Modeling multifunctionality of genes with secondary gene co-expression networks in human brain provides novel disease insights

Juan A. Sánchez[1+], Ana L. Gil-Martinez[2+], Alejandro Cisterna[1], Sonia García-Ruíz[2], Alicia Gómez-Pascual[1], Regina H. Reynolds[2], Mike Nalls[3,4], John Hardy[2], Mina Ryten[2]*, Juan A. Botía[1,2]*

[1]Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, Murcia, Spain

[2]Department of Neurodegenerative Diseases, UCL Institute of Neurology, London, UK

[3]Molecular Genetics Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, USA

[4]Data Tecnica International, Glen Echo, MD 20812, USA

+,* These authors contributed equally

*To whom correspondence should be addressed.

Abstract

Motivation: Co-expression networks are a powerful gene expression analysis method to study how genes co-express together in clusters with functional coherence that usually resemble specific cell type behaviour for the genes involved. They can be applied to bulk-tissue gene expression profiling and assign function, and usually cell type specificity, to a high percentage of the gene pool used to construct the network. One of the limitations of this method is that each gene is predicted to play a role in a specific set of coherent functions in a single cell type (i.e. at most we get a single <gene, function, cell type> for each gene). We present here GMSCA (Gene Multifunctionality Secondary Co-expression Analysis), a software tool that exploits the co-expression paradigm to increase the number of functions and cell types ascribed to a gene in bulk-tissue co-expression networks.

Results: We applied GMSCA to 27 co-expression networks derived from bulk-tissue gene expression profiling of a variety of brain tissues. Neurons and glial cells (microglia, astrocytes and oligodendrocytes) were considered the main cell types. Applying this approach, we increase the overall number of predicted triplets <gene, function, cell type> by 46.73%. Moreover, GMSCA predicts that the SNCA gene, traditionally associated to work mainly in neurons, also plays a relevant function in oligodendrocytes.
Availability: The tool is available at GitHub, https://github.com/drlaguna/GMSCA as open-source software.
Contact: juanbot@um.es for software issues and use and mina.ryten@ucl.ac.uk for results.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gene co-expression networks (GCN) are a combination of gene clusters and graph networks, based on the correlation of mRNA levels from gene expression profiling(Botía, Juan A. et al., 2017; Langfelder and Horvath, 2008; Miller et al., 2010; Oldham et al., 2008). Genes appearing together in a cluster or as neighbors in the network are said to be co-expressed. GCN analysis is a powerful tool for determining genes associated with

molecular mechanisms underlying biological processes of interest and for defining the function of a gene in a cell type using bulk-tissue transcriptomic data. GCNs provide insights into gene function in specific cell types by detecting gene clusters (i.e. modules) enriched for cell type markers. By applying the "guilt-by-association" (GBA) heuristic(van Dam et al., 2017; Wolfe et al., 2005) to GCNs, all genes in a cell-type enriched module are then predicted to relate to a single cell type within which, they share the same function.

However, there is ample evidence to suggest that a single gene may have different biological functions in different cellular contexts. For example, the tumor suppressor gene, *TP53*, which encodes tumor protein p53 has different roles depending on its interaction partners and consequently is implicated not only in DNA damage and repair, but also in the initiation of apoptosis and senescence(Gillis and Pavlidis, 2011). Yet current GCN analyses cannot capture such complexity even if it is reflected in transcriptomic data because commonly used forms of GCN analysis assign a single gene to a single module (see supplementary methods, section 2.1., for a detailed discussion on this). This is a significant limitation, especially when considering tissues with high cellular heterogeneity such as human brain tissue, where cellular context is likely to be of key importance to the understanding of gene function. Thus, elucidating the different roles of critical genes, while accounting for different contexts, such as expression in different cell types, could reveal new insights into the comprehension of gene function(Eating Disorders Working Group of the Psychiatric Genomics Consortium et al., 2020; Hekselman and Yeger-Lotem, 2020; Muratore et al., 2017; Reynolds et al., 2019).

While cell-specific transcriptomic analyses may provide a means of addressing that issue, there are significant challenges associated with the construction of co-expression-based approaches in this context. Although single-cell technology is evolving rapidly, the gene expression data generated remains sparse in nature, with low sensitivity on the reads per gene and cell types detected when using single-cell and single-nucleus RNA-sequencing approaches, respectively. Other issues include drop-outs or transcriptional bursting (i.e. the time depending variation of transcription activity)(Buettner et al., 2015; Handley et al., 2015; Haque et al., 2017; Hicks et al., 2018; Kolodziejczyk et al., 2015; Zhu et al., 2018). Furthermore, cell-specific human brain gene expression data sets generated from large numbers of individuals are rare due to the associated costs, and are largely limited to single-nucleus RNA-sequencing (snRNAseq) data as tissue is primarily sampled from post-mortem tissue. These snRNAseq data sets not only tend to be particularly sparse, but also have systematic biases, such as the under-representation of genes expressed within the neuropil (i.e. the synaptically enriched area of the central nervous system). To overcome some of these challenges, alternative techniques have been developed to study specific cell types in bulk tissue, for example by deconvoluting the specific contribution of each cell type to gene expression(Baron et al., 2016; Dong et al., 2020; Newman et al., 2015, 2019; Tsoucas et al., 2019; Wang et al., 2019).

Thus, the development of tools to maximize the value of the large quantities of deeply sequenced and publicly available human brain bulk-tissue transcriptomic datasets remains important. In this study, we develop a new method to investigate gene multifunctionality in bulk-tissue transcriptomic datasets. Here, we define multifunctionality as the association of a gene to multiple biological functions within a tissue as a result of its different cellular contexts. Particularly, we are primarily interested in the cell type context of genes. And our aim is to uncover different cell types and functions for the same genes and in this way opening new ways to study diseases with a genetic basis. We propose Gene Multifunctionality in Secondary Co-expression network Analysis

(GMSCA) to investigate gene multifunctionality on gene expression profiling from bulk tissue (see Figure 1a and b). GMSCA is applicable when single-cell based data is not available yet but also in settings where there are paired bulk-tissue and single-cell samples. Importantly, GMSCA has been developed not to focus on producing estimates of each cell contribution to gene expression but on predicting gene function in each cell type the gene is expressed in. This process involves two steps. Firstly, GMSCA constructs a primary gene co-expression network (PGCN) from a gene expression matrix to produce a primary set of triplets <gene, cell type, function> from the annotated PGCN (see section 2.2. of supplementary methods). All genes found in those triplets are said to be typed. Secondly, GMSCA creates secondary gene co-expression networks (SGCN) for each of the target cell types. For each cell type and modules enriched for that cell type within the PGCNs, GMSCA removes the contribution of that cell type from the expression matrix. GMSCA now constructs, with the newly created gene expression matrix (which includes all the original gene pool), a SGCN and extracts new triplets <gene, cell type, function> (see Figure 1c and supplementary methods). When a gene was not identified within a triplet in the PGCN, but now appears in any triplet from the SGCNs, we say the gene is activated. If, on the contrary, it appeared within a triplet generated from the PGCN but not within any of the SGCN triplets, it is termed deactivated (see Figure 1 b for details).

