EPiC Series in Computing

Volume 70, 2020, Pages 49–58

Proceedings of the 12th International Conference
on Bioinformatics and Computational Biology

EPiC Computing

PhGC: A Machine Learning Based Workflow for Phenotype-Genotype Co-analysis on Autism

Safa Shubbar[1] *,†; Chen Fu[2] *; Zhi Liu[3], Anthony Wynshaw-Boris[2], and Qiang Guan[1]

[1] Department of Computer Science, Kent State University, Kent, OH, USA
{sshubbar, qguan}@kent.edu
[2] Case Western Reserve University, OH, USA
{chen.fu, ajw168}@case.edu
[3] iLambda, Aurora, OH, USA
{zliu}@ilambda.ai

Abstract

Autism spectrum disorder (ASD) is a heterogeneous disorder, diagnostic tools attempt to identify homogeneous subtypes within ASD. Previous studies found many behavioral/physiological commodities for ASD, but the clear association between commodities and underlying genetic mechanisms remains unknown. In this paper, we want to leverage machine learning to figure out the relationship between genotype and phenotype in ASD. To this purpose, we propose PhGC pipeline to leverage machine learning approach to to identify behavioral phenotypes of ASD based on their corresponding genomics data. We utilize unsupervised clustering algorithms to extract the core members of each clusters and profile the core member subsets to explore the characteristics using genotype data from the same dataset. Our genome annotation results showed that most of the alleles with different frequency among clusters were represented by the core members.

1 Introduction

The genetic basis for Autism Spectrum Disorder (ASD) remains obscure even after great effort were made in the past decades. One important reason is the phenotypic heterogeneity. ASD patients showed highly variable comorbidities, including but not limited to seizures, macrocephaly, intelligence deficiency, etc. For patients with abnormality of same behavioral phenotypes, the degree of severeness also varies [9]. For this reason, GWAS approach, a main genetic tool to explore genetic basis of complex traits, requires huge amount of samples (ASD patients and normal controls, each for more than several hundred, even several thousand) to be genotyped [17, 4]. However, the number of significant GWAS loci called in each of these studies was low. One important reason for the low power was that without classifying ASD patients to more specific subtypes, the number of patients share similar clinical phenotypes (and potentially,

[*] First two authors share equal contributions

[†] University Of Al-Qadisiyah, Al Diwaniyah, Iraq

Q. Ding, O. Eulenstein and H. Al-Mubaid (eds.), BICOB 2020 (EPiC Series in Computing, vol. 70), pp. 49–58

common genetic basis) was small even the total number included in the GWAS was high. Some efforts had been made to classify the ASD samples before input to GWAS [8]. The conclusion for that research was negative, i.e., reduction of phenotype complexity did not increase power to detect significant GWAS loci. However, we think the reason for this outcome could be, at least, partially attributed to the subjective instead of objective way of sample classification. Also, the score cutoff to classify samples, e.g., as high verbal IQ was selected subjectively in that research. We propose in this study, to first classify ASD samples into subgroups in a data-driven way. The genetic basis in each of the subgroups should have higher similarity, which enable us to detect loci with high alternative allele frequency or homozygosity and likely to be related with the pathology of ASD in each of the subgroups. A number of recent studies, including a variety of genetic conditions (e.g Williams syndrome [21], fragile X syndrome [14] and neurofibromatosis [27]) have indicated genetic disorder-specific behavioral profiles existence, that encourage further efforts in this field. Genetic variants with main effect are present in probably 10%-20% of autism spectrum disorder (ASD) cases [3], more than 100 high confidence genes have been discovered to be associated with ASD [7].

