Identification of Schizophrenia-Associated Gene Polymorphisms Using Hybrid Filtering Feature Selection with Structural Information

Yingying Wang[1], Zichun Zeng[1,2], and Yunpeng Cai[1(✉)]

[1] Shenzhen Institutes of Advance Technology, Chinese Academy of Sciences, Shenzhen, People's Republic of China
yp.cai@siat.ac.cn
[2] University of Science and Technology of China, Hefei, People's Republic of China

Abstract. Schizophrenia is a complex and severe neurological disorder that affects lots of people worldwide. Despite its strong evidence of heritability revealed by lots of genetic studies, research for locating of schizophrenia associated genes remains frustrating as numerous efforts had failed to identify biomarkers that could strongly impact the diagnosis and prognosis of schizophrenia. The major challenge lies in the weak discrimination of single gene marker and the enormous number of gene variants that exist in human genome. In this paper we propose a hybrid feature selection method that utilizes the biological structural information of the gene variants to tackle this problem. A set of statistical techniques are developed to encourage the clustering of multiple informative SNP variants on the same gene, which boost the probability of finding biologically meaningful features and suppresses false discoveries. As a result, the proposed method achieves significantly better performance on a published schizophrenia human genome data set compared with previous studies, with an area-under-ROC-curve of 65% and an odd ratio of 2.82 (95%CI: 1.80 – 4.40). 36 gene markers are discovered to be associated with the onset of schizophrenia with many of which verified directly or indirectly by previous literature. The method proposed in this paper can be also adopted for efficient control of false discoveries in finding biomarkers from genomic data.

Keywords: Schizophrenia · Biomarkers · SNP · Feature selection

1 Introduction

Schizophrenia is a heterogeneous and multi-factored disease that affects approximately 0.5-1.2% of individual worldwide [1]. Schizophrenia is very complex partly due to the complicating of brain and the enormous neuronal interconnections and permutations thereof in humans. It is thought to be caused by both genetic and environmental factors and the interactions between them [2]. Though genetic factors are considered to be the main issues since schizophrenia has a heritability of about 80%, the research based on genetics has been frustrating because numerous efforts had failed to identify biomarkers that could strongly impact the diagnosis and prognosis of schizophrenia. However, with the development of high-throughput genotyping technologies, many biomarkers especially single nucleotide polymorphisms (SNPs, also termed as

X. Yin et al. (Eds.): HIS 2015, LNCS 9085, pp. 174–184, 2015.
DOI: 10.1007/978-3-319-19156-0_18

common genetic variants) have been identified to be associated with schizophrenia as some studies shown [3-6].

Considering the major goal of schizophrenia genetic research is to choose a list of genetic loci with significant biomarkers, machine learning methods become good choices since they have been applied in many biological-related researches successfully such as microarray analyses, etc. In the field of schizophrenia, few studies had adopted machine learning methods from different aspects. One study identified 36 SNPs related to schizophrenia using logit linear models to represent the relationship between genotype and risk of schizophrenia. Results indicated that a Bayesian approach could identify genes possibly involved in the etiology of schizophrenia [7]. To identify relationships between brain structure volumes and cognitive performance, and the differences of these relationships between control and schizophrenia patients, a study used a Bayesian decision-theoretic method to find morphological biomarker features that best explained neuropsychological test scores in the context of a multivariate response linear model with interactions [8]. A study paid attention to the brain cortical thickness in order to investigate possible subtypes of schizophrenia patients using Lloyd's k-means cluster analysis and found no subtypes specific to patients [9]. Another study used a hybrid machine learning method for fusing fMRI and SNP data to classify schizophrenia patients and healthy controls [10]. These studies showed that machine learning methods can identify biomarkers (such as SNPs) with biological significance for the deep researching of schizophrenia.

