

CellWhere: graphical display of interaction networks organized on subcellular localizations

Lu Zhu, Apostolos Malatras, Matthew Thorley, Idonnya Aghoghogbe, Arvind Mer, Stéphanie Duguez, Gillian Butler-Browne, Thomas Voit, William Duddy

▶ **To cite this version:**

Lu Zhu, Apostolos Malatras, Matthew Thorley, Idonnya Aghoghogbe, Arvind Mer, et al.. CellWhere: graphical display of interaction networks organized on subcellular localizations. Nucleic Acids Research, 2015, 43 (W1), pp.W571-W575. 10.1093/nar/gkv354 . hal-01277971

HAL Id: hal-01277971

https://hal.sorbonne-universite.fr/hal-01277971

Submitted on 23 Feb 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Published online 16 April 2015 *Nucleic Acids Research, 2015, Vol. 43, Web Server issue* ***W571–W575***
doi: 10.1093/nar/gkv354

CellWhere: graphical display of interaction networks organized on subcellular localizations

Lu Zhu[1,2,†], Apostolos Malatras[2,†], Matthew Thorley[2], Idonnya Aghoghogbe[2,3], Arvind Mer[4], Stéphanie Duguez[2], Gillian Butler-Browne[2], Thomas Voit[2] and William Duddy[2,*]

[1]Bioinformatics Department, Bielefeld University, Bielefeld, D-33501, Germany, [2]Center for Research in Myology, Sorbonne Universités, UPMC Univ Paris 06, INSERM UMRS975, CNRS FRE3617, 47 Boulevard de l'hôpital, 75013 Paris, France, [3]Orthopaedics and Musculoskeletal Science, University College London, London, WC1E 6BT, UK and [4]Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, SE-17177, Sweden

Received January 30, 2015; Revised March 27, 2015; Accepted April 02, 2015

ABSTRACT

Given a query list of genes or proteins, CellWhere produces an interactive graphical display that mimics the structure of a cell, showing the local interaction network organized into subcellular locations. This user-friendly tool helps in the formulation of mechanistic hypotheses by enabling the experimental biologist to explore simultaneously two elements of functional context: (i) protein subcellular localization and (ii) protein–protein interactions or gene functional associations. Subcellular localization terms are obtained from public sources (the Gene Ontology and UniProt—together containing several thousand such terms) then mapped onto a smaller number of CellWhere localizations. These localizations include all major cell compartments, but the user may modify the mapping as desired. Protein–protein interaction listings, and their associated evidence strength scores, are obtained from the Mentha interactome server, or power-users may upload a pre-made network produced using some other interactomics tool. The Cytoscape.js JavaScript library is used in producing the graphical display. Importantly, for a protein that has been observed at multiple subcellular locations, users may prioritize the visual display of locations that are of special relevance to their research domain. CellWhere is at http://cellwhere-myology.rhcloud.com.

INTRODUCTION

In the analysis of omics data, a researcher is often confronted with a short list of genes and, by extension, their encoded proteins. This list may simply contain differentially expressed genes from a single experimental comparison, or it may result from some secondary analysis, such as functional enrichment using a tool such as DAVID (1), leading edge analysis of Gene Set enrichments (2), clustering of transcripts based on the correlation of their expression profiles (3) or network clustering based on gene functional associations (4,5). More generally, based on specialist knowledge, genes/proteins may be listed based on their special interest to a particular research project. In any case, to interpret such a list and to formulate mechanistic hypotheses, it is useful to explore previously published data concerning two areas of functional context: (i) subcellular locations at which the proteins have been reported and (ii) interactions between proteins, both within the list and with other proteins outside the list. These two types of data are now available from various public sources, but a tool to combine them in an informative and user-friendly way has not existed in the public domain.