GMSCA increases the number of triplets obtained by 46.73% on average compared to conventional GCN, so notably increasing its utility. Therefore, secondary co-expression networks can augment existing tissue deconvolution approaches as they generate additional functional annotation of different cell types. These new functional annotations include multiple annotations for single genes (i.e., they shape gene´s multifunctionality), creating new applications for this well-established analysis.

2   Methods

We created three network families in the form of R packages as resources: CoExp10UKBEC, CoExpROSMAP and CoExpGTEx. Networks were created with WGCNA and refined with k-means with CoExpNets R package(Botía, Juan A. et al., 2017).

CoExp10UKBEC comprises GCNs created from confirmed non pathological human brain samples of 10 different brain areas. Samples were profiled for gene expression with Affymetrix Human Exon v2. Microarrays. Identical gene pool was used to construct co-expression networks (details of GCN construction here (Forabosco et al., 2013)).

ROSMAP is a transcriptomics human frontal cortex resource with RNA-seq based gene expression profiling for 640 samples, corrected for batch effect using ComBat. Age, sex, RIN and PMI (Post Mortem Interval) were regressed out of individual gene expression. GCNs were constructed from residuals. The 640 samples were arranged into four different groups to create 4 GCNs. One with all samples and three depending on neuritic plaques deposition postmortem diagnostic (a group for values 1, a second group for values 2 and 3 and a third group for values 4) with 200, 158, and 221 respectively.

We developed 13 co-expression networks for the 13 available brain areas in GTEx RNA-seq samples. We corrected samples for batch effect with ComBat. Then, we generated surrogated variables to model unknown effects in data and regressed those, age, sex, PMI and RIN from the overall

gene expression. GCNs were constructed from residuals. See the supplementary for the effect of sample size in GCN construction.

The procedure for creating both PGCNs and SGNCs is identical. The smooth parameter that satisfies scale free topology is found. An adjacency matrix is calculated and the subsequent topology overlap matrix (TOM) is obtained. The clustering is performed using 1-TOM as distance. Gene clusters are obtained by the cuTree algorithm with 100 as the minimum number of genes. A refinement step based on the k-means algorithm with 50 iteration steps is applied afterwards. Genes are ranked within the modules through the Pearson correlation of their expression and the 1st principal component of gene expression of the whole module, the module membership, MM. Gene clusters are annotated with function using gProfiler (see details here (Botía, Juan A. et al., 2017)). All networks are annotated for cell type as follows.

Module Annotation for function and cell type

To annotate network modules for specific brain cell types, we tested modules for brain specific gene markers for the four main brain cell types (neuron, microglia, astrocytes and oligodendrocytes, (see Figure 1d) using manually curated cell marker datasets found at the CoExpNets package. We redesigned the original Fisher´s Exact test employed in CoExpNets to assess the significance of the overlap between genes at modules and genes at marker sets to account for module size as there is significant association between network module size and -log10() transformation of p-values from the tests in all PGCNs, mean R^2 0.28, min 0.17, max 0.45 (supplementary Figure 4 a, b). After applying Bonferroni correction, to account for multiple testing, we regressed out gene set size effect from p-values significance (supplementary Figure 4 c, d, supplementary methods, section 2.3.2.) reducing the number of positive tests from an average of 22 per network to just a mean of 11.03. Note that GMSCA is not restricted to either to the cell types considered here, nor the gene marker sets from CoExpNets. It can be tailored to any tissue and its corresponding cell types of interest provided that there are suitable cell type specific markers.

Human Phenotype Ontology Enrichment Analysis

Human Phenotype Ontology (HPO) terms associated with each human gene were downloaded from HPO (Köhler et al., 2019). To determine the enriched Human Phenotype Ontology (HPO) terms we have selected the relevant overrepresented phenotypes in each module, these are phenotypes that are in at least 2% of the genes from the module and Fold Change (FC) is a measure of how many times that phenotype is more likely to be found in the module than by chance (HPO ontology)

Secondary co-expression networks generation

After generating the PGCN of a gene expression matrix, GMSCA detects modules enriched for the cell types of interest. For each gene in a module enriched for cell type ct, GMSCA generates a prediction triplet <gene, ct, function> where *function* refers to the Gene Ontology based annotation for all genes in the module. Then it builds SGCNs by applying a transformation to the gene expression of genes at modules enriched for the cell types of interest. GMSCA´s model of gene expression for any gene g as follows

$$e = \sum_{c \in CT}' \alpha_c x_c + \beta \, x_0,$$

where e is the gene expression for a specific gene, CT is the set of all cell types considered of interest for the tissue, x_c is the particular contribution per sample of cell type c to the gene, and x_o represents other unknown factors contributing to the expression. GMSCA assumes this additive model to decompose the gene expression matrix of each cell type enriched module as follows. Let us suppose that module m is enriched for a particular cell type. GMSCA removes the contribution of that cell type to gene expression from module m by starting with the $M_{S,m}$ matrix (i.e. gene expression of genes in module m for all samples S), it then obtains the set of PCAs to explain 90% of the variance, and compounds a new matrix with them, $T_{m,m}$, the transformed gene expression matrix. It then removes the 1st PCA from that matrix to create $T'_{m-1,m-1}$. Note that the 1st PCA is assumed to be the cell type contribution to gene expression that we remove. $T'_{m-1,m-1}$ is finally used to reconstruct the original expression, using all PCA axes but the 1st. The current matrix, $M'_{S,m}$, is now free of the detected cell type contribution. This matrix is then used by GMSCA to create a new GCN that we call secondary, i.e. SGCN. GMSCA constructs, in this way, one SGCN for each cell type. Note that it is extremely rare to find gene modules enriched for more than one cell type. And in those cases, the enrichment of one of the cell types is almost marginal, while the other is highly significant. GMSCA drops out the marginal signal.

