bioRxiv preprint doi: https://doi.org/10.1101/133363; this version posted May 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Reverse Engineering of Trascriptional Networks Uncovers Candidate Master Regulators Governing Neuropathology of Schizophrenia

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

[1]Institute for Genomic Medicine, Columbia University, New York, 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.

bioRxiv preprint doi: https://doi.org/10.1101/133363; this version posted May 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Tissue-specific reverse engineering of transcriptional networks has led to groundbreaking discoveries uncovering underlying regulators of cellular networks in various diseases. However, whether these approaches can be applied to complex psychiatric diseases is largely unexplored, partly due to the general lack of appropriate cellular models for mental disorders. In this study, using a recently published high quality RNA-seq data on dorsolateral prefrontal cortex from 307 Schizophrenia (SCZ) patients and 245 controls, we deconvoluted the transcriptional network aiming at the identification of master regulators mediating expression of a large body of 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 five candidate master regulators (MRs), including TCF4, NR1H2, HDAC9, ZNF10, and ZNF436. TCF4 was previously identified as a SCZ susceptibility gene, but its regulatory subnetworks had been elusive. Other genes have not been convincingly associated with SCZ in previous studies. Additional analysis of predicted transcription factor binding site, ChIP-Seq data and ATAC-Seq data confirmed many predicted regulatory targets by the identified MRs. Our study uncovered a few candidate master regulators for SCZ that affects a collection of genes, and these master regulators may serve as therapeutic targets for intervention.

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. Despite a vast amount of research workforce, its etiology has a long way to be fully elucidated. Over the past decade, a considerable effort has been put in establishing genetic basis of SCZ to uncover 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) [5] and common variants [6].

Thus far, post-mortem gene-expression profiling has been the back-bone of a large portion of research studies conducted on SCZ. This technique bears several limitations, making it challenging when interpreting the experimental outcomes [7-9]. 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 [10]. On the other hand, SCZ is a complex disorder causing genetic disturbances across a wide range of genes while SCZ patients have diverse genetic backgrounds and personality traits [11]. 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 [12, 13] and temporal cortices [14, 15] and hippocampus [16, 17]. In this regard, several points should be mentioned including: (1) SCZ is associated with genomic disturbances across several cortical regions affecting a large spectrum of genes [16]; (2) multiple cell types host transcriptome changes comprised of subclasses of principal neurons [18, 19], interneurons [12, 20], and olygodendrocytes [12] 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 [13, 21], energy metabolism [22, 23], immune response [24, 25], and inhibitory neurotransmission [21, 28]; (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 [9, 10, 26]. Despite considerable number of research studies conducted on case-control gene-expression samples, a great deal of the reported findings

bioRxiv preprint doi: https://doi.org/10.1101/133363; this version posted May 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 [27], 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 development, 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 [28], cellular networks possess the scale-free property mostly observed in protein-protein interaction and metabolic networks [29, 30]. One of the most commonly used network-based representation approaches of cellular processes is co-expression networks bearing the scale-free property [31, 32]. Nevertheless, co-expression networks and other similar means are not comprehensive enough to fully recapitulate the entire underlying molecular structures affecting the disease phenotype. In principle, currently existing expression network analysis approaches fall into four categories including [28]: optimization methods [33, 34], which usually optimize the network based upon a certain criterion; regression techniques [35, 36], which fit the data to a priori models; Multi-omics integrative approaches [37], and statistical methods [38]. 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 [28]. 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., [28] and its successful applications in a wide range of complex diseases such as cancer [39, 40], prostate differentiation [41], and neurodegenerative diseases [42], the way to infer causal relationships between transcription factors and their downstream regulon was paved.

A powerful substitute to microarray technology to accurately characterize transcription at the gene level is RNA sequencing. 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 Cultured primary Neuronal cells derived from Olfactory Neuroepithelium (CNON) of individuals with and without SCZ.

Results

Analysis on the CMC data sets

The CMC data contains 15971 transcripts extracted from 552 brain samples 307 of which 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

bioRxiv preprint doi: https://doi.org/10.1101/133363; this version posted May 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Reconstruction of Accurate Cellular Networks) as powerful and versatile tool to reconstructing cellular networks [28, 43]. 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 process is performed by applying Data Processing Inequality (DPI), a well-known term in data transmission theory. 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 [43]) 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 that will be called MRs. We curated a long list of known human TFs from three sources including FANTOM5 consortium [44], a curated set by Vaquerizas [45], and TRRUST [46]. A total of 2198 TFs 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 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. These TFs can be shortlisted as follows: JUND [47]; KCNIP3 [48]; LHX6 [49]; NRG1 and GSK3B [50]; NFE2 and MZF1 [51]; NPAS3 [52]; DISC1 [53, 54]. 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.

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) [40].

