The discovery of integrated gene networks for autism and related disorders

Fereydoun Hormozdiari, Osnat Penn, Elhanan Borenstein, et al.

Genome Res. published online November 5, 2014
Access the most recent version at doi:10.1101/gr.178855.114

The discovery of integrated gene networks for autism and related disorders

Fereydoun Hormozdiari[1]*, Osnat Penn[1]*, Elhanan Borenstein[1], Evan. E. Eichler[1, 2]

* Joint First Authors

[1] Department of Genome Sciences, University of Washington, Seattle, WA 98195

[2] Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195

Correspondence to:

Evan Eichler, PhD

Department of Genome Sciences
University of Washington School of Medicine
Foege S-413A, Box 355065
3720 15th Ave NE
Seattle, WA 98195

Email: eee@gs.washington.edu

ABSTRACT

Despite considerable genetic heterogeneity underlying neurodevelopmental diseases, there is compelling evidence that many disease genes will map to a much smaller number of biological subnetworks. We developed a computational method, termed MAGI (Merging Affected Genes into Integrated-networks), that simultaneously integrates protein-protein interaction and RNA-seq expression profiles during brain development to discover “modules” enriched for *de novo* mutations in probands. We applied this method to recent exome sequencing of 1116 autism and intellectual disability patients discovering two distinct modules that differ in their properties and associated phenotypes. The first module consists of 80 genes associated with Wnt, Notch, SWI/SNF and NCOR complexes and shows the highest expression early during embryonic development (8–16 post-conception weeks, pcw). The second module consists of 24 genes associated with synaptic function, including long-term potentiation and calcium signaling with higher levels of postnatal expression. Patients with *de novo* mutations in these modules are more significantly intellectually impaired and carry more severe missense mutations when compared to probands with *de novo* mutations outside of these modules. We used our approach to define subsets of the network associated with higher functioning autism as well as greater severity with respect to IQ. Finally, we applied MAGI independently to epilepsy and schizophrenia exome sequencing cohorts and find significant overlap as well as expansion of these modules suggesting a core set of integrated neurodevelopmental networks common to seemingly diverse human diseases.

INTRODUCTION

There has been considerable progress in the discovery of *de novo* mutations and candidate genes in patients with neurodevelopmental and neuropsychiatric diseases, such as autism spectrum disorders (ASD) (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012a; Sanders et al., 2012), intellectual disability (ID) (de Ligt et al., 2012; Rauch et al., 2012), epilepsy (Allen et al., 2013) and schizophrenia (Fromer et al., 2014; Gulsuner et al., 2013). The excess of severe, truncating mutations but the low frequency of recurrence in the same genes has led to the prediction of hundreds to thousands of genes (Iossifov et al., 2012; O'Roak et al., 2012a; Purcell et al., 2014; Sanders et al., 2012) underlying sporadic cases of disease. Despite this genetic heterogeneity, there is emerging evidence that subsets of these genes are highly connected in protein-protein interaction (PPI) or co-expression networks or modules working in concert toward similar biological functions (Summerized in Allen et al., 2013; Gulsuner et al., 2013; Mitra et al., 2013; O'Roak et al., 2012b; Parikshak et al., 2013; Willsey et al., 2013). Such networks are anticipated to be the future targets of disease therapy (Stessman et al., 2014).

Although few methods allow the use of both the PPI and a co-expression network (e.g., Lin et al., 2010), most studies on neurological diseases focus primarily on one type of network, while ignoring information from other networks or using such data in a post-hoc fashion to further refine and filter genes. O'Roak, for example, focused exclusively on the PPI network when deriving his *CHD8*-beta-catenin network and restricted his analysis to only 39% of the most severe *de novo* mutations discovered in ASD patients (O'Roak et al., 2012b). Similarly, the AXAS approach focused on a specific set of predefined candidate genes and their first-order neighbors, based on PPI networks, transcription factor, and miRNA binding sites (Cristino et al., 2014). NETBAG also uses the PPI network, together with KEGG pathways and many other various descriptors, in an integrative approach in order to detect modules of genes affected by CNVs in autism (Gilman et al., 2011) and a combination of CNVs and SNVs in Schizophrenia (Gilman et al., 2012). However, there are several known limitations for approaches that depend exclusively on PPI datasets. First, PPI datasets are far from complete (Hart et al., 2006), especially when only highly confident edges are considered (Supplementary Table 1). Second, PPI datasets are biased due to emphases on published literature (Hakes et al., 2008). Also, PPI are often limited to a single splicing isoform (Corominas et al., 2014). Co-expression data are less biased and therefore the incorporation of co-expression with PPI helps to mitigate such limitations.

In contrast, Parikshak et al. (2013), applied the WGCNA (Weighted Gene Co-expression Network Analysis) method (Horvath et al., 2006) to find a set of large, mutually exclusive modules (average size >600 genes) that are highly co-expressed during normal brain development and then selected a subset of these modules enriched for *de novo* mutations in autism. The approach was initially unguided by the mutations in cases and controls and did not leverage protein interaction data. Enrichment for *de novo* mutations as well as PPI was tested post hoc to the detection of modules. Although the final significant modules were quite large

(overall >4400 genes across five modules), they did not include some of the most significant genes associated with ASD, such as *CHD8*, *DYRK1A*, and *GRIN2B* (O'Roak et al., 2012b). The DAWN method (Liu et al., 2014) also employs WGCNA as part of a hidden Markov random field algorithm in order to predict risk ASD genes based on TADA scores (He et al., 2013). These risk ASD genes are later used to identify co-expression or PPI subnetworks for ASD.

Other studies (e.g. Gulsuner et al., 2013; Willsey et al., 2013) focus only on specific subsets of genes, group them all into a common set, and search for other genes that share similar expression characteristics. Using this strategy, Willsey et al. (2013) were able to suggest specific neurodevelopmental subtissues and time points critical for autism. They restricted their seeds to nine high-confidence autism genes seen across multiple ASD studies and then expanded this set into modules by adding for each one of them 20 more genes showing the highest co-expression across different brain regions and different time points. The approach identified four significant networks covering 437 genes and it is not clear how it scales with more samples sequenced. Moreover, there is the tacit assumption that the selected genes work together and can therefore be expanded into a single module. Other studies aim at finding modules not specific to any certain disease (Ulitsky and Shamir, 2007) or pathways dysregulated in cases versus controls based on differential expression analysis (Chowdhury et al., 2011; Ulitsky et al., 2010).

Recent studies on ASD and schizophrenia have found that genes with putative loss-of-function (LoF) mutations in cases are not only more densely connected in PPI networks but also demonstrate higher co-expression with each other (Gulsuner et al., 2013; O'Roak et al., 2012b). Motivated by this observation, we have developed a novel method that simultaneously integrates information from both PPI and co-expression networks to identify highly connected modules in both types of networks that are also enriched in mutations in cases and not in controls. We call this method MAGI, short for **M**erging **A**ffected **G**enes into **I**ntegrated-networks. MAGI is based on a combinatorial optimization algorithm that aims to maximize the number of mutations in the modules while accounting for gene length and distribution of putative LoF and missense mutations in cases and controls. MAGI is generic and can be applied to any disease, given a list of *de novo* mutations in cases and relevant co-expression information. Using neurodevelopmental RNA-seq data from the BrainSpan Atlas (http://www.brainspan.org/), we have applied it to exome sequence data generated from ASD, ID, epilepsy and schizophrenia, providing a comprehensive comparison of common and specific gene modules for these diseases.

RESULTS

Algorithm

We define a “disease module” as a set of genes that are enriched in *de novo* mutations in cases compared to controls and show evidence of both a high number of protein interactions and high co-expression during brain development (see Methods for formal problem definition). We

initially considered the union of *de novo* mutations obtained from four published ASD (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012a; Sanders et al., 2012) and two ID (de Ligt et al., 2012; Rauch et al., 2012) studies as input cases with truncated variants obtained from the NHLBI Exome Sequencing Project (ESP) (http://evs.gs.washington.edu/EVS/) as controls. Similar to previous studies (Gulsuner et al., 2013; Krumm et al., 2014; O'Roak et al., 2012b; Parikshak et al., 2013; Willsey et al., 2013), we also found that genes with *de novo* mutations in probands are more likely to be connected by PPI and have higher co-expression when compared to *de novo* mutations in siblings indicating that comparisons between affected and unaffected siblings from the same family provide a powerful approach to test the validity of our method. In addition, using a permutation test (n = 10,000) we found that genes with *de novo* LoF or missense mutations in probands have a significantly higher number of protein interactions connecting them ($p < 0.00029$, Supplementary Figure 40). For the purpose of this study, we use the HPRD (Keshava Prasad et al., 2009) and STRING (Szklarczyk et al., 2011) databases for PPI and RNA-seq data from the BrainSpan Atlas for co-expression analyses.

A naïve approach would be to simultaneously consider all possible subnetworks that show a significant number of protein interactions and high co-expression. Since the optimization version of this problem is computationally *NP*-hard (see Methods for formal problem definition), we employed a two-stage heuristic (Figure 1). Using the color-coding algorithm (Alon et al., 1995), we first define a large set of small "seed pathways", each including 5–8 genes, enriched in *de novo* mutations. For this, we utilize a scoring function that integrates both *de novo* LoF and missense mutations, taking into account the length of the genes. Other scoring functions, such as TADA (He et al., 2013), may be implemented. Every two genes along these paths are required to be highly co-expressed (the top 5% of all co-expression values) and have a PPI edge connecting them.

Second, the method merges these seed pathways into larger modules (Figure 1 and Methods). For this purpose, we developed a random-walk approach that starts with a random seed pathway and continually merges it with other seed pathways, as long as the score of the resulting module does not decrease and the constraints regarding the PPI density and co-expression are satisfied. Repeating this step produces a set of potentially overlapping modules. Last, a local search routine is performed, in which single genes are added or removed from the module, to improve the total score and find a local maximal module (Figure 1). Among all the local optimal solutions, we denote the module with the highest score as the *Best Module (M_Best)*. To account for other suboptimal modules with high scores (e.g., modules with scores within the top one percentile) that overlap *M_Best*, MAGI constructs an "ensemble" of genes, i.e., a union of the genes in these suboptimal modules. For this ensemble of genes we can calculate how many times each gene appears in different suboptimal solutions and finally assign a "confidence score" to every gene. For simplicity we denote genes that appear in more than 5% of the suboptimal modules as *M_Extended*.

