CORE
Metadata, citation and similar papers at core.ac.uk
Provided by University of Birmingham Research Portal

UNIVERSITY OF BIRMINGHAM

Research at Birmingham

Identifying novel protein interactions

Goncalves Carneiro, Vítor; Clarke, Thomas; Davies, Clare; Bailey, Dalan

DOI:
10.1016/j.ymeth.2015.08.022

License:
Creative Commons: Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)

Document Version
Peer reviewed version

Citation for published version (Harvard):
Gonçalves Carneiro, V, Clarke, T, Davies, CC & Bailey, D 2015, 'Identifying novel protein interactions: proteomic methods, optimisation approaches and data analysis pipelines', Methods.
https://doi.org/10.1016/j.ymeth.2015.08.022

Link to publication on Research at Birmingham portal

Publisher Rights Statement:
Eligibility for repository: Checked on 29/10/2015

General rights
Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

• Users may freely distribute the URL that is used to identify this publication.
• Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
• User may use extracts from the document in line with the concept of ‘fair dealing’ under the Copyright, Designs and Patents Act 1988 (?)
• Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

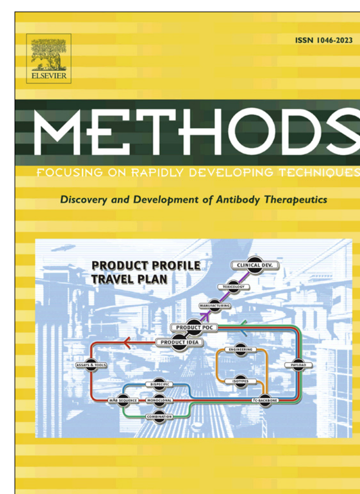
When citing, please reference the published version.

Take down policy
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

Download date: 01. Feb. 2019

Accepted Manuscript



Identifying novel protein interactions: proteomic methods, optimisation approaches and data analysis pipelines

Daniel Gonçalves Carneiro, Thomas Clarke, Clare C. Davies, Dalan Bailey

PII: S1046-2023(15)30066-9
DOI: http://dx.doi.org/10.1016/j.ymeth.2015.08.022
Reference: YMETH 3779

To appear in: *Methods*

Received Date: 30 June 2015
Revised Date: 26 August 2015
Accepted Date: 27 August 2015

Please cite this article as: D.G. Carneiro, T. Clarke, C.C. Davies, D. Bailey, Identifying novel protein interactions: proteomic methods, optimisation approaches and data analysis pipelines, *Methods* (2015), doi: http://dx.doi.org/10.1016/j.ymeth.2015.08.022

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

1 **Identifying novel protein interactions: proteomic methods, optimisation approaches and data**
2 **analysis pipelines**

3 Daniel Gonçalves Carneiro[1], Thomas Clarke[1], Clare C Davies*[1] and Dalan Bailey*[1]

4

5 [1] College of Medical and Dental Sciences
6 University of Birmingham
7 Edgbaston
8 Birmingham
9 B15 2TT
10 UK

11 * Co-corresponding authors: email d.bailey@bham.ac.uk; c.c.davies@bham.ac.uk

12 Telephone: +441214146854

13 **Abstract:**

14 The technological revolution in high-throughput nucleic acid and protein analysis in the last 15 years
15 has launched the field of 'omics and led to great advances in our understanding of cell biology.
16 Consequently the study of the cellular proteome and protein dynamics, in particular interactomics,
17 has been a matter of intense investigation, specifically the determination and description of complex
18 protein interaction networks in the cell, not only with other proteins but also with RNA and DNA. The
19 analysis of these interactions, beginning with their identification and ultimately resulting in structural
20 level examination, is one of the cornerstones of modern biological science underpinning basic
21 research and impacting on applied biology, biomedicine and drug discovery. In this review we
22 summarise a selection of emerging and established techniques currently being applied in this field
23 with a particular focus on affinity-based purification systems and their optimisation, including tandem
24 affinity purification (TAP) tagging, isolation of proteins on nascent DNA (IPOND) and RNA-Protein
25 immunoprecipitation in tandem (RIPiT). The recent application of quantitative proteomics to improve
26 stringency and specificity is also discussed, including the use of metabolic labelling by stable
27 isotope labelling by amino acids in cell culture (SILAC), localization of organelle proteins by isotope
28 tagging (LOPIT) and proximity-dependent biotin identification (BioID). Finally, we describe a range
29 of software resources that can be applied to interactomics, both to handle raw data and also to
30 scrutinise its broader biological context. In this section we focus especially on open-access online
31 interactomic databases such as Reactome and IntAct.

32 **Keywords:** interactomics, protein-protein interaction, RNA-protein interaction, DNA-protein
33 interactions, affinity purification, TAP-tagging

ACCEPTED MANUSCRIPT

34 **Introduction:**

35 Fundamentally, cellular life depends on a continuing network of molecular interactions among
36 proteins, nucleic acids and small molecules. Interactions of this nature underpin the key processes
37 of transcription and translation and are essential for supporting cellular structure, metabolism, cell
38 signalling and all other aspects of cell biology. Aberrant interactions, caused by genetic or
39 environmental stimuli, are the cause of most human diseases such as cancer or auto-immunity,
40 while all pathogens, in one way or another, hijack these interactions to promote their own growth
41 over that of the host. As a result the description of these interactions is, to some degree, the central
42 principle of all molecular biology research. Their isolation, identification and characterisation can be
43 either straightforward or complex; this tenet being dependent on the very nature of the interaction
44 itself.

45 For the purposes of this review we have focused on examining techniques designed to identify three
46 of the major types of interactions found in the cell: Protein-protein, DNA-protein and RNA-protein.
47 This obviously excludes other interactions such as small molecules and proteins; however, these
48 have been excellently addressed elsewhere [1-3]. In this article we will particularly focus on recent
49 advancements in these fields, considerations that should be made on the choice of methodology as
50 well as potential limitations to bear in mind when designing experiments (overview; Figure 1). We
51 will also provide a basic overview of these techniques and their application as a starting point for
52 researchers, e.g. structural biologists, looking to implement these technologies in their research
53 along with a list of resources to facilitate downstream data processing.

54 **Methods For Identifying Protein-Protein Interactions:**

55 The two commonly used approaches for discovery of protein-protein interactions are: (1) the use of
56 recombinant protein libraries that can be screened with a bait protein or (2) the isolation and *de*
57 *novo* detection of protein complexes using mass spectrometry (MS).

58

59 **Yeast-Two-Hybrid:** Pioneering interactomic techniques, such as yeast two-hybrid (Y2H), were
60 important initial tools for dissecting protein-protein interactions in cells [4]. In this system, two
61 recombinant fusion proteins are generated: the bait is fused with a DNA-binding domain and the
62 prey with a transcription activation domain. These two elements are expressed in yeast along with a
63 reporter gene whose regulation is under a promoter recognised by the recombinant transcription
64 activation domain. If these two proteins interact with each other, the transcription activation domain
65 becomes functional leading to expression of the reporter gene. Despite being a robust assay, this
66 system is limited by the preference towards high affinity interactions that happen in the nucleus, by
67 protein complexes that form in cellular membranes or cases where proteins require more than one
68 binding partner [5]. Recent advances in the field have led to the development of variations of Y2H,

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

69 specialising in the recruitment of the bait and prey to the plasma membrane [6], cytosol [7, 8] or the
70 endoplasmic reticulum [9], and also expanding the number of proteins included, such as multiple
71 baits [10]. Nonetheless, this technique is still largely limited by the choice of library. Identifying
72 protein-protein interactions relies on specific properties of proteins beyond their peptide sequence,
73 such as protein abundance, cellular localization, post-translational modifications and isoforms or
74 variants; most of which can be missed by surrogate systems as Y2H [11]. Some of these limitations
75 have been addressed with the development of more powerful hybrid technologies that can
76 successfully isolate macromolecular protein assemblies, i.e. by using split-ubiquitin membrane yeast
77 two-hybrid (MYTH) where bait and prey proteins are tagged with split ubiquitin [12]. Specific
78 protein-protein interactions result in reconstitution of the ubiquitin and cleavage – a process that
79 releases a fused transcription factor and leads to reporter protein activation. This approach has
80 proved particularly successful for isolating membrane protein interactions since normal Y2H
81 targeting of hydrophobic proteins to the nucleoplasm can be inefficient. Subsequent improvements
82 have included the use of genomically integrated baits (iMYTH) [13] to provide more endogenous
83 protein levels as well as adaptation of the system to mammalian cells (MaMTH) [14].
84
85 **Protein microarrays**: The advent of high-throughput technology has allowed the rapid screening of
86 protein interactions in different organisms. Protein microarrays, in particular, are increasingly being
87 used to study protein interactions with applications being extended to diagnostic [15] or enzymatic
88 discovery [16]. In a protein microarray, which is a derivation of the classical DNA chip for detection
89 of RNA, a library of complete or functional domains of recombinant proteins is printed onto a slide
90 [17]. A tagged bait protein is then passed over the slide allowing it to specifically interact with a
91 relevant prey protein, before detection by tag-specific immunoblotting. Protein microarrays have
92 allowed the determination of not only protein-protein interactions in the context of diverse network
93 modalities and functional modifications (reviewed in [18]) but also RNA-protein [19] and DNA-protein
94 interactions [20, 21]. Although an increasing diversity of microarray chips are commercially available
95 at competitive prices, these methods are limited by the necessity to reproduce a putative protein
96 interaction *in vitro* and the recapitulation of the physiochemical and stoichiometric environment
97 necessary for the interaction.
98
99 **Mass spectrometry-based approaches:** Most affinity purification techniques rely on the isolation
100 of material for subsequent MS-based protein identification. In this process, a complex protein
101 sample is separated by size and isoelectric point - both 2D gel electrophoresis and liquid
102 chromatography (LC) are common practices in proteomics. This is followed by proteolysis often
103 through tryptic digestion, a process that generates small peptides which are subsequently exposed
104 to ionizing radiation and analysed by a mass spectrometer. During this procedure each peptide is
105 broken into smaller ions whose mass-to-charge (m/z) ratios are used to extrapolate the peptide

