Supplementary Material

Gene Ontology Curation of Neuroinflammation Biology Improves the Interpretation of Alzheimer's Disease Gene Expression Data

MATERIALS AND METHODS

Quantification of key annotation results

The annotations contributed by this project to the GO resource [1,2] are attributed to ARUK-UCL and included in the GO Consortium annotation files available through various ftp sites, including the EMBL-EBI ftp site (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/) as well as the GO browsers QuickGO [3] and AmiGO [4]. In order to identify and quantify the annotations contributed by ARUK-UCL as a part of this project, rather than our previous ARUK-funded annotation initiatives (also attributed to ARUK-UCL), the QuickGO annotation browser was used. Specifically, all annotations contributed for the human priority entities (taxon: 9606) and their mouse and rat orthologues (taxa: 10090 and 10116, respectively) were identified between 21 May 2018 and 19 May 2019. The PubMed identifiers (PMIDs) associated with the annotations of the priority entities and their two rodent orthologues were extracted. These PMIDs were used to identify all our PMID-referenced annotations to human, mouse, and rat entities, thus identifying annotations for both priority entities and any other annotated entities. QuickGO statistics were then checked to find ARUK-UCL attributed annotations to other taxa contributed between 21 May 2018 and 19 May 2019. Finally, all our annotations based on ISS evidence (sequence similarity evidence used in manual assertion; ECO:0000250) [5], which had a GO_REF rather than a PMID reference, contributed between 21 May 2018 and 19 May 2019, were identified and counted. The summary of these annotations is presented in Tables 1 and 2.

Additional GO term-specific (Supplementary Table 3) filtering options were used in QuickGO [6] in order to quantify annotation data specific to neuroinflammatory processes (Fig. 1).

Protein-protein interaction (PPI) network construction

In order to support analyses of the miR- and target-centered miR-target molecular interaction networks, we constructed a network of experimentally-validated protein-protein interactions (PPIs) in Cytoscape 3.7.1 [7]. The 17 proteins shown in Figure 2 were used as seeds, and PPI data was imported from 5 PSICQUIC [8] interaction files, including 4 IMEx standard datasets [9], which contain experimentally-demonstrated PPIs: IntAct [10], MINT [11], UniProt [12], and BHF-UCL (IMEx standard PPI data submitted by BHF-funded biocurators [13]) and one non-IMEx EBI-GOA-nonIntAct file, derived from manually curated GO annotations based on experimental physical interaction evidence ('IPI': physical interaction evidence used in manual assertion (ECO:0000353)) [5]. All PPI interaction files were accessed from directly within Cytoscape on 1 July 2019.

RefSeq to UniProt identifier mapping

The majority of RefSeq [14] transcript identifiers, previously identified in the original publication by Avramopoulos et al. [15], were not automatically mapped to UniProt identifiers in PANTHER (322 unmapped IDs in the 'Higher in AD' dataset and 315 unmapped identifiers in the 'Lower in AD' dataset, accessed 12 August 2019). Consequently, in order to re-analyze the 'Higher in AD' and 'Lower in AD' datasets, the UniProt 'Retrieve/ID Mapping' tool [16] was used to retrieve almost all of the UniProt [12] identifiers for both gene lists. Of a total of 52 identifiers, which could not be mapped, 14 were identified manually, 2 were mapped to

RNAcentral identifiers, 35 had been removed from the NCBI RefSeq database, and one was a pseudogene.