Subcellular localization

Owing to the annotation efforts of model organism databases, high-quality subcellular localization information for the proteins of many organisms can be obtained from two carefully curated sources: UniProt (6) and the Gene Ontology (GO) (7). UniProt stores this information in its 'Subcellular location' field for each protein, while GO annotates proteins to the Cellular Component branch of its ontology. In both annotation systems, terms may vary from low specificity (e.g. 'Membrane') to higher specificity (e.g. 'Gap junction'), and a given protein may be annotated to multiple terms. As of writing, some 1283 terms are in use by UniProt, and 3812 by GO. Terms are fewer in UniProt because they are applied conservatively: in general only the more classically recognized location(s) of a given protein are noted, whereas GO is structured toward a systematic listing of all of the known (published) locations of a protein, even those that are rarely observed. For example, the protein Dystrophin is most studied at the membrane

*To whom correspondence should be addressed. Tel: +33 6 21 81 28 94; Fax: +33 1 42 16 57 00; Email: william.duddy@upmc.fr
†These authors contributed equally to the paper as first authors.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://nar.oxfordjournals.org/ at UPMC on February 23, 2016

of muscle cells and its Uniprot Subcellular location is restricted to this location. However, GO lists several related and sometimes more specific Cellular Components including the 'dystrophin-associated glycoprotein complex' and 'Z disc', but also 'Filopodium' which has been reported not in muscle cells but in platelets. Thus, the UniProt subcellular location field is useful to obtain the 'classically' described location(s) of a protein, whereas GO can suggest locations that are of special interest to a specific research area.

Protein–protein interactions and gene functional associations

Interaction networks are usually based on empirical data from direct physical protein–protein binding assays (such as co-immunoprecipitation or yeast 2-hybrid experiments) and/or from indirect 'functional associations' such as gene co-expression or genetic interactions, but may also incorporate derived knowledge of signaling pathways. Various user-friendly interactome exploration tools are easily accessible to the bench researcher (a few examples include: NetGestalt (5), GeneMANIA (4), PathwayLinker (8), STRING (9), IntAct (10) and Mentha (11)). These tools vary in the types of experimental data that they include. For example, IntAct consists of curated listings of direct protein–protein interactions (PPIs) or colocalizations conforming to the MIMIx standard (for the minimum information required for reporting a molecular interaction experiment (12)), whereas GeneMANIA includes multiple functional association types from a large set of selected publications. Tools may combine data from stringent curation of individual experiments but also from text-mining and predictive approaches. Mentha is one of the more stringent: similarly to IntAct, it limits itself to direct physical PPIs curated by members of the International Molecular Exchange consortium (IMEx; (13)). It is also unique in having both a powerful interface for programmatic access and a simple scoring function that allows query cut-offs based on the strength of the interaction evidence. Each of these interactomics tools may serve different purposes and in CellWhere we make direct use of Mentha to build PPI networks from query lists, but we also facilitate (via Cytoscape 3 (14,15)) the import of networks created using other tools.

RATIONALE

The purpose of CellWhere is to enable bench researchers to quickly explore the reported subcellular locations of a list of genes/proteins, and to put these subcellular locations into the context of previously identified physical interactions that could be occurring between them and other genes/proteins within the cell. As such, CellWhere was created with three goals in mind:

(i) To aid in the formulation of mechanistic hypotheses by showing where proteins are typically described to locate in the cell and what their most strongly evidenced interactions are.
(ii) To act as a screening tool to show whether proteins and their interactors could be at selected locations of special interest.
(iii) To add subcellular location information to gene association networks that have been created using other tools.

METHODS: NETWORK GENERATION, PROTEIN LOCALIZATION AND GRAPH ORGANIZATION

UniProt compiles a downloadable data file that includes both UniProt and GO localizations for all manually annotated (i.e. Swiss-Prot) and non-redundant protein sequences. Mentha maintains a file listing protein interactions. CellWhere downloads these files automatically within 24 h of each UniProt or Mentha update. Identifiers, localizations and interactions are then parsed and organized together with mapping information in a relational database. A CellWhere query begins by first mapping submitted gene symbols or other identifiers to the Swiss-Prot accession of the corresponding protein.

Mentha data are queried to obtain (i) evidence scores for interactions between proteins of the query list and (ii) proteins that interact with the query list, selected based on the strength of the evidence score. In this way a network is created and grown, up to a maximum size set by the user. Certain proteins (for example, Ubiquitins and heat shock proteins) form a great many interactions due to general functions that are unlikely to be pertinent to a specific mechanistic pathway. To filter out such 'promiscuous interactors', CellWhere pre-processes the Mentha data, making interaction counts for every protein. By default, when adding interactors, CellWhere ignores proteins that bind more than 100 partners. This corresponds to 1271 (1.6%) of the 81 919 proteins currently documented by Mentha. The user may adjust this cut-off as desired.