Many studies have identified varied behavioral phenotypes within ASD. A survey conducted by Beglinger and Smith to review research efforts to subtype ASD based on level of functioning and social abilities [6]. Obafemi-Ajayi and colleagues [24] applied varied comprehensive cluster analysis techniques to facial surface measurements. They assert that facial morphology constitutes viable biomarkers between groups of 62 boys with ASD and matched controls and that subgroups with distinctive facial morphology could be identified. Three ASD subgroups were found, Clinical phenotype of the first subgroup subjects is described by 50 % (9/18) Autistic Disorder, 44 % (8/18) with Asperger Syndrome, and 6 % (1/18) with PDD-NOS. While subgroup 2 subjects demonstrate the most coherent clinical phenotype with 79 % (11/14) described as Autistic Disorder, 14 % (2/ 14) as PDD-NOS, and 7 % (1/14) as Asperger Syndrome. Finally, subgroup 3 (the largest subgroup) appears to represent the broad composition of children diagnosed with ASD, 47 % (14/30) of the boys described as Autistic Disorder, 33 % (10/30) as Asperger Syndrome, and 20 % as PDD-NOS. Most recently, Five ASD subgroups with performance ranging from severe deficits to minimal impairments were identified, a clustering method was utilized based on task performance of reading emotions and mental states by Lombardo and colleagues [22].Therefore, assessments able to classify subgroups of ASD may improve the outcomes of targeted treatments of subtypes of ASD [25].

To this purpose, PhGC was used to leverage machine learning in the form of Kmeans, PAM, and Hierarchical clustering techniques to identify behavioral phenotypes of ASD from a sample of 658 ASD patients (563 of them were also subjected to genotyping) using data from a detailed assessment of skills across developmental domains. Most of the phenotype clinic data are redundant. We first apply data pre-processing and then utilize unsupervised clustering algorithm to the cleaned phenotype data. After clustering, core members of each clusters are selected and their genotyping (from whole exome sequencing) data were used to identify genetic loci with significant difference in allele frequency among clusters. These loci were then annotated to identify genes related to them. ANNOVAR is a command line tool too which means we can use it across any operating system (e.g. Linux, Windows, Mac, etc.). The tool we used for annotation is ANNOVAR tool [30]. It is not only efficient for the gene annotation, but also has the ability to download the database directly from (UCSC browser, 1000 genome project or ANNOVAR website) with a simple command, which could help us to link loci we found with known pathological alleles. This study is the largest and most extensive of its type to date, results indicate that machine learning not only successfully extracted behavioral subtypes and the relationships among subtypes, but our annotation result also yielded that the different

