INAUGURAL – DISSERATION

zur
Erlangung der Doktorwürde
der
Gesamtfakultät für Mathematik, Ingenieur- und Naturwissenschaften
der
Ruprecht – Karls – Universität
Heidelberg

vorgelegt von

GEORGIEVA, Magdalena Marin, MMath
aus Bulgarien

Tag der mündlichen Prüfung : …………………

Statistical analysis and modelling of proteomic and genetic network data illuminate hidden roles of proteins and their connections

Betreuer : Prof. Dr. Anna Marciniak-Czochra
Prof. Dr. Robert Russell

Acknowledgements

I would have liked to be brief, but there are a lot of people who I am immensely grateful to for helping me get to this point. I would like to start with my two supervisors, Prof. Anna Marciniak-Czochra and Prof. Rob Russell, who created this cross-discipline research opportunity for me to learn and grow towards what I envisioned. They supported me with absolutely anything that I have asked for. I am also very grateful for all my colleagues in both groups, who were always willing to help. And especially: JC, Gurdeep, Torsten, Jiao (all four doctors already) and Benni for providing feedback on my thesis drafts, and Filip for always being extremely helpful and supportive.

Biology is best learnt through practice and I am extremely grateful for the collaborators which I visited and who took time to teach me some of the experimental methods. Specifically, Dr. Karsten Boldt, Shibu Antony, Prof. Esben Lorentzen, Nevin Zacharia and all the great colleagues I met in the Ueffing group in the Institute for ophthalmic research in Tübingen and in the Lorentzen group in Aarhus University. Another special thanks goes to Dr. Narcis Petriman who, beyond teaching me how to write an article, is a great collaborator.

I was fortunate to be a part of the SCilS European Training Network through which I got very useful training, many opportunities to present my research, collaborate and get feedback. I would like to thank my colleagues in the network who, through trusting me to analyse their data, have enriched my understanding about biological experiments. For that, I would also like to recognise the Roepman lab in Radboud University and, specifically, Prof. Ronald Roepman and Dr. Mariam Aslanyan who encouraged my involvement in analysing the projects of their lab.

I came here because of several great teachers in the past and this is not a biography, so I only mention the ones who put me on this specific (PhD) path, Julia Haralambieva and Dr. Andrew Mellor. Thank you!

I am grateful to my friends, who were on the same journey with me, helped me with the equations or supported me from outside the academic track. And finally, I am grateful to my family for supporting me and, especially, to Gogi, Vanko and Nikola for always being there and helping me push through.

Abstract

While many stable protein complexes are known, the dynamic interactome is still underexplored. Experimental techniques such as single-tag affinity purification, aim to close the gap and identify transient interactions, but need better filtering tools to discriminate between true interactors and noise.

This thesis develops and contrasts two complementary approaches to the analysis of protein-protein interaction (PPI) networks derived from noisy experiments. The majority of data used for the analysis come from *in vitro* experiments aggregated from known databases (IntAct, BioGRID, BioPlex), but is also complemented by experiments done by our collaborators from the Ueffing group in the Institute of Ophthalmic Research, Tübingen University (Germany).

Chapter 3 presents the statistical approach to the data analysis. It focuses on the case of a single dataset with target and control data in order to determine the significant interactions for the target. The procedure follows an expected trajectory of preprocessing, quality control, statistical testing, correction and discussion of results. The approach is tailored to the specific dataset, experiment design and related assumptions. This is specifically relevant for the missing value imputation where multiple approaches are discussed and a new method, building upon a previous method, is proposed and validated.

Chapter 4 presents a different approach for the filtering of experimental results for PPIs. It is a statistic, WeSA (weighted socio-affinity), which improves upon previous methods of scoring PPIs from affinity proteomics data. It uses network analysis techniques to analyse the full PPI network without the need for controls. WeSA is tested on protein-protein networks of varying accuracy, including the curated IntAct dataset, the unfiltered records in BioGRID, and the large BioPlex dataset. The model is also tested against the previous same-goal method. While the function itself proves superior, another major advantage is that it can efficiently combine and compare observations across studies and can therefore be used to aggregate and clean results from incoming experiments in the context of all already available data.

In the final part, uses of WeSA beyond wild-type PPI networks are analysed. The framework is proposed as a novel way to effectively compare mechanistic differences between variants of the same protein (e.g. mutant vs wild type). I also explore the use of WeSA to study other biological and non-biological networks such as genome-wide association studies (GWAS) and gene-phenotype associations, with encouraging results.

In conclusion, this work presents and compares a variety of mathematical, statistical and computational approaches adapted, combined and/or developed specifically for the task of obtaining a better overview of protein-protein interaction networks. The novel methods performance is validated and, specifically, WeSA, is extensively tested and analysed, including beyond the field of PPI networks.

Zusammenfassung

Während viele stabile Proteinkomplexe bekannt sind, ist das dynamische Interaktom noch wenig erforscht. Experimentelle Techniken wie die Single-Tag-Immunpräzipitation und die Affinitätsreinigung zielen darauf ab, die Lücke zu schließen und flüchtige Wechselwirkungen zu identifizieren, benötigen aber bessere Filterwerkzeuge, um zwischen echten Interaktoren und Rauschen zu unterscheiden.

In dieser Arbeit werden zwei komplementäre Ansätze zur Analyse von Protein-Protein-Interaktionsnetzwerken (PPI) aus verrauschten Experimenten entwickelt und gegenübergestellt. Die meisten Daten, die für die Analyse verwendet werden, stammen aus In-vitro-Experimenten, die aus bekannten Datenbanken (IntAct, BioGRID, BioPlex) zusammengetragen wurden, werden aber auch durch Experimente unserer Mitarbeiter aus der Gruppe Ueffing am Institut für Augenheilkunde der Universität Tübingen (Deutschland) ergänzt.

Kapitel 3 stellt den statistischen Ansatz für die Datenanalyse vor. Es konzentriert sich auf den Fall eines einzelnen Datensatzes mit Ziel- und Kontrolldaten, um die signifikanten Wechselwirkungen für das Ziel zu bestimmen. Das Verfahren folgt einem erwarteten Ablauf von Vorverarbeitung, Qualitätskontrolle, statistischer Prüfung, Korrektur und Diskussion der Ergebnisse. Der Ansatz ist auf den spezifischen Datensatz, die Versuchsanordnung und die damit verbundenen Annahmen angepasst. Dies gilt insbesondere für die Imputation fehlender Werte, wofür mehrere Ansätze diskutiert werden und eine neue Methode, die auf einer früheren „Tail"-Imputationsmethode aufbaut, vorgeschlagen und validiert wird.

Kapitel 4 stellt ein anderer Ansatz für die Filterung von Versuchsergebnissen für PPIs vor. Es handelt sich um eine Statistik, WeSA (gewichtete Sozioaffinität), die die bisherigen Methoden zur Bewertung von PPIs aus Affinitätsproteomikdaten verbessert. Sie nutzt Techniken der Netzwerkanalyse, um das gesamte PPI-Netzwerk zu analysieren, ohne Notwendigkeit für Kontrollgruppen. WeSA wurde an Protein-Protein-Netzwerken mit unterschiedlicher Genauigkeit getestet, darunter der kuratierte IntAct-Datensatz,

die ungefilterten Datensätze in BioGRID und der große BioPlex-Datensatz. Das Modell wird auch im Vergleich zur vorherigen Methode mit dem gleichen Ziel getestet. Während sich die Funktion selbst als überlegen erweist, besteht ein weiterer großer Vorteil darin, dass sie Beobachtungen über Studien hinweg effizient kombinieren und vergleichen kann und daher verwendet werden kann, um die Ergebnisse neuer Experimente im Kontext aller bereits verfügbaren Daten zu aggregieren und zu bereinigen.

Im letzten Teil wird die Verwendung von WeSA außer Wildtyp-PPI-Netzwerke analysiert. Das Framework wird als neuartiger Weg zum effektiven Vergleich mechanistischer Unterschiede zwischen Varianten desselben Proteins (z. B. Mutante vs. Wildtyp) zu vergleichen. Ich erforsche auch den Einsatz von WeSA zur Untersuchung anderer biologischer und nicht-biologischer Netzwerke wie genomweite Assoziationsstudien (GWAS) und Gen-Phänotyp-Assoziationen, mit vielversprechenden Ergebnissen. Abschließend werden in dieser Arbeit verschiedene mathematische, statistische und rechnerische Ansätze vorgestellt und miteinander verglichen, die speziell für die Aufgabe, einen besseren Überblick über Protein-Protein-Interaktionsnetze zu erhalten, angepasst, kombiniert und/oder entwickelt wurden. Die Leistung der neuartigen Methoden wird validiert und, insbesondere WeSA, wird ausführlich getestet und analysiert, auch über den Bereich der PPI-Netzwerke hinaus.

Contents

1 Introduction

“A mathematician is a device
for turning coffee into theorems.”
— Alfréd Rényi

Mathematics has found various applications to connect and extract information from otherwise seemingly unrelated fields. It is widely used in biology supporting all directions of research in the area from public health and epidemiology to phylogeny and genetics. Established mathematical models and computational methods have aided the understanding of the increasing flood of biological data [14, 69, 22]. For instance, Fisher tests, t-tests and p-values are an integral part of many biological research papers [5, 66]. Moreover, the advances in mathematical modelling and statistical analysis have allowed research in biology to develop bolder protocols, which result in faster data acquisition, bigger outputs and larger impact [FR12, FR13, FR19]. And the interchange of information and techniques between biology and mathematics works in a bootstrapping way where the progress of one field leads to subsequent progress in the other [FR8].

It is a common occurrence for mathematics to expand in response to novelties in the fields of application and the work laid out here is a part of this simultaneous complementary progress of the two fields of biology and mathematics. It is powered by recent biological advances which require better mathematical approaches [10]. In particular, we have collaborated with experimental researchers who develop new protocols for investigating the human interactome, the network of protein-protein interactions, and based on close communication, we have developed multiple techniques to complement and analyse their experiments.

The human interactome is a central area of research as the discovery of the network of protein-protein interactions (PPIs) can elucidate the mechanism of human diseases and guide diagnostics and the development of treatments [48]. Alongside data from other species and model organisms, experiments on human cell lines are also growing in numbers. Many low-throughput studies as well as a few large-scale projects are focused on the problem of finding the comprehensive and accurate human interactome [37, 49], while resources are simultaneously directed at coordinating and systematically unifying current knowledge [52, 53, 63]. These efforts are still ongoing and to our knowledge there is no resource which maintains and analyses the current information while also allowing for dynamic update and analysis of the PPI network.

The work presented here exploits statistical, mathematical and computational techniques to gain insights into the interactome. It contrasts two main models: a statistical approach and a mathematical model and evaluates their performance in deciphering meaningful networks from noisy biological experiments.

The first approach implements a statistical workflow of hypothesis testing. Its performance is investigated depending on the missing value imputation method where we implement, improve and contrast the most popular approaches based on Singular Value Decomposition, normal distribution imputation and distribution sampling.

The complementary second method is a scoring function based on observed-to-expected ratios. It uses combinatorics calculations and the theory of the configuration model to compute edge weights. To test the method we formulate several biologically sensible hypotheses and test them through a combination of ROC analysis, statistical testing and Markov clustering.

As an expansion, the modelling approach is investigated further as it is a technique which is not specific to only PPI networks but can be useful in other seemingly different (biological) contexts. We test its usability in network analysis more generally through investigation of other gene networks related to phenotypes and disease.

1.1 Introduction to the methods and overview of the main results

The majority of the work focuses on the analysis of human protein interactions. The initial experimental research is performed in laboratory grown cell lines. For any experiment which aims to detect an interaction, normally at least the protein of interest would be modified through a process of fusion. This genetically modified protein can be introduced in cell lines through transfection, the addition of the synthetic nucleic acid sequence. The so modified cell culture is later used for experiments such as immunoprecipitation (IP) and the affinity purification (AP) experiments discussed in a large portion of the thesis.

Affinity purification and similar methods jointly referred to as AP-like are experiments aiming to determine at once the neighbourhood of a target protein called bait. Specifically, these experiments immobilise the bait and pull the proteins from the cell which somehow directly or indirectly stick to it (prey proteins).

Using a network representation, it is clear that even without additional problems like contaminants, the method does not find only the immediate neighbours (direct interactors) of the focus node (bait). Often the approach identifies entire assemblies or complexes involving the bait-protein, including proteins that are not in direct physical contact with the bait [4]. The work presented here describes two complementary methods to start from the raw experimental results which lead to a better understanding of the underlying PPI network.

1.1.1 Statistical workflow

The most common approach in AP experiment analysis is to compare the target to a control protein and use statistical testing.

During this doctoral work we collaborated within the European Training Network SCilS, with 11 institutions in which several partners did experiments on PPIs. We were able to test our statistical procedure on their data and add to the standard method by an additional comparison study of missing value (MV) imputation methods.

The pre-processing includes an imputation step where we propose a modification of normal tail imputation to complete a normal distribution. This sampling procedure employs rejection sampling in order to sample from a custom-designed distribution.

The results are inspected via Principal Component Analysis (PCA) against the performance of the imputation method based on singular value decomposition (SVD). The latter has been a popular choice in the context of AP-like experiment imputation [65, 46, 22, 67]. The PCA shows better clustering of the customised method.

After the MV imputation, the analysis proceeds normally with hypothesis testing for a difference between the distributions of the measurements of the target and control. A main drawback of the statistical analysis is the need for a reliable and unbiased control. For this reason, we propose a complementary or altogether alternative approach where a model function is used to rank interactors.

1.1.2 Scoring function of network edge weights

In the second approach, we devise a scoring function that works on the corpus of PPI data. Instead of requiring control experiments and going through the analysis of experiments separately, we use the full existing data to draw inference from the complete known network. This involves handling data obtained using heterogeneous experimental methods on large protein networks with up to 8.2 million edges.

For each potentially interacting pair the scoring function determines a weight for their connecting edge. To do so, it uses the particular nodes and the network degree distribution and compares their specificity to a random graph. A helpful strategy for constructing random graphs with a specified degree distribution, which is also exploited here, is the configuration model [51]. This model allows us to apply probability and combinatorics to study and compare the expected number of multiple edges (or edge weights).

We test the model using a combination of techniques. We first use the popular receiver operating characteristic (ROC) analysis to assess performance. We then validate the scores using published results or formulate biologically sensible hypotheses which we verify via statistical testing. The network is clustered using Markov clustering to obtain groups which are observed to overlap with the known community structures of protein complexes.

1.1.3 Extensions to the model

Several model modifications and adaptations are tested. The first one uses a modified network in which edge weights can be decimal. In the original formulation of our score, interactions are binary (observed or not) and some are possibly observed multiple times resulting in multiple edges or, equivalently, integer-weighted edges. In the first model modification further experimental evidence is incorporated in the form of rational confidence weights. These are compared via ROC analysis to the original model.

Other adaptations involve applying the model to a bipartite graph. Two cases in which bipartite networks are explored through the examples of gene-to-trait associations captured by Genome-wide association studies (GWAS) studies and gene-phenotype associations seen in mouse genetics data. GWAS studies are obtained by sequencing a sample population, recording the traits and diseases those individuals report and testing their significance. Individual studies are again subject to missing insight from what is already known, while in addition there is a stacking bias towards (interesting) discoveries.

The second bipartite network we explore is based on phenotyping research in mouse (*Mus musculus*) recorded in the Mouse Genome Database (MGI, MGD) [25]. In such studies, a mouse genome is modified to introduce an allele of interest and some or all resulting phenotypes are recorded. While the mice are a main model organism in the study of the human genome and understanding human phenotypes and disease, heterogeneity of mouse research presents a challenge to the unbiased analysis from multiple studies [13]. To our knowledge there is no study which analyses the comprehensive MGD resource as a whole.

Both bipartite networks provide essential insight into the functional role of genes and the severity of effects of genetic variation. However, the two databases are formed by individual studies and annotations which pose bias towards interesting research and results and are an obstacle to systematic analysis. However, we propose that the network structure lends itself to analysis using our scoring function and we propose two model adaptations for the bipartite setting: one scores the original edges while the second weights the one-mode projection graph through shared neighbours.

Several methods are employed to validate or disprove the models. Literature research confirms some of our obtained results and gives us insight for expected hypotheses. In addition, we use statistical analysis and testing and investigate correlations to understand the models and their relevance.

Our analysis shows that the modelling function can be applied successfully to distinguish likely connections from noise. It is especially useful when a node of interest has many connections but the focus should be drawn to just the few essential ones. In the context of the projection graph we also observe correlation with similarity scores from another established database, STRING [63]. This confirms our suspicion and opens up possibilities for further investigation.

1.2 Outline

The thesis is structured in six chapters. After the introduction, Chapter 2 presents a brief literature review outlining the main research laying the foundations for this work. The following three chapters are devoted to the main methods and results of our work.

More specifically, Chapter 3 presents our statistical framework for analysis of PPI networks. It presents the specifics of the experiments analysed both in this and the next section. Then it lays out the steps of the statistical analysis in order and discusses the challenges encountered and improvements made during the process.

Chapter 4 contrasts a new aggregating model to the previously discussed case-by-case approach of statistical analysis. In this chapter we present the scoring function WeSA that we have developed to work with big network data. We discuss several databases that we use for the analysis and testing in the chapter. The results of this section are split into two subsections: performance in terms of metrics and biological insight and results. The penultimate section is an extension to the method presented in Chapter 4. It introduces two separate further applications of WeSA in the respective sections 5.1 and 5.2. Both of them are structured in the same way: they present the data on which the extension works, they present context-specific modifications to the model function and then evaluate performance and relevance to biological research.

Chapter 6 summarises the key findings and discusses the possibilities for further research.

2 Determining biological networks

2.1 Recognising true edges in biological networks

What do we mean by biological networks? Most often we use the term to refer to physical connections between molecules. Usually these are protein-protein interactions, but also there are protein-chemical [FR26], protein-nucleotide [FR57], etc. Beyond physical, there are other indirect relations often studied in networks, such as gene-regulatory networks that (most often) relate transcription factors to the genes whose expression they affect [FR61].

However, many things (beyond molecules) can be put into networks for particular tasks. For instance, drug-targets can be linked to disease indications or side-effects (e.g. [FR6]). In the context of this thesis (Chapter 5) we wished, for example, to obtain a network of genes associated to the phenotypes observed as a result of their genetic perturbation (either in mouse experiments or genome-wide association studies).

Protein-protein interaction networks

PPI networks are based on results from experiments. The experimental protocols which search for direct physical links between proteins could loosely be divided into two categories which we refer to as pairwise experiments and AP-like experiments.

With the term pairwise study we refer to experiments between two specified proteins of interest in a controlled environment. Such studies examine only the single interaction. Among the most popular pairwise methods are split protein methods such as Yeast Two-Hybrid (Yeast Two-Hybrid) system [14] or Protein Complementation Assay (Protein Complementation Assay) [FR11]. In those methods a split signalling protein in yeast is attached to two proteins of interest which are tested for interaction. The assumption is that a signal is transmitted only if the target proteins interact. These cellular approaches experience a unique false positive driver which is the possibly frequent reconstruction of the fragments [FR11]. Alternatively, because Y2H studies are conducted in yeast, there is no way to control for interactions with other proteins. If the two tested proteins have a shared interactor within yeast, they can bind to it which can put them in sufficient proximity to output a positive (interaction) signal.

For many proteins the need to express them in a certain environment for a pairwise study is problematic. For example, in Y2H the signal readout is based on an interaction in the nucleus, but many investigated proteins are not native to the nucleus and, specifically, membrane proteins are known to be difficult to examine accurately [14]. Beyond, cellular compartment localisation, expression in different host organisms can change the behaviour and characteristics of proteins. Tyrosine signalling, for instance, is ob- served to be a characteristic of eukaryotes and despite tyrosine kinase activity being present in yeast, phosphorylation is low when compared to to human [FR9].

The issue of non-native environments extends beyond split-protein experiments to other pairwise methods such as purified protein pull-downs. A major challenge in pull-down experiments studies is optimising the experimental conditions in order to mimic the proteins functional environment [55]. For instance, salt concentration in experiment buffers should be determined carefully since a low salt concentration can disrupt protein stability while high salt concentration is more likely to break bonds between proteins [FR51]. All pairwise methods, however, have in common the inability to detect proteins supporting indirect interactions and (competitor) proteins interfering and preventing interactions.

Protein interactions in nature happen competitively, which makes AP methods better suited for unbiased probing of the interactome. Here, a target protein is attached to a special peptide (i.e. tag) which is known to bind to particular molecules called beads[1]. Because of this known and controlled connection, one is able to retrieve all proteins which interact with the target protein [14]. By using an AP method one can obtain a list of proteins that are somehow physically connected to the target protein. However, AP experiments typically contain significant noise (background). A major generator of the noise is the unknown nature of the relationships (direct physical interactions or interactions through linking proteins) which requires further verification, e.g. pairwise validation [49]. Additionally, background accumulates through problems with experimental components, e.g. through non-specific binding affinity of the beads [FR51].

The issues with noise are exacerbated as the once popular “tandem” affinity methods

[1] Beads are tag-specific.

are now being replaced by single-tag AP experiments. The former uses a two-part tag that (after capture of the target with its interactors) is separated from the bait in two distinct washing steps [FR43],[32]. As washing disrupts weak interactions, tandem-tag AP is substituted with AP experiments using a single Strep or FLAG tag and, respectively, a single washing step [10]. This allows for more of the connections formed in the precipitate to be preserved, but also leaves many more contaminants.

Furthermore, it is clear that PPI studies, which rely on material from multiple cells, suffer from inter-cellular, tissue or organism variability. Even cells grown as technical replicates in the lab are never identical and, moreover, processes in the cell are not always in sync. For example, transcription suffers from random transcription bursting, i.e. interruptions of the steady Poisson-modelled process of mRNA transcription [24], which leads to stochastic changes in the content of the cell. In addition, replication of experimental conditions is arguably a bigger factor accounting for experimental variability. This is a multifaceted challenge encompassing details, which are overlooked or impossible to define precisely, such as pressure, room temperature, temperature of reagents, researchers' accuracy, etc.

An emerging technique to counter inter-cell variability, is the development of single cell analysis, which allows for the examination of a single cell and subsequent cell-cell comparison [FR20]. Its potential drives discoveries and fast-paced progress in the field [45], but it still has its challenges both due to the lack of standard protocols for different tissues and types of experiment or due to limitations of resolution/detectability [FR58].

There are similarly technical issues related to detection limits, particularly in proteomics. Specifically, the instruments that detect peptides during mass-spectrometry identification depend on a certain abundance in order that biological signals can be distinguished from background (this is elaborated on in section 3.3). In part, the process is stochastic also due to the variable capacity of the machine to efficiently ionise the submitted sample. In practise, this means that certain proteins are simply not seen, or not seen consistently during replicates in these experiments.

Different interaction detection designs, equipment, conditions and methods significantly affect the results and have largely prevented previous research on the aggregate data. Instead, studies work on their own target and control samples to produce new PPI information (e.g. [10]). Yet, PPI analysis has seen some range of aggregate studies and those are presented in the next section on mathematical and computational methods.

Gene-phenotype networks

Study methods to characterise genome-to-phenome or genome-to-disease links rely on collecting heterogeneous data from many individuals. The GWAS approach to find significant associations between genes and phenotypes builds a network of genes and traits by expanding from the trait nodes. GWAS studies collect data on all individuals with a specific phenotype and examine the variants which are related [FR52]. Mice studies can also focus on finding relationships but they start, conversely, from genes and link them to phenotypes [FR18]. Data from GWAS and mouse phenotypic studies create a network of all links between genes and phenotypes. However, cross-organism heterogeneity in all such studies is not straightforward to control, particularly, there is conflicting evidence whether laboratory inbred mice have lower variability than outbred mice [66]. The study of Tuttle *et al.* combines literature, data on inbred mice and data on mice from the diversity outbred population [FR28] to measure variability across phenotypes. They model outbred mice variability as dependent on both genetics and environment, whereas only the latter affects their inbred mice model. They find out that the two sources of variance somewhat equate and confirm unpredictable diversity despite attempts to control the population.

Some studies aiming to improve the irreproducibility problem introduced by organism variability argue that overly controlled environments are part of the problem. Driven by the idea of blending the microbiome in different samples of mice, research suggests modifications such as changes in breeding environment, growing samples physically together. A particularly notable idea is combining results from different laboratories. In their research [FR54], Voelkl *et al.* simulate multiple treatment analysis scenarios and propose that expanding a study to as few as 2-4 labs, as opposed to just a single one, can increase effect capture.

Despite this finding that a mixture of laboratory environments can improve a study, to our knowledge, efforts have not been focused on expanding beyond a single study. In both gene- disease association studies and animal phenotyping studies, results from publications have been considered mostly on their own. Research on the whole corpus of data has been limited to overviews or clustering without attempts on filtering out noise [FR16].

Finally, none of the network exploration studies are immune to technical noise [24]. In gene-disease association studies, technical noise can occur from problems with sequencing, while in phenotyping studies, issues may be due to variation in phenotypical labelling or measurements, to name a few.

Despite the limitations discussed above, both in how interactions can be sufficiently detected and how typical sources of variability and signal background can be addressed, we have seen indications that the current range of network exploration methods can be useful. We propose that the analysis of existing networks can lead to economical changes in future study design.

2.2 The mathematics of working with networks and filtering network noise in literature

Networks, which in this thesis are taken to mean the same as graphs, are graphical representations of the relationships between constituents in a biological system. A nodes-edges tuple summarises the information for the graph. In this thesis I develop theory working with the network of PPIs, but in Chapter 5 we also see that the model we develop can be applied beyond PPI networks. In the main case of Chapter 4, a network is a tuple $(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of genes and $\mathcal{E}$ is the edge list or list of observed gene-gene pairs in interaction studies. Information from an edge list can equivalently be presented in an adjacency matrix, but lists (also called adjacency lists) are convenient when the graph (respectively, the matrix) is sparse.

In the following chapter expanding on applications we work with bipartite networks $\mathcal{G} = (\mathcal{V}_1 \cup \mathcal{V}_2, \mathcal{E})$. The graphs in those cases are naturally split into two partitions: a set of genes ($\mathcal{V}_1$) and a set of traits, diseases or phenotypes ($\mathcal{V}_2$). The two partitions are disjoint, i.e. $\mathcal{V}_1 \cap \mathcal{V}_2 = \emptyset$, but when discussing the bipartite graphs we also relate them to their one-mode projections. The one-mode projection graphs over just one vertex partition (in our case we looked at the projections onto $\mathcal{V}_1$) is the graph obtained by connecting nodes from the partition if they share a neighbour and allowing for multiple edges corresponding to multiple neighbours.

While summary statistics about the networks in the project may not be important, it is useful to know that they are normally sparse. In addition, in the case of PPI networks, we have observed scale-free characteristics of the degree distribution, meaning, existence of hub nodes or proteins with disproportionately many connections compared to the average.

There are some already established approaches aiming to elucidate the specific community structure of biological networks. Particularly, for PPI networks there are a growing number of studies and some important methods are presented below. Gene-trait networks have also attracted some interest in analysis. For both cases the majority of the approaches use some form of clustering.

Network approaches for protein complex prediction can be grouped in two main categories: purely computational and biological-information enriched [74]. The former are based solely on the raw experimental network data, while the latter also incorporate further biological knowledge.

Studies that go beyond network information include in their predictions, for instance, information on expression data, under the assumption that proteins which are expressed (i.e. active) simultaneously are more likely to interact [FR62]. As another example, Yu *et al.* [FR60] developed a regression model using the network information as well as weights of node similarity based on annotations of proteins biological and molecular function or cellular localisation. The marginal improvements and performance which still allows for much improvement[2] which result from such models, however, point to possible problems. We specifically want to underscore that protein information is heterogeneous and very incomplete, so relying on further annotations could introduce bias towards more researched proteins and their associated parts of the network.

On the other side of the spectrum are placed all clustering approaches that do not take into account any information beyond the network. The most popular of those is probably the Markov Clustering (MCL) approach [69]. It is a divisive approach starting from the whole network and dividing it into clusters by exploiting the concept of flow through a network which is taught in courses on Interacting Particle Systems. In particular, Van Dongen introduces the concepts of expansion (squaring) and contraction (column-wise re-scaling of the powers) of a flow matrix and argues that the two operations iteratively enhance the difference between within-cluster and inter-cluster edge weights. The process converges for adjacency matrices with sparse bounds and results in an idempotent matrix (i.e. matrix which is invariant to squaring). In the obtained matrix representation, clusters are built around ‘attractor’ nodes. Attractors are all nodes which, according to the output matrix, have positive flow to themselves, while nodes which lead to attractors are ‘attracted’. Attractors may also be attracted to other attractors in which case they co-exist in a cluster.

The algorithm has been benchmarked in the context of protein complex discovery in

[2] F-measure scores improve in the 0.5-0.6 range in [FR60] and up to 0.6 in [FR62].

comparison to other algorithms. Its performance can be slightly inferior to information-enriched methods [FR60, FR62]. However, it is comparable or better to other purely computational approaches such as the greedy algorithm, PC2P, [FR35] and the agglomerative Core&Peel method [FR38]. It is an important note here that the models are largely similar and presenting one or another as superior is dependant on the measure chosen for testing. For instance, MCL is superior to Core&Peel when modularity is computed, but lags behind in F-measure. Yet, all network clustering approaches tested by [FR35] are observed to have recall below 0.72 which is attributed to sparse complexes. While the studies are clustering the yeast interactome, this can still present an interesting comparison to the performance of our own method in Chapter 4.

There has also been some clustering analysis performed on gene-trait association networks from GWAS. These are focused on only few traits at a time as opposed to looking at the whole network. For instance, research on sex hormones SHBG and testosterone has developed individual linear mixed models for the prediction of important genetic drivers for the respective hormone levels [FR47]. As an additional perspective in case of overlapping features, they have used normalised association numbers to produce hierarchical clustering and identify the genetic variants contributing to effects on sex hormones. In a different study focused on obesity, Grant *et al.* [FR16] aim to identify differences in mechanism on a more-detailed level. It is another representative of the approach to GWAS studies, namely, focusing on a trait of interest and tailoring the analysis according to it. Here, the focus is on body-mass index (BMI) and only nine additional traits picked for their proven association to BMI were selected to use in further classification and clustering. In general, GWAS analysis has seen some clustering approaches and modelling applied to gain insights, but that has been supplemented by manual curation of the input and has been limited to looking at few traits at a time.

So far the methods presented have been at two opposite ends: either strictly focused on the clustering algorithm from a computational perspective or heavily supplemented by biological information. The approach we present below does not incorporate any supplementary biological information, which makes it general. However, it differs from purely algorithmic clustering approaches as it uses more of the information hidden in the network. This approach has been proposed first by Gavin *et al.* [30].

Instead of working with an unweighted network of binary observations, the study of Gavin *et al.* on the yeast interactome proposes the socio-affinity (SA) weight, a score weighting observations based on a comparison to their expectation. The weighted graph is then clustered using three different methods: single linkage, complete linkage and unweighted pair group method with arithmetic mean. The paper introduces two interesting ideas incorporated in the weight of each edge. The first one is to use the observed-to-expected ratio which scales each interaction relative to the observations of the proteins in the pair. That can down-weight links of abundant proteins and highlight those that are rare but specific. The second important point introduced by the SA score is that it is calculated symmetrically for both proteins and includes a third component for any matrix information. With the term matrix (information) we refer to every observation which is not direct; in a networks setting, that translates to pairs which share a neighbour. With these two points made, the SA formula is:

$$SA(i,j) = S_{i,j} + S_{j,i} + M_{i,j}$$

where the two symmetric terms $S_{i,j}$ denote the log-ratio of observed-to-expected edges between i and j and $M_{i,j}$ is the log-ratio for indirect matrix terms. The score clearly uses no additional information apart from the network structure, but is able to effectively scale the importance of each observation. The method has not been tested tested against representative reference sets, but was observed to retrieve accurate biological structures.

The SA method has two close adaptations published by Collins *et al.* [20] and Schelhorn *et al.* [60] again for the *Saccharomyces cerevisiae* interactome. The former integrates Bayesian setup (testing) within the SA framework to incorporate also negative interactions. Negatives should always be treated with caution given that some experiments may be inaccurate in detecting interactions. The authors propose a similar three-term structure of the score, but instead of working with observed-to-expected ratio, they look at probability ratios of the form:

$$\frac{\mathbb{P}(ij \text{ is observed}|ij \text{ is true})}{\mathbb{P}(ij \text{ is observed}|ij \text{ is false})}$$

In their method Collins *et al.* propose that the probability of a true association to be preserved and detected, is estimated individually for each input dataset. Specifically, their paper integrates three studies and calculates this probability through the observed frequency of successful purification over a very high confidence set of interactions[3]. The second main modification is the use of probability of negatives, the probability that a given bait-prey pair would be observed for nonspecific reasons. It is calculated using

[3] For Krogan *et al.* [45], Gavin *et al.* [30], and Ho *et al.* [FR21] data, this gave values of 0.51, 0.62, and 0.265, respectively.

a Poisson process (i.e. exponential distribution). The main advantage of the method is that it manages to down-weight negatives. Despite the other study-specific tailoring, their results have varying accuracy across datasets and their attempt to combine datasets leads to approximately 50% coverage of the gold-standard reference set.

The second adaptation of the score, named ISA and developed by Schelhorn *et al.* [60], retracts the matrix term as allegedly obscuring true physical interactions[4]. We argue, later, that matrix term does improve predictions. ISA also modifies the two remaining terms to use a probability instead of the observed-to-expected ratio. That is, each term is a logarithm of an expression of the form $\mathbb{P}(S_{ij} \geq s_{ij})^{-1}$, where S and s denote the random variable and its sample (observed) value, respectively; the variable represents the count of evidence for interaction between a pair of proteins. This probability modification is done in order to reduce the 'diminishing returns' behaviour of SA as evidence accumulates[5]. In our model, presented in section 4, similar effect is achieved instead by the term weights. Testing is performed comparing ISA scores to several other methods including the aforementioned SA and PE scores, while reference sets include pairwise interactions and confirmed 3D structures. ISA and SA are observed to outperform the other methods in those tests, but the authors do not find any notable superiority in performance of ISA compared to SA.

Finally, the advancement in technology and availability of resources brings us closer to the possibility of building an accurate interactome by probing every interaction. While still far from the goal, two notable studies presenting a snapshot of the current progress are presented below. The first direction represented by the papers of Rolland *et al.* [58] and Luck *et al.* [49] is experimental. As a recent alternative, Evans *et al.* [26] and Burke *et al.* [FR5] set the foundations for computational probing of the PPI network.

The experimental papers [58, 49] provide an updated map of the human interactome as part of The Human Reference Protein Interactome Mapping Project[6] (HuRI). They analyse and test their pairwise-experimentally validated interaction set extensively and further provide correlations to GO terms[7] and diseases, among others. An attempt is made to probe pairs uniformly and decrease popularity or other biases, via analysing also the number of publications attached to the proteins. Interestingly, Rolland *et al.*

[4] Not to be confused with appearance in the same complex which may be re-framed as indirect interactions.

[5] In other words, as more evidence is added, the rate of increase is SA score slows down.

[6] Description of the general project: http://www.interactome-atlas.org/about/

[7] GO terms are annotations of proteins in the Gene Ontology (GO) database in three categories: molecular function, biological process and cellular component.

compare the recovery rates of pairs detected in one pairwise study only to a random control and to those observed in at least two studies. They detect only a slight difference in recovery to random controls and a huge drop compared to rates of pairs with multiple supporting evidence. While the conclusion is drawn from evidence including at least one pairwise experiment for each pair, this result is important as we think about the reliability of AP-like experiments as well. Ahead for the HuRI project, there is still a lot of ground to be covered if the full human interactome is to be covered; efforts are ongoing and as of 5th May 2023 they have found 64,006 interactions among 9,094 proteins.

All of the above experimental and computational approaches are aimed at revealing *whether* particular biological entities interact or are in a more indirect relationship. However, they do not, normally, reveal any molecular details about *how* the interactions occur. Certain key developments in the last few years, which built on many decades of structural biology, provide the means to identify or predict such molecular details.

In particular, there are now extremely accurate computational approaches to predict protein structures. For instance, AlphaFold2 (AF) [FR23] is a protein structure prediction model developed by DeepMind, which disrupted the structural research landscape. It participated in the 2021 edition of the biennial challenge on protein structure prediction CASP14 in which teams of modellers compete to make the most accurate predictions of yet-unpublished protein structures.

AF uses neural networks but is able to communicate between two otherwise separate information sources in the form of matrices: sequence alignments and positional information. The transmitted updates are used to further modify the two matrices. Via transformers which refocus attention on important rows of those matrices fine-tuning is improved. Ultimately, this has led to their model having 0.96 Å root-mean-square deviation at 95% residue coverage (r.m.s.d.$_{95}$) of the backbone. In comparison, the second best performer had an average backbone prediction accuracy of two carbon atoms or 2.8 Å.

The AF method spread quickly and currently UniProt displays within its pages AF structure predictions for all human proteins. The algorithm was also extended to be able to predict not only single structures but complexes [26]. This already gave rise to faster interaction determination (presented in Chapter 4) and large-scale efforts to probe the interactome [FR5].

3 Statistical workflow

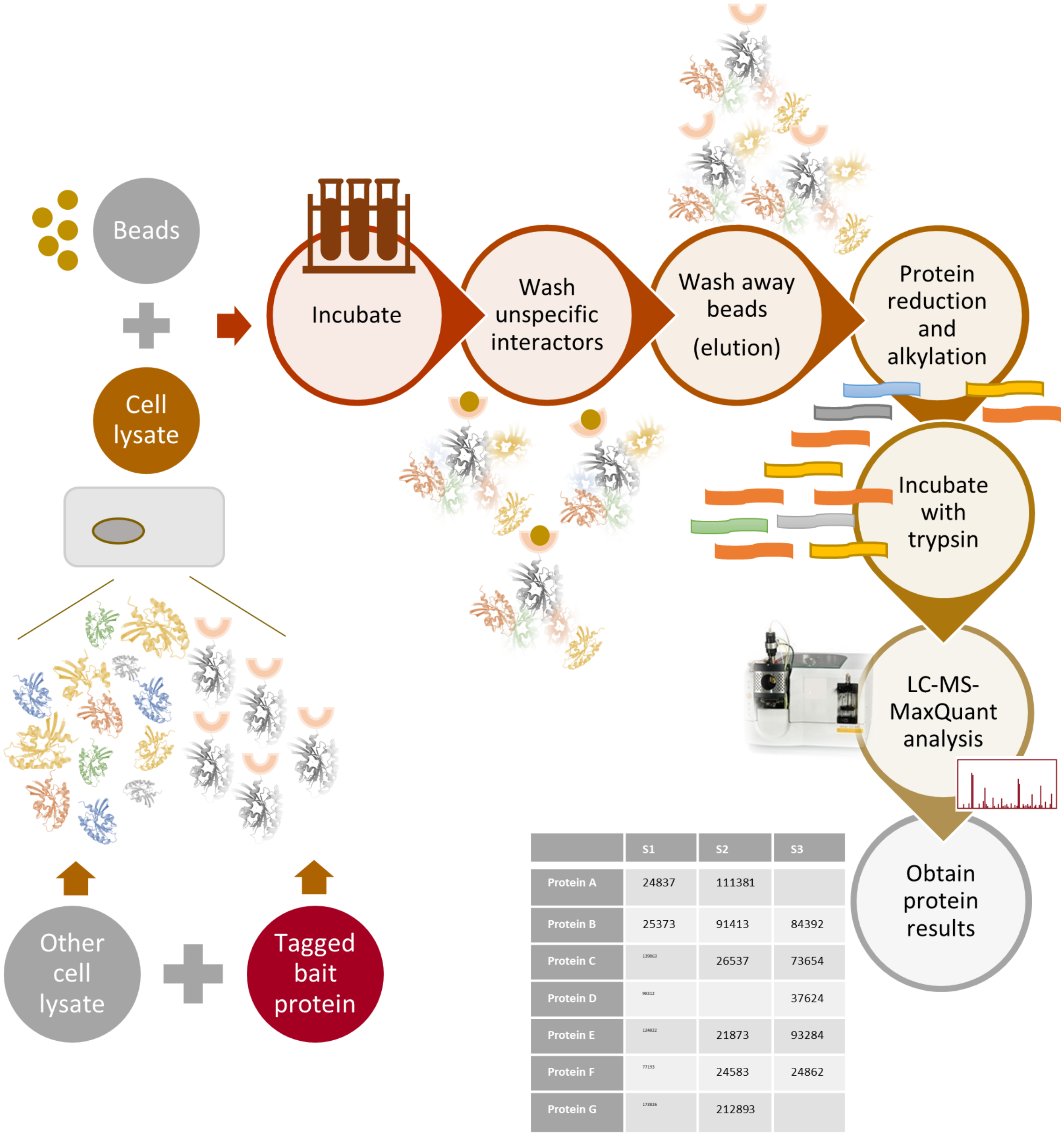
Proteomics methods aim to identify weaker connections between proteins. Some experimental approaches mentioned in the previous chapter are just a few of many examples [FR46, FR3, FR1, FR2, FR14, FR22] which showcase changes in modern protocols to accommodate for the search of weak connections. There is thus a growing challenge to try and distinguish true interactions from noise when signals are particularly weak. In this chapter, are discussed the challenges in the follow-up analysis of the single-tag AP experiments which were mentioned. As an example to supplement the theoretical procedure, we use an AP experiment in which the single tag attached to the bait is Strep (Strep-AP).

3.1 Data

3.1.1 Sketch Protocol: Affinity purification coupled with Mass-Spectrometry (AP-MS)

Most data which we have applied the method from this chapter on is yet unpublished Strep-AP-MS data from our collaborators in Tübingen. The focus in this chapter is on the general principles we have observed when working with this type of data and alongside those, to illustrate the procedure we present the details of one example. This example is an experiment I have conducted with help from Shibu Antony in the Ueffing Group during my visit in the Institute for Ophthalmic Research, Tübingen. Specifically, we have followed the protocol below to produce Strep-AP samples for RAB7 wild type and RAF1 at two time points aiming to investigate the mechanistic change driving cilium disassembly.

Running such an experiment requires standard buffers, beads functionalised to bind the tag and cell lysate from HEK293T cells modified to express the target protein together



	S1	S2	S3
Protein A	24837	111381	
Protein B	25373	91413	84392
Protein C	139863	26537	73654
Protein D	98312		37624
Protein E	124822	21873	93284
Protein F	77193	24583	24862
Protein G	173826	212893	

Figure 3.1: Steps of affinity purification experiment after having the input components: cell lysate (including tagged protein) and binding beads. The schema shows the detailed process up to the submission for mass-spectrometry.

with the Strep tag[1]. Then the process goes through 5 main steps before the result is obtained (figure 3.1).