Nevertheless, despite these attempts, the genetic origin of schizophrenia remains almost unrevealed. Specifically, most of the above works merely discovers a set of weak associated biomarkers (most of which have statistical significance $p > 0.001$), without giving a complete prediction model. Most of them are not cross-validated or have only marginal separation on the validation data. Although in [10] the authors claimed to achieve 87% accuracy in leave-one-out cross validation of a small data set, they actually used the information of the validation sample (with label information) during the feature selection phase, hence the result is not a real cross-validation accuracy, but should be regarded as a training accuracy which is not externally validated and the replicability is questionable. Hence, finding a discriminative model for separating schizophrenia vs. normal genotypes will still be a milestone in the research of schizophrenia genomics.

The major challenge in finding replicable schizophrenia gene markers is that this disease is likely caused by a collection of genetic factors but no gene is discovered to be strongly informative. Due to the large number of human gene variation factors (for example, there are approximately 10 million SNPs on the human genome) and the restricted number of samples under study, a large number of false-positive biomarkers will be selected which seems to be informative on the training data but will fail in the validation data. Most existing feature selection methods are not powerful enough to control the false-positive rate in such a high dimension and weak indicator case.

In this paper we propose a feature selection strategy to tackle this problem by taking into account the distribution information of the selected factors on the genomic structure. Specifically, the clustering of multiple informative features on the same gene or chromosome will provide an additional indication that these features are not likely random. For example, some previous works have found a set of schizophrenia-associated SNPs clustered on the same gene [6] or a set of genes on the same chromosome [11]. By exploiting this character and favor clustered features, we are

able to select features that are more reliable and control the false positive rate. A crucial issue is to quantitatively describe the degree of enrichment for features on a gene or a chromosome. We employ the idea of multi-dimensional chi-square test and design a hybrid pipeline to pick out a set of gene variants that are enriched on a few genes. As a result, we derive a prediction model that achieves a C-statistics accuracy of 0.65 on a ten-fold cross-validation test, which is significantly higher than previous results. A set of schizophrenia-associated genes, and SNPs on them, are identified, which are proved to be with rational biology interpretation.

2 Material and Methods

2.1 Dataset

We downloaded SNP array data GSE27923 [11] from NCBI GEO [12, 13]. The dataset contained 120 schizophrenia patient-parents trio samples. In all the 360 persons detected, 128 were schizophrenia patients (including 120 schizophrenia patients, 6 of these patients' father and 2 of these patients' mother were also schizophrenia patients) and 232 were healthy controls. Four SNP array platforms were used for each person: Affymetrix Human Mapping 50K Hind240 SNP Array, Affymetrix Human Mapping 50K Xba240 SNP Array, Affymetrix Human CentHindAv2 SNP Array and Affymetrix Human CentXbaAv2 SNP Array. The SNP probes could be mapped to 115,117 NCBI dbSNP [14] entries altogether, which scatter on all 22 human chromosomes and each SNP entry contains three sub-genotypes.

2.2 Overall Framework

Fig.1 demonstrates the framework of the overall procedure in this paper. We split the original data in to ten folds randomly. In each iteration, 9 folds of the data are used for

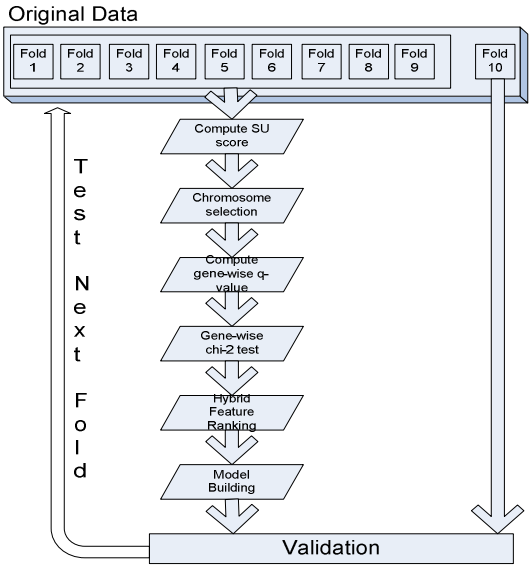

Fig. 1. Framework of the feature selection and validation procedure

model building and one fold for model validation. The procedure is iterated until all folds are used for validation. The compound result of all validations is used to evaluate the performance of the model. In order to deal with the high dimensional nature of the data and the small sample size, we propose a hybrid feature selection scheme combining multiple steps, which will be introduced in the next section.

