Association of ultra-rare genetic variants with epilepsy

Dissertation

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

der Mathematisch-Naturwissenschaftlichen Fakultät
und
der Medizinischen Fakultät
der Eberhard-Karls-Universität Tübingen

vorgelegt von

Mahmoud Eltayeb Koko Musa
aus Portsudan, Sudan

2023

Tag der mündlichen Prüfung:	10. Januar 2023
Dekan der Math.-Nat. Fakultät:	Prof. Dr. Thilo Stehle
Dekan der Medizinischen Fakultät:	Prof. Dr. Bernd Pichler
1. Berichterstatter:	Prof. Dr. Holger Lerche
2. Berichterstatter:	Prof. Dr. Olaf Rieß
3. Berichterstatter:	Prof. Dr. Anna-Elina Lehesjoki
Prüfungskommission:	Prof. Dr. Holger Lerche
	Prof. Dr. Olaf Rieß
	Prof. Dr. Bernd Antkowiak
	Prof. Dr. Stephan Ossowski

Erklärung / Declaration:

Ich erkläre, dass ich die zur Promotion eingereichte Arbeit mit dem Titel: *"Association of ultra-rare genetic variants with epilepsy"* selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

I hereby declare that I have produced the work entitled: *"Association of ultra-rare genetic variants with epilepsy"*, submitted for the award of a doctorate, on my own (without external help), have used only the sources and aids indicated and have marked passages included from other works, whether verbatim or in content, as such. I swear upon oath that these statements are true and that I have not concealed anything. I am aware that making a false declaration under oath is punishable by a term of imprisonment of up to three years or by a fine.

Tübingen, den ....

Datum / Date Unterschrift /Signature

Summary

Epilepsy represents a wide spectrum of phenotypes with various etiologies and comorbidities. Genetic predisposition to epilepsy is conferred by rare variants and common risk alleles. Ultra-rare variants (URVs) – those not seen in healthy population controls – are thought to underlie a substantial part of the risk mediated by coding variants. In this dissertation, the role of URVs was studied in several cohorts of individuals with common epilepsy syndromes, aiming to identify new genetic etiologies underlying epileptogenesis.

Multiple approaches based on whole exome sequencing were utilized, scaling from studies of single families to populations and from genes to gene sets. First, five closely consanguineous Sudanese families, in which multiple siblings (whose parents are cousins) were diagnosed with a genetic epilepsy, were examined to touch upon the role of rare bi-allelic coding variation in familial epilepsies. There was no evidence to support a key role for recessive inheritance in less severe epilepsies. However, the results expanded the phenotypic spectrum of biallelic ultra-rare *PRRT2* variants, previously linked to movement disorders, to include mild self-limited epilepsy.

Second, sequencing data from individuals diagnosed with genetic generalized epilepsy (GGE; n = 1,928 cases vs. 8,578 ancestry-matched controls of European descent) were analyzed using gene and gene set collapsing approaches to identify key URV associations. Separate analyses of familial GGE (n = 945 cases vs. 8,626 controls) or sporadic GGE (n = 1,005 cases vs. 8,621 controls) were also performed. URVs in *GABRG2* showed an association with familial GGE (approaching study-wide significance) but not with sporadic GGE. Additionally, a higher enrichment of URVs affecting genes encoding $GABA_A$ receptors and GABAergic pathway genes was seen in familial vs. sporadic GGE.

Third, the burden of URVs in a comprehensive range of gene sets was studied in the exomes of individuals diagnosed with GGE (n = 3,064), non-acquired focal epilepsy (NAFE; n = 3,522) or developmental and epileptic encephalopathy (DEE; n = 1,003), compared to 3,962 ancestry-matched controls. In GGE, the burden of URVs in constrained genic regions – those devoid of variations in the general population – was higher in gene sets important for inhibitory signaling vs. in gene sets representative of excitatory signaling. Conversely, there was a relatively higher burden in excitatory vs. inhibitory gene sets in NAFE.

In summary, this dissertation presents novel findings pertaining to the role of ultra-rare coding variation in epileptic disorders, providing new insights into the spectrum of key genes and gene sets related to epileptogenesis.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACMG	American College of Medical Genetics and Genomics
AD	Allele Depth
AMP	Association for Molecular Pathology
ASM	Anti-Seizure Medication
ATAV	Analysis Tool for Annotated Variants platform
ATVB	Italian Atherosclerosis, Thrombosis, and Vascular Biology study
BH-FDR	Benjamini-Hochberg False Discovery Rate
CAE	Childhood Absence Epilepsy
CBZ	Carbamazepine
CCR	Consensus Coding Regions score
CDS	Coding Sequence
CENet	Canadian Epilepsy Network
CI	Confidence Interval
ClinGen	Clinical Genome Resource
CNV	Copy Number Variant
DEE	Developmental and Epileptic Encephalopathy
DP	Depth (of a variant call)
DRAGEN	Dynamic Read Analysis for GENomics platform
EEG	Electroencephalography
EGMA	Epilepsy with Grand Mal seizures on Awakening
EGTCS	Epilepsy with Generalized Tonic-Clonic Seizures alone
EOAE	Early Onset Absence Epilepsy
FAME	Familial Adult Myoclonic Epilepsy
FDR	False Discovery Rate
FET	Fisher's Exact Test
FS	Fisher's Strand bias score (of a variant call)
GA4GH	Global Alliance for Genomics and Health
GATK	Genome Analysis Toolkit
GEFS+	Generalized Epilepsy with Febrile Seizures Plus
GO	Gene Ontology
GQ	Genotype Quality (of a variant call)
GRCh	Genome Reference Consortium Human genome build
GTCS	Generalized Tonic-Clonic Seizure(s)
gVCF	File in Genome Variant Call Format
GWAS	Genome-wide Association Study
HGNC	HUGO [Human Genome Organization] Gene Nomenclature Committee
Hom-Het	Homozygous-heterozygous variants ratio
HPO	Human Phenotype Ontology
ID	Intellectual Disability
IGE	Idiopathic Generalized Epilepsy
IGM	Institute for Genomic Medicine
ILAE	International League Against Epilepsy
INDEL	Insertion-Deletion variant
Ins-Del	Insertions-Deletions ratio
JAE	Juvenile Absence Epilepsy
JME	Juvenile Myoclonic Epilepsy

KEGG	Kyoto Encyclopedia of Genes and Genomes
KING	Kinship-based Inference for GWAS software
LCSB	Luxembourg Centre for Systems Biology
MAC	Minor Allele Count
MAF	Minor Allele Frequency
MCD	Malformation(s) of Cortical Development
MDS	Multidimensional Scaling
MIGen	Myocardial Infarction Genetics consortium
MIM	Mendelian Inheritance in Man catalog of human genetic disorders
MPC	Missense-badness Polyphen and Constraint score
MQ	Mapping Quality (of a variant call)
MQRS	Mapping Quality Rank Sum score (of a variant call)
MRI	Magnetic Resonance Imaging
MTR	Missense Tolerance Ratio score
NAFE	Non-Acquired Focal Epilepsy
NDD	Neurodevelopmental Disorder
NFE	Non-Finnish European population
OMIM	Online Mendelian Inheritance in Man database
OR	Odds Ratio
PB	Phenobarbitone
PCA	Principal Component Analysis
pLoF	predicted Loss-of-Function
PME	Progressive Myoclonic Epilepsy
PPh2	Polyphen-2 Diversity based score
PTV	Protein Truncating Variant
Q/D	Quality/Depth (of a variant call)
QQ	Quantile-Quantile (plot)
QUAL	Quality (of a variant call)
QV	Qualifying Variant
REVEL	Rare Exome Variant Ensemble Learner score
RPRS	Read Position Rank Sum score (of a variant call)
RVAS	Rare Variant Association Study
SNV	Single Nucleotide Variant
SOR	Strand Odds Ratio (of a variant call)
S-SCZ	Swedish Schizophrenia study
SVM	Support Vector Machine
TiTv	Transition-Transversion ratio
TLE	Temporal Lobe Epilepsy
URV	Ultra-rare Variant
VCF	File in Variant Call Format
VEP	Variant Effect Predictor tool
VPA	Valproate
VQS	Variant Quality Score
VQSLOD	Variant Quality Score Log Odds
VQSR	Variant Quality Score Recalibration
VUS	Variant of Uncertain Significance
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

Chapter 1: Introduction

1.1 Epilepsy

1.1.1 The disease and its burden

Epilepsy is a common brain disease hallmarked by a predisposition to recurrent seizures.[1] A seizure is an intermittent abnormality of the central nervous system physiology characterized by a transient occurrence of an abnormal, excessive or synchronous neuronal activity in the brain.[1] Seizures may manifest as visible alterations (e.g., in muscle tone, movement, or behavior), as a sensation that only the affected individual perceives or as recorded changes in brain activity on electroencephalography (EEG).[2] Although seizures may also be non-epileptic,[3] there is a considerable overlap between the use of the two terms *epilepsy* and *seizures* in the common language as well as in professional medical terminology; seizure(s), epileptic seizure, and epilepsy share a common identifier (HP:0001250) in the Human Phenotype Ontology (HPO), a standardized vocabulary of phenotypic abnormalities encountered in human disease.[4] Epilepsy shows a wide phenotypic and genetic variability.[5–7] The disease varies widely in the age at onset, severity and comorbidity.[8,9] Developmental, behavioral and psychiatric comorbidities (e.g., autism spectrum disorder, intellectual disability) are particularly common.[10,11] The incidence of epilepsy in younger children (below 3 years of age) is estimated around 2.4 per 1,000 children (95% Confidence Interval (CI): 2.2–2.6).[12] The prevalence shows a bimodal distribution that peaks in children below ten years and adults above 80 years.[13] The disease has an estimated lifetime prevalence exceeding 5 in 1000 individuals, with studies from different populations providing estimates around 6.2 per 1000 persons (95% CI: 5.4–7.4) and 7.6 per 1,000 persons (95% CI: 6.2–9.4).[13,14]

1.1.2 Diagnosis and treatment

The diagnosis of epilepsy is largely clinical.[1] The International League Against Epilepsy (ILAE) defines epilepsy, in the scope of clinical practice, as the presence of any of these three conditions:[1] (I) At least two unprovoked or reflex seizures occurring more than 24 hours apart; (II) One unprovoked or reflex seizure and a probability of further seizures with a recurrence risk (i.e., risk of seizures occurring again over the next 10 years) similar to the general recurrence risk after two unprovoked seizures; (III) A diagnosis of an epilepsy syndrome (see below for *epilepsy syndromes*). EEG is the standard investigation to document, classify and monitor seizure activity.[2] Brain imaging using Magnetic Resonance Imaging (MRI) is particularly useful to identify focal lesions and other structural abnormalities that may

cause the disease.[15] The frequency of epileptic seizures can range from a few occurrences per a lifetime to frequent and uncontrollable daily seizures, and temporal fluctuations in seizure frequency and clustering of episodes are well known.[16] Pharmacotherapy is the main treatment option, typically tailored to the phenotypes of affected individuals. Since these medications (known as Anti-convulsant or Anti-Epileptic Drugs) primarily confer symptomatic control, they are increasingly referred to as Anti-Seizure Drugs or Anti-Seizure Medications (ASMs). These medications achieve symptomatic control in two thirds of individuals with epilepsy (*pharmacoresponsive*).[17,18] In the remaining third, epilepsy does not respond to the currently available ASMs (*drug-resistant*).[19] Surgical approaches, and other novel approaches targeting mechanisms, offer additional therapeutic hope.[20,21] The disease may resolve spontaneously (*self-limited*), even without treatment[1]; epilepsy is *resolved* for individuals who have remained seizure-free for ten consecutive years (off ASM for at least five consecutive years) or who had an age-dependent epilepsy syndrome (see the examples of *epilepsy syndromes* below) but are now past the risk age.[1]

1.1.3 Classification

The most recent approach to epilepsy classification by the ILAE adopts three diagnostic levels: *seizures*, *epilepsies*, and *epileptic syndromes*.[22,23] Additionally, the classification accommodates the *etiology* (structural, genetic, infectious, immune, metabolic, a mixture of many, or unknown) and *comorbidities* (e.g., autism spectrum disorder or intellectual disability) in all three levels. Once the seizures in an individual are identified as epileptic, the entry classification level (*seizure type*) includes a description of the seizure onset (focal onset, generalized onset, or unknown onset). Further classification based on the nature of seizures (motor vs. non-motor), evolution (focal to bilateral), awareness (impaired awareness or aware) is made when possible. Several seizure types may co-exist in the same individual.[24] If sufficient information is available, a second level of classification (*epilepsy type*) can be attempted, in which the epilepsy is described as focal epilepsy (focal onset seizure types), generalized epilepsy (generalized onset seizure types), or combined generalized and focal epilepsy (co-existence of both types).[24] Otherwise, the epilepsy type is unknown. These two levels of the diagnostic and classification framework differ slightly for neonates (e.g., to highlight the key role of EEG, the predominance of focal onset and the preponderance of certain etiologies).[23] A third diagnostic level is to describe an *epilepsy syndrome*. Epilepsy syndromes are clinical entities that show a group of features usually occurring together (types of seizures commonly

seen, age when seizures commonly begin, part of the brain involved, usual course, genetic etiology, or other features).[22]

Multiple syndromes are well recognized in clinical practice.[25] These reflect the clinical course, etiology, EEG findings, and the treatment outcomes, thus helping in laying out management and counselling plans.[25] Common epilepsy syndromes encompass generalized onset and focal onset epilepsies previously referred to as "idiopathic." Since indicators of a genetic etiology are increasingly identified in these categories, they are currently considered *genetic epilepsies*. Together, these Genetic Generalized Epilepsies (GGEs) and Non-acquired Focal Epilepsies (NAFEs) comprise the overwhelming majority of the epilepsy diagnoses (more than two thirds).[26,27] Self-limited (previously: benign) neonatal and infantile epilepsies, Rolandic epilepsy (Childhood Epilepsy with Centro-Temporal Spikes) and occipital epilepsies are common focal onset epilepsy syndromes, whereas common generalized epilepsy syndromes include Childhood Absence Epilepsy (CAE), Juvenile Myoclonic Epilepsy (JME), Juvenile Absence Epilepsy (JAE), and Epilepsy with Generalized Tonic-Clonic Seizures Alone (EGTCS) (also known as Epilepsy with Grand Mal seizures on Awakening (EGMA)).[25] Several syndromes may constitute a phenotypic continuum with various etiologies (e.g., absence seizures in Early Onset Absence Epilepsy (EOAE), CAE and JAE or focal onset seizures in self-limited neonatal, neonatal-infantile and infantile epilepsies).[28,29]

Epilepsies are also encountered as severe, rare, disorders associated with cognitive impairment and possibly the presence of an encephalopathy; a diagnosis of a Developmental and Epileptic Encephalopathy (DEE) indicates that developmental abnormalities and/or a continuous and usually severe seizure activity had resulted in an encephalopathy and cognitive decline.[30,31] There is also a large spectrum of developmental aberrations (e.g., developmental delay, regression or plateauing) and early onset neurodegenerative changes that may or may not be accompanied with epilepsy. These are usually referred to as Neurodevelopmental Disorders (NDDs) with Epilepsy.[32] Different severe and less severe epilepsies may constitute phenotypic spectra with a shared etiology but variable severity (e.g., Dravet syndrome and Generalized Epilepsy with Febrile Seizures Plus (GEFS+)).[33,34] The general approach to epilepsy diagnosis and classification is outlined in Figure 1.1.

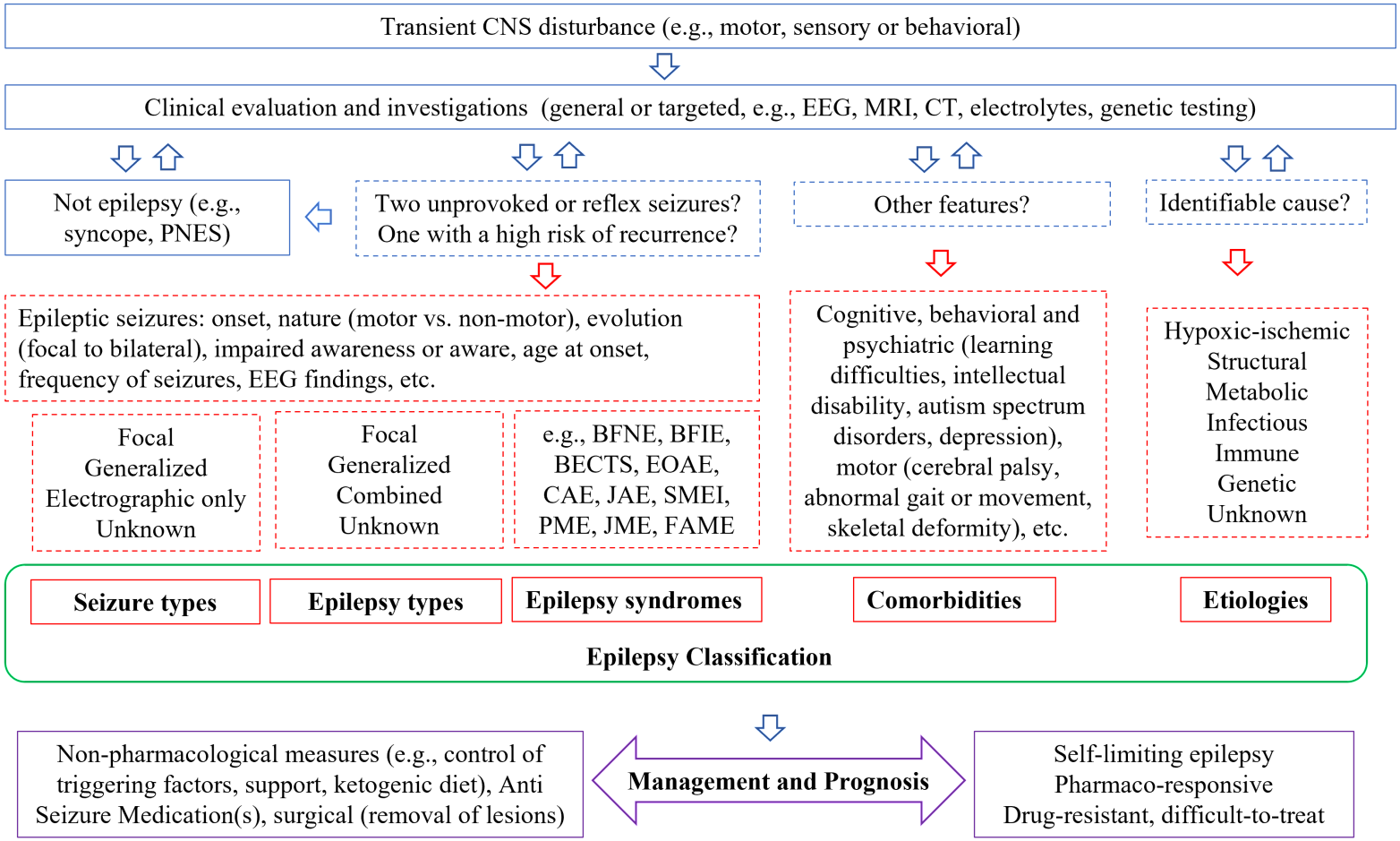


Figure 1.1: Approach to epilepsy diagnosis, classification and management. The classification reflects the ILAE frameworks (see sections 1.1.2 and 1.1.3). BECTS: Self limiting (formerly Benign) Epilepsy with Centro-Temporal Spikes. BFNE/BFIE: Self limiting (formerly Benign) Familial Neonatal/Infantile Epilepsy. CAE: Childhood Absence Epilepsy. EOAE: Early Onset Absence Epilepsy. FAME: Familial Adult Myoclonic Epilepsy. JAE: Juvenile Absence Epilepsy. JME: Juvenile Myoclonic Epilepsy. PME: Progressive Myoclonic Epilepsy. PNES: Psychogenic non-epileptic seizures. SMEI: Severe Myoclonic Epilepsy of Infancy.

1.2 Genetic epilepsies

To label epilepsy as genetic, the underlying genetic cause does not need to be identified; instead, the clinician making the diagnosis is rather expected to evaluate the overall evidence for all etiologies.[22] Certain clues could point towards a genetic etiology, like the presence of a family history or a diagnosis of a certain rare epilepsy syndrome known to be genetic. Evidence from twin studies and large-scale population studies for the common, genetically complex, types of epilepsies justifies a diagnosis of a genetic epilepsy in other individuals diagnosed with the same type of epilepsy.[35,36] More than one etiology could be present according to the ILAE framework.[22,23] Additional analyses are typically sought to determine the inheritance pattern and the mode of genetic causality.[35] Hereafter, a few basic concepts that lay a foundation to understanding genetic epilepsies will be presented followed by a brief summary of the current understanding of epilepsy risk genes and risk variants.

1.2.1 Predisposition in individuals, families, and populations

Although there are no clear boundaries between genetic studies of individuals, families, and populations, these constitute three major targets for genetic workup aimed at identifying new determinants (*discovery*[37,38]) or recognizing known risk determinants (*testing*[39]). Identifying risk genes in individuals is typically synonymous with genetic testing rather than genetic discovery.[40] For discovery, it is customary to study a group of individuals or families with similar presentations; Single multi-generational families can be studied of their own accord, especially in the context of genetic epidemiology.[38] Familial epilepsies, which constitute about 10% – 40% of the overall disease prevalence depending on the type of epilepsy, show a considerable genetic and phenotypic heterogeneity that remains to be explained,[35,41–43] necessitating large scale analyses in populations. Novel risk determinants can be suspected in a single individual or a single family (e.g., based on existing frameworks for variant prioritization[44]) but would require validation and replication in other unrelated individuals or families. Understanding the risk profiles in the population (of known and novel genes) is typically achieved using association studies.[33,45] Population studies can include individuals with or without a family history; these usually include unrelated subjects but recent analytic advances now allow the inclusion of related individuals.[46]

Studies of individuals, families and populations can be integrated. Selected individuals or families from larger cohorts (e.g., carrying variants in candidate genes) can be studied further

(e.g., using segregation or functional analysis).[47] When investigating a single individual, a *priori* knowledge of risk determinants usually exists, and the aim is typically to reach a diagnosis/stratify the risk in this individual.[39,48] However, it is likely that individuals are investigated along with other family members since the evaluation of variant pathogenicity relies partially on their inheritance profiles.[49] Individuals with variants in known disease genes can be included in *post hoc* analyses of large cohorts (e.g., to describe genotype-phenotype correlations or to study individuals with a relatively rare diagnosis).[50,51] Similarly, individuals without a genetic diagnosis can be included in large cohorts to identify new genes.[40]

It is also possible to integrate results from multiple populations to improve the chances of novel gene discovery, particularly for of rare epilepsies; in addition to an increase in the sample size, this offers the chance to examine phenotypic spectrums of known genes as well as the effects of modifiers of monogenic disorders.[52,53] Although genome-wide association studies (GWAS) – investigating common variant – tend to show population specific profiles (in terms of individual risk variants/alleles), there is no strong evidence to suggest notable differences between populations in the profile of genes predisposing to epilepsy (i.e., population-specific risk genes). For instance, it was possible to replicate GWAS loci from large-scale analysis of European samples in studies targeting non-European populations.[54] Similarly, single gene defects (caused by rare rather than common variants) are not typically population specific; this has become evident specially with the implementation of international networks for data sharing like the Matchmaker Exchange.[55] Few studies suggested the presence of population-specific genetic associations, however, without validation.[56,57]

1.2.2 Models of genetic predisposition in epilepsy

Genetic predisposition in epilepsy varies widely between two ends of a spectrum: monogenic epilepsy syndromes (single gene disorders) and epilepsies with complex inheritance (likely polygenic disorders).[58–60] Monogenic, oligogenic and polygenic causality/predisposition reflect the number of genes presumed to be necessary to cause the genetic disease.[61,62] The use of the terms *Mendelian*, *non-Mendelian*, and *complex* to describe the inheritance of an epilepsy phenotype is ideally based on the inheritance patterns in a pedigree (or several pedigrees).[61,62] Monogenic epilepsies can be extremely rare or relatively common, inherited (Mendelian) or sporadic (e.g., with *de novo* variants), mild or severe diseases. These are increasingly named after the causative/predisposing gene (e.g., *SCN1A*-related epilepsy, *GABRG2*-related epilepsy, *PRRT2*-related epilepsy, *PCDH19*-related

epilepsy). Polygenic inheritance is assumed to explain the majority of cases with less severe presentations (and typically with a sporadic nature or complex inheritance), possibly with an additional contribution from environmental modifiers (although this contribution has not been validated).[61–63]

There is also a distinction to be made between two ends of a spectrum of disease-related variants: variants with high effect size (that are usually coding and ultra-rare) and variants with low effect (typically non-coding and common variants). High effect variants are typically hypothesized to cause Mendelian inheritance (or in a broader sense, monogenic phenotypes, as the disease can be sporadic with *de novo* high effect variants), whereas low to moderate effect variants are thought to cause various degrees of genetic predisposition;[64,65] These can underlie Mendelian inheritance with low penetrance or complex inheritance (e.g., familial clustering without a clear inheritance pattern), or predispose to non-Mendelian sporadic phenotypes. Terms like risk variants/alleles are usually used to indicate a variant that does not associate with a clear inheritance pattern but is thought to play a major role in predisposition based on statistical or functional evidence.[66] Notations like "pathogenic" or "causative" entail that a variant has an effect that is high enough to explain why an individual would show the phenotype and are thus used mostly with monogenic phenotypes.[39]

De novo variants are typically seen in sporadic DEE cases whereas recessive inheritance (10 – 15% of total cases with a genetic diagnosis) is a likely possibility when several siblings (with healthy parents) are affected or in those individuals with a background of parental consanguinity.[32,67] Syndromes with mild to moderate phenotypes may be single gene disorders as well. A few NAFE syndromes have an identifiable monogenic cause (e.g., Self-limited (Benign) Familial Infantile Epilepsy,[68] Autosomal Dominant Nocturnal Frontal Lobe Epilepsy).[59] Although the underlying inheritance models in most GGE syndromes are complex and not completely understood,[35] a few cases carry variants in dominant epilepsy genes.[47,51] It is noteworthy that less severe monogenic epilepsies are typically part of a disease spectrum that includes DEE or NDD. Different mechanisms may underlie severe vs. mild presentations.[50,69] Otherwise, heterozygous variants with similar functional defects may cause a spectrum of phenotypes, sometimes seen with phenotypic heterogeneity (i.e., subjects carrying the same variant showing a range of disease severities); [65,69–71] cases with bi-allelic variants in dominant genes are typically associated with more severe presentations.[69,72,73]

1.2.3 Performing a genetic workup in epilepsy

The yield of genetic workup in individuals with an epilepsy diagnosis depends largely on the severity of the epilepsy phenotype as well as the nature of applied methods.[74,75] A genetic cause can be identified in around a third to a half of individuals with NDD with Epilepsy or DEE.[76] Though up to 80% of GGEs (a lower percentage of NAFEs) are thought to have a major genetic component,[77] variants with high effects have been identified in less than 10% of affected individuals.[78] Most clinically relevant variants are short coding alterations.[79] Other types of genetic variants explain a very small fraction of the cases remaining unresolved (i.e., without pathogenic and likely pathogenic short coding variants), e.g., copy number alterations, intronic variants and repeat expansions.[80–83] Copy Number Variants (CNVs) are a recognized cause of monogenic DEEs (5–10% of resolved cases),[73,84] whereas up to 3% of patients with common epilepsies carry epilepsy-associated CNVs.[85] Intronic variants affecting alternative splicing are a rare cause of developmental disorders and epilepsy.[80] Repeat expansions, traditionally associated with Progressive Myoclonic Epilepsy (PME), have been recently implicated in dominantly inherited, non-progressive, Familial Adult Myoclonic Epilepsy (FAME).[81–83]

Various methodologies and approaches can be adopted to identify genes and variants with high effect. The methods of genetic studies have evolved hand in hand with the methods of genotyping/sequencing and the approach (e.g., choice of tests, analysis strategy) for gene identification in a certain epilepsy phenotype is different for discovery or diagnostic purposes. Dominant epilepsy syndromes have been a natural target for linkage studies and family-based re-sequencing studies.[72] Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) are now increasingly used as a first-tier investigation both in individuals and large-scale cohorts, and have also replaced Sanger sequencing and positional cloning in investigating known linkage loci to identify the causative genes.[75,86–89] Recent advances in sequencing technologies led to the resolution of few robust linkage results that remained without identifiable genes for a long time. For instance, long-read sequencing targeting linkage loci in which short coding variants were not originally identified have uncovered underlying intronic repeat expansions in FAME.[81–83]

In severe or rare DEEs/NDDs, sequencing of trios (affected proband and parents) or quartets (affected sibs and parents) is the preferable approach for both genetic testing and discovery.[32,90] Apart from discovery/research, genetic evaluation in common/less severe epilepsies is limited and usually targets early onset phenotypes that are either difficult to treat,

that have additional comorbidities, or that show familial clustering.[86] Less severe epilepsies that are thought to show Mendelian(-like) inheritance in large pedigrees are investigated using family-based sequencing studies, where several individuals are studied using WES/WGS (with or without linkage analysis) followed by segregation analysis and possibly functional validation.[72] The interpretation of segregation results can be challenging when asymptomatic carriers are observed; although possible,[91] statistical evaluation/quantification of disease association based on segregation analysis in single or few families is not typically performed. Further supportive evidence is obtained through the identification of multiple unrelated individuals or families with similar genetic alterations (e.g., the identification of pathogenic variants in $GABA_A$ receptor subunit encoding genes in several families with different types of epilepsy).[71]

Large-scale analyses are becoming increasingly necessary, since susceptibility genes yet to be discovered are likely quite rare or cause a moderate increase in disease risk. Promising but rare candidate genes in which only limited cases are found (e.g., genes coding for synaptic proteins, ion channels, neurotransmitter receptors/transporters or proteins directly interacting with products of known epilepsy genes) can be screened in large cohorts through patient repositories and collaborative networks.[55] Reverse phenotyping has been successfully employed to expand disease-gene association, particularly in genes originally defined in a rare/severe monogenic epilepsy and later implicated in susceptibility to mild and intermediate epilepsy phenotypes.[50,69] Beside these "guided" approaches, association designs have been established as hypothesis-free approaches (i.e., exome-wide or genome-wide) both for common and rare variants. [45,47,54,60,92–94] A few genome-wide significant risk loci were identified at a population scale using GWAS (e.g., *SCN1A*).[45,60,92] Rare Variant Association Studies (RVASs) provided valuable insights into the genetics of many genetic disorders including epilepsy.[33,34,78] Testing rare variants individually requires prohibitively large sample numbers to achieve significance.[95] Therefore, analytic methods for RVAS usually group variants per gene (Figure 1.2). Grouping variants in larger units of several genes (gene set analyses) aggregates the signal from multiple related genes, aiming to achieve statistical significance.[47,96]

In RVASs, only a subset of variants, named *qualifying variants* (QVs), is evaluated. These variants are defined using a set of criteria that are meant to enrich the analysis for true disease associated alleles (e.g., filtering based on allele frequency and deleteriousness). To be considered definitive risk genes, candidate genes identified in RVAS need to have biological

relevance as well as to achieve study-wide significance (ideally, after correction for multiple testing at an appropriate probability threshold, e.g., $= 2.5 \times 10^{-6}$ when testing 20,000 protein coding genes). In practice, genes in which rare deleterious variants are frequent enough to reach exome-/genome-wide significance (e.g., *SCN1A*, *DEPDC5*) already have a known role in epilepsy, thus do not require additional functional validation.[33,78] Discovery of novel susceptibility genes or gene sets is otherwise corroborated with subsequent replication in independent cohorts and with functional validation.[47] Functional validation typically necessitates zooming in again to evaluate the individual carriers and thus may require more stringent variant evaluation processes compared to those used to define qualifying variants (see below).

Frameworks have been developed to define and evaluate the validity and relevance of genes identified in discovery studies, e.g., the Mendelian disease genes evaluation framework proposed by the Clinical Genome Resource (ClinGen).[97] These frameworks thus guide the transition in genetic testing from research to diagnostics. As mentioned, genetic screening for clinical purposes is adopted primarily for DEEs, NDDs or severe/difficult-to-treat phenotypes. It mostly employs exome sequencing (panel sequencing, clinical exomes, whole exomes) as first-tier choices with other scans being reserved for specific scenarios (e.g., WGS, array-based CNV scans and homozygosity mapping, multiplex-ligation probe-amplification).[39,75,86–89] Unresolved cases in standard clinical testing can be re-evaluated for novel discovery, thus linking diagnostics and research.[40,98] The evaluation of rare variants in individuals with rare epilepsies aims in general to maximize the odds of identifying those variants likely to have true disease relevance (i.e., substantially increasing disease risk).

Assessing the pathogenicity of a variant in an individual assumes an established disease-gene relationship and usually follows different standards depending on test providers and regulatory authorities.[99] For instance, the American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) framework[49] relies on a scoring system based on several genetic and functional characteristics to classify the variants in groups of (likely-)pathogenic variants or (likely-)benign variants. Variants without sufficient evidence to be classified in these groups are denoted variants of uncertain significance (VUS). Risk alleles (from GWAS loci) are typically interpreted in terms of the relative risk/odds ratio of the associated phenotype. Attempts to calculate polygenic risk scores from GWAS loci are promising.[66] These are nonetheless far from clinical implementation.

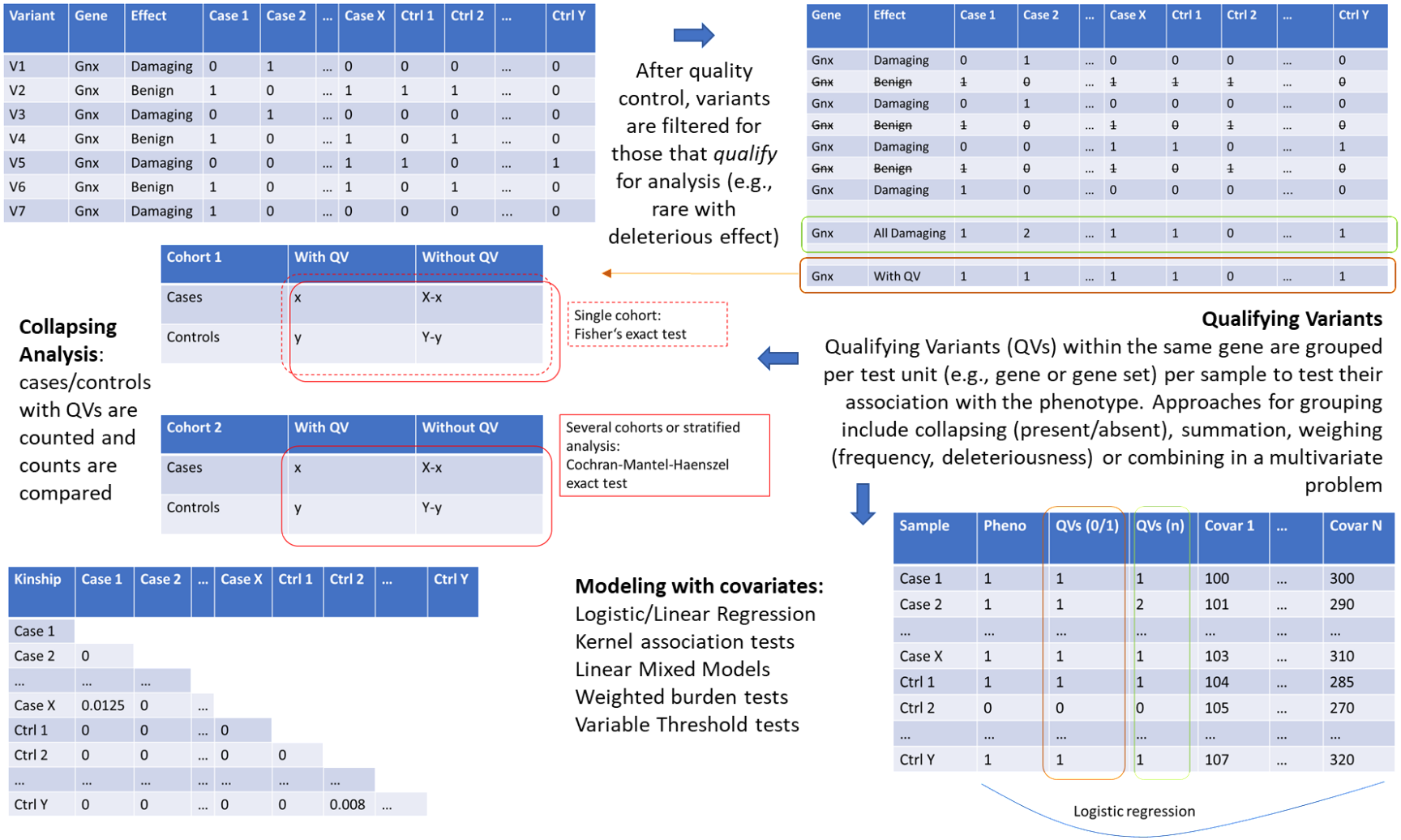


Figure 1.2: Key approaches to rare variant association analysis. Collapsing variants (assuming equal weight) followed by Fisher's exact test are appropriate for analyzing variants predicted to be deleterious, in well-matched cohorts. Cochran-Mantel-Haenszel exact test can be used for stratified analyses (e.g., separate analyses of males and females as a proxy for including sex as a covariate). Regression analyses allows for including covariates. Kernel association tests were designed to test variant sets (e.g., genes) and combining rare and common variants. Linear Mixed Models allow for modelling kinship and thus including related samples. Testing variants at different minor allele frequency thresholds can be done using Variable Threshold tests. Similarly, variants can be assigned different weights in Weighted Burden tests.

1.3 Epilepsy genes and variants

1.3.1 Epilepsy-related genes

The Online Mendelian Inheritance in Man (OMIM) database[100] lists (end of 2021) a hundred DEE genes (Phenotypic Series: PS308350). More than four hundred genes are linked to various genetic epilepsy syndromes in Genomics England PanelApp (https://panelapp.genomicsengland.co.uk/), a tool for collaborative gene panel sharing and review (panel 402, v2.2). Most of these are DEE or NDD genes (as detailed above, there is robust evidence for monogenic causality in rare and severe epilepsies). On the other hand, very few genes have an established link with common or mild epilepsy syndromes (though, not necessarily a statistical association), e.g., ion channels, neurotransmitter receptors and synaptic proteins. Notably, most genes implicated in GGE or NAFE are also DEE genes.[65,69–71] Some were identified independently in rare and common epilepsies whereas others were implicated in susceptibility to less severe phenotypes after their initial discovery in individuals diagnosed with GEFS+, DEE or NDD – through detailed phenotypic characterization of large series of individuals carrying pathogenic and likely pathogenic variants.[50,69]

Whereas variants in a handful of genes segregate in families with GGE (e.g., *GABRG2*, *GABRA1*, *GABRA5*, *GABRB2*), single genes did not show significant association with GGE so far, including in the most recent and largest RVAS in epilepsy which analyzed samples from a wide range of populations.[33,34,78] *SCN1A* is the commonest gene in large scale RVAS in DEEs, reaching study-wide significance.[33] *DEPDC5* is a major risk determinant in NAFE with exome-wide significance.[78] *KCNQ2*, *KCNQ3*, *SCN2A*, *SCN8A* and *PRRT2* variants are rather prevalent in self-limited, less severe focal epilepsies diagnosed in neonates and infants, though they are implicated in severe phenotypes too.[29] Several genes encoding subunits of excitatory receptors (e.g., *CHRNA2*, *CHRNA4*, *CHRNB2*) are implicated in familial focal onset epilepsies that could have a later onset. Despite original suggestive evidence, the contribution of several candidate genes to the etiology of GGE (e.g., *CACNA1H*, *EFHC1*, *ICK*) has been disputed or refuted.[56,101,102] Studies examining the mutational burden in gene sets have shown an increased burden in deleterious URVs in genes encoding $GABA_A$ receptor subunits and GABAergic pathway genes particularly in GGE.[33,47] Of note was an increased burden in genes associated with dominant epilepsy syndromes, DEE genes, and NDD-Epilepsy genes both in common GGEs/NAFEs and rare epilepsies, emphasizing a shared genetic component.[33,34,78]

1.3.2 Ultra-rare variants in epilepsy

Ultra-rare variants (URVs) are those variants seen in one or a few patients while absent from the general population.[33,78] Collectively, these constitute most variants seen in humans (e.g., in large sequencing studies, biobanks and databases[103,104]). Their extremely low frequency reflects either a recent origin (the allele did not propagate through enough generations to be common) or negative selection due to reduced fitness.[64,105] URVs have shown enrichment in several – particularly, neurological – diseases, reflecting their high effect size.[106–108] In DEE and NDD genes, the role of URVs with predicted or proven functional effects is well-established – sometimes with consequences on precision treatments.[21,48,73,109–111] Since the risk in siblings of individuals with common or mild epilepsies – though higher than the risk in the general population – is lower than what is expected from monogenic inheritance,[65,112] it seems likely that a combination of ultra-rare, rare and common variants, polygenic and environmental modifiers have a substantial role in less severe phenotypes. Interestingly, URVs but not rare coding variants have shown replicable association with these common epilepsies in several gene sets; most patients who carry functional variants (predicted or validated, that segregate with an epilepsy phenotype or are *de novo* variants) have ultra-rare, rather than rare or common, short coding alterations.[47,64,78]

1.4 Rationale and objectives

To summarize, the current paradigm in explaining the complex inheritance of many epilepsy syndromes is that of "several interacting or additive common risk elements with low effect, which can be overlayed by ultra-rare *de novo* or inherited, typically heterozygous, variants with high effect".[38] High effect URVs are typically associated with severe or familial disease presentations. Therefore, the investigation of single independent families or few families with very similar phenotypes is the preferred approach to identify presumably monogenic variants with large effect size.[65] Sequencing-based analyses, the contemporary method for gene discovery, led to a surge of discoveries, implicating various genes coding for ion channels, neurotransmitter receptors, and synaptic proteins as well as enzymes and structural proteins in various epilepsies.[59] In these genes, heterozygous short coding URVs are the commonest type of alterations linked to disease predisposition. Deleterious coding URVs in *SCN1A* (in DEE) and *DEPDC5* (in NAFE) are major risk determinants (showing study-wide significance).[33,34,78] A high burden of URVs in genes encoding $GABA_A$ receptor subunits and

in GABAergic pathway genes points to the importance of genes critical for inhibitory pathways.[33,34,47,78] Comparisons across the phenotypic spectrum (DEE, GGE, NAFE) revealed a high URV burden in gene sets of known epilepsy genes, suggesting a shared genetic component. Further comparisons between coding variants not seen in the general population vs. those seen at low frequencies suggested that URVs have a predominant role in predisposing to various epilepsies.[33,78] Additional comparisons of genetic predisposition in familial vs. sporadic NAFE indicated a higher burden of coding URVs in individuals with family history (vs. controls) compared to those with sporadic disease (vs. identical controls).[78]

Nonetheless, understanding the elements of genetic predisposition to epilepsy is yet unachievable in many individuals, particularly in common, genetically complex epilepsies. Bi-allelic inheritance – which has an established role in DEEs and NDDs – has been suggested to play a role in predisposing to the commoner, less severe forms of epilepsy, but the evidence remains very limited and the influence of bi-allelic variants on the genetic risk remains therefore poorly understood.[113,114] Also, prior RVAS of GGE have resulted in different lead candidates in GGE (top-ranked genes), whereas *DEPDPC5* was replicated in two RVAS of NAFE.[33,78] It remains unknown whether the some of the top-ranked genes in previous GGE studies, or possibly different genes, might reach study-wide significance through meta-analyses of existing cohorts. Comparisons of genetic predisposition in familial vs. sporadic disease has been performed in focal but not generalized epilepsy,[78] although familial disease is more frequent in generalized epilepsy compared to focal epilepsy. Similarly, the association of numerous biologically informed gene sets with plausible relevance to the epileptogenesis (i.e., presumed to be important for the pathogenesis of epilepsy) with the disease is yet to be investigated and compared between GGE and NAFE. Further questions naturally remain, but these are a few with relevance to this dissertation.

Accordingly, the *general objective* of this doctoral work was to study the association of ultra-rare coding variants with familial and sporadic epilepsy. The *specific objectives* were:

1. To assess the contribution of bi-allelic variants, compared to heterozygous URVs, to the genetic risk of familial epilepsies in consanguineous families from Sudan.
2. To measure the association of ultra-rare coding variants with common generalized epilepsy in the presence and absence of family history.
3. To estimate the burden of ultra-rare coding variants in key gene sets in common generalized epilepsies in comparison to common focal epilepsies and rare encephalopathies.

1.5 Research Approach and Dissertation Structure

1.5.1 Approach

In practice, addressing the abovementioned objectives requires targeted approaches to enrich the analyses for disease-relevant variants. Bi-allelic variants are best studied in consanguineous families. Large sample sizes necessary for studies of heterozygous URVs in complex epilepsies (in which single major genes are not expected to be common) can be achieved using joint analysis of existing datasets collected by international collaborations, particularly those enriched for familial cases. Variant prioritization strategies based on *in silico* scores specifically designed to prioritize rare variants serve to enrich gene set burden analyses with high effect variants. Taking advantage of these approaches, these studies were performed:

- **Studying the association of bi-allelic coding variants with familial epilepsy in Sudanese families**

The aim of this family-based genetics study in a Sudanese cohort was to identify the variants and genes underlying epilepsy in several consanguineous families using exome sequencing, with focus on bi-allelic inheritance. We investigated five families with two or more siblings diagnosed with epilepsy, whose parents are cousins. The nature of the genetic ancestry of these families offered insights into one of the understudied populations (African population).

- **Studying the association of coding variation in protein coding genes with familial and sporadic generalized epilepsies**

Five cohorts of individuals diagnosed with GGE collected and sequenced by the Canadian Epilepsy Network, Epi4K Consortium, Epilepsy Phenome/Genome Project, EpiPGX Consortium, and EuroEPINOMICS-CoGIE Consortium were jointly analyzed and compared to ancestry-matched controls using rare variant collapsing association analysis. This was followed by separate analyses in individuals with a positive family history (familial GGEs) and those without a family history (sporadic GGEs) to highlighted differences between the disease forms.

- **Studying the association of coding variation in biologically informed gene sets with common and rare epilepsies**

A cohort of European individuals with epilepsy collected by the Epi25 Collaborative was examined for evidence of enrichment of ultra-rare conserved and constrained variants in

key gene sets. Patients from DEE, GGE, and NAFE cohorts were compared to matched population controls to establish key novel associations of gene sets and pathways with epilepsy.

1.5.2 Dissertation structure

To facilitate an easy navigation of methods and results as well as an accurate description of my contributions to the outlined collaborative research projects, this monograph presents the studies highlighted above in separate chapters (chapters 2 – 4) that I adapted from published articles or articles in preparation. These chapters follow the conventional structure: summary, background, methods, results, and discussion sections. These three chapters are followed with concluding remarks where I highlight the relevance of the findings in the context of the overall topic of association of ultra-rare variants with epilepsy. To acknowledge the increasingly collaborative nature of genetic research, this monograph presents the scientific findings primarily using the plural pronoun *we*, and a statement of contributions is provided at the end of this dissertation. I confirm that the presented monograph is my own work and follows the licensing agreements of published materials.

Chapter 2: The association of bi-allelic variants with epilepsy in Sudanese families

This chapter was adapted from a manuscript in preparation (**Koko** *et al*. 2022). See the statement of contributions at the end of this dissertation.

2.1 Summary

Background: We studied the role of recessively inherited variations in the etiology of epilepsy in consanguineous Sudanese families.

Methods: We investigated five families in which epilepsy was the main presentation in two or more siblings whose parents are cousins. To identify candidate disease variants, whole exome sequencing (WES) in one affected individual was coupled with homozygosity mapping in several siblings and their parents. Homozygous, compound heterozygous and heterozygous alterations were considered. These were then evaluated using segregation analysis. Copy number variant (CNV) analysis was performed using WES and array data.

Results: A homozygous pathogenic *PRRT2* splice-site variant (IVS1-1G>A) was detected in three siblings presenting with familial infantile epilepsy, a phenotype that is typical of monoallelic but not biallelic *PRRT2* alterations. A heterozygous likely pathogenic missense variant in *PCDH19* (p.(Asp375Val)) was identified in another family, with a phenotype within the spectrum of known *PCDH19*-related presentations. Additional missense variants of uncertain significance were identified in *SPTAN1*, *GRIN2B*, and *SCN3A*. Two previously reported rare *EFHC1* variants were seen in two families but were classified benign. No disease related CNVs were identified.

Interpretation: We did not find sufficient evidence to support a common role for recessive inheritance in the etiology of common genetic epilepsies. Nonetheless, we expanded the phenotypes of homozygous *PRRT2* variants to include mild epilepsy without movement disorders.

2.2 Background

Single gene defects cause a wide spectrum of developmental and epileptic encephalopathies as well as neurodevelopmental disorders with epilepsy.[32] Less severe genetic epilepsies are typically complex, with few individuals or families harboring pathogenic variants in single genes.[115–117] These complex familial epilepsies allow identifying genes mediating high disease risk through linkage or sequencing studies, especially when the pedigrees show several affected siblings with homogeneous phenotypes.[38] Early discoveries implicated heterozygous variants in several genes in predisposing to various genetically complex epilepsy syndromes.[65] A substantial role for recessive inheritance has been stipulated only in few studies, but definitive evidence is still lacking.[113,118–120] As consanguinity is not prevalent in European populations, it is possible that studies from non-European populations could help decipher the genetic background in seemingly recessive epilepsies. Not much is known about the genetics of epilepsy in African populations, especially those with high consanguinity. The Sudanese population has a particularly high rate of consanguinity and extended families.[121–123] Prior epidemiological studies showed a considerable burden of epilepsy in Sudan.[124,125] More recently, a genetic study in Sudanese families showed a potential for identifying new risk genes predisposing to genetic epilepsies.[126] Hypothesizing that Sudanese families are suitable to touch upon the role of bi-allelic inheritance in less severe epilepsies, we studied several families with multiple siblings diagnosed with a genetic epilepsy, in which the parents are cousins.

2.3 Methods

2.3.1 Overview of the study design

We performed a series of family-based studies to investigate the genetics of neurological disorders in Sudanese families (the Sudanese Neurogenetics Project). More than two hundred individuals were investigated during the first phase of the project (2012 – 2018), while a second phase is currently ongoing. For these studies, individuals were considered for inclusion during their visits to several (tertiary) neurology clinics in Khartoum, Sudan. The phenotypes and family history were provisionally assessed, and the parents/families of these individuals were then approached to inquire about their willingness to participate. Selected families were then investigated using either exome panels or whole exome sequencing.

Multiple novel findings were published, which guided our understanding of the genetics of neurological diseases in Sudan.[122,127]

As part of this project, we aimed to examine the genetic etiology in individuals diagnosed with a familial generalized epilepsy, who had a background of parental consanguinity. Sixty individuals in our cohort (seen between 2012 – 2018) were documented to have had epilepsy as the primary presentation. Those index patients were evaluated for suitability for further family-based genetic analysis based on these criteria: (1) a clinically diagnosed mild genetic epilepsy, with generalized seizures as part of the presentation (2) at least one additional sibling diagnosed with epilepsy, and (3) parents who were first- or second-degree cousins. The families of 14 probands were reachable for further inquiries about the family history and agreed to a genetic workup, including five probands (with consanguineous parents) who had a developmental and epileptic encephalopathy or a neurodevelopmental/early-onset neurodegenerative disorder with epilepsy (families E1, E6, E10, E12, E13, E14), three with a less severe epilepsy but not with a background of close parental consanguinity (E4, E9, E15), and six that matched the criteria above (E2, E3, E5, E7, E8, E11). This chapter presents the results obtained from the workup performed in five families (E2, E3, E5, E8, E11), whereas one family (E7) was excluded as contact was lost prior to sampling. The methods used in our genetic workup are outlined in Figure 2.1.

