UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À

L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN SCIENCES CLINIQUES ET BIOMÉDICALES

PAR

ALEXANDRE GIRARD

UNRAVELING THE ROLE OF NON-CODING RARE VARIANTS IN EPILEPSY AND ITS SUBTYPES WITH DEEP LEARNING

AOÛT 2023

Résumé

La découverte de nouveaux variants a vu son rythme diminuer dans les dernières années dans les études sur l'épilepsie, malgré l'utilisation de cohortes de très grandes tailles. Conséquemment, la majorité de l'héritabilité reste inexpliquée. Les variants rares non-codants ont été largement ignorés dans les études sur l'épilepsie, même si ces variants peuvent avoir un impact significatif sur l'expression des gènes. Nous avons accès au séquençage du génome complet (WGS) de 247 patients épileptiques et 377 témoins. Pour déterminer l'impact fonctionnel des variants non-codants, ExPecto, un algorithme d'apprentissage profond a été utilisé pour prédire le changement d'expression dans des tissus cérébraux. Nous avons comparé le fardeau des variants rares non-codants délétères des cas et des témoins. Les variants rares non-codants hautement délétères étaient significativement enrichis pour l'épilepsie génétique généralisée (GGE), mais pas pour l'épilepsie focale non-acquise (NAFE) ou pour tous les cas d'épilepsie comparés aux contrôles. Cette étude a permis de démontrer que les variants rares non-codants délétères sont associés à l'épilepsie, plus spécifiquement pour les patients GGE. De plus grande cohortes de WGS en épilepsie seront requises pour investiguer ces effets avec une plus grande résolution. Néanmoins, nous avons démontré l'importance d'étudier les régions non-codantes en épilepsie, une maladie où les nouvelles découvertes sont rares.

Abstract

The discovery of new variants has slowed down in recent years in epilepsy studies, despite the use of very large cohorts. Consequently, most of the heritability is still unexplained. Rare non-coding variants have been largely ignored in studies on epilepsy, although non-coding single nucleotide variants can have a significant impact on gene expression. We had access to whole genome sequencing (WGS) from 247 epilepsy patients and 377 controls. To assess the functional impact of non-coding variants, ExPecto, a deep learning algorithm was used to predict expression change in brain tissues. We compared the burden of rare non-coding deleterious variants between cases and controls. Rare non-coding highly deleterious variants were significantly enriched with Genetic Generalized Epilepsy (GGE), but not with Non-Acquired Focal Epilepsy (NAFE) or all epilepsy cases when compared with controls. In this study, we showed that rare non-coding deleterious variants are associated with epilepsy, specifically with GGE. Larger WGS epilepsy cohort will be needed to investigate those effects at a greater resolution. Nevertheless, we demonstrated the importance of studying non-coding regions in epilepsy, a disease where new discoveries are scarce.

Table of Contents

List of Tables

List of Figures

List of Abbreviations, Symbols and Acronyms

AI……Artificial Intelligence

CENet……Canadian Epilepsy Network

CVS…… Constraint Violation Score

DEE…… Developmental Epileptic Encephalopathy

eQTL…… expression Quantitative Trait Loci

ENCODE…… Encyclopedia of DNA Elements

gDNA……Genomic DNA

GGE…… Genetic Generalized Epilepsy

GTEx…… Genotype-Tissue Expression project

GWAS……Genome-Wide Association Study

ILAE…… International League Against Epilepsy

MAF…… Minor Allele Frequency

MRI……Magnetic Resonance Imaging

NAFE…… Non-Acquired Focal Epilepsy

OR…… Odds Ratio

PCA…… Principal Component Analysis

SNV……Single Nucleotide Variant

TWAS……Transcriptome-Wide Association Study

UMAP…… Uniform Manifold Approximation and Projection

WES……Whole Exome Sequencing

WGS…… Whole Genome Sequencing

Acknowledgments

I want to start by thanking all the participants of this study, without whom none of this work would have been possible. Next, I want to thank the FRQS for the scholarship that allowed me to achieve this master's degree. I want to thank IVADO and the CIHR for funding this study. Finally, I want to thank my director for his support throughout the project.

Foreword

The goal of this study was to investigate the role of non-coding genomic regions in the etiology of epilepsy. I strived to overcome the challenge of assessing the functional impact of non-coding variants. To do so I used deep learning as a tool to predict the effect of such variants, thus prioritizing variants of greater importance. Doing so suggested an important impact of those regions in the disease. Moreover, the method that was used is easily transposable to other studies no matter the trait of interest. The limitations of this study reside in the fact that the predictions made with deep learning are not experimentally validated since brain tissue from the patients are unavailable. However, this could be partly addressed by using cellular models from epileptic mice. Another limitation is the fact that the sample size is relatively small, which limits the resolution of new discoveries.

Introduction

Epilepsy

Epidemiology

Epilepsy is a neurological disorder, for which there are several subtypes with large phenotypic variability, but a common symptom to all epilepsies are spontaneous unprovoked seizures[1]. Overall, around 3% of the population will be affected by epilepsy at some point in their lifetime[2,3]. Indeed, epilepsy is a disease with no particular age of onset and symptoms can eventually stop or last a lifetime. For many patients the affliction will be a burden for their entire life. Epilepsy's consequences have many ramifications. The disease has a significant impact on the life of those affected by it. People affected by the disease have a higher mortality, extensive medication, frequent medical consultations and some cannot drive to name a few of the consequences of epilepsy[4]. Some of these consequences have major social and economic repercussions, such as a high cost associated with medication and medical care as well as frequent absence from work. Epilepsy is ranked as the third neurological disease that has the highest disability-adjusted life years as of 2017 after Alzheimer and chronic migraine[5,6]. Moreover, epilepsy can be a fatal illness and affected individuals have a two to threefold increase in their mortality rate compared to healthy individuals[7,8].

Overall, the prevalence of epilepsy is 6.38 per 1000 people and the incidence is 61.44 per 100 000 people every year[9]. Additionally, between 20 to 30% of patients have drug-resistant epilepsy[10,11]. Those patients don't respond to anti-seizure drugs and suffer even more from their condition. It may take longer for them to be treated as they have to try a large spectrum of drugs, sometimes in combinations to find a solution. If all of this doesn't work, they must resort to surgery, neuromodulation and/or vagal nerve stimulation[12]. This makes

treating drug-resistant epilepsy a gruesome process for the patients, their loved ones and the health care providers.

Disease pathophysiology

A seizure is a sudden and uncontrolled burst of electrical activity in the brain. This is caused by a specific type of cell, the neuron. Neurons are cells that can fire electric signals along an elongated arm called an axon[13]. The signal can be propagated by the opening of ion channels, proteins that can allow specific charged molecules to go through when they are opened. Ion channels only open when there is a shift in the surrounding electric charges, thus contributing to the propagation of the electric signal through the axon[14]. Neurons can send signals to each other by releasing chemicals at junction points called synapse. Those chemicals, called neurotransmitters, are released by the presynaptic neuron, and quickly bind to specific receptors on the postsynaptic neuron[15]. Different neurotransmitters will have different effects on the postsynaptic neuron. Some neurotransmitters will have an excitatory effect (excitatory conductance), others will have an inhibitory effect (inhibitory conductance)[16]. A postsynaptic neuron can receive inputs from multiple presynaptic neurons, where their combined effect will determine if the postsynaptic neuron also fires an electric signal or not.

The mechanism by which seizures are usually provoked is an imbalance between excitatory and inhibitory conductance[17]. This has been confirmed by using pharmaceutical agents. Such an imbalance is not permanently present in epilepsy patients, thus showing that a more complex mechanism is responsible for triggering seizures. The reason for this is that most patients rarely experience seizures (<1% of brain activity except for the most severe cases) which means that most of the time there is no such imbalance in their brain[18].

Furthermore, there are hundreds of genes associated with the disease, most of which are not directly related to neuronal excitation or inhibition, which implies that multiple indirect, complex and subtle molecular mechanisms are behind epileptic seizures potentially in parallel and interacting with each other[17]. As of today, there are several hypotheses as to how genetic variation can lead to rare spontaneous imbalance of activity in the brain and it remains a certain challenge in the field. Especially since a hypothesis could be true for a specific subtype of epilepsy, but not another.

Disease phenotype

Classification of different epilepsy phenotypes is made in accordance with the International League Against Epilepsy (ILAE) guidelines. It is generally made according to the type of seizure, for example focal or generalized and how the brain might be affected, such as developmental abnormalities[1,19]. For the present work, two types of epilepsy are of interest, Non-Acquired Focal Epilepsy (NAFE) and Genetic Generalized Epilepsy (GGE). Those types were selected because they are the most common type of epilepsy and are not caused by brain lesions or other external factors, thus showing stronger heritability[20,21].

NAFE is a focal epilepsy meaning that seizures originate from a network of neurons in only one hemisphere of the brain[22]. NAFE is a common type of epilepsy that is classified as non-acquired, which implies that the disorder is not caused by an acquired factor like a trauma, an infection, or a stroke. To diagnose a patient with NAFE, they must have had at least two unprovoked seizures occurring over 24 hours apart in the 6 months prior to the beginning of treatment and they must have a brain Magnetic Resonance Imaging (MRI) that shows no sign of epileptic lesion as per the current classification by the ILAE[22].

GGE is a generalized epilepsy, therefore seizures spread to networks of neurons in both cerebral hemispheres. As its name implies, GGE is a form of epilepsy with a genetic cause. Patients have GGE if their clinical and electroencephalography characteristics are in line with the ILAE syndrome definition. For the current project, patients were diagnosed according to the 1989 ILAE guideline (latest guideline at the start of recruitment)[23]. An MRI of the brain is not required for diagnosis.

Other characteristics may help to refine the diagnosis to a more specific type of epilepsy, such as length of seizures, frequency of seizures, the time at which seizures occur, etc. This procedure is called fine phenotyping and although it has been done for some individuals in our cohort, this information will not be used to further stratify the sample as it would create groups too small to have enough statistical power. Namely, some of the patients in this study have Jeavons syndrome, an idiopathic generalized type of reflex epilepsy with childhood onset, a specific seizure manifestation, striking light sensitivity and the possibility of tonic-clonic seizures[24]. Those patients fall into the broader category of GGE and will be included in this group.

Genetics in epilepsy

Around 2% of epilepsy are monogenic, which means that they follow either recessive or dominant Mendelian transmission. For instance cortical dysplasia-focal epilepsy[25] is a recessive form of the disease whereas autosomal dominant nocturnal frontal lobe epilepsy[26] is a dominant form of the disease. The remaining 98% of genetic epilepsies are complex traits[2]. Those traits are defined as traits for which a plethora of genomic regions are associated with the phenotype, often involving interactions and multiple biological pathways[27,28]. For this reason, it is possible that variants that affect a gene that is seemingly unrelated to the trait

of interest can indeed play a role in the expression of the phenotype through a myriad of biological interactions[29].

In recent years, the main focus of genomic research in epilepsy has been on genome-wide association studies (GWAS). A GWAS is a statistical approach that is used to test if there is an association between every individual variant and the phenotype of interest[30]. Traditionally, GWAS are performed with genotype data of a case-control cohort. This model often restricts findings to common variants. In epilepsy GWAS have historically been the main tool to make new discoveries[2,3,11,31–34].

The ILAE is a leader in GWAS for epilepsy. Over the past decade they conducted three meta-analyses, which allowed them to discover most of the newly associated loci and genes[3,11,33]. The first of those studies was published in 2014[3]. They combined data from 12 cohorts for a total of 8 696 cases and 26 157 controls, which was by far the greatest sample size ever assembled for epilepsy at the time, surpassing the 3 445 cases and 6 935 controls of a 2010 study[35]. In the end, this meta-analysis associated two loci with all epilepsy. The first at 2q24.3 centered in SCN1A a gene already associated with monogenic forms of the disease (Table 1). The other at 4p15.1 which overlapped with PCDH7 a gene never associated with epilepsy before. They also identified a locus specific to GGE located at 2p16.1 which contained VRK2 and FANCL. VRK2 had already been suggested as a risk gene for epilepsy and it was the first time that FANCL was associated with the disease. ILAE second meta-analysis was published in 2018[33]. In this study, the sample size was of 15 212 cases and 29 677 controls. When combining all epilepsy cases together they identified a new association at 16q12.1 and replicated the associations of the previous study. Additionally, they were able to show that the 2q24.3 locus contains a second independent signal. This locus

was also the only significant signal among NAFE patients. GGE only analysis revealed 11 significant loci, out of which 7 were associated with epilepsy for the first time. All the significant loci were mapped to 146 genes, every gene was attributed a score based on a variety of criteria to identify candidate risk genes. 21 genes were targeted as likely risk genes for epilepsy. Finally, the most recent meta-analysis of the ILAE was published in 2022 and is still in a preprint state[11]. This time the sample size has grown to 29 944 cases and 52 538 controls. Four loci were significantly associated to all epilepsies, two of which are novel associations. Of all their previous associations only the 2q24.3 locus was replicated in this study. No locus reached significance for NAFE cases. For GGE, 22 loci were significantly associated with the condition, 13 of which are novel. By using a similar method as in their previous GWAS, they mapped 282 genes to the significant loci and used a prioritization method to identify 29 candidate genes for epilepsy. This time they conducted a transcriptome-wide association study (TWAS). TWAS are a family of analysis that aim to combine GWAS data with expression mapping studies (like the Genotype-Tissue Expression project (GTEx)[36] and the Encyclopedia of DNA Elements (ENCODE)[37]) to investigate potential regulatory mechanisms that could be caused by non-coding SNVs[38]. By using a TWAS method named FUSION[39] the ILAE identified 27 genes that are associated with epilepsy and differential expression in the brain. Out of those 27, 19 were outside of the associated GWAS loci. Next, with a method named SMR[40] they demonstrated a potentially causal link between brain expression of RMI1 and all epilepsy. For GGE they showed a potentially causal link for RMI1 CDK5RAP3 and TVP23B. Finally, they estimated that the required sample size to identify enough SNV to account for 90% of SNV based heritability in GGE would be of around 2.5 million cases. To create such a cohort would be highly

impractical, not only would it require an astronomical funding, but it would also mean that almost 4% of the worldwide epilepsy population would have to be recruited[41].

Table 1. List of loci associated with epilepsy from major GWAS.

Locus	Phenotype	Gene	Reference
2q24.3	All	SCN1A	ILAE consortium on complex epilepsies 2014.
	All and NAFE	SCN3A, SCN2A, TTC21B and SCN1A	ILAE consortium on complex epilepsies 2018.
	All and GGE	SCN1A	ILAE consortium on complex epilepsies 2022.
	All	SCN1A and TTC21B	Song *et al.* 2021.
4p15.1	All	PCDH7	ILAE consortium on complex epilepsies 2014.
2p16.1	GGE	VRK2 and FANCL	ILAE consortium on complex epilepsies 2014.
	All and GGE	FANCL and BCL11A	ILAE consortium on complex epilepsies 2018.
	All and GGE	BCL11A	ILAE consortium on complex epilepsies 2022.
16q12.1	All	HEATR3 and BRD7	ILAE consortium on complex epilepsies 2018.
2p24.1	GGE	None	ILAE consortium on complex epilepsies 2018.
2q32.3	GGE	STAT4	ILAE consortium on complex epilepsies 2018.

4p15.1	GGE	PCDH7	ILAE consortium on complex epilepsies 2018.
	GGE	PCDH7	ILAE consortium on complex epilepsies 2022.
4p12	GGE	GABRA2	ILAE consortium on complex epilepsies 2018.
5q22.3	GGE	KCNN2	ILAE consortium on complex epilepsies 2018.
	GGE	KCNN2	ILAE consortium on complex epilepsies 2022.
6p22.3	GGE	ATXN1	ILAE consortium on complex epilepsies 2018.
6q22.33	GGE	None	ILAE consortium on complex epilepsies 2018.
	GGE	PTPRK	ILAE consortium on complex epilepsies 2022.
17q21.32	GGE	PNPO	ILAE consortium on complex epilepsies 2018.
	GGE	CDK5RAP3	ILAE consortium on complex epilepsies 2022.
21q22.11	GGE	GRIK1	ILAE consortium on complex epilepsies 2018.
	GGE	GRIK1	ILAE consortium on complex epilepsies 2022.
9q21.13	All	RORB	ILAE consortium on complex epilepsies 2022.

10q24.32	All and GGE	KCNIP2	ILAE consortium on complex epilepsies 2022.
1q43	GGE	RYR2 and CHRM3	ILAE consortium on complex epilepsies 2022.
2q12.1	GGE	POU3F3	ILAE consortium on complex epilepsies 2022.
2q32.2	GGE	GLS	ILAE consortium on complex epilepsies 2022.
3p22.3	GGE	STAC	ILAE consortium on complex epilepsies 2022.
3p21.31	GGE	CACNA2D2	ILAE consortium on complex epilepsies 2022.
5q31.2	GGE	SPOCK1	ILAE consortium on complex epilepsies 2022.
7p14.1	GGE	SUGCT	ILAE consortium on complex epilepsies 2022.
9q21.32	GGE	RMI1	ILAE consortium on complex epilepsies 2022.
12q13.13	GGE	SCN8A	ILAE consortium on complex epilepsies 2022.
16p13.3	GGE	RBFOX1	ILAE consortium on complex epilepsies 2022.
17p13.1	GGE	ARHGEF15	ILAE consortium on complex epilepsies 2022.
19p13.3	GGE	AP3D1	ILAE consortium on complex epilepsies 2022.

