

# Characterisation of model species interactome available from primary molecular interaction databases

*Vitalii Kleshchevnikov*

*21 December 2016*

# Contents

<b>1 Outline</b>	<b>3</b>
<b>2 Abstract</b>	<b>4</b>
<b>3 Introduction</b>	<b>5</b>
Defining interactome . . . . .	6
Experimental approaches for discovering interactions . . . . .	6
Challenges of interactomics . . . . .	8
<b>4 Motivation for this study</b>	<b>10</b>
<b>5 Aims of the study</b>	<b>11</b>
<b>6 Methods - data processing and analysis</b>	<b>12</b>
Getting proteome from UniProtKB . . . . .	12
Getting and transforming interactome data from IMEx databases and BioGRID . . . . .	12
Physicochemical properties and disordered region prediction . . . . .	13
Gene Ontology enrichment analysis . . . . .	13
<b>7 Results</b>	<b>15</b>
1. How well available interactome covers the proteome . . . . .	15
Available interactome covers substantial fraction of the reviewed proteome of main model species . . . . .	15
2. BioGRID database (as obtained from Menta) overlaps significantly with IMEx databases . . . . .	17
3. Mouse and human proteins are commonly combined for interaction experiments . . . . .	17
4. Which proteins are missing interaction evidence? . . . . .	20
Olfactory receptors are a major group of human proteins not represented in IntAct . . . . .	20
Human proteins with no available interactions are on average shorter than the proteins with interactions available . . . . .	22
5. Do major protein interaction detection methods (two-hybrid and AP-MS) exhibit a bias towards biochemical properties of the proteins? . . . . .	23
AP-MS is biased towards longer proteins . . . . .	23

Major experimental role types (bait/prey/neutral component) are not substantially influenced by the mass of the protein . . . . .	25
<b>8 Supplementary figures</b>	<b>26</b>
Supplementary figure 1 . . . . .	26
Supplementary figure 2 . . . . .	26
Supplementary figure 3 . . . . .	26
Supplementary figure 4 . . . . .	26

The report was published on 2017-05-16. Data used in the report was available on 2017-04-24.

# Chapter 1

## Outline

1. Abstract
2. Introduction
3. Methods
4. Results and discussion
  - how available interactome covers the proteome
  - IntAct(IMEx) vs Biogrid
  - which proteins are missing
  - interaction detection biases
  - study bias: the more articles, the more interactions
  - which protein function are interaction databases and high-throughput datasets enriched in?
5. Conclusion

# **Chapter 2**

## **Abstract**

# Chapter 3

## Introduction

The structure and the function of the cell arise from interactions between molecules inside and outside it [Hein et al., 2015]. Though proteins, nucleic acids, lipids and small molecules can all form important interactions, studies and literature focus mainly on interactions between proteins and other macromolecules. We can discover and study these molecular interactions using a number of experimental and computational techniques. This study aims to describe the coverage and the biases of currently available molecular interaction data. We focus on molecular interactions identified in the experimental setting most of which are represented (in the literature and databases) by protein-protein interactions (although, there is a considerable amount of data on protein-DNA interactions, for example, ChIP-Seq data, which is traditionally incorporated into the genomic or specialist databases). We focus on specifically on protein-protein interaction data.

Molecular interactions can be classified using multiple criteria: the information interaction detection methods can provide towards ground truth, interaction detection method, biological role of the interaction (covalent binding, enzymatic reaction, e.g.) which can be explored in molecular interaction ontology [Hermjakob et al., 2004, <http://www.ebi.ac.uk/ols/ontologies/mi/>]. A standard way of describing interaction allows record published interactions into databases, assign interactions a score based on the evidence and reliability and, least but not last, reuse interaction data for computational analyses, gaining insight into the novel function of proteins and generate hypothesis. In the recent years, the fact that biases in molecular interaction data can mislead network-driven studies has become evident [Schaefer et al., 2015], which motivates the need to study the coverage and bias of currently available molecular interaction data, identify missing proteins and molecular functions those proteins perform. These results may help in selecting appropriate network data for data integration studies.

We investigate the data in the IntAct database [Orchard et al., 2014] which is provider of high-depth manually curated molecular interaction data, a part of IMEx consortium. The IMEx consortium [Orchard et al., 2012] is an international collaboration between a group of major public interaction data providers who share curation effort. This study can aid literature curation effort by IntAct group by pointing to the publications containing interaction

data on proteins with no interactions currently deposited in the IntAct database.

interac we can classify interactions into different types: binary interactions and associations. Binary interactions are the interactions between two components, for example, two specific proteins, some detection methods (e.g. two-hybrid) identify those. To understand associations, we need to imagine we know proteins A, B and C constitute a complex and interact as shown in a figure 1 A. When we conduct an experiment, we choose the bait (the molecule experimentally treated to capture its interacting partners - called preys) to be protein A, and by detection method (e.g. affinity-purification mass spectrometry) we get both protein B and protein C detected as preys. Next step is to translate bait-prey relationship into a model of reality like the one shown in the figure 1 A. We call interactions between A-B and A-C associations because we cannot infer the true relationship between A, B, and C from this experiment design. In the other words, establishing that proteins are in direct physical contact is really challenging. However, to represent associations in a tabular format with each row corresponding to one interaction (e.g. A-B) we need to expand those. Two ways are commonly used to expand interactions, matrix and spoke expansion, both shown in the figure 1 B.

## Defining interactome

The aggregation of all components and their interactions into a single network results in what we call interactome, the whole of all molecular interactions. You can also look into the subset of this network, for example, you can select only those proteins that are expressed in the brain, and only the interactions between these proteins identified experimentally in the brain cells. This example reflects the complexity and the diversity of the interactome - which is what you would expect from a system underlying the complexity and the diversity of the cell types, cellular behaviours, and functions. For the same reason, only by studying these interactions and how they change in specific cell types and under specific circumstances in combination with the functional analysis we can decipher cellular regulatory networks. The ultimate goal of the research in the field would be to capture all physical interactions and thoroughly describe them while avoiding false positive results.

## Experimental approaches for discovering interactions

Numerous experimental protein interaction detection methods are currently widely used. The classification we provide is far from comprehensive but gives a short description of the methods analysed in this study. Based on the evidence provided and possibility to scale up to high-throughput studies methods can be classified into 3 main categories.

The first category is formed by methods using affinity purification to capture all proteins associated with the bait. Only proteins that have a direct or indirect physical connection with the bait will be purified. Following purification procedure, those proteins can be identified using western-blotting and specific antibody staining or using mass-spectrometry, latter can be done in a high-throughput manner. The main advantage of these methods is the potential

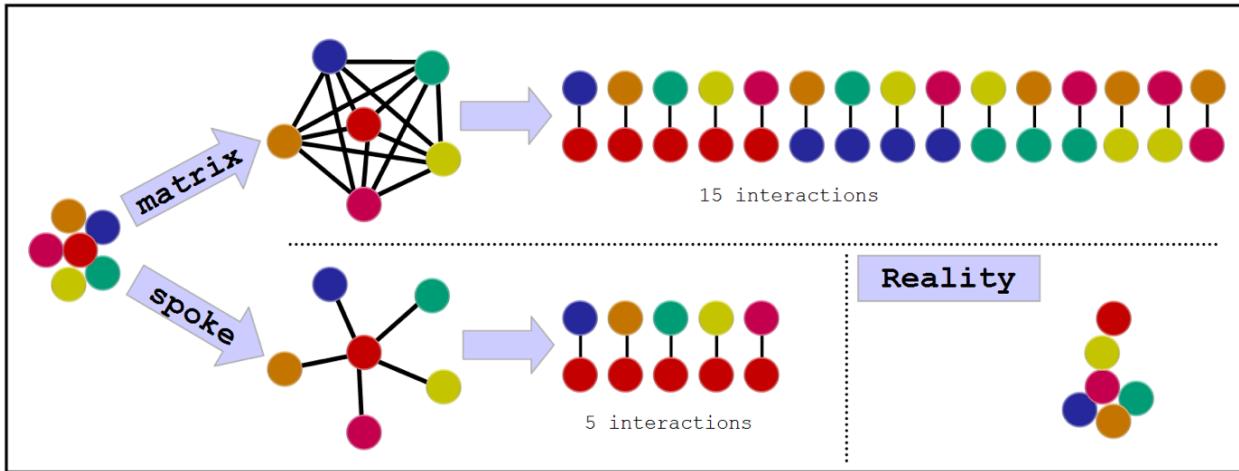


Figure 3.1: Figure 1-B. Association and physical associations identified using AP-MS approaches fundamentally don't provide binary interactions. Binary interactions have to be inferred from the list of proteins represented as cluster on the left. Matrix expansion links every protein to every other protein in the AP-MS-derived cluster. Spoke expansion only links the bait with all other proteins in a cluster. As you can see, none of these methods generate the exact set of binary interactions that occur in reality. Spoke expansion tends to generate less false-positives and is used more often.

ability capture any protein, in different cellular contexts and to quantitatively characterise interaction properties [Hein et al., 2015]. The main disadvantage of these techniques comes from the fact that experiments indentify both direct and indirect interactions between the bait and captured proteins and no way of distinguishing those (although, one may delete identified proteins one by one from the cell and decipher direct interactions). This type of interactions in called associations and do be represented in the network requires the use of expansion methods (Figure 1-B, adopted from IntAct website). The other drawback of this method is dependency on the availability and quaility of antibodies for affinity-purification step. We will call these methods affinity-purification followed mass-spectrometry (AP-MS) across this report.

As defined by molecular interaction ontology [Hermjakob et al., 2004, <http://www.ebi.ac.uk/ols/ontologies/mi/>], an association is an interaction between molecules that may participate in formation of one, but possibly more, physical complexes, association will be called physical if experiments show enough evidence that proteins are in the same physical complex but don't show direct interaction. The second category of methods is formed by protein complementation techniques which include two-hybrid (transcription factor complementation), the most widely used interaction detection method (including high-throughput experiments). In this method, pairs of proteins are tested for interaction and therefore discovered interactions are more likely to be direct (the main advantage of this method). Classic implementation of two-hybrid is performed in yeast cells and requires studied proteins to be non-membrane, however, two-hybrid for membrane proteins and for mammalian cells was also developed [Lemmens et al., 2015, Saraon et al. [2017]]. The main disadvantage of two-hybrid methods is that every protein has to be cloned into a plasmid or other vector and exogenously expressed.

