Convergent and distributed effects of the schizophrenia-associated 3q29 deletion on the human neural transcriptome

Esra Sefik[1,4]*, Ryan H. Purcell[2,3]*, Megan Merritt-Garza[2,3], Sridhar Karne[2,3], Jessica Randall[5], The Emory 3q29 Project[1], Elaine F. Walker[4], Gary J. Bassell[2,3], Jennifer G. Mulle[1,6#]

[1]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA

[2]Department of Cell Biology, Emory University School of Medicine, Atlanta, GA, USA

[3]Laboratory of Translational Cell Biology, Emory University School of Medicine, Atlanta, GA, USA

[4]Department of Psychology, Emory University, Atlanta, GA, USA

[5]Emory Integrated Computational Core, Woodruff Health Sciences Center, Emory University, Atlanta, GA, USA

[6]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, USA

*These authors contributed equally

[#]Correspondence: jmulle@emory.edu

Abstract

The 1.6Mb 3q29 deletion (3q29Del) confers >40-fold increased risk for schizophrenia (SZ) and is also a risk factor for intellectual disability (ID) and autism spectrum disorders (ASD). No single gene in this interval is definitively associated with SZ, ID, or ASD, prompting the hypothesis that neurodevelopmental sequelae emerge upon loss of multiple functionally-connected genes. However, 3q29 interval genes are unevenly annotated and the impact of the 3q29Del on the human neural transcriptome is not known. To formulate unbiased hypotheses, we engaged a systems-level approach using the adult human cortical transcriptome to predict novel biological roles, functional inter-relations and disease associations for individual 3q29 genes. Weighted gene co-expression network analysis showed that the 21 protein-coding genes located in the 3q29 interval segregate into seven clusters, demonstrating both convergent and distributed effects across the transcriptomic landscape. Analysis of 3q29 gene-containing clusters revealed nervous-system specific functions, such as regulation and organization of synaptic signaling, as well as core aspects of cell biology, including transcriptional regulation, chromatin remodeling, protein modifications, and mitochondrial metabolism. Top network neighbors of 3q29 interval genes show significant overlap with known SZ, ASD and ID-risk genes, suggesting that 3q29Del biology is relevant to idiopathic disease. By leveraging “guilt by association”, we propose nine 3q29 genes, including one hub gene, as prioritized drivers of neuropsychiatric risk. To test these predictions, we generated the first human cellular model of 3q29Del syndrome and show that neural progenitor cell (NPC) lines derived from 3q29Del carriers empirically validate network-derived targets, highlighting both the biological relevance of this *in silico* analysis and this novel NPC model of 3q29Del. These results provide foundational information to formulate testable hypotheses on causal drivers and mechanisms of the largest known genetic risk factor for SZ.

Introduction

Copy number variants (CNVs) offer tractable entry points to investigate the genetic contributions to risk for complex neuropsychiatric diseases. The recurrent 1.6Mb deletion of the 3q29 interval (3q29Del) is robustly associated with schizophrenia spectrum disorders (SZ) [1-4] and is the strongest known risk allele for the disease with an estimated odds ratio > 40 [5]. Because of this large effect size, and its defined genomic boundaries, 3q29Del is an attractive target for investigations into biological pathways relevant for SZ and related neurodevelopmental and neuropsychiatric illnesses. However, it is not currently known which gene(s) in this interval are responsible for the increased susceptibility.

Like most CNVs associated with neuropsychiatric disorders, 3q29Del produces phenotypic pleiotropy. As compared to the general population, it is associated with a 16-fold and 34-fold increased risk for autism spectrum disorders (ASD) in males and females, respectively [6]. Emerging evidence suggests that attention deficit hyperactivity disorder (ADHD), intellectual disability (ID), and anxiety disorders are also common in 3q29Del carriers [6-8]. Despite the strength of this genetic association, the biological mechanisms of the neuropsychiatric risk associated with this locus are unclear. Indeed, the individual genetic drivers within the 3q29Del interval have not been identified. No single gene within the interval has been definitively associated with any of the aforementioned neuropsychiatric sequelae, prompting the hypothesis that haploinsufficiency of more than one of the 21 protein-coding genes located in the 3q29 interval is required [9]. However, most 3q29 genes are poorly annotated with limited functional data in existing gene ontology knowledge bases. The paucity of information regarding the biological roles of these genes severely hampers the development of evidence-based hypotheses directed at deciphering the link between 3q29Del and neuropsychiatric pathology. In addition, no transcriptomic investigation of 3q29Del in humans has been reported previously and it is unclear what impact hemizygous loss of these genes might have in the nervous system.

Among the 21 protein-coding genes of the 3q29Del locus, several have been proposed as drivers of the behavioral phenotypes [10], yet the clinical evidence for their individual association with neuropsychiatric disease remains suggestive. *DLG1* produces a synaptic scaffold protein (aka SAP97) that interacts with AMPA and NMDA-type glutamate receptors [11-14], the latter of which have been hypothesized to be involved in SZ pathogenesis [15]. However, the mouse-specific behavioral phenotypes of 3q29Del are not recapitulated by haploinsufficiency of *Dlg1* alone [16]. Another prominent candidate, *PAK2* encodes a brain-expressed protein

kinase that is involved in cytoskeletal dynamics [17] and dendritic spine morphology [18]. Both *DLG1* and *PAK2* are homologous to genes linked to intellectual disability [19, 20] and evidence from *Drosophila* indicates that haploinsufficiency of both genes may be required for synaptic defects rather than each individually [21]. A recent study generated combinatorial knockdowns of select homologs of 3q29 interval genes in *Drosophila* and *Xenopus laevis* and proposed that a component of the nuclear cap-binding complex, *NCBP2* [22], mediates neurodevelopmental defects in 3q29Del syndrome [9]. However, in the 3q29Del mouse model, NCBP2 is not decreased at the protein level in brain tissue, dampening enthusiasm for this gene as a causal element [23]. It remains unclear which genes and their potential interactions are responsible for 3q29-associated phenotypes.

Overall, the literature provides very little additional guidance regarding what role other genes in the 3q29Del locus may have in neuropsychiatric disease risk or how heterozygosity of multiple genes may lead to clinical phenotypes in humans. Recent advances have shown that abnormal phenotypes linked to gene mutations are often detected in biological processes that were not initially anticipated or hypothesized [24], suggesting that the few existing annotations of 3q29 interval genes likely reflect an incomplete functional profile of this locus. Finally, annotation-driven approaches to identify candidate genes are notoriously susceptible to publication bias [25].

To avoid bias and address the knowledge gap for 3q29 interval genes, we employed a systems biology approach, rooted in graph theory [26-28], to generate evidence-based predictions that relate individual 3q29 genes to specific biological functions and disease phenotypes. Accumulating evidence indicates that genes work in conjunction, rather than in isolation, to realize most cellular functions that determine the behavior of complex biological systems [9, 21, 29]. In this holistic context, functional information about well-characterized genes can be leveraged to infer novel biological functions and disease associations for genes that are poorly characterized [30-32]. This *in silico* approach to the investigation of unknown biology stems from the “guilt by association” paradigm [33], which is predicated on the assumption that protein products of highly co-expressed (i.e., coordinately transcribed) genes share functionality, forming multi-gene subsystems within a parent biological system [34-36]. Weighted gene co-expression network analysis (WGCNA) [27, 37] is a technique that has been successfully deployed in this context to study how genes jointly affect complex human diseases [38-42]. We sought to apply this paradigm to 3q29Del syndrome to glean biological insights into disease-associated driver genes and functional mechanisms. To assess the validity of these network-based predictions, we generated novel induced-pluripotent stem cell (iPSC) lines from 3q29Del cases and matched controls, differentiated these

cell lines to neural progenitors, and determined RNA expression profiles. Empirical validation of our predictions using biomaterials derived from 3q29Del study subjects highlights the value of this approach, and provides foundational information to formulate rigorous, targeted and testable hypotheses on the causal drivers and mechanisms underlying the largest known single genetic risk factor for SZ.

Results

Unbiased gene co-expression network analysis reveals convergent and distributed effects of 3q29 interval genes across the adult human cortical transcriptome.

Our WGCNA-based unsupervised network analysis approach [27, 37], applied to publicly-available data from the Genotype Tissue Expression Project (GTEx) [43] and including 18 410 protein coding genes, revealed that the protein-coding transcriptome of the healthy adult human prefrontal cortex (PFC) can be organized into a co-expression network of 19 modules (labeled by color) (Fig. 1A, B). The identified modules group genes with highly similar expression profiles into densely interconnected clusters, which likely represent shared function and co-regulation. The resulting module sizes ranged from 43 (steel blue) to 4 746 (green) genes, with an average module size of 1 014 genes, excluding the grey module (151 genes), which contained genes that could not be unequivocally assigned to any module. To ensure the reproducibility of our network analysis results, we tested the preservation of the identified modules in an independent dataset obtained from the BrainSpan Developmental Transcriptome Project [44]. Results from the module preservation and quality analyses revealed the highly replicable and robust nature of the network (Supp. Fig. S6-S7, Supp. Table S2.4), allowing commencement of its interrogation for novel insights into 3q29Del.

The 21 protein-coding genes located in the 3q29 interval were found to cluster into seven modules (Fig. 1B, “3q29 modules”), which include black (size = 1 170 genes; one 3q29 gene), brown (size = 1 972 genes; 4 3q29 genes), dark turquoise (size = 496 genes; 1 3q29 gene), green (size = 4 746 genes; 6 3q29 genes), magenta (size = 1 437 genes; 1 3q29 gene), midnight blue (size = 1 414 genes; 3 3q29 genes), and turquoise (size = 3 319 genes; 5 3q29 genes) modules. The union of all nodes (genes) co-clustering in the identified 3q29 modules amounts to a broad transcriptomic network of 14 554 genes; we refer to this unique subset as the broad 3q29 network. Within the broad 3q29 network, 18 of the 21 3q29 interval genes were found to concentrate into just four modules (brown, green, midnight blue, turquoise) (Fig. 1B), suggesting that the haploinsufficiency of the

3q29 locus may perturb discrete biological processes via multiple hits, cumulatively disrupting redundancy and compensatory resiliency in the normative regulation of cellular functions (Fig. 1F). Simultaneously, leading presumptive candidates *DLG1* (black module) and *PAK2* (dark turquoise module) were found in opposite branches of the network (Fig. 1C), demonstrating the distributed effects of this CNV across the transcriptomic landscape (Fig. 1G).

The identified modules further clustered into three higher level meta-modules (clusters of highly correlated modules), detected as major branches of the resulting eigengene dendrogram (Fig. 1B, D). The midnight blue and black modules were found to co-cluster within the first meta-module, grouping 2 584 total genes, including four 3q29 interval genes. The magenta, green and turquoise modules clustered together within the second meta-module, grouping 9 502 total genes, including 12 3q29 interval genes. Finally, the brown and dark turquoise modules clustered together within the third meta-module, grouping 2 468 total genes, with five 3q29 interval genes. The observed grouping, as well as the segregation, of 3q29 modules into meta-modules represents a higher-order transcriptomic organization of the 3q29 locus, which likely reflects dependencies and interactions between biological pathways involving distinct sets of 3q29 interval genes. This finding further suggests functional convergence within the 3q29 interval. A list of gene-sets for each identified module is provided in Supp. Table S2.2.

Pathway analysis points to functional involvement of the 3q29 locus in nervous-system functions and core aspects of cell biology.

The functional enrichment analysis of individual 3q29 modules showed that the constituent genes of each identified module converge onto canonical biological processes, indicating that these co-expression-based 3q29 modules are biologically meaningful (Fig. 1E). For example, the turquoise module (harboring *CEP19, FBXO45, PIGX, TCTEX1D2, NRROS*) showed overrepresentation of roles specific to the neuronal system (adjusted P = 2.98E-12), including functions in protein-protein interactions at synapses (adjusted P = 7.12E-07), neurexins and neuroligins (adjusted P = 3.58E-06), transmission across chemical synapses (adjusted P = 6.31E-06), neurotransmitter receptors and postsynaptic signal transmission (adjusted P = 9.47E-05), serotonin receptors (REAC:R-HSA-390666, adjusted P = 6.42E-03), and unblocking of N-methyl-D-aspartate (NMDA) receptors, glutamate binding and activation (adjusted P = 1.41E-02). Similarly, the green module (harboring *BDH1, PIGZ,*

PCYT1A, SMCO1, SLC51A, MFI2), revealed overarching functional characteristics related to nervous system function, including neuronal system (adjusted *P* = 7.76E-07), potassium channels (adjusted *P* = 2.99E-03), L1 cell adhesion molecule (L1CAM) interactions (adjusted *P* = 8.69E-03), transmission across chemical synapses (adjusted *P* = 1.66E-02), G protein-coupled receptor (GPCR) ligand binding (adjusted *P* = 1.74E-02), G alpha (q) signaling events (REAC:R-HSA-416476, adjusted *P* = 2.37E-02), recycling pathway of L1 (adjusted *P* = 3.93E-02), and coat protein complex I (COPI)-mediated anterograde transport (REAC:R-HSA-6807878, adjusted *P* = 4.27E-02).

The magenta module (harboring *RNF168*) was predominantly enriched for a role in controlling the modification, turnover and spatiotemporal colocalization of substrates necessary for a variety of intracellular interactions. The top biological pathways that were found to be overrepresented in this module include post-translational protein modification (adjusted *P* = 1.22E-07), gene expression (transcription) (adjusted *P* = 7.98E-06), MHC-I mediated antigen processing and presentation (adjusted *P* = 1.69E-05), antigen processing: ubiquitination and proteasome degradation (adjusted *P* = 2.78E-05), RNA polymerase II transcription (adjusted *P* = 1.05E-04), signaling by TGF-beta family members (adjusted *P* = 2.87E-03), protein ubiquitination (adjusted *P* = 1.14E-02), and sumoylation (adjusted *P* = 1.94E-02). A link was identified between magenta module genes and the initiation of MHC class I (MHC-I)-dependent immune responses, which is driven by a genomic locus that has been implicated in the etiology of SZ [45, 46]. MHC-I antigen presentation has been shown to strictly depend on peptide supply by the ubiquitin-proteasome system to initiate an effective adaptive immune response [47]. Thus, the simultaneous enrichment of these interacting processes in a single module supports the biological relevance of the identified pattern of clustering.