To get cell lysate, pre-grown cells (14 cm dishes) were scraped in the presence of 1

[1] I used pre-prepared cells by Shibu Antony, who did modifications to the cells (transfection) by adding a PEI-DNA mixture to the medium of the cells and, then, allowing cells to grow for at least 48 hours.

ml LB (Merck; 492016-500ML), incubated (40 minutes, $4°$C, 35rpm), centrifuged (10 minutes, $4°$C, 10000 g) and supernatant was collected. Beads (Serva, 37283.01) were washed with TBS (45 ml Millipore and 5 ml 10x TBS AppliChem, A1086,1000), LB and twice with WB (Merck, 492016-500ML). Beads concentration in the resulting solution was measured by a Bradford sample.

In the first step, beads and lysate are mixed together and incubated for sufficient time at proper temperature (here, $4°$C for up to 2 hours on a 35rpm shaker). This step makes sure that the target protein binds well to its interactors and its tag sticks to the beads.

The sample is then washed three times. The washing procedure comprises a centrifuge-discard supernatant loop as after centrifugation the beads, with the proteins bound to them, precipitate to the bottom of the test tube. All remaining supernatant is unbound protein which ought to be discarded, although contaminants can remain.

During the elution step which follows the beads are separated from the tag and removed from the samples using a Strep-tag elution buffer (IBA, 2-1000-025; concentration: 1:10). The only remaining components in the sample after this step are the proteins we are interested in: the protein network in which the target protein is involved.

In the protein reduction and alkylation and trypsin-incubation steps the proteins are separated and denatured. This is done, by adding a mix of ABC(30 μl,50 mM; Sigma, A6141-25G)/RapiGest(4 μl; Waters, 186001861)/DTT(1 μl; Merck, 1.11474.0025) (final DTT concentration: 2.7 mM) and incubating (10 min., 60 °C, 500 rpm, Thermoblock Thermo-Shaker Incubator MT-100 manufactured by Universal Labortechnik GmbH KO.KG) and then, by adding the proteinase trypsin (Serva, 37283.01) in a concentration up to 4.7% and leaving for 2 hours at $37°$C. As an end result, the proteins are broken down to short peptides which the machine can detect.

The sample was then submitted for label-free mass-spectrometry quantification which was performed as in [34]. The preparation, mass spectrometry and computational reconstruction[2], which is involved, goes through the steps of separation of peptides by liquid chromatography, ionisation, measurement of mass-to-charge ratios of the ions (MS1[3]), fragmentation of ions and detection of fragments (MS2), identification of peptides based on MS2 by comparison to theoretical spectra, mapping of these peptides

[2] Computational mapping to a reference proteome database based on number of peptides and protein coverage to distinguish the proteins in the analysed sample is done using MaxQuant software [FR10].

[3] MS1 and MS2 are the specific mass-spectrometer spectra characterisation steps which are then used to computationally identify the proteins.

to proteins and thereby identifying and quantifying them [1].

The MS analysis is reliant on the success of each of those steps: the peptide separation and ionisation, the sensitivity of MS1 and MS2, the completeness and specificity of the comparison database. Stochastic errors on every step accumulate and can produce some errors or omissions which will be discussed in the next sections.

The output of the MS analysis is a table with individual label-free quantification readings (LFQ values) of all proteins detected in the samples. Usually, the LFQ values are log-transformed in order to achieve normality of the overall log-LFQ distribution. Then the samples are ready for further (statistical) analysis.

In the rest of this work, when I refer to Affinity Purification, I refer to this specific procedure.

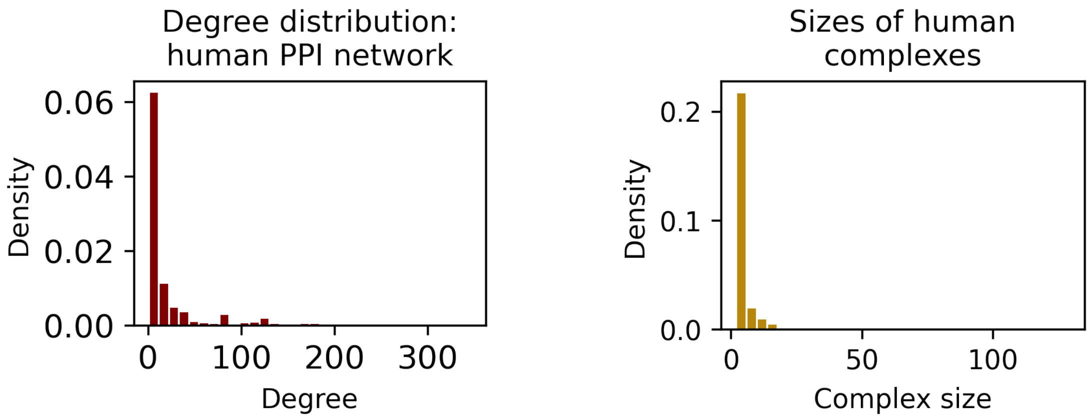
3.1.2 Description of the data

There can be thousands of proteins identified in an AP experiment with the average number from the 21 experiments we have analysed being 3,644. In contrast, our estimate based on the large CORUM dataset, which contains a curated set of protein complexes [31], is that a protein can directly or indirectly (as part of a complex) be involved in 21 interactions on average[4]. This estimation is based on the network created from the data on human proteins on CORUM by connecting proteins recorded to be in the same complex. The degree distribution for all 4,427 proteins is presented in figure 3.2a. Past literature combining databases seems to validate this number somewhat. One study aggregating validated interactions estimate 32 interactions on average [2], while an older one extends the experimental knowledge for PPIs by inference and still describes 38 neighbours per protein on average [72].

Furthermore, research so far has agreed on protein complexes comprising from 2 to over a hundred proteins. The two largest recorded human complexes on CORUM, the Spliceosome, E and A complexes, respectively contain 129 and 113 protein subunits. However, the complex size distribution is heavily skewed with median and mean 3 and 4.13 (figure 3.2b).

It is possible that a protein participates in multiple complexes which are captured all at once. However, we hope that at the precise time point of experiment the target

[4] Median: 6, 75% quantile: 16, max: 346



(a) Degree distribution density for the protein-protein ‘interaction‘ network as built from CORUM records on complexes. The network of interactions, direct or indirect within complexes, involve 4,427 human proteins in total.

(b) Density histogram of individual complex sizes for all complexes comprising human proteins. In total, 3,538 complexes were identified for the distribution.

Figure 3.2

protein is only (or at least mostly) involved in a predetermined process. Thus, we hope spontaneous binding in other conformations is reduced or almost eliminated.

In our RAB7 sample there were 985 separate entries (proteins) identified. Of those, 13% (or 130) could not be identified uniquely, but were matched to 2 or more polypeptides. This still leaves us with 855 distinct proteins, which is drastically higher than the average protein complex size.

It is clear that the method captures more than a single protein complex. It is possible that the protein in focus participates in multiple complexes and that some are simultaneously occurring. Moreover, in the cell, some complexes are situationally impossible to occur (e.g. a ciliary and a nucleus protein interacting), whereas in *in vitro* experiments such as these ones unfeasible connections are just as easy to form as all others as long as the proteins are a match in more basic factors such as shape and composition. A third possibility is that the true aim of the experiment is achieved successfully and weak connections between protein complexes hold the big structures together and allow them to co-purify. Thus, the nodes (proteins) in the connected components are observed more thoroughly, but the edges are unclear. Finally, due to fewer cleaning steps, there are possibly more contaminants still included as well as proteins which bind non-specifically. Indeed, the last is very likely given that a stated goal of modifications to affinity proteomics over the last decade has been simply to provide *more* proteins simply by removing one of the washing steps during purification.

Due to these issues and the desire to present a clean network of the proteome, it is clear that rigorous analysis is needed.

3.2 Experiment goal, assumption and controls

RAB7 is described as a GTPase regulating endosomal trafficking and has been found in several organelle membranes. Its best researched location is the late endosome membrane which is notably found in microtubule sorting centres.

In this experiment, RAB7 is investigated specifically because of its recently discovered role in cilia disassembly. Wang *et al.* describe that knockdown of the protein has been observed to lead to cilia elongation which is rescued by re-supplying the cell with RAB7 [70]. Furthermore, in the RAB7-knockdown cells, cilia disassembly is blocked in the presence of disassembly-inducing serum which is confirmed to be reversible through supplementation of active RAB7 Q67L mutant. No other previous research has positioned RAB7 within the cilium or having any ciliary involvement.

This RAB7 AP experiment thus aims to find the interactions which RAB7 forms in the cilium throughout the cilia disassembly phase. APs were conducted in two time points: normal conditions (labelled “0 hours”) and serum induced disassembly (“2 hours”).

In order to do statistical analysis, further experiments with a specifically chosen control are done. The control is chosen in such a way, so that background noise is similar, but specific interactions do not overlap. The statistical analysis then assumes that the two sets of proteins – the set of proteins in the network surrounding the target protein and that for the control protein – have an intersection which is an empty set.

The control is unique to the bait and aim of the experiment. For instance, experiments which aim to identify interactions in the cilium would normally use as control a protein with no ciliary function. Some experiments use as control an empty tag, which would identify the non-specific binding of the tag and beads.

In this case, since we would like to identify the role of RAB7 in cilia upon disassembly serum supplementation, the control protein should not be involved in any ciliary interactions. It needs, however, to be present in all other regions where the background noise of the experiment may come from.

RAF1 has been used in many ciliary studies as control. On the one hand, it is present in most of the cell; according to UniProt, it localises to the cytosol, nucleus, cell membrane and mitochondria. On the other hand, it had been claimed that the protein is well-researched and no ciliary function had been found making the prospect of the existence of such function unlikely. These reasons made it a standard control in cilia studies [27, 34, 10].

3.3 Statistics: Data preprocessing

Normally, experiments are done in triplicates and some may even do more than three replicates. Our RAB7 and RAF1 proteins data has 6 replicates for each protein and each time point, 3 done by me and 3 done by Shibu Antony. We have used two time points: 0 hours and 2 hours, as explained in section 3.2.

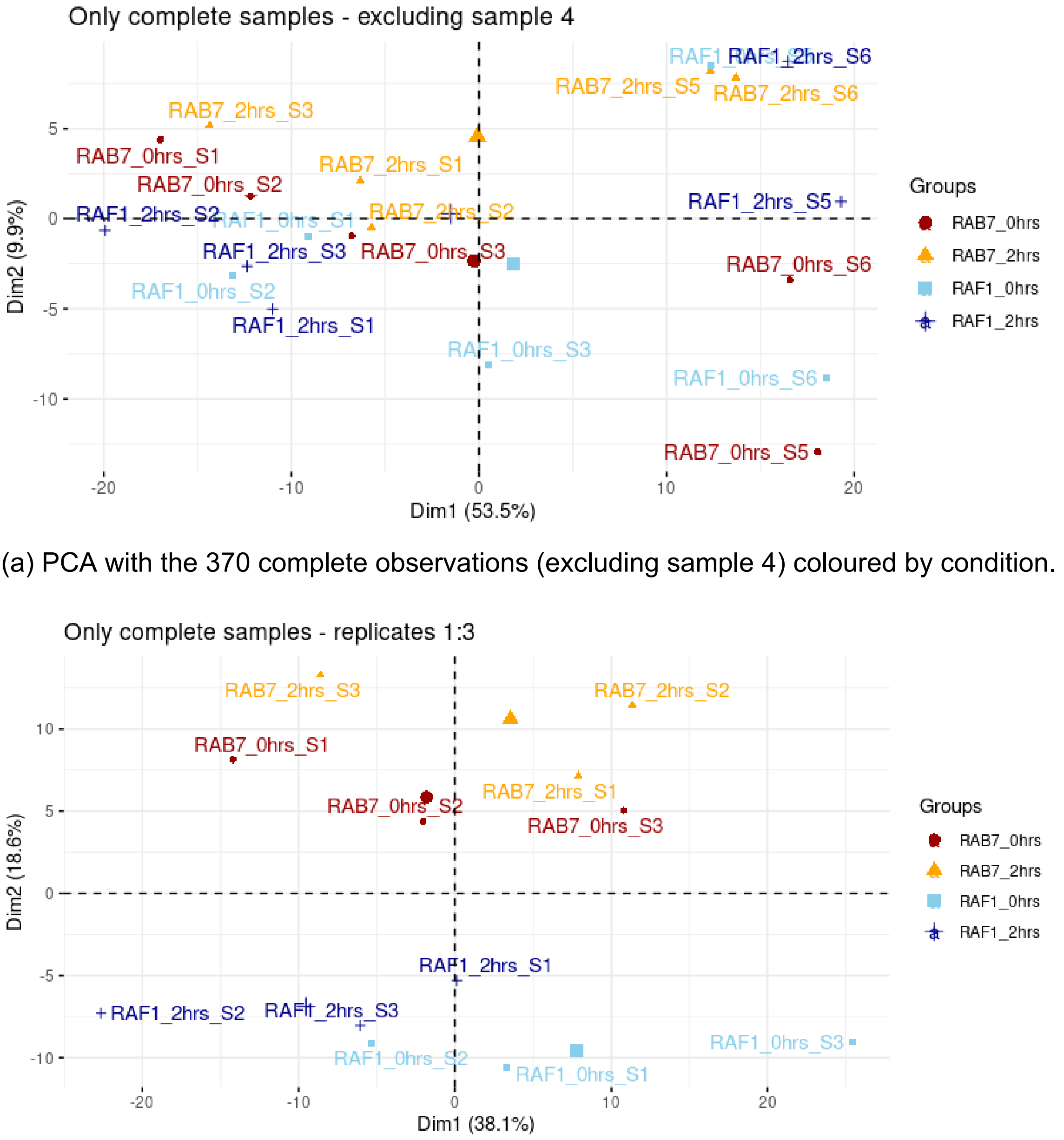
For each time point and sample (columns in table representation) and proposed protein interactor (rows), the results from the complete procedure described in 3.1.1 is a specific LFQ value reflecting the intensity of detection of that proposed protein interactor. If a protein is not found in a replicate, a missing value is inserted[5]. We use this table of LFQ values in the statistical analysis presented here to reduce background. As mentioned previously, those values are log-transformed.

First, we checked if the variability accumulated from the experimental procedure between samples was not preventing cumulative analysis to be performed on the grouped samples. As a quality control to establish that the data are correct, we verified the groups through principal component analysis of the complete observations. If the data are acceptable, the 4 different conditions (two time points 0 and 2 hours for two proteins RAB7 and RAF1) should satisfy two things:

- the 2 conditions performed with RAB7 should separate from each other and from the control, RAF1, in the PCA plot;
- the samples within a condition should cluster together.

Due to a great number of missing values it is impossible to make a PCA plot of the complete observations from the raw data (only 5 rows are complete across samples),

[5] Therefore, the final table includes a separate row for every protein which is found in any of the replicates and n (respectively, 3 for RAB7 and Strep-binding beads) columns in which the LFQ values corresponding to the n replicates are recorded.



(a) PCA with the 370 complete observations (excluding sample 4) coloured by condition.

(b) PCA with the 435 complete observations for samples 1, 2 and 3 only, coloured by condition.

Figure 3.3

so we investigate the data further in order to find a way to increase the input data for PCA. For this dataset, we observed a sharp contrast between replicates 1-3 and 4-6. The first three are much more complete. In particular, ignoring samples 4-6 results in 435 complete observations across conditions which can be used for PCA. Under even closer investigation, the culprit seems to be sample 4, which compared to the rest, contains too many MVs. Excluding that single sample results in 370 complete observations with which we performed the initial PCA with results shown in figure 3.3a.

While much of the variance is still unexplained, the lack of clear clustering of the conditions suggests an inherent problem with the data. In fact, PC1 differentiates between samples 1-3 and samples 5-6.

This aligns with the experimental design in which I have performed the first 3 replicates with beads binding to Strep tag, while my colleague Shibu Antony performed the experiments in samples 4-6 with beads binding to FLAG tag. Thus, I iterated the quality control step, but this time only on samples 1-3. This time the conditions are somewhat separate (figure 3.3b), so only the first three samples are used to demonstrate the procedure for analysis further. This new dataset (with only samples 1-3 per condition) will be referred to as RAB7-Strep or just RAB7.

Characterisation of missing values

There is no information on the exact quantity of missing values in AP-MS proteomics, but they are a well-recognised challenge due to their high frequency of occurrence in any dataset. Literature works on other proteomics datasets give us some guidance on the fractions of observations resulting in missing values. An overview of microarray experiments has found that missing values can reach 10% and additionally, studies with several experimental conditions can have omissions in almost all genes with MV for genes reaching close to 95% for the 122 conditions in Tibshirani *et al.* [64]. While microarray data can serve as a lower bound it is believed that the AP-MS experiments have many more missing values. Another review by Albrecht *et al.* on gel proteomics [3], suggest electrophoresis and staining methods can have up to 50% missing values, while, similar to the case with microarray experiments, up to 90% of the represented genes can have at least one missing observation.

In the raw datasets we have worked with throughout my PhD studies, the work of scientists in the Ueffing lab in Tübingen, the mean percentage of missing values in baseline wild-type (WT) experiments is 35%. They have produced 21 datasets using 16 distinct bait proteins and an AP protocol. Every experiment has between 3 and 6 technical replicates and up to 29 conditions. Perhaps due to the high number of conditions in some datasets there are also an average of 48% of prey proteins which are missing at least one LFQ value (range 31-75%).

The RAB7 dataset is cleaner in comparison and it contains fewer missing values than others. Out of the 855 prey-proteins, 567 have complete observation numbers at the 0 time point, the others have at least one missing value. The total proportion of missing values is almost 21% for the same time point and 175 proteins (20%) have missing values for at least 2 out of the 3 replicates.

It has so far become clear, that missing values across replicates are a main problem for analysis. Moreover, they can be so many that the simple procedure of dropping observations with missing information is not appropriate. The solutions would be to either understand the experiments better and decrease the proportion of missing values or use an appropriate way to impute the missing values. First, I discuss the former by presenting the multitude of reasons creating missing values which are also impossible to avoid in the MS-related steps.

Missing values can be due to various causes which previous studies separate into two main classes: missing completely at random (MCAR) or missing not at random (MNAR) ([46, 42]).

MCAR values arise from random variability and errors throughout the whole process. For instance, insufficient fragmentation of the protein peptides or an error reading a peptide would lead to an error in quantification and subsequent detection. These errors are stochastic and universal to any kind of data, meaning, the methods developed to deal with them normally apply regardless of the specifics of the data.

On the other hand, MNAR values depend on the experimental design. Their imputation will affect differently the results depending on the imputation method chosen. Specifically, when choosing imputation method which is tailored to MNAR values, one should consider the particular reasons for their occurrence.

Not random causes for missing values are usually related to protein abundances. Every MS machine has a threshold limiting its ability to detect rare peptides and so peptides and, by extension, the proteins containing them, which have low counts close to or below that threshold will inevitably contain missing values. It is important to note that this detection limit is fuzzy as per the definition of Burgin of fuzzy limits [16]. In the context of the MS experiments this means:

Constraint 1: For any mass-spectrometer, there exists a detection limit L which is a fuzzy limit.

This means that every experiment has its own detection limit a_n and for the sequence $\{a_n\}_{n=1}^{\infty}$ it is true that it fuzzy converges to the detection limit L. In more detail, for any candidate limit a, we can define its upper defect as $\delta(a) = \text{inf}\{r : \exists N > 0, s.t. |a - a_n| \leq$

$r + \epsilon, \forall n \geq N, \forall \epsilon \in R^+\}$. With this definition, a is a fuzzy limit when its upper defect is bounded, i.e. $\delta(a) < \infty$.

It is clear that fuzzy limits are not always unique, but rather usually comprise a set. For a MS there is a lower bound which can theoretically be obtained as the minimum number of ions which can be detected and measured properly (by design of the machine) [FR34]. However, not all peptides are properly ionised and, in addition, there is also plenty of noise to obscure the weak signals. Due to these reasons, we do not know the exact machine limitation, but even if we did, there is no certainty that abundance slightly above the threshold will be detected.

While peptide abundance is the major cause to which MNAR values are attributed, it is worth mentioning that undetected proteins may also be due to peptides not identifying proteins uniquely. Moreover, a third cause for decreased recognition rate is the incompleteness of the database against which proteins are searched. Every lab can choose what reference database to search proteins against, but even after efforts to achieve the best possible coverage, there is no way to identify a protein uniquely if it is not known.

Having presented the types of missing values, it is clear that they cannot significantly be reduced by means of improving the conduct of the experiment. Therefore, one needs to apply some statistical means in order to handle missing values in the raw data.

Handling missing values

Normally, the preprocessing steps related to missing values that we do are two: removing the outliers (bad observations) and then imputation. By the former, we mean that results which are confirmed in less than half of the replicates are normally deemed invalid. Specifically, in the case of proteomics, proteins which are captured in less than half of the samples are not considered “observed” in later steps of the analysis.

Then follows the imputation step for which we have the aforementioned constraints. There are methods targeting effective imputation when MNAR values are present, or general techniques aimed at MCAR values. There are a range of studies comparing the performance of differently tailored approaches [46, 71, 41] and the researchers agree that, depending on the specific situation and, especially, the distribution of missing values across the two main categories, there is no single method which outperforms the other ones.

Next, I present two foundational methods for missing value imputation based on the two types of missing values we choose to target (MCAR or MNAR). The strategy for picking one direction or the other should be built after considering what we expect the missing values to be.

MNAR-tailored missing value imputation methods assume that missing values mainly arise from the inherent left-truncation (or relative sparsity) of the data because of the machine limitation. Thus, simple methods would assign to the missing values 0 or another constant corresponding to the suggested detection limit, e.g. the minimum of the detected frequencies. This idea can be built-upon further by substituting on a case-by-case basis by taking into account also the gene-specific minimum [19]. Another approach is to define a normal distribution on the left tail of the distribution and pick from it [36].

Two of the best received imputation methods among the MNAR ones, in our experience, are the zero imputation and the normal imputation, so we elaborate on how they work before testing them on the data.

The zero imputation is, as mentioned above, not more than replacing all MVs by zero. This creates an artificial mode at 0 and a 'cliff' between the minimal observed and the imputed values. This design can be very unrealistic, while, in addition, it is also problematic for the testing afterwards if testing takes into account the mean.

The normal imputation chooses a normal distribution corresponding to low-LFQ values to impute from. That is, a normal distribution for sampling is defined by choosing a mean some distance to the left of the original mean and an appropriate variance. For example, in their Perseus software, Tyanova *et al.* [67] suggest default parameters μ_1, σ_1 to be defined as $mu_1 = \mu - 1.8\sigma$ and $\sigma_1 = 0.3\sigma$ where μ and σ are the mean and standard deviation of the original distribution. A main weakness of this parameter estimation is the fact that the sample mean would normally be shifted to the right. That is, if we assume that the MVs are missing from the left tail, not having them will shift the weight towards the values present (i.e. higher). Another discussion point if whether it is reasonable to assume that the values are normally distributed in the so-defined region on the lower tail.

Below, the RAB7 dataset is used to implement a MV imputation method for MNAR values. The method is based on the same assumption as the normal imputation. That is, the method assumes that the missing values occur for low intensities. However, we have addressed the issues mentioned about the normal distribution by defining a data-customised density and taking into account the mean bias. The general assumption for the method we develop and describe below is that there are no missing values from the upper half of the distribution.

The procedure is outlined next and is supplemented by figure 3.4 which implements it on the RAB7 data.

1. From the data: determine the median, μ (including the MVs set below the minimum value). This is the dashed blue line in the figure.
   Here, it is important to note that the missing values need to be less than half of the data. The condition is always satisfied if we have removed observations with at least 50% MVs in the replicates before starting the imputation.

2. Reflect the upper 50% of data through the median and determine the standard deviation, σ, of the new dataset.

3. Create the fitted normal distribution with mean μ and variance σ^2 (green line in figure 3.4)

4. Define the difference as:

   $$\delta(x) = \begin{cases} \big(f_N(x) - f(x)\big)\mathbb{1}_{x<\mu} \text{ when } f(x) < f_N(x) \\ 0 \text{ otherwise} \end{cases}$$

   where $f(x)$ is the density of the data (excluding missing values) as shown in by the red line in figure 3.4 and $f_N(x)$ is the density of a normal distribution $\mathcal{N}(\mu, \sigma^2)$. The dotted line represents $\delta(x)$.

5. Define, Ω, the domain of $f(x)$ as the region between inf$\{f(x) > 0\}$ and μ. Find the area under the curve

   $$a = \int_{x\in\Omega} \delta(x)\, dx$$

   .

6. Construct a probability density function (pdf), $p(x)$, by rescaling $\delta(x)$ to satisfy the pdf property that $\int_{x\in\Omega} p(x)\, dx = 1$. That is, define

   $$p(x) = \frac{\delta(x)}{a}$$

   In the figure, $p(x)$ is shown as a thick black dashed line.

7. Sample the MVs from the distribution with pdf $p(x)$. Since the distribution is not defined by any parameters it is not possible to sample directly from it. Instead, we have used rejection sampling the details of which are also given below.

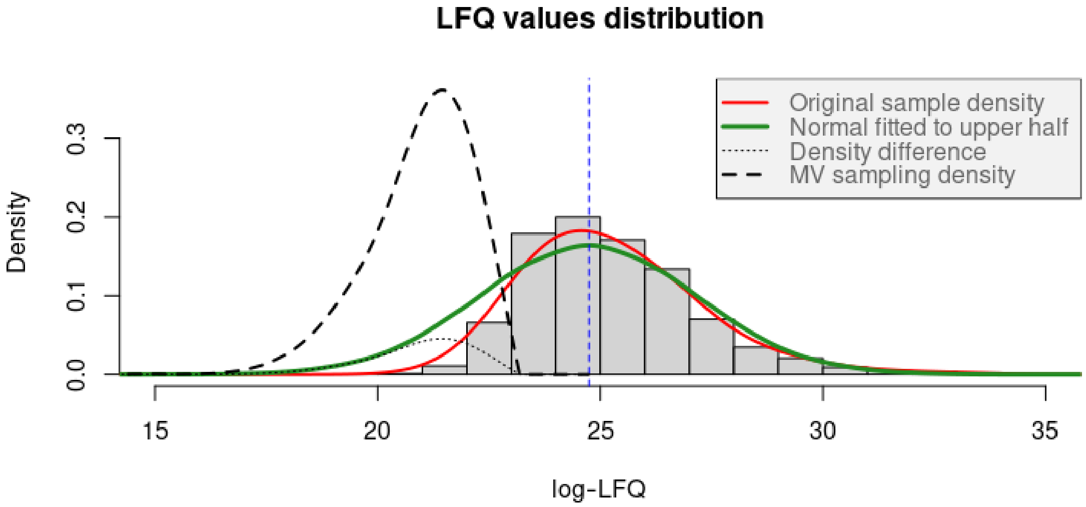


Figure 3.4: Histogram of the log-LFQ intensity values. On top of it is overlaid the density of the data shown (red line). The median of this data is presented as a blue dashed line. The green line depicts a density of a normal distribution with mean which is the same as the median of the red distribution; its standard deviation is estimated from the upper 50% of the data. The dotted line presents the absolute difference between the two densities when the normal distribution density is larger than the density of the data; this is only on the right side of the data median and is going to be used for sampling the missing values. This difference curve is converted into a probability density function (for sampling MVs from) through division by the area under the curve.

Rejection sampling It can simulate a random variable X with a given probability density function $p(x)$ over its support Ω. If $\exists M \in \mathbb{R}$ and a p.d.f. $q(x)$ defined on Ω such that $M \geq \frac{p(x)}{q(x)}, \forall x \in \Omega$, then the rejection sampling algorithm returns a sample $x \sim p(x)$.

The algorithm itself goes through two steps:

Step 1. Sample $y \sim q(x)$ and $u \sim U[0,1]$

Step 2. If $u \leq \frac{p(x)}{Mq(x)}$, return $x = y$. Else, repeat from step 1.

The full procedure can be found in the pseudocode for algorithm 5 in the appendix and the resulting log-LFQ distribution is shown in comparison to the original in figure 3.5.

As described and seen in the plots above, the couple of approaches focusing on MNAR MVs imputation will substitute all missing values with something towards the low side

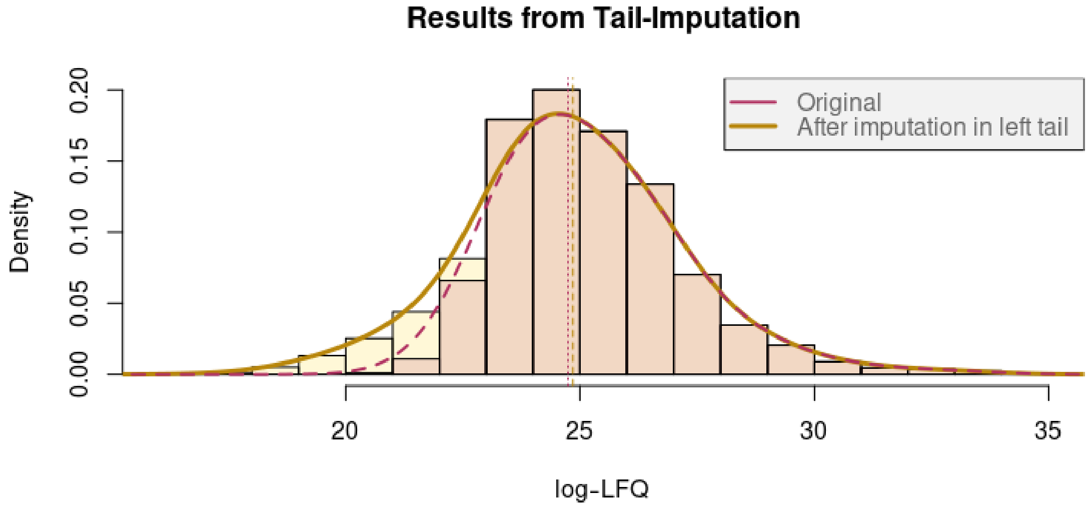


Figure 3.5: Distribution of the original log-LFQ values in the data (red line and red-histogram bars) and distribution of the log-LFQ values after the imputation (yellow). The median of the original distribution and the mean of the complete distribution after the imputation are presented with vertical dashed lines (they almost overlap).
Note, MVs are counted towards the size of the first distribution so the area of the shown red density is not equal to one.

of the spectrum and may not be suitable if the data is expected to be mostly missing at random, so next we discuss procedures which are better suited for randomly missing data.

MCAR-tailored missing value imputation methods are instead suggested to be more effective in such cases when most values are missing at random. They are non-specific to the context in which the data is gathered, thus, there are many statistical developments which apply here. Standard imputation studied in statistics is mean imputation; that is, if data is continuous, any missing value is substituted with the mean of the sample, whereas for categorical data, substitution is with the mode. This basic technique assumes, among others, normality and homogeneity of the data. In proteomics and results from LFQ classification data is not coming from one general distribution, but rather, every protein has its own profile according to the target. Therefore, this and similar methods based on estimating a universal distribution to impute from would not apply in this case.

There are other techniques, many of which based on some machine learning (ML), which have a somewhat individual prediction for every missing value. Perhaps the earliest such method is the linear regression estimation by Buck [15]. This method relies on enough completely filled out entries on which to calculate the regression coefficients for models to estimate the missing values. The method clearly implies a consistent pattern across columns and is, thus, unsuitable for data on replicates.

Nevertheless, there are some general imputation methods for MCAR values which can work well for the protein data. For the purpose of brevity, we elaborate on two such techniques, the ML method of k-Nearest Neighbours (kNN) [21, 7] and the statistical approach of SVDimpute [65] based on the singular value decomposition (SVD). These are relevant in recent research on the topic [22, 38, 46, 71] and have been implemented in popular analysis software [67].

The kNN method, as the name suggests, involves finding the complete observations which have the closest characteristic to the point of interest. The algorithm needs only two decisions: the distance measure and the value of k. For continuous variables such as the LFQ values an appropriate measure is the standard Euclidean distance. The choice of k is more complicated as it should balance between bias and variance of the predictions (imputations). On the one hand, Cover and Hart [21] prove that $k = 1$ is admissible and the lower k is, the lower the bias. However, variance for low values of k is large and decreases as k increases. After deciding on the algorithm procedure, the imputation is done for each missing value separately, thus, using a lot of computational resource and scaling poorly, which is the main drawback of the method.

The SVDimpute method uses SVD to consecutively decompose the initial matrix of LFQ records and, then, predict an approximation based on either the top or top k singular values and their corresponding singular vectors. This requires a complete initial matrix which is initialised by substituting MVs with their respective row averages. The algorithm relies on the fact that the biggest singular values account for most of the variability. Therefore, predictions which are made can predict as much variance as possible without overfitting. As is the case with the kNN algorithm, here again there is a danger of not predicting enough variance if k is low and so, a balance should be found. Results from [65] *et al.* from trying different thresholds and datasets suggest that best results are achieved when approximately 20% of the singular values are chosen for predictions.

The kNN and SVDimpute methods are suggested to work similarly well and being able to handle up to 20% missing values without a significant drop in quality. SVDimpute displays slight superiority above kNN when data has some structure to it (e.g. time-

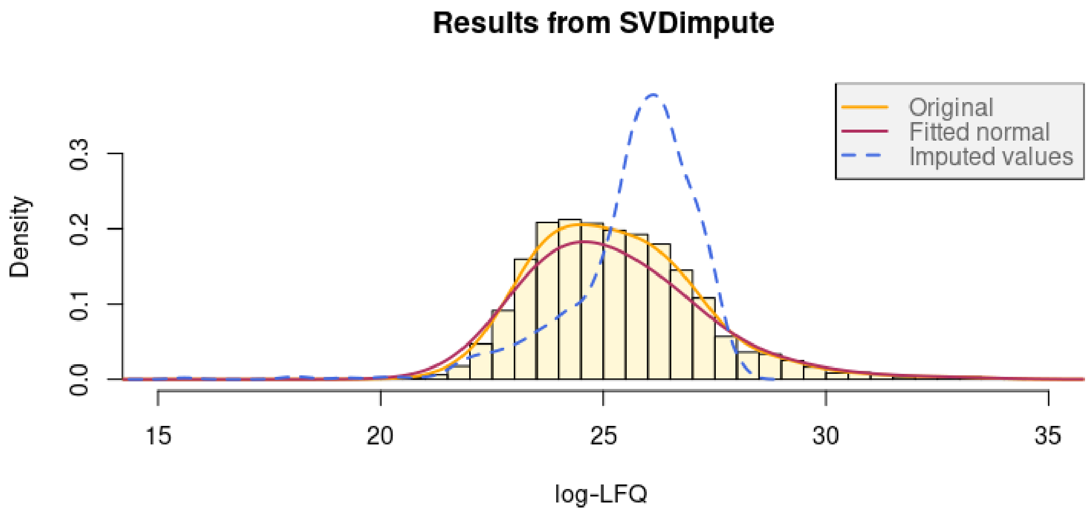


Figure 3.6: Distribution of the original log-LFQ values in the data (red line) and distribution of the log-LFQ values after the SVD-based imputation (orange line and histogram bars).
Note, MVs are counted towards the size of the distribution but are missing in the plot, so the areas under the density curves shown are not equal to one.

series or multiple conditions) while kNN is slightly better when no such patterns exist [65, 75, 29]. Pseudocode for both the kNN and SVDimpute algorithms presenting some more detail for the steps involved is presented in the appendix.

In the case of the RAB7 data, we have already seen that the sample of LFQ values has a right-skewed distribution (figure 3.4). The general assumption is that log-LFQ intensity values would have a normal distribution. This is depicted by the normal distribution in the figure (green line).

The fact that the original distribution is right-skewed may indicate that observations from the lower tail are missing, thus, pointing to the values being MNAR. However, a taller-than-expected peak suggests random reading errors are possible. Thus, we have decided to implement both an MNAR and an MCAR method and visually assess and compare those to determine the method which is best for the specific circumstance.

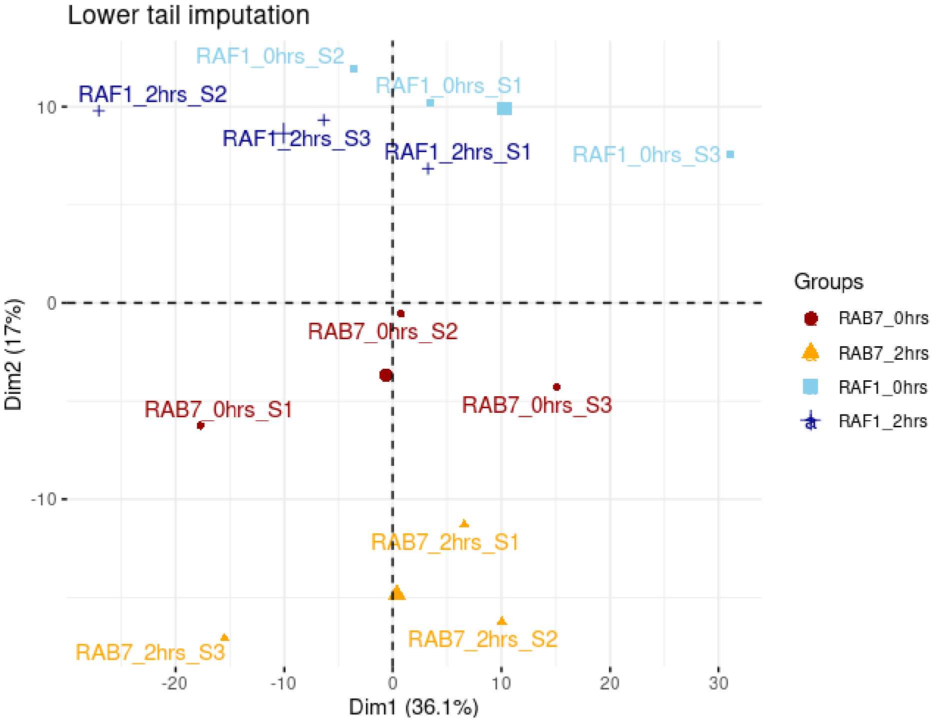
From the presented MCAR methods, the SVD-based imputation is well-suited to this particular dataset. The RAB7 data has some structure due to the two proteins and time points and the method can take advantage of that. Since the log-transformation reduces variance and does not preserve the exact distances, we do the imputation before transforming the data.

The results are shown in figure 3.6 which presents the original distribution alongside the log-LFQ values after the imputation. The distribution of the imputed values is included as well (as a blue dashed line). From the latter it is clear that, opposite to the MNAR method, in this case the imputed values are mostly not in the lower-ranges and even show a strongly left-skewed density. Overall, this pushes the new log-LFQ density to look similar to a normal distribution, but the density is ‘filled in‘ on the top side rather than the bottom.

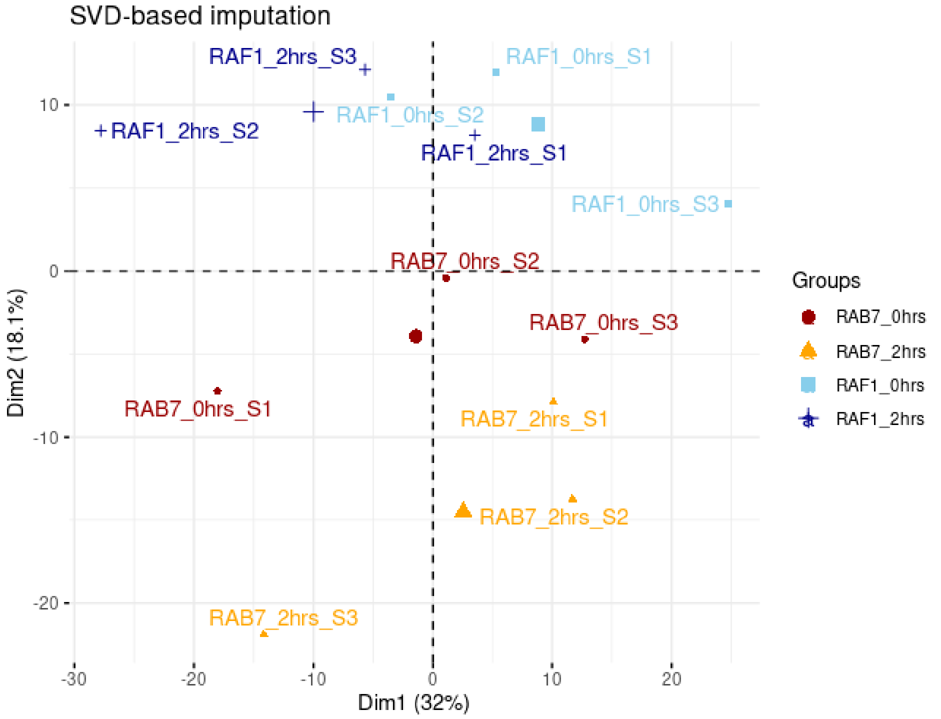
Comparison and inspection of results Visual inspection of the distributions can help to ascertain that the missing values are imputed in sensible ranges when we know what those are. However, it is a good additional step to also run principal component analysis to determine if the replicate and condition structure is preserved. It is worth noting that this step can serve as a way to compare the imputation methods. Yet, even if the imputation method is chosen, it can still serve as a quality check. If the PCA does not show any abnormalities, we proceed with analysis. If not, measures should be taken to ensure quality is good and, for instance, if there is a problematic replicate, it might be discarded.

The PCA plots for RAB7 with the two imputation methods are presented side-by-side in figure 3.7. They indicate some separation of the conditions and time points which is especially true for the lower tail imputation method (left subplot) and the RAB7 time points[6]. Therefore, in this case there is no need for further investigation and the lower-tail imputation method is preferred over the SVD-based one. The data obtained after this imputation is taken for further analysis.

[6] It is worth pointing out that any separation between the two RAF1 conditions suggests some distinctive activity related to the choice of time points and respectively, cilia disassembly. As discussed previously, RAF1 is normally assumed to not have any ciliary role, so no wide separation between the two time points is expected but can be viewed as a line of further study (beyond the scope of this thesis). This does not seem to be the case (the samples are mixed), but it is difficult to assess whether there is a difference due to the low total explained variance by PC1 and PC2 and small number of replicates.



(a) PCA plot for the data after imputation with the lower-tail imputation method. The y-axis is inverted for easier visual comparison with (b).