21q21.1	GGE	TMPRSS15	ILAE consortium on complex epilepsies 2022.
22q13.32	GGE	FAM19A5	ILAE consortium on complex epilepsies 2022.
7q21.11	All	GRM3	Song *et al.* 2021.
8p23.1	All	TNKS	Song *et al.* 2021.

Another notable GWAS was conducted by Song *et al.*[31] in 2021. They did a meta-analysis based on data from the 2018 ILAE GWAS[33], UK biobank[42], Japanese population[43] and FINNGEN[44] for a total of 26 352 cases and 774 517 controls. They discovered three significant risk loci. One at 2q24.3 which was identified in every ILAE GWAS. They also had significant association at 7q21.11 and 8p23.1, both novel associations. Risk genes in those regions include GRM3 and TNKS. A TWAS was also conducted and significant association between brain expression regulation of TNKS, TTC21B and RP11-375N15.2 and epilepsy was detected.

Rare genetic variations have also been studied in epilepsy. However, this was only done for coding regions through whole exome sequencing (WES). Notably, a study by Feng *et al.*[32] conducted in 2019 screened the WES of 9 170 cases and 8 436 controls. Although this study did not discover any new risk loci or genes, they showed that epilepsy patients had an enrichment of ultra-rare deleterious variants based on pLI score[45]. In the end this study showed that rare coding variants are likely to be risks factor for the disease, and that to deepen our understanding of epilepsy there is a need to investigate more than common variants.

There is a need for cohorts to be as large as possible to allow new discoveries, reaching as much as tens of thousands of individuals. However, even with large cohorts new discoveries are scarce, and studies tend to mostly replicate previous findings. This is especially surprising since the heritability of epilepsy is high, estimated at around 80% in a twin study[46], yet our current knowledge allows us to explain about 32% of the heritability in GGE and 9% in NAFE[33]. Indeed, a large proportion of the heritability remains unexplained by common variants and coding variants. Despite this, rare non-coding variants have never been studied in epilepsy. Similarly, asthma has an estimated heritability of around 70%[47,48] and a SNP-based heritability estimated from the UK Biobank GWAS of 14%[49]. The same pattern is encountered where a lot of common variants of small effects are associated with the disease, leading to a lack of power to detect those variants[50]. However, efforts have been made to use WGS to study rare non-coding SNV and those efforts were successful[51–53]. Such studies have yet to be performed in epilepsy.

Finally, throughout most genetic studies of epilepsy GGE and NAFE are analyzed separately. The reason for this is that those syndromes seem to have drastically different underlying genetic mechanisms. Indeed, most genes associated to epilepsy are specific to either GGE or NAFE with little overlap between the phenotype[3,11,31–33,54–56]. Furthermore, as mentioned above a greater proportion of the heritability is explained for GGE and in most studies more new discoveries were made for GGE than NAFE. Because of this, it is thought that GGE have a greater genetic burden[2,3,11,33,57].

Non-coding regions and their functional impact

The human genome can be divided in two parts, coding, and non-coding regions. Coding regions are the sequences that compose the exons and that have the potential to be

transcribed and translated into protein. Those regions compose a little less than 2% of the genome. The rest is called non-coding DNA and those regions will never be translated into protein. Historically, non-coding regions were largely ignored, even earning the title of 'Junk DNA' as they were regarded as having no importance[58]. However, in the last decade more and more studies showed the functional impact that those regions can have on gene expression level, thus having a non-negligible role on phenotype determination[27–29].

Even though non-coding single nucleotide variants (SNV) don't alter the structure of protein they can impact gene expression through multiple mechanisms. A change of the sequence in a transcription factor binding region can alter the affinity of the transcription factor, therefore increasing or decreasing gene transcription depending on the change in affinity. A change in non-coding sequence can also impact the histone marks that will be bound to the histones, which will affect the availability of the DNA to the transcription complex, hence altering gene expression. Those are a couple of examples from a large number of ways in which non-coding variants can have a functional impact on gene expression[27,28,36,37,59,60]. Studies on psychiatric disorders have already demonstrated an enrichment of both *de novo* mutations and inherited variants. For example, in autism spectrum disorder most WGS studies looked at *de novo* mutations in regulatory regions, while a couple of studies were able to show enrichment in inherited variants[61–64]. In developmental disorder *de novo* mutations were also enriched in regulatory regions[65]. In schizophrenia, the sequencing of open chromatin regions (ATAC-seq) led to the identification of regulatory risk variants[66]. The effects of non-coding regulatory variants have also been studied and demonstrated in traits unrelated to the nervous system, like cardiovascular disease[67–69], arthritis[70,71], plasma protein levels[72] and height[73].

The effect of many non-coding variants on gene expression have been quantified. In the last decade multiple large-scale projects like GTEx[36] and ENCODE[37] have identified and quantified the tissue specific effect of thousands of expression Quantitative Trait Loci (eQTL) on gene expression. An eQTL is a SNV that influences gene expression. A property of eQTLs is that their expression change effects are tissue specific. This is because of the natural variation between the different tissues in transcription factors, histone marks, epigenetic signature, etc. More specifically this means that the same eQTL can have a drastically different impact in the brain cortex, than in muscle tissue, which is why tissue specificity is a crucial feature to consider when studying eQTLs.

Finally, studying non-coding regions is a notable challenge due to the gargantuan size of those regions coupled with the fact that it is harder to predict the pathogenicity of non-coding SNV. As opposed to coding variants, where it is easy to see if the variant will cause a missense or a loss of function, finding eQTLs requires large tissue collection effort, which makes it especially hard to work with neurological disorders, since brain tissue is primarily available post-mortem[74].

Artificial Intelligence in genetics

In recent years artificial intelligence (AI) has been blossoming. With a wide range of applications, it was only a question of time before AI could be used in genetics. An interesting use of AI is to predict the effect of SNV on gene expression, in other words to predict how a change in the nucleotide sequence might affect the transcription level of neighbouring genes. Accomplishing this is no small feat, mainly because of the numerous ways by which genomic sequence can interact with gene expression[27–29]. Indeed, to make reliable predictions an AI would require a massive amount of data to be trained on, like chromatin

immunoprecipitation, RNA-seq, Hi-C, eQTL, etc. Those are all methods that yield data on chromatin accessibility, transcription level, DNA-protein interaction and expression change. Multiple algorithms that are able to predict expression change have been developed in the past 5 years[59,75–77]. They can make tissue specific expression fold change predictions for any SNV. These algorithms are of great value for the study of non-coding genomic regions. They can be used to prioritize variants of great functional interest, thus simplifying the study of those large regions. Indeed, those regions are vast, which leads to a significant loss of power due to multiple comparisons correction. Furthermore, there is little to no information on the role of non-coding regions, which is another challenge to assess the clinical impact of non-coding variants. Finally, the aforementioned algorithms have the benefit of yielding informative data without having to acquire tissue samples from patients. This characteristic is particularly useful in neurological disorders because of how arduous it is to access brain tissue[74].

Application of deep learning in genetics

Examples of algorithms that can make predictions of tissue specific expression change caused by variants include ExPecto[59], Enformer[75], Xpresso[76] and Basenji[77]. They have one common characteristic, and it is that they all use deep learning, which is a specific type of AI. Deep learning works by doing computation with a neural network[78]. A neural network can be pictured as a giant grid of nodes that are connected with each other, similarly to the way that neurons are connected in the brain, hence the name (Figure 1). A neural network uses layers to achieve its computations. There are three types of layers, the input layer, the hidden layers and the output layer[79,80]. The input layer is the first layer of a neural network, it contains a number of nodes equal to the number of input variables. This layer

receives the data and sends it to the hidden layers. The hidden layers are the layers where all the computations occur. At every node some computations are made on the data, and it is then sent to the next layer. Every time that data is transferred between nodes a coefficient is used to weight the data that is transferred. That weight is what is calibrated when the model is trained. The number of hidden layers depends on the complexity of what the model tries to predict, the closer the prediction is to a linear relation the smaller is the required number of layers (between one to three). The more complex the relation, the greater is the required number of layers (from 5 to dozens). Finally, the output layer is the layer which holds the information of interest that we wanted our model to predict.

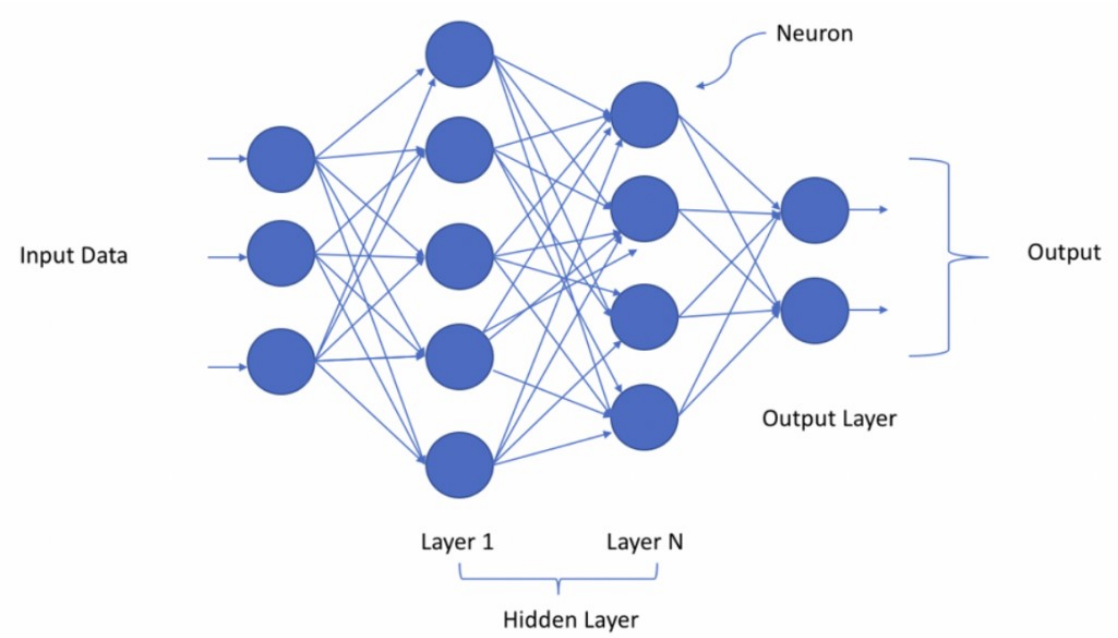


Figure 1. Schematic visualisation of a deep learning neural network.[81]

To be trained a model needs data corresponding to the expected input and have the associated output to evaluate the performance. The algorithm will then use these data to test

different weights combinations. It will do so randomly at first, but it will slowly progress toward an optimized solution by considering the previous attempts. In the end, the trained model will be tested on data that it has never been exposed to. This procedure is there to ensure that the performance of the model is not reliant on prior exposure to the data. This phenomenon can be caused by underfitting, overfitting during training, or poor training data. Underfitting is what happens when the model finds a solution that is too simplistic. A simple example would be that the model finds a linear solution to a problem with a higher degree of complexity. In that case the trained model would not be able to properly consider all the variables and their relations, thus resulting in poor performance. This seldom occurs when using deep learning due to the inherent complexity of the training process, but it can happen, mainly if there is a bias in the training data. Overfitting occurs when the model overoptimizes and finds a solution so specific and specialized that it can only be applied to the training data. Any attempt to use the model on other data lead to bad performance because the high specificity prevents generalization. Overfitting is often caused by a poor choice of parameters in the training algorithm. Therefore, the model needs to be trained again with different parameters. Finally, if the training data are bad the model will be equally as bad. The performance is heavily dependent on the quality of the data. If there is a bias in the training set it will translate to the model, thus hampering the accuracy of its predictions.

Objectives

The main objective of this study is to assess the impact of non-coding regions on the genetic etiology of epilepsy. More specifically whether these regions have an effect in a specific subtype of epilepsy (GGE or NAFE). To do so, the functional impact of rare non-coding variants have to be assessed with a deep learning algorithm. Therefore, one of the

objective of this project is to use the deep learning algorithm ExPecto[82] to predict gene expression change for brain and neurological tissues for every SNV in our dataset.

Materials and Methods

Please note that the following chapter expands upon the 'Materials and methods' section of the article. Therefore, readers who go through this chapter will have access to detailed explanations of the methodology, but they may find the 'Materials and methods' section of the article redundant.

Cohort

Description

The cohort used in this study is the Canadian Epilepsy Network (CENet) cohort[54,83–86]. It has a multitude of genetic data from genotyping to WGS. For the purpose of this project only WGS data from unrelated GGE and NAFE patients was used. Controls come from the same cohort, a part of the original CENet project consisted of WGS trios of Developmental Epileptic Encephalopathy (DEE) composed of both parents and the affected child. Those parents are good controls since the main contributing factor for DEE are de novo mutations[65,86–89], thus parents are not expected to carry a strong genetic burden that could reduce statistical power. Moreover, there are advantages to choosing the DEE parents as controls. First, the parents from the DEE trios are more representative of the patients' population because of the Québec founder effect[90,91] and at the time the study was conducted no population reference cohort of WGS existed for the Québec population. Additionally, it ensures that there is no batch effect in the data. Both of those factors remove potential bias and sources of false positive error from the analyses.

The WGS part of the CENet cohort is composed of 1155 total samples including patients and controls. After data cleaning, phenotype-based filtering and removing related individuals (which are all described below), 624 samples were remaining (Table 2). Some of the patients are labeled as having a mixed phenotype. What this means is that those people

are either NAFE or GGE, but their phenotype is different from other family members. For instance, a GGE patient who comes from a family of NAFE patients.

Table 2. Number of individuals for each phenotype

Phenotype	Male	Female	Total	Mean age	Median age
Controls	190	187	377	-	-
All Cases	112	135	247	45	41
GGE	52	71	123	45	41
NAFE	50	62	112	40	36
Mixed	10	2	12	54	59
Total	302 (48%)	322 (52%)	624		

Number of individuals by phenotype and by sex. Age data was not available for controls and was available for only a portion of the cases. Age was calculated in 2023.

Phenotyping

Patients' diagnosis was conducted by epileptologists in centre hospitalier de l'Université de Montréal research center in Montréal. Blood samples were collected at this site. Control blood samples were collected in centre hospitalier universitaire Sainte-Justine in Montréal and Hospital for Sick Children in Toronto. ILAE guidelines[22,23] were used for diagnosis as described in the 'Introduction', in the 'Epilepsy' section, under 'Disease phenotype'.

Sequencing

DNA was extracted from the blood samples and was sent to Genome Quebec Innovation Center in Montreal for sequencing. WGS was made at 30X coverage. Genomic DNA (gDNA) was cleaned using ZR-96 DNA Clean & ConcentratorTM-5 Kit (Zymo) prior to being quantified using the Quant-iTTM PicoGreen dsDNA Assay Kit (Life Technologies) and its integrity assessed on agarose gels. Libraries were generated using the TruSeq DNA PCR-Free Library Preparation Kit (Illumina) according to the manufacturer's recommendations. Libraries were quantified using the Quant-iTTM PicoGreen Double Stranded DNA (dsDNA) Assay Kit (Life Technologies) and the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (Kapa Biosystems). The average size fragment was determined using a LabChip GX (PerkinElmer) instrument. The libraries were denatured in 0.05N NaOH and diluted to 8pM using HT1 buffer. The clustering was done on an Illumina cBot and the flowcell was run on a HiSeq 2500 for 2×125 cycles (paired-end mode) using v4 chemistry and following the manufacturer's instructions. A phiX library was used as a control and mixed with libraries at 0.01 level. The Illumina control software used was HCS 2.2.58 and the real-time analysis program used was RTA v. 1.18.64. bcl2fastq v1.8.4 was used to demultiplex samples and generate fastq reads. The filtered reads were aligned to reference Homo_sapiens assembly b37. Each readset was aligned using BWA-MEM version 0.7.10 to create a Binary Alignment Map file (.bam). Bam files were processed to gvcf files and we performed joint calling of gvcf files that were merged into a single vcf file using GATK version 3.7-0[92]. The vcf file was recalibrated and filtered following the GATK best practice guidelines.

Statistical analyses

Generating Data

ExPecto[82] was used in order to generate expression change data for every SNV in our cohort (over 55 million). The algorithm was used without any modification to the source code. ExPecto runs in two phases. The first one consists of the deep learning neural network. It uses Python 3.8 and will predict the effect of the variation in the sequence on the general level of transcription. To do so the neural network was trained on 2 002 different histone marks, transcription factors and DNA accessibility profiles from over 200 tissues and cell types. Training performed by Zhou *et al.*[82] was used for this phase. The output of the neural network is spatial feature transformation data, that will be used in the next step. The second phase is a tissue specific regularized model that uses the spatial feature transformation data to compute the tissue specific expression change data in natural log fold change. By combining those two steps the model can predict expression change for 218 tissues and cell types that come from either GTEx[36] or ENCODE[37]. Predictions were made for the 26 tissues that are related to the brain or neural cells and 3 outgroup tissues of different embryonic origin to be used for comparison purposes (Artery Aorta, Colon Transverse and Skin of Body) (Figure 2).