2.3.2 Ethical approvals, phenotyping and sampling

Ethical approvals were obtained from the local committees of the participating institutions. Informed written consent (and assent when applicable) was obtained from adult participants or parents of children. The probands were evaluated at the Pediatrics Neurology Clinic, Soba University Hospital (Khartoum, Sudan) or during house visits when needed. Saliva samples were collected from the participants using Oragene OG-500 kits (DNA Genoteck, Canada). DNA was extracted according to the PrepIT L2P protocol provided by the manufacturer.

2.3.3 Whole exome sequencing and genome-wide genotyping

Whole exome sequencing (WES) was performed for one proband per family (selected based on the quality of extracted DNA). The enrichment was done using the Agilent Sure Select XT Human All Exon V6 kits (Agilent, US), and multiplexed paired-end sequencing was then performed on Illumina HiSeq4000 or NovaSeq platforms (Illumina, US). The average sequencing depth of coverage ranged approximately between 50- and 150-fold. Four to six samples per family (all available siblings diagnosed with epilepsy, their parents, and one or

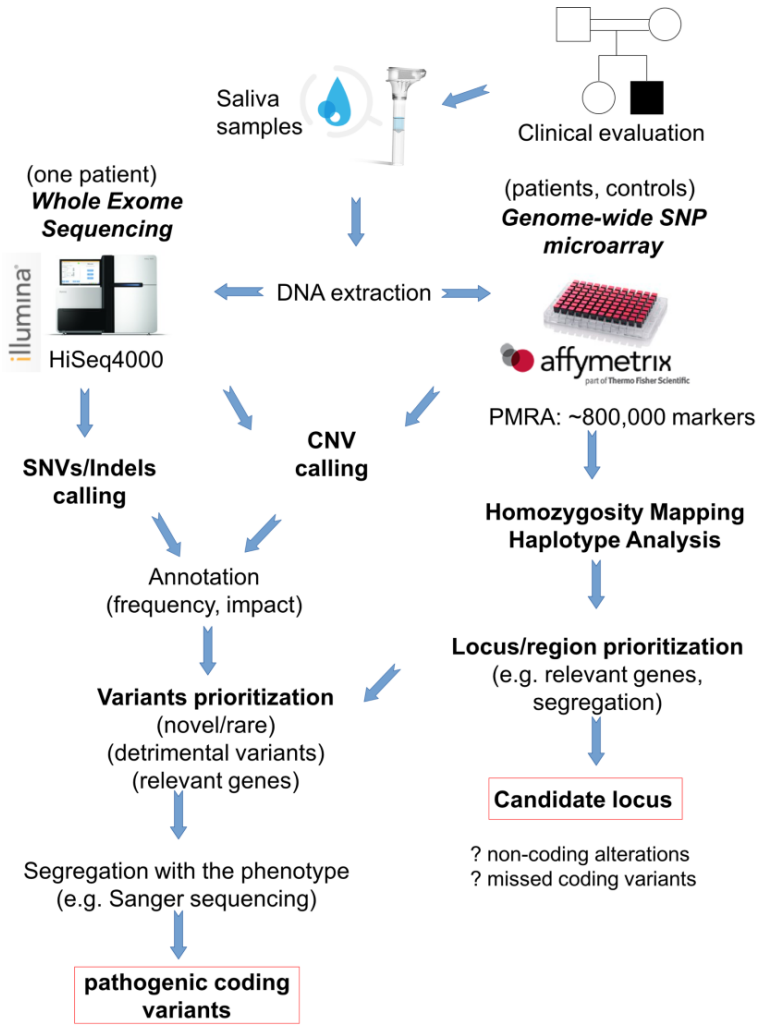


Figure 2.1: Flowchart of the methods used to investigate the genetics of epilepsy in Sudanese families.
PMRA: Precision Medicine Research Array.

two elder siblings not diagnosed with epilepsy) were genotyped using the Affymetrix Precision Medicine Research Array (Affymetrix Inc. - Thermo Fisher Scientific, Dreieich, Germany). These arrays contain around 800,000 genome-wide probes. These experiments were performed at Cologne Center for Genomics (CCG, Cologne, Germany).

2.3.4 Exome and array data processing

The alignment of WES reads on the Genome Reference Consortium human genome build 38 (GRCh38) was performed using bwa kit v0.7.15.[128] Read sorting, duplicate marking, base quality score recalibration, and haplotype calling were performed for individual samples using the Genome Analysis Toolkit (GATK, v4.1.4.1).[129] These data were then processed jointly with additional WES data from 126 Sudanese individuals from 68 unrelated families (internal database). All files in genome Variant Call Format (gVCF) were imported to a GATK GenomicsDB database and genotyping was then performed jointly on all samples, limited to the exonic regions and immediate exon-intron junctions of protein coding genes as derived from consensus coding sequence (release 37).[130] Variants with multiple alleles were split, and all variants were normalized and sorted using vt[131] (v0.57) and htslib/bcftools[132] (v1.10). Variant effect annotation (affected genes, consequences on mRNA and protein for RefSeq/Ensembl transcripts) was carried out using Variant Effect Predictor (VEP, Ensembl release 104).[133] Additional annotations were obtained including minor allele frequencies from several databases (gnomAD[103] r3.1.2, TOPMed[134] Freeze 8, and DiscovEHR[135] Freeze 50) and several *in silico* metrics of deleteriousness, conservation and constraint (GERP++ RS,[136] CADD,[137] MTR,[138] REVEL,[139] M-CAP,[140] ClinPred,[141] ada,[142] rf,[142] TraP,[143] MaxEntScan[144]). Annotations not available through VEP or its dbNSFP/dbscSNV extensions[145] were obtained directly from their online servers or repositories. The genome-wide array data was processed according to the manufacturer's protocol (Affymetrix Inc.). Genotype calling was performed using BRLLM method implemented in Axiom Power Tools.[146] Homozygosity mapping was performed using Homozygosity Mapper.[147] Haplotype analysis around candidate heterozygous variants was carried out manually. Additionally, copy number analysis was performed using PennCNV.[148]

2.3.4 Variant prioritization

Heterozygous, homozygous and hemizygous variants annotated as predicted loss-of-function/protein truncating variants (nonsense, frameshift, and canonical splice-site alterations), in-frame indels (not overlapping known repeat sites) and missense variants (with a CADD[137] score > 20 and GERP++ RS[149] score > 2) were considered. To examine for

recessive inheritance, we filtered for rare variants (minor allele frequency less than 0.5% and less than two homozygous calls) that lied in a homozygous run or in a gene with shared haplotypes in all affected siblings. For dominant inheritance models, we filtered for variants with an alternate allele count < 3 in gnomAD.[103] Afterwards, we filtered the variants for those located in genes with known or suspected association with epilepsy or developmental disorders with epilepsy by performing literature search for genes with candidate variants.[150] Additionally, we evaluated ClinVar[151] release 2020-12-26 variants regardless of their allele frequency. Finally, we explored the allele frequency of our candidate variants in our dataset of jointly called exomes from Sudanese individuals to ensure that the candidate variants are not common population-specific polymorphisms. We used leave-one-out allele frequencies from 130 individuals, excluding the sample under investigation (i.e., examining the frequency in four other individuals from this study, 77 individuals with neurological phenotypes and 49 controls including individuals with non-neurological phenotypes not related to seizure disorders). Copy number variants were examined among family members as well as in the Database of Genomic Variants[152] and gnomAD.[103]

2.3.5 Variant validation and classification

The candidate variants were validated using Sanger sequencing (GATC Biotech – Eurofins Genomics, Cologne, Germany or LGC Genomics, Berlin, Germany). The segregation of the final candidate variants was evaluated using Sanger sequencing or microarray data. To accommodate both mono-allelic, bi-allelic inheritance, dosage effects and incomplete penetrance, segregation with the phenotype was considered positive when (1) all individuals with epilepsy carried the variant in heterozygous or homozygous state and (2) none of the controls had the variant in a homozygous state (heterozygous carriers without the phenotype did not rule out segregation). The variants were then classified according to the American College for Medical Genetics and Genomics and Association for Molecular Pathology (ACMG/AMP) framework applied using the Genetic Variant Interpretation Tool.[49,153] The following criteria were utilized: BP1 (Missense variant in a gene for which primarily truncating variants are known to cause disease), BP6 (Reputable source recently reports variant as benign but the evidence is not available to the laboratory to perform an independent evaluation), BS1 (Allele frequency is greater than expected for disorder), BS4 (Lack of segregation in affected members of a family), PB5 (Variant found in a case with an alternate molecular basis for disease), PM1 (Located in a mutational hot spot and/or critical and well-established functional domain without benign variation), PM2 (Absent from controls, or at extremely low frequency

and absent in homozygous state if recessive, in gnomAD, TOPMed BRAVO, and DiscovEHR), PM5 (Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before), PP1 (Cosegregation with disease in multiple affected family members in a gene definitively known to cause the disease, considered Moderate evidence), PP2 (Co-segregation with disease in multiple affected family members in a gene definitively known to cause the disease), PP3 (GERP, CADD and two additional scores support a deleterious effect on the gene or gene), and PVS1 (null variant in a gene in which LOF is a known mechanism of disease).

2.4 Results

We identified one homozygous splice acceptor ultra-rare variant (URV; not seen in population controls) and several heterozygous missense URVs in epilepsy-related genes. No rare homozygous variants (seen at a low frequency in population controls) with plausible disease relevance were identified, whereas two rare heterozygous variants previously reported in ClinVar were found. Copy number analysis was unremarkable.

2.4.1 Pathogenic and likely pathogenic epilepsy-related variants

A pathogenic homozygous variant in *PRRT2* (c.-65-1G>A) segregated in three siblings (Family 11) with a homogenous phenotype of generalized (tonic, atonic) seizures with infantile onset (Figure 2.2A). The presenting symptom in all three siblings was tonic seizures in the first year of life, described by their mother as episodes of generalized body stiffness with sudden onset, lasting for seconds up to a minute, followed by full recovery. Additionally, two elder siblings had atonic seizures described as episodes of floppiness, of abrupt onset, lasting for a few minutes. In all three siblings, these seizure attacks were infrequent, associated with confusion and followed by full recovery. There was no association with fever or other illnesses. Also, there was no history of abnormal movements with or without the seizure episodes. Pregnancy and birth history were uneventful, and all siblings achieved developmental milestones appropriate for their age, without a history of developmental delay. The elder siblings at school age had satisfactory performance, and there was no impression of an intellectual disability (formal testing was not performed).

Following the first occurrence (tonic seizures at 3 months of age), the eldest sibling (8 years of age, female) experienced two further episodes of tonic seizures (two weeks apart). Additionally, four episodes of atonic seizures happened over a period of three months.

Available EEG reports (sleep EEG at 5 months of age) indicated a generalized brain slowing with activity in the delta/theta range, scanty alpha waves, without spike activity. She was started on valproate treatment (VPA) at the time with no further seizures for two years. A trial to stop VPA was followed by recurrence (4 episodes of seizures in one day) which was controlled with continued use (no further seizures). A follow up awake EEG record (at the age of 6 years old) showed normal brain activity and no epileptiform discharge was seen. Another trial to discontinue VPA treatment (approximately after another two years) was successful.

The second sibling (5 years of age, male) had eight episodes of tonic seizures in one day as the first presentation (3 months of age) with full recovery in between. He experienced a temporary regression of milestones at disease onset (loss of head support) followed by full recovery over a few weeks. A sleep EEG record at the time was normal. He was started on VPA treatment, during which further six episodes of atonic seizures occurred over a period of three months. VPA treatment was successfully tapered off after two years of continued use. A second sleep EEG at the age of 2.5 years was normal. The third sibling (male) was seen a few weeks after the first episode of tonic seizures (at the age of 4 months), which affected the upper limbs only. The mother observed an episode of rhythmic tongue jerking lasting for several seconds without involvement of jaw or face muscles. A similar episode followed two days after, whereas a third one after another day involved the whole body. A sleep EEG record was normal. He was just started on phenobarbital (PB) at the time of evaluation.

The mother of these siblings did not recall having experienced any episodes of seizures or abnormal movements, whereas the father reported an episode (around the age of 30 years) of brief loss of body tone and consciousness (falling on a table in front of him while sitting) which was not preceded with any prodromal symptoms. It is thought to have lasted a few minutes, without further occurrence of similar episodes till the time of evaluation. No EEG was performed, and cardiac evaluation was reported to be normal. There was a family history of seizure disorders (cousins of the proband). It was not possible to obtain an accurate description of the phenotypes (patients who were not available for evaluation). However, the disease was reported to have a comparatively severe nature.

Exome sequencing revealed a splice site variant (IVS1-1G>A) in *PRRT2* which was in a homozygous run in all three patients. The segregation of this canonical splice-site variant in these siblings was confirmed using Sanger sequencing. An unaffected sibling did not carry the variant, whereas their parents were heterozygous carriers (Figure 2.2A). This alteration in the junction of the first intron and second exon is predicted to disrupt the splicing and was classif-

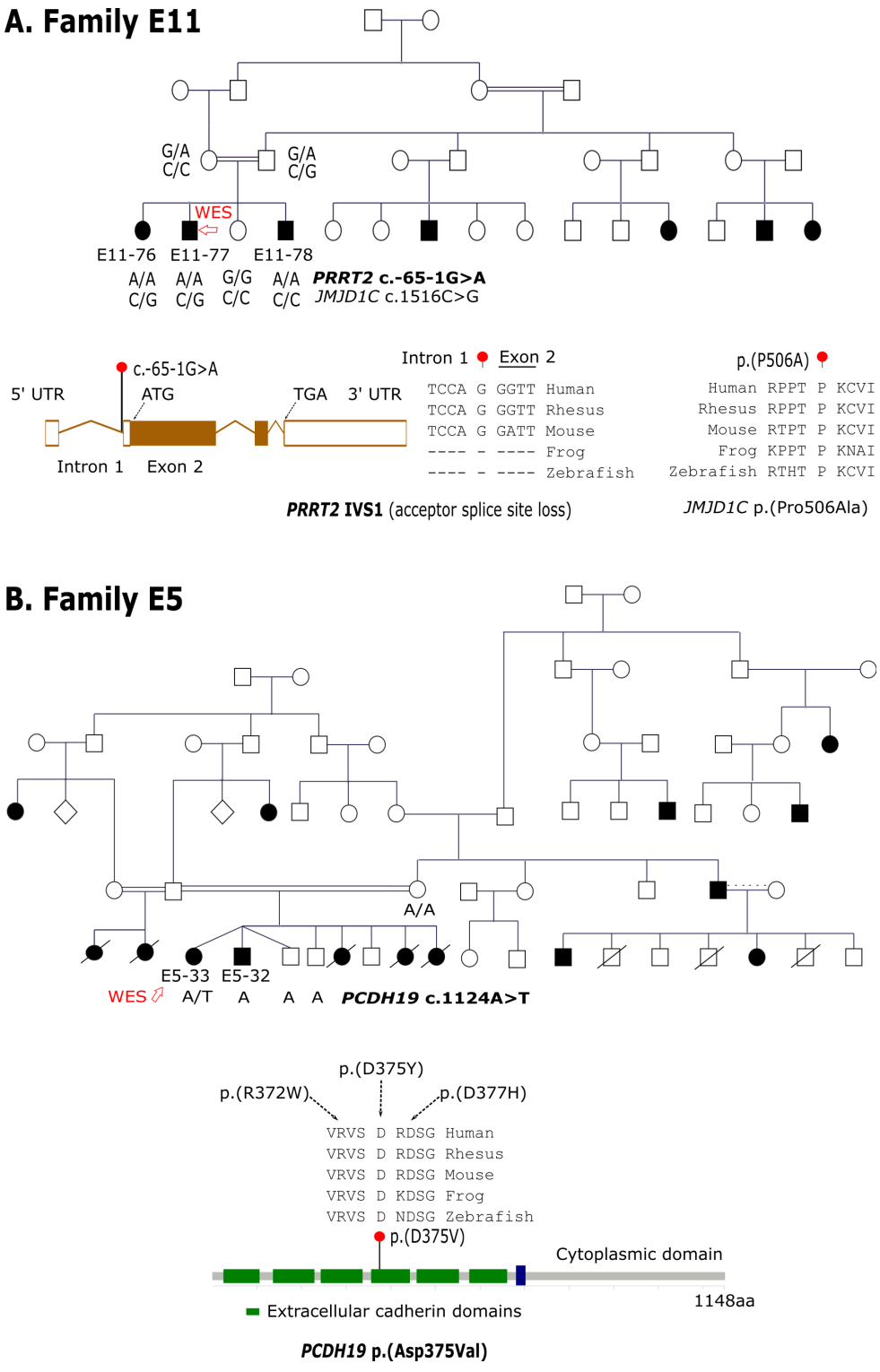


Figure 2.2: Pathogenic variants identified in dominant epilepsy genes in Families E11 and E5. A. Pedigree of Family E11 showing the segregation of two variants in *PRRT2* (homozygous; pathogenic) and *JMJD1C* (heterozygous). *PRRT2* variant (classified as pathogenic) is in an exon-intron junction, whereas *JMJD1C* missense variant (classified as benign) changes a conserved amino acid. **B.** Pedigree of Family E5 showing a missense variant in *PCDH19* (heterozygous) identified in a proband. This variant (p.(D375V); classified as likely pathogenic) is located extracellularly in a cadherin domain where several pathogenic variants were previously identified (black arrows). The individual identifiers of investigated patients are shown (see Table 2.1 for the phenotypes) and probands investigated using whole exome sequencing (WES) are indicated with a red arrow.

ified as pathogenic (ACMG/AMP framework). A second, heterozygous missense variant in *JMJD1C* (p.(Pro506Ala)) was paternally inherited in two affected siblings (Tables 2.1 and 2.2); it was classified as likely benign.

A likely pathogenic variant in *PCDH19* (p.(Asp375Val)) was identified in a 19-year-old female in Family E5 who was diagnosed with epilepsy in the first year of life. Generalized tonic-clonic seizures (GTCS) were the only recalled seizure type but the seizure onset (focal or generalized) was not known. No febrile seizures were described, and the developmental history was normal. The exact number of seizures was not known but these were described as infrequent. Available records indicated normal neurological findings at the time of diagnosis and a neurological examination at the time of sampling was normal apart from subtle ptosis on the left side. She did not have an EEG recording and was successfully treated with Carbamazepine (CBZ), which was stopped after four years. She had one brother and five sisters diagnosed with epilepsy (Figure 2.2B). According to the family, all five female siblings died in early childhood following respiratory tract infections that were not associated with seizures (three full sisters and two other paternal sisters). The brother had a late disease onset at the age of 12 years. Seizure attacks (GTCS) were usually preceded with numbness felt in the upper limb on the right side. Neurological examination showed mild weakness and hyperesthesia (pain and vibration sense) on the left side. He also had mild spasticity bilaterally (upper and lower limbs) and hyperreflexia (with diffused biceps and adductor reflexes). An MRI report indicated the presence of symmetrical parietal abnormal signal intensity (peri-Rolandic cortex) and small high parietal areas of encephalomalacia. A sleep EEG showed infrequent generalized sharp waves. At the time of sampling, he was on treatment with CBZ for four years (no seizures on treatment). He, as well as the mother of these patients, did not carry the variant and a sample from the father was not available to validate the origin of the allele (i.e., a paternal or a *de novo* origin are possible).

2.4.3 Additional variants of uncertain significance

We identified several URVs in the remaining families (Table 2.1). In Family E2, a heterozygous missense variant in *SPTAN* (p.(Gly1793Cys)) was identified in a 7-year-old and her brother (5 years old), who were both diagnosed with epilepsy in the first year of life. The proband had the first seizure attack (GTCS) at the age of three days, followed two days later with a status epilepticus. An EEG performed at the time showed recurrent runs of generalized spike and wave complexes. Following another status epilepticus at the age of one year, she had a temporary developmental regression (motor milestones, speech) and developed a squint.

Table 2.1: Summary of presentations and genetic findings in siblings diagnosed with epilepsy from five Sudanese consanguineous families.

ASM: Anti-Seizure Medications. CBZ: Carbamazepine. FID: Family Identifier. GTCS: Generalized Tonic-Clonic Seizures. ID: Intellectual Disability. IID: Individual Identifier. PB: Phenobarbitone. VPA: Valproate. WES: Whole Exome Sequencing. Onset: age at first seizure in days (d), months (m) or years (y).

FID	IID	Age, Sex	Onset	Presenting symptom	Related symptoms	EEG	ASM	Response	Genetic Investigation	Genes with a candidate variant
E2	11	7y, F	3d	GTCS	-	Generalized spike-and-wave complexes	VPA	Responder (no seizures)	WES	*SPTAN1, GRIN2B*
	13	5y, M	1y	GTCS	Inattentive, overactive	NA	VPA	Responder (no seizures)	Segregation	*SPTAN1*
E3	16	23y, F	13y	Myoclonic seizures	GTCS	NA	CBZ, VPA	Responder (no seizures)	Segregation	*EFHC1*
	17	19y, M	13y	Myoclonic seizures	Tonic seizures	NA	VPA	Responder (infrequent seizures)	Segregation	*EFHC1*
	18	14y, M	13y	Myoclonic seizures	GTCS	NA	VPA	Responder (no seizures)	WES	*SCN3A, EFHC1*
E5	33	19y, F	1y	GTCS	-	NA	CBZ	Responder (off ASM)	WES	*PCDH19*
	32	19y, M	12y	GTCS	Left-sided weakness and hyperesthesia	Generalized sharp waves	CBZ	Responder (no seizures)	Segregation	-
E8	62	16y, M	12y	Myoclonic seizures	-	Focal epileptic discharges (left temporal)	VPA	Responder (> 50% reduction in seizures)	Segregation	-
	61	11y, F	9y	Myoclonic seizures	Atonic seizures	Focal epileptic discharge (right occipital)	LEV	Responder (infrequent seizures)	WES	*DEPDC5, EFHC1*
E11	76	8y, F	3m	Tonic seizures	Atonic seizures	Generalized slowing	VPA	Responder (off ASM)	Segregation	*PRRT2, JMJD1C*
	77	5y, M	3m	Tonic seizures	Atonic seizures	Normal	VPA	Responder (off ASM)	WES	*PRRT2, JMJD1C*
	78	6m, M	3m	Tonic seizures	-	NA	PB	Unclassified (just started on ASM)	Segregation	*PRRT2*

Table 2.2: Frequency, deleteriousness, and segregation of candidate variants in epilepsy-related genes.

Note that the association of *EFHC1* variants with genetic epilepsy has been disputed (see the discussion) but the variants were recurrent. Candidate variants were filtered based on minor allele frequency, CADD deleteriousness score >20 and GERP++ RS conservation score > 2 (see the methods). Additional supplemental prediction and scores are given for missense variants (MTR, REVEL, ClinPred, M-CAP) and splice-site variants (ada, TraP, rf, MaxEntScan); "D" indicates a score in the pathogenic/deleterious/damaging range whereas "T" indicates score in the tolerated/benign range. DEE: Developmental and Epileptic Encephalopathy. GGE: Genetic Generalized Epilepsy. HGNC: HUGO Gene Nomenclature Committee genes names. ID: Intellectual Disability. JME: Juvenile Myoclonic Epilepsy. MAF: Minor Allele Frequency. MCD: Malformations of Cortical Development. NAFE: Non-acquired Focal epilepsy. NDD: Neurodevelopmental disorder. pLOF: predicted loss-of-function.

Family ID	Variant					MAF				*In silico* predictions	
	Genomic change	HGNC Gene	Related pheno-types	coding change	protein change	gnom-AD	TOP-Med	Disco-vEHR	130 Sd.	GERP++ RS & CADD	Other scores
E2	Chr9-128617674-G-T (heterozygous)	*SPTAN1*	DEE, NDD/ID	NM_003127 c.5377G>T	p.(G1793C)	0	0	0	0	GERP: 5.8 CADD: 31.0	MTR: 0.40 (D) REVEL: 0.60 (D) ClinPred: 1.00 (D) M-CAP: 0.13 (D)
	Chr12-13563804-T-C (heterozygous)	*GRIN2B*	DEE, NDD/ID	NM_000834 c.3434A>G	p.(H1145R)	0	0	0	0	GERP:5.2 CADD: 25.0	MTR: 0.53 (D) REVEL: 0.70 (D) ClinPred: 0.98 (D) M-CAP: 0.32 (D)
E3	Chr2-165146925-A-C (heterozygous)	*SCN3A*	DEE, MCD	NM_006922 c.1485T>G	p.(S495R)	0	0	0	0	GERP: 5.6 CADD: 24	MTR: 0.70 (D) REVEL: 0.66 (D) ClinPred: 1.00 (D) M-CAP: 0.25 (D)
	Chr6-52454102-G-A (heterozygous)	*EFHC1*	GGE/JME	NM_018100 c.731G>A	p.(R244Q)	1.86×10^{-4}	0.91×10^{-4}	<0.001	0	GERP: 5.9 CADD: 34	MTR: 1.03 (T) REVEL: 0.73 (D) ClinPred: 0.90 (D) M-CAP: 0.16 (D)
E5	ChrX-100407474-T-A (heterozygous)	*PCDH19*	DEE, GEFS+, ASD	NM_020766 c.1124A>T	p.(D375V)	0	0	0	0	GERP: 6 CADD: 26.5	MTR: 0.66 (D) REVEL: 0.94 (D) ClinPred: 1.00 (D) M-CAP: 0.96 (D)

Table 2.2: Continued.

Family ID	Variants					MAF				*In silico* predictions	
	Genomic change	HGNC Gene	Known pheno-types	coding change	protein change	gnom-AD	TOP-Med	Disco-vEHR	130 Sd.	GERP++ RS & CADD	Other scores
E8	Chr22-31815135-C-T (heterozygous)	*DEPDC5*	NAFE, MCD	NM_014662 c.1589C>T	p.(A530V)	0	0	0	0	GERP: 6 CADD: 23.6	MTR: 1.09 (T) REVEL: 0.08 (T) ClinPred: 0.90 (D) M-CAP: 0.02 (T)
	Chr6-52465035-C-T (heterozygous)	*EFHC1*	GGE/JME	NM_018100 c.1057C>T	p.(R353W)	1.24×10^{4}	1.24×10^{4}	<0.001	0	GERP: 3.5 CADD: 27.3	MTR: 1.03 (T) REVEL: 0.31 (T) ClinPred: 0.13 (T) M-CAP: 0.02 (T)
E11	Chr16-29812989-G-A (homozygous)	*PRRT2*	BFIE, PKD	NM_145239 c.-65-1G>A	pLOF	0	0	0	0	GERP: 4.6 CADD: 31	rf: 0.94 (D) ada: 1.0 (D) MaxEnt: 8.75 (D) TraP: 0.96 (D)
	Chr10-63214651-G-C (heterozygous)	*JMJDC1*	NDD, ASD	NM_032776 c.1516C>G	p.(P506A)	0	0	0	0	GERP: 5.1 CADD: 23.3	MTR: 1.08 (T) REVEL: 0.23 (T) ClinPred: 0.99 (D) M-CAP: 0.01 (B)

A convergent squint (right side) was observed on neurological examination which was otherwise normal. Her younger brother also had an infantile epilepsy, with the first episode of seizures occurring on the first day of life (GTCS). No EEG was performed. He was described as inattentive and overactive and had a mild delay in achieving motor milestones. His neurological examination did not show additional findings. Both were on treatment with VPA (no seizures for more than one year) but compliance was poor (due to availability and cost). In addition to the VUS in *SPTAN1*, the proband carried another variant in *GRIN2B* (p.(His1145Arg)) that was not detected in her brother (Figure 2.3). These URVs were predicted to be deleterious, affecting amino acids that are conserved among orthologs. However, both variants were found in homozygous status in an unaffected individual (*SPTAN1* in the father and *GRIN2B* in a sibling without epilepsy).

Another VUS in *SCN3A* (p.(Ser495Arg)) was identified in the youngest of three siblings with epilepsy in Family E3 but did not segregate with the phenotype (Figure 2.4A). This deleterious URV affected an amino acid in the linker region between domains I and II of Nav1.3 that was conserved among genes encoding neuronal Nav1.x channels (*SCN1A*, *SCN2A*, *SCN8A*, *SCN9A*) but not the remaining Nav1.x encoding genes. Starting in early childhood (between 3 and 5 years), the proband carrying this variant (14 years old; male) and his elder brother (19 years old; not carrying the variant) developed what was described as jerking episodes during sleep, but the exact nature of these episodes was unclear. These were not reported in the eldest affected sibling (23 years old; female; not carrying the variant). The proband as well as this eldest sibling had otherwise a similar disease presentation as they developed frequent bilateral myoclonic seizures at the age of 13 years, occurring mostly early in the morning, and infrequent attacks of GTCS (eldest) or tonic (youngest). Their brother developed myoclonic seizures at the same age of 13 years, though occurring on the left side. He also had infrequent occurrences of GTCS as well as tonic seizures. All siblings had a normal developmental history. There were no additional symptoms or findings on neurological examination. No EEG was available. Following initial treatment in the eldest patient with CBZ, complete seizure control was achieved with add-on VPA therapy. The other siblings were on monotherapy with VPA (complete seizure control in the youngest, a few seizures precipitated by sleep deprivation in the other). Moreover, a rare missense variant in *EFHC1* (p.(Arg225Gln)), was identified in these three probands and was classified “Benign”. The variant was heterozygous in the three patients as well as their healthy parents, and homozygous in two elder siblings not diagnosed with epilepsy (Figure 2.4A). Although reported in individu-

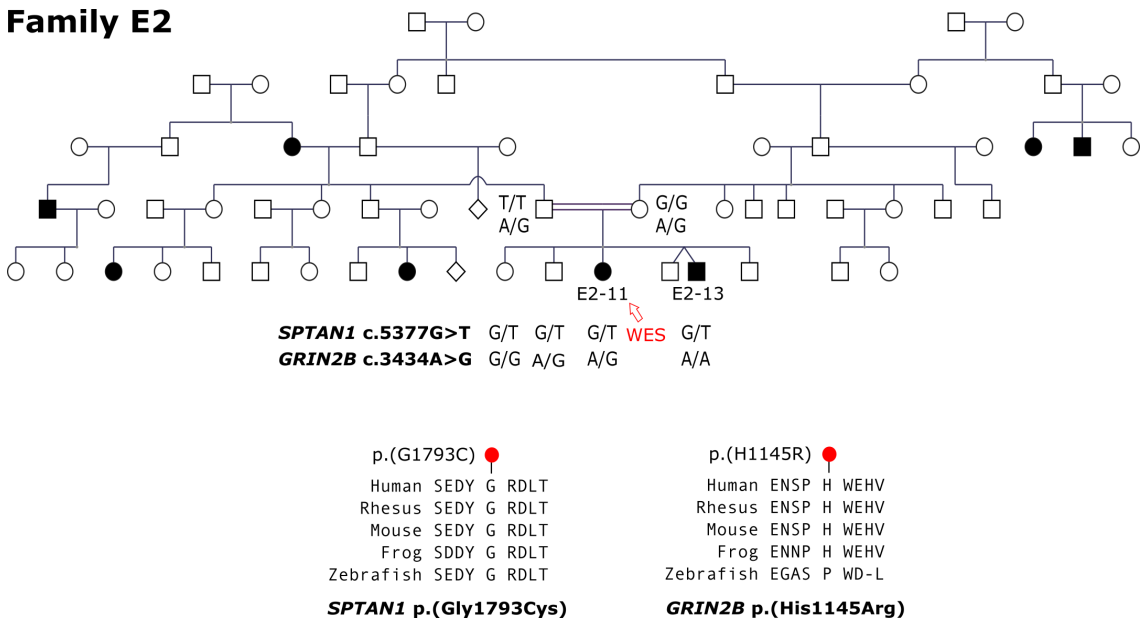


Figure 2.3: Variants of Uncertain Significance in dominant epilepsy genes in Family E2. Pedigree of Family E2 showing the segregation of two variants of uncertain significance (VUS). The individual identifiers of investigated patients are shown (see Table 2.1 for the phenotypes) and a single proband investigated using whole exome sequencing (WES) is indicated with a red arrow. The amino acid conservation around the affected sites in several ortholog genes is also shown.

als with epilepsy in ClinVar (classified as VUS), it was seen in three different population databases with frequencies < 0.1% (Table 2.2).

An additional rare missense variant in *EFHC1* (p.(Arg353Trp)) – that was also reported as VUS in ClinVar but seen in different population databases – was identified in one of two siblings diagnosed with epilepsy in Family E8 (Table 2.2). It was inherited from a healthy father (Figure 2.4B). We did not identify other variants with likely disease relevance or with uncertain significance in this family. However, the proband also carried a heterozygous URV in *DEPDC5* (p.(Ala530Val)). Both variants (in *EFHC1* and *DEPDC5*) did not segregate with the phenotype (were absent in an affected sibling) and were classified as benign and likely benign, respectively. The carrier (11 years old, female) was diagnosed with epilepsy at the age of 9 years, presenting with myoclonic and atonic seizures (less than 10 attacks in the month prior to ASM initiation). Her neurological examination was normal, and her EEG showed intermittent spikes and sharp waves over the right occipital area with bilateral spreading. At the time of sampling, she was on monotherapy with Levetiracetam (LEV) with reduction in seizure frequency (> 50%). Her elder brother (16 years), who did not carry the variants, was diagnosed with epilepsy at the age of 12 years, presenting with myoclonic seizures as well. His EEG at the time showed bilateral intermittent discharges of high amplitude spikes and sharp waves suggestive of generalized seizures. He was responsive to VPA monotherapy (> 50% reduction in seizure frequency). A second EEG (1 year later) showed epileptiform activity that is localized to the left temporal area with brief spreading. The ACMG/AMP criteria used for the classification of all variants are presented in Table 2.3.

2.5. Discussion

Here, we report findings from a family-based study in five consanguineous Sudanese families with genetic epilepsies. The nature of the genetic ancestry of these families offered insights into one of the understudied African populations and the current observations underscore the complexity of inheritance in common epilepsies. In this small cohort, we identified a pathogenic variant in *PRRT2* predisposing to one of the commonest "single-gene" epilepsies.[154] Pathogenic variants in *PRRT2* are frequently identified in individuals with familial infantile epilepsy, paroxysmal movement disorders with and without epilepsy, and less frequently in those with sporadic infantile epilepsies, hemiplegic migraine and episodic ataxia.[68,155–158] The epilepsy phenotype reported here was consistent with a diagnosis of self-

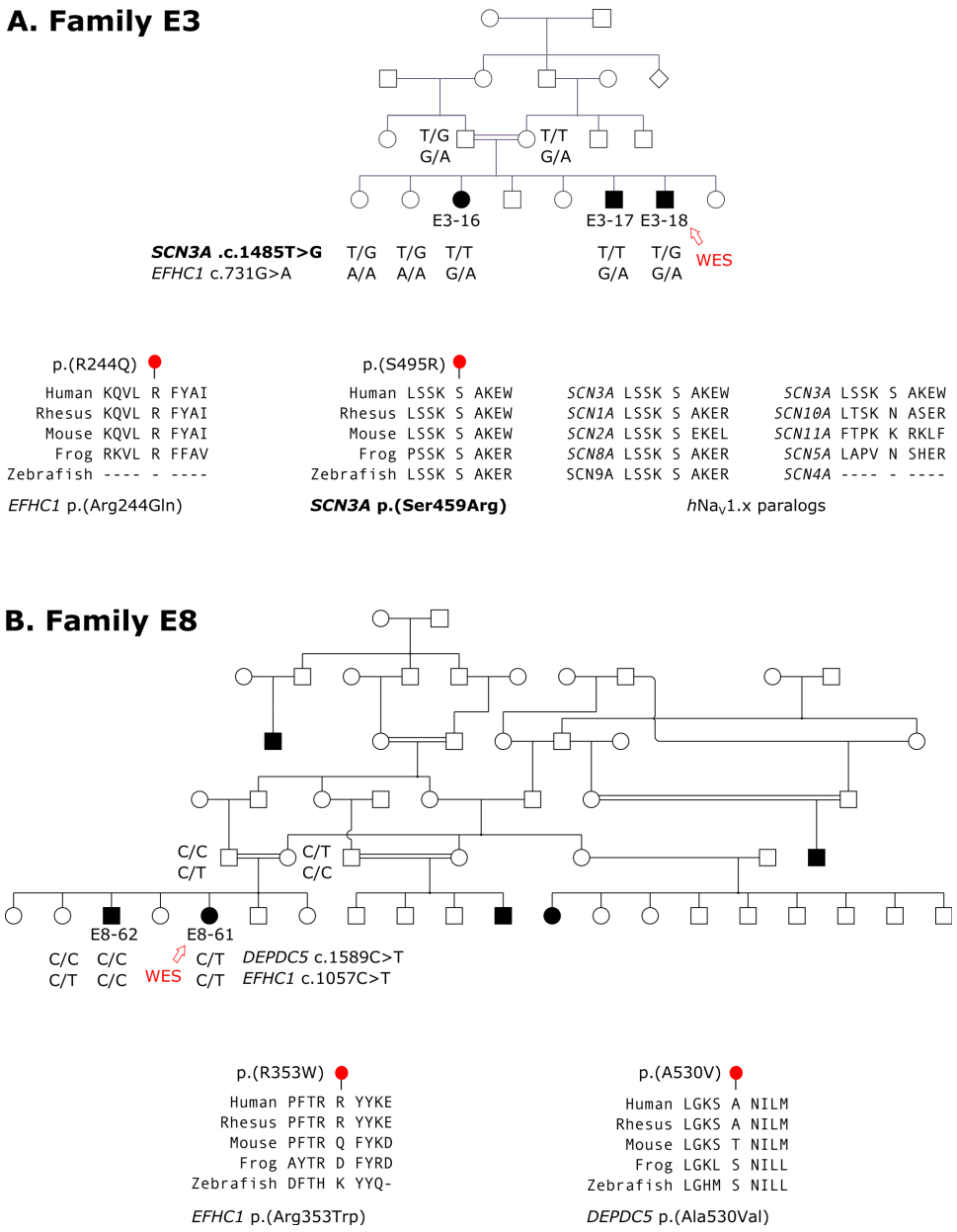


Figure 2.4: Benign variants in *EFHC1* identified in Families E3 and E8. A. Pedigree of Family E3 showing the segregation of a benign variant in *EFHC1* and a variant of uncertain significance (VUS) in *SCN3A*. **B.** Pedigree of Family E8 showing the segregation of two (likely) benign variants in *EFHC1* and *DEPDC5*. The amino acid conservation in orthologs in and around affected sites is shown for all variants. The conversation of an amino acid in *SCN3A* where a VUS was detected (in bold font) is shown additionally in paralogs. The individual identifiers of investigated patients are shown (see Table 2.1 for the phenotypes) and probands investigated using whole exome sequencing (WES) are indicated with a red arrow.

Table 2.3: Evaluation of the disease relevance of identified candidate variants.

The classification was based on the American College for Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) framework (see the methods for the explanation of the criteria). Strong computational evidence indicates pathogenic/deleterious prediction in more than 3/6 tools (see Table 2.2). URV: Ultra-rare variant (i.e., not seen in population controls; see Table 2.2).

FID	Gene	Transcript and Variant	Predicted change	URV	Computatio-nal evidence	Segreg-ation	ACMG/AMP criteria	Classification	Previously in ClinVar
E11	*PRRT2*	NM_145239:c.-65-1G>A	IVS1-1	Yes	Strong	Yes	PVS1, PM2, PP1M, PP3	Pathogenic	No
E5	*PCDH19*	NM_020766:c.1124A>T	p.(D375V)	Yes	Strong	No	PM1, PM2, PM5, PP3, BS4	Likely Pathogenic	No
E2	*SPTAN1*	NM_003127:c.5377G>T	p.(G1793C)	Yes	Strong	No	PM2, PP2, PP3, BS4	Uncertain Significance	No
E2	*GRIN2B*	NM_000834:c.3434A>G	p.(H1145R)	Yes	Strong	No	PM2, PP2, PP3, BS4	Uncertain Significance	No
E3	*SCN3A*	NM_006922:c.1485T>G	p.(S495R)	Yes	Strong	No	PM2, PP2, PP3, BS4	Uncertain Significance	No
E8	*DEPDC5*	NM_014662:c.1589C>T	p.(A530V)	Yes	Weak	No	PM2, PP3, BP1, BS4	Likely Benign	No
E11	*JMJD1C*	NM_032776:c.1516C>G	p.(P506A)	Yes	Weak	No	PM2, PB5, BS4	Likely Benign	No
E3	*EFHC1*	NM_018100:c.731G>A	p.(R244Q)	No	Strong	No	PP3, BS1, BS4	Benign	VCV000205388.5
E8	*EFHC1*	NM_018100:c.731G>A	p.(R353W)	No	Weak	No	PP3, BP6, BS1, BS4	Benign	VCV000205401.5

limited familial infantile epilepsy and thus concordant with known *PRRT2*-related presentations.[68,159] Bilateral tonic seizures as seen in three individuals from Family 11 are a known disease presentation.[160] Atonic seizures, seen additionally in two of the three patients, are also a known – though relatively infrequent – presentation.[68]

Whereas heterozygous *PRRT2* variants are detectable in around 80% of individuals with self-limited familial infantile epilepsy,[68,154] homozygous variants typically underlie a severe presentation with prolonged episodes of paroxysmal movement disorders, learning difficulties or developmental delay in addition to mild or severe epilepsy.[157,159–162] However, the presentation in the reported siblings from Family 11, who carried a bi-allelic splice-site variant, was rather mild in comparison and was reminiscent of presentations seen with heterozygous loss-of-function variants. For instance, Döring and colleagues reported a proband with *PRRT2*-related epilepsy with a similar presentation consisting of clusters of bilateral tonic seizures beginning at four months of age, who carried the most recurrent pathogenic frameshift variant c.649dupC, in a heterozygous state.[160] This could be explained by an arguably weak effect of the IVS1-1G>A change on splicing or by the existence of a mixture of multiple functional and non-functional aberrant transcripts (thus partial loss-of-function).[163] Other variants not detected through exome sequencing (e.g., other intronic and regulatory variants) may have additional effect on the splicing.[164] RNA samples were not available from our patients to examine these effects. Since the pathogenicity of canonical splice variants is widely accepted,[49] a mini-gene assay[165] was not used to evaluate this variant further (as it does not capture the *cis*- or *trans*- regulatory effects of other variants in the haplotype, nor the temporal and spatial tissue specific expression patterns).

Although we did not identify further homozygous variants with likely disease relevance in the remaining families, we detected other deleterious URVs in several genes causing or predisposing to dominant epilepsy syndromes or related neurodevelopmental disorders, a finding that is consistent with the high burden in damaging ultra-rare coding variation in common epilepsies.[33,78] Despite the strong computational evidence of pathogenicity, only one of these variants was classified as likely pathogenic (*PCDH19* p.(Asp375Val)), highlighting one of the challenges of rare variant evaluation in complex diseases. It was not feasible to study the segregation of this variant reliably as potential carriers were deceased (Figure 2.2B), but the phenotype associated with it was consistent with the wide presentations of *PCDH19*-related female-limited epilepsy.[166,167] It was in a mutational hotspot and a previous report implicated a *de novo* variant affecting the same amino acid (p.(Asp375Tyr)) in a DEE with a Dravet-like

presentation.[168] Charge neutralizing changes at calcium-coordinating residues in this region (e.g., Asp375) may decrease affinity for calcium and reduce protein instability.[169]

The contribution of other identified VUS (in *SPTAN1, GRIN2B* and *SCN3A*) to the disease is likely insignificant, as these genes typically underlie developmental phenotypes that encompass seizure disorders (DEEs or NDDs with epilepsy) but not common (complex) familial epilepsy syndromes.[170–172] Moreover, these variants either did not segregate in all individuals with epilepsy or were seen in homozygous state in other family members without epilepsy. Two rare *EFHC1* variants had limited evidence of pathogenicity, although these were seen previously reported in ClinVar. The relatively high population frequency of these variant in several population databases (Table 2.2) compared to what is typical of true epilepsy-related variants[64] is consistent with the evidence arguing against a strong role for *EFHC1* in predisposing to GGE syndromes.[101] Notably, we did not identify candidate variants in *ADGRV1,* a gene recently suggested to underlie predisposition to GGE syndromes in Sudanese families,[126] in this small cohort of only five families.

Although the investigated individuals had parents who are cousins, the inheritance pattern in most families, when examined over generations or in extended pedigrees, is not typical of bi-allelic/recessive inheritance. These pedigrees feature several individuals without consanguineous parents or affected parent-child pairs. Although genetic heterogeneity (several monogenic loci in one pedigree) cannot be ruled out, these pedigrees are rather consistent with complex (non-Mendelian) inheritance.[61] Restrictions in funding and accessibility to samples from extended families resulted in some limitations particularly in terms of assessing oligogenic modes of inheritance (e.g., through WES in several family members and comprehensive segregation analysis in extended families and multiple generations). Nonetheless, inspection of other pedigrees in similar studies shows similar observations and oligogenic inheritance was frequently suggested in these studies.[118,126,173]

In summary, we did not find sufficient evidence to indicate that recessive inheritance could explain the missing heritability in common epilepsies. However, we show that the phenotypes of homozygous *PRRT2* variants may include an isolated mild epilepsy without paroxysmal movement disorders or developmental aberrations. Several individuals had variants of uncertain significance in epilepsy-related genes. These findings highlight the complexity of epilepsy inheritance in consanguineous families.

Chapter 3: The association of coding variants with familial and sporadic generalized epilepsy

This chapter was adapted from: **Koko** *et al*. 2022. *Epilepsia*. PMID:35032048. See the statement of contributions at the end of this dissertation.

3.1 Summary

Background: We aimed to identify genes associated with genetic generalized epilepsy (GGEs) by combining large cohorts enriched with individuals with a positive family history. Additionally, we aimed to compare the association of genes independently with familial and sporadic GGE.

Methods: We performed a case-control whole exome sequencing study in unrelated individuals of European descent diagnosed with GGE (n = 2,203 before quality control (QC)) and ancestry matched controls who were previously recruited and sequenced through multiple international collaborations. The association of ultra-rare variants with epilepsy (URVs; in 18,834 protein coding genes) was examined in 1,928 individuals with GGEs (vs. 8,578 controls; after QC), then separately in 945 individuals from the same cohort who had GGE and a positive family history (vs. 8,626 controls; after QC), and finally in 1,005 individuals from our cohort with sporadic GGE (vs. 8,621 controls; after QC). We additionally examined the association of URVs with familial and sporadic GGE in two gene sets important for inhibitory signaling (19 genes encoding $GABA_A$ receptors, 113 genes representing the GABAergic pathway).

Results: *GABRG2* was associated with GGE ($p = 1.8 \times 10^{-5}$), approaching study-wide significance in familial GGEs ($p = 3.0 \times 10^{-6}$), whereas no gene approached significance in sporadic GGEs. Deleterious URVs in the most intolerant sub-genic regions in genes encoding $GABA_A$ receptors were associated with familial GGE (OR = 3.9, 95% CI = 1.9 – 7.8, FDR-adjusted p = 0.0024), whereas their association with sporadic GGEs had lower odds (OR = 3.1, 95% CI = 1.3 – 6.7, FDR-adjusted p = 0.022). URVs in GABAergic pathway genes were associated with familial GGE (OR=1.8, 95% CI = 1.3 – 2.5, FDR-adjusted p = 0.0024) but not with sporadic GGE (OR = 1.3, 95% CI = 0.9 – 1.9, FDR-adjusted p = 0.19).

Interpretation: URVs in *GABRG2* are an important risk factor for familial GGE. The association of gene sets of GABAergic signaling with familial GGE is more prominent than with sporadic GGE.

3.2 Background

The genetic risk factors of generalized epilepsies have proven challenging to decipher despite growing evidence supporting its existence from twin and family studies.[112,174] Both genome-wide association studies[45,60] and rare variant association studies[33,34,47,78] investigated increasingly larger cohorts of genetic generalized epilepsies (GGEs). These studies provided key insights into the heritability as well as the nature of variants, genes and gene sets underlying it, thus partially explaining a complex inheritance profile that likely spans ultra-rare coding variants (URVs),[33,34,78] common variants,[45,60,66] copy number alterations,[85,175,176] and repeat expansions.[81–83] Prior large-scale sequencing studies of individuals with familial GGEs failed to show statistically significant associations in single genes.[47,78] Familial non-acquired focal epilepsies (NAFEs) demonstrate a markedly higher burden of URVs compared to sporadic NAFEs.[78] This, however, was not investigated so far in GGE. Nonetheless, gene set burden analyses in these studies demonstrated that URVs in multiple phenotypically and biologically informed gene sets (e.g., dominant epilepsy and developmental epileptic and encephalopathy genes, genes encoding $GABA_A$ receptors) are associated with an increased risk of seizures. These patterns were later replicated in independent case-control studies of predominantly sporadic GGE cases, which found that – despite much larger cohorts – only a few single genes have approached study-wide significance.[33,34] Aiming to identify protein coding genes in which URVs are significantly associated with an increased risk of generalized epilepsy, we performed a combined analysis of multiple cohorts of individuals with GGE and ancestry-matched controls. To improve the power of genetic discovery, we enriched our analysis with individuals with a positive family history of the disease and examined this subset of familial GGEs separately. Consequently, we investigated individuals with sporadic GGE to understand if familial and sporadic GGE had different genetic architectures.

3.3 Methods

3.3.1 Overview of the study design

In this case-control rare-variant association study, we investigated the association of ultra-rare and rare genetic variants with epilepsy in individuals with a diagnosis of a GGE and matched controls of European descent. We jointly analyzed whole exome sequencing (WES) data from two independent datasets encompassing GGE patients previously studied by (1) the

Epi4K Consortium and the Epilepsy Phenome/Genome Project[78] (referred to hereafter, along with matched controls, as the first dataset) or (2) the Canadian Epilepsy Network (CENet) and the Epicure, EpiPGX, and EuroEPINOMICS-CoGIE Consortia[47] (referred to, with their matched controls, as the second dataset). Control cohorts were obtained for the first dataset from local collections available at the Institute for Genomic Medicine[34,177] (IGM) (New York, USA), and for the second dataset from controls available at the Luxembourg Centre for Systems Biology (LCSB) (Esch-sur-Alzette, Luxembourg) obtained from the database of Genotypes and Phenotypes[178] or the Epi25 Collaborative.[33]

Ethical approvals from Institutional Review Boards and relevant Ethics Committees and written informed consent procedures were previously obtained and detailed elsewhere.[47,78] The details of the recruitment or acquisition of analyzed case or control cohorts, diagnostic and inclusion criteria were also previously described.[33,47,78] Here, we intended primarily to identify genes significantly increasing the risk of GGE by combining these cohorts. To that aim, we analyzed data from 2,203 affected individuals (1,214 from the first dataset and 989 from the second dataset; before quality control). Subsequently, we examined the strength of the association separately in 1,035 individuals (659 from the first dataset; 376 from the second dataset) with a positive family history of epilepsy. Afterwards, we went on to assess the remaining 1,168 individuals (555 from the first dataset; 613 from the second dataset) without a family history or with an unknown family history status. The analysis strategy is outlined in Figure 3.1.

