Deconvolution of Transcriptional Networks Identifies TCF4 as a Master Regulator in Schizophrenia

Abolfazl Doostparast Torshizi[1,2#], Chris Armoskus[3,4], Siwei Zhang[5,6], Hanwen Zhang[5], Tade Souaiaia[3,4], Marc P. Forrest[7], Oleg V. Evgrafov[3,4], James A. Knowles[3,4], Jubao Duan[5,6*], Kai Wang[1,2,4,#*]

[1]Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[2]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[3]College of Medicine, SUNY Downstate Medical Center, Brooklyn, NY 11203, USA

[4]Zilkhe Neurogenetic Institute, University of Southern California, Los Angeles, CA 90089, USA

[5]Center for Psychiatric Genetics, North Shore University Health System, Evanston, IL 60201, USA

[6]Department of Psychiatry and Behavioral Neurosciences, University of Chicago, Chicago, IL 60015, USA

[7]Department of Physiology, Northwestern University, Chicago, IL 60611, USA

[#] Prior address: Department of Biomedical Informatics and Institute for Genomic Medicine, Columbia University, New York, NY 10032, USA

* To whom correspondence should be addressed

Abstract

Tissue-specific reverse engineering of transcriptional networks has uncovered master regulators (MRs) of cellular networks in various cancers, yet the application of this method to neuropsychiatric disorders is largely unexplored. Here, using RNA-Seq data on postmortem dorsolateral prefrontal cortex (DLPFC) from schizophrenia (SCZ) patients and control subjects, we deconvoluted the transcriptional network to identify transcriptional MRs that mediate expression of a large body of target genes. Together with an independent RNA-Seq data on cultured primary neuronal cells derived from olfactory neuroepithelium, we identified *TCF4*, a leading SCZ risk locus implicated by genome-wide association studies, as a candidate MR dysregulated in SCZ. We validated the dysregulated *TCF4-related* transcriptional network through examining the transcription factor binding footprints inferred from human induced pluripotent stem cell (hiPSC)-derived neuronal ATAC-Seq data and direct binding sites obtained from ChIP-seq data in SH-SY5Y cells. The predicted *TCF4* regulons were enriched for genes showing transcriptomic changes upon knockdown of *TCF4* in hiPSC-derived neural progenitor cells (NPC) and glutamatergic neurons (Glut_N), in which the hiPSC cell line was sampled from a SCZ patient. The altered *TCF4* gene network perturbations in NPC, as compared to that in Glut_N, was more similar to the expression differences in the *TCF4* gene network observed in the DLPFC of individuals with SCZ. Moreover, *TCF4*-associated gene expression changes in NPC were more enriched for pathways involved in neuronal activity, genome-wide significant SCZ risk genes, and SCZ-associated *de novo* mutations. Our results suggest that *TCF4* serves as a MR of a gene network that confers susceptibility to SCZ at early stage of neurodevelopment, highlighting the importance of network dysregulation involving hundreds of genes in conferring susceptibility to neuropsychiatric diseases.

Introduction

Schizophrenia (SCZ) is a debilitating neuropsychiatric disease, which affects about 1% of adults[1-3]. SCZ is a highly heritable disease, with heritability estimated to be up to 73%-83% in twin studies[4-6]; however, it's underlying genetic architecture remains incompletely understood[7]. Recent genome wide approaches have identified a plethora of common disease risk variants[8], rare copy-number variants (CNVs) of high penetrance[9, 10] and rare protein-altering mutations[11] that confer susceptibility to SCZ. Although exciting, the biology underlying these genetic findings remains poorly understood, prohibiting the development of targeted therapeutic strategies. A major challenge is the polygenic nature of SCZ, where risk alleles in many genes, as well as non-coding regions of the genome, across the full allelic frequency spectrum are involved, and likely act in interacting networks[7]. Therefore, given the large number of genes potentially involved, it has been challenging to identify which are the core set of genes that play major roles in the pathways or networks. It even becomes more challenging if the recent hypothesis of omnigenic model is true for SCZ, in which case almost all the genes expressed in a disease-relevant cell type may confer risk to the disease through widespread network interactions with a core set of genes[12]. Thus, it is imperative to identify disease-relevant core gene networks and possibly, the network master regulators (MRs), which are more likely to be targeted for therapeutic interventions[13-15], if they exist.

Transcriptional networks, as a harmonized orchestration of genomic and regulatory interactions, play a central role in mediating cellular processes through regulating gene expression. One of the most commonly used network-based modeling approaches of cellular processes are scale-free co-expression networks[16, 17]. However, co-expression networks and other similar networks are not comprehensive enough to fully recapitulate the entire underlying molecular interactions driving the disease phenotype[18]. Despite the wide adoption of co-expression network analysis approaches, these methods have several limitations[19], such as the inability to infer or incorporate causal regulatory relationships, the difficulty to handle mammalian genome-wide networks with many genes, and the presence of high false positive rates due to indirect connections. In contrast, information-theoretic deconvolution techniques[18] aim to infer causal relationships between transcription factors and their downstream regulon, and have been recently successfully applied in a wide range of complex diseases such as cancers[20], prostate differentiation[21], and neurodegenerative diseases[22].

Inferring disease-relevant transcriptional gene networks requires well-powered transcriptomic case-control datasets from disease-relevant tissues or cell types. For SCZ, although abnormal gene expression in multiple brain regions[23] and different types of neuronal cells[24-27] may be involved in disease pathogenesis, abnormalities in cellular and neurochemical functions of dorsolateral prefrontal cortex (DLPFC) have been demonstrated in SCZ patients[28]. Furthermore, DLPFC controls high-level cognitive functions, many of which are disturbed in SCZ. The Common Mind Consortium (CMC) RNA-Seq data on DLPFC, which includes 307 SCZ cases and 245 controls, is currently the largest genomic data set on postmortem brain samples for neuropsychiatric disorders[29]. The CMC study validates ~20% of SCZ loci having variants that can potentially contribute to the altered gene expression, where five of which involve only a single gene such as *FURIN*, *TSNARE1*, and *CNTN4*[29]. A gene co-expression analysis in the CMC study implicated a gene network relevant to SCZ, but does not address the question of whether a MR potentially orchestrates transcriptional architecture of SCZ. In this study, using CMC RNA-Seq data,

we reverse-engineered the regulatory processes mediating SCZ to identify critical MRs and to infer their role in orchestrating cellular transcriptional networks. We then examined the CMC-based prediction of MRs and gene networks using another independent SCZ case-control RNA-Seq data on Cultured Neuronal cells derived from Olfactory Neuroepithelium (CNON)[30]. Among the top candidate MRs identified from both datasets, we selected *TCF4*, a leading SCZ risk gene from a genome-wide association study (GWAS)[8], for empirical validation of predicted network activity using human induced pluripotent stem cell (hiPSC)-derived neurons. A general overview of both the computational and experimental stages of this study is provided in Figure 1. Based on the results, we identified *TCF4* as a MR in SCZ that likely contributes to SCZ susceptibility at the early stage of neurodevelopment.

Results

Deconvolution of transcriptional gene network of CMC data uncovers master regulators

The CMC transcriptomic study implicated 693 differentially expressed genes in SCZ postmortem brains[29]. However, the magnitude of case-control expression differences were very small, posting a challenge to inferring their biological relevance. We thus reverse-engineered the transcriptional regulatory networks to infer the MRs and their gene targets (regulons), from which we constructed their corresponding sub-networks. To achieve this, we employed Algorithm for Reconstruction of Accurate Cellular Networks (ARACNe) to reconstruct cellular networks[18, 19] (see Materials and Methods). This approach first identifies gene-gene co-regulatory patterns by the information theoretic measure Mutual Information, followed by network pruning through removing indirect relationships in which two genes are co-regulated through one or more intermediate entities. This enables us to observe network relationships that are more likely to represent direct transcriptional interactions (e.g., transcription factor binding to a target gene) or post-transcriptional interactions (e.g., microRNA-mediated gene inhibition). We identified 1,466 genes as hubs by computational analysis of the CMC RNA-Seq data (Supplementary Table 1), where the corresponding subnetworks for these hub genes contain 24,548 transcriptional interactions (Supplementary Table 2).

Using the reconstructed network, we next performed a virtual protein activity analysis to investigate the activity of the identified hub genes by taking into account the expression patterns of their downstream regulons through a dedicated probabilistic algorithm called VIPER (Virtual Inference of Protein-activity by Enriched Regulon analysis)[20]. This method exploits the regulator mode of action, the confidence of the regulator-target gene interactions, and the pleiotropic features of the target regulation. We fed the generated network into VIPER to evaluate whether any of the identified hub genes in the network has significant regulatory role in the expression degrees of its downstream target genes. We then ranked the hub genes by VIPER-adjusted activity P-values (representing the statistical significance of being a MR; see Methods and Materials). We further defined a short list of potential MRs (N=93) with the adjusted activity P-values less than 0.005.

