Published online 14 June 2021 *Nucleic Acids Research, 2021, Vol. 49, Web Server issue* ***W153–W161***
https://doi.org/10.1093/nar/gkab449

DGLinker: flexible knowledge-graph prediction of disease–gene associations

Jiajing Hu[1,2,†], Rosalba Lepore[3,†], Richard J.B. Dobson[1,4,5], Ammar Al-Chalabi[2,6], Daniel M. Bean[1,4,†] and Alfredo Iacoangeli[1,2,7,*,†]

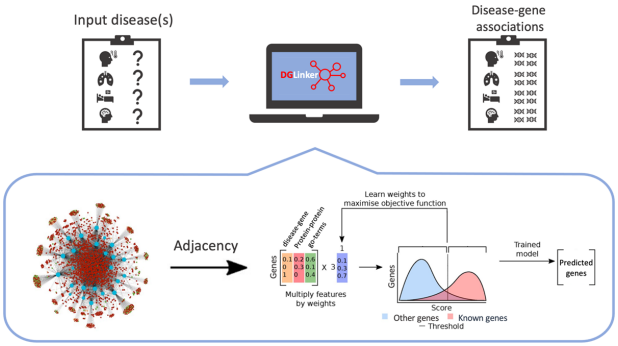
[1]Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology & Neuroscience, King's College London, SE5 8AF, London, UK, [2]Department of Basic and Clinical Neuroscience, Maurice Wohl Clinical Neuroscience Institute, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, SE5 9RT, UK, [3]BSC-CNS Barcelona Supercomputing Center, Barcelona, 08034, Spain, [4]Health Data Research UK London, University College London, London, WC1E 6BT, UK, [5]Institute of Health Informatics, University College London, London, NW1 2DA, UK, [6]King's College Hospital, Bessemer Road, Denmark Hill, London, SE5 9RS, UK and [7]National Institute for Health Research Biomedical Research Centre and Dementia Unit at South London and Maudsley NHS Foundation Trust and King's College London, London, SE5 8AF, UK

Received March 04, 2021; Revised April 30, 2021; Editorial Decision May 10, 2021; Accepted May 17, 2021

ABSTRACT

As a result of the advent of high-throughput technologies, there has been rapid progress in our understanding of the genetics underlying biological processes. However, despite such advances, the genetic landscape of human diseases has only marginally been disclosed. Exploiting the present availability of large amounts of biological and phenotypic data, we can use our current understanding of disease genetics to train machine learning models to predict novel genetic factors associated with the disease. To this end, we developed DGLinker, a webserver for the prediction of novel candidate genes for human diseases given a set of known disease genes. DGLinker has a user-friendly interface that allows non-expert users to exploit biomedical information from a wide range of biological and phenotypic databases, and/or to upload their own data, to generate a knowledge-graph and use machine learning to predict new disease-associated genes. The webserver includes tools to explore and interpret the results and generates publication-ready figures. DGLinker is available at https://dglinker.rosalind.kcl.ac.uk. The webserver is free and open to all users without the need for registration.

GRAPHICAL ABSTRACT



INTRODUCTION

Thanks to the establishment of high-throughput technologies as a common tool in the biomedical field, vast amounts of biological and phenotype information are currently available. Machine learning (ML) is a powerful tool for exploiting this heterogeneous source of knowledge for the prediction of novel associations between biological factors (e.g. genes) and phenotypes. Such predictions can be used for a multitude of purposes including the prioritization of disease genes. Given the large number of targets that high-throughput experiments provide, their individual validation, let alone all the possible interactions between them, is time-consuming and expensive. In some cases, for example for the hundreds of millions of variants from human whole-genome sequencing experiments, this can be prohibitive. In this context, gene prioritization can play an important role (1,2).

*To whom correspondence should be addressed. Tel: +44 7434885640; Email: alfredo.iacoangeli@kcl.ac.uk
†The authors wish it to be known that, in their opinion, these authors contributed equally to the study and manuscript.

The use of ML for the prediction of novel disease-gene associations presents several challenges including the interpretation of the predictions, the selection of appropriate data to generate the model, and an adequate choice of the known disease genes for the training. These are key for usable non-trivial predictions and to avoid model bias (3). Currently available methods generally lack tools for both the interpretation of the predictions and for the evaluation of the model. Moreover, they tend to provide limited flexibility over the data that can be used (1,3–8).

We therefore developed DGLinker, a webserver for the prediction of novel candidate genes for human diseases. DGLinker has a user-friendly interface that allows non-expert users to select a customizable set of databases, and use our in-house ML method (3,9) to predict new candidate genes on the basis of genes that are known to be associated with the target disease (method overview in Figure 1).

The webserver includes utilities to explore and interpret the results including a network visualization tool for a graphical exploration of the interactions between the disease genes and other biological factors in the knowledge-graph (KG) that contributed to their classification. It also performs gene enrichment analysis to test the overrepresentation among the predictions of genes associated with specific biological processes (10). Via its user-friendly interface, DGLinker allows users to select a set of databases, including protein-protein interaction, disease-gene (DG) association, transcriptomics, gene function, text mining of scientific literature, and upload their own data for the generation of the KG. The control over the data used in the model can favour the minimization of trivial predictions and hidden biases, factors that can limit the applicability of this class of methods. DGLinker produces a number of publication-ready figures and graphs. The outputs can be downloaded as csv files for use with spreadsheet programs as well as image files of the graphs and figures. On such basis we believe DGLinker to be a novel and promising resource for human disease research in the era of big biological data and precision medicine.

RESULTS

Webserver overview

DGLinker is a web-based server extension of our previously published knowledge-based ML method (3,9) for the prediction of candidate disease genes. In order to maximise its usability, DGLinker has a user-friendly interface that requires no informatics skills and provides a highly flexible analysis framework that gives the user control over the data used for the generation of the knowledge-graph and the training of the predictive model. Moreover, the webserver provides utilities for the evaluation of the model and the interpretation of the results. These are key aspects that are often overlooked and limit the use of this class of methods in the biomedical field. It is freely accessible and there is no login requirement. The DGLinker pipeline consists of four main steps: (i) specification of known disease associated genes, (ii) selection (and/or upload) of the data to generate the KG, (iii) ML training and DG predictions, (iv) results visualization and evaluation (Figure 2). More details are provided in the corresponding sections below.

