



FeatSNP: An Interactive Database for Brain-Specific Epigenetic Annotation of Human SNPs

Chun-yu Ma¹, Pamela Madden², Paul Gontarz¹, Ting Wang^{3*} and Bo Zhang^{1*}

OPEN ACCESS

Edited by:

Shandar Ahmad,
Jawaharlal Nehru University, India

Reviewed by:

Wei-Hua Chen,
Huazhong University of Science
and Technology, China
Sandeep Kumar Dhanda,
La Jolla Institute for Immunology (LJI),
United States

*Correspondence:

Ting Wang
twang@wustl.edu
Bo Zhang
bzhang29@wustl.edu

Specialty section:

This article was submitted to
Bioinformatics and Computational
Biology,
a section of the journal
Frontiers in Genetics

Received: 19 November 2018

Accepted: 08 March 2019

Published: 02 April 2019

Citation:

Ma C-y, Madden P, Gontarz P,
Wang T and Zhang B (2019)
FeatSNP: An Interactive Database
for Brain-Specific Epigenetic
Annotation of Human SNPs.
Front. Genet. 10:262.
doi: 10.3389/fgene.2019.00262

¹ Center of Regenerative Medicine, Department of Developmental Biology, Washington University School of Medicine, St. Louis, MO, United States, ² Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, United States, ³ Department of Genetics, The Edison Family Center for Genome Sciences & Systems Biology, Washington University School of Medicine, St. Louis, MO, United States

FeatSNP is an online tool and a curated database for exploring 81 million common SNPs' potential functional impact on the human brain. FeatSNP uses the brain transcriptomes of the human population to improve functional annotation of human SNPs by integrating transcription factor binding prediction, public eQTL information, and brain specific epigenetic landscape, as well as information of Topologically Associating Domains (TADs). FeatSNP supports both single and batched SNP searching, and its interactive user interface enables users to explore the functional annotations and generate publication-quality visualization results. FeatSNP is freely available on the internet at FeatSNP.org with all major web browsers supported.

Keywords: SNP, database, epigenetics, brain, transcription factor

INTRODUCTION

Genome-wide association studies (GWAS) and expression quantitative trait loci (eQTL) analyses have identified thousands of genetic variants that are associated with a wide range of human phenotypes, shedding lights on the understanding of the genetic effect to human diseases. However, a key challenge for scientists in the human genetics community is to understand the molecular mechanism connecting significant genetic variant and specific phenotype. More than 90% of SNPs associated with human phenotypes are located in non-protein-coding regions, and cannot be explained by alteration of amino acid sequence of proteins (Welter et al., 2014). Recently, mounting evidence suggests that disease-associated non-coding SNPs are highly enriched in tissue-specific regulatory elements including enhancers, which can be detected and defined by specific chromatin modifications (Carey et al., 2015;

Zhou et al., 2015; Agrawal et al., 2018). Moreover, some non-coding SNPs are found to be located within transcription factor (TF) binding motifs, which affect the TF binding affinity and result in allele switching and/or allele-specific regulation of target genes (Andersson et al., 2014; Roadmap Epigenomics et al., 2015; Nelson et al., 2016). These evidences underscore the potential causal role of non-coding genetic variants in affecting human diseases and phenotypes through regulation of gene expression (Claussnitzer et al., 2015).

Here we introduce FeatSNP, an online tool and database which provides an interactive user interface (UI) for inquiring brain-specific functional and epigenetic annotation of human SNPs. Unlike traditional SNP functional annotation databases, such as RegulomeDB (Boyle et al., 2012) and HaploReg (Ward and Kellis, 2012), FeatSNP focuses on the collection and curation of brain-specific functional genomics data, including epigenomes, transcriptomes, and eQTL data, to better annotate the regulatory potential of single SNP. Specifically, FeatSNP supplies a series of new features to facilitate research understanding the functional annotation of SNP on human brain (**Supplementary Table S1**). FeatSNP uses human brain transcriptomes to improve and refine the prediction of allele-specific TF binding motifs. The expression correlation between SNP-associated gene and predicted SNP-associated TFs was used to determine the best allele-associated TF candidate. The interactive UI allows the users easily to browse functional annotation and generate analysis results and high quality figures.