ACCEPTED MANUSCRIPT

106 sequence [22]. The identification and quantification of these peptides, achieved through
107 computational comparison with protein databases, allows the reconstruction of a putative protein
108 profile for a sample. However, the identification of a specific protein can lead to false-positives due
109 to uncategorised isoforms or protein variants (since a peptide can originate from more than one
110 protein) [23, 24]. The use of LC-MS to study protein-protein interactions has two main limitations: (1)
111 the relative quantification and identification of specific versus non-specific interactions and (2) the
112 overall complexity of the protein profile in the sample. Other limitations do also exist: the
113 identification of low-yield proteins, the transient state of protein interactions, and the requirement for
114 co-factors and/or post-translation modifications; however, these are often related to specific
115 interactions and will not be addressed in any detail in this review. Of note, with regards to identifying
116 low-yield proteins, the improved sensitivity of modern MS technology is somewhat rectifying these
117 issues, as is the availability of improved affinity purification reagents and techniques. Nevertheless,
118 care should always be taken when designing interactomics experiments to consider the amount of
119 starting material required, often cell number.

120 **Immunoprecipitation techniques:** The application of new protein-protein interaction techniques
121 combined with affordable LC and MS approaches has revolutionized interactomics research. The
122 most commonly used technique for isolating protein interactors is affinity purification, usually
123 employing ectopically-expressed tagged (e.g. GFP, HA, Flag) bait proteins. This process has
124 advantages over Y2H in that it allows the efficient study of proteins in more physiological situations
125 such as mammalian cells. In this methodology, the protein of interest and any complex it forms in
126 the cell are recovered by the use of high affinity reagents (e.g. antibodies or peptides coupled to
127 agarose beads) to the tag. This reliable and flexible approach has been applied in a range of
128 biological contexts to study both small protein complexes as well as global protein-protein
129 interaction networks [25, 26]. The clear strength of this method is that it allows for the enrichment of
130 final protein complexes that occur less frequently or have low abundancy in the cell; however,
131 optimization of the purification protocol is essential to simultaneously maintain the protein
132 complexes but also to eliminate possible contamination through non-specific protein interactions
133 with the bait, tag or affinity matrix.

134 **Tandem Affinity Purification:** Issues with interaction stringency during immunoprecipitation can be
135 resolved using tandem affinity purification (TAP) [27]. This technology relies on the application of
136 dual affinity tagged-fusion proteins containing a strong antigenic region (such as Protein G, GFP,
137 etc.) fused to a separate tag, such as streptavidin or calmodulin-binding peptide. The two tags are
138 normally separated by a specific exogenous cleavage site, such as the Tobacco Etch Virus or
139 PreScission (GE Healthcare) or Rhinovirus 3C protease cleavage sites. After expression of a TAP-
140 tagged bait, protein complexes are isolated by two separate affinity purifications. The cleavable
141 linkers and small peptide tags used in the second purification step allow specific elutions to be

ACCEPTED MANUSCRIPT

142 performed after each purification further increasing the specificity of the pull-downs, reviewed in
143 more detail here [28, 29]. This technique has been used to study protein-protein interactions in a
144 variety of fields, such as virology, oncology and cell biology and is one of the most popular methods
145 for identifying new protein binding partners [30-33].

146 Despite its broad application, large-scale analysis of the interactome using TAP tagging is time
147 consuming and expensive. In addition, due to the thorough purification technique, weaker or
148 transient binding partners may not be detected in the final protein preparation. Moreover, the use of
149 recombinant proteins, generally overexpressed from complementary DNA, can affect localization,
150 endogenous abundance and hence protein complex stoichiometry, turnover and the general
151 function of the protein of interest, ultimately influencing the nature of the protein-protein interaction.
152 It is therefore essential that any protein interaction identified should be subsequently confirmed
153 under more physiological conditions, i.e. with endogenous untagged protein. One emerging
154 technique to counter some of these issues is to normalise the expression of the tagged protein to a
155 more physiologically relevant level. Bacterial artificial chromosomes (BACs), large enough to
156 contain the gene of interest as well as its endogenous regulatory sequences, have recently been
157 applied to assist this normalisation. In fact, BACs libraries fused with TAP tags have facilitated high-
158 throughput interactomic studies of different cells [34, 35].

159 **Methods For Identifying DNA-Protein Interactions:**

160 The identification of DNA-protein interactions underpins basic research in a number of crucial
161 biological fields including DNA replication, DNA repair and chromatin remodelling. As such there is a
162 great interest in developing new techniques to examine this in as close to a physiological setting as
163 possible.

164 **IPOND:** There are a range of emerging techniques to identify proteins that bind DNA; one popular
165 method in recent years has been isolation of proteins on nascent DNA (IPOND). This technique
166 utilises the thymidine analogue 5-ethynyl-2′-deoxyuridine (EdU) which, when added to cells growing
167 in tissue culture, is rapidly incorporated into newly synthesised DNA [36]. Crucially, this occurs
168 without interfering with the replication machinery or causing any measurable activation of the cell's
169 DNA damage response and is therefore not considered deleterious to cell viability and function [37].
170 EdU labelled cells can then be treated with a reversible cross-linking agent to facilitate the isolation
171 of any DNA-interacting proteins [38]. The functional alkyne group in EdU also allows the addition of
172 a biotin tag (through click chemistry) to the nascent DNA which can facilitate purification and the
173 subsequent isolation of protein interactors by immunoblotting or MS.

174 ***Xenopus laevis* eggs:** Another increasingly popular method, particularly in the study of DNA
175 replication, is the application and manipulation of *Xenopus laevis* eggs for DNA-protein

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

176 interactomics. These soluble egg extracts are an attractive model as they undergo the same nuclear 177 transitions of the eukaryotic cell cycle *in vitro* as they do naturally *in vivo* [39]. *Xenopus* eggs are 178 typically around 1mm in diameter and contain a vast array of RNAs and proteins. These are 179 required to synthesise up to 4000 nuclei as a result of 12 synchronous rounds of cell cycle; a 180 process which occurs immediately after fertilisation in the absence of subsequent rounds of 181 transcription [40]. Typically, before fertilisation occurs, *Xenopus* egg extracts are biologically 182 arrested in metaphase of meiosis II [41]. Exogenous DNA added to the egg extracts is first 183 organised into chromatin and then into structures that correspond to those of cells in interphase, 184 thereby enabling DNA duplication to occur [39]. Given that the *Xenopus* egg extract is able to 185 support such a wide range of biological activity, ranging from *in vitro* DNA replication to mitosis, they 186 represent a unique tool for interrogating the dynamics and mechanisms that underpin such 187 processes. In addition, the egg extracts form a soluble soup-like mixture that is amenable to 188 immuno-depletion or supplementation with recombinant protein. Likewise, small molecule inhibitors 189 to various enzymes or DNA damaging agents can be easily added. This further enhances their 190 suitability as a tool to identify and investigate DNA-protein interactions. These techniques have been 191 applied to isolate chromatin and chromatin-associated proteins from egg extracts with DNA-protein 192 interactions subsequently being identified by MS [42] or immunoblotting [43]. In addition, Moreno *et* 193 *al.,* have recently utilised this method to study the role of polyubiquitylation in the termination of DNA 194 replication, demonstrating that inhibition of polyubiquitylation caused a prolonged association of the 195 active helicase with replicating chromatin [44]. As such these methods can answer key questions 196 relating not only to DNA and protein interactions but also to the duration of these interactions and 197 their significance for biological processes.