(b) PCA plots for the final after imputation with the SVD-based method.

Figure 3.7: PCA plots of the data after imputation.

3.4 Statistics: Analysis and results

As a first analysis step, the data is compared to a control to understand the significance of the results. The control is chosen as mentioned previously. For any specific prey from the dataset, higher LFQ intensities in the replicates of the target protein compared to those for the control, would mean enrichment of that prey in binding with the target and would be labelled as specific binding. To make this comparison robust, we do statistical testing to understand significance.

While we do not know the distribution of the LFQ intensities, we assumed normality of the log-LFQ values and imputed missing values under this assumption. Hence, working with log-LFQ values from now on, we can assume their distribution to be normal.

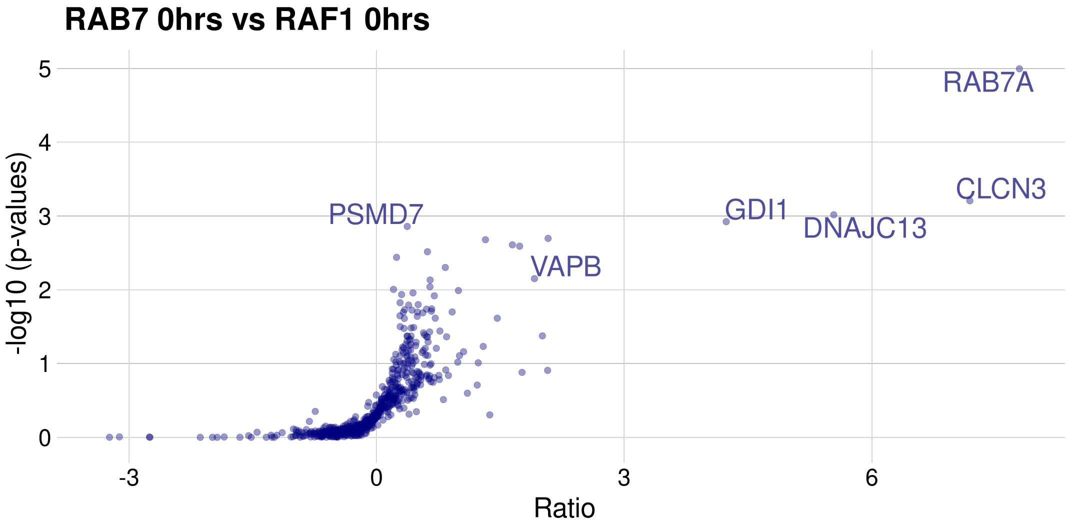
With this assumption, it is possible to apply the Student t-test to compare target and control. In particular, we use the one-sided t-test with the following null and alternative hypotheses:

H_0: The mean log-LFQ of the target is no larger than the mean log-LFQ of the control.
H_1: The mean log-LFQ of the target is greater than the mean log-LFQ of the control.

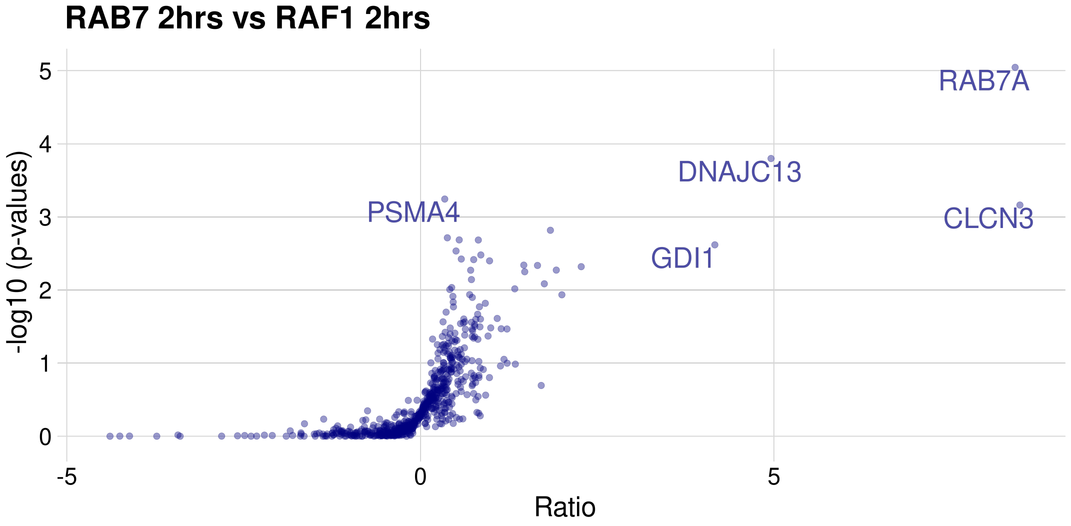
This t-test is conducted for every gene in the sample separately. The goal is to discover proteins which are over-represented among the preys of the target compared to the control which would mean they are specific interactors for the gene tested.

The RAB7 dataset we use to demonstrate the procedure has two time points and we apply the same testing separately for each one. On its own, each comparison shows which genes are significantly enriched at the specific time point. The results are presented in figure 3.8.

A useful quality check at this stage is to look for RAB7. It should, if the experiments were conducted properly, be enriched in its respective samples and imputed for the rest. Optimally, RAB7 should only be among the results from the RAB7 pull-down, meaning, its log-LFQ values for the bait RAF1 should be imputed and so, in the lower range of the distribution. It may happen that RAF1 captures RAB7, but in this case, still the log-LFQ intensity should be much lower than in the RAB7-baited experiments. That is, in terms of the log-ratio between the intensities in the RAB7 pull-down vs the RAF1 one, it should have a high positive ratio. If that is not the case, it means that there may be a strong connection between the two proteins which makes RAF1 inappropriate general control. Notably, RAB7 has the highest log-ratio which provides confidence in the experimental design and results.



(a) Plot when the baits are RAB7 and RAF1 and the time point is 0 hours.



(b) Plot when the baits are RAB7 and RAF1 and the time point is 2 hours.

Figure 3.8: Plot of p-values log-ratio (base 2) of RAB7-to-RAF1 LFQ intensities against the negative logarithm of the p-values resulting from t-test comparison as described in section 3.4. Specifically, log-LFQ intensities from one bait are compared to the other.

It is wrong, however, to leave the testing at the stage of raw p-values from the t-tests. As the testing is done separately on many genes, a finalising step is to apply a multiple testing correction. We have normally applied the Benjamini-Hochberg correction [8] which keeps the false discovery rate (FDR) low. It admits some false positives, but is more powerful than other methods. Given that the current aim in the affinity proteomics experiments which we analyse is to capture as many significant observations as possible, rather than finding only few very certain interactors, with this correction we am sacrificing precision slightly in order to achieve much better recall.

However, in the case of the 3-replicate RAB7 experiment and 680 total tests for each tested pair, the power of most individual tests is too small. The median power falls below 20%. Summary representation of the power densities is shown in figure 3.9.

This is reflected in the p-values which are relatively high. Moreover, readjusting them with a multiple-testing correction leaves only two significant observations in the RAF1 comparison at time point 0 (RAB7 and CLCN3). While this may not be an issue, we designed the experiment to find a difference at the second time point, cilia disassembly. We expect to have differences at that point between RAB7 and RAF1, but the only significant distinctions are in the LFQ values of four proteins: RAB7, CLCN3, DNAJC13 and PSMA4 (p-values: 0.006, 0.036, 0.036, 0.097). All other proteins do not have so large a difference in their log-LFQ values to overcome the power problem. Moreover, none of these four candidates is significantly different in its concentrations between the time points in the RAB7 experiments.

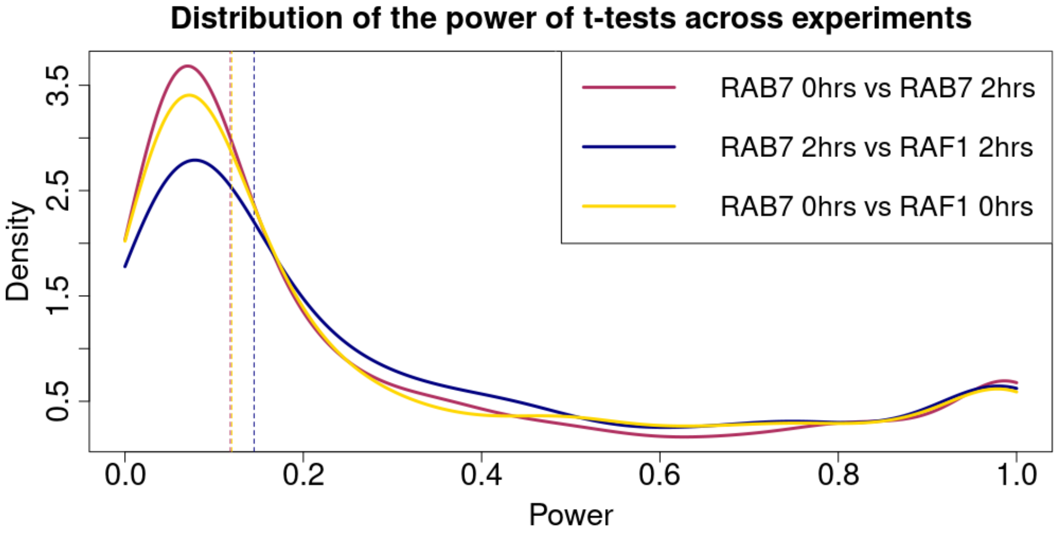


Figure 3.9: Densities for the power of the t-tests for the three different comparisons. Dashed lines represent the medians.

Some researchers rely on additional data to help with cleaning the results before starting any statistical analysis. Perhaps the most prominent such dataset is the CRAPome [50]. It consists of proteins which are common contaminants and unspecific binders. Excluding them in advance can reduce the number of tests performed and, thus, increase the number of significant results based on the adjusted p-value. The approach is controversial as, by constantly filtering them out, we will never find the CRAPome proteins even if they are true interactors sometimes. Another approach is increasing the number of replicates performed but this can raise costs and may not be as sustainable on a large scale.

Finally, in other analyses, where there are significantly enriched proteins precipitating with the target, but not the control, there can be an additional step comparing conditions of the target. For example, there can be another time point or another experiment with a mutated target protein. In those cases, the goal is to determine whether the mutation, or new condition, brings about a significant change. For this purpose, we conduct a second student t-test to compare the WT protein results to the mutated protein results. It depends on the goal of the experiment, but normally the test we do is 2-sided in this case, as we are interested in both proteins which stop being active and those who appear in the complex from one condition to the next. The results from the tests should again be corrected for multiple testing.

In the case with the RAB7 experiments, there are results at time point 0 hours and 2 hours and the ultimate goal of the experiment would be to find out which genes change in their binding between the time points serving as mechanistic insight into the process of ciliogenesis.

This last step can only be completed once we have done the comparisons with the control as it is important to overlay the significant genes from both analyses. The reason is, that only changes in genes which were significantly hypothesised to interact with the target, i.e. significantly enriched in comparison to the control experiment, are important. If we do not have initial or subsequent (at 2 hours) information on difference with the control, any difference between the time points is irrelevant.

3.5 Limitations and discussion

As we have just seen, the power of the test is a major limitation to the usefulness of the end results. This is affected by the decisions in all previous steps, starting from the experimental design and accuracy. If the experiment is replicated more times, the random fluctuations of the log-LFQ distribution of each individual prey and condition will decrease. More accurate machine and better breaking down and ionisation of proteins can also improve downstream analysis. The benefits are even two-fold: improved accuracy of already detected quantities and reduction in the number of missing values. Missing values are a great part of this chapter since the way to handle them influences the results significantly. Yet, it is clear that diminishing their numbers would be followed by increased accuracy.

The way to handle missing values is still unclear. While we have observed more realistic results using a left-tail imputation method, this is not suitable every time. Furthermore, *'realistic'* is dependent on assumption we make about the overall distribution of the log-LFQ intensities and the hypothesised reasons for their occurrence. In the analysis we rely on the two assumptions that:

- all log-LFQ intensity values are normally distributed in aggregate;
- MVs occur for non-random (abundance) reasons and their true measurements are concentrated along the lower-end of the distribution.

With regard to the specific testing, there is also the consideration that all imputed values will influence the means tested by the Student t-test. If imputed values are close to the rest of the samples, this may obscure the mean difference in comparisons. Another approach is to assume missing values to be completely unobserved observations (substituting with zeros for the log-LFQ values). This can cause the difference between the means to be more pronounced but it suffers from, for instance, the bias of skewing observations and breaking the normality assumption for the t-test.

Other tests can be applied to compare the conditions. It is thought that the data is normal, so t-test is more powerful than Wilcoxon test. However, if imputations lead to non-normal distribution, the latter test can be more appropriate. Another idea is to consider the mass-spectrometer in a similar way as a Geiger counter, detecting particles or not with some probability. Then the LFQ values will represent a discrete distribution and may be treated with a χ^2-statistic. Perhaps other models based on counting can be developed.

However, with most imputation methods the LFQ values do not form normal samples and, thus, cannot be treated with that statistic since it assumes normal distribution of each observation (row). The question of whether the lower-tail imputation that we presented in this chapter can complement a χ^2 analysis in a non-biased way, and whether using the statistic can be taken advantage of can be a point for further testing.

Finally, as we are doing the comparison statistics here, regardless of the specific test used, a major problem with the technique is that there is no perfectly annotated protein. Controls are chosen so that they do not have a known overlap with the function of the target protein, which involves tremendous effort searching, but even if a control is not known to have anything in common with the target protein, more complicating functions may be uncovered in the future.

For instance, RAF1 is often used as non-ciliary control [27, 34, 10], but recent research speculates that it may have some ciliary involvement specifically for the process of cilia disassembly. In particular, a study by Dong *et al.* published after our experiments were conducted, reveal that the drug dabrafenib inhibits cilia disassembly. That drug is specific inhibitor to BRAF, but has also been observed to work together with RAF1 in some settings[7] [44]. This study creates some doubt around the assumption that RAF1 is not related to the ciliary landscape and, in turn, poses the question whether the control is appropriate or it would have obscured some true interactions.

In general, the method of target-to-control comparison relies on the highly limiting assumption that the direct and indirect binding of the control is non-specific to the aims of the experiment and to remove proteins with similar discovery levels in target and control means removing background noise. In this way, any overlaps, even when they have genuine relevance to the target, will not be detected as true and significant possible interactors to the bait protein.

In the following chapter we discuss a different method which we designed in order to avoid the need for controls, thus, reducing the problems and assumptions surrounding that and analysing the data in a less biased way.

[7] Specifically, it was observed that for the activation of the MAPK pathway in regular RAF cell lines, dabrafenib requires RAF1. This was observed to not be the case with mutated BRAF.

4 WeSA

A main goal of my doctoral project has been to design, refine and test a statistic for interrogating protein-protein interaction networks. Presenting the main theory and results addressing this goal is the focus of the current chapter.

The statistic we present and call WeSA is designed to not only respond to the challenge that handling large datasets is, but also leverage the growing amount of information. In attempts to uncover the complex relationships in the human interactome, researchers develop experimental protocols which output huge datasets needing analysis. Instead of requiring more resource expenditure to ensure control data, we design WeSA - a score analysing the new data, based on all previously available information from other studies.

In addition to being highly efficient, the score provides a less biased alternative to the regular case-control experiments as discussed in the previous chapter (e.g. [10]). This is due to the fact that the concept of a control experiment is substituted by large data from a variety of proteins, thus, eliminating the bias of choice and significantly reducing the bias of incomplete knowledge.

As we present the model function and testing, we will also elaborate on its advantage over other scoring functions. In contrast to static databases like STRING [63], WeSA allows for dynamic updating of information. Furthermore, heterogeneity between datasets and methods usually requires individual adaptations based on the method, e.g. fitting of parameters or incorporating a measure of estimated accuracy for each individual study [20]. In contrast, WeSA performs well in combining datasets even without the presence of additional information about each dataset. With WeSA one can add to the data which is already there and analyse their own interactions in this context. In section 4.2.4, we discuss in detail the implementation of the WeSA computation algorithm and an associated web tool.

In the results section of the chapter, we test the performance of WeSA against several hypotheses using multiple testing methods, datasets and mathematical models.

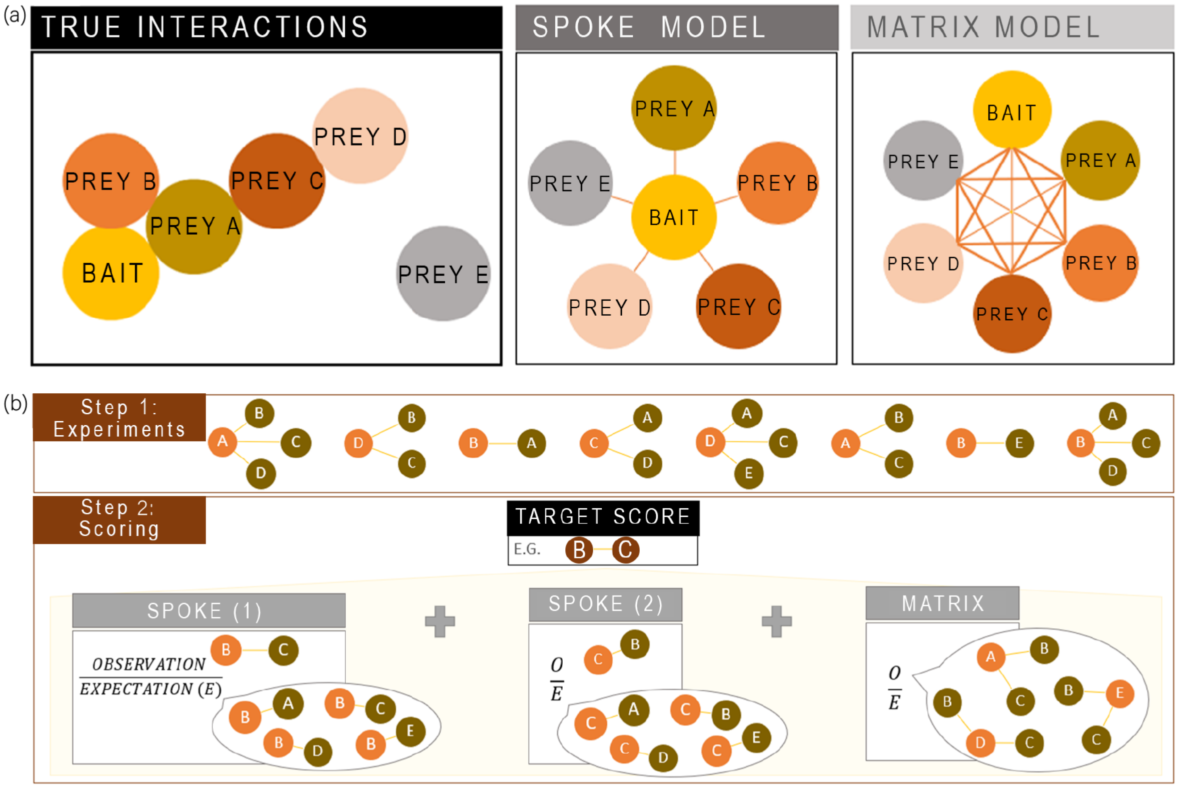


Figure 4.1: (a) An example arrangement of proteins which are captured by AP. Some of them may exist in nature as a complex, others (prey E) are contaminants. The two models (spoke and matrix) of recording results from AP experiments are shown in the top and right panels. (b) Illustration of the steps for calculating WeSA scores. Proteins are shown as nodes of a network and connections between them are drawn as edges. The result of 8 imagined AP experiments are shown; the bright orange proteins are the baits and their preys are connected only to the baits (spoke model). The bottom panel shows the O:E terms with the observations from step 1, which are relevant for calculating each one.

4.1 Datasets

Intact

We retrieved IntAct data on 22 April 2021. It contains 1,156,386 records of protein-protein interactions about half of which are human. From human interactions, we also make an effort to exclude direct pairwise studies of interactions such as yeast two-hybrid and protein complementation assays. This is done, because there are many interactions which, while existing, are random (non-specific) with respect to function [14] and they do not carry information for protein complexes as a whole. Since the idea of the WeSA score described below is to separate interactions by their specificity, random pairwise testing is not of much use.

The filtered IntAct data contains 291.6 thousand pairs of human protein-protein interactions from affinity-purification-like experiments [52]. Since IntAct records are done using a spoke model (figure 4.1a, centre), a matrix expansion (as in figure 4.1a, right) resulted in more than 14 million pairs whose marginal data provided more context for the SA score. We left splicing unchanged in this context but different splicing forms can be considered as equivalent if needed.

We tested the scores against the complexes contained in CORUM. The proteins for more than 96% of the CORUM complexes listed there were also covered in IntAct and, therefore, included in the evaluation.

BioPlex

A critique of interactions recorded in IntAct is their bias towards overly stable or high-affinity interactions [4]. AP methods used until recently, and which predominate in IntAct, aimed to wash away artefacts and had a tendency to remove weaker interaction partners. As mentioned in Chapter 3 proteins tend to have between 2 and not more than 100 partners. This is indeed the case in Intact: there are 13 records of interactions pulling down more than 500 proteins and only 5 with more than 1000 proteins in the final precipitate.

Despite these few outliers, generally, the captured preys are fewer in number and stably attached to the baits. That is why to understand transient relations, experiments are changing towards less washing and larger prey spectra.

To our knowledge, the BioPlex Interactome [37] is the largest study on the human interactome in which only a single tag (and thus a single washing stage) is used and so, filtering is significantly reduced to allow the observation of weaker interactions. In total BioPlex comprises 15,650 experiments and has identified 7,766,619 unique bait-prey pairs. The average number of preys per experiment in BioPlex surpasses 620, which is incomparable to the previously discussed IntAct data. These experiments can lead to new insight into protein-protein networks, but require tools to filter the noise out.

BioGRID

The third database we use throughout this section to load experimental interaction data is the Biological General Repository for Interaction Datasets (BioGRID) [53]. We retrieved the entire database on 13.12.2022.

Compared to the above databases, BioGRID is less filtered and consists of more data than IntAct. There is no complete curation for the Homo Sapiens data, so the landscape still contains a lot of noise. On the other hand, BioGRID contains fewer interactions per experiment than BioPlex; the interactions reported in BioGRID have been pre-filtered.

The full database records information across species and experiment types, but we have filtered for relevance. As with previous experiments, we only focused on human proteins. For the calculations of WeSA scores described in the section, we also filtered out any targeted experiments so the data we take into account comes from some type of bait-prey experiments.

BioGRID contains 667,413 relevant protein pairs 520,101 of which - unique. They come from 30,200 separate publications, corresponding to an average of around 22 pairs per publication.

CORUM

We compared results from the analysis of the three databases to the data on complexes recorded in the Comprehensive Resource of Mammalian Protein Complexes (CORUM) database (retrieved: 28.09.2022) [59].

There are 3,538 human complexes in CORUM. Their sizes range from 2 to 129 (see table 6.1 in the Appendix). The largest of those relate to the spliceosome and ribosome. Specifically, 4 of the biggest 10 are complexes E, A, C and the smaller B complex which are all part of the splicing cycle. Equally many among the top 10 are the records from the ribosome including the ribosome complex alongside its 55S, 39S and 60S ribosome subunits. The smallest complexes are simply pairs of interacting proteins annotated to function together in a stable complex.

We considered only human proteins. Proteins from other organisms were also excluded even if they have a human ortholog. For example, the IFT-B complex is recorded to contain the mouse gene Cluap1. In the particular instance, Cluap1 is dropped from the complex, because our analysis focuses solely on human genes.

CORUM is updated regularly and archive files such as our first retrieval in 2020 [31] show that the database has increased significantly. From 2,171 in 2020, the number of recorded complexes has increased by more than 60% while also big units such as the spliceosome are broken down into smaller and more specific components. This enrichment of information is expected to continue in the future giving us an even better benchmark and protein complexes landscape.

These datasets are not entirely independent. In particular, data from BioGRID and IntAct contributes (obviously) to how human complexes are determined and annotated inside of CORUM. We did not attempt to correct for this, but it is important to bear in mind when observing the performance of (in particular) BioGRID and IntAct in the Results section below.

4.2 Methods

4.2.1 WeSA score

Inspired by the likelihood ratio statistic, we use a similar approach to modify the socio-affinity metric [30], which will be referred to as weighted socio-affinity (WeSA, W_{ij}) and expressed as:

$$W_{ij} = n_{ij} \times S_{ij} + n_{ji} \times S_{ji} + n_{ij}^{M} \times M_{ij} \tag{4.1}$$

where n_{ij} is the observed number of retrievals of j, when i is the experiment bait, and n_{ij}^{M} is the scaled observed number of joint retrievals of i and j when both are captured in the matrix of another bait.

The higher the WeSA score, the better is the likelihood that the interaction is real as it is observed more times than expected. Conversely, low scores mean fewer than expected occurrences and a possible lack of specificity.

The main terms S_{ij},S_{ji},M_{ij} are the components of the SA score based on the observed-to-expected ratio of interactions between protein pairs. They are shown in figure 4.1b.

Let X_{ij} is the random variable modelling the number of times a protein i is bait and a protein j is prey in the same experiment and let X_{ji} model the analogous situation when the roles of i and j are reversed.[1] The observed values for X_{ij} and X_{ji} from the sample (the collected data of experiments) are known and we denote them by O_{ij} and O_{ji}. Let $E_{ij} = \mathbb{E}[X_{ij}]$ and $E_{ji} = \mathbb{E}[X_{ji}]$. With this notation, we define the first two terms of equation 4.1, also labelled 'spoke' terms, as:

$$S_{ij} = \log \frac{O_{ij}}{E_{ij}} \qquad S_{ji} = \log \frac{O_{ji}}{E_{ji}} \tag{4.2}$$

Let Y_{ij} model the number of times a protein i and a protein j were prey proteins in the precipitate of the same bait. The matrix term M_{ij} is calculated as the ratio $M_{ij} = \log \frac{O^M_{ij}}{E^M_{ij}}$, where O^M_{ij} is the observed value of Y_{ij} from the sample data and $E^M_{ij} = \mathbb{E}[Y_{ij}]$.

The expectation in all three cases is calculated as would be expected from a directed configuration model, where we fix the degrees of bait nodes to the average number of preys they 'caught' across experiments in the sample data.

The configuration model is explained in more detail in the Appendix, but the idea is sketched briefly here.

The configuration model is a common model for simulating random networks with a specified degree distribution. The end result is an arbitrarily big collection of different random networks with the same degree distribution as a specified input. This allows us to approximate expectation (and other useful parameters if needed) using the sample mean of the collection of configuration networks.

We use the sample network, to infer the degree distribution for the nodes with two modifications. We keep each experiment (and each bait, respectively) as a separate node and we distinguish between two categories, baits and preys, in order to keep track of the terms of WeSA. Also, we simplify the model by standardising the degree of baits which are the same. Instead of having the degree of each bait exactly the same as in experiments, all baits which are the same protein have the same degree that is equal to the average degree of the protein when it was bait.

Expected number and mathematical estimation of parameters In general, the expectation can be calculated based on the configuration graph. This is the only option

[1] In all modelling, the proteins i and j may have multiple copies and be observed together in various different combinations.

when multiple conditions restrict the model beyond its degree distribution. However, this method is time-consuming and, in many cases, relies on excessive computational power. In the case of protein-protein interaction networks, the randomisations can be avoided by computing the closed-form of the expectation presented below.

For brevity, we present the case for X_{ij}, but all spoke terms are calculated analogously. Let us define $Z_1, \ldots, Z_{n_{bait}}$ to be the independent identically distributed (i.i.d.) random variables (r.v.) modelling the number of times we see the bait i and the prey j together in one experiment. Then,

$$X_{ij} = \sum_{t=1}^{n_{bait}} Z_t$$

Let K be the discrete r.v. of choosing one bait from the sample space of all baits, Ω, and let J be the r.v. modelling the number of times that j is prey in a randomly chosen experiment. With this notation:

$$\mathbb{E}[Z_t] = \mathbb{E}[J \mid K = i]\mathbb{P}(K = i)$$

We can estimate $\mathbb{P}(K = i)$ using the frequency of i as bait in the sample data, f_i^{bait}. The probability of choosing a prey which is j can also be approximated from the sample as the frequency of j among the preys in the sample, f_j^{prey}. J depends on K for the total number of connections it can form. If n_i^{prey} is the average number of preys captured by the bait i, then $J \mid K = i \sim Binomial(n_i^{prey}, f_j^{prey})$. Therefore, the closed form of the expectation is:

$$\begin{aligned} E_{ij} &= \mathbb{E}[\sum_{t=1}^{n_{bait}} Z_t] = \sum_{t=1}^{n_{bait}} \mathbb{E}[Z_t] = \sum_{t=1}^{n_{bait}} \mathbb{E}[J \mid K = i]\mathbb{P}(K = i) \\ &= \sum_{t=1}^{n_{bait}} (n_i^{prey} f_j^{prey}) f_i^{bait} = n_{bait} n_i^{prey} f_j^{prey} f_i^{bait} \end{aligned} \tag{4.3}$$

This derivation does not exclude the possibility of capturing the same prey multiple times in the same experiment. This is done to keep the total sum of all spoke expectations constant. Moreover, databases of experimental results do not normally report the quantities of individual preys in an experiment and instead we only know if the prey is present or not. This is likely to have an effect on the estimation of the average number of preys per bait (i.e. the estimate n_i^{prey} in our model may be lower than what it could be if abundances from each experiment are used). In human PPI networks, all human proteins can be preys, so the frequency f_j^{prey} is normally very small compared to n_i^{prey}, which makes the occurrence of multiple links between i and j in the same experiment unlikely. However, if there was a more precise database from which to infer our estimates or the probabilities were bigger, instead of E_{ij} defined above, we could use the expected number of experiments in which i is seen as bait and j is seen as prey at least once. This expectation has the form:

$$n_{bait} f_i^{bait}\big(1-(1-f_j^{prey})^{n_i^{prey}}\big) = n_{bait} f_i^{bait}\Big(f_j^{prey} n_i^{prey} - \mathcal{O}\big((f_j^{prey} n_i^{prey})^2\big)\Big) \tag{4.4}$$

The use of the binomial expansion shows the condition for $f_j^{prey} n_i^{prey}$ which decreases the possibility of multiple edges forming between prey j and the bait i in a single experiment. When $f_j^{prey} n_i^{prey}$ is sufficiently small, the equations 4.4 and 4.3 are approximately equal.

We define I (similarly to J) to be the r.v. for the number of times i is observed as prey in a randomly chosen experiment. The number of edges between i and j in the matrix for a single experiment in which there are I preys i and J preys j, is equal to $\frac{IJ}{2}$. Let $k_1, \ldots, k_{n_{bait}}$ be the n_{bait} baits in the sample experiment data, then:

$$E_{ij}^M = \mathbb{E}\left[\sum_{k=k_1}^{k=k_{n_{bait}}} \left(\frac{IJ}{2} \mid K=k\right)\right] = \frac{1}{2}\sum_{k=k_1}^{k=k_{n_{bait}}} \mathbb{E}[IJ \mid K=k] \tag{4.5}$$

Conditional on $K=k$, we can model the distribution of the preys in an experiment as $Multinomial(n_k^{prey}, p)$, where p is the vector of probabilities for the occurrence of each prey protein. The expectation and covariance from the multinomial distribution are:

$$\begin{aligned} \mathbb{E}[I \mid K=k] &= n_k^{prey} f_i^{prey} \\ \mathbb{E}[J \mid K=k] &= n_k^{prey} f_j^{prey} \\ cov(I,J \mid K=k) &= -n_k^{prey} f_i^{prey} f_j^{prey} \end{aligned}$$

By definition,

$$\begin{aligned} cov(I,J) &= \mathbb{E}[IJ] - \mathbb{E}[I]\mathbb{E}[J] \\ \Rightarrow \quad \mathbb{E}[IJ \mid K=k] &= cov(I,J \mid K=k) + \mathbb{E}[I \mid K=k]\mathbb{E}[J \mid K=k] \\ &= -n_k^{prey} f_i^{prey} f_j^{prey} + (n_k^{prey} f_i^{prey})(n_k^{prey} f_j^{prey}) \\ &= n_k^{prey}(n_k^{prey}-1) f_i^{prey} f_j^{prey} \end{aligned}$$

Finally, equation 4.5 becomes:

$$E_{ij}^{M} = f_i^{prey} f_j^{prey} \sum_{k=k_1}^{k=k_{n_{bait}}} \binom{n_k^{prey}}{2} \tag{4.6}$$

Weights in the WeSA formula based on the observed counts in the data make the final WeSA sum less prone to errors with the increase of observation numbers. That is expected since the number of times a co-occurrence is seen relates to higher confidence in the results being true.

Special attention here needs to be paid to the experiment precipitate containing too many proteins which can cause the matrix term to overshadow spoke terms contribution to WeSA. To counteract this effect scaling is done on the matrix term to bring it to the scale of the distribution of n_{ij}. If the matrix term is denoted as n_{ij}^{M}, n_{ij}^{M} is obtained from the raw matrix count, O_{ij}^{M}, of co-occurrences of the ij-pair, by scaling using the estimator, $\mu(n_{ij})$, of the mean number of observations in the spoke terms and $\mu(O_{ij}^{M})$, the mean number of observations in the matrix:

$$n_{ij}^{M} = O_{ij}^{M} \times \frac{\mu(n_{ij})}{\mu(O_{ij}^{M})}$$

The estimators for those averages are obtained from the sample means.

Experimental confidence extension. The score allows for the incorporation of additional information about the confidence in each experimental result. What is meant here is to use the knowledge of each researcher performing the experiments to assign a probability to the outcome of each experiment being indeed true and then use that probability instead of the binary classification of pairs as observed or not observed.

IntAct has records of these probabilities labelled ‘experimental confidence’ and they can be retrieved together with the data. They are not given for all experiments and the distribution is left-truncated around 0.3 while also not exhibiting normality around the truncation limit. This makes the records unfavourable to work with and we will only briefly touch on the possibility of using experimental confidence.

It is important to note that the derivation of WeSA terms remains similar even when using the experimental confidence. The only change is in the probability $\mathbb{P}(j^p \mid i^b)$, which, considering the confidence weights (c), becomes:

$$\mathbb{P}(j^p \mid i^b) = \frac{\sum c_{j=prey}}{\sum c}$$

This same note applies to the calculation of the matrix term in 4.6, where both f_i^{prey} and f_j^{prey} change.

4.2.2 ROC analysis

After the calculation of WeSA scores, we can compare them to CORUM complexes. This is done using receiver operating characteristic (ROC) analysis. The main tool in this analysis used below is the ROC curve which plots the true positive rate (TPR) against the false positive rate (FPR).

Its characteristic area under the curve (AUC) is the standard measure of the quality of the model. A ROC curve above the diagonal (1,1) vector corresponds to a better-than-random model performance. Respectively, AUC ≥ 0.5 is associated with predictions (or a model) better than random.

The ROC curve is a suitable testing metric and can be applied beyond this chapter of the thesis.

Threshold defined from ROC analysis. On several occasions in this work, after plotting the ROC curves, we also look at the optimal threshold. This is defined as the cutoff of the 'best balance' between TPR and FPR.

There are multiple popular ways to calculate this threshold [68]. We have implemented three of the most popular ones, namely, the Youden index [73], the concordance method [47] and the closest-to-(0,1) method [54]. In our implementations they are all close to overlapping with each other, so for simplicity in this written work we have only presented the closest-to-(0,1)-defined threshold.

This threshold is defined from the point on the ROC curve which is closest to the top left edge of the graph, i.e. the point with coordinates (0,1). That is, the threshold is defined as the value minimising the distance:

$$\sqrt{(1 - TPR)^2 + FPR^2}$$

Precision-Recall curve. The Precision-Recall (PR) curve plots precision against the recall. In the analysis of protein-protein interactions, there is no good benchmark, especially for negatives. As our test set of positives is based on CORUM complexes, we assume as negatives all other interactions. This is imprecise and probably a lot of the ‘negatives’ are real interactions which have not been discovered or are not part of a complex. For this reason, the PR curve would present a much lower estimate of the precision curve and is not regarded as highly informative.

4.2.3 Statistical testing: Mann-Whitney U test

In further analysis, data is grouped according to various criteria. In this chapter, there are multiple instances of testing across categories, but such testing is also required in subsequent chapters. As an example, in figure 4.10 of the chapter, WeSA scores for interactions confirmed in pairwise experiments (category 1) are compared to overall scores (category 2). To compare the continuous distributions in these two categories, we used Mann-Whitney U (MWU) test, also called the Wilcoxon rank-sum test.

This test is used to compare whether two distributions differ significantly by comparing their means. In the example mentioned above, it is applied to binned-distributions of the pairwise-confirmed set of interactions and the full set as the two distributions. An important characteristic is that the power of the test decreases with decreasing the sample size. For samples with fewer than 8 elements, it is impossible to achieve $p < 0.05$[2].

In comparison to other possible statistical tests for the purpose, below we briefly mention the t-test, chi-squared test, Binomial test and Fisher exact test. While every context is unique, we have found that in this MWU is applicable to the situations in this chapter.

- t-test: MWU is superior to t-test when the distribution is not normal. This is expected to be the case for most groups, as they can rely on thresholds and truncation. For instance, confirmed interactions are supposed to have WeSA scores from the higher end of the WeSA distribution (not normal).
- χ^2, Binomial and Fisher exact tests: These tests are applied to discrete variables and their counts across categories (in contrast to tests on continuous random variables). They test independence of categories.

[2] https://www.graphpad.com/guides/prism/latest/statistics/how_the_mann-whitney_test_works.htm

4.2.4 Technical methods

Everything in this chapter is built in Python 3.8 (Jupyter Notebook). A full list of used libraries for the calculation of WeSA scores is given in the Appendix.

The main WeSA calculation algorithm: running time, complexity, data types and storage

Due to its centrality to this thesis, I explain here the main algorithm used to calculate WeSA scores.

The full pseudocode of the simplest version of the algorithm is presented with all dependent functions in Algorithm 1 below. The three functions on which the algorithm depends are presented immediately after the main function in algorithms 2, 3, 4.

Algorithm 1: Calculating WeSA scores

```
Data: data
/* data has at least columns [bait, prey, identifier]                          */
Result: A dataframe A containing SA and WeSA scores
1 Begin
      /* Computing spoke terms in dataframe s:                                 */
2     s, result = compute_spoke(data)
3     Remove rows from s where bait is the same as prey
      /* Calculating n_prey = number of preys observed with a particular bait (excluding
         itself):                                                              */
4     n_id_prey = DataFrame with columns [bait, prey, n_id_prey] where n_id_prey
        equals the sum of n_i_prey grouped by [bait, identifier] from result
5     n_prey = DataFrame with columns [bait, term0] from n_id_prey, where term0 =
        preys * (preys-1)/2
6     n_prey = DataFrame with columns [bait, term] from n_prey, where term is sum of
        term0 grouped by bait
7     total_sum = sum of term column from n_prey
8     total_purifications = number of all unique [bait, identifier] pairs extracted from data
```

After calculating the dataframe with the spoke information, the algorithm goes through preparation for calculating matrix components for the same protein pairs as in s.

Then we calculated consecutively the observed matrix numbers O_matrix (rows 13-14), f_j_prey (line 15) and intermittent combination of matrix terms (lines 16-19) to enable calculations of more terms contributing to the matrix (lines 20-23) before ultimately concluding with the computation of the full matrix terms contributing to SA (line 24) and WeSA (line 25).

```
    /* Initialising prerequisites to combine all matrix terms: */
9   m = from s keep columns [bait, prey] but rename as [prey_x, prey_y]
10  data1 = dataframe of all unique [prey, identifier] pairs from data
11  dict_prey_ids = create_dict_prey_ids(data1)
12  matrix_counts = create_matr_list(s, dict_prey_ids)
```

```
    /* O_matrix = number of times that $i$ and $j$ are seen together in matrix: */
13  m[interactors] = join rowwise [prey_x, prey_y] from m
14  m[O_matrix] = Map interactors from m to counts from matrix_counts
15  f_j_prey = DataFrame with columns [prey, f_j_prey] from s (with a unique prey per row)
```

```
    /* Combining matrix data: */
16  m = Merge(m, f_j_prey) on prey_x based on existence in m
17  m = Merge(m, f_j_prey) on prey_y based on existence in m
18  m = Merge(m, n_prey) on prey_x based on existence in m AND rename the column term as term_x
19  m = Merge(m, n_prey) on prey_y based on existence in m AND rename the column term as term_y
```

```
    /* Computing binomial term in the expectation expression: */
20  m[binomial] = total_sum - (term_x + term_y) rowwise from m
    /* Compute matrix expectation: */
21  E_matrix = Multiply(f_x_prey, f_y_prey, binomial) rowwise from m
    /* Compute matrix scaling weight: */
22  E_Om = average of the counts in matrix_list
23  E_Os = average of O_spoke from s
24  m_weight = E_Os/E_Om
```

```
    /* Compute SA and WeSA terms for the matrix: */
24  m_ij = log(O_matrix/E_matrix) rowwise from m AND Replace NAs with 0
25  m[Lambda_m] = m_weight * m[O_matrix] * m[m_ij]
```

At line 25 both spoke and matrix terms have been individually combined. Therefore, the last few lines merge everything to output the final SA and WeSA scores for each protein pair.

```
    /* Combine all terms (spoke and matrix):                                        */
    s[interactors] = join rowwise [bait, prey] from s
    A = Merge(s, s) based on shared interactors in forward and reversed order
    A = Merge(A, m) on interactors AND drop rows with duplicated interactors
    A[SA] = s_ij + s_ji + m_ij rowwise from A
    A[WeSA] = Lambda_ij + Lambda_ji + Lambda_m rowwise from A
    return A
```

The algorithm consists of two main parts:

1. calculating the components of the spoke term given as:

$$S(ij) = O_{spoke} \log \frac{O_{spoke}}{E_{spoke}}$$

   where

$$E_{spoke} = f_i^{bait} n_{bait} f_j^{prey} n_i^{prey}$$

   This is done through the supplementary function compute_spoke.

2. calculating the matrix term contributors. The matrix term is given by the formula:

$$M(ij) = w_{matrix} O_{matrix} \log \frac{O_{matrix}}{E_{matrix}}$$

   where w_{matrix} is the scaling weight, O_{matrix} is the observed co-occurrence number and E_{matrix} is given by the formula:

$$E_{matrix} = f_i^{prey} f_j^{prey} \sum_{k \neq i,j} \binom{n_k^{prey}}{2}$$

The names of variables in the scripts shown are kept as close to the notation in these equations as possible.