Other 3q29 modules point to less-obvious pathways that may be involved in 3q29 pathology. Pathways that were overrepresented in the black module (harboring *DLG1*) mainly encompass terms related to regulation of gene expression and maintenance of the integrity of the cellular genome. Specifically, the top Reactome biological pathways that were overrepresented in the black module include metabolism of RNA (adjusted *P* = 2.45E-07), processing of capped intron-containing pre-mRNA (adjusted *P* = 7.74E-04), chromatin organization (adjusted *P* = 6.36E-03), mRNA splicing (adjusted *P* = 2.11E-02), post-translational protein modification (adjusted *P* = 2.33E-02), DNA Repair (adjusted *P* = 4.02E-02) and tRNA processing (adjusted *P* = 4.97E-02). Similarly, the midnight-blue module (harboring *UBXN7, SENP5, WDR53*), was enriched for roles in DNA repair

and transcriptional and translational regulation. Intriguingly, an important functional signature that set the midnight blue module apart from the black module was its specific enrichment for a role in cell cycle regulation. Specifically, the top biological pathways that were overrepresented in the midnight blue module include gene expression (transcription) (adjusted *P* = 2.10E-39), metabolism of RNA (adjusted *P* = 1.64E-06), DNA double-strand break repair (adjusted *P* = 9.79E-04), mRNA 3'-end processing (adjusted *P* = 5.49E-04), cell cycle (adjusted *P* = 3.26E-06), mRNA splicing (major pathway) (adjusted *P* = 5.33E-03), cell cycle checkpoints (adjusted *P* = 5.73E-03), transport of mature mRNA derived from an intron-containing transcript (adjusted *P* = 1.53E-02), and transcriptional regulation by the tumor suppressor TP53 (adjusted *P* = 2.10E-02).

Assessment of shared function among constituent genes of the brown module (harboring *NCBP2, TFRC, TM4SF19, ZDHHC19*) revealed primary enrichment for roles in cellular metabolism and mitochondrial function. The top biological pathways that were overrepresented in the brown module include fatty acid metabolism (adjusted *P* = 7.59E-08), mitochondrial fatty acid beta-oxidation (adjusted *P* = 8.33E-05), chondroitin sulfate/dermatan sulfate metabolism (adjusted *P* = 1.42E-03), peroxisomal protein import (adjusted *P* = 1.79E-03), metabolism of fat-soluble vitamins (adjusted *P* = 4.99E-03), peptide hormone biosynthesis (adjusted *P* = 8.72E-03), solute-carrier (SLC)-mediated transmembrane transport (adjusted *P* = 8.94E-03), and diseases associated with glycosaminoglycan metabolism (adjusted *P* = 1.30E-02). Notably, the brown module was also found to be enriched for two canonical KEGG-annotated pathways: the Hippo signaling pathway (adjusted *P* = 4.82E-03) and the Wnt signaling pathway (adjusted *P* = 4.96E-02), which play crucial roles in growth and developmental pathways with substantial cross-talk [48].

Concurrently, the dark turquoise module (harboring *PAK2*) was found to coalesce genes involved in epigenetic regulation of gene expression and in signal transduction pathways that are mediated by Rho GTPases. This latter function is at least in part attributable to *PAK2*, which encodes a known Rho GTPase effector. Intriguingly, this module was also found to be enriched for a functional role in estrogen receptor-mediated signaling, suggesting a potential mechanism for sex specific effects. In particular, the top terms that were overrepresented in the dark turquoise module include estrogen-dependent gene expression (adjusted *P* = 5.08E-06), estrogen receptor (ESR)-mediated signaling (adjusted *P* = 9.52E-06), Rho GTPase effectors (adjusted *P* = 1.35E-05), SIRT1 negatively regulates rRNA expression (adjusted *P* = 2.09E-05), gene silencing by RNA (adjusted *P* = 3.95E-05), nucleosome assembly (adjusted *P* = 4.75E-05), chromatin modifying enzymes

(adjusted *P* = 5.12E-05), signaling by nuclear receptors (adjusted *P* = 8.46E-05), meiotic synapsis (adjusted *P* = 8.73E-05), and epigenetic regulation of gene expression (adjusted *P* = 1.44E-04). Notably, the dark turquoise module was found to share the same meta-module as the metabolism-related brown module, suggesting the involvement of several 3q29 interval genes in a hierarchical transcriptomic control structure that interconnects Rho GTPase-mediated signaling cascades and estrogen-regulated signal transduction pathways with metabolic regulation.

Taken together, functional characterization of the 3q29 modules point to novel mechanisms of shared or synchronized action for sets of 3q29 interval genes that cluster in the same network module and further coalesce in the same meta-module. Simultaneously, the variety of biological pathways that were found to be enriched in 3q29 modules suggests distributed involvement of this locus in not only nervous-system specific functions, such as regulation and organization of synaptic signaling and components, but also in core aspects of cell biology, including mitochondrial metabolism, transcriptional regulation, protein modifications, and cell cycle regulation. For more detail, see Supp. Information Section B: Extended Results and Supp. Table S2.5.

UBXN7 is a highly connected cortical hub-gene predicted to play a crucial role in the neurodevelopmental and psychiatric sequela of 3q29Del.

Previous work has shown that targeted disruption of a highly-connected “hub” gene has a more deleterious effect on the ability of the network to function and leads to a larger number of phenotypic outcomes than the disruption of randomly selected genes or targeted deletion of less connected genes [49, 50]. Hence, hub genes are often regarded as crucial components of the overall function of their module [51]. Therefore, we sought to measure how strongly connected individual 3q29 interval genes are to their assigned network modules using the eigengene-based module connectivity measure (kME) of each 3q29 interval gene for its respective module (Fig. 1C, Supp. Table S2.3). To generate rigorous WGCNA-based predictions, we adopted a conservative criterion that defines hub genes as nodes with kME > 0.8 ($P < 0.05$), a stringent threshold used by several other studies [52-54]. This analysis revealed that only one 3q29 interval gene is a hub gene: *UBXN7* (kME = 0.84, *P* = 8.33E-30), which encodes a ubiquitin ligase-substrate adaptor [55, 56]. This finding is particularly striking, since *UBXN7* is understudied and poorly annotated, with a total of 13 citations in PubMed as of May 2020.

Furthermore, evaluation of the module membership strengths of 3q29 interval genes revealed that *SMCO1* (kME = 0.11, *P* = 0.25), *SLC51A* (kME = 0.17, *P* = 0.09) and *MFI2* (kME = 0.09, *P* = 0.35) have non-significant kMEs for their assigned module, suggesting poor module connectivity. The mRNA expression summaries obtained from the Human Protein Atlas [57] (HPA) (Supp. Table S2.6) indicate nearly negligible or very low mRNA expression levels for *SMCO1*, *SLC51A*, and *MFI2* in the human cerebral cortex. These data indicate the low abundance of these 3q29 interval genes in our tissue of interest, which likely relates to their peripheral network assignments in our analysis. Consequently, *SMCO1, SLC51A* and *MFI2* were excluded from downstream analysis to ensure accurate refinement of tight network connections formed by 3q29 interval genes.

Nine 3q29 interval genes form transcriptomic subnetworks enriched for known SZ, ASD and IDD-risk genes.

We next sought to identify a refined subset of network targets, along with co-clustering 3q29 interval genes, as the top-neighbor-based 3q29 subnetwork, which represents a distilled subgraph of the broad 3q29 network identified earlier in this study. A top neighbor was defined as any node with an expression profile that shows moderate-to-high pairwise correlation (Spearman's rho (ρ) ≥ 0.5, *P* < 0.05) with a 3q29 interval gene within the same module. Based on this criterion, we reduced 3q29 modules to strongly connected top neighbors of individual 3q29 genes. Our results revealed that several 3q29 interval genes are among the top neighbors of one another within the same module. *FBXO45* (ρ = 0.5, *P* = 5.43E-09) and *PIGX* (ρ = 0.6, *P* = 1.24E-10) were identified as top-neighbors of *CEP19* in the turquoise module. Similarly, *SENP5* and *WDR53* were top-neighbors of each other (ρ = 0.5, *P* = 1.05E-07) in the midnight blue module. This finding further suggests that the correlated activity of subsets of 3q29 interval genes may converge onto the same or synchronized multicomponent biological processes in the adult PFC.

TM4SF19 (0 top neighbors) and *ZDHHC19 (3 top neighbors)* were found to have no or < 5 intra-modular partners in the brown module that met the correlation threshold to qualify as top neighbors. Similar to *SMCO1, SLC51A* and *MFI2,* the mRNA expression summaries obtained from the HPA [57] indicate low abundance of these two 3q29 interval genes in our tissue of interest (Supp. Table S2.6), which likely reflects the reason behind their lack of strongly connected top neighbors in our network analysis. Hence, *TM4SF19* and *ZDHHC19* were excluded from our downstream disease-association analysis, along with *SMCO1, SLC51A* and *MFI2.*

Overall, we identified 5 715 unique top neighbors of 3q29 interval genes combined across seven modules (Supp. Table S3.1). According to the "local hypothesis" proposed by the emerging paradigm of network medicine, the human transcriptome and proteome demonstrate non-random topological characteristics, where disease genes tend to interact with other disease genes and play distinct roles in disrupting the same biochemical process underlying a common pathophenotype [58]. Concordant with this prediction, within this top-neighbor-based network, there are several genes that have been extensively implicated in the etiology of neuropsychiatric disease. These include *MECP2*, *NRXN1*, *GRIN2A*, *GRIN2B*, *CHD8*, *SATB2*, *CNTNAP2*, *FOXP1*, *PTEN*, and *SCN2A*. Motivated by this observation, we sought to ask whether top neighbors of individual 3q29 genes contain a significant overlap with known SZ, ASD or intellectual/developmental disability (IDD)-risk genes.

The probability of a substantive connection between any given node and disease-risk increases with the number of connections identified between a node and network partners with known disease association [58]. Thus, we used a hypergeometric test to gauge the level of overlap between the intra-modular top neighbors of 16 3q29 interval genes and six curated lists of evidence-based IDD [59], ASD or SZ-risk genes [60-64], spanning a wide range of the allele frequency spectrum (Supp. Table S3.2). Hypergeometric test results, corrected for multiple testing, revealed a significant overrepresentation of one or more of these established risk gene-sets among the top neighbors of nine 3q29 interval genes (adjusted $P < 0.05$) (Fig. 2).

In the black module, a substantial proportion of the top neighbors of *DLG1* was found to overlap with known SZ-risk genes from the exonic *de novo* mutations gene-set (overlap size = 18, adjusted P = 1.11E-05). Concurrently, in the midnight blue module, all three constituent 3q29 genes had top-neighbor-based subnetworks enriched for IDD, ASD and/or SZ-risk genes. Particularly, top neighbors of *UBXN7,* a hub gene of the midnight blue module (kME = 0.84, P = 8.33E-30) and the only hub gene within the 3q29 interval, showed a significant overlap with known IDD (overlap size = 14, adjusted P = 3.05E-04), ASD (overlap size = 15, adjusted P = 5.46E-05), and SZ-risk genes from the CommonMind case-control gene-set (overlap size = 45, adjusted P = 4.20E-03). Similarly, top neighbors of *SENP5* (gene-set size = 713) and *WDR53* (gene-set size = 278) were disproportionately represented among known IDD (*SENP5* overlap size = 11, adjusted P = 5.05E-03; *WDR53* overlap size = 6, adjusted P = 1.51E-02) and ASD-risk genes (*SENP5* overlap size = 12, adjusted P = 1.28E-03; *WDR53* overlap size = 6, adjusted P = 1.51E-02) in the midnight blue module. In the green module, only two out of the six constituent 3q29 interval genes were found to have top neighbors that showed a significant

overrepresentation of SZ-risk genes. Specifically, top neighbors of *BDH1* (gene-set size = 1008) and *PIGZ* (gene-set size = 995) were enriched for known SZ-risk genes from the CommonMind case-control gene-set (*BDH1* overlap size = 66, adjusted *P* = 4.49E-06; *PIGZ* overlap size = 58, adjusted *P* = 7.03E-04). In the magenta module, top neighbors of *RNF168* (gene-set size = 556) showed a significant overlap with known IDD (overlap size = 9, adjusted *P* = 1.07E-02) and SZ-risk genes from the CommonMind case-control gene-set (overlap size = 65, adjusted *P* = 5.21E-17). In the turquoise module, only two out of the five constituent 3q29 interval genes had top-neighbor-based networks enriched for known IDD, ASD and/or SZ-risk genes. Specifically, top neighbors of *CEP19* (gene-set size = 1161) were enriched for known IDD (overlap size = 16 , adjusted *P* = 1.64E-03), ASD (overlap size = 15, adjusted *P* = 1.64E-03), and SZ-risk genes from the exonic *de novo* mutations gene-set (overlap size = 29, adjusted P = 3.32E-02). In this module, top neighbors of *FBXO45* (gene-set size = 1101) were also enriched for known IDD-risk genes (overlap size = 14, adjusted *P* = 1.35E-02). Finally, there was no statistically significant evidence for overrepresentation of known disease genes of interest among the top neighbors of 3q29 genes that clustered in the brown or dark turquoise modules.

To evaluate the specificity of the identified disease enrichment patterns, we also tested the top neighbors of 16 3q29 interval genes for overlap with known Parkinson's Disease (PD) [65], late-onset Alzheimer's Disease (AD) [66], and inflammatory bowel disease (IBD) risk genes [67]. These disease phenotypes have no known link to 3q29Del syndrome, thus, genetic risk loci associated with these conditions were considered negative controls in this network analysis. In addition, a large list of common variants associated with height [68] were included in our analysis as a fourth negative control to rule out a potential bias that could be introduced to our analysis by differences in the sizes of curated gene-sets. We found no statistically significant evidence for overrepresentation of AD or IBD-risk genes among the top neighbors of interrogated 3q29 interval genes (Fig. 2). Only the top neighbors of *SENP5* (gene-set size = 713) showed a significant overlap with height-associated genes (overlap size = 30, adjusted *P* = 2.36E-02), and the top neighbors of *NRROS* (gene-set size = 68), which did not show an enrichment for known IDD, ASD, or SZ risk genes, exhibited a small but significant overlap with known PD-risk genes (overlap size = 3, adjusted *P* = 2.00E-02) (Fig. 2).

