DATA ANALYTICS AND MACHINE LEARNING

TRANSPARENT: a Python tool for designing transcription factor regulatory networks

Carlo Derelitto[1] · Daniele Santoni[1]

Abstract

Transcription factors are proteins able to selectively bind DNA short traits, namely transcription factors binding sites, in order to regulate gene expression in terms of both repression and activation. Despite plenty of studies focusing on transcription factors and on the role they play in specific biological tasks or diseases, is available in the literature, to our knowledge there is no tool able to automatically provide a list of transcription factors involved in this task and the associated interaction network through a solid computational analysis. TRANScriPtion fActor REgulatory NeTwork (TRANSPARENT) is a user-friendly Python tool designed to help researchers in studying given biological tasks or given diseases in human, by identifying transcription factors controlling and regulating the expression of genes associated with that task or disease. The tool takes in input a list of gene IDs and provides (1) a set of transcription factors that are significantly associated with the input genes, (2) the correspondent P values (i.e., the probability that this observed association was driven by chance) and (3) a transcription factor network that can be directly visualized through STRING database. The effectiveness and reliability of the tool were assessed by applying it to two different test cases: schizophrenia and autism disorders. The obtained results clearly show that identified TFs, for both datasets, are significantly associated with those disorders, in terms of both gene enrichment and coherence with the literature. The proposed tool TRANSPARENT can be a useful instrument to investigate transcription factor networks and unveil the role that TFs play in given biological tasks and diseases.

Keywords Transcription factors · Transcription factors binding sites · Protein networks · Gene regulation · Disease-associated genes

1 Introduction

Transcription factors (TFs) are proteins involved in the regulation of gene expression. They are able to selectively bind DNA short traits, namely transcription factors binding sites (TFBSs), often located in the promoter regions of genes, to regulate gene expression in terms of both repression and activation. A large collection of experimental datasets related to TFBSs (Zhang et al. 2020; Yevshin et al. 2017)—mainly coming from Chip-seq experiments—are available as well as a large amount of prediction data coming from computational tools designed and trained on experimental data (Tan and Lenhard 2016; Jayaram et al. 2016). TFs often act together and their binding to DNA sites of given promoters is tightly orchestrated in order to facilitate or impede gene expression depending on the need of the cell at a given time (Cumboo et al. 2018). The design of TF regulatory networks is a key point to understand the complex mechanisms underlying the regulation of gene expression in biological tasks and pathways (Wilkinson et al. 2017; Neph et al. 2012).

The study of TFs networks can also play a crucial role in designing therapeutic intervention to identify specific targets as shown by Karamouzis and Papavassiliou in the context of Breast cancer (Karamouzis and Papavassiliou 2011).

Chen and colleagues studied how TFs coordinate gene expression in a combinatorial fashion, through cliques of self-regulated core TFs controlling cell identity and cell state. They also studied the complex and interconnected feed-forward transcriptional loops building core transcriptional regulatory circuitry in cancer (Chen et al. 2020).

✉ Daniele Santoni
daniele.santoni@iasi.cnr.it

Carlo Derelitto
derelittocarlo@yahoo.it

[1] Istituto di Analisi dei Sistemi ed Informatica - Antonio Ruberti - CNR, Via dei Taurini 19, 00185 Rome, Italy

Lots of studies focused on identifying TFs and their interaction networks in different contexts such as self-renewability and pluripotency of embryonic stem cells (Nakai-Futatsugi and Niwa 2013), hematopoiesis (Wilson et al. 2011), environmental stress response (Song et al. 2016), T-cell development and differentiation (Collins et al. 2009) among the others.

Cui et al. (2010) developed a software package to identify TFs involved in biological processes using both gene expression data and existing knowledge base.

Despite a large number of studies, focusing on TFs and on the role they play in specific biological tasks or diseases, is available in the literature, to our knowledge there is no tool able to automatically provide a list of TFs involved in those tasks, through a solid computational analysis and the corresponding interaction networks exclusively basing on promoter TFBS enrichment.

Indeed MEME suite (Bailey et al. 2015) faces up a similar task but from a different point of view and so far it was not thought and focused on TFs and promoter sequences. Given a set of sequences (promoter sequences of given genes in this case), it is able to provide, if any, common consensus sequences occurring more than expected. The user can also provide his own consensus to look for in the sequences, but the software is not structured so that one can provide a position weight matrix (PWM) commonly used to characterize the TFBSs of a given TF (we recall that a position weight matrix reports, for a collection of sequences, the frequency of each nucleotide occurring in each position). Thus MEME does not use ad hoc algorithm designed to find similarity based on PWMs like the software we used (matchPWM of the Biostring R library, see Sect. 2).

