAG epitope using the anti-FLAG M2 monoclonal antibody was thought to be a highly sensitive and specific. Specific signal corresponding to full-length FLAG-DAPK-1 was absent in all

life stages tested (embryos through to 10-day old adults). Based on the mRNA data available *dapk-1* expression is highest during embryonic and L2-L3 larval stages (Figure 3.12) and further evidence suggests expression heightens during ageing, marginally at around day 6 and then substantially post day 16 [203]. Therefore, ageing *C. elegans* for longer may help in detecting FLAG-DAPK-1.

To troubleshoot this issue of FLAG detection, nematode samples were lysed in lysis buffers of increasing stringency, first using WLB2 which contains Triton X-100 (non-ionic detergent), then RIPA buffer which contains SDS and sodium deoxycholate (ionic detergents) and finally lysis directly in denaturation buffer containing β -mercaptoethanol (Figure 3.10). Despite this approach, specific FLAG signal corresponding to full-length DAPK-1 was not detected and using the less refined, yet more harsh, lysis method of nematode lysis in denaturation buffer, a lot of non-specific signal was detected. In addition, protein quantity was adjusted by lysing a greater number of nematodes. In some case whereby the protein concentration was estimated, approximately 30 μ g of protein was loaded into each well of the gel, however even at high protein quantities, no FLAG specific signal was observed. Protein extraction was successful when using these three lysis methods, as shown by intense specific signal when probing samples with anti- β -actin antibody. The protein detected here was most likely ACT-2, the *C. elegans* β -actin ortholog, which shares a high degree of sequence similarity to the immunogen used for antibody production (93% coverage, 86% identity).

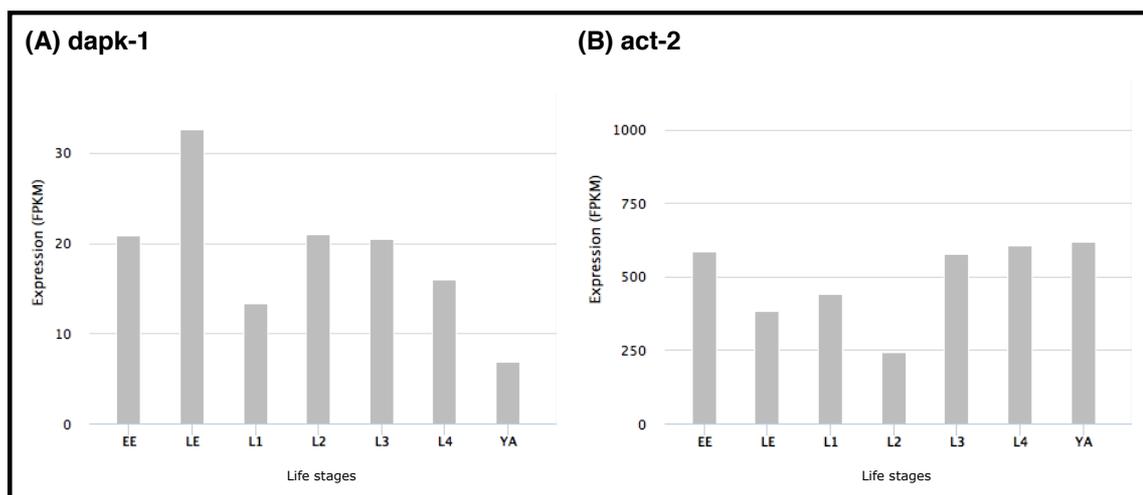


Figure 3.12 – *dapk-1* and *act-2* mRNA expression through the developmental stages of *C. elegans*. Median mRNA expression values from the PolyA+ and Ribozero modENCODE datasets accessed via WormBase. (A) *dapk-1* expression, (B) *act-2* expression. Life stage abbreviations: EE, early embryo; LE, late embryo; L1-L4, larval stages 1-4; YA, young adult.

Since specific signal corresponding to full-length FLAG-DAPK1 was undetectable *via* Western blot analysis, *dapk-1* transcripts were assessed by qPCR. This was to gather insight into first, whether the FLAG sequence was transcribed and second, whether *dapk-1* expression was affected by the genome editing. It was reassuring to identify that the FLAG sequence was being transcribed (Figure 3.11A), largely because a number of *dapk-1* splice variants lack exon 1 (Figure 3.11B), which is where the FLAG epitope is fused. However, splicing events may impact on the representation of FLAG-*dapk-1* in the *dapk-1* transcript pool (including all isoforms). Therefore, further work is required to decipher *dapk-1* RNA processing and to gather insight into the ratios of expression levels between *dapk-1* isoforms. Based on studies to map the splicing patterns in another ROCO gene of similar size and complexity, human *LRRK2*, this type of analysis is challenging, especially with low abundance targets [204].

An indication that *dapk-1* expression is reduced in the FLAG-*dapk-1* strain (Figures 3.11C and 3.11D) may have contributed to the challenge of detecting FLAG-DAPK-1 at the protein level. The expression of the already lowly expressed *dapk-1* may be suppressed due to the genomic modification. Alternative strategies to pursue *in vivo* DAPK-1 interactome studies include *dapk-1* transgenesis or stimulation of *dapk-1* expression, however these approach possess limitations with regard to non-physiological expression and therefore pursuing validation of the FLAG-*dapk-1* strain is favoured.

3.4.5 Conclusions

The analyses underlying this chapter provide a route for studying PPIs in a model organism whereby the current PPI landscape for a protein of interest is sparse and a clear ortholog has been identified. In this case, the analyses were centred on *C. elegans* DAPK-1. Predicted PPI networks were constructed based on orthologous inferences from reported PPIs in other species and this was coupled with reported *C. elegans* PPIs to assess the potential interaction profile of DAPK-1 in the wider cellular context. This was largely facilitated by the implementation of a *C. elegans* PPI data query option in PINOT. These hybrid networks, incorporating both predicted and evidence-based PPI data, were complemented by DAPK-1 PPI detection using the yeast two-hybrid approach which yielded six novel DAPK-1 interactors. In addition, a FLAG-*dapk-1* *C. elegans* strain was engineered for further DAPK-1 interactomic studies, however technical validation of this strain is ongoing.

Furthermore, this research provides the foundations for developments and future investigations in several directions. Improvements to the *C. elegans* PPI query aspect of PINOT are already

under consideration, in relation to implementation of the Alliance of Genome Resources API and incorporation of data captured by the PSICQUIC. At several stages in these analyses interactors were prioritised for further validation, which streamlines future 'wet-lab' studies. In addition, once validated, the FLAG-*dapk-1* strain represents a powerful tool for mapping the *C. elegans* DAPK-1 interaction profile. Overall, the resources developed and results obtained will guide and strengthen future DAPK-1 interactome and functional studies.

CHAPTER FOUR

Developing Novel Mutant *dapk-1 C. elegans* Models

4.1 Introduction

The *dapk-1* gene in *C. elegans* was identified from a genetic screen for mutant strains associated with epidermal morphogenesis defects, first reported in 2004 [205]. This gene was initially termed *mor-3*, a *C. elegans* nomenclature based on observed phenotype, whereby *mor* corresponds to 'morphological: rounded nose'. Based on *C. elegans* tissue gene expression profiling, *dapk-1* is expressed in the epidermal layer, specifically the hypodermis, and the body wall musculature [165,206,207]. Throughout the life stages of *C. elegans*, *dapk-1* gene expression increases during embryonic development and is most highly expressed during late embryogenesis (approximately 10 hours into embryonic development), then *dapk-1* is expressed at lower levels throughout larval development and adulthood [208] (Figure 3.12A). The pattern of expression *dapk-1* expression in aged nematodes requires further investigation, however insights from the data available suggests *dapk-1* expression is relatively low up to adult day 14 then increases within days 16-19 [203].

Since the identification and genetic mapping of *dapk-1*, a number of studies have assessed phenotypic characteristics of *C. elegans* upon genetic manipulation of *dapk-1* [164–167]. One approach has been to utilise *dapk-1* mutant strains (Table 4.1), two of which are available from the Caenorhabditis Genetics Centre (CGC; *gk219* and *ju4*). The *gk219* allele is a *dapk-1* gene knockout derived from a deletion in the 5' UTR region and exon 1 of the gene. This strain was generated and isolated by the *C. elegans* Knockout Consortium [209] as part of a large-scale chemical mutagenesis effort, using trimethylpsoralen (TMP) or ethyl methane-sulphonate (EMS), for creating gene knockouts across the *C. elegans* genome. The *ju4* allele also originated from an EMS based chemical mutagenesis screen, whereby mutants were assessed in relation to defects in epidermal morphogenesis [165]. In addition, hundreds of other mutant *dapk-1* strains have been generated, as part of the million mutation project [210], but not characterised.

Of the mutant *dapk-1* strains with reported phenotypes, the *ju4* allele which harbours a substitution mutation in the kinase domain (S179L), displays the most prominent phenotype. Of note, this residue is not conserved in relation to human DAPK1 (Figure 4.1B). This mutation results in a highly penetrant cuticle morphology defect evident in multiple body areas (head, tail, vulva, dorsal midline) from L3 larval development onwards, with progressive severity [165]. This defect is characterised by a thickened cuticle and epidermal degeneration which triggers innate immune responses in the nematode. The penetrance of this phenotype in other mutant *dapk-1* strains is considerably lower (Table 4.1) and is temperature sensitive, with a higher incidence within populations maintained at 25°C in comparison to 15°C, across all four listed

mutant alleles [165]. The molecular underpinnings of this epidermal homeostasis dysregulation are largely unknown.

Table 4.1 – Mutant *dapk-1* *C. elegans* strains with reported cuticle defect phenotype. Penetrance of this phenotype is temperature sensitive, percentages reported correspond to maintaining the nematodes at 20°C [165]

| Mutant | Mutation | Impact on protein sequence | Cuticle defect penetrance |
|--------------------------------|--|--|---------------------------|
| <i>dapk-1</i> (<i>gk219</i>) | 612bp deletion 5' UTR and most of exon 1 Gene knockout | Not translated | 19% |
| <i>dapk-1</i> (<i>ju4</i>) | C > T substitution Position 1308062 | S179L in kinase domain | 100% |
| <i>dapk-1</i> (<i>ju469</i>) | 25bp deletion Exon 9 – intron 9 junction | Predicted truncation after residue 862 (in ROC-COR) | 16% |
| <i>dapk-1</i> (<i>ju557</i>) | C > T substitution Position 1321809 | Q691stop in ROC domain | 33% |

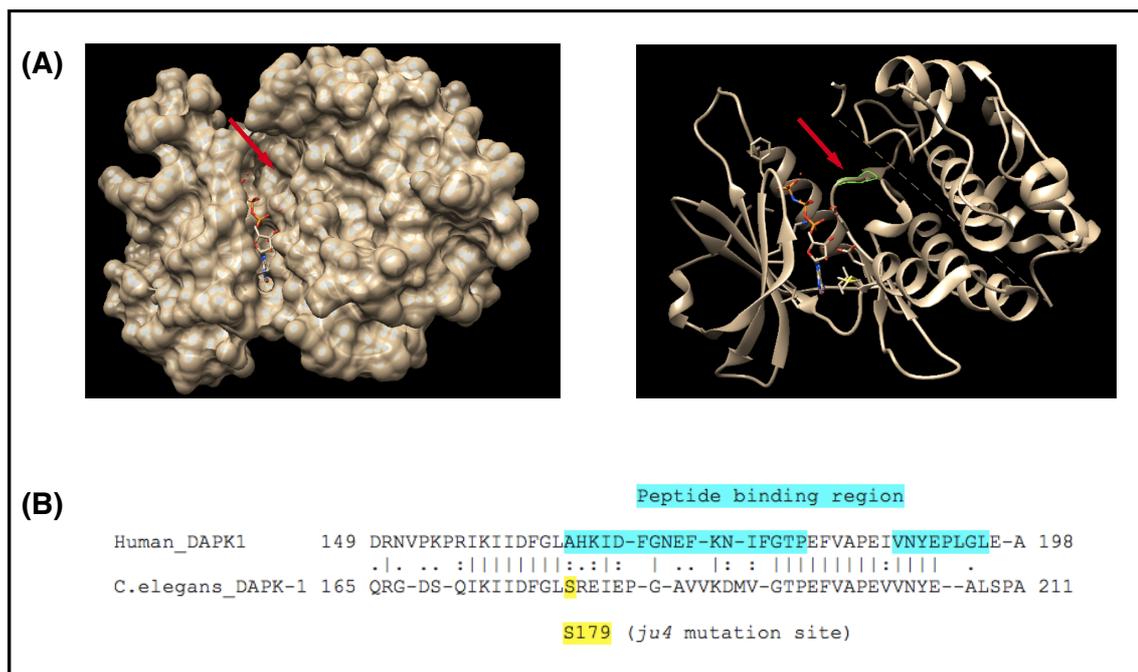


Figure 4.1 – Structural inference of *ju4* S179L mutation site based on human DAPK1 kinase structure. (A) Crystal structure of human DAPK1 kinase domain with ATP analogue, at a 1.62Å resolution. Protein Data Bank ID: 1JKL, [24]. S179 aligned residue A165 (in human structure) highlighted by red arrow. (B) Human DAPK1 *C. elegans* DAPK-1 sequence alignment at the *ju4* S179L locus, with peptide binding region shown in blue (human sequence).

In addition, transgenic approaches to exogenously express *dapk-1*, wild-type and mutant forms, in *C. elegans* have also been adopted to assess the role of DAPK-1. Transgenic expression using green fluorescent protein (GFP) under the transcriptional control of the *dapk-1* promoter, has been a method used to pinpoint *dapk-1* expression within the epidermis *in vivo* [165]. Mutant *dapk-1* lacking specific domains or harbouring point mutations, have been expressed in *C. elegans* as transgenes to assess phenotypic consequences on the wild-type and mutant *dapk-1* genetic backgrounds [166]. The only construct variant which induced the cuticle morphology phenotype on the wild-type background was the *ju4* allele, although exogenous expression of the constructs lacking the kinase domain did enhance this phenotype on the *ju469* allele genetic background. Exogenous expression of full-length *dapk-1* rescued this phenotype on the *ju4* allele genetic background. Furthermore, nematode lethality was evident when exogenously expressing the *ju4* allele and constructs lacking the kinase domain on the *ju4* allele background.

Another study using transgenesis of *dapk-1*, but in the context of excitotoxicity, reported that *dapk-1* overexpression, by incorporating a heat shock protein (*hsp-16*) promoter upstream of the gene, enhanced neuronal toxicity in an established nematode excitotoxicity model strain (*glt-3;nuls5*). In contrast, crossing the *dapk-1* knockout allele (*gk219*) into this strain had the opposite effect, neuronal toxicity was suppressed [167]. This research therefore positions DAPK-1 as a regulator of excitotoxic mechanisms in *C. elegans*.

A further genetic tool which has been used to study the role of *dapk-1* is targeted RNAi to knockdown the expression of the *dapk-1* transcript. A study investigating the links between autophagy regulation and survival in *C. elegans* fed *dapk-1* RNAi to a mutant *gpb-2* strain which is hypersensitive to starvation stress and autophagy hyperactive [164]. As a result, starvation-induced autophagy was reduced, using survival as a readout. Upon generating a double mutant *gpb-2 dapk-1* strain, which harboured the *gk219 dapk-1* knockout allele, starvation-induced autophagy was again reduced, using LGG-1 as a marker for autophagy. This study suggested a role for DAPK-1 in muscarinic signalling in relation to autophagy regulation [164].

Despite a range of genetic tools which have been utilised to manipulate *dapk-1* for gathering insight into the function of the translated protein, the current understanding of how DAPK-1 functions as a molecule and how it influences signalling in the wider cellular context remains unclear. With advances in precision genome editing, such as the CRISPR-Cas9 approach, generating targeted point mutations in the *C. elegans* genome have become an invaluable tool for assessing protein function. The key benefits of this approach are that modifications are carried out with a high degree of precision and at the endogenous locus [211], hence the

mutated gene is under endogenous transcriptional control. In addition, this approach had not been performed in the context of *dapk-1* gene manipulation on the wild-type genetic background. With this in mind, this was the strategy adopted in this project.

4.1.1 Project aims

The goal of the research outlined in this chapter was to develop novel *C. elegans* models for understanding DAPK-1 function. These models would enhance the mutant *dapk-1* strain toolbox due to the generation of novel mutant nematodes whereby key residues in relation to the proposed DAPK-1 catalytic activities were targeted. Furthermore, these models would provide a foundation for *in vivo* validation and signalling-related characterisation of predicted DAPK-1 interactors and Y2H hits presented in the previous chapter. The initial aim was to generate and begin to characterise these novel mutant *dapk-1* strains.

To achieve this, mutagenesis would be carried out by a specialist *C. elegans* genome editing service provider. Two residues were selected for CRISPR-Cas9 mutagenesis, K57W and T715N, which were hypothesised to result in kinase inactivity and GTP binding deficiency, respectively, based on sequence conservation and biochemical analysis of human DAPK1 and other human ROCO proteins [30,34,56,212,213]. Therefore, my first aim was to validate the mutant genotypes generated from these mutagenesis events. Second, I aimed to outcross these newly created strains against the wild-type genetic background in order to reduce the likelihood of genetic variation elsewhere in the genome (not at the desired locus), this may arise due to potential off-target effects of the CRISPR-Cas9 mutagenesis process.

My third aim in relation to these novel mutant *dapk-1* *C. elegans* strains was to gather preliminary phenotypic data. In the first instance, in relation to known mutant *dapk-1* phenotypes, i.e. the head morphology abnormality. Then, to assess general features of *C. elegans* fitness, such as determining lifespan profiles and performing progeny counts for the mutant strains. These initial phenotyping analyses would provide an indication of the impact a kinase inactive form and a GTP binding deficient form of DAPK-1 has on the physiology of *C. elegans in vivo*.

4.2 Materials and Methods

The maintenance of *C. elegans* for the research outlined in this chapter was as previously described in Chapter Three – 3.2.6.1 *C. elegans* maintenance. N2, *dapk-1 (gk219)* and *dapk-1 (ju4)* strains were obtained from CGC.

4.2.1 Generating novel *dapk-1* mutant strains

In addition to the mutant *dapk-1* strains available from CGC (*gk219* and *ju4*), two novel mutations were introduced into *dapk-1* by CRISPR-Cas9 mutagenesis. The modified genomic sequences are presented below (4.2.1.1 – Genomic modifications resulting in *dapk-1* K57W and T715N mutant strains). At the protein level, these mutations correspond to K57W and T715N, hypothesised to result in kinase inactivity and GTP binding deficiency, respectively. The generation of these mutant strains was outsourced to Knudra Transgenics (USA). Similarly to the genome engineering outsourced to SunyBiotech, the service provided by Knudra Transgenics covered sgRNA design, plasmid construction, microinjection, screening and validation of the homozygous mutant nematodes. Two independent microinjections were performed for each desired mutant strain.

4.2.1.1 Genomic modifications of the *dapk-1* K57W and T715N mutant strains

Genomic sequence at the locus (exon 2 [K57W] and exon 9 [T715N]) of the novel *dapk-1* mutant strains (*knu483 K57W*, *knu484 K57W*, *knu450 T715N*, *knu458 T715N*) is listed below. Modified sequence is highlighted: syntron (blue); recoded exon (yellow); codon corresponding to mutation (pink). Exonic and intronic sequence is represented by uppercase and lowercase, respectively.

dapk-1 (knu483 and knu484 K57W):

```
ctaaaatgttgctctttgaatcaacaccacaccttaaacgcagtcgaaaattggtaattttacagCGGCCA
GTTTGCTGTTGTTTCGTCGTGTTTCGCGACCGCAAGACTGGAGAGAAGTACGCTGCTTGGTTCATTAAGAAA
CGCCGCTATGCTACCTCCCGCCGCGGTGTTACCCGCCAAAACATTGAGCGCGAGGTCCGTGTTCTTCAAA
AAATCCGCGGTAACCTCAATGTGGTGGAGCTTCATGCAGTTTATGAAA
```

dapk-1 (*knu450* and *knu458 T715N*):

TCTCCCAACTCTACCCACTGGACACCTCTC **TTCGTCGTATCAAACCTAAACTTCTTGGACACTCTCAGTC**
CGGCAAGAATCGTCTTGTACAGACACTTCACTCGTCACGTGGAA

4.2.2 Validation of *dapk-1* mutant strains

4.2.2.1 DNA extraction

DNA extraction and PCR was performed on each mutant *dapk-1* strain to validate the presence of the mutation at the genomic level. As previously described, DNA was crudely extracted from *C. elegans* by picking 10 adult (day 1) nematodes into 20µl worm lysis buffer 1 (see '3.2.1 general laboratory materials and media/buffer compositions' for buffer composition) supplemented with 1mg/ml proteinase K, per strain. N2 nematodes were processed alongside mutant strains to act as a control for wild-type *dapk-1* and extractions were carried out with a minimum of two replicates (i.e. a minimum of 20 nematodes sampled). Samples were lysed and DNA extracted by incubation at 65°C for 1 hour followed by 95°C for 15 minutes.

4.2.2.2 PCR and restriction digestion

Genotyping was carried out using the cleaved amplified polymorphic sequences (CAPS) approach, whereby PCR is performed to amplify a region of genomic DNA in the proximity of a mutation site followed by characterisation of the amplicons by restriction digestion and gel electrophoresis. The PCR and restriction digestion experimental designs for this genotyping are illustrated in Figure 4.2. Primer sequences are listed in Table 4.2. PCRs were performed in a total volume of 15µl: 3µl 5x OneTaq reaction buffer, 0.3µl 10mM dNTPs, 0.3µl 10µM primer mix (forward and reverse), 0.1µl OneTaq Hot Start polymerase, 3µl DNA extract, 8.3µl nuclease-free water. PCR reagents sourced from New England Biolabs (NEB). The thermocycler parameters for PCR are outlined in Table 4.3. Samples were incubated in a BioRad T100 Thermal Cycler.

PCR products derived from *ju4*, T715N and K57W reactions were subjected to restriction digestion, see Table 4.4 for the corresponding restriction endonuclease and reaction buffer. Digest reactions were performed in a total volume of 20µl: 0.3µl restriction endonuclease, 2µl 10x corresponding reaction buffer, 7µl PCR amplified product, 10.7µl nuclease-free water. For digestion, samples were incubated at 37°C for 2 hours. DNA fragment size was visualised by 2%

(w/v) agarose gel electrophoresis (80V for 1 hour) in TAE buffer, using 0.5x SYBR safe stain for DNA detection and a 100bp or 2-log DNA ladder. Gels were imaged on a U:Genius 3 (Syngene).

Table 4.2 – Primer sequences for mutant *dapk-1* validation by PCR. Primers supplied by Eurofins.

| Primer | Primer sequence (5' to 3') |
|------------------------|----------------------------|
| <i>gk219</i> forward 1 | CTGAGAATTTTGAAAACCCG |
| <i>gk219</i> forward 2 | CTAGACTCATTTCCTCCC |
| <i>gk219</i> reverse 1 | CTTTGTCTCTGCTTCCTTGC |
| <i>ju4</i> forward | TCGACCACGTCTGTGCAAAGAATGCC |
| <i>ju4</i> reverse | CCACTGCTCCCGGCTCAATTTCTCTA |
| T715N mutant forward | CTAAACTGACTCACATGG |
| T715N mutant reverse | GTAGTAACTAAAAGAGCCG |
| K57W mutant forward 1 | CCTAGTACAAAATCCACGG |
| K57W mutant reverse 1 | CACCATGTTCCACTTTTCTC |
| K57W mutant forward 2 | TTCAGGTATTCTGGTTCCG |
| K57W mutant reverse 2 | CTAACGTAGATACGAGCTGC |

Table 4.3 – PCR thermocycler programme for *gk219*, *ju4*, T715N and K57W genotyping

| Temperature | Time |
|-------------|----------|
| 94°C | 30 secs |
| 94°C | 30 secs* |
| ** | 30 secs* |
| 68°C | 1 min* |
| 68°C | 5 mins |

*these three steps were cycled 40 times

** annealing temperature varied depending on primer combination: *gk219*, 49°C; *ju4*, 62°C; T715N, 48°C; K57W primer set 1, 50-55°C; K57W primer set 2, 50°C.

Table 4.4 – Restriction endonucleases used for genotype validation. All sourced from New England Biolabs (NEB). Cleavage site annotations, Y corresponds to C or T, R corresponds to A or G.

| Enzyme | Cleavage site | Incubation temperature | Reaction buffer | Template validation |
|--------------|--|------------------------|-----------------|---------------------|
| <i>Xba</i> I | 5'... T [▼] CTAGA... 3' 3'... AGATC [▲] T... 5' | 37°C | CutSmart | <i>ju4</i> |
| <i>Ava</i> I | 5'... C [▼] YCGRG... 3' 3'... GRGCY [▲] C... 5' | 37°C | CutSmart | T715N |
| <i>Nru</i> I | 5'... TCG [▼] CGA... 3' 3'... AGC [▲] GCT... 5' | 37°C | NEB 3.1 | K57W |

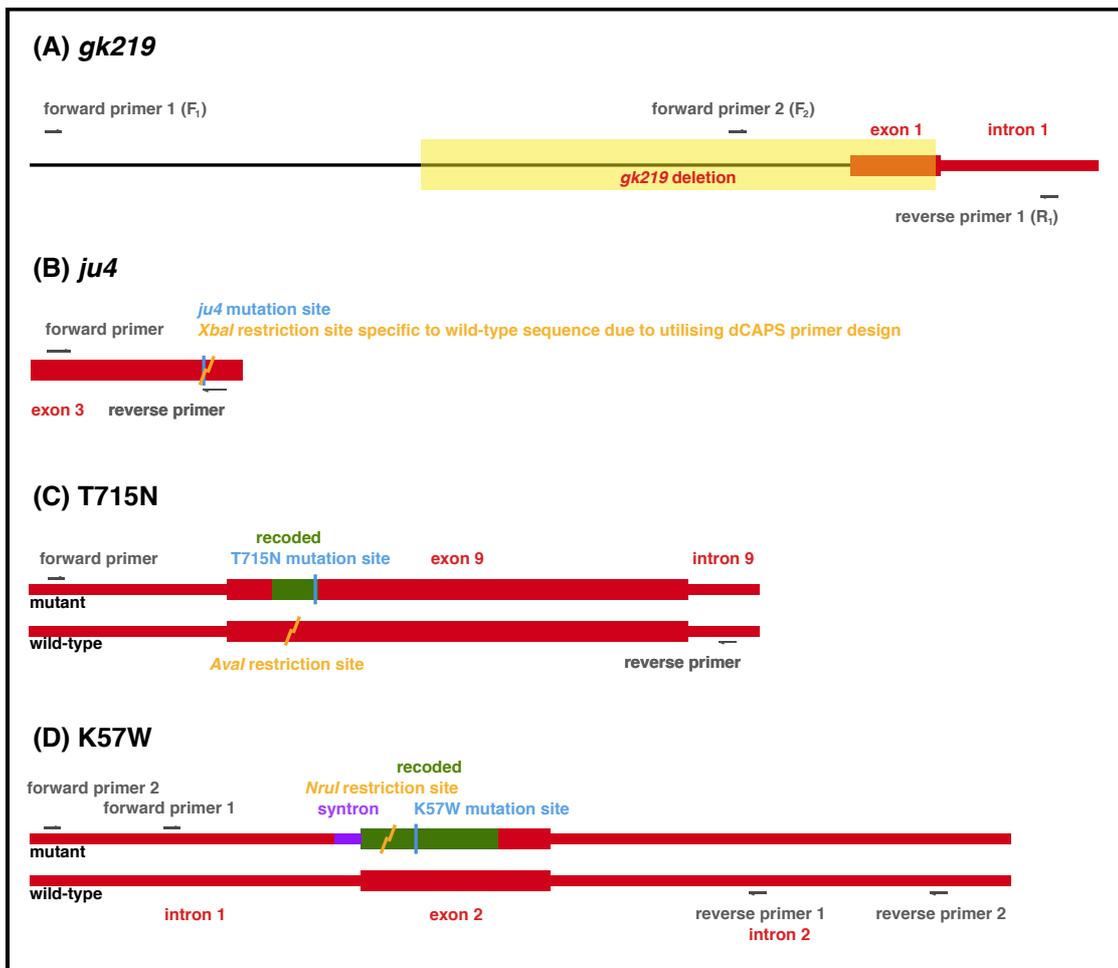


Figure 4.2 – Mutant *dapk-1* genotyping experimental design using PCR and restriction digestion. (A) Three primer reaction to validate the 5' UTR deletion in the *gk219* strain. (B) Derived cleavage amplified polymorphic sequence (dCAPS) primer design to incorporate *XbaI* restriction site specific to N2 strain. (C) Primer design in relation to mutant T715N and N2 genomic templates, and *AvaI* cleavage site which is specific to the wild-type amplicon. (D) Two primer pairs designed to amplify the genomic region containing the K57W mutation site with *NruI* restriction site annotated which is specific to the mutant amplicon.

4.2.3 Outcrossing novel *dapk-1* mutant strains

Once validated the novel mutant *dapk-1* strains were outcrossed back onto the wild-type genetic background to reduce the likelihood of off-target genetic modifications from the genome editing persisting in the genome, whilst retaining the desired modification. This outcrossing was performed with homozygous mutant hermaphrodite nematodes and N2 male nematodes. Since the frequency of male nematodes in *C. elegans* N2 populations is less than 0.2% [100], the generation of male nematodes was induced to obtain a sufficient number of males for establishing genetic crosses.

4.2.3.1 Generating male *C. elegans*

Non-disjunction of the X chromosome during meiosis which results in the generation of male *C. elegans* was achieved by heat stress. L4 N2 hermaphrodite nematodes were picked onto NGM plates, 6 plates in total, 10 nematodes per plate. These plates were incubated at 30°C for 5 hours, 5.5 hours and 6 hours, two plates per incubation time, then returned to 20°C. The progeny population of these L4 nematodes contained an increased proportion of males (approximately 5%), which were identifiable once this generation of nematodes were approximately L4 stage. Following the generation of males by heat shock, populations of N2 male nematodes were maintained by picking 2 L4 N2 hermaphrodites and 10 adult N2 males onto an NGM plate which would result in predominately cross-fertilisation and a 30-50% proportion of males in the progeny population.

4.2.3.2 Outcrossing

Mutant *dapk-1* strain nematodes were outcrossed by picking 2 L4 mutant hermaphrodites and 10 adult N2 males onto an NGM plate. The progeny from this cross (F1 generation), derived from cross-fertilisation, were heterozygous for the mutant allele. Two L4 heterozygous mutant hermaphrodites and 10 adult N2 males were then picked onto an NGM plate, for further outcrossing. The progeny population from this second cross (F2 generation) were singled onto individual NGM plates, incubated under standard condition until embryos were laid, then lysed for genotyping. Progeny derived from self-fertilisation of heterozygote hermaphrodites would result in a proportion of homozygous mutant nematodes (25% in theory). These homozygotes and subsequent generations derived from self-fertilisation were considered 2 x outcrossed. This process was repeated to achieve a 4 x outcrossed *dapk-1* T715N strain.

The genotype of the nematodes, in relation to the *dapk-1* mutation, was traced by PCR and restriction digestion as previously described in '4.2.2 Validation of *dapk-1* mutant strains' with the modification that single worms were lysed in 10µl WLB1 supplemented with 1mg/ml proteinase K, as oppose to 20µl.

4.2.4 Phenotyping *dapk-1* mutant strains

All phenotyping was performed on hermaphrodite nematodes. The *dapk-1* T715N mutant strain (*dapk-1 knu458 T715N*) used for phenotype analysis was 4x outcrossed

4.2.4.1 Scoring head morphology defects

The presence of head morphology defects in *C. elegans* was scored using a stereo microscope, observing adult day 1 nematodes on NGM plates. Representative images of the head morphology defect phenotype were captured by immobilising nematodes in 25mM sodium azide on 2% (w/v) agarose pads set on glass microscope slides. These nematodes were imaged using a AMG EVOS FL imaging system with a 4x objective lens.

4.2.4.2 Lifespan assay

The impact of *dapk-1* mutant genotypes on lifespan was assessed by monitoring nematodes daily, under the stereo microscope, to identify the number of nematodes that died on each day of the lifespan time course. Approximately 120 L4 nematodes were picked onto NGM plates supplemented with 50µM FUDR on day 0 (three or four plates per genotype, each containing 30-35 nematodes, were setup in parallel) and maintained under standard conditions. At 24-hour intervals, each plate was monitored to identify dead nematodes. If the nematodes were not obviously alive (e.g. mobile or visible pharynx pumping), the plate was lightly tapped on the bench and if still indistinguishably alive, the nematode was lightly touched with a platinum wire. If the nematode was unresponsive to these stimuli, it was reported as dead. Dead nematodes were removed from the assay plates and discarded.

Nematodes which were lost (e.g. dried up on the side of plates) or had died as a result of age-related vulval integrity defects (AVID; i.e. vulval rupturing which leads to premature death [214]) were censored from the survival plots. This assay was repeated (as previously described) with the addition of 100µg/ml ampicillin supplemented into the NGM, which was seeded with 10x concentrated heat-killed OP50. OP50 was heat-killed by incubation at 75°C for 90 minutes. To verify the bacteria was heat-killed, a sample of the culture was streaked on an LA plate and

incubated at 37°C overnight. This addition of ampicillin was largely implemented to reduce foreign bacterial contamination.

4.2.4.3 Progeny counts

The production and viability of progeny derived from self-fertilisation in *dapk-1* mutant hermaphrodite *C. elegans* was evaluated by counting the number of nematodes which survived post-embryonic development for individual parent nematodes. Single L4 nematodes were picked onto NGM plates (1 per plate) and maintained under standard conditions. The nematodes were monitored the following day to determine when the first embryos were laid. From that point, the adult nematode on each plate (not progeny nematodes) was transferred onto a fresh NGM plate every 24 hours for 5 days. At day 5, the adult nematodes were maintained for a further 2 days to ensure progeny production had ceased. Once the progeny had developed to L3/L4 stage, the nematodes on each plate were counted. These counts corresponded to a 24 hour period of embryo laying.

