Integration of DNA Methylation, Copy Number Variation, and Gene Expression for Gene Regulatory Network Inference and Application to Psychiatric Disorders

Dong-Chul Kim*, Mingon Kang†, Baoju Zhang‡, Xiaoyong Wu‡, Chunyu Liu§ and Jean Gao†
*Department of Computer Science, University of Texas at Pan American, Edinburg, Texas 78541, USA
†Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, Texas 76019, USA
‡College of Electronics and Communication Engineering, Tianjin Normal University, Tianjin, China 300074
§Department of Psychiatry, University of Illinois at Chicago, Chicago, IL 66012, USA
Email: kimd1@utpa.edu, mingon.kang@mavs.uta.edu, liucy@uic.edu, gao@uta.edu

***Abstract*—Biological network inference is a crucial problem to solve in Bioinformatics as most of biological process are based on biomolecular interactions. Many researchers have worked on especially the inference of gene regulatory networks where a node and edge represent a gene and regulation relationship respectively assuming that a gene can regulate another gene indirectly. However, a gene expression level can be influenced by not only genes and proteins but also other biological factors. Therefore, the inference could be more effective if those factors are considered in gene regulatory network inferences. In this paper, we propose an integrative approach to infer gene regulatory networks where a gene can be regulated by not only gene and but also DNA Methylation and copy number variation. It is assumed that a gene can be directly regulated by a single DNA Methylation and copy number variation at most. The simulation results show that our method outperforms popular and state-of-the-art methods of biological network inference. In addition, we applied the proposed method to psychiatric disorder data. The inferred networks provide the relationships within a set of genes that are more likely to be regulated by DNA Methylation and copy number variation of the genes.**

I. Introduction

Biological process is based on gene expression programs that regulate the transcription of thousands of genes. Understanding gene regulation is crucial to gain insights in biological processes of disease. In order to understand gene regulatory mechanism, a number of researches have been conducted by using high-throughput microarray expression data. Diverse computational and statistical approaches [1] has been proposed to decipher regulatory relationships between genes. However, inferring Gene Regulatory Network (GRN) is still a challenge because gene expressions are regulated by many complex factors such as copy number variation (CNV) and DNA Methylation (DM) of the genes. GRN inferences were normally based on gene expression data only but not other data types that could be related to gene expression. For better quality of network inference, it is demanded to integrate heterogeneous data with gene expression.

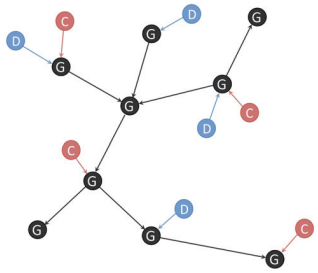


Figure 1: Example of DCGRN: blue, red, and black nodes indicate DM, CNV, and gene respectively

Presently, it is known that DM could play a role in regulating gene expression but its association with gene expression is not fully discovered yet. Since DM can prevent TF from binding to promoter, negative correlation has been generally observed with gene expression, but it is also discovered that the increased DM is positively correlated with high level gene expression in some cases [2]. CNV can also influence gene expression level as another important factor of gene expression. Many researches have been performed to elucidate the association between CNV and gene expression in target disease [3]. It is similar to DM that there are both positive and negative correlation between CNV and gene expression [4]. We note that these two factors can be a direct regulator of the gene in GRN, and network inference with CNV and DM data could perform better than gene expression data only. To verify the potentials of DM and CNV in GRN inference, we explore the impact of integrating gene expression and these two factors in gene regulatory network inference in this paper.

