Deep learning prediction of attention-deficit hyperactivity disorder in African Americans by copy number variation

Yichuan Liu[1], Hui-Qi Qu[1], Xiao Chang[1], Kenny Nguyen[1], Jingchun Qu[1], Lifeng Tian[1], Joseph Glessner[1], Patrick MA Sleiman[1,2,3] and Hakon Hakonarson[1,2,3,4]

[1]Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; [2]Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA; [3]Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; [4]Division of Pulmonary Medicine, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

Corresponding authors: Yichuan Liu. Email: liuy5@email.chop.edu; Hakon Hakonarson. Email: hakonarson@email.chop.edu

Impact statement

Deep learning algorithms have achieved successes in many fields, including sequencing data. ADHD is a prevalent psychiatric disorder, and research show no single mutations or variants cause the disorder; in other words, ADHD is more likely caused or impacted by underlying pathways instead of single gene or variants based on previous studies. In this study, we collect 524 African Americans whole genome sequencing data, including both ADHD and controls, then applied multiple layer deep learning algorithm, using copy number variations as feature vectors to capture the genetic susceptibility and molecular underpinnings of ADHD. The accuracy of the prediction was evaluated using two-fold random shuffle tests, as well as an independent whole genome sequencing dataset that contain 351 European Americans. We found a superior and powerful labeling ability compared to traditional clustering methods; meanwhile, population-specific performance difference cannot be ruled out.

Abstract

Current understanding of the underlying molecular network and mechanism for attention-deficit hyperactivity disorder (ADHD) is lacking and incomplete. Previous studies suggest that genomic structural variations play an important role in the pathogenesis of ADHD. For effective modeling, deep learning approaches have become a method of choice, with ability to predict the impact of genetic variations involving complicated mechanisms. In this study, we examined copy number variation in whole genome sequencing from 116 African Americans ADHD children and 408 African American controls. We divided the human genome into 150 regions, and the variation intensity in each region was applied as feature vectors for deep learning modeling to classify ADHD patients. The accuracy of deep learning for predicting ADHD diagnosis is consistently around 78% in a two-fold shuffle test, compared with ∼50% by traditional k-mean clustering methods. Additional whole genome sequencing data from 351 European Americans children, including 89 ADHD cases and 262 controls, were applied as independent validation using feature vectors obtained from the African American ethnicity analysis. The accuracy of ADHD labeling was lower in this setting (∼70–75%) but still above the results from traditional methods. The regions with highest weight overlapped with the previously reported ADHD-associated copy number variation regions, including genes such as *GRM1* and *GRM8*, key drivers of metabotropic glutamate receptor signaling. A notable discovery is that structural variations in non-coding genomic (intronic/intergenic) regions show prediction weights that can be as high as prediction weight from variations in coding regions, results that were unexpected.

Keywords: Deep learning, African Americans, attention-deficit hyperactivity disorder, copy number variations, whole genome sequencing

***Experimental Biology and Medicine* 2021; 246: 2317–2323. DOI: 10.1177/15353702211018970**

Introduction

Attention-deficit hyperactivity disorder (ADHD) is a common psychiatric disorder with prevalence of 6–8% in children, with symptoms persisting into adulthood in over two-thirds of cases, causing significant life-long impairments.[1–3] While enrichment of certain copy number variations (CNVs) have shown associations with genetic susceptibility of ADHD,[4] the genetic underpinnings of ADHD remain largely unknown in part due to the high degree of ADHD heterogeneity, suggesting that the

ISSN 1535-3702

molecular mechanisms underlying ADHD are complex, and are likely to constitute multiple gene networks not readily addressed with current genotyping and sequencing platforms.[5] In addition, minority ethnicities such as African Americans (AA) have been less well studied compared to European Americans (EA), and the impact of non-coding genomic structural variations, e.g. CNVs, inversions, and translocations, has been understudied in human psychiatric disorders, including ADHD, hampering the advancement of the field.[6,7]

Machine learning methods, especially the multiple layer deep learning algorithm, have been applied on complex biological data to explore the underlying molecular factors.[8,9] To capture the genetic susceptibility and molecular underpinnings of ADHD,[5,10] deep learning model has multiple advances that are particularly useful when analyzing complex data. In this study, we performed whole genome sequencing (WGS) with high read depth (>30× coverage) for 524 AA children, including 116 ADHD patients and 408 controls, and we established a multi-layer perceptron (MLP) neuronal network using CNVs and other structural variation intensities from different genomic regions as feature vectors in order to separate the ADHD cases from controls. The accuracy of the prediction was evaluated using two-fold random shuffle tests, as well as an independent WGS dataset that contain 351 EA children (89 ADHD vs. 262 controls). Compared to traditional clustering methods, our results support a superior and powerful labeling ability of the deep learning algorithm compared to conventional methods, with non-coding structural variations including CNVs demonstrating particular robustness to the classification, beyond expectations.

