Building a schizophrenia genetic network:

Evidence that Transcription Factor 4 regulates schizophrenia risk genes

Hanzhang Xia[1], Fay M. Jahr[2], Nak-Kyeong Kim[3], Linying Xie[1], Andrey A. Shabalin[4], Julien Bryois[5], Douglas H. Sweet[6], Mohamad M. Kronfol[2], Preetha Palasuberniam[2], MaryPeace McRae[2], Brien P. Riley[6], Patrick F. Sullivan[5,8], Edwin J. van den Oord[1], Joseph L. McClay[2*].

[1] Center for Biomarker Research and Precision Medicine, Virginia Commonwealth University, Richmond, Virginia, United States.

[2] Department of Pharmacotherapy and Outcomes Science, Virginia Commonwealth University, Richmond, Virginia, United States.

[3] Department of Biostatistics, Virginia Commonwealth University, Richmond, Virginia, United States,

[4] Department of Psychiatry, University of Utah, Salt Lake City, Utah, United States

[5] Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden.

[6] Department of Pharmaceutics, Virginia Commonwealth University, Richmond, Virginia, United States.

[7] Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, Virginia, United States

[8] Departments of Genetics and Psychiatry, University of North Carolina School of Medicine, Chapel Hill, North Carolina, United States.

* Correspondence to: Dr. Joseph L. McClay, Department of Pharmacotherapy and Outcomes Science, Virginia Commonwealth University School of Pharmacy, Medical College of Virginia campus, 410 N 12th St, Richmond, VA 23298-0533, USA.

ABSTRACT

Background: The transcription factor 4 (*TCF4*) locus is one of the most robust association findings with schizophrenia (SZ). However, little is known about the genes regulated by the encoded transcription factor.

Methods: We conduct chromatin immunoprecipitation sequencing (ChIP-seq) of TCF4 in neural-derived (SH-SY5Y) cells to identify genome-wide TCF4 binding sites, followed by data integration with functional experiments and SZ association findings.

Results: We identified 11,322 TCF4 binding sites present in two ChIP-seq experiments. These binding sites implicated a gene set that was significantly enriched in 1) Gene Ontology categories including axon/neuronal development and insulin signaling, 2) genes preferentially expressed in brain, in particular pyramidal neurons of the somatosensory cortex, 3) genes down-regulated in TCF4 siRNA knockdown experiments, 4) genes down-regulated in post-mortem brain tissue from SZ patients (OR=2.8, permutation $p<4x10^{-5}$). Considering genomic alignments, TCF4 binding sites significantly overlapped those for neural DNA binding proteins such as FOXP2 and the SZ-associated EP300. TCF4 binding sites were modestly enriched among SZ risk loci from the Psychiatric Genomic Consortium (OR=1.56, p=0.03). In total, 130 TCF4 binding sites occurred in 39 of the 108 regions published in 2014. Thirteen genes within the 108 loci had both a TCF4 binding site ±10kb and were differentially expressed in siRNA knockdown experiments of TCF4, suggesting direct TCF4 regulation.

Conclusions: These findings confirm TCF4 as an important regulator of neural genes and point towards functional interactions with potential relevance for SZ.

Abstract word count = 232

INTRODUCTION

Large-scale genome-wide assocation studies (GWAS) have converged on specific risk loci for schizophrenia (SZ)[1]. One of the most robust findings is the *TCF4* (transcription factor 4) region on chromosome 18q21.2[2]. First discovered in GWAS meta-analysis[3], the finding remained significant in a follow-up study[4] and a large family-based replication study[5]. Most pertinently, single nucleotide polymorphisms (SNPs) at *TCF4* were among the top findings (p=3.34 x 10^{-12}) in the 2014 Psychiatric Genomics Consortium (PGC) mega-analysis of SZ[1]. Congruent with the association with SZ, *TCF4* has also been associated with SZ endophenotypes such as neurocognition and sensorimotor gating[6–8].

The biology of *TCF4* suggests a plausible role in CNS disorders: 1) *TCF4* encodes a transcription factor abundantly expressed in brain that has been implicated in neuronal development[9] and function[10,11]. 2) Mutations at *TCF4* causes Pitt-Hopkins Syndrome (PHS), a rare genetic disorder characterized by neurological deficits including mental retardation[2,12–14]. 3) Balanced chromosomal rearrangements in patients with neurodevelopmental disorders have encompassed *TCF4*[15]. 4) *TCF4* is a target for transcriptional regulation by microRNA 137 (gene ID: *MIR137*)[16], which is also a top association finding for SZ[1]. 5) Transgenic mice that overexpress *TCF4* have cognitive and sensorimotor impairments[17], which mirror deficits observed in SZ patients. Overall, the biological rationale linking *TCF4* to the CNS and SZ is compelling[2], suggesting that further study of this gene could advance our understanding of SZ pathogenesis.

The protein encoded by *TCF4* is a basic helix-loop-helix (bHLH) transcription factor (TF) that recognizes an Ephrussi-box ('E-box') binding site ('CANNTG')[2,9,18]. However, this motif is too small and non-specific to accurately predict *TCF4* binding computationally. A precise map of binding sites is vital for deciphering the gene regulatory networks under the influence of a TF[19]. In recent years, systematic mapping of TF binding has been enabled via chromatin immunoprecipitation coupled to next-generation sequencing (ChIP-seq)[20]. ChIP-seq works by precipitating the desired protein-DNA complex out of cell lysate using an antibody complementary to the protein of interest. After removal of the protein, the

liberated DNA fragments are sequenced and mapped back to the reference genome to yield a map of regions bound by the protein[21] (Figure 1a).

The ENCODE Consortium, which aims to map all regulatory elements in the human genome, has used ChIP-seq extensively to map binding profiles of many TFs in human cell lines[22]. TF binding is tissue-specific, so separate experiments are required to characterize binding in cells from each tissue of interest. At its outset, ENCODE did not have a major CNS focus[23]. The consortium attempted to map TCF4 binding sites in bone marrow-derived K562 cells, (https://www.encodeproject.org/experiments/ENCSR000FCF/). However, these data were revoked shortly after release. To redress the lack of information on TCF4 binding sites, here we describe TCF4 ChIP-seq in a CNS-derived cell line. To probe the relationship with SZ, we tested the TCF4 gene network for overlap with SZ risk genes from GWAS and gene expression studies.

MATERIALS AND METHODS

***TCF4* nomenclature** Transcription Factor 4, located on chromosome 18, was previously known by the aliases *E2-2*, *ITF-2*, *PTHS*, and *SEF-2* (http://www.ncbi.nlm.nih.gov/gene/6925). *TCF4* and *TCF7L2* (Gene ID: 6934) are often confused because they share the *TCF4* alias[9, 25]. *TCF7L2*, located on chromosome 10, encodes "T-Cell Factor 4", an effector of β-catenin signaling[36]. However, it has not been linked to SZ. In this study, all references to *TCF4* are to the gene with co-ordinates chr18:52889562-53332018 (hg19) and these identifiers: HUGO name=*TCF4*, ENTREZ=6925, HGNC=11634, ENSEMBL=ENSG00000196628, and UNIPROT=P15884.

Antibody selection At the start of this study, no ChIP-grade antibodies were available for TCF4, so we selected candidate antibodies from commercial vendors. Due to the confusion in nomenclature with T-Cell Factor 4, we discovered several mislabeled antibodies. Therefore, we limited ourselves to antibodies with published epitope sequences that could be confirmed as TCF4 via protein BLAST (NCBI). Eight antibodies were selected, of which three passed initial QC and were used for ChIP-seq: polyclonal anti-TCF4 antibodies K-12 (sc-48947) and N-16 (sc-48949) from Santa Cruz Biotechnology (Dallas, TX) and monoclonal anti-TCF4 antibody 1G4 (Novus, Littleton, CO).

