35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Sparse Generalized Canonical Correlation Analysis for Biological Model Integration: A Genetic Study of Psychiatric Disorders

Mingon Kang[1], Baoju Zhang[3], Xiaoyong Wu[3], Chunyu Liu[2], and Jean Gao[1]

Abstract— **In the post-genomic era, unveiling causal traits in the complex mechanisms that involve a number of diseases has been highlighted as one of the key goals. Much research has recently suggested integrative approaches of both genome-wide association studies (GWAS) and gene expression profiling-based studies provide greater insight of the mechanism than utilizing only one. In this paper, we propose a novel method, sparse generalized canonical correlation analysis (SGCCA), to integrate multiple biological data such as genetic markers, gene expressions, and disease phenotypes. The proposed method provides a powerful approach to comprehensively analyze complex biological mechanism while utilizing the multiple data simultaneously. The new method is also designed to identify a few of the elements significantly involved in the system among a large number of elements within the variable sets. The advantage of the method as well lies in the output of easily interpretable solutions. To verify the performance of SGCCA, we performed experiments with simulation data and human brain data of psychiatric diseases. Its capability to detect significant elements of the sets and the relations of the complex system is assessed.**

I. INTRODUCTION

Discovering causal traits in complex mechanisms involving multi diseases is one of the key goals in the post-genomic era. GWAS have achieved successes in uncovering significantly important genetic markers of interests. Nevertheless, the majority of GWAS perform the analysis with individual loci independently, which makes a single genetic variant explanation only a small proportion of the phenotypic variations, and often fails to find statistically significant variations after multiple testing corrections. Alternatively, a number of research studies for expression quantitative trait loci (eQTL) has uncovered the genetic traits via gene sets or pathway association analysis. However due to two specific limitations, first that global assays of gene expressions tend to be biased toward more expressed measurements, and secondly that the relationships between gene expressions and disease phenotypes are ignored in the model, the integrative approaches taking into account multiple forms of data can provide great insights to capture associations comprehensively.

The integrations of GWAS and gene expression profiling-based studies (e.g., eQTL) have been attempted recently, and they have emphasized the important roles of the methodology [1], [2], [3], [4]. Most of the integrative approaches used step-by-step processes to integrate the data. Hsu et al. performed a four-stage approach, in which meta-analysis performed separately in each stage to prioritize the significant genes of interests [2]. Xiong et al. computed scores from expression-based tests and SNP (single-nucleotide polymorphism)-based tests separately and integrated the scores using z-score sum, Fisher's method, and rank sum [3]. However, the limitations of these attempted integration methods are that some significant elements can be filtered out through the various stages while analyzing the data in pieces.

In this paper, we propose a sparse version of generalized canonical correlation analysis (SGCCA) to focus on two problems. First, we need a method to integrate biological data such as genetic markers, gene expressions, and disease phenotypes while analyzing them simultaneously. Rather than the investigation of each study independently and combining them in the traditional integrative methods, the proposed integrative approach can leverage the power to identify the pathway of the biological system by complementing the lack each has. For the problem, we adapted the generalized canonical correlation analysis (GCCA) that provides a powerful approach to dissect complicated mechanisms in which multiple variables cooperate associatively. Secondly, we need to provide easily interpretable solutions of GCCA as well as taking into account the group effects of the data since only a few elements of the data sets tend to be significantly associated to the traits of interests. Moreover GCCA lacks the power, especially when the number of independent variables is much larger than the number of samples. For this reason we propose a novel sparse method of GCCA (SGCCA) and take advantage of the integration of biological data and knowledge to conclusively identify the traits associated in the mechanism.

II. METHODS

Let $\mathbf{X}$ be K observable variables concerning the relationships between the blocks, i.e., $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K\}$. The observations are obtained from n samples, and can be represented by latent variables ($\boldsymbol{\xi}$) which are the linear combinations of observable variables.

$$\boldsymbol{\xi}_k = \mathbf{X}_k \mathbf{v}_k, \tag{1}$$

where $\mathbf{v}_k$ is a coefficient vector of the $\mathbf{X}_k$. p_k indicates the number of independent variables of each block, e.g., $\mathbf{X}_k \in \Re^{n \times p_k}$.