Cell-type specific gene expression estimates for Barres

We downloaded supplementary table 4 from (Zhang et al., 2016) which contains expression levels for neurons, mature astrocytes, microglia and oligodendrocytes. Within each cell type, we averaged all available values for each gene. We scale and log transform that matrix. Then we generate a boolean matrix of specific gene expression in cell types, $M_{g \times ct}$ with genes in rows and cell types in columns such that M[i,j] is set to TRUE when gene i has greater expression than the mean expression from the whole expression matrix.

To assess the multifunctionality prediction overlap between Barres data and GMSCA, we looked at GMSCA predictions from PGCNs and SGCNs, and assessed the overlap in the corresponding cell type at the Barres´ data. We report a Fisher´s exact test on this overlap. In each comparison, we remove known gene markers to avoid optimistic and unfair overlap assessments.

Cell-type specific gene expression estimates for sc-ROSMAP

We downloaded the sc-ROSMAP filtered count matrix from the Synapse Web site. In sc-ROSMAP, we can find 48 samples of human frontal cortex tissue arranged into 24 Alzheimer's disease cases and 24 controls. We used control, dropped all zero count cell samples and generated a Boolean matrix of gene expression in neurons, microglia, mature astrocytes and oligodendrocytes, $M_{g \times ct}$ with genes in rows and cell types in columns such that M[i,j] is TRUE when gene i has a fold change of 3 of expression in cell type j, with respect to the overall mean expression within that gene and cell type.

We assess the significance (Fisher's exact test) of the overlap between all four ROSMAP network predictions and single cell observations. We removed known gene markers to avoid optimistic and unfair overlap assessments and aggregated excitatory and inhibitory neurons into a single type.

Using EWCE to replicate cell-specific expression enrichment from GSMCA predictions

EWCE was used to assess whether multifunctional predictions from GMSCA show any enrichment in expression in equivalent cell types from mouse. We tested each predicted cell type gene set from GMSCA for enrichment in the equivalent cell type at EWCE with tye bootstrap.enrichment.test() from the EWCE package, we converted human genes in our networks into MGI symbols using the mouse_to_human_homologs data table from EWCE. Gene markers were removed from the analysis. We used all MGI genes at the mouse_to_human_homologs table as the background gene set and 10000 permutations for each test. Results with P < 0.05 were reported as significant.

3   Results

To investigate the utility of GMSCA for the prediction of gene multifunctionality in brain, we used 27 co-expression networks from three brain-derived gene expression datasets (see Table 1), including 13 bulk RNA-seq networks from GTEx's (The GTEx Consortium et al., 2015) brain regions and 4 bulk RNA-seq networks from the ROSMAP (A. Bennett et al., 2012a, 2012b) project, all variations of frontal cortex samples. Further, we reused 10 bulk AffyExon Array networks derived from UKBEC's 10 brain regions (UK Brain Expression Consortium et al., 2014). We refer to them as the GTEx, ROSMAP and UKBEC network families respectively.

The average gene pool for the three network families is 18682 and the average number of modules (clusters) is 30. As GMSCA uncovers multifunctionality, it focuses on modules enriched for markers of the cell types of interest. The average cell type enriched module size is 622 genes. Note that a percentage of those genes are cell type markers themselves; on average 158 genes, i.e., 25% of the whole module. Therefore, 75% of the genes in those modules generate new <gene, cell type, function> triplets in that cell type. Overall, the average PGCN generates more than 4500 of those prediction triplets. We focus on the major cell types in brain: neurons, microglia, astrocytes and oligodendrocytes (see Figure 1 d). We found that 44% of the PGCN modules enriched for markers of the major cell types were tagged as neurons, 21% as microglia, 18% as astrocytes and 17% as oligodendrocytes. Thanks to this approach, we gained the capability of studying gene multifunctionality by discovering new gene multifunctionality in specific cells with an additional 46.73% of the overall gene pool. 37% of the gene pool refers to newly activated genes and 9.73% are additional cell types for genes that become multifunctional (see Figure 1 c).

GCN modules enriched for cell type markers are reliable

Predictions made by GMSCA come from cell marker enriched modules. Then we assessed whether those modules were replicable (i.e., preserved), and therefore reliable, across brain areas. If a gene module is preserved in a similar tissue, then the module is credible and replicable. If the module is not preserved, it might be due to two reasons. Either the module is specific for that tissue, and hence reliable, or the module is not reliable, i.e., most genes are found there just by chance. We used WGNCA's preservation analysis based on the estimation of a Z statistic called Z

summary (see [24] for a detailed explanation). Values of Z over 10 for a module suggest strong module preservation. Values over 2 suggest some preservation. Modules with Z under 2 are not preserved at the tissue tested (see Figure 2).

We first tested GTEx PGCNs´s modules enriched for cell type-specific markers for preservation across the remaining GTEx PGCNs. This resulted in 560 preservation tests. 8.6% of those tests were negative (i.e. Z < 2) suggesting that those modules were not preserved in the remaining brain PGCNs, while 40.4% of the tests showed signs of preservation (Z >= 2) and 50.8% showed strong preservation (Z >= 10). We performed 630 tests at UKBEC networks and 1.9% were negative. 13.3% pointed to weak preservation and the remaining 84.7% indicated strong preservation. All ROSMAP target modules were very well preserved as we may expect: one of the PGCNs includes all ROSMAP samples, and the other three include subsets of the samples used within that PGCN.

Next, we assessed the 52 GTEx SGCNs (each PGCN generates four new GCNs) for preservation. This resulted in 5157 tests being performed of which 4.4% were negative (a cell type enriched module was not preserved in a GTEx tissue), 25.9% implied weak preservation and 69.6% yielded strong preservation. We performed the same analysis for the 40 UKBEC SGCNs. In this case, 2310 preservation tests were performed and 1.12% of the tests were negative, whereas 17.7% of the tests showed signs of preservation and 81.12% showed strong preservation. All preservation tests in ROSMAP SGCNs showed strong preservation.