2.3 Feature Selection

2.3.1 Symmetric Uncertainty

Because the gene SNP features are expressed in categorical variables, traditional filtering methods for continuous variables such as Student's t-test are not able to apply to the data directly. Although one can convert categorical variables into continuous one using encoding techniques, due to the small number of samples and imbalance distribution of attribute values, traditional filtering methods usually have poor performance after conversion. On the other hand, chi-square test is often used for detecting the statistical significance between groups of categorical attributes but is known to be too sensitive to the variance of the data, hence is not suitable for feature selection. In this paper we use the symmetric uncertainty (SU) [15] as the metric for filtering feature selection, which is expressed by the following equation:

$$\mathrm{SU}(\mathrm{X},\mathrm{Y}) = \frac{2I(X,Y)}{H(X)+H(Y)} \tag{1}$$

Where X is the values of samples on the studied variable and Y is the class labels of sample, I is the mutual information between X and Y, $H(X)$ and $H(Y)$ are the entropy of X and Y, respectively. A higher SU score indicates a higher distinction between different classes of samples on feature X.

Symmetric uncertainty has been previously employed in some feature selection methods (e.g., FCBF [16]) but in our experiment we found that these methods perform poorly on the schizophrenia data, because by using the symmetric uncertainty as the only metric for selection, many false positive features are included in the model. In this paper, we propose multiple statistical technologies to tackle this problem and avert the short-coming of previous approaches.

2.3.2 Chromosome Ranking and Selection

The high dimensional nature of the gene SNP array and the limited number of samples poses a severe challenge to feature selection. In order to pick out reliable features, a trade-off should be considered between the completeness of the feature set and the control of false discoveries. In this paper we apply a conservative strategy that we only consider genes on a chromosome with a significant number of informative features. Although some informative features will be lost by using this strategy, it is discovered that the false discovery rate is efficiently lowered. The detailed procedure of chromosome selection is described below:

1. Compute the SU score of each SNP variable;
2. Sort all variables by their SU score in descending order;
3. For a given variable i, suppose r_i is the rank order of the variable, the rank score for the variable is given by $rs_i = \max(100\text{-}r_i,0)$;
4. The rank score for a chromosome is calculated as:

$$S_{chr} = \frac{\sum_{v_i \in \Xi} rs_{v_i}}{|\Xi|} \tag{2}$$

where Ξ is the set of all SNP variables located on the chromosome *chr* and $|\Xi|$ is the number of SNP varialbles on that chromosome

5. Chromosomes with rank score >0.1 are selected and the SNP variables located on the chromosomes are all included for next feature selection steps.

2.3.3 Gene-Wise Benjamini–Hochberg Correction

The Benjamini–Hochberg procedure [17] is a technique for false discovery rate control which take into account the total number of variables under consideration. A larger number of variables considered will increase the risk of false discovery and thus more stringent criteria should be applied to control the risk. On the other hand, a larger number of informative features will imply more chance of finding true positive. The BH procedure applied a balancing strategy by re-calculating the measurement of significance as:

$$q_k = \min\left(\frac{m}{k} p_k, q_{k+1}\right), \quad k = m-1 \ldots 1, \quad q_m = p_m \tag{3}$$

where $\{p_k\}$ is the ordered set of the p-values for all variables derived from a normal statistical test (such as Student's t-test or chi-square test), satisfying $p_1 \le p_2 \le \cdots \le p_m$, and m is the total number of variables in a given data set.

One thing to address is that unlike Student's t-test or chi-square test, there is no traditional measurement of significance for symmetric uncertainty. Nevertheless, we empirically observe that the distribution of SU score approximates the chi-square distribution. Hence we used the following method to calculate the significance (p-value) of SU score for each variable:

$$p_{SU}(X_k) = 1 - F\left(\frac{SU(X_k,Y)}{mean_{i \in \Omega}(SU(X_i,Y))}, 1\right) \tag{4}$$

where $F(x,1)$ is the cumulative distribution function for the chi-square distribution with degree of freedom 1, Ω is the entire set of all variables and X_k is the variable under investigation.

