Original citation:

Permanent WRAP URL:
http://wrap.warwick.ac.uk/80160

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work of researchers of the University of Warwick available open access under the following conditions.

This article is made available under the Creative Commons Attribution 3.0 International license (CC BY 3.0) and may be reused according to the conditions of the license. For more details see: http://creativecommons.org/licenses/by/3.0/

A note on versions:
The version presented in WRAP is the published version, or, version of record, and may be cited as it appears here.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk
A LIR appears as a hallmark of the Atg8-interacting proteins. Previously described as the \textbf{xLIR} motif (where \textit{x} can be any amino acid), we and others recently extended this sequence to 6 amino acids based on the multiple alignment of LIR sequences from proteins described to interact in a LIR-dependent manner with Atg8-proteins.\cite{10,16,17} Based on the in silico analysis of experimentally verified functional LIR motifs, we redefined the sequence of the LIR motif. The resulting consensus sequence—referred to as the \textbf{xLIR} motif—is (ADEFGPLRSK)(DEGMSTV) (\textit{WFY})(DEILQTV)(ADEFHIKLMPSTV)(ILV), where the residues marked in bold (positions 3 and 6) correspond to the most crucial residues for the interaction with Atg8-family proteins. An \textbf{xLIR} motif overlapping a region with the potential to transit from a disordered to an ordered state provides a reliable candidate for a functional binding motif.\cite{10,17,18}

In addition to selective autophagy receptors, Atg8-family proteins can bind a variety of proteins in an LIR-dependent manner. Indeed, many LIR motif-containing proteins (LIRCPs) are required for the formation of the autophagosome or vesicular transport or they are proteins that are directly targeted to the phagophore for autophagic clearance.\cite{16,19-22,23,24} It is worth mentioning that LIR motif-independent modes of interaction with Atg8-family proteins have also been reported.

**CONTACT** Ioannis P. Nezis Ioannis.P.Nezis@warwick.ac.uk
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/kaup.

**Supplemental data for this article can be accessed on the publisher’s website.**
© 2016 Anne-Claire Jacomin, Siva Samavedam, Vasilis Promponas, and Ioannis P. Nezis. Published with license by Taylor & Francis.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The moral rights of the named author(s) have been asserted.
both in selective autophagy receptors and in other autophagy-related proteins.²⁸

In this report, we describe the use of the iLIR server¹⁷ combined with a Gene Ontology (GO) term analysis to sort the genes from 8 model organisms (Arabidopsis thaliana, Caenorhabditis elegans, Danio rerio, Gallus gallus, Homo sapiens, Mus musculus, Rattus norvegicus and Saccharomyces cerevisiae) encoding proteins containing at least one xLIR motif inside an intrinsically disordered region. The data have been collected in the iLIR database (https://ilir.warwick.ac.uk), with the aim to provide a useful resource to researchers interested in studying the Atg8-family proteins interactome. Additionally, a curated text-mining analysis of the literature permitted us to sort human and mouse proteins known to be a part of the Atg8-family proteins interactome or to be involved in pathways linked to autophagy, and also to identify novel putative LIRCPs that have not been associated with autophagy previously.

Results and discussion

Content of the iLIR database

The iLIR database is a web resource freely available at https://ilir.warwick.ac.uk. The website has been designed to give the user an easy way to browse available data and perform BLAST-based searches using a protein sequence of interest against part or all the sequences available in the database for proteins containing a similar xLIR motif. The website also provides hyperlinks to the UniProt database for each entry and the possibility to download the data.

Within the iLIR database different functionalities are organized under specific menus. The 'LIRCPs' menu gives access to the full list of putative LIRCPs listed in the database for the different model organisms analyzed. For a specific organism, data are presented in a table containing the following information for each entry: (i) the UniProtKB accession of the protein, (ii) the position, sequence and position-specific scoring matrix score of the xLIR,¹⁷ (iii) similar LIR motif in experimentally characterized LIRCPs (if any), (iv) the name of the protein, and (v) the UniProt derived GO terms associated with the molecular function, biological process and cellular component classes. The full table of data can be downloaded as an Excel file (Fig. 1).