[1]M. Kang is with the Department of Computer Science and Engineering, University of Texas at Arlington, TX 76019, USA {mingon.kang, gao} at mavs.uta.edu

[2]C. Liu is with Faculty of the Department of Psychiatry, University of Illinois at Chicago, IL 60612, USA liucy at uic.edu

[3]Baoju Zhang and Xiaoyong Wu are with the School of Physics and Electronic Information, Tianjin Normal University, Tianjin, China

978-1-4577-0216-7/13/$26.00 ©2013 European Union

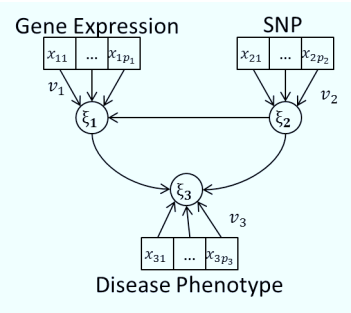


Fig. 1: A conceptual model of the integration of multiple data such as SNPs, gene expressions, and disease phenotypes. Independent variables are represented by squares and latent variables by circles

The conceptual model for the system must be designed in advance as shown in Fig. 1. In the model, C indicates the connectivity between the latent variables ($C \in \Re^{K\times K}$), and similarly, $h_{k,j}$ shows the set of all blocks such that the arrow points from l block to k block.

The model includes two internal relation models such as outer relations and inner relations. The outer relations takes into account the relations between the independent variables and their latent variable (2), while the inner relations represents the relations between latent variables of the blocks (3).

$$\mathbf{X}_k = \mathbf{b}_k^0 + \boldsymbol{\xi}_k \mathbf{b}_k^\top \quad (2)$$

$$\boldsymbol{\xi}_k = \boldsymbol{\gamma}_k^0 + \sum_{k\neq l, l=1}^{K} h_{k,l}\gamma_{k,l}\boldsymbol{\xi}_l, \quad (3)$$

where $\mathbf{b}_k^0$ and $\boldsymbol{\gamma}_k^0$ are the intercepts of the relations.

We find the associations of significant elements of the blocks maximizing the total correlations. The solutions of GCCA have been proposed [5], [6], [7]. In order to estimate the latent variables, we use the Wold's procedure rather than Lohmoller's since the Wold's procedure guarantees its convergence [7].

In the Wold's procedure, we introduce the elastic net penalization into the multiple regression when computing $\mathbf{v}_k^{(s)}$ for the sparsity and the group effect [8].

$$\mathbf{v}_k^{(s)} = \arg\min_{\mathbf{v}_k^{(s)}} |\boldsymbol{\xi}_k - \mathbf{X}_k\mathbf{v}_k^{(s)}|^2 + \lambda_1 \sum |\mathbf{v}_k^{(s)}| + \lambda_2 \sum ||\mathbf{v}_k^{(s)}||^2 \quad (4)$$

where λ_1 and λ_2 are the penalty parameters respectively.

In the association studies, cis-regularization plays an important role in uncovering polymorphic regions including transcription factor binding sites (TFBS), enhancers and promoters which have been reported to directly control gene expressions [9], [10], [11]. In order to take advantage of the knowledge, we also introduced a variable for a prior weight when computing the coefficient, $\mathbf{v}_k^{(s)}$.

$$\boldsymbol{\upsilon}_k^{(s)} = diag(\boldsymbol{\omega}_k) \cdot \mathbf{v}_k^{(s)}, \quad (5)$$

where $\boldsymbol{\omega}_k$ is the diagonal matrix for the weighting coefficients of each block (i.e., $\boldsymbol{\omega}_k \in \Re^{p_k\times p_k}$). We assume that the cis-acting SNPs are located within 1kb of the corresponding genes as the majority of research has shown [10].

