Title

Sub-Genic Intolerance, ClinVar and the Epilepsies:
A Whole Exome Sequencing Study of
29,165 Individuals

Author list

Epi25 Collaborative

Affiliations

See Supplemental Subjects and Methods for full author list and affiliations

Corresponding Authors

Joshua E. Motelow, MD, PhD
Department of Genetics and Development, Columbia University Irving Medical Center
Hammer Health Sciences
701 West 168th Street
New York, New York 10032, USA

David B. Goldstein, PhD
Department of Genetics and Development, Columbia University Irving Medical Center
Hammer Health Sciences
701 West 168th Street
New York, New York 10032, USA

Correspondence Emails

jm4279@cumc.columbia.edu, dg2875@cumc.columbia.edu

Abstract

Both mild and severe epilepsies are influenced by variants in the same genes yet an explanation for the resulting phenotypic variation is unknown. As part of the ongoing Epi25 Collaboration, we performed a whole-exome sequencing analysis of 13,487 epilepsy-affected individuals and 15,678 controls. While prior Epi25 studies focused on gene-based collapsing analyses, we asked how the pattern of variation within genes differs by epilepsy type. Specifically, we compared the genetic architectures of severe developmental and epileptic encephalopathies (DEEs) and two generally less severe epilepsies, genetic generalized epilepsy and non-acquired focal epilepsy (NAFE). Our gene-based rare variant collapsing analysis used geographic ancestry-based clustering which included broader ancestries than previously possible and revealed novel associations. Using the missense intolerance ratio (MTR), we found that variants in DEE cases are in significantly more intolerant genic sub-regions than those in NAFE cases. Only previously reported pathogenic variants absent in available genomic datasets showed a significant burden in epilepsy cases compared to controls, and the ultra-rare pathogenic variants associated with DEE were located in more intolerant genic sub-regions than variants associated with non-DEE epilepsies. MTR filtering improved the yield of ultra-rare pathogenic variants in cases compared to controls. Finally, analysis of variants in genes without a disease association revealed a significant burden of loss-of-function variants in the genes most intolerant to such variation, indicating additional epilepsy-risk genes yet to be discovered. Taken together, our study suggests that genic and sub-genic intolerance are critical characteristics for interpreting the effects of variation in genes that influence epilepsy.

Introduction

Epilepsy is a clinical diagnosis in which the individual has an enduring predisposition to seizures. Although the most severe types most commonly begin in childhood with profound impact, epilepsies can begin at any age with a cumulative incidence approaching 4%.[1-3] While the genetics of the epilepsies are complex, uncovering pathogenic variants can, in some cases, provide opportunities for targeted or precision medicines.[4; 5] Whole exome sequencing case-control studies have led to multiple insights into the epilepsies such as the contribution of *de novo* variants in developmental and epileptic encephalopathy (DEE [MIM: 308350]), the role of the GABA pathway in genetic generalized epilepsy (GGE [MIM: 600669]), and the link between non-acquired focal epilepsy (NAFE [MIM: 604364, 245570]) with GATOR1 complex genes.[6-10] DEEs are a severe form of early onset, intractable epilepsy associated with developmental delay.[8; 11-14] In contrast, GGE and NAFE, characterized by generalized seizures and focal seizures, respectively, are more common and generally less severe.[1; 2; 15-17] Yet, exome sequencing has revealed that a set of 43 genes typically associated with DEE also harbor ultra-rare variants in milder epilepsies.[7; 9]

It is unknown how these variants cause such different epilepsy phenotypes despite being drawn from a set of shared genes and even from within the same gene. The likelihood of a gene being associated with disease can be predicted *in silico*, in part, by a given gene's intolerance to functional variation in the general population.[18-20] Epilepsy causing variants tend to be rare in the general population and located in the least tolerant genes.[7; 9; 18; 21] While genic intolerance may help determine the likelihood of a gene-disease association, it does not clarify the differential impact of variants within the same gene.[22] Variants within the same gene may lead to widely different epilepsy phenotypes.[23-29] To predict the differential effects of two variants within the same gene requires an understanding of sub-genic intolerance as different regions or domains will have varied importance for the protein's function and may therefore contribute differentially to disease phenotype or severity.[22] Consistent with this idea, distributions of disease mutations often cluster in specific genic sub-regions.[30] In general, epilepsy variants cluster in the most intolerant genic sub-regions.[22; 31-33] The relationship between the severity of epilepsy caused by *SCN2A v*ariants and sub-genic intolerance has been explored,[32] but a more systematic study of the association of sub-genic intolerance and epilepsy severity has not been undertaken. Given that a single variant may lead to variable phenotypes,[34-37] we do not expect sub-genic intolerance to explain all severity variability, but a deeper investigation will add to our understanding of the complex sequelae of a single variant.

The Epi25 Collaborative (Epi25) is the largest epilepsy exome analysis to date with more than 200 partners from 40 research cohorts contributing exome and phenotype data from more than 19,000 individuals with epilepsy (see Web Resources). The aspiration of the collaborative is that extensive exome data combined with accurate phenotypic data will allow for well-matched cohorts and clarify genotype-phenotype relationships in epilepsy and has already yielded rich results for rare variants in the epilepsies.[9] A dataset of this magnitude and detail allows us to examine the presence of curated variants from a clinical database such as ClinVar.[38; 39] Similarly, we are able to test for the burden of damaging variants in the ~15,000 genes not-yet-associated with Mendelian disease to detect the potential for epilepsy-gene discovery. Combining expansive genetic data from Epi25 and recently developed sub-genic intolerance metrics, we show that in a set of genes harboring missense variants in both milder and more severe epilepsies, variants in more severe epilepsies are preferentially located in less tolerant genic sub-regions. Furthermore, only ultra-rare (i.e. not found in a public database) Pathogenic/Likely Pathogenic[40] ClinVar variants are increased in our cohort, and our sub-genic intolerance finding is replicated in these ultra-rare variants. Finally, there likely remain undiscovered epilepsy or epilepsy-risk genes among the genes most intolerant to loss-of-function variation.

Subjects and Methods

Study Design and Participants

As described previously, we collected DNA and detailed phenotyping data on individuals with epilepsy from 40 sites in Europe, North America, Australasia, and Asia (Table S1).[9] Here we analyzed individuals with DEEs (n = 2,007), GGE (also known as idiopathic generalized epilepsy; n = 5,771), and NAFE (n = 7,489) accounting for the first three years of enrollment in Epi25. A subset of the data is available on dbGaP: phs001489. Following sample quality control, relatedness testing (see Sample and Variant Quality Control [QC]), and clustering (see Clustering), the combined epilepsy analysis included 13,171 cases (1,782, 5,048, and 6,341 subjects with DEE, GGE, and NAFE, respectively) along with 14,100 controls (2,048 genomes and 12,052 exomes). In the included clusters in the individual epilepsy analyses, 1,835 subjects with DEE were compared to 13,978 controls, 5,303 subjects with GGE were compared to 15,677 controls and 6,439 subjects with NAFE were compared to 15,678 controls. Control samples were aggregated from local collections at the Institute of Genomic Medicine at Columbia University Irving Medical Center. Controls who passed the same quality control and who were not known to have phenotypes overlapping DEE, GGE, or NAFE, or be related to a proband with epilepsy were analyzed following geographic ancestry clustering (Figure S1, Table S2).

