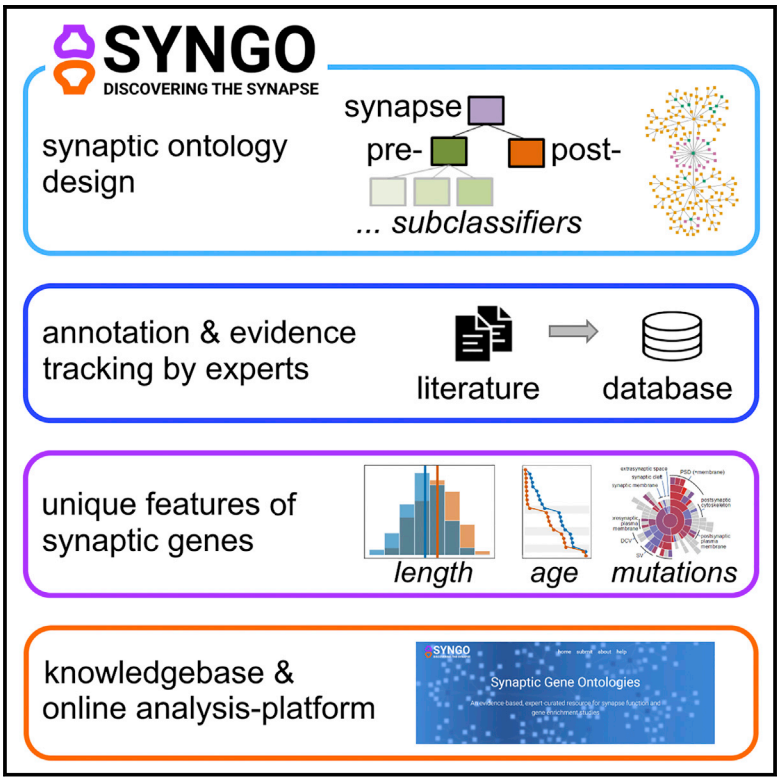
NeuroResource

Neuron

SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse

Graphical Abstract



Highlights

- SynGO is a public knowledge base and online analysis platform for synapse research
- SynGO has annotated 1,112 genes with synaptic localization and/or function
- SynGO genes are exceptionally large, well conserved, and intolerant to mutations
- SynGO genes are strongly enriched among genes associated with brain disorders

Authors

Frank Koopmans, Pim van Nierop, Maria Andres-Alonso, ..., Paul D. Thomas, August B. Smit, Matthijs Verhage

Correspondence

guus.smit@cncr.vu.nl (A.B.S.), matthijs@cncr.vu.nl (M.V.)

In Brief

The SynGO consortium presents a framework to annotate synaptic protein locations and functions and annotations for 1,112 synaptic genes based on published experimental evidence. SynGO reports exceptional features and disease associations for synaptic genes and provides an online data analysis platform.

Koopmans et al., 2019, Neuron *103*, 217–234
July 17, 2019
https://doi.org/10.1016/j.neuron.2019.05.002

CellPress

Neuron
NeuroResource

SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse

Frank Koopmans,[1,2] Pim van Nierop,[2] Maria Andres-Alonso,[28,29] Andrea Byrnes,[31] Tony Cijsouw,[18] Marcelo P. Coba,[3] L. Niels Cornelisse,[1] Ryan J. Farrell,[35] Hana L. Goldschmidt,[23] Daniel P. Howrigan,[31] Natasha K. Hussain,[23,24] Cordelia Imig,[19] Arthur P.H. de Jong,[26] Hwajin Jung,[27] Mahdokht Kohansalnodehi,[25] Barbara Kramarz,[4] Noa Lipstein,[19] Ruth C. Lovering,[4] Harold MacGillavry,[22] Vittoria Mariano,[14,15] Huaiyu Mi,[5] Momchil Ninov,[25] David Osumi-Sutherland,[6] Rainer Pielot,[21] Karl-Heinz Smalla,[21] Haiming Tang,[5] Katherine Tashman,[31] Ruud F.G. Toonen,[1] Chiara Verpelli,[36] Rita Reig-Viader,[16,17] Kyoko Watanabe,[33,34] Jan van Weering,[1] Tilmann Achsel,[14,15] Ghazaleh Ashrafi,[35] Nimra Asi,[31] Tyler C. Brown,[31] Pietro De Camilli,[7] Marc Feuermann,[8] Rebecca E. Foulger,[4] Pascale Gaudet,[8] Anoushka Joglekar,[13] Alexandros Kanellopoulos,[14,15] Robert Malenka,[9] Roger A. Nicoll,[10] Camila Pulido,[35] Jaime de Juan-Sanz,[35]

(Author list continued on next page)

[1]Department of Functional Genomics, CNCR, VU University and UMC Amsterdam, 1081 HV Amsterdam, the Netherlands
[2]Department of Molecular and Cellular Neurobiology, CNCR, VU University and UMC Amsterdam, 1081 HV Amsterdam, the Netherlands
[3]Zilkha Neurogenetic Institute and Department of Psychiatry and Behavioral Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA 90333, USA
[4]Functional Gene Annotation, Institute of Cardiovascular Science, UCL, London WC1E 6JF, UK
[5]Division of Bioinformatics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA
[6]European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK
[7]Departments of Neuroscience and Cell Biology, HHMI, Kavli Institute for Neuroscience, Yale University School of Medicine, 295 Congress Avenue, New Haven, CT 06510, USA
[8]SIB Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland
[9]Nancy Pritzker Laboratory, Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305, USA
[10]Departments of Cellular and Molecular Pharmacology and Physiology, University of California, San Francisco, San Francisco, CA 94158, USA
[11]Department of Neuroscience, Genentech, South San Francisco, CA 94080, USA
[12]Department of Molecular and Cellular Physiology, Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305, USA
[13]Brain and Mind Research Institute and Center for Neurogenetics, Weill Cornell Medicine, New York, NY, USA
[14]Department of Fundamental Neurosciences, University of Lausanne, 1006 Lausanne, Switzerland
[15]Department of Biomedicine and Prevention, University of Rome Tor Vergata, 00133 Rome, Italy
[16]Molecular Physiology of the Synapse Laboratory, Biomedical Research Institute Sant Pau, 08025 Barcelona, Spain
[17]Universitat Autònoma de Barcelona, 08193 Bellaterra, Cerdanyola del Vallès, Spain
[18]Department of Neuroscience, Tufts University School of Medicine, Boston, MA 02111, USA

(Affiliations continued on next page)

SUMMARY

Synapses are fundamental information-processing units of the brain, and synaptic dysregulation is central to many brain disorders ("synaptopathies"). However, systematic annotation of synaptic genes and ontology of synaptic processes are currently lacking. We established SynGO, an interactive knowledge base that accumulates available research about synapse biology using Gene Ontology (GO) annotations to novel ontology terms: 87 synaptic locations and 179 synaptic processes. SynGO annotations are exclusively based on published, expert-curated evidence. Using 2,922 annotations for 1,112 genes, we show that synaptic genes are exceptionally well conserved and less tolerant to mutations than other genes. Many SynGO terms are significantly overrepresented among gene variations associated with intelligence, educational attainment, ADHD, autism, and bipolar disorder and among *de novo* variants associated with neurodevelopmental disorders, including schizophrenia. SynGO is a public, universal reference for synapse research and an online analysis platform for interpretation of large-scale -omics data (https://syngoportal.org and http://geneontology.org).

INTRODUCTION

Synapses are information-processing units of the brain that provide the foundation for higher-level information integration in dendrites, neurons, and networks. Use-dependent changes in synaptic strength (synaptic plasticity) are firmly established as main underlying principles of cognitive processes, such as memory formation and retrieval, perception, sensory processing,

Morgan Sheng,[11] Thomas C. Südhof,[12] Hagen U. Tilgner,[13] Claudia Bagni,[14,15] Àlex Bayés,[16,17] Thomas Biederer,[18] Nils Brose,[19] John Jia En Chua,[20] Daniela C. Dieterich,[21] Eckart D. Gundelfinger,[21] Casper Hoogenraad,[22] Richard L. Huganir,[23,24] Reinhard Jahn,[25] Pascal S. Kaeser,[26] Eunjoon Kim,[27] Michael R. Kreutz,[28,29] Peter S. McPherson,[30] Ben M. Neale,[31] Vincent O'Connor,[32] Danielle Posthuma,[33,34] Timothy A. Ryan,[35] Carlo Sala,[36] Guoping Feng,[31] Steven E. Hyman,[31] Paul D. Thomas,[5] August B. Smit,[2,37,*] and Matthijs Verhage[1,37,38,*]

[19]Department of Molecular Neurobiology, Max Planck Institute of Experimental Medicine, 37075 Göttingen, Germany
[20]Department of Physiology, Yong Loo Lin School of Medicine and Neurobiology/Ageing Program, Life Sciences Institute, National University of Singapore and Institute of Molecular and Cell Biology, A*STAR, Singapore, Singapore
[21]Leibniz Institute for Neurobiology, CBBS and Medical Faculty, Otto von Guericke University, 39120 Magdeburg, Germany
[22]Cell Biology, Department of Biology, Faculty of Science, Utrecht University, 3584 CH Utrecht, the Netherlands
[23]Solomon H. Snyder Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[24]Kavli Neuroscience Discovery Institute, Johns Hopkins University, Baltimore, MD 21205, USA
[25]Department of Neurobiology, Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany
[26]Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA
[27]Center for Synaptic Brain Dysfunctions, IBS, and Department of Biological Sciences, KAIST, Daejeon 34141, South Korea
[28]RG Neuroplasticity, Leibniz Institute for Neurobiology, 39118 Magdeburg, Germany
[29]Leibniz Group "Dendritic Organelles and Synaptic Function," ZMNH, University MC, Hamburg, 20251, Germany
[30]Department of Neurology and Neurosurgery, Montreal Neurological Institute, McGill University, Montreal, QC H3A 2B4, Canada
[31]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[32]Biological Sciences, University of Southampton, Southampton SO17 1BJ, UK
[33]Department Complex Trait Genetics, CNCR, Neuroscience Campus Amsterdam, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, the Netherlands
[34]Department of Clinical Genetics, UMC Amsterdam, 1081 HV Amsterdam, the Netherlands
[35]Department of Biochemistry, Weill Cornell Medicine, New York, NY 10065, USA
[36]CNR Neuroscience Institute Milan and Department of Biotechnology and Translational Medicine, University of Milan, 20129 Milan, Italy
[37]Senior author
[38]Lead Contact
*Correspondence: guus.smit@cncr.vu.nl (A.B.S.), matthijs@cncr.vu.nl (M.V.)
https://doi.org/10.1016/j.neuron.2019.05.002

attention, associative learning, and decision making (Abdou et al., 2018; Groschner et al., 2018; Kandel, 2001; Petersen and Crochet, 2013; Ripollés et al., 2018). Based on both genetic and neurobiological evidence, synaptic dysregulation is widely recognized as an important component of risk in many brain disorders (termed "synaptopathies"; Boda et al., 2010; Bourgeron, 2015; Grant, 2012; Monday and Castillo, 2017), such as autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD), schizophrenia, Alzheimer's disease, and Parkinson's disease (Arnsten et al., 2012; Bourgeron, 2015; De Rubeis et al., 2014; Fromer et al., 2014; Heutink and Verhage, 2012; Hong et al., 2016; Selkoe, 2002; Soukup et al., 2018; Spires-Jones and Hyman, 2014; Südhof, 2008). Despite these intense investigations and a variety of research efforts focused on synaptic proteins and their subcellular organization and specific functions, only sparse efforts have been made to establish systematic resources for synapse biology in health and disease. In particular, the ontology of synaptic processes has been poorly defined, which has precluded systematic annotation of synaptic genes.

Gene Ontology (GO) is the most widely used resource for gene function annotations. The resource has two components: (1) the ontology, a framework of definitions called "terms" to describe gene functions and locations and their relationships, and (2) GO annotations, statements linking genes to specific terms (Ashburner et al., 2000; The Gene Ontology Consortium, 2019). The ontology is divided into three aspects: (1) molecular function (MF), defining the molecular activities of gene products; (2) cellular component (CC), defining where they are active; and (3) biological process, defining the processes they carry out. Relationships between CC terms and between biological process (BP) terms generally specify how smaller structures are parts of larger ones. The accuracy of GO annotations critically depends on how well experimental evidence supports the annotations.

Using existing annotations to synaptic GO terms and synaptic gene sets, several studies have shown that synaptic genes (i.e., genes encoding synaptic proteins) are significantly enriched in genetic variations associated with several brain traits (Savage et al., 2018; Zwir et al., 2018) and have produced valuable leads to understand the role of synapse function and dysfunction in these traits (De Rubeis et al., 2014; Fromer et al., 2014; Mattheisen et al., 2015; Pedroso et al., 2012; Thapar et al., 2016). However, the lack of systematic annotation of synaptic genes also limits progress. Available resources, including GO, have limited representations of synapse biology and lack a comprehensive ontology of synaptic processes and locations. Existing resources are biased by uneven and patchy coverage of different aspects of synapse biology. Moreover, existing resources include data that have not been curated by experts, and a large fraction of the data has been aggregated in an unsupervised manner; e.g., by automated text mining or large-scale experiments that result in high rates of false positives, such as bulk proteomics and yeast two-hybrid studies. Thresholds for inclusion are not systematically defined and are typically low. Together, these shortcomings limit the effect of such resources and may engender incorrect conclusions.

To overcome these limitations, we established SynGO, a partnership between the GO Consortium and 15 synapse expert laboratories in Europe, North America, and Asia, for systematic annotation of synaptic proteins. SynGO experts have developed an extensive ontology to represent synaptic locations (87 terms) and processes (179 terms) and generated almost 3,000 annotations of synaptic genes to these terms based on a novel comprehensive evidence-tracking system that classifies evidence using only published datasets. Using SynGO, we observed that synaptic genes are exceptionally well conserved, highly intolerant to mutations, and associated with many brain traits, such as IQ and educational attainment, and brain disorders, such as ASD, ADHD, and bipolar disorder. SynGO provides a unique, publicly accessible knowledge base (https://syngoportal.org) as a universal reference for synapse research and education and for enrichment studies of genomic associations, mRNA profiles, and proteomics data.