Rewrite (4) with cis-regularization,

$$\arg\min |\boldsymbol{\xi}_k - \mathbf{X}_k\boldsymbol{\upsilon}_k^{(s)}|^2 + \lambda_1 \sum |\boldsymbol{\upsilon}_k^{(s)}| + \lambda_2 \sum ||\boldsymbol{\upsilon}_k^{(s)}||^2. \quad (6)$$

Then, it can be re-written introducing a scaling factor $(1+\lambda_2)$ to prevent double shrinkage [8], [12],

$$\arg\min \boldsymbol{\upsilon}_k^{(s)\top}\Big(\frac{\mathbf{X}_k^\top\mathbf{X}_k + \lambda_2 I}{1+\lambda_2}\Big)\boldsymbol{\upsilon}_k^{(s)} - 2\boldsymbol{\xi}_k\mathbf{X}_k\boldsymbol{\upsilon}_k^{(s)} + \lambda_1|\boldsymbol{\upsilon}_k^{(s)}|. \quad (7)$$

For efficient computing, the univariate soft-thresholding (UST) strategy is applied, setting $\lambda_2 \to \infty$ [8], [12].

$$\boldsymbol{\upsilon}_k^{(s)} = \Big(|\boldsymbol{\xi}_k^\top\mathbf{X}_i\boldsymbol{\upsilon}_k^{(s)}| - \frac{\lambda_1}{2}\Big)_+ sign\Big(\boldsymbol{\xi}_k^\top\mathbf{X}_i\boldsymbol{\upsilon}_k^{(s)}\Big) \quad (8)$$

where $(F)_+ = F$ if $F > 0$ and $(F)_+ = 0$ if $F \leq 0$.

After estimating the latent variables, SGCCA estimates coefficients of both the inner relations and the outer relations models as well as their intercepts. The parameters $\mathbf{b}_k$ of the inner relations can be estimated by regressing $\boldsymbol{\xi}_k$ on other latent variables which point to $\boldsymbol{\xi}_k$. For $\boldsymbol{\gamma}_k$ of the outer relations, only the columns of $\mathbf{X}_k$, where the corresponding $\mathbf{v}_k$ is non-zero, are computed by regressing on $\boldsymbol{\xi}_k$ to preserve the sparsity. The algorithm is described in detail in Algorithm 1.

A. Tuning penalty parameters

K-fold cross-validation is mainly used to optimize the penalty parameters [8], [12], [13]. However, it needs very large memory spaces, and it is computationally costly to run due to the large number of variables. For efficiency, we designed an algorithm using a K-mean clustering based on the ideas that insignificant coefficients tend to place near zero value. The method initializes two classes setting both zero and the biggest coefficient values as initial values, and performs the traditional UST method repeatedly until a desired number of significant coefficients (η) leave. Then, it finds the optimal parameter such that it makes a maximum correlation in the multiple regressions among the remaining significant coefficients.

III. EXPERIMENT RESULTS

The goal of this study is to identify the actual significant elements of the sets among large numbers of putative elements in the complex system where multiple heterogenous data cooperate closely and estimate the internal relations of the mechanism. To assess the performance of the proposed method, we conducted the experiments with simulation data.

A. Simulation study

In the simulation study, we generated the simulation data on the underlying designed system in Fig. 2, where only five elements of both $\mathbf{X}$ and $\mathbf{Y}$ are designated as significant, which means non-zero coefficients of the corresponding loading vectors are co-related to $\mathbf{Z}$. A various number of the zero-mean errors are added up into the sets of both $\mathbf{X}$ and $\mathbf{Y}$. We tested how many of the five elements designated as ground