To examine whether other distinct modules can be identified, we remove the genes found in *M_Best* from the PPI and co-expression networks and rerun MAGI. This process can be iterated multiple times so that at each iteration *i* module *Mi* is generated. Modules found to be non-significant or show high overlap with modules detected in previous iterations are filtered (see Supplementary Material).

Simulations are used in order to assess the significance of the modules. For this, we shuffle the mutations seen in cases using three different null models: The first model (Null-1) is based on the length of the gene while the other two adjust for transition/transversion ratios from the ESP (Null-2) (Tennessen et al., 2012) and from whole-genome *de novo* mutation rates (Null-3) (Kong et al., 2012). The latter was previously used for an enrichment analysis of genes carrying *de novo* mutations (O'Roak et al., 2012a). To eliminate potential bias, we enforced an approximately similar degree of PPI in our simulations as seen in the observed data (Null-2 and Null-3). Similar simulations were performed by shuffling the mutations observed in unaffected siblings and controls using the same null models. The full details of MAGI and the simulation approach are provided (Methods and Supplementary Material).

MAGI was applied to exome *de novo* mutation datasets (Table 1) obtained from nine recently published studies on ASD (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012a; Sanders et al., 2012), ID (de Ligt et al., 2012; Rauch et al., 2012), epilepsy (Allen et al., 2013) and schizophrenia (Fromer et al., 2014; Gulsuner et al., 2013).

Autism and intellectual disability

Due to the considerable comorbidity between diagnoses, we applied MAGI to the union of *de novo* mutations among autism and ID probands (n = 877, Table 1). We note that the highest scoring seed pathway (comprised of eight genes: *STXBP1*, *SYNGAP1*, *GRIN2B*, *DLG4*, *STX1B*, *PRKCB*, *GRIN1*, and *DLG3*) is significantly enriched (p < 3.2e-2, Bonferroni correction) in the KEGG long-term potentiation signaling pathway and the GO annotations synaptic function (regulation of synaptic transmission, regulation of transmission of nerve impulse, and regulation of neuronal synaptic plasticity) (p < 0.035, Bonferroni correction). All eight genes are associated with neurodevelopmental disease—OMIM (www.omim.org) and AISS (Krumm et al., 2013)—although only four have *de novo* mutations in probands (3/4 are LoF). This top-scoring seed highlights the potential power of using both PPI and co-expression data simultaneously to discover sets of genes that have a similar function with disease relevance. Combining the seed pathways (Figure 1) and reiterating module discovery, we identified two distinct disjoint modules (M1 and M2) associated with autism and ID (Figures 2 and 3, respectively).

For module 1 (Figure 2), *M1_Best* consists of 48 genes, while *M1_Extended* is comprised of 80 genes corresponding to 28 LoF and 39 missense mutations (Figure 2). *M1_Best* is significantly (p < 0.005) enriched in *de novo* mutations (in both overall score of genes and total number of truncating mutations) in comparison to *all* three null models (Figure 2b and 2c and

Supplementary Figure 23), suggesting that indeed *de novo* mutations observed in cases can be clustered into a highly connected and co-expressed module, as opposed to the same number of randomly shuffled mutations. In addition, *M1_Best* is significantly enriched in previously described autism/neurodevelopmental genes (Table 2). This is in stark contrast to the best module found using unaffected sibling data, which does not differ significantly to *any* of the three null models and is not enriched in autism/neurodevelopmental genes (Figure 2d, 2e and Table 2). *M1_Extended* is strongly associated with chromatin remodeling (Supplementary Figure 27), while both *M1_Best* and *M1_Extended* are significantly enriched for the Wnt (p < 3.2E-3 and p < 7.1E-5) and the Notch (*M1_Extended*; p < 1.4E-3) signaling pathways (after Bonferroni correction). We also note overlap with other SWI/SNF (*PBRM1*, *ARID1B*, *SMARCC1* and *SMARCC2*), NCOR and TCF complexes (Figure 2a)—multi-protein complexes critical for normal neuronal development (De Ferrari and Moon, 2006; Ille and Sommer, 2005; Ronan et al., 2013). Probands with *de novo* LoF mutations in *M1_Extended* were found to have significantly lower IQ than all other probands with LoF mutations (p < 0.018). Similarly, probands with *de novo* LoF or missense mutations in *M1_Extended* were found to have significantly lower IQ compared to other probands with the same types of mutations (p < 0.016). The *M1_Extended* module genes are highly expressed for all distinct subtissues of the brain early during development with peak expression between 8–16 pcw (Figure 4b and Supplementary Figures 29 and 31a). We confirmed this pattern using expression data from the Gene Atlas (Gene Expression Omnibus accession number GSE1133), which showed the highest expression in fetal brain when compared to the adult brain or all other brain tissues (Supplementary Figure 32).

After excluding mutations and genes associated with *M1_Best*, we repeated MAGI and discovered a second module (Figure 3). *M2_Extended* includes 24 genes comprising a total of 10 LoF and 11 missense mutations in ASD+ID probands. Six genes carry *de novo* LoF mutations, six genes harbor exclusively missense mutations, and 12 have neither missense nor LoF mutations within currently published datasets but are predicted to be highly related by co-expression and PPI data. The module is enriched for synaptic plasticity genes (*CALM1*, *GRIN2A*, *GRIN2B*, *MAPK1*, *PRKCB*, and *RPS6KA3*; p < 1.6e-5 after Bonferroni correction) (Supplementary Figure 28) and for known neurodevelopmental disease genes (Table 2; a sevenfold enrichment, p < 1.01e-13). We found that among probands with LoF or missense mutations in *M2_Best* or *M2_Extended* the number of individuals with IQ<70 was highly enriched (p < 0.006) compared to probands with LoF or missense mutations outside these modules (Table 2). Similar to the M1 module, the *M2_ Extended* genes show a significant higher level of expression in the fetal brain (p < 0.001) when compared to other tissues. In contrast to M1, the average level of expression of the genes is low prenatally, with a dramatic rise in expression postnatally (Figure 4b and Supplementary Figures 30-32).

More than 50% of the mutations represented within these two modules occur only as *de novo* missense mutations. Since the pathological significance for missense mutations is less clear than truncating mutations, we assessed their severity using the C-score measure (Kircher et al., 2014).

We find that proband missense mutations in *M1_Best* have significantly ($p < 0.01$) higher C-scores (n = 26, median = 20.15) when compared to *de novo* mutations of probands outside the module (n = 667, median = 16.99) (Table 2). Similarly, we find that missense mutations for M2 have significantly higher C-scores ($p < 0.0002$) (n = 11, median = 24.6) when compared to probands' *de novo* missense mutations outside of the module (n = 682, median = 17.025). The most severe mutations (C-scores >= 32) occur in six genes in M1 (*TCF4*, *SUPT16H*, *CASK*, *GPS1*, *HDAC9*, and *DHX9*) and two genes in M2 (*KCNH1* and 3 missense mutations in *STXBP1*) (Figure 4a).

Although the goal of this project was to consider all the brain regions simultaneously, we have also applied MAGI to expression data of specific brain regions separately. The regions considered were the same as the ones analyzed by (Willsey et al., 2013), namely: (1) Primary visual cortex - superior temporal cortex, or V1C-STC cluster; (2) Prefrontal and primary motor-somatosensory cortex or PFC-MSC cluster; (3) Striatum (STR), hippocampus (HIP), and amygdaloid (AMY); (4) Mediodorsal nucleus of thalamus - cerebellar cortex or MD-CBC cluster. For each of these regions, we calculated the co-expression (Pearson correlation coefficient) between every pair of genes considering the time points from 8 pcw to 1 year after birth, and run MAGI to generate *M1_Best* and *M2_Best*. The overlap between the genes found in these four modules and the original modules produced using the full data is provided (Supplementary Figures 34 and 37). Modules produced for the V1C-STC cluster and STR, HIP, and AMY were most similar (over 70% overlap) to the original modules while the one produced for the MD-CBC cluster was the most different (~0.55 overlap). Interestingly, some of these region-specific modules include genes previously associated with autism (e.g., *TRIP12*, *POGZ*, and *NOTCH3*).

Phenotypic associations and overlap with schizophrenia and epilepsy

We investigated the phenotypes associated with these modules and their potential relationship to other neuropsychiatric and neurological diseases. First, we divided the ASD+ID samples into two distinct sets based on IQ, namely: probands with ID (IQ < 70, with or without reported ASD) and samples with ASD but no ID (IQ ≥ 70 or high-functioning autism). We reran MAGI independently on each set and constructed gene modules as described previously (Table 1). Although this treatment reduces the input sample size and power, it focuses network construction on a more homogenous set of disease phenotypes (Table 1). For IQ < 70 we found two significant modules (based on the three null models, $p < 0.005$ for both modules) that highly overlap the previous two ASD+ID modules (Figure 5a). However, for "high-functioning" autism (IQ ≥ 70), we found that significant modules only overlap with the ASD+ID M1 module. This suggests that mutations in the M2 module (i.e., long-term potentiation pathway/synaptic plasticity) are associated with lower IQ but less likely to be found in ASD probands with IQ ≥ 70. While it is clear that M1 genes may be associated with either phenotypic category (Figure 5), it is interesting that the proportion of truncating mutation differs. Only 27% (12/45) of the high-functioning autism patients carry a *de novo* truncating mutation in this module as compared to

40% (17/43) of patients with ID. This finding is consistent with the observation that probands with *de novo* LoF mutations in *M1_Extended* have significantly lower IQ than other probands with LoF mutations ($p < 0.018$).

