Pipeline for GWAS analysis with DeepSEA

Pradeep Varathan

PhD Student, Indiana University Purdue Univerisity Indianapolis,
INFO-I590 AI in Health Care
Project Report

* ppugale@iu.edu

Introduction

In the present day research environment, we have a lot of big data on genotype information for specific diseases. This provides us with an opportunity to use deep learning method to train networks to recognize the functionality of specific parts of the sequences. GWAS or genome wide analysis summary statistics allows us to interpret important single nucleotide polymorphisms that are related to a specific disease and also to a specific population. DeepSEA has been developed using proteome and motif-sequence data which includes nearly all the promoters and enhancers in the genome and can successfully predict this particular sequence has any functionality to words regulating specific function on gene .this allows us to interpret the impact of SNP variants in the same sequences. Most variants in a population are not recognized in their phenotype because they go mute when the central dogma takes place, that is, when the variants from DNA get transcribed to mRNA and then translated to protein amino acid sequences, they do not change the amino acid in that particular position and thus, the mutations do not affect the final protein. This is also called as silent mutations. While very few and rare scenarios where the variant significantly affects the protein, these cases can be accumulated in a population due to evolutionary genetic drifts. Therefore, using GWAS which enables us to tabulate the highly significant variants present within a population and also calculating the odds ratio in accordance with a disease, provides us the effect alleles and the alternate alleles. With the usage of DeepSEA network, we are able to look into how much these alleles can make a difference in the protein functionality based on the predictions.

Materials and methods

Dataset Used

For this methodology, we used the IGAP (International Genomics of Alzheimer's Project) summary analysis from it's GWAS performed on cohorts of Alzheimer's disease patients. IGAP is a large two-stage study based upon genome-wide association studies (GWAS) on individuals of European ancestry. In stage 1, IGAP used genotyped and imputed data on 7,055,881 single nucleotide polymorphisms (SNPs) to meta-analyse four previously-published GWAS datasets consisting of 17,008 Alzheimer's disease cases and 37,154 controls (The European Alzheimer's disease Initiative – EADI the Alzheimer Disease Genetics Consortium – ADGC The Cohorts for Heart and Aging Research in Genomic Epidemiology consortium – CHARGE The Genetic and Environmental Risk in AD consortium – GERAD). In stage 2, 11,632 SNPs were genotyped and tested for

association in an independent set of 8,572 Alzheimer's disease cases and 11,312 controls. Finally, a meta-analysis was performed combining results from stages 1 & 2. From this, we filtered the top 2000 SNPs based on their p-values and used for the whole pipeline.

Methodology

We used the DeepSEA algorithm, which uses three convolutional layers with a ReLu activation function and a maxpooling layer to reduce the dimensionality further, while the linear neural network at the final layer completes the reduction to fit the 919 chromatin types from different cell types as the output. The input of DeepSEA would be a one hot encoded sequence of about 1000 base pairs long and thus would have a 4x1000 matrix as the initial input. The one hot encoding is done prior to feeding into the DeepSEA. The first convolutional layer has an output channel size of 320 with a kernel size of 8, and with the maxpooling of kernel size 4 and stride 4,and is followed by convolutional layers that have 480 and 960 channels respectively with the same kernel size, and each passing through a maxpooling layer and a ReLu activation function.

To obtain the sequences, the first step was to filter the GWAS summary statistics from IGAP based on the p-values and sort them in the ascending order. After that, the top 2000 SNPs were filtered out for this current study due to memory challenges in the post analysis step. Using the biopython module which connects to the Entrez database from NCBI, I pulled the sequence information in FASTA format that was later converted to string format so that it can be one hot encoded for the deep learning network. These steps were then made into a simple profiler which I called C-SHAL that is updated on GitHub and can be used to perform the whole process just by providing the basic input commands and also has a tutorial section.

The final analysis of the results begin with performing a log change on the effect allele sequence and the alternate allele sequence of the SNP variant. This is tabulated across all 919 chromatin cell types and the SNPs taken into account. Further, I performed a simple biclustering analysis for the whole matrix as such, and also divided the matrices into DNase I–hypersensitive sites, Transcription factor binding sites and histone marks , for which I performed the above analysis individually to get a better separation among specific cell types. A general heatmap also reveals a similar trend on the blocks of correlated areas that shall discussed below. The post analysis was done in R with the help of BiClust and Biclustermd packages while the rest of the analysis was completely done on Python, with which a tool for terminal usage, C-SHAL was also created and available at https://github.com/PradoVarathan/C-SHAL-DeepSEA/

