A unified genome constraint, pathogenicity, and pLoF model identifies new genes associated with epilepsy

Oscar Aguilar[1]*, *Mijail Rivas*[2], *Manuel A. Rivas*[3]*

Abstract

Background: Epilepsy is a highly heterogeneous disorder thought to have strong genetic components. However, identifying these risk factors using whole-exome sequencing studies requires very large sample sizes and good signal-to-noise ratio in order to assess the association between rare variants in any given gene and disease.

Methods: We present a novel approach for predicting constraint in the human genome – sections of the genome where any mutation can cause a severe disorder. Through application of a Hidden Markov Model (HMM) to the Regeneron Genetics Center Million Exome dataset and the AllofUs whole genome sequencing data, we predict the probability of observing no variants across the population for each position in the genome. Next, we aggregate the constraint predictions by gene and assess its association to epilepsy. Finally, we extend our analysis model to incorporate pathogenicity predictions from AlphaMissense (AM) and pLoFs, and compare against published results.

Results: We identified a set of ($p < 1x10^{-4}$) genes with stronger signals than previously published studies including KDM5B, KCNQ2, CACNA1A, CACNA1B, RYR2, and ATP2B2. Our models allow us to evaluate the contribution of constraint, protein structure based pathogenicity prediction from AM, and pLoFs jointly.

Conclusion: We showed that relatively simple sequence-dependent constraint prediction models can complement structure-based missense variant pathogenicity predictions and pLoFs for population cohort studies which require additional statistical power in the identification of gene-based signals for neurogenetic and psychiatric disorders.

[1] Department of Management Science & Engineering, Stanford University
[2] National Institute of Neurology and Neurosurgery, Mexico and Clinical Epileptology Fellowship UNAM
[3] Department of Biomedical Data Science, Stanford University
* Corresponding authors. E-mails: osthoag@stanford.edu; mrivas@stanford.edu

Introduction

Epilepsy is a highly heterogeneous neurological disorder affecting a significant proportion of the global population. The complex etiologies of this condition involve a substantial genetic component, prompting extensive research to identify the underlying risk factors. Epilepsy, characterized by a predisposition to recurrent epileptic seizures, presents a prevalence of 4-10 per 1,000 individuals globally, making it a prevalent neurological condition. While the role of genetic contributions to epilepsy has long been acknowledged, the comprehensive understanding of the complete spectrum of genetic causes remains a formidable challenge.

Whole-exome sequencing (WES) has emerged as a powerful tool for investigating the genetic basis of epilepsy, enabling the identification of rare variants associated with various diseases. Recent efforts in this domain include an unprecedented study by the Epi25 Collaborative [1] with a sample size of over 54,000 individuals, including 20,979 epilepsy cases and 33,444 controls across diverse genetic ancestries. While this study marks a crucial step forward in unraveling the rare variant risk underlying a spectrum of epilepsy syndromes, it also underscores the ongoing challenges and limitations inherent in current genetic analyses of complex disorders. The need for larger sample sizes, the incorporation of diverse genetic ancestries, and the exploration of alternative analytical approaches are crucial to advancing our understanding of the genetic architecture of epilepsy, as well as related neurogenetic and psychiatric disorders.

Recent advancements in the field of structure-based missense variant pathogenicity predictions have significantly contributed to unraveling the complexities of genetic factors associated with epilepsy. A groundbreaking development in this domain comes from Google DeepMind's AlphaMissense [2], a deep learning model which leverages AlphaFold2 to predict the pathogenicity of single amino acid changes in proteins. Missense variants, which alter the amino acid sequence of proteins, play a crucial role in disrupting protein function and, consequently, have implications for organismal fitness. AlphaMissense's predictions for the entire human proteome are provided as a community resource which holds promise for uncovering previously unknown disease-causing genes, enhancing the diagnostic yield of rare genetic diseases, and advancing our understanding of the genetic basis of epilepsy.

