BMJ Paediatrics Open

NeuroCNVscore: A tissue specific framework to prioritizing the pathogenicity of CNVs in neurodevelopmental disorders

Journal:	*BMJ Paediatrics Open*
Manuscript ID	bmjpo-2023-001966
Article Type:	Original research
Date Submitted by the Author:	17-Mar-2023
Complete List of Authors:	Liu, Xuanshi; Beijing Children's Hospital Xu, Wenjian; Beijing Children's Hospital, Biology Leng, Fei; Beijing Children's Hospital Zhang, Peng; Beijing Children's Hospital Guo, Ruolan; Beijing Children's Hospital Zhang, Yue; Beijing Children's Hospital Hao, Chanjuan; Beijing Children's Hospital Ni, Xin; Beijing Children's Hospital, Department of Otolaryngology, Head and Surgery Li, Wei; Beijing Children's Hospital
Keywords:	Genetics

Title: NeuroCNVscore: A Tissue Specific Framework to Prioritize the Pathogenicity of CNVs in Neurodevelopmental Disorders

Short title: Prioritizing the pathogenicity of CNVs

Authors: Xuanshi Liu[1], Wenjian Xu[1], Fei Leng[1], Peng Zhang[1], Ruolan Guo[1], Yue Zhang[1], Chanjuan Hao[1*], Xin Ni[2*], Wei Li[1*]

Affiliations:

[1]Beijing Key Laboratory for Genetics of Birth Defects, Beijing Paediatric Research Institute; MOE Key Laboratory of Major Diseases in Children; Genetics and Birth Defects Control Centre, Beijing Children's Hospital, Capital Medical University, National Centre for Children's Health, Beijing, China;

[2]Department of Otolaryngology, Head and Surgery, Beijing Children's Hospital, Capital Medical University, National Centre for Children's Health, Beijing, China.

[*] **Corresponding authors.** Emails: liwei@bch.com.cn (Li W.), nixin@bch.com.cn (Ni X.), hchjhchj@163.com (Hao C.).

Word Count: 4364

Fundings: This work was partially supported by the Ministry of Science and Technology of China (2019YFA0802104; 2016YFC1000306); the National Natural Science Foundation of China (31830054); the Beijing Natural Science Foundation (5222007); and the Beijing Municipal Health Commission (JingYiYan 2018-5).

Abstract

Background: Neurodevelopmental disorders (NDDs) are associated with altered development of the brain especially in childhood. Copy number variants (CNVs) play a crucial role in the genetic etiology of NDDs by disturbing gene expression directly at linear sequence or remotely at three-dimensional genome level which can exert in a tissue-specific manner. There are tools for prioritizing the pathogenicity of CNVs, but none focuses specifically on NDDs, although the increased number of NDD studies using whole-genome sequencing has generated a large amount of CNVs. **Methods:** Using an XGBoost classifier, we integrated 189 features that represent genomic sequences, gene information, and functional/genomic segments for evaluating genome-wide CNVs in a neuro/brain-specific manner. We utilized Human Phenotype Ontology to construct an independent NDD-related set. **Results:** Our neuroCNVscore framework (https://github.com/lxsbch/neuroCNVscore) achieved high predictive performance (PR = 0.82; AUC = 0.85) and outperformed an existing reference method SVScore. Predicted pathogenic CNVs were enriched in known autism associated genes. **Conclusions**: The neuroCNVscore prioritizes functional, deleterious and pathogenic CNVs in NDDs at whole genome-wide level, which is important for genetic studies and clinical genomic screening of NDDs as well as for providing novel biological insights into NDDs.

Key Words: Neurodevelopmental disorder; Copy number variant; Pathogenicity; Tissue specificity; Gene expression

Key Messages:

- **What is already known on this topic**

CNVs are important in the genetic etiology of NDDs. Systematic identification of CNV pathogenicity by virtue of their size, number and impact on genome is challenge. Several tools are available to evaluate CNVs or structural variants, but none on CNVs specific in NDDs.

- **What this study adds**

The neuroCNVscore is a useful tool in prioritizing functional and/or pathogenic CNVs in NDDs at whole genome-wide level in a neuro/brain-specific manner.

- **How this study might affect research, practice or policy**

Given the expanding studies on NDDs and the usage of sequencing in clinical practice, our neuroCNVscore speeds up the screening on pathogenic CNVs, which facilitates the clinical diagnoses of CNVs with unknown significant, and thus may provide novel biological insights into NDDs.

Introduction

Neurodevelopmental disorders (NDDs) are characterized by the inability to achieve cognitive, emotional, and motor developmental milestones including autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD) and schizophrenia. It is estimated to affect over 11.3%, and 15% of the population in low and middle-income countries [1] and US, [2] respectively. NDD's heritability is high that has been estimated from twin and family studies as 50% to 90% in ASD, [3] 88% in ADHD [4] and 85% in schizophrenia. [5] Genomic alterations are commonly found in children with NDDs. However, the explained genetic etiology of NDDs accounts for only a small proportion.

Copy number variants (CNVs) have been shown to be important for NDD genetic etiology. [6, 7] However, systematic identification of CNV pathogenicity by virtue of their number, size and impact on the genome is still a challenge. It is approximately 1,000 CNVs per genome ranging in size from 50 base pairs (bp) to several mega bases (Mb). CNVs, by definition, result in gain or loss of DNA segments (copy number loss and copy number gain), that are exerted by altering the dosage of gene regions [8] as well as by disrupting non-coding areas, [7, 9] which requires various genomic assays in both tissue-specific and non-specific manner to dissect. Growing number of studies by whole genome sequencing (WGS) and the complexity of identifying pathogenic CNVs make computational prediction an appropriate tool.

Many assessing tools have been developed to evaluate the pathogenicity of single nucleotide variants (SNVs), [10] [11] but fewer studies have systematically focused on assessing the pathogenic CNVs, especially none in NDD related CNVs. Recently, SVScore, [12] SVFX, [13] SVPath, [14] and AnnotSV [15] have been developed to interpret the SVs by integrating results from prediction matrices of SNPs, using cancer related SVs as inputs, counting SVs with overlapped exons, or integrating multiple sources to annotate SVs. However, the aggregated effects on SNPs, somatic impacts of SVs, or only overlapping exons without tissue-specific information may bias the effects of CNVs, and germline variations are the major focus in NDDs.

We here present a novel supervised machine learning framework, named as neuroCNVScore (https://github.com/lxsbch/neuroCNVscore), to score the pathogenicity of CNVs related to NDDs. We hypothesize that the computational prediction on pathogenic CNVs would benefit from a set of comprehensive tissue-specific features covering the whole genomic regions. Hence, we utilized cleaned germline CNVs from published NDD studies, [16-19] and gene lists together with a comprehensive set of neuro/brain-specific data on non-coding regions from ENCODE, [20] Roadmap, [21] EpiMap [22] and PsychENCODE [23] to train our models. Moreover, we constructed an NDD disease associated independent dataset using Human Phenotype Ontology (HPO) to validate trained models. The performance of neuroCNVScore was compared with a reference method SVScore. [12] This neuroCNVScore is designed for assessing the pathogenicity of CNVs in NDDs generated from association studies or clinical diagnoses.

Methods

Data collection and pre-processing/harmonization

The training set (identified by genomic coordinates) was gathered from several case-control based NDD studies. We assigned CNVs from cases as likely pathogenic (LP). In contrast, the CNVs from unaffected individuals and parents served as the control. Together, we collected 86,694 CNVs in the LP and 786,058 in the control set from four data sources, respectively (**Error! Reference source not found.. 1**).

Initial data filtering and harmonization were performed on all autosomal chromosome CNVs in three major steps. We first removed CNVs <50 bp and divided CNVs into copy number loss and copy number gain giving their potential impacts on the genome. Next, we deleted CNVs which had 90% reciprocal overlapped between LP and control. Finally, we applied an empirical cumulative distribution function with bin size of 60 to generate size matched LP and control to overcome the amount of disparity on CNVs. For each type, we sampled the same number of LP CNVs and matched the number of control CNVs in every bin. For training, we retained 13,857 cleaned LP CNVs and 13,859 cleaned control CNVs.

Next, we constructed the independent test set by assembling 51,819 disease associated variations from ClinVar and 136,181 common CNVs from GnomAD 2.1. For the NDD related set, we retained CNVs with length > 50 bp, germline, pathogenic, and the record of HPO: 0012759 (neurodevelopmental abnormality associated genes). For common CNVs, we kept CNVs with quality record PASS, and allele frequency > 0.1. To avoid over estimation, we removed those CNVs with 90% reciprocal overlap with the training dataset under the same variant type.

Finally, we collected several NDD related gene lists to test the biological validity and robustness of neuroCNVscore including CHD8 target genes, [24] human postsynaptic density (PSD) proteins [25] and ASD risk genes (FDR < 0.3). [18] The overall workflow is outlined in **Figure 1**.

This study has been approved by the Ethics Committee of Beijing Children's Hospital, Capital Medical University (2018-k-62).

A comprehensive tissue-specific feature collection and feature matrix construction

For each CNV, a broad range of features was compiled into a feature matrix. We leveraged 189 features in total from three different levels: (1) gene level (Gen), (2) functional/genomic segment level (Fun), and (3) sequence level (Seq). The description of features is shown in **Table S1**.

In brief, a set of gene level features (N = 62) that capture gene essentiality, dosage sensitivity and neurodevelopmental phenotype associated genes were collected. Since non-coding CNVs can disrupt regulatory regions affecting gene expression and translation in a linear or 3D manner, we obtained a regulatory cascade catalogue (N = 120 at functional/genomic segment level) by integrating multi-omics data covering experimentally identified or computational predicted regulatory regions at a tissue-specific manner. Lastly, the features at sequence level (N = 7) comprised of GC content, cross species conservation score (phylop46way and phastcon46way which are derived from phyloP or Hidden Markov Model via multiple alignment of 45 vertebrate genomes to the human genome), heterochromatin positions, collapsed repeat regions (DacMapExclude, DukeMapExclude are genomic regions calculated by different algorithms) retrieved from UCSC, and human accelerated regions accessed from Doan *et al.*.[26] These features could facilitate the identification of functional genomic regions and/or filter the genomic regions which may cause artefacts by downstream segments.

Based on various features, annotations were performed in three distinct ways: (1) sum up the number of overlapped features with a given CNV, (2) a discrete value that denotes the number of the features which has >50% reciprocal overlapped regions with a given CNV, (3) average value of overlapped regions between the feature and a given CNV. After initial annotation, we divided the entire feature matrix into length of each CNV and then applied min-max scaling. Considering the differences in features, e.g. triplosensitivity is a measurement only for the copy number gain, we kept 172 features out of 189 for the copy number loss model and 172 features in the copy number gain model, respectively.

Design of XGBoost model and the training strategy

To choose an appropriate model, we compared the performances among different algorithms (Naïve Bayes, Logistic Regression, Support Vector Machine, and XGBoost), and found XGBoost had the best performance in the python framework from Scikit 0.22.1 with the binary logistic objective function. A total of 80%/20% of the variant sets was used as training/test sets, respectively. Next, we trained the XGBoost model with optimized parameters by using grid search and evaluated our models through an independent test set. Additionally, we assessed the performance by comparing our model with SVScore.

Statistics

Statistical analyses were performed using Python (version 2.7). The performance was measured by precision-recall (PR) and receiver operating characteristic (ROC) curves. For individual feature comparison, we applied two-tailed Wilcoxon rank-sum tests. All genomic data is in GRCh37 genome build. Figures were generated by the ggplot package in R (version 3.6.1) or matplotlib in Python.

Patient and public involvement

Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research. No ethical issues are involved in this study as this paper only used the data deposited in the public accessible databases.

Results

Individual feature analyses highlight the importance to collect a comprehensive feature set

To understand the characteristics of CNVs in NDDs, we investigated distribution of features between LP and control sets. In total, we observed 121 and 106 significant features at the threshold of $P = 0.05$ in copy number loss and copy number gain models, respectively (**Table S2**). This demonstrated a large spectrum of features showing significant differences between sets, and an integrated feature framework prone to the pathogenic status of CNVs that were functionally relevant.

Among these significant features, functional/genomic segment features ranked higher than the others. Most of the highly ranked features were related to histone modification markers (e.g. H3K27me3, H3K27ac) and 3D chromatin related features (e.g. enhancers) (**Figure 2**). This is expected since noncoding regions account for 98% of the human genome and CNVs can affect the genome by interrupting the regulatory regions.

Comparisons among four algorithms show that XGBoost outperforms others

To find an optimal model for discriminating pathogenic CNVs, we evaluated the predictive performance of Naïve Bayes, Logistic Regression, Support Vector Machine (SVM) and XGBoost on the test sets (**Figure 3**). XGBoost model performed the best (average precision (AP) and area under curve (AUC) were 0.82, 0.85 for copy number loss; AP and AUC were 0.80, 0.84 for copy number gain). Therefore, we applied the XGBoost to construct the neuroScoreCNV.

Accuracy assessments reveal better performance of neuroScoreCNV

We evaluated the performance of neuroScoreCNV and SVScore by the independent set as described in the flowchart (**Fig. 1**). neuroScoreCNV achieved relatively higher performance compared to SVScore (**Figure 4B, D**). For two different types of models, we observed AP = 0.88, AUC = 0.93 at copy number loss (**Figure 4A, B**, orange line), and AP = 0.68, AUC = 0.67 at copy number gain model (**Figure 4C, D**, orange line). The different performances between models are in agreement with a previous study. [13]

Moreover, we investigated the biological validity and robustness from two aspects. It was shown interruptions at conserved regions could cause diseases since these regions are normally functional. [27] Therefore, we first computed the CNV pathogenic scores generated with the new feature matrices in which a conservation score (i.e. PhyloP46way, one of the commonly used conservation score that considering individual base conservation) was excluded. We observed higher CNV pathogenic scores (≥ 0.7) tended to have higher conservation scores after correlating log$_{10}$(PhyloP46way) and new pathogenic scores (**Figure 5A, B**). Then, we checked if our predicted scores were capable of prioritizing CNVs with known NDD associated genes. LP CNVs covered significantly (*P* < 0.05) more NDD related genes than the control group (**Figure 5B**). Overall, our approach achieved higher performance in discriminating LP CNVs from control or benign CNVs.

Feature importancy highlights the important role of regulatory regions in NDDs

We computed the feature importancy by permutation. We categorized model features into three groups: functional/genomic level (Fun), gene level (Gen) and sequence level (Seq) (**Figure 6**, **Table S3**). The most important features were genes with haploinsufficiency scores (PHI) and triplosensitivity scores (PTS). PHI reflects the probability of one single functional copy to be sufficient to maintain function, whereas PTS suggests the probability of an additional copy of a gene for generating phenotypes. PHI and PTS are important parameters for evaluating the pathogenicity in clinical diagnoses based on the ACMG guidelines. [28] This is also true in neuroCNVScore. In NDDs, several studies found pathogenic CNVs were sensitive to dosage. [29]

Additionally, we noticed several phenotypes were prominent such as HPO: 000717 (autism associated genes), HPO: 0002960 (autoimmunity associated genes) and HPO: 0025031 (abnormality of the digestive system associated genes). It is known that immune system abnormalities and/or gastrointestinal symptoms can co-occur with

ASD [30] and schizophrenia. [31] Compelling evidence demonstrated autoimmune response was important in ASD. [32] Purified IgG containing antibodies from the mothers of children with ASD can cause abnormal behaviours in animal models. [33, 34]

Among important features at the functional/genomic segment level, we observed several key players in 3D chromatin conformation including enhancers and TADs. Meanwhile, DNase-Seq which suggests active regulatory elements at open chromatin was also an important feature. The emerging evidence has highlighted the role of 3D chromatin conformation in relation to NDDs. [23, 35] Collectively, studying the interaction between CNVs and the higher order of chromatin conformation could provide novel insights into the etiology of NDDs and explain the missing heredity of NDDs.

Discussion

In this work, we introduced a novel framework, neuroCNVscore, to ascertain the pathogenicity of CNVs in NDDs. NeuroCNVscore outperformed a commonly used tool SVScore on independent datasets from ClinVar and gnomAD. Importantly, neuroCNVscore has unique ability to prioritize the functional, deleterious and pathogenic CNVs derived from either NDD's association studies or clinical diagnoses, which may provide biological new insights into NDDs, especially at the three-dimensional genome level.

There are several factors contribute to the accuracy and robustness of neuroCNVscore. First, we used a high-quality set of germline CNVs from published NDD studies as the training set, which assures that our model is of high quality. Secondly, we validated our models at an NDD associated independent dataset and outperformed a published tool, SVScore. Furthermore, we created a comprehensive feature collection (N = 189) at gene, functional genomic, and sequence levels. Specifically, we incorporated a significant amount of tissue-specific functional genomic data. As a result, we can not only identify the genes disrupted by CNVs, but also the disrupted regulatory elements that act in a tissue-specific manner during development. This is especially important for the studies in NDD since brain tissue is normally hard to access.

While the neuroCNVscore performed well, it may be improved by incorporating expert-curated CNVs from whole genome sequencing studies in NDDs and healthy controls. Along with the increased knowledge and functional genomics data on non-coding regions, additional informative features can be integrated into the model to better address the hidden mechanisms. Moreover, we developed neuroCNVscore based on XGBoost, but it is worth exploring deep learning algorithms in the future.

Together, our neuroCNVscore performed well and is a useful tool for generating hypotheses in genome wide association studies in NDDs and could facilitate the understanding of genetic etiology of NDDs.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

XL designed the study, performed the analysis and drafted the manuscript. WX and FL participated in the design and interpretation of the data and revised the manuscript. PZ, RG and YZ participated in the interpretation of data. CH coordinated the project and supervised the study. XN coordinated the project and acquisition the funding. WL coordinated the project, supervised the study, critically reviewed and revised the manuscript. All authors read and approved the final manuscript.

Availability of Data and Materials

All features analysed during this study are collected from public datasets. Sources can be found from https://github.com/macarthur-lab/gene_lists. All CNV training data are included in these publications [16-19] and testing data are from the ClinVar database. The source code is available at https://github.com/lxsbch/neuroCNVscore.

Acknowledgements

We thank MacArthur's Lab for sharing the comprehensive collections of gene lists. We thank Dr. Sree Rohit Raj Kolora for reviewing, revising the manuscript and useful discussion.

References

1. Bitta M, Kariuki SM, Abubakar A, et al. Burden of neurodevelopmental disorders in low and middle-income countries: A systematic review and meta-analysis. *Wellcome Open Res* 2017;2:121. doi: 10.12688/wellcomeopenres.13540.3

2. America's Children and the Environment. Health: Neurodevelopmental Disorders – Report Contents, 2019.

3. Gaugler T, Klei L, Sanders SJ, et al. Most genetic risk for autism resides with common variation. *Nat Genet* 2014;46:881-5. doi: 10.1038/ng.3039

4. Larsson H, Chang Z, D'Onofrio BM, et al. The heritability of clinically diagnosed attention deficit hyperactivity disorder across the lifespan. *Psychol Med* 2014;44:2223-9. doi: 10.1017/S0033291713002493

5. Cardno AG, Marshall EJ, Coid B, et al. Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. *Arch Gen Psychiatry* 1999;56:162-8. doi: 10.1001/archpsyc.56.2.162

6. Marshall CR, Howrigan DP, Merico D, et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat Genet*

2017;49:27-35. doi: 10.1038/ng.3725

7. Brandler WM, Antaki D, Gujral M, et al. Paternally inherited cis-regulatory structural variants are associated with autism. *Science* 2018;360:327-31. doi: 10.1126/science.aan2261

8. Coe BP, Stessman HAF, Sulovari A, et al. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat Genet* 2019;51:106-16. doi: 10.1038/s41588-018-0288-4

9. Devanna P, Chen XS, Ho J, et al. Next-gen sequencing identifies non-coding variation disrupting miRNA-binding sites in neurological disorders. *Mol Psychiatry* 2018;23:1375-84. doi: 10.1038/mp.2017.30

10. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7:248-9. doi: 10.1038/nmeth0410-248

11. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 2019;176:535-48 e24. doi: 10.1016/j.cell.2018.12.015

12. Ganel L, Abel HJ, FinMetSeq C, et al. SVScore: an impact prediction tool for structural variation. *Bioinformatics* 2017;33:1083-85. doi: 10.1093/bioinformatics/btw789

13. Kumar S, Harmanci A, Vytheeswaran J, et al. SVFX: a machine learning framework to quantify the pathogenicity of structural variants. *Genome Biol* 2020;21:274.

doi: 10.1186/s13059-020-02178-x

14. Yang Y, Wang X, Zhou D, et al. SVPath: an accurate pipeline for predicting the pathogenicity of human exon structural variants. *Brief Bioinform* 2022;23: bbac014 doi: 10.1093/bib/bbac014

15. Geoffroy V, Guignard T, Kress A, et al. AnnotSV and knotAnnotSV: a web server for human structural variations annotations, ranking and analysis. *Nucleic Acids Res* 2021;49:W21-W28. doi: 10.1093/nar/gkab402

16. Coe BP, Witherspoon K, Rosenfeld JA, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* 2014;46:1063-71. doi: 10.1038/ng.3092

17. Cooper GM, Coe BP, Girirajan S, et al. A copy number variation morbidity map of developmental delay. *Nat Genet* 2011;43:838-46. doi: 10.1038/ng.909

18. Sanders SJ, He X, Willsey AJ, et al. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* 2015;87:1215-33. doi: 10.1016/j.neuron.2015.09.016

19. Zarrei M, Burton CL, Engchuan W, et al. A large data resource of genomic copy number variation across neurodevelopmental disorders. *NPJ Genom Med* 2019;4:26. doi: 10.1038/s41525-019-0098-3

20. Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* 2018;46:D794-D801. doi: 10.1093/nar/gkx1081

21. Roadmap Epigenomics C, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518:317-30. doi: 10.1038/nature14248

22. Boix CA, James BT, Park YP, et al. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* 2021;590:300-07. doi: 10.1038/s41586-020-03145-z

23. Wang D, Liu S, Warrell J, et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* 2018;362:eaat8464. doi: 10.1126/science.aat8464

24. Sugathan A, Biagioli M, Golzio C, et al. CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proc Natl Acad Sci U.S.A.* 2014;111:E4468-77. doi: 10.1073/pnas.1405266111

25. Bayes A, van de Lagemaat LN, Collins MO, et al. Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nat Neurosci* 2011;14:19-21. doi: 10.1038/nn.2719

26. Doan RN, Bae BI, Cubelos B, et al. Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. *Cell* 2016;167:341-54 e12. doi: 10.1016/j.cell.2016.08.071

27. Kellis M, Wold B, Snyder MP, et al. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U.S.A.* 2014;111:6131-8. doi: 10.1073/pnas.1318948111

28. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405-24. doi: 10.1038/gim.2015.30

29. Han X, Chen S, Flynn E, et al. Distinct epigenomic patterns are associated with haploinsufficiency and predict risk genes of developmental disorders. *Nat Commun* 2018;9:2138. doi: 10.1038/s41467-018-04552-7

30. Hughes HK, Mills Ko E, Rose D, et al. Immune Dysfunction and Autoimmunity as Pathological Mechanisms in Autism Spectrum Disorders. *Front Cell Neurosci* 2018;12:405. doi: 10.3389/fncel.2018.00405

31. Severance EG, Prandovszky E, Castiglione J, et al. Gastroenterology issues in schizophrenia: why the gut matters. *Curr Psychiatry Rep* 2015;17:27. doi: 10.1007/s11920-015-0574-0

32. Wu S, Ding Y, Wu F, et al. Family history of autoimmune diseases is associated with an increased risk of autism in children: A systematic review and meta-analysis. *Neurosci Biobehav Rev* 2015;55:322-32. doi: 10.1016/j.neubiorev.2015.05.004

33. Bauman MD, Iosif AM, Ashwood P, et al. Maternal antibodies from mothers of children with autism alter brain growth and social behavior development in the rhesus monkey. *Transl Psychiatry* 2013;3:e278. doi: 10.1038/tp.2013.47

34. Hertz-Picciotto I, Croen LA, Hansen R, et al. The CHARGE study: an

epidemiologic investigation of genetic and environmental factors contributing to autism. *Environ Health Perspect* 2006;114:1119-25. doi: 10.1289/ehp.8483

35. Won H, de la Torre-Ubieta L, Stein JL, et al. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* 2016;538:523-27. doi: 10.1038/nature19847

Figure Legends

Figure 1. The flowchart of neuroCNVscore development and evaluation in this study. In Data Sets, the sources of training set and test set are listed. The training set was derived from four NDDs studies under the case-control design, while the validation set was from ClinVar and GnomAD. The numbers of raw and cleaned CNVs in the brackets are indicated. LP, likely pathogenic. In Neuro-features, comprehensive neuro/brain related features were gathered at gene, sequence, and functional/genomic segments levels. In Prediction and Validation, biological validations were performed in two ways: 1) correlations between phyloP46way and the pathogenic scores generated by the new model where phyloP46way was excluded from the feature matrix; 2) using the independent set of NDD related gene lists including PSD genes to cognition, CHD8 targets, and ASD risk genes.