For comparison, we ran MAGI on two other sets of recently published *de novo* mutations reported for adult schizophrenia and encephalopathy epilepsy trios (Table 1), generating two significant modules for schizophrenia and one for epilepsy (see Supplementary Data 1). The epilepsy *M1_Extended* is enriched for genes associated with SNARE interaction and vesicular transport pathways with $p < 0.03$ after Benferroni correction (KEGG pathways). We also observed a very strong enrichment of autism and ID (ASD+ID) and epilepsy modules for FMRP targets (Darnell et al., 2011). More than 70% (17/24) of *M2_Extended* genes (ASD+ID) are known FMRP targets ($p < 0.00001$), while the epilepsy *M1_Extended* shows more than 61% (22/36) overlap, representing a 9.5-fold enrichment in FMRP targets. In agreement with a recent publication (Fromer et al., 2014), we also find overlap among ASD, ID and schizophrenia networks. The overlap is particularly pronounced among the M1 modules where 27% (12/45) of schizophrenia *M1_Extended* genes overlap the high-functioning autism and ID modules. Genes such as *CUL3*, *ZMYND11*, *SMARCC2*, and *GRIN2A* are noteworthy as they are covered by more than one disease module (Figure 5a and 5b) and have indeed mutated sporadically in different disease studies. In addition, one gene, *POGZ*, shows very high co-expression with *M1_Extended* from ASD and ID studies and has been seen to have multiple *de novo* mutations in ASD and ID as well as schizophrenia. Combined, the data suggest either comorbidity or underlying common neurological pathways with diverse disease outcomes. The networks we have defined provide a framework to explore these possibilities.

DISCUSSION

Detecting PPI, co-expression and gene-ontology networks related to neurodevelopmental and neuropsychiatric diseases is an active area of research (Ben-David and Shifman, 2012; Gilman et al., 2011; Sakai et al., 2011; Voineagu et al., 2011). MAGI differs from previous approaches in that it simultaneously considers both PPI and co-expression data while trying to cover genes that are enriched in mutations in probands when compared to controls. Comparing our results to other network approaches (e.g., AXAS (Cristino et al., 2014), NETBAG (Gilman et al., 2011), and DAWN (Liu et al., 2014)) highlights the importance of using both data sources (Supplementary Figures 12-21, Supplementary Tables 2-4). Modules that were generated using PPI-based approaches seem to have a substantial number of gene pairs with low co-expression during brain development. Comparison to other recently reported co-expression based networks (Parikshak et al., 2013; Willsey et al., 2013) shows overlap as well as substantial differences in network membership: our predicted ASD modules are smaller (Supplementary Table 2) and show a greater enrichment with known neurodevelopmental disease genes (Supplementary Figures 14

and 15). As more samples are sequenced and additional *de novo* variants are discovered, MAGI modules will become further refined in addition to revealing previously undiscovered modules.

An advantage of MAGI stems from its reliance on the *de novo* mutations to directly guide the generation of the modules, as opposed to first defining modules and then testing for a significant enrichment of mutations. There are also, however, limitations. Not all biological interactions will involve protein interactions (i.e., RNA-protein or protein-DNA) and current protein interaction network databases are largely incomplete leading to missing edges. As a result, some of the genes with multiple *de novo* mutations (*POGZ*, *SETBP1*, *ADNP*, and *SCN2A*) demonstrate high co-expression with the detected modules but fail to incorporate into specific modules due to lack of sufficient PPI edges. Indeed, we find a significant enrichment of truncating *de novo* mutations in genes outside of these two modules, which are co-expressed with modules M1 and M2 (Supplementary Figure 41). An area of future development, then, will be to extend membership (with modified penalty parameters) to genes that are highly co-expressed but not directly connected to the module by a protein interaction.

Our analysis of ASD and ID suggests two fundamentally distinct modules with different properties and phenotypic manifestations. The *M1_Extended* module is significantly enriched in genes with chromatin remodeling function and includes many genes encoding the SWI/SNF complex as well as genes associated with NCOR/HDAC3, Notch and Wnt signaling pathways. Genes within this module show the highest level of expression early in development (8–16 pcw). In contrast, the *M2_Extended* module is enriched in synaptic genes associated with long-term potentiation/calcium signaling and shows the highest level of expression postnatally (birth to 1 year) as has been observed for other networks (Willsey et al., 2013). Patients with LoF mutations within M2 are much more likely to be intellectually disabled (IQ < 70) when compared to M1. Although the *M1_Extended* module is more strongly associated with autism, the proportion of *de novo* truncating mutations in that module (27%) among high-functioning autism individuals (IQ > 70) decreases when compared to those that are intellectual disabled (40%).

In our analysis, we find that *de novo* missense mutations within M1 or M2 genes show significantly greater severity when compared to *de novo* missense mutations outside of the gene networks (Figure 4). Thus, one important application of MAGI is to discriminate possible disease-associated missense mutations—a current bottleneck of most exome sequencing studies of parent–child trios. A clear-cut example is the gene *STXBP1* (syntaxin-binding protein 1) identified as part of ASD+ID M2 and epilepsy M1 modules. *STXBP1* is known to play a role in the release of neurotransmitters and is a regulatory protein for the SNARE complex; truncating mutations in it contribute to early infantile epilepsy (Hamdan et al., 2009). In our analysis, three severe *de novo* missense mutations were identified among ASD and ID probands; and four missense mutations were identified in epilepsy probands, three of which carry a high C-score (>30). We propose module membership and the severity of missense mutations as criteria to select *de novo* missense mutations in candidate genes for further prioritization. We note that several genes within the modules have, as of yet, no known truncating or missense mutations in

these studies. Mutations in such highly interconnected genes may be incompatible with life or associated with more severe syndromic forms of developmental delay. *EP300* and *CREBBP* (ASD+ID M1 module), for example, are critical for embryonic development and mutations in them result in Rubinstein-Taybi syndrome. Similarly, *SMAD4* (Wnt pathway) and *SMARCB1* (SWI/SNF complex) are neurodevelopmental genes associated with the Myhre and Coffin-Siris syndromes, respectively (Kleefstra et al., 2012; Tsurusaki et al., 2012). Moreover, recent targeted resequencing of a subset of novel genes within the modules (Coe et al., 2014) (*BCL11A*, *DLL1*, *NCKAP1*, *RAB2A*, *TBR1*, and *ZMYND11*), as well as a cross-validation analysis (Supplementary Figure 45), provide evidence of an excess of disruptive mutations in ASD and ID highlighting the functional utility in the discovery of new disease genes.

A comparison of the modules for ASD/ID, epilepsy and schizophrenia suggests considerable overlap among gene candidates and networks underlying seemingly diverse neurological diseases. In several cases, *de novo* truncating mutations have been observed in the same gene but arise in different disease cohorts. Comparing studies of ASD, ID and schizophrenia, for example, reveals recurrent LoF mutations for *CHD8*, *ZMYND11*, and *SMARCC2*. This finding is in agreement with a recent schizophrenia exome sequencing study that found additional *de novo* mutations in genes such as *CHD8*, *MECP2*, and *AUTS2*—all highly associated with ASD and ID (McCarthy et al., 2014). There are two possible explanations. One hypothesis is that different neurological diseases share common neurodevelopmental pathways such that disruption may lead to different pathologies depending on the genetic background of the patient. An alternative, but not mutually exclusive, explanation may be disease comorbidity. The high comorbidity of ID and ASD is well established with 60% of children with a diagnosis of autism being intellectually disabled. Similar comorbidities have been reported for ID and schizophrenia and epilepsy (Amiet et al., 2008; Rapoport et al., 2009). The epilepsy mutations were discovered among children with epileptic encephalopathies—a severe form of early onset epilepsy frequently associated with disturbances in cognition and behavior. Similarly, many schizophrenics with *de novo* LoF mutations also showed poor school performance consistent with mild ID (Fromer et al., 2014). Ideally, patient recontact and comparison of the phenotypes with disruptions of the same mutation across diverse neurological diseases will be required to determine the true extent of distinct diagnoses and the importance of these nosological divisions (Stessman et al., 2014).

METHODS

Databases and datasets: The *de novo* mutations were collected from nine different studies: four ASD (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012b; Sanders et al., 2012), two ID (de Ligt et al., 2012; Rauch et al., 2012), one epilepsy (Allen et al., 2013), and two schizophrenia cohorts (Fromer et al., 2014; Gulsuner et al., 2013). LoF mutations are defined as likely gene-disruptive or loss-of-function mutations observed in cases. For controls, we used siblings and normal trios from Iossifov et al. (2012); Neale et al. (2012); O'Roak et al. (2012b); Sanders et al.

(2012); Gulsuner et al. (2013). The PPI network is comprised of the union of StringDB v9.05 (Szklarczyk et al., 2011) human (organism ID 9606) interactions that are experimentally verified (experimental scores > 400) and have high confidence scores (>700), together with the complete HPRD database (http://www.hprd.org/). The co-expression network was constructed as a complete graph using the normalized RNA-seq RPKM values from the BrainSpan Atlas (http://www.brainspan.org). Pearson correlation coefficient r was calculated between every pair of genes across all the subtissues and time points, and r^2 were set as the edges' weights.

MAGI algorithm: As shown in Figure 1, MAGI includes two main steps: the first involves finding relatively short seed pathways with high scores and the second is merging them into much larger clusters. In the first step, high-scoring seed pathways are detected, defined as sets of five to eight genes that form a connected simple path in the PPI network, are highly co-expressed, and are enriched in patient mutations. Then in the second step, pathways that are highly connected in the PPI and also co-expressed with each other are combined together to create modules, by applying a random walk on a graph where nodes represent seed pathways and edges represent nodes that can be merged together. The random walk iteratively merges nodes into a module until no more seed pathways can be added to it without reducing its score. Next, a local search is applied on the module to further improve the score allowing for single genes to be included or removed. The clustering process can be repeated many times to create a set of modules that satisfy the constraints, from which the local optimal with highest score module is picked (*M_Best*). In practice we found that there are many suboptimal local modules that partially overlap the highest score local module. To address the high-scoring local maximal modules, the method reports, in addition to the highest scoring local optimal module, the union of the top one percentile of solutions that have been found.