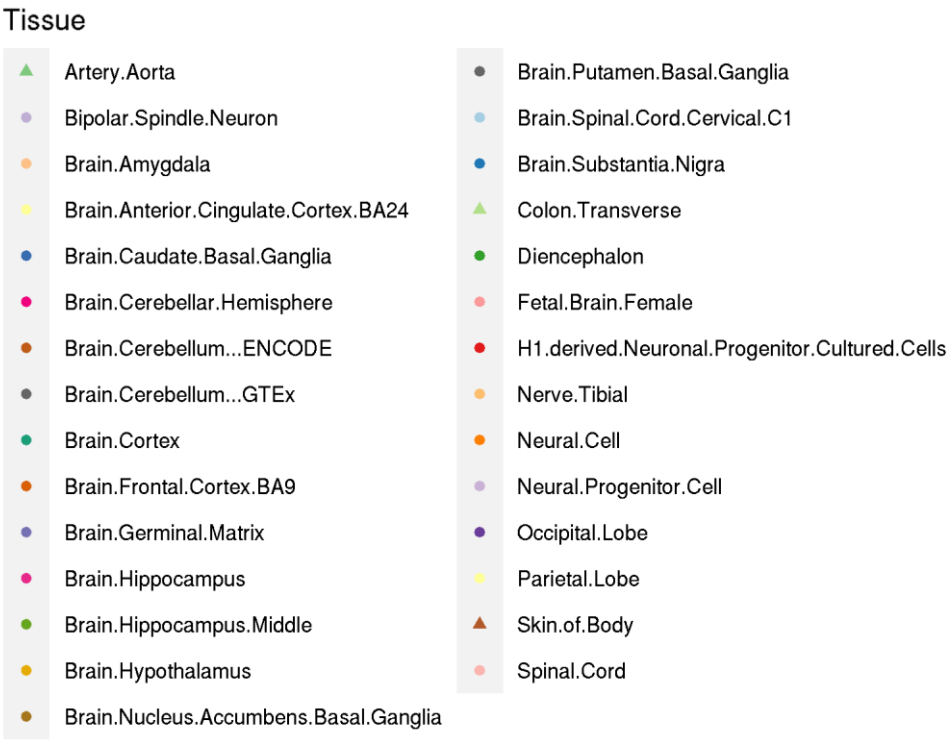


Figure 2. List of tissues for which predictions were made with ExPecto.

Tissues marked with a triangle are from the outgroup. Colors serve as a legend for figure 3.

In the end, 29 predictions were made for over 55 million SNVs, which means that around 1,6 billion tissue specific expression change predictions were made (Figure 3).

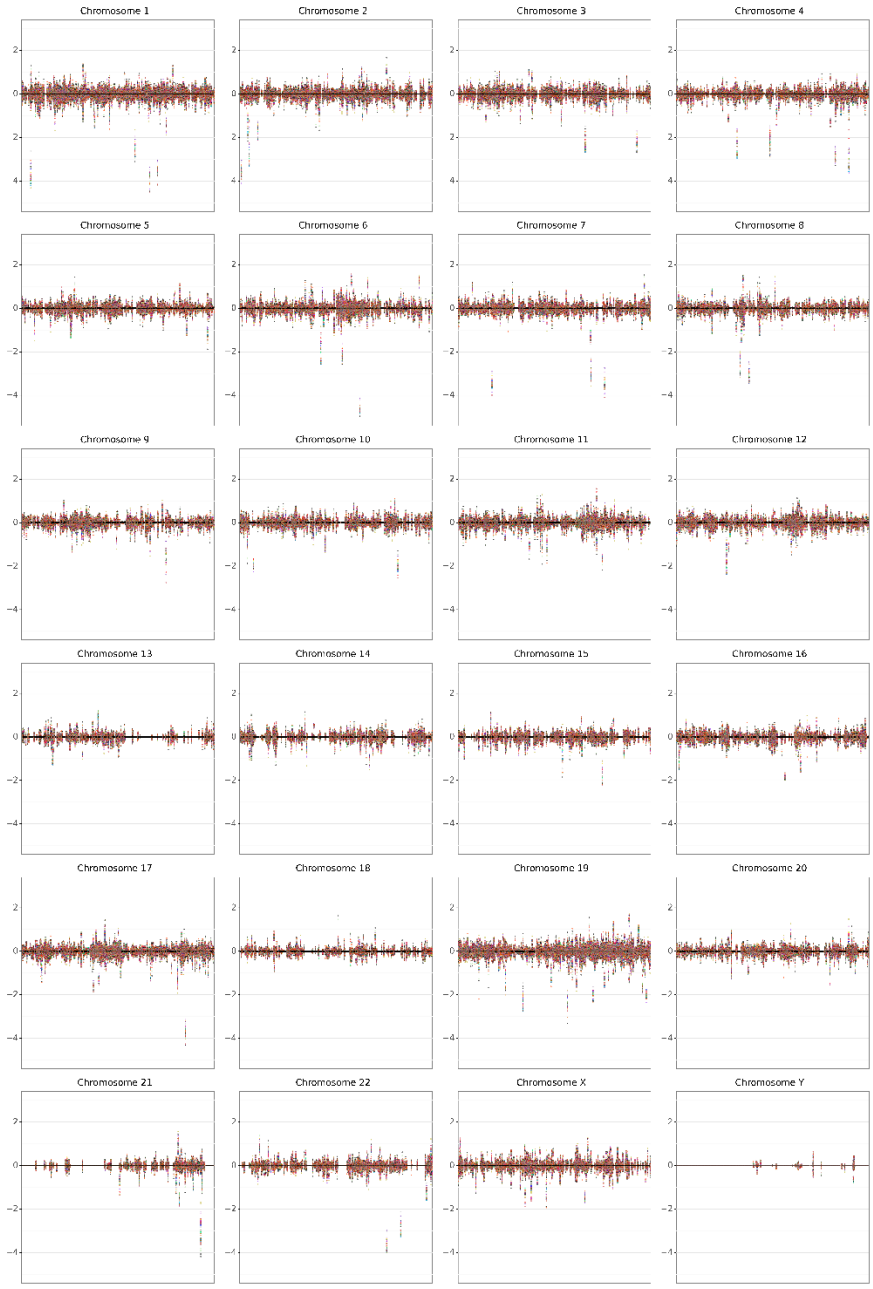


Figure 3. Visual representation of around 1,6 billion expression change values across the genome.

The y axis corresponds to the natural log expression fold change. The x axis corresponds to genomic positions.

ExPecto validation

To ensure that the results are accurate, the model was tested on known eQTLs from the GTEx v6p database[36]. To do this the prediction's directionality was compared to the observed directionality. The directionality is defined as the direction of the change in expression, in other words it corresponds to whether the transcription level increases or decreases. Then the proportion of variants for which the directionality of the prediction and the observation match is the accuracy. This can be plotted against another measure called magnitude. Magnitude corresponds to the absolute expression change, meaning that it is strictly positive. By doing this it is possible to see the relation between accuracy and magnitude, as well as a way to identify a critical magnitude value above which the accuracy is optimal (figure 4). The accuracy is perfect for a magnitude greater than 0.2, thus only variants with a magnitude above this threshold were used in the following analyses.

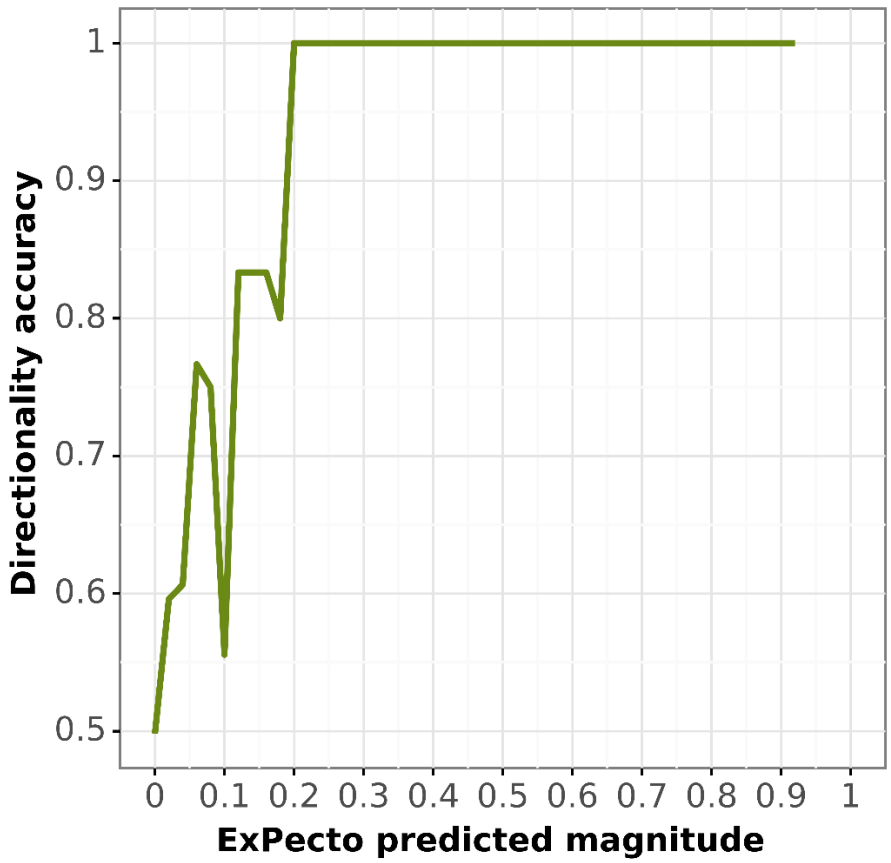


Figure 4. Accuracy of predictions' directionality on known GTEx eQTLs

Directionality accuracy was computed according to ExPecto's predicted magnitude in natural log fold change.

Constraint Violation Score

As mentioned earlier, it is a challenge to assess whether a change in gene expression will have a neutral, beneficial, or deleterious effect[29,93]. To achieve this, expression change values are transformed into constraint violation scores (CVS). A CVS considers selective pressure to quantify the effect of gene expression. Computation of CVS requires a tissue specific expression change prediction for a gene and that gene's variation potential. The variation potential is the sum of the predicted gene expression change in the natural log for

all possible SNV 1 kb upstream and downstream of the transcription start site. Since a variant that decreases expression will have a negative predicted value and a variant that increases expression will have a positive value, the sum will tend towards 0 if there is no imbalance between the two. However, if there is a selective pressure on the gene that results in a high expression level in the tissue of interest, most variants will cause transcription levels to drop because it is already optimized for high expression. Therefore, there will be more negative terms in the sum leading to a negative variation potential. The higher the selective pressure, the higher the variation potential will be. The inverse is also true, thus if a gene is under selective pressure for low expression in the tissue, most variants will increase its expression, leading to a positive variation potential. The CVS is obtained by multiplying the predicted gene expression change with the variation potential of the associated gene for the relevant tissue. The CVS can either be positive or negative. A positive CVS means that the variant is deleterious because the direction of the change goes against the selective pressure. A negative CVS means that the variant is beneficial because the direction of the change goes with the selective pressure. The fact that it is obtained by a multiplication ensures that the magnitude of the expression change, and the magnitude of the selective pressure are both considered.

To ensure that the CVS had good tissue specificity, CVS distribution was compared in patients with epilepsy between the three outgroup tissues and the three tissues associated with epilepsy (hippocampus, amygdala and brain cortex)[11,94–99]. All three epilepsy related tissues had a distinctive density pike around CVS 30 that is absent from the outgroup tissues (Figure 5). This confirmed that the tissue sensitivity of the model is good.

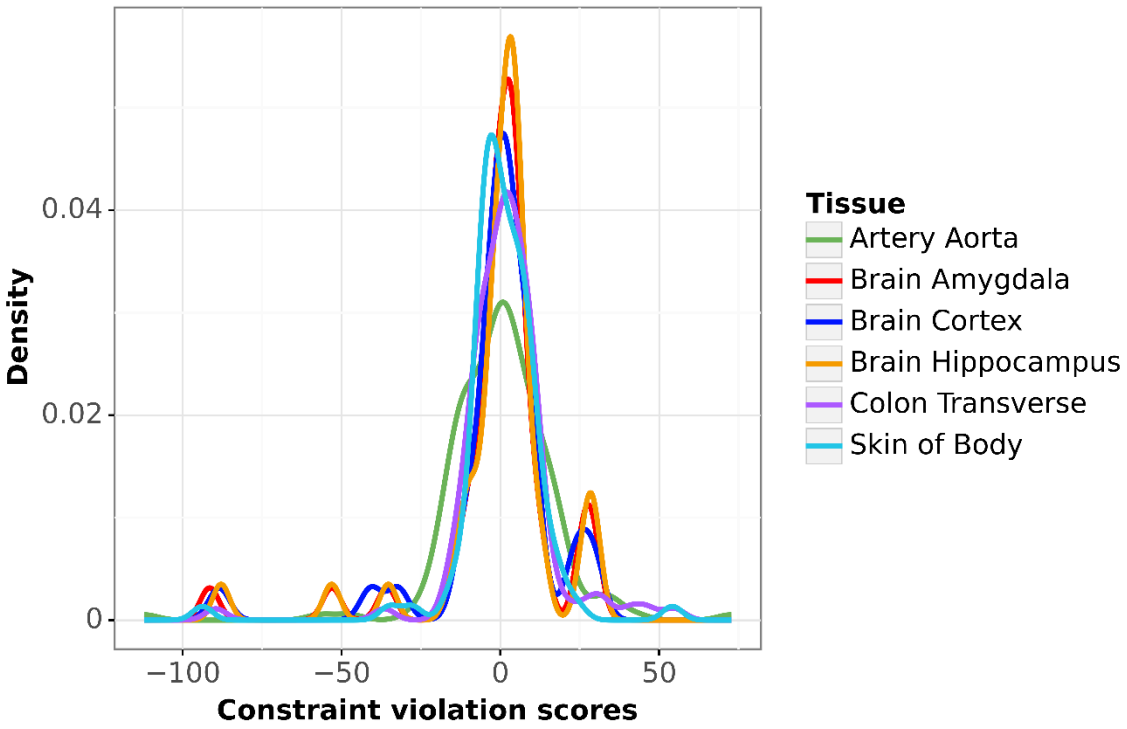


Figure 5. Density curve of the constraint violation scores for three epilepsy related tissues and three outgroup tissues.

Data cleaning

Cleaning was made using plink v2.0[100]. First, Single Nucleotide Variants (SNVs) with a call rate below 98% were removed using '--geno 0.02'. Afterward, individuals with a genotype rate below 98% were excluded using '--mind 0.02'. Next, SNV that did not follow the Hardy-Weinberg equilibrium were removed using '--hwe 0.001'. Finally, individuals with unknown biological sex were excluded. SNVs were filtered based on their minor allele frequency (maf), only rare variants (maf <0.01[101–103]) as assessed in our cohort as well as in gnomAD were kept [104].

Population structure

To ensure that population structure didn't cause bias to subsequent analyses, a two-dimension Uniform Manifold Approximation and Projection (UMAP) was made to be used

as a covariable. A UMAP is created from a principal component analysis (PCA) and is a method used to reduce the number of dimensions and capture more complexity with fewer variables. To make the PCA, further cleaning was required. The PCA was made with plink v2.0[100]. Only common variants were used in the PCA (maf > 0.05[102]) and SNVs that were not in linkage disequilibrium by using '--indep-pairwise 50 5 0.2'. A plot of PC1 and PC2 shows a logical separation of self-declared ethnicity, and it does not show signs of bias (Figure 6).

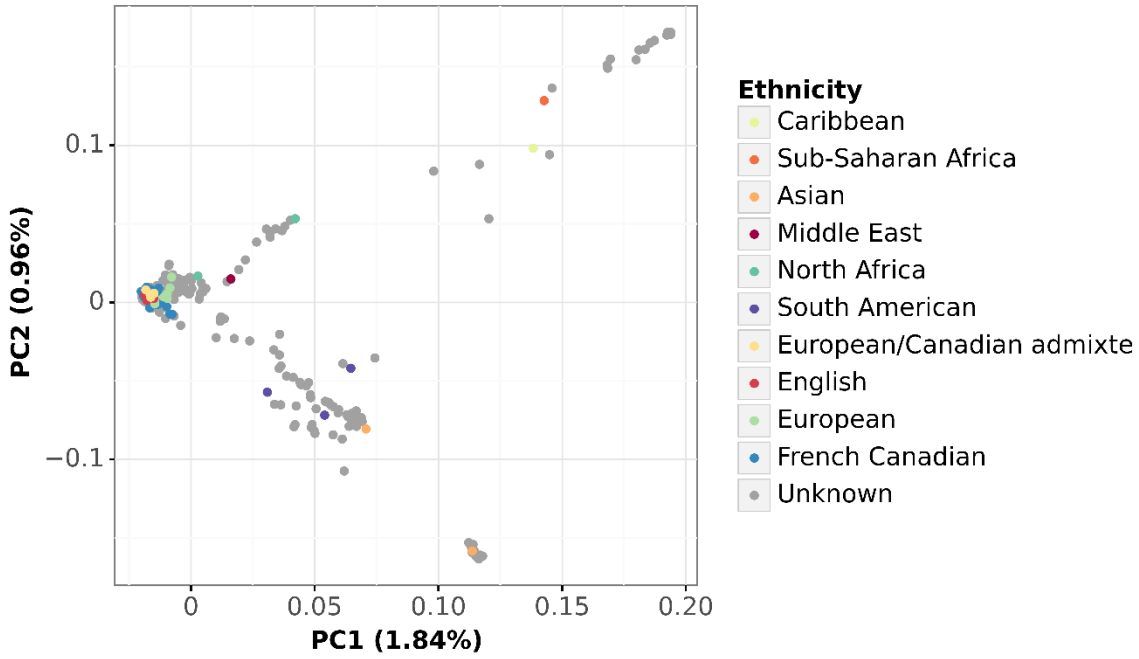


Figure 6. Principal component analysis of PC1 and PC2.
Variance explained by the PC is in the name of the axis. Colors represent self-reported ethnicity.