Adding to this landscape is the Regeneron Genetics Center Million Exome (RGC-ME) dataset [3], representing the largest catalog of human protein-coding variation to date, derived from exome sequencing of 985,830 individuals with diverse ancestry. This comprehensive resource includes approximately 10.5 million missense variants (54% novel) and 1.1 million predicted loss-of-function (pLOF) variants (65% novel, 53% observed only once). Understanding the constraints in coding regions and genes is vital for assessing their role in diseases. The RGC-ME dataset, with its unprecedented sample size, enhances the precision of constraint scores, enabling the identification of highly constrained genes even among those lacking known disease associations.

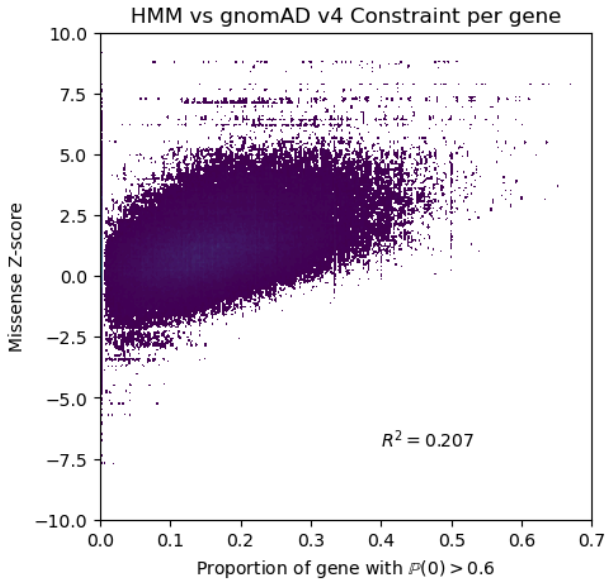
Results

Study overview

In this study, we implement a novel approach to predict genomic constraint using a Hidden Markov Model (HMM) applied to the RGC-ME dataset. Our model estimates the probability of observing no variants across the population for each position in the genome, providing insight into regions where mutations are likely to cause severe disorders. We applied this constraint inference model to investigate the genetic basis of epilepsy. Leveraging the predictions generated, we integrate pathogenicity predictions from AlphaMissense (AM) and information on predicted Loss-of-Function (pLoF) variants into a meta-regression model with the target of estimating the effect size of empirical cases for each gene/group. Through weighted least squares regression, we evaluate the contribution of HMM constraint probabilities, pathogenicity predictions from AM, and pLoFs to the identification of new genes associated with epilepsy.

Constraint per gene vs gnomAD

In order to validate the HMM predictions as a measure of constraint, we compare it to another measure called the Missense Tolerance Ratio (MTR), calculated as the observed proportion of missense variants divided by the expected proportion relative to the number of all possible variants in a given window [4]. Specifically, we compute the proportion of each gene with predicted probability of 0 above a certain threshold and evaluate it against the z-score of the observed count of missense variants relative to the expected count as obtained from gnomAD. The joint distribution of z-scores versus constraint proportion with thresholds 0.6 and 0.8 are shown below. The r-squared value is obtained by fitting a GLM between the two measures using ordinary least-squares regression. Decreasing the threshold from 0.8 to 0.6 increases the correlation from 0.186 to 0.207 and yields a more normal joint distribution.



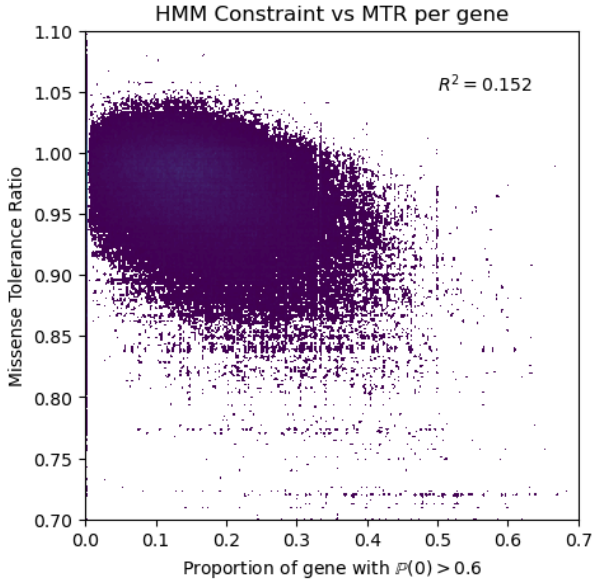


Figure 1: Joint distribution of z-score of observed vs expected missense variants per gene (left) and missense tolerance ratio (right) versus proportion of gene with constraint probability > 0.6