RESULTS

SynGO Ontologies Provide Comprehensive Frameworks for Synaptic Gene Annotation

To systematically annotate synaptic genes, we designed a generic synapse model as a conceptual starting point, defining locations at the synapse and processes related to the synapse, and refined this model iteratively until consensus was reached among expert laboratories worldwide (Figure 1). Subsequently, we created GO terms for CCs and BPs for synapses and defined their relationships. At the top level of the CC hierarchy (Figure 2A), synaptic proteins can be described as localized to the presynapse, postsynapse, synaptic cleft, extra-synaptic space, and synaptic membranes (the latter term is used when no distinction is possible between pre- and postsynaptic membranes). From these high-level terms, up to 4 additional hierarchical levels were defined for pre- or postsynaptic cytosol or membrane or organelles within these compartments. The SynGO CC ontology adds substantial precision to the pre-existing GO ontology that contained 13 terms directly connected to the central "synapse" term (and 19 additional terms). SynGO maintained two of these 13 terms (Figure 2A, green symbols) and excluded 11 (Figure 2A, purple symbols). Some of the GO terms were replaced by similar but more precise terms, and others were replaced with more specific terms further down in the hierarchical SynGO ontology. Altogether, 142 SynGO CC ontology terms were designed for accurate annotation of synaptic localizations (Table S2). To visualize this elaborate ontology hierarchy and provide a standardized visualization of SynGO annotations, all CC terms populated with gene annotations in SynGO 1.0 (92 of 142 terms) were plotted in a circular fashion, with the highest hierarchical term (synapse) in the center and each layer of subclasses in outward concentric rings (Figure 2C; see Table S2 for all term names). SynGO did not define mitochondria as part of a specific synaptic CC because mitochondrial proteins are already well annotated (Calvo et al., 2016; Smith and Robinson, 2019).

BP terms for synaptic processes and their relationships were also defined consistently with existing GO terms, with pre- and postsynaptic processes, synaptic organization, synaptic signaling, axonal and dendritic transport, and metabolism as main terms with up to 5 levels of subclasses (Figure 2B). In total, the BP ontology features 256 terms of which 212 are new. 192 of these BP ontology terms were populated with gene annotations in SynGO 1.0 and visualized as a sunburst plot (Figure 2D, analogous to Figure 2B; see Table S2 for all term names). Hence, these novel CC and BP ontologies provide substantial innovation and increased precision for the ontology of the synapse. Together, these ontologies provide a comprehensive structure for systematic annotation of synaptic genes.

SynGO Is Based on Expert Annotation and Systematic Evidence Tracking

Currently available synaptic protein lists contain many unsupervised inclusions, in particular from large-scale, automated experiments expected to have substantial false positive rates. SynGO established a systematic evidence tracking protocol and annotation by synapse experts only, based exclusively on published experimental data (PubMed). The SynGO workflow (Figure S1) was implemented in a web interface and used by the experts to annotate synaptic genes. To systematically track evidence, classifications were designed for the model systems used (Figure S2). For synaptic localization (CC), microscopy and biochemical studies were defined as the main experimental classes, each with several subclasses. For functional studies, experimental classes were defined based on perturbation type and methodology (assay) used to detect the consequences, again with several subclasses (Figure S2). These classifications were made coherent with the evidence and conclusions ontology (ECO) (Giglio et al., 2019), and new ECO terms were defined. Together, these three dimensions of evidence—(1) model system and/or preparation, (2) experimental perturbation, and (3) assay—provide a systematic, coherent, and detailed definition of the evidence to annotate synaptic genes.

Annotations completed by expert laboratories first passed through a quality control pipeline by the SynGO support team (Figure S1) and were then added either directly to the SynGO database (https://syngoportal.org) or returned to the expert laboratories for further editing. These annotations were also deposited in the GO annotation repository (http://geneontology.org) as Gene Ontology causal activity models (GO-CAM; The Gene Ontology Consortium, 2019). Together, this evidence tracking system, including detailed references to the evidence (PubMed ID [PMID], figure, and panel), provides an excellent framework for comprehensive transparent annotation of synaptic genes.

SynGO 1.0 Provides 2,922 Expert-Curated Annotations on 1,112 Synaptic Genes

Using the three dimensions of evidence tracking (model system and/or preparation, experimental perturbation, and assay), 2,922 expert-curated annotations were generated using cumulative candidate synaptic gene lists from published (Lips et al., 2012; Ruano et al., 2010) and unpublished data resources (the European Union [EU]-funded projects: European consortium on synaptic protein networks [EUROSPIN] and Synapse and Systems Biology Consortium [SYNSYS]; see Acknowledgments), proteomics data, and specific input from expert laboratories. The annotations were subjected to quality control and, typically after iterative optimization, deposited in the SynGO

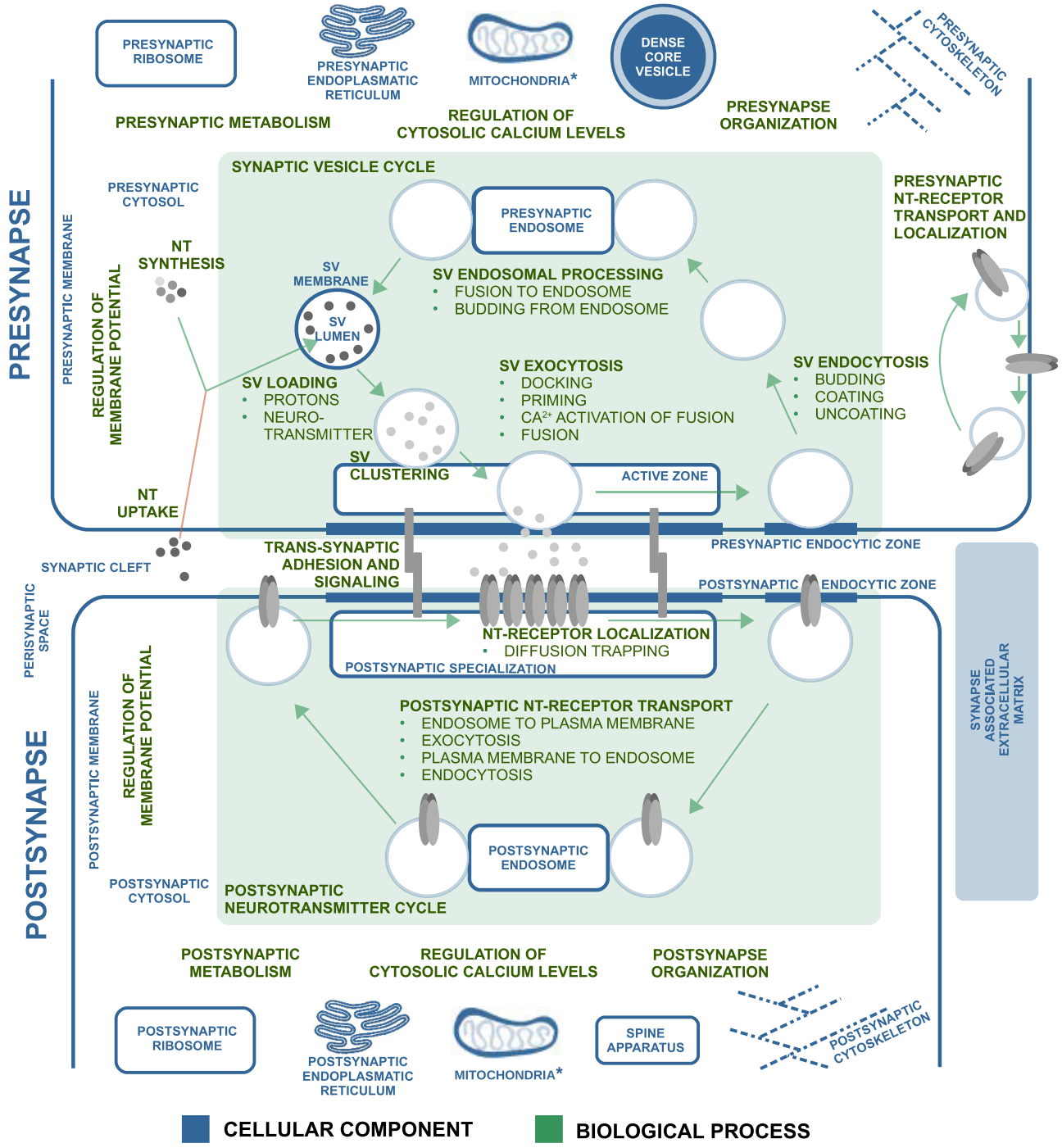


Figure 1. Conceptual Framework of Synapse Ontology in SynGO

The top-level CC (location, shown in green) and biological process (function, shown in blue) terms are depicted in a schematic representation of a synapse. For the full set of ontology terms, which also include all subclassifiers that further specialize the terms shown here, see Figure 2 and Table S2. The mitochondrion is depicted for completeness but is not part of SynGO ontology (see text).

database and the central GO knowledge base (The Gene Ontology Consortium, 2019; Figure S1). We found compelling evidence for 1,112 unique synaptic genes. These were admitted to the SynGO 1.0 knowledge base. The full list of 1,112 genes can be downloaded from https://syngoportal.org. For most genes, both subcellular localization (CC) and BP evidence was found (60%; Figure S3A); for the remaining 40%, evidence was lacking for either CC or BP, and only one term was included.

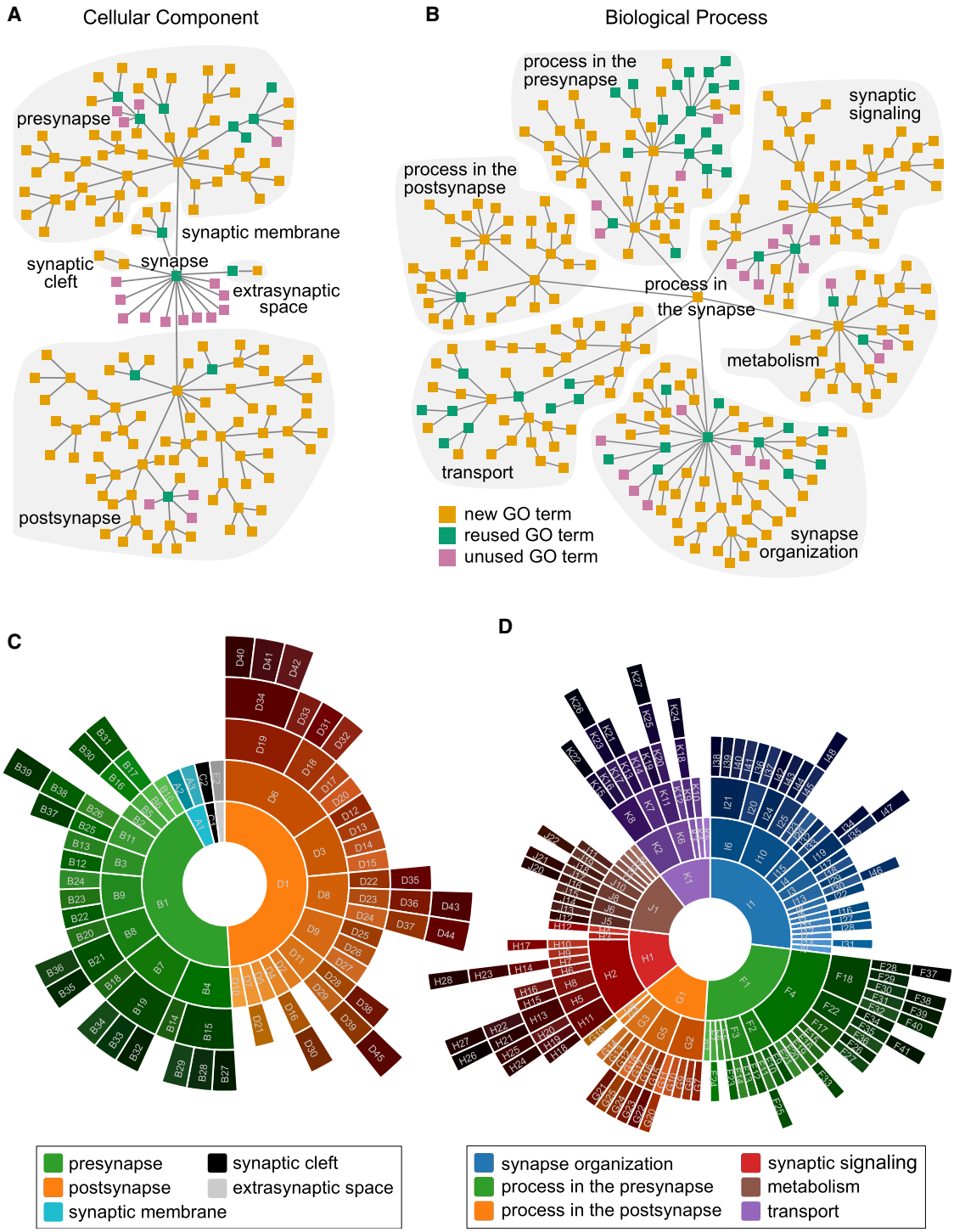


Figure 2. Increased Resolution in Synaptic Ontology Terms

(A and B) Comparison between new terms in SynGO (orange) and pre-existing synapse ontology terms in GO (green and purple) for (A) CCs (locations) and (B) biological processes (BPs; functions). SynGO adds resolution by creating increasingly detailed terms in a consistent system for CC (129 new terms) and BP (212 new terms). Some existing GO terms identical to SynGO ontologies were re-used (green nodes; 13 for CC and 44 for BP), and some existing GO synapse-related terms that did not overlap with the SynGO ontologies were discarded or replaced (purple nodes; 18 for CC and 22 for BP). Table S1 contains a complete list of pre-existing GO terms indicated in green and purple.