Materials and methods

ADHD individuals' selection and WGS processing

The patients were recruited as the Philadelphia Neurodevelopmental Cohort (PNC), archived in the biobank of the Center for Applied Genomics (CAG) at the Children's Hospital of Philadelphia (CHOP); more details could be found in the eMerge project stage III.[7,11,12] Additional information for the 205 ADHD patients, including 116 AA and 89 EA, and 670 controls, such as gender and age could be found in Table 1. All methods were carried out in accordance with relevant guidelines and regulations, and all experimental protocols were approved by the Institutional Review Board (IRB) of Children's Hospital of Philadelphia (CHOP). Informed consent was obtained from all subjects or, if subjects are under 18, from a parent and/or legal guardian with assent from the child if seven years or older.

WGS based on the Illumina platform were processed using the standard pipeline. More specifically, the data were aligned to the GRCh37 reference using bwa v0.7.10[13] and BEDTools 2.17.0.[14] Aligned SAM files were processed with Samtools v0.1.19.[15] More details were described in our previous publications.[7,12] The CNVs were detected by MANTA,[16] and the CNVs that passed the default threshold were categorized into different classes based on genomic annotations, including "exonic," "intronic," and "intergenic."[11]

Genomics feature vector selections for deep learning models

The human genome region was divided into 150 pieces (20 M bp/piece) based on genomic coordinates. The occurrence counts of nine different types of variations, including deletions, duplications, and other types of variations in exonic, intronic, and intergenic regions (Table 2), were calculated, respectively, for each piece, which was applied as a feature vector in the deep learning model. The processes were repeated for all individuals in the study. A random forest algorithm was introduced to reduce the feature vectors by computing relative importance or contribution of each genomic piece. Feature vectors that have weight with zero importance were removed for different types of variations. The programming codes are in Python language built on the Scikit-learn package (version 0.21.3).[17]

Table 2. Predictive accuracy for 351 EA individuals based on 574 AA for different types of CNVs.

CNV types	Accuracy in 351 EA
Exonic deletion	71.3%
Exonic duplication	75.4%
Other SVs in exonic	75.4%
Intergenic deletion	75.4%
Intergenic duplication	75.4%
Other SVs in intergenic	75.4%
Intronic deletion	75.4%
Intronic duplication	75.4%
Other SVs in Intronic	75.4%

CNV: copy number variation; EA: European Americans.

Table 1. General information of the research subjects.

			Age		Gender	
Ethnicity	Phenotype	Number	Mean±STD	Median	Female	Male
AA	Cases	116	22.2 ± 3.3	22.0	42 (36.2%)	74 (63.8%)
	Controls	408	23.7 ± 3.5	24.0	232 (56.9%)	176 (43.1%)
EA	Cases	89	22.5 ± 3.7	21.6	29 (32.6%)	60 (67.4%)
	Controls	262	24.2 ± 3.7	24.3	129 (49.2%)	133 (50.8%)

AA: African Americans; EA: European Americans.

Deep learning parameters, random shuffled two-fold tests, and traditional clustering methods

MLP from the Scikit-learn package (version 0.21.3)[17] was applied as the deep learning model based on 15 different types of mutations. Parameters for deep learning model, including maximum iterations, alpha value in L2 regularization, activation functions, solvers, learning rate, number of layers, and numbers of neurons per layer, were optimized using "gp_minimize" function from the scikit-optimise 0.7.2 python library. Most activation functions for CNVs of MLP are "relu" except intronic deletion CNVs, which is "logistic"; the solver include "sgd" and "lbfgs," while neurolayers ranged from seven to nine.

A two-fold random shuffle test was applied to test the predictive abilities. Five hundred and twenty-four AA patients, including 116 ADHD patients were split into equal ratio randomly. One dataset was applied as the training set and the other one used as the testing set. The procedures were repeated for 50 times independently. Feature vectors in genome were selected as described in the previous paragraph in order to build the deep learning model, and then individuals in the testing set were labeled as ADHD or controls. To compare with traditional clustering methods, we used k-means algorithm since it has stable and excellent performance if the number of clusters is known, which is 2 in our case.

