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

Abolfazl Doostparast Torshizi[1], Chris Armoskus[2], Siwei Zhang[3], Hanwen Zhang[3], Oleg V Evgrafov[2], James A Knowles[2], Jubao Duan[*3], Kai Wang[*1]

[1]Institute for Genomic Medicine, Columbia University Medical Center, New York, NY 10032, USA.

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

[3]Center for Psychiatric Genetics, North Shore University Health System and The University of Chicago, Evanston, IL 60201, USA.

* to whom correspondence should be addressed

Abstract

Tissue-specific reverse engineering of transcriptional networks has led to discoveries uncovering underlying regulators of cellular networks in several cancers and other diseases. However, whether these approaches can be applied to complex psychiatric disorders is largely unexplored, partly due to the general lack of data on appropriate tissues for mental disorders. In this study, using RNA-Seq data on dorsolateral prefrontal cortex from 307 schizophrenia (SCZ) cases and 245 controls, we deconvoluted the transcriptional network to identify transcriptional “master regulators” mediating expression of a large body of target genes. Together with an independent RNA-Seq data on cultured primary neuronal cells derived from olfactory neuroepithelium from a cohort of 143 SCZ cases and 112 controls, we identified TCF4 as a candidate master regulator. TCF4 was previously identified as a SCZ susceptibility gene from genome-wide association studies, but its regulatory subnetworks had been elusive. We confirmed many predicted regulatory targets by TCF4 through additional analysis of predicted transcription factor binding sites and ATAC-Seq data (p=3.251e-05). We further knocked down TCF4 expression in human induced pluripotent stem cells (hiPSC) at 3 and 14 days post neuronal differentiation, confirmed that TCF4 affects axon guidance pathway, and that the collection of predicted target genes were dysregulated on day 3 (p=1.55e-06) and day 14 (p=1.05e-08). In conclusion, multiple lines of evidence suggested that TCF4 serves as a master regulator in SCZ, demonstrating a clear example where network dysregulation involving hundreds of genes can lead to susceptibility to complex psychiatric diseases.

Introduction

As a debilitating neurocognitive disease, schizophrenia (SCZ) affects almost 0.7% [1-4] of adults. Mostly diagnosed at the onset of adulthood, SCZ causes severe neurocognitive and neurophysiological dysfunctions. SCZ is highly heritable disorder [5] in which thousands of common alleles with small effects contribute to the disease risk [6]. Over the past a few decades, a considerable effort has been put in identifying genetic risk factors as the underlying drivers of this crippling disease, but only a fraction of these studies have conclusively identified some genetic risk factors such as rare copy-number variants (CNVs) [7-10], rare protein-altering mutations [11], and common variants [12]. Despite a vast amount of research studies, the genetic basis of SCZ is still poorly understood, prohibiting the development of targeted therapeutic strategies.

Thus far, post-mortem gene-expression profiling has been the back-bone of a large number of research studies conducted on the molecular basis of SCZ. This technique bears several limitations, making it challenging when interpreting the experimental outcomes [13-15]. There are also logistical issues collecting post-mortem brain tissues due to time consuming process of building brain banks where only a few portion of the tissues may be related to SCZ-affected people [16] though it has been highly addressed with the recent large-scale SCZ transcriptional studies. On the other hand, SCZ is a complex disorder causing genetic disturbances across a wide range of genes given that SCZ patients bear diverse genetic backgrounds and personality traits [17]. Given the fact that mining gene expression patterns from brain-tissue instead of peripheral blood is a primary means in studying transcription regulatory mechanisms, previous studies indicate that gene expression changes in SCZ affects multiple regions in the brain including the prefrontal [18, 19] and temporal cortices [20, 21] along with hippocampus [22, 23]. In this regard, several points should be made including: (1) SCZ is associated with genomic disturbances across several cortical regions affecting a large spectrum of genes [22]; (2) multiple cell types host transcriptome changes comprised of subclasses of principal neurons [24, 25], interneurons [18, 26], and olygodendrocytes [18] making any inference quite challenging due to the pathophysiological processes that each cell types goes through; (3) expression of transcripts related to the genes involved in various intracellular processes are directly affected in SCZ where interrelations between these transcripts are not well-established yet. Some of these transcripts include: synaptic transmission [19, 27], energy metabolism [28, 29], immune response [30, 31], and inhibitory neurotransmission [5]; (4) there might be links between gene expression changes and susceptible genetic loci playing important roles in developing SCZ while despite emerging empirical evidences, these links are not well-investigated yet [15, 16, 32]. Despite considerable number of case-control gene-expression studies, a great deal of the reported findings shows a small amount of overlap, though many of them share the same data sets [4]. Along with the instinctive limitations of microarray-based studies [33], it appears that these

methods are underpowered to capture the subtle regulatory patterns in cell-specific studies where multiple brain regions are directly involved.

Due to the complexities of the underlying signatures governing mechanistic processes in SCZ, it makes perfect sense to take advantage of molecular networks to uncover such complexities. Molecular networks (*interactome*) as a harmonized orchestration of genomic interactions, play a central role in mediating cellular processes through regulating expression of the genes or formation of transcriptional complexes. According to [34], cellular networks possess the scale-free property mostly observed in protein-protein interaction and metabolic networks [35, 36]. One of the most commonly used network-based representation approaches of cellular processes is co-expression networks bearing the scale-free property [37, 38]. Nevertheless, co-expression networks and other similar means are not comprehensive enough to fully recapitulate the entire underlying molecular structures driving the disease phenotype. In principle, currently existing expression network analysis approaches fall into four categories including [34]: optimization methods [39, 40], which usually optimize the network based upon a certain criterion; regression techniques [41, 42], which fit the data to a priori models; multi-omics integrative approaches [43], and statistical methods [44]. Despite vast applicability of these sets of approaches, there are critical issues making them less efficient in dealing with eukaryotic organisms bearing complex cellular structures. These shortcomings can be listed as follows: connecting genes having direct interactions leaving their mutual causal effects aside, overfitting when dealing with small number of samples, suffering from curse of dimensionality, and not being able to reverse-engineer the mammalian genome-wide cellular networks [34]. On the other hand, many of the currently existing network-based approaches are not context-specific leading to high false positive rates. With the emergence of information-theoretic deconvolution techniques initiated by Basso et al., [34] and its successful applications in a wide range of complex diseases such as cancer [45, 46], prostate differentiation [47], and neurodegenerative diseases [48], the way to infer causal relationships between transcription factors and their downstream regulon was paved.

As the largest genomic data bank in brain samples obtained from autopsies of individuals with and without psychiatric disorders, Common Mind Consortium (CMC) [4] has provided researchers with a rich database orders of magnitude larger than the current similar databases. In this study, using RNA-seq data obtained from prefrontal cortex, we will be conducting the reverse engineering of the regulatory processes mediating SCZ to elucidate critical Master Regulators (MRs) and to infer their role in orchestrating cellular transcriptional processes. The prefrontal cortex was chosen for two reasons: first, it controls high-level cognitive functions many of which are disturbed in SCZ; second, years of study point out this important region due to its abnormalities in cellular and neurochemical functions in SCZ sufferers. Using the reverse-engineered extracted networks, we will further analyze the activity of the detected MRs in developing Schizophrenia and will validate the computational results

both by reviewing the published transcription factors (TFs) and published in-vitro experiments. Our findings of these experiments were then also validated using another independent RNA-Seq data from a previous study using Cultured primary Neuronal cells derived from Olfactory Neuroepithelium (CNON) of individuals with and without SCZ. The rationale behind using CNON lies in the fact that since most neuronal cells in adult brains are not dividing or differentiating, then adult brains are not adequate to characterize gene expression alterations influencing neurodevelopment; however, CNON cells are able to divide and migrate both in 2D and 3D cultures [49]. After assembling the SCZ interactome using the separate gene expression datasets followed by virtual protein activity analysis, we will focus on highly enriched MRs shared by the two interactomes and investigate their regulatory role in development of SCZ. A general overview of the both computational and experimental stages of this research has been provided in Figure 1.

Results

Analysis on the CMC data sets

The CMC data contains 15971 transcripts extracted from 552 brain samples, including 307 cases and 245 controls. Our main goal was to reverse-engineer the transcriptional regulatory networks based on generated RNA-seq data to infer the MRs and their blanket of regulated targets (regulon) along with constructing their corresponding sub-networks. To do so, we employed ARACNe (Algorithm for Reconstruction of Accurate Cellular Networks) as a powerful and versatile tool to reconstructing cellular networks [34, 50] (see methods and materials). In this method, first gene-gene co-regulatory patterns are identified by an information theoretic method called Mutual Information (MI). In the next step, the constructed networks are pruned by removing indirect relationships in which two genes are co-regulated through one or more intermediate entities. This allows us to observe relationships bearing significantly high probabilities of representing direct interactions or mediated interactions through post-transcriptional agents not being detected from gene expression profiles.

The P-value threshold of 1e-08 using DPI=0.1 (as recommended in [50]) lead to a repertoire of 102473 interactions ranging from 1 to 106 interactions for each individual transcript. Our goal is to focus on hub Transcription Factor (TF) genes and transcriptional-affecting genes that will be called MRs. We curated a list of such genes from three sources including FANTOM5 consortium [51], a curated set by Vaquerizas [52], and TRRUST [53]. A total of 2198 TFs and expression regulatory genes were curated from these sources. Among the entire hub genes of the constructed network, 1466 TFs were found as hub genes. We will further focus on the TFs and expression-altering genes and their respective sub-network topology. TFs subnetwork is provided in Supplementary Table 1. 1466 TFs were found in the network which meet the

ARACNe set up criteria. This subnetwork contains 24548 interactions accounting for almost 24% of the entire interactions in the constructed network. We annotated these transcripts to their genes where 57 TFs were Differentially Expressed (DE) between cases and controls at p=0.01 (FDR<0.05). The full network on CMC data is provided in Supplementary Table 2.