It should be noted that the calculation of p-value is uniform for all variables, while the calculation of the q-value is applied on SNPs located on each gene separately. By this means, genes with a large number of SNP mutations are filtered with a more strict criteria and the overall chance of false discovery is suppressed, leading the selected features to be more stable.

2.3.4 Gene-Wise Chi-square Test

The BH procedure is powerful in reducing the number of false features. However, it does not provide a mechanism to boost informative genes. In biology, the phenomenon that a significant number of disease-associated SNPs clustering on the same gene provide a strong indicator that the gene should play an important role on the development of the disease. To quantitatively measure this phenomenon and enough clustered feature, we apply a multi-dimensional chi-square model:

$$p(G) = \min_{j=1..|G|}(1 - F(\frac{\sum_{k=1..j} SU(X_k,Y)}{mean_{i\in\Omega}(SU(X_i,Y))}, j)) \quad (5)$$

where $G = \{X_1, X_2, \ldots, X_{|G|}\}$ is a set of SNP variables on the same gene which are sorted in increasing values of their single-SNP significance $p_{SU}(X_k)$. Eq. 5 seeks for an optimal number of variables which maximizes the statistical significance of the model, and use it as the significance of the gene. In this manner the clustering effect of informative variables on the same gene is encouraged.

2.3.5 Hybrid Feature Selection Scheme

With the above procedure we got three kinds of evaluation scores for each SNP variable: SU score, the q-value of the SNP with the gene, and the chi-square p-value of the gene that the SNP belongs to (SNPs belonging to the same gene are all assigned the same p-value). We then rank each score individually for all variables selected in section 2.3.3, with SU score in descending order and p-value or q-value in ascending order. The largest rank of the three ranks is used as the rank of the variable. After that, the 30 top-ranked variables in each fold of training dataset are selected for model building.

2.4 Model Building

The Naïve Bayes Classifier

$$\mathrm{P}(\mathrm{Y} = \mathrm{y}_0|\mathrm{X}) = \mathrm{P}(\mathrm{Y} = \mathrm{y}_0) \prod_{i=1}^{m} P(x_i\,|Y = \mathrm{y}_0) \quad (6)$$

is adopted as the prediction model for clinical outcomes of gene variables. Here y_0=0 indicates a healthy outcome and y_0=1 indicates a disease outcome. $X = \{x_1, \ldots, x_m\}$ is a test sample with *m* feature variables selected on the above steps. The prior probability *P(Y)* and conditional probabilities $P(x_i|Y)$ are computed from the training data set with Laplace smoothing applied [18].

In each round of cross-validation test, each validation test sample is assigned a predicted probability. After the entire ten-fold cross-validation is finished, the prediction results are merged together using the predicted probability as the unified outcome. The probability P=0.5 is then used as a cut-off to separate the samples into high risk (P>=0.5) and low risk (P<0.5) groups. The odds ratio between the two groups is

computed, and a receiver operating characteristic (ROC) curve [19] is plotted based on the data to evaluate the performance of the derived models.

3 Results

Fig.2 depicts the odds ratio obtained by the models created in this paper and its comparison to previous results. With the probability cut-off P=0.5, our model achieves an odds ratio of 2.82 (95% CI: 1.80 – 4.40) which is significantly better than previous reported results on cross-validation data. The odds ratio results of Purcell [3] and Shi [6] are taken directly from the original report of their papers, with ratios converted to be always larger than 1. The result of Yamada [11] is computed using their published data (which is also the material used in this paper) using the best result of the selected biomarkers reported in their paper. Among twenty SNP variables reported in [11], only two variables have the 95% CI of the odds ratios completely larger than 1 (1.98 [CI: 1.22-3.21] for SNP rs10496761, and 1.89 [CI: 1.18-3.04] for SNP rs1048076, respectively). Hence, although the calculated odds ratio of [11] is relatively high, it is not obtained on a validation set and hence the comparison is biased to their results. Even so, they are still surpassed by our results. Moreover, these two SNPs only identify 24% and 35% of the patients, respectively. When combining them together, the number of false positive patients grows a lot and the odds ratio drop to near 1. Thus the gene markers reported in previous study [11] is not capable of making accurate prediction.