Figure 2. Comparisons of top three features between control and LP (likely pathogenic)

sets. The top three significant features between control and LP sets in copy number loss (A) and copy number gain (B). The X-axis shows the significant feature types. Fun_level, Function/genomic segment level. The Y-axis is the value of log transformed feature matrices. Unpaired *t*-tests were applied and significant levels were. **** *P* < 0.0001.

Figure 3. Performances on CNVs among Naïve Bayes, Logistic Regression, Support Vector Machine (SVM) and XGBoost algorithms. XGBoost showed the best performance by precision-recall curve and ROC curve for both copy number loss (A, B) and copy number gain (C, D). AP: average precision; AUC: area under curve.

Figure 4. Performances on neuroCNVscore and SVScore in the independent set as described in the flowchart of Figure 1. Precision-Recall (A) and ROC (B) curves calculated with copy number loss from the independent dataset; Precision-Recall (C) and ROC (D) curves calculated with copy number gain from the independent dataset.

Figure 5. Biological validation of neuroCNVscore. The plot (A) shows the comparisons between PhyloP scores (log10(PhyloP46way)) and pathogenic scores generated by excluding PhyloP46way from the original neuroCNVscore model, regions with higher pathogenic scores tend to have higher PhyloP scores. The number of NDD related genes (B) between predicted LP and control groups in both copy number loss

and copy number gain models shows that more NDD related genes are found in LP. To present the figures in a clearer way, PhyloP46way and count were log-transformed. **P* < 0.05.

Figure 6. Top 20 features from feature importance analyses. Highly important features of copy number loss model (A) and copy number gain model (B) are listed. All the feature names were colored and formatted as following: feature type (Fun_/Gen_ /Seq_feature names (original sources)_tissue type (if applicable). Fun: Function, in blue; Gen: Gene, in green; Seq: Sequence, in purple.

Supplementary Tables

Table S1. A detailed feature description.

Table S2. Individual feature comparisons.

Table S3. Feature importancy.

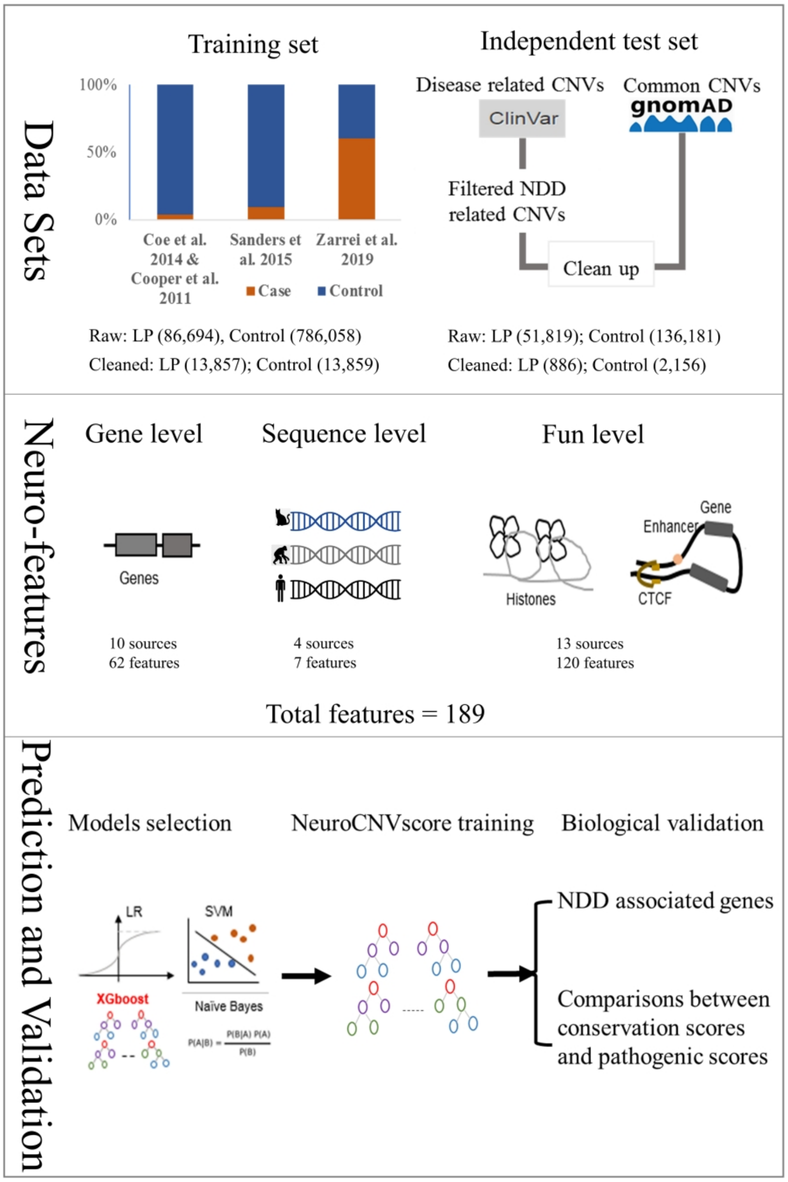

Figure 1-flowchart of data analysis

140x202mm (600 x 600 DPI)

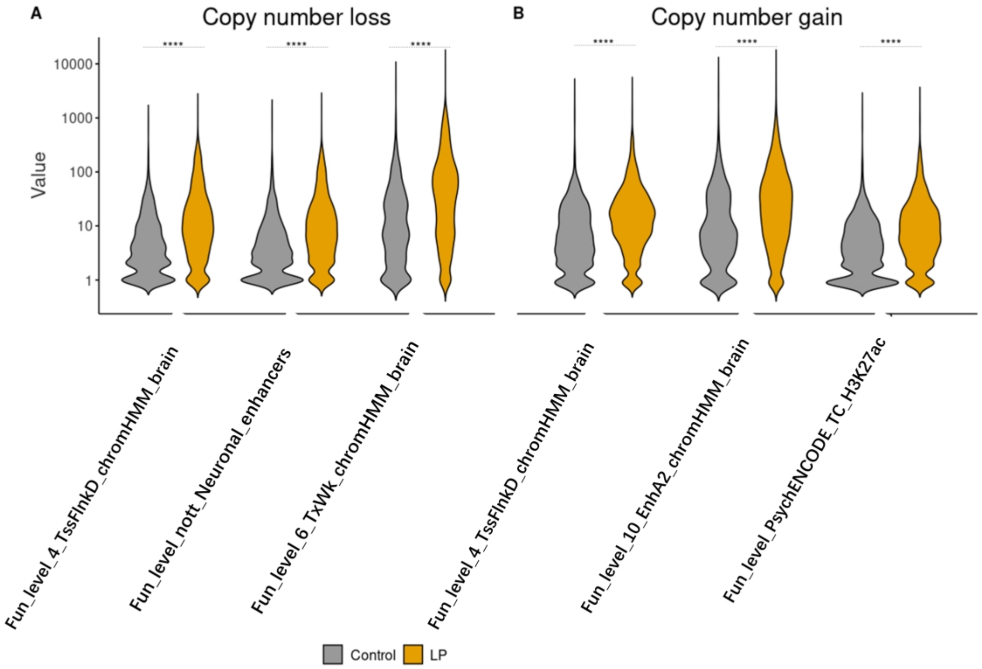

Figure 2-Top features

198x137mm (600 x 600 DPI)

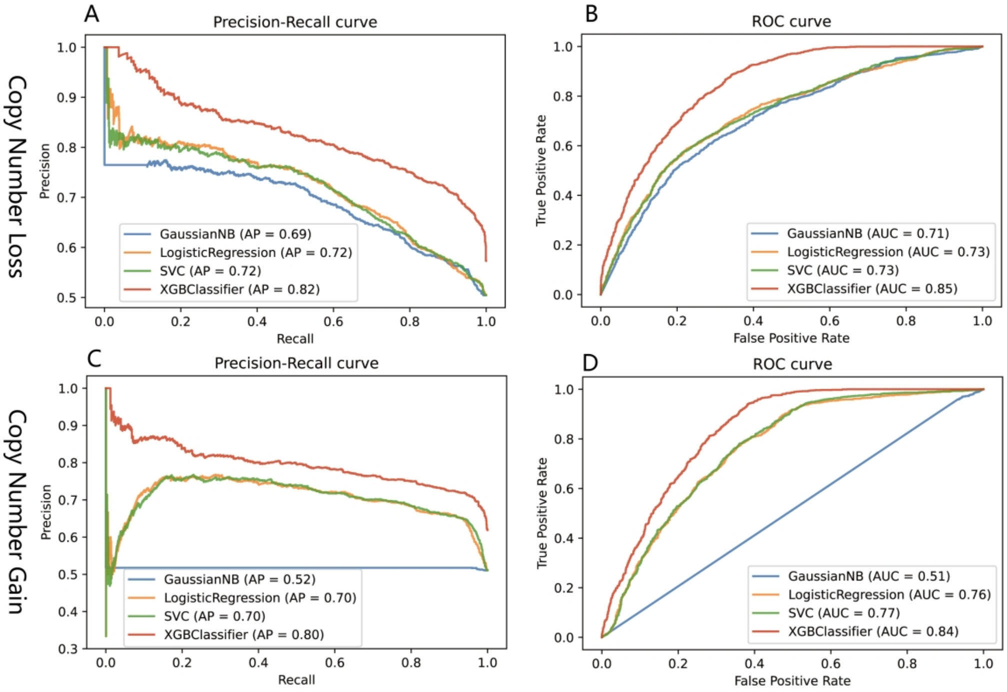

Figure 3-XGBooster performance

252x172mm (300 x 300 DPI)

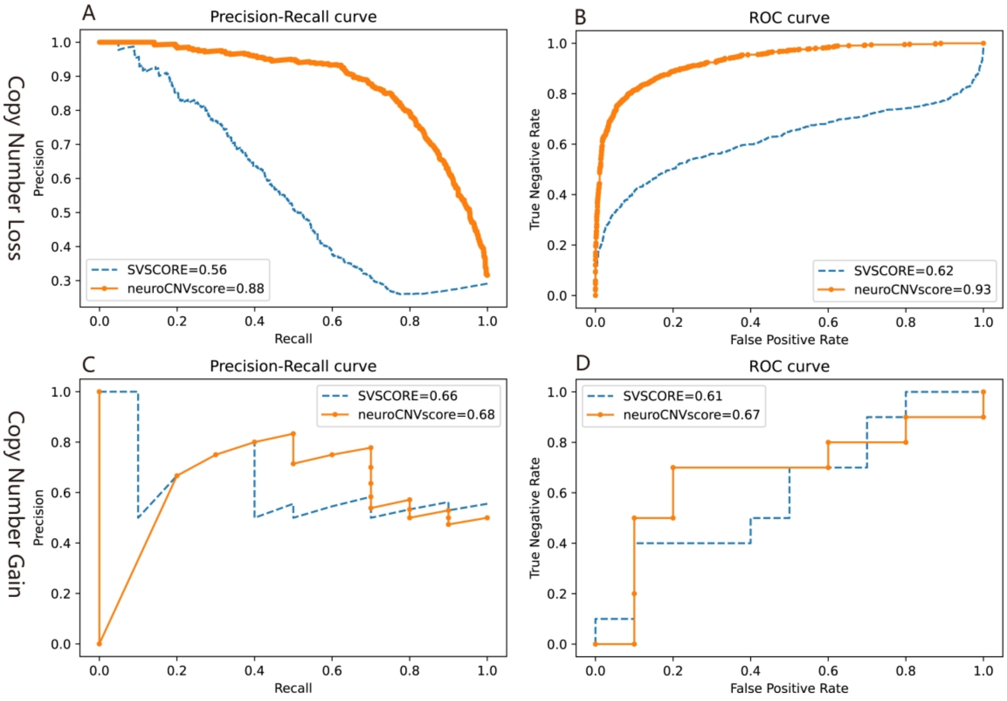

Figure 4-neuroCNV vs. SV

276x190mm (300 x 300 DPI)

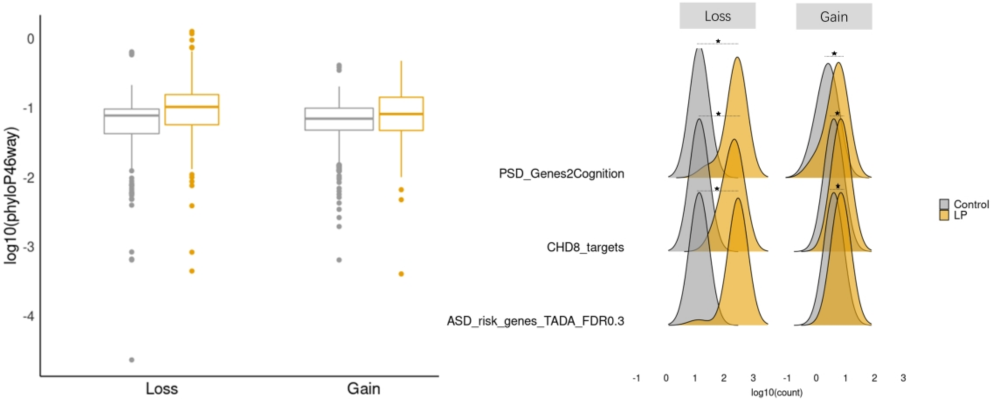

Figure 5-biological validation

293x122mm (600 x 600 DPI)

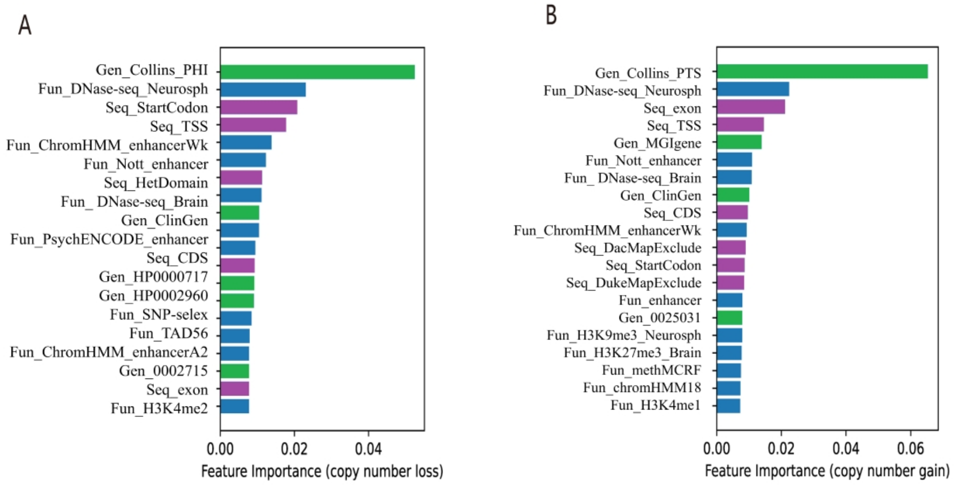

Figure 6-feature importancy

2040x1120mm (96 x 96 DPI)

Supplementary Tables

Table S1. A detailed feature description. This table includes all features used in our model. These features are grouped into three levels: gene, functional/genomic segment and sequence. A brief description along with references is described on each feature.

Feature category	**Feature set**	**Description**	**Feature type**	**References**
Gene level (N = 61)	Cell essential and nonessential genes	CRISPR/Cas9 screens identified essential genes in human cell lines. Curated in[2]	discrete	[1]
	ClinGen curated genes and genomic regions	Genes and genomics regions were rated from 0 to 3, indicating an increased evidence on dosage sensitivity. Additional two levels (40,30) suggest unlikely dosage sensitive and genes associated with autosomal recessive phenotype.	discrete	[3]
	DDG2P database	A curated list of genes linked to developmental disorders compiled by clinicians as part of the DDD study to facilitate clinical feedback on likely causal variants	discrete	[4]
	Dosage sensitive genes	Predicted score on dosage sensitive genes (i.e., haploinsufficiency or triplosensitivity)	discrete	[5, 34]
	FDA proved drug target	Genes with protein products that are mechanistic targets of FDA-approved drugs. Curated in [2]	discrete	[6]
	G protein-coupled receptor	GPCR list curated in[2]	discrete	[32, 33, 35]
	Mouse heterozygous LoF lethal	Genes that are lethal in mouse models when inactivated heterozygous. Curated by [2]	discrete	[7]
	Neurodevelopmental process related genes	Genes associated with various phenotypes from HPO: Abnormality of the nervous system (HP:0000707)-associated genes Abnormality of nervous system physiology (HP:0012638)-associated	discrete	[8]

		genes Behavioral abnormality (HP:0000708)-associated genes Abnormality of nervous system morphology (HP:0012639)-associated genes Abnormality of the immune system (HP:0002715)-associated genes Neurodevelopmental abnormality (HP:0012759)-associated genes Autoimmunity (HP:0002960)-associated genes Morphological abnormality of the central nervous system (HP:0002011)-associated genes Schizophrenia (HP:0100753)-associated genes Autistic behavior (HP:0000729)-associated genes Abnormality of movement (HP:0100022)-associated genes Seizures (HP:0001250)-associated genes Autism (HP:0000717)-associated genes Hyperactivity (HP:0000752)-associated genes Abnormality of prenatal development or birth (HP:0001197)-associated genes Impairment in personality functioning (HP:0031466)-associated genes Abnormality of the digestive system (HP:0025031)-associated genes Growth abnormality (HP:0001507)-associated genes Abnormal fear/anxiety-related behavior (HP:0100852)-associated genes Abnormality of brain morphology (HP:0012443)-associated genes Abnormality of higher mental function (HP:0011446)-associated genes		
	Olfactory receptors	Any HUGO-recognized family of olfactory receptor genes	discrete	9
	SFARI gene	Genes implicated in autism susceptibility	discrete	10

Functional/genomic segment level (N = 121)	Chromatin states	Brain related chromatin states inferred by the extended 18-way ChromHMM model across 98 tissues from the Roadmap Epigenomics Project	discrete	11
	CTCF binding sites	Genome wide observed CTCF binding sites from Brain	continuous	12
		Genome wide CTCF binding sites from 7 cell lines generated by ChIP-seq. Curated by UCSC	continuous	13
	DNA Accessibility	ATAC-seq from brain and neurosph.	continuous	13
	DNase hypersensitive sites	Observed DNase I hypersensitive areas from brain and neurosph.	continuous	13
		DNase hypersensitive sites assayed from a collection of cell types. Download from UCSC table browser NAR 2004	continuous	14
		RoadmapDNasePromCount	discrete	15
	Enhancers	Brain cell type-specific enhancers identified by PLAC-seq	discrete	16
		dbSUPER: Super enhancers from Brain Angular Gyrus; Brain Anterior Caudate; Brain Cingulate Gyrus; Brain Hippocampus Middle; Brain Inferior Temporal Lobe	discrete	17
		EpiMap: enhancers from the brain and neurosph.	discrete	12
		EnhancerAtlas 2.0: Enhancer predictions in 197 human cell lines & tissues	discrete	18
		FANTOM Enhancers: Enhancer predictions for human tissues and cell types from the FANTOM5 consortium	discrete	19
		HACER: Active enhancer predictions in human cell lines & tissues based on PRO-seq, GRO-seq, or CAGE data	discrete	20
		PsychENCODE: PEC EnhancersDER-03a_hg19_PEC_enhancers_clean.bed	discrete	21
		SEA: Super enhancer predictions from 143 human cell lines and tissues (mapped back to hg19 using liftOver with minimum 75% match)	discrete	23

		Sedb: Super enhancer and typical enhancer predictions from 541 human cell lines and tissues	discrete	22
		VISTA: Experimentally-validated mammalian enhancers	discrete	24
	Genomic segmentations	All autosomal, protein-coding genes; CDS; exon; Selenocysteine; start_codon; stop_codon; transcript UTR	discrete	25
	Histone markers	H2AFZ, H2AK5ac, H2AK9ac, H2BK120ac, H2BK12ac, H2BK15ac, H2BK20ac, H2BK5ac, H3F3A, H3K27ac, H3K4ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me1, H3K9me2, H3K9me3 from the brain or neurosph	continuous	12
		H3K27ac peaks for the Prefrontal Cortex, the Temporal Cortex, and the Cerebellar Cortex	continuous	21
	Long range probable genes	Target genes by prediction on GWAS hits and 3D chromatin structures	discrete	26
	Loop anchors and topological associated domains in higher-order chromatin structure	TAD boundaries (defined as the start and end coordinates for each TAD ± 5kb) from 30 samples meeting our ENCODE data inclusion criteria available for download from the ENCODE Data Portal	continuous	14
		Selected "derived" datasets from PsychENCODE Integrated Analysis Package, including cortex enhancers, transcriptionally active regions, TAD boundaries, and H3k27ac peaks	continuous	21
		Yue labs	continuous	27
	Methylation	MeDIP/MRE (mCRF) methylation calls	continuous	15
	Transcript active regions	Cortex Transcriptionally Active Regions are found within at least 70% of the individuals	continuous	21
	Transcript factor binding sites	SNP-SELEX	discrete	28
	Transcript starting sites	The 2000bp flanking regions about transcript starting sites	discrete	36

Sequence level (N = 7)	Blacklisted regions	Genome regions have anomalous, unstructured, high signal/read counts (DacMapExclude), problematic regions for short sequence tag signal detection (DukeMapExclude)	discrete	[29]
	Cross species conservation score	The conservation scoring (phylop46way, phastcon46way) for multiple alignments of 45 vertebrate genomes to the human genome	continuous	[29]
	GC content	GC content calculated with a "span" size of 5 bases	continuous	[29]
	Heterochromatin positions	It is calculated based on H3K9me3 enrichment regions	discrete	[30]
	Human accelerated regions	Human accelerated regions are conserved genomic loci with elevated divergence in humans	discrete	[31]

Table S2. Individual feature comparisons. This table compares all of the features used in the copy number loss and copy number gain models. The comparisons were made using the two-tailed Wilcoxon rank-sum test, with a significant cut off of *P* = 0.05. All the feature names were reformatted as followed: feature type (Fun_level/Gen_level/Seq_level)_feature names(original sources)_tissue type (if applicable). Fun: Function; Gen: Gene; Seq: Sequence.