Making use of ARACNe, in addition to identifying new MRs, we also re-identified many of the TFs previously published in the literature that may be potentially involved in the pathogenesis of SCZ. These TFs include: JUND [54]; KCNIP3 [55]; LHX6 [56]; NRG1 and GSK3B [57]; NFE2 and MZF1 [58]; NPAS3 [59]; DISC1 [60, 61]. We should mention that the obtained MRs are the direct result of unbiased data-driven analysis of transcription data and no latent biological knowledge were incorporated in their acquisition.

Virtual Protein Activity Analysis Using RNA-seq Data

We managed further network analysis to deeply analyze the activity of the identified MRs by taking into account the expression pattern of their downstream regulon through a dedicated probabilistic algorithm. This method exploits the regulator mode of action, the regulator-target gene interaction confidence and the pleiotropic features of the target regulation. We used VIPER (Virtual Inference of Protein-activity by Enriched Regulon analysis) [46].

We fed the output of the previous stage (ARACNe) network into VIPER to check whether any of the identified MRs have significant regulatory role in expression level of their downstream target genes. We reconstructed the network using 1466 MRs consisting 24548 interactions. To account for the matter of co-regulation in the transcriptional, we provided a null model by using the z-score of the gene expression signatures obtained from two-sided t-test by permuting the samples at random (see Methods and Materials). The computational results of the ten top MRs are depicted in Figure 2a. According to Figure 2a, the first column on the left represents the expression of the regulon of the MR where blue (red) bars indicate repressed (activated) targets. In the second column, gene symbols are represented. The Act and Exp columns represent the inferred activity and expression degrees of the identified MRs.

The activity of the MRs is inferred based on the enrichment of their closet regulated targets. In order to find out which genes are the enriched targets in the genetic signatures i.e., the z-score previously computed, we employed the leading-edge analysis [62] proposed by Subramanian *et al*. to identify the genes driving the enrichment of a gene set on the signatures based on Gene Set Enrichment Analysis (GSEA). Among these top MRs, all of them were captured by ARACNe as hub genes during the network deconvolution process.

According to [63, 64], significant activation of MRs based on its regulon analysis can cause confounding effects since many of their regulated targets might have been regulated by a bona fide activated TF. This phenomenon is called shadow effect. This is even more serious in

transcriptional regulations because they are highly pleiotropic. To address this, we penalized the contribution of the pleiotropically regulated targets to the enrichment score. Since we had previously addressed the pleiotropic effects in network generation stage, we expect to observe a small number of pleiotropic connections. No shadow connections were observed. Further, in order to predict synergistic interactions between the regulators, we computed the enrichment of co-regulons. This was defined as the intersections between the targets. Our expectation was that a gene expression signature is synergistically regulated by a combination of regulators when their corresponding co-regulons show a significantly higher enrichment on the signature than the union of the corresponding regulons [65]. We computed the enrichment of the co-regulons for the top 10 co-regulators. Out of the top 10 co-regulating sets, 8 co-regulating sets are observed that significantly co-regulate their downstream targets and were identified also by VIPER. Two other transcripts can be observed here including ENSG00000114631 annotated to PODXL2 and ENSG00000163655 annotated to GMPS.

Methylation patterns of TMEM9 [66], has been reported to be associated with Parkinson's disease though it has not been referred to as directly regulating SCZ. KLHL36 has been reported to be likely of associating with SCZ but has not been directly referred to. CRH is a protein coding gene associated with Alzheimer's disease, depression, and SCZ [67, 68]. ARPP19 has been identified to be DE between two large SCZ cohorts associated with nerve terminal function [69]. HDAC9 plays an important role in transcriptional regulation, cell cycle progression and developmental events. HDAC9 is a histone deacetylase inhibitor that has been shown to potentiate hipocampal-dependent memory and synaptic plasticity leading to different neuropsychiatric conditions [70]. TCF4 is another important finding that may play an important role in nervous system development. It is associated with Pitt Hopkins Syndrome and mental retardation [71]. This gene has also been reported to host a SNP in its intron that might be associated with SCZ [72-74]. TCF4 has also been reported to host genetic variants contributing to SCZ development [12]. In the following sections, we will conduct similar network analysis and will discuss common findings between these two numerical experiments.

Results on CNON Data

As an independent data, we made use of RNA-Seq data from Cultured primary Neuronal cells derived from Olfactory Neuroepithelium (CNON) of individuals with and without SCZ generated in our recent study. The CNON data contain 23920 transcripts extracted from 255 individual samples 143 of which cases and 112 controls. Adopting the same set up parameters, we ran ARACNe [34, 50] on the new data set and reverse engineered the transcriptional network. The P-value threshold of 1e-8 using DPI=0.1 (as recommended in [50]) lead to a repertoire of 173524 interactions ranging from 1 to 330 interactions for each individual transcript. The full constructed network is provided in Supplementary Table 3. Using our curated TF list, we observed 1836 TF nodes in the constructed network. The subnetwork

of these TFs is provided in Supplementary Table 4. This subnetwork accounts for 34757 interactions and covers almost 20% of the entire significant existing interactions in the network. Conducting the activity analysis of the identified hub genes, VIPER was run on the constructed network. The top 10 MRs were obtained. Representations of the MRs is provided in Supplementary Fig. 1.

Common Findings

Using the top 15 MRs identified in the CMC data (Figure 2a), we checked the reverse-engineered network of CNON data to see if any of these MRs also act as hub genes in the CNON network. Five MRs were found in the CNON network including: TCF4, NR1H2, HDAC9, ZNF436, and ZNF10. The entire target genes being regulated by these MRs are being listed in Supplementary Table 5. Overall, these five MRs regulate 102, 68, 36, 66, and 43 genes, respectively. For each MR, we conducted Pathway Enrichment Analysis (PEA) using WebGestalt [75] (FDR-corrected p-values are reported). PEA of the targets of TCF4 revealed the following pathways (multiple test adjusted) including: Notch signaling pathway (p=0.0001), ErbB signaling pathway (p=0.0009), Purine metabolism (p=0.0004), long-term depression (p=0.0008), circadian rhythm (p=0.001). Targets of NR1H2 were enriched in the following pathways: bladder cancer (p=0.0015) and pathways in cancer (p=0.0117). Chronic myeloid leukemia (p=0.00045), TGF-beta signaling pathway (p=0.0060), p53 signaling pathway (p=0.0040), pathways in cancer (p=0.0033) were observed for the targets of ZNF436. Oxidative phosphorylation (p=0.0060) and Metabolic pathways (p=0.0077) refer to the ZNF10 targets and finally, insulin signaling pathway (p=0.0056) and focal adhesion (p=0.0115) were observed for the targets of HDAC9.

Among the identified MRs, TCF4 and NR1H2 are TFs, ZNF10 and ZNF436 are zinc finger proteins that may be involved in transcriptional regulation, and HDAC9 is a histone modification gene which is involved in peters anomaly [76] and congenital malformations [77]. In order to check whether the identified MRs are enriched in the promoter region of their targets, we conducted TF binding enrichment analysis (TFBEA) using JASPAR [78] (Figure 3) on TCF4 and NR1H2. First, sequence motif of these genes were extracted. Then, we extracted the sequence of the target genes 2000 base pairs upstream and 1000 base pairs downstream of their Transcription Start Sites (TSSs). The significance threshold of 0.80 was chosen on the relative enrichment scores of the target genes (Figure 3). To make sure that the TFBEA results are projecting the correct enrichment scores, for both TCF4 and NR1H2, we re-iterated the TFBEA process using a random list of unrelated genes which were not in their subnetworks and computed the relative enrichment scores. Enrichment scores for the real target genes of these MRs and the random target genes is provided in Figure 3. For both TCF4 and NR1H2 the enrichment scores are significantly higher than random enrichment scores implicating high confidence in finding correct binding positions in the target genes. Our

analysis below focused on TCF4 given its known involvement in SCZ from large-scale genome-wide association studies. To illustrate how TCF4 interactome in the both prefrontal cortex and CNON gene expression data contribute to the phenotype, GSEA analysis was conducted revealing the expression pattern of TCF4 inferred target genes in both expression data (Figure 2b-c).

Experimental Validation

In order to gain a deeper insight into the real impacts of the identified MRs on their regulons, we have conducted a series of experiments on the regulatory impacts of MRs if they bind to the putative targets. We first examined whether the predicted MRs bind to their regulon genes by using our ATAC-seq-based empirical TF-binding footprints (inferred by PIQ tool) data sets in iPSC-derived glutamatergic neurons [79]. Out of the top 5 MRs, only TCF4 and NR1H2 are among the TFs analyzed by PIQ. Using the three sets of ATAC-seq footprint data derived from day 30, day 41 neurons, and iPS cells, we extracted the corresponding BED files. All footprints have been annotated using the TF matrix with the names of different TF factors annotated in the BED files. For each sample, footprints are 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 (and 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 hg19 reference genome. 29 target genes out of 102 regulons of TCF4 (total TCF4 targets in ATAC-seq data=3276) were validated to be bound by TCF4, FET=3.251e-05 (Fisher's Exact Test).

TCF4 Knockdown

To examine regulatory role in driving phenotypic transition in SCZ patients, we chose TCF4 as it has been picked up in the computational stage of this research. TCF4 is a susceptible SCZ risk gene whose role in SCZ development has been investigated mostly through genome wide association studies (GWAS) or similar approaches with the hope of uncovering its corresponding susceptible genetic loci [12, 80]. From another molecular perspective, TCF4 knockdown has been investigated in SH-SY5Y neuroblastoma cells that may not be relevant with respect to the tissue of interest [81]. Based on this rationale, we empirically validated the predicted TCF4 interactome using RNA-seq data produced from human induced pluripotent stem cell (hiPSC) model via knockdown expression of TCF4 (see methods and materials). The hiPSC line used for deriving neurons was generated from cryopreserved lymphocytes (CPLs) of a 29 year-old male SCZ patient (cell ID: 07C65853) from the MGS collection ([82]). Neural cells on day 3 and 14 were used for RNA-seq analysis. At each time point, three ITF-2 (TCF4) knockdown samples and three controls were generated. The TCF4 knockdown in hiPSC-derived neuronal precursor cells (NCS) has been depicted in Figure 5. In this figure,

hiPSCs were stained positive for pluripotency markers NANOG, TRA-1-60, OCT4, and SSEA-4. Further description can be found in the figure caption. We measured the TCF4 knockdown efficiency by qPCR in NPCs three days after plating and in glutamatergic neurons fourteen days post neuronal induction, respectively (Figure 5 f-g).