Overall, 19 out of 96 hypergeometric tests (20%) revealed a significant overrepresentation of SZ, ASD, and/or IDD-risk gene-sets among the same top neighbors, amounting to a proportion that is an order of magnitude larger than that of the negative controls (Fig. 2, Supp. Table S3.3). By contrast, only 2 out of 64 (3%)

hypergeometric tests indicated a significant overlap between the top neighbors of interrogated 3q29 interval genes and negative control gene-sets (Fig. 2, Supp. Table S3.3). The substantial margin observed between these two enrichment ratios supports the high specificity and validity of our network-derived inferences for uncovering biology relevant to 3q29Del. By leveraging the guilt by association principle, we prioritize *BDH1, CEP19, DLG1, FBXO45, PIGZ, RNF168, SENP5, UBXN7, and WDR53*, along with their 284 unique SZ, ASD, and/or IDD-related top neighbors from significant overlap tests as likely drivers of the neurodevelopmental and psychiatric consequences of 3q29Del (Fig. 3A, Supp. Table S3.3). See Supp. Table S3.2 for a full list of curated gene-sets.

Disease-relevant driver genes prioritized by network analysis load onto key biological pathways linked to neuropsychiatric disorders.

To formulate testable hypotheses about the key biological mechanisms linking the 3q29 locus to major neuropsychiatric phenotypes associated with 3q29Del syndrome, we interrogated whether the prioritized driver genes identified in our network analysis assemble into known biological pathways and processes that are annotated in major gene ontology databases. Functional enrichment analysis on the union of our 293 prioritized driver genes (including nine 3q29 genes) (Fig. 3A) revealed their significant overrepresentation in several key biological processes.

Specifically, our findings indicate enrichment of our prioritized driver genes in eight biological pathways annotated by the Reactome and KEGG databases (Fig. 3B). These are axon guidance (adjusted P = 3.64E-03), post-translational protein modifications (adjusted P = 5.24E-03), long-term potentiation (adjusted P = 7.29E-03), diseases of signal transduction (adjusted P = 1.00E-02), regulation of actin cytoskeleton (adjusted P = 1.17E-02), deubiquitination (adjusted P = 2.42E-02), chromatin organization (adjusted P = 3.32E-02), and diseases associated with glycosylation precursor biosynthesis (adjusted P = 4.06E-02). This analysis also revealed the enrichment of our prioritized driver genes for several fundamental biological processes annotated by the Gene Ontology Project (GO:BP), including chromosome organization (adjusted P = 3.81E-09), histone modification (adjusted P = 3.31E-08), cellular component morphogenesis (adjusted P = 4.57E-06), regulation of organelle organization (adjusted P = 6.40E-06), DNA metabolic process (adjusted P = 3.83E-05), regulation of telomere maintenance (adjusted P = 6.62E-05), neuron differentiation (adjusted P = 1.88E-04), protein modification by

small protein conjugation or removal (adjusted *P* = 2.34E-04), neuron projection morphogenesis (adjusted *P* = 4.46E-04), neurogenesis (adjusted *P* = 1.89E-03), chemical synaptic transmission, postsynaptic (adjusted *P* = 2.37E-03), post-embryonic development (adjusted *P* = 2.66E-03), synapse organization (adjusted *P* = 3.20E-03), protein acetylation (adjusted *P* = 5.48E-03), cell surface receptor signaling pathway involved in cell-cell signaling (adjusted *P* = 6.41E-03), and excitatory postsynaptic potential (adjusted *P* = 8.97E-03). We hypothesize that the disruption of one or more of these biological pathways and processes, some of which have been demonstrated to be altered in idiopathic SZ and ASD [2, 69], lie on the casual pathway to neuropsychopathology in 3q29Del syndrome.

The top 20 biological processes and pathways enriched among our prioritized driver genes is shown in Fig. 3B. For clear illustration of our findings, we organized all identified GO:BP terms into a network of related functional annotation categories in Fig. 3C. Detailed results of this functional enrichment analysis are provided in Supp. Table S3.4.

Network-derived targets predict differentially expressed genes in human NPCs derived from 3q29Del carriers.

To test the validity of our WGCNA-derived predictions, we generated induced pluripotent stem (iPS) cell lines from 3q29Del carriers (N=7 lines from 6 individuals) and matched controls (N=6 lines from 6 individuals) (Fig. 4A) and differentiated these lines to neural progenitor cell (NPC) cultures. These lines were characterized at both iPSC and NPC stages; we found no evidence of an effect of 3q29Del on reprogramming efficiency or capacity for neural differentiation (Fig. 4B, Supp. Fig. S9-S10). This approach to empirical validation is predicated on the assumption that perturbation of 3q29 interval gene dosage in neural tissue should lead to the differential expression of the true transcriptomic network partners of 3q29 genes. Following this approach, we tested the enrichment of the network targets identified in this study for differential expression in 3q29Del NPCs compared to matched control lines. This analysis revealed 253 protein-coding differentially expressed genes (DEGs) ($P < 0.05$), 16 of which were identified as 3q29 interval genes (*BDH1, CEP19, DLG1, FBXO45, MFI2, NCBP2, PAK2, PCYT1A, PIGX, RNF168, SENP5, SLC51A, TCTEX1D2, TFRC, UBXN7, WDR53*) (Fig. 3C) (Supp. Table S3.5). The scaled expression of these 16 3q29 interval genes showed a consistent reduction in 3q29Del NPCs that was proportional to gene copy number (Fig. 3D).

All 253 protein-coding DEGs were tested for enrichment of genes corresponding to the three levels of network interconnectedness identified in this study to conjoin the 3q29 interval with the global protein-coding genome of the adult human PFC. These levels correspond to i) the broad 3q29 network (14 544 genes), ii) the top-neighbor-based 3q29 subnetwork (5 736 genes), and iii) the prioritized drivers (293 genes). Hypergeometric tests revealed significant enrichment of the interrogated DEGs for network-derived ties at all three levels of this analysis ($p < 0.05$) (Fig. 4E). Specifically, 206 out of 253 protein-coding DEGs were found to overlap with the broad 3q29 network (P = 2.8E-9). 84 out of 253 protein-coding DEGs were found to overlap with the top-neighbor-based 3q29 subnetwork (P = 3.5E-03). Finally, 13 out of 253 protein-coding DEGs were found to overlap with prioritized drivers predicted to be associated with the neurodevelopmental and psychiatric consequences of 3q29Del (P = 3.0E-5). See Supp. Tables S2.2, and S3.1, 3, and 5 for gene lists at each level.

Finally, to estimate the developmental age of the NPCs generated in this study, we compared the average gene expression profiles of our case and control NPCs to publicly available data obtained from the Allen BrainSpan Atlas (http://www.brainspan.org). Using Spearman's rank correlation analysis, we found that, in both groups, gene expression in NPCs showed the highest correlation with the transcriptomic profile of the human brain during 8_{th} ($\rho = 0.80$, $P < 0.05$) and 9_{th} weeks after conception ($\rho = 0.82$, $P < 0.05$) (Fig. 4F). Hence, we estimate the developmental stage of our NPCs to correspond to the early fetal phase of human brain development, recapitulating previous reports indicating the high resemblance of *in vitro* iPSC-derived NPCs to fetal brain tissue [70]. Given the stark difference between the estimated age of our NPCs and age-at-death for the post-mortem samples used to conduct network analysis, this empirical validation of network-derived targets in this new NPC model of 3q29Del proves even more striking.

Discussion

The 3q29Del is an exciting genetic variant for the study of neuropsychiatric disorders, with a high risk for SZ in particular, and may offer key biological insights to advance our understanding of complex mental illness. Currently, the driver genes and affected biological pathways that link 3q29Del to neuropsychiatric pathology remain unknown. To avoid bias introduced by annotation-based criteria in the formulation of mechanistic hypotheses, we engaged a system-level vantage point, taking into consideration the collective behavior of large ensembles of genes [32], for predicting potential biological roles, functional inter-relations and disease

associations for 3q29 interval genes. We interrogated patterns of co-variation in the expression of this locus in relation to the global protein-coding transcriptome of the healthy human brain. We used publicly available transcriptomic data from the GTEx Project [43] to perform WGCNA [27, 37] on postmortem cortical samples obtained from donors with no known history of psychiatric or neurological disease. Given that precise actions of genes are often dependent on their tissue and age context [44], we focused our analysis on the adult PFC, a brain region most consistently implicated in the pathophysiology of SZ [71], which typically strikes during young adulthood.

We topologically analyzed the resulting network model to identify the modular properties and undirected connectivity patterns of the protein coding elements of the 3q29 interval. To ensure the reproducibility of our network analysis results, we tested the preservation of the modules identified in this study with respect to an independent dataset obtained from the BrainSpan Developmental Transcriptome Project [44]. Ensuing functional characterization of the modules that were found to harbor 3q29 interval genes revealed novel insights into biological pathways and processes that may be perturbed by 3q29Del in the human brain. Evaluation of the network connectivity strengths of 3q29 genes led to the identification of a hub gene in the interval that likely plays an influential role in maintaining systemic equilibrium. We also determined refined sub-networks of genes that are strongly correlated with 3q29 gene expression patterns. We tested these sub-networks for enrichment of risk genes known to be associated with SZ, ASD and IDD, which constitute three major neuropsychiatric phenotypes disproportionately observed in 3q29Del carriers compared to the general population. Ample evidence demonstrates that disease-associated genes tend to be more closely connected to each other than random pairs of nodes in omics-based networks, amounting to a screening approach that has enabled the identification of novel disease risk genes in other applications [72-76]. Supported by these findings, 3q29 gene sub-networks that were found to be enriched for known risk genes of disease relevance led to the prioritization of a subset of 3q29 genes, which we propose to be likely drivers of 3q29Del-associated neuropsychiatric phenotypes. Finally, we developed and characterized a new neural cell model of 3q29Del by generating iPSC-derived NPCs from 3q29 probands and matched controls to assess the validity of our graph-based predictions. Our findings provide foundational information to formulate rigorous, targeted and testable hypotheses on the causal drivers and mechanisms underlying the largest known single genetic risk factor for SZ.

Genomic studies have identified several recurrent CNVs that confer high risk for neuropsychiatric disorders [2, 8]. The challenge going forward is to understand which genes within these loci are the major drivers of risk. The candidate gene era of psychiatric genetics did not yield clear targets into the genetic basis of SZ [77]. Nonetheless, it is still tempting to choose candidate drivers within CNV intervals based on known gene functions. Of all genes in the 3q29 locus, *DLG1* and *PAK2* have been most often proposed as candidate drivers of neuropsychiatric phenotypes [10, 21, 78]. Indeed, a recent literature search indicated that there have been more publications related to these two genes than of all other 3q29Del genes combined (see Supp. Fig. S8). Surprisingly, our analysis does not support inclusion of *PAK2* among candidate drivers for the neuropsychiatric phenotypes associated with 3q29Del syndrome. Instead, our results lend support to the hypothesis that nine other 3q29 interval genes, most of which are largely understudied, may be the primary drivers of the neurodevelopmental and psychiatric consequences of 3q29Del. *BDH1, CEP19*, *DLG1*, *FBXO45, PIGZ, RNF168, SENP5, UBXN7* and *WDR53* were found to form transcriptomic subnetworks that show a significant overlap with known SZ, ASD and/or IDD-risk genes curated from evidence-based literature. Several top neighbors within these tightly connected subnetworks include targets that have been extensively studied in the context of neuropsychiatric disease, including *MECP2*, *NRXN1*, *GRIN2A*, *GRIN2B*, *CHD8*, *SATB2*, *CNTNAP2*, *FOXP1*, *PTEN*, and *SCN2A*. By leveraging the guilt by association principle, we prioritize these nine 3q29 interval genes, along with their 284 SZ, ASD, and/or IDD-related top neighbors from significant overlap tests as primary drivers of the neuropsychiatric phenotype in 3q29Del.

It is currently unknown whether the neuropsychiatric pathology that emerges in carriers of recurrent CNVs, such as 3q29Del, overlaps with that of individuals who share the same clinical diagnosis but do not share the same rare genetic variant. Our findings suggest that, at a system-level, molecular perturbations caused by the hemizygous deletion of select 3q29 interval genes may, in part, overlap, with the genetic etiologies contributing to the manifestation of idiopathic forms of SZ, ASD and ID in the wider population. Disease-relevant driver genes prioritized by our network analysis were found to load onto key biological pathways, such as neurogenesis, neuron differentiation, synapse organization, excitatory postsynaptic potential, long-term potentiation, axon guidance, regulation of actin cytoskeleton, signal transduction, post-translational protein modifications, chromatin organization and histone modification. We hypothesize that the disruption of one or more of these biological pathways and processes, some of which have been demonstrated to be altered in idiopathic SZ and

ASD [2, 69], lie on the casual pathway to neuropsychopathology in 3q29Del syndrome. Given the observed convergence between network targets of 3q29Del genes and known disease risk genes, our results also support the notion that targeted investigation of our evidence-based predictions for affected pathways in 3q29Del may concurrently elucidate mechanisms of relevance to severe neuropsychiatric disease more widely.

To our knowledge, no single gene within the 3q29 interval has been definitively associated with any of the aforementioned neuropsychiatric sequelae, prompting the hypothesis that neuropathology in 3q29Del syndrome emerges upon loss of multiple genes that are functionally connected. Indeed, our unsupervised analysis of the modular organization of the adult human PFC revealed evidence for convergence of 3q29 interval genes into co-expression modules that show overrepresentation of overlapping functions. The 21 protein-coding 3q29 interval genes clustered into seven modules, with 18 concentrated in just four modules. Simultaneously, the seven modules harboring 3q29 interval genes were found to segregate into three higher-level meta-modules, in which, leading presumptive candidates *DLG1* and *PAK2* were distributed into clusters located in opposite branches of the network. These findings suggest that 3q29 interval genes display both convergent and distributed effects across the adult human cortical transcriptome. Thus, rather than functioning as independent non-interacting agents, sets of 3q29 interval genes may have shared and/or synchronized function at both the module and meta-module levels of transcriptomic organization, and likely constitute interacting sources of pathology in 3q29Del syndrome. Our approach here cannot predict whether the consequences of haploinsufficiency are more deleterious when spread across the transcriptomic landscape or when concentrated within only a few discrete modules; yet, we found evidence in our analysis for a potential mechanism of 3q29Del action that works through both of these models. It is conceivable that the consequences of the haploinsufficiency arise through the weakening of multiple distinct pathways that normally provide protective redundancies (distributed model), and/or through multiple insults to a functionally connected module that cumulatively disrupt resiliency (convergent biology model). Consequently, these hypotheses warrant further testing in this context.

