Design of ASD subtyping approach based on multi-omics data to promote personalized healthcare

Tao Chen
Wuhan University
chentao1979@126.com

Peixin Lu
Wuhan University
2018lupx@gmail.com

Li Chen
Chongqing Medical University
chenli2012@126.com

Jie Chen
Chongqing Medical University
jchen010@foxmail.com

Tingyu Li
Chongqing Medical University
tyli@vip.sina.com

Tanya Froehlich
Cincinnati Children's Hospital Medical Center
Tanya.Froehlich@cchmc.org

Long Lu
Wuhan University
bioinfo@gmail.com

Abstract

Autism spectrum disorder (ASD) is a heterogeneous neurodevelopmental disorder that has been confirmed to be related to some genetics risk factors which can lead to different clinical phenotypes. At present, ASD is mainly diagnosed based on some behavior and cognitive scales, which can not reveal the mechanism of disease occurrence, development and prognosis. In recent years, some studies have applied omics techniques into ASD research, but these studies are only based on single omics data source such as genomics, proteomics or transcriptomic without investigating ASD subtypes from integration of multi-omics data. In this study, we proposed an ASD subtyping framework that integrates clinical and multi-omics data to identify and analyze ASD subtypes at the molecular level. Due to the heterogeneity of different data modalities, a fusion clustering strategy was used to produce more accurate and interpretable clusters. Based on ASD subtyping results, we also proposed a classification framework to predict the subtype of new ASD patients. Deep learning method was used to extract features from each data modality, then all extracted features were integrated by the multiple kernel learning method to improve the classification accuracy.

1. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by impaired communication and social interaction as well as restricted and repetitive interests and behaviors [1]. ASD is now diagnosed mainly based on Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) criteria and various behavior scales such as Autism Diagnostic Observation Schedule (ADOS) , Autism Diagnostic Interview, Revised (ADI-R) [2]. However, this behavior-based diagnosis method does not define any subtypes that can reflect phenotypic heterogeneity of ASD and facilitate better understanding of etiology of the condition. In recent years, some ASD subtyping studies based on clinical manifestation or behavioral symptoms have emerged [3, 4], but the symptoms on ASD patients change significantly with the increase of age, resulting in subjective and unreliable subtyping results that are also difficult to elucidate the etiology of each ASD phenotype. With the wide application of artificial intelligence on medical image analysis, a large body of research has emerged using multiple Magnetic Resonance Imaging (MRI) techniques finding impaired functional and structural connectivity between brain regions. And these abnormalities in connectivity sometimes can be used as biomarkers for the diagnosis of ASD in clinic.

However, the biomarkers obtained in these imaging studies are not always consistent, making it impossible to achieve clinical reliability by relying only on medical imaging [5, 6]. The reason for this inconsistence is that each ASD subtype has its distinct characterizations in brain, while most imaging studies do not take into account the heterogeneity between different ASD subtypes. So if considering patient subtypes in imaging analysis, it can improve the reliability of ASD clinical diagnosis and treatment. At present, there are studies explaining the imaging heterogeneity of ASD patients from molecular perspective. For example, Qureshi et al found that 16p11.2 chromosome deletion leads to an increase in brain volume, while its duplication leads to a decrease in brain volume [7]. Some studies also revealed how ASD risk genes (such as CNTNAP2, MET) affect brain structure and function networks [8, 9]. These

URI: https://hdl.handle.net/10125/64150
978-0-9981331-3-3

studies have emphasized the important role of inheritability in ASD diseases. And combining genes and imaging can facilitate the discovery of more accurate ASD imaging biomarkers.

Behavioral assessment and imaging diagnosis both have certain limitations and neither one can reveal the etiology of ASD. With the continuous advancement in basic scientific research especially in the fields of genetics and neuroscience, researchers have determined that ASD is a kind of diffuse developmental disorder of the central nervous system caused by a variety of environmental factors under the influence of certain genetic factors. In 2009, the NIMH conducted a research project called Research Domain Criteria (RDoC) [10], which aims to transcend the limits of traditional classification method, combining genetics, neuroscience, behavioral science and other methods to clarify the neural basis and biomarkers of mental illness. This project implies that we can find more clinically relevant ASD subtypes and biomarkers by integrating multiple data sources such as genetic, neuroimaging, behavioral assessment, and patient clinical history to achieve personalized healthcare and improved prognosis for patients with ASD.