Results

Phenotype prediction accuracy in 574 AA children

A two-fold random shuffle test was applied to assess the ADHD labeling prediction for 50 rounds. Reduced feature vectors, which are based on the random forest algorithm, show a reproducible prediction accuracy (average accuracy 78%, with ~3% standard deviations) in classifying ADHD individuals versus controls using the deep learning model (Figure 1), using optimized parameters as described in the method section. In contrast, the accuracy of the traditional k-means clustering based on the same set of feature vectors was only ~50%, or equal to random analysis. Interestingly, and a notable observation is that structural variations in non-coding genomes include intronic and intergenic regions and showed similar level of predictive accuracy compared to structural variations in coding regions.

Phenotype prediction accuracy in an independent dataset of 351 EA children

To verify the accuracy of the deep learning model, WGS data from 351 EA individuals, including 89 ADHD children, were investigated as an independent testing set. The feature vectors were selected based on the data from the 524 AA subjects. While the accuracy of labeling is slightly reduced, it is still above 70% (Table 1). Of note, the size of the training set is larger than the testing set (524 vs. 351), which implies that prediction using feature vectors is more robust compared to the two-fold tests. However, the accuracy level was decreased, which suggests that there are genomic level differences between the two ethnicities for ADHD patients, as well as a potential for population-specific performance of the deep learning model.

Genomics regions with high weights based on the deep learning model

The weight or the contribution for each genomic region (feature vector) is based on 524 AA individuals and calculated using the Random Forest algorithm, as described in the method section. The genomics regions (as feature vectors) containing CNVs show different weights in the prediction model (Figure 2(a)). Some regions containing coding CNVs have significantly higher weight and overlap with the previously reported ADHD-associated CNVs (Table 3). Among these regions, the feature vectors at chr7:1–20000000 and chr6:140000001–160000000, which ranked as the 5th and 17th highest weight (2.1%, 1.8%) in coding region CNV deletions, overlap with the *GRM8* and *GRM1* gene regions as reported previously (deletion at chr7: 126,525,124–126,536,202 and duplication at chr6: 146,657,076–146,694,047).[4] *GRM8* and *GRM1* are key drivers of metabotropic glutamate receptor (mGluR) neurotransmitter signaling, a pathway shown in a previous study to harbor CNVs that are enriched in both ADHD and autism cases[4] who responded in a significant way (CGI-I and Vanderbilt rating scale assessment, $p < 0.001$) to an mGluR activator drug in a clinical trial setting for ADHD, with seven of the children in the trial having autism as comorbid symptom.[11] Genomic regions with non-coding CNVs (intronic/intergenic) showed more uniformed weight distribution compared to coding region CNVs (Figure 2(b) and (c)), which could be explained by non-coding region CNVs mainly serving as a biomarker of ADHD genetic susceptibility conferred by functional genetic variations in each region.

Discussion

Growing evidence indicate that genomic structural variations, most notably CNVs, are likely to associate with and influence the development of psychiatric disorders, including ADHD.[18,19] However, to date, CNVs have not yet to be applied as predictive feature vectors in labeling phenotypic status of the patients. One of the major obstacles is the structural variations especially in large non-coding regions. While most CNVs in the human genome are benign, the functional studies for CNVs are challenging and expensive. The current solution has been to search for over-represented CNVs that are enriched in ADHD patients compared to controls. One of the main deficiencies of this approach is that structural variations that contribute to the essential underlying network which has less significant p value may be missed, especially if residing in non-coding regions.

Deep learning algorithms and models have been proven to be effective when applied on complex biological data, including WGS,[9] and have been successfully applied in multiple studies, such as prediction of the drug resistance and classifications of primary/metastatic cancers.[20,21] In this study, we recruited 524 individuals from the minority group of AA, including 116 ADHD patients and 408