4.3 Results

4.3.1 Validation of mutant *dapk-1* *C. elegans* strains

For assessing *C. elegans* phenotypes in the context of gathering insight into the potential role of DAPK-1, several mutant *dapk-1* strains were obtained. Two previously established mutant strains were purchased, the *dapk-1* (*gk219*) gene knockout strain and the *dapk-1* (*ju4*) S179L point mutation variant. In addition, novel mutant *dapk-1* strains were engineered by Knudra Transgenics (USA), the *dapk-1* (*knu483* K57W; *knu484* K57W) and *dapk-1* (*knu450* T715N; *knu458* T715N) strains, designed based on the hypothesis that these mutants would result in kinase inactivity and GTP binding deficiency in DAPK-1, respectively. The first task once obtaining these strains was to validate their genotype in relation to the *dapk-1* gene. This was to ensure results from downstream experimentation with these nematodes could be interpreted in the context of validated mutant *dapk-1* genotypes.

4.3.1.1 *dapk-1 gk219, ju4, knu450 T715N* and *knu458 T715N* genotype validation

Genotype validation was performed using PCR targeting the mutation site, followed by restriction digestion (in some cases) and agarose gel electrophoresis to determine amplified (and digested) DNA fragments. See Figure 4.2 for a visual representation of the experimental design with the expected DNA fragment sizes following each series of reactions.

The PCR for validating the *dapk-1* (*gk219*) strain (*dapk-1* deletion) contained two forward primers, one which targeted the 5' UTR of *dapk-1* approximately 940bp upstream of the gene (forward 1) and one which targeted the 5' UTR which was reported to be deleted in the *dapk-1* (*gk219*) strain (forward 2). Since shorter amplicons are preferentially enriched in PCR due to the amplification taking less time, plus the extension period in the PCR thermocycles limits the length of an amplicon which can be produced, amplification primed by forward primer 1 and the reverse primer on the N2 genetic background would not result in a product, due to the size of the potential amplicon. Hence, on the N2 template background, the expected amplicon was 392bp, primed by forward primer 2 and the reverse primer. Whereas, on the *dapk-1* (*gk219*) genetic background, which lacks the region whereby primer 2 anneals to the genomic template, the expected amplicon was 591bp, primed by forward primer 1 and the reverse primer. These expected results match the results obtained across duplicate samples of each strain (Figure 4.3A). Therefore, these results validate the genomic deletion within the 5' UTR and exon 1 of *dapk-1* within the *gk219* strain.

The PCR experimental design for validating the *dapk-1 (ju4)* genotype utilised the derived cleavage amplified polymorphic sequence (dCAPS) primer strategy [215]. This approach was used since the *ju4*-associated mutation did not create or disrupt a restriction site with the genomic sequence. Therefore, using the dCAPS method, a mismatch in the reverse primer sequence created an *XbaI* restriction site specific to the N2 genomic template. Upon PCR and restriction digestion with *XbaI*, the expected result was the production of a 210bp amplicon, which was cleaved into 23bp and 187bp fragments on the N2 genetic background only. This result was not observed when performing the PCR and restriction digestion (Figure 4.3B). Amplicons of the expected size were evident (210bp), however fragments corresponding to the expected *XbaI* cleavage were not observed. These reactions were repeated and products separated in a high concentration of agarose (3% w/v as oppose to 2% w/v) in an attempt to obtain better separation between the expected 210bp and 187bp bands, however the same result was obtained. These results indicate that the expected cleavage did not occur.

Unlike the *dapk-1 (gk219)* and *dapk-1 (ju4)* strains, the *dapk-1 (knu450 T715N)* and *dapk-1 (knu458 T715N)* strains were engineered as part of this project and therefore it was especially critical that this genotype was validated prior to further experimentation. Knudra Transgenics performed verification of the mutation by sequencing, as part of the service provided. However, robustly performing in-house validation of the genotype was important for tracking the mutation in later genetic outcross experiments.

As part of the mutagenesis process, performed by Knudra Transgenics, a 55bp region which included the mutation site was recoded (see '4.2.1.1 Genomic modifications of the *dapk-1* K57W and T715N mutant strains'), however only the T715 residue was altered at the protein level. Within this region on the N2 genetic background there is an *AvaI* restriction site, however this was disrupted upon recoding this region in the T715N strains. Therefore, a 810bp region was amplified at this locus and the *AvaI* restriction site was utilised for genotyping. Upon restriction digestion with this nuclease it was expected that the N2 amplicons would be cleaved into 290bp and 520bp fragments, whereas T715N strain derived amplicons would not be cleaved. This expected result aligned with the result obtained for both of the independently generated T715N strains, *knu450 T715N* and *knu458 T715N* (Figure 4.3C). Hence, this validated the T715N strain genotype.

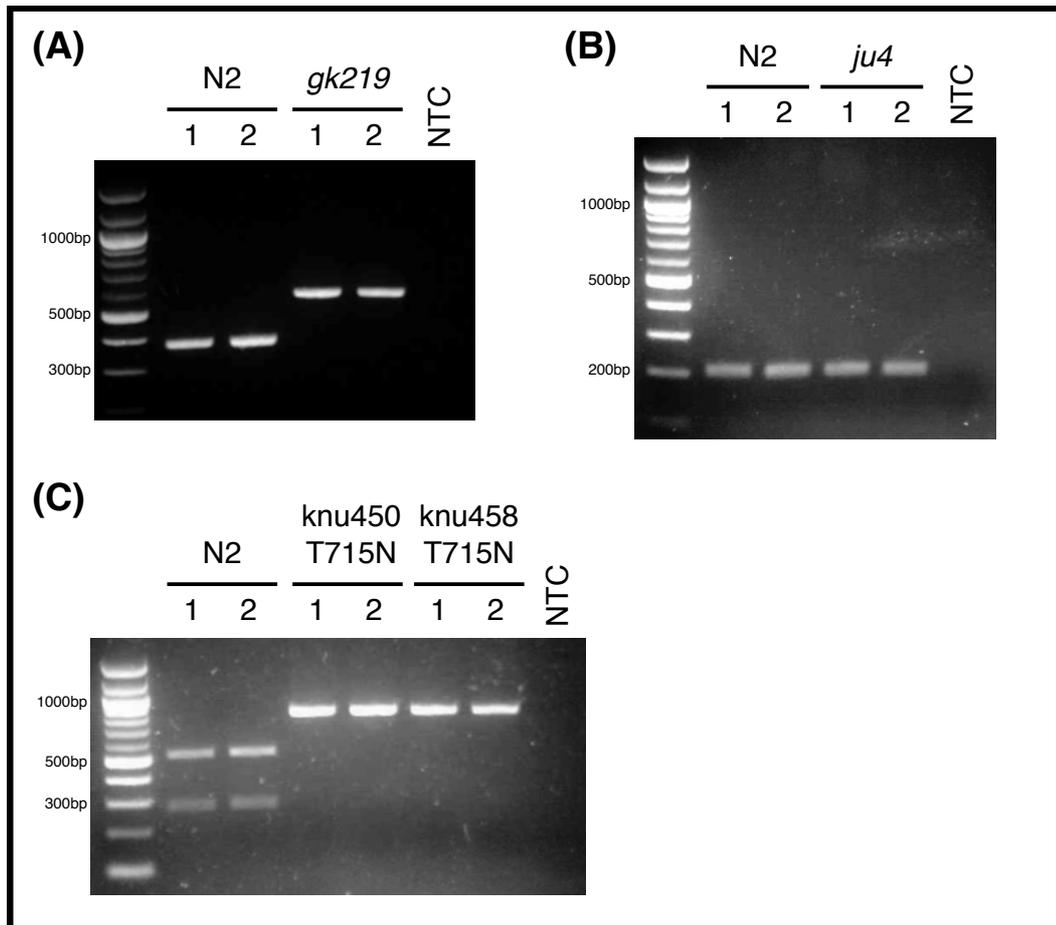


Figure 4.3 – *dapk-1* mutant genotype validation for *gk219*, *ju4* and *T715N* strains. PCR of genomic sequence in the proximity of the *dapk-1* mutation site to verify the mutant genotype. (A) Agarose gel image of PCR products amplified from a 3 primer combination reaction targeting the 5' UTR of *dapk-1*, which is deleted within the *dapk-1* (*gk219*) mutant strain. (B) Agarose gel image of DNA fragments derived from PCR of a region flanking the mutation site within the *dapk-1* (*ju4*) mutant strain followed by restriction digestion by *Xba*I. (C) Agarose gel image of PCR products subjected to restriction digestion with *Ava*I, following amplification of a genomic region in the proximity of the *T715N* mutation.

4.3.1.2 Optimising K57W genotype validation

Validation of the *dapk-1* K57W genotype by PCR and restriction digestion proved to be challenging. Knudra Transgenics had verified the desired mutation by sequencing a 435bp region which encompassed the mutation site, however upon attempting to amplify a wider genomic region at this locus for in-house validation, no PCR product was observed from either of the K57W mutant strains following agarose gel electrophoresis. The verification data provided by Knudra Transgenics indicated the insertion of a 34bp syntron (synthetic intron) in substitution of a 23bp intronic region, followed by a 164bp recoded sequence which corresponds to exon 2 and includes the mutation site (Figure 4.2D; and see '4.2.1.1 Genomic modifications of the *dapk-1* K57W and T715N mutant strains'). As mentioned regarding the T715N mutagenesis, this recoded region only altered the desired residue (K57) at the protein level.

The initial PCRs targeting this locus were designed to amplify a 690bp region (in relation to the K57W genotype which included the syntron insertion) and these reactions were first performed with an annealing temperature of 55°C within the PCR cycle. The expected PCR product size was observed on the N2 and *dapk-1* T715N template backgrounds, but the lack of signal evident in lanes corresponding to both independently derived *dapk-1* K57W (*knu483* and *knu484*) strains suggested no amplification occurred on these template backgrounds. These PCRs were repeated, with the incorporation of lowered annealing temperatures in the PCR cycle (55°C, 53°C, 50°C) to assess whether lowering the annealing temperature, and therefore the primer binding stringency, would result in the expected amplification on the *dapk-1* K57W template background. This optimisation strategy did not clearly result in the expected amplicon size for *dapk-1* K57W derived samples, although weak signals corresponding to fragments of length 500-600bp were evident in the *knu483* K57W samples when performing the PCR with an annealing temperature of 50°C and an even weaker signal was present at the same (50°C) annealing temperature for sample *knu484* K57W 2 at the expected fragment size (Figure 4.4A). The expected amplicon size was observed for N2 and *dapk-1* T715N derived samples, across the three annealing temperature trialled.

Further optimisation of this PCR was attempted with regard to the MgCl₂ concentration within the reaction mix. The reaction buffer supplied with the OneTaq Hot Start polymerase (NEB) used contains 1.8mM MgCl₂ once diluted to the recommended 1x concentration. This is likely to be the optimal MgCl₂ concentration for the efficiency of this specific polymerase. However, increasing MgCl₂ concentration in PCRs tends to increase the activity of the polymerase and therefore increase the amplicon yield, although this is at the expense of the enzyme fidelity

[216]. Since fidelity was not paramount in this PCR application and polymerase activity could potentially be enhanced, these PCRs were repeated with further MgCl₂ supplementation: 1.8mM, 2.5mM and 3.2mM final concentrations. At 2.5mM MgCl₂ signal of an amplicon corresponding to a fragment slightly larger than expected was evident in the *knu483* K57W 1 sample, this was paired with an off-target amplicon of around 500bp which was also observed in one of the T715N samples. At the higher MgCl₂ concentration (3.2mM), the expected signal was lost in the N2 samples and reproducible signal in K57W derived samples remained absent (Figure 4.4B).

Although the primers tested for this locus performed as expected on the N2 and *dapk-1* T715N template backgrounds, the lack of expected amplification on the K57W template background led to a third approach. This was to design a new primer pair for amplifying this genomic region. The expected amplicon size with this new primer pair, forward 2 and reverse 2, was 1044bp and 1033bp on the K57W and N2 template backgrounds, respectively. In addition, upon restriction digestion with *Nru*I, the 1044bp fragment would be cleaved into 660bp and 384bp fragments, whereas cleavage was not expected to occur in the N2 amplicons. Weak signal was evident in the *knu484* K57W sample at approximately 660bp which would indicate expected cleavage and the corresponding PCR product was subsequently run on an agarose gel which showed the expected 1044bp size (Figure 4.4C). However, this observation is coupled with signal in the no template control (NTC) lane of this experiment which indicates nucleic acid contamination and therefore overall, this result cannot be interpreted as evidence in support of validating the K57W genotype.

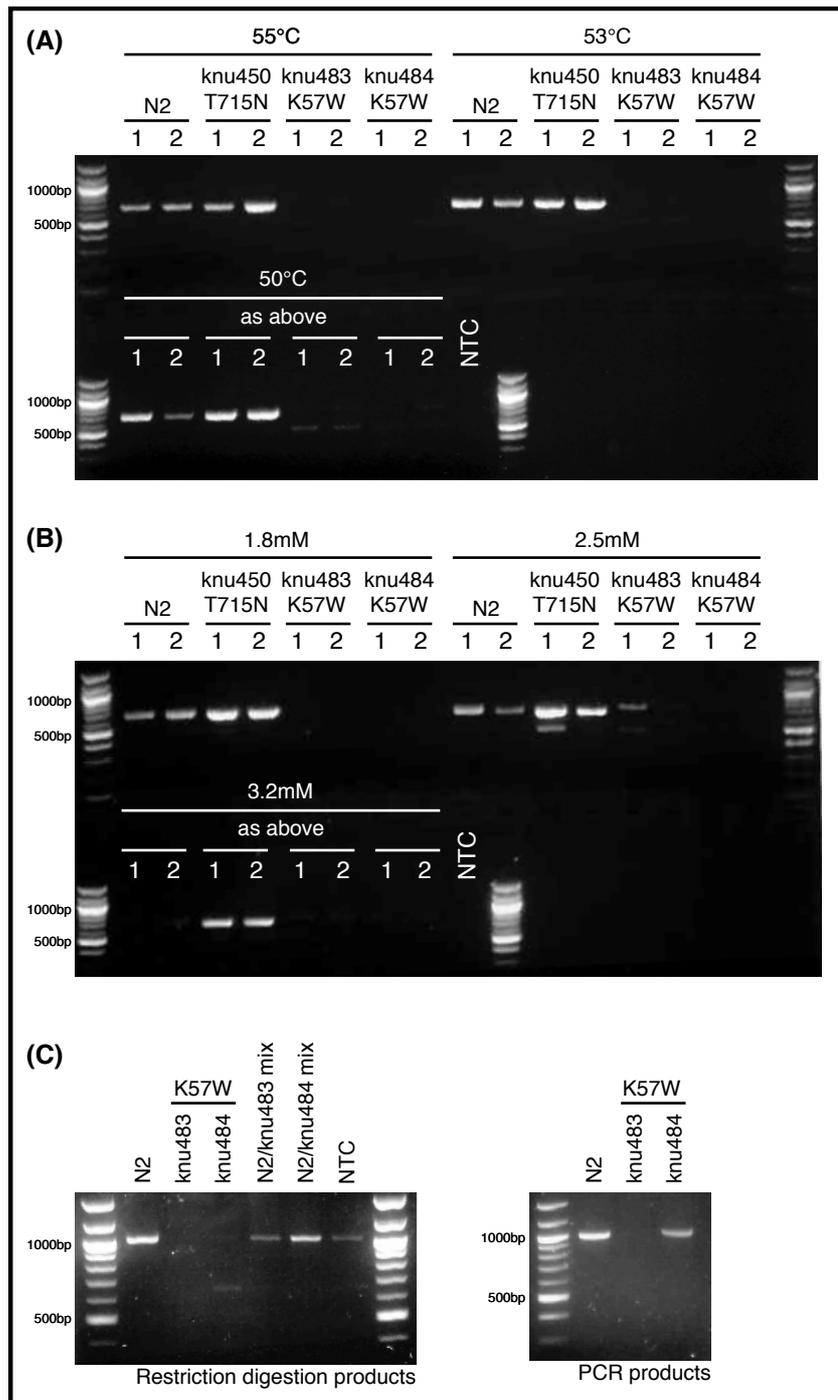


Figure 4.4 – Optimising mutation validation of *dapK-1* K57W strain. (A) Impact of decreasing the annealing temperature within the PCR cycle on amplified fragment production using forward 1 (F_1) and reverse 1 (R_1) primer combination. (B) Effect of adjusting the $MgCl_2$ concentration within the reaction mix on PCR amplification using F_1 and R_1 primer combination. Numbers correspond to duplicate sample preparations. (C) New primer combination (forward 2 and reverse 2) targeting a wider region around the K57W mutation site, left image: fragments following restriction digestion with *Nru*I; right image: PCR products corresponding to DNA fragments used in the restriction digestion. NTC, no template control.

4.3.2 Outcrossing the *dapk-1 (knu458 T715N)* strain

Following successful genotype validation of the *dapk-1* T715N strains, the *dapk-1 (knu458 T715N)* strain of this genotype was outcrossed onto the N2 genetic background to reduce any background genetic variation within the genome which may impact future phenotypic analysis. The crosses were setup to enable cross-fertilisation from male N2 nematodes to hermaphrodite mutant nematodes and the genotype of the resulting progeny was traced using PCR and restriction digestion. Figure 4.5 provides a representation of how the genotypes are traced, these gel images are derived from the 2x outcrossed nematodes. As previously described, the expected fragment size indicating the mutant allele which does not undergo cleavage by *AvaI* is 810bp whilst the N2 allele is cleaved to produce 290bp and 520bp fragments.

As shown in Figure 4.5A, a number of parent nematodes from the 2x outcrossed population were individually genotyped and this resulted in the identification of homozygotes and heterozygotes in relation to the mutant allele. Of note, the nematodes which were homozygous for the mutant allele at this stage (e.g. nematode 4 and nematode 12) suggests that self-fertilising hermaphrodite nematodes persisted through these two generations of the crossing process and therefore these nematodes were not of outcrossed origin. The genotypes of interest here are the heterozygotes (e.g. nematode 8 and nematode 10) since this genotype indicates cross-fertilisation has occurred. When these heterozygotes produce progeny by self-fertilisation, a subset (25% in theory) would be homozygous for the mutant allele. These nematodes were identified by genotyping the progeny of nematode 8 and nematode 10 (Figure 4.5B), since the parent nematodes laid embryos prior to lysis. Several progeny nematodes tested positive for the homozygous mutant genotype (e.g. nematode 5 and nematode 13 derived from parent nematode 8). Nematode 5 from parent nematode 8 was taken forward to establish populations of a 2x outcrossed *dapk-1 (knu458 T715N)* strain and this process was repeated to achieve a 4x outcrossed *dapk-1 (knu458 T715N)* strain.

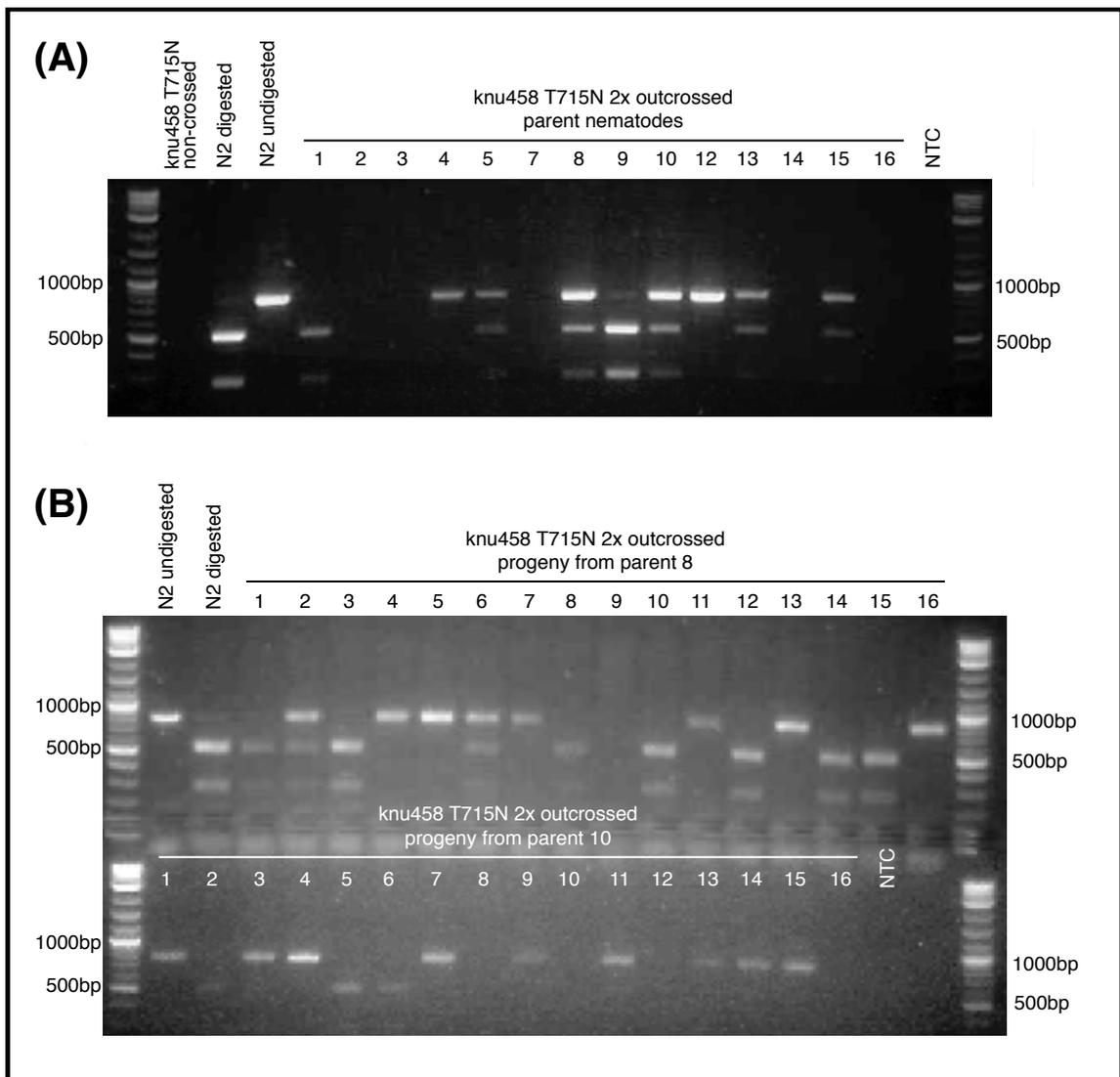


Figure 4.5 – Representative genotype verification upon outcrossing the *dapk-1* (*knu458 T715N*) strain onto the N2 background by PCR and restriction digestion. Images show the resulting genotypes of single nematodes from the 2x outcrossing process. Restriction digestion using *AvaI*. (A) Genotyping parent hermaphrodite nematodes which have undergone 2x outcrossing to identify heterozygotes. (B) Genotyping the progeny population of nematodes derived from heterozygotes identified from the previous image (nematode 8 and nematode 10), to identify homozygotes in relation to the mutation. NTC, no template control.

4.3.3 Phenotypic analysis

4.3.3.1 Head morphology defect

Initial analysis of mutant *dapk-1 C. elegans* revealed that the *dapk-1 (ju4)* mutant strain developed an abnormal head morphology which manifested from an L4 stage (Figure 4.6A). This phenotype was visible using the stereo microscope and was also evident in some of the other mutant *dapk-1* strains. Since this phenotype had been previously reported in relation to the *dapk-1 (gk219)* and *dapk-1 (ju4)* strains with associated penetrance data [165], the penetrance of this phenotype was further assessed in relation to the *dapk-1 (gk219)*, *dapk-1 (ju4)* (replicate assessment), *dapk-1 (knu458 T715N)*, *dapk-1 (knu483 K57W)* and *dapk-1 (knu484 K57W)* strains (Figure 4.6B and 4.6C). The penetrance data collected for the *dapk-1 (gk219)* and *dapk-1 (ju4)* strains aligned closely with the previous report, specifically *dapk-1 (gk219)*: 20% (n=210), compared to a previously reported 19% (n=281); *dapk-1 (ju4)*: 98% (n=442), which was previously reported as 100% (n=309). In relation to the newly generated 4x outcrossed *dapk-1 (knu458 T715N)* mutant strain, this phenotype was only observed in 1 nematode, which equated to less than a 0.3% penetrance amongst the nematodes scored (n=383). This phenotype was absent in the N2 strain.

Since this phenotype was evident in the novel K57W mutant strains, the penetrance of the defect was also quantified. These preliminary results indicate a penetrance of 16% (n=199) and 47% (n=30) for the *dapk-1 (knu483 K57W)* and *dapk-1 (knu484 K57W)* strains, respectively (Figure 4.6C). Of note, in-house genotype validation of these strains was absent and hence these strains had not been outcrossed, however genomic validation performed by Knudra Transgenics supports the desired mutant genotype.

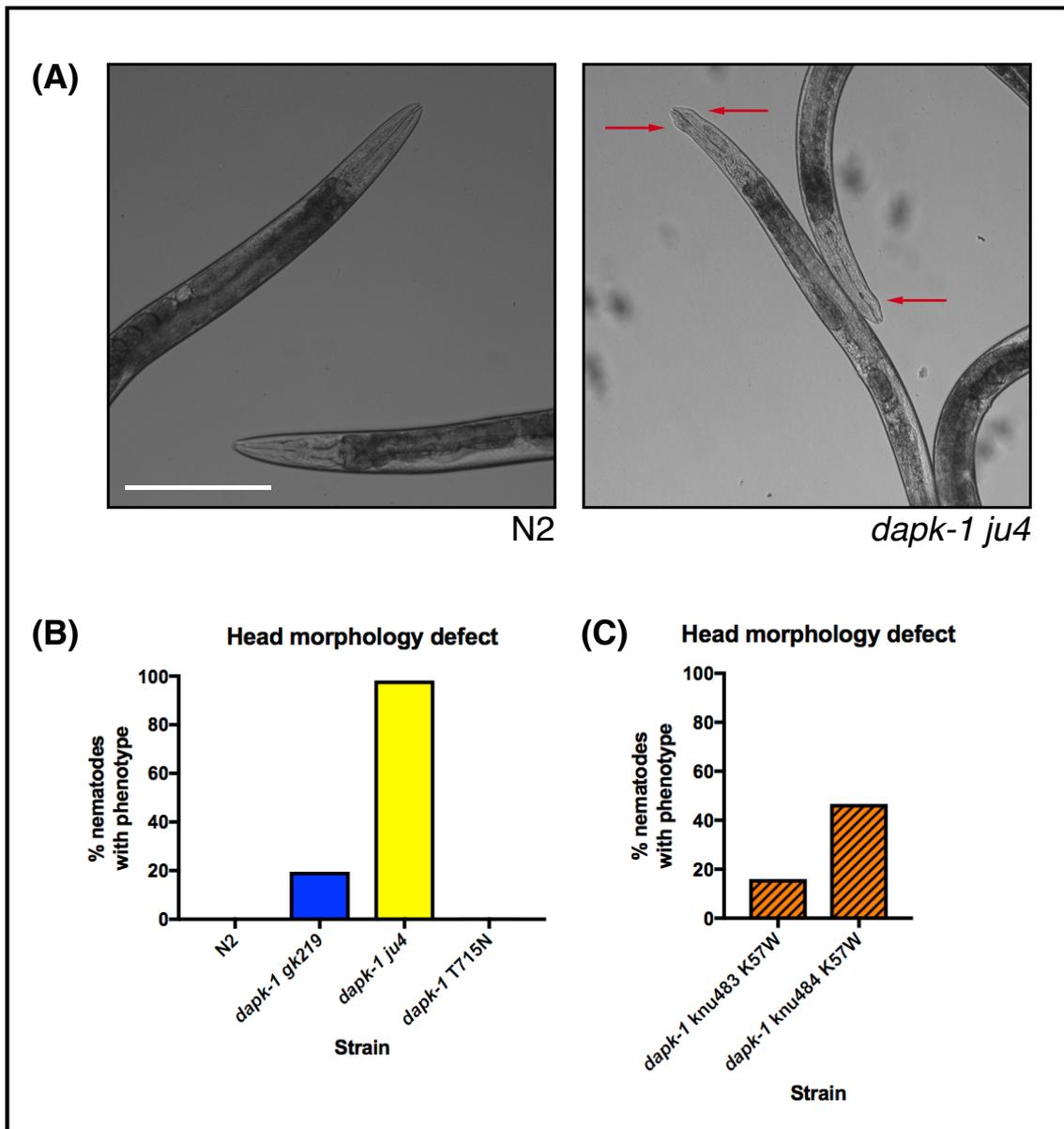


Figure 4.6 – Head morphology defect in mutant *dapk-1 C. elegans* strains. Abnormal head morphology in mutant *dapk-1* strains is penetrant to varying extents dependent on *dapk-1* mutation. (A) Representative transmitted light images of the head region in adult day 1 *C. elegans*, N2 and *dapk-1 (ju4)* genotypes. Red arrows indicate head morphology defect. Imaged on AMG EVOS FL microscope, with 4x objective lens. Scale bar represents 200 μ m. (B) Percentage of *C. elegans* sampled which exhibited head morphology abnormalities, as displayed in the right image above, for several mutant *dapk-1* strains: N2, n=72; *dapk-1 (gk219)*, n=210; *dapk-1(ju4)*, n=442; *dapk-1 (knu458 T715N)*, n=383. (C) Percentage of *dapk-1 K57W* mutant *C. elegans*, from both generated strains, which show head morphology defects. The results for these strains are very preliminary, these nematode populations were not outcrossed and in-house validation of this mutant genotype was incomplete. *dapk-1 (knu483 K57W)*, n=199; *dapk-1 (knu484 K57W)*, n=30.

4.3.3.2 Lifespan assessment

The lifespan of mutant *dapk-1 C. elegans* was assessed and statistically analysed in comparison to the N2 strain. Using an experimental setup whereby the nematodes were maintained on standard NGM, the median lifespan of N2 *C. elegans* was 14 days (Figure 4.7A). The lifespan of *dapk-1 (gk219)* and *dapk-1 (ju4)* mutant *C. elegans* was significantly extended compared to N2 nematodes, to a median lifespan of 16 days and 18 days, respectively (adjusted $p < 0.0004$ in both cases [log-rank Mantel-Cox test with Bonferroni correction]). In particular, the *dapk-1 (ju4)* mutant strain lifespan profile represents very few nematode deaths until day 16 in the assay (Figure 4.7A). The median lifespan of the *dapk-1 (knu458 T715N)* strain nematodes was extended to 16 days, although the difference was not statistically significant across the entire lifespan (adjusted $p = 0.2960$ [log-rank Mantel-Cox test with Bonferroni correction]). The FLAG-*dapk-1* strain was included in this assay and the lifespan profile for these nematodes was similar to that of N2 *C. elegans*, there was no statistically significant difference (adjusted $p > 0.9999$ [log-rank Mantel-Cox test with Bonferroni correction]).

When this assay was repeated but with the supplementation of 100µg/ml ampicillin in the NGM and with heat-killed OP50 (as opposed to a live bacterial culture), the lifespan of nematodes increased across all genotypes tested (Figure 4.7B and 4.7C). Both ampicillin supplementation and the use of dead OP50 have been previously reported to extend lifespan, independently [217–219]. Interestingly, the lifespans of mutant *dapk-1* strains were shortened in comparison to N2 *C. elegans* in this experimental setup (Figure 4.7B). This shortening of lifespan was strongly statistically significant for the *dapk-1 (gk219)* and *dapk-1 (ju4)* strains (adjusted $p < 0.0004$ in both cases [log-rank Mantel-Cox test with Bonferroni correction]) and was the converse of lifespan observations in the absence of ampicillin and presence of live OP50. As previously observed, no statistically significant difference was apparent for the *dapk-1 (knu458 T715N)* strain in comparison to the N2 strain (adjusted $p = 0.1860$ [log-rank Mantel-Cox test with Bonferroni correction]). Upon comparing the two assay results for each genotype, the greatest difference in median lifespan was evident in the N2 strain, whereby the median lifespan was extended by 13 days (Figure 4.7C). The lesser differences in median lifespan were observed in the *dapk-1 (gk219)* and *dapk-1 (ju4)* strains, 1 day and 5 days, respectively.