Algorithm complexity The average complexity of the full function is $\mathcal{O}(n^2)$, where n is the number of protein pairs in the input data. This is due to the two supplementary functions, algorithms 3- 4, for creating the dictionaries used to compute the matrix terms. Working with Python dictionaries is preferred to the initial dataframe implementation, since retrieval from a dictionary is constant, while querying a dataframe normally has

Algorithm 2: Function compute_spoke

```
Data: data
Result: Two dataframes with results of spoke term calculations.
Begin
    /* f_i_bait = fraction of purifications where protein $i$ was bait and n_bait = total
       purifications */
    bait_identifier = DataFrame with all unique [identifier, bait] pairs extracted from data
    total_purifications = number of rows of bait_identifier
    n_i_bait = DataFrame with columns [bait, count (of the bait)] extracted from
      bait_identifier
    f_i_bait = DataFrame with columns [bait, f_i_bait (frequency of the bait)] extracted
      from n_i_bait

    /* Compute n_i_prey = number of preys retrieved for each particular bait $i$ */
    purified_preys = DataFrame with columns [bait, identifier, n_i_prey (corresponding
      number of preys)] extracted from data
    result = Merge (bait_identifier, purified_preys) on columns [identifier, bait]
    n_i_prey = from result sum n_i_prey grouped by bait to DataFrame with columns
      [bait, n_i_prey]
    n_i_prey = Merge (n_i_prey, n_i_bait) on column bait
    n_i_prey[n_i_prey] = n_i_prey/count

    /* Compute f_j_prey = fraction of all retrieved preys that were protein $j$ */
    f_j_prey = DataFrame with columns [prey, count (of the prey)] extracted from data
    f_j_prey[f_j_prey] = frequency from f_j_prey (calculated count/sum(counts))

    /* Compute $n_{i,j}$ = number of times that $i$ retrieves $j$ when $i$ is tagged */
    sa = DataFrame with columns [bait, prey, O_spoke (observed number of the pair)]
      extracted from data grouped by [bait, prey]
```

linear complexity.

The quadratic complexity is reduced further after the first computation by the separation of the culprit algorithms 3- 4 and storing their results on the non-volatile memory. This is added to algorithm 1 as a simple *if* statement encompassing lines 11-12 and running them only if the respective dictionaries have not been calculated yet. Instead, if they have been saved already, they are simply loaded back. This detail can be found in the unmodified Python script for the function in the Appendix.

```
    /* Summary: putting it all together for the spoke terms. */
    s = Merge(sa, f_j_prey) on prey, how = left
    s = Merge(s, n_i_prey) on bait, how = left
    s = Merge(s, f_i_bait) on bait, how = left
    s[E_spoke] = Multiply rowwise [f_i_bait, f_j_prey, n_i_prey, total_purifications]
    s[s_ij] = log(O_spoke/E_spoke) rowwise from s
    s[Lambda_ij] = O_spoke * s_ij rowwise from s
    return s, result
```

Algorithm 3: Function create_dict_prey_ids

```
Data: data
/* data has at least columns [bait, prey] to get the matrix terms from */
Result: A dictionary dict_prey_ids to keep the preys (keys) and their identifiers (values
        as lists).
Begin
    /* A function which creates a dictionary of prey proteins ad their identifiers. */
    INITIALISE dict_prey_ids
    for prey in the unique preys do
        dict_prey_ids[prey] = [list of all identifiers for that prey]
    end
    return dict_prey_ids /* The number of keys in the dictionary is equal to the number
        of unique preys (i.e. at most equal to data size) */
```

Algorithm 4: Function create_matr_list

```
Data: data; dict_prey_ids
Result: A dictionary matrix_list to keep the pairs of proteins (keys) and their co-
        occurrence counts in the matrix (values).
Begin
    /* A function which creates a dictionary of prey proteins and their identifiers. */
    INITIALISE matrix_list
    for $i$ going over the indices of data do
        if bait[$i$] from data is in dict_prey_ids then
            matr_list[ 'bait[$i$] ; prey[$i$]' ] = LENGTH of overlap between identifiers from
              dict_prey_ids corresponding to bait[$i$] and prey[$i$] from data
        end
    end
    return matr_list /* For pairs with spoke term (not only matrix), size of this dictionary
        is at most equal to the size of data. */
```

With the saving of the two matrix dictionaries, the complexity reduces to the complexity of the next bottlenecks, which are the merging operations on dataframes. There are several of those both in the calculation of spoke (alg. 2) and matrix components (alg. 1, lines 16-19 and 27-28). They all have average run time as $\mathcal{O}(n \log n)$. However, it should be noted that to achieve this for the matrix term calculations, we have restricted them to only pairs for which at least one protein was tested as a bait, too.

Updating the data. New information (experiments) can be added to the data easily without re-calculating the matrix dictionaries for the full data. Instead, the dictionaries from algorithms 3, 4 are created only for the new data and then summed by key with the ones for the old data. This would not increase the $\mathcal{O}(n \log n)$ complexity if the size of the newly added data is smaller than $\sqrt{n \log n}$.

The opportunity to update the matrix dictionaries in two steps by creating a new smaller dictionary and combining it with the old records is foundational for the build-up of the webtool presented next.

Webtool – tools for building options and interface

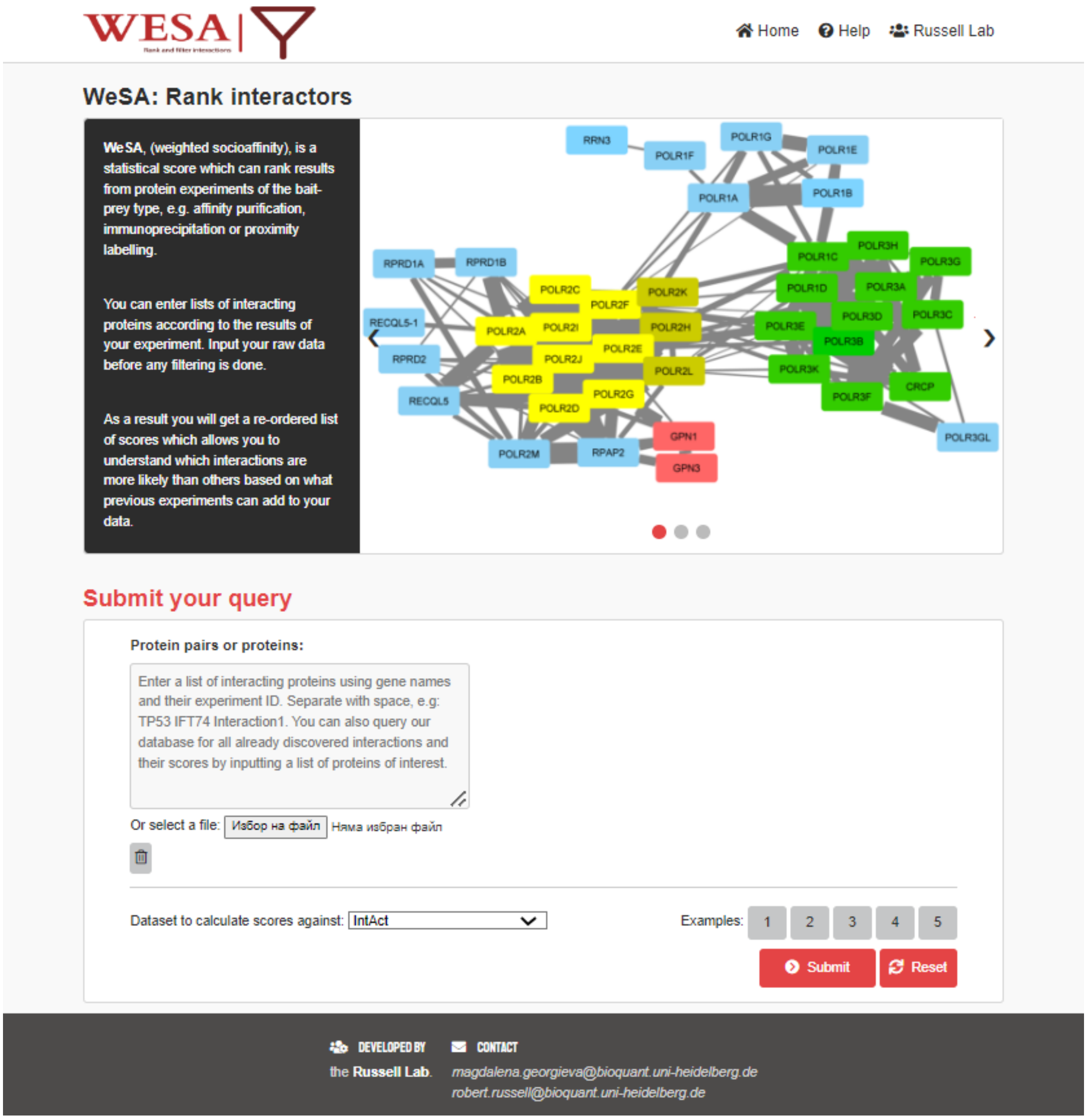
The webtool[3] was built with the Python flask package, which integrates back-end operations in Python with front-end interface creation with CSS and HTML.

The webtool gives the opportunity to experimental researchers to combine their data with the results reported in popular databases. It has two functionalities:

- users can submit own results. The tool will take in the new experimental results, add them to existing data in a selected interaction database and calculate the WeSA scores based on the joint data;
- users can query the existing database by submitting a protein or a list of a few proteins and the tool will return a WeSA-ranked list of their previously identified interactions.

Submission and options Figure 4.3 presents the portion of the homepage related to making queries. When submitting their query, users can paste their data in a submission box (fig. 4.3, option 1) or present a file (option 2). The input should be space-separated. Depending on whether there are 1, 2 or 3 columns in the submission, the

[3] wesa.russelllab.org



WeSA: Rank interactors

WeSA, (weighted socioaffinity), is a statistical score which can rank results from protein experiments of the bait-prey type, e.g. affinity purification, immunoprecipitation or proximity labelling.

You can enter lists of interacting proteins according to the results of your experiment. Input your raw data before any filtering is done.

As a result you will get a re-ordered list of scores which allows you to understand which interactions are more likely than others based on what previous experiments can add to your data.

Submit your query

Protein pairs or proteins:

Enter a list of interacting proteins using gene names and their experiment ID. Separate with space, e.g: TP53 IFT74 Interaction1. You can also query our database for all already discovered interactions and their scores by inputting a list of proteins of interest.

Or select a file: Избор на файл Няма избран файл

Dataset to calculate scores against: IntAct

Examples: 1 2 3 4 5

the Russell Lab.

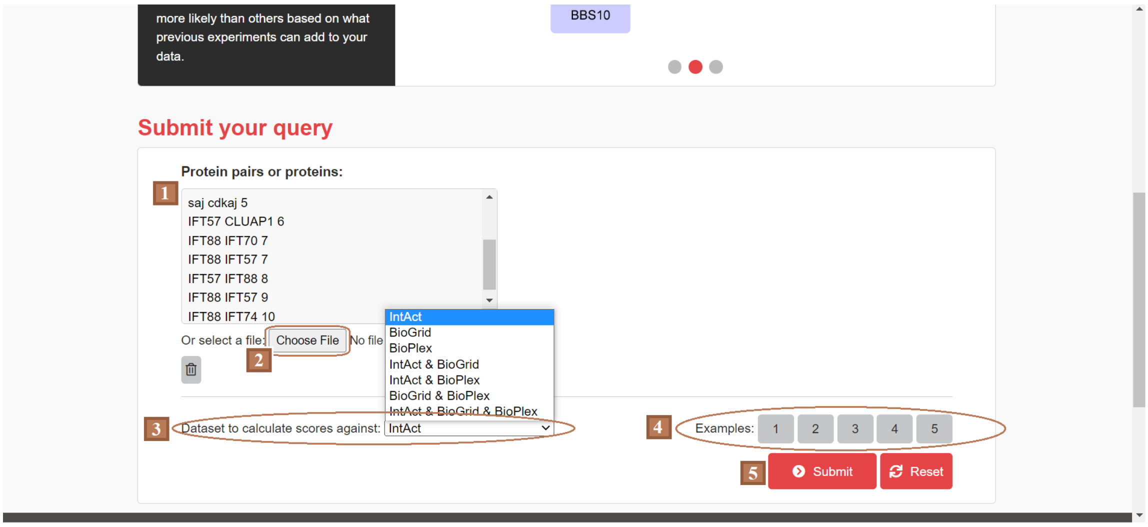
magdalena.georgieva@bioquant.uni-heidelberg.de
robert.russell@bioquant.uni-heidelberg.de

Figure 4.2: Homepage of WeSA webtool.

program continues by calculating the WeSA scores for the new submissions or by returning pre-calculated records.

With the submission, users can choose which repository to use for calculations or querying (option 3). For calculations of updated scores based on user-submitted results, the data from the database they choose will be used as background to enrich the submission and calculate scores based on the combined information. For querying already discovered protein interactions, the results will be displayed only from the dataset chosen. The options are all used databases in our analysis (IntAct, BioGRID, BioPlex) and their various combinations.

Finally, there is an option for users to see different examples (option 4) instead of submitting their query (options 1,2). This can help them understand the submission format through examples as well as to form some expectations about the results.



more likely than others based on what previous experiments can add to your data.

Submit your query

Protein pairs or proteins:

```
saj cdkaj 5
IFT57 CLUAP1 6
IFT88 IFT70 7
IFT88 IFT57 7
IFT57 IFT88 8
IFT88 IFT57 9
IFT88 IFT74 10
```

Or select a file: Choose File No file

Dataset to calculate scores against: IntAct

Figure 4.3: Submission interface and options. Options and buttons which are mentioned in the text are numbered.

When users are ready, they can submit their query and they will be redirected to the output page, where their results are displayed in the form of a table (figure 4.4). The output presents the pairs of proteins together with their calculated WeSA and SA scores and the raw observation numbers from the whole data, separately for each spoke model and the matrix.

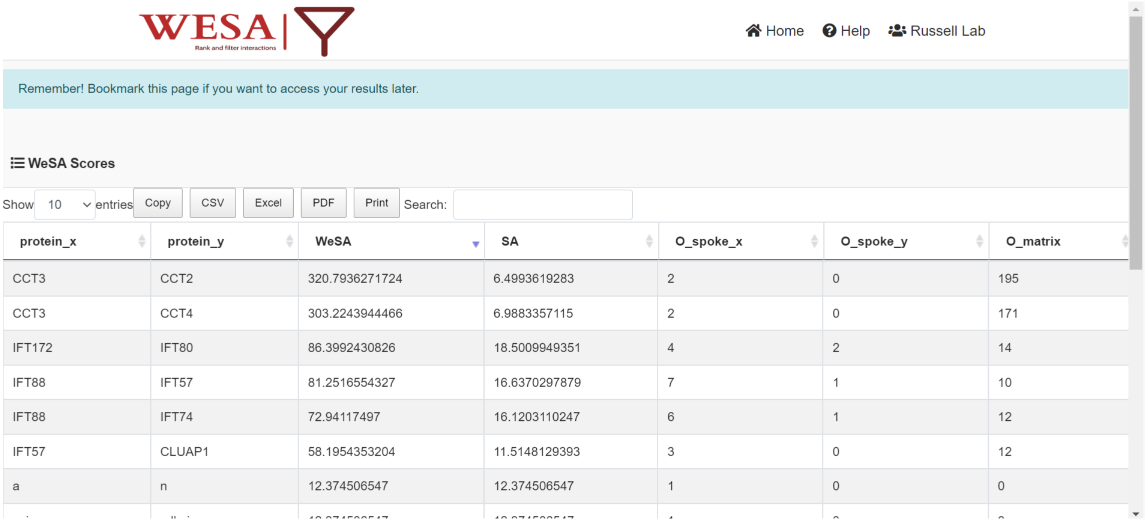
The table can be sorted by any chosen column, so users can focus on the top of the list, proteins with the highest scores. Results are also searchable by writing a (partial) string in the search box. Finally, the results can be printed or exported in a couple of different formats including saved as .csv, .pdf or Excel files, or directly copied and pasted in another file.

Query type 1: Submission of new data for scoring. Submissions of this type have two or three space-separated columns of the form:

bait prey experiment-identifier

The experiment identifier is optional in the case when the results from a single experiment are submitted.

As explained previously, the algorithm then computes some matrix information for this



Remember! Bookmark this page if you want to access your results later.

protein_x	protein_y	WeSA	SA	O_spoke_x	O_spoke_y	O_matrix
CCT3	CCT2	320.7936271724	6.4993619283	2	0	195
CCT3	CCT4	303.2243944466	6.9883357115	2	0	171
IFT172	IFT80	86.3992430826	18.5009949351	4	2	14
IFT88	IFT57	81.2516554327	16.6370297879	7	1	10
IFT88	IFT74	72.94117497	16.1203110247	6	1	12
IFT57	CLUAP1	58.1954353204	11.5148129393	3	0	12
a	n	12.374506547	12.374506547	1	0	0

Figure 4.4: Results page of the webtool.

data before merging it with the chosen background database and completing the WeSA calculations. The returned output contains scores based on the selected dataset updated with the new information. It is helpful to re-rank results from big and noisy experiments and focus on validating the most likely interacting pairs.

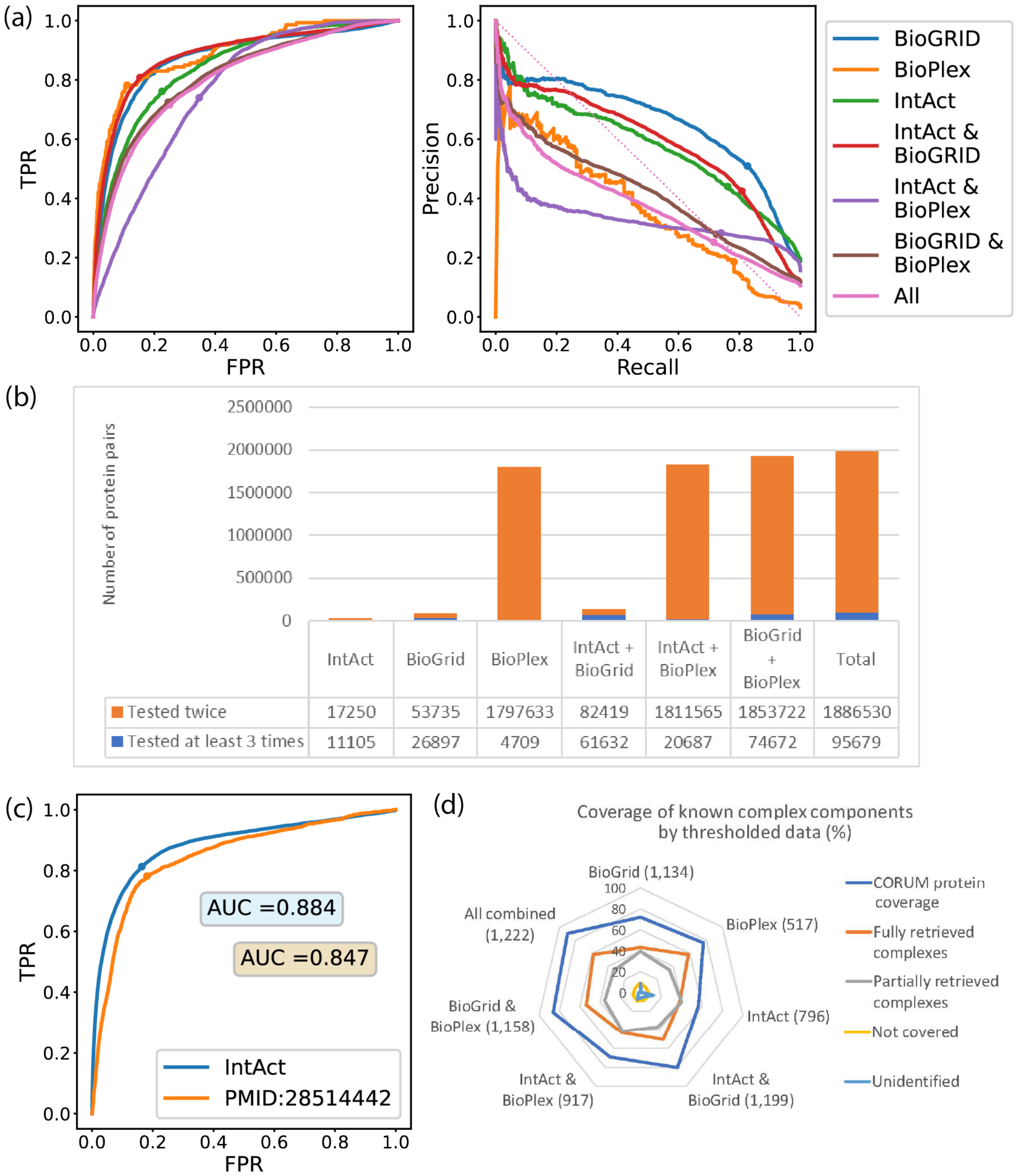
Query type 2: Retrieving all known information about a protein or a protein list. This type of submissions has only one column listing all proteins for which information is desired. The algorithm then searches through the selected pre-recorded dataset which was selected and outputs only pairs in which the queried proteins are present. This is useful if a researcher wants to obtain an unbiased overview of the data which is already published with scores helping them to rank the most likely interactors and focus on them.

4.3 Results

4.3.1 Performance (algorithmic advance), comparisons and predictions

WeSA retrieves many known complexes

As no best way for assessing performance is known, we looked at our results from multiple perspectives. First, since one of the main contributions of WeSA should be its



	IntAct	BioGrid	BioPlex	IntAct + BioGrid	IntAct + BioPlex	BioGrid + BioPlex	Total
■ Tested twice	17250	53735	1797633	82419	1811565	1853722	1886530
■ Tested at least 3 times	11105	26897	4709	61632	20687	74672	95679

Figure 4.5: (a) ROC curves for WeSA with CORUM control on all datasets – BioGRID, Intact, BioPlex, all combined. Only protein pairs observed at least three times are presented. (b) Table and bar chart showing the exact number of observed protein pairs across datasets. (c) ROC curves comparing full IntAct data with the single largest study recorded in IntAct (PMID:28514442). (d) Diagrams showing protein coverage of CORUM and complex coverage overlap between complexes from CORUM and those established using WeSA (as percentages). Complexes are counted as partially covered if at least 50% of the components were identified together.

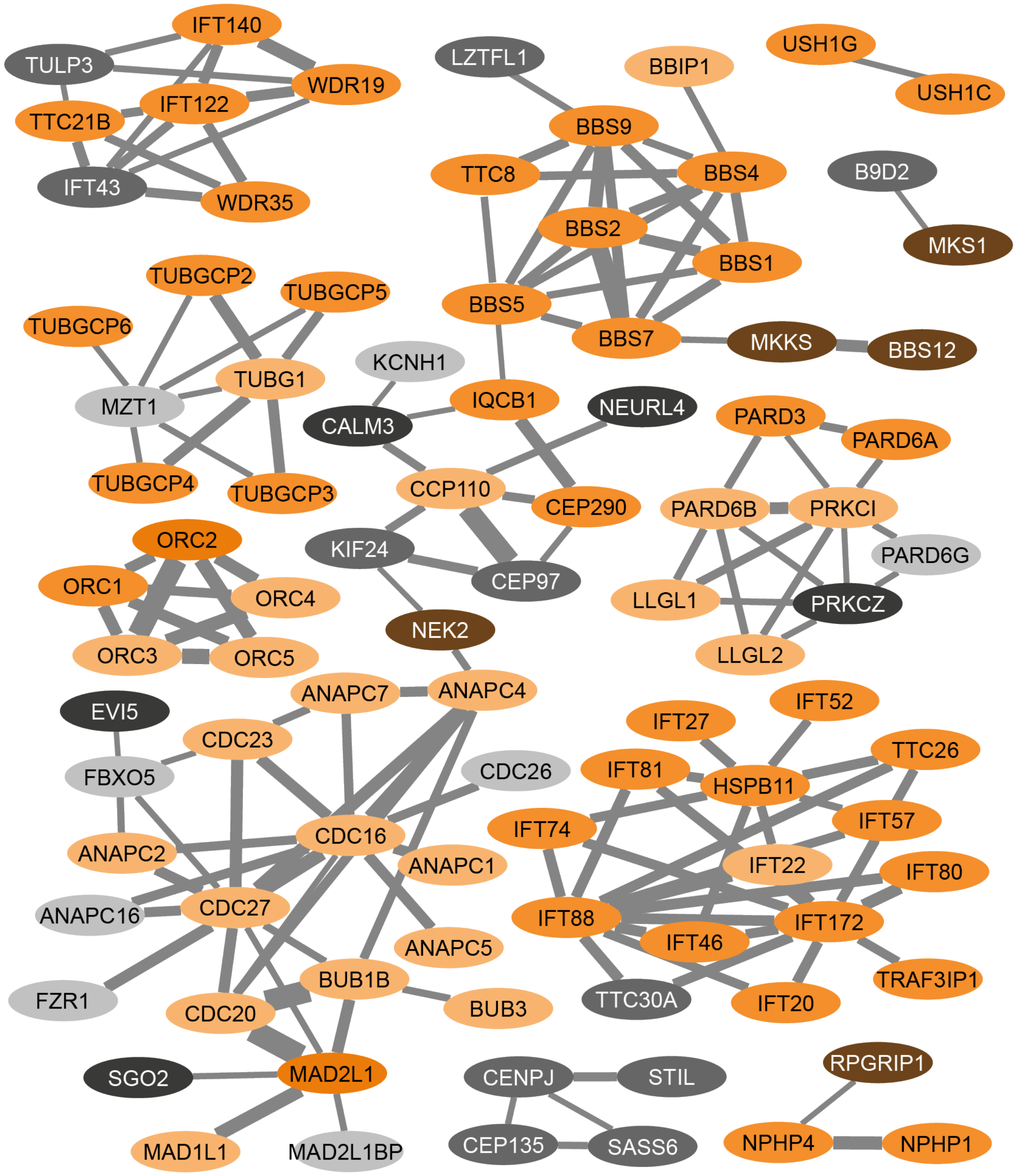


Figure 4.6: Examples of complexes which are retrieved after thresholding the IntAct-scored network.

ability to find complexes, we performed ROC analysis to assess how well it captures complex pairs (figure 4.5a). For the scoring we considered only protein pairs which were observed at least three times (figure 4.5b). We used the CORUM database and expanded complexes as complete graphs. Interactions between pairs of proteins within the same complex are considered true, while the rest are considered false. Since the current CORUM release contains 1,578 complexes and ignores inter-complex links and spontaneous interactions, this control set will miss many possibly true interactions. Thus, it cannot be used as a benchmark for accuracy, but can still be useful, especially in quantifying TPR.

It is expected that CORUM-complex retrieval is best with homogeneous data, such as the BioPlex database. However, we did not have enough data to assess this since in BioPlex baits are not often repeated and we ignore scores of pairs which were only observed once or twice. BioGRID and IntAct, however, show similarly good results while also containing more ‘eligible’ pairs (seen at least 3 times).

Interestingly, the ROC curve for WeSA scores of the combined IntAct and BioGRID data has higher AUC than the individual components, which suggests that the method works despite noise coming from experiment variability. Another confirmation of that claim is the presented figure 4.5c, which contrasts the results from a single study to the results from all data in IntAct. While the difference is small, the ROC analysis again favours the non-homogeneous but richer dataset.

These results are also associated with a high CORUM protein coverage (figure 4.5d). The proteins from each dataset which are also recorded in some complex on CORUM are normally more than two thirds, except for IntAct, which has 57% coverage.

This high coverage 4.5d justifies and reinforces the results given by the ROC curves (figure 4.5a). The former shows that most complex overlaps are at least partially covered. Specifically, the combined data from all three datasets have just above 90% overlap with CORUM proteins and in addition, out of the 1,222 WeSA defined complexes, more than 90% overlap with already confirmed CORUM data at least partially; more than half of those are retrieved in full. There are some complexes which are unexpected and rarely (just 7 cases) have both proteins unidentified in any complex.

Some examples of complexes defined by WeSA scoring, clustering and thresholding are presented in figure 4.6. The examples focus on ciliary proteins and the components they participate in. Nodes coloured in orange show a complex confirmed in CORUM, while grey nodes are not yet identified in any CORUM complex. This presents an opportunity for speculation and prediction which will be explored later.

A single experiment can be scored resulting in a ranked list of potential interactions

For inherently noisy experiments such as frameworks similar to the BioPlex study [37] or single-wash method like FLAG-tag affinity purification [10], a control is normally used to clean contaminants and unspecific binding. However, in addition to doubling the work, a good control protein can be hard to identify, especially given our incomplete knowledge of protein function.

Given the finding from the previous section, that combining different experiments together improves scores, then WeSA can provide an efficient alternative to the aforementioned control-reliant method. The benefits are two-fold. First, if a researcher does multiple replicate experiments with the same bait, they already have a lot of evidence which should add to what is already there and, in some novel cases, should be properly handled in an unbiased way. The second advantage is provided by the fact that if one experiment contains a lot of noise due to its design and aims, there is no good way to filter that noise out with a single control, but that noise may be neutralised from the joint information of all previous data.

Since it does not rely on any additional experiments, provided there exists a large enough and suitable dataset to combine with, a single experiment can be scored based solely on the raw experimental data that is already out there. To provide further evidence to this statement, we test the performance of WeSA on the full dataset and compare that to the single largest study recorded in IntAct (figure 4.5c). While one might expect that scores from the latter are better, given the homogeneity of the method for it, they are very similar to the full dataset even outperforming the single study. Thus, the ROC curves suggest that the availability of data, rather than its homogeneity improve predictions more. The WeSA score seems to improve robustness, so data from multiple studies can be combined effectively. This result opens up the possibility to add a single experiment to the already existing data and scoring it in this context, which should give experimental researchers a better tool to filter out noise from their observations.

WeSA retrieves more within-complex interactions than previous socio-affinity-like methods or experimental methods alone

We start by comparing our results to current experimental confidence metrics to show that WeSA manages to re-arrange the list of interactions and enrich the top (high WeSA scores) in true positives better. For this comparison we used confidence scores as recorded in IntAct (figure 4.7 and 4.8, top left panels).

We also confirm previous hypothesis [30, 12, 45] that integrating spoke and matrix information improves filtering predictions. The results can be found in the appendix, in figure 6.1. Despite this conclusion and while establishing that data on co-occurrence in the matrix is beneficial, even without that scores can be calculated.

One of the aims of WeSA is to provide a fast way to analyse data from one or more experiments based on all other data which is available. By using this information, further analysis can be focused to a narrower list of top-scoring protein pairs. The ratio observed-to-expected applied to recent and past evidence, can reduce the need for controls or replicates, but data on replicates can be utilised well as every observation adds further evidence to make WeSA scores more accurate.

This goal sets apart the method from other statistical techniques which are performed on a single dataset (e.g. [22, 17]). Moreover, the idea of only using information from the experimental precipitate avoids further biases arising from exponentially scaling inaccuracy with the addition of further complexity (e.g. [18]). These methods which rely on well curated details, can form a part of further analysis, while WeSA can save time by shortening the list of potential interactions and overcoming biases such as dependencies between the target and control or between the target and the CRAPome [50].

	≥ 3 observations					
	WeSA			SA		
Dataset	AUC	TPR	FPR	AUC	TPR	FPR
IntAct	0.842	0.762	0.226	0.825	0.777	0.231
BioGRID	0.87	0.826	0.195	0.812	0.764	0.269
BioPlex	0.89	0.782	0.11	0.834	0.735	0.235
IntAct + BioGRID	0.883	0.808	0.153	0.837	0.758	0.222
IntAct + BioPlex	0.762	0.74	0.347	0.793	0.717	0.285
BioGRID + BioPlex	0.81	0.725	0.246	0.729	0.614	0.296
All	0.8	0.716	0.25	0.729	0.622	0.301

Table 4.1: Statistics from the ROC analysis displayed in figure 4.7. For each dataset the AUC is presented for analysis based on WeSA and SA scores. Alongside that are displayed the TP and FP rates with the best threshold at each case.

For these reasons, in figure 4.7 and its adjacent table 4.1, as well as in figure 4.8, we compare the performance of WeSA against the same-purpose method of socio-affinity [30]. In 5 out of 7 cases, WeSA performed better than SA; in two cases, their perfor-

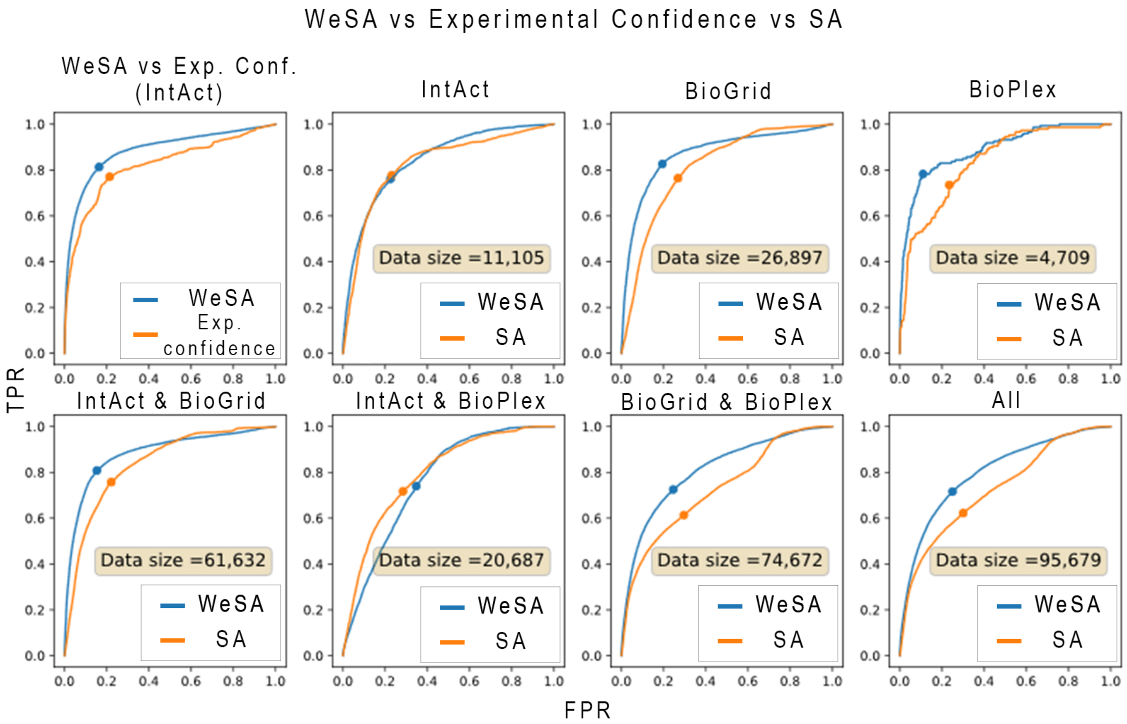


Figure 4.7: ROC curves of WeSA vs experimental confidence (top left); rest: WeSA vs SA for different datasets. Minimum number of baited observations is set to 3.
Full statistics about the ROC parameters in this figure are presented in table 4.1.

mance is comparable. If we do not impose the at-least-3-observations rule, optimal TPR and FPR for WeSA are much better in all cases (figure 4.8). However, accuracy in those cases is very poor.

Validation from pairwise experiments

As we have discussed earlier, there is no single benchmark for prediction performance, so in the following sections, we verify WeSA scores by comparing them to other metrics which can be expected to be related. As a start, in this part we look at validation from pairwise studies. We show two perspectives: one for positives and one for hypothesised negatives.

As a benchmark, we used the interactions discovered through pairwise experiments deposited in IntAct. The total number of pairs in this set is 47,846.

For each dataset we consider two groups. In group one (blue in figure 4.9) are the pairs of proteins which overlap with the benchmark set. In the second group (orange in the figure) are all pairs. The distribution of WeSA scores in the two groups across

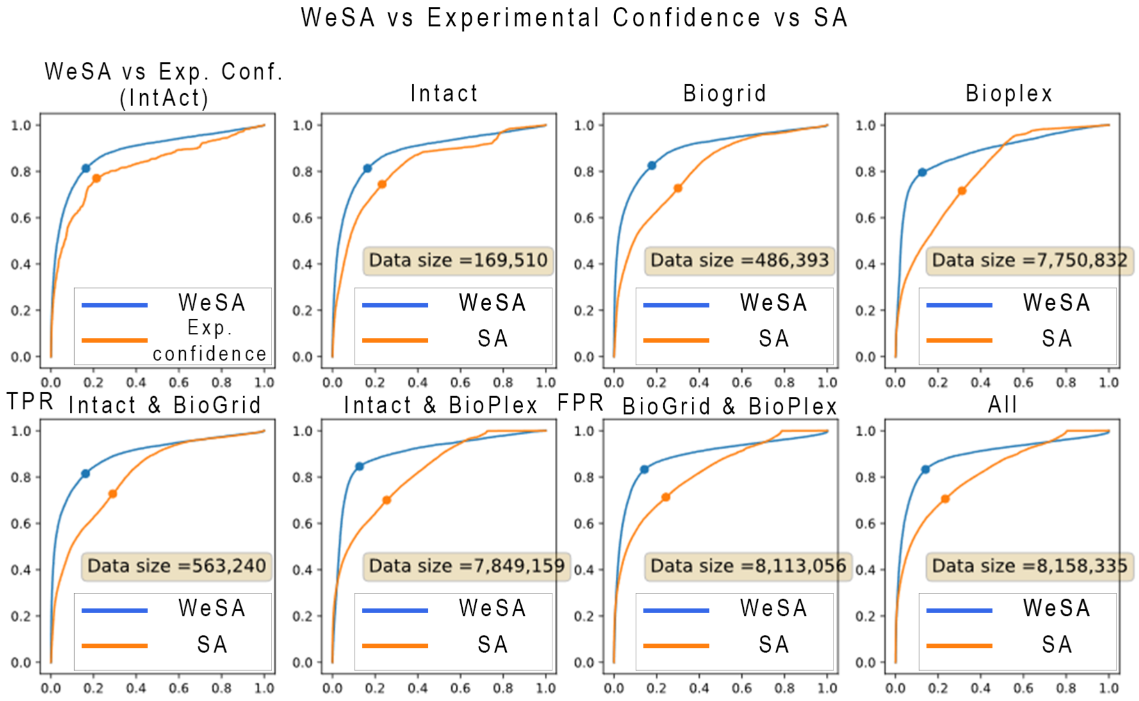


Figure 4.8: ROC curves of WeSA vs experimental confidence (top left); rest: WeSA vs SA for different datasets. Minimum number of baited is set to 1.

datasets is presented in the boxplots in figure 4.9. The mean unfiltered WeSA scores are significantly higher for group one than for the total distribution averages. This is quantified and confirmed by a MWU with results shown in the figure.

If we apply the best threshold to the WeSA scored pairs, we find just under 70% coverage of the interactions confidently (experimental confidence > 0.8) observed in pairwise studies. We do not expect a complete match since pairwise studies sometimes test unnaturally occurring protein connections, “losers” in competition for binding or connections separated from the real world in some other way. However, there is still positive verification of some links from pairwise studies. That can be a reason to believe that those pairs of proteins for which WeSA scores are high, but do not fall in a CORUM complex can be the bridges between complexes.

3D structures

Another way to look at validation is by comparing prediction WeSA scores to actual 3-dimensional protein interaction structures. There are two ways to do this which are discussed here: validation using data from the Protein Data Bank (PDB) [9] and pre-

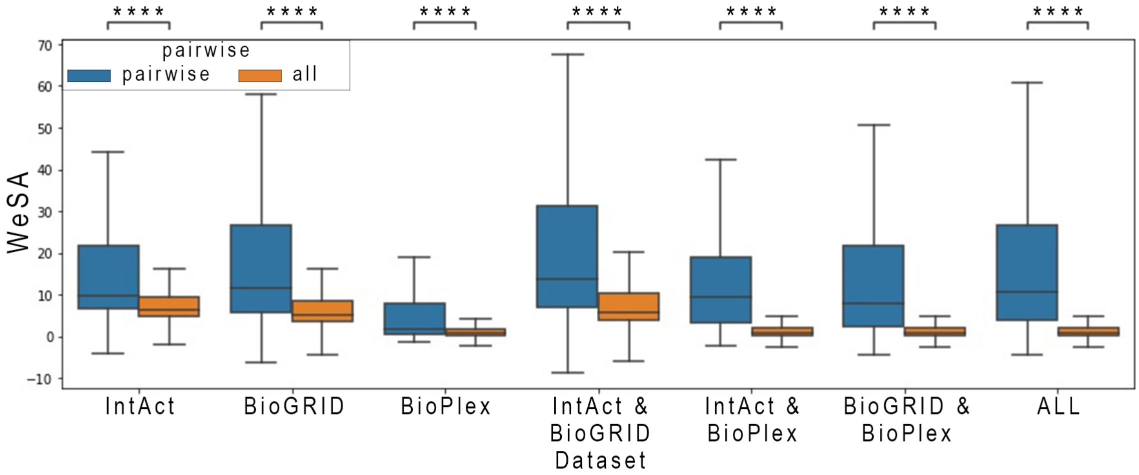


Figure 4.9: Scores from pairwise experiments vs total WeSA score distributions by subset. Comparison by MWU test where **** denote very high significance ($p \leq 0.0001$).

dicted interactions with AF-Multimer (AF-M) [26].

Structures of interacting proteins deposited in the Protein Data Bank (PDB) [9] provide observation of the mechanism of interaction in great detail. Their precision allows for validation for the WeSA filtering method. Examining the structures from the most recent PDB update (retrieved: 24.10.2022) we have obtained the size of the interacting surface (in number of residues) for 21,497 protein pairs. In addition, we have used the full Negatome dataset [11] containing 6,532 protein pairs which are suggested not to interact. We then compared coverage and looked at the WeSA scores for those pairs.

We define four groups of interactions based on the size of the interaction interface: small, containing links between proteins with interface of 1 or 2 amino-acids; medium of size between 3 and 20 residues and big (larger than 20 residues). In addition to those, we defined a fourth category ‘none’ for proteins from the Negatome which do not have an interacting interface.

The distribution of scores in these 4 groups and the comparison with overall scores is presented in figure 4.10. It is clear, that the complexes are enriched in high scoring pairs. Furthermore, testing the score distributions within the groups (with MWU test) highlights the consistently higher WeSA scores across datasets for interactions with larger interfaces. Specifically, in every dataset, groups of pairs with big interfaces have very significantly higher scores compared to groups of pairs from the Negatome which do not interact directly.