Validation on AllofUs whole exome sequence data

We further validate our methodology of constraint inference via HMM using an independent dataset of 250k individuals from AllofUs (AoU). We use the HMM trained on chromosome 2 of RGC-ME to generate separate predictions for all chromosomes (except chromosome 2) of RGC-ME and AoU. **Figure 2** shows the joint distribution of constraint probabilities across both datasets. Although most of the frequency mass is concentrated in the low probability regions of the marginal distributions (prob < 0.1), there is an overlap in the positions where both datasets assign high probability of constraint (top right). In addition, we compare this observed joint distribution to the expected joint distribution if constraint predictions were independent across both datasets. See the Methods section for the full analysis. Overall, the most significant overlap occurs in the high and low probability regions, suggesting that useful constraint information is recovered by the model independent of the dataset.

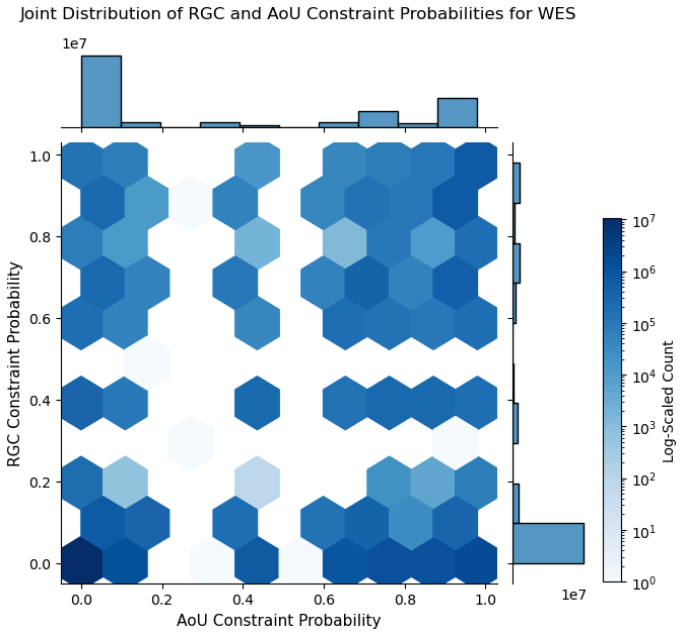


Figure 2: Joint distribution of WES constraint predictions from AoU and RGC.

Application to epilepsy

Having validated the utility of the HMM as a relatively simple sequence-dependent constraint prediction model, we further incorporate structure-based missense variant pathogenicity predictions from AlphaMissense and existing information on pLoF variants. The HMM constraint predictions, AM pathogenicity predictions, and indicators for pLoF are combined into a meta-regression model in order to estimate the effect size of empirical cases. The result is a unified model for the identification of gene-based signals for epilepsy.

We apply this analysis to the Epi25 Collaborative [1] dataset of 54k individuals with and without epilepsy. The table below shows the p-value from each component of the unified model for a subset of the genes with overall p-value < 0.0001. For each of these genes, we observe that the meta-regression model detects a signal ($p < 0.0001$) where the original analysis from the population cohort study does not have sufficient statistical power. In addition, we can identify how each component of the model contributes to this signal.

Gene Name	Group	Constraint p-value	Pathogenicity p-value	pLoF p-value	Constant p-value	Unified model p-value
KCNQ2	DEE	0.002238	0.003110	NaN	3.763469e-48	8.382517e-08
CACNA1A	GGE	0.000005	0.067895	0.000245	9.986918e-32	3.718246e-09
CACNA1B	GGE	0.062667	0.071230	0.000451	1.589863e-40	0.000033
ATP2B2	DEE	0.000271	0.561085	0.000079	1.873766e-57	2.308754e-07
KDM5B	DEE	0.900668	0.005757	0.000068	8.676487e-93	0.000073
RYR2	GGE	0.818109	0.408430	4.728223e-10	1.742691e-77	2.647377e-09

Table 1: P-values from each component of the unified constraint, pathogenicity, and pLOF model for a subset of the genes with final p-value < 0.00001.