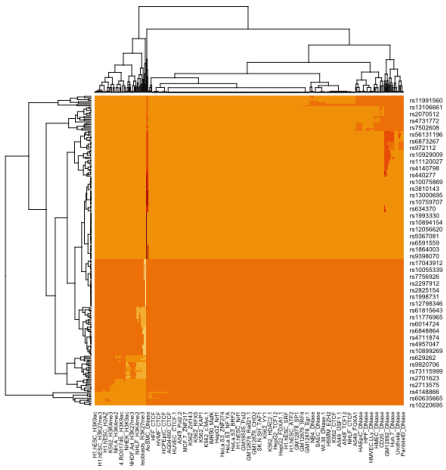

Fig 1. Heat map visualization of the top values after exclusion of lesser variation SNPs

Further, to visualize and build bi-clusters on the observed data, I used R programming and its packages biclust, QUBIC and XGR for the bi-cluster analysis and network analysis of the SNPs obtained. Moreover, we also did analyze the SNPs and genes that are related to these top SNPs based on their positions in the genome.

Results

The results from the C-SHAL pipeline as described in the methodology section provided with the log changes for all chromatin types from the various cell types and also for the SNPs provided. For the analysis, I took values that are greater and smaller than 35 percent of the absolute maximum in the log2 fold change and plotted the heat map to visualize ever so slight variations in the chromatin changes. This heat map clearly showed two major separation, one with higher fold change across while the other with slightly lower fold change as seen in the Fig1.

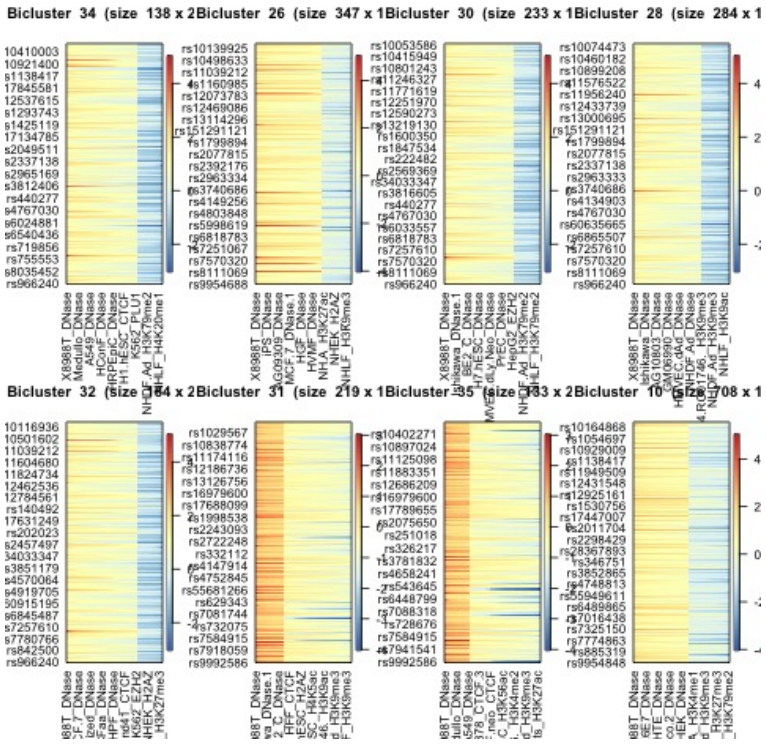

Fig 2. Visualization of the top 8 clusters based on the score from GWAS

Further, taking the above filtered data and analyzed them with the biclustering package called QUBIC, which provided with about 35 biclusters with its biclustering algorithm. To check on the validity of each biclusters, I analyzed the significance of each cluster by calcualting the average p value of the SNPs from the GWAS data and assigning them as a score for each cluster. Further, visualizing these cluster scores and arranging them in order, I could observe a pattern in the top significant clusters, which are provided in the Fig2.

The most significant bicluster was bicluster number 34 as seen in the figure which accounted for 138 SNP across 218 chromatin features. To analyze the variants that were clustered together, I used the XGR package that was based on the GWAS catalog and since the IGAP was implemented in the European population, I chose the EUR catalog

from the GWAS for the linkage disequilibrium scoring and analysis. While analysing the variants relation with disease, we got a high odds ratio shared between Alzheimer's disease and tauopathy with extremely high significance. Moreover, the variants rs2965169, rs7255063, rs1160984 and rs440277 had a high LD-score and shred high significance among the genes in the cluster. The variants were distributed among chromosomes 3, 5, 7, 11 ,14, 15 and 22 with most of them present in chromosome 14 and 5.

As for the 218 chromatin features, I found that all 125 DNases were present in this cluster as in most of the clusters cementing the role of DNases in the disease modification by variants. Moreover, out of the 93 other chromatin features, higher order of those are present in the histones section. The results of these analysis are found in the Cluster Data folder in the repository.