Phenotyping Procedures

As described previously, epilepsies were clinically diagnosed by epileptologists (see below for criteria DEEs, GGE, NAFE) in accordance with the International League Against Epilepsy (ILAE) classification at the time of diagnosis and recruitment.[2; 9] De-identified (non-PHI [protected health information]) phenotyping data were entered into the Epi25 Data repository (hosted at the Luxembourg Centre for Systems Biomedicine) via online case record forms based on the RedCAP platform. De-identified data for subjects of previous coordinated efforts with phenotyping (e.g., the

Epilepsy Phenome/Genome Project[41] and the EpiPGX Project, see Web Resources) which were already entered into a database were accessed and transferred to the new platform. Phenotyping data underwent review for uniformity among sites and quality control, and inconsistencies were reviewed by the phenotyping committee.

Case Definitions

Epilepsy diagnoses and classification for Epi25 have been described previously.[9] Briefly, DEE diagnosis required severe refractory epilepsy of unknown etiology with developmental plateau or regression and epileptiform features on EEG. Exclusion criteria included epileptogenic lesions on MRI. GGE diagnosis required a history of generalized seizure types with generalized epileptiform discharges on EEG. Exclusion criteria include focal seizures, moderate-to-severe intellectual disability and epileptogenic lesions found on neuroimaging (when available). Diagnosis of NAFE required a history of focal seizures with either focal epileptiform discharges or normal finds on EEG. Exclusion criteria included neuroimaging lesions (except hippocampal sclerosis), a history of generalized seizures, and moderate-to-severe intellectual disability.

Informed Consent

Adult subjects or the legal guardian for enrolled children signed informed consent at participating centers per the ethical requirements of the local rules at the time of enrollment.[9] The consent must not exclude data sharing to be included in the study. Consent forms for samples collected after January 25, 2015 required specific language according to the National Institutes of Health’s Genomic Data Sharing policy (see Web Resources). For control individuals, protocols were approved by Columbia University’s institutional review board and participants provided informed consent for the use of DNA in genetic research.

Next-Generation Sequencing Data Generation

All Epi25 samples were sequenced at the Broad Institute of Harvard and the Massachusetts Institute of Technology (MIT) on the Illumina HiSeq X platform, with the use of 151 bp paired-end reads. Exome capture was performed with Illumina Nextera Rapid Capture or TruSeq Rapid Exome enrichment kit (target size 38 Mb). FastQ files were transferred to the Institute for Genomic Medicine at Columbia University (IGM - Columbia University, New York, NY, USA).

Next-generation sequencing of controls was performed at the IGM and were a mixture of whole genome sequencing and whole exome sequencing. Exomes were captured with multiple capture kits and sequenced according to standard protocols on Illumina's HiSeq 2000, HiSeq 2500 and NovaSeq 6000 (Illumina, San Diego, CA, USA) platform with 150 bp paired-end reads. Genomes were sequenced according to standard protocols on Illumina's HiSeq 2000, HiSeq 2500 and NovaSeq 6000 (Illumina, San Diego, CA, USA) platform.

Variant Calling

Both cases and controls were processed with the same IGM bioinformatic pipeline for variant calling. Reads were aligned to human reference GRCh37 using DRAGEN (Edico Genome, San Diego, CA, USA)[42] and duplicates were marked with Picard (Broad Institute, Boston, MA, USA). Variants were called according to the Genome Analysis Toolkit (GATK - Broad Institute, Boston, MA, USA) Best Practices recommendations v3.6.[43; 44] Finally, variants were annotated with ClinEff[45] and custom annotations including Genome Aggregation Database (gnomAD) v2.1 frequencies[20], regional-intolerance metrics,[31; 32] *in silico* filters,[46] and ClinVar (as of 10/20/2020)[38; 39] clinical annotation were added using the IGM's in-house ATAV platform.[47]

Sample and Variant Quality Control (QC)

Only samples with at least 90% of the consensus coding sequence (CCDS release 20)[48] covered at a minimum of 10x, less or equal 2% contamination levels according to VerifyBamID[49], and single nucleotide variants (SNVs) and indels overlapping the Single Nucleotide Polymorphism database (dbSNP)[50] at least 85% and 80%, respectively were included. Samples with a discordance between self-declared and sequence-derived gender were removed to prevent phenotype-genotype mismatch. KING was used to detect related individuals and removed one of each pair that had an inferred relationship of second-degree or closer while favoring the inclusion of cases over controls and well-covered over poorly-covered.[51]

Analyses were restricted to variants within the CCDS inclusive of two base intronic extensions to accommodate canonical splice variants. All included variants had to fulfill the following criteria to be included: i) at least 10x coverage of the site, ii) quality score (QUAL) ≥ 50, iii) genotype quality score (GQ) ≥ 20, iv) quality by depth score (QD) ≥ 5, v) mapping quality score (MQ) ≥ 40, vi) read position rank sum score (RPRS) ≥ -3, vii) mapping quality rank sum score (MQRS) ≥ -10, viii) Fisher’s strand bias score (FS) ≤ 60 (SNVs) or ≤ 200 (indels), ix) strand odds ratio (SOR) ≤ 3 (SNVs) or ≤ 10 (indels), x) GATK Variant Quality Score Recalibration filter “PASS”, and xi), alternate allele fraction for heterozygous calls ≥ 0.3. Known sequencing artifacts as described previously[52] as well as low quality variants per Exome Aggregation Consortium[53], gnomAD[20], or the Exome Variant Server were excluded (see Web Resources).

Clustering

As previously described by Cameron-Christie and colleagues, Principal Component Analysis (PCA) for dimensionality reduction was performed on a set of predefined variants to capture population structure.[54] The Louvain method of community detection using the first six principal components (PCs) as input was applied to identify clusters within the data that reflect the geographic ancestry of the samples as previously described.[55; 56] To check the quality of the clusters, we performed further dimensionality reduction using the Uniform Manifold Approximation and Projection (UMAP)[57] on the first six PCs (Figure S1A-C) to disentangle subcontinental structure, which is then reflected in the cluster membership.[58; 59] A neural-network pre-trained on samples with known geographic ancestry generated probability estimates for each of six groups (European, African, Latino, East Asian, South Asian and Middle Eastern). A 95% probability cut-off was used to assign a geographic ancestry label to each sample. Samples that did not reach 95% for any of the ancestry groups were labelled “Admixed” (Figure S1).