The UMAP was made from the first five PCs of the PCA (figure 7). This decision was made since after those PCs the variance explained by each PC tends to remain stable,

thus indicating that the PCs beyond the first five explain mainly individual relations instead of populational relations (Figure 8).

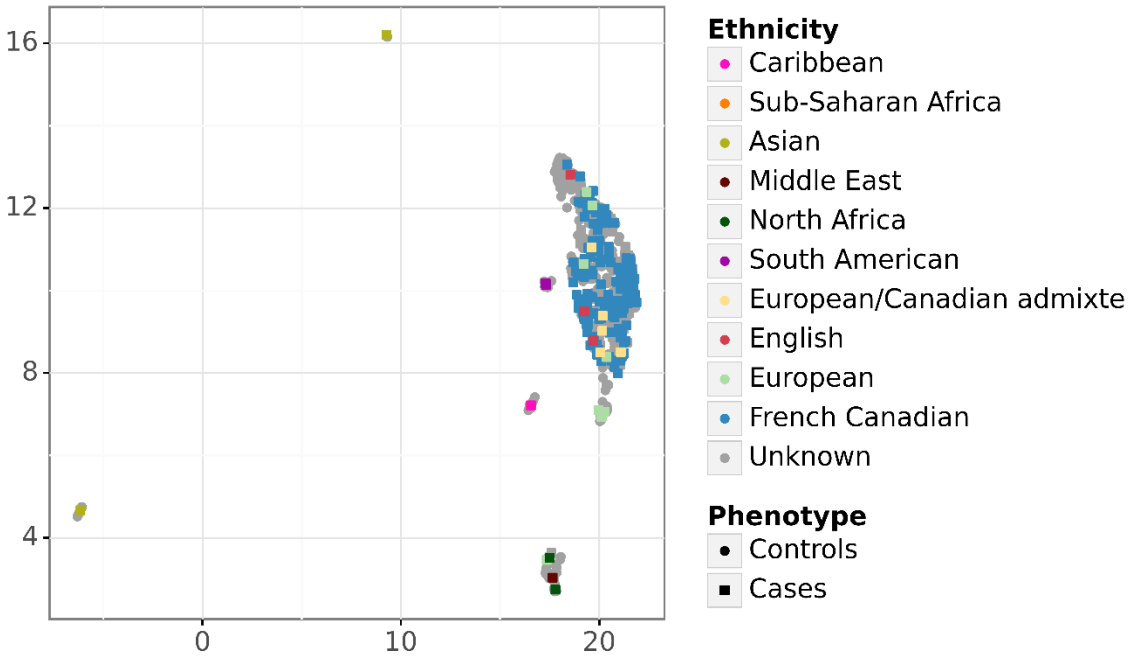


Figure 7. UMAP of ethnicity for the epilepsy patients and controls.

Colors represent self-reported ethnicity and shape represent the phenotype.

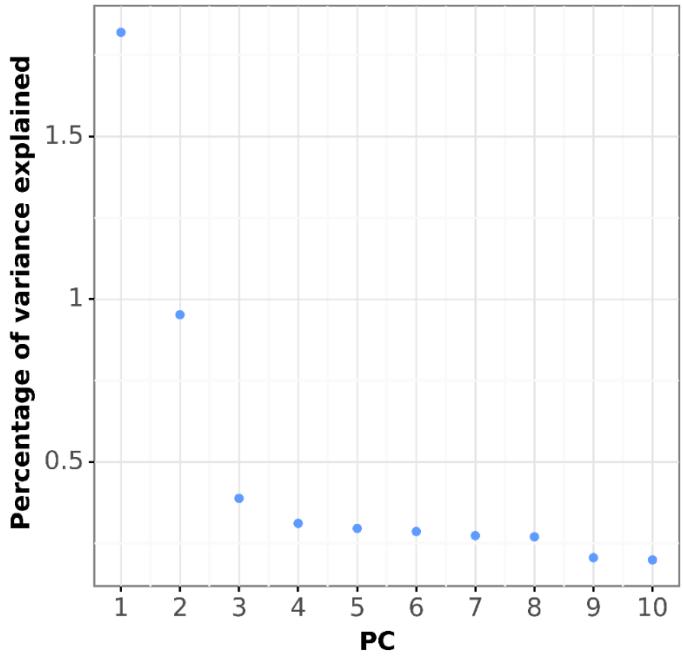


Figure 8. Scree plot for the first 10 PCs.

The python package umap-learn v0.5.1 was used to make the UMAP. Parameters were set to n_neighbours = 624 and min_dsit = 0. Maximizing the number of neighbours considered favorized the emergence of population structure over relations between small groups of people. Minimizing the minimum distance was unrestricting and did not force the algorithm to create distance between individuals, thus giving a better depiction of reality.

Logistic regressions

To compare the burden of rare non-coding variants between cases and controls, logistic regressions were performed at different CVS windows. For each window a patient was considered as exposed if he had at least one variant for which the median CVS score for the hippocampus, amygdala and brain cortex is inside that window. The analysis was performed for windows of]10, 20];]20, 30];]30, 40]; > 40. These analyses were performed for all cases (GGE, NAFE and mixed) against controls, GGE against controls, NAFE against

controls and GGE against NAFE. Statsmodels v0.12.2 was used to do the logistic regressions. Both UMAP dimensions and sex were used as covariables. Analyses were repeated with only individuals of European descent to ensure that signals were not the result of an ethnic bias (Article - S4 Fig). Analyses were also repeated by using the median of the outgroup tissues (artery aorta, colon transverse and skin of body) to confirm that significant signals are tissue specific (Article - S2 Fig).

Article - Unraveling the role of non-coding rare variants in epilepsy

Article's context

The following article presents the aforementioned study in a more concise manner. The paper was submitted to PLOS ONE on April 27, 2023. This version of the paper was resubmitted on July 26, 2023 following the review process. Supplementary figures can be found after the **References** section of the Article. The supplementary table is not directly included in this work due to its size, however it can be accessed through the preprint of an earlier and shorter version of the paper (https://doi.org/10.1101/2022.12.13.22283363).

Personal contribution to the paper

Patients' selection, sample extractions and sequencing were all performed before the start of this project, consequently I did not partake in those steps of the study. I played a role in the conception of the study. I conducted all of the mentioned analyses and produced all of the figures. I wrote the first version of the paper and improved it with the input of the coauthors.

Unraveling the role of non-coding rare variants in epilepsy

Alexandre Girard[1], Claudia Moreau[1], Jacques L. Michaud[2, 3], Berge Minassian[4, 5], Patrick Cossette[6, 7], Simon L. Girard[1, 8]

[1]University of Quebec in Chicoutimi, Centre Intersectoriel en Santé Durable, Saguenay, Canada

[2]CHU Sainte-Justine, Montréal, Canada

[3]University of Montreal, Department of Neurosciences and Department of Pediatrics, Montréal, Canada

[4]The Hospital for Sick Children, Department of Pediatrics, Toronto, Canada

[5]University of Texas Southwestern Medical School, Department of Pediatrics, Dallas, United States of America

[6]CHUM Research Center, Montréal, Canada

[7]University of Montreal, Department of Neurosciences, Montréal, Canada

[8]Laval University, CERVO Research Center, Québec, Canada

Corresponding author

Simon Girard, PhD

E-mail: simon2_girard@uqac.ca

Résumé

La découverte de nouveaux variants a atteint un plateau dans les dernières années dans les études sur l'épilepsie, malgré l'utilisation de cohortes de très grandes tailles. Conséquemment, la majorité de l'héritabilité reste inexpliquée. Les variants non-codants rares ont été largement ignorés dans les études sur l'épilepsie, même si les variants non-codants d'un seul nucléotide peuvent avoir un impact significatif sur l'expression des gènes. Nous avons eu accès à du séquençage de génome complet (WGS) de 247 patients épileptiques et 377 témoins. Pour déterminer l'impact fonctionnel des variants non-codants, ExPecto, un algorithme d'apprentissage profond a été utilisé pour prédire le changement d'expression dans des tissus cérébraux. Nous avons comparé le fardeau des variants rares délétères non-codants entre les cas et les témoins. Les variants rares hautement délétères non-codants étaient significativement enrichis pour l'épilepsie génétique généralisée (GGE), mais pas pour l'épilepsie focale non-acquise (NAFE) ou tous les cas d'épilepsie comparés aux contrôles. Dans cette étude nous avons démontré que les variants rares délétères non-codants sont associés à l'épilepsie, spécifiquement pour les GGE. De plus grandes cohortes de WGS en épilepsie seront requises pour investiguer ces effets avec une plus grande résolution. Néanmoins, nous avons démontré l'importance d'étudier les régions non-codantes en épilepsie, une maladie où les nouvelles découvertes sont rares.

Abstract

The discovery of new variants has leveled off in recent years in epilepsy studies, despite the use of very large cohorts. Consequently, most of the heritability is still unexplained. Rare non-coding variants have been largely ignored in studies on epilepsy, although non-coding single nucleotide variants can have a significant impact on gene expression.

We had access to whole genome sequencing (WGS) from 247 epilepsy patients and 377 controls. To assess the functional impact of non-coding variants, ExPecto, a deep learning algorithm was used to predict expression change in brain tissues.

We compared the burden of rare non-coding deleterious variants between cases and controls. Rare non-coding highly deleterious variants were significantly enriched in Genetic Generalized Epilepsy (GGE), but not in Non-Acquired Focal Epilepsy (NAFE) or all epilepsy cases when compared with controls.

In this study we showed that rare non-coding deleterious variants are associated with epilepsy, specifically with GGE. Larger WGS epilepsy cohort will be needed to investigate those effects at a greater resolution. Nevertheless, we demonstrated the importance of studying non-coding regions in epilepsy, a disease where new discoveries are scarce.

Introduction

Epilepsy is a neurological disorder characterized by epileptic seizures and spontaneous episodes of abnormal neuronal activity [1,2]. Approximately 3% of all individuals will be affected during their lifetime [3,4]. The vast majority of genetic epilepsies are complex traits (>98%); traits that are affected by a plethora of genomic regions. With such a large array of contributing signals, it is an arduous task to detect significant associations. Several studies used familial trios to investigate de novo mutations, thus leading to the association of multiple genes with the disease [5–7]. However, new variant discoveries rarely meet expectations notably in recent epilepsy studies. Only large cohorts composed of tens of thousands of individuals had some success [4,8–13]. These studies mainly focused on common or coding variants. The overwhelming majority of studies in epilepsy use either genotyping or exome sequencing to investigate the genetic causes of the disease. Consequently, little is known concerning the implication of non-coding regions in the etiology of the disease [4,8–13]. However, these regions were shown to have an important impact on the phenotype of an individual as they affect the expression of neighboring genes [14–18]. As a large portion of the heritability of epilepsy remains unexplained, there is a glaring need to study the impact of rare non-coding variants in epilepsy.

Since non-coding regions are so vast, a strategy to prioritize variants of interest is to investigate the impact of expression quantitative trait loci (eQTL). Studying eQTL in neurological disease is a notable challenge, mainly because eQTLs effects are tissue specific and brain tissues are mostly available post-mortem [19]. Nevertheless, progress in deep learning now allows us to predict the functional effects of variants from sequencing data

without having to sample tissues from our patients [20–22]. In this study, we aimed to characterize the role of rare non-coding variants in epilepsy based on their functional effect in brain tissues using one of the most powerful deep learning algorithm, ExPecto [21]. We used whole genome sequencing (WGS) data from the Canadian Epilepsy Network (CENet) cohort to investigate deleterious rare functional variants in epileptic patients [23].

Materials and methods

Cohort phenotyping

The CENet cohort is composed of patients with Genetic Generalized Epilepsy (GGE) or Non-Acquired Focal Epilepsy (NAFE) collected in CHUM Research Center in Montreal and controls (unaffected Developmental Epileptic Encephalopathy (DEE) trio parents) collected in CHU Ste-Justine in Montreal and the Hospital for Sick Children in Toronto [20–23]. The patients were recruited between 2002 and 2014. Patients were diagnosed by epileptologists. The clinical epilepsy phenotype was classified according to the current classification by the International League against Epilepsy (ILAE) [24]. More specifically for NAFE, patients were at least five years of age and have experienced at least two unprovoked seizures in the six months prior to starting treatment, an MRI scan of the brain that did not demonstrate any potentially epileptogenic lesions, other than mesial temporal sclerosis. Patients with clinical and EEG characteristics meeting the 1989 ILAE syndrome definitions for GGE were included. An MRI of the brain was not required for participation. All patients were at least four years of age at the time of diagnosis. In GGE, we also included patients with Jeavons syndrome, which is an idiopathic generalized form of reflex epilepsy characterized by childhood onset, unique seizure manifestations, striking light sensitivity and possible occurrence of generalized tonic-clonic seizures. Certain cases were found with an epilepsy

phenotype different from the other affected family members, hence they were marked as 'mixed'. Only one affected GGE or NAFE patient was used per family, therefore all the individuals in this study are unrelated. We used WGS from 377 controls and 247 patients, 123 GGE, 112 NAFE and 12 mixed patients (Table 1). This study was approved by the CHUM research Center (CRCHUM) ethics committee and written informed consent was obtained for all patients (2003-1394, ND 02.058 -BSP (CA)). We did not have access to information that could allow us to identify the patients.

Table 1. Number of individuals for each phenotype

Phenotype	Male	Female	Total
Controls	190	187	377
All Cases	112	135	247
GGE	52	71	123
NAFE	50	62	112
Mixed	10	2	12

Sequencing

DNA was extracted at the time of recruitment. All samples were sequenced at the same time in 2015. Samples were sequenced for the whole genome at 30X coverage at Genome Quebec Innovation Center in Montreal. gDNA was cleaned using ZR-96 DNA Clean & ConcentratorTM-5 Kit (Zymo) prior to being quantified using the Quant-iTTM PicoGreen

dsDNA Assay Kit (Life Technologies) and its integrity assessed on agarose gels. Libraries were generated using the TruSeq DNA PCR-Free Library Preparation Kit (Illumina) according to the manufacturer's recommendations. Libraries were quantified using the Quant-iTTM PicoGreen dsDNA Assay Kit (Life Technologies) and the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (Kapa Biosystems). The average size fragment was determined using a LabChip GX (PerkinElmer) instrument. The libraries were denatured in 0.05N NaOH and diluted to 8pM using HT1 buffer. The clustering was done on an Illumina cBot and the flowcell was run on a HiSeq 2500 for 2×125 cycles (paired-end mode) using v4 chemistry and following the manufacturer's instructions. A phiX library was used as a control and mixed with libraries at 0.01 level. The Illumina control software used was HCS 2.2.58 and the real-time analysis program used was RTA v. 1.18.64. bcl2fastq v1.8.4 was used to demultiplex samples and generate fastq reads. The filtered reads were aligned to reference Homo_sapiens assembly b37. Each readset was aligned using BWA-MEM version 0.7.10 to create a Binary Alignment Map file (.bam). Bam files were processed to gvcf files and we performed joint calling of gvcf files that were merged into a single vcf file using GATK version 3.7-0 [25]. The vcf file was recalibrated and filtered following the GATK best practice guidelines.

Data cleaning

Cleaning was made using plink v2.0 [26]. First, Single Nucleotide Variants (SNV) with a call rate below 98% were removed using '--geno 0.02'. Afterward, individuals with a genotype rate below 98% were excluded using '--mind 0.02'. Next, SNV that did not follow the Hardy-Weinberg equilibrium were removed using '--hwe 0.001'. Finally, individuals

with unknown biological sex were excluded. To make the principal component analysis, further cleaning was required. Only common variants were used in the PCA (maf > 0.05) and SNV that were not in linkage disequilibrium by using '--indep-pairwise 50 5 0.2'. Single nucleotide variants (SNVs) were filtered based on their minor allele frequency (maf), only rare variants (maf <0.01) as assessed in our cohort as well as in gnomAD were kept [27].

Statistical Analyses

We applied ExPecto on rare variants (MAF<0.01) for three tissues related to epilepsy: hippocampus, amygdala and brain cortex (GTEx V6) for which we calculated the gene expression change median [16,31]. ExPecto computes gene expression changes by using a neural network to predict the effect of variants on features such as transcription factors, histone marks and DNA accessibility. It then transforms those feature predictions in tissue specific gene expression changes with L2-regularized linear regression models. Gene expression changes calculated by ExPecto were used to compute a Constraint Violation Score (CVS) in accordance with the methods described in Zhou et al. [21]. The CVS quantifies how deleterious the gene expression change is: the higher the score, the more deleterious the variant.