Anyway to use MEME to this aim the user should provide any single consensus for each known TF (TRANSPARENT analyzes 626 different TFs) to identify significant TFs.

TRANSPARENT is implemented so that all available PWMs related to known and reliable TFs are considered and automatically included in the analysis; moreover, the tool also includes and manages the different transcripts associated to the genes in the considered list. The user has just to upload his own gene list and the computational analysis is completely *transparent* providing final results in textual mode and with a link to furtherly customize transcription factor network analysis through STRING web site.[1]

TRANSPARENT is a user-friendly Python tool, designed to help researchers to analyze TFs involved in the regulation of specific genes associated to a given task or a given disease in human. The tool was successfully applied to two different test cases: schizophrenia and autism disorders, identifying a set of TFs involved in the considered diseases and their interaction networks.

---

[1] https://string-db.org/.
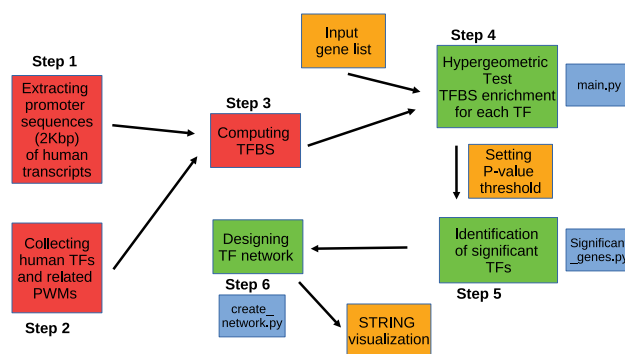
Fig. 1 TRANSPARENT pipeline. Red boxes represent precomputed steps, green boxes represent core steps of the tool computed through Python scripts (cyan boxes). Yellow boxes represent input data provided by users and instruments for the visualization (STRING database)

2 Materials and methods

TRANSPARENT (TRANScriPtion fActor REgulatory NeTwork) is a Python tool designed to identify TFs associated to a pool of genes responsible for a given task or associated to a given disease and to build an interaction network of selected TFs. The pipeline of the tool is depicted in Fig 1. Steps 1–3 (red boxes) are precomputed and data are already included in the package in order to minimize computational time and resources. Step 4–6 are computed on sample instance providing as a result a list of TFs associated to the uploaded gene list and an interaction network that can be directly visualized and managed through STRING site.

The six steps TRANSPARENT pipeline are reported in the following:

- Step 1—*Extracting human promoter sequences*
  A complete list of human genes and related transcripts, linked to the different isoforms of gene products, are selected. A number of 23,459 genes and 73,432 transcripts are collected. Promoter sequences (2000 base pairs upstream the transcription start site are considered according to Cumboo et al. 2018) of those gene/transcripts were retrieved through the package "TxDb Hsapiens UCSC.hg19.KnownGene" version 3.2.2 of R software.

- Step 2—*Collecting human TFs and related PWMs*
  The set of available 626 human TFs is selected and the related consensus pattern sequences, expressed in terms of position weight matrices (PWMs), are retrieved through JASPAR database (Fornes et al. 2020).

- Step 3—*Computing TFBSs*
  TFBSs associated to each considered human TF are com-

puted through the matchPWM() function, integrated into the Biostrings R library,[2] setting a threshold of 0.90.

- Step 4—*TFBS enrichment for each TF and hypergeometric test*

  Statistical tests are performed to assess the association between a given TF and the input gene set. Input file must contain a list of Entrez gene IDs, one ID for each line; an input file sample is included in the directory testset of the package. A hypergeometric test is performed for each considered TF, by comparing the number of genes in the pool set showing at least one TFBS in the promoter region (over all the transcripts) and the expected number, computed on the whole gene set. Obtained *P* values are then adjusted using Bonferroni's correction. A complete list of TFs, their associated *P* value and adjusted *P* value is made available in the output directory.

- Step 5—*Identification of significant TFs*

  TFs providing low *P* values (according to a threshold set by the user) are identified as potential regulatory factors of genes of the pool since they show a significant TFBS enrichment in the promoter sequences of those genes. A list of significant TFs is made available in the output directory together with the list of genes showing TFBSs related to a given TF.