Cell culture SH-SY5Y cells (ATCC, Manassas, VA) were cultured according to supplier's standard protocols.

Antibody validation We closely followed published ENCODE guidelines for antibody validation[24]. In addition to Western blotting of SH-SY5Y cell lysates, we used immunoprecipitation (IP) of TCF4 protein followed by Western blotting or protein mass spectrometry to characterize the proteins captured in IP (see Supplementary Methods).

Chromatin Immunoprecipitation sequencing (ChIP-seq) Each ChIP assay used approximately $1.2x10^7$ SH-SY5Y cells and was performed with the SOLiD ChIP-Seq Kit (Life Technologies) according to manufacturer's specifications, with some adjustments (Supplementary Methods). ChIP-Seq libraries were validated using the BioAnalyzer high-sensitivity chip assay (Agilent) prior to multiplexed high throughput sequencing on the SOLiD 5500 platform. 50 bp single end reads were generated, with a target read number of 25 million tags per sample. In addition to ChIP samples using anti-TCF4 antibodies, their respective IgG controls and input DNA controls were sequenced.

ChIP-seq data analysis Reads were aligned to the human genome (build hg19) using BioScope 1.2 (Life Technologies). Multi-mapping reads were discarded and only stringent single alignments retained. Sample files were output as .bam files using BioScope. PCR duplicates were removed and alignment files were written in tagAlign.gz format using Samtools[25] and Bedtools[26]. To call peaks and assess experiment quality, we used the SPP peak caller[21] distributed with phantompeakqualtools[27] (https://code.google.com/archive/p/phantompeakqualtools/). A false discovery rate (FDR) threshold of 1%, as implemented in SPP, was used to call peaks. Any peaks that mapped to hg19 ENCODE blacklisted regions (ftp://encodeftp.cse.ucsc.edu/users/akundaje/rawdata/blacklists/hg19/ wgEncodeHg19ConsensusSignalArtifactRegions.bed.gz) were removed.

Data Integration Bioinformatics analysis was conducted in R (www.r-project.org). Gene lists were obtained from refGene via UCSC Genome Browser download (August 10th 2017), followed by elimination of transcripts with ambiguous mapping and pruning entries by maximum boundary to yield a single non-redundant locus per gene. Gene pathway analysis used GREAT (Genomic Regions Enrichment of Annotations Tool)[28] version 3.0.0, assigning proximal regulatory domains ± 10 kb. Tissue-specific expression of the top TCF4 genes was tested using the "Gene2Func" mode in FUMA (Functional Mapping and Annotation)[29].

For cell-specific expression analysis, we obtained single cell RNA-seq data from five brain regions in mice (9970 single cells) that were previously clustered into 24 different cell types[30]. Normalization factors were computed for each of the 9970 single cells using the scran R package[31,32] using the 50% of the genes with mean expression higher than the median. The normalization factors were computed after clustering cells using the scran quickcluster() function to account for cell type heterogeneity. We then performed 24 differential expression analyses using BPSC[33] testing each cell type against the 23 other cell types using the normalization factor as a covariate. For each differential expression analysis, the t-statistics were then transformed to a standard normal distribution. Finally, for each cell type, we used linear regression to test if the standard normalized t-statistics for genes in the TCF4 gene set were significantly higher than for genes not in the gene set.

Enrichment testing of TCF4 binding sites in significant genes from siRNA knockdown expression studies[34,35] was carried out by mapping genes ± 10 kb, followed by one-sided Fisher exact tests. Permutation testing of significant findings used the shiftR package (https://github.com/andreyshabalin/shiftR), as outlined previously[36]. Test of overlap between TCF4 peaks and genomic annotations used the LOLA (Locus Overlap Analysis) Bioconductor package[37], with all mappings obtained from the LOLACore annotation set (databio.org/regiondb). The background set for this analysis was the default DNase hypersensitive sites (DHS) in multiple tissues ("activeDHS" set), which captures known human regulatory regions[37,38]. Testing for overlap with Psychiatric Genomics Consortium SZ findings was also based on genomic locus, rather than gene. SZ-associated loci were obtained by download of the "scz2.anneal.108" file from the PGC website (https://www.med.unc.edu/pgc/results-and-downloads). Background sets for this analysis were either "activeDHS" as above or DNase hypersensitive sites specifically for SK-N-SH cells ("wgEncodeOpenChromDnaseSknshPk" track, Duke DHS from ENCODE). No DHS data for SH-SY5Y were available, but SH-SY5Y are a subline of SK-N-SH isolated from the same donor[39]. After matching to the background sets, enrichment testing used Fisher exact tests followed by permutation as above.

RESULTS

ChIP-seq and peak calling

At the start of the project, no validated ChIP antibodies were available for TCF4. Of eight candidate anti-TCF4 antibodies identified, three passed initial immunoblot testing according to ENCODE guidelines, whereby a single band of the appropriate mass accounting for >50% of the total lane intensity was observed[24] (Figure 1b). We performed IP cross-reactivity studies using these antibodies, where IP with anti-TCF4 antibodies was used as the substrate for Western blot with a different anti-TCF4 antibody. Figure 1c shows that anti-TCF4 antibodies K-12 and N-16 immunoprecipitated a protein of the appropriate mass that was detected by monoclonal anti-TCF4 antibody 1G4, indicating that all three antibodies were likely detecting the same protein. We also conducted mass spectrometry-based proteomics analysis of the IP of all three antibodies and successfully detected TCF4 peptides in each case (Supplementary Figures S1 and S2), while we did not detect any in IgG control (mock IP) experiments. We therefore proceeded with ChIP-seq using these antibodies.

Each ChIP experiment involved isolating TCF4-bound DNA from approximately 1.2×10^7 SH-SY5Y cells. Average DNA yield was within expected parameters (~10 ng per replicate), but DNA was also recovered from ChIP using the control IgG antibodies. This indicated that non-specific material was captured and that mock IP ChIP-seq was the appropriate background experiment to call peaks. We conducted each ChIP in triplicate and met/exceeded ENCODE guidelines for a minimum of 10^7 uniquely-aligned reads per ChIP replicate[24]. However, only one replicate each for two of the three antibodies (K-12 and N-16) showed good enrichment. Figure 1d shows the summary statistics for these assays and cross-correlation plots are provided in Supplementary Figures S3 and S4.

After filtering out ENCODE blacklisted regions, the number of peaks called in each of the two successful experiments was 77,167 for antibody K-12 and 70,551 for antibody N-16, using a stringent False Discovery Rate (FDR) threshold of 1% (Figure 1d). This large number of peaks is typical in ChIP-seq experiments, where peaks can number in the tens or hundreds of thousands[40,24].

Peak consistency and genomic distribution

We examined overlap between peaks called in our two experiments. First, we plotted the difference in starting position between peaks observed in K-12 ChIP-seq and the closest peak in N-16 ChIP-seq. Figure 2a shows that there was a clear enrichment for peaks with start sites ± 200 bp in each experiment. This is the approximate peak size (~225 bp) in the 'narrowpeak' output format used by the SPP peak caller. Therefore, we specified that the peaks called in the two separate experiments had to overlap (at least 1bp in common) to count as "replicated"; neighboring, non-overlapping peaks were not considered. Applying this criterion resulted in 11,322 TCF4 binding sites present in both experiments. As expected, overlap was greater for peaks called with higher confidence. Sorting SPP signalValue, 65% of peaks in the top 100 in the K-12 experiment had a matching peak in the N-16 data; 50% in the top 250; 45% in the top 500; and 29% in the top 5000.