Part of the difference in non-preserved modules between GTEx and UKBEC can be explained due to PGCNs module size. The Z preservation value correlates positively with module size (R^2 0.38, supplementary Figure 1) (Langfelder et al., 2011). And maximum module size in GTEx PGCNs is 1663, and 2953 in UKBEC. Additionally, we wanted to further investigate whether the non-preserved GTEx modules could have had some biological meaning. We performed a functional enrichment analysis on the genes at those modules, based on the Gene Ontology, REACTOME and KEGG pathway databases using gProfileR R package (Reimand et al., 2007). On average, we obtained 164 annotation terms per non-preserved module, suggesting these modules were biologically relevant. Their lack of preservation may indicate that these modules were specific to their tissue.

GMSCA multifunctional predictions replicate well in cell-type-specific datasets

To assess whether GMSCA multifunctionality predictions replicate in cell-specific brain datasets we considered two cell-specific gene expression datasets in human brain cortex, sc-ROSMAP and Barres Lab data, which are different in nature and a mouse brain expression dataset (Zeisel et al., 2018) through the EWCE (Skene and Grant, 2016) tool.

One of the human datasets is sc-ROSMAP (Mathys et al., 2019), from single cell transcriptomics in a fraction of bulk-RNAseq ROSMAP paired brain cortex samples (24 AD cases and 24 controls). The other is based on immunopanning of temporal lobe cortex samples (Mathys et al., 2019; Zhang et al., 2016) from Barres Lab. This involved purification of specific cell types using cell surface markers followed by gene expression profiling on these purified cell types. As we already noted, sc-ROSMAP and Barres' datasets are of different nature but both were reduced to a list of

cell type specific genes for each of the cell types we are using in this paper (see methods). These lists are used then to compare with the cell multifunctionality predictions generated by GMSCA. Cell markers within GMSCA are removed from the analysis to avoid optimistically biased results.

Multifunctionality predictions from all four ROSMAP PGCNs yielded significant overlap with sc-ROSMAP data (see supplementary File 2). For example, in the NotAD network the overlaps are as follows: Fisher´s Exact Test (FET): P < 4.27E-36 for neurons, P < 3.27E-46 for oligodendrocytes, 0.017 for microglia and 2.94E-42 for astrocytes. SGCN predictions generated overlaps with the following significance: FET P < 2.02E-36 for neurons, P < 2.4E-7 for oligodendrocytes, P < 2.3E-3 for Microglia and P < 1.28E-3 for astrocytes.

Barres Lab based predictions were assessed for overlap with our cortex networks (see supplementary file 2). Focusing, for example, on the same tissue used within the Barres paper, temporal lobe (the 10UKBEC TCTX GCN), GMSCA generated highly significant overlaps in primary predictions with P < 4.51E-67 for neurons, P < 1.94E-31 for oligodendrocytes, P < 7.2E-13 for microglia and P < 1.01E-84 for astrocytes. Secondary predictions yielded significant overlaps for neurons (P < 1.19E-73), oligodendrocytes (P < 4.09E-42) and astrocytes (P 2.02E-12), but not for microglia (P 0.91).

GMSCA predictions were assessed for replication on a different-species and more comprehensive single-cell brain transcriptomic dataset, through EWCE (Expression Weighted Cell Enrichment). This tool enables easy integration with mouse brain transcriptomics from 19 brain regions (Zeisel et al., 2015) and can be used to test whether a given gene list provided by the user presents higher expression levels in each cell type from the reference dataset than expected by chance. We tested all cell type predictions from PGCNs and SGCNs to assess whether EWCE recapitulated similar enrichments. This analysis showed that 100% of the GMSCA gene sets predicted to be functional in a given cell type were also found to be significantly enriched in expression by EWCE in the corresponding cell type (see methods, supplementary File 5)

GMSCA multifunctional predictions show concordance across the three network families in cortex-like tissues

As we used three brain network families, we could compare same-tissue networks between families to assess the agreement of predictions across their networks. Cortex is the only tissue present in all three families but also putamen and substantia nigra in GTEx and UKBEC (remind that ROSMAP is only frontal cortex). We assessed the significance of each pairwise overlap between (gene, cell-type) predictions from both primary and secondary networks, for each cell type and pair of frontal cortex tissues. All Fisher's exact tests performed on the pairwise overlaps were highly significant, being the higher p-value of 1.3E-318 (see supplementary File 4 and supplementary figure 2).

In regard to the putamen brain area, 35% of the neuron predictions made from the UKBEC putamen networks (primary and secondary), are also found as neuron in the GTEx putamen networks. Note that the putamen UKBEC PGCN generates no microglia related predictions (it generates 2686 predictions in the SGCN though). The agreements for

oligodendrocytes and astrocytes are 22.01% and 49.76%, respectively. For the substantia nigra brain area, the numbers for neuron, microglia, oligodendrocytes and astrocytes are 28.62%, 35.98%, 56.26%, and 42.91% respectively (see supplementary File 4 and supplementary figure 3).

GMSCA links the SNCA gene to oligodendrocytes

The *SNCA* gene encodes the alpha-synuclein protein, located at the long arm of chromosome 4 at position 4q22.1. Mutations in this gene cause Parkinson disease (PD) (Appel-Cresswell et al., 2013; Krüger et al., 1998; Lesage et al., 2013; Polymeropoulos, 1997; Zarranz et al., 2004). *SNCA* is also linked to other neuro-degenerative diseases, mainly Alzheimer's disease (Hashimoto and Masliah, 1999). The alpha-synuclein protein is involved in the regulation of neurotransmitter release, synaptic function and plasticity of dopaminergic neurons (Cheng et al., 2011). *SNCA* is also are associated with neuronal dysfunction. Therefore, it has been traditionally seen as highly relevant to neurons. Thanks to applying GMSCA to our networks, the generated multifunctionality predictions included an association of *SNCA* to oligodendrocytes in 4 SGCNs, see below (supplementary File 1).