The accuracy of ExPecto predictions was validated using known eQTLs from the GTEx V6p release (S1 Fig) [16]. To do so, we used two parameters, the prediction's directionality and magnitude. Directionality is defined as whether the Single Nucleotide Variant (SNV) increases or decreases gene expression. Magnitude is defined as the absolute size of the effect in gene expression fold change (natural log). We determined that the magnitude above which

the accuracy of the prediction's directionality was perfect was 0.2 so we kept only variants with a median (for the cortex, hippocampus, and amygdala) above this threshold (S1 Fig). Analyses were replicated with the median of three non-neurological tissues: artery aorta, colon transverse and skin of body to validate the tissue specificity of the model (S2 Fig).

A binomial logistic regression was performed with the python package statsmodels v0.12.2. Sex was used as a covariate as well as a two-dimension UMAP (umap-learn v0.5.1) based on the first five principal components (plink v2.0) (S3 Fig). We had access to self-declared ethnicity from the affected individuals, which allowed us to confirm that the UMAP was accurate. Analyses were replicated by using only individuals of French-Canadian and European descent to validate that the findings were not related to population structure (S4 Fig). Those individuals were selected based on the cluster that corresponded with self-declared French-Canadians and Europeans.

Results

We compared the proportion of patients and controls who had at least one rare variant (MAF<0.01) at different CVS thresholds using a binomial logistic regression analysis to compute odds ratios (OR). We repeated the analysis to compare GGE with controls, NAFE with controls and GGE with NAFE (mixed patients were removed from these analyses). Variants found in the final analysis are available in S1 Table. Our analyses showed no significant difference when comparing cases and controls (Fig 1A). Nevertheless, the OR tends to increase with the CVS threshold and reaches a peak for variants with a CVS above 40 (OR 1.54; 95% CI 0.77-3.11). However, there is a significant difference between GGE

and controls for CVS above 40 (OR 2.74; 95% CI 1.20-6.22) (Fig 1B). On the other hand, NAFE and controls show no significant difference, meaning that the trend observed when comparing all cases and controls was solely driven by the GGE (Fig 1C). Next, GGE and NAFE have a significantly different burden for CVS thresholds of]20, 30] (OR 2.47; 95% CI 1.10-5.52) and above 40 (OR 3.19; 95% CI 1.002-10.13) (Fig 1D). Finally, it is worth noting that none of those signals are significant after multiple tests correction, with the lowest adjusted p-value of 0.066 at CVS over 40 for GGE against controls (Bonferroni correction). Nevertheless, this particular signal is robust to multiple tests correction when using only individuals of French-Canadian and European descent with an adjusted p-value of 0.044 (S4 Fig). No significant signals weres observed when repeating the analyses with non-neurological tissues, thus demonstrating the reliability and power of the tissue specific model (S2 Fig).

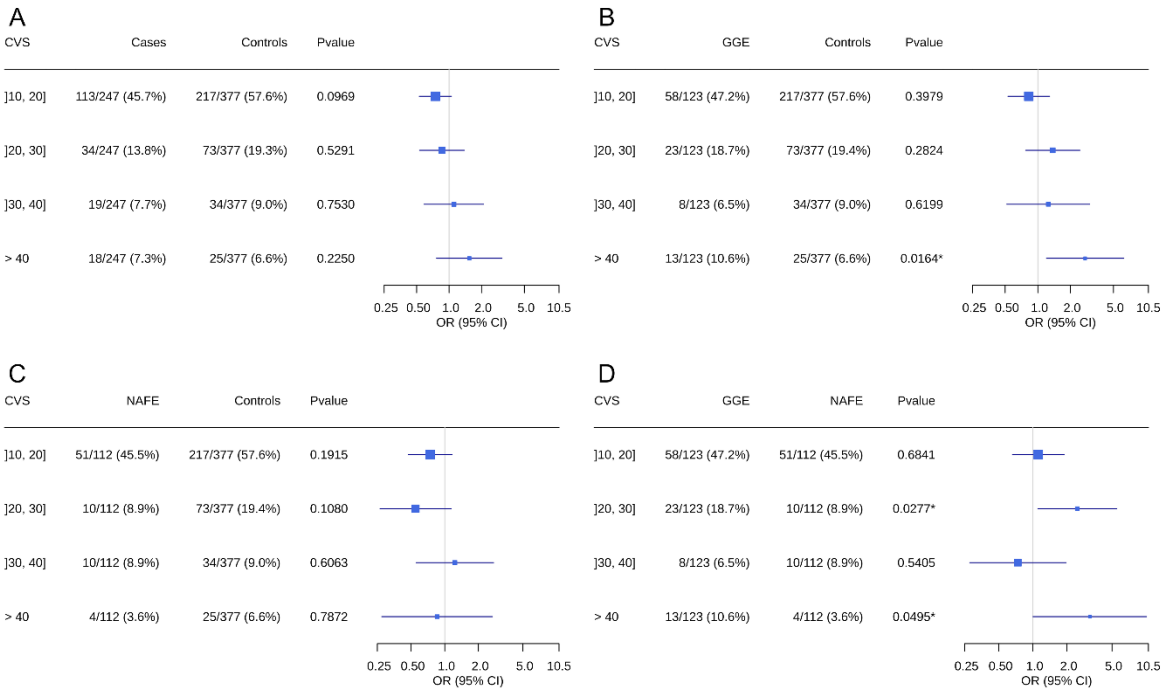


Fig 1. Burden of variants for different CVS thresholds across epilepsy phenotypes. Odds ratios and p-value were calculated using a binomial logistic regression for variants of different Constraint Violation Score (CVS) thresholds. Lines represent 95% confidence intervals. Comparisons were made for cases and controls (A), Genetic Generalized Epilepsy (GGE) and controls (B), Non-Acquired Focal Epilepsy (NAFE) and controls (C) and GGE and NAFE (D).

Discussion

We found strong evidence that deleterious non-coding rare variants with a higher CVS are enriched in GGE. However, it is not the case for NAFE who are known to have a smaller genetic burden [4,8–11]. Once again, our findings demonstrate the importance of separating GGE and NAFE when studying epilepsy, which is also supported by most associated genes

being specific to subphenotypes [23,32,33]. Furthermore, the differences observed when directly comparing GGE to NAFE highlights the varying impact of non-coding variants on those subphenotypes.

We are the first to highlight the potential impact of rare non-coding SNV on a genome-wide scale in epilepsy. This discovery reveals the importance of studying non-coding regions which may explain a part of the missing heritability in epilepsy [9]. Moreover, it showcases the need to conduct more studies on WGS in order to make discoveries at a higher resolution in non-coding regions.

In addition to the contribution that we bring to the field of epilepsy, the method used in this study could be applicable to other diseases to provide a better understanding of the role of rare non-coding SNV in various pathologies.

Limitations

The study has two main limitations. First, the use of deep learning, as useful as it is, has the limitation of being predictions, not observations. Nonetheless, we validated that those predictions were accurate by using experimental data from GTEx (S1 fig). Additionally, ExPecto is not able to compute long range (>40kb) sequence effects on gene expression, which limits this work to the study of short-range interactions. The second limitation is our small sample size. Despite this, we were successful in identifying a genome-wide effect in individuals of European descent, but we lacked power to investigate those effects at a gene or variant level resolution.

Conclusion

This study reveals the importance of non-coding regions in the etiology of epilepsy. The effect was specific for GGE, whereas NAFE showed no significant difference with controls. Therefore, our results indicate that the differences between those subphenotypes extends to non-coding genetic mechanisms. Larger WGS cohorts will be needed to deepen our understanding of the role of non-coding regions in epilepsy.

Data Availability

Raw whole genome sequences of all epilepsy patients for which we have appropriate consent have been deposited in the European Genome-phenome Archive, under the accession code EGAS00001002825.

Acknowledgements

We would like to thank all the participants of this study, without whom none of this work would have been possible.

References

1. Fisher RS, Acevedo C, Arzimanoglou A, Bogacz A, Cross JH, Elger CE, et al. ILAE Official Report: A practical clinical definition of epilepsy. Epilepsia. 2014;55(4):475–82.

2. Chang BS, Lowenstein DH. Epilepsy. New England Journal of Medicine. 2003 Sep 25;349(13):1257–66.

3. Perucca P, Bahlo M, Berkovic SF. The Genetics of Epilepsy. Annual Review of Genomics and Human Genetics. 2020;21(1):205–30.

4. The International League Against Epilepsy Consortium. Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. The Lancet Neurology. 2014;13(9):893.

5. Zhu X, Padmanabhan R, Copeland B, Bridgers J, Ren Z, Kamalakaran S, et al. A case-control collapsing analysis identifies epilepsy genes implicated in trio sequencing studies focused on de novo mutations. PLOS Genetics. 2017 Nov 29;13(11):e1007104.

6. Wong JKL, Gui H, Kwok M, Ng PW, Lui CHT, Baum L, et al. Rare variants and de novo variants in mesial temporal lobe epilepsy with hippocampal sclerosis. Neurology Genetics [Internet]. 2018 Jun 1 [cited 2023 Jul 6];4(3). Available from: https://ng.neurology.org/content/4/3/e245

7. Consortium ER, Project EP, Consortium E. De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. American journal of human genetics. 2014 Oct 1;95(4):360–70.

8. Feng YCA, Howrigan DP, Abbott LE, Tashman K, Cerrato F, Singh T, et al. Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals. The American Journal of Human Genetics. 2019 Aug 1;105(2):267–82.

9. The International League Against Epilepsy Consortium on Complex Epilepsies, Abou-Khalil B, Auce P, Avbersek A, Bahlo M, Balding DJ, et al. Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. Nat Commun. 2018 Dec 10;9(1):5269.

10. Song M, Liu J, Yang Y, Lv L, Li W, Luo XJ. Genome-Wide Meta-Analysis Identifies Two Novel Risk Loci for Epilepsy. Front Neurosci. 2021 Aug 12;15:722592.

11. International League Against Epilepsy Consortium on Complex Epilepsies, Berkovic SF, Cavalleri GL, Koeleman BP. Genome-wide meta-analysis of over 29,000 people with epilepsy reveals 26 loci and subtype-specific genetic architecture [Internet]. medRxiv; 2022

[cited 2022 Jul 26]. p. 2022.06.08.22276120. Available from: https://www.medrxiv.org/content/10.1101/2022.06.08.22276120v1

12. Ishigaki K, Akiyama M, Kanai M, Takahashi A, Kawakami E, Sugishita H, et al. Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. Nat Genet. 2020 Jul;52(7):669–79.

13. Suzuki T, Koike Y, Ashikawa K, Otomo N, Takahashi A, Aoi T, et al. Genome-wide association study of epilepsy in a Japanese population identified an associated region at chromosome 12q24. Epilepsia. 2021;62(6):1391–400.

14. Pagni S, Mills JD, Frankish A, Mudge JM, Sisodiya SM. Non-coding regulatory elements: Potential roles in disease and the case of epilepsy. Neuropathology and Applied Neurobiology. 2022;48(3):e12775.

15. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, et al. Coordinated Effects of Sequence Variation on DNA Binding, Chromatin Structure, and Transcription. Science. 2013 Nov 8;342(6159):744–7.

16. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013 Jun;45(6):580–5.

17. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep;489(7414):57–74.

18. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017 Oct;550(7675):204–13.

19. Lewis DA. The Human Brain Revisited: Opportunities and Challenges in Postmortem Studies of Psychiatric Disorders. Neuropsychopharmacol. 2002 Feb;26(2):143–54.

20. Agarwal V, Shendure J. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. Cell Reports [Internet]. 2020 May 19 [cited 2022 Nov 26];31(7). Available from: https://www.cell.com/cell-reports/abstract/S2211-1247(20)30616-1

21. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet. 2018 Aug;50(8):1171–9.

22. Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. 2018 May;28(5):739–50.

23. Moreau C, Rébillard RM, Wolking S, Michaud J, Tremblay F, Girard A, et al. Polygenic risk scores of several subtypes of epilepsies in a founder population. Neurol Genet. 2020 Mar 27;6(3):e416.

24. Hamdan FF, Myers CT, Cossette P, Lemay P, Spiegelman D, Laporte AD, et al. High Rate of Recurrent De Novo Mutations in Developmental and Epileptic Encephalopathies. Am J Hum Genet. 2017 Nov 2;101(5):664–85.

25. Monlong J, Girard SL, Meloche C, Cadieux-Dion M, Andrade DM, Lafreniere RG, et al. Global characterization of copy number variants in epilepsy patients from whole genome sequencing. PLoS Genet. 2018 Apr;14(4):e1007285.

26. Moreau C, Tremblay F, Wolking S, Girard A, Laprise C, Hamdan FF, et al. Assessment of burden and segregation profiles of CNVs in patients with epilepsy. Ann Clin Transl Neurol. 2022 Jun 8;9(7):1050–8.

27. Berg AT, Berkovic SF, Brodie MJ, Buchhalter J, Cross JH, van Emde Boas W, et al. Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE Commission on Classification and Terminology, 2005-2009. Epilepsia. 2010 Apr;51(4):676–85.

28. van der Auwera G, O'Connor BD. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra [Internet]. O'Reilly Media, Incorporated; 2020. Available from: https://books.google.ca/books?id=wwiCswEACAAJ

29. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.

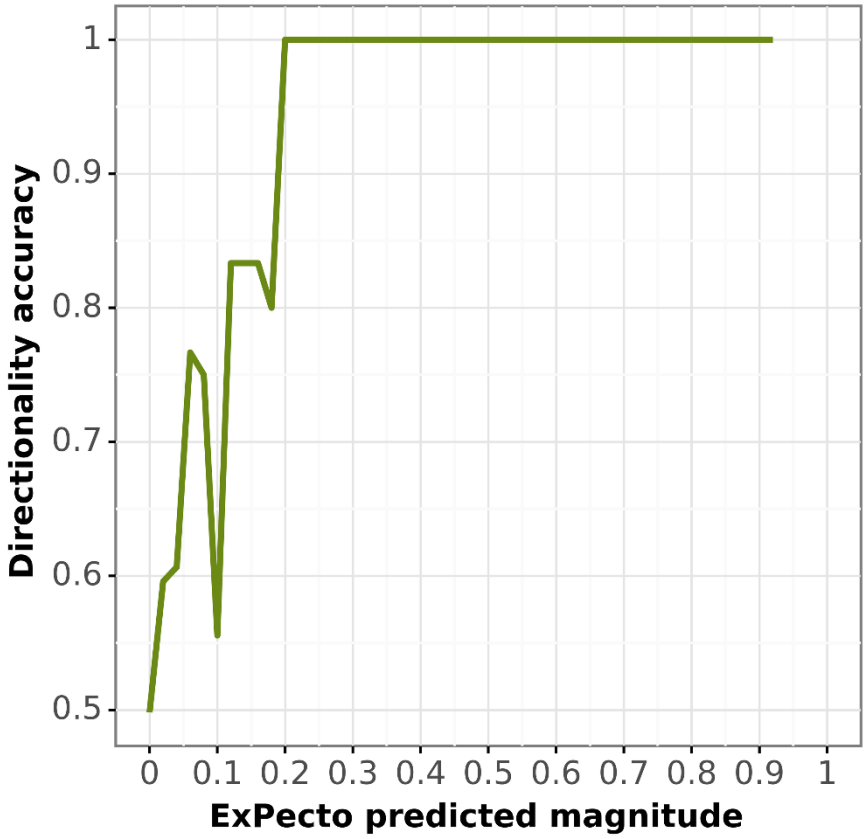
30. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434–43.

31. Turrin NP, Rivest S. Innate immune reaction in response to seizures: implications for the neuropathology associated with epilepsy. Neurobiology of Disease. 2004 Jul 1;16(2):321–34.

32. Koko M, Krause R, Sander T, Bobbili DR, Nothnagel M, May P, et al. Distinct gene-set burden patterns underlie common generalized and focal epilepsies. eBioMedicine [Internet]. 2021 Oct 1 [cited 2022 Nov 18];72. Available from: https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(21)00381-9/fulltext

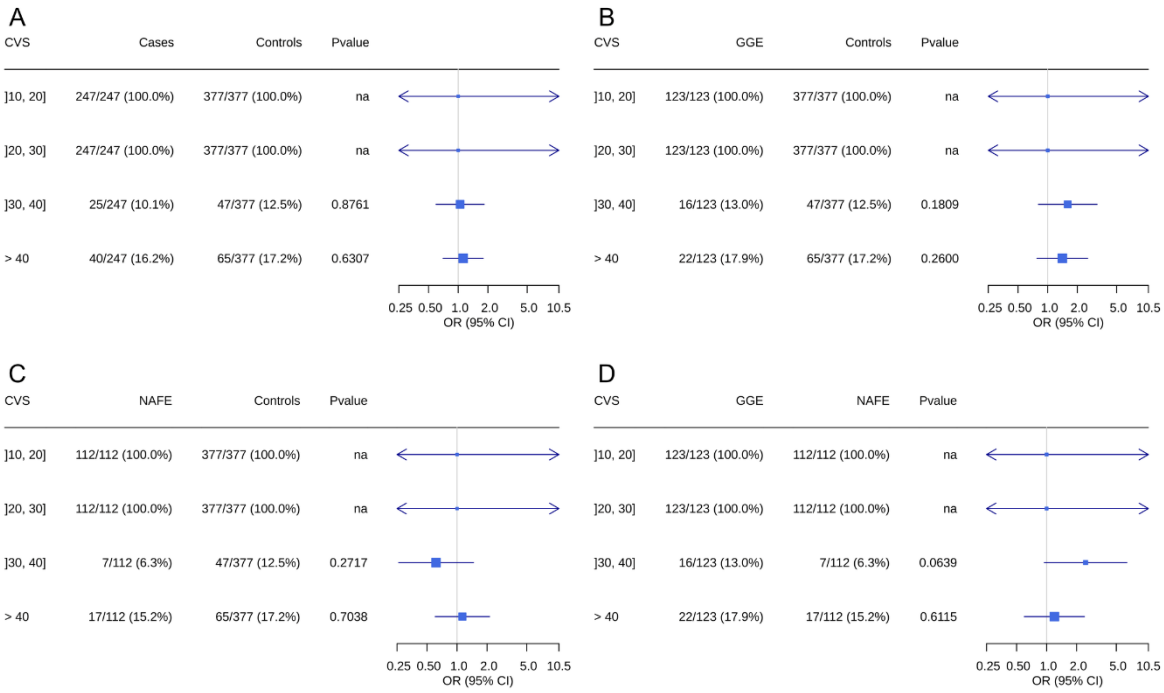
33. Motelow JE, Povysil G, Dhindsa RS, Stanley KE, Allen AS, Feng YCA, et al. Sub-genic intolerance, ClinVar, and the epilepsies: A whole-exome sequencing study of 29,165 individuals. The American Journal of Human Genetics. 2021 Jun 3;108(6):965–82.