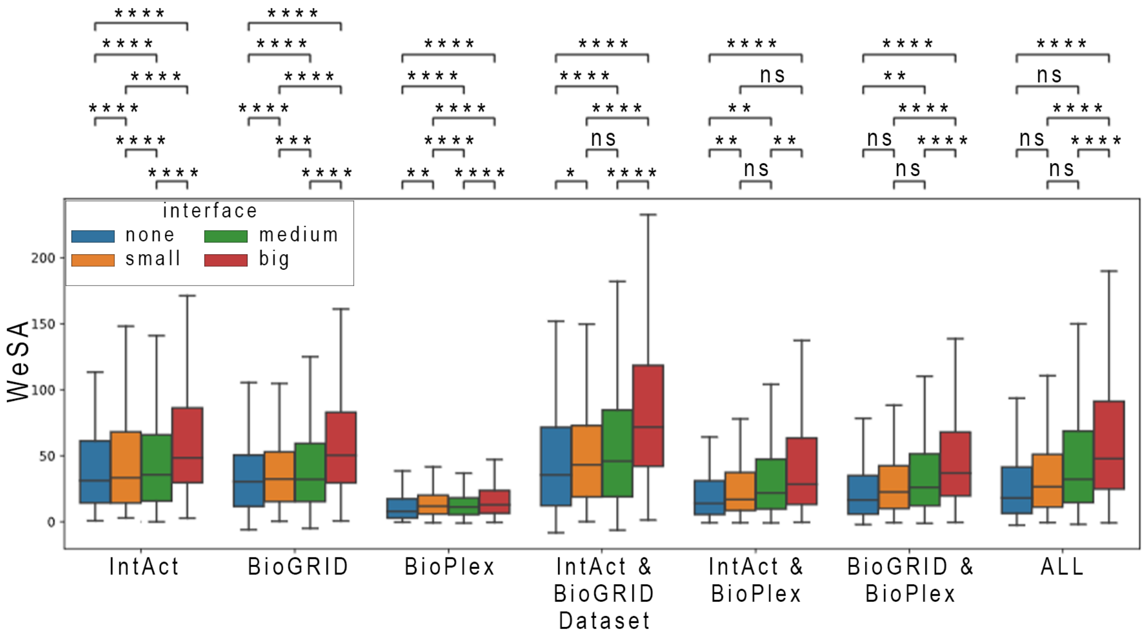


Figure 4.10: Boxplots of WeSA scores in the four interface size categories (big > 20 interacting residues; 20 ≥ medium ≥ 3; small < 3; none = 0 residues). Plots are presented and tested separately for each dataset. Significance annotations based on Mann-Whitney test. p-value annotation legend: ns: $p > 0.05$; *: $0.05 > p \geq 0.01$; **: $0.01 > p \geq 0.001$; ***: $0.001 > p \geq 0.0001$; ****: 0.0001 > p

In addition, the advance in predictive complex models brought the creation of AF [FR23] followed by the AF-M modification [26]. On the basis of iterations over interacting neural networks, these models predict the structure of individual proteins or their interactions, respectively.

A case study of 3D structures involving ciliary proteins

Since AF-M is a prediction algorithm, it still requires experimental validation to verify interactions. However, the high performance of the AF algorithm, as mentioned in section 2, lends credibility to the method and can be used as another verification.

Here, we present a case study based on our collaboration with the Lorentzen lab (Aarhus University) on the structural model of the IFT-B complex [55]. The paper validates the position of 15 proteins and their interactions within the IFT-B complex: IFT81, IFT74, IFT46, IFT52, IFT88, TTC30A, IFT20, TRAF3IP1, CLUAP1, IFT80, IFT172, IFT57, IFT22, HSPB1, IFT27. This complex also largely overlaps with our findings from scoring and clustering in figure 4.6.

In more detail, first, let us look at the scores of the IFT-B proteins. A query in the webtool to retrieve all WeSA scored interactions from IntAct for the key interactors[4] retrieves 281 pairs of proteins. This includes all observations including those with low scores. The top 85 scoring pairs, however, all include proteins only from the IFT-B complex. Moreover, out of the 34 direct interactions, 30 pairs are within those top-scoring. Missing are only the direct interactions between CLUAP1 and IFT52, IFT20 and TRAF3IP1 as well as the pair IFT20-TRAF3IP1. To investigate what causes the omissions, we look at the raw counts. After investigating those more closely, it looks as if CLUAP1 has not been used very often as a bait (only 9 times in BioGRID) and is acting unexpectedly as such (capturing seemingly random proteins). This may be expected given its long α-helical structure which possibly creates problems for its independent purification and stabilisation.

Apart from validating the scores of the full IFT-B complex, this work also highlighted the predictive power of WeSA. While at the beginning of the collaboration, the structural biologists in Aarhus were unaware of the direct interaction between IFT80 and IFT172, the two proteins had the highest WeSA score among their respective interactors. This was largely due to results from affinity purification experiments including observations in the matrix which were not validated or completely lost in the format of the recording databases. However, the high score is what prompted us to design a validating experiment.

The workflow we followed, which can be applied in general for validation is: applying AF as the first validation step and then validating experimentally. For the former, we ran 5 AF models out of which all confirmed a large interaction interface between the two proteins. In the biochemical validation step, I worked together with Nevin Zacharia from the Lorentzen group to perform a pull-down experiment of the two purified proteins. The full details of the experiment are detailed in the reference paper.

As outlined here, the validation procedure using AF cannot be used on its own but has to be combined with experimental validation. However, it is mentioned here as a possibility given the high accuracy of AF and the otherwise low availability of real experimental structural information.

[4] The query is for: IFT81, IFT74, IFT52, IFT88, TTC30A, IFT20, TRAF3IP1, CLUAP1, IFT172

4.3.2 Biological relevance – network and examples

For illustration, we generated a network of WeSA results. The network resulting from the combination of BioGRID and IntAct records is produced and edge weights corresponding to the computed WeSA scores are assigned. An optimal threshold is applied to clean low-scoring edges.

The network is clustered using MCL, which groups nodes according to their connections and WeSA scores. We analysed the combined IntAct and BioGRID data and for it, the threshold corresponding to the ROC optimisation revealed 1,187 clusters, covering 76.3% of the proteins on CORUM. Some of the resulting clusters are presented in figures 4.11-4.13.

It is worth mentioning that clustering allows us to group proteins which are more tightly linked together even if they have high-scoring interactors outside the cluster. For example, a chaperonin complex has a specific role in the folding of a few specific proteins. This can be a cause for strongly scoring interactions with those few proteins. However, the links within the complex are much stronger. Thus, the clustering algorithm would likely be able to separate the chaperonin complex from the proteins it is helping despite their possible high individual connections.

Above the optimal threshold, we see clear separation of some already known individual complexes such as IFT-B, IFTA-TULP3, BBSome (with additional BBS10 and LZTFL1), Multisynthetase complex (figure 4.11a-4.11d). As is the case with the BBSome complex (figure 4.11c), there are some complexes which cluster with protein attachments or other complexes. However, even if complexes have overlaps, they can still be jointly identified and cleaned from additional contaminants or non-specific interactions. For instance, after thresholding and clustering, the RNA polymerase II and III complexes are separated from the other proteins which overcrowd the interaction network such as UXT, ATF4 or the MED complex (figure 4.11e). All 15 Polymerase III and 12 Polymerase II (Pol II) proteins are retrieved (components identified by Ramsay *et al.* [57] and Kershnar *et al.* [43]), even though they share 3 proteins. Moreover, both complexes contain strong links to the shared POLR2H, POLR2L and POLR2K.

In addition, the network is cleaner: the number of edges between Pol II/III and other proteins is halved (from 229 to 115) and the number of immediate neighbour nodes is reduced from 52 to 17. Yet, the resulting network contains unannotated links, which comprises a possibility for predictions. For instance, the Polymerase cluster includes some proteins which are not recorded as parts of the complex in CORUM. Further

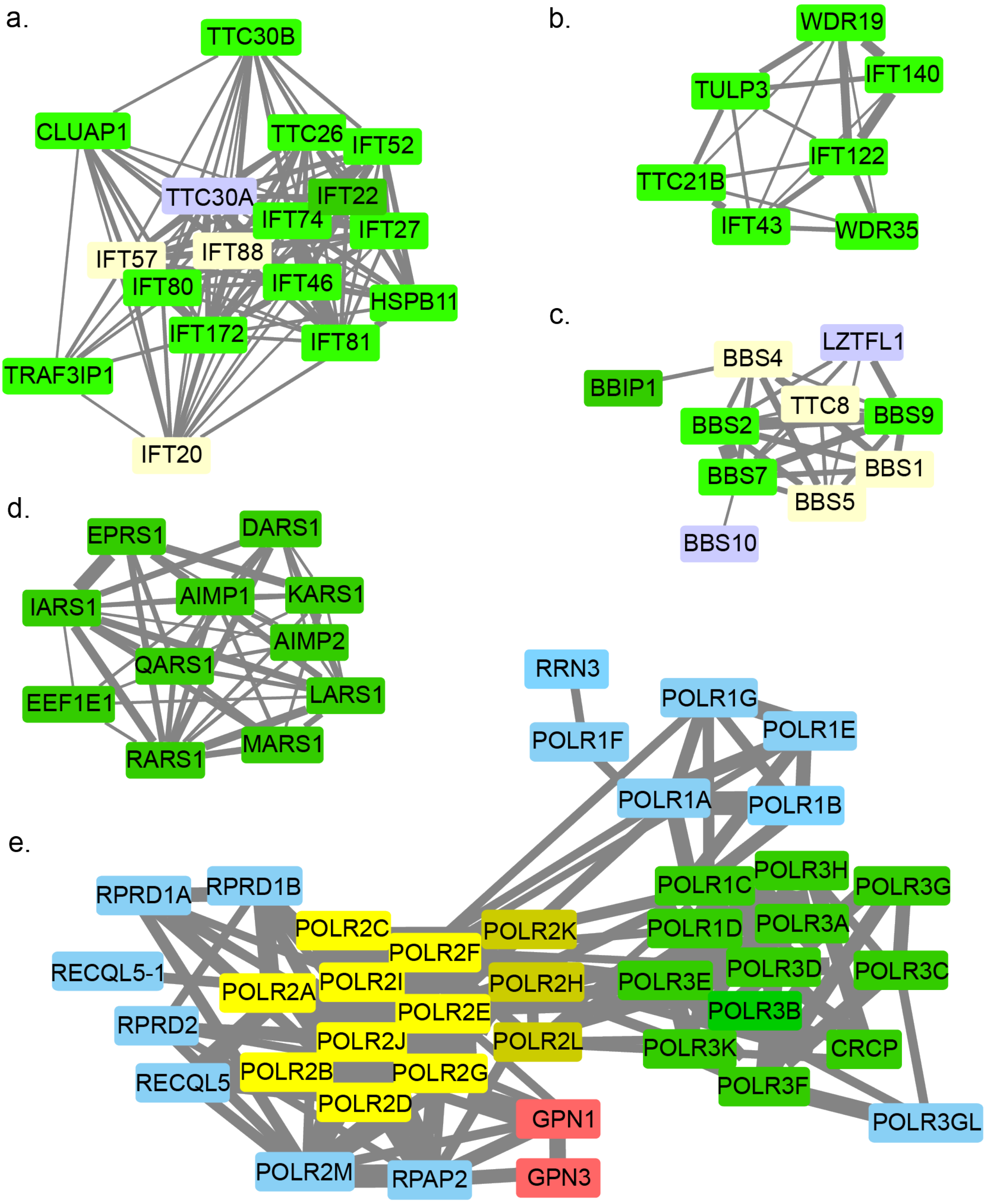


Figure 4.11: Network clusters after computing WeSA scores for the combined IntAct-BioGRID data, filtering out low-scoring interactions and clustering. a. Cluster including the IFTB complex. It does not contain any additional components. b. Cluster including the IFTA complex. TULP3 is also present. c. Cluster including the BBSome complex. There are the two additional proteins BBS10 and LZTFL1. d. Cluster including the Multisynthetase complex. e. RNA polymerase II core complex (yellow) and RNA polymerase III complex (green). Shared proteins are coloured lime-green.

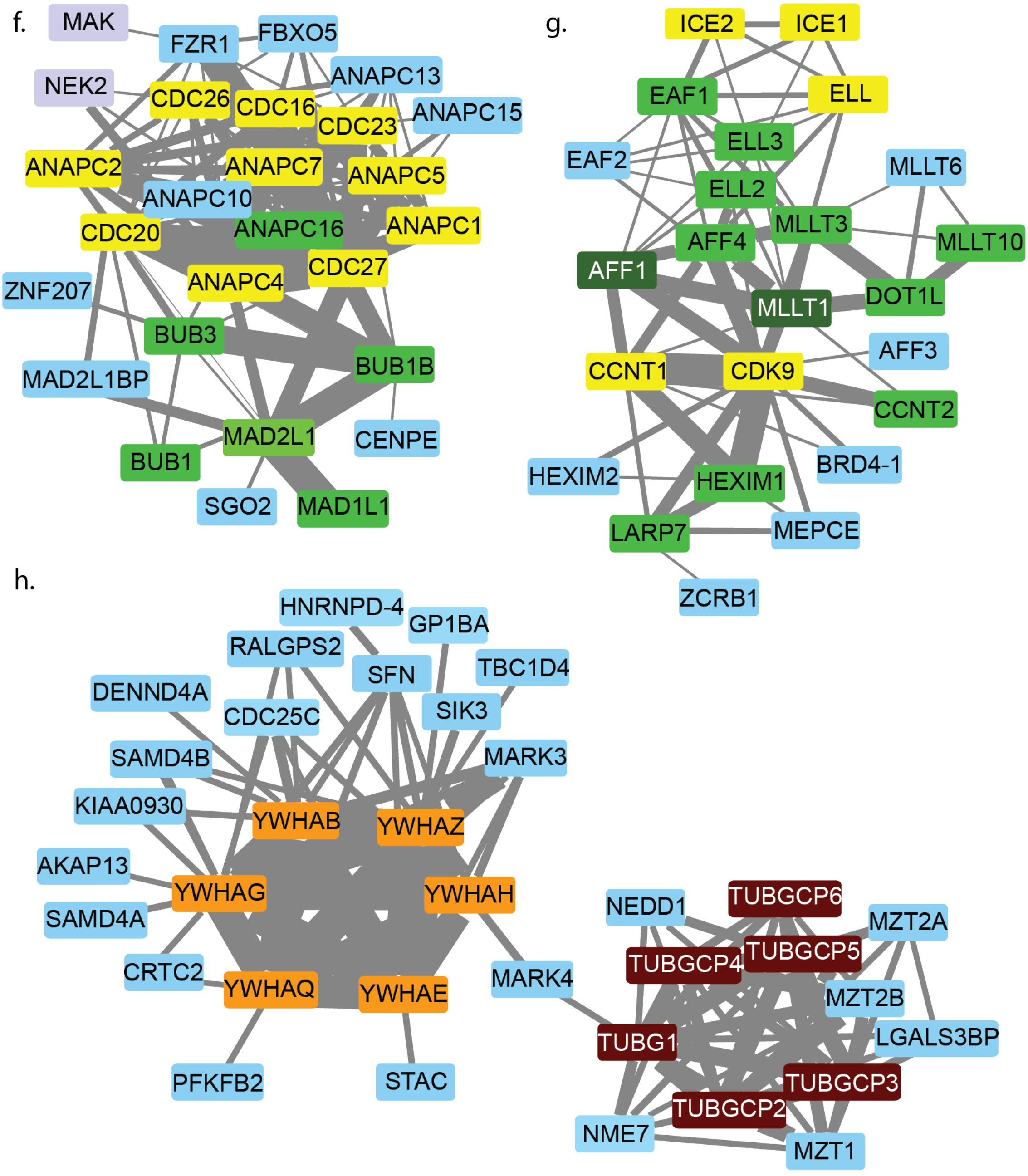


Figure 4.12: f. Anaphase promoting complex (APC/C) complex (yellow); APC/C-ANAPC16 complex; CDC20-MAD2 complex; Mitotic checkpoint complex; BUB1-BUB3 complex; MAD1L1-MAD2L1; CENPE, SGO2, MAD2L1BP, ZNF207. g. HEXIM1, AFF4, DOT1L. h. Gamma-tubulin complex (brown) weakly connected through MARK4 to part of the Kinase maturation complex (orange).

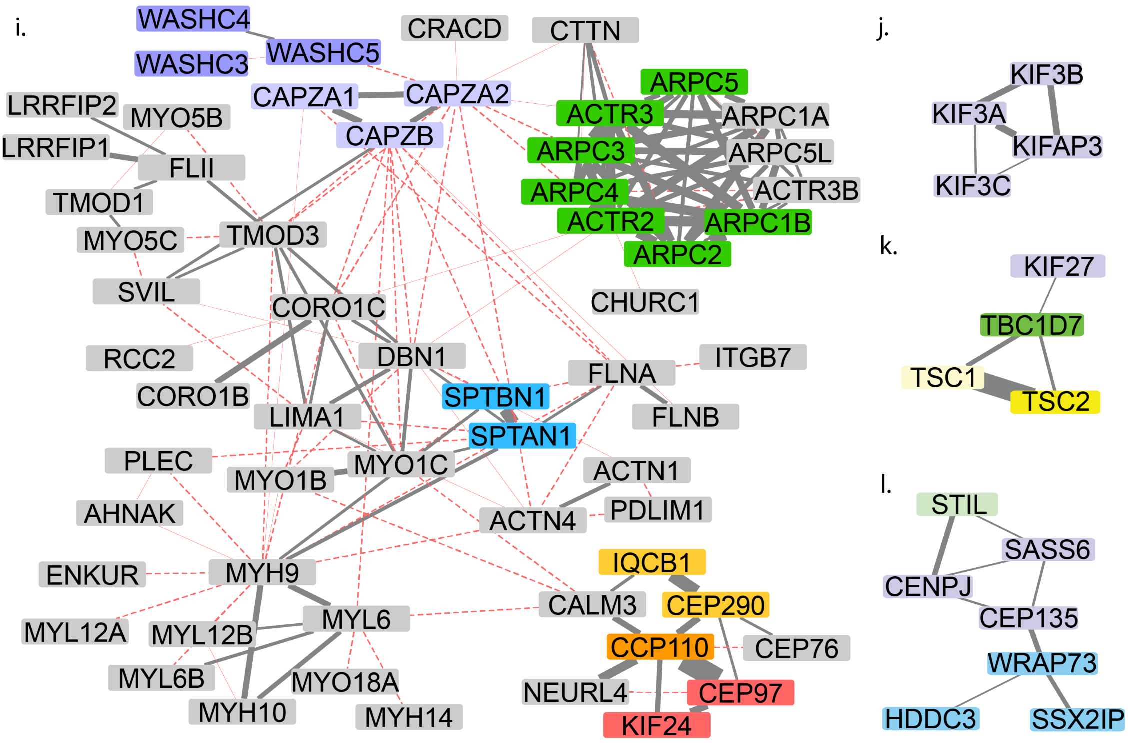


Figure 4.13: i. A 61-protein cluster containing CAPZalpha/beta (light purple), partial WASH complex (purple), SPTAN1-SPTBN1 heterodimer (blue), the full Arp 2/3 complex (green), mechanisms for ciliogenesis (yellow-orange) and its negative regulation (red-orange). Thin red edges correspond to WeSA scores <41; dashed red lines indicate scores between 41 and 49. Gray nodes correspond to nodes for which there is no complex information. j-l. Components which are not annotated in CORUM but can be potentially real, especially given the participating genes have been previously annotated as ciliary [12]; only TSC2 and the blue nodes are not known ciliary genes, but maybe they have a ciliary connection.

investigation into those proteins, however, provides evidence in support of the additional interactions: isoform 1 of POLR2M (also known as Gdown1) is recognised as a Pol II protein enhancing Mediator regulation [35], while a tandem affinity purification study confirms RPAP1 attachment to the complex [39], a protein which recruits RPAP2 [FR29].

Among the attachments to Pol III we observe also RRN3 and POLR1F, components of the Polymerase I complex. Search for the other Pol I proteins in the network reveals high WeSA scores within the complexes SL1/TIF-IB and eIF-3 and missing links to the other subunits. Thus, WeSA manages to put the proteins in their respective 3 complexes.

While the optimal threshold gives a good indication of the protein complexes, varying this threshold can illuminate other details. Reducing it can show more transient interactions, whereas increasing it will leave even more likely complexes. Figure 4.13i shows how increasing the threshold can separate a big component into smaller pieces and further remove possible contaminants. WeSA scores agree with the observation that a weak WASH-CAPZalpha/beta complex [40] can form between the two smaller complexes, but those subunits are separated if we look at WeSA scores bigger than 49.

There are still many unknown complexes and proteins whose role is unknown. We can try to understand some of them through this clustered network. Figures 4.13j-l show a few components which may form complexes. This is supported by the fact that most proteins within the complexes were previously recorded as either ciliary gold-standard or likely candidates [12].

4.4 Discussion

As observed in this section, socio-affinity methods are efficient as they take advantage of all previously known information. Compared to the statistical method presented in the previous chapter, this approach does not need controls which reduces significantly the resources required. Compared to other metrics for protein-protein interaction identification, it is unbiased and does not require any further information about the interactions, the study's accuracy or its positioning relative to other studies.

Our tests show good performance in capturing true positives and pronounced difference between scores of unconfirmed and confirmed interactions through other methods such as pairwise experiments and observed interaction structures.

By adding the weights, WeSA improves greatly on SA. This is the case, especially for observations of low-counts, which until now were only intuitively perceived as less confident.

The method is still limited in accuracy and especially the small counts naturally contribute to bigger variation. However, doing WeSA analysis as a first filtering step can narrow down the space which needs to be explored and confirmed with bigger certainty.

We have also found that the metric is robust in adding more observations even if they do not come from the same experimental techniques. This fact allows us to combine datasets or update them with new data. The significance of it is for the work of researchers who can now make use of previous knowledge and analyse the results of their experiments in the context of past experiments. The scoring is now accessible to everyone to make use of in their analysis through our web tool.

The weights in the WeSA terms have been chosen as the raw counts. Due to that, WeSA can be confused with the Pearson's chi-squared statistic which would have a χ^2 distribution and indeed, its robust form was used for inspiration of the method. Specifically, for a sample space Ω with events i, such that $\cup_{i\in\Omega} i = \Omega$ and $i \cap j = \emptyset$ for $i, j \in \Omega$, the Pearson statistic is given as:

$$P = \sum_{i\in\Omega} \frac{(O_i - E_i)^2}{E_i} \approx 2 \sum_{i\in\Omega} O_i \log \frac{O_i}{E_i}$$

The last approximation comes from the Taylor expansion of the log terms on the right hand side together with the fact that $\sum\limits_{i\in\Omega} O_i = \sum\limits_{i\in\Omega} E_i$.

The difference between this statistic and WeSA is the matrix term, which does not complete the sample space together with the spoke terms. Since WeSA terms do not span a sample space, the equality $\sum\limits_{i\in\Omega} O_i = \sum\limits_{i\in\Omega} E_i$ does not hold and the approximation to a sum of squared normal variables is not possible. Yet, we have found that this form is useful to work with our biological data showing superior performance to other tested weights, precisely, comparisons to no-weight, logged-weight and square-rooted weight.

Further directions We have previously explored how experimental confidence scores perform against WeSA scores. These are only given in IntAct, but we can attempt to create equivalent measures from other databases and experiments. For example, mass spectrometry results may include intensities or LFQ values, quantifying the amount of protein captured. They might provide some insight on frequency of interactions. However, one should be mindful of the input as, for example, cell lysate has widely different protein concentrations and it is expected that the results will accordingly have different abundances. While seemingly impossible to put in an idealised mathematical model, if initial concentrations are normalised, this can give rise to some further exploration.

Experimental confidence scores for protein pairs can further be integrated into WeSA by substituting binary experimental results with probabilities as mentioned in the methods in section 4.2.1. We briefly explore this here by making confidence-adjustments using the IntAct dataset and discuss ways of expansion to sets such as BioPlex where confidence is not directly recorded. The procedure uses the input of the experimental confidence to weigh protein pairs before inputting them in the WeSA computation. The results are presented in figure 4.14.

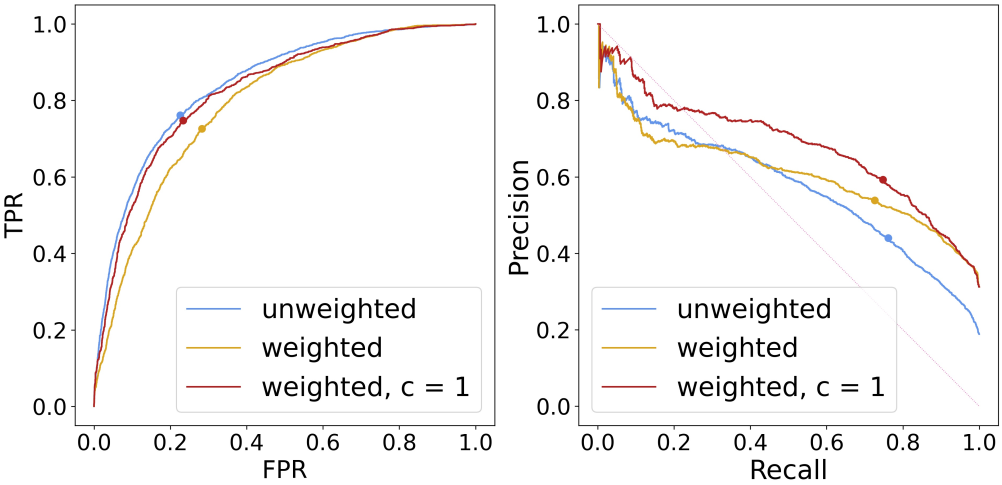


Figure 4.14: In this figure ‘weighting’ refers to weighting the observation counts by the experimental confidence. The panels present ROCs (left) and Precision-Recall curves (right) resulting from scoring the IntAct data in three ways: 1) ordinary way as presented by equations 4.1 and 4.2 (without weighting based on experimental confidence or pseudocounts as from equation 4.7); 2) WeSA scores incorporating experimental confidence but no pseudocounts; 3) WeSA scores incorporating both experimental confidence and pseudocounts $c = 1$.

In IntAct experimental confidence is part of the records, but its calculation depends on the individual approaches. It may be that this heterogeneity of IntAct confidence data caused confidence-weighted scores to perform worse than binary records in terms of predicting CORUM complexes. Below, an improvement due to pseudocounts will be discussed.

As a further extension, for datasets such as BioPlex, where experimental confidence may be related to observation quantities, we may want to use that information. In BioPlex raw data on peptide spectral measurements (PSMs) is reported and one may attempt to approximate confidence by inferring probabilities from the PSMs. Specifically, this can be achieved by PSM normalisation followed by finding the corresponding cumulative distribution.

While it looks like raw data with no confidence assignment is more successful in integrating different data, there may be situations where such additional weighting can be useful, even if, for example, only to focus attention on same-method or same-tissue experiments. Moreover, the PR curve with integrated experimental confidence information shows better performance despite the worse ROC curve. If there is a reason to believe that confidence scores should be taken into account, it is suggested to incorporate also pseudocounts (c) [23].

As observed in figure 4.14, a standard addition of pseudocounts of 1, improve the experimental-confidence-weighted WeSA performance. The pseudocounts are integrated in the WeSA formula via the modification:

$$W_{ij} = n_{ij} \times \log \frac{O_{ij} + c}{E_{ij} + c} + n_{ji} \times \log \frac{O_{ji} + c}{E_{ji} + c} + n_{ij}^{M} \times \log \frac{O_{ij}^{M} + c}{E_{ij}^{M} + c} \tag{4.7}$$

Pseudocounts reduce the effects of random noise and data variation, so they have a positive effect in the integration of the heterogeneous information, particularly when it is further enriched with different confidence measures. The pseudocounts are particularly effective to decrease scores for rare observations of abundant proteins. We have tested several possible pseudocount values close to the medians of the observed and expected numbers (results are presented in figure 4.15 and in the appendix figure 6.2). Note that $c = 0$ is equivalent to the ordinary WeSA formula without modifications. Both the Laplace's rule of $c = 1$ and the addition of $c = 0.5$ [FR32] fall within the tested range and are observed to improve scores. With the ROC analysis, there seems to only be negligible benefit to using a specific c, but this can be explored further.

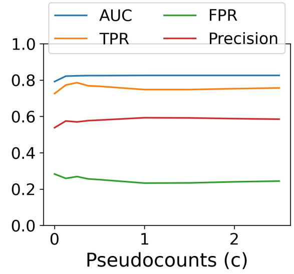


Figure 4.15: Results from ROC analysis of WeSA with varying pseudocounts $c \in$ {0, 0.1, 0.2, 0.3, 0.4, 0.5, 1, 1.5, 2, 2.5}. WeSA scores are weighted by the experimental confidence. Data are taken from IntAct.

The numbers in the input are not unbiased; specifically, IntAct confidence scores are, as mentioned previously, truncated around 0.3 and, moreover, exhibit some bias around 0.4. This can be one reason for a poorer performance of the weighting. It can also point to intrinsic incompatibility of experiments with vastly different mean prey numbers and the fact that they should not be over-analysed together. Instead, the simplest form of WeSA may be better at managing and overcoming biases which arise with complexity.

Further, we propose using WeSA to identify changes in the interactome resulting from mutations whenever such experiments are performed. Decreases in WeSA scores, especially below an optimal threshold, would point to a probable decrease in interaction affinity or complete disruption. On the contrary, pairs whose score increases to surpass the WeSA threshold, probably appear due to the mutation.

In figure 4.16 we present this idea through an example implemented on experimental data on IFT140. The data is obtained from AP experiments conducted using the protocol in figure 3.1 in the Institute for Ophthalmic research, Tübingen University. As bait were used both IFT140 WT and the variant T484M. Calculating the WeSA scores for both WT and mutant resulted in either a complete disappearance or a sharp plunge of the WeSA scores of several possibly interacting pairs. Intentionally the results are presented in full before clustering, in order to highlight the contrast between identified complex components and the rest of the network. WeSA scores for the rest of the network remain relatively similar while the focus of change is within the identified IFT complex lending credibility to this result and the potential of the method.

A challenge for such use is limited data related to interactions with mutated proteins. Moreover, such information is collected in cells which normally are unchanged except for the bait which is modified. That means that scores for an interaction containing a variant would be calculated based on only the single spoke term and thus, could be much less reliable than other better-informed scores containing also information from the reversed spoke model and the matrix.

The similarity to two other known measures, risk ratio and odds ratio, makes it a possible future direction to explore such substitution of the log-term. The benefit of odds ratios (OR) can be in deriving confidence intervals (CI) around the scores. While we

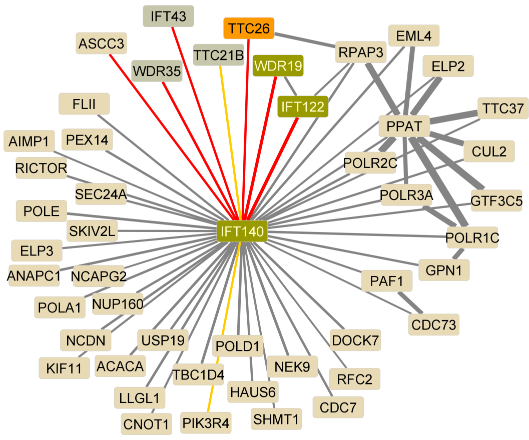


Figure 4.16: Network representing the comparison between IFT140 WT and IFT140 T484M variant. The edges are the links to the 30 top WeSA-scoring interactors of IFT140 WT. No clustering is applied. Red edges are completely lost in the mutant and the yellow edges have significantly decreased in their WeSA score for the variant (and are unlikely to be present). Nodes in olive-green are the core IFT-A complex components and in light grey are the three peripheral proteins in the same complex. TTC26 is a component of the connected IFT-B complex.

have implemented CI calculation for WeSA based on an estimation of a confidence interval for the expectation, the ones for the OR are based on an established mathematical formulation and, thus, can be more informative.

Alternatively, the risk ratio (RR), also called relative risk, is already widely used in the clinical settings for measuring disease risk. In that setting, it is found to be more accurate than the OR when diseases are not rare [FR17]. In PPI networks, some links *are* rare, but in other contexts where the numbers are higher, it can be a statistic worth exploring.

Both the aforementioned measures are critiqued for losing information on baseline risk[5], which can be re-framed as the link to the original numbers. For example, RR, which is a ratio of probabilities can increase by 300% but if the underlying baseline probability is 0.001, the increase to 0.003 may not be very important. While in WeSA the formula involves observation numbers and rare proteins are of no less importance than the more abundant proteins, we still make sure to report the raw observation numbers in each term category (the two spoke and the matrix term) separately.

With the recent breakthrough in protein interaction prediction and the development of AF-M, there is a possibility of accelerating the characterisation of the full interactome. While the application of AF on the full PPI landscape requires immense resource and, though attempted [FR5], should still be guided by already known information on the PPIs. We propose that, as seen in the section on 3D structures (within section 4.3.1), WeSA can be applied as a first-level filter for noisy datasets and experiments. This can reduce the needed resource by narrowing down the space for a second-level verification by AF-M. Only after these two need a complex or interaction be tested in the lab to ascertain it and add to a corpus of gold-standard interaction data.

Finally, WeSA can be applied more generally to probe other biological and non-biological networks in which a lot of noisy data is present and in need of cleaning. In the next chapter, we present some such applications.

[5] *baseline risk* - in a context of getting a disease, this is the risk in the unexposed (control) group; in a proteomics context, this would be the probability of the target prey to interact with proteins other than the target conditional on the total number of links of those alternative baits.

5 Applications of WeSA

5.1 Genome-wide association studies

With the advancement in technology and specifically, exome and genome sequencing, researchers have been able to study the genetic code of individuals and attempt to link it to personal features and diseases. The first such studies, the genome-wide scans using positional cloning [56] assumed that every disease is connected to unique regions in the DNA and aimed at finding them through sequencing of patients and their families. The method worked for diseases such as cystic fibrosis which are characterised by well-known Mendelian inheritance [76, 5], but the unique-loci assumption was disputed. In particular, a meta-study by Altmüller *et al.* [5] on 101 scans for 31 complex diseases such as Alzheimer's disease and asthma is one of the first analyses of the accuracy and universality of the method. It suggested that studies are combined to achieve better coverage of the gene-disease associations.

Indeed, the field of gene-disease links discovery has grown to be able to accommodate for more flexibility, including the discovery of variants of only moderate impact or multiple combinations of variants which lead to similar and/or simultaneous effects. This is what Genome-wide Association Studies (GWAS) do. Early studies started by looking at a specific disease and making comparisons to a control group to identify single-nucleotide polymorphisms (SNPs) associated with the specific disease [33]. With improved resource availability, however, large projects which have the means to investigate thousands of exomes with a mix of traits have appeared (e.g. the UK Biobank [6]).

5.1.1 Data

All GWAS results are recorded in the unifying database of the GWAS Catalog [62] and are updated continuously. We retrieved the full data for this thesis on 27.01.2023.

These consist of around 470 thousand associations between a trait and positions on the genome out of which 264,088 are within regions linked to a gene. Only these observations are considered further; the other approximately one third of the data which is in intergenic regions is filtered out.

The data covers 17,661 unique traits ranging from diseases such as breast cancer and inguinal hernia to features like BMI and height. In total, those link to 11,012 unique human genes. Yet, the variants are many more, the unique variants are almost 140 thousand out of which over 15 thousand are observed at least 3 times and 54 variants are repeatedly observed at least 100 times.

In terms of studies, at the moment of retrieval, there were 5,018 papers contributing information on gene-trait associations from GWAS. The majority reported on a single disease or trait and only about 6% observed links to more than 5 traits.

5.1.2 How to apply WeSA to GWAS

Gene-wide association studies are an example of networks with two types of nodes. One type of nodes comprises human features and diseases which we will refer to, jointly and for short, as traits. The other type of nodes consists of genes or variants linked to those traits. The traits in these studies are usually the equivalent of baits in proteomics, whereas variants (or genes) can be thought of as preys.

The obvious difference is that this graph is bipartite, since no variants relate directly to each other as is also the case for traits. This eliminates one of the spoke terms in the WeSA equation. There are also no matrix terms which can be calculated for variant-disease associations. Thus, a score for the association between a trait T and a variant V can be calculated as:

$$WeSA(T \leftrightarrow V) = O_{T \leftarrow V} \times \log \frac{O_{T \leftarrow V}}{E_{T \leftarrow V}} \tag{5.1}$$

The expectation is computed mathematically as in Chapter 4.

There is a second perspective on this dataset which can score the indirect association between nodes of the same type. This can be done through the matrix term both for gene-gene association based on their shared traits and for trait-trait associations based on shared genes. In this case a score between two traits T_1 and T_2 (analogously for gene-gene or variant-variant associations), is calculated as:

$$WeSA(T_1 \leftrightarrow T_2) = \frac{O_{T_1-T_2}}{\sum\limits_{X,Y:traits} O_{X-Y}} \times \log \frac{O_{T_1-T_2}}{E_{T_1-T_2}} \tag{5.2}$$

The dash [$-$] signifies, as before, an indirect contact, i.e. a shared neighbour. In this application, scaling of the weight is not necessary as the score only consists of this single term.

5.1.3 Results

There is no dataset which can give precise information of the true positives, but we can test it by validating hypotheses, comparing the results to other research and looking at case studies.

Media bias is down-scored

One of the aims of WeSA is to reduce non-specificity. In the case of GWAS, this includes links between *trending* traits and genes. Specifically, the top 5 traits with the most GWAS links are: height, educational attainment, BMI, smoking initiation and white blood cell count (full details in table 6.3 in the Appendix).

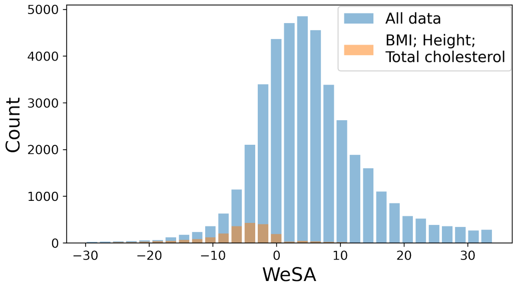
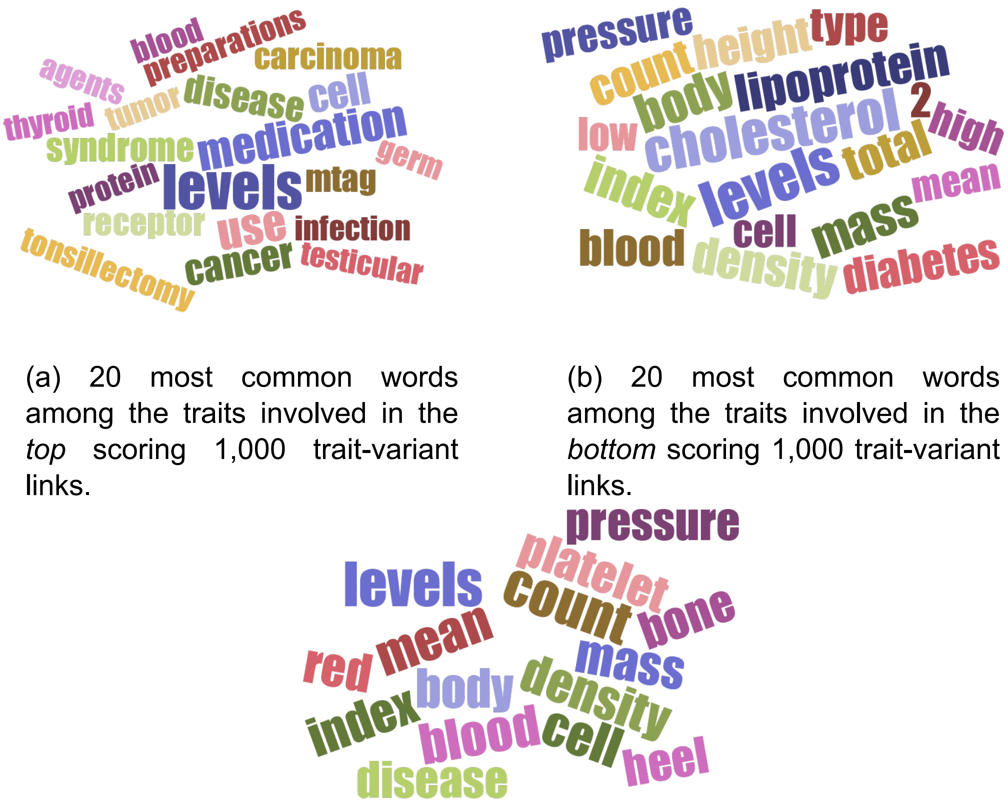


Figure 5.1: Histogram of the computed WeSA scores.

After scoring the connections in the database, these popular traits are confirmed to score lowest (figure 5.1). BMI is observed 152 times among the bottom 1,000 scores. Notably, only BMI for infants had higher scores (linking to variant rs2767486 in the locus of the leptin receptor gene LEPR), whereas no variant scores high for BMI in later years. The other most common traits at the bottom of the scores table are ‘Total cholesterol levels’ (106 occurrences), ‘Height’ (87), ‘Type 2 diabetes’ (83) and ‘Triglyceride levels’ (68).

At the other end, with the highest scores are normally diseases. Among the top 1,000 traits the most common words include ‘disease’, ‘syndrome’, ‘cancer’ and ‘tumor’ (figure 5.2a). Notably, those are not observed among the words from low-ranked traits (figure 5.2b).



(a) 20 most common words among the traits involved in the *top* scoring 1,000 trait-variant links.

(b) 20 most common words among the traits involved in the *bottom* scoring 1,000 trait-variant links.

(c) 15 most common words among all traits in the GWAS database (with repetition for different variants and studies they link to).

Figure 5.2: Most common words among traits in different categories of GWAS data. Generated using: https://www.jasondavies.com/wordcloud/

Trait	Gene	Observed number	WeSA score
Colorectal cancer	SMAD7	23	22.55
Breast cancer	FGFR2	21	17.59
Type 2 diabetes	CDKAL1	36	-44.46
Type 2 diabetes	FTO	22	-56.56
Type 2 diabetes	TCF7L2	44	-57.87
Body mass index	FTO	38	-83.1

Table 5.1: Scores of trait-gene links with more than 20 observations for the specific pairing.