The UniProt accessions of the network are mapped to localization terms from UniProt and/or GO. These terms are then mapped to CellWhere's own localizations. This is achieved by means of a manually created mapping file, which maps UniProt/GO terms to 50 CellWhere localizations. These 50 localizations include all major cell compartments, and 50 is a sufficiently small number that the output visual display does not become overly crowded by different localizations. However, if desired the user may modify the mapping file to add more CellWhere localizations. CellWhere currently maps all Uniprot and GO localization terms that have been applied to more than 25 proteins. This covers more than 99% of all protein localization annotations (1 258 337 out of a total of 1 269 645) and includes the most frequently used 1013 of the 3812 terms that comprise the GO Cellular Component namespace, and 422 of the 1283 terms parsed from the Uniprot Subcellular location field. The user may modify these mappings, including to add mappings for the rarely used terms.

An example localization procedure is shown in Table 1. For a given query gene or protein, all of its retrieved localization terms and their mappings to CellWhere localizations can be downloaded from CellWhere in tabular format, but for proteins with multiple CellWhere locations only selected locations are chosen at which to display the protein in CellWhere's interactive graph display. The user has two options regarding how these locations are chosen: a generic option, in which the most frequently annotated locations

Downloaded from http://nar.oxfordjournals.org/ at UPMC on February 23, 2016

are selected, and a prioritization option, in which a location is selected based on its user-specified score (which the user may set, for example, in accordance with the relevance of the location to their research project). Table 1 shows several examples, indicating (by the red coloring) which CellWhere location would be selected using either of the two options. Selection using the generic option is according to the 'frequency' column, whereas selection by the prioritization option is according to the score set in a user-defined 'flavor' (in this example, we set priority scores according to the location's relevance to muscle physiology, the 'Muscle flavor priority'). Using the Generic option would place RRAD, EMILIN2 and ACTC1, primarily at the Membrane, ECM and Cytoplasm, respectively. Whereas, using the muscle prioritization flavor, ACTC1 would be placed into the 'Focal adhesion' location, because the muscle flavor sets a high priority score on this location, due to its being of special interest to muscle research. The user may choose whether the graph will show only the most frequently annotated location or also show duplicate nodes at alternative locations.

Several pre-made flavors for prioritization scoring are provided, but the user can customize their own flavor by creating and uploading a new mapping file. Instructions to do this are given in drop-down information on CellWhere's front page, and in the User Guide section. As described on the site, users are encouraged to email their flavors to us to be included on the drop-down list available as a pre-made option to all users.

After localizations have been obtained, automated spatial organization of the graph is then achieved using a limited vocabulary that was created to tell CellWhere how to place locations relative to the boundaries of the cell. This vocabulary includes terms such as 'IN Cytoplasm', 'UNDER Membrane' and 'ACROSS Membrane' and is explained in more detail in the user guide. Spatial relation mappings are provided, but may also be set by the user. The Cytoscape.js JavaScript library (http://cytoscape.github.io/cytoscape.js/) is used to produce the graphical display in an html format that is readable by all common web browsers. Cytoscape.js was chosen over other graphing platforms—such as Cytoscape Web (16), D3 (http://d3js.org/) or sigmajs (http://sigmajs.org/)—in large part because of its built-in support for compound nodes (used by CellWhere to group proteins into their localizations), but also its shared philosophy with the Cytoscape desktop application.

RESULTS: VISUAL DISPLAY AND INTERFACE

The user can submit a list of query IDs—several identifier types are supported—or upload a pre-made network (in xgmml format from Cytoscape 3) and has the option to retrieve localizations from UniProt, GO or both. Localizations may be prioritized based on their annotation frequency ('generic') or by priority scores (a user-created 'flavor', or one of those already provided). If the option to add Mentha interactions and interactors is selected, then the maximum size of the network can be selected. If a pre-made network is uploaded, then a parameter (e.g. 'fold-change') may be used on which to color the nodes of the network.

Example output is shown in Figure 1. The localized network is graphed to resemble a physical map of the cell, placing proteins in a way that can help to hypothesize and interpret mechanistic links between genes or proteins of interest. Edges connecting the nodes are thicker when the Mentha evidence to support the interaction is stronger. The graph is interactive: edges can be selected to list Mentha evidence, and links are provided to supporting publications in PubMed; protein nodes link to their UniProt page; nodes and localization groupings can be moved around by the user.