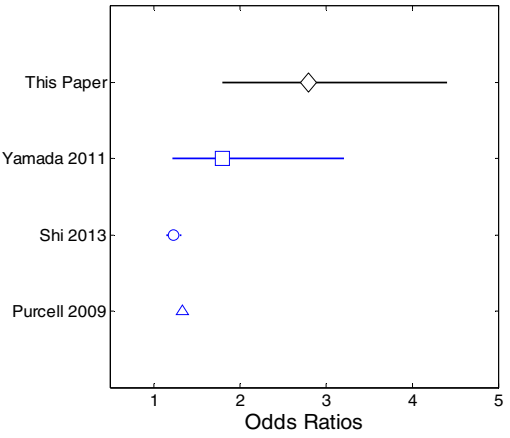

Fig. 2. Comparison of the odd ratios obtained by the result of this paper and previous studies in Purcell [3], Shi [6], Yamada [11], respectively. The markers show the positions of the mean odds ratio and the lines show the positions of the 95% confidence intervals (CI). The CI of the Purcell marker was not reported.

Fig.3 depicts the obtained ROC curve following the ten-fold cross validation procedure, the area under ROC curve (AUC) is 0.65. As a comparison, using traditional filter selection methods such as t-test or chi-square test achieved a poor AUC of near 0.5. By carrying out a permutation test using 1000 permuted datasets, we further confirmed that the result obtained on real dataset is superior to those on permuted data ($p<0.001$) and thus the discovered model is not likely a false discovery. Although the discrimination is not very strong for the patient and healthy group, the result already outperforms previous ones where only marginal separations were obtained.

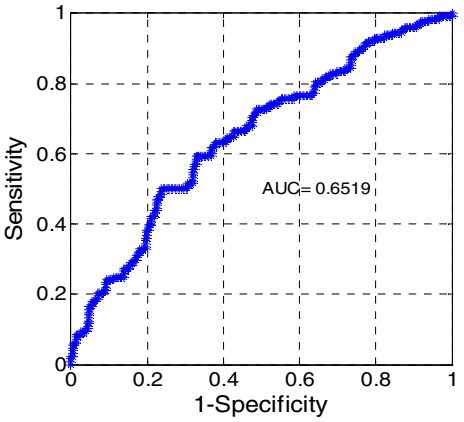

Fig. 3. The receiver operating characteristic (ROC) curve obtained from ten-fold cross validation results on the studied dataset using the proposed method

In ten-fold cross-validation, altogether 74 SNP variables are selected for model building in different folds. Genes with at least 10 hits in total during the ten rounds are chosen for biological analysis. Table 1 listed the 36 SNP variables selected and their biological literature previously reported. We see that most SNP variables are located on the X chromosome and autosome No.1, despite that usually 6-8 chromosomes will be considered after the chromosome selection step 2.3.3, and many SNPs are clustered on the same gene, which exactly as the algorithm expected. Through literature search, we found that 6 out of the 16 total genes have been known to be biomarkers for schizophrenia, and an additional 3 genes are linked with the gene regulation or metabolic functions regarding brain functions, which all persuades that our method is powerful of discovering biological relevant gene variants. Moreover, our study also discovered some new informative genes that help to predict the onset of the disease. The biological meaning of these genes will be further validated through further external validations with more data.

Table 1. List of selected gene biomarkers selected by the methods proposed in this paper. The symbol * denotes that the gene is known to be directly connected with schoziphrenia in literature, and + denotes that it is involved in some processes related to the disease or brain function disorders.