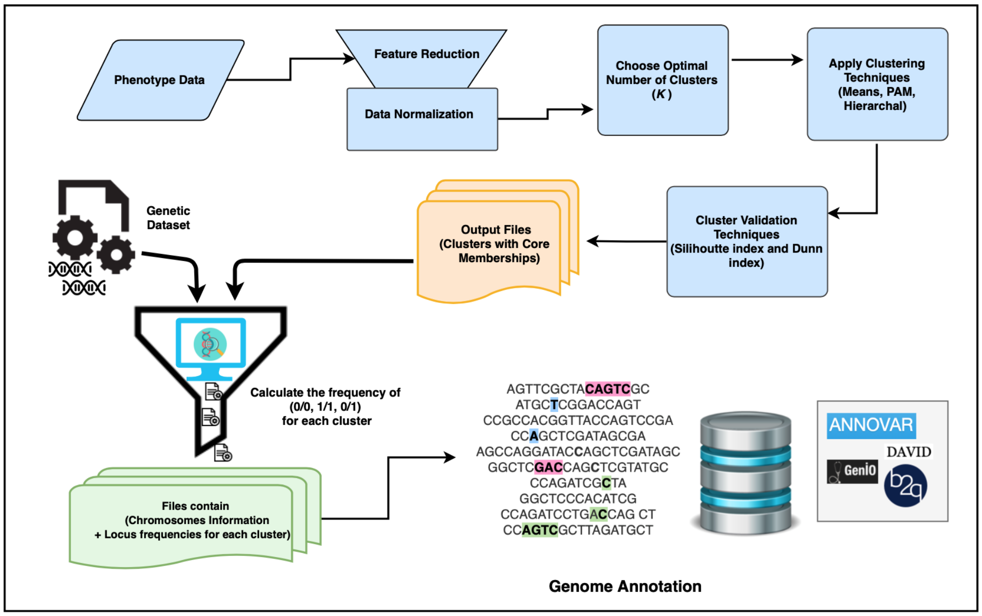

Phenotype Data
Feature Reduction
Data Normalization
Choose Optimal Number of Clusters (*K*)
Apply Clustering Techniques (Means, PAM, Hierarchal)
Cluster Validation Techniques (Silihoutte index and Dunn index)
Output Files (Clusters with Core Memberships)
Genetic Dataset
Calculate the frequency of (0/0, 1/1, 0/1) for each cluster
Files contain (Chromosomes Information + Locus frequencies for each cluster)
AGTTCGCTACAGTCGC
ATGCTCGGACCAGT
CCGCCACGGTTACCAGTCCGA
CCAGCTCGATAGCGA
AGCCAGGATACCAGCTCGATAGC
GGCTCGACCAGCTCGTATGC
CCAGATCGCTA
GGCTCCCACATCG
CCAGATCCTGACCAG CT
CCAGTCGCTTAGATGCT
ANNOVAR
DAVID
GeniO
b2q
Genome Annotation

Figure 1: PhGC Workflow

resulted clusters are also presented by different groups of genes.

2 PhGC Workflow System

The workflow of the proposed PhGC is illustrated in (Fig.1). The methodology is built around two main datasets, each of them consisting of different types and amounts of data:

- Set I, the phenotypic dataset contains subjects and behavioral variables for each subject. No class labels, however, it has the advantage of being relatively large and easier or less expensive to obtain.
- Set II, genetic dataset, contains subjects with chromosomes information and locus for each subject.

2.1 DataSet

Data we used is from NDAR (national database for autism research). The study from which the data collected was "UIC ACE: Translational Studies of Insistence on Sameness in Autism", included phenotype data for 658 ASD patients and 563 of them were also subjected to genotyping using Illumina Human_660 SNP arrays. The 563 patients have both phenotypic data and genotypic data and subjected to following analysis.

2.2 Machine Learning Techniques

The previous section described the structure of the data used be our PhGC approach. This section provides details of all the steps implemented in our methodology. These main steps are presented in (Algorithm.1).

2.2.1 Variable Reduction/Data Transformation

Rstudio version 2018 was used for the data pre-processing. The entire dataset provided 125 variables (attributes) that required reduction in number before performance of a cluster analysis. Variables with missing data were excluded immediately. Variables that were reduced by selection of variables chosen to reflect certain parameters (e.g. total behavioral measurements),

Algorithm 1: PhGC Framework

```
Input: Set I: Phenotypic dataset. No class Labels
       Set II: Genetic dataset
Output: Multiple gene sets explain the sub-categories of Autism
1 apply data pre-processing techniques on Set I.
2 determine the optimal number of clusters k in Set I
3 apply clustering techniques on Set I
4 validate the generated clusters in terms of compactness, separation, and connectivity
5 generates output files of clusters with core members
6 compare resulted files with Set II, for each cluster:
    (a) extract subjects with both phenotypic and genetic data
    (b) calculate the alternative allele frequency ($F_{1/1}$) of each locus
    (c) calculate the frequency of homozygous for alternative loci ($H_{1/1}$) genotype (1/1) for each locus
7 perform genotype annotation on resulted data from step 6
```

also we have benefited from the expertise of medical experts. Duplicated data were excluded (e.g. some subjects had multiple interviews in different age/date).

Data pre-processing is a preliminary practice performed on the raw data before using any data exploration algorithms to enhance the results performance [11]. Data Normalization standardize the raw data by converting them to fall in a small specified range using a linear transformation which can generate good quality clusters and improve the accuracy of clustering algorithms. The data normalization methods includes Z-score, Min-Max and Decimal scaling [13]. Normalization before clustering is necessary for distance metric, like the Euclidean distance that are sensitive to variations within the magnitude or scales from the attributes [19].

To extract latent clusters from the data, a data matrix, D, was constructed. D is represented as a collection of vectors D $=X_1, X_2, ..., X_m$. Each vector,X_i, represents a unique data instance (Subjectkey), and each vector element,X_{ij}, represents a specific measurement (attribute) for that point corresponding to proficiency in one of the five behavioral measurements including (irritability, lethargy, hyperactivity, stereotype and inappropriate speech). This proficiency was determined by summing the number of skills assessment questions answered affirmatively, resulting in a 461×5 dimensional matrix. Subjects were required to have all 5 variables to be included in the cluster analysis.