Ability to clone protein-coding sequence and, in case of yeast two-hybrid, correct protein folding, are the limiting factors for two-hybrid but not AP-MS techniques. As a side note, the lack of antibodies for a specific bait protein may force researchers to tag, clone and express protein which is a subject to similar problems, however, AP-MS would allow identification of the binding partners, no cloning needed. Final category is formed by structure-based methods (co-crystallisation and X-ray crystallography, e.g.). These methods can provide valuable information on how exactly physical interaction occurs, however, these methods are extremely labor-intensive and, therefore, non-scalable and will always need complementary experiments showing if the proteins actually interact in the cellular context. To conclude, all protein interaction detection methods have their strengths and weaknesses, so it is important to accept that every protein interaction detection method has its limitations in the ability to identify true physical interaction and serves as evidence we use to infer protein interaction. By combining different methods to identify protein interaction we can gain more confidence in our findings and by disrupting interaction under specific cellular context we can identify its function.

## Challenges of interactomics

Four big challenges substantially complicate the study of molecular interactions, especially on the whole organism scale. The first being that we don't know the true nature of underlying our experimental results (all assays provide evidence that interaction is possible and some can provide quantitative description, but all are prone to error and the problem described in the figure 1 A) which lead to the necessity of combining interaction data from multiple experiments and complex statistical evaluation of how probable the interaction is based on that data (such as Bayesian approach, Braun et al. [2009], Zhang et al. [2011]) rather than receiving confident yes-or-no result from a single experiment. Interaction databases make an effort to score the interactions based on supporting evidence, however, this is usually done with non-probabilistic heuristic approaches, such as MI score implemented in IntAct [Villaveces et al., 2015]. Every database that aggregates interaction data from other resources will develop an algorithm to score interactions. The challenge is to identify when to put the threshold between high and low confidence interactions or when to say "I am confident the interaction exists".

The second big challenge is the problem of "noise" - or the problem of false positives. Different interaction detection experiments are prone to these errors for different reasons, for example, in-vitro experiments (e.g. TAP-MS) may allow the interaction between proteins which are normally separated between different cellular compartments. Specific groups of proteins (based on their physical or chemical properties, abundance) may have a higher susceptibility to false positives, for example, highly abundant proteins are easier to detect and may also be less efficiently diluted during the affinity purification procedure, which may lead to artifactual results. Contamination is another common problem for AP-MS experiments which have motivated the creation of contaminant database, CRAPOME [Mellacheruvu et al., 2013]. A more general problem of noise can be addressed by proteome-scale interactomics experiments (which can include enough samples to guarantee low false positive rate while still identifying

interactions).

The third big challenge is that our knowledge of interactome is incomplete which arises from the fact that experimental approaches have low statistical power and often miss out some real interactions. Many efforts to reproduce protein interactions find little overlap between the new and the original study [Braun et al., 2009]. Also, many proteins, especially in non-model species have no known interactions.

The final challenge contributes to the “incomplete interactome” problem but is grounded in the fact that not all protein interaction discovered and published are included in protein interaction databases. In other words, this is database curation problem. More than 100 public databases containing protein interactions are available now [<http://pathguide.org/>]. These databases differ: - by the types of data they include (e.g. computational prediction, manual curation experimental data from research papers (primary databases), aggregated data from many primary databases (secondary databases)),  
- the level of detail captured from articles to describe interactions - how often and if databases are updated with new data, identifiers and annotation.

The level of detail ranges from mentioning only the pairs of interacting partners and heuristic scores assigned to them (for instance, STRING) to the ones containing experiment details (detection method, bait/prey status, if available - quantitative data, experiment setup, protein variants), such as IntAct [Orchard et al., 2014]. The amount of interaction data generated per year is growing exponentially making manual curation of all this data into primary databases a daunting task. To prioritise curation efforts and reduce redundancy between databases (to curate different data using the same standards) IMEx consortium was formed in 2012 [Orchard et al., 2012]. IMEx-compliant databases include primary databases such as IntAct and MINT, but doesn’t include BioGRID (which curates at the lower level of detail) and no longer active databases such as HPRD [Peri et al., 2004] and BIND [Isserlin et al., 2011].

# **Chapter 4**

## **Motivation for this study**

Solving some of these challenges may be easier than the others. In particular, to solve the last challenge we can prioritise curation efforts for already published interactions to cover unrepresented proteins and we can encourage authors to submit their results to the databases prior to publishing. We can also encourage research of underrepresented parts of the interactome. This motivates the need to characterise the interactome already present in interaction databases. More specifically, we need to measure how available interactome covers the proteome of main model species, if there are any biases towards proteins with or without available interactions and if major protein interaction detection methods exhibit biases towards specific groups of proteins. These results can be also used to inform data integration projects and show the need to adjust for biases when studying protein-protein interactions network.

# Chapter 5

## Aims of the study

1. Describe the coverage of the proteome of main model species. Considering either all UniProtKB or SwissProt entries only as the proteome (canonical identifiers as well as protein isoforms). Consider all interactions from IMEx-compliant databases as interactome.
2. Compare the coverage of the proteome by interaction evidence from IMEx to the interaction evidence from BioGRID (the other major primary database).
3. Investigate if proteins with no available interactions stand out by specific functions (Gene Ontology, GO: biological process and molecular function), cellular localisation (GO), molecular mass, or protein evidence status from SwissProt
4. Find out if major protein interaction detection methods (two-hybrid and AP-MS) exhibit any bias towards biochemical properties of the proteins involved (mass, disordered regions, hydropathy, the fraction of charged residues)
5. Investigate the relationship between the number of interactions or MI score and the number of publications or GO terms per protein.

# **Chapter 6**

## **Methods - data processing and analysis**

### **Getting proteome from UniProtKB**

Whole proteome (all UniProtKB) for each species was downloaded programmatically in R using UniProt rest API. SwissProt-proteome was subset from the whole proteome by reviewed status column. UniProt identifies proteins by UniProtKB/AC (e.g. P04637, accession) which does not distinguish between protein isoforms. UniProt aggregates isoform information and identifiers (e.g. P04637-4) in a separate column with zero to many isoforms per each UniProtKB accession. To generate proteome list which includes protein isoforms, isoform accessions were extracted and combined with the list of generic accessions. In this analysis, protein evidence status and protein mass are only attributed to generic accessions.

### **Getting and transforming interactome data from IMEx databases and BioGRID**

Interactome from all IMEx databases was downloaded programmatically in R using PSIQUIC package from Bioconductor [Shannon, 2017]. The list of interactions (pairs of interactors) was transformed into the list of interactors preserving interactor identifiers, the type of interactor identifier, species information and the database interaction originates from. Only unique proteins were IMEx databases contain interactions between proteins, RNA, DNA and small molecules, moreover, these interaction may involve molecules originating from different species. Therefore, to perform by species interactome/proteome comparison there is a need to remove non-UniProtKB/AC molecule identifiers (which removes non-protein molecules, although, may also remove a small fraction of proteins which have no UniProtKB/AC) and there is a need to remove proteins originating from other species. Also, entries in IMEx databases has to be cleaned of tags and textual descriptions (“taxid:9606(human-h1299)|taxid:9606(Homo sapiens lung lymph node carcinoma)” to “9606”) to make further analysis easier and cleaner. Next, when provided in the research articles protein isoform information is always included in

IMEx databases, so to perform analysis excluding isoform information UniProtKB/AC were cleaned of -N suffix (P04637-4 to P04637). IMEx consortium databases such as IntAct, MINT, BHF-UCL, MPIDB, MatrixDB, HPIDb, I2D-IMEx, InnateDB-IMEx, MolCon, UniProt, MBInfo are currently integrated into IntAct [Orchard et al., 2014]. Data from all these databases was used in the analysis and referred to as IntAct data or IMEx data.

## Physicochemical properties and disordered region prediction

Information on disordered region content and physicochemical properties of individual proteins were obtained from the dataset generated by Vincent and Schnell in 2015 [Vincent and Schnell, 2016a]. Briefly, Vincent and Schnell used a number of disorder prediction algorithms (IUPred and DisEMBL) and their consensus to generate disordered region predictions for each protein which was used to calculate the fraction of disordered regions in a protein. In addition, Vincent and Schnell used localCIDER version 0.1.7 (Classification of Intrinsically Disordered Ensemble Regions) to calculate physical properties such as fraction of charged residues, mean hydropathy or charge separation for each protein. This was done for 10 eukaryotic proteomes and written to SQLite-database which was made available online through dryad service [Vincent and Schnell, 2016b].

## Gene Ontology enrichment analysis

Enrichment analysis can identify groups of genes selectively enriched or depleted in a specific dataset. Genes can be grouped by common annotations describing their function, location in the cell or membership in a pathway, such as those described in KEGG [Kanehisa et al., 2016] or Reactome [Fabregat et al., 2016]. Gene Ontology Consortium (GO) [Gene Ontology, 2015] annotates gene products (proteins, non-coding RNA, protein complexes) by molecular function they perform, cellular component they are located in and biological process they are involved in. This annotation is first done manually by a curator based on evidence provided in the research paper and then expanded to other proteins using computational annotation methods, for example, mouse orthologues of human protein are likely to have similar function. Computational annotation is less reliable however it allows to generate annotations for non-studied proteins or in poorly studied organisms. More information on the evidence codes for Gene Ontology annotations can be found on the GO Consortium website [<http://geneontology.org/page/guide-go-evidence-codes>]. All analyses in this study included all Gene Ontology annotations.

Enrichment analysis uses statistical tests and protein to GO term associations to identify functions and compartments overrepresented in a particular list of proteins as compared to background list. Statistical test is performed for every GO term and uses hypergeometric distribution to estimate the probability that a number of proteins in our list is annotated by a specific GO term by chance based on the fraction of proteins annotated with that term in the background list. Calculated probability tells how likely (by chance) is that our list is enriched in proteins with function described by that specific GO term. If that's highly unlikely to occur by chance we consider GO function to be enriched. Multiple hypothesis

testing correction is applied to adjust probabilities for higher probability of getting extreme result by chance due to testing thousands of GO terms. All enrichment test were performed and visualised using R package clusterProfiler [Yu et al., 2012]. GO annotations used in the clusterProfiler package are compiled from GO consortium into R package GO.db [Carlson, 2017] provided through Bioconductor project [Gentleman et al., 2004].

# Chapter 7

## Results

### 1. How well available interactome covers the proteome

#### Available interactome covers substantial fraction of the reviewed proteome of main model species

In this section, we compare the proteome, the list of all proteins in annotated Uniprot [The UniProt, 2017], to the list of proteins which have interacting partners annotated in IMEx consortium databases. We compare the coverage across seven model species. We have also compared interactome coverage of the manually reviewed SwissProt (shown on the Figure 7.1) and the reviewed SwissProt and unreviewed trEMBL (all UniProtKB) combined shown on the Supplementary Figure 1.

Non-canonical protein isoform identifiers (P32054-1) were converted to canonical identifiers (P32054) resulting in interactions “any isoform to any isoform” (Venn diagrams on the right) or isoform identifiers were preserved (Venn diagrams on the left). You can see that adding isoform information adds more proteins to the SwissProt list but not so many proteins to the IMEx list.