Of the set of significant genes with stronger signals than previously published studies, many have a well established relationship with epilepsy. Mutations in the KCNQ2 gene are associated with a spectrum of neonatal-onset epilepsy syndromes, including benign familial neonatal seizures (BFNS) and developmental and epileptic encephalopathy (DEE) [5]. Similarly, the CACNA1A gene is associated with various types of epilepsy including DEE [6], and mutations in the CACNA1B gene are linked to various forms of epilepsy ranging from rare episodic ataxia syndromes to genetic generalized epilepsy (GGE) [7]. KCNQ2 and CACNA1A/CACNA1B play important roles in neuronal functions by providing instructions for making potassium and calcium channels, respectively. Within the unified model, information on constraint, pathogenicity, and pLoF each play different roles in the signal for each gene. For KCNQ2, constraint prediction from the HMM and pathogenicity predictions from AM each contribute to the overall model p-value of 8.38×10^{-8}. While pLoF variant information contributes to both CACNA1A and CACNA1B, our constraint predictions provide much larger value to the model for CACNA1A (p-value of 3.72×10^{-9}) over the model for CACNA1B (p-value of 3.3×10^{-5}).

Many other genes identified in our analysis have been linked to epilepsy, but do not have causal mechanisms which are well understood in the existing literature. De novo variants in ATP2B2, for instance, have been associated with variable neurodevelopmental disorders, including seizures [8]. Another study found an association in expression of ATP2B2 and epilepsy ($p=0.049$), along with a high correlation between ATP2B2 and its related long non-coding RNA (lnc-MTR-1) in patients with epilepsy [9]. Mutations in KDM5B have likewise been associated with variable neurodevelopmental disorders including intellectual disability (ID), and has been shown to regulate memory consolidation in the hippocampus of mice [10]. However, the link between genes in the KDM5 family and epilepsy in humans is limited to case reports of an individual with a de nuovo frameshift variant in KDM5B [11] and a Korean family with a deleted

region of genes including KDM5A [12]. Finally, there is an emerging relationship between certain types of epilepsy and mutations in RYR2, another gene with a crucial role in calcium homeostasis and signaling. Two recent studies have identified novel RYR2 mutations in five children with benign epilepsy of childhood with centrotemporal spikes (BECTS) [13] and a child with focal epilepsy [14]. The unified models for ATP2B2, KDM5B, and RYR2 each receive significant signals from pLoFs, with overall p-values of 2.31×10^{-7}, 7.3×10^{-5}, and 2.65×10^{-9} respectively.