198 **DNA baits:** Using DNA directly as the bait for identifying DNA-protein interactions is also a popular 199 technique. One such application is the identification of guanine quadruplex (G-quadruplex) binding 200 proteins. G-quadruplexes are unusual secondary structures which typically occur as a result of 201 guanine rich regions within DNA or RNA [45]. Their recent identification in the promoter regions of 202 genes implicated in cancer suggests a regulatory role in transcription [46]. Binding proteins have 203 been identified using pull down assays where G-quadruplexes serve as the bait. In this instance 204 specificity can be determined using mutated oligonucleotides which are unable to form defined G-205 quadruplex structures. In these experiments, which aim to examine how these complex structures 206 affect cellular processes such as transcription and translation, isolated proteins are identified by MS 207 [47].

208 **Methods For Identifying RNA-Protein Interactions:**

209 RNA processing, including alternative mRNA splicing and editing, is central for the increased 210 proteome complexity associated with higher eukaryotes. Consequently, more than 1000 RNA-

ACCEPTED MANUSCRIPT

211 binding proteins (RBPs) are encoded by the genome. However, with the explosion of technological
212 advances enabling whole genome transcriptional analysis, it is becoming increasingly clear that
213 RNA is not solely an intermediate between DNA and protein production. Indeed, whilst only 2% of
214 our genome encodes for proteins, up to 70% is transcribed into non-coding RNA, including long
215 non-coding RNAs (lncRNAs). Both mutations in RBPs and aberrant expression of lncRNAs have
216 been implicated in human disease, such as neurodegeneration and cancer [48]. Despite this, little
217 is known about how RBPs control cell-type specific alternative splicing of mRNAs and lncRNA
218 biology. Interaction studies and structural insights will undoubtedly be central in providing a deeper
219 understanding of these processes. The "omics" era has led to a rapid expansion in methodology to
220 study the RNA-protein landscape. Traditionally, two main approaches are employed: protein-
221 centred, whereby specific proteins serve as a bait to isolate associated RNA which is identified by
222 sequencing; or RNA-centred involving the expression of affinity-coupled RNA and MS to identify
223 proteins that bind a specific RNA species.

224 **Protein-centred approaches:** Chromatin immunoprecipitation (ChIP) is a well-established
225 technique that enables the isolation of protein:DNA complexes via antibody-mediated purification of
226 endogenous or tagged proteins. RNA immunoprecipitation (RIP) is analogous to chromatin
227 immunoprecipitation (ChIP), but in this case, RBPs are affinity purified and their associated RNAs
228 isolated. RNA is then converted to cDNA and target sequences identified by quantitative PCR,
229 cDNA libraries, microarray analysis or RNA-Seq [49]. Unfortunately, RBPs are relatively
230 promiscuous when overexpressed, thus it is important to generate stable cell lines that express
231 levels close to that of endogenous protein. Accordingly, CRISPR technology enabling the
232 introduction of tag sequences within gene sequences upstream of the open reading frame will
233 undoubtedly facilitate this. Of note, His-tagged sequences are not recommended as the positive
234 charge can alter the RNA binding properties of the RBP.

235 Like ChIP, RIP can be performed under native or denaturing conditions; however, whilst native RIP
236 potentially offers physiological isolation of complexes, extensive non-specific interactions that can
237 occur between isolated RBPs and RNA after cell lysis can be problematic [50]. To circumvent this,
238 RNA-protein complexes are cross-linked by treating cells with ultraviolet C (UVC) light, enabling
239 denaturation of samples and stringent lysis conditions (cross-linking and immunoprecipitation, CLIP)
240 [51]. A further development on this strategy is photoactivatable ribonucleoside-enhanced (PAR)-
241 CLIP, enabling single-nucleotide resolution of protein interaction sites and the definition of sequence
242 motifs [52]. However, UVC-induced cross-linking is far more efficient when the RNA-protein
243 interface is composed of nucleic acid bases and aromatic amino acid side chains, which is often but
244 not always the case for RBPs [53]. CLIP techniques are therefore inherently biased towards sites
245 that are more effectively cross-linked, and thus may not be a true representation of the RNA-binding
246 landscape. To overcome this, a combinatorial method of RIP and CLIP has recently been devised,

ACCEPTED MANUSCRIPT

247 RIPiT (RNA-Protein immunoprecipitation in tandem) [54], that involves two sequential
248 immunoprecipitations, and is thus highly analogous to the isolation of protein-protein complexes by
249 TAP-tagging. Importantly, RNA-protein complexes are cross-linked with formaldehyde rather than
250 UVC and the noise-to-signal ratio of RNA-binding footprints is greatly enhanced by the two-stage
251 purification process. Indeed, whilst direct RNA-protein interactions would need verification, the
252 ability to perform sequential immunoprecipitations of two different proteins within an RNA-bound
253 complex will enable detailed and dynamic analysis of multi-subunit RBPs.

254 **RNA-centred approaches:** A reverse but complementary approach to study ribonucleoprotein
255 (RNP) complexes involves identifying trans-regulatory proteins that interact with a specific RNA
256 sequence. Since RNP complexes are highly dynamic and multifaceted, a robust *in vivo* system is
257 required. The "Ribotrap" method, developed by Keene and colleagues [55], was one of the first to
258 exploit the phenomenon of naturally occurring RNA-RBP protein interactions, such as the
259 bacteriophage MS2 viral coat protein that recognises with high specificity an MS2-RNA stem loop
260 structure. By engineering and expressing two constructs, one that will express the MS2-RNA stem
261 loop conjugated to the RNA bait and the other that will express the MS2 protein conjugated to an
262 affinity tag (e.g. GST, protein A), *in vivo* RNA-RBP complexes can be isolated by affinity purification.
263 This approach has been successfully applied to identify mRNAs [56], and more notably in the
264 identification of lncRNA-interacting proteins that contribute to breast cancer metastasis [57].
265 However, one caveat of this system is the use of affinity tags that reduce specificity and limit the
266 choice of elution buffers. A recent development has been the use of an engineered Cys4 CRISPR
267 endoribonuclease protein that specifically binds with high affinity to a short 16-nt hairpin sequence
268 that can be conjugated to the RNA bait via *in vitro* transcription [58]. Importantly, two point mutations
269 in Cys4 significantly enhanced its competencies for RNP complex capture. The first mutation, Ser50
270 to cysteine, enables coupling to biotin, and thus immobilisation of the Cys4-RNP complex on avidin
271 resin. The second one, His29 to alanine, inactivates RNA strand scission, which can be reversed
272 with imidazole. This modified Cys4 thus has the ability to selectively tether RNA bait whilst enabling
273 the release of RNA-protein complexes via inducible imidazole-mediated cleavage of the 16-nt
274 hairpin sequence. Crucially, direct comparison of Cys4-based and biotinylated RNA affinity
275 purification strategies revealed a dramatic decrease in the amount of background contamination to
276 such an extent that the gel purification step after affinity purification can be omitted [58]. Although
277 the expression of Cys4-tagged transcripts in cells has yet to be used to purify native RNPs, one can
278 anticipate that such a method will certainly provide new insight into the composition of RNA-protein
279 complexes, particularly in the context of lncRNA biology.

280 **Methods For Improving Stringency and Specificity During Affinity Purification:**

281

ACCEPTED MANUSCRIPT

282 **Relative quantification:** Broadly speaking the relative quantification of proteins in a sample
283 provides a wider understanding of the proteome. This is especially true within the context of
284 interactomics analysis where relative quantification can be applied to increase confidence during the
285 identification of protein interactions [59]. This is achieved by mixing differentially labelled lysates
286 during affinity purification to identify specific and non-specific interactions with the tag, bait and/or
287 affinity matrix. This labelling is achieved through marking of specific amino acids and can be
288 performed either before collection of the sample - *metabolic labelling*, when labels are incorporated
289 into the organism or tissue through modification of the nutrients - or at the moment of sample
290 preparation - *chemical labelling*, in which labels are added *in vitro* to the existent protein pool. Stable
291 isotope labelling by amino acids in cell culture (SILAC) [60] and its derivatives, such as ^{15}N-labelling
292 [61, 62] and NeuCode SILAC [63], are technologies in which cells, or even entire organisms [64,
293 65], are grown in media supplemented with arginine and lysine residues containing combinations of
294 ^{13}C and ^{15}N, so proteins with altered masses can be generated. At the point of LC-MS, each
295 tryptically digested peptide contains at least one of these labelled amino acids, inducing small shifts
296 in the detected m/z ratio and enabling relative quantification of peptides from differently labelled
297 samples, e.g. tag-alone versus tagged-bait. These SILAC-based approaches for studying protein-
298 protein interactions have been successfully applied in a variety of fields, such as infectious diseases
299 [66, 67] and cancer [68, 69]. Established SILAC reagents have only allowed, until recently, the
300 analysis of three samples simultaneously, generically known as light, medium and heavy labels;
301 however, new developments in the field have increased this number greatly, i.e. NeuCode (neutron
302 encoded) SILAC uses a collection of isotopologues of lysine with subtle mass differences achieved
303 by the incorporation of deuterium into its side chain, and can be used to analyse up to 18 different
304 samples in one experiment [63, 70].