Editing of Cytoscape miR-target network files prior to functional GO term enrichment analyses

By default, miR-target Cytoscape 3.7.1 [7] networks include RNAcentral [17] identifiers for miRs, and Ensembl [18] gene identifiers for their protein-coding target genes. GO annotation files (here specifically the 'gene_association.goa_human' used for the functional GO term enrichment analyses using BiNGO [19]) by default include RNAcentral [17] identifiers for miRs, and UniProt [12] identifiers for protein-coding genes. Consequently, in order to perform GO term enrichment analyses on each miR-target network it was first necessary to edit the Cytoscape node table and replace all Ensembl gene identifiers in the 'name' column with UniProt identifiers. Each network was exported from Cytoscape as an xgmml file, opened in a text editor, copied and pasted into a Microsoft Excel spreadsheet and edited using standard functions and formulas. The UniProt mapping tool [16] was used to convert the Ensembl gene identifiers to UniProt identifiers. The edited network file was then copied and pasted back into a text editor and saved in the original xgmml file format recognized by Cytoscape. Finally, the xgmml file was imported into Cytoscape and the 'name' column in the node table was checked to ensure that the identifiers had been updated.

Editing of Cytoscape miR-target network files to improve data visualization

By default, the miR-target Cytoscape 3.7.1 [7] networks include RNAcentral [17] identifiers in the 'Human Readable Label' column of a Cytoscape node table. In order to improve data display, the Cytoscape node table of each miR-target network was exported from Cytoscape and the 'Human Readable Label' column was edited in a Microsoft Excel spreadsheet to replace the RNAcentral identifiers with miR names. The TarBase file available from the RNAcentral database mappings ftp site (ftp://ftp.ebi.ac.uk/pub/databases/RNAcentral/current_release/id_mapping/database_mappings/) was used as reference. Two columns: the 'shared name' column and the 'Human Readable Label' column (including column title rows) were then copied from the spreadsheet, pasted into a text editor and saved as a text file. The edited table was imported back into the Cytoscape network and checked to ensure that the 'Human Readable Label' had been updated.

For Supplementary Tables 1-10, see the Excel file.

REFERENCES

[1] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29.

[2] The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330-D338.

[3] Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R (2009) QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **25**, 3045-3046.

[4] Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, AmiGO Hub; Web Presence Working Group (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**, 288-289.

[5] Giglio M, Tauber R, Nadendla S, Munro J, Olley D, Ball S, Mitraka E, Schriml LM, Gaudet P, Hobbs ET, Erill I, Siegele DA, Hu JC, Mungall C, Chibucos MC (2019) ECO, the Evidence & Conclusion Ontology: community standard for evidence information. *Nucleic Acids Res* **47**, D1186-D1194.

[6] Huntley RP, Binns D, Dimmer E, Barrell D, O'Donovan C, Apweiler R (2009) QuickGO: a user tutorial for the web-based Gene Ontology browser. *Database (Oxford)* **2009**, bap010.

[7] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504.

[8] del-Toro N, Dumousseau M, Orchard S, Jimenez RC, Galeota E, Launay G, Goll J, Breuer K, Ono K, Salwinski L, Hermjakob H (2013) A new reference implementation of the PSICQUIC web service. *Nucleic Acids Res* **41**, W601-606.

[9] Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, Bridge A, Briganti L, Brinkman FS, Brinkman F, Cesareni G, Chatr-aryamontri A, Chautard E, Chen C, Dumousseau M, Goll J, Hancock RE, Hancock R, Hannick LI, Jurisica I, Khadake J, Lynn DJ, Mahadevan U, Perfetto L, Raghunath A, Ricard-Blum S, Roechert B, Salwinski L, Stümpflen V, Tyers M, Uetz P, Xenarios I, Hermjakob H (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods* **9**, 345-350.

[10] Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC,

Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H (2014) The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**, D358-363.

[11] Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P (2016) The Reactome pathway Knowledgebase. *Nucleic Acids Res* **44**, D481-487.

[12] The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506-D515.

[13] Lovering RC, Dimmer EC, Talmud PJ (2009) Improvements to cardiovascular gene ontology. *Atherosclerosis* **205**, 9-14.

[14] Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40**, D130-135.

[15] Avramopoulos D, Szymanski M, Wang R, Bassett S (2011) Gene expression reveals overlap between normal aging and Alzheimer's disease genes. *Neurobiol Aging* **32**, 2319.e2327-2334.