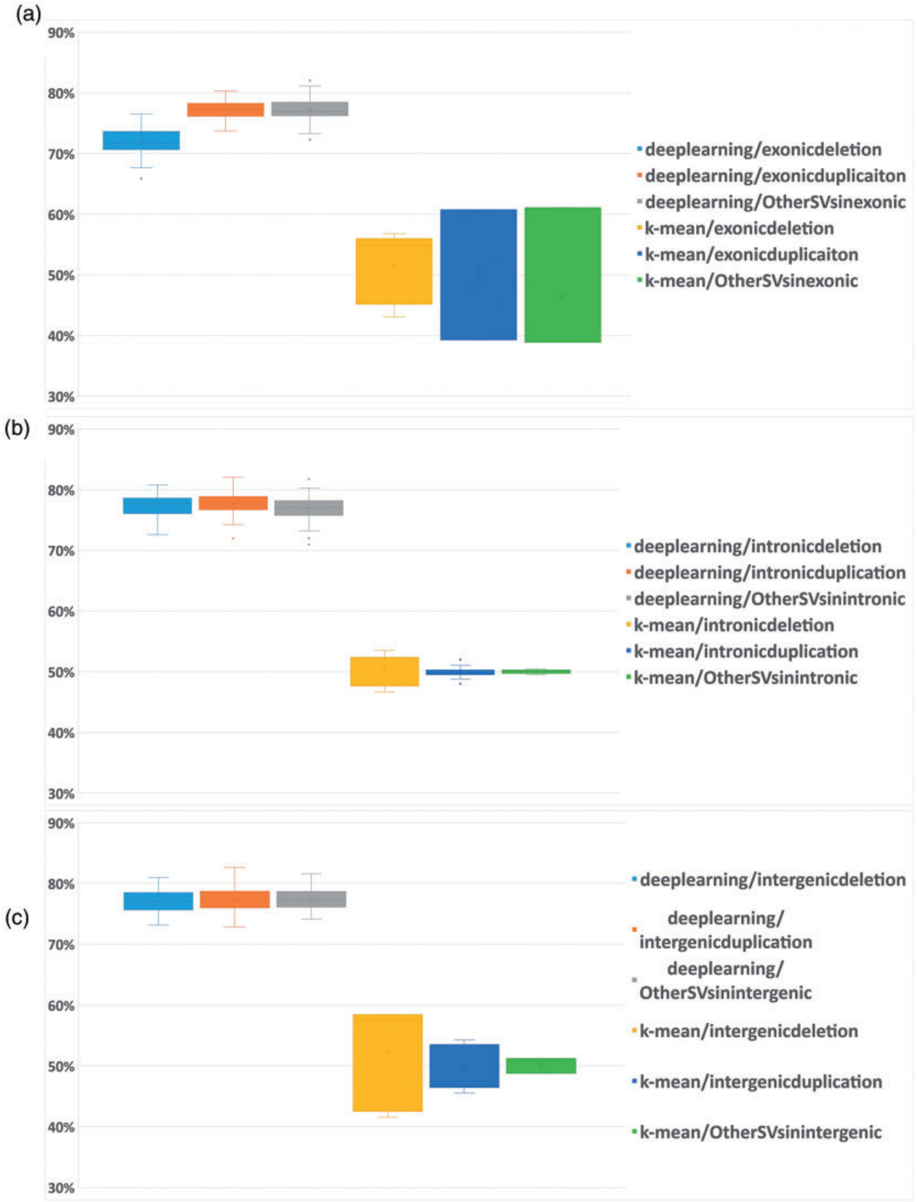

Figure 1. Prediction accuracy boxplots of deep learning model compared to traditional K-mean clustering in two-fold random shuffle tests for AA. (a) CNVs and other type of structural variations located in exonic regions; (b) in intronic regions; (c) in intergenic regions.

healthy controls, and generated >30× coverage WGS data. Instead of focusing on an overall enrichment analysis for structural variations, we divided the entire human genome into 150 regions and calculated the structural variations' intensity in each genomic region, then a MLP neuronal networks, with an advantage of solving extremely complex problem,[17] were applied using CNV intensity in each region as a feature vector. The model's accuracy was first evaluated through a two-fold random shuffle testing and repeated 50 times. The results (Figure 1) showed that the accuracy is reproducible and stable around 78%, which is a significant improvement compared to traditional clustering method, such as k-mean, which in this dataset performed at random (50%). To test the model for more general applications and to explore the potential differences between AA and EA, additional WGS data from 351 EA children, including 89 ADHD and 262 controls, were generated using exactly the same protocols and pipelines. The 524 AA samples were applied as the training data, while 351 EA samples were used as the testing data. The accuracy was reduced but still above 70% (Table 1), indicating there may be potential genomic differences between the two population ethnicities since more training data resulted in a less accuracy level compared to the AA random shuffle tests. A population-specific performance difference of the deep learning model for disease prediction cannot be ruled out either.

