The discovery of integrated gene networks for autism and related disorders

Fereydoun Hormozdiari,1,5 Osnat Penn,1,5 Elhanan Borenstein,1,2,3 and Evan E. Eichler1,4

1Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA; 2Departments of Computer Science and Engineering, University of Washington, Seattle, Washington 98102, USA; 3Santa Fe Institute, Santa Fe, New Mexico 87501, USA; 4Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA;

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 interactions 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 patients with autism and intellectual disability, 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 found significant overlap as well as expansion of these modules, suggesting a core set of integrated neurodevelopmental networks common to seemingly diverse human diseases.

[Supplemental material is available for this article.]

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 (Gulsuner et al. 2013; Fromer et al. 2014). 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; Sanders et al. 2012; Purcell et al. 2014) 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 coexpression networks or in modules working in concert toward similar biological functions (for summaries, see O’Roak et al. 2012b; Allen et al. 2013; Gulsuner et al. 2013; Mitra et al. 2013; Parikshak et al. 2013; Willsley 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 PPI and coexpression network analyses (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 et al. (2012b), for example, focused exclusively on the PPI network when deriving their CHD8–beta-catenin network and restricted their analysis to only 39% of the most severe de novo mutations discovered in ASD patients. 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 data sets. First, PPI data sets are far from complete (Hart et al. 2006), especially when only highly confident edges are considered (Supplemental Table 1).

Second, PPI data sets are biased due to emphases on published literature (Hakes et al. 2008). Also, PPI networks are often limited to a single splicing isoform (Corominas et al. 2014). Coexpression data are less biased and therefore the incorporation of coexpression with PPI helps to mitigate such limitations.

In contrast, Parikshak et al. (2013) applied the WGCNA (weighted gene coexpression network analysis) method (Horvath et al. 2006) to find a set of large, mutually exclusive modules (average size > 600 genes) that are highly coexpressed 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.

© 2015 Hormozdiari et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.
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 coexpression or PPI subnetworks for ASD.

Other studies (e.g., Gulsuner et al. 2013; Willsley 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, Willsley et al. (2013) were able to suggest specific neurodevelopmental subtraits 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 coexpression across different brain regions and different time points. This 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 at finding pathways disregulated in cases versus controls based on differential expression analysis (Ulitsky et al. 2010; Chowdhury et al. 2011).

Recent studies on ASD and schizophrenia have found that genes with putative loss-of-function (LoF) mutations in cases not only are more densely connected in PPI networks but also demonstrate higher coexpression with each other (O’Roak et al. 2012b; Gulsuner et al. 2013). Motivated by this observation, Willsley et al. (2013) developed a novel method that simultaneously integrates in-...
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; Supplemental Material).

MAGI was applied to exome de novo mutation data sets (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 (Gulsuner et al. 2013; Fromer et al. 2014).

**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, GRN1, and DLG3) is significantly enriched in the KEGG long-term potentiation signaling pathway ($P < 3.2 \times 10^{-2}$, Bonferroni correction) and the GO annotation 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 AIMS (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 coexpression data simultaneously to discover sets of genes that have a similar function with disease relevance. By combining the seed pathways (Fig. 1) and reiterating module discovery, we identified two distinct disjoint modules (M1 and M2) associated with autism and ID (Figs. 2 and 3, respectively).

For module 1 (Fig. 2), M1_Best consists of 48 genes, while M1_Extended is comprised of 80 genes corresponding to 28 LoF and 39 missense mutations (Fig. 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 (Fig. 2B,C; Supplemental Fig. 23), suggesting that indeed de novo mutations observed in cases can be clustered into a highly connected and coexpressed 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 (Fig. 2D,E; Table 2). M1_Extended is strongly associated with chro-

![Figure 1](https://example.com/figure1.png)

**Figure 1.** Flowchart of MAGI. Given PPI and coexpression networks 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.

| **Table 1.** Phenotype distribution of de novo mutations/modules | Samples | De novo mutations | Significant modules |
|---|---|---|---|
| **Samples** | **Missense** | **LoF or indels** | **M1_Extended**<sup>a</sup> | **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 | — | — |

<sup>a</sup>Number of genes (and in parentheses number of distinct samples with de novo mutations) in the modules.