In the past few decades, scientists have been actively exploring the genetic basis of ASD using methods such as cytogenetic research, linkage analysis, and association analysis of candidate genes [11]. The findings show that ASD is a complex disease with high level of genetic heterogeneity. In recent years, great progress has been made in the identification of ASD genetic pathogenicity loci. At present, there are hundreds of pathogenic genes associated with ASD, most of which are rare and have various types of mutations [12]. In addition to the mutations inherited from parents, many ASD de novo mutations have been discovered by using the whole-exome sequencing technique, and the number of identified mutations is up to more than 1000 [13]. Iossifov et al. used whole-exome sequencing find that in a family with only one ASD patient, new gene mutations such as copy number variant (CNV) accounted for approximately 30% of ASD cases [14]. Common genetic variants such as single nucleotide polymorphism (SNP) also influence the onset of ASD. Although many ASD-related pathogenic genes have been found, the high level of inheritance heterogeneity is far from explaining the pathogenesis of ASD. Transcriptomics studies use cDNA microarrays and high-throughput RNA sequencing technology to obtain gene expression data, combining with gene co-expression network analysis to discover the molecular mechanisms related to ASD [15, 16]. Proteomics studies have found that the clinical manifestations of ASD may be related to changes in proteins, such as mutations leading to changes in protein and amino acid sequences, and mutations in gene regulatory regions leading to changes in protein expression or abnormal modification of proteins [17, 18]. In recent years, the acquisition of specific biomarkers through metabolomics research is effective for early ASD screening and diagnosis. Currently, some analytical techniques such as gas chromatography-mass spectrometry (Gc-Ms), capillary electrophoresis-time-of-flight mass spectrometry (CE-TOF), liquid chromatography-mass spectrometry (Lc-Ms), nuclear magnetic resonance (NMR) have been widely used in ASD urine or blood metabolite analysis. By comparison with the healthy controls, the significant differences in metabolites are mainly related to intestinal microbial metabolism, energy metabolism, oxidative stress [19-21]. At present, it has been found that ruminococcus, clostridium and desulfovibrio in the gastrointestinal tract of ASD are significantly different from healthy controls [22, 23]. The damage to normal nerve development caused by gastrointestinal microflora and its metabolites is one of the ASD etiology research directions. Through bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP) technique, DNA is directly extracted from feces of ASD patients for sequencing to determine specific bacterial composition, and then all microorganisms in the intestine can be identified [24].

In summary, most of the above studies are based on single omics data source and did not systematically analyze and explore ASD subtypes. The characterization of each omics technique on ASD subtype level has not yet been studied as well as the interaction between different omics data and their distinct regulatory role on signaling pathways. To solve this problem, we proposed an ASD subtyping framework by fusing multi-omics and clinical assessment data using the unsupervised clustering method. Then for subtype prediction of a new unknown patient, we proposed a classification framework based on the multi-omics data and clustering results. Due to the complexity and high dimensionality of omics data, deep learning method was used to automatically extract discriminative features from all omics data. Then we used the multiple kernel learning (MKL) method to explore complementary information among these features and discover different contribution of relevant omics features.

2. ASD subtyping based on multi-omics data

Studies have shown that ASD is highly heritable, and that each ASD-related gene mutation or CNV will result in a different clinical phenotype. However, due to the etiology complexity of ASD, it is generally possible to find pathogenic genes distributed on different chromosomes in patients. At the same time, due to the difference in gene expression and the interaction of multiple genes in a common path, ASD appears in clinical phenotype with high level of heterogeneity. Therefore, we propose an ASD subtyping framework by integrating clinical and multi-omics data to discriminate among distinct ASD subtypes.