A major goal of this study was to infer unknown functions for understudied 3q29 interval genes by leveraging the known functions of well-studied genes co-clustering in network modules. Pathway analysis of modules harboring 3q29 interval genes revealed likely functional involvement of the 3q29 locus in not only nervous-system specific functions, but also in core aspects of cell biology that are non-specific to an organ system. For example, the closely-related black (harboring *DLG1*) and midnight blue (harboring *UBXN7, WDR53, SENP5*) modules

were significantly associated with gene ontology (GO) terms related to regulation of gene expression, chromatin organization and DNA repair. The green (harboring *BDH1, PIGZ, PCYT1A, SMCO1, SLC51A, MFI2)* and turquoise (harboring *CEP19, FBXO45, PIGX, TCTEX1D2, NRROS)* modules were both associated with GO terms related to nervous system development and function, and in particular, regulation and organization of synaptic signaling and components. This finding is surprising because 10 out of 11 3q29 interval genes located in these latter two modules have not been identified as synaptic genes. Similarly, the 3q29 interval genes that clustered in the black and midnight blue modules have not been implicated in gene regulatory pathways or DNA repair. We maintain that biological functions of poorly annotated genes can be inferred through the graph-based modeling of inter-gene relationships employed in this study. Thereby, we predict novel roles for individual 3q29 genes in functions related to synaptic transmission, modulation of neurotransmission, synapse structure and function, mitochondrial metabolism, transcriptional and translational regulation, chromatin remodeling, cell cycle regulation, and protein modification, localization and turnover. We propose that the subset of predicted functions that are non-specific to an organ system likely contribute to global developmental outcomes in 3q29Del.

Within-module analysis of membership strength revealed that *UBXN7* is a hub gene of the midnight blue module, with top neighbors enriched for risk genes associated with all three major neuropsychiatric phenotypes of 3q29Del. This finding suggests that *UBXN7* may play a critical role in maintaining systemic equilibrium in biological pathways and processes that are likely disease relevant. Previous observations have led to the hypothesis that hub nodes of biological networks are associated with human disease [79, 80]. One study has shown that disease-genes, identified from OMIM's Morbid Map of the Human Genome, disproportionately exhibit hub-gene characteristics, with protein products participating in more known protein-protein interactions than that of non-disease genes [81]. Supported by this literature, we predict that (1) *UBXN7* exerts critical influence on a large network of co-expressed genes, and (2) loss-of-function (LoF) mutations in *UBXN7* can cause major dysfunction in the biological pathways involving this gene (indeed, its pLI score = 0.99, i.e. extremely intolerant to LoF [82]). Consequently, we prioritize *UBXN7* as a major driver with likely disease relevance in 3q29Del. *UBXN7* has not been previously linked to neuropsychiatric disorders or proposed as a candidate driver of 3q29Del syndrome. However, *UBXN7* has been reported to regulate the ASD-associated E3 ubiquitin ligase Cullin-3 (*CUL3*) [83], an interaction that may deserve more attention in light of our findings [56]. Additionally, *UBXN7* was recently found to be recruited to ubiquitinated histones with p97 to allow for the re-initiation of

transcription following mitosis: a process that is important in maintaining cellular identity in progenitor pools [84] and may explain how *UBXN7*, a protein typically involved in ubiquitination, may strongly affect gene transcription.

Notably, a recent report identified the *Drosophila* homolog of *NCBP2* as a mediator of neurodevelopmental defects [9]. While we did not find support for inclusion of *NCBP2* among prioritized driver genes, our analysis does not preclude the possibility that *NCBP2* exacerbates other 3q29 locus haploinsufficiencies. Similarly, while the brown (harboring *NCBP2* et al.) and dark turquoise (harboring *PAK2*) modules did not harbor prioritized driver genes, the genes clustering in these modules may moderate phenotypic expressivity. For example, the dark turquoise module was significantly associated with estrogen receptor-mediated signaling and an intriguing emerging feature of 3q29Del syndrome is the markedly reduced sex bias in risk for ASD [6]. Additional studies will be required to assess the drivers of sex-specific phenotypes of 3q29Del syndrome.

We hypothesized that hemizygosity of the 3q29 locus would principally impact the expression of genes in the broad 3q29 network and the top-neighbor-based 3q29 subnetwork identified in this study. We tested this hypothesis by producing and characterizing the first-reported neural cell model of 3q29Del and determining the differential transcriptomic profile of these lines compared with matched controls. The co-expression network described here was built from adult human PFC gene expression data; yet, we sought to validate the identified network by determining its overlap with differentially expressed genes in human NPCs, which most closely match embryonic developmental stages [85]. Notwithstanding this considerable difference in developmental phase, we found highly significant enrichment of network-derived targets among differentially expressed NPC genes in 3q29Del carriers. These results indicate that the transcriptomic network connections formed by the 3q29 locus may be relatively stable through development and that iPSC-derived neural cells will be a valuable model for targeted follow-up studies.

One limitation of the current study is its singular focus on the protein-coding elements of the 3q29 interval summarized at the gene-level. The locus also contains three antisense transcripts, one long noncoding RNA, and one microRNA. Inclusion of non-coding RNAs and splice variants in network construction increases network size by 100-fold [86], which in turn dramatically increases the computational resources necessary for analysis. Although this high computational demand can theoretically be overcome by building a co-expression network in a block-by-block fashion [37], previous work has shown that networks generated by block-wise design reflect only an approximation of networks generated by single-block design, which negatively influences the accuracy

of subsequent module detection [86]. To avoid these shortcomings, and pursue the most parsimonious hypothesis, we limited the scope of this study to protein-coding genes, with expression values summarized at the gene-level. How the non-coding elements of the interval, along with splice variants, integrate into the mechanistic predictions formulated in this study is a question ripe for future investigation.

Now that recurrent, highly penetrant CNV loci have been identified as important risk factors for neuropsychiatric disorders, determination of the component genes driving this risk is the next step toward deciphering mechanisms. We used an unbiased systems biology approach that leveraged the power of open data to infer unknown functions for understudied 3q29 interval genes, and to refine the 3q29 locus to nine prioritized driver genes, including one hub gene, with potential disease relevance. Importantly, this approach can partially overcome barriers to formulation of relevant hypotheses that are introduced by poor annotation of interval genes, without requiring laborious, expensive, and time-consuming experiments to functionally characterize all genes within the interval. Our results reveal the power of this approach for prioritization of putative drivers. Ongoing and future studies will be directed at understanding how these genes work in concert and how multiple haploinsufficiencies confer risk for neuropsychiatric disease.

Materials and Methods

Weighted gene co-expression network analysis (WGCNA)

To interrogate the functional consequences of the 3q29Del at the individual gene level, we employed unbiased weighted gene co-expression network analysis (WGCNA package in R [37, 87], version 1.68) and organized the adult human cortical transcriptome into clusters (modules) of highly correlated genes (nodes). The network was constructed on publicly available RNA-Seq data obtained from the GTEx Project [43], using samples from the PFC (Brodmann Area 9) of male and female adults (age range = 20-79) with no known history of psychiatric or neurological disorder. Details on sample attributes and donor phenotypes are available in Supp. Table S2.1 and Supp. Fig. S1, S3. After processing the data (see Supp. Information Section A: Extended Methods and Supp. Fig. S2), 18 410 protein-coding genes from 107 samples constituted the final dataset for construction of the network.

The single-block pipeline implemented in the WGCNA R package was employed to build a signed and weighted gene co-expression network. Co-expression similarity was defined as the biweight midcorrelation

coefficient between the expression profiles of gene-pairs calculated for all possible comparisons. To capture the continuous nature of pairwise interactions in biological systems and accentuate strong positive correlations, the resulting co-expression similarity matrix was transformed into a signed and weighted adjacency matrix. A soft-thresholding power (β) of 8 was identified as the lowest possible β yielding a power-law degree distribution that approximately fits one with a scale-free network topology, (signed R_2 fit index = 0.8), while maintaining a relatively high mean connectivity (mean k > 100) (Supp. Fig. S4). The rationale behind our choice to establish a scale-free topology is the demonstrated biological relevance of its degree distribution as a unifying property of many biological networks in nature, where a few hub nodes have many connections while most nodes have few connections [88-91].

A signed adjacency matrix was chosen over an unsigned adjacency matrix to re-scale the underlying pairwise correlations from the [−1, 1] interval to the [0, 1] interval, as opposed to treating negative correlations as positive. The resulting signed and weighted adjacency matrix was transformed into a topological overlap (TO) matrix to capture not only the correlation between pairs of genes but also the connections among "neighborhoods" of genes[92, 93]. We summarized the gene expression profiles of individual modules by eigengenes [37], identified as the first principal component of the expression data in a given module. Construction of eigengenes amounted to a network-based data reduction method that served as a means to effectively correlate entire modules. Leveraging this approach, we amalgamated modules with very similar expression profiles to eliminate spurious assignment of highly co-expressed genes into separate modules. To this end, we conducted average linkage hierarchical clustering of module eigengenes (ME) on a correlation-based dissimilarity metric and merged modules that are strongly correlated (cut height = 0.2) (Supp. Fig. S5). For detailed methodology and results of the module preservation and quality analyses, see Supp. Information Section A: Extended Methods, Section B: Extended Results and Supp. Fig. S6, S7.

Functional characterization of network modules harboring 3q29 genes

Co-expression modules found to harbor at least one 3q29 interval gene ("3q29 modules") were subjected to functional enrichment analyses, run on the g:Profiler webserver (http://biit.cs.ut.ee/gprofiler; Ensembl version 96, Ensembl Genomes version 43) by using GO:BP terms, and Reactome (REACT) and Kyoto Encyclopedia of Genes and Genomes (KEGG) biological pathways. Top 10 biological pathways (REACT) found to be enriched

in individual 3q29 modules were ranked by statistical significance level (adjusted-p< 0.05) and are shown in Fig, 1E. Detailed information on methodology is provided in Supp. Information Section A: Extended Methods.

Identification of prioritized driver genes

Six gene-sets were curated for this purpose from the following sources: 1) 93 IDD-risk (Deciphering Developmental Disorders Study [59]); 2) 86 “high-confidence” ASD-risk genes categorized by the Simons Foundation Autism Research Initiative (SFARI); 3) 651 SZ-related genes identified by Meng et al. (2018) [62] to demonstrate significant differential expression in the postmortem brain tissue of SZ cases/controls, identified from the PsychENCODE BrainGVEX dataset [63]; 4) 636 SZ-related genes shown by Fromer et al. (2016) to demonstrate significant differential expression in the postmortem dorsolateral PFC tissue of SZ cases/controls, identified from the CommonMind Consortium dataset [60], 5) 340 SZ-risk genes adjacent to SZ-associated genetic loci, identified by the most recent genome-wide association study (GWAS) [64] conducted by the Psychiatric Genomics Consortium (PGC); 6) 290 SZ-risk genes with exonic *de novo* mutations, identified via the Neuropsychiatric Disorder *De Novo* Mutations Database [61]. Four negative control gene-sets were curated from the following sources: 1) 25 AD-risk genes identified by the largest published GWAS meta-analysis conducted by the Genetic and Environmental Risk in AD/Defining Genetic, Polygenic and Environmental Risk for AD Consortium (GERAD/PERADES) [66]; 2) 67 PD-risk genes identified by the largest published GWAS meta-analysis conducted by the International PD Genomics Consortium [65]; 3) 98 IBD-related genes identified by the International IBD Genetics Consortium as “strong positional candidate genes” in GWAS-identified risk loci [67]; 4) 479 height-associated genes identified by the largest published GWAS meta-analysis conducted by the Genetic Investigation of Anthropometric Traits Consortium (GIANT) [68]. The background list for the hypergeometric test was determined as the total number of unique genes used to conduct WGCNA. The *GeneOverlap* package [94] in R (version 1.20.0) was used for this analysis. The hypergeometric p-values obtained by overlap analysis were corrected for multiple testing using the Benjamini-Hochberg procedure (n = 10 gene-sets). 3q29 genes whose top-neighbor-based subnetworks were found to show a significant overlap with SZ, ASD and/or IDD risk genes (adjusted $P < 0.05$) were prioritized as driver genes, along with their SZ, ASD, and/or IDD-related top neighbors from the corresponding enriched disease gene-set.

Experimental validation of WGCNA-based predictions in iPSC model of 3q29Del

Four cases with 3q29Del were ascertained from the Emory 3q29 Project registry (3q29deletion.org) [7]. Participants and their families traveled to Atlanta, GA, where phenotypic assessments were conducted using gold-standard diagnostic instruments administered by expert clinicians as part of a comprehensive and systematic phenotyping protocol developed by the Emory 3q29 Project[95]. This study has been approved by the Emory University Institutional Review Board and all participants (or legal guardians) provided informed consent. Blood samples from four 3q29Del probands (lines 19-1001, 20-1031, 21-1031, 22-1031) were collected for establishment of iPSC lines. Two additional 3q29Del cases were ascertained from RUCDR Infinite Biologics (Piscataway, NJ; lines 00C04409, 04C38098), originally phenotyped as part of the Molecular Genetics of SZ study[96]. Controls were recruited to be age and sex matched to cases. See Supp. Table S1.3 for full sample information. All deletion carriers had a clinically confirmed diagnosis of 3q29Del.

Whole blood samples of 5-10 mL were collected in EDTA Vacutainer test tubes and processed for the isolation of erythroid progenitor cells (EPCs) using the Erythroid Progenitor Kit (StemCell Technologies). EPCs were reprogrammed using Sendai particles (CytoTune-iPS 2.0 Reprogramming kit, Invitrogen) and plated onto Matrigel coated six-well plates (Corning). Cultures were transitioned from erythroid expansion media to ReproTesR (StemCell Technologies) and then fed daily with ReproTesR until clones were isolated. iPSCs were maintained on Matrigel coated tissue culture plates with mTeSR Plus (StemCell Technologies). Lines 00C04409, 04C38098, and MH0200876 were acquired from RUCDR Infinite Biologics (Piscataway, NJ) and were cultured and characterized by the authors at Emory University.