305 *In vitro* chemical labelling offers, in general, higher multiplicity of samples and flexibility in the design
306 of the experimental set-up. In this process, proteins are collected and covalently linked to a specific
307 tag that directly alters the m/z ratio of targeted peptides. Common practises include dimethyl-
308 labelling in which free amino groups (such as side chains of lysines and the N-terminus of the
309 peptides) react with isotopologues of cyanoborohydride and formaldehyde, generating
310 dimethylamines, whose mass shift is detected by MS [71]. This technique is inexpensive, easy to
311 establish and has comparable accuracy to SILAC [72]. Other tags have also been developed to
312 generate higher levels of multiplicity. The most commonly used compounds include two derivatives
313 of *N*-hydroxysuccinimide: tandem mass tags (TMT) [73], which use a combination of amino-reactive
314 compounds to analyse up to 10 different samples, and isobaric tags for relative and absolute
315 quantification (iTRAQ) [74] for up to 8 samples.

316 **Reducing protein complexity:** Another frequent problem in the study of protein interactions is the
317 overall complexity of the final sample. Several approaches can be taken to focus the specificity,

ACCEPTED MANUSCRIPT

318 frequency and localization of protein interactions including cell fractionation and labelling by
319 proximity. The advantages of these approaches are that they reduce the potential number of hits in
320 interactomic screens to more manageable levels. For instance using fractionation, it is possible to
321 restrict protein populations to a particular region of the cell where the tagged or endogenous protein
322 of interest is localised, therefore facilitating the identification of interacting partners and the reduction
323 in background noise or non-specific interactions. The development of physical separation protocols,
324 such as differential and gradient centrifugation, in combination with biochemical assays - that rely on
325 the solubility of certain proteins in a variety of detergents, salt concentrations and buffers – allows
326 the isolation of organelles or specific regions of the cell, such as lipid rafts [75] or the centrosome
327 [76]. A method of particular relevance is protein correlation profiling (PCP) which relies on the
328 proteomic characterization of different cellular fractions separated by gradient centrifugation and
329 identified by immunoblotting [77]. Using these fractionation systems, the enrichment of the protein
330 profile of a specific organelle or large protein complexes can be used to study protein-protein
331 interactions; however, there are disadvantages including insufficient lysis, dissociation of complexes
332 and/or re-localization of proteins to other fractions during lysis or as a result of changes in ion
333 concentrations, pH, etc. [78, 79]. Emerging technologies to control for these limitations include
334 localization of organelle proteins by isotope tagging (LOPIT). In this method *in vitro* chemical
335 labelling is used to quantify the abundance of proteins along a density gradient. This pattern can
336 then be compared to the position of reported identified proteins in the same gradient, assuming that
337 proteins of the same organelle will localise in the same fraction [80-82].

338 **Proximity labelling:** Other methods use localised protein-tagging to mark and differentiate protein
339 complexes in the cell. The success of this approach relies on the restricted modification of proteins
340 in a specific region of the cell or proteins surrounding the bait. The most common reaction catalysed
341 in this approach is the biotinylation of protein complexes that can be easily recovered after
342 fractionation or stringent purification. In 2013, Rhee and colleagues used endogenous biotinylation
343 to determine the proteome of human mitochondria [83]. In this study, ascorbate peroxidase, an
344 enzyme that oxidises phenolic compounds, was retargeted to the inner surface of the mitochondrion
345 to specifically biotinylate proteins in close proximity after exposure to biotin-phenol. After
346 fractionation the mitochondria were recovered and biotinylated proteins from these organelles were
347 purified using streptavidin beads, highlighting the power of this approach (see Figure 2).

348 Other recent advancements in labelling include proximity-dependent biotin identification (BioID)
349 which allows the biotinylation of lysines in proteins surrounding a bait tagged with biotin-ligase from
350 *E. coli* [84]. Biotinylation is triggered by incubation with biotin and labelled proteins can be
351 stringently purified due to the strong interaction of streptavidin with biotin. This method has been
352 used to identify interaction partners of several protein substrates as well as larger protein complexes
353 in the context of infection [85] and cancer [86, 87]. Other techniques for labelling by proximity

ACCEPTED MANUSCRIPT

354 include the use of cross-linking compounds combined with MS. These approaches are ideal for
355 identifying protein interactions and studying the structure of large protein complexes [88, 89].

356 **Data handling: integration, visualisation and network/pathway analysis**

357 There are various tools currently available for analysing the results of protein interaction studies.
358 These are often designed with broad spectrum researchers in mind to provide an easy-to-use and
359 accessible platform for data analysis. This review aims to provide a brief introduction to some of the
360 powerful tools available online, or via download, but is by no means exhaustive. Comprehensive
361 lists of available software can also be found online, e.g. at the ExPASy website [90] (also, see Table
362 1).

363 **Handling mass spectrometry data:** Raw spectra and peptide identification data from MS can be
364 both large and complicated and it is therefore essential to establish an analysis pipeline for its
365 handling. In most instances the initial steps in this process will be handled by a local proteomics
366 service provider; however, as submission of raw proteomics data during publication is becoming
367 more widely required it is advisable to have some understanding of this. A useful and expansive list
368 of available software can be found online (see Table 1); however, some of the most popular tools
369 include Max Quant [91] and Mascot (Matrix Science). These packages handle the identification,
370 characterisation and quantification of peptides and are suitable for large mass-spectrometric data
371 sets. Software is used to match identified peptides to *in silico* digested proteomes for protein
372 identification and is also capable of quantitative analysis, i.e. label-free or stable isotope labelling in
373 tissue culture (SILAC) technologies. Another popular piece of software for visualising the results of
374 MS is Scaffold, from Proteome Software, which can provide information on protein function and sub-
375 cellular localisation and also allows easy comparison between biological samples.

376 **Data repositories and storage:** Proteomics data in its raw format can be stored (i.e. post-
377 publication) in the Proteomics Identifications (PRIDE) database [92]. This curated repository is
378 hosted by the European Bioinformatics Institute (EBI) and stores detailed information such as
379 peptide identifications, post-translational modifications (PTMs) as well as the associated spectra.
380 Interactomics data can also be correlated to gene expression studies, e.g.microarrays; the two
381 popular repositories for these studies are Gene Expression Omnibus (GEO) [93], part of the
382 National Center for Biotechnology (NCBI) and EBI's Array Express [94]. These databases can be
383 used to access published experimental data sets and both websites have useful online tools to
384 probe gene expression levels. This cross-comparison with a user's own data can provide quick,
385 cheap and interesting additional information on proteins of interest.

386 **What is protein X?** Some of the first questions about any identified protein are the most basic, e.g.
387 what is this protein and what does it do? Manually curated databases such as the Universal Protein

ACCEPTED MANUSCRIPT

388 Resource (Uniprot) [95], a collaboration between the EBI, Swiss Institute of Bioinformatics (SIB) and
389 the Protein Information Resource (PIR), are a vital tool at this point and can provide essential
390 information on isoform sequences, post translational modifications, molecular functions, taxonomy,
391 etc. Another useful website is the Weizmann Institute's GeneCards [96] which is very useful for
392 identifying known aliases for any given protein.

393 **Identifying protein interactions and additional network analysis:** Identifying the interaction
394 between a protein and its substrate, binding partner or small molecule is obviously an area of
395 intense interest in modern molecular biology. As such there are a range of tools, both freeware and
396 available under licence, to probe and map these interactions (established or predicted). Essentially
397 these tools allow the user to examine large interactome databases, providing a broad overview of
398 what may be known about an input and its biology. Two particularly popular online tools are IntAct
399 [97], another service from the EBI which catalogues molecular interaction data, and String [98], a
400 database of known and predicted protein interactions run by a cross-institute team including the
401 University of Copenhagen, European Molecular Biology Laboratory (EMBL) and SIB. These
402 databases include direct and indirect associations derived from a range of experimental and text-
403 based sources and are very useful for developing a broader understanding of the molecular
404 interactions of a protein. Also popular is the commercially available software Ingenuity Pathways
405 Analysis (IPA) which allows subsequent integration with gene ontology networks and biological
406 pathway analysis.

407 **Identifying broader biological context:** Interactomic data sets can be complicated by the number
408 of individual proteins identified, be it from the use of crude approaches or the very nature of the
409 baits themselves. In this instance it can be helpful to analyse the data at a more systematic level to
410 look for enrichment of biological pathways. Two popular tools for this are Gorilla [99], a gene
411 ontology enrichment analysis and visualisation tool, as well as the Reactome Pathway Database
412 [100] organised by Ontario Institute for Cancer Research (OICR), New York University Medical
413 Centre (NYUMC) and the EBI. Both tools are easily adapted to lists of proteins, rather than genes,
414 and are designed for pathway enrichment analysis, identifying statistically over-represented gene
415 members in input data. Reactome's online database provides an extensive library of curated
416 pathways that can be analysed at the molecular level, providing a unique insight into the potential
417 biological context of any identified proteins. Such tools are particularly useful when multi-partite
418 complexes of proteins are identified rather than single binding partners.