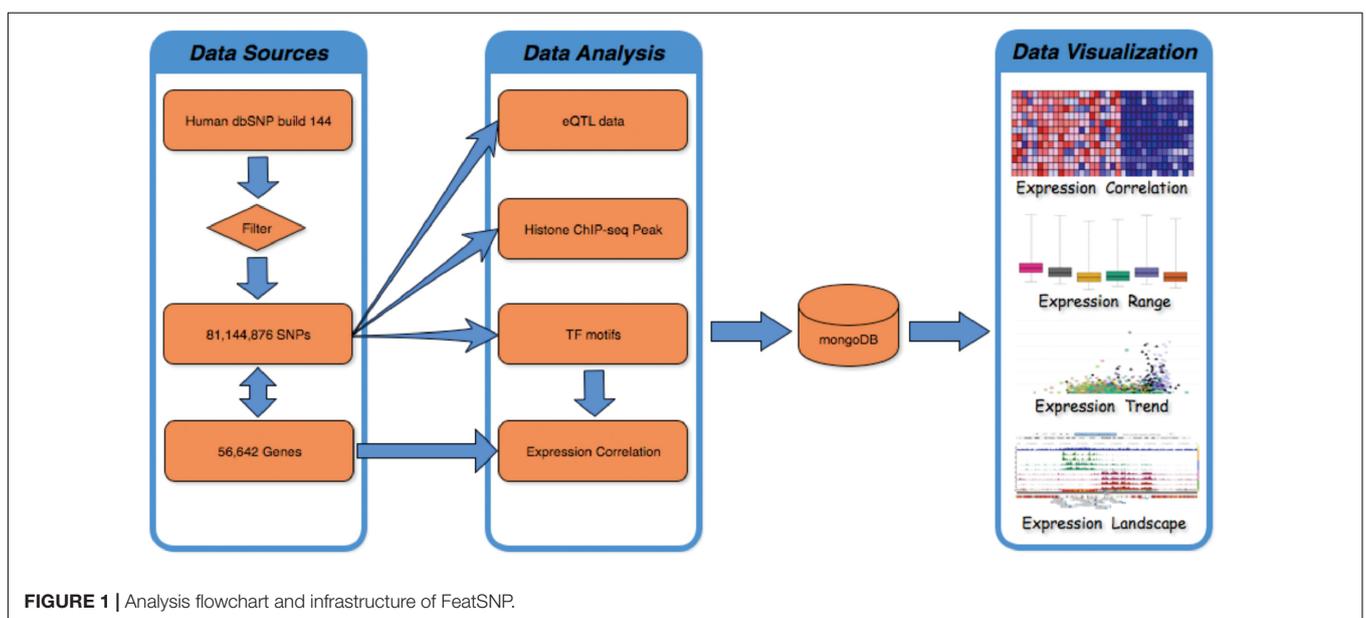
METHODS

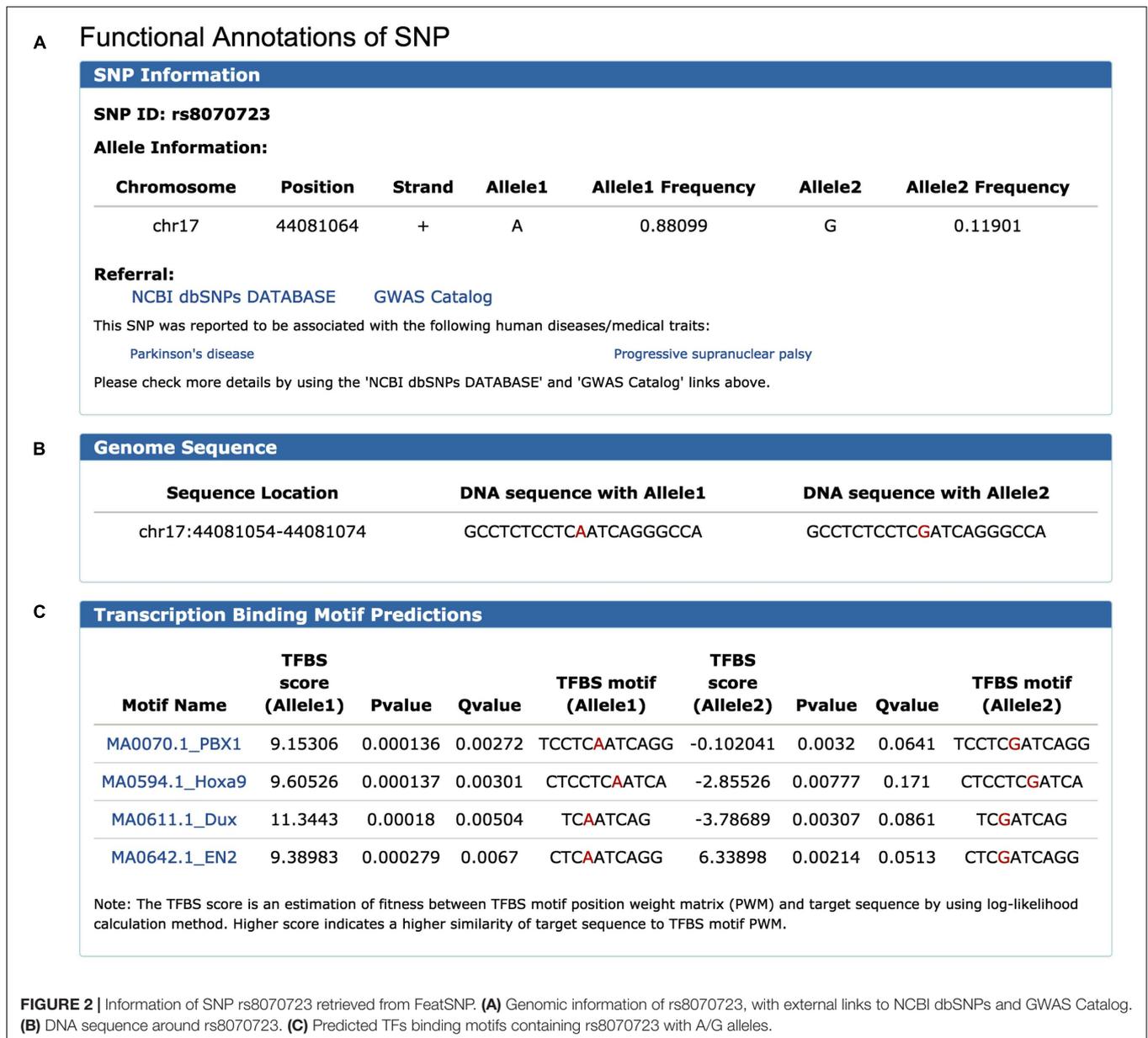
FeatSNP consists of a front end UI implemented with HTML/PHP/JavaScript, and a backend NoSQL database implemented with MongoDB (v3.2.7) as shown in **Figure 1**. The current SNP dataset contains 81,144,876 bi-allelic SNPs from

dbSNP (V144), with SNP accession number as unique identifier in the database. Human dbSNP build 144 was downloaded from <ftp.ncbi.nih.gov/snp>, which includes 84,435,229 SNPs records, 1,591,294 insertions records, 2,595,517 deletions records, 33,234 indel records, and 110 Multiple Nucleotide Polymorphisms (MNPs) records. After filtering redundant records, 81,144,876 of 84,435,229 biallelic SNPs were used to generate functional annotations and were curated by the FeatSNP database. The genome coordinates (hg19) of 81,144,876 SNPs were used to associate the SNPs with their nearest genes based on 56,642 records of GENCODE gene annotation Release 19 (GRCh37.p13).

To predict impact of allele-specific TF binding affinity by SNPs, the Position Weight Matrix (PWM) of 519 vertebrate TFs were collected from JASPAR (Core Vertebrate 2016) (Mathelier et al., 2016). After evaluating the motif weight PWM of 519 TFs at base-pair resolution (**Supplementary Figure S2**), the reference and alternate alleles for every SNP with flanking 10 bp of genomic sequences both upstream and downstream were obtained from the UCSC Genome Browser. FIMO (Grant et al., 2011) was used to scan the 21 bp sequence to identify binding motifs matching any of the 519 TF PWMs, and calculate the TFBS motif scores. Only instances where a motif in the sequence (i) passed the threshold of $P < 1e-2$ in either the reference or the alternate allele, and (ii) contained the SNP location and (iii) the difference of motif scores between the reference and the alternate allele was greater than 2, were recorded in the database.