Algorithm 1 SGCCA

$r \leftarrow 1$
For all k, standardize $\mathbf{X}_k$
$\mathbf{v}_k^{(0)} = \sqrt{n}\frac{\mathbf{v}_k^{(0)}}{\|\mathbf{X}_k\mathbf{v}_k^{(0)}\|}$
$\boldsymbol{\xi}_k^{(0)} = \mathbf{X}_k\mathbf{v}_k^{(0)}$
$s \leftarrow 0$
repeat
  For all k,
  Update prior weight $\boldsymbol{\omega}_k$.
  $\mathbf{r}_{k,l}^{(s)} = \begin{cases} corr(\mathbf{X}_k\boldsymbol{\upsilon}_k^{(s)}, \mathbf{X}_l\boldsymbol{\upsilon}_l^{(s+1)}) & \text{if } l < k \\ corr(\mathbf{X}_k\boldsymbol{\upsilon}_k^{(s)}, \mathbf{X}_l\boldsymbol{\upsilon}_l^{(s)}) & \text{if } l > k. \end{cases}$
  $\mathbf{w}_{k,l}^{(s)} = \text{sign}(r_{k,l}^{(s)})$
  $\boldsymbol{\xi}_k^{(s)} = \sum_{l=1}^{k-1} c_{k,l}w_{k,l}^{(s)}\mathbf{X}_l\boldsymbol{\upsilon}_l^{(s+1)} + \sum_{l=k+1}^{N} c_{k,l}w_{k,l}^{(s)}\mathbf{X}_l\boldsymbol{\upsilon}_l^{(s)}$
  $\boldsymbol{\upsilon}_k^{(s)} = \left(|\boldsymbol{\xi}^\top\mathbf{X}_i\boldsymbol{\upsilon}_k^{(s)}| - \frac{\lambda_1}{2}\right)_+ sign\left(\boldsymbol{\xi}^\top\mathbf{X}_i\boldsymbol{\upsilon}_k^{(s)}\right)$
  $\mathbf{v}_k^{(s+1)} = \sqrt{n}\frac{\mathbf{v}_k^{(s+1)}}{\mathbf{X}_k\mathbf{v}_k^{(s+1)}}$
  $\boldsymbol{\xi}_k^{(s+1)} = \mathbf{X}_k\mathbf{v}_k^{(s+1)}$
  $s \leftarrow s + 1$
until $\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} corr(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)$ converge
Regress $\boldsymbol{\xi}_k$ on $\sum_{k\neq l, l=1}^{K} h_{k,l}\gamma_{k,l}\boldsymbol{\xi}_l$
Regress only the columns of $\mathbf{X}_k$, which are non-zero columns of $\mathbf{v}_k$, on $\boldsymbol{\xi}_k$
Estimate the intercepts $\mathbf{b}_k^0$ and $\boldsymbol{\gamma}_k^0$
if $\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} corr(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) \leq \rho$ **then**
  Exit
else
  $X_k \leftarrow X_k - (\mathbf{1}(\mathbf{b}_k^0)^\top + \boldsymbol{\xi}_k(\mathbf{b}_k)^\top)$
  $r \leftarrow r + 1$
  Goto step 2
end if

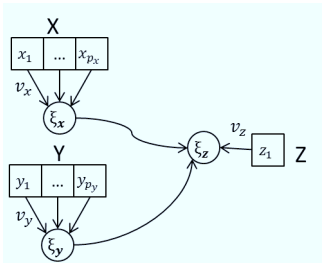
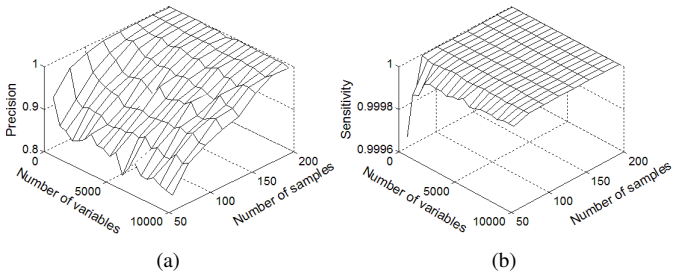


Fig. 2: A simple model with three blocks of variables for the simulation study.

truth among the large number of whole sets are identified by performing the proposed method. The precision and the sensitivity were measured by varying the size of variables and samples, and the results are depicted in Fig. 3. The precision appears to increase as the number of samples increase, and the number of variables decrease. However, the result shows the powerful performance of SGCCA with higher precision than at least 0.8. The sensitivity is interestingly very high no matter what the size is, since the false negatives of the experiments were very small (≈ 0). The experiment results prove the powerful capability of the method to detect



(a) (b)

Fig. 3: The precision and sensitivity assessment with the simulation data for evaluating the performance of SGCCA.

significant elements in the complex systems where multiple models are combined.

B. Human brain data for psychiatric diseases

We applied the proposed method to human brain data comprising of psychiatric diseases such as: schizophrenia, bipolar disorder, and major depression from 131 patients including 44 controls. The human brain data includes 852,963 SNPs and 25,833 gene expression measurements for each individual.

First, we performed GWAS and eQTL studies separately using the traditional method. No significant SNPs were observed after multiple testing corrections in GWAS, while only two of *FGF5* and *DYNC1I2* were identified as significant genes in eQTL. By combining the results of the step-by-step integrative models (not literal integration), it may fail to discover the significant causal traits.