Seed pathways detection: To minimize the effect of edges that are found in the PPI but may not hold in human brain tissues, each of the seed pathways' edges is also required to show a significantly high co-expression (top 5% in the co-expression network, in practice $-r^2 > 0.37$). To detect these pathways, we use an approach based on the color-coding algorithm (Alon et al. 1995) that outputs a set of high-scoring paths in addition to the maximum-score path. The color-coding approach is an efficient method for finding simple paths of size $h \leq \log(|V|)$ in polynomial time. A simple extension of this method allows for finding a path that maximizes the summation of scores assigned to each node. In short, this approach involves two steps: (a) random coloring of the graph's nodes with h different colors and (b) a dynamic programming algorithm for finding the colorful path (i.e., a simple path that covers all h colors exactly once) that maximizes the score. Iterations of these two steps are needed since the optimal path is not necessarily colorful at each iteration. It was shown that an expected $O(e^h)$ iterations is enough to find the optimal path with high probability. We have modified the dynamic programming step (by adding an extra dimension) in order to limit the total number of LoF mutations (Δ) found in controls (see Supplementary Material for more details). In practice, we run 1000 iterations for each threshold (Δ = 0, 1, 2 and 3) and possible path length (h = 5, 6, 7 or 8) to produce a total of

16,000 potential pathways seeds. We then define “high-scoring seeds” as having a score higher than half the score of the optimal seed of the same category. These paths are used in the next step to create the modules (see Supplementary Material for exact definition of the constraints).

Clustering seed pathways: Seeds are merged into high-scoring clusters that satisfy a stringent set of constraints (Supplementary Material). The clustering is modeled as a random walk on the graph of seeds. Each high-scoring seed found in the previous step is considered a node in this new graph, and there are edges between nodes if the union of the two seeds satisfies the constraints. Each random walk on this graph creates a single module U while traversing the graph. We start with a random seed and continually merge it with other neighboring seeds to increase the total score of the module (module score defined similar to (Ideker et al., 2002)) while keeping the constraints satisfied. Finally, after reaching a node in the graph that cannot be further traversed to any other node, we apply a local search on module U. The local search steps are (a) random gene removal, (b) random gene addition, and (c) random gene swap. We continue to apply these steps as long as the new set satisfies the constraints and the total score increases. Although this local search in theory might take many steps, in practice it reaches a local maximum very quickly.

Software availability: MAGI is free for public use. The program, source code (written in C), as well as the input files are publicly available for download at http://eichlerlab.gs.washington.edu/MAGI/.

REFERENCES

Allen, A.S., Berkovic, S.F., Cossette, P., Delanty, N., Dlugos, D., Eichler, E.E., Epstein, M.P., Glauser, T., Goldstein, D.B., Han, Y.J., *et al.* (2013). De novo mutations in epileptic encephalopathies. Nature *501*, 217-221.

Alon, N., Yuster, R., and Zwick, U. (1995). Color-Coding. J Assoc Comput Mach *42*, 844-856.

Amiet, C., Gourfinkel-An, I., Bouzamondo, A., Tordjman, S., Baulac, M., Lechat, P., Mottron, L., and Cohen, D. (2008). Epilepsy in autism is associated with intellectual disability and gender: evidence from a meta-analysis. Biological psychiatry *64*, 577-582.

Ben-David, E., and Shifman, S. (2012). Networks of neuronal genes affected by common and rare variants in autism spectrum disorders. PLoS genetics *8*, e1002556.

Chowdhury, S.A., Nibbe, R.K., Chance, M.R., and Koyuturk, M. (2011). Subnetwork state functions define dysregulated subnetworks in cancer. Journal of computational biology : a journal of computational molecular cell biology *18*, 263-281.

Corominas, R., Yang, X., Lin, G.N., Kang, S., Shen, Y., Ghamsari, L., Broly, M., Rodriguez, M., Tam, S., Trigg, S.A., *et al.* (2014). Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. Nature communications *5*, 3650.

Cristino, A.S., Williams, S.M., Hawi, Z., An, J.Y., Bellgrove, M.A., Schwartz, C.E., Costa Lda, F., and Claudianos, C. (2014). Neurodevelopmental and neuropsychiatric disorders represent an interconnected molecular system. Molecular psychiatry *19*, 294-301.

Darnell, J.C., Van Driesche, S.J., Zhang, C., Hung, K.Y., Mele, A., Fraser, C.E., Stone, E.F., Chen, C., Fak, J.J., Chi, S.W., *et al.* (2011). FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. Cell *146*, 247-261.

De Ferrari, G.V., and Moon, R.T. (2006). The ups and downs of Wnt signaling in prevalent neurological disorders. Oncogene *25*, 7545-7553.

de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., *et al.* (2012). Diagnostic exome sequencing in persons with severe intellectual disability. The New England journal of medicine *367*, 1921-1929.

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.

Gilman, S.R., Chang, J., Xu, B., Bawa, T.S., Gogos, J.A., Karayiorgou, M., and Vitkup, D. (2012). Diverse types of genetic variation converge on functional gene networks involved in schizophrenia. Nature neuroscience *15*, 1723-1728.

Gilman, S.R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M., and Vitkup, D. (2011). Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron *70*, 898-907.

Gulsuner, S., Walsh, T., Watts, A.C., Lee, M.K., Thornton, A.M., Casadei, S., Rippey, C., Shahin, H., Consortium on the Genetics of, S., Group, P.S., *et al.* (2013). Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. Cell *154*, 518-529.

Hakes, L., Pinney, J.W., Robertson, D.L., and Lovell, S.C. (2008). Protein-protein interaction networks and biology--what's the connection? Nature biotechnology *26*, 69-72.

Hamdan, F.F., Piton, A., Gauthier, J., Lortie, A., Dubeau, F., Dobrzeniecka, S., Spiegelman, D., Noreau, A., Pellerin, S., Cote, M., *et al.* (2009). De novo STXBP1 mutations in mental retardation and nonsyndromic epilepsy. Annals of neurology *65*, 748-753.

Hart, G.T., Ramani, A.K., and Marcotte, E.M. (2006). How complete are current yeast and human protein-interaction networks? Genome biology *7*, 120.

He, X., Sanders, S.J., Liu, L., De Rubeis, S., Lim, E.T., Sutcliffe, J.S., Schellenberg, G.D., Gibbs, R.A., Daly, M.J., Buxbaum, J.D., *et al.* (2013). Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS genetics *9*, e1003671.

Horvath, S., Zhang, B., Carlson, M., Lu, K.V., Zhu, S., Felciano, R.M., Laurance, M.F., Zhao, W., Qi, S., Chen, Z., *et al.* (2006). Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proceedings of the National Academy of Sciences of the United States of America *103*, 17402-17407.

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A.F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics *18 Suppl 1*, S233-240.

Ille, F., and Sommer, L. (2005). Wnt signaling: multiple functions in neural development. Cellular and molecular life sciences : CMLS *62*, 1100-1108.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.H., Narzisi, G., Leotta, A., *et al.* (2012). De novo gene disruptions in children on the autistic spectrum. Neuron *74*, 285-299.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., *et al.* (2009). Human Protein Reference Database--2009 update. Nucleic acids research *37*, D767-772.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics *46*, 310-315.

Kleefstra, T., Kramer, J.M., Neveling, K., Willemsen, M.H., Koemans, T.S., Vissers, L.E., Wissink-Lindhout, W., Fenckova, M., van den Akker, W.M., Kasri, N.N., *et al.* (2012). Disruption of an EHMT1-associated chromatin-modification module causes intellectual disability. American journal of human genetics *91*, 73-82.

Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., *et al.* (2012). Rate of de novo mutations and the importance of father's age to disease risk. Nature *488*, 471-475.

Krumm, N., O'Roak, B.J., Karakoc, E., Mohajeri, K., Nelson, B., Vives, L., Jacquemont, S., Munson, J., Bernier, R., and Eichler, E.E. (2013). Transmission disequilibrium of small CNVs in simplex autism. American journal of human genetics *93*, 595-606.

Krumm, N., O'Roak, B.J., Shendure, J., and Eichler, E.E. (2014). A de novo convergence of autism genetics and molecular neuroscience. Trends in neurosciences *37*, 95-105.

Lin, C.C., Hsiang, J.T., Wu, C.Y., Oyang, Y.J., Juan, H.F., and Huang, H.C. (2010). Dynamic functional modules in co-expressed protein interaction networks of dilated cardiomyopathy. BMC systems biology *4*, 138.

Liu, L., Lei, J., Sanders, S.J., Willsey, A.J., Kou, Y., Cicek, A.E., Klei, L., Lu, C., He, X., Li, M., *et al.* (2014). DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. Molecular autism *5*, 22.

McCarthy, S.E., Gillis, J., Kramer, M., Lihm, J., Yoon, S., Berstein, Y., Mistry, M., Pavlidis, P., Solomon, R., Ghiban, E., *et al.* (2014). De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. Molecular psychiatry.

Mitra, K., Carvunis, A.R., Ramesh, S.K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. Nature reviews Genetics *14*, 719-732.

Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.F., Stevens, C., Wang, L.S., Makarov, V., *et al.* (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242-245.

O'Roak, B.J., Vives, L., Fu, W., Egertson, J.D., Stanaway, I.B., Phelps, I.G., Carvill, G., Kumar, A., Lee, C., Ankenman, K., *et al.* (2012a). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science *338*, 1619-1622.

O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., *et al.* (2012b). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246-250.

Parikshak, N.N., Luo, R., Zhang, A., Won, H., Lowe, J.K., Chandran, V., Horvath, S., and Geschwind, D.H. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell *155*, 1008-1021.

Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kahler, A., *et al.* (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature *506*, 185-190.