On day 3, 4891 (3363) genes were differentially expressed (DE) including 2330 upregulated and 2561 downregulated genes at FDR=0.05 (FDR=0.01) (see methods and materials). By contrast, on day 14, 3152 at FDR<0.05 genes were DE including 1862 (1290) upregulated (downregulated) genes at FDR=0.05 At FDR<0.01, 1897 genes were DE. The full list of the DE genes along with complementary statistical information can be found in Supplementary Table 7. We checked the expression patterns of the identified TCF4 regulon in the produced RNA-Seq data at both time points based on their respective log fold change between the knockdown and control samples (Figure 4a). expression patterns of the TCF4 regulon follow almost the same pattern with some differences in the log fold change at different time points. However, expression of some targets looks remarkably different between day 3 and 14. For example, NT5E and ATP10A were slightly downregulated at day 3 but they are highly upregulated at day 14. In order to check whether the expression of the identified targets are associated with the phenotype (here knockdown and control samples) we conducted GSEA using the RNA-Seq on the both time points over the TCF4 regulon. In both cases, expression patterns represent significant association with the phenotypes (Figure 4c-d). Using the identified DE genes at each time point, we further probed the underlying perturbed biological pathways at each separate time point (Supplementary Table 6 and Supplementary Files). We will further discuss them in the next section.

Discussion

In this study, with the aim of uncovering the gene expression mediating drivers that may contribute to the onset of SCZ based on a data-driven perspective, we managed to assemble tissue-specific transcriptional regulatory networks using ARACNe on two independent RNA-Seq data from dorsolateral prefrontal cortex and cultured primary neuronal cells derived from olfactory neuroepithelium. Focusing on expression regulatory hub genes, we computationally inferred the activity of such proteins leading to identification of TCF4, NR1H2, HDAC9, ZNF10, and ZNF436. These MRs were independently identified on the deconvoluted networks generated from the two separate data. Further, using existing signatures, we hypothesized the role of TCF4 to exert regulatory effects over a large body of genes across the deconvoluted network as it has already been picked up as a SCZ susceptible rick factor. Significant lines of evidence bolstered the modulating role of TCF4 to drive a long list of genes whose dysregulation contribute to SCZ phenotypic variations. Probing this hypothesis, we knocked down TCF4 to experimentally survey the computational

findings and to analyze the expression patterns of its regulon upon inactivation of TCF4.

TCF4 knockdown in neuronal stem cells led to recognition a large set of DE genes resulting in activation of a set of biological pathways. Although several pathways between the two time points are shared, but several distinctive pathways can be observed, too. Among the top 25 pathways (FDR<0.01), several pathways were shared including Apotosis (P-day3=9.83e-03, P-day14=5.94e-5), Focal adhesion (P-day3=2.01e-06, P-day14=5.18e-12), p53 signaling pathway (P-day3=7.01e-07, P-day14=4.16e-07), Axon guidance (P-day3=1.52e-06, P-day14= 5.94e-05), MAPK signaling pathway (P-day3=1.38e-03, P-day14=3.37e-05), Regulation of actin cytoskeleton (P-day3=1.52e-03, P-day14=2.72e-09), Pathways in cancer (P-day3=1.98e-04, P-day14=5.44e-06), and Colorectal cancer (P-day3=5.23e-03, P-day14=3.02e-04). As the cell death mechanism, apoptosis navigates neurodevelopment and its role in diverse neuropathological conditions is increasingly recognized [83]. At both time points, most of the components of apoptosis pathway were upregulated. Despite activation of focal adhesion pathway in both time steps, regulation patterns look different. In day 3, 32 focal adhesion genes are upregulated and 54 were downregulated while the same numbers on day 14 are 69 and 12, respectively. Focal adhesion has been remarked in to be altered in SCZ [84].

We observe that depleted TCF4 effectively alters axon guidance (FDR-day3=1.52e-06, FDR-day14=5.94e-05). It can be stated that TCF4 can partially exert its influence on the phenotype through defecting neuronal activity. At both time points, autism associated genes expressed in anterior cingulate cortex [85] were observed including (P-values are FDR corrected): ROBO1 (P-day3=6.09e-24, P-day14=0.003), ROBO2 (P-day3=5.21e-22, P-day14=1.96e-06), ROBO3 (P-day3=3.78e-09, P-day14=3.44e-07), EFNB2 (P-day3=1.71e-07, P-day14=1.32e-12), ENFB3 (P-day3=4.36e-09), PLXNA4 (P-day3=0.00033). Of the pathways altered exclusively at day 3 can be named long-term potentiation (P=5.04e-05), neurotrophin signaling pathway (P=2.61e-05), mTOR signaling pathway (1.98e-04), endocrine resistance (P=1.52e-04), endocytosis (P=2.28e-03). On the other hand, some other pathways were exclusively altered at day 14 such as: TNF signaling pathway (P=6.49e-06), FoxO signaling pathway (P=1.92e-04), AGE-RAGE signaling pathway in diabetic complications (P=1.93e-06), PI3K-Akt signaling pathway (P=3.74e-04) (see Supplementary Figure 6).

TCF4 is known to host a genetic aberration significantly associated with SCZ. However, downstream regulon of TCF4 highly associated with development of SCZ or other neurodevelopmental or neuropsychiatric impairments is still illusive. Using the credible SCZ associated genes in [12], we sought to check whether any of these genes were picked up when depleting TCF4. 52 genes were DE across the two time points (Figure 4b). In this figure, the first two columns on the right indicate whether their corresponding

genes are DE at day 3 or 14 in the way that black cells mean ‘yes’ and white cells indicate ‘no’. For the DE genes at each time point, the FDR level and Log Fold Change (LFC) are provided in the figure. An important point is that although all of these genes have been proven to host SCZ susceptible loci, we demonstrated that their expression is highly dependent on TCF4 proving that they are in fact a major part of TCF4 interactome. Among these genes, some of them were DE considerably higher than the others including: GPM6A (P=1.26e-20), PLCL1 (P=8.31e-15), SREBF2 (P=8.9e-13), GNL3 (P=2.58e-10). From standpoint of classical neurobiological knowledge of neuronal synaptic transmission as well as regulation of electrical excitability, a recent study [86] has curated 92 SCZ associated single nucleotide polymorphisms (SNPs) for the presence of genes related to neuronal electrical excitability and neurotransmission. We observed a large proportion, 44 out of 92 (FET=2.45e-10), of these genes to be DE after TCF4 knockdown that can be considered as minor network constituents of TCF4 interactome. In addition, the most recent GWAS on SCZ [5] reports 30 novel genetic loci accompanied by their associated genes. We have picked up several members of this gene list to be highly dysregulated after TCF4 knockdown (Figure 4e-f).

Comparing the list of DE genes after TCF4 knockdown and the list of TCF4 regulon identified by ARACNe, 39 genes were enriched in the DE genes (FET=1.55e-06) on day 3. On day 14, when the number of DE genes are smaller, 33 genes (FET=1.05e-08) were experimentally shown to be dysregulated after TCF4 knockdown. Compared to TCF4 knockdown in neuroblastoma cells [81], the number of DE genes are considerably larger in our experiments. We can observe 316 shared genes on day 3 (FET=2.39e-17) and 206 shared genes on day 14 (FET=1.6e-11). This difference confirms the importance of tissue specific investigations that leads to expression of genes that may not be expressed in unrelated tissues. Studying the contribution of the identified interactome to expression of TCF4, we collected the GSEA results on three expression data including the CMC and CNON data along with the TCF4 knockdown RNA-Seq data Figure 6. According to Figure 6a, most of the target genes represent a rather higher enrichment score in TCF4 expression in the CMC and CNON data compared to the same enrichment score of the same gene in day 3 and day 14. Yet, FHL2, NT5E, and SLC7A5 show almost the highest GSEA rank metric score in the entire datasets being depicted in Figure 6b. a recent study indicates several SNPs in SLC7A5 to be highly associated with SCZ [87].

Conclusions

In this study, using high-quality RNA-seq data from prefrontal cortex of a large cohort of SCZ patients, we assembled the transcriptional networks followed by virtual protein

activity analysis. Applying the same procedure on an independent RNA-seq data, we picked up five highly effective MRs governing expression of a large set of genes whose expression patterns contributed significantly to the onset of SCZ including: TCF4, NR1H2, HDAC9, ZNF10, and ZNF436. TCF4 has recently been under investigations due its role in carrying SCZ-associated genetic mutations according to GWAS. To further probe the underlying interactome of TCF4 to uncover how its expression can regulate transcriptional networks in neuronal cells, we empirically validated its role using RNA-seq data produced from human induced pluripotent stem cell (hiPSC) model via knockdown expression of TCF4. Our findings indicate how TCF4 can alter many of neuronal and signaling pathways and give us deeper insights into the etiology of TCF4 transcription.

Methods and Materials

RNA-seq data of Dorsolateral Prefrontal Cortex (DLPFC)

RNA-seq data of DLPFC (Dorsolateral Prefrontal Cortex) from individuals 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/). Brain tissues in this portal are 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).

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 bot 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. MI was thresholded 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 1e-08 using DPI=0.1 (as recommended in [50]) were used in running ARACNe.

Virtual protein activity analysis

The regulon enrichment on gene expression signatures was tested by 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 order to generate a quantitative measurement of difference between the groups any method can be used including: fold change, Student's t-test, Mann-Whitney U test, etc. As an alternative, single- sample-based gene expression signatures can be obtained by comparing the expression levels of each feature in each sample against a set of reference samples by any suitable method, including for example Students t-test, z-score transformation or fold change; or relative to the average expression level across all samples when clear reference samples are not available. 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. Further information about VIPER can be accessed in [46].