TCF4 binds the 'Ebox' sequence motif ('CANNTG'). Overall, 86.4% of the 11,322 binding sites encompassed an Ebox. There was a marginal enrichment among the top findings, whereby 90% of the top 500 sites encompassed an Ebox compared to 85.6% for the bottom 500. However, Ebox motifs are small, redundant and occur frequently in the genome (>14 million in hg19). Therefore they are a non-specific indicator of ChIP-seq success.

Figure 2b shows the distribution of all 11,322 binding sites by chromosome. No binding sites were obtained for the chrY because SH-SY5Y is genetically female. If the binding sites were random, most would occur on Chr1, the largest chromosome, with decreasing numbers to Chr21, the smallest. However, we observe a distinctly nonrandom distribution with relatively large numbers of peaks detected on chrs 7 and 17. The full set of 11,322 binding sites are provided in Supplementary Table S1.

Genes associated with TCF4 binding sites and their tissue-specific expression

To obtain insight into TCF4's functional role, we used GREAT[28] to conduct enrichment tests for GO categories and MSigDB pathways. GREAT first assigns genomic loci to regulatory domains associated with genes. Only 34 TCF4 binding sites could not be assigned to any genes (Supplementary

Figure S5). The results from GREAT are shown in Table 1. Significant results in the GO cellular component category had a strong neuronal theme, while insulin signaling and axon/neuronal development were notable pathway findings.

The full set of 11,322 TCF4 binding sites implicated 6528 unique genes ±10 kb of the gene body (Supplementary Table S2), which is approximately one quarter of all genes in refGene, including non-coding RNAs and genes with provisional nomenclature designations. We used FUMA[29] to test for enrichment in tissue-specific differentially expressed gene sets (Supplementary Figure S6). The TCF4 gene set was most enriched in genes up-regulated in the brain and pituitary, and genes down-regulated in the heart and blood vessels. To further probe the expression patterns of the TCF4 gene set, we looked at single cell RNA-seq data for specific CNS cell types. We observed that the TCF4 gene set was significantly over-expressed in pyramidal neurons from the somatosensory cortex ($P = 5.2 \times 10^{-5}$, Figure 3).

Differential regulation of TCF4 genes in siRNA knockdown experiments and post-mortem data

We integrated our ChIP-seq data with relevant datasets to refine our findings and identify SZ risk genes under TCF4 control. We first tested for overlap between the TCF4 gene set and gene expression data from a *TCF4* knockdown experiment using small interfering RNA (siRNA), also conducted in SH-SY5Y cells[34]. We expected that genes differentially expressed following *TCF4* knockdown should be enriched for those with TCF4 binding sites from ChIP-seq. A significant enrichment was observed and this was driven by genes downregulated following TCF4 knockdown (Table 2). No significant enrichment was observed in the upregulated genes. We also tested the TCF4 gene set for enrichment in a second TCF4 siRNA knockdown experiment, this time in cortical neuron progenitor cells[35] (Table 2). Once again downregulated genes, but not upregulated genes, were significantly enriched. This consistency was encouraging. All genes overlapping between ChIP-seq and siRNA studies are provided in Supplementary Tables S4 and S5.

We next tested if the TCF4 gene set was enriched among genes differentially expressed in post-mortem brain tissue from SZ patients. Here we used findings from Fromer et al. (2016)[41], who conducted bulk RNA-seq of dorsolateral prefrontal cortex (DLPFC) from 258 SZ cases and 279 controls. Once again, downregulated genes were significantly enriched with almost half possessing a TCF4 binding site, compared to just 12.5% of the up-regulated genes. This highly significant enrichment in down-regulated genes was surprising because these are downregulated in SZ, not as a direct result of *TCF4* knockdown. We hypothesized that *TCF4* is driving the down-regulation of genes in SZ by being downregulated itself, analogous to the situation in the siRNA experiments. However, *TCF4* expression was significantly *upregulated* in the SZ patients (fold enrichment = 1.16, Fromer et. al, their Supplementary Data File 3[41]). Further analysis of these gene sets showed that few genes were shared in common. That is, only seven genes with TCF4 binding sites from ChIP-seq were down-regulated in the SH-SY5Y siRNA study[34] *and* down-regulated in the SZ expression study[41]. These were *ANKMY1*, *FAM78A*, *IGF2*, *MXRA8*, *NT5M*, *PELI3* and *TNS3*. Notably, *IGF2* had the largest fold-change reduction of all genes in the Fromer et al. study. A further two genes with TCF4 binding sites, *DBNL* and *PNPLA7*, showed downregulation in the neural progenitor cell siRNA study[35] *and* the SZ gene expression study.

Overlap of TCF4 binding sites with other TFs and SZ risk loci

We next considered overlap of TCF4 binding sites with UCSC genome browser features. TCF4 binding sites were highly enriched (OR=7.08) in CpG islands, classified according to the Weizmann Institute CpG evolution model[42], and in transcription start sites (OR=3.19) from the SwitchGear Genomics library (www.switchgeargenomics.com) (Table 2). These are consistent with the binding profile of a TF. We also looked at overlap with binding profiles for other TFs from ENCODE. Notable overlapping factors were FOXP2, a neural transcription factor involved in speech development[43], and p300, a chromatin remodeler encoded by the *EP300* gene that is associated with SZ[1].

Our final data integration analysis was with the 108 PGC2 SZ risk loci published in 2014[1]. For this analysis, we used genomic locus-based enrichment rather than gene sets. This is important because many SZ-associated genes are large and may be more likely to overlap genomic annotations by chance than a randomly selected gene set[44]. Therefore, annotations can be confounded with gene size, leading to erroneous conclusions of enrichment. Testing for TCF4 binding site enrichment in the associated genomic regions, and not considering occurrence in genes, obviates this potential bias. Overall, 39 of the 108 PGC SZ loci contained one or more TCF4 binding sites, with 130 sites in total falling within their boundaries. Permutation testing of enrichment using all known human regulatory regions as the background set revealed a non-significant overlap (p=0.082). Narrowing the background set to relevant regulatory regions for the cell type used in the ChIP-seq experiments revealed a nominally significant enrichment (p=0.035) (Table 1). PGC SZ genes with TCF4 binding sites ±10 kb that were also differentially expressed in either of the siRNA experiments described above were *APH1A*, *C1orf54*, *CENPM*, *CHRNA5*, *DFNA5*, *GFOD2*, *GRAMD1B*, *LRP1*, *MPP6*, *PDCD11*, *SEZ6L2*, *TLE1*, and *XRCC3*. Of these, both *LRP1* and *XRCC3* had three binding sites in our ChIP-seq data, *PDCD11* had two, while the remainder had one each.

DISCUSSION

We obtained data on TCF4 binding from two separate ChIP-seq experiments, which allowed us to probe the TCF4 gene network for association with SZ risk genes. In our ChIP-seq experiments, only two out of three antibodies yielded usable data. The ENCODE Consortium attempted TCF4 ChIP-seq in K562 cells but revoked their data because of poor quality. Therefore, consistently obtaining high quality TCF4 ChIP-seq data may be a challenge. The simplest explanation may be that none of the antibodies so far used in TCF4 ChIP-seq are particularly well suited to this purpose. It is accepted that antibodies that pass initial characterization may still fail to yield good ChIP-seq data[24]. Yet between the ENCODE experiments and ours, at least four antibodies have been used in TCF4 ChIP-seq, with only two yielding acceptable data. We speculate that the complex nature of *TCF4* itself may contribute to this variability. The *TCF4* locus can give rise to 18 potentially distinct isoforms via alternative splicing[18]. Some isoforms are exclusively nuclear while others rely on heterodimerization partners. Furthermore, different isoforms may contain one or two transcription activation domains that act synergistically in combination[18]. It is therefore plausible that relative expression of different isoforms varies significantly over time, in conjunction with the expression of functional partner proteins. Further work will be needed to understand the impact of this complexity for SZ.