(legend continued on next page)

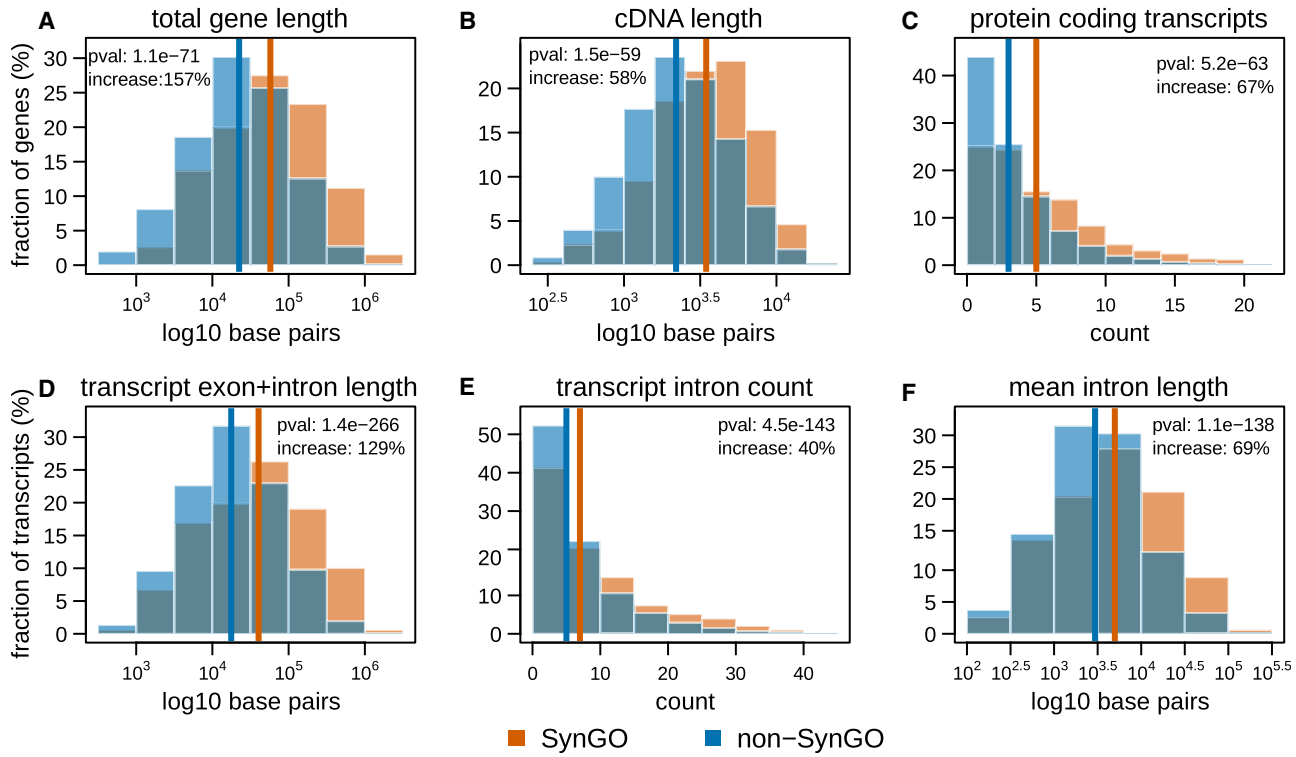


Figure 3. Gene Features Compared between Synaptic Genes and the Rest of the Genome

(A–F) Total gene length (A), cDNA length (B), number of known protein-coding splice variants (C), total length of protein-coding transcripts (D), number of introns in protein-coding transcripts (E), and mean length of introns in protein-coding transcripts (F). Vertical lines indicate median values for respective data distributions, which were also used to compute the percentage increase for synaptic genes. Two-sample Student's t test was applied to log-transformed data to confirm that overall distributions were significantly distinct, and a Wilcoxon rank-sum test was used for the count data in (C) and (E); "pval" in each panel denotes the resulting p values. Analogous comparison between SynGO and brain-enriched or brain most-expressed genes is shown in Figure S4.

A core set of synaptic proteins was annotated to 3 or more CC or BP terms (Figure S3B). Most evidence was obtained from studies of rodent species (Figure S3C) of either intact tissue or cultured neurons (Figure S3D). Microscopy and biochemical fractionation were the two main assay types used to make CC annotations, whereas BP annotations were based on a larger array of assay types assessing synaptic function (Figure S3E). Together, these 2,922 expert-curated annotations on 1,112 synaptic genes, with a core set annotated to 3 or more CC or BP terms, provide an excellent annotation collection for descriptive studies, functional analyses of synaptic genes, and gene enrichment studies.

The Structure of Synaptic Genes Is Very Different from Other Genes

As a first descriptive analysis, we compared basic structural features of SynGO-annotated synaptic genes with other genes. Human gene features were extracted from BioMart (GRCh38.p12) and Ensembl. Interestingly, synaptic genes were found to be different from other (non-SynGO) genes in many respects. Synaptic genes were, on average, more than twice as long as other genes (2.6-fold of non-SynGO genes; Figure 3A), with 1.6-fold longer cDNA (Figure 3B). The number of known protein-coding transcripts was 1.7-fold higher (Figure 3C), and the sequence of introns and exons (immature transcript length) for protein-coding transcripts was more than 2-fold longer (Figure 3D). Protein-coding transcripts for synaptic genes also contained 1.4-fold more introns (Figure 3E), and these were 1.7-fold longer (Figure 3F).

To compare SynGO genes with other brain-expressed genes, we defined two control gene sets: (1) brain-enriched genes (6,600 genes with maximal expression difference between brain and other tissues; Ganna et al., 2016) and (2) "top N" genes most highly expressed in the brain, with N equal to the number of unique genes in the SynGO set (1,112). Differences between SynGO genes and control sets A and B were generally smaller

(C and D) SynGO ontology terms shown in (A) and (B) (in orange or green) that were populated with at least one gene annotation in SynGO v1.0 were visualized as "sunburst plots," an alternative representation of tree structures, for (C) CCs and (D) BPs. The top-level terms in these CC and BP ontology trees, "synapse" and "process in the synapse," respectively, are represented by a white circle in the center of the sunburst. Terms on the second level of the ontology term tree, previously highlighted in (A) and (B), are color-coded as indicated in the legend. Subclasses in outer circles are shown in progressively darker colors. Table S2 contains the complete list of SynGO ontology terms matching the sunburst plots.

in comparisons of gene size, introns, and cDNA length but still highly significant (Figures S4A–S4L). Finally, we tested the possibility that SynGO-annotated genes have a higher structural and topological complexity than other genes, especially more transmembrane regions (TMRs), and that this may explain the observed differences between SynGO genes and others. A TMR prediction algorithm (Krogh et al., 2001) indicated that SynGO-annotated genes indeed encode significantly more proteins with at least one TMR (35.2% versus 29.7% for the whole genome; $p = 6.1e{-}5$, using a two-sided Fisher's exact test). However, when comparing SynGO annotated proteins with all membrane proteins, SynGO proteins are still significantly different to a similar extent and in all aspects indicated in Figure 3 and Figures S4A–S4L (see Figures S4M–S4R).

We also investigated the complexity of isoform expression of synaptic genes in cerebellar neurons using recently published full-length RNA sequencing data (Gupta et al., 2018). Synaptic genes expressed a higher number of distinct isoforms compared with non-SynGO genes, per equal read counts, than non-synaptic genes (Figure S5).

We also analyzed the number of posttranslational modifications as important determinants of cell signaling by testing the number of experimentally verified modifications obtained from dbPTM (Huang et al., 2016) and UniProt (The UniProt Consortium, 2017) per protein and per amino acid (to correct for differences in average protein length; Figure S6). The incidence of all major modifications (phosphorylation, ubiquitination, acetylation, and S-nitrosylation) appears to be significantly higher in synaptic proteins compared with other proteins. However, these observations might emerge (in part) from the fact that synaptic proteins are more extensively studied.

Synaptic Genes Emerged Earlier in Evolution Than Other Genes, Primarily in Three Major Waves

We tested when SynGO genes emerged in evolution relative to other genes. We found that their evolution follows a pattern that differs substantially from the overall pattern for all human genes (Figure 4A). Specifically, SynGO genes evolved primarily in three "waves" of innovation during which modern-day synaptic genes were gained at a faster rate than other human genes. The first wave of emergence of SynGO genes was prior to the last eukaryotic common ancestor (LECA), approximately 1,800 million years ago (mya) (Kumar et al., 2017). Although the LECA was unicellular and obviously did not form synapses, it did possess cellular machinery that would later be co-opted for the synapse, such as vesicle trafficking, exocytosis, and signal reception. The second wave was prior to the last common ancestor of the eumetazoa (multicellular animals) and corresponds to the first appearance of the synapse. Among SynGO genes gained during this wave, we found strong enrichments for pre- and postsynaptic membranes and the postsynaptic density (Figure S7B) and weak enrichments for a few synaptic processes (Figure S7C). The third wave was prior to the last common ancestor of vertebrates, suggesting significant synaptic evolution in this period. SynGO genes gained during this last wave are again enriched for the postsynaptic density and now also the active zone and for more specific, largely regulatory processes (Figure S7E). By approximately 450 mya, about 95% of all SynGO genes were already in place, with very few additional synaptic genes appearing after that point. A similar trend, albeit with smaller differences, was observed when gene duplication events were not weighted (Figure S7). Figure 4B shows one of the few exceptions to this rule: the carnitine palmitoyltransferase gene family expanded via a gene duplication prior the last common ancestor of placental mammals, resulting in an additional, neuron-specific paralog found only in placental mammals (CPT1C), whereas other amniotes have only two paralogs (CPT1A and CPT1B) expressed primarily in other tissues. CPT1C is localized to the endoplasmic reticulum in neurons and has been shown to directly regulate the levels of α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptors in the postsynapse (Fadó et al., 2015). Overall, however, our analysis indicates that the synapse is highly conserved among modern vertebrates, as suggested previously (Emes et al., 2008), and that 95% of the human synaptic genes in SynGO 1.0 are shared among vertebrates. Because the invertebrates *C. elegans* and *D. melanogaster* are important model organisms in synapse biology, we also explored how many paralogs emerged in these invertebrates and how many in the vertebrate lineage (until humans) for any shared gene. For both invertebrates, we found that almost 30% of all genes have a 1:1 relationship with human genes (one paralog identified in each species; Figure 4C). For most genes, more than a single paralog is identified ("many"), with one a single paralog in *C. elegans* and *D. melanogaster* (many-to-1) or more than one in all species (many-to-many, Figure 4C). Interestingly for synaptic genes, we found fewer 1:1 relationships and more many-to-1 and many-to-many (Figure 4C). This indicates that synaptic genes underwent gene duplication at a higher rate than other genes after the vertebrate-invertebrate bifurcation.

Synaptic Gene Expression Is Enriched in the Brain

We predicted that the expression levels of SynGO genes are higher in the brain than in other tissues. To test this, we compared tissue-specific expression using different gene sets in GTEx v.7 (Battle et al., 2017). Brain enrichment was computed (STAR Methods) and plotted against the expression level of this transcript in the brain. As shown in Figure S8A, expression of SynGO genes is generally higher in the brain than in other tissues, although some SynGO genes are de-enriched in the brain. SynGO genes with high expression levels in the brain are, on average, enriched to a similar extent as those with lower expression levels in the brain (Figures S8A and S8B).

We compared brain expression enrichment for different SynGO CC and BP terms. Several terms within these ontologies, especially in BP, are predicted to be highly brain specific, e.g., *trans*-synaptic signaling, active zone assembly, or postsynaptic density organization, whereas others are expected to be similar to terms outside of the synapse and outside of the brain; e.g., phosphatase and kinase pathways. Indeed, analyses of individual SynGO terms in CC and BP ontologies revealed a large degree of heterogeneity among proteins annotated for different terms (Figures S8C and S8D). The pre- and postsynaptic plasma membranes and especially the postsynaptic density contain proteins that are highly

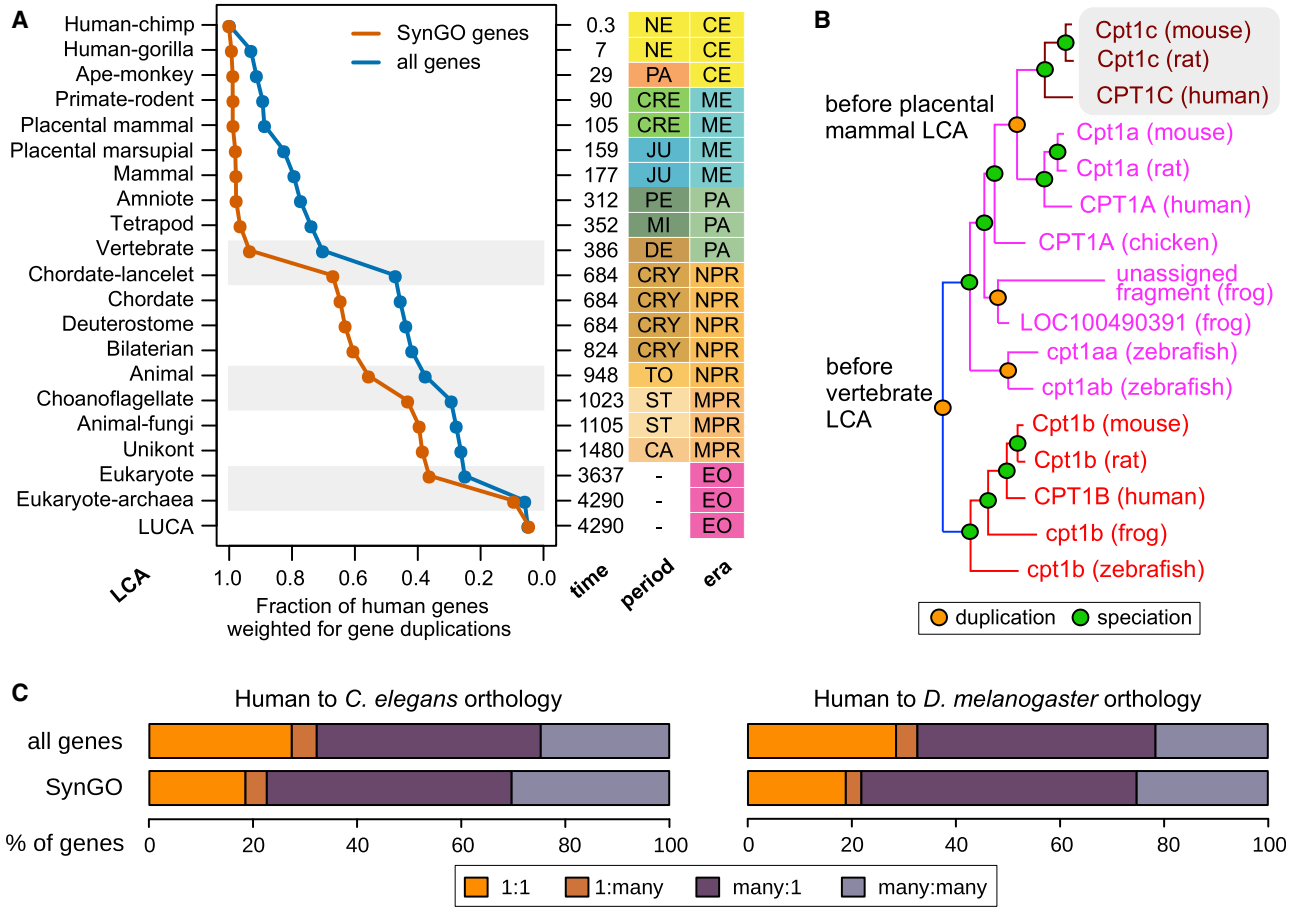


Figure 4. Synaptic Genes Are Exceptionally Well Conserved

(A) Cumulative distribution of synaptic genes (orange) and all human genes (blue) by gene age. Highlighted areas (gray) show periods of rapid gain of synaptic genes. Ages (time in million years ago) are obtained from dating of gene duplication events (relative to speciation events) in PANTHER gene trees (Mi et al., 2018). Clades are shown on the y axis, their names on the left, and estimated speciation times on the right. LCA, last common ancestor; LUCA, last universal common ancestor. Note that, unlike the phylostratigraphic approach (Domazet-Lošo et al., 2007), ages reflect not simply the oldest traceable gene age but explicitly consider gene duplication by adding a fractional count for each duplication event along the evolutionary path to a modern gene (see STAR Methods for details and abbreviations). This is critical because of the prevalence of gene duplication in the evolution of eukaryotic genomes.