In order to avoid any false positive findings and to make sure that the identified potential MRs are biologically proven to exert regulatory effects on their targets, we intersected the short list of potential

MRs with a comprehensive set of 2,198 genes curated from three sources[31-33], which are either transcription factors (TFs) or other transcriptional regulatory genes (Supplementary Fig. 1). This process gave us 5 potential MRs: *TCF4*, *NR1H2*, *HDAC9*, *ZNF436*, and *ZNF10.* Based on previous studies[34, 35], identification of activated MRs in a disease state may be confounded, when several of its regulons are all targets of a different transcriptional factor, that is, the "shadow effect" [34, 35]. Because of highly pleiotropic nature of transcriptional regulations, we evaluated the shadow effects in the constructed networks. We should note that during the process of network deconvolution, most of the potential co-regulatory edges have been trimmed to preserve the scale-free topology of the network, so we expect to observe no shadow effect. As expected, no shadow connections were observed in this analysis indicating a high confidence in the constructed transcriptional network to reflect likely real regulatory connection between a MR and its targets.

Network deconvolution of CNON data independently identifies *TCF4* as a MR

We next attempted a replication of the identified MRs in an independent RNA-Seq data set (with 23,920 transcripts) from CNON of 143 SCZ cases with 112 controls[30]. Using the similar set up criteria as used on the CMC data, we observed 1,836 TF or expression regulatory nodes in the constructed network, including 34,757 predicted interactions (Supplementary Table 4). To further analyze the activity of the identified hub genes, we ran VIPER on the constructed network and identified the top MRs.

The five identified MRs in the CMC network (*TCF4*, *NR1H2*, *HDAC9*, *ZNF436*, and *ZNF10)* were re-confirmed to be MRs within the network created from CNON data. Based on the deconvolution analysis, these five MRs were predicted to regulate the expression of 101, 68, 36, 66, and 43 regulons combined, respectively (Supplementary Table 5). Among these genes, *TCF4* is a well-established SCZ-associated risk factor identified by GWAS[8, 36], which has been validated in several additional studies[37-41]. Rare variants in *TCF4* have also been implicated in Pitt Hopkins Syndrome and intellectual disability[39, 42]. However, its regulatory mechanisms at a systems level is poorly understood. Therefore, we focused on *TCF4* for additional analysis.

To further evaluate the connection between *TCF4* dysregulation and SCZ, we conducted pathway enrichment analysis using WebGestalt[43]. The predicted *TCF4*-associated regulons were enriched for several pathways such as Notch signaling pathway (FDR= 1×10^{-4}) and long-term depression (FDR= 8×10^{-4}). Additionally, to check whether the promoter region of predicted *TCF4* targets are enriched for *TCF4* binding motif sequence, we conducted TF binding enrichment analysis (TFBEA) using JASPAR[44]. The *TCF4* motif sequence is extracted from JASPAR and is used for binding enrichment analysis (Figure 2a). We extracted the sequence of the target genes 2 kb upstream and 1 kb downstream of their Transcription Start Sites (TSS), and used a significance threshold of 0.80 to test the relative enrichment scores of its target genes. We found that the observed binding enrichment score is significantly higher than expected by chance (two-sided t-test $P=2.2 \times 10^{-16}$), based on a random list of genes outside of the *TCF4* subnetwork (Figure 2b), suggesting that genes within the *TCF4* regulon can be directly targeted by *TCF4* binding events.

Inferred *TCF4* regulons are significantly enriched for *TCF4* targets predicted from ATAC-seq and ChIP-seq experiments

Our observed enrichments of TF binding motif of TCF4 in their regulons was purely a DNA sequence-based prediction, so we next tested whether *TCF4* can physically interact with its regulons. To address this question, we first identified the target genes of TCF4 by examining the inferred empirical TF occupancy (i.e, TF-binding footprints) in an Assay for Transposase Accessible Chromatin (ATAC-Seq) data in human induced pluripotent stem cell (hiPSC) and hiPSC-derived glutamatergic neurons[45, 46]. 29 target genes out of 101 regulons of *TCF4* were found to be bound by *TCF4* in ATAC-seq (total *TCF4* targets in ATAC-seq data=3,276) (enrichment P=3.3 x 10^{-5} by Fisher's Exact Test). Furthermore, with the list of empirically determined *TCF4* targets in a recent *TCF4* ChIP-Seq study on human SH-SY5Y neuroblastoma cells[47] (a total of 10,604 peaks for 5,436 unique candidate target genes), we observed an overlap of 41 *TCF4* target between *TCF4* ChIP-seq data and our inferred *TCF4* regulons, representing a significant enrichment (P=5.28 x 10^{-5} by Fisher's exact test). Altogether, both ATAC-Seq and ChIP-Seq analyses demonstrated that our list of predicted regulons are significantly enriched for *TCF4* transcriptional regulatory targets (Figure 2c).

Network perturbation by *TCF4* Knockdown in hiPSC-derived neurons supports *TCF4* as a MR and a regulator of neurodevelopment

To further investigate the functional relevance of *TCF4* dysregulation in schizophrenia, we examined how the predicted *TCF4* transcriptional subnetwork changes with decreased *TCF4* expression in hiPSC-derived neuronal cells. We acknowledge that similar experiments have been performed on the SH-SY5Y neuroblastoma cell line[47, 48], which is a cell type that may be less relevant to SCZ. We first derived relatively pure NPCs (NESTIN+/PAX6+/SOX2+) from hiPSCs (TRA-1-60+/SSEA4+/NANOG+/OCT4+) (Figure 3a-d), which were further differentiated into glutamatergic neurons (Glut_N) (VGLUT1+/MAP2+) (Figure 3e). hiPSCs were captured from a SCZ patient (see Methods and Materials). To capture possible differential effects of *TCF4* on gene expression during different developmental stages, we carried out *TCF4* knockdown by lentiviral shRNA in NPCs of day 3 after initiating neural differentiation (Figure 3c-d) and in early stage of Glut_N of day 14 after initiating differentiation (Figure 3e). We achieved significant *TCF4* knockdown in both cell stages: *TCF4* expression in the knockdown group was reduced to 32.2% and 48.6% of that in the control groups, respectively (as measured by qPCR; Figure 3f-g, $P < 0.001$). We also performed RNA-Seq analysis on the same set of cells with three biological replicates at both time points in both cell types, and found that the reduction of *TCF4* transcript level was consistent to the qPCR results (Figure 3h).

Transcriptomic analysis of the RNA-Seq data of *TCF4* knockdown revealed 4,891 DE genes in the day 3 group (NPC data, Figure 3i), including 2,330 upregulated and 2,561 downregulated genes (FDR<0.05) (see Materials and Methods). Day 14 group (Glut_N data, Figure 3j) showed 3,152 DE genes, of which 1,862 were upregulated and 1,290 were downregulated (FDR<0.05) (Supplementary Table 7). To examine whether the gene network perturbed by *TCF4* knockdown in hiPSC-derived neurons is concordant with the predicted *TCF4* regulons, we compared the list of genes affected by *TCF4* knockdown and the list of 101 *TCF4* regulons from the CMC/CNON data. We found that 39 predicted *TCF4* regulons showed altered expression by *TCF4* knockdown in NPCs (enrichment P=1.55 x 10^{-6} by

Fisher's exact test), while 33 predicted *TCF4* regulons were dysregulated after *TCF4* knockdown in Glut_N (enrichment P =1.05 x 10^{-8} by Fisher's exact test). This difference between the enrichment P-values of the conducted experiments compared to the neuroblastoma cell line confirms the importance of tissue-specific investigation of transcriptional regulation (even though the set of regulons may be more conserved across tissues), and further supports the role of *TCF4* as a MR in a transcriptional network relevant to neurodevelopment (Figure 2c).

Although our deconvolution analysis identified 101 genes in the *TCF4* subnetwork, we reasoned that it is likely part of a much larger transcriptional network that would contain neurobiological pathways relevant to schizophrenia. Previous studies in mice[49] and neuroblastoma cells[48] have demonstrated that *TCF4* dysregulation affects a large number of genes and pathways however, *TCF4*-related pathways have never been explored in human NPCs or glutamatergic neurons, a much stronger model for studying psychiatric disorders. To uncover the cellular pathways regulated by *TCF4* in these cell types we performed gene ontology (GO) and pathway analyses on the genes perturbed by *TCF4* knockdown.