- Step 6—*Designing TF network*

  A link to STRING database (Szklarczyk et al. 2021) visualizing the network of significant TFs is provided in the output directory. Default view is designed with a stringent interaction threshold but can be changed by the user in the STRING database. STRING visualization allows an at-a-glance view of connected significant TFs associated to the considered gene pool. The network of TFs and linked genes, initially submitted by the user, is also available through STRING database (when the -l flag is set).

3 Results

Several gene sets associated to given diseases were considered as test cases to assess the effectiveness and reliability of the tool. In this section, results associated to two sample gene lists are reported: the former is made of genes associated to schizophrenia disorder, and the latter is made of genes associated to autism disorder.

Gene lists were obtained through DisGeNET (Piñero et al. 2017), setting a threshold of 0.3 on the score associated to the likelihood of the link between gene and disease.
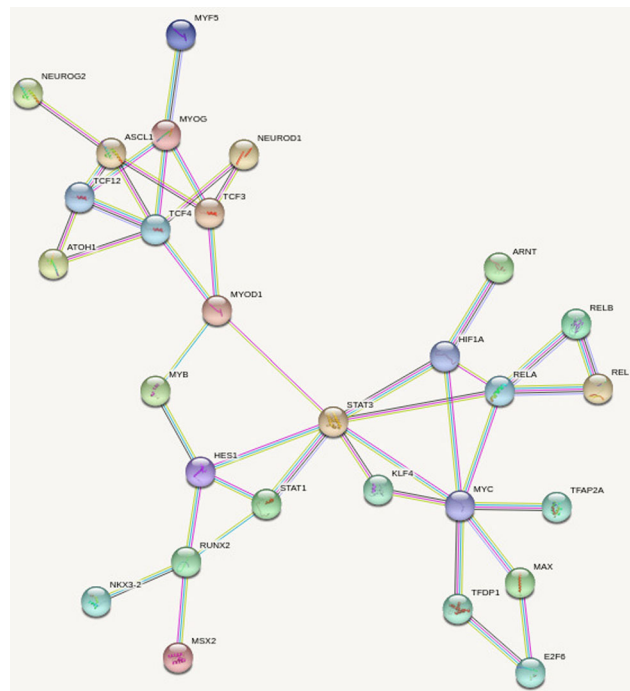
[2] https://bioconductor.org/packages/release/bioc/html/Biostrings.html.

Fig. 2 TF network associated to schizophrenia

3.1 First case study: schizophrenia disorder

A list of 1026 genes associated to schizophrenia disorder was downloaded from Disgenet (likelihood score higher than 0.3). The software identified 76 TFs (80 PWM models) showing a significant TFBS enrichment—adjusted *P* value smaller than 10^{-2}—in the promoter sequences of the 1026 schizophrenia-associated genes (101 TFs-107 PWM models—when considering an adjusted *P* value smaller than 5×10^{-2}). Potential interactions among the 76 identified TFs were analysed through STRING database (Szklarczyk et al. 2021). 28 TFs, out of the initial 76 ones, were found to be connected considering a stringent threshold on interaction likelihood ($T = 0.9$). The network, made available by TRANSPARENT software through STRING, is reported in Fig. 2. The number of connected TFs considering a smaller threshold on interaction likelihood is 38 for $T = 0.7$ and 66 for $T = 0.4$. The extended network, considering first (or second) neighbor nodes can be analyzed through STRING database in terms of both biological composition and clusters of the networks and can be customized setting the interaction likelihood threshold.

Enrichment disease analysis of identified TFs (adjusted P-value smaller than (5×10^{-2}) was performed through DAVID tool (Jiao et al. 2012). Several classes, coming from GAD Disease database (Becker et al. 2004), were found to be significantly enriched; among them:

Table 1 Identified TFs associated to schizophrenia in GAD disease database

Gene symbol	Gene name
ASCl1	Achaete-scute family bHLH transcription factor 1
FOXP2	Forkhead box P2
KLF5	Kruppel-like factor 5
RUNX2	Runt-related transcription factor 2
TBX3	T-box 3
TCF4	Transcription factor 4
TFAP2A	Transcription factor AP-2 alpha
TFAP2B	Transcription factor AP-2 beta

- Schizophrenia disorder (8 TFs—P value 6.9×10^{-2} considering best 76 TFs)
- Parkinson's disease (6 TFs P value 2×10^{-3})
- Depression (5 TFs—P value 1.7×10^{-2})
- Neurological Disease Class (27 TFs—P value 3.1×10^{-2})
- Antisocial behavioral traits (2 TFs—P value 6.1×10^{-2})
- PSYCH disease Class (19 TFs—P value 6.6×10^{-2})
- Schizophrenia/bipolar disorder (2 TFs—P value 9.1×10^{-2}) considering best 101 TFs.