1,259 transcriptome datasets of 13 brain tissues generated by the GTEx consortium (Gibson, 2015) were used to calculate the Pearson correlation between each SNP associated gene and predicted binding TFs. The lowly expressed gene and TFs (expression of all samples in one tissue less than 0.2RPKM) were removed. The correlation and gene expression in 13 brain tissues were visualized by using JavaScript package Highcharts (v5.0.2).





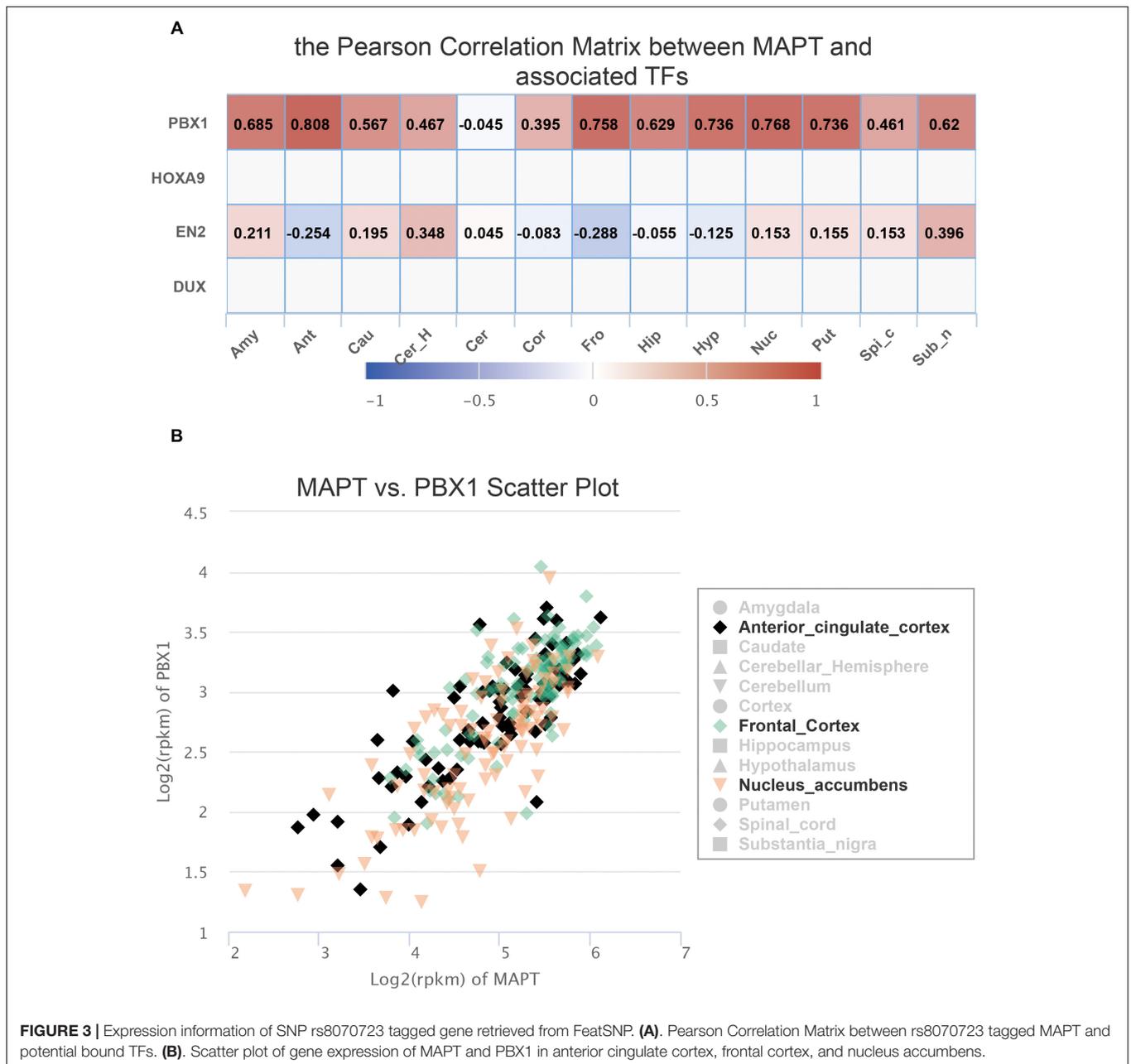
eQTL data of 10 brain tissues generated by GTEx consortium were negative-log₁₀ transformed and further visualized by using Highcharts (v5.0.2).