SNCA is predominantly expressed in brain tissue (Figure 3 a), according to GTEx control samples bulk expression data V8 (The GTEx Consortium et al., 2015). Moreover, *SNCA* is mainly expressed in neurons and oligodendrocytes (Zhang et al., 2016) in the human cortex (Figure 3 b-c), as the Barres Lab data shows. In fact, *SNCA* was found in neuron-enriched modules in 10 out of the 13 GTEx PGCNs, all 4 ROSMAP networks and in 8 out of the 10 UKBEC networks. The only alternative cell type linked to *SNCA* is oligodendrocytes through the GTEx SGCNs of substantia nigra, putamen and hippocampus and the UKBEC temporal cortex.

The substantia nigra is a central tissue in PD. Therefore we focused the subsequent analysis on this tissue (Figure 3 d). *SNCA* is found in the orangered2 module of the PGCN (module membership, MM = 0.84, see methods). That module is enriched for dopaminergic neuron markers (P < 4.59E-25) and GO terms like catecholamine and dopamine biosynthetic processes (GO:0042423 P < 1.1E-5; GO:0042416 P < 9.9E-5), locomotory behavior (GO:0007626 P < 4.1E-4), neurotransmitter transport (GO:0006836 P < 1.9E-5), axonogenesis (GO:0007409 P < 6.4E-4), dopaminergic neuron differentiation (GO:0071542 P < 8.0E-3) and ferric ion transport (GO:0015682 P < 4.6E-4). Dopaminergic neuronal loss is the hallmark of PD. Moreover, the aforementioned biological processes are highly relevant to PD (Aggarwal et al., 2019; Nutt et al., 2004; Sawada et al., 2013). Interestingly, the darkorange2 module includes six more genes linked to PD (supplementary table 1). All this evidence suggests this network module is linked to PD.

When we applied GMSCA to create the corresponding SGCN, *SNCA* is found in the substantia nigra midnightblue module (MM = 0.53), which is enriched for markers of oligodendrocytes (P < 1.48E-4). This finding is replicated using the putamen and hippocampus co-expression networks, belonging to the same network family, GTEx. Moreover, themidnightblue module shared 27 genes with the *SNCA* SGCN module in putamen (Fisher´s exact test on the overlap significant, P < 5.3E-11) and 45 with the *SNCA* SGCN module at the hippocampus, P < 2.2E-16, suggesting function similarity of these modules across the three tissues linking *SNCA* to oligodendrocytes. The midnightblue module (Figure 3 e) was enriched for the following GO terms: synaptic plasticity (GO:0048167 P < 1.3E-4), neuronal plasticity (GO:0048168 P < 1.9E-3), regulation of synapse

organization (GO:0050807 P < 3.5E-3), axonal fasciculation (GO:0007413 P < 7.2E-3), neuron projection fasciculation (GO:0106030 P < 7.2E-3), axon development (GO:0061564 P < 3.5E-3), neuron recognition (GO:0008038 P < 4.1E-3), regulation of neurotransmitter levels (GO:0001505 P < 5.5E-3), microtubule polymerization (GO:0046785 P < 1.4E-3), regulation of glutamate receptor signaling pathway (GO:1900449 P < 3.6E-3) and synaptic vesicle exocytosis (GO:0016079 P < 2.1E-3). Note that BP terms like "regulation of neurotransmitter levels" and "synaptic vesicle exocytosis" are processes found in the orangered2 module of the substantia nigra PGCN. Terms specific to the SGCN module, and also linked to oligodendrocytic activity, were: regulation of synapse organization (Eroglu and Barres, 2010), regulation of glutamate receptor signaling pathway(Gautier et al., 2015), axon development, axonal fasciculation, neuron projection fasciculation, neuron recognition or microtubule polymerization [50]. Interestingly, the midnightblue module included *APP* (MM = 0.84). Mutations in *APP* increase the risk of Alzheimer's disease (Köhler et al., 2019). *APP*'s expression pattern in specific brain cells is similar to *SNCA*'s pattern, mainly expressed in neurons and oligodendrocytes (see Figure 3 c). The *APP* gene has been reported to have a role in regulating axonal myelination in oligodendrocytes (Truong et al., 2019).

A gene based analysis of the phenotypes linked to the midnightblue module (see methods) uncovered HPO terms like Dementia (HP:0000726 fold change 2.55), Myoclonus (HP:0001336 FC 2.13), Variable Expressivity (HP:0003828 FC 1.85), Anxiety (HP:0000739 FC 1.65), Intellectual severe disability (HP:0010864 FC 1.56), Depressivity (HP:0000716 FC 1.24), Tremor (HP:0001337 FC 1.22), Ataxia (HP:0001251 FC 1.2) and Dystonia (HP:0001332 FC: 1.16).

4   Discussion

We propose GMSCA (Gene Multifunctionality Co-expression Analysis), as a method and software tool to uncover additional co-expression profiles for genes, apart from those we obtain with conventional co-expression analyses. GMSCA significantly enhances the power of GCN analyses, delivering the capability to study gene multifunctionality in specific cells, as models of gene co-expression in bulk tissue. GMSCA needs the same inputs as conventional co-expression analysis on bulk tissue such as gene expression profiles, gene marker sets for the cell type under study and gene sets enrichment analysis tools as gProfileR for pathway or GO based module annotation. This paper's experiments are based on brain samples. Therefore, the SGCN obtained were focused on neurons, microglia, astrocytes and oligodendrocytes cell types. Interestingly, this approach increased the number of predictions of gene multifunctionality in specific cells with an additional 46.73% of the overall gene pool.

To demonstrate the potential of this method, we firstly assessed the reliability of GMSCA prediction triplets by means of its application to three network families for three different transcriptomics-based projects, UKBEC, GTEx and ROSMAP using WGCNA's preservation analysis. We demonstrated that the cell-type enriched modules that GMSCA used to generate multifunctionality predictions were stable, including those in SGCNs. Part of the difference we obtained in non-preserved modules between GTEx and UKBEC could be explained by PGCNs module size.

Secondly, we showed a high level of replication of predictions in cell-type specific external data sets. As we already noted, sc-ROSMAP and Barres' datasets are of different nature but both were reduced to a list of cell type

specific genes for each of the cell types. Multifunctionality predictions from all four ROSMAP PGCNs yielded significant overlaps with the single nucleus ROSMAP dataset. We assessed prediction replication in Barres Lab data by looking at how our cortex networks' predictions replicated in Barres cell specific gene expression profiles. All cell types but microglia showed highly significant levels of agreement. Moreover, thanks to the EWCE tool we were capable of assessing GMSCA predictions on mouse single-cell transcriptomics. 100% of both PGCN and SGCN gene sets yielded significant expression enrichment in equivalent brain cells.