Input options

DGLinker bases its prediction on a set of genes known to be associated with the target phenotype(s). Therefore, the user needs to provide both a list of genes and the phenotypes they are associated with. To do so the following three options are available on the webserver. (i) *Select phenotype(s)*: this option allows the user to provide one or more input phenotypes. DGLinker will automatically retrieve all genes reported to be associated with them in the selected DG databases. Currently, DGLinker includes all disease and phenotype terms, and DG associations from DisGeNet (11), OMIM (12), Clinvar (13) and HPO (14). (ii) *Select phenotype(s) associated genes*: This option allows the user to provide a set of genes and the phenotypes they are associated with. If this option is used, DGLinker will replace the corresponding DG relations in the database with the ones provided by the user. The definition of which genes are associated with a target phenotype strongly affects the predictions and varies greatly among DG databases. This is largely dependent on the evidence used to support their association (3). For example, for genes whose variants can increase disease risk, one might consider the results of a genome-wide association study (GWAS) sufficient while others could require evidence of segregation with the disease in families. Neither choices are right or wrong in general and might depend on the study design and aim. As a consequence, it is very common for a user working in the biomedical field, to have their own curated list of DG associations optimised for their specific study. Input option (ii) is designed to facilitate this common scenario. (iii) *Select genes*: This third and last option allows the user to provide a set of genes without selecting a specific phenotype. This is suitable for studying phenotypes that are not present in the DGLinker database. We recommend the users to search for the target phenotype using option (i) or (ii) before using this option. Where this option is used, it is not a requirement that the associated phenotype is necessarily a disease, for example, a user could specify genes linked to a specific biological pathway or drug response.

Available databases

DGLinker has a wide range of databases of biological and phenotypic relations available to generate the knowledge-graph (Table 1). These include a selection of 20 databases grouped in following classes: disease-gene associations, protein–protein interactions, gene pathways, expression data, gene function and biological interactions mined from literature. By default, the latest versions of DisGeNet, Gene ontology (15) and IntAct (16) are selected. Users can also upload their own dataset(s). These have to be in comma delimited csv format and include one ‘Gene’ column. The HGNC nomenclature (17) must be used. Currently, there is a limit of 100Mb to upload datasets, however, this limit can be increased, and new databases can be added on demand.

After submission

After submission, the user is directed onto the waiting page where the unique job ID is displayed. This can be used in

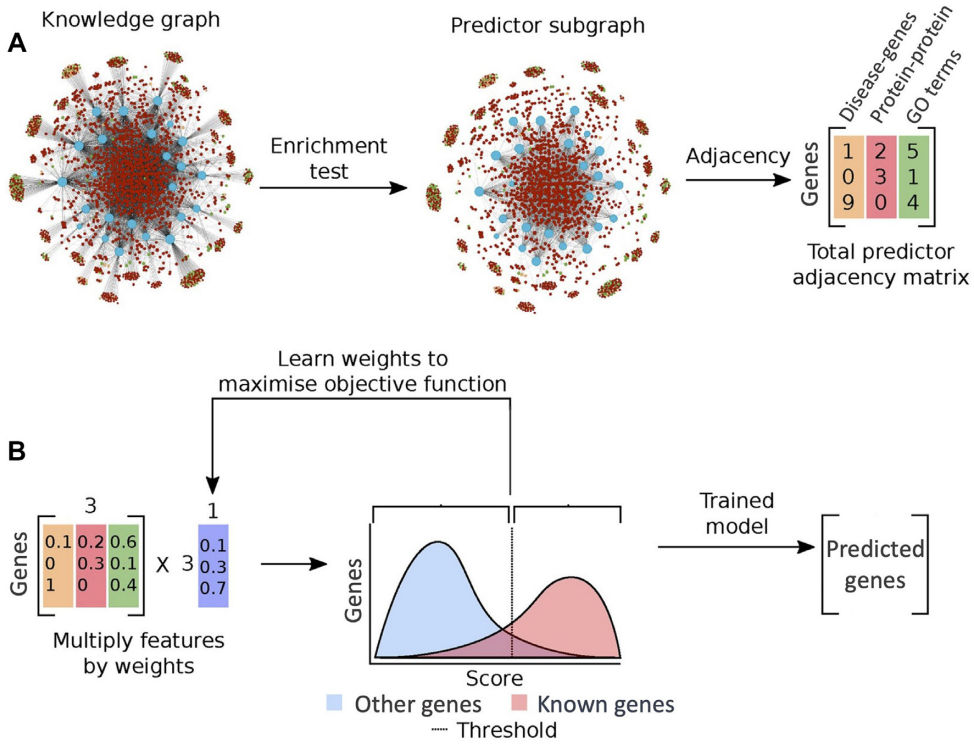


Figure 1. Method overview: The method takes as input a graph of known data related to the prediction task, in this case, gene-disease links, gene functions, and others, and returns a list of predicted edges missing from that graph. (**A**) Starting from a knowledge graph, an enrichment test is used to identify predictive features of the genes known to be associated with the target phenotype(s). The total adjacency of every gene with all predictors of each type (the columns of the matrix) is calculated from the graph. Blue nodes are genes, red nodes are proteins, orange nodes are diseases, green nodes are GO terms. (**B**) The features (adjacency matrix from (a)) are scaled and weighted to produce a final score for every gene. The optimum weighting and score threshold are learned from the set of known associated genes. In other words, to predict new genes linked to a target phenotype, the algorithm compares all genes known to be linked to the target to all other genes and builds a predictive profile based on a weighted combination of existing relationships in the graph. Every gene is then scored for its similarity to this profile. Predictions are made by applying a threshold to this similarity score, with all genes above the threshold predicted as candidate genes. Adapted from Bean *et al.*(9).

the homepage to retrieve the job results. If a valid email address was provided at submission, the job ID and a link to the results page are also emailed. The waiting page automatically refreshes every 10 seconds until the job is completed. The user is then redirected to the results page. A standard job with default data sources takes about 10 minutes to be completed. However, jobs can take up to a few hours as the processing time depends on the number of genes and databases used, as well as whether the cross-validation protocol is used. If the cross-validation is selected, DGLinker performs a standard *N*-fold cross validation protocol (36) (where *N* is selected by the user but $\leq$5) using the input disease genes, and the results are reported in the subsequent model evaluation tab of the results page. Although the *N*-fold cross validation can be a useful tool to evaluate the model performance, it does increase the job processing time by approximately a factor *N*.