The 'Search' menu offers the user to screen their sequence of interest for the presence of LIR (xLIR and WxxL) motifs using the iLIR server as described elsewhere.¹⁷ In addition, the user has the possibility to search in the database using specific keywords: gene name, protein description or UniProt identifier. The user may also look directly for the presence of similar proteins with the 'BLAST' page using PSI-BLAST.²⁹ The search can be run against Swiss-Prot and TrEMBL entries from the UniProt database (a total of 276,499 FASTA sequences). The results page shows pattern positions in the query sequence and the corresponding matching positions in the subject sequences from the database along with the alignments between them. Red asterisks match the position of the conserved xLIR motif in the subject sequences. Subject sequences matched are named by their UniProtKB accession number and a link permits the redirection to the UniProtKB page for each entry (Fig. 2).

Finally, the 'GO Annotation' menu provides pre-computed information relative to the GO terms distribution for the LIRCPs identified for each organism. Three types of analyses are available: (i) The 'GO Slim' submenu directs users to a list of reduced GO terms and their abundance for each category in

Figure 1. Screenshot of an iLIR database data page. In the ‘LIRCPs’ menu, the user can access the full data available in the database for each model organism. The data are arranged in a table giving various information for each entry, such as the Uniprot Accession ID and protein name, the position and sequence of the xLIR as well as the position-specific scoring matrix (PSSM) score and the similarity of other validated LIR motifs. The data can be downloaded directly.
a specific organism. The user can sort the entries based on their counts or adjusted p-value. (ii) The ‘Distribution’ submenu directs users to a bar chart view of the GO terms distribution for each organism. (ii) The ‘Enrichment’ submenu permits the visualization of the proportion of entries for each GO term for the LIRCPs for any pair of species available in the iLIR database (Fig. S1).

Prediction of the LIR-containing proteins (LIRCP) in the proteome of model organisms

Using iLIR, a computational approach for predicting LC3-interaction regions in proteins, we identified putative LIRCPs from 8 model organisms (see Methods for details). We found that the proportion of putative LIRCPs varies between 4% to
5% of the total ORFs for each organism but we observed no correlation between the proportion of LIRCPs and the size of the proteome (number of ORFs) (Fig. 3 and Table 1).

**Text-mining analysis for the identification of novel LIRCPs in mammals**

In order to further investigate novel putative LIRCPs in mammals, we first concentrated on the human and mouse proteomes. Our batch analysis lead to the identification of 6087 and 4218 entries, respectively. Consecutively to the application of the statistical significance for each GO slim category for these organisms, we decided to eliminate the entries sorted as ‘non significant’ (adjusted p-value > 0.1) from the rest of the analysis. This procedure permitted us to sort a total of 1766 and 1976 entries for the human and mouse proteome, respectively, with a low to high significance level (p-adj ≤ 0.1). We made use of these significant hits for further analysis.

Previous studies have identified and described 31 proteins encoded by the human, yeast and *Arabidopsis thaliana* genomes involved in autophagy through their interaction with at least one protein belonging to the Atg8-family and containing a functional, verified LIR motif. However, the LIR motifs of a few of these proteins are not contained within an intrinsically disordered region such as human ATG4B or yeast Atg3 and Atg19. From the 31 verified LIRCPs, all 21 proteins with a LIR motif within an anchor region have been successfully identified in our computational analysis, thus validating the sorting procedure (these proteins constitute the group ‘A’ in the rest of the text) (Table S1).