The sets of SNPs and genes identified by the proposed method are listed in Table. I and Table. II, where the top eleven ranked elements are listed among a total of both 93 genes and 691 SNPs. Both of *FGF5* and *DYNC1I2*, which were discovered by the traditional method, are also identified by the proposed method. The results of the experiment revealed that the interaction of 93 genes and 691 SNPs play an important role in the mechanism of psychiatric disorders. *FGF5* gene was identified as an oncogene, as well as involving various biological processes [14]. *MDGA1* gene, which was not identified in the traditional method, was reported to confer risk to schizophrenia and bipolar disorder [15].

SGCCA takes into account grouping effects of SNPs. For instance, the SNPs of *RS11192242*, *RS4918142*, *RS2140837*, and *RS2177744* are located nearby each other and show linkage disequilibrium.

All parameters of the model were estimated, and the putative biological model for psychiatric disorders is illustrated in Fig. 4, where the numbers above arrows show the correlation between the blocks. The model appeared to have strong correlations of -0.78, -0.52, and 0.72, between the blocks.

IV. CONCLUSIONS

In this paper we proposed a novel method of the sparse generalized canonical correlation analysis. SGCCA has the power to detect significant elements of the multi-block

TABLE I: The top eleven ranked genes among a total of 93 genes identified by the proposed method.

Genes	Chr.	Start	Stop	Coef.	P-value*
FGF5	4	81406766	81431195	-0.925	0.001
STXBP5	6	147566568	147748588	-0.772	0.003
MDGA1	6	37708262	37773744	-0.765	0.012
TOB2	22	40159438	40172973	-0.572	0.032
SMARCD2	17	59263176	59274083	-0.550	0.050
TPCN1	12	112143652	112220770	-0.478	0.068
HIF3A	19	51492145	51538530	-0.464	0.096
C14orf159	14	90650164	90761456	-0.455	0.064
DIP2C	10	310130	725608	-0.452	0.009
DHDDS	1	26631360	26670384	0.439	0.080
DYNC1I2	2	172252226	172313167	0.432	0.001
⋮	⋮	⋮	⋮	⋮	⋮

Chr.: chromosome; Coef.: coefficient; P-value*: p-value $\times 10^{-3}$

TABLE II: The top eleven ranked SNPs among a total of 691 SNPs identified by the proposed method.

RS ID	Chr.	Pos.	Coef.	P-value*
RS11192242	10	106694535	0.263	8.073
RS4918142	10	106695763	0.263	8.073
RS2140837	10	106706765	0.263	8.073
RS16966294	17	36205020	-0.253	0.049
RS2177744	10	106703623	0.238	8.095
RS6835683	4	110633716	0.219	0.043
RS28681408	4	110641689	0.219	0.043
RS12648965	4	110631312	0.217	0.061
RS434157	5	112219541	-0.216	0.649
RS11680510	2	101577543	0.206	1.511
RS5909746	23	116086767	-0.196	0.328
⋮	⋮	⋮	⋮	⋮

Pos.: position; P-value*: p-value $\times 10^{-3}$

variables, as many studies have sought the sparse solutions out for methodologies such as PCA, PLS, and CCA on machine learning in order to provide easily interpretable solutions for the models. Not only introducing the solution of SGCCA, we also emphasize the potential power of the method to comprehensively analyze the complex biological systems which cooperate associatively. The proposed method can model complex biological systems integrating existing models and data. Combining the increasing complementary biological data and knowledge such as gene ontology, DNA methylation, mRNA, and microRNA expressions, can provide more accurate insights into the discovery of the complicated biological systems.

The performance of SGCCA was assessed with simulation experiments. As a practical application we applied the method to human brain data of psychiatric disorders. The integrated model of GWAS and cis-eQTL, including the multiple biological data such as genetic markers, gene expressions, and disease phenotypes, is proposed.

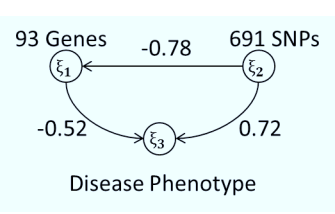


Fig. 4: The model had strong relationships between SNPs, genes, and phenotypes. The numbers above arrows show the correlation between the blocks.