Chromosome	SNP ID	Gene Name	Gene Function	Literature
X	rs996106	PPP1R2P9	protein phosphatase 1, regulatory (inhibitor) subunit 2 pseudogene 9	[20]+
	rs723028			
	rs205869			
	rs205870			
	rs4986541			
	rs2410977			
X	rs2285634		intron variant TRAPPC2/OFD1	[21]*
X	rs5980419	IDS	iduronate 2-sulfatase	
	rs6540313			
	rs7065976			
X	rs431207	DMD	dystrophin	[25]*
	rs725979			
	rs1921386			
	rs1921395			
	rs436628			
X	rs6522686	NAP1L3	nucleosome assembly protein 1-like 3	
1	rs2282729	TNR	tenascin R	[23][24]*
	rs10489316			
	rs10492392			
	rs1385540			
	rs3766680			
	rs4570382			
	rs10489311			
1	rs10493026	RUNX3	runt-related transcription factor 3	
1	rs10489202	MPC2/BRP44	mitochondrial pyruvate carrier 2	[10]*
1	rs149912	DCAF6	DDB1 and CUL4 associated factor 6	[10]*
3	rs1348990		no info	
3	rs879161	PHC3	polyhomeotic homolog 3	
	rs7638400			
3	rs7619166	ACTRT3	actin-related protein T3	
3	rs10510897	CADPS	Ca++-dependent secretion activator	[27]+
17	rs3815341	CCL11	chemokine (C-C motif) ligand 11	[22]*
17	rs10515122	ANKFN1	ankyrin-repeat and fibronectin type III domain containing 1	
	rs7207271			
19	rs3810137	ZNF225	zinc finger protein 225	[26]+
	rs9304639			

4 Conclusion

The enormous number of gene variants that exist in human genome poses a major challenge to genomics studies and the discovery of disease-related gene biomarkers, especially when in the situations where no single strong correlated genes exist. In this

paper we proposed a hybrid feature selection method that utilizes the biological structural information of the gene variants, and adopted a set of statistical techniques to make use of the clustering feature of multiple informative SNP variants on the same gene, thus boost the probability of finding biologically meaningful against false discoveries. Our study showed that the proposed method achieved significantly better performance on the discovery of schizophrenia associated gene markers. In the future, the proposed method will be also applied in other types of genomic data mining for efficient control of false discoveries in biomarkers discoveries.

Acknowlegements. This work was supported by Shenzhen Municipal Science and Technology Research Development and Funds and Platform Construction Plan Key Laboratory Program (CXB201111250113A), Shenzhen Basic Research Fund (JCYJ2013-0329155553732), the Promotion Funds for Key Laboratory in Shenzhen (ZDSY-20120617113021359) and Shenzhen Innovation Funding for Advanced Talents (KQCX20130628112914291).

References

1. Takahashi, S.: Heterogeneity of schizophrenia: Genetic and symptomatic factors. Am. J. Med. Genet. B Neuropsychiatr. Genet. **162B**, 648–652 (2013)
2. Rethelyi, J.M., Benkovits, J., Bitter, I.: Genes and environments in schizophrenia: The different pieces of a manifold puzzle. Neurosci. Biobehav. Rev. **37**, 2424–2437 (2013)
3. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., Sklar, P.: Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. **460**, 748–752 (2009)
4. Shatz, C.J.: MHC class I: an unexpected role in neuronal plasticity. Neuron. **64**, 40–45 (2009)
5. Kwon, E., Wang, W., Tsai, L.H.: Validation of schizophrenia-associated genes CSMD1, C10orf26, CACNA1C and TCF4 as miR-137 targets. Mol. Psychiatry **18**, 11–12 (2013)
6. Shi, Y.Y., Li, Z.Q., Xu, Q., Wang, T., Li, T., et al.: Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. Nature Genetics **43**, 1224–1227 (2011)
7. Hall, H., Lawyer, G., Sillen, A., Jonsson, E.G., Agartz, I., Terenius, L., Arnborg, S.: Potential genetic variants in schizophrenia: a Bayesian analysis. World J. Biol. Psychiatry **8**, 12–22 (2007)
8. Laywer, G., Nyman, H., Agartz, I., Arnborg, S., Jonsson, E.G., Sedvall, G.C., Hall, H.: Morphological correlates to cognitive dysfunction in schizophrenia as studied with Bayesian regression. BMC Psychiatry **6**, 31 (2006)
9. Lawyer, G., Nesvag, R., Varnas, K., Frigessi, A., Agartz, I.: Investigating possible subtypes of schizophrenia patients and controls based on brain cortical thickness. Psychiatry Res. **164**, 254–264 (2008)
10. Yang, H., Liu, J., Sui, J., Pearlson, G., Calhoun, V.D.: A Hybrid Machine Learning Method for Fusing fMRI and Genetic Data: Combining both Improves Classification of Schizophrenia. Front. Hum. Neurosci. **4**, 192 (2010)
11. Yamada, K., Iwayama, Y., Hattori, E., Iwamoto, K., Toyota, T., Ohnishi, T., Ohba, H., Maekawa, M., Kato, T., Yoshikawa, T.: Genome-wide association study of schizophrenia in Japanese population. PLoS One **6**, e20468 (2011)