From these proteins, we extracted their associated GO slim categories for the 3 GO classes (Molecular Function, Biological Process and Cellular Component). Totally, 26 different GO terms were obtained (6 for the Molecular Function class, 8 for the Biological Process class and 12 for the Cellular Component class) (Fig. 4 and Table S1). We noticed that only 4 of these proteins have been assigned to the GO term ‘GO:0006914’–autophagy’ as a Biological Process; other proteins have been assigned to GO terms that can be related to autophagy such as GO:0005739–mitochondrion, GO:0030904–retromer complex (Cellular Component), GO:0006810–transport (Biological Process), GO:0005515–protein binding and GO:0042277–peptide binding (Molecular Function). Additionally, various GO terms not directly related to autophagy have been pinpointed such as GO:0005634–nucleus, GO:0005576–extracellular region, GO:0009986–cell surface, GO:0006457–protein folding, GO:0007049–cell cycle, GO:0004871–signal transducer activity or GO:0042562–hormone binding. This suggests that many proteins whose original function is not related to autophagy might interact with Atg8-family proteins in a way that remains unknown. In order to test this assumption, we decided to screen all the putative LIRCPs with a significant adjusted p-value (sorted as previously described) for the human and mouse proteomes, which are associated with at least one of the 26 GO terms correlated with the 21 experimentally validated human and yeast LIRCPs. Over 1,000 entries have thus been filtered. A manually curated search of these entries using PubMed, permitted us to sort 18 proteins already described to interact with an Atg8-family protein, irrespective of further evidence of a direct interaction (referred to hereafter as group ‘B’, Table S2). Three of these proteins—GPSM1/AGS3, NCOA4 and MAPK8IP1/JIP1—had been shown to interact directly (i.e., through in vitro studies) with some members of the Atg8-family. The 15 remaining proteins—PICALM, PCM1, STAT1, STAT3, UBQLN1 and UBQLN2, PEG3, HTT, SYNP2, UBRA4, MAP1S, BCL10, ODF1, FNIP2, APC, and CSPG4—have been identified to function in complexes containing Atg8-family proteins in *cellulo* by co-immunoprecipitation and/or colocalization experiments (Table S2). In line with the functions of the LIRCPs containing experimentally verified LIR motifs (Table S1), it appears that

<table>
<thead>
<tr>
<th># ORF</th>
<th># LIRCPs</th>
<th>LIR motifs</th>
<th>% LIRCPs</th>
<th>Ratio #LIRs:LIRCP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Homo sapiens</td>
<td>88479</td>
<td>5204</td>
<td>6087</td>
</tr>
<tr>
<td>Mouse</td>
<td>Mus musculus</td>
<td>51130</td>
<td>3599</td>
<td>4218</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>Danio rerio</td>
<td>41102</td>
<td>2571</td>
<td>3007</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>Arabidopsis thaliana</td>
<td>33350</td>
<td>1458</td>
<td>1592</td>
</tr>
<tr>
<td>Rat</td>
<td>Rattus norvegicus</td>
<td>28649</td>
<td>2068</td>
<td>2425</td>
</tr>
<tr>
<td>Worm</td>
<td>Caenorhabditis elegans</td>
<td>26595</td>
<td>1290</td>
<td>1526</td>
</tr>
<tr>
<td>Chicken</td>
<td>Gallus gallus</td>
<td>17864</td>
<td>1215</td>
<td>1400</td>
</tr>
<tr>
<td>Yeast</td>
<td>Saccharomyces cerevisiae</td>
<td>6693</td>
<td>332</td>
<td>363</td>
</tr>
</tbody>
</table>
Figure 4. Distribution of the GO terms of the 21 human proteins listed in Kalvari et al. which have a verified xLIR in an intrinsically disordered region (see also Table S1). MF, Molecular Function; BP, Biological Process; CC, Cellular Component.