Source	**Features in copy number loss model**	***P* value**	**Source**	**Features in copy number gain model**	***P* value**
Functional/ genomic segment level	*Fun_level_significant features are 80 out of 120*		**Functional/ genomic segment level**	*Fun_level_significant features are 75 out of 120*	
Chromatin states from Roadmap Epigenomics Project	Fun_level_1_TssA_chromHMM_brain	6.63E-124	Chromatin states from Roadmap Epigenomics Project	Fun_level_1_TssA_chromHMM_brain	3.09E-194
	Fun_level_10_EnhA2_chromHMM_brain	2.51E-232		Fun_level_10_EnhA2_chromHMM_brain	2.15E-288
	Fun_level_11_EnhWk_chromHMM_brain	5.72E-305		Fun_level_11_EnhWk_chromHMM_brain	~0
	Fun_level_12_ZNF_chromHMM_brain	1.51E-06		Fun_level_12_ZNF_chromHMM_brain	3.38E-04
	Fun_level_13_Het_chromHMM_brain	1.98E-08		Fun_level_13_Het_chromHMM_brain	1.03E-09
	Fun_level_14_TssBiv_chromHMM_brain	1.05E-01		Fun_level_14_TssBiv_chromHMM_brain	2.36E-05
	Fun_level_15_EnhBiv_chromHMM_brain	2.14E-05		Fun_level_15_EnhBiv_chromHMM_brain	1.46E-33
	Fun_level_16_ReprPC_chromHMM_brain	1.60E-08		Fun_level_16_ReprPC_chromHMM_brain	9.91E-01
	Fun_level_17_ReprPCWk_chromHMM_brain	4.98E-02		Fun_level_17_ReprPCWk_chromHMM_brain	5.47E-07
	Fun_level_18_Quies_chromHMM_brain	1.10E-42		Fun_level_18_Quies_chromHMM_brain	2.14E-96
	Fun_level_2_TssFlnk_chromHMM_brain	1.43E-91		Fun_level_2_TssFlnk_chromHMM_brain	1.95E-146
	Fun_level_3_TssFlnkU_chromHMM_brain	1.68E-157		Fun_level_3_TssFlnkU_chromHMM_brain	5.89E-253
	Fun_level_4_TssFlnkD_chromHMM_brain	3.15E-199		Fun_level_4_TssFlnkD_chromHMM_brain	2.89E-297
	Fun_level_5_Tx_chromHMM_brain	4.59E-133		Fun_level_5_Tx_chromHMM_brain	1.22E-198
	Fun_level_6_TxWk_chromHMM_brain	4.35E-290		Fun_level_6_TxWk_chromHMM_brain	~0
	Fun_level_7_EnhG1_chromHMM_brain	9.33E-114		Fun_level_7_EnhG1_chromHMM_brain	9.76E-164

	Fun_level_8_EnhG2_chromHMM_brain	2.69E-53		Fun_level_8_EnhG2_chromHMM_brain	9.04E-71
	Fun_level_9_EnhA1_chromHMM_brain	1.50E-188		Fun_level_9_EnhA1_chromHMM_brain	1.04E-253
Enhancers	Fun_level_dbsuper_Brain_Angular_Gyrus	4.27E-09	Enhancers	Fun_level_dbsuper_Brain_Angular_Gyrus	3.51E-07
	Fun_level_dbsuper_Brain_Anterior_Caudate	3.54E-14		Fun_level_dbsuper_Brain_Anterior_Caudate	3.70E-09
	Fun_level_dbsuper_Brain_Cingulate_Gyrus	6.08E-15		Fun_level_dbsuper_Brain_Cingulate_Gyrus	8.29E-15
	Fun_level_dbsuper_Brain_Hippocampus_Middle_150	9.98E-15		Fun_level_dbsuper_Brain_Hippocampus_Middle_150	7.05E-11
	Fun_level_dbsuper _Brain_Hippocampus_Middle	3.29E-19		Fun_level_dbsuper _Brain_Hippocampus_Middle	9.35E-14
	Fun_level_dbsuper_Brain_Inferior_Temporal_Lobe	1.30E-17		Fun_level_dbsuper_Brain_Inferior_Temporal_Lobe	1.62E-14
	Fun_level_dbsuper_Brain_Mid_Frontal_Lobe	2.93E-02		Fun_level_dbsuper_Brain_Mid_Frontal_Lobe	2.66E-02
	Fun_level_famton_astrocyte	3.96E-04		Fun_level_famton_astrocyte	2.30E-02
	Fun_level_famton_brain	4.89E-01		Fun_level_famton_brain	6.62E-01
	Fun_level_famton_CL:0000127	9.29E-05		Fun_level_famton_CL:0000127	8.25E-06
	Fun_level_famton_count	1.29E-170		Fun_level_famton_count	1.99E-252
	Fun_level_famton_neuronal_stem_cell	3.28E-01		Fun_level_famton_neuronal_stem_cell	6.99E-01
	Fun_level_famton_permssive	2.26E-129		Fun_level_famton_permssive	3.25E-147
	Fun_level_enhancerAtlas_Astrocyte_EP	3.90E-40		Fun_level_enhancerAtlas_Astrocyte_EP	1.32E-89
	Fun_level_enhancerAtlas_Cerebellum_EP	9.56E-61		Fun_level_enhancerAtlas_Cerebellum_EP	5.46E-104
	Fun_level_enhancerAtlas_ESC_neuron_EP	3.51E-25		Fun_level_enhancerAtlas_ESC_neuron_EP	2.70E-37
	Fun_level_gene_enhancer_links_brain_enhcenter	2.04E-277		Fun_level_gene_enhancer_links_brain_enhcenter	~0
	Fun_level_gene_enhancer_links_neurosph_enhcenter	6.76E-239		Fun_level_gene_enhancer_links_neurosph_enhcenter	~0
	Fun_level_hacer_T1	1.24E-73		Fun_level_hacer_T1	5.71E-136
	Fun_level_SE_ele	2.61E-39		Fun_level_SE_ele	2.03E-79

	Fun_level_SEA00101	1.16E-43		Fun_level_SEA00101	6.05E-66
	Fun_level_nott_Astrocyte_enhancers	1.69E-155		Fun_level_nott_Astrocyte_enhancers	1.75E-222
	Fun_level_nott_Astrocyte_promoters	1.42E-87		Fun_level_nott_Astrocyte_promoters	3.97E-149
	Fun_level_nott_H3K4me3_around_TSS	6.28E-84		Fun_level_nott_H3K4me3_around_TSS	1.46E-139
	Fun_level_nott_Microglia_enhancers	1.03E-79		Fun_level_nott_Microglia_enhancers	4.05E-123
	Fun_level_nott_Microglia_promoters	5.38E-67		Fun_level_nott_Microglia_promoters	3.48E-125
	Fun_level_nott_Neuronal_enhancers	1.52E-301		Fun_level_nott_Neuronal_enhancers	~0
	Fun_level_nott_Neuronal_promoters	3.73E-87		Fun_level_nott_Neuronal_promoters	1.89E-137
	Fun_level_nott_Oligo_enhancers	6.43E-164		Fun_level_nott_Oligo_enhancers	7.10E-221
	Fun_level_nott_Oligo_promoters	1.42E-92		Fun_level_nott_Oligo_promoters	2.70E-151
	Fun_level_nott_superEnhancer	1.00E+00		Fun_level_nott_superEnhancer	1.00E+00
	Fun_level_vista	9.93E-07		Fun_level_vista	9.38E-07
CTCF binding sites	Fun_level_ctcf	2.25E-65	CTCF binding sites	Fun_level_ctcf	6.96E-112
	Fun_level_CTCF_observed_Brain	~0		Fun_level_CTCF_observed_Brain	~0
DNase hypersensitive sites	Fun_level_DNaselClusterd	1.82E-49	DNase hypersensitive sites	Fun_level_DNaselClusterd	1.69E-61
	Fun_level_DnaseMaster	1.49E-73		Fun_level_DnaseMaster	2.90E-96
	Fun_level_DNase-seq_observed_Brain	~0		Fun_level_DNase-seq_observed_Brain	~0
	Fun_level_DNase-seq_observed_Neurosph	~0		Fun_level_DNase-seq_observed_Neurosph	~0
Genomic segmentations from Gencode	Fun_level_EncodeAwgTfbsBroadNhaCtcf	3.30E-76	Genomic segmentations from Gencode	Fun_level_EncodeAwgTfbsBroadNhaCtcf	4.70E-150
	Fun_level_EncodeRegTfbsClustered	2.01E-45		Fun_level_EncodeRegTfbsClustered	3.01E-147
	Fun_level_gencode_CDS	3.68E-17		Fun_level_gencode_CDS	4.60E-62
	Fun_level_gencode_exon	5.44E-01		Fun_level_gencode_exon	5.12E-07
	Fun_level_gencode_gene	2.45E-22		Fun_level_gencode_gene	1.60E-26
	Fun_level_gencode_Selenocysteine	5.45E-01		Fun_level_gencode_Selenocysteine	6.28E-01
	Fun_level_gencode_start_codon	2.45E-01		Fun_level_gencode_start_codon	6.60E-19

	Fun_level_gencode_stop_codon	7.36E-06		Fun_level_gencode_stop_codon	1.77E-25
	Fun_level_gencode_transcript	8.59E-01		Fun_level_gencode_transcript	8.33E-01
	Fun_level_gencode_UTR	2.94E-12		Fun_level_gencode_UTR	3.42E-41
	Fun_level_miRNA	1.84E-125		Fun_level_miRNA	5.54E-245
	Fun_level_non-codingRNAs	1.24E-15		Fun_level_non-codingRNAs	1.00E-03
Histone markers	Fun_level_ATAC-seq_observed_Brain	~0	Histone markers	Fun_level_ATAC-seq_observed_Brain	~0
	Fun_level_H2AFZ_imputed_Brain	~0		Fun_level_H2AFZ_imputed_Brain	~0
	Fun_level_EP300_imputed_Brain	~0		Fun_level_EP300_imputed_Brain	~0
	Fun_level_EP300_imputed_Neurosph	~0		Fun_level_EP300_imputed_Neurosph	~0
	Fun_level_H2AFZ_imputed_Neurosph	~0		Fun_level_H2AFZ_imputed_Neurosph	~0
	Fun_level_H2AFZ_observed_Brain	~0		Fun_level_H2AFZ_observed_Brain	~0
	Fun_level_H3k27ac	7.28E-91		Fun_level_H3k27ac	6.19E-90
	Fun_level_H3K27ac_imputed_Brain	~0		Fun_level_H3K27ac_imputed_Brain	~0
	Fun_level_H3K27ac_imputed_Neurosph	~0		Fun_level_H3K27ac_imputed_Neurosph	~0
	Fun_level_H3K27ac_observed_Brain	~0		Fun_level_H3K27ac_observed_Brain	~0
	Fun_level_H3K27ac_observed_Neurosph	~0		Fun_level_H3K27ac_observed_Neurosph	~0
	Fun_level_H3K27me3_imputed_Brain	5.90E-280		Fun_level_H3K27me3_imputed_Brain	~0
	Fun_level_H3K27me3_imputed_Neurosph	7.54E-278		Fun_level_H3K27me3_imputed_Neurosph	~0
	Fun_level_H3K27me3_observed_Brain	5.33E-109		Fun_level_H3K27me3_observed_Brain	1.41E-239
	Fun_level_H3k4me1	1.12E-95		Fun_level_H3k4me1	1.55E-83
	Fun_level_H3K4me1_imputed_Brain	~0		Fun_level_H3K4me1_imputed_Brain	~0
	Fun_level_H3K4me1_imputed_Neurosph	~0		Fun_level_H3K4me1_imputed_Neurosph	~0
	Fun_level_H3K4me1_observed_Brain	~0		Fun_level_H3K4me1_observed_Brain	~0
	Fun_level_H3K4me1_observed_Neurosph	~0		Fun_level_H3K4me1_observed_Neurosph	~0
	Fun_level_H3K4me2_observed_Brain	~0		Fun_level_H3K4me2_observed_Brain	~0

	Fun_level_H3k4me3	2.24E-26		Fun_level_H3k4me3	8.36E-17
	Fun_level_H3K4me3_imputed_Brain	5.80E-02		Fun_level_H3K4me3_imputed_Brain	4.68E-01
	Fun_level_H3K4me3_imputed_Neurosph	~0		Fun_level_H3K4me3_imputed_Neurosph	~0
	Fun_level_H3K4me3_observed_Brain	5.92E-02		Fun_level_H3K4me3_observed_Brain	4.71E-01
	Fun_level_H3K4me3_observed_Neurosph	~0		Fun_level_H3K4me3_observed_Neurosph	~0
	Fun_level_H3K9ac_imputed_Brain	~0		Fun_level_H3K9ac_imputed_Brain	~0
	Fun_level_H3K9ac_imputed_Neurosph	~0		Fun_level_H3K9ac_imputed_Neurosph	~0
	Fun_level_H3K9me3_imputed_Brain	3.25E-75		Fun_level_H3K9me3_imputed_Brain	9.60E-177
	Fun_level_H3K9me3_imputed_Neurosph	2.39E-122		Fun_level_H3K9me3_imputed_Neurosph	7.66E-231
	Fun_level_H3K9me3_observed_Brain	4.50E-61		Fun_level_H3K9me3_observed_Brain	3.22E-160
	Fun_level_H3K9me3_observed_Neurosph	1.74E-20		Fun_level_H3K9me3_observed_Neurosph	3.66E-53
	Fun_level_H4K20me1_imputed_Neurosph	~0		Fun_level_H4K20me1_imputed_Neurosph	~0
	Fun_level_H4K20me1_observed_Brain	~0		Fun_level_H4K20me1_observed_Brain	~0
	Fun_level_POLR2A_imputed_Neurosph	~0		Fun_level_POLR2A_imputed_Neurosph	~0
	Fun_level_RAD21_imputed_Brain	~0		Fun_level_RAD21_imputed_Brain	~0
	Fun_level_RAD21_imputed_Neurosph	~0		Fun_level_RAD21_imputed_Neurosph	~0
	Fun_level_SMC3_imputed_Brain	~0		Fun_level_SMC3_imputed_Brain	~0
	Fun_level_SMC3_imputed_Neurosph	~0		Fun_level_SMC3_imputed_Neurosph	~0
Long range probable genes	Fun_level_liu_csbj_targetgene	1.62E-22	Long range probable genes	Fun_level_liu_csbj_targetgene	1.14E-43
Methylation	Fun_level_methMCRF	1.06E-154	Methylation	Fun_level_methMCRF	1.46E-257
Loop anchors and topological	Fun_level_PsychENCODE_CBC_H3K27ac	8.71E-57	Loop anchors and topological	Fun_level_PsychENCODE_CBC_H3K27ac	4.13E-65
	Fun_level_PsychENCODE_HiC_EP	1.44E-53		Fun_level_PsychENCODE_HiC_EP	7.19E-84
	Fun_level_PsychENCODE_loops_interRegion	6.15E-07		Fun_level_PsychENCODE_loops_interRegion	1.89E-01

associated domains in higher-order chromatin structure	Fun_level_PsychENCODE_PEC_Enhancers	2.61E-158	associated domains in higher-order chromatin structure	Fun_level_PsychENCODE_PEC_Enhancers	3.36E-192
	Fun_level_PsychENCODE_PFC_H3K27ac	3.08E-152		Fun_level_PsychENCODE_PFC_H3K27ac	2.05E-184
	Fun_level_PsychENCODE_TAR	6.89E-41		Fun_level_PsychENCODE_TAR	1.40E-83
	Fun_level_PsychENCODE_TC_H3K27ac	5.58E-204		Fun_level_PsychENCODE_TC_H3K27ac	1.86E-265
	Fun_level_TAD56	7.60E-149		Fun_level_TAD56	1.18E-172
DNase hypersensitive sites	Fun_level_RoadmapDNasePromCount	7.33E-34	DNase hypersensitive sites	Fun_level_RoadmapDNasePromCount	7.61E-74
Transcript factor binding sites from snp-selex	Fun_level_snp_selex	8.60E-02	Transcript factor binding sites from snp-selex	Fun_level_snp_selex	4.03E-02
Transcript starting sites	Fun_level_tss2000bp	2.78E-06	Transcript starting sites	Fun_level_tss2000bp	1.45E-01
Higher-order chromatin structure from Yue lab	Fun_level_yue_loops_hippo	4.56E-133	Higher-order chromatin structure from Yue lab	Fun_level_yue_loops_hippo	1.26E-135
Gene level	*The Gen_level_significant features are 34 out of 45*		**Gene level**	*The Gen_level_significant features are 25 out of 45*	
ClinGen curated	Gen_level_ClinGen_haploinsufficiency_gene_0	1.29E-03	ClinGen curated	Gen_level_ClinGen_region_curation_Triplosensitivity_0	4.45E-01

genes and genomic regions	Gen_level_ClinGen_haploinsufficiency_gene_1	8.63E-03	genes and genomic regions	Gen_level_ClinGen_region_curation_Triplosensitivity_1	4.03E-01
	Gen_level_ClinGen_haploinsufficiency_gene_2	5.95E-01		Gen_level_ClinGen_region_curation_Triplosensitivity_2	6.03E-02
	Gen_level_ClinGen_haploinsufficiency_gene_3	6.19E-07		Gen_level_ClinGen_region_curation_Triplosensitivity_3	3.88E-01
	Gen_level_ClinGen_haploinsufficiency_gene_30	1.07E-12		Gen_level_ClinGen_region_curation_Triplosensitivity_40	4.33E-12
	Gen_level_ClinGen_haploinsufficiency_gene_40	7.01E-01		Gen_level_ClinGen_triplosensitivity_gene	7.80E-01
	Gen_level_ClinGen_region_curation_Haploinsufficiency_0	8.02E-01		Gen_level_ClinGen_triplosensitivity_gene_0	1.10E-30
	Gen_level_ClinGen_region_curation_Haploinsufficiency_1	8.83E-01		Gen_level_ClinGen_triplosensitivity_gene_1	9.07E-01
	Gen_level_ClinGen_region_curation_Haploinsufficiency_2	8.13E-01		Gen_level_ClinGen_triplosensitivity_gene_2	9.41E-01
	Gen_level_ClinGen_region_curation_Haploinsufficiency_3	1.09E-03		Gen_level_ClinGen_triplosensitivity_gene_3	1.00E+00
	Gen_level_ClinGen_region_curation_Haploinsufficiency_30	9.29E-01		Gen_level_ClinGen_triplosensitivity_gene_30	1.00E+00
	Gen_level_ClinGen_region_curation_Haploinsufficiency_40	2.03E-19		Gen_level_ClinGen_triplosensitivity_gene_40	1.00E+00
	Gen_level_loss_of_function_score1	2.61E-03		Gen_level_gain_activating_score1	8.03E-01
	Gen_level_loss_of_function_score2	2.72E-06		Gen_level_gain_activating_score2	6.81E-01
	Gen_level_loss_of_function_score3	9.89E-25		Gen_level_gain_activating_score3	2.84E-01

Dosage sensitive genes	Gen_level_Collins_rCNV_PLIgenes_PHI	0∞	Dosage sensitive genes	Gen_level_Collins_rCNV_PLIgenes_PTS	0∞
DDG2P database	Gen_level_ddg2p_loss	2.66E-55	DDG2P database	Gen_level_ddg2p_gain	6.90E-02
Cell essential and nonessential genes	Gen_level_Essential_in_culture_CRISPR	9.29E-05	Cell essential and nonessential genes	Gen_level_Essential_in_culture_CRISPR	6.25E-13
	Gen_nonEssential_in_culture_CRISPR	3.02E-01		Gen_nonEssential_in_culture_CRISPR	6.10E-01
FDA proved drug target	Gen_level_FDA-approved_drug_targets	1.30E-03	FDA proved drug target	Gen_level_FDA-approved_drug_targets	9.60E-06
G protein-coupled receptor	Gen _level_gpcr_union	4.01E-01	G protein-coupled receptor	Gen_level_gpcr_union	2.83E-04
Neurodevelopmental process related genes	Gen_level_HP_0000707	4.37E-67	Neurodevelopmental process related genes	Gen_level_HP_0000707	2.00E-76
	Gen_level_HP_0000708	1.53E-35		Gen_level_HP_0000708	1.08E-34
	Gen_level_HP_0000717	7.64E-04		Gen_level_HP_0000717	3.17E-03
	Gen_level_HP_0000729	2.70E-11		Gen_level_HP_0000729	4.92E-07
	Gen_level_HP_0000752	8.38E-09		Gen_level_HP_0000752	1.85E-05
	Gen_level_HP_0001197	3.76E-08		Gen_level_HP_0001197	1.66E-12
	Gen_level_HP_0001250	1.96E-33		Gen_level_HP_0001250	3.00E-31
	Gen_level_HP_0001507	1.02E-47		Gen_level_HP_0001507	6.01E-56
	Gen_level_HP_0002011	3.35E-49		Gen_level_HP_0002011	5.05E-55
	Gen_level_HP_0002715	3.12E-29		Gen_level_HP_0002715	7.10E-33
	Gen_level_HP_0002960	7.04E-01		Gen_level_HP_0002960	3.47E-01

	Gen_level_HP_0011446	5.78E-55		Gen_level_HP_0011446	2.19E-59
	Gen_level_HP_0012443	8.08E-50		Gen_level_HP_0012443	4.97E-53
	Gen_level_HP_0012638	2.46E-65		Gen_level_HP_0012638	4.90E-75
	Gen_level_HP_0012639	4.94E-52		Gen_level_HP_0012639	2.73E-58
	Gen_level_HP_0012759	6.99E-62		Gen_level_HP_0012759	1.26E-61
	Gen_level_HP_0025031	2.95E-46		Gen_level_HP_0025031	1.56E-57
	Gen_level_HP_0031466	2.42E-09		Gen_level_HP_0031466	1.77E-08
	Gen_level_HP_0100022	2.12E-38		Gen_level_HP_0100022	2.70E-44
	Gen_level_HP_0100753	1.32E-01		Gen_level_HP_0100753	8.25E-01
	Gen_level_HP_0100852	3.88E-04		Gen_level_HP_0100852	8.91E-04
Mouse heterozygous LoF lethal	Gen_level_mgi_essential_gene	1.42E-56	Mouse heterozygous LoF lethal	Gen_level_mgi_essential_gene	7.08E-81
Olfactory receptors	Gen_level_Olfactory_receptors_mainland	6.60E-03	Olfactory receptors	Gen_level_Olfactory_receptors_mainland	6.10E-01
Sfari gene	Gen_level_sfari_gene	1.81E-30	Sfari gene	Gen_level_sfari_gene	8.84E-22
Sequence level	*The Seq_level_significant features are 7 out of 7*		**Sequence level**	*The Seq_level_significant features are 6 out of 7*	
Blacklisted regions	Seq_level_DacMapExclude	4.23E-15	Blacklisted regions	Seq_level_DacMapExclude	5.88E-31
Sfari gene	Seq_level_DukeMapExclude	2.42E-26	Sfari gene	Seq_level_DukeMapExclude	3.62E-60
GC content	Seq_level_GC	7.26E-10	GC content	Seq_level_GC	8.61E-30
Human accelerated	Seq_level_HAR	8.53E-03	Human accelerated	Seq_level_HAR	2.01E-01

regions Heterochromatin positions			regions Heterochromatin positions		
Human accelerated regions Heterochromatin positions Cross species conservation score	Seq_level_HetDomain	3.54E-63	Human accelerated regions Heterochromatin positions Cross species conservation score	Seq_level_HetDomain	1.50E-60
	Seq_level_phastCons46way	1.04E-29		Seq_level_phastCons46way	1.07E-91
Human accelerated regions	Seq_level_phyloP46way	3.62E-19	Human accelerated regions	Seq_level_phyloP46way	1.82E-15

Table S3. Feature importancy. In the copy number loss and copy number gain models, we calculated feature importancy. $P = 0.05$ is set as the significant level. All the feature names were reformatted as feature names (original sources)_tissue type (if applicable).