12. Edgar, R., Domrachev, M., Lash, A.E.: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. **30**, 207–210 (2002)
13. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al.: NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. **41**, D991–995 (2013)
14. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., Sirotkin, K.: dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. **29**, 308–311 (2001)
15. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Amsterdam (2011). ISBN: 978-0-12-374856-0
16. Lei, Y., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proc. Intl. Conf. Mach. Learn. **3**, 856–863 (2003)
17. Benjamini, Y.: HochbergY: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Royal Stat. Soc. B **57**(1), 289–300 (1995)
18. Manning, C.D., Raghavan, P., Schutze, M.: Introduction to Information Retrieval. Cambridge University Press (2008)
19. Fawcett, T.: An Introduction to ROC Analysis. Pattern Recognition Letters **7**(8), 861–874 (2006)
20. Hakak, Y., Walker, J.R., Li, C., Wong, W.H., Davis, K.L., Buxbaum, J.D., et al.: Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. Proc. Natl. Acad Sci. **98**(8), 4746–4751 (2001)
21. Zong, M., Wu, X.G., Chan, C.W.L., Chio, M.Y., Chan, H.S., Tanner, J.A., Yu, S.: The Adaptor Function of TRAPPC2 in Mammalian TRAPPs Explains TRAPPC2-Associated SEDT and TRAPPC9-Associated Congenital Intellectual Disability. PLOS ONE **6**(8), e23350 (2011)
22. Teixeira, A.L., et al.: Increased serum levels of CCL11/eotaxin in schizophrenia. Prog. Neuropsychopharmacol Biol. Psychiatry **32**, 710–714 (2008)
23. Morawski, M., et al.: Tenascin-R promotes assembly of the extracellular matrix of perineuronal nets via clustering of aggrecan. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **369** (2014)
24. Kahler, A.K., et al.: Candidate gene analysis of the human natural killer-1 carbohydrate pathway and perineuronal nets in schizophrenia: B3GAT2 is associated with disease risk and cortical surface area. Biol. Psychiatry **69**, 90–96 (2011)
25. Lindor, N.M., Sobell, J.L., Heston, L.L., Thibodeau, S.N., Sommer, S.S.: Screening the dystrophin gene suggests a high rate of polymorphism in general but no exonic deletions in schizophrenics. American journal of medical genetics **54**(1), 1–4 (1994)
26. Bowden, N.A., Weidenhofer, J., Scott, R.J., Schall, U., Todd, J., Michie, P.T., Tooney, P.A.: Preliminary investigation of gene expression profiles in peripheral blood lymphocytes in schizophrenia. Schizophrenia research **82**(2), 175–183 (2006)
27. Hattori, K., Tanaka, H., Wakabayashi, C., Yamamoto, N., Uchiyama, H., Teraishi, T., et al.: Expression of Ca2+-dependent activator protein for secretion 2 is increased in the brains of schizophrenic patients. Prog. in Neuro-Psycho. & Biol. Psyc. **35**(7), 1738–1743 (2011)