SPATIO-TEMPORAL GENE DISCOVERY FOR AUTISM SPECTRUM DISORDER

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

By
Utku Norman
May 2018

Spatio-Temporal Gene Discovery for Autism Spectrum Disorder
By Utku Norman
May 2018

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

A. Ercüment Çiçek (Advisor)

Mehmet Koyutürk

Öznur Taştan Okan

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

SPATIO-TEMPORAL GENE DISCOVERY FOR AUTISM SPECTRUM DISORDER

Utku Norman
M.S. in Computer Engineering
Advisor: A. Ercüment Çiçek
May 2018

Whole Exome Sequencing (WES) studies for Autism Spectrum Disorder (ASD) could identify only around six dozen risk genes to date, because the genetic architecture of the disorder is highly complex. To speed the gene discovery process up, a few network-based ASD gene discovery algorithms were proposed. Although these methods use static gene interaction networks, functional clustering of genes is bound to evolve during neurodevelopment and disruptions are likely to have a cascading effect on the future associations. Thus, approaches that disregard the dynamic nature of neurodevelopment are limited. Here, we present a spatio-temporal gene discovery algorithm for ASD, which leverages information from evolving gene coexpression networks of neurodevelopment. The algorithm solves a prize-collecting Steiner forest based problem on coexpression networks, adapted to model neurodevelopment and transfer information from precursor neurodevelopmental windows. The decisions made by the algorithm can be traced back, adding interpretability to the results. We apply the algorithm on WES data of 3,871 samples and identify risk clusters using BrainSpan coexpression networks of early- and mid-fetal periods. On an independent dataset, we show that incorporation of the temporal dimension increases the predictive power: Predicted clusters are hit more (i.e. they contain genes with more disruptive mutations on them) and show higher enrichment in ASD-related functions compared to the state of the art.

Keywords: spatio-temporal networks, gene discovery, prize-collecting Steiner forest problem, autism spectrum disorder.

ÖZET

OTİZM SPEKTRUM BOZUKLUĞU İÇİN ZAMAN-MEKANSAL GEN KEŞFİ

Utku Norman
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: A. Ercüment Çiçek
Mayıs 2018

Otizm Spektrum Bozukluğu'nun (OSB) kalıtsal yapısının karmaşıklığından dolayı Tüm Ekzom Dizileme (Whole Exome Sequencing ya da WES) çalışmaları ile günümüze değin sadece altı düzine kadar risk geni belirlenebilmiştir. Gen keşif sürecini hızlandırabilmek amacıyla ağ temelli birkaç yöntem geliştirilmiştir. Bu yöntemlerin kullandıkları ağlar, durağan türden gen-gen etkileşim ağlarındandır. Gelgelelim, genlerin işlevsel kümelenmeleri sinir sisteminin gelişimiyle evrilir. Ayrıca, gen işleyişlerindeki aksaklıklar kimi zaman sonraki gen-gen etkileşimleri üzerinde katlanarak artan bozulmalara neden olur. Bu nedenle, sinir sistemi gelişiminin değişken ve devingen doğasını göz önünde bulundurmayan yaklaşımlar sınırlı kalacaktır. Çalışmamızda sinir sistemi gelişimi bağlamında evrimleşen gen-gen ortak ifade (coexpression) ağlarının zaman-mekansal bilgisini kullanan ST-Steiner adını verdiğimiz bir gen keşif algoritması sunulmaktadır. Bu algoritma, sinir sistemi gelişimini modelleyecek şekilde uyarlanmış ödül toplayan Steiner ormanı (prize-collecting Steiner forest) temelli bir problemi, öncül sinir-gelişimsel pencerelerdeki bilgiyi taşıyarak, ortak ifade ağlarında çözmektedir. Algoritmanın verdiği kararların izleri geriye doğru sürülebilmekte; bu da sonuçların yorumlanabilirliğini arttırmaktadır. Çalışmamızda ST-Steiner, 3871 örnekten oluşan WES verisine uygulanmakta; erken ve orta cenin dönemlerinin BrainSpan ortak ifade ağlarından risk geni kümeleri belirlenmektedir. Ayrıca, bağımsız bir veri kümesinde, zamansal bilgiyi eklemenin öngörü gücünü arttırdığı gösterilmektedir: Belirlenen kümeler en gelişkin yöntemler (state of the art) ile karşılaştırıldığında hem daha fazla isabet görmekte ,yani daha fazla yıkıcı değişinim (mutasyon) içeren genlerden oluşmakta, hem de OSB ile ilişkili işlevlerde daha çok zenginleşme (enrichment) göstermektedir.

Anahtar sözcükler: zaman-mekansal ağlar, gen keşfi, ödül toplayan Steiner ormanı problemi, otizm spektrum bozukluğu.

To Gülseren—
моя любимая.

Acknowledgement

First and foremost, I am truly grateful to Asst. Prof. A. Ercüment Çiçek, for his guidance in how to carry out research, his patience and constant support, always accompanied by swift responses to my inquiries. He greeted me with this project that I could dive into, in a field that I wished to explore. Thank you.

I would like to thank Mehmet Koyutürk and Öznur Taştan Okan for accepting to be in my thesis committee. I am particularly grateful for the assistance given by Serhan Yilmaz, his comments have been invaluable. I thank Fereydoun Hormozdiari and Li Liu for help with their methods, and Lambertus Klei for the help on processing BrainSpan data.

I would like to offer my special thanks to Gülseren Şen, who has always been by my side, encouraging me, sustaining my perseverance. I am indepted to her.

Last but not least, I would like to thank everyone I am yet to mention. In particular, I would like to express my deepest gratitude to my family. They have always supported me with love and understanding.

This work was supported by the very much appreciated Simons Foundation Autism Research Initiative Explorer Grant (416835) to Asst. Prof. A. Ercüment Çiçek.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Autism Spectrum Disorder (ASD) is a common neurodevelopmental disorder that affects ∼1.5% of the children in the US [1]. Recent whole exome sequencing (WES) efforts have paved the way for identification of dozens of ASD risk genes [2, 3, 4, 5, 6, 7, 8]. Unfortunately, this number corresponds to only a small portion of the large genetic puzzle, which is expected to contain around a thousand genes [9].

Detection of *de novo* loss-of-function mutations (dnLoFs) has been the key for gene discovery due to their high signal-to-noise ratio. However, such mutations are rare and they affect a diverse set of genes. Thus, for most of the genes, the rarity and diversity of variants prevent statistically significant signals from being observed. Therefore, despite analyzing thousands of trios, our yield of discovered genes is still low. The journey towards getting the full picture of the genetic architecture will take a long time and will be financially costly.

Guilt-by-association based gene discovery techniques came in handy to predict ASD risk genes. They assume these genes are working as a functional cluster and utilize networks to capture the functional connectivity. Starting from already known risk genes, these techniques predict a cluster of closely interacting genes using networks.

There are only a few network-based ASD-tailored gene discovery algorithms in the literature: (i) NETwork-Based Analysis of Genetic associations (NETBAG) algorithm [10, 11], (ii) Detecting Associations With Networks (DAWN) [12, 13] algorithm, and (iii) Merging Affected Genes into Integrated networks (MAGI) [14] algorithm.

NETBAG constructs an integrated network through a comprehensive analysis of many annotation resources such as Gene Ontology (GO) and protein domains, as well as numerous available interaction networks—e.g. Protein-Protein Interactions (PPIs), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. On the network, the method assigns each gene pair an interaction score that signifies the likelihood of those genes participating in a genetic phenotype. Then, by following a greedy seed-and-extend procedure, it generates many clusters of genes and returns the cluster with the maximal score that is significant with respect to a permutation test.

DAWN estimates partial coexpression networks using BrainSpan, but for only small windows of the neurodevelopment that are indicated as hotspots for ASD [15]. Unlike NETBAG and MAGI, which use phenotype-neutral networks, DAWN's approach favors links to already known ASD genes. It predicts ASD genes by assigning a posterior risk score to each gene based on its interactions with other genes, using a hidden Markov random field based approach.

MAGI uses two PPIs: Search Tool for the Retrieval of Interacting Genes / Proteins (STRING) [16] and Human Protein Reference Database (HPRD) [17]. It also makes use of the coexpression network generated by using full (all brain regions and neurodevelopmental periods) data from the BrainSpan dataset [18]. MAGI follows a seed-and-extend based approach similar to NETBAG to generate seed pathways which are enriched with dnLoFs in cases, compared to controls. Then, it merges the pathways as long as the cluster score is improved.

We also would like to mention the method by [19], which uses a data-driven tissue-specific network, where thousands of experiments from the literature are integrated in a Bayesian framework [20]. Instead of predicting a module like the other described methods, this method uses the known ASD genes and their connectivity patterns as features to train a Support Vector Machine (SVM) classifier, and then assign every gene a probability of being associated with ASD.

Clearly, none of the above-mentioned methods consider the fact that gene interactions (coexpressions) evolve over time. Despite having fundamental differences in their approaches, all of these methods have one point in common: The biological gene interaction networks that they use are static. Yet, it is demonstrated that different neurodevelopmental spatio-temporal windows have different topologies and consequently, the clustering of ASD susceptibility genes changes drastically [15]. Moreover, dysregulation of pathways in earlier periods has cascading effects on the circuitry of the future time periods. For instance, it is shown that ASD risk increases 36-fold with a damage in early cerebellar circuitry, but a damage in adulthood does not cause dysfunction in social interactions [21]. Thus, early cerebellar damage is demonstrated to result in a cascade of long-term deficits in cerebellum and cause ASD [22]. Hence, we argue that the state-of-the-art methods are limited in predictive power, since static networks they use would fail to capture the dynamic nature of neurodevelopment.

In this work, we propose a novel ASD gene discovery algorithm termed ST-Steiner. The algorithm modifies the prize-collecting Steiner forest (PCSF) problem and extends it to spatio-temporal networks in order to mimic neurodevelopment. Instead of performing gene discovery on a single network or any number of networks separately, ST-Steiner solves an optimization problem progressively over a cascade of spatio-temporal networks while leveraging information coming from earlier neurodevelopmental periods.

The algorithm ST-Steiner has three novel aspects: (i) For the first time, the problem is solved on a cascade spatio-temporal coexpression networks so that the dynamic nature of neurodevelopment is taken into account; (ii) The results are more interpretable compared to other methods in the literature, since the decisions made by the algorithm can be traced back on the spatio-temporal cascade; (iii) As the problem is formulated as a PCSF prediction problem, the algorithm predicts only the genes that are essential for the connectivity of known risk genes. Note that, this is in contrast to the other approaches which can return redundant paths between known risk genes.

We apply the algorithm ST-Steiner on exome sequencing data from [6], and identify gene clusters using 2 gene coexpression network cascades of early-fetal and mid-fetal periods. For the first time, we benchmark the performances of the network-based ASD gene discovery algorithms on common training data. We validate the predicted clusters with independent exome sequencing data from [7] and find that ST-Steiner's predictions (i) identify more risk genes with less predictions compared to the state-of-the-art methods, (ii) overlap more with targets of known ASD-related transcription factors and (iii) enriched more in ASD-related KEGG pathways. We also show in various controlled settings that using temporal information boosts the predictive power, which supports our claim that the clustering of risk genes is spatio-temporal. We demonstrate that prediction of genes like NOTCH2 and NOTCH3 can be traced back to their role in neuron differentiation in embryonic period. Finally, with the incorporation of the time dimension, we predict new genes that point to a disruption of organization of microtubules in the embryonic period.

Chapter 2

Methods

2.1 Overview and Background of ST-Steiner

The method we propose to remedy the shortcoming posed by using static networks is built on prize-collecting Steiner forest (PCSF) problem.

The original Steiner tree problem is concerned with finding a tree to connect a given set of distinguished (seed) nodes [23, 24]. In biological context, it is first used over node-weighted biological networks to detect regulatory subnetworks [25].

The prize-collecting version of the Steiner tree problem aims to find a tree that maximizes the sum of the prizes of the selected nodes while penalizing the total cost of the connecting edges. In the biology domain, PCST has proven useful for applications like detecting functional modules in PPIs [26], signaling pathway prediction [27, 28], metabolic pathway prediction [29] and drug repositioning [30].

Prize-collecting Steiner forest (PCSF) problem is a relaxation of PCST that allows multiple disconnected components (trees). PCSF has been used to identify multiple independent signaling pathways on a single network [31]. Later, PCSF is extended to predict a single tree shared among multiple samples (networks with identical topology) with different mutation profiles (different seed genes) [32].

Among the PCSF-variant approaches, we are only aware of two applications of PCSF on temporal data. The first one uses a temporal cascade like ours, however, tries to satisfy k connectivity demands (e.g. vertex1 must be connected to vertex2) [33]. This is different than our application, which aims to incorporate information from past time points without imposing such constraints. The second work uses fold changes in temporal expression data, but by solving PCSF on static PPI networks [34].

In ST-Steiner, selection of similar genes across multiple networks is rewarded in a similar manner to [32]. However, in contrast, (i) ST-Steiner works with networks of different topologies, (ii) networks are organized in a temporal hierarchy, (iii) reward mechanism is weighted by the prize of a node and only affects networks of future time windows, and, (iv) networks represent spatio-temporal windows in brain development rather than samples with different mutation profiles, (v) multiple brain regions in the same time window can be simultaneously analyzed without constraining to be similar to each other, where all selected genes in a given time frame can be assigned an additional prize in the analysis of the next time frame.

A motivating toy example for ST-Steiner and its decision making process is illustrated in Figure 2.1. Consider the coexpression network of spatio-temporal window 1. By solving PCSF on this network, the known risk genes (black) are minimally connected by selecting a set of genes (red-bordered). This selection affects which genes will be chosen on the coexpression network of spatio-temporal window 2. Assume genes X, Y and Z are equally likely to be selected (equal prizes and related edge costs) to connect the seed genes. Then, the algorithm prefers gene X, because it is selected in the earlier period and its prize is increased. Next, we formally define PCST, PCSF and then ST-Steiner.

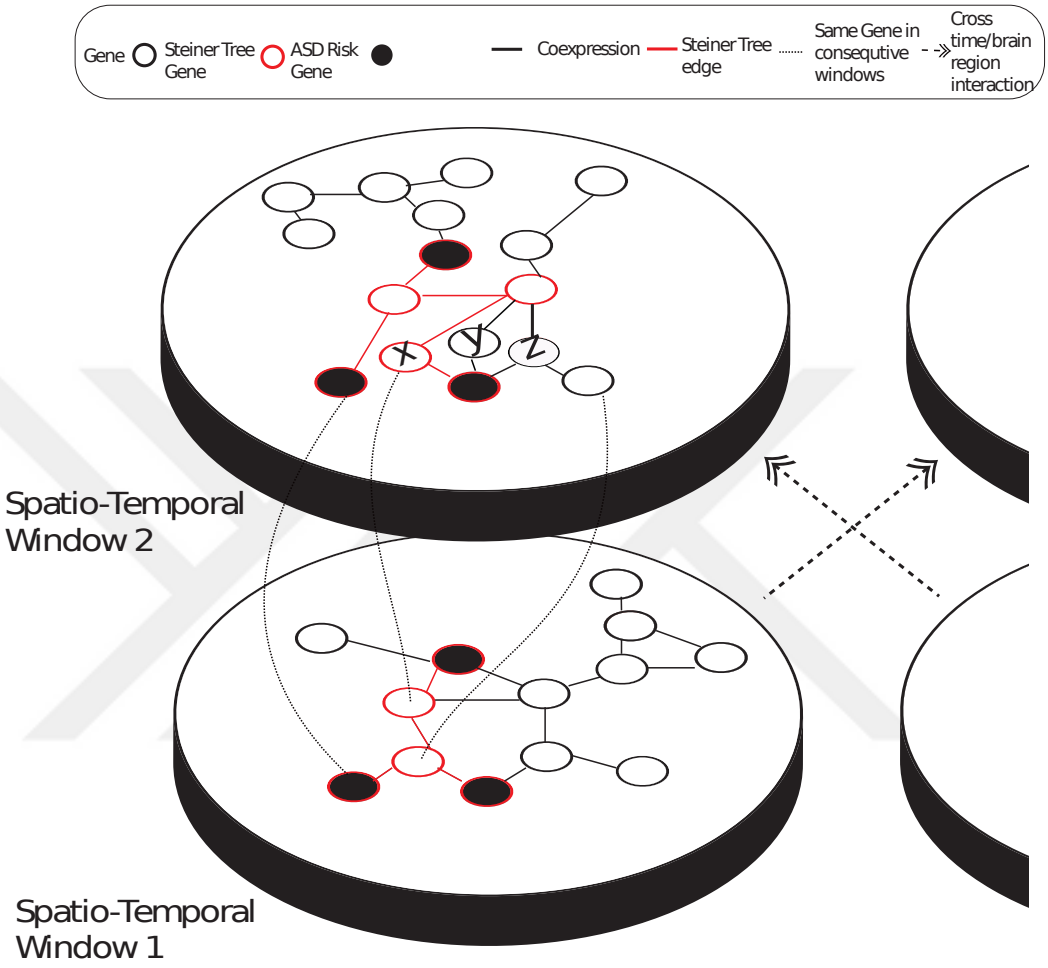

Figure 2.1: Figure shows two spatio-temporal windows (plates) and respective coexpression networks along with a parallel brain region and its plates (partially shown, on the right). Circles represent genes and black edges represent pairs of genes that are coexpressed. Red-bordered nodes form the Steiner tree found on plate 1 (linked with red edges), which minimally connects black seed genes. In ST-Steiner, genes that are selected in plate 1 are more likely to be selected in plate 2. Curved lines between windows show the mapping of selected genes from plate 1 to plate 2. On the second plate ST-Steiner can pick X, Y or Z to connect the seed genes. Assuming that they all have identical priors and identical edge costs, the algorithm would pick X, because it is selected in the prior window and its prize is increased. If other brain regions in the first temporal window are also considered, then selected genes in those regions would also be used (from the plate on the right).

2.2 Prize-collecting Steiner Tree Problem (PCST)

Let $G(V, E)$ denote an undirected vertex- and edge-weighted graph. V is the set of nodes, $E \subseteq V \times V$ is the set of edges that connects node pairs. $p : V \to \mathbb{R}_{\geq 0}$ is the node prize function and $c : E \to \mathbb{R}_{\geq 0}$ is the edge cost function for G. The task is to find a connected subgraph $T(V_T, E_T)$ of a graph G, that minimizes

$$o_T(T) = \sum_{e \in E_T} c(e) + \beta \sum_{v \notin V_T} p(v), \quad \beta \geq 0, \tag{2.1}$$

where β is a sparsity parameter that adjusts the trade-off between node prizes and edge costs. This trade-off corresponds to collecting the prize on a node by including the node and evading its edge cost by excluding it. An optimal solution is a tree, since if a cycle exists in T, any edge on the cycle can be removed to obtain another tree T' with $o_T(T') \leq o_T(T)$.