Rapoport, J., Chavez, A., Greenstein, D., Addington, A., and Gogtay, N. (2009). Autism spectrum disorders and childhood-onset schizophrenia: clinical and biological contributions to a relation revisited. Journal of the American Academy of Child and Adolescent Psychiatry *48*, 10-18.

Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N., *et al.* (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet *380*, 1674-1682.

Ronan, J.L., Wu, W., and Crabtree, G.R. (2013). From neural development to cognition: unexpected roles for chromatin. Nature reviews Genetics *14*, 347-359.

Sakai, Y., Shaw, C.A., Dawson, B.C., Dugas, D.V., Al-Mohtaseb, Z., Hill, D.E., and Zoghbi, H.Y. (2011). Protein interactome reveals converging molecular pathways among autism disorders. Science translational medicine *3*, 86ra49.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., *et al.* (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237-241.

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 research *13*, 2498-2504.

Stessman, H.A., Bernier, R., and Eichler, E.E. (2014). A genotype-first approach to defining the subtypes of a complex disease. Cell *156*, 872-877.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., *et al.* (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic acids research *39*, D561-568.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., *et al.* (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64-69.

Tsurusaki, Y., Okamoto, N., Ohashi, H., Kosho, T., Imai, Y., Hibi-Ko, Y., Kaname, T., Naritomi, K., Kawame, H., Wakui, K., *et al.* (2012). Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. Nature genetics *44*, 376-378.

Ulitsky, I., Krishnamurthy, A., Karp, R.M., and Shamir, R. (2010). DEGAS: de novo discovery of dysregulated pathways in human diseases. PloS one *5*, e13367.

Ulitsky, I., and Shamir, R. (2007). Identification of functional modules using network topology and high-throughput data. BMC systems biology *1*, 8.

Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., Tian, Y., Horvath, S., Mill, J., Cantor, R.M., Blencowe, B.J., and Geschwind, D.H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature *474*, 380-384.

Willsey, A.J., Sanders, S.J., Li, M., Dong, S., Tebbenkamp, A.T., Muhle, R.A., Reilly, S.K., Lin, L., Fertuzinhos, S., Miller, J.A., *et al.* (2013). Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell *155*, 997-1007.

Tables.

Table 1. Phenotype distribution of *de novo* mutations/modules

		***De Novo* Mutations**		**Significant Modules**	
	Samples	**Missense**	**LoF or Indels**	**M1_Extended***	**M2_Extended**
ASD+ID	1116	696	181	80 (66)	24 (20)
ASD (IQ≥70)	603	383	93	78 (45)	-
ID (IQ<70)	417	278	80	68 (43)	19 (14)
Schizophrenia	722	466	86	52 (33)	26 (11)
Epilepsy	264	194	35	35 (13)	-
Siblings_Control	697	368	61	-	-

***** Number of genes (and in parentheses number of distinct samples with *de novo* mutations) in the modules.

Table 2. Modules enrichment

	OMIM/AISS Enrichment			C-Score		IQ	
Modules (ASD+ID)	**Genes in OMIM+AISS**	**Enrichment OMIM AISS**	***p*-Value OMIM AISS**	**Inside Module Outside Module**	***p*-value**	**Inside Module Outside Module***	***p*-value**
M1_Best (n=48)	*ARID1B*[2], *CASK*[2], *CHD8*[1], *CREB1*, *CTNNB1*[2], *DYRIK1A*[2], *EP300*, *EPHB2*, *HDAC2*, *MECP2*[2], *SMC3*, *TCF4*, *TUBA1A*	9.6x 2.02x	4.2E-5 0.00267	n=26, median=20.15 n=667, median=16.99	0.01	n= 47, low/high ratio=1.474 n=501, low/high ratio=0.722	0.021
M1_Extended (n=80)	*ARID1B*[2], *CASK*[2], *CHD8*[1], *CREB1*, *CREBBP*, *CTNNB1*[2], *DYRIK1A*[2], *EP300*, *EPHB2*, *GTF2IRD1*, *HDAC2*, *MECP2*[2], *SIRT1*, *SMAD4*, *SMARCB1*[2], *SMC3*, *TCF4*, *TUBA1A*	6.67x 2.3x	1.0E-4 0.002	n=39, median=20.6 n=654, median=16.93	4E-04	n= 63, low/high ratio=1.1 n=485, low/high ratio=0.732	0.138
M2_Best (n=21)	*CALM1*, *DLG3*[2], *DLG4*, *DLGAP1*, *GRIN2A*[3], *GRIN2B*[2], *HTR2A*, *KCNMA1*, *MAPK1*, *RIMS1*, *RPS6KA3*[2], *SHANK2*[1], *STX1A*, *STXBP1*[3], *SV2B*, *SYNGAP1*[2], *SYT1*	25.8x 7.1x	4.7E-9 4.71E-13	n=11, median=24.6 n=682, median=17.025	2E-04	n= 17, low/high ratio=3.25 n=531, low/high ratio=0.735	0.006
M2_Extended (n=24)	*CALM1*, *DLG3*[2], *DLG4*, *DLGAP1*, *GRIN2A*[3], *GRIN2B*[2], *HTR2A*, *KCNMA1*[3], *PRKCB*, *RIMS1*, *RPS6KA3*[2], *SHANK2*[1], *STX1A*, *STXBP1*[3], *SV2B*, *SYNGAP1*[2], *SYT1*	27.1x 6.4x	2.2E-8 6.2E-12	n=11, median=24.6 n=682, median=17.025	2E-04	n= 17, low/high ratio=3.25 n=531, low/high ratio=0.735	0.006
Sib1_Best (n=43)	*AKT1*, *ATRX*[2], *DNMT1*, *EP300*, *HDAC2*, *MBD3*, *SIRT1*, *SMC3*	1.64x 1.59x	0.45 0.13	n=19, median=18.08 n=347, median=15.53	0.66	-	-
Sib1_Extended (n=59)	*AKT1*, *ARTX*[2], *AXIN1*, *DNMT1*, *EP300*, *HDAC2*, *MBD3*, *MECP2*[2], *SIRT1*, *SMC3*	2.5x 1.6x	0.189 0.092	n=19, median=18.08 n=347, median=15.53	0.66	-	-

List of genes in different modules with known neurodevelopmental diseases (using OMIM and AISS+OMIM datasets). The superscript above each gene represents the

annotated disease from OMIM. 1: autism, 2: intellectual disability, 3: epilepsy. * the low/high ratio indicates the ratio of number of probands carrying *de novo* LoF or missense mutations in the module with ID (IQ<70) over the ones without ID (IQ≥70). *P-value*s for enrichment in OMIM and AISS are calculated by Fisher's Exact Test. *P-values* for higher C-scores of probands missense mutations in the modules are calculated by the Mann Whitney test on the raw C-scores. *P*-values for enrichment of individuals with IQ<70 among probands with LoF or missense mutations in the modules are calculated by Fisher's exact test.

Figure Legends

Figure 1. Flowchart of MAGI. Given a PPI and co-expression network and case and control mutations, MAGI detects highly connected modules that are enriched for mutations in cases. The first phase calculates a score for each gene in the networks and selects seed pathways with high scores based on an extension of the color-coding algorithm (Alon et al., 1995). In the second phase, MAGI merges the seeds into modules using a random-walk approach and improves each one of them by applying a local search. The output consists of the best module detected, as well as a set of suboptimal modules. Each gene is assigned a "confidence score" according to its frequency within suboptimal modules.

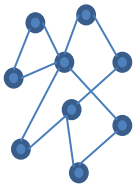
Figure 2. **Module *M1*.** a) Genes detected as part of module *M1_Extended* are displayed as graph nodes using Cytoscape (Shannon et al., 2003). Node colors reflect the score of each gene based on the number and type of *de novo* mutations: the more intense red color indicates a higher score while bright gray indicates a score of zero (no *de novo* mutations observed). Edges (black lines) between two nodes represent genes that interact with each other according to the PPI network and are also highly co-expressed (Pearson correlation coefficient $r^2 > 0.37$, i.e. the genes are included in the top 5% of gene pair co-expression during brain development). The innermost circle contains genes detected in more than 99% of the suboptimal solutions. Subsequent concentric circles display genes found in more than 80%, 20%, and 5% (*M1_Extended*) of the suboptimal solutions, respectively (see Methods). Nodes with black outlines are the ones detected in the optimal module detected (*M1_Best*). See Supplementary Figure 38 for a force-directed layout of this module. b) The *M1_Best* score (dashed black line) shown in comparison to the top-scored module of 200 simulations using null model Null-1. c) The number of LoF mutations covered by the top-scoring module (*M1_Best*) found using proband mutations versus the number of simulated LoF mutations covered by the top-scoring modules found under the same simulation. d) The score of the top module found using siblings and control mutations (dashed black line) in comparison to the top-scored module of 200 simulations with the same number of mutations using Null-1. The siblings simulations were performed without using the ESP constraint (although similar results were obtained when an ESP constraint was applied). e) The number of LoF mutations covered by the top-scoring module found using siblings' mutations versus the number of simulated LoF mutations covered by the top-scoring modules found using Null-1. Sibling simulations were performed without filtering based on ESP controls.

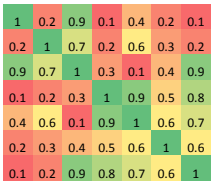
Figure 3. Module *M2*. MAGI reiteration after M1 components removed. See legend of Figure 2 for details. See Supplementary Figure 39 for a force-directed layout of this module.

Figure 4. Severity of missense mutations and differences in temporal patterns of expression. a) The distribution of C-scores of probands' missense mutations found inside *M1_Extended* or *M2_Extended*, outside *M1_Extended* and *M2_Extended*, and unaffected siblings' missense mutations. b) Average normalized expression of all brain subtissues for *M1_Extended* and *M2_Extended* during brain development. Error bars represent the mean ± SEM.