We fed the output of the previous stage (ARACNe) network to VIPER in order to check whether or not any of the identified MRs have significant regulatory role in expression level of their downstream regulon. We reconstructed the network using 1466 MRs consisting 24548 interactions. One of the unrealistic assumptions in analyzing regulatory networks is to consider a uniform distribution of the targets as a prior for analyzing the regulatory effects of MRs. This is due to the fact that the degree of co-regulation in transcriptional networks are high and the assumption of statistical independence of gene expression is unrealistic. To account for these correlations between the genes, 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. The computational results of the ten top MRs are depicted in Figure 1.

According to Figure 1, the first column on the left represents the expression of the regulon of the MR where blue means repression and red means activation. 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 [55] proposed by Subramanian et al. to identify the

bioRxiv preprint doi: https://doi.org/10.1101/133363; this version posted May 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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. Table 1 denotes these MRs along with the number of their target genes and their DE p-value between cases and controls at FDR<0.05.

According to [56, 57], 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 [58]. We computed the enrichment of the co-regulons for the top 10 co-regulators (Figure 2).

According to Figure 2, 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 [59], 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 [60, 61]. ARPP19 has been identified to be DE between two large SCZ cohorts associated with nerve terminal function [62]. 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 [63]. 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 [64]. This gene has also been reported to host a SNP in its intron that might be associated with SCZ [65-67]. SCZ has also been reported to host genetic variants contributing to SCZ development [6]. In the following sections, we will conduct similar network analysis and will discuss common findings between these two numerical experiments.

Results on CNON Data

The CNON data contain 23920 transcripts extracted from 255 brain samples 143 of which cases and 112 controls. Adopting the same set up parameters, we ran ARACNe [28, 43] 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 [43]) 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. The entire target genes of the five MRs in both CMC and CNON datasets are represented in Figure 3.

bioRxiv preprint doi: https://doi.org/10.1101/133363; this version posted May 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Common Findings

Using the top 15 MRs identified in the CMC data (Figure 1), 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 having 48, 23, 11, 38, and 14 gene targets. 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 [68]. 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 [69] and congenital malformations [70]. 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 [71] (Figure 4) on TCF4 and NR1H2. First, sequence motif of these genes were extracted. Second, 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 4). 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 4. For both TCF4 and NR1H2 the enrichment scores are significantly larger than random enrichment scores implicating high confidence in finding correct binding positions in the target genes. These results will further be validated *in vivo*.

Conducting two-sided t-test (FDR<01.05) between cases and controls in the CMC data, the following MRs were Differentially Expressed (DE) including: NR1H2 (p=0.0015), ZNF436 (p=1.67e-05), ZNF10 (p=0.0043), and HDAC (p=1.93e-07). Also TCF4 was partially DE with p=0.0180. Looking at the target genes of these MRs, we note that most of the target genes are DE.

expression Quantitative Trait Loci (eQTL) Analysis

In order to probe the role of genetic variants on expression of the identified MRs, we analyzed the associations between cis and trans acting Single Nucleotide Polymorphisms (SNPs) to identify SNPs associated with the expression of the MRs of interest. We used the eQTL analysis results provided by Fromer et al [4] on the CMC data and extracted the associated SNPs for the candidate MRs. It should be mentioned that the eQTL analysis conducted in [4] replicates eQTLs characterized by a variety of RNA-seq and microarray studies from different sources such as GTEx v6 [72], Harvard Brain Bank [73], BrainCloud [74], the National Institute of Health (NIH) [75], and the UK Britain Expression Consortium [76]. Associated SNPs having distance over 1 Mbp from each gene were called trans-acting. The list of detected associations between the SNPs and the MRs is are depicted in Figure 5. The false discovery threshold of FDR ≤ 0.05 was applied to filter out insignificant associations. In this figure, SNPs that passed the FDR threshold have been marked by star sign.

As a trans-acting SNP, rs4713722 was associated with TCF4 (p=9.04e-08). Two cis-acting SNPs were

bioRxiv preprint doi: https://doi.org/10.1101/133363; this version posted May 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

significantly associated with HDAC9 including rs2528407 (p=0.00047997) and rs4413678 (p=0.000698039). We also identified three susceptible cis-acting loci associated with ZNF436 including rs11589432 (p=0.000328594), rs11588134 (p=0.000329618), and rs3795295 (p=0.000331596). No association had passed the FDR correction filter for NR1H2 and ZNF10. Of these SNPs, rs4413678 has already been captured to be associated with major depression disorder but no similar finding for the other SNPs were observed in the literature. A conclusion that can be made here is that the identified loci may have direct impact not only on their associated genes, but exert particular expression patterns on the entire regulon of the MRs.

We checked to see if the corresponding SNPs to each of the MRs were in LD. On, ZNF436, 100 pairwise SNPs were in LD having ($R^2 \geq 0.86$). We then looked at the SNPs that were in LD with the identified significant candidate loci in this gene. The three candidate SNPs in ZNF436 were in LD having ($R^2 \geq 0.966$). In HDAC9, 150 pairs of SNPs were in LD but none of the two candidate SNPs rs2528407 and rs4413678 were in LD with others.