Exome sequencing data generation for the case and control cohorts was previously described. [33,47,78] In compliance with privacy regulations, the genotypes from the two datasets were processed in parallel at the IGM and the LCSB. A neural network predictive model was used to exclude individuals unlikely to be of a non-Finnish European descent. We removed one sample from each pair of duplicates/related individuals within each dataset and one sample from each pair of duplicates between the two datasets. We also performed quality control procedures to remove low quality samples/variants to harmonize the coverage and call rate between the cases and controls within each dataset. Contingent on case-control matching, the final number of cases or controls included in each analysis (*all, familial* and *sporadic GGEs* analyses) differed slightly across analyses (see results).

Figure 3.1: Flow chart summarizing the analysis strategy used in this study. IGM: Institute of Genomic Medicine, New York, USA. LCSB: Luxembourg Centre for Systems Bioscience, Esch-sur-Alzette, Luxembourg. DRAGEN: Dynamic Read Analysis for Genomic platform. GATK: Genome Analysis Toolkit. ATAV: Analysis Tool for Annotated Variants. ULHPC: University of Luxembourg High Performance Computing Cluster. GGE: Genetic Generalized Epilepsy. QC: Quality control. GQ: Genotype Quality. AD: Allele Depth. DP: Depth. VQSR: Variant Quality Score Recalibration. CCDS: Consensus Coding Sequence. CMH: Cochran Mantel Haenszel test. Details on ATAV: https://github.com/igm-team/atav. Details on ULHPC: https://hpc.uni.lu/.

3.3.2 Sequence data generation and quality control in the first dataset

Participants whose sequencing data formed the first dataset included 1,214 GGE patients recruited by the Epi4K Consortium and Epilepsy Phenome/Genome Project as previously described,[37,78] and sequenced at the IGM at Columbia University (New York, USA). The diagnosis of a GGE syndrome required the patients to have generalized epilepsy with absence, myoclonic or tonic-clonic seizures and generalized spike-and-wave discharge on EEG. To qualify for the familial analysis, patients were required to have at least one relative (up to the third degree) who had been diagnosed with epilepsy. Ancestry matched controls (*n* = 14,100 before quality control) were selected from multiple collections of control cohorts at the IGM.[34]

WES of DNA samples from participants forming the first dataset was performed at IGM and sequenced using Illumina's HiSeq 2000, HiSeq 2500 or NovaSeq 6000 platforms (Illumina, San Diego, CA, USA) following enrichment with Agilent All Exon Enrichment kits (Agilent Technologies, Santa Clara, CA, USA), NimbleGen SeqCap EZ Exome Enrichment kit (Roche NimbleGen, Madison, WI, USA), Twist Human Core Exome (Twist Bioscience, San Francisco, CA, USA) or IDT xGen Exome Research Panel (Integrated DNA Technologies, Coralville, IA, USA). The sequence data from all cases and controls were processed according to the IGM bioinformatics pipeline.[78,177] Sequencing reads were aligned to the human reference genome build 37 (GRCh37) using Illumina's Dynamic Read Analysis for GENomics (DRAGEN) Bio-IT Platform.[179,180] Picard (https://broadinstitute.github.io/picard/) and the Genome Analysis Tool Kit v3.6 (GATK[181]) were used to perform duplicate read marking, base quality scores recalibration, indel realignment and haplotype calling. The samples were processed individually at different time points and the variants obtained from single sample calling were imported and integrated in the Analysis Tool for Annotated Variants (ATAV) Database.[177]

For this study, samples with possible contamination (heterozygosity exceeding 2%) determined using VerifyBamID,[182] with discordance between self-declared and sequence-derived sex, or with low coverage (less than 85% of the consensus coding sequence[130] release 20 (CCDS r20) targets covered at a minimum of 10x) were removed. Related individuals were identified using Kinship-based Inference for GWAS[183] (KING). One of each pair that had an inferred relationship of third-degree or closer was dropped, preferentially retaining affected over control individuals and samples with higher coverage. EIGENSTRAT[184] was then used

to remove ethnicity outliers to minimize the effects of residual population stratification (Figure 3.2).

The following variant-level parameters (hard filters) were enforced: Quality/Depth (QD) > 5, Quality (QUAL) > 50, Mapping Quality (MQ) > 40, Strand Odds Ratio (SOR) < 3 (SNVs) or < 10 (indels), Fisher's Strand bias score (FS) < 60 (SNVs) or < 200 (indels), Read Position Rank Sum score (RPRS) < -3, and Mapping Quality Rank Sum score (MQRS) < -10. Variants were required to pass GATK Variant Quality Score Recalibration (VQSR) filter. Known artifacts and variants failing quality filters in population databases (Exome Variant Server, ExAC, gnomAD) were excluded. Low quality genotype calls with total allelic depth (DP) < 10 or genotype quality (GQ) < 20 were filtered. Heterozygous calls had a minimum alternate allele fraction (AD/DP) of 0.3. As previously described, a coverage harmonization procedure was employed to remove the variants that were differentially covered across the cases and controls.[78] Briefly, this was based on plotting the cumulative difference in site coverage between cases and controls to identify a filtering cut-off that will minimize this difference while allowing the largest possible number of variants to be retained.

3.3.3 Sequence data generation and quality control in the second dataset

The individuals with generalized epilepsy analyzed here were selected from 2,524 individuals recruited by the Epicure/EuroEPINOMICS-CoGIE Consortia, EpiPGX Consortium, and CENet as described previously.[47] For the purpose of this work, we used the sequence data of 989 individuals ascertained to have classical GGE phenotypes (childhood or juvenile absence epilepsy (CAE or JAE), juvenile myoclonic epilepsy (JME), or epilepsy with generalized tonic-clonic seizures alone (EGTCS)), early-onset absence epilepsy (age of onset < 3 years), epilepsy with myoclonic absences, or unclassified GGE. Familial cases had more than one self-reported affected first-degree relative. The controls for this dataset (n = 4,904 before quality control) were obtained from the database of Genotypes and Phenotypes[178] (dbGAP studies: MIGen Ottawa Heart Study controls, Rotterdam study controls and Alzheimer Disease Genetics Study controls) or from the Epi25 Collaborative.[33]

Sequencing of Epicure/EuroEPINOMICS-CoGIE cases was done with the Illumina HiSeq 2000 using NimbleGen SeqCap EZ Human Exome Library (NimbleGen, Madison, WI, USA) at Cologne Center for Genomics (Cologne, Germany). WES for the EpiPGX cohort was done at deCODE genetics (Reykjavik, Iceland) on the Illumina HiSeq 2500 with Nextera Rapid Capture Expanded Exome kit (Illumina, San Diego, CA, USA). WES for individuals recruited

by CENet was performed by the McGill University and Génome Québec Innovation Center (MAGQUIC, Québec, Canada) on Illumina HiSeq sequencing platforms using TruSeq or Roche Nimblegen EZ libraries. Controls from the Epi25 Collaborative were sequenced at the Broad Institute of Harvard and the Massachusetts Institute of Technology on the Illumina HiSeq platform using Illumina Nextera Rapid Capture or TruSeq Rapid Exome enrichment kit. The Rotterdam Study controls were sequenced using Illumina HiSeq 2000 by use of Roche Nimblegen EZ Human Exome Library. The Alzheimer study controls were sequenced over multiple time points at the Broad Institute using different capture kits. Fastq files were aligned to GRCh37 as previously described[47,185] and jointly called using DRAGEN Bio-IT Platform.[179,180]

The sample-level call rate, autosomal and chrX inbreeding coefficients were collected using Plink[186] 1.9. and Picard from GATK[129] v4. 1.4.1. Samples with phenotypes other than GGEs or without appropriate permissions for inclusion and samples with extremely low variant counts (< 10,000 non-missing calls) were removed. Samples with genotyping rates lower than 80%, outlier samples on autosomal heterozygosity (> 4 median absolute deviations on autosomal inbreeding coefficient estimates), and samples with discordant or ambiguous sequencing sex based on chromosome X inbreeding co-efficient estimates ($F < 0.3$ for female and $F > 0.7$ for male predicted sequencing sex) were excluded. The remaining samples were scanned for relatedness (third degree) using KING.[183] For duplicates and pairs with matching phenotypes, the sample with the higher genotyping rate was retained. Otherwise, cases were preferentially retained.

Next, multi-dimensional scaling was used to project the major continental ancestry of the study samples using "1000 Genomes" data (2,504 samples) using KING. The top principal components were visualized and used to classify the ancestry with a support vector machine (SVM) using R package *e1071*.[187] Samples with predicted European ancestry were retained. Following the baseline variant filtering steps outlined below, the variant calling metrics were re-examined to exclude any additional sample outliers. All samples with SNV counts < 15,000 were filtered (this removed all Rotterdam controls and most Alzheimer controls). Outliers beyond three standard deviations per cohort on key variant calling metrics (Heterozygous-Homozygous calls ratio, Transitions-Transversions ratio, and Insertions-Deletions ratio) were filtered. To ensure adequate case control matching and the removal of ancestry outliers, PCA analysis using EIGENSTRAT[184] was employed (Figure 3.2).

The variants were filtered for those located in the CCDS exonic coding regions (padded on each side to accommodate canonical splice sites and masked for low-complexity regions) using bcftools[188] v1.9. The variants were decomposed, normalized, and sorted using bcftools and vt[131] v0.5. Low quality genotypes were filtered by setting calls with total allelic depth < 10 or genotype quality < 20 to missing. Heterozygous calls had a minimum alternate allele depth fraction (AD/DP) 0.25. This genotype filtering was performed using bcftools. A combination of hard filtering and filtering based on recalibrated variant quality scores on was employed to remove low quality variants. Variant calls with low quality were filtered (SNVs: QUAL < 10, QD < 2, MQ < 30, FS > 60, MQRankSum < -12.5, RPRS < -8; Indels: QUAL < 10, QD < 2, RPRS < -20, FS > 200). Variant Quality Score Recalibration (VQSR) was performed on the normalized and genotype-filtered call set using GATK based on these annotations: QD, FS, SOR, MQRankSum, and RPRS. SNVs and Indels failing VQSR Tranche 99.0 filter were removed.

Since the datasets were sequenced using different capture kits, we performed additional harmonization steps to limit our analysis to the coding regions covered in all kits & to minimize the spurious effects caused differences in capture kits. Variants were retained only if they had genotyping rates ≥ 90% both in EpiPGX cases (largest case dataset; representing Illumina capture targets) and MIGen Ottawa controls (largest controls dataset; representing Agilent capture targets). After removal of sample outliers (see above), a final round of call rate harmonization was then performed where the variant call rate was calculated among the remaining cases and controls and variants with call rates below 95% in cases or controls were filtered. Also, the cumulative difference of call rate between cases and controls was plotted and 9.4% of the variants were removed to minimize this difference while retaining the largest possible number of variants.

3.3.4 Duplicates and ancestry harmonization across cohorts

To maximize the ancestry matching between the two analyzed patient cohorts, the ancestry prediction among the cases was harmonized in our two cases datasets by using the same ancestry prediction model to ensure homogeneity in ancestry assignment. Principal components analysis was performed on genotypes of previously defined well covered exonic autosomal polymorphic markers.[78]

Figure 3.2: Principal Component Analysis for ancestry matching. The plot shows eigenvectors on the first and second principal components from 1055 individuals with GGE vs. 6814 controls from the first dataset and 829 individuals with GGE vs. 1764 controls from the second dataset.

A neural network model that uses the first five principal component axes as the independent variables, trained on more than two thousand individuals with pre-evaluated genetic ancestry from six ethnic groups (European, Middle Eastern, Hispanic, East Asian, South Asian, and African), was then used to predict the probability of a European ancestry. Cases with < 95% probability were excluded.

To exclude likely duplicates between the two case cohorts, a genotype hashing approach adopted from the Gencrypt method described[189] was used to avoid the need for genotype sharing across the two study sites. A group of variants with minor allele frequency > 0.1 and genotyping rate > 98% in both cohorts was identified. From this pool, two hundred sets were created, each consisting of randomly selected non-overlapping 150 SNPs. For each sample, the genotypes over each set were concatenated keeping their order and converted to *sha256* cryptographic hashes. The hashes were exchanged and compared between cohorts. In total, 57 cases shared one or more hashes (according, likely to have identical genotypes in ≥ 150 randomly selected polymorphic markers) were considered possible duplicates. These were retained only in the first dataset and were removed from the second set.

3.3.5 Variant annotations

The analysis was limited to coding variants located in the exons of 18,834 protein-coding genes from CCDS r20, extended with two bases on each side to accommodate canonical intronic splice sites. Variant effects were annotated using ClinEff[190] v1.0c. Population allele frequencies were estimated from gnomAD[103] r2.1 and DiscovEHR[135] database v1. Since a portion of our control samples overlapped with gnomAD exomes, gnomAD allele frequencies were based on gnomAD genomes. Missense variants were further annotated with three *in silico* deleteriousness and intolerance scores (selected based on our previous work[34,78]): Polyphen2 (PPh2) Human Diversity based score,[191] the Rare Exome Variant Ensemble Learner (REVEL) score[139] and the Missense Tolerance Ratio (MTR) score.[138] The population allele frequencies and *in silico* missense deleteriousness and intolerance scores were annotated for the first dataset (and its matched controls) using ATAV[177] and for the second dataset (and its matched controls) using Annovar[192] and bcftools.[188]

3.3.6 Qualifying variants' distribution and Quantile-Quantile (QQ) plots

To ensure that we achieved an adequate case control matching and coverage/call rate harmonization in each dataset, we examined the distribution plots of qualifying variants tallies. Variant tallies were examined separately for each study dataset and collectively for the final merged dataset. The significance of the differences in the distribution density of ultra-rare

synonymous variants was examined using Wilcoxon Rank Sum test with continuity correction as implemented in R (Figure 3.3).

The QQ plots for the combined analysis (see *gene-level association analysis* below) show the *p* values from those genes with at least one qualifying variant in the joint case cohort and the expected *p* values from a uniform distribution. The negative $\log_{10}$ of the observed *p* values was plotted against the negative $\log_{10}$ of an equal number of uniformly distributed *p* values ($-log_{10}((k-0.5)/n)$, where *k* is the gene rank and *n* the total genes). The confidence intervals for the expected *p* values were based on values drawn from a beta distribution ($-log_{10}(qbeta(\ /2, k, n-k))$ and $-log_{10}(qbeta(1-\ /2, k, n-k))$, where = 0.05 for a 95% confidence interval) using the *stats* package[193] in R 3.3. The genomic inflation factors ( ) was calculated using the regression method implemented in the function *estlambda2()* from R package *QQperm*.[194]

3.3.7 Analysis models

We defined three primary analysis models to examine the association of functional coding variation with GGE, based on a combination of three filtering criteria: minor allele frequency, variant types (effects), and *in silico* predictions (specifically for missense variants). We targeted URVs which we defined as those with a minor allele frequency (MAF) < 0.05% in our test datasets (internal MAF) and not seen in independent gnomAD & DiscovEHR population reference datasets (external MAF). Functional variants (i.e., presumed to affect the function of protein coding gene products) included those with predicted Loss-of-Function (pLoF: canonical splice-site, stop-gain & frameshift variants), in-frame insertions and deletions and missense variants. For each of the three models, missense variants were filtered further based on their expected (*in silico*) deleteriousness predicted using PPh2, REVEL or based on REVEL in combination MTR to capture the degree of sub-genic intolerance of the affected site. The analysis model targeting functional URVs while limiting missense variants to those with damaging PPh2 prediction is like the prior model we used to analyze a subset of our samples,[78] thus allowing for comparisons of outcomes with the increase in sample size. The latter approaches based on REVEL & MTR (i.e., analysis of deleterious variants identified with an ensemble method designed for rare variants in combination with sub-genic intolerance limiting) were recently shown to improve pathogenicity prediction in epilepsy.[34,138] A control model targeting synonymous URVs presumed to have a neutral effect was used to assess potential biases in cases vs. controls comparisons that are unlikely to be unrelated to disease risk. We supplemented our primary analyses with additional secondary models to examine the

association of (i) rare functional variants (defined as those with both internal and external MAFs lower than 0.1%) with and without URVs and (ii) pLoF variants without other types of functional variants (as these represent a class of high effect variants). Altogether, eight models were investigated (one control model, three primary models, and five secondary models) as summarized in Table 3.1.

3.3.8 Gene-level associations

As adopted in our previous studies,[78] we performed gene collapsing analyses by assigning a 1 or 0 indicator in a *gene by sample* matrix to indicate the presence or absence (respectively) of qualifying variants. Qualifying variants (QVs) were defined as variants matching the criteria for each analysis model in each gene and study individual (assuming dominant inheritance). The collapsing analysis was performed separately in our two independent study datasets and a Cochran–Mantel–Haenszel exact test (CMH) was then used to quantify the gene-level association between case status and QV carrier status by comparing the counts of cases and controls with QVs in the two datasets while accounting for cohort stratification.[34] Separate comparisons were performed for *all, familial* and *sporadic* GGEs, each against their ancestry-matched controls. We adopted a Bonferroni multiple testing correction for gene-level *p* values (= 0.05) accounting for three phenotypic groups, three primary analysis models and 18,834 protein-coding genes with a study-wide significance cut-off of 2.9×10^{-7}. The homogeneity in the observed odds between the two data sets was examined using Breslow-Day and Woolf tests. The genomic inflation factor ( ) was estimated as detailed in the supplements. The collapsing and subsequent joint statistical analyses were performed using ATAV[177] or R *data.table*,[195] R *tidyverse,*[196] and R *stats* on R[193] v3.3.

3.3.9 Gene set association analyses

We also studied two gene sets that are important for inhibitory signaling in which GGEs had previously shown an increased burden of deleterious URVs. This association was established in a subset of our current samples[47] and was later validated in additional datasets.[33] However, a stratified analysis based on family history was not performed in our previous work. Here, we examined the association of URVs in these gene sets with familial GGEs (vs. controls), with sporadic GGEs (vs. controls), and directly between individuals with familial GGE vs. those with a sporadic GGE. We complemented these comparisons with an analysis of all GGEs vs. controls (as a positive control). To measure the association, we did gene set collapsing analyses by collapsing QVs across all genes in the investigated gene set (i.e., a case/

Table 3.1: Overview of association analysis models.

MAF: Minor Allele Frequency. PPh2: Polyphen 2 Human Diversity based prediction. REVEL: Rare Exome Variant Ensemble Learner. MTR: Missense Tolerance Ratio score. pLoF: predicted loss-of-function variants. pLoF variants included stop-gain & stop-loss variants, frameshift insertions & deletions, and canonical splice-site variants. Functional variants included pLoF, in-frame insertions & deletions, and missense variants (the missense variants were filtered using PPh2, REVEL, and MTR predictions as indicated). MAF from gnomAD were based on the 'genomes' subset. The cut-offs for REVEL & MTR scores were based on Ref.[34]

Models	**Primary models**				**Secondary models**				
	Control	**Ultra-rare functional variants**			**Rare functional variants**		**Loss of Function variants**		
	Synonymous	**PPh2**	**REVEL**	**MTR**					
Minor Allele Frequencies									
Internal MAF	<0.0005	<0.0005	<0.0005	<0.0005	<0.001	without URVs	<0.001	without URVs	<0.0005
DiscovEHR MAF	0	0	0	0	<0.001		<0.001		0
gnomAD r2 MAF	0	0	0	0	<0.001		<0.001		0
Classes of Variants									
ClinEff Effects	Synonymous	Functional	Functional	Functional	Functional	Functional	pLoF	pLoF	pLoF
Missense variants filters									
PPh2 prediction	-	"Probably"	-	-	"Probably"	"Probably"	-	-	-
REVEL score	-	-	≥ 0.5	≥ 0.5	-	-	-	-	-
MTR score	-	-	-	≤ 0.78	-	-	-	-	-

control was a carrier if they harbored a QV in any gene in the gene set) followed by the CMH test. *P* values from the analyses of functional variants were adjusted for twenty-four multiple tests (four phenotypic comparisons, three URV analysis models, and two gene sets) using a Benjamini-Hochberg False Discovery Rate (FDR) procedure to maximize the power (as opposed to Familywise Error Rate control). In addition to the primary gene set association testing, we did further secondary analyses to explore the relative contribution of single genes to the overall association seen in gene sets of interest. We tested the association in a stepwise manner, starting at the top ranked gene and then adding one gene at a time from a ranked list of genes forming the gene set (ranked based on their gene level association) until reaching the complete set. As the change in the direction of effect was the main outcome we intended to investigate, the outcomes from these secondary analyses were not corrected for multiple testing. The list of genes composing the two gene sets of inhibitory signaling tested in this study were obtained from previously published work.[47]

Genes encoding $GABA_A$ receptors: *GABRA1, GABRA2, GABRA3, GABRA4, GABRA5, GABRA6, GABRB1, GABRB2, GABRB3, GABRD, GABRE, GABRG1, GABRG2, GABRG3, GABRP, GABRQ, GABRR1, GABRR2, GABRR3*. **GABAergic pathway genes**: $GABA_A$ and *ABAT, ADCY1, ADCY2, ADCY3, ADCY4, ADCY5, ADCY6, ADCY7, ADCY8, ADCY9, ANK2, ANK3, ARHGEF9, DISC1, DLC1, DNAI1, FGF13, GABARAP, GABARAPL1, GABARAPL2, GABBR1, GABBR2, GAD1, GAD2, GLS, GLS2, GLUL, GNAI1, GNAI2, GNAI3, GNAO1, GNB1, GNB2, GNB3, GNB4, GNB5, GNG10, GNG11, GNG12, GNG13, GNG2, GNG3, GNG4, GNG5, GNG7, GNG8, GNGT1, GNGT2, GPHN, HAP1, KCNB2, KCNC1, KCNC2, KCNC3, KCNJ6, KIF5A, KIF5B, KIF5C, MAGI1, MKLN1, MTOR, MYO5A, NLGN2, NRXN1, NSF, PFN1, PLCL1, PRKACA, PRKACB, PRKACG, PRKCA, PRKCB, PRKCG, RDX, SCN1A, SCN1B, SCN2B, SCN3A, SCN8A, SEMA4D, SLC12A2, SLC12A5, SLC32A1, SLC38A1, SLC38A2, SLC38A3, SLC38A5, SLC6A1, SLC6A11, SLC6A13, SRC, STARD13, TRAK1, TRAK2*.

3.3.10 Overrepresentation of known disease genes among top-ranked genes

The Online Mendelian Inheritance in Man (OMIM, https://www.omim.org/) database was used to obtain a list of genes associated with susceptibility to GGE (IGE, CAE & JME; phenotypic series: PS600669, PS254770 and PS600131) or causing Developmental and Epileptic Encephalopathies (DEE; phenotypic series: PS308350). A hypergeometric test was employed to examine the probability that *n* genes from a gene set of *N* genes appear by chance

among the top-ranked *k* genes when examining a total of 18,834 protein coding genes. The enrichment was tested at each rank *k* occupied by a gene from the gene set using a *phyper* function from R *stats* package as follows: *phyper*(*n*-1, *N*, 18834-*N*, *k*, lower.tail= FALSE). No correction of multiple testing was performed as the direction of change was the main outcome we intended to test.

3.4 Results

We studied the association of coding URVs with generalized epilepsy in a cohort of 1,928 unrelated individuals diagnosed with GGE and 8,578 matched controls of European descent, and then in a subset of 945 individuals diagnosed with a familial GGE (studied against 8,626 matched controls). Afterwards, we proceeded to compare the observed outcomes to those seen in 1,005 individuals with a diagnosis of a sporadic GGE (studied against 8,621 matched controls; all counts after quality control). The sample counts from the two study cohorts are detailed in Table 3.2. The total number of samples in the analysis of all GGEs was slightly lower than the sum of familial and sporadic GGEs as few samples were removed during the case-control matching process. We did not detect a prominent deviation of observed *p* values from expected *p* values in synonymous variant association testing (= 0.86 – 1.06, Figure 3.3) indicating adequate population substructure matching between individuals with epilepsy (cases) and without epilepsy (controls).

Table 3.2: Numbers of analyzed samples from the study cohorts.

Datasets		Analysis		
		All GGEs	**Positive Family History**	**Negative Family History**
First	Individuals with epilepsy	1,099	629	492
	Controls	6,814	6,862	6,857
Second	Individuals with epilepsy	829	316	513
	Controls	1,764	1,764	1,764
Total	Individuals with epilepsy	1,928	945	1,005
	Controls	8,578	8,626	8,621

3.4.1 *GABRG2* is the top-ranked gene associated with GGE

GABRG2 (MIM: 137164) was the top-ranked gene in the analysis of all GGEs, showing prominent association in two primary models: the PPh2 model (p = 1.8x10^{-5}) examining the association of functional URVs while filtering missense variants based on a damaging Polyphen2 prediction (Figure 3.4) and MTR model (p = 1.2x10^{-5}) combining sub-genic intolerance with REVEL (Figure 3.5). Limiting the cases to individuals with a family history of epilepsy strengthened the association with *GABRG2* in the PPh2 model (p = 3.0x10^{-6}). Using REVEL combined with MTR did not outperform the PPh2 model in terms of significance in the analysis of familial GGEs (p = 1.4x10^{-5}). Nonetheless, it maximized the separation between cases and controls, resulting in higher odds (Table 3.3) by preferentially filtering all *GABRG2* variants seen in our control sets – in line with recent findings suggesting that sub-genic intolerance filtering might be particularly effective for analyses geared towards specificity as opposed to sensitivity.[34] The analysis of sporadic GGEs was generally unremarkable for *GABRG2* (p = 0.15 – 0.015) and the top-ranked genes did not include biologically meaningful candidates (Table 3.3). In general, secondary analyses of rare functional and pLoF variants neither captured significantly associated single genes nor strong novel candidates with biological relevance (Tables 3.4 – 3.5 and Figures 3.6 – 3.8).

3.4.2 *GABRG2* qualifying variants

Most URVs in *GABRG2* were missense and few were recurrent whereas rare variant analyses (up to a MAF of 0.1%) resulted in the inclusion of additional *GABRG2* variants exclusively in the control cohorts (Table 3.6). A variant disrupting a canonical intronic splice donor site (IV6SD) detected in this study was previously associated with familial CAE and febrile seizures,[197] phenotypes that were also prominent in earlier *GABRG2* families featuring an overlap of GGE and Generalized Epilepsy with Febrile Seizures Plus[198–200] (GEFS+). Here, absence epilepsy with no history of febrile seizures or an affected family member was reported. A second variant seen in an individual with familial GGE (p.Met199Val) segregated in a previous study[201] with a phenotype of GEFS+ and was also reported in an individual with NAFE in the first Epi25 Collaborative study.[33] Sample overlap or relatedness to these previously reported individuals was not investigated (genetically) but it is unlikely based on our patients' clinical and family histories. Lastly, a variant seen in the familial epilepsy cohort (p.Arg177Pro) affected a codon for which an identical change (p.Arg177Gly) was seen previously in a family with febrile seizures.[202] The allelic origin of these variants (*de novo* vs. inherited) was not validated.

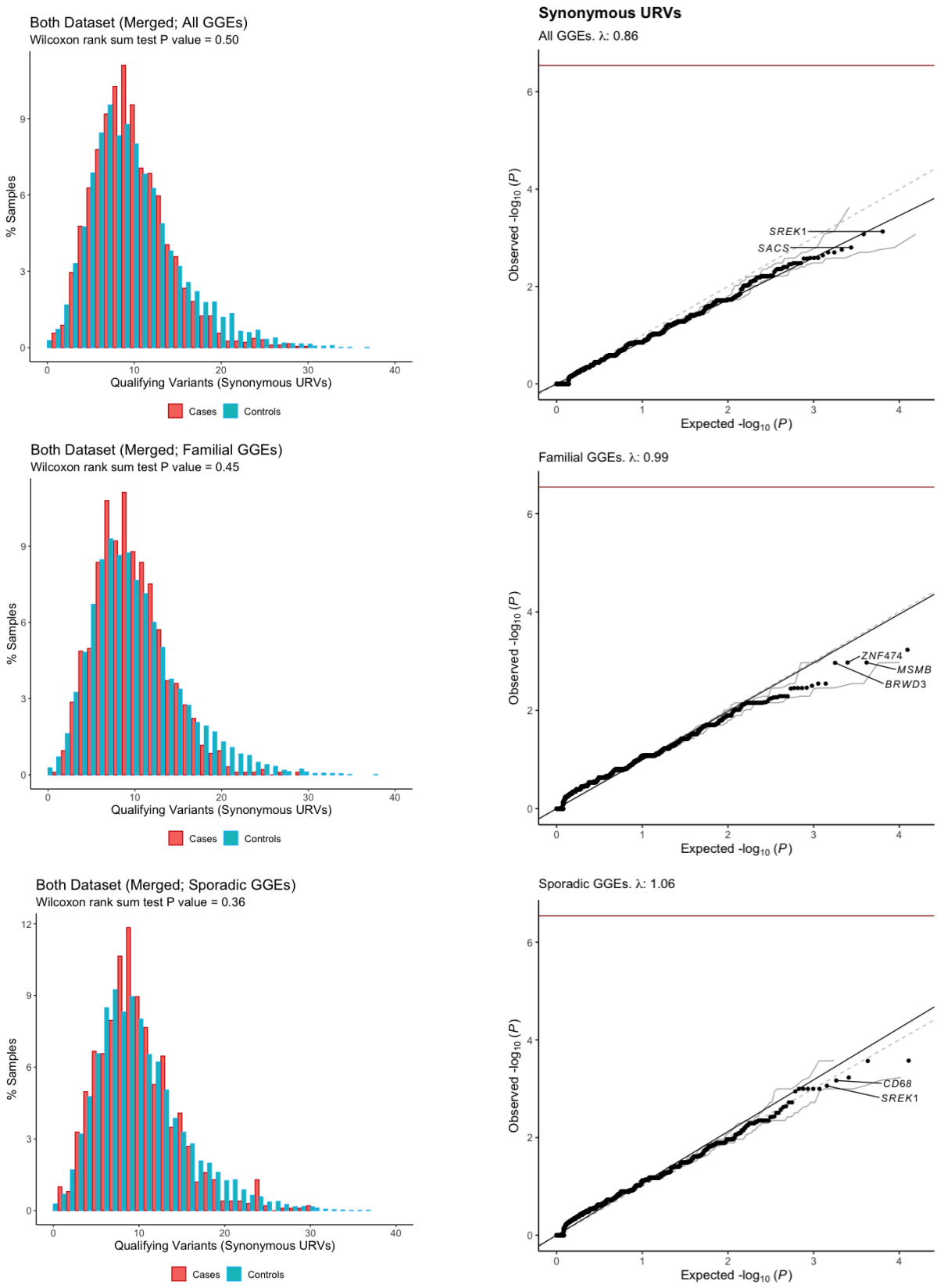


Figure 3.3: Balance of ultra-rare synonymous qualifying variants tallies between cases and controls in the total dataset. The QQ plots show the negative $\log_{10}$ of observed *p* values *vs.* the expected *p* values from a uniform distribution. *P* values were obtained from a two-sided Cochran Mantel Haenszel exact test of the association of ultra-rare synonymous qualifying variants. The 95% confidence intervals are shown as grey solid lines. The slope of the solid black line indicates the genomic inflation factor, whereas the slope of the dotted line equals 1.

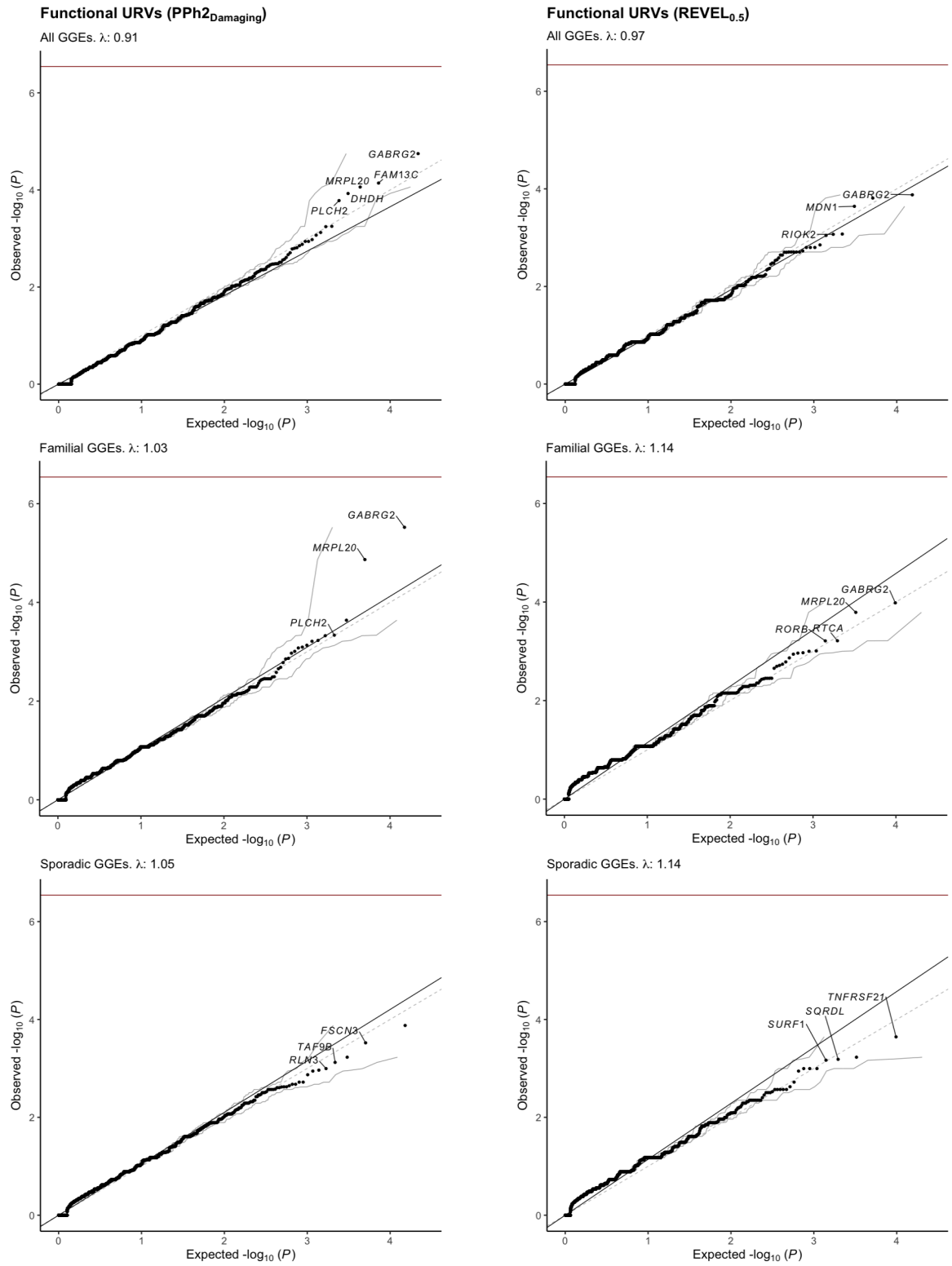


Figure 3.4: Association of ultra-rare deleterious variants with genetic generalized epilepsy. The quantile-quantile plots compare observed *p* values (Cochran-Mantel-Haenszel exact test) and expected *p* values (drawn from a uniform distribution) in analyses of 1,928 individuals with genetic generalized epilepsy (GGEs) vs. 8,578 controls and subsets of familial GGEs (945 cases vs. 8,626 controls) or sporadic GGEs (1,005 cases vs. 8,621 controls). The 95% confidence intervals are shown as grey solid lines. The slope of the solid black line indicates the genomic inflation factor, whereas the slope of the dotted line equals 1. Labels: genes that are enriched in cases in both datasets among the five top-raking genes. Exome-wide significance after Bonferroni correction (dark red line) was defined by a *p* value < 2.9×10^{-7}.

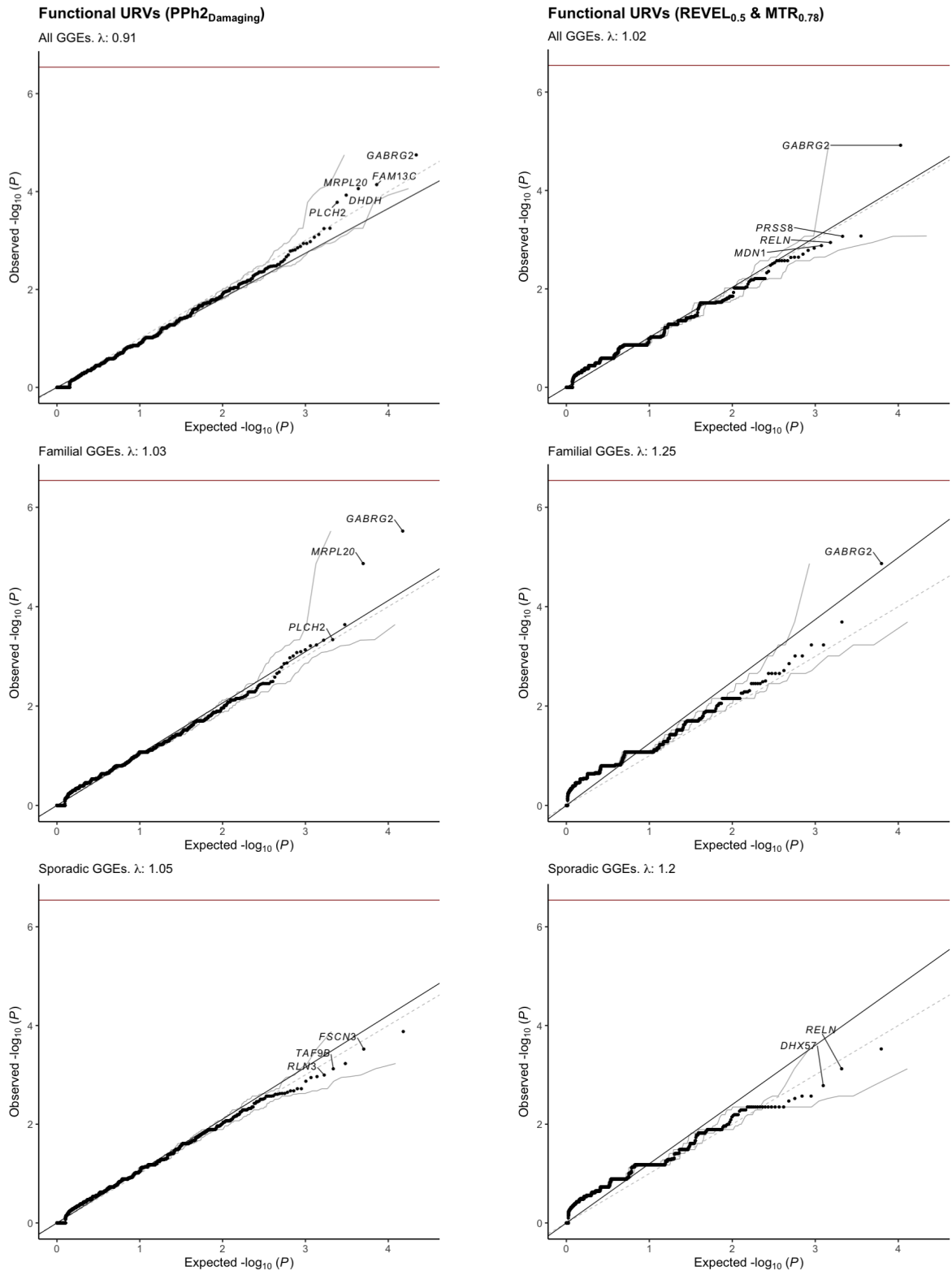


Figure 3.5: Association of ultra-rare deleterious and intolerant variants with genetic generalized epilepsy. The quantile-quantile plots compare observed *p* values (Cochran-Mantel-Haenszel exact test) and expected *p* values (drawn from a uniform distribution) in analyses of 1,928 individuals with genetic generalized epilepsy (GGEs) vs. 8,578 controls and subsets of familial GGEs (945 cases vs. 8,626 controls) or sporadic GGEs (1,005 cases vs. 8,621 controls). The 95% confidence intervals are shown as grey solid lines. The slope of the solid black line indicates the genomic inflation factor, whereas the slope of the dotted line equals 1. Labels: genes that are enriched in cases in both datasets among the five top-raking genes. Exome-wide significance after Bonferroni correction (dark red line) was defined by a *p* value < 2.9×10^{-7}.

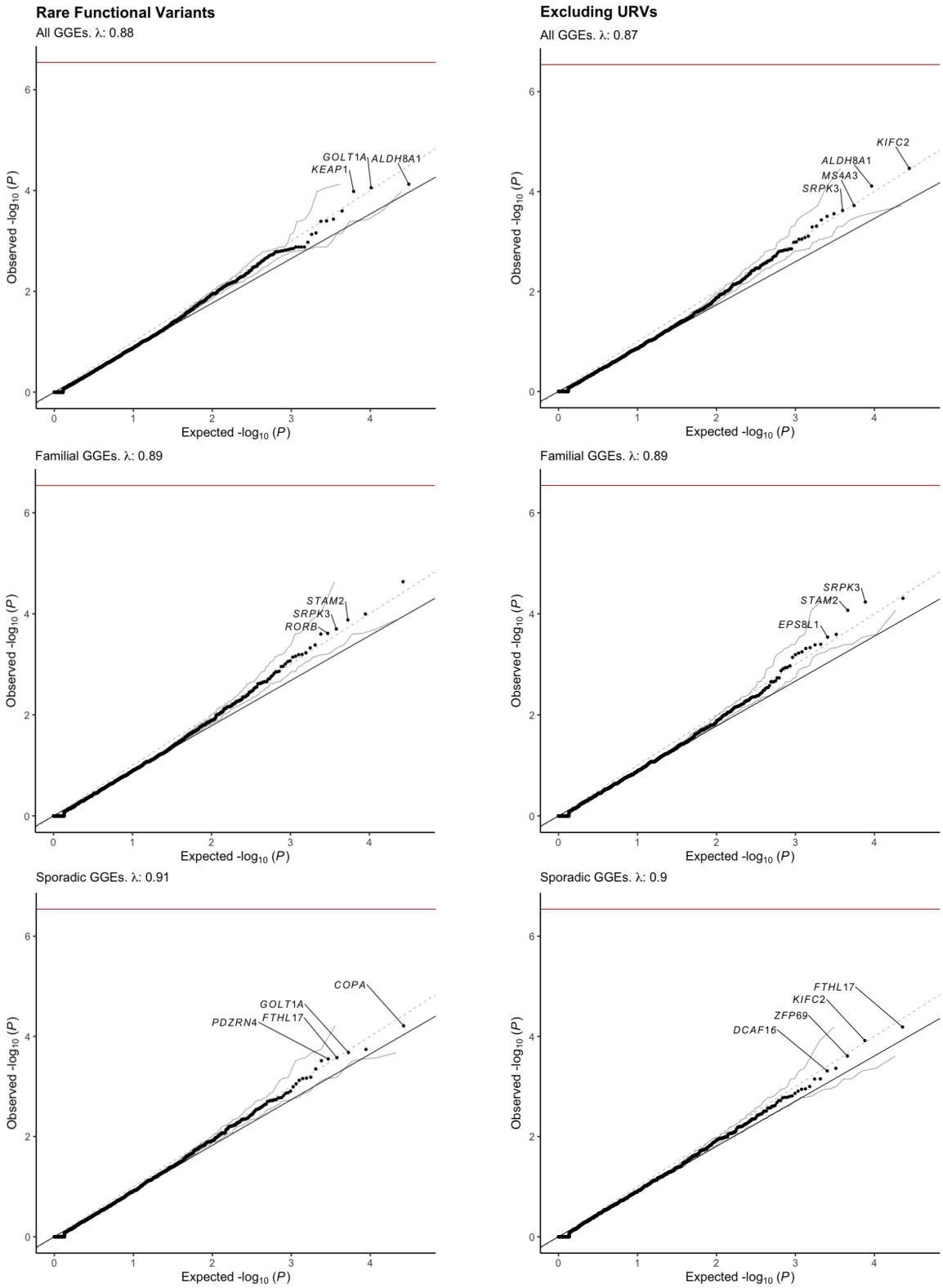


Figure 3.6: Association of rare deleterious variants with genetic generalized epilepsy. The quantile-quantile plots compare observed *p* values (Cochran-Mantel-Haenszel exact test) and expected *p* values (drawn from a uniform distribution) in analyses of 1,928 individuals with genetic generalized epilepsy (GGEs) vs. 8,578 controls and subsets of familial GGEs (945 cases vs. 8,626 controls) or sporadic GGEs (1,005 cases vs. 8,621 controls). The 95% confidence intervals are shown as grey solid lines. The slope of the solid black line indicates the genomic inflation factor, whereas the slope of the dotted line equals 1. Labels: genes that are enriched in cases in both datasets among the five top-raking genes. Exome-wide significance after Bonferroni correction (dark red line) was defined by a *p* value < 2.9×10^{-7}.

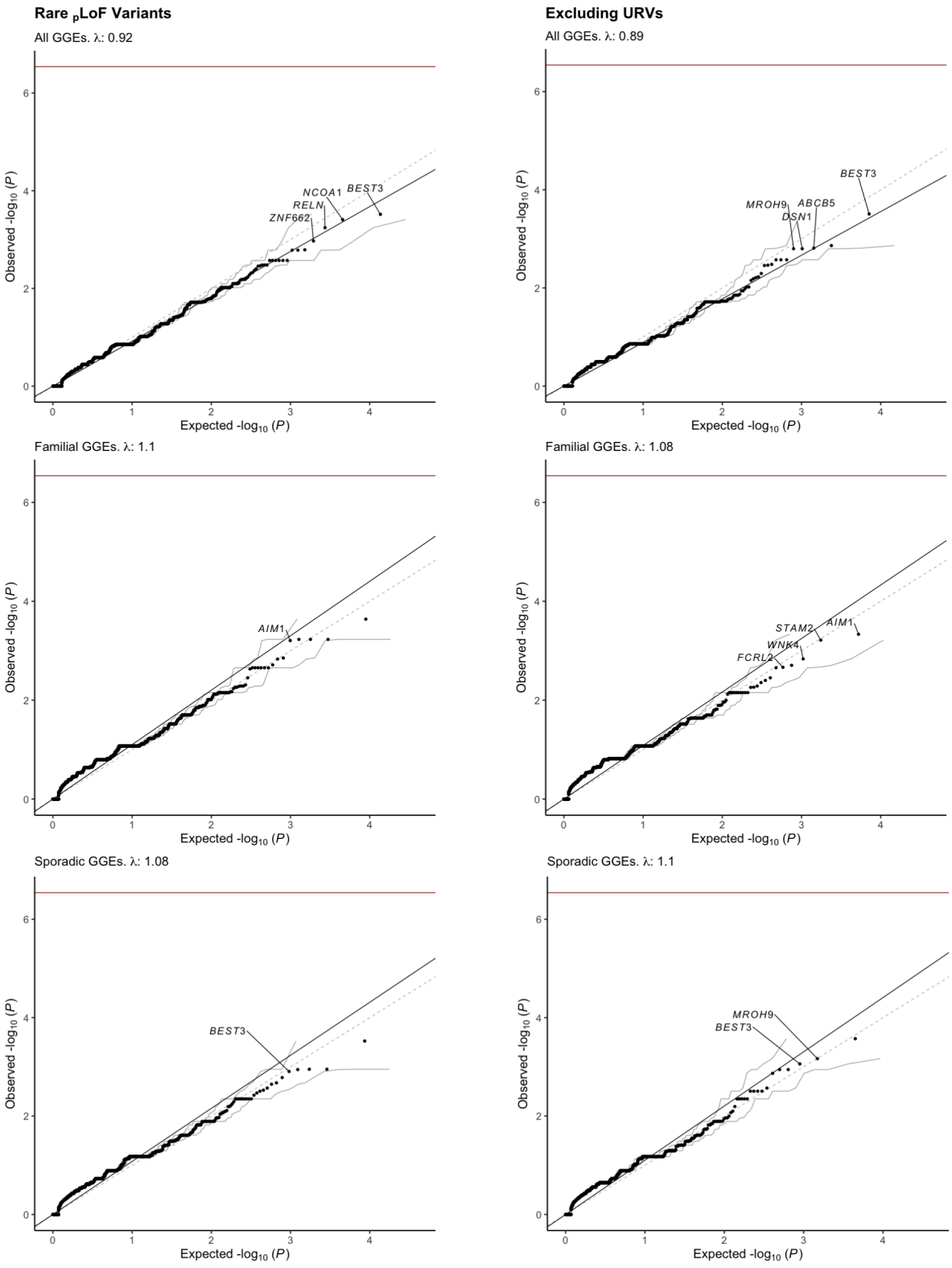


Figure 3.7: Association of rare predicted loss of function variants with genetic generalized epilepsy. The quantile-quantile plots compare observed *p* values (Cochran-Mantel-Haenszel exact test) and expected *p* values (drawn from a uniform distribution) in analyses of 1,928 individuals with genetic generalized epilepsy (GGEs) vs. 8,578 controls and subsets of familial GGEs (945 cases vs. 8,626 controls) or sporadic GGEs (1,005 cases vs. 8,621 controls). The 95% confidence intervals are shown as grey solid lines. The slope of the solid black line indicates the genomic inflation factor, whereas the slope of the dotted line equals 1. Labels: genes that are enriched in cases in both datasets among the five top-raking genes. Exome-wide significance after Bonferroni correction (dark red line) was defined by a *p* value < 2.9×10^{-7}.