Histone modification ChIP-seq data of 10 brain tissues were downloaded from NIH Roadmap Epigenomics data portal. Bedtools was used to identify SNPs residing in peaks of 7 histone modification marks (H3K4me3, H3K36me3, H3K27me3, H3K4me1, H3K27ac, H3K9me3, and H3K9ac) that were identified by macs2 (Zhang et al., 2008) with default parameters. To enhance the user experience, the WashU epigenome browser (Zhou et al., 2015) was embedded in the UI to display epigenetic landscape in a 200 bp region surrounding each SNP. The browser also displays DNA methylation data (Whole Genome Bisulfite Sequencing) of 4 neuronal progenitor and

brain tissues generated by Roadmap Epigenomics Project, enhanced epilogos visualization¹ of all 127 epigenomes, and topologically associating domains (TAD) data of GM12878, IMR90, and Hap1 cell lines (Rao et al., 2014; Sanborn et al., 2015). eQTL data of 10 brain tissues generated by GTEx consortium were also visualized on the embedded WashU epigenome browser.

The association records of SNP and human disease/traits (V1.0.2) were downloaded from GWAS Catalog. 33,894 associations with *p*-value smaller than 5E-8 were kept and classified based on 1,374 human disease/traits categories. The functional annotations of these 33,894 SNPs were

¹epilogos.altiusinstitute.org



reported on FeatSNP.org/html_file/disease_classification.html (**Supplementary Figure S3**).

RESULTS

To illustrate the use of FeatSNP, we performed the analysis using rs8070723 as an example. rs8070723 is an intronic A/G SNP (major allele A frequency 0.881, minor allele G frequency 0.119) in *MAPT*, the gene that encodes the microtubule-associated protein tau, and is associated with Progressive Supranuclear Palsy (Hoglinger et al., 2011) and with Parkinson's Disease (UK Parkinson's Disease Consortium et al., 2011).

To better understand the regulatory potential of this human disease-associated SNP, we inquired the epigenetic annotation of rs8070723 in FeatSNP through Single SNP ID Searching function on SNP Query Page (**Supplementary Figure S4**). The database first reported the basic information of SNP rs8070723, including genomic location, allelic frequency, surrounding DNA sequence, and associated gene (**Figures 2A–C**). Users can further access the genetic information and associated human disease or traits of inquired SNPs on dbSNP and GWAS Catalog through external links.

FeatSNP found four potential TF binding motifs harboring rs8070723 with A allele, including *PBX1*, *Hoxa9*, *Dux*, and *EN2*. All four TF binding motifs had high TFBS scores in A allele,