(B) Evolution of the family of genes containing CPT1C (highlighted in gray), a synaptic gene annotated in SynGO. There are three tissue-specific isoforms in this family: CPT1A (liver), CPT1B (muscle), and CPT1C (brain). The latter is only found in placental mammals.

(C) Orthology relations between human genes and their counterparts in *C. elegans* and *D. melanogaster* were classified by the number of paralogs matching respective organisms. For example, the many-to-1 group contains all human genes that have undergone gene duplication from their ancestral gene whereas the given model organism has not.

significantly enriched in the brain (Figure S8C). Active zones and synaptic vesicles, but not dense core vesicles, also contain significantly enriched proteins (Figure S8C). For BP, a strong enrichment was observed for most major synaptic processes except metabolism and transport (Figure S8D). Taken together, these data indicate that expression of SynGO genes is higher in brain than in other tissues, especially for "synapse-specific" locations or functions.

Synaptic Proteins Are Exceptionally Intolerant to Mutations

The frequency of coding variants in the general population is an indication of the functional constraints on these genes. To test whether SynGO genes have the same loss-of-function mutation incidence as other genes, we used the probability of being loss-of-function-intolerant (pLI) obtained from the Exome Aggregation Consortium (ExAC; Karczewski et al., 2017). The pLI was compared between all SynGO genes and other genes. A major difference in loss-of-function intolerance was observed; SynGO genes are exceptionally intolerant to loss-of-function mutations relative to non-SynGO, brain-enriched, and top N most highly brain-expressed control genes (Figures 5A–5C). The distribution of high pLI values was similar among different CC and BP terms (Figures 5D and 5E). In the CC ontology, pLI scores were particularly high (mean value, $\geq$0.7) for postsynaptic density (PSD) and active-zone genes. Interestingly, synaptic vesicle and dense core vesicle annotated genes showed much lower pLI scores (mean value, $\leq$0.5). Taken together, these data indicate that

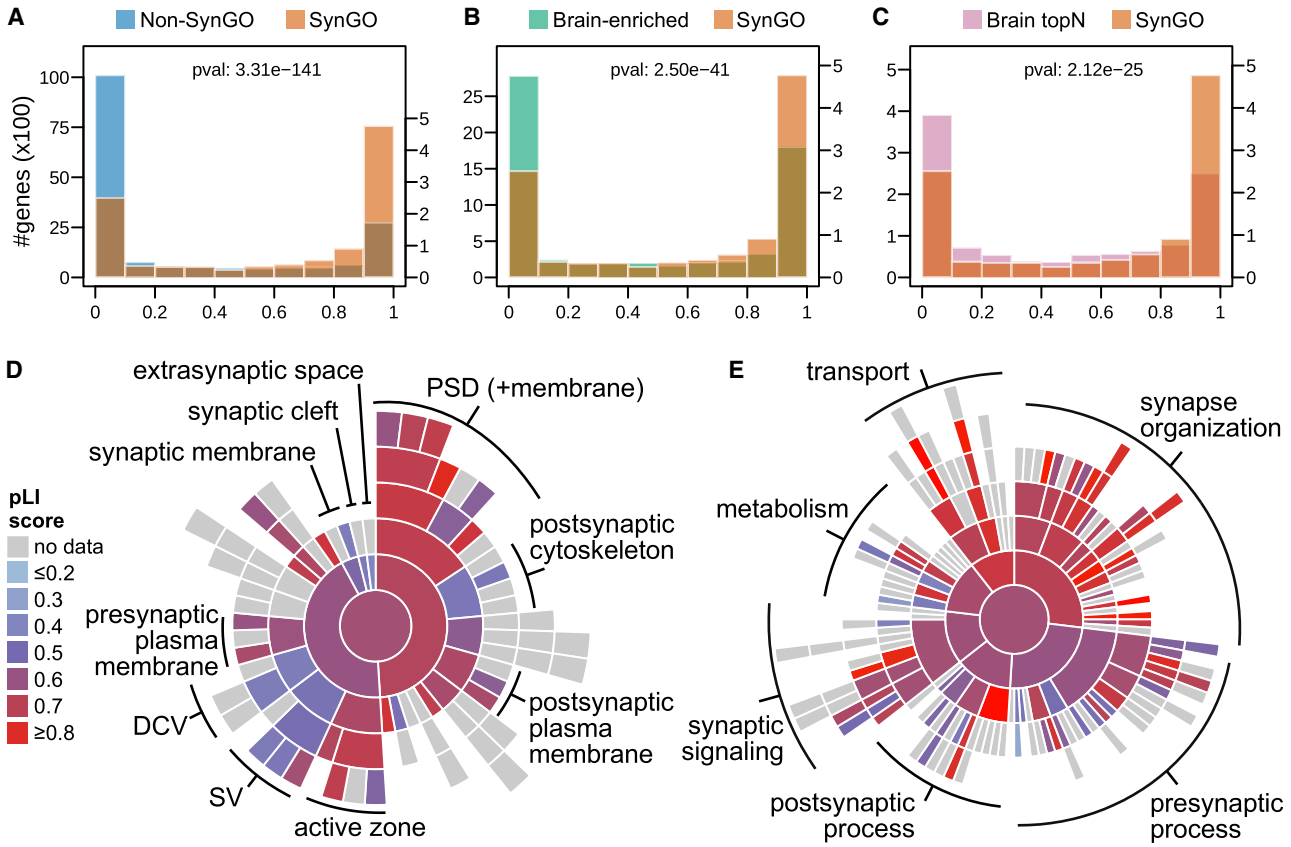


Figure 5. Gene pLI Scores, Indicating Probability of Intolerance to Loss-of-Function Mutation

(A–C) pLI scores compared between synaptic genes and (A) the rest of the genome, (B) brain enriched genes, and (C) the 1,112 genes most highly expressed in the brain. Two-sample Wilcoxon signed-rank test p values indicate that overall distributions are significantly different (denoted as "pval").

(D and E) Mean pLI scores for respective synaptic genes annotated against (D) SynGO CC terms and (E) BP terms are visualized in a sunburst plot for terms with at least 5 unique annotated genes with a pLI score. Terms where annotated genes are typically loss-of-function (LoF)-tolerant are shown in blue, whereas terms with mostly LoF-intolerant genes are shown in red. Note that the CC and BP sunburst plots are aligned with Figures 2C and 2D, respectively.

synaptic genes are exceptionally intolerant to loss-of-function mutations, suggesting that functional constraints and evolutionary selection pressure on synaptic genes are much stronger than for other genes.

Synaptic Proteins Annotated to Closely Related SynGO Terms Are More Likely to Interact

SynGO proteins annotated to the same ontology term or to closely related terms are predicted to often be in the same protein complexes or involved in the same process and are thus more likely to interact. This prediction was tested using protein-protein interaction data available through StringDB v.10.5 (Jeanquartier et al., 2015) using the "high confidence" interaction filter. Proteins reported to be in the same protein complexes were significantly overrepresented in synaptic genes annotated against the same CC term in SynGO (Figure S9A) and also for the same BP term (Figure S9B). Hence, synaptic proteins annotated for the same CC or BP term are much more likely to interact, and, vice versa, interacting synaptic proteins are much more likely to have the same localization or be part of a similar process.

Different Synaptic Preparations Contain Largely Overlapping Synaptic Protein Collections

SynGO enables analysis of existing, large-scale proteomics data from biochemical preparations enriched for synaptic components. We extracted data from 19 well-described and quantitative proteomics studies on 3 biochemical preparations enriched for synaptic components: (1) synaptosome fractions (7 studies), (2) PSD fractions (6 studies), and (3) active-zone or docked vesicle fractions (5 studies) (see STAR Methods for data sources). Synaptosome studies have identified between 894 and 3,331 proteins (Figure 6A). These protein collections contained between 17% and 39% of the SynGO CC-annotated proteins. Together, 80% of proteins with a SynGO CC annotation were detected in at least one of the synaptosome preparations. PSD analyses typically identified smaller numbers of components, up to 1,207 (Roy et al., 2018).

A consensus set of proteins identified in at least three proteomics datasets per compartment contains 2,621 unique proteins for the synaptosome, 791 for the PSD, and 88 for the active zone. The PSD components showed a large degree of overlap (90%) with the synaptosome consensus set, with only 76 proteins

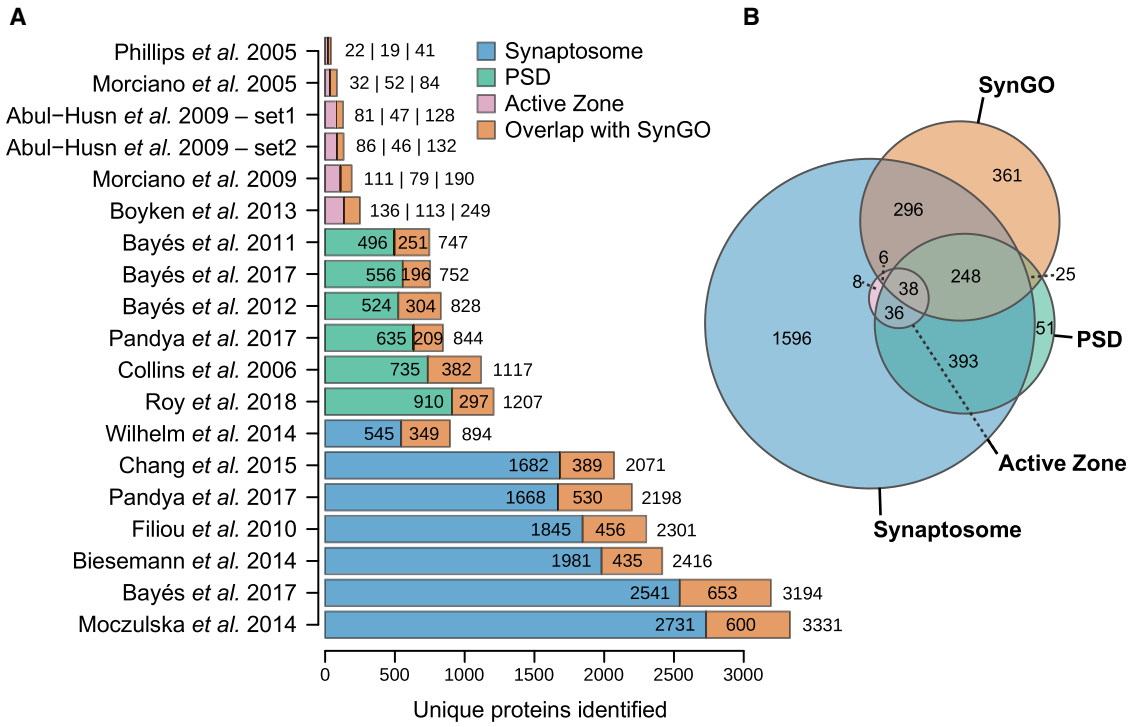


Figure 6. Representation of SynGO Proteins in Large-Scale Proteomic Analyses of Synaptic (Sub)Fractions

Proteins identified in a selection of published proteomics analyses of biochemically purified synaptic fractions (synaptosomes, postsynaptic densities [PSDs], and active zone) were analyzed for SynGO-annotated proteins.

(A) The number of unique proteins detected in the selected studies. Blue, synaptosomes; green, PSDs; pink, active zone; orange, subset of proteins that are CC-annotated in SynGO.

(B) Overlap among SynGO CC-annotated proteins (orange) and "consensus sets" for synaptosome (blue), PSD (green), or active zone (pink), defined as proteins identified in at least three datasets described in (A) (matching respective compartments).

Table S4 details the selected proteomics studies and their identified proteins.

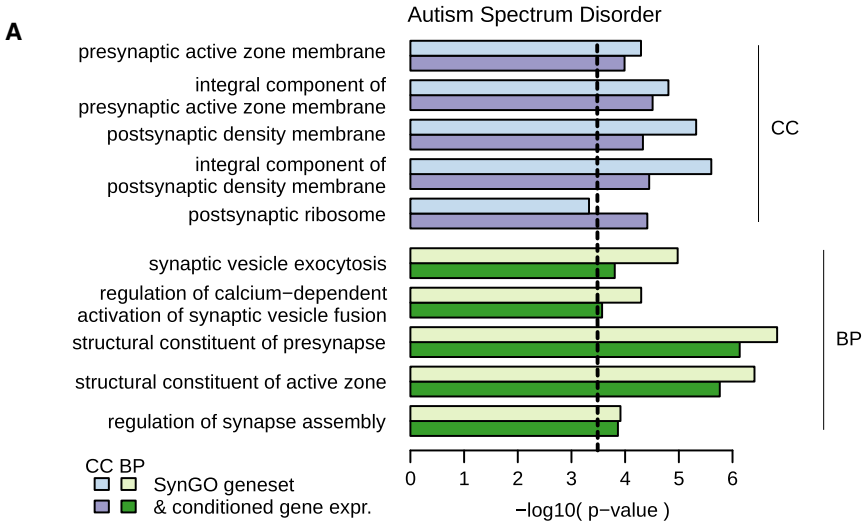
exclusively identified in the PSD consensus set (Figure 6B). 73% (1,906 proteins) of the synaptosome consensus set is not found in the PSD consensus set, 78% (2033 proteins) is not found in SynGO 1.0, and, in total, 61% (1,596 proteins) of the synaptosome consensus set was not found in either the PSD, active zone, or the SynGO database.