Supporting information



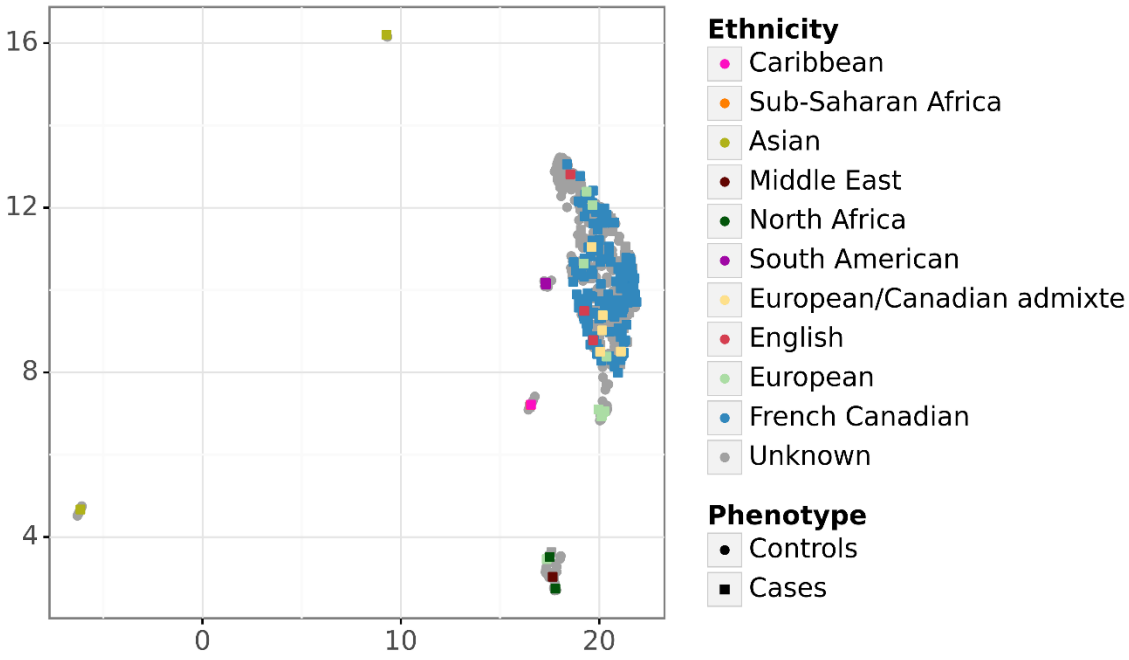
S1 Fig. Accuracy of predictions' directionality on known GTEx eQTLs

Directionality accuracy was computed according to ExPecto's predicted magnitude in natural log fold change.



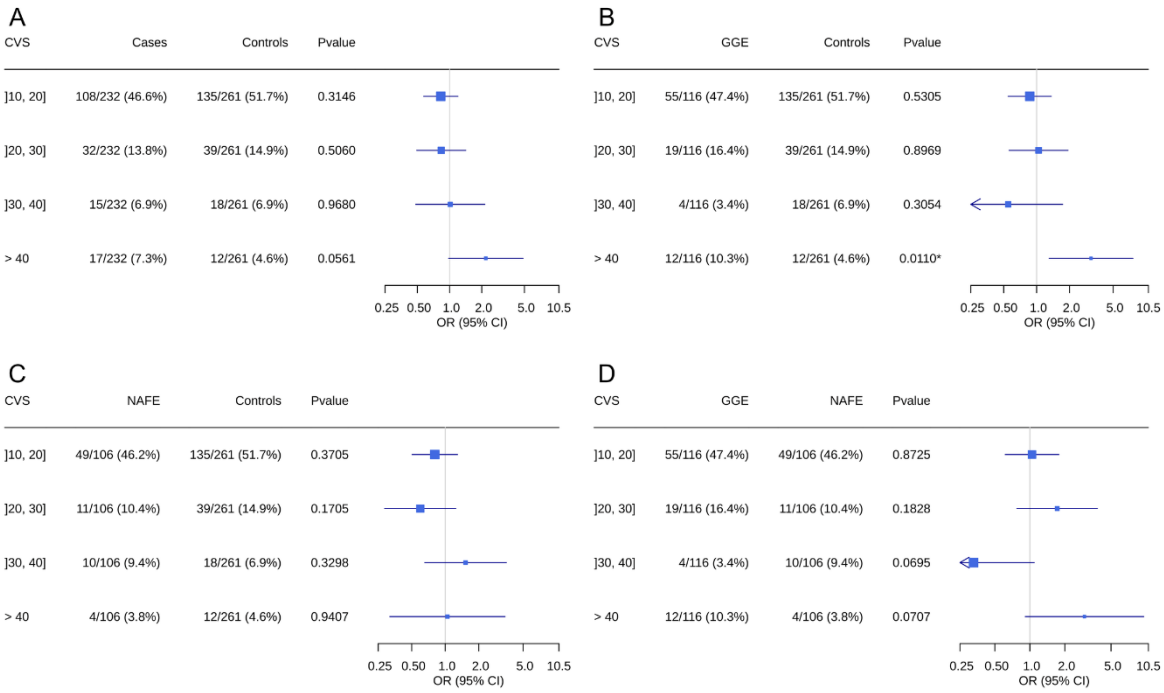
S2 Fig. Burden of variants for different CVS thresholds across epilepsy phenotypes when using non-neurological tissues.

Odds ratios and p-value were calculated using a binomial logistic regression for variants of different Constraint Violation Score (CVS) thresholds. Lines represent 95% confidence intervals. Comparisons were made for cases and controls (A), Genetic Generalized Epilepsy (GGE) and controls (B), Non-Acquired Focal Epilepsy (NAFE) and controls (C) and GGE and NAFE (D). Tissues that were used are artery aorta, colon transverse and skin of body.



S3 Fig. UMAP of ethnicity for the epilepsy patients and controls.

The UMAP was made with 'umap-learn v0.5.1' and based on the first 5 principal components.



S4 Fig. Burden of variants for different CVS thresholds across epilepsy phenotypes whit only individuals of European descent.

Odds ratios and p-value were calculated using a binomial logistic regression for variants of different Constraint Violation Score (CVS) thresholds. Lines represent 95% confidence intervals. Comparisons were made for cases and controls (A), Genetic Generalized Epilepsy (GGE) and controls (B), Non-Acquired Focal Epilepsy (NAFE) and controls (C) and GGE and NAFE (D).

S1 Table. List of variants included in the final analysis.

Discussion

Logistic regression interpretation

All cases against controls

In the analyses that compare all cases to controls, no significant difference was observed for any CVS windows (Article - Fig. 1A). Even though it was expected that epilepsy patients would have a greater burden of rare non-coding variants than controls, it does not come as a surprise that the sample size used in this study is too small to achieve significance when combining epilepsy subtypes. Indeed, even studies with sample sizes over 100 times greater than the one in this project yield a few significant associations when comparing all patients to controls[3,11,31–33]. By assuming that the computed odds ratio is right a sample size of 2 641 individuals would have been required to achieve 80% power, whereas such a level of power was reached when comparing only GGE to controls with a sample size of 500 (G*Power v3.1.9.7). This demonstrates that due to the great heterogeneity between GGE and NAFE, analyses have more power when they consider only one subtype of the disease at a time[2,34].

Nevertheless, a clear trend can be observed, as the CVS windows increase the corresponding odds ratio increases too. This indicates that with a greater sample size, significance level could have been reached for current windows and it could be possible to look at higher CVS windows to investigate the effect of even more deleterious variants.

Genetic generalized epilepsy against controls

When comparing only GGE with controls, odds ratios were higher for all CVS windows, thus indicating that the previously discussed trend was mainly driven by the genetic burden of GGE patients (Article – Fig. 1B). This was expected since studies conducted on GGE with genotyping or exome sequencing have both shown that GGE patients have a

stronger genetic component than NAFE, as well as a stronger heritability[2,3,11,31–34]. Therefore, it is of no surprise that this genetic disparity extends to rare non-coding variants.

Furthermore, GGE patients have a significantly higher genetic burden for highly deleterious variants (CVS > 40). This result is strong evidence for the role of rare non-coding variants in the etiology of GGE. This is a hallmark in the study of epilepsy as it is the first time that the effect of rare non-coding variants has been studied in the disease. This is surely the first of many steps in demonstrating the importance of those variants for GGE and deepening our understanding of the disease.

Non-acquired focal epilepsy against controls

The comparison of NAFE to controls revealed no significant enrichment (Article - Fig. 1C). As opposed to previous analyses, there is no trend of increasing odds ratios with increasing CVS windows. As of now, there is no indication that rare non-coding variants are associated with NAFE. However, genetic heterogeneity is greater between NAFE subtypes than between GGE subtypes, which means that there could be a significant loss of power due to divergent genetic mechanisms across the different NAFE subtypes[11]. With the current sample size, it was not possible to further stratify the NAFE patients into more phenotypes. Despite this, the heritability is nonetheless lower in NAFE, which could be explained by the recent evidences that demonstrate that somatic mutations have a significant role in the development of focal epilepsies[105–111]. Even though the entire picture is still unclear, based on current knowledge it is undeniable that NAFE heritability is smaller than GGE and that NAFE patients don't seem to have a genetic burden associated with rare non-coding variants.

Genetic generalized epilepsy against non-acquired focal epilepsy

Comparing GGE to NAFE highlights the main CVS windows that differ between those two epilepsy subtypes (Article – Fig. 1D). For highly deleterious variants (CVS > 40), there is a significant difference between GGE and NAFE, which was expected due to the very different genetic burden discussed above. The confidence interval is quite large, which is a consequence of the small sample size.

The genetic burden of variants with CVS from 30 to 40 is very similar between the two phenotypes, but there is once again a significant difference when looking at CVS from 20 to 30. This is unexpected, since windows of significance are separated by a non-significant window. To definitively answer this question, findings will have to be replicated in larger WGS cohorts to increase power and resolution. Since it is the first time that rare non-coding variants are investigated in epilepsy, no literature exists on the subject. A hypothesis was formulated: It is possible that the observation is due to the fact that GGE don't necessarily have a higher number of deleterious rare non-coding variants than NAFE, but that a higher proportion of those variants are highly deleterious which increases their risk of developing the disease. The significance at CVS 20 to 30 could be due to the same phenomenon, which cause GGE patients to have more moderately deleterious variants (CVS 20 to 30), but not necessarily more lightly deleterious variants (CVS 10 to 20). Once again, more studies will need to be conducted on the subject before the mechanism causing this disparity in the type of variants enriched in GGE can be better understood.

Principal contributions

Importance in the field

This research brings an important contribution to the field of epilepsy, it is the first time that rare non-coding variants are studied in the disease. The results show a significant enrichment of those variants in GGE, thus demonstrating that a portion of the missing heritability is likely explained by those variants. This work lays the foundations for more and bigger WGS cohort to delve into non-coding regions. Indeed, larger sample size will be needed to study those regions at a higher resolution by using statistical tools developed for large-scale WGS data set[112–114]. In the future it will be possible to conduct meta-analyses when more WGS datasets will be created. A new tool that can achieve variant specific resolution for rare variants meta-anlysis was recently published[115]. Until such power can be obtained, single cohort studies of larger size will have to use genetic burden across shifting windows to associate loci to the disease, as it has been historically done for the study of rare variants[113,116–123]. Although the cost of WGS is still considerably greater than genotyping, the continuously decreasing price of WGS combined with the development of new technologies, might spark an increase in the utilization of WGS[124–128]. With methods that keep getting more cost-efficient and faster WGS becomes a real possibility when creating a cohort[129–131]. In the end, with the growing accessibility of WGS and the biological insight it can provide to better understand the genetic risk of complex diseases, it is of the uttermost importance that other research groups consider creating WGS cohorts of epileptic patients.

From a clinical standpoint, WGS are rarely used, and non-coding variants are often ignored. However, annotation of non-coding variants has improved enough to give valuable insight on the role of those variants[132]. Indeed, the functional role of many non-coding

regions have been put into light in recent years[133–135]. With this, new guidelines are being established to identify non-coding pathogenic variants which can help with fine diagnosis of diseases[136]. Additionally long non-coding RNAs have been recognized as drug targets in certain diseases[137]. Those applications further emphasize the importance of studying non-coding regions in epilepsy.

Exportability of the method

The method used in this project can easily be applied to other cohorts, of epilepsy patients or of people afflicted with another disease. Indeed, the only required data to conduct those analyses are WGS data of the patients. Study on other diseases could benefit from using this method as it would allow them to investigate the effect of rare non-coding variants. Examples of diseases for which the method would yield useful information are schizophrenia[138], Alzheimer[139] and diabetes[140] to name a few. In those diseases for which WGS studies are scarce, a large proportion of the heritability is missing[141–143], and past studies used WGS mainly to study *de novo* mutations[144–149], copy number variants[150–157] and non-coding RNA[158–163]. Similarly, to epilepsy a part of the missing heritability of those diseases probably resides in rare non-coding variants. The method proposed in this project allows for efficient prioritization of non-coding variants with functional effect, thus simplifying the study of non-coding regions. Furthermore, it opens the possibility of having tissue specific expression data without having to take tissue sample from patients. This is especially beneficial when the organ of interest is hard to access like the brain. Therefore, this method is of great interest for neurological disorders as they are complex traits, they often have high heritability and a high proportion of it is missing[164,165].

Limitations

Artificial Intelligence

Even though artificial intelligence is a method with revolutionary potential it also has limitations[166,167]. Indeed, an AI model's performances are limited by the amount and quality of available data. If there is a bias in the data, the model will be biased as well. Additionally, if the model encounters a situation that is not covered by the training data it will perform poorly. However, due to validations made by comparing predicted data to experimental data, both in this study and in the original ExPecto paper[82], predictions made in this study seem to be robust. It is certain that future algorithms will have increased performance as computing power continues to grow and more data is available to train models. For this reason, research group interested in applying this method to their dataset should probably use the state-of-the-art algorithm at the time of their study instead of ExPecto, which might be deemed too old at some point in the future.

Sample Size

The most limiting factor of this study is its small sample size. The cohort used in this project is only a fraction of the size of those used in other epilepsy studies[3,11,31–34,43]. The sample size is certainly limiting when it comes to power. It is hard to make new discoveries in epilepsy, mainly because of the small effect size of individual variants and the genetic heterogeneity between patients[168]. Therefore, with the sample size of this study it was not possible to look at the genome with a variant or locus specific resolution. This is why only genome-wide enrichment results were presented. Moreover, it was not possible to further stratify epilepsy patients in more subphenotypes than GGE and NAFE. This would have been especially beneficial for NAFE because of the greater heterogeneity between patients. It is

also hard to estimate the required sample size to overcome those limitations since there is no precedent for it in the disease. Therefore, any power analyses would be based on mostly arbitrary assumptions.

Low mappability regions

A limitation of WGS is that some regions have a low mappability. Mappability is the process of assembling sequencing reads by 'mapping' them to a reference genome[169]. When regions are unique it is easy to correctly assemble the reads. However, some genomic regions are composed of repeats, which can be longer than the reads. These regions have a low mappability because if the read is shorter than the repeat it is impossible to map[170]. On the GRCh37 reference genome around 7.6% of the genome has a low mappability. This means that even though WGS was used, some regions were still out of reach of the analyses.

Conclusion

This work's objective was to assess the role of rare non-coding variants in epilepsy. To do so, a deep learning algorithm named ExPecto was used. This algorithm can predict the tissue specific expression change effect of any variant in the genome. By using this information, it was determined that highly deleterious rare non-coding variants are enriched in epilepsy patients compared to controls, specifically in GGE. On the other hand, NAFE patients have no difference compared with controls, which indicates a rather different underlying genetic mechanism between those epilepsy subtypes. The results presented in this study showed, like previous studies, that GGE epilepsy has a stronger genetic burden than NAFE[3,11,31–33]. Even though the sample size limited power and resolution of discoveries, it is clear that rare non-coding variants play a role in the etiology of the disease. In the end, further studies involving WGS with larger sample sizes will be necessary to thoroughly explore the role of non-coding regions with increased statistical power and resolution. The findings of this research provide a basis for investigating these regions in epilepsy, as they could potentially account for a portion of the unexplained heritability.