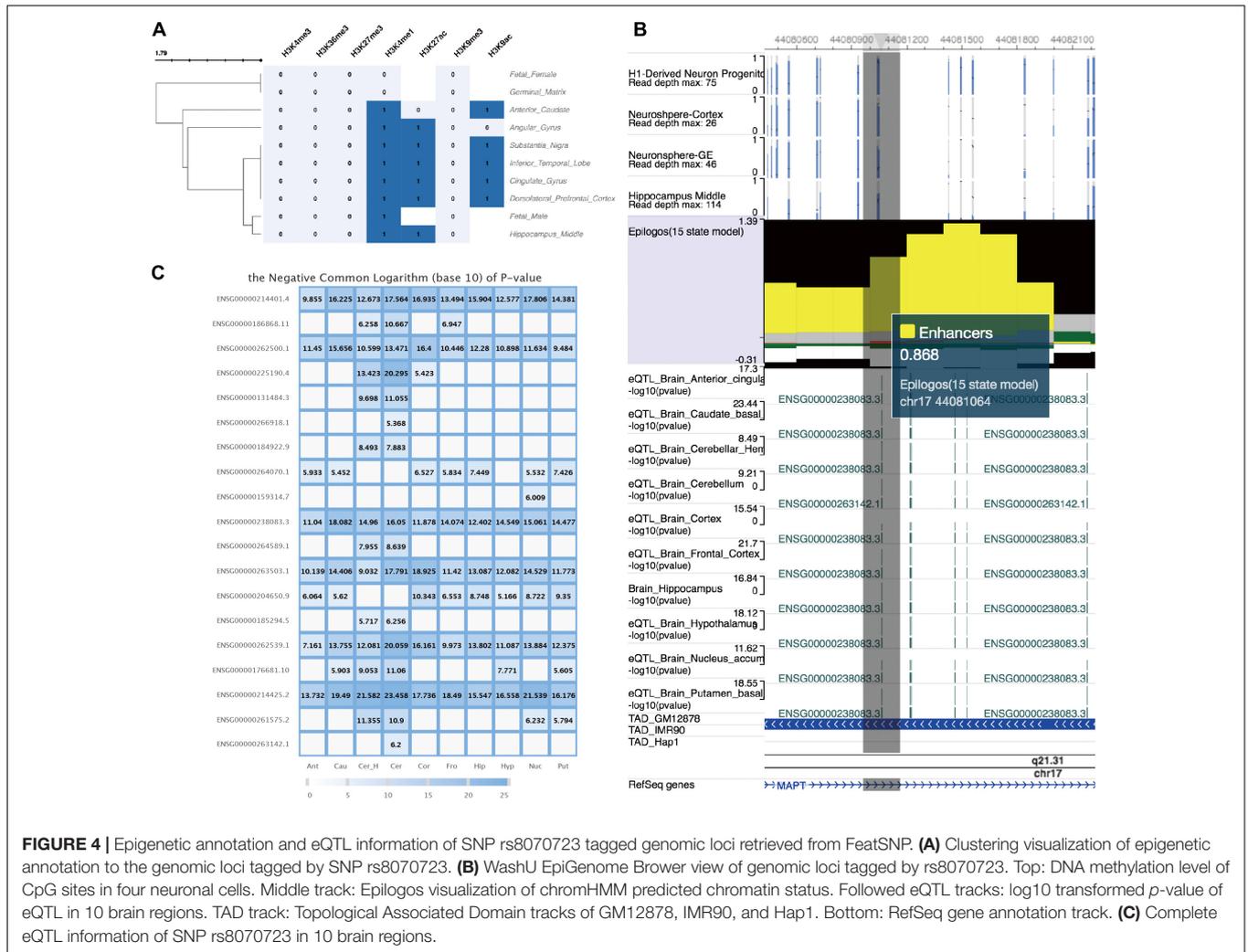


FIGURE 4 | Epigenetic annotation and eQTL information of SNP rs8070723 tagged genomic loci retrieved from FeatSNP. **(A)** Clustering visualization of epigenetic annotation to the genomic loci tagged by SNP rs8070723. **(B)** WashU EpiGenome Browser view of genomic loci tagged by rs8070723. Top: DNA methylation level of CpG sites in four neuronal cells. Middle track: Epilogos visualization of chromHMM predicted chromatin status. Followed eQTL tracks: log10 transformed p-value of eQTL in 10 brain regions. TAD track: Topological Associated Domain tracks of GM12878, IMR90, and Hap1. Bottom: RefSeq gene annotation track. **(C)** Complete eQTL information of SNP rs8070723 in 10 brain regions.

and the TFBS motifs were destroyed with G allele with low TFBS scores (Figure 2C). *PBX1* encodes a nuclear protein that belongs to the *PBX* homeobox family of transcriptional factors, and studies suggested *PBX1* regulates the patterning of the cerebral cortex (Golonzhka et al., 2015) and its transcriptional network controls dopaminergic neuron development in Parkinson’s disease (Villaescusa et al., 2016). *EN2* encodes homeodomain-containing proteins and has been implicated in the control of pattern formation during development of the central nervous system (Genestine et al., 2015). *Hoxa9* is an important homeobox transcription factor and plays important roles in myeloid leukemogenesis (Siriboonpiputtana et al., 2017). Dux-family transcription factors were recently identified to regulate zygotic genome activation in placental mammals (De Iaco et al., 2017). Thus, *PBX1* and *EN2* could be the potential master TFs affected by the SNP rs8070723.