The best interactome annotated by IMEx databases is baker’s yeast, 2nd best interactome is *E.coli*. All other interactomes cover less than the half of their respective proteome (all UniProtKB, Supplementary figure 1). The overlap between the interactome and reviewed proteome (SwissProt) is significantly better.

A large fraction of human, mouse, arabidopsis proteins and more than a half of drosophila and *C.elegans* proteins are absent in SwissProt (but included in trEMBL) – suggesting under-annotation by the SwissProt-Uniprot team.

Protein isoforms (in multicellular model organisms) are almost not annotated in the interactome. Human is an exception: out of 22049 protein isoforms annotated by SwissProt 2656 protein isoforms have interacting partners. Other molecular interaction databases (active BioGRID, inactive HPRD) do not record isoform information.

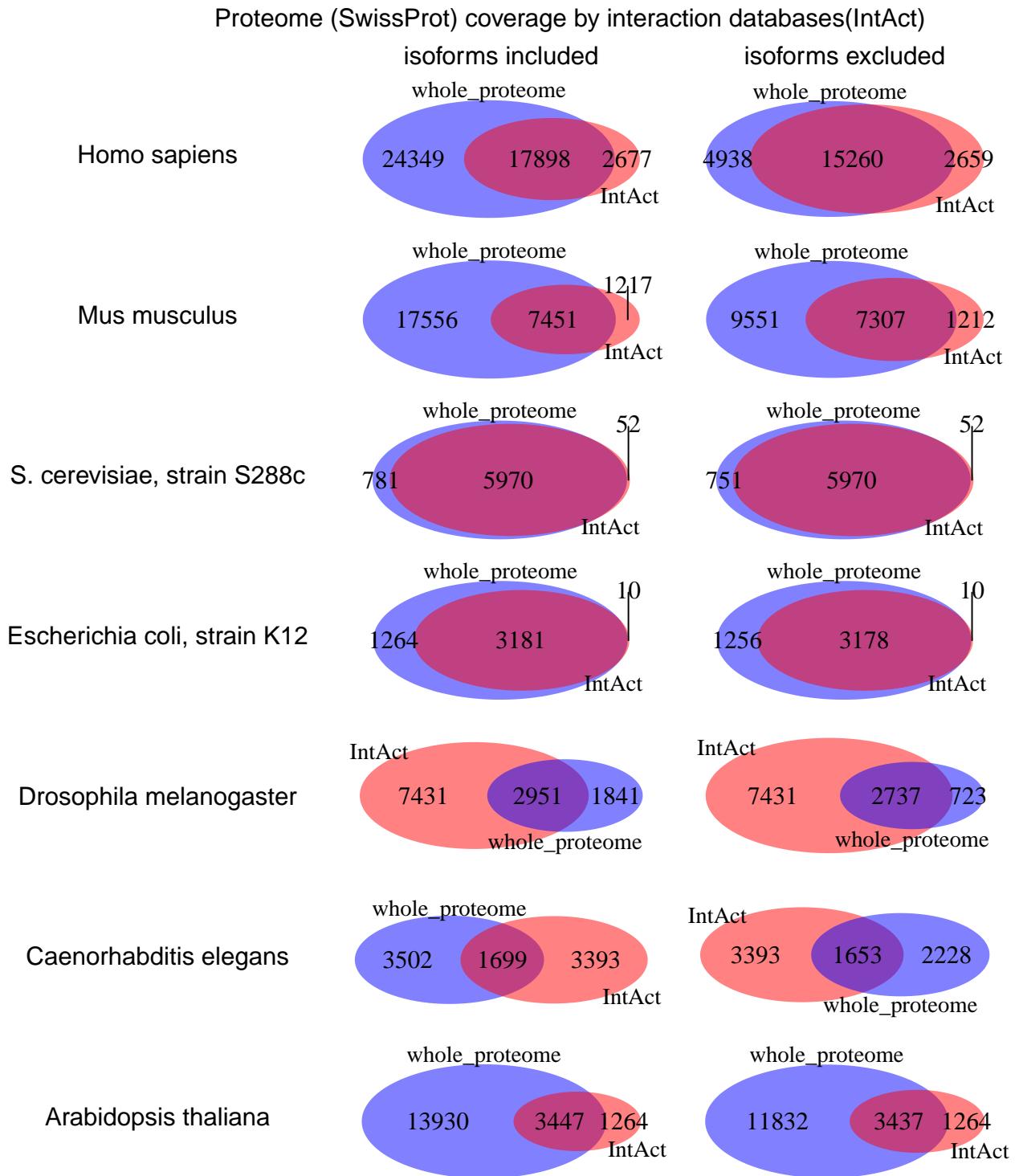


Figure 7.1: The best coverage is observed for yeast, E.coli and human. Isoform coverage is limited, but still significant for human. Red circles represent proteins which have interacting partners curated in IntAct, blue circles represent proteins in SwissProt, taking into account isoforms (left) or canonical identifiers (right).

## 2. BioGRID database (as obtained from Mentha) overlaps significantly with IMEx databases

BioGRID database is a major primary protein interaction database [Chatr-Aryamontri et al., 2017]. IntAct and BioGRID combined contain all interaction information which has been curated to currently active public databases (major inactive databases include HPRD [Peri et al., 2004] and BIND [Isserlin et al., 2011]). BioGRID uses shallow curation level (retains only some information about the interaction and the experiment) and identifies proteins using Entrez Gene ID while IntAct uses UniProtKB identifiers. This has allowed BioGRID to curate more interaction data, for example, BioGRID currently contains interaction data from 25307 publications and compared to only 11130 in IMEx (human data only considered), more information can be found: [https://github.com/vitkl/darkspaceproject/blob/master/BioGRID/BioGRID\\_dsgen.md](https://github.com/vitkl/darkspaceproject/blob/master/BioGRID/BioGRID_dsgen.md)).

One of the main drawbacks of BioGRID curation model is that the use of Entrez Gene ID does not allow to distinguish between different protein encoded by the same gene or different protein isoforms, specify sufficient-to-bind regions of the protein. In our analysis, this difference introduces additional identifier mapping step (Gene ID to UniProtKB), which generates redundancy and ambiguity: several proteins can be coded by one gene and several genes can code single proteins. Mentha database [Calderone et al., 2013] has imported all BioGRID-stored interactions and has mapped Gene ID to UniProtKB, so we used Mentha to get BioGRID-stored interactions. Mentha doesn't import interactions for *E.coli*, therefore *E.coli* is not present in this analysis.

BioGRID has recently incorporated a large-scale study aimed, in part, to find interacting partners of the proteins with no known interactions [Huttl et al., 2015], which increases observed coverage by BioGRID.

Overall, BioGRID overlaps with IMEx to a large extent. Nonetheless, for all of the species we have looked at, BioGRID has annotated some proteins (and their interactions) which are not annotated in IMEx. BioGRID stores substantially more interactions for arabidopsis and a considerable fraction of human and mouse interactions.

## 3. Mouse and human proteins are commonly combined for interaction experiments

The fact that researchers tend to put proteins from other species (mostly human) into mouse experiments or tend to put mouse proteins into cells from other species (mostly human) is also common for interaction detection experiments and is clearly seen in Figure 7.3: half of the mouse interactors are from the other species. This holds true both for IMEx databases (Figure 7.3) and for BioGRID (Supplementary figure 2); however, this analysis doesn't show which proteins (mouse or human) were used as bait to capture interactions in which cells (mouse or human).

Interchangeable use of mouse and human proteins generates interaction data which is hard

### Proteome coverage by interaction databases, overlap between IntAct and BioGRID

UniprotKB, isoforms excluded SwissProt, isoforms excluded

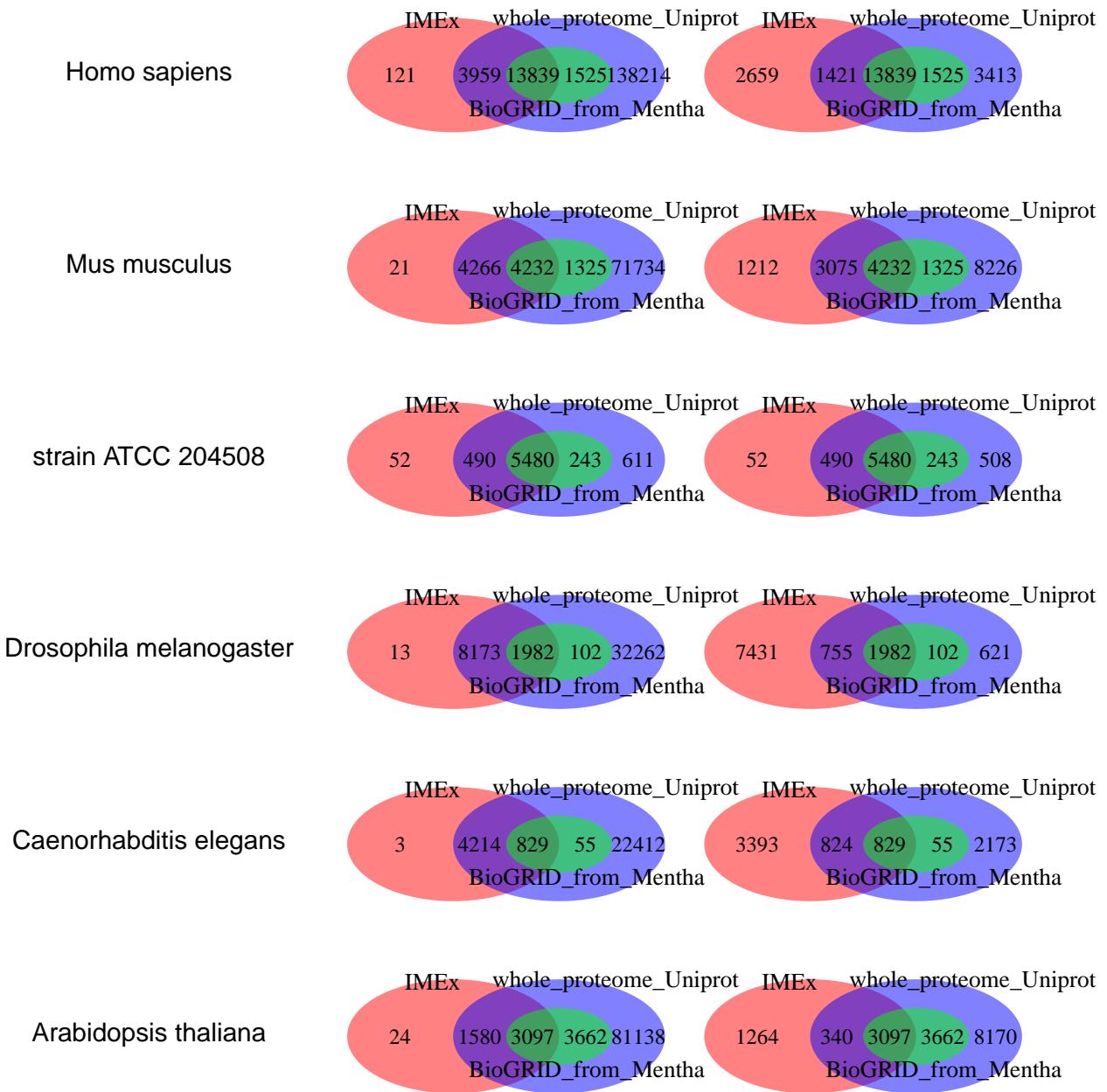


Figure 7.2: BioGRID and IMEx share a considerable fraction of proteins with interaction data. Red circles represent proteins which have interacting partners curated in IntAct, blue circles represent proteins in SwissProt, green circles represent proteins which have interacting partners curated in BioGRID, taking into account all UniProtKB (left) or only manually reviewed SwissProt (right) proteins, only canonical identifiers considered.