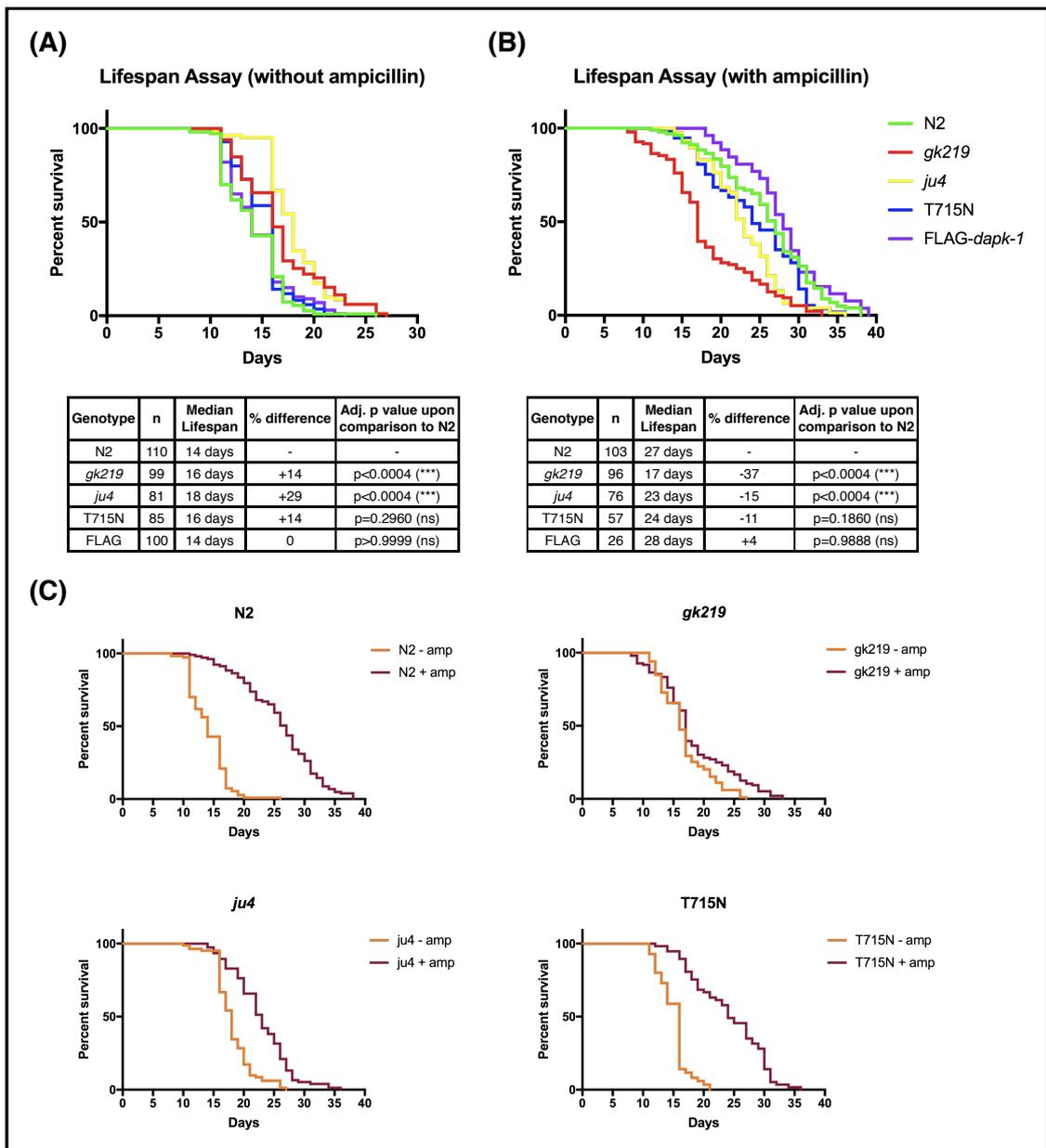


Figure 4.7 – Analysis of lifespan in mutant *dapk-1* *C. elegans* strains. The lifespan of nematode strains are presented as Kaplan-Meier survival plots with the log-rank Mantel-Cox statistical test and Bonferroni multiple testing correction to assess for statistically significant differences. (A) Survival plot of various mutant *dapk-1* *C. elegans* strains maintained on NGM lacking ampicillin. (B) Survival plot of various mutant *dapk-1* *C. elegans* strains maintained on NGM supplemented with 100µg/ml ampicillin. In both cases the summary statistics are presented in the table below the plots. (C) Survival plots for each strain separately, displaying both conditions with and without ampicillin supplemented in the media, data derived from the same assays as presented in plots above. Abbreviations: Adj, adjusted; amp, ampicillin; ns, not significant

4.3.3.3 Viable progeny counts

A further assessment of mutant *dapk-1 C. elegans* fitness was performed by counting the number of viable offspring produced by individual nematodes during the period of reproductive activity (day 1-7 of adulthood). This dataset tested non-normal (on the whole; D'Agostino and Pearson normality test) and therefore the non-parametric Mann-Whitney U test with Bonferroni multiple testing correction was performed to evaluate statistical significance across genotypes. A statistically significant reduction in the total number of viable offspring produced was observed in the *dapk-1 (ju4)* and *dapk-1 (knu458 T715N)* mutant strains, in comparison to N2 nematodes (Figure 4.8A). The greatest difference was evident in the *dapk-1 (ju4)* strain, whereby these nematodes produced 74 less viable progeny (mean average), 195 (*dapk-1 [ju4]*) compared to 269 (N2) (adjusted $p < 0.0003$ [Mann-Whitney U test with Bonferroni correction]).

Since these counts were obtained on a daily basis it was possible to identify on which days there were considerable disparity in viable progeny produced (Figure 4.8B). The greatest difference was observed during day 1 of reproductive maturity, which was also when the greatest number of viable progeny was produced. On average 156 viable progeny were produced by N2 nematodes during day 1 whereas only 77 viable progeny were produced by *dapk-1 (ju4)* mutant *C. elegans* on this day (averages reported as means). From days 2 to 7 the number of progeny produced followed a similar trend throughout the genotypes, whereby fewer progeny were produced each day. Furthermore, on day 3 it was evident the *dapk-1 (gk219)* and *dapk-1 (ju4)* strains produced a greater number of progeny than the N2 nematodes. This is interesting because an observation of delayed development in larval stages of *dapk-1 (gk219)* and *dapk-1 (ju4)* nematodes (which is yet to be quantified) may persist into reproductive maturity whereby reproductive and embryonic developmental processes may also be slowed.

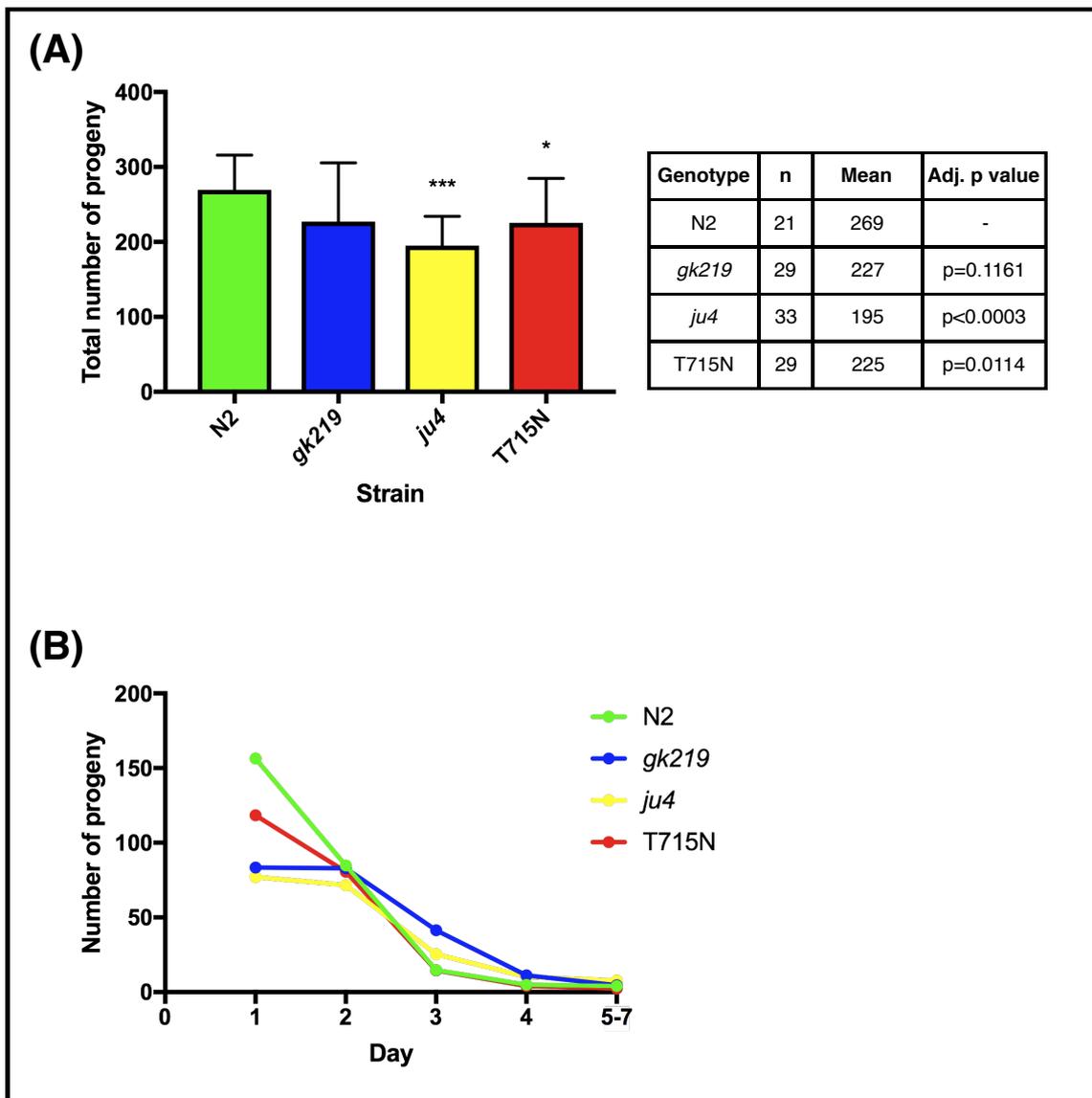


Figure 4.8 – Viable progeny assessment in mutant *dapk-1* *C. elegans* strains. (A) Total number of progeny produced from self-fertilisation of hermaphrodite nematodes across 7 days of reproductive maturity. Data plotted as mean of offspring number with standard deviation. Statistical significance determined by the Mann-Whitney U test with Bonferroni multiple testing correction, comparing each mutant *dapk-1* strain against wild-type (N2). (B) Number of progeny produced day-by-day over the assessment period. Data points indicate the mean average number of progeny produced across all nematodes sampled for each strain. Abbreviations: Adj, adjusted.

4.4 Discussion

The validation and preliminary phenotyping data generated and presented in this chapter provides insight into novel mutant *dapk-1 C. elegans* models and broadens the knowledgebase of previously reported mutant strains. The creation of new mutant strains using CRISPR-Cas9 technology demonstrates the application of genome editing for functional studies in *C. elegans* and the suggested steps to validate and establish a novel mutant strain for phenotypic analysis. Also highlighted are the challenges faced during the genotype validation process.

4.4.1 Further insight into mutant *dapk-1* associated cuticle abnormalities

Abnormalities of the cuticle in mutant *dapk-1 C. elegans*, which manifests prominently in the head region of day 1 adult nematodes, have been previously reported in relation to the *dapk-1 (gk219)* and *dapk-1 (ju4)* strains [165]. Interestingly, the penetrance of this phenotype is considerably greater in the *dapk-1 (ju4)* strain than any other mutant *dapk-1* strains (Table 4.1 and Figure 4.6B). As described, penetrance data obtained during this study for this phenotype align with previous findings and the novel *dapk-1 (knu458 T715N)* mutant nematodes possess this defect at an extremely low penetrance (<0.3%). Mutant *dapk-1* strains with C-terminal DAPK-1 truncations which lack the death domain (*ju469* and *ju557*) also display this phenotype at a moderate penetrance (Table 4.1) [165].

Characterisation of this phenotype has shown that the cuticle thickens by up to 10 times in the *dapk-1 (ju4)* strain expanding both outwards from the body wall and inwards into the underlying epidermal tissue. Moreover, this cuticle accumulation resembles wounding derived scar formation [220]. This led Tong *et al.* to investigate the role of DAPK-1 in relation to innate immune responses which resulted in the proposed function of this protein as a negative regulator of tissue repair in the epidermis and a repressor of immune responses upon wounding. It is speculated that DAPK-1 negatively regulates the p38 MAPK cascade upstream of TIR-1 [165]. In addition, it was identified that antimicrobial peptides (AMPs), NLP-29 and NLP-30, which are downstream of the p38 MAPK pathway, were upregulated in the *dapk-1 (ju4)* strain. However, how DAPK-1 functions at a molecular level remains unclear.

Furthermore, this phenotype is induced by exogenous expression of *dapk-1 S179L* (the *ju4* allele mutation) on the N2 genetic background and therefore it has been suggested that the *ju4* allele results in a *dapk-1* gain-of-function [166]. Due to the position of the underlying mutation, S179L in the kinase domain within a region referred to as the peptide binding ledge, it has been

proposed that this mutation causes alterations to phosphorylation substrate specificity [166]. Using structural inference from the human DAPK1 kinase domain (Figure 4.1) [24], this hypothesis is plausible since this residue aligns within the putative peptide binding region in human DAPK1 and based on three-dimensional modelling of the human structure, appears to be externally orientated.

The explanation of how *dapk-1* genotype influences this phenotype is likely to be complex, largely due to the multidomain structure of DAPK-1 and its potential intrinsic regulatory mechanisms [30] i.e. domain cross-talk. Nevertheless, the kinase domain appears to play a role. This is supported by preliminary results from the newly generated *dapk-1 (knu483 K57W)* and *dapk-1 (knu484 K57W)* strains (in the absence of in-house genotype verification and outcrossing), whereby this phenotype is observed at a 16-47% penetrance. However, transgenic expression of an equivalently kinase inactive form of DAPK-1 (K57A mutant) does not induce or enhance this phenotype on the N2 and *dapk-1 (ju4)* genetic backgrounds, respectively [166]. In addition, it appears GTPase activity of the ROC domain does not significantly impact this phenotype.

4.4.2 Exploring new *dapk-1* phenotype avenues

Further preliminary assessment of mutant *dapk-1* phenotypes was carried out in relation to lifespan and progeny yield. Upon lifespan assessment in the absence of ampicillin in the NGM, the lifespan of the *dapk-1 (gk219)* and *dapk-1 (ju4)* strains was significantly extended. This is particularly interesting in the context of the suggested innate immunity role for DAPK-1. A consequence of *C. elegans* ageing is a decline in immune function, referred to as immunosenescence, due to reduced p38 MAPK signalling [221], a pathway which DAPK-1 is thought to influence [165]. With this process, nematodes become increasingly susceptible to pathogen infection, such routes of infection can be *via* wound sites and the intestine, since intestinal tissue integrity deteriorates with aging [165,222]. OP50, the *C. elegans* bacterial food source, can become an opportunistic pathogen, proliferating in the intestine of aged nematodes and invading surrounding tissues as tissue integrity declines [219]. Since the innate immune response in the gain-of-function *dapk-1 (ju4)* strain is thought to be constitutively active in the epidermis [165], it may be that this strain can evade potential pathogen infection for an extended period throughout aging. This would be an intriguing hypothesis to test in a future study.

When this assay was repeated in the presence of ampicillin and heat-killed OP50, the impact of mutant *dapk-1* genotype on lifespan had the opposite effect compared to N2 nematodes. On average, N2 nematodes outlived mutant *dapk-1* nematodes. In this assay, concentrated heat-killed OP50 was fed to the worms and therefore there would be no growth of this bacteria on the plates. Consequently, the likelihood of OP50 becoming pathogenic *via* intestinal and surrounding tissue inhabitation was much less and hence this is a likely contributor to the extended lifespan across all genotypes. The comparably shortened lifespan in mutant *dapk-1* strains in this experimental setup requires further investigation. Of note, the difference in lifespan profile for the *dapk-1 (gk219)* strain in the presence and absence of ampicillin and live OP50 is relatively slight, compared to differences across other genotypes. This suggests that DAPK-1 plays a role in the extended lifespan evident in the presence of ampicillin and heat-killed OP50, which is notably observed in the N2 strain (Figure 4.7C). In addition, mortality in these mutant nematodes may be due to a mechanism unrelated to alterations in immune responses. One potential route for future investigation is the role of autophagy in longevity of mutant *dapk-1 C. elegans*, since *dapk-1* knockdown has been associated with reduced starvation-induced autophagy [164] and autophagy is a crucial process for *C. elegans* longevity [223]. Moreover, human DAPK1 regulates autophagy induction *via* beclin 1 [49,84], which strengthens motivation to pursue this line of investigation further.

Progeny yield or brood size was another phenotype assessed in these mutant *dapk-1* nematodes. The results obtained, when testing the *dapk-1 (ju4)* and *dapk-1 (knu458 T715N)* mutant strains, show that brood size is significantly compromised in these nematodes in comparison to N2 nematodes. Again, the *dapk-1 (ju4)* strain displayed the most prominent phenotype. Although these are preliminary findings, these data provides a foundation for exploring this phenotype in further detail. Indications from these data suggest both the kinase and GTPase activities of DAPK-1 play a role in biological processes which lead to this phenotype. At a functional level this observation could be due to defects in various physiological processes including gamete production, fertilisation and embryogenesis, or a secondary consequence of an indirect insult to nematode physiology. These results taken together with extensive evidence that human DAPK1 is involved in apoptosis [49] and that apoptosis is a fundamental process in *C. elegans* embryonic development [224], positions apoptosis during embryogenesis as a candidate process to investigate in the context of mutant *dapk-1 C. elegans* strains. This is also supported by life-stage mRNA expression data, whereby a peak in *dapk-1* expression during late embryogenesis is evident (Figure 3.12).

4.4.3 Genotype validation challenges

Using PCR and restriction digestion, genotype validation of two of the mutant *dapk-1* strains was inconclusive. First, validation of the *ju4* allele using the dCAPS primer method followed by restriction digestion with *XbaI* resulted in the expected PCR amplification, however the expected amplicon cleavage, specific to the N2 template background, did not occur. Despite troubleshooting this experiment in relation to increasing both the agarose gel percentage for electrophoresis and the restriction digestion incubation time, the expected result was not observed. Further optimisation was not pursued since this nematode strain exhibited the head morphology defect phenotype as previously described [165]. This phenotype acted as a proxy for genotype validation, however verification of the genomic sequence in exon 3 of *dapk-1* in this strain is required prior to further experimentation.

In-house genotype validation of the newly generated *dapk-1* (*knu483 K57W*) and *dapk-1* (*knu484 K57W*) strains was also challenging. Despite sequencing derived verification data from Knudra Transgenics and a number of PCR optimisation approaches trialled, evidence of this genotype was not robustly reproducible. Performing in-house validation of this strain by PCR and restriction digestion was imperative for this strain to be outcrossed for phenotypic analysis, since this is the method utilised for tracking the genotype through the outcrossing process. Thoughts of potential off-target issues of the CRISPR-Cas9 approach [225], such as off-target genomic DNA cleavage and sequence insertion at the undesired locus for example, or even on-target genomic damage [226], have not been eliminated. However, when performing PCR using the forward 2 and reverse 2 primer combination, which was designed to amplify a region of approximately 1kb, indications of the expected amplicon and digested products were evident. Therefore, further PCR optimisation with this primer pair is worthwhile pursuing.

4.4.4 Conclusions

This body of research provides the initial direction for expanding our understanding of DAPK-1 in *C. elegans*. Novel mutant *dapk-1* strains were engineered whereby key residues in relation to the hypothesised catalytic activities of DAPK-1 were precisely targeted, enabling phenotypic assessment of nematodes possessing kinase inactive or GTP binding deficient forms of DAPK-1. These genetic tools form a strong basis to enable dissection of DAPK-1 function *in vivo*, specifically in relation to the catalytic activities of this protein. The genotype of the GTPase deficient T715N mutant was successfully validated and this strain was 4x outcrossed in

preparation for phenotypic characterisation. In-house genotype validation of the kinase inactive K57W strain is incomplete. Further optimisation is required to firmly verify this mutant genotype, which then requires outcrossing onto the wild-type genetic background prior to any detailed phenotypic analysis.

In addition, the initial assessment of phenotypes associated with mutant *dapk-1 C. elegans* will guide future investigation into defining and characterising a clear *dapk-1* related phenotype. Once a defined phenotype has been established, this adds a readout to the toolbox for understanding DAPK-1 function. In particular, this would be of considerable value for validating DAPK-1 interactors (discussed in the previous chapter), whereby knockdown or overexpression of these potential interactors could be used as an approach to assess the modulation of an described phenotype. The results obtained build on the previously proposed role of DAPK-1 in innate immune responses, with suggestions that the absence of DAPK-1 GTPase activity does not impact the head morphology defect. Furthermore, this preliminary phenotyping data also highlights routes for potential examination, such as the role of DAPK-1 in autophagy and apoptosis.

CHAPTER FIVE

General Discussion and Suggested Future Work

5.1 General Discussion

Due to the multidomain structure and dual catalytic activities of DAPK1, the role and regulation of this protein is likely to be considerably complex. The basis to this complexity can be deciphered through defining the proximal interactome of DAPK1. Efforts to delineate the interaction profile of DAPK1 have been fruitful, however there remains scope for further developing our understanding of where DAPK1 fits within the wider subcellular landscape. Since this protein, and more broadly the human ROCO proteins which share structural homology, are linked to numerous diseases, this is also substantial interest in understand the function of these proteins from a therapeutic perspective.

The overarching goal of the research presented in this thesis was to gain a deeper understanding of the physiological function of DAPK1, taking a bioinformatic and functional modelling approach. A programme of research was established which encompassed resource development and investigative analyses to collectively contribute towards this principal goal. Each results chapter (Chapters Two, Three and Four) describe and discuss focussed projects with this convergent theme of defining the DAPK1 interactome, initially in relation to the human proteome, then due to the genetically tractability of this protein and the opportunity to explore its interaction landscape *in vivo*, in relation to the *C. elegans* proteome.

5.1.1 Research overview

An underpinning aspect to a considerable proportion of the research presented was the development and application of the Weighted Protein-Protein Interaction Network Analysis (WPPINA) pipeline which was further developed into an online PPI query resource, PINOT (Protein Interaction Network Online Tool) [121]. This bioinformatic tool maximises the utility of freely available literature-derived human and *C. elegans* PPI data through a systematic data processing pipeline. Data is extracted and collated from numerous repositories, subjected to quality control measures and confidence scored based on method detection and publication records, providing the user with an easy-to-interpret output file for downstream applications. Two notable features of PINOT which provides enhanced value are that data is extracted at the time of query (for human PPIs), and all entries in the data output are of verifiable, fully traceable back to the source publication. This pipeline was utilised at multiple stages throughout this thesis.

In terms of investigating the DAPK1 PPI profile, the first analyses focussed on the ROCO protein family as a whole [12,227]. This project aimed to gather interactomic and functional insight into the commonalities and distinctions between these structurally related multidomain proteins. PPI network analyses was performed using literature-derived data (*via* the WPPINA pipeline) and novel protein microarray data, and in combination these datasets were utilised to prioritise potential interactors for validation and functional studies. These prioritised interactors include numerous common nodes between multiple ROCO proteins. Functional enrichment analysis shed light on a potential shared function of the ROCO proteins centred on the stress response, although it appears these proteins have largely evolved divergent roles, based on the overall degree of overlap between their interaction profiles and functional characterisation to date.

Exploring the DAPK1 interaction profile was then translated into the context of *C. elegans* for predictive and experimental DAPK-1 PPI mapping. Predicted DAPK-1 PPI networks were generated utilising two distinct approaches, the MIST interolog feature and PINOT coupled with Ortholist 2.0, which both rely on orthologous inference. Reported interactors were then added to PINOT/Ortholist-derived network to incorporate evidence based PPI data and position the potential DAPK-1 interactome in a wider cellular context. These approaches demonstrate the utility of data integration strategies, in particular for prioritising interactors for future experimentation. Furthermore in this body of research, six novel DAPK-1 interactors were identified by an outsourced yeast two-hybrid (Y2H) screen. Of these hits, MEP-1, SYD-9 and UNC-14 appear to be intriguing candidate proteins for follow-up studies, based on their prior characterisation.

The identification of UNC-14 as a DAPK-1 interactor is of particular interest in the context of autophagy. UNC-14 is reported to interact with UNC-51 [194,198], which is orthologous to human ULK1, a kinase involved in the regulation of phagophore formation during the initiation of autophagy [228]. In addition, it is reported that human DAPK1 phosphorylates beclin 1 [84], which promotes assembly of the class III phosphatidylinositol 3-kinase (PI3KC3) complex I required for the induction of autophagy [49,228]. This positions DAPK1 as a potential coregulator of the ULK1 and beclin 1/PI3KC3 pathways in autophagy induction. Therefore, testing whether DAPK1 regulates ULK1 in the mammalian system provides an exciting route for future investigation.

Finally, a number of novel *dapk-1 C. elegans* strains were engineered and initial validation and phenotypic analysis is underway. These include a FLAG-*dapk-1* strain whereby a 3xFLAG tag was knocked-in to the *dapk-1* locus using CRISPR-Cas9 genome editing. The rationale for generating

this strain was to perform *in vivo* DAPK-1 PPI interaction analysis, such as AP-MS. Furthermore, K57W and T715N mutant *dapk-1* strains were generated and these are hypothesised to result in kinase inactive and GTP binding deficient forms of DAPK-1, respectively. Of note, further work is required to validate the genotype of the K57W strain. These strains will aid in deciphering how the biochemical activity of DAPK-1 relates to the interaction profile and ultimately the function of DAPK-1 *in vivo*. Preliminary phenotypic analysis of these strains, in conjunction with other available mutant *dapk-1* strains, *dapk-1 (gk219 [dapk-1 knockout])* and *dapk-1 (ju4 [proposed dapk-1 gain-of-function [166]])*, have yielded some interesting results to pursue for future research. In particular, reduced progeny counts in mutant *dapk-1* strains were observed, providing a line of enquiry, potentially in relation to apoptosis in embryogenesis. The route for further examination is supported by extensive evidence relating to the role of DAPK1 in apoptotic pathways [49].

5.1.2 Outlook

The research outcomes of the work presented in this thesis contribute novel findings in relation to the interaction and functional profiles of DAPK1 in human (in the context of the ROCO protein family), and DAPK-1 in *C. elegans*. In addition, this work will guide future strands of investigation, for both, DAPK1/DAPK-1 research and more broadly, the development of PINOT for example, will assist in PPI analyses in general.

5.1.2.1 The human DAPK1 and *C. elegans* DAPK-1 interactomes

The literature-derived ROCO protein interaction network analysis revealed the extent of the research into defining the human DAPK1 interaction profile. Progress into mapping this interactome had been fruitful and the number of potential DAPK1 binding partners which possess functional connections to a diverse range of signalling cascades demonstrates the complex nature of understanding the role and regulation of DAPK1 [75]. However, when considering this data with an aspect of confidence-weighting, interactors and thus signalling pathways which are more confidently associated with DAPK1 can be distinguished from the expanse of detected DAPK1 PPIs. Since there is therapeutic interest in targeting DAPK1 [229], identifying *bona fide* substrates and developing the necessary tools to dissect DAPK1 signalling (such as high quality antibodies and specific chemical modulators) is of importance. In addition, visualising this PPI data collectively in conjunction with PPI data relating to the other ROCO proteins enabled identification of the potential overlap in interaction profiles of these partially homologous proteins [12]. Therefore, this PPI analysis approach enabled a vision of the DAPK1 and more broadly the ROCO protein interaction landscape from a new angle.

The predicted *C. elegans* DAPK-1 PPI network and the putative DAPK-1 interactors identified by Y2H provide new leads for defining the role of DAPK-1 in *C. elegans*. The DAPK-1 interactome was largely undefined prior to embarking on this research project, and hence these potential interactors will facilitate in generating insight into the molecular mechanisms underlying DAPK-1 function. Although only CMD-1 was common to both the predicted and identified DAPK-1 interactors, the predicted interactors are of importance for determining whether the structural orthology between the human and *C. elegans* forms of DAPK1/DAPK-1 persists into interactome and in turn, functional orthology. For example, it has been proposed that the role of *C. elegans* DAPK-1 in innate immune responses is analogous to the role of human DAPK1 in inflammation [161]. If this is the case, inference from *C. elegans* *dapk-1* models for understanding DAPK1 in humans and *vice versa*, becomes a powerful approach. Moreover, if it is the case it poses the question, how translational is it to other structurally orthologous proteins?

From phenotypic analysis it is clear the DAPK-1 plays an important role in *C. elegans* homeostasis, most notably in the epidermis [165]. However, although these phenotypic links have been observed, how DAPK-1 is regulated as a molecule and how it impacts the wider landscape of subcellular signalling is unknown. Once the novel mutant *dapk-1* strains (K57W and T715N) are established, the outlook for contributing evidence to these unknowns is promising. These strains are proposed to act as tools for assessing the kinase and GTPase activities of DAPK-1, independently, therefore enabling phenotype analysis in the context of specific DAPK-1 catalytic activities. Furthermore, a distinct yet complementary approach for evaluating the role of DAPK-1 kinase and GTPase activities in *C. elegans* cell signalling would be to chemically manipulate these catalytic activities, however this is largely reliant on the development of novel compounds in order to eliminate concerns of off-target effects derived from current inhibitors/activators tested in relation to human DAPK1 [91].

5.1.2.2 PINOT and PPI data curation

It is hoped that the development of PINOT will have a positive impact on the wider research community, since this resource is applicable and useful for PPI analyses across the human and *C. elegans* proteomes and adds value to the current selection of PPI query resources. The utility of PINOT is diverse, several applications are presented here, with respect to providing an overview of the ROCO protein interaction profiles, prioritising hits from novel high-throughput PPI screens and predicting the proximal interactome of DAPK-1. In addition, this approach has been applied to numerous other projects, including mapping human disease related PPI

networks [13,14]. The PINOT data output provide a valuable foundation of confidence scored PPIs collated from decades of published research for a multitude of downstream analyses.

As discussed in Chapter Two, there are limitations to utilising literature-derived PPI data, those already discussed include the incompleteness of the data, the ascertainment bias and the reliance on high quality database curation. Unfortunately, these limitations will persist into the future. The incompleteness of data falls into two categories, PPIs that have not be identified and also PPIs that have been detected but not curated into repositories. This latter scenario is particularly pertinent in the current era of high-throughput screens and therefore database curators, such as the team at IntAct, are attempting to reduce this void in curation by encouraging research groups to deposit data in repositories as standard procedure, especially targeting this message at research groups known to generate masses of PPI data. Noteworthy progress in the molecular interaction curation field with regard to data quality and consistency is due efforts in standardisation led by the IMEx consortium [6].

In addition, there are many considerations centred around the experimental setup of PPI detection and physiological relevance of an identified PPI, that require thought within the existent PPI datasets. For example, the post-translational modification (PTM) status of a protein will likely influence its interaction profile, this has been demonstrated for LRRK2 [230]. However, often this data is either not known or not readily available to be an included annotation in the curation process. Furthermore, the temporal and spatial expression patterns of proteins will impact the occurrence of a PPI, this is an area partially addressed in relation to data integration of tissue-specific mRNA expression measures in the analysis of common ROCO protein interactors. With advances in detecting cell type specific expression levels [231], these strategies could be implemented with further precision. Emerging evidence also suggests tRNA abundance, which is variable in different tissues, is a key player in protein translation efficiency and in turn, protein levels [232]. In addition, subcellular localisation of potential binding partners is a further consideration, for example FADD which is a reported interactor of DAPK1 and LRRK2 has distinct functional roles depending on its cellular localisation and phosphorylation state [233], this will also be reflected in distinctions in its interaction profiles. Therefore, expression and localisation data integration strategies will be of significant benefit and add a layer of physiological relevance to PPI network analysis.

5.2 Future Work

In addition to the novel findings and developments presented in this thesis, this work provides the foundation for future investigation in a number of directions. Many of these routes for further work have been suggested within each results chapter. A repeated theme from the PPI network analyses is the prioritisation of interactors for validation. Suggested future work of particular interest is highlighted below.

First, in relation to further defining the human DAPK1 interaction network, novel experimental PPI data generated by distinct method detection strategies will continue to broaden our understanding of the DAPK1 interactome. Interactors detected by several distinct methodological approaches are of particular interest for identifying novel *bona fide* DAPK1 interactors for functional characterisation experiments. Similarly, follow-up experimentation of common interactors between ROCO proteins provide an intriguing direction for future study. These nodes are of particular interest because although it appears the ROCO proteins largely occupy separate functional spheres, commonality in their interaction profiles suggests potential cross-talk in their proximal interactomes. Additionally, and in relation to progressing the *in silico* analysis, the integration of further data types, for example protein expression and localisation data (as mentioned in the previous section), would be of significant benefit for better interpreting the physiological relevance of binding partners.

With regard to avenues for further investigation of DAPK-1 in *C. elegans*, validating the physical interaction of the hits identified *via* Y2H, in particular MEP-1, SYD-9 and UNC-14, would be a valuable next step. Testing the validity of these interactors would be best achieved using a distinct method detection technique, such as co-immunoprecipitation, and further characterisation in relation to phosphorylation mapping *via* mass spectrometry approaches would elucidate whether these proteins are DAPK-1 substrates. If these proteins are genuine DAPK-1 interactors they represent important tools for deciphering DAPK-1 function *in vivo*. Next, further work centred around the validation and characterisation of novel *dapk-1 C. elegans* models is necessary. In particular, establishing phenotypic readouts in the K57W and T715N *dapk-1* models would result in a powerful tool to assess proposed kinase inactivity and GTP binding deficiency of DAPK-1 *in vivo*. These strains could then be utilised for assessing the effect of manipulating the expression (i.e. knock-down, overexpress) of the putative DAPK-1 interactors, that were identified *via* the Y2H screen, on an established *dapk-1* related phenotype. Furthermore, *C. elegans* is widely used for compound screening [234] and therefore compounds, such as DAPK inhibitors and activators [91,229], could be tested on nematodes in

order to expand the DAPK-1 manipulation toolbox to chemical approaches for developing our understanding of the role and regulation of DAPK-1. Ultimately, results from these *in vivo* assessments could form the basis of biochemical and functional inferences from *C. elegans* DAPK-1 to human DAPK1.

Finally, future work to maintain and develop PINOT is in the pipeline. Although the *C. elegans* query option encompasses a wide coverage of PPI data, the widest coverage of complete data entries compared with alternative tools, there is scope for further widening this data capture. For example *via* incorporation of the PSICQUIC API in addition to utilising WormBase derived data. In addition, one feature PINOT lacks compared to alternative PPI query tools is a network visualisation alongside the data output and therefore developing this feature would strengthen the resource further.