Gene expression data (e.g. DNA microarray) allow network inference by studying the regulatory patterns between expression levels of two genes. In order to construct a network, many computational methods have been suggested including Mutual Information (MI), Bayesian Network (BN), and L1 regularized linear regression (Lasso). These popular methods have their own advantages and disadvantages. For example, MI is fast and simple limited to indirected net-

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.71

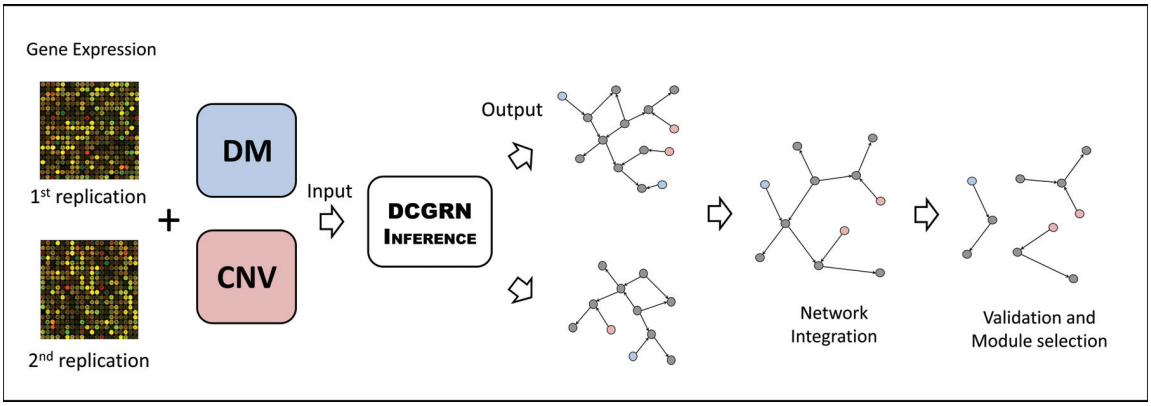


Figure 2: Overview of DCGRN inference for psychiatric disorder

work. BN provides directed acyclic network inference but requires high computational cost and a number of samples. In this paper, we employed Lasso-based network inference complemented advantages without indirection limitation and computational weakness. Given a network model where each node represents gene, DM, and CNV, Lasso gives sparsity in network inference (the gene regulatory network with DM and CNV is referred as to DCGRN in this paper). The strength of sparsity is determined by performing cross validation and edge weight is optimized by using coordinate descent algorithm. In addition, DCGRN model assumes that DM and CNV can linearly regulate gene expression but not be regulated by genes. The example of DCGRN is described in Figure 1. The algorithm we proposed is also based on statistical test to select significant edges from DN and CNV to gene. In the experimental results section, the performance of the method was tested with synthetic DCGRNs and simulated data sets. DCGRN model and Lasso-based inference method are applied to psychiatric disorder data that consist of CNV, DM, and two replications of gene expression. Figure 2 describes the procedures of DCGRN inference. First, two DCGRNs are independently inferred by using two replication data and common DM and CNV. Secondly, two DCGRNs are integrated by including only edges that two networks commonly have. Lastly, significant edges are selected the integrated network by statistical test and validated by using related literatures and database. Final results provide potential biomarkers associated with mental disorders as well as regulatory relationships between gene, DM, and CNV.

There are three contributions in this work. First, DCGRN is the first network model for the integration of DM, CNV, and gene expression. DCGRN provides not only gene regulatory relationships but also effect of DNA methylation and copy number alterations on gene expression. Secondly, the lasso-based method to infer DCGRN is evaluated by using simulation data and perform effectively. Lastly, our network model and method is applied to psychiatric disorder data and demonstrate the ability to provide regulatory relationships of biomarkers by integrating DM and CNV with gene expression data.

II. Method

A. DCGRN Model

DCGRN is a directed network that encodes the regulatory relationships over a set of genes, DMs, and CNVs. Let $X \in \mathbb{R}^{m\times n}$ denote the matrix of gene expression levels of m genes and n samples. The matrix $D \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{m\times n}$ also denotes DMs and CNVs of the genes. The three data matrix are defined as $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]$, $D = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_m]$, and $C = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_m]$ where $\mathbf{x}_i, \mathbf{d}_i, \mathbf{c}_i$ are ith row vector of data matrix X, D, and C respectively. DCGRN is defined as follows.