Transcription factor binding site enrichment analysis

Human reference genome (version GRCh37.p13) 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 [88] and scanned 3000 upstream and 1000 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. JASPAR database contains Position Weight Matrices (PWM) for each TF. These matrices have four rows representing A, C, G, and T bases and each column represents the relative weight of that pair in that position obtained from the frequency of that base in that position. 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.

RNA sequencing

RNA-seq libraries were prepared from total RNA from collected neuronal cells of 12 samples. 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 (q-PCR). The quantified libraries were then fed into HiSeq/MiSeq sequencers after pooling according to its effective concentration and expected data volume.

RNA-seq data analysis

The sequenced reads for each sample were obtained as fastq files. Initial quality check of the raw data was performed using FastQC. All of fastq files passed the quality control stage including removing low quality bases, adapters, short sequences, and checking for rRNA contamination. High quality reads were mapped to the reference genome GRCh38 using STAR [89]. Sorting and counting was also performed using STAR based on the procedure recommended by the developers. To detect DE genes, the output BAM file was fed to edgeR software package [90]. Normalization, and differential expression analysis was conducted using the default recommended settings.

hiPSC and 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 SZ patient (cell ID: 07C65853) from the MGS collection ([82]). The NorthShore University HealthSystem Institutional Review Board (IRB) approved the study. hiPSCs were cultured using the feeder-free method on Geltrex (Thermofisher)-coated plate in mTeSR1 medium (StemCell). Media were changed daily, and cells were passaged as clumps every 5-day using ReLeSR (StemCell) with the presence of 5 µp ROCK inhibitor (R&D Systems). hiPSCs were characterized by positive IF staining of pluripotent stem cell markers OCT4, TRA-1-60, NANOG, SSEA4.

Human neural stem cells (NSCs) and neuron differentiation

NSCs were derived using PSC Neural Induction Medium (Thermofisher) with modifications. In general, iPSCs were re-plated as clumps into Geltrex (Thermofisher)-coated 6-well plates (Thermofisher) in mTeSR1 (StemCell) on Day 0. From Day 2, the media was switched to PSC Neural Induction media and changed daily, with frequent removal of non-neuronal cells. On Day 11, P0 NSCs were harvested using Neural Rosette Selection Reagent (StemCell). NSCs were maintained in Neural Expansion Medium (Thermofisher) and passaged every 4-6 day with the presence of 5 μM ROCK inhibitor (R&D Systems) until P4. NSCs could be cryopreserved in Neural Expansion Medium with 10% DMSO after P1. NSCs were characterized by positive IF staining for Nestin and Pax6.

Neuron differentiation and ITF-2 (TCF4) knockdown

hiPSC-derived NSCs (passage 3) were plated at 3 x 105 per well in 12-well plates pre-coated with Geltrex for 2 hours. The neural differentiation media consists of neural basal media, B-27 supplement, L-glutamine, 10 ng/ml BDNF, 10 ng/ml GDNF, and 10 ng/ml NT-3 ([79]) was used and the NSCs were allowed to grow and differentiate for 3 or 14 days, respectively. At the corresponding day, 20 μl (1.0 x 105 infection units of virus, IFU) of either ITF-2 or shRNA lentiviral particles (Santa Cruz sc-61657-V for ITF-2, contains a pool of 3 target-specific, propriety constructs targeting human ITF-2 gene locus at 18q21.2; and sc-108080 for control, which encodes a scrambled non-specific shRNA sequence) were added to each well (3 replicates) and the cells were further cultured for 48 hours before collection. The cells were washed with PBS in wells, and 600 μl Buffer RLT Plus were directly added to each well to collect homogenised cell suspension. Total RNAs were extracted from the collected cell suspension using RNEasy Plus mini kit (Qiagen) and the quality of extracted total RNA were examined using Nanodrop spectrum analysis and agarose gel electrophoresis. The obtained total RNAs were used for RNA-seq after verifying the knockdown efficiency by qPCR. The day-14 excitatory neurons were characterized by positive IF staining for VGLUT1 and PSD95.

qPCR to validate the knockdown efficiency of ITF-2

cDNA was reverse-transcribed from total RNAs using the High-Capacity Reverse Transcription Kit with RNase Inhibitor (ThermoFisher). Briefly, 50 ng of total RNA from Day 3 Neuron and 100 ng total RNA from day 14 Neuron were used for each reverse-transcription reaction (20 μl) according to the manufacturer's instructions. cDNA was diluted 15-fold (280 μl of nuclease-free water mixed with 20 μl of RT product) before applying for qPCR. Subsequent qPCR analysis was performed on a Roche LightCycler 480 II real-time PCR machine using TaqMan probes against ITF-2 and GAPDH (as internal reference) respectively. Briefly, 5 μl

diluted RT product was applied for each 10 µl qPCR reaction using the TaqMan Universal Amplification kit (ThermoFisher/AppliedBiosystems) with customized TaqMan probes (Hs.PT.58.21450367 for ITF-2, 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 Roche 480 IInalinbuild analysis software with relative quantification (ΔCt method applied).

Pathway enrichment and GO analysis

Pathway enrichment and GO analysis were conducted using WebGestalt [75]. KEGG was used as the functional database and human reference genome was used as the background. Maximum (minimum) number of genes for each category was set to 2000 (5) 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.

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.).

Conflict of Interest

The authors declare no conflicts of interest.

References

1. Liu, S., et al., *The early growth response protein 1-miR-30a-5p-neurogenic differentiation factor 1 axis as a novel biomarker for schizophrenia diagnosis and treatment monitoring.* Transl Psychiatry, 2017. **7**(1): p. e998.
2. McGrath, J., et al., *Schizophrenia: a concise overview of incidence, prevalence, and mortality.* Epidemiol Rev, 2008. **30**: p. 67-76.
3. Torrey, E.F., *Prevalence studies in schizophrenia.* Br J Psychiatry, 1987. **150**: p. 598-608.
4. Fromer, M., et al., *Gene expression elucidates functional impact of polygenic risk for schizophrenia.* Nat Neurosci, 2016. **19**(11): p. 1442-1453.
5. Li, Z., et al., *Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia.* Nat Genet, 2017.
6. International Schizophrenia, C., et al., *Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.* Nature, 2009. **460**(7256): p. 748-52.
7. Kirov, G., *CNVs in neuropsychiatric disorders.* Hum Mol Genet, 2015. **24**(R1): p. R45-9.
8. Marshall, C.R., et al., *Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects.* Nat Genet, 2017. **49**(1): p. 27-35.
9. Stefansson, H., et al., *CNVs conferring risk of autism or schizophrenia affect cognition in controls.* Nature, 2014. **505**(7483): p. 361-6.
10. Genovese, G., et al., *Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia.* Nature Neuroscience, 2016. **19**(11): p. 1433-1441.
11. Singh, T., et al., *Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders.* Nat Neurosci, 2016. **19**(4): p. 571-7.
12. Schizophrenia Working Group of the Psychiatric Genomics, C., *Biological insights from 108 schizophrenia-associated genetic loci.* Nature, 2014. **511**(7510): p. 421-7.
13. Mirnics, K., P. Levitt, and D.A. Lewis, *Critical appraisal of DNA microarrays in psychiatric genomics.* Biol Psychiatry, 2006. **60**(2): p. 163-76.
14. Mirnics, K. and J. Pevsner, *Progress in the use of microarray technology to study the neurobiology of disease.* Nature Neuroscience, 2004. **7**(5): p. 434-439.
15. Mirnics, K., P. Levitt, and D.A. Lewis, *DNA microarray analysis of postmortem brain tissue.* Int Rev Neurobiol, 2004. **60**: p. 153-81.
16. Horvath, S., Z. Janka, and K. Mirnics, *Analyzing schizophrenia by DNA microarrays.* Biol Psychiatry, 2011. **69**(2): p. 157-62.
17. Harrison, P.J. and D.R. Weinberger, *Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence.* Mol Psychiatry, 2005. **10**(1): p. 40-68; image 5.
18. Hakak, Y., et al., *Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia.* Proc Natl Acad Sci U S A, 2001. **98**(8): p. 4746-51.
19. Mirnics, K., et al., *Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex.* Neuron, 2000. **28**(1): p. 53-67.
20. Aston, C., L. Jiang, and B.P. Sokolov, *Microarray analysis of postmortem temporal cortex from patients with schizophrenia.* J Neurosci Res, 2004. **77**(6): p. 858-66.
21. Bowden, N.A., R.J. Scott, and P.A. Tooney, *Altered gene expression in the superior temporal gyrus in schizophrenia.* BMC Genomics, 2008. **9**: p. 199.
22. Haroutunian, V., et al., *Variations in oligodendrocyte-related gene expression across multiple cortical regions: implications for the pathophysiology of schizophrenia.* Int J Neuropsychopharmacol, 2007. **10**(4): p. 565-73.
23. Altar, C.A., et al., *Deficient hippocampal neuron expression of proteasome, ubiquitin, and*