Results page

The results page consists of the following five tabs: Job description, Model Performance, Results Table, Enrichment Analysis and Gene Networks. The Job description tab reports all job details including the job ID, the complete list of the input genes and phenotypes, the number of input and predicted genes and the databases used. In the Model Performance tab (Figure 2A) the user can find a set of metrics useful for the assessment of its quality. These include the results of the cross validation, standard metrics accuracy, F1 and area under the ROC curve (37), and a pie chart representation of the overall contribution of each data source to the model. The Results Table (Figure 2B) displays all the genes, both known and predicted, that the model classified as disease associated, ranked by their score. The score for each gene is its similarity to the learned profile of the known disease-associated genes. The absolute value of the score is not meaningful *per se*, only the relative values between genes of the same model are. Links to external resources, such as Gene cards (38), Clinvar (39), KEGG (40), GTEx (20) and STRING (35) are provided for each gene together with a pie chart representation of the contribution of each data source to the gene score.

The following two tabs, Enrichment Analysis and Gene Networks, are dedicated to tools for the interpretation of the results. The Enrichment Analysis tab (Figure 2c) displays and allows the download of the gene enrichment analysis results that DGLinker generates automatically. This tests the overrepresentation of gene sets from nine databases among the input known disease genes, the predicted genes and their union (10). These databases are the

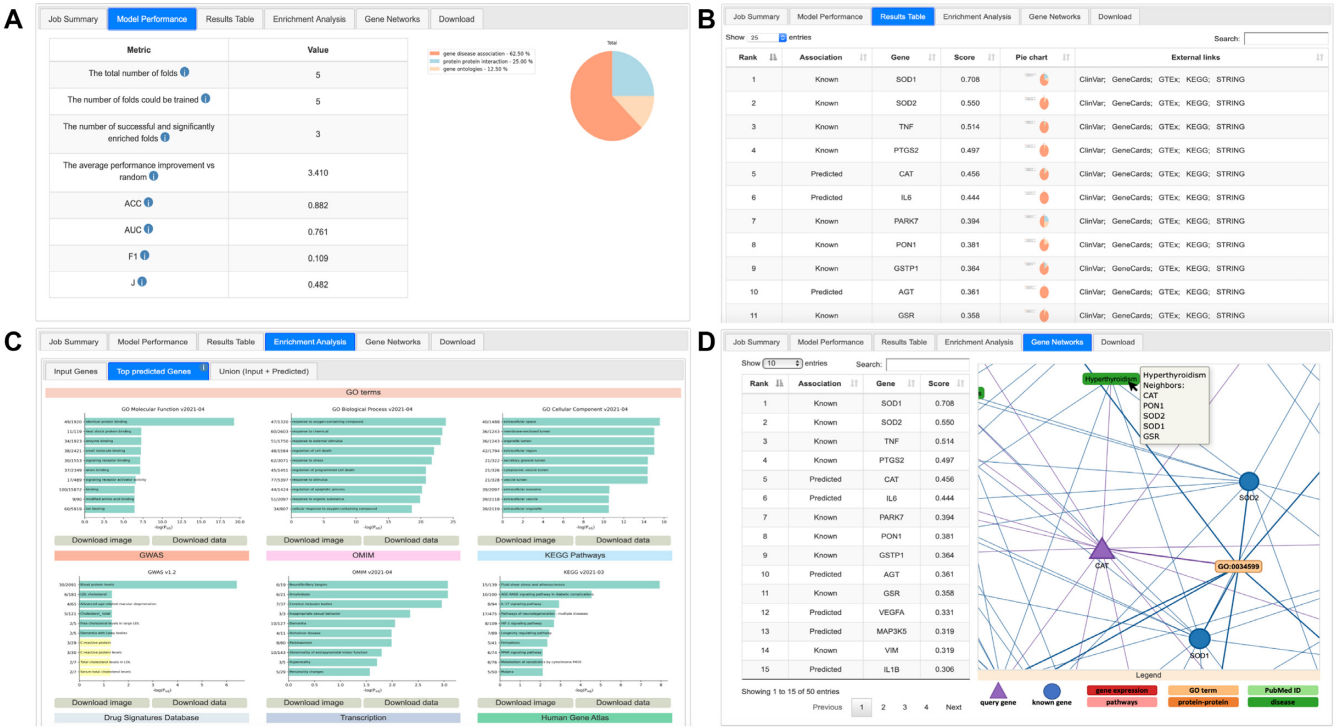


Figure 2. Results visualization and evaluation tabs example. The four panels of this figure display the (**A**) Model Performance tab, (**B**) Results Table, (**C**) Enrichment Analysis and (**D**) Gene Networks for the results of the job example available on the DGLinker website.