Our data was generated in the SH-SY5Y cell line, which is commonly used in CNS research. However, it is a transformed line that will have differences in gene expression from healthy neurons *in vivo* and it is well-known that TF binding is highly cell-specific[22,45]. Unfortunately, transformed cells are a practical necessity. The very low DNA yields in ChIP-seq mean that millions of cells are needed for each experiment. These numbers are extremely difficult to obtain for primary cells, particularly human neurons. Methods that yield high quality data from very low DNA input quantities are urgently needed in this area. Nevertheless, our data did indicate that genes implicated by our TCF4 binding sites were enriched in brain expressed gene sets. Furthermore, using single cell RNA-seq data, we observed enrichment of the TCF4 gene set among genes expressed in pyramidal neurons of the somatosensory cortex, a cell type that was previously associated with schizophrenia[30].

Several genes that we identified as having TCF4 binding sites have been functionally linked with *TCF4* in prior studies. For example, TCF4 has also been shown to affect neural excitability via repression of potassium channel *KCNQ1*[11]. In our data, this gene had 19 unique TCF4 binding sites, the 16th largest number for any gene. Conversely, given the problems with "TCF4" nomenclature, delineating what is *not* seen in our data may be of value. Several articles describe the interaction between "TCF4", β-catenin and p300[46]. This relates to T Cell Factor 4, encoded by *TCF7L2*. Beta-catenin is encoded by *CTNNB1* and we did not detect a TCF4 binding site at this gene. Furthermore, even though the binding profiles of p300 (encoded by *EP300*) and TCF4 strongly overlap (Table 2), TCF4 does not appear to regulate *EP300*, at least in our data. The co-occurrence of their binding sites does, however, imply that they may be involved in the regulation of a partially overlapping set of genes in CNS cells. *EP300* is a PGC2 SZ risk locus[1] and was associated with emotional processing in functional neuroimaging experiments[47]. Further analysis into the overlapping pathways regulated by these two SZ-associated DNA binding proteins may be relevant.

Several genes with TCF4 binding sites in our ChIP-seq data were significantly enriched among those down-regulated in *TCF4* siRNA knockdown experiments. However, the ORs of enrichment were fairly modest, even for siRNA knockdown that used the same cell line (SH-SY5Y) as our ChIP-seq experiments. This can be explained because not all genes differentially expressed in a TCF4 siRNA knockdown experiment will be directly under TCF4 control[35]. Those genes that TCF4 directly affects can, in turn, affect the expression of others. Similarly, ChIP-seq data can be noisy and the number of peaks can exceed the numbers of genes a TF directly regulates, possibly as the result of sporadic binding[40]. By identifying the overlap between ChIP-seq and siRNA data, this allows us to cancel out some of the noise inherent to each strategy. This should provide a better indicator of genes under direct TCF4 control, since it combines direct physical interaction of the TF with a functional effect on expression.

A subset of genes with TCF4 binding sites were differentially expressed in siRNA studies *and* dysregulated in post-mortem brain tissue from SZ patients. Among these, *IGF2* (insulin-like growth

factor 2) was the most down-regulated gene in postmortem SZ brain[41]. *IGF2* may regulate neural plasticity to modulate behavior and memory[50]. Furthermore, deficits in hippocampal neurogenesis in a mouse model of 22q11.2 deletion-associated SZ can be rescued by *IGF2*[51]. A significant amount of work has been conducted on the role of *IGF2* in the brain, yet few studies address the role of *IGF2* in SZ etiology. One issue is that *IGF2* was not detected among PGC SZ risk loci[1]. This may indicate that down-regulation of *IGF2* in SZ is a consequence of risk variants at other loci or environmental factors, rather than as a result of risk variants at the *IGF2* locus itself.

On the other hand, several PGC SZ risk genes did contain TCF4 binding sites, although the enrichment was modest. Thirteen such genes also showed differential expression in *TCF4* siRNA experiments. The risk loci identified by the PGC may span several hundred kb and contain many genes. It is often not apparent which genes in these regions should be selected for further study. *PDCD11* is likely regulated by TCF4, exhibiting both TCF4 binding sites and differential expression in a *TCF4* siRNA study. It is also expressed in brain[53] and located on chromosome 10, in the third most significant region identified by the PGC (chr10:104423800-105165583, P-value = 6.12E-19)[1]. There are 14 genes in this region and *PDCD11* is poorly characterized. However, its likely regulation by TCF4 may suggest further work is merited to characterize this interaction.

With robust risk loci implicated in SZ and the shift of focus to functional experiments and data integration, the onus is now on interpreting overlap between genome-wide profiles and identifying themes in complex datasets involving hundreds of relevant elements. Dissecting these complex interactions to tease out the causal networks relevant for psychiatric phenotypes is of paramount importance.

Conflict of Interest

The authors declare no financial conflicts of interest.

Acknowledgement

Thanks to Derek Blake at the MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, United Kingdom for providing siRNA gene expression data, and to Jens Hjerling-Leffler, Sten Linnarsson, Ana Munoz Manchado and Amit Zeisel of the Karolinska Institute, Stickholm, Sweden for providing single-cell RNA-sequencing data. This work was funded through grant R21 MH099419 from the US National Institute of Mental Health to JL McClay.

Author contributions

JM conceived of and designed the study. HX, FJ and LX conducted the main experiments. DS, MK, PP, MM and BR contributed to or advised on experimental procedures. PS and EO advised on aspects of experimental design and data integration. JM, NK, AS, JB and EO conducted the data analysis. All authors read and approved the manuscript.

REFERENCES

1. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).
2. Blake, D. J. *et al.* TCF4, schizophrenia, and Pitt-Hopkins Syndrome. *Schizophr Bull* **36,** 443–447 (2010).
3. Stefansson, H. *et al.* Common variants conferring risk of schizophrenia. *Nature* **460,** 744–747 (2009).
4. Steinberg, S. *et al.* Common variants at VRK2 and TCF4 conferring risk of schizophrenia. *Hum. Mol. Genet.* **20,** 4076–4081 (2011).
5. Aberg, K. A. *et al.* A comprehensive family-based replication study of schizophrenia genes. *JAMA Psychiatry* **70,** 573–581 (2013).
6. Lennertz, L. *et al.* Novel schizophrenia risk gene TCF4 influences verbal learning and memory functioning in schizophrenia patients. *Neuropsychobiology* **63,** 131–136 (2011).
7. Lennertz, L. *et al.* Impact of TCF4 on the genetics of schizophrenia. *Eur Arch Psychiatry Clin Neurosci* **261 Suppl 2,** S161-165 (2011).
8. Quednow, B. B. *et al.* The schizophrenia risk allele C of the TCF4 rs9960767 polymorphism disrupts sensorimotor gating in schizophrenia spectrum and healthy volunteers. *J. Neurosci.* **31,** 6684–6691 (2011).
9. Flora, A., Garcia, J. J., Thaller, C. & Zoghbi, H. Y. The E-protein Tcf4 interacts with Math1 to regulate differentiation of a specific subset of neuronal progenitors. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 15382–15387 (2007).