By using WebGestalt[43] we identified a number of shared gene pathways, including focal adhesion, axon guidance, MAPK signaling pathway, and apoptosis (Supplementary Table 6 and Supplementary Figs. 2-6). In the Supplementary Table 6, we showed the altered pathways at the both time points as a result of *TCF4* knockdown, and the detailed list of up/down regulated genes in each pathway. We found almost similar enrichment for "axon guidance" (FDR =1.521 x 10^{-6} in NPCs vs FDR =5.941 x 10^{-5} in Glut_Ns), while observing some unique pathways (long-term potentiation, neurotrophin signaling, and mTOR signaling) relevant to neuronal activity. An independent pathway analysis using Metacore (see Methods and Materials, Supplementary Figs. 7-8) also showed consistent results. The top gene expression changes in NPCs converged on early neurodevelopmental processes such as axon guidance, neurogenesis and attractive and repulsive receptors, which guide neuronal growth and axon targeting during development. In contrast, the most significant categories in Glut_Ns involved cell adhesion and cell-matrix interactions, suggesting that *TCF4* regulates different subsets of genes during development. Interestingly, the MetaCore analysis revealed that the upregulated genes had distinct functions in NPC and Glu_Ns, involved in mRNA translation and cell-matrix interactions respectively. However, down regulated genes shared a number of categories including axon guidance as the most significant.

For both upregulated and downregulated genes by *TCF4* knockdown, NPCs showed more enriched neuronal-activity related gene pathways than in Glut_Ns and with greater significance (Supplementary Tables 8-9). It is also noteworthy that most altered neuronal-related pathways are enriched in the down-regulated set of genes upon *TCF4* knockdown. These results suggest that *TCF4* gene targets act at early stage of neurogenesis and may be more relevant to SCZ biology.

Higher *TCF4* gene network expression activity in NPCs is more reflective of SCZ case-control difference in prefrontal cortex

Although analyses of *TCF4* knockdown in NPC and Glut_N support the role of *TCF4* as a MR, the *TCF4* gene network (interactome) expression activity, i.e., the correlation between *TCF4* expression and its regulon's expression patterns, may be different in either NPCs and Glut-N cellular stages. Such cell-type specific expression of the network activity may inform when the temporal expression of *TCF4* is most

relevant to SCZ. Towards this end, we carried out gene set enrichment analyses (GSEA) in the CMC/CNON data, as well as the *TCF4* knockdown RNA-Seq data in NPCs and in Glut_Ns. We found that the SCZ case/control expression differences of *TCF4* regulons in CMC data (i.e., prefrontal cortex) appeared to be more correlated with the *TCF4* regulons expression changes upon *TCF4* knockdown in NPCs than in Glut_Ns. To further demonstrate the closer similarity of the *TCF4* interactome expression activity in prefrontal cortex and in NPCs, we depicted the GSEA enrichment scores in three data sets (CMC/CNON, NPC and Glut_N) in a Circos plot (Supplementary Fig. 9a). We found that the pattern of SCZ case/control expression differences of *TCF4* and its regulon in prefrontal cortex appeared to be more similar to the pattern of higher *TCF4* network expression activity in NPCs than in Glut_Ns. Similarly, the GSEA rank metric score of the *TCF4* regulons also showed that the SCZ case/control differential *TCF4* interactome expression pattern in prefrontal cortex was more comparable to that in NPCs than in Glut_Ns (Supplementary Fig. 9b). These analyses thus suggest that higher *TCF4* expression and its positively correlated regulon expression activity at early stage of neurodevelopment and neurogenesis may be a risk factor of SCZ.

TCF4 gene network in NPCs is more relevant to SCZ disease biology

Given that the expression pattern of *TCF4* and it regulons in prefrontal cortex are more correlated to NPCs than Glut_Ns, we reasoned that *TCF4* gene network in NPCs may be more relevant to SCZ disease biology. To test this hypothesis, we have examined the enrichments of genes regulated by *TCF4* in NPCs, as compared to Glut_Ns for disease-relevant gene pathways, SCZ GWAS risk genes, and *de novo* SCZ mutations.To test whether the *TCF4* gene network acting in NPCs is more relevant to the genetic etiology of SCZ, we compared a list of known SCZ susceptibility genes from genome-wide significant risk loci[8] that showed gene expression change upon *TCF4* knockdown in NPCs and in Glut_Ns. We found that 52 of those genes (out of 148 reported in reference[8]) were altered by *TCF4* across at least one of two time points in either NPCs or Glut_Ns. Of these 52 SCZ credible genes that showed *TCF4*-altered expression, 63.5% were uniquely found in NPCs (*N*=33) while 25% (*N*=13) were uniquely found in Glut_Ns (Figure 3k). Furthermore, for the 92 SCZ-associated single nucleotide variants (SNVs) that were indexed for genes related to neuronal electrical excitability and neurotransmission[50], 45 had altered expression by *TCF4* knockdown either in NPCs or Glut_Ns. Among them, 24 genes (53.34%) were uniquely found in NPCs while only 7 genes (15.6%) were uniquely found in Glut_Ns (Figure 3l). Lastly, we conducted SCZ *de novo* SNV (dnSNV) enrichment analysis to probe the overlap of *TCF4*-associated gene expression changes and dnSNV in SCZ (Figure 4a). We found that while *TCF4*-affected genes in NPCs showed significant enrichment for both loss-of-function (LoF) and missense SCZ dnSNVs, and coding dnSNVs in general (Figure 4b), only LoF dnSNVs were enriched in Glut_Ns (Figure 4c). Thus, the analyses of both common variants and dnSNVs associated with SCZ in *TCF4*-affected genes nominated NPC as a more disease-relevant cell-type and stage of development for *TCF4* gene network. Altogether, our results suggest that *TCF4* and its regulons in early stage of neurodevelopment and neurogenesis may be more relevant for SCZ pathogenesis.

Discussion

Like other complex disorders, SCZ is polygenic or even possibly omnigenic[12] with hundreds or even thousands of susceptibility genes each with small effect size. A major challenge for understanding the biological implications of genetics findings is to identify the core gene networks in disease relevant tissues or cell types[12]. In this study, with the aim of uncovering the gene expression-mediating drivers that may contribute to SCZ pathogenesis, we have employed a data-driven approach to assemble tissue-specific transcriptional regulatory networks on two independent RNA-Seq datasets from dorsolateral prefrontal cortex (CMC data) and cultured primary neuronal cells derived from olfactory neuroepithelium (CNON data). We identified *TCF4* as a MR based on the deconvoluted transcriptional networks from the two independent data sets. For the predicted gene network of *TCF4*, a known leading SCZ GWAS locus, we have empirically validated the predicted *TCF4* gene targets by analyzing both TF-binding footprints in neuronal cells and by transcriptomic profiling of *TCF4*-associated gene expression changes in both hiPSC-derived NPCs and Glut_Ns upon *TCF4* knockdown.

Compared to the commonly used weighted gene co-expression network analysis (WGCNA)[51], the method used in our study aims to find MRs rather than identifying co-expression patterns as well as bearing a number of unique features, including the ability to elucidate potentially causal interactions, to illustrate the hierarchical structure of the identified transcriptional networks, and to identify potential feed-forward loops between MRs that may cause synergistic effects in the network. The confidence of the identified MRs and their subnetworks was further strengthened by their presence in both transcriptomic datasets, and experimentally validated in hiPSC-derived neuronal cells using ATAC-Seq data, ChIP-Seq data, and transcriptomic profiling.

TCF4 is one of the leading SCZ risk genes identified by GWAS[8]. However, the molecular mechanism underlying the genetic association remain elusive. Here, we have shown computationally that *TCF4* is one of the top MRs of SCZ, and this was empirically validated in hiPSC model of mental disorders. Although the downstream gene targets of *TCF4* have been previously profiled in a neuroblastoma cell lines[48], the hiPSC cellular model used here enabled us to examine the neurodevelopmental relevance of *TCF4* gene network. By examining the *TCF4*-pertubred gene networks in hiPSC-derived NPC (early neuronal development/neurogenesis stage) and in differentiated Glut_N, we found that *TCF4*-altered genes in NPC are more enriched for gene pathways that are related to neuronal activities. Furthermore, we have shown that *TCF4*-altered genes in NPC are more enriched for credible SCZ GWAS risk genes and for SCZ-associated dnSNVs. Thus, multiple lines of evidence from our study suggest that *TCF4* gene network expression activity in the early stage of neurodevelopment and neurogenesis may be more important for SCZ pathogenesis.