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 regulons by using our ATAC-seq-based empirical TF-binding footprints (inferred by PIQ tool) data sets in iPSC-derived glutamatergic neurons. 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 d30, d41 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) (in an R package). 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. 20 target genes out of 102 regulons of TCF4 (total TCF4 targets in ATAC-seq data=3947) were validated to be bound by TCF4 (Fisher's exact test P=0.175). These genes are listed in Table 2.

To make further investigations, we downloaded the TCF4 ChIP-seq data in HEK293 cell line from ENCODE [77]. After retrieving the data, we annotated the data using ANNOVAR [78] to obtain the enriched regions by TCF4. 46 TCF4 peaks (total targets= 7008) were observed within the promoter region of the target genes (Fisher's exact test P=1.702e-04) which are listed in Table 3.

In another effort, we investigated several ChIP-seq experiments available on Cistrome.org. Two MRs of NR1H2 and ZNF436 were available for human subjects in blood tissue. 61 (out of 68) (P=1.42e-17, total target number= ~10000) and 20 (out of 66) (Fisher's exact test P=0.124, 7582 targets) target genes were validated to be affected by NR1H2 and ZNF436, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/133363; this version posted May 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

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 pruned 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 ($MI(x,y) < \min(MI(x,z), MI(z,y))$).

VIPER

The regulon enrichment on gene expression signatures was tested by VIPER 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, Students 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) that will be discussed below. 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.

Analytic rank-based enrichment analysis (aREA)

aREA tests for a global shift in the positions of each regulon genes when projected on the rank-sorted gene expression signature. Following up on the work in [79, 80], we used the mean of the quantile-transformed rank positions as test statistic (enrichment score). The enrichment score is computed twice: first by a one-tail approach, based on the absolute value of the gene expression signature obtained from statistical t-test; and then by a two-tail approach, where the positions of the genes whose expression is repressed by the regulator (R) are inverted in the gene expression signature before computing the enrichment score. The one-tail and two-tail enrichment score estimates are integrated while weighting their contribution based on the estimated mode of regulation. The contribution of each target gene from a given regulon to the enrichment score is also weighted based on the regulator-target gene interaction confidence. At last, the statistical significance for the enrichment score is estimated by comparison to a null model generated by permuting the samples uniformly at random or by an analytic approach equivalent to shuffle the genes in the signatures uniformly at random.

Transcription factor binding site enrichment analysis

Human reference genome (version GRCh37.p13) was used to extract the DNA sequence around TSS for transcription binding enrichment analysis. We obtained the gene coordinates from Ensembl database 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 TF. JASPAR database contains Position Weight Matrices (PWM) for each TF. Then, using a modified Needleman-Wunsch algorithm and the corresponding PWM of the TF, input sequence is

bioRxiv preprint doi: https://doi.org/10.1101/133363; this version posted May 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

scanned to check whether or not the motif is enriched in the sequence.

Acknowledgements

We are grateful for 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 development team of the ARACNe software for comments on our software usage, and thank the Wang lab members for helpful suggestions and discussions. This study was supported by NIH grant MH108728 (K.W.) and MH086874 (O.E. and J.A.K.).

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. Kirov, G., CNVs in neuropsychiatric disorders. Hum Mol Genet, 2015. 24(R1): p. R45-9.
6. Schizophrenia Working Group of the Psychiatric Genomics, C., Biological insights from 108 schizophrenia-associated genetic loci. Nature, 2014. 511(7510): p. 421-7.
7. Mirnics, K., P. Levitt, and D.A. Lewis, Critical appraisal of DNA microarrays in psychiatric genomics. Biol Psychiatry, 2006. 60(2): p. 163-76.
8. 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.
9. Mirnics, K., P. Levitt, and D.A. Lewis, DNA microarray analysis of postmortem brain tissue. Int Rev Neurobiol, 2004. 60: p. 153-81.
10. Horvath, S., Z. Janka, and K. Mirnics, Analyzing schizophrenia by DNA microarrays. Biol Psychiatry, 2011. 69(2): p. 157-62.
11. 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.
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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.
22. Middleton, F.A., et al., Gene expression profiling reveals alterations of specific metabolic pathways in schizophrenia. J Neurosci, 2002. 22(7): p. 2718-29.
23. 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.
24. 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.
25. Saetre, P., et al., Inflammation-related genes up-regulated in schizophrenia brains. BMC Psychiatry, 2007. 7: p. 46.
26. Dean, B., et al., Recent advances in postmortem pathology and neurochemistry in schizophrenia.