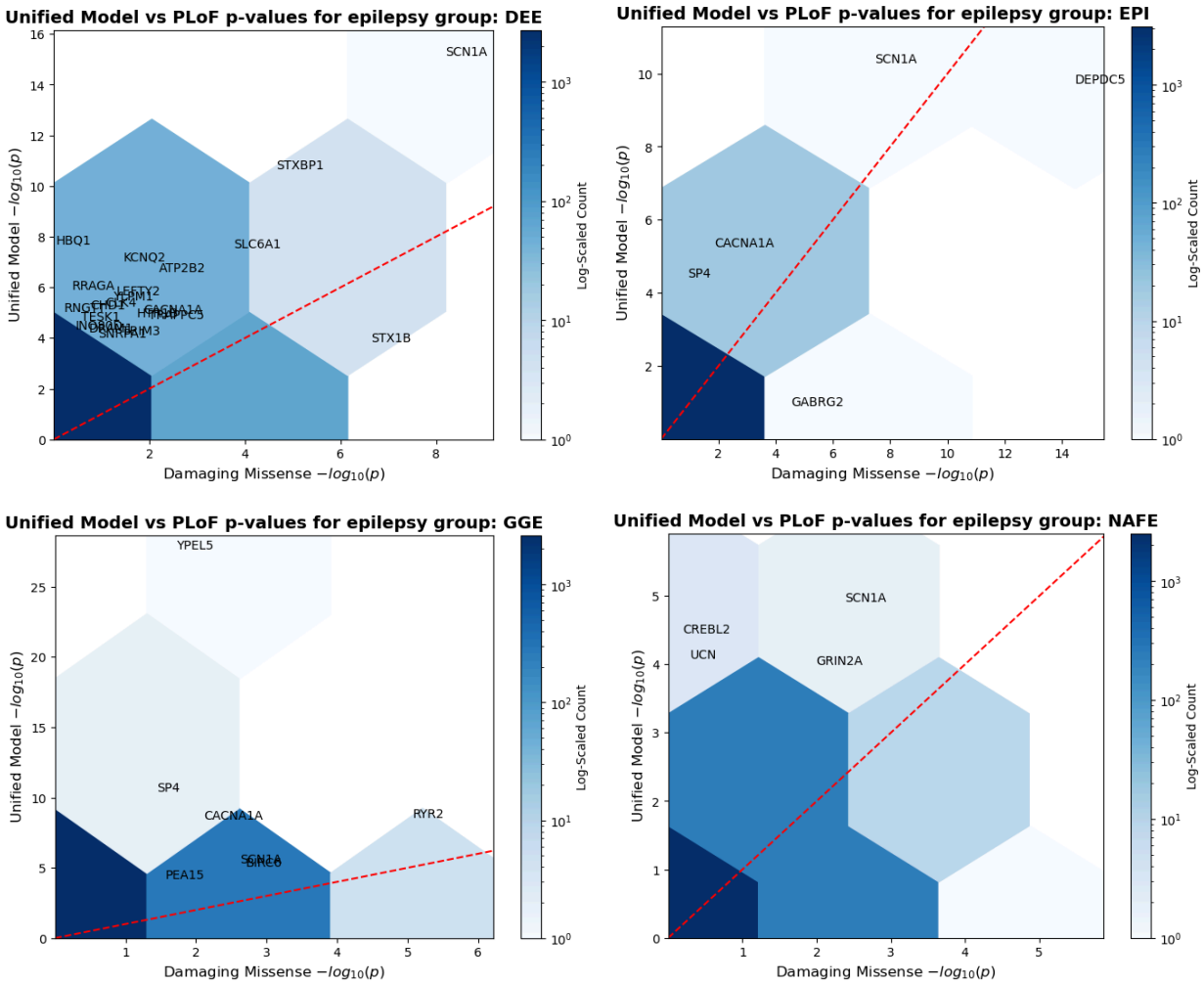


Figure 3: Joint distribution of (log-scaled) p-values from our unified model vs published (log-scaled) p-values for various types of epilepsy. Of the published results, we use the minimum p-value from damaging missense variants and pLOF variants for comparison (x-axis). Shaded areas indicate the (log-scaled) frequency of joint p-values. Significant ($p < 0.0001$) genes where results differ by over two orders of magnitude are annotated.

Discussion

Our study presents a novel approach to understanding the genetic underpinnings of epilepsy through a unified model incorporating genome constraint, pathogenicity predictions, and predicted Loss-of-Function (pLoF) variants. The integration of Hidden Markov Model (HMM)-based constraint predictions with AlphaMissense (AM) pathogenicity predictions and pLoF data provided a robust framework for identifying new candidate genes.

Our findings underscore the complementary roles of constraint, pathogenicity, and pLoF in elucidating the genetic basis of epilepsy. For instance, the KCNQ2 gene, associated with neonatal-onset epilepsy syndromes, demonstrated significant contributions from both constraint predictions and AM pathogenicity predictions. Similarly, while pLoF information was pivotal for genes like CACNA1A and CACNA1B, the constraint predictions provided additional value, particularly for CACNA1A. Moreover, our study revealed novel insights into genes with previously underexplored associations with epilepsy. For example, ATP2B2, KDM5B, and RYR2 showed significant signals in our unified model, despite limited existing literature on their causal mechanisms in epilepsy. These findings suggest potential new avenues for research into the genetic basis of epilepsy and other neurogenetic disorders.