10. Kennedy, A. J. *et al.* Tcf4 Regulates Synaptic Plasticity, DNA Methylation, and Memory Function. *Cell Rep* **16,** 2666–2685 (2016).

11. Rannals, M. D. *et al.* Psychiatric Risk Gene Transcription Factor 4 Regulates Intrinsic Excitability of Prefrontal Neurons via Repression of SCN10a and KCNQ1. *Neuron* **90,** 43–55 (2016).

12. Brockschmidt, A. *et al.* Severe mental retardation with breathing abnormalities (Pitt-Hopkins syndrome) is caused by haploinsufficiency of the neuronal bHLH transcription factor TCF4. *Hum. Mol. Genet.* **16,** 1488–1494 (2007).

13. Zweier, C. *et al.* Haploinsufficiency of TCF4 causes syndromal mental retardation with intermittent hyperventilation (Pitt-Hopkins syndrome). *Am. J. Hum. Genet.* **80,** 994–1001 (2007).

14. Amiel, J. *et al.* Mutations in TCF4, encoding a class I basic helix-loop-helix transcription factor, are responsible for Pitt-Hopkins syndrome, a severe epileptic encephalopathy associated with autonomic dysfunction. *Am. J. Hum. Genet.* **80,** 988–993 (2007).

15. Talkowski, M. E. *et al.* Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell* **149,** 525–537 (2012).

16. Kwon, E., Wang, W. & Tsai, L.-H. Validation of schizophrenia-associated genes CSMD1, C10orf26, CACNA1C and TCF4 as miR-137 targets. *Mol. Psychiatry* **18,** 11–12 (2013).

17. Brzózka, M. M., Radyushkin, K., Wichert, S. P., Ehrenreich, H. & Rossner, M. J. Cognitive and sensorimotor gating impairments in transgenic mice overexpressing the schizophrenia susceptibility gene Tcf4 in the brain. *Biol. Psychiatry* **68,** 33–40 (2010).

18. Sepp, M., Kannike, K., Eesmaa, A., Urb, M. & Timmusk, T. Functional diversity of human basic helix-loop-helix transcription factor TCF4 isoforms generated by alternative 5’ exon usage and splicing. *PLoS ONE* **6,** e22138 (2011).

19. Farnham, P. J. Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.* **10,** 605–616 (2009).

20. Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10,** 669–680 (2009).

21. Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26,** 1351–1359 (2008).

22. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

23. Kavanagh, D. H., Dwyer, S., O'Donovan, M. C. & Owen, M. J. The ENCODE project: implications for psychiatric genetics. *Mol. Psychiatry* **18,** 540–542 (2013).

24. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22,** 1813–1831 (2012).

25. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

26. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

27. Marinov, G. K., Kundaje, A., Park, P. J. & Wold, B. J. Large-scale quality analysis of published ChIP-seq data. *G3 (Bethesda)* **4,** 209–223 (2014).

28. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28,** 495–501 (2010).

29. Watanabe, K., Taskesen, E., Bochoven, A. van & Posthuma, D. FUMA: Functional mapping and annotation of genetic associations. *bioRxiv* 110023 (2017). doi:10.1101/110023

30. Skene, N. G. *et al.* Genetic Identification Of Brain Cell Types Underlying Schizophrenia. *bioRxiv* 145466 (2017). doi:10.1101/145466

31. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* **5,** 2122 (2016).

32. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17,** 75 (2016).

33. Vu, T. N. *et al.* Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* **32,** 2128–2135 (2016).

34. Forrest, M. P., Waite, A. J., Martin-Rendon, E. & Blake, D. J. Knockdown of human TCF4 affects multiple signaling pathways involved in cell survival, epithelial to mesenchymal transition and neuronal differentiation. *PLoS ONE* **8,** e73169 (2013).

35. Hill, M. J. *et al.* Knockdown of the schizophrenia susceptibility gene TCF4 alters gene expression and proliferation of progenitor cells from the developing human neocortex. *J Psychiatry Neurosci* **42,** 181–188 (2017).

36. McClay, J. L. *et al.* High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction. *Genome Biol.* **16,** 291 (2015).

37. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32,** 587–589 (2016).

38. Sheffield, N. C. *et al.* Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res.* **23,** 777–788 (2013).

39. Kovalevich, J. & Langford, D. Considerations for the use of SH-SY5Y neuroblastoma cells in neurobiology. *Methods Mol. Biol.* **1078,** 9–21 (2013).

40. MacQuarrie, K. L., Fong, A. P., Morse, R. H. & Tapscott, S. J. Genome-wide transcription factor binding: beyond direct target regulation. *Trends Genet.* **27,** 141–148 (2011).

41. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19,** 1442–1453 (2016).

42. Cohen, N. M., Kenigsberg, E. & Tanay, A. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell* **145,** 773–786 (2011).

43. Graham, S. A. & Fisher, S. E. Decoding the genetics of speech and language. *Curr. Opin. Neurobiol.* **23,** 43–51 (2013).

44. Raychaudhuri, S. *et al.* Accurately Assessing the Risk of Schizophrenia Conferred by Rare Copy-Number Variation Affecting Genes with Brain Function. *PLOS Genetics* **6,** e1001097 (2010).

45. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489,** 91–100 (2012).

46. Lévy, L. *et al.* Acetylation of beta-catenin by p300 regulates beta-catenin-Tcf4 interaction. *Mol. Cell. Biol.* **24,** 3404–3414 (2004).

47. Erk, S. *et al.* Functional neuroimaging effects of recently discovered genetic risk loci for schizophrenia and polygenic risk profile in five RDoC subdomains. *Transl Psychiatry* **7,** e997 (2017).

48. Otto, S. J. *et al.* A new binding motif for the transcriptional repressor REST uncovers large gene networks devoted to neuronal functions. *J. Neurosci.* **27,** 6729–6739 (2007).

49. Ballas, N., Grunseich, C., Lu, D. D., Speh, J. C. & Mandel, G. REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis. *Cell* **121,** 645–657 (2005).

50. Fernandez, A. M. & Torres-Alemán, I. The many faces of insulin-like peptide signalling in the brain. *Nat. Rev. Neurosci.* **13,** 225–239 (2012).

51. Ouchi, Y. *et al.* Reduced adult hippocampal neurogenesis and working memory deficits in the Dgcr8-deficient mouse model of 22q11.2 deletion-associated schizophrenia can be rescued by IGF2. *J. Neurosci.* **33,** 9408–9419 (2013).

52. Quednow, B. B. *et al.* Schizophrenia risk polymorphisms in the TCF4 gene interact with smoking in the modulation of auditory sensory gating. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 6271–6276 (2012).

53. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell Proteomics* **13,** 397–406 (2014).

Table 1. Functional annotation and pathway analysis of TCF4 genomic binding sites using GREAT.