The weight of each genomic region was computed using the random forest algorithm based on the contributions for 524 AA children, as described in the method section. Multiple regions with high rank/weight contain coding region CNVs that are associated with ADHD or have

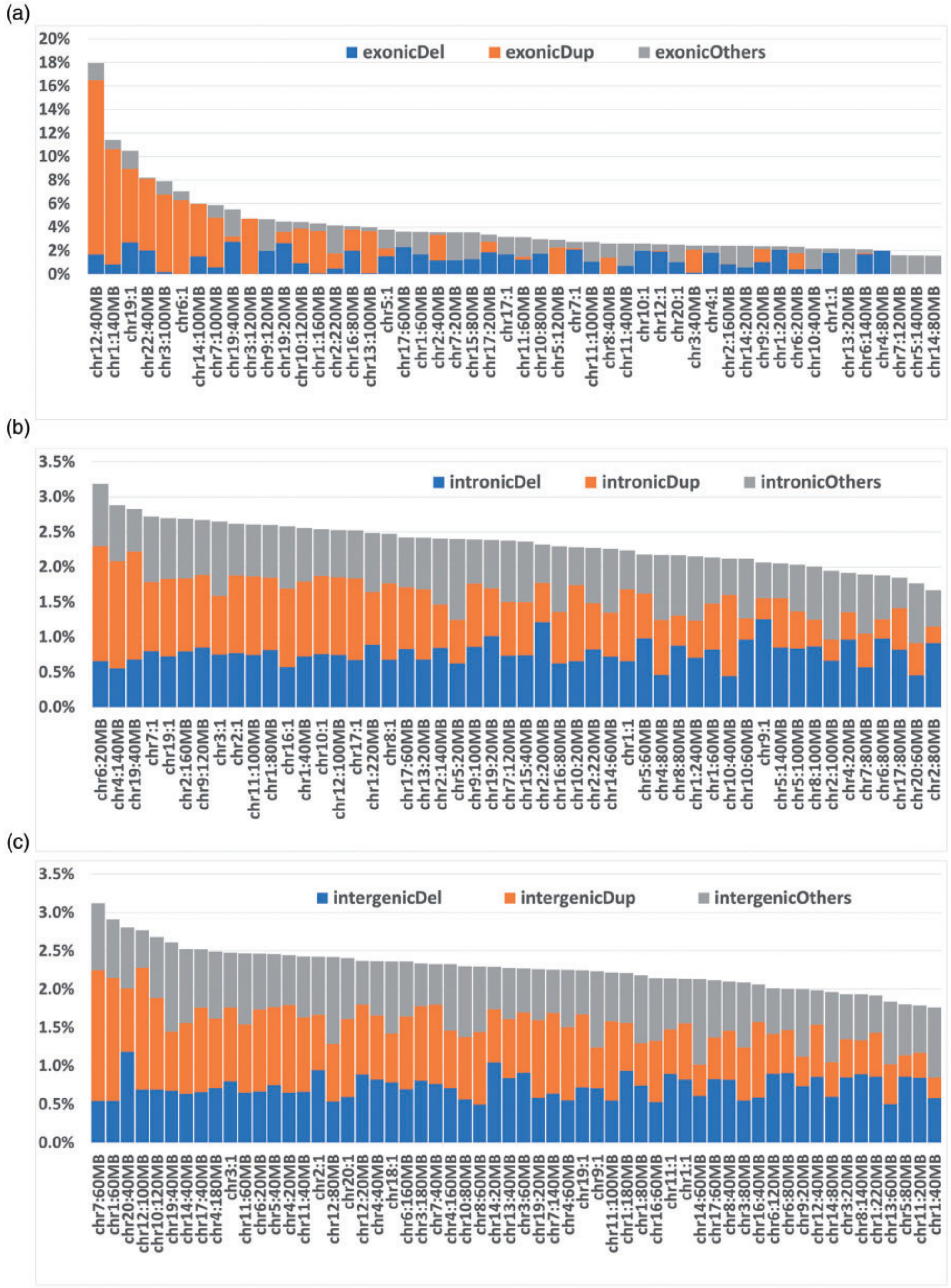

Figure 2. High weights/contribution genomic regions selected based on random forest algorithm. (a) CNVs and other type of structural variations located in exonic regions; (b) in intronic regions; (c) in intergenic regions.

functional roles within the ADHD gene networks established by previous studies (Table 3). A representative example is the high weight regions that contain the *GRM1* and *GRM8* CNVs. These two genes belong to the mGluR neurotransmitter signaling network, as highlighted in a recent clinical trial report, demonstrating that the mGluR activator NFC-1 could improve the ADHD symptoms significantly in adolescents with CNVs that reside within the glutamatergic gene (mGluR) networks and disrupt the mGluR neurotransmitter signaling.[22]

Two regions (chr12:40000001–60000000 and chr1: 140000001–160000000) showed high weight (15%/9%, respectively) in terms of intensity of exonic duplications. None of the regions contains any previously reported ADHD associated CNVs; however, previous studies suggest tendency of significance in the QTL study of deletions for the region of chr12:40000001–60000000,[23,24] while the region of chr1: 140000001–160000000 contains duplications with incomplete penetrant phenotype and psychiatric problems according to two independent case reports, suggesting a potential functional role in ADHD.[25,26]

A notable observation regarding the ADHD labeling is that CNVs and other structural variations in non-coding genomic regions, such as the intronic and intergenic

Table 3. High weighted genomics regions that overlapped with the previously reported ADHD-associated CNVs.