Features in copy number loss model	**Feature importancy**	**Features in copy number gain model**	**Feature importancy**
1_TssA_chromHMM_brain	5.34E-03	1_TssA_chromHMM_brain	5.58E-03
10_EnhA2_chromHMM_brain	7.75E-03	10_EnhA2_chromHMM_brain	4.92E-03
11_EnhWk_chromHMM_brain	1.38E-02	11_EnhWk_chromHMM_brain	9.25E-03
12_ZNF_chromHMM_brain	4.61E-03	12_ZNF_chromHMM_brain	4.92E-03
13_Het_chromHMM_brain	5.81E-03	13_Het_chromHMM_brain	5.78E-03
14_TssBiv_chromHMM_brain	4.27E-03	14_TssBiv_chromHMM_brain	4.13E-03
15_EnhBiv_chromHMM_brain	5.65E-03	15_EnhBiv_chromHMM_brain	6.17E-03
16_ReprPC_chromHMM_brain	4.25E-03	16_ReprPC_chromHMM_brain	4.11E-03
17_ReprPCWk_chromHMM_brain	5.42E-03	17_ReprPCWk_chromHMM_brain	5.54E-03
18_Quies_chromHMM_brain	6.22E-03	18_Quies_chromHMM_brain	7.34E-03
2_TssFlnk_chromHMM_brain	4.47E-03	2_TssFlnk_chromHMM_brain	4.31E-03
3_TssFlnkU_chromHMM_brain	4.67E-03	3_TssFlnkU_chromHMM_brain	4.55E-03
4_TssFlnkD_chromHMM_brain	7.33E-03	4_TssFlnkD_chromHMM_brain	5.30E-03
5_Tx_chromHMM_brain	4.64E-03	5_Tx_chromHMM_brain	5.06E-03
6_TxWk_chromHMM_brain	4.98E-03	6_TxWk_chromHMM_brain	4.75E-03
7_EnhG1_chromHMM_brain	4.54E-03	7_EnhG1_chromHMM_brain	4.40E-03
8_EnhG2_chromHMM_brain	4.72E-03	8_EnhG2_chromHMM_brain	5.37E-03
9_EnhA1_chromHMM_brain	5.45E-03	9_EnhA1_chromHMM_brain	4.83E-03
ATAC-seq_observed_Brain	5.82E-03	ATAC-seq_observed_Brain	5.96E-03
Brain_Angular_Gyrus_dbsuper	4.18E-03	Brain_Angular_Gyrus_dbsuper	3.47E-03
Brain_Anterior_Caudate_dbsuper	4.33E-03	Brain_Anterior_Caudate_dbsuper	5.08E-03
Brain_Cingulate_Gyrus_dbsuper	3.99E-03	Brain_Cingulate_Gyrus_dbsuper	3.93E-03

Brain_Hippocampus_Middle_150_dbsuper	4.54E-03	Brain_Hippocampus_Middle_150_dbsuper	6.33E-03
Brain_Hippocampus_Middle_dbsuper	3.94E-03	Brain_Hippocampus_Middle_dbsuper	4.19E-03
Brain_Inferior_Temporal_Lobe_dbsuper	3.47E-03	Brain_Inferior_Temporal_Lobe_dbsuper	5.37E-03
Brain_Mid_Frontal_Lobe_dbsuper	4.99E-03	Brain_Mid_Frontal_Lobe_dbsuper	7.03E-03
ClinGen_haploinsufficiency_gene_0	3.71E-03	ClinGen_region_curation_Triplosensitivity_0	2.24E-03
ClinGen_haploinsufficiency_gene_1	5.91E-03	ClinGen_region_curation_Triplosensitivity_1	7.05E-03
ClinGen_haploinsufficiency_gene_2	0	ClinGen_region_curation_Triplosensitivity_2	6.10E-03
ClinGen_haploinsufficiency_gene_3	9.43E-03	ClinGen_region_curation_Triplosensitivity_3	5.38E-03
ClinGen_haploinsufficiency_gene_30	4.80E-03	ClinGen_region_curation_Triplosensitivity_40	1.01E-02
ClinGen_haploinsufficiency_gene_40	1.14E-03	ClinGen_triplosensitivity_gene	0
ClinGen_region_curation_Haploinsufficiency_0	4.63E-03	ClinGen_triplosensitivity_gene_0	6.16E-03
ClinGen_region_curation_Haploinsufficiency_1	0	ClinGen_triplosensitivity_gene_1	0
ClinGen_region_curation_Haploinsufficiency_2	2.59E-03	ClinGen_triplosensitivity_gene_2	0
ClinGen_region_curation_Haploinsufficiency_3	4.81E-03	ClinGen_triplosensitivity_gene_3	0
ClinGen_region_curation_Haploinsufficiency_30	0	ClinGen_triplosensitivity_gene_30	0
ClinGen_region_curation_Haploinsufficiency_40	1.05E-02	ClinGen_triplosensitivity_gene_40	0
Collins_rCNV_PLIgenes_PHI	5.24E-02	Collins_rCNV_PLIgenes_PTS	6.53E-02
ctcf	5.27E-03	ctcf	4.46E-03
CTCF_observed_Brain	4.34E-03	CTCF_observed_Brain	5.10E-03
DacMapExclude	5.65E-03	DacMapExclude	8.93E-03
ddg2p_loss	7.51E-03	ddg2p_gain	3.54E-03
DNaselClusterd	4.23E-03	DNaselClusterd	4.51E-03
DnaseMaster	5.52E-03	DnaseMaster	4.98E-03
DNase-seq_observed_Brain	1.11E-02	DNase-seq_observed_Brain	1.08E-02
DNase-seq_observed_Neurosph	2.30E-02	DNase-seq_observed_Neurosph	2.24E-02

DukeMapExclude	6.33E-03	DukeMapExclude	8.47E-03
EncodeAwgTfbsBroadNhaCtcf	5.10E-03	EncodeAwgTfbsBroadNhaCtcf	4.79E-03
EncodeRegTfbsClustered	4.86E-03	EncodeRegTfbsClustered	5.60E-03
enhancerAtlas_Astrocyte_EP	3.41E-03	enhancerAtlas_Astrocyte_EP	6.82E-03
enhancerAtlas_Cerebellum_EP	4.23E-03	enhancerAtlas_Cerebellum_EP	4.91E-03
enhancerAtlas_ESC_neuron_EP	3.34E-03	enhancerAtlas_ESC_neuron_EP	4.45E-03
EP300_imputed_Brain	4.79E-03	EP300_imputed_Brain	4.00E-03
EP300_imputed_Neurosph	5.06E-03	EP300_imputed_Neurosph	5.47E-03
Essential_in_culture_CRISPR	3.85E-03	Essential_in_culture_CRISPR	5.13E-03
famton_astrocyte	3.93E-03	famton_astrocyte	6.99E-03
famton_brain	6.68E-03	famton_brain	5.28E-03
famton_CL:0000127	3.45E-03	famton_CL:0000127	3.77E-03
famton_count	5.37E-03	famton_count	6.22E-03
famton_neuronal_stem_cell	4.61E-03	famton_neuronal_stem_cell	3.17E-03
famton_permssive	4.56E-03	famton_permssive	5.03E-03
FDA-approved_drug_targets	4.03E-03	FDA-approved_drug_targets	4.57E-03
GC	6.47E-03	gain_activating_score1	0
gencode_CDS	9.24E-03	gain_activating_score2	0
gencode_exon	7.74E-03	gain_activating_score3	1.80E-03
gencode_gene	7.42E-03	GC	5.89E-03
gencode_Selenocysteine	0	gencode_CDS	9.60E-03
gencode_start_codon	2.08E-02	gencode_exon	2.11E-02
gencode_stop_codon	3.54E-03	gencode_gene	7.17E-03
gencode_transcript	4.69E-03	gencode_Selenocysteine	4.83E-04
gencode_UTR	5.77E-03	gencode_start_codon	8.60E-03

gene_enhancer_links_brain_enhcenter	5.55E-03	gencode_stop_codon	4.29E-03
gene_enhancer_links_neurosph_enhcenter	6.37E-03	gencode_transcript	5.05E-03
gpcr_union	3.95E-03	gencode_UTR	4.27E-03
H2AFZ_imputed_Brain	4.28E-03	gene_enhancer_links_brain_enhcenter	4.33E-03
H2AFZ_imputed_Neurosph	4.71E-03	gene_enhancer_links_neurosph_enhcenter	6.14E-03
H2AFZ_observed_Brain	5.17E-03	gpcr_union	3.52E-03
H3k27ac	4.88E-03	H2AFZ_imputed_Brain	4.46E-03
H3K27ac_imputed_Brain	4.59E-03	H2AFZ_imputed_Neurosph	5.95E-03
H3K27ac_imputed_Neurosph	5.14E-03	H2AFZ_observed_Brain	5.61E-03
H3K27ac_observed_Brain	4.54E-03	H3k27ac	5.37E-03
H3K27ac_observed_Neurosph	5.79E-03	H3K27ac_imputed_Brain	6.32E-03
H3K27me3_imputed_Brain	5.02E-03	H3K27ac_imputed_Neurosph	6.64E-03
H3K27me3_imputed_Neurosph	6.51E-03	H3K27ac_observed_Brain	4.71E-03
H3K27me3_observed_Brain	6.65E-03	H3K27ac_observed_Neurosph	6.18E-03
H3k4me1	6.11E-03	H3K27me3_imputed_Brain	5.09E-03
H3K4me1_imputed_Brain	4.14E-03	H3K27me3_imputed_Neurosph	6.37E-03
H3K4me1_imputed_Neurosph	5.30E-03	H3K27me3_observed_Brain	7.70E-03
H3K4me1_observed_Brain	5.63E-03	H3k4me1	7.23E-03
H3K4me1_observed_Neurosph	4.32E-03	H3K4me1_imputed_Brain	5.71E-03
H3K4me2_observed_Brain	5.20E-03	H3K4me1_imputed_Neurosph	6.01E-03
H3k4me3	7.73E-03	H3K4me1_observed_Brain	5.83E-03
H3K4me3_imputed_Brain	6.09E-03	H3K4me1_observed_Neurosph	5.33E-03
H3K4me3_imputed_Neurosph	5.11E-03	H3K4me2_observed_Brain	5.29E-03
H3K4me3_observed_Brain	5.38E-03	H3k4me3	5.32E-03
H3K4me3_observed_Neurosph	5.40E-03	H3K4me3_imputed_Brain	5.26E-03

H3K9ac_imputed_Brain	5.52E-03	H3K4me3_imputed_Neurosph	4.10E-03
H3K9ac_imputed_Neurosph	5.54E-03	H3K4me3_observed_Brain	6.84E-03
H3K9me3_imputed_Brain	5.66E-03	H3K4me3_observed_Neurosph	6.45E-03
H3K9me3_imputed_Neurosph	5.85E-03	H3K9ac_imputed_Brain	4.98E-03
H3K9me3_observed_Brain	5.53E-03	H3K9ac_imputed_Neurosph	6.92E-03
H3K9me3_observed_Neurosph	4.70E-03	H3K9me3_imputed_Brain	6.78E-03
H4K20me1_imputed_Neurosph	4.40E-03	H3K9me3_imputed_Neurosph	6.60E-03
H4K20me1_observed_Brain	5.01E-03	H3K9me3_observed_Brain	6.58E-03
hacer_T1	5.06E-03	H3K9me3_observed_Neurosph	7.87E-03
HAR	4.44E-03	H4K20me1_imputed_Neurosph	5.54E-03
HetDomain	1.13E-02	H4K20me1_observed_Brain	5.35E-03
HP_0000707	3.42E-03	hacer_T1	6.26E-03
HP_0000708	4.66E-03	HAR	5.29E-03
HP_0000717	9.17E-03	HetDomain	6.86E-03
HP_0000729	3.13E-03	HP_0000707	4.05E-03
HP_0000752	3.17E-03	HP_0000708	4.61E-03
HP_0001197	4.68E-03	HP_0000717	4.45E-03
HP_0001250	3.83E-03	HP_0000729	5.33E-03
HP_0001507	4.22E-03	HP_0000752	5.64E-03
HP_0002011	5.61E-03	HP_0001197	4.52E-03
HP_0002715	7.74E-03	HP_0001250	2.60E-03
HP_0002960	9.04E-03	HP_0001507	5.80E-03
HP_0011446	5.33E-03	HP_0002011	4.32E-03
HP_0012443	6.80E-03	HP_0002715	4.48E-03
HP_0012638	3.39E-03	HP_0002960	3.96E-03

HP_0012639	5.11E-03	HP_0011446	6.29E-03
HP_0012759	7.35E-03	HP_0012443	2.78E-03
HP_0025031	4.49E-03	HP_0012638	4.53E-03
HP_0031466	5.46E-03	HP_0012639	2.91E-03
HP_0100022	6.53E-03	HP_0012759	5.83E-03
HP_0100753	0	HP_0025031	7.89E-03
HP_0100852	4.28E-03	HP_0031466	5.76E-03
liu_csbj_targetgene	5.18E-03	HP_0100022	3.54E-03
loss_of_function_score1	5.91E-03	HP_0100753	4.64E-03
loss_of_function_score2	3.25E-03	HP_0100852	5.19E-03
loss_of_function_score3	5.53E-03	liu_csbj_targetgene	4.99E-03
methMCRF	6.51E-03	methMCRF	7.48E-03
mgi_essential_gene	4.35E-03	mgi_essential_gene	1.39E-02
miRNA	4.85E-03	miRNA	5.29E-03
non-codingRNAs	5.81E-03	non-codingRNAs	4.88E-03
nonEssential_in_culture_CRISPR	4.74E-03	nonEssential_in_culture_CRISPR	4.26E-03
nott_Astrocyte_enhancers	3.95E-03	nott_Astrocyte_enhancers	5.09E-03
nott_Astrocyte_promoters	4.51E-03	nott_Astrocyte_promoters	6.32E-03
nott_H3K4me3_around_TSS	6.04E-03	nott_H3K4me3_around_TSS	5.61E-03
nott_Microglia_enhancers	4.68E-03	nott_Microglia_enhancers	5.14E-03
nott_Microglia_promoters	5.68E-03	nott_Microglia_promoters	5.16E-03
nott_Neuronal_enhancers	1.23E-02	nott_Neuronal_enhancers	1.09E-02
nott_Neuronal_promoters	4.69E-03	nott_Neuronal_promoters	5.36E-03
nott_Oligo_enhancers	4.86E-03	nott_Oligo_enhancers	5.79E-03
nott_Oligo_promoters	6.11E-03	nott_Oligo_promoters	4.81E-03

nott_superEnhancer	0	nott_superEnhancer	0
Olfactory_receptors_mainland	6.80E-03	Olfactory_receptors_mainland	2.50E-03
phastCons46way	6.22E-03	phastCons46way	6.99E-03
phyloP46way	5.13E-03	phyloP46way	5.65E-03
POLR2A_imputed_Neurosph	6.71E-03	POLR2A_imputed_Neurosph	4.29E-03
PsychENCODE_CBC_H3K27ac	6.66E-03	PsychENCODE_CBC_H3K27ac	4.85E-03
PsychENCODE_HiC_EP	4.96E-03	PsychENCODE_HiC_EP	4.61E-03
PsychENCODE_loops_interRegion	4.33E-03	PsychENCODE_loops_interRegion	4.30E-03
PsychENCODE_PEC_Enhancers	1.04E-02	PsychENCODE_PEC_Enhancers	7.93E-03
PsychENCODE_PFC_H3K27ac	6.18E-03	PsychENCODE_PFC_H3K27ac	4.74E-03
PsychENCODE_TAR	5.81E-03	PsychENCODE_TAR	5.18E-03
PsychENCODE_TC_H3K27ac	4.70E-03	PsychENCODE_TC_H3K27ac	4.70E-03
RAD21_imputed_Brain	5.53E-03	RAD21_imputed_Brain	6.13E-03
RAD21_imputed_Neurosph	5.69E-03	RAD21_imputed_Neurosph	6.04E-03
RoadmapDNasePromCount	4.78E-03	RoadmapDNasePromCount	4.88E-03
SE_ele	4.32E-03	SE_ele	5.16E-03
SEA00101	4.63E-03	SEA00101	5.57E-03
sfari_gene	4.56E-03	sfari_gene	4.32E-03
SMC3_imputed_Brain	6.01E-03	SMC3_imputed_Brain	5.35E-03
SMC3_imputed_Neurosph	4.65E-03	SMC3_imputed_Neurosph	6.88E-03
snp_selex	8.40E-03	snp_selex	4.48E-03
TAD56	7.91E-03	TAD56	7.23E-03
tss2000bp	1.78E-02	tss2000bp	1.46E-02
vista	4.74E-03	vista	4.54E-03
yue_loops_hippo	5.13E-03	yue_loops_hippo	6.64E-03

References

1. Hart T, Tong AHY, Chan K, et al. Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3 (Bethesda)* 2017;7:2719-27. doi: 10.1534/g3.117.041277

2. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581:434-43. doi: 10.1038/s41586-020-2308-7

3. Strande NT, Riggs ER, Buchanan AH, et al. Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. *Am J Hum Genet* 2017;100:895-906. doi: 10.1016/j.ajhg.2017.04.015

4. Wright CF, Fitzgerald TW, Jones WD, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* 2015;385:1305-14. doi: 10.1016/S0140-6736(14)61705-0

5. Collins RL, Glessner JT, Porcu E, et al. A cross-disorder dosage sensitivity map of the human genome. *Cell* 2022;185:3041-3055.e25. doi: 10.1101/2021.01.26.21250098

6. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;46:D1074-D82. doi: 10.1093/nar/gkx1037

7. Motenko H, Neuhauser SB, O'Keefe M, et al. MouseMine: a new data warehouse for MGI. *Mamm Genome* 2015;26:325-30. doi: 10.1007/s00335-015-9573-z

8. Kohler S, Gargano M, Matentzoglu N, et al. The Human Phenotype Ontology in 2021. *Nucleic Acids Res* 2021;49:D1207-D17. doi: 10.1093/nar/gkaa1043

9. Braschi B, Denny P, Gray K, et al. Genenames.org: the HGNC and VGNC resources in 2019. *Nucleic Acids Res* 2019;47:D786-D92. doi: 10.1093/nar/gky930

10. Abrahams BS, Arking DE, Campbell DB, et al. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism* 2013;4:36. doi: 10.1186/2040-2392-4-36

11. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012;9:215-6. doi: 10.1038/nmeth.1906

12. Boix CA, James BT, Park YP, et al. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* 2021;590:300-07. doi: 10.1038/s41586-020-03145-z

13. Sabo PJ, Hawrylycz M, Wallace JC, et al. Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc Natl Acad Sci U.S.A.* 2004;101:16837-42. doi: 10.1073/pnas.0407387101

14. Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* 2018;46:D794-D801. doi: 10.1093/nar/gkx1081

15. Roadmap Epigenomics C, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518:317-30. doi: 10.1038/nature14248

16. Nott A, Holtman IR, Coufal NG, et al. Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science* 2019;366:1134-39. doi: 10.1126/science.aay0793

17. Khan A, Zhang X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res* 2016;44:D164-71. doi: 10.1093/nar/gkv1002

18. Gao T, Qian J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res* 2020;48:D58-D64. doi: 10.1093/nar/gkz980

19. Andersson R, Gebhard C, Miguel-Escalada I, et al. An atlas of active enhancers across human cell types and tissues. *Nature* 2014;507:455-61. doi: 10.1038/nature12787

20. Wang J, Dai X, Berry LD, et al. HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res* 2019;47:D106-D12. doi: 10.1093/nar/gky864

21. Wang D, Liu S, Warrell J, et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* 2018;362:eaat8464. doi: 10.1126/science.aat8464

22. Jiang Y, Qian F, Bai X, et al. SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res* 2019;47:D235-D43. doi: 10.1093/nar/gky1025

23. Chen C, Zhou D, Gu Y, et al. SEA version 3.0: a comprehensive extension and update of the Super-Enhancer archive. *Nucleic Acids Res* 2020;48:D198-D203. doi: 10.1093/nar/gkz1028

24. Visel A, Minovitsky S, Dubchak I, et al. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* 2007;35:D88-92. doi: 10.1093/nar/gkl822

25. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 2012;22:1760-74. doi: 10.1101/gr.135350.111

26. Liu X, Xu W, Leng F, et al. Prioritizing long range interactions in noncoding regions using GWAS and deletions perturbed TADs. *Comput Struct Biotechnol J* 2020;18:2945-52. doi: 10.1016/j.csbj.2020.10.014