The 8 identified TFs belonging to the GAD disease class schizophrenia are reported in Table 1.

Identified TFs with the lowest adjusted P values are:

- MAZ (adjusted P value $< 10^{-21}$)
- KLF5 (adjusted P value $< 10^{-18}$)
- KLF15 (adjusted P value $< 10^{-17}$)
- VEZF1 (adjusted P value $< 10^{-16}$)
- ZNF148 (adjusted P value $< 10^{-15}$).

Remarkably 29 TFs showed an adjusted P value lower than 10^{-5}. The table with all the TFs with P value smaller than 5×10^{-2} is reported as Supplemental Material (TabS1). Interestingly KLF5 (adjusted P value $< 10^{-18}$), KLF15 (adjusted P value $< 10^{-17}$), KLF4 (adjusted P value $< 10^{-13}$) and KLF2 (adjusted P value $< 10^{-11}$), belonging to the family Kruppel-like factor, were identified as highly significant, in particular KLF5 was found to be highly related to schizophrenia and downregulated in schizophrenic subjects (Yanagi et al. 2008). Moreover 5 TFs, belonging to the TCF family, TCF3 (P value $< 10^{-6}$), TCF12 (P value $< 10^{-6}$), TCFL5 (P value $< 10^{-3}$), TCF7 (P value $< 10^{-3}$) and TCF4 (P value $< 10^{-2}$) were found as highly significant. Particularly TCF4 was found to regulate genes involved in neuronal development and schizophrenia risk (Xia et al. 2018; Zakharyan 2016. Also FOS (adjusted P value $< 10^{-2}$) is known to be involved in schizophrenia disorder (Zakharyan 2016).
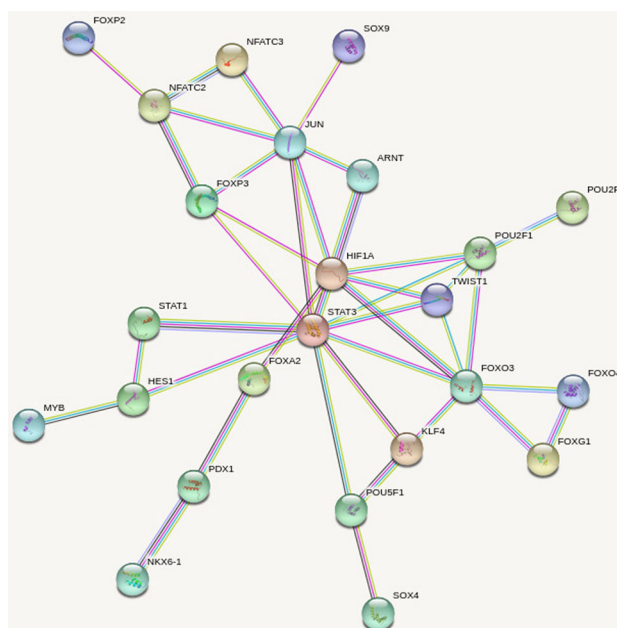
Fig. 3 TF network associated to autism disorder in GAD disease database

3.2 Second case study: autism disorder

A list of 1112 genes associated to autism disorder was downloaded from Disgenet (no threshold on the likelihood was set). The software identified 181 TFs showing a significant TFBS enrichment—adjusted P value smaller than 0.01—in the promoter sequences of the 1112 autism-associated genes (214 TFs when considering an adjusted P value smaller than 0.05). Potential interactions among the 181 identified TFs were analyzed through STRING database (Szklarczyk et al. 2021). 24 TFs, out of the initial 181, were found to be in the first connected component, considering a stringent threshold on interaction likelihood ($T = 0.9$). The related network, made available by TRANSPARENT software, is reported in the Fig. 3. The number of TSs in the connected network considering a smaller threshold on interaction likelihood is 97 for $T = 0.7$ and 172 for $T = 0.4$.