2.3 Prize-collecting Steiner Forest Problem (PCSF)

PCSF is an extension of PCST that lifts the connectedness constraint on the desired subgraph: Instead of a single tree, the goal is to find a forest. Given G, the problem is to find a subgraph $F(V_F, E_F)$ of the graph G that minimizes

$$o_F(F) = \sum_{e \in E_F} c(e) + \beta \sum_{v \notin V_F} p(v) + \omega \kappa_F, \quad \beta \geq 0,\ \omega \geq 0, \tag{2.2}$$

where $\kappa_F \in \mathbb{N}$ is the number of connected subgraphs (trees) in the subgraph F and ω is a parameter that adjusts its penalty. PCSF is a generalized version of PCST and reduces to PCST when $\omega = 0$. An instance of PCSF can be solved as a PCST instance by adding an artificial node v_0 to V and edges $E_0 = \{v_0 v_i |\ v_i \in V\}$ with cost ω to E. Solving PCST on this new graph, and afterwards, removing v_0 and E_0, yields a minimal solution for the original PCSF instance [35].

2.4 Spatio-temporal Modelling

In order to model the spatio-temporal dynamics of neurodevelopment, we consider a spatio-temporal system $\boldsymbol{G} = (\boldsymbol{G_1}, \boldsymbol{G_2}, \cdots, \boldsymbol{G_T})$, a list of T consecutive time windows.

The i^{th} time window $\boldsymbol{G_i} = \{G_i^1, G_i^2 \cdots, G_i^n\}$ is a set of spatio-temporal networks, with a cardinality of n. The network $G_i^j \in \boldsymbol{G_i}$ (with node prize function p_i^j and edge cost function c_i^j), captures the topological state of the system $\boldsymbol{G}$ for the j^{th} spatial region in the i^{th} temporal window.

In the context of spatio-temporal gene discovery for ASD, a network G_i^j represents the coexpression of genes during human brain development at the j^{th} brain region cluster out of n regions in total during the i^{th} time interval $[t_i, t_i + \tau]$, where $\tau \in \mathbb{N}$ is the granularity parameter.

2.5 Spatio-temporal Prize-collecting Steiner Forest Problem (ST-Steiner)

Given a spatio-temporal system $\boldsymbol{G}$, the problem is finding a minimum spatio-temporal sub-system $\boldsymbol{F} = (\boldsymbol{F_1}, \boldsymbol{F_2}, \cdots, \boldsymbol{F_T})$. $\boldsymbol{F_i} = \{F_i^1, F_i^2 \cdots F_i^n\}$ derives from the i^{th} time window $\boldsymbol{G_i}$, and $F_i^j(V_{F_i}^j, E_{F_i}^j)$ is a subgraph of graph $G_i^j(V_i^j, E_i^j)$. An optimal sub-system $\boldsymbol{F}$ minimizes the objective function

$$o(\boldsymbol{F}) = \sum_{i=1}^{T}\sum_{j=1}^{n} o_F(F_i^j) + \sum_{i=2}^{T}\sum_{j=1}^{n} \lambda_i^j \sum_{\substack{v \in V_i^j \\ v \notin V_{F_i}^j}} \phi(\alpha, v, p_i^j, \boldsymbol{F_{i-1}}) \tag{2.3}$$

where (i) $o_F(F_i^j)$ refers to the objective function for a single forest F_i^j, shown in Equation 2.2, (ii) ϕ is an artificial prize function that promotes the selection of nodes which are selected in forests of the previous time window $\boldsymbol{F_{i-1}}$, and finally, (iii) $\lambda_i^j \geq 0$ is a parameter that adjusts the impact of the artificial prize.

The artificial prize function is defined as

$$\phi(\alpha, v, p, \boldsymbol{F_c}) = p(v) \left(\sum_{F \in \boldsymbol{F_c}} \frac{I(V_F, v)}{|\boldsymbol{F_c}|} \right)^{\alpha}, \quad \begin{matrix} F = (V_F, E_F), \\ \alpha \geq 1, \end{matrix} \tag{2.4}$$

where (i) α adjusts the nonlinearity between the artificial prize for node v and the fraction of inclusion of node v among the set of subgraphs $\boldsymbol{F_c}$, and (ii) $I(V_F, v)$ is an indicator function that has value 1 if $v \in V_F$, 0 otherwise. This definition of ϕ is similar to the definition in [32], but here, each node gets an artificial prize proportional to its prize. Note that, the use of function ϕ corresponds to increasing the prize $p_i^j(v)$ by an artificial prize, for all time windows $i > 1$ and each node $v \in V_i^j$, such that $v \in V_{F_{i-1}}^j$ and $F_{i-1}^j \in \boldsymbol{F_{i-1}}$.

Chapter 3

Results

3.1 Datasets and Generation of the Spatio-temporal Networks

In order to model neurodevelopment, we use the BrainSpan microarray dataset of the Allen Brain Atlas [18] and generate a spatio-temporal system (cascade) of coexpression networks. To partition the dataset into developmental periods and clusters of brain regions, we follow the practice in [15] as described next.

Brain regions are clustered according to their similarity and four clusters are obtained: (i) primary visual cortex and superior temporal cortex (V1C) (ii) prefrontal cortex and primary motor-somatosensory cortex (PFC), (iii) striatum, hippocampal anlage/hippocampus, and amygladia (SHA), and (iv) mediodorsal nucleus of the thalamus and cerebellar cortex (MDCBC). The time windows which are associated with these brain regions: 1–3 which corresponds to early fetal period, 3–5 and 4–6 which correspond to mid-fetal periods ($\tau = 2$). Note that these time windows represent early neurodevelopment, which is an important stage for ASD. Each graph G_i^j is a spatio-temporal coexpression network, where i denotes one of the time intervals and j denotes one of the 4 brain region clusters.

In this work, a spatio-temporal window of neurodevelopment and its corresponding coexpression network is denoted by the abbreviation for its brain region cluster followed by the time window of interest, e.g. "PFC(1-3)" represents the region PFC at the specific time interval 1–3.

We report our results on the following two spatio-temporal cascades: (i) All spatial regions in time window 1–3, i.e. SHA(1-3), V1C(1-3), MDCBC(1-3) and PFC(1-3), as precursor networks for PFC(3-5) and (ii) all spatial regions in time window 1–3 as precursor networks for PFC(4-6). The target networks of interest, PFC(3-5) and PFC(4-6), are suggested as hotspots for ASD risk [15]. Furthermore, these are also the subject matter of DAWN [13], which allows us to directly compare our results to theirs.

An edge between two nodes is created if their absolute Pearson correlation coefficient $|r|$ is greater than or equal to 0.7 in the related portion of BrainSpan data. This threshold has also been used in the literature [12, 13, 15]. Each edge between a pair of genes is assigned a cost of $1 - r^2$. In order to set node prizes, as a measure of association with ASD, we make use of Transmission And De novo Association test (TADA) q-values applied on ASD mutation data. TADA is a statistical framework that assigns each gene a prior risk by integrating information from *de novo* and inherited mutations as well as small deletions [8, 9]. We obtain TADA q-values calculated on ASC WES cohort—reported on 17 sample sets consisting of 16,098 DNA samples, 3,871 ASD cases and also 9,937 ancestry-matched/parental controls [6]. We set node (gene) prizes to the negative natural logarithm of the TADA q-values. Thus, in all experiments, the prize function is identical for all networks.

3.2 Parameter Selection Procedure for ST-Steiner

The algorithm has 4 parameters for each spatio-temporal network. The first one is ω, which adjusts the number of trees in an estimated forest. This parameter is set to 0 to obtain a single tree per network, which is in line with the common

assumption in the literature that ASD genes are part of a connected, functional cluster. α is the second parameter which adjusts the artificial prize and is set to 2 in all our tests to add modest non-linearity to the artificial prize, as done in [32]. The third parameter β adjusts the trade-off between the node prizes and the edge costs. The final parameter is λ, which controls the magnitude of the artificial prize.

In order to select β, we make use of *efficiency ratio* (ε), which is defined as $\varepsilon = \frac{S \cap S_T}{S \setminus S_T}$, where S_T is a set of a ground truth (target) nodes to be covered by selecting a set of nodes S [36]. A high efficiency ratio is desirable, as it means that more target genes are covered by including less new predictions. In this work, we use the predicted set of 107 ASD risk genes in [6] as our ground truth set S_T. We experiment with two ε values and select β to yield $\varepsilon = 0.5 \pm t$ or $\varepsilon = 0.67 \pm t$, with tolerance $t = 0.05$. Note that $\varepsilon = 0.5$ means that in S, there exist two additional genes that are not in S_T for each target gene in S_T; similarly, $\varepsilon = 0.67$ means three extra genes are included in order to add two target genes. If no such network is found within a pre-specified number of runs (i.e. 10), the tolerance is relaxed to $t = 0.1$.

For each spatio-temporal network, to find a β that yields a desired level of ε, we first set λ to 0 (i.e. no artificial prize so the time dimension is ignored). Then, we perform binary search for β on the fixed interval of $[0.0, 2.0]$ to find a β that yields a desired ϵ. Note that, this procedure is performed for all coexpression networks of a given cascade.

After fixing the β, using the same binary search procedure, we select a λ that promotes the addition of a small number of genes that are selected in the previous time window. For this purpose, a selected subgraph $F'(V_{F'}, E_{F'})$ is targeted to have number of nodes $|V_{F'}| = (1 + \rho)|V_F| \pm t$, where tree-growth parameter $\rho \geq 0$ determines a new size. We experiment with $\rho = 0.05$ and 0.1, which correspond to growing the network in size by 5% and 10%, respectively. This overall procedure (first selecting β when $\lambda = 0$, and then selecting λ) is performed consecutively in temporal order for all coexpression networks. Note that for the first time window it is always the case that $\lambda = 0$, since there is no previous window.

3.3 Comparison with the State of the Art

We compare the performance of ST-Steiner with three state-of-the-art network-based ASD gene discovery algorithms which predict a module of ASD genes: NETBAG [10, 11], DAWN [12, 13] and MAGI [14].

3.3.1 Input Training Data

The data from the ASC WES cohort from [6] is input to all three methods. ST-Steiner makes use of TADA values. NETBAG utilizes a list of genetic events: We treated each gene with one or more dnLoF as if it was hit by a separate event targeting that gene only. DAWN uses TADA values as in ST-Steiner. MAGI, by design, uses *de novo* counts. In addition, MAGI uses a control cohort: We use the control data reported in and made available[1] with their paper.

3.3.2 Input Networks

ST-Steiner uses the two cascades that are explained in Section 3.1. For the other three methods, the suggested networks in the corresponding papers are used. That is, NETBAG is run with its own phenotype network. DAWN uses the PFC(3-5) and PFC(4-6) partial coexpression networks obtained from BrainSpan, as reported in [13]. MAGI is run using the STRING [16] and HPRD [17] PPI networks and the full coexpression network obtained from the BrainSpan dataset [18].

[1] https://eichlerlab.gs.washington.edu/MAGI/

3.3.3 Input Parameters and Implementation Details

ST-Steiner parameters are selected as described in Section 3.2. Here, for both cascades, we report results obtained by using $\varepsilon = 0.5$ and $\rho = 0.1$ unless stated otherwise. The ST-Steiner results obtained by using different values for these parameters are stated explicitly while denoting the result. The exact parameter values for β, λ, ε and ρ for every ST-Steiner result in this work are reported in Tables A.1, A.2 and A.3. We use a message-passing based implementation named msgsteiner(v1.3)[2] to solve PCST [28]. We fix the solver's parameters as follows: The maximum search depth to 30, noise to 0 and the reinforcement parameter to $1e-3$. The rest are set to their defaults.

NETBAG is run with its latest available code and data, dated January 2016[3]. We follow the steps presented in the README file provided with the code package. We prepare a genetic events file that lists every gene with at least one dnLoF mutation observed in the input WES data as separate events. Then, we perform 10,000 null distribution trials, as in the original paper [10]. For the gene metrics data required by the algorithm, we supply the data file that contains the average connectivity of the top 20 genes, which is included in the NETBAG code package and is used in [11]. For DAWN, since the reported gene sets in [13] are also generated with the same input WES data, we use the lists as they are. MAGI is run with its publicly available code (v0.1 Alpha[1]). We set the upper bound on the size of a module to 200. For the rest of the required parameters, values from the paper are used [14]. Optional parameters are set to their default values as supplied in their manual file. MAGI is run for 310,000+ iterations to obtain the first set of predictions (M1_Best and M1_Extended). After M1_Best is removed, MAGI is run for another 310,000+ iterations to obtain M2_Best and M2_Extended—see [14] for details.

[2]http://staff.polito.it/alfredo.braunstein/code/
[3]http://netbag.org/

3.3.4 Resulting Predictions

ST-Steiner predicts a network of 234 genes on the first cascade: This resulting network is referred to as ST-St. Every(1-3)+PFC(3-5). On the second cascade, it predicts a network of 256 genes and it is referred to as ST-St. Every(1-3)+PFC(4-6). See the gene lists in Tables A.4 and A.5, respectively. NETBAG selects a cluster of 87 genes, with p-value 0.048. The list of genes is given in Table A.6. DAWN outputs two networks provided in [13]: (i) DAWN PFC(3-5) (246 genes), (ii) DAWN PFC(4-6) (218 genes). MAGI predicts a core network with 47 genes (M1_Best—which is then extended to 104 genes (M1_ Extended—referred to as MAGI Ext1). Second core network (M2_Best—referred to as MAGI Best2) contains 19 genes which is then extended to module M2_Extended (MAGI Ext2, 80 genes). Gene lists as selected by MAGI for the first and second iterations are provided in Tables A.7 and A.8, respectively. Note that MAGI Best1 $\subset$ MAGI Ext1, but MAGI Best2 $\not\subset$ MAGI Ext2. Overlaps between the predictions are given in Table A.9.

3.3.5 Validation Experiments on Independent WES Data

We validate the predicted gene networks by each method, using the dnLoFs observed in 1,643 samples in the Simons Simplex Collection (SSC) WES cohort which are not included in the input training data—which is Autism Sequencing Consortium (ASC) WES cohort. A dnLoF mutation has a very high signal-to-noise ratio, and the genes that are hit are considered to be high risk genes. Therefore, since the two cohorts are completely independent, this is a powerful validation experiment and constitutes a benchmark for all methods, as also done in [13]. Note that we are comparing all methods when they are provided with identical training data in terms of dnLoF mutations (of ASC WES cohort): In this sense, this work is the first benchmark of the state-of-the-art network-based ASD gene discovery algorithms when the training data is kept the same. There are 251 genes that are hit in the SSC cohort, and we call them validated genes from this point on.

Table 3.1: Overlap ($\cap$) of predicted gene set S of each method with the 251 genes that have at least 1 dnLoF in 1,643 samples from [7] which are not included in the training set. Fisher's exact test is used to assess significance. The most significant result is marked in **bold**. As for the background 18,736 genes are used—the union with TADA genes in [6] and an extra gene included by MAGI.

gene set name	p-value	$\lvert\cap\rvert/\lvert S\rvert$
ST-Steiner Every(1-3)+PFC(3-5)	**6.958e-12**	**21/234**
ST-Steiner Every(1-3)+PFC(4-6)	1.827e-09	19/256
NETBAG	2.476e-06	9/87
DAWN PFC(3-5)	3.842e-08	17/246
DAWN PFC(4-6)	9.017e-10	18/218
MAGI Best1	3.679e-05	6/47
MAGI Ext1	7.865e-05	8/104
MAGI Best2	2.628e-02	2/19
MAGI Ext2	4.396e-03	5/80

Inspecting Table 3.1, we see that ST-St. Every(1-3)+PFC(3-5) returns the most significant overlap by recovering 21 of those genes with only 234 predictions. DAWN PFC(4-6) gets the second most significant result by identifying 18 validated genes with 218 predictions and ST-St. Every(1-3)+PFC(4-6) is ranked third with 19 validated genes but with 256 predictions. On the other hand, MAGI and NETBAG are not as successful as ST-Steiner or DAWN in identifying these genes.

One important observation is that despite using the same spatio-temporal windows and having high overlaps (56.8% and 67.4%) with DAWN PFC(3-5) and DAWN PFC(4-6) networks, ST-Steiner's predictions recover more validated genes. This can be attributed to both (i) using the cascaded information coming from a preceding neurodevelopmental window and (ii) ST-Steiner predicting a tree (rather than a forest), which only includes high prize low cost genes that are essential for connectivity. We explore the benefit of utilizing the time dimension further in Section 3.4.

Aside from the state-of-the-art methods described above, we also compare the same predictions of ST-Steiner with other predicted ASD-risk gene modules from the literature: NETBAG [10], AXAS [37], and coexpression based modules

from [15] and [38]. As done in [14], we take their outputs as is for these comparisons. Results are given in Table A.10. We note that none of these previously predicted modules get close to the precision of ST-Steiner.

3.3.6 Validation Experiments on ASD-related Gene Sets

While the exact set of functional circuits and a full picture for the genetic architecture of ASD is far from complete, there are some leads: We compare the overlap of predicted gene sets by each method with gene sets that are indicated as relevant to ASD in external resources. Fisher's exact test is used to assess significance. We present the results in Table 3.2 and indicate the most significant overlap value in bold.

First, we compare the overlaps of the results of each method with Category 1 and Category 2 genes as curated by Simons Foundation Autism Research Initiative (SFARI)[4]. Category 1 includes 24 ASD-associated genes with high confidence that are confirmed by independent experiments. Category 2 consists of 59 genes with strong confidence—without requiring independent replication of association. The gene lists are available in Tables A.11 and A.12, respectively, as accessed on January 15, 2018. ST-St. Every(1-3)+PFC(3-5) predicts the largest number of Category 1 genes (16 out of 24), while DAWN PFC(4-6) predicts more genes in Category 2 (21 out of 59).

Next, we consider the targets of two genes which are previously shown to be related to ASD. FMRP is a transcription factor whose targets have been shown to be associated with ASD [3]. Its targets are used as extra covariates in DAWN algorithm and are found to significantly contribute to the prediction performance [13]. Furthermore, the identified ASD-risk genes have significant overlaps with these targets [6]. RBFOX, on the other hand, is a splicing factor which was shown to be dysregulated in ASD [39, 40]. Its targets have also been demonstrated to overlap with ASD genes [6].

[4]https://gene.sfari.org/database/gene-scoring, as accessed on January 15, 2018

We compare the overlaps between the results of each method and (i) 842 FMRP targets reported in [41], (ii) 1,048 RBFOX targets reported in [39] with significant HITS-CLIP peaks, and finally (iii) 587 genes which are RBFOX targets with alternative splicing events [6, 39, 40]. For all of these target sets, ST-Steiner's predictions have the most significant overlaps compared to the sets predicted by NETBAG, DAWN and MAGI. Furthermore, ST-St. Every(1-3)+PFC(4-6) identifies 25% more FMRP targets than the closest result. This finding stands out in these set of comparisons as the most significant overlap.