The number of interacting partners which have UniprotKB identifiers and belong to the species of interest

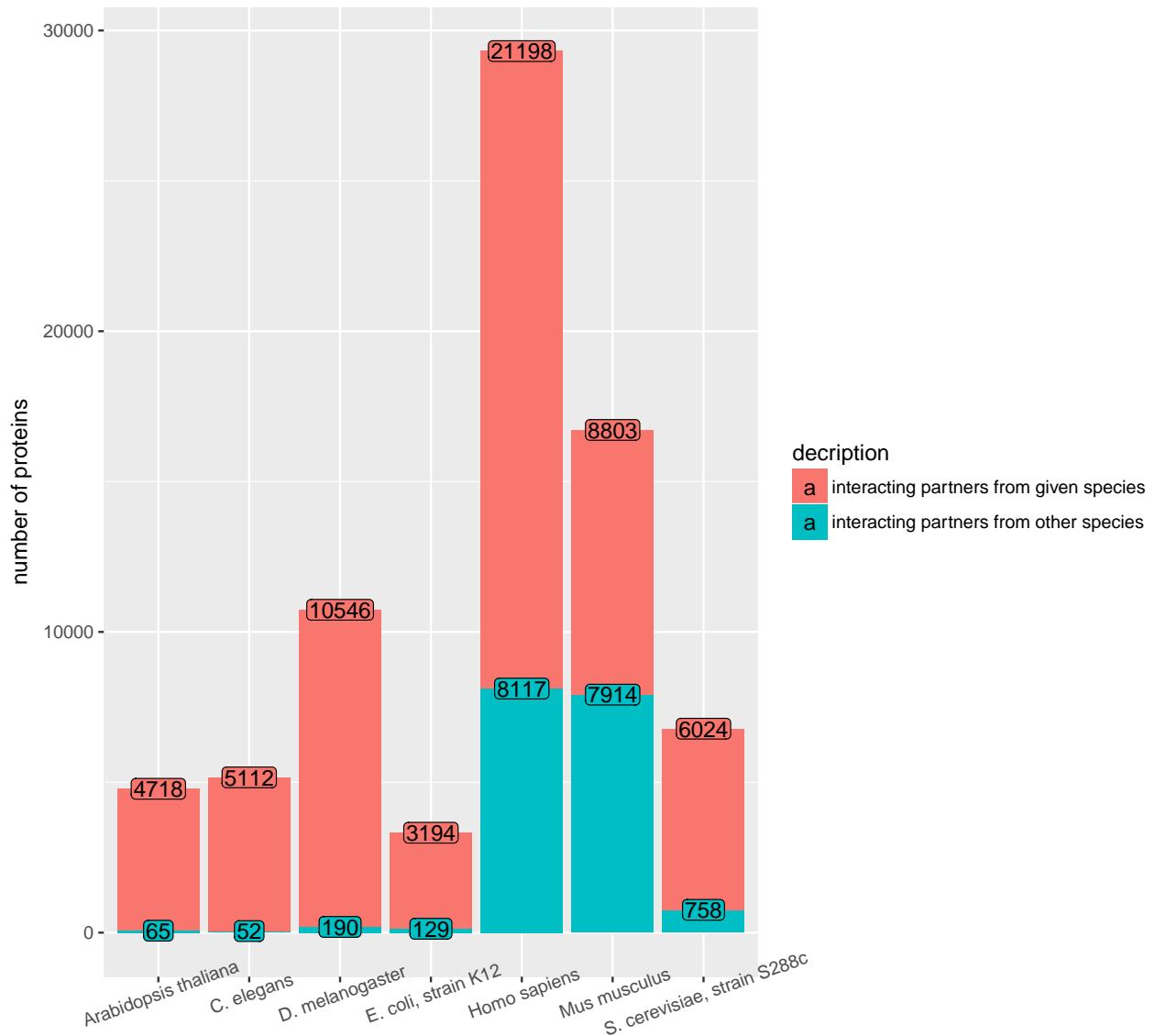


Figure 7.3: Mouse and human proteins are commonly combined for interaction experiments. Red bar shows the number of proteins from a given species, blue bar shows the number of proteins from another species. Red group proteins form interactions both with red and blue group proteins, blue group proteins form interactions only with red group proteins

to reuse and introduces imprecision due to the fact that it requires the mapping between homologous proteins; however, this may not be the biggest problem with studying the interactions between mouse and human proteins and trying to correctly interpret results. Recent studies of intrinsically disordered proteins show that linear amino acid motifs located in disordered regions frequently mediate protein-protein interactions [Babu, 2016], for example, the disordered region of p53 mediates its ability to recruit transcription-activating proteins to the promoter [Buljan et al., 2012]. More importantly, these linear amino acid motifs can evolve quickly, for example, allowing cancer cells to escape control by P53 [Buljan et al., 2012]. So, while the interaction between mouse protein A and human protein B can exist, that might not be true for the interaction between human protein A and human protein B, and vice-versa.

On the other hand, some researchers advocate that interactions important for the cellular function should be conserved between species [Li et al., 2017].

Surprisingly, 19438 interactions between mouse and human proteins were discovered in human rather than mouse cells (only 1233) suggesting that researchers use mouse rather than human proteins as baits (1158 mouse baits total, 5606 human preys total, including isoforms, from 441 publications) to find interactions directly relevant to human interactome research, including human disease.

## 4. Which proteins are missing interaction evidence?

Characterising the properties of proteins missing interaction evidence can help prioritise curation efforts. By looking for proteins missing interaction evidence and involved in particular biological function (as described by Gene Ontology) we can complete missing part of the interactome.

### Olfactory receptors are a major group of human proteins not represented in IntAct

We have performed Gene Ontology enrichment analysis of proteins with no available interactions as compared to whole proteome (SwissProt protein list). We have looked at overrepresented biological processes, molecular fuctions and cellular compartment localisation of proteins. Here only overrepresented biological processes are shown.

Most of the proteins missing interactions are smell receptors, ion channels, adhesion molecules, lipid synthesys enzymes, taste receptors. A minor category of proteins involving sperm-egg recognition is also overrepresented. As you might infer, all these proteins are transmembrane or are connected to the membrane, which is also supported by enrichment analysis for GO cellular component. We can speculate that olfactory receptors are numerous and according to our current knowledge, similar in function, therefore this group has not attracted enough interest to study it's protein-protein interactions. Nonetheless, transmembrane proteins are challenging to perform protein-protein interactions studies on.

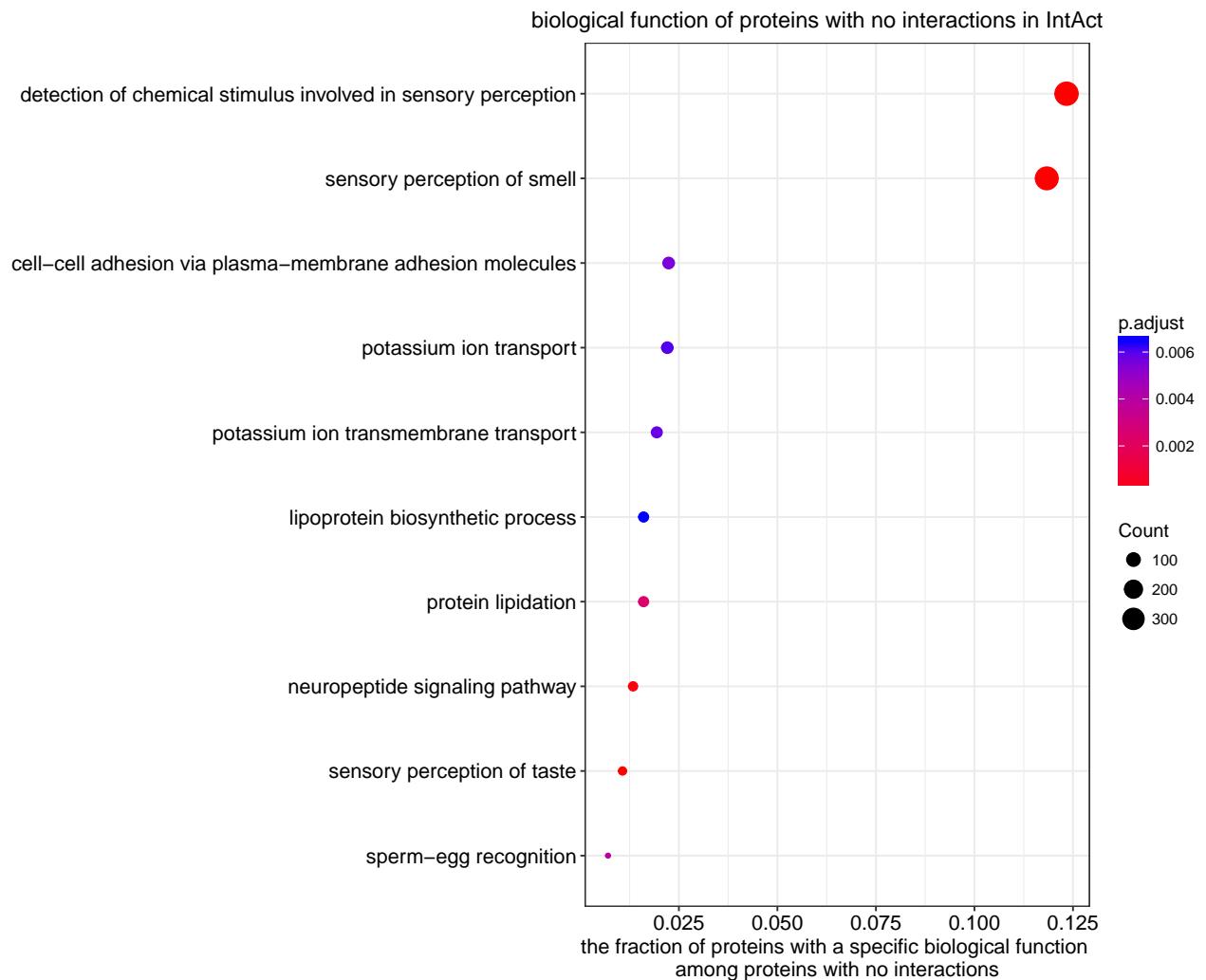


Figure 7.4: Overrepresented biological processes. The size of the dot shows the number of proteins of that category in the dataset (proteins without interaction), x-axis represents the fraction of proteins of that category in the dataset. Colour shows the probability of that overlap occurring by chance.