The utility of each component of the model is limited by the quality of the data. The pathogenicity predictions from AlphaMissense, for instance, contribute most to genes with better understanding of the proteins they encode. Future work will focus on expanding our unified approach to incorporate additional features and datasets. A larger constraint inference model, such as an LSTM or LLM, could incorporate other information in addition to the binary sequence of observed mutations. Similarly, nonlinear models with more features can take advantage of the large number of samples for each gene to better estimate effect size. Finally, our methodology can be applied to other neurodevelopmental disorders such as autism, schizophrenia, and bipolar disorder.

In conclusion, our unified model provides a powerful tool for the identification of genes associated with epilepsy. By integrating genome constraint, pathogenicity predictions, and pLoF data, we offer a comprehensive framework that can enhance the detection of gene-disease associations and provide new insights into the genetic architecture of these complex disorders.

Methods

Hidden Markov Model for whole exome sequence constraint inference

The Hidden Markov Model (HMM) takes as input a binary sequence for each chromosome. Positions encoded “1” indicate that at least one variant has been observed within the cohort. When no mutation exists in the population for that position, it is encoded with a “0”. As output, the HMM estimates the probability of observing a zero at that position. In other words, the HMM predicts the likelihood that the genome is “constrained” at that position in the sequence. The HMM is trained using an expectation-maximization (Baum-Welch) algorithm, which iteratively updates the transition probability matrix and the emission probability matrix for the model’s hidden states. We used the *hmmlearn* library in Python to train a model with two hidden states.

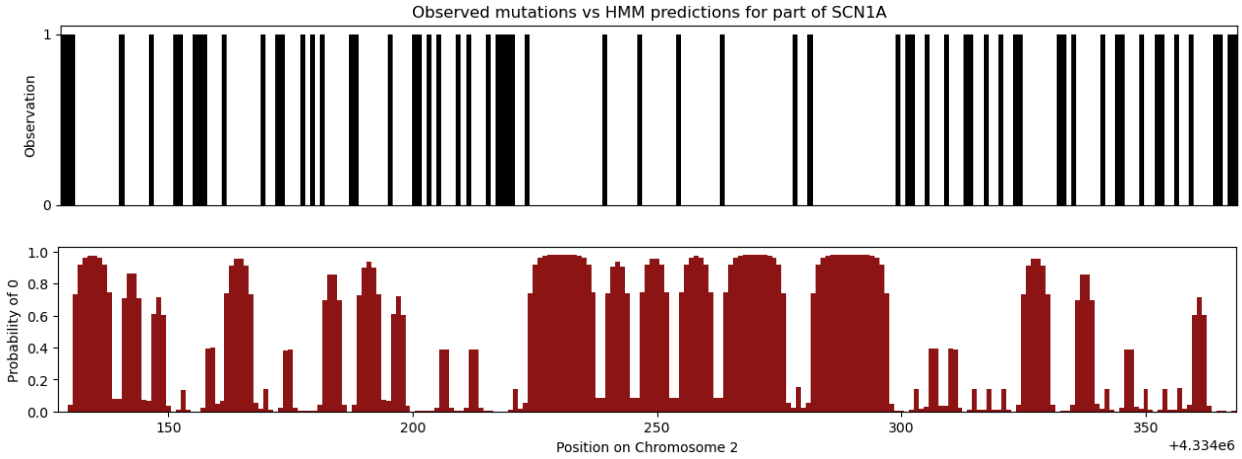


Figure 4: Example of an observed binary sequence vs the predictions made by the HMM

Using the Regeneron Genetic Center Million Exome (RGC-ME) dataset, we trained an HMM using the mutation patterns of chromosome 2 and generated constraint predictions for chromosomes 1 through 22. Before training/predicting on any chromosome, we filtered for positions with over 20% coverage for at least 90% of individuals as a form of quality control. In addition, we restricted our analysis to the protein-coding regions of the exome. The resulting predictions were used as a constraint signal for the identification of genes linked to epilepsy.