[16] Pundir S, Martin MJ, O'Donovan C, Consortium U (2016) UniProt Tools. *Curr Protoc Bioinformatics* **53**, 1.29.21-15.

[17] The RNAcentral Consortium (2019) RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Res* **47**, D221-D229.

[18] Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, To JK, Laird MR, Lavidas I, Liu Z, Loveland JE, Maurel T, McLaren W, Moore B, Mudge J, Murphy DN, Newman V, Nuhn M, Ogeh D, Ong CK, Parker A, Patricio M, Riat HS, Schuilenburg H, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Zadissa A, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Cunningham F, Yates A, Flicek P (2018) Ensembl 2018. *Nucleic Acids Res* **46**, D754-D761.

[19] Maere S, Heymans K, Kuiper M (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448-3449.

[20] Huntley RP, Kramarz B, Sawford T, Umrao Z, Kalea A, Acquaah V, Martin MJ, Mayr M, Lovering RC (2018) Expanding the horizons of microRNA bioinformatics. *RNA* **24**, 1005-1017.

[21] Garcia O, Saveanu C, Cline M, Fromont-Racine M, Jacquier A, Schwikowski B, Aittokallio T (2007) GOlorize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. *Bioinformatics* **23**, 394-396.

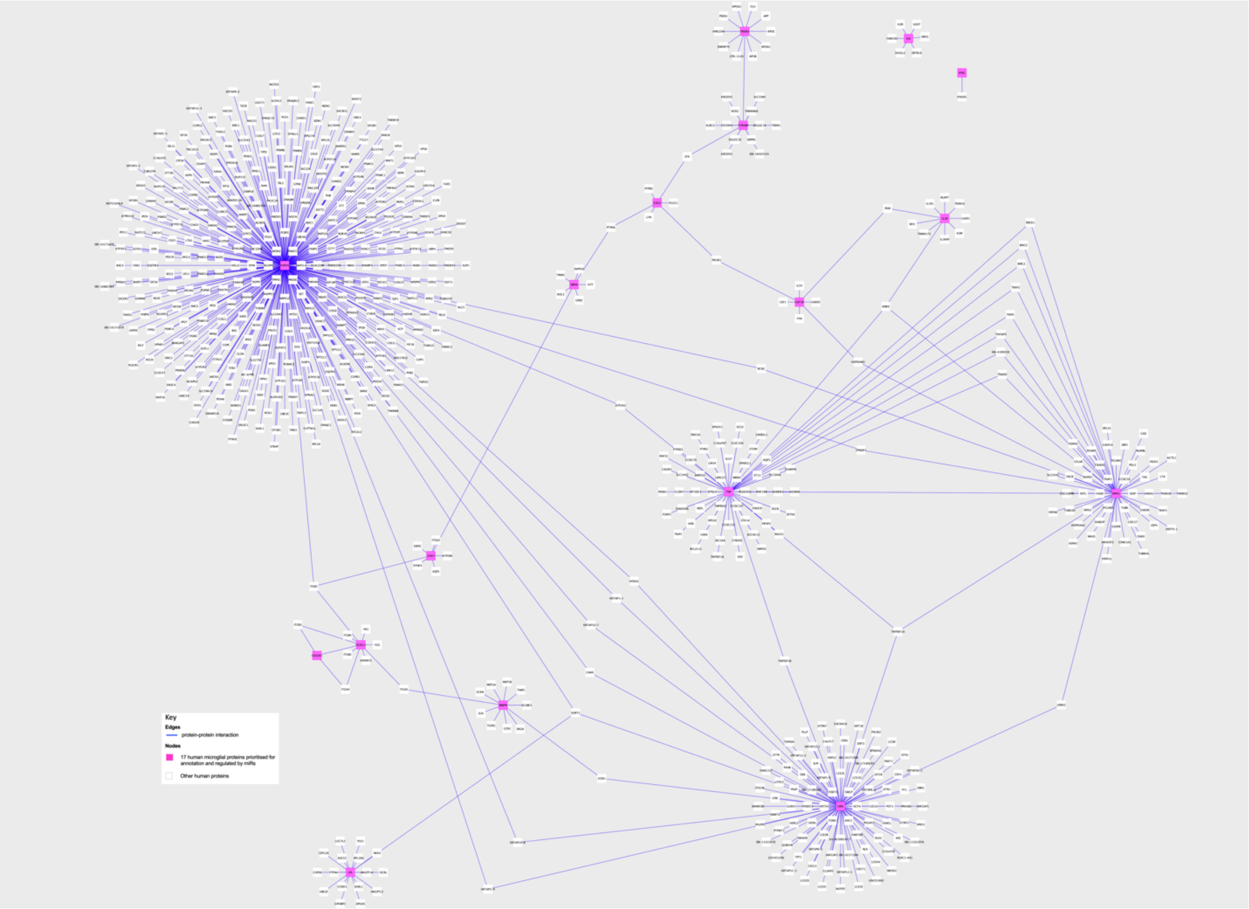
[22] Gray KA, Yates B, Seal RL, Wright MW, Bruford EA (2015) Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res* **43**, D1079-1085.