## Human proteins with no available interactions are on average shorter than the proteins with interactions available

Protein length or mass are physical properties of a protein which can, in theory, influence its usage as a bait in experiments and its detection in case methods depend on protein length. Protein length is also important biologically. Longer proteins can have multiple functional domains and, therefore, more interactions. The distribution of protein mass has a very long right tail - there are much more big proteins than a normal distribution would predict (Supplementary figure 3), which only allows using non-parametric statistical tests (Wilcoxon test). Log10 transformation of protein mass, though, makes extreme values less extreme and is approximately normally distributed.

The distribution of protein mass for proteins with and without interacting partners curated in IMEx

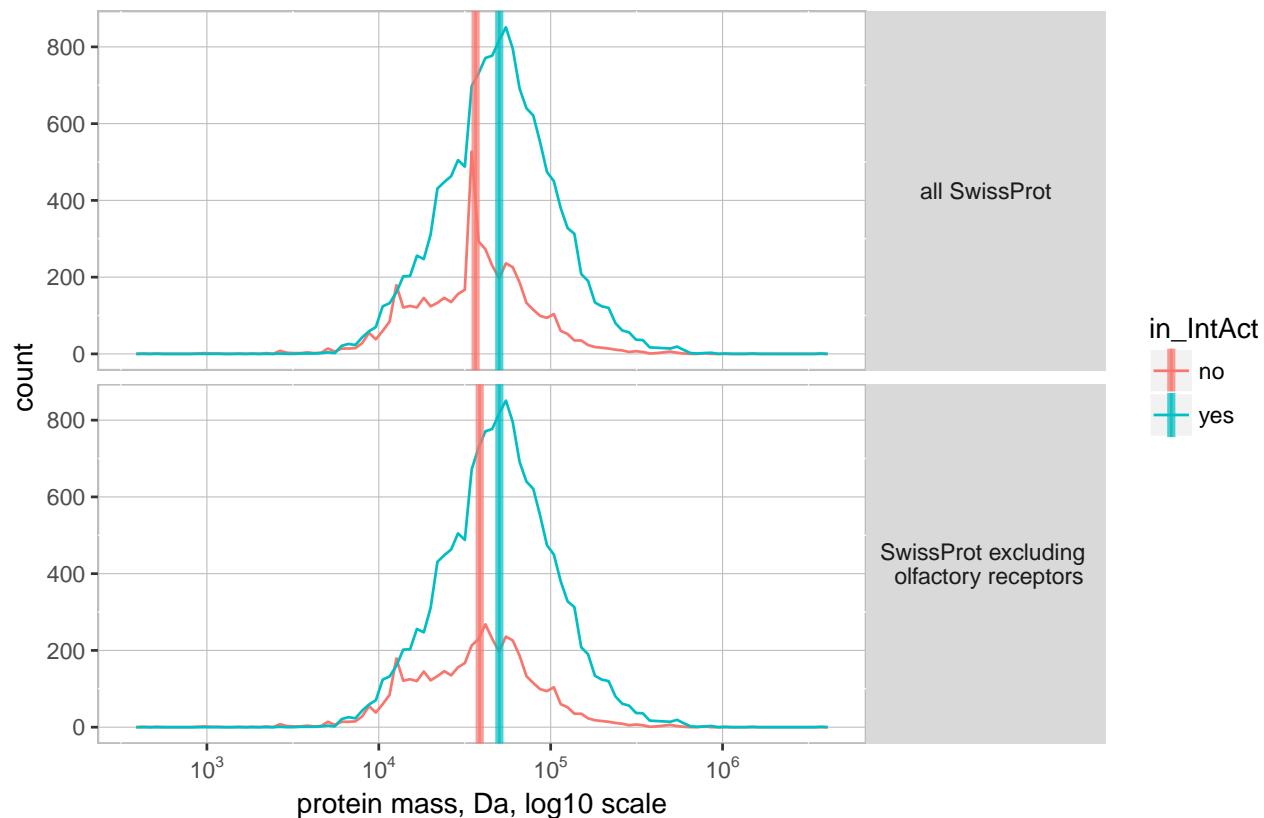


Figure 7.5: Human proteins with no available interactions are on average shorter than the proteins with interactions available. The plot compares the distribution of protein mass across SwissProt proteins with or without interacting partners in IntAct. Vertical line highlights the median.

This difference in protein mass between proteins present and absent in the interactome is highly unlikely to occur by chance (Wilcoxon rank test (Mass, Da, 95% confidence interval: -13000, -11100, p-value: 6.76e-142) and Student t-test(log10 of Mass, Da, 95% confidence interval: -0.153, -0.132, p-value: 4.68e-143) on the whole population of proteins, Monte-Carlo

sampling (is it useful?), permutation of labels followed by Wilcox rank test (is it useful?) - Supplementary figure 3,4,5). Removing 416 olfactory receptors, evidently, does not change this trend (Wilcox rank test on Mass, Da, 95% confidence interval: -12800, -10800, p-value: 2.98e-121).

## 5. Do major protein interaction detection methods (two-hybrid and AP-MS) exhibit a bias towards biochemical properties of the proteins?

The problem of bias in interactomics is often discussed in research papers and in the community. Now we will focus on the bias of the experimental detection methods. We choose to compare two major families of methods: two-hybrid and affinity-purification followed by mass-spectrometry (AP-MS).

We identify two-hybrid method using PSI-MI ontology: detection method = “transcriptional complementation assay” (MI:0018) - all methods which belong to this type (which are children terms in the ontology). We identify AP-MS method using two PSI-MI ontology terms: detection method = “affinity chromatography technology” (MI:0004) and participant identification method = “partial identification of protein sequence” (MI:0433). The use of ontology terms for searching interaction allows to only specify a single term as opposed to listing every individual detection method included in those groups. One can browse ontology lookup service to identify which exact approaches are included under the umbrella of two-hybrid or affinity-purification-mass-spectrometry.

We fetch interactions detected using these methods by querying molecular interaction databases using PSICQUIC service.

### AP-MS is biased towards longer proteins

We have compared interaction detection method bias towards protein mass across four species: *Homo sapiens*, *Mus musculus*, *S. cerevisiae* strain S288c, *Escherichia coli* strain K12. We observe that affinity-purification followed by mass-spectrometry (AP-MS) seems to capture a higher proportion of larger proteins as compared to two-hybrid methods. This pattern is present across four species that we looked at, however, is manifested to a different extent

The statistical significance of the difference in protein mass across protein detection methods was tested using linear model approach. Linear model offers way to perform multiple statistical tests (ANOVA and posthoc ttests) in R. Linear model takes a group assignment for a protein (only two-hybrid, only AP-MS, both two-hybrid and AP-MS, other methods - a vector 0s and 1s where 1 show which group a protein belongs to) and the mass of that protein. Using these parameters for each of the proteins linear model can learn the mean of each group as well as provide a way to calculate errors for each of the between-group comparisons. The robust linear model used for this test is less sensitive to the extreme values. The figure X shows that the mass of the proteins which were identified interacting using AP-MS technique

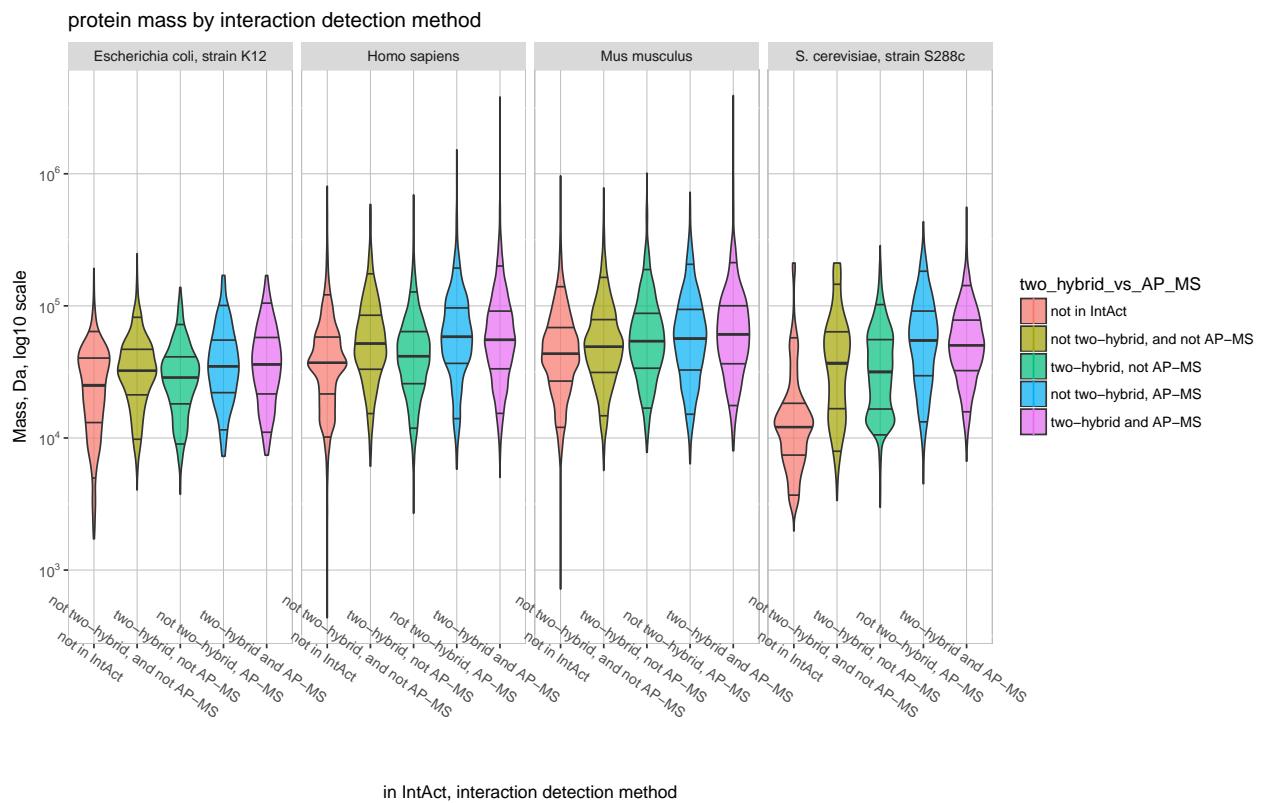


Figure 7.6: Plots show the distribution of protein mass across four model species separated by interaction detection method. Details in text.

is on average higher than the mass of the proteins identified using two-hybrid or the other methods.