Table 1. Current set of databases available on DGLinker

Database	version	Type of data	Source website	citation
Gene Ontology	v2021-02	Gene function (GOterms)	www.current.geneontology.org	Gene Ontology Consortium, 2021 (18)
ArrayExpress Atlas (experiment E-MTAB-513 Illumina body map)	v2021-02	Expression	www.ebi.ac.uk	R. Petryszak *et al.*, 2013 (19)
GTex, Tissue specific gene expression and eQTLs	v8	Expression	www.gtexportal.org	J. Lonsdale *et al.*, 2013 (20)
Human Protein Atlas (HPA)	v20.1	Expression	www.proteinatlas.org	M. Uhlen *et al.*, 2017 (21)
KEGG	v2021-03	Gene pathways	www.genome.jp	M. Kanehisa, *et al.*, 2020 (22)
Reactome	v76	Gene pathways	www.reactome.org	B. Jassal *et al.*, 2020 (23)
ClinVar	v2021-04	Gene-Disease Association	www.ncbi.nlm.nih.gov	M. J. Landrum *et al.*, 2018 (13)
DisGeNet	v7.0	Gene-Disease Association	www.disgenet.or	J. Piñero *et al.*, 20 (24)
HPO	v2021-04	Gene-Disease Association	www.hpo.jax.org	S. Köhler *et al.*, 2021 (25)
OMIM	v2021-04	Gene-Disease Association	www.omim.org	J. S. Amberger, *et al.*, 2015 (12)
BioGrid	v4.3.196	Protein-Protein Interaction	www.thebiogrid.org	A. Chatr-Aryamontri *et al.*, 2017 (26)
IMEx	v2021-02	Protein-Protein Interaction	www.ebi.ac.uk	S. Orchard *et al.*, 2012 (27)
InnateDB	v2021-02	Protein-Protein Interaction	www.ebi.ac.uk	B. Karin *et al.*, 2013 (28)
IntAct	v2021-04	Protein-Protein Interaction	www.ebi.ac.uk	S. Orchard *et al.*, 2014 (29)
MatrixDB	v2021-02	Protein-Protein Interaction	www.ebi.ac.uk	C. Olivier *et al.*, 2019 (30)
Mentha	v2021-02	Protein-Protein Interaction	www.ebi.ac.uk	A. Calderone, *et al.*, 2013 (31)
MINT	v2021-02	Protein-Protein Interaction	www.ebi.ac.uk	A. Chatr-Aryamontri *et al.*, 2011 (32)
UniProt	v2021-02	Protein-Protein Interaction	www.ebi.ac.uk	T. U. Consortium, 2021 (33)
NCBI PubMed	v2021-04	Publications	www.ncbi.nih.gov	NCBI Resource Coordinators, 2018 (34)
String	v11.0	Publications	www.string-db.org	D. Szklarczyk *et al.*, 2016 (35)

GO gene sets for Biological processes (15), Cellular components and Molecular functions (15), the GWAS catalog (41), the OMIM database (gene–disease associations) (42), KEGG (biological pathways) (40), DSigDB (drug signatures) (43), the Encode and Chea consensus database (transcription) (44,45), and the Human Gene Atlas (46)). The enrichment analysis can help researchers gain insight into the phenotype and biological processes underlying the results. The tab is designed to also allow for a direct comparison between the input known genes and the predictions. In the Gene Networks tab (Figure 2D) the user can visualize the interaction network of each individual gene. The visualization of the individual interaction networks allows for the inspection of the biological and phenotypical factors that contributed to the prediction of a given gene, and of which known disease genes such factors are linked to. Finally, all results, including graphs and figures, gene lists and raw data, can be downloaded as a zip archive from the Download button.

Comparison to other available DG prediction webservers

By reviewing the available tools for the prediction of novel candidate DG associations given a target phenotype and a set of known associated genes, that (i) have a user-friendly web interface, (ii) are publicly available and (iii)

are currently functioning, we have identified two such tools, Phen2Gene (7) and Phenolyzer (4). Additionally, three tools, GeneMANIA (6), MaxLink (47) and ToppGenet (48), despite not allowing the direct input of a target phenotype, can be used to predict DG associations by manually providing a set of disease genes selected autonomously by the user, for example by using an external database of DG associations like OMIM, DisGeNet, or ClinVar (39). In comparison to these tools, DGLinker offers a number of advantages in terms of flexibility and availability of the data to build the model, evaluation of the model and interpretation of the predictions (Table 2).

Although GeneMANIA, MaxLink and ToppGenet can make predictions of DG associations, they require the user to perform an extra manual step that is not trivial. Furthermore, GeneMANIA and ToppGenet make use of databases of DG associations for their prediction, as a result, the user-defined input disease genes are likely to largely overlap with DG sets in their databases, leading to trivial and potentially misleading predictions. MaxLink does not use DG databases to build the model, and GeneMANIA is flexible in regard to the databases used so that DG databases can be excluded. However, considering that many human diseases have genetic causes that overlap to some extent or underlie common biological mechanisms, DG databases are a powerful source of information for the prediction of novel DG associations. Therefore, excluding them from the model could impact their performance.

Performance evaluation

We used the DisGeNet data to simulate prospective prediction performance using a temporal hold-out as an external validation set. We trained the model on DG associations up to and including 2018 and evaluated on all subsequent data as of DisGeNet v7.0 (Table 3). The KG contained DG associations from DisGeNet (v7 2018), protein-protein interactions from IntAct (2020-11-06) and gene function from Gene Ontology (2020-11-17). These three databases are the default setting in DGLinker. Assessing the performance of DG predictions presents several challenges. Due to our limited knowledge of the genetic landscape of most human diseases, complete sets of true positives and true negatives are generally not available. As a consequence, classic metrics such as precision and recall, might not be adequate in this context. Instead, we assessed the model performance using a hypergeometric test for enrichment of newly associated genes in the set of predictions from the model vs the background of all genes in the knowledge graph. Enrichment was considered significant if $P < 0.05$ for hypergeometric test for enrichment following 5% false discovery rate correction. 1131 diseases had at least one new associated gene by 2020 in DisGeNet. For 91% (1024) of these, at least one of the new associated genes was in the KG. For 804 diseases, the model could be trained and made predictions of new disease associated genes. The predictions were significantly enriched for new genes for 184 diseases (22.9%). During training, the model is optimising the J statistic defined as sensitivity + specificity – 1. The top-scoring models (training achieved $J \geq 0.9$, Supplementary Figure S1) were significant in 42.2% of cases (146/346).