mitochondrial genes in multiple schizophrenia cohorts. Biol Psychiatry, 2005. **58**(2): p. 85-96.
24. O'Connor, J.A. and S.E. Hemby, *Elevated GRIA1 mRNA expression in Layer II/III and V pyramidal cells of the DLPFC in schizophrenia.* Schizophr Res, 2007. **97**(1-3): p. 277-88.
25. Arion, D., et al., *Infragranular gene expression disturbances in the prefrontal cortex in schizophrenia: signature of altered neural development?* Neurobiol Dis, 2010. **37**(3): p. 738-46.
26. Hashimoto, T., et al., *Alterations in GABA-related transcriptome in the dorsolateral prefrontal cortex of subjects with schizophrenia.* Mol Psychiatry, 2008. **13**(2): p. 147-61.
27. Vawter, M.P., et al., *Microarray analysis of gene expression in the prefrontal cortex in schizophrenia: a preliminary study.* Schizophr Res, 2002. **58**(1): p. 11-20.
28. Middleton, F.A., et al., *Gene expression profiling reveals alterations of specific metabolic pathways in schizophrenia.* J Neurosci, 2002. **22**(7): p. 2718-29.
29. Iwamoto, K., M. Bundo, and T. Kato, *Altered expression of mitochondria-related genes in postmortem brains of patients with bipolar disorder or schizophrenia, as revealed by large-scale DNA microarray analysis.* Hum Mol Genet, 2005. **14**(2): p. 241-53.
30. Arion, D., et al., *Molecular evidence for increased expression of genes related to immune and chaperone function in the prefrontal cortex in schizophrenia.* Biol Psychiatry, 2007. **62**(7): p. 711-21.
31. Saetre, P., et al., *Inflammation-related genes up-regulated in schizophrenia brains.* BMC Psychiatry, 2007. **7**: p. 46.
32. Dean, B., et al., *Recent advances in postmortem pathology and neurochemistry in schizophrenia.* Curr Opin Psychiatry, 2009. **22**(2): p. 154-60.
33. Hitzemann, R., et al., *Introduction to sequencing the brain transcriptome.* Int Rev Neurobiol, 2014. **116**: p. 1-19.
34. Basso, K., et al., *Reverse engineering of regulatory networks in human B cells.* Nat Genet, 2005. **37**(4): p. 382-90.
35. Han, J.D., et al., *Evidence for dynamically organized modularity in the yeast protein-protein interaction network.* Nature, 2004. **430**(6995): p. 88-93.
36. Jeong, H., et al., *The large-scale organization of metabolic networks.* Nature, 2000. **407**(6804): p. 651-4.
37. Jordan, I.K., et al., *Conservation and coevolution in the scale-free human gene coexpression network.* Mol Biol Evol, 2004. **21**(11): p. 2058-70.
38. Lukashin, A.V., M.E. Lukashev, and R. Fuchs, *Topology of gene expression networks as revealed by data mining and modeling.* Bioinformatics, 2003. **19**(15): p. 1909-1916.
39. Friedman, N., et al., *Using Bayesian networks to analyze expression data.* Journal of Computational Biology, 2000. **7**(3-4): p. 601-620.
40. Gat-Viks, I. and R. Shamir, *Chain functions and scoring functions in genetic networks.* Bioinformatics, 2003. **19 Suppl 1**: p. i108-17.
41. Gardner, T.S., et al., *Inferring genetic networks and identifying compound mode of action via expression profiling.* Science, 2003. **301**(5629): p. 102-105.
42. Yeung, M.K.S., J. Tegner, and J.J. Collins, *Reverse engineering gene networks using singular value decomposition and robust regression.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(9): p. 6163-6168.
43. Ideker, T., et al., *Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.* Science, 2001. **292**(5518): p. 929-34.
44. Butte, A.J., et al., *Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(22): p. 12182-12186.
45. Aytes, A., et al., *Cross-Species Regulatory Network Analysis Identifies a Synergistic Interaction*

between FOXM1 and CENPF that Drives Prostate Cancer Malignancy. Cancer Cell, 2014. **25**(5): p. 638-651.
46. Alvarez, M.J., et al., *Functional characterization of somatic mutations in cancer using network-based inference of protein activity.* Nature Genetics, 2016. **48**(8): p. 838-+.
47. Dutta, A., et al., *Identification of an NKX3.1-G9a-UTY transcriptional regulatory network that controls prostate differentiation.* Science, 2016. **352**(6293): p. 1576-1580.
48. Brichta, L., et al., *Identification of neurodegenerative factors using translatome-regulatory network analysis.* Nature Neuroscience, 2015. **18**(9): p. 1325-+.
49. Evgrafov, O.V., et al., *Olfactory neuroepithelium-derived neural progenitor cells as a model system for investigating the molecular mechanisms of neuropsychiatric disorders.* Psychiatric Genetics, 2011. **21**(5): p. 217-228.
50. Margolin, A.A., et al., *Reverse engineering cellular networks.* Nat Protoc, 2006. **1**(2): p. 662-71.
51. Lizio, M., et al., *Gateways to the FANTOM5 promoter level mammalian expression atlas.* Genome Biol, 2015. **16**: p. 22.
52. Vaquerizas, J.M., et al., *A census of human transcription factors: function, expression and evolution.* Nat Rev Genet, 2009. **10**(4): p. 252-63.
53. Han, H., et al., *TRRUST: a reference database of human transcriptional regulatory interactions.* Sci Rep, 2015. **5**: p. 11432.
54. Boyajyan, A.S., S.A. Atshemyan, and R.V. Zakharyan, *Association of schizophrenia with variants of genes that encode transcription factors.* Molecular Biology, 2015. **49**(6): p. 875-880.
55. Van Schijndel, J.E. and G.J.M. Martens, *Gene Expression Profiling in Rodent Models for Schizophrenia.* Current Neuropharmacology, 2010. **8**(4): p. 382-393.
56. Volk, D.W., J.R. Edelson, and D.A. Lewis, *Cortical inhibitory neuron disturbances in schizophrenia: role of the ontogenetic transcription factor Lhx6.* Schizophr Bull, 2014. **40**(5): p. 1053-61.
57. Emamian, E.S., *AKT/GSK3 signaling pathway and schizophrenia.* Front Mol Neurosci, 2012. **5**: p. 33.
58. Guo, A.Y., et al., *A novel microRNA and transcription factor mediated regulatory network in schizophrenia.* BMC Syst Biol, 2010. **4**: p. 10.
59. Pickard, B.S., et al., *Disruption of a brain transcription factor, NPAS3, is associated with schizophrenia and learning disability.* American Journal of Medical Genetics Part B-Neuropsychiatric Genetics, 2005. **136b**(1): p. 26-32.
60. Millar, J.K., et al., *Disruption of two novel genes by a translocation co-segregating with schizophrenia.* Hum Mol Genet, 2000. **9**(9): p. 1415-23.
61. 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.* Molecular Psychiatry, 2001. **6**(2): p. 173-178.
62. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
63. Lefebvre, C., et al., *A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers.* Molecular Systems Biology, 2010. **6**.
64. Jiang, Z. and R. Gentleman, *Extensions to gene set enrichment.* Bioinformatics, 2007. **23**(3): p. 306-13.
65. Carro, M.S., et al., *The transcriptional network for mesenchymal transformation of brain tumours.* Nature, 2010. **463**(7279): p. 318-U68.
66. Masliah, E., et al., *Distinctive patterns of DNA methylation associated with Parkinson disease: identification of concordant epigenetic changes in brain and peripheral blood leukocytes.*

Epigenetics, 2013. **8**(10): p. 1030-8.
67. Rehman, H.U., *Role of CRH in the pathogenesis of dementia of Alzheimer's type and other dementias.* Curr Opin Investig Drugs, 2002. **3**(11): p. 1637-42.
68. Bennett, A.O.M., *Stress and anxiety in schizophrenia and depression: glucocorticoids, corticotropin-releasing hormone and synapse regression.* Aust N Z J Psychiatry, 2008. **42**(12): p. 995-1002.
69. Maycox, P.R., et al., *Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function.* Mol Psychiatry, 2009. **14**(12): p. 1083-94.
70. Lopez-Atalaya, J.P., et al., *Genomic targets, and histone acetylation and gene expression profiling of neural HDAC inhibition.* Nucleic Acids Res, 2013. **41**(17): p. 8072-84.
71. 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, 2016. **59**(6-7): p. 310-4.
72. Blake, D.J., et al., *TCF4, schizophrenia, and Pitt-Hopkins Syndrome.* Schizophr Bull, 2010. **36**(3): p. 443-7.
73. Wirgenes, K.V., et al., *TCF4 sequence variants and mRNA levels are associated with neurodevelopmental characteristics in psychotic disorders.* Transl Psychiatry, 2012. **2**: p. e112.
74. Quednow, B.B., M.M. Brzozka, and M.J. Rossner, *Transcription factor 4 (TCF4) and schizophrenia: integrating the animal and the human perspective.* Cell Mol Life Sci, 2014. **71**(15): p. 2815-35.
75. Wang, J., et al., *WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W77-83.
76. Happ, H., et al., *8q21.11 microdeletion in two patients with syndromic peters anomaly.* Am J Med Genet A, 2016. **170**(9): p. 2471-5.
77. Parsa, C.F. and M.P. Robert, *Thromboembolism and congenital malformations: from Duane syndrome to thalidomide embryopathy.* JAMA Ophthalmol, 2013. **131**(4): p. 439-47.
78. Mathelier, A., et al., *JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles.* Nucleic Acids Res, 2016. **44**(D1): p. D110-5.
79. 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, 2017. **21**(3): p. 305-318 e8.
80. Lennertz, L., et al., *Impact of TCF4 on the genetics of schizophrenia.* Eur Arch Psychiatry Clin Neurosci, 2011. **261 Suppl 2**: p. S161-5.
81. Forrest, M.P., et al., *Knockdown of human TCF4 affects multiple signaling pathways involved in cell survival, epithelial to mesenchymal transition and neuronal differentiation.* PLoS One, 2013. **8**(8): p. e73169.
82. Shi, J., et al., *Common variants on chromosome 6p22.1 are associated with schizophrenia.* Nature, 2009. **460**(7256): p. 753-7.
83. Jarskog, L.F., et al., *Apoptotic mechanisms in the pathophysiology of schizophrenia.* Prog Neuropsychopharmacol Biol Psychiatry, 2005. **29**(5): p. 846-58.
84. Fan, Y., et al., *Focal adhesion dynamics are altered in schizophrenia.* Biol Psychiatry, 2013. **74**(6): p. 418-26.
85. Suda, S., et al., *Decreased expression of axon-guidance receptors in the anterior cingulate cortex in autism.* Mol Autism, 2011. **2**(1): p. 14.
86. Devor, A., et al., *Genetic evidence for role of integration of fast and slow neurotransmission in schizophrenia.* Mol Psychiatry, 2017. **22**(6): p. 792-801.
87. Comasco, E., et al., *Genetic and Functional Study of L-Type Amino Acid Transporter 1 in*

Schizophrenia. Neuropsychobiology, 2016. **74**(2): p. 96-103.
88. Yates, A., et al., *Ensembl 2016.* Nucleic Acids Res, 2016. **44**(D1): p. D710-6.
89. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
90. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.