Feature vector CNV type	Genomics region	Rank/ weight	Overlapped ADHD-associated CNVs	Genes in previous ADHD-associated studies
Coding region deletions	chr7:1–20000000	5/2.1%	chr7: 126,525,124–126,536,202 (DEL)	*GRM8*
	chr6:140000001–160000000	17/1.8%	chr6: 146,657,076–146,694,047 (DUP)	*GRM1*
	chr1:60000001–80000000	19/1.7%	chr1: 72,317,292–72,328,395 (DUP)	*NEGR1*
	chr1:40000001–60000000	53/0.8%	chr1: 56,053,497–56,064,495 (DEL)	*USP24*
	chr19:20000001–40000000	3/2.6%	chr19: 38,427,720–38,444,834 (DEL)	*SLC7A10*
	chr19:1–20000000	2/2.7%	chr19: 15,992,679–15,997,923 (Denovo DEL)	*LOC126536*
	chr17:60000001–80000000	4/2.3%	chr17: 71,112,486–71,120,734 (Denovo DEL)	*KIAA1783*
	chr12:40000001–60000000	18/1.7%	chr12: 55,902,280–55,923,860 (Denovo DEL)	*NDUFA4L2, NXPH4, SHMT2, STAC3*
	chr19:40000001–59128983	1/2.7%	chr19: 59,423,491–59,428,132 (Denovo DUP)	*LILRB3, LIR-3*
	chr16:80000001–90354753	8/2%	chr16: 87,694,595–87,778,383 (Denovo DEL)	*AX748415, CDH15, LOC197322*
Coding region duplications	chr7:140000001–159138663	22/1%	chr7: 153,495,598–153,564,827 (DUP)	*DPP6*

CNV: copy number variation; ADHD: attention-deficit hyperactivity disorder.

regions, provide relatively comparable accuracy level to structural variations in exonic regions (Figure 1). It indicates that CNVs located in non-coding genomic regions may either be of functional/regulational impact themselves or may be tagging the genetic susceptibility of ADHD at nearby regions, including coding regions. Thus, the genetic association may be from functional impacts of the CNVs, or from the linkage disequilibrium with functional variations in other regions. For example, we found chr3:1–20000000 region weighed 1% in prediction (2nd rank) when using intronic structural variations as feature vectors. The region contains the *CNTN4* gene, which is an essential interactor gene in the mGluR pathway and mutation positive patients were found of clinically meaningful and statistically significant response to the treatments targeting mGluR signaling.[22]

The major limitation of this study is the sample size due to the cost of WGS with deep coverage. To avoid overfitting problem in the deep learning model, we had to keep the number of feature vectors around 150, which lead to relatively large genomic regions (~20 M bp for each) for the fixed human genome length (~3 billion bp). In other words, the resolutions of highly weighted regions are relatively modest. To increase the resolution warrants further research efforts and methodological development.

AUTHORS' CONTRIBUTIONS

Conceptualization: HH and YL; literature search: YL and HQ; data analysis: YL, HQ, and CX; data interpretation: YL, HQ, CX, KN, JQ, LT, JG, PS, and HH; original draft writing: YL and HQ; review and revision: YL, HQ, and HH; supervision: HH.

DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICAL APPROVAL

This study was approved by the Children's Hospital of Philadelphia (CHOP) Institutional Review Board (IRB). All the participants signed the informed consent form.

FUNDING

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The study was supported by Institutional Development Funds from the Children's Hospital of Philadelphia to the Center for Applied Genomics, The Children's Hospital of Philadelphia Endowed Chair in Genomic Research to HH.