[23] Kozomara A, Birgaoanu M, Griffiths-Jones S (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res* **47**, D155-D162.

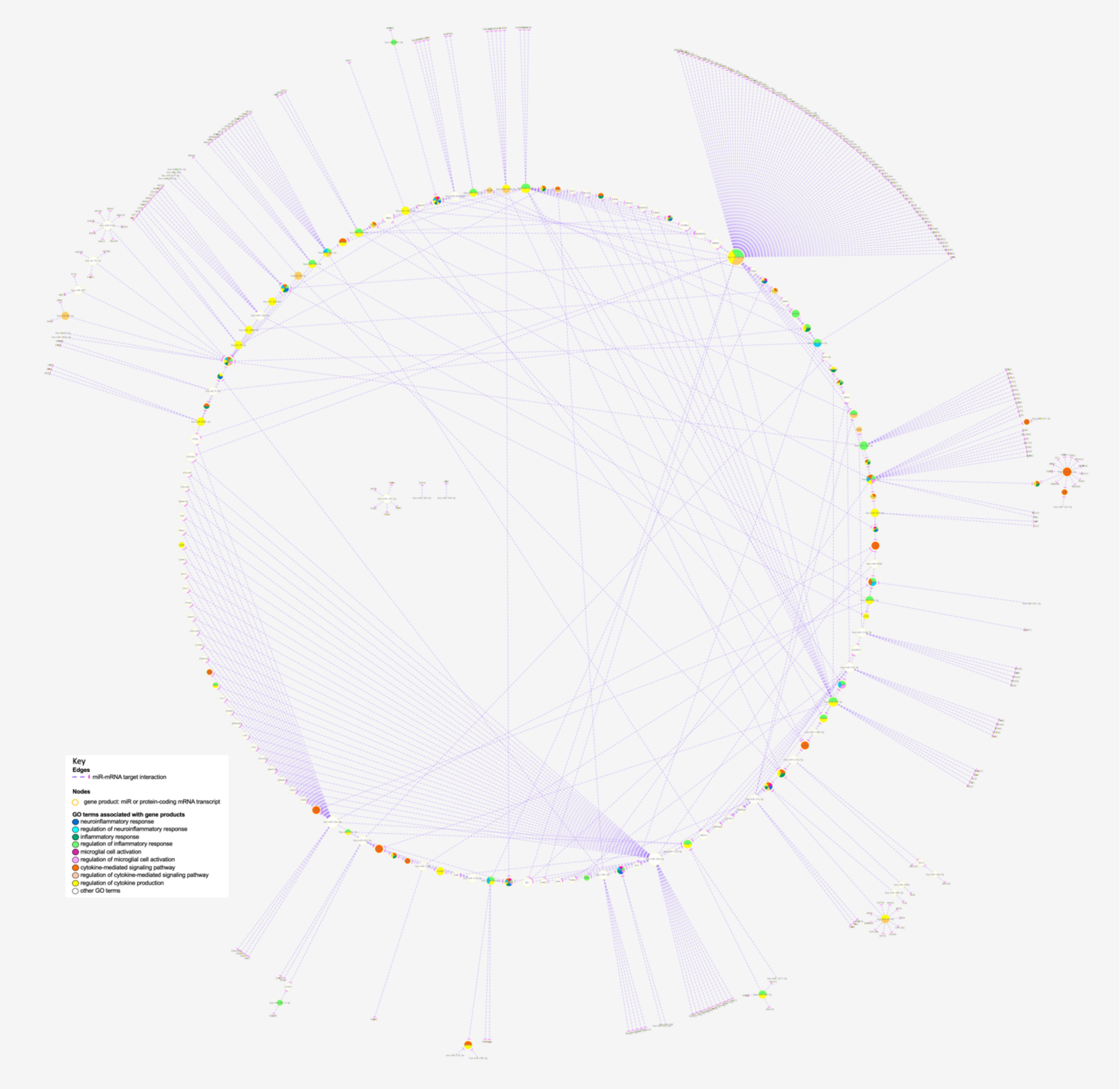
[24] Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD (2019) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* **47**, D419-D426.