Given the low number of new associated genes for many phenotypes (median = 2), in some cases even predicting all of them could not result in a significant test. This might result in the underestimation of the model performance. We therefore also reported the overall performance considering only those models that could be significantly enriched given the number of predicted and predictable genes (Table 3, 'validation has sufficient power'). In these cases, the predictions were enriched for new disease genes in 39.1% (200/512) of all models and 45.2% (146/323) of the models whose training achieved $J \geq 0.9$. For an overview of how the performance varies with J please see Supplementary Figure S1 and Table S1.

We also performed a cross validation study on these diseases. For diseases with at least five known genes in the 2018 data we performed 5-fold cross validation. 599 diseases met this condition, with median 17 known genes (interquartile range = 38.5). For 537 (∼90%) of diseases the cross-validation model was significantly enriched for the held-out genes in at least 3 folds. Although informative, it is important to remark that cross validation is likely to over-estimate the external performance as highly similar genes can be separated across validation folds.

Software documentation and data availability

The DGLinker website (http://DGLinker.rosalind.kcl.ac.uk) provides an extensive tutorial section in which step by step instructions with figures guide the user through the steps necessary to perform DG predictions and utilize the tools for model evaluation and results interpretation. The Downloads sub-section of the tutorial provides links to all external resources and software used, as well as the links to the GitHub repositories (49) where the KG-ML method code is available under the GPLv3 licence. The ML method is well documented and also available as an open-source python package (https://pypi.org/project/edgeprediction/). The data used and generated in the evaluation of the tool performance are publicly available on GitHub (https://github.com/KHP-Informatics/DGLinker-validation).

Usage example: Amyotrophic Lateral Sclerosis

The example section presents the results of an application derived from our recent publication (3) in which our method was used for the prediction of novel candidate genes in Amyotrophic Lateral Sclerosis (ALS) using data from early 2019. ALS is a rare (lifetime risk ∼1 in 400 in Europeans), late-onset, fatal disease whose genetic causes are highly heterogenic among patients and largely unknown. Moreover, there is not a complete consensus among ALS experts regarding which genes are implicated with the disease, and as a result, the ALS genes reported in public DG databases vary greatly, ranging from 20 to over 130 (50,51). In this landscape, we have used the DGLinker method to predict candidate ALS genes using four gene sets from as many sources, DisGeNet (101 genes), AL-

Table 2. Comparison of available tools for DG predictions. The webservers are compared in terms of characteristics related to their general design, allowed input, data sources used and results section. A traffic light colour system was used for a rapid visual evaluation. [1]Predictions are not necessarily disease specific, but DG data are used in the model. [2]Only Human Phenotype Ontology terms (HPOs) are allowed, and they have to be retrieved externally by the user. GG stands for gene–gene (interactions among biological factors), DG stands for disease-gene, and TM stands for text mining (associations mined from scientific literature)

	General					Input				Data sources			Predictive Model	Results		
Tool name	DG specific	non-human organisms	non-disease predictions	open-source method	API	phenotypes	genes	DG links	custom	types	user selection	user upload	model evaluation	rank	interpretation tools	figures/graphs generation
Phen2Gene(7)	yes	no	no	yes	yes	yes (HPO)[2]	no	no	no	GG,DG	no	no	no	yes	no	no
Phenolyzer(4)	yes	no	no	yes	no	yes	no	yes	yes	GG,DG	no	no	no	yes	limited	limited
GeneMANIA(6)	no	yes	yes	no	no	no	yes	no	no	GG,DG	yes	yes	no	yes	limited	no
MaxLink(47)	no	yes	yes	yes	no	no	yes	no	no	GG,DG	no	no	no	Yes	no	no
ToppGenet(48)	no	no	yes[1]	no	yes	no	yes	no	no	GG	yes	no	no	yes	no	no
DGLinker	yes	no	yes[1]	yes	no	yes	yes	yes	yes	GG,DG, TM	yes	yes	yes	yes	yes	yes

Table 3. Temporal external validation of predictive performance using DisGeNet. All disease-gene associations in DisGeNet up to 2018 were used to predict associations added by 2020. Significance is determined at threshold $P < 0.05$ after 5% false discovery rate correction for multiple comparisons. 'Overall' = values for all models, 'Significant' = values for all models that were significant in validation, 'Not significant' = values for all models that were not significant in validation. 'IQR' = Interquartile range

	Overall					Significant			Not Significant		
Criteria	Number of diseases	Significant (N, %)	Median (IQR) validation genes in KG	Median (IQR) predictions validated	Median (IQR) training genes	Median (IQR) validation genes in KG	Median (IQR) predictions validated	Median (IQR) training genes	Median (IQR) validation genes in KG	Median (IQR) predictions validated	Median (IQR) training genes
At least 1 new gene by 2020 At least 1 new gene is in KG	1024	170 (16.6%)	1 (2)	1 (1)	6 (19.25)	4 (5)	2 (4)	9 (20)	1 (1)	0 (1)	6 (19)
At least 1 new gene by 2020 At least 1 new gene is in KG Model made at least 1 prediction	804	184 (22.9%)	1 (2)	1 (2)	9 (31)	3 (5)	2 (3.25)	10 (19.25)	1 (1)	1 (1)	8 (32.5)
At least 1 new gene by 2020 At least 1 new gene is in KG Model made at least 1 prediction Training J >= 0.9	346	146 (42.2%)	2 (2)	1 (1)	10 (15.75)	3 (4.75)	2 (2)	7 (13)	1 (1)	0 (0)	12 (20.5)
At least 1 new gene by 2020 At least 1 new gene is in KG Model made at least 1 prediction Validation has sufficient power	512	200 (39.1%)	2 (3)	1 (2)	13 (34)	3 (5)	2 (3)	10.5 (20.25)	2 (2)	0 (1)	16 (38.25)
At least 1 new gene by 2020 At least 1 new gene is in KG Model made at least 1 prediction Validation has sufficient power Training J >= 0.9	323	146 (45.2%)	2 (3)	1 (1)	9 (14)	3 (4.75)	2 (2)	7 (13)	1 (1)	0 (0)	11 (14)