Then, we focus on two among the most well-established disrupted functionalities in ASD, which are also the main theme of [6] in addition to shaping its title: In ASD, the genes related to chromatin modification and synapses are disrupted. We obtain the list of 152 histone modifier genes (HMGs) used in [6] for enrichment analyses. ST-St. Every(1-3)+PFC(3-5) identifies 12 HMGs with 234 predictions and provides the most significant result. DAWN, NETBAG and MAGI return a maximum of 9, 5, and 8 HMGs, respectively.

Finally, we construct a large set of synaptic genes by getting the union of synapse-related Gene Ontology (GO) Biological Process terms—see the caption of Table 3.2 for the list of terms. MAGI has the most significant overlap in this category, by identifying 9 genes with only 19 predictions on the core of the second module—MAGI Best2. Although ST-St. Every(1-3)+PFC(4-6) identifies 27 synaptic genes with 256 predictions, it is not as significant due to its larger gene set size.

Comparisons of ST-Steiner with other predicted clusters from the literature for ASD-related gene sets are presented in Table A.13 [10, 15, 37, 38]. Out of many modules, M1 and M2 of [37] and M16 of [38] show higher enrichment in HMGs, synaptic genes and FMRP/RBFOX targets, respectively. However, these are very large sets (containing 700, 680 and 492 genes, respectively) and show very little overlap with known or likely ASD genes: SFARI Category 1 (0, 4 and 1 overlaps, respectively), SFARI Category 2 (9, 8 and 3 overlaps) and validated genes from [7] (13, 7 and 7 overlaps).

Table 3.2: Table shows the significance of the intersection ($\cap$) of the gene set S predicted by each method and ASD-related gene sets as indicated in external resources. Significance is assessed using Fisher's exact test. As for the background 18,736 genes are used (the union genes with TADA values in [6] and 1 extra gene included by MAGI). The most significant result is marked in **bold**. The first two lists are SFARI Gene's Category 1 (24 genes) and Category 2 (59 genes) gene sets. The third list is the list of transcription factor FMRP's targets (842 genes). Fourth and the fifth gene sets are the targets of the splicing factor RBFOX: *RBFOX-peak* denotes 1,048 target genes with significant HITS-CLIP peaks and *RBFOX-splicing* denotes 587 genes RBFOX targets with alternative splicing events. Histone modifiers (152 genes) is the set of histone modifier genes used in [6]. Finally, Synaptic genes (873 genes) is the union of synapse-related-GO terms' gene sets which are GO:0021681, 0021680, 0021694, 0021692, 0060076, 0060077, 0051965, 0090129, 0032230, 0051968, 0045202, 0098794, 0098793, 0048169, 0051963, 0090128, 0032225, 0032228, 0051966, 0007271, 0001963, 0035249.

Predicted Gene Set S	SFARI Category 1		SFARI Category 2		FMRP Targets		RBFOX - peak		RBFOX - splice		Histone Modifiers		Synaptic Genes	
	p-value	$\lvert\cap\rvert/\lvert S\rvert$	p-value	$\lvert\cap\rvert/\lvert S\rvert$	p-value	$\lvert\cap\rvert/\lvert S\rvert$	p-value	$\lvert\cap\rvert/\lvert S\rvert$	p-value	$\lvert\cap\rvert/\lvert S\rvert$	p-value	$\lvert\cap\rvert/\lvert S\rvert$	p-value	$\lvert\cap\rvert/\lvert S\rvert$
ST-St. Every(1-3)+PFC(3-5)	**1.397e-25**	**16/234**	2.906e-22	19/234	3.198e-22	52/234	2.196e-09	38/234	**1.120e-11**	**31/234**	**4.709e-07**	**12/234**	1.195e-04	25/234
ST-St. Every(1-3)+PFC(4-6)	8.334e-23	15/256	5.890e-20	18/256	**8.684e-27**	**60/256**	**1.735e-11**	**44/256**	3.649e-08	27/256	7.890e-06	11/256	7.622e-05	27/256
NETBAG	5.062e-18	10/87	1.422e-13	10/87	1.750e-10	21/87	2.870e-04	14/87	8.199e-05	11/87	7.103e-04	5/87	2.409e-03	11/87
DAWN PFC(3-5)	5.479e-19	13/246	**4.242e-25**	**21/246**	3.985e-18	48/246	2.740e-09	39/246	3.774e-09	28/246	1.835e-04	9/246	3.274e-04	25/246
DAWN PFC(4-6)	7.067e-24	15/218	1.252e-19	17/218	7.762e-19	46/218	2.735e-10	38/218	2.231e-08	25/218	7.347e-05	9/218	1.744e-03	21/218
MAGI Best1	3.554e-07	4/47	9.615e-03	2/47	1.031e-08	14/47	3.943e-03	8/47	5.946e-04	7/47	2.100e-06	6/47	2.813e-01	4/47
MAGI Ext1	1.866e-07	5/104	3.220e-04	4/104	3.600e-06	17/104	1.500e-02	12/104	2.137e-05	13/104	2.072e-06	8/104	9.394e-03	11/104
MAGI Best2	2.652e-04	2/19	1.000e+00	0/19	2.321e-12	12/19	3.127e-03	5/19	2.512e-03	4/19	1.000e+00	0/19	**6.258e-08**	**9/19**
MAGI Ext2	3.064e-06	4/80	2.049e-03	3/80	1.165e-08	18/80	1.442e-03	12/80	3.735e-05	11/80	1.000e+00	0/80	3.862e-06	15/80

3.3.7 KEGG Pathway and GO Enrichment Analyses

We compare the performance of the methods with respect to their enrichment in Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and Gene Ontology (GO) Biological Process (BP) terms (i) associated with known or likely ASD genes and (ii) that the predictions of the methods are enriched in.

The first set of experiments are carried out on the KEGG pathways and GO BP terms that significantly overlap with SFARI Category 1 and 2 genes (83 genes). We obtain 6 such KEGG pathways and 15 such GO BPs by utilizing Enrichr tool[5] with its gene-set libraries KEGG 2016 and GO Biological Process 2017 [42]. Using Fisher's exact test, we assess the significance of the overlap of the predictions of each method (ST-Steiner, NETBAG, DAWN and MAGI) with these pathways and terms. Results indicate that ST-Steiner has the most significant overlap in 3 out of 6 associated KEGG pathways and in 5 out of 15 associated GO terms, whereas MAGI outweighs in only 1 KEGG pathway and in 7 GO terms—see Figures B.2 and B.3 for KEGG and GO comparisons, respectively.

The second set of experiments focus on the top-3 most enriched KEGG pathways and GO BP terms for each method. We compare every method on the union of these pathways and terms, similar to [14]. Three notable findings related to ASD in these figures are for the KEGG pathways (i) *GABAergic Synapse* (ST-Steiner is better), (ii) *WNT Signaling Pathway* (MAGI is better) and for the GO BP (iii) *Notch Signaling Pathway* (MAGI is slightly better than ST-Steiner)—see Figures B.4 and B.5 for KEGG and GO comparisons, respectively.

MAGI's GO BP enrichment performance is better compared to other methods. ST-Steiner picks more relevant genes in related KEGG pathways. In this sense, MAGI and ST-Steiner perform comparably.

[5]https://amp.pharm.mssm.edu/Enrichr/, via the Application Programming Interface (API) documented at https://amp.pharm.mssm.edu/Enrichr/help#api

3.4 Alternative Uses of Spatial and Temporal Information

In this section, we compare the results of ST-Steiner under different settings by discussing how the parameters and number of precursor windows affect the performance. Then, we evaluate the benefit of using temporal hierarchy. We use the same input data and parameters as described in Section 3.3. We consider two new cascades: (i) ST-St. PFC(1-3)+(3-5), where PFC(1-3) is the precursor window for PFC(3-5) and (ii) ST-St. PFC(1-3)+(4-6)—please see Tables A.14 and A.15 for the detailed results on all upcoming comparisons in this section, and Tables A.16–A.21 for the referenced ST-Steiner predictions.

3.4.1 Effect of Parameters

First, we run ST-Steiner by setting ρ to 0.05, which decreases the number of time induced genes. Secondly, we increase ε to 0.67, which shrinks the size of the network. In both cases, the number of validated genes that are identified do not improve. Also, the number of genes predicted in our chosen setting is comparable with the literature (in particular, with DAWN). Hence, we fix our setting to $\varepsilon = 0.5$ and $\rho = 0.1$.

3.4.2 Effect of Using Multiple Spatial Regions

We assess the effect of changing the number of spatial (brain) regions used as precursors. Results show that using a single precursor coexpression network rather than multiple reduces the overlaps with the lists from external sources. Therefore, we prefer using multiple precursors.

3.4.3 ST-Steiner vs. Union of Independent Results

One way of joining information in two temporal windows is to analyze them independently and then, take the union of the selected genes. Yet, this would not predict a connected network of genes, which defies the initial assumption that the genes are part of a functional cluster. Still, we run ST-Steiner on PFC(1-3), PFC(3-5) and PFC(4-6) independently, with no time effect ($\lambda = 0,\ \rho = 0,\ \beta$ is selected such that $\varepsilon = 0.5$) and then, obtain ST-St. PFC(1-3) $\cup$ ST-St. PFC(4-6). This results in a gene set of 293 genes, which does not predict any new validated genes, when compared to ST-St. PFC(1-3)+(4-6). ST-St. PFC(1-3) $\cup$ ST-St. PFC(3-5) has one more validated gene retained in the union set which includes 275 genes. However, due to its large size, it not as significant as ST-St. PFC(1-3)+(3-5) ($p = 1.472e-10$ vs. $7.0e-12$). Thus, our scheme is better in terms of the prediction performance and also for returning biologically interpretable results as the gene set is connected.

3.4.4 ST-Steiner vs. Single Analysis with Coarser Granularity

To combine information in two windows, another alternative is to perform the analysis on a coexpression network that spans both windows (i.e. instead of PFC(1-3) and PFC(4-6), use PFC(1-6)). To assess this approach, we obtain network ST-St. PFC(1-6) with no time effect ($\lambda = 0,\ \rho = 0,\ \beta$ is selected such that $\varepsilon = 0.5$). This network contains 197 genes and identifies only 14 genes that are validated ($p = 4.480e-7$). In contrast, our stepwise approach ST-St. PFC(1-3)+(4-6) identifies 19 validated genes by predicting 256 genes ($p = 1.836e-9$). Similarly, ST-St. PFC(1-5) with no time effect also predicts only 14 validated genes and ST-St. PFC(1-3)+(3-5) identifies 21 validated genes. These indicate that finer granularity coexpression networks are more potent.

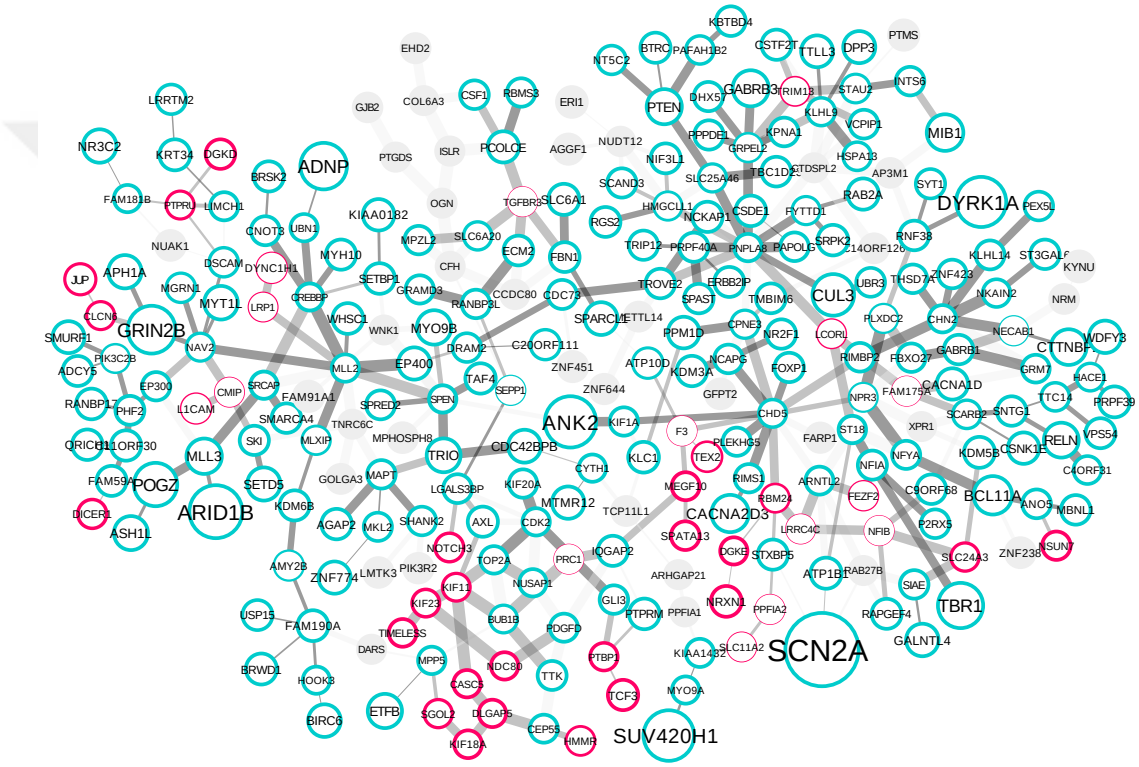

Figure 3.1: Figure visualizes ST-St. PFC(1-3)+(3-5) network which is laid over ST-St. PFC(3-5)+10%. Turquoise bordered genes are common in ST-St. PFC(1-3)+(3-5) and ST-St. PFC(3-5)+10%. Pink bordered genes are only in ST-St. PFC(1-3)+(3-5), highlighting the effect of using the temporal dimension and information coming from PFC(1-3). Gray genes are present only in PFC(3-5)+10% and are not included by ST-St. PFC(1-3)+(3-5). Size of a node indicates its significance w.r.t. its TADA q-value in [6] (the larger, the more significant). The thickness of the border of a node indicates its robustness. The thickness of an edges represents the correlation coefficient between the gene pair (the thicker, the higher). Visualized using the CoSE layout [43] in Cytoscape [44].

3.4.5 ST-Steiner vs. No Temporal Effect

To see if information coming from a precursor improves the performance, we compare ST-St. PFC(1-3)+(3-5) and ST-St. PFC(1-3)+(4-6) with ST-Steiner results only on PFC(3-5) and PFC(4-6), respectively. For this purpose, we obtain ST-St. PFC(3-5) and ST-St. PFC(4-6) independently, with no time effect ($\lambda = 0$, $\rho = 0$, β is selected such that $\varepsilon = 0.5$). The former identifies 20 validated genes by making 211 predictions ($p = 8.272e - 12$), while ST-St. PFC(1-3)+(3-5) identifies 21 validated genes by choosing 234 genes ($p = 7.0e - 12$). PFC(4-6) identifies 18 validated genes by predicting 235 genes ($p = 3.018e - 9$), whereas ST-St. PFC(1-3)+(4-6) identifies 19 validated genes by predicting 256 ($p = 1.836e - 9$). The increase in the number of validated genes demonstrate that ST-Steiner can leverage temporal information.

Due to ρ, the network size is increased to include more genes by considering the time dimension. Therefore, the above comparison is slightly unfair. What we want is to see if the leveraged information comes just from network getting bigger or is due to the spatio-temporal analysis. To investigate this, we obtain two additional results: ST-St. PFC(3-5)+10% and ST-St. PFC(4-6)+10% with the following parameters: $\lambda = 0$, β is selected such that they have comparable sizes to PFC(1-3)+(3-5) and PFC(1-3)+(4-6), respectively. These results contain 230 and 266 genes, respectively. We see that PFC(3-5)+10% and ST-St. PFC(4-6)+10% identify 20 and 19 genes respectively, and are less significant. This suggests that the prize mechanism can successfully promotes selecting ASD genes.

We also perform a robustness analysis similar to the one in [12]. That is, we remove the genetic signal from 1/30 randomly selected genes and rerun ST-Steiner. This is repeated 30 times for each fold to see how frequently each gene is selected. The visualization of ST-St. PFC(1-3)+(3-5) in comparison to ST-St. PFC(3-5)+10% is illustrated in Figure 3.1 along with these robustness values—see Figure B.1 for a similar comparison of ST-St. PFC(1-3)+(4-6) and PFC(4-6)+10%.

3.5 Time and Memory Performance

All tests are performed on a server with 2 CPUs and 20 cores in total (Intel Xeon Processor E5-2650 v3) with 256GB memory. ST-Steiner's running time scales with the size and connectivity of the input network as well as the number of networks considered in the cascade. The largest network we experiment with is PFC(1-3), which has 11,891 genes and 4,987,350 connections: ST-Steiner took $\sim$6 hours with peak memory usage of 12 GB on maximum 40 threads.

Chapter 4

Discussion

We first argue that the results indicate ST-Steiner is capable of predicting new risk genes, by using temporal information coming from precursor time windows. Then, we inspect the predicted gene lists by focusing on specific genes, and discuss whether ST-Steiner's selections are traceable to the precursor windows.

4.1 On ST-Steiner Predicting New Genes with Temporal Information

We evaluate genes that are included both in ST-St. PFC(1-3)+(3-5) and ST-St. PFC(1-3)+(4-6), but excluded from PFC(3-5)+10% or PFC(4-6)+10%, i.e. nodes that are predicted due to temporal information. There are 9 such genes and we focus on the 6 which are detected neither by DAWN nor by MAGI. Furthermore, these genes are not pointed out in WES studies: They all have low scores in TADA, and ranked at most as Category 4 in SFARI Gene which stands for "minimal evidence", if included at all—see Table A.22 for the list of 339 genes that are classified as Category 4.

The first one is KIF23, a kinesin family member protein. Kinesins are known to transport cargo to dendritic spines undergoing synaptic plasticity over microtubules [45]. ST-St. PFC(1-3)+(3-5) selects 5 KIF genes. Note that TADA q-values are available only for 6 KIF genes. Kinesins are also known to play a role in organization of spindle microtubules during mitosis [46]. We detect two more genes, NDC80 and SGOL2, which take part in kinetochore-microtubule attachment during cell division. These two genes and above-mentioned KIF genes closely interact in Figure 3.1. We do not find any evidence of association with ASD for these 3 genes. These results suggest that an early disruption of these circuitries can be related to ASD.

The following genes have subtle clues in the literature. The fourth gene MEGF10 is also related to cell proliferation and adhesion, and recently variants in its regulatory region have tied it to ASD [47]. The fifth gene is CMIP, which is a signaling protein previously associated with language delay. There are two studies (one very recent) indicating haploinsufficiency of CMIP leads to ASD [48, 49]. Selection of CMIP in ST-St. PFC(1-3)+(3-5) enables inclusion of another gene L1CAM with a link to CMIP, which is related to neurite outgrowth and cell migration. The final gene, which is time induced (ironically), is TIMELESS. It is related to circadian rhythm, and it is also linked to ASD before in a study [50].

4.2 On Predicted Genes Providing Traceable Information

Here, we focus on genes that are relatively more established in literature compared to the genes in the previous subsection. However, them not being selected by ST-Steiner without the time information allows us to trace back the information and show that ST-Steiner is able to also return more interpretable results.