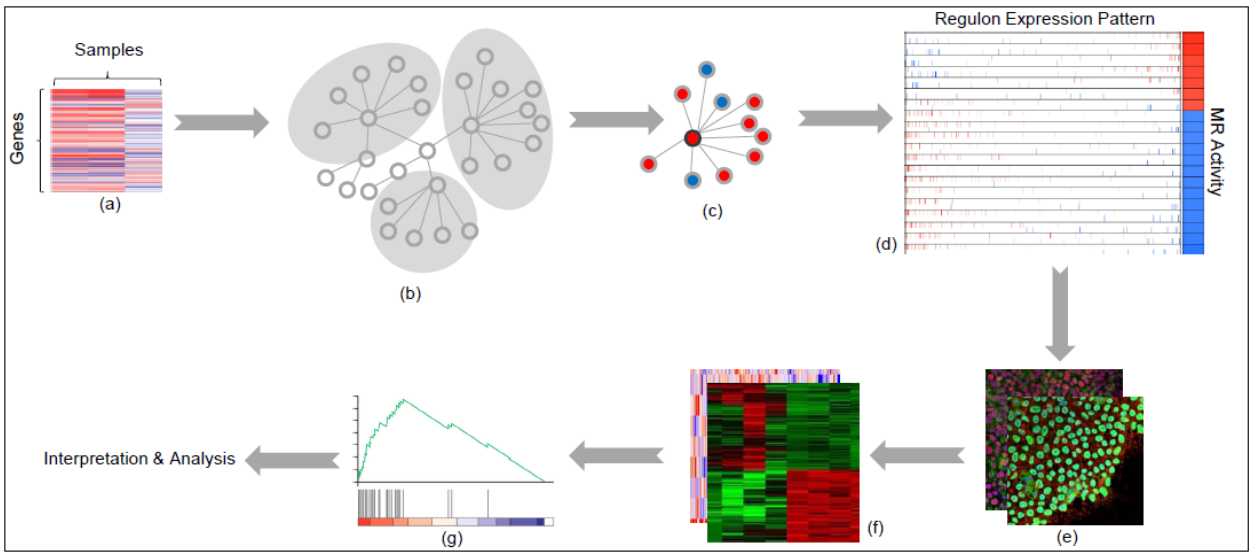


Figure 1. A schematic view of this study; (a) obtaining gene expression data, (b) transcriptional network reconstruction using the provided RNA-Seq data using ARACNe, (c) identification of highly influential genes (MRs) through virtual protein activity analysis using VIPER, (d) selecting the top ranked identified MRs shared between the two independent interactomes, (e) knocking down TCF4 expression in human induced pluripotent stem cells (hiPSC), (f) RNA-sequencing and data processing, (g) TCF4 regulon expression analysis and validation.

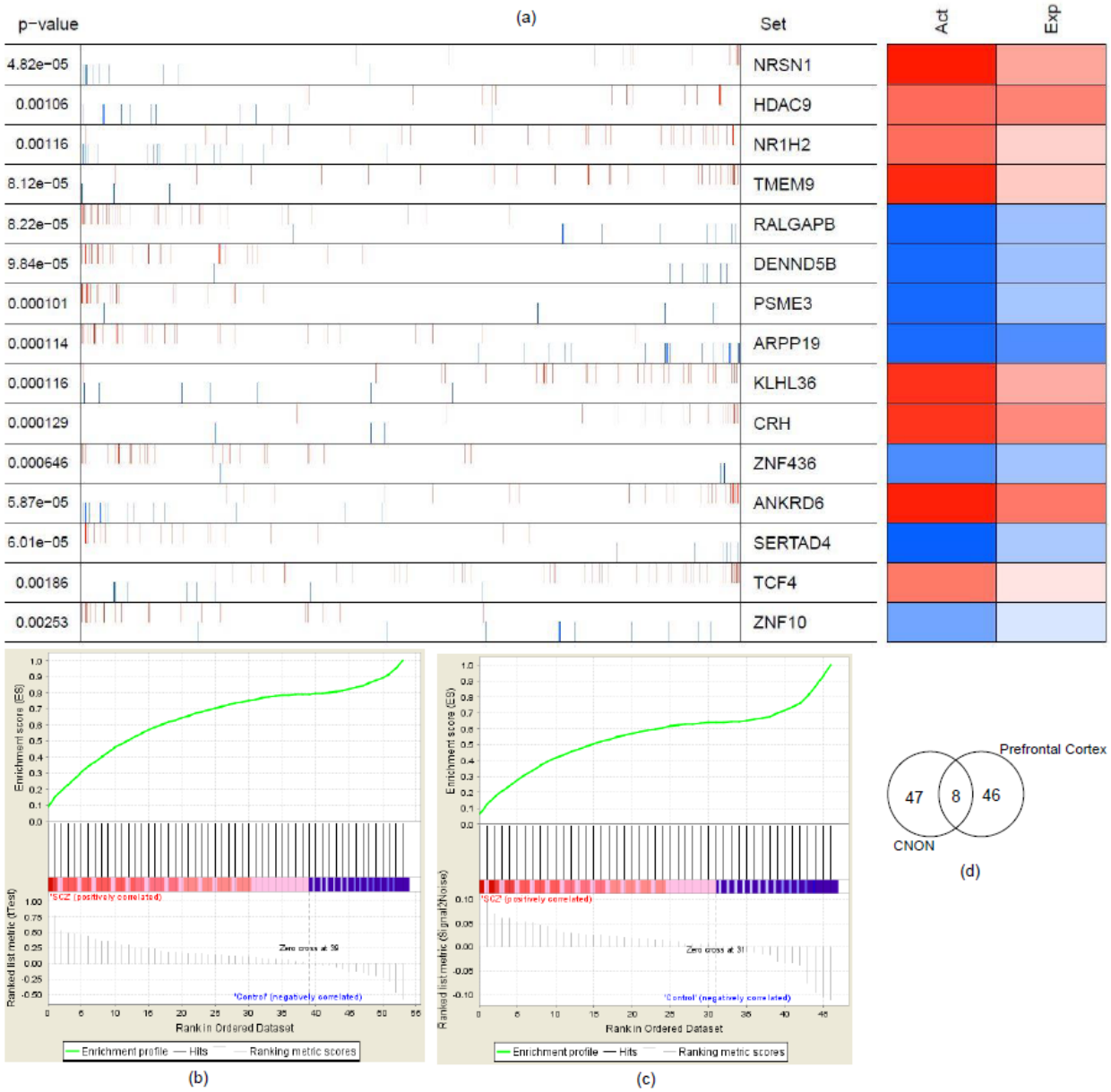


Figure 2. (a) Virtual Inference of Protein-activity by Enriched Regulon analysis; (b) Contribution of TCF4 interactome expression to SCZ in prefrontal cortex RNA-Seq data; (c) Contribution of TCF4 interactome expression to SCZ in CNON RNA-Seq data; (d) Overlap of interactome constructed from prefrontal cortex and CNON gene expression.

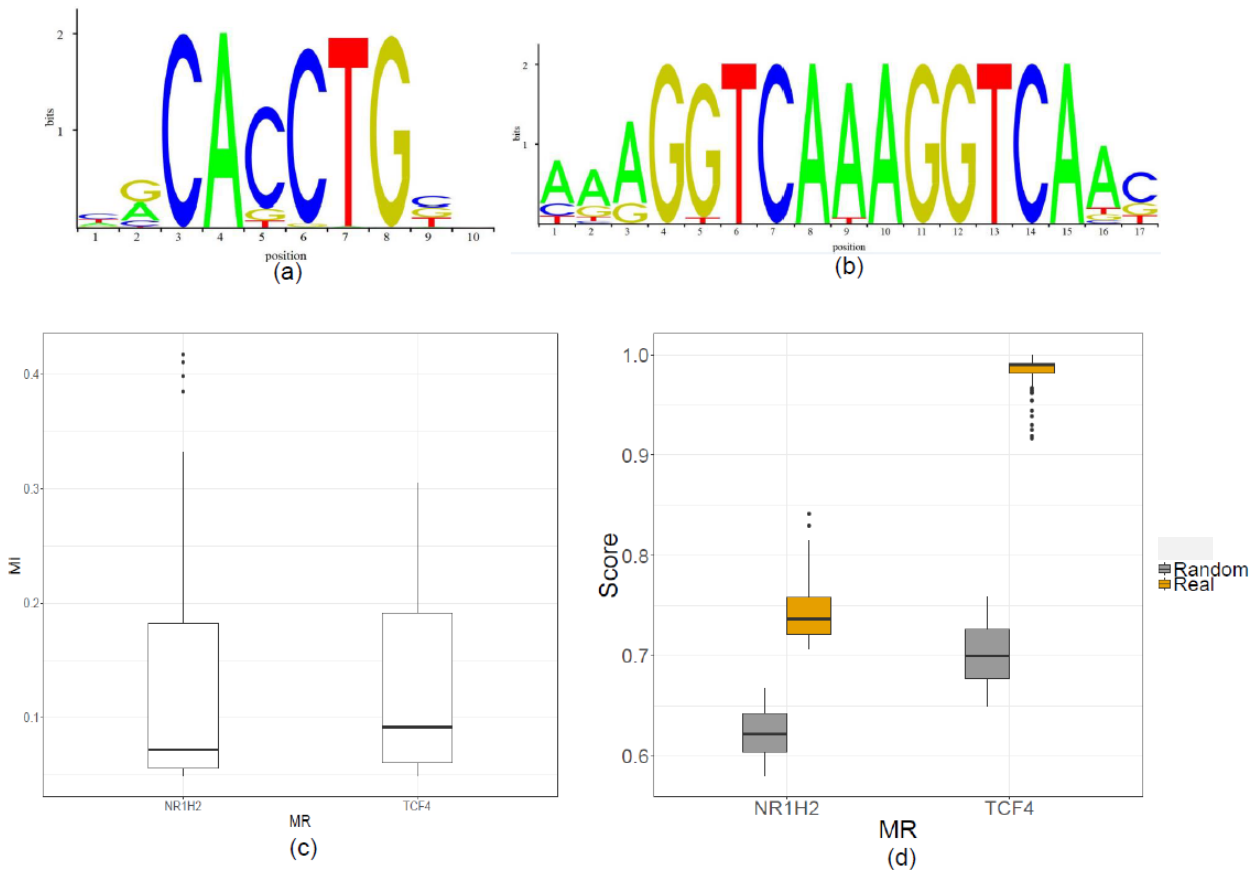


Figure 3. Transcription binding enrichment analysis of TCF4 and NR1H2 using the union of the targets obtained from both independent datasets; (a) TCF4 motif, (b) NR1H2 motif, (c) Mutual information values, (d)Transcription binding enrichment scores.