Cell lines were characterized for stem cell markers by RT-PCR and immunocytochemistry after at least 10 passages in culture. Total RNA was isolated from each cell line with the RNeasy Plus Kit (Qiagen) according to manufacturer's protocol. mRNA was reverse transcribed into cDNA using the High Capacity cDNA Synthesis Kit (Applied Biosystems). Expression of pluripotency genes *OCT4*, *SOX2*, *REX1* and *NANOG* was determined by RT-PCR. Sendai virus inactivity was confirmed using Sendai genome specific primers. For immunocytochemistry, cells were plated and cultured on Matrigel coated Nunc Permanox chamber slides. The cells were fixed in 4% paraformaldehyde and stained with primary antibodies for the common stem cell transcription factors Nanog (R&D Systems), Sox2 (Invitrogen) and Oct4 (Cell Signaling) as well as stem cell surface markers Tra-1-81(Millipore), Tra-1-61 (Millipore) and SSEA4 (Cell Signaling) (Supp. Fig. S9). Full

antibody information is provided in Supp. Table S1.1. Images were acquired on an Olympus FV-1000 confocal microscope or a Keyence BZ-X810 fluorescence microscope under 20X objectives. In addition, each cell line was genotyped by CytoSNP-850K microarrays to ascertain any additional CNVs that may have been present in the study participant or arisen in culture. Copy number variation was called using PennCNV[97] and any CNVs smaller than 200 Kb were disregarded. For the remaining CNVs, gene content was identified via the UCSC Genome Browser (*genome.ucsc.edu*) and possible pathogenesis was assessed using the OMIM and ClinVar databases. Four variants of unknown significance (VUS) were identified in 3q29Del NPC lines along with the seven deletions at the 3q29 locus on hg38 coordinates (Supp. Table S1.4). See Supp. Information Section A: Extended Methods for further details.

Differentiation of iPSCs to NPCs

Characterized iPSC cultures demonstrating no obvious spontaneous differentiation were differentiated to NPCs using STEMdiff SMADi Neural Induction Medium (StemCell Technologies). Briefly, iPSC cultures were dissociated with Accutase (StemCell Technologies, Vancouver). $2x10_6$ cells were plated in Induction Medium + 10uM Y-27632 (Reprocell, Beltsville, MD) in each well of a Matrigel-coated 6-well plate and re-fed daily (without Y-27632). Cells were passaged with Accutase after 7, 14, and 21 days and re-plated at a density of $1.4x10_6$ cells/well each time. At the 3rd plating (day 21), cells were transitioned to Neural Progenitor Medium (StemCell Technologies) and designated NPC passage 0. Morphology was monitored; rosette-like structures typically formed by NPC passage 1. Between NPC p0-p2 cells were plated on Matrigel-coated Nunc chamber slides for characterization by immunofluorescence as described above with antibodies to SOX2 (Invitrogen), PAX6 (Biolegend), and Nestin (Abcam) (Supp. Fig. S10). Slides were imaged on a Keyence BZ-X810 microscope under a 20X objective and quantified in ImageJ (example images in Fig. 3 were acquired on an Olympus FV1000 confocal microscope). Cells were harvested for RNA isolation at similar morphological stages.

RNA-sequencing, alignment and quantification of reads

NPC RNA was isolated from passage 1-3 NPC cultures with the RNeasy Plus Kit (Qiagen). RNA was processed according to standard protocols. After library generation, 50M paired-end 150bp read sequencing was performed on an Illumina platform. Sequences were quality-checked using FastQC for completeness, depth

and read quality and aligned to the hg38 reference genome using STAR aligner[98]. Gene quantification was conducted by using HTSeq-count [99].We used two analysis tools (DESeq2 [100] and edgeR [101]) using the negative binomial model to identify DEGs between case and control NPCs. Only the consensus DEGs that were concurrently identified by DESeq2 and edgeR at nominal significance level ($P < 0.05$) were carried into downstream analysis. We determined statistically significant differences in gene expression at the nominal level of significance ($P < 0.05$). We also evaluated the effect of Benjamini-Hochberg correction on raw p-values to account for multiple hypothesis testing. All steps of the differential expression analysis were conducted in R. Heatmaps in Fig. 3 were generated with the R package *pheatmap* [102]. Differential gene expression of 3q29 interval genes was also quantified by real-time quantitative reverse transcription PCR (qRT-PCR; TaqMan assays listed in Supp. Table S1.2) in a separate set of NPC cultures from the same cohort of cell lines (Supp. Fig. S11).

Age estimation of control and 3q29Del NPCs

To benchmark the approximate developmental age of the NPCs generated in this study, we compared the average gene expression profiles of our control and 3q29Del NPCs to publicly available RNA-Seq data obtained from the Allen BrainSpan Atlas [44]. Using Spearman's rank correlation analysis, we identified the developmental timepoints showing the strongest correlation with the average gene expression values observed in our NPCs for all detected genes (Fig. 4F). A similar approach to characterizing maturity in iPSC-derived NPCs was taken by previous studies[70], [103].

Comparison of empirically-derived DEGs and network-derived predictions

Protein-coding NPC DEGs (gene-type annotated by GENCODE v19) were tested for enrichment of genes identified in i) the broad 3q29 network; ii) top-neighbor-based 3q29 subnetwork, and iii) prioritized drivers using hypergeometric tests. The *GeneOverlap* package[94] in R was used for this analysis. The background list for the hypergeometric test was determined as the total number of unique genes detected to have at least one read in one NPC line. Hypergeometric $P < 0.05$ was used as the significance cut-off for defining the enrichment-based validation of WGCNA-derived predictions in experimental data. The list of NPC DEGs, including the subsets intersecting WGCNA-derived predictions are provided in Supp. Table S3.5. Full list of gene symbols

corresponding to the three levels of network interconnectedness are provided in Supp. Tables S2.2, and S3.1, 3, and 5. A detailed account of the methodology used to identify the levels of network interconnectedness summarized above is provided in Supp. Information Section A: Extended Methods.

Acknowledgments

This study was supported by NIH R01MH110701 and Emory University School of Medicine. This study was also supported in part by the Emory Integrated Genomics Core (EIGC), Emory Integrated Computational Core (EICC), Emory Integrated Cellular Imaging (ICI), and Emory Stem Cell Core (ESCC), which are subsidized by the Emory University School of Medicine.

The GTEx data (release version 6) used for the analyses described in this manuscript were downloaded from the GTEx portal on 01/08/2019 (http://www.gtexportal.org/home/datasets/, file name: “GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_rpkm.gct.gz”); dbGaP accession number: phs000424.v6.p1. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

The BrainSpan data (Developmental Transcriptome Dataset) used to construct the test network were downloaded from the Allen Brain Atlas portal on 01/12/2019 (https://www.brainspan.org/static/download/, file name: “RNA-Seq Gencode v10 summarized to genes”); dbGaP accession number: phs000755.v2.p1.

Competing Interests Statement

The authors declare no competing interests.

References

1. Levinson DF, Duan J, Oh S, Wang K, Sanders AR, Shi J *et al.* Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. *Am J Psychiatry* 2011; **168**(3)**:** 302-316.

2. Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS *et al.* Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat Genet* 2017; **49**(1)**:** 27-35.

3. Mulle JG, Dodd AF, McGrath JA, Wolyniec PS, Mitchell AA, Shetty AC *et al.* Microdeletions of 3q29 confer high risk for schizophrenia. *Am J Hum Genet* 2010; **87**(2)**:** 229-236.

4. Szatkiewicz JP, O'Dushlaine C, Chen G, Chambert K, Moran JL, Neale BM *et al.* Copy number variation in schizophrenia in Sweden. *Mol Psychiatry* 2014; **19**(7)**:** 762-773.

5. Mulle JG. The 3q29 deletion confers >40-fold increase in risk for schizophrenia. *Mol Psychiatry* 2015; **20**(9)**:** 1028-1029.

6. Pollak RM, Murphy MM, Epstein MP, Zwick ME, Klaiman C, Saulnier CA *et al.* Neuropsychiatric phenotypes and a distinct constellation of ASD features in 3q29 deletion syndrome: results from the 3q29 registry. *Mol Autism* 2019; **10:** 30.

7. Glassford MR, Rosenfeld JA, Freedman AA, Zwick ME, Mulle JG, Unique Rare Chromosome Disorder Support G. Novel features of 3q29 deletion syndrome: Results from the 3q29 registry. *Am J Med Genet A* 2016; **170A**(4)**:** 999-1006.

8. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* 2015; **87**(6)**:** 1215-1233.

9. Singh MD, Jensen M, Lasser M, Huber E, Yusuff T, Pizzo L *et al.* NCBP2 modulates neurodevelopmental defects of the 3q29 deletion in Drosophila and Xenopus laevis models. *PLoS Genet* 2020; **16**(2)**:** e1008590.

10. Carroll LS, Williams HJ, Walters J, Kirov G, O'Donovan MC, Owen MJ. Mutation screening of the 3q29 microdeletion syndrome candidate genes DLG1 and PAK2 in schizophrenia. *Am J Med Genet B Neuropsychiatr Genet* 2011; **156B**(7)**:** 844-849.

11. Gardoni F, Mauceri D, Fiorentini C, Bellone C, Missale C, Cattabeni F *et al.* CaMKII-dependent phosphorylation regulates SAP97/NR2A interaction. *J Biol Chem* 2003; **278**(45)**:** 44745-44752.

12. Leonard AS, Davare MA, Horne MC, Garner CC, Hell JW. SAP97 is associated with the alpha-amino-3-hydroxy-5-methylisoxazole-4-propionic acid receptor GluR1 subunit. *J Biol Chem* 1998; **273**(31)**:** 19518-19524.

13. Lin EI, Jeyifous O, Green WN. CASK regulates SAP97 conformation and its interactions with AMPA and NMDA receptors. *J Neurosci* 2013; **33**(29)**:** 12067-12076.

14. Zhou W, Zhang L, Guoxiang X, Mojsilovic-Petrovic J, Takamaya K, Sattler R *et al.* GluR1 controls dendrite growth through its binding partner, SAP97. *J Neurosci* 2008; **28**(41)**:** 10220-10233.

15. Moghaddam B, Javitt D. From revolution to evolution: the glutamate hypothesis of schizophrenia and its implication for treatment. *Neuropsychopharmacology* 2012; **37**(1)**:** 4-15.

16. Rutkowski TP, Purcell RH, Pollak RM, Grewenow SM, Gafford GM, Malone T *et al.* Behavioral changes and growth deficits in a CRISPR engineered mouse model of the schizophrenia-associated 3q29 deletion. *Mol Psychiatry* 2019.

17. Bokoch GM. Biology of the p21-activated kinases. *Annu Rev Biochem* 2003; **72:** 743-781.

18. Wang Y, Zeng C, Li J, Zhou Z, Ju X, Xia S *et al.* PAK2 Haploinsufficiency Results in Synaptic Cytoskeleton Impairment and Autism-Related Behavior. *Cell Rep* 2018; **24**(8)**:** 2029-2041.

19. Allen KM, Gleeson JG, Bagrodia S, Partington MW, MacMillan JC, Cerione RA *et al.* PAK3 mutation in nonsyndromic X-linked mental retardation. *Nat Genet* 1998; **20**(1)**:** 25-30.

20. Tarpey P, Parnau J, Blow M, Woffendin H, Bignell G, Cox C *et al.* Mutations in the DLG3 gene cause nonsyndromic X-linked mental retardation. *Am J Hum Genet* 2004; **75**(2)**:** 318-324.

21. Grice SJ, Liu JL, Webber C. Synergistic interactions between Drosophila orthologues of genes spanned by de novo human CNVs support multiple-hit models of autism. *PLoS Genet* 2015; **11**(3)**:** e1004998.

22. Pabis M, Neufeld N, Shav-Tal Y, Neugebauer KM. Binding properties and dynamic localization of an alternative isoform of the cap-binding complex subunit CBP20. *Nucleus* 2010; **1**(5)**:** 412-421.

23. Rutkowski TP, Purcell RH, Pollak RM, Grewenow SM, Gafford GM, Malone T *et al.* Behavioral changes and growth deficits in a CRISPR engineered mouse model of the schizophrenia-associated 3q29 deletion. *Molecular Psychiatry* 2019.

24. Barbaric I, Miller G, Dear TN. Appearances can be deceiving: phenotypes of knockout mice. *Brief Funct Genomic Proteomic* 2007; **6**(2)**:** 91-103.

25. Johnson EC, Border R, Melroy-Greif WE, de Leeuw CA, Ehringer MA, Keller MC. No Evidence That Schizophrenia Candidate Genes Are More Associated With Schizophrenia Than Noncandidate Genes. *Biol Psychiatry* 2017; **82**(10)**:** 702-708.

26. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. Protein function in the post-genomic era. *Nature* 2000; **405**(6788)**:** 823-826.

27. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005; **4:** Article17.

28. Ma'ayan A, Blitzer RD, Iyengar R. Toward predictive models of mammalian cells. *Annu Rev Biophys Biomol Struct* 2005; **34:** 319-349.

29. Qiu Y, Arbogast T, Lorenzo SM, Li H, Tang SC, Richardson E *et al.* Oligogenic Effects of 16p11.2 Copy-Number Variation on Craniofacial Development. *Cell Rep* 2019; **28**(13)**:** 3320-3328 e3324.

30. Wolfe CJ, Kohane IS, Butte AJ. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics* 2005; **6:** 227.

31. Horan K, Jang C, Bailey-Serres J, Mittler R, Shelton C, Harper JF *et al.* Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiol* 2008; **147**(1)**:** 41-57.

32. Ahn AC, Tewari M, Poon CS, Phillips RS. The limits of reductionism in medicine: could systems biology offer an alternative? *PLoS Med* 2006; **3**(6)**:** e208.

33. Oliver S. Guilt-by-association goes global. *Nature* 2000; **403**(6770)**:** 601-603.

34. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. *Genome Res* 2004; **14**(6)**:** 1085-1094.

35. Moreau Y, Tranchevent LC. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet* 2012; **13**(8)**:** 523-536.

36. Oti M, Brunner HG. The modular nature of genetic diseases. *Clin Genet* 2007; **71**(1)**:** 1-11.

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

38. Radulescu E, Jaffe AE, Straub RE, Chen Q, Shin JH, Hyde TM *et al.* Identification and prioritization of gene sets associated with schizophrenia risk by co-expression network analysis in human brain. *Mol Psychiatry* 2020; **25**(4)**:** 791-804.