## Major experimental role types (bait/prey/neutral component) are not substantially influenced by the mass of the protein

Unlike the experimental role type, whether a protein interaction is detected using a particular interaction detection method is influenced by the mass of that protein. The experimental role of the protein (bait/prey/neutral component) can, in theory, be influenced by protein mass. For example, it may be easier to clone shorter proteins for use as bait/prey. So it is important to see if the experimental role is really influenced by protein mass.

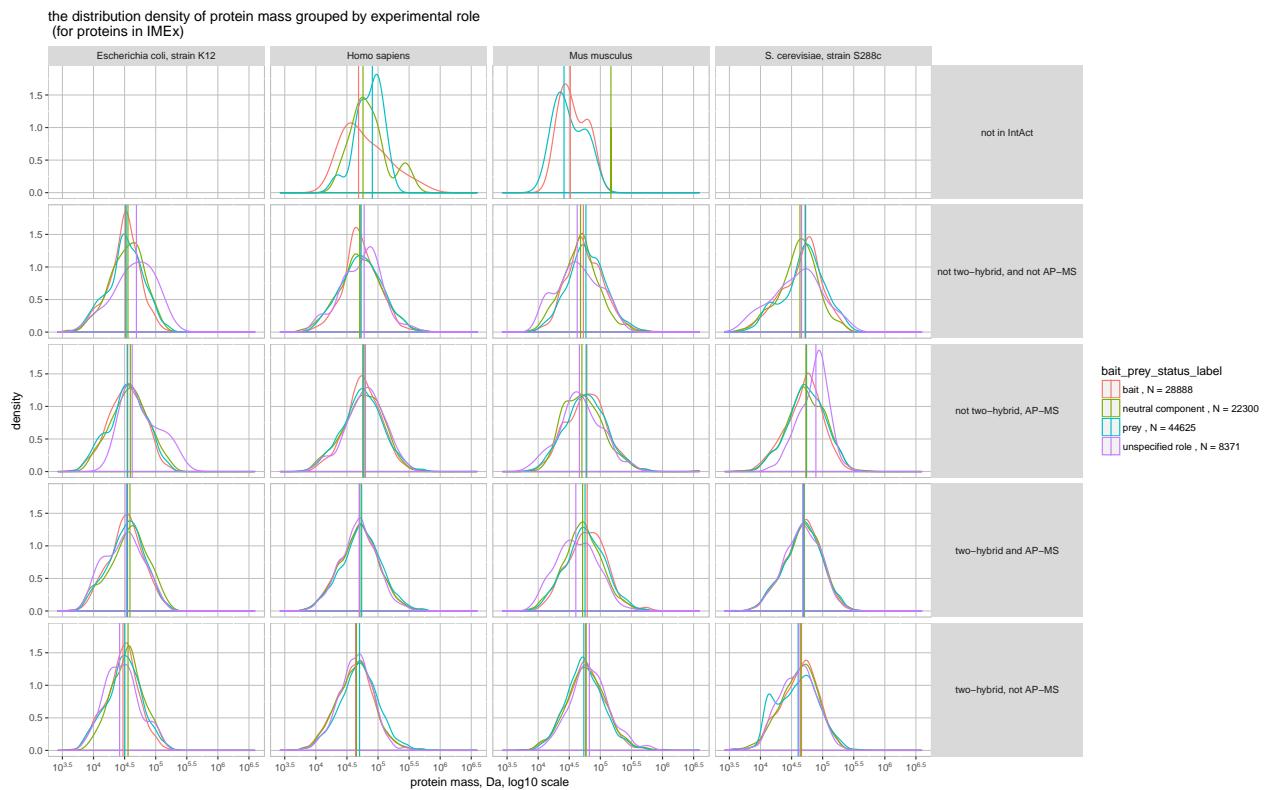


Figure 7.7: Experimental role of a protein is not substantially influenced by the mass of the protein.

Evidently, the experimental role of a protein is not substantially influenced by the mass of the protein. Nonetheless, this analysis doesn't factor in the interaction detection method, which will be done later.

# **Chapter 8**

## **Supplementary figures**

**Supplementary figure 1**

**Supplementary figure 2**

**Supplementary figure 3**

**Supplementary figure 4**

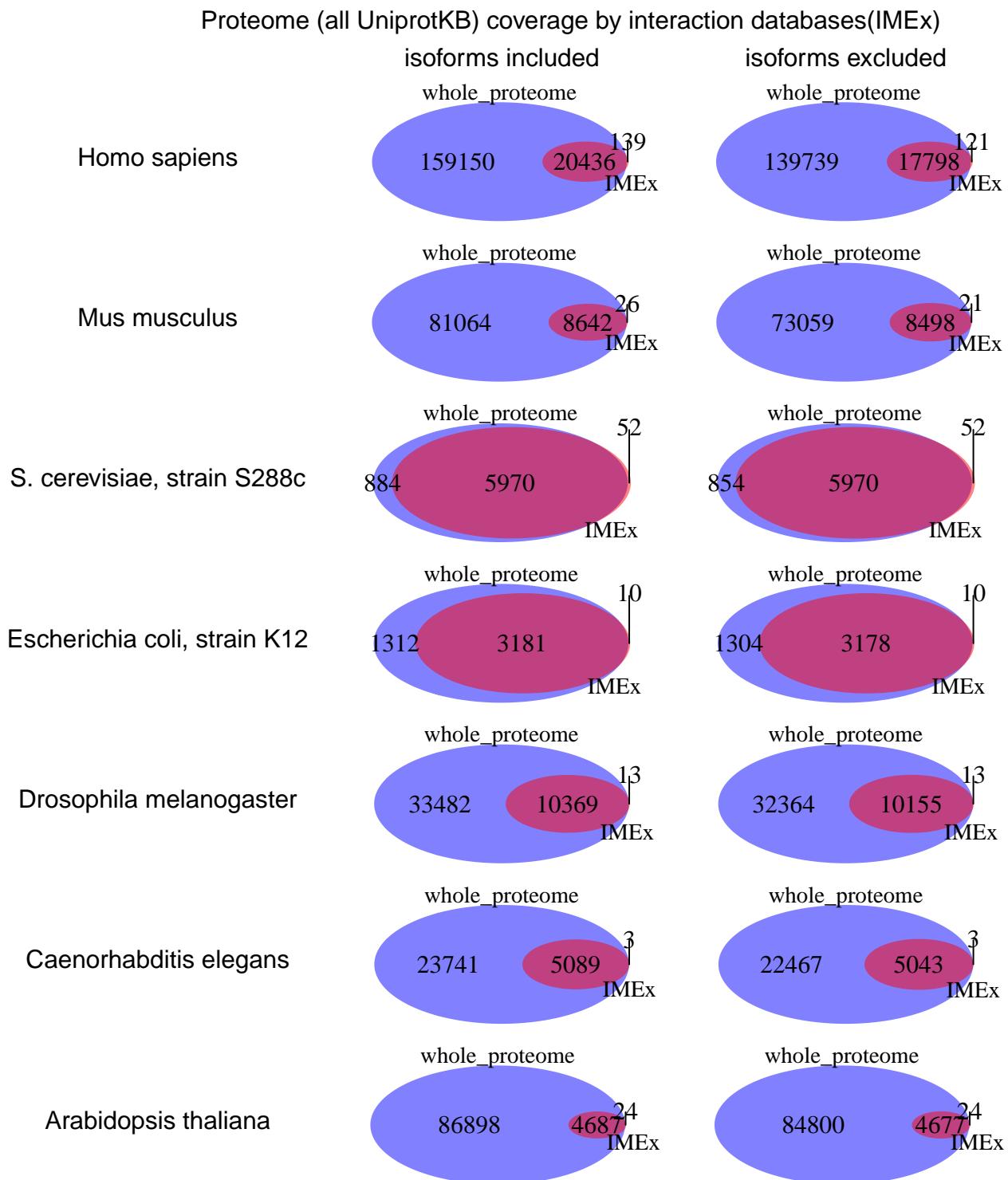


Figure 8.1: Supplementary figure 1. The overlap between proteins with known interactions (interactome) and all proteins included in UniprotKB (proteome)

The number of interacting partners which have UniprotKB identifiers and belong to the species of interest

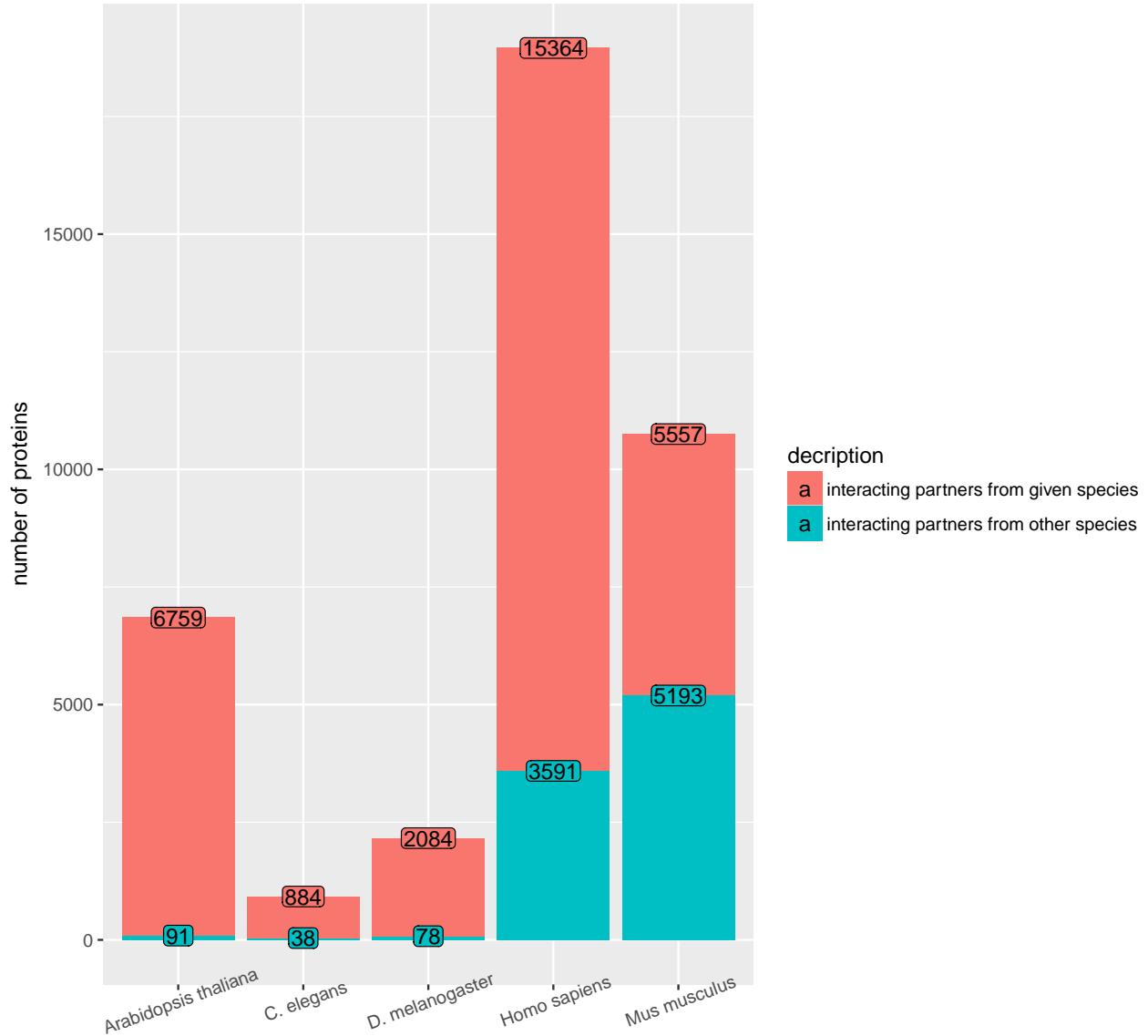


Figure 8.2: Supplementary figure 2. The number of interacting proteins from given species or the other species in BioGRID