Cluster analysis is a major analytical method for disease subtypes based on high-dimensional omics data. It can classify functionally related genes (or other omics data) according to the consistent trend or proximity of expression levels. It has already been widely used in cancer subtyping [25]. Cluster analysis generally needs to establish an accurate measure of the "distance" between samples. The measurement methods include Euclidean distance, Chebyshev distance, Manhattan distance, Minkowski distance, angle cosine, Pearson correlation coefficient and so on. Since the clustering results are susceptible to a large amount of noise in the omics data, and the omics data is usually very high in dimension, it is necessary to use data dimensionality reduction methods before cluster analysis such as Principal Component Analysis (PCA) , Partial Least Square (PLS), etc. removing data containing useless information. In the field of bioinformatics analysis, common clustering algorithms include Hierarchical Clustering, K-means, Fuzzy C-means, and Self- Organizing Maps and so on.

The above cluster analysis methods are mainly based on the statistical theory that the knowledge in the biological field is rarely used, so that the clustering result does not produce the reasonable biological interpretation that people need to understand the disease. In addition, it is more important to combine and fuse different kind of omics data effectively for clustering. The easiest way is to concatenate feature matrix of each omics data source directly, but this approach will result in a further lower signal-to-noise ratio. Another way is to cluster individually for each omics data source and then merge the clustering results. However, the potential inconsistencies between each clustering results may cause merge errors. Currently, there are two methods for the fusion clustering of different genomics data in the field of tumor research, including the iCluster [26] method based on the joint latent variable model and the SNF [27] method based on the sample similarity network. Studies have shown that SNF has better performance in tumor subtyping than iCluster [27].

In ASD subtyping, we can also use the above two methods to perform fusion cluster analysis on various omics data (such as gene expression, protein expression, etc.) and compare the different performance between the two methods. Since cluster analysis is an unsupervised learning method, its clustering results can not be tested by ground truth. The biostatistical method can be used to perform statistical tests on clustering results, including survival analysis, contour coefficients and clustering statistical significance test. The multi-omics based ASD subtyping framework is shown in Figure 1.

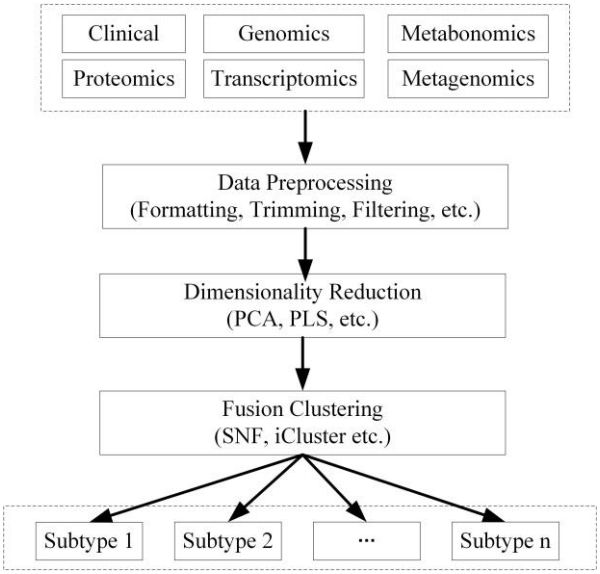

Figure 1. Framework of multi-omics based ASD subtyping

3. Subtype prediction of new ASD patients

After ASD subtyping, it is also necessary to construct a classification model that can distinguish between different ASD subtypes. A wide range of supervised learning methods are available to build a classifier that is used to predict the subtype of a new ASD patient. Since the multi-omics and clinical evaluation data are integrated when producing ASD subtype clusters, it is also beneficial to improve the model performance that integrating different data modalities to obtain complementary information when building an ASD subtype prediction model. Deep learning method is suitable for extracting more discriminative features from high-dimensional and complex data. The extracted features from each data modality then can be integrated to further improve the model performance. The ASD subtype prediction framework is shown in Figure 2.

Convolutional Neural Network (CNN) is a feedforward neural network derived from biological nerves. It is a multi-layer perceptron model designed to recognize two-dimensional images and has certain invariance to image transformation. CNN is composed of three components including the convolutional layer, the pooling layer, and the fully connected layer, and it has some advantages such as network weight sharing and direct input of the original image. Although CNN is mainly used in the field of computer vision, its powerful feature analysis capabilities can also be used for the processing of biological omics data [28]. There are many CNN models such as LeNet-5, Inception v3 etc. that can be used to extract features from multi-omics and clinical data.