REFERENCES

- [1] Rolland, T., Tas, M., Sahni, N., Yi, S., et al., A Proteome-Scale Map of the Human Interactome Network. *Cell* 2014, 159, 1212–1226.
- [2] Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., et al., The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* 2015, 162, 425–440.
- [3] Huttlin, E.L., Bruckner, R.J., Paulo, J.A., Cannon, J.R., et al., Architecture of the human interactome defines protein communities and disease networks. *Nature* 2017, 545, 505–509.
- [4] Chatr-aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., et al., The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* 2015, 43, D470–D478.
- [5] Kerrien, S., Aranda, B., Breuza, L., Bridge, A., et al., The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* 2012, 40, D841-6.
- [6] Orchard, S., Kerrien, S., Abbani, S., Aranda, B., et al., Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods* 2012, 9, 345–350.
- [7] Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 2015, 43, D1049-56.
- [8] Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., et al., The Reactome pathway Knowledgebase. *Nucleic Acids Res.* 2016, 44, D481-7.
- [9] Slenter, D.N., Kutmon, M., Hanspers, K., Riutta, A., et al., WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* 2018, 46, D661–D667.
- [10] Oliver, S., Guilt-by-association goes global. *Nature* 2000, 403, 601–603.
- [11] Taipale, M., Tucker, G., Peng, J., Krykbaeva, I., et al., A quantitative chaperone interaction network reveals the architecture of cellular protein homeostasis pathways. *Cell* 2014, 158, 434–448.
- [12] Tomkins, J.E., Dihanich, S., Beilina, A., Ferrari, R., et al., Comparative Protein Interaction Network Analysis Identifies Shared and Distinct Functions for the Human ROCO Proteins. *Proteomics* 2018, 18, e1700444.
- [13] Ferrari, R., Lovering, R.C., Hardy, J., Lewis, P.A., Manzoni, C., Weighted Protein Interaction Network Analysis of Frontotemporal Dementia. *J. Proteome Res.* 2017, 16, 999–1013.
- [14] Ferrari, R., Kia, D.A., Tomkins, J.E., Hardy, J., et al., Stratification of candidate genes for Parkinson’s disease using weighted protein-protein interaction network analysis. *BMC Genomics* 2018, 19, 452.

- [15] Panga, V., Raghunathan, S., A cytokine protein-protein interaction network for identifying key molecules in rheumatoid arthritis. *PLoS One* 2018, 13, e0199530.
- [16] Bis-Brewer, D.M., Danzi, M.C., Wuchty, S., Zuchner, S., A network biology approach to unraveling inherited axonopathies. *Sci. Rep.* 2019, 9, 1692.
- [17] Huang, K., Fingar, D.C., Growing knowledge of the mTOR signaling network. *Semin. Cell Dev. Biol.* 2014, 36, 79–90.
- [18] Liu, C., Liu, L., Zhou, C., Zhuang, J., et al., Protein-protein interaction networks and different clustering analysis in Burkitt's lymphoma. *Hematology* 2018, 23, 391–398.
- [19] Cong, Q., Anishchenko, I., Ovchinnikov, S., Baker, D., Protein interaction networks revealed by proteome coevolution. *Science (80-.).* 2019, 365, 185–189.
- [20] Lewis, P.A., The function of ROCO proteins in health and disease. *Biol. Cell* 2009, 101, 183–191.
- [21] Civiero, L., Dihanich, S., Lewis, P.A., Greggio, E., Genetic, structural, and molecular insights into the function of ras of complex proteins domains. *Chem. Biol.* 2014, 21, 809–818.
- [22] Bosgraaf, L., Van Haastert, P.J.M., Roc, a Ras/GTPase domain in complex proteins. *Biochim. Biophys. Acta - Mol. Cell Res.* 2003, 1643, 5–10.
- [23] Marín, I., van Egmond, W.N., van Haastert, P.J.M., The Roco protein family: a functional perspective. *FASEB J.* 2008, 22, 3103–3110.
- [24] Tereshko, V., Teplova, M., Brunzelle, J., Watterson, D.M., Egli, M., Crystal structures of the catalytic domain of human protein kinase associated with apoptosis and tumor suppression. *Nat Struct Biol* 2001, 8, 899–907.
- [25] Shiloh, R., Bialik, S., Kimchi, A., The DAPK family: A structure-function analysis. *Apoptosis* 2014, 19, 286–297.
- [26] Marin, I., The Parkinson disease gene LRRK2: evolutionary and structural insights. *Mol. Biol. Evol.* 2006, 23, 2423–2433.
- [27] West, A.B., Moore, D.J., Choi, C., Andrabi, S.A., et al., Parkinson's disease-associated mutations in LRRK2 link enhanced GTP-binding and kinase activities to neuronal toxicity. *Hum. Mol. Genet.* 2007, 16, 223–232.
- [28] Wauters, L., Versees, W., Kortholt, A., Roco Proteins: GTPases with a Baroque Structure and Mechanism. *Int. J. Mol. Sci.* 2019, 20.
- [29] Carlessi, R., Levin-Salomon, V., Ciprut, S., Bialik, S., et al., GTP binding to the ROC domain of DAP-kinase regulates its function through intramolecular signalling. *EMBO Rep.* 2011, 12, 917–23.

- [30] Jebelli, J.D., Dihanich, S., Civiero, L., Manzoni, C., et al., GTP binding and intramolecular regulation by the ROC domain of Death Associated Protein Kinase 1. *Sci. Rep.* 2012, 2, 695.
- [31] Korr, D., Toschi, L., Donner, P., Pohlenz, H.-D., et al., LRRK1 protein kinase activity is stimulated upon binding of GTP to its Roc domain. *Cell. Signal.* 2006, 18, 910–920.
- [32] Guo, L., Gandhi, P.N., Wang, W., Petersen, R.B., et al., The Parkinson's disease-associated protein, leucine-rich repeat kinase 2 (LRRK2), is an authentic GTPase that stimulates kinase activity. *Exp. Cell Res.* 2007, 313, 3658–3670.
- [33] Dihanich, S., Civiero, L., Manzoni, C., Mamais, A., et al., GTP binding controls complex formation by the human ROCO protein MASL1. *FEBS J.* 2014, 281, 261–274.
- [34] Ito, G., Okai, T., Fujino, G., Takeda, K., et al., GTP binding is essential to the protein kinase activity of LRRK2, a causative gene product for familial Parkinson's disease. *Biochemistry* 2007, 46, 1380–1388.
- [35] Klein, C.L., Rovelli, G., Springer, W., Schall, C., et al., Homo- and heterodimerization of ROCO kinases: LRRK2 kinase inhibition by the LRRK2 ROCO fragment. *J. Neurochem.* 2009, 111, 703–715.
- [36] Reyniers, L., Del Giudice, M.G., Civiero, L., Belluzzi, E., et al., Differential protein-protein interactions of LRRK1 and LRRK2 indicate roles in distinct cellular signaling pathways. *J. Neurochem.* 2014, 131, 239–250.
- [37] Hanafusa, H., Ishikawa, K., Kedashiro, S., Saigo, T., et al., Leucine-rich repeat kinase LRRK1 regulates endosomal trafficking of the EGF receptor. *Nat. Commun.* 2011, 2, 158.
- [38] Hanafusa, H., Yagi, T., Ikeda, H., Hisamoto, N., et al., LRRK1 phosphorylation of Rab7 at S72 links trafficking of EGFR-containing endosomes to its effector RILP. *J. Cell Sci.* 2019, 132.
- [39] Hanafusa, H., Kedashiro, S., Tezuka, M., Funatsu, M., et al., PLK1-dependent activation of LRRK1 regulates spindle orientation by phosphorylating CDK5RAP2. *Nat. Cell Biol.* 2015, 17, 1024–1035.
- [40] Morimoto, K., Baba, Y., Shinohara, H., Kang, S., et al., LRRK1 is critical in the regulation of B-cell responses and CARMA1-dependent NF-kappaB activation. *Sci. Rep.* 2016, 6, 25738.
- [41] Xing, W.R., Goodluck, H., Zeng, C., Mohan, S., Role and mechanism of action of leucine-rich repeat kinase 1 in bone. *Bone Res.* 2017, 5, 17003.
- [42] Iida, A., Xing, W., Docx, M.K.F., Nakashima, T., et al., Identification of biallelic LRRK1 mutations in osteosclerotic metaphyseal dysplasia and evidence for locus heterogeneity. *J. Med. Genet.* 2016, 53, 568–574.
- [43] Guo, L., Girisha, K.M., Iida, A., Hebbar, M., et al., Identification of a novel LRRK1 mutation in a family with osteosclerotic metaphyseal dysplasia. *J. Hum. Genet.* 2017, 62, 437–441.

- [44] Wallings, R., Manzoni, C., Bandopadhyay, R., Cellular processes associated with LRRK2 function and dysfunction. *FEBS J.* 2015, 282, 2806–2826.
- [45] Kluss, J.H., Mamais, A., Cookson, M.R., LRRK2 links genetic and sporadic Parkinson's disease. *Biochem. Soc. Trans.* 2019, 47, 651–661.
- [46] Dihanich, S., MASL1: a neglected ROCO protein. *Biochem. Soc. Trans.* 2012, 40, 1090–1094.
- [47] Zhong, J., Wang, H., Chen, W., Sun, Z., et al., Ubiquitylation of MFHAS1 by the ubiquitin ligase praja2 promotes M1 macrophage polarization by activating JNK and p38 pathways. *Cell Death Dis.* 2017, 8, e2763.
- [48] Kumkhaek, C., Aerbajinai, W., Liu, W., Zhu, J., et al., MASL1 induces erythroid differentiation in human erythropoietin-dependent CD34+ cells through the Raf/MEK/ERK pathway. *Blood* 2013, 121, 3216–3227.
- [49] Singh, P., Ravanan, P., Talwar, P., Death Associated Protein Kinase 1 (DAPK1): A Regulator of Apoptosis and Autophagy. *Front. Mol. Neurosci.* 2016, 9, 46.
- [50] Deiss, L.P., Feinstein, E., Berissi, H., Cohen, O., Kimchi, A., Identification of a novel serine/threonine kinase and a novel 15-kD protein as potential mediators of the gamma interferon-induced cell death. *Genes Dev.* 1995, 9, 15–30.
- [51] Feinstein, E., Druck, T., Kastury, K., Berissi, H., et al., Assignment of DAP1 and DAPK--genes that positively mediate programmed cell death triggered by IFN-gamma--to chromosome regions 5p12.2 and 9q34.1, respectively. *Genomics* 1995, 29, 305–307.
- [52] Benderska, N., Schneider-Stock, R., Transcription control of DAPK. *Apoptosis* 2014, 19, 298–305.
- [53] Pulling, L.C., Grimes, M.J., Damiani, L.A., Juri, D.E., et al., Dual promoter regulation of death-associated protein kinase gene leads to differentially silenced transcripts by methylation in cancer. *Carcinogenesis* 2009, 30, 2023–30.
- [54] Chen, H.Y., Lin, Y.M., Chung, H.C., Lang, Y.D., et al., MiR-103/107 promote metastasis of colorectal cancer by targeting the metastasis suppressors DAPK and KLF4. *Cancer Res.* 2012, 72, 3631–3641.
- [55] Jin, Y., Blue, E.K., Dixon, S., Hou, L., et al., Identification of a new form of death-associated protein kinase that promotes cell survival. *J. Biol. Chem.* 2001, 276, 39667–39678.
- [56] Jin, Y., Blue, E.K., Gallagher, P.J., Control of death-associated protein kinase (DAPK) activity by phosphorylation and proteasomal degradation. *J. Biol. Chem.* 2006, 281, 39033–39040.
- [57] Lin, Y., Stevens, C., Hrstka, R., Harrison, B., et al., An alternative transcript from the death-associated protein kinase 1 locus encoding a small protein selectively mediates membrane blebbing. *FEBS J.* 2008, 275, 2574–84.

- [58] de Diego, I., Kuper, J., Bakalova, N., Kursula, P., Wilmanns, M., Molecular basis of the death-associated protein kinase-calcium/calmodulin regulator complex. *Sci. Signal.* 2010, 3, ra6.
- [59] Zimmermann, M., Atmanene, C., Xu, Q., Fouillen, L., et al., Homodimerization of the death-associated protein kinase catalytic domain: development of a new small molecule fluorescent reporter. *PLoS One* 2010, 5, e14120.
- [60] Shani, G., Marash, L., Gozuacik, D., Bialik, S., et al., Death-Associated Protein Kinase Phosphorylates ZIP Kinase , Forming a Unique Kinase Hierarchy To Activate Its Cell Death Functions. *Mol. Cell. Biol.* 2004, 24, 8611–8626.
- [61] Citri, A., Harari, D., Shohat, G., Ramakrishnan, P., et al., Hsp90 recognizes a common surface on client kinases. *J. Biol. Chem.* 2006, 281, 14361–14369.
- [62] Zhang, L., Nephew, K.P., Gallagher, P.J., Regulation of death-associated protein kinase. Stabilization by HSP90 heterocomplexes. *J. Biol. Chem.* 2007, 282, 11795–804.
- [63] Cohen, O., Feinstein, E., Kimchi, A., DAP-kinase is a Ca²⁺/calmodulin-dependent, cytoskeletal-associated protein kinase, with cell death-inducing functions that depend on its catalytic activity. *EMBO J.* 1997, 16, 998–1008.
- [64] Shohat, G., Spivak-Kroizman, T., Cohen, O., Bialik, S., et al., The pro-apoptotic function of death-associated protein kinase is controlled by a unique inhibitory autophosphorylation-based mechanism. *J. Biol. Chem.* 2001, 276, 47460–7.
- [65] Gozuacik, D., Bialik, S., Raveh, T., Mitou, G., et al., DAP-kinase is a mediator of endoplasmic reticulum stress-induced caspase activation and autophagic cell death. *Cell Death Differ.* 2008, 15, 1875–1886.
- [66] Guenebeaud, C., Goldschneider, D., Castets, M., Guix, C., et al., The dependence receptor UNC5H2/B triggers apoptosis via PP2A-mediated dephosphorylation of DAP kinase. *Mol. Cell* 2010, 40, 863–76.
- [67] Widau, R.C., Jin, Y., Dixon, S. a, Wadzinski, B.E., Gallagher, P.J., Protein phosphatase 2A (PP2A) holoenzymes regulate death-associated protein kinase (DAPK) in ceramide-induced anoikis. *J. Biol. Chem.* 2010, 285, 13827–38.
- [68] Bialik, S., Bresnick, a R., Kimchi, A., DAP-kinase-mediated morphological changes are localization dependent and involve myosin-II phosphorylation. *Cell Death Differ.* 2004, 11, 631–44.
- [69] Chen, C.-H., Wang, W.-J., Kuo, J.-C., Tsai, H.-C., et al., Bidirectional signals transduced by DAPK?ERK interaction promote the apoptotic effect of DAPK. *EMBO J.* 2005, 24, 294–304.

- [70] Lee, Y.-R., Yuan, W.-C., Ho, H.-C., Chen, C.-H., et al., The Cullin 3 substrate adaptor KLHL20 mediates DAPK ubiquitination to control interferon responses. *EMBO J.* 2010, 29, 1748–61.
- [71] Raveh, T., Berissi, H., Eisenstein, M., Spivak, T., Kimchi, A., A functional genetic screen identifies regions at the C-terminal tail and death-domain of death-associated protein kinase that are critical for its proapoptotic activity. *Proc. Natl. Acad. Sci. U. S. A.* 2000, 97, 1572–7.
- [72] Lai, M.-Z., Chen, R.-H., Regulation of inflammation by DAPK. *Apoptosis* 2014, 19, 357–63.
- [73] Kuo, J.-C., Wang, W.-J., Yao, C.-C., Wu, P.-R., Chen, R.-H., The tumor suppressor DAPK inhibits cell motility by blocking the integrin-mediated polarity pathway. *J. Cell Biol.* 2006, 172, 619–631.
- [74] Tian, J.-H., Das, S., Sheng, Z.-H., Ca²⁺-dependent phosphorylation of syntaxin-1A by the death-associated protein (DAP) kinase regulates its interaction with Munc18. *J. Biol. Chem.* 2003, 278, 26265–26274.
- [75] Bialik, S., Kimchi, A., The DAP-kinase interactome. *Apoptosis* 2014, 19, 316–328.
- [76] Green, D.R., Llambi, F., Cell Death Signaling. *Cold Spring Harb. Perspect. Biol.* 2015, 7.
- [77] Galluzzi, L., Vitale, I., Aaronson, S.A., Abrams, J.M., et al., Molecular mechanisms of cell death: recommendations of the Nomenclature Committee on Cell Death 2018. *Cell Death Differ.* 2018, 25, 486–541.
- [78] Cohen, O., Inbal, B., Kissil, J.L., Raveh, T., et al., DAP-kinase participates in TNF-alpha- and Fas-induced apoptosis and its function requires the death domain. *J. Cell Biol.* 1999, 146, 141–148.
- [79] Jang, C.-W., Chen, C.-H., Chen, C.-C., Chen, J., et al., TGF-beta induces apoptosis through Smad-mediated expression of DAP-kinase. *Nat. Cell Biol.* 2002, 4, 51–58.
- [80] Pelled, D., Raveh, T., Riebeling, C., Fridkin, M., et al., Death-associated protein (DAP) kinase plays a central role in ceramide-induced apoptosis in cultured hippocampal neurons. *J. Biol. Chem.* 2002, 277, 1957–1961.
- [81] Raveh, T., Droguett, G., Horwitz, M.S., DePinho, R. A., Kimchi, A., DAP kinase activates a p19ARF/p53-mediated apoptotic checkpoint to suppress oncogenic transformation. *Nat. Cell Biol.* 2001, 3, 1–7.
- [82] Shamloo, M., Soriano, L., Wieloch, T., Nikolich, K., et al., Death-associated protein kinase is activated by dephosphorylation in response to cerebral ischemia. *J. Biol. Chem.* 2005, 280, 42290–42299.
- [83] Inbal, B., Bialik, S., Sabanay, I., Shani, G., Kimchi, A., DAP kinase and DRP-1 mediate membrane blebbing and the formation of autophagic vesicles during programmed cell death. *J. Cell Biol.* 2002, 157, 455–468.

- [84] Zalckvar, E., Berissi, H., Mizrachy, L., Idelchuk, Y., et al., DAP-kinase-mediated phosphorylation on the BH3 domain of beclin 1 promotes dissociation of beclin 1 from Bcl-XL and induction of autophagy. *EMBO Rep.* 2009, 10, 285–292.
- [85] Eisenberg-Lerner, a, Kimchi, a, DAP kinase regulates JNK signaling by binding and activating protein kinase D under oxidative stress. *Cell Death Differ.* 2007, 14, 1908–1915.
- [86] Wang, W.J., Kuo, J.C., Yao, C.C., Chen, R.H., DAP-kinase induces apoptosis by suppressing integrin activity and disrupting matrix survival signals. *J. Cell Biol.* 2002, 159, 169–179.
- [87] Bialik, S., Berissi, H., Kimchi, A., A high throughput proteomics screen identifies novel substrates of death-associated protein kinase. *Mol. Cell. Proteomics* 2008, 7, 1089–1098.
- [88] Tu, W., Xu, X., Peng, L., Zhong, X., et al., DAPK1 Interaction with NMDA Receptor NR2B Subunits Mediates Brain Damage in Stroke. *Cell* 2010, 140, 1–24.
- [89] Pei, L., Shang, Y., Jin, H., Wang, S., et al., DAPK1-p53 interaction converges necrotic and apoptotic pathways of ischemic neuronal death. *J. Neurosci.* 2014, 34, 6546–56.
- [90] Houle, F., Poirier, A., Dumaresq, J., Huot, J., DAP kinase mediates the phosphorylation of tropomyosin-1 downstream of the ERK pathway, which regulates the formation of stress fibers in response to oxidative stress. *J. Cell Sci.* 2007, 120, 3666–3677.
- [91] Huang, Y., Chen, L., Guo, L., Hupp, T.R., Lin, Y., Evaluating DAPK as a therapeutic target. *Apoptosis* 2014, 19, 371–86.
- [92] Yang, B., Yang, Z.-G., Gao, B., Shao, G.-G., Li, G.-H., 5-Aza-CdR can reverse gefitinib resistance caused by DAPK gene promoter methylation in lung adenocarcinoma cells. *Int. J. Clin. Exp. Pathol.* 2015, 8, 12961–12966.
- [93] Ren, X., Li, H., Song, X., Wu, Y., Liu, Y., 5-Azacytidine treatment induces demethylation of DAPK1 and MGMT genes and inhibits growth in canine mammary gland tumor cells. *Onco. Targets. Ther.* 2018, 11, 2805–2813.
- [94] Nair, S., Hagberg, H., Krishnamurthy, R., Thornton, C., Mallard, C., Death associated protein kinases: Molecular structure and brain injury. *Int. J. Mol. Sci.* 2013, 14, 13858–13872.
- [95] Kim, B.M., You, M.-H., Chen, C.-H., Lee, S., et al., Death-associated protein kinase 1 has a critical role in aberrant tau protein regulation and function. *Cell Death Dis.* 2014, 5, e1237.
- [96] Kim, B.M., You, M.-H., Chen, C.-H., Suh, J., et al., Inhibition of death-associated protein kinase 1 attenuates the phosphorylation and amyloidogenic processing of amyloid precursor protein. *Hum. Mol. Genet.* 2016, 25, 2498–2513.
- [97] Henshall, D.C., Schindler, C.K., So, N.K., Lan, J.-Q., et al., Death-associated protein kinase expression in human temporal lobe epilepsy. *Ann. Neurol.* 2004, 55, 485–494.

- [98] Byerly, L., Cassada, R.C., Russell, R.L., The life cycle of the nematode *Caenorhabditis elegans*. I. Wild-type growth and reproduction. *Dev. Biol.* 1976, 51, 23–33.
- [99] Tissenbaum, H.A., Using *C. elegans* for aging research. *Invertebr. Reprod. Dev.* 2015, 59, 59–63.
- [100] Corsi, A.K., A biochemist's guide to *Caenorhabditis elegans*. *Anal. Biochem.* 2006, 359, 1–17.
- [101] Sulston, J.E., Horvitz, H.R., Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev. Biol.* 1977, 56, 110–156.
- [102] Lai, C.H., Chou, C.Y., Ch'ang, L.Y., Liu, C.S., Lin, W., Identification of novel human genes evolutionarily conserved in *Caenorhabditis elegans* by comparative proteomics. *Genome Res.* 2000, 10, 703–713.
- [103] Harris, T.W., Arnaboldi, V., Cain, S., Chan, J., et al., WormBase: a modern Model Organism Information Resource. *Nucleic Acids Res.* 2019.
- [104] Snider, J., Kotlyar, M., Saraon, P., Yao, Z., et al., Fundamentals of protein interaction network mapping. *Mol. Syst. Biol.* 2015, 11, 848.
- [105] Wilson, L.J., Linley, A., Hammond, D.E., Hood, F.E., et al., New Perspectives, Opportunities, and Challenges in Exploring the Human Protein Kinome. *Cancer Res.* 2018, 78, 15–29.
- [106] Gerwert, K., Mann, D., Kotting, C., Common mechanisms of catalysis in small and heterotrimeric GTPases and their respective GAPs. *Biol. Chem.* 2017, 398, 523–533.
- [107] Manzoni, C., Denny, P., Lovering, R.C., Lewis, P.A., Computational analysis of the LRRK2 interactome. *PeerJ* 2015, 3, e778.
- [108] Porras, P., Duesbury, M., Fabregat, A., Ueffing, M., et al., A visual review of the interactome of LRRK2: Using deep-curated molecular interaction data to represent biology. *Proteomics* 2015, 15, 1390–1404.
- [109] Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F.S.L., et al., PSICQUIC and PSIScore: accessing and scoring molecular interactions. *Nat. Methods* 2011, 8, 528–529.
- [110] Breuer, K., Foroushani, A.K., Laird, M.R., Chen, C., et al., InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. *Nucleic Acids Res.* 2013, 41, D1228-33.
- [111] Licata, L., Briganti, L., Peluso, D., Perfetto, L., et al., MINT, the molecular interaction database: 2012 update. *Nucleic Acid Res.* 2012, 40, 857–861.
- [112] Mellacheruvu, D., Wright, Z., Couzens, A.L., Lambert, J.-P., et al., The CRAPome: a Contaminant Repository for Affinity Purification Mass Spectrometry Data. *Nat. Methods* 2013, 10, 730–736.

- [113] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., et al., Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 2003, 13, 2498–2504.
- [114] GTEx Consortium, The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 2013, 45, 580–585.
- [115] Reimand, J., Arak, T., Adler, P., Kolberg, L., et al., g:Profiler — a web server for functional interpretation of gene lists (2016 update). *Nucleic Acid Res.* 2016, 44, 83–89.
- [116] Huang, D.W., Sherman, B.T., Lempicki, R.A., Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 2009, 37, 1–13.
- [117] Laukens, K., Naulaerts, S., Berghe, W. Vanden, Bioinformatics approaches for the functional interpretation of protein lists: from ontology term enrichment to network analysis. *Proteomics* 2015, 15, 981–996.
- [118] Mi, H., Huang, X., Muruganujan, A., Tang, H., et al., PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acid Res.* 2017, 45, D183–D189.
- [119] Wang, J., Vasaikar, S., Shi, Z., Greer, M., Zhang, B., WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res.* 2017, 45, W130–W137.
- [120] Reimand, J., Kull, M., Peterson, H., Hansen, J., Vilo, J., g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* 2007, 35, W193–200.
- [121] Tomkins, J.E., Ferrari, R., Vavouraki, N., Hardy, J., et al., PINOT: An Intuitive Resource for Integrating Protein-Protein Interactions. *bioRxiv* 2019.
- [122] Cookson, M.R., LRRK2 Pathways Leading to Neurodegeneration. *Curr. Neurol. Neurosci. Rep.* 2015, 15, 42.
- [123] Mouasni, S., Tourneur, L., FADD at the Crossroads between Cancer and Inflammation. *Trends Immunol.* 2018, 39, 1036–1053.
- [124] Zhuang, H., Wang, X., Zha, D., Gan, Z., et al., FADD is a key regulator of lipid metabolism. *EMBO Mol. Med.* 2016.
- [125] Hartman, M.A., Finan, D., Sivaramakrishnan, S., Spudich, J.A., Principles of unconventional myosin function and targeting. *Annu. Rev. Cell Dev. Biol.* 2011, 27, 133–155.
- [126] Kabbage, M., Dickman, M.B., The BAG proteins: a ubiquitous family of chaperone regulators. *Cell. Mol. Life Sci.* 2008, 65, 1390–1402.

- [127] Wang, X., Guo, J., Fei, E., Mu, Y., et al., BAG5 Protects against Mitochondrial Oxidative Damage through Regulating PINK1 Degradation. *PLoS One* 2014, 9, e86276.
- [128] Stricher, F., Macri, C., Ruff, M., Muller, S., HSPA8/HSC70 chaperone protein: structure, function, and chemical targeting. *Autophagy* 2013, 9, 1937–1954.
- [129] Petsalaki, E., Zachos, G., Clks 1, 2 and 4 prevent chromatin breakage by regulating the Aurora B-dependent abscission checkpoint. *Nat. Commun.* 2016, 7, 11451.
- [130] Carmena, M., Ruchaud, S., Earnshaw, W.C., Making the Auroras glow: regulation of Aurora A and B kinase function by interacting proteins. *Curr. Opin. Cell Biol.* 2009, 21, 796–805.
- [131] Nguyen, A.L., Drutovic, D., Vazquez, B.N., El Yakoubi, W., et al., Genetic Interactions between the Aurora Kinases Reveal New Requirements for AURKB and AURKC during Oocyte Meiosis. *Curr. Biol.* 2018, 28, 3458–3468.e5.
- [132] Uzor, S., Zorzou, P., Bowler, E., Porazinski, S., et al., Autoregulation of the human splice factor kinase CLK1 through exon skipping and intron retention. *Gene* 2018, 670, 46–54.
- [133] Shimizu, H., Nagamori, I., Yabuta, N., Nojima, H., GAK, a regulator of clathrin-mediated membrane traffic, also controls centrosome integrity and chromosome congression. *J. Cell Sci.* 2009, 122, 3145–3152.
- [134] Roosen, D.A., Blauwendraat, C., Cookson, M.R., Lewis, P.A., DNAJC proteins and pathways to parkinsonism. *FEBS J.* 2019, 286, 3080–3094.
- [135] Prunier, C., Prudent, R., Kapur, R., Sadoul, K., Lafanechere, L., LIM kinases: cofilin and beyond. *Oncotarget* 2017, 8, 41749–41763.
- [136] Duan, X., Zhang, Y., Chen, K.-L., Zhang, H.-L., et al., The small GTPase RhoA regulates the LIMK1/2-cofilin pathway to modulate cytoskeletal dynamics in oocyte meiosis. *J. Cell. Physiol.* 2018, 233, 6088–6097.
- [137] Chi, H., Sarkisian, M.R., Rakic, P., Flavell, R.A., Loss of mitogen-activated protein kinase kinase kinase 4 (MEKK4) results in enhanced apoptosis and defective neural tube development. *Proc. Natl. Acad. Sci. U. S. A.* 2005, 102, 3846–3851.
- [138] Craig, E.A., Stevens, M. V., Vaillancourt, R.R., Camenisch, T.D., MAP3Ks as central regulators of cell fate during development. *Dev. Dyn.* 2008, 237, 3102–3114.
- [139] Chen, Y., Chen, C.-F., Riley, D.J., Chen, P.-L., Nek1 kinase functions in DNA damage response and checkpoint control through a pathway independent of ATM and ATR. *Cell Cycle* 2011, 10, 655–663.
- [140] Fry, A.M., O'Regan, L., Sabir, S.R., Bayliss, R., Cell cycle regulation by the NEK family of protein kinases. *J. Cell Sci.* 2012, 125, 4423–4433.

- [141] Brieno-Enriquez, M.A., Moak, S.L., Holloway, J.K., Cohen, P.E., NIMA-related kinase 1 (NEK1) regulates meiosis I spindle assembly by altering the balance between alpha-Adducin and Myosin X. *PLoS One* 2017, 12, e0185780.
- [142] Shalom, O., Shalva, N., Altschuler, Y., Motro, B., The mammalian Nek1 kinase is involved in primary cilium formation. *FEBS Lett.* 2008, 582, 1465–1470.
- [143] Guo, L., Wang, Z.-B., Wang, H.-H., Zhang, T., et al., Nek11 regulates asymmetric cell division during mouse oocyte meiotic maturation. *Biochem. Biophys. Res. Commun.* 2016, 474, 667–672.
- [144] Bainbridge, T.W., DeAlmeida, V.I., Izrael-Tomasevic, A., Chalouni, C., et al., Evolutionary divergence in the catalytic activity of the CAM-1, ROR1 and ROR2 kinase domains. *PLoS One* 2014, 9, e102695.
- [145] Diaz-Horta, O., Abad, C., Sennaroglu, L., Foster, J. 2nd, et al., ROR1 is essential for proper innervation of auditory hair cells and hearing in humans and mice. *Proc. Natl. Acad. Sci. U. S. A.* 2016, 113, 5993–5998.
- [146] Sutt, S., Cansby, E., Paul, A., Amrutkar, M., et al., STK25 regulates oxidative capacity and metabolic efficiency in adipose tissue. *J. Endocrinol.* 2018, 238, 187–202.
- [147] Thompson, B.J., Sahai, E., MST kinases in development and disease. *J. Cell Biol.* 2015, 210, 871–882.
- [148] Roder, J., Linstid, B., Oliveira, C., Improving the power of gene set enrichment analyses. *BMC Bioinformatics* 2019, 20, 257.
- [149] Steger, M., Tonelli, F., Ito, G., Davies, P., et al., Phosphoproteomics reveals that Parkinson's disease kinase LRRK2 regulates a subset of Rab GTPases. *Elife* 2016, 5.
- [150] Sanna, G., Del Giudice, M.G., Crosio, C., Iaccarino, C., LRRK2 and vesicle trafficking. *Biochem. Soc. Trans.* 2012, 40, 1117–1122.
- [151] Pfeffer, S.R., LRRK2 and Rab GTPases. *Biochem. Soc. Trans.* 2018, 46, 1707–1712.
- [152] Biosa, A., Trancikova, A., Civiero, L., Glauser, L., et al., GTPase activity regulates kinase activity and cellular phenotypes of Parkinson's disease-associated LRRK2. *Hum. Mol. Genet.* 2013, 22, 1140–1156.
- [153] So, J., Pasculescu, A., Dai, A.Y., Williton, K., et al., Integrative analysis of kinase networks in TRAIL-induced apoptosis provides a source of potential targets for combination therapy. *Sci. Signal.* 2015, 8, rs3.
- [154] Ko, H.S., Bailey, R., Smith, W.W., Liu, Z., et al., CHIP regulates leucine-rich repeat kinase-2 ubiquitination, degradation, and toxicity. *Proc. Natl. Acad. Sci. U. S. A.* 2009, 106, 2897–2902.
- [155] Ding, X., Goldberg, M.S., Regulation of LRRK2 stability by the E3 ubiquitin ligase CHIP. *PLoS One* 2009, 4, e5949.

- [156] Tomczak, A., Mortensen, J.M., Winnenburger, R., Liu, C., et al., Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. *Sci. Rep.* 2018, 8, 5115.
- [157] Rossi, G., Redaelli, V., Contiero, P., Fabiano, S., et al., Tau Mutations Serve as a Novel Risk Factor for Cancer. *Cancer Res.* 2018, 78, 3731–3739.
- [158] Alanis-Lobato, G., Andrade-Navarro, M.A., Schaefer, M.H., HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res.* 2017, 45, D408–D414.
- [159] Hu, Y., Vinayagam, A., Nand, A., Comjean, A., et al., Molecular Interaction Search Tool (MIST): an integrated resource for mining gene and protein interaction data. *Nucleic Acids Res.* 2018, 46, D567–D574.
- [160] Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., et al., STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 2019, 47, D607–D613.
- [161] Chuang, M., Chisholm, A.D., Insights into the functions of the death associated protein kinases from *C. elegans* and other invertebrates. *Apoptosis* 2014, 19, 392–397.
- [162] van Egmond, W.N., van Haastert, P.J.M., Characterization of the Roco protein family in *Dictyostelium discoideum*. *Eukaryot. Cell* 2010, 9, 751–761.
- [163] Cnops, G., Neyt, P., Raes, J., Petrarulo, M., et al., The TORNADO1 and TORNADO2 genes function in several patterning processes during early leaf development in *Arabidopsis thaliana*. *Plant Cell* 2006, 18, 852–866.
- [164] Kang, C., You, Y., Avery, L., Dual roles of autophagy in the survival of *Caenorhabditis elegans* during starvation. *Genes Dev.* 2007, 21, 2161–2171.
- [165] Tong, A., Lynn, G., Ngo, V., Wong, D., et al., Negative regulation of *Caenorhabditis elegans* epidermal damage responses by death-associated protein kinase. *Proc. Natl. Acad. Sci. U. S. A.* 2009, 106, 1457–1461.
- [166] Chuang, M., Hsiao, T.I., Tong, A., Xu, S., Chisholm, A.D., DAPK interacts with Patronin and the microtubule cytoskeleton in epidermal development and wound repair. *Elife* 2016, 5.
- [167] Del Rosario, J.S., Feldmann, K.G., Ahmed, T., Amjad, U., et al., Death Associated Protein Kinase (DAPK) -mediated neurodegenerative mechanisms in nematode excitotoxicity. *BMC Neurosci.* 2015, 16, 25.
- [168] Lai, T.W., Zhang, S., Wang, Y.T., Excitotoxicity and stroke: identifying novel targets for neuroprotection. *Prog. Neurobiol.* 2014, 115, 157–188.