Then, we tested the level of agreement in GMSCA predictions on the same tissues across the three networks families (frontal cortex tissue across the three families and substantia nigra and putamen tissues across GTEx and UKBEC) and demonstrated that agreement was high. We compared same-tissue networks between families and assessed the agreement level of predictions for cell marker enriched modules across their networks, in PGCNs and SGCNs, for preservation across brain areas.

We used this SGCN to get details about the biological function of *SNCA* within specific cell types, such as how relevant *SNCA* was in the corresponding module and what genes *SNCA* co-expressed with, e.g. APP linked to dementia and Alzheimer. In this study, *SNCA* was found in oligodendrocytes through the GTEx SGCNs of substantia nigra, putamen and hippocampus and through the UKBEC SGCNs of temporal cortex. The relationship of *SNCA* and oligodendrocytes is supported by previous studies that have observed alpha-synuclein-containing inclusions ("coiled bodies") in oligodendrocytes in parkinsonian brains (Wakabayashi et al., 2000). Recently, the link between oligodendrocytes and Parkinson's disease was reinforced by integrating GWAS results with single-cell transcriptomic data (Eating Disorders Working Group of the Psychiatric Genomics Consortium et al., 2020) through testing the genes tagged by the GWAS outcome into the EWCE tool. Thus, GMSCA has the potential to shed light on gene function within specific cell types and molecular processes, particularly in a disease context. To conclude, we would like to emphasize that the utility of GMCSA is not limited to brain tissue, but this method could be easily tailored to address questions around gene multifunctionality in any other tissue assuming the availability of expression profiling data and high quality, relevant cell-specific markers.

References

A. Bennett, D., A. Schneider, J., Arvanitakis, Z., and S. Wilson, R. (2012a). Overview and Findings from the Religious Orders Study. Curr. Alzheimer Res. *9*, 628–645.

A. Bennett, D., A. Schneider, J., S. Buchman, A., L. Barnes, L., A. Boyle, P., and S. Wilson, R. (2012b). Overview and Findings from the Rush Memory and Aging Project. Curr. Alzheimer Res. *9*, 646–663.

Aggarwal, A., Reichert, H., and VijayRaghavan, K. (2019). A locomotor assay reveals deficits in heterozygous Parkinson's disease model and proprioceptive mutants in adult *Drosophila*. Proc. Natl. Acad. Sci. *116*, 24830–24839.

Appel-Cresswell, S., Vilarino-Guell, C., Encarnacion, M., Sherman, H., Yu, I., Shah, B., Weir, D., Thompson, C., Szu-Tu, C., Trinh, J., et al. (2013). Alpha-synuclein p.H50Q, a novel pathogenic mutation for Parkinson's disease: α-Synuclein p.H50q, A Novel Mutation For Pd. Mov. Disord. *28*, 811–813.

Baron, M., Veres, A., Wolock, S.L., Faust, A.L., Gaujoux, R., Vetere, A., Ryu, J.H., Wagner, B.K., Shen-Orr, S.S., Klein, A.M., et al. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst. *3*, 346-360.e4.

Botía, Juan A., Vandrovcova, J., Forabosco, P., Guelfi, S., D'Sa, K., Hardy, J., Lewis, C.M., Ryten, M., and Weale, M.E. (2017). An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks. BMC Syst. Biol. *11*.

Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol. *33*, 155–160.

Cheng, F., Vivacqua, G., and Yu, S. (2011). The role of alpha-synuclein in neurotransmission and synaptic plasticity. J. Chem. Neuroanat. *42*, 242–248.

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., and de Magalhães, J.P. (2017). Gene co-expression analysis for functional classification and gene–disease predictions. Brief. Bioinform. bbw139.

Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C.M., Zou, F., and Jiang, Y. (2020). SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. Brief. Bioinform.

Eating Disorders Working Group of the Psychiatric Genomics Consortium, International Headache Genetics Consortium, 23andMe Research Team, Bryois, J., Skene, N.G., Hansen, T.F., Kogelman, L.J.A., Watson, H.J., Liu, Z., Brueggeman,

L., et al. (2020). Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson's disease. Nat. Genet. *52*, 482–493.

Eroglu, C., and Barres, B.A. (2010). Regulation of synaptic connectivity by glia. Nature *468*, 223–231.

Forabosco, P., Ramasamy, A., Trabzuni, D., Walker, R., Smith, C., Bras, J., Levine, A.P., Hardy, J., Pocock, J.M., Guerreiro, R., et al. (2013). Insights into TREM2 biology by network analysis of human brain gene expression data. Neurobiol. Aging *34*, 2699–2714.

Gautier, H.O.B., Evans, K.A., Volbracht, K., James, R., Sitnikov, S., Lundgaard, I., James, F., Lao-Peregrin, C., Reynolds, R., Franklin, R.J.M., et al. (2015). Neuronal activity regulates remyelination via glutamate signalling to oligodendrocyte progenitors. Nat. Commun. *6*.

Gillis, J., and Pavlidis, P. (2011). The Impact of Multifunctional Genes on "Guilt by Association" Analysis. PLoS ONE *6*, e17258.

Handley, A., Schauer, T., Ladurner, A.G., and Margulies, C.E. (2015). Designing Cell-Type-Specific Genome-wide Experiments. Mol. Cell *58*, 621–631.

Haque, A., Engel, J., Teichmann, S.A., and Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. *9*.

Hashimoto, M., and Masliah, E. (1999). Alpha-synuclein in Lewy Body Disease and Alzheimer's Disease. Brain Pathol. *9*, 707–720.

Hekselman, I., and Yeger-Lotem, E. (2020). Mechanisms of tissue and cell-type specificity in heritable traits and diseases. Nat. Rev. Genet. *21*, 137–150.

Hicks, S.C., Townes, F.W., Teng, M., and Irizarry, R.A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics *19*, 562–578.

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J.-P., Gargano, M., Harris, N.L., Matentzoglu, N., McMurry, J.A., et al. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. *47*, D1018–D1027.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015). The Technology and Biology of Single-Cell RNA Sequencing. Mol. Cell *58*, 610–620.