The first two genes are NOTCH2 and NOTCH3, membrane receptors of NOTCH signaling pathway. The former gene is selected in both ST-St. PFC(1-3)+(3-5) and ST-St. PFC(1-3)+(4-6), the latter is selected in ST-St. PFC(1-3)+(4-6). This pathway is important for neuron differentiation [51]. More importantly, it is active during embryogenesis (time-point 1) [52]. Thus, its disruption is expected to have a cascaded effect during time window 3–5.

Another gene that is retained is TCF3, which has been highlighted in [6] as one of the few hub transcription factors to regulate many ASD-risk genes along with many histone modifiers. Similar to NOTCH signaling pathway, which regulates neuron differentiation, TCF3 is found to promote differentiation in embryonic stem cells [53, 54]. More importantly, it represses neuron differentiation in neural precursor cells [55, 56], which again corresponds to the time window 1–3. Thus, in line with our hypothesis that clustering of genes is spatio-temporal, ST-Steiner predicts these genes by considering the effect of the earlier time window. This also means that one can trace back the information source which enables the selection of these genes: This adds further interpretability to our results.

Chapter 5

Conclusion

ASD is a common neurodevelopmental disorder which is a life-long challenge for many families all around the world. Gaining an understanding of the cause of ASD and opening a way to the development of new treatments would certainly have a major impact on the lives of many. Even though we have long ways to go to understand the etiology of the disorder, network-based gene discovery algorithms have proven useful for gene discovery. In this work, we propose a novel ASD gene discovery algorithm that for the first time models the cascading effect of disrupted functional circuits in neurodevelopment.

We show that ST-Steiner is more precise than the state-of-the-art methods. While other methods output a graph of genes, ST-Steiner outputs a tree/forest. That is, no multiple paths exist between any pair of genes and some genes can be left out if they are not essential for connectivity due to having low prior risk or high-cost edges. Precision is also achieved via rewarding selected genes from earlier periods, which makes the algorithm more confident on the predictions it makes. We show that once the temporal information is employed, predictive power increases, which supports our hypothesis that the clustering of genes is spatio-temporal rather than static.

We restrict our tests to ASD and only to well known ASD-related spatio-temporal windows (PFC(3-5) and PFC(4-6)) for a comprehensive comparison with other methods. However, ST-Steiner can be used for any disorder given a progression model and corresponding cascade of coexpression networks. One example is intellectual disability (ID). As a matter of fact, ID and ASD are known to be highly comorbid and to share a genetic component [57]. It is an interesting problem to identify periods of divergence and convergence using ST-Steiner. One other target could be neurodegenerative disorders such as Alzheimer's for which adulthood Brain coexpression networks from BrainSpan can be used.

We also think there is still room for improvement in ST-Steiner's formulation. Considering that PPI networks give MAGI an orthogonal source of information, one future extension direction would be the incorporation of information from this source. Note that the goal of this work is to show that the clustering of ASD genes are dynamic rather than static. This is the reason why we do not consider PPIs, but restrict the problem definition and our tests to observe the cascaded effects of dysregulation in earlier time windows.

Bibliography

[1] Autism and D. D. M. N. S. Y. . P. Investigators, "Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, united states, 2010," *Morbidity and Mortality Weekly Report: Surveillance Summaries*, vol. 63, no. 2, pp. 1–21, 2014.

[2] B. J. O'Roak, L. Vives, S. Girirajan, E. Karakoc, N. Krumm, B. P. Coe, R. Levy, A. Ko, C. Lee, J. D. Smith, E. H. Turner, I. B. Stanaway, B. Vernot, M. Malig, C. Baker, B. Reilly, J. M. Akey, E. Borenstein, M. J. Rieder, D. A. Nickerson, R. Bernier, J. Shendure, and E. E. Eichler, "Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations," *Nature*, vol. 485, 2012.

[3] I. Iossifov, M. Ronemus, D. Levy, Z. Wang, I. Hakker, J. Rosenbaum, B. Yamrom, Y. H. Lee, G. Narzisi, A. Leotta, J. Kendall, E. Grabowska, B. Ma, S. Marks, L. Rodgers, A. Stepansky, J. Troge, P. Andrews, M. Bekritsky, K. Pradhan, E. Ghiban, M. Kramer, J. Parla, R. Demeter, L. L. Fulton, R. S. Fulton, V. J. Magrini, K. Ye, J. C. Darnell, and R. B. Darnell, "De novo gene disruptions in children on the autistic spectrum," *Neuron*, vol. 74, 2012.

[4] B. M. Neale, Y. Kou, L. Liu, A. Ma'ayan, K. E. Samocha, A. Sabo, C. F. Lin, C. Stevens, L. S. Wang, V. Makarov, P. Polak, S. Yoon, J. Maguire, E. L. Crawford, N. G. Campbell, E. T. Geller, O. Valladares, C. Schafer, H. Liu, T. Zhao, G. Cai, J. Lihm, R. Dannenfelser, O. Jabado, Z. Peralta, U. Nagaswamy, D. Muzny, J. G. Reid, I. Newsham, and Y. Wu, "Patterns

and rates of exonic de novo mutations in autism spectrum disorders," *Nature*, vol. 485, 2012.

[5] S. J. Sanders, M. T. Murtha, A. R. Gupta, J. D. Murdoch, M. J. Raubeson, A. J. Willsey, A. G. Ercan-Sencicek, N. M. DiLullo, N. N. Parikshak, J. L. Stein, M. F. Walker, G. T. Ober, N. A. Teran, Y. Song, P. El-Fishawy, R. C. Murtha, M. Choi, J. D. Overton, R. D. Bjornson, N. J. Carriero, K. A. Meyer, K. Bilguvar, S. M. Mane, N. Sestan, R. P. Lifton, M. Gunel, K. Roeder, D. H. Geschwind, B. Devlin, and M. W. State, "De novo mutations revealed by whole-exome sequencing are strongly associated with autism," *Nature*, vol. 485, 2012.

[6] S. De Rubeis, X. He, A. P. Goldberg, C. S. Poultney, K. Samocha, A. E. Cicek, Y. Kou, L. Liu, M. Fromer, S. Walker, *et al.*, "Synaptic, transcriptional and chromatin genes disrupted in autism," *Nature*, vol. 515, no. 7526, pp. 209–215, 2014.

[7] I. Iossifov, B. J. O'Roak, S. J. Sanders, M. Ronemus, N. Krumm, D. Levy, H. A. Stessman, K. T. Witherspoon, L. Vives, K. E. Patterson, *et al.*, "The contribution of de novo coding mutations to autism spectrum disorder," *Nature*, vol. 515, no. 7526, pp. 216–221, 2014.

[8] S. J. Sanders, X. He, A. J. Willsey, A. G. Ercan-Sencicek, K. E. Samocha, A. E. Cicek, M. T. Murtha, V. H. Bal, S. L. Bishop, S. Dong, *et al.*, "Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci," *Neuron*, vol. 87, no. 6, pp. 1215–1233, 2015.

[9] X. He, S. J. Sanders, L. Liu, S. D. Rubeis, E. T. Lim, J. S. Sutcliffe, G. D. Schellenberg, R. A. Gibbs, M. J. Daly, J. D. Buxbaum, and et al., "Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes," *PLoS Genetics*, vol. 9, no. 8, 2013.

[10] S. R. Gilman, I. Iossifov, D. Levy, M. Ronemus, M. Wigler, and D. Vitkup, "Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses," *Neuron*, vol. 70, no. 5, pp. 898–907, 2011.

[11] S. R. Gilman, J. Chang, B. Xu, T. S. Bawa, J. A. Gogos, M. Karayiorgou, and D. Vitkup, "Diverse types of genetic variation converge on functional gene networks involved in schizophrenia," *Nature neuroscience*, vol. 15, no. 12, pp. 1723–1728, 2012.

[12] L. Liu, J. Lei, S. J. Sanders, A. J. Willsey, Y. Kou, A. E. Cicek, L. Klei, C. Lu, X. He, M. Li, R. A. Muhle, A. Ma'ayan, J. P. Noonan, N. Šestan, K. A. McFadden, M. W. State, J. D. Buxbaum, B. Devlin, and K. Roeder, "Dawn: a framework to identify autism genes and subnetworks using gene expression and genetics," *Molecular Autism*, vol. 5, p. 22, Mar 2014.

[13] L. Liu, J. Lei, and K. Roeder, "Network assisted analysis to reveal the genetic basis of autism," *The annals of applied statistics*, vol. 9, no. 3, p. 1571, 2015.

[14] F. Hormozdiari, O. Penn, E. Borenstein, and E. E. Eichler, "The discovery of integrated gene networks for autism and related disorders," *Genome Research*, vol. 25, pp. 142–154, May 2014.

[15] A. J. Willsey, S. J. Sanders, M. Li, S. Dong, A. T. Tebbenkamp, R. Muhle, S. K. Reilly, L. Lin, S. Fertuzinhos, J. A. Miller, M. Murtha, C. Bichsel, W. Niu, J. Cotney, A. G. Ercan-Sencicek, J. Gockley, A. R. Gupta, W. Han, X. He, E. J. Hoffman, L. Klei, J. Lei, W. Liu, L. Liu, C. Lu, X. Xu, Y. Zhu, S. M. Mane, E. S. Lein, and L. Wei, "Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism," *Cell*, vol. 155, 2013.

[16] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, *et al.*, "The string database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D561–D568, 2010.

[17] T. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, *et al.*, "Human protein reference database-2009 update," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D767–D772, 2008.

[18] S. M. Sunkin, L. Ng, C. Lau, T. Dolbeare, T. L. Gilbert, C. L. Thompson, M. Hawrylycz, and C. Dang, "Allen brain atlas: an integrated spatio-temporal portal for exploring the central nervous system," *Nucleic acids research*, vol. 41, no. D1, pp. D996–D1008, 2012.

[19] A. Krishnan, R. Zhang, V. Yao, C. L. Theesfeld, A. K. Wong, A. Tadych, N. Volfovsky, A. Packer, A. Lash, and O. G. Troyanskaya, "Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder," *Nature neuroscience*, vol. 19, no. 11, pp. 1454–1462, 2016.

[20] C. S. Greene, A. Krishnan, A. K. Wong, E. Ricciotti, R. A. Zelaya, D. S. Himmelstein, R. Zhang, B. M. Hartmann, E. Zaslavsky, S. C. Sealfon, *et al.*, "Understanding multicellular function and disease with human tissue-specific networks," *Nature genetics*, vol. 47, no. 6, p. 569, 2015.

[21] S. S.-H. Wang, A. D. Kloth, and A. Badura, "The cerebellum, sensitive periods, and autism," *Neuron*, vol. 83, no. 3, pp. 518–532, 2014.

[22] L. de la Torre-Ubieta, H. Won, J. L. Stein, and D. H. Geschwind, "Advancing the understanding of autism disease mechanisms through genetics," *Nature medicine*, vol. 22, no. 4, pp. 345–361, 2016.

[23] S. E. Dreyfus and R. A. Wagner, "The steiner problem in graphs," *Networks*, vol. 1, no. 3, pp. 195–207, 1971.

[24] P. Winter, "Steiner problem in networks: a survey," *Networks*, vol. 17, no. 2, pp. 129–167, 1987.

[25] M. S. Scott, T. Perkins, S. Bunnell, F. Pepin, D. Y. Thomas, and M. Hallett, "Identifying regulatory subnetworks for a set of genes," *Molecular & Cellular Proteomics*, vol. 4, no. 5, pp. 683–692, 2005.

[26] M. T. Dittrich, G. W. Klau, A. Rosenwald, T. Dandekar, and T. Müller, "Identifying functional modules in protein–protein interaction networks: an integrated exact approach," *Bioinformatics*, vol. 24, no. 13, pp. i223–i231, 2008.

[27] S.-s. C. Huang and E. Fraenkel, "Integration of proteomic, transcriptional, and interactome data reveals hidden signaling components," *Science signaling*, vol. 2, no. 81, p. ra40, 2009.

[28] M. Bailly-Bechet, C. Borgs, A. Braunstein, J. Chayes, A. Dagkessamanskaia, J.-M. François, and R. Zecchina, "Finding undetected protein associations in cell signaling by belief propagation," *Proceedings of the National Academy of Sciences*, vol. 108, no. 2, pp. 882–887, 2011.

[29] K. Faust, P. Dupont, J. Callut, and J. Van Helden, "Pathway discovery in metabolic networks by subgraph extraction," *Bioinformatics*, vol. 26, no. 9, pp. 1211–1218, 2010.

[30] Y. Sun, P. N. Hameed, K. Verspoor, and S. Halgamuge, "A physarum-inspired prize-collecting steiner tree approach to identify subnetworks for drug repositioning," *BMC systems biology*, vol. 10, no. 5, p. 128, 2016.

[31] N. Tuncbag, A. Braunstein, A. Pagnani, S.-S. C. Huang, J. Chayes, C. Borgs, R. Zecchina, and E. Fraenkel, "Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem," *Journal of computational biology*, vol. 20, no. 2, pp. 124–136, 2013.

[32] A. Gitter, A. Braunstein, A. Pagnani, C. Baldassi, C. Borgs, J. T. Chayes, R. Zecchina, and E. Fraenkel, "Sharing information to reconstruct patient-specific pathways in heterogeneous diseases," in *Biocomputing 2014: Proceedings of the Pacific Symposium, Kohala Coast, Hawaii, USA, January 3-7, 2014* (R. B. Altman, A. K. Dunker, L. Hunter, T. E. Klein, and M. D. Ritchie, eds.), pp. 39–50, 2014.

[33] A. Khodaverdian, B. Weitz, J. Wu, and N. Yosef, "Steiner network problems on temporal graphs," *arXiv preprint arXiv:1609.04918*, 2016.

[34] G. Budak, O. Eren Ozsoy, Y. Aydin Son, T. Can, and N. Tuncbag, "Reconstruction of the temporal signaling network in salmonella-infected human cells," *Frontiers in microbiology*, vol. 6, p. 730, 2015.

[35] N. Tuncbag, A. Braunstein, A. Pagnani, S.-S. C. Huang, J. Chayes, C. Borgs, R. Zecchina, and E. Fraenkel, "Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem," *Lecture Notes in Computer Science Research in Computational Molecular Biology*, pp. 287–301, 2012.

[36] S.-s. C. Huang, D. C. Clarke, S. J. C. Gosline, A. Labadorf, C. R. Chouinard, W. Gordon, D. A. Lauffenburger, and E. Fraenkel, "Linking Proteomic and Transcriptional Data through the Interactome and Epigenome Reveals a Map of Oncogene-induced Signaling," *PLoS Comput Biol*, vol. 9, pp. e1002887+, Feb. 2013.

[37] A. Cristino, S. Williams, Z. Hawi, J. An, M. Bellgrove, C. Schwartz, *et al.*, "Neurodevelopmental and neuropsychiatric disorders represent an interconnected molecular system," *Molecular Psychiatry*, vol. 19, pp. 294–301, 2014.

[38] N. N. Parikshak, R. Luo, A. Zhang, H. Won, J. K. Lowe, V. Chandran, S. Horvath, and D. H. Geschwind, "Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism," *Cell*, vol. 155, 2013.

[39] S. M. Weyn-Vanhentenryck, A. Mele, Q. Yan, S. Sun, N. Farny, Z. Zhang, C. Xue, M. Herre, P. A. Silver, M. Q. Zhang, *et al.*, "Hits-clip and integrative modeling define the rbfox splicing-regulatory network linked to brain development and autism," *Cell reports*, vol. 6, no. 6, pp. 1139–1152, 2014.

[40] I. Voineagu, X. Wang, P. Johnston, J. K. Lowe, Y. Tian, S. Horvath, J. Mill, R. M. Cantor, B. J. Blencowe, and D. H. Geschwind, "Transcriptomic analysis of autistic brain reveals convergent molecular pathology," *Nature*, vol. 474, no. 7351, pp. 380–384, 2011.

[41] J. C. Darnell, S. J. Van Driesche, C. Zhang, K. Y. S. Hung, A. Mele, C. E. Fraser, E. F. Stone, C. Chen, J. J. Fak, S. W. Chi, *et al.*, "Fmrp stalls ribosomal translocation on mrnas linked to synaptic function and autism," *Cell*, vol. 146, no. 2, pp. 247–261, 2011.

[42] E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan, "Enrichr: interactive and collaborative html5 gene list enrichment analysis tool," *BMC Bioinformatics*, vol. 14, p. 128, Apr 2013.

[43] U. Dogrusoz, E. Giral, A. Cetintas, A. Civril, and E. Demir, "A layout algorithm for undirected compound graphs," *Information Sciences*, vol. 179, no. 7, pp. 980–994, 2009.

[44] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[45] D. P. McVicker, A. M. Awe, K. E. Richters, R. L. Wilson, D. A. Cowdrey, X. Hu, E. R. Chapman, and E. W. Dent, "Transport of a kinesin-cargo pair along microtubules into dendritic spines undergoing synaptic plasticity," *Nature communications*, vol. 7, p. 12741, 2016.

[46] S. C. Ems-McClung and C. E. Walczak, "Kinesin-13s in mitosis: Key players in the spatial and temporal organization of spindle microtubules," in *Seminars in cell & developmental biology*, vol. 21, pp. 276–282, Elsevier, 2010.

[47] Z. Wu, J. Qin, Y. You, Y. Ma, M. Jia, L. Wang, T. Lu, W. Yue, Y. Ruan, D. Zhang, *et al.*, "Genetic variants in the transcription regulatory region of megf10 are associated with autism in chinese han population," *Scientific reports*, vol. 7, no. 1, p. 2292, 2017.

[48] N. Van der Aa, G. Vandeweyer, E. Reyniers, S. Kenis, L. Dom, G. Mortier, L. Rooms, and R. F. Kooy, "Haploinsufficiency of cmip in a girl with autism spectrum disorder and developmental delay due to a de novo deletion on chromosome 16q23. 2," *Autism Research*, vol. 5, no. 4, pp. 277–281, 2012.

[49] M. Luo, J. Fan, T. L. Wenger, M. H. Harr, M. Racobaldo, S. Mulchandani, H. Dubbs, E. H. Zackai, N. B. Spinner, and L. K. Conlin, "cmip haploinsufficiency in two patients with autism spectrum disorder and co-occurring gastrointestinal issues," *American Journal of Medical Genetics Part A*, vol. 173, no. 8, pp. 2101–2107, 2017.

[50] Z. Yang, A. Matsumoto, K. Nakayama, E. F. Jimbo, K. Kojima, K.-i. Nagata, S. Iwamoto, and T. Yamagata, "Circadian-relevant genes are highly polymorphic in autism spectrum disorder patients," *Brain and Development*, vol. 38, no. 1, pp. 91–99, 2016.

[51] E. Cau and P. Blader, "Notch activity in the nervous system: to switch or not switch?," *Neural development*, vol. 4, no. 1, p. 36, 2009.

[52] J. Wolter, "The notch signaling pathway in embryogenesis," *Embryo Project Encyclopedia*, 2013.