---

Hormozdiari et al.
Figure 2. The discovery of integrated gene networks for ASD

A

B

Proband M1

D

Siblings M1

C

Proband M1

E

Siblings M1

Figure 2. (Legend on next page)
matin remodeling (Supplemental Fig. 27), while both M1_Best and M1_Extended are significantly enriched for Wnt (P < 3.2 × 10^{-3} and P < 7.1 × 10^{-3}) and the Notch (M1_Extended; P < 1.4 × 10^{-3}) signaling pathways (after Bonferroni correction). We also note overlap with other SWISSNP (PBMR1, ARID1B, SMARC1, and SMARC2C), NCO, and TCF complexes (Fig. 2A): multiprotein complexes critical for normal neuronal development (Ille and Sommer 2005; De Ferrari and Moon 2006; 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 with other probands with the same types of mutations (P < 0.016). The M1_Extended module genes are highly expressed for all distinct subterritories of the brain early during development, with peak expression between 8–16 pcw (Fig. 4B; Supplemental Figs. 29, 31a). We confirmed this pattern using expression data from the Gene Atlas (Gene Expression Omnibus accession number GSE11333), which showed the highest expression in fetal brain compared to the adult brain or all other brain tissues (Supplemental Fig. 32).

After excluding mutations and genes associated with M1_Best, we repeated MAGI and discovered a second module (Fig. 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 data sets but are predicted to be highly related by coexpression and PPI data. The module is enriched for synaptic plasticity genes (CALM1, GRIN2A, GRIN2B, MAPK1, PRKCB, and RPS6KB3; P < 1.6 × 10^{-5} after Bonferroni correction) (Supplemental Fig. 28) and for known neurodevelopmental disease genes (Table 2; a sevenfold enrichment, P = 1.01 × 10^{-3}). 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 significantly higher level of expression in the fetal brain (P < 0.001) 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 post-natally (Fig. 4B; Supplemental Figs. 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 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) 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 (TCFC4, SUPT16H, CASK, GPS1, HDAC9, and DHX9) and two genes in M2 (KCNJ1 and three missense mutations in STXBP1) (Fig. 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 (Willsye et al. 2013), namely, (1) primary visual cortex–superior temporal cortex, or V1-STM cluster; (2) prefrontal and primary motor-somatosensory cortex, or PFC-MSC cluster; (3) striatum (STR), hippocampus (HIP), and amygdaloid (AMY); and (4) mediiodorsal nucleus of thalamus–cerebellar cortex, or MD-CBC cluster. For each of these regions, we calculated the coexpression (Pearson correlation coefficient) between every pair of genes considering the time points from 8 pcw to 1 yr after birth, and ran 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 (Supplemental Figs. 34, 37). Modules produced for the V1C-STM cluster and STR, HIP, and AMY were most similar (>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 (Fig. 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

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 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 coexpressed (Pearson correlation coefficient r > 0.37, i.e., the genes are included in the top 5% of gene pair coexpression during brain development). The innermost circle contains genes detected in >99% of the suboptimal solutions. Subsequent concentric circles display genes found in >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). For a force-directed layout of this module, see Supplemental Figure 3B. (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.