bioRxiv preprint doi: https://doi.org/10.1101/133363; this version posted May 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Curr Opin Psychiatry, 2009. 22(2): p. 154-60.
27. Hitzemann, R., et al., Introduction to sequencing the brain transcriptome. Int Rev Neurobiol, 2014. 116: p. 1-19.
28. Basso, K., et al., Reverse engineering of regulatory networks in human B cells. Nat Genet, 2005. 37(4): p. 382-90.
29. Han, J.D., et al., Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature, 2004. 430(6995): p. 88-93.
30. Jeong, H., et al., The large-scale organization of metabolic networks. Nature, 2000. 407(6804): p. 651-4.
31. 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.
32. 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.
33. Friedman, N., et al., Using Bayesian networks to analyze expression data. Journal of Computational Biology, 2000. 7(3-4): p. 601-620.
34. Gat-Viks, I. and R. Shamir, Chain functions and scoring functions in genetic networks. Bioinformatics, 2003. 19 Suppl 1: p. i108-17.
35. Gardner, T.S., et al., Inferring genetic networks and identifying compound mode of action via expression profiling. Science, 2003. 301(5629): p. 102-105.
36. 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.
37. Ideker, T., et al., Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science, 2001. 292(5518): p. 929-34.
38. 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.
39. 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.
40. 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-+.
41. 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.
42. Brichta, L., et al., Identification of neurodegenerative factors using translatome-regulatory network analysis. Nature Neuroscience, 2015. 18(9): p. 1325-+.
43. Margolin, A.A., et al., Reverse engineering cellular networks. Nat Protoc, 2006. 1(2): p. 662-71.
44. Lizio, M., et al., Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol, 2015. 16: p. 22.
45. Vaquerizas, J.M., et al., A census of human transcription factors: function, expression and evolution. Nat Rev Genet, 2009. 10(4): p. 252-63.
46. Han, H., et al., TRRUST: a reference database of human transcriptional regulatory interactions. Sci Rep, 2015. 5: p. 11432.
47. 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.
48. 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.
49. 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.
50. Emamian, E.S., AKT/GSK3 signaling pathway and schizophrenia. Front Mol Neurosci, 2012. 5: p. 33.
51. Guo, A.Y., et al., A novel microRNA and transcription factor mediated regulatory network in schizophrenia. BMC Syst Biol, 2010. 4: p. 10.
52. 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.
53. 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.
54. 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.
55. 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.
56. 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.
57. Jiang, Z. and R. Gentleman, Extensions to gene set enrichment. Bioinformatics, 2007. 23(3): p. 306-13.
58. Carro, M.S., et al., The transcriptional network for mesenchymal transformation of brain tumours. Nature, 2010. 463(7279): p. 318-U68.
59. 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.
60. 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.
61. Bennett, A.O.M., Stress and anxiety in schizophrenia and depression: glucocorticoids, corticotropin-

bioRxiv preprint doi: https://doi.org/10.1101/133363; this version posted May 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

releasing hormone and synapse regression. Aust N Z J Psychiatry, 2008. 42(12): p. 995-1002.
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, 2009. 14(12): p. 1083-94.
63. 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.
64. 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.
65. Blake, D.J., et al., TCF4, schizophrenia, and Pitt-Hopkins Syndrome. Schizophr Bull, 2010. 36(3): p. 443-7.
66. 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.
67. 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.
68. Wang, J., et al., WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res, 2013. 41(Web Server issue): p. W77-83.
69. 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.
70. 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.
71. 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.
72. Ardlie, K.G., et al., The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science, 2015. 348(6235): p. 648-660.
73. Zhang, B., et al., Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell, 2013. 153(3): p. 707-20.
74. Colantuoni, C., et al., Temporal dynamics and genetic control of transcription in the human prefrontal cortex. Nature, 2011. 478(7370): p. 519-23.
75. Gibbs, J.R., et al., Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet, 2010. 6(5): p. e1000952.
76. Ramasamy, A., et al., Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat Neurosci, 2014. 17(10): p. 1418-28.
77. Sloan, C.A., et al., ENCODE data at the ENCODE portal. Nucleic Acids Research, 2016. 44(D1): p. D726-D732.
78. Wang, K., M. Li, and H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res, 2010. 38(16): p. e164.
79. Kim, S.Y. and D.J. Volsky, PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics, 2005. 6: p. 144.
80. Tian, L., et al., Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci U S A, 2005. 102(38): p. 13544-9.

bioRxiv preprint doi: https://doi.org/10.1101/133363; this version posted May 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