$$\mathbf{x}_i = \mathbf{b}_i X + \mathbf{w}_i D + \mathbf{f}_i C + \mu_i + \varepsilon_i, \tag{1}$$

where $\mathbf{b}_i$, $\mathbf{w}_i$, and $\mathbf{f}_i$ denotes ith row vector of adjacency matrix $B\in\mathbb{R}^{m\times m}$, $W\in\mathbb{R}^{m\times m}$, and $F\in\mathbb{R}^{m\times m}$ respectively;μ_i is a model bias; and ε_i is a residual. As we assume there is no self-regulation (self-loop edge), $b_{ii} = 0, \forall i = 1, ..., m$ where b_{ii} denotes ith element of $\mathbf{b}_i$. It is also assumed that there is no cycle. Especially, the case of two nodes cycle (i.e. both b_{ij} and b_{ji} are non-zero) is not considered in DCGRN. The parameter b_{ij} represents the activation(positive) or deactivation(negative) weight of edge from jth gene to ith gene. If b_{ij} is zero, there is no edge from from jth gene to ith gene. In addition, it is assumed that a gene can be directly influenced by only DM and CNV that belong to the gene but no other genes. It means that only diagonal elements of D and C can be non-zero. w_{ii} and f_{ii} represent the regulatory weight of DM and CNV of ith gene. Our goal in this model is to find B, W, and F that best fit to observed gene expression, DM, and CNV data. In other words, the problem is to estimate $\mathbf{b}_i$, $\mathbf{w}_i$, and $\mathbf{f}_i$ that minimize ε_i. After the bias is removed by mean centering, (1) can be restated in a least square minimization problem defined as

$$\min_{\mathbf{b}_i, \mathbf{w}_i, \mathbf{f}_i} ||\mathbf{x}_i - \mathbf{b}_i X - \mathbf{w}_i D - \mathbf{f}_i C||_2^2 \tag{2}$$

where $||\cdot||_2$ denotes 2 norm.

B. Sparse DCGRN

Regression tends to include as many genes as possible to explain the expression level of target gene. In order to avoid the overfitting and exclude most of edges that are not most likely to be true edge, many sparse linear models have been used in gene regulatory network inference. The sparse penalty forces b_{ij} to be zero when gene i and regulator j are not more likely interacting than other possible regulator genes. For Sparse DCGRN (SDCGRN), 1 norm sparse penalty is also applied to all edges (B,W,F) as following:

$$\min_{\mathbf{b}_i,\mathbf{w}_i,\mathbf{f}_i} ||\mathbf{x}_i - \mathbf{b}_i X - \mathbf{w}_i D - \mathbf{f}_i C||_2^2 + \lambda_1||\mathbf{b}_i||_1 + \lambda_2||\mathbf{w}_i||_1 + \lambda_3||\mathbf{f}_i||_1 \quad (3)$$

where λs are penalty weights. Since only diagonal elements of W and F are non-zero, (3) can be re-expressed as

$$L(\boldsymbol{\beta}_i) = \min_{\beta_i} ||\mathbf{x}_i - \boldsymbol{\beta}_i Y||_2^2 + \lambda||\boldsymbol{\beta}_i||_1 \quad (4)$$

where

$$\boldsymbol{\beta}_i = [b_{i1}, b_{i2}, \ldots, b_{ii-1}, b_{ii+1}, \ldots, b_{im}, w_{ii}, f_{ii}], \quad (5)$$

$$Y = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_m, \mathbf{d}_i, \mathbf{c}_i]. \quad (6)$$

Since the objective function (4) is convex, $\boldsymbol{\beta}_i$ can be optimized by using coordinate descent algorithm given λ. To find the optimal $\boldsymbol{\beta}_i$, the derivative of (4) with respect to β_{ij} is induced as follows.