Integrating multiple heterogeneous omics data sources such as genomic, proteomic, metabolomic, etc. can lead to better classification than simply using one data source alone. Support Vector Machine (SVM) is a supervised learning method and commonly used in classification problems. The optimal performance of SVMs is highly dependent on the kernel function used. Cross validation is the standard approach to select the best kernel function among a set of candidates such as linear kernel, polynomial kernel and gaussian kernel. However, SVM is not suitable for analyzing multiple data sources using a single kernel function. We can use Multiple Kernel Learning (MKL) method to combine kernels calculated on different input data modality to obtain better predictive performance [29,30].

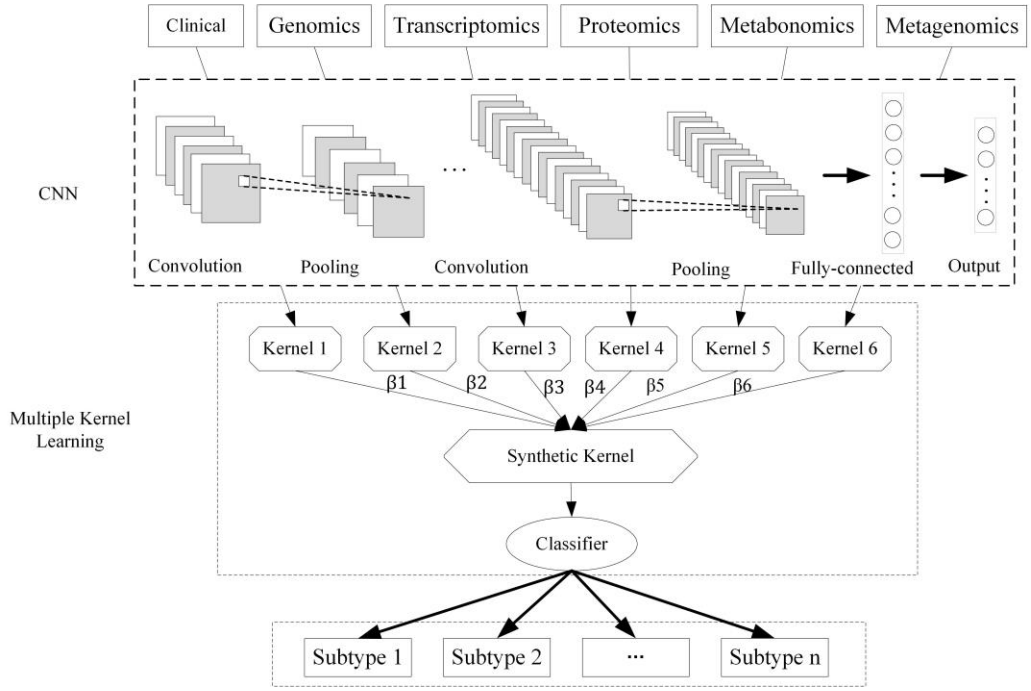

Figure 2. ASD subtype prediction framework

4. Discussion

The molecular subtyping has been successfully applied to cancer research and there are many publicly available repositories such as the Cancer Genome Atlas (TCGA), Multi-Omics Profiling Expression Database (MOPED), Oncomine, ArrayExpress, COSMIC etc. In the field of ASD research, the National Database for Autism Research (NDAR) allows a researcher to associate a single research participant's anonymized genetic, imaging, clinical assessment and other information to analyze the ASD and its subtypes [31]. This paper proposed an ASD subtyping framework based on molecular multi-omics data which can be obtained from the public databases or from the private medical institutions. Unsupervised clustering methods were used to fuse the clinical and multi-omics data to group them into different clusters. Once the ASD subtypes are generated, they can be characterized by combining the clusters with existing pathway or interaction network knowledge to increase the interpretability of the generated ASD subtypes. Another framework this paper proposed is the classification model to accurately predict the subtypes of new ASD patients. In this framework,

CNN network was used to extract discriminative features from various data modalities. In order to differentiate the contributions or weights to the classification of each modality, we intended to leverage the MKL technique to learn the optimal combination coefficients of different kernel functions, which is expected to improve the subtype prediction accuracy. Of note, there are challenges and impediments for ASD subtyping due to data heterogeneity, diversity of standards and analysis tools, variability of experimental procedures. However, compared to ever-changing psychiatric nosological definitions such as DSM-5, ASD subtyping from molecular perspective holds the promise of better personalized healthcare and precision medicine for ASD patients with specific manifestations.