References

1. Chang BS, Lowenstein DH. Epilepsy. *N Engl J Med*. 2003;349(13):1257-1266. doi:10.1056/NEJMra022308

2. Perucca P, Bahlo M, Berkovic SF. The Genetics of Epilepsy. *Annu Rev Genomics Hum Genet*. 2020;21(1):205-230. doi:10.1146/annurev-genom-120219-074937

3. The International League Against Epilepsy Consortium. Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. *Lancet Neurol*. 2014;13(9):893.

4. Puteikis K, Mameniškienė R. Mortality among People with Epilepsy: A Retrospective Nationwide Analysis from 2016 to 2019. *Int J Environ Res Public Health*. 2021;18(19):10512. doi:10.3390/ijerph181910512

5. Murray CJL, Vos T, Lozano R, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*. 2012;380(9859):2197-2223. doi:10.1016/S0140-6736(12)61689-4

6. Kyu HH, Abate D, Abate KH, et al. Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*. 2018;392(10159):1859-1922. doi:10.1016/S0140-6736(18)32335-3

7. Tomson T. Mortality in epilepsy. *J Neurol*. 2000;247(1):15-21. doi:10.1007/s004150050004

8. Beghi E. The Epidemiology of Epilepsy. *Neuroepidemiology*. 2020;54(2):185-191. doi:10.1159/000503831

9. Fiest KM, Sauro KM, Wiebe S, et al. Prevalence and incidence of epilepsy: A systematic review and meta-analysis of international studies. *Neurology*. 2017;88(3):296-303. doi:10.1212/WNL.0000000000003509

10. Dalic L, Cook MJ. Managing drug-resistant epilepsy: challenges and solutions. *Neuropsychiatr Dis Treat*. 2016;12:2605-2616. doi:10.2147/NDT.S84852

11. International League Against Epilepsy Consortium on Complex Epilepsies, Berkovic SF, Cavalleri GL, Koeleman BP. Genome-wide meta-analysis of over 29,000 people with epilepsy reveals 26 loci and subtype-specific genetic architecture. Published online June 14, 2022:2022.06.08.22276120. doi:10.1101/2022.06.08.22276120

12. Yoo JY, Panov F. Identification and Treatment of Drug-Resistant Epilepsy. *Contin Lifelong Learn Neurol*. 2019;25(2):362. doi:10.1212/CON.0000000000000710

13. Rama S, Zbili M, Debanne D. Signal propagation along the axon. *Curr Opin Neurobiol*. 2018;51:37-44. doi:10.1016/j.conb.2018.02.017

14. Freeman SA, Desmazières A, Fricker D, Lubetzki C, Sol-Foulon N. Mechanisms of sodium channel clustering and its influence on axonal impulse conduction. *Cell Mol Life Sci*. 2016;73(4):723-735. doi:10.1007/s00018-015-2081-1

15. Süudhof TC. Neurotransmitter Release. In: Südhof TC, Starke K, eds. *Pharmacology of Neurotransmitter Release*. Handbook of Experimental Pharmacology. Springer; 2008:1-21. doi:10.1007/978-3-540-74805-2_1

16. Kobayashi R, Nishimaru H, Nishijo H. Estimation of excitatory and inhibitory synaptic conductance variations in motoneurons during locomotor-like rhythmic activity. *Neuroscience*. 2016;335:72-81. doi:10.1016/j.neuroscience.2016.08.027

17. Staley K. Molecular mechanisms of epilepsy. *Nat Neurosci*. 2015;18(3):367-372. doi:10.1038/nn.3947

18. Moran NF, Poole K, Bell G, et al. Epilepsy in the United Kingdom: seizure frequency and severity, anti-epileptic drug utilization and impact on life in 1652 people with epilepsy. *Seizure - Eur J Epilepsy*. 2004;13(6):425-433. doi:10.1016/j.seizure.2003.10.002

19. Fisher RS, Acevedo C, Arzimanoglou A, et al. ILAE Official Report: A practical clinical definition of epilepsy. *Epilepsia*. 2014;55(4):475-482. doi:10.1111/epi.12550

20. Koeleman BPC. What do genetic studies tell us about the heritable basis of common epilepsy? Polygenic or complex epilepsy? *Neurosci Lett*. 2018;667:10-16. doi:10.1016/j.neulet.2017.03.042

21. Speed D, O'Brien TJ, Palotie A, et al. Describing the genetic architecture of epilepsy through heritability analysis. *Brain*. 2014;137(10):2680-2689. doi:10.1093/brain/awu206

22. Berg AT, Berkovic SF, Brodie MJ, et al. Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE Commission on Classification and Terminology, 2005-2009. *Epilepsia*. 2010;51(4):676-685. doi:10.1111/j.1528-1167.2010.02522.x

23. Proposal for revised classification of epilepsies and epileptic syndromes. Commission on Classification and Terminology of the International League Against Epilepsy. *Epilepsia*. 1989;30(4):389-399. doi:10.1111/j.1528-1157.1989.tb05316.x

24. Zawar I, Knight EP. Epilepsy With Eyelid Myoclonia (Jeavons Syndrome). *Pediatr Neurol*. 2021;121:75-80. doi:10.1016/j.pediatrneurol.2020.11.018

25. Strauss KA, Puffenberger EG, Huentelman MJ, et al. Recessive Symptomatic Focal Epilepsy and Mutant Contactin-Associated Protein-like 2. *N Engl J Med*. 2006;354(13):1370-1377. doi:10.1056/NEJMoa052773

26. Combi R, Dalprà L, Tenchini ML, Ferini-Strambi L. Autosomal dominant nocturnalfrontal lobe epilepsy. *J Neurol*. 2004;251(8):923-934. doi:10.1007/s00415-004-0541-x

27. Pagni S, Mills JD, Frankish A, Mudge JM, Sisodiya SM. Non-coding regulatory elements: Potential roles in disease and the case of epilepsy. *Neuropathol Appl Neurobiol*. 2022;48(3):e12775. doi:10.1111/nan.12775

28. Kilpinen H, Waszak SM, Gschwind AR, et al. Coordinated Effects of Sequence Variation on DNA Binding, Chromatin Structure, and Transcription. *Science*. 2013;342(6159):744-747. doi:10.1126/science.1242463

29. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*. 2017;169(7):1177-1186. doi:10.1016/j.cell.2017.05.038

30. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*. 2013;9(1):29. doi:10.1186/1746-4811-9-29

31. Song M, Liu J, Yang Y, Lv L, Li W, Luo XJ. Genome-Wide Meta-Analysis Identifies Two Novel Risk Loci for Epilepsy. *Front Neurosci*. 2021;15:722592. doi:10.3389/fnins.2021.722592

32. Feng YCA, Howrigan DP, Abbott LE, et al. Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals. *Am J Hum Genet*. 2019;105(2):267-282. doi:10.1016/j.ajhg.2019.05.020

33. The International League Against Epilepsy Consortium on Complex Epilepsies, Abou-Khalil B, Auce P, et al. Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. *Nat Commun*. 2018;9(1):5269. doi:10.1038/s41467-018-07524-z

34. Suzuki T, Koike Y, Ashikawa K, et al. Genome-wide association study of epilepsy in a Japanese population identified an associated region at chromosome 12q24. *Epilepsia*. 2021;62(6):1391-1400. doi:10.1111/epi.16911

35. Kasperaviciūte D, Catarino CB, Heinzen EL, et al. Common genetic variation and susceptibility to partial epilepsies: a genome-wide association study. *Brain J Neurol*. 2010;133(Pt 7):2136-2147. doi:10.1093/brain/awq130

36. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580-585. doi:10.1038/ng.2653

37. Dunham I, Kundaje A, Aldred SF, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74. doi:10.1038/nature11247

38. Zhu H, Zhou X. Transcriptome-wide association studies: a view from Mendelian randomization. *Quant Biol*. Published online June 17, 2020. doi:10.1007/s40484-020-0207-4

39. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*. 2016;48(3):245-252. doi:10.1038/ng.3506

40. Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet*. 2016;48(5):481-487. doi:10.1038/ng.3538

41. Mehndiratta MM, Wadhai SA. International Epilepsy Day - A day notified for global public education & awareness. *Indian J Med Res*. 2015;141(2):143-144.

42. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779

43. Ishigaki K, Akiyama M, Kanai M, et al. Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nat Genet*. 2020;52(7):669-679. doi:10.1038/s41588-020-0640-3

44. Kurki MI, Karjalainen J, Palta P, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*. 2023;613(7944):508-518. doi:10.1038/s41586-022-05473-8

45. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-291. doi:10.1038/nature19057

46. Kjeldsen MJ, Corey LA, Christensen K, Friis ML. Epileptic seizures and syndromes in twins: the importance of genetic factors. *Epilepsy Res*. 2003;55(1):137-146. doi:10.1016/S0920-1211(03)00117-7

47. Thomsen SF, Van Der Sluis S, Kyvik KO, Skytthe A, Backer V. Estimates of asthma heritability in a large twin sample. *Clin Exp Allergy*. 2010;40(7):1054-1061. doi:10.1111/j.1365-2222.2010.03525.x

48. Ullemar V, Magnusson PKE, Lundholm C, et al. Heritability and confirmation of genetic association studies for childhood asthma in twins. *Allergy*. 2016;71(2):230-238. doi:10.1111/all.12783

49. Vicente CT, Revez JA, Ferreira MAR. Lessons from ten years of genome-wide association studies of asthma. *Clin Transl Immunol*. 2017;6(12):e165. doi:10.1038/cti.2017.54

50. Kim KW, Ober C. Lessons Learned From GWAS of Asthma. *Allergy Asthma Immunol Res*. 2018;11(2):170-187. doi:10.4168/aair.2019.11.2.170

51. Hecker J, Chun S, Samiei A, et al. FGF20 and PGM2 variants are associated with childhood asthma in family-based whole-genome sequencing studies. *Hum Mol Genet*. 2023;32(4):696-707. doi:10.1093/hmg/ddac258

52. Campbell CD, Mohajeri K, Malig M, et al. Whole-Genome Sequencing of Individuals from a Founder Population Identifies Candidate Genes for Asthma. *PLOS ONE*. 2014;9(8):e104396. doi:10.1371/journal.pone.0104396

53. Mak ACY, White MJ, Eckalbar WL, et al. Whole-Genome Sequencing of Pharmacogenetic Drug Response in Racially Diverse Children with Asthma. *Am J Respir Crit Care Med*. 2018;197(12):1552-1564. doi:10.1164/rccm.201712-2529OC

54. Moreau C, Rébillard RM, Wolking S, et al. Polygenic risk scores of several subtypes of epilepsies in a founder population. *Neurol Genet*. 2020;6(3):e416. doi:10.1212/NXG.0000000000000416

55. Koko M, Krause R, Sander T, et al. Distinct gene-set burden patterns underlie common generalized and focal epilepsies. *eBioMedicine*. 2021;72. doi:10.1016/j.ebiom.2021.103588

56. Motelow JE, Povysil G, Dhindsa RS, et al. Sub-genic intolerance, ClinVar, and the epilepsies: A whole-exome sequencing study of 29,165 individuals. *Am J Hum Genet*. 2021;108(6):965-982. doi:10.1016/j.ajhg.2021.04.009

57. Ellis CA, Petrovski S, Berkovic SF. Epilepsy genetics: clinical impacts and biological insights. *Lancet Neurol*. 2020;19(1):93-100. doi:10.1016/S1474-4422(19)30269-8

58. Palazzo AF, Gregory TR. The Case for Junk DNA. *PLOS Genet*. 2014;10(5):e1004351. doi:10.1371/journal.pgen.1004351

59. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet*. 2018;50(8):1171-1179. doi:10.1038/s41588-018-0160-6

60. Aguet F, Brown AA, Castel SE, et al. Genetic effects on gene expression across human tissues. *Nature*. 2017;550(7675):204-213. doi:10.1038/nature24277

61. De Rubeis S, He X, Goldberg AP, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*. 2014;515(7526):209-215. doi:10.1038/nature13772

62. Turner TN, Coe BP, Dickel DE, et al. Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell*. 2017;171(3):710-722.e12. doi:10.1016/j.cell.2017.08.047

63. Werling DM, Brand H, An JY, et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat Genet*. 2018;50(5):727-736. doi:10.1038/s41588-018-0107-y

64. Brandler WM, Antaki D, Gujral M, et al. Paternally inherited cis-regulatory structural variants are associated with autism. *Science*. 2018;360(6386):327-331. doi:10.1126/science.aan2261

65. Myers CT, Stong N, Mountier EI, et al. De Novo Mutations in PPP3CA Cause Severe Neurodevelopmental Disease with Seizures. *Am J Hum Genet*. 2017;101(4):516-524. doi:10.1016/j.ajhg.2017.08.013

66. Fullard JF, Giambartolomei C, Hauberg ME, et al. Open chromatin profiling of human postmortem brain infers functional roles for non-coding schizophrenia loci. *Hum Mol Genet*. 2017;26(10):1942-1951. doi:10.1093/hmg/ddx103

67. Pérez-Agustín A, Pinsach-Abuin M, Pagans S. Role of Non-Coding Variants in Brugada Syndrome. *Int J Mol Sci*. 2020;21(22):8556. doi:10.3390/ijms21228556

68. Villar D, Frost S, Deloukas P, Tinker A. The contribution of non-coding regulatory elements to cardiovascular disease. *Open Biol*. 2020;10(7):200088. doi:10.1098/rsob.200088

69. van Ouwerkerk AF, Bosada FM, van Duijvenboden K, et al. Identification of atrial fibrillation associated genes and functional non-coding variants. *Nat Commun*. 2019;10(1):4755. doi:10.1038/s41467-019-12721-5

70. Okada Y, Eyre S, Suzuki A, Kochi Y, Yamamoto K. Genetics of rheumatoid arthritis: 2018 status. *Ann Rheum Dis*. 2019;78(4):446-453. doi:10.1136/annrheumdis-2018-213678

71. Wang Q, Martínez-Bonet M, Kim T, et al. Identification of a regulatory pathway governing TRAF1 via an arthritis-associated non-coding variant. Published online December 15, 2022:2022.12.15.520628. doi:10.1101/2022.12.15.520628

72. Kierczak M, Rafati N, Höglund J, et al. Contribution of rare whole-genome sequencing variants to plasma protein levels and the missing heritability. *Nat Commun*. 2022;13(1):2532. doi:10.1038/s41467-022-30208-8

73. Trynka G, Westra HJ, Slowikowski K, et al. Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *Am J Hum Genet*. 2015;97(1):139-152. doi:10.1016/j.ajhg.2015.05.016

74. Lewis DA. The Human Brain Revisited: Opportunities and Challenges in Postmortem Studies of Psychiatric Disorders. *Neuropsychopharmacology*. 2002;26(2):143-154. doi:10.1016/S0893-133X(01)00393-1

75. Avsec Ž, Agarwal V, Visentin D, et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods*. 2021;18(10):1196-1203. doi:10.1038/s41592-021-01252-x

76. Agarwal V, Shendure J. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep*. 2020;31(7). doi:10.1016/j.celrep.2020.107663

77. Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res*. 2018;28(5):739-750. doi:10.1101/gr.227819.117

78. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539

79. Roberts DA, Yaida S, Hanin B. *The Principles of Deep Learning Theory*. Cambridge University Press Cambridge, MA, USA; 2022.