$$\partial L/\partial\beta_{ij} = -\mathbf{y}_j(\mathbf{x}_i^\top - Y^\top\boldsymbol{\beta}_i^\top) + \lambda\partial_{\beta_{ij}}||\boldsymbol{\beta}_i||_1 \quad (7)$$

$$= -\mathbf{y}_j(\mathbf{x}_i^\top - Y_{(-j)}^\top\boldsymbol{\beta}_{i(-j)}^\top - \mathbf{y}_j^\top\beta_{ij}) + \lambda\partial_{\beta_{ij}}||\boldsymbol{\beta}_i||_1 \quad (8)$$

where $Y_{(-j)}$ denotes matrix Y whose jth row is removed, and $\mathbf{y}_j$ is jth row vector of Y. Then, (8) can be simplified to

$$-c_{ij} + a_{ij}\beta_{ij} + \lambda\partial_{\beta_{ij}}||\boldsymbol{\beta}_i||_1 \quad (9)$$

where c_{ij} denotes $\mathbf{y}_j(\mathbf{x}_i^\top - Y_{(-j)}^\top\boldsymbol{\beta}_{i(-j)}^\top)$ and a_{ij} denotes $\mathbf{y}_j\mathbf{y}_j^\top$. Given a λ, β_{ij} can be calculated by using the coordinate descent algorithm based on subgradient as following:

$$\beta_{ij} = \begin{cases} (-c_{ij} - \lambda)/a_{ij} & \text{if } c_{ij} < \text{-}\lambda, \\ 0 & \text{if } |c_{ij}| \le \lambda, \\ (-c_{ij} + \lambda)/a_{ij} & \text{if } c_{ij} > \lambda. \end{cases} \quad (10)$$

In Algorithm 1, the overall procedure to solve (4) is described.

C. Tests of Statistical Significance for Edge Selection

Although the sparsity is applied to $\boldsymbol{\beta}_i$, w_{ii} and f_{ii} (i.e. β_{im} and β_{im+1}) could be non-zero with small coefficient value as false positive regulation. In order to reduce the false positive rate of inferred edges from DM and CNV to corresponding gene, w_{ii} and f_{ii} are re-estimated when they are non-zero. First, a vector $\boldsymbol{\beta}_i'$ is set with only non-zero coefficients of the estimated β_i, then ordinary least square is applied to $\boldsymbol{\beta}_i'$. The corresponding β_{ij}' for non zero w_{ii} or f_{ii} is statistically tested whether coefficient might be zero (i.e. null hypothesis). If p-value of t-score in t distribution is not less than 0.0001 as a significance level, we set the coefficient to zero.

Algorithm 1 Sparse DCGRN algorithm

```
procedure SDCGRN(x_i, Y, λ, ε)
    initialize β_i to zero vector
    while err > ε do
        β_i^old ← β_i
        for j ← 1, M_s do
            Update β_ij via (10)
        end for
        err ← ||β_i^old − β_i||_2
    end while
    return β_i
end procedure
```

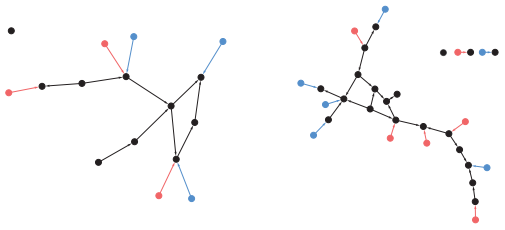
D. DCGRN Inference

DCGRN inference consists of three steps. First, initial network is inferred by using SDCGRN. Secondly, all edges from DM and CNV to gene in the sparse network are statistically tested. Lastly, SDCGRN is performed again with only non-zero elements of B, W, and F of second step. Then, non-zero coefficient element of $\boldsymbol{\beta}_i$ is re-estimated by applying ordinary least square since SDCGRN is only to select edges like feature selection but not estimate coefficient value. The re-estimated B is finalized by removing edges whose coefficient value is less than heuristically predefined threshold. Although it is assumed that there is no two node cycle, the final result could include the cycle as DCGRN inference does not have the ability to avoid the cycle. In addition, the effect of third step depends on how many genes are regulated by DM and CNV, and it is assumed that a gene can be regulated by only a single DM and CNV of the gene. The indirect regulation by DM and CNV of other genes is not considered in DCGRN inference.