It is also confirmed that the most popular traits are commonly scoring low (figures 5.2b and 5.2c). This is expected since popular traits such as body mass index (after infancy), while much researched, may not have clear genetic origins. It is worth noting that there is also more variety among the traits with highest scores. This serves as another confirmation of the specificity of the genes linked to those traits. Finally, high observation numbers are not correlated with the WeSA scores. Particularly, there are 6 trait-gene links with more than 20 observations as shown in table 5.1. Out of those, four score very negatively and two score positively, suggesting that absolute observation count is not always indicative of significance.

There are confirmed variant-disease links among the top WeSA scorers

In this section we briefly explore two directions for validation: summary statistics for the whole ranked dataset and specific case studies.

We calculated gene-trait scores based on GWAS pairs and compared their pattern to UniProt records of diseases. UniProt provides manually curated annotations about genes' involvement in disease. We used data retrieved on 28.02.2023 where we found 30.9% of the genes on GWAS have annotations on UniProt (13,157 genes). UniProt is continuously expanded, so the coverage and amount of annotations per gene may increase. Notably, however, we compared gene-traits that are supported by UniProt and those that are not and discovered significantly higher WeSA scores among the former (MWU test, $p < 10^{-25}$, figure 5.3). Since UniProt rarely contains non-disease trait information, and otherwise is often focused on very widely studied diseases (cancer), there is an inherent bias in its information which makes these results inconclusive, but encouraging enough to warrant further investigation.

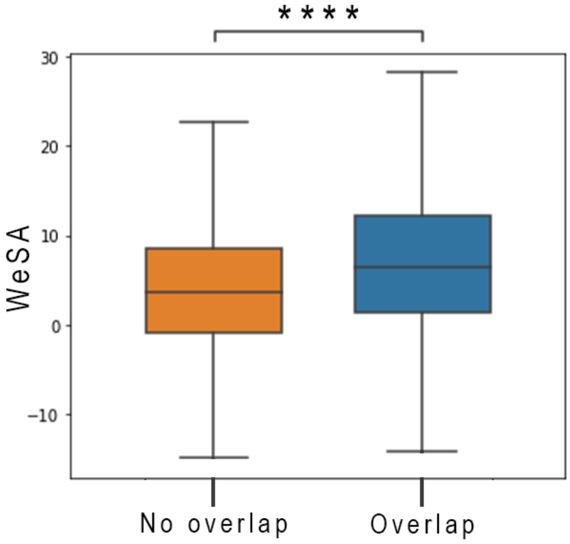


Figure 5.3: Boxplot showing the computed WeSA scores of trait-gene relationships. The groups are separated based on whether the trait-gene relationship has any overlap with UniProt records.

It is beyond the scope of this thesis to do a detailed investigation of all trait-variant links. Moreover, for a completely non-biased analysis, more complicated procedures should account for linkage disequilibrium[1] relating SNPs. Therefore, the following is only the start of such study to indicate the promise of the method.

In this second validation attempt, we look at additional evidence to confirm some of the

[1] Linkage disequilibrium (LD) refers to the non-random simultaneous variation in two alleles. A high LD means correlation between them.

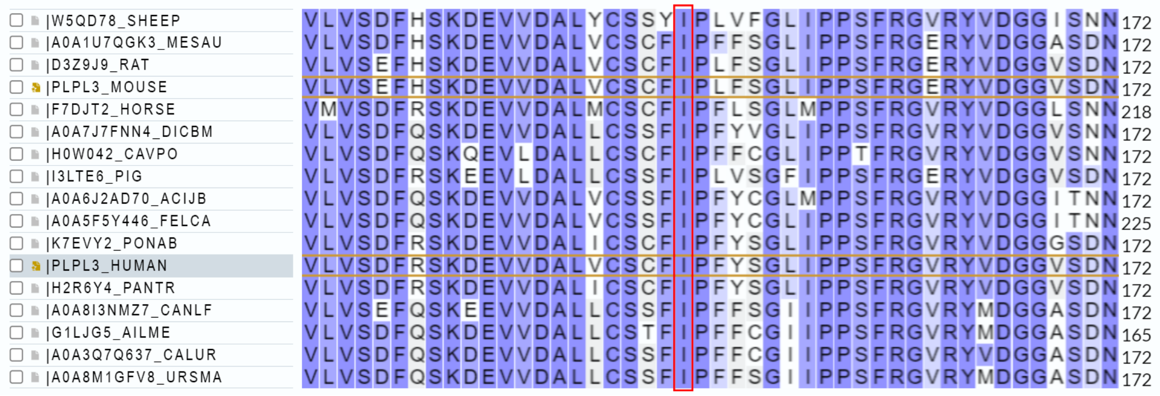
top-scoring disease-variant links. We go over four case studies below. Full results for those are listed in the Appendix,in tables 6.2-6.8.

The top ranked trait-variant association has supporting evidence from human and mice studies. The connection between variant rs738409 and non-alcoholic fatty liver disease has the highest WeSA score (Appendix, tables 6.4, 6.5). As a reminder, this is calculated based on GWAS records, specifically, observation counts and computed expectation numbers.

The specific rs738409 variant is mapped to an isoleucine to methionine change at position 148 of the gene PNPLA3 (PNPLA3/I148M). PNPLA3 is a catalyst for the process of synthesising phosphatidic acid.

This top scorer is a well-studied variant outside GWAS studies. In human studies targeting adults [FR53] and adolescents [FR44], patients with liver problems and, particularly non-alcoholic fatty liver disease, PNPLA3/I148M was studied. Both studies find the I148M variant is significantly more common in target populations compared to control groups.

Although GWAS records only human studies, the isoleucine at this protein position is conserved in many other species (figure 5.4). Particularly, the corresponding *Mus musculus* protein is slightly shorter, but has 84% coverage and 67.65% sequence identity.



```
☐ |W5QD78_SHEEP        VLVSDFHSKDEVVDALYCSSYIPLVFGLIPPSFRGVRYVDGGISNN 172
☐ |A0A1U7QGK3_MESAU    VLVSDFHSKDEVVDALVCSCFIPFFSGLIPPSFRGERYVDGGASDN 172
☐ |D3Z9J9_RAT          VLVSEFHSKDEVVDALVCSCFIPLFSGLIPPSFRGERYVDGGVSDN 172
☐ |PLPL3_MOUSE         VLVSEFHSKDEVVDALVCSCFIPLFSGLIPPSFRGERYVDGGVSDN 172
☐ |F7DJT2_HORSE        VMVSDFRSKDEVVDALMCSCFIPFLSGLMPPSFRGVRYVDGGLSNN 218
☐ |A0A7J7FNN4_DICBM    VLVSDFQSKDEVVDALLCSSFIPFYVGLIPPSFRGVRYVDGGVSNN 172
☐ |H0W042_CAVPO        VLVSDFQSKQEVLDALLCSCFIPFFCGLIPPTFRGVRYVDGGVSNN 172
☐ |I3LTE6_PIG          VLVSDFRSKEEVLDALLCSSFIPLVSGFIPPSFRGERYVDGGVSDN 172
☐ |A0A6J2AD70_ACIJB    VLVSDFQSKDEVVDALVCSSFIPFYCGLMPPSFRGVRYVDGGITNN 172
☐ |A0A5F5Y446_FELCA    VLVSDFQSKDEVVDALVCSSFIPFYCGLIPPSFRGVRYVDGGITNN 225
☐ |K7EVY2_PONAB        VLVSDFRSKDEVVDALICSCFIPFYSGLIPPSFRGVRYVDGGGSDN 172
☐ |PLPL3_HUMAN         VLVSDFRSKDEVVDALVCSCFIPFYSGLIPPSFRGVRYVDGGVSDN 172
☐ |H2R6Y4_PANTR        VLVSDFRSKDEVVDALICSCFIPFYSGLIPPSFRGVRYVDGGVSDN 172
☐ |A0A8I3NMZ7_CANLF    VLVSEFQSKEEVVDALLCSSFIPFFSGIIPPSFRGVRYMDGGASDN 172
☐ |G1LJG5_AILME        VLVSDFQSKDEVVDALLCSTFIPFFCGIIPPSFRGVRYMDGGASDN 165
☐ |A0A3Q7Q637_CALUR    VLVSDFQSKDEVVDALLCSSFIPFFCGIIPPSFRGVRYVDGGASDN 172
☐ |A0A8M1GFV8_URSMA    VLVSDFQSKDEVVDALLCSSFIPFFSGIIPPSFRGVRYMDGGASDN 172
```

Figure 5.4: Results from multiple sequence alignment of the closest corresponding genes of 17 species between positions 136 and 152 of the human PNPLA3 protein. I148 is highlighted.

The mouse gene is interesting because of a study on mice confirming the association of the PNPLA3 gene to non-alcoholic fatty liver disease and outlining the mechanism of action [FR27]. In particular, Kumari *et al.* observe that murine with the I148M mutation experience phosphatidic acid synthesis from lysophosphatidic acid (LPA) at twice the rate of the wild-type mice. This leads to lipid accumulation which is speculated to be the cause of non-alcoholic fatty liver disease.

There is related evidence for variant rs1047891 to affect homoarginine levels. In GWAS, variant rs1047891 which is positioned within the gene CPS1 (T1406N) has been observed with 28 distinct traits. However, after applying WeSA, the variant scores high only in links with glycine-related levels and homoarginine levels (Appendix, table 6.6). Interestingly, an independent study on infants has previously confirmed correlation of the variant with arginine levels [FR37]. In particular, homozygosity for the variant in the studied population was observed to correlate with higher arginine concentrations, which is linked to susceptibility to neonatal pulmonary hypertension. Since homoarginine is the homolog of arginine, this observation can be interpreted as a confirmation of the WeSA score.

The top scores for variant rs603424 are related in a study on swine. This variant is a common (minor allele frequency = 0.37) guanine to adenine substitution in an intron region mapped to the PKD2L1 gene. According to GWAS, it has associations to 16 different traits, but the top are levels of different fatty acids: myristoleate, palmitoleate, 5-dodecenoate and 1-palmitoleoyl-GPC (Appendix, table 6.7). In particular, the acids corresponding to the former two bases have been observed to act together in a swine study [FR48]. In this independent study, it was observed that the specific combination of both myristoleic and palmitoleic acid could raise low-density lipoprotein levels in growing swine.

The top 1% of all WeSA scores identify other confirmed relationships. In particular, looking at the links of colorectal cancer, four variants get scored in the top 1% (Appendix, tables 6.2, 6.8). These are rs6983267 (intergenic related to CASC8, CCAT2), rs3802842 (intergenic between COLCA1 and COLCA2), rs4939827 (SMAD7) and rs704017 (ZMIZ1-AS1 intron), all identified in multiple genome-wide association studies. Three of the four are also confirmed by other case-control statistical analyses to be significantly associated with colorectal cancer or tumours [FR42, FR7, FR41, FR49, FR45]. Specifically, the top-scoring variant is associated with gene CCAT2, which apart from human studies, has been researched by Chen *et al.* [FR7] in mouse organoids and confirmed to promote carcinogenesis when overexpressed.

One-mode projection network of the matrix term

Interesting in this case of the bipartite GWAS network becomes also the matrix term. It is a measure for the indirect association between genes as detected via their common neighbours. Specifically, we think that the higher a term is, the more often the two genes co-occur in traits and thus there may be a link between them. On the other hand, low matrix scores are probably enriched in proteins which do not link to each other.

If we define genes which are linked to many of the same diseases as similar, we can propose that diseases which are unique to each of two similar genes can potentially be yet undiscovered but shared diseases, too. On the other hand, if there are some diseases which are not shared for two extremely similar genes, it can be an interesting case exploring what is the cause of distinction.

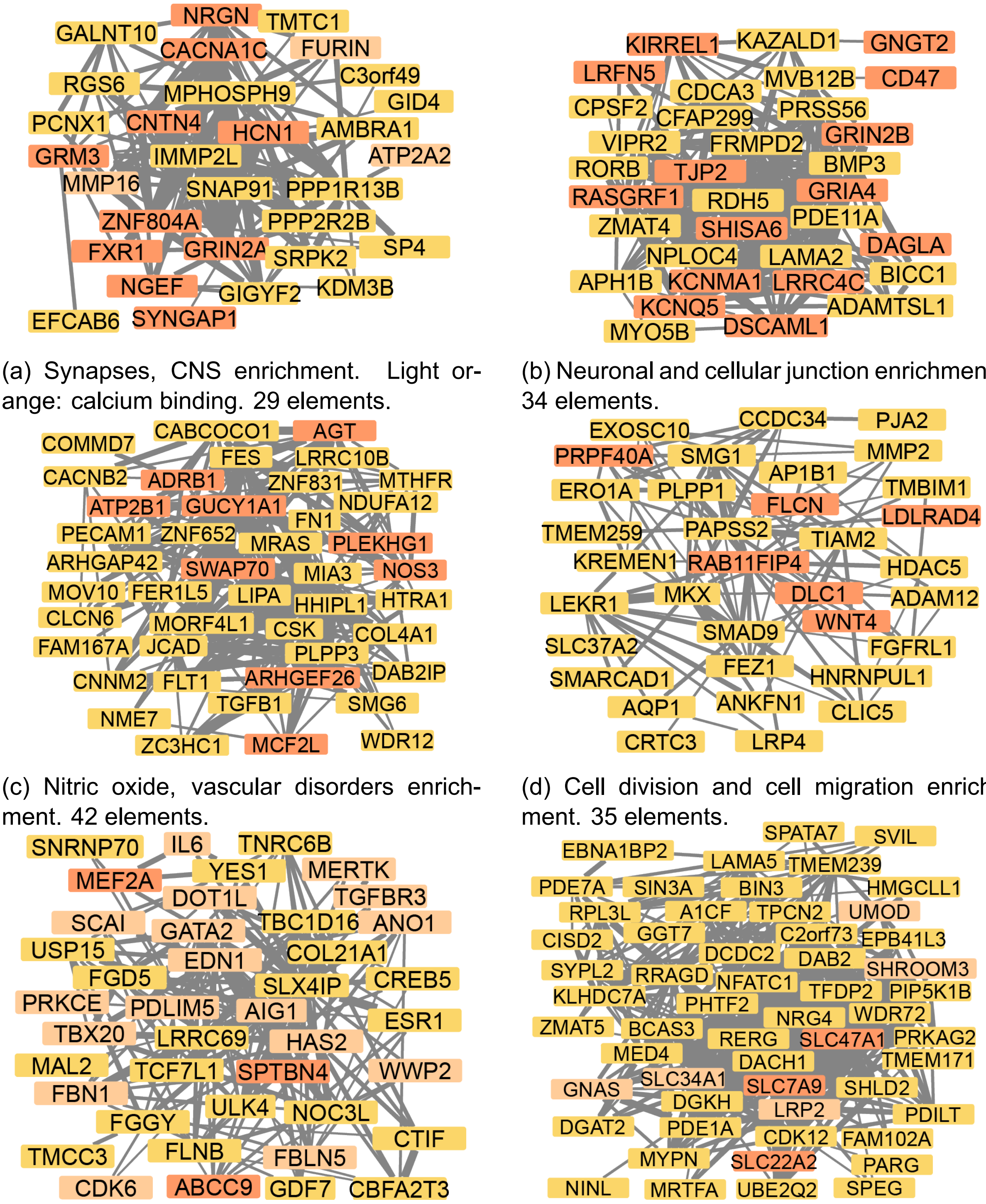
This is an adaptation of the 'guilt by association' idea where we assume that if items are similar, they link to overlapping sets of features. The approach has already been observed to work in a study about inflammatory bowel diseases (IBD) from Franke *et al.* [28], which illuminates new loci associated to Crohn's disease. The main idea is that IBD consists of sub-phenotypes, particularly, ulcerative colitis and Crohn's disease, and they found new genes that were previously only associated with ulcerative colitis but turn out to be Crohn's disease genes as well.

If we assume that the matrix term of WeSA can provide a measure of similarity between genes or variants (higher scores link to more of the same traits), then we can cluster the projection network and try to suggest novel effects of those genes. In figure 5.5 and in the rest of this chapter, we present a few examples to illustrate the idea.

To get the gene-gene network, we calculated the matrix terms of the WeSA scores as presented previously. Thus, scores of connections are based on shared diseases. Clustering is then performed using MCL clustering on the weighted network.

We observe 459 clusters in total with sizes up to 695 genes, excluding the giant component. However, the median cluster size is 3 and only 11 of the clusters are medium-sized, meaning, they contain between 25 and 60 elements. We pick those sizes to look for meaningful enrichment of GO terms and pathways through GetGo [12].

All 6 of the clusters shown in figure 5.5 are enriched in genes from specific biological processes. 10 out of the 29 genes in figure 5.5a are found significantly in neuronal components: postsynaptic density, dendritic spine, glutamatergic synapse, axon and



(a) Synapses, CNS enrichment. Light orange: calcium binding. 29 elements.

(b) Neuronal and cellular junction enrichment. 34 elements.

(c) Nitric oxide, vascular disorders enrichment. 42 elements.

(d) Cell division and cell migration enrichment. 35 elements.

(e) Cardiovascular enrichment from biological processes in the gene ontology. 40 elements. Light orange nodes are showing cardiovascular disease enrichment.

(f) Enrichment in membrane proteins (light peach), L-alpha-amino acid transmembrane transport (orange). The group is specifically enriched in kidney proteins. 54 elements.

Figure 5.5: Six of the medium-sized clusters resulting from clustering the weighted gene-gene network based on WeSA matrix scores calculated from GWAS. The genes in orange contribute to the specified enrichment.

growth cone. Some of them are also found to be leading to enrichment for genes regulating synaptic plasticity. According to GWAS data, those genes are overwhelmingly related to schizophrenia but had been also linked to 377 other traits including anorexia, bipolar disorder, smoking initiation and educational attainment. It could be speculated that diseases which are common for the cluster are actually more widespread. Specifically, given schizophrenia is so common among these genes, could those not found through GWAS also be associated to the disease? Out of the highlighted “neuronal” genes, all but two are associated to Schizophrenia in GWAS. It is then our hypothesis that the other two, CACNA1C and GRM3, are also schizophrenia associated genes.

In the cluster in figure 5.5c, there is enrichment in regulation-related genes, specifically nitric oxide mediated signal transduction, regulation of blood pressure and peptidyl-tyrosine phosphorylation. There are also enriched signalling pathways, such as MAPK and RAS signalling and signalling by SCF-KIT and NGF. In the raw GWAS data, the genes from this group are associated mostly to vascular traits with the top 3 being systolic and diastolic blood pressure and coronary artery disease. Out of the 9 genes highlighted for their regulatory function, 8 are already associated with coronary artery disease in GWAS. Drawing on the guilt by association hypothesis again, we can speculate that at least the ninth one, ADRB1, is also associated to the disease. This gene codes for the beta-1 adrenergic receptor, the target for drugs mediating many heart problems.

Similar cases can be made for the other clusters. Cluster 5.5b is enriched in many neuronal and cellular junction terms, while in GWAS it is associated to traits of the eye: refractive error, myopia, spherical equivalent. The two together suggest a mechanism of disease and can again give rise to gene-trait association predictions. The clusters in 5.5d and 5.5f are less specific but we still observe enrichment. The cluster in 5.5d is enriched in cell division and migration-annotated genes, which can be connected to the top GWAS traits for the group: heel bone density and height. The last plot, 5.5f, is enriched in membrane proteins. 11 of the proteins[2] are found in kidney tissue which also makes this enriched. The latter is in agreement with GWAS, in which the top associations of genes within this group are with glomerular filtration rate and creatinine levels.

Finally, the case of the cluster in 5.5e can be interesting for their association to cardiovascular diseases which have not (yet) been annotated to cardiovascular processes.

[2] Specifically: KLHDC7A, LRP2, PDE7A, PIP5K1B, SLC22A2, SLC34A1, SLC7A9, SVIL, TFDP2, UMOD, WDR72

The genes in the cluster show some association to vascular traits (pulse and blood pressure). The GetGo gene-function enrichment tool shows enrichment in two biological processes: cardiac transduction (proteins: ABCC9, MEF2A, SPTBN4) and SMAD protein phosphorylation and binding (proteins: GDF7, TGFBR3, USP15, MEF2A). The two perspectives (processes and diseases) can be linked and we may be able to predict processes from diseases and vice versa.

5.1.4 Discussion

We have ranked the gene-trait and SNP-trait associations using the spoke term of the WeSA score. In section 5.1.3 we give examples of some validated case studies. We discuss the circumstances and expectations from GWAS studies and the inherent media bias which creates ‘hubs’ of genes associated to diseases that have been studied more extensively than others. For instance, in absolute numbers, more genes are discovered in connection to ‘popular’ traits. This, however, seems to be counteracted by the WeSA ratio which compares observation to expectation numbers. Therefore we expect that “hot” traits will be among those with many connections but low WeSA numbers.

We have also discussed the projection graph idea, which consists in ranking the similarity of genes (or traits) based on the matrix term in the WeSA score. This can be used to predict gene-disease links, or other gene-gene links, as for example pathways. The idea is that we can use the high similarity of genes (high WeSA matrix terms) to predict similarity in functions, interaction, folding, or even disease mechanism.

As an expansion, another interesting trajectory to explore is a strength of connection based not only on WeSA scores, but on disease similarity as well. That is, if we can define a hypergraph, in which every hypernode is a cluster of traits, then we can classify a strength of relation based on the weight of the hyperedge between the gene and the hypernode. This can be explored by naïve counting or it can again be put into the WeSA observed-to-expected framework.

Limiting to this analysis is the higher complexity and impossibility to standardise all information. It should be noted that mutations are often *tissue-specific*, *outside coding regions* of the genome, they can be (trans-) mutations which can act very *far from the gene* relevant to the phenotype or be in regions of *linkage disequilibrium*. To clarify the third point, regulatory trans- mutations are defined as those acting on a gene more than 5 Mb away (incl. on another chromosome). There are many SNPs outside a gene-encoding region and it is not always known what the effect of such mutations is, if any. Lastly, attempts to annotate linkage disequilibrium (LD) have been made [FR15], but the standardisation of resulting records is still slow [FR50, FR36]. In view of this, further efforts can lead to incorporation of the LD knowledge and subsequent refinement of the WeSA scores.

GWAS trait handling is also still underdeveloped. While for diseases there is hierarchy, for traits there is none[34]. A researcher who wants to group traits or focus on high-level traits would only need to build their own ontology based on the data.

5.2 Mouse Genome Informatics (MGI)

The continuous development of new technologies and methods in genetic research and related biology fields have been foundational for the understanding of the genome. While the human genome was only sequenced last year [FR33], deciphering of the mouse genome is possible since 2002 [FR56]. As an essential model organism mice have been of particular scientific interest and the research community has been accumulating mice genetic data at a rapid pace making use also of newer high-throughput experimental methods. Collected experimental information about the laboratory mouse (*Mus musculus*), including a database of mouse genes, genomic sequences, and phenotypic information is recorded by Mouse Genome Informatics (MGI).

In this section, we discuss the use of WeSA to analyse MGI records from genomic studies investigating phenotypes. In particular, phenotypic studies in mice consist of generating genetically modified mice using methods such as CRISPR/Cas9 [FR59, FR55] followed by analysis of the mice for phenotypic changes. Phenotypic changes include differences in their behaviour, morphology and physiological functions, and are validated to confirm that the observed phenotypes are directly related to the targeted gene (as opposed to being due to off-target effects or other confounding factors) [FR18]. However, there still remains bias towards ubiquitous and/ or non-specific phenotypes which, we argue, can be decreased by the application of WeSA.

[3] https :// biobank.ctsu.ox.ac.uk/ crystal /

[4] https :// biobank.ctsu.ox.ac.uk/ crystal /browse.cgi?id=−2&cd=field_list

5.2.1 Data

For this study we have used the MGI dataset as retrieved and pre-processed by Benedetta Leoni, a fellow in the EU SCilS consortium. The final data which are presented here were retrieved on 08.03.2023 and cleaned to leave only four columns:

- allele which was tested;
- phenotype which was observed for the instance;
- reference publication (as a PubMed ID);
- system in which the phenotype manifests as classified in MGI. This encompasses 29 general categories which act as parents to the annotated phenotypes, e.g. immune system, adipose tissue, skeleton, embryo.

The data consist of around 468.5 thousand association triplets of a mouse gene, linked phenotype and study. It covers more than 10 thousand unique phenotypes[5]. In total, those come from almost 35 thousand studies and link to 15,845 unique mouse genes. The phenotypes are categorised into 29 general systems, which are of significantly unequal sizes (table 5.2). For example, the olfactory and taste functions and their corresponding phenotypes have not been reported at the same order of magnitude as the most common systems in the database: immune, hematopoietic and nervous systems.

5.2.2 How to apply WeSA to MGI

MGI is another example of a bipartite network, as it contains two groups or types of nodes: phenotypes or systems and genes. The genes in these studies are usually the equivalent of baits in proteomics, since the analysis starts by mutating a chosen gene. Then, the resulting effect on the phenotype is observed, which can be thought of as preys. Phenotypes are, however, very specific and 30% of them have been observed at most 5 times, making repetition of gene-phenotype pair unlikely and any analysis quite noisy. In response, instead of using phenotypes, we have used the more general categories of ‘systems’ as preys (table 5.2) on which to apply WeSA.

Same as when applied to GWAS, WeSA can be calculated separately for two different purposes. In the first approach, a score is calculated for the association between every

[5] Notably, only phenotypes are recorded. There is no disease information.

System	No. of links	System	No. of links
Immune system	55,014	Craniofacial	14,578
Hematopoietic system	47,533	Renal/urinary system	12,306
Nervous system	43,444	Neoplasm	10,598
Homeostasis/metabolism	39,159	Muscle	9,874
Growth/size/body	32,178	Digestive/alimentary system	9,826
Cardiovascular system	30,611		
Skeleton	29,380	Limbs/digits/tail	8,577
Behaviour/neurological	27,515	Hearing/vestibular/ear	8,143
Mortality/ageing	27,475	Respiratory system	8,119
Reproductive system	26,061	Liver/biliary system	6,770
Cellular	23,705	Pigmentation	5,076
Endocrine/exocrine/glands	18,747	Adipose/tissue	4,096
Embryo	17,595	Normal/phenotype	3,991
Vision/eye	17,375	Not analyzed	2,095
Integument	16,437	Taste/olfaction	494

Table 5.2: List of systems and the number of times a phenotype within that system has been observed to link to an allele as recorded in MGI.

system and gene. This is computed exactly in the same way as in equation 5.1, but substituting a trait (T) for a gene and a variant (V) for a system to obtain gene-system scores.

The second approach would be, analogously to equation 5.2 from section 5.1, to score the indirect association between nodes of the same type. Of particular interest can be the matrix term score for gene-gene association based on shared classification. In this case a score between two genes, G_1 and G_2, is calculated as:

$$WeSA(G_1 \leftrightarrow G_2) = \frac{O_{G_1-G_2}}{\sum\limits_{X,Y:genes} O_{X-Y}} \times \log \frac{O_{G_1-G_2}}{E_{G_1-G_2}}$$

The dash [$-$] signifies, as before, an indirect contact, i.e. a shared neighbour. We hypothesise that this score is related to a similarity score of genes, which will be investigated in the next section.

5.2.3 Results

In the case of the mouse genome to phenotype relationships, there is no possibility to obtain accurate and certain information on gold standard links. Instead of a strict comparison and a benchmark, we discuss our expectations and compare the top scores to diseases or the STRING similarity score. In the main part of this section, we look at validation strategies for the bipartite model. In the discussion, we expand on the topic by considering the one-mode projection network with some possible validations.

Throughout this chapter, we normally use the second level of the MGI phenotype ontology (consisting of 29 classification terms) and the phenotypes at that level are referred to as "system (MGI)s". All levels are reported starting from level 1, which comprises "mammalian phenotype" only (the most general level).

Validating gene-system scores based on annotated diseases

After obtaining the standard WeSA gene-system scores, we can compare them to disease annotations for human proteins reported by UniProt. UniProt links a gene to a disease if either variants in that specific gene are found to cause the disease or the gene is confirmed as essential in the steps leading to the disease. These annotations are not provided for mouse genes, but we can infer them from their human counterparts. The workflow for this approach is the following:

1. calculate all WeSA scores for gene-system pairs;
2. go through genes of interest in two steps:
   i. observe the scores of a gene X (of interest) and find out if it scores high in association with any systems (WeSA scores). Note that some genes may have more systems which associate strongly with them, while others may have only one or even none;

   ii. obtain the human ortholog (i.e. evolutionary equivalent), X_h of gene X;

   iii. extract the diseases linked to gene X_h from UniProt;

   iv. validate if the disease corresponds to the systems which were observed to score high in i.

This validation is clearly based on manual validation. However, there is a possibility for automation. The missing link preventing us from devising a computational testing pipeline is the lack of means to compare diseases and systems automatically. That is, if there is a map annotating diseases to the main systems they affect, the accuracy of the scores could be measured by a categorical test, e.g. how many times there is an overlap between the disease systems and the top 3 highest WeSA scoring scoring systems.

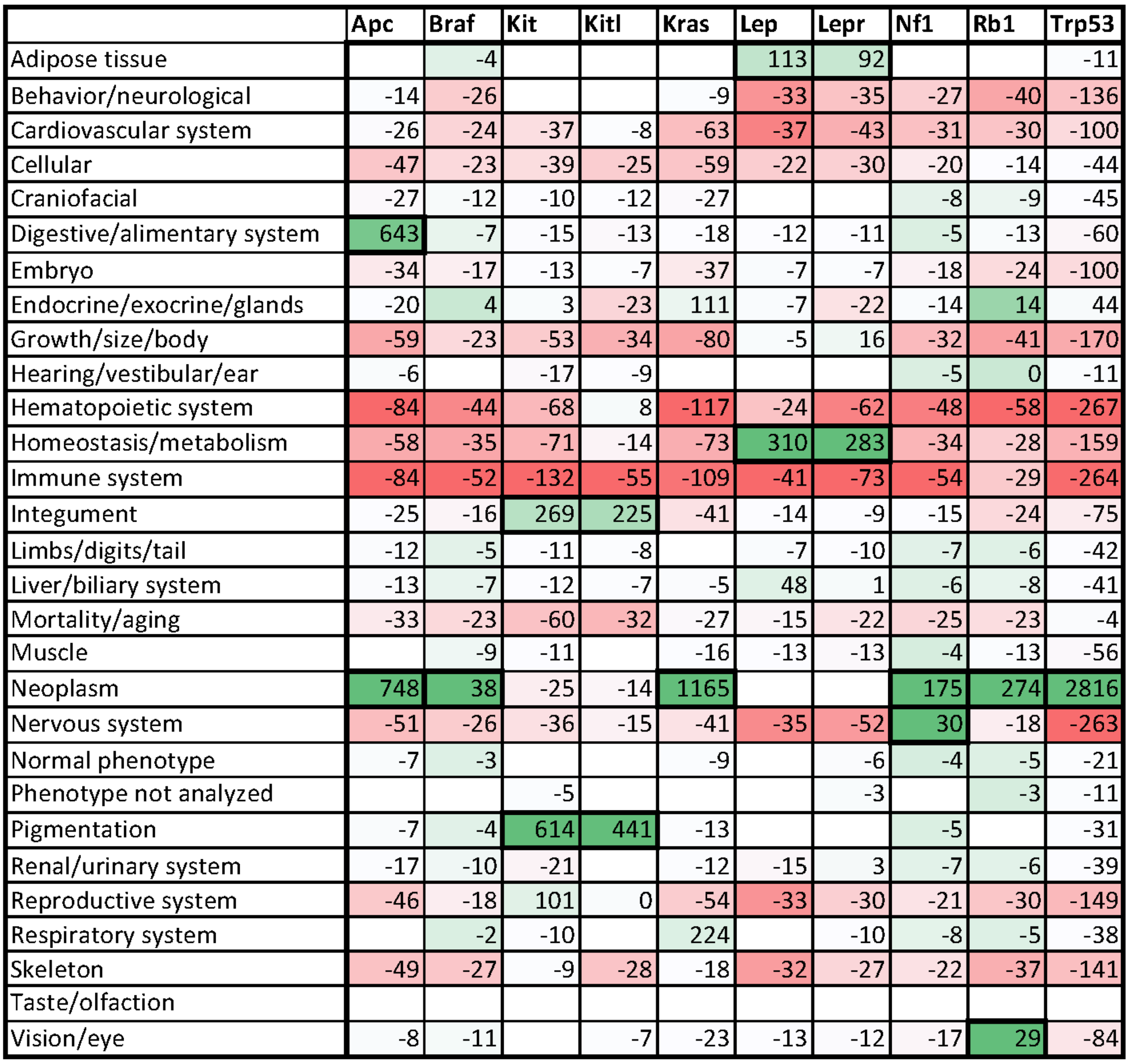
WeSA works well in narrowing down the list of candidate-systems which are affected by a gene. Many genes in MGI are annotated with many different phenotypes to the point that sometimes these correspond to most of the systems at the same time, thus reducing their significance. Here, we applied WeSA to significantly reduce the list of systems associated to each gene to output more relevant associations. We started with a subset of genes that according to MGI are linked to at least 70% of the phenotype systems present (i.e. at least 21 of the 29 systems). There are 457 such genes and, unsurprisingly, many of them do not exhibit high WeSA scores. Only 91 have a link to at least one system (WeSA-)scoring above the mean.

If we look more closely into those 91 genes, we can see that despite their many annotations, WeSA reorders the strength of their effect to different systems quite well. In figure 5.6 we highlight some case studies. For example, the genes Apc, Braf, Kras, Nf1, Rb1 and Tpr53 are scoring highest for ‘neoplasm’ and indeed their human equivalent genes are all well-established cancer genes. Moreover, genes such as Apc, Nf1 and Rb1 have also secondary effects which has been correctly captured by their second-highest WeSA scores.

Among the selected 91 proteins, there are also two gene pairs (Kit and Kitl, Lep and Lepr) which interact to form a functional ligand-receptor unit. Their WeSA scores show overlap in the highest scoring systems for the two proteins in each of these pairs. Lep and Lepr score high in Homeostasis and metabolism, and Adipose tissue; Kit and Kitl both score highly for Pigmentation and Integument, terms that refer to biological process and organ that are affected by the diseases (fig. 5.6). These results are in agreement with the diseases annotated to their human equivalent genes, e.g. for Kit/Kitl: mastocytosis, Waardenburg syndrome, hyperpigmentation.

The other ligand-receptor pair we observe is Lep and Lepr which WeSA scores categorise as related to homeostasis/ metabolism and adipose tissue. Notably, these proteins have more than 20 system annotations, but WeSA is able to distinguish the top ones.

WeSA is able to downweight the systems which are unspecific. Since, every mouse experiment is resource-expensive, not detecting an obvious resulting phenotype might encourage the performance of additional less important or conclusive experiments until



	Apc	Braf	Kit	Kitl	Kras	Lep	Lepr	Nf1	Rb1	Trp53
Adipose tissue		-4				113	92			-11
Behavior/neurological	-14	-26			-9	-33	-35	-27	-40	-136
Cardiovascular system	-26	-24	-37	-8	-63	-37	-43	-31	-30	-100
Cellular	-47	-23	-39	-25	-59	-22	-30	-20	-14	-44
Craniofacial	-27	-12	-10	-12	-27			-8	-9	-45
Digestive/alimentary system	643	-7	-15	-13	-18	-12	-11	-5	-13	-60
Embryo	-34	-17	-13	-7	-37	-7	-7	-18	-24	-100
Endocrine/exocrine/glands	-20	4	3	-23	111	-7	-22	-14	14	44
Growth/size/body	-59	-23	-53	-34	-80	-5	16	-32	-41	-170
Hearing/vestibular/ear	-6		-17	-9				-5	0	-11
Hematopoietic system	-84	-44	-68	8	-117	-24	-62	-48	-58	-267
Homeostasis/metabolism	-58	-35	-71	-14	-73	310	283	-34	-28	-159
Immune system	-84	-52	-132	-55	-109	-41	-73	-54	-29	-264
Integument	-25	-16	269	225	-41	-14	-9	-15	-24	-75
Limbs/digits/tail	-12	-5	-11	-8		-7	-10	-7	-6	-42
Liver/biliary system	-13	-7	-12	-7	-5	48	1	-6	-8	-41
Mortality/aging	-33	-23	-60	-32	-27	-15	-22	-25	-23	-4
Muscle		-9	-11		-16	-13	-13	-4	-13	-56
Neoplasm	748	38	-25	-14	1165			175	274	2816
Nervous system	-51	-26	-36	-15	-41	-35	-52	30	-18	-263
Normal phenotype	-7	-3			-9		-6	-4	-5	-21
Phenotype not analyzed			-5				-3		-3	-11
Pigmentation	-7	-4	614	441	-13			-5		-31
Renal/urinary system	-17	-10	-21		-12	-15	3	-7	-6	-39
Reproductive system	-46	-18	101	0	-54	-33	-30	-21	-30	-149
Respiratory system		-2	-10		224		-10	-8	-5	-38
Skeleton	-49	-27	-9	-28	-18	-32	-27	-22	-37	-141
Taste/olfaction										
Vision/eye	-8	-11		-7	-23	-13	-12	-17	29	-84

Figure 5.6: Heatmap of WeSA scores of genes discussed in the text. Scores in darker shades of green are considered to indicate strong associations according to WeSA analysis.

finally detecting one. As a result, there is a bias, corroborated by the fact that ‘normal’ and ‘unobserved’ phenotype terms are among the least frequent annotations (bottom in Table 5.6)

We expect that some cellular phenotypes are often unspecific as consequence of rare and forced conditions, e.g. difference in cellular expression or abnormal mortality due to intentional aneurysm rupture. The ‘mortality/ ageing’ system has a whole category of induced mortality through different means such as ‘abnormal susceptibility to thrombosis induced morbidity/mortality’ or ‘abnormal susceptibility to colitis induced morbidity/mortality’. Additionally, cellular phenotypes represent changes at the cellular level that often may not translate into macroscopic phenotypes such as significant changes in tissue or organ morphology or physiology. Such phenotypes, we expect, would normally be annotated in mice which fare well beforehand and, thus, are under an abstract modification of the survivorship bias.

Our results presented in figure 5.7 confirm this suggestion. The two aforementioned

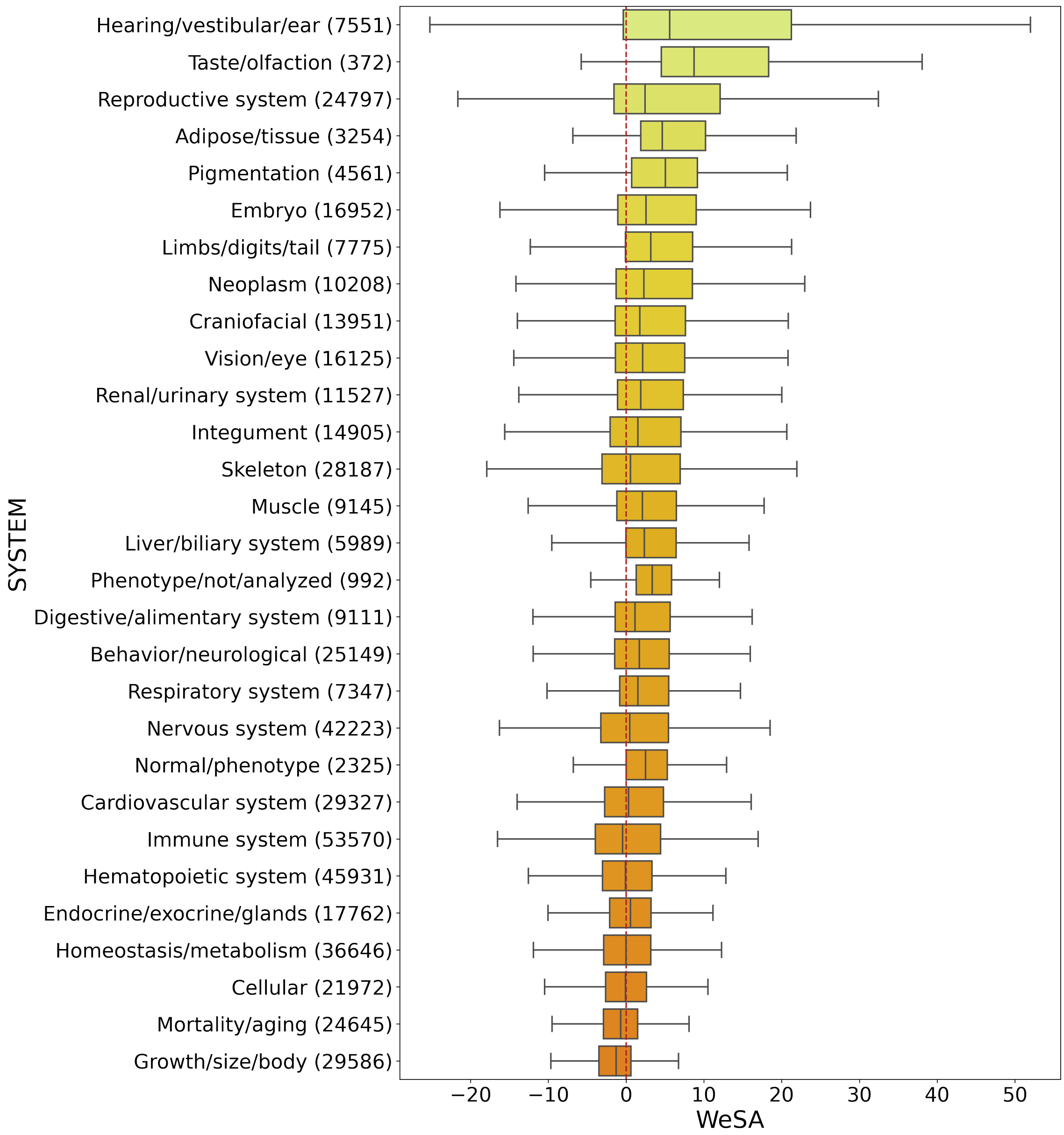


Figure 5.7: Boxplots (excluding outliers) of the distribution of WeSA scores within systems ordered by their 75% quartile. The dashed line illustrates WeSA = 0. Next to each system in brackets is given the number of times a phenotype of that system has been recorded.

systems have median WeSA scores comparable to random, while simultaneously having relatively low variance around the median showing that WeSA rarely detects these systems as specific to a gene.

On the contrary, the systems with highest Q3 (75% quartile) are normally very clear to detect. The most intuitive example is pigmentation which, when present is clearly visible and there is no specificity which makes the category artificially inflated.

Reproductive system case study: WeSA scores correlate with human diseases. As mentioned previously, there is no clear mapping of human diseases to systems. This is an obstacle before the creation of a full testing pipeline from the perspective of disease-system-phenotype match. Here we present a rough attempt to create such mapping from the disease ontology for the reproductive system and test our WeSA scores against it.