Clustering was performed on the combined epilepsies as previously described.[56] Clusters containing at least 20 cases in each epilepsy type (DEE, GGE, NAFE) and 20 controls were kept (Figure S1C, Table S3). Each epilepsy type/control group separately underwent clustering again to optimize ancestry matching for each epilepsy type (Figure S1D-L). The individual epilepsy clustering was used for individual epilepsy QQ plots (Figure 1), the analysis of common enrichment among DEE genes (Figure 2), and associated supplementary figures and tables. The combined epilepsy analysis was used for the combined epilepsy collapsing analysis, sub-genic comparisons, and ClinVar Pathogenic/Likely Pathogenic analyses (Figures 3-4, control data in Figure 5) and associated supplementary figures and tables. The individual epilepsy clusters were also used to demonstrate potential for gene discovery (Figure 6) with associated supplementary figures and tables. The individual epilepsy type analyses were used for all other Epi25 analyses. All clusters underwent coverage harmonization (see Coverage Harmonization in Methods).

Coverage Harmonization

As described previously[52], coverage differences between cases and controls introduce a bias because no variants can be called without sufficient coverage. To reduce the influence of coverage differences caused by different capture kits or sequencing depth in general, we used a site-based pruning approach and removed sites where the absolute difference in percentages of cases compared to controls with at least 10x coverage was greater than 7.0%. Each cluster (see Clustering above) underwent independent coverage harmonization. This resulted in four sets coverage maps (Figure S1).

Qualifying Variant

In the context of collapsing analyses, qualifying variants have been defined in order to identify a set of variants that are enriched for real variant calls and variants with strong functional effects.[60] Here we defined a qualifying variant (QV) as a variant passing both QC filters (see Sample and Variant Quality Control) and model-specific filters (Table S4) such as variant effect filters, pathogenicity predictors, and internal and external minor allele frequency (MAF) filters. Variants could be drawn from three pools: i) variants from Epi25 data and matched controls blinded to ClinVar status (Figures 1-2, 6), ii) variants from Epi25 data and matched controls designated Pathogenic/Likely Pathogenic (P/LP) in ClinVar as of 10/20/2020 or iii) all published P/LP ClinVar variants as of 10/20/2020. For analyses of variants in Epi25 data and matched controls blinded to ClinVar status (i) (Figures 1, 2, 3, control data in 5 and 6, Table 1), the following filtering was applied in addition to the variant QC filtering (see Sample and Variant Quality Control [QC]): a) all variants are “ultra-rare” meaning they are not found in any non-neuro gnomAD population, b) all protein truncating variants (PTVs) were filtered with Loss-Of-Function Transcript Effect Estimator (LOFTEE) to remove likely false-positive PTVs,[20] c) all variants located in region with highly repetitive elements were removed to reduce false-positive variants,[61] d) all variants in regions with a proportion expression across transcripts (pext) value less than 1/10 the maximum pext value for that gene were removed as they are unlikely to affect translated mRNA[62] and e) variants were excluded with an internal allele frequency greater than 0.05% applied to the combined case-control call set by cluster excluding one allele to allow for clusters in which one allele might exceed that allele frequency threshold.[62] PTV effects included stop gain, frameshift, splice acceptor, and splice donor variants.

For P/LP variants found in Epi25 and matched controls (Figure 4) (ii) and all published P/LP variants (Figure 5, non-control data) (iii), no universal filtering was applied beyond variant QC. ClinVar variants could additionally be filtered by ClinVar "review status", which attempts to capture the level of review supporting the assertion of clinical significance for the variant with increasing number of "gold stars" from 0 to 4.[63-65]

In addition to the filtering applied above, we defined the following categories of missense variants to be utilized in the study. For "damaging" missense variants, REVEL[46] filter ≥ 0.5 (when defined) was applied. For "intolerant" missense variants, a missense tolerance ratio (MTR) filter ≤ 0.78 (when defined) was applied which represents a variant in the most intolerant quartile of all regions in the exome to missense variation (see Web Resources).[32] To further enhance missense variants for those located in intolerant genic-subregions, we utilized a separate model in which we added an exon-based LIMBR percentile < 25. LIMBR is a sub-genic intolerance score previously shown to enhance selection for missense variants associated with DEEs.[31]

Gene-Based Collapsing

As described previously[7; 52; 56], we performed gene-based collapsing to test whether there is a significant enrichment of cases harboring a QV in a given gene compared to controls. For each gene within each cluster, an indicator variable (1/0 states) was assigned to each individual based on the presence of at least one qualifying variant in the gene (state 1) or no qualifying variants in that gene (state 0) to create a gene-by-subject matrix for each cluster. From the collapsing matrices of the individual clusters, we extracted the number of cases/controls with and without a QV per gene and used the exact two-sided Cochran-Mantel-Haenszel (CMH) test[66; 67] to test for an association between disease status and QV status (Table S4) while controlling for cluster membership. Finally, we created quantile-quantile (QQ) plots (described below). We defined a study-wide Bonferroni multiplicity-adjusted significance threshold of $p < 1.6 \times 10^{-7}$ (0.05 / [18650 CCDS genes × 17 non-synonymous models]).

The synonymous model was used as a putatively negative control (Figure S2, S3A, Tables S4, S6-S8, S16). Additional details for the 17 non-synonymous models can be found in Table S4. The top 200 ranked genes for each analysis can be found in the supplemental tables (Tables S6-26). The membership of each gene in the following gene-sets is also indicated: (D) 43 dominant genes associated with DEE in the Online Mendelian Inheritance in Man (OMIM, see Web Resources) (see Gene-Set Enrichment Testing), (P) 101 dominant genes with epilepsy or related terms in its OMIM phenotype, (L) the 1,920 genes most intolerant to loss-of-function variation in the general population (see Gene-Set Enrichment Testing), top 200 ranked gene in prior Epi25 DEE (D25), GGE (G25) or NAFE (N25) association analyses,[9] or top 300 ranked gene in prior GGE (G4K) or NAFE (N4K) Epi4K association analyses.[7] Epi4K was a large WES epilepsy project completed prior to Epi25.

Quantile-Quantile Plots and Genomic Inflation Factor λ

Quantile-quantile (QQ) plots were generated using empirical (permutation-based) expected probability distributions using a previously described method.[7; 52] For each collapsing model and cluster, the original case and control labels were randomly permuted while the rest of the gene by sample matrix was kept fixed. For each cluster we extracted the number of newly sampled cases/controls with and without a QV per gene and used the Cochran-Mantel-Haenszel (CMH) test to test for an association between case/control status (see Gene-Based Collapsing) and QV status (see Qualifying Variant) while controlling for cluster membership. This process was repeated 1,000 times and for each permutation the p-values were ordered. The mean of each rank-ordered estimate across the 1,000 permutations (i.e., the average 1st order statistic, the average 2nd order statistic, etc.) represent the empirical estimates of the expected ordered p-values. The negative logarithm of the permutation-based expected distribution relative to the observed ordered statistic was plotted to get permutation-based QQ plots. The permutation-based expected p-values were also used to estimate the genomic inflation factor λ based on the regression method as described previously.[7; 52] Genes labeled in black are known epilepsy genes based on manual review of the literature while genes labeled in color are candidate epilepsy genes.