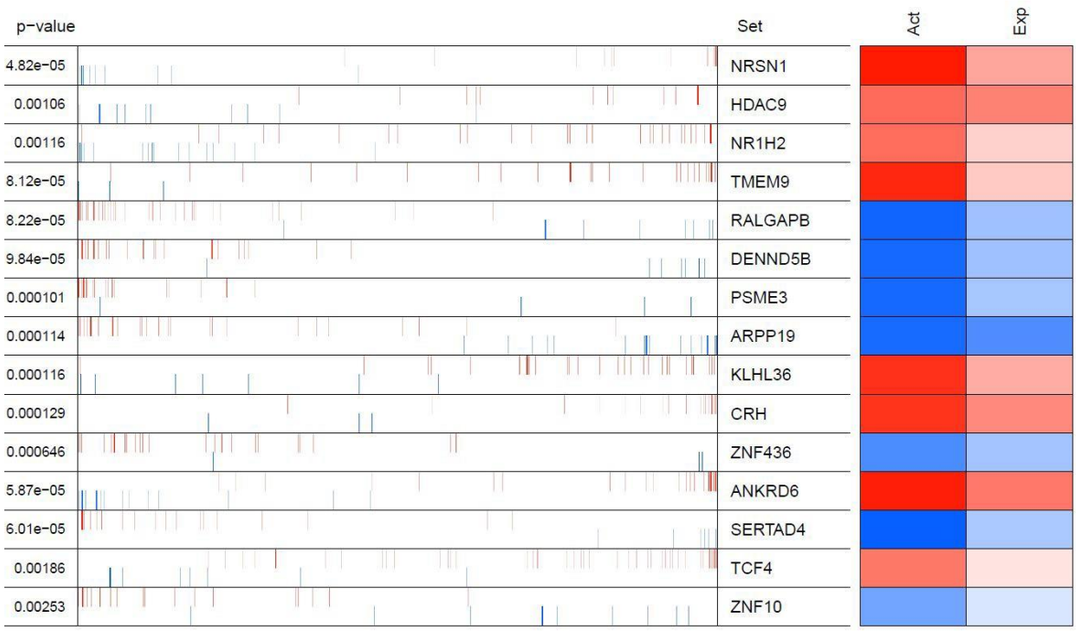


Figure 1. Virtual Inference of Protein-activity by Enriched Regulon analysis.

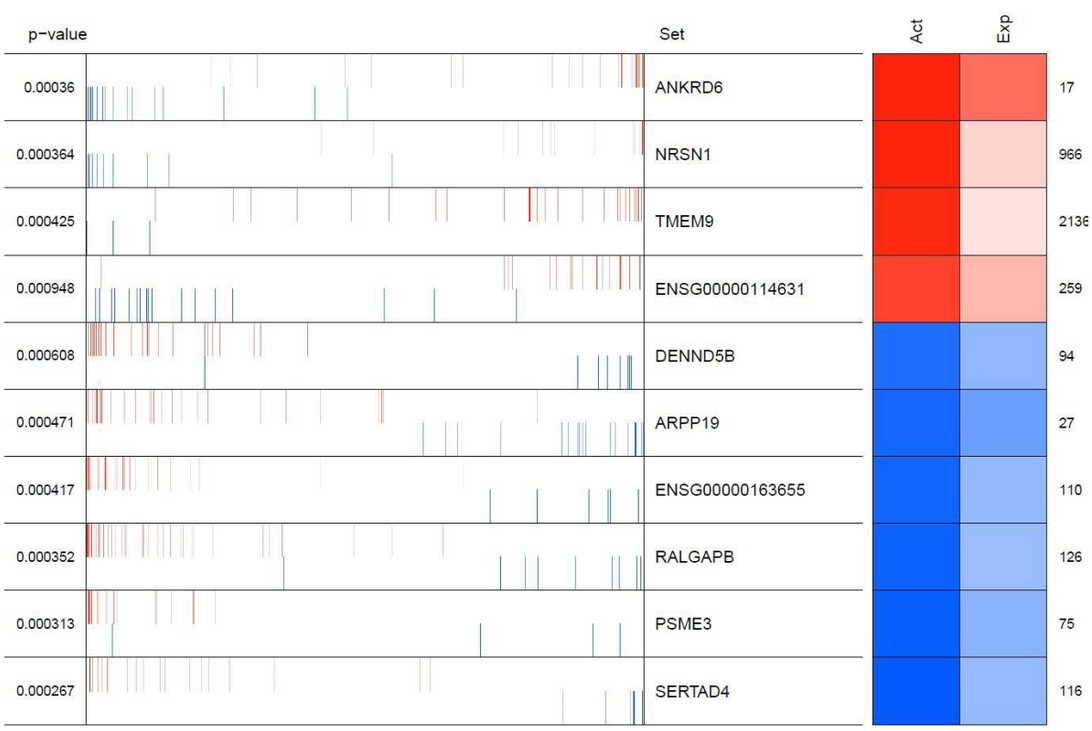
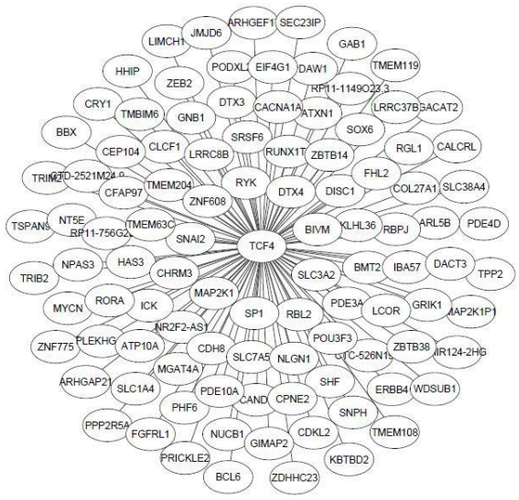
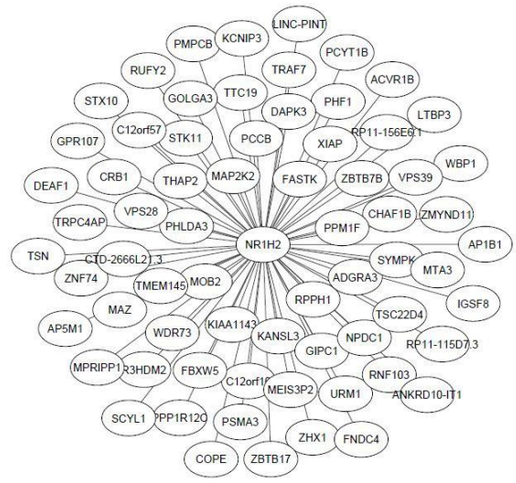