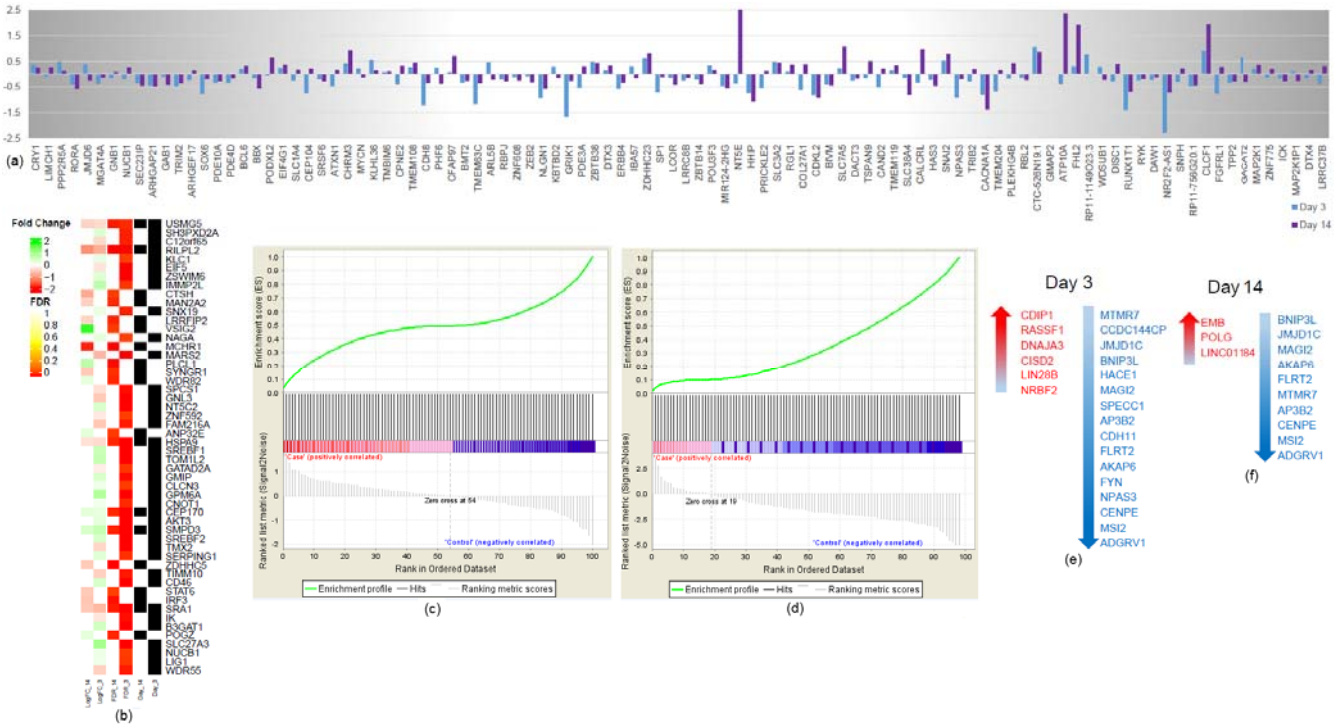


Figure 4. (a) Log fold change of the identified TCF4 regulon on days 3 and 14; (b) expression patterns of credible SCZ susceptibility genes ([12]) affected by TCF4 knockdown; (c) gene set enrichment of TCF4 regulon based on the RNA-Seq data from TCF4 knockdown at day 3; (d) gene set enrichment of TCF4 regulon based on the RNA-Seq data from TCF4 knockdown at day 14; (e) List of newly GWAS discovered genes ([5]) significantly expressed after TCF4 knockdown at day 3; (f) List of newly GWAS discovered genes ([5]) significantly expressed after TCF4 knockdown at day 14.

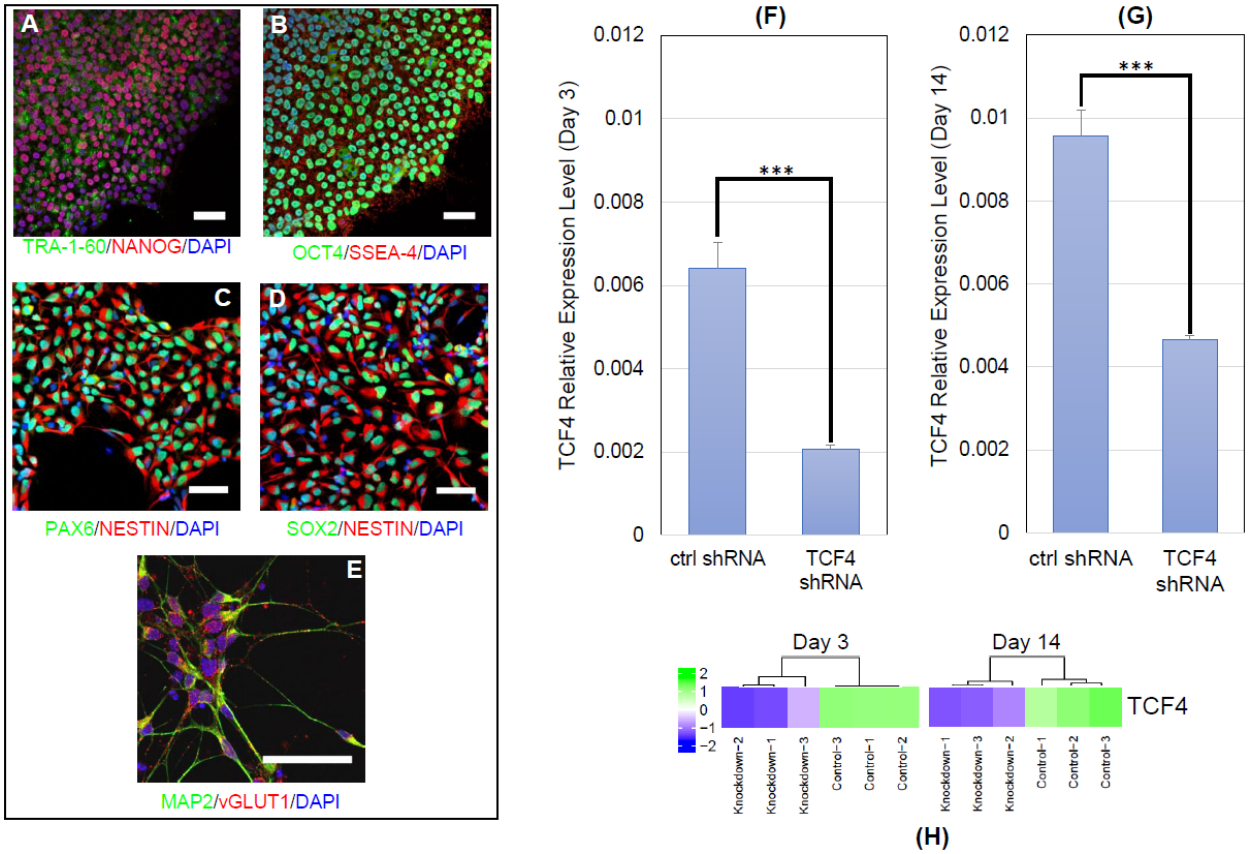
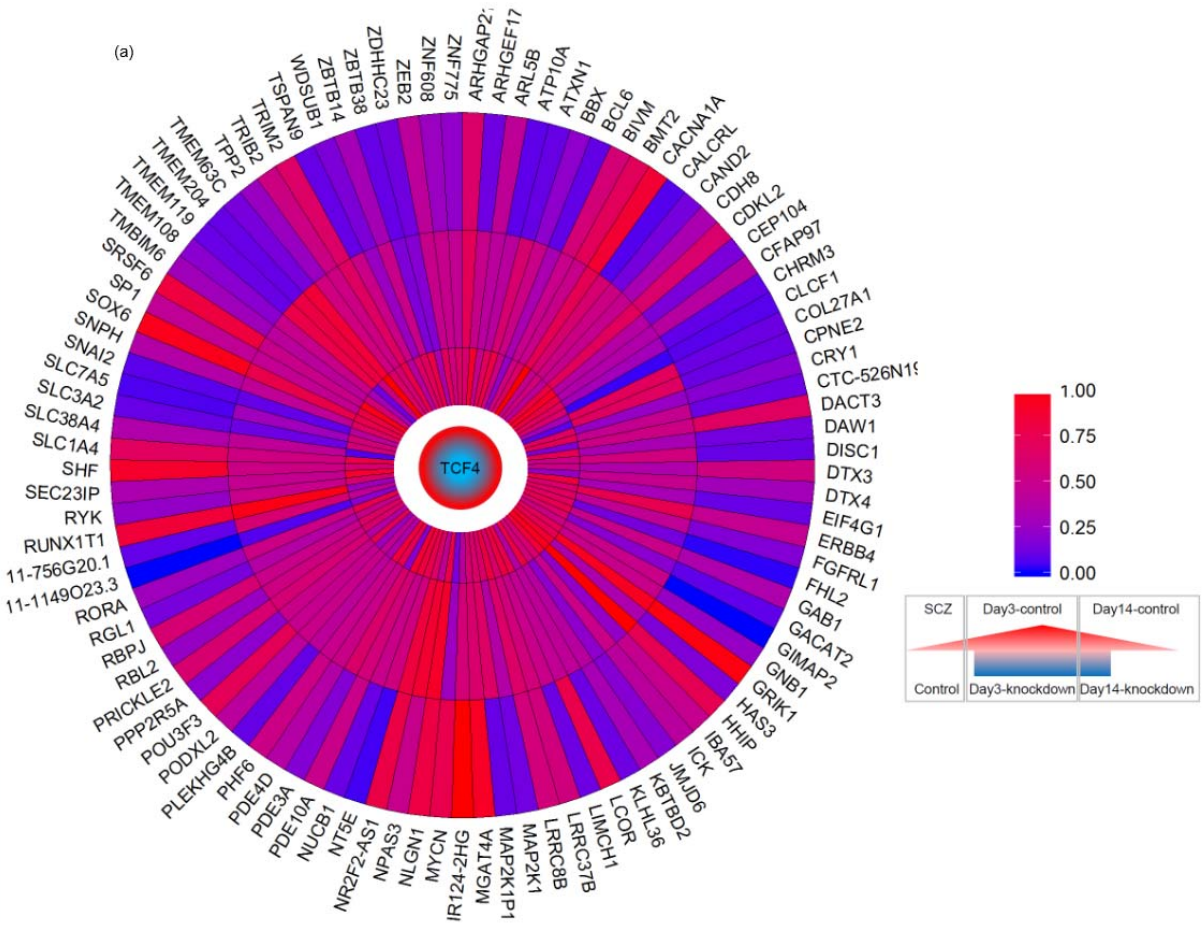