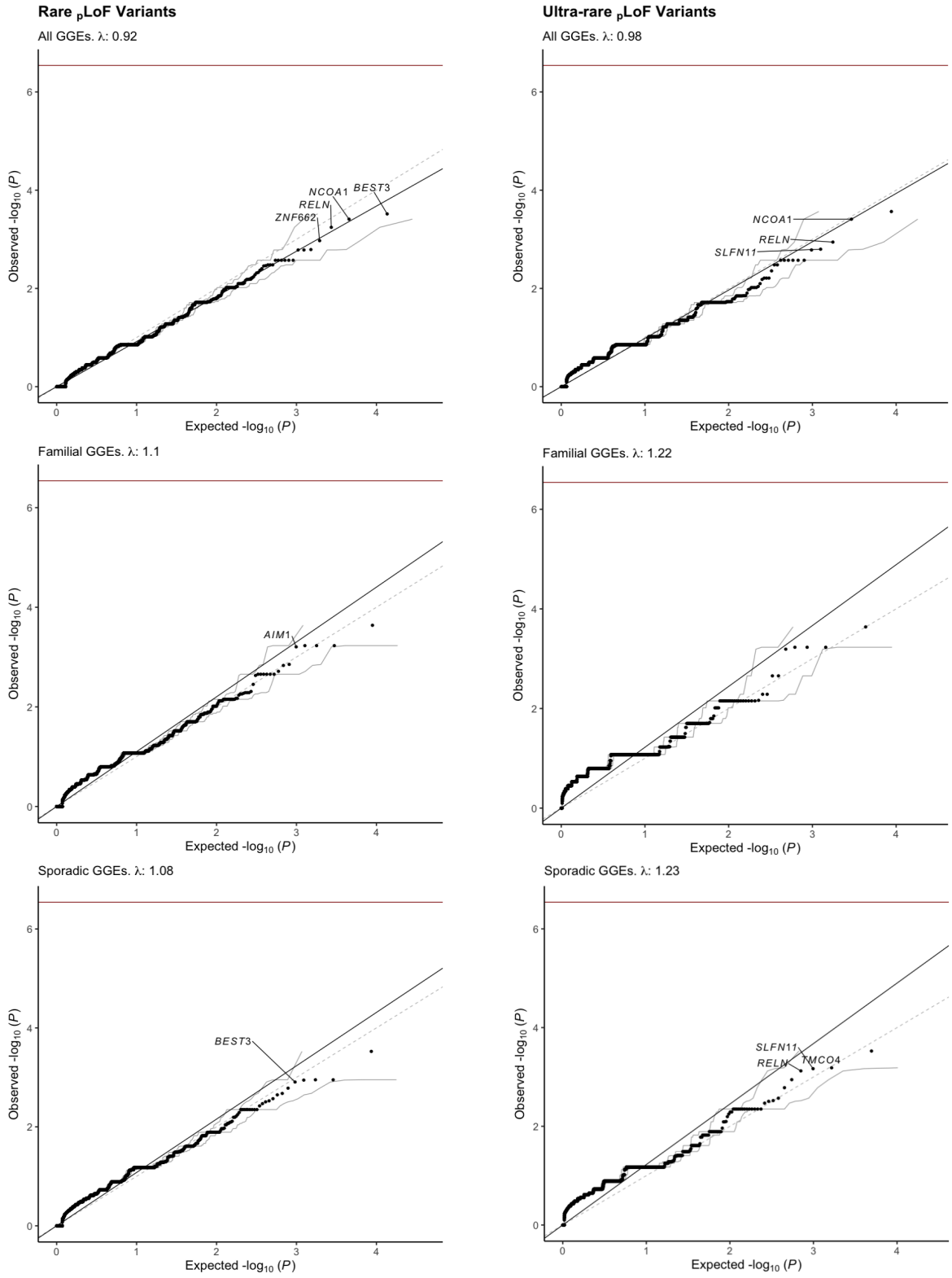


Figure 3.8: Association of rare predicted loss of function variants with genetic generalized epilepsy. The quantile-quantile plots compare observed *p* values (Cochran-Mantel-Haenszel exact test) and expected *p* values (drawn from a uniform distribution) in analyses of 1,928 individuals with genetic generalized epilepsy (GGEs) vs. 8,578 controls and subsets of familial GGEs (945 cases vs. 8,626 controls) or sporadic GGEs (1,005 cases vs. 8,621 controls). The 95% confidence intervals are shown as grey solid lines. The slope of the solid black line indicates the genomic inflation factor, whereas the slope of the dotted line equals 1. Labels: genes that are enriched in cases in both datasets among the five top-raking genes. Exome-wide significance after Bonferroni correction (dark red line) was defined by a *p* value < 2.9×10^{-7}.

Table 3.3: Top-ranked genes in the primary analyses of ultra-rare functional variants.

Odds Ratio (OR) and *p* values are given from a Cochran-Mantel-Haenszel exact test. The accompanying homogeneity *p* value indicates the lowest *p* value from Breslow-Day & Woolf tests for homogeneity of odds, where *p* values < 0.05 indicate significantly different odds between the two analysis datasets. CI: Confidence Interval. HGNC: HUGO Gene Nomenclature Committee genes names. QVs: qualifying variants. See Table 3.1 for the details of the PPh2, REVEL & MTR analysis models.

Analysis	URVs	HGNC	Epilepsy gene	Qualifying Cases			Qualifying Controls			OR (95% CI)	*P* value (homogeneity)
				1st Dataset	2nd Dataset	Both Datasets	1st Dataset	2nd Dataset	Both Datasets		
All GGEs	PPh2	*GABRG2*	yes	7 (0.64%)	3 (0.36%)	10 (0.52%)	3 (0.04%)	1 (0.06%)	4 (0.05%)	12.1 (3.4 – 54.1)	1.8×10^{-5} (0.54)
	REVEL			4 (0.36%)	3 (0.36%)	7 (0.36%)	0 (0.00%)	1 (0.06%)	1 (0.01%)	28.3 (3.4 – 1307.3)	1.3×10^{-4} (0.15)
	MTR			4 (0.36%)	3 (0.36%)	7 (0.36%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	∞ (6.1 – ∞)	1.2×10^{-5} (0.53)
Familial GGEs	PPh2	*GABRG2*	yes	6 (0.95%)	2 (0.63.%)	8 (0.85%)	3 (0.04%)	1 (0.06%)	4 (0.05%)	18.9 (5 – 86.5)	3×10^{-6} (0.63)
	REVEL			3 (0.48%)	2 (0.63%)	5 (0.53%)	0 (0.00%)	1 (0.06%)	1 (0.01%)	40.6 (4.4 – 1934.3)	1.0×10^{-4} (0.19)
	MTR			3 (0.48%)	2 (0.63%)	5 (0.53%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	∞ (7.9 – ∞)	1.4×10^{-5} (0.64)
Sporadic GGEs	PPh2	*FAM13C*	-	5 (0.81%)	0 (0.00%)	5 (0.50%)	4 (0.05%)	0 (0.00%)	4 (0.05%)	17.6 (3.8 – 89.0)	1.3×10^{-4} (0.44)
	REVEL	*TNFRSF21*	-	2 (0.47%)	2 (0.39%)	4 (0.40%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	∞ (5.1 – ∞)	2.3×10^{-4} (0.52)
	MTR	*TRPV5*	-	3 (0.70%)	0 (0.00%)	3 (0.30%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	∞ (5.8 – ∞)	3.0×10^{-4} (0.18)

Table 3.4: Top-ranked genes in the secondary analyses of rare functional variation.

Odds Ratio (OR) and *p* values are given from a Cochran-Mantel-Haenszel exact test. The accompanying homogeneity *p* value indicates the lowest *p* value from Breslow-Day & Woolf tests for homogeneity of odds, where *p* values < 0.05 indicate significantly different odds between the two analysis datasets. CI: Confidence Interval. GGE: Genetic Generalized Epilepsy. HGNC: HUGO Gene Nomenclature Committee genes names. OMIM: Online Mendelian Inheritance in Man database. URVs: Ultra-rare variants. OMIM phenotypes: *ZIC3*: VACTERL (vertebral defects, anal atresia, cardiac defects, tracheo-esophageal fistula, renal anomalies, and limb abnormalities), *COPA*: Autoimmune interstitial lung, joint, and kidney disease.

Analysis	Variants		HGNC (OMIM gene)	Qualifying Cases			Qualifying Controls			OR (95% CI)	*P* value (homogeneity)
				1st Dataset	2nd Dataset	Both Datasets	1st Dataset	2nd Dataset	Both Datasets		
All GGEs	Rare	+ URVs	*ALDH8A1* (no)	14 (1.3%)	3 (0.4%)	17 (0.9%)	22 (0.3%)	2 (0.1%)	24 (0.3%)	3.9 (1.9 – 7.6)	7.5×10^{-5} (0.82)
		- URVs	*KIFC2* (no)	8 (0.7%)	1 (0.1%)	9 (0.5%)	5 (0.07%)	0 (0%)	5 (0.06%)	10.5 (3.1 – 40.4)	3.4×10^{-5} (0.64)
Familial GGEs	Rare	+ URVs	*ZIC3* (yes)	6 (1%)	0 (0%)	6 (1%)	3 (0.04%)	0 (0%)	3 (0.03%)	22.0 (4.7 – 136.7)	2.3×10^{-5} (0.53)
		- URVs	*ZIC3* (yes)	4 (0.6%)	0 (0%)	4 (0.5%)	0 (0%)	0 (0%)	0 (0%)	∞ (7.2 – ∞)	4.9×10^{-5} (0.23)
Sporadic GGEs	Rare	+ URVs	*COPA* (yes)	6 (1.4%)	7 (1.4%)	13 (1.3%)	19 (0.3%)	4 (0.2%)	23 (0.3%)	5.0 (2.2 – 10.7)	6.2×10^{-5} (0.69)
		- URVs	*FTHL17* (no)	5 (1.2%)	3 (0.6%)	8 (0.8%)	5 (0.07%)	2 (0.1%)	7 (0.08%)	10.4 (3.1 – 35.7)	6.5×10^{-5} (0.37)

Table 3.5: Top-ranked genes in the secondary analyses of predicted Loss of Function (pLoF) variants.

Odds Ratio (OR) and *p* values are given from a Cochran-Mantel-Haenszel exact test. The accompanying homogeneity *p* value indicates the lowest *p* value from Breslow-Day & Woolf tests for homogeneity of odds, where *p* values < 0.05 indicate significantly different odds between the two analysis datasets. CI: Confidence Interval. GGE: Genetic Generalized Epilepsy. HGNC: HUGO Gene Nomenclature Committee genes names. OMIM: Online Mendelian Inheritance in Man database. URVs: Ultra-rare variants.

Analysis	pLoF Variants		HGNC (OMIM gene)	Qualifying Cases			Qualifying Controls			OR (95% CI)	*P* value (homogeneity)
				1st Dataset	2nd Dataset	Both Datasets	1st Dataset	2nd Dataset	Both Datasets		
All GGEs	URVs only		*CEP350* (no)	5 (0.5%)	0 (0%)	5 (0.3%)	1 (0.01%)	0 (0%)	1 (0.01%)	31.1 (3.5 – 1460.3)	2.7×10^{-4} (0.28)
	Rare	+ URVs	*BEST3* (no)	8 (0.7%)	2 (0.2%)	10 (0.5%)	10 (0.1%)	0 (0%)	10 (0.1%)	5.6 (2.1 – 15.3)	3×10^{-4} (0.64)
		- URVs	*BEST3* (no)	7 (0.6%)	1 (0.1%)	8 (0.4%)	6 (0.09%)	0 (0%)	6 (0.07%)	7.7 (2.3 – 27.4)	3.1×10^{-4} (0.59)
Familial GGEs	URVs only		*CEP350* (no)	4 (0.6%)	0 (0%)	4 (0.5%)	1 (0.01%)	0 (0%)	1 (0.01%)	43.9 (4.3 – 2132.8)	2.3×10^{-4} (0.42)
	Rare	+ URVs	*CPA3* (no)	4 (0.6%)	0 (0%)	4 (0.5%)	1 (0.01%)	0 (0%)	1 (0.01%)	43.9 (4.3 – 2132.8)	2.3×10^{-4} (0.42)
		- URVs	*AIM1* (no)	6 (1%)	3 (1%)	9 (1%)	10 (0.1%)	5 (0.3%)	15 (0.2%)	5.2 (2.0 – 12.9)	4.6×10^{-4} (0.45)
Sporadic GGEs	URVs only		*GRIK5* (no)	3 (0.7%)	0 (0%)	3 (0.3%)	0 (0%)	0 (0%)	0 (0%)	∞ (5.8 – ∞)	3×10^{-4} (0.18)
	Rare	+ URVs	*GRIK5* (no)	3 (0.7%)	0 (0%)	3 (0.3%)	0 (0%)	0 (0%)	0 (0%)	∞ (5.8 – ∞)	3×10^{-4} (0.18)
		- URVs	*DSN1* (no)	4 (0.9%)	0 (0%)	4 (0.4%)	2 (0.02%)	0 (0%)	2 (0.02%)	28.1 (4.0 – 309.8)	2.7×10^{-4} (0.35)

Table 3.6: *GABRG2* variants identified in cases and controls.

The variants were analyzed up to an external minor allele frequency (MAF) of 0.1%. All those detected in the cases had an internal MAF and external MAF = 0 (i.e., ultra-rare variants). No rare variants were seen in individuals with epilepsy. Two variants (p.T58A, p.D231N) were seen in two control individuals. The remaining variants were found only once. CAE: Childhood Absence Epilepsy. GEFS: Generalized Epilepsy Febrile Seizures. FS: Febrile Seizures. NAFE: Non-Acquired Focal Epilepsy. TM: Transmembrane segment. PPh2, REVEL and MTR scores considered damaging/deleterious/intolerant are underlined.

Group		Variant & Transcript				*In silico* predictions			Previously reported phenotypes
		NM_000816	NM_198904	NM_198903	Location	PPh2	REVEL	MTR	
(A) *GABRG2* variants in individuals diagnosed with genetic generalized epilepsy (GGE)									
Familial GGEs	**Ultra-rare**	c.478G>T p.A160S	c.478G>T p.A160S	c.478G>T p.A160S	N-terminus	1	0.42	0.50	-
		c.530G>C p.R177P	c.530G>C p.R177P	c.530G>C p.R177P	N-terminus	1	0.69	0.61	FS (p.R177G)
		c.595A>G p.M199V	c.595A>G p.M199V	c.595A>G p.M199V	N-terminus	1	0.91	0.47	GEFS (p.M199V); NAFE (p.M199V)
		c.639T>A p.Y213*	c.639T>A p.Y213*	c.759T>A p.Y253*	N-terminus	NA	NA	NA	-
		c.755T>C p.V252A	c.755T>C p.V252A	c.875T>C p.V292A	N-terminus	0.98	0.35	0.67	-
		c.1213G>A p.G405S	c.1237G>A p.G413S	c.1357G>A p.G453S	Cytoplasmic (TM3-TM4)	0.99	0.35	0.78	-
		c.1324G>T p.D442Y	c.1348G>T p.D450Y	c.1468G>T p.D490Y	Cytoplasmic (TM3-TM4)	1	0.94	0.76	-
		c.1370A>G p.N457S	c.1394A>G p.N465S	c.1514A>G p.N505S	TM4	1	0.84	0.56	-

Sporadic GGEs	**Ultra-rare**	c.259+2T>G IVS2SD	c.259+2T>G IVS2SD	c.259+2T>G IVS2SD	N-terminus	NA	NA	NA	-
		c.769+2T>G IVS6SD	c.769+2T>G IVS6SD	c.889+2T>G IVS7SD	N-terminus	NA	NA	NA	Familial CAE (IVS6SD)

(B) *GABRG2* variants in individuals without epilepsy (controls)

Controls	**Rare**	c.173C>A p.T58N	c.173C>A p.T58N	c.173C>A p.T58N	N-terminus	1	0.384	0.97	-
		c.530G>A p.R177Q	c.530G>A p.R177Q	c.530G>A p.R177Q	N-terminus	0.98	0.466	0.61	GGE (p.R177P); FS (p.R177G)
		c.691G>A p.D231N	c.691G>A p.D231N	c.811G>A p.D271N	N-terminus	1	0.485	0.70	-
		c.748G>A p.E250K	c.748G>A p.E250K	c.868G>A p.E290K	N-terminus	0.99	0.456	0.67	-
		c.1087C>T p.R363W	c.1087C>T p.R363W	c.1207C>T p.R403W	Cytoplasmic (TM3-TM4)	1	0.584	0.70	-
		c.1113_1115del p.K372del	c.1113_1115del p.K372del	c.1233_c.1235del p.K412del	Cytoplasmic (TM3-TM4)	NA	NA	NA	-
		c.1148G>A p.R383H	c.1172G>A p.R391H	c.1292G>A p.R431H	Cytoplasmic (TM3-TM4)	1	0.301	1.02	-
Controls	**Ultra-rare**	c.172A>G p.T58A	c.172A>G p.T58A	c.172A>G p.T58A	N-terminus	1.00	0.30	0.97	-
		c.571C>A p.Q191K	c.571C>A p.Q191K	c.571C>A p.Q191K	N-terminus	0.99	0.39	0.56	-
		NA	c.1130T>C p.L377P	c.1250T>C p.L417P	Cytoplasmic (TM3-TM4)	0.91	0.44	0.86	-
		c.1309C>T p.R437C	c.1333C>T p.R445C	c.1453C>T p.R485C	Cytoplasmic (TM3-TM4)	1.00	0.71	8.5	-

3.4.3 Overlap between top hits in large-scale studies

Although not study-wide significant, *GABRG2* achieved a higher rank than in our prior URVs analysis[78] in 640 familial GGEs vs. 3,877 controls using an analysis model comparable to the current PPh2 model (rank 7, $p = 9.2 \times 10^{-4}$). Its rank was higher than that seen in two recent large-scale analyses from the Epi25 Collaborative[33,34] in 3,108 GGEs vs. 8,436 controls (rank 3, $p = 6.2 \times 10^{-4}$) vs. 5,303 GGEs and 15,677 controls (rank = 37, $p = 6.1 \times 10^{-3}$). Apart from *GABRG2,* there was little overlap between the leading associations in the recent analyses[33,34,78] and this study (Table 3.7). *CACNA1B* [MIM: 601012], the top hit in our prior analysis[5] ($p = 1.7 \times 10^{-5}$), showed a less prominent association than previously seen (rank 5 in the MTR model/familial GGEs; $p = 0.00098$). Our analysis also did not recapture two genes previously seen as top hits with suggestive association[33,34] (*CACNA1G* [MIM: 604065] with $p = 2.5 \times 10^{-4}$ and *SLC6A1* [MIM: 137165] with $p = 2.1 \times 10^{-6}$). *GABRA1* [MIM: 137160] was among few shared top hits, achieving comparable ranks in all studies (rank 9 in the MTR model analysis of all GGEs with $p = 0.0023$; rank 8 in the 1st Epi25 Collaborative study[33] with $p = 0.0022$; rank 9 in the 2nd study[34] with $p = 0.0013$).

3.4.4 Overrepresentation of disease genes among the top-ranked genes

Multiple genes suggested to increase the susceptibility to GGE had limited evidence of association (Table 3.8). On the other hand, genes underlying dominant DEE syndromes were among the top-ranked genes (Table 3.9), as expected from the known enrichment of such URVs in genes causing dominant DEE in generalized epilepsies.[33,34,78]

3.4.5 Association of GABAergic gene sets with familial and sporadic GGE

Few $GABA_A$ encoding genes had p values < 0.05 (Table 3.10). The association of URVs in two gene sets important for GABAergic signaling (genes encoding $GABA_A$ receptors and GABAergic pathway genes) with the phenotype was not prominent in the analysis of deleterious URVs, whereas the incorporation of sub-genic intolerance in the definition of QVs improved the power[34,138] and unraveled clear association signals in the analysis of all GGEs (Figure 3.9A) and familial GGEs (Figure 3.9B). It also aided the identification of an association between genes encoding $GABA_A$ receptors and sporadic GGEs, though weaker than what was seen in comparisons of familial GGEs vs. controls (Figure 3.9C). We did not detect an association between GABAergic pathway genes and sporadic GGE as expected from previous findings,[33] possibly due to insufficient power or differences in the analysis models (Figure

3.9C). The outcome of a direct comparison of 945 individuals with familial GGE vs. 1,005 individuals with sporadic GGE was unremarkable and likely underpowered (Fig. 3.9D).

Table 3.7: Comparisons of top-ranked genes with three previous large-scale rare variant association studies of genetic generalized epilepsy.

A. Association of top his from recent studies in the current analysis						
Genes from recent studies				**Outcomes in this study (All GGEs analysis)**		
Study	**Rank**	**Gene**	***P* value**	**URVs PPh2**	**URVs REVEL**	**URVs MTR**
		Top-ranked genes		***P* values (Rank if ≤ 10) in this study**		
Epi4K & EP/GP	1	*CACNA1B*	0.000017	0.011	0.0028	0.0015 **(rank 6)**
	2	*KEAP1*	0.000056	0.0016	0.15	0.052
	3	*COPB1*	0.00022	0.039	0.089	0.0096
	4	*PHTF1*	0.00030	0.0071	0.0019	0.079
	5	*KCNQ2*	0.00040	0.61	1	1
	5	*SLC9A2*	0.00040	0.36	0.14	0.14
	7	*ATP1A3*	0.00092	0.035	0.016	0.025
	7	*GABRG2*	0.00092	0.000018 **(rank 1)**	0.00013 **(rank 1)**	0.000012 **(rank 1)**
	9	*ZNF100*	0.0010	0.041	0.039	0.039
	10	*CUX1*	0.0013	0.0066	0.017	0.0029
	10	*SCN1A*	0.0013	0.043	0.071	0.012
Epi25 Years 1&2	1	*CACNA1G*	0.00025	0.51	0.19	0.17
	2	*EEF1A2*	0.00038	0.21	0.57	0.57
	3	*GABRG2*	0.00062	0.000018	0.00013	0.000012 **(rank 1)**
	3	*UNC79*	0.00062	0.064	1	1
	5	*ALDH4A1*	0.0014	0.68	0.68	1
	6	*SLC6A1*	0.0020	1	0.57	0.57
	7	*RC3H2*	0.0020	0.75	1	1
	8	*GABRA1*	0.0022	0.0053	0.0023	0.0023 **(rank 9)**
	9	*LRRFIP1*	0.0052	0.76	0.59	1
	9	*DNAJC13*	0.0052	0.38	0.60	1
	9	*ZBTB2*	0.0052	1	1	1
Epi25 Years 1 – 3	1	*SLC6A1*	0.0000021	1	0.57	0.57
	2	*SCN1A*	0.000034	0.043	0.071	0.012
	3	*MYH8*	0.000262	0.31	0.33	0.065
	4	*FBXO42*	0.000447	1	1	1
	5	*DAW1*	0.000619	0.34	0.34	0.50
	6	*GRIN2A*	0.000862	1	1	1
	7	*NUP98*	0.000863	0.22	0.039	0.22
	8	*MYO5C*	0.001199	0.49	0.37	0.69
	9	*GABRA1*	0.001348	0.0053	0.0023	0.0023 **(rank 9)**
	10	*KCNK18*	0.001599	1	1	1

B. **Association of top his from the current analysis in recent studies**

Genes from this study (All GGEs analysis)				Outcomes in previous studies		
Analysis	**Rank**	**Gene**	***P* value**	**Epi4K & EP/GP**	**Epi25 Years 1&2**	**Epi25 Years 1 – 3**
		Top-ranked genes		***P* values (Rank ≤ 10) in recent studies**		
URVs PPh2	1	*GABRG2*	0.000018	0.00092 **(rank 7)**	0.00062 **(rank 3)**	0.0061
	2	*FAM13C*	0.000072	0.06	1	> 0.03
	3	*MRPL20*	0.000087	1	1	> 0.03
	4	*DHDH*	0.00012	0.05	1	> 0.03
	5	*PLCH2*	0.00017	1	1	> 0.03
	6	*ACSF2*	0.00056	0.26	1	> 0.03
	7	*KCNMA1*	0.00057	1	1	> 0.03
	8	*COL5A3*	0.00075	0.41	0.18	> 0.03
	9	*RLN3*	0.00085	1	0.27	> 0.03
	10	*TANC2*	0.0011	0.01	1	> 0.03
URVs REVEL	1	*GABRG2*	0.00013	0.00092 **(rank 7)**	0.00062 **(rank 3)**	0.0061
	2	*PDE1A*	0.00015	0.05	1	> 0.03
	3	*MDN1*	0.00023	0.66	0.18	> 0.03
	4	*WDR83*	0.00084	0.0098	1	> 0.03
	5	*RIOK2*	0.00085	0.0018	1	> 0.03
	6	*CEP350*	0.00090	0.045	0.27	> 0.03
	7	*TTC21B*	0.0014	0.46	1	> 0.03
	8	*PLEKHM3*	0.0016	0.023	0.57	> 0.03
	8	*FKBP10*	0.0016	0.26	1	> 0.03
	8	*SURF1*	0.0016	0.60	0.57	> 0.03
URVs MTR	1	*GABRG2*	0.000012	0.00092 **(rank 7)**	0.00062 **(rank 3)**	0.0061
	2	*CEP350*	0.00084	0.045	0.27	> 0.03
	3	*PRSS8*	0.00085	0.26	1	> 0.03
	4	*RELN*	0.0011	0.36	0.62	> 0.03
	5	*MDN1*	0.0013	0.66	0.18	> 0.03
	6	*CACNA1B*	0.0015	0.000017 **(rank 1)**	0.18	> 0.03
	7	*PRSS12*	0.0016	1	0.71	> 0.03
	8	*ZNF662*	0.0020	0.092	0.47	> 0.03
	9	*GABRA1*	0.0023	0.055	0.0022 **(rank 8)**	0.0013 **(rank 9)**
	9	*TCN2*	0.0023	0.26	0.66	> 0.03

Table 3.8: Association of OMIM genes implicated in susceptibility to generalized epilepsy.

Results from the primary analyses limited to genes with *p* values < 0.05.

Analysis	Gene	Qualifying Cases	Frequency in cases	Qualifying Controls	Frequency in controls	Direction of association	Association *P* value	Rank	Hypergeometric test *P* value
				All GGEs					
URVs PPh2	*GABRG2*	10	0.00519	4	0.00047	Cases	0.000018	1	7.4e-04
	KCNMA1	6	0.00311	2	0.00023	Cases	0.00057	7	1.1e-05
	GABRA1	5	0.00259	2	0.00023	Cases	0.00527	48	5.5e-06
	RORB	5	0.00259	2	0.00023	Cases	0.01538	133	2.3e-06
URVs REVEL	*GABRG2*	7	0.00363	1	0.00012	Cases	0.00013	1	7.4e-04
	GABRA1	5	0.00259	1	0.00012	Cases	0.00227	19	8.7e-05
	RORB	5	0.00259	2	0.00023	Cases	0.01538	83	2.9e-05
URVs MTR	*GABRG2*	7	0.00363	0	0.00000	Cases	0.000012	1	7.4e-04
	GABRA1	5	0.00259	1	0.00012	Cases	0.00227	9	1.8e-05
	RORB	4	0.00207	2	0.00023	Cases	0.03883	182	3.0e-04
	GABRB3	2	0.00104	0	0.00000	Cases	0.04440	216	1.5e-05
				Familial GGEs					
URVs PPh2	*GABRG2*	8	0.00847	4	0.00046	Cases	3.0e-06	1	7.4e-04
	RORB	5	0.00529	2	0.00023	Cases	0.00061	7	1.1e-05
	KCNMA1	4	0.00423	2	0.00023	Cases	0.00098	11	5.4e-08
	GABRA1	3	0.00317	2	0.00023	Cases	0.0076	57	7.4e-08

Table 3.8: Continued.

Analysis	Gene	Qualifying Cases	Frequency in cases	Qualifying Controls	Frequency in controls	Direction of association	Association *P* value	Rank	Hypergeometric test *P* value
					Familial GGEs				
URVs REVEL	*GABRG2*	5	0.00529	1	0.00012	Cases	0.00010	1	7.4e-04
	RORB	5	0.00529	2	0.00023	Cases	0.00061	4	3.1e-06
	GABRA1	3	0.00317	1	0.00012	Cases	0.0035	17	2.2e-07
URVs MTR	*GABRG2*	5	0.00529	0	0.00000	Cases	1.4e-05	1	7.4e-04
	RORB	4	0.00423	2	0.00023	Cases	0.0031	13	4.0e-05
	GABRA1	3	0.00317	1	0.00012	Cases	0.0035	17	2.2e-07
					Sporadic GGEs				
URVs REVEL	*KCNMA1*	4	0.00398	5	0.00058	Cases	0.0071	43	3.2e-02
	GABRB3	2	0.00199	1	0.00012	Cases	0.033	188	8.3e-03
URVs MTR	*GABRB3*	2	0.00199	0	0.00000	Cases	0.015	71	5.2e-02
	GABRG2	2	0.00199	0	0.00000	Cases	0.015	72	1.3e-03

Table 3.9: Association of OMIM genes implicated autosomal dominant developmental and epileptic encephalopathies.

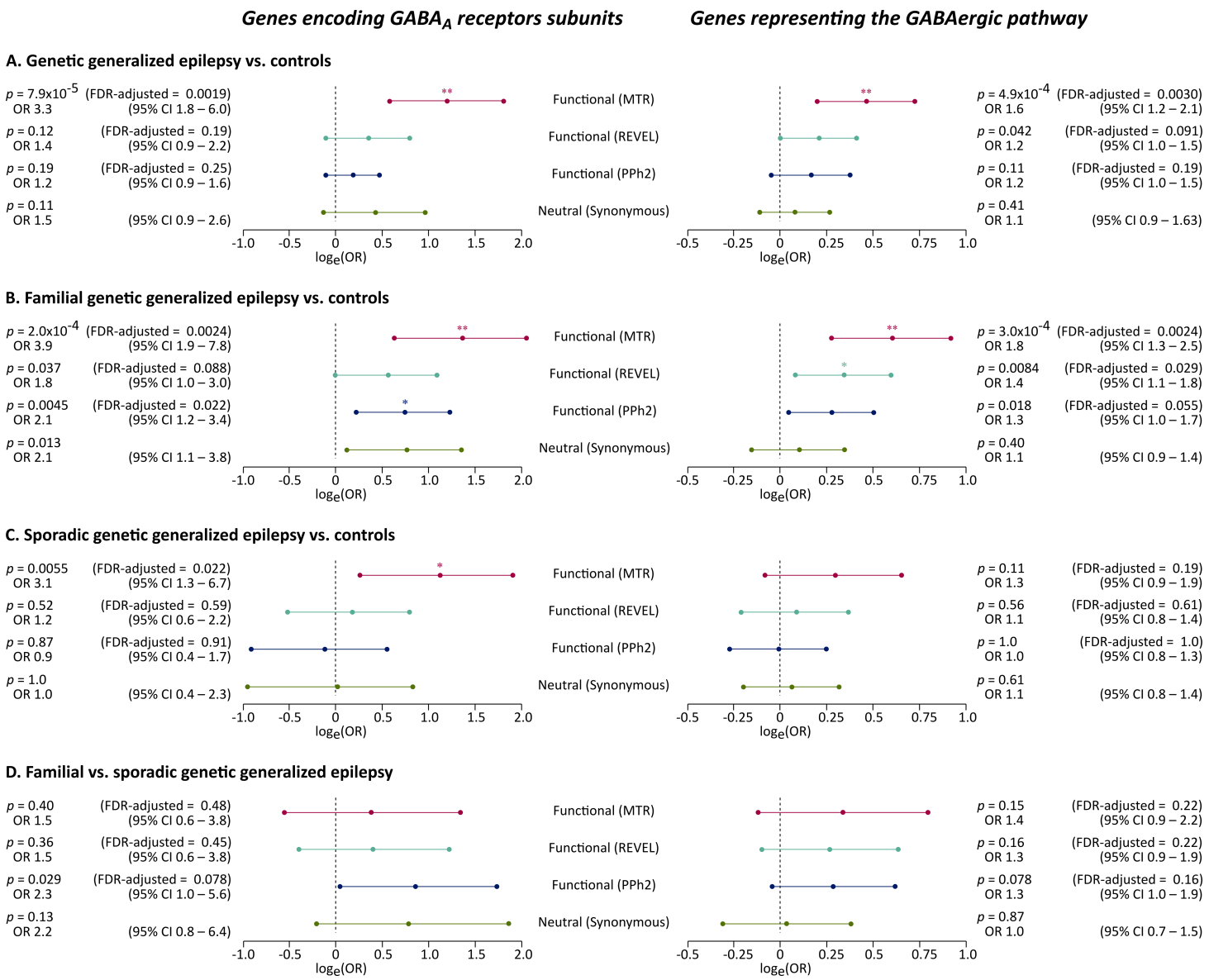
Analysis	Gene	Qualifying Cases	Frequency in cases	Qualifying Controls	Frequency in controls	Direction of association	Association *P* value	Rank	Hypergeometric test *P* value
					All GGEs				
URVs PPh2	*GABRG2*	10	0.00519	4	0.00047	Cases	0.000018	1	2.4e-03
	CACNA1A	11	0.00571	21	0.00245	Cases	0.00370	32	2.6e-03
	GABRA1	5	0.00259	2	0.00023	Cases	0.00527	48	2.0e-04
	KCNA2	5	0.00259	4	0.00047	Cases	0.00656	55	8.9e-06
	SCN1A	11	0.00571	25	0.00291	Cases	0.04277	376	2.0e-03
URVs REVEL	*GABRG2*	7	0.00363	1	0.00012	Cases	0.00013	1	2.4e-03
	GABRA1	5	0.00259	1	0.00012	Cases	0.00227	19	9.3e-04
	KCNA2	5	0.00259	4	0.00047	Cases	0.00656	43	1.5e-04
	GRIN2B	3	0.00156	1	0.00012	Cases	0.00958	58	1.1e-05
	HCN1	3	0.00156	1	0.00012	Cases	0.01860	103	4.6e-06
	KCNB1	4	0.00207	5	0.00058	Cases	0.03897	272	4.4e-05
	CACNA1A	8	0.00415	18	0.00210	Cases	0.04283	283	4.5e-06
URVs MTR	*GABRG2*	7	0.00363	0	0.00000	Cases	0.000012	1	2.4e-03
	GABRA1	5	0.00259	1	0.00012	Cases	0.00227	9	2.0e-04
	KCNA2	5	0.00259	4	0.00047	Cases	0.00656	32	6.0e-05
	GRIN2B	3	0.00156	1	0.00012	Cases	0.00958	42	3.0e-06
	SCN1A	9	0.00467	13	0.00152	Cases	0.01180	52	1.5e-07
	HCN1	3	0.00156	1	0.00012	Cases	0.01860	76	2.5e-08
	CHD2	3	0.00156	2	0.00023	Cases	0.02148	132	2.6e-08
	CACNA1A	6	0.00311	11	0.00128	Cases	0.03028	140	1.3e-09
	GABRB3	2	0.00104	0	0.00000	Cases	0.04440	216	1.8e-09
					Familial GGES				
URVs PPh2	*GABRG2*	8	0.00847	4	0.00046	Cases	3.0e-06	1	2.4e-03
	GABRA1	3	0.00317	2	0.00023	Cases	0.0076	57	8.2e-03
	GABRB2	3	0.00317	3	0.00035	Cases	0.013	92	1.4e-03

	SCN1A	7	0.00741	25	0.00290	Cases	0.030	202	1.4e-03
	HCN1	2	0.00212	2	0.00023	Cases	0.038	258	3.6e-04
URVs REVEL	*GABRG2*	5	0.00529	1	0.00012	Cases	0.00010	1	2.4e-03
	GABRA1	3	0.00317	1	0.00012	Cases	0.0035	17	7.4e-04
	GABRB2	3	0.00317	3	0.00035	Cases	0.013	96	1.6e-03
	CHD2	3	0.00317	4	0.00046	Cases	0.016	103	1.1e-04
	GRIN2B	2	0.00212	1	0.00012	Cases	0.020	123	1.1e-05
	HCN1	2	0.00212	1	0.00012	Cases	0.020	124	4.8e-07
	KCNB1	3	0.00317	5	0.00058	Cases	0.030	190	3.1e-07
	SCN1A	7	0.00741	25	0.00290	Cases	0.033	196	1.8e-08
	CACNA1A	5	0.00529	17	0.00197	Cases	0.037	201	9.5e-10
URVs MTR	*GABRG2*	5	0.00529	0	0.00000	Cases	1.4e-05	1	2.4e-03
	GABRA1	3	0.00317	1	0.00012	Cases	0.0035	17	7.4e-04
	CHD2	3	0.00317	2	0.00023	Cases	0.0052	21	1.6e-05
	SCN1A	6	0.00635	13	0.00151	Cases	0.0089	46	4.3e-06
	GRIN2B	2	0.00211	1	0.00012	Cases	0.020	92	2.6e-06
	HCN1	2	0.00211	1	0.00011	Cases	0.020	93	8.6e-08
	KCNB1	3	0.00317	4	0.00046	Cases	0.021	115	9.9e-09
	CACNA1A	4	0.00423	11	0.00128	Cases	0.037	151	2.4e-09
					Sporadic GGEs				
URVs PPh2	*CACNA1A*	6	0.00597	18	0.00209	Cases	0.0062	48	1.1e-01
	KCNA2	3	0.00299	4	0.00046	Cases	0.017	141	4.5e-02
URVs REVEL	*KCNA2*	3	0.00299	4	0.00046	Cases	0.017	114	2.4e-01
	GABRB3	2	0.00199	1	0.00012	Cases	0.033	188	7.4e-02
URVs MTR	*GABRB3*	2	0.00199	0	0.00000	Cases	0.015	71	1.6e-01
	GABRG2	2	0.00199	0	0.00000	Cases	0.015	72	1.3e-02
	KCNA2	3	0.00299	4	0.00046	Cases	0.017	83	1.0e-03

Table 3.10: Association of genes encoding $GABA_A$ receptor subunits.

Analysis	Gene	Qualifying Cases	Frequency in cases	Qualifying Controls	Frequency in controls	Direction of association	Association *P* value	Rank	Hypergeometric test *P* value
					All GGEs				
URVs PPh2	*GABRG2*	10	0.00519	4	0.00047	Cases	0.000018	1	1.0e-03
	GABRA1	5	0.00259	2	0.00023	Cases	0.0053	48	1.1e-03
URVs REVEL	*GABRG2*	7	0.00363	1	0.00012	Cases	0.00013	1	1.0e-03
	GABRA1	5	0.00259	1	0.00012	Cases	0.0023	19	1.6e-04
URVs MTR	*GABRG2*	7	0.00363	0	0.00000	Cases	0.000012	1	1.0e-03
	GABRA1	5	0.00259	1	0.00012	Cases	0.0023	9	3.5e-05
	GABRB3	2	0.00104	0	0.00000	Cases	0.044	216	1.3e-03
					Familial GGEs				
URVs PPh2	*GABRG2*	8	0.00847	4	0.00046	Cases	3.0e-06	1	1.0e-03
	GABRA1	3	0.00317	2	0.00023	Cases	0.0076	57	1.5e-03
	GABRB2	3	0.00317	3	0.00035	Cases	0.013	92	1.0e-04
URVs REVEL	*GABRG2*	5		1	0.00012	Cases	0.00010	1	1.0e-03
	GABRA1	3	0.00317	1	0.00012	Cases	0.0035	17	1.3e-04
	GABRB2	3	0.00317	3	0.00035	Cases	0.013	96	1.2e-04
URVs MTR	*GABRG2*	5	0.00529	0	0.00000	Cases	1.4e-05	1	1.0e-03
	GABRA1	3	0.00317	1	0.00012	Cases	0.0035	17	1.3e-04
					Sporadic GGEs				
URVs REVEL	*GABRB3*	2	0.00199	1	0.00012	Cases	0.033	188	1.7e-01
URVs MTR	*GABRB3*	2	0.00199	0	0.00000	Cases	0.015	71	6.9e-02
	GABRG2	2	0.00199	0	0.00000	Cases	0.015	72	2.4e-03

Figure 3.9: Association of ultra-rare variation in genes encoding $GABA_A$ receptors and GABAergic pathway genes with genetic generalized epilepsy (GGE). The forest plots show the association of ultra-rare deleterious and intolerant variants with the phenotype in analyses of 1,928 individuals with GGE *vs.* 8,578 controls **(A)**, 945 individuals with familial GGE *vs.* 8,626 controls **(B)**, 1,005 individuals with sporadic GGE *vs.* 8,621 controls **(C)**, and a direct comparison of 945 individuals with familial GGE *vs.* 1,005 individuals with sporadic GGE **(D)**. Four (primary and control) ultra-rare variant models are shown (*y* axis). The association in each analysis is displayed as the natural logarithm of stratified odds ratio from a Cochran-Mantel-Haenszel exact test (*x* axis). Errors bars indicated the logarithm of the 95% confidence intervals (CI). The corresponding odds ratios and associated *p* values, and False Discovery Rate (FDR) adjusted *p* values are displayed on the side. The tests for synonymous variants were not adjusted for multiple testing.



3.5 Discussion

Here, we add to the evidence indicating that deleterious URVs in *GABRG2* are a frequent risk factor for generalized epilepsies. Notably, rare variants (seen in external population controls) did not contribute to the observed association, emphasizing the role of URVs in less severe epilepsies.[33,34,47,64,78] This work extends our prior analysis[78] and corroborates the association of coding variation in *GABRG2* with familial GGE. The current analysis benefits from a higher number of individuals with familial GGE and a balanced distribution of familial and sporadic cases compared to recent large-scale analyses[33,34] enriched for sporadic GGEs. Our attempts to integrate multiple cohorts from independent studies to achieve this larger sample size came with some limitations. Quality control and harmonization measures mandated the exclusion of putative qualifying variants in genes of interest. The restrictions in genotype sharing across study sites limited the possibilities to invoke analysis methods incorporating covariates to handle residual population stratification. Also, the use of phenotypic definitions and classifications from independent studies might have resulted inadvertently in minor inconsistencies in sample stratification across the familial and sporadic cohorts.

The lack of study-wide significance in rare variant association studies in GGE and the failure to reproduce multiple leading associations speak to the marked genetic heterogeneity. The exact extent of the contribution of rare coding variation in GGE heritability is largely unknown. It remains, therefore, difficult to speculate on the interpretation of any negative findings, and on whether a further increase in statistical power might corroborate suggestive associations. Using a similar study design to the one used to examine the current set (slightly exceeding 10,000 samples), we estimate that a total sample size exceeding 16,000 samples would be required to achieve study-wide significance in a gene with rates of qualifying variants similar to those observed in *GABRG2*. These carrier rates seem, however, to be an upper-bound estimate due to the multitude of familial GGEs included here; the sample size required is probably much larger when examining sporadic GGEs.[33,34]

Nonetheless, the observed association of *GABRG2* with GGE further validates the outcomes of an analysis performed by the Epi25 Collaborative[33] (albeit, with partial overlap in the control datasets). The prominent difference in *GABRG2* rank in a second iteration[34] of the Epi25 study with an expanded sample size might be explained by the familial origin of *GABRG2* variants, since both studies had considerably lower ratios of familial to sporadic

GGEs (approximately 1:7 ratio). *GABRG2* was also the lead association in a burden analysis of pLoF URVs (p value = 6.9×10^{-5}) in a recent study investigating the exomes of 3,999 individuals with epilepsy (without further phenotypic sub-classification) vs. 277,586 controls from the UK Biobank[104] (https://genebass.org/). The recurrence of the same *GABRG2* variants in different epilepsy types (GGE & DEE or GGE & NAFE) as well as in familial and sporadic GGEs with overlapping phenotypes underscores a considerable genetic overlap and possibly a complex inheritance. For instance, the *GABRG2* locus was recently found to be associated with febrile seizures,[203] highlighting the role of common variants in a phenotype that was prominent in earlier families with an increased susceptibility to GGE and GEFS+ linked to rare *GABRG2* variants.[198–200]

Burden analysis also pointed out the presence of shared patterns of risk determinants between severe epilepsies (DEE) and common epilepsies (GGE & NAFE) in gene sets that are key for inhibitory signaling.[33,47] A former analysis (in 3,108 individuals with GGE) did not capture a considerable change in URVs burden in genes encoding $GABA_A$ receptors or GABAergic pathway genes upon the exclusion of a relatively small subset (n=380) of familial samples.[33] Conversely, we found a more prominent association between ultra-rare coding variation in GABAergic pathway genes and familial GGE in comparison to its association with sporadic GGE, albeit not demonstrable in direct (familial vs. sporadic) comparisons. Direct comparisons with sufficient power could help confirm the subtle differences in risk profiles.

In summary, we show that URVs in *GABRG2* are an important risk factor for GGE. Future work on epilepsy cohorts enriched with familial cases, extending the analysis to additional types of genetic variation (e.g., alterations in copy numbers and repeats, rare intronic and regulatory variants, and common risk alleles), could further our understanding of the genetic heterogeneity in GGE and the evidently complex inheritance.

Chapter 4: The association of coding variation in biologically informed gene sets with common and rare epilepsies

This chapter was adapted from: **Koko** *et al*. *EBioMedicine* 2021. PMID: 34571366. See the statement of contributions at the end of this dissertation.

4.1 Summary

Background: Analyses of few gene sets in epilepsy showed a potential to unravel key disease associations. We set out to investigate the burden of ultra-rare variants (URVs) in a comprehensive range of biologically informed gene sets presumed to be implicated in epileptogenesis.

Methods: The burden of 12 URV types in 92 gene sets was compared between cases and controls using whole exome sequencing data from individuals of European descent with developmental and epileptic encephalopathies (DEE, n = 1,003), genetic generalized epilepsy (GGE, n = 3,064), or non-acquired focal epilepsy (NAFE, n = 3,522), collected by the Epi25 Collaborative, compared to 3,962 ancestry-matched controls.

Results: Missense URVs in highly constrained regions were enriched in neuron-specific and developmental genes, whereas genes not expressed in the brain were not affected. GGE featured a higher burden in gene sets derived from inhibitory vs. excitatory neurons or associated receptors, whereas the opposite was found for NAFE, and DEE featured a burden in both. Top-ranked susceptibility genes from recent genome-wide association studies (GWAS) and gene sets derived from generalized vs. focal epilepsies revealed specific enrichment patterns of URVs in GGE vs. NAFE.

Interpretation: Missense URVs affecting highly constrained sites differentially impact genes expressed in inhibitory vs. excitatory pathways in generalized vs. focal epilepsies. The excess of URVs in top-ranked GWAS risk-genes suggests a convergence of rare deleterious and common risk-variants in the pathogenesis of generalized and focal epilepsies.

4.2 Background

Dismantling the genetic architecture behind epilepsy is yet to be within reach in many individuals. The role of genetic causality is apparent in the developmental and epileptic encephalopathies (DEEs),[48,73,109] sometimes with consequences on precision treatments.[21,110,111,204] In contrast, only few individuals with familial or sporadic genetic generalized epilepsies (GGEs) or non-acquired focal epilepsies (NAFEs) harbor monogenic causative variations.[33,34,47,78] Therefore, methodologies investigating the mutational burden of neurobiologically meaningful gene sets improve the prospects to dissect the joint effects of multiple genetic factors underlying the complex genetic architecture of these common epilepsy syndromes. Such 'gene set' analysis approaches are likely to provide valuable insights into the role of certain gene sets and pathways in epilepsy. Recent gene set burden analyses have shown an enrichment in ultra-rare deleterious and intolerant variants both in common and rare epilepsies in genes associated with dominant epilepsy syndromes, DEE genes, and neuro-developmental disorders (NDDs) with epilepsy genes, emphasizing a shared genetic component.[33,34,78] Evidence for the enrichment of rare missense variants in genes encoding $GABA_A$ receptors and GABAergic pathway genes in genetic generalized epilepsies pointed to the importance of the inhibitory pathway.[33,47] We used the large-scale dataset collected by the Epi25 Collaborative[33] for a comprehensive, exome-based case-control study to examine the burden of ultra-rare variants (URVs) in a large number of candidate gene sets for three different epilepsy forms (DEE, GGE, NAFE), aiming to understand the specific roles of deleterious URVs in key pathways implicated in epileptogenesis. Focusing on regional constraint and paralog conservation, we identified relevant and specific gene set associations in these three epilepsy forms.

4.3 Methods

4.3.1 Overview of the study design

The Epi25 Collaborative collected phenotyping data and generated exome sequencing data from individuals with different subtypes of epilepsy.[33] We analyzed subjects from recruitment years 1 and 2 (n=13,197 before filtering) targeting individuals diagnosed with DEE (n=1,474), GGE (n=4,510), NAFE (n=5,321) as outlined in Figure 4.1.

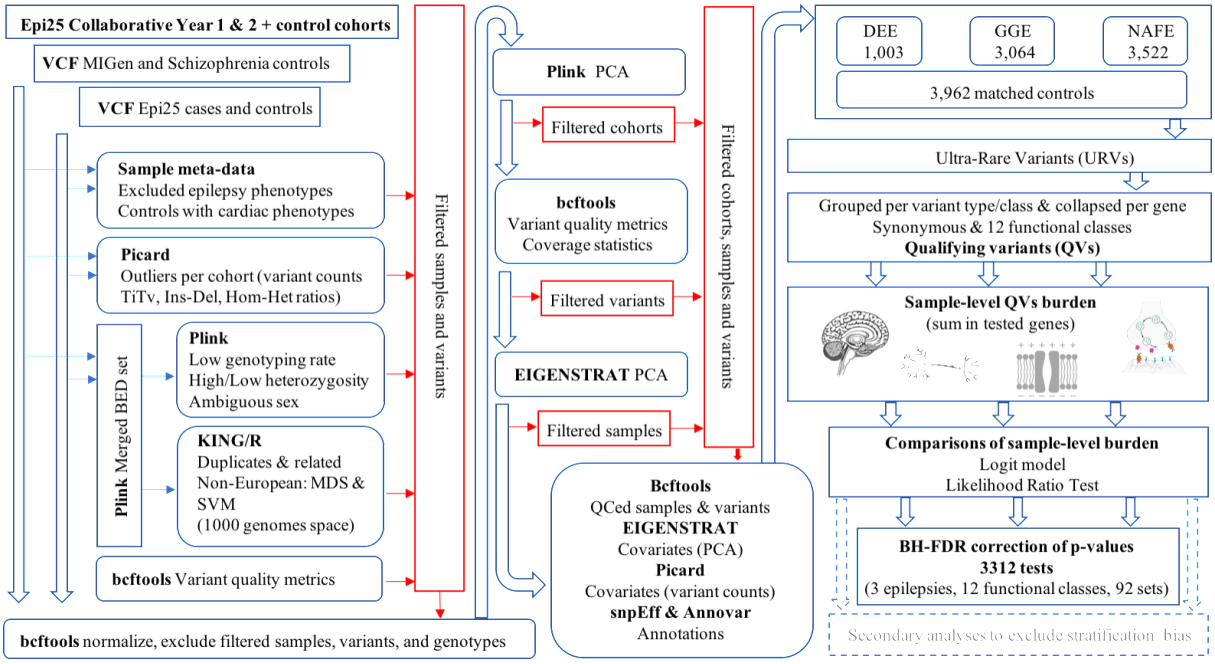


Figure 4.1: Outlines of the burden analysis method. Thirteen (twelve functional/non-synonymous and one synonymous) variants classes/types with focus on missense variants in constrained or paralog-conserved sites were tested in the three epilepsy phenotypes against a shared set of matched controls. The burden was examined in 92 gene sets (detailed in Table 1) using a logistic regression model with the count of qualifying variants per sample as a predictor and sample sex, ten principal components, singletons and exome-wide variant counts as covariates. Secondary analyses: an analysis restricting the genes in all gene sets to autosomal genes (to exclude bias introduced by male-to-female ratio imbalances), an analysis testing the controls prepared for exome sequencing using Illumina ICE capture kits against controls prepared with Agilent SureSelect capture kits (to exclude bias caused by differences in enrichment kits) coupled with an analysis of randomly selected cases and controls (500 permutations) to ensure adequate power, and a direct comparison of GGEs vs. NAFEs using highly constrained variants. BED: PLINK binary biallelic genotype table. BH-FDR: Benjamini-Hochberg False Discovery Rate. DEE: Developmental and Epileptic Encephalopathies. GGE: Genetic Generalized Epilepsies. Hom-Het: Homozygous-Heterozygous. Ins-Del : Insertion-Deletion. MDS: Multi-dimensional scaling. NAFE: Non-Acquired Focal Epilepsies. PCA: Principal Component Analysis. QCed: Quality-controlled. SVM: Support Vector Machine. TiTv: Transition-Transversion. VCF: Variant Call Format file.