Validation on AllofUs whole exome sequence data

In order to validate our constraint model, we extend its application to the whole exome sequence data from AllofUs (AoU), an independent dataset of 245,388 individuals of mixed ancestry [15]. After applying the same coverage and protein-coding filters to the AoU WES data, we use the model trained on RGC-ME to generate predictions for chromosomes 1 through 22. Next, we evaluate the joint distribution of the constraint predictions across both datasets

(excluding chromosome 2) in order to gauge the consistency of our model. The frequencies of the resulting observed joint distribution is compared in **Figure 5** to the expected frequencies under the assumption of independence between the two marginal distributions. The plot of χ^2 statistics between the observed and expected distributions shows that the shared constraint information is concentrated in the high and low probability regions. This observation, along with the R^2 of 0.153 between the constraint probabilities for AoU and RGC, indicates that useful constraint information is recovered by the HMM independent of the given dataset.

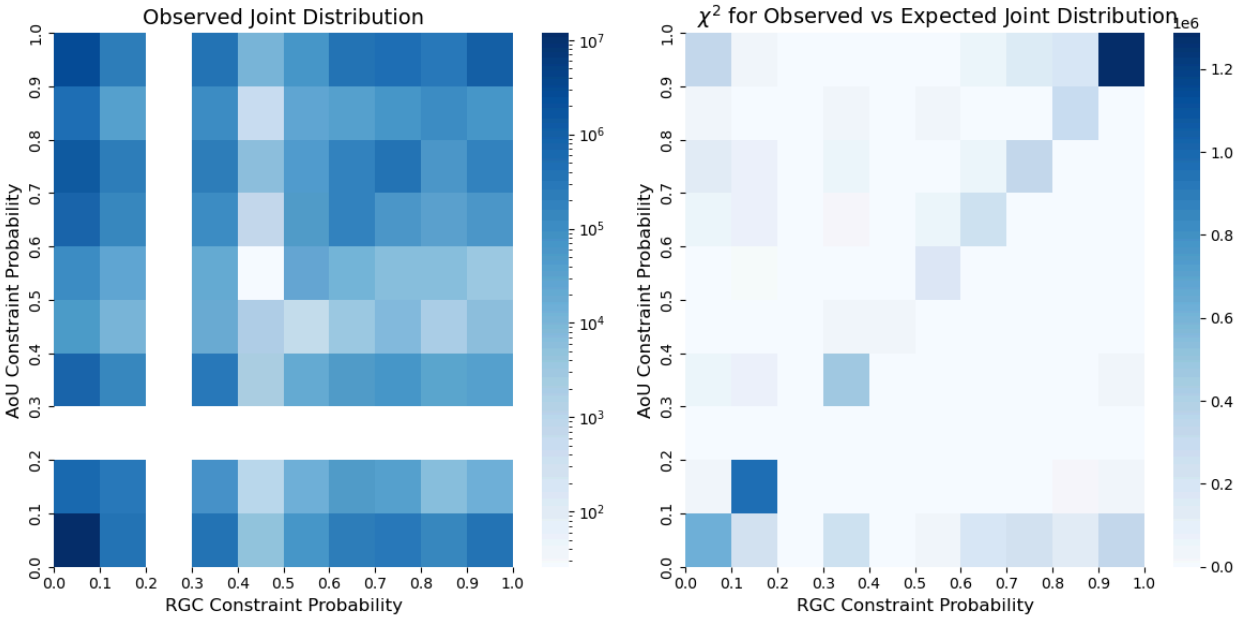


Figure 5: Observed joint distribution (left) of WES constraint predictions (excluding chromosome 2) from AoU and RGC, and distribution of corresponding χ^2 statistics (right)

Unified constraint, pathogenicity, and pLoF model

In order to integrate information from Alpha Missense (AM) and pLoF variants into a unified model, we build a meta regression model to estimate the effect size of empirical cases for each gene/group. For each gene, we only consider positions in the exome with at least one allele number in the study population's cases and controls and at most 5 allele counts in the full cohort. For every position in the sequence for that gene, we take the maximum pathogenicity prediction from AM across all variants, the constraint probability from our model and an indicator of whether or not the variant is a pLoF. The effect size and its variance are computed from the observed cases/controls in the cohort population using the allele frequencies and counts. Finally, we use a weighted least squares regression model with constraint probability, max pathogenicity, pLoF indicator, and constant term as predictors for effect size weighted by variance. The p values for each predictor variable were recorded along with the overall p value of the model. Together, these provide for a comprehensive look into the various signals we use in our identification of new genes associated with epilepsy.