Krüger, R., Kuhn, W., Müller, T., Woitalla, D., Graeber, M., Kösel, S., Przuntek, H., Epplen, J.T., Schols, L., and Riess, O. (1998). AlaSOPro mutation in the gene encoding α-synuclein in Parkinson's disease. Nat. Genet. *18*, 106–108.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

Langfelder, P., Luo, R., Oldham, M.C., and Horvath, S. (2011). Is My Network Module Preserved and Reproducible? PLoS Comput. Biol. *7*, e1001057.

Lesage, S., Anheim, M., Letournel, F., Bousset, L., Honoré, A., Rozas, N., Pieri, L., Madiona, K., Dürr, A., Melki, R., et al. (2013). G51D α-synuclein mutation causes a novel Parkinsonian-pyramidal syndrome: SNCA G51D in Parkinsonism. Ann. Neurol. *73*, 459–471.

Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J.Z., Menon, M., He, L., Abdurrob, F., Jiang, X., et al. (2019). Single-cell transcriptomic analysis of Alzheimer's disease. Nature *570*, 332–337.

Miller, J.A., Horvath, S., and Geschwind, D.H. (2010). Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proc. Natl. Acad. Sci. *107*, 12698–12703.

Muratore, C.R., Zhou, C., Liao, M., Fernandez, M.A., Taylor, W.M., Lagomarsino, V.N., Pearse, R.V., Rice, H.C., Negri, J.M., He, A., et al. (2017). Cell-type Dependent Alzheimer's Disease Phenotypes: Probing the Biology of Selective Neuronal Vulnerability. Stem Cell Rep. *9*, 1868–1884.

Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457.

Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat. Biotechnol. *37*, 773–782.

Nutt, J.G., Carter, J.H., and Sexton, G.J. (2004). The dopamine transporter: Importance in Parkinson's disease. Ann. Neurol. *55*, 766–773.

Oldham, M.C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., and Geschwind, D.H. (2008). Functional organization of the transcriptome in human brain. Nat. Neurosci. *11*, 1271–1282.

Polymeropoulos, M.H. (1997). Mutation in the -Synuclein Gene Identified in Families with Parkinson's Disease. Science *276*, 2045–2047.

Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007). g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. *35*, W193–W200.

Reynolds, R.H., Botía, J., Nalls, M.A., International Parkinson's Disease Genomics Consortium (IPDGC), System Genomics of Parkinson's Disease (SGPD), Hardy, J., Gagliano Taliun, S.A., and Ryten, M. (2019). Moving beyond neurons: the role of cell type-specific gene regulation in Parkinson's disease heritability. NPJ Park. Dis. *5*, 6.

Sawada, H., Oeda, T., and Yamamoto, K. (2013). Catecholamines and Neurodegeneration in Parkinson's Disease—From Diagnostic Marker to Aggregations of α-Synuclein. Diagnostics *3*, 210–221.

Skene, N.G., and Grant, S.G.N. (2016). Identification of Vulnerable Cell Types in Major Brain Disorders Using Single Cell Transcriptomes and Expression Weighted Cell Type Enrichment. Front. Neurosci. *10*.

The GTEx Consortium, Ardlie, K.G., Deluca, D.S., Segre, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science *348*, 648–660.

Truong, P.H., Ciccotosto, G.D., Merson, T.D., Spoerri, L., Chuei, M.J., Ayers, M., Xing, Y.L., Emery, B., and Cappai, R. (2019). Amyloid precursor protein and amyloid precursor-like protein 2 have distinct roles in modulating myelination, demyelination, and remyelination of axons. Glia *67*, 525–538.

Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., and Yuan, G.-C. (2019). Accurate estimation of cell-type composition from gene expression data. Nat. Commun. *10*.

UK Brain Expression Consortium, North American Brain Expression Consortium, Ramasamy, A., Trabzuni, D., Guelfi, S., Varghese, V., Smith, C., Walker, R., De, T., Coin, L., et al. (2014). Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat. Neurosci. *17*, 1418–1428.

Wakabayashi, K., Hayashi, S., Yoshimoto, M., Kudo, H., and Takahashi, H. (2000). NACP/α-synuclein-positive filamentous inclusions in astrocytes and oligodendrocytes of Parkinson's disease brains. Acta Neuropathol. (Berl.) *99*, 14–20.

Wang, X., Park, J., Susztak, K., Zhang, N.R., and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat. Commun. *10*.

Wolfe, C.J., Kohane, I.S., and Butte, A.J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics *6*, 227.

Zarranz, J.J., Alegre, J., Gómez-Esteban, J.C., Lezcano, E., Ros, R., Ampuero, I., Vidal, L., Hoenicka, J., Rodriguez, O., Atarés, B., et al. (2004). The new mutation, E46K, of α-synuclein causes parkinson and Lewy body dementia: New α-Synuclein Gene Mutation. Ann. Neurol. *55*, 164–173.

Zeisel, A., Munoz-Manchado, A.B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.

Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Häring, M., Braun, E., Borm, L.E., La Manno, G., et al. (2018). Molecular Architecture of the Mouse Nervous System. Cell *174*, 999-1014.e22.

Zhang, Y., Sloan, S.A., Clarke, L.E., Caneda, C., Plaza, C.A., Blumenthal, P.D., Vogel, H., Steinberg, G.K., Edwards, M.S.B., Li, G., et al. (2016). Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. Neuron *89*, 37–53.