Term Name	Binom Rank	Binom Raw P-Value	B. FDR Q-Val	B. Fold Enrichment	B. Observed Region Hits	B. Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	H. Observed Gene Hits	H. Total Genes	H. Gene Set Coverage
GO Molecular Function												
14-3-3 protein binding	16	1.71E-31	3.95E-29	4.75	88	0.008	82	3.52E-02	2.06	13	16	0.002
protein kinase A catalytic subunit binding	92	1.18E-14	4.73E-13	2.82	75	0.007	53	2.44E-03	2.54	11	11	0.002
ankyrin binding	120	1.68E-12	5.17E-11	2.78	64	0.006	41	7.91E-04	2.27	17	19	0.002
GO Cellular Component												
cell cortex part	24	2.39E-33	1.26E-31	2.40	243	0.021	29	1.68E-04	1.57	63	102	0.009
axon terminus	29	3.20E-27	1.39E-25	2.31	212	0.019	26	8.05E-05	1.76	43	62	0.006
neuron projection terminus	39	7.87E-22	2.55E-20	2.05	218	0.019	24	6.54E-05	1.73	47	69	0.007
cortical cytoskeleton	71	6.04E-13	1.08E-11	2.09	117	0.010	39	7.79E-03	1.59	37	59	0.005
GO Biological Process												
tooth mineralization	69	2.07E-35	3.13E-33	5.27	90	0.008	442	5.36E-03	2.54	9	9	0.001
insulin receptor signaling pathway	79	4.36E-34	5.76E-32	2.05	349	0.031	601	2.60E-02	1.31	77	149	0.011
enamel mineralization	111	9.82E-30	9.23E-28	4.98	79	0.007	594	2.58E-02	2.54	7	7	0.001
regulation of cell size	171	2.90E-24	1.77E-22	2.00	260	0.023	531	1.46E-02	1.46	47	82	0.007
spinal cord dorsal/ventral patterning	180	2.74E-23	1.59E-21	2.97	115	0.010	670	4.00E-02	1.87	14	19	0.002
CD4-positive or CD8-positive, alpha-beta T cell lineage commitment	196	3.14E-22	1.67E-20	5.36	54	0.005	594	2.58E-02	2.54	7	7	0.001
spinal cord patterning	212	4.64E-21	2.29E-19	2.76	116	0.010	685	4.50E-02	1.81	15	21	0.002
regulation of axon extension	225	1.69E-20	7.85E-19	2.16	182	0.016	675	4.00E-02	1.56	27	44	0.004
negative regulation of osteoblast differentiation	233	6.95E-20	3.12E-18	2.43	137	0.012	489	8.67E-03	1.77	23	33	0.003
ventral spinal cord interneuron specification	406	1.13E-14	2.92E-13	2.84	74	0.007	598	2.56E-02	2.29	9	10	0.001
T cell selection	540	8.99E-12	1.74E-10	2.05	110	0.010	379	2.42E-03	1.88	23	31	0.003
positive T cell selection	549	1.37E-11	2.60E-10	2.50	71	0.006	638	3.47E-02	1.94	13	17	0.002
positive regulation of adenylate cyclase activity	576	2.69E-11	4.87E-10	2.08	102	0.009	492	9.45E-03	1.74	24	35	0.003
cellular response to estradiol stimulus	898	6.76E-08	7.86E-07	2.24	56	0.005	617	2.76E-02	2.03	12	15	0.002
cellular response to estrogen stimulus	1016	4.37E-07	4.49E-06	2.03	61	0.005	685	4.50E-02	1.81	15	21	0.002
positive regulation of neurotransmitter transport	1022	4.67E-07	4.77E-06	2.79	32	0.003	382	2.44E-03	2.54	10	10	0.001
prostaglandin biosynthetic process	1426	5.10E-05	3.74E-04	2.12	35	0.003	582	2.38E-02	1.90	15	20	0.002
positive regulation of neurotransmitter secretion	1504	8.97E-05	6.22E-04	2.50	23	0.002	594	2.58E-02	2.54	7	7	0.001
MSigDB Canonical Pathways												
Insulin signaling pathway	4	9.43E-30	3.11E-27	2.04	309	0.027	54	2.07E-02	1.34	73	138	0.010
Sonic Hedgehog (Shh) Pathway	6	7.70E-28	1.69E-25	3.74	101	0.009	52	1.99E-02	2.06	13	16	0.002
Genes involved in Class B/2 (Secretin family receptors)	9	6.90E-23	1.01E-20	2.16	203	0.018	39	1.59E-02	1.46	50	87	0.007
Vibrio cholerae infection	24	1.56E-17	8.56E-16	2.36	126	0.011	82	3.89E-02	1.50	32	54	0.005
LKB1 signaling events	29	2.51E-16	1.14E-14	2.44	109	0.010	23	3.62E-03	1.73	32	47	0.005
Regulation of RhoA activity	61	6.48E-11	1.40E-09	2.00	107	0.009	68	3.12E-02	1.57	29	47	0.004
p38 signaling mediated by MAPKAP kinases	88	3.11E-09	4.67E-08	2.58	51	0.005	49	1.82E-02	1.93	16	21	0.002
Ras activation uopn Ca2+ infux through NMDA receptor	93	5.53E-09	7.85E-08	2.33	60	0.005	26	5.23E-03	2.22	14	16	0.002
fl-arrestin-dependent Recruitment of Src Kinases in GPCR Signaling	96	7.81E-09	1.07E-07	2.89	40	0.004	70	3.08E-02	2.03	12	15	0.002
CREB phosphorylation through the activation of CaMKII	98	1.09E-08	1.47E-07	2.37	56	0.005	41	1.65E-02	2.18	12	14	0.002
Genes involved in Signalling to RAS	129	3.16E-07	3.23E-06	2.33	46	0.004	21	3.48E-03	1.98	21	27	0.003

Table 2. Integration of TCF4 binding profile with functional and SZ datasets.

Gene-based enrichment via permutation analysis

TCF4 knockdown experiments	Set size	Overlap	Odds Ratio	Fisher P-value	Permutation P-value
SH-SY5Y siRNA knockdown (all)	921	267	1.30	0.0003	0.0005
Upregulated	396	100	1.06	0.3159	n/a
Downregulated	525	167	1.48	3.28E-05	0.0001
Cortical precursor siRNA knockdown (all)	502	138	1.20	0.0434	0.0438
Upregulated	240	47	0.76	0.9603	n/a
Downregulated	262	91	1.68	6.78E-05	<4.11E-05*
SZ gene expression Fromer et. al (2016) (all)	623	190	1.38	0.0002	0.0001
Upregulated	296	37	0.45	1	n/a
Downregulated	326	153	2.77	2.20E-16	<4.11E-05*

Genomic locus enrichment tests using LOLA

UCSC annotations	bed file	overlap	Odds Ratio	P-value
Evo CpG islands	evoCpg.bed	3706	7.08	~0
CpG islands	cpgIslandExt.bed	960	2.76	2.30E-151
Switch DB TSSs	switchDbTss.bed	584	3.19	2.46E-118
LaminBq	laminB1.bed	4404	1.49	1.46E-94
Simple repeats	simpleRepeat.bed	836	1.99	3.46E-68

Top 15 overlapping DNA binding protein profiles from ENCODE

Experiment	antibody	overlap	Odds Ratio	P-value
ChIP SK-N-SH	NRSF	587	5.75	2.90E-231
ChIP GM12878	Pol2	637	2.26	5.54E-71
ChIP MCF-7	CTCF	561	2.26	3.53E-63
ChIP H1-hESC	CTCF	437	2.40	3.38E-56
ChIP K562	Max	580	2.09	3.11E-54
ChIP HEK293	ZNF263	508	2.18	4.30E-53
ChIP HepG2	Pol2	425	2.36	8.00E-53
ChIP PFSK-1	FOXP2	305	2.73	2.83E-50
ChIP K562	ZBTB7A	357	2.48	2.88E-49
ChIP GM12891	CTCF	422	2.26	5.42E-48
ChIP A549	CTCF	505	2.02	3.23E-44
ChIP SK-N-SH	p300	494	2.03	6.72E-44
ChIP K562	Egr-1	403	2.18	7.15E-43
ChIP K562	MAZ	450	2.04	8.06E-41
ChIP SH-SY5Y	GATA-2	413	2.09	9.19E-40