III. Experimental Results

A. Simulation Studies

1) Random network and synthetic data sets: The evaluation of the DCGRN inference is performed by simulations based on random network and synthetic data. m indicates the number of genes and is set to 10, 20, and 30. $m\times m$ matrix B is initialized to zero matrix, then elements of B are randomly selected avoiding any cycle. The parameter E_G decides the number of inbound edges per gene on average, which means higher E_G make the network more complex. The selected b_{ij} has random coefficient value that is uniformly distributed over 0.5$\sim$1 or -0.5$\sim$-1. Similarly W and F are initialized to zero matrix, then diagonal elements (wii and



(a) m=10,E_G=1,DC=0.3 (b) m=20,E_G=1,DC=0.3

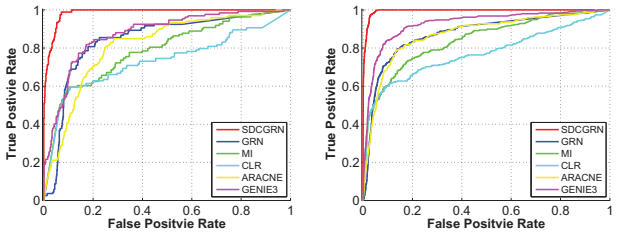
Figure 3: Example of random networks with different parameter settings: m, E_G, and DC indicate the number of genes, average number of edges per node, and percentage of nodes that have DM and CNV respectively.

fii) are randomly selected. The parameter DC indicates the percentage of nodes that are regulated by DM and CNV. For example, if m and DC are 10 and 0.3 respectively, three (10×0.3) DMs and CNVs regulate corresponding genes (i.e. three diagonal elements of W and F are non-zero). The selected w_{ii} has random value like b_{ij}, and selected f_{ii} is set to 1. d_{ij} has random value that is uniformly distributed over 0$\sim$1. c_{ij} is randomly set as 0, 1, 2, 3, or 4 with the probabilities 0.03, 0.07, 0.87, 0.02 and 0.01, respectively. X is generated by calculating $X = (I-B)^{-1}(WD+FC+E)$ where E_{ij} is generated from Gaussian distribution with zero mean and variance 0.01. The number of samples for each network size is N=100, 200, 300, 400, 500. E_G is set to 1, 2, and 3. DC is set to 0.1, 0.3, 0.5, 0.7, and 0.9. Given data X, D and C, $\hat{B}$, $\hat{W}$, $\hat{F}$ are inferred, and then they are compared to true edges of B, W, F. Figure 3 displays the examples of random network with different parameters, m, E_G, and DC.

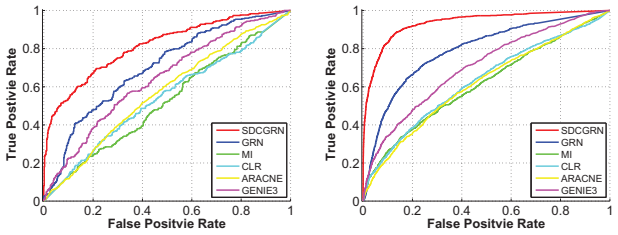
2) Evaluation results: For the evaluation, true positive (TP), false positive (FP), true negative (TN), and false negative (FN) edges are counted to measure the accuracy criteria, True Positive Rate (TPR) and False Discovery Rate (FDR) The performance of proposed method are compared to other network inference methods as follows.