Although our empirical validation of MR has focused on *TCF4*, other MRs that we identified here suggest that additional gene networks are relevant to SCZ. Indeed, the identified MRs contain many of the TFs previously reported in the literature that may be potentially involved in the pathogenesis of SCZ, such as *JUND*[52], *KCNIP3*[53], *LHX6*[54], *NRG1* and *GSK3B*[55], *NFE2* and *MZF1*[56], *NPAS3*[57], and *DISC1*[58]. Among those newly identified MRs, methylation patterns of *TMEM9*[59] has been reported to be associated with Parkinson's disease, yet *CRH* is a protein-coding gene associated with Alzheimer's disease, depression,

and SCZ[60, 61], and *ARPP19* has been identified to be associated with nerve terminal function[62]. A particularly noteworthy new MR is *HDAC9*, a histone deacetylase inhibitor that is important for chromatin remodeling. It has been shown to regulate mouse cortical neuronal dendritic complexities[63] as well as hippocampal-dependent memory and synaptic plasticity under different neuropsychiatric conditions[64]. In our deconvoluted gene network of *HDAC9*, its regulons are enriched for genes related to focal adhesion (*P*=0.0115), a biological function that has been shown to be altered in SCZ[65]. Of note, *TCF4*-altered genes in both NPCs and Glut-Ns also showed enrichment of focal adhesion pathway. The MRs we identified, and their subnetworks may provide a rich resource of SCZ-relevant gene networks that can be perturbed to increase our understanding of SCZ disease pathophysiology.

In summary, by employing both a computational approach and a hiPSC neuronal developmental model, we have identified *TCF4* as a MR that likely contributes to SCZ susceptibility at the early stage of neurodevelopment. Our study suggests that MRs in SCZ can be identified by transcriptional network deconvolution, and that MRs such as *TCF4* may constitute convergent gene networks that confer disease risk in a spatial and temporal manner. Transcriptional network deconvolution in larger and more SCZ-relevant cell types/stages, combined with empirical network perturbation, will further deepen our understanding of the genetic contribution to SCZ disease biology.

Methods and Materials

RNA-Seq data of Dorsolateral Prefrontal Cortex (DLPFC)

RNA-seq data of DLPFC (Dorsolateral Prefrontal Cortex) were downloaded from "normalized SVA corrected" directory of the CommonMind Consortium (CMC) Knowledge Portal (https://www.synapse.org/#!Synapse:syn5609499) using Synapse Python client (http://docs.synapse.org/python/). The data being used in this study contains 307 SCZ cases and 245 controls where the paired-end RNA-Seq data has been generated on an Illumina HiSeq 2500. Brain tissues in this dataset were obtained from several brain bank collections including: Mount Sinai NIH Brain and Tissue Repository, University of Pittsburgh NeuroBioBAnk and Brain and Tissue repositories, University of Pennsylvania Alzheimer's Disease Core Center, and the NIMH Human Brain Collection Core. Detailed procedure of tissue collection, sample preparation, data generation and processing can be found in the CommonMind Consortium Knowledge Portal Wiki page (https://www.synapse.org/#!Synapse:syn2759792/wiki/69613).

RNA-Seq data of Cultured Primary Neuronal cells derived from Olfactory Neuroepithelium (CNON)

The normalized RNA-Seq read counts from 143 SCZ cases and 112 controls extracted from cultured primary neuronal cells derived from neuroepithelium were generated as previously described[30]. Similar to the CMC data, the raw CNON data consists of 100 bp paired-end reads. The data has been pre-processed to exclude rRNA and mitochondrial genes. Later, normalized counts were calculated by the DESeq2 software. These analyses are available at:
https://www.biorxiv.org/content/early/2017/10/26/209197

ARACNe network reconstruction

ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks), an information-theoretic algorithm for inferring transcriptional interactions, was used to identify candidate transcriptional regulators of the transcripts annotated to genes both in CMC and CNON data. First, mutual interaction between a candidate TF(x) and its potential target (y) was computed by pairwise mutual information, MI(x, y), using a Gaussian kernel estimator. A threshold was applied on MI based on the null-hypothesis of statistical independence (P<0.05, Bonferroni corrected for the number of tested pairs). Second, the constructed network was trimmed by removing indirect interactions the data processing inequality (DPI), a property of the MI. Therefore, for each (x, y) pair, a path through another TF(z) was considered and every path pertaining the following constraint were removed ($MI(x,y) < \min(MI(x,z), MI(z,y))$). P-value threshold of 1 x 10^{-8} using DPI=0.1 (as recommended[19]) were used when running ARACNe.

Virtual protein activity analysis

The regulon enrichment on gene expression signatures was tested by the VIPER (Virtual Inference of Protein-activity by Enriched Regulon analysis) algorithm. First, the gene expression signature is obtained by comparing two groups of samples representing distinctive phenotypes or treatments. In the next step, regulon enrichment on the gene expression signature can be computed using Analytic rank-based enrichment analysis (aREA). At the end, significance values (P-value and normalized enrichment score) are computed by comparing each regulon enrichment score to a null model generated by randomly and uniformly permuting the samples 1,000 times. The output of VIPER is a list of highly active MRs as well as their activity scores and their enrichment P-values. Further information about VIPER can be accessed in reference[20].

Transcription factor binding site enrichment and footprint analysis

Human reference genome (hg19) was used to extract the DNA sequence around transcript start sites (TSSs) for transcription binding enrichment analysis. We obtained the gene coordinates from Ensembl BioMart tool[66] and scanned 2000 bp upstream and 1000 bp downstream of the TSS. The motifs of the TFs were obtained from JASPAR and the extracted sequences of each target were then fed into JASPAR and analyzed versus their corresponding TFs. Then, using the PWM of the TF, JASPAR employs the modified Needleman-Wunsch algorithm to align the motif sequence with the target sequence in that the input sequence is scanned to check whether or not the motif is enriched. The output is the enrichment score of the input TF in the designated target genes.

We used PIQ to assess the local TF occupancy footprint from ATAC-Seq data[45, 67]. We extracted the corresponding BED files for TF footprint analysis using the PIQ R package. All the footprints were annotated using the TF matrix with the names of different TFs annotated in the BED files. For each sample, footprints were generated using three different PIQ purity scores (0.7, 0.8, or 0.9; equivalent to FDR of 0.3, 0.2, or 0.1, respectively). The corresponding files are then extracted using the MR list and the peak names/coordinates containing TCF4 gene are collected as a subset of the original BED file. Such subsets of genomic coordination are then annotated using the findPeaks.pl included in the HOMER package with the hg19 reference genome.

hiPSC generation and cell culture

The hiPSC line used for deriving neurons was generated by Sendai virus method at Rutgers University Cell and DNA Repository (RUCDR)-NIMH Stem Cell Center from cryopreserved lymphocytes (CPLs) of a 29 year-old male SCZ patient (cell ID: 07C65853) from the MGS collection[68]. The NorthShore University HealthSystem Institutional Review Board (IRB) approved the study. hiPSCs were cultured using the feeder-free method in Geltrex (Thermofisher)-coated plates in mTeSR1 medium (StemCell). The media were changed daily, and cells were passaged as clumps every 5 days using ReLeSR (StemCell) in the presence of 5 µm ROCK inhibitor (R&D Systems). hiPSCs were characterized by positive IF staining of pluripotent stem cell markers OCT4, TRA-1-60, NANOG, and SSEA4 (Figure 3a-b).

hiPSC-derived neural progenitor cells (NPCs)

NPCs were differentiated from hiPSC using the PSC Neural Induction Medium (Thermofisher) with modifications. In brief, hiPSCs were re-plated as clumps in Geltrex (Thermofisher)-coated 6-well plates (Thermofisher) in mTeSR1 (StemCell) on Day 0. From Day 2, the medium was switched to PSC Neural Induction Medium and changed daily. On Day 11, NPCs were harvested using Neural Rosette Selection Reagent (StemCell). NPCs were maintained in Neural Expansion Medium (Thermofisher) and passaged every 4-6 day in the presence of 5 µM ROCK inhibitor (R&D Systems) until the 4th passage (P4). P4 NPCs were characterized by positive IF staining for Nestin and Pax6 (Figure 3c-d).