Enrichment analysis of the 214 identified TFs (adjusted P value smaller than 0.05) was performed through DAVID tool (Jiao et al. 2012). Several classes, coming from GAD Disease database (Becker et al. 2004), were found to be significantly enriched; among them:

- Autism (8 TFs—P value $< 9.5 \times 10^{-2}$ considering best 181 TFs)
- Neurodevelopmental psychiatric disorders (3 TFs—P value $< 1.7 \times 10^{32}$)
- Parkinson's Disease (7 TFs—P value $< 1.2 \times 10^{-2}$)
- Depression (7 TFs—P value $< 1.6 \times 10^{-2}$) considering best 214 TFs.

Table 2 Identified TFs associated to autism disorder in GAD disease database

Gene symbol	Gene name
CUX2	Cut-like homeobox 2
EN2	Engrailed homeobox 2
FOXP2	Forkhead box P2
HES1	Hes family bHLH transcription factor 1
HOXA1	Homeobox A1
MAZ	Myc-associated zinc finger protein
NFIL3	Nuclear factor, interleukin 3 regulated
POU6F2	POU class 6 homeobox 2

The 8 identified TFs belonging to the GAD disease class autism are reported in Table 2.

Identified TFs with the lowest P values are:

- VEZF1 (P value $< 10^{-17}$)
- MZF1 (P value $< 10^{-17}$)
- KFL15 (P value $< 10^{-15}$)
- MAZ (P value $< 10^{-15}$)
- ZNF148 (P value $< 10^{-14}$).

Forty-five TFs showed an adjusted P value lower than 10^{-5}. The table with all the TFs with P value smaller than 10^{-2} is reported as Supplemental Material (TabS2). Interestingly, many TFs-19—belonging to the family FOX (forkhead box), were identified as highly significant; among them: FOXP2 (adjusted P value $< 10^{-8}$), FOXH1 (adjusted P value $< 10^{-8}$), FOXK2 (adjusted P value $< 10^{-7}$), FOXP1 (adjusted P value $< 10^{-7}$), FOXA3 (adjusted P value $< 10^{-6}$), FOXO3 (P value $< 10^{-5}$), FOXD1 (adjusted P value $< 10^{-5}$). Those findings are consistent with the related literature, providing a strong evidence of the link between FOX genes (expressed in the central nervous system that are involved in brain development as well as the evolution of language) and autism spectrum disorder, regulating genes implicated in this disorder (Bowers and Konopka 2012). Interestingly a significant association between FOXP2 single nucleotide polymorphisms and autistic disorder was found in Gong et al. (2004). Moreover FOXO subfamily is known to be involved in age-progressive axonal degeneration and associated to several neurological and neurodevelopmental disorders, such as epilepsy, microcephaly, and autism (Hwang et al. 2018). Similarly, several TFs belonging to the homeobox family (HOX) and to Basic Helix-Loop-Helix (BHLH), in particular ASCL1, were identified by the software as highly significant in line with previous works claiming the association between those TF families and autism disorders (Rylaarsdam and Guemez-Gamboa 2019).

4 Conclusion

The tool presented in this work is a user-friendly software designed to help researchers in analysing gene sets associated to a given task or a given disease. It allows to extract useful information regarding transcription factors involved in the expression regulation of given gene sets providing the related TF network that can be directly visualized and further customized through STRING web resource. The effectiveness and reliability of the tool was assessed trough two different test cases: schizophrenia and autism disorder. Obtained results clearly show that identified TFs, for both datasets, are significantly associated with given disorders, in terms of both gene enrichment and coherence with the literature. TRANSPARENT is based on a simple but straight computational analysis; to the best of our knowledge there is no available tool able to provide clear and easy to use data associated to transcription factor regulation of a given gene set. In conclusion we are confident that TRANSPARENT can be a useful instrument to investigate transcription factor networks and unveil the role that TFs play in given biological tasks and diseases.

Declarations

Declarations

References

Bailey TL, Johnson J, Grant CE, Noble WS (2015) The MEME Suite. Nucl Acids Res 43(W1):W39-49

Becker KG, Barnes KC, Bright TJ, Wang SA (2004) The genetic association database. Nat Genet 36:431–432

Bowers JM, Konopka G (2012) The role of the FOXP family of transcription factors in ASD. Dis Markers 33(5):251–60. https://doi.org/10.3233/DMA-2012-0919

Chen Y, Xu L, Lin RY, Müschen M, Koeffler HP (2020) Core transcriptional regulatory circuitries in cancer. Oncogene 39:6633–6646

Collins A, Littman DR, Taniuchi I (2009) RUNX proteins in transcription factor networks that regulate T-cell lineage choice. Nat Rev Immunol 9(2):106–115