146 Genome Research www.genome.org
are associated with lower IQ but less likely to be found in ASD probands with IQ $\geq 70$. While it is clear that M1 genes may be associated with either phenotypic category (Fig. 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 Supplemental

**Figure 3.** Module M2. MAGI reiteration after M1 components removed. For details, see legend of Figure 2. For a force-directed layout of this module, see Supplemental Figure 3.
| Modules (ASD + ID) | Genes in OMIM + AISS | Enrichment | C-score | IQ |
|------------------|----------------------|------------|---------|----|
| **OMIM/AISS enrichment** | | **P-value** | Inside module; outside module | **P-value** | Inside module; outside module | **P-value** |
| **M1_Best (n = 48)** | ARID1B², CASK², CHD8³, CREB1, CTNNB1², DYSK1A², EP300, EPHB2, HDAC2, MECP²², SMC3, TCF4, TUBA1A | 9.6| 2.02× | 4.2× 10⁻⁵ | 2.02× | 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)** | ARID1B², CASK², CHD8³, CREB1, CREBBP, CTNNB1², DYSK1A², EP300, EPHB2, GTF2IRD1, HDAC2, MECP²², SIRT1, SMAD4, SMARCBI², SMC3, TCF4, TUBA1A | 6.67| 2.3× | 1.0× 10⁻⁴ | 2.3× | 0.002 | n = 39, median = 20.6; n = 654, median = 16.93 | 4× 10⁻⁶ | n = 63, low/high ratio = 1.1; n = 485, low/high ratio = 0.732 | 0.138 |
| **M2_Best (n = 21)** | CALM1, DLG3², DLG4, DLGAP1, GRIN²A³, GRIN²B³, HTRA²A, KCNNMA1, MAPK1, RIMS1, RPS6KA3², SHANK2³, STX1A, STXB1², SV2B, SYNAP1², SYT1 | 25.8| 7.1× | 4.7× 10⁻⁹ | 7.1× | 4.71× 10⁻¹³ | n = 11, median = 24.6; n = 682, median = 17.025 | 2× 10⁻⁴ | n = 17, low/high ratio = 3.25; n = 531, low/high ratio = 0.735 | 0.006 |
| **M2_Extended (n = 24)** | CALM1, DLG3², DLG4, DLGAP1, GRIN²A³, GRIN²B³, HTRA²A, KCNNMA1, MAPK1, RIMS1, RPS6KA3², SHANK2³, STX1A, STXB1², SV2B, SYNAP1², SYT1 | 27.1| 6.4× | 2.2× 10⁻⁸ | 6.4× | 6.2× 10⁻¹² | n = 11, median = 24.6; n = 682, median = 17.025 | 2× 10⁻⁴ | n = 17, low/high ratio = 3.25; n = 531, low/high ratio = 0.735 | 0.006 |
| **Sib1_Best (n = 43)** | AKT1, ATRX², DNMT1, EP300, HDAC2, MBDS, SIRT1, SMC3 | 1.6× | 1.59× | 0.45 | 0.13 | n = 19, median = 18.08; n = 347, median = 15.53 | 0.66 | — | — |
| **Sib1_Extended (n = 59)** | AKT1, ATRX², AXIN1, DNMT1, EP300, HDAC2, MBDS, MECP²², SIRT1, SMC3 | 2.5× | 1.6× | 0.19 | 0.092 | n = 19, median = 18.08; n = 347, median = 15.53 | 0.66 | — | — |

List of genes in different modules with known associated neurodevelopmental diseases (using OMIM and AISS). The superscript above each gene represents the annotated disease from OMIM: 1, autism; 2, intellectual disability; 3, epilepsy.

P-values 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 U-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.

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).
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 siblings’ and controls’ missense mutations. (B) Average normalized expression of all brain subtissues for M1_Extended and M2_Extended during brain development. Error bars represent 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 schizophrenia, and the first module of epilepsy. The $p$-value reported for each disease is the maximum $p$-value of the three null models.
Data 1). The epilepsy M1_Extended is enriched for genes associated with SNARe interaction and vesicular transport pathways with $P < 0.03$ after Bonferroni 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 > 61% (22/36) overlap, representing a 9.5-fold enrichment in FMRP targets. In agreement with a recent publication (Framer 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 (Fig. 5A,B) and have indeed mutated sporadically in different disease studies. In addition, one gene, POGZ, shows very high coexpression 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, coexpression, and gene-ontology networks related to neurodevelopmental and neuropsychiatric diseases is an active area of research (Gilman et al. 2011; Sakai et al. 2011; Voineagu et al. 2011; Ben-David and Shifman 2012). MAGI differs from previous approaches in that it simultaneously considers both PPI and coexpression 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 (Supplemental Figs. 12–21; Supplemental Tables 2–4).