39. Etemadikhah M, Niazi A, Wetterberg L, Feuk L. Transcriptome analysis of fibroblasts from schizophrenia patients reveals differential expression of schizophrenia-related genes. *Sci Rep* 2020; **10**(1)**:** 630.

40. Gordon A, Forsingdal A, Klewe IV, Nielsen J, Didriksen M, Werge T *et al.* Transcriptomic networks implicate neuronal energetic abnormalities in three mouse models harboring autism and schizophrenia-associated mutations. *Mol Psychiatry* 2019.

41. Chen C, Cheng L, Grennan K, Pibiri F, Zhang C, Badner JA *et al.* Two gene co-expression modules differentiate psychotics and controls. *Mol Psychiatry* 2013; **18**(12)**:** 1308-1314.

42. Oron O, Getselter D, Shohat S, Reuveni E, Lukic I, Shifman S *et al.* Gene network analysis reveals a role for striatal glutamatergic receptors in dysregulated risk-assessment behavior of autism mouse models. *Transl Psychiatry* 2019; **9**(1)**:** 257.

43. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015; **348**(6235)**:** 648-660.

44. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M *et al.* Spatio-temporal transcriptome of the human brain. *Nature* 2011; **478**(7370)**:** 483-489.

45. Mokhtari R, Lachman HM. The Major Histocompatibility Complex (MHC) in Schizophrenia: A Review. *J Clin Cell Immunol* 2016; **7**(6).

46. Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* 2016; **530**(7589)**:** 177-183.

47. Warnatsch A, Bergann T, Kruger E. Oxidation matters: the ubiquitin proteasome system connects innate immune mechanisms with MHC class I antigen presentation. *Mol Immunol* 2013; **55**(2)**:** 106-109.

48. Varelas X, Miller BW, Sopko R, Song S, Gregorieff A, Fellouse FA *et al.* The Hippo pathway regulates Wnt/beta-catenin signaling. *Dev Cell* 2010; **18**(4)**:** 579-591.

49. Cooper TF, Morby AP, Gunn A, Schneider D. Effect of random and hub gene disruptions on environmental and mutational robustness in Escherichia coli. *BMC Genomics* 2006; **7:** 237.

50. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* 2008; **322**(5898)**:** 104-110.

51. Almaas E. Biological impacts and context of network theory. *J Exp Biol* 2007; **210**(Pt 9)**:** 1548-1558.

52. Sanati N, Iancu OD, Wu G, Jacobs JE, McWeeney SK. Network-Based Predictors of Progression in Head and Neck Squamous Cell Carcinoma. *Front Genet* 2018; **9:** 183.

53. Laine VN, Verhagen I, Mateman AC, Pijl A, Williams TD, Gienapp P *et al.* Exploration of tissue-specific gene expression patterns underlying timing of breeding in contrasting temperature environments in a song bird. *BMC Genomics* 2019; **20**(1)**:** 693.

54. Li W, Wang L, Wu Y, Yuan Z, Zhou J. Weighted gene coexpression network analysis to identify key modules and hub genes associated with atrial fibrillation. *Int J Mol Med* 2020; **45**(2)**:** 401-416.

55. Alexandru G, Graumann J, Smith GT, Kolawa NJ, Fang R, Deshaies RJ. UBXD7 binds multiple ubiquitin ligases and implicates p97 in HIF1alpha turnover. *Cell* 2008; **134**(5)**:** 804-816.

56. Tao S, Liu P, Luo G, Rojo de la Vega M, Chen H, Wu T *et al.* p97 Negatively Regulates NRF2 by Extracting Ubiquitylated NRF2 from the KEAP1-CUL3 E3 Complex. *Mol Cell Biol* 2017; **37**(8).

57. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A *et al.* Proteomics. Tissue-based map of the human proteome. *Science* 2015; **347**(6220)**:** 1260419.

58. Chan SY, Loscalzo J. The emerging paradigm of network medicine in the study of human disease. *Circ Res* 2012; **111**(3)**:** 359-374.

59. Deciphering Developmental Disorders S. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* 2017; **542**(7642)**:** 433-438.

60. Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* 2016; **19**(11)**:** 1442-1453.

61. Li J, Cai T, Jiang Y, Chen H, He X, Chen C *et al.* Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Mol Psychiatry* 2016; **21**(2)**:** 290-297.

62. Meng Q, Wang K, Brunetti T, Xia Y, Jiao C, Dai R *et al.* The DGCR5 long noncoding RNA may regulate expression of several schizophrenia-related genes. *Sci Transl Med* 2018; **10**(472).

63. Psych EC, Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ *et al.* The PsychENCODE project. *Nat Neurosci* 2015; **18**(12)**:** 1707-1712.

64. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014; **511**(7510)**:** 421-427.

65. Chang D, Nalls MA, Hallgrimsdottir IB, Hunkapiller J, van der Brug M, Cai F *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat Genet* 2017; **49**(10)**:** 1511-1516.

66. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. *Nat Genet* 2019; **51**(3)**:** 414-430.

67. Momozawa Y, Dmitrieva J, Theatre E, Deffontaine V, Rahmouni S, Charloteaux B *et al.* IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat Commun* 2018; **9**(1)**:** 2427.

68. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN *et al.* Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum Mol Genet* 2018; **27**(20)**:** 3641-3649.

69. Guan J, Cai JJ, Ji G, Sham PC. Commonality in dysregulated expression of gene sets in cortical brains of individuals with autism, schizophrenia, and bipolar disorder. *Transl Psychiatry* 2019; **9**(1)**:** 152.

70. Brennand K, Savas JN, Kim Y, Tran N, Simone A, Hashimoto-Torii K *et al.* Phenotypic differences in hiPSC NPCs derived from patients with schizophrenia. *Mol Psychiatry* 2015; **20**(3)**:** 361-368.

71. Selemon LD. Regionally diverse cortical pathology in schizophrenia: clues to the etiology of the disease. *Schizophr Bull* 2001; **27**(3)**:** 349-377.

72. Ala U, Piro RM, Grassi E, Damasco C, Silengo L, Oti M *et al.* Prediction of human disease genes by human-mouse conserved coexpression analysis. *PLoS Comput Biol* 2008; **4**(3)**:** e1000043.

73. Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. *Nat Genet* 2004; **36**(10)**:** 1090-1098.

74. Torkamani A, Dean B, Schork NJ, Thomas EA. Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia. *Genome Res* 2010; **20**(4)**:** 403-412.

75. van Dam S, Cordeiro R, Craig T, van Dam J, Wood SH, de Magalhaes JP. GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases. *BMC Genomics* 2012; **13:** 535.

76. McCarroll SA, Murphy CT, Zou S, Pletcher SD, Chin CS, Jan YN *et al.* Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat Genet* 2004; **36**(2)**:** 197-204.

77. Farrell MS, Werge T, Sklar P, Owen MJ, Ophoff RA, O'Donovan MC *et al.* Evaluating historical candidate genes for schizophrenia. *Mol Psychiatry* 2015; **20**(5)**:** 555-562.

78. Willatt L, Cox J, Barber J, Cabanas ED, Collins A, Donnai D *et al.* 3q29 microdeletion syndrome: clinical and molecular characterization of a new syndrome. *Am J Hum Genet* 2005; **77**(1)**:** 154-160.

79. Wachi S, Yoneda K, Wu R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* 2005; **21**(23)**:** 4205-4208.

80. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. *Bioinformatics* 2006; **22**(18)**:** 2291-2297.

81. Xu J, Li Y. Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics* 2006; **22**(22)**:** 2800-2805.

82. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016; **536**(7616)**:** 285-291.

83. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 2012; **485**(7397)**:** 246-250.

84. Oh E, Mark KG, Mocciaro A, Watson ER, Prabu JR, Cha DD *et al.* Gene expression and cell identity controlled by anaphase-promoting complex. *Nature* 2020; **579**(7797)**:** 136-140.

85. Griesi-Oliveira K, Acab A, Gupta AR, Sunaga DY, Chailangkarn T, Nicol X *et al.* Modeling non-syndromic autism and the impact of TRPC6 disruption in human neurons. *Mol Psychiatry* 2015; **20**(11)**:** 1350-1365.

86. van Dam S, Vosa U, van der Graaf A, Franke L, de Magalhaes JP. Gene co-expression analysis for functional classification and gene-disease predictions. *Brief Bioinform* 2018; **19**(4)**:** 575-592.

87. Langfelder P, Horvath S. Fast R Functions for Robust Correlations and Hierarchical Clustering. *J Stat Softw* 2012; **46**(11).

88. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004; **5**(2)**:** 101-113.

89. Fu J, Keurentjes JJ, Bouwmeester H, America T, Verstappen FW, Ward JL *et al.* System-wide molecular evidence for phenotypic buffering in Arabidopsis. *Nat Genet* 2009; **41**(2)**:** 166-167.

90. Ouma WZ, Pogacar K, Grotewold E. Topological and statistical analyses of gene regulatory networks reveal unifying yet quantitatively different emergent properties. *PLoS Comput Biol* 2018; **14**(4)**:** e1006098.

91. Lachowiec J, Queitsch C, Kliebenstein DJ. Molecular mechanisms governing differential robustness of development and environmental responses in plants. *Ann Bot* 2016; **117**(5)**:** 795-809.

92. Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics* 2007; **8:** 22.

93. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL. Hierarchical organization of modularity in metabolic networks. *Science* 2002; **297**(5586)**:** 1551-1555.

94. Shen L. GeneOverlap: Test and visualize gene overlaps. R package version 1.20.0 edn2019.

95. Murphy MM, Lindsey Burrell T, Cubells JF, Espana RA, Gambello MJ, Goines KCB *et al.* Study protocol for The Emory 3q29 Project: evaluation of neurodevelopmental, psychiatric, and medical symptoms in 3q29 deletion syndrome. *BMC Psychiatry* 2018; **18**(1)**:** 183.

96. Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* 2009; **460**(7256)**:** 753-757.

97. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 2007; **17**(11)**:** 1665-1674.

98. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; **29**(1)**:** 15-21.

99. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015; **31**(2)**:** 166-169.

100. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014; **15**(12)**:** 550.

101. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; **26**(1)**:** 139-140.

102. Kolde R. pheatmap: Pretty Heatmaps. 1.0.12 edn2019.

103. Mehta SR, Tom CM, Wang Y, Bresee C, Rushton D, Mathkar PP *et al.* Human Huntington's Disease iPSC-Derived Cortical Neurons Display Altered Transcriptomics, Morphology, and Maturation. *Cell Rep* 2018; **25**(4)**:** 1081-1096 e1086.

Figure captions

Figure 1. Unbiased weighted gene co-expression network analysis (WGCNA) reveals convergent and distributed effects of 3q29Del across the adult human cortical transcriptome. (**A**) Data analysis workflow underlying WGCNA-derived predictions for functional interrogation of the 3q29Del interval. Reference and test datasets from the GTEx and BrainSpan Projects were used for network construction and establishment of module preservation, respectively. Network schematics represent the three levels of network interconnectedness identified in this study to conjoin the 3q29 interval with the global protein-coding genome of the adult human prefrontal cortex (PFC). BA9: Brodmann area 9. (**B**) Hierarchical clustering of module eigengenes summarizing the modules identified via cluster analysis. 21 protein coding genes located in the 3q29 interval were found to organize into seven modules (black boxes). (**C**) Eigengene-based connectivity heatmap of 3q29 genes. Connectivity strength (kME) was defined by the Pearson's correlation between genes and module eigengenes (ME). *UBXN7* was identified as a hub-gene (red box; kME > 0.8, P = 8.33E-30). (**D**) Heatmap representing the strength of Pearson's correlation (r) between MEs. The seven modules harboring 3q29 interval genes further clustered into three meta-modules, corresponding to the blue squares along the diagonal. (**E**) Top 10 biological pathways (Reactome database) enriched in modules harboring 3q29 interval genes (adjusted $P < 0.05$; capped at $-\log_{10}$ adjusted $P = 10$). (**F**) Predicted model of 3q29Del biology, working through convergent and distributed mechanisms of action in the adult human PFC. Nodes represent the 21 protein-coding 3q29 interval genes, including one hub gene. Colors and dashed lines reflect module and meta-module assignments, respectively.

Figure 2. 3q29 interval genes form transcriptomic subnetworks enriched for known schizophrenia, autism and intellectual / developmental disability-risk genes. Adjusted p-values from hypergeometric tests identifying the significance of the overlap between intra-modular top neighbors of individual 3q29 genes and known risk genes for autism spectrum disorders (ASD), intellectual / developmental disability (IDD) and schizophrenia (SZ) across the allelic frequency spectrum. Risk gene-sets for three traits with no known association to the 3q29Del syndrome were also tested for over-representation as negative controls. Common variants associated with height were included as another negative control to rule out a potential bias introduced by gene-set size. Nine protein coding genes from the 3q29 interval formed transcriptomic subnetworks that are

significantly enriched for known SZ, ASD and/or IDD risk genes (orange highlight, adjusted $P < 0.05$). The proportion of hypergeometric tests with significant overrepresentation of SZ, ASD, and IDD gene-sets (19/96) was found to be an order of magnitude larger than that of the negative control tests (2/64), demonstrating the high specificity of the identified disease enrichment patterns.

Figure 3. Network of prioritized drivers predicted to contribute to the neurodevelopmental and psychiatric sequelae of 3q29Del. (**A**) Nine protein coding genes from the 3q29 interval formed transcriptomic subnetworks that were significantly enriched for known schizophrenia (SZ), autism spectrum disorder (ASD) and intellectual / developmental disability (IDD)-risk genes (adjusted $P < 0.05$). Black and red nodes illustrated in this network diagram represent these 9 3q29 interval genes and their 284 disease-associated top neighbors, respectively. The union of these prioritized genes constitute 293 genetic drivers predicted to contribute to the neurodevelopmental and psychiatric phenotypes of 3q29Del. The color of network edges that connect node-pairs represents module assignment. (**B**) Top 20 biological processes and pathways with significant enrichment among prioritized drivers (adjusted $P < 0.05$). The identified Gene Ontology biological processes (GO: BP) and Reactome and KEGG biological pathways point to key mechanisms through which 3q29Del is predicted to influence susceptibility to SZ, ASD, and IDD. (**C**) Organization of all statistically significant biological processes enriched in candidate drivers into a network of related functional annotation categories. GO:BP terms are connected if they have a high overlap (share many genes); edge width represents magnitude of the overlap.