hiPSC differentiation and shRNA-mediated TCF4 (ITF-2) knockdown

hiPSC-derived NPCs were plated at the density of 3 x 10^5 per well in a 12-well plate pre-coated with Geltrex for 2 hours. The differentiation media used for induced differentiation of NPCs contains neural basal medium, B-27 supplement, L-glutamine, 10 ng/ml BDNF, 10 ng/ml GDNF, and 10 ng/ml NT-3[45]. NPCs were allowed to grow and differentiate for 3 or 14 days post induction, respectively. At the corresponding day, 20 µl (1.0 x 10e5 infection units of virus, IFU) of either *TCF4* or control shRNA lentiviral particles (Santa Cruz sc-61657-V for *TCF4*, which contains a pool of 3 target-specific, propriety constructs targeting human *TCF4* gene locus at 18q21.2; and sc-108080 for control, which contains a set of scrambled non-specific shRNA sequence) were added into each well (3 replicates for each group), and the transduced cells were further cultured for 48 hours before collection. Total RNAs were extracted from homogenized cell lysate using RNEasy Plus mini kit (Qiagen), and the quality of extracted total RNA were examined using Nanodrop spectrum analysis and agarose gel electrophoresis. The day-14 excitatory neurons were characterized by positive IF staining for VGLUT1 and PSD95 (Figure 3e).

qPCR validation of TCF4 shRNA knockdown efficiency

We used qPCR to verify the knockdown efficiency of shRNA before applying RNA-Seq. Briefly, cDNAs were reverse-transcribed from total RNAs using the High-Capacity Reverse Transcription Kit with RNase Inhibitor (ThermoFisher). 50 ng of total RNA were used for each reverse-transcription reaction (20 µl) according to the manufacturer's instructions. cDNA was diluted 15-fold before applying to qPCR. Subsequent qPCR analysis was performed on a Roche LightCycler 480 II real-time PCR machine using TaqMan probes against *TCF4* and *GAPDH* (as internal reference) respectively. Briefly, 5 µl diluted RT product was applied to each 10 µl qPCR reaction using the TaqMan Universal Amplification kit (ThermoFisher/AppliedBiosystems) with customized TaqMan probes (Hs.PT.58.21450367 for TCF4,

Integrated DNA Technologies; and Hs03929097_g1 for GAPDH, AppliedBiosystems). Cycle parameters: 95°C, 10 min; 95°C, 15 sec, 60°C 1 min, 45 cycles. Data analysis was performed using the build-in analysis software of Roche 480 II with relative quantification (ΔCt method applied).

RNA-seq and data analysis

RNA-seq libraries were prepared from total RNAs from the collected neuronal cells of 12 samples (three biological replicates, i.e., independent cell cultures, of four distinct groups). These sample groups are for *TCF4* expression knockdown and control at hiPSC-derived NPC stage (day 3 post-differentiation) and Glut-N stage (day 14 post-differentiation). Total RNAs were extracted as described above. Three main methods for quality control (QC) of RNA samples were conducted including: (1) preliminary quantification; (2) testing RNA degradation and potential contamination (Agarose Gel Electrophoresis); (3) checking RNA integrity and quantification (Agilent 2100). After the QC procedures, library was constructed and library QC was conducted consisting three steps including: (1) testing the library concentration (Qubit 2.0); (2) testing the insert size (Agilent 2100); (3) precise quantification of library effective concentration (qPCR). The quantified libraries were then sequenced using Illumina HiSeq/MiSeq sequencers with 150 bp paired-end reads, after pooling according to its effective concentration and expected data volume.

The paired-end sequenced reads for each sample were obtained as FASTQ files. Initial quality check of the raw data was performed using FastQC. All FASTQ files passed the quality control stage including removing low quality bases, adapters, short sequences, and checking for rRNA contamination. The average number of pre-processed reads per sample was ~22,740,000. High quality reads were mapped to the reference genome GRCh38 using STAR[69]. Sorting and counting was also performed using STAR based on the procedure recommended by the developers. To detect DE genes, the output read counts were fed to edgeR software package[70]. Normalization, and differential expression analysis was conducted using the recommended settings.

Pathway enrichment and GO analysis

Pathway enrichment and GO analysis were conducted using WebGestalt[43]. KEGG was used as the functional database and human reference genome was used as the background. The maximum and minimum number of genes for each category were set to 2000 and 5, respectively based on the default setting. Bonferroni-Hochberg (BH) multiple test adjustment was applied to the enrichment output. FDR significance threshold was set to 0.05.

MetaCore pathway analysis

Ensembl transcript IDs of differentially expressed genes were imported in Metacore (v 6.32) and processed using the 'Enrichment Analysis' workflow. For each time point (3 day or 14 day), all the upregulated (FDR<0.05), all downregulated (FDR<0.05) and the top 2000 differentially expressed genes (FDR<0.05) were analyzed. Each list was compared to a defined background set of genes consisting of all the expressed transcripts (>1 FPKM) at the relevant time point (3 day or 14 day). The 'Pathway Maps' or 'Process networks' that were enriched in each analysis were considered significant when FDR<0.05.

Enrichment analysis on de novo mutations in schizophrenia

Data on *de novo* variants in schizophrenia and controls was obtained from http://denovo-db.gs.washington.edu/denovo-db [accessed 11/04/2017][71]. Variants identified in schizophrenia and control subjects were stratified into four categories (all protein coding, missense, loss-of-function and synonymous) based on annotations from the denovo-db. To generate a list of variants affecting protein coding ('all protein coding'), *de novo* variants outside exonic coding regions were excluded (including *de novos* in 5'-UTRs, 3'-UTRs, intronic and intergenic regions and non-coding exons) as well as synonymous variants. The 'loss-of-function' list included frameshift, stop gained, stop lost, start lost, splice acceptor and donor variants. We then compiled a total of six lists containing different sets of differentially expressed genes from the *TCF4* knockdown (KD) experiments at day 3 and day 14, as follows: day3_all_sig (all differentially expressed genes FDR<0.05; 4898 genes), day3_up_sig (all upregulated genes FDR<0.05; 2332 genes), day3_down_sig (all downregulated genes FDR<0.05; 2566), day14_all_sig (all differentially expressed genes FDR<0.05; 3156 genes), day14_up_sig (all upregulated genes FDR<0.05; 1864 genes), day14_down_sig (all downregulated genes FDR<0.05; 1292). Each of the *de novo* lists for schizophrenia and controls were intersected with the relevant list of differentially expressed genes from the *TCF4* KD experiments to determine the shared number of genes in each comparison. A hypergeometric test was used to test the statistical significance of each overlap (number of shared genes between lists) and using all protein coding genes as a background set (20,338, human genome build GRCh38.p10). The data is displayed as the –log10 (p value) for each comparison.

Acknowledgements

We appreciate the CommonMind Consortium to provide the RNA-seq data on dorsolateral prefrontal cortex from patients and control subjects (The data were generated as part of the CommonMind Consortium supported by funding from Takeda Pharmaceuticals Company Limited, F. Hoffman-La Roche Ltd and NIH grants R01MH085542, R01MH093725, P50MH080405, R01MH097276, R01MH075916, P50MH096891, P50MH084053S1, R37MH057881 and R37MH057881S1, HHSN271201300031C, AG02219, AG05138 and MH06692). We thank the developers of the ARACNe software for comments on the software usage, and thank the Wang lab members for helpful suggestions and discussions. This study was supported by NIH grant MH108728 (K.W.), HG006465 (K.W.), MH086874 (O.E. and J.A.K.), MH102685 and MH106575 (J.D.).

Competing interests

The authors declare no competing interests.

Author Contributions

A.D.T. conceived the study, designed the study, performed the computational experiments, analyzed the data, and wrote the manuscript. C.A., O.V.E., T.S., and J.A.K. generated and pre-processed the CNON data and J.A.K. edited the manuscript. S.Z. and H.Z. generated the hiPSC cells, conducted *TCF4* knockdown and other experiments as well as writing the relevant Methods sections. M.P.F. advised on

the study and performed additional enrichment analysis on *TCF4*. K.W. conceived the study, wrote and edited the manuscript, and supervised the project. J.D. supervised the experimental sections of the study, wrote and edited the manuscript.