Gene-Set Enrichment Testing

As described previously[7], biologically informed gene-sets can reveal important pathways or gene characteristics by aggregated signal across related genes (Table S5). We utilized the following gene sets (GS-1 to GS-6) informed by their OMIM disease associations, inheritance patterns and genic intolerance.

GS-1) 43 established dominant (e.g. autosomal dominant or x-linked dominant) DEE genes drawn from OMIM Phenotypic Series PS308350 and PS617711 on 10/9/2020

GS-2) 24 genes drawn from the 43 genes in (GS-1) for which in all three epilepsies have a damaging missense variant

GS-3) 101 established dominant genes associated with OMIM phenotypes containing epilepsy and epilepsy related terms on 02/16/2021

GS-4) 14 genes harboring ultra-rare missense variants associated with both DEE and with epilepsy but not DEE in ClinVar (*SZT2*, *SCN2A*, *SCN1A*, *HCN1, GABRA1*, *GABRG2*, *KCNQ3*, *SPTAN1*, *KCNT1*, *GRIN2B*, *GABRB3*, *CHD2*, *TBC1D24*, *KCNQ2)* as of 10/20/2020.

GS-5) 10 gene sets representing the genes without a confirmed disease phenotype in OMIM on 02/16/2021 (18,852 CCDS genes – 3,964 genes = 14,888 genes) distributed into

10 groups by their loss-of-function observed/expected upper bound fraction (LOEUF) decile were created.[20] LOEUF is the 90% upper bound of the confidence interval of the observed/expected ratio of predicted loss-of-function variants in gnomAD and can be used to bin genes into deciles of approximately 1,920 genes each.

GS-6) 10 gene sets representing the genes without a confirmed phenotype in OMIM on 02/16/2021 (18,852 CCDS genes – 3,964 genes = 14,888 genes) distributed into 10 groups by their missense Z score were created.[19; 20; 68] Missense Z score captures the number of observed missense variants in a gene compared to the expected number of missense variants in the general population. The score was used to bin genes into deciles of approximately 1,920 genes each.

For a gene-set analysis, we extracted the number of cases/controls with and without at least one QV among any of the genes in the gene-set and used the exact two-sided CMH test[66; 67] to test for an association between disease status and QV status while controlling for cluster membership. To examine association with LOEUF deciles (Figure 6), only controls without a disease association in our database (“Controls” and “Healthy Family Members”) were used (Table S2). We used a false discovery rate (FDR) correction for multiple comparisons. We performed 123 CMH tests to determine odds ratios for gene-set enrichment testing, and defined a significant enrichment at FDR < 0.05. For forest plots, odds ratios and *p*-values were displayed for associations with an unadjusted *p*-value < 0.05.

Sub-Genic Intolerance Comparison

We examined sub-genic intolerance scores (MTR) in multiple ways. We compared the raw MTR and MTR domain percentiles scores across epilepsies and controls directly using the Kruskal-Wallis test by rank. For groups with *p*-value < 0.05, we performed pair-wise comparisons using the Wilcoxon signed-rank test. This method may not be an adequate comparison because, despite enriching for damaging missense variants with REVEL, controls with qualifying variants remain (which are unlikely to be true positives) indicating that some of the qualifying variants found in cases may also be benign. Direct comparison of sub-genic intolerance scores among epilepsies is therefore difficult to interpret because the QV burden is different among epilepsies (see Results) and the true positive rate among these QVs is unknown.

To compare MTR among epilepsies, it was necessary to estimate and compare the "true positive" distribution of scores for each epilepsy. To achieve this, we created a weighted average of the cumulative distribution function (CDF) of MTR scores for ultra-rare damaging missense variants in each epilepsy (CDF_{DEE}, CDF_{GGE}, and CDF_{NAFE}) and the CDF of ultra-rare damaging missense variants in our controls (CDF_{CTRL}) to obtain the "true positive" CDF for each epilepsy (CDF_{DEE_TP}, CDF_{GGE_TP}, and CDF_{NAFE_TP}). Only damaging missense variants with defined MTR scores were considered.

At a given MTR value, the "true positive" CDF is a weighted average of the epilepsy and control CDF with the weights determined by the QV rate of the control population at that MTR value. For example, if, at a MTR score of 0.5, 4% of DEE cases have an ultra-rare damaging missense variant and 1% of control cases of have an ultra-rare damaging missense variant, then $CDF_{DEE\_TP}(0.5) = 0.75 \times CDF_{DEE}(0.5) + 0.25 \times CDF_{CTRL}(0.5)$.[69] We then used a Kolmogorov–Smirnov test (statistic D) to compare the distribution of "true positive" MTR CDFs of each epilepsy pair. Given that we did not know the distribution of D, for each comparison, we performed a permutation test with 10,000 permutations. We assessed significance at $p < 0.05$.

To compare sub-genic intolerance scores by gene, we compared the "true positive" mean MTR by gene for DEE compared to NAFE and compared to GGE. In a given gene, the "true positive" mean MTR is a weighted average of the epilepsy mean MTR and control mean MTR scores with the weights determined by the QV rate of the control population in that gene. For example, if, in gene X, 4% of DEE cases have an ultra-rare damaging missense variant and 1% of control cases of have an ultra-rare damaging missense variant, then $\text{Mean}_{\text{DEE\_TP}}(X) = 0.75 \times \text{Mean}_{\text{DEE}}(X) + 0.25 \times \text{Mean}_{\text{CTRL}}(X)$. For those genes with no control variants, the means were calculated without weighting. We measured the number of genes where DEE had a lower weighted mean MTR and measured significance with a binomial test with the null hypothesis that DEE variants had a lower Mean_{TP} in half of the genes in the tested gene set.

To compare the MTR values of published ClinVar variants (i.e. not drawn from our cases or controls), we divided the variants into those associated with DEE and non-DEE epilepsy. ClinVar variants with phenotypes containing “epilepsy” or “epileptic” were considered associated with epilepsy. Those with phenotypes containing “West”, “Dravet”, “Lennox-Gastaut“, ”infantile spasm”, “Ohtahara”, “myoclonic”, or “glut 1” were considered associated with DEE while the remainder were classified as non-DEE epilepsy. There were inadequate number of variants specifically associated with GEE and NAFE to further sub-divide them. For variants with multiple clinical associations, the most severe association was assigned. We looked at only ultra-rare variants with a defined MTR value. We limited our analysis to only those genes harboring variants in both epilepsy groups (see Gene-Set Enrichment Testing). The control variant set was drawn from the combined epilepsy analysis (Figure S1A-C). We used a two-sample Wilcoxon test to assess significance. We measured the number of genes where DEE had a lower mean MTR and measured significance with a binomial test with the null hypothesis that DEE variants had a lower mean MTR in half of the genes in the tested gene set.