Psychiatric Genomics Consortium risk loci overlap

Data file	background	overlap	Odds Ratio	Fisher P-value	Permutation P-value
scz2.anneal.108	ActiveDHS[+]	158	1.38	8.88E-05	0.0816
	SK-N-SH DHS[+]	55	1.56	0.0017	0.035

* 4.11E-05 is the minimum P-value obtainable when permuting over 24,358 genes (includes non-coding RNAs and provisional IDs)

[+] DHS, DNaseI Hypersensitivity Sites

FIGURE LEGENDS

Figure 1. ChIP-seq outline and quality control

a) Flow diagram showing the ChIP-seq procedure in cultured cells.

b) Western blotting in SH-SY5Y cell extract using three different anti-TCF4 antibodies (1G4, K-12 and N-16) identifies a single clear band for TCF4.

c) Crossover IP for TCF4. In this experiment, anti-TCF4 antibodies K-12 and N-16, plus control IgG, were used to immunoprecipitate proteins that were subsequently probed in a Western blot using anti-TCF4 antibody 1G4. Using different anti-TCF4 antibodies in the IP and Western blotting increases confidence that TCF4, rather than a non-specific protein, is detected. The presence of the TCF4 band in the bound (IP) fraction, and not in the bound IgG (control) fraction, indicates that the K-12 and N-16 antibodies successfully immunoprecipitated TCF4.

d) ChIP-seq output summary. The number of uniquely aligning reads for each ChIP-seq experiment are shown. The Normalized Strand Cross-correlation coefficient (NSC) and Relative Strand Cross-correlation coefficient (RSC) values indicate enrichment in ChIP (Figure 1d), with higher values indicating more enrichment. NSC values less than 1.1 are relatively low and the minimum possible value is 1 (no enrichment). The minimum possible RSC value is 0 (no signal), highly enriched experiments have values greater than 1, and values much less than 1 may indicate low quality (see genome.ucsc.edu/ENCODE/qualityMetrics.html). The values for our experiments indicate moderately good enrichment.

Figure 2.

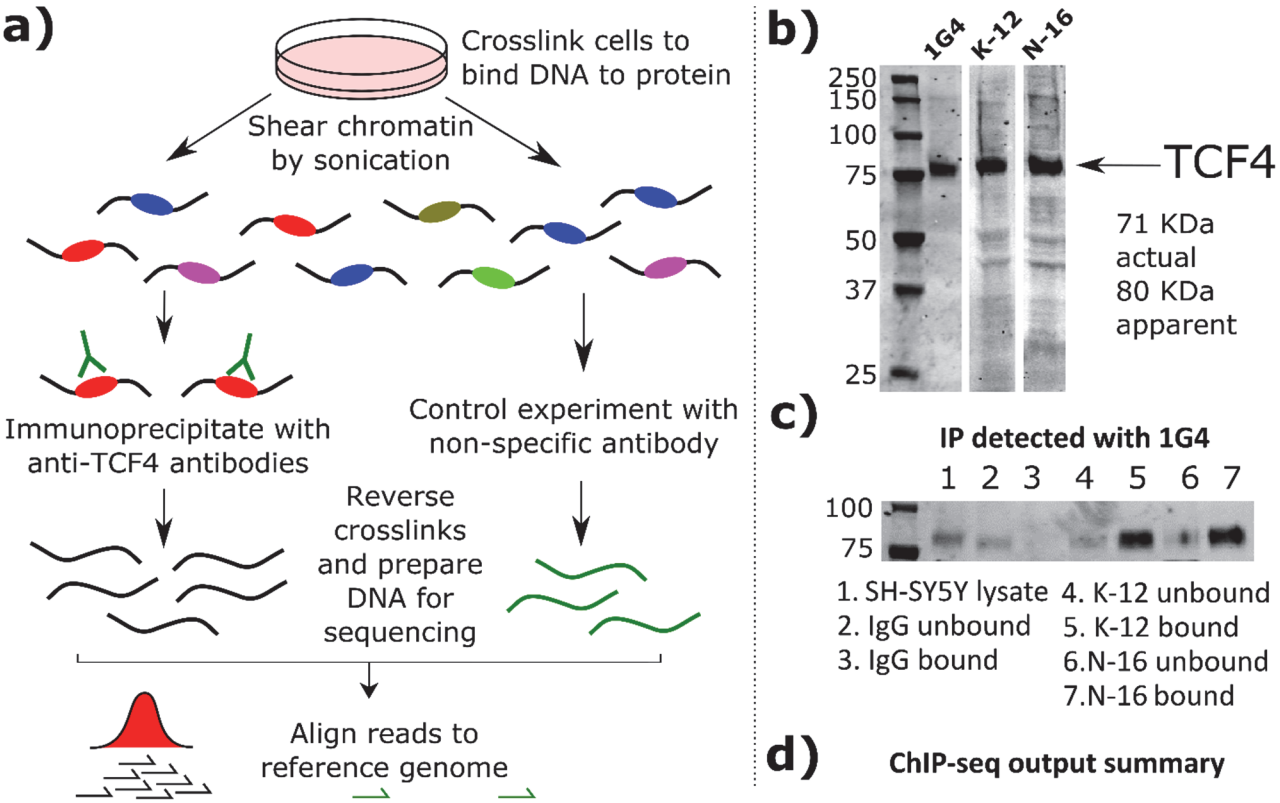
a) For our two TCF4 ChIP-seq experiments, we matched peaks in one experiment to the closest peak in the other and plotted the frequency distribution of the distances between the start sites. The histogram shows a clear enrichment for peaks that starting within 200 bp of each other, indicating agreement between the two experiments.

b) Numbers of TCF4 peaks per chromosome, using genome build hg19. Chromosome 23 is the X chromosome.

Figure 3.

Enrichment of TCF4 gene set expression in specific cell types as determined by single cell RNA-seq[30]. We tested if expression of genes in the TCF4 gene set were significantly higher than for genes not in the gene set for each cell type. The bold line at ~2.7 on the x-axis is the Bonferroni-adjusted $-\log_{10}$(p-value) for multiple testing (α=0.05/24).

Figure 1



d) ChIP-seq output summary

Anti-body	Unique reads	NSC	RSC	Peaks (FDR<1%)
K-12	14,034,896	1.12	1.049	77,167
N-16	16,190,153	1.13	1.048	70,551