References

1. McGrath, J., Saha, S., Chant, D. & Welham, J. Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiol Rev* **30**, 67-76 (2008).
2. Kessler, R.C*., et al.* The prevalence and correlates of nonaffective psychosis in the National Comorbidity Survey Replication (NCS-R). *Biol Psychiatry* **58**, 668-676 (2005).
3. Saha, S., Chant, D., Welham, J. & McGrath, J. A systematic review of the prevalence of schizophrenia. *PLoS Med* **2**, e141 (2005).
4. Hilker, R*., et al.* Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register. *Biol Psychiatry* **83**, 492-498 (2018).
5. Sullivan, P.F., Kendler, K.S. & Neale, M.C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch Gen Psychiatry* **60**, 1187-1192 (2003).
6. Cardno, A.G*., et al.* Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. *Arch Gen Psychiatry* **56**, 162-168 (1999).
7. Rees, E., O'Donovan, M.C. & Owen, M.J. Genetics of schizophrenia. *Current Opinion in Behavioral Sciences* **2**, 8-14 (2015).
8. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427 (2014).
9. Marshall, C.R*., et al.* Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat Genet* **49**, 27-35 (2017).
10. Genovese, G*., et al.* Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nature Neuroscience* **19**, 1433-1441 (2016).
11. Singh, T*., et al.* Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat Neurosci* **19**, 571-577 (2016).
12. Boyle, E.A., Li, Y.I. & Pritchard, J.K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186 (2017).
13. Mondragon, E. & Maher, L.J., 3rd. Anti-Transcription Factor RNA Aptamers as Potential Therapeutics. *Nucleic Acid Ther* **26**, 29-43 (2016).
14. Zhang, Y*., et al.* Rapid single-step induction of functional neurons from human pluripotent stem cells. *Neuron* **78**, 785-798 (2013).
15. Corbett, B.F*., et al.* DeltaFosB Regulates Gene Expression and Cognitive Dysfunction in a Mouse Model of Alzheimer's Disease. *Cell Rep* **20**, 344-355 (2017).
16. Lukashin, A.V., Lukashev, M.E. & Fuchs, R. Topology of gene expression networks as revealed by data mining and modeling. *Bioinformatics* **19**, 1909-1916 (2003).
17. Jordan, I.K., Marino-Ramirez, L., Wolf, Y.I. & Koonin, E.V. Conservation and coevolution in the scale-free human gene coexpression network. *Mol Biol Evol* **21**, 2058-2070 (2004).
18. Basso, K*., et al.* Reverse engineering of regulatory networks in human B cells. *Nat Genet* **37**, 382-390 (2005).
19. Margolin, A.A*., et al.* Reverse engineering cellular networks. *Nat Protoc* **1**, 662-671 (2006).
20. Alvarez, M.J*., et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature Genetics* **48**, 838-+ (2016).

21. Dutta, A.*, et al.* Identification of an NKX3.1-G9a-UTY transcriptional regulatory network that controls prostate differentiation. *Science* **352**, 1576-1580 (2016).
22. Brichta, L.*, et al.* Identification of neurodegenerative factors using translatome-regulatory network analysis. *Nat Neurosci* **18**, 1325-+ (2015).
23. Haroutunian, V., Katsel, P., Dracheva, S., Stewart, D.G. & Davis, K.L. Variations in oligodendrocyte-related gene expression across multiple cortical regions: implications for the pathophysiology of schizophrenia. *Int J Neuropsychopharmacol* **10**, 565-573 (2007).
24. Hashimoto, T.*, et al.* Alterations in GABA-related transcriptome in the dorsolateral prefrontal cortex of subjects with schizophrenia. *Mol Psychiatry* **13**, 147-161 (2008).
25. Hakak, Y.*, et al.* Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. *Proc Natl Acad Sci U S A* **98**, 4746-4751 (2001).
26. O'Connor, J.A. & Hemby, S.E. Elevated GRIA1 mRNA expression in Layer II/III and V pyramidal cells of the DLPFC in schizophrenia. *Schizophr Res* **97**, 277-288 (2007).
27. Arion, D., Horvath, S., Lewis, D.A. & Mirnics, K. Infragranular gene expression disturbances in the prefrontal cortex in schizophrenia: signature of altered neural development? *Neurobiol Dis* **37**, 738-746 (2010).
28. Guillozet-Bongaarts, A.L.*, et al.* Altered gene expression in the dorsolateral prefrontal cortex of individuals with schizophrenia. *Mol Psychiatry* **19**, 478-485 (2014).
29. Fromer, M.*, et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* **19**, 1442-1453 (2016).
30. Evgrafov, O.V.*, et al.* Gene expression in patient-derived neural progenitors provide insights into neurodevelopmental aspects of schizophrenia. *bioRxiv* (2017).
31. Lizio, M.*, et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* **16**, 22 (2015).
32. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**, 252-263 (2009).
33. Han, H.*, et al.* TRRUST: a reference database of human transcriptional regulatory interactions. *Sci Rep* **5**, 11432 (2015).
34. Lefebvre, C.*, et al.* A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol Syst Biol* **6** (2010).
35. Jiang, Z. & Gentleman, R. Extensions to gene set enrichment. *Bioinformatics* **23**, 306-313 (2007).
36. Stefansson, H.*, et al.* Common variants conferring risk of schizophrenia. *Nature* **460**, 744-747 (2009).
37. Steinberg, S.*, et al.* Common variants at VRK2 and TCF4 conferring risk of schizophrenia. *Hum Mol Genet* **20**, 4076-4081 (2011).
38. Schizophrenia Psychiatric Genome-Wide Association Study, C. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* **43**, 969-976 (2011).
39. Kharbanda, M.*, et al.* Partial deletion of TCF4 in three generation family with non-syndromic intellectual disability, without features of Pitt-Hopkins syndrome. *Eur J Med Genet* **59**, 310-314 (2016).
40. Blake, D.J.*, et al.* TCF4, schizophrenia, and Pitt-Hopkins Syndrome. *Schizophr Bull* **36**, 443-447 (2010).
41. Quednow, B.B., Brzozka, M.M. & Rossner, M.J. Transcription factor 4 (TCF4) and schizophrenia: integrating the animal and the human perspective. *Cell Mol Life Sci* **71**, 2815-2835 (2014).
42. Forrest, M.P., Hill, M.J., Quantock, A.J., Martin-Rendon, E. & Blake, D.J. The emerging roles of TCF4 in disease and development. *Trends Mol Med* **20**, 322-331 (2014).
43. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res* **41**, W77-83 (2013).

44. Mathelier, A.*, et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **44**, D110-115 (2016).
45. Forrest, M.P.*, et al.* Open Chromatin Profiling in hiPSC-Derived Neurons Prioritizes Functional Noncoding Psychiatric Risk Variants and Highlights Neurodevelopmental Loci. *Cell Stem Cell* **21**, 305-318 e308 (2017).
46. Zhang, S.*, et al.* Open chromatin dynamics reveals stage-specific transcriptional networks in hiPSC-based neurodevelopmental model. *Stem Cell Res* **29**, 88-98 (2018).
47. Forrest, M.P.*, et al.* The Psychiatric Risk Gene Transcription Factor 4 (TCF4) Regulates Neurodevelopmental Pathways Associated With Schizophrenia, Autism, and Intellectual Disability. *Schizophrenia Bulletin*, sbx164-sbx164 (2017).
48. Forrest, M.P., Waite, A.J., Martin-Rendon, E. & Blake, D.J. Knockdown of human TCF4 affects multiple signaling pathways involved in cell survival, epithelial to mesenchymal transition and neuronal differentiation. *PLoS One* **8**, e73169 (2013).
49. Kennedy, A.J.*, et al.* Tcf4 Regulates Synaptic Plasticity, DNA Methylation, and Memory Function. *Cell Rep* **16**, 2666-2685 (2016).
50. Devor, A.*, et al.* Genetic evidence for role of integration of fast and slow neurotransmission in schizophrenia. *Mol Psychiatry* **22**, 792-801 (2017).
51. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
52. Boyajyan, A.S., Atshemyan, S.A. & Zakharyan, R.V. Association of schizophrenia with variants of genes that encode transcription factors. *Mol Biol+* **49**, 875-880 (2015).
53. Van Schijndel, J.E. & Martens, G.J.M. Gene Expression Profiling in Rodent Models for Schizophrenia. *Curr Neuropharmacol* **8**, 382-393 (2010).
54. Volk, D.W., Edelson, J.R. & Lewis, D.A. Cortical inhibitory neuron disturbances in schizophrenia: role of the ontogenetic transcription factor Lhx6. *Schizophr Bull* **40**, 1053-1061 (2014).
55. Emamian, E.S. AKT/GSK3 signaling pathway and schizophrenia. *Front Mol Neurosci* **5**, 33 (2012).
56. Guo, A.Y., Sun, J., Jia, P. & Zhao, Z. A novel microRNA and transcription factor mediated regulatory network in schizophrenia. *BMC Syst Biol* **4**, 10 (2010).
57. Pickard, B.S., Malloy, M.P., Porteous, D.J., Blackwood, D.H.R. & Muir, W.J. Disruption of a brain transcription factor, NPAS3, is associated with schizophrenia and learning disability. *Am J Med Genet B* **136b**, 26-32 (2005).
58. Millar, J.K.*, et al.* Genomic structure and localisation within a linkage hotspot of Disrupted In Schizophrenia 1, a gene disrupted by a translocation segregating with schizophrenia. *Mol Psychiatr* **6**, 173-178 (2001).
59. Masliah, E., Dumaop, W., Galasko, D. & Desplats, P. Distinctive patterns of DNA methylation associated with Parkinson disease: identification of concordant epigenetic changes in brain and peripheral blood leukocytes. *Epigenetics* **8**, 1030-1038 (2013).
60. Rehman, H.U. Role of CRH in the pathogenesis of dementia of Alzheimer's type and other dementias. *Curr Opin Investig Drugs* **3**, 1637-1642 (2002).
61. Bennett, A.O.M. Stress and anxiety in schizophrenia and depression: glucocorticoids, corticotropin-releasing hormone and synapse regression. *Aust N Z J Psychiatry* **42**, 995-1002 (2008).
62. Maycox, P.R.*, et al.* Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function. *Mol Psychiatry* **14**, 1083-1094 (2009).
63. Sugo, N.*, et al.* Nucleocytoplasmic translocation of HDAC9 regulates gene expression and dendritic growth in developing cortical neurons. *Eur J Neurosci* **31**, 1521-1532 (2010).
64. Lopez-Atalaya, J.P., Ito, S., Valor, L.M., Benito, E. & Barco, A. Genomic targets, and histone acetylation and gene expression profiling of neural HDAC inhibition. *Nucleic Acids Res* **41**, 8072-8084 (2013).