Figure 4. WGCNA network predicts differentially expressed genes in 3q29Del neural progenitors. (**A**) Induced pluripotent stem cell (iPSC) lines were generated from 3q29Del carriers (N=7) and matched controls (N=6). (**B**) iPSC lines were differentiated to forebrain neural progenitor cell (NPC) cultures by dual-SMAD pathway inhibition. Scale bar = 50µm. RNA-seq was performed on NPC lines (**C-F**). (**C**) 253 protein-coding genes were found to be differentially expressed (black label) ($P < 0.05$) and all 10 protein-coding genes that survived the FDR adjusted $P < 0.05$ threshold (blue label) reside in the 3q29Del locus. (**D**) Gene-scaled expression of all detected protein-coding genes within the 3q29 interval shows a consistent reduction in 3q29Del NPCs proportional to gene copy number. Genes in gray were not detected by RNA-Seq. Asterisks indicate level of significance (*$P < 0.05$, **FDR adjusted $P < 0.05$). (**E**) For empirical validation, all 253 protein-coding DEGs were

tested for enrichment of genes found in the broad 3q29 network (all genes in 3q29 modules), top-neighbor based 3q29 subnetwork, and disease-associated prioritized drivers. A significant enrichment was found at each level of network interconnectedness ($P < 0.05$). (**F**) Spearman rank correlations (ρ) of average control and 3q29Del NPC gene expression profiles compared to Brainspan data spanning 31 developmental timepoints of the human brain transcriptome. For both groups, 8_{th} ($\rho = 0.80$, $P < 0.05$) and 9_{th} ($\rho = 0.80$, $P < 0.05$) weeks post conception (pcw) were the most similar time points. m: months, y: years of age.

Figure 1

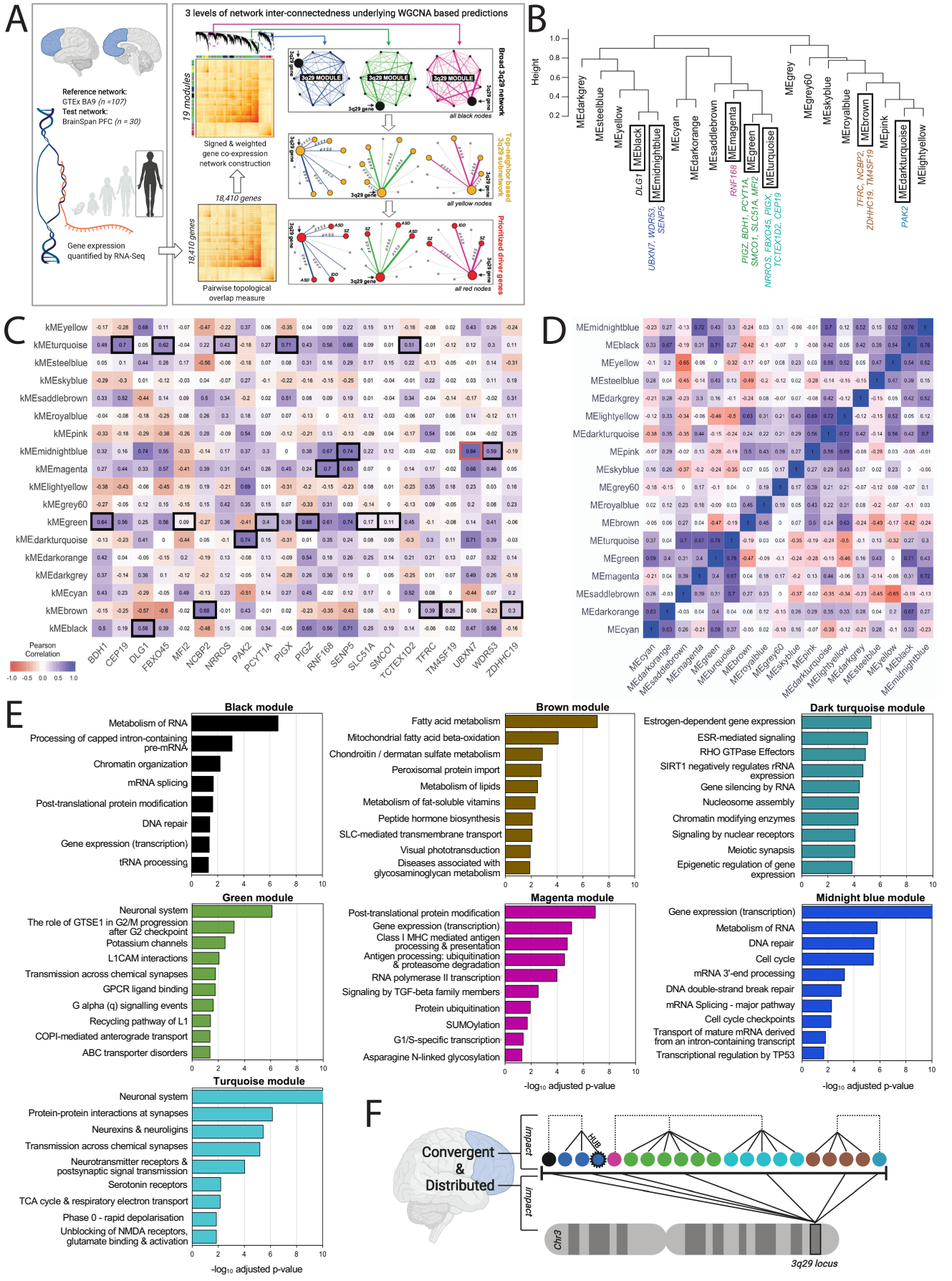

A
Reference network:
GTEx BA9 (n =107)
Test network:
BrainSpan PFC (n = 30)
Gene expression quantified by RNA-Seq
3 levels of network inter-connectedness underlying WGCNA based predictions
19 modules
Signed & weighted gene co-expression network construction
18,410 genes
18,410 genes
Pairwise topological overlap measure
3q29 MODULE
3q29 MODULE
3q29 MODULE
3q29 gene
all black nodes
Broad 3q29 network
all yellow nodes
Top-neighbor based 3q29 subnetwork
all red nodes
Prioritized driver genes
ASD
SZ
IDD
B
Height
MEdarkgrey
MEsteelblue
MEyellow
MEblack
MEmidnightblue
MEcyan
MEdarkorange
MEsaddlebrown
MEmagenta
MEgreen
MEturquoise
MEgrey
MEgrey60
MEskyblue
MEroyalblue
MEbrown
MEpink
MEdarkturquoise
MElightyellow
DLG1
UBXN7, WDR53, SENP5
RNF168
PIGZ, BDH1, PCYT1A, SMCO1, SLC51A, MF2, NRROS, FBXO45, PIGX, TCTEX1D2, CEP19
TFRC, NCBP2, ZDHHC19, TM4SF19
PAK2
C
kMEyellow
kMEturquoise
kMEsteelblue
kMEskyblue
kMEsaddlebrown
kMEroyalblue
kMEpink
kMEmidnightblue
kMEmagenta
kMElightyellow
kMEgrey60
kMEgreen
kMEdarkturquoise
kMEdarkorange
kMEdarkgrey
kMEcyan
kMEbrown
kMEblack
BDH1
CEP19
DLG1
FBXO45
MF2
NCBP2
NRROS
PAK2
PCYT1A
PIGX
PIGZ
RNF168
SENP5
SLC51A
SMCO1
TCTEX1D2
TFRC
TM4SF19
UBXN7
WDR53
ZDHHC19
Pearson Correlation
-1.0 -0.5 0.0 0.5 1.0
D
MEmidnightblue
MEblack
MEyellow
MEsteelblue
MEdarkgrey
MElightyellow
MEdarkturquoise
MEpink
MEskyblue
MEgrey60
MEroyalblue
MEbrown
MEturquoise
MEgreen
MEmagenta
MEsaddlebrown
MEdarkorange
MEcyan
E
Black module
Metabolism of RNA
Processing of capped intron-containing pre-mRNA
Chromatin organization
mRNA splicing
Post-translational protein modification
DNA repair
Gene expression (transcription)
tRNA processing
Brown module
Fatty acid metabolism
Mitochondrial fatty acid beta-oxidation
Chondroitin / dermatan sulfate metabolism
Peroxisomal protein import
Metabolism of lipids
Metabolism of fat-soluble vitamins
Peptide hormone biosynthesis
SLC-mediated transmembrane transport
Visual phototransduction
Diseases associated with glycosaminoglycan metabolism
Dark turquoise module
Estrogen-dependent gene expression
ESR-mediated signaling
RHO GTPase Effectors
SIRT1 negatively regulates rRNA expression
Gene silencing by RNA
Nucleosome assembly
Chromatin modifying enzymes
Signaling by nuclear receptors
Meiotic synapsis
Epigenetic regulation of gene expression
Green module
Neuronal system
The role of GTSE1 in G2/M progression after G2 checkpoint
Potassium channels
L1CAM interactions
Transmission across chemical synapses
GPCR ligand binding
G alpha (q) signalling events
Recycling pathway of L1
COPI-mediated anterograde transport
ABC transporter disorders
Magenta module
Post-translational protein modification
Gene expression (transcription)
Class I MHC mediated antigen processing & presentation
Antigen processing: ubiquitination & proteasome degradation
RNA polymerase II transcription
Signaling by TGF-beta family members
Protein ubiquitination
SUMOylation
G1/S-specific transcription
Asparagine N-linked glycosylation
Midnight blue module
Gene expression (transcription)
Metabolism of RNA
DNA repair
Cell cycle
mRNA 3'-end processing
DNA double-strand break repair
mRNA Splicing - major pathway
Cell cycle checkpoints
Transport of mature mRNA derived from an intron-containing transcript
Transcriptional regulation by TP53
Turquoise module
Neuronal system
Protein-protein interactions at synapses
Neurexins & neuroligins
Transmission across chemical synapses
Neurotransmitter receptors & postsynaptic signal transmission
Serotonin receptors
TCA cycle & respiratory electron transport
Phase 0 - rapid depolarisation
Unblocking of NMDA receptors, glutamate binding & activation
-log10 adjusted p-value
F
Convergent & Distributed
impact
impact
HUB
Chr3
3q29 locus

Figure 2

	BDH1	CEP19	DLG1	FBXO45	NCBP2	PAK2	PCYT1A	PIGX	RNF168	SENP5	TCTEX1D2	TFRC	UBXN7	WDR53	PIGZ	NRROS
IDD risk *(McRae et al., 2017)* gene-set size = 93	0.568	0.002*	0.085	0.014*	1.000	0.673	0.699	0.080	0.011*	0.005*	0.677	1.000	< 0.000*	0.015*	0.150	1.000
ASD risk *(SFARI categories 1 & 2)* gene set size = 86	0.568	0.002*	0.128	0.070	1.000	0.754	0.699	0.258	0.160	0.001*	0.991	1.000	< 0.000*	0.015*	0.849	1.000
Schizophrenia risk *(BrainGVEX case-control)* gene-set size = 651	0.769	0.665	1.000	0.940	1.000	0.673	1.000	0.635	1.000	1.000	0.420	1.000	0.807	0.875	0.849	1.000
Schizophrenia risk *(CommonMind case-control)* gene-set size = 636	< 0.000*	0.966	1.000	0.953	1.000	0.673	1.000	0.999	< 0.000*	0.060	0.991	1.000	0.004*	0.283	0.001*	1.000
Schizophrenia risk *(GWAS identified common risk variants)* gene-set size = 340	0.283	0.158	0.362	0.776	1.000	0.754	0.699	0.258	1.000	1.000	0.677	0.590	0.771	0.822	0.253	1.000
Schizophrenia risk *(Exonic de novo mutations)* gene-set size = 290	0.769	0.033*	< 0.000*	0.147	1.000	0.754	0.699	0.143	0.640	1.000	0.991	1.000	0.738	0.919	0.868	1.000
Alzheimer's disease risk *(GWAS identified common risk variants)* gene-set size = 25	0.769	0.665	1.000	0.940	1.000	0.673	1.000	0.999	1.000	1.000	1.000	0.590	0.807	1.000	0.849	1.000
Parkinson's disease risk *(GWAS identified common risk variants)* gene-set size = 67	0.568	0.514	0.493	0.278	1.000	0.754	0.699	0.143	0.640	1.000	1.000	1.000	0.807	1.000	0.849	0.020*
IBD risk *(GWAS identified common risk variants)* gene-set size = 98	0.769	0.665	1.000	0.912	1.000	1.000	0.699	0.679	1.000	1.000	0.991	1.000	0.738	1.000	0.849	1.000
Height *(GWAS identified common associated variants)* gene-set size = 479	0.175	0.771	0.128	0.940	1.000	0.702	0.699	0.999	0.640	0.024*	0.991	1.000	0.738	0.283	0.435	1.000