[53] B. J. Merrill, U. Gat, R. DasGupta, and E. Fuchs, "Tcf3 and lef1 regulate lineage differentiation of multipotent stem cells in skin," *Genes & development*, vol. 15, no. 13, pp. 1688–1705, 2001.

[54] H. Nguyen, M. Rendl, and E. Fuchs, "Tcf3 governs stem cell features and represses cell fate determination in skin," *Cell*, vol. 127, no. 1, pp. 171–183, 2006.

[55] A. Kuwahara, H. Sakai, Y. Xu, Y. Itoh, Y. Hirabayashi, and Y. Gotoh, "Tcf3 represses wnt–β-catenin signaling and maintains neural stem cell population during neocortical development," *PloS one*, vol. 9, no. 5, p. e94408, 2014.

[56] T. Ohtsuka, H. Shimojo, M. Matsunaga, N. Watanabe, K. Kometani, N. Minato, and R. Kageyama, "Gene expression profiling of neural stem cells and identification of regulators of neural differentiation during cortical development," *Stem Cells*, vol. 29, no. 11, pp. 1817–1828, 2011.

[57] J. McCarthy, "Children with autism spectrum disorders and intellectual disability," *Current Opinion in Psychiatry*, vol. 20, no. 5, pp. 472–476, 2007.

Appendix A

Supplementary Tables

Table A.1: This table shows the parameter values used to obtain the ST-Steiner runs which consider a single spatio-temporal window (coexpression network). The β value corresponding to the targeted ϵ value is shown. For a given ϵ, after selecting a β, the realized ϵ value is given along with the resulting gene set size. Please see Section 3.2 for explanation on the parameters and their selection procedures.

gene list name	ε targeted	β selected	ε realized	size
ST-St. PFC(1-5)	0.5	0.375	0.517986	211
ST-St. PFC(1-6)	0.5	0.398438	0.503817	197
ST-St. PFC(3-5)	0.5	0.48	0.496454	211
ST-St. PFC(4-6)	0.5	0.5	0.525974	235
ST-St. PFC(3-5) ε0.67	0.67	0.41	0.666667	165
ST-St. PFC(4-6) ε0.67	0.67	0.460938	0.666667	195
ST-St. PFC(1-3) ε0.67	0.67	0.09375	0.674419	72
ST-St. PFC(1-3)	0.5	0.125	0.513158	115
ST-St. V1C(1-3)	0.5	0.15625	0.586207	138
ST-St. SHA(1-3)	0.5	0.25	0.566372	177
ST-St. MDCBC(1-3)	0.5	0.0258789	0.5	39

Table A.2: This table shows the parameter values used to obtain the ST-Steiner runs which consider a cascade of spatio-temporal windows (coexpression networks). Given input β value found by ST-Steiner on this coexpression network without the time dimension (see Table A.1) and a target ρ value, the selected λ value along with the realized ρ value are shown. Last column shows the resulting gene set size. Please see Section 3.2 for explanation on the parameters and their selection procedures.

gene list name	input β	ρ targeted	λ selected	ρ realized	size
ST-St. Every(1-3)+PFC(3-5)	0.48	0.1	0.59375	0.109005	234
ST-St. Every(1-3)+PFC(4-6)	0.5	0.1	0.75	0.0893617	256
ST-St. PFC(1-3)+(3-5)	0.48	0.1	0.375	0.104265	233
ST-St. PFC(1-3)+(4-6)	0.5	0.1	0.4375	0.110638	261
ST-St. PFC(1-3)+(3-5) ρ0.05	0.48	0.05	0.25	0.056872	223
ST-St. PFC(1-3)+(4-6) ρ0.05	0.5	0.05	0.375	0.0468085	246
ST-St. PFC(1-3)+(3-5) ε0.67, ρ0.05	0.41	0.05	0.289062	0.0545455	174
ST-St. PFC(1-3)+(4-6) ε0.67, ρ0.05	0.460938	0.05	0.09375	0.0512821	205
ST-St. PFC(1-3)+(4-6) ε0.67, ρ0.1	0.460938	0.1	0.25	0.0974359	214
ST-St. PFC(1-3)+(3-5) ε0.67, ρ0.1	0.41	0.1	0.5	0.10303	182

Table A.3: The selected β values used to enlarge ST-St. PFC(3-5) and ST-St. PFC(4-6) networks by 10% from their original sizes. Targeted and realized gene set sizes are also shown. See Table A.1 for the parameters used to obtain ST-St. PFC(3-5) and ST-St. PFC(4-6).

gene list name	original size	size targeted	β selected	size realized
ST-St. PFC(3-5)+10%	211	233	0.525	230
ST-St. PFC(4-6)+10%	235	259	0.53125	266

Table A.4: Gene list for ST-St. Every(1-3)+PFC(3-5), which is the result of ST-Steiner run on PFC(3-5) coexpression network with the precursor coexpression networks MDCBC(1-3), PFC(1-3), SHA(1-3) and V1C(1-3). The predicted gene sets for these precursors are given in Tables A.16, A.17, A.18 and A.19, respectively. This list contains 234 genes.

ADCY5	C14ORF126	DHX57	GLI3	KIF20A	MPZL2	PAPOLG	RBM24	SPAST	TRIP12
ADNP	C20ORF111	DLGAP5	GOLGA3	KIF23	MTMR12	PBX1	RBMS3	SPEN	TROVE2
AGAP2	CACNA1D	DOCK1	GPR124	KLC1	MYH10	PCOLCE	RELN	SPRED2	TTC14
AGGF1	CACNA2D3	DPP3	GRAMD3	KLHL9	MYO9A	PDGFD	RGS2	SRCAP	TTK
AMY2B	CASC5	DRAM2	GRIA2	KLHL14	MYO9B	PEX5L	RIMBP2	SRPK2	TTLL3
ANK2	CDC42BPB	DSCAM	GRIN2B	KPNA1	MYT1L	PHF2	RIMS1	ST3GAL6	UBN1
ANO5	CDC73	DYRK1A	GRM7	KRT34	NAV2	PIK3C2B	RNF38	ST18	UBR3
AP3M1	CEP55	ECM2	GRPEL2	KYNU	NCAPG	PLEKHG5	SCAND3	STAU2	USP15
APH1A	CHD5	EP300	HACE1	L1CAM	NCKAP1	PNPLA8	SCARA3	STXBP5	VCPIP1
ARHGAP21	CHN2	EP400	HIST2H2AB	LCORL	NDC80	POGZ	SCARB2	SUV420H1	VPS54
ARID1B	CNOT3	ERBB2IP	HMGCLL1	LGALS3BP	NECAB1	PPFIA1	SCN2A	SYT1	WDFY3
ARNTL2	CPNE3	ETFB	HOOK3	LIMCH1	NFIA	PPM1D	SETBP1	TAF4	WHSC1
ASH1L	CREBBP	FAM59A	HSPA13	LRRTM2	NFYA	PPPDE1	SETD5	TBC1D23	YPEL5
ATP1B1	CSDE1	FAM181B	INTS6	MAPT	NIF3L1	PRPF39	SHANK2	TBL1XR1	ZNF238
ATP10D	CSF1	FAM190A	IQGAP2	MBNL1	NKAIN2	PRPF40A	SIAE	TBR1	ZNF423
AXL	CSNK1E	FAT4	KBTBD4	MGRN1	NLGN2	PTBP1	SKI	TCF3	ZNF451
BCL11A	CSTF2T	FBN1	KDM3A	MIB1	NPR3	PTEN	SLC6A1	TCP11L1	ZNF644
BIRC6	CTDSPL2	FBXO27	KDM5B	MIER3	NR2F1	PTPRM	SLC6A20	THSD7A	ZNF774
BRSK2	CTTNBP2	FOXP1	KDM6B	MKL2	NR3C2	PTPRU	SLC24A3	TIMELESS	
BRWD1	CUL3	FYTTD1	KIAA0182	MLL2	NRM	QRICH1	SLC25A46	TMBIM6	
BTRC	CYTH1	GABRA1	KIAA1432	MLL3	NRXN1	RAB2A	SMARCA4	TNRC6C	
C4ORF31	DARS	GABRB1	KIF1A	MLL4	NT5C2	RANBP3L	SMURF1	TOP2A	
C9ORF68	DGKD	GABRB3	KIF11	MLXIP	P2RX5	RANBP17	SNTG1	TRIM13	
C11ORF30	DGKE	GALNTL4	KIF18A	MPP5	PAFAH1B2	RAPGEF4	SPARCL1	TRIO	

Table A.5: Gene list for ST-St. Every(1-3)+PFC(4-6) which is the result of ST-Steiner run on PFC(3-5) coexpression network with the precursor coexpression networks MDCBC(1-3), PFC(1-3), SHA(1-3) and V1C(1-3). The predicted gene sets for these precursors are given in Tables A.16, A.17, A.18 and A.19, respectively. This list consists of 256 genes.

AADAC	C9ORF68	DHX57	GABRB1	KIF11	MYO9B	PAPOLG	RGS2	SMURF1	TRIO
ACACB	C11ORF30	DIP2C	GABRB3	KIF18A	MYT1L	PBX1	RIMBP2	SNTG1	TROVE2
ACP2	C20ORF111	DLGAP5	GALNTL4	KIF20A	NCAPG	PCDH10	RIMS1	SPARCL1	TTC14
ADCY5	CACNA1B	DNAH8	GFPT2	KLC1	NCKAP1	PCOLCE	RNF38	SPAST	TTC18
ADNP	CACNA1D	DPP3	GGNBP2	KLHL9	NCOA6	PDGFD	ROS1	SPEN	TTK
AGAP2	CACNA2D3	DPY19L3	GLI3	KLHL14	NDC80	PEX5L	RSBN1L	SRCAP	TTLL3
ANGPTL3	CARKD	DRAM2	GRIA1	KPNA1	NDRG4	PHF2	SCARA3	SRPK2	UTP6
ANK2	CASC5	DYNC1H1	GRIA2	KRT34	NECAB1	PIGZ	SCN1A	ST18	VPS54
ANO5	CD96	DYRK1A	GRIN2B	KYNU	NFASC	PIK3R2	SCN2A	STXBP5	WDFY3
APH1A	CDC42BPB	ECM2	GRM7	L1CAM	NFE2L2	PLA1A	SCN7A	SUV420H1	WDR63
ARHGEF11	CDC73	EHD2	HAPLN3	LGALS3BP	NFIA	PLEKHG5	SEC31A	SV2B	WHSC1
ARID1A	CECR2	EP300	HDLBP	LMTK3	NFYA	PNPLA8	SEC31B	SYNE2	XPO5
ARID1B	CEP55	EP400	HIST1H2AE	LRFN5	NIF3L1	POGZ	SERPINB9	SYNGAP1	ZBED4
ARNTL2	CEP164	EPB49	HSPA13	MAK	NINL	PPM1D	SETBP1	SYT1	ZNF238
ASH1L	CHD5	ERBB2IP	IQGAP2	MAPT	NKAIN2	PPPDE1	SETD5	TAF4	ZNF410
ASPM	CHD8	ERI1	ISLR	MBD5	NOL6	PRPF39	SHANK2	TAS2R20	ZNF451
ASPN	CHN2	EXOSC2	ITGA5	MBNL1	NOP14	PRPF40A	SHB	TBC1D23	ZNF462
ASZ1	CNOT3	FAM59A	JUP	ME1	NOTCH3	PTEN	SIAE	TBR1	ZNF528
ATP1B1	COBL	FAM190A	KBTBD4	MGRN1	NR2F1	PTMS	SKI	TCF3	ZNF559
ATP10D	CORO1B	FBN1	KDM3A	MIB1	NR3C2	PTPRK	SLC2A13	TGIF2	ZNF594
AXL	CSDE1	FBXL13	KDM4B	MLL2	NRG4	PTPRM	SLC6A1	TIMELESS	ZNF638
BCL11A	CSNK1E	FBXO27	KDM5B	MLL3	NRXN1	QRICH1	SLC24A3	TMBIM6	ZNF774
BIRC6	CSTF2T	FEZF2	KDM6B	MPZL2	NSUN7	RAB2A	SLC35B2	TMEM178	
BRSK2	CTTNBP2	FOXG1	KIAA0182	MSH4	NUSAP1	RAD54B	SLCO1B3	TNPO3	
BRWD1	CUL3	FOXP1	KIAA1407	MTMR12	OGN	RANBP17	SLCO6A1	TOP2A	
BUB1B	DBX2	GABRA1	KIF1A	MYH10	P2RX5	RAPGEF4	SMARCA4	TRAPPC4	

Table A.6: Gene list for the cluster selected by NETBAG when run on on *de novo* counts of [6] by using code and network of NETBAG dated January 2016. The cluster has global p-value 0.0483, local p-value 0.0088, z-score 0.22086 and cluster score 139.486. This list contains 87 genes.

ADNP	CBX4	DGKD	GRIN2B	KMT2C	MYO9B	PRPF39	RPS6KA3	SVIL	TTK
APH1A	CDC42BPB	DST	HGF	KMT2E	NAA15	PTEN	SH3D19	TAF4	TTN
ARID1B	CDC73	DYRK1A	ILF2	L1CAM	NOL6	PTPRR	SKI	TCF3	UBR3
ASH1L	CECR2	EP300	IQGAP2	LDB1	NUAK1	RAPGEF3	SMARCC2	TCF4	VCP
AXL	CHD8	EP400	JADE2	LEO1	OXSR1	RAPGEF4	SMARCE1	TMPO	VCPIP1
BIRC3	CNOT3	EPHB2	JARID2	MAD1L1	PBX1	REPIN1	SMURF1	TRIM17	ZMYND11
BIRC6	CUL1	GABRB1	KDM4B	MAP3K14	PHF2	RIMS1	SPAST	TRIM37	
BRSK2	CUL3	GABRB3	KDM5B	MGRN1	POLRMT	RNF38	SRPK2	TRIO	
BTRC	DDX3X	GRIA2	KMT2A	MYH10	PPM1D	RPL12	SRSF11	TROVE2	

Table A.7: MAGI genes found by running MAGI on *de novo* counts of [6] as case data and the control data provided in [14] only. MAGI Best1 (M1_Best in MAGI's terminology) is a module with maximum score and contains 47 genes (highlighted with *). MAGI Ext1 (M1_Extended) is the union of high-scoring suboptimal modules and includes 104 genes. See [14] for details. Algorithm is run for 310,000+ iterations. The genes included in both Best1 and Ext1 are in **bold**. Note that MAGI Best1 $\subset$ MAGI Ext1.

ACTB*	**CBX5***	**CUL1***	GTPBP4	**MAP2K6***	PPP1R8	RPL12	**SMAD2***	SNW1	UBE2I
ACTL6A	CCT4	**DHX15***	HDAC2	MAPK14	**PRPF6***	RSF1	SMAD4	SP1	UBE2N
ADNP*	CCT5	**DMAP1***	HNRNPK	MDM2	PSMD3	RTF1	SMARCA5	SRC	**WDR5***
AKT1	**CDC73***	DVL2	**HNRNPR***	MECP2	PSMD9	RUVBL1	SMARCB1	SRPK2	YWHAG
ARID1B*	**CHD8***	DYNC1H1	ILF2	**MLL***	PTEN	**SF3A2***	**SMARCC1***	**SRRM2***	ZFAND5
ATF2	CHMP1A	**DYRK1A***	JUP	**MLL3***	RB1	**SF3B1***	**SMARCC2***	STRN3	
AXIN1*	CREB1	**EP300***	KAT5	MSH2	**RBBP5***	**SF3B2***	SMARCD1	**TAF4***	
BAZ1B*	**CSNK1E***	FAM40A	**KIAA1967***	**MYC***	**RELA***	SFRS4	**SMARCE1***	**TCF3***	
BCL11A*	CSNK2A1	**FBXW11***	LDB1	**NCOR1***	REST	**SIN3A***	**SMURF1***	**TCF4***	
BRAF	CSNK2A2	GORAB	**LEO1***	NFKB1	RLIM	**SIRT1***	SMURF2	TEX10	
CASP2	**CTNNB1***	**GSK3B***	**LRP5***	**PIAS1***	RPL11	**SKI***	SNIP1	TRAF6	

Table A.8: MAGI genes found by running MAGI on *de novo* counts of [6] as case data and the control data provided in [14] only, after removing the genes in MAGI Best1. The algorithm MAGI is run iteratively to obtain 310,000+ clusters in this second setting. Among these, a maximum scoring cluster MAGI Best2 cluster contains 19 genes (highlighted with *). Best2 is then used by the algorithm to generate MAGI Ext2, which includes 80 genes. The genes included in both Best1 and Ext1 are in **bold**. Note that MAGI Best2 $\not\subset$ MAGI Ext2.

ABI2*	C7ORF64	DKC1	GORAB	L1CAM	PPP1R8	RPL5	SHB	UBE2I	**YWHAG***
ANK2*	CASP2	**DLG3***	**GRB2***	**LPHN1***	PRPF40A	RPL10	SLC8A1	UBE2N	ZFAND5
AP1G1*	CBLL1	**DLG4***	GTPBP4	MAP4K5	PSMD3	RPL12	**SPTBN1***	UBN1	ZNF451
AP2A2*	CCT3	**DLGAP4***	HIRA	MDM2	PTBP2	RPL14	SRC	UBR5	
ARHGEF2	CFL1	DYNC1H1	**HTT***	NOP56	PTEN	RPL30	STRN3	UPF1	
ARHGEF7*	CHMP1A	EIF2B5	ILF2	**NUMB***	PUF60	RPL35	**SYNGAP1***	USP7	
ATF2	CLINT1	ELAVL1	JUP	PAFAH1B1	RABEP1*	SETD5	TRAF2	WASF1*	
BMI1	CSNK2A1	FAM40A	KIF1B	PAFAH1B2	RNPS1	SFRS12	TRAF6	XIAP	
BRAF*	DAG1	GIT1*	KPNA1	POLR1C	RPL4	SHANK2	UBC	YWHAB	

Table A.9: Overlap coefficient of the predicted gene sets by each method, where overlap coefficient between sets A and B is defined as $o(A, B) = \frac{|A \cap B|}{min(|A|,|B|)}$.

gene list name	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) ST-St. Every(1-3)+PFC(3-5)	1.000	0.679	0.517	0.568	0.528	0.255	0.135	0.053	0.125
(2) ST-St. Every(1-3)+PFC(4-6)	0.679	1.000	0.517	0.573	0.674	0.277	0.163	0.105	0.150
(3) NETBAG	0.517	0.517	1.000	0.506	0.460	0.319	0.230	0.000	0.050
(4) DAWN PFC(3-5)	0.568	0.573	0.506	1.000	0.601	0.319	0.163	0.053	0.075
(5) DAWN PFC(4-6)	0.528	0.674	0.460	0.601	1.000	0.234	0.144	0.105	0.138
(6) MAGI Best1	0.255	0.277	0.319	0.319	0.234	1.000	1.000	0.000	0.000
(7) MAGI Ext1	0.135	0.163	0.230	0.163	0.144	1.000	1.000	0.105	0.288
(8) MAGI Best2	0.053	0.105	0.000	0.053	0.105	0.000	0.105	1.000	0.789
(9) MAGI Ext2	0.125	0.150	0.050	0.075	0.138	0.000	0.288	0.789	1.000

Table A.10: Intersection $\cap$ of each of the predicted gene set S that follows with the 251 genes that have at least 1 dnLoF in 1,643 samples from [7] are shown: (i) ST-Steiner's predicted gene sets, (ii) PPI-based AXAS modules in [37], (iii) co-expression based modules from [38] and (iv) from [15], and (v) NETBAG 2G module reported in [10] (the module that selects at most two genes per *de novo* CNV). Fisher's exact test is used to assess significance and and the most significant results is marked in **bold**. The union of TADA results of [6] and the gene sets, amounting to 19,830 genes, is used for the background. Willsey03 and Parikshak13 prefixes refer to the modules reported in [15] and [38], respectively. AXAS modules are obtained via personal communication.