65. Fan, Y.*, et al.* Focal adhesion dynamics are altered in schizophrenia. *Biol Psychiatry* **74**, 418-426 (2013).
66. Yates, A.*, et al.* Ensembl 2016. *Nucleic Acids Res* **44**, D710-716 (2016).
67. Sherwood, R.I.*, et al.* Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* **32**, 171-178 (2014).
68. Shi, J.*, et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460**, 753-757 (2009).
69. Dobin, A.*, et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
70. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
71. Turner, T.N.*, et al.* denovo-db: a compendium of human de novo variants. *Nucleic Acids Res* **45**, D804-D811 (2017).

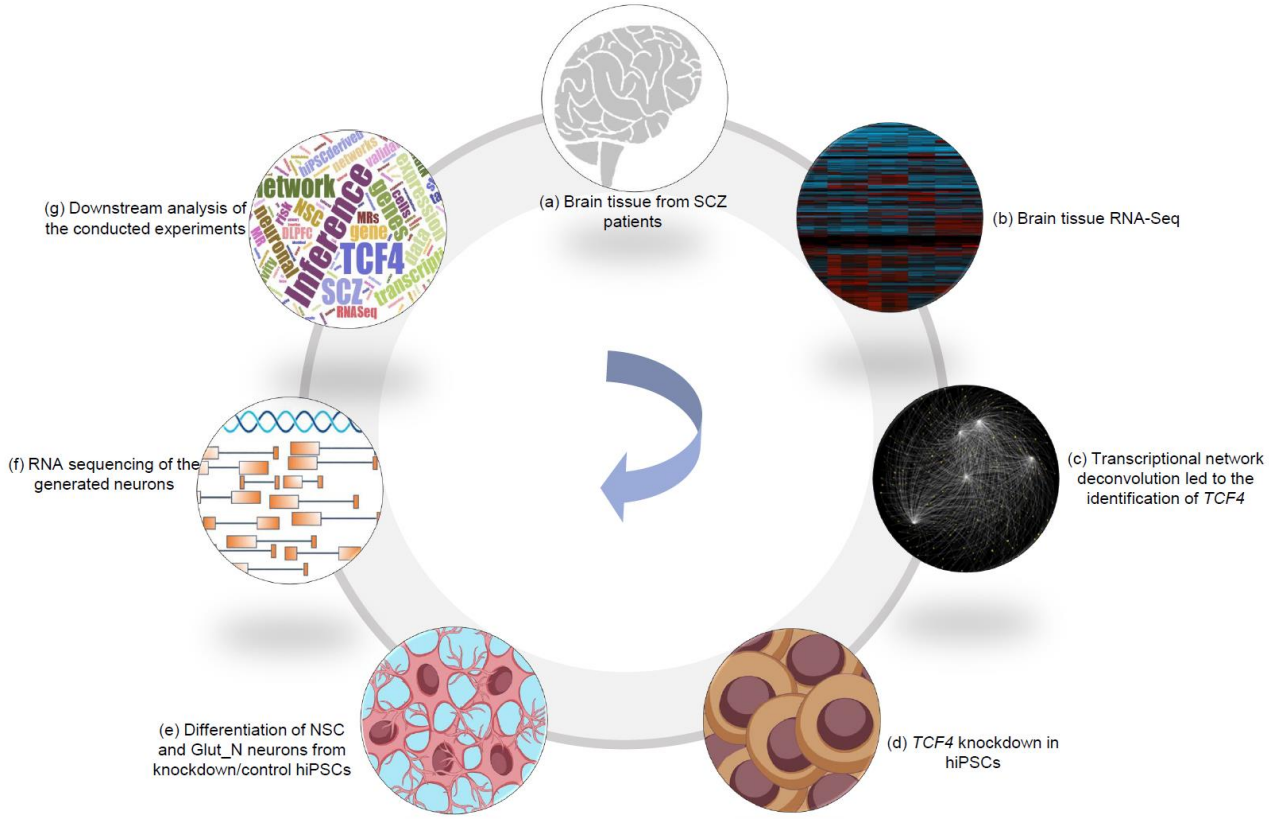

Figure 1. An overview of the current study; (a, b) obtaining gene expression data generated from DLPFC (CMC data) and later CNON data as an independent data, (c) transcriptional network reconstruction based on the provided RNA-Seq data using ARACNe and identification of TCF4 as a disease-relevant master regulator (MR) through virtual protein activity analysis, (d) capturing human induced pluripotent stem cells (hiPSCs) from a SCZ patient, (e) TCF4 expression knockdown in hiPSC-derived neurons, (f) RNA-sequencing of the generated neuronal cells with and without TCF4 knockdown as well processing the generated data, (g) downstream inferences and analyses of the conducted experiments.

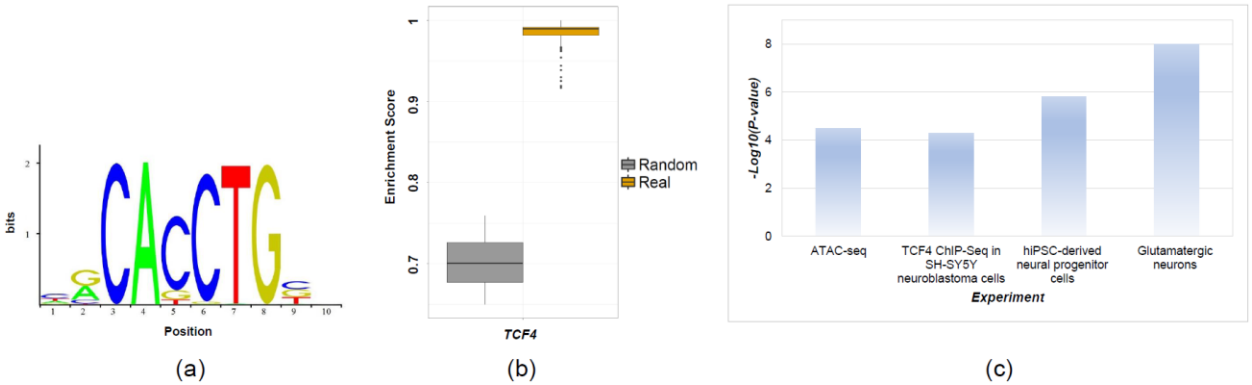

Figure 2. Summary of TCF4 regulons in the CMC and CNON data. (a) TCF4 motif from JASPAR database used in the analysis; (b) TCF4 binding site enrichment scores (based on TCF4 motifs from the JASPAR database) among the TCF4 regulons compared to that of a set of random genes; (c) Enrichment p-values of the TCF4 regulons, compared with several gene sets, including the predicted TCF4 sites from ATAC-Seq, the differentially expressed genes from our TCF4 knockdown experiments, and the predicted TCF4 binding sites in neuroblastoma cells by ChIP-Seq (P-value obtained from Fisher's exact test).