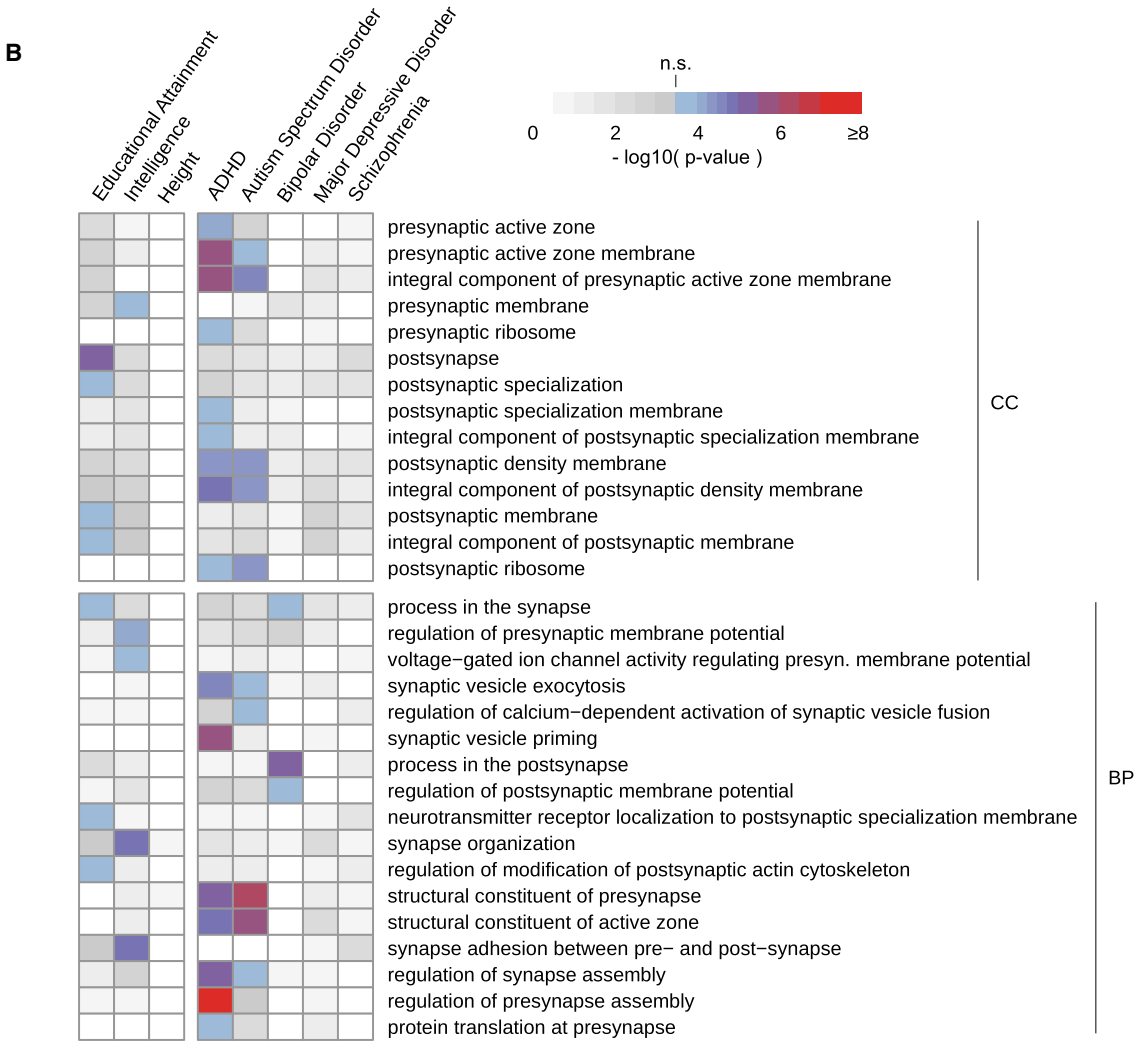
Active-zone preparations yielded smaller numbers of proteins, maximally 249 (Figure 6A). These protein collections contained between 35% and 62% of SynGO-annotated proteins, slightly more than synaptosome and postsynaptic density percentages. A total of 2,084 proteins currently lacking SynGO 1.0 CC annotation were identified in at least three proteomics datasets of synaptosome, active zone, or PSD subcellular fractions (Figure 6B).

Taken together, these data indicate that SynGO aids in dissecting overlap and differences in large synaptic protein sets that were purified in different synaptic preparations. Many proteins identified in such fractions await experimental validation before they can be annotated to SynGO CC and BP terms.

Synaptic Genes Are Enriched among Genes Associated with Various Brain Traits

Results from large-scale genetic studies are often used to test for association of a trait of interest with a set of functionally related genes. Such tests gain power with a higher confidence definition of the gene sets used. We predicted that expert-curated, evidence-based SynGO genes show robust associations with experimental data on brain traits and that SynGO gene sets are more strongly associated than existing synapse gene sets. We tested this prediction on genome-wide association study (GWAS) data for three continuous traits—educational attainment (EA; Lee et al., 2018), IQ (Savage et al., 2018), and human height (Wood et al., 2014)—and for five brain disorders—ADHD (Demontis et al., 2016), ASD (Grove et al., 2019), schizophrenia (Pardiñas et al., 2018), bipolar disorder (Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011), and major depression (Wray et al., 2018). The association with gene sets based on SynGO genes and previously annotated synaptic genes in GO were compared with three control gene sets for these traits: all other genes, other genes with similar brain-enriched expression, and genes with similar (high) conservation. Two analysis methods were used: Multi-marker Analysis of GenoMic Annotation (MAGMA; de Leeuw et al., 2015) and linkage disequilibrium score (LDSC) regression analysis (Bulik-Sullivan et al., 2015). LDSC tests for enrichment of SNP-based heritability for various traits in gene sets, whereas MAGMA tests whether gene-level genetic association with the various traits is stronger in specific gene sets. Both methods account for confounders like gene size and linkage disequilibrium in different ways.





(legend on next page)

Figure 7A shows gene set analyses using MAGMA for ASD. We observed a highly significant association of the sets involving the presynaptic active zone and the postsynaptic density (CC terms) for presynaptic functions and synapse assembly (BP terms; Figure 7A). These associations remained significant, albeit typically less strongly, when conditioned on brain gene expression values (Figure 7A, dark colors) or on homology conservation scores (Figures S10 and S11). Interestingly, one set of SynGO genes, postsynaptic ribosome genes, was not significant compared with all other genes but became significant when conditioned on brain-expressed genes. Hence, gene set analysis for SynGO genes in ASD GWAS data reveals new and highly significant associations with pre- and postsynaptic compartments and presynaptic processes.

Similar analyses were performed for all other traits listed above (Figure 7B). SynGO genes were significantly associated with educational attainment, especially genes annotated with postsynaptic localizations and processes. Five SynGO ontology terms were associated with intelligence, but none were associated with human height. Furthermore, many ontology terms were associated with ADHD, especially ontologies involving locations and functions related to the presynaptic active zone and presynaptic assembly (Figure 7B). Finally, strong associations of both pre- and postsynaptic terms were observed for ASD and for postsynaptic processes with bipolar disorder (Figure 7B). Very similar conclusions were reached when additionally conditioning on homology conservation scores (Figures S10 and S11) and when LDSC regression analysis was used instead of MAGMA (Figures S12 and S13).

Taken together, SynGO genes are strongly enriched in GWAS results for brain-related traits, with new links becoming manifest between ASD and the synapse, ADHD and presynaptic genes, educational attainment and postsynaptic processes, and several other links between synaptic genes and bipolar disorder or intelligence.

Synaptic Genes Are Enriched among *De Novo* Protein-Coding Variants for Four Brain Disorders

In addition to GWAS studies, exome sequence studies of *de novo* coding variations have recently become available, allowing us to perform enrichment studies in SynGO genes among all *de novo* coding variations detected from several brain disorder patient populations. We tested for enrichment in SynGO genes of protein-truncating variants (PTV) and missense mutations that were previously reported to be associated with 4 brain diseases: developmental delay (DD; 4,293 trios), intellectual disability (ID; 971 trios), ASD (3,982 trios), and schizophrenia (SCZ, 1,024 trios), with non-syndromic congenital heart defect (CHD; 1,487 trios) and unaffected siblings (UNAFF SIB; 2,216 trios) as non-affected classes (see Table S7 for all references). PTV and missense mutations were filtered when they were present in the ExAC reference database (Lek et al., 2016), and *de novo* enrichment in each group was compared against a mutation model that estimates the expected mutation rate among each gene set. SynGO gene enrichment was compared with previously annotated synaptic genes in GO and with matched brain-enriched genes: control gene sets with similar brain enrichment and gene size exactly matching SynGO genes. SynGO genes were robustly enriched for all 4 disease classes (Figures 8A and 8B), most strongly for ID (>2-fold enriched) but also for DD (1.6-fold enriched), ASD (1.4-fold enriched), and SCZ (1.3-fold enriched). All of these enrichments for SynGO genes were substantially stronger than for synaptic genes previously annotated in GO, especially for DD and ID (Figure 8A). PTVs and missense mutations in SynGO genes were not enriched for coronary heart disease-non-syndromic (CHD-NS) and in unaffected siblings (Figure 8A).

To test the distribution of these enrichments within SynGO ontology terms, we plotted the enrichment p values for each term as false color values in SynGO CC and BP ontologies (Figures 8C and 8D; Table S7). Highly enriched gene sets were unevenly distributed among locations and processes. For subcellular locations (CC), the strongest associations were observed in the postsynaptic density and active zone together with pre- and postsynaptic plasma membrane terms (Figure 8C). For BPs, the strongest associations accumulated in synaptic vesicle exocytosis and generation of the presynaptic membrane potential, with further association in postsynaptic processes and synapse organization (Figure 8D). Together, these data show that SynGO genes were strongly enriched for *de novo* PTV and missense variations in all four brain disorders. Importantly, SynGO genes are more robustly enriched than GO genes previously annotated to the synapse.

DISCUSSION

This study describes SynGO, the first comprehensive knowledge base that provides an expert community consensus ontology of the synapse. The ontology and annotations accumulated in SynGO provide a comprehensive definition of synapses, new unique features of synapses, new links between synapses and brain disorders, and excellent future perspectives as an up-to-date interactive community resource. Using SynGO 1.0, we analyzed gene/protein properties, evolutionary conservation, mRNA expression, loss-of-function tolerance, protein-protein interaction, and enrichment in GWAS data for brain-related traits and brain disorders and in rare *de novo* coding variations for neurodevelopmental disorders.

Figure 7. Enrichment Study of SynGO Gene Sets in GWASs

(A) MAGMA analysis of autism spectrum disorder revealed enrichment of SynGO CC (light blue) and BP (light green). Conditioning by gene expression values (GTEx) typically reduced the signal, except for the postsynaptic ribosome, as visualized in dark blue and dark green. Only SynGO ontology terms significant after Bonferroni correction at α 0.05 (P_{bon} = 0.05 divided by154, vertical dashed line) in the latter analysis are shown.

(B) Overview of significantly enriched SynGO ontology terms in various GWASs. The p values from the MAGMA analysis, with conditioning by gene expression values, were color-coded from blue to red for all ontology terms significant after Bonferroni correction at α 0.05. Additional studies are available in Figures S10–S13 and Table S6.

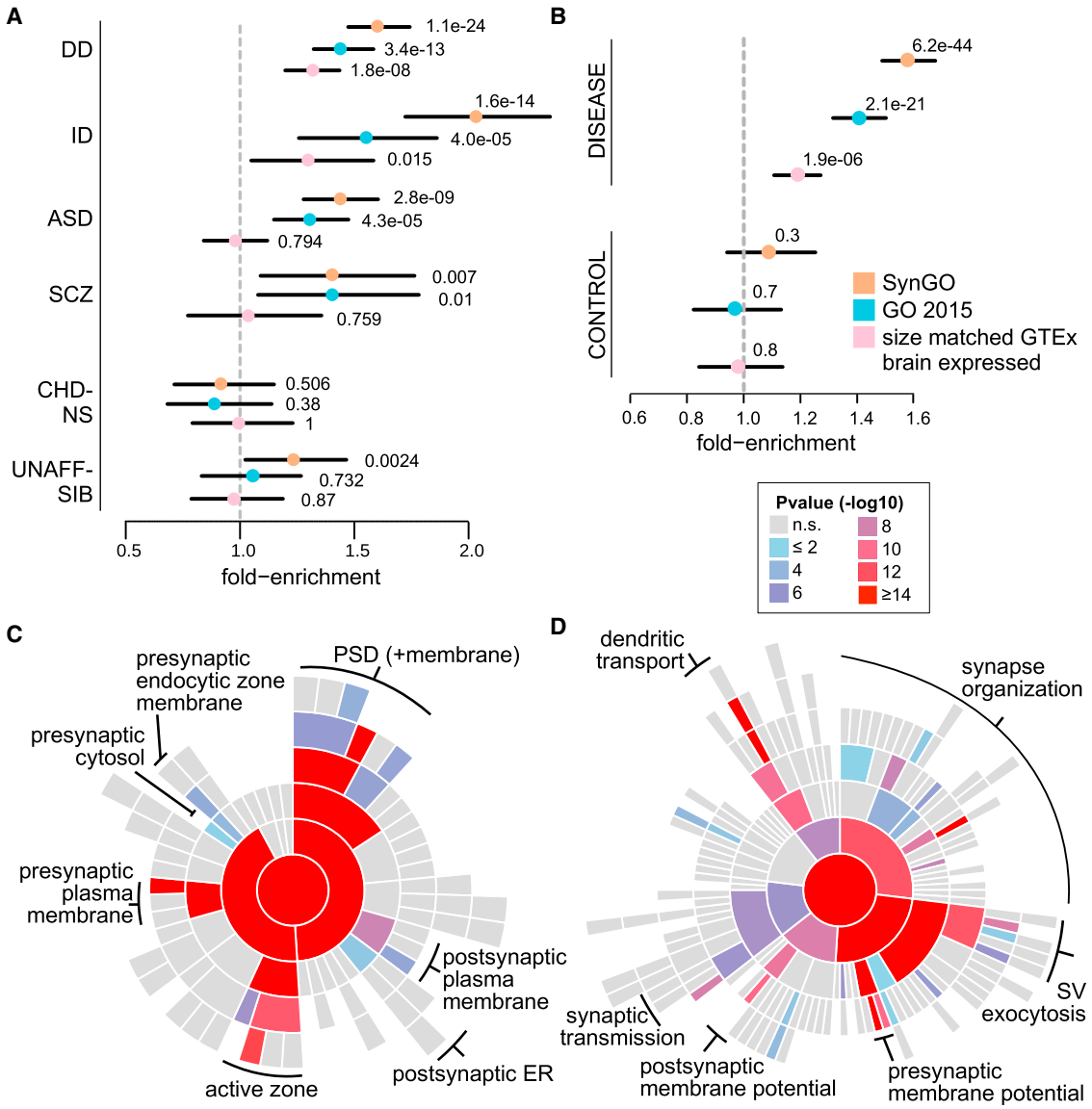


Figure 8. Enrichment for PTV and Missense Mutations in SynGO Genes

(A) Synaptic genes are more enriched for PTV and missense mutations among patients with brain disorders compared with the control set of GTEx brain-expressed genes of equal size and compared with pre-existing synaptic annotations in GO. For each comparison, the p values from a binomial test against mutation model expectation are shown as text, their median fold enrichment as a circle (color-coded by gene set), and the 10%∼90% quantile of fold enrichment as a horizontal line. Patient populations with brain disorders: developmental delay (DD), intellectual disability (ID), autism spectrum disorder (ASD), and schizophrenia (SCZ). As a control group, we included patient populations with non-syndromic coronary heart disease (CHD-NS) or unaffected siblings (UNAFF-SIB).

(B) Group-level effects were tested for the patient populations described in (A).