Table A.10: (text on the previous page)

gene list name	p-value	$\lvert\cap\rvert/\lvert S\rvert$
ST-St. Every(1-3)+PFC(3-5)	**2.445e-12**	**21/234**
ST-St. Every(1-3)+PFC(4-6)	7.302e-10	19/256
NETBAG 2G	2.313e-01	2/72
Willsey03 MDCBC(8-10)	2.882e-05	11/188
Willsey03 PFC(3-5)	2.376e-04	9/163
Willsey03 PFC(4-6)	1.324e-03	8/167
Parikshak13 M2	4.655e-04	27/1,039
Parikshak13 M3	3.076e-01	16/996
Parikshak13 M13	6.201e-02	5/870
Parikshak13 M16	6.813e-01	7/492
Parikshak13 M17	8.865e-01	12/1,041
AXAS M1	1.646e-01	13/700
AXAS M2	7.267e-01	7/680
AXAS M3	1.681e-01	8/395
AXAS M4	9.477e-06	14/270
AXAS M5	4.134e-02	6/199
AXAS M6	1.000e+00	2/198
AXAS M7	2.719e-01	0/155
AXAS M8	4.489e-01	3/154
AXAS M9	6.644e-01	2/119
AXAS M10	6.573e-01	2/114
AXAS M11	6.560e-01	2/113
AXAS M12	1.000e+00	1/111
AXAS M13	1.000e+00	1/98
AXAS M14	8.580e-02	3/82
AXAS M15	2.782e-01	2/82
AXAS M16	5.465e-01	1/62
AXAS M17	1.000e+00	0/61
AXAS M18	5.288e-01	1/59
AXAS M19	5.227e-01	1/58
AXAS M20	4.714e-01	1/50
AXAS M21	4.293e-01	1/44
AXAS M22	1.000e+00	0/36
AXAS M23	6.868e-02	2/34
AXAS M24	2.859e-02	2/21
AXAS M25	1.000e+00	0/20
AXAS M26	1.526e-01	1/13

Table A.11: Gene list for Simons Foundation Autism Research Initiative (SFARI) Category 1, that is consisted of 24 genes associated with ASD with high confidence[4].

ADNP	ASH1L	CUL3	GRIN2B	KMT5B	POGZ	SCN2A	SYNGAP1		
ANK2	ASXL3	DSCAM	KATNAL2	MYT1L	PTEN	SETD5	TBR1		
ARID1B	CHD8	DYRK1A	KMT2A	NAA15	RELN	SHANK3	TRIP12		

Table A.12: Gene list for Simons Foundation Autism Research Initiative (SFARI) Category 2, that contains 59 genes associated with ASD with strong confidence[4].

ANKRD11	CACNA2D3	DDX3X	GIGYF2	KAT2B	MECP2	NLGN3	SCN9A	SRSF11	UPF3B
BAZ2B	CHD2	DEAF1	GRIA1	KDM5B	MED13	NRXN1	SHANK2	TBL1XR1	USP7
BCKDK	CIC	DIP2C	GRIP1	KDM6A	MED13L	PHF3	SLC6A1	TCF20	USP15
BCL11A	CNTN4	ERBIN	ILF2	KMT2C	MET	PTCHD1	SMARCC2	TNRC6B	WAC
CACNA1D	CNTNAP2	FOXP1	INTS6	MAGEL2	MSNP1AS	RANBP17	SPAST	TRIO	WDFY3
CACNA1H	CTNND2	GABRB3	IRF2BPL	MBOAT7	NCKAP1	RIMS1	SRCAP	UBN2	

Table A.13: Table shows the significance of the intersection ($\cap$) of the gene set S predicted by each method and ASD-related gene sets as indicated in external resources. AXAS modules are from [37] and obtained via personal communication. Willsey03 and Parikshak13 prefixes refer to the modules reported in [15] and [38], respectively. Willsey03 PFC(3-5), PFC(4-6) and MDCBC(8-10) are the modules enriched in probable ASD genes. The Parikshak13 modules that are included are the significant modules as chosen for comparison in [14]. NETBAG 2G is the cluster in [10] that selects at most two genes per *de novo* CNV. Significance is assessed using Fisher's exact test. For the background, 19,830 genes are used (the union of genes with TADA values in [6] and the compared gene sets). The most significant results are in **bold**. The first two lists are SFARI Gene's Category 1 (24 genes) and Category 2 (59 genes) gene sets. The third is the list of transcription factor FMRP's targets (842 genes). The fourth and the fifth gene sets are the targets of the splicing factor RBFOX: *RBFOX-peak* denotes 1,048 target genes with significant HITS-CLIP peaks and *RBFOX-splicing* denotes 587 genes RBFOX targets with alternative splicing events. Histone modifiers (152 genes) denotes the set of histone modifier genes used in [6]. Finally, Synaptic genes (873 genes) stands for the union of synapse-related-GO terms' gene sets which are GO:0021681, 0021680, 0021694, 0021692, 0060076, 0060077, 0051965, 0090129, 0032230, 0051968, 0045202, 0098794, 0098793, 0048169, 0051963, 0090128, 0032225, 0032228, 0051966, 0007271, 0001963 and 0035249.

Table A.13: (text on the previous page)

	SFARI Category 1		SFARI Category 2		FMRP Targets		RBFOX - peak		RBFOX - splice		Histone Modifiers		Synaptic Genes	
	p-value	$\|\cap\|/\|S\|$	p-value	$\|\cap\|/\|S\|$	p-value	$\|\cap\|/\|S\|$	p-value	$\|\cap\|/\|S\|$	p-value	$\|\cap\|/\|S\|$	p-value	$\|\cap\|/\|S\|$	p-value	$\|\cap\|/\|S\|$
ST-St. Every(1-3)+PFC(3-5)	**5.682e-26**	**16/234**	**1.018e-22**	**19/234**	2.637e-23	52/234	4.747e-10	38/234	**2.723e-12**	**31/234**	2.604e-07	12/234	6.757e-05	25/234
ST-St. Every(1-3)+PFC(4-6)	3.590e-23	15/256	2.190e-20	18/256	4.701e-28	60/256	2.818e-12	44/256	1.159e-08	27/256	4.656e-06	11/256	3.640e-05	27/256
NETBAG 2G	1.000e+00	0/72	6.248e-05	4/72	9.131e-06	13/72	1.241e-03	11/72	4.823e-05	10/72	2.276e-03	4/72	7.554e-08	16/72
Willsey03 PFC(3-5)	1.152e-11	8/163	1.408e-03	4/163	5.142e-14	33/163	3.712e-07	26/163	3.668e-07	19/163	1.300e-01	3/163	9.195e-04	17/163
Willsey03 PFC(4-6)	1.400e-11	8/167	8.836e-02	2/167	1.785e-14	34/167	2.827e-09	30/167	4.079e-03	12/167	1.369e-01	3/167	1.814e-01	11/167
Willsey03 MDCBC(8-10)	8.047e-15	10/188	2.370e-03	4/188	5.070e-06	23/188	1.257e-02	18/188	1.367e-02	12/188	1.029e-04	8/188	3.136e-05	22/188
Parikshak13 M2	3.445e-02	4/1,039	3.383e-02	7/1,039	4.682e-07	79/1,039	6.531e-03	36/1,039	3.455e-01	25/1,039	8.722e-03	16/1,039	6.160e-04	25/1,039
Parikshak13 M3	3.412e-01	2/996	2.774e-02	7/996	3.741e-01	36/996	7.969e-02	40/996	1.022e-01	38/996	2.141e-06	23/996	4.671e-04	23/996
Parikshak13 M13	6.246e-01	0/870	5.215e-01	1/870	3.283e-03	55/870	3.880e-03	65/870	3.568e-01	30/870	7.027e-02	2/870	8.508e-17	96/870
Parikshak13 M16	4.530e-01	1/492	1.802e-01	3/492	**1.511e-36**	**95/492**	**5.615e-24**	**87/492**	4.739e-06	34/492	5.946e-01	2/492	5.114e-17	68/492
Parikshak13 M17	1.000e+00	1/1,041	5.519e-01	4/1,041	3.727e-06	76/1,041	7.207e-01	57/1,041	5.105e-01	34/1,041	5.175e-03	1/1,041	9.962e-09	87/1,041
AXAS M1	1.000e+00	0/700	2.090e-04	9/700	2.771e-09	65/700	2.466e-02	50/700	1.616e-02	32/700	**5.071e-25**	**41/700**	2.231e-01	24/700
AXAS M2	8.427e-03	4/680	8.677e-04	8/680	6.132e-20	86/680	6.457e-09	73/680	1.353e-11	55/680	1.098e-01	9/680	**1.545e-86**	**175/680**
AXAS M3	1.201e-03	4/395	6.337e-01	0/395	7.670e-03	28/395	1.581e-02	32/395	7.635e-01	10/395	1.021e-08	17/395	6.217e-01	15/395
AXAS M4	2.805e-01	1/270	4.644e-02	3/270	6.074e-14	43/270	4.286e-05	31/270	1.136e-10	31/270	4.657e-01	3/270	2.567e-13	43/270
AXAS M5	1.000e+00	0/199	4.489e-01	1/199	7.284e-02	14/199	1.011e-02	19/199	2.437e-03	14/199	1.000e+00	1/199	7.454e-05	22/199
AXAS M6	1.000e+00	0/198	4.473e-01	1/198	5.103e-02	3/198	6.284e-01	8/198	3.920e-01	8/198	1.000e+00	1/198	4.367e-03	18/198
AXAS M7	1.000e+00	0/155	1.000e+00	0/155	8.400e-01	7/155	4.670e-01	10/155	6.334e-01	3/155	6.356e-01	0/155	4.277e-01	9/155
AXAS M8	1.000e+00	0/154	3.691e-01	1/154	8.394e-01	7/154	9.301e-03	16/154	4.668e-01	6/154	6.351e-01	0/154	3.289e-01	4/154
AXAS M9	1.000e+00	0/119	2.992e-01	1/119	1.618e-03	13/119	4.670e-05	18/119	2.487e-02	8/119	6.008e-01	1/119	4.282e-02	10/119
AXAS M10	1.293e-01	1/114	1.000e+00	0/114	1.089e-03	13/114	8.947e-04	15/114	3.933e-01	5/114	5.737e-02	3/114	1.510e-04	15/114
AXAS M11	1.282e-01	1/113	4.643e-03	3/113	5.007e-11	24/113	2.652e-04	16/113	1.887e-02	8/113	5.818e-01	1/113	1.079e-32	46/113
AXAS M12	1.000e+00	0/111	1.000e+00	0/111	2.398e-01	7/111	3.866e-01	8/111	3.873e-01	1/111	5.753e-01	1/111	2.983e-05	16/111
AXAS M13	1.000e+00	0/98	1.000e+00	0/98	6.114e-01	5/98	1.742e-01	8/98	5.389e-01	4/98	1.731e-01	2/98	3.291e-01	2/98
AXAS M14	1.000e+00	0/82	1.000e+00	0/82	1.000e+00	3/82	1.000e+00	4/82	1.000e+00	2/82	1.306e-01	2/82	1.000e+00	3/82
AXAS M15	9.470e-02	1/82	1.000e+00	0/82	7.834e-03	9/82	8.015e-01	3/82	7.363e-01	3/82	4.685e-01	1/82	1.018e-02	9/82
AXAS M16	1.000e+00	0/62	1.000e+00	0/62	4.709e-02	6/62	5.667e-01	4/62	7.058e-01	2/62	3.798e-01	1/62	1.991e-01	5/62
AXAS M17	1.000e+00	0/61	1.000e+00	0/61	1.869e-01	5/61	1.000e+00	3/61	2.071e-03	7/61	3.750e-01	1/61	5.248e-03	8/61
AXAS M18	1.000e+00	0/59	1.000e+00	0/59	1.698e-04	10/59	1.291e-01	6/59	2.990e-02	5/59	1.000e+00	0/59	1.072e-03	9/59
AXAS M19	1.000e+00	0/58	1.000e+00	0/58	1.795e-01	0/58	7.508e-02	0/58	1.000e+00	1/58	1.000e+00	0/58	1.000e+00	2/58
AXAS M20	1.000e+00	0/50	1.000e+00	0/50	4.708e-01	3/50	1.480e-02	7/50	5.967e-02	4/50	1.000e+00	0/50	1.000e+00	2/50
AXAS M21	1.000e+00	0/44	1.000e+00	0/44	4.345e-01	3/44	7.872e-02	5/44	3.746e-01	2/44	4.475e-02	2/44	1.000e+00	1/44
AXAS M22	1.000e+00	0/36	1.000e+00	0/36	1.951e-01	3/36	1.175e-01	4/36	2.099e-02	4/36	1.000e+00	0/36	2.446e-08	12/36
AXAS M23	1.000e+00	0/34	1.000e+00	0/34	1.735e-01	3/34	6.983e-01	2/34	2.657e-01	2/34	1.000e+00	0/34	4.036e-01	0/34
AXAS M24	1.000e+00	0/21	1.762e-03	2/21	2.231e-01	2/21	1.000e+00	1/21	4.673e-01	1/21	1.492e-01	1/21	1.000e+00	0/21
AXAS M25	1.000e+00	0/20	1.000e+00	0/20	1.000e+00	0/20	6.229e-01	0/20	1.000e+00	0/20	1.000e+00	0/20	1.000e+00	0/20
AXAS M26	1.000e+00	0/13	1.000e+00	0/13	1.000e+00	0/13	5.033e-01	1/13	1.000e+00	0/13	1.000e+00	0/13	4.448e-01	1/13

Table A.14: This table compares the performance of running ST-Steiner in different settings. The first column indicates the context that a run was mentioned. For each setting, the table presents the intersection $\cap$ of the predicted gene set S with the 251 genes that have at least 1 *de novo* LoF mutation in 1,643 samples from [7], which are not in the training set. Fisher's exact test is used to assess significance. The most significant results are shown in **bold**. The union of TADA results of [6] and the gene sets, amounting to 18,736 genes, is used for the background.

used in	gene list name	p-value $\lvert\cap\rvert/\lvert S\rvert$
comparisons with other methods	ST-St. Every(1-3)+PFC(3-5)	7.000e-12 21/234
	ST-St. Every(1-3)+PFC(4-6)	1.836e-09 19/256
evaluating the effect of spatio-temporal analysis	ST-St. PFC(1-3)+(3-5)	**6.447e-12 21/233**
	ST-St. PFC(1-3)+(4-6)	3.858e-10 20/261
evaluating the effect of parameter choices	ST-St. PFC(1-3)+(3-5) $\varepsilon 0.5$, $\rho 0.05$	2.284e-11 20/223
	ST-St. PFC(1-3)+(4-6) $\varepsilon 0.5$, $\rho 0.05$	1.347e-10 20/246
	ST-St. PFC(1-3)+(3-5) $\varepsilon 0.67, \rho 0.05$	1.980e-10 17/174
	ST-St. PFC(1-3)+(3-5) $\varepsilon 0.67, \rho 0.1$	4.657e-11 18/182
	ST-St. PFC(1-3)+(4-6) $\varepsilon 0.67, \rho 0.05$	4.145e-11 19/205
	ST-St. PFC(1-3)+(4-6) $\varepsilon 0.67, \rho 0.1$	6.716e-10 18/214
ST-Steiner vs union of independent results	ST-St. PFC(1-3) $\cup$ PFC(3-5) $\varepsilon 0.5, \lambda 0$	1.472e-10 21/275
	ST-St. PFC(1-3) $\cup$ PFC(4-6) $\varepsilon 0.5, \lambda 0$	2.897e-09 20/293
ST-Steiner vs. coarser granularity	ST-St. PFC(1-5)	1.025e-06 14/211
	ST-St. PFC(1-6)	4.480e-07 14/197
ST-Steiner vs. no time effect	ST-St. PFC(3-5)	8.272e-12 20/211
	ST-St. PFC(4-6)	3.018e-09 18/235
	ST-St. PFC(3-5)+10%	4.008e-11 20/230
	ST-St. PFC(4-6)+10%	3.467e-09 19/266

Table A.15: This table compares the performance of ST-Steiner with itself, when run in different settings. First column indicates the context these runs were mentioned. It shows the significance of the intersection $\cap$ of the predicted gene sets S and ASD-related gene sets indicated in external resources. Significance is assessed using Fisher's exact test. As for the background, 18,737 genes are used (the union genes with TADA values in [6] and extra 2 genes included by MAGI). The most significant result is marked in **bold**. The first two lists are SFARI Gene's Category 1 (24 genes) and Category 2 (59 genes) gene sets. Category 1 indicates the high-confidence risk genes, followed by Category 2 (https://gene.sfari.org/database/gene-scoring accessed on January 15, 2018, available in Tables A.11 and A.12, respectively)). The third list is the list of targets for the transcription factor FMRP (842 genes) [41]. The fourth and the fifth gene sets are the targets of the splicing factor RBFOX: *RBFOX-peak* denotes 1,048 target genes with significant HITS-CLIP peaks and *RBFOX-splicing* denotes 587 genes RBFOX targets with alternative splicing events [6, 39, 40]. Histone modifiers (152 genes) stands for the set of histone modifier genes used in [6]. Finally, Synaptic genes (873 genes) denotes the union of synapse-related-GO terms' gene sets which are GO:0021681, 0021680, 0021694, 0021692, 0060076, 0060077, 0051965, 0090129, 0032230, 0051968, 0045202, 0098794, 0098793, 0048169, 0051963, 0090128, 0032225, 0032228, 0051966, 0007271, 0001963, 0035249.