2.2.2 Clusters Number Determination

In the absence of domain information to suggest the correct number of clusters k from a given data set [28]. Several mechanisms exist for statistically determining that help to make a decision, including two methods direct and statistical testing. Direct methods are optimization of the criterion within cluster sum of squares (e.g elbow and silhouette methods)[33]. Statistical testing methods consist of comparing evidence against the null hypothesis (e.g gap statistic) [32].

2.2.3 Cluster Analysis

Further analyzed using the K-means [15], Partition around medoids (PAM)[29], and Hierarchical [10] clustering algorithms. In some biological applications it is difficult or impossible to define a labeled data set, which exclude the use of supervised machine-learning methods. In such cases, unsupervised machine-learning methods can be used to detect clusters or individual outlier objects of objects that differ from the control group in a data set [31]. Cluster analysis is the task of identifying the groups of the population or data points that are cohesive and separated from other data points in other groups [12]. K-means is a widely used and very simple partition based clustering method [16]. K-means algorithm automatically partition a data set into k groups. K-means algorithm requires three parameters specified by user: number of clusters K, cluster initialization, and distance metric.

The PAM clustering algorithm is also a partition based clustering method. PAM algorithm [29] is very similar to K-means mostly because both work by trying to minimize the error. However, outlier sensitivity is a known fault of K-means while PAM works with medoids to avoid the outlier sensitivity [26].

The hierarchical clustering technique is also one of the popular clustering techniques in Machine Learning. Hierarchical clustering is an algorithm that groups similar objects into a set of groups called (clusters), where each cluster is distinct from each other cluster, and the points (objects) within each cluster are largely similar to each other[18]. Although various cluster results were obtained from applying different input parameters of the three clustering algorithms when applied to the ASD data. The question remains: how many clusters actually in the data and which set of clusters is valid? Cluster validation refers to to design the procedures that evaluate the goodness of cluster analysis results in a quantitative and objective fashion [16]. The goal of cluster validation is to find the cluster partition set which is the most suitable to the input dataset. Compactness and Separateness are the two measurement criteria used in cluster validation for evaluating and selecting an optimal clustering scheme [20]. A good cluster algorithm result should yield well separated and compact clusters. There are different cluster validity indices types that measure the quality of clustering results. Validation indices based on internal criteria estimate the fit between the data by itself and the expected structure by the clustering algorithm (clusters)[5].

Since the underlying structure of the data is unknown two internal criteria validation indices were used for the evaluation of the multiple clustering results obtained on the ASD study population to measure the goodness of the clusters. Large values of Silhouette and Dunn indices correspond to a better overall quality of the clustering result. Feature selection/extraction is an essential aspect of all cluster analysis [23]. Using a large number of features (126 in our case) increases the probability of feature redundancy. The feature selection goal is to remove irrelevant features by finding the minimal feature subset necessary and sufficient to support the target concept. The feature subset should improve prediction accuracy. To determine which ASD features were significant and discriminant among the 126 features, we have benefited from the expertise of medical experts. We expected that the chosen 5 discriminant features would improve the prediction strength of the models.

3 Results

3.1 Clustering Results

Since our phenotype dataset is unlabeled, Unsupervised clustering algorithms were utilized to cluster this dataset and extract the core members of each clusters and profile the core member subsets to explore the characteristics using genotype data from the same dataset. The techniques used (in section 2.2.2) estimated 3 clusters within the dataset. we also reran our clustering algorithms with different k values (number of clusters) from 2 to 6. (we did not go beyond 6 due to the limited size of the ASD data). The best k-means results (as determined by cluster validity indices) was for k = 3. Based on that, we selected the K-means output with

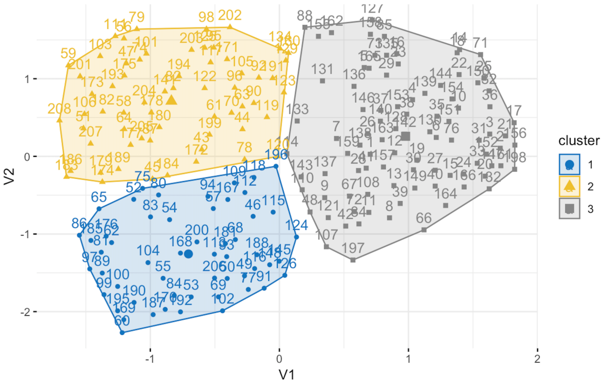


Figure 2: Clustering results visualization

k = 3 as the optimal cluster configuration, for the remaining two clustering techniques (PAM and hierarchical)we varied k from 3 to 6. (Fig.2) shows the 3 clusters identified within the 461 subjects ASD data set.