27. Wang Y, Song F, Zhang B, et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol* 2018;19:151. doi: 10.1186/s13059-018-1519-9

28. Yan J, Qiu Y, Ribeiro Dos Santos AM, et al. Systematic analysis of binding of transcription factors to noncoding variants. *Nature* 2021;591:147-51. doi: 10.1038/s41586-021-03211-0

29. Karolchik D, Hinrichs AS, Furey TS, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 2004;32:D493-6. doi: 10.1093/nar/gkh103

30. Ho JW, Jung YL, Liu T, et al. Comparative analysis of metazoan chromatin organization. *Nature* 2014;512:449-52. doi: 10.1038/nature13415

31. Doan RN, Bae BI, Cubelos B, et al. Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. *Cell* 2016;167:341-54 e12. doi: 10.1016/j.cell.2016.08.071

32. Harding SD, Sharman JL, Faccenda E, et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. *Nucleic Acids Res* 2018;46:D1091-D106. doi: 10.1093/nar/gkx1121

33. Alexander SP, Christopoulos A, Davenport AP, et al. THE CONCISE GUIDE TO PHARMACOLOGY 2017/18: G protein-coupled receptors. *Br J Pharmacol* 2017;174 Suppl 1:S17-S129. doi: 10.1111/bph.13878

34. Collins RL, Glessner JT, Porcu E, et al. A cross-disorder dosage sensitivity map of the human genome. *Cell* 2022;185:3041-55.e25. doi: 10.1016/j.cell.2022.06.036

35. UniProt Consortium T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2018;46:2699. doi: 10.1093/nar/gky092

36. Howe KL, Achuthan P, Allen J, et al. Ensembl 2021. *Nucleic Acids Res* 2021;49:D884-D91. doi: 10.1093/nar/gkaa942

BMJ Paediatrics Open

NeuroCNVscore: A tissue-specific framework to prioritize the pathogenicity of CNVs in neurodevelopmental disorders

Journal:	*BMJ Paediatrics Open*
Manuscript ID	bmjpo-2023-001966.R1
Article Type:	Original research
Date Submitted by the Author:	05-Jun-2023
Complete List of Authors:	Liu, Xuanshi; Beijing Children's Hospital Xu, Wenjian; Beijing Children's Hospital, Biology Leng, Fei; Beijing Children's Hospital Zhang, Peng; Beijing Children's Hospital Guo, Ruolan; Beijing Children's Hospital Zhang, Yue; Beijing Children's Hospital Hao, Chanjuan; Beijing Children's Hospital Ni, Xin; Beijing Children's Hospital, Department of Otolaryngology, Head and Surgery Li, Wei; Beijing Children's Hospital
Keywords:	Genetics, Neurology

Title: NeuroCNVscore: A Tissue-Specific Framework to Prioritize the Pathogenicity of CNVs in Neurodevelopmental Disorders

Short title: Prioritizing the pathogenicity of CNVs

Authors: Xuanshi Liu[1], Wenjian Xu[1], Fei Leng[1], Peng Zhang[1], Ruolan Guo[1], Yue Zhang[1], Chanjuan Hao[1*], Xin Ni[2*], Wei Li[1*]

Affiliations:

[1]*Beijing Key Laboratory for Genetics of Birth Defects, Beijing Paediatric Research Institute; MOE Key Laboratory of Major Diseases in Children; Genetics and Birth Defects Control Centre, Beijing Children's Hospital, Capital Medical University, National Centre for Children's Health, Beijing, China*;

[2]*Department of Otolaryngology, Head and Surgery, Beijing Children's Hospital, Capital Medical University, National Centre for Children's Health, Beijing, China.*

[*] **Corresponding authors.** Emails: liwei@bch.com.cn (Li W.), nixin@bch.com.cn (Ni X.), hchjhchj@163.com (Hao C.).

Word Count: 2360 words through Introduction to Discussion.

Fundings: This work was partially supported by the Ministry of Science and Technology of China (2019YFA0802104; 2016YFC1000306); the National Natural Science Foundation of China (31830054); the Beijing Natural Science Foundation (5222007); and the Beijing Municipal Health Commission (JingYiYan 2018-5).

Abstract

Background: Neurodevelopmental disorders (NDDs) are associated with altered development of the brain especially in childhood. Copy number variants (CNVs) play a crucial role in the genetic aetiology of NDDs by disturbing gene expression directly at linear sequence or remotely at three-dimensional genome level in a tissue-specific manner. Despite the substantial increase in NDD studies employing whole-genome sequencing, there is no specific tool for prioritizing the pathogenicity of CNVs in the context of NDDs. **Methods:** Using an XGBoost classifier, we integrated 189 features that represent genomic sequences, gene information, and functional/genomic segments for evaluating genome-wide CNVs in a neuro/brain-specific manner, to develop a new tool, neuroCNVscore. We utilized Human Phenotype Ontology to construct an independent NDD-related set. **Results:** Our neuroCNVscore framework (https://github.com/lxsbch/neuroCNVscore) achieved high predictive performance (PR = 0.82; AUC = 0.85) and outperformed an existing reference method SVScore. Notably, the predicted pathogenic CNVs showed enrichment in known genes associated with autism. **Conclusions**: NeuroCNVscore prioritizes functional, deleterious and pathogenic CNVs in NDDs at whole genome-wide level, which is important for genetic studies and clinical genomic screening of NDDs as well as for providing novel biological insights into NDDs.

Key Words: Neurodevelopmental disorder; Copy number variant; Pathogenicity; Tissue specificity; Gene expression

Key Messages:

- **What is already known on this topic**

CNVs are important in the genetic aetiology of NDDs. Systematic identification of CNV pathogenicity by virtue of their size, number and impact on genome is challenge. Several tools are available to evaluate CNVs or structural variants, but none on CNVs specific for NDDs.

- **What this study adds**

NeuroCNVscore is a useful tool in prioritizing functional and/or pathogenic CNVs in NDDs at whole genome-wide level in a neuro/brain-specific manner.

- **How this study might affect research, practice or policy**

Given the expanding studies on NDDs and the usage of sequencing in clinical practice, our neuroCNVscore speeds up the screening on pathogenic CNVs, which facilitates the clinical diagnoses of CNVs with unknown significant, and thus may provide novel biological insights into NDDs.

Introduction

Neurodevelopmental disorders (NDDs) are characterized by the inability to achieve cognitive, emotional, and motor developmental milestones including autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD) and schizophrenia. It is estimated to affect over 11.3%, and 15% of the population in low and middle-income countries (1) and US, (2) respectively. NDD's heritability is high that has been estimated from twin and family studies as 50% to 90% in ASD, (3) 88% in ADHD (4) and 85% in schizophrenia. (5) Genomic alterations are commonly found in children with NDDs. However, the explained genetic aetiology of NDDs accounts for only a small proportion.

Copy number variants (CNVs) are structural variants (SVs) in the genome that involve the gain or loss of large segments of DNA, which have been implicated in NDDs. (6,7) Systematic identification of CNV pathogenicity by virtue of their number, size and impact on the genome is still a challenge. It is approximately 1,000 CNVs per genome ranging in size from 50 base pairs (bp) to several mega bases (Mb). CNVs make effects by altering the dosage of gene regions (8) as well as by perturbing non-coding areas. (7,9) Growing number of studies by whole genome sequencing (WGS) and the complexity of identifying pathogenic CNVs call for computational prediction tools.

Many assessing tools have been developed to evaluate the pathogenicity of single nucleotide variants (SNVs), (10,11) but fewer studies have systematically focused on assessing the pathogenic CNVs, especially none in NDD-related CNVs. Recently, SVScore, (12) SVFX, (13) SVPath, (14) and AnnotSV (15) have been developed to interpret the SVs by integrating results from prediction matrices of SNPs, using cancer related SVs as inputs, counting SVs with overlapped exons, or integrating multiple sources to annotate SVs. However, the aggregated effects on SNPs, somatic impacts of SVs, or only overlapping exons without tissue-specific information may bias the effects of CNVs. As germline variations are the major focus in NDDs, a specific tool is needed for assessing the effects of CNVs on NDDs.

We here present a novel supervised machine learning framework, named as neuroCNVScore (https://github.com/lxsbch/neuroCNVscore), to score the pathogenicity of CNVs related to NDDs. We hypothesize that the computational prediction on pathogenic CNVs would benefit from a set of comprehensive tissue-specific features covering the whole genomic regions. Hence, we employed germline CNVs obtained from published NDD studies, (16-19) and curated gene lists together with a comprehensive set of neuro/brain-specific data on non-coding regions from ENCODE, (20) Roadmap, (21) EpiMap (22) and PsychENCODE (23) to train our models. Moreover, we constructed an independent dataset associated with NDDs by filtering the phenotypes from Human Phenotype Ontology (HPO, https://hpo.jax.org/) to evaluate the performance of our trained models. The performance of neuroCNVScore was compared with a reference method SVScore. (12) This neuroCNVScore is designed for assessing the pathogenicity of CNVs in NDDs generated from association studies or genetic tests.

Methods

Data collection and pre-processing/harmonization

We developed neuroCNVscore, which utilized XGBoost and comprehensive genome-wide features to evaluate the likelihood that a given CNV contributes to the development or manifestation of NDDs. To assess the pathogenicity associated with CNV in NDDs, we gathered training set (identified by genomic coordinates) from several case-control NDD studies. We assigned CNVs from cases as likely pathogenic (LP). In contrast, the CNVs from unaffected individuals and parents served as the control. Together, we collected 86,694 CNVs in the LP set and 786,058 in the control set from four data sources, respectively (**Fig. 1**).

Initial data filtering and harmonization were performed on all autosomal chromosome CNVs in three major steps. Firstly, we excluded CNVs with a size smaller than 50 base pairs, and the remaining CNVs were categorized into two groups based on their impact on the genome: copy number loss and copy number gain. Next, we deleted CNVs which had 90% reciprocal overlap between LP and control. Finally, we applied an empirical cumulative distribution function with bin size of 60 to generate size matched LP and control to overcome the amount of disparity between groups. For each CNV type, we sampled an equal number of LP CNVs ensuring the matching of control CNVs in each bin. For training process, we retained 13,857 cleaned LP CNVs and 13,859 cleaned control CNVs.

Next, we constructed an independent test set by assembling 51,819 disease associated variations from ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/) and 136,181 common CNVs from GnomAD 2.1 (http://www.gnomad-sg.org/). For the NDD related set, we retained CNVs with length > 50 bp, germline, pathogenic, and the term of HPO: 0012759 (neurodevelopmental abnormality associated genes). For common CNVs, we kept CNVs with quality record PASS, and allele frequency > 0.1. To avoid over-estimation, we removed those CNVs with 90% reciprocal overlap within the training dataset under the same variant type.

Finally, we collected several NDD related gene lists to evaluate the biological validity and robustness of neuroCNVscore including CHD8 target genes, (24) human postsynaptic density (PSD) proteins (25) and ASD risk genes (FDR < 0.3). (18) The overall workflow is outlined in **Fig. 1**.

A comprehensive tissue-specific feature collection and feature matrix construction

For each CNV, a broad range of features was compiled into a feature matrix. We leveraged 189 features in total from three different levels: (1) gene level (Gen), (2) functional/genomic segment level (Fun), and (3) sequence level (Seq). The description of features is shown in **Table S1**.

In brief, a set of gene level features (N = 62) that contain gene entity, dosage sensitivity and neurodevelopmental phenotype were collected. Since non-coding CNVs may disrupt regulatory regions to compromise gene expression and translation in a linear or 3D manner, we obtained a regulatory cascade catalogue (N = 120 at functional/genomic segment level). This catalogue integrated multi-omics data encompassing experimentally identified or computational predicted regulatory regions with a focus on tissue-specific annotation. Finally, the sequence level features (N = 7) comprised of information of GC content, cross species conservation score (phylop46way and phastcon46way which are derived from phyloP or Hidden Markov Model via multiple alignment of 45 vertebrate genomes to the human genome), heterochromatin positions, collapsed repeat regions (DacMapExclude, DukeMapExclude are genomic regions calculated by different algorithms) retrieved from the UCSC genome browser (http://genome.ucsc.edu/), and human accelerated regions accessed by Doan *et al.*. (26) These features were instrumental in identifying functional genomic regions and/or filtering out the genomic regions which may cause artefacts from downstream segments.

Based on a variety of features, annotations were performed in three distinct ways: (1) counting the number of overlapped features with a given CNV, (2) assessing a discrete value that denotes the number of the features which has >50% reciprocal overlapped regions with a given CNV, (3) calculating the average value of overlapped regions between the feature and a given CNV. After initial annotation, we divided the entire feature matrix based on the length of each CNV and then applied min-max scaling. Considering the differences in features, e.g. triplosensitivity is a measurement only for the copy number gain, we kept 172 features out of 189 for the copy number loss model and 172 features out of 189 in the copy number gain model, respectively.

Design of XGBoost model and the training strategy

To choose an appropriate model, we compared the performances among different algorithms (Naïve Bayes, Logistic Regression, Support Vector Machine, and XGBoost), and we found that XGBoost had the best performance in the python framework from Scikit 0.22.1 with the binary logistic objective function. A total of 80%/20% of the variant sets was used as training/test sets, respectively. Next, we trained the XGBoost model with optimized parameters by using grid search and evaluated our models through an independent test set. Additionally, we assessed the performance by comparing our model with SVScore, which can evaluate various types of SV including CNV.

Statistics

Statistical analyses were performed using Python (version 2.7). The performance was measured by precision-recall (PR) and receiver operating characteristic (ROC) curves.

For individual feature comparison, we applied two-tailed Wilcoxon rank-sum tests. All genomic data is in GRCh37 genome build. Figures were generated by the ggplot package in R (version 3.6.1) or matplotlib in Python.

Patient and public involvement

Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research. No ethical issues are involved in this study as this paper only used the data deposited in the public accessible databases.

Results

Feature analyses pinpoint comprehensive feature sets

To understand the characteristics of CNVs in NDDs, we investigated the distribution of features between LP and control sets. In total, we observed 121 and 106 significant features at the threshold of $P = 0.05$ in copy number loss and copy number gain models, respectively (**Table S2**). These findings demonstrated that a large spectrum of features have significant differences between sets.

Among these significant features, functional/genomic segment features ranked higher than the others. Most of the highly ranked features were related to histone modification markers (e.g. H3K27me3, H3K27ac) and 3D chromatin related features (e.g. enhancers) (**Fig. 2**). This is as expected since noncoding regions account for 98%

of the human genome and CNVs can affect the gene function by interrupting the regulatory regions.

Comparisons among four algorithms reveal the superior performance of XGBoost

To find an optimal model for identifying pathogenic CNVs, we evaluated the predictive performance of Naïve Bayes, Logistic Regression, Support Vector Machine (SVM) and XGBoost on the test sets (**Fig. 3**). The XGBoost model showed the highest performance (average precision (AP) and area under curve (AUC) were 0.82, 0.85 for copy number loss; AP and AUC were 0.80, 0.84 for copy number gain). Therefore, we applied the XGBoost model to construct our neuroScoreCNV framework.

Accuracy assessments reveal better performance of neuroScoreCNV than SVScore

We evaluated the performance of neuroScoreCNV and SVScore by an independent set as described in the flowchart (**Fig. 1**). NeuroScoreCNV achieved relatively better performance evaluated by both AP and AUC values compared to SVScore (**Fig. 4**). The different performances between models are in agreement with a previous study. (13)

Moreover, we investigated the biological validity and robustness from two aspects. It was shown that interruptions at conserved regions could cause diseases since these regions are normally functional. (27) Therefore, we first computed the CNV pathogenic scores generated with the new feature matrices in which a conservation score (i.e.

PhyloP46way, one of the commonly used conservation score that considering individual base conservation) was excluded. We observed that higher CNV pathogenic scores (≥0.7) tended to have higher conservation scores, as indicated by the correlation between $\log_{10}$(PhyloP46way) and the new pathogenic scores (**Fig. 5A, B**). Then, we checked if our predicted scores were capable of prioritizing CNVs with known NDD-associated genes. LP CNVs covered significantly ($P < 0.05$) more NDD-related genes than the control group (**Fig. 5B**). Overall, our approach achieved higher performance in discriminating LP CNVs from control or benign CNVs.

Feature importancy highlights the important role of regulatory regions in NDDs

We categorized model features into three groups: functional/genomic level (Fun), gene level (Gen) and sequence level (Seq) and computed the feature importancy by permutation. (**Fig. 6Figure 6**, **Table S3**). The most important features were genes with haploinsufficiency scores (PHI) and triplosensitivity scores (PTS). PHI reflects the probability of one single functional copy to be sufficient to maintain function, whereas PTS suggests the probability of an additional copy of a gene for generating phenotypes. PHI and PTS are important parameters for evaluating the pathogenicity in clinical diagnoses based on the ACMG guidelines. (28) This is also true in neuroCNVScore. In NDDs, several studies found pathogenic CNVs were sensitive to dosage. (29)

Additionally, we noticed several prominent phenotypes such as HPO: 000717 (autism associated genes), HPO: 0002960 (autoimmunity associated genes) and HPO:

0025031 (abnormality of the digestive system associated genes). It is known that immune system abnormalities and/or gastrointestinal symptoms can co-occur with ASD (30) and schizophrenia. (31) Compelling evidence has demonstrated the importance of autoimmune response in ASD. (32) Purified IgG containing antibodies from the mothers of children with ASD can cause abnormal behaviours in animal models. (33,34)

Among the important features at the functional/genomic segment level, we observed several key players in 3D chromatin conformation including enhancers and topologically associated domains (TADs). Meanwhile, DNase-Seq which suggests active regulatory elements at open chromatin was also an important feature. The emerging evidence has highlighted the role of 3D chromatin conformation in relation to NDDs. (23, 35) Collectively, studying the interaction between CNVs and the higher order of chromatin conformation could provide novel insights into the aetiology of NDDs and explain the missing heredity of NDDs.

Discussion

In this study, we have introduced a novel framework, neuroCNVscore, to evaluate the pathogenicity of CNVs in NDDs. NeuroCNVscore outperformed a commonly used tool SVScore on independent datasets from ClinVar and gnomAD. Importantly, neuroCNVscore has unique ability to prioritize the functional, deleterious and pathogenic CNVs derived from either NDD's association studies or clinical diagnoses, which may provide biological insights into NDDs, especially at the three-dimensional genome level.

There are several factors contribute to the accuracy and robustness of neuroCNVscore. First, we used a high-quality set of germline CNVs from published NDD studies as the training set, ensuring the high reliability of this model. Secondly, we validated our models by using an independent dataset associated with NDD, which outperformed a published tool, SVScore. Furthermore, we curated a comprehensive feature collection (N = 189) at gene, functional genomic, and sequence levels. Specifically, we incorporated a significant amount of tissue-specific functional genomic data, enabling the identification of disrupted genes and regulatory elements that act in a tissue-specific manner during development. This is especially important for the studies in NDD since brain tissue is normally hard to access.

While the neuroCNVscore performed well, it may be improved by incorporating expert-curated CNVs from whole genome sequencing studies in NDDs and healthy controls. Along with the increased knowledge and functional genomics data on non-coding regions, additional informative features can be integrated into the model to better address the underlying mechanisms. Moreover, we developed neuroCNVscore based on XGBoost, but it is worth exploring deep learning algorithms in future investigation.

In summary, our neuroCNVscore is a useful tool for generating hypotheses in genome-wide association studies in NDDs and could facilitate the understanding of genetic aetiology of NDDs.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

XL designed the study, performed the analysis and drafted the manuscript. WX and FL participated in the design and interpretation of the data and revised the manuscript. PZ, RG and YZ participated in the interpretation of data. CH coordinated the project and supervised the study. XN coordinated the project and acquisition the funding. WL coordinated the project, supervised the study, critically reviewed and revised the manuscript. All authors read and approved the final manuscript.

Availability of Data and Materials

All features analysed during this study are collected from public datasets. Sources can be found from https://github.com/macarthur-lab/gene_lists. All CNV training data are included in these publications [16-19] and testing data are from the ClinVar database. The source code is available at https://github.com/lxsbch/neuroCNVscore.

Ethics Statement

This study has been approved by the Ethics Committee of Beijing Children's Hospital, Capital Medical University (2018-k-62).

Acknowledgements

We thank MacArthur's Lab for sharing the comprehensive collections of gene lists. We thank Dr. Sree Rohit Raj Kolora for reviewing, revising the manuscript and useful discussion.