Lollipop and MTR Plots

Lollipop mutation diagrams were generated for the 24 genes analyzed for the sub-genic intolerance comparison (GS-2) using lollipops-v.1.5.3.[70] All 614 missense variants (DEE = 100, GGE = 133, and NAFE = 153, Control = 228) were displayed across the linear gene structure of the associated gene. For each gene, the MTR distribution with missense variant locations plotted was juxtaposed against the lollipop mutation diagram. MTR data were downloaded from the MTR-Viewer website (see Web Resources).[71]

Comparison of Evolutionary Constrained Regions

Evolutionary constraint for missense variants was assessed at three levels. For base-level scores, we used the GERP++ "rejected substitution" (RS) score in which higher scores correspond to greater constraint.[72; 73] For exonic and domain constraint, we used exonic and domain subGERP scores, respectively.[22] We compared scores across epilepsies and controls directly using the Kruskal-Wallis test by rank. No group reached statistical significance (*p*-value < 0.05) so no pair-wise comparisons were performed.

Candidate Non-OMIM Epilepsy Genes

To ascertain additional potential epilepsy-gene associations not found in OMIM, we highlighted genes which are (1) in the most intolerant decile to LOF variation in the general population by LOEUF rank, (2) not associated with a disease in OMIM, (3) harbor PTVs with LOFTEE filtering in more than one case, and (4) harbor no control PTVs with LOFTEE filtering.

Data Analysis and Display

Unless otherwise noted in the methods, data analysis and visualization were performed using R (version 3.6.0).[74] Notches in boxplots indicate 1.58 * interquartile range / sqrt(n), which approximates the 95% confidence interval.[75]

Results

Gene-Based Collapsing in Three Types of Epilepsies

The results of the gene-based collapsing should be viewed through the lens of prior rare-variant association analyses of epilepsy data, and specifically, Epi25 data. The data in this analysis are a superset of the data used in prior Epi25 analyses.[9] The cluster-based collapsing analysis allows for the inclusion of multiple ancestries as each geographic ancestry matched cluster is analyzed separately (Figure S1). The results are then combined using the Cochran-Mantel-Haenszel test (see Statistical Analyses in Methods) accounting for population sub-structure.[56] The sample size increased in all three epilepsies (1,835 from 1,021 DEE cases, 5,303 from 3,108 GEE cases, 6,349 from 3,597 NAFE cases) due to increased enrollment in Epi25 and the inclusion of cases with non-European geographic ancestry. Other differences include a different control set and different *in silico* methods of indicating qualifying variant (QV) status. We ran gene-based collapsing (Tables S6-S26) for gene-discovery counting PTVs and damaging missense variants for all three epilepsies (Figure 1, Tables S9, S12, S14) and all epilepsies combined (Figure S3B, Table S17). There was expected overlap among the top ranked genes from prior Epi25 analyses as well as the suggestion of candidate genes not previously associated with epilepsy (Tables S11 – S26).

In the DEE collapsing analysis (Figure 1A, Table S9), the top two ranked genes were the same as in the prior Epi25 analysis, but now *SCN1A* ([MIM: 182389] OR = 7.1, $p = 4.4 \times 10^{-8}$) and *NEXMIF* (previously known as *KIAA2022* [MIM: 300524] OR 26.5, $p = 8.6 \times 10^{-8}$) both achieve study-wide significance. In contrast to prior Epi25 analyses, nine of the top ten ranked genes are known epilepsy genes[76-87] demonstrating the strength of the increased sample size and clustering methodology. The remaining gene, *AP3S2* ([MIM: 602416] OR = 70.5, $p = 2.7 \times 10^{-4}$), is a component of the AP3 complex, an adaptor-related complex with no prior association to epilepsy although it was a top 200 hit in the prior Epi25 DEE analysis.[9; 88] Hermansky-Pudlak syndrome 10

(MIM: 617050), which is notable for infantile onset of immunodeficiency and intractable seizures, is caused by biallelic mutations in *AP3D1* (MIM: 607246), a different component of the same AP3 complex.[89] To highlight candidate genes, we removed DEE cases in Figure 1A that harbored a qualifying variant in any of the 101 dominant genes with epilepsy or related terms in the OMIM phenotype and re-ran the collapsing analysis (Figure S4, Table S11). The 5th ranked gene, *SRCAP* ([MIM: 611421] OR = 6.8, $p = 1.6 \times 10^{-3}$), is highly intolerant to loss-of-function variants (LOUEF = 0.1) and is associated with Floating-Harbor syndrome (MIM: 136140) which can include seizures.[90; 91] In summary, this enlarged DEE analyses with cases of non-European geographic ancestry produced results that more consistently elevated known epilepsy genes and importantly, proposed genes without prior epilepsy associations (*AP3S2, SRCAP*).

Four of the top ten ranked genes in the gene-based collapsing analysis for GGE (Figure 1B, Table S12) were previously associated with epilepsy (*SLC6A1* [MIM: 137165], *SCN1A, GRIN2A* [MIM: 138253], *GABRA1* [137160]).[92-95] The top hit is *SLC6A1* (OR = 16.6, $p = 2.1 \times 10^{-6}$) which was a top 200 gene in the prior Epi25 GGE analysis but now approaches study-wide significance.[9] *SCL6A1* was initially implicated in DEE, but its role in generalized epilepsies has only been more recently revealed.[95; 96] Among the remaining genes, there are two promising candidates. (1) *FBXO42* ([MIM: 609109] OR = 13.6, $p = 4.5 \times 10^{-4}$) is a highly intolerant gene (LOEUF = 0.27) important in the regulation of p53 and not yet implicated in disease but was a top 200 GGE associated gene in the prior Epi25 analysis[9], and (2) *KCNK18* ([MIM: 613655] OR = Inf, $p = 1.6 \times 10^{-3}$) is a potassium channel implicated in migraine pathology.[97; 98] Promising candidate genes for GGE from the prior Epi25 analysis (*CACNA1G* [MIM: 604065] and *UNC79* [MIM: 616884]) were not among the top 200 associated genes, which may be related to the different method of filtering missense variants.[9] Further limiting missense variants to intolerant as well as damaging (Figure S5B, Table S13) elevated *CACNA1B* ([MIM: 601012], OR = 5.5, $p = 3.5 \times 10^{-4}$). Bi-allelic LOF variants in *CACNA1B* cause severe epilepsy.[99] *CACNA1B* was the top gene associated with GGE in Epi4K[7], a large WES epilepsy project prior to Epi25. No association was found in the prior Epi25 analysis and there is limited other literature linking *CACNA1B* to GGE. This new GGE Epi25 collapsing analysis did not confirm promising candidate genes from the prior Epi25 analysis but did provide additional support for the association between *CACNA1B* and GGE and proposed candidate genes (*FBXO42* and *KCNK18*).