(C and D) The median disease p value per ontology term (with at least 5 unique annotated genes) was visualized for (C) CC and (D) BP. Note that the CC and BP sunburst plots are aligned with Figures 2C and 2D, respectively.

SynGO Is a Major Step Forward in Defining Synapses

Adequately defining a biological system like the synapse requires a coherent and logical definition of its components, their relationships, and how biological functions emerge from these. The SynGO ontology is the first to provide such definitions coherently for the synapse. The SynGO 1.0 ontology has defined 87 CC and 179 BP terms, designed in consensus by expert laboratories worldwide. Previous models suffered from the lack of a coherent, top-down design of synapse-related ontology terms and relations. Consequently, many heterogeneous terms, both specific and general, were positioned directly under the master term "synapse" (Figures 2A and 2B).

Adequately defining synapses also requires the underlying annotations to be accurate and reliable. SynGO is exclusively based on published, expert-curated evidence and detailed classification of this evidence. This is a substantial innovation that provides

accountability for decisions made by experts and allows structured discussions and resolving annotation disputes, in particular in the web-based SynGO resource (https://syngoportal.org). Moreover, different types of evidence can now be integrated in statistical models in a differential manner. For instance, evidence that is considered very strong can be given a higher weight than evidence less so. Finally, providing evidence-tracking tools to (future) expert contributors engages the synapse research community, ensuring that SynGO annotations are based on solid evidence. Hence, the new SynGO evidence tracking system is a fundamental step forward regarding annotation accuracy, transparency, and expert engagement and a solid basis for future refinements in a biology-driven overall synaptic ontology framework.

Using SynGO 1.0 annotations, we show that the SynGO ontology indeed adequately defines the synapse. We show that (1) SynGO genes are indeed more evolutionarily conserved than other genes (Figure 4), (2) that brain-specific aspects of synapses are particularly enriched (Figure S8), (3) that SynGO proteins documented to interact are much more likely to be annotated to the same ontology terms (Figure S9), and (4) that enrichment of synaptic genes among genes associated with traits in GWAS data (Figure 7) and among rare variants causing neurodevelopmental disorders (Figure 8) is, without exception, stronger for SynGO genes than for gene sets previously annotated to the synapse. Together, these four groups of observations confirm that SynGO adequately defines synapses, consistent with previous findings, and consistently outperforms previous gene set resources used in gene set analyses.

Although the definition of a synapse is now becoming accurate and reliable, the definition of synaptic genes remains precarious. No cellular compartment operates in isolation. Components move in and out, and no gene product, also not of SynGO genes, is expressed exclusively in the synapse. Because GO annotations for location (CC) and process (BP) are independent, genes that regulate synaptic function do not necessarily have to be located in the synapse. In principle, this opens the possibility of annotating, for instance, transcription factors that regulate the expression of synaptic genes. SynGO 1.0 currently only lists a few of these examples, but it will eventually be useful to include such genes in SynGO annotation. Such genes can be easily excluded from an analysis by filtering for CC terms; i.e., only genes that have a confirmed synaptic location will be retained. Other regulatory aspects of synapse function may include proteins derived from the extracellular matrix, axon, dendrite, or glia, which are not yet accommodated in SynGO 1.0.

Taken together, SynGO provides a comprehensive definition of the synapse with new elaborate and consensus ontologies, accurate and transparent evidence tracking, and close to 3,000 validated annotations. SynGO is ready to serve as a universal reference in synapse biology and for enrichment studies using -omics data and to form a fundamental component of future computational models to help understand synaptic computation principles in the brain and their dysregulation in disease.

SynGO Discovers Unique Features of Synaptic Genes and New Disease Links

In addition to adequately defining synapses, SynGO also allowed us to identify several novel features of synapses and synaptic genes. We show that (1) synaptic genes are structurally very different from other genes (Figure 3); (2) that nearly all synaptic genes have evolved prior to the last common ancestor of all vertebrates, much earlier than the average for other human genes (Figure 4); and (3) that synaptic genes are exceptionally intolerant to mutations (Figure 5). The accumulation of more coding and non-coding sequences may have served to expand their transcriptional regulatory repertoire and diversification of functions of the encoded proteins. Larger genes with more intron-exon boundaries may have given rise to more alternatively spliced variants; a prediction that may be validated with the introduction of new long-read RNA sequencing. Also, mechanisms of gene duplication and splicing have generated expansion of synaptic gene diversity. Interestingly, because synaptic genes have been found to be highly intolerant to mutation, this diversification must have come with incorporating new essential synaptic functions, such as in features of plasticity, contributing to the accelerating computational capabilities of the brain during evolution.

Synaptic dysregulation is central to many brain disorders (synaptopathies). The SynGO analyses described here strengthen the links between synapses and many brain traits (Figures 7 and 8). Many SynGO CC and/or BP terms are enriched among genes associated with educational attainment, intelligence, ADHD, ASD, and bipolar disorder. In particular, analysis of SynGO suggests a link between educational attainment and postsynaptic processes, between ADHD and both pre- and postsynaptic genes, between ASD and presynaptic genes (in addition to the well-known links to the PSD; see Bourgeron, 2015), and between bipolar disorder and postsynaptic genes. One informative achievement of SynGO analyses is that, because of the detailed structure of the SynGO ontology, the genetic risk for each disease was mapped to specific synaptic locations and processes. The mapping resolution to specific terms is currently limited by the small number of genes annotated in some subclasses in level 3 and down. More synapse research is necessary to drive this refinement to saturation and allow more specific and definitive associations between genetic risk for brain disorders and distinct synaptic locations and processes.

SynGO Is Expected to Grow as an Expert Community Effort

Although SynGO 1.0 contains 2,922 annotations, this is still only a fraction of all relevant information available in scientific literature. Only for a core set of proteins, SynGO 1.0 contains three or more annotations per protein. A concerted effort of experts involved in synapse research will help to uncover a larger fraction of available information on synapses and further improve the effect of SynGO. The publicly accessible SynGO portal has been optimized to make such efforts with a user-friendly interface and stored credits for each annotator.

SynGO 1.0 contains 2,922 annotations against 1,112 genes, but proteomics studies of synaptic preparations implicate a few thousand proteins in synapses (Figure 6). An unknown fraction of these synaptic candidate proteins will prove to be *bona fide* synaptic, for which the experimental evidence is currently lacking. It is important to note that biochemical purifications

cannot purify synapses or synaptic compartments to completeness, and some candidate proteins will remain false positives. SynGO 1.0 does not include these candidates by default to avoid low-confidence analyses with SynGO data. However, they can be downloaded from the SynGO database for validation studies. SynGO is also working together with UniProt (The UniProt Consortium, 2017) to accumulate information on available antibodies to facilitate this validation.

Using the public SynGO interface (https://syngoportal.org), SynGO ontologies and gene annotations can be used for enrichment analyses of any new dataset (genomic, mRNA, or protein), and differences between experimental and control groups can be computed and visualized using SynGO visualization tools (Figures 1, 2C, and 2D). The SynGO ontologies and annotations are also fully integrated into the central GO resource (http://geneontology.org) and are made available as part of standard GO releases so that this information is automatically included in all analysis tools that use GO. SynGO annotations are available as both standard GO annotations (http://geneontology.org/docs/go-annotations/) and as GO-CAM models (https://geneontology.cloud/browse/g:SynGO).

Proteins that function in different types of synapses are systematically annotated in SynGO. However, SynGO 1.0 and currently published data do not yet provide sufficient resolution to define individual synaptic proteomes (synaptomes) down to specific synapse populations, which will be important to predict function (e.g., being facilitating or depressing or being inhibitory or excitatory) and to identify changes in disease. Biochemical purifications or other systematic studies of specific synapse populations will be required to establish such specific synaptomes. Until such data become available, the currently available single-cell mRNA resources can be a proxy to define which synaptic genes are expressed in specific neuronal populations. Hence, continued research in the synapse field provides excellent opportunities to further improve and expand SynGO, whereas, conversely, SynGO can provide the conceptual framework and be a key hypothesis generator for such future studies.

The approach described here, including the novel evidence tracking and multimodal analyses, may also provide a foundation for higher-fidelity annotation of other systems, other parts of neurons, other brain cells, or non-neuronal cells and systems. Eventually, such efforts will provide a more complete picture of biological processes and common themes; e.g., in secretion principles or signal detection and integration between synapses and other systems.

Conclusion

Taken together, SynGO provides the scientific community with a public data resource for universal reference in synapse research that is fully integrated in the Gene Ontology resource (http://geneontology.org) for online gene enrichment analyses. By engagement of the synapse research community, SynGO aims to reach saturation to establish a truly comprehensive definition of the synapse. SynGO already brings together many expert laboratories but actively seeks the participation of additional experts to annotate new synaptic genes and/or refine existing annotations. A user-friendly interface (https://syngoportal.org) supports submission of such contributions, which will be reviewed by domain experts before being admitted to SynGO.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - Synaptic gene ontologies and integration into GO
  - Annotation systematics
  - Gene expression data
  - Gene features
  - Isoform counts from full-length RNA sequencing
  - PTM data
  - Conservation of synaptic genes
  - Large-scale protein-protein interaction data
  - Proteomics of synaptic fractions
  - GWAS datasets
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - MAGMA gene-set analysis
  - LDSC gene set analysis
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.neuron.2019.05.002.

ACKNOWLEDGMENTS

SynGO was funded by The Stanley Center for Psychiatric Research at The Broad Institute of MIT and Harvard. SynGO was built on previous efforts (synaptic parts lists) funded by the European Union (EUROSPIN HEALTH-F2-2009-241498 and SYNSYS HEALTH-F4-2010-242167). A.B.P. was supported by Spanish grants BFU2012-34398 and BFU2015-69717-P (partially funded by FEDER funds of the European Union), Ramón y Cajal fellowship RYC-2011-08391, the European FP People Marie Curie Action career integration grant 304111, and the CERCA Program/Generalitat de Catalunya. M.R.K. was supported by DFG CRC779 Project B08, EU-JPND STAD, and Leibniz Foundation SAW. P.D.C. was supported by NIH NS36251. E.D.G. and D.C.D. were supported by DFG CRC779 Project B09. N.B. was supported by the German Federal Ministry of Education and Research (ERA-NET Neuron Synpathy) and an ERC advanced grant from the European Union (SynPrime). M.V. was supported by an ERC advanced grant from the European Union (ERC-ADG-322966-DCVfusion).

AUTHOR CONTRIBUTIONS

G.F., S.E.H., F.K., P.v.N., P.D.T., A.B.S., and M.V. designed the study. All authors designed ontologies and reached consensus. R.E.F., B.K., R.C.L., H. Mi, P.G., and D.O.-S. implemented ontologies and evidence in GO, GO training, and quality control. M.A.-A., J.J.E.C., T.C., L.N.C., R.J.F., H.L.G., P.S.M., C.I., A.P.H.d.J., H.J., M.K., N.L., H. MacGillavry, P.v.N., M.N., V.O., R.P., K.-H.S., R.F.G.T., C.V., R.R.-V., and J.v.W. annotated more than 50 synaptic genes. C.B., À.B., T.B., N.B., J.J.E.C., D.C.D., E.D.G., C.H., R.L.H., R.J., P.S.K., E.K., M.R.K., P.S.M., V.O., T.A.R., and C.S. supervised annotations. F.K. and P.v.N. performed annotation QC. M.F., H. Mi, and P.G. performed phylogenetic annotation. A.B., D.P.H., F.K., H.T., K.T., and K.W. performed bioinformatics analyses. B.M.N., D.P., P.D.T., A.B.S., and M.V. supervised bioinformatics analyses. F.K., with input from P.v.N., A.B.S., and M.V.,

designed and built the SynGO portal. F.K., with input from A.B., D.P.H., P.v.N., K.T., and K.W., generated figures. M.V., with input from T.C.B., F.K., A.B.S., P.D.T., and all expert laboratories, wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests. M.S. and C.H. were employees of Genentech, a member of the Roche Group. S.E.H. serves on the Boards of Voyagers Therapeutics and Q-State Biosciences and on the scientific advisory boards of Janssen and BlackThorn.

Received: February 19, 2019
Revised: April 2, 2019
Accepted: April 30, 2019
Published: June 3, 2019

REFERENCES

Abdou, K., Shehata, M., Choko, K., Nishizono, H., Matsuo, M., Muramatsu, S.I., and Inokuchi, K. (2018). Synapse-specific representation of the identity of overlapping memory engrams. Science *360*, 1227–1231.

Abul-Husn, N.S., Bushlin, I., Morón, J.A., Jenkins, S.L., Dolios, G., Wang, R., Iyengar, R., Ma'ayan, A., and Devi, L.A. (2009). Systems approach to explore components and interactions in the presynapse. Proteomics *9*, 3303–3315.

Arnsten, A.F., Wang, M.J., and Paspalas, C.D. (2012). Neuromodulation of thought: flexibilities and vulnerabilities in prefrontal cortical network synapses. Neuron *76*, 223–239.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Battle, A., Brown, C.D., Engelhardt, B.E., and Montgomery, S.B.; GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis &Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; eQTL manuscript working group (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

Bayés, A., van de Lagemaat, L.N., Collins, M.O., Croning, M.D., Whittle, I.R., Choudhary, J.S., and Grant, S.G. (2011). Characterization of the proteome, diseases and evolution of the human postsynaptic density. Nat. Neurosci. *14*, 19–21.

Bayés, A., Collins, M.O., Croning, M.D., van de Lagemaat, L.N., Choudhary, J.S., and Grant, S.G. (2012). Comparative study of human and mouse postsynaptic proteomes finds high compositional conservation and abundance differences for key synaptic proteins. PLoS ONE *7*, e46683.

Bayés, À., Collins, M.O., Reig-Viader, R., Gou, G., Goulding, D., Izquierdo, A., Choudhary, J.S., Emes, R.D., and Grant, S.G. (2017). Evolution of complexity in the zebrafish synapse proteome. Nat. Commun. *8*, 14613.

Biesemann, C., Grønborg, M., Luquet, E., Wichert, S.P., Bernard, V., Bungers, S.R., Cooper, B., Varoqueaux, F., Li, L., Byrne, J.A., et al. (2014). Proteomic screening of glutamatergic mouse brain synaptosomes isolated by fluorescence activated sorting. EMBO J. *33*, 157–170.

Bipolar Disorder and Schizophrenia Working Group of the Psychiatric Genomics Consortium (2018). Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. Cell *173*, 1705–1715.e16.

Boda, B., Dubos, A., and Muller, D. (2010). Signaling mechanisms regulating synapse formation and function in mental retardation. Curr. Opin. Neurobiol. *20*, 519–527.

Bourgeron, T. (2015). From the genetic architecture to synaptic plasticity in autism spectrum disorder. Nat. Rev. Neurosci. *16*, 551–563.