References

1. Bitta M, Kariuki SM, Abubakar A, et al. Burden of neurodevelopmental disorders in low and middle-income countries: A systematic review and meta-analysis. *Wellcome Open Res* 2017;2:121. doi: 10.12688/wellcomeopenres.13540.3

2. America's Children and the Environment. Health: Neurodevelopmental Disorders – Report Contents, 2019.

3. Gaugler T, Klei L, Sanders SJ, et al. Most genetic risk for autism resides with common variation. *Nat Genet* 2014;46:881-5. doi: 10.1038/ng.3039

4. Larsson H, Chang Z, D'Onofrio BM, et al. The heritability of clinically diagnosed attention deficit hyperactivity disorder across the lifespan. *Psychol Med* 2014;44:2223-9. doi: 10.1017/S0033291713002493

5. Cardno AG, Marshall EJ, Coid B, et al. Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. *Arch Gen Psychiatry* 1999;56:162-8. doi: 10.1001/archpsyc.56.2.162

6. Marshall CR, Howrigan DP, Merico D, et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat Genet* 2017;49:27-35. doi: 10.1038/ng.3725

7. Brandler WM, Antaki D, Gujral M, et al. Paternally inherited cis-regulatory structural variants are associated with autism. *Science* 2018;360:327-31. doi: 10.1126/science.aan2261

8. Coe BP, Stessman HAF, Sulovari A, et al. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat Genet* 2019;51:106-16. doi: 10.1038/s41588-018-0288-4

9. Devanna P, Chen XS, Ho J, et al. Next-gen sequencing identifies non-coding variation disrupting miRNA-binding sites in neurological disorders. *Mol Psychiatry* 2018;23:1375-84. doi: 10.1038/mp.2017.30

10. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting

damaging missense mutations. *Nat Methods* 2010;7:248-9. doi: 10.1038/nmeth0410-248

11. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 2019;176:535-48 e24. doi: 10.1016/j.cell.2018.12.015

12. Ganel L, Abel HJ, FinMetSeq C, et al. SVScore: an impact prediction tool for structural variation. *Bioinformatics* 2017;33:1083-85. doi: 10.1093/bioinformatics/btw789

13. Kumar S, Harmanci A, Vytheeswaran J, et al. SVFX: a machine learning framework to quantify the pathogenicity of structural variants. *Genome Biol* 2020;21:274. doi: 10.1186/s13059-020-02178-x

14. Yang Y, Wang X, Zhou D, et al. SVPath: an accurate pipeline for predicting the pathogenicity of human exon structural variants. *Brief Bioinform* 2022;23: bbac014 doi: 10.1093/bib/bbac014

15. Geoffroy V, Guignard T, Kress A, et al. AnnotSV and knotAnnotSV: a web server for human structural variations annotations, ranking and analysis. *Nucleic Acids Res* 2021;49:W21-W28. doi: 10.1093/nar/gkab402

16. Coe BP, Witherspoon K, Rosenfeld JA, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* 2014;46:1063-71. doi: 10.1038/ng.3092

17. Cooper GM, Coe BP, Girirajan S, et al. A copy number variation morbidity map of

developmental delay. *Nat Genet* 2011;43:838-46. doi: 10.1038/ng.909

18. Sanders SJ, He X, Willsey AJ, et al. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* 2015;87:1215-33. doi: 10.1016/j.neuron.2015.09.016
19. Zarrei M, Burton CL, Engchuan W, et al. A large data resource of genomic copy number variation across neurodevelopmental disorders. *NPJ Genom Med* 2019;4:26. doi: 10.1038/s41525-019-0098-3
20. Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* 2018;46:D794-D801. doi: 10.1093/nar/gkx1081
21. Roadmap Epigenomics C, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518:317-30. doi: 10.1038/nature14248
22. Boix CA, James BT, Park YP, et al. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* 2021;590:300-07. doi: 10.1038/s41586-020-03145-z
23. Wang D, Liu S, Warrell J, et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* 2018;362:eaat8464. doi: 10.1126/science.aat8464
24. Sugathan A, Biagioli M, Golzio C, et al. CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proc*

Natl Acad Sci U.S.A. 2014;111:E4468-77. doi: 10.1073/pnas.1405266111

25. Bayes A, van de Lagemaat LN, Collins MO, et al. Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nat Neurosci* 2011;14:19-21. doi: 10.1038/nn.2719

26. Doan RN, Bae BI, Cubelos B, et al. Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. *Cell* 2016;167:341-54 e12. doi: 10.1016/j.cell.2016.08.071

27. Kellis M, Wold B, Snyder MP, et al. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U.S.A.* 2014;111:6131-8. doi: 10.1073/pnas.1318948111

28. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405-24. doi: 10.1038/gim.2015.30

29. Han X, Chen S, Flynn E, et al. Distinct epigenomic patterns are associated with haploinsufficiency and predict risk genes of developmental disorders. *Nat Commun* 2018;9:2138. doi: 10.1038/s41467-018-04552-7

30. Hughes HK, Mills Ko E, Rose D, et al. Immune Dysfunction and Autoimmunity as Pathological Mechanisms in Autism Spectrum Disorders. *Front Cell Neurosci* 2018;12:405. doi: 10.3389/fncel.2018.00405

31. Severance EG, Prandovszky E, Castiglione J, et al. Gastroenterology issues in

schizophrenia: why the gut matters. *Curr Psychiatry Rep* 2015;17:27. doi: 10.1007/s11920-015-0574-0

32. Wu S, Ding Y, Wu F, et al. Family history of autoimmune diseases is associated with an increased risk of autism in children: A systematic review and meta-analysis. *Neurosci Biobehav Rev* 2015;55:322-32. doi: 10.1016/j.neubiorev.2015.05.004

33. Bauman MD, Iosif AM, Ashwood P, et al. Maternal antibodies from mothers of children with autism alter brain growth and social behavior development in the rhesus monkey. *Transl Psychiatry* 2013;3:e278. doi: 10.1038/tp.2013.47

34. Hertz-Picciotto I, Croen LA, Hansen R, et al. The CHARGE study: an epidemiologic investigation of genetic and environmental factors contributing to autism. *Environ Health Perspect* 2006;114:1119-25. doi: 10.1289/ehp.8483

35. Won H, de la Torre-Ubieta L, Stein JL, et al. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* 2016;538:523-27. doi: 10.1038/nature19847

Figure Legends

Figure 1. The flowchart of neuroCNVscore development and evaluation in this study. In Data Sets, the sources of training set and test set are listed. The training set was derived from four NDDs studies under the case-control design, while the validation set

was from ClinVar and GnomAD. The numbers of raw and cleaned CNVs in the brackets are indicated. LP, likely pathogenic. In Neuro-features, comprehensive neuro/brain related features were gathered at gene, sequence, and functional/genomic segments levels. In Prediction and Validation, biological validations were performed in two ways: 1) correlation analyses between phyloP46way and the pathogenic scores generated by the new model where phyloP46way was excluded from the feature matrix; 2) utilization of an independent set of NDD related gene lists including PSD genes to cognition, CHD8 targets, and ASD risk genes.

Figure 2. Comparisons of top three features between control and LP (likely pathogenic) sets. The top three significant features between control and LP sets in copy number loss (A) and copy number gain (B). The X-axis shows the types of significant features. Fun_level, Function/genomic segment level. The Y-axis displays the values of log-transformed feature matrices. Unpaired *t*-tests were applied and significant levels were. **** $P < 0.0001$.

Figure 3. Performances of Naïve Bayes, Logistic Regression, Support Vector Machine (SVM) and XGBoost algorithms in evaluating CNVs. XGBoost showed superior performance demonstrated by precision-recall curves and ROC curves for both copy number loss (A, B) and copy number gain (C, D). AP: average precision; AUC: area under curve.

Figure 4. Performances of neuroCNVscore and SVScore in an independent set as described in the flowchart of Figure 1. Precision-Recall (A) and ROC (B) curves were calculated with copy number loss from the independent dataset; Precision-Recall (C) and ROC (D) curves were calculated with copy number gain from the independent dataset.

Figure 5. Biological validation of neuroCNVscore. The plot (A) shows the comparisons between PhyloP scores (log10(PhyloP46way)) and pathogenic scores generated by excluding PhyloP46way from the original neuroCNVscore model, regions with higher pathogenic scores tend to have higher PhyloP scores. The number of NDD related genes (B) between the predicted LP and control groups in both copy number loss and copy number gain models shows that more NDD related genes are found in LP groups. For better presentation, log transformations were applied to PhyloP46way scores and the gene counts. **P* < 0.05.

Figure 6. Top 20 features obtained from feature importance analyses. Highly important features of copy number loss model (A) and copy number gain model (B) are listed. All the feature names were color-coded and formatted as following: feature type (Fun_/Gen_/Seq_feature names (original sources)_tissue type (if applicable). Fun: Function, in blue; Gen: Gene, in green; Seq: Sequence, in purple.

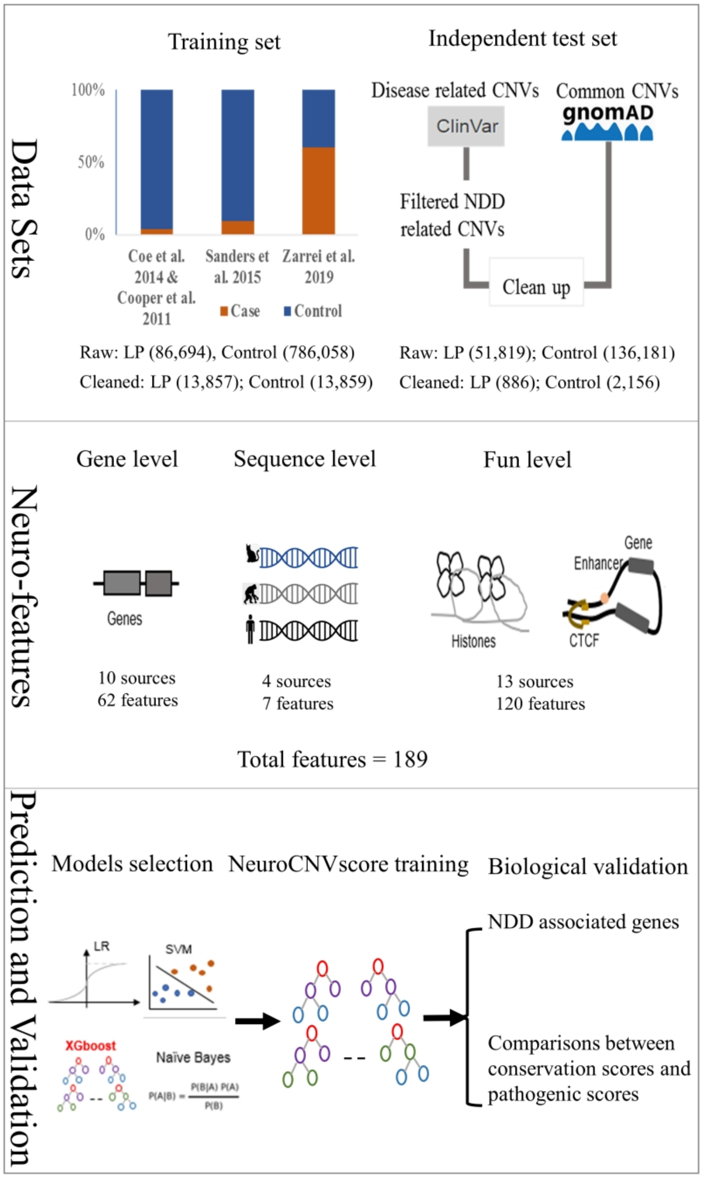

Figure 1

190x319mm (300 x 300 DPI)

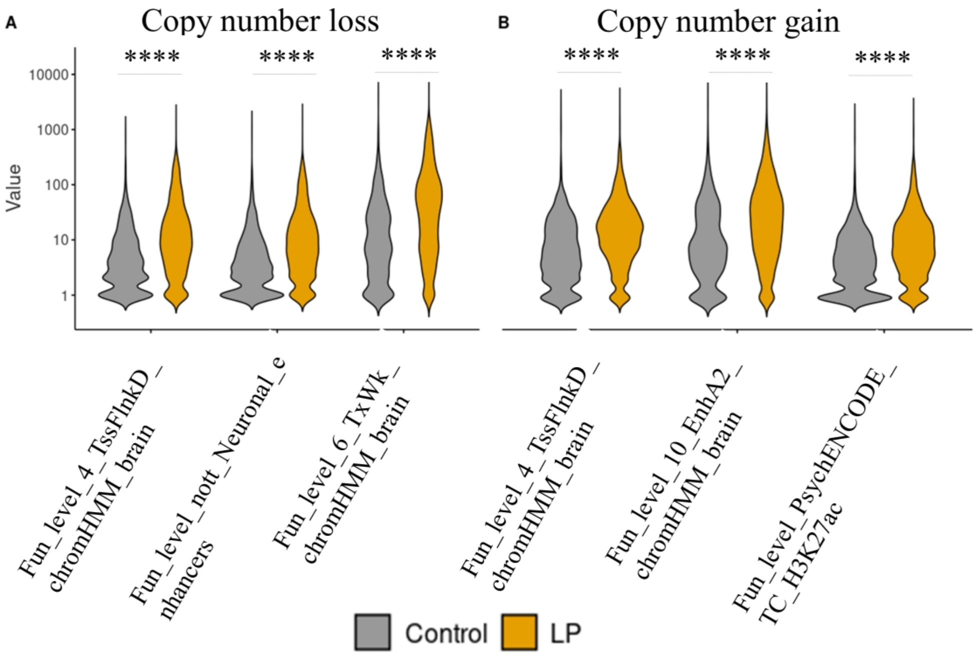

Figure 2

194x133mm (300 x 300 DPI)

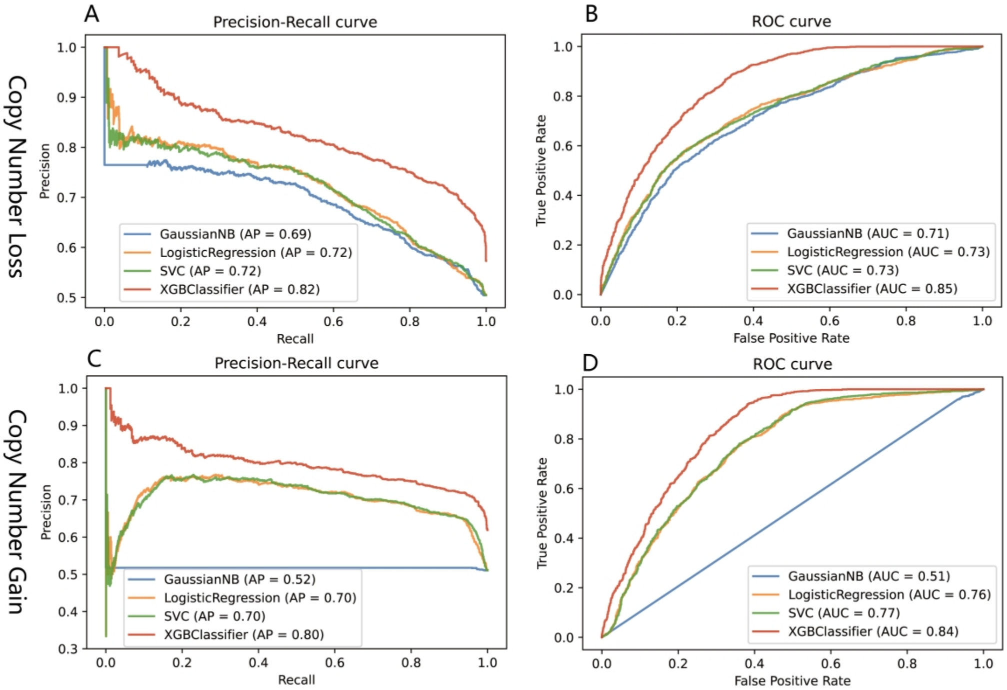

Figure 3-XGBooster performance

252x172mm (300 x 300 DPI)

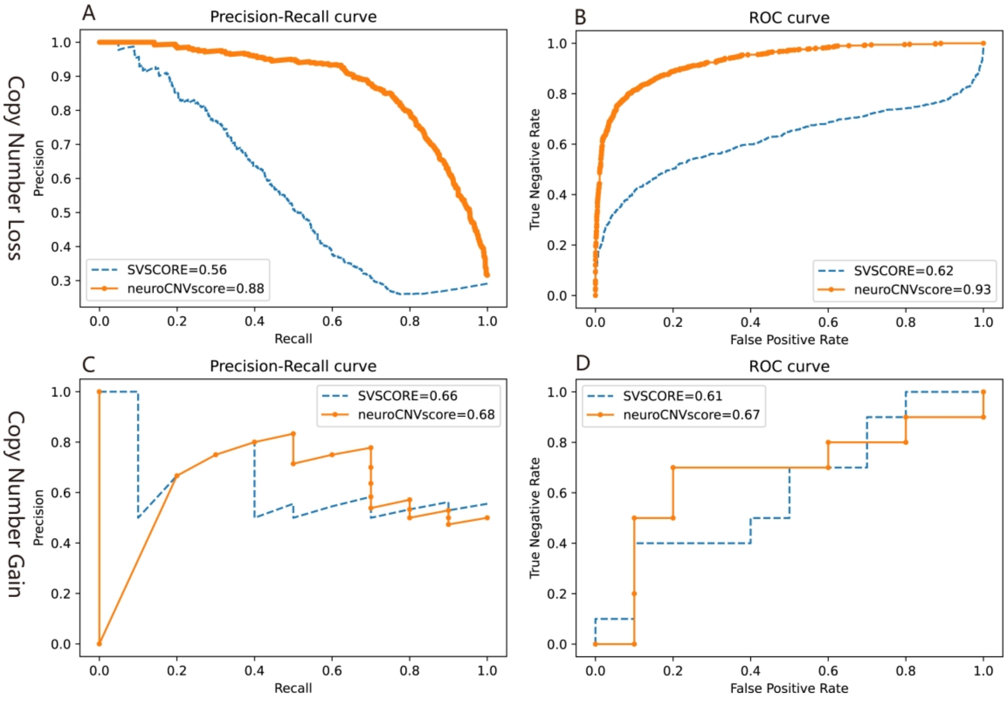

Figure 4-neuroCNV vs. SV

276x190mm (300 x 300 DPI)

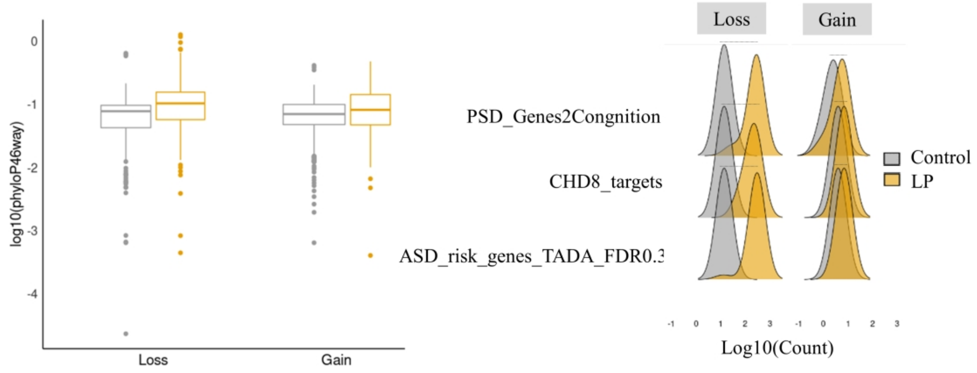

Figure 5

323x125mm (300 x 300 DPI)

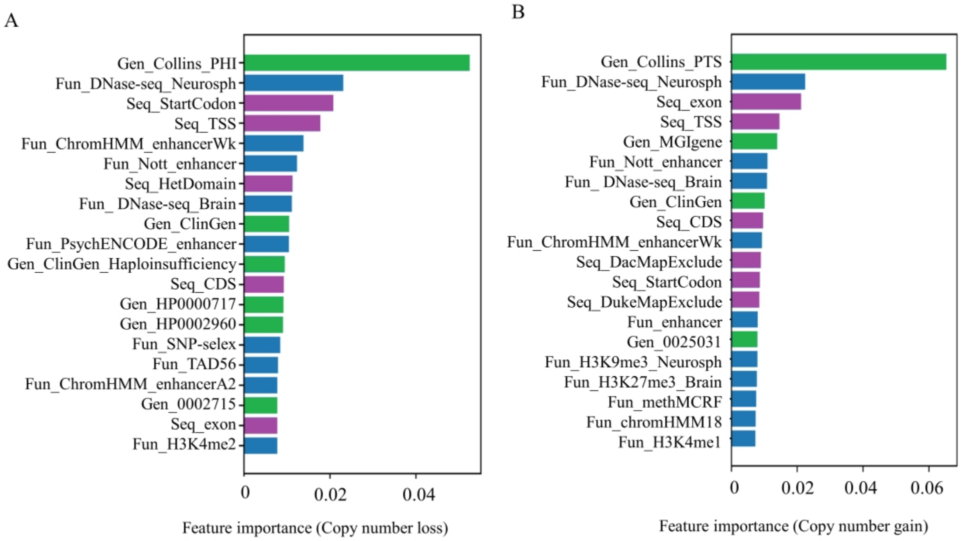

Figure 6

338x190mm (300 x 300 DPI)

Supplementary Tables

Table S1. A detailed feature description. This table includes all features used in our model. These features are grouped into three levels: gene, functional/genomic segment and sequence. A brief description along with references is described on each feature.

Feature category	**Feature set**	**Description**	**Feature type**	**References**
Gene level (N = 61)	Cell essential and nonessential genes	CRISPR/Cas9 screens identified essential genes in human cell lines. Curated in[2]	discrete	[1]
	ClinGen curated genes and genomic regions	Genes and genomics regions were rated from 0 to 3, indicating an increased evidence on dosage sensitivity. Additional two levels (40,30) suggest unlikely dosage sensitive and genes associated with autosomal recessive phenotype.	discrete	[3]
	DDG2P database	A curated list of genes linked to developmental disorders compiled by clinicians as part of the DDD study to facilitate clinical feedback on likely causal variants	discrete	[4]
	Dosage sensitive genes	Predicted score on dosage sensitive genes (i.e., haploinsufficiency or triplosensitivity)	discrete	[5, 34]
	FDA proved drug target	Genes with protein products that are mechanistic targets of FDA-approved drugs. Curated in [2]	discrete	[6]
	G protein-coupled receptor	GPCR list curated in[2]	discrete	[32, 33, 35]
	Mouse heterozygous LoF lethal	Genes that are lethal in mouse models when inactivated heterozygous. Curated by [2]	discrete	[7]
	Neurodevelopmental process related genes	Genes associated with various phenotypes from HPO: Abnormality of the nervous system (HP:0000707)-associated genes Abnormality of nervous system physiology (HP:0012638)-associated	discrete	[8]

		genes Behavioral abnormality (HP:0000708)-associated genes Abnormality of nervous system morphology (HP:0012639)-associated genes Abnormality of the immune system (HP:0002715)-associated genes Neurodevelopmental abnormality (HP:0012759)-associated genes Autoimmunity (HP:0002960)-associated genes Morphological abnormality of the central nervous system (HP:0002011)-associated genes Schizophrenia (HP:0100753)-associated genes Autistic behavior (HP:0000729)-associated genes Abnormality of movement (HP:0100022)-associated genes Seizures (HP:0001250)-associated genes Autism (HP:0000717)-associated genes Hyperactivity (HP:0000752)-associated genes Abnormality of prenatal development or birth (HP:0001197)-associated genes Impairment in personality functioning (HP:0031466)-associated genes Abnormality of the digestive system (HP:0025031)-associated genes Growth abnormality (HP:0001507)-associated genes Abnormal fear/anxiety-related behavior (HP:0100852)-associated genes Abnormality of brain morphology (HP:0012443)-associated genes Abnormality of higher mental function (HP:0011446)-associated genes		
	Olfactory receptors	Any HUGO-recognized family of olfactory receptor genes	discrete	[9]
	SFARI gene	Genes implicated in autism susceptibility	discrete	[10]

Functional/genomic segment level (N = 121)	Chromatin states	Brain related chromatin states inferred by the extended 18-way ChromHMM model across 98 tissues from the Roadmap Epigenomics Project	discrete	11
	CTCF binding sites	Genome wide observed CTCF binding sites from Brain	continuous	12
		Genome wide CTCF binding sites from 7 cell lines generated by ChIP-seq. Curated by UCSC	continuous	13
	DNA Accessibility	ATAC-seq from brain and neurosph.	continuous	13
	DNase hypersensitive sites	Observed DNase I hypersensitive areas from brain and neurosph.	continuous	13
		DNase hypersensitive sites assayed from a collection of cell types. Download from UCSC table browser NAR 2004	continuous	14
		RoadmapDNasePromCount	discrete	15
	Enhancers	Brain cell type-specific enhancers identified by PLAC-seq	discrete	16
		dbSUPER: Super enhancers from Brain Angular Gyrus; Brain Anterior Caudate; Brain Cingulate Gyrus; Brain Hippocampus Middle; Brain Inferior Temporal Lobe	discrete	17
		EpiMap: enhancers from the brain and neurosph.	discrete	12
		EnhancerAtlas 2.0: Enhancer predictions in 197 human cell lines & tissues	discrete	18
		FANTOM Enhancers: Enhancer predictions for human tissues and cell types from the FANTOM5 consortium	discrete	19
		HACER: Active enhancer predictions in human cell lines & tissues based on PRO-seq, GRO-seq, or CAGE data	discrete	20
		PsychENCODE: PEC EnhancersDER-03a_hg19_PEC_enhancers_clean.bed	discrete	21
		SEA: Super enhancer predictions from 143 human cell lines and tissues (mapped back to hg19 using liftOver with minimum 75% match)	discrete	23

		Sedb: Super enhancer and typical enhancer predictions from 541 human cell lines and tissues	discrete	22
		VISTA: Experimentally-validated mammalian enhancers	discrete	24
	Genomic segmentations	All autosomal, protein-coding genes; CDS; exon; Selenocysteine; start_codon; stop_codon; transcript UTR	discrete	25
	Histone markers	H2AFZ, H2AK5ac, H2AK9ac, H2BK120ac, H2BK12ac, H2BK15ac, H2BK20ac, H2BK5ac, H3F3A, H3K27ac, H3K4ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me1, H3K9me2, H3K9me3 from the brain or neurosph	continuous	12
		H3K27ac peaks for the Prefrontal Cortex, the Temporal Cortex, and the Cerebellar Cortex	continuous	21
	Long range probable genes	Target genes by prediction on GWAS hits and 3D chromatin structures	discrete	26
	Loop anchors and topological associated domains in higher-order chromatin structure	TAD boundaries (defined as the start and end coordinates for each TAD ± 5kb) from 30 samples meeting our ENCODE data inclusion criteria available for download from the ENCODE Data Portal	continuous	14
		Selected "derived" datasets from PsychENCODE Integrated Analysis Package, including cortex enhancers, transcriptionally active regions, TAD boundaries, and H3k27ac peaks	continuous	21
		Yue labs	continuous	27
	Methylation	MeDIP/MRE (mCRF) methylation calls	continuous	15
	Transcript active regions	Cortex Transcriptionally Active Regions are found within at least 70% of the individuals	continuous	21
	Transcript factor binding sites	SNP-SELEX	discrete	28
	Transcript starting sites	The 2000bp flanking regions about transcript starting sites	discrete	36

Sequence level (N = 7)	Blacklisted regions	Genome regions have anomalous, unstructured, high signal/read counts (DacMapExclude), problematic regions for short sequence tag signal detection (DukeMapExclude)	discrete	29
	Cross species conservation score	The conservation scoring (phylop46way, phastcon46way) for multiple alignments of 45 vertebrate genomes to the human genome	continuous	29
	GC content	GC content calculated with a "span" size of 5 bases	continuous	29
	Heterochromatin positions	It is calculated based on H3K9me3 enrichment regions	discrete	30
	Human accelerated regions	Human accelerated regions are conserved genomic loci with elevated divergence in humans	discrete	31

Table S2. Individual feature comparisons. This table compares all of the features used in the copy number loss and copy number gain models. The comparisons were made using the two-tailed Wilcoxon rank-sum test, with a significant cut off of *P* = 0.05. All the feature names were reformatted as followed: feature type (Fun_level/Gen_level/Seq_level)_feature names(original sources)_tissue type (if applicable). Fun: Function; Gen: Gene; Seq: Sequence.