Gene-based collapsing analysis for NAFE (Figure 1C, Table S14) showed a familiar top hit, *DEPDC5* ([MIM: 614191] OR = 5.4, $p = 1.3 \times 10^{-6}$) and four additional genes (*GRIN2A, SCN1A, SCN8A* [600702]*,* and *NPRL2* [607072]), which have previously been implicated in NAFE.[7; 9; 80; 84; 92; 100; 101] Renin, the protein encoded by *REN* ([MIM: 179820] OR = 12.7, $p = 4.2 \times 10^{-4}$), is produced by juxtaglomerular cells of the kidney but has been implicated as a target of adjuvant therapy for epilepsy.[102; 103] *ADORA2B* ([MIM: 600446] OR = Inf, $p = 4.5 \times 10^{-4}$), is a small gene encoding an adenosine receptor not associated with disease but being explored for its role in epileptogenesis.[104; 105] *DAW1* ([MIM: N/A] OR = 30.0, $p = 1.8 \times 10^{-4}$), a little understood gene, supports cilia function.[106] The increased sample size did not further support promising genes from the prior Epi25 analysis such as *TRIM3* (MIM: 605493), *PPFIA3* (MIM: 603144), and *KCNJ3* (MIM: 601534).[9] Further limiting missense variants to intolerant as well as damaging (Figure S5C, Table S15) removed all control enriched genes from the top ten ranked genes and elevated known epilepsy genes. Interestingly, the 7th ranked gene, *TSC1* ([MIM: 605284], OR = 14, $p = 1.7 \times 10^{-3}$), is typically associated with focal epilepsy in the context of tuberous sclerosis-1 (MIM: 191100) or focal cortical dysplasia, type II, somatic (MIM: 607341) although the individuals with focal epilepsy in this study do not have a lesion on MRI.[107; 108] Like the GGE collapsing analysis, the NAFE collapsing analysis proposed different candidate genes rather than confirming those from prior Epi25 analyses.

Milder Epilepsies Remain Enriched for Ultra-Rare Variants in a Limited Gene-Set

Our group has previously observed that more mild epilepsies are enriched in genes also associated with severe phenotypes.[7; 9] To limit the degree to which individual genes in the gene-set drove that finding and facilitate comparisons of variants across epilepsies, we recapitulated that analysis but narrowed the gene set of dominant DEE-associated genes to include only those 24 genes containing at least one damaging missense variant in all three epilepsies (Figure 2, Tables S5, S27). DEE (CMH pooled odds ratio [OR] = 2.1, FDR-adjusted p value [*adj.p*] = 1.9 × 10^{-9}) and NAFE (CMH pooled odds ratio [OR] = 1.3, FDR-adjusted p value [*adj.p*] = 1.2 × 10^{-3}) are enriched for all missense variants. All three epilepsies are enriched for damaging missense variants (DEE OR = 3.7, *adj.p* = 6.8 × 10^{-17}, GGE OR = 1.7, *adj.p* = 1.2 × 10^{-4}, NAFE OR = 1.7, *adj.p* = 6.4 × 10^{-5}), and removing the damaging filter, all three epilepsies are also enriched for variants in intolerant genic sub-regions (DEE OR = 3.5, *adj.p* = 1.6 × 10^{-14}, GGE OR = 1.7, *adj.p* = 1.2 × 10^{-4}, NAFE OR = 1.6, *adj.p* = 3.5 × 10^{-4}). Combining both improves enrichment in all three epilepsies (DEE OR = 5.5, *adj.p* = 8.1 × 10^{-19}, GGE OR = 2.2, *adj.p* = 1.0 × 10^{-6}, NAFE OR = 2.0, *adj.p* = 1.8 × 10^{-5}). Only DEE and GGE were enriched for loss-of-function variants (DEE OR = 12.7, *adj.p* = 1.9 × 10^{-9}, GGE OR = 3.8, *adj.p* = 4.6 × 10^{-4}), which is consistent with prior analyses.[9] In summary, despite restricting our DEE-associated gene-set further to ensure that at least one case per epilepsy harbored a damaging missense variant in each gene and enlarging our samples to include individuals of non-European ancestry, a familiar pattern of enrichment exists in the milder epilepsies.

Ultra-Rare DEE Variants in Epi25 are Located in Intolerant Genic Sub-Regions

After demonstrating that more mild epilepsies (GGE, NAFE) were enriched for ultra-rare damaging missense variants in the same gene set as severe epilepsies (DEE) (Figure 2), we tested the hypothesis that variants associated with DEE were located in more intolerant sub-regions than those associated with GGE or NAFE. Despite filtering for pathogenicity with REVEL, there remains a background rate of enrichment of ultra-rare damaging missense variants in the control population (Figure 2, Table S29). This suggests that a portion of the ultra-rare damaging missense variants in our epilepsy cases are also benign, which makes direct comparison of the sub-genic intolerance score among epilepsy subtypes (Figure S6A) difficult to interpret as the burden of damaging missense variants in DEE cases is higher than those of GGE or NLFE (CMH, DEE-GGE OR = 2.2, *adj.p* = 7.8 × 10^{-7}, DEE-NAFE OR = 2.3, *adj.p* = 9.4 × 10^{-8}, Table S28). Instead, we estimated the distribution of MTR scores of "true positive" ultra-rare damaging missense variants in each epilepsy and made pair-wise comparisons using a Kolmogorov–Smirnov (K-S) test (see Sub-Genic Intolerance Comparison in Methods, Figure 3). Consistent with our hypothesis, the distribution of MTR scores for DEE variants was significantly different from NAFE ("true positive" median MTR DEE = 0.670 vs. NAFE = 0.721, K-S, $p < 0.0156$) while the difference from GGE did not achieve statistical significance ("true positive" median MTR DEE = 0.670 vs. GGE = 0.710, K-S, $p = 0.38$). On a per gene basis, the MTR scores of DEE variants are not uniformly more intolerant than GGE and NAFE (Figure S7). Though the above analysis demonstrates that DEE variants lay in more intolerant genic sub-regions than NAFE variants, it does not account for the possible differential contribution of specific genes to specific epilepsies among the 24 genes. To address this concern, we performed a second analysis which compared the weighted mean MTR of DEE compared to NAFE and to GGE (Table S29). The weighted mean MTR scores of the DEE variants was lower in 15 of the 24 genes compared to NAFE (binomial test, $p = 0.31$) and 15 of the 24 genes compared to GGE (binomial test, $p = 0.31$).