The output may be downloaded in html format for sharing, or as a network for more advanced manipulation in the Cytoscape 3 desktop application (or any other tool capable of importing xgmml format). For each query, a complete unfiltered list of retrieved localizations and their CellWhere location mappings can be downloaded.

As well as the user-friendly interface, there is also an API for programmatic access using the http POST method. This is explained in detail, listing input fields and example code, in the developer guide section of the help menu.

RESULTS: COMPARISON WITH RELATED TOOLS

A general-purpose network manipulation and analysis tool such as Cytoscape can facilitate the integration and visualization of many types of information, including to group or color nodes according to subcellular localizations, but localization information must first be obtained and summarized. Further work is then required if the user desires to organize the graphical layout based on these localizations, such that the network resembles a schematic of the cell. Besides Cytoscape, there are several popular free-to-access tools that are focused on biological network analysis and/or visualization (some of which are listed above), but these tools generally lack the automatic integration of subcellular localization information. However, the localization-related functionalities of CellWhere have limited overlap with two existing pieces of software, one publicly available (Cerebral viewer (17)) and the other commercial (Ingenuity IPA (QIAGEN Redwood City, www.qiagen.com/ingenuity)).

Cerebral viewer is a plug-in currently only available for the older version 2.8 of the Cytoscape desktop application. It facilitates a stratified graph layout based on localizations, but does not provide localization or interaction information, which must be provided by the user (discussed further in Supplementary note S1).

Ingenuity IPA is a data integration and exploration tool for omics data analysis. Provided as part of its network-based clustering approach is a graph output in which genes are positioned into stratified localizations. IPA identifies interactions and localizations using a proprietary knowledgebase derived from Ingenuity's in-house literature curation. To highlight the similarities and differences between IPA and CellWhere, we re-analyzed in CellWhere a previously published (18) network that was produced using IPA (Supplementary note S1). A notable difference was that, whereas the IPA-generated network is limited to primary compartments (nucleus, cytoplasm, membrane, extracellular), CellWhere can display and automatically position numerous sub-compartments. If we ignore sub-compartments

Downloaded from http://nar.oxfordjournals.org/ at UPMC on February 23, 2016

Table 1. In this example of CellWhere's localization procedure, three query IDs are submitted

INPUT		RETRIEVAL			OUTPUT		
Query ID	Uniprot ACC	Localization source	Localization term	Description	CellWhere localization	Frequency	Muscle flavor priority
RRAD	P55042	GO	GO:0005886	plasma membrane	Membrane	100%	2500
		UniProt	Cell membrane	Cell membrane			
		GO	GO:0016020	membrane			
EMILIN2	Q9BXX0	GO	GO:0031012	extracellular matrix	Extracellular matrix	57%	4200
		GO	GO:0005578	proteinaceous extracellular matrix			
		UniProt	extracellular matrix	extracellular matrix			
		GO	GO:0005581	collagen			
		GO	GO:0005576	extracellular region	Extracellular	43%	4100
		UniProt	Secreted	Secreted			
		UniProt	extracellular space	extracellular space			
ACTC1	P68032	GO	GO:0005925	focal adhesion	Focal adhesion	8%	7250
		GO	GO:0031674	I band	Sarcomere	17%	7200
		GO	GO:0030017	sarcomere			
		GO	GO:0005884	actin filament	Actin cytoskeleton	8%	6200
		GO	GO:0070062	extracellular vesicular exosome	Vesicular exosome	8%	4900
		GO	GO:0005615	extracellular space	Extracellular	8%	4100
		GO	GO:0016020	membrane	Membrane	8%	2500
		GO	GO:0005856	cytoskeleton	Cytoplasm	42%	2000
		GO	GO:0005737	cytoplasm			
		GO	GO:0005829	cytosol			
		UniProt	cytoskeleton	cytoskeleton			
		UniProt	Cytoplasm	Cytoplasm			

Localizations are retrieved for their corresponding UniProt (Swiss-Prot) accessions and mapped to CellWhere localizations. For each query protein a single CellWhere localization is selected for display on the network graph. Selection may be 'generic'—based on annotation frequency—or by localization 'flavor'—based on priority scores (provided or set by the user) to select localizations that are of special interest to a particular research domain (in the example, muscle research is chosen).