- [169] Lee, T.H., Chen, C.-H., Suizu, F., Huang, P., et al., Death-associated protein kinase 1 phosphorylates Pin1 and inhibits its prolyl isomerase activity and cellular function. *Mol. Cell* 2011, 42, 147–59.
- [170] Farboud, B., Targeted genome editing in *Caenorhabditis elegans* using CRISPR/Cas9. *Wiley Interdiscip. Rev. Dev. Biol.* 2017, 6.
- [171] Cho, J. Grove, C.A. Van Auken, K. Chan, J. Gao, S. Sternberg, P.W., 2018 Update on Protein-Protein Interaction Data in WormBase. *microPublication Biol.* 2018, 10.17912/micropub.biology.000074.
- [172] Kim, W., Underwood, R.S., Greenwald, I., Shaye, D.D., OrthoList 2: A New Comparative Genomic Analysis of Human and *Caenorhabditis elegans* Genes. *Genetics* 2018, 210, 445–461.
- [173] Reddy Chichili, V.P., Kumar, V., Sivaraman, J., Linkers in the structural biology of protein-protein interactions. *Protein Sci.* 2013, 22, 153–167.
- [174] Mitchell, D.H., Stiles, J.W., Santelli, J., Sanadi, D.R., Synchronous growth and aging of *Caenorhabditis elegans* in the presence of fluorodeoxyuridine. *J. Gerontol.* 1979, 34, 28–36.
- [175] Kunitz, M., Crystalline desoxyribonuclease; isolation and general properties; spectrophotometric method for the measurement of desoxyribonuclease activity. *J. Gen. Physiol.* 1950, 33, 349–362.
- [176] Hu, Y., Flockhart, I., Vinayagam, A., Bergwitz, C., et al., An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics* 2011, 12, 357.
- [177] Nichio, B.T.L., Marchaukoski, J.N., Raittz, R.T., New Tools in Orthology Analysis: A Brief Review of Promising Perspectives. *Front. Genet.* 2017, 8, 165.
- [178] Li, S., Armstrong, C.M., Bertin, N., Ge, H., et al., A map of the interactome network of the metazoan *C. elegans*. *Science (80-.)*. 2004, 303, 540–543.
- [179] Huang, X.-T., Zhu, Y., Chan, L.L.H., Zhao, Z., Yan, H., An integrative *C. elegans* protein-protein interaction network with reliability assessment based on a probabilistic graphical model. *Mol. Biosyst.* 2016, 12, 85–92.
- [180] Qadota, H., Matsunaga, Y., Bagchi, P., Lange, K.I., et al., Protein phosphatase 2A is crucial for sarcomere organization in *Caenorhabditis elegans* striated muscle. *Mol. Biol. Cell* 2018, 29, 2084–2097.
- [181] Reichman, R., Shi, Z., Malone, R., Smolikove, S., Mitotic and Meiotic Functions for the SUMOylation Pathway in the *Caenorhabditis elegans* Germline. *Genetics* 2018, 208, 1421–1441.

- [182] Busch, J.M.C., Erat, M.C., Blank, I.D., Musgaard, M., et al., A dynamically interacting flexible loop assists oligomerisation of the *Caenorhabditis elegans* centriolar protein SAS-6. *Sci. Rep.* 2019, 9, 3526.
- [183] Lange, K.I., Suleman, A., Srayko, M., Kinetochore Recruitment of the Spindle and Kinetochore-Associated (Ska) Complex Is Regulated by Centrosomal PP2A in *Caenorhabditis elegans*. *Genetics* 2019, 212, 509–522.
- [184] Shen, X., Valencia, C.A., Gao, W., Cotten, S.W., et al., Ca(2+)/Calmodulin-binding proteins from the *C. elegans* proteome. *Cell Calcium* 2008, 43, 444–456.
- [185] Karabinos, A., Bussing, I., Schulze, E., Wang, J., et al., Functional analysis of the single calmodulin gene in the nematode *Caenorhabditis elegans* by RNA interference and 4-D microscopy. *Eur. J. Cell Biol.* 2003, 82, 557–563.
- [186] Sonnichsen, B., Koski, L.B., Walsh, A., Marschall, P., et al., Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. *Nature* 2005, 434, 462–469.
- [187] van der Voet, M., Berends, C.W.H., Perreault, A., Nguyen-Ngoc, T., et al., NuMA-related LIN-5, ASPM-1, calmodulin and dynein promote meiotic spindle rotation independently of cortical LIN-5/GPR/Galpha. *Nat. Cell Biol.* 2009, 11, 269–277.
- [188] Unhavaithaya, Y., Shin, T.H., Miliaras, N., Lee, J., et al., MEP-1 and a homolog of the NURD complex component Mi-2 act together to maintain germline-soma distinctions in *C. elegans*. *Cell* 2002, 111, 991–1002.
- [189] Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., et al., Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* 2003, 421, 231–237.
- [190] Leight, E.R., Glossip, D., Kornfeld, K., Sumoylation of LIN-1 promotes transcriptional repression and inhibition of vulval cell fates. *Development* 2005, 132, 1047–1056.
- [191] Takasaki, T., Liu, Z., Habara, Y., Nishiwaki, K., et al., MRG-1, an autosome-associated protein, silences X-linked genes and protects germline immortality in *Caenorhabditis elegans*. *Development* 2007, 134, 757–767.
- [192] Wang, Y., Gracheva, E.O., Richmond, J., Kawano, T., et al., The C2H2 zinc-finger protein SYD-9 is a putative posttranscriptional regulator for synaptic transmission. *Proc. Natl. Acad. Sci. U. S. A.* 2006, 103, 10450–10455.
- [193] Axang, C., Rauthan, M., Hall, D.H., Pilon, M., Developmental genetics of the *C. elegans* pharyngeal neurons NSML and NSMR. *BMC Dev. Biol.* 2008, 8, 38.
- [194] Ogura, K.-I., Goshima, Y., The autophagy-related kinase UNC-51 and its binding partner UNC-14 regulate the subcellular localization of the Netrin receptor UNC-5 in *Caenorhabditis elegans*. *Development* 2006, 133, 3441–3450.

- [195] Sakamoto, R., Byrd, D.T., Brown, H.M., Hisamoto, N., et al., The *Caenorhabditis elegans* UNC-14 RUN domain protein binds to the kinesin-1 and UNC-16 complex and regulates synaptic vesicle localization. *Mol. Biol. Cell* 2005, 16, 483–496.
- [196] Lai, T., Garriga, G., The conserved kinase UNC-51 acts with VAB-8 and UNC-14 to regulate axon outgrowth in *C. elegans*. *Development* 2004, 131, 5991–6000.
- [197] Reece-Hoyes, J.S., Pons, C., Diallo, A., Mori, A., et al., Extensive rewiring and complex evolutionary dynamics in a *C. elegans* multiparameter transcription factor network. *Mol. Cell* 2013, 51, 116–127.
- [198] Ogura, K., Shirakawa, M., Barnes, T.M., Hekimi, S., Ohshima, Y., The UNC-14 protein required for axonal elongation and guidance in *Caenorhabditis elegans* interacts with the serine/threonine kinase UNC-51. *Genes Dev.* 1997, 11, 1801–1811.
- [199] Callebaut, I., de Gunzburg, J., Goud, B., Mornon, J.P., RUN domains: a new family of domains involved in Ras-like GTPase signaling. *Trends Biochem. Sci.* 2001, 26, 79–83.
- [200] Hopp, T.P., Prickett, K.S., Price, V.L., Libby, R.T., et al., A Short Polypeptide Marker Sequence Useful for Recombinant Protein Identification and Purification. *Nat. Biotechnol.* 1988, 6, 1204–1210.
- [201] Einhauer, A., Jungbauer, A., The FLAG peptide, a versatile fusion tag for the purification of recombinant proteins. *J. Biochem. Biophys. Methods* 2001, 49, 455–465.
- [202] Lobbstaël, E., Reumers, V., Ibrahim, A., Paesen, K., et al., Immunohistochemical detection of transgene expression in the brain using small epitope tags. *BMC Biotechnol.* 2010, 10, 16.
- [203] Lund, J., Tedesco, P., Duke, K., Wang, J., et al., Transcriptional profile of aging in *C. elegans*. *Curr. Biol.* 2002, 12, 1566–1573.
- [204] Vlachakis, D., Labrou, N.E., Iliopoulos, C., Hardy, J., et al., Insights into the Influence of Specific Splicing Events on the Structural Organization of LRRK2. *Int. J. Mol. Sci.* 2018, 19.
- [205] Tong, A., Moseley, S., Goncharov, A., Jin, Y., Chisholm, A., *mor-3* encodes a *C. elegans* death-associated protein kinase (DAP kinase) that functions in post-embryonic epidermal morphogenesis. *Meet. Abstr. West Coast Worm Meet.* 2004.
- [206] Hunt-Newbury, R., Viveiros, R., Johnsen, R., Mah, A., et al., High-throughput in vivo analysis of gene expression in *Caenorhabditis elegans*. *PLoS Biol.* 2007, 5, e237.
- [207] Kaletsky, R., Yao, V., Williams, A., Runnels, A.M., et al., Transcriptome analysis of adult *Caenorhabditis elegans* cells reveals tissue-specific gene and isoform expression. *PLoS Genet.* 2018, 14, e1007559.
- [208] Celnik, S.E., Dillon, L.A.L., Gerstein, M.B., Gunsalus, K.C., et al., Unlocking the secrets of the genome. *Nature* 2009, 459, 927–930.

- [209] Moerman, D.G., Barstead, R.J., Towards a mutation in every gene in *Caenorhabditis elegans*. *Brief. Funct. Genomic. Proteomic.* 2008, 7, 195–204.
- [210] Thompson, O., Edgley, M., Strasbourger, P., Flibotte, S., et al., The million mutation project: a new approach to genetics in *Caenorhabditis elegans*. *Genome Res.* 2013, 23, 1749–1762.
- [211] Friedland, A.E., Tzur, Y.B., Esvelt, K.M., Colaiacovo, M.P., et al., Heritable genome editing in *C. elegans* via a CRISPR-Cas9 system. *Nat. Methods* 2013, 10, 741–743.
- [212] Taymans, J.-M., Vancaenenbroeck, R., Ollikainen, P., Beilina, A., et al., LRRK2 kinase activity is dependent on LRRK2 GTP binding capacity but independent of LRRK2 GTP binding. *PLoS One* 2011, 6, e23207.
- [213] Daniels, V., Vancaenenbroeck, R., Law, B.M.H., Greggio, E., et al., Insight into the mode of action of the LRRK2 Y1699C pathogenic mutant. *J. Neurochem.* 2011, 116, 304–315.
- [214] Leiser, S.F., Jafari, G., Primitivo, M., Sutphin, G.L., et al., Age-associated vulval integrity is an important marker of nematode healthspan. *Age (Omaha)*. 2016, 38, 419–431.
- [215] Neff, M.M., Neff, J.D., Chory, J., Pepper, A.E., dCAPS, a simple technique for the genetic analysis of single nucleotide polymorphisms: experimental applications in *Arabidopsis thaliana* genetics. *Plant J.* 1998, 14, 387–392.
- [216] Lorenz, T.C., Polymerase chain reaction: basic protocol plus troubleshooting and optimization strategies. *JoVE* 2012, e3998.
- [217] Wilson, M.A., Shukitt-Hale, B., Kalt, W., Ingram, D.K., et al., Blueberry polyphenols increase lifespan and thermotolerance in *Caenorhabditis elegans*. *Aging Cell* 2006, 5, 59–68.
- [218] Gems, D., Riddle, D.L., Genetic, behavioral and environmental determinants of male longevity in *Caenorhabditis elegans*. *Genetics* 2000, 154, 1597–1610.
- [219] Garigan, D., Hsu, A.-L., Fraser, A.G., Kamath, R.S., et al., Genetic analysis of tissue aging in *Caenorhabditis elegans*: a role for heat-shock factor and bacterial proliferation. *Genetics* 2002, 161, 1101–1112.
- [220] Pujol, N., Cypowyj, S., Ziegler, K., Millet, A., et al., Distinct innate immune responses to infection and wounding in the *C. elegans* epidermis. *Curr. Biol.* 2008, 18, 481–489.
- [221] Youngman, M.J., Rogers, Z.N., Kim, D.H., A decline in p38 MAPK signaling underlies immunosenescence in *Caenorhabditis elegans*. *PLoS Genet.* 2011, 7, e1002082.
- [222] McGee, M.D., Weber, D., Day, N., Vitelli, C., et al., Loss of intestinal nuclei and intestinal integrity in aging *C. elegans*. *Aging Cell* 2011, 10, 699–710.
- [223] Nakamura, S., Yoshimori, T., Autophagy and Longevity. *Mol. Cells* 2018, 41, 65–72.
- [224] Conradt, B., Wu, Y.-C., Xue, D., Programmed Cell Death During *Caenorhabditis elegans* Development. *Genetics* 2016, 203, 1533–1562.

- [225] Zhang, X.-H., Tee, L.Y., Wang, X.-G., Huang, Q.-S., Yang, S.-H., Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. *Mol. Ther. - Nucleic Acids* 2015, 4, e264.
- [226] Kosicki, M., Tomberg, K., Bradley, A., Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.* 2018, 36, 765–771.
- [227] Taymans, J.-M., Chartier-Harlin, M.-C., In silico and Wet Bench Interactomics Sheds Light on the Similarities and Differences between Human ROCO Proteins. *Proteomics* 2018, 18, e1800103.
- [228] Dikic, I., Elazar, Z., Mechanism and medical implications of mammalian autophagy. *Nat. Rev. Mol. Cell Biol.* 2018, 19, 349–364.
- [229] Farag, A.K., Roh, E.J., Death-associated protein kinase (DAPK) family modulators: Current and future therapeutic outcomes. *Med. Res. Rev.* 2019, 39, 349–385.
- [230] Li, X., Wang, Q.J., Pan, N., Lee, S., et al., Phosphorylation-dependent 14-3-3 binding to LRRK2 is impaired by common mutations of familial Parkinson’s disease. *PLoS One* 2011, 6, e17153.
- [231] McClure, C.D., Southall, T.D., Getting Down to Specifics: Profiling Gene Expression and Protein-DNA Interactions in a Cell Type-Specific Manner. *Adv. Genet.* 2015, 91, 103–151.
- [232] Hernandez-Alias, X., Benisty, H., Schaefer, M.H., Serrano, L., Translational efficiency across healthy and tumor tissues is proliferation-related. *bioRxiv* 2019.
- [233] Tourneur, L., Chiocchia, G., FADD: a regulator of life and death. *Trends Immunol.* 2010, 31, 260–269.
- [234] O’Reilly, L.P., Luke, C.J., Perlmutter, D.H., Silverman, G.A., Pak, S.C., *C. elegans* in high-throughput drug discovery. *Adv. Drug Deliv. Rev.* 2014, 69–70, 247–253.

APPENDIX

Supporting Files Conferences and Publications

7.1 Supporting files

Supporting files corresponding to analyses in Chapter Two are available at:

<https://www.dropbox.com/sh/6rxxf2v0grl5s3o/AADNTUFC8YtOUxEuMmz3LFSua?dl=0>

Programming scripts which underlie data processing in PINOT are available at:

https://www.reading.ac.uk/bioinf/downloads/PINOT_scripts/

7.2 Conference Presentations

Tomkins, J. E., Manzoni, C., Dihanich, S., Beilina, A., Cookson, M. R., and Lewis, P. A. An insight into the ROCO protein interactome. *LRRK2: Ten Years Along the Road to Therapeutic Intervention*. 2016. Henley-on-Thames, UK. (Poster presentation)

Tomkins, J. E., Dihanich, S., Beilina, A., Cookson, M. R., Lewis, P. A., and Manzoni, C. An insight into the ROCO protein interactome. *Omics Strategies to Study the Proteome*. 2017. Breckenridge, CO, USA. (Poster presentation)

Tomkins, J. E., Dihanich, S., Beilina, A., Ferrari, R., Ilacqua, N., Cookson, M. R., Lewis, P. A., and Manzoni, C. Comparative Protein Interaction Network Analysis of the Human ROCO Proteins. *Integrating Systems Biology: From Networks to Mechanisms to Models*. 2018. Heidelberg, Germany. (Poster presentation)

Tomkins, J. E., Dihanich, S., Beilina, A., Ferrari, R., Ilacqua, N., Cookson, M. R., Lewis, P. A., and Manzoni, C. Comparative Protein Interaction Network Analysis of the Human ROCO Proteins. *Biennial International LRRK2 Meeting*. 2018. Padua, Italy. (Poster presentation)

Tomkins, J. E., Ferrari, R., Vavouraki, N., Hardy, J., Lovering, R. C., Lewis, P. A., McGuffin, L. J., and Manzoni, C. PINOT: A Transparent Data Mining Pipeline for Mapping Confidence-Weighted Protein-Protein Interaction Networks. *Network Inference in Biology and Disease*. 2019. Pozzuoli, Italy. (Short talk and poster presentation)

Tomkins, J. E., Ferrari, R., Vavouraki, N., Hardy, J., Lovering, R. C., Lewis, P. A., McGuffin, L. J., and Manzoni, C. PINOT: A Novel Data Mining Platform for Mapping Confidence-Weighted Protein-Protein Interaction Networks for *Caenorhabditis elegans*. *UK Worm Meeting 2019*. 2019. London, UK. (Poster presentation)

7.3 Publications

Lubbe, S.J., Escott-Price, V., Gibbs, J.R., Nalls, M.A., et al., Additional rare variant analysis in Parkinson's disease cases with and without known pathogenic mutations: evidence for oligogenic inheritance. *Hum. Mol. Genet.* 2016, 25, 5483–5489.

*Tomkins, J.E., Dihanich, S., Beilina, A., Ferrari, R., et al., Comparative Protein Interaction Network Analysis Identifies Shared and Distinct Functions for the Human ROCO Proteins. *Proteomics* 2018, 18, e1700444.

Ferrari, R., Kia, D.A., Tomkins, J.E., Hardy, J., et al., Stratification of candidate genes for Parkinson's disease using weighted protein-protein interaction network analysis. *BMC Genomics* 2018, 19, 452.

Leksmono, C.S., Manzoni, C., Tomkins, J.E., Lucchesi, W., et al., Measuring Lactase Enzymatic Activity in the Teaching Lab. *JoVE* 2018.

Mitchell, R., Mellows, B., Sheard, J., Antonioli, M., et al., Secretome of adipose-derived mesenchymal stem cells promotes skeletal muscle regeneration through synergistic action of extracellular vesicle cargo and soluble proteins. *Stem Cell Res. Ther.* 2019, 10, 116.

*Tomkins, J.E., Ferrari, R., Vavouraki, N., Hardy, J., et al., PINOT: An Intuitive Resource for Integrating Protein-Protein Interactions. *bioRxiv* 2019.

Under review at *Cell Communication and Signaling*

*publications which relate directly to this thesis (provided in full on the following pages)

Comparative Protein Interaction Network Analysis Identifies Shared and Distinct Functions for the Human ROCO Proteins

James E. Tomkins, Sybille Dihanich, Alexandra Beilina, Raffaele Ferrari, Nicolò Ilacqua, Mark R. Cookson, Patrick A. Lewis, and Claudia Manzoni*

Signal transduction cascades governed by kinases and GTPases are a critical component of the command and control of cellular processes, with the precise outcome partly determined by direct protein–protein interactions (PPIs). Here, we use the human ROCO proteins as a model for investigating PPI signaling events—taking advantage of the unique dual kinase/GTPase activities and scaffolding properties of these multidomain proteins. PPI networks are reported that encompass the human ROCO proteins, developed using two complementary approaches. First, using the recently developed weighted PPI network analysis (WPPINA) pipeline, a confidence-weighted overview of validated ROCO protein interactors is obtained from peer-reviewed literature. Second, novel ROCO PPIs are assessed experimentally via protein microarray screens. The networks derived from these orthologous approaches are compared to identify common elements within the ROCO protein interactome; functional enrichment analysis of this common core of the network identified stress response and cell projection organization as shared functions within this protein family. Despite the presence of these commonalities, the results suggest that many unique interactors and therefore some specialized cellular roles have evolved for different members of the ROCO proteins. Overall, this multi-approach strategy to increase the resolution of protein interaction networks represents a prototype for the utility of PPI data integration in understanding signaling biology.

understanding how they influence the physiology of biological processes is important in determining how subtle alterations in protein function may lead to disease. Since protein–protein interactions (PPIs) are central to these processes, and as interacting proteins are likely to be involved in the same or related pathway,^[1] searching for proteins that physically interact with each other represents a means to achieve deeper insight into the highly interconnected landscape of cellular functions. The importance of elucidating protein interactors within cell signaling events is illustrated in our understanding of the mTOR complexes,^[2] whereby the assembly of particular protein interactors differentially initiates a diverse range of functional pathways. The human ROCO protein family (Figure 1) consists of four multidomain cell signaling proteins, death-associated protein kinase 1 (DAPK1), leucine-rich repeat kinase 1 (LRRK1), leucine-rich repeat kinase 2 (LRRK2), and malignant fibrous histiocytoma amplified sequence 1 (MASL1 or MFHAS1), which are characterized by a tandem ROC (Ras of complex proteins)-COR (C-terminal of ROC) supra-domain.^[3] Although the ROCO proteins are defined by this conserved region, the domain topology surrounding the ROC-COR unit (which includes numerous protein interaction motifs) is diverse and dissimilar between ROCO

1. Introduction

The subcellular environment hosts a dynamic network of molecular events that regulates cell homeostasis and coordinates signal transduction. Defining these regulatory mechanisms and

(C-terminal of ROC) supra-domain.^[3] Although the ROCO proteins are defined by this conserved region, the domain topology surrounding the ROC-COR unit (which includes numerous protein interaction motifs) is diverse and dissimilar between ROCO

J. E. Tomkins, Dr. P. A. Lewis, N. Ilacqua, Dr. C. Manzoni
 School of Pharmacy
 University of Reading
 Whiteknights Campus
 Reading, UK
 E-mail: c.manzoni@reading.ac.uk

See the associated Commentary by Jean-Marc Taymans,
<https://doi.org/10.1002/pmic.201800103>

© 2018 The Authors. *Proteomics* Published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/pmic.201700444

Dr. S. Dihanich, Dr. R. Ferrari, Dr. C. Manzoni, Dr. P. A. Lewis
 Department of Molecular Neuroscience
 UCL Institute of Neurology
 London, UK

Dr. A. Beilina, Dr. M. R. Cookson
 Laboratory of Neurogenetics
 National Institute on Aging
 National Institutes of Health
 Bethesda, USA

N. Ilacqua
 Department of Biology
 University of Padova
 Padova, Italy

Significance Statement

This research demonstrates the utility of the extensive collection of PPI data already in the public domain via peer-reviewed publication to complement novel PPI datasets, in order to identify similarities and differences in PPI and functional profiles of related multidomain proteins and prioritize interactors to pursue for validation in the laboratory. The human ROCO proteins are an attractive protein family for utilizing this approach since their primary structure consists of a conserved region flanked by a diverse range of PPI domains within a single open reading frame.

Our literature mining pipeline implemented in this analysis, WPPINA, ensures a wide coverage of reported PPIs from multiple data repositories which maximizes the usefulness of novel data integration, such as protein microarray as is used in this study. The significance of this strategy is that novel datasets are not just stand-alone results and can be interpreted in combination with decades of research into PPIs of particular proteins of interest, by adopting this straightforward approach to support further investigations.

proteins. Three of the four ROCO proteins (DAPK1, LRRK1, and LRRK2) also harbor active kinase domains in addition to the GT-Pase activity of the ROC domain, an arrangement that is exclusive to these three proteins, within the human proteome. The combination of multiple enzymatic activities coupled with a range of PPI domains within the same open reading frame positions the ROCO proteins as a unique protein family to investigate the functional commonalities and differences of structurally related proteins. The presence of several interaction domains within the primary structure of these proteins may reduce the requirement for adaptor proteins to complex with ROCO proteins. Therefore, direct interactors are likely to be functionally relevant effector proteins and hence the analysis of direct ROCO protein

interactors will provide important functional insight into this family of proteins. Thus, the human ROCO proteins are an attractive protein family to utilize as a model for PPI network analysis, to explore the link between PPI profiles and functional fates. This approach has been previously used for LRRK2 in isolation,^[4,5] the DAPK1 interactome has been reviewed,^[6] and the comparison between LRRK1 and LRRK2 has been attempted.^[7] However, the collective PPI network analysis of the entire human ROCO protein family is a novel contribution.

Despite their sequence similarity (Figure 1), the human ROCO proteins appear to be associated with different cellular processes. DAPK1 is linked to cell death pathways^[8–10] and is also involved in inflammation.^[11] LRRK1 has been associated with numerous distinct cellular mechanisms, which include EGFR trafficking,^[12] mitotic spindle orientation,^[13] and humoral immunity.^[14] LRRK2 has been implicated in a diverse range of cellular processes, including macroautophagy, cytoskeletal dynamics, and mitochondrial function.^[15] Finally MASL1, the least studied of the human ROCO proteins, has functional connections to macrophage polarization^[16,17] and erythropoiesis.^[18] These proteins also have disease relevance: DAPK1, LRRK1, and MASL1 in cancers,^[19–21] while mutations in LRRK2 are a common genetic contributor to familial Parkinson's Disease (PD)^[22] and LRRK2 has been associated with numerous other human diseases.^[23] However, significant gaps in our understanding of ROCO protein biology persist, which have implications for drug development in human disease.^[19,24] In addition, fundamental questions relating to why such similar proteins are differentially involved in health and disease, and how the complex enzymatic functions of these proteins fit with the biochemical regulation of cellular signaling pathways, remain to be addressed.

Since key components underlying the functional divergence evident between the ROCO proteins will reside in their proximal interactomes,^[7,23] we set out to investigate these interactomes using two orthologous approaches to determine PPI networks across the human ROCO protein family. We first used an in-house data mining approach which enabled

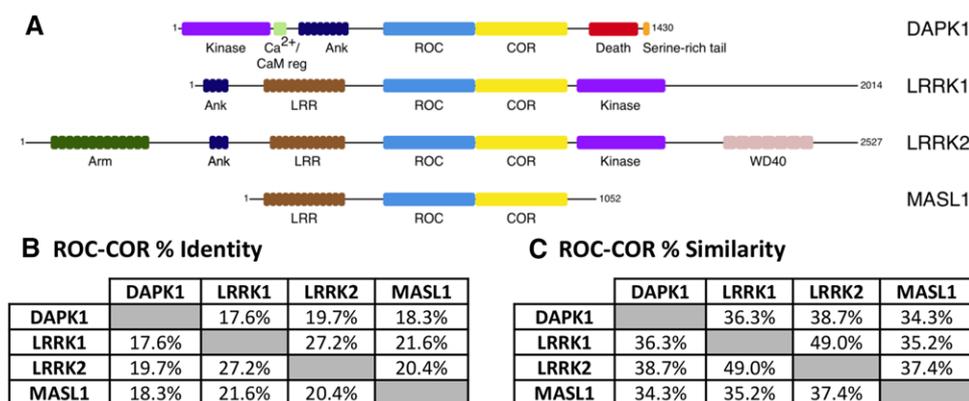


Figure 1. Domain topology of the human ROCO proteins and ROC-COR supra-domain sequence similarity. A) Multidomain structure of the human ROCO proteins which are characterized by a conserved tandem ROC-COR domain. Abbreviations: Ank, ankyrin repeats; Arm, armadillo repeats; Ca²⁺/CaM reg, calcium/calmodulin regulatory domain; COR, C-terminal of ROC; DAPK1, death-associated protein kinase 1; LRR, leucine-rich repeats; LRRK1, leucine-rich repeat kinase 1; LRRK2, leucine-rich repeat kinase 2; MASL1, malignant fibrous histiocytoma amplified sequence 1; ROC, Ras of complex proteins. B) Peptide sequence identity and C) similarity of the ROC-COR supra-domain across ROCO proteins. ROC-COR region defined as residues 612–1225 (DAPK1), 574–1143 (LRRK1), 1271–1790 (LRRK2) and 345–972 (MASL1). Please note that the presence of a WD40 domain in LRRK1 is still a matter of scientific debate.^[50,52]

identification of PPIs reported in the published literature, to generate a weighted protein–protein interaction network analysis (WPPINA).^[25] Second, to complement WPPINA, we used protein microarray screens to construct an experimental network, enabling hypothesis-free discovery of novel protein interactors. We compared these two ROCO protein networks to validate interactors across the approaches and prioritize interactors for further investigation. Functional insight into these networks was obtained by utilizing gene ontology (GO) functional annotations. These results highlight a subset of interactors common to multiple ROCO proteins, but also numerous interactors specific to particular ROCO proteins, supporting the hypothesis that these proteins have evolved largely independent cellular functions.

Furthermore, we demonstrate that the use of WPPINA to query a high-throughput-derived PPI dataset (such as data obtained by protein microarray screens) represents a novel, rapid, and effective tool to prioritize protein interactors for further experimental validation based on the functional knowledge that is readily available in the published literature.

2. Experimental Section

2.1. Literature-Derived Network Data Download

Protein–protein interaction data was obtained by querying the PSICQUIC online interface^[26] (available at <http://www.ebi.ac.uk/Tools/webservices/psicquic/view/main.xhtml>) for DAPK1, LRRK1, LRRK2, and MASL1, independently. Data was downloaded on January 12, 2017, in a MITAB 2.5 format, from six primary database sources: IntAct,^[27] BioGRID,^[28] InnateDB, Innate-DB-All, InnateDB-IMEx,^[29] and MINT^[30] to ensure a wide capture of reported PPIs.

2.2. Construction of the Literature-Derived Network

The literature-derived ROCO PPI network was constructed as previously described.^[25] In brief, datasets from primary PPI databases were processed to obtain format and protein identifier (ID) consistency, utilizing a dictionary dataset of all human proteins (developed from a UniProt search of human proteins obtained on January 13, 2017). Data from the six datasets were then merged into a single file and repeated equivalent interaction data entries (i.e., interactions derived from the same publication and annotated in multiple databases) were removed.

A series of filtering steps were applied. First, non-protein interactors, such as chemical and miRNA, and protein ID terms corresponding to non-reviewed automatic annotations, which include UniProt TrEMBL IDs, were removed. In addition, transcript-specific information was removed. Next, non-human interactors, which included seed orthologs, were discarded. Filtered datasets were then subjected to method detection reassignment, which grouped similar detection methods based on the EBI Molecular Interactions Ontology, available at <http://www.ebi.ac.uk/ols/ontologies/mi> (File 1, Supporting Information).

A confidence value was assigned to each interaction based on three parameters: method score (MS), the number of different methods used to detect a specific interaction (one method scored a value of 1, multiple methods scored a value of 2); publication score (PS), the number of publications that report a specific interaction (one publication scored a value of 1, multiple publications scored a value of 2); and CRAPome score (CS), the likelihood that the interaction is an affinity purification mass spectrometry (APMS) contaminant. The CS utilizes the CRAPome^[31] (version 1.1), a known contaminant repository for APMS experiments, which contained 411 datasets at the time of scoring (January 18, 2017). Each interactor that was detected by APMS was queried against the CRAPome and if the protein was a positive hit in >50% of the CRAPome datasets and had only been detected by APMS, the protein was scored a value of –1; if the protein was a positive hit in >50% of the CRAPome datasets but had also been detected by another non-APMS method or was a positive hit in 30–50% of the CRAPome dataset and had only been detected by APMS, the protein was scored a value of –0.5; and if the protein was a positive hit for <50% of the CRAPome datasets and had also been detected by another non-APMS method or was a positive hit in <30% of the CRAPome datasets, the protein was scored a value of 0.

The sum of the three scoring parameters then formed the basis of a confidence scale and only interactions that scored <2 were retained for constructing the network. This <2 score threshold ensures that nodes of the network represent interactors that have been independently replicated, by method and/or publication.

2.3. Protein Production and Purification

HEK293T cells were transfected with 3xFLAG tagged DAPK1, LRRK1, LRRK2, MASL1, or GFP plasmids using PEI reagent, collected 24 h after transfection and cells were lysed in the buffer: 20 mM Tris (pH 7.5), 150 mM NaCl, 1 mM EDTA, 1% Triton, 10% Glycerol, protease inhibitor cocktail (Roche), and 1x Halt phosphatase inhibitor cocktail (Thermo Scientific). Lysates were pre-cleared by centrifugation at 20 000 × g for 10 min and incubated for 1 h at 4 °C with EZview Red Protein G beads (Sigma) to remove proteins non-specifically binding to agarose. After pre-clear with protein G beads, lysates were incubated for 1 h at 4 °C with EZview Red Anti-FLAG M2 Agarose (Sigma) that is suitable for the immunoprecipitation of FLAG fusion proteins. Beads were washed six times with the wash buffer: 20 mM Tris (pH 7.5), 400 mM NaCl, 1% Triton and proteins were eluted in 25 mM Tris (pH 7.5), 150 mM NaCl, and 100 μg mL⁻¹ 3xFLAG peptide (Sigma). Protein yields and purity were estimated by staining gels with Coomassie brilliant blue staining (Thermo Scientific, Figure 1, Supporting Information).