The computation of the silhouette score and dunn for this cluster assignment yielded the values 0.45 and 0.068, respectively. Following the clustering procedure, the k-means method was executed four times, varying the number of cluster from two to six. The silhouettes and Dunn index scores for these clustering allocations are shown in table.1, Best number according to each index is highlighted in bold.

Table 1: Results of k-means executions with dierent number of clusters.

Num. of clusters	Silhouette score	Dunn score
3	**0.45**	**0.068**
4	0.38	0.034
5	0.42	0.052
6	0.39	0.062

It is remarkable that this was precisely the number of clusters that resulted from the clusters number determination techniques, therefore providing a robust support for choosing three clusters as the suitable number k of groups. (Fig.3) shows the silhouette graph of k-mean clustering strategy k-means with three clusters. The cluster 3 shows the widest and highest silhouette, hence it is the strongest cluster and it comprises the most samples. The clusters 1 and 2 are a little narrower in comparison to the cluster 3 and they include less samples. large value of the silhouette score (around 1) for a given element indicates that there is no doubt the observations are very well clustered. On the contrary, a low silhouette score (around 0) suggests an observation belonging to the intersection of two clusters. However, observations with negative values are probably placed in the wrong cluster.

The next step, core members of each clusters are selected and their corresponding genomics data are extracted for annotation.

3.2 Genotype Annotation

First we clustered the patients based on their phenotypes. The original research took in total of 58 behavioral measurements in 5 categories, including irritability, lethargy, hyperactivity,

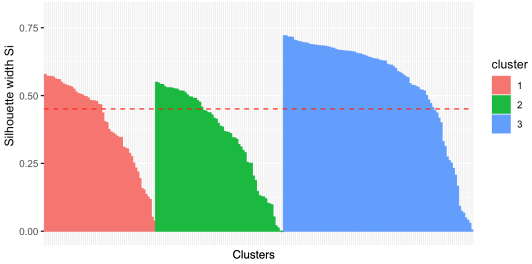
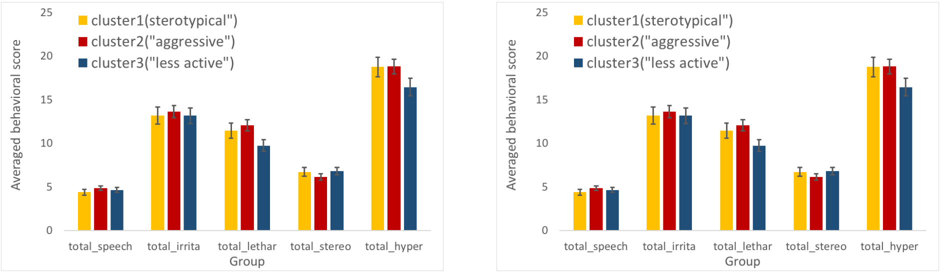


Figure 3: Silhouette values graph of kmeans clustering strategy



(a) All Samples

(b) Core Samples

Figure 4: Distribution of Clusters on the five behavioral measures

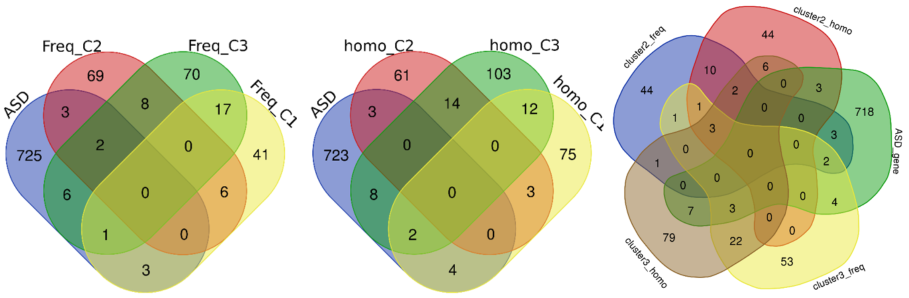
stereotype and Inappropriate Speech. We removed patients with too many missing measurements, and 461 patients were included in the final clustering results.