*Figure 5. TCF4 knockdown in hiPSC-derived neuronal precursor cells (NSC) and glutamatergic neurons. (A and B) Representative immunofluorescence (IF)-staining images of hiPSCs. hiPSCs are stained positive for pluripotency markers TRA-1-60 (green) and NANOG (red) (in A), OCT4 (green) and SSEA-4 (red) (in B). (C and D) Representative IF-staining images of NPCs (day 3 post plating cells). NPCs are stained positive for PAX6 (green) and NESTIN (red) (in C), and SOX2 (green) (in D). (E) Representative image of day-14 neurons stained positive for MAP2 (green) and vGLUT1 (red). Scale bars: 50 um in all images. Cell nucleus are stained with DAPI (blue) (A-E). (F and G) TCF4 knockdown efficiency measured by qPCR in NPCs (3 days after plating) and in glutamatergic neurons (day 14 post neuronal induction), respectively. For knocking down TCF4, cells were transducted with either control or TCF4 shRNA lentiviral particles and TCF4 expression was assayed by qPCR 48 hours post transduction. GAPDH is used as the endogenous control to normalize the TCF4 expression. Error bars: mean ± SD (n=4). ***: P < 0.001, Student's t-test, two-tailed, heteroscedastic; (h) TCF4 expression levels at days 3 and 14 based on the RNA-Seq generated data.*


(a)
TCF4
ARHGAP21
ARHGEF17
ARL5B
ATP10A
ATXN1
BBX
BCL6
BIVM
BMT2
CACNA1A
CALCRL
CAND2
CDH8
CDKL2
CEP104
CFAP97
CHRM3
CLCF1
COL27A1
CPNE2
CRY1
CTC-526N19
DACT3
DAW1
DISC1
DTX3
DTX4
EIF4G1
ERBB4
FGFRL1
FHL2
GAB1
GACAT2
GIMAP2
GNB1
GRIK1
HAS3
HHIP
IBA57
ICK
JMJD6
KBTBD2
KLHL36
LCOR
LIMCH1
LRRC37B
LRRC8B
MAP2K1
MAP2K1P1
MGAT4A
IR124-2HG
MYCN
NLGN1
NPAS3
NR2F2-AS1
NT5E
NUCB1
PDE10A
PDE3A
PDE4D
PHF6
PLEKHG4B
PODXL2
POU3F3
PPP2R5A
PRICKLE2
RBL2
RBPJ
RGL1
RORA
11-1149O23.3
11-756G20.1
RUNX1T1
RYK
SEC23IP
SHF
SLC1A4
SLC38A4
SLC3A2
SLC7A5
SNAI2
SNPH
SOX6
SP1
SRSF6
TMBIM6
TMEM108
TMEM119
TMEM204
TMEM63C
TPP2
TRIB2
TRIM2
TSPAN9
WDSUB1
ZBTB14
ZBTB38
ZDHHC23
ZEB2
ZNF608
ZNF775
1.00
0.75
0.50
0.25
0.00
SCZ
Day3-control
Day14-control
Control
Day3-knockdown
Day14-knockdown

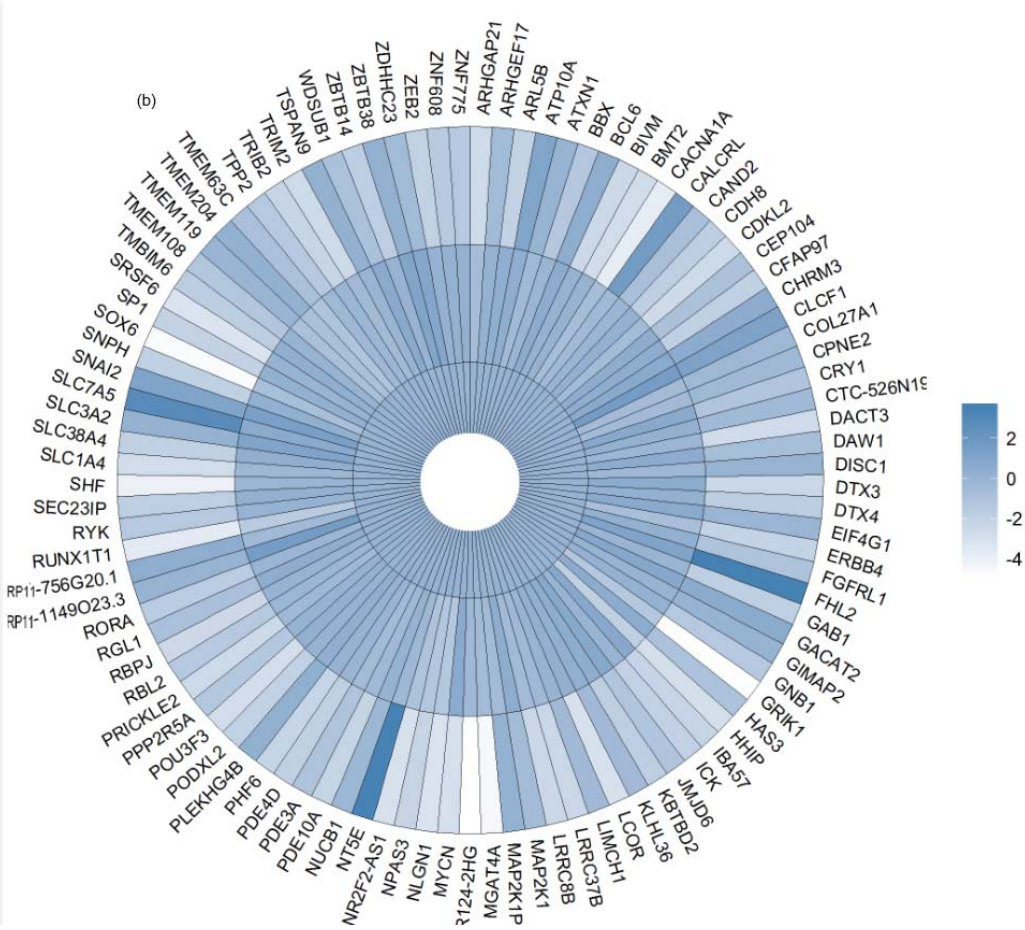

(b)
ARHGAP21
ARHGEF17
ARL5B
ATP10A
ATXN1
BBX
BCL6
BIVM
BMT2
CACNA1A
CALCRL
CAND2
CDH8
CDKL2
CEP104
CFAP97
CHRM3
CLCF1
COL27A1
CPNE2
CRY1
CTC-526N19
DACT3
DAW1
DISC1
DTX3
DTX4
EIF4G1
ERBB4
FGFRL1
FHL2
GAB1
GACAT2
GIMAP2
GNB1
GRIK1
HAS3
HHIP
IBA57
ICK
JMJD6
KBTBD2
KLHL36
LCOR
LIMCH1
LRRC37B
LRRC8B
MAP2K1
MAP2K1P
MGAT4A
R124-2HG
MYCN
NLGN1
NPAS3
NT5E
NR2F2-AS1
NUCB1
PDE10A
PDE3A
PDE4D
PHF6
PLEKHG4B
PODXL2
POU3F3
PPP2R5A
PRICKLE2
RBL2
RBPJ
RGL1
RORA
RP11-1149O23.3
RP11-756G20.1
RUNX1T1
RYK
SEC23IP
SHF
SLC1A4
SLC38A4
SLC3A2
SLC7A5
SNAI2
SNPH
SOX6
SP1
SRSF6
TMBIM6
TMEM108
TMEM119
TMEM204
TMEM63C
TPP2
TRIB2
TRIM2
TSPAN9
WDSUB1
ZBTB14
ZBTB38
ZDHHC23
ZEB2
ZNF608
ZNF775
2
0
-2
-4

Figure 6. The identified TCF4 interactome and their contributions to expression of TCF4. The inner track represents the expression contribution in CMC and CNON data, the middle track represents the TCF4 knockdown RNA-Seq data at day 3 and the outer track represents the TCF4 knockdown RNA-Seq data at day 14; (a) GSEA enrichment score of the TCF4 regulon in three different states; (b) rank metric score of the regulon representing the contribution of each gene to expression pattern of TCF4.