REFERENCES

[1] A. C. Nica, S. B. Montgomery, A. S. Dimas, B. E. Stranger, C. Beazley, I. Barroso, and E. T. Dermitzakis, "Candidate causal regulatory effects by integration of expression qtls with complex trait genetic associations," *PLoS Genet*, vol. 6, p. e1000895, 04 2010.

[2] Y.-H. Hsu, M. C. Zillikens, S. G. Wilson, C. R. Farber, S. Demissie, N. Soranzo, E. N. Bianchi, E. Grundberg, L. Liang, J. B. Richards, K. Estrada, Y. Zhou, A. van Nas, M. F. Moffatt, G. Zhai, A. Hofman, J. B. van Meurs, H. A. P. Pols, R. I. Price, O. Nilsson, T. Pastinen, L. A. Cupples, A. J. Lusis, E. E. Schadt, S. Ferrari, A. G. Uitterlinden, F. Rivadeneira, T. D. Spector, D. Karasik, and D. P. Kiel, "An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility loci for osteoporosis-related traits," *PLoS Genet*, vol. 6, p. e1000977, 06 2010.

[3] Q. Xiong, N. Ancona, E. R. Hauser, S. Mukherjee, and T. S. Furey, "Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets," *Genome Research*, vol. 22, no. 2, pp. 386–397, 2012.

[4] L. Conde, P. Bracci, R. Richardson, S. Montgomery, and C. Skibola, "Integrating gwas and expression data for functional characterization of disease-associated snps: An application to follicular lymphoma," *American journal of human genetics*, vol. 92, pp. 126–130, 2013.

[5] H. Wold, "Partial least squares," *Encyclopedia of the Statistical Sciences*, pp. 581–591, 1985.

[6] J. A. Wegelin, "A survey of partial least squares (pls) methods, with emphasis on the two-block case," tech. rep., 2000.

[7] M. Hanafi, "Pls path modelling: computation of latent variables with the estimation mode b," *Computational Statistics*, vol. 22, pp. 275–292, 2007.

[8] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 301–320, 2005.

[9] V. G. Cheung, R. R. Nayak, I. X. Wang, S. Elwyn, S. M. Cousins, M. Morley, and R. S. Spielman, "Polymorphic *Cis*- and *Trans*-regulation of human gene expression," *PLoS Biol*, vol. 8, p. e1000480, 09 2010.

[10] P. J. Wittkopp and G. Kalay, "Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence," *Nature Reviews Genetics*, vol. 13, pp. 59–69, 2012.

[11] L. W. M. Loo, I. Cheng, M. Tiirikainen, A. Lum-Jones, A. Seifried, L. M. Dunklee, J. M. Church, R. Gryfe, D. J. Weisenberger, R. W. Haile, S. Gallinger, D. J. Duggan, S. N. Thibodeau, G. Casey, and L. Le Marchand, "*cis*-expression qtl analysis of established colorectal cancer risk variants in colon tumors and adjacent normal tissue," *PLoS ONE*, vol. 7, p. e30477, 02 2012.

[12] S. Waaijenborg, P. C. Verselewel de Witt Hamer, and A. H. Zwinderman, "Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 7, 2008.

[13] A. Lykou and J. Whittaker, "Sparse cca using a lasso with positivity constraints," *Comput. Stat. Data Anal.*, vol. 54, pp. 3144–3157, Dec. 2010.

[14] S. Allerstorfer, G. Sonvilla, , H. Fischer, S. Spiegl-Kreinecker, C. Gauglhofer, U. Setinek, T. Czech, C. Marosi, J. Buchroithner, J. Pichler, R. Silye, T. Mohr, K. Holzmann, B. Grasl-Kraupp, B. Marian, M. Grusch, J. Fischer, M. Micksche, and W. Berger, "Fgf5 as an oncogenic factor in human glioblastoma multiforme: autocrine and paracrine activities," *Oncogene*, vol. 27, pp. 4180 – 4190, 2008.

[15] J. Li, J. Liu, G. Feng, T. Li, Q. Zhao, Y. Li, Z. Hu, L. Zheng, Z. Zeng, L. He, T. Wang, and Y. Shi, "The mdga1 gene confers risk to schizophrenia and bipolar disorder," *Schizophrenia Research*, vol. 125, no. 2-3, pp. 194–200, 2011.