Modules that were generated using PPI-based approaches seem to have a substantial number of gene pairs with low coexpression during brain development. Comparison to other recently reported coexpression-based networks (Parikhshak et al. 2013; Willsey et al. 2013) shows overlap as well as substantial differences in network membership: Our predicted ASD modules are smaller (Supplemental Table 2) and show a greater enrichment with known neurodevelopmental disease genes (Supplemental Figs. 14, 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 coexpression 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 coexpressed with modules M1 and M2 (Supplemental Fig. 41). An area of future development, then, will be to extend membership (with modified penalty parameters) to genes that are highly coexpressed 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 NCOA2/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 yr) 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 compared to those that are intellectually 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 (Fig. 4A). 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 Robinow-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 (Supplemental Fig. 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 (Froner 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 data sets

The de novo mutations were collected from nine different studies: four ASD (lossisov 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 (Gulsuner et al. 2013; Froner et al. 2014). LoF mutations are defined as likely gene-disruptive or loss-of-function mutations observed in the cases. For controls, we used unaffected siblings and normal trios from lossisov et al. (2012), Neale et al. (2012), O’Roak et al. (2012b), Sanders et al. (2012), and Gulsuner et al. (2013). The PPI network is composed 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 coexpression network was constructed as a complete graph using the normalized RNA-seq RPKM values from the BrainSpan Atlas (http://www.brainspan.org). A Pearson correlation coefficient $r$ was calculated between every pair of genes across all the subtissues and time points, and $r^2$ was 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 coexpressed, and are enriched in patient mutations. Then in the second step, pathways that are highly connected in the PPI and also coexpressed 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 coexpression (top 5% in the coexpression 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 $h = \log{|V|}$ 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: (1) random coloring of the graph’s nodes with $h$ different colors and (2) 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 $O(d^h)$ iteration 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 ($\Delta$) found in controls (for more details, see Supplemental Material). In practice, we run 1000 iterations for each threshold ($\Delta = 0, 1, 2, 3$) and possible path length ($h = 5, 6, 7, 8$) 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 (for exact definition of the constraints, see Supplemental Material).

Clustering seed pathways

Seeds are merged into high-scoring clusters that satisfy a stringent set of constraints (Supplemental 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 Iodéké 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 (1) random gene removal, (2) random gene addition, and (3) 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), and input files are publicly available for download in the Supplemental Material and at http://eichlerlab.gs.washington.edu/MAGI/.
Competing interest statement

E.E. is on the scientific advisory board for DNAnexus, Inc.

Acknowledgments

This work was supported by the Simons Foundation Autism Research Initiative (SFARI) 303241 and NIH/NIMH R01MH101221 to E.E.E. O.P. was supported by the Howard Hughes Medical Institute. We are thankful for useful discussions with Drs. Ron Shamit, Eran Halperin, and Phuong Dang regarding the computational definition of the problem. We are also grateful to T. Brown for help with manuscript preparation.

References

Allen AS, Berkovic SF, Cossette P, Delanty N, Dlugos D, Eichler EE, Epstein MP, Glauser T, Goldstein DB, Han YJ, et al. 2013. De novo mutations in epileptic encephalopathies. Nature 501: 217–221.

Alon N, Yuster R, Zwick U. 1995. Color-coding. J Comput Biol 2(1–2) 131–137.

Allen AS, Berkovic SF, Cossette P, Delanty N, Dlugos D, Eichler EE, Epstein MP, Glauser T, Goldstein DB, Han YJ, et al. 2013. De novo mutations in epileptic encephalopathies. Nature 501: 217–221.

Alon N, Yuster R, Zwick U. 1995. Color-coding. J Comput Biol 2(1–2) 131–137.