5. References

[1] E. Simonoff, A. Pickles, T. Charman, S. Chandler, T. Loucas, and G. Baird, "Psychiatric disorders in children with autism spectrum disorders: prevalence, comorbidity, and associated factors in a population-derived sample," J Am Acad Child Adolesc Psychiatry, 2008, 47(8), pp. 921-9.

[2] J. A. Reaven, S. L. Hepburn, and R. G. Ross, "Use of the ADOS and ADI-R in children with psychosis: importance of clinical judgment," Clin Child Psychol Psychiatry, 2008, 13(1), pp. 81-94.

[3] S. Georgiades, P. Szatmari, L. Zwaigenbaum, E. Duku, S. Bryson, W. Roberts, and W. Mahoney, "Structure of the autism symptom phenotype: A proposed multidimensional model," J Am Acad Child Adolesc Psychiatry, 2007, 46(2), pp. 188-96.

[4] A. E. Lane, R. L. Young, A. E. Baker, and M. T. Angley, "Sensory processing subtypes in autism: association with adaptive behavior," J Autism Dev Disord, 2010, 40(1), pp. 112-22.

[5] M. A. Just, V. L. Cherkassky, T. A. Keller, and N. J. Minshew, "Cortical activation and synchronization during sentence comprehension in high-functioning autism: evidence of underconnectivity," Brain, 2004, 127(Pt 8), pp. 1811-21.

[6] K. McFadden and N. J. Minshew, "Evidence for dysregulation of axonal growth and guidance in the etiology of ASD," Front Hum Neurosci, 2013, 7, p. 671.

[7] A. Y. Qureshi, S. Mueller, A. Z. Snyder, P. Mukherjee, J. I. Berman, T. P. Roberts, and R. L. Buckner, "Opposing brain differences in 16p11.2 deletion and duplication carriers," J Neurosci, 2014, 34(34), pp. 11199-211.

[8] E. L. Dennis, N. Jahanshad, J. D. Rudie, J. A. Brown, K. Johnson, K. L. McMahon, and P. M. Thompson, "Altered structural brain connectivity in healthy carriers of the autism risk gene, CNTNAP2," Brain Connect, 2011, 1(6), pp. 447-59.

[9] J. D. Rudie, L. M. Hernandez, J. A. Brown, D. Beck-Pancer, N. L. Colich, P. Gorrindo, and M. Dapretto, "Autism-associated promoter variant in MET impacts functional and structural brain networks," Neuron, 2012, 75(5), pp. 904-15.

[10] T. Insel, B. Cuthbert, M. Garvey, R. Heinssen, D. S. Pine, K. Quinn, and P. Wang, "Research domain criteria (RDoC): toward a new classification framework for research on mental disorders," Am J Psychiatry, 2010, 167(7), pp. 748-51.

[11] D. Ma, D. Salyakina, J. M. Jaworski, I. Konidari, P. L. Whitehead, A. N. Andersen, and M. A. Pericak-Vance, "A genome-wide association study of autism reveals a common novel risk locus at 5p14.1," Ann Hum Genet, 2009, 73(Pt 3), pp. 263-73.

[12] J. A. S. Vorstman, J. R. Parr, D. Moreno-De-Luca, R. J. L. Anney, J. I. Nurnberger, Jr., and J. F. Hallmayer, "Autism genetics: opportunities and challenges for clinical translation," Nat Rev Genet, 2017, 18(6), pp. 362-376.

[13] J. Chang, S. R. Gilman, A. H. Chiang, S. J. Sanders, and D. Vitkup, "Genotype to phenotype relationships in autism spectrum disorders," Nat Neurosci, 2015, 18(2), pp. 191-8.