After the 461 ASD patients were clustered based on the behavioral phenotypes, the average score of each of the 5 measurements were plotted in Fig.4 with the Fig.4(a) shows all samples case and Fig.4(b) shows only core sample case (only top 10 closest samples to the center of each clusters) The alternative allele frequency ($F_{1/1}$) of each of the 1,329,577 SNV/INDEL loci was calculated for each cluster. Also, the frequency of homozygous for alternative loci ($H_{1/1}$) genotype (1/1) for each cluster was calculated for each locus. $H_{1/1} = n_{1/1}/(n_{all})$. $F_{1/1} = (n_{1/1} * 2 + n_{0/1})/(2 * n_{all})$. Loci with too many un-genotyped individuals (missing genotype $> 50\%$) would be excluded from the analysis In this way, for each locus we had six measurements $F_{1/1}1, F_{1/1}2, F_{1/1}3, H_{1/1}1, H_{1/1}2$ and$H_{1/1}3$, corresponding to the 3 clusters.

To find loci with uniquely high $H_{1/1}$ or $F_{1/1}$ for each cluster, we used the formula (saying to find loci with high homozygosity of alternative allele in cluster 1): $\delta H1 = H_{1/1}1 - (H_{1/1}2 + H_{1/1}3)/2$. Similarly, for loci with high alternative frequency in cluster 1, $\delta F1 = F_{1/1}1 - (F_{1/1}2 + F_{1/1}3)/2$. The cutoff we selected for δF was 0.15 and 0.2 for δH. Given the sample size (n=461), this cutoff gave us reasonable balance between sensitivity and false positive rate. The loci were then annotated using ANNOVAR [30] with hg19. Genes with the loci passed $\delta F > 0.12$ will be identified, genes with loci of $\delta H > 0.15$ will also be identified. The overlap between the two gene list and known ASD related gene (from SFARI) were found using online Venn diagram tool [2]. Results are shown in Fig.5.

Gene ontology (GO) and pathway analysis were performed using GSEA (Gene Set Enrichment Analysis)[1].

On the behavioral level, cluster 3 showed lowest score in lethargy and hyperactivity (Fig.4(a)). We found cluster 2 have slightly higher irritability score. Cluster 1 showed lowest speech score. We labeled cluster 1 as “speech“, cluster 2 as “irritability” and cluster 3 as “lethality“. The average score for 10 most representative patients from each cluster showed more significant



(a) (b) (c)

Figure 5: Venn diagrams for overlap between known ASD related genes and genes in each cluster harboring loci with distinct alternative allele frequency

difference among clusters (Fig.4(b)), and the direction of the difference were consistent with those observed from all patients. There were 109, 218 and 134 patients in the 3 clusters, respectively. The contrast between cluster 3 versus cluster 1 and 2 reach statistical significance for the measurement lethargy (T_test, p=0.013) and marginal for hyperactivity (T_test, p=0.053).

We next found the loci with high alternative allele frequency for each cluster. With cutoff of $\delta F > 0.12$(see method part), there were 134, 150 and 224 significant loci in cluster 1 to cluster 3, respectively . With $\delta H > 0.15$, there were 197, 142 and 287 significant loci in cluster 1 to cluster 3, respectively . After annotation, SNV/INDELs affect gene related regions (variations in intergenic, upstream, downstream regions were excluded) were linked to genes, results in gene list of 68,88,104 for $\delta F > 0.12$ in cluster 1 to 3, respectively. For $\delta H > 0.15$ in cluster 1 to 3, the gene list was 96, 80, 184, respectively. GO analysis and pathway analysis showed three GO terms were significant for C3_freq gene list, which are RESPONSE_TO_ABIOTIC_STIMULUS, TRANSMEMBRANE_TRANSPORT and PROTEIN_UBIQUITINATION . There were 85 GO terms significant for C3_homo list, with the top 5 are shown in table.2.

Table 2: Top five significant genes ontology (GO).