Data availability

The data has been uploaded to the database of Genotypes and Phenotypes (dbGaP, https://www.ncbi.nlm.nih.gov/gap/) with the accession number phs001165

ORCID iDs

Yichuan Liu https://orcid.org/0000-0003-2023-072X
Hui-Qi Qu https://orcid.org/0000-0001-9317-4488
Jingchun Qu https://orcid.org/0000-0003-1974-2496

REFERENCES

1. Polanczyk GV, Willcutt EG, Salum GA, Kieling C, Rohde LA. ADHD prevalence estimates across three decades: an updated systematic review and meta-regression analysis. *Int J Epidemiol* 2014;**43**:434–42
2. Visser SN, Danielson ML, Bitsko RH, Holbrook JR, Kogan MD, Ghandour RM, Perou R, Blumberg SJ. Trends in the parent-report of health care provider-diagnosed and medicated attention-deficit/hyperactivity disorder: United States, 2003-2011. *J Am Acad Child Adolesc Psychiatry* 2014;**53**:34–46 e2
3. Barbaresi WJ, Colligan RC, Weaver AL, Voigt RG, Killian JM, Katusic SK. Mortality, ADHD, and psychosocial adversity in adults with childhood ADHD: a prospective study. *Pediatrics* 2013;**131**:637–44
4. Elia J, Glessner JT, Wang K, Takahashi N, Shtir CJ, Hadley D, Sleiman PM, Zhang H, Kim CE, Robison R, Lyon GJ, Flory JH, Bradfield JP, Imielinski M, Hou C, Frackelton EC, Chiavacci RM, Sakurai T, Rabin C, Middleton FA, Thomas KA, Garris M, Mentch F, Freitag CM, Steinhausen HC, Todorov AA, Reif A, Rothenberger A, Franke B, Mick EO, Roeyers H, Buitelaar J, Lesch KP, Banaschewski T, Ebstein RP, Mulas F, Oades RD, Sergeant J, Sonuga-Barke E, Renner TJ, Romanos M, Romanos J, Warnke A, Walitza S, Meyer J, Palmason H, Seitz C, Loo SK, Smalley SL, Biederman J, Kent L, Asherson P, Anney RJ, Gaynor JW, Shaw P, Devoto M, White PS, Grant SF, Buxbaum JD, Rapoport JL, Williams NM, Nelson SF, Faraone SV, Hakonarson H. Genome-wide copy number variation study associates metabotropic glutamate receptor gene networks with attention deficit hyperactivity disorder. *Nat Genet* 2011;**44**:78–84
5. Connolly JJ, Glessner JT, Elia J, Hakonarson H. ADHD & pharmacotherapy: past, present and future: a review of the changing landscape of drug therapy for attention deficit hyperactivity disorder. *Ther Innov Regul Sci* 2015;**49**:632–42
6. Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, Baldursson G, Belliveau R, Bybjerg-Grauholm J, Baekvad-Hansen M, Cerrato F, Chambert K, Churchhouse C, Dumont A, Eriksson N, Gandal M, Goldstein JI, Grasby KL, Grove J, Gudmundsson OO, Hansen CS, Hauberg ME, Hollegaard MV, Howrigan DP, Huang H, Maller JB, Martin AR, Martin NG, Moran J, Pallesen J, Palmer DS, Pedersen CB, Pedersen MG, Poterba T, Poulsen JB, Ripke S, Robinson EB, Satterstrom FK, Stefansson H, Stevens C, Turley P, Walters GB, Won H, Wright MJ, Consortium AWGotPG, Early L, Genetic Epidemiology C, and Me Research T, Andreassen OA, Asherson P, Burton CL, Boomsma DI, Cormand B, Dalsgaard S, Franke B, Gelernter J, Geschwind D, Hakonarson H, Haavik J, Kranzler HR, Kuntsi J, Langley K, Lesch KP, Middeldorp C, Reif A, Rohde LA, Roussos P, Schachar R, Sklar P, Sonuga-Barke EJS, Sullivan PF, Thapar A, Tung JY, Waldman ID, Medland SE, Stefansson K, Nordentoft M, Hougaard DM, Werge T, Mors O, Mortensen PB, Daly MJ, Faraone SV, Borglum AD, Neale BM. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat Genet* 2019;**51**:63–75
7. Liu Y, Chang X, Qu H, Glessner J, Tian L, Li D, Qiu H, Sleiman PMA, Hakonarson H. Non-coding structural variation differentially impacts attention-deficit hyperactivity disorder (ADHD) gene networks in African American vs Caucasian children. *Sci Rep* 2020;**10**:15252
8. Schmidt B, Hildebrandt A. Deep learning in next-generation sequencing. *Drug Discov Today* 2021;**26**:173–80
9. Xu C, Jackson SA. Machine learning and complex biological data. *Genome Biol* 2019;**20**:76
10. Akutagava-Martins GC, Rohde LA, Hutz MH. Genetics of attention-deficit/hyperactivity disorder: an update. *Expert Rev Neurother* 2016;**16**:145–56