The epilepsy classification, phenotyping and consent procedures have been previously described.[33] Five control cohorts[33,178] were available for this analysis (*n*=13,299), including Italian controls from the Epi25 Collaborative (*n*=300), the Swedish Schizophrenia Study controls (*n*=6,242), and three Myocardial Infarction Genetics (MIGen) Consortium cohorts: Leicester UK Heart Study (*n*=1,165), Ottawa Heart Study (*n*=1,915) and the Italian Atherosclerosis, Thrombosis, and Vascular Biology (ATVB) Study (*n*=3,677). The ethical approval and consents procedures for the individual cohorts were reported by the Epi25 Collaborative.[33] Subjects investigated by the Epi25 Collaborative provided signed informed consent at the participating centers according to local national ethical requirements and their standards at the time of collection. Approval for data reuse and analysis was obtained from the Epi25 Collaborative (cases) and dbGAP (controls). The data generation process has been previously described.[33] We considered Non-Finnish European (NFE) individuals diagnosed with DEE, GGE, or NAFE and matched controls for this analysis. The ancestry was predicted based on 1000 Genomes data[205] using a Support Vector Machine (SVM), removing 1,911 individuals with epilepsy and 146 controls. The quality control procedures[181,183,184,186,188,206–208] aimed to ensure adequate cases control matching and minimize the coverage and call rate differences between cohorts. The study methodology is summarized in Figure 4.1. Following quality control and harmonization steps outlined hereafter, 58% of the initial cases (Table 4.1) and 30% of the control samples (Table 4.2) were included in the analysis. The final analysis set included 7,589 cases (DEE=1,003, GGE=3,064, NAFE=3,522) and 3,962 matched controls (ATVB = 1,673, Leicester=1,082, Ottawa=924, Epi25 Italian=283).

Table 4.1: Epilepsy samples analyzed in this study.

Phenotype group	Total	Phenotype review	Initial QC	Final QC
Developmental and Epileptic Encephalopathy	1,474	1,467	1,040 (71%)	1,003 (68%)
Genetic Generalized Epilepsy	4,510	4,471	3,183 (71%)	3,064 (68%)
Non-Acquired Focal Epilepsy	5,321	5,304	3,616 (68%)	3,522 (66%)
Febrile Seizures and GEFS spectrum	301	Not considered		
Symptomatic / Lesional	1,434	Not considered		
Other epilepsies, unclassified epilepsies, non-epileptic seizures or not available	157	Not considered		
Total	13,197	11,242	7,839 (59%)	7,589 (58%)

Table 4.2: Control datasets analyzed in this study.

Control set	Capture kits	In dbGAP	In gnomAD	Phenotype	Total	Initial QC	Final QC
Epi25 Collaborative controls (Italy)	Illumina TruSeq/Nextera	No	No	Unaffected	300	283 (94%)	283 (94%)
Leicester Heart study (UK)	Illumina ICE	phs001000.v1.p1	Yes	Unaffected	1,100	1082 (98%)	1,082 (98%)
				Coronary Artery Disease	65	-	-
Ottawa Heart Study (Canada)	Agilent SureSelect	phs000806.v1.p1	Yes	Unaffected	987	946 (96%)	924 (94%)
				Coronary Artery Disease	928	-	-
Atherosclerosis Thrombosis & Vascular Biology study (Italy)	Agilent SureSelect	phs001592.v1.p1	No	Unaffected	1,802	1,673 (93%)	1,673 (93%)
				Coronary Artery Disease	1,875	-	-
Swedish Schizophrenia Study (Sweden)	Agilent SureSelect	phs000473.v2.p2	Yes	Unaffected	6,242	4,838 (78%)	-
Total					13,299	8,822 (66%)	3,962 (30%)

4.3.2 Data preparation and quality control

4.3.2.1 Baseline sample quality control

Access to two sets of variant calls (separate but jointly called VCF files) mapped to the human genome build GRCh37 was granted by the Epi25 Collaborative. The first set (*n*=13,497) contained calls from patients (*n*=13,197) and controls (*n*=300) collected by the Epi25 Collaborative. The second set (*n*=12,999) included controls from the Swedish Schizophrenia Study (S-SCZ; dbGAP accession number phs000473.v2.p2), patients and controls from MIGen Consortium cohorts (dbGAP accession numbers: phs000814.v1.p1, phs001000.v1.p1, phs000806.v1.p1) with access permission granted from dbGAP. The data generation process has been previously described.[33] The exome sequencing was performed on an Illumina HiSeq 2000 or 2500 (Illumina, USA) at the Broad Institute (different patches or timepoints) and utilized Illumina TruSeq/Nextera (Epi25), Illumina's ICE Capture (MIGen), or Agilent SureSelect Human All Exon Kits (MIGen and S-SCZ) (Agilent, USA). The sample counts are given in Tables 4.1 and 4.2 (see above).

Cases with a diagnosis other than DEE, GGE or NAFE were removed. The case definitions from the Epi25 Collaborative can be accessed online (http://epi-25.org/epi25-data-checks). Controls from MIGen cohorts with a coronary artery disease were not included in the analysis to avoid any prominent overlap in genetic predisposition. Gencode[206] coding sequence (CDS) boundaries (v33 lifted to b37) were padded with 10 bp and masked for low complexity and repeat regions (stratification files dated March 9, 2017), obtained from the Global Alliance for Genomics and Health (GA4GH) resource,[207] using bedtools v2.29.2.[209] All subsequent sample quality control, variant quality control and final analysis was performed over these regions (totaling 38Mb).

The variant calling metrics were gathered for the two datasets over the CDS boundaries described above using the Genome Analysis ToolKit[129] (GATK) v4.1.4.1. Outliers beyond four absolute deviations (per cohort) on total single nucleotide variants (snvs) count, insertions-deletions (indels) count, transition-transversion ratio (TiTv ratio), insertions-deletions ratio (Ins-Del ratio), or homozygous-heterozygous variants ratio (Hom-Het ratio) were filtered (Figure 4.2). The VCF files were converted to PLINK[186] v1.9 binaries and merged. The genotyping rate per sample was then calculated over the target CDS boundaries. Samples with genotyping rate less than 90% were filtered (Figure 4.2). PLINK was used to select a set of informative SNPs with missingness less than 0.01, minor allele frequency (MAF) exceeding 0.05, and in Hardy-Weinberg Equilibrium. These were then pruned (r 0.5) and used to estimate autosomal heterozygosity using the F-statistic. Outliers beyond three standard deviations were filtered (Figure 4.3). Informative SNPs (as detailed above) located on chrX were split, pruned, and then used to estimate the F-statistic over chrX. Samples with ambiguous ($0.2 < F < 0.6$; cut-offs determined following visualization of distribution) or discordant sequencing and reported sex were filtered (Figure 4.3).

KING[183] v2.2.4 was used to detect duplicate samples and estimate the relatedness. For each pair from duplicates and related samples up to the third degree (Figure 4.3), the sample with the lower genotyping rate was filtered. KING was also used to perform multidimensional scaling (MDS) analysis (5 principal components) on genotyping data from 2,451 samples from the "1000 Genomes Project" [205] followed by projection of the case and control samples into the "1000 Genomes" space. A subset of variants (n=73,080) that are called both in the "1000 Genomes" data and our dataset were selected for projection. The eigenvectors (five principal components) from a randomly selected subset of "1000 Genomes" samples (80% of samples)

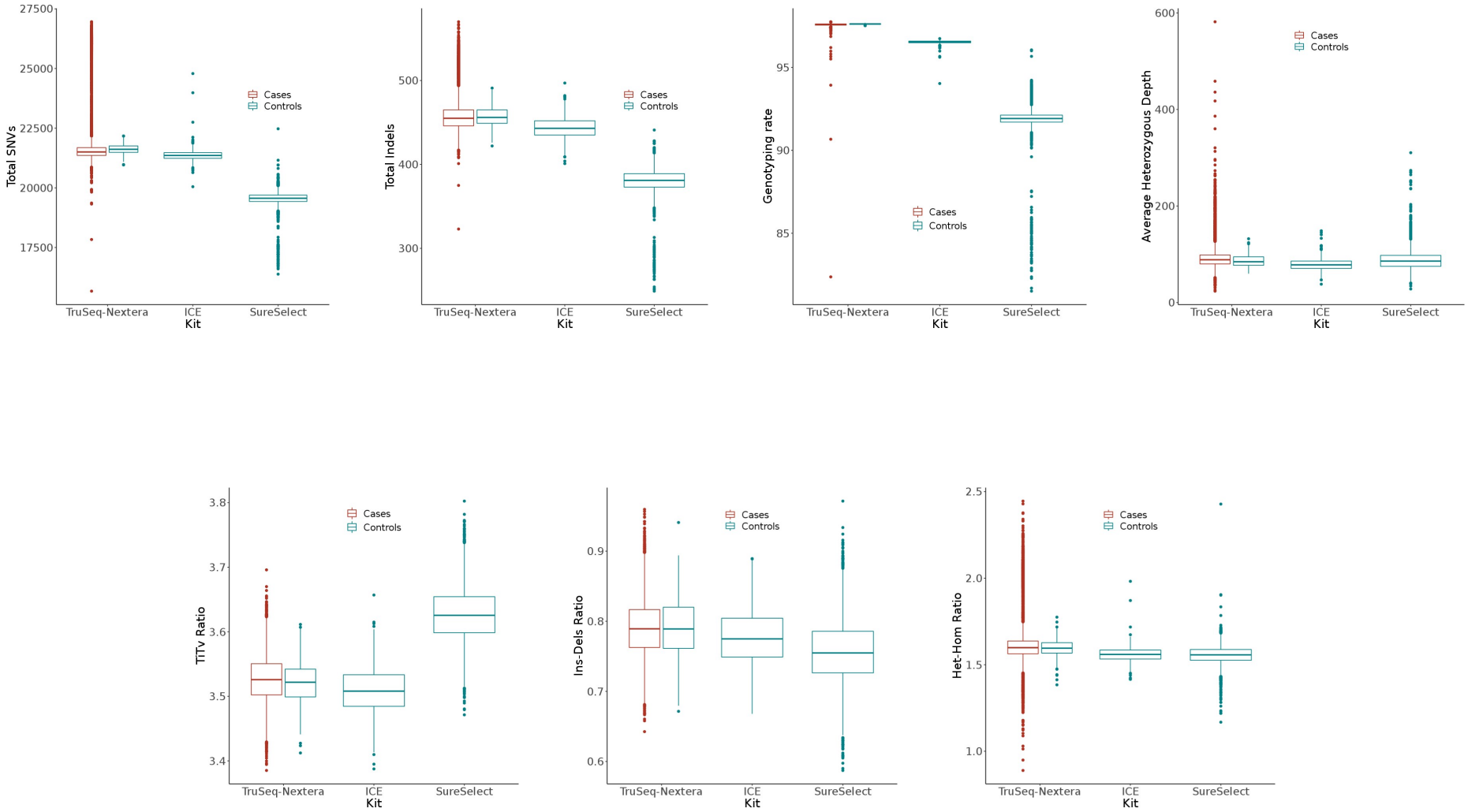


Figure 4.2: Variant calling metrics of sequencing cohorts grouped by capture kits. These metrics were collected over Gencode coding regions, padded with 10 bps and masked for regions of repeats and low complexity. Samples with genotyping rate < 90% and outliers (> 4 absolute deviations per sequencing cohort) on SNV/Indel counts, TiTv ratio, Ins-Del ratio and Het-Hom ratio were filtered. SNVs: single nucleotide variations. Indels: insertions and deletions. TiTv: transitions-transversions. Ins-Del: insertions-deletions. Het-Hom: heterozygous-homozygous.

were used to train an SVM, as implemented in R package *e1071*.[187] A radial kernel was used to recognize four major continental ancestry groups: Europeans (excluding Finnish), African, Admixed American, South and East Asian. The SVM was tested on the remaining "1000 Genomes" samples (20%), where it correctly recalled all samples from the European ancestry group, then used to classify the cases and control study samples (Figure 4.4). Samples with a predicted ancestry other than European were filtered. These filtering steps removed 7,511 samples.

To maximize the case-control matching among the remaining 18,985 samples, MDS (10 principal components) was repeated on a subset of samples from "1000 Genomes", of reported non-Finnish European ancestry (n=500, Northern and Western Europeans from Utah, British in England and Scotland, Tuscany in Italy, Iberian from Spain, Finnish in Finland). The ancestry projection of the study samples labelled as European by the SVM (variants selected as indicated above) was repeated on this MDS space of European "1000 Genomes" populations (Figure 4.4). Upon visualization of the first two principal components, samples clustering with Finnish Europeans were removed (PC1 > 0.04). To remove poorly matched cases and controls, the Euclidean distance between all pairs of remaining case and control samples (on PC1/PC2) was calculated. Outlier samples (beyond three median absolute deviations) were filtered. The final set of baseline-filtered samples constituted 7,836 cases and 8,822 controls (n=16,661) as detailed in Table 4.3.

Table 4.3: Summary of baseline sample-level quality control.

Criteria	Filter	Failing/Total (%)
Phenotype	Cases other than DEE, GGE, NAFE	1,773/13,197 (13.4%)
	Controls with cardiac phenotype	2,867/13,299 (21.6%)
Variant calling metrics	Outliers > 4 absolute deviations on key metrics	1,088/26,496 (4.1%)
Genotyping rate	< 90% in called variants in coding regions	66/26,496 (0.2%)
Autosomal heterozygosity	Outliers > 3 standard deviations	1011/26,496 (3.8%)
ChrX heterozygosity	0.2 < F < 0.6 or discordant reported/predicted sex	255/26,496 (1.0%)
Kinship	Duplicate, twin or related up to the 3rd degree	331/26,496 (1.2%)
Major continental ancestry	Non-European ancestry predictions from SVM trained on 1000 Genomes samples	2,057/26,496 (7.8%)
Samples failing one or multiple filters		7,511/26,496 (28.3%)
		18,985 samples remaining
Matching	Finnish (MDS)/outliers on PCA (PC1/2)	2,324/18,985 (12.2%)
		16,661 samples remaining

4.3.2.2 Baseline variant quality control

Two VCFs (see above) containing 6,429,324 jointly called sites were merged using bcftools/htslib[188] v1.10.2 and sites located outside the target CDS boundaries (see *Baseline sample quality control* above) and sites with low recalibrated variant quality scores (SNPs VQS Lod < -3.0998 and Indel VQS Lod < 0.8107 corresponding to VQSR sensitivity tranche 99.6 and 95.0, respectively) were filtered. The variants were allele-split and normalized using bcftools[188] and vt[131] v0.57721. The merged and normalized VCF was subset to the baseline filtered samples identified as detailed above. Genotype calls with depth < 10, quality < 20, or half-missing calls were set to missing. Heterozygous genotypes with allele depth to total depth ratio < 0.25 were set to reference calls. Variants with allele count equal to 0 were removed. These filtering steps were performed on a binary VCF stream piped between the outlined filtering commands. Afterwards, the depth of coverage per variant was calculated. Only variants covered at a minimum depth of ten-folds in 95% of the baseline filtered cases and control sets were kept. Additionally, the distribution of the difference in mean coverage and the percentage of samples covered at depth ten-folds was visualized. All outlier variants beyond three standard deviations were filtered. The statistical calculations were performed in R[193] v3.3. This quality control process resulted in a well-harmonized coverage between cases and controls (Figure 4.5).

4.3.2.3 Residual stratification

To maximize the cohort, sample, and variant matching, we performed multiple rounds of PCA coupled with coverage harmonization among cohorts. To remove poorly matching sample cohorts, a baseline round of PCA (10 principal components) was performed using PLINK[186] on a set of pruned variants (pruning was performed as described above). A cohort of Swedish controls (n=4,838) clustered poorly with the rest of the study samples on the top principal components (PC1/2) (Figure 4.6) and was therefore excluded. We then calculated the variant call rates across the remaining cohorts (Epi25, Leicester, Ottawa, ATVB), and removed all variants where any given cohort had a coverage < 95% (defined as the number of samples with non-missing genotype calls divided by the total number of samples in the cohort) or if the difference in coverage between any two given cohorts exceeded 0.5%. Variants not in Hardy-Weinberg equilibrium (p value < 10^{-6}) were identified and filtered.

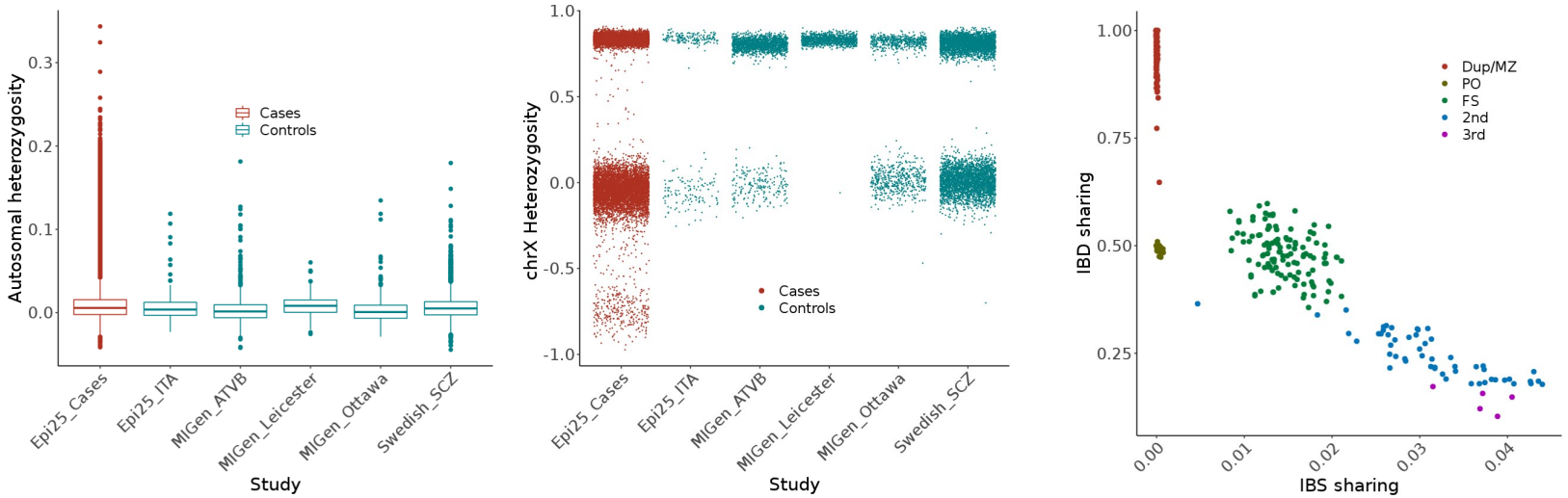


Figure 4.3: Heterozygosity and kinship filtering. A set of common, pruned variants with high genotyping rate was used to calculate the F-statistic in autosomes (left) and chrX (center) using PLINK. Samples with low or excess autosomal heterozygosity (> 3 standard deviations) were filtered. For sex prediction (SNP-sex), cut-offs of 0.2 and 0.6 were used to separate female and male clusters from samples with ambiguous sequencing sex prediction. Integrated kinship predictions (right) using KING identified pairs of duplicates/twins and related samples. One sample from each pair was filtered. IBD: Identity by descent. IBS: identity by state. Dup/MZ: duplicates or monozygotic twins. PO: parent-offspring. FS: full-sibling. 2nd: second degree. 3rd: third degree.

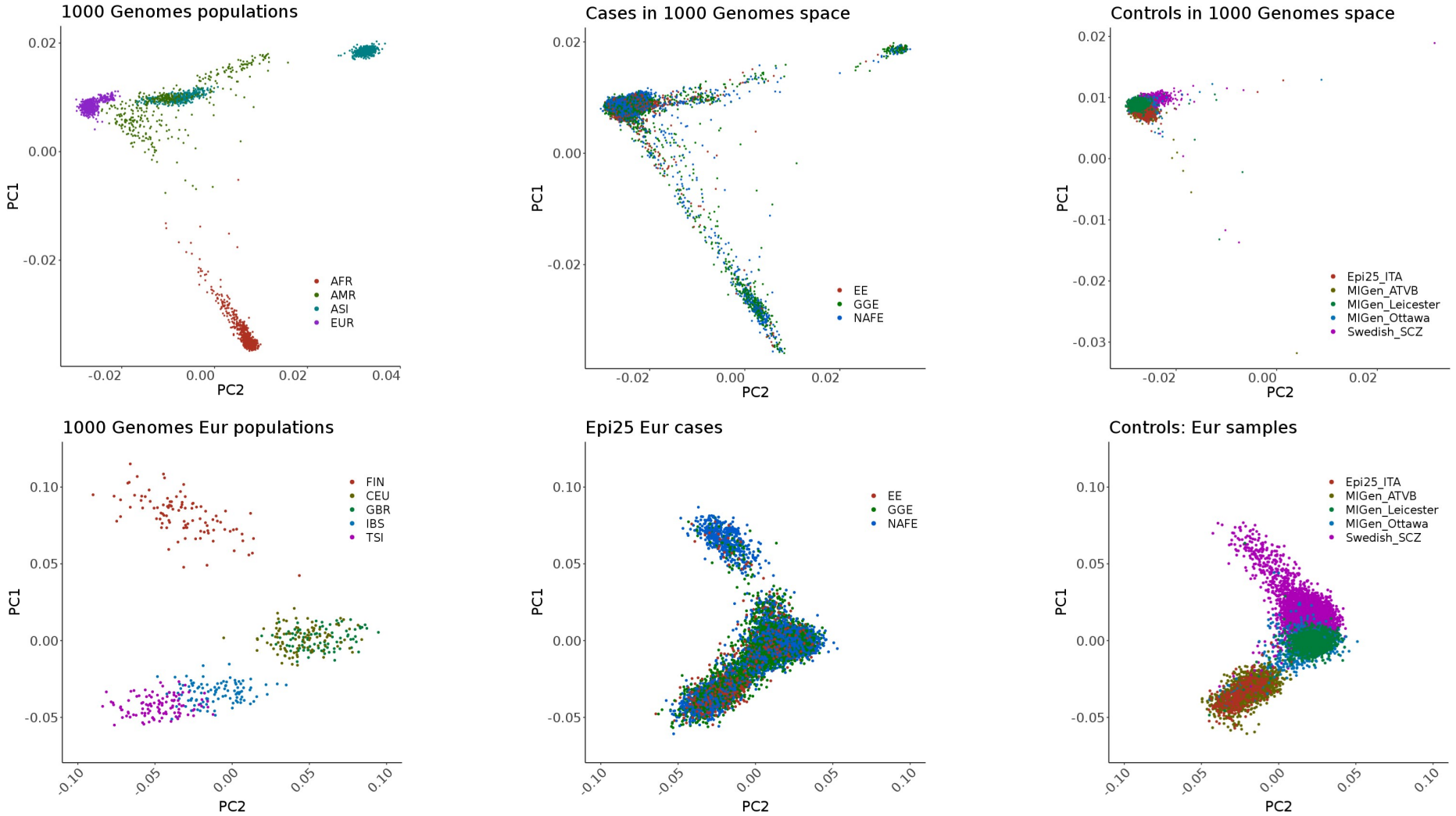


Figure 4.4: Continental ancestry groups. KING multidimensional scaling (MDS) projection in the 1000 Genomes space (top left panel) was used to estimate the major ancestry components. The cases (top center panel) showed wide variability in continental ancestry. The controls (top right panel) were mostly of European ancestry. A support vector machine was trained on 1000 Genomes sample labels and used to identify Epi25 and control samples with likely European ancestry (bottom center and right panels). A second round of MDS was performed to project the principal components of 500 samples of European ancestry from the 1,000 Genomes (bottom left) on the Epi25 cases and control samples classified as European (bottom center and right panel) . See Figure 4.5 for subsequent case-control matching.

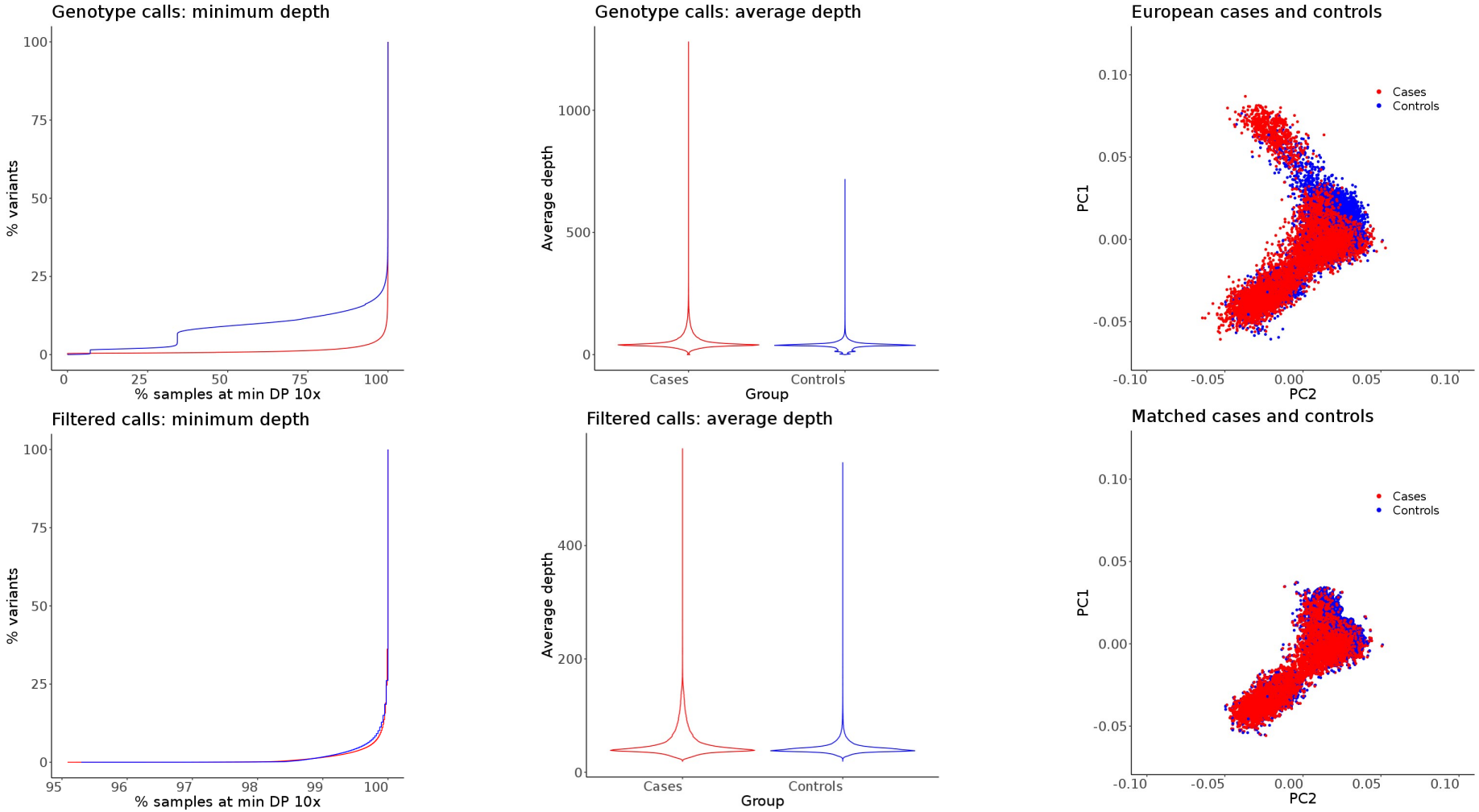


Figure 4.5: Baseline case-control matching and variant harmonization. The precent of samples covered at a minimum depth of 10x (top left) and the average depth (top center) are shown for the cases (red) and controls (blue). Multidimensional scaling was used to estimate the major ancestral components (top-right; see Figure 4.4 for details). To harmonize the ancestry and the variant calls, (a) about 20% of variants were removed where the percent of covered cases and controls was lower that 95%; (b) the difference in the average depth in cases and controls was calculated and outliers (> 3 standard deviations) were pruned out; (c) the difference in the percent of samples covered at depth 10x was calculated and variants with extreme differences (> 3 standard deviations) were also pruned; and (d) Poorly matched cases and controls on the top principal components PC1/PC2 and those of likely Finnish ancestry (PC1 > 0.04) were removed. This resulted in a homogeneous variant call rate (plots in bottom panels).

This filtering insured that the top principal components would capture the ancestry and not the exome capture kits differences (Figure 4.6). A second round of PCA (10 principal components) using EIGENSTRAT[184,208] v6.1.4 was performed (outlier vectors = 2, outliers sigma = 6, iterations = 5) complemented by removal of extreme outliers identified upon visual inspection (PC1/PC2). A small subset of poorly matched samples (*n*=272) was subsequently removed. A third and final round of PCA with identical EIGENSTRAT parameters showed a well-matched case-control cohort (Figure 4.6). The variant calling metrics were balanced for this set (Figure 4.7). The numbers of samples and variants in this final dataset are given in Tables 4.4 and 4.5.

Table 4.4: Final sample counts.

Group	Cohort	Samples		Females (%)	
Cases	DEE	7,589	1,003	4,067 (53.6%)	462 (46%)
	GGE		3,064		1,764 (57.6%)
	NAFE		3,522		1,841 (52.3%)
Controls	Epi25	3,962	283	767 (19.4%)	116 (40%)
	Ottawa		924		458 (49.7%)
	Leicester		1,082		1 (0.1%)
	ATVB		1,673		192 (11.5%)
All		11,551		4,834 (41.8%)	

Table 4.5: Final variant statistics.

Quality Control (QC)	Variants Count			Samples
Unfiltered	Jointly called sites in all samples		6,481,248	26,496
Baseline QC	In coding regions, normalized, genotype-filtered, filtered on variant quality score logarithm-of-odds, not in low-complexity regions, allele count > 0 in baseline-filtered samples		2,224,099	16,661
	Depth and call-rate harmonization		1,674,222	
Final QC	Allele count > 0 in final case-control set, cohort-level call-rate harmonization, in Hardy-Weinberg Equilibrium.		1,267,392	11,551
	Total variants per variant category	SNVs	1,247,342	
		Indels	20,050	
	Variants with allele frequency < 0.5 %		1,203,350	
	Variants with allele counts 1-3		1,054,919	
	Singleton variants		806,046	

4.3.3 Qualifying variants, gene collapsing analysis and genomic inflation

The variants were annotated using snpEff[190] v4.3 and Annovar[192] v20191024. We focused on URVs as these have shown a strong burden of deleterious pathogenic variants in

multiple studies of epilepsy and other neurological disorders.[33,34,64,78,106,108,210,211] URVs were defined based on their Minor Allele Counts (MACs) in the study dataset (internal allele count/frequency) and their estimated frequency in the general population (external MAF). Specifically, we examined variants that are: (i) Seen in less than three cases and controls (MAC ≤ 3); (ii) Not seen in DiscovEHR[135] (MAF in DiscovEHR = 0; (iii) Seen at a very low allele frequency in gnomAD[103] r2.1 database (MAF in gnomAD ≤ 2x10-5). We performed three separate analyses for the three epilepsy phenotypes; therefore, MACs were calculated independently in each analysis. This was intended to provide a better control for inflation compared to calculating MACs from all cases and controls. Accordingly, the reported variant counts in the control sets may differ slightly between the three analyses. Since our controls overlapped partially with gnomAD r2.1, we did not require complete absence of variants in gnomAD. The genotypes and annotations were queried using bcftools[188] or snpEff[190] and imported for statistical analysis in R[193] v3.3. These were collapsed in a dominant model (reference as 0, heterozygous, homozygous, and hemizygous as 1) to obtain a matrix of *samples vs. genes* where the cells contained 0/1 indicators for the presence or absence of a qualifying variant (QV) in each given sample and gene.

Single gene collapsing analysis was performed using Fisher's Exact Test (FET). The Genomic Inflation Factor ( ) was estimated using *QQ-perm*[194] by comparing observed vs. expected *p* values from a synonymous dominant model. Observed *p* values were calculated by performing a gene-level collapsing analysis for synonymous QVs using FET. Permutation-based *p* values were obtained from one thousand permutations (shuffling of case-control labels followed by FET). This was performed with a parallel implementation of the *QQ-perm* method using *parallel* package. The resulting *p* values were ordered and the mean values per rank from these 1000 permutations were taken as the expected *p* values for ordered ranks, and the 2.5th – 97.5th centiles were taken as 95% confidence intervals. The negative $\log_{10}$ of the observed *p* values was plotted against the negative $\log_{10}$ of the mean permutation *p* values to obtain the *Quantile-Quantile* plots of synonymous variants collapsing analysis. The synonymous collapsing analysis showed minimum inflation (genomic inflation factor = 1.06 – 1.1; Figure 4.8).

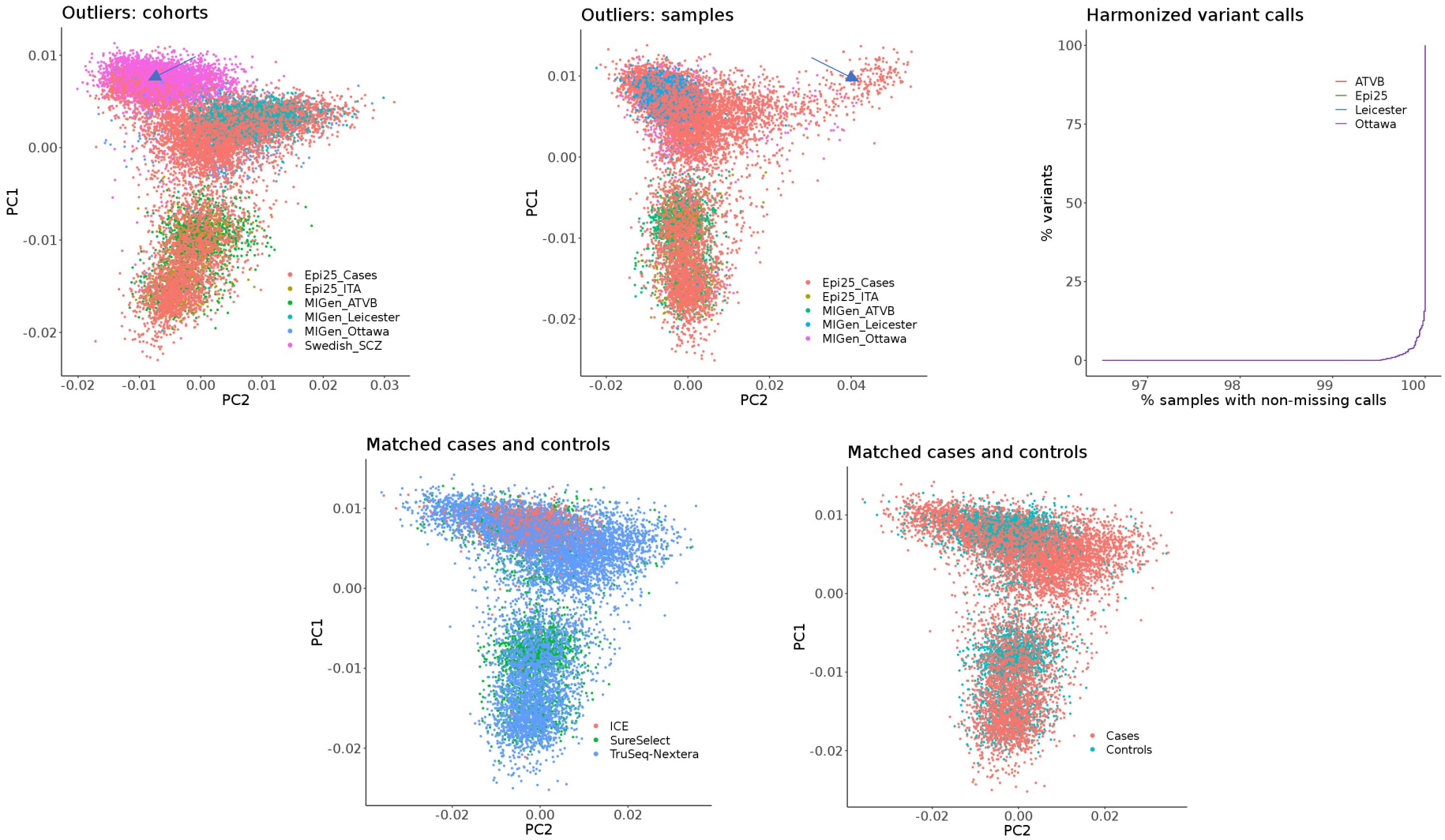


Figure 4.6: Final case-control matching. Principal component analysis of baseline-filtered cases and controls showed residual population and cohort stratification (top left). Swedish controls (outliers on the first round of PCA; arrow in top left panel) and additional poorly matched samples (outliers on the second round of PCA; arrow in top center panel) were filtered. The call-rate was harmonized between different sequencing cohorts (top right) by removing all variants where the difference in call rate between pairs of individual cohorts exceeds 0.5%. These measures minimized the patch effects (bottom left). The first and second principal components of the final matched case control set (bottom right panel) capture the northern-southern and eastern-western European geographical axis, respectively.

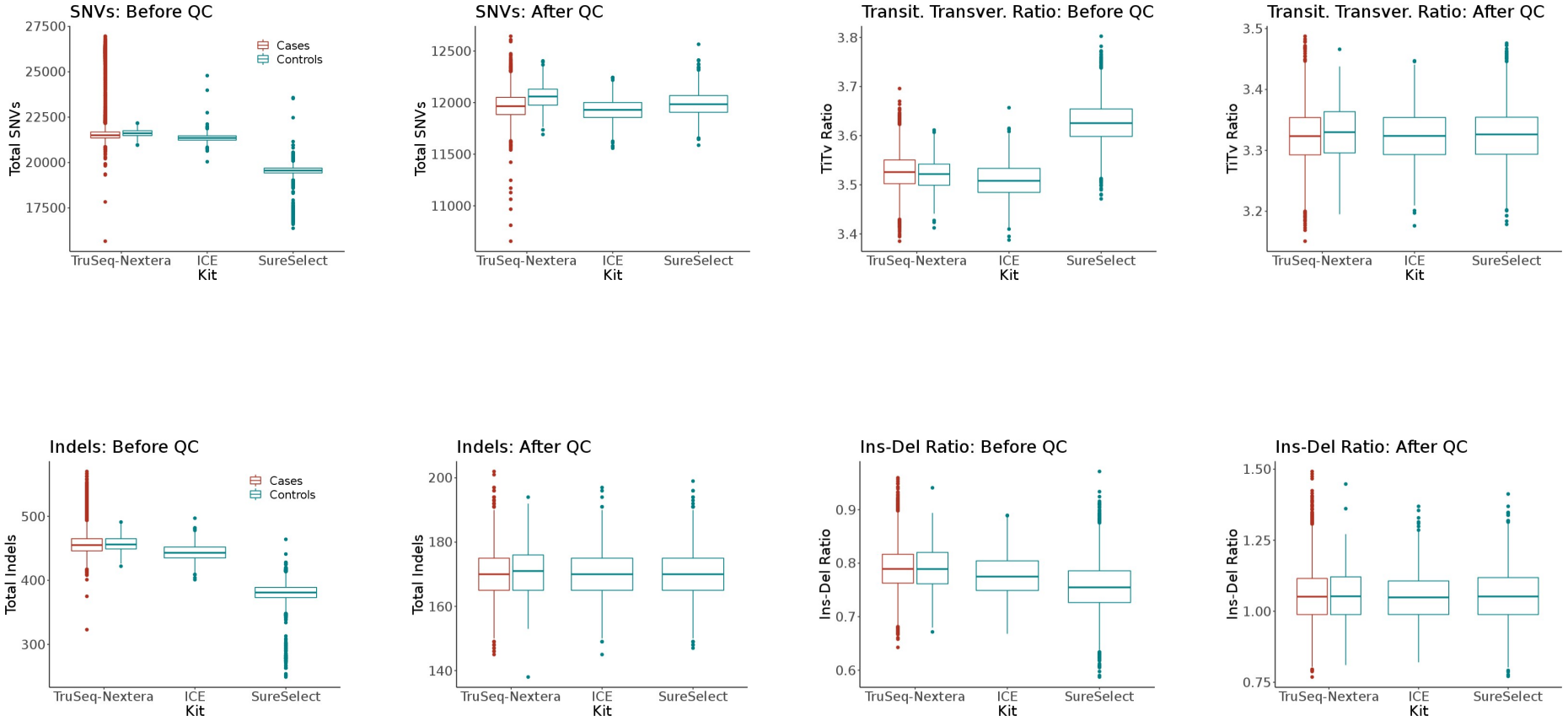


Figure 4.7: Variant counts and calling metrics in the final sample set. Quality control and coverage harmonization processes ensured inclusion of variants with adequate coverage across capture kits, eventually minimizing the possibility of spurious outcomes from differences between capture kits.

4.3.4 URVs classes

URVs were categorized further into multiple classes based on their functional consequences and collapsed by gene as QVs. We considered twelve non-synonymous variant classes including protein-truncating variants (presumed loss-of-function) and multiple groups of missense variants (mix of neutral, loss-, and gain-of-function mechanisms) as well as a (thirteenth) synonymous control classes of variants (presumed neutral). The grouping of missense QVs in multiple (partially overlapping) classes focused on three perspectives: conventional *in silico* deleteriousness, constraint, and paralog conservation. It was based on multiple predictions, namely, PolyPhen-2[191] (PPh2), Sorting Intolerant From Tolerant[212] (SIFT), Missense Badness Polyphen and Constraint[213] (MPC), Missense Tolerance Ratio[138] (MTR), and Constrained Coding Regions[214] (CCR). While MPC and MTR scores are scaled down to individual missense alterations, CCR score aims to identify coding regions that are completely devoid of variation in the population. Additionally, we used para-Z-score for paralog conservation[215] since it has been proposed that most disease genes in humans have paralogs.[216]

The analyzed functional classes of variants (Table 4.6) were: (i) Benign missense variants: as predicted by PPh2 and SIFT. (ii) Damaging missense variants: as predicted by PPh2 and SIFT. (iii) Protein Truncating Variants (PTVs): pLoF variants that included stop-gained, start-lost, frameshift, splice-donor, and splice-acceptor variants. (iv) All functional variants: combined PTVs, in-frame indels, and damaging missense variants. (v) "MPC 1" missense variants: constrained missense with MPC score ≥ 1. (vi) "MPC 2" missense variants: highly constrained missense with MPC score ≥ 2 (enriched for *de novo* variants). (vii) "MTR ClinVar" missense variants: constrained missense with MTR score ≤ 0.825 which is the median for ClinVar variants not denoted as *de novo*. (viii) "MTR De Novo" missense variants: highly constrained missense with MTR score ≤ 0.565 which is the median for ClinVar *de novo* variants. (ix) "CCR 80" missense variants: highly constrained missense variants in regions with CCR score ≥ 80, with MPC score ≥ 1, and MTR score ≤ 0.825. (x) "paralog-non-conserved": missense variants located in sites not conserved across paralog genes as indicated by a para-Z-score ≤ 0. (xi) "paralog-conserved": missense variants located in sites conserved across paralog genes as indicated by a para-Z-score > 0. (xii) "paralog highly conserved": missense variants in highly conserved sites between paralog genes with para-Z-score ≥ 1. (xiii) "Synonymous" variants that served as a control class for inflation.

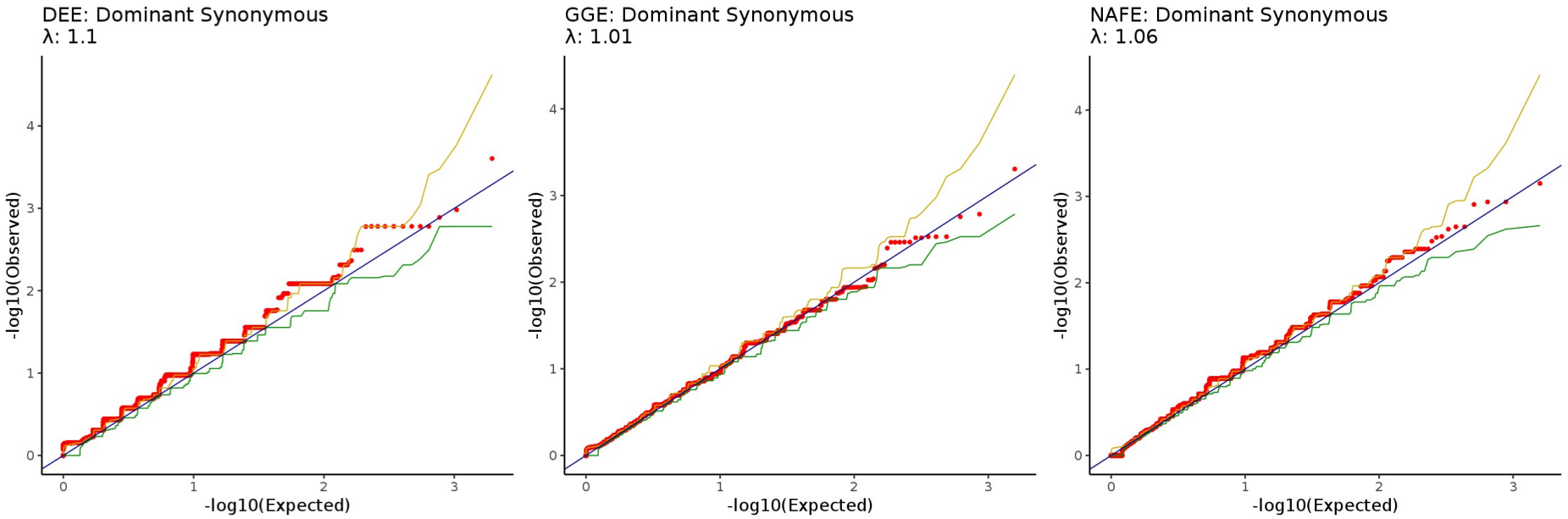


Figure 4.8: Quantile-Quantile plots of gene collapsing analysis of ultra-rare synonymous variants. Observed *p* values are obtained from testing the significance of the difference in the number of cases and controls carrying qualifying variants in each gene using Fisher's Exact Test. Expected *p* values are obtained from 1,000 permutations of case and control labels followed by Fisher's Exact Test. Grey area indicates 2.5th to 97.5th centiles of permutation *p* values. Genomic Inflation Factor estimates (λ) were calculated from a comparison of the observed and mean permutation *p* values. DEE: developmental and epileptic encephalopathies. GGE: genetic generalized epilepsies. NAFE: non-acquired focal epilepsies.

4.3.5 Gene sets

In total, ninety-two gene sets were tested. In addition to exome-wide burden testing (one gene set of all protein coding genes), we defined additional 91 specific gene sets as follows: (a) 34 sets based on gene expression patterns in the brain and genic intolerance;[217,218] (b) 28 functional groups including ion channels,[78] $GABA_A$ receptors,[47] excitatory receptors,[47] GABAergic pathway,[47] Postsynaptic Density protein 95 (PSD-95) interactors,[78] Gene Ontology (GO) gene sets of GABAergic and glutamatergic synapses,[219–221] neuronal pathways from Kyoto Encyclopedia of Genes and Genomes[222] (KEGG) and neuronal gene sets from Reactome[223] database; (c) 14 gene sets of known disease-related genes including monogenic epilepsy-causing genes,[33,47,78] epilepsy GWAS top-ranked genes from positional mapping within a window of 250 kb around significant loci plus mapping based on chromatin interaction between gene promoters and the significant locus,[60] and co-regulated genes in the brain;[224,225] and (d) 15 non-neuronal gene sets.[222] The gene sets are outlined in Table 4.7. The genes in each gene set are available as an online supplement.[226] The construction of gene sets leveraged multiple sources as detailed in the online supplementary. To ensure homogeneity between gene sets obtained from different sources and snpEff annotations used in this study, each gene set was limited to those genes annotated with snpEff as protein coding genes using Ensembl[133] gene identifiers on GRCh37.75. When available, Ensembl gene identifiers were obtained from sources of gene sets. Otherwise, *biomaRt*[227] package and gProfiler[228] were used to map HUGO [Human Gene Organization] Gene Nomenclature Committee (HGNC) gene names and gene name synonyms to their Ensembl gene identifiers. *biomaRt* was also used to map mouse genes to their human paralogues for two gene sets.

4.3.6 Gene set burden analysis

We examined the burden of QVs in thirteen variant classes (Table 4.6) for 92 gene sets (Table 4.7) in three epilepsy phenotypes (DEE, GGE, and NAFE) against a set of matched controls. Gene set burden testing was done using logistic regression by regressing case-control status on the individual QVs counts. In each sample, URVs that matched the specific analysis criteria were collapsed by gene into QVs (each sample was a assigned a status indicator: 1 for the presence of a QV or 0 for its absence) and these QVs were aggregated (summed per sample) across a target gene set to get a burden score (assuming equal weights and direction of effects) which was used as a predictor in a binomial model while adjusting for additional covariates (sex, top ten principal components, exome-wide variant count, and exome wide singletons

Table 4.7: Variant types and conditions used for the gene group burden analysis.

Variant conditions	Control		Deleterious variants			Missense constraint					Paralog conservation		
Effects (Sequence Ontology terms)	Synonymous	Benign Missense	Damaging Missense	PTV	All Functional	MPC1	MPC2	MTR ClinVar	MTR DeNovo	CCR 80	Paralog non-conserved	Paralog conserved	Paralog highly conserved
synonymous_variant	+	-	-	-	-	-	-	-	-	-	-	-	-
missense_variant (additional filters)	-	+	+	-	+	+	+	+	+	+	+	+	+
	-	PPh2 & SIFT benign	PPh2 & SIFT damaging	-	PPh2 & SIFT damaging	MPC ≥1	MPC ≥ 2	MTR ≤ 0.825	MTR ≤ 0.565	CCR ≥ 80 MPC ≥1 MTR ≤ 0.825	Para-Z-score ≤ 0	Para-Z-score > 0	Para-Z-score ≥ 1
stop_gained	-	-	-	+	+	-	-	-	-	-	-	-	-
splice_acceptor_variant	-	-	-	+	+	-	-	-	-	-	-	-	-
splice_donor_variant	-	-	-	+	+	-	-	-	-	-	-	-	-
exon_loss_variant	-	-	-	+	+	-	-	-	-	-	-	-	-
frameshift_variant	-	-	-	+	+	-	-	-	-	-	-	-	-
start_lost	-	-	-	+	+	-	-	-	-	-	-	-	-
stop_lost	-	-	-	-	+	-	-	-	-	-	-	-	-
conservative_inframe_insertion	-	-	-	-	+	-	-	-	-	-	-	-	-
disruptive_inframe_insertion	-	-	-	-	+	-	-	-	-	-	-	-	-
conservative_inframe_deletion	-	-	-	-	+	-	-	-	-	-	-	-	-
disruptive_inframe_deletion	-	-	-	-	+	-	-	-	-	-	-	-	-

count) using *glm*() function from *stats* package.[193]

We presumed equal weights and direction of effects for the variants in the classes under analysis by taking the sum of QVs in a specific gene set per sample as a predictor for a binary phenotype in a regression model. While this assumption is reasonable for highly deleterious variants, it is rather simplistic for milder genetic alterations. This approach is also not ideal to estimate the odds in data sets with low counts. However, the computational ease, the clarity in setting up the analysis parameters in comparison to other variance componence-based and hybrid methods, e.g., skat-o,[229] are key advantages that motivated this choice. The use of similar regression models has been shown to capture the major signals in gene set burden analysis in epilepsy and other neurological diseases.[33,78,106] Likelihood ratio test (LRT) from *lmtest* package[230] was used to compare a model with QVs burden and covariates as predictors against a null model (covariates only). The null model was *glm*(sex + variant counts + singletons + PC1…PC10) and the test model was *glm*(QV burden + sex + variant counts + singletons + PC1… PC10). Log-odds from LRT and their respective 95% confidence intervals and *p* values are presented here as a measure of enrichment in tested gene sets. Multiple gene sets were tested in parallel using *parallel* package.[193]

We employed a Benjamini-Hochberg false discovery rate (FDR) multiple testing adjustment for *p* values that accounted for 3,312 tests (92 gene sets x 3 epilepsy phenotypes x 12 non-synonymous variant classes, excluding synonymous variants used as a control class) as implemented in *p.adjust*(method = "BH") function from *stats* package.[193] The *p* values from the analysis of the synonymous class of variants were not FDR-adjusted, similar to previous analysis approaches.[33] The cut-off for substantial enrichment was defined as FDR-adjusted *p* value < 0.05. For simplicity, *p* values (FDR-adjusted except for synonymous variants) are indicated throughout the presented plots using stars as follows: no star > 0.05, * < 0.05, ** < 0.005, *** < 0.0005, **** < 0.00005. Data handling steps were performed in R v3.3 using R base, *data.table* and *tidyverse* packages.[193,195,196]

4.3.7 Secondary analysis

Four secondary analyses were performed to explore the extent of the observed differences between GGEs and NAFEs and to exclude potential bias (e.g., introduced by the imbalance in male-to-female ratios between cases and controls or residual differences in variant coverage and quality metrics).