Rank	Genes
1	CELLULAR_POTASSIUM_ION_HOMEOSTASIS
2	ESTABLISHMENT_OR_MAINTENANCE_OF_TRANSMEMBRANE_ELECTROCHEMICAL_-GRADIENT
3	SODIUM_ION_EXPORT
4	POTASSIUM_ION_HOMEOSTASIS
5	CELLULAR_SODIUM_ION_HOMEOSTASIS

All these terms are related with neuronal excitability. Moreover, ATP1A4 and ATP1A2, two genes related with ATP metabolism and import/export ions, were in both C3_homo and C3_freq list. And mutation on ATP1A2 had been associated with hemiplegia [17]. Other GO terms significant for C3_homo list includes NEURON_PROJECTION_DEVELOPMENT, CELL_DEVELOPMENT, BIOLOGICAL_ADHESION and POSITIVE_REGULATION_OF_CELL_PROLIFERATION. 5 KEGG Pathways significant for C3_homo genes list, included KEGG_ALDOSTERONE_REGULATED_SODIUM_REABSORPTION, which confirmed the GO analysis results. Also, CELL_ADHESION_MOLECULES_CAMS was significant, suggested the importance of mutations in cell adhesion molecules in ASD. (see below for more discussion). TRANSMEMBRANE_TRANSPORT was the only GO term significant for C2_freq gene list. For C2_homo gene list, two GO terms, EXTRACELLULAR_STRUCTURE_ORGANIZATION and POSITIVE_REGULATION_OF_HYDROLASE_ACTIVITY were significant. KEGG pathways significant for C2_homo gene list included FOCAL_ADHESION and ECM_RECEPTOR_-

INTERACTION. These results were consistent with previous report of the importance of cell adhesion molecules in brain development and pathogenesis of Autism [17, 4]. No significant GO or pathways were found using C1_freq and C1_homo gene lists. This is consistent with the intermediate phenotypic score of cluster 1 in most of the measurements.

4 Conclusion

In this paper, we present a new approach of mining the phenotype and genotype bio-medical data. We treat the phenotype data as the unlabeled data and apply the clustering algorithm to extract the core members of each cluster that can represent the features of each clusters. These grouping features are verified by using genomic data for annotation. The annotation results provide a better understanding of the sub-categories of Autism. Multiple gene sets that can best explain the sub-categories were identified. In next step, we will set up automatic pipeline based on current work to identify alleles/genes/GOs for more ASD subtypes using other datasets from NDAR. Also, the alleles/genes identified for clusters identified using clinical phenotypes will be associated with phentypes to find strong and weak associations between genotype and phenotype, which can help to identify more efficient and reliable biomarkers for ASD diagnosis. The genes with strong association with specific subgroup would also be potential treatment targets for specific sub-type of ASD.

References

[1] gsea. http://software.broadinstitute.org/gsea/index.jsp). Accessed: 2019-02-30.

[2] Venn. http://bioinformatics.psb.ugent.be/webtools/Venn/). Accessed: 2019-03-30.

[3] Brett S Abrahams and Daniel H Geschwind. Advances in autism genetics: on the threshold of a new neurobiology. *Nature reviews genetics*, 9(5):341, 2008.

[4] Richard J.L. Anney and et al. Meta-analysis of gwas of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Acta Veterinaria Scandinavica*, 8(1), 5 2017.

[5] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, JesúS M PéRez, and IñIgo Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013.

[6] Leigh J Beglinger and Tristram H Smith. A review of subtyping in autism and proposed dimensional classification model. *Journal of Autism and Developmental Disorders*, 31(4):411–422, 2001.

[7] Catalina Betancur. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain research*, 1380:42–77, 2011.

[8] Pauline Chaste and et al. A genome-wide association study of autism using the simons simplex collection: Does reducing phenotypic heterogeneity in autism increase genetic homogeneity? *Biological Psychiatry*, 77(9):775–784, 5 2015.

[9] Eric Courchesne, Tiziano Pramparo, Vahid H. Gazestani, Michael V. Lombardo, Karen L. Pierce, and Nathan E. Lewis. The asd living biology: From cell proliferation to clinical phenotype. In *Molecular Psychiatry*, 2018.

[10] William HE Day and Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24, 1984.