Boyken, J., Grønborg, M., Riedel, D., Urlaub, H., Jahn, R., and Chua, J.J. (2013). Molecular profiling of synaptic vesicle docking sites reveals novel proteins but few differences between glutamatergic and GABAergic synapses. Neuron *78*, 285–297.

Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.

Calvo, S.E., Clauser, K.R., and Mootha, V.K. (2016). MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. Nucleic Acids Res. *44* (D1), D1251–D1257.

Chang, R.Y., Etheridge, N., Nouwens, A.S., and Dodd, P.R. (2015). SWATH analysis of the synaptic proteome in Alzheimer's disease. Neurochem. Int. *87*, 1–12.

Collins, M.O., Husi, H., Yu, L., Brandon, J.M., Anderson, C.N., Blackstock, W.P., Choudhary, J.S., and Grant, S.G. (2006). Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. J. Neurochem. *97* (*Suppl 1*), 16–23.

de Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput. Biol. *11*, e1004219.

De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al.; DDD Study; Homozygosity Mapping Collaborative for Autism; UK10K Consortium (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209–215.

Demontis, D., Lescai, F., Børglum, A., Glerup, S., Østergaard, S.D., Mors, O., Li, Q., Liang, J., Jiang, H., Li, Y., et al. (2016). Whole-exome sequencing reveals increased burden of rare functional and disruptive variants in candidate risk genes in individuals with persistent attention-deficit/hyperactivity disorder. J. Am. Acad. Child Adolesc. Psychiatry *55*, 521–523.

Domazet-Lošo, T., Brajković, J., and Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends in Genetics *23*, 533–539.

Emes, R.D., Pocklington, A.J., Anderson, C.N., Bayes, A., Collins, M.O., Vickers, C.A., Croning, M.D., Malik, B.R., Choudhary, J.S., Armstrong, J.D., and Grant, S.G. (2008). Evolutionary expansion and anatomical specialization of synapse proteome complexity. Nat. Neurosci. *11*, 799–806.

Fadó, R., Soto, D., Miñano-Molina, A.J., Pozo, M., Carrasco, P., Yefimenko, N., Rodríguez-Álvarez, J., and Casals, N. (2015). Novel regulation of the synthesis of α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor subunit GluA1 by carnitine palmitoyltransferase 1C (CPT1C) in the hippocampus. J. Biol. Chem. *290*, 25548–25560.

Filiou, M.D., Bisle, B., Reckow, S., Teplytska, L., Maccarrone, G., and Turck, C.W. (2010). Profiling of mouse synaptosome proteome and phosphoproteome by IEF. Electrophoresis *31*, 1294–1301.

Fromer, M., Pocklington, A.J., Kavanagh, D.H., Williams, H.J., Dwyer, S., Gormley, P., Georgieva, L., Rees, E., Palta, P., Ruderfer, D.M., et al. (2014). De novo mutations in schizophrenia implicate synaptic networks. Nature *506*, 179–184.

Finucane, H., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genetics *47*, 1228–1235.

Ganna, A., Genovese, G., Howrigan, D.P., Byrnes, A., Kurki, M., Zekavat, S.M., Whelan, C.W., Kals, M., Nivard, M.G., Bloemendal, A., et al. (2016). Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. Nat. Neurosci. *19*, 1563–1565.

Gaudet, P., Livstone, M.S., Lewis, S.E., and Thomas, P.D. (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Brief. Bioinform. *12*, 449–462.

Gazal, S., Finucane, H., Furlotte, N., Loh, P., Palamara, P., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B., Gusev, A., et al. (2017). Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nature Genetics *49*, 1421–1427.

Giglio, M., Tauber, R., Nadendla, S., Munro, J., Olley, D., Ball, S., Mitraka, E., Schriml, L.M., Gaudet, P., Hobbs, E.T., et al. (2019). ECO, the Evidence & Conclusion Ontology: community standard for evidence information. Nucleic Acids Res. *47* (D1), D1186–D1194.

Grant, S.G. (2012). Synaptopathies: diseases of the synaptome. Curr. Opin. Neurobiol. *22*, 522–529.

Groschner, L.N., Chan Wah Hak, L., Bogacz, R., DasGupta, S., and Miesenbock, G. (2018). Dendritic integration of sensory evidence in perceptual decision-making. Cell *173*, 894–905.e13.

Grove, J.S.R., Ripke, S., Als, T.D., Mattheisen, M., Walters, R., Won, H., Pallesen, J., Agerbo, E., Andreassen, O.A., Anney, R., et al. (2019). Common risk variants identified in autism spectrum disorder. bioRxiv. https://doi.org/10.1101/224774.

Gupta, I., Collier, P.G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A.B., Sloan, S.A., et al. (2018). Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. Nat. Biotechnol. Published online October 15, 2018. https://doi.org/10.1038/nbt.4259.

Heutink, P., and Verhage, M. (2012). Neurodegeneration: new road leads back to the synapse. Neuron *75*, 935–938.

Hong, S., Beja-Glasser, V.F., Nfonoyim, B.M., Frouin, A., Li, S., Ramakrishnan, S., Merry, K.M., Shi, Q., Rosenthal, A., Barres, B.A., et al. (2016). Complement and microglia mediate early synapse loss in Alzheimer mouse models. Science *352*, 712–716.

Huang, K.Y., Su, M.G., Kao, H.J., Hsieh, Y.C., Jhong, J.H., Cheng, K.H., Huang, H.D., and Lee, T.Y. (2016). dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. Nucleic Acids Res. *44* (D1), D435–D446.

Jeanquartier, F., Jean-Quartier, C., and Holzinger, A. (2015). Integrated web visualizations for protein-protein interaction databases. BMC Bioinformatics *16*, 195.

Kandel, E.R. (2001). The molecular biology of memory storage: a dialogue between genes and synapses. Science *294*, 1030–1038.

Karczewski, K.J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D.M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K.E., Cummings, B.B., et al.; The Exome Aggregation Consortium (2017). The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. *45* (D1), D840–D845.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. *305*, 567–580.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. Mol. Biol. Evol. *34*, 1812–1819.

Lee, J.J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T.A., Bowers, P., Sidorenko, J., Karlsson Linnér, R., et al.; 23andMe Research Team; COGENT (Cognitive Genomics Consortium); Social Science Genetic Association Consortium (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat. Genet. *50*, 1112–1121.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Lips, E.S., Cornelisse, L.N., Toonen, R.F., Min, J.L., Hultman, C.M., Holmans, P.A., O'Donovan, M.C., Purcell, S.M., Smit, A.B., Verhage, M., et al.; International Schizophrenia Consortium (2012). Functional gene group analysis identifies synaptic gene groups as risk factor for schizophrenia. Mol. Psychiatry *17*, 996–1006.

Martin, J., Walters, R.K., Demontis, D., Mattheisen, M., Lee, S.H., Robinson, E., Brikell, I., Ghirardi, L., Larsson, H., Lichtenstein, P., et al.; 23andMe Research Team; Psychiatric Genomics Consortium: ADHD Subgroup; iPSYCH–Broad ADHD Workgroup (2018). A genetic investigation of sex bias in the prevalence of attention-deficit/hyperactivity disorder. Biol. Psychiatry *83*, 1044–1053.

Mattheisen, M., Samuels, J.F., Wang, Y., Greenberg, B.D., Fyer, A.J., McCracken, J.T., Geller, D.A., Murphy, D.L., Knowles, J.A., Grados, M.A., et al. (2015). Genome-wide association study in obsessive-compulsive disorder: results from the OCGAS. Mol. Psychiatry *20*, 337–344.

Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P.D. (2018). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res *47* (D1), D419–D426.

Moczulska, K.E., Pichler, P., Schutzbier, M., Schleiffer, A., Rumpel, S., and Mechtler, K. (2014). Deep and precise quantification of the mouse synaptosomal proteome reveals substantial remodeling during postnatal maturation. J. Proteome Res. *13*, 4310–4324.

Monday, H.R., and Castillo, P.E. (2017). Closing the gap: long-term presynaptic plasticity in brain function and disease. Curr. Opin. Neurobiol. *45*, 106–112.

Morciano, M., Burré, J., Corvey, C., Karas, M., Zimmermann, H., and Volknandt, W. (2005). Immunoisolation of two synaptic vesicle pools from synaptosomes: a proteomics analysis. J. Neurochem. *95*, 1732–1745.

Morciano, M., Beckhaus, T., Karas, M., Zimmermann, H., and Volknandt, W. (2009). The proteome of the presynaptic active zone: from docked synaptic vesicles to adhesion molecules and maxi-channels. J. Neurochem. *108*, 662–675.

Pandya, N.J., Koopmans, F., Slotman, J.A., Paliukhovich, I., Houtsmuller, A.B., Smit, A.B., and Li, K.W. (2017). Correlation profiling of brain sub-cellular proteomes reveals co-assembly of synaptic proteins and subcellular distribution. Sci. Rep. *7*, 12107.

Pardiñas, A.F., Holmans, P., Pocklington, A.J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S.E., Bishop, S., Cameron, D., Hamshere, M.L., et al.; GERAD1 Consortium; CRESTAR Consortium; GERAD1 Consortium; CRESTAR Consortium; GERAD1 Consortium; CRESTAR Consortium (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat. Genet. *50*, 381–389.

Pedroso, I., Lourdusamy, A., Rietschel, M., Nöthen, M.M., Cichon, S., McGuffin, P., Al-Chalabi, A., Barnes, M.R., and Breen, G. (2012). Common genetic variants and gene-expression changes associated with bipolar disorder are over-represented in brain signaling pathway genes. Biol. Psychiatry *72*, 311–317.

Petersen, C.C., and Crochet, S. (2013). Synaptic computation and sensory processing in neocortical layer 2/3. Neuron *78*, 28–48.

Phillips, G.R., Florens, L., Tanaka, H., Khaing, Z.Z., Fidler, L., Yates, J.R., 3rd, and Colman, D.R. (2005). Proteomic comparison of two fractions derived from the transsynaptic scaffold. J. Neurosci. Res. *81*, 762–775.

Psychiatric GWAS Consortium Bipolar Disorder Working Group (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat. Genet. *43*, 977–983.

Ripollés, P., Ferreri, L., Mas-Herrero, E., Alicart, H., Gómez-Andrés, A., Marco-Pallares, J., Antonijoan, R.M., Noesselt, T., Valle, M., Riba, J., and Rodriguez-Fornells, A. (2018). Intrinsically regulated learning is modulated by synaptic dopamine signaling. eLife *7*, e38113.

Roy, M., Sorokina, O., Skene, N., Simonnet, C., Mazzo, F., Zwart, R., Sher, E., Smith, C., Armstrong, J.D., and Grant, S.G.N. (2018). Proteomic analysis of postsynaptic proteins in regions of the human neocortex. Nat. Neurosci. *21*, 130–138.

Ruano, D., Abecasis, G.R., Glaser, B., Lips, E.S., Cornelisse, L.N., de Jong, A.P., Evans, D.M., Davey Smith, G., Timpson, N.J., Smit, A.B., et al. (2010).

Functional gene group analysis reveals a role of synaptic heterotrimeric G proteins in cognitive ability. Am. J. Hum. Genet. *86*, 113–125.

Savage, J.E., Jansen, P.R., Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C.A., Nagel, M., Awasthi, S., Barr, P.B., Coleman, J.R.I., et al. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. Nat. Genet. *50*, 912–919.

Selkoe, D.J. (2002). Alzheimer's disease is a synaptic failure. Science *298*, 789–791.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M.H., Baldock, R., Barbiera, G., et al. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. *43* (W1), W589-98.

Smith, A.C., and Robinson, A.J. (2019). MitoMiner v4.0: an updated database of mitochondrial localization evidence, phenotypes and diseases. Nucleic Acids Res. *47* (D1), D1225–D1228.

Soukup, S.F., Vanhauwaert, R., and Verstreken, P. (2018). Parkinson's disease: convergence on synaptic homeostasis. EMBO J. *37*, e98960.

Spires-Jones, T.L., and Hyman, B.T. (2014). The intersection of amyloid beta and tau at synapses in Alzheimer's disease. Neuron *82*, 756–771.

Südhof, T.C. (2008). Neuroligins and neurexins link synaptic function to cognitive disease. Nature *455*, 903–911.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447–D452.

Thapar, A., Martin, J., Mick, E., Arias Vásquez, A., Langley, K., Scherer, S.W., Schachar, R., Crosbie, J., Williams, N., Franke, B., et al. (2016). Psychiatric gene discoveries shape evidence on ADHD's biology. Mol. Psychiatry *21*, 1202–1207.

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

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Research *45*, D158–D169.

Wilhelm, B.G., Mandad, S., Truckenbrodt, S., Kröhnert, K., Schäfer, C., Rammner, B., Koo, S.J., Claßen, G.A., Krauss, M., Haucke, V., et al. (2014). Composition of isolated synaptic boutons reveals the amounts of vesicle trafficking proteins. Science *344*, 1023–1028.

Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al.; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; LifeLines Cohort Study (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.

Wray, N.R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E.M., Abdellaoui, A., Adams, M.J., Agerbo, E., Air, T.M., Andlauer, T.M.F., et al.; eQTLGen; 23andMe; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat. Genet. *50*, 668–681.

Yates, B., Braschi, B., Gray, K.A., Seal, R.L., Tweedie, S., and Bruford, E.A. (2017). Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res. *45* (D1), D619–D625.

Zhu, X., and Stephens, M. (2018). Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. Nature Comm. *9*.

Zwir, I., Arnedo, J., Del-Val, C., Pulkki-Råback, L., Konte, B., Yang, S.S., Romero-Zaliz, R., Hintsanen, M., Cloninger, K.M., Garcia, D., et al. (2018). Uncovering the complex genetics of human temperament. Mol. Psychiatry. Published online October 2, 2018. https://doi.org/10.1038/s41380-018-0264-5.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
All data produced by SynGO consortium	This paper	https://syngoportal.org
Software and Algorithms		
MAGMA	de Leeuw et al., 2015	https://ctg.cncr.nl/software/magma
Stratified LD-Score Regression (S-LDSC)	Finucane et al., 2015; Gazal et al., 2017	https://github.com/bulik/ldsc
biomaRt	Smedley et al., 2015	https://bioconductor.org/packages/biomaRt
iGraph R package	https://cran.r-project.org/web/packages/igraph/index.html	https://igraph.org/
Other		
Collected data from published synaptic proteomic datasets	This paper	See Table S4
Collected samples and data for PTV and missense mutations	This paper	See Table S7

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Matthijs Verhage (matthijs@cncr.vu.nl).