Figure 2. Representation of the enrichment of co-regulons on the gene expression signatures.

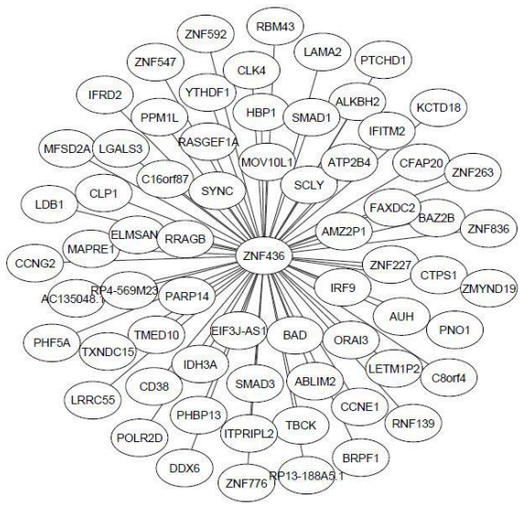
bioRxiv preprint doi: https://doi.org/10.1101/133363; this version posted May 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



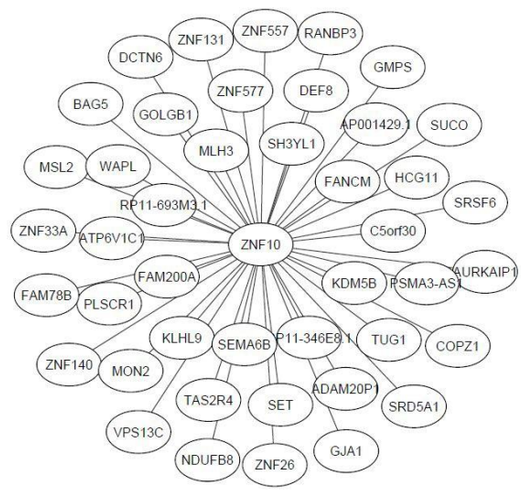
(a) TCF4 target genes.



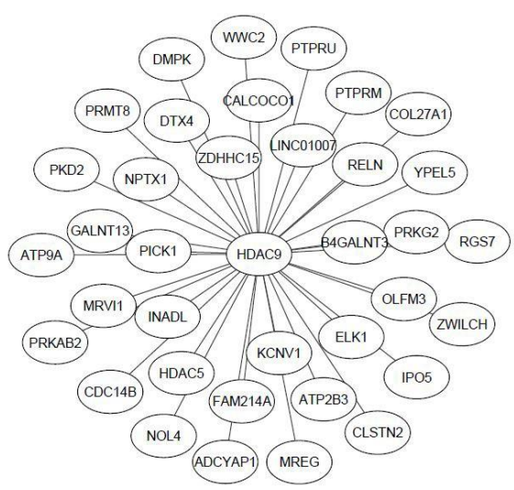
(b) NR1H2 target genes.



(c) ZNF436 target genes.



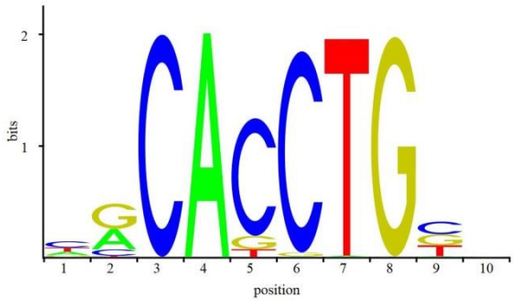
(d) ZNF10 target genes.



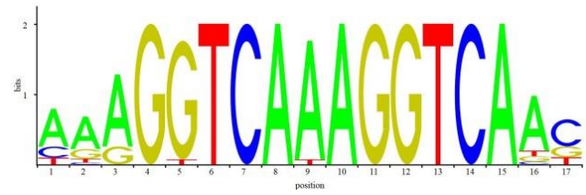
(e) HDAC9 target genes.

Figure 3. Target genes of the five identified master regulators.