Source	**Features in copy number loss model**	***P* value**	**Source**	**Features in copy number gain model**	***P* value**
Functional/ genomic segment level	*Fun_level_significant features are 80 out of 120*		**Functional/ genomic segment level**	*Fun_level_significant features are 75 out of 120*	
Chromatin states from Roadmap Epigenomics Project	Fun_level_1_TssA_chromHMM_brain	~0	Chromatin states from Roadmap Epigenomics Project	Fun_level_1_TssA_chromHMM_brain	~0
	Fun_level_10_EnhA2_chromHMM_brain	~0		Fun_level_10_EnhA2_chromHMM_brain	~0
	Fun_level_11_EnhWk_chromHMM_brain	~0		Fun_level_11_EnhWk_chromHMM_brain	~0
	Fun_level_12_ZNF_chromHMM_brain	~0		Fun_level_12_ZNF_chromHMM_brain	0.0003
	Fun_level_13_Het_chromHMM_brain	~0		Fun_level_13_Het_chromHMM_brain	~0
	Fun_level_14_TssBiv_chromHMM_brain	0.1050		Fun_level_14_TssBiv_chromHMM_brain	~0
	Fun_level_15_EnhBiv_chromHMM_brain	~0		Fun_level_15_EnhBiv_chromHMM_brain	~0
	Fun_level_16_ReprPC_chromHMM_brain	~0		Fun_level_16_ReprPC_chromHMM_brain	0.9910
	Fun_level_17_ReprPCWk_chromHMM_brain	0.0498		Fun_level_17_ReprPCWk_chromHMM_brain	~0
	Fun_level_18_Quies_chromHMM_brain	~0		Fun_level_18_Quies_chromHMM_brain	~0
	Fun_level_2_TssFlnk_chromHMM_brain	~0		Fun_level_2_TssFlnk_chromHMM_brain	~0
	Fun_level_3_TssFlnkU_chromHMM_brain	~0		Fun_level_3_TssFlnkU_chromHMM_brain	~0
	Fun_level_4_TssFlnkD_chromHMM_brain	~0		Fun_level_4_TssFlnkD_chromHMM_brain	~0
	Fun_level_5_Tx_chromHMM_brain	~0		Fun_level_5_Tx_chromHMM_brain	~0
	Fun_level_6_TxWk_chromHMM_brain	~0		Fun_level_6_TxWk_chromHMM_brain	~0
	Fun_level_7_EnhG1_chromHMM_brain	~0		Fun_level_7_EnhG1_chromHMM_brain	~0

	Fun_level_8_EnhG2_chromHMM_brain	~0		Fun_level_8_EnhG2_chromHMM_brain	~0
	Fun_level_9_EnhA1_chromHMM_brain	~0		Fun_level_9_EnhA1_chromHMM_brain	~0
Enhancers	Fun_level_dbsuper_Brain_Angular_Gyrus	~0	Enhancers	Fun_level_dbsuper_Brain_Angular_Gyrus	~0
	Fun_level_dbsuper_Brain_Anterior_Caudate	~0		Fun_level_dbsuper_Brain_Anterior_Caudate	~0
	Fun_level_dbsuper_Brain_Cingulate_Gyrus	~0		Fun_level_dbsuper_Brain_Cingulate_Gyrus	~0
	Fun_level_dbsuper_Brain_Hippocampus_Middle_150	~0		Fun_level_dbsuper_Brain_Hippocampus_Middle_150	~0
	Fun_level_dbsuper _Brain_Hippocampus_Middle	~0		Fun_level_dbsuper _Brain_Hippocampus_Middle	~0
	Fun_level_dbsuper_Brain_Inferior_Temporal_Lobe	~0		Fun_level_dbsuper_Brain_Inferior_Temporal_Lobe	~0
	Fun_level_dbsuper_Brain_Mid_Frontal_Lobe	0.0293		Fun_level_dbsuper_Brain_Mid_Frontal_Lobe	0.0266
	Fun_level_famton_astrocyte	0.0004		Fun_level_famton_astrocyte	0.0230
	Fun_level_famton_brain	0.4890		Fun_level_famton_brain	0.6620
	Fun_level_famton_CL:0000127	0.0001		Fun_level_famton_CL:0000127	~0
	Fun_level_famton_count	~0		Fun_level_famton_count	~0
	Fun_level_famton_neuronal_stem_cell	0.3280		Fun_level_famton_neuronal_stem_cell	0.6990
	Fun_level_famton_permssive	~0		Fun_level_famton_permssive	~0
	Fun_level_enhancerAtlas_Astrocyte_EP	~0		Fun_level_enhancerAtlas_Astrocyte_EP	~0
	Fun_level_enhancerAtlas_Cerebellum_EP	~0		Fun_level_enhancerAtlas_Cerebellum_EP	~0
	Fun_level_enhancerAtlas_ESC_neuron_EP	~0		Fun_level_enhancerAtlas_ESC_neuron_EP	~0
	Fun_level_gene_enhancer_links_brain_enhcenter	~0		Fun_level_gene_enhancer_links_brain_enhcenter	~0
	Fun_level_gene_enhancer_links_neurosph_enhcenter	~0		Fun_level_gene_enhancer_links_neurosph_enhcenter	~0
	Fun_level_hacer_T1	~0		Fun_level_hacer_T1	~0
	Fun_level_SE_ele	~0		Fun_level_SE_ele	~0

	Fun_level_SEA00101	~0
	Fun_level_nott_Astrocyte_enhancers	~0
	Fun_level_nott_Astrocyte_promoters	~0
	Fun_level_nott_H3K4me3_around_TSS	~0
	Fun_level_nott_Microglia_enhancers	~0
	Fun_level_nott_Microglia_promoters	~0
	Fun_level_nott_Neuronal_enhancers	~0
	Fun_level_nott_Neuronal_promoters	~0
	Fun_level_nott_Oligo_enhancers	~0
	Fun_level_nott_Oligo_promoters	~0
	Fun_level_nott_superEnhancer	1
	Fun_level_vista	~0
CTCF binding sites	Fun_level_ctcf	~0
	Fun_level_CTCF_observed_Brain	~0
DNase hypersensitive sites	Fun_level_DNaselClusterd	~0
	Fun_level_DnaseMaster	~0
	Fun_level_DNase-seq_observed_Brain	~0
	Fun_level_DNase-seq_observed_Neurosph	~0
Genomic segmentations from Gencode	Fun_level_EncodeAwgTfbsBroadNhaCtcf	~0
	Fun_level_EncodeRegTfbsClustered	~0
	Fun_level_gencode_CDS	~0
	Fun_level_gencode_exon	0.5440
	Fun_level_gencode_gene	~0
	Fun_level_gencode_Selenocysteine	0.5450
	Fun_level_gencode_start_codon	0.2450

	Fun_level_SEA00101	~0
	Fun_level_nott_Astrocyte_enhancers	~0
	Fun_level_nott_Astrocyte_promoters	~0
	Fun_level_nott_H3K4me3_around_TSS	~0
	Fun_level_nott_Microglia_enhancers	~0
	Fun_level_nott_Microglia_promoters	~0
	Fun_level_nott_Neuronal_enhancers	~0
	Fun_level_nott_Neuronal_promoters	~0
	Fun_level_nott_Oligo_enhancers	~0
	Fun_level_nott_Oligo_promoters	~0
	Fun_level_nott_superEnhancer	1
	Fun_level_vista	~0
CTCF binding sites	Fun_level_ctcf	~0
	Fun_level_CTCF_observed_Brain	~0
DNase hypersensitive sites	Fun_level_DNaselClusterd	~0
	Fun_level_DnaseMaster	~0
	Fun_level_DNase-seq_observed_Brain	~0
	Fun_level_DNase-seq_observed_Neurosph	~0
Genomic segmentations from Gencode	Fun_level_EncodeAwgTfbsBroadNhaCtcf	~0
	Fun_level_EncodeRegTfbsClustered	~0
	Fun_level_gencode_CDS	~0
	Fun_level_gencode_exon	~0
	Fun_level_gencode_gene	~0
	Fun_level_gencode_Selenocysteine	0.6280
	Fun_level_gencode_start_codon	~0

	Fun_level_gencode_stop_codon	~0		Fun_level_gencode_stop_codon	~0
	Fun_level_gencode_transcript	0.8590		Fun_level_gencode_transcript	0.8330
	Fun_level_gencode_UTR	~0		Fun_level_gencode_UTR	~0
	Fun_level_miRNA	~0		Fun_level_miRNA	~0
	Fun_level_non-codingRNAs	~0		Fun_level_non-codingRNAs	0.0010
Histone markers	Fun_level_ATAC-seq_observed_Brain	~0	Histone markers	Fun_level_ATAC-seq_observed_Brain	~0
	Fun_level_H2AFZ_imputed_Brain	~0		Fun_level_H2AFZ_imputed_Brain	~0
	Fun_level_EP300_imputed_Brain	~0		Fun_level_EP300_imputed_Brain	~0
	Fun_level_EP300_imputed_Neurosph	~0		Fun_level_EP300_imputed_Neurosph	~0
	Fun_level_H2AFZ_imputed_Neurosph	~0		Fun_level_H2AFZ_imputed_Neurosph	~0
	Fun_level_H2AFZ_observed_Brain	~0		Fun_level_H2AFZ_observed_Brain	~0
	Fun_level_H3k27ac	~0		Fun_level_H3k27ac	~0
	Fun_level_H3K27ac_imputed_Brain	~0		Fun_level_H3K27ac_imputed_Brain	~0
	Fun_level_H3K27ac_imputed_Neurosph	~0		Fun_level_H3K27ac_imputed_Neurosph	~0
	Fun_level_H3K27ac_observed_Brain	~0		Fun_level_H3K27ac_observed_Brain	~0
	Fun_level_H3K27ac_observed_Neurosph	~0		Fun_level_H3K27ac_observed_Neurosph	~0
	Fun_level_H3K27me3_imputed_Brain	~0		Fun_level_H3K27me3_imputed_Brain	~0
	Fun_level_H3K27me3_imputed_Neurosph	~0		Fun_level_H3K27me3_imputed_Neurosph	~0
	Fun_level_H3K27me3_observed_Brain	~0		Fun_level_H3K27me3_observed_Brain	~0
	Fun_level_H3k4me1	~0		Fun_level_H3k4me1	~0
	Fun_level_H3K4me1_imputed_Brain	~0		Fun_level_H3K4me1_imputed_Brain	~0
	Fun_level_H3K4me1_imputed_Neurosph	~0		Fun_level_H3K4me1_imputed_Neurosph	~0
	Fun_level_H3K4me1_observed_Brain	~0		Fun_level_H3K4me1_observed_Brain	~0
	Fun_level_H3K4me1_observed_Neurosph	~0		Fun_level_H3K4me1_observed_Neurosph	~0
	Fun_level_H3K4me2_observed_Brain	~0		Fun_level_H3K4me2_observed_Brain	~0

	Fun_level_H3k4me3	~0		Fun_level_H3k4me3	~0
	Fun_level_H3K4me3_imputed_Brain	0.0580		Fun_level_H3K4me3_imputed_Brain	0.4680
	Fun_level_H3K4me3_imputed_Neurosph	~0		Fun_level_H3K4me3_imputed_Neurosph	~0
	Fun_level_H3K4me3_observed_Brain	0.0592		Fun_level_H3K4me3_observed_Brain	0.4710
	Fun_level_H3K4me3_observed_Neurosph	~0		Fun_level_H3K4me3_observed_Neurosph	~0
	Fun_level_H3K9ac_imputed_Brain	~0		Fun_level_H3K9ac_imputed_Brain	~0
	Fun_level_H3K9ac_imputed_Neurosph	~0		Fun_level_H3K9ac_imputed_Neurosph	~0
	Fun_level_H3K9me3_imputed_Brain	~0		Fun_level_H3K9me3_imputed_Brain	~0
	Fun_level_H3K9me3_imputed_Neurosph	~0		Fun_level_H3K9me3_imputed_Neurosph	~0
	Fun_level_H3K9me3_observed_Brain	~0		Fun_level_H3K9me3_observed_Brain	~0
	Fun_level_H3K9me3_observed_Neurosph	~0		Fun_level_H3K9me3_observed_Neurosph	~0
	Fun_level_H4K20me1_imputed_Neurosph	~0		Fun_level_H4K20me1_imputed_Neurosph	~0
	Fun_level_H4K20me1_observed_Brain	~0		Fun_level_H4K20me1_observed_Brain	~0
	Fun_level_POLR2A_imputed_Neurosph	~0		Fun_level_POLR2A_imputed_Neurosph	~0
	Fun_level_RAD21_imputed_Brain	~0		Fun_level_RAD21_imputed_Brain	~0
	Fun_level_RAD21_imputed_Neurosph	~0		Fun_level_RAD21_imputed_Neurosph	~0
	Fun_level_SMC3_imputed_Brain	~0		Fun_level_SMC3_imputed_Brain	~0
	Fun_level_SMC3_imputed_Neurosph	~0		Fun_level_SMC3_imputed_Neurosph	~0
Long range probable genes	Fun_level_liu_csbj_targetgene	~0	Long range probable genes	Fun_level_liu_csbj_targetgene	~0
Methylation	Fun_level_methMCRF	~0	Methylation	Fun_level_methMCRF	~0
Loop anchors and topological	Fun_level_PsychENCODE_CBC_H3K27ac	~0	Loop anchors and topological	Fun_level_PsychENCODE_CBC_H3K27ac	~0
	Fun_level_PsychENCODE_HiC_EP	~0		Fun_level_PsychENCODE_HiC_EP	~0
	Fun_level_PsychENCODE_loops_interRegion	~0		Fun_level_PsychENCODE_loops_interRegion	0.1890

associated domains in higher-order chromatin structure	Fun_level_PsychENCODE_PEC_Enhancers	~0	associated domains in higher-order chromatin structure	Fun_level_PsychENCODE_PEC_Enhancers	~0
	Fun_level_PsychENCODE_PFC_H3K27ac	~0		Fun_level_PsychENCODE_PFC_H3K27ac	~0
	Fun_level_PsychENCODE_TAR	~0		Fun_level_PsychENCODE_TAR	~0
	Fun_level_PsychENCODE_TC_H3K27ac	~0		Fun_level_PsychENCODE_TC_H3K27ac	~0
	Fun_level_TAD56	~0		Fun_level_TAD56	~0
DNase hypersensitive sites	Fun_level_RoadmapDNasePromCount	~0	DNase hypersensitive sites	Fun_level_RoadmapDNasePromCount	~0
Transcript factor binding sites from snp-selex	Fun_level_snp_selex	0.0860	Transcript factor binding sites from snp-selex	Fun_level_snp_selex	0.0403
Transcript starting sites	Fun_level_tss2000bp	~0	Transcript starting sites	Fun_level_tss2000bp	0.1450
Higher-order chromatin structure from Yue lab	Fun_level_yue_loops_hippo	~0	Higher-order chromatin structure from Yue lab	Fun_level_yue_loops_hippo	~0
Gene level	*The Gen_level_significant features are 34 out of 45*		**Gene level**	*The Gen_level_significant features are 25 out of 45*	
ClinGen curated	Gen_level_ClinGen_haploinsufficiency_gene_0	0.0013	ClinGen curated	Gen_level_ClinGen_region_curation_Triplosensitivity_0	0.4450

genes and genomic regions	Gen_level_ClinGen_haploinsufficiency_gene_1	0.0086	genes and genomic regions	Gen_level_ClinGen_region_curation_Triplosensitivity_1	0.4030
	Gen_level_ClinGen_haploinsufficiency_gene_2	0.5950		Gen_level_ClinGen_region_curation_Triplosensitivity_2	0.0603
	Gen_level_ClinGen_haploinsufficiency_gene_3	~0		Gen_level_ClinGen_region_curation_Triplosensitivity_3	0.3880
	Gen_level_ClinGen_haploinsufficiency_gene_30	~0		Gen_level_ClinGen_region_curation_Triplosensitivity_40	~0
	Gen_level_ClinGen_haploinsufficiency_gene_40	0.7010		Gen_level_ClinGen_triplosensitivity_gene	0.7800
	Gen_level_ClinGen_region_curation_Haploinsufficiency_0	0.8020		Gen_level_ClinGen_triplosensitivity_gene_0	~0
	Gen_level_ClinGen_region_curation_Haploinsufficiency_1	0.8830		Gen_level_ClinGen_triplosensitivity_gene_1	0.9070
	Gen_level_ClinGen_region_curation_Haploinsufficiency_2	0.8130		Gen_level_ClinGen_triplosensitivity_gene_2	0.9410
	Gen_level_ClinGen_region_curation_Haploinsufficiency_3	0.0011		Gen_level_ClinGen_triplosensitivity_gene_3	1.0000
	Gen_level_ClinGen_region_curation_Haploinsufficiency_30	0.9290		Gen_level_ClinGen_triplosensitivity_gene_30	1.0000
	Gen_level_ClinGen_region_curation_Haploinsufficiency_40	~0		Gen_level_ClinGen_triplosensitivity_gene_40	1.0000
	Gen_level_loss_of_function_score1	0.0026		Gen_level_gain_activating_score1	0.8030
	Gen_level_loss_of_function_score2	~0		Gen_level_gain_activating_score2	0.6810
	Gen_level_loss_of_function_score3	~0		Gen_level_gain_activating_score3	0.2840

Dosage sensitive genes	Gen_level_Collins_rCNV_PLIgenes_PHI	~0	Dosage sensitive genes	Gen_level_Collins_rCNV_PLIgenes_PTS	~0
DDG2P database	Gen_level_ddg2p_loss	~0	DDG2P database	Gen_level_ddg2p_gain	0.0690
Cell essential and nonessential genes	Gen_level_Essential_in_culture_CRISPR	0.0001	Cell essential and nonessential genes	Gen_level_Essential_in_culture_CRISPR	~0
	Gen_nonEssential_in_culture_CRISPR	0.3020		Gen_nonEssential_in_culture_CRISPR	0.6100
FDA proved drug target	Gen_level_FDA-approved_drug_targets	0.0013	FDA proved drug target	Gen_level_FDA-approved_drug_targets	~0
G protein-coupled receptor	Gen_level_gpcr_union	0.4010	G protein-coupled receptor	Gen_level_gpcr_union	0.0003
Neurodevelopmental process related genes	Gen_level_HP_0000707	~0	Neurodevelopmental process related genes	Gen_level_HP_0000707	~0
	Gen_level_HP_0000708	~0		Gen_level_HP_0000708	~0
	Gen_level_HP_0000717	0.0008		Gen_level_HP_0000717	0.0032
	Gen_level_HP_0000729	~0		Gen_level_HP_0000729	~0
	Gen_level_HP_0000752	~0		Gen_level_HP_0000752	~0
	Gen_level_HP_0001197	~0		Gen_level_HP_0001197	~0
	Gen_level_HP_0001250	~0		Gen_level_HP_0001250	~0
	Gen_level_HP_0001507	~0		Gen_level_HP_0001507	~0
	Gen_level_HP_0002011	~0		Gen_level_HP_0002011	~0
	Gen_level_HP_0002715	~0		Gen_level_HP_0002715	~0

	Gen_level_HP_0002960	0.7040		Gen_level_HP_0002960	0.3470
	Gen_level_HP_0011446	~0		Gen_level_HP_0011446	~0
	Gen_level_HP_0012443	~0		Gen_level_HP_0012443	~0
	Gen_level_HP_0012638	~0		Gen_level_HP_0012638	~0
	Gen_level_HP_0012639	~0		Gen_level_HP_0012639	~0
	Gen_level_HP_0012759	~0		Gen_level_HP_0012759	~0
	Gen_level_HP_0025031	~0		Gen_level_HP_0025031	~0
	Gen_level_HP_0031466	~0		Gen_level_HP_0031466	~0
	Gen_level_HP_0100022	~0		Gen_level_HP_0100022	~0
	Gen_level_HP_0100753	0.1320		Gen_level_HP_0100753	0.8250
	Gen_level_HP_0100852	~0		Gen_level_HP_0100852	0.0009
Mouse heterozygous LoF lethal	Gen_level_mgi_essential_gene	~0	Mouse heterozygous LoF lethal	Gen_level_mgi_essential_gene	~0
Olfactory receptors	Gen_level_Olfactory_receptors_mainland	0.0066	Olfactory receptors	Gen_level_Olfactory_receptors_mainland	0.6100
Sfari gene	Gen_level_sfari_gene	~0	Sfari gene	Gen_level_sfari_gene	~0
Sequence level	*The Seq_level_significant features are 7 out of 7*		**Sequence level**	*The Seq_level_significant features are 6 out of 7*	
Blacklisted regions	Seq_level_DacMapExclude	~0	Blacklisted regions	Seq_level_DacMapExclude	~0
Sfari gene	Seq_level_DukeMapExclude	~0	Sfari gene	Seq_level_DukeMapExclude	~0
GC content	Seq_level_GC	~0	GC content	Seq_level_GC	~0
Human	Seq_level_HAR	0.0085	Human	Seq_level_HAR	0.2010

accelerated regions Heterochromatin positions			accelerated regions Heterochromatin positions		
Human accelerated regions Heterochromatin positions Cross species conservation score	Seq_level_HetDomain	~0	Human accelerated regions Heterochromatin positions Cross species conservation score	Seq_level_HetDomain	~0
	Seq_level_phastCons46way	~0		Seq_level_phastCons46way	~0
Human accelerated regions	Seq_level_phyloP46way	~0	Human accelerated regions	Seq_level_phyloP46way	~0

p-values: when $P < 1{\times}10^{-4}$, it is shown as ~0, and when $1{\times}10^{-4} < P < 1$, it is shown as a decimal mode.