[14] I. Iossifov, B. J. O'Roak, S. J. Sanders, M. Ronemus, N. Krumm, D. Levy, and M. Wigler, "The contribution of de novo coding mutations to autism spectrum disorder," Nature, 2014, 515(7526), pp. 216-21.

[15] S. Gupta, S. E. Ellis, F. N. Ashar, A. Moes, J. S. Bader, J. Zhan, and D. E. Arking, "Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism," Nat Commun, 2014, 5, p. 5748.

[16] M. Quesnel-Vallieres, R. J. Weatheritt, S. P. Cordes, and B. J. Blencowe, "Autism spectrum disorder: insights into convergent mechanisms from transcriptomics," Nat Rev Genet, 2019, 20(1), pp. 51-63.

[17] A. G. Ngounou Wetie, K. L. Wormwood, S. Russell, J. P. Ryan, C. C. Darie, and A. G. Woods, "A Pilot Proteomic Analysis of Salivary Biomarkers in Autism Spectrum Disorder," Autism Res, 2015, 8(3), pp. 338-50.

[18] J. Yang, Y. Chen, X. Xiong, X. Zhou, L. Han, L. Ni, and C. Huang, "Peptidome Analysis Reveals Novel Serum Biomarkers for Children with Autism Spectrum Disorder in China," Proteomics Clin Appl, 2018, 12(5), p. e1700164.

[19] X. Ming, T. P. Stein, V. Barnes, N. Rhodes, and L. Guo, "Metabolic perturbance in autism spectrum disorders: a metabolomics study," J Proteome Res, 2012, 11(12), pp. 5856-62.

[20] P. Emond, S. Mavel, N. Aidoud, L. Nadal-Desbarats, F. Montigny, F. Bonnet-Brilhault, and C. R. Andres, "GC-MS-based urine metabolic profiling of autism spectrum disorders," Anal Bioanal Chem, 2013, 405(15), pp. 5291-300.

[21] M. Guo, J. Zhu, T. Yang, X. Lai, Y. Lei, J. Chen, and T. Li, "Vitamin A and vitamin D deficiencies exacerbate symptoms in children with autism spectrum disorders," Nutr Neurosci, 2019, 22(9), pp. 637-647.

[22] S. M. Finegold, "Therapy and epidemiology of autism--clostridial spores as key elements," Med Hypotheses, 2008, 70(3), pp. 508-11.

[23] S. M. Finegold, "Desulfovibrio species are potentially important in regressive autism," Med Hypotheses, 2011, 77(2), pp. 270-4.

[24] Y. Song, C. Liu, and S. M. Finegold, "Real-time PCR quantitation of clostridia in feces of autistic children," Appl Environ Microbiol, 2004, 70(11), pp. 6459-65.

[25] D. Wang and J. Gu, "Integrative clustering methods of multi-omics data for molecule-based cancer classifications," Quantitative Biology, 2016, 4(1), pp. 58-67.

[26] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," Bioinformatics, 2009, 25(22), pp. 2906-12.

[27] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," Nat Methods, 2014, 11(3), pp. 333-7.

[28] H. Yu, D. C. Samuels, Y. Y. Zhao, and Y. Guo, "Architectures and accuracy of artificial neural network for disease classification from omics data," BMC Genomics, 2019, 20(1), p. 167.

[29] A. Rahimi and M. Gonen, "Discriminating early- and late-stage cancers using multiple kernel learning on gene sets," Bioinformatics, 2018, 34(13), pp. i412-i421.

[30] M. Tao, T. Song, W. Du, S. Han, C. Zuo, Y. Li, and Z. Yang, "Classifying Breast Cancer Subtypes Using Multiple Kernel Learning Based on Omics Data," Genes (Basel), 2019, 10(3),

[31] N. Payakachat, J. M. Tilford, and W. J. Ungar, "National Database for Autism Research (NDAR): Big Data Opportunities for Health Services Research and Health Technology Assessment," Pharmacoeconomics, 2016, 34(2), pp. 127-38.