19 / 96 tests

2 / 64 tests

Figure 3

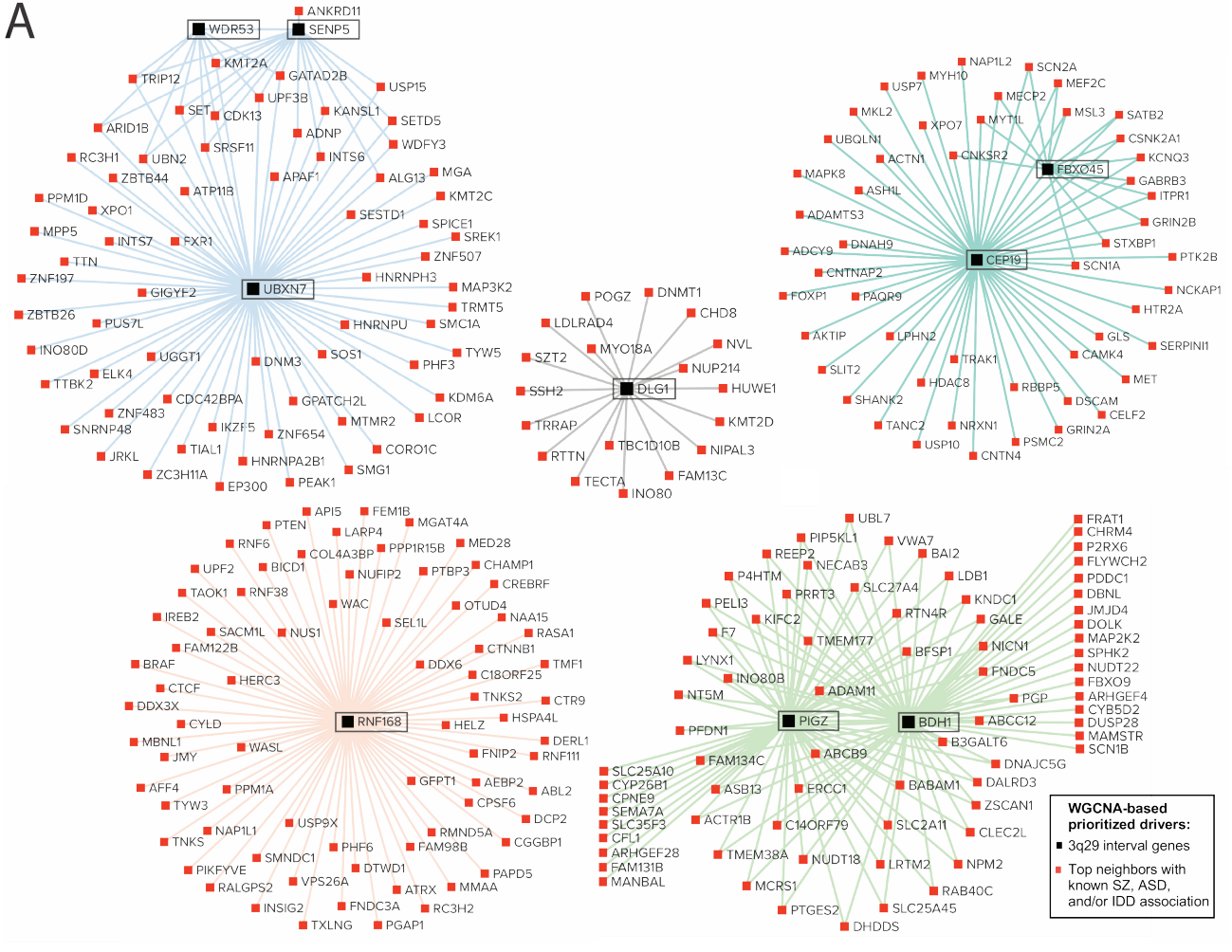
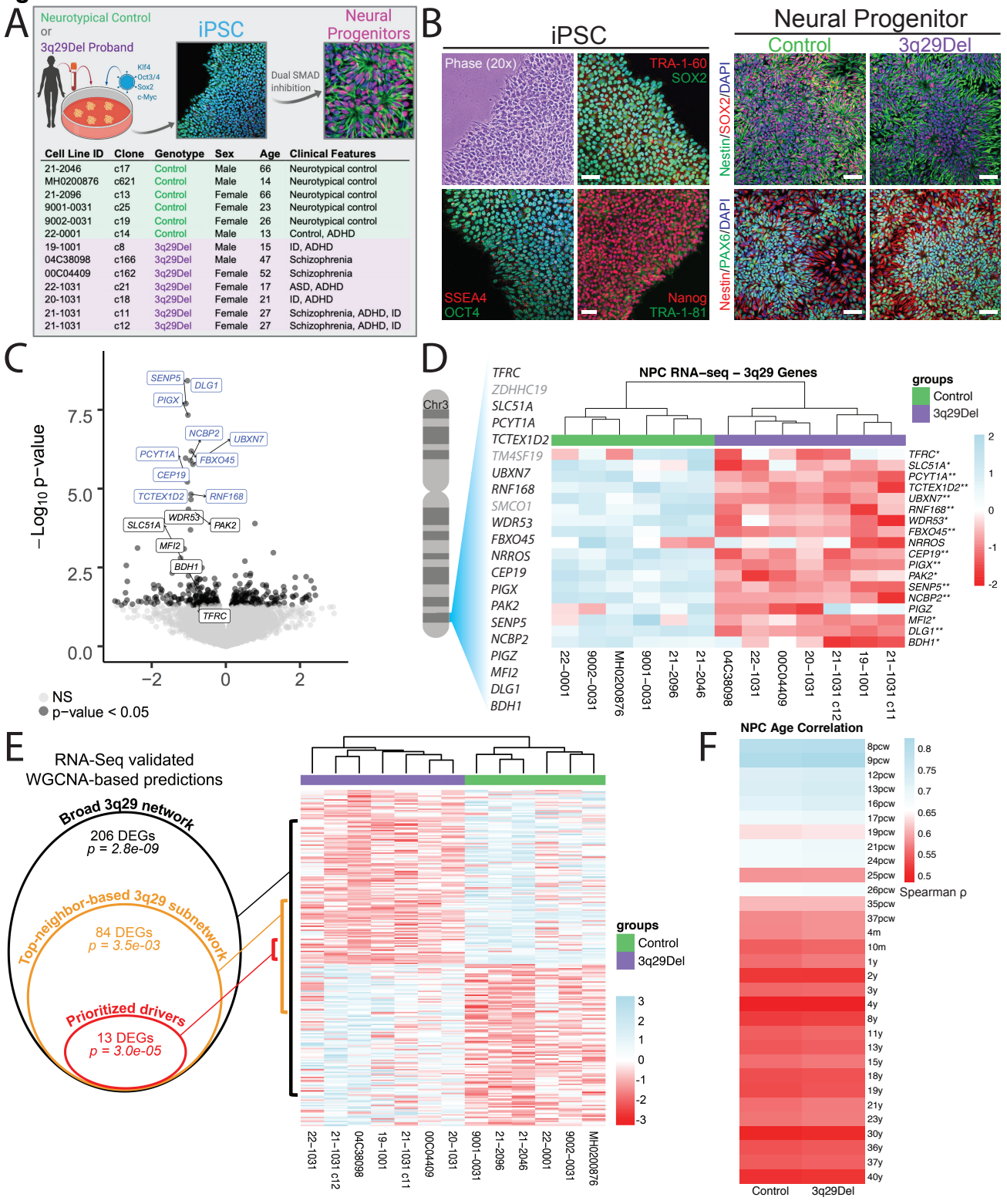

A
WDR53
SENP5
UBXN7
ANKRD11
KMT2A
GATAD2B
TRIP12
USP15
UPF3B
SET
CDK13
KANSL1
SETD5
ARID1B
ADNP
WDFY3
SRSF11
RC3H1
UBN2
INTS6
ZBTB44
APAF1
ALG13
MGA
ATP11B
PPM1D
KMT2C
XPO1
SESTD1
SPICE1
MPP5
INTS7
FXR1
SREK1
TTN
ZNF507
ZNF197
HNRNPH3
GIGYF2
MAP3K2
ZBTB26
TRMT5
PUS7L
HNRNPU
SMC1A
INO80D
UGGT1
DNM3
SOS1
TYW5
ELK4
PHF3
TTBK2
CDC42BPA
GPATCH2L
KDM6A
ZNF483
MTMR2
LCOR
SNRNP48
IKZF5
ZNF654
TIAL1
CORO1C
JRKL
HNRNPA2B1
SMG1
ZC3H11A
EP300
PEAK1
POGZ
DNMT1
CHD8
LDLRAD4
MYO18A
NVL
SZT2
NUP214
SSH2
DLG1
HUWE1
TRRAP
KMT2D
TBC1D10B
NIPAL3
RTTN
FAM13C
TECTA
INO80
NAP1L2
SCN2A
MYH10
MEF2C
USP7
MECP2
MKL2
MYT1L
MSL3
SATB2
XPO7
UBQLN1
CSNK2A1
ACTN1
CNKSR2
KCNQ3
FBXO45
MAPK8
GABRB3
ASH1L
ITPR1
ADAMTS3
GRIN2B
ADCY9
DNAH9
STXBP1
CEP19
PTK2B
CNTNAP2
SCN1A
NCKAP1
FOXP1
PAQR9
HTR2A
AKTIP
LPHN2
GLS
SERPINI1
SLIT2
TRAK1
CAMK4
HDAC8
RBBP5
MET
SHANK2
DSCAM
TANC2
NRXN1
CELF2
GRIN2A
USP10
PSMC2
CNTN4
API5
FEM1B
PTEN
MGAT4A
LARP4
RNF6
COL4A3BP
PPP1R15B
MED28
UPF2
BICD1
NUFIP2
PTBP3
CHAMP1
TAOK1
RNF38
CREBRF
WAC
OTUD4
IREB2
SEL1L
NAA15
SACM1L
NUS1
RASA1
FAM122B
CTNNB1
BRAF
DDX6
TMF1
HERC3
C18ORF25
CTCF
TNKS2
CTR9
DDX3X
CYLD
RNF168
HELZ
HSPA4L
MBNL1
DERL1
WASL
JMY
FNIP2
RNF111
AFF4
PPM1A
GFPT1
AEBP2
ABL2
TYW3
CPSF6
USP9X
DCP2
NAP1L1
RMND5A
TNKS
PHF6
FAM98B
CGGBP1
SMNDC1
PIKFYVE
DTWD1
PAPD5
RALGPS2
VPS26A
ATRX
MMAA
INSIG2
FNDC3A
RC3H2
TXLNG
PGAP1
UBL7
PIP5KL1
VWA7
REEP2
BAI2
NECAB3
P4HTM
LDB1
SLC27A4
PRRT3
KNDC1
PELI3
RTN4R
KIFC2
GALE
F7
TMEM177
BFSP1
NICN1
LYNX1
FNDC5
INO80B
ADAM11
NT5M
PGP
PIGZ
BDH1
ABCC12
PFDN1
B3GALT6
ABCB9
DNAJC5G
FAM134C
BABAM1
DALRD3
ASB13
ERCC1
ZSCAN1
ACTR1B
C14ORF79
SLC2A11
CLEC2L
TMEM38A
NUDT18
LRTM2
NPM2
MCRS1
RAB40C
PTGES2
SLC25A45
DHDDS
SLC25A10
CYP26B1
CPNE9
SEMA7A
SLC35F3
CFL1
ARHGEF28
FAM131B
MANBAL
FRAT1
CHRM4
P2RX6
FLYWCH2
PDDC1
DBNL
JMJD4
DOLK
MAP2K2
SPHK2
NUDT22
FBXO9
ARHGEF4
CYB5D2
DUSP28
MAMSTR
SCN1B
WGCNA-based prioritized drivers:
3q29 interval genes
Top neighbors with known SZ, ASD, and/or IDD association
B
Chromosome organization
Histone modification
Cellular component morphogenesis
Regulation of organelle organization
DNA metabolic process
Regulation of telomere maintenance
Neuron differentiation
Protein modification by small protein conjugation or removal
Neuron projection morphogenesis
Neurogenesis
Chemical synaptic transmission, postsynaptic
Post-embryonic development
Long-term potentiation
Regulation of actin cytoskeleton
Axon guidance
Post-translational protein modification
Diseases of signal transduction
Deubiquitination
Chromatin organization
Diseases associated with glycosylation precursor biosynthesis
0
2
4
6
8
10
-log10 adjusted p-value
C
regulation of nervous system development
synapse organization
regulation of cell differentiation
neuron differentiation
head development
negative regulation of synaptic transmission
neuron projection morphogenesis
post-embryonic development
chemical synaptic transmission, postsynaptic
excitatory postsynaptic potential
cellular component morphogenesis
neurogenesis
regulation of postsynaptic membrane potential
regulation of N-methyl-D-aspartate selective glutamate receptor activity
regulation of cytoskeleton organization
regulation of neurotransmitter receptor activity
cell surface receptor signaling pathway involved in cell-cell signaling
regulation of organelle organization
chromosome organization
cellular macromolecule catabolism
regulation of telomere maintenance
regulation of sister chromatid cohesion
histone modification
protein acylation
protein acetylation
DNA methylation or demethylation
positive regulation of gene expression
macromolecule methylation
DNA metabolism
RNA localization
protein modification by small protein conjugation or removal
positive regulation of macromolecule biosynthesis
1
9
-log10 adjusted p-value

Figure 4


A
Neurotypical Control
or
3q29Del Proband
iPSC
Neural Progenitors
Klf4
Oct3/4
Sox2
c-Myc
Dual SMAD inhibition
B
iPSC
Phase (20x)
TRA-1-60
SOX2
SSEA4
OCT4
Nanog
TRA-1-81
Neural Progenitor
Control
3q29Del
Nestin/SOX2/DAPI
Nestin/PAX6/DAPI
C
−Log10 p−value
7.5
5.0
2.5
0.0
−2
0
2
SENP5
DLG1
PIGX
NCBP2
UBXN7
PCYT1A
FBXO45
CEP19
TCTEX1D2
RNF168
WDR53
SLC51A
PAK2
MFI2
BDH1
TFRC
NS
p−value < 0.05
D
Chr3
NPC RNA−seq – 3q29 Genes
groups
Control
3q29Del
TFRC
ZDHHC19
SLC51A
PCYT1A
TCTEX1D2
TM4SF19
UBXN7
RNF168
SMCO1
WDR53
FBXO45
NRROS
CEP19
PIGX
PAK2
SENP5
NCBP2
PIGZ
MFI2
DLG1
BDH1
TFRC*
SLC51A*
PCYT1A**
TCTEX1D2**
UBXN7**
RNF168**
WDR53*
FBXO45**
NRROS
CEP19**
PIGX**
PAK2*
SENP5**
NCBP2**
PIGZ
MFI2*
DLG1**
BDH1*
22−0001
9002−0031
MH0200876
9001−0031
21−2096
21−2046
04C38098
22−1031
00C04409
20−1031
21−1031 c12
19−1001
21−1031 c11
E
RNA-Seq validated
WGCNA-based predictions
Broad 3q29 network
206 DEGs
p = 2.8e-09
Top-neighbor-based 3q29 subnetwork
84 DEGs
p = 3.5e-03
Prioritized drivers
13 DEGs
p = 3.0e-05
groups
Control
3q29Del
F
NPC Age Correlation
8pcw
9pcw
12pcw
13pcw
16pcw
17pcw
19pcw
21pcw
24pcw
25pcw
26pcw
35pcw
37pcw
4m
10m
1y
2y
3y
4y
8y
11y
13y
15y
18y
19y
21y
23y
30y
36y
37y
40y
Spearman ρ
Control
3q29Del