Gene list name	SFARI Category 1		SFARI Category 2		FMRP Targets		RBFOX - peak		RBFOX - splice		Histone Modifiers		Synaptic Genes	
	p-value	$\|\cap\|/\|S\|$	p-value	$\|\cap\|/\|S\|$	p-value	$\|\cap\|/\|S\|$	p-value	$\|\cap\|/\|S\|$	p-value	$\|\cap\|/\|S\|$	p-value	$\|\cap\|/\|S\|$	p-value	$\|\cap\|/\|S\|$
ST-St. Every(1-3)+PFC(3-5)	1.402e-25	16/234	**2.920e-22**	**19/234**	3.244e-22	52/234	2.212e-09	38/234	1.129e-11	31/234	4.725e-07	12/234	1.200e-04	25/234
ST-St. Every(1-3)+PFC(4-6)	8.361e-23	15/256	5.918e-20	18/256	**8.829e-27**	**60/256**	**1.750e-11**	**44/256**	3.672e-08	27/256	7.914e-06	11/256	7.659e-05	27/256
ST-St. PFC(1-3)+(3-5)	1.307e-25	16/233	1.072e-20	18/233	1.695e-21	51/233	5.518e-10	39/233	1.107e-09	28/233	3.223e-06	11/233	7.853e-04	23/233
ST-St. PFC(1-3)+(4-6)	8.411e-25	16/261	8.395e-20	18/261	1.801e-25	59/261	4.239e-10	42/261	7.353e-07	25/261	5.478e-05	10/261	9.050e-04	25/261
ST-St. PFC(1-3)+(3-5) $\rho 0.05$	6.348e-26	16/223	4.820e-21	18/223	8.899e-21	49/223	1.958e-09	37/223	8.292e-11	29/223	2.108e-06	11/223	1.521e-04	24/223
ST-St. PFC(1-3)+(4-6) $\rho 0.05$	3.188e-25	16/246	2.876e-20	18/246	4.587e-26	58/246	6.288e-11	42/246	2.406e-07	25/246	3.324e-05	10/246	3.281e-04	25/246
ST-St. PFC(1-3)+(3-5) $\varepsilon 0.67, \rho 0.05$	3.766e-23	14/174	4.988e-18	15/174	9.226e-19	41/174	1.780e-09	32/174	1.979e-10	25/174	**1.797e-07**	**11/174**	1.601e-03	18/174
ST-St. PFC(1-3)+(4-6) $\varepsilon 0.67, \rho 0.05$	2.746e-24	15/205	4.360e-20	17/205	4.408e-25	52/205	2.459e-09	35/205	7.106e-06	20/205	6.859e-06	10/205	3.065e-04	22/205
ST-St. PFC(1-3)+(4-6) $\varepsilon 0.67, \rho 0.1$	5.329e-24	15/214	9.144e-20	17/214	1.960e-22	50/214	7.842e-09	35/214	1.041e-06	22/214	1.000e-05	10/214	1.514e-03	21/214
ST-St. PFC(1-3)+(3-5) $\varepsilon 0.67, \rho 0.1$	7.210e-23	14/182	5.541e-21	17/182	2.242e-21	45/182	8.618e-11	35/182	2.686e-09	24/182	2.829e-07	11/182	1.073e-03	19/182
ST-St. PFC(1-5)	2.549e-26	16/211	2.694e-18	16/211	9.984e-23	50/211	5.747e-06	29/211	2.078e-07	23/211	8.838e-06	10/211	2.110e-06	27/211
ST-St. PFC(1-6)	8.191e-27	16/197	3.287e-17	15/197	1.442e-21	47/197	1.373e-07	31/197	5.876e-08	23/197	4.823e-06	10/197	**1.556e-07**	**28/197**
ST-St. PFC(1-3) $\cup$ PFC(3-5) $\varepsilon 0.5$	**7.081e-29**	**18/275**	6.469e-21	19/275	5.093e-25	60/275	6.901e-09	41/275	**1.686e-12**	**35/275**	1.549e-05	11/275	1.403e-03	25/275
ST-St. PFC(1-3) $\cup$ PFC(4-6) $\varepsilon 0.5$	5.565e-24	16/293	6.729e-19	18/293	7.531e-26	63/293	4.811e-10	45/293	1.270e-08	30/293	1.424e-04	10/293	1.844e-03	26/293
ST-St. PFC(3-5)	2.549e-26	16/211	1.752e-21	18/211	9.984e-23	50/211	1.879e-08	34/211	2.114e-11	29/211	1.229e-06	11/211	1.382e-03	21/211
ST-St. PFC(4-6)	2.250e-23	15/235	1.253e-20	18/235	6.111e-23	53/235	7.133e-10	39/235	9.942e-08	25/235	2.250e-05	10/235	8.492e-04	23/235
ST-St. PFC(3-5)+10%	1.056e-25	16/230	3.154e-19	17/230	1.397e-22	52/230	1.342e-09	38/230	1.759e-10	29/230	2.844e-06	11/230	1.409e-03	22/230
ST-St. PFC(4-6)+10%	1.503e-22	15/266	3.425e-21	19/266	2.088e-23	57/266	6.466e-11	44/266	8.155e-08	27/266	6.420e-05	10/266	2.405e-05	29/266

Table A.16: Gene list for ST-St. MDCBC(1-3), which is the result of ST-Steiner run on MDCBC(1-3) coexpression network. The list is consisted of 39 genes.

AHI1	CACNA2D3	DACH1	GALNTL4	LHX1	NTNG1	RELN	SCN2A	SLITRK5	TRPC3
ARID1B	CD55	ERC2	GNAQ	MBNL2	NTRK3	RGS2	SEMA3D	SLITRK6	ZFHX3
BCL11A	CTTNBP2	FREM3	GRIN2B	MEIS1	PDE1C	RPS6KA3	SIAH3	SPARCL1	ZNF238
C11ORF30	CYP26B1	GAD2	KCNN2	NCAPD2	PTPRM	SCARA3	SLC6A1	SYT4	

Table A.17: Gene list for ST-St. PFC(1-3), which is the result of ST-Steiner run on PFC(1-3) coexpression network. The list is consisted of 115 genes.

AAK1	BUB1B	CKAP2L	DYRK1A	HIVEP2	LGALS3BP	NCAPH	PRC1	SMARCA2	TCF3
ACBD7	C11ORF30	CLIP3	E2F8	HMMR	LRRTM2	NDC80	PRKD2	SMARCA4	TEX2
ADCY5	C15ORF42	CMIP	EP400	HMP19	MAPT	NEDD1	PTBP1	SMURF1	TIMELESS
ADNP	CACNA1D	CTTNBP2	ESPL1	IQGAP2	MBNL1	NOTCH3	PTEN	SPAG5	TOP2A
AGAP2	CACNA2D3	CUL3	ETFB	IREB2	MCAM	NRXN1	RACGAP1	SPATA13	TRIO
ANK2	CASC5	CXORF57	FAM59A	JUP	MEGF10	NSUN7	RGS2	SPC24	TTK
ANO5	CDC42BPB	DGKD	FAT1	KDM3A	MELK	NUSAP1	SCN2A	SPEN	XRCC2
ARID1B	CDCA2	DICER1	FOXM1	KIF11	MIB1	PBX1	SEMA4G	SRPK2	
ASH1L	CDCA7L	DLGAP5	GABRB3	KIF18A	MLL3	PDGFD	SETBP1	STXBP5	
AXL	CDK2	DSCAM	GLI3	KIF20A	MYO9B	PHLDA1	SGOL2	SUV420H1	
BCL11A	CEP55	DUSP10	GRIN2B	KIF23	MYT1L	PITPNC1	SKA1	SYNGAP1	
BTG3	CHD8	DYNC1H1	HIST1H2AB	L1CAM	NCAPD2	POGZ	SLC1A1	TBR1	

Table A.18: Gene list for ST-St. SHA(1-3), which is the result of ST-Steiner run on SHA(1-3) coexpression network. The list contains 177 genes.

ACACB	C3ORF15	CUL3	FOXP1	IQGAP2	LYST	NFIA	RELN	SLCO1B3	TCF3
ADCY5	C13ORF28	CYP4F12	FRY	ITGA5	MAK	NOTCH3	REPIN1	SLCO6A1	TCF4
ADNP	C14ORF135	DAB2	GABRA1	KATNAL2	MAPT	NR2F1	RGS2	SMURF1	TEX2
AGAP2	C20ORF195	DBX2	GABRB1	KCNT2	MBD5	NRXN1	RIMS1	SNTG1	TMBIM6
ANGPTL3	CACNA1D	DGKD	GABRB3	KIAA0182	MGRN1	NUF2	RPA1	SORCS3	TRIO
ANK2	CACNA2D3	DOCK3	GALNTL4	KIAA0319	MIB1	OPRL1	RXRG	SPARCL1	TROVE2
ARHGEF7	CARKD	DPP3	GLI3	KIF1A	MLL3	PARP6	SCAMP5	SRBD1	TTC39A
ARID1B	CDC42BPA	DRD3	GLT8D2	KIF18A	MPP5	PBX1	SCARA3	SRPK2	TTK
ASH1L	CDC42BPB	DYNC1H1	GNA12	KIF20A	MPZL2	PCOLCE	SCN1A	STMN4	TTLL3
ATP1B1	CECR2	DYRK1A	GNAO1	KIRREL3	MSH4	PCTP	SCN2A	STXBP1	VIM
ATP10D	CEP55	EIF2C1	GRIA1	KLC1	MTSS1	PHF15	SCN3B	STXBP5	WDFY3
ATXN1	CHD3	EP400	GRIA2	KYNU	MYH10	POGZ	SETBP1	SUPT16H	ZFPM2
AXL	CHD4	ERBB2IP	GRIK2	L1CAM	MYO9B	PPM1D	SETD5	SUV420H1	ZNF238
BCL11A	CHD8	ETFB	GRIN2B	LGALS3BP	MYT1L	PTEN	SGIP1	SYNGAP1	ZNF423
BIRC6	CHEK1	FAM190A	GRM7	LMTK3	NCAM1	PTPRM	SH3RF2	SYNPR	ZNF774
BRSK2	CNTN5	FEZF2	HDLBP	LRRC4	NCAPG	QKI	SLC5A7	SYT1	
BRWD1	COLEC12	FGFR1	HIVEP3	LRRTM2	NEUROD6	RAB2A	SLC6A1	SYT5	
C2ORF73	CTTNBP2	FOXG1	HMP19	LUM	NFATC4	RAPGEF4	SLC41A2	TBR1	

Table A.19: Gene list for ST-St. V1C(1-3), which is the result of ST-Steiner run on V1C(1-3) coexpression network. The list is consisted of 138 genes.

ABCA9	CACNA2D3	CSNK1E	ETFB	KATNAL2	MAPT	NR3C2	PTEN	SLC6A1	STXBP5
ADNP	CASC5	CTTNBP2	EXO1	KDM5B	MCM10	NR4A2	PTPRB	SLC6A13	SUV420H1
AGAP2	CATSPERB	CUL3	FAM59A	KIAA0182	MEGF10	NRG4	RAG2	SLC6A20	SYNGAP1
ANK2	CCDC80	CYP1B1	FAT1	KIF2C	MIB1	NRXN1	RANBP3L	SLC7A2	TBR1
ARID1B	CDC20	DAB2	FBN1	KIF11	MKI67	NSUN7	RBMS3	SLCO1B3	TCF3
ASAH1	CDC42BPB	DIP2C	FBXL7	KIF18A	MLL3	PBX1	RGS2	SLCO6A1	TIMELESS
ASH1L	CDCA2	DLGAP5	FEM1B	KIF20A	MPZL2	PCOLCE	SCARA3	SMURF1	TNIK
ATP1B1	CDCA8	DSCAM	FGFR3	KYNU	MSH4	PDGFD	SCN2A	SORCS3	TRIO
AXL	CEP55	DSP	GABRB3	L1CAM	MYO9B	PEX5L	SCN7A	SPARCL1	TTC18
BCL11A	CHD8	DST	GLI3	LAMA2	MYT1L	PHF2	SETBP1	SPAST	TTK
BRWD1	CIT	DYRK1A	GRAMD3	LGALS3BP	NCKAP1	PHKA1	SETD5	SPATA13	VWA5A
BUB1B	CREB3L2	ECM2	GRIA2	LIX1	NDC80	POGZ	SGOL2	SPC24	ZIC2
C9ORF84	CSDE1	EHD2	GRIN2B	LRIG1	NOTCH2	PPM1D	SKA1	SPC25	
CA14	CSF1	EP400	IQGAP2	LRRTM2	NOTCH3	PTBP1	SKI	SRPK2	

Table A.20: Gene list for ST-St. PFC(1-3)+(3-5) which is the result of ST-Steiner run on PFC(3-5) coexpression network, when PFC(1-3) coexpression network is used as precursor. ST-Steiner's predicted gene set for the PFC(1-3) coexpression network is given in Table A.17. This list consists of 233 genes.

ADCY5	CACNA2D3	DLGAP5	GLI3	KLC1	MYH10	PAPOLG	RAPGEF4	SMURF1	TOP2A
ADNP	CASC5	DPP3	GRAMD3	KLHL9	MYO9A	PCOLCE	RBM24	SNTG1	TRIM13
AGAP2	CDC42BPB	DRAM2	GRIN2B	KLHL14	MYO9B	PDGFD	RBMS3	SPARCL1	TRIO
AMY2B	CDC73	DSCAM	GRM7	KPNA1	MYT1L	PEX5L	RELN	SPAST	TRIP12
ANK2	CDK2	DYNC1H1	GRPEL2	KRT34	NAV2	PHF2	RGS2	SPATA13	TROVE2
ANO5	CEP55	DYRK1A	HACE1	L1CAM	NCAPG	PIK3C2B	RIMBP2	SPEN	TTC14
APH1A	CHD5	ECM2	HMGCLL1	LCORL	NCKAP1	PLEKHG5	RIMS1	SPRED2	TTK
ARID1B	CHN2	EP300	HMMR	LGALS3BP	NDC80	PLXDC2	RNF38	SRCAP	TTLL3
ARNTL2	CLCN6	EP400	HOOK3	LIMCH1	NECAB1	PNPLA8	SCAND3	SRPK2	UBN1
ASH1L	CMIP	ERBB2IP	HSPA13	LRP1	NFIA	POGZ	SCARB2	ST3GAL6	UBR3
ATP1B1	CNOT3	ETFB	INTS6	LRRC4C	NFIB	PPFIA2	SCN2A	ST18	USP15
ATP10D	CPNE3	F3	IQGAP2	LRRTM2	NFYA	PPM1D	SEPP1	STAU2	VCPIP1
AXL	CREBBP	FAM59A	JUP	MAPT	NIF3L1	PPPDE1	SETBP1	STXBP5	VPS54
BCL11A	CSDE1	FAM175A	KBTBD4	MBNL1	NKAIN2	PRC1	SETD5	SUV420H1	WDFY3
BIRC6	CSF1	FAM181B	KDM3A	MEGF10	NOTCH3	PRPF39	SGOL2	SYT1	WHSC1
BRSK2	CSNK1E	FAM190A	KDM5B	MGRN1	NPR3	PRPF40A	SHANK2	TAF4	ZNF423
BRWD1	CSTF2T	FBN1	KDM6B	MIB1	NR2F1	PTBP1	SIAE	TBC1D23	ZNF774
BTRC	CTTNBP2	FBXO27	KIAA0182	MKL2	NR3C2	PTEN	SKI	TBR1	
BUB1B	CUL3	FEZF2	KIAA1432	MLL2	NRXN1	PTPRM	SLC6A1	TCF3	
C4ORF31	CYTH1	FOXP1	KIF1A	MLL3	NSUN7	PTPRU	SLC6A20	TEX2	
C9ORF68	DGKD	FYTTD1	KIF11	MLXIP	NT5C2	QRICH1	SLC11A2	TGFBR3	
C11ORF30	DGKE	GABRB1	KIF18A	MPP5	NUSAP1	RAB2A	SLC24A3	THSD7A	
C20ORF111	DHX57	GABRB3	KIF20A	MPZL2	P2RX5	RANBP3L	SLC25A46	TIMELESS	
CACNA1D	DICER1	GALNTL4	KIF23	MTMR12	PAFAH1B2	RANBP17	SMARCA4	TMBIM6	

Table A.21: Gene list for ST-St. PFC(1-3)+(4-6) which is the result of ST-Steiner run on PFC(4-6) coexpression network, when PFC(1-3) coexpression network is used as precursor. ST-Steiner's predicted gene set for the PFC(1-3) coexpression network is given in Table A.17. This list consists of 261 genes.

AADAC	C20ORF111	DIP2C	GABRB1	KIF18A	MYO9B	OGN	RAPGEF4	SMURF1	TTC18
ABCA5	CACNA1B	DLGAP5	GABRB3	KIF20A	MYT1L	P2RX5	RBM39	SNTG1	TTK
ACP2	CACNA1D	DNAH8	GALNTL4	KIF23	NCAPG	PAPOLG	RGS2	SPARCL1	TTLL3
ADCY5	CACNA2D3	DPP3	GFPT2	KLC1	NCKAP1	PBX1	RIMBP2	SPAST	UTP6
ADNP	CASC5	DPY19L3	GGNBP2	KLHL9	NCOA6	PCDH10	RIMS1	SPATA13	VPS54
AGAP2	CBX4	DRAM2	GLI3	KLHL14	NDC80	PCOLCE	RNF38	SPEN	WDFY3
ANGPTL3	CD96	DSCAM	GRB14	KPNA1	NDRG4	PDGFD	ROS1	SRCAP	WDR63
ANK2	CDC42BPB	DYNC1H1	GRIA1	KRT34	NEBL	PEX5L	SCARA3	SRPK2	WHSC1
ANO5	CDC73	DYRK1A	GRIN2B	KYNU	NECAB1	PHF2	SCN1A	STXBP5	XPO5
APH1A	CECR2	ECM2	GRM7	L1CAM	NFASC	PIGZ	SCN2A	SUV420H1	ZBED4
ARHGEF11	CEP55	EDNRA	HAPLN3	LGALS3BP	NFE2L2	PIK3R2	SCN7A	SV2B	ZNF238
ARID1A	CEP164	EHD2	HDLBP	LMTK3	NFIA	PLA1A	SEC31B	SYNE2	ZNF426
ARID1B	CHD5	EP300	HIST1H2AE	LRFN5	NFYA	PLEKHG5	SERPINB9	SYNGAP1	ZNF451
ARNTL2	CHD8	EP400	HSPA13	MAK	NIF3L1	PNPLA8	SETBP1	TAF4	ZNF462
ASH1L	CHN2	EPB49	IQGAP2	MAP7	NINL	POGZ	SETD5	TAS2R13	ZNF559
ASZ1	CLCA3P	ERBB2IP	ISLR	MAPT	NKAIN2	PPM1D	SGOL2	TAS2R20	ZNF594
ATP1B1	CMIP	ERI1	ITGA5	MBD5	NOL6	PPPDE1	SHANK2	TBC1D23	ZNF638
ATP10D	CNOT3	EXOSC2	JUP	MBNL1	NOP14	PRKD2	SHB	TBR1	ZNF774
AXL	COBL	FAM59A	KBTBD4	MCAM	NOTCH2	PRPF39	SIAE	TCF3	
BCL11A	CORO1B	FAM190A	KDM3A	ME1	NOTCH3	PRPF40A	SKI	TIMELESS	
BIRC6	CSDE1	FBN1	KDM4B	MEGF10	NR2F1	PTEN	SLC6A1	TMBIM6	
BRSK2	CSNK1E	FBXL13	KDM5B	MGRN1	NR3C2	PTMS	SLC24A3	TMEM178	
BRWD1	CSTF2T	FBXO27	KDM6B	MIB1	NRG4	PTPRM	SLC25A39	TNPO3	
BUB1B	CTTNBP2	FEZF2	KIAA0182	MLL3	NRM	QRICH1	SLC35B2	TOP2A	
C1ORF173	CUL3	FOXG1	KIAA1407	MSH4	NRXN1	RAB2A	SLCO1B3	TRIO	
C9ORF68	DBX2	FOXP1	KIF1A	MTMR12	NSUN7	RAD54B	SLCO6A1	TROVE2	
C11ORF30	DHX57	GABRA1	KIF11	MYH10	NUSAP1	RANBP17	SMARCA4	TTC14	

Table A.22: Gene list for Simons Foundation Autism Research Initiative (SFARI) Category 4[4]. The list contains 339 genes with minimal evidence.