SoD (126 genes) (52), ClinVar (44 genes), a manually-curated list (40 genes) (51) and the union of all these sets (199 genes). In total, 651 genes were predicted. The enrichment analysis highlighted that the predictions were enriched for genes associated with biological processes known to be affected by the ALS pathogenesis, such as angiogenesis (53), lipid metabolism (54), mitochondria activity (55), protein kinase activity (56), superoxide metabolism (57,58), vesicle-trafficking (59), neurotransmitter regulation (60), and with other neurodegenerative diseases for which evidence of phenotypic and genetic overlap with ALS exist, such as Charcot-Marie-Tooth disease, Parkinson's disease, Frontotemporal dementia, Schizophrenia and Alzheimer's Disease. Moreover, the predicted genes were significantly enriched ($P = 0.012$) for genes that were identified to be associated with ALS in subsequent genetic studies, i.e. they were not yet present in the DG databases used in the experiment. These were *ATXN1* (61), *ATXN3* (62), *SCFD1* (62), *CAV1* (63) and *SPTLC1* (64). Only *ACSL5* (62) and *GLT8D1 (65)* were not present among the predicted genes. An extensive discussion and in depth analysis of the predictions can be found in our recent publication (3). The example on the DGLinker website shows the results obtained using the ALS associated genes from DisGeNet as input known disease genes, and the most recent versions of DisGeNet, IntAct and Gene Ontology.

DISCUSSION

As our understanding of disease grows, it becomes possible to predict missing DG links with increasing accuracy. The DGLinker webserver aims to make this predictive capability widely available by automating data pre-processing, providing a range of data and allowing the results to be analysed directly. Although there have been a number of studies to date predicting DG association, DGLinker is the only current webserver tool to automate this increasingly powerful process while providing the necessary flexibility and making the results available for downstream validation.

In some cases, DGLinker does not make new predictions. The primary reasons are a lack of sufficient training data or limited overlap of the selected datasets. In these cases, we recommend adjusting the selection of databases and input genes accordingly.

At present DGLinker includes a number of datasets and analysis tools. We will continue to develop the platform making additional databases and methods for the analysis of the results available. To this end, we would welcome requests of specific databases and tools by the users. Considering the heterogeneity of the genetic architecture and of the underlying biology of human diseases, we recognise the importance of in-depth testing of the method for specific diseases. Following our work on ALS (3), we will perform

studies on single or subgroups of diseases, to explore the performance of DGLinker and provide guidance and custom protocols for such cases via new tutorials on the website or open-access publications.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

UK Research and Innovation; Medical Research Council; South London and Maudsley NHS Foundation Trust; MND Scotland; Motor Neurone Disease Association; National Institute for Health Research; China Scholarship Council; Spastic Paraplegia Foundation. Funding for open access charge: UKRI. D.B. is funded by a UKRI Innovation Fellowship (Health Data Research UK MR/S00310X/1). A.I is funded by the Motor Neurone Disease Association. J.U. is funded by the King's-China Scholarship Council PhD Scholarship programme. This is an EU Joint Programme-Neurodegenerative Disease Research (JPND) project. The project is supported through the following funding organizations under the aegis of JPND-http://www.neurodegenerationresearch.eu/ (United Kingdom, Medical Research Council MR/L501529/1 to A.A.-C., principal investigator [PI] and MR/R024804/1 to A.A.-C., PI]; Economic and Social Research Council ES/L008238/1 to A.A.-C. [co-PI]) and through the Motor Neurone Disease Association. This study represents independent research partly funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The work leading up to this publication was funded by the European Community's Horizon 2020 Programme (H2020-PHC-2014-two-stage; grant 633413). We acknowledge use of the research computing facility at King's College London, Rosalind (https://rosalind.kcl.ac.uk), which is delivered in partnership with the National Institute for Health Research (NIHR) Biomedical Research Centres at South London & Maudsley and Guy's & St. Thomas' NHS Foundation Trusts and part-funded by capital equipment grants from the Maudsley Charity (award 980) and Guy's and St Thomas' Charity (TR130505). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, King's College London, or the Department of Health and Social Care.

Conflict of interest statement. None declared.