No clear relationship exists between gene, protein domain, and epilepsy type (Figure S8). Despite the large Epi25 dataset, we likely remain underpowered to untangle the epilepsy by protein space relationship on an individual gene level.[33] MTR is calculated on a sliding window making it independent of known gene structures. Domain-based MTR showed a smaller difference among the epilepsies (Figure S6A-B) suggesting that the sub-genic intolerance differences among the epilepsies is at least partially independent from gene structures.[32] We also examined whether missense variants associated with DEE were located in more evolutionary constrained bases, exons or domains than milder epilepsies (Figure S6C-E). No comparison met statistical significance. This was true despite both evolutionary constrained and intolerant domains harboring pathogenic variants although differences in domains may be difficult to assess given the limited number per gene.[22]

Only Ultra-Rare Pathogenic/Likely Pathogenic ClinVar Variants are Enriched in Epi25

The sample size of Epi25 allows us to assess the representation of variants found in ClinVar, a heavily used clinical database of curated variants, in our three epilepsy sub-groups and investigate whether sub-genic intolerance might add clinically useful information.[38; 39] Using a set of 101 genes with epilepsy or related terms in their OMIM phenotypes (Table S5), we examined the burden of P/LP variants in our cases compared to controls (Figure 4A, Table S30). Given the prior findings that epilepsy cases are enriched with ultra-rare variants but not more common variants[7], we divided our ClinVar analysis into variants not found in the non-neuro gnomAD populations (ultra-rare) and variants seen in the general population (public). Consistent with prior reports, there was an increased burden of ultra-rare P/LP variants in our epilepsy cases compared to controls irrespective of epilepsy type (CMH, DEE OR = 84.5, *adj.p* = 8.9×10^{-38}, GGE OR = 14.5, *adj.p* = 1.8×10^{-11}, NAFE OR = 14.4, *adj.p* = 6.9×10^{-13}). There was no enrichment in public variants (Figure 4A). Epilepsy variants in ClinVar also found in gnomAD or future public datasets may require additional investigation to confirm pathogenicity.

Severe Pathogenic/Likely Pathogenic ClinVar Variants are Located in Intolerant Genic Sub-Regions

Among ultra-rare ClinVar variants, we sought to determine if we could further differentiate epilepsy variants from control variants (Figure 4B, Table S31). ClinVar "review status" attempts to capture the level of review supporting the assertion of clinical significance for the variant with increasing number of "gold stars" from zero to four.[63-65] Filtering ultra-rare P/LP ClinVar based on review status did not improve discrimination in a dose dependent fashion. In all three epilepsies, there were no zero star controls but the enrichment of variants with more than one star exceeded the enrichment of variants with one star (CMH, DEE OR = 47.5, $adj.p = 7.8 \times 10^{-12}$ → OR = 91.6, $adj.p = 1.4 \times 10^{-21}$, GGE OR = 9.1, $adj.p = 5.6 \times 10^{-4}$ → OR = 17.2, $adj.p = 6.6 \times 10^{-7}$, NAFE OR = 8.2, $adj.p = 2.0 \times 10^{-3}$ → OR = 10.7, $adj.p = 1.3 \times 10^{-4}$). We next examined whether sub-genic intolerance filtering could further improve discrimination of cases compared to controls. After filtering with MTR, the OR of ultra-rare missense variants increased in all three epilepsies (CMH, DEE OR = 92.5, $adj.p = 3.4 \times 10^{-32}$ → OR = 335.4, $adj.p = 1.4 \times 10^{-25}$, GGE OR = 14.9, $adj.p = 1.2 \times 10^{-9}$ → OR = 59.6, $adj.p = 3.9 \times 10^{-10}$, NAFE OR = 12.3, $adj.p = 3.8 \times 10^{-9}$ → OR = 34.7, $adj.p = 9.2 \times 10^{-8}$). All three epilepsies were enriched with ultra-rare PTVs in ClinVar (DEE OR = 49.8, $adj.p = 3.4 \times 10^{-6}$, GGE OR = 11.0, $adj.p = 0.045$, NAFE OR = 24.7, $adj.p = 1.6 \times 10^{-4}$). Among the few public variants, only missense variants filtered with MTR were statistically enriched in NAFE cases, and overall, MTR filtering removed all 12 control missense variants but only four of ten epilepsy variants (Table S32). In summary, sub-genic intolerance filtering improved discrimination of both ultra-rare and public variants in ClinVar, suggesting sub-genic intolerance provides additive information to identify potential false-positive or variable penetrance variants in ClinVar.

Using ultra-rare P/LP ClinVar variants, we sought to confirm our Epi25 finding (Figure 3) that missense variants in severe epilepsies are located in more intolerant genic sub-regions than milder epilepsies. We compared median sub-genic intolerance scores between DEE and non-DEE epilepsies (see Sub-Genic Intolerance Comparison in Methods) in genes with missense variants in both epilepsy groups (Figure 5, Tables S5, S33). The median MTR score was lower (more intolerant) for published ClinVar DEE variants compared to non-DEE epilepsy ClinVar variants (median DEE MTR = 0.57 vs. median non-DEE MTR = 0.70, Wilcoxon signed-rank test, $p < 6.7 \times 10^{-3}$). When examined by gene, the mean MTR score for the DEE variants was lower than the non-DEE variants in 11 of 14 genes tested, (binomial test, $p = 0.057$). Reassuringly, both DEE and non-DEE variants existed in more intolerant regions than ultra-rare control variants (median control MTR 0.83, same control set as Figure 2).

Epilepsy Genes Remain to be Discovered and are Likely Loss-Of-Function Intolerant

There are ~3,900 genes identified in OMIM as being causative or a risk factor for disease.[109] Analyzing likely damaging variants in non-OMIM genes may give a sense of as-yet to be discovered epilepsy genes (Figure 6, Tables S34-S35). GGE and NAFE revealed a significant burden of PTVs in the intersection of non-OMIM genes with the decile of genes most intolerant to loss-of-function variation in the general population (GGE OR = 1.3, *adj.p* = 2.7×10^{-4}, NAFE OR = 1.2, *adj.p* = 0.013) (Figures 6B-6C). We highlighted the top four genes in the most intolerant decile associated with GGE and NAFE that had more than one case PTV and no control PTVs (Table 1). The most significant GGE candidate gene, *NLGN2* (MIM: 606479, 3 cases), encodes neuroligin 2, which is a trans-synaptic adhesion molecules important in the synapse.[110] The most significant NAFE candidate gene was *WDR18* (MIM:N/A, 4 cases) whose protein product forms the PELP1-TEX10-WDR18 complex important in ribosomal maturation.[111] A table of potential DEE, GGE and NAFE genes are included in the supplement (Tables S36-S38). Finally, to investigate additional candidate genes, we performed rare variant collapsing analysis with only PTVs (Figure S9, Table S18-S20), only damaging missense variants (Figure S10, Tables S21-23) and PTVs combined with damaging and intolerant missense variants further limited to intolerant LIMBR exons (see Qualifying Variant in Methods, Figure S11, Tables S24-S26).[31]

DEE cases also revealed a trend towards increased burden in the intersection of non-OMIM genes with the 7th most intolerant decile (DEE OR = 1.1, *adj.p* = 0.14, Figure 6A), which may reflect genes associated with recessive epilepsies.[20] None of the epilepsies revealed a significant burden of damaging and intolerant missense variants in missense intolerant genes (Table S35).