[25] Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, Thomas PD (2019) Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat Protoc* **14**, 703-721.

Supplementary Figure 1. AD-relevant microglial protein-protein interaction (PPI) network. The network was seeded in Cytoscape 3.7.1 [7] with UniProt [12] identifiers of the 17 AD-relevant microglial proteins whose expression is shown as being regulated by microRNAs in Figure 2; the magenta nodes correspond to the 17 proteins. Interaction data for the network was imported from the IntAct [10], BHF-UCL [13], MINT [10], UniProt [12], and EBI-GOA-nonIntAct files (accessed 1 July 2019). The 'yFiles Organic Layout' was applied and adjusted manually. The network file was exported from Cytoscape in the xgmml format and is provided in Supplementary Table 7.


Key
Edges
protein-protein interaction
Nodes
17 human microglial proteins prioritised for annotation and regulated by miRs
Other human proteins

Supplementary Figure 2. miR-centered molecular interaction network. The network was constructed in Cytoscape 3.7.1 [7] by seeding with RNAcentral identifiers of microRNAs prioritised for annotation (Supplementary Table 2) and importing molecular interaction data from the EBI-GOA-miR file [20] (accessed 19 August 2019). MiR nodes are embedded within flattened diamond shapes; the other nodes, labelled with the HGNC-approved gene symbols, represent protein-coding targets of miR regulation. All purple dashed edges represent experimentally demonstrated associations between miRs and their targets; the cap on the purple edge faces the target. The BiNGO [19] and GOlorize [21] plugins were implemented for GO term enrichment analysis and visualization of results, respectively. The colors of the nodes' fragments correspond to GO terms shown in the key. The 'Allegro Fruchterman-Reingold Layout' was applied in Cytoscape 3.7.1 [7] and adjusted manually. The network file was exported from Cytoscape in the xgmml format and is provided in Supplementary Table 9. All results of the BiNGO enrichment analysis are provided in Supplementary Table 10.


Key
Edges
miR-mRNA target interaction
Nodes
gene product: miR or protein-coding mRNA transcript
GO terms associated with gene products
neuroinflammatory response
regulation of neuroinflammatory response
inflammatory response
regulation of inflammatory response
microglial cell activation
regulation of microglial cell activation
cytokine-mediated signaling pathway
regulation of cytokine-mediated signaling pathway
regulation of cytokine production
other GO terms