The Disease Ontology (DO) [61] has been developed with the aim to be useful in mapping model systems to human diseases. It contains a hierarchy of diseases and their terminology annotations. We used OMIM[6] codes to map the diseases from the disease ontology onto the diseases reported by UniProt. OMIM terms are simply standard terminology (codes) for human diseases.

If we go back to the pipeline introduced at the start of the results section, we see that it is designed to check our results element-wise, but we can modify 2. as follows:

i. convert all mouse genes to their human orthologs;

ii. extract the diseases associated to the human genes from UniProt; these are given alongside their respective OMIM terms;

iii. map UniProt diseases to DO through their OMIM terms;

iv. map disease ontology terms to systems;

v. test how WeSA scores behave within systems; are they better for some systems than for others?

While points i-iii. above are clear, the mapping between DO and systems is not yet available. Some DO terms seem to correspond to the systems in MGI but there are

[6] Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). World Wide Web URL: https://omim.org/.

discrepancies. The latter is an obstacle to the production of full mapping, but we believe that the link between MGI ‘Reproductive system’ and DO ‘Disease of anatomical entity-Reproductive system disease’ is unambiguous and we mapped them to each other. In total, DO contains 91 distinct diseases with OMIM annotations which fall within the reproductive system diseases. We use this mapping as a benchmark for the investigation in this section.

> ***Notation** [reproductive system annotations]:* First, WeSA scores for gene-system (incl. Reproductive system) links are calculated from MGI and the MGI phenotypes. In contrast, the disease-mapped ‘reproductive system’ category ($\mathrm{RS_D}$) contains genes (without any scores) which are mapped to the reproductive system through the human diseases they are linked to.

We mapped 3,628 of our scored genes to the DO through steps i-iii. The rest either did not have a human gene equivalent, did not correspond to any observed disease or the disease did not have any OMIM annotation. Out of the total, we mapped 84 genes to the reproductive system through steps i.-iv.

We used chi-squared test with the 2×2 contingency table 5.3 and confirmed that the ‘reproductive system’ annotations from MGI are more inside $\mathrm{RS_D}$ than they are in general (Chi squared statistic = 61.154 and p-value < .0001).

	Reproductive system	Overall
Number of genes	84	3,628
Number of MGI annotations in Reproductive system	76	985

Table 5.3: Statistics of the Reproductive system group.

Using the set of ‘positives’ established from the diseases of the reproductive system and setting as negatives everything outside of that set, we compare the distributions of the WeSA scores of gene associations to the reproductive system. In figure 5.8 we display the distribution boxplots and the computed ROC curve.

We compared the two shown distributions using a MWU test and found a highly significant difference ($p < 10^{-16}$) for a one sided hypothesis. In particular, this confirms the hypothesis that the distribution of WeSA scores for genes in the $\mathrm{RS_D}$ category is greater than the distribution of WeSA scores outside that category.

The second panel of figure 5.8 shows the ROC analysis and the optimal threshold which achieves near-perfect balance between TPR and FPR. Currently only 0.06% of the non-

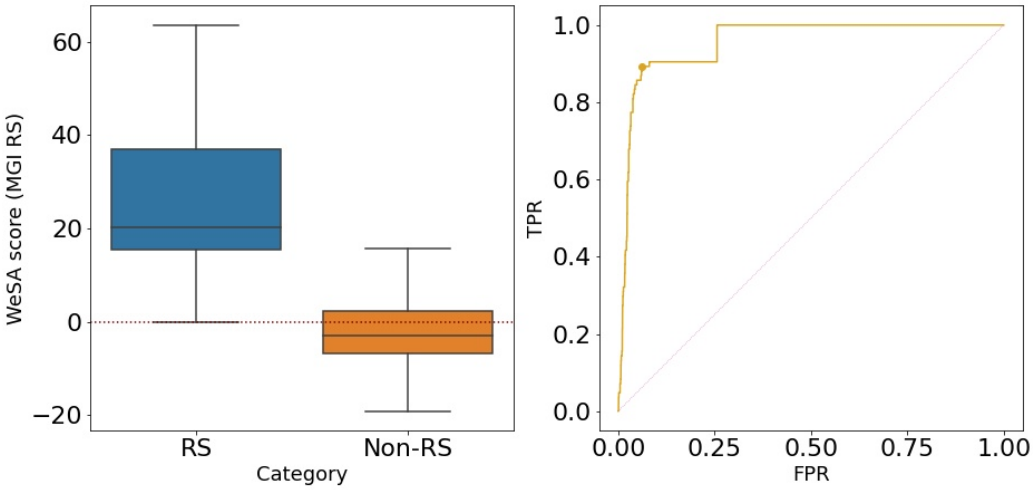


Figure 5.8: *Left panel:* Boxplots of the distributions of WeSA scores for links in MGI to the Reproductive system. On the x-axis the two groups of genes separated by their disease association (x-axis): either to diseases of the reproductive system (RS_D) or beyond (non-RS_D). *Right panel:* ROC curve based on the WeSA scores for MGI-RS associations. The benchmark are the genes which are in the RS_D category due to the diseases their human orthologs are lined to. The optimal threshold is marked as a point on the ROC curve and corresponds to TPR = 0.89 and FPR = 0.06.

RS_D category falls above the threshold (i.e. is FP), which corresponds to 59 genes due to the high numbers in the non-RS_D category. Hence, while the results comprise an encouraging pilot study, further testing is needed when more data becomes available.

As an extension, due to the low ratio of members in the benchmark group, compared to the total, the precision is still low (0.25). As we have discussed also in Chapter 4, what is currently labelled as false positives can still be related to the category of interest. Undiscovered gene-disease links also remain uncounted, in this case, weaker evidence from genes which are not solely or directly responsible for a disease in the RS_D can obscure the statistics. This is why precision in all cases is hard or impossible to measure and is likely an extremely conservative lower bound rather than an accurate estimate.

Gene-gene scores seem to correlate with similarity information from STRING

A natural extension to the bipartite scoring is to score the projected graph in which gene nodes are connected by weighted edges based on the shared system connections they have. Here we used the same method to calculate scores as presented in section 5.1

and 5.2.2.

In this final analysis, we investigate whether the gene-gene WeSA score is related to the similarity of genes. To test this hypothesis, we first calculated the gene-gene WeSA scores solely on the basis of shared phenotypes from MGI. Excluding those gene pairs that were linked only by a single shared phenotype (the large majority), we obtained a the scores for a total of 18,945 unique gene pairs and we compared those scores to the association scores from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) [63].

The STRING database is, to our knowledge, the most comprehensive database aggregating information about the association between proteins from multiple sources. There are seven sources of information which are referred to as channels (listed in table 5.4). One of the STRING features is the association score, which is a combination of the seven individual scores given for each channel. These scores measure the strength of evidence linking a protein pair: the higher the score, the stronger the evidence in support of an association.

The majority of evidence comes from physical interactions (i.e. experimental channel), which is not what we expect gene-gene scores to correlate the most with; the fact that they are responsible for phenotypes from the same system does not imply direct physical interaction. STRING also incorporates channels which may be closer correlating to the obtained WeSA scores. Specifically, out of the seven components contributing to the total association score[7] we believe the scores may correlate to pathways. That is why we look for strongest correlation with database scores which summarise curated evidence from pathway databases.

Given the still insufficient evidence and variability across sources, we should note that not all information channels are populated equally and indeed, fusion, co-occurrence and neighbourhood, which rely on orthologs and evolutionary data have scarce evidence. Overall, only 37% of the WeSA-scored gene pairs are found on STRING with the percentage sinking much lower for the evidence in some channels.

To measure correlation, we calculate the Pearson's correlation coefficient (Pearson's

[7] The seven channels are: 1) Neighbourhood, i.e. how frequently the proteins are located in the same vicinity on the genome; this is mainly useful for species with smaller genomes and less redundancy than *Mus musculus* and *Homo sapiens*; 2) Fusion score - estimated from the amount of fusion events, i.e. how often the proteins combined to form a single new protein; 3) co-occurrence - marks the tendency of orthologs to occur under similar conditions and distributions; 4) coexpression scores - obtained from microarray and RNAseq experiments and correspond to the correlation in expression of a pair of proteins across experiments; 5) literature mining; 6) experimental; 7) databases.

r). It is specifically a measure of strength of the association of the data to the line of best fit derived from the linear model. High absolute values of the Pearson's r correspond to strong correlation. The statistics are given in table 5.4.

Channel	Number of records	Pearson's r
neighbourhood	184	-0.137
fusion	12	-0.602
cooccurence	180	-0.149
coexpression	4372	0.09
experimental	3466	0.344
database	2356	0.351
textmining	6504	0.16
combined score	7006	0.454

Table 5.4: Statistics for the correlation of each STRING score with the WeSA score. The number of pairs which have both scores (WeSA and the respective STRING score) is given in the second column.

With the exception of the three channels neighbourhood, fusion and cooccurence, for which information is so limited that the STRING data is provided for less than 1% of data points, all other records show positive correlation between STRING scores and WeSA scores. Based on the Pearson's r, the combined score has a medium strong linear correlation to WeSA. The database scores are the most strongly correlated individual channel, but both the experimental channel and the database channel have medium correlation with the WeSA scores. Evidence from textmining which is recorded for more than a third of the gene pairs is also showing some weak positive correlation with the WeSA scores.

An argument can be made that, from a physical perspective, WeSA scores do not linearly correlate with the strength of association between gene pairs. Rather, there may be diminishing returns or a logarithmic slowing down of the growth of connection strength relative to the number of observed connections. Due to that argument, we also checked the difference between WeSA scores for pairs which score low or high on STRING. For all three channels and for the combined score for which positive correlation was found previously, there is also a good separation between the two categories. This is established using a MWU test and is illustrated through the boxplots in figure 5.9.

It is worth noting that STRING scores also have some predictive nature and information for them is incomplete. With expansion of research we may get more evidence in support of the WeSA score.

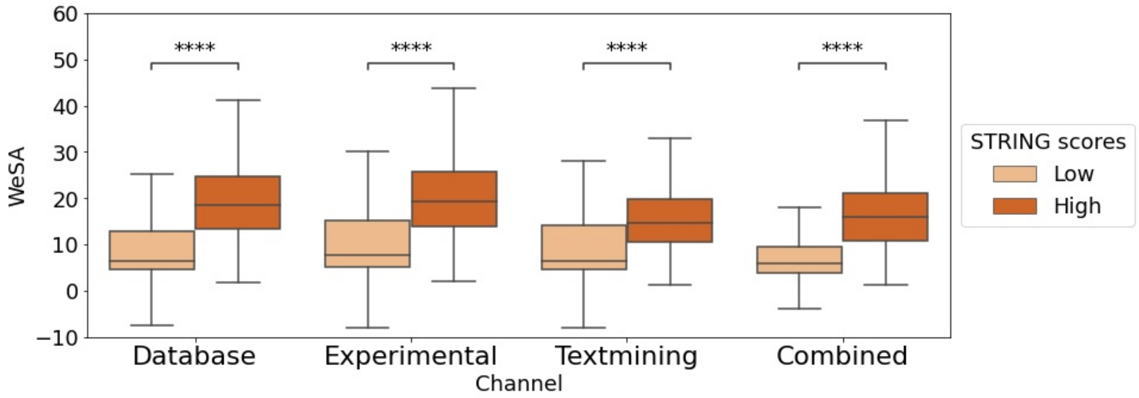


Figure 5.9: Boxplot of STRING scores for the two categories: high-scoring STRING gene-gene links (score more than 500 out of 1000) and low-scoring (score less than 500). Categories for the STRING score are on the x-axis. Comparisons are done using a MWU test and '****' corresponds to very significance $p < 0.0001$.

5.2.4 Discussion

In this section, we looked at mouse models which are crucial for understanding human systems and disease. As a simpler mammalian system, they are a main model organism [FR40]. Their relevance makes the analysis of MGI data valuable and here we present how WeSA can be useful to enrich the toolkit for MGI analysis.

What is still missing in most biological research, but also, particularly in mice studies is standardisation. Brown *et al.* [13] reviewed the progress made by the International Mouse Phenotyping Consortium (IMPC) towards the goal of generating a comprehensive resource for modified mice phenotype data. They underscore the impact which such database can have. However, the paper also highlights the importance of standardising the experimental protocols and phenotype inspection. In order for data to be analysable in aggregate and get the most precise results, there is a need for uniformity in procedures everywhere. However, we have seen here that WeSA can somewhat circumvent this obstacle.

WeSA scores weighting the relationship between genes and systems on MGI seem to correlate well to human diseases. We have seen that WeSA can usually focus the researcher's attention to the most important systems for the gene based on the known information. In cases where a gene is annotated to relate to many of the systems, such knowledge can navigate future research effectively.

Another implication of studies like Brown's [FR24, FR39, FR31, FR30, 13] is that phenotyping may be biased in various ways in different research. Despite all efforts, there are still many irreproducible results demanding more scrutiny of published discoveries. Our own hypothesis, which we also based on the observed numbers (fig. 5.7) is that genetic modifications are rarely annotated as neutral or having no effect. This are likely explained by inconsistent recording or overtesting. The former is meant to highlight that lack of effect is, in many cases, not recorded at all, so there are disproportionately more records on observed phenotypic changes. On the other hand, there is a possibility of overtesting, which is the search among all phenotypes, however insignificant, and probably finding something by random chance or because of the accumulated error of the multiple scenarios.

As mentioned, in the Reproductive system case study section, it is not surprising that most of those highly annotated genes have low WeSA scores; we expect that due to the bias from overtesting. We have seen, for example, that mortality phenotypes are common among the low-scorers. Some of those phenotypes are very likely to force death and could have very low baseline risk (or low survival) which means that the real-world importance of changes may be less important, while the errors due to the low numbers can be more frequent. This is among our results which confirm the overtesting bias and illustrate the power of WeSA in overcoming that bias.

Finally, we explored a possible connection between genes based on the number of shared systems they attach to via their annotated phenotypes. While we did find some positive correlation of those WeSA scores to the STRING similarity scores, there is room for improvement. Some further refinements may explore connecting the genes not based on shared systems, but based on a less general level of the MGI hierarchy. Depending on the goal, other test comparisons can also be chosen, while the STRING based ones can improve with the availability of more data. It still remains to be discovered if factors with very little information at present such as gene fusion are related to the WeSA scores we have presented here.

MGI studies allow us to gain insight into the function of genes and the phenotypical effects they have on mice, which can further aid understanding of human systems and human disease mechanisms. Thus, making sense of the information which is already available is important and WeSA provides us with an additional tool to do so.

6 Discussion and conclusion

This work has focused on analysis of biological networks with a particular focus on the human PPI network. A lot of effort is devoted to experiments which can speed-up the process of interaction discovery, but the use and analysis of their data is still suboptimal. Our goal has been to develop approaches to optimise both the study design and data collection step as well as the analysis itself.

In our first project we focused on improving the statistical pipeline when experiments are compared to a control. This analysis is essential as experiments can be conducted in a range of specifically designed conditions; they can examine PPIs in different cell types or different stages of the cell cycle, to name a few. Statistical comparison of target replicates to a carefully chosen, appropriate control can be very accurate in filtering out noise and detecting only the set of specific interactions.

The main challenge in such analysis after the obtaining of raw results is the presence of missing values. They often appear stochastically for both technical and biological reasons. We have adapted the major MV imputation method for MNAR missing values to sample the MVs from a custom distribution. The method alongside the full subsequent analysis and comparison is resented. In our affinity purification study of RAB7 our newer imputation method has performed better than the top MCAR comparison method SVDimpute both qualitatively and quantitatively in the PCA.

Case-control studies are not without difficulties in their design. There are two main difficulties which we highlight, namely, the problem with control choice and resources. In order for statistical analysis to be useful, the control for comparisons needs to be able to capture the same noise as the target, but without overlapping in any of the functional links. Moreover, the constant effort to both search for such controls and perform the control experiments is resource-intense and provides justification to any development which can reduce those costs.

The second part of the project, thus, disrupts the current experimental setup and develops a model function, WeSA, which weights the evidence for every PPI relative to past research instead of a specific control. We do not complicate data collection or reduce the reach of the analysis by incorporating any additional information. Rather, our model scales the raw number of multiple edges observed in databases according to their expectation. The framework can be incorporated as a preparation step before established clustering procedures such as MCL to convert the unweighted graph to a weighted one and improve the input of the clustering.

We find that the WeSA scoring procedure identifies protein complexes well. It performs better than the original SA method. With WeSA we can score whole datasets and reach TPR above 80% while keeping the FPR below 20% and obtaining precision better than 50% despite the lack of a well-defined negatives set. The good coverage of already existing sets is encouraging and bodes well for future applications of the framework more generally to future experiments (whether high-throughput or small-scale).

The score has been observed to possess additional benefits and to correlate with other reference sets. We have used it in predictions and suggest further predictions which could be made. Notably, however, we have found out that information can be added to already existing data and WeSA scores can be effectively updated allowing for instant feedback to researchers as well.

As WeSA is not specific to the context of PPI networks or even to biology, we have applied it to two other networks to test its universality. The two types of data we use are gene-trait associations from GWAS and gene-phenotype relationships from MGI. In both cases the relevant networks are bipartite and we propose additional modifications to the WeSA score reflecting the opportunities provided by bipartite networks. Specifically, we apply the score to weight a one-mode projection of the bipartite graph, thus deriving weighted gene-gene networks which can be clustered to uncover pathways or gene similarities.

In addition, the simple application of WeSA to the bipartite networks is also observed to work through different empirical comparisons. We have observed correlation to curated disease knowledge from literature and databases, including to human orthologs for the mouse data in MGI. We also confirmed some biologically sensible hypotheses, but many others can be formed in future work.

Science progresses quickly towards newer ways of data collection in all fields and much of that data comes in the form of intractable networks. We envision more applications of WeSA in the context of other big networks in biology and beyond. For instance, it can be applied to ChIP-Seq data to identify essential binding sites or as a filter to annotations in Gene Ontology.

A main challenge with such network filtering studies is the identification of a test set. As with our analysis on MGI, the design of tests requires the integration of external knowledge and sources, and our future efforts will also be directed towards the definition of a complete reference dataset of disease-system annotations. As genetic disease models, including mice models, attract research attention currently and effort is put towards the understanding of the causal genome-phenome relationships, gold-standard sets of disease relationships are crucial. A dataset of disease-system annotations, as mentioned previously, would allow immediate and unbiased performance evaluation beyond our MGI study and can be relevant to all methods working on the genome-phenome causality problem.

Finally, there are still various biases in all of the genetic experiments. In PPI discovery, these are introduced due to proteins being difficult to work with, impossible to detect or for any other reason overly popular or seldom studied. While newer research and high-throughput projects aim at universal coverage, it is not clear what the conditions of those are. In the case of proteins, are new proteins getting identified with the latest methods and at what rates is the coverage expansion happening? There are also open questions around the sufficiency of evidence: when will we know enough and how much is enough? Providing the answers to these fundamental theoretical questions can guide the efficient distribution of resources. Moreover, knowing what is necessary and sufficient for WeSA to have a beneficial impact on analysis can help expand the directions of data collection.

As we have proposed, it is possible to use the WeSA scores to make comparisons between networks observed under different conditions, such as networks varying over time or changing in response to variation. As long as there is sufficient input information to go into the model, such comparisons open up possibilities to understand the mechanisms of genetic disease, protection and beyond.

Appendix

CORUM

Table 6.1: Biggest and smallest human complexes from CORUM [59] alongside their respective sizes. The full data contains 3,538 different complexes.

Complex Name	Size
Spliceosome, E complex	129
Spliceosome, A complex	113
Nop56p-associated pre-rRNA complex	104
C complex spliceosome	80
55S ribosome, mitochondrial	78
39S ribosomal subunit, mitochondrial	48
60S ribosomal subunit, cytoplasmic	47
Respiratory chain complex I (holoenzyme), mitochondrial	44
Spliceosome, B complex	43
...	
REEP1-ZFYVE27 complex	2
DZIP1-IFT88 complex	2
TSC2-HERC1 complex	2
CEP164-DZIP1 complex	2
BCLAF1-TET2 complex	2

Python libraries used in WeSA script Main libraries:

- pandas
- numpy
- re
- scipy.stats

- collections
- itertools
- sklearn

For plotting:

- matplotlib
- seaborn
- mpl_toolkits
- statannot
- adjustText

Other libraries:

- pickle - for storage of dictionaries;
- timeit - for timing parts of the algorithm during testing
- networkx - for visualising graphs and/ or doing network analysis also beyond the section of protein-protein interaction analysis

I have additionally created a library of functions of my own which are useful on their own in different parts of the analysis, but some combine in the calculation of WeSA.

The matrix information benefits the WeSA score. Specifically, it improves the AUC by 17% (up from 0.72). Detailed statistics for the optimal threshold also show increase in TPR (up by 0.1 to 0.76) and decrease of FPR (from 0.34 to 0.23). Even the PR curve improves with precision at the threshold increasing by 42% of its level without the matrix term.

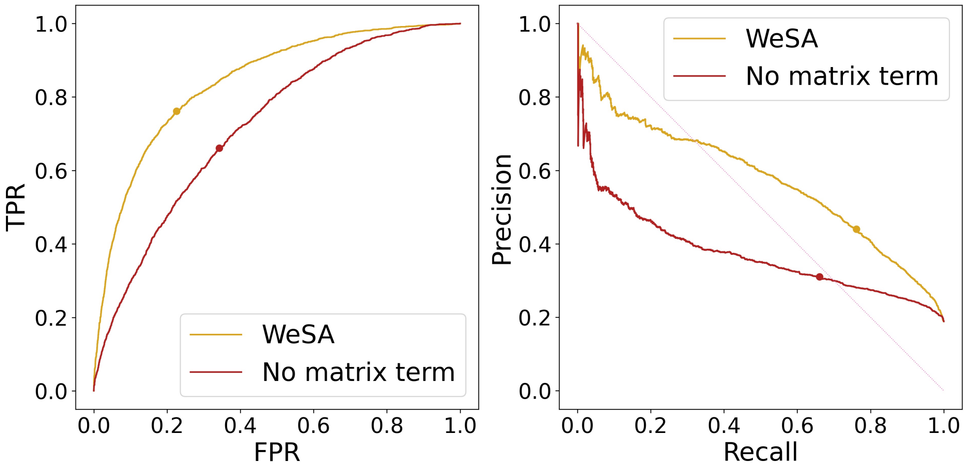


Figure 6.1: ROC (left) and PR curves (right) for the data in IntAct scored in two ways: 1) according to the WeSA score in equation 4.1; 2) based on only the spoke models, i.e. excluding the matrix term. The data in IntAct as cleaned in all other analysis is used to produce the plots.

Pseudocounts ROC analysis

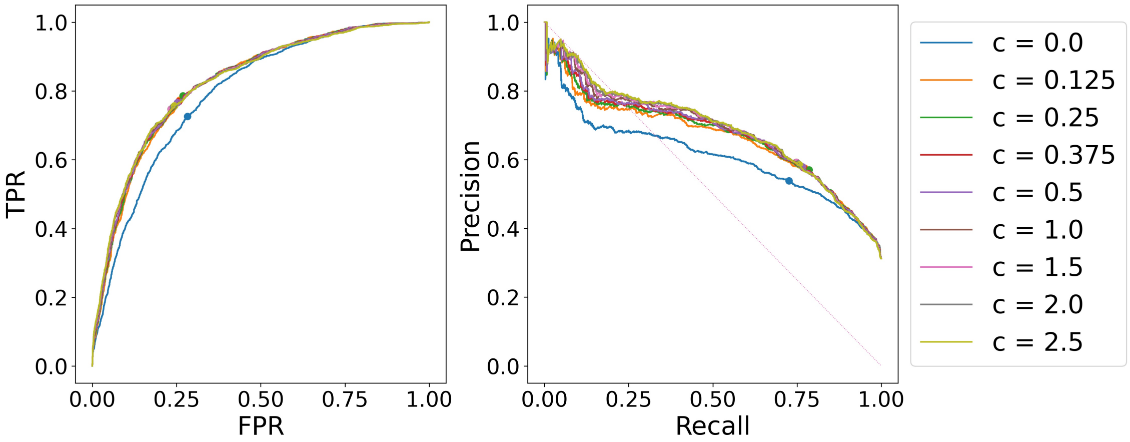


Figure 6.2: Results from ROC analysis of WeSA with varying pseudocounts $c \in$ {0,0.1,0.2,0.3,0.4,0.5,1,1.5,2,2.5}. ROC curves (left) and Precision-Recall curves (right) for the various values of the parameter c. The scores are weighted by experimental confidence and the data is taken from IntAct.

GWAS results

Table 6.2: Summary statistics of GWAS scores of links between traits and SNPs

	SA	WeSA
count	14,535	14,535
mean	2.1	4.65
std	2.17	5.12
min	-5.066	-38.76
25%	0.52	1.25
50%	1.86	4.29
75%	3.15	7.42
95%	6.22	13.47
99%	8.61	19.19
99.9%	10.29	28.06
max	10.69	34.1

Table 6.3: Top 20 traits with the most links in GWAS

Trait	Gene
Height	2225
Educational attainment	1956
Body mass index	1365
Smoking initiation	1170
White blood cell count	1020
Systolic blood pressure	1013
Heel bone mineral density	1003
Red blood cell count	969
Protein quantitative trait loci (liver)	947
Schizophrenia	942
Waist circumference adjusted for body mass index	942
Platelet count	926
Mean corpuscular hemoglobin	902
Type 2 diabetes	896
Waist-to-hip ratio adjusted for BMI	895
Metabolite levels	885
Hip circumference adjusted for BMI	875
Mean corpuscular volume	840
Total cholesterol levels	833
Blood protein levels	810

Table 6.4: Scores for all edges of variant rs738409

index	trait	variant	WeSA	SA	No. records
0	Nonalcoholic fatty liver disease	rs738409	34.1	3.79	9
1576	Percent liver fat	rs738409	10.74	5.37	2
5440	Liver enzyme levels (alanine transaminase)	rs738409	5.64	2.82	2
8288	Alanine aminotransferase levels	rs738409	3.53	0.71	5
8327	Total testosterone levels	rs738409	3.5	1.75	2
9294	Gout	rs738409	2.71	1.36	2
9513	Aspartate aminotransferase levels	rs738409	2.58	0.52	5
12557	Hemoglobin	rs738409	-0.09	-0.04	2
13353	Sex hormone-binding globulin levels	rs738409	-1.28	-0.64	2
13642	Hematocrit	rs738409	-1.88	-0.63	3
14218	Mean corpuscular hemoglobin	rs738409	-4.66	-1.55	3
14229	Triglyceride levels	rs738409	-4.77	-2.38	2
14342	Total cholesterol levels	rs738409	-6.18	-2.06	3
14370	Platelet count	rs738409	-6.52	-1.63	4

Table 6.5: Scores for all edges connecting to the trait Nonalcoholic fatty liver disease.

index	trait	variant	WeSA	SA	No. records
0	Nonalcoholic fatty liver disease	rs738409	34.1	3.79	9
1552	Nonalcoholic fatty liver disease	rs2143571	10.75	5.38	2
2697	Nonalcoholic fatty liver disease	rs73001065	8.73	4.36	2
5368	Nonalcoholic fatty liver disease	rs58542926	5.7	1.9	3
7461	Nonalcoholic fatty liver disease	rs1260326	4.17	1.04	4
8440	Nonalcoholic fatty liver disease	rs429358	3.41	1.14	3

Table 6.6: Scores for all edges of variant rs1047891

index	trait	variant	WeSA	SA	No. records
137	Glycine levels	rs1047891	19.23	2.14	9
383	Homoarginine levels	rs1047891	15.86	5.29	3
1416	Gamma-glutamylglycine levels	rs1047891	11.15	5.57	2
1417	N-palmitoylglycine levels	rs1047891	11.15	5.57	2
1586	N-acetylglycine levels	rs1047891	10.7	5.35	2
4639	Serine levels	rs1047891	6.24	3.12	2
9066	Urinary albumin-to-creatinine ratio	rs1047891	2.89	1.45	2
9276	Blood urea nitrogen levels	rs1047891	2.74	0.91	3
10839	Serum 25-Hydroxyvitamin D levels	rs1047891	1.31	0.65	2
11269	Estimated glomerular filtration rate (creatinine)	rs1047891	0.85	0.43	2
12334	Appendicular lean mass	rs1047891	0.11	0.06	2
12696	Apolipoprotein A1 levels	rs1047891	-0.28	-0.14	2
13012	Alanine aminotransferase levels	rs1047891	-0.64	-0.16	4
13311	Creatinine levels	rs1047891	-1.14	-0.38	3
13313	Estimated glomerular filtration rate	rs1047891	-1.15	-0.23	5
13868	Sex hormone-binding globulin levels	rs1047891	-2.62	-0.87	3
13955	HDL cholesterol	rs1047891	-2.97	-1.48	2
13993	Lymphocyte count	rs1047891	-3.12	-1.56	2
14009	Neutrophil count	rs1047891	-3.22	-1.61	2
14021	HDL cholesterol levels	rs1047891	-3.3	-0.83	4
14055	Red cell distribution width	rs1047891	-3.53	-1.76	2
14169	Mean platelet volume	rs1047891	-4.28	-1.43	3
14244	Red blood cell count	rs1047891	-4.91	-2.45	2
14428	White blood cell count	rs1047891	-8.26	-2.06	4
14432	Mean corpuscular hemoglobin	rs1047891	-8.42	-1.68	5
14439	Mean corpuscular volume	rs1047891	-8.88	-1.48	6
14475	Systolic blood pressure	rs1047891	-10.48	-2.62	4
14491	Platelet count	rs1047891	-11.98	-1.71	7

Table 6.7: Scores for all edges of variant rs603424

index	trait	variant	WeSA	SA	No. records
561	Myristoleate (14:1n5) levels	rs603424	14.39	7.2	2
562	Palmitoleate (16:1n7) levels	rs603424	14.39	7.2	2
563	5-dodecenoate (12:1n7) levels	rs603424	14.39	7.2	2
564	1-palmitoleoyl-GPC (16:1) levels	rs603424	14.39	7.2	2
1059	lysoPhosphatidylcholine acyl C16:1 levels	rs603424	12.19	6.1	2
1060	1-(1-enyl-palmitoyl)-2-palmitoleoyl-GPC (P-16:0/16:1) levels	rs603424	12.19	6.1	2
1735	Palmitoleic acid (16:1n-7) levels	rs603424	10.36	5.18	2
10515	Serum metabolite levels	rs603424	1.64	0.82	2
13065	Heel bone mineral density	rs603424	-0.78	-0.19	4
13446	Serum alkaline phosphatase levels	rs603424	-1.43	-0.72	2
13645	Lymphocyte count	rs603424	-1.89	-0.63	3
13839	Red cell distribution width	rs603424	-2.5	-0.83	3
13891	Mean platelet volume	rs603424	-2.71	-0.9	3
13978	Coronary artery disease	rs603424	-3.04	-1.01	3
14210	Diastolic blood pressure	rs603424	-4.63	-2.32	2
14250	Low density lipoprotein cholesterol levels	rs603424	-4.98	-1.66	3

Table 6.8: Scores for all edges linking to Colorectal cancer

index	trait	variant	WeSA	SA	No. records
2	Colorectal cancer	rs6983267	31.32	1.84	17
23	Colorectal cancer	rs3802842	25.36	2.54	10
47	Colorectal cancer	rs4939827	21.88	1.99	11
84	Colorectal cancer	rs704017	20.66	2.07	10
448	Colorectal cancer	rs12241008	15.21	2.54	6
449	Colorectal cancer	rs11196172	15.21	2.54	6
450	Colorectal cancer	rs3824999	15.21	2.54	6
451	Colorectal cancer	rs10411210	15.21	2.54	6
529	Colorectal cancer	rs6066825	14.59	2.08	7
943	Colorectal cancer	rs10774214	12.68	2.54	5
944	Colorectal cancer	rs7229639	12.68	2.54	5
945	Colorectal cancer	rs647161	12.68	2.54	5
1208	Colorectal cancer	rs1078643	11.77	2.35	5
1825	Colorectal cancer	rs2427308	10.14	2.54	4
2284	Colorectal cancer	rs3217810	9.25	2.31	4
2691	Colorectal cancer	rs10505477	8.74	1.75	5
2837	Colorectal cancer	rs11874392	8.52	2.13	4
2838	Colorectal cancer	rs1800469	8.52	2.13	4
3465	Colorectal cancer	rs113569514	7.61	2.54	3
3466	Colorectal cancer	rs73376930	7.61	2.54	3
3467	Colorectal cancer	rs7398375	7.61	2.54	3
3468	Colorectal cancer	rs12603526	7.61	2.54	3
3469	Colorectal cancer	rs2732875	7.61	2.54	3
4182	Colorectal cancer	rs17094983	6.74	2.25	3
5598	Colorectal cancer	rs10811654	5.53	1.84	3
5763	Colorectal cancer	rs10936599	5.43	1.36	4
6286	Colorectal cancer	rs11108175	5.07	2.54	2
6287	Colorectal cancer	rs10911251	5.07	2.54	2
6288	Colorectal cancer	rs61510274	5.07	2.54	2
6289	Colorectal cancer	rs4546885	5.07	2.54	2
6290	Colorectal cancer	rs12022676	5.07	2.54	2
6291	Colorectal cancer	rs812481	5.07	2.54	2
6292	Colorectal cancer	rs826732	5.07	2.54	2

6293	Colorectal cancer	rs4711689	5.07	2.54	2
6294	Colorectal cancer	rs73208120	5.07	2.54	2
6295	Colorectal cancer	rs77969132	5.07	2.54	2
6296	Colorectal cancer	rs77776598	5.07	2.54	2
6297	Colorectal cancer	rs4948317	5.07	2.54	2
6298	Colorectal cancer	rs201395236	5.07	2.54	2
6299	Colorectal cancer	rs12143541	5.07	2.54	2
6300	Colorectal cancer	rs9929218	5.07	2.54	2
6301	Colorectal cancer	rs992157	5.07	2.54	2
6302	Colorectal cancer	rs2070699	5.07	2.54	2
6303	Colorectal cancer	rs12412391	5.07	2.54	2
6304	Colorectal cancer	rs3217901	5.07	2.54	2
6305	Colorectal cancer	rs12818766	5.07	2.54	2
6799	Colorectal cancer	rs6584283	4.66	1.55	3
6800	Colorectal cancer	rs7014346	4.66	1.55	3
7336	Colorectal cancer	rs12979278	4.26	2.13	2
7338	Colorectal cancer	rs3217874	4.26	2.13	2
7339	Colorectal cancer	rs16878812	4.26	2.13	2
7340	Colorectal cancer	rs1741640	4.26	2.13	2
7342	Colorectal cancer	rs7226855	4.26	2.13	2
7344	Colorectal cancer	rs62404966	4.26	2.13	2
7609	Colorectal cancer	rs35107139	4.0	1.33	3
8024	Colorectal cancer	rs17816465	3.69	1.84	2
8025	Colorectal cancer	rs3830041	3.69	1.84	2
8026	Colorectal cancer	rs45597035	3.69	1.84	2
8027	Colorectal cancer	rs3087967	3.69	1.84	2
10962	Colorectal cancer	rs11692435	1.18	0.59	2
13945	Colorectal cancer	rs174537	-2.91	-0.58	5
14103	Colorectal cancer	rs3184504	-3.84	-1.92	2

Configuration model for the directed bait-prey network In the general case of a configuration model, it preserves a set degree distribution. To do so, graphically we can imagine that every node has “stubs” attached to it corresponding to the number of edges the node initially has. Those stubs are then connected in pairs to make edges. This does not exclude the possibility of self-loops and multiple edges, but it only preserves the degree distribution.

In the case of a directed graph as is this one, the “stubs” have to be adapted to correspond to incoming and outgoing edges. Then those should be connected to make a proper directed edge. That is, care should be taken to avoid connecting two outgoing halves or two incoming ones; two stubs can be connected only if one is outgoing and the other one is incoming.

Algorithms

Algorithm 5: Rejection sampling from $p(x)$

```
Data: $\omega$, $p(\omega)$
/* $\omega$ is a discrete subset of $\Omega$ (uniformly sampled)                */
Result: A sample with density $p(x)$
begin
    /* Initialisation:                                                        */
    interval = | $\omega$ |
    up_limit = max($p(\omega)$) /* I'll draw samples from discrete uniform with size $interval$,
        so $q(y) = 1/interval$.                                               */
    /* M $\geq p/q \Leftrightarrow$ M $\geq$ up_limit $\times$ interval       */
    M = interval $\times$ up_limit
    good_sample = FALSE /* Iterations (count = number of MVs):                */
    while good_sample == False do
        y = sample Uniformly($\omega$)
         u = sample U(0,1)
         accept_prob = p(y)/up_limit
         good_sample = (u < accept_prob)
    Return: y
```

Algorithm 6: kNN algorithm - pseudocode

```
Data: $n \times m$ table with LFQ values, $n =$ number of preys; $m =$ number of replicates
Result: A value is assigned to all MV-observations using kNN
begin
    /* Initialisation:                                                              */
    Load data  choose $k$ /* $k$ can be chosen by optimising the accuracy of predictions if
        we have (or define) a test set                                              */
    /* Iterations (count = number of MVs):                                          */
    for point in MVs do
        i. compute the Euclidean distance to all complete observations
        ii. record in a list and sort in ascending order
        iii. pick the k NN (observations with the smallest $k$ distances)
        iv. substitute the MV using mean substitution
    Return completed table
```

Algorithm 7: SVDimpute algorithm - pseudocode

```
Data: $n \times m$ matrix ($A$) with LFQ values, $n =$ number of preys; $m =$ number of replicates
    or conditions
Result: A value is assigned to all MV-observations using SVDimpute
begin
    /* Initialisation:                                                              */
    Load data in matrix $A$
     Record all MV indices in a list of tuples $L$
    /* Preprocessing                                                                */
    for point $(i, j)$ in $L$ do
        Substitute $A(i, j) = mean($row $i$ from $A)$
    /* Iterations: initialise a change variable and do iterations until they result in a
        change $> 0.01$ (or other threshold)                                        */
    while change > 0.01 do
        i. compute the SVD for $A = U\Sigma V^T$
            ii. find largest singular value, $\sigma$, and singular vectors, $u$ and $v$, corresponding
            to it
            /* The algorithm can also use the top $k$ singular values and corresponding
            vectors, then $u$ and $v$ are matrices of dimensions $n \times k$ and $k \times m$, respectively,
            and $\sigma$ is a diagonal $k \times k$ matrix.                          */
        iii. predict $A$ as $\tilde{A} = u \times \sigma \times v$
        iv. update $A$ by substituting in positions $L$ the entries from $\tilde{A}$
        v. calculate change due to substitution
    Return completed table
```

List of Acronyms

AF AlphaFold. 16

AF-M AlphaFold-Multimer. 70

AP Affinity Purification. 3

AUC Area Under the Curve. 52

BioGRID Biological General Repository for Interaction Datasets. 46

BMI Body-mass Index. 13

CI Confidence Interval. 81

CORUM Comprehensive Resource of Mammalian Protein Complexes. 46

DO Disease Ontology. 101

FDR False Discovery Rate. 37

FP False Positives. 67

FPR false positive rate. 52

GO Gene Ontology. 15

GWAS Genome-wide association studies. vi, 5

HuRI The Human Reference Protein Interactome Mapping Project. 15

i.i.d. independent identically distributed. 49

IBD Inflammatory Bowel Disease. 90

IFT Intraflagellar Transport. 46

IP Immunoprecipitation. 3

kNN k-Nearest Neighbours. 32

LD Linkage Disequilibrium. 94

LFQ Label-free Quantification. 20

MCAR Missing Completely at Random. 26

MCL Markov Clustering. 12

MGD Mouse Genome Database. 5

MGI Mouse Genome Informatics. x, 94

ML Machine Learning. 31

MNAR Missing not at Random. 26

MS Mass-spectrometry. 17

MV Missing Values. 3

MWU Mann-Whitney U test. 53

OMIM Online Mendelian Inheritance in Man. 101

OR Odds Ratio. 81

PCA Protein Complementation Assay. 7

PCA Principal Component Analysis. 3

PDB Protein Data Bank. 69

PPI Protein-Protein Interaction. 2

PR Precision-Recall. 53

PSM Peptide Spectral Measurements. 79

r.v. random variable. 49

RR Risk Ratio. 82

RS_D Reproductive System (disease-mapped). 102, 103

SA Socio-affinity. 13

SCilS European Training Network SCilS. 3

SNP Single-Nucleotide Polymorphism. 83

STRING Search Tool for the Retrieval of Interacting Genes/Proteins. 104

SVD Singular Value Decomposition. 3

TP True Positives. 67

TPR true positive rate. 52

WeSA Weighted Socio-affinity. 47

WT Wild-Type. 25

Y2H Yeast Two-Hybrid. 7

Glossary of Terms

adjacency matrix a matrix representation of a network. 11

Affinity Purification experimental protocol to identify all proteins sticking to a target; here, often also Strep-AP. 20

AP-like experiments aiming to determine at once the neighbourhood of a target bait protein; they use immobilised bait and from a supplemented lysate pull down the proteins which somehow directly or indirectly stick to the bait. 3

bait In affinity proteomics, normally, the protein to which a tag is attached (can also be “empty”). In other methods without a tag, a bait is the protein starting the experiment; for instance, in a BioID experiment that is the protein fused with the biotin ligase which can biotinylate the proteins which come in close proximity. 3

beads molecules which bind to a specific tag. 8

BioGRID a database recording experimental interaction data. 46

BioPlex a study of the human interactome using a single-tag protocol; release v.3.0 identifies more than 15 thousand proteins. 45

bipartite network graphs whose nodes are split into two disjoin partitions with no inner edges. 11

cell culture laboratory methods to grow cells. 3

cell line standard laboratory grown cells. 2

configuration model a model for constructing random graphs with a specified degree distribution. 4

CRAPome Published dataset of contaminants in proteomics experiments. 38

expression data information on proteins which are expressed (active) at the tested timepoint. 12

FPR proportion of negatives which were falsely classified as positives, FP/(FP+TN). 52

fusion protein modification, fusing peptides together. 3

Gene Ontology a database with annotations of proteins in three categories: molecular function, biological process and cellular component. 15

GWAS here, normally, a database containing variant-disease records from genome-wide association studies. 5

hub node network node with a lot of connections. 11

idempotent matrix matrix which is invariant to squaring. 12

IntAct a database recording experimental interaction data. 44

interactome the network of protein-protein interactions. 1

mass-spectrometry A method for protein/ peptide identification. 19

matrix indirect connections between pairs of prey proteins from the same AP experiment; modelled as a complete graph. 14, 44, 45

MGI, MGD a database containing mouse experimental records; specifically of interest are genotype-phenotype records. 5

one-mode projection graph the graph obtained by connecting nodes from just one partition of a bipartite graph if they share a neighbour. 11

optimal threshold threshold achieving best balance between TPR and FPR, here using the closest-to-(0,1) method. 52

pairwise study an experiment which is designed to test a specific single pair of proteins for direct interaction. 7

PDB a database of protein and protein-interaction structures derived from experiments. 69

Pearson's correlation coefficient (Pearson's r) a measure of strength of the association of the data to the line of best fit derived from the linear model. 104

precision true positives as a fraction of all predicted positives, TP/(TP+FP). 38

Precision-Recall curve a plot of recall (x-axis) and precision (y-axis) as the threshold varies. 53

prey In affinity proteomics, normally, the proteins which stick to the bait after the washing steps. In other methods relying on a protein change rather than proteins sticking together, such as biotinylation in BioID experiments, all changed (in the example, biotinylated) proteins are referred to as preys. 3

Principal Component Analysis Principal Component Analysis captures the directions (principal components) which capture the most of the variance of the data. 3

Protein Complementation Assay Experimental methods for testing a pairwise interaction using split signal proteins. 7

pseudocounts additional artificial counts. 80

pull-down a pairwise experiment working with purified protein. 8

recall equal to TPR, TP/(TP+FN). 38

ROC curve A plot of TPR versus FPR under a varying threshold. 52

scale-free the property of a graph degree distribution to follow a power-law. 11

single cell analysis allows for the examination of a single cell and subsequent cell-cell comparison. 9

spoke connections between prey proteins and the bait in AP experiments; modelled as a star graph. 44, 45

Strep-AP Affinity purification with a single Strep tag. 17

STRING A database of protein connections derived from 7 different channels: experimental, databases, neighbourhood, fusion, co-occurrence, coexpression, literature mining.. 104

supernatant unbound protein obtained from a washing step in IP experiment. 19

SVDimpute a method for MV imputation which uses SVD to consecutively decompose the initial matrix and, then, predict an approximation based on a certain amount of the top singular values and their corresponding singular vectors. 32

system (MGI) phenotypic classifications from the MGI ontology of phenotypes; it is the first level of diversification in the ontology. 97

tag special peptide known to bind to particular molecules called beads. 8, 9

TPR see recall. 52

trait here, disease or feature annotations from GWAS. 84

transcription bursting interruptions of the steady Poisson-modelled process of mRNA transcription. 9

transfection lab. introduction of protein in the cell. 3

Yeast Two-Hybrid Experimental system for testing pairwise interactions in yeast. 7

Bibliography

[1] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422:198–207, 04 2003.