2.4. Protein Microarrays

3xFLAG tagged, full-length DAPK1, LRRK1, LRRK2, MASL1, and GFP control proteins were purified as previously described.^[32] Six micrograms of each purified 3xFLAG tagged protein were used to probe protein microarrays (Protoarray,

version 4.1; Invitrogen) according to the manufacturer's instructions with the modification that after 3xFLAG tagged protein probing, arrays were probed with monoclonal ANTI-FLAG BioM2–Biotin, Clone M2 (Sigma-Aldrich) antibody, followed by probing with Alexa Fluor 647 streptavidin (Invitrogen).^[33] Arrays were imaged using an Axon GenePix 4000B fluorescence scanner and images were analyzed using GenePix Pro software. ProtoArray Prospector software was used to analyze the microarray data acquired from GenePix Pro and identify the significant hits. Binding strength was estimated as Z-scores, that is, numbers of standard deviations above background fluorescence on the array. Each protein on the array was spotted in duplicate, hence reported values were averaged for both spots. Signals considered as potential interactions were determining using a Z-score threshold of $Z > 3$. ROCO protein positive hit interactors were determined by filtering against GFP (negative control) interactions to identify proteins that bound DAPK1, LRRK1, LRRK2, or MASL1 but not GFP.

2.5. Functional Annotation

To gather insight into the cellular processes that are influenced by the proteins within the networks, functional enrichment analysis was performed. This analysis is based on gene ontology (GO) annotations and determines enrichment of biological process (BP) annotations within a query protein list (ROCO protein interactors in this case), by a comparison against annotations for the entire human genome. Functional enrichment analysis was undertaken using g:Profiler g:GOST (available at <http://biit.cs.ut.ee/gprofiler/index.cgi>), on June 23, 2017. Statistical significance was determined using Fisher's one-tailed test with a g:Profiler g:SCS algorithm to correct for multiple testing; $p < 0.05$ was set as the significance threshold and output data was not subjected to hierarchical filtering. Results were confirmed by replication of the functional enrichment analysis using WebGestalt^[34] (<http://webgestalt.org/option.php>) and Panther^[35] (<http://www.pantherdb.org/>) on November 22, 2017 (File 10, Supporting Information); the statistical testing underlying the enrichment analysis for these alternative portals is different, thus replication by this means provides reinforcement of the result obtained using g:Profiler.

All algorithms used for data processing were developed in R version 3.2.2. Networks were generated and visualized using Cytoscape^[36] version 3.3.0 and graphs were produced in GraphPad Prism 7.0.

3. Results

We here present an insight into the protein interaction network of the ROCO protein family. The four human ROCO proteins, DAPK1, LRRK1, LRRK2, and MASL1 were used as seed proteins. The term “interactome” refers to the group of proteins that directly bind to a particular seed protein.

3.1. Construction of the Literature-Derived Network

The literature-derived PPI network (Figure 2A) was constructed by collecting the reported PPIs of ROCO proteins, utilizing

our recently developed pipeline (WPPINA),^[25] which collates data from several databases within the IMEx consortium.^[37] Data were quality checked and a confidence threshold was applied to retain only interactions that have been replicated by a minimum of two experimental methods and/or reported in at least two peer-reviewed publications. Therefore, this network provides a confidence-weighted visual overview of state-of-the-art PPI knowledge centered on the human ROCO proteins.

The network topology indicated a strong bias toward the LRRK2 interactome with 113 interactors, compared to the 38, 14, and 4 interactors for DAPK1, LRRK1, and MASL1, respectively (Figure 2B). This differential recovery of PPIs is likely driven by literature bias toward proteins with known disease associations. For example, LRRK2 is the focus of many investigations within PD research,^[38] whereas MASL1 is relatively understudied.^[21] Interestingly, this trend differs when considering the interactomes prior to applying the confidence threshold (i.e., when retaining all reported interactors regardless of replication; Figure 2B). Of the 57 DAPK1 interactors reported within the literature, 38 were retained when the confidence threshold was applied. This relatively high (66.7%) retention of interactors indicates that the majority of interactors that have been identified for DAPK1 have been replicated. Four of the seven (57.1%) reported MASL1 interactors were also replicated observations and hence were retained for constructing the network. In contrast, only 16.5% of LRRK1 interactors and 23.5% of LRRK2 interactors were retained after confidence thresholding, showing limiting replication of the interactors identified. These results suggested that the expanse of PPI data for more widely studied proteins does not directly reflect increased confidence or robustness of the related interactome.

Considering only the confidence thresholded network, our results indicated common interactors between ROCO proteins: three interactors common to both DAPK1 and LRRK2 (FADD, MYO1B, and MYO1D), and two interactors common to both LRRK1 and LRRK2 (BAG5 and HSPA8; Figure 2A). Functional insight into these common interactors is summarized in Table 1, Supporting Information. In addition, from this analysis it was shown that DAPK1, LRRK1, and LRRK2 can exist as homo- and hetero-dimers, conformations that may be critical for the functions of these proteins.^[39,40] In contrast, the MASL1 interactome was fully detached from the other ROCO protein interactomes within this network, indicating a lack of common interactors between MASL1 and the other ROCO proteins on the basis of the existing literature.

3.2. Generating the Experimental Network

To address the biases in literature coverage for the human ROCO proteins, we performed protein microarray experiments as a hypothesis-free approach for identifying potential ROCO protein interaction partners. This approach formed the basis of the experimental network (Figure 3A). We limited false-positive hits in each interactome by setting a Z-score threshold to distinguish positive hits from background signals and by filtering ROCO protein hits against GFP hits as a negative control for non-specific binding.

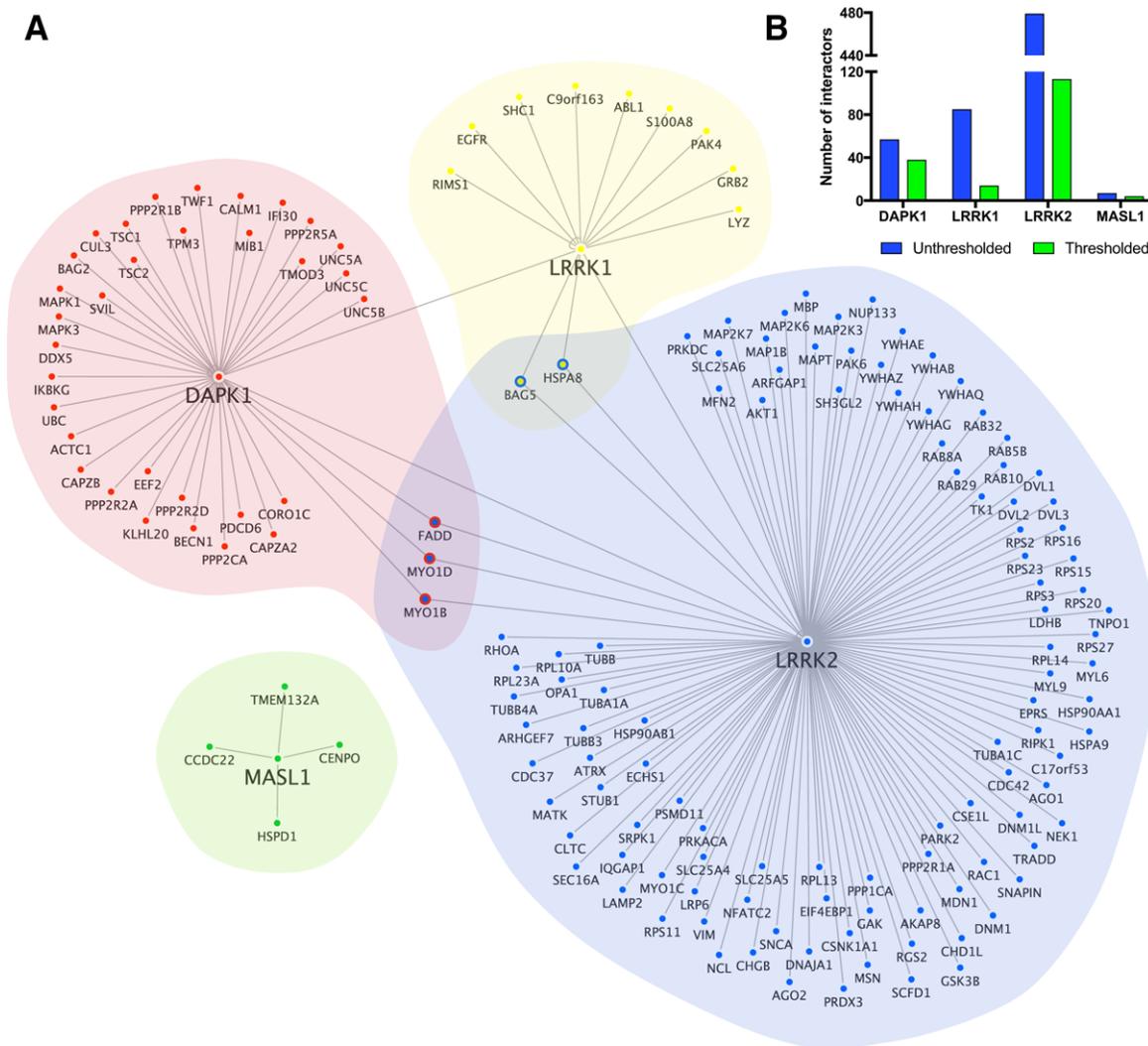


Figure 2. Literature-derived ROCO protein interaction network. A) Network visualization of the ROCO protein interaction partners following data processing via the WPPINA pipeline. B) Quantification of the interactors retained following confidence score thresholding.

In contrast to the literature-derived network, this network displayed a more even distribution of interactors around each seed protein (Figure 3B). Specifically, we identified 87 (DAPK1), 51 (LRRK1), 78 (LRRK2), and 87 (MASL1) positive hits for each seed protein, respectively (File 2, Supporting Information). Of note, numerous kinases have been identified as potential MASL1 interactors (Table 1), six of which, CLK1, LIMK1, MAP3K4, NEK11, ROR1, and STK25, appear to be specific interacting partners of MASL1.

A remarkable finding from mapping this protein microarray data was that 23.5% of the entire network consisted of common node connections between two or more seed proteins. Furthermore, 8.4% of the nodes in the network were common to three or more seed proteins and five nodes (2.2% of the network) were common to all four seed proteins (Figure 3A,C). This suggested that the overlap between seed protein interactomes might in fact be greater than previously reported. However, it is important to note that further validation

of these interactors is required to increase confidence in their veracity.

3.3. Identification of ROCO Protein Common Interactors

A particular advantage of applying two orthologous network analysis approaches is to compare and combine the networks to minimize the burden of approach-specific limitations and maximize the capacity of available data. To achieve this, we merged the literature-derived network and the microarray data with the aim of validating via the literature some of the experimentally obtained, but not replicated, hits. Many nodes were common to both networks (Figure 4; referred to as the common core network). These common nodes can be categorized into three groups: i) interactors of the same seed protein that are cross-supported by both networks (e.g., ARFGAP1, CHGB, and GAK which are common to LRRK2 in both networks); ii) interactors

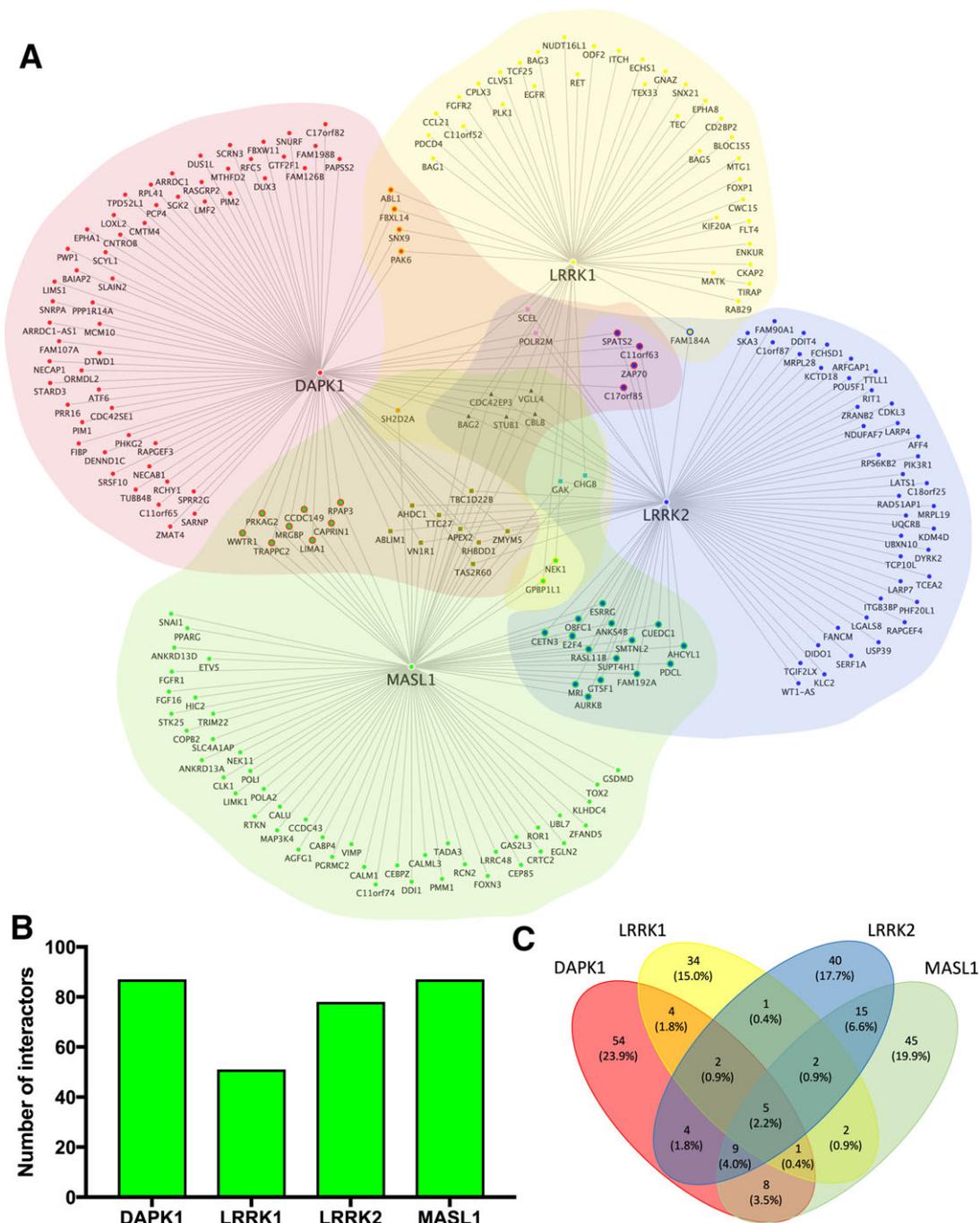


Figure 3. Experimental ROCO protein interaction network. A) ROCO protein interaction network analysis using protein microarray screens. B) Quantification of positive interactors identified by protein microarray for each ROCO protein. Three hundred and three interactions identified across 226 nodes. C) Extent of common nodes within the experimental network. Number of interactors and percentage of entire experimental network reported.

common to both literature-derived data and the experimental network but within different seed protein interactomes (Figure 4), and iii) interactors that are common to both literature-derived and experimental datasets associated with the same seed protein, but only if the confidence threshold is removed from the literature-derived data (Table 2, Supporting Information). Interactors from iii) do not exceed the confidence threshold in place within the WPPINA pipeline to support replication of

interactors, however with integration of the protein microarray data these interactors would exceed this threshold due to acquiring independent replication from the protein microarray experiments.

Considering the overlap between the literature-derived and the experimental networks, 14 common interactors were identified (Figure 4B). When the non-thresholded literature-derived data and the protein microarray network were examined, 48

Table 1. MASL1-interacting kinases. Kinases that were identified as interactors of MASL1 in the protein microarray screen, with functional associations. Of note, cell cycle-related functions appear to be a common functional theme.

| Kinase interactor | | Additional seed interaction | Functional enrichment contribution | Further functional detail |
|----------------------|--|-----------------------------|--|--|
| Abbreviated name | Full protein name | | | |
| AURKB | Aurora kinase B | LRRK2 | - | <ul style="list-style-type: none"> • Interacts with CLK1,^[53] another MASL1-interacting kinase identified in this protein microarray screen • Phosphorylated AURKB localizes to kinetochores in prometaphase cells^[54] • Functional role in mitotic cell division, specifically as a catalytic unit of the chromosomal passenger complex (CPC)^[54] • Dysregulation associated with tumorigenesis^[55] |
| CLK1 ^{a)} | CDC2-like kinase 1 | - | - | <ul style="list-style-type: none"> • Associates and phosphorylates AURKB,^[53] another MASL1-interacting kinase identified within this protein microarray screen • Dual specificity kinase that localizes to the nucleus^[56] • Involved in alternative splicing and neuronal differentiation^[56–58] • Potential drug target for Influenza and Alzheimer's disease (AD)^[59,60] |
| GAK | Cyclin-G-associated kinase | LRRK1 and LRRK2 | Development, transport, intracellular organization, protein metabolism | <ul style="list-style-type: none"> • Androgen receptor-interacting transcriptional coactivator^[61] • Localizes to the trans-Golgi network^[62] • Involved in clathrin-mediated membrane trafficking and metaphase mitotic progression^[63] • Disease links to cancer and Parkinson's disease (PD)^[61,64,65] |
| LIMK1 ^{a)} | LIM domain kinase 1 | - | - | <ul style="list-style-type: none"> • Regulates microtubule dynamics, specifically mitotic spindle structure and positioning • Acts downstream of several Rho-family GTPase signal transduction pathways^[66] |
| MAP3K4 ^{a)} | Mitogen-activated protein kinase kinase kinase 4 | - | - | <ul style="list-style-type: none"> • Mediator in stress-activated p38/MAPK and JNK signaling pathways^[67] • Involved in tumor suppression and epithelial-mesenchymal transition^[68] • Loss of MAP3K4 is associated with defective neural tube development^[69] |
| NEK1 | NIMA-related kinase 1 | LRRK1 | Cell Cycle, Intracellular Organization, Protein Metabolism, Response to Stimulus | <ul style="list-style-type: none"> • Associated with axial spondylometaphyseal dysplasia^[70] • Involved in DNA damage response and cell cycle control; suggested role in post-mitotic cilia assembly • Mutations in NEK1 are associated with ciliopathy and polycystic kidney disease (PKD)^[71] |
| NEK11 ^{a)} | NIMA-related kinase 11 | - | - | <ul style="list-style-type: none"> • Involved in DNA damage and genotoxic stress responses • Highly expressed throughout S phase of the cell cycle to the G2-M transition • Activated by phosphorylation by ATM and ATR kinases^[71] |

(Continued)

interactors were common to both datasets (Figure 4 and File 3, Supporting Information).

To further investigate the likelihood that the common core network reports true interactions, we added an additional score considering tissue-specific gene expression. Using expression data derived from GTEx^[41] and a gene expression threshold of

three reads per kilobase of transcript per million mapped reads (RPKM), co-expression analysis identified distinct tissues where specific interactor mRNA were expressed together with specific seed protein mRNA (Table 3, Supporting Information).

Concerning pairwise interactions between ROCO proteins and interactors from the common core network (48 interactors and a

Table 1. Continued.

| Kinase interactor | | Additional seed interaction | Functional enrichment contribution | Further functional detail |
|---------------------|---|-----------------------------|------------------------------------|---|
| Abbreviated name | Full protein name | | | |
| ROR1 ^{a)} | Receptor tyrosine kinase-like orphan receptor 1 | - | - | <ul style="list-style-type: none"> ● Pseudokinase ● Non-canonical Wnt transmembrane receptor^[72] ● Highly upregulated in chronic lymphocytic leukemia (CLL)^[73] and other blood cancers^[74] |
| STK25 ^{a)} | Serine/threonine kinase 25 | - | - | <ul style="list-style-type: none"> ● Associates with Golgi apparatus ● Dominant negative STK25 causes dispersal of the Golgi apparatus and inhibits cell migration^[75] ● Involved in glucose homeostasis^[76] ● Regulates lipid release from lipid droplets and induces NAFLD/NASH pathogenesis^[77] |

a) kinases specific to MASL1

total of 115 pairwise interactions; Table 3, Supporting Information), only 1 protein out of 48 interactors, DUX3, was not found in the GTEx database used for co-expression analysis. On average, co-expression within nine tissues was evident, whilst in ten cases co-expression was found in 12 tissues (out of 13 analyzed, Table 3, Supporting Information). Although most tissues included at least one co-expressed interaction pair, one tissue (skeletal muscle) did not show any co-expressed interaction pairs due to an absence of significant ROCO protein expression in this tissue. The highest proportion of co-expressed interaction pairs was seen in the reproductive apparatus (96% of co-expressed interaction pairs), followed by two tissues: brain and intestine (92% and 90% of co-expressed interaction pairs, respectively; Table 3, Supporting Information).

3.4. Functional Insight Into the Common Core Network

The literature-derived and common core networks were subjected to functional enrichment analysis based on gene ontology (GO) functional annotations. Particularly, we used biological process (BP) terms to gather functional insight into these networks. The significantly enriched BP terms were grouped into functional blocks defined by more specific semantic classes (using a curated dictionary list to match GO terms with a custom grouped ontology) based on semantic similarity (Files 4 and 8, Supporting Information). This enabled an overview of significantly enriched functions (see Figure 2, Supporting Information for a summary of functions associated with the ROCO literature-derived and common core networks, and Files 4–9, Supporting Information for a breakdown of each functional block, including semantic class- and GO term-specificity).

Within the common core network, which consists of 48 common interactors plus the seed protein nodes (Figure 4), a total of 26 GO BP terms were significantly enriched, representing a specific subset of the whole 516 functionally diverse terms significantly enriched within the literature-derived network (Figure 2, Supporting Information and File 4, Supporting

Information). The predominantly enriched terms within this refined analysis indicated “response to stimulus” and “intracellular organization” functional blocks supported by “stress” and “cell projections” semantic classes, respectively (Table 2). Functional associations for specific ROCO proteins were also explored by functional enrichment analysis of the individual interactomes within the literature-derived network identifying “cell death” and “development” as distinct functional themes for DAPK1 and LRRK1, respectively, and “intracellular organization” and “transport” for LRRK2 (Table 2).

4. Discussion

The human ROCO proteins are defined by a ROC-COR supra-domain which contains highly conserved motifs and substantial sequence similarity (Figure 1), a tandem domain organization that can be evolutionarily traced from prokaryotic organisms.^[42] This domain homology is paralleled by flanking domain dissimilarity, driving a twofold interest into the proximal interactors of these proteins and their potential effects on subcellular functions: first from a fundamental biology perspective in relation to the complex domain organization of these proteins, and second from a drug discovery perspective due to the involvement of these proteins in human diseases.

In the current study, we used a combination of bioinformatic literature-based analysis (WPPINA) and an experimentally derived protein microarray dataset, to expand our insight into the ROCO protein interactomes, specifically into common and distinct interactors, and functional pathways regulated by this family of proteins.

Although literature-derived PPI networks are incomplete by definition as they are affected by ascertainment bias^[43] and depend solely on existing experimental findings (i.e., many interactors may exist that are yet to be discovered and/or relatively newly discovered protein interactors will be neglected in comparison to the more studied ones), the WPPINA analysis reported here represents a comprehensive literature review of reported ROCO

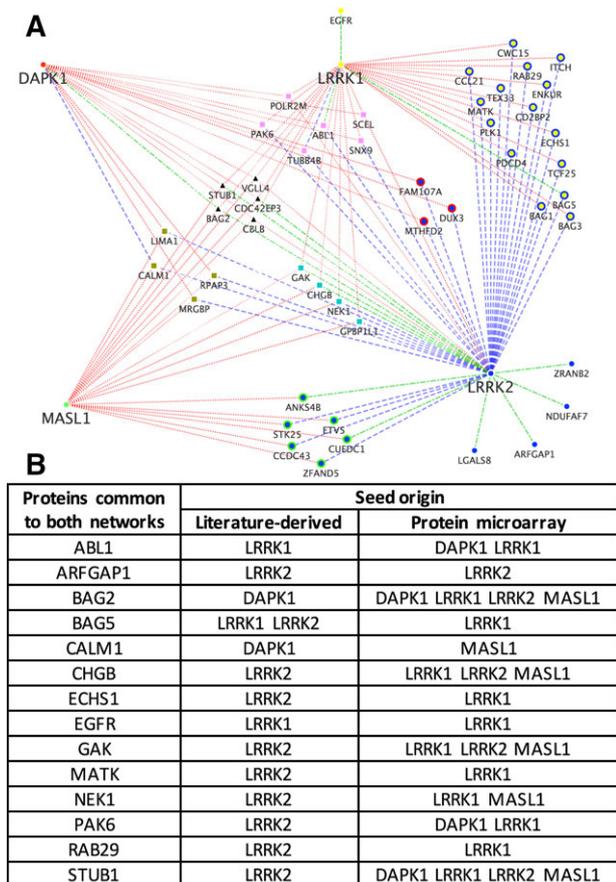


Figure 4. Common nodes across literature-derived and experimental ROCO PPI data. A) The network is specifically depicted to highlight interactors that are common to both non-thresholded literature-derived data and protein microarray data. Dotted edges indicate interactions deriving from protein microarray experiments; dashed edges indicate interactions described in literature; dotted and dashed edges are interactions replicated between the two datasets. Seeds are represented with a circular node. Common interactors are represented with a double circular node if they are common to two seeds, square node if they are common to three seeds, and triangular node if they are common to all four seeds. B) Common nodes across the literature-derived and experimental networks when considering literature-derived data after thresholding.

PPIs^[25] and ensures an extensive and weighted coverage of primary literature sources comparatively to currently used literature mining and prediction network mapping tools.

In the case of the ROCO proteins, the literature-derived PPI network incorporates a potential bias toward the LRRK2 interactome, due to extensive investigation into LRRK2 in relation to PD.^[44,45] This is evidenced by the nearly twofold increase in the number of LRRK2 interactors (from 62 to 113) compared to a previous analysis performed in 2014 using an earlier version of the same data processing pipeline.^[4] Conversely, the distribution of nodes amongst the other ROCO proteins highlights the comparative neglect of research into characterizing the DAPK1, LRRK1, and MASL1 interactomes.^[21] However, it is worth considering that all interactions reported through WPPINA are experimentally proven, replicated, and cleared from type-I error.

The domain topology and primary structures of the ROCO proteins are dissimilar outside of the ROC-COR region (Figure 1), hence common interactors may provide hints toward ROC-COR-specific interactions. The common interactors identified within the literature-derived network include: FADD, MYO1B, and MYO1D (between DAPK1 and LRRK2), and BAG5 and HSPA8 (between LRRK1 and LRRK2). Functional insight into these common interactors is provided in Table 1, Supporting Information. Of note, two common interactors (MYO1B and MYO1D) are unconventional myosin proteins involved in vesicle trafficking, a critical function for many cellular processes and ultimately cell survival. Interestingly, Rab proteins have a regulatory role in myosin motor function, which combined with evidence of Rab proteins as LRRK2 substrates^[46] and myosins as LRRK2 interactors, supports a key role for LRRK2 in the regulation of intracellular vesicle transport.^[47] This WPPINA approach allows for the straightforward identification of these mutual connections which could easily be overlooked when reviewing literature using alternative strategies. By removing the confidence threshold within the WPPINA pipeline, we increased the number of interactors reported within the literature-derived ROCO PPI network; however, the additional interactors have to be considered carefully since there is no evidence of replication within the peer-reviewed literature.

The experimental network, which is based on protein microarray data, provides novel insight into the ROCO protein interactomes. This network is not biased toward a specific seed protein since all are equally evaluated utilizing a hypothesis-free approach and is complete in relation to the extensive range of proteins immobilized on the microarray (9480 proteins). However, the experimental network is not as robust as the literature-derived network due to technical biases (i.e., intrinsic limitations to this experimental procedure, including the choice of baits for the microarray; alterations of physiological protein conformations [non-physiological environment, absence of lipidic membranes, tagged preys]; variations of posttranslational modifications as evidence suggests that the phosphorylation state of LRRK2 impacts the protein interaction profile of the protein).^[48] Consequently, interactions reported in the experimental network require replication by alternative interaction detection methods to overcome the technical biases and ensuring validity of the protein microarray positive hits.

Nevertheless, this high-throughput approach allows for the identification of potential novel interactors, expanding the current landscape of the ROCO protein interaction network, particularly for the less studied ROCO proteins. For example, many potential MASL1 interactors have been identified, which include numerous kinases (Table 1). MASL1 (unlike the other ROCO proteins) lacks an intrinsic kinase domain (Figure 1), therefore it can be hypothesized that its GTPase activity within the ROC domain may influence an extrinsic kinase domain.^[39,49,50] The novel potential MASL1-interacting kinases identified in this screen may be downstream effectors of the switch-like GTPase activity of MASL1 and thus part of a reciprocal regulatory relationship.

To address the intrinsic biases of these two approaches, we integrated the literature-derived and experimentally derived data. The advantage of this strategy is that the microarray data will dilute the ascertainment bias of the literature-derived network, while the literature-derived network will supply the reproducibil-

Table 2. Most significantly enriched terms from functional enrichment analysis of each dataset.

| Datasets | <i>p</i> -value | GO term | Semantic class |
|-------------------------------|-----------------|---|--|
| Literature-derived Network | 4.6E-36 | Cellular component organization or biogenesis | Intracellular organization |
| | 2.44E-30 | Intracellular transport | Transport—intracellular |
| | 4.31E-30 | Cellular component organization | Intracellular organization |
| DAPK1 | 0.000000385 | Cell death | Cell death |
| | 0.000000439 | Apoptotic process | Cell death—apoptosis |
| | 0.00000136 | Programmed cell death | Cell death |
| LRRK1 | 0.0000349 | Neuron projection development | Development—neuronal—axon |
| | 0.0000382 | Cell development | Development |
| | 0.0000911 | Neurogenesis | Development—neuronal |
| LRRK2 | 2.75E-29 | Cellular component organization or biogenesis | Intracellular organization |
| | 4.36E-29 | Intracellular transport | Transport—intracellular |
| | 1.61E-26 | Establishment of localization in cell | Protein metabolism—localization |
| Common core Network | 0.0000192 | Regulation of cellular response to stress | Response to stimulus—stress |
| | 0.0000539 | Plasma membrane bounded cell projection organization | Intracellular organization—cell projections |
| | 0.0000752 | Cell projection organization | Intracellular organization—cell projections |

ity element and aid prioritization of positive hits from the microarray experiments. By overlaying this data, numerous interactors common to both datasets became evident, including interactors from the same seed origin, reinforcing confidence in the protein microarray data, and interactors that were replicated between the two approaches but in association with different seeds of origin (Figure 4), opening new avenues for future functional investigation. These common nodes across both approaches were used to construct the common core network (Figure 4), which illustrates the potential overlap in ROCO protein interaction profiles.

The probability of proteins interacting within the cellular environment is subject to a number of important variables, including both temporal and spatial patterns of expression. Therefore, we subjected the interactors of the common core network to tissue-specific gene expression profiling using data from GTEx (Table 3, Supporting Information). Although this represents a crude type of analysis (i.e., temporal expression and intracellular localization are not taken into account), it provided another way to assess the probability of the interactions reported in the common core network based on co-existence of protein transcripts in human tissues. Particularly, we gathered that the highest frequency of co-expressed interaction pairs was in the reproductive apparatus, followed by brain and intestine, whilst only skeletal muscle did not show any co-expression. Additionally, ten proteins (ABL1, CALM1, CBLB, CDC42EP3, GAK, MRGBP, RPAP3, SNX9, STUB1, and TUBB4B) were co-expressed with ROCO proteins in 12 out of the 13 tissues analyzed. This insight into tissue-specific co-expression supports the likelihood of the majority of pairwise interactions that have been reported in the literature and that have been assessed in a functional context.

To obtain functional insight into the ROCO protein interaction network, we performed functional enrichment analysis for the literature-derived and common core networks, independently. The analysis of the former evidenced a diverse range of cellular functions (Figure 2A, Supporting Information), which support the concept of the ROCO proteins as hubs for a multitude of signaling cascades and hence challenging targets for therapeutic development.^[4] The analysis of the latter suggested a limited range of associated functional blocks: cell death, intracellular organization (particularly cell projections), protein metabolism, and response to stimulus (particularly stress response; Table 2 and Figure 2B, Supporting Information). In addition, functional enrichment analysis of individual ROCO protein interactomes indicated distinct functional themes for each seed protein (Table 2). In combination, these enrichment analyzes provide an overview of cellular functions associated with the ROCO proteins, suggesting potential convergent and divergent roles of these proteins within the cell, thus guiding future detailed assessments of ROCO protein function.

These analyzes provide a valuable foundation for understanding the ROCO protein interaction network. We here integrated peer-reviewed literature, microarray, and co-expression datasets to isolate common and distinct interactors of the ROCO proteins. We constructed a ROCO protein common core network highlighting the extent of commonality in the interaction profiles of these proteins. Using functional analysis approaches, we showed that, the ROCO proteins share a structurally conserved unit, which may be responsible for shared interactions (such as those with BAG2, CBLB, CDC42EP3, STUB1, and VGLL4) and as consequence, may influence the involvement of the ROCO proteins in common pathways identified (such as stress response

and cell projection organization). However, despite this domain conservation, the ROCO proteins seem to have evolved largely divergent interactomes and associated functions within the cell (Table 2). This supports previous research into the functions of LRRK1 and LRRK2.^[7] The diversification of interactomes and biological functions of the ROCO proteins may reflect an evolutionary pressure toward phylogenetic differentiation of a single ancestral ROCO gene^[51] and may justify why the human ROCO proteins are differently associated with disease.