11. Calkins ME, Merikangas KR, Moore TM, Burstein M, Behr MA, Satterthwaite TD, Ruparel K, Wolf DH, Roalf DR, Mentch FD, Qiu H, Chiavacci R, Connolly JJ, Sleiman PMA, Gur RC, Hakonarson H, Gur RE. The Philadelphia neurodevelopmental cohort: constructing a deep phenotyping collaborative. *J Child Psychol Psychiatry* 2015;**56**:1356–69
12. Liu Y, Chang X, Qu HQ, Tian L, Glessner J, Qu J, Li D, Qiu H, Sleiman P, Hakonarson H. Rare recurrent variants in noncoding regions impact Attention-Deficit hyperactivity disorder (ADHD) gene networks in children of both African American and European American ancestry. *Genes (Basel)* 2021;**12**:310
13. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**:1754–60
14. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2
15. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. Genome project data processing S. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9
16. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, Cox AJ, Kruglyak S, Saunders CT. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 2016;**32**:1220–2
17. Pedregosa FV, Gael Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in python. J Mach Learn Res 2011;**12**:2825–30
18. Faraone SV, Larsson H. Genetics of attention deficit hyperactivity disorder. *Mol Psychiatry* 2019;**24**:562–75
19. Gudmundsson OO, Walters GB, Ingason A, Johansson S, Zayats T, Athanasiu L, Sonderby IE, Gustafsson O, Nawaz MS, Jonsson GF, Jonsson L, Knappskog PM, Ingvarsdottir E, Davidsdottir K, Djurovic S, Knudsen GPS, Askeland RB, Haraldsdottir GS, Baldursson G, Magnusson P, Sigurdsson E, Gudbjartsson DF, Stefansson H, Andreassen OA, Haavik J, Reichborn-Kjennerud T, Stefansson K. Attention-deficit hyperactivity disorder shares copy number variant risk with schizophrenia and autism spectrum disorder. *Transl Psychiatry* 2019;**9**:258
20. Deelder W, Christakoudi S, Phelan J, Benavente ED, Campino S, McNerney R, Palla L, Clark TG. Machine learning predicts accurately Mycobacterium tuberculosis drug resistance from whole genome sequencing data. *Front Genet* 2019;**10**:922
21. Jiao W, Atwal G, Polak P, Karlic R, Cuppen E, Subtypes PT, Danyi A, de Ridder J, van Herpen C, Lolkema MP, Steeghs N, Getz G, Morris Q, Stein LD, Consortium P; Clinical Translation Working G. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat Commun* 2020;**11**:728
22. Elia J, Ungal G, Kao C, Ambrosini A, De Jesus-Rosario N, Larsen L, Chiavacci R, Wang T, Kurian C, Titchen K, Sykes B, Hwang S, Kumar B, Potts J, Davis J, Malatack J, Slattery E, Moorthy G, Zuppa A, Weller A, Byrne E, Li YR, Kraft WK, Hakonarson H. Fasoracetam in adolescents with ADHD and glutamatergic gene network variants disrupting mGluR neurotransmitter signaling. *Nat Commun* 2018;**9**:4
23. Fisher SE, Francks C, McCracken JT, McGough JJ, Marlow AJ, MacPhie IL, Newbury DF, Crawford LR, Palmer CG, Woodward JA, Del'Homme M, Cantwell DP, Nelson SF, Monaco AP, Smalley SL. A genomewide scan for loci involved in attention-deficit/hyperactivity disorder. *Am J Hum Genet* 2002;**70**:1183–96
24. Elia J, Gai X, Xie HM, Perin JC, Geiger E, Glessner JT, D'Arcy M, deBerardinis R, Frackelton E, Kim C, Lantieri F, Muganga BM, Wang L, Takeda T, Rappaport EF, Grant SF, Berrettini W, Devoto M, Shaikh TH, Hakonarson H, White PS. Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. *Mol Psychiatry* 2010;**15**:637–46
25. Xavier J, Zhou B, Bilan F, Zhang X, Gilbert-Dussardier B, Viaux-Savelon S, Pattni R, Ho SS, Cohen D, Levinson DF, Urban AE. Laurent-Levinson C. 1q21.1 microduplication: large verbal-nonverbal performance discrepancy and ddPCR assays of HYDIN/HYDIN2 copy number. *NPJ Genom Med* 2018;**3**:24
26. Kaymakçalan H, Li P. 1q21.1 deletions and duplications in 2 siblings with psychiatric problems. *Indian J Pediatr* 2019;**86**:1068

(Received March 5, 2021, Accepted May 1, 2021)