419 **Visualising and modelling protein interactions at the molecular level:** A comprehensive list of
420 available software can be found on the International Union of Crystallography's website (see Table
421 1); however, briefly, structural information is available for download via the Research Collaboratory
422 for Structural Bioinformatics' (RCSB) Protein Data Bank (PDB) [101]. This powerful curated

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

423 database provides access to structural data from a range of sources and is invaluable for accessing
424 published crystal structures and information on established interactions. Two popular packages for
425 visualising structural data are Chimera from the University of California San Francisco (UCSF) [102]
426 and Schrödinger's PyMOL. Both tools provide high-powered and interactive visualisation of 3D
427 molecular structures and include features such as imaging supramolecular assemblies and results
428 of molecular docking studies. Finally, there are also easily accessible tools for user-defined
429 homology-modelling such as Swiss-Model [103] from SIB, which automates the process allowing
430 easy access to a protein modelling pipeline.

431 **Final Comments**

432 The interactomics techniques and data analysis tools described above have all successfully been
433 applied in the laboratory setting to identify and characterise protein interactions. In most situations,
434 they are relatively easy to establish and common enough to be available commercially. However, it
435 is worth highlighting that many of these methodologies are now being adapted for high throughput
436 analysis, enabling global analysis of the protein-protein, DNA-protein and RNA-protein binding
437 landscape at an unprecedented scale. As with many emerging 'omics platforms and technologies,
438 e.g. microarrays and next-generation sequencing, the available interactomics data 'online' now
439 outstrips the analytical ability of most labs. As larger scale interactomics technologies come on
440 board (e.g. affordable and robust protein microarrays) this trend will doubtlessly continue. It is
441 therefore becoming more and more important to examine the available data online before
442 conducting any interactomics study, as many existing large-scale studies can easily be reinterpreted
443 without investing time in developing new systems or technical approaches.

444 **Acknowledgements:**

445 DB and CD are both independent investigators funded by the University of Birmingham Fellowship
446 scheme. DGC and TC are PhD students in these labs funded by the same scheme. CD is the
447 recipient of an MRC New Investigator Award (MR/M009912/1).

448 **References:**

449 1. McFedries, A., A. Schwaid, and A. Saghatelian, *Methods for the elucidation of protein-small molecule*
450 *interactions.* Chem Biol, 2013. **20**(5): p. 667-73.
451 2. Silvian, L., I. Enyedy, and G. Kumaravel, *Inhibitors of protein-protein interactions: new methodologies*
452 *to tackle this challenge.* Drug Discov Today Technol, 2013. **10**(4): p. e509-15.
453 3. Arkin, M.R., Y. Tang, and J.A. Wells, *Small-molecule inhibitors of protein-protein interactions:*
454 *progressing toward the reality.* Chem Biol, 2014. **21**(9): p. 1102-14.
455 4. Fields, S. and O. Song, *A novel genetic system to detect protein-protein interactions.* Nature, 1989.
456 **340**(6230): p. 245-6.
457 5. Brückner, A., et al., *Yeast Two-Hybrid, a Powerful Tool for Systems Biology.* International Journal of
458 Molecular Sciences, 2009. **10**(6): p. 2763-2788.

ACCEPTED MANUSCRIPT

459 6. Ehrhard, K.N., et al., *Use of G-protein fusions to monitor integral membrane protein-protein*
460 *interactions in yeast.* Nat Biotechnol, 2000. **18**(10): p. 1075-9.
461 7. Dirnberger, D., M. Messerschmid, and R. Baumeister, *An optimized split-ubiquitin cDNA-library*
462 *screening system to identify novel interactors of the human Frizzled 1 receptor.* Nucleic Acids Res,
463 2008. **36**(6): p. e37.
464 8. Johnsson, N. and A. Varshavsky, *Split ubiquitin as a sensor of protein interactions in vivo.* Proc Natl
465 Acad Sci U S A, 1994. **91**(22): p. 10340-4.
466 9. Urech, D.M., P. Lichtlen, and A. Barberis, *Cell growth selection system to detect extracellular and*
467 *transmembrane protein interactions.* Biochim Biophys Acta, 2003. **1622**(2): p. 117-27.
468 10. Serebriiskii, I., V. Khazak, and E.A. Golemis, *A two-hybrid dual bait system to discriminate specificity*
469 *of protein interactions.* J Biol Chem, 1999. **274**(24): p. 17080-7.
470 11. Koegl, M. and P. Uetz, *Improving yeast two-hybrid screening systems.* Briefings in Functional
471 Genomics & Proteomics, 2007. **6**(4): p. 302-312.
472 12. Iyer, K., et al., *Utilizing the split-ubiquitin membrane yeast two-hybrid system to identify protein-*
473 *protein interactions of integral membrane proteins.* Sci STKE, 2005. **2005**(275): p. pl3.
474 13. Paumi, C.M., et al., *Mapping protein-protein interactions for the yeast ABC transporter Ycf1p by*
475 *integrated split-ubiquitin membrane yeast two-hybrid analysis.* Mol Cell, 2007. **26**(1): p. 15-25.
476 14. Petschnigg, J., et al., *The mammalian-membrane two-hybrid assay (MaMTH) for probing membrane-*
477 *protein interactions in human cells.* Nat Methods, 2014. **11**(5): p. 585-92.
478 15. Kwon, J.A., et al., *High diagnostic accuracy of antigen microarray for sensitive detection of hepatitis C*
479 *virus infection.* Clin Chem, 2008. **54**(2): p. 424-8.
480 16. Zhu, H., et al., *Analysis of yeast protein kinases using protein chips.* Nat Genet, 2000. **26**(3): p. 283-9.
481 17. Zhu, H., et al., *Global Analysis of Protein Activities Using Proteome Chips.* Science, 2001. **293**(5537):
482 p. 2101-2105.
483 18. Uzoma, I. and H. Zhu, *Interactome mapping: using protein microarray technology to reconstruct*
484 *diverse protein networks.* Genomics, proteomics & bioinformatics, 2013. **11**(1): p. 18-28.
485 19. Zhu, J., et al., *RNA-binding proteins that inhibit RNA virus infection.* Proc Natl Acad Sci U S A, 2007.
486 **104**(9): p. 3129-34.
487 20. Berger, M.F. and M.L. Bulyk, *Universal protein binding microarrays for the comprehensive*
488 *characterization of the DNA binding specificities of transcription factors.* Nat Protoc, 2009. **4**(3): p.
489 393-411.
490 21. Ho, S.W., et al., *Linking DNA-binding proteins to their recognition sequences by using protein*
491 *microarrays.* Proc Natl Acad Sci U S A, 2006. **103**(26): p. 9940-5.
492 22. Yates, J.R., C.I. Ruse, and A. Nakorchevsky, *Proteomics by mass spectrometry: approaches, advances,*
493 *and applications.* Annu Rev Biomed Eng, 2009. **11**: p. 49-79.
494 23. Cargile, B.J., J.L. Bundy, and J.L. Stephenson, Jr., *Potential for false positive identifications from large*
495 *databases through tandem mass spectrometry.* J Proteome Res, 2004. **3**(5): p. 1082-5.
496 24. Kumar, C. and M. Mann, *Bioinformatics analysis of mass spectrometry-based proteomics data sets.*
497 FEBS Letters, 2009. **583**(11): p. 1703-1712.
498 25. Jager, S., et al., *Global landscape of HIV-human protein complexes.* Nature, 2012. **481**(7381): p. 365-
499 370.
500 26. Krogan, N.J., et al., *Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.*
501 Nature, 2006. **440**(7084): p. 637-43.
502 27. Rigaut, G., et al., *A generic protein purification method for protein complex characterization and*
503 *proteome exploration.* Nat Biotechnol, 1999. **17**(10): p. 1030-2.
504 28. Bailey, D., et al., *Identification of Protein Interacting Partners Using Tandem Affinity Purification.*
505 Journal of Visualized Experiments : JoVE, 2012(60): p. 3643.
506 29. Goodfellow, I. and D. Bailey, *Detection of protein-protein interactions using tandem affinity*
507 *purification.* Methods Mol Biol, 2014. **1177**: p. 121-33.
508 30. Puig, O., et al., *The tandem affinity purification (TAP) method: a general procedure of protein*
509 *complex purification.* Methods, 2001. **24**(3): p. 218-29.