Discussion

In this, the largest Epi25 exome study of epilepsies to date including individuals of non-European geographic descent, we reaffirm that ultra-rare variants contribute to the 3 major epilepsy groups (Figure 1). Our collapsing analyses proposed epilepsy-genes (*AP3S2, SRCAP*, *FBXO42, KCNK18, REN*, and *ADORA2B*) requiring future confirmation. These associations reveal the power of increasing sample size with Epi25 and our clustering technique allowing the inclusion of non-European populations. The *p*-values in DEE analyses must be regarded in light of the smaller sample size of individuals with DEE (1,835 DEE compared to 5,303 GGE and 6,379 NAFE). We were unable to confirm several promising candidate genes from the prior Epi25 analysis which may be secondary to different control groups, different *in silico* filters and a larger sample size.[9] We confirmed enrichment of ultra-rare variants in GGE and NAFE in genes associated with DEE even when limited to genes in which all epilepsies have a damaging missense variant to limit single and distinct genes driving associations with different epilepsies (Figure 2).

Sub-genic intolerance has broad implications. It has been shown to help improve discrimination between pathogenic and benign variants and confirm the pathogenicity of new variants.[22; 32; 112-115] Pathogenic variants may cluster in areas of regional intolerance,[31; 32; 116] and sub-genic intolerance scores may inform biochemical exploration, yielding novel insights into protein function.[117] To our knowledge, this is the broadest demonstration that sub-genic intolerance scores might not only be different between case compared to control but also affect disease severity (Figure 3 and 5).[32] This discrepancy may broadly inform the functional similarities of mutations leading to more severe disease across genes or interestingly, across gene families.[118]

Using the large Epi25 dataset allowed us to assess variants documented in ClinVar (Figure 4). Allele frequency is known to be inversely associated with pathogenicity, and, among Epi25 cases, only ultra-rare variants were enriched in cases compared to controls (Figure 4A). Previous analyses have used population based MAFs to reclassify variants as benign.[32; 68; 119; 120] The evolving nature of ClinVar classifications has been noted previously as more population-wide control data become available.[63; 64; 121] Within the ultra-rare MAF bin, review status did not provide additional enrichment in a dose-dependent manner in our data (Figure 4B) although it has indicated higher true positive value in other studies focused on more common variants.[63-65; 122] One and two star ultra-rare pathogenic variants in ClinVar have been reported as possible false-positives,[122] although no study to our knowledge has systematically evaluated ultra-rare P/LP ClinVar variants for false-positivity or incomplete penetrance. Finally, four of the five ultra-rare and all 12 public missense P/LP variants harbored by controls were located in more tolerant regions of the exome (Figure 4B, Tables S31-S32). The enrichment of ClinVar variants with MTR filtering suggests that regional intolerance may provide additional information to clinicians assessing ClinVar variants.

There likely remain genes that will ultimately be associated with a disease although the pace of discovery may be slowing.[109] In this Epi25 cohort, GGE and NAFE contained an increased burden of PTVs in the non-OMIM genes most intolerant to loss-of-function variation in the general population (Figure 6). No increase was seen for individuals with DEE, suggesting that gene discovery for DEE is advanced compared to the milder epilepsies. There are several genes with PTVs in multiple cases but in no controls that are potential epilepsy or epilepsy-risk genes (Tables 1, S37-S39). With increased sample size, these genes may become more prominent in future collapsing analyses.

Limitations of this study are that individuals with epilepsy were enrolled at variable ages, leaving open the possibility that a case may evolve from one epilepsy to another. While we posit that variant location determines the severity of the variant and therefore determines the phenotype, this does not address variants which have one autosomal dominant phenotype and a different autosomal recessive phenotype. The sub-genic intolerance score by gene interaction (Figure S6-S7) may be secondary to different numbers of variants per gene, MTR not completely capturing all sub-genic intolerance information, or other factors contribute to epilepsy severity. Examining the collective sub-genic intolerance scores of variants from multiple genes does not take into account within gene comparisons (i.e. sub-genic intolerance distributions differ per gene as do the epilepsy type-by-gene burdens). We attempted to address these confounds (Tables S29, S33) but were under-powered. Future studies will be needed to understand the gene-by-intolerance score interaction. Finally, segregation analysis of variants in candidate epilepsy genes (Table 1) could weaken or bolster the proposed relationships. Unfortunately, we do not have access to Epi25 family member data. As the Epi25 enrollment increases, we look forward to the increased power allowing for the further elucidation of the genetic architectures of the epilepsies.

Consortia

Joshua E. Motelow, Gundula Povysil, Ryan S. Dhindsa, Kate E. Stanley, Andrew S. Allen, Yen-Chen Anne Feng, Daniel P. Howrigan, Liam E. Abbott, Katherine Tashman, Felecia Cerrato, Caroline Cusick, Tarjinder Singh, Henrike Heyne, Andrea E. Byrnes, Claire Churchhouse, Nick Watts, Matthew Solomonson, Dennis Lal, Namrata Gupta, Benjamin M. Neale, Gianpiero L. Cavalleri, Patrick Cossette, Chris Cotsapas, Peter De Jonghe, Tracy Dixon-Salazar, Renzo Guerrini, Hakon Hakonarson, Erin L. Heinzen, Ingo Helbig, Patrick Kwan, Anthony G. Marson, Slavé Petrovski, Sitharthan Kamalakaran, Sanjay M. Sisodiya, Randy Stewart, Sarah Weckhuysen, Chantal Depondt, Dennis J. Dlugos, Ingrid E. Scheffer, Pasquale Striano, Catharine Freyer, Roland Krause, Patrick May, Kevin McKenna, Brigid M. Regan, Caitlin A. Bennett, Costin Leu, Stephanie L. Leech, Terence J. O'Brien, Marian Todaro, Hannah Stamberger, Danielle M. Andrade, Quratulain Zulfiqar Ali, Tara R. Sadoway, Heinz Krestel, André Schaller, Savvas S. Papacostas, Ioanna

Kousiappa, George A. Tanteles, Christou Yiolanda, Katalin Štěrbová, Markéta Vlčková, Lucie Sedláčková, Petra Laššuthová, Karl Martin Klein, Felix Rosenow, Philipp S. Reif, Susanne Knake, Bernd A. Neubauer, Friedrich Zimprich, Martha Feucht, Eva Reinthaler, Wolfram S. Kunz, Gábor Zsurka, Rainer Surges, Tobias H. Baumgartner, Randi von Wrede, Manuela Pendziwiat, Hiltrud Muhle, Annika Rademacher, Andreas van Baalen, Sarah von Spiczak, Ulrich Stephani, Zaid Afawi, Amos D. Korczyn, Moien Kanaan, Christina Canavati, Gerhard Kurlemann, Karen Müller-Schlüter, Gerhard Kluger, Martin Häusler, Ilan Blatt, Johannes