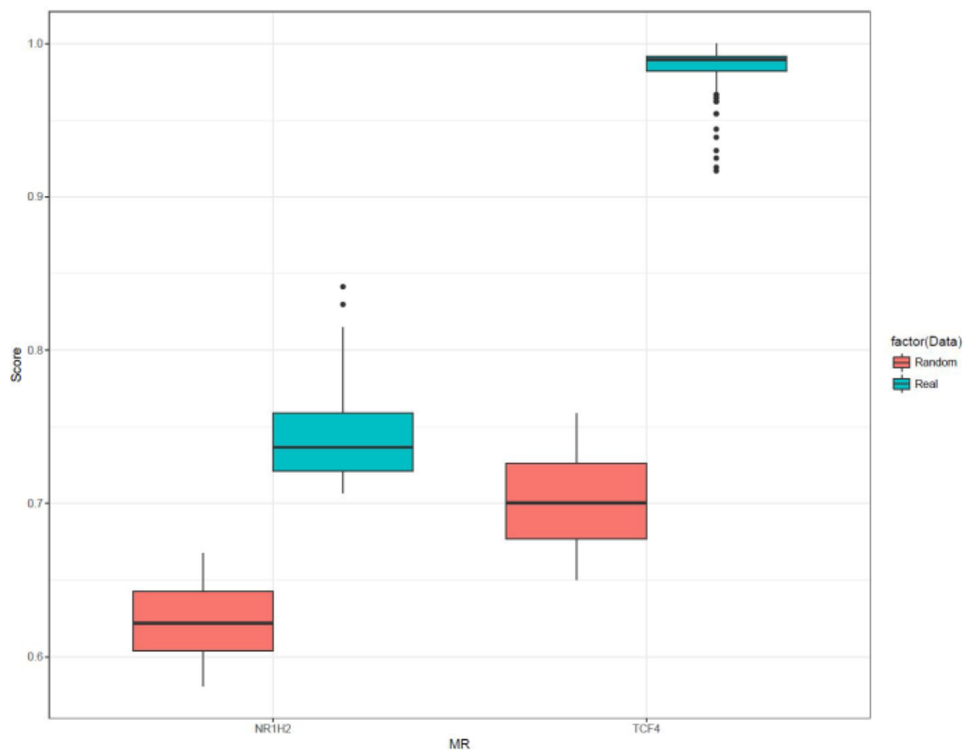
bioRxiv preprint doi: https://doi.org/10.1101/133363; this version posted May 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



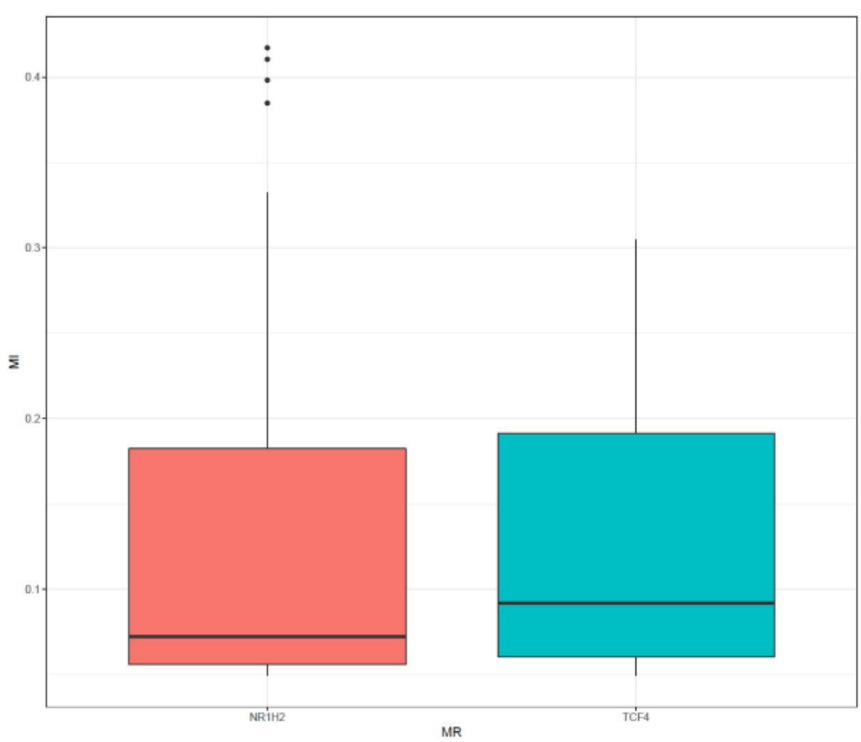
(a) TCF4 motif.



(b) NR1H2 motif.



(c) Transcription binding enrichment scores.