Since FeatSNP curated 1,259 transcriptome data of 13 brain tissues generated by the GTEx consortium (Gibson, 2015), we were able to further check the expression level of *PBX1* and *EN2* in multiple brain regions in FeatSNP database. *EN2* was only expressed in the cerebellum of the

brain (Supplementary Figure S1A) and did not correlate with expression level of *MAPT* (Figure 3A). We found that *PBX1* highly expressed in different brain regions (Supplementary Figure S1B), and we also found the expression of *MAPT* had strong and specific correlation with *PBX1* in multiple brain regions (Figure 3A), especially in anterior cingulate cortex ($r = 0.808$), nucleus accumbens ($r = 0.768$), and frontal cortex ($r = 0.768$) (Figure 3B), which were considered as major affected regions of Progressive Supranuclear Palsy (Salmon et al., 1997).

We further explored the epigenetic annotation of the genomic regions tagged by rs8070723 in 10 brain regions by using epigenome data generated from Roadmap Consortium, which were also curated in FeatSNP database. We found the regions tagged by SNP rs8070723 enriched for strong active histone modification signals including H3K4me1, H3K9ac, and H3K27ac in 8 brain tissues (Figure 4A). Such active histone modifications were generally associated with active enhancer and promoter functions. Chromatin epigenetic status prediction based on chromHMM (Ernst and Kellis, 2012) suggested that the regions tagged by SNP rs8070723 could be considered

as strong enhancers (Figure 4B). Finally, we explored the eQTL data in 13 brain tissues, and found rs8070723 was associated with several genes' expression, including *MAPT* (Figures 4B,C). *MAPT* gene mutations have been associated with several neurodegenerative disorders such as Alzheimer's disease and Parkinson's disease. Our result suggests that rs8070723 G allele might influence *MAPT* expression level by reducing the binding affinity of upstream regulatory protein *PBX1*, therefore providing a mechanistic association with neurodegenerative diseases including Progressive Supranuclear Palsy and Parkinson's Disease.

CONCLUSION

In summary, FeatSNP is an interactive database providing brain-specific functional genomics resources to investigate the regulatory potential of human SNPs. This database provides a multitude types of functional annotations, including TF binding motif prediction, epigenetic landscape, expression correlation and eQTL information. We anticipate that this database will facilitate scientists to investigate the functional impact of their candidate genetic variants in a more streamlined, rapid, and efficient fashion.

REFERENCES

- Agrawal, A., Chou, Y. L., Carey, C. E., Baranger, D. A. A., Zhang, B., Sherva, R., et al. (2018). Nabis dependence. *Mol. Psychiatry* 23, 1293–1302. doi: 10.1111/den.13366
- Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461. doi: 10.1038/nature12787
- Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 22, 1790–1797. doi: 10.1101/gr.137323.112
- Carey, C. E., Agrawal, A., Zhang, B., Conley, E. D., Degenhardt, L., Heath, A. C., et al. (2015). Monoacylglycerol lipase (MGLL) polymorphism rs604300 interacts with childhood adversity to predict cannabis dependence symptoms and amygdala habituation: evidence from an endocannabinoid system-level analysis. *J. Abnorm. Psychol.* 124, 860–877. doi: 10.1037/abn0000079
- Claussnitzer, M., Dankel, S. N., Kim, K. H., Quon, G., Meuleman, W., Haugen, C., et al. (2015). FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* 373, 895–907. doi: 10.1056/NEJMoa1502214
- De Iaco, A., Planet, E., Coluccio, A., Verp, S., Duc, J., and Trono, D. (2017). DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat. Genet.* 49, 941–945. doi: 10.1038/ng.3858
- Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* 9, 215–216. doi: 10.1038/nmeth.1906
- Genestine, M., Lin, L., Durens, M., Yan, Y., Jiang, Y., Prem, S., et al. (2015). Engrailed-2 (En2) deletion produces multiple neurodevelopmental defects in monoamine systems, forebrain structures and neurogenesis and behavior. *Hum. Mol. Genet.* 24, 5805–5827. doi: 10.1093/hmg/ddv301
- Gibson, G. (2015). Human genetics. GTEX detects genetic effects. *Science* 348, 640–641. doi: 10.1126/science.aab3002

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: <http://www.roadmapepigonomics.org/>.