In summary, we utilized a confidence-weighted data processing pipeline (WPPINA) to prioritize high-throughput experimental results. Importantly, this approach provides the flexibility to incorporate data from a wide range of sources, and in the future could be further complemented by findings from yeast two-hybrid and stable isotope labelling with amino acids in cell culture (SILAC) screens, for example. Together, this analysis highlights the value of a multi-layered approach, combining bioinformatics with novel experimental data to better inform and accelerate laboratory investigations.

Abbreviations

DAPK1, death-associated protein kinase 1; LRRK1, leucine-rich repeat kinase 1; LRRK2, leucine-rich repeat kinase 2; MASL1, malignant fibrous histiocytoma amplified sequence 1; WPPINA, weighted protein–protein interaction network analysis

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

J.E.T. is supported by BBSRC CASE studentship with BC Platforms, BB/M017222/1. P.A.L. is supported by the MRC (grants MR/N026004/1; MR/L010933/1), Parkinson's UK (Fellowship F1002), and the Michael J. Fox Foundation for Parkinson's Research. R.F. is supported by funding from Alzheimer's Society (grant number 284). This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

LRRK2, LRRK1, DAPK1, MASL1/MAFHAS1, protein networks, ROCO proteins, protein microarrays

Received: November 28, 2017
Revised: February 5, 2018

- [1] S. Oliver, *Nature* **2000**, 403, 601.
- [2] K. Huang, D. C. Fingar, *Semin. Cell Dev. Biol.* **2014**, 36, 79.
- [3] L. Civiero, S. Dihanich, P. A. Lewis, E. Greggio, *Chem. Biol.* **2014**, 21, 809.
- [4] C. Manzoni, P. Denny, R. C. Lovering, P. A. Lewis, *PeerJ* **2015**, 3, e778.
- [5] P. Porras, M. Duesbury, A. Fabregat, M. Ueffing, S. Orchard, C. J. Gloeckner, H. Hermjakob, *Proteomics* **2015**, 15, 1390.
- [6] S. Bialik, A. Kimchi, *Apoptosis* **2014**, 19, 316.
- [7] L. Reyniers, M. G. Del Giudice, L. Civiero, E. Belluzzi, E. Lobbstaer, A. Beilina, G. Arrigoni, R. Derua, E. Waelkens, Y. Li, C. Crosio, C. Iaccarino, M. R. Cookson, V. Baekelandt, E. Greggio, J.-M. Taymans, *J. Neurochem.* **2014**, 131, 239.
- [8] L. P. Deiss, E. Feinstein, H. Berissi, O. Cohen, A. Kimchi, *Genes Dev.* **1995**, 9, 15.
- [9] B. Inbal, O. Cohen, S. Polak-Charcon, J. Kopolovic, E. Vadai, L. Eisenbach, A. Kimchi, *Nature* **1997**, 390, 180.
- [10] V. Levin-Salomon, S. Bialik, A. Kimchi, *Apoptosis* **2014**, 19, 346.
- [11] M.-Z. Lai, R.-H. Chen, *Apoptosis* **2014**, 19, 357.
- [12] H. Hanafusa, K. Ishikawa, S. Kedashiro, T. Saigo, S.-I. Iemura, T. Natsume, M. Komada, H. Shibuya, A. Nara, K. Matsumoto, *Nat. Commun.* **2011**, 2, 158.
- [13] H. Hanafusa, S. Kedashiro, M. Tezuka, M. Funatsu, S. Usami, F. Toyoshima, K. Matsumoto, *Nat. Cell Biol.* **2015**, 17, 1024.
- [14] K. Morimoto, Y. Baba, H. Shinohara, S. Kang, S. Nojima, T. Kimura, D. Ito, Y. Yoshida, Y. Maeda, H. Sarashina-Kida, M. Nishide, T. Hosokawa, Y. Kato, Y. Hayama, Y. Kinehara, T. Okuno, H. Takamatsu, T. Hirano, Y. Shima, M. Narazaki, T. Kurosaki, T. Toyofuku, A. Kumanogoh, *Sci. Rep.* **2016**, 6, 25738.
- [15] R. Wallings, C. Manzoni, R. Bandopadhyay, *FEBS J.* **2015**, 282, 2806.
- [16] W. Chen, Y. Xu, J. Zhong, H. Wang, M. Weng, Q. Cheng, Q. Wu, Z. Sun, H. Jiang, M. Zhu, Y. Ren, P. Xu, J. Chen, C. Miao, *Oncotarget* **2016**, 7, 78726.
- [17] J. Zhong, H. Wang, W. Chen, Z. Sun, J. Chen, Y. Xu, M. Weng, Q. Shi, D. Ma, C. Miao, *Cell Death Dis.* **2017**, 8, e2763.
- [18] C. Kumkhaek, W. Aerbajinai, W. Liu, J. Zhu, N. Uchida, R. Kurlander, M. M. Hsieh, J. F. Tisdale, G. P. Rodgers, *Blood* **2013**, 121, 3216.
- [19] Y. Huang, L. Chen, L. Guo, T. R. Hupp, Y. Lin, *Apoptosis* **2014**, 19, 371.
- [20] B. Titz, T. Low, E. Komisopoulou, S. S. Chen, L. Rubbi, T. G. Graeber, *Oncogene* **2010**, 29, 5895.
- [21] S. Dihanich, *Biochem. Soc. Trans.* **2012**, 40, 1090.
- [22] C. Klein, A. Westenberger, *Cold Spring Harb. Perspect. Med.* **2012**, 2, a008888.
- [23] P. A. Lewis, C. Manzoni, *Sci. Signal.* **2012**, 5, pe2.
- [24] J.-M. Taymans, E. Greggio, *Curr. Neuropharmacol.* **2016**, 14, 214.
- [25] R. Ferrari, R. C. Lovering, J. Hardy, P. A. Lewis, C. Manzoni, *J. Proteome Res.* **2017**, 16, 999.
- [26] H. H. Aranda, H. BrunoBlankenburg, S. Kerrien, F. S. Brinkman, A. Ceol, E. Chautard, J. M. Dana, J. De Las Rivas, M. Dumousseau, E. Galeota, A. Gaulton, J. Goll, R. E. Hancock, R. Isserlin, R. C. Jimenez, J. Kerssemakers, J. Khadake, D. J. Lynn, M. Michaut, G. O'Kelly, K. Ono, S. Orchard, *Nat. Methods* **2011**, 8, 528.
- [27] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeifferberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard, H. Hermjakob, *Nucleic Acids Res.* **2012**, 40, D841.
- [28] A. Chat-aryamontri, B.-J. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke, D. Chen, C. Stark, A. Breitkreutz, N. Kolas, L. O'Donnell, T. Reguly, J. Nixon, L. Ramage, A. Winter, A. Sellam, C. Chang, J. Hirschman, C. Theesfeld, J. Rust, M. S. Livstone, K. Dolinski, M. Tyers, *Nucleic Acids Res.* **2015**, 43, D470.

- [29] K. Breuer, A. K. Foroushani, M. R. Laird, C. Chen, A. Sribnaia, R. Lo, G. L. Winsor, R. E. W. Hancock, F. S. L. Brinkman, D. J. Lynn, *Nucleic Acids Res.* **2013**, *41*, D1228.
- [30] L. Licata, L. Briganti, D. Peluso, L. Peretto, M. Iannucelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardoza, E. Santonico, L. Castagnoli, G. Cesareni, *Nucleic Acid Res.* **2012**, *40*, 857.
- [31] D. Mellacheruvu, Z. Wright, A. L. Couzens, J.-P. Lambert, N. St-Denis, T. Li, Y. V. Miteva, S. Hauri, M. E. Sardi, T. Y. Low, V. A. Halim, R. D. Bagshaw, N. C. Hubner, A. Al-Hakim, A. Bouchard, D. Faubert, D. Fermin, W. H. Dunham, M. Goudreault, Z.-Y. Lin, B. G. Badillo, T. Pawson, D. Durocher, B. Coulombe, R. Aebersold, G. Superti-Furga, J. Colinge, A. J. R. Heck, H. Choi, M. Gstaiger, S. Mohammed, I. M. Cristea, K. L. Bennett, M. P. Washburn, B. Raught, R. M. Ewing, A.-C. Gingras, A. I. Nesvizhskii, *Nat. Methods* **2013**, *10*, 730.
- [32] L. Civiero, R. Vancraenenbroeck, E. Belluzzi, A. Beilina, E. Lobbstaël, L. Reyniers, F. Gao, I. Micetic, M. De Maeyer, L. Bubacco, V. Baekelandt, M. R. Cookson, E. Greggio, J.-M. Taymans, *PLoS One* **2012**, *7*, e43472.
- [33] A. Beilina, I. N. Rudenko, A. Kaganovich, L. Civiero, H. Chau, S. K. Kalia, *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 2626.
- [34] J. Wang, D. Duncan, Z. Shi, B. Zhang, *Nucleic Acids Res.* **2013**, *41*, W77.
- [35] H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, P. D. Thomas, *Nucleic Acid Res.* **2017**, *45*, D183.
- [36] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, *Genome Res.* **2003**, *13*, 2498.
- [37] S. Orchard, S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, A. Bridge, L. Briganti, F. S. L. Brinkman, G. Cesareni, A. Chatr-aryamontri, E. Chautard, C. Chen, M. Dumousseau, J. Goll, R. E. W. Hancock, L. I. Hannick, I. Jurisica, J. Khadake, D. J. Lynn, U. Mahadevan, L. Peretto, A. Raghunath, S. Ricard-Blum, B. Roechert, L. Salwinski, V. Stumpflen, M. Tyers, P. Uetz, I. Xenarios, H. Hermjakob, *Nat. Methods* **2012**, *9*, 345.
- [38] M. R. Cookson, *Curr. Neurol. Neurosci. Rep.* **2015**, *15*, 42.
- [39] J. D. Jellibell, S. Dihanich, L. Civiero, C. Manzoni, E. Greggio, P. A. Lewis, *Sci. Rep.* **2012**, *2*, 695.
- [40] Z. Berger, K. A. Smith, M. J. Lavoie, *Biochemistry* **2010**, *49*, 5511.
- [41] GTEx Consortium, *Nat. Genet.* **2013**, *45*, 580.
- [42] L. Bosgraaf, P. J. M. Van Haastert, *Biochim. Biophys. Acta Mol. Cell Res.* **2003**, *1643*, 5.
- [43] S. N. Simoes, D. C. J. Martins, C. A. B. Pereira, R. F. Hashimoto, H. Brentani, *BMC Bioinformatics* **2015**, *16 Suppl 1*, S9.
- [44] A. R. Esteves, R. H. Swerdlow, S. M. Cardoso, *Exp. Neurol.* **2014**, *261*, 206.
- [45] B. D. Lee, V. L. Dawson, T. M. Dawson, *Trends Pharmacol. Sci.* **2012**, *33*, 365.
- [46] M. Steger, F. Tonelli, G. Ito, P. Davies, M. Trost, M. Vetter, S. Wachter, E. Lorentzen, G. Duddy, S. Wilson, M. A. Baptista, B. K. Fiske, M. J. Fell, J. A. Morrow, A. D. Reith, D. R. Alessi, M. Mann, *eLife* **2016**, *5*, <https://doi.org/10.7554/eLife.12813>
- [47] G. Sanna, M. G. Del Giudice, C. Crosio, C. Iaccarino, *Biochem. Soc. Trans.* **2012**, *40*, 1117.
- [48] X. Li, Q. J. Wang, N. Pan, S. Lee, Y. Zhao, B. T. Chait, Z. Yue, *PLoS One* **2011**, *6*, e17153.
- [49] A. Biosa, A. Trancikova, L. Civiero, L. Glauser, L. Bubacco, E. Greggio, D. J. Moore, *Hum. Mol. Genet.* **2013**, *22*, 1140.
- [50] D. Korr, L. Toschi, P. Donner, H.-D. Pohlenz, B. Kreft, B. Weiss, *Cell. Signal.* **2006**, *18*, 910.
- [51] I. Marin, *Mol. Biol. Evol.* **2006**, *23*, 2423.
- [52] N. D. Jorgensen, Y. Peng, C. C.-Y. Ho, H. J. Rideout, D. Petrey, P. Liu, W. T. Dauer, *PLoS One* **2009**, *4*, e8463.
- [53] E. Petsalaki, G. Zachos, *Nat. Commun.* **2016**, *7*, 11451.
- [54] M. Carmena, S. Ruchaud, W. C. Earnshaw, *Curr. Opin. Cell Biol.* **2009**, *21*, 796.
- [55] R. Mehra, I. G. Serebriiskii, B. Burtness, I. Atsaturov, E. A. Golemis, *Lancet. Oncol.* **2013**, *14*, e425.
- [56] B. E. Aubol, G. Wu, M. M. Keshwani, M. Movassat, L. Fattet, K. J. Hertel, X.-D. Fu, J. A. Adams, *Mol. Cell* **2016**, *63*, 218.
- [57] J. Prasad, K. Colwill, T. Pawson, J. L. Manley, *Mol. Cell. Biol.* **1999**, *19*, 6991.
- [58] M. P. Myers, M. B. Murphy, G. Landreth, *Mol. Cell. Biol.* **1994**, *14*, 6954.
- [59] M. Zu, C. Li, J.-S. Fang, W.-W. Lian, A.-L. Liu, L.-S. Zheng, G.-H. Du, *Molecules* **2015**, *20*, 19735.
- [60] P. Jain, C. Karthikeyan, N. S. H. N. Moorthy, D. K. Waiker, A. K. Jain, P. Trivedi, *Curr. Drug Targets* **2014**, *15*, 539.
- [61] M. R. Ray, L. A. Wafa, H. Cheng, R. Snoek, L. Fazli, M. Gleave, P. S. Rennie, *Int. J. Cancer* **2006**, *118*, 1108.
- [62] S. Kametaka, K. Moriyama, P. V. Burgos, E. Eisenberg, L. E. Greene, R. Mattera, J. S. Bonifacio, *Mol. Biol. Cell* **2007**, *18*, 2991.
- [63] H. Shimizu, I. Nagamori, N. Yabuta, H. Nojima, *J. Cell Sci.* **2009**, *122*, 3145.
- [64] A. Dumitriu, C. D. Pacheco, J. B. Wilk, K. E. Strathearn, J. C. Latourelle, S. Goldwurm, G. Pezzoli, J.-C. Rochet, S. Lindquist, R. H. Myers, *Hum. Mol. Genet.* **2011**, *20*, 1478.
- [65] M. A. Nalls, N. Pankratz, C. M. Lill, C. B. Do, D. G. Hernandez, M. Saad, A. L. DeStefano, E. Kara, J. Bras, M. Sharma, C. Schulte, M. F. Keller, S. Arepalli, C. Letson, C. Edsall, H. Stefansson, X. Liu, H. Pliner, J. H. Lee, R. Cheng, M. A. Ikram, J. P. A. Ioannidis, G. M. Hadjigeorgiou, J. C. Bis, M. Martinez, J. S. Perlmutter, A. Goate, K. Marder, B. Fiske, M. Sutherland, G. Xiromerisiou, R. H. Myers, L. N. Clark, K. Stefansson, J. A. Hardy, P. Heutink, H. Chen, N. W. Wood, H. Houlden, H. Payami, A. Brice, W. K. Scott, T. Gasser, L. Bertram, N. Eriksson, T. Foroud, A. B. Singleton, *Nat. Genet.* **2014**, *46*, 989.
- [66] C. Prunier, R. Prudent, R. Kapur, K. Sadoul, L. Lafanechere, *Oncotarget* **2017**, *8*, 41749.
- [67] B. T. Bettinger, D. C. Amberg, *J. Cell. Biochem.* **2007**, *101*, 34.
- [68] L.-X. Yang, Q. Gao, J.-Y. Shi, Z.-C. Wang, Y. Zhang, P.-T. Gao, X.-Y. Wang, Y.-H. Shi, A.-W. Ke, G.-M. Shi, J.-B. Cai, W.-R. Liu, M. Duan, Y.-J. Zhao, Y. Ji, D.-M. Gao, K. Zhu, J. Zhou, S.-J. Qiu, Y. Cao, Q.-Q. Tang, J. Fan, *Hepatology* **2015**, *62*, 1804.
- [69] H. Chi, M. R. Sarkisian, P. Rakic, R. A. Flavell, *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 3846.
- [70] Z. Wang, E. Horemuzova, A. Iida, L. Guo, Y. Liu, N. Matsumoto, G. Nishimura, A. Nordgren, N. Miyake, E. Tham, G. Grigelioniene, S. Ikegawa, *J. Hum. Genet.* **2017**, *62*, 503.
- [71] A. M. Fry, L. O'Regan, S. R. Sabir, R. Bayliss, *J. Cell Sci.* **2012**, *125*, 4423.
- [72] T. W. Bainbridge, V. I. DeAlmeida, A. Izrael-Tomasevic, C. Chalouni, B. Pan, J. Goldsmith, A. P. Schoen, G. A. Quinones, R. Kelly, J. R. Lill, W. Sandoval, M. Costa, P. Polakis, D. Arnott, B. Rubinfeld, J. A. Ernst, *PLoS One* **2014**, *9*, e102695.
- [73] A. H. Daneshmanesh, E. Mikaelsson, M. Jeddi-Tehrani, A. A. Bayat, R. Ghods, M. Ostadkarampour, M. Akhondi, S. Lagercrantz, C. Larsson, A. Osterborg, F. Shokri, H. Mellstedt, H. Rabbani, *Int. J. Cancer* **2008**, *123*, 1190.
- [74] H. Karvonen, W. Niininen, A. Murumagi, D. Ungureanu, *Biochem. Soc. Trans.* **2017**, *45*, 457.
- [75] C. Preisinger, B. Short, V. De Corte, E. Bruyneel, A. Haas, R. Kopajtich, J. Gettemans, F. A. Barr, *J. Cell Biol.* **2004**, *164*, 1009.
- [76] E. Cansby, M. Amrutkar, L. M. Holm, A. Nerstedt, A. Reyahi, E. Stenfeldt, J. Boren, P. Carlsson, U. Smith, J. R. Zierath, M. Mahlapuu, *FASEB J.* **2013**, *27*, 3660.
- [77] M. Amrutkar, E. Cansby, E. Nunez-Duran, C. Pirazzi, M. Stahlman, E. Stenfeldt, U. Smith, J. Boren, M. Mahlapuu, *FASEB J.* **2015**, *29*, 1564.

PINOT: An Intuitive Resource for Integrating Protein-Protein Interactions

Tomkins JE¹, Ferrari R², Vavouraki N¹, Hardy J², Lovering RC³, Lewis PA^{1,2}, McGuffin LJ^{4*}, and Manzoni C^{1,2*}

1. School of Pharmacy, University of Reading, Whiteknights, Reading, RG6 6AP, United Kingdom; 2. Department of Molecular Neuroscience, UCL Institute of Neurology, Queen Square, London, WC1B 5EH, United Kingdom; 3. Functional Gene Annotation, Institute of Cardiovascular Science, University College London, WC1E 6JF, United Kingdom; 4 School of Biological Sciences, University of Reading, Whiteknights, Reading, RG6 6AS, United Kingdom.

*Corresponding authors: c.manzoni@reading.ac.uk; l.j.mcguffin@reading.ac.uk

Abstract (350)

Background: The past decade has seen the rise of omics data, for the understanding of biological systems in health and disease. This wealth of data includes protein-protein interaction (PPI) derived from both low and high-throughput assays, which is curated into multiple databases that capture the extent of available information from the peer-reviewed literature. Although these curation efforts are extremely useful, reliably downloading and integrating PPI data from the variety of available repositories is challenging and time consuming.

Methods: We here present a novel user-friendly web-resource called PINOT (Protein Interaction Network Online Tool; available at http://www.reading.ac.uk/bioinf/PINOT/PINOT_form.html) to optimise the collection and processing of PPI data from the IMEx consortium associated repositories (members and observers) and from WormBase for constructing, respectively, human and *C. elegans* PPI networks.

Results: Users submit a query containing a list of proteins of interest for which PINOT will mine PPIs. PPI data is downloaded, merged, quality checked and confidence scored based on the number of distinct methods and publications in which each interaction has been reported. Examples of PINOT applications are provided to highlight the performance, the ease of use and the potential applications of this tool.

Conclusions: PINOT is a tool that allows users to survey the literature, extracting PPI data for a list of proteins of interest. The comparison with analogous tools showed that PINOT was able to extract similar numbers of PPIs while incorporating a set of innovative features. PINOT processes both small and large queries, it downloads PPIs live through the PSICQUIC and it applies quality control filters on the downloaded PPI annotations (i.e. removing the need of manual inspection by the user). PINOT provides the user with information on detection methods and publication history for each of the downloaded interaction data entries and provides results in a table format that can be easily further customised and/or directly uploaded in a network visualization software.

Keywords: protein interaction, protein network, network, data mining, protein database

Background

During the past two decades the use of omics data to understand biological systems has become an increasingly valued approach (1). This includes extensive efforts to detect protein-protein interactions (PPIs) on an almost proteome-wide scale (2, 3). The utility of such data has been greatly supported by primary database curation and the International Molecular Exchange (IMEx) Consortium, which promotes collaborative efforts in standardising and maintaining high quality data curation across the major molecular interaction data repositories (4). The primary databases, such as IntAct (5) and BioGRID (6), are rich data resources providing a comprehensive record of published PPI literature. PPI data are critical to describe connections among proteins, which in turn supports both inference of new functions for proteins (based on the guilt by association principle (7)) and visualization of protein connectivity via shared interactors, thus shedding light on communal pathways involving proteins of interest (8-10). Additionally, literature extracted PPI data can support the prioritization of interactions from high-throughput experiments which generate large lists of potential PPI hits, therefore assisting the selection of candidates for further analysis/validation (11).

However, the process of collating PPI data from multiple sources is currently hampered by the fact that no single data source encompasses the full extent of PPIs reported in the literature, requiring users to merge (partial) information mined from different primary databases. Furthermore, merging such data is not straightforward

due to inconsistencies in data format and differences in data curation across the PPI databases (IMEx members vs non-members).

To optimize the use of PPI data from the public domain, we developed a user-friendly tool that assists PPI data extraction and processing: the Protein Interaction Network Online Tool (PINOT). This tool represents the development (and automation) of our previous PPI analysis framework (i.e. weighted protein-protein interaction network analysis - WPPINA) (9, 11-15). Through PINOT, PPI data is downloaded directly (i.e. downloaded “live” at the time of the query) from seven databases using the Proteomics Standard Initiative Common Query Interface (PSICQUIC) and integrated to ensure a wide coverage of the PPIs available from these repositories (16). These data are scored through a simple and transparent procedure based on ‘method detection’ and ‘publication records’ and allows the user to further apply customized confidence thresholds. PINOT is fully automated and available online as an open access resource. Output data are provided as a summary table (directly online or emailed to the user), which summarizes the most comprehensive current knowledge of the PPI landscape for the protein(s)-of-interest submitted in the query list. Of note, the R scripts which underlie PINOT can be freely downloaded from GitHub and can be customized by the users.

Methods

Protein Interaction Network Online Tool (PINOT)

PINOT can be run automatically at http://www.reading.ac.uk/bioinf/PINOT/PINOT_form.html (hereafter referred to as “webservice”). A choice of parameters is integrated by default as explained further below and in Supplementary Materials (S1). Alternatively, R scripts can be downloaded from [GitHub – [http...](http://www.github.com)] (hereafter referred to as “standalone tool”, since parameters can be modified as *per* user choice).

A list of proteins of interest (seeds) can be queried to identify their literature-reported interactors that have been curated into PPI databases (Figure 1).

(A) PINOT Interface

PINOT: Protein Interaction Network Online Tool (Version 1.0)

This form allows you to run the PINOT pipeline on our servers.

Required - EITHER upload a file containing a single column list of UniProt, HUGO or WormBase identifiers here:

Choose File no file selected

OR paste a single column list of UniProt, HUGO or WormBase identifiers here: [Help](#)

Required - Select organism [Help](#)
Homo sapiens

Required - Select filter level [Help](#)
 Stringent
 Lenient

Optional - E-mail address (you will be sent a link to your results and email attachments) [Help](#)

Optional - Short subject name for your submission [Help](#)

Reset Submit

(B) Query Input Examples

Q55007
Q385D2
P53355
P37840
P10636
Swiss-Prot UniProt ID

OR

LRRK2
LRRK1
DAPK1
SNCA
MAPT
HGNC approved symbol

OR

WBGene00016600
WBGene00007303
WBGene00000272
WBGene00001578
WBGene00004927
WormBase gene ID

(C) Result Output Example

| NameA | SwissA | EntrezA | NameB | SwissB | EntrezB | MethodScore | Method | PublicationScore | PMIDS | FinalScore |
|-------|--------|---------|------------|--------|---------|-------------|---------------------------|------------------|---------------------------------|------------|
| DAPK1 | P53355 | 1612 | ABL1 | P05519 | 25 | 1 | Array | 1 | pubmed:25113927 | 2 |
| DAPK1 | P53355 | 1612 | ABL1M1 | O14639 | 3983 | 1 | Array | 1 | pubmed:25113927 | 2 |
| DAPK1 | P53355 | 1612 | ACTC1 | P68032 | 70 | 2 | Chromatography/CoIP | 1 | pubmed:25852190 | 3 |
| DAPK1 | P53355 | 1612 | AHDG1 | Q5TGY3 | 27245 | 1 | Array | 1 | pubmed:25113927 | 2 |
| DAPK1 | P53355 | 1612 | ARPC2 | Q9U924 | 27301 | 1 | Array | 1 | pubmed:25113927 | 2 |
| DAPK1 | P53355 | 1612 | ARRDC1 | Q8NS52 | 92714 | 1 | Array | 1 | pubmed:25113927 | 2 |
| DAPK1 | P53355 | 1612 | ARRDC1-AS1 | Q9H2J1 | 85026 | 1 | Array | 1 | pubmed:25113927 | 2 |
| DAPK1 | P53355 | 1612 | ATF6 | P18850 | 23926 | 1 | Array | 1 | pubmed:25113927 | 2 |
| DAPK1 | P53355 | 1612 | BAG2 | Q9P816 | 9532 | 3 | Chromatography/Array/CoIP | 2 | pubmed:25852190;pubmed:25113927 | 5 |
| DAPK1 | P53355 | 1612 | BAIAP2 | Q9UQ88 | 10458 | 1 | Array | 1 | pubmed:25113927 | 2 |

(D) Discarded Proteins Log File Example

final_network_log.txt
proteins_dropped
NO proteins removed

(E) Network Providers Log File Example

final_network_providers.txt
BioGrid
bhf-ucl
IntAct
MINT
UniProt
MBInfo
InnateDB

FIGURE 1 – PINOT user interface

A. Screenshot of the PINOT webpage, B. Examples of the text file to be uploaded or list to be populated into the text box of query seeds (i.e. proteins for which protein interactors will be extracted from primary databases that manually curate the literature), C. Example result output file from PINOT, containing the extracted and processed PPI data (only the file's header is reported as an example), D. Example of the discarded proteins log file from PINOT, a text file reporting all the seeds for which interactions are not returned to the user, and E. Example of the network providers log file from PINOT containing a list of active databases that were utilised for downloading PPI data.

For *Homo sapiens* (taxonomy ID: 9606) the seed identifiers submitted into the query field must be in an approved HUGO Gene Nomenclature Committee (HGNC) gene symbol or valid Swiss-Prot UniProt ID format. Upon query submission, PPI data are extracted directly (via API: Shannon, P. (2018) PSICQUIC R package, DOI: 10.18129/B9.bioc.PSICQUIC (17)) from seven primary databases, all of which directly annotate PPI data from peer-reviewed literature: bhf-ucl, BioGRID (6), InnateDB (18), IntAct (5), MBIInfo (<https://www.mechanobio.info>), MINT (19) and UniProt (20). The downloaded protein interaction data are then parsed, merged, filtered and scored (Figure 2) automatically by PINOT. Detailed description of the PINOT pipeline can be found in the supplementary materials (S1). The user can select to run PINOT with lenient or stringent filter parameters. The output of PINOT (Figure 1C-E) consists of: i) a network file (final_network.txt), which is a tab-spaced text file containing the processed PPI data in relation to the seeds in the initial query list; ii) a log file

(final_network_log.txt) reporting proteins that have been discarded from the initial query list, and; iii) a log file (final_network_providers.txt) indicating the PPI databases used by the API at download. The output dataset is available for download and/or emailed to the user.

For *Caenorhabditis elegans* (taxonomy ID: 6239) the seed identifiers must be in an approved WormBase gene ID (21) format, “WBGene” followed by 8 numerical digits. Upon submission PPI data are downloaded from an internal network stored within PINOT and created (following similar criteria applied for the human PPIs - details in S1) based on the WormBase PPI catalogue (Alliance_molecular_interactions.tar file downloaded from the Alliance of Genome Resources on 15th April 2019). The user can apply stringent or lenient filtering options. The output of PINOT for a *C. elegans* query consists of: i) a network file (final_network.txt), which is a tab-spaced text file containing the processed PPIs for the seeds in the initial query list; and ii) a log file (final_network_log.txt) reporting proteins that have been discarded from the initial query list.

Software

The PINOT pipeline is coded in R and runs on a Linux server at the University of Reading, with java servlets processing user’s submissions *via* the web interface.

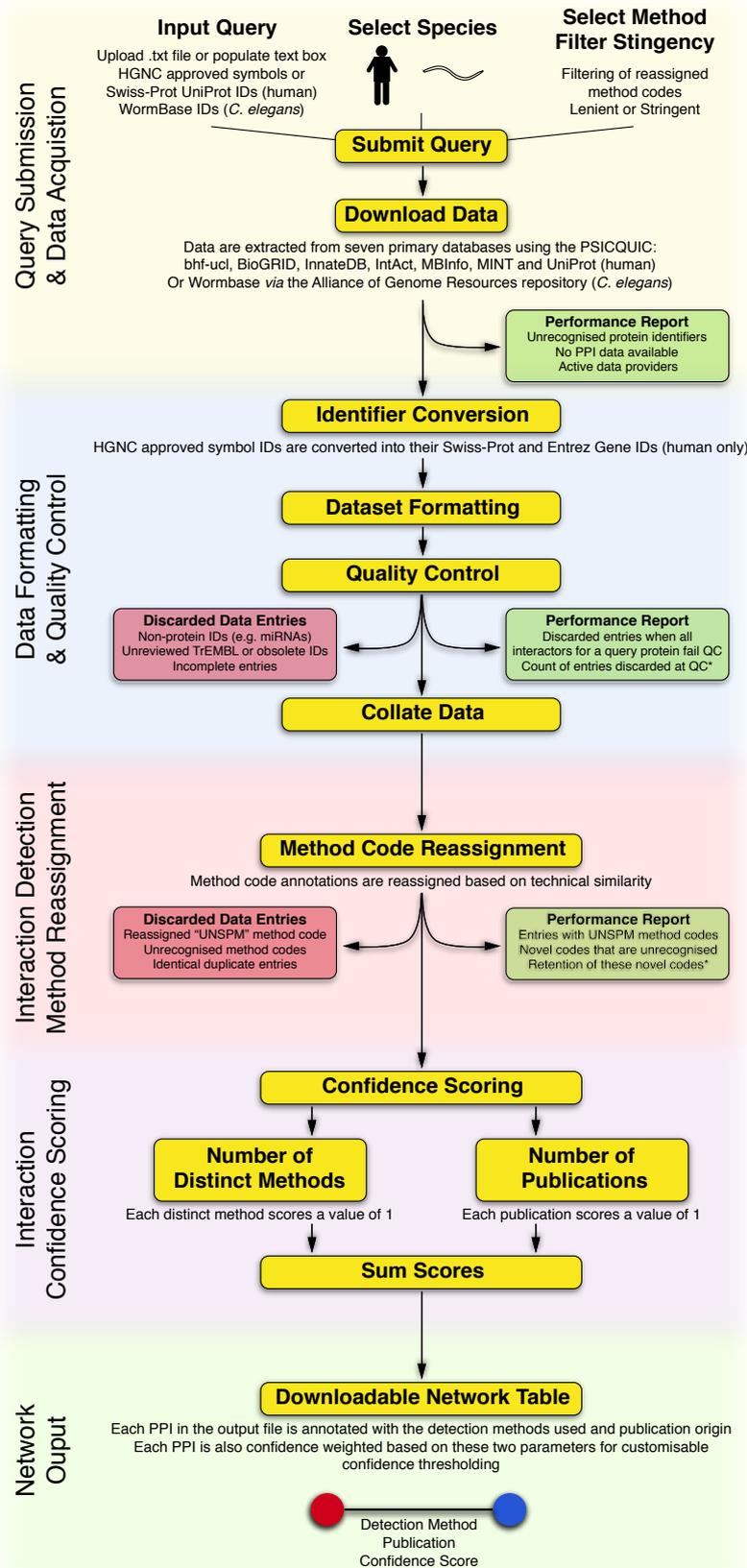
PINOT quality control

We have tested the PINOT pipeline using multiple input query lists structured as follows: i) small input lists = 6 sets of 1 to 5 proteins, selected randomly or in association with typical processes suspected to be functionally relevant for Parkinson’s Disease (PD); and ii) large input list = 941 proteins, the mitochondrial proteome as reported by MitoCarta2.0 (22).

PINOT was compared to two other related online tools. For this analysis, searching parameters were selected (where possible) to maximize the extraction of protein interactions: the Human Integrated Protein-Protein Interaction Reference (HIPPIE) was used with confidence score = 0 and no filters on confidence level, interaction type or tissue expression; and the Molecular Interaction Search Tool (MIST) was used with no filtering rank to download only protein protein interactions. Importantly and of note, files from HIPPIE and MIST required manual parsing after download to remove entries that were associated to no PMID and/or no conversion method code. Data were downloaded on 18th September 2019.



Protein Interaction Network Online Tool



*features available when running the pipeline manually

FIGURE 2 – PINOT pipeline

A stepwise insight into the process which underlies the PINOT pipeline. Performance reports (green boxes) are generated and data are discarded (red boxes) at numerous stages within the pipeline to ensure high quality and transparent data processing.