(d) Mutual information values

Figure 4. Transcription binding enrichment analysis of TCF4 and NR1H2.

bioRxiv preprint doi: https://doi.org/10.1101/133363; this version posted May 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

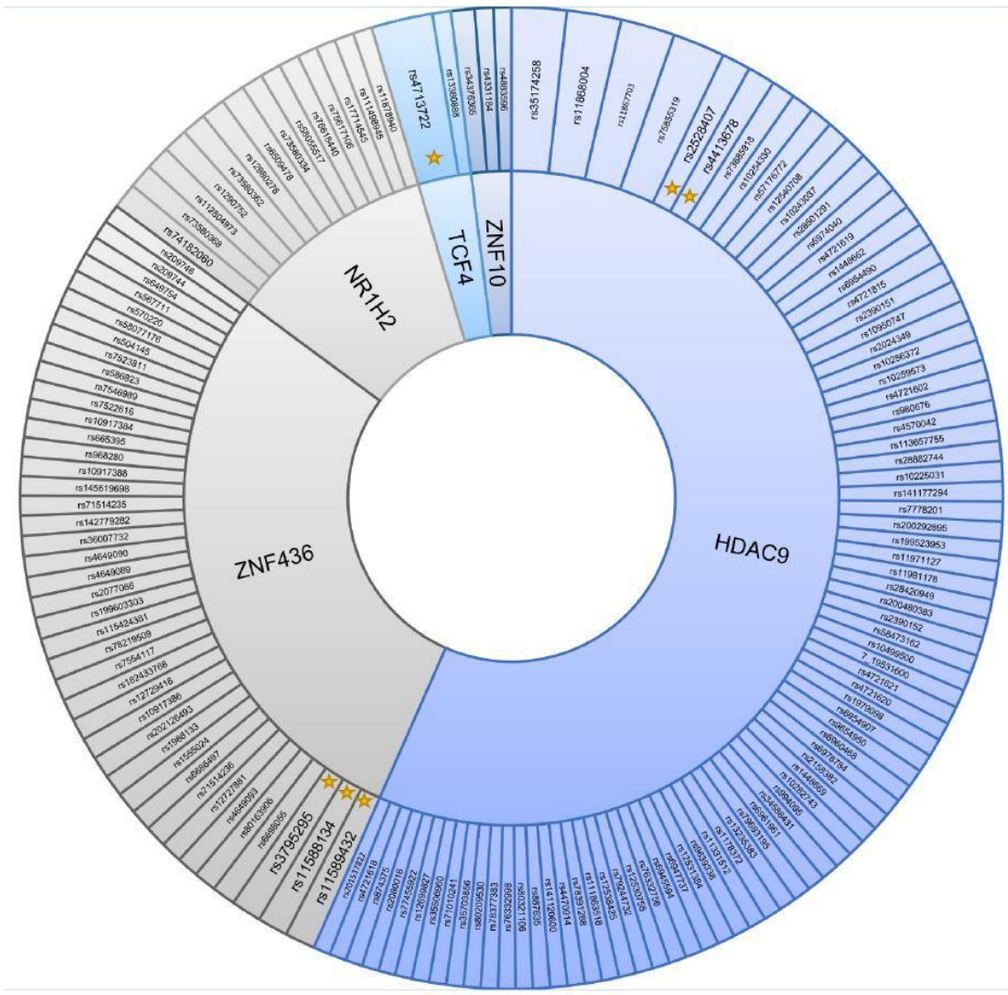


Figure 5. SNP-MR associations.

bioRxiv preprint doi: https://doi.org/10.1101/133363; this version posted May 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1. Topology of the MRs in the extracted networks combining the entire case and control samples

Gene	***#Targets***	***P-value***	***Gene***	***#Targets***	***P-value***
NRSN1	25	9.5e-6	ZNF10	29	0.004276522
ANKRD6	43	6.29e-8	TCF4	54	0.01824
TMEM	29	0.00078	NR1H2	46	0.00149
KLHL36	36	2.56e-5	HDAC9	26	1.93E-07
CRH	27	5.04e-7	ZNF436	28	1.67E-05
ARPP19	49	4.22e-9	SERTAD4	26	4.75e-5
PSME3	28	2.43e-5	RALGAPB	48	1.45e-5
DENND5B	32	1.23e-5			

Table 2. Validated TCF4 targets based on ATAC-seq experiments.

CDKL2	TSPAN9	CALCRL	SNAI2	NPAS3
RUNX1T1	RYK	MAP2K1	MGAT4A	GAB1
SOX6	PDE4D	CHRM3	CDH8	BMT2
RBPJ	GRIK1	ZBTB38	ERBB4	DACT3

Table 3. Validated TCF4 targets using TCF4 ENCODE ChIP-seq data from HEK293 cell line.

HHIP	PRICKL	SLC3A	RGL1	COL2	TSPAN9	SNAI
NPAS3	FHL2	WDSU	DISC	RYK	NR2F2-	ZNF
CRY1	LIMCH1	PPP2R	ROR	JMJD6	MGAT4	GNB
ARHGAP21	GAB1	TRIM2	SOX6	PDE1	PDE4D	BCL
SRSF6	ATXN1	CHRM	TMBI	CDH8	PHF6	CFA
RBPJ	ZNF608	ZEB2	NLGN	KBTB	GRIK1	PDE
ZBTB38	ERBB4	LCOR	POU3F3			