REFERENCES

1. Zolotareva,O. and Kleine,M. (2019) A survey of gene prioritization tools for mendelian and complex human diseases. *J. Integr. Bioinformatics*, **16**, 4.
2. Iacoangeli,A., Al Khleifat,A., Sproviero,W., Shatunov,A., Jones,A., Morgan,S., Pittman,A., Dobson,R., Newhouse,S. and Al-Chalabi,A. (2019) DNAscan: personal computer compatible NGS analysis, annotation and visualisation. *BMC Bioinformatics*, **20**, 213.
3. Bean,D.M., Al-Chalabi,A., Dobson,R.J. and Iacoangeli,A. (2020) A knowledge-based machine learning approach to gene prioritisation in amyotrophic lateral sclerosis. *Genes*, **11**, 668.
4. Yang,H., Robinson,P.N. and Wang,K. (2015) Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods*, **12**, 841–843.
5. Hwang,S., Kim,C.Y., Yang,S., Kim,E., Hart,T., Marcotte,E.M. and Lee,I. (2019) HumanNet v2: human gene networks for disease research. *Nucleic Acids Res.*, **47**, D573–D580.
6. Warde-Farley,D., Donaldson,S.L., Comes,O., Zuberi,K., Badrawi,R., Chao,P., Franz,M., Grouios,C., Kazi,F. and Lopes,C.T. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–W220.
7. Zhao,M., Havrilla,J.M., Fang,L., Chen,Y., Peng,J., Liu,C., Wu,C., Sarmady,M., Botas,P. and Isla,J. (2020) Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR Genomics Bioinformatics*, **2**, lqaa032.
8. Kotera,M., Yamanishi,Y., Moriya,Y., Kanehisa,M. and Goto,S. (2012) GENIES: gene network inference engine based on supervised analysis. *Nucleic Acids Res.*, **40**, W162–W167.
9. Bean,D.M., Wu,H., Iqbal,E., Dzahini,O., Ibrahim,Z.M., Broadbent,M., Stewart,R. and Dobson,R.J. (2017) Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Sci. Rep.*, **7**, 16416.
10. Klopfenstein,D., Zhang,L., Pedersen,B.S., Ramírez,F., Vesztrocy,A.W., Naldi,A., Mungall,C.J., Yunes,J.M., Botvinnik,O. and Weigel,M. (2018) GOATOOLS: a Python library for Gene Ontology analyses. *Sci. Rep.*, **8**, 10872.
11. Piñero,J., Bravo,À., Queralt-Rosinach,N., Gutiérrez-Sacristán,A., Deu-Pons,J., Centeno,E., García-García,J., Sanz,F. and Furlong,L.I. (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.*, **45**, D833–D839.
12. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
13. Landrum,M.J., Lee,J.M., Benson,M., Brown,G.R., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D. and Jang,W. (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.
14. Köhler,S., Gargano,M., Matentzoglu,N., Carmody,L.C., Lewis-Smith,D., Vasilevsky,N.A., Danis,D., Balagura,G., Baynam,G., Brower,A.M. *et al.* (2021) The human phenotype ontology in 2021. *Nucleic Acids Res.*, **49**, D1207–D1217.
15. Dolinski,K., Dwight,S., Eppig,J., Harris,M., Hill,D., Issel-Tarver,L., Kasarskis,A., Lewis,S., Matese,J. and Richardson,J. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25–29.
16. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P. and Valencia,A. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, **32**, D452–D455.
17. Povey,S., Lovering,R., Bruford,E., Wright,M., Lush,M. and Wain,H. (2001) The HUGO gene nomenclature committee (HGNC). *Hum. Genet.*, **109**, 678–680.
18. Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
19. Petryszak,R., Burdett,T., Fiorelli,B., Fonseca,N.A., Gonzalez-Porta,M., Hastings,E., Huber,W., Jupp,S., Keays,M. and Kryvych,N. (2014) Expression Atlas update—a database of gene and transcript expression from microarray-and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **42**, D926–D932.
20. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F. and Young,N. (2013) The genotype-tissue expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
21. Uhlen,M., Zhang,C., Lee,S., Sjöstedt,E., Fagerberg,L., Bidkhori,G., Benfeitas,R., Arif,M., Liu,Z. and Edfors,F. (2017) A pathology atlas of the human cancer transcriptome. *Science*, **357**, eaan2507.
22. Kanehisa,M., Furumichi,M., Sato,Y., Ishiguro-Watanabe,M. and Tanabe,M. (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, **49**, D545–D551.
23. Jassal,B., Matthews,L., Viteri,G., Gong,C., Lorente,P., Fabregat,A., Sidiropoulos,K., Cook,J., Gillespie,M. and Haw,R. (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.