Results

PINOT is a webtool that takes a list of proteins/genes (seeds) as input and returns a table containing a comprehensive list of PPIs - published in peer-reviewed literature – centred upon the seeds. This table consists of a variable number of rows and 11 columns (Figure 1 and 3). Each row represents a binary interaction between one of the seeds (interactor A) and one of its specific protein interactors (interactor B). The 11 columns contain: the gene name, the Swiss-Prot protein ID and the Entrez gene ID for interactor A and B (“NameA”, “SwissA”, “EntrezA”, “NameB”, “SwissB”, “EntrezB”); the number and type of different methods through which the interaction has been identified (“Method.Score”, “Method”); and the number of different publications reporting the interaction and the corresponding PubMed IDs (“Publication.Score”, “PMIDS”). The final column (“Final.Score”) contains a confidence score based on the number of different methods + the number of different publications reporting the interaction. PPIs with a final score of 2 are reported in literature by 1 publication and detected by 1 technique; these PPIs are considered “suggestive” (but are clearly not “replicated”). They might be either: i) false positives, or ii) true novel interactions that have not yet been replicated in additional studies. A final score >2 suggests a degree of replication that can be either or both: multiple publications reporting the PPI and multiple techniques used to detect the interaction. It is not possible to obtain a final score <2 since every PPI annotation – to be retained in PINOT – has to be supported by at least 1 interaction detection method and 1 PMID; if this condition is not met, the PPI is discarded by PINOT and not shown in the output file.

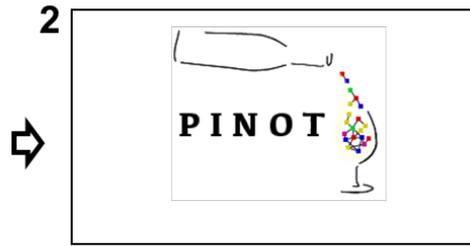
The PINOT output can be imported into Cytoscape (23) or yED (<https://www.yworks.com/products/yed>) for network visualization by selecting the “NameA” and “NameB” columns as source and target nodes, respectively.

PINOT: Example of application

In Figure 3 PINOT has been used to download PPIs for a limited selection of human protein products of genes mutated in familial PD: ATP13A2, FBXO7, GBA, PINK1, SMPD1 and VPS35. PINOT quickly retrieved a table containing 327 interactions from

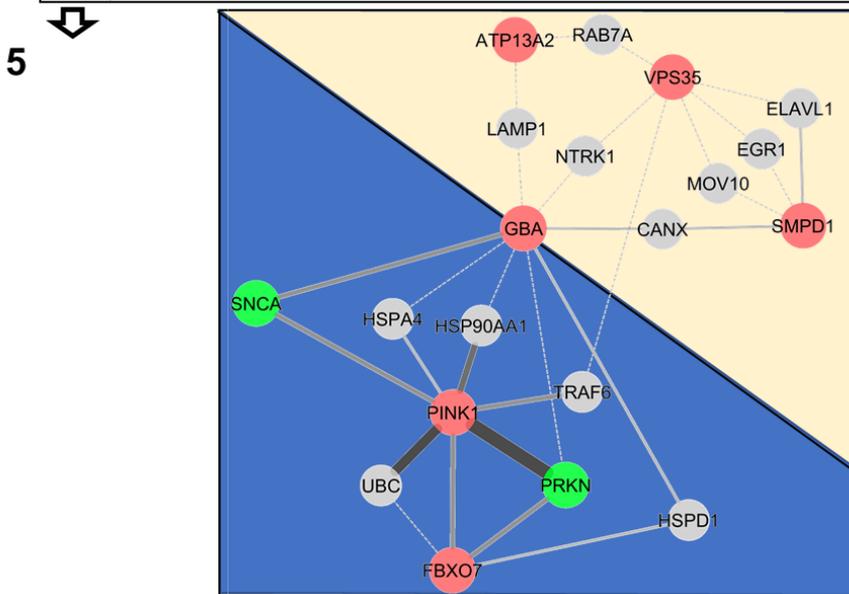
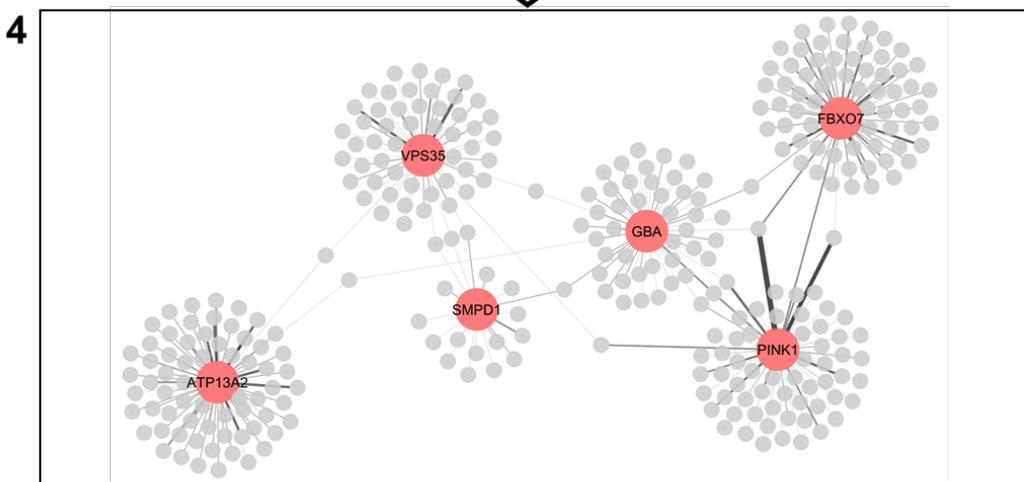
peer-reviewed literature (with associated PMIDs) thus supporting and simplifying otherwise time-consuming classical literature mining. The PINOT output was imported into Cytoscape and PPIs were visualized in a network (“NameA” = source and “NameB” = target), the seeds were highlighted in dark-red and the edges (interactions between each protein) were coded based on the “Final.Score” field, thus highlighting the confidence (number of methods + number of publications) of the interaction. Since we were interested in interactors that were common to the seeds in the initial list - and not interactors of just one seed - the network was filtered retaining only the nodes (interactors) that bridged two or more seeds. The obtained core-network revealed that among the common interactors of the seeds (PD proteins) there were 2 proteins (SNCA and PRKN), which are products of 2 additional genes known for being mutated in familial PD as it was for those in the initial seed list used for the query. Additionally, topological analysis (based on the number and strength of the edges) suggested the core network could be subdivided into 2 distinct clusters respectively including PINK1, FBXO7 and the newly identified PRKN and SNCA in the first cluster, while ATP13A2, VPS35 and SMPD1 were more closely associated in the second cluster, with GBA a bridge seed between the 2 clusters. This observation suggested a dichotomy, based on the protein interactomes, of the seeds included in the initial input list. Based on the guilt-by-association principle we hypothesised that the proteins contributing to these clusters could be associated with different cellular functions and components. We therefore performed functional enrichment analysis (based on Gene Ontology (GO) Cellular Component (CC) annotations) using g:Profiler (24) revealing that indeed, clusters 1 and 2 are associated with mitochondria and vacuoles/lysosomes/endosomes, respectively.

- 1
- | | |
|---|---------|
| 1 | ATP13A2 |
| 2 | FBXO7 |
| 3 | GBA |
| 4 | PINK1 |
| 5 | SMPD1 |
| 6 | VPS35 |
- or
- | | |
|---|--------|
| 1 | Q9NQ11 |
| 2 | Q9Y3I1 |
| 3 | P04062 |
| 4 | Q9BXM7 |
| 5 | P17405 |
| 6 | Q96QK1 |



3

| | A | B | C | D | E | F | G | H | I | J | K |
|----|---------|--------|---------|----------|--------|---------|--------------|------------------------|-------------------|-------------------------|-------------|
| 1 | NameA | SwissA | EntrezA | NameB | SwissB | EntrezB | Method.Score | Method | Publication.Score | PMIDS | Final.Score |
| 2 | ATP13A2 | Q9NQ11 | 23400 | LAMP1 | P11279 | 3916 | 1 | I | 1 | pubmed:22768177 | 2 |
| 3 | ATP13A2 | Q9NQ11 | 23400 | RAB7A | P51149 | 7879 | 1 | I | 1 | pubmed:22768177 | 2 |
| 4 | FBXO7 | Q9Y3I1 | 25793 | HSPD1 | P10809 | 3329 | 2 | MSI;COIP | 1 | pubmed:26496610 | 3 |
| 5 | PINK1 | Q9BXM7 | 65018 | HSP90AA1 | P07900 | 3320 | 1 | MSI | 4 | ed:18003639;pubmed:2293 | 5 |
| 6 | PINK1 | Q9BXM7 | 65018 | PRKN | O60260 | 5071 | 7 | APMS;MSI;E;G;FT;COIP;I | 18 | ed:19966284;pubmed:2611 | 25 |
| 7 | FBXO7 | Q9Y3I1 | 25793 | PINK1 | Q9BXM7 | 65018 | 3 | MSI;APMS;COIP | 1 | pubmed:23933751 | 4 |
| 8 | FBXO7 | Q9Y3I1 | 25793 | PRKN | O60260 | 5071 | 3 | MSI;APMS;COIP | 1 | pubmed:23933751 | 4 |
| 9 | FBXO7 | Q9Y3I1 | 25793 | UBC | P0CG48 | 7316 | 1 | MSI | 1 | pubmed:16196087 | 2 |
| 10 | GBA | P04062 | 2629 | CANX | P27824 | 821 | 2 | MSI;COIP | 1 | pubmed:26496610 | 3 |
| 11 | GBA | P04062 | 2629 | HSP90AA1 | P07900 | 3320 | 1 | MSI | 1 | pubmed:22160715 | 2 |



- 6
- | | | |
|---|--|--|
| perinuclear region neuron projection inclusion body | cell body neuronal cell body | lysosome lysosomal membrane |
| mitochondrion mitochondrial envelope mitochondrial membrane | lytic vacuole lytic vacuole membrane vacuole vacuolar membrane vacuolar part | endosome endosomal part endosome membrane late endosome |

Figure 3 – PINOT: An example application

A stepwise insight into the potential use of PINOT. 1. A submission list is created as a text file using gene names as per HGNC approved symbols or Swiss-Prot IDs; the submission list can be uploaded as file or pasted into the PINOT interface. 2. PINOT downloads from PSICQUIC the human PPIs (in this example, stringent filters applied) 3. PPIs are provided back to the user *via* email or from the webpage; results are in a parsable file that can be opened by a text reader application and imported into Microsoft Excel, for example. 4. The interactions can be visualized in a network format by opening the PINOT output through Cytoscape. Connections between nodes (edges) are coded with increased line width based on the final score that interaction was assigned by PINOT. The wider the edge – the more confident PINOT is about the interactions. 5. The interactions can be further processed according to the user's research question, in this case, only interactors that are communal to at least 2 of the initial query proteins have been retained, generating a core network (in dark-red the initial seeds; in bright-green the identified common interactors that are proteins mutated in PD). Based on the network topology the seeds and their interactors can be visually clustered into group 1 (depicted in gold) and group 2 (depicted in blue). 6. Specific functional enrichment (GO CC terms) for groups 1 and 2 after filtering out the less represented terms. Analyses performed on the 22/08/2019.

Human-PINOT performance

The performance of PINOT was compared to that of alternative resources for both a small and large number of seeds. Regarding the former, five different small seed lists were used as input for PPI query in HIPPIE (25) and MIST (26), two alternative online and freely available resources. It should be noted that, despite apparent similarities, each of these tools has been developed differently. All three resources (PINOT, HIPPIE and MIST) have distinguishing features for addressing different research questions (Table 1). The results of the different queries have been compared, by evaluating the total number of interactors provided in the output (Figure 4A).

PINOT, HIPPIE and MIST retrieved a similar number of PPIs. PINOT with stringent filtering applied, was always extracting fewer interactions; this is an expected outcome since this filter option is built with the purpose of retaining only annotations that have survived stringent screening, largely based on completeness of curated data entries.

The large input list was queried in PINOT and HIPPIE, the only two servers that allowed for processing more than 900 seeds within the submission list. In fact, MIST submission needed to be divided into multiple small lists to allow the browser to properly process the query. Additionally, the downloaded table(s) were not parsable (in an automated fashion), thus making MIST counterintuitive for the processing of large input lists. The number of retrieved interactors was slightly higher for HIPPIE in

comparison with PINOT when the stringent QC filter was applied. PINOT with lenient filtering applied retrieved more interactions than HIPPIE (Figure 4B). Additionally, the vast majority of downloaded interactions were similar from using the two resources, suggesting that PINOT is able to extract specific interactions from literature (Figure 4C).

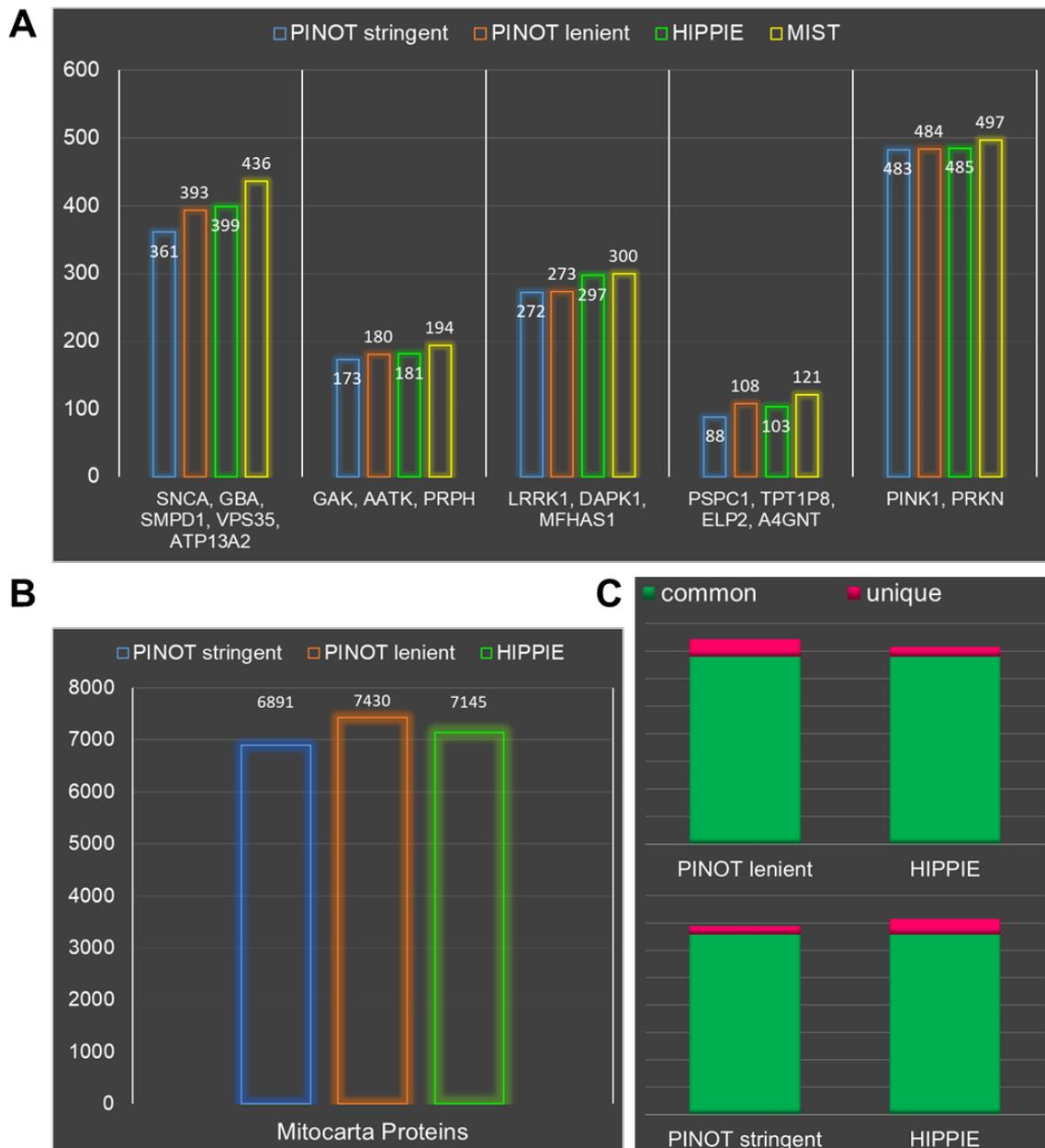


Figure 4 – PINOT: Performance & Sensitivity

A. PINOT performance was evaluated by counting the number of interactors retrieved (gene names) upon submission of the reported query lists to human-PINOT (with stringent and lenient filtering), HIPPIE and MIST (on 18th September 2019). The databases were set to

retrieve the maximum number of interactions (by removing all possible filters). The HIPPIE and MIST outputs were manually cleaned to remove interactions with i) no interaction detection method; ii) no PMID; iii) multiple Entrez IDs. The number of retained interactions retrieved is reported on top of each bar. B. PINOT (with stringent and lenient filtering) and HIPPIE were queried to retrieve PPIs for a seed list of 941 protein from Mitocarta 2.0. C. Comparison between PINOT and HIPPIE showing that the vast majority of interactors (Entrez IDs) downloaded by the two tools was identical: 6790 common interactors for PINOT lenient (640 unique interactors) vs HIPPIE (355 unique interactors); 6572 common interactors for PINOT stringent (319 unique interactors) vs HIPPIE (573 unique interactors).

C. elegans-PINOT performance

The performance of PINOT for querying *C. elegans* PPI data was tested alongside the *C. elegans* query option in MIST, assessing interaction networks of different dimensions (Figure 5). The data acquisition strategy underlying these two resources differs slightly, PINOT extracts data from the latest release of WormBase molecular interaction data, whereas MIST utilises data from numerous sources, including WormBase, BioGRID and IMEx associated repositories.

Similarly to comparisons across the human PPI query capacity, PINOT and MIST performed comparably in terms of the number of PPI data entries extracted. More specifically and as previously described with human data, PINOT extracting slightly fewer across these test query cases (Figure 5). However, upon assessing the completeness of these extracted data entries, in terms of interaction detection method and/or PMID annotations, there was a striking difference in performance. Since the PINOT pipeline focusses particular emphasis on the QC of data, all data entries within the output dataset were complete, whereas incomplete data entries persisted in the MIST output dataset. In the more abundant PPI data pools, for example when querying the ATP and CED *C. elegans* proteins (Figure 5), incomplete data entries accounted for the majority of the output dataset in MIST.

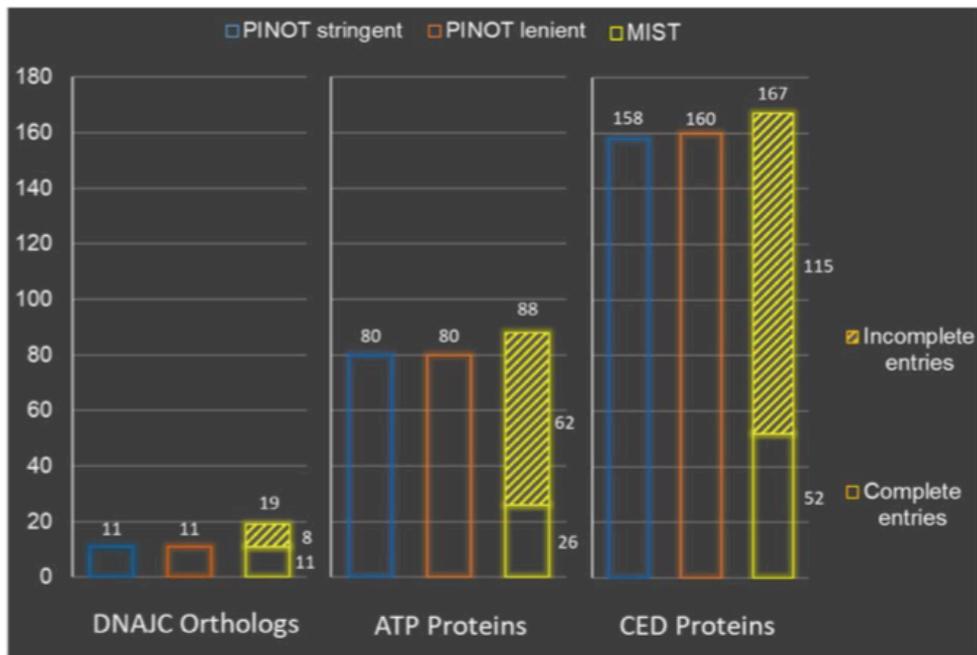


Figure 5 – PINOT and MIST performance comparison for *C. elegans* PPI data

The performance of PINOT (with stringent and lenient filter options) and MIST was assessed in terms of the number of PPI data entries extracted upon querying different protein lists (on 24th September 2019). The output dataset was evaluated in relation to the number of complete and incomplete (lacking interaction detection method and/or PubMed ID annotations) data entries extracted. The query lists were PD-associated DNAJC orthologs: DNJ-14, DNJ-25, DNJ-27, Y73B6BL.12, K07F5.16, RME-8 and GAKH-1; ATP proteins: ATP-1, ATP-2, ATP-3, ATP-4, ATP-5 and ATP-6; and CED proteins: CED-1, CED-2, CED-3, CED-4, CED-5, CED-6, CED-7, CED-8, CED-9, CED-10, CED-11, CED-12 and CED-13. The input format used for PINOT was the WormBase gene ID, the common gene name (as listed here) was used for MIST querying and no filter by rank parameter was set.

Discussion

PINOT can be used as a tool to quickly and effectively survey the literature and download the most up-to-date PPI data available for a given set of proteins/genes of interest. This is particularly useful to assist anyone attempting to mine overwhelming abundant literature targeting certain proteins/genes, in relation identifying reported PPIs.

The PPI data downloaded through PINOT can be used as a reference list from literature for experimental PPI data resulting from high-throughput experiments (protein microarrays; yeast 2 hybrid screens, etc) helping in prioritisation of

experimental results for validation. PINOT can also be useful to evaluate interactors of different proteins/genes of interest within an input seed list simultaneously. The analysis of the combined interactomes of such seeds can reveal the existence of communal interactors, can provide a base to cluster the seeds into groups and can support further functional analysis to better characterize the functional landscape of seeds of interest.

Alternative tools that appear to be similar to PINOT are HIPPIE and MIST. STRING (27) is a conceptually different tool; it does not report 'interaction detection methods' nor 'Publication IDs' for PPIs which are crucial pieces of information for the evaluation and interpretation of PPI data. Additionally, the reported interactions are not focused only on the proteins in the input list, interactions of interactors are also reported, thus making it difficult to parse the output table. HIPPIE implements a tailored confidence score for different methodological approaches; MIST provides a valuable resource for users interested in mapping PPIs across species (i.e. interologs); PINOT focusses on high quality PPI data output by implementing multiple QC steps to remove problematic or non-univocal annotations. PINOT performance was comparable to that of HIPPIE and MIST both in terms of number and identity of downloaded interactions. However, there are some unique features of PINOT that are not, at the moment, integrated within the other databases. Human PPIs in PINOT are directly downloaded from PSICQUIC at every query submission. In contrast, PPIs in HIPPIE and MIST are recovered from a central built-in repository within the servers, implying that both HIPPIE and MIST require constant updates in order to retrieve the latest PPI data available. This is clearly demonstrated by searching for interactors of LRRK2, where (at the time of submission) 1 high-throughput publication is updated in PSICQUIC, while both HIPPIE and MIST do not contain this full annotation yet (Figure 6).

PINOT has access to the most up-to-date interactions that could be retrieved at a given time from PSICQUIC (however, it has to be considered that each database is responsible for updating their PSICQUIC service and therefore discrepancies might exist with the central database).

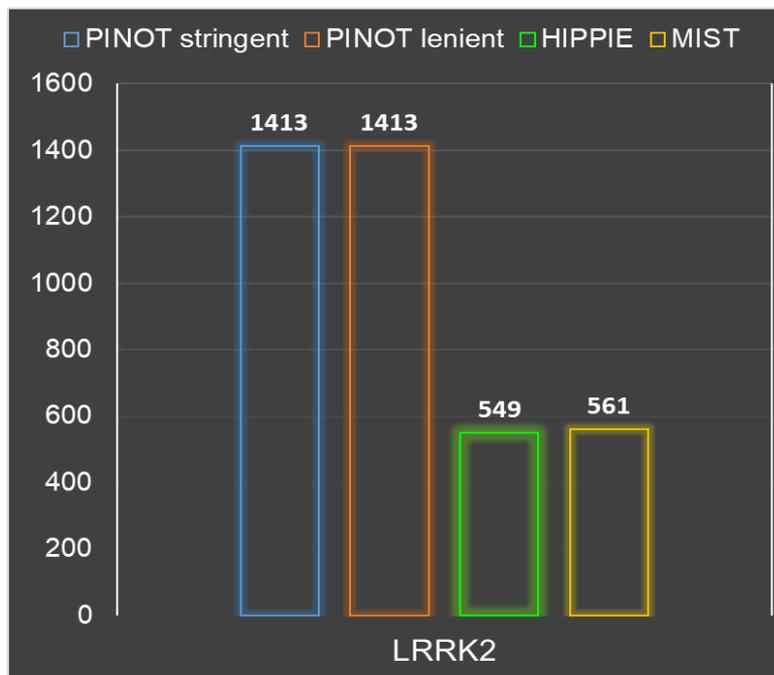


Figure 6 – LRRK2 interactome

PINOT performance was evaluated by counting the number of interactors retrieved (gene names) for LRRK2 by using PINOT (with stringent and lenient filtering), HIPPIE and MIST. The databases have been set to retrieve the maximum number of interactions (by removing all possible filters). HIPPIE and MIST output were manually cleaned to remove interactions with i) no interaction detection method; ii) no PMID; iii) multiple Entrez IDs. The number of the surviving interaction retrieved is reported on top of each bar (18-09-2019).

PINOT implements QC filtering which involves discarding PPI data entries that are curated without a PMID and/or the interaction detection method annotation. Therefore the output file from PINOT does not require further QC by the user, while lists from MIST and HIPPIE require manual parsing to remove incomplete data entries through a time consuming, post-hoc processing procedure.

Another distinctive feature of the PINOT pipeline is the implementation of a unique interaction detection method conversion step. During this step, the interaction detection method annotation for each downloaded interaction data entry is converted based on a conversion table (S2) that is available for download from the PINOT web-portal. During this conversion, technically similar methods are grouped together. For example: “Two Hybrid - MI:0018”, “Two Hybrid Array - MI:0397” and “Two Hybrid Pooling Approach - MI:0398” are grouped together into the “Two Hybrid” method

category. This step of 'method clustering and reassignment' is critical to assess the actual number of distinct methods used to describe a particular interaction and to dilute the bias caused in the event of the same technique being annotated under slightly different method codes in different PPI databases.

Interaction scores are provided in different formats for the three tools. HIPPIE incorporates a filtering system based on a confidence score between 0 and 1 that can be set either before or after the analysis. This is a complex scoring system, which takes into consideration multiple parameters, such as the number of publications that report a specific interaction and a semi-computational quality score based on the experimental approach (for example, imaging techniques would score less than direct interactions etc.) (28). MIST similarly has an option for filtering interactions pre- or post-analysis; however, this is based on fixed ranking values defined as low, medium (interaction supported by other species), or high (supported by multiple experimental methods and/or reported in multiple publications). In the case of PINOT, two different scores are provided: the interaction detection method score (MS) reports the number of different methods used (after conversion), while the publication score (PS) counts the number of different publications which report the interaction. Finally, PINOT coding scripts are fully available for download. They are coded in R to make them accessible to a large research audience; additionally a read me text file helps customization of the scripts according to the users' needs. Some of the divergent features across PINOT, HIPPIE, MIST and STRING are reported in Table 1.

| | PINOT | HIPPIE | MIST | STRING |
|-----------------------|-------|--------|------|--------|
| Live PPI data | yes | no | no | no |
| Large Submission | yes | yes | no | no |
| Parsable Table | yes | yes | no | yes |
| PPIs for seeds only | yes | yes | yes | no |
| Visualization app | no | yes | yes | yes |
| Other Species PPIs | yes | no | yes | yes |
| Score | yes | yes | yes | yes |
| Pubmed ID (PMID) | yes | yes | yes | no |
| Detection Method | yes | yes | yes | no |
| QC on method and PMID | yes | no | no | - |
| Entrez ID | yes | yes | yes | no |
| Swiss-Prot ID | yes | no | no | no |
| Codes available | yes | no | no | no |

Table 1 – Comparison of features available within the PINOT, HIPPIE, MIST and STRING resources.

List of Abbreviations

PPI = protein protein interaction; PINOT = protein interaction network online tool; PD = Parkinson's Disease; PMID = Pubmed ID; QC = Quality Control

Declarations

Consent for Publication: Not applicable

Availability of data: This resource is available as a fully automated web-server at: http://www.reading.ac.uk/bioinf/PINOT/PINOT_form.html; R scripts, which underlie this bioinformatics pipeline, are free for download.

Competing Interests: The authors declare that they have no competing interests

Funding: This work was supported by Biotechnology and Biological Sciences Research Council CASE studentship with BC Platforms [grant number BB/M017222/1]; Engineering and Physical Sciences Research Council [grant

number EP/M508123/1]; the Medical Research Council [grant numbers MR/N026004/1; MR/L010933/1 to JH and PAL]; Parkinson's UK to PAL [grant number F1002]; Alzheimer's Society to RF [grant number 284]; Alzheimer Research UK to RCL [grant number ARUK-NAS2017A-1]; the National Institute for Health Research University College London Hospitals Biomedical Research.

Author's Contribution: TEJ, MC and FR, elaborated the pipeline and wrote the manuscript; TEJ, MC and VN wrote the scripts and tested the pipeline, MJL implemented the pipeline for the website, HJ, LCR, LAP offered critical advice for the implementation of the pipeline, critically reviewed the manuscript and obtained financial support.

Acknowledgements: We thank WormBase for providing the list of PPIs used for the generation of the internal *C. elegans* network that is queried when *C. elegans* genes/proteins are queried by PINOT users.

Supplementary Files:

S1 = supplementary materials and methods

S2 = supplementary file 1: 'interaction detection method conversion'

References

1. Manzoni C, Kia DA, Vandrovцова J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief Bioinform.* 2018;19(2):286-302.
2. Hein MY, Hubner NC, Poser I, Cox J, Nagaraj N, Toyoda Y, et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell.* 2015;163(3):712-23.
3. Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, et al. Architecture of the human interactome defines protein communities and disease networks. *Nature.* 2017;545(7655):505-9.
4. Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, et al. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods.* 2012;9(4):345-50.
5. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, et al. The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* 2012;40(Database issue):D841-6.
6. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* 2015;43(Database issue):D470-8.
7. Bonetta L. Protein-protein interactions: Interactome under construction. *Nature.* 2010;468(7325):851-4.
8. Blokhuis AM, Koppers M, Groen EJ, van den Heuvel DM, Dini Modigliani S, Anink JJ, et al. Comparative interactomics analysis of different ALS-associated proteins identifies converging molecular pathways. *Acta Neuropathol.* 2016;132(2):175-96.

9. Ferrari R, Kia DA, Tomkins JE, Hardy J, Wood NW, Lovering RC, et al. Stratification of candidate genes for Parkinson's disease using weighted protein-protein interaction network analysis. *BMC Genomics*. 2018;19(1):452.
10. Hwang S, Son SW, Kim SC, Kim YJ, Jeong H, Lee D. A protein interaction network associated with asthma. *J Theor Biol*. 2008;252(4):722-31.
11. Tomkins JE, Dihanich S, Beilina A, Ferrari R, Ilacqua N, Cookson MR, et al. Comparative Protein Interaction Network Analysis Identifies Shared and Distinct Functions for the Human ROCO Proteins. *Proteomics*. 2018;18(10):e1700444.
12. Bonham LW, Steele NZR, Karch CM, Manzoni C, Geier EG, Wen N, et al. Protein network analysis reveals selectively vulnerable regions and biological processes in FTD. *Neurol Genet*. 2018;4(5):e266.
13. Ferrari R, Lovering RC, Hardy J, Lewis PA, Manzoni C. Weighted Protein Interaction Network Analysis of Frontotemporal Dementia. *J Proteome Res*. 2017;16(2):999-1013.
14. Rossi G, Redaelli V, Contiero P, Fabiano S, Tagliabue G, Perego P, et al. Tau Mutations Serve as a Novel Risk Factor for Cancer. *Cancer Res*. 2018;78(13):3731-9.
15. Taymans JM, Chartier-Harlin MC. In silico and Wet Bench Interactomics Sheds Light on the Similarities and Differences between Human ROCO Proteins. *Proteomics*. 2018;18(13):e1800103.
16. Aranda B, Blankenburg H, Kerrien S, Brinkman FS, Ceol A, Chautard E, et al. PSICQUIC and PSIScore: accessing and scoring molecular interactions. *Nat Methods*. 2011;8(7):528-9.
17. Shannon P. PSICQUIC R package DOI: 10.18129/B9biocPSICQUIC. 2018.
18. Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, et al. InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. *Nucleic Acids Res*. 2013;41(Database issue):D1228-33.
19. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, et al. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res*. 2012;40(Database issue):D857-61.
20. UniProt C. UniProt: a hub for protein information. *Nucleic Acids Res*. 2015;43(Database issue):D204-12.
21. Lee RYN, Howe KL, Harris TW, Arnaboldi V, Cain S, Chan J, et al. WormBase 2017: molting into a new stage. *Nucleic Acids Res*. 2018;46(D1):D869-D74.
22. Calvo SE, Clauser KR, Mootha VK. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res*. 2016;44(D1):D1251-7.
23. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-504.
24. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res*. 2019;47(W1):W191-W8.
25. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res*. 2017;45(D1):D408-D14.
26. Hu Y, Vinayagam A, Nand A, Comjean A, Chung V, Hao T, et al. Molecular Interaction Search Tool (MIST): an integrated resource for mining gene and protein interaction data. *Nucleic Acids Res*. 2018;46(D1):D567-D74.
27. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res*. 2017;45(D1):D362-D8.
28. Schaefer MH, Fontaine JF, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA. HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PLoS One*. 2012;7(2):e31826.