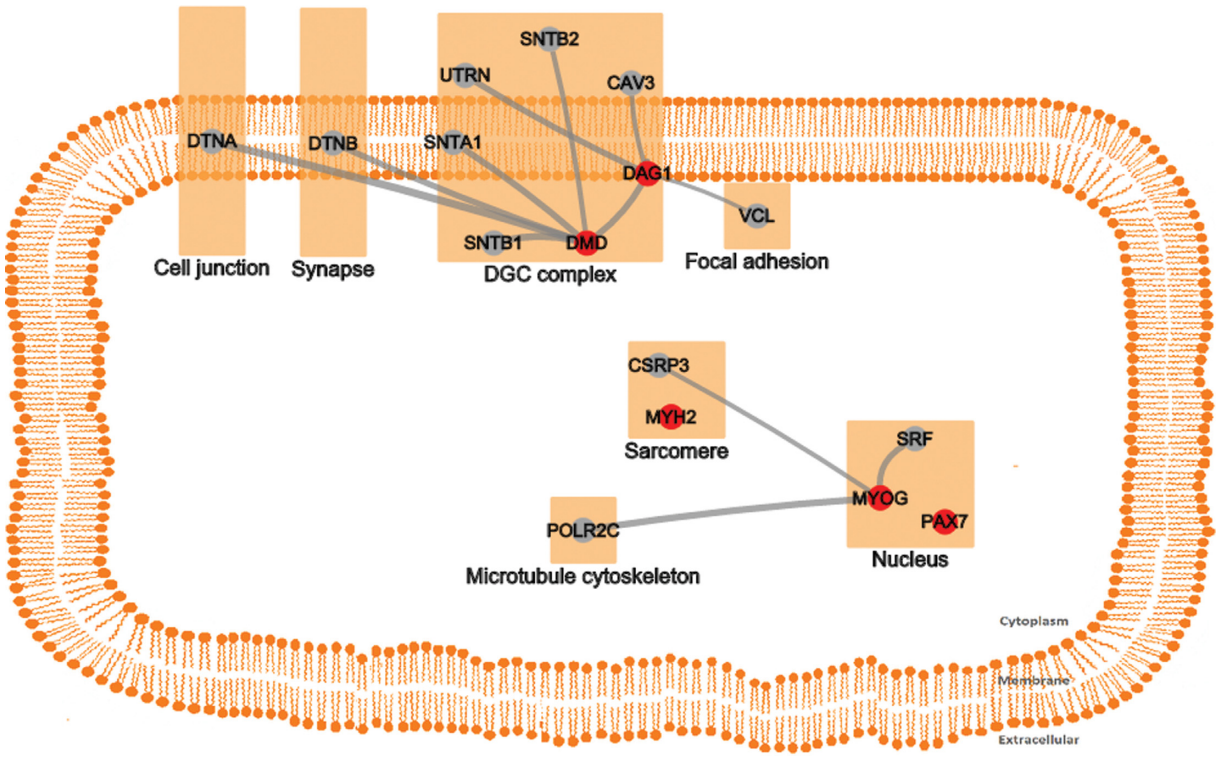


Figure 1. Screenshot of an interactive graph generated by submitting CellWhere's pre-loaded example query. Proteins can be placed into their classically reported locations or, as in this example, the user can prioritize locations that are of special interest to their research area (in this case, muscle; DGC = Dystroglycan complex—a complex of glycoproteins that interact with Dystrophin, located at the muscle cell membrane).

Downloaded from http://nar.oxfordjournals.org/ at UPMC on February 23, 2016

and compare the tools' placements of proteins into just the four primary compartments and use CellWhere's 'generic' localization option, then the placement outputs were similar (87% of proteins were placed into the same compartment). Where differences arose, it was usually among proteins that were annotated to multiple locations by UniProt and GO, usually with clear biological basis: for example, heat shock factor protein 2 (HSF2) is cytoplasmic (where it was placed by CellWhere) during normal growth and moves to the nucleus (where it was placed by IPA) upon activation (more examples are given in Supplementary note S1).