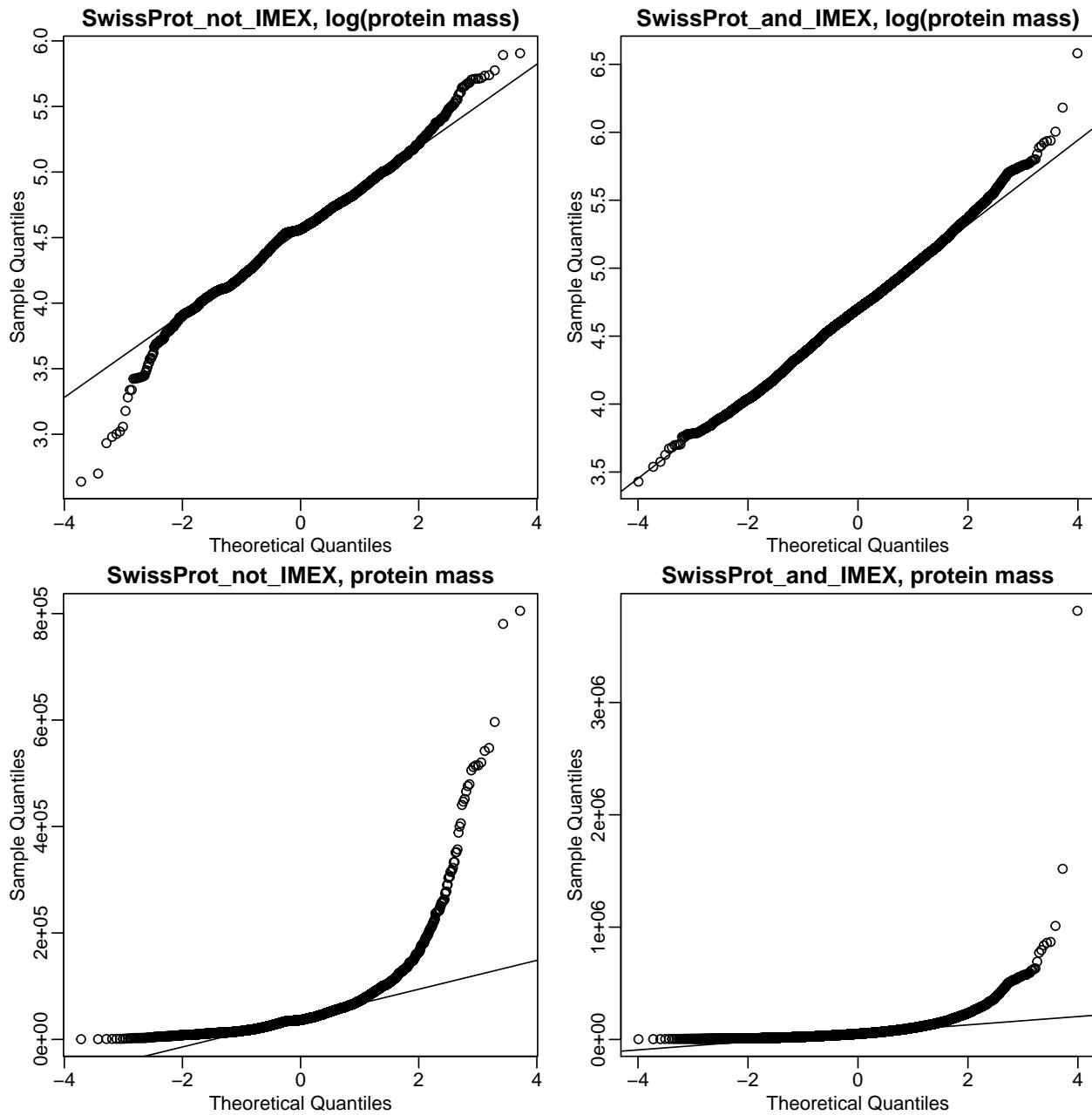


Figure 8.3: Supplementary figure 3. Proteomes have much more long proteins than normal distribution would predict (proteins, dots, are above the line in qqplot). The distribution of logarithm base 10 of protein mass is approximately normal

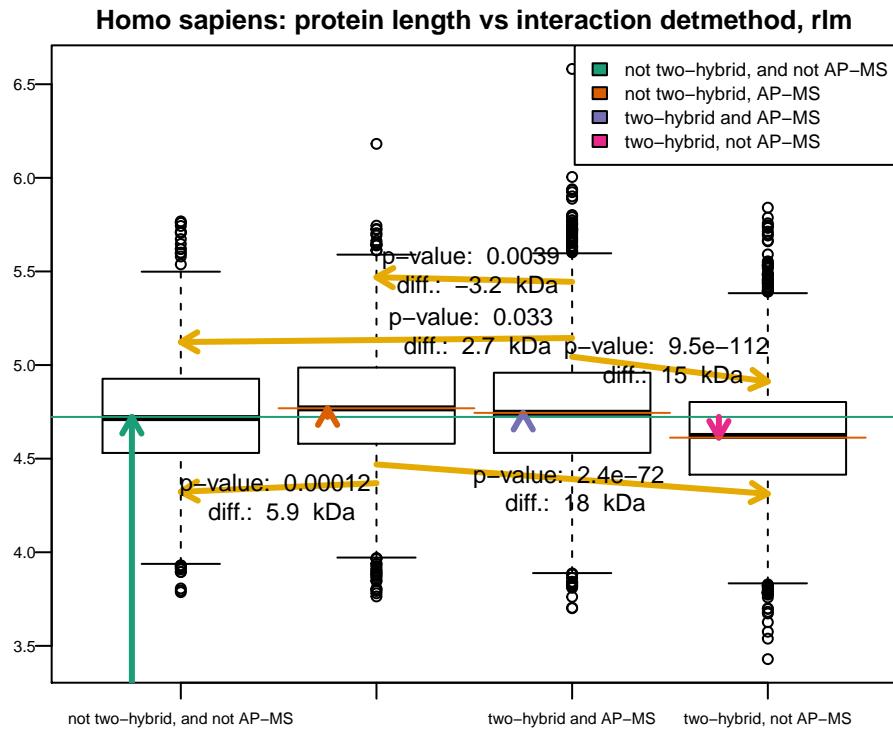


Figure 8.4: Supplementary figure 4. The results of statistical tests can be visualised on a plot. In the box plot below, each vertical arrow points to the median mass of the proteins in a particular group. The middle line of the boxplot represents a median which happens to coincide with mean because the distribution of logarithm base 10 of mass is approximately normally distributed. Each yellow arrow shows the p-value and the difference in protein mass between two corresponding groups.

# Bibliography

M Madan Babu. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem Soc Trans*, 44(5):1185–1200, Oct 2016. ISSN 1470-8752 (Electronic); 0300-5127 (Linking). doi: 10.1042/BST20160172.

Pascal Braun, Murat Tasan, Matija Dreze, Miriam Barrios-Rodiles, Irma Lemmens, Haiyuan Yu, Julie M Sahalie, Ryan R Murray, Luba Roncari, Anne-Sophie de Smet, Kavitha Venkatesan, Jean-Francois Rual, Jean Vandenhoute, Michael E Cusick, Tony Pawson, David E Hill, Jan Tavernier, Jeffrey L Wrana, Frederick P Roth, and Marc Vidal. An experimentally derived confidence score for binary protein-protein interactions. *Nat Methods*, 6(1):91–97, Jan 2009. ISSN 1548-7105 (Electronic); 1548-7091 (Linking). doi: 10.1038/nmeth.1281.

Marija Buljan, Guilhem Chalancon, Sebastian Eustermann, Gunter P Wagner, Monika Fuxreiter, Alex Bateman, and M Madan Babu. Tissue-specific splicing of disordered segments that embed binding motifs rewrites protein interaction networks. *Mol Cell*, 46(6):871–883, Jun 2012. ISSN 1097-4164 (Electronic); 1097-2765 (Linking). doi: 10.1016/j.molcel.2012.05.039.

Alberto Calderone, Luisa Castagnoli, and Gianni Cesareni. mentha: a resource for browsing integrated protein-interaction networks. *Nat Methods*, 10(8):690–691, Aug 2013. ISSN 1548-7105 (Electronic); 1548-7091 (Linking). doi: 10.1038/nmeth.2561.

Marc Carlson. *GO.db: A set of annotation maps describing the entire Gene Ontology*, 2017. R package version 3.4.1.

Andrew Chatr-Aryamontri, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K Kolas, Lara O'Donnell, Sara Oster, Chandra Theesfeld, Adnane Sellam, Chris Stark, Bobby-Joe Breitkreutz, Kara Dolinski, and Mike Tyers. The biogrid interaction database: 2017 update. *Nucleic Acids Res*, 45(D1):D369–D379, Jan 2017. ISSN 1362-4962 (Electronic); 0305-1048 (Linking). doi: 10.1093/nar/gkw1102.

Antonio Fabregat, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, Steven Jupe, Florian Korninger, Sheldon McKay, Lisa Matthews, Bruce May, Marija Milacic, Karen Rothfels, Veronica Shamovsky, Marissa Webber, Joel Weiser, Mark Williams, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Res*, 44(D1):

D481–7, Jan 2016. ISSN 1362-4962 (Electronic); 0305-1048 (Linking). doi: 10.1093/nar/gkv1351.