METHOD DETAILS

Synaptic gene ontologies and integration into GO

Ontology terms in SynGO v1.0 were compared to pre-existing synaptic ontologies in the GO database prior to the starting date of SynGO (2015-01-01). A snapshot of the GO database representing the state at 2015-01-01 was obtained from http://purl.obolibrary.org/obo/go/releases/2014-12-22/go.obo (the last release in 2014) and converted into a directed graph using the iGraph R package (https://igraph.org/). To construct the CC and BP graphs in Figure 2 we first created a tree from the SynGO v1.0 ontologies and classified terms that were present in the GO snapshot as 'reused'. Next, pre-existing synapse related terms that were not used by SynGO, indicated as purple nodes in Figure 2, were defined as subclassifiers of these 'reused' terms within the GO snapshot. Finally, we restricted resulting terms to match the scope of SynGO v1.0 (typical glutamatergic and GABA-ergic synapses). Terms that further specialize parent terms into serotonergic-, dopaminergic-, cholinergic-synapses, neuromuscular junctions, or 'regulation of' terms, were not taken into account in this evaluation of candidate terms for re-use by SynGO. Graphs in Figure 2 were visualized using a force-directed layout algorithm in Cytoscape (Shannon et al., 2003).

SynGO ontologies and annotations were integrated into the existing ontologies within the GO database and will continuously be updated as the SynGO project expands synaptic ontologies and adds annotations in the future. These GO ontologies are available in the 'goslim_synapse' subset, its most recent version is always available at http://purl.obolibrary.org/obo/go/subsets/goslim_synapse.obo. Respective SynGO annotations are translated when exported to GO, e.g., annotations against 'process in the presynapse' are stored in GO as 'biological_process(GO:0008150) occurs_in presynapse(GO:0098793)'. The identifier of such terms that only exist in SynGO starts with "SYNGO":, whereas terms also available in GO have identifiers that start with "GO": (as seen in the SynGO terms list in Table S2). SynGO annotations as integrated into GO are available through existing GO tools and websites, analysis on the SynGO subset is possible by filtering for annotations with the 'contributor = SynGO' property. All data from the SynGO consortium together with purpose-built analysis tools and community engagement are available through the SynGO website at https://syngoportal.org.

Annotation systematics

Detailed reference to the three dimensions of evidence was stored as part of each annotation (PubMed ID, figure numbers, panels, see Table S3), providing a detailed rationale for each annotation, which can be reviewed by SynGO users. For any given study,

annotations were made for the species used and these were subsequently mapped to the consensus human ortholog using HUGO Gene Nomenclature Committee (HGNC) data resource (Yates et al., 2017). Annotations for orthologous genes in different species were possible and encouraged, yielding multiple annotations to the same consensus human ortholog originating from different species. In addition, we applied SynGO annotations in GO Phylogenetic Annotation (Gaudet et al., 2011) to infer annotations to evolutionarily-related genes, using the experimentally-supported SynGO annotations as evidence. In this process, an expert biocurator reviewed all experimentally-supported GO annotations for all members of a gene family in >100 species in the context of a phylogenetic tree and inferred functions of experimentally uncharacterized genes in tens of other organisms. In the current SynGO 1.0 we did not systematically annotate different splice forms of single genes, because systematic evidence for splice site-specific subcellular localizations or functions is currently sparse. In cases where studies used different approaches to reach the same conclusion, multiple annotations for the same gene to the same CC or BP terms were made frequently and were encouraged. Similarly, when evidence existed for annotating a single gene to multiple CC or BP terms (multiple locations or functions), multiple annotations were made and encouraged. Following standard GO annotation practice, the same gene/protein may be annotated at different levels along the SynGO hierarchical ontology tree. For instance, initial evidence may indicate that a protein is involved in synaptic transmission (SynGO term *chemical synaptic transmission*; GO:0007268), a subsequent study may reveal the protein regulates presynaptic secretion (SynGO term *synaptic vesicle exocytosis*; GO:0016079) and the most recent study may show that the protein regulates vesicle priming (SynGO term *synaptic vesicle priming*; GO:0016082).

Gene expression data

The ''brain-expressed'' control set consists of genes that were expressed in significantly higher levels in brain compared to other tissues in Genotype Tissue Expression Consortia (GTEx) data (Ganna et al., 2016). The control set with ''brain topN'' was defined as the N highest expressed genes in brain, where N was set to the number of unique genes annotated in SynGO v1.0. The highest expressed genes were computed by ranking the average gene-expression levels (in RPKM) from all brain samples in GTEx (Battle et al., 2017) version 6 (*GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_rpkm.gct.gz*).

For the brain enrichment analysis of synaptic genes in Figure S8 we computed the mean fold change comparing brain to all other tissues for each gene in the GTEx (version 7) dataset. To examine enrichment, we applied a Wilcoxon Rank-Sum test for each SynGO ontology containing at least 5 genes. We used a one-sided hypothesis test in order to test whether the genes in the annotation are more brain expressed than expected under the null.

Gene features

Gene features described in Figures 3 and S4 were extracted from the BioMart (Smedley et al., 2015) Ensembl Human genes GRCh38.p12 dataset and the Ensembl REST API Endpoints (release 95). Total gene length was computed using the start_position and end_position BioMart attributes (gene start and end, in base pairs). All known splice variants per gene were obtained through BioMart, from which the number of protein coding splice variants were counted using the transcript_biotype attribute. cDNA length was extracted from gene sequences provided through the Ensembl REST API with 'mask_feature = 1' parameter, and analogously all transcript exonic and intronic regions were obtained.

Isoform counts from full-length RNA sequencing

From our recent publication (Gupta et al., 2018) we isolated full-length long reads that were expressed in neuronal subtypes, namely external granular layer neurons, internal granular layer neurons and Purkinje cells and had been attributed to a spliced protein coding gene. Subsequently, we considered only genes that had 20 or more such reads and split this gene list into two subsets: those annotated in SynGO and its complement. These groups differed substantially in the number of reads per gene. In order to normalize this, we randomly selected 10 full-length reads for each gene, resulting in two gene lists (SynGO and non-SynGO) with exactly 10 reads each. We then counted the number of distinct isoforms that these 10 reads described for each gene and repeated this subsampling process 1000 times.

PTM data

We acquired PTM data from the uniprot.org webservice by selecting all active records for proteome *up000005640* (*Homo sapiens*). For each modification, we count the number of positions reported. For instance, to investigate Phosphorylation in some protein accession's record we count the number of times we find any matches to ''phospho*'' (eg; phosphoserine, phosphothreonine or phosphotyrosine) in the column 'modified residue'. The dbPTM data for each PTM was retrieved by downloading respective files from http://dbptm.mbc.nctu.edu.tw/download.php. To compare across both data sources and account for unevenly distributed numbers of splice variants, data were collapsed from each protein isoform to respective genes by selecting the highest PTM count. For PTM counts normalized by protein length, gene-level aggregation was also performed by max value.

Conservation of synaptic genes

Cumulative distribution of genes by gene age: Gene trees, covering ~95% of human genes, were obtained from the PANTHER resource (Mi et al., 2018). Gene duplication events were dated relative to the earliest speciation node descending from the duplication. Trees were then pruned to contain only human paralogs, and the root of the tree (this ensures that fractional gene counts will add

up to the total number of human genes). Each human gene was then traced back through the pruned tree to the root of the tree, and the number of branches was counted; this gives the total number of duplications (plus one, for the root) along the path to the root. Then, for each human gene, for each duplication (and root node) along the path from the gene to the root, a fractional count of 1/total was added to the count of genes that evolved at the date of that node. This process yields a count of human genes gained over each period of evolution, including gene duplication events. Estimated speciation times were taken from the TimeTree resource (Kumar et al., 2017). Abbreviations used in Figure 4 are: Eras; CE: Cenozoic, ME: Mesozoic, PA: Paleozoic, NPR: Neo-Proterozoic, MPR: Meso-Proterozoic, EO: Eoarchean. Periods; NE: Neogene, PA: Paleogene, CRE: Cretaceous, JU: Jurassic, PE: Pennsylvanian, MI: Mississipian, DE: Devonian, CRY: Cryogenian, TO: Tonian, ST: Stenian, CA: Calymmian. The tree of CPT1C-related genes was obtained from the PANTHER website and can be accessed, together with additional information about the sequences and a multiple sequence alignment, at http://pantherdb.org/treeViewer/treeViewer.jsp?book=PTHR22589&species=agr. For enrichment analysis of synaptic genes at different periods of evolution, we extracted reconstructed ancestral genomes from the Ancestral Genomes resource [PMID: 30371900], and used the set of human "proxy genes" for each ancestral gene. The specific ancestral genomes were obtained from the following URLs:

- http://ancestralgenomes.org/species/genes/(list:genes/Metazoa-Choanoflagellida/Homo%20sapiens)
- http://ancestralgenomes.org/species/genes/(list:genes/Bilateria/Homo%20sapiens)
- http://ancestralgenomes.org/species/genes/(list:genes/Craniata-Cephalochordata/Homo%20sapiens)
- http://ancestralgenomes.org/species/genes/(list:genes/Euteleostomi/Homo%20sapiens)

For each ontology term we applied a 1-sided Fisher exact test with 'greater than' hypothesis to compare genes only found in the 'after' set with all genes in the 'before' set. To find enriched terms within the entire SynGO ontology, we first selected the most specific term where each 'gene cluster' (unique set of genes) is found and then applied multiple testing correction using False Discovery Rate (FDR) on the subset of terms that contain these 'gene clusters'. For human-*C. elegans* and human-*D. melanogaster* orthologs, we used the "ancestral genome comparison" functions available in the Ancestral Genomes resource, to obtain the genes in each genome (e.g., human) that descend from each gene in the bilaterian common ancestor ("inparalogs"). We used this information to match up inparalog groups in the two genomes being compared, to obtain sets of orthologs between those genomes; e.g., the inparalog group of human gene(s) that descend from a given bilaterian ancestral gene are all orthologs of the inparalog group of *C. elegans* gene(s) that descend from that same ancestral gene. We classified each ortholog set as either 1:1, 1-to-many, many-to-1 or many-to-many depending on the number of inparalogs in each organism (i.e., whether there were gene duplications after speciation). We then calculated the proportion of genes (either all genes, or only SynGO genes, with at least one ortholog between human and a given model organism) that are in each type of ortholog set.

Large-scale protein-protein interaction data

StringDB (Szklarczyk et al., 2015) 10.5 human interactions ("9606.protein.links.detailed.v10.5.txt") were filtered by combined score (700, high confidence) and experimental evidence (400, medium confidence). StringDB PPIs then were matched to SynGO HGNC annotated genes by gene symbol, or alternative names ("9606.protein.aliases.v10.5.txt") for cases without a match. The distance between a pair of SynGO genes was defined as their path distance. For the CC model, the path distance between a membrane term and it's integral, anchored or extrinsic sub-classes (e.g., from SV membrane to anchored component of SV membrane) was set to zero. For the null distribution we computed all path distances within the CC or BP graph between any pair of all SynGO genes.

Proteomics of synaptic fractions

Proteins identified in selected proteomics studies shown in Figure 6 were taken from the following published sources: (A) synaptosome fractions (7 studies: Bayés et al., 2017; Biesemann et al., 2014; Chang et al., 2015; Filiou et al., 2010; Moczulska et al., 2014; Pandya et al., 2017; Wilhelm et al., 2014); (B) postsynaptic density fractions (PSD, 6 studies: Bayés et al., 2011, 2012, 2017; Collins et al., 2006; Pandya et al., 2017; Roy et al., 2018) and (C) active zone or docked vesicle fractions (5 studies: Abul-Husn et al., 2009; Boyken et al., 2013; Morciano et al., 2005, 2009; Phillips et al., 2005). Identified proteins were mapped to human gene identifiers (HGNC) using the https://www.uniprot.org ID mapping service and mapping tables provided through https://www.genenames.org (Table S4). Keratins were considered an external contaminant and therefore excluded from downstream analysis. The Venn diagram was generated using the 'eulerr' R package.

GWAS datasets

GWAS summary statistics for 8 traits were collected from the following resources; ADHD (Martin et al., 2018), Autism Spectrum Disorder, Bipolar Disorder (Bipolar Disorder and Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2018) and Major Depressive Disorder (Wray et al., 2018) from https://www.med.unc.edu/pgc/results-and-downloads, Educational Attainment (Lee et al., 2018) from https://www.thessgac.org/data, Height (Wood et al., 2014) from https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files, Intelligence (Savage et al., 2018) from https://ctg.cncr.nl/software/summary_statistics, Schizophrenia (Pardiñas et al., 2018) from https://walters.psycm.cf.ac.uk/.

QUANTIFICATION AND STATISTICAL ANALYSIS

MAGMA gene-set analysis

First MAGMA gene analysis (de Leeuw et al., 2015) was performed using the 1000 Genomes Phase3 reference panel for European population by assigning SNPs to genes within a 2kb upstream and 1kb downstream window for 20,319 genes. The default model (SNP-wide mean) was used. Then MAGMA gene-set analyses were then performed for SynGO and original synaptic GO terms. For SynGO, one additional set with all SynGO genes was added, and in total 154 terms with at least 5 annotated (unique) genes were tested. For original GO, 5 additional sets; all synaptic genes, all BP genes, all CC genes, presynapse and postsynapse were added, and in total 96 terms with at least 5 annotated (unique) genes were tested. The gene set analyses were performed with the following three conditions for each trait: 1) no additional covariate, 2) conditioning on brain and average expression across all tissue types based on GTEx v7 RNA-seq dataset (Battle et al., 2017), 3) conditioning on brain and average expression, and the level of conservation of the genes. GTEx v7 RNA-seq data were obtained from https://gtexportal.org. The homology conservation scores in Figures S10–S13 represent the level of conservation of genes, measured by the number of species with homolog genes using 65 species available through BioMart. Bonferroni correction was performed for each analysis separately (P_{bon} = 0.05/154 for SynGO and 0.05/96 for GO). Statistical results are available in Table S6.

LDSC gene set analysis

To assess the contribution of each SynGO term to disease/phenotype heritability, we applied Stratified LD-Score Regression (S-LDSC) (Finucane et al., 2015; Gazal et al., 2017) to binary gene set annotations constructed with a ±100KB window around each gene as done in previous work (Finucane et al., 2015; Zhu and Stephens, 2018). In our analyses, we conditioned on the 75 functional annotations in the baseline-LD model (Gazal et al., 2017), an annotation containing all 23,987 protein-coding genes with a ±100KB window, as well as brain-enriched genes (see above), and a continuous annotation representing the conservation score of each gene. For each gene set from SynGO or pre-existing synaptic GO annotations, we assessed the statistical significance of the gene set annotations standardized effect size τ^*, (defined as the proportionate change in per-SNP heritability associated to a one standard deviation increase in the value of the annotation, conditioned on other annotations included in the model; Gazal et al., 2017) based on Bonferroni correction. Statistical results are available in Table S6.

DATA AND SOFTWARE AVAILABILITY

All data from the SynGO consortium together with purpose-built analysis tools and community engagement are available through the SynGO website at https://syngoportal.org.