We further examine the relationship between protein structure based pathogenicity prediction from AM and constraint data. A binomial regression analysis was performed to

evaluate correlation between the binary sequence of mutation observations and the maximum predicted pathogenicity. The resulting McFadden pseudo-R^2 value of 0.0081 indicates limited information overlap between AM predictions and mutation observations, especially when compared to the McFadden pseudo-R^2 of 0.3834 when using HMM predictions.

Bibliography

1. Epi25 Collaborative, Chen S, Neale BM, Berkovic SF. Shared and distinct ultra-rare genetic risk for diverse epilepsies: A whole-exome sequencing study of 54,423 individuals across multiple genetic ancestries. medRxiv. 2023. doi:10.1101/2023.02.22.23286310

2. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381: eadg7492.

3. Sun KY, Bai X, Chen S, Bao S, Kapoor M, Zhang C, et al. A deep catalog of protein-coding variation in 985,830 individuals. bioRxivorg. 2023. doi:10.1101/2023.05.09.539329

4. Traynelis J, Silk M, Wang Q, Berkovic SF, Liu L, Ascher DB, et al. Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res. 2017;27: 1715–1729.

5. KCNQ2. In: Epilepsy Foundation [Internet]. [cited 29 Apr 2024]. Available: https://www.epilepsy.com/causes/genetic/kcnq2

6. What Is CACNA1A. In: Epilepsy Foundation [Internet]. [cited 29 Apr 2024]. Available: https://www.epilepsy.com/causes/genetic/cacna1a-related-epilepsy

7. Epi4K consortium, Epilepsy Phenome/Genome Project. Ultra-rare genetic variation in common epilepsies: a case-control sequencing study. Lancet Neurol. 2017;16: 135–143.

8. Poggio E, Barazzuol L, Salmaso A, Milani C, Deligiannopoulou A, Cazorla ÁG, et al. ATP2B2 de novo variants as a cause of variable neurodevelopmental disorders that feature dystonia, ataxia, intellectual disability, behavioral symptoms, and seizures. Genet Med. 2023;25: 100971.

9. Taheri M, Pourtavakoli A, Eslami S, Ghafouri-Fard S, Sayad A. Assessment of expression of calcium signaling related lncRNAs in epilepsy. Sci Rep. 2023;13: 17993.

10. Perez-Sisques L, Bhatt S, Matuleviciute R, Gileadi T, Kramar E, Graham A, et al. The intellectual disability risk gene regulates long term memory consolidation in the hippocampus. J Neurosci. 2024. doi:10.1523/JNEUROSCI.1544-23.2024

11. Mangano GD, Antona V, Calì E, Fontana A, Salpietro V, Houlden H, et al. A complex epileptic and dysmorphic phenotype associated with a novel frameshift KDM5B variant and deletion of SCN gene cluster. Seizure. 2022;97: 20–22.

12. Han JY, Park J. Variable Phenotypes of Epilepsy, Intellectual Disability, and Schizophrenia Caused by 12p13.33-p13.32 Terminal Microdeletion in a Korean Family: A Case Report and Literature Review. Genes . 2021;12. doi:10.3390/genes12071001

13. Ma M-G, Liu X-R, Wu Y, Wang J, Li B-M, Shi Y-W, et al. Mutations Are Associated With Benign Epilepsy of Childhood With Centrotemporal Spikes With or Without Arrhythmia. Front Neurosci. 2021;15: 629610.

14. Hu J, Gao X, Chen L, Zhou T, Du Z, Jiang J, et al. A novel mutation in ryanodine receptor 2

() genes at c.12670G>T associated with focal epilepsy in a 3-year-old child. Front Pediatr. 2022;10: 1022268.

15. All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. Nature. 2024;627: 340–346.