80. Xiao C, Sun J. *Introduction to Deep Learning for Healthcare*. Springer; 2021.

81. Kumar A. Deep Learning Explained Simply in Layman Terms. Data Analytics. Published September 17, 2020. Accessed July 27, 2023. https://vitalflux.com/deep-learning-neural-network-explained-simply-layman-terms/

82. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet*. 2018;50(8):1171-1179. doi:10.1038/s41588-018-0160-6

83. Moreau C, Tremblay F, Wolking S, et al. Assessment of burden and segregation profiles of CNVs in patients with epilepsy. *Ann Clin Transl Neurol*. 2022;9(7):1050-1058. doi:10.1002/acn3.51598

84. Monlong J, Cossette P, Meloche C, Rouleau G, Girard SL, Bourque G. Human copy number variants are enriched in regions of low mappability. *Nucleic Acids Res*. 2018;46(14):7236-7249. doi:10.1093/nar/gky538

85. Monlong J, Girard SL, Meloche C, et al. Global characterization of copy number variants in epilepsy patients from whole genome sequencing. *PLoS Genet*. 2018;14(4):e1007285. doi:10.1371/journal.pgen.1007285

86. Hamdan FF, Myers CT, Cossette P, et al. High Rate of Recurrent De Novo Mutations in Developmental and Epileptic Encephalopathies. *Am J Hum Genet*. 2017;101(5):664-685. doi:10.1016/j.ajhg.2017.09.008

87. Jiang X, Raju PK, D'Avanzo N, et al. Both gain-of-function and loss-of-function de novo CACNA1A mutations cause severe developmental epileptic encephalopathies in the spectrum of Lennox-Gastaut syndrome. *Epilepsia*. 2019;60(9):1881-1894. doi:10.1111/epi.16316

88. Malerba F, Alberini G, Balagura G, et al. Genotype-phenotype correlations in patients with de novo KCNQ2 pathogenic variants. *Neurol Genet*. 2020;6(6). doi:10.1212/NXG.0000000000000528

89. Rydzanicz M, Zwoliński P, Gasperowicz P, et al. A recurrent de novo variant supports KCNC2 involvement in the pathogenesis of developmental and epileptic

encephalopathy. *Am J Med Genet A*. 2021;185(11):3384-3389. doi:10.1002/ajmg.a.62455

90. Gagnon L, Moreau C, Laprise C, Vézina H, Girard SL. Deciphering the genetic structure of the Quebec founder population using genealogies. *Eur J Hum Genet*. Published online April 4, 2023:1-7. doi:10.1038/s41431-023-01356-2

91. Moreau C, Vézina H, Labuda D. [Founder effects and genetic variability in Quebec]. *Med Sci*. 2007;23(11):1008-1013. doi:10.1051/medsci/200723111008

92. van der Auwera G, O'Connor BD. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media, Incorporated; 2020. https://books.google.ca/books?id=wwiCswEACAAJ

93. Lappalainen T, Scott AJ, Brandt M, Hall IM. Genomic Analysis in the Age of Human Genome Sequencing. *Cell*. 2019;177(1):70-84. doi:10.1016/j.cell.2019.02.032

94. Turrin NP, Rivest S. Innate immune reaction in response to seizures: implications for the neuropathology associated with epilepsy. *Neurobiol Dis*. 2004;16(2):321-334. doi:10.1016/j.nbd.2004.03.010

95. Huberfeld G, Blauwblomme T, Miles R. Hippocampus and epilepsy: Findings from human tissues. *Rev Neurol (Paris)*. 2015;171(3):236-251. doi:10.1016/j.neurol.2015.01.563

96. Aroniadou-Anderjaska V, Fritsch B, Qashu F, Braga MFM. Pathology and pathophysiology of the amygdala in epileptogenesis and epilepsy. *Epilepsy Res*. 2008;78(2):102-116. doi:10.1016/j.eplepsyres.2007.11.011

97. Guerrini R. Genetic Malformations of the Cerebral Cortex and Epilepsy. *Epilepsia*. 2005;46(s1):32-37. doi:10.1111/j.0013-9580.2005.461010.x

98. Pires G, Leitner D, Drummond E, et al. Proteomic differences in the hippocampus and cortex of epilepsy brain tissue. *Brain Commun*. 2021;3(2):fcab021. doi:10.1093/braincomms/fcab021

99. Young JC, Vaughan DN, Nasser HM, Jackson GD. Anatomical imaging of the piriform cortex in epilepsy. *Exp Neurol*. 2019;320:113013. doi:10.1016/j.expneurol.2019.113013

100. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet*. 2007;81(3):559-575.

101. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet*. 2010;11(11):773-785. doi:10.1038/nrg2867

102. Saint Pierre A, Génin E. How important are rare variants in common disease? *Brief Funct Genomics*. 2014;13(5):353-361. doi:10.1093/bfgp/elu025

103. Momozawa Y, Mizukami K. Unique roles of rare variants in the genetics of complex diseases in humans. *J Hum Genet*. 2021;66(1):11-23. doi:10.1038/s10038-020-00845-2

104. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7

105. Ye Z, McQuillan L, Poduri A, et al. Somatic mutation: The hidden genetics of brain malformations and focal epilepsies. *Epilepsy Res*. 2019;155:106161. doi:10.1016/j.eplepsyres.2019.106161

106. Lim JS, Kim W il, Kang HC, et al. Brain somatic mutations in MTOR cause focal cortical dysplasia type II leading to intractable epilepsy. *Nat Med*. 2015;21(4):395-400. doi:10.1038/nm.3824

107. Møller RS, Weckhuysen S, Chipaux M, et al. Germline and somatic mutations in the MTOR gene in focal cortical dysplasia and epilepsy. *Neurol Genet*. 2016;2(6). doi:10.1212/NXG.0000000000000118

108. Kim JK, Cho J, Kim SH, et al. Brain somatic mutations in MTOR reveal translational dysregulations underlying intractable focal epilepsy. *J Clin Invest*. 2019;129(10):4207-4223. doi:10.1172/JCI127032

109. Sim NS, Ko A, Kim WK, et al. Precise detection of low-level somatic mutation in resected epilepsy brain tissue. *Acta Neuropathol (Berl)*. 2019;138(6):901-912. doi:10.1007/s00401-019-02052-6

110. Kim S, Baldassari S, Sim NS, et al. Detection of Brain Somatic Mutations in Cerebrospinal Fluid from Refractory Epilepsy Patients. *Ann Neurol*. 2021;89(6):1248-1252. doi:10.1002/ana.26080

111. Sim NS, Seo Y, Lim JS, et al. Brain somatic mutations in SLC35A2 cause intractable epilepsy with aberrant N-glycosylation. *Neurol Genet*. 2018;4(6). doi:10.1212/NXG.0000000000000294

112. Li Z, Li X, Zhou H, et al. A framework for detecting noncoding rare variant associations of large-scale whole-genome sequencing studies. *Nat Methods*. 2022;19(12):1599-1611. doi:10.1038/s41592-022-01640-x

113. Chen H, Huffman JE, Brody JA, et al. Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. *Am J Hum Genet*. 2019;104(2):260-274. doi:10.1016/j.ajhg.2018.12.012

114. Kelly TN, Sun X, He KY, et al. Insights From a Large-Scale Whole-Genome Sequencing Study of Systolic Blood Pressure, Diastolic Blood Pressure, and Hypertension. *Hypertension*. 2022;79(8):1656-1667. doi:10.1161/HYPERTENSIONAHA.122.19324

115. Li X, Quick C, Zhou H, et al. Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nat Genet*. 2023;55(1):154-164. doi:10.1038/s41588-022-01225-6

116. Li Z, Li X, Liu Y, et al. Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole-Genome Sequencing Studies. *Am J Hum Genet*. 2019;104(5):802-814. doi:10.1016/j.ajhg.2019.03.002

117. He Z, Xu B, Buxbaum J, Ionita-Laza I. A genome-wide scan statistic framework for whole-genome sequence data analysis. *Nat Commun*. 2019;10(1):3018. doi:10.1038/s41467-019-11023-0

118. Li X, Li Z, Zhou H, et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet*. 2020;52(9):969-983. doi:10.1038/s41588-020-0676-4

119. Povysil G, Petrovski S, Hostyk J, Aggarwal V, Allen AS, Goldstein DB. Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat Rev Genet*. 2019;20(12):747-759. doi:10.1038/s41576-019-0177-4

120. Yang Y, Basu S, Zhang L. A Bayesian hierarchically structured prior for rare-variant association testing. *Genet Epidemiol*. 2021;45(4):413-424. doi:10.1002/gepi.22379

121. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet*. 2019;104(3):410-421. doi:10.1016/j.ajhg.2019.01.002

122. Zhou W, Zhao Z, Nielsen JB, et al. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat Genet*. 2020;52(6):634-639. doi:10.1038/s41588-020-0621-6

123. Gogarten SM, Sofer T, Chen H, et al. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics*. 2019;35(24):5346-5348. doi:10.1093/bioinformatics/btz567

124. Kumar KR, Cowley MJ, Davis RL. Next-Generation Sequencing and Emerging Technologies. *Semin Thromb Hemost*. 2019;45(7):661-673. doi:10.1055/s-0039-1688446

125. Lou RN, Jacobs A, Wilder AP, Therkildsen NO. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Mol Ecol*. 2021;30(23):5966-5993. doi:10.1111/mec.16077

126. Hess JF, Kohl TA, Kotrová M, et al. Library preparation for next generation sequencing: A review of automation strategies. *Biotechnol Adv*. 2020;41:107537. doi:10.1016/j.biotechadv.2020.107537

127. Rezapour A, Souresrafil A, Barzegar M, Sheikhy-Chaman M, Tatarpour P. Economic evaluation of next-generation sequencing techniques in diagnosis of genetic disorders: A systematic review. *Clin Genet*. 2023;103(5):513-528. doi:10.1111/cge.14313

128. McCombie WR, McPherson JD. Future Promises and Concerns of Ubiquitous Next-Generation Sequencing. *Cold Spring Harb Perspect Med*. 2019;9(9):a025783. doi:10.1101/cshperspect.a025783

129. Mantere T, Kersten S, Hoischen A. Long-Read Sequencing Emerging in Medical Genetics. *Front Genet*. 2019;10. Accessed April 4, 2023. https://www.frontiersin.org/articles/10.3389/fgene.2019.00426

130. Guo J, Cheng T, Xu H, Li Y, Zeng J. An efficient and cost-effective method for primer-induced nucleotide labeling for massive sequencing on next-generation sequencing platforms. *Sci Rep*. 2019;9(1):3125. doi:10.1038/s41598-019-38996-8

131. Levy SE, Boone BE. Next-Generation Sequencing Strategies. *Cold Spring Harb Perspect Med*. 2019;9(7):a025791. doi:10.1101/cshperspect.a025791

132. Gloss BS, Dinger ME. Realizing the significance of noncoding functionality in clinical genomics. *Exp Mol Med*. 2018;50(8):1-8. doi:10.1038/s12276-018-0087-0

133. Bae B, Miura P. Emerging Roles for 3′ UTRs in Neurons. *Int J Mol Sci*. 2020;21(10):3413. doi:10.3390/ijms21103413

134. Pereira-Castro I, Moreira A. On the function and relevance of alternative 3′-UTRs in gene expression regulation. *WIREs RNA*. 2021;12(5):e1653. doi:10.1002/wrna.1653

135. Li Q, Zhang C, Wen J, et al. Transcriptome Analyses Show Changes in Gene Expression Triggered by a 31-bp InDel within OsSUT3 5′UTR in Rice Panicle. *Int J Mol Sci*. 2023;24(13):10640. doi:10.3390/ijms241310640

136. Ellingford JM, Ahn JW, Bagnall RD, et al. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Med*. 2022;14(1):73. doi:10.1186/s13073-022-01073-3

137. Chen Y, Li Z, Chen X, Zhang S. Long non-coding RNAs: From disease code to drug role. *Acta Pharm Sin B*. 2021;11(2):340-354. doi:10.1016/j.apsb.2020.10.001

138. Trifu SC, Kohn B, Vlasie A, Patrichi BE. Genetics of schizophrenia (Review). *Exp Ther Med*. 2020;20(4):3462-3468. doi:10.3892/etm.2020.8973

139. Bertram L. Next Generation Sequencing in Alzheimer's Disease. In: Castrillo JI, Oliver SG, eds. *Systems Biology of Alzheimer's Disease*. Methods in Molecular Biology. Springer; 2016:281-297. doi:10.1007/978-1-4939-2627-5_17

140. Fareed M, Chauhan W, Fatma R, Din I, Afzal M, Ahmed Z. Next-generation sequencing technologies in diabetes research. *Diabetes Epidemiol Manag*. 2022;7:100097. doi:10.1016/j.deman.2022.100097

141. Plooster M, Brennwald P, Gupton SL. Endosomal trafficking in schizophrenia. *Curr Opin Neurobiol*. 2022;74:102539. doi:10.1016/j.conb.2022.102539

142. Andrews SJ, Renton AE, Fulton-Howard B, Podlesny-Drabiniok A, Marcora E, Goate AM. The complex genetic architecture of Alzheimer's disease: novel insights and future directions. *eBioMedicine*. 2023;90. doi:10.1016/j.ebiom.2023.104511

143. Pang H, Lin J, Luo S, et al. The missing heritability in type 1 diabetes. *Diabetes Obes Metab*. 2022;24(10):1901-1911. doi:10.1111/dom.14777

144. Rees E, Kirov G, O'Donovan MC, Owen MJ. De Novo Mutation in Schizophrenia. *Schizophr Bull*. 2012;38(3):377-381. doi:10.1093/schbul/sbs047

145. Fromer M, Pocklington AJ, Kavanagh DH, et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature*. 2014;506(7487):179-184. doi:10.1038/nature12929

146. Alkelai A, Greenbaum L, Docherty AR, et al. The benefit of diagnostic whole genome sequencing in schizophrenia and other psychotic disorders. *Mol Psychiatry*. 2022;27(3):1435-1447. doi:10.1038/s41380-021-01383-9

147. Zhao G, Li K, Li B, et al. Gene4Denovo: an integrated database and analytic platform for de novo mutations in humans. *Nucleic Acids Res*. 2020;48(D1):D913-D926. doi:10.1093/nar/gkz923

148. Nicolas G, Veltman JA. The role of de novo mutations in adult-onset neurodegenerative disorders. *Acta Neuropathol (Berl)*. 2019;137(2):183-207. doi:10.1007/s00401-018-1939-3

149. Al-Khawaga S, Mohammed I, Saraswathi S, et al. The clinical and genetic characteristics of permanent neonatal diabetes (PNDM) in the state of Qatar. *Mol Genet Genomic Med*. 2019;7(10):e00753. doi:10.1002/mgg3.753

150. Bergen SE, O'Dushlaine CT, Ripke S, et al. Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. *Mol Psychiatry*. 2012;17(9):880-886. doi:10.1038/mp.2012.73

151. Kirov G, Pocklington AJ, Holmans P, et al. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol Psychiatry*. 2012;17(2):142-153. doi:10.1038/mp.2011.154

152. Tansey KE, Rees E, Linden DE, et al. Common alleles contribute to schizophrenia in CNV carriers. *Mol Psychiatry*. 2016;21(8):1085-1089. doi:10.1038/mp.2015.143

153. Lee WP, Conery M, Valladares O, et al. Copy number variation (CNV) identification and association study on 3,928 Alzheimer's disease whole genome sequencing data from the Alzheimer's Disease Sequencing Project (ADSP). *Alzheimers Dement*. 2021;17(S3):e052721. doi:10.1002/alz.052721

154. Zhang B. Integrative analysis identifies copy number variations and their controlled causal molecular networks in Alzheimer's disease. *Alzheimers Dement*. 2020;16(S3):e043341. doi:10.1002/alz.043341

155. Grayson BL, Smith ME, Thomas JW, et al. Genome-Wide Analysis of Copy Number Variation in Type 1 Diabetes. *PLOS ONE*. 2010;5(11):e15393. doi:10.1371/journal.pone.0015393

156. Prabhanjan M, Suresh RV, Murthy MN, Ramachandra NB. Type 2 diabetes mellitus disease risk genes identified by genome wide copy number variation scan in normal populations. *Diabetes Res Clin Pract*. 2016;113:160-170. doi:10.1016/j.diabres.2015.12.015

157. Berberich AJ, Huot C, Cao H, et al. Copy Number Variation in GCK in Patients With Maturity-Onset Diabetes of the Young. *J Clin Endocrinol Metab*. 2019;104(8):3428-3436. doi:10.1210/jc.2018-02574

158. Gibbons A, Udawela M, Dean B. Non-Coding RNA as Novel Players in the Pathophysiology of Schizophrenia. *Non-Coding RNA*. 2018;4(2):11. doi:10.3390/ncrna4020011

159. Merelo V, Durand D, Lescallette AR, et al. Associating schizophrenia, long non-coding RNAs and neurostructural dynamics. *Front Mol Neurosci*. 2015;8. Accessed April 5, 2023. https://www.frontiersin.org/articles/10.3389/fnmol.2015.00057

160. Chen L, Guo X, Li Z, He Y. Relationship between long non-coding RNAs and Alzheimer's disease: a systematic review. *Pathol - Res Pract*. 2019;215(1):12-20. doi:10.1016/j.prp.2018.11.012

161. Wang E, Lemos Duarte M, Rothman LE, Cai D, Zhang B. Non-coding RNAs in Alzheimer's disease: perspectives from omics studies. *Hum Mol Genet*. 2022;31(R1):R54-R61. doi:10.1093/hmg/ddac202

162. Sun X, Wong D. Long non-coding RNA-mediated regulation of glucose homeostasis and diabetes. *Am J Cardiovasc Dis*. 2016;6(2):17.

163. An T, Zhang J, Ma Y, et al. Relationships of Non-coding RNA with diabetes and depression. *Sci Rep*. 2019;9(1):10707. doi:10.1038/s41598-019-47077-9

164. Matthews LJ, Turkheimer E. Three legs of the missing heritability problem. *Stud Hist Philos Sci*. 2022;93:183-191. doi:10.1016/j.shpsa.2022.04.004

165. van Calker D, Serchov T. The "missing heritability"—Problem in psychiatry: Is the interaction of genetics, epigenetics and transposable elements a potential solution? *Neurosci Biobehav Rev*. 2021;126:23-42. doi:10.1016/j.neubiorev.2021.03.019

166. Naudé W. Artificial intelligence vs COVID-19: limitations, constraints and pitfalls. *AI Soc*. 2020;35(3):761-765. doi:10.1007/s00146-020-00978-0

167. Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat Rev Genet*. 2023;24(2):125-137. doi:10.1038/s41576-022-00532-2

168. Helbig I, Lowenstein DH. Genetics of the epilepsies: where are we and where are we going? *Curr Opin Neurol*. 2013;26(2):179-185. doi:10.1097/WCO.0b013e32835ee6ff

169. Li W, Freudenberg J. Mappability and read length. *Front Genet*. 2014;5:381. doi:10.3389/fgene.2014.00381

170. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep*. 2019;9(1):9354. doi:10.1038/s41598-019-45839-z

Appendix

This project has been subjected to an ethical certification. The reference number for the certificate is 2020-402.