[2] Monica Agrawal, Marinka Zitnik, and Jure Leskovec. Large-scale analysis of disease pathways in the human interactome. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 23:111–122, 01 2018.

[3] Daniela Albrecht-Eckardt, Olaf Kniemeyer, Axel Brakhage, and Reinhard Guthke. Missing values in gel-based proteomics. *Proteomics*, 10:1202–11, 03 2010.

[4] Patrick Aloy and Robert Russell. The third dimension for protein interactions and complexes. *Trends in biochemical sciences*, 27:633–8, 01 2003.

[5] Janine Altmüller, Lyle Palmer, Guido Fischer, Hagen Scherb, and Matthias Wjst. Genomewide scans of complex human diseases: True linkage is hard to find. *American journal of human genetics*, 69:936–50, 12 2001.

[6] Josh Backman, Alexander Li, Anthony Marcketta, Dylan Sun, Joelle Mbatchou, Michael Kessler, Christian Benner, Daren Liu, Adam Locke, s Balasubramanian, Ashish Yadav, Nilanjana Banerjee, Christopher Gillies, Amy Damask, Simon Liu, Xiaodong Bai, Alicia Hawes, Evan Maxwell, Lauren Gurski, and Manuel Ferreira. Exome sequencing and analysis of 454,787 uk biobank participants. *Nature*, 599, 11 2021.

[7] Gustavo E. A. P. A. Batista and Maria Carolina Monard. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5-6):519–533, 2003.

[8] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57:289–300, 1 1995.

[9] Helen Berman, John Westbrook, Zukang Feng, Helge Weissig, Ilya Shindyalov, and Philip Bourne. The protein data bank. *Nucleic Acids Research*, 28, 08 2000.

[10] Tina Beyer, Franziska Klose, Anna Kuret, Felix Hoffmann, Robert Lukowski, Marius Ueffing, and Karsten Boldt. Tissue- and isoform-specific protein complex analysis with natively processed bait proteinse. *Journal of Proteomics*, 231:103947, 08 2020.

[11] Philipp Blohm, Goar Frishman, Smialowski Pawel, Florian Goebels, Benedikt Wachinger, Andreas Ruepp, and Dmitrij Frishman. Negatome 2.0: A database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic acids research*, 42, 11 2013.

[12] Karsten Boldt, Jeroen Reeuwijk, Qianhao Lu, Konstantinos Koutroumpas, Thanh-Minh Nguyen, Yves Texier, Sylvia van Beersum, Nicola Horn, Jason Willer, Dorus Mans, Gerard Dougherty, Ideke Lamers, Karlien Coene, Heleen Arts, Matthew Betts, Tina Beyer, Emine Bolat, Christian Gloeckner, Khatera Haidari, Robert B Russell, and Ronald Roepman. An organelle-specific protein landscape identifies novel diseases and molecular mechanisms. *Nature communications*, 7:11491, 05 2016.

[13] Steve D M Brown and Mark W Moore. The international mouse phenotyping consortium: past and future perspectives on mouse phenotyping. *Mammalian genome*, 23(9-10):632–640, 2012.

[14] Anna Brückner, Cécile Polge, Nicolas Lentze, Daniel Auerbach, and Uwe Schlattner. Yeast Two-Hybrid, a Powerful Tool for Systems Biology. *Int J Mol Sci.*, 10(6):2763–2788, 2009.

[15] S. F. Buck. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society: Series B (Methodological)*, 22:302–306, 7 1960.

[16] Mark Burgin. Theory of fuzzy limits. *Fuzzy Sets and Systems*, 115(3):433–443, 2000.

[17] Mengfei Cao, Hao Zhang, Jisoo Park, Noah Daniels, Mark Crovella, Lenore Cowen, and Benjamin Hescott. Going the distance for protein function prediction: A new distance metric for protein interaction networks. *PloS one*, 8:e76339, 10 2013.

[18] Hyunghoon Cho, Bonnie Berger, and Jian Peng. Compact integration of multi-network topology for functional analysis of genes. *Cell Systems*, 3, 11 2016.

[19] Timothy Clough, Safia Thaminy, Susanne Ragg, Ruedi Aebersold, and Olga Vitek. Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs. *BMC bioinformatics*, 13 Suppl 16:S6, 11 2012.

[20] Sean Collins, Patrick Kemmeren, Xue-Chu Zhao, Jack Greenblatt, Forrest Spencer, Frank Holstege, Jonathan Weissman, and Nevan Krogan. Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae. *Molecular & cellular proteomics : MCP*, 6:439–50, 04 2007.

[21] T. Cover and P. Hart. Nearest neighbor pattern classification, 1967.

[22] Juergen Cox and Matthias Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26:1367–72, 12 2008.

[23] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[24] Nils Eling, Michael Morgan, and John Marioni. Challenges in measuring and understanding biological noise. *Nature Reviews Genetics*, 20, 05 2019.

[25] Janan Eppig, Joel Richardson, James Kadin, Martin Ringwald, Judith Blake, and Carol Bult. Mouse Genome Informatics (MGI): reflecting on 25 years. *Mammalian genome : official journal of the International Mammalian Genome Society*, 26, 08 2015.

[26] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Timothy Green, Augustin Žídek, Russell Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, and Demis Hassabis. Protein complex prediction with AlphaFold-Multimer. 10 2021.

[27] Siebren Faber, Stef J. F. Letteboer, Katrin Junger, Rossano Butcher, Trinadh V. Satish Tammana, Sylvia E. C. van Beersum, Marius Ueffing, Rob W. J. Collin, Qin Liu, Karsten Boldt, and Ronald Roepman. PDE6D Mediates Trafficking of Prenylated Proteins NIM1K and UBL3 to Primary Cilia. *Cells*, 12(2), 2023.

[28] Franke, A and McGovern, D and Barrett, J and Wang, Kai and Radford-Smith, G and Ahmad, Tariq and Lees, C and Balschun, Tobias and Lee, James and Roberts, R and Anderson, Carl and Bis, J and Bumpstead, S and Ellinghaus, David and Festen, Eleonora and Georges, Mirlande and Green, T and Haritunians, Talin and Jostins, L and Parkes, M. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet*, 42:1118–1125, 2010.

[29] Xiangchao Gan, Alan Wee-Chung Liew, and Hong Yan. Microarray missing data imputation based on a set theoretic framework and biological knowledge. *Nucleic acids research*, 34:1608–19, 02 2006.

[30] Anne-Claude Gavin, Patrick Aloy, Paola Grandi, Roland Krause, Markus Boesche, Martina Marzioch, Christina Rau, Lars Jensen, Sonja Bastuck, Birgit Dümpelfeld, Angela Edelmann, Marie-Anne Heurtier, Verena Hoffman, Christian Hoefert, Karin Klein, Manuela Hudak, Anne-Marie Michon, Malgorzata Schelder, Markus Schirle, Marita Remor, Tatjana Rudi, Sean Hooper, Andreas Bauer, Tewis Bouwmeester, Georg Casari, Gerard Drewes, Gitte Neubauer, Jens M Rick, Bernhard Kuster, Peer Bork, Robert B Russell, and Giulio Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–6, 04 2006.

[31] Madalina Giurgiu, Julian Reinhard, Barbara Brauner, Irmtraud Dunger, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic acids research*, 47, 10 2018.

[32] Christian Gloeckner, Karsten Boldt, Annette Schumacher, and Marius Ueffing. Tandem Affinity Purification of Protein Complexes from Mammalian Cells by the Strep/FLAG (SF)-TAP Tag. *Methods in molecular biology (Clifton, N.J.)*, 564:359–72, 02 2009.

[33] Jonathan Haines, Michael Hauser, Silke Schmidt, William Scott, Lana Olson, Paul Gallins, Kylee Spencer, Shu Kwan, Maher Noureddine, John Gilbert, Nathalie Schnetz-Boutaud, Anita Agarwal, Eric Postel, and Margaret Pericak-Vance. Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration. *Science (New York, N.Y.)*, 308:419–21, 05 2005.

[34] Felix Hoffmann, Sylvia Bolz, Katrin Junger, Franziska Klose, Timm Schubert, Franziska Woerz, Karsten Boldt, Marius Ueffing, and Tina Beyer. TTC30A and

TTC30B Redundancy Protects IFT Complex B Integrity and Its Pivotal Role in Ciliogenesis. *Genes*, 13(7), 2022.

[35] Xiaopeng Hu, Sohail Malik, Costin Negroiu, Kyle Hubbard, Chidambaram Velalar, Brian Hampton, Dan Grosu, Jennifer Catalano, Robert Roeder, and Averell Gnatt. A Mediator-responsive form of metazoan RNA polymerase II. *Proceedings of the National Academy of Sciences of the United States of America*, 103:9506–11, 07 2006.

[36] Nina Hubner, Alexander Bird, Juergen Cox, Bianca Splettstoesser, Peter Bandilla, Ina Poser, Anthony Hyman, and Matthias Mann. Quantitative proteomics combined with bac transgeneomics reveals. *The Journal of cell biology*, 189:739–54, 05 2010.

[37] Edward Huttlin, Raphael Bruckner, José Navarrete-Perea, Joe Cannon, Kurt Baltier, Fana Gebreab, Melanie Gygi, Alexandra Thornock, Gabriela Zárraga, Stanley Tam, John Szpyt, Alexandra Panov, Hannah Parzen, Sipei Fu, Arvene Golbazi, Eila Maenpaa, Keegan Stricker, Sanjukta Thakurta, Ramin Rad, and Steven Gygi. Dual Proteome-scale Networks Reveal Cell-specific Remodeling of the Human Interactome. 01 2020.

[38] Sebastian Jäger, Arndt Allhorn, and Felix Biessmann. A benchmark for data imputation methods. *Frontiers in Big Data*, 4, 07 2021.

[39] Celia Jeronimo, Marie-France Langelier, Mahel Zeghouf, Marilena Cojocaru, Dominique Bergeron, Dania Baali, Diane Forget, Sanie Mnaimneh, Armaity Davierwala, Jeff Pootoolal, Mark Chandy, Veronica Canadien, Bryan Beattie, Dawn Richards, Jerry Workman, Timothy Hughes, Jack Greenblatt, and Benoit Coulombe. RPAP1, a Novel Human RNA Polymerase II-Associated Protein Affinity Purified with Recombinant Wild-Type and Mutated Polymerase Subunits. *Molecular and cellular biology*, 24:7043–58, 09 2004.

[40] Da Jia, Timothy Gomez, Zoltan Metlagel, Junko Umetani, Zbyszek Otwinowski, Michael Rosen, and Daniel Billadeau. WASH and WAVE actin regulators of the Wiskott–Aldrich syndrome protein (WASP) family are controlled by analogous structurally related complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 107:10442–7, 06 2010.

[41] Liang Jin, Yingtao Bi, Chenqi Hu, Jun Qu, Shichen Shen, Xue Wang, and Yu Tian. A comparative study of evaluating missing value imputation methods in label-free proteomics. *Scientific Reports*, 11:1760, 1 2021.

[42] Yuliya V Karpievitch, Alan R Dabney, and Richard D Smith. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics*, 13:S5, 11 2012.

[43] Eric Kershnar, Szu-Yuan Wu, and Cheng-Ming Chiang. Immunoaffinity Purification and Functional Characterization of Human Transcription Factor IIH and RNA Polymerase II from Clonal Cell Lines That Conditionally Express Epitope-tagged Subunits of the Multiprotein Complexes. *The Journal of biological chemistry*, 273:34444–53, 01 1999.

[44] Alastair King, Marc Arnone, Maureen Bleam, Katherine Moss, Jingsong Yang, Kelly Fedorowicz, Kimberly Smitheman, Joseph Erhardt, Angela Hughes-Earle, Laurie Kane-Carson, Robert Sinnamon, Hongwei Qi, Tara Rheault, David Uehling, and Sylvie Laquerre. Dabrafenib; Preclinical Characterization, Increased Efficacy when Combined with Trametinib, while BRAF/MEK Tool Combination Reduced Skin Lesions. *PloS one*, 8:e67583, 07 2013.

[45] Nevan Krogan, Gerard Cagney, Haiyuan Yu, Gouqing Zhong, Xinghua Guo, Alexandr Ignatchenko, Joyce Li, Shuye Pu, Nira Datta, Aaron Tikuisis, Thanuja Punna, José Peregrín-Alvarez, Michael Shales, Xin Zhang, Mike Davey, Mark Robinson, Alberto Paccanaro, James Bray, Anthony Sheung, and Jack Greenblatt. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440:637–643, 04 2006.

[46] Cosmin Lazar, Laurent Gatto, Myriam Ferro, Christophe Bruley, and Thomas Burger. Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *Journal of Proteome Research*, 15:1116–1125, 4 2016.

[47] Xinhua Liu. Classification accuracy and cut point?selection. *Statistics in medicine*, 31:2676–86, 10 2012.

[48] Haiying Lu, Qiaodan Zhou, Jun He, Zhongliang Jiang, Cheng Peng, Rongsheng Tong, and Jianyou Shi. Recent advances in the development of protein-protein interactions modulators: mechanisms and clinical trials. *Signal transduction and targeted therapy*, 5:213, 09 2020.

[49] Kim-DK. Lambourne L. et al. Luck, K. A reference map of the human binary protein interactome. *Nature*, 580:402–408, 2020.

[50] Dattatreya Mellacheruvu, Zach Wright, Amber Couzens, Jean-Philippe Lambert, Nicole St-Denis, Tuo Li, Yana Miteva, Simon Hauri, Mihaela Sardiu, Teck Low, Vincentius Halim, Richard Bagshaw, Nina Hubner, Abdallah Al-Hakim, Annie Bouchard, Denis Faubert, Damian Fermin, Wade Dunham, Marilyn Goudreault, and Alexey Nesvizhskii. The CRAPome: a Contaminant Repository for Affinity Purification Mass Spectrometry Data. *Nature methods*, 10:730–6, 08 2013.

[51] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[52] Sandra Orchard, Mohamed Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy Campbell, Gayatri Chavali, Carol Chen, Noemi Del Toro Ayllón, Margaret Duesbury, Marine Sivade, Eugenia Galeota, Ursula Hinz, Marta Iannuccelli, Sruthi Jagannathan, Rafael Jimenez, Jyoti Khadake, Astrid Lagreid, and Henning Hermjakob. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42, 11 2013.

[53] Oughtred, Rose and Rust, Jennifer and Chang, Christie and Breitkreutz, Bobby-Joe and Stark, Chris and Willems, Andrew and Boucher, Lorrie and Leung, Genie and Kolas, Nadine and Zhang, Frederick and Dolma, Sonam and Coulombe-Huntington, Jasmin and Chatr-aryamontri, Andrew and Dolinski, Kara and Tyers, Mike. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic and chemical interactions. *Protein Science*, 30, 10 2020.

[54] Margaret Sullivan Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*, volume 28. Oxford University Press, Oxford, UK, 2003.

[55] Narcis Petriman, Marta Loureiro-López, Michael Taschner, Nevin Zacharia, Magdalena Georgieva, Niels Boegholm, Jiaolong Wang, Andre Mourao, Robert Russell, Jens Andersen, and Esben Lorentzen. Biochemically validated structural model of the 15-subunit intraflagellar transport complex IFT-B. *The EMBO Journal*, 41, 11 2022.

[56] Aldamaria Puliti, Gianluca Caridi, Roberto Ravazzolo, and Gian Ghiggeri. Teaching molecular genetics: Chapter 4 - positional cloning of genetic disorders. *Pediatric nephrology (Berlin, Germany)*, 22:2023–9, 01 2008.

[57] Ewan Ramsay, Guillermo Abascal-Palacios, Julia Daiß, Helen King, Jerome Gouge, Michael Pilsl, Fabienne Beuron, Edward Morris, Philip Gunkel, Christoph

Engel, and Alessandro Vannini. Structure of human RNA polymerase III. *Nature Communications*, 11, 12 2020.

[58] Thomas Rolland, Murat Taşan, Benoit Charloteaux, Samuel Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, Irma Lemmens, Celia Fontanillo, Roberto Mosca, Atanas Kamburov, Susan Ghiassian, Xinping Yang, Lila Ghamsari, Dawit Balcha, Bridget Begg, Pascal Falter-Braun, Marc Brehme, Martin Broly, and Marc Vidal. A Proteome-Scale Map of the Human Interactome Network. *Cell*, 159:1212–1226, 11 2014.

[59] Andreas Ruepp, Corinna Montrone, Barbara Brauner, Gisela Fobo, Goar Frishman, Madalina Giurgiu, and George Tsitsiridis. Corum: the comprehensive resource of mammalian protein complexes–2022. *Nucleic Acids Research*, 11 2022.

[60] Sven-Eric Schelhorn, Julián Mestre, Mario Albrecht, and Elena Zotenko. Inferring Physical Protein Contacts from Large-Scale Purification Data of Protein Complexes. *Molecular & cellular proteomics : MCP*, 10:M110.004929, 2011.

[61] Lynn Schriml, Richard Lichenstein, Katharine Bisordi, Cynthia Bearer, J. Allen Baron, and Carol Greene. Modeling the enigma of complex disease etiology. *Journal of Translational Medicine*, 21:148, 02 2023.

[62] Elliot Sollis, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, Tudor Groza, Osman Güneş, Peggy Hall, James Hayhurst, Arwa Ibrahim, Yue Ji, Sajo John, Elizabeth Lewis, Jacqueline MacArthur, Aoife McMahon, David Osumi-Sutherland, Kalliope Panoutsopoulou, Zoë Pendlington, and Laura W Harris. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51, 11 2022.

[63] Damian Szklarczyk, Annika L. Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T. Doncheva, John H. Morris, Peer Bork, Lars J. Jensen, and Christian von Mering. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 2019.

[64] R Tibshirani, J Parker, T Hastie, JS Marron, A Nobel, S Deng, H Johnsen, R Pesich, S Geisler, J Demeter, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci US A*, 100(14):8418–23, 2003.

[65] Olga Troyanskaya, Mike Cantor, Gavin Sherlock, Trevor Hastie, Rob Tibshirani, David Botstein, and Russ Altman. Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics*, 17:520–525, 07 2001.

[66] Alexander Tuttle, Vivek Philip, Elissa Chesler, and Jeffrey Mogil. Comparing phenotypic variation between inbred and outbred mice. *Nature Methods*, 15, 11 2018.

[67] Stefka Tyanova, Tikira Temu, Pavel Sinitcyn, Arthur Carlson, Marco Hein, Tamar Geiger, Matthias Mann, and Juergen Cox. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods*, 13, 06 2016.

[68] Ilker Unal. Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach. *Computational and Mathematical Methods in Medicine*, 2017:1–14, 05 2017.

[69] Stijn van Dongen. *Graph clustering by flow simulation*. PhD thesis, University of Utrecht, 2000.

[70] Guang Wang, Huai-Bin Hu, Yan Chang, Yan Huang, Zeng-Qing Song, Shi-Bo Zhou, Liang Chen, Yu-Cheng Zhang, Min Wu, Hai-Qing Tu, Jin-Feng Yuan, Na Wang, Xin Pan, Ai-Ling Li, Tao Zhou, Xue-Min Zhang, Kun He, and Hui-Yan Li. Rab7 regulates primary cilia disassembly through cilia excision. *The Journal of Cell Biology*, 218:jcb.201811136, 12 2019.

[71] Bobbie-Jo M. Webb-Robertson, Holli K. Wiberg, Melissa M. Matzke, Joseph N. Brown, Jing Wang, Jason E. McDermott, Richard D. Smith, Karin D. Rodland, Thomas O. Metz, Joel G. Pounds, and Katrina M. Waters. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *Journal of Proteome Research*, 14(5):1993–2001, 2015. PMID: 25855118.

[72] Guanming Wu, Xin Feng, and L Stein. A human functional protein interaction network and its application to cancer data analysis. *Genome biology*, 11:R53, 05 2010.

[73] Walter J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

[74] Javad Zahiri, Ali Emamjomeh, Samaneh Bagheri, Asma Ivazeh, Ghasem Mahdevar, Hessam Sepasi Tehrani, Mehdi Mirzaie, Barat Fakheri, and Morteza Mohammad-Noori. Protein complex prediction: A survey. *Genomics*, 112, 01 2019.

[75] Xinshan Zhu, Jiayu Wang, Biao Sun, Chao Ren, Ting Yang, and Jie Ding. An efficient ensemble method for missing value imputation in microarray gene expression data. *BMC Bioinformatics*, 22, 04 2021.

[76] Julian Zielenski and Lap-Chee Tsui. Cystic fibrosis: Genotypic and phenotypic variations. *Annual review of genetics*, 29:777–807, 02 1995.

Further references

[FR1] Juan S. Bonifacino, David C. Gershlick, and Esteban C. Dell'Angelica. Immunoprecipitation. *Current Protocols in Cell Biology*, 71, 6 2016.

[FR2] Georg Borner, Marco Hein, Jennifer Hirst, James Edgar, Matthias Mann, and Margaret Robinson. Fractionation profiling: A fast and versatile approach for mapping vesicle proteomes and protein-protein interactions. *Molecular biology of the cell*, 25, 08 2014.

[FR3] Tess Branon, Justin Bosch, Ariana Sanchez, Namrata Udeshi, Tanya Svinkina, Steven Carr, Jessica Feldman, Norbert Perrimon, and Alice Ting. Efficient proximity labeling in living cells and organisms with TurboID. *Nature Biotechnology*, 36, 10 2018.

[FR4] Steve D M Brown and Mark W Moore. The international mouse phenotyping consortium: past and future perspectives on mouse phenotyping. *Mammalian genome*, 23(9-10):632–640, 2012.

[FR5] David Burke, Patrick Bryant, Inigo Barrio-Hernandez, Danish Memon, Gabriele Pozzati, Aditi Shenoy, Wensi Zhu, Alistair Dunham, Pascal Albanese, Andrew Keller, Richard Scheltema, James Bruce, Alexander Leitner, Petras Kundrotas, Pedro Beltrao, and Arne Elofsson. Towards a structurally resolved human protein interaction network. *Nature Structural & Molecular Biology*, 30:1–10, 01 2023.

[FR6] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science (New York, N.Y.)*, 321:263–6, 08 2008.

[FR7] Baoqing Chen, Mihnea Dragomir, Linda Fabris, Recep Bayraktar, Erik Knutsen, Xu Liu, Changyan Tang, Yongfeng Li, Tadanobu Shimura, Tina Catela Ivkovic, Mireia Cruz De Los Santos, Simone Anfossi, Masayoshi

Shimizu, Maitri Shah, Hui Ling, Peng Shen, Asha Multani, Barbara Pardini, Jared Burks, and George Calin. The Long Noncoding RNA CCAT2 induces chromosomal instability through BOP1 - AURKB signaling. *Gastroenterology*, 159, 08 2020.

[FR8] Joel Cohen. Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS biology*, 2:e439, 01 2005.

[FR9] Thomas Corwin, Jonathan Woodsmith, Federico Apelt, Miguel Andrade, Bryan Ballif, Ulrich Correspondence, Jean-Fred Fontaine, David Meierhofer, Johannes Helmuth, Arndt Grossmann, and Ulrich Stelzl. Defining Human Tyrosine Kinase Phosphorylation Networks Using Yeast as an In Vivo Model Substrate. *cell system*, 5:128–139, 08 2017.

[FR10] Juergen Cox, Ivan Matic, Maximiliane Hilger, Nagarjuna Nagaraj, Matthias Selbach, Jesper Olsen, and Matthias Mann. A practical guide to the maxquant computational platform for silac-based quantitative proteomics. *Nature protocols*, 4:698–705, 02 2009.

[FR11] Taylor Dolberg, Anthony Meger, Jonathan Boucher, William Corcoran, Elizabeth Schauer, Alexis Prybutok, Srivatsan Raman, and Joshua Leonard. Computation-guided optimization of split protein systems. *Nature Chemical Biology*, 17:1–9, 05 2021.

[FR12] Christine Engeland, Johannes Heidbuechel, Robyn Araujo, and Adrianne Jenner. Improving immunovirotherapies: the intersection of mathematical modelling and experiments. *ImmunoInformatics*, 6:100011, 06 2022.

[FR13] Neil Ferguson, Christl Donnelly, and Roy Anderson. The foot-and-mouth epidemic in great britain: Pattern of spread and impact of interventions. *Science (New York, N.Y.)*, 292:1155–60, 06 2001.

[FR14] Leonard Foster, Carmen Hoog, Yanling Zhang, Yong Zhang, Xiaohui Xie, Vamsi Mootha, and Matthias Mann. A mammalian organelle map by protein correlation profiling. *Cell*, 125:187–99, 05 2006.

[FR15] Ellen Goode and Gail Jarvik. Assessment and implications of linkage disequilibrium in genome-wide single-nucleotide polymorphism and microsatellite panels. *Genetic epidemiology*, 29 Suppl 1:S72–6, 01 2005.

[FR16] Andrew Grant, Dipender Gill, Paul Kirk, and Stephen Burgess. Noiseaugmented directional clustering of genetic association data identifies distinct mechanisms underlying obesity. *PLOS Genetics*, 18:e1009975, 01 2022.

[FR17] R L Grant. Converting an odds ratio to a range of plausible relative risks for better communication of research findings. *BMJ*, 348:f7450, 2014.

[FR18] Thomas Gridley and Stephen Murray. *Mouse mutagenesis and phenotyping to generate models of development and disease*, volume 148. 03 2022.

[FR19] Alan Hastings and Margaret Palmer. Mathematics and biology. a bright future for biologists and mathematicians? *Science (New York, N.Y.)*, 299:2003–4, 04 2003.

[FR20] Lukas Heumos, Anna Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte Lücken, Daniel Strobl, Juan Henao, Fabiola Curion, Herbert Schiller, Fabian Theis, and Christian Müller. Best practices for single-cell analysis across modalities. *Nature reviews. Genetics*, 03 2023.

[FR21] Yuen Ho, Albrecht Gruhler, Adrian Heilbut, Gary Bader, Lynda Moore, Sally-Lin Adams, Anna Millar, Paul Taylor, Keiryn Bennett, Kelly Boutilier, Lingyun Yang, Cheryl Wolting, Ian Donaldson, Soeren Schandorff, Juanita Shewnarane, Mai Vo, Joanne Taggart, Marilyn Goudreault, Brenda Muskat, and Mike Tyers. Systematic Identification of Protein Complexes in Saccharomyces cerevisiae by Mass Spectrometry. *Nature*, 415:180–3, 02 2002.

[FR22] Won-Ki Huh, James Falvo, Luke Gerke, Adam Carroll, Russell Howson, Jonathan Weissman, and Erin O'Shea. Global analysis of protein localization in budding yeast. *Nature*, 425:686–691, 11 2003.

[FR23] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon Kohl, Andrew Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:1–11, 08 2021.

[FR24] Neri Kafkafi, Joseph Agassi, Elissa Chesler, John Crabbe, Wim Crusio, David Eilam, Robert Gerlai, Ilan Golani, Alex Gomez-Marin, Ruth Heller, Fuad Iraqi, Iman Jaljuli, Natasha Karp, Hugh Morgan, George Nicholson, Donald Pfaff, S Richter, Philip Stark, Oliver Stiedl, and Yoav Benjamini. Reproducibility and replicability of rodent phenotyping in preclinical studies. *Neuroscience and Biobehavioral Reviews*, 87:218–232, 01 2018.

[FR25] Nevan Krogan, Gerard Cagney, Haiyuan Yu, Gouqing Zhong, Xinghua Guo, Alexandr Ignatchenko, Joyce Li, Shuye Pu, Nira Datta, Aaron Tikuisis, Thanuja Punna, José Peregrín-Alvarez, Michael Shales, Xin Zhang, Mike Davey, Mark Robinson, Alberto Paccanaro, James Bray, Anthony Sheung, and Jack Greenblatt. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440:637–643, 04 2006.

[FR26] Michael Kuhn, Christian von Mering, Monica Campillos, Lars Jensen, and Peer Bork. Stitch: Interaction networks of chemicals and proteins. *Nucleic acids research*, 36:D684–8, 02 2008.

[FR27] Manju Kumari, Gabriele Schoiswohl, Chandramohan Chitraju, Margret Paar, Irina Cornaciu, Ashraf Rangrez, Nuttaporn Wongsiriroj, Harald Nagy, Pavlina Ivanova, Sarah Scott, Oskar Knittelfelder, Gerald Rechberger, Ruth Birner-Gruenberger, Sandra Eder, H Brown, Guenter Haemmerle, Monika Oberer, Achim Lass, Erin Kershaw, and Rudolf Zechner. Adiponutrin Functions as a Nutritionally Regulated Lysophosphatidic Acid Acyltransferase. *Cell metabolism*, 15:691–702, 05 2012.

[FR28] Ryan Logan, Raymond Robledo, Jill Recla, Vivek Philip, Jason Bubier, Jeremy Jay, Carter Harwood, Troy Wilcox, Daniel Gatti, Carol Bult, Gary Churchill, and Elissa Chesler. High-precision genetic mapping of behavioral traits in the diversity outbred mouse population. *Genes, brain, and behavior*, 02 2013.

[FR29] Cian Lynch, Raquel Bernad, Isabel Calvo, Sandrina Nobrega-Pereira, Sergio Ruiz Macias, Nuria Ibarz, Ana Martinez del Val, Osvaldo Graña, Gonzalo Gómez-López, Eduardo León, Vladimir Angarica, Antonio Sol, Sagrario Ortega, Oskar Fdez-Capetillo, Enrique Rojo, Javier Muñoz, and Manuel Serrano. The RNA Polymerase II Factor RPAP1 Is Critical for Mediator-Driven Transcription and Cell Identity. *Cell Reports*, 22:396–410, 01 2018.

[FR30] Holger Maier, Stefanie Leuchtenberger, Helmut Fuchs, Valerie Gailus-Durner, and Martin Angelis. Big data in large-scale systemic mouse phenotyping. *Current Opinion in Systems Biology*, 4, 08 2017.

[FR31] Silvia Mandillo, Valter Tucci, Sabine Hölter, Hamid Meziane, Mumna Al Banchaabouchi, Magdalena Kallnik, Heena Lad, Patrick Nolan, Abdel-Mouttalib

Ouagazzal, Emma Coghill, Karin Gale, Elisabetta Golini, Sylvie Jacquot, Wojtek Krezel, Andrew Parker, Fabrice Riet, Ilka Schneider, Daniela Marazziti, Johan Auwerx, and Wolfgang Wurst. Reliability, robustness, and reproducibility in mouse behavioral phenotyping: A cross-laboratory study. *Physiological genomics*, 34:243–55, 06 2008.

[FR32] Niklas Norén, Andrew Bate, Roland Orre, and Ivor Edwards. Extending the methods used to screen the who drug safety database towards analysis of complex associations and improved accuracy for rare events. *Statistics in medicine*, 25:3740–57, 11 2006.

[FR33] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey Bzikadze, Alla Mikheenko, Mitchell Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah Hoyt, Mark Diekhans, Glennis Logsdon, Michael Alonge, Stylianos Antonarakis, Matthew Borchers, Gerry Bouffard, Shelise Brooks, and Adam Phillippy. The complete sequence of a human genome. *Science*, 376:44–53, 04 2022.

[FR34] Jonathon O'Brien, Harsha Gunawardena, Joao Paulo, Xian Chen, Joseph Ibrahim, Steven Gygi, and Bahjat Qaqish. The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. *The annals of applied statistics*, 12:2075–2095, 12 2018.

[FR35] Sara Omranian, Angela Angeleska, and Zoran Nikoloski. Pc2p: Parameter-free network-based prediction of protein complexes. *Bioinformatics*, 37, 01 2021.

[FR36] Bogdan Pasaniuc and Alkes Price. Dissecting the genetics of complex traits using summary association statistics. *Nature reviews. Genetics*, 18, 11 2016.

[FR37] DeLinda Pearson, Sheila Dawling, William Walsh, Jonathan Haines, Brian Christman, Amy Bazyk, Nathan Scott, and Marshall Summar. Neonatal Pulmonary Hypertension — Urea-Cycle Intermediates, Nitric Oxide Production, and Carbamoyl-Phosphate Synthetase Function. *The New England journal of medicine*, 344:1832–8, 06 2001.

[FR38] Marco Pellegrini, Miriam Baglioni, and Filippo Geraci. Protein complex prediction for large protein protein interaction networks with the Core&Peel method. *BMC Bioinformatics*, 17:37–58, 11 2016.

[FR39] Steve Perrin. Preclinical research: Make mouse studies work. *Nature*, 507:423–5, 03 2014.

[FR40] Megan Phifer-Rixey and Michael W Nachman. The natural history of model organisms: Insights into mammalian biology from the wild house mouse *Mus musculus*. *eLife*, 4:e05959, apr 2015.

[FR41] Alan Pittman, Emily Webb, Luis Carvajal Carmona, Kimberley Howarth, Maria Chiara Di Bernardo, Peter Broderick, Sarah Spain, Axel Walther, Amy Price, Kate Sullivan, Philip Twiss, Sarah Fielding, Andrew Rowan, Emma Jaeger, Jayaram Vijayakrishnan, Ian Chandler, Steven Penegar, Mobshra Qureshi, Steven Lubbe, and Richard Houlston. Refinement of the basis and impact of common 11q23.1 variation to the risk of developing colorectal cancer. *Human molecular genetics*, 17:3720–7, 09 2008.

[FR42] Jenny Poynter, Jane Figueiredo, David Conti, Kathleen Kennedy, Steven Gallinger, Kimberly Siegmund, Graham Casey, Stephen Thibodeau, Mark Jenkins, John Hopper, Graham Byrnes, John Baron, Ellen Goode, Maarit Tiirikainen, Noralane Lindor, John Grove, Polly Newcomb, Jeremy Jass, Joanne Young, and Loic Marchand. Variants on 9p24 and 8q24 Are Associated with Risk of Colorectal Cancer: Results from the Colon Cancer Family Registry. *Cancer research*, 67:11128–32, 01 2008.

[FR43] Guillaume Rigaut, Anna Shevchenko, Berthold Rutz, Matthias Wilm, Matthias Mann, and B.A. Séraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nature biotechnology*, 17:1030–2, 11 1999.

[FR44] Stefano Romeo, Federica Sentinelli, Valentina Cambuli, Michela Incani, Tiziana Congiu, Vanessa Matta, Sabrina Pilia, Isabel Huang-Doran, Efisio Cossu, Sandro Loche, and Marco Baroni. The 148M allele of the PNPLA3 gene is associated with indices of liver damage early in life. *Journal of hepatology*, 53:335–8, 08 2010.

[FR45] Emel Rothzerg, Xuan Ho, Jiake Xu, David Wood, Aare Martson, and Sulev Kõks. Upregulation of 15 Antisense Long Non-Coding RNAs in Osteosarcoma. *Genes*, 12:1132, 07 2021.

[FR46] Kyle Roux, Dae Kim, Brian Burke, and Danielle May. *BioID: A Screen for*

Protein-Protein Interactions: BioID Screen for Protein-Protein Interactions, volume 91, pages 19.23.1–19.23.15. 02 2018.

[FR47] Katherine Ruth, Felix Day, Jessica Tyrrell, Deborah Thompson, Andrew Wood, Anubha Mahajan, Robin Beaumont, Laura Wittemans, Susan Martin, Alexander Busch, A. Mesut Erzurumluoglu, Benjamin Hollis, Tracy O'Mara, Mark McCarthy, Claudia Langenberg, Douglas Easton, Nicholas Wareham, Stephen Burgess, Anna Murray, and John Perry. Using human genetics to understand the disease impacts of testosterone in men and women. *Nature Medicine*, 26:1–7, 02 2020.

[FR48] Dana Smith, Darrell Knabe, H. Russell Cross, and Stephen Smith. A diet containing myristoleic plus palmitoleic acids elevates plasma cholesterol in young growing swine. *Lipids*, 31:849–58, 09 1996.

[FR49] Nan Song, Aesun Shin, Ji Won Park, Jeongseon Kim, and Jae Oh. Common risk variants for colorectal cancer: An evaluation of associations with age at cancer onset. *Scientific Reports*, 7:40644, 01 2017.

[FR50] Patel N. Turcotte M. et al. Tam, V. Benefits and limitations of genome-wide association studies. *Nat Rev Genet*, 20:467–484, 2019.

[FR51] Michael Taschner, Kristina Weber, Andre Mourao, Melanie Korntner-Vetter, Mayanka Awasthi, Marc Stiegler, Sagar Bhogaraju, and Esben Lorentzen. Intraflagellar transport proteins 172, 80, 57, 54, 38, and 20 form a stable tubulin-binding IFT-B2 complex. *The EMBO Journal*, 35, 02 2016.

[FR52] Emil Uffelmann, Qinqin Huang, Nchangwi S Munung, Jantina Vries, Yukinori Okada, Alicia Martin, Hilary Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1, 12 2021.

[FR53] Luca Valenti, Ahmad Al-Serri, Ann Daly, Enrico Galmozzi, Raffaela Rametta, Paola Dongiovanni, Valerio Nobili, Enrico Mozzi, Giancarlo Roviaro, Ester Vanni, Elisabetta Bugianesi, Marco Maggioni, Anna Ludovica Fracanzani, Silvia Fargion, and Christopher Day. Homozygosity for the Patatin-Like Phospholipase-3/Adiponutrin I148M Polymorphism Influences Liver Fibrosis in Patients with Nonalcoholic Fatty Liver Disease. *Hepatology (Baltimore, Md.)*, 51:1209–17, 04 2010.

[FR54] Bernhard Voelkl, Lucile Vogt, Emily Sena, and Hanno Würbel. Reproducibility of preclinical animal research improves with heterogeneity of study samples. *PLOS Biology*, 16:e2003693, 02 2018.

[FR55] Hong-Xia Wang, Mingqiang Li, Ciaran Lee, Syandan Chakraborty, Hae-Won Kim, Gang Bao, and Kam Leong. CRISPR/Cas9-Based Genome Editing for Disease Modeling and Therapy: Challenges and Opportunities for Nonviral Delivery. *Chemical Reviews*, 117, 06 2017.

[FR56] Robert Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep Abril, Pankaj Agarwal, Richa Agarwala, Rachel Ainscough, Marina Alexandersson, Peter An, Stylianos Antonarakis, John Attwood, Robert Baertsch, Jonathon Bailey, Karen Barlow, Stephan Beck, Eric Berry, Bruce Birren, Toby Bloom, and Eric Lander. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–62, 01 2003.

[FR57] Chase Weidmann, Anthony Mustoe, Parth Jariwala, J. Calabrese, and Kevin Weeks. Analysis of RNA–protein networks with RNP-MaP defines functional hubs on RNA. *Nature Biotechnology*, 39:1–10, 03 2021.

[FR58] Lu Wen and Fuchou Tang. Recent advances on single-cell sequencing technologies. *Precision Clinical Medicine*, 5, 01 2022.

[FR59] Hui Yang, Haoyi Wang, and Rudolf Jaenisch. Generating genetically modified mice using CRISPR/Cas-mediated genome engineering. *Nature protocols*, 9:1956–1968, 08 2014.

[FR60] Fang-Yuan Yu, Zhi-Hua Yang, Nan Tang, Hong Lin, Jian-Hua Luo, Shuai Jiang, Yong-Sheng Ding, Lu-Xia Wang, and Hong-Wen Deng. Predicting protein complex in protein interaction network - a supervised learning based method. *BMC Systems Biology*, 8(Suppl 3):S4, 2014.

[FR61] Shilu Zhang, Saptarshi Pyne, Stefan Pietrzak, Spencer Halberg, Sunnie Grace McCalla, Alireza Fotuhi Siahpirani, Rupa Sridharan, and Sushmita Roy. Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. *Nature Communications*, 14, 05 2023.

[FR62] Zehua Zhang, Jian Song, Jijun Tang, and Fei Guo. Detecting complexes from edge-weighted PPI networks via genes expression analysis. *BMC Systems Biology*, 12:40, 10 2018.