Figure 5. Distribution of the GO terms of the 756 human entries that have not been linked to autophagy-associated processes (see also Table S4). MF, Molecular Function; BP, Biological Process; CC, Cellular Component.
the proteins interacting with Atg8-family members we sorted can be related to the autophagy process in various ways. Some of these Atg8-interacting proteins are selective autophagy receptors for the targeting of specific cargos (e.g., NCOA4, PICALM, PCM1, STAT1), whereas others are degraded themselves by autophagy (e.g., BCL10, ODF1). Yet some others are implicated in the regulation of the autophagic process (e.g., GPSM1/AG33, MAPK8IP1/JIP1, UBQ1L, PEG3, HTT, SYNP02, UBR4, MAP1S, FNIP2) (Table S2)32,34,39-47,49

In addition, our text-mining analysis permitted us to sort 256 supplementary entries corresponding to proteins that have been demonstrated to be involved in the regulation of autophagy, the degradation of specific substrates, or to be themselves degraded by autophagy without any evidence of interaction with Atg8-family proteins (referred to hereafter as group ‘C’, Table S3). These proteins have been described to take part in a broad range of processes related to autophagy, such as immunity (NFKBIA/IκBα), protein component class, GO:0005515, further investigation. suggestion that these proteins are promising candidates for (4.76%) for the Biological Process class (Fig. 5). This observa-
j shows that our database includes already experimentally vali-

The GO slim category of the model organisms using a hypergeometric test, employed through the Perl module Math::Pari (Search.cpan.org, (2014). Math-Pari-2.010808 Retrieved from: http://search.cpan.org/CPAN/authors/id/I/IL/ILYAZ/modules/Math-Pari-2.010808.zip.) based on the following criteria:

- Number of proteins assigned to a particular GO slim category in the model organism (n)
- Number of putative LIRCPs assigned to the same GO slim category in the model organism (x)
- Total number of proteins in the model organism (N)
- Total number of putative LIRCPs in the model organism (k)

Of course, there are some limitations to the iLIR database. At the moment, the iLIR server is not able to predict the nonca-

Methods

Proteomes of model organisms and prediction of the LIR-containing proteins (LIRCPs)

We selected 8 model organisms: Arabidopsis thaliana, Caenorhabditis elegans, Danio rerio, Gallus gallus, Homo sapiens, Mus musculus, Rattus norvegicus and Saccharomyces cerevisiae. The protein sequences encoding the complete genomes of these model organisms were obtained from the UniProt database (Uniprot.org, (2014). UniProt. [online] Available at: http://www.uniprot.org/ [Accessed 06 February 2014]). A stand-alone version of iLIR was employed to process the data in batch mode and predict LIRCPs based on the presence of at least one xLIR within an intrinsically disordered region.

Gene Ontology (GO) enrichment analysis

The GO enrichment analysis was performed by downloading the ID (identifiers) mapping data for each organism from UniProt. These data contains cross-references for a given UniProt identifier with mappings to multiple databases such as EntrezGene, RefSeq, GI, PDB, GO, PIR, NCBI-taxon, UniGene etc. each recorded as an identifier of the respective database. We also downloaded the Gene Ontology Protein Information Resource slim generic categories from the online GO database (Geneontology.org, (2014). GO Database. [online] Available at: http://www.geneontology.org/ontology/subsets/goslim_generic.obo [Accessed 19 June 2014].)

Using the mapping, GO slim and UniProt files together with the list of LIRCPs for a model organism, we generated GO class distribution files with counts of proteins having a particular GO slim category. One distribution file for each GO top level hierarchy (i.e., Biological Process, Cellular Component and Molecular Function) has been generated.