1. A secondary analysis was performed on the ninety-two gene sets but limited to autosomal genes (excluding all genes on chromosome X). The aim was to estimate the bias created by male-to-female ratios imbalance.
2. Another secondary analysis was performed using MIGen Leicester samples (Illumina ICE capture kits) as cases vs. MIGen Ottawa/ATVB samples as controls (Agilent SureSelect capture kits) to exclude the presence of significant residual stratification between capture kits. Comparisons between samples prepared using Illumina Nextera/TruSeq and Illumina ICE or Agilent SureSelect were not performed as these are almost identical to the primary analysis of epilepsy cases (Nextera/TruSeq) vs. controls (ICE & SureSelect) analysis.
3. Randomly selected GGEs (n=1,100) and controls (n=2,789) were tested to examine if these numbers are enough to capture the main signals, to confirm the validity of the control-control testing. We did five hundred permutations, using the CCR80 class of variants, taking the mean of the odds, 2.5^{th}/97.5^{th} centiles of odds and average *p* values per tested gene set as an outcome of this permutation analysis. The random selection of samples and final summarization of outcomes was done using R base functions.
4. A limited secondary analysis directly comparing the CCR80 class of variants between individuals with GGE and NAFE to validate the patterns observed in case vs. control comparisons.

Table 4.7: Gene sets investigated in a study of ultra-rare variants burden in epilepsy.

The number of gene sets in each category is given in parenthesis.

Group of all protein coding genes (1):		
-all genes annotated by snpEff as protein coding.		
Groups based on brain expression (34): Expression in the brain, regional, cellular, and sub-cellular expression patterns.		
Brain-expressed LOF-intolerant genes: excluding genes with no expression in the cortex/hippocampus -pLI > 0.995. -pLI 0.9-0.995. -pLI 0.8-0.9.	Cortical and hippocampal expression level: -High, Moderate, Low in the cortex. -High, Moderate, Low in the hippocampus. Brain development: -Brain development genes (Gene-Ontology group). -Brain developmental genes (extended group). -Early developmental genes. -Late developmental genes.	Cell-type-specific enrichment: -Neurons -glial cells. -Excitatory neurons. -Inhibitory neurons. -Astrocytes. -Microglia -Oligodendrocytes. -Endothelium. Neuronal Localization: -Axon Initial Segment. -Synaptic (curated group). -Synaptic (extended group).
Brain-expressed missense-intolerant genes: excluding genes with no expression in the cortex/hippocampus		

-Z-score > 3.09. -Z-score 2.5-3.09. -Z-score 2-2.5.	Enrichment in the brain: -Brain-enriched. -Brain-enhanced.	-Synaptic vesicle and active zone. -Pre-synaptic. -Post-synaptic. -Pre-synaptic only. -Post-synaptic only.
Functional gene sets (28): Ion channels, transporters, synaptic cycles, pathways and neurotransmitter cycles.		
Ion channels, neurotransmitter receptors and related genes: -Voltage-gated ion channels. -Voltage-gated cation channels. -Brain-specific voltage-gated ion channels. -$GABA_A$ receptors. -GABAergic pathway. -Excitatory receptors. -NMDAR & ARC. -PSD-95 interactors. GABAergic/Glutamatergic synapses (GO groups): -GABAergic synapse. -Glutamatergic synapse. -only in GABAergic. -only in glutamatergic. -shared genes.	GABAergic/Glutamatergic pathways (KEGG database): -GABAergic pathway. -Glutamatergic pathway. -only in GABAergic. -only in glutamatergic. -shared genes. Additional neuronal pathways (KEGG): -Cholinergic pathway. -Dopaminergic pathway. -mTOR pathway. -Synaptic vesicle cycle.	GABA/glutamate cycles (Reactome database; pooled from multiple groups): -GABA release, receptor activation, and clearance -Glutamate release, uptake, and clearance cycle. Additional neuronal groups (Reactome database): -Presynaptic depolarization. -Neurexins and Neuroligins. -Synaptic Adhesion molecules. -Receptor-type Protein Tyrosine Phosphatases.
Disease-associated and intolerant genes (14): Genes and gene sets with known associations with epilepsy and related neurological diseases		
Monogenic disease-causing genes: -Generalized epilepsy genes. -Focal epilepsy genes. -Dominant epilepsy genes -DEE genes. -NDD with epilepsy genes. -FMRP targets. -MGI seizure genes.	Top-ranking 100 genes in ILAE2 GWAS: -Generalized epilepsy GWAS. -Focal epilepsy GWAS. -All epilepsies GWAS. Brain co-expression module: -Co-expressed module identified in non-diseased post-mortem brain tissues. (enriched for *de novo* variants in DEE).	Regulatory and co-expression modules in epilepsy: -Co-expression network identified in brain tissues of Temporal Lobe Epilepsy patients. -Two modules within this network.
Control groups (15):		
Genes not expressed in the brain: -RNA not detected in cortex, in hippocampus, or all GTEx regions. -Protein is depleted in the brain.	KEGG metabolic pathways: -Type II Diabetes. -Carbohydrate Absorption & Digestion. -Protein Absorption & Digestion. -Fat Absorption & Digestion.	KEGG cancer pathways: -CA Breast, CA Lung, CA Colon, CA Prostate, Renal Cell Ca, CA Pancreas, Hepatocellular Ca.

4.4 Results

4.4.1 URVs excess in brain-expressed genes

The distribution of benign missense variants reflected the ancestry groups of our samples with a slight excess of synonymous variation in the controls (Figure 4.9). Missense variants in intolerant sites were in excess in the cases (Figures 4.10 and 4.11). The use of a combination of three intolerance metrics (to identify highly deleterious variants in functionally critical genic regions) showed a considerable difference between the cases and controls; about half of the cases, in contrast to roughly one-fourth of controls, harbored one or more QVs in highly constrained regions (Figure 4.12). These differences in variant burden were examined further using binary logistic regression.

First, we investigated the burden of URVs across all protein coding genes. This revealed a clear enrichment in constrained missense variants that was maximum in consensus constrained coding regions predicted by MPC, MTR and CCR scores (Figure 4.13A). A previous similar analysis of this and related datasets[33,34] examined loss-of-function intolerant genes and demonstrated an increased burden in ultra-rare constrained as well as PTVs. Here, the examination of brain-expressed intolerant genes showed, similarly, a marked enrichment in PTVs in addition to a burden in highly constrained missense variants that is comparable to what is seen exome-wide (Figures 4.13B and 4.14).

When we examined protein coding genes grouped by their relative brain expression, damaging missense variants were only substantially enriched in genes highly expressed in the cortex or hippocampus, whereas those expressed at medium or low levels only showed an enrichment for the most constrained missense variants (Figure 4.15A). Genes showing a higher expression in the adult brain compared to other tissues (brain-enriched & brain-enhanced) were also preferentially enriched, as well as genes associated with brain development. Genes related to late rather than early development showed a slightly higher enrichment in all three phenotypic groups (Figure 4.15B).

4.4.2 Burden in neuronal genes and pathways

Focusing further on cell-type specific expression, we found that neuron-specific genes were preferentially affected compared to those enriched in glial cells, particularly in GGE (Figure 4.16A). To obtain further insight into the nature of this neuronal enrichment, we used sets of genes representing paralogs of mouse genes found to be enriched in excitatory or inhib-

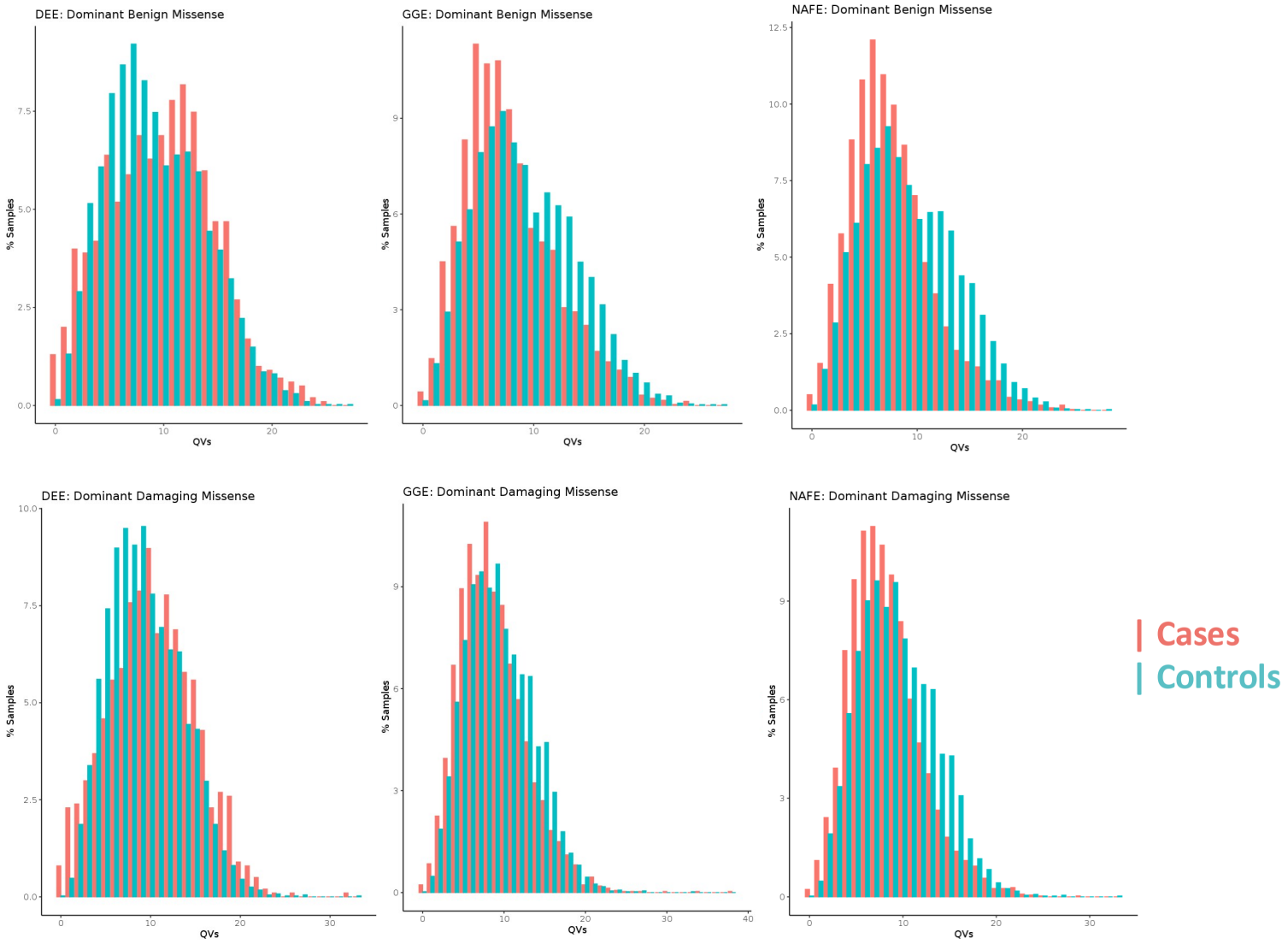


Figure 4.9: Distribution of benign and damaging missense qualifying variants in cases and controls. Plots from the analysis of benign (top) and damaging (bottom) missense variants are shown. DEE: developmental and epileptic encephalopathies. GGE: genetic generalized epilepsies. NAFE: non-acquired focal epilepsies.

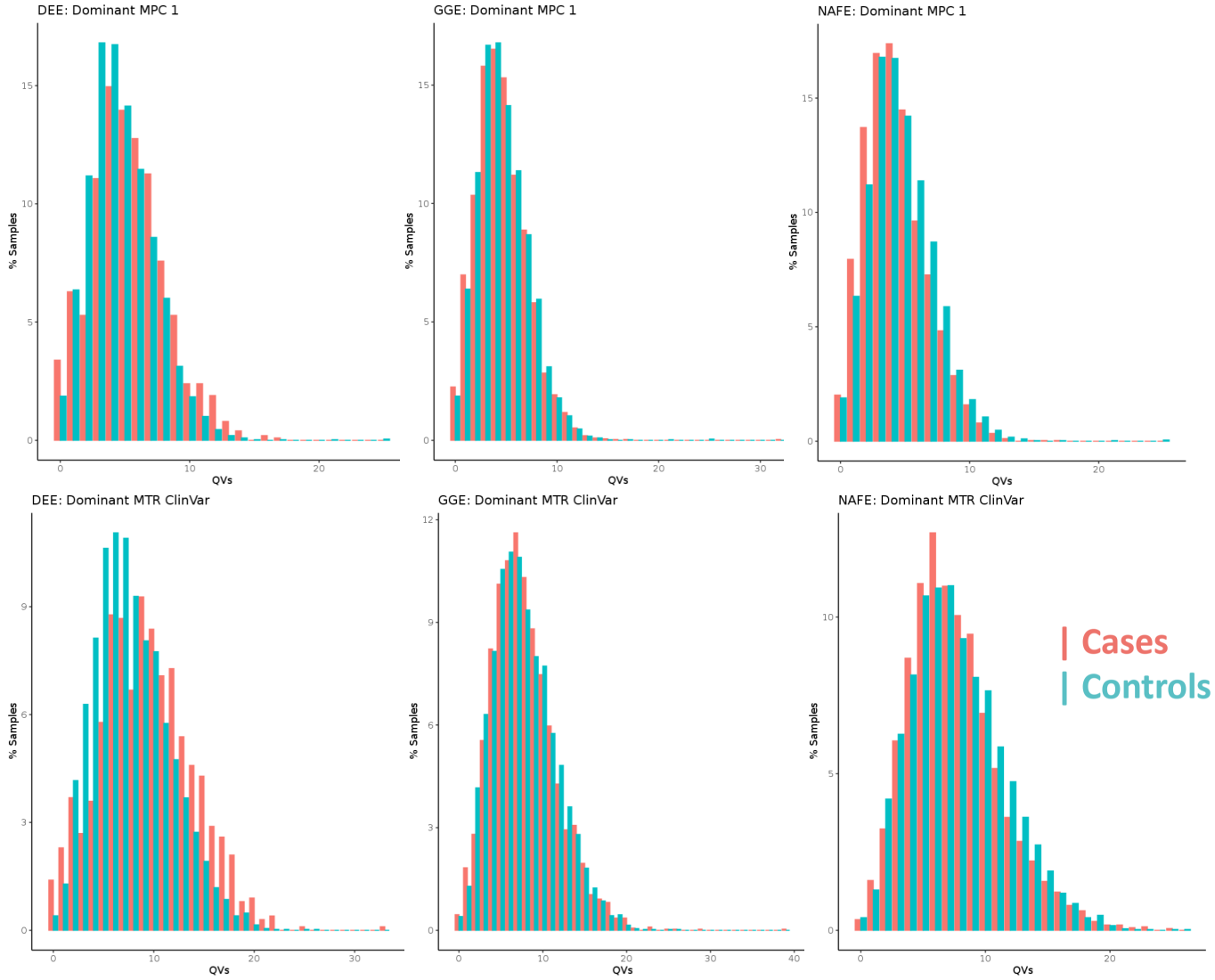


Figure 4.10: Distribution of missense qualifying variants in moderately constrained sites in cases and controls. Plots from the analysis of missense variants in moderately constrained sites are shown (top: MPC 1 class, bottom: MTR ClinVar class). DEE: developmental and epileptic encephalopathies. GGE: genetic generalized epilepsies. NAFE: non-acquired focal epilepsies.

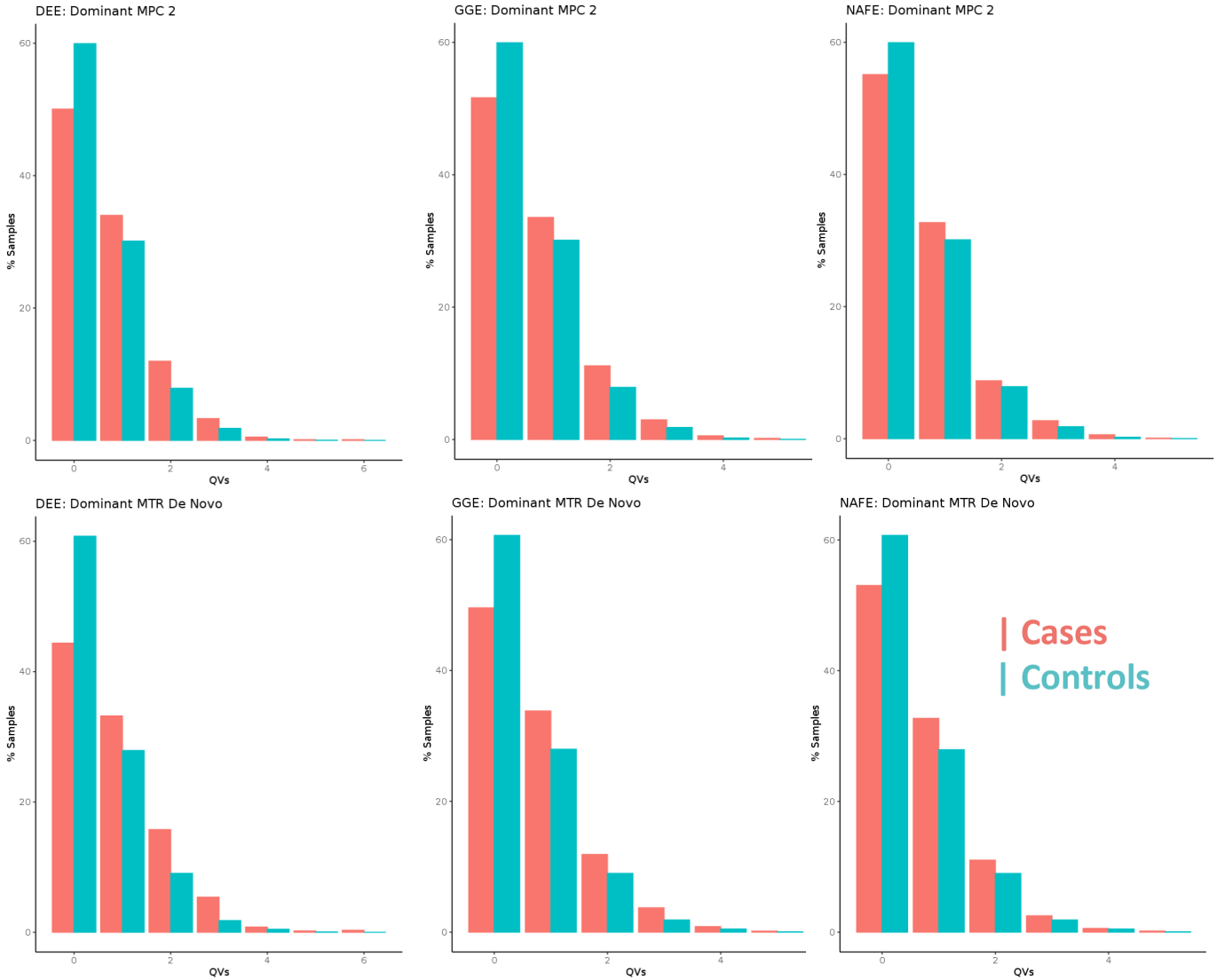


Figure 4.11: Distribution of missense qualifying variants in highly constrained sites in cases and controls. Plots from the analysis of missense variants in highly constrained sites (top: MPC 2 class; bottom: MTR De Novo class). DEE: developmental and epileptic encephalopathies. GGE: genetic generalized epilepsies. NAFE: non-acquired focal epilepsies.

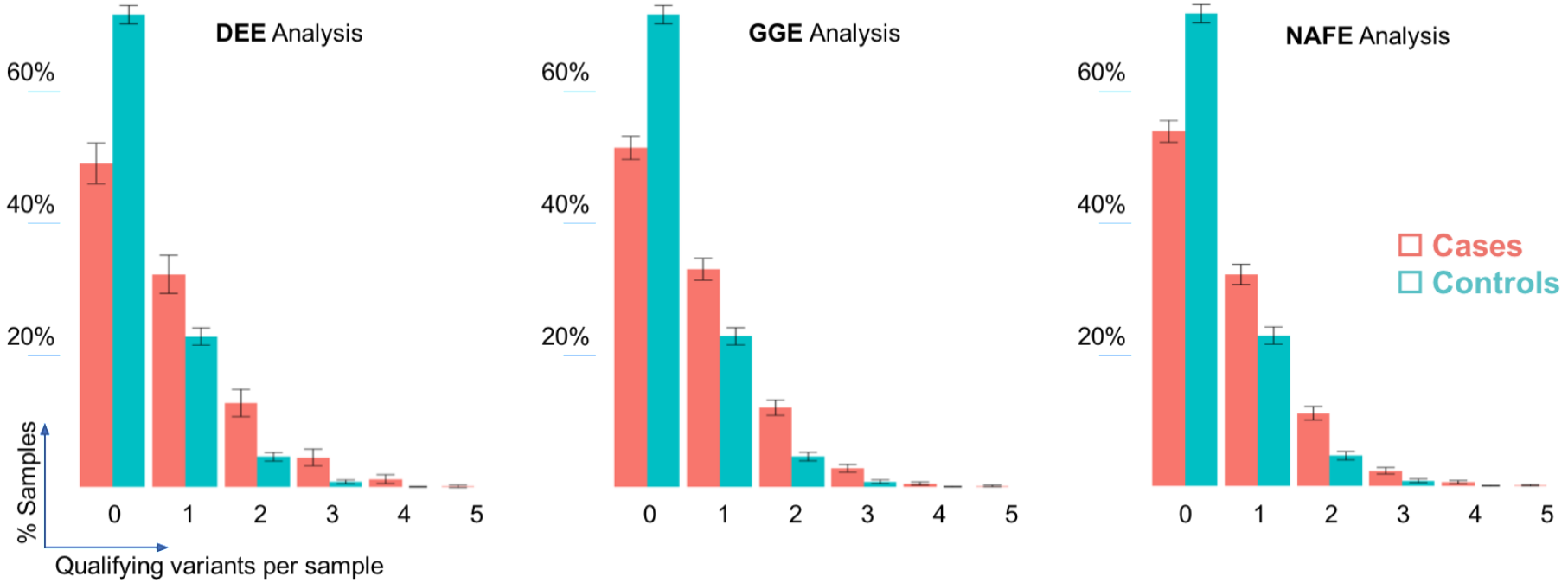


Figure 4.12: Distribution of missense qualifying variants in constrained coding regions (CCR > 80) in cases and controls. Roughly, half of the cases compared to one fourth of the controls harbor one or more qualifying variant per exome in highly constrained sites. Error bars indicate the 95% confidence intervals calculated as follows: ! $\pm 1.96\times\sqrt{!\,(1-!\,)/*}$ where p is the proportion of samples and n is the total number of samples. DEE: developmental and epileptic encephalopathies. GGE: genetic generalized epilepsies. NAFE: non-acquired focal epilepsies.

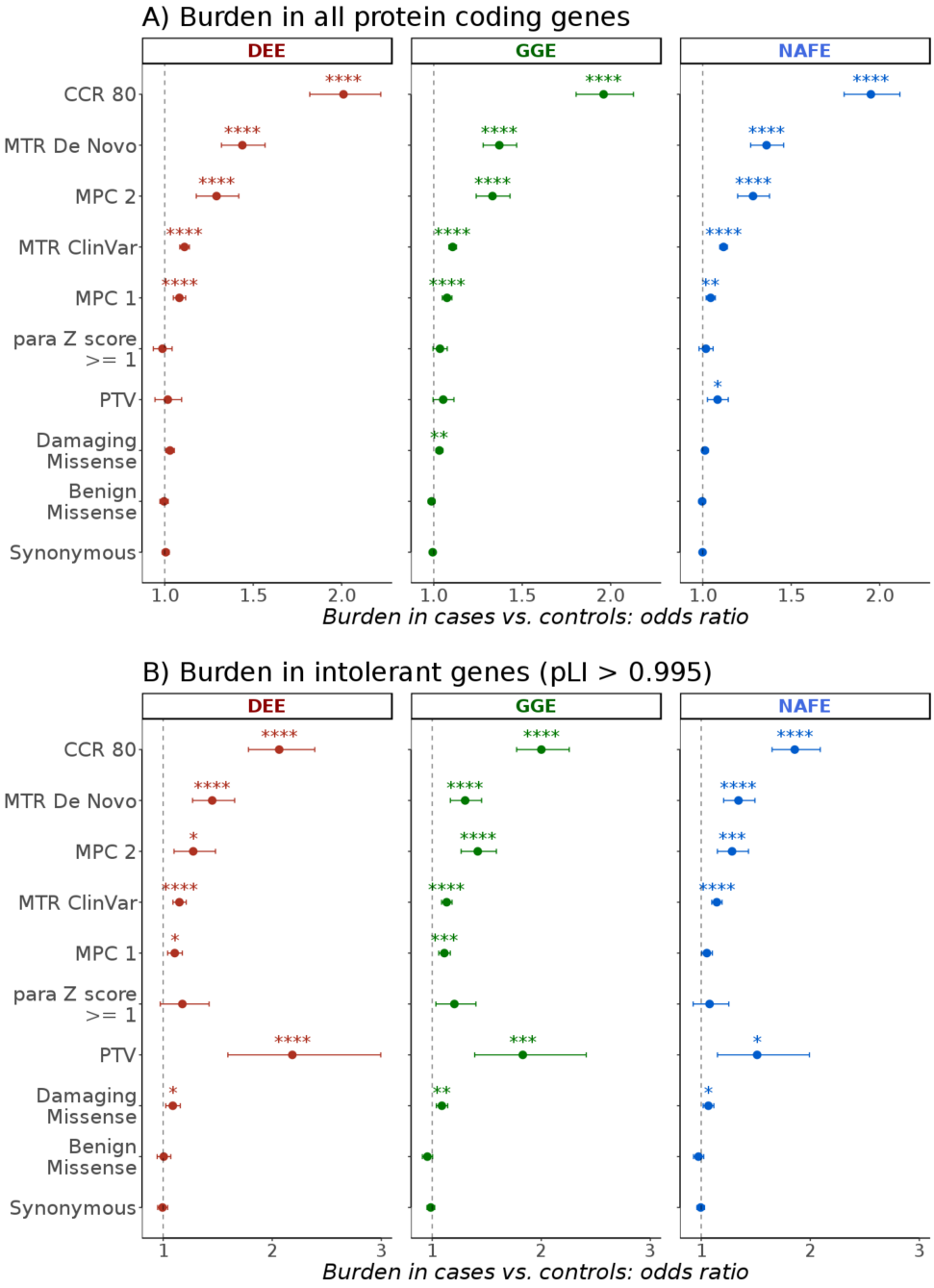


Figure 4.13: Exome-wide burden of ultra-rare variants in the epilepsies. The burden in developmental and epileptic encephalopathies (DEE), genetic generalized epilepsies (GGE) and non-acquired focal epilepsies (NAFE) in **(A)** 19,402 protein coding genes and **(B)** 1,743 genes with probability of loss-of-function intolerance (pLI) score > 0·995 is shown in multiple classes of variants (y-axis; see methods) as odds ratio (x-axis) from Likelihood Ratio Test (bars indicate 95% confidence intervals). False-Discovery-Rate-adjusted *p* values (synonymous variants analysis *p* values were not adjusted) are indicated with stars as follows: no star > 0.05, * < 0.05, ** < 0.005, *** < 0.0005, **** < 0.00005.

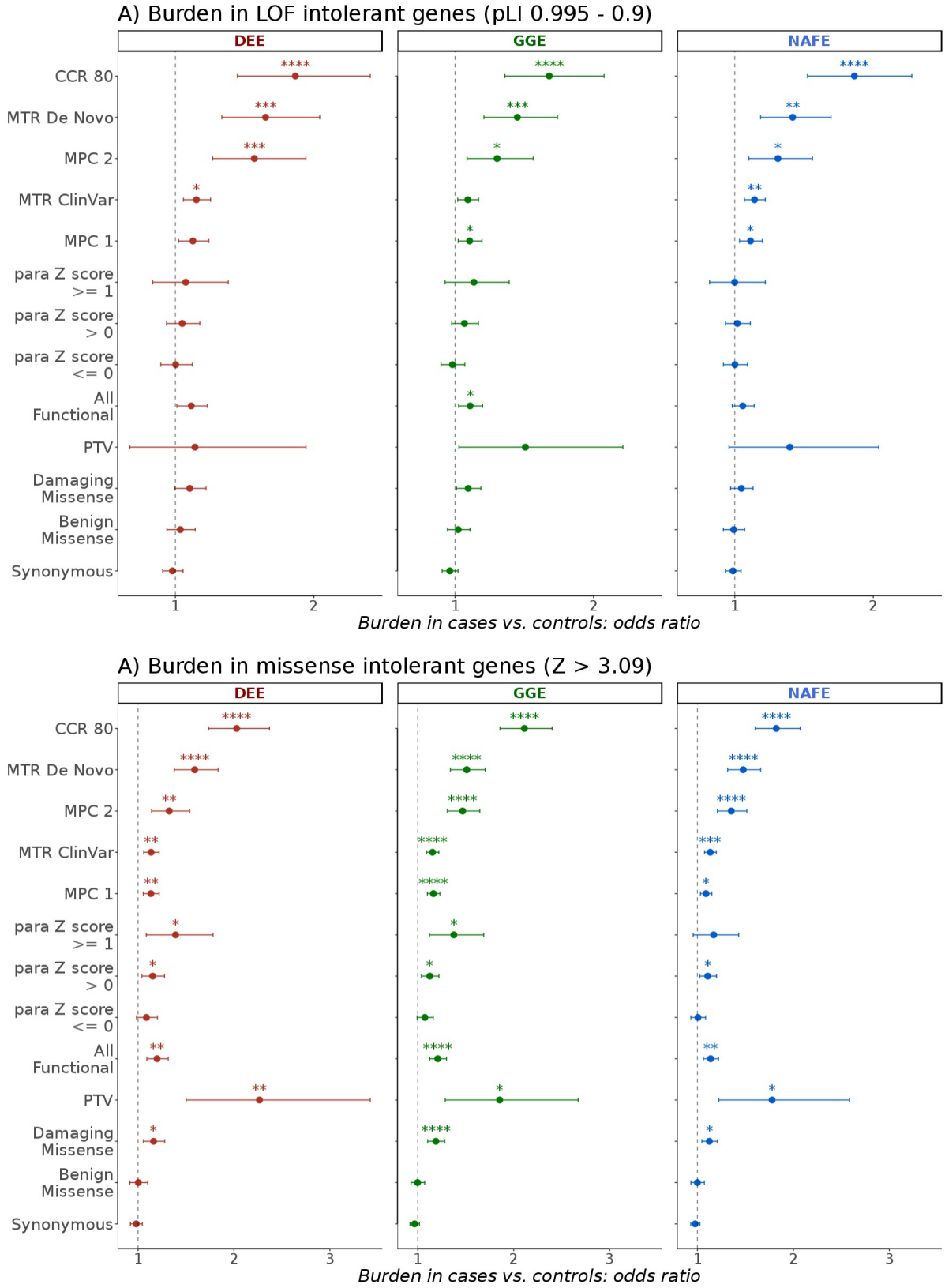


Figure 4.14: Burden of ultra-rare variants in intolerant genes. y axis: variant classes. x axis: odds ratio from regression analysis of individual burden of qualifying variants. Stars indicate False Discovery Rate-adjusted p values: * < 0.05, ** < 0.005, *** < 0.0005, **** < 0.00005. Error bars indicate 95% confidence intervals of odds. DEE: developmental and epileptic encephalopathies. GGE: genetic generalized epilepsies. NAFE: non-acquired focal epilepsies. pLI: probability of loss-of-function intolerance.

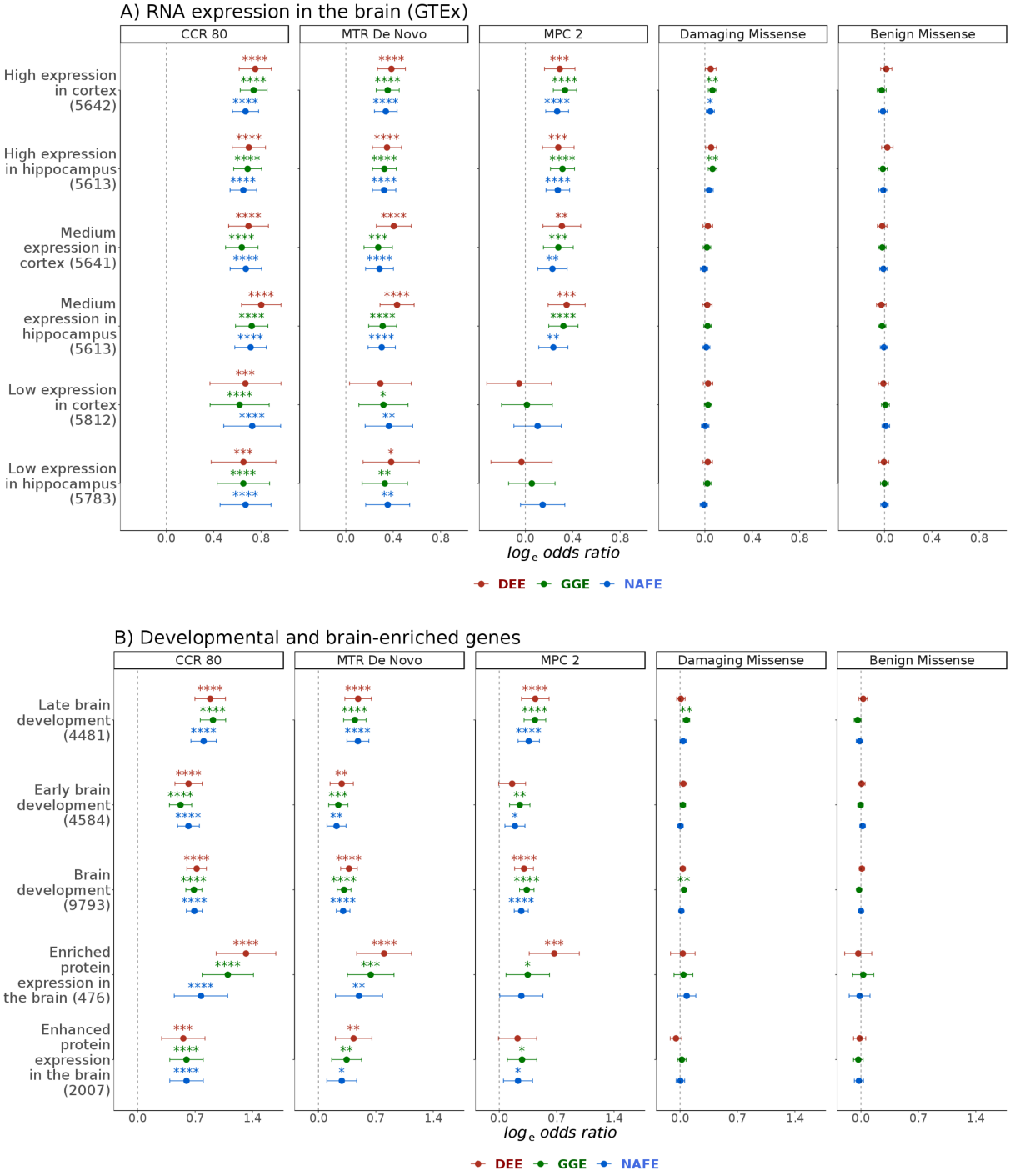


Figure 4.15: Burden of ultra-rare missense variants in brain expressed and developmental genes. The burden of benign or damaging missense variants and missense variants in highly paralog-conserved or highly constrained sites in developmental and epileptic encephalopathies (DEE), genetic generalized epilepsies (GGE) and non-acquired focal epilepsies (NAFE) is shown in gene sets based on levels of RNA/protein expression in the cortex and hippocampus **(A)** or enrichment in adult or developing brain **(B)**. gene sets are shown on the *y*-axis (number of genes between brackets). Log odds ratio (Likelihood Ratio Test) are shown on the *x*-axis (error bars indicate 95% confidence intervals). The variant conditions are shown in vertical panels. False-Discovery-Rate-adjusted *p* values (synonymous condition not adjusted) are indicated with stars as follows: no star > 0.05, * < 0.05, ** < 0.005, *** < 0.0005, **** < 0.00005. High, medium and low expression was based on expression levels in Gene Tissue Expression Project portal (GTEx). Brain-enriched (with more than four-fold expression in the brain compared to other tissues) and brain-enhanced genes (higher but less than four-fold expression) were obtained from the Human Protein Atlas.

itory neurons. Interestingly, genes preferentially expressed in inhibitory neurons showed an increased burden only in GGE, whereas those preferentially expressed in excitatory neurons showed a more prominent signal in NAFE. Next, we examined functional gene sets that could, more specifically, underlie the observed enrichment in neuronal and synaptic genes. Ion channels, neurotransmitter receptors and transporters are widely implicated in epilepsy, especially in monogenic and familial forms, displaying considerable phenotypic heterogeneity and presenting as mild or severe epilepsies.[70,71,231] Variants in $GABA_A$ receptors were enriched in GGE but not in DEE or NAFE while those in gene sets representing genes encoding N-Methyl-D-Aspartate receptor and Activity-Regulated Cytoskeleton protein[78] (NMDAR-ARC) interactors were enriched in NAFE and DEE. A comprehensive gene set for the GABAergic pathway genes[47] showed a prominent signal in GGE and DEE, and less in NAFE. In contrast, a gene set representing PSD-95 interactors showed comparable enrichment in NAFE and GGE (Figure 4.16B). Brain-expressed ion channels were found to be enriched for highly constrained missense variants (CCR 80 class of variants) in common as well as rare epilepsies (Figure 4.16B).

Since well-established epilepsy genes, like ion channels and receptors, show differential distributions in different neuronal compartments,[232,233] we examined further sets of genes based on subcellular localization. We found that pre- and postsynaptic genes were enriched with variants in cases vs. controls, as well as a small set of 17 genes located in axon initial segments (most prominent in DEE) (Figure 4.17A). Genes encoding neurexins and neuroligins, important elements of pre- and post-synaptic interaction promoting adhesion between dendrites and axons,[234] were enriched in DEE (Figure 4.17B). Also, the synaptic vesicle cycle pathway (KEGG) showed a prominent signal in both DEE and GGE. We also examined the burden in the mTOR pathway (KEGG), hypothesizing that it could have potential relevance to focal epilepsies, but did not detect a substantial enrichment (Figure 4.17B). Interestingly, NAFE analysis displayed a burden in endothelial and astrocyte-specific genes in highly constrained genic regions (Figure 4.16A).

4.4.3 Gene sets representative of excitatory and inhibitory signaling pathways

We then compared the patterns of URVs burden in genes involved in the GABAergic (main inhibitory) pathway and synapse against those in the glutamatergic (main excitatory) pathway and synapse in the brain, by examining their unique and overlapping genes based on KEGG pathways[222] or GO synaptic gene sets[220] and sets of specific receptors (Figure 4.18).

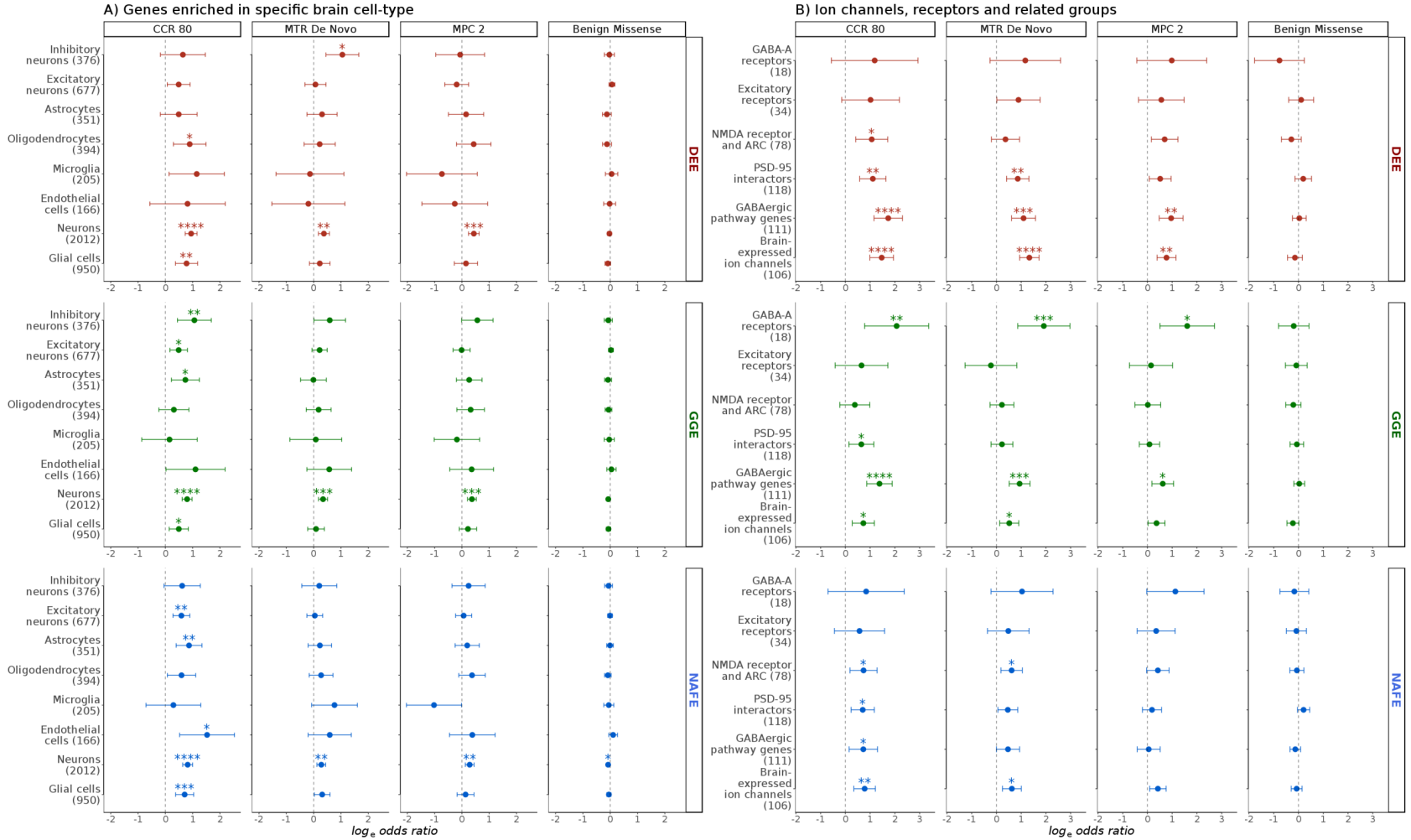


Figure 4.16: Burden in neuronal and glial cells, ion channels, receptors and related interactors. The burden in developmental and epileptic encephalopathies (DEE), genetic generalized epilepsies (GGE) and non-acquired focal epilepsies (NAFE) in shown on the *x*-axis (log-odds from Likelihood Ratio Test; error bars indicate 95% confidence intervals). Gene sets are shown on the *y*-axis (number of genes between brackets). The variant conditions are shown in vertical panels. False-Discovery-Rate-adjusted *p* values (synonymous condition not adjusted) are indicated with stars as follows: no star > 0.05, * < 0.05, ** < 0.005, *** < 0.0005, **** < 0.00005. **(A)** Burden in genes enriched in specific brain cells including neuron- or glia-enriched genes and their subtypes. **(B)** Burden in key biologically informed neuronal gene sets with known or suspected relation to epilepsy. NMDA: N-Methyl-Dextro-Aspartate. ARC: neuronal activity-regulated cytoskeleton-associated protein. PSD-95: Post-Synaptic-Density protein 95.

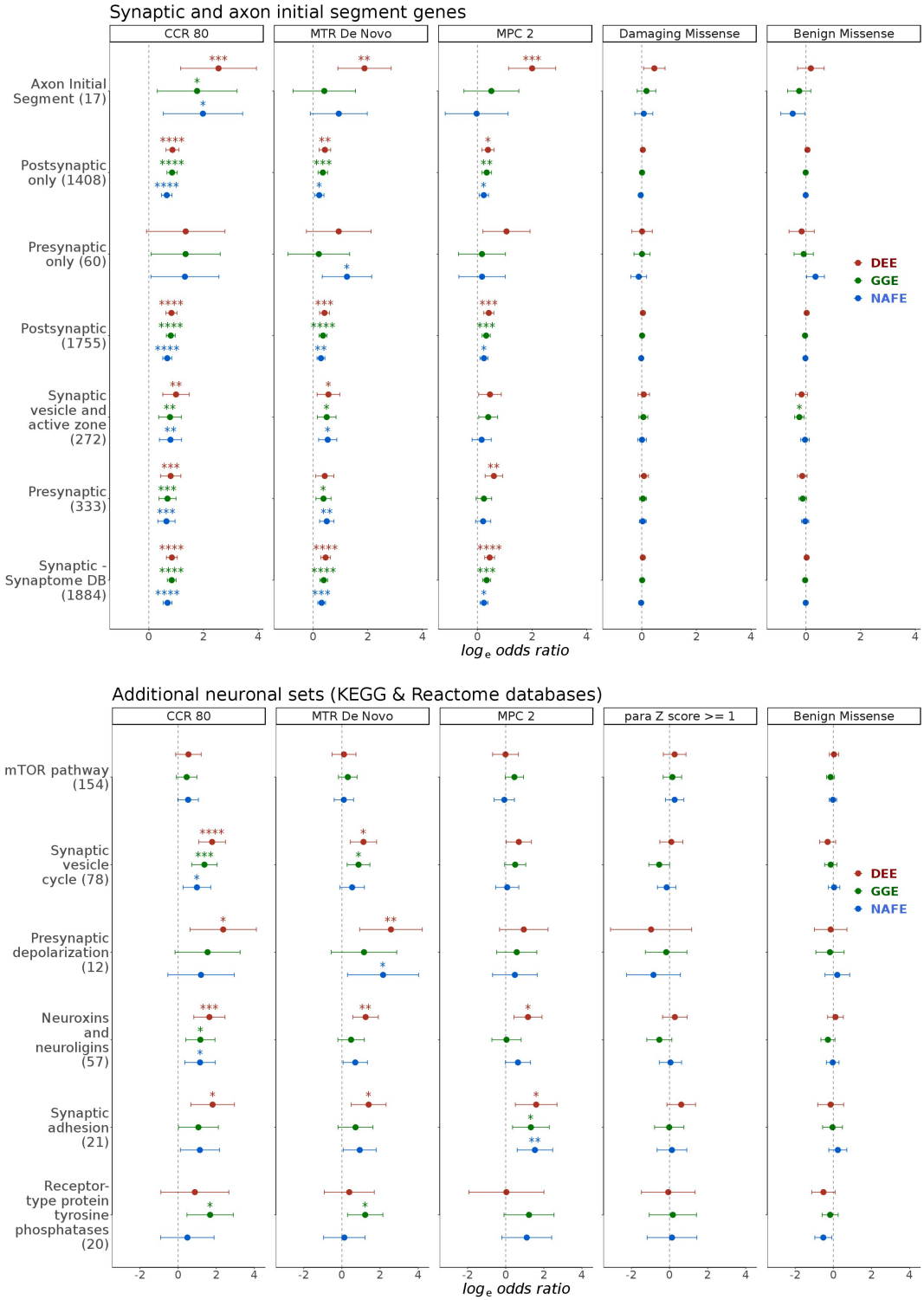


Figure 4.17: Burden in groups of axon initial segment genes, synaptic genes and additional neuronal gene sets. Panels: variant classes. *y* axis: gene sets (genes count between parenthesis). *x* axis: log odds ratio from regression analysis of individual burden of qualifying variants. Stars indicate False Discovery Rate-adjusted *p* values: * < 0.05, ** < 0.005, *** < 0.0005, **** < 0.00005. Error bars indicate 95% confidence intervals of odds. DEE: developmental and epileptic encephalopathies. GGE: genetic generalized epilepsies. NAFE: non-acquired focal epilepsies.

GGE showed a higher burden in GABAergic vs. glutamatergic synapse (GO) and pathway (KEGG) genes, in $GABA_A$ receptors vs. excitatory receptors/NMDAR-ARC interactors genes, and in GABAergic pathway genes (comprehensive gene set) vs. PSD-95 interactors, thus matching the higher burden in genes representing inhibitory vs. excitatory neuronal signaling. The CCR 80 analysis of GO gene sets in NAFE showed a higher burden in glutamatergic vs. GABAergic synapse genes, akin to the pattern seen in genes enriched in excitatory vs. inhibitory neurons. The analysis of KEGG glutamatergic vs. GABAergic pathway genes did not confirm this finding (Figure 4.18B).

Altogether, these comparisons of the burden in missense variants in highly constrained sites between GGE and NAFE (Figures 4.16 and 4.18) suggest the following patterns: (i) brain-expressed ion channels, genes enriched in excitatory neurons, enriched in astrocytes, PSD-95 interactors, GABAergic and glutamatergic synapse/pathway genes show an increased burden in cases vs. controls both in GGE & NAFE; (ii) in GGE, this enrichment is coupled with a stronger enrichment in inhibitory neuronal genes, in $GABA_A$ receptors and in GABAergic synapse-specific genes (higher burden in inhibitory vs. excitatory gene sets); and (iii) in NAFE, this is accompanied by an absence of enrichment in the later gene sets and increased burden in the NMDAR-ARC gene set (higher burden in excitatory vs. inhibitory gene sets). A direct comparison of GGEs vs. NAFEs supported the observation of a substantially higher burden of highly constrained variants (CCR 80 class of missense variants) in GABAergic pathway genes in GGEs (Figure 4.19).