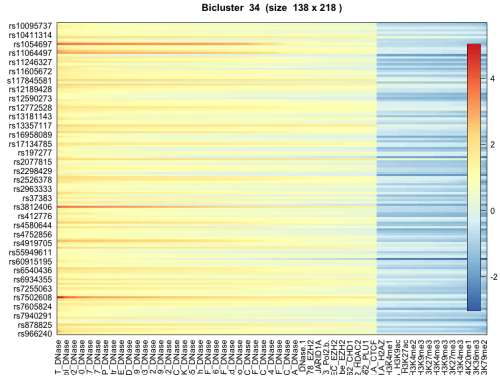

Fig 3. Heatmap visualization of the most significant cluster based on the score from GWAS

Discussion

As we can see in the results section, we targeted the SNP variants that belonged to the most significant cluster among the 35 biclusters obtained from the QUBIC package. Among the chromatin features, all the DNases were present in the cluster and the heatmap shows a significant difference in the log change between DNases and the rest of the chromatin features.I concentrated on analyzing the SNPs that were obtained from these clusters first and using the EUR population GWAS catalog, a network analysis profile was done on the variants. It resulted in Alzheimer's Disease as a top network relation as expected. The next step was to build the network based on the SNPs and their genes from the KEGG and String database. Surprisingly, I was able to only get one perfect gene from the whole network, which was PICALM gene.

This particular gene has been extensively studied in Alzheimer's Disease and has been related to the cognitive decline and many cases used as a biomarker. Though the SNP variant rs3851179 has been studied in this gene, the same allele major allele variant was not found in the IGAP GWAS data but rather four new variants for the gene resulted from LD analysis with the highest score. Those include rs2965169, rs7255063, rs1160984 and rs440277, which have never been studied before in Alzheimer's disease. The SNP variant rs2965169 has been studied as a viable proto-oncogene variant but not in relation with cognitive decline. PICALM, playing an integral role in clathirin mediated pathway, is also dominant in tau-plauqe generation and thus, can be a great target and a biomarker for Alzheimer's disease with its involvement in both the tau and amyloid plaque generation.

The histones from the cluster were all under the negative log change with the alternate allele but the rest of the chromatin features including DNases and transcription factors were under the positive log change with the alternate allele, as seen in Fig 3. This again might implement that these variants are more influential in DNases and might affect the functionality of the histone chromatin features.

Conclusion

On the whole, we can see the using DeepSEA deep learning pipeline to predict the probability scores of the 919 important chromatin features on a GWAS summary statistics SNP variants with their major allele and alternate allele provides us clues to which variants and what type of chromatic features these variants affect and by how much. From the analysis, I see that the SNPs included in the cluster 34 i.e 138 SNPs can be considered important with respect to changes in the chromatin features. Among those SNPs, rs2965169, rs7255063, rs1160984 and rs440277 have been predicted to be most significant and has not been researched yet. The gene PICALM has also been important since most of the variants analyzed from the biclustering analysis belonged to this particular gene. All results along with a working pipeline is present in the github repository https://github.com/PradoVarathan/C-SHAL-DeepSEA/.

Acknowledgments

We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the patients, and their families. The i–Select chips was funded by the French National Foundation on Alzheimer's disease and related disorders. EADI was supported by the LABEX (laboratory of excellence program investment for the future) DISTALZ grant, Inserm, Institut Pasteur de Lille, Université de Lille 2 and the Lille University Hospital. GERAD was supported by the Medical Research Council (Grant n° 503480), Alzheimer's Research UK (Grant n° 503176), the Wellcome Trust (Grant n° 082604/2/07/Z) and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grant n° 01GI0102, 01GI0711, 01GI0420. CHARGE was partly supported by the NIH/NIA grant R01 AG033193 and the NIA AG081220 and AGES contract N01–AG–12100, the NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by the NIH/NIA grants: U01 AG032984, U24 AG021886, U01 AG016976, and the Alzheimer's Association grant ADGC–10–196728.

References

1. Ponomareva NV, Andreeva TV, Protasova MA, Filippova YV, Kolesnikova EP, Fokin VF, Illarioshkin SN, Rogaev EI. Genetic Association between Alzheimer's Disease Risk Variant of the PICALM Gene and Auditory Event-Related Potentials in Aging. Biochemistry (Mosc). 2018 Sep;83(9):1075-1082. doi: 10.1134/S0006297918090092. PMID: 30472946.

2. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12(10):931-934. doi:10.1038/nmeth.3547

3. Fang, H., Knezevic, B., Burnham, K.L. et al. XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. Genome Med 8, 129 (2016). https://doi.org/10.1186/s13073-016-0384-y