ABAT	CACNA1F	DAGLA	FAM47A	HTR1B	MARK1	NLGN4Y	PPP2R5D	SIN3A	TBL1X
ABCA7	CACNA1G	DDX53	FAN1	HTR3A	MBD1	NPAS2	PRKDC	SLC1A1	TBX1
ABCA10	CACNA1I	DIXDC1	FBXO33	HTR3C	MBD3	NR1D1	PRPF39	SLC4A10	TDO2
ABCA13	CADM1	DLGAP2	FBXO40	HYDIN	MBD4	NR2F1	PSD3	SLC6A4	TECTA
ACE	CADPS2	DLX2	FCRL6	ICA1	MCM4	NRCAM	PTBP2	SLC6A8	TM4SF19
ADCY5	CAMK2A	DLX6	FEZF2	IL1R2	MCPH1	NRP2	PTGS2	SLC7A7	TOP3B
ADK	CAMK4	DMXL2	FHIT	IL1RAPL1	MDGA2	NRXN2	PTPN11	SLC9A9	TPO
ADORA2A	CAPN12	DNAH3	FRK	IL1RAPL2	MEF2C	NTNG1	QRICH1	SLC22A9	TRIM33
ADORA3	CARD11	DNAH10	GABRA4	IMMP2L	MEGF10	NTRK1	RAB39B	SLC22A15	TSHZ3
ADRB2	CASC4	DNAH17	GALNT13	INPP1	MEGF11	NTRK3	RAD21L1	SLC25A12	TSPAN7
AFF2	CASK	DNER	GALNT14	IQGAP3	MIB1	NUP133	RAPGEF4	SLC25A39	TSPAN17
AGAP1	CCDC64	DOCK1	GAS2	IQSEC2	MKL2	NXPH1	REEP3	SLC27A4	TTC25
AGBL4	CCDC88C	DPP4	GLIS1	ITPR1	MNT	ODF3L2	RERE	SLCO1B3	TTN
AGMO	CCDC91	DPP6	GLO1	JMJD1C	MSANTD2	OFD1	RFX3	SMC3	TUBGCP5
AGTR2	CD99L2	DPYD	GNB1L	KATNAL1	MSR1	OR1C1	RGS7	SMG6	UBE2H
APBA2	CD276	DRD3	GPC6	KCNK7	MTF1	OR2M4	RHOXF1	SNAP25	UNC13A
AR	CDH8	DST	GPR37	KCNMA1	MTHFR	OR2T10	RIMS3	SND1	UNC80
ARHGAP32	CDH9	DUSP15	GPR85	KCTD13	MUC4	OTX1	RIT2	SNTG2	USH2A
ARHGEF9	CDH10	DVL3	GPX1	KDM4B	MUC12	OXT	RNF38	SPP2	VASH1
ARNT2	CDH22	DYDC1	GRID2	KHDRBS2	MYH4	PCDH9	RNF135	SRGAP3	VSIG4
ASMT	CELF6	DYDC2	GRIN2A	KLF16	MYO5C	PCDH10	RP11-1407O15.2	SSPO	WNK3
ATP6V0A2	CHD1	EIF3G	GRM7	KRR1	MYO16	PCDH11X	RPL10	SSRP1	WNT1
ATRX	CHRM3	EIF4E	GSTM1	KRT26	NAALADL2	PCDH15	RPS6KA2	ST7	XPO1
AVPR1B	CLN8	ELAVL2	GTF2I	LAMA1	NACC1	PDCD1	RPS6KA3	STK39	YWHAE
AZGP1	CLTCL1	EN2	GUCY1A2	LEP	NBEA	PER1	SASH11	STX1A	ZNF292
BIRC6	CMIP	EP300	HLA-A	LILRB2	NCKAP5	PIK3CG	SATB2	STYK1	ZNF462
BRCA2	CNGB3	EPC2	HLA-B	LRBA	NCOR1	PITX1	SCFD2	SYAP1	ZNF517
BRD4	CNTNAP5	EPPK1	HLA-G	LRFN2	NDNL2	PLN	SCN4A	SYN1	ZNF548
BST1	COL28A1	ERG	HNRNPH2	LRFN5	NDUFA5	PLXNA3	SCP2	SYN2	ZNF559
BZRAP1	CPT2	ERMN	HNRNPU	LRP2	NEO1	POLA2	SDC2	SYNE1	ZNF626
C3orf58	CTNNA3	ESR2	HOMER1	LRRC1	NFIA	POMT1	SETDB1	SYNJ1	ZNF713
C4B	CX3CR1	ESRRB	HRAS	LZTS2	NIPA1	POT1	SETDB2	SYT17	ZNF774
C15orf62	CYLC2	EXT1	HS3ST5	MAOA	NIPA2	PPP1R1B	SEZ6L2	TANC2	ZWILCH
CA6	CYP11B1	FABP5	HSD11B1	MAPK3	NLGN2	PPP2R1B	SGSM3	TBC1D5	

Appendix B

Supplementary Figures

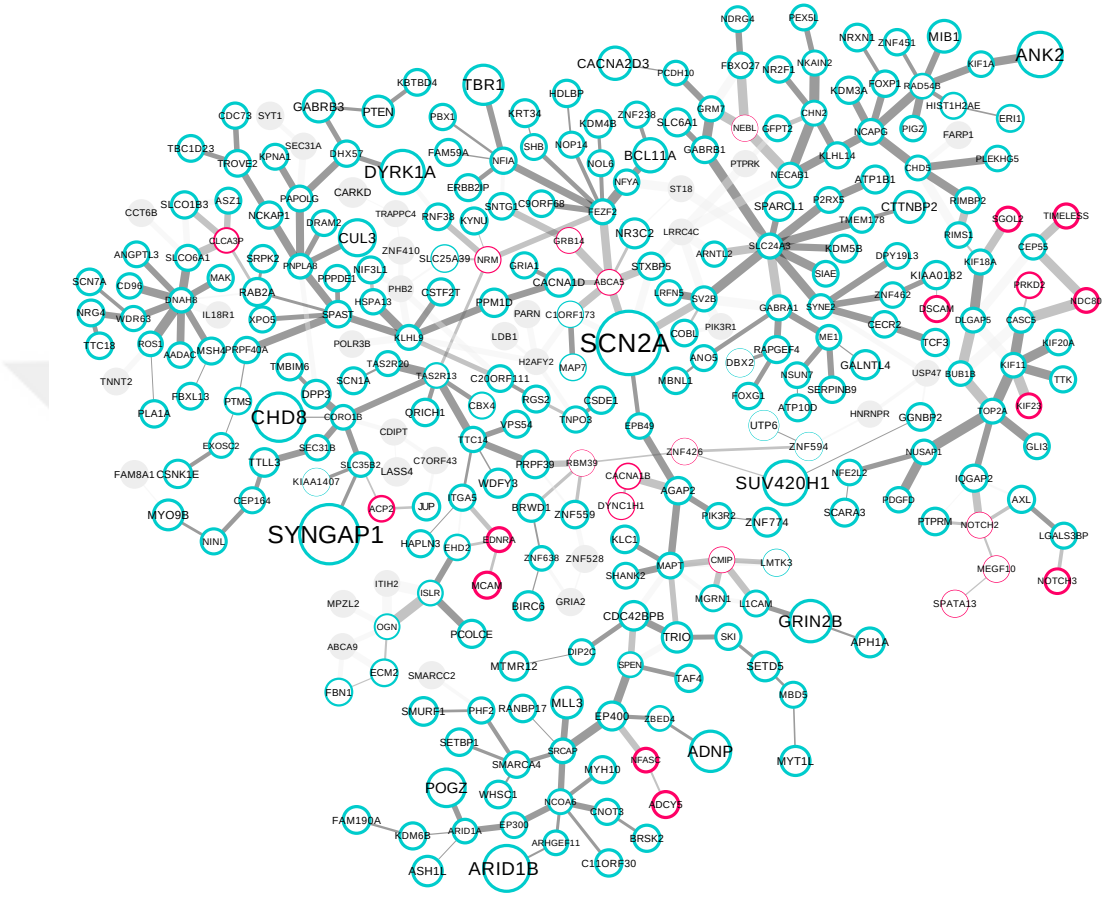

Figure B.1: Figure visualizes ST-St. PFC(1-3)+(4-6) network when it is laid over ST-St. PFC(4-6)+10%, by taking their union and coloring according to membership. Turquoise bordered genes are common in ST-St. PFC(1-3)+(4-6) and ST-St. PFC(4-6)+10%. Pink bordered genes are only in ST-St. PFC(1-3)+(4-6), highlighting the effect of using the temporal dimension and information coming from PFC(1-3). Grey genes are only in PFC(4-6)+10% and were not included by ST-St. PFC(1-3)+(4-6). Size of a node indicates its significance w.r.t. its TADA q-value from [6] (the larger, the more significant). The thickness of an edges represents the correlation coefficient between the gene pair (the thicker, the higher). Visualized using CoSE layout [43] in Cytoscape [44].

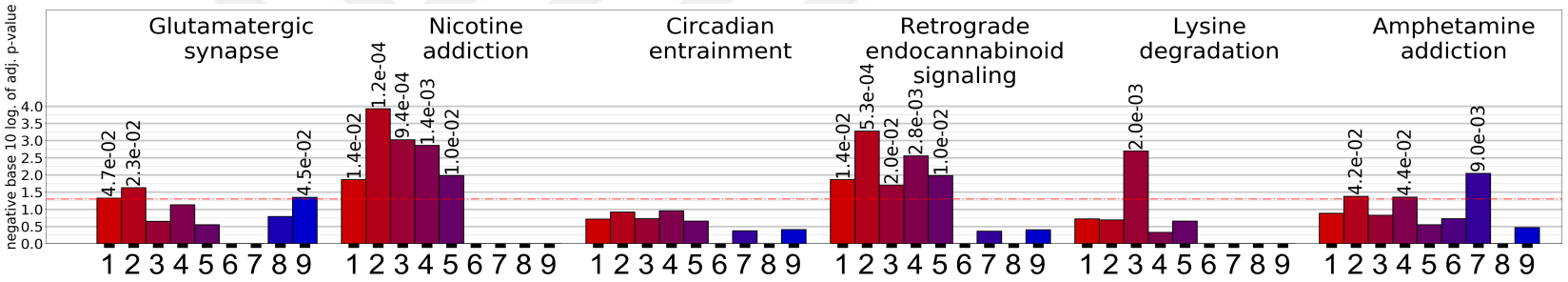

Figure B.2: Figure compares enrichment of the predicted gene sets of ST-Steiner, NETBAG, MAGI, and DAWN in 6 KEGG pathways. 1: ST-Steiner Every(1-3)+PFC(3-5); 2: ST-Steiner Every(1-3)+PFC(4-6); 3: NETBAG; 4: DAWN PFC 3–5; 5: DAWN PFC 4–6; 6: MAGI Best1; 7: MAGI Ext1; 8: MAGI Best2; 9: MAGI Ext2. The pathways shown here are all of the enriched (i.e. adjusted p-value < 0.05) pathways for the union of SFARI Category 1 and 2 genes (83 genes), in the order of increasing adjusted p-value from left to right. The red, dashed line shows the significance threshold 0.05. Enrichr tool and reported adjusted p-values for the gene-set library KEGG 2016 are used.

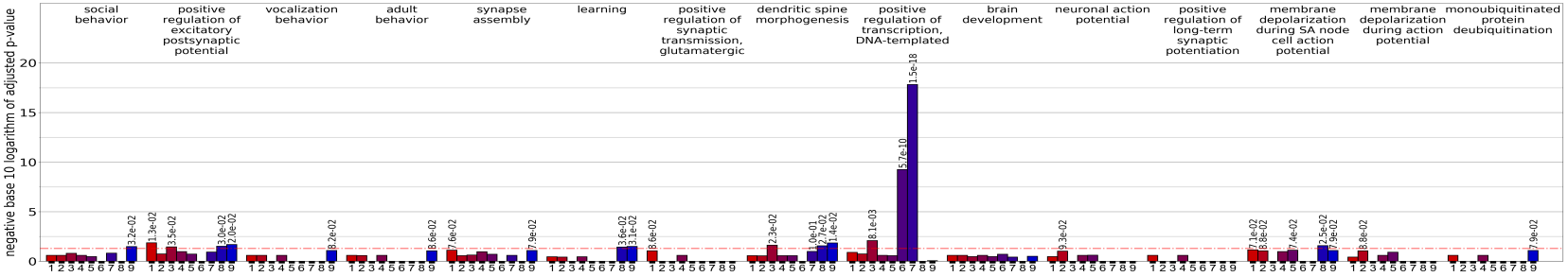

Figure B.3: Figure compares enrichment of the predicted gene sets of ST-Steiner, NETBAG, MAGI and DAWN in 15 GO Biological Process terms. The terms given here are the top-15 terms that SFARI Category 1 and 2 genes are enriched in (i.e. adjusted p-value < 0.05), in the order of increasing adjusted p-value from left to right. 1: ST-Steiner Every(1-3)+PFC(3-5); 2: ST-Steiner Every(1-3)+PFC(4-6); 3: NETBAG; 4: DAWN PFC 3–5; 5: DAWN PFC 4–6; 6: MAGI Best1; 7: MAGI Ext1; 8: MAGI Best2; 9: MAGI Ext2. The red, dashed line shows the significance threshold 0.05. Enrichr and reported adjusted p-values for the gene-set library GO Biological Process 2017 are used.

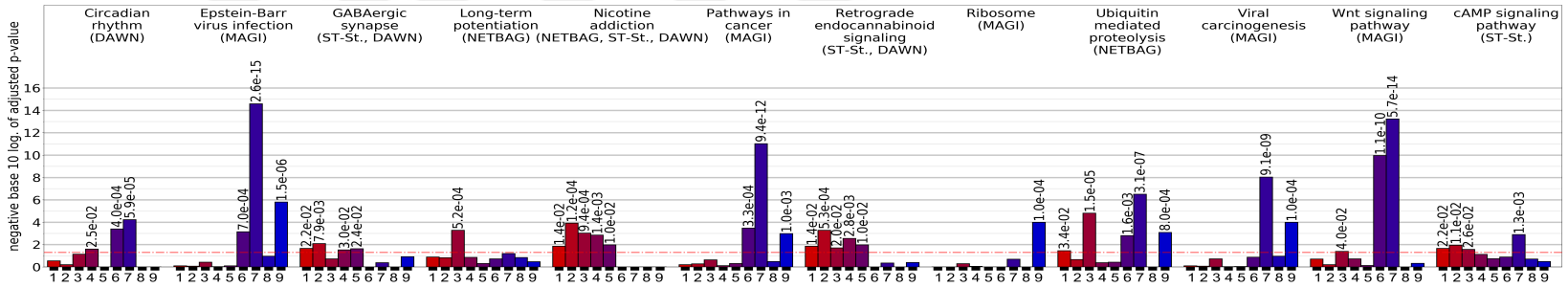

Figure B.4: Figure compares enrichment of the predicted gene sets of ST-Steiner, NETBAG, MAGI and DAWN in 12 KEGG pathways. 1: ST-Steiner Every(1-3)+PFC(3-5); 2: ST-Steiner Every(1-3)+PFC(4-6); 3: NETBAG; 4: DAWN PFC 3–5; 5: DAWN PFC 4–6; 6: MAGI Best1; 7: MAGI Ext1; 8: MAGI Best2; 9: MAGI Ext2. The pathways shown here are the union of the top-3 pathways in which the gene sets are individually enriched, excluding MAGI Best1 and Best2 for fairness (resulting in 2 gene sets per method determining the pathway list, except for NETBAG which contributes with 1 prediction only). For every pathway, we indicate the method or methods which have at least one gene set that picks the pathway in parenthesis. The red, dashed line shows the significance threshold 0.05. (Enrichr tool and reported adjusted p-values for the gene-set library KEGG 2016 are used).

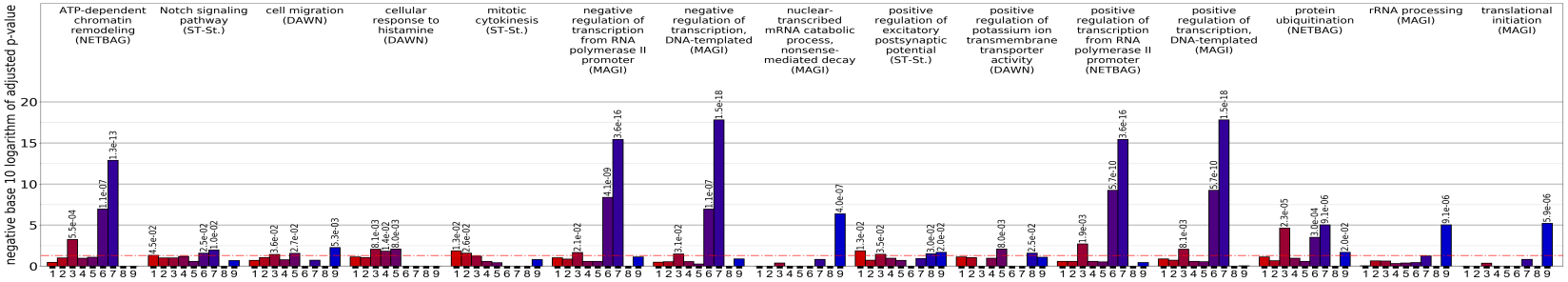

Figure B.5: Figure compares enrichment of the predicted gene sets of ST-Steiner, NETBAG, MAGI and DAWN in 15 GO Biological Process terms. 1: ST-Steiner Every(1-3)+PFC(3-5); 2: ST-Steiner Every(1-3)+PFC(4-6); 3: NETBAG; 4: DAWN PFC 3–5; 5: DAWN PFC 4–6; 6: MAGI Best1; 7: MAGI Ext1; 8: MAGI Best2; 9: MAGI Ext2. The terms shown here are the union of the top-3 terms in which the gene sets are individually enriched, excluding MAGI Best1 and Best2 for fairness (resulting in 2 gene sets per method, except for NETBAG which contributes with 1 prediction only). For each term, the methods that pick the term in at least one of their top-3 lists are indicated in parenthesis. The red, dashed line shows the significance threshold 0.05. (Enrichr tool and reported adjusted p-values for the gene-set library GO Biological Process 2017 are used).