4.4.4 Burden in top GWAS hits, co-expression modules and known epilepsy-related genes

Recent efforts from the ILAE consortium on complex epilepsies identified multiple associations in a large GWAS of common epilepsies.[60] To examine the hypothesis that genes located near the top GWAS hits are also affected by rare variants, we tested the enrichment in sets of the 100 top-ranked genes derived from the ILAE GWAS in generalized, focal, and all epilepsies. Interestingly, when limiting the analysis to Consensus Coding Regions (CCR80 class of variants), top-ranked genes derived from the GWAS of either generalized or focal epilepsies were preferentially enriched for rare variants in the respective phenotypic groups of GGE and NAFE (Figure 4.20A). Although the observed enrichment was rather subtle, this result was corroborated by a similar pattern for two, rather small, sets of known epilepsy genes that are predominantly associated with either generalized or focal epilepsy.[47]

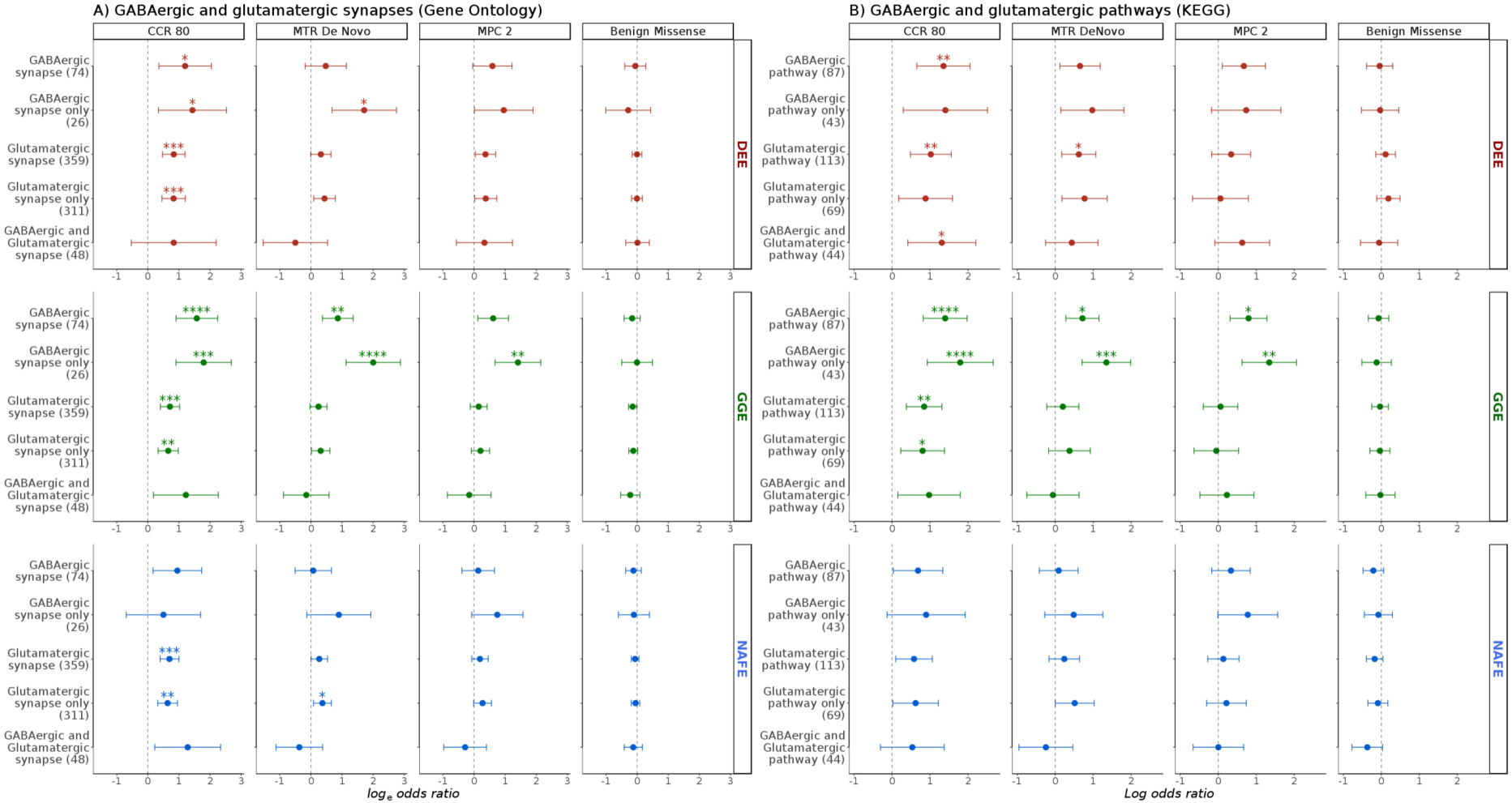


Figure 4.18: Enrichment in major neuronal synapses and pathways. Panels show comparison of enrichment patterns in developmental and epileptic encephalopathies (DEE), genetic generalized epilepsies (GGE) and non-acquired focal epilepsies (NAFE) in GABAergic and glutamatergic synapses and pathway genes based on **(A)** Gene Ontology (GO) and **(B)** Kyoto Encyclopedia for Genes and Genomes (KEGG). The burden is shown on the *x*-axis (log-odds from Likelihood Ratio Test; error bars indicate 95% confidence intervals). Gene sets are shown on the *y*-axis (number of genes between brackets). The variant conditions are shown in vertical panels. False-Discovery-Rate-adjusted *p* values (synonymous condition not adjusted) are indicated with stars as follows: no star > 0.05, * < 0.05, ** < 0.005, *** < 0.0005, **** < 0.00005. Complete groups, genes specific to one of the two synapses/pathways as well as their intersection were tested.

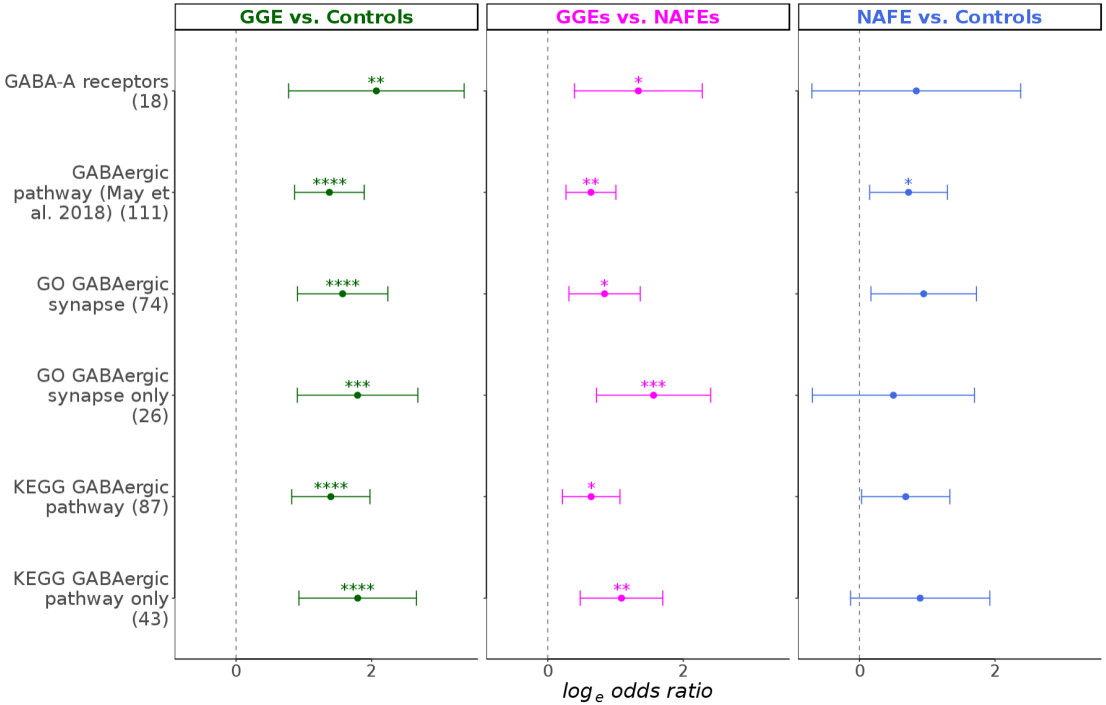


Figure 4.19: Gene sets with substantial differences in URVs burden in a direct comparison of GGEs vs. NAFEs. All gene sets with p values < 0.01 (corresponds to a False Discovery Rate (FDR)-adjusted p value of 0.05 in the primary analysis) are shown. Panels: variant classes. y axis: gene sets (genes count between parenthesis). x axis: log odds ratio from regression analysis of individual burden of qualifying variants. Stars indicate FDR-adjusted p values: * < 0.05, ** < 0.005, *** < 0.0005, **** < 0.00005. Error bars indicate 95% confidence intervals of odds. GGE: genetic generalized epilepsies. NAFE: non-acquired focal epilepsies.

We also aimed to touch upon the role of brain co-expression modules identified in post-mortem brain tissues from healthy individuals[225] and contrast these to the networks and modules identified in brain tissue derived from epilepsy patients.[224] A brain expression module was found to be substantially enriched for rare deleterious variants in an independent cohort of DEE.[225] A link to common epilepsy phenotypes was also inferred, but a burden in URVs was not examined so far. This module showed a non-specific enrichment in all three epilepsy subtypes with highest odds in DEE. In resected hippocampi of individuals with Temporal Lobe Epilepsy (TLE), Johnson and colleagues identified two co-expression modules within a gene-regulatory transcriptional network.[224] A subtle enrichment was seen in these modules in DEE and GGE, but not NAFE (Figure 4.20B).

The previous Epi25 Collaborative analyses[33,34] demonstrated a high burden of missense variants in constrained (intolerant) sites in DEE, GGE, and NAFE, seen in dominant epilepsy genes, DEE genes, and NDD-Epilepsy genes. We observed similar enrichment patterns (Figure 4.21) in MPC 2 and MTR De Novo classes of variants (enriched for *de novo* mutations). Limiting the analysis to highly constrained genic regions (CCR 80 class of variants) resulted in a marked increase in URVs burden, as was the trend in all the tested gene sets so far. Testing these sets also unraveled strong enrichment in PTVs and missense variants in paralog-conserved sites. PTVs and missense variants in paralog-conserved sites did not show substantial enrichment in exome-wide analysis and most of other expression-based, localization-based, or pathway-based gene sets. However, we saw a modest increase in PTV burden in highly intolerant genes with probability of Loss-of-function Intolerance (pLI) > 0.995 in all epilepsies (Figure 4.13). The choice of the pLI score cut-off was based on the outcomes of a previous analysis[10] which demonstrated that the burden in PTVs in genes with pLI > 0.9 is driven primarily by genes with pLI > 0.995 rather than 0.9 – 0.995. In a gene set of known DEE genes, in which highly intolerant genes are rather prevalent, we saw a prominent enrichment in PTVs burden in DEE. Also, there was an increased burden in missense variants in paralog-conserved sites in sets of epilepsy-related disease genes (DEE genes, dominant Epilepsy genes, NDD-Epilepsy genes). This burden was strong in DEE but not as remarkable in GGE and NAFE (Figure 4.21).

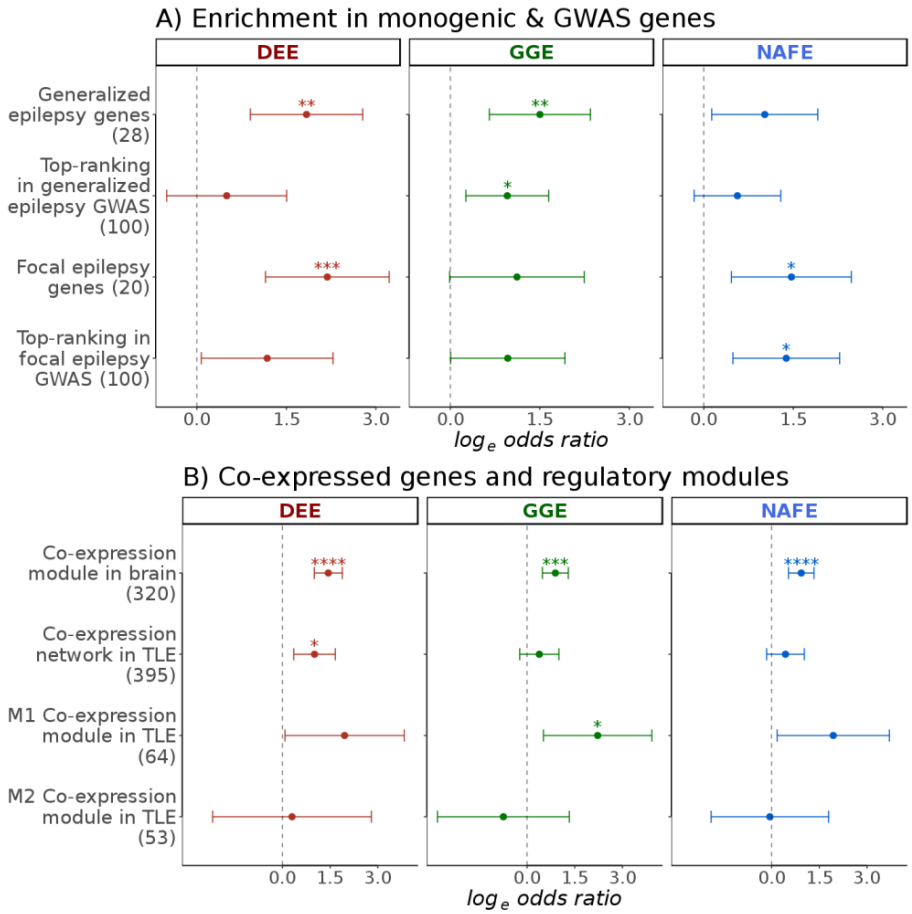


Figure 4.20: Risk elements in GWAS top-ranked genes and co-expression modules. The burden of missense variants in highly constrained sites (log-odds on the *x*-axis; error bars indicate 95% confidence intervals) in developmental and epileptic encephalopathies (DEE), genetic generalized epilepsies (GGE) and non-acquired focal epilepsies (NAFE) is shown in gene sets (*y*-axis; number of genes in parenthesis) representing **(A)** Generalized or Focal epilepsy (presumed monogenic) genes as well as top-ranked 100 genes from GWAS of generalized and focal epilepsies, and **(B)** co-expressed genes identified in post-mortem brain tissues of healthy individuals (module of 320 genes) or in brain tissues from Temporal Lobe Epilepsy (TLE) patients (network of 395 genes) as well as two sub-modules of this network (M1 and M2). False-Discovery-Rate adjusted *p* values are indicated with stars as follows: no star > 0.05, * < 0.05, ** < 0.005, *** < 0.0005, **** < 0.00005.

4.4.5 Control analyses to exclude bias and inflation

Examination of control classes and control gene sets that are not expected to show an enrichment supported the validity of our analysis. The analysis for synonymous variants did not show more substantial enrichment than expected by chance, indicating sufficient control for inflation, particularly in exome-wide models and gene sets with considerable number of genes. In this control analysis (synonymous variants), few tests showed p values < 0.05 (15 out of 276 tests of 92 gene sets and 3 phenotypes: 5.4%). The analysis for benign missense variants, another class that is not expected to show an increased burden in cases vs. controls,[33] did not show substantial enrichment as well. Nine out of 276 tests for benign missense variants (3.2%) showed p values < 0.05 (only 2 with FDR-adjusted p values < 0.05). Possible alternative explanations for such subtle signals include residual population stratification, differences in exome capture not adjusted by covariates and the presence of synonymous variants with functional consequences.[235] However, these proportions are close to the limit expected by chance under a true null hypothesis (5% with = 0.05).

Four sets of genes not expressed in the brain that were tested (high confidence genes with depleted RNA and protein expression in the brain, genes with no RNA detected in the cortex, the hippocampus, or any brain tissue) were not substantially enriched in most the tested variant classes (Figure 4.22). Also, we examined eleven cancer and metabolic pathways (KEGG) to have additional insights into the specificity of the observed signals to neuronal processes and genes (Figures 4.23 and 4.24). Among 540 tests targeting functional variants in these non-neuronal gene sets (3 epilepsy subtypes, 15 sets representing genes not expressed in the brain, KEGG metabolic and cancer pathways, 12 non-synonymous functional classes of variants), 18 tests (3.3%) had an FDR-adjusted p values < 0.05. At least for some of those, the enrichment could be explained by an overlap with genes known to play a role in epilepsy. For instance, genes forming the Type II Diabetes KEGG pathway are substantially enriched in DEE (FDR-adjusted p values of 0.007 for MTR DeNovo and 0.01 for CCR 80 class of variants). This pathway contains two genes that are known to cause DEE, namely, *CACNA1A*[236] and *CACNA1E*.[237] The enrichment was no longer prominent (p values > 0.05) after the removal of these two genes (Figure 4.24).

A potential source of bias in our burden testing was the imbalance in male-to-female ratios between cases and controls (Table 4.5). We provide results from a secondary analysis that excluded all genes located on chromosome X, which shows that any bias not captured by

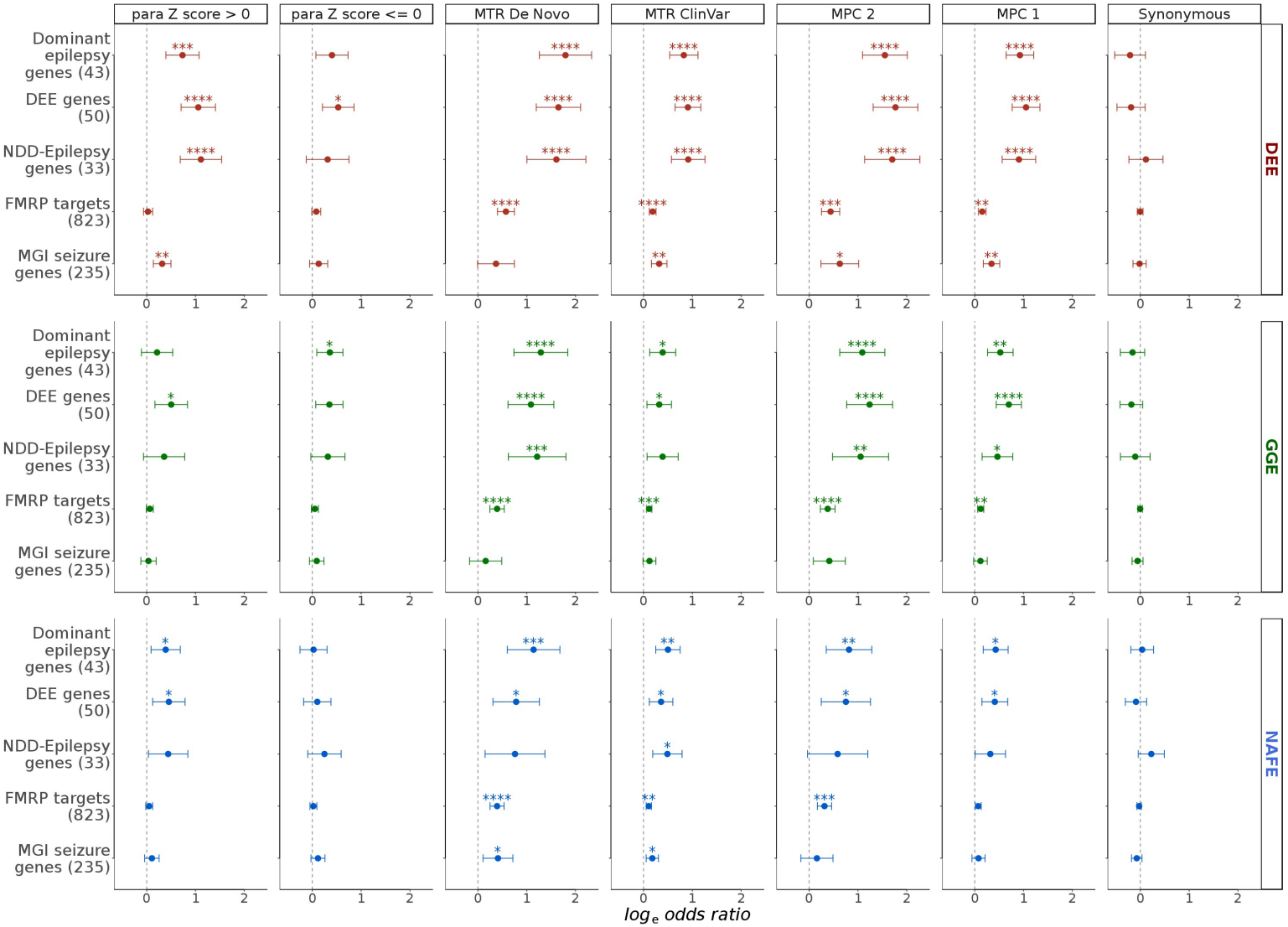


Figure 4.21: Burden of ultra-rare variants in groups of epilepsy-related known disease genes. The burden in five gene sets (*y*-axis; number of genes between brackets) in developmental and epileptic encephalopathies (DEE), genetic generalized epilepsies (GGE) and non-acquired focal epilepsies (NAFE) (horizontal panel) in selected variant classes (vertical panels) is shown on the *x*-axis (log odd ratios from Likelihood Ratio Test; error bars indicate 95% confidence intervals). False-Discovery-Rate-adjusted *p* values (synonymous variants analysis *p* values were not adjusted) are indicated with stars as follows: no star > 0.05, * < 0.05, ** < 0.005, *** < 0.0005, **** < 0.0005. NDD-Epilepsy: neurodevelopmental disorders with epilepsy. FMPR: Fragile-X Mental Retardation Protein targets. MGI: Mouse Genome Informatics database.

the inclusion of sample sex as a covariate is likely marginal (online supplementary[226]). To exclude any major residual stratification resulting from the use of different enrichment kits, we additionally performed a controls-only analysis in which we compared control samples enriched with Illumina ICE capture kits (from Leicester study) to controls enriched using Agilent SureSelect kits (ATVB study and Ottawa study). This analysis reflected a good control for any potential bias introduced by different exome capture systems and demonstrated that the mixing of controls included (Leicester and Ottawa) or not included (ATVB) in gnomAD is unlikely to have affected our main outcomes (Figure 4.25 and online supplementary[226]).

4.5 Discussion

By analyzing the sequencing data of 11,551 unrelated European individuals (1,003 individuals with DEE, 3,064 individuals with GGE, and 3,522 individuals with NAFE vs. 3,962 controls), we show an increased burden in ultra-rare missense variants in highly constrained sites in epilepsy cases compared to controls, not only in intolerant and known epilepsy-related genes, as previously shown,[33,34] but also exome-wide in all protein coding genes. Similar to the observations made in several other phenotypes, the burden in PTVs was most prominent in known disease genes and brain-expressed loss-of-function intolerant genes.[103,107,217] Consistent with their enrichment in neurodevelopmental disorders,[215] the burden in missense variants in paralog-conserved sites was prominent in DEEs. The lower burden of these variants in GGEs and NAFEs may reflect a true disparity between rare and common epilepsies. The presented results are also consistent with previous analyses of missense variants in a small number of gene sets examined in similar cohorts.[33,34,47,78]

The systematic analysis of additional gene sets and a wider variety of classes of variants revealed interesting findings about the neurobiology of distinct types of epilepsy. Although associated with higher odds ratios of an epilepsy phenotype, enriched variants are not deterministic on their own, since about one-fourth of the controls also carry qualifying variants in the CCR 80 analysis. As such, the phenotype is determined by a constellation of other factors, possibly including the severity of variants,[34] patterns of multiple variations, oligogenic contribution from rare variants,[238] and polygenic risk from common variants.[66] Developmental genes were key drivers in all epilepsies suggesting that the impairment of developmental processes is not limited to DEEs with marked developmental deficits.[22]

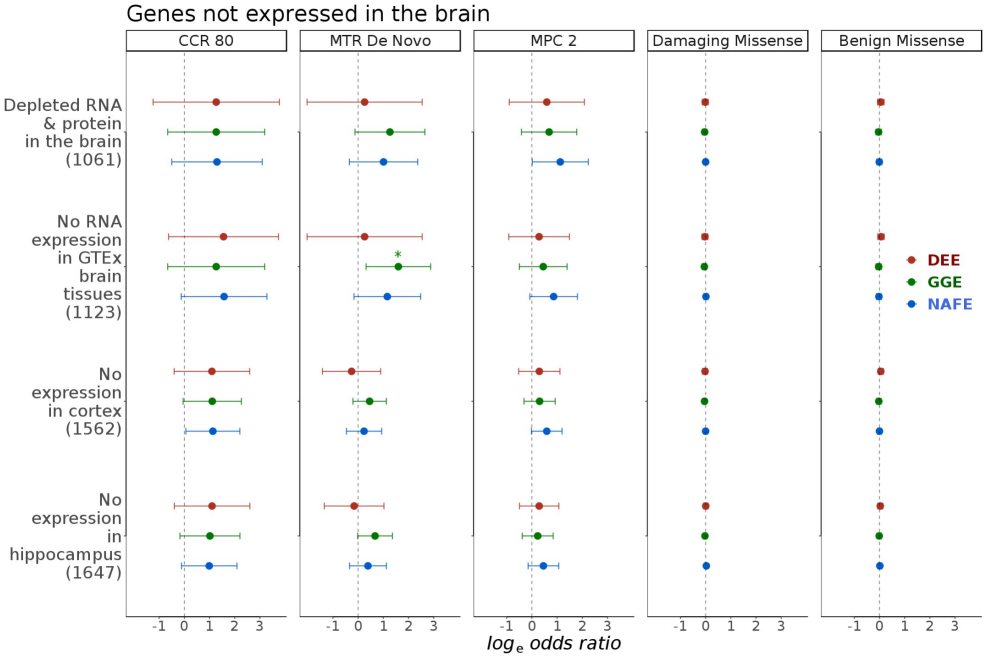


Figure 4.22: Burden in groups of genes not expressed in the brain. Panels: variant classes. *y* axis: gene sets (genes count between parenthesis). *x* axis: log odds ratio from regression analysis of individual burden of qualifying variants. Stars indicate False Discovery Rate-adjusted *p* values: * < 0.05, ** < 0.005, *** < 0.0005, **** < 0.00005. Error bars indicate 95% confidence intervals of odds. DEE: developmental and epileptic encephalopathies. GGE: genetic generalized epilepsies. NAFE: non-acquired focal epilepsies.

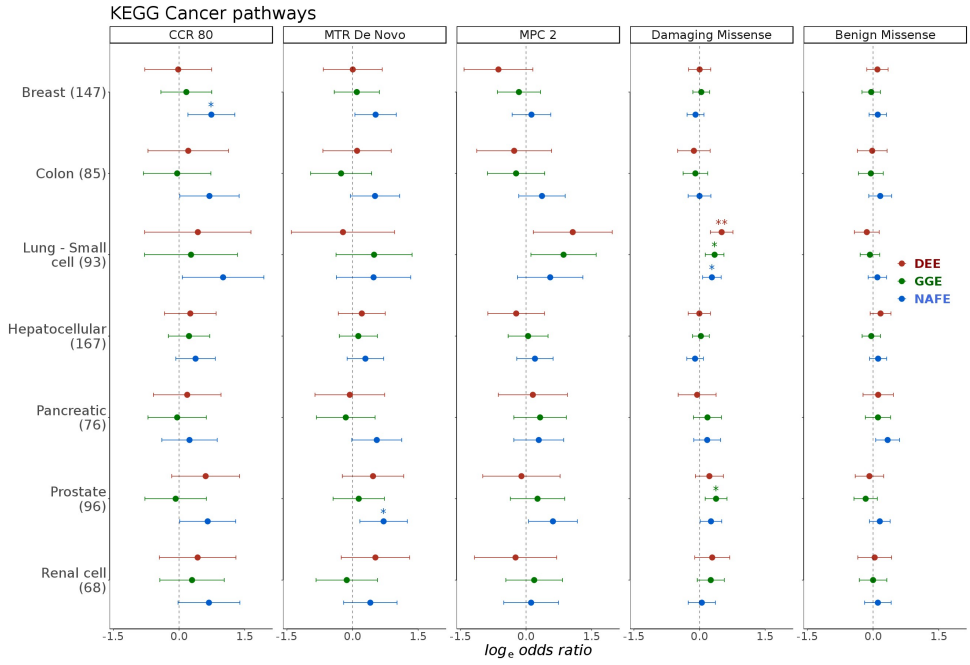


Figure 4.23: Burden in gene sets from KEGG cancer pathways. Panels: variant classes. *y* axis: gene sets (genes count between parenthesis). *x* axis: log odds ratio from regression analysis of individual burden of qualifying variants. Stars indicate False Discovery Rate-adjusted *p* values: * < 0.05, ** < 0.005, *** < 0.0005, **** < 0.00005. Error bars indicate 95% confidence intervals of odds. DEE: developmental and epileptic encephalopathies. GGE: genetic generalized epilepsies. NAFE: non-acquired focal epilepsies.

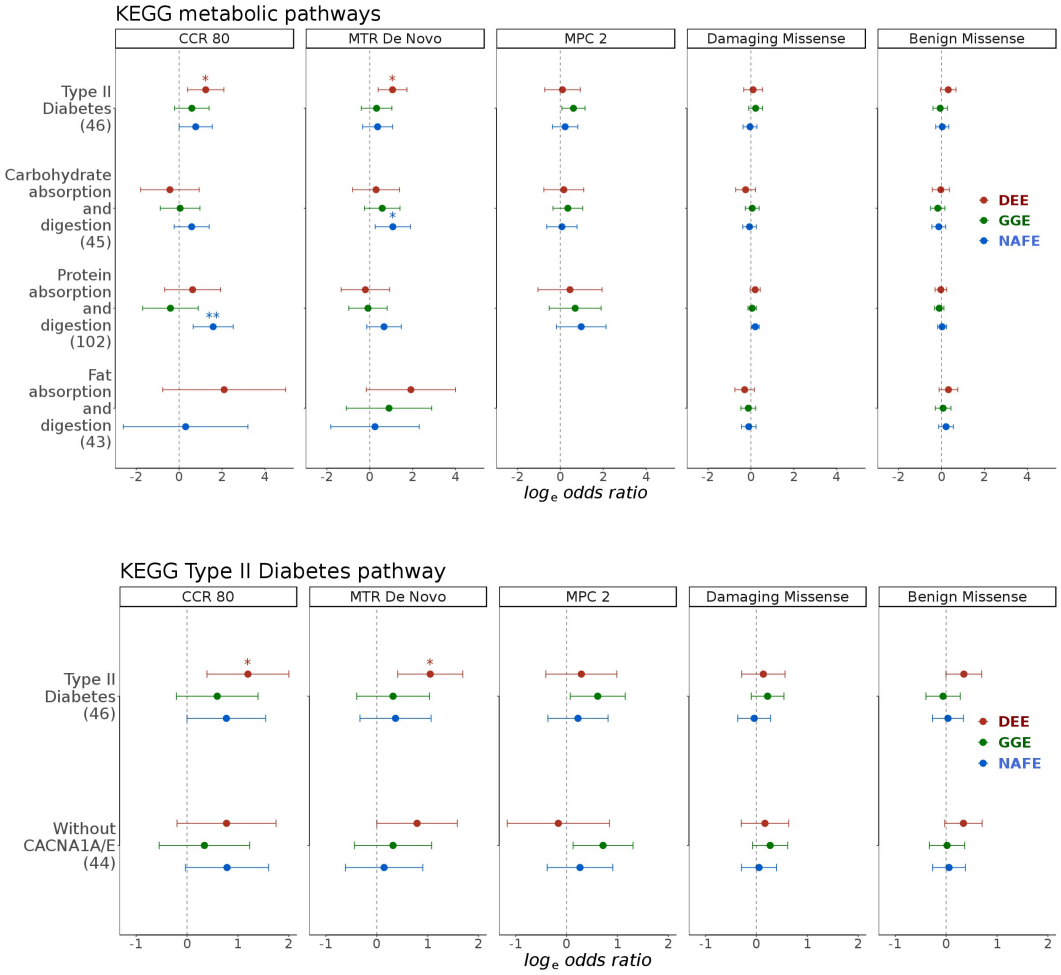


Figure 4.24: Burden in gene sets from KEGG metabolic pathways. The burden in selected gene sets representative of metabolic processes (absorption and digestion) or a metabolic disorder (type II diabetes) is shown in the top panel. The enrichment of variants in DEE in the type II diabetes pathway can be explained by the presence of two DEE genes (*CACNA1A* & *CACNA1E*) as shown in the bottom panel. Sub-panels: variant classes. *y* axis: gene sets (genes count between parenthesis). *x* axis: log odds ratio from regression analysis of individual burden of qualifying variants. Stars indicate False Discovery Rate-adjusted *p* values: * < 0.05, ** < 0.005, *** < 0.0005, **** < 0.00005. Error bars indicate 95% confidence intervals of odds. DEE: developmental and epileptic encephalopathies. GGE: genetic generalized epilepsies. NAFE: non-acquired focal epilepsies.

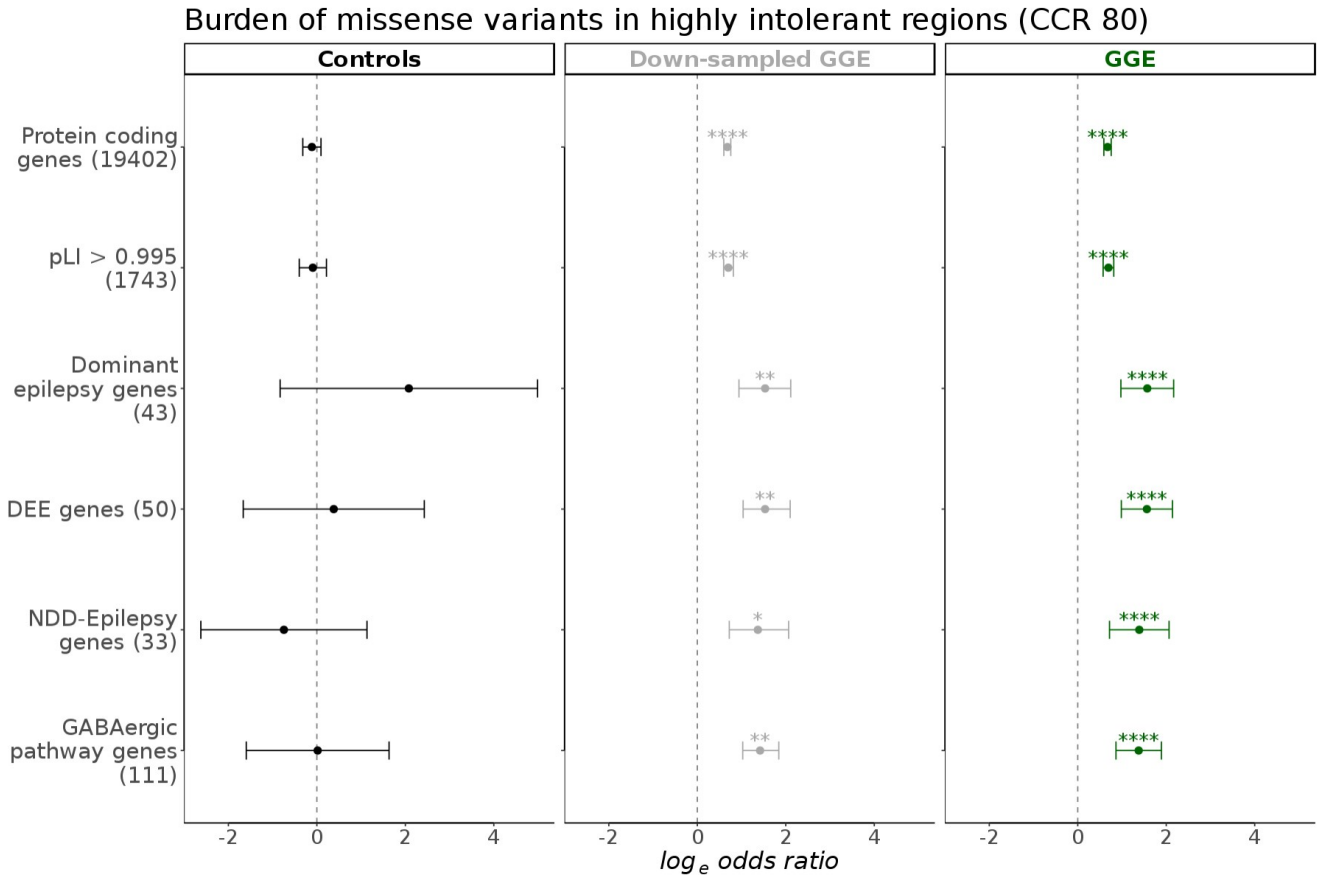


Figure 4.25: Secondary analyses to exclude capture kit artifacts. An analysis of the burden of missense variants in regions with CCR score equal to or exceeding 80 in six key gene sets in 1,100 controls prepared using Illumina ICE capture kits (in gnomAD) *vs.* 2,789 controls prepared using Agilent SureSelect kit (not in gnomAD) did not show any substantial enrichment. These numbers are likely sufficient to detect an enrichment in these genes sets based on an analysis of an equal number of randomly selected GGE cases *vs.* controls. The results of the analysis of all GGEs vs. all controls in these gene sets are shown for comparison. Panels: variant classes. *y* axis: gene sets (genes count between parenthesis). *x* axis: log odds ratio from regression analysis of individual burden of qualifying variants. Stars indicate False Discovery Rate-adjusted *p* values: * < 0.05, ** < 0.005, *** < 0.0005, **** < 0.00005. Error bars indicate 95% confidence intervals of odds. For down-sampling: odds and *p* values were averaged over 500 permutation and error bars indicate 2.5th and 97.5th centiles of odds. DEE: developmental and epileptic encephalopathies. GGE: genetic generalized epilepsies. NDD: Neurodevelopmental disorders. pLI: probability of Loss-of-function intolerance score.

The enrichment in synaptic genes is another shared feature between the epilepsies that has also been observed in neurodevelopmental disorders with epilepsy,[32,239] schizophrenia,[106] and autism.[240] This highlights a shared genetic architecture not only between epilepsy subtypes but also with other related neurological disorders, as has been shown previously for common variants.[241]

Despite the common genetic and phenotypic features, DEEs, GGEs and NAFEs represent well-recognized phenotypic clusters with defined electro-encephalographic and clinical characteristics. Given the phenotypic severity of DEEs, the prevalence of *de novo* variants and monogenic cases in DEE (those with pathogenic and likely pathogenic variants in known monogenic genes), and the description of phenotypic spectra for genes involved in DEE that also span the milder GGE or NAFE, the distinction between severe and mild epilepsies could be attributed, at least to some extent, to the severity of the genetic defects, their functional effects or their localization within certain channel regions.[34,51,69,71,242,243] The distinction between GGE and NAFE, however, is probably functional, at least in part, as suggested by previous work demonstrating the centrality of GABAergic genes in generalized epilepsies.[33,47] Also, it is well recognized that few genes present with focal epilepsy and are not linked generalized epilepsy syndromes.[59] Here, phenotype-specific patterns in gene sets representing neuronal inhibitory vs. excitatory signaling were observed in comparisons of GGE and NAFE.

Additional disparities in key gene sets (genes implicated in monogenic generalized & focal epilepsy, the one hundred top-ranked genes associated with GWAS hits in generalized & focal epilepsy) point to a possible genetic-functional divergence, so that a common background of shared risk seems to be overlaid by specific risk entities. The enrichment of rare variants in GWAS genes also supports the convergence of ultra-rare and common variants in conferring epilepsy risk, in concordance with the observed enrichment of epilepsy GWAS hits for monogenic epilepsy genes.[60] According to our findings, a link between common and rare variants is likely to be also relevant for the phenotypic heterogeneity observed in seizure disorders. Notably, polygenic risk scores also pointed out the specificity of the risk profiles in common epilepsies.[66] Based on previous findings of an increased URV burden in DEEs[225] and the current findings in GGEs and NAFEs, it is also conceivable that differentially expressed genes in individuals with epilepsy, representing closely orchestrated networks with possible functional correlations, would highlight modules in which altered transcription, URVs, or both contribute to cause both rare and common epilepsies.

The associations presented in this work should be interpreted with the caveats of gene set testing in mind.[244] Pathways and molecular processes are not consistently defined in different resources. These differences may explain the discrepancies in enrichment patterns in the same pathway. We examined multiple overlapping gene sets from different sources to corroborate the findings that underscore a genuine biological relevance. Our analysis has additional limitations which we aimed to overcome using stringent analysis and quality control strategies. The limited use of about half of the controls from the primary analysis affected the overall power. Nevertheless, we were able to reproduce most of the major signals from gene sets with large effect sizes, the latter thereby acting as positive controls. Multiple secondary analyses suggested that the impalpable of male-to-female ratios in our case and control sets and the use of sequencing data from ExAC,[245] gnomAD[103] or DiscovEHR[135] to develop, train or validate *in silico* algorithms used for estimating constraint[138,213,214] do not seem to have introduced a substantial bias (online supplementary[226]). The overlap between the controls used in this study and gnomAD controls created some challenges in defining URVs. For population frequency filtering, we allowed around five alleles in gnomAD (allele frequency of $2x10^{-5}$) to retain URVs from our control that are also seen in gnomAD while still filtering common variants and prevalent sequencing artifacts.

In conclusion, missense URVs affecting constrained sites in brain-expressed genes show distinct signatures in epilepsy. Enrichment patterns of URVs-affected genes suggest a preferential involvement of inhibitory genes in GGE and excitatory genes in focal epilepsies. Genes implicated by common GWAS variants may also be disrupted by URVs in various epilepsy phenotypes, suggesting a convergence of rare disruptive variants, and common variants in the pathogenesis of epilepsy.

Chapter 5: Concluding remarks

5.1 The overall context

Although not originally intended to answer the same research question, the presented research efforts funnel together to highlight the complex and likely polygenic nature of common epilepsies. In hindsight, it is therefore reasonably accurate to describe the theme of this doctoral work as the discovery and evaluation of the association of URVs in protein coding genes with epilepsy. Previously, URVs showed the strongest evidence for an association with less severe epilepsies compared to variants with a higher frequency.[33] This has been established in rare variant association studies using collapsing approaches assuming dominant inheritance. Here, we reiterate the relevance of URVs to epilepsy risk and add to the understanding of the genetic risk mediated through URVs particularly in genetically complex epilepsies.

We performed a small family-based study aiming to investigate the extent to which bi-allelic variants might contribute the risk of familial epilepsies in consanguineous families (Chapter 2). Our findings highlight the complexity of genetic risk determination as a homozygous pLoF variant in *PRRT2* was found to underlie a similar phenotype like heterozygous alleles. We could not find sufficient evidence to suggest a major contribution from bi-allelic alterations. Although this conclusion cannot be readily generalized given the exceedingly small size of our cohort, it does argue against a substantial role for bi-allelic inheritance. Concomitant work in a relatively larger cohort of 20 unrelated Sudanese families did not find clear evidence of recessive inheritance in GGE syndromes as well.[126] Other prior family studies investigating recessive variants did not result in findings that could be replicated, whereas collapsing analyses using recessive models were underpowered.[33]

Otherwise, we identified heterozygous URVs in known dominant disease genes in a few families. Additionally, our related work on the genetics of rare NDD with epilepsy in Sudan revealed several novel URVs, sometimes with a founder effect.[127,246–251] These epilepsy genes were well established across different populations, echoing the notion that epilepsy genes are unlikely to be population specific, although the individual risk variants (apart from recurrent variants in mutational hotspots) are likely dependent on the population structure. Two previously reported VUS in *EFHC1* were classified as benign in this study (Table 2.3). Notably, *EFHC1* and *ICK* have been hypothesized to convey a population specific risk to common epilepsies but this association is debatable. Both were identified in the same linkage locus and have been linked to GGE syndromes (particularly JME) in Hispanic, European-American and Japanese populations.[57,252] These findings could not be replicated.[101,56]

URVs in several genes supported by previous family-based and functional studies but failing to achieve study-wide significance individually (e.g., genes encoding several $GABA_A$ receptors) cannot be set apart from genes that do not have sufficient biological evidence yet with comparable *p* values (e.g., Figure 3.4). This highlights one of the major challenges in performing RVAS. Apart from the examination of data sets from various populations, increasing the sample size (accordingly, the power to detect smaller effect sizes) is the most direct approach to overcome this challenge. To this end, our work (Chapter 3) attempted to achieve a considerable increase in the sample size of investigated epilepsy cohorts through a meta-analysis of multiple existing data sets while enriching the analysis for familial cases. This dissertation also touched upon few other workarounds, including adopting gene set based analysis methods coupled with constraint metrics to capture high effect variants (Chapter 4).

The results added to our understanding of the pathogenesis of the different epilepsies, particularly GGE. We found that familial GGE (vs. controls) showed a proportionally higher burden of deleterious variants (predicted to have a high effect size) compared to sporadic GGEs (vs. controls). Despite the lack of study-wide significance, several top-ranked genes were linked to DEEs, corroborating previous obseravations.[33,34,78] URVs in known epilepsy genes seem to underlie a small fraction of GGEs, thus constituting a relatively rare predisposing factor at a population scale. Also, there seems to be additional contribution from genes yet to be implicated in the so-called *monogenic epilepsies*. To that end, several gene sets not based on known disease genes, but on biological entities, were found to show an enrichment in deleterious URVs (Chapter 4). Moreover, there seems to be a distinction between the epileptogenic mechanisms between GGE and focal onset epilepsies that can be attributed – partially – to the nature of affected pathways, as seen in comparisons of gene sets important for inhibitory vs. excitatory signaling.

An overarching limitation of these studies which employed exome sequencing is the restricted analysis of short coding variants. Since the added value of employing whole genome sequencing over whole exomes in genetic diagnostics is still very limited,[253] this drawback is unlikely to have affected the validity of the major outcomes of this work. We investigated a single-proband per family in these studies. This could potentially miss variation specific to single family members. Our aim was to capture variation with strong effect while accounting for possible genetic heterogeneity (i.e., high effect URVs regardless of their nature as private

variants or inherited variants shared between affected family members), which lends validity to our approach based on single probands.

We also assumed an equal weight and a single direction of effect for qualifying variants (i.e., variants within a single class were considered equally deleterious). To overcome the limitations of inferring the severity of functional defects from *in silico* scores, we resorted to using different models (e.g., PPh2 vs. REVEL) or different cut-offs (MPC1 vs. MPC2). Such collapsing approaches are computationally inexpensive (for instance, compared to generalized mixed models) with a well-controlled type I error rate, as shown in recent comparisons,[254] and therefore well-suited for such study designs. Our association studies also assumed a dominant collapsing model (in which both heterozygous and homozygous variants are considered dominant qualifying variants). This is rather justified given epilepsy is not among a few diseases with a considerable contribution from bi-allelic variants to disease risk.[255] Other common limitations of the presented association studies are the use of analysis models that assumes homogeneity in the phenotypic characteristics among the patients. The effect of individuals' age (at onset and sampling) was not considered, mostly due to incomplete data from our controls (Chapters 3 and 4). Similarly, the relative severity of the disease was not accounted for. Both (age and severity) are key determinants of the yield of genetic testing.[256] Including such detailed metrics would have also required much larger cohorts achieve reasonable power. Future work, as will be discussed next, can address these limitations.

5.2 Future directions

(1) Levering the increased availability of samples through international collaboratives, biobanks and public databases:

The most natural continuation of the presented work is the investigation of larger cohorts enriched with individuals with a positive family history and individuals from under-represented populations. This serves to increase the overall power to confirm tentative associations (not reaching study-wide significance) and to replicate findings in independent cohorts. Additional disease descriptors and severity scores could be incorporated into the association analysis, e.g., to examine the enrichment of URVs in different age groups or to include the age of onset as a covariate. Also, comparisons of affected siblings might offer additional insights into disease modifiers. If parental data are available, it might be possible to

capture additional, more specific, effects (e.g., parent-of-origin[257]). Merging datasets across existing cohorts can improve the outcomes of genetic analyses but harmonizing the phenotypic information across cohorts is a considerable challenge. The use of controlled vocabularies (e.g., HPO[4]) and consensus case definitions (e.g., ILAE definitions[22]) will allow merging data from diverse sources with more efficiency. The availability of appropriate control datasets may hinder large case-control analyses. However, data from population databases and biobanks are increasing considerably. With the growing availability of sample-level genotypes (e.g., UK Biobank) or haplotypes (e.g., gnomAD), it will be possible to combine case datasets from case-control studies with publicly available control datasets with increasing efficiency. Also, different approaches have been developed to combine external case datasets with publicly available controls using summary statistics (without the need for genotype-level data).[258–260]

(2) Leveraging recent advances in sequencing technologies and genetic analysis models:

With the advances in short and long read sequencing technologies, it will be meaningful to investigate the contribution of other types of coding and non-coding variation that are so far difficult to capture (e.g., structural variants and repeat expansions). Similarly, the advances in genetic modeling of kinship and ancestry using generalized mixed models could be leveraged to include related individuals and multiple ancestries.[46] Frameworks have also been developed to combined case-control and family-based sequencing data,[261] potentially improving the statistical power. Ultimately, Bayesian approaches incorporating existing evidence as prior probabilities in association testing might highlight additional differences.[262] The advances in exome capture technologies will make joining dataset from different cohorts more efficient, as high-quality data is easier to combine and homogenize.

(3) An in-depth analysis of the association seen in gene sets:

Last, it would be meaningful to identify the contribution of single genes within each gene set to the overall burden. This is likely to reflect differences in genes driving the association in different epilepsies. A few genes might drive the enrichment in a gene set. Simple methodological approaches based on ranking genes on the frequency of QVs can highlight such genes, e.g., though the identification of ranks in which the enrichment is maximized (e.g., Table 3.8) or through leave-out cross-validation by removing top-ranked genes (e.g., Figure 2.24). Testing a comprehensive range of gene sets in sufficiently large cohorts of familial and

sporadic epilepsies could further improve our understanding of the overlap and distinctions between the two disease forms. Since the mechanisms of compensation for single gene defects within pathways are not always clear, examining smaller gene sets representing protein complexes or directly interacting proteins might be another relevant approach. Moving from self-contained gene set analyses to competitive gene set enrichment testing – where the enrichment in gene sets is examined relative to the background burden attributed to all other genes not in the gene set – could offer further biologically relevant insights.[244]

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. 2014;475–82.
2. Britton JW, Frey LC, Hopp JL, Korb P, Koubeissi MZ, Lievens WE, et al. EEG in the Epilepsies. In: Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants. St. Louis EK, Frey LC, editors. Chicago: American Epilepsy Society. 2016.
3. Birca V, Keezer MR, Chamelian L, Lortie A, Nguyen DK. Recognition of Psychogenic Versus Epileptic Seizures Based on Videos. Can J Neurol Sci. 2021 Jun 21;1–9.
4. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Research. 2021 Jan 8;49(D1):D1207–17.
5. Winawer MR. Phenotype definition in epilepsy. Epilepsy Behav. 2006 May;8(3):462–76.
6. Guerrini R, Buchhalter JR. Epilepsy phenotypes and genotype determinants: Identical twins teach lessons on complexity. Neurology. 2014 Sep 16;83(12):1038–9.
7. Myers KA, Johnstone DL, Dyment DA. Epilepsy genetics: Current knowledge, applications, and future directions. Clin Genet. 2019 Jan;95(1):95–111.
8. Nicolson A, Chadwick DW, Smith DF. A comparison of adult onset and “classical” idiopathic generalised epilepsy. J Neurol Neurosurg Psychiatry. 2004 Jan;75(1):72–4.
9. Gaitatzis A, Carroll K, Majeed A, W Sander J. The epidemiology of the comorbidity of epilepsy in the general population. Epilepsia. 2004 Dec;45(12):1613–22.
10. Srivastava S, Sahin M. Autism spectrum disorder and epileptic encephalopathy: common causes, many questions. J Neurodev Disord. 2017;9:23.
11. Ewen JB, Marvin AR, Law K, Lipkin PH. Epilepsy and Autism Severity: A Study of 6,975 Children. Autism Res. 2019 Aug;12(8):1251–9.
12. Symonds JD, Elliott KS, Shetty J, Armstrong M, Brunklaus A, Cutcutache I, et al. Early childhood epilepsies: epidemiology, classification, aetiology, and socio-economic determinants. Brain. 2021 Oct 22;144(9):2879–91.
13. Beghi E, Giussani G, Nichols E, Abd-Allah F, Abdela J, Abdelalim A, et al. Global, regional, and national burden of epilepsy, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet Neurology. 2019 Apr;18(4):357–75.
14. Fiest KM, Sauro KM, Wiebe S, Patten SB, Kwon C-S, Dykeman J, et al. Prevalence and incidence of epilepsy: A systematic review and meta-analysis of international studies. Neurology. 2017 Jan 17;88(3):296–303.
15. Roy T, Pandit A. Neuroimaging in epilepsy. Ann Indian Acad Neurol. 2011 Apr;14(2):78–80.