Araujo AM, Coutinho-Filho A, Bouzamondo A, Tordjman M, Lechat P, Mottron L, Cohen D. 2008. Epilepsy in autism is associated with intellectual disability and gender: evidence from a meta-analysis. Biol Psychiatry 64: 577–582.

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

Chowdhury SA, Nibbe RK, Chance MR, Koyuturk M. 2011. Subnetwork De Ferrari GV, Moon RT. 2006. The ups and downs of Wnt signaling in glioblastoma. Trends Neurosci 29: 574–581.

Corominas R, Yang X, Lin GN, Kang S, Shen Y, Ghamsari L, Broly M, Rodriguez M, Tamm S, Pachter LS. 2010. Dynamic and non-obvious birth of gene networks in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. N Engl J Med 363: 1921–1929.

De Ferrari GV, Moon RT. 2006. The ups and downs of Wnt signaling in glioblastoma. Trends Neurosci 29: 574–581.

Darnell JC, Van Driesche SJ, Zhang C, Hung KY, Mele A, Fraser CE, Stone EF, Coe BP, Witherspoon K, Rosenfeld JA, van Bon BW, Vulto-van Silfhout AT, Ille F, Sommer L. 2005. Wnt signaling: multiple functions in neural development. Cell Mol Life Sci 62: 1100–1108.

Darnell JC, Van Driesche SJ, Zhang C, Hung KY, Mele A, Fraser CE, Stone EF, Coe BP, Witherspoon K, Rosenfeld JA, van Bon BW, Vulto-van Silfhout AT, Ille F, Sommer L. 2005. Wnt signaling: multiple functions in neural development. Cell Mol Life Sci 62: 1100–1108.

Eisenberg MS, Casari G, Muratet M, Goyette P, Bienfait M, Fritsch L, Kunkel LM, et al. 2003. GeneCards: a comprehensive gene database with functional and computational annotations. Genome Res 13: 1980–1990.

Eisenberg MS, Casari G, Muratet M, Goyette P, Bienfait M, Fritsch L, Kunkel LM, et al. 2003. GeneCards: a comprehensive gene database with functional and computational annotations. Genome Res 13: 1980–1990.

Eisenberg MS, Casari G, Muratet M, Goyette P, Bienfait M, Fritsch L, Kunkel LM, et al. 2003. GeneCards: a comprehensive gene database with functional and computational annotations. Genome Res 13: 1980–1990.
Ronan JL, Wu W, Crabtree GR. 2013. From neural development to cognition: unexpected roles for chromatin. Nat Rev Genet 14: 347–359.

Sakai Y, Shaw CA, Dawson BC, Dugas DV, Al-Mohtaseb Z, Hill DE, Zoghbi HY. 2011. Protein interactome reveals converging molecular pathways among autism disorders. Sci Transl Med 3: 86ra49.

Sanders SJ, Murtha MT, Gupta A, Murdoch JD, Raubeson MJ, Willsey AJ, Erkan-Sencicek AG, Dilullo NM, Parikshak NN, Stein JL, 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 NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504.

Stessman HA, Bernier R, Eichler EE. 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 Res 39: D561–D568.

Tennessen JA, Bigham AW, O’Connor TD, Fu W, Kenny EE, 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. Nat Genet 44: 376–378.

Ulitsky I, Shamir R. 2007. Identification of functional modules using network topology and high-throughput data. BMC Syst Biol 1: 8.

Ulitsky I, Krishnamurthy A, Karp RM, Shamir R. 2010. DEGAS: de novo discovery of dysregulated pathways in human diseases. PLoS ONE 5: e13367.

Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor RM, Blencowe BJ, Geschwind DH. 2011. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature 474: 380–384.

Willsey AJ, Sanders SJ, Li M, Dong S, Tebbenkamp AT, Muhle RA, Reilly SK, Lin L, Fertuzzinos S, Miller JA, et al. 2013. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell 155: 997–1007.

Received May 22, 2014; accepted in revised form October 31, 2014.