Figure 5. Overlap of the modules identified for different neurodevelopmental diseases. a) Venn diagram representing the overlap between genes that carry *de novo* mutations and are detected as part of the first modules when analyzing ASD without ID (IQ $\geq$ 70), ID with or without reported ASD (IQ < 70), and schizophrenia. Genes with LoF mutations are colored in red. Genes with only missense mutations are colored in black. Asterisks indicate genes for which mutations have been observed in two different groups. b) Venn diagram representing the overlap between genes that carry *de novo* mutations and are detected as part of the second module of ID, the second module of schizophrenia, and the first module of epilepsy. The *p*-value reported for each disease is the maximum *p*-value of the three null models.

Figure 1.





Gene	LoF	Missense
CHD8	3	0
CTNNB1	1	1
EP300	0	0
ARID1B	1	0

Gene	Severe (ESP)
CHD8	0
CTNNB1	0
EP300	0
ARID1B	1

PPI network

Co-expression network

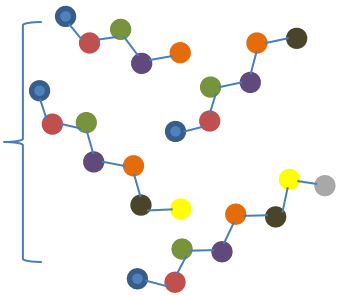
Cases

Control

Calculate gene scores

Generate seed pathways (Color-coding)

16,000 seeds

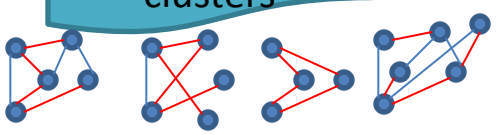


Create clusters:
1. Merge (Random walk)
2. Improve (Local search)

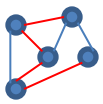
100,000 clusters

Confidence score for each gene

Top 1% suboptimal clusters



Best module



M_Best

M_Extended

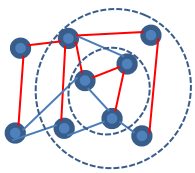
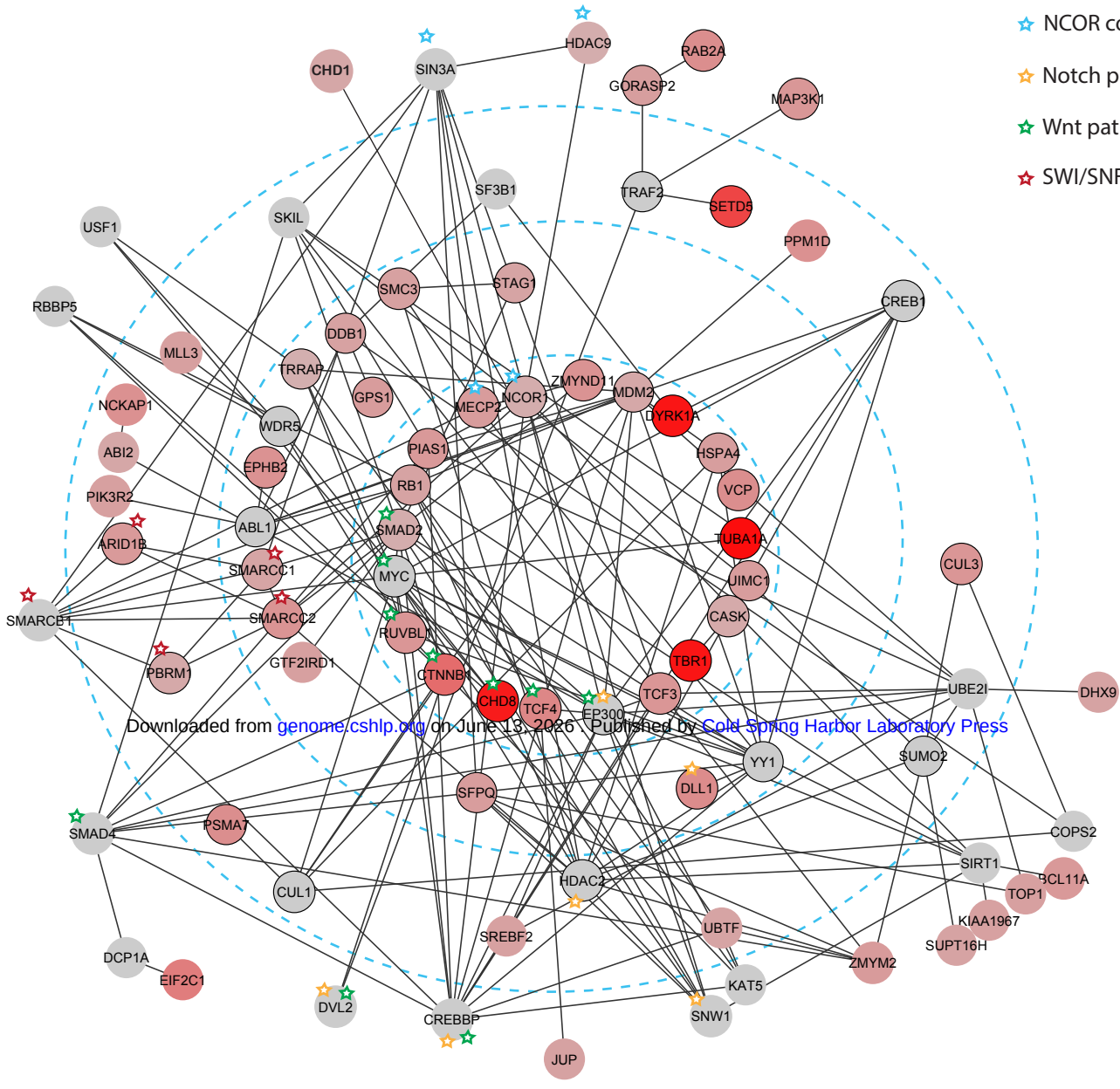


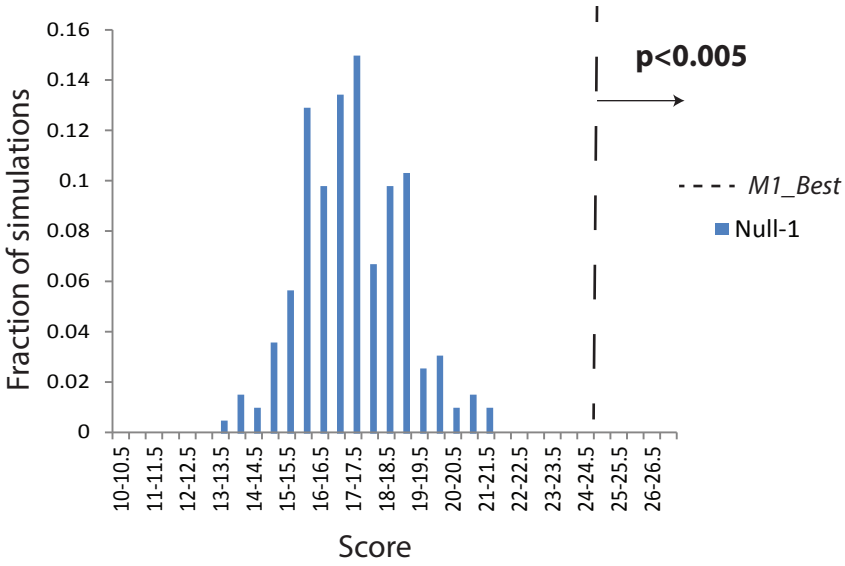
Figure 2.

a)

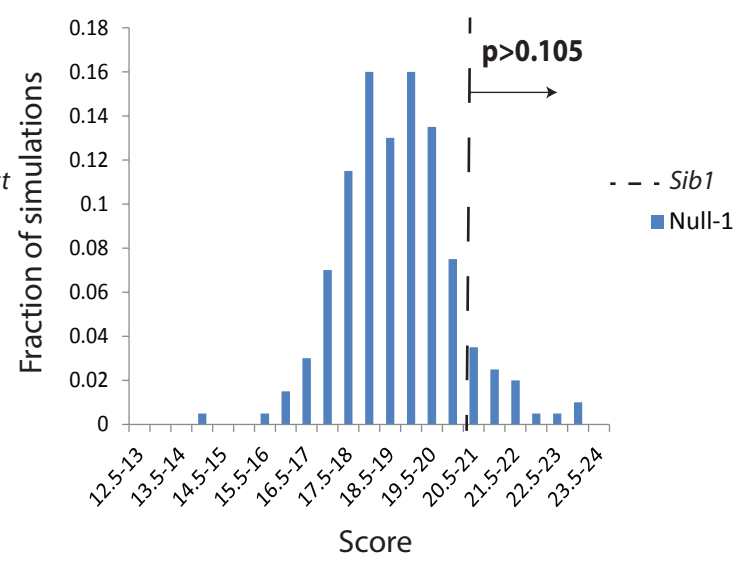


b) Probands M1

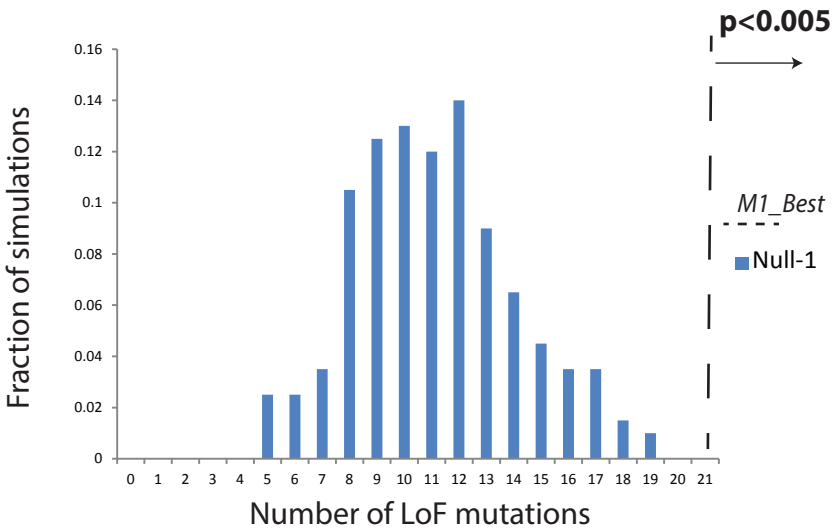
d) Siblings M1



c)



e)