24. Piñero,J., Ramírez-Anguita,J.M., Saüch-Pitarch,J., Ronzano,F., Centeno,E., Sanz,F. and Furlong,L.I. (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.*, **48**, D845–D855.
25. Köhler,S., Gargano,M., Matentzoglu,N., Carmody,L.C., Lewis-Smith,D., Vasilevsky,N.A., Danis,D., Balagura,G., Baynam,G. and Brower,A.M. (2021) The human phenotype ontology in 2021. *Nucleic Acids Res.*, **49**, D1207–D1217.
26. Chatr-Aryamontri,A., Oughtred,R., Boucher,L., Rust,J., Chang,C., Kolas,N.K., O'Donnell,L., Oster,S., Theesfeld,C. and Sellam,A. (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res.*, **45**, D369–D379.
27. Orchard,S., Kerrien,S., Abbani,S., Aranda,B., Bhate,J., Bidwell,S., Bridge,A., Briganti,L., Brinkman,F.S. and Cesareni,G. (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods*, **9**, 345–350.
28. Breuer,K., Foroushani,A.K., Laird,M.R., Chen,C., Sribnaia,A., Lo,R., Winsor,G.L., Hancock,R.E., Brinkman,F.S. and Lynn,D.J. (2013) InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Res.*, **41**, D1228–D1233.
29. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C. and Del-Toro,N. (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.
30. Clerc,O., Deniaud,M., Vallet,S.D., Naba,A., Rivet,A., Perez,S., Thierry-Mieg,N. and Ricard-Blum,S. (2019) MatrixDB: integration of new data with a focus on glycosaminoglycan interactions. *Nucleic Acids Res.*, **47**, D376–D381.
31. Calderone,A., Castagnoli,L. and Cesareni,G. (2013) Mentha: a resource for browsing integrated protein-interaction networks. *Nat. Methods*, **10**, 690–691.
32. Chatr-aryamontri,A., Ceol,A., Palazzi,L., Nardelli,G., Schneider,M., Castagnoli,L. and Cesareni,G. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **35**, D572–D574.
33. UniProt Consortium (2021) UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
34. (2018) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **46**, D8–D13.
35. Szklarczyk,D., Morris,J.H., Cook,H., Kuhn,M., Wyder,S., Simonovic,M., Santos,A., Doncheva,N.T., Roth,A. and Bork,P. (2016) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.*, **45**, D362–D368.
36. Browne,M.W. (2000) Cross-validation methods. *J. Math. Psych.*, **44**, 108–132.
37. Obuchowski,N.A. (2005) ROC analysis. *Am. J. Roentgenol.*, **184**, 364–372.
38. Rebhan,M., Chalifa-Caspi,V., Prilusky,J. and Lancet,D. (1998) GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, **14**, 656–664.
39. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*, **42**, D980–D985.
40. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
41. MacArthur,J., Bowler,E., Cerezo,M., Gil,L., Hall,P., Hastings,E., Junkins,H., McMahon,A., Milano,A. and Morales,J. (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, **45**, D896–D901.
42. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
43. Yoo,M., Shin,J., Kim,J., Ryall,K.A., Lee,K., Lee,S., Jeon,M., Kang,J. and Tan,A.C. (2015) DSigDB: drug signatures database for gene set analysis. *Bioinformatics*, **31**, 3069–3071.
44. Wang,J., Zhuang,J., Iyer,S., Lin,X.-Y., Greven,M.C., Kim,B.-H., Moore,J., Pierce,B.G., Dong,X. and Virgil,D. (2012) Factorbook. org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res.*, **41**, D171–D176.
45. Lachmann,A., Xu,H., Krishnan,J., Berger,S.I., Mazloom,A.R. and Ma'ayan,A. (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, **26**, 2438–2444.
46. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M. and Kreiman,G. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 6062–6067.
47. Guala,D., Sjölund,E. and Sonnhammer,E.L. (2014) MaxLink: network-based prioritization of genes tightly linked to a disease seed set. *Bioinformatics*, **30**, 2689–2690.
48. Chen,J., Bardes,E.E., Aronow,B.J. and Jegga,A.G. (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.*, **37**, W305–W311.
49. Dabbish,L., Stuart,C., Tsay,J. and Herbsleb,J. (2012) In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. ACM, pp. 1277–1286.
50. Radunovic,A. and Leigh,P.N. (1999) ALSODatabase: database of SOD1 (and other) gene mutations in ALS on the Internet. European FALS Group and ALSOD Consortium. *Amyotroph. Lateral Scler Other Motor Neuron Disord.*, **1**, 45–49.
51. Iacoangeli,A., Al Khleifat,A., Sproviero,W., Shatunov,A., Jones,A.R., Opie-Martin,S., Naselli,E., Topp,S.D., Fogh,I. and Hodges,A. (2019) ALSgeneScanner: a pipeline for the analysis and interpretation of DNA sequencing data of ALS patients. *Amyotrophic Lateral Sclerosis Frontotemp. Degener.*, **20**, 207–215.
52. Wroe,R., Wai-Ling Butler,A., Andersen,P.M., Powell,J.F. and Al-Chalabi,A. (2008) ALSOD: the Amyotrophic Lateral Sclerosis Online Database. *Amyotroph. Lateral Scler.*, **9**, 249–250.
53. Oosthuyse,B., Moons,L., Storkebaum,E., Beck,H., Nuyens,D., Brusselmans,K., Van Dorpe,J., Hellings,P., Gorselink,M. and Heymans,S. (2001) Deletion of the hypoxia-response element in the vascular endothelial growth factor promoter causes motor neuron degeneration. *Nat. Genet.*, **28**, 131–138.
54. Adibhatla,R.M. and Hatcher,J.F. (2007) Role of lipids in brain injury and diseases. *Future Lipidol.*, **2**, 403–422.
55. Smith,E.F., Shaw,P.J. and De Vos,K.J. (2019) The role of mitochondria in amyotrophic lateral sclerosis. *Neurosci. Lett.*, **710**, 132933.
56. Liscic,R.M. (2015) Molecular basis of ALS and FTD: implications for translational studies. *Arh. Hig. Rada Toksikol.*, **66**, 285–290.
57. Barber,S.C. and Shaw,P.J. (2010) Oxidative stress in ALS: key role in motor neuron injury and therapeutic target. *Free Radic. Biol. Med.*, **48**, 629–641.
58. Bowling,A.C., Schulz,J.B., Jr,Brown and Beal,M.F. (1993) Superoxide dismutase activity, oxidative damage, and mitochondrial energy metabolism in familial and sporadic amyotrophic lateral sclerosis. *J. Neurochem.*, **61**, 2322–2325.
59. Nishimura,A.L., Mitne-Neto,M., Silva,H.C., Richieri-Costa,A., Middleton,S., Cascio,D., Kok,F., Oliveira,J.R., Gillingwater,T. and Webb,J. (2004) A mutation in the vesicle-trafficking protein VAPB causes late-onset spinal muscular atrophy and amyotrophic lateral sclerosis. *Am. J. Hum. Genet.*, **75**, 822–831.
60. Foerster,B.R., Pomper,M.G., Callaghan,B.C., Petrou,M., Edden,R.A., Mohamed,M.A., Welsh,R.C., Carlos,R.C., Barker,P.B. and Feldman,E.L. (2013) An imbalance between excitatory and inhibitory neurotransmitters in amyotrophic lateral sclerosis revealed by use of 3-T proton magnetic resonance spectroscopy. *JAMA Neurol.*, **70**, 1009–1016.
61. Tazelaar,G.H., Boeynaems,S., De Decker,M., van Vugt,J.J., Kool,L., Goedee,H.S., McLaughlin,R.L., Sproviero,W., Iacoangeli,A. and Moisse,M. (2020) ATXN1 repeat expansions confer risk for amyotrophic lateral sclerosis and contribute to TDP-43 mislocalization. *Brain Commun.*, **2**, fcaa064.
62. Iacoangeli,A., Lin,T., Al Khleifat,A., Jones,A.R., Opie-Martin,S., Coleman,J.R., Shatunov,A., Sproviero,W., Williams,K.L. and Garton,F. (2020) Genome-wide meta-analysis finds the ACSL5-ZDHHC6 locus Is associated with ALS and links weight loss to the disease genetics. *Cell Rep.*, **33**, 108323.
63. Cooper-Knock,J., Zhang,S., Kenna,K.P., Moll,T., Franklin,J.P., Allen,S., Nezhad,H.G., Iacoangeli,A., Yacovzada,N.Y. and Eitan,C.

(2020) Rare variant burden analysis within enhancers identifies CAV1 as an ALS risk gene. *Cell Rep.*, **33**, 108456.
64. Dunn-Giroux,T., Gable,K., Gupta,S.D., Mohassel,P., Nalls,M., Donkervoort,S., Piccus,Z., Majumder,S., Proia,R.L. and Le Pichon,C.E. (2020) SPTLC1 mutations associated with early onset amyotrophic lateral sclerosis. *FASEB J.*, **34**, 1.
65. Cooper-Knock,J., Moll,T., Ramesh,T., Castelli,L., Beer,A., Robins,H., Fox,I., Niedermoser,I., Van Damme,P. and Moisse,M. (2019) Mutations in the glycosyltransferase domain of GLT8D1 are associated with familial amyotrophic lateral sclerosis. *Cell Rep.*, **26**, 2298–2306.