Consortium Gene Ontology. Gene ontology consortium: going forward. *Nucleic Acids Res*, 43(Database issue):D1049–56, Jan 2015. ISSN 1362-4962 (Electronic); 0305-1048 (Linking). doi: 10.1093/nar/gku1179.

Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004. ISSN 1474-760X (Electronic); 1474-7596 (Linking). doi: 10.1186/gb-2004-5-10-r80.

Marco Y Hein, Nina C Hubner, Ina Poser, Jurgen Cox, Nagarjuna Nagaraj, Yusuke Toyoda, Igor A Gak, Ina Weisswange, Jorg Mansfeld, Frank Buchholz, Anthony A Hyman, and Matthias Mann. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell*, 163(3):712–723, Oct 2015. ISSN 1097-4172 (Electronic); 0092-8674 (Linking). doi: 10.1016/j.cell.2015.09.053.

Henning Hermjakob, Luisa Montecchi-Palazzi, Gary Bader, Jerome Wojcik, Lukasz Salwinski, Arnaud Ceol, Susan Moore, Sandra Orchard, Ugis Sarkans, Christian von Mering, Bernd Roechert, Sylvain Poux, Eva Jung, Henning Mersch, Paul Kersey, Michael Lappe, Yixue Li, Rong Zeng, Debashis Rana, Macha Nikolski, Holger Husi, Christine Brun, K Shanker, Seth G N Grant, Chris Sander, Peer Bork, Weimin Zhu, Akhilesh Pandey, Alvis Brazma, Bernard Jacq, Marc Vidal, David Sherman, Pierre Legrain, Gianni Cesareni, Ioannis Xenarios, David Eisenberg, Boris Steipe, Chris Hogue, and Rolf Apweiler. The hupo psi's molecular interaction format—a community standard for the representation of protein interaction data. *Nat Biotechnol*, 22(2):177–183, Feb 2004. ISSN 1087-0156 (Print); 1087-0156 (Linking). doi: 10.1038/nbt926.

Edward L Huttlin, Lily Ting, Raphael J Bruckner, Fana Gebreab, Melanie P Gygi, John Szpyt, Stanley Tam, Gabriela Zarraga, Greg Colby, Kurt Baltier, Rui Dong, Virginia Guarani, Laura Pontano Vaites, Alban Ordureau, Ramin Rad, Brian K Erickson, Martin Wuhr, Joel Chick, Bo Zhai, Deepak Kolippakkam, Julian Mintseris, Robert A Obar, Tim Harris, Spyros Artavanis-Tsakonas, Mathew E Sowa, Pietro De Camilli, Joao A Paulo, J Wade Harper, and Steven P Gygi. The biplex network: A systematic exploration of the human interactome. *Cell*, 162(2):425–440, Jul 2015. ISSN 1097-4172 (Electronic); 0092-8674 (Linking). doi: 10.1016/j.cell.2015.06.043.

Ruth Isserlin, Rashad A El-Badrawi, and Gary D Bader. The biomolecular interaction network database in psi-mi 2.5. *Database (Oxford)*, 2011:baq037, 2011. ISSN 1758-0463 (Electronic); 1758-0463 (Linking). doi: 10.1093/database/baq037.

Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Kegg as a reference resource for gene and protein annotation. *Nucleic Acids Res*, 44(D1):D457–62, Jan 2016. ISSN 1362-4962 (Electronic); 0305-1048 (Linking). doi: 10.1093/nar/gkv1070.

Irma Lemmens, Sam Lievens, and Jan Tavernier. Mappit, a mammalian two-hybrid method for in-cell detection of protein-protein interactions. *Methods Mol Biol*, 1278:447–455, 2015. ISSN 1940-6029 (Electronic); 1064-3745 (Linking). doi: 10.1007/978-1-4939-2425-7{\\_\}29.

Taibo Li, Rasmus Wernersson, Rasmus B Hansen, Heiko Horn, Johnathan Mercer, Greg Slodkowicz, Christopher T Workman, Olga Rigina, Kristoffer Rapacki, Hans H Staerfeldt, Soren Brunak, Thomas S Jensen, and Kasper Lage. A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat Methods*, 14(1):61–64, Jan 2017. ISSN 1548-7105 (Electronic); 1548-7091 (Linking). doi: 10.1038/nmeth.4083.

Dattatreya Mellacheruvu, Zachary Wright, Amber L Couzens, Jean-Philippe Lambert, Nicole A St-Denis, Tuo Li, Yana V Miteva, Simon Hauri, Mihaela E Sardiu, Teck Yew Low, Vincentius A Halim, Richard D Bagshaw, Nina C Hubner, Abdallah Al-Hakim, Annie Bouchard, Denis Faubert, Damian Fermin, Wade H Dunham, Marilyn Goudreault, Zhen-Yuan Lin, Beatriz Gonzalez Badillo, Tony Pawson, Daniel Durocher, Benoit Coulombe, Ruedi Aebersold, Giulio Superti-Furga, Jacques Colinge, Albert J R Heck, Hyungwon Choi, Matthias Gstaiger, Shabaz Mohammed, Ileana M Cristea, Keiryn L Bennett, Mike P Washburn, Brian Raught, Rob M Ewing, Anne-Claude Gingras, and Alexey I Nesvizhskii. The crapome: a contaminant repository for affinity purification-mass spectrometry data. *Nat Methods*, 10(8):730–736, Aug 2013. ISSN 1548-7105 (Electronic); 1548-7091 (Linking). doi: 10.1038/nmeth.2557.

Sandra Orchard, Samuel Kerrien, Sara Abbani, Bruno Aranda, Jignesh Bhate, Shelby Bidwell, Alan Bridge, Leonardo Briganti, Fiona S L Brinkman, Gianni Cesareni, Andrew Chatr-aryamontri, Emilie Chautard, Carol Chen, Marine Dumousseau, Johannes Goll, Robert E W Hancock, Linda I Hannick, Igor Jurisica, Jyoti Khadake, David J Lynn, Usha Mahadevan, Livia Perfetto, Arathi Raghunath, Sylvie Ricard-Blum, Bernd Roechert, Lukasz Salwinski, Volker Stumpflen, Mike Tyers, Peter Uetz, Ioannis Xenarios, and Henning Hermjakob. Protein interaction data curation: the international molecular exchange (imex) consortium. *Nat Methods*, 9(4):345–350, Apr 2012. ISSN 1548-7105 (Electronic); 1548-7091 (Linking). doi: 10.1038/nmeth.1931.

Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi del Toro, Margaret Duesbury, Marine Dumousseau, Eugenia Galeota, Ursula Hinz, Marta Iannuccelli, Sruthi Jagannathan, Rafael Jimenez, Jyoti Khadake, Astrid Lagreid, Luana Licata, Ruth C Lovering, Birgit Meldal, Anna N Melidoni, Mila Milagros, Daniele Peluso, Livia Perfetto, Pablo Porras, Arathi Raghunath, Sylvie Ricard-Blum, Bernd Roechert, Andre Stutz, Michael Tognoli, Kim van Roey, Gianni Cesareni, and Henning Hermjakob. The mintact project—intact as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res*, 42(Database issue):D358–63, Jan 2014. ISSN 1362-4962 (Electronic); 0305-1048 (Linking). doi: 10.1093/nar/gkt1115.

Suraj Peri, J Daniel Navarro, Troels Z Kristiansen, Ramars Amanchy, Vineeth Surendranath, Babylakshmi Muthusamy, T K B Gandhi, K N Chandrika, Nandan Deshpande, Shubha Suresh, B P Rashmi, K Shanker, N Padma, Vidya Niranjan, H C Harsha, Naveen Talreja,

B M Vrushabendra, M A Ramya, A J Yatish, Mary Joy, H N Shivashankar, M P Kavitha, Minal Menezes, Dipanwita Roy Choudhury, Neelanjana Ghosh, R Saravana, Sreenath Chandran, Sujatha Mohan, Chandra Kiran Jonnalagadda, C K Prasad, Chandan Kumar-Sinha, Krishna S Deshpande, and Akhilesh Pandey. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*, 32(Database issue):D497–501, Jan 2004. ISSN 1362-4962 (Electronic); 0305-1048 (Linking). doi: 10.1093/nar/gkh070.

Punit Saraon, Ingrid Grozavu, Sang Hyun Lim, Jamie Snider, Zhong Yao, and Igor Stagljar. Detecting membrane protein-protein interactions using the mammalian membrane two-hybrid (mamth) assay. *Curr Protoc Chem Biol*, 9(1):38–54, Mar 2017. ISSN 2160-4762 (Electronic); 2160-4762 (Linking). doi: 10.1002/cpch.15.

Martin H Schaefer, Luis Serrano, and Miguel A Andrade-Navarro. Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Front Genet*, 6:260, 2015. ISSN 1664-8021 (Linking). doi: 10.3389/fgene.2015.00260.

Paul Shannon. *PSICQUIC: Proteomics Standard Initiative Common QUery InterfaCe*, 2017. R package version 1.14.0.

Consortium The UniProt. Uniprot: the universal protein knowledgebase. *Nucleic Acids Res*, 45(D1):D158–D169, Jan 2017. ISSN 1362-4962 (Electronic); 0305-1048 (Linking). doi: 10.1093/nar/gkw1099.

J M Villaveces, R C Jimenez, P Porras, N Del-Toro, M Duesbury, M Dumousseau, S Orchard, H Choi, P Ping, N C Zong, M Askenazi, B H Habermann, and Henning Hermjakob. Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. *Database (Oxford)*, 2015, 2015. ISSN 1758-0463 (Electronic); 1758-0463 (Linking). doi: 10.1093/database/bau131.

Michael Vincent and Santiago Schnell. A collection of intrinsic disorder characterizations from eukaryotic proteomes. *Sci Data*, 3:160045, Jun 2016a. ISSN 2052-4463 (Electronic); 2052-4463 (Linking). doi: 10.1038/sdata.2016.45.

MP Vincent and S Schnell. Data from: A collection of intrinsic disorder characterizations from eukaryotic proteomes, 2016b. URL <http://dx.doi.org/10.5061/dryad.sm107>.

Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterProfiler: an r package for comparing biological themes among gene clusters. *OMICS*, 16(5):284–287, May 2012. ISSN 1557-8100 (Electronic); 1536-2310 (Linking). doi: 10.1089/omi.2011.0118.

Wangshu Zhang, Fengzhu Sun, and Rui Jiang. Integrating multiple protein-protein interaction networks to prioritize disease genes: a bayesian regression approach. *BMC Bioinformatics*, 12 Suppl 1:S11, Feb 2011. ISSN 1471-2105 (Electronic); 1471-2105 (Linking). doi: 10.1186/1471-2105-12-S1-S11.