- GRN: Lasso without DM and CNV
- MI: edge is scored by mutual information
- CLR: context likelihood relatedness [5].
- ARACNE: [6]
- GENIE3: top rank in DREAM 4 [7]
- DCGRN: the proposed method

Figure 4 is ROC curves to measure the performance of network inference in different parameters of network size and complexity (DC and number of samples are fixed to 0.3 and 100 respectively). In all parameter settings, the proposed method outperforms other methods. The performance of all methods is much better in E_G=1 (Figure 4(a), (b)) than E_G=3 (Figure 4(c), (d)) because indirect regulation could be more in more complex network. Regarding to network size, the performance of all methods tends to increase when the network size increases because TN is likely to be increase



(a) m=10,E_G=1,DC=0.3 (b) m=30,E_G=1,DC=0.3



(c) m=10,E_G=3,DC=0.3 (d) m=30,E_G=3,DC=0.3

Figure 4: ROC curve of the propsed method and compared methods with different parameter settings: m, E_G, and DC indicate the number of genes, average number of inbound edges per node, and percentage of nodes that have DM and CNV respectively. The number of samples is set to 100.

when the network size increases.

B. Application to Psychiatric Disorder Data

The psychiatric disorder data consists of gene expressions (two replications), DM, and CNV data. There are 87 case and 43 control samples measured from human brain. From 25833 genes of expression data, we selected 175 genes, which have both corresponding CNV and DM. When a gene is selected for a given CNV, it is assumed that a CNV can regulate a gene if the CNV location is overlapped with the location of the gene. In addition, if multiple genes can belong to a CNV, the gene that has the highest correlation coefficient with CNV is selected.

Figure 5 is the integrated network of the two replications with 16 genes that have high correlation coefficient ($>$0.2)between gene and DM. There are two DMs that are connected to genes, CSMD1 and CYP2E1. Also, there is a CNV that is connected the gene, SLC2A3. A number of literature support that CSMD1 and CYP2E1 are associated with schizophrenia but there is no report about associations with DM of the two genes [8], [9]. Although SLC2A3 is not known as a biomarker of psychiatric disorders, it has been reported that SLC2AS is related to brain disease, Huntington's disease [10]. We mainly note four interconnected genes, SELENBP1, FKBP5, CTNNA3, and SLC35F2 as potential biomarkers for psychiatric disorders. SELENBP1 are significantly interconnected to other genes. Many studies reported SELENBP1 as a biomarker for psychiatric disorders [11]. FKBP5 that is connected SELENBP1 is also known

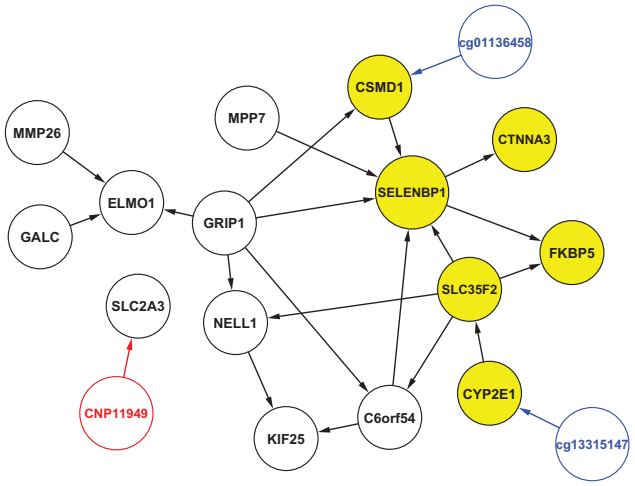


Figure 5: The integrated network of two replication data with 16 selected genes that are high correlated with corresponding DM and CNV. Yellow colored genes are the interconnected biomarkers for psychiatric disorder and brain disease.

as a biomarker for Depression [12]. Another two genes, CTNNA3 and SLC35F2, are not known as biomarker for psychiatric disorders but brain-related disease, Alzheimers disease [13], [14].

IV. Conclusion

In this paper, two biological factors, DM and CNV, are considered as a regulator that influence gene expression. In order to infer what DM and CNV could regulate a gene, the novel network inference method is proposed. In other words, our method provides gene regulatory network where DM and CNV are involved as a regulator of gene. The method is based on three steps of linear regression and statistical test. To evaluate the performance of the method, the synthetic data set are generated by using random networks. The performance is better than state-of-the-art methods in different parameter settings of random networks. We also applied the method to psychiatric disorder data in order to explore the network of genes that are likely to be regulated by DM and CNV.