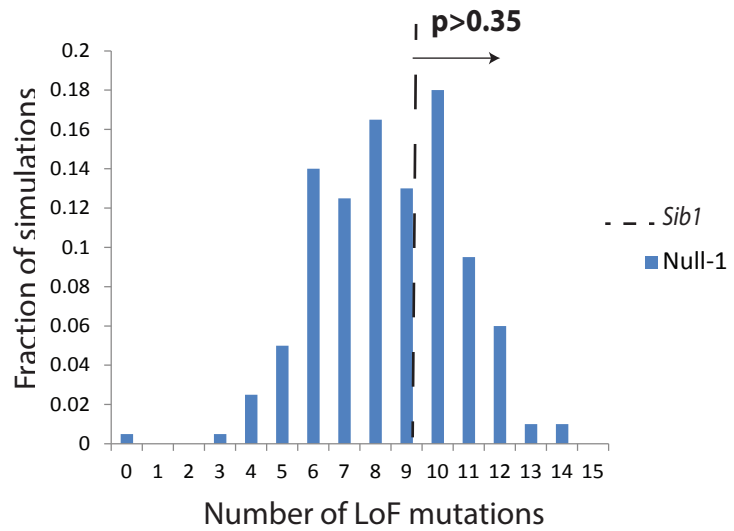
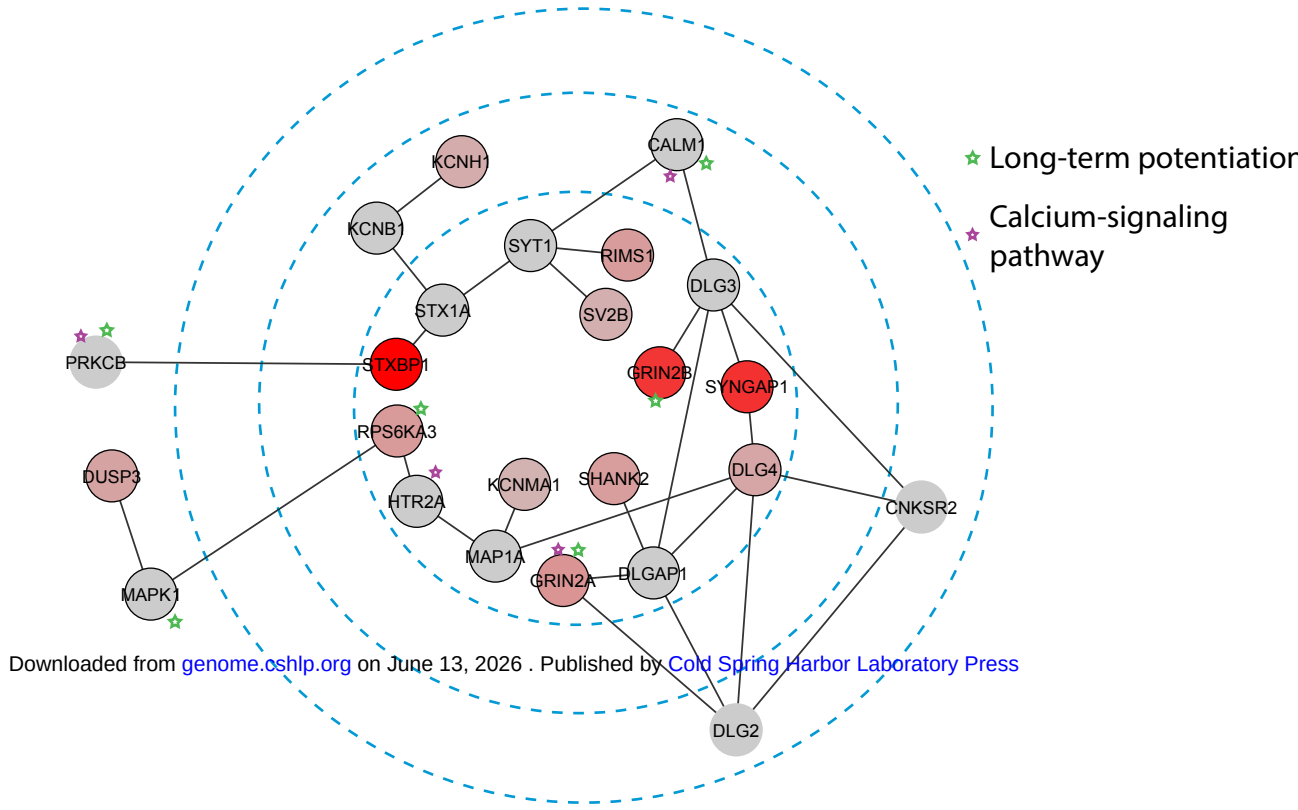
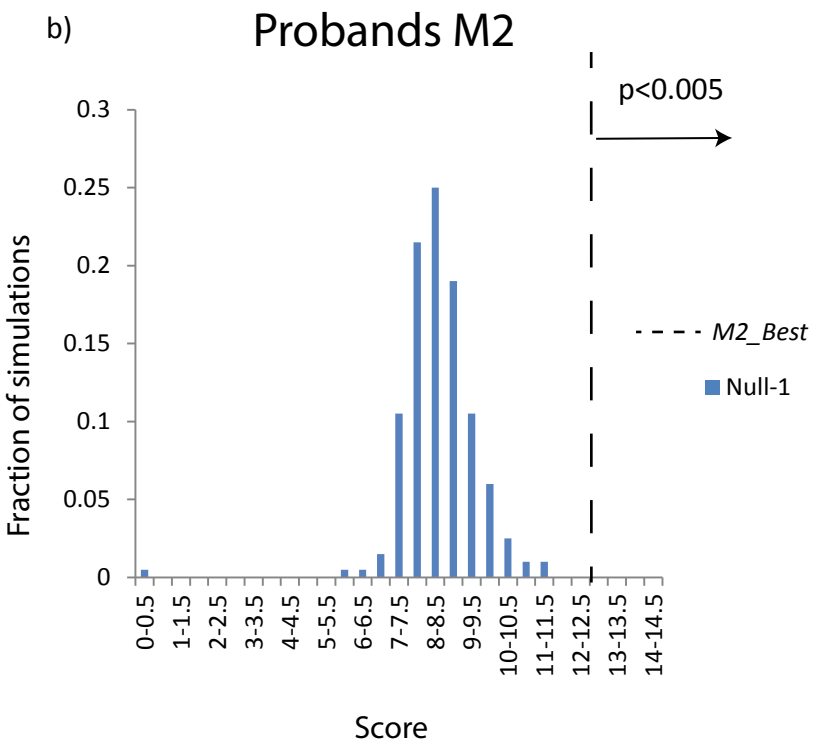


Figure 3.

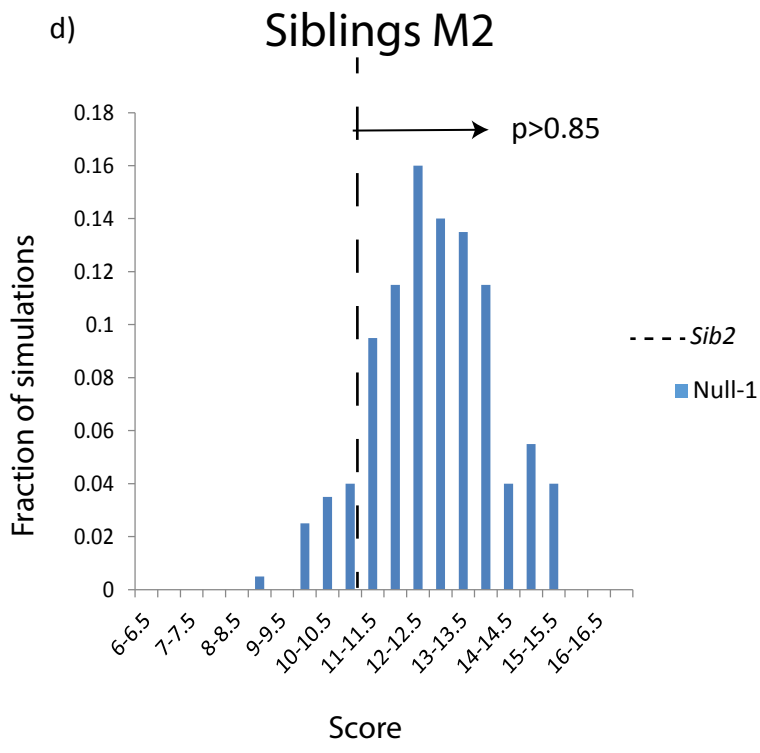
a)



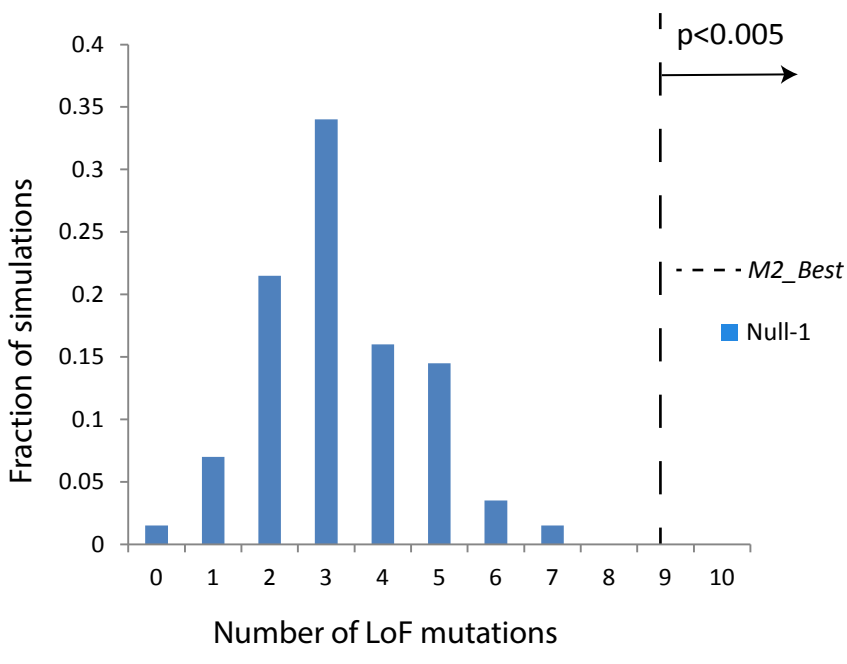
b)



d)



c)



e)