Zhu, L., Lei, J., Devlin, B., and Roeder, K. (2018). A unified statistical framework for single cell and bulk RNA sequencing data. Ann. Appl. Stat. *12*, 609–632.
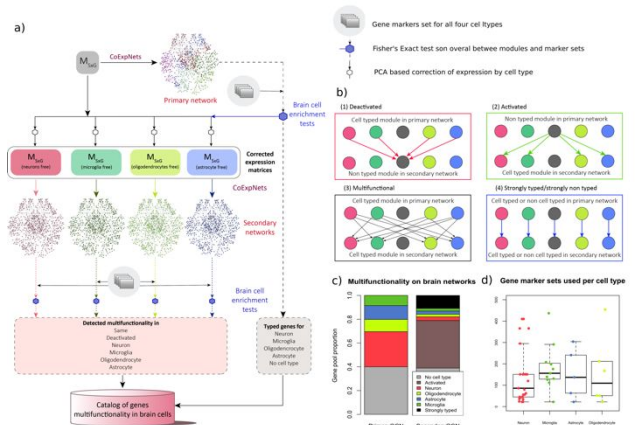
Figure 1. a) GMSCA generates a list of triplets <gene, cell type, function> for the genes included in the initial gene expression profiling matrix, MSxG. First, a WGCNA+k-means co-expression network called the primary network is created. Its modules are then tested (Fisher´s Exact Test, FET) for enrichment of cell type markers, in this case with brain cell type marker sets for neurons (in red), microglia (in green), oligodendrocytes (in light green) and astrocytes (in blue). Those modules with a clear signal (FET P < 0.05 on just a single cell type) are selected and their corresponding cell signal removed (see Methods) from expression to generate a new M'SxG. GMSCA creates a new co-expression network (these are called secondary) for each expression matrix and annotates their modules in the same way. Cell-type enriched modules in both primary and secondary networks generate as many triplets (gene, cell type, function) as genes in the module. b) Any gene found in a cell type enriched module is tagged by GMSCA as "typed". When a gene in a primary co-expression network shifts from a cell-type enriched module to a non-cell-type enriched module in the secondary co-expression network, we say the gene is deactivated (red arrows). When it goes from a non-cell-type enriched module to a cell-type enriched module, we say the gene is activated (green arrows). A gene is multifunctional when it goes from a cell-type enriched module to a module with a different cell type enrichment (black arrows). A gene is strongly typed when it goes from a cell-type enriched module to another module with the same cell-type enrichment. It is strongly non-typed when the primary module and secondary module are both non-cell-type enriched. c) An average primary co-expression network tags 37% of genes as non-typed, another 30% as neuronal, 10% as oligodendrocytic, 13% as astrocytic and 8% as microglial. In a secondary network, 37% of non-typed genes become typed and 42.5% of typed genes become deactivated. Also, as 9.1% of the genes are tagged as pleiotropic in a single network, we can gain up to 36.4% annotations of pleiotropy with all four secondary networks GMSCA creates in this case. c) Number and size of gene markers set used by GMSCA for each cell type, in this paper.
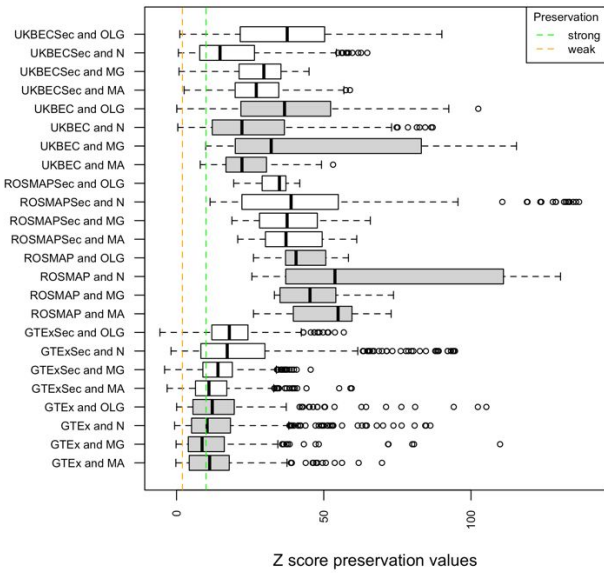

Figure 2. distribution of preservation values for all cell-type enriched modules as detected by GMSCA, for primary (in grey) and secondary (in white) networks. Each box plot corresponds to all modules within a family found to be enriched for the indicated cell type (MA for mature astrocytes, MG for microglia, N for neuron and OLG for oligodendrocytes). Families ending with "Sec" refer to secondary network modules. The vertical green dashed line marks the strong preservation limit (modules are easily replicable in other brain tissues and therefore highly reliable). The vertical orange dashed line marks the weak preservation limit (modules show signs or preservation and therefore some evidence of reliability). All preservation tests are performed within each network family. 98.1% of the preservervation tests sustain network reliability in UKBEC and 95.5% in GTEx. All tests yield strong preservation within the ROSMAP family.
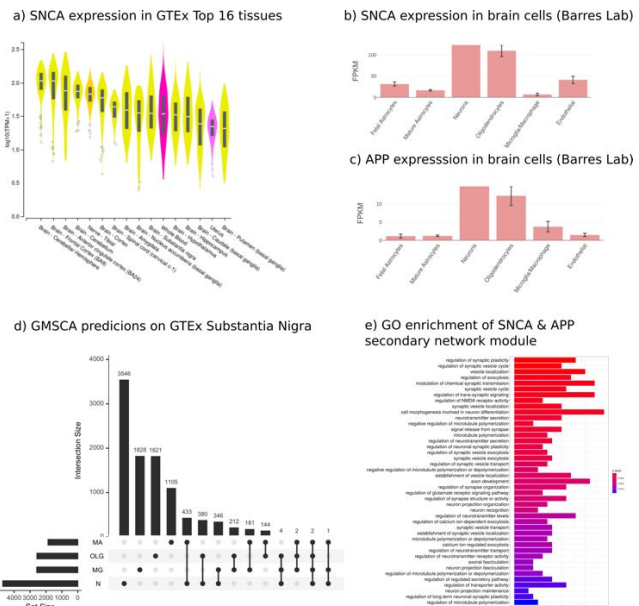

Figure 3. a) SNCA is predominantly expressed in brain, as the violin plots show. All 13 brain tissues are within the top 16 GTEx tissues expressing

SNCA. b) Immunopanning data on brain cell expression from Barres Lab shows SNCA predominantly expresses in neurons and oligodendrocytes. GMSCA tags SNCA as neuronal and oligodendrocytic in the substantia nigra, putamen and hippocampus. c) APP, genetically linked to Alzheimer, is another interesting gene found in the same module as SNCA. Barres Lab's data confirms it is predominantly expressed in neurons and oligodendrocytes. GMSCA tags it as neuronal and oligodendrocytic. d) UpSet plot (UpSet R package) on the genes from the predictions by GMSCA on the GTEx substantia nigra samples. SNCA and APP are found at the intersection between N (neurons) and OLG (oligodendrocytes) with 378 genes more.