References

[1] W. C. Young, A. E. Raftery, and K. Y. Yeung, "Fast bayesian inference for gene regulatory networks using scanbma," *BMC systems biology*, vol. 8, no. 1, p. 47, 2014.

[2] J. R. Wagner, S. Busche, B. Ge, T. Kwan, T. Pastinen, and M. Blanchette, "The relationship between dna methylation, genetic and expression inter-individual variation in untransformed human fibroblasts," *Genome biology*, vol. 15, no. 2, p. R37, 2014.

[3] G. E. Liu, L. Xu, and K. S. Huang, "Recent advances in studying of copy number variation and gene expression." *Gene Expression to Genetical Genomics*, no. 7, 2014.

[4] C. N. Henrichsen, E. Chaignat, and A. Reymond, "Copy number variants, diseases and gene expression," *Human molecular genetics*, vol. 18, no. R1, pp. R1–R8, 2009.

[5] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, "Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles," *PLoS biology*, vol. 5, no. 1, p. e8, 2007.

[6] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano, "Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC bioinformatics*, vol. 7, no. Suppl 1, p. S7, 2006.

[7] A. Irrthum, L. Wehenkel, P. Geurts *et al.*, "Inferring regulatory networks from expression data using tree-based methods," *PloS one*, vol. 5, no. 9, p. e12776, 2010.

[8] G. Donohoe, J. Walters, A. Hargreaves, E. Rose, D. Morris, C. Fahey, S. Bellini, E. Cummins, I. Giegling, A. Hartmann *et al.*, "Neuropsychological effects of the csmd1 genome-wide associated schizophrenia risk variant rs10503253," *Genes, Brain and Behavior*, vol. 12, no. 2, pp. 203–209, 2013.

[9] V. M. Steen, C. Nepal, K. M. Ersland, R. Holdhus, M. Nævdal, S. M. Ratvik, S. Skrede, and B. Håvik, "Neuropsychological deficits in mice depleted of the schizophrenia susceptibility gene csmd1," *PloS one*, vol. 8, no. 11, p. e79501, 2013.

[10] A. Vittori, C. Breda, M. Repici, M. Orth, R. A. Roos, T. F. Outeiro, F. Giorgini, E. J. Hollox *et al.*, "Copy-number variation of the neuronal glucose transporter gene slc2a3 and age of onset in huntington's disease," *Human molecular genetics*, vol. 23, no. 12, pp. 3129–3137, 2014.

[11] T. Kanazawa, G. Chana, S. J. Glatt, H. Mizuno, E. Masliah, H. Yoneda, M. T. Tsuang, and I. P. Everall, "The utility of selenbp1 gene expression as a biomarker for major psychotic disorders: replication in schizophrenia and extension to bipolar disorder with psychosis," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 147, no. 6, pp. 686–689, 2008.

[12] P. Zimmermann, T. Brückl, A. Nocon, H. Pfister, E. B. Binder, M. Uhr, R. Lieb, T. E. Moffitt, A. Caspi, F. Holsboer *et al.*, "Interaction of fkbp5 gene variants and adverse life events in predicting depression onset: results from a 10-year prospective community study," *American Journal of Psychiatry*, vol. 168, no. 10, pp. 1107–1116, 2011.

[13] A. Miyashita, H. Arai, T. Asada, M. Imagawa, E. Matsubara, M. Shoji, S. Higuchi, K. Urakami, A. Kakita, H. Takahashi *et al.*, "Genetic association of ctnna3 with late-onset alzheimer's disease in females," *Human molecular genetics*, vol. 16, no. 23, pp. 2854–2869, 2007.

[14] E. Bacchelli, F. Ceroni, D. Pinto, S. Lomartire, M. Giannandrea, P. DAdamo, E. Bonora, P. Parchi, R. Tancredi, A. Battaglia *et al.*, "A ctnna3 compound heterozygous deletion implicates a role for αt-catenin in susceptibility to autism spectrum disorder," *Journal of neurodevelopmental disorders*, vol. 6, no. 1, pp. 1–11, 2014.