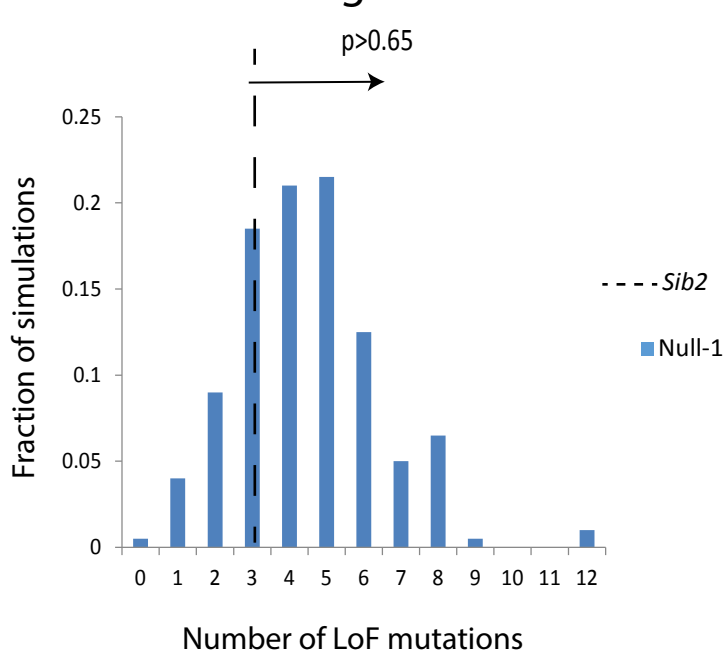


Figure 4.

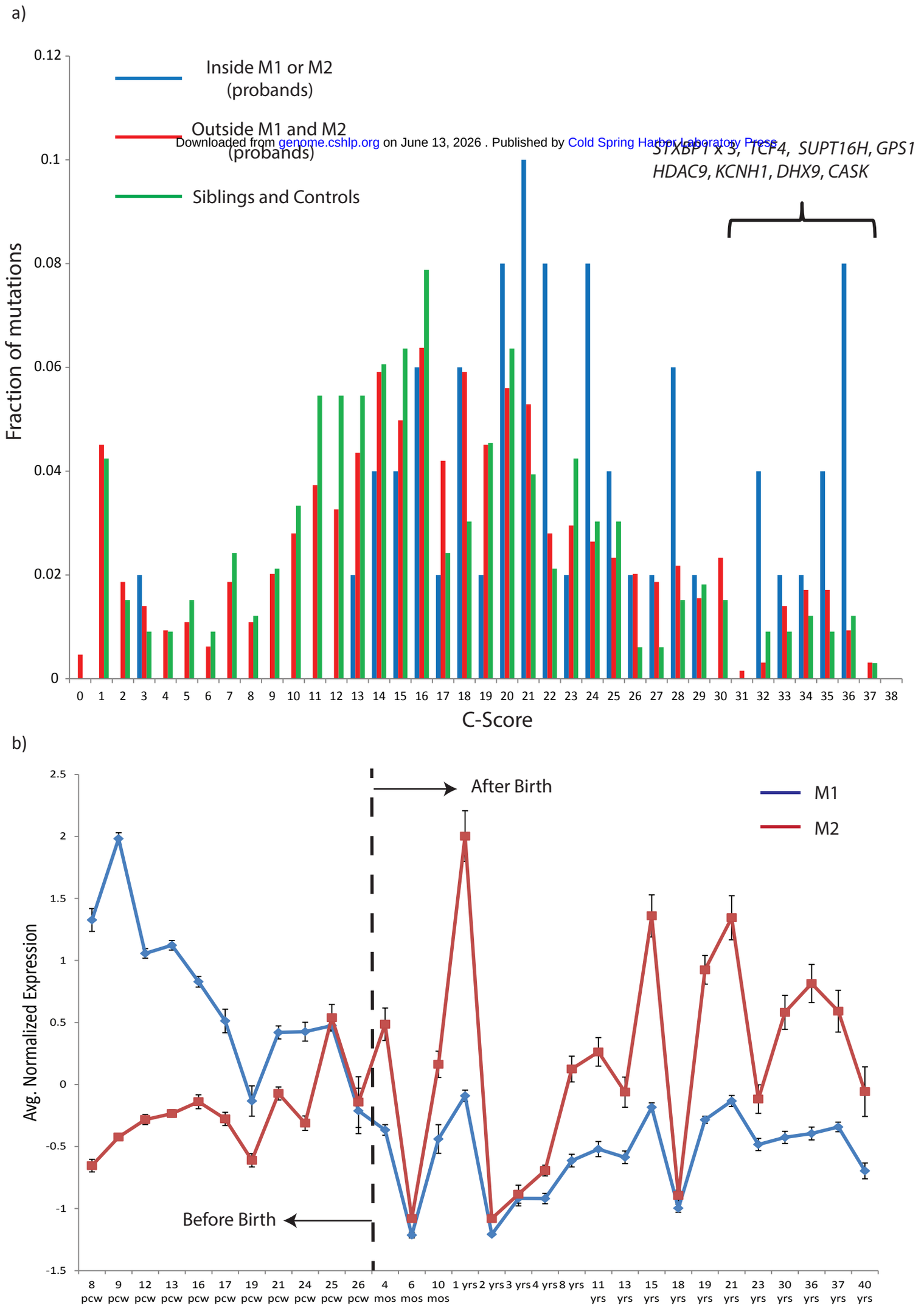
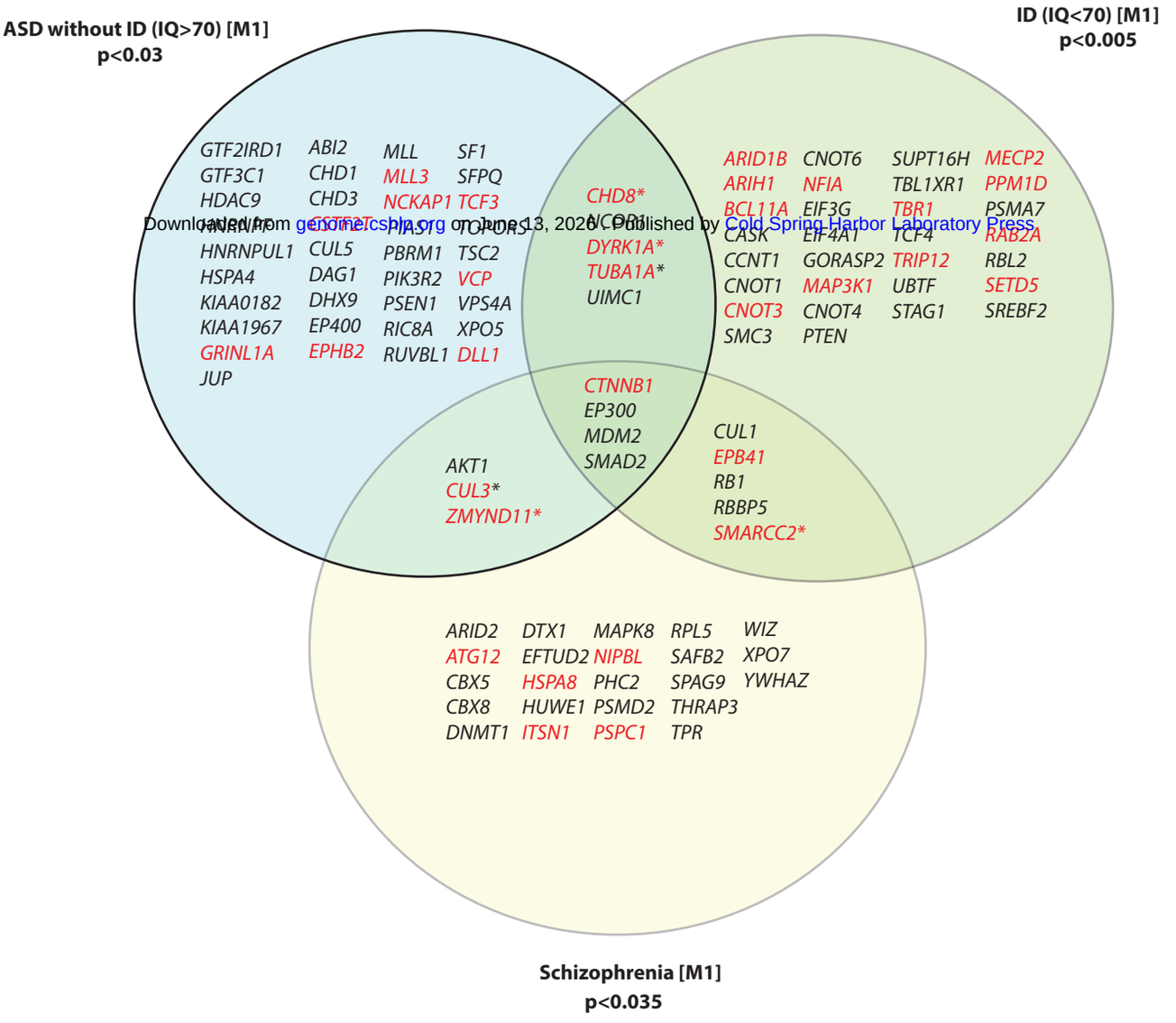


Figure 5.

a)



b)