[11] Pedro M Domingos. A few useful things to know about machine learning. *Commun. acm*, 55(10):78–87, 2012.

[12] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631, 2002.

[13] Ian Gibson and Christopher Amies. Data normalization techniques, July 10 2001. US Patent 6,259,456.

[14] Scott S Hall, Amy A Lightbody, Melissa Hirt, Ava Rezvani, and Allan L Reiss. Autism in fragile x syndrome: a category mistake? *Journal of the American Academy of Child & Adolescent Psychiatry*, 49(9):921–933, 2010.

[15] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[16] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[17] Arija G. Jansen, Gwen C. Dieleman, Philip R. Jansen, Frank C. Verhulst, Danielle Posthuma, and Tinca J. C. Polderman. Psychiatric polygenic risk scores as predictor for attention deficit/hyperactivity disorder and autism spectrum disorder in a clinical child and adolescent sample. *Behavior Genetics*, Jul 2019.

[18] Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[19] Seo Young Kim, Jae Won Lee, and Jong Sung Bae. Effect of data normalization on fuzzy clustering of dna microarray data. *BMC bioinformatics*, 7(1):134, 2006.

[20] Ferenc Kovács, Csaba Legány, and Attila Babos. Cluster validity measurement techniques. In *6th International symposium of hungarian researchers on computational intelligence*, page 35. Citeseer, 2005.

[21] Alan J Lincoln, Yvonne M Searcy, Wendy Jones, and Catherine Lord. Social interaction behaviors discriminate young children with autism and williams syndrome. *Journal of the American Academy of Child & Adolescent Psychiatry*, 46(3):323–331, 2007.

[22] Michael V Lombardo, Meng-Chuan Lai, Bonnie Auyeung, Rosemary J Holt, Carrie Allison, Paula Smith, Bhismadev Chakrabarti, Amber NV Ruigrok, John Suckling, Edward T Bullmore, et al. Unsupervised data-driven stratification of mentalizing heterogeneity in autism. *Scientific Reports*, 6:35333, 2016.

[23] Hiroshi Motoda and Huan Liu. Feature selection, extraction and construction. *Communication of IICM (Institute of Information and Computing Machinery, Taiwan) Vol*, 5(67-72):2, 2002.

[24] Tayo Obafemi-Ajayi, Judith H Miles, T Nicole Takahashi, Wenchuan Qi, Kristina Aldridge, Minqi Zhang, Shi-Qing Xin, Ying He, and Ye Duan. Facial structure analysis separates autism spectrum disorders into meaningful clinical subgroups. *Journal of autism and developmental disorders*, 45(5):1302–1317, 2015.

[25] Opal Ousley and Tracy Cermak. Autism spectrum disorder: defining dimensions and subgroups. *Current developmental disorders reports*, 1(1):20–28, 2014.

[26] Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341, 2009.

[27] Natalie A Pride, Jonathan M Payne, and Kathryn N North. The impact of adhd on the cognitive and academic functioning of children with nf1. *Developmental neuropsychology*, 37(7):590–600, 2012.

[28] Stan Salvador and Philip Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *16th IEEE International Conference on Tools with Artificial Intelligence*, pages 576–584. IEEE, 2004.

[29] Frederik T Verleysen and Arie Weeren. Clustering by publication patterns of senior authors in the social sciences and humanities. *Journal of Informetrics*, 10(1):254–272, 2016.

[30] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164, 07 2010.

[31] Wei Wu, Eugene Bleecker, Wendy Moore, William W Busse, Mario Castro, Kian Fan Chung, William J Calhoun, Serpil Erzurum, Benjamin Gaston, Elliot Israel, et al. Unsupervised phenotyping of severe asthma research program participants using expanded lung data. *Journal of Allergy and Clinical Immunology*, 133(5):1280–1288, 2014.

[32] Mingjin Yan and Keying Ye. Determining the number of clusters using the weighted gap statistic. *Biometrics*, 63(4):1031–1037, 2007.

[33] Ying Zhang, Semu Moges, and Paul Block. Optimal cluster analysis for objective regionalization of seasonal precipitation in regions of high spatial–temporal variability: application to western ethiopia. *Journal of Climate*, 29(10):3697–3717, 2016.