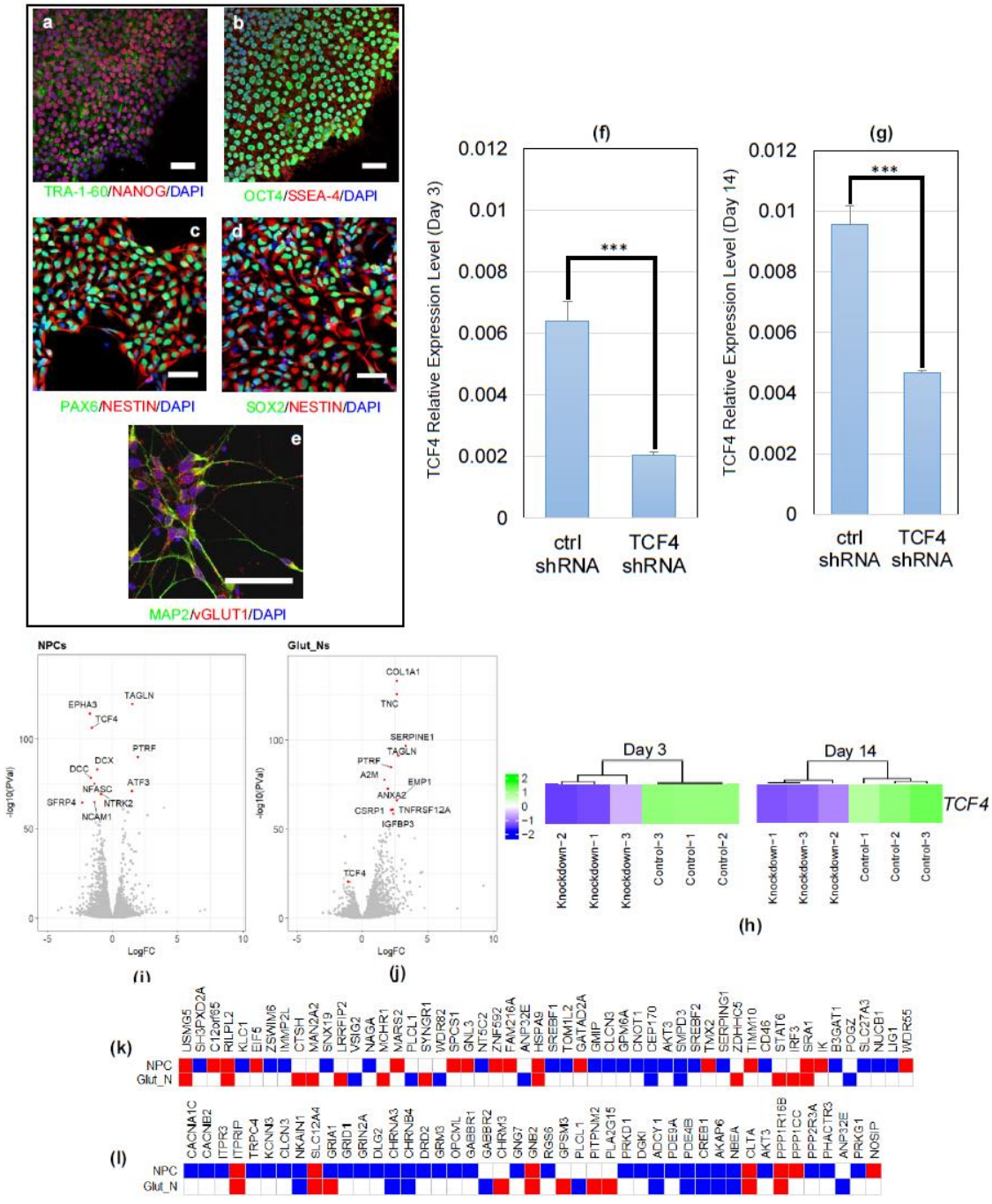

*Figure 3. TCF4 knockdown in hiPSC-derived neuronal precursor cells (NPCs) and glutamatergic neurons. Scale bars: 50 um in all images, cell nucleus are stained with DAPI (blue) (a-e), GAPDH is used as the endogenous control to normalize the TCF4 expression for qPCR, Error bars: mean ± SD (n=4). ***: P < 0.001, Student's t-test, two-tailed, heteroscedastic. (a) Representative immunofluorescence (IF)-staining images of hiPSCs. hiPSCs are stained positive for pluripotency markers TRA-1-60 (green) and NANOG (red); (b) Representative immunofluorescence (IF)-staining images of hiPSCs. hiPSCs are stained positive for OCT4 (green) and SSEA-4 (red); (c) Representative IF-staining images of NPCs (day 3 post plating cells): NPCs are stained positive for PAX6 (green) and NESTIN (red); (d)*

Representative IF-staining images of NPCs (day 3 post plating cells): NPCs are stained positive for NESTIN (red) and SOX2 (green); (e) Representative image of day-14 neurons stained positive for MAP2 (green) and vGLUT1 (red); (f) TCF4 knockdown efficiency measured by qPCR in NPCs (3 days after plating); (g) TCF4 knockdown efficiency measured by qPCR in glutamatergic neurons (day 14 post neuronal induction); (h) TCF4 expression levels at days 3 and 14 in RNA-Seq data; (i) volcano plot of DE genes in NPCs; (j) volcano plot of DE genes in Glut_Ns; (k) overlap of the DE genes upon TCF4 knockdown with a list of GWAS-implicated SCZ risk genes (up/down regulated genes are shown in red/blue, respectively. White cells indicates that the gene is not DE); (l) overlap of the DE genes upon TCF4 knockdown with a list of credible GWAS risk loci (up/down regulated genes are shown in red/blue, respectively. White cells indicates that the gene is not DE).

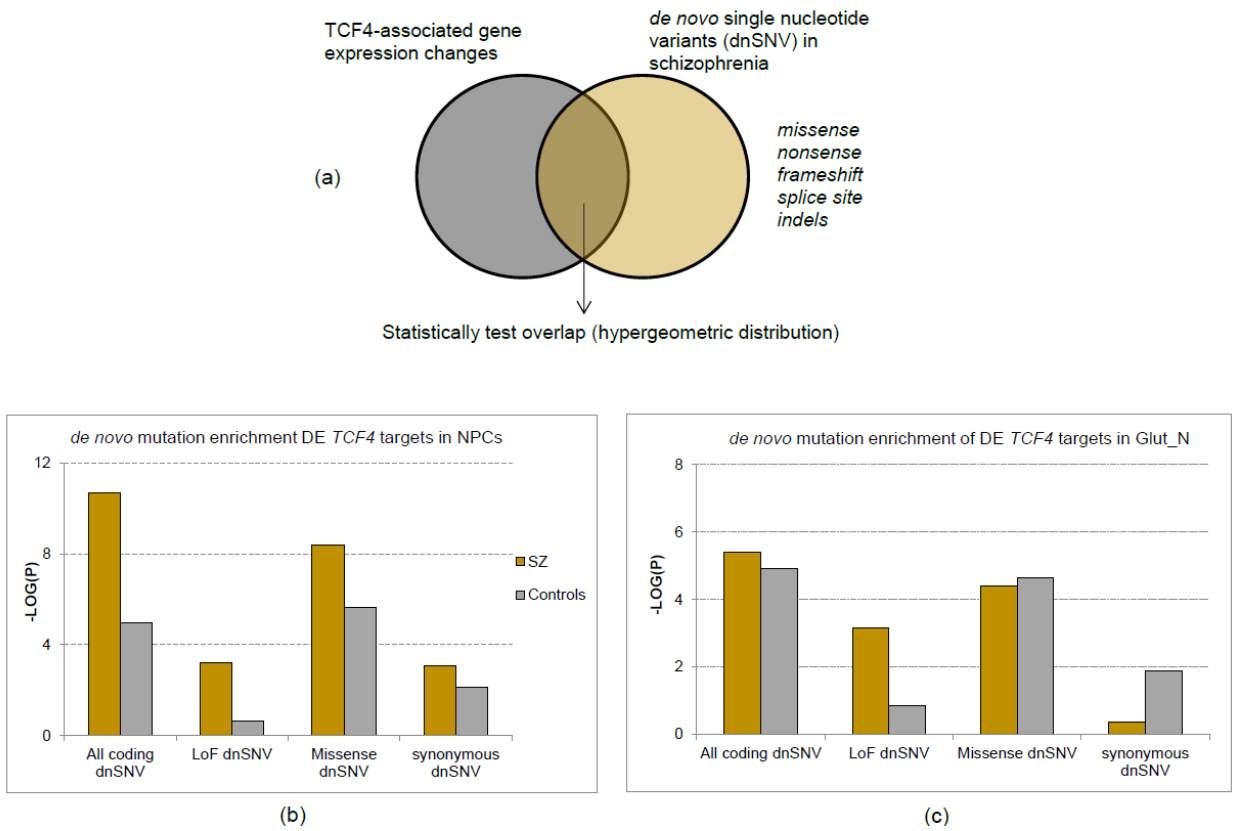

Figure 4. Enrichment of SCZ de novo SNVs in TCF4-associated genes expression changes; (a) analysis framework; (b) All DE genes at day 3 (NPCs), The P-value on the y-axis was obtained from a hypergeometric test was used to test the statistical significance of each overlap (number of shared genes between lists) and using all protein coding genes as a background set; (c) All DE genes at day 14 (Glut_N).