16. Haut SR, Shinnar S, Moshé SL. Seizure clustering: risks and outcomes. Epilepsia. 2005 Jan;46(1):146–9.
17. Thijs RD, Surges R, O'Brien TJ, Sander JW. Epilepsy in adults. Lancet. 2019 Feb 16;393(10172):689–701.
18. Asadi-Pooya AA, Beniczky S, Rubboli G, Sperling MR, Rampp S, Perucca E. A pragmatic algorithm to select appropriate antiseizure medications in patients with epilepsy. Epilepsia. 2020 Aug;61(8):1668–77.
19. Kwan P, Arzimanoglou A, Berg AT, Brodie MJ, Allen Hauser W, Mathern G, et al. Definition of drug resistant epilepsy: consensus proposal by the ad hoc Task Force of the ILAE Commission on Therapeutic Strategies. Epilepsia. 2010 Jun;51(6):1069–77.
20. Ryvlin P, Cross JH, Rheims S. Epilepsy surgery in children and adults. Lancet Neurol. 2014 Nov;13(11):1114–26.
21. Lerche H. Drug-resistant epilepsy — time to target mechanisms. Nat Rev Neurol. 2020 Nov;16(11):595–6.
22. Scheffer IE, Berkovic S, Capovilla G, Connolly MB, French J, Guilhoto L, et al. ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. Epilepsia. 2017 Apr;58(4):512–21.
23. Pressler RM, Cilio MR, Mizrahi EM, Moshé SL, Nunes ML, Plouin P, et al. The ILAE classification of seizures and the epilepsies: Modification for seizures in the neonate. Position paper by the ILAE Task Force on Neonatal Seizures. Epilepsia. 2021 Mar;62(3):615–28.
24. Ellis CA, Ottman R, Epstein MP, Berkovic SF, Epi4K Consortium. Generalized, focal, and combined epilepsies in families: New evidence for distinct genetic factors. Epilepsia. 2020 Dec;61(12):2667–74.
25. Katyayan A, Diaz-Medina G. Epilepsy: Epileptic Syndromes and Treatment. Neurol Clin. 2021 Aug;39(3):779–95.
26. Jallon P, Latour P. Epidemiology of idiopathic generalized epilepsies. Epilepsia. 2005;46(9):10–4.
27. Banerjee PN, Filippi D, Allen Hauser W. The descriptive epidemiology of epilepsy-a review. Epilepsy Res. 2009 Jul;85(1):31–45.
28. Giordano L, Vignoli A, Cusmai R, Parisi P, Mastrangelo M, Coppola G, et al. Early onset absence epilepsy with onset in the first year of life: A multicenter cohort study. Epilepsia. 2013 Oct;54:66–9.
29. Lee EH. Epilepsy syndromes during the first year of life and the usefulness of an epilepsy gene panel. Korean J Pediatr. 2018;61(4):101.

30. Scheffer IE, Liao J. Deciphering the concepts behind “Epileptic encephalopathy” and “Developmental and epileptic encephalopathy.” Eur J Paediatr Neurol. 2020 Jan;24:11–4.
31. Specchio N, Curatolo P. Developmental and epileptic encephalopathies: what we do and do not know. Brain. 2021 Feb 12;144(1):32–43.
32. Heyne HO, Singh T, Stamberger H, Abou Jamra R, Caglayan H, Craiu D, et al. De novo variants in neurodevelopmental disorders with epilepsy. Nat Genet. 2018 Jul;50(7):1048–53.
33. Epi25 Collaborative. Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals. The American Journal of Human Genetics. 2019 Aug;105(2):267–82.
34. Epi25 Collaborative. Sub-genic intolerance, ClinVar, and the epilepsies: A whole-exome sequencing study of 29,165 individuals. The American Journal of Human Genetics. 2021 Jun;108(6):965–82.
35. Marini C, Scheffer IE, Crossland KM, Grinton BE, Phillips FL, McMahon JM, et al. Genetic architecture of idiopathic generalized epilepsy: clinical genetic analysis of 55 multiplex families. Epilepsia. 2004 May;45(5):467–78.
36. Vadlamudi L, Milne RL, Lawrence K, Heron SE, Eckhaus J, Keay D, et al. Genetics of epilepsy: The testimony of twins in the molecular era. Neurology. 2014 Sep 16;83(12):1042–8.
37. The Epi4K Consortium. Epi4K: Gene discovery in 4,000 genomes. Epilepsia. 2012 Aug;53(8):1457–67.
38. Ottman R, Risch N. Genetic Epidemiology and Gene Discovery in Epilepsy. In: Jasper’s Basic Mechanisms of the Epilepsies. Noebels JL, et al., editors. 2012;1–14.
39. Poduri A, Sheidley BR, Shostak S, Ottman R. Genetic testing in the epilepsies-developments and dilemmas. Nature reviews Neurology. 2014;10(5):293–9.
40. Bertoli-Avella AM, Kandaswamy KK, Khan S, Ordonez-Herrera N, Tripolszki K, Beetz C, et al. Combining exome/genome sequencing with data repository analysis reveals novel gene–disease associations for a wide range of genetic disorders. Genet Med. 2021 Aug;23(8):1551–68.
41. Gesche J, Hjalgrim H, Rubboli G, Beier CP. The clinical spectrum of familial and sporadic idiopathic generalized epilepsy. Epilepsy Res. 2020 Sep;165:106374.
42. Abou-Khalil B, Krei L, Lazenby B, Harris PA, Haines JL, Hedera P. Familial genetic predisposition, epilepsy localization and antecedent febrile seizures. Epilepsy Res. 2007 Jan;73(1):104–10.
43. Callenbach PM, Geerts AT, Arts WF, van Donselaar CA, Peters AC, Stroink H, et al. Familial occurrence of epilepsy in children with newly diagnosed multiple seizures: Dutch Study of Epilepsy in Childhood. Epilepsia. 1998 Mar;39(3):331–6.

44. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. Nat Rev Genet. 2017 Oct;18(10):599–612.
45. Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. The Lancet Neurology. 2014 Sep;13(9):893–903.
46. Zhou W, Zhao Z, Nielsen JB, Fritsche LG, LeFaive J, Gagliano Taliun SA, et al. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. Nat Genet. 2020 Jun;52(6):634–9.
47. May P, Girard S, Harrer M, Bobbili DR, Schubert J, Wolking S, et al. Rare coding variants in genes encoding $GABA_A$ receptors in genetic generalised epilepsies: an exome-based case-control study. The Lancet Neurology. 2018 Aug;17(8):699–708.
48. Hebbar M, Mefford HC. Recent advances in epilepsy genomics and genetic testing. F1000Res. 2020 Mar 12;9:185.
49. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015 May;17(5):405–23.
50. Johannesen KM, Liu Y, Koko M, Gjerulfsen CE, Sonnenberg L, Schubert J, et al. Genotype-phenotype correlations in *SCN8A*-related disorders reveal prognostic and therapeutic implications. Brain. 2021 Aug 25;awab321.
51. Johannesen KM, Gardella E, Linnankivi T, Courage C, de Saint Martin A, Lehesjoki A-E, et al. Defining the phenotypic spectrum of *SLC6A1* mutations. Epilepsia. 2018 Feb;59(2):389–402.
52. Fahed AC, Wang M, Homburger JR, Patel AP, Bick AG, Neben CL, et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. Nat Commun. 2020 Dec;11(1):3635.
53. Rahit KMTH, Tarailo-Graovac M. Genetic Modifiers and Rare Mendelian Disease. Genes. 2020 Feb 25;11(3):239.
54. Kaibara FS, de Araujo TK, Araujo PAORA, Alvim MKM, Yasuda CL, Cendes F, et al. Association Analysis of Candidate Variants in Admixed Brazilian Patients With Genetic Generalized Epilepsies. Front Genet. 2021 Jul 8;12:672304.
55. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al. The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. Human Mutation. 2015 Oct;36(10):915–21.
56. Lerche H, Berkovic SF, Lowenstein DH; EuroEPINOMICS-CoGIE Consortium; EpiPGX Consortium; Epi4K Consortium/Epilepsy Phenome/Genome Project. Intestinal-Cell Kinase and Juvenile Myoclonic Epilepsy. N Engl J Med. 2019 Apr 18;380(16):e24.

57. Bailey JN, de Nijs L, Bai D, Suzuki T, Miyamoto H, Tanaka M, et al. Variant Intestinal-Cell Kinase in Juvenile Myoclonic Epilepsy. N Engl J Med. 2018 Mar 15;378(11):1018–28.
58. Helbig I, Lowenstein DH. Genetics of the epilepsies: where are we and where are we going? Current Opinion in Neurology. 2013 Apr;26(2):179–85.
59. Wang J, Lin Z-J, Liu L, Xu H-Q, Shi Y-W, Yi Y-H, et al. Epilepsy-associated genes. Seizure. 2017 Jan;44:11–20.
60. The International League Against Epilepsy Consortium on Complex Epilepsies. Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. Nat Commun. 2018 Dec;9(1):5269.
61. Badano JL, Katsanis N. Beyond Mendel: an evolving view of human genetic disease transmission. Nat Rev Genet. 2002 Oct;3(10):779–89.
62. Antonarakis SE, Chakravarti A, Cohen JC, Hardy J. Mendelian disorders and multifactorial traits: the big divide or one for all? Nat Rev Genet. 2010 May;11(5):380–4.
63. Hunter DJ. Gene–environment interactions in human diseases. Nat Rev Genet. 2005 Apr;6(4):287–98.
64. Bennett CA, Petrovski S, Oliver KL, Berkovic SF. ExACtly zero or once: A clinically helpful guide to assessing genetic variants in mild epilepsies. Neurol Genet. 2017 Aug;3(4):e163.
65. Helbig I, Heinzen EL, Mefford HC, the ILAE Genetics Commission. Primer Part 1-The building blocks of epilepsy genetics. Epilepsia. 2016 Jun;57(6):861–8.
66. Leu C, Stevelink R, Smith AW, Goleva SB, Kanai M, Ferguson L, et al. Polygenic burden in focal and generalized epilepsies. Brain. 2019 Nov 1;142(11):3473–81.
67. Martin HC, Jones WD, McIntyre R, Sanchez-Andrade G, Sanderson M, Stephenson JD, et al. Quantifying the contribution of recessive coding variation to developmental disorders. Science. 2018 Dec 7;362(6419):1161–4.
68. Schubert J, Paravidino R, Becker F, Berger A, Bebek N, Bianchi A, et al. *PRRT2* mutations are the major cause of benign familial infantile seizures. Hum Mutat. 2012 Oct;33(10):1439–43.
69. Wolking S, May P, Mei D, Møller RS, Balestrini S, Helbig KL, et al. Clinical spectrum of *STX1B* -related epileptic disorders. Neurology. 2019 Mar 12; 92(11):e1238-e1249.
70. Oyrer J, Maljevic S, Scheffer IE, Berkovic SF, Petrou S, Reid CA. Ion Channels in Genetic Epilepsy: From Genes and Mechanisms to Disease-Targeted Therapies. Pharmacol Rev. 2018 Jan;70(1):142–73.
71. Maljevic S, Møller RS, Reid CA, Pérez-Palma E, Lal D, May P, et al. Spectrum of $GABA_A$ receptor variants in epilepsy. Current Opinion in Neurology. 2019 Apr;32(2):183–90.

72. Schubert J, Siekierska A, Langlois M, May P, Huneau C, Becker F, et al. Mutations in *STX1B*, encoding a presynaptic protein, cause fever-associated epilepsy syndromes. Nat Genet. 2014 Dec;46(12):1327–32.
73. Happ HC, Carvill GL. A 2020 View on the Genetics of Developmental and Epileptic Encephalopathies. Epilepsy Curr. 2020 Mar;20(2):90–6.
74. Sánchez Fernández I, Loddenkemper T, Gaínza-Lein M, Sheidley BR, Poduri A. Diagnostic yield of genetic tests in epilepsy: A meta-analysis and cost-effectiveness study. Neurology. 2019 Jan 29;92(5):e418–28.
75. Jiang Y, Song C, Wang Y, Zhao J, Yang F, Gao Q, et al. Clinical Utility of Exome Sequencing and Reinterpreting Genetic Test Results in Children and Adults With Epilepsy. Front Genet. 2020 Dec 18;11:591434.
76. Mefford HC. The Road to Diagnosis: Shortening the Diagnostic Odyssey in Epilepsy. Epilepsy Curr. 2019 Sep;19(5):307–9.
77. Kjeldsen MJ, Corey LA, Christensen K, Friis ML. Epileptic seizures and syndromes in twins: the importance of genetic factors. Epilepsy Res. 2003 Jul;55(1–2):137–46.
78. The Epi4K Consortium, The Epilepsy Phenome/Genome Project. Ultra-rare genetic variation in common epilepsies: a case-control sequencing study. The Lancet Neurology. 2017 Feb;16(2):135–43.
79. Sheidley BR, Malinowski J, Bergner AL, Bier L, Gloss DS, Mu W, et al. Genetic testing for the epilepsies: A systematic review. Epilepsia. 2021 Dec 10;epi.17141.
80. Carvill GL, Engel KL, Ramamurthy A, Cochran JN, Roovers J, Stamberger H, et al. Aberrant Inclusion of a Poison Exon Causes Dravet Syndrome and Related *SCN1A*-Associated Genetic Epilepsies. The American Journal of Human Genetics. 2018 Dec;103(6):1022–9.
81. Ishiura H, Doi K, Mitsui J, Yoshimura J, Matsukawa MK, Fujiyama A, et al. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. Nat Genet. 2018 Apr;50(4):581–90.
82. Florian RT, Kraft F, Leitão E, Kaya S, Klebe S, et al. Unstable TTTTA/TTTCA expansions in *MARCH6* are associated with Familial Adult Myoclonic Epilepsy type 3. Nat Commun. 2019 Dec;10(1):4919.
83. Corbett MA, Kroes T, Veneziano L, Bennett MF, Florian R, Schneider AL, et al. Intronic ATTTC repeat expansions in *STARD7* in familial adult myoclonic epilepsy linked to chromosome 2. Nat Commun. 2019 Dec;10(1):4920.
84. Hirabayashi K, Uehara DT, Abe H, Ishii A, Moriyama K, Hirose S, et al. Copy number variation analysis in 83 children with early-onset developmental and epileptic encephalopathy after targeted resequencing of a 109-epilepsy gene panel. J Hum Genet. 2019 Nov;64(11):1097–106.

85. Niestroj L-M, Perez-Palma E, Howrigan DP, Zhou Y, Cheng F, Saarentaus E, et al. Epilepsy subtype-specific copy number burden observed in a genome-wide study of 17 458 subjects. Brain. 2020 Jul 1;143(7):2106–18.
86. Weber YG, Biskup S, Helbig KL, Von Spiczak S, Lerche H. The role of genetic testing in epilepsy diagnosis and management. Expert Review of Molecular Diagnostics. 2017 Aug 3;17(8):739–50.
87. Helbig KL, Farwell Hagman KD, Shinde DN, Mroske C, Powis Z, Li S, et al. Diagnostic exome sequencing provides a molecular diagnosis for a significant proportion of patients with epilepsy. Genet Med. 2016 Sep;18(9):898–905.
88. Ostrander BEP, Butterfield RJ, Pedersen BS, Farrell AJ, Layer RM, Ward A, et al. Whole-genome analysis for effective clinical diagnosis and gene discovery in early infantile epileptic encephalopathy. Genomic Med. 2018 Dec;3(1):22.
89. Thodeson DM, Park JY. Genomic testing in pediatric epilepsy. Cold Spring Harb Mol Case Stud. 2019 Aug;5(4):a004135.
90. Yuen RKC, Thiruvahindrapuram B, Merico D, Walker S, Tammimies K, Hoang N, et al. Whole-genome sequencing of quartet families with autism spectrum disorder. Nat Med. 2015 Feb;21(2):185–91.
91. Mohammadi L, Vreeswijk MP, Oldenburg R, van den Ouweland A, Oosterwijk JC, van der Hout AH, et al. A simple method for co-segregation analysis to evaluate the pathogenicity of unclassified variants; *BRCA1* and *BRCA2* as an example. BMC Cancer. 2009 Dec;9(1):211.
92. EPICURE Consortium, EMINet Consortium, Steffens M, Leu C, Ruppert A-K, Zara F, et al. Genome-wide association analysis of genetic generalized epilepsies implicates susceptibility loci at 1q43, 2p16.1, 2q22.3 and 17q21.32. Human Molecular Genetics. 2012 Dec 15;21(24):5359–72.
93. Zhang Y, Qu J, Mao C-X, Wang Z-B, Mao X-Y, Zhou B-T, et al. Novel Susceptibility Loci were Found in Chinese Genetic Generalized Epileptic Patients by Genome-wide Association Study. CNS Neurosci Ther. 2014 Nov;20(11):1008–10.
94. Wang M, Greenberg DA, Stewart WCL. Replication, reanalysis, and gene expression: *ME2* and genetic generalized epilepsy. Epilepsia. 2019 Feb 4;epi.14654.
95. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. The American Journal of Human Genetics. 2014 Jul;95(1):5–23.
96. Das S, McClain CJ, Rai SN. Fifteen Years of Gene Set Analysis for High-Throughput Genomic Data: A Review of Statistical Approaches and Future Challenges. Entropy. 2020 Apr 10;22(4):427.

97. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen — The Clinical Genome Resource. N Engl J Med. 2015 Jun 4;372(23):2235–42.
98. Kaplanis J, Samocha KE, Wiel L, Zhang Z, Arvai KJ, et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature. 2020 Oct 29;586(7831):757–62.
99. Strande NT, Riggs ER, Buchanan AH, Ceyhan-Birsoy O, DiStefano M, Dwight SS, et al. Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. Am J Hum Genet. 2017 Jun 1;100(6):895–906.
100. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). Online Mendelian Inheritance in Man, OMIM® [Internet]. 2021.
101. Subaran RL, Conte JM, Stewart WCL, Greenberg DA. Pathogenic *EFHC1* mutations are tolerated in healthy individuals dependent on reported ancestry. Epilepsia. 2015 Feb;56(2):188–94.
102. Calhoun JD, Huffman AM, Bellinski I, Kinsley L, Bachman E, Gerard E, et al. *CACNA1H* variants are not a cause of monogenic epilepsy. Human Mutation. 2020 Jun;41(6):1138–44.
103. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May 28;581(7809):434–43.
104. Karczewski KJ, Solomonson M, Chao KR, Goodrich JK, Tiao G, Lu W, et al. Systematic single-variant and gene-based association testing of 3,700 phenotypes in 281,850 UK Biobank exomes. medRxiv. 2021 Jun.
105. Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, et al. Using high-resolution variant frequencies to empower clinical genome interpretation. Genetics in Medicine. 2017 Oct;19(10):1151–8.
106. Genovese G, Fromer M, Stahl EA, Ruderfer DM, Chambert K, Landén M, et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. Nat Neurosci. 2016 Nov;19(11):1433–41.
107. Ganna A, Satterstrom FK, Zekavat SM, Das I, Kurki MI, Churchhouse C, et al. Quantifying the Impact of Rare and Ultra-rare Coding Variation across the Phenotypic Spectrum. Am J Hum Genet. 2018 Jun 7;102(6):1204–11.
108. Wilfert AB, Turner TN, Murali SC, Hsieh P, Sulovari A, Wang T, et al. Recent ultra-rare inherited variants implicate new autism candidate risk genes. Nat Genet. 2021 Jul 26.
109. McTague A, Howell KB, Cross JH, Kurian MA, Scheffer IE. The genetic landscape of the epileptic encephalopathies of infancy and childhood. The Lancet Neurology. 2016 Mar;15(3):304–16.

110. Striano P, Minassian BA. From Genetic Testing to Precision Medicine in Epilepsy. Neurotherapeutics. 2020 Apr;17(2):609–15.
111. Sisodiya SM. Precision medicine and therapies of the future. Epilepsia. 2021 Mar;62(S2).
112. Berkovic SF, Howell RA, Hay DA, Hopper JL. Epilepsies in twins: Genetics of the major epilepsy syndromes. Ann Neurol. 1998 Apr;43(4):435–45.
113. Zara F, Gennaro E, Stabile M, Carbone I, Malacarne M, Majello L, et al. Mapping of a Locus for a Familial Autosomal Recessive Idiopathic Myoclonic Epilepsy of Infancy to Chromosome 16p13. The American Journal of Human Genetics. 2000 May;66(5):1552–7.
114. Baykan B, Madia F, Bebek N, Gianotti S, Güney AI, Cine N, et al. Autosomal recessive idiopathic epilepsy in an inbred family from Turkey: identification of a putative locus on chromosome 9q32-33. Epilepsia. 2004 May;45(5):479–87.
115. Berkovic SF. Genetics of Epilepsy in Clinical Practice: Genetics of Epilepsy in Clinical Practice. Epilepsy Curr. 2015 Jul;15(4):192–6.
116. Thomas RH, Berkovic SF. The hidden genetics of epilepsy—a clinically important new paradigm. Nat Rev Neurol. 2014 May;10(5):283–92.
117. Robinson R. Current topic: Genetics of childhood epilepsy. Archives of Disease in Childhood. 2000 Feb 1;82(2):121–5.
118. Chentouf A, Dahdouh A, Guipponi M, Oubaiche ML, Chaouch M, Hamamy H, et al. Familial epilepsy in Algeria: Clinical features and inheritance profiles. Seizure. 2015 Sep;31:12–8.
119. De Falco FA, Majello L, Santangelo R, Stabile M, Bricarelli FD, Zara F. Familial Infantile Myoclonic Epilepsy: Clinical Features in a Large Kindred with Autosomal Recessive Inheritance. Epilepsia. 2001 Dec;42(12):1541–8.
120. DiFrancesco JC, Barbuti A, Milanesi R, Coco S, Bucchi A, Bottelli G, et al. Recessive Loss-of-Function Mutation in the Pacemaker *HCN2* Channel Causing Increased Neuronal Excitability in a Patient with Idiopathic Generalized Epilepsy. Journal of Neuroscience. 2011 Nov 30;31(48):17327–37.
121. Bittles AH, Black ML. Consanguinity, human evolution, and complex diseases. Proceedings of the National Academy of Sciences. 2010 Jan 26;107(1):1779–86.
122. Elsayed LEO, Mohammed IN, Hamed AAA, Elseed MA, Johnson A, Mairey M, et al. Hereditary spastic paraplegias: identification of a novel *SPG57* variant affecting TFG oligomerization and description of HSP subtypes in Sudan. Eur J Hum Genet. 2017 Jan;25(1):100–10.
123. Dobon B, Hassan HY, Laayouni H, Luisi P, Ricaño-Ponce I, Zhernakova A, et al. The genetics of East African populations: a Nilo-Saharan component in the African genetic landscape. Sci Rep. 2015 Sep;5(1):9996.

124. Mohammed IN, Abdel Moneim M, Abdel Rahman A. The profile of childhood epilepsy in Sudan. Khartoum Medical Journal. 2010;03(02):444–7.
125. Mohamed IN, Osman AH, Mohamed S, Hamid EK, Hamed AA, Alsir A, et al. Intelligence quotient (IQ) among children with epilepsy: National epidemiological study - Sudan. Epilepsy Behav. 2020 Feb;103(Pt A):106813.
126. Dahawi M, Elmagzoub MS, A. Ahmed E, Baldassari S, Achaz G, Elmugadam FA, et al. Involvement of *ADGRV1* Gene in Familial Forms of Genetic Generalized Epilepsy. Front Neurol. 2021 Oct 21;12:738272.
127. Elsayed LEO, Drouet V, Usenko T, Mohammed IN, Hamed AAA, Elseed MA, et al. A Novel Nonsense Mutation in *DNAJC6* Expands the Phenotype of Autosomal-Recessive Juvenile-Onset Parkinson's Disease. Ann Neurol. 2016 Feb;79(2):335–7.
128. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013 May 26.
129. Auwera GAV de, O'Connor BD. Genomics in the cloud: using Docker, GATK, and WDL in Terra. First edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly; 2020. 467 p.
130. Pujar S, O'Leary NA, Farrell CM, Loveland JE, Mudge JM, Wallin C, et al. Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. Nucleic Acids Research. 2018 Jan 4;46(D1):D221–8.
131. Tan A, Abecasis GR, Kang HM. Unified representation of genetic variants. Bioinformatics. 2015 Jul 1;31(13):2202–4.
132. Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, et al. HTSlib: C library for reading/writing high-throughput sequencing data. GigaScience. 2021 Jan 29;10(2):giab007.
133. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016 Dec;17(1).
134. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021 Feb 11;590(7845):290–9.
135. Dewey FE, Murray MF, Overton JD, Habegger L, Leader JB, Fetterolf SN, et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science. 2016 Dec 23;354(6319):aaf6814.
136. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLoS Computational Biology. 2010 Dec;6(12):e1001025.
137. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research. 2019 Jan 8;47(D1):D886–94.

138. Traynelis J, Silk M, Wang Q, Berkovic SF, Liu L, Ascher DB, et al. Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res. 2017 Oct;27(10):1715–29.
139. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. The American Journal of Human Genetics. 2016 Oct;99(4):877–85.
140. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016 Dec;48(12):1581–6.
141. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. The American Journal of Human Genetics. 2018 Oct;103(4):474–83.
142. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Research. 2014 Dec 16;42(22):13534–44.
143. Gelfman S, Wang Q, McSweeney KM, Ren Z, La Carpia F, Halvorsen M, et al. Annotating pathogenic non-coding variants in genic regions. Nat Commun. 2017 Aug 9;8(1):236.
144. Eng L, Coutinho G, Nahas S, Yeo G, Tanouye R, Babaei M, et al. Nonclassical splicing mutations in the coding and noncoding regions of the *ATM* Gene: Maximum entropy estimates of splice junction strengths: NONCLASSICAL ATM SPLICING MUTATIONS. Hum Mutat. 2004 Jan;23(1):67–76.
145. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020 Dec;12(1):103.
146. Affymetrix Inc. BRLMM: An Improved Genotype Calling Method for the Genechip Human Mapping 500k Array Set [White Paper]. 2006.
147. Seelow D, Schuelke M, Hildebrandt F, Nurnberg P. HomozygosityMapper--an interactive approach to homozygosity mapping. Nucleic Acids Research. 2009 Jul 1;37(Web Server):W593–9.
148. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, et al. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Research. 2007 Nov 1;17(11):1665–74.
149. Goode DL, Cooper GM, Schmutz J, Dickson M, Gonzales E, Tsai M, et al. Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. Genome Research. 2010 Mar 1;20(3):301–10.

150. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2013 Jan;41(Database issue):D8–20.
151. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research. 2018 Jan 4;46(D1):D1062–7.
152. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucl Acids Res. 2014 Jan;42(D1):D986–92.
153. Kleinberger J, Maloney KA, Pollin TI, Jeng LJB. An openly available online tool for implementing the ACMG/AMP standards and guidelines for the interpretation of sequence variants. Genet Med. 2016 Nov;18(11):1165.
154. Scheffer IE, Grinton BE, Heron SE, Kivity S, Afawi Z, Iona X, et al. *PRRT2* phenotypic spectrum includes sporadic and fever-related infantile seizures. Neurology. 2012 Nov 20;79(21):2104–8.
155. Gardiner AR, Bhatia KP, Stamelou M, Dale RC, Kurian MA, Schneider SA, et al. *PRRT2* gene mutations: From paroxysmal dyskinesia to episodic ataxia and hemiplegic migraine. Neurology. 2012 Nov 20;79(21):2115–21.
156. Riant F, Roze E, Barbance C, Meneret A, Guyant-Marechal L, Lucas C, et al. *PRRT2* mutations cause hemiplegic migraine. Neurology. 2012 Nov 20;79(21):2122–4.
157. Labate A, Tarantino P, Viri M, Mumoli L, Gagliardi M, Romeo A, et al. Homozygous c.649dupC mutation in *PRRT2* worsens the BFIS/PKD phenotype with mental retardation, episodic ataxia, and absences: *Homozygous PRRT2 Mutation and Mental Retardation*. Epilepsia. 2012 Dec;53(12):e196–9.
158. Landolfi A, Barone P, Erro R. The Spectrum of *PRRT2*-Associated Disorders: Update on Clinical Features and Pathophysiology. Front Neurol. 2021;12:629747.
159. Ebrahimi-Fakhari D, Saffari A, Westenberger A, Klein C. The evolving spectrum of *PRRT2*-associated paroxysmal diseases. Brain. 2015 Dec;138(Pt 12):3476–95.
160. Döring JH, Saffari A, Bast T, Brockmann K, Ehrhardt L, Fazeli W, et al. The Phenotypic Spectrum of *PRRT2*-Associated Paroxysmal Neurologic Disorders in Childhood. Biomedicines. 2020 Oct 28;8(11):E456.
161. El Achkar CM, Rosen Sheidley B, O'Rourke D, Takeoka M, Poduri A. Compound heterozygosity with *PRRT2*: Pushing the phenotypic envelope in genetic epilepsies. Epilepsy & Behavior Case Reports. 2019;11:125–8.
162. Delcourt M, Riant F, Mancini J, Milh M, Navarro V, Roze E, et al. Severe phenotypic spectrum of biallelic mutations in *PRRT2* gene. J Neurol Neurosurg Psychiatry. 2015 Jul;86(7):782–5.

163. Abramowicz A, Gos M. Splicing mutations in human genetic disorders: examples, detection, and confirmation. J Appl Genetics. 2018 Aug;59(3):253–68.

164. Raponi M, Baralle D. Alternative splicing: good and bad effects of translationally silent substitutions: Alternative splicing. FEBS Journal. 2010 Feb;277(4):836–40.

165. Gaildrat P, Killian A, Martins A, Tournier I, Frébourg T, Tosi M. Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants. Methods Mol Biol. 2010;653:249–57.

166. Smith L, Singhal N, El Achkar CM, Truglio G, Rosen Sheidley B, Sullivan J, et al. *PCDH19* - related epilepsy is associated with a broad neurodevelopmental spectrum. Epilepsia. 2018 Mar;59(3):679–89.

167. Liu A, Xu X, Yang X, Jiang Y, Yang Z, Liu X, et al. The clinical spectrum of female epilepsy patients with *PCDH19* mutations in a Chinese population: Clinical spectrum of female epilepsy patients with *PCDH19* mutations. Clin Genet. 2017 Jan;91(1):54–62.

168. van Harssel JJT, Weckhuysen S, van Kempen MJA, Hardies K, Verbeek NE, de Kovel CGF, et al. Clinical and genetic aspects of *PCDH19*-related epilepsy syndromes and the possible role of *PCDH19* mutations in males with autism spectrum disorders. Neurogenetics. 2013 Feb;14(1):23–34.

169. Cooper SR, Jontes JD, Sotomayor M. Structural determinants of adhesion by Protocadherin-19 and implications for its role in epilepsy. eLife. 2016 Oct 26;5:e18529.

170. Syrbe S, Harms FL, Parrini E, Montomoli M, Mütze U, Helbig KL, et al. Delineating *SPTAN1* associated phenotypes: from isolated epilepsy to encephalopathy with progressive brain atrophy. Brain. 2017 Sep 1;140(9):2322–36.

171. Zaman T, Helbig KL, Clatot J, Thompson CH, Kang SK, Stouffs K, et al. *SCN3A*-Related Neurodevelopmental Disorder: A Spectrum of Epilepsy and Brain Malformation. Ann Neurol. 2020 Aug;88(2):348–62.

172. Platzer K, Yuan H, Schütz H, Winschel A, Chen W, Hu C, et al. *GRIN2B* encephalopathy: novel findings on phenotype, variant clustering, functional consequences and treatment aspects. J Med Genet. 2017 Jul;54(7):460–70.

173. Landoulsi Z, Laatar F, Noé E, Mrabet S, Ben Djebara M, Achaz G, et al. Clinical and genetic study of Tunisian families with genetic generalized epilepsy: contribution of *CACNA1H* and *MAST4* genes. Neurogenetics. 2018 Aug;19(3):165–78.

174. Peljto AL, Barker-Cummings C, Vasoli VM, Leibson CL, Hauser WA, Buchhalter JR, et al. Familial risk of epilepsy: a population-based study. Brain. 2014 Mar;137(3):795–805.

175. de Kovel CGF, Trucks H, Helbig I, Mefford HC, Baker C, Leu C, et al. Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. Brain. 2010 Jan 1;133(1):23–32.
176. Mefford HC, Muhle H, Ostertag P, von Spiczak S, Buysse K, Baker C, et al. Genome-Wide Copy Number Variation in Epilepsy: Novel Susceptibility Loci in Idiopathic Generalized and Focal Epilepsies. Frankel WN, editor. PLoS Genet. 2010 May 20;6(5):e1000962.
177. Ren Z, Povysil G, Hostyk JA, Cui H, Bhardwaj N, Goldstein DB. ATAV: a comprehensive platform for population-scale genomic analyses. BMC Bioinformatics. 2021 Dec;22(1):149.
178. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucl Acids Res. 2014 Jan;42(D1):D975–9.
179. Goyal A, Kwon HJ, Lee K, Garg R, Yun SY, Hee Kim Y, et al. Ultra-Fast Next Generation Human Genome Sequencing Data Processing Using DRAGEN™ Bio-IT Processor for Precision Medicine. OJGen. 2017;07(01):9–19.
180. Zhao S, Agafonov O, Azab A, Stokowy T, Hovig E. Accuracy and efficiency of germline variant calling pipelines for human genome data. Sci Rep. 2020 Dec;10(1):20222.
181. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010 Sep 1;20(9):1297–303.
182. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, et al. Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. The American Journal of Human Genetics. 2012 Nov;91(5):839–48.
183. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010 Nov 15;26(22):2867–73.
184. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genet. 2006;2(12):e190.
185. Wolking S, Moreau C, McCormack M, Krause R, Krenn M, EpiPGx Consortium, et al. Assessing the role of rare genetic variants in drug-resistant, non-lesional focal epilepsy. Ann Clin Transl Neurol. 2021 Jul;8(7):1376–87.
186. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaSci. 2015 Dec;4(1):7.
187. Meyer D, Dimitriadou E, Hornik K, Weingessel A, and Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-3. The Comprehensive R Archive Network. 2019.
188. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

189. Turchin MC, Hirschhorn JN. Gencrypt: one-way cryptographic hashes to detect overlapping individuals across samples. Bioinformatics. 2012 Mar 15;28(6):886–8.
190. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. 2012 Apr;6(2):80–92.
191. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–9.
192. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research. 2010 Sep 1;38(16):e164–e164.
193. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2017.
194. Petrovski S, Wang Q. QQperm: Permutation Based QQ Plot and Inflation Factor Estimation. 2016.
195. Dowle M, Srinivasan A. data.table: Extension of `data.frame`. The Comprehensive R Archive Network. 2019.
196. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. JOSS. 2019 Nov 21;4(43):1686.
197. Kananura C, Haug K, Sander T, Runge U, Gu W, Hallmann K, et al. A Splice-Site Mutation in *GABRG2* Associated With Childhood Absence Epilepsy and Febrile Convulsions. Arch Neurol. 2002 Jul 1;59(7):1137.
198. Baulac S, Huberfeld G, Gourfinkel-An I, Mitropoulou G, Beranger A, Prud'homme J-F, et al. First genetic evidence of $GABA_A$ receptor dysfunction in epilepsy: a mutation in the γ2-subunit gene. Nat Genet. 2001 May;28(1):46–8.
199. Wallace RH, Marini C, Petrou S, Harkin LA, Bowser DN, Panchal RG, et al. Mutant $GABA_A$ receptor γ2-subunit in childhood absence epilepsy and febrile seizures. Nat Genet. 2001 May;28(1):49–52.
200. Marini C, Harkin LA, Wallace RH, Mulley JC, Scheffer IE, Berkovic SF. Childhood absence epilepsy and febrile seizures: a family with a $GABA_A$ receptor mutation. Brain. 2003 Jan 1;126(1):230–40.
201. Boillot M, Morin-Brureau M, Picard F, Weckhuysen S, Lambrecq V, Minetti C, et al. Novel *GABRG2* mutations cause familial febrile seizures. Neurol Genet. 2015 Dec;1(4):e35.
202. Audenaert D, Schwartz E, Claeys KG, Claes L, Deprez L, Suls A, et al. A novel *GABRG2* mutation associated with febrile seizures. Neurology. 2006 Aug 22;67(4):687–90.

203. Skotte L, Fadista J, Bybjerg-Grauholm J, Appadurai V, Hildebrand MS, Hansen TF, et al. Genome-wide association study of febrile seizures identifies seven new loci implicating fever response and neuronal excitability genes. Brain. 2022;145(2):555–68.
204. EpiPM Consortium. A roadmap for precision medicine in the epilepsies. The Lancet Neurology. 2015 Dec;14(12):1219–28.
205. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74.
206. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research. 2019 Jan 8;47(D1):D766–73.
207. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, et al. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol. 2019 May;37(5):555–60.
208. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006 Aug;38(8):904–9.
209. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.
210. Demontis D, Satterstrom K, Duan J, Lescai F, Dinesen Østergaard S, Lesch K-P, et al. The Role of Ultra-Rare Coding Variants In ADHD. European Neuropsychopharmacology. 2019;29:S724–5.
211. Singh T, Walters JTR, Johnstone M, Curtis D, et al. The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. Nat Genet. 2017 Aug;49(8):1167–73.
212. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Research. 2012 Jul 1;40(W1):W452–7.
213. Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, et al. Regional missense constraint improves variant deleteriousness prediction. bioRxiv. 2017 Jun.
214. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome. Nat Genet. 2019 Jan;51(1):88–95.
215. Lal D, May P, Perez-Palma E, Samocha KE, Kosmicki JA, et al. Gene family information facilitates variant interpretation and identification of disease-associated genes in neurodevelopmental disorders. Genome Med. 2020 Dec;12(1):28.

216. Dickerson JE, Robertson DL. On the Origins of Mendelian Disease Genes in Man: The Impact of Gene Duplication. Molecular Biology and Evolution. 2012 Jan 1;29(1):61–9.
217. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014 Sep;46(9):944–50.
218. Genotype Tissue Expression Portal. [cited 2021 Jan 31]. Available from: https://www.gtexportal.org
219. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25–9.
220. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Research. 2021 Jan 8;49(D1):D325–34.
221. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems. 2015 Dec;1(6):417–25.
222. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Research. 2021 Jan 8;49(D1):D545–51.
223. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Research. 2019 Nov 6;gkz1031.
224. Johnson MR, Behmoaras J, Bottolo L, Krishnan ML, Pernhorst K, Santoscoy PLM, et al. Systems genetics identifies Sestrin 3 as a regulator of a proconvulsant gene network in human epileptic hippocampus. Nat Commun. 2015 May;6(1):6031.
225. Delahaye-Duriez A, Srivastava P, Shkura K, Langley SR, Laaniste L, Moreno-Moral A, et al. Rare and common epilepsies converge on a shared gene regulatory network providing opportunities for novel antiepileptic drug discovery. Genome Biol. 2016 Dec;17(1):245.
226. Koko M, Krause R, Sander T, Bobbili DR, Nothnagel T, May P, et al. Supplements to the article: Distinct gene-set enrichment patterns underlie common generalized and focal epilepsies. Mendeley; 2021.
227. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009 Aug;4(8):1184–91.
228. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler—a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 2016 Jul 8;44(W1):W83–9.
229. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. The American Journal of Human Genetics. 2012 Aug;91(2):224–37.
230. Zeileis, Hothorn. Diagnostic Checking in Regression Relationships. Rnews. 2002.

231. Lerche H, Shah M, Beck H, Noebels J, Johnston D, Vincent A. Ion channels in genetic and acquired forms of epilepsy: Ion channels in epilepsy. The Journal of Physiology. 2013 Feb;591(4):753–64.
232. Vacher H, Mohapatra DP, Trimmer JS. Localization and Targeting of Voltage-Dependent Ion Channels in Mammalian Central Neurons. Physiological Reviews. 2008 Oct;88(4):1407–47.
233. Martenson JS, Tomita S. Synaptic localization of neurotransmitter receptors: comparing mechanisms for AMPA and $GABA_A$ receptors. Current Opinion in Pharmacology. 2015 Feb;20:102–8.
234. Craig AM, Kang Y. Neurexin–neuroligin signaling in synapse development. Current Opinion in Neurobiology. 2007 Feb;17(1):43–52.
235. Takata A, Ionita-Laza I, Gogos JA, Xu B, Karayiorgou M. De Novo Synonymous Mutations in Regulatory Elements Contribute to the Genetic Etiology of Autism and Schizophrenia. Neuron. 2016 Mar 2;89(5):940–7.
236. Myers CT, McMahon JM, Schneider AL, Petrovski S, Allen AS, Carvill GL, et al. De Novo Mutations in *SLC1A2* and *CACNA1A* Are Important Causes of Epileptic Encephalopathies. The American Journal of Human Genetics. 2016 Aug;99(2):287–98.
237. Helbig KL, Lauerer RJ, Bahr JC, Souza IA, Myers CT, Uysal B, et al. De Novo Pathogenic Variants in *CACNA1E* Cause Developmental and Epileptic Encephalopathy with Contractures, Macrocephaly, and Dyskinesias. The American Journal of Human Genetics. 2018 Nov;103(5):666–78.
238. Takata A, Nakashima M, Saitsu H, Mizuguchi T, Mitsuhashi S, Takahashi Y, et al. Comprehensive analysis of coding variants highlights genetic complexity in developmental and epileptic encephalopathy. Nat Commun. 2019 Dec;10(1):2506.
239. Appenzeller S, Balling R, Barisic N, Baulac S, Caglayan H, Craiu D, et al. De Novo Mutations in Synaptic Transmission Genes Including *DNM1* Cause Epileptic Encephalopathies. The American Journal of Human Genetics. 2014 Oct;95(4):360–70.
240. De Rubeis S, He X, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014 Nov;515(7526):209–15.
241. The Brainstorm Consortium. Analysis of shared heritability in common disorders of the brain. Science. 2018 Jun 22;360(6395):eaap8757.
242. Liu Y, Schubert J, Sonnenberg L, Helbig KL, Hoei-Hansen CE, Koko M, et al. Neuronal mechanisms of mutations in *SCN8A* causing epilepsy or intellectual disability. Brain. 2019 Feb 1;142(2):376–90.

243. Vardar G, Gerth F, Schmitt XJ, Rautenstrauch P, Trimbuch T, Schubert J, et al. Epilepsy-causing *STX1B* mutations translate altered protein functions into distinct phenotypes in mouse neurons. Brain. 2020 Jul 1;143(7):2119–38.

244. de Leeuw CA, Neale BM, Heskes T, Posthuma D. The statistical properties of gene-set analysis. Nat Rev Genet. 2016 Jun;17(6):353–64.

245. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug;536(7616):285–91.

246. Koko M, Yahia A, Elsayed LE, Hamed AA, Mohammed IN, Elseed MA, et al. An identical-by-descent novel splice-donor variant in *PRUNE1* causes a neurodevelopmental syndrome with prominent dystonia in two consanguineous Sudanese families. Annals of Human Genetics. 2021 Jun 10;ahg.12437.

247. Elsayed LEO, Mohammed IN, Hamed AAA, Elseed MA, Salih MAM, Yahia A, et al. Novel Homozygous Missense Mutation in the *ARG1* Gene in a Large Sudanese Family. Front Neurol. 2020 Oct 29;11:569996.

248. Yahia A, Elsayed L, Babai A, Salih MA, El-Sadig SM, Amin M, et al. Intra-familial phenotypic heterogeneity in a Sudanese family with *DARS2*-related leukoencephalopathy, brainstem and spinal cord involvement and lactate elevation: a case report. BMC Neurol. 2018 Dec;18(1):175.

249. Cauley ES, Hamed A, Mohamed IN, Elseed M, Martinez S, Yahia A, et al. Overlap of polymicrogyria, hydrocephalus, and Joubert syndrome in a family with novel truncating mutations in *ADGRG1*/*GPR56* and *KIAA0556*. Neurogenetics. 2019 May;20(2):91–8.

250. Amin M, Bakhit Y, Koko M, Ibrahim MOM, Salih MA, Ibrahim M, et al. Rare variant in *LAMA2* gene causing congenital muscular dystrophy in a Sudanese family. A case report. Acta Myol. 2019 Mar;38(1):21–4.

251. Amin M, Vignal C, Hamed AAA, Mohammed IN, Elseed MA, Drunat S, et al. Novel variants causing megalencephalic leukodystrophy in Sudanese families. J Hum Genet. 2021 Sep 10.

252. Bailey JN, Patterson C, de Nijs L, Durón RM, Nguyen V-H, Tanaka M, et al. *EFHC1* variants in juvenile myoclonic epilepsy: reanalysis according to NHGRI and ACMG guidelines for assigning disease causality. Genet Med. 2017 Feb;19(2):144–56.

253. Battke F, Schulte B, Schulze M, Biskup S. The question of WGS's clinical utility remains unanswered. Eur J Hum Genet. 2021 May;29(5):722–3.

254. Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, Tachmazidou I, et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. Nature. 2021 Aug 10.

255. Heyne HO, Karjalainen J, Karczewski KJ, Lemmelä SM, Zhou W, FinnGen, et al. Mono- and bi-allelic effects of coding variants on disease in 176,899 Finns. medRxiv. 2021 Nov 11.

256. Symonds JD, Zuberi SM, Stewart K, McLellan A, O'Regan M, MacLeod S, et al. Incidence and phenotypes of childhood-onset genetic epilepsies: a prospective population-based national cohort. Brain. 2019 Aug 1;142(8):2303–18.
257. Pal DK, Durner M, Klotz I, Dicker E, Shinnar S, Resor S, et al. Complex inheritance and parent-of-origin effect in juvenile myoclonic epilepsy. Brain and Development. 2006 Mar;28(2):92–8.
258. Guo MH, Plummer L, Chan Y-M, Hirschhorn JN, Lippincott MF. Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. The American Journal of Human Genetics. 2018 Oct;103(4):522–34.
259. Lee S, Kim S, Fuchsberger C. Improving power for rare-variant tests by integrating external controls. Genet Epidemiol. 2017 Nov;41(7):610–9.
260. Hendricks AE, Billups SC, Pike HNC, Farooqi IS, Zeggini E, Santorico SA, et al. ProxECAT: Proxy External Controls Association Test. A new case-control gene region association test using allele frequencies from public controls. Borecki I, editor. PLoS Genet. 2018 Oct 16;14(10):e1007591.
261. He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet. 2013;9(8):e1003671.
262. Venkataraman GR, DeBoever C, Tanigawa Y, Aguirre M, Ioannidis AG, Mostafavi H, et al. Bayesian model comparison for rare-variant association studies. The American Journal of Human Genetics. 2021 Dec;108(12):2354–67.

Statement of Contributions

This dissertation presents results from collaborative studies in which I made significant contributions to the study design, data collection, analysis, interpretation, and writing. Unless otherwise specified with appropriate citations, the presented text, tables, and figures constitutes original content that I wrote, designed, or created. The original work from which several chapters were adapted benefited from valuable and generous intellectual input contributed by my advisors, my colleagues, and my collaborators, as will be outlined below. Prof. Dr. Holger Lerche, Dr. Ulrike Hedrich-Klimosch, and Johanna Krüger assisted with proofreading this thesis and made several valuable suggestions regarding its content.

Studying the association of bi-allelic coding variants with familial epilepsy in Sudanese families

Study framework: This part presents a family-based study of several Sudanese families using exome sequencing and genome-wide genotyping. This chapter was adapted from: Koko *et al*. In preparation. 2022.

Contributors: Mahmoud Koko (MK), Ashraf Yahia (AY), Liena Elsayed (LE), Ahlam Hamed (AH), Maha Elseed (ME), Inaam Mohamed (IM), Janine Altmüller (JA), Mohamed Toliat (MT), Julian Schubert (JS), Holger Lerche (HL), and the Sudanese neurogenetics research group (NGS) members and collaborators. The members/collaborators of NGS are listed as in these articles: PMID:34111303; PMID:30352563, and PMID:29739362.

Contributions: Conception, planning and design: AY, LE, HL, MK, JS. Patients' evaluation: AH, ME, IM, AY, MK, and others from NGS. Sampling and data acquisition: AY, MK, and others from NGS. Exome sequencing and genotyping: MK, JS, JA, MT. Data analysis: MK. Data interpretation: MK, JS, HL. Writing – first draft: MK. Writing – revisions: MK, HL, AY, LE, ME.

Studying the association of coding variants with familial and sporadic generalized epilepsy

Study framework: The study presents a joint analysis of several exome datasets from individuals with epilepsy previously recruited by the Canadian Epilepsy Network (CENet), Epi4K Consortium (Epi4K), Epilepsy Phenome/Genome Project (EP/GP), EpiPGX

Consortium (EpiPGX), and EuroEPINOMICS CoGIE Consortium (CoGIE), and controls from the National Center for Biotechnology Information database of Genotypes and Phenotypes (NCBI dbGAP) and Institute for Genomic Medicine (IGM, NY, USA). This chapter was adapted from: Koko *et al*. *Epilepsia*. 2022. PMID:35032048.

Contributors: Mahmoud Koko (MK), Joshua E. Motelow (JEM), Kate E. Stanley (KES), Dheeraj R. Bobbili (DRB), Ryan S. Dhindsa (RSD), Patrick May (PM), Simon Girard and Patrick Cossette for CENet, Samuel F. Berkovic (SFB), Daniel H. Lowenstein (DHL) and David B. Goldstein (DBG) for the Epi4K and EP/GP, Holger Lerche (HL) for EpiPGX and CoGIE. The members of the consortia are listed here: PMID:35032048.

Contributions: Conception, planning, and design: SFB, HL, DBG, DHL, PM, KES, RSD. Patients' and sequence data: CENet, Epi4K, EP/GP, EpiPGX, CoGIE. Exome data acquisition: KES, RSD, DB, PM, SG. Data analysis: MK, JEM, KES, DRB, RSD, PM. Data interpretation: MK, JEM, KES, RSD, PM, HL, DBG, SFB, DHL. Writing – first draft: MK. Writing – revisions: MK, JEM, KES, PM, HL, SFB, DHL.

Studying the association of coding variation in biologically informed gene sets with common and rare epilepsies

Study framework: The study is a secondary analysis of an exome dataset from individuals with epilepsy recruited by the Epi25 Collaborative. This chapter was adapted from: Koko *et al*. *EBioMedicine*. 2021. PMID:34571366.

Contributors: Mahmoud Koko (MK), Roland Krause (RK), Thomas Sander (TS), Dheeraj Reddy Bobbili (DRB), Michael Nothnagel (MN), Patrick May (PM), Holger Lerche (HL), and Epi25 Collaborative (Epi25). The members of Epi25 are listed here: PMID:34571366.

Contributions: Conception, planning and design: HL, PM, MK. Patients' and sequence data: Epi25. Data acquisition: RK, HL, PM, MK. Data analysis: MK, PM. Data interpretation: MK, HL, PM, TS, MN. Writing – first draft: MK. Writing – revisions: MK, HL, PM, TS, MN, RK DRB.

Acknowledgement

This doctoral work was supported by personal funding scholarships from the German Academic Exchange Office (DAAD Research Grants - Doctoral Programmes in Germany, April 2016 – September 2020) and Eberhard Karls University of Tübingen (Universitätsklinikum Tübingen, November 2020 – October 2021).