ACCEPTED MANUSCRIPT

510 31. Gunzl, A. and B. Schimanski, *Tandem affinity purification of proteins.* Curr Protoc Protein Sci, 2009.
511 **Chapter 19**: p. Unit 19 19.
512 32. Chung, L., et al., *Norovirus Translation Requires an Interaction between the C Terminus of the*
513 *Genome-linked Viral Protein VPg and Eukaryotic Translation Initiation Factor 4G.* The Journal of
514 Biological Chemistry, 2014. **289**(31): p. 21738-21750.
515 33. Haura, E.B., et al., *Using iTRAQ® Combined with Tandem Affinity Purification to Enhance Low-*
516 *abundance Proteins Associated with Somatically-mutated EGFR Core Complexes in Lung Cancer.*
517 Journal of proteome research, 2011. **10**(1): p. 182-190.
518 34. Hubner, N.C., et al., *Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo*
519 *protein interactions.* J Cell Biol, 2010. **189**(4): p. 739-54.
520 35. Poser, I., et al., *BAC TransgeneOmics: a high-throughput method for exploration of protein function*
521 *in mammals.* Nature methods, 2008. **5**(5): p. 409-415.
522 36. Salic, A. and T.J. Mitchison, *A chemical method for fast and sensitive detection of DNA synthesis in*
523 *vivo.* Proc Natl Acad Sci U S A, 2008. **105**(7): p. 2415-20.
524 37. Sirbu, B.M., et al., *Analysis of protein dynamics at active, stalled, and collapsed replication forks.*
525 Genes Dev, 2011. **25**(12): p. 1320-7.
526 38. Sirbu, B.M., et al., *Identification of proteins at active, stalled, and collapsed replication forks using*
527 *isolation of proteins on nascent DNA (iPOND) coupled with mass spectrometry.* J Biol Chem, 2013.
528 **288**(44): p. 31458-67.
529 39. Gillespie, P.J., A. Gambus, and J.J. Blow, *Preparation and use of Xenopus egg extracts to study DNA*
530 *replication and chromatin associated proteins.* Methods, 2012. **57**(2): p. 203-13.
531 40. Cross, M.K. and M.A. Powers, *Learning about cancer from frogs: analysis of mitotic spindles in*
532 *Xenopus egg extracts.* Dis Model Mech, 2009. **2**(11-12): p. 541-7.
533 41. Masui, Y. and C.L. Markert, *Cytoplasmic control of nuclear behavior during meiotic maturation of frog*
534 *oocytes.* J Exp Zool, 1971. **177**(2): p. 129-45.
535 42. Ma, L., et al., *Identification of small ubiquitin-like modifier substrates with diverse functions using the*
536 *Xenopus egg extract system.* Mol Cell Proteomics, 2014. **13**(7): p. 1659-75.
537 43. Leach, C.A. and W.M. Michael, *Ubiquitin/SUMO modification of PCNA promotes replication fork*
538 *progression in Xenopus laevis egg extracts.* J Cell Biol, 2005. **171**(6): p. 947-54.
539 44. Moreno, S.P., et al., *Polyubiquitylation drives replisome disassembly at the termination of DNA*
540 *replication.* Science, 2014. **346**(6208): p. 477-81.
541 45. Murat, P. and S. Balasubramanian, *Existence and consequences of G-quadruplex structures in DNA.*
542 Curr Opin Genet Dev, 2014. **25**: p. 22-9.
543 46. Huppert, J.L., *Structure, location and interactions of G-quadruplexes.* FEBS J, 2010. **277**(17): p. 3452-
544 8.
545 47. von Hacht, A., et al., *Identification and characterization of RNA guanine-quadruplex binding proteins.*
546 Nucleic Acids Res, 2014. **42**(10): p. 6630-44.
547 48. Yang, G., X. Lu, and L. Yuan, *LncRNA: a link between RNA and cancer.* Biochimica et biophysica acta,
548 2014. **1839**(11): p. 1097-109.
549 49. Keene, J.D., J.M. Komisarow, and M.B. Friedersdorf, *RIP-Chip: the isolation and identification of*
550 *mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts.*
551 Nature protocols, 2006. **1**(1): p. 302-7.
552 50. Mili, S. and J.A. Steitz, *Evidence for reassociation of RNA-binding proteins after cell lysis: implications*
553 *for the interpretation of immunoprecipitation analyses.* RNA, 2004. **10**(11): p. 1692-4.
554 51. Ule, J., et al., *CLIP identifies Nova-regulated RNA networks in the brain.* Science, 2003. **302**(5648): p.
555 1212-5.
556 52. Hafner, M., et al., *Transcriptome-wide identification of RNA-binding protein and microRNA target*
557 *sites by PAR-CLIP.* Cell, 2010. **141**(1): p. 129-41.
558 53. Andersen, C.B.F., et al., *Structure of the exon junction core complex with a trapped DEAD-box ATPase*
559 *bound to RNA.* Science, 2006. **313**(5795): p. 1968-1972.
560 54. Singh, G., E.P. Ricci, and M.J. Moore, *RIPiT-Seq: a high-throughput approach for footprinting*
561 *RNA:protein complexes.* Methods, 2014. **65**(3): p. 320-32.

ACCEPTED MANUSCRIPT

562 55. Beach, D.L. and J.D. Keene, *Ribotrap : targeted purification of RNA-specific RNPs from cell lysates*
563 *through immunoaffinity precipitation to identify regulatory proteins and RNAs.* Methods in molecular
564 biology, 2008. **419**: p. 69-91.
565 56. Yoon, J.H., S. Srikantan, and M. Gorospe, *MS2-TRAP (MS2-tagged RNA affinity purification): tagging*
566 *RNA to identify associated miRNAs.* Methods, 2012. **58**(2): p. 81-7.
567 57. Gumireddy, K., et al., *Identification of a long non-coding RNA-associated RNP complex regulating*
568 *metastasis at the translational step.* Embo Journal, 2013. **32**(20): p. 2672-2684.
569 58. Lee, H.Y., et al., *RNA-protein analysis using a conditional CRISPR nuclease.* Proceedings of the
570 National Academy of Sciences of the United States of America, 2013. **110**(14): p. 5416-21.
571 59. Mann, M., *Functional and quantitative proteomics using SILAC.* Nat Rev Mol Cell Biol, 2006. **7**(12): p.
572 952-8.
573 60. Ong, S.E., et al., *Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate*
574 *approach to expression proteomics.* Mol Cell Proteomics, 2002. **1**(5): p. 376-86.
575 61. Ippel, J.H., et al., *In vivo uniform (15)N-isotope labelling of plants: using the greenhouse for structural*
576 *proteomics.* Proteomics, 2004. **4**(1): p. 226-34.
577 62. Washburn, M.P., et al., *Analysis of Quantitative Proteomic Data Generated via Multidimensional*
578 *Protein Identification Technology.* Analytical Chemistry, 2002. **74**(7): p. 1650-1657.
579 63. Hebert, A.S., et al., *Neutron-encoded mass signatures for multiplexed proteome quantification.* Nat
580 Methods, 2013. **10**(4): p. 332-4.
581 64. Kruger, M., et al., *SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor*
582 *for red blood cell function.* Cell, 2008. **134**(2): p. 353-64.
583 65. Larance, M., et al., *Stable Isotope Labeling with Amino acids in Nematodes.* Nature methods, 2011.
584 **8**(10): p. 849-851.
585 66. Munday, D.C., et al., *Using SILAC and quantitative proteomics to investigate the interactions between*
586 *viral and host proteomes.* Proteomics, 2012. **12**(4-5): p. 666-72.
587 67. Barrero, C.A., et al., *HIV-1 Vpr Modulates Macrophage Metabolic Pathways: A SILAC-Based*
588 *Quantitative Analysis.* PLoS ONE, 2013. **8**(7): p. e68376.
589 68. Geiger, T., et al., *Super-SILAC mix for quantitative proteomics of human tumor tissue.* Nat Methods,
590 2010. **7**(5): p. 383-5.
591 69. Zanivan, S., et al., *In Vivo SILAC-Based Proteomics Reveals Phosphoproteome Changes during Mouse*
592 *Skin Carcinogenesis.* Cell Reports. **3**(2): p. 552-566.
593 70. Rose, C.M., et al., *Neutron encoded labeling for peptide identification.* Anal Chem, 2013. **85**(10): p.
594 5129-37.
595 71. Boersema, P.J., et al., *Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics.*
596 Nat Protoc, 2009. **4**(4): p. 484-94.
597 72. Lau, H.T., et al., *Comparing SILAC- and stable isotope dimethyl-labeling approaches for quantitative*
598 *proteomics.* J Proteome Res, 2014. **13**(9): p. 4164-74.
599 73. Thompson, A., et al., *Tandem mass tags: a novel quantification strategy for comparative analysis of*
600 *complex protein mixtures by MS/MS.* Anal Chem, 2003. **75**(8): p. 1895-904.
601 74. Ross, P.L., et al., *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive*
602 *isobaric tagging reagents.* Mol Cell Proteomics, 2004. **3**(12): p. 1154-69.
603 75. Foster, L.J., C.L. De Hoog, and M. Mann, *Unbiased quantitative proteomics of lipid rafts reveals high*
604 *specificity for signaling factors.* Proc Natl Acad Sci U S A, 2003. **100**(10): p. 5813-8.
605 76. Andersen, J.S., et al., *Proteomic characterization of the human centrosome by protein correlation*
606 *profiling.* Nature, 2003. **426**(6966): p. 570-4.
607 77. Foster, L.J., et al., *A mammalian organelle map by protein correlation profiling.* Cell, 2006. **125**(1): p.
608 187-99.
609 78. Huber, L.A., K. Pfaller, and I. Vietor, *Organelle proteomics: implications for subcellular fractionation*
610 *in proteomics.* Circ Res, 2003. **92**(9): p. 962-8.
611 79. Tinnikov, A.A. and H.H. Samuels, *A novel cell lysis approach reveals that caspase-2 rapidly*
612 *translocates from the nucleus to the cytoplasm in response to apoptotic stimuli.* PLoS One, 2013.
613 **8**(4): p. e61085.