AUTHOR CONTRIBUTIONS

C-yM and BZ performed the data analysis, C-yM and PG developed the database and website. PM, TW, and BZ designed and supervised the study.

FUNDING

This work was supported by National Institutes of Health grant DA027995, HG007175, HG007354, and Goldman Sachs Philanthropy Fund.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2019.00262/full#supplementary-material>

- Golonzhka, O., Nord, A., Tang, P. L. F., Lindtner, S., Ypsilanti, A. R., Ferretti, E., et al. (2015). Pbx regulates patterning of the cerebral cortex in progenitors and postmitotic neurons. *Neuron* 88, 1192–1207. doi: 10.1016/j.neuron.2015.10.045
- Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018. doi: 10.1093/bioinformatics/btr064
- Hoglinger, G. U., Melhem, N. M., Dickson, D. W., Sleiman, P. M., Wang, L. S., Klei, L., et al. (2011). Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nat. Genet.* 43, 699–705. doi: 10.1038/ng.859
- Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C. Y., Denay, G., Lee, J., et al. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 44, D110–D115. doi: 10.1093/nar/gkv1176
- Nelson, E. C., Agrawal, A., Heath, A. C., Bogdan, R., Sherva, R., Zhang, B., et al. (2016). Evidence of CNH3 involvement in opioid dependence. *Mol. Psychiatry* 21, 608–614. doi: 10.1038/mp.2015.102
- Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., and Robinson, J. T. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665–1680. doi: 10.1016/j.cell.2014.11.021
- Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330. doi: 10.1038/nature14248
- Salmon, E., Van der Linden, M. V., and Franck, G. (1997). Anterior cingulate and motor network metabolic impairment in progressive supranuclear palsy. *Neuroimage* 5, 173–178. doi: 10.1006/nimg.1997.0262
- Sanborn, A. L., Rao, S. S., Huang, S. C., Durand, N. C., Huntley, M. H., Jewett, A. L., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U.S.A.* 112, E6456–E6465. doi: 10.1073/pnas.1518552112
- Siriboonpipittana, T., Zeisig, B. B., Zarowiecki, M., Fung, T. K., Mallardo, M., Tsai, C. T., et al. (2017). Transcriptional memory of cells of origin overrides beta-catenin requirement of MLL cancer stem cells. *EMBO J.* 36, 3139–3155. doi: 10.15252/embj.201797994

- UK Parkinson's Disease Consortium, Wellcome Trust Case Control Consortium, Spencer, C. C., Plagnol, V., Strange, A., Gardner, M., et al. (2011). Dissection of the genetics of Parkinson's disease identifies an additional association 5' of SNCA and multiple associated haplotypes at 17q21. *Hum. Mol. Genet.* 20, 345–353. doi: 10.1093/hmg/ddq469
- Villaescusa, J. C., Li, B., Toledo, E. M., Rivetti di Val Cervo, P., Yang, S., Stott, S. R., et al. (2016). A PBX1 transcriptional network controls dopaminergic neuron development and is impaired in Parkinson's disease. *EMBO J.* 35, 1963–1978. doi: 10.15252/embj.201593725
- Ward, L. D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 40, D930–D934. doi: 10.1093/nar/gkr917
- Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–D1006. doi: 10.1093/nar/gkt1229
- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoutte, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9:R137. doi: 10.1186/gb-2008-9-9-r137
- Zhou, X., Li, D., Zhang, B., Lowdon, R. F., Rockweiler, N. B., Sears, R. L., et al. (2015). Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. *Nat. Biotechnol.* 33, 345–346. doi: 10.1038/nbt.3158

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Ma, Madden, Gontarz, Wang and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.