We assessed the statistical significance of each GO slim category of the model organisms using a hypergeometric test. We assessed the statistical significance of each GO slim category of the model organisms using a hypergeometric test, employed through the Perl module Math::Pari (Search.cpan.org, (2014). Math-Pari-2.010808 Retrieved from: http://search.cpan.org/CPAN/authors/id/I/IL/ILYAZ/modules/Math-Pari-2.010808.zip.) based on the following criteria:

- Number of proteins assigned to a particular GO slim category in the model organism (n)
- Number of putative LIRCPs assigned to the same GO slim category in the model organism (x)
- Total number of proteins in the model organism (N)
- Total number of putative LIRCPs in the model organism (k)
The formula used for predicting the probability using hyper-geometric test \((h)\) is given below: \[ h(x; N, n, k) = \frac{[kC_n][N-kC_{n-x}]}{[NC_n]} \]

To control the false discovery rate, we have also generated p-adjusted values employing the Benjamini-Hochberg method from Perl’s Statistics::Multtest module (Search.cpan.org, (2014). Statistics-Multtest-0.13. Retrieved from: http://search.cpan.org/CPAN/authors/id/J/JO/JOKERGOO/Statistics-Multtest-0.13.tar.gz). Following the hypergeometric test and false discovery rate correction, the GO distribution files were updated with p-value and p-adjusted values. Then, the GO slim categories data of model organisms was further classified based on different cut-offs for p-adjusted (p-adj) values as:

(i) Highly significant, \(p\text{-adj } < 0.01\)
(ii) Significant \((p\text{-adj } > 0.01 \text{ and } \leq 0.05)\)
(iii) Low significance \((p\text{-adj } > 0.05 \text{ and } p\text{-adj } < 0.1)\)
(iv) Not significant \((p\text{-adj } > 0.1)\)

**Web application**

The iLIR database has been developed for making the list of putative LIRCPs from the complete proteome of selected model organisms available for researchers worldwide. This web resource is based on well-established web technologies, including HTML, CSS, JavaScript, PHP (v5.3.28), JpGraph (v3.5.0b1) (Jpgraph.net, (2014). v3.5.0b1 Retrieved from: http://jpgraph.net/download/download.php?p=5) and the Apache web server technologies to develop and serve the web application.

**Abbreviations**

AIM | Atg8-interacting motif  
GO | gene ontology  
LIR | LC3-interacting region  
LIRCP | LIR-containing protein  
LRS | LC3 recognition sequence  
MAP1LC3/LC3 | microtubule associated protein 1 light chain 3  
xLIR | extended LIR motif

**Disclosure of potential conflicts of interest**

No potential conflicts of interest were disclosed.

**Funding**

This work is supported by BBSRC (grant BB/L006324/1 to I.P.N.).

**References**

A.-C. JACOMIN ET AL.


[26] Xie R, Nguyen S, McKeehan WL, Liu L. Microtubule-associated protein 1S (MAP1S) bridges autophagic components with microtubules and mitochondria to affect autophagic and PI3P to mediate microtubule plus end-directed vesicle transport. J Cell Biol 2010; 188:253-69; PMID:20100911; http://dx.doi.org/10.1038/jcb.200907015


[29] Pankiv S, Alemu EA, Brech A, Bruun JA, Lamark T, Overvatn A, Bjorkoy G, Johansen T. FYCO1 is a Rab7 effector that binds to LC3, which is a Rab7 effector that binds to LC3, targets Alz-25126726; http://dx.doi.org/10.4161/auto.29640


[38] Paul S, Kashyap AK, Jia W, He YW, Schaefer BC. Selective autophagy component, interacts with GABARAP and is regulated by ULK1 phosphorylation. Autophagy 2014; 10:1749-60; PMID:25212676; http://dx.doi.org/10.4161/aut.29640


[58] Esteve JM, Armengod ME, Knecht E. BRCA1 negatively regulates tumorigenesis in TSC and LAM. Autophagy 2011; 7:1400-1; PMID:21646846; http://dx.doi.org/10.1080/17514120.2011.559123


[64] Li L, Guan KL. Microtubule-associated protein/microtubule affinity-regulating kinase 4 (MARK4) is a negative regulator of the mammalian target of rapamycin complex 1 (mTORC1). J Biol Chem 2013; 288:703-8; PMID:23184942; http://dx.doi.org/10.1074/jbc.C112.396903