ACCEPTED MANUSCRIPT

614 80. Dunkley, T.P., et al., *Localization of organelle proteins by isotope tagging (LOPIT).* Mol Cell
615 Proteomics, 2004. **3**(11): p. 1128-34.
616 81. Dunkley, T.P., et al., *Mapping the Arabidopsis organelle proteome.* Proc Natl Acad Sci U S A, 2006.
617 **103**(17): p. 6518-23.
618 82. Christoforou, A., A.M. Arias, and K.S. Lilley, *Determining protein subcellular localization in*
619 *mammalian cell culture with biochemical fractionation and iTRAQ 8-plex quantification.* Methods
620 Mol Biol, 2014. **1156**: p. 157-74.
621 83. Rhee, H.-W., et al., *Proteomic Mapping of Mitochondria in Living Cells via Spatially-Restricted*
622 *Enzymatic Tagging.* Science (New York, N.Y.), 2013. **339**(6125): p. 1328-1331.
623 84. Roux, K.J., et al., *A promiscuous biotin ligase fusion protein identifies proximal and interacting*
624 *proteins in mammalian cells.* J Cell Biol, 2012. **196**(6): p. 801-10.
625 85. Chen, A.L., et al., *Novel components of the Toxoplasma inner membrane complex revealed by BioID.*
626 MBio, 2015. **6**(1): p. e02357-14.
627 86. Coyaud, E., et al., *BioID-based identification of SCF beta-TrCP1/2 E3 ligase substrates.* Mol Cell
628 Proteomics, 2015.
629 87. Dingar, D., et al., *BioID identifies novel c-MYC interacting partners in cultured cells and xenograft*
630 *tumors.* J Proteomics, 2015. **118**: p. 95-111.
631 88. Schmidt, C., et al., *Comparative cross-linking and mass spectrometry of an intact F-type ATPase*
632 *suggest a role for phosphorylation.* Nat Commun, 2013. **4**.
633 89. Liu, F., et al., *An enhanced protein crosslink identification strategy using CID-cleavable chemical*
634 *crosslinkers and LC/MS(n) analysis.* Proteomics, 2012. **12**(3): p. 401-5.
635 90. Artimo, P., et al., *ExPASy: SIB bioinformatics resource portal.* Nucleic Acids Res, 2012. **40**(Web Server
636 issue): p. W597-603.
637 91. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range*
638 *mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.
639 92. Vizcaino, J.A., et al., *The PRoteomics IDEntifications (PRIDE) database and associated tools: status in*
640 *2013.* Nucleic Acids Res, 2013. **41**(Database issue): p. D1063-9.
641 93. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets--update.* Nucleic Acids Res,
642 2013. **41**(Database issue): p. D991-5.
643 94. Kolesnikov, N., et al., *ArrayExpress update--simplifying data submissions.* Nucleic Acids Res, 2015.
644 **43**(Database issue): p. D1113-6.
645 95. UniProt, C., *UniProt: a hub for protein information.* Nucleic Acids Res, 2015. **43**(Database issue): p.
646 D204-12.
647 96. Safran, M., et al., *GeneCards Version 3: the human gene integrator.* Database (Oxford), 2010. **2010**:
648 p. baq020.
649 97. Orchard, S., et al., *The MIntAct project--IntAct as a common curation platform for 11 molecular*
650 *interaction databases.* Nucleic Acids Res, 2014. **42**(Database issue): p. D358-63.
651 98. Szklarczyk, D., et al., *STRING v10: protein-protein interaction networks, integrated over the tree of*
652 *life.* Nucleic Acids Res, 2015. **43**(Database issue): p. D447-52.
653 99. Eden, E., et al., *GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene*
654 *lists.* BMC Bioinformatics, 2009. **10**: p. 48.
655 100. Croft, D., et al., *Reactome: a database of reactions, pathways and biological processes.* Nucleic Acids
656 Res, 2011. **39**(Database issue): p. D691-7.
657 101. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
658 102. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J
659 Comput Chem, 2004. **25**(13): p. 1605-12.
660 103. Arnold, K., et al., *The SWISS-MODEL workspace: a web-based environment for protein structure*
661 *homology modelling.* Bioinformatics, 2006. **22**(2): p. 195-201.

662

663 **Figure Legends:**

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

664 **Figure 1: Affinity purification of protein complexes.** Experimental design for interactomics
665 experiments can be improved through the application of novel techniques such as protein
666 correlation profiling and endogenous biotinylation; methods that can reduce the complexity of
667 samples to be analysed by mass spectrometry. Various techniques for affinity purification of protein
668 complexes including tandem affinity purification and RNA-protein immunoprecipitation in tandem
669 can then be applied to identify protein complexes, although this is dependent on the type of bait
670 used (protein, DNA or RNA). Finally, quantification of proteins can also be applied to reduce
671 complexity, increase stringency and also to provide statistical confidence that identified proteins are
672 bait- specific.

673 **Figure 2: Schematic representation of spatially-restricted endogenous enzymatic tagging.**
674 The ascorbate peroxidase (APEX) protein can be retargeted to different sub-cellular locations in the
675 cell by the addition of organelle specific signal peptides, e.g. mitochondrial localisation signals. Re-
676 engineered APEX expression and localisation can then be determined by microscopy. Enzymatic
677 tagging proceeds after addition of biotin-phenol and H_2O_2 to cells for a short period (1 minute).
678 During this incubation the enzymatic activity of APEX produces short-lived biotin-phenolic radicals
679 that react only with proteins in close proximity, tagging them with biotin. After cell lysis, biotinylated
680 protein complexes can be purified using high stringency streptavidin beads and subsequently
681 analysed by mass spectrometry.

682

683

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

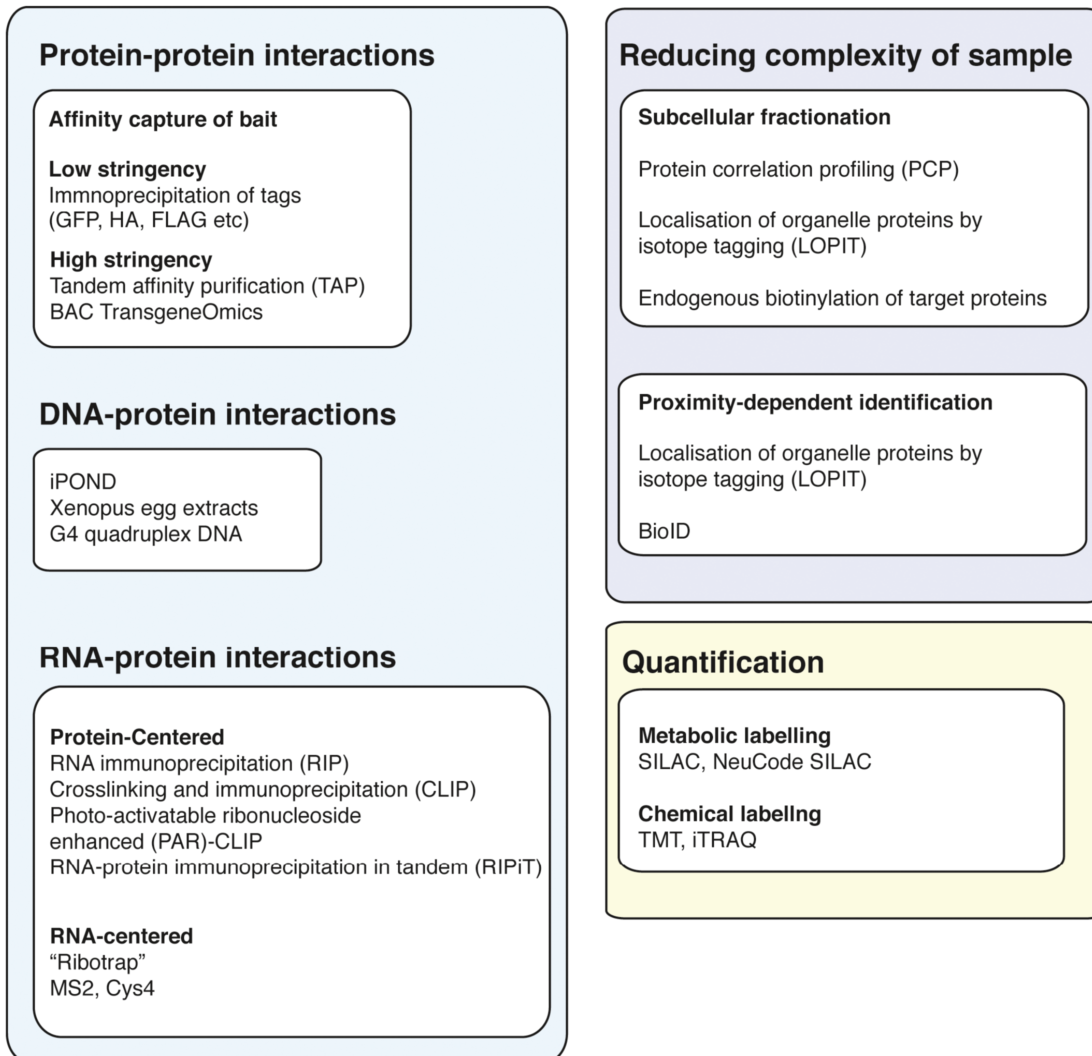
684 **Table 1:** Online resources and software for interactomics researchers.