Figure 2

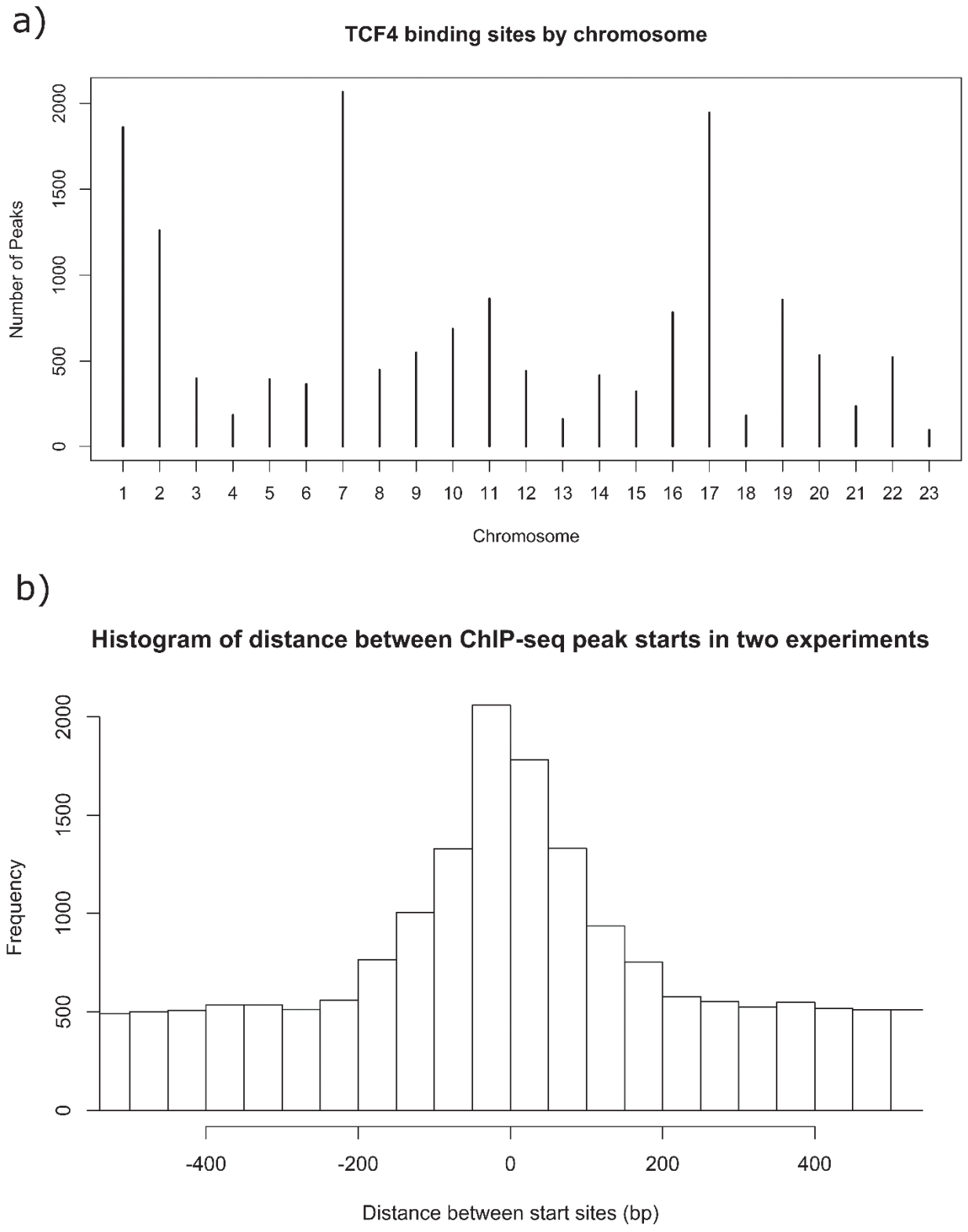
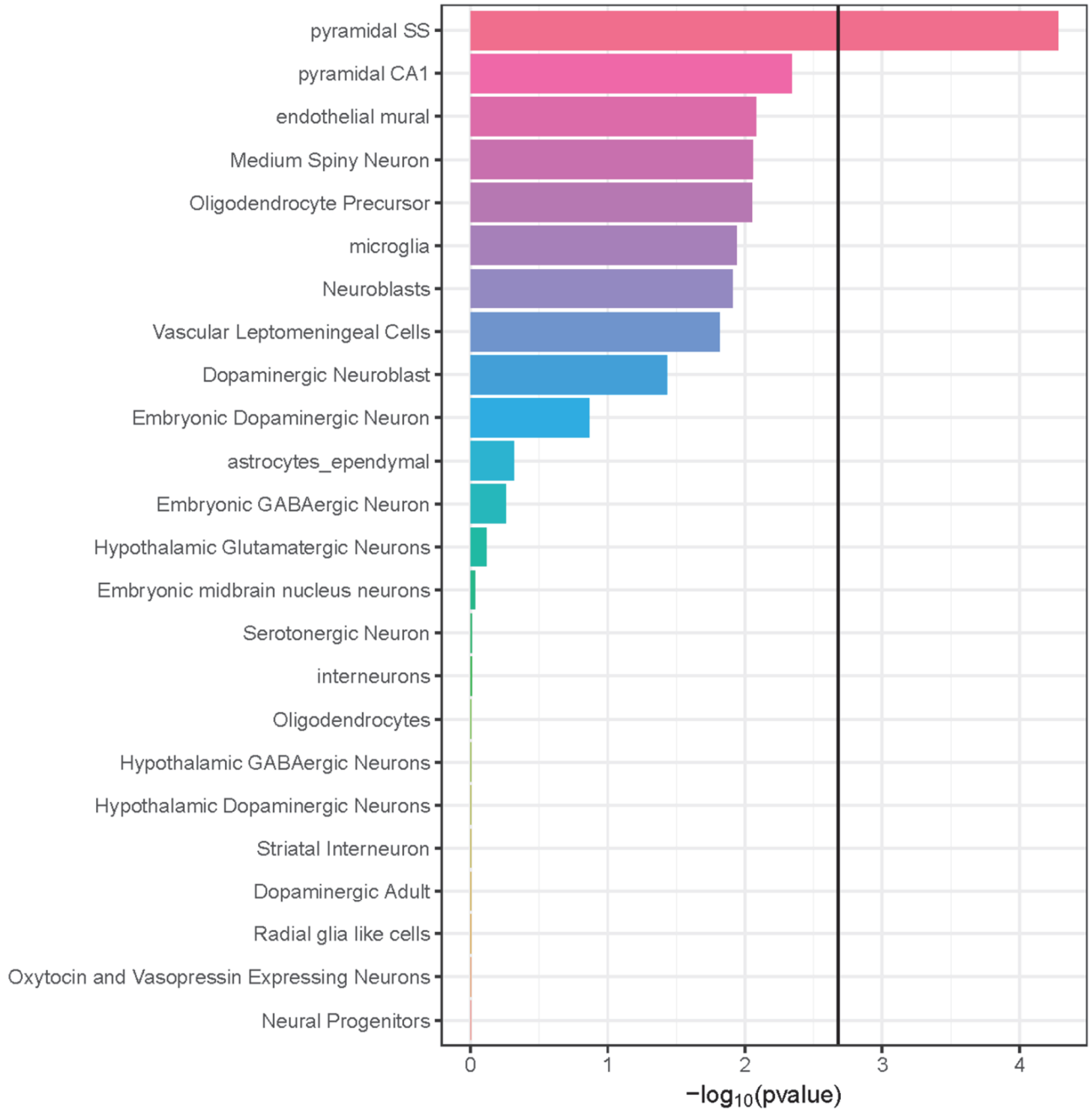

a)
TCF4 binding sites by chromosome
Number of Peaks
0
500
1000
1500
2000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Chromosome
b)
Histogram of distance between ChIP-seq peak starts in two experiments
Frequency
0
500
1000
1500
2000
-400
-200
0
200
400
Distance between start sites (bp)

Figure 3


pyramidal SS
pyramidal CA1
endothelial mural
Medium Spiny Neuron
Oligodendrocyte Precursor
microglia
Neuroblasts
Vascular Leptomeningeal Cells
Dopaminergic Neuroblast
Embryonic Dopaminergic Neuron
astrocytes_ependymal
Embryonic GABAergic Neuron
Hypothalamic Glutamatergic Neurons
Embryonic midbrain nucleus neurons
Serotonergic Neuron
interneurons
Oligodendrocytes
Hypothalamic GABAergic Neurons
Hypothalamic Dopaminergic Neurons
Striatal Interneuron
Dopaminergic Adult
Radial glia like cells
Oxytocin and Vasopressin Expressing Neurons
Neural Progenitors
0
1
2
3
4
−log10(pvalue)