Thus CellWhere showed strong agreement of generic localizations with the IPA-generated network, but it provides a more deeply resolved representation of protein sublocations within the cell, and in addition it provides the facility to highlight rare localizations according to the user's interests.

CONCLUSION

Tools such as Cerebral viewer and Ingenuity IPA have shown that it can be informative and useful to integrate a summary of subcellular localization into an interaction network. Integrated displays can help to suggest mechanistic links between parts of the network. CellWhere is the first free-to-access public tool to summarize subcellular localizations and integrate this information with the local interactome. CellWhere can be used to visually structure a network based on the classically known locations of proteins. Notably, it can also be used as a screening tool to identify proteins (and their interactors) that may be present at locations of special interest to a specific research project or domain.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank David Lynn for kind advice and for sharing details of InnateDB (http://www.innatedb.com), which inspired the localization mapping approach used by CellWhere. We also thank the reviewers for their insightful and constructive comments.

FUNDING

The French Muscular Dystrophy Association; the MyoGrad International Graduate School for Myology, University Pierre and Marie Curie [to M.T. and A.M.]. Funding for open access charge: Association Institut de Myologie.
Conflict of interest statement. None declared.

REFERENCES

1. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
2. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
3. Michalopoulos,I., Pavlopoulos,G.A., Malatras,A., Karelas,A., Kostadima,M.-A., Schneider,R. and Kossida,S. (2012) Human gene correlation analysis (HGCA): a tool for the identification of transcriptionally co-expressed genes. *BMC Res. Notes*, **5**, 265.
4. Zuberi,K., Franz,M., Rodriguez,H., Montojo,J., Lopes,C.T., Bader,G.D. and Morris,Q. (2013) GeneMANIA prediction server 2013 update. *Nucleic Acids Res.*, **41**, W115–W122.
5. Shi,Z., Wang,J. and Zhang,B. (2013) NetGestalt: integrating multidimensional omics data over biological networks. *Nat. Methods*, **10**, 597–598.
6. Activities at the Universal Protein Resource (UniProt). (2014) *Nucleic Acids Res.*, **42**, D191–D198.
7. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
8. Farkas,I.J., Szántó-Várnagy,A. and Korcsmáros,T. (2012) Linking proteins to signaling pathways for experiment design and evaluation. *PLoS One*, **7**, e36202.
9. Franceschini,A., Szklarczyk,D., Frankild,S., Kuhn,M., Simonovic,M., Roth,A., Lin,J., Minguez,P., Bork,P., von Mering,C. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.
10. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., Del-Toro,N. *et al.* (2014) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.
11. Calderone,A., Castagnoli,L. and Cesareni,G. (2013) mentha: a resource for browsing integrated protein-interaction networks. *Nat. Methods*, **10**, 690–691.
12. Orchard,S., Salwinski,L., Kerrien,S., Montecchi-Palazzi,L., Oesterheld,M., Stümpflen,V., Ceol,A., Chatr-aryamontri,A., Armstrong,J., Woollard,P. *et al.* (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat. Biotechnol.*, **25**, 894–898.
13. Orchard,S., Kerrien,S., Abbani,S., Aranda,B., Bhate,J., Bidwell,S., Bridge,A., Briganti,L., Brinkman,F.S.L., Brinkman,F. *et al.* (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods*, **9**, 345–350.
14. Cline,M.S., Smoot,M., Cerami,E., Kuchinsky,A., Landys,N., Workman,C., Christmas,R., Avila-Campilo,I., Creech,M., Gross,B. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.*, **2**, 2366–2382.
15. Smoot,M.E., Ono,K., Ruscheinski,J., Wang,P.-L. and Ideker,T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.
16. Lopes,C.T., Franz,M., Kazi,F., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.
17. Barsky,A., Gardy,J.L., Hancock,R.E.W. and Munzner,T. (2007) Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics*, **23**, 1040–1042.
18. Nghiem,P.P., Hoffman,E.P., Mittal,P., Brown,K.J., Schatzberg,S.J., Ghimbovschi,S., Wang,Z. and Kornegay,J.N. (2013) Sparing of the dystrophin-deficient cranial sartorius muscle is associated with classical and novel hypertrophy pathways in GRMD dogs. *Am. J. Pathol.*, **183**, 1411–1424.

Downloaded from http://nar.oxfordjournals.org/ at UPMC on February 23, 2016