Table S3. Feature importancy. All the feature names were shown as feature names (original sources)_tissue type (if applicable). p-values: when $P < 1\times10^{-4}$, it is shown as ~0, and when $1\times10^{-4} < P < 1$, it is shown as a decimal mode. $P = 0.05$ is set as the significant level.

Features in copy number loss model	**Feature importancy**	**Features in copy number gain model**	**Feature importancy**
1_TssA_chromHMM_brain	0.0053	1_TssA_chromHMM_brain	0.0056
10_EnhA2_chromHMM_brain	0.0078	10_EnhA2_chromHMM_brain	0.0049
11_EnhWk_chromHMM_brain	0.0138	11_EnhWk_chromHMM_brain	0.0093
12_ZNF_chromHMM_brain	0.0046	12_ZNF_chromHMM_brain	0.0049
13_Het_chromHMM_brain	0.0058	13_Het_chromHMM_brain	0.0058
14_TssBiv_chromHMM_brain	0.0043	14_TssBiv_chromHMM_brain	0.0041
15_EnhBiv_chromHMM_brain	0.0057	15_EnhBiv_chromHMM_brain	0.0062
16_ReprPC_chromHMM_brain	0.0043	16_ReprPC_chromHMM_brain	0.0041
17_ReprPCWk_chromHMM_brain	0.0054	17_ReprPCWk_chromHMM_brain	0.0055
18_Quies_chromHMM_brain	0.0062	18_Quies_chromHMM_brain	0.0073
2_TssFlnk_chromHMM_brain	0.0045	2_TssFlnk_chromHMM_brain	0.0043
3_TssFlnkU_chromHMM_brain	0.0047	3_TssFlnkU_chromHMM_brain	0.0046
4_TssFlnkD_chromHMM_brain	0.0073	4_TssFlnkD_chromHMM_brain	0.0053
5_Tx_chromHMM_brain	0.0046	5_Tx_chromHMM_brain	0.0051
6_TxWk_chromHMM_brain	0.0050	6_TxWk_chromHMM_brain	0.0048
7_EnhG1_chromHMM_brain	0.0045	7_EnhG1_chromHMM_brain	0.0044
8_EnhG2_chromHMM_brain	0.0047	8_EnhG2_chromHMM_brain	0.0054
9_EnhA1_chromHMM_brain	0.0055	9_EnhA1_chromHMM_brain	0.0048
ATAC-seq_observed_Brain	0.0058	ATAC-seq_observed_Brain	0.0060
Brain_Angular_Gyrus_dbsuper	0.0042	Brain_Angular_Gyrus_dbsuper	0.0035
Brain_Anterior_Caudate_dbsuper	0.0043	Brain_Anterior_Caudate_dbsuper	0.0051
Brain_Cingulate_Gyrus_dbsuper	0.0040	Brain_Cingulate_Gyrus_dbsuper	0.0039

Brain_Hippocampus_Middle_150_dbsuper	0.0045	Brain_Hippocampus_Middle_150_dbsuper	0.0063
Brain_Hippocampus_Middle_dbsuper	0.0039	Brain_Hippocampus_Middle_dbsuper	0.0042
Brain_Inferior_Temporal_Lobe_dbsuper	0.0035	Brain_Inferior_Temporal_Lobe_dbsuper	0.0054
Brain_Mid_Frontal_Lobe_dbsuper	0.0050	Brain_Mid_Frontal_Lobe_dbsuper	0.0070
ClinGen_haploinsufficiency_gene_0	0.0037	ClinGen_region_curation_Triplosensitivity_0	0.0022
ClinGen_haploinsufficiency_gene_1	0.0059	ClinGen_region_curation_Triplosensitivity_1	0.0071
ClinGen_haploinsufficiency_gene_2	~0	ClinGen_region_curation_Triplosensitivity_2	0.0061
ClinGen_haploinsufficiency_gene_3	0.0094	ClinGen_region_curation_Triplosensitivity_3	0.0054
ClinGen_haploinsufficiency_gene_30	0.0048	ClinGen_region_curation_Triplosensitivity_40	0.0101
ClinGen_haploinsufficiency_gene_40	0.0011	ClinGen_triplosensitivity_gene	~0
ClinGen_region_curation_Haploinsufficiency_0	0.0046	ClinGen_triplosensitivity_gene_0	0.0062
ClinGen_region_curation_Haploinsufficiency_1	~0	ClinGen_triplosensitivity_gene_1	~0
ClinGen_region_curation_Haploinsufficiency_2	0.0026	ClinGen_triplosensitivity_gene_2	~0
ClinGen_region_curation_Haploinsufficiency_3	0.0048	ClinGen_triplosensitivity_gene_3	~0
ClinGen_region_curation_Haploinsufficiency_30	~0	ClinGen_triplosensitivity_gene_30	~0
ClinGen_region_curation_Haploinsufficiency_40	0.0105	ClinGen_triplosensitivity_gene_40	~0
Collins_rCNV_PLIgenes_PHI	0.0524	Collins_rCNV_PLIgenes_PTS	0.0653
ctcf	0.0053	ctcf	0.0045
CTCF_observed_Brain	0.0043	CTCF_observed_Brain	0.0051
DacMapExclude	0.0057	DacMapExclude	0.0089
ddg2p_loss	0.0075	ddg2p_gain	0.0035
DNaselClusterd	0.0042	DNaselClusterd	0.0045
DnaseMaster	0.0055	DnaseMaster	0.0050
DNase-seq_observed_Brain	0.0111	DNase-seq_observed_Brain	0.0108
DNase-seq_observed_Neurosph	0.0230	DNase-seq_observed_Neurosph	0.0224

DukeMapExclude	0.0063	DukeMapExclude	0.0085
EncodeAwgTfbsBroadNhaCtcf	0.0051	EncodeAwgTfbsBroadNhaCtcf	0.0048
EncodeRegTfbsClustered	0.0049	EncodeRegTfbsClustered	0.0056
enhancerAtlas_Astrocyte_EP	0.0034	enhancerAtlas_Astrocyte_EP	0.0068
enhancerAtlas_Cerebellum_EP	0.0042	enhancerAtlas_Cerebellum_EP	0.0049
enhancerAtlas_ESC_neuron_EP	0.0033	enhancerAtlas_ESC_neuron_EP	0.0045
EP300_imputed_Brain	0.0048	EP300_imputed_Brain	0.0040
EP300_imputed_Neurosph	0.0051	EP300_imputed_Neurosph	0.0055
Essential_in_culture_CRISPR	0.0039	Essential_in_culture_CRISPR	0.0051
famton_astrocyte	0.0039	famton_astrocyte	0.0070
famton_brain	0.0067	famton_brain	0.0053
famton_CL:0000127	0.0035	famton_CL:0000127	0.0038
famton_count	0.0054	famton_count	0.0062
famton_neuronal_stem_cell	0.0046	famton_neuronal_stem_cell	0.0032
famton_permssive	0.0046	famton_permssive	0.0050
FDA-approved_drug_targets	0.0040	FDA-approved_drug_targets	0.0046
GC	0.0065	gain_activating_score1	~0
gencode_CDS	0.0092	gain_activating_score2	~0
gencode_exon	0.0077	gain_activating_score3	0.0018
gencode_gene	0.0074	GC	0.0059
gencode_Selenocysteine	~0	gencode_CDS	0.0096
gencode_start_codon	0.0208	gencode_exon	0.0211
gencode_stop_codon	0.0035	gencode_gene	0.0072
gencode_transcript	0.0047	gencode_Selenocysteine	0.0005
gencode_UTR	0.0058	gencode_start_codon	0.0086

gene_enhancer_links_brain_enhcenter	0.0056	gencode_stop_codon	0.0043
gene_enhancer_links_neurosph_enhcenter	0.0064	gencode_transcript	0.0051
gpcr_union	0.0040	gencode_UTR	0.0043
H2AFZ_imputed_Brain	0.0043	gene_enhancer_links_brain_enhcenter	0.0043
H2AFZ_imputed_Neurosph	0.0047	gene_enhancer_links_neurosph_enhcenter	0.0061
H2AFZ_observed_Brain	0.0052	gpcr_union	0.0035
H3k27ac	0.0049	H2AFZ_imputed_Brain	0.0045
H3K27ac_imputed_Brain	0.0046	H2AFZ_imputed_Neurosph	0.0060
H3K27ac_imputed_Neurosph	0.0051	H2AFZ_observed_Brain	0.0056
H3K27ac_observed_Brain	0.0045	H3k27ac	0.0054
H3K27ac_observed_Neurosph	0.0058	H3K27ac_imputed_Brain	0.0063
H3K27me3_imputed_Brain	0.0050	H3K27ac_imputed_Neurosph	0.0066
H3K27me3_imputed_Neurosph	0.0065	H3K27ac_observed_Brain	0.0047
H3K27me3_observed_Brain	0.0067	H3K27ac_observed_Neurosph	0.0062
H3k4me1	0.0061	H3K27me3_imputed_Brain	0.0051
H3K4me1_imputed_Brain	0.0041	H3K27me3_imputed_Neurosph	0.0064
H3K4me1_imputed_Neurosph	0.0053	H3K27me3_observed_Brain	0.0077
H3K4me1_observed_Brain	0.0056	H3k4me1	0.0072
H3K4me1_observed_Neurosph	0.0043	H3K4me1_imputed_Brain	0.0057
H3K4me2_observed_Brain	0.0052	H3K4me1_imputed_Neurosph	0.0060
H3k4me3	0.0077	H3K4me1_observed_Brain	0.0058
H3K4me3_imputed_Brain	0.0061	H3K4me1_observed_Neurosph	0.0053
H3K4me3_imputed_Neurosph	0.0051	H3K4me2_observed_Brain	0.0053
H3K4me3_observed_Brain	0.0054	H3k4me3	0.0053
H3K4me3_observed_Neurosph	0.0054	H3K4me3_imputed_Brain	0.0053

H3K9ac_imputed_Brain	0.0055	H3K4me3_imputed_Neurosph	0.0041
H3K9ac_imputed_Neurosph	0.0055	H3K4me3_observed_Brain	0.0068
H3K9me3_imputed_Brain	0.0057	H3K4me3_observed_Neurosph	0.0065
H3K9me3_imputed_Neurosph	0.0059	H3K9ac_imputed_Brain	0.0050
H3K9me3_observed_Brain	0.0055	H3K9ac_imputed_Neurosph	0.0069
H3K9me3_observed_Neurosph	0.0047	H3K9me3_imputed_Brain	0.0068
H4K20me1_imputed_Neurosph	0.0044	H3K9me3_imputed_Neurosph	0.0066
H4K20me1_observed_Brain	0.0050	H3K9me3_observed_Brain	0.0066
hacer_T1	0.0051	H3K9me3_observed_Neurosph	0.0079
HAR	0.0044	H4K20me1_imputed_Neurosph	0.0055
HetDomain	0.0113	H4K20me1_observed_Brain	0.0054
HP_0000707	0.0034	hacer_T1	0.0063
HP_0000708	0.0047	HAR	0.0053
HP_0000717	0.0092	HetDomain	0.0069
HP_0000729	0.0031	HP_0000707	0.0041
HP_0000752	0.0032	HP_0000708	0.0046
HP_0001197	0.0047	HP_0000717	0.0045
HP_0001250	0.0038	HP_0000729	0.0053
HP_0001507	0.0042	HP_0000752	0.0056
HP_0002011	0.0056	HP_0001197	0.0045
HP_0002715	0.0077	HP_0001250	0.0026
HP_0002960	0.0090	HP_0001507	0.0058
HP_0011446	0.0053	HP_0002011	0.0043
HP_0012443	0.0068	HP_0002715	0.0045
HP_0012638	0.0034	HP_0002960	0.0040

HP_0012639	0.0051	HP_0011446	0.0063
HP_0012759	0.0074	HP_0012443	0.0028
HP_0025031	0.0045	HP_0012638	0.0045
HP_0031466	0.0055	HP_0012639	0.0029
HP_0100022	0.0065	HP_0012759	0.0058
HP_0100753	~0	HP_0025031	0.0079
HP_0100852	0.0043	HP_0031466	0.0058
liu_csbj_targetgene	0.0052	HP_0100022	0.0035
loss_of_function_score1	0.0059	HP_0100753	0.0046
loss_of_function_score2	0.0033	HP_0100852	0.0052
loss_of_function_score3	0.0055	liu_csbj_targetgene	0.0050
methMCRF	0.0065	methMCRF	0.0075
mgi_essential_gene	0.0044	mgi_essential_gene	0.0139
miRNA	0.0049	miRNA	0.0053
non-codingRNAs	0.0058	non-codingRNAs	0.0049
nonEssential_in_culture_CRISPR	0.0047	nonEssential_in_culture_CRISPR	0.0043
nott_Astrocyte_enhancers	0.0040	nott_Astrocyte_enhancers	0.0051
nott_Astrocyte_promoters	0.0045	nott_Astrocyte_promoters	0.0063
nott_H3K4me3_around_TSS	0.0060	nott_H3K4me3_around_TSS	0.0056
nott_Microglia_enhancers	0.0047	nott_Microglia_enhancers	0.0051
nott_Microglia_promoters	0.0057	nott_Microglia_promoters	0.0052
nott_Neuronal_enhancers	0.0123	nott_Neuronal_enhancers	0.0109
nott_Neuronal_promoters	0.0047	nott_Neuronal_promoters	0.0054
nott_Oligo_enhancers	0.0049	nott_Oligo_enhancers	0.0058
nott_Oligo_promoters	0.0061	nott_Oligo_promoters	0.0048

nott_superEnhancer	~0	nott_superEnhancer	~0
Olfactory_receptors_mainland	0.0068	Olfactory_receptors_mainland	0.0025
phastCons46way	0.0062	phastCons46way	0.0070
phyloP46way	0.0051	phyloP46way	0.0057
POLR2A_imputed_Neurosph	0.0067	POLR2A_imputed_Neurosph	0.0043
PsychENCODE_CBC_H3K27ac	0.0067	PsychENCODE_CBC_H3K27ac	0.0049
PsychENCODE_HiC_EP	0.0050	PsychENCODE_HiC_EP	0.0046
PsychENCODE_loops_interRegion	0.0043	PsychENCODE_loops_interRegion	0.0043
PsychENCODE_PEC_Enhancers	0.0104	PsychENCODE_PEC_Enhancers	0.0079
PsychENCODE_PFC_H3K27ac	0.0062	PsychENCODE_PFC_H3K27ac	0.0047
PsychENCODE_TAR	0.0058	PsychENCODE_TAR	0.0052
PsychENCODE_TC_H3K27ac	0.0047	PsychENCODE_TC_H3K27ac	0.0047
RAD21_imputed_Brain	0.0055	RAD21_imputed_Brain	0.0061
RAD21_imputed_Neurosph	0.0057	RAD21_imputed_Neurosph	0.0060
RoadmapDNasePromCount	0.0048	RoadmapDNasePromCount	0.0049
SE_ele	0.0043	SE_ele	0.0052
SEA00101	0.0046	SEA00101	0.0056
sfari_gene	0.0046	sfari_gene	0.0043
SMC3_imputed_Brain	0.0060	SMC3_imputed_Brain	0.0054
SMC3_imputed_Neurosph	0.0047	SMC3_imputed_Neurosph	0.0069
snp_selex	0.0084	snp_selex	0.0045
TAD56	0.0079	TAD56	0.0072
tss2000bp	0.0178	tss2000bp	0.0146
vista	0.0047	vista	0.0045
yue_loops_hippo	0.0051	yue_loops_hippo	0.0066

References

1. Hart T, Tong AHY, Chan K, et al. Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3 (Bethesda)* 2017;7:2719-27. doi: 10.1534/g3.117.041277
2. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581:434-43. doi: 10.1038/s41586-020-2308-7
3. Strande NT, Riggs ER, Buchanan AH, et al. Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. *Am J Hum Genet* 2017;100:895-906. doi: 10.1016/j.ajhg.2017.04.015
4. Wright CF, Fitzgerald TW, Jones WD, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* 2015;385:1305-14. doi: 10.1016/S0140-6736(14)61705-0
5. Collins RL, Glessner JT, Porcu E, et al. A cross-disorder dosage sensitivity map of the human genome. *Cell* 2022;185:3041-3055.e25. doi: 10.1101/2021.01.26.21250098
6. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;46:D1074-D82. doi: 10.1093/nar/gkx1037
7. Motenko H, Neuhauser SB, O'Keefe M, et al. MouseMine: a new data warehouse for MGI. *Mamm Genome* 2015;26:325-30. doi: 10.1007/s00335-015-9573-z
8. Kohler S, Gargano M, Matentzoglu N, et al. The Human Phenotype Ontology in 2021. *Nucleic Acids Res* 2021;49:D1207-D17. doi: 10.1093/nar/gkaa1043
9. Braschi B, Denny P, Gray K, et al. Genenames.org: the HGNC and VGNC resources in 2019. *Nucleic Acids Res* 2019;47:D786-D92. doi: 10.1093/nar/gky930
10. Abrahams BS, Arking DE, Campbell DB, et al. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism* 2013;4:36. doi: 10.1186/2040-2392-4-36
11. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012;9:215-6. doi: 10.1038/nmeth.1906
12. Boix CA, James BT, Park YP, et al. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* 2021;590:300-07. doi: 10.1038/s41586-020-03145-z
13. Sabo PJ, Hawrylycz M, Wallace JC, et al. Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc Natl Acad Sci U.S.A.* 2004;101:16837-42. doi: 10.1073/pnas.0407387101
14. Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* 2018;46:D794-D801. doi: 10.1093/nar/gkx1081

15. Roadmap Epigenomics C, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518:317-30. doi: 10.1038/nature14248

16. Nott A, Holtman IR, Coufal NG, et al. Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science* 2019;366:1134-39. doi: 10.1126/science.aay0793

17. Khan A, Zhang X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res* 2016;44:D164-71. doi: 10.1093/nar/gkv1002

18. Gao T, Qian J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res* 2020;48:D58-D64. doi: 10.1093/nar/gkz980

19. Andersson R, Gebhard C, Miguel-Escalada I, et al. An atlas of active enhancers across human cell types and tissues. *Nature* 2014;507:455-61. doi: 10.1038/nature12787

20. Wang J, Dai X, Berry LD, et al. HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res* 2019;47:D106-D12. doi: 10.1093/nar/gky864

21. Wang D, Liu S, Warrell J, et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* 2018;362:eaat8464. doi: 10.1126/science.aat8464

22. Jiang Y, Qian F, Bai X, et al. SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res* 2019;47:D235-D43. doi: 10.1093/nar/gky1025

23. Chen C, Zhou D, Gu Y, et al. SEA version 3.0: a comprehensive extension and update of the Super-Enhancer archive. *Nucleic Acids Res* 2020;48:D198-D203. doi: 10.1093/nar/gkz1028

24. Visel A, Minovitsky S, Dubchak I, et al. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* 2007;35:D88-92. doi: 10.1093/nar/gkl822

25. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 2012;22:1760-74. doi: 10.1101/gr.135350.111

26. Liu X, Xu W, Leng F, et al. Prioritizing long range interactions in noncoding regions using GWAS and deletions perturbed TADs. *Comput Struct Biotechnol J* 2020;18:2945-52. doi: 10.1016/j.csbj.2020.10.014

27. Wang Y, Song F, Zhang B, et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol* 2018;19:151. doi: 10.1186/s13059-018-1519-9

28. Yan J, Qiu Y, Ribeiro Dos Santos AM, et al. Systematic analysis of binding of transcription factors to noncoding variants. *Nature* 2021;591:147-51. doi: 10.1038/s41586-021-03211-0

29. Karolchik D, Hinrichs AS, Furey TS, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 2004;32:D493-6. doi: 10.1093/nar/gkh103

30. Ho JW, Jung YL, Liu T, et al. Comparative analysis of metazoan chromatin organization. *Nature* 2014;512:449-52. doi: 10.1038/nature13415

31. Doan RN, Bae BI, Cubelos B, et al. Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. *Cell* 2016;167:341-54 e12. doi: 10.1016/j.cell.2016.08.071

32. Harding SD, Sharman JL, Faccenda E, et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. *Nucleic Acids Res* 2018;46:D1091-D106. doi: 10.1093/nar/gkx1121

33. Alexander SP, Christopoulos A, Davenport AP, et al. THE CONCISE GUIDE TO PHARMACOLOGY 2017/18: G protein-coupled receptors. *Br J Pharmacol* 2017;174 Suppl 1:S17-S129. doi: 10.1111/bph.13878

34. Collins RL, Glessner JT, Porcu E, et al. A cross-disorder dosage sensitivity map of the human genome. *Cell* 2022;185:3041-55.e25. doi: 10.1016/j.cell.2022.06.036

35. UniProt Consortium T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2018;46:2699. doi: 10.1093/nar/gky092

36. Howe KL, Achuthan P, Allen J, et al. Ensembl 2021. *Nucleic Acids Res* 2021;49:D884-D91. doi: 10.1093/nar/gkaa942