Cui X, Wang T, Chen HS et al (2010) TF-finder: a software package for identifying transcription factors involved in biological processes using microarray data and existing knowledge base. BMC Bioinform 11:425

Cumboo F, Vergni D, Santoni D (2018) Investigating transcription factor synergism in humans. DNA Res 25(1):103–112

Fornes O, Castro-Mondragon JA, Khan A, van der Lee R et al (2020) JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucl Acids Res 48(D1):D87–D92

Gong X, Jia M, Ruan Y, Shuang M, Liu J, Wu S, Guo Y, Yang J, Ling Y, Yang X, Zhang D (2004) Association between the FOXP2 gene and autistic disorder in Chinese population. Am J Med Genet B Neuropsychiatr Genet 127B(1):113–6. https://doi.org/10.1002/ajmg.b.20162

Hwang I, Oh H, Santo E, Kim DY, Chen JW, Bronson RT, Locasale JW, Na Y, Lee J, Reed S, Toth M, Yu WH, Muller FL, Paik J (2018) FOXO protects against age-progressive axonal degeneration. Aging Cell 17(1):e12701. https://doi.org/10.1111/acel.12701

Jayaram N, Usvyat D, Martin AC (2016) Evaluating tools for transcription factor binding site prediction. BMC Bioinf 17(1):547

Jiao X, Sherman BT, Huang DW, Stephens R, Baseler MW, Lane HC, RichardLempicki RA. DAVID-WS: a stateful web service to facilitate gene/protein list analysis

Karamouzis MV, Papavassiliou AG (2011) Transcription factor networks as targets for therapeutic intervention of cancer: the breast cancer paradigm. Mol Med 17(11–12):1133–1136

Nakai-Futatsugi Y, Niwa H (2013) Transcription factor network in embryonic stem cells: heterogeneity under the stringency. Biol Pharm Bull 36(2):166–170

Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA (2012) Circuitry and dynamics of human transcription factor regulatory networks. Cell 150(6):1274–1286

Piñero J, Bravo A, Queralt-Rosinach N, Gutíerrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucl Acids Res 45(D1):D833–D839

Rylaarsdam L, Guemez-Gamboa A (2019) Genetic causes and modifiers of autism spectrum disorder. Front Cell Neurosci 13:385

Song L, Huang SC, Wise A, Castanon R, Nery JR, Chen H, Watanabe M, Thomas J, Bar-Joseph Z, Ecker JR (2016) A transcription factor hierarchy defines an environmental stress response network. Science 354(6312):aag1550

Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, Jensen LJ, von Mering C (2021) The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucl Acids Res 49(D1):D605-12

Tan G, Lenhard B (2016) TFBSTools: an R/Bioconductor package for transcription factor binding site analysis. Bioinformatics 32:1555–1556

Wilkinson AC, Nakauchi H, Göttgens B (2017) Mammalian transcription factor networks: recent advances in interrogating biological complexity. Cell Syst 5(4):319–331. https://doi.org/10.1016/j.cels.2017.07.004

Wilson NK, Calero-Nieto FJ, Ferreira R, Göttgens B (2011) Transcriptional regulation of haematopoietic transcription factors. Stem Cell Res Therapy 2:6

Xia H, Jahr FM, Kim NK, Xie L, Shabalin AA, Bryois J, Sweet DH, Kronfol MM, Palasuberniam P, McRae M, Riley BP, Sullivan PF, van den Oord EJ, McClay JL (2018) Building a schizophrenia genetic network: transcription factor 4 regulates genes involved in neuronal development and schizophrenia risk. Hum Mol Genet 27(18):3246–3256. https://doi.org/10.1093/hmg/ddy222

Yanagi M, Hashimoto T, Kitamura N, Fukutake M, Komure O, Naoki NN, Kawamata T, Maeda K, Shirakawa O (2008) Expression of Kruppel-like factor 5 gene in human brain and association of the gene with the susceptibility to schizophrenia. Schizophr Res 100:291–301

Yevshin I, Sharipov R, Valeev T, Kel A, Kolpakov F (2017) GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. Nucl Acids Res 45(D1):D61–D67

Zakharyan R (2016) Transcription factors in schizophrenia: a current view of genetic aspects. Sci J Genetics Gen Ther 2(1):17–21

Zhang Q, Liu W, Zhang H, Xie G, Miao Y, Xia M, Guo A (2020) hTFtarget: a comprehensive database for regulations of human transcription factors and their targets. Genom Proteom Bioinform 18(2):120–128