Resource	Description	URL	Reference
Swiss Institute of Bioinformatics Expert Protein Analysis System	Software and program access/list	http://www.expasy.org/	Artimo *et al.*, 2012
Molecular biology tools	List of online (often freeware) analysis tools	http://molbiol-tools.ca/	n/a software and program list, Dr. Andrew Kropinski
PRIDE Proteomics Identifications	Proteomics data repository	http://www.ebi.ac.uk/pride/archive/	Vizcaino *et al.*, 2013
National Center for Biotechnology Gene Expression Omnibus (GEO)	Gene expression experimental data repository	http://www.ncbi.nlm.nih.gov/geo/	Barrett *et al.*, 2013
European Bioinformatics Institute Array Express	Archive of functional genomics data	http://www.ebi.ac.uk/arrayexpress/	Kolesnikov *et al.*, 2015
IntAct	Protein interactions database	http://www.ebi.ac.uk/intact/	Orchard *et al.*, 2014
String	Protein interactions database	http://string-db.org/	Szklarczyk *et al.*, 2015
Ingenuity Pathways Analysis (IPA)	Commercially available biological pathway analysis software	http://www.ingenuity.com/products/ipa	n/a, commercial software
Mass spectrometry utilities	Software and program list for handling mass spectrometry data	http://www.ms-utils.org/wiki/pmwiki.php/Main/SoftwareList	n/a, software and program list
Matrix Science Mascot	Software to process mass spectrometry data	http://www.matrixscience.com/index.html	n/a, commercial software
Max Quant	Software to process mass spectrometry data	http://141.61.102.17/maxquant_doku/doku.php?id=start	Cox & Mann, 2008
Scaffold	Software to handle mass spectrometry data	http://www.proteomesoftware.com/products/scaffold/	n/a, commercial software
The Universal Protein Resource (Uniprot)	Curated database of protein information	http://www.uniprot.org/	Uniprot Consortium, 2015
Weizmann Institute's GeneCards	Database of genetic information	http://www.genecards.org/	Safran *et al.*, 2010
Reactome	Online tool for analysing biological pathways	http://www.reactome.org/	Croft *et al.*, 2011
Gorilla: Gene Ontology enrichment analysis and visualization tool	Online tool for identifying enriched pathways in data sets	http://cbl-gorilla.cs.technion.ac.il/	Eden *et al.*, 2009
International Union of Crystallography Resources Page	Extensive list of structural software	http://www.iucr.org/resources/other-directories/software	n/a, software and program list
RCSB PDB database	Database of protein structures	http://www.rcsb.org/pdb/home/home.do	Berman *et al.*, 2000
University of California San Francisco's Chimera	Software for analysing structural data	http://www.cgl.ucsf.edu/chimera/	Pettersen et al., 2004
PyMOL	Software for analysing structural data	https://www.pymol.org/	n/a, commercial software
SIB Swiss-Model	Online tool for protein modelling	http://swissmodel.expasy.org/	Arnold *et al.*, 2006

685

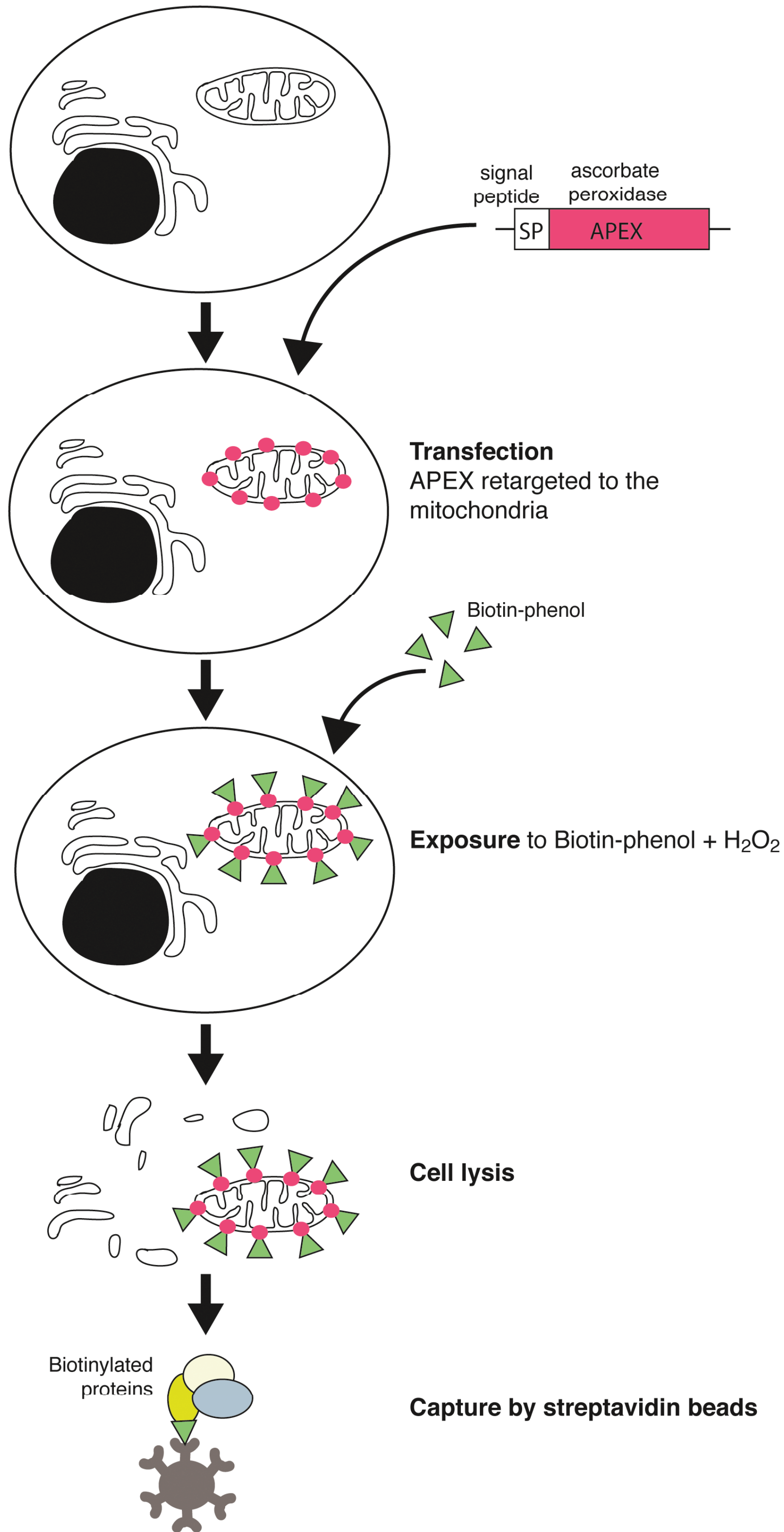
686

ACCEPTED MANUSCRIPT



687

Figure 1: Gonçalves Carneiro et al.



689

Figure 2: Gonçalves Carneiro et al.

ACCEPTED MANUSCRIPT

690

691 Highlights:

692 1. **Introducing interactomics techniques:** We have provided an introductory review of a broad range of
693 methodologies used to identify protein interactions, particularly emerging technologies in the fields
694 of protein-protein, DNA-protein and RNA-protein interactomics.
695 2. **Improving experimental set-up:** We describe some of the powerful approaches researchers can
696 apply to interactomics to improve stringency and specificity in their experiments. These include
697 quantitative labelling, proximity labelling and sub-cellular fractionation.
698 3. **Downstream processing:** There are numerous online tools and freeware for interrogating and
699 storing interactomics data. We have provided an overview of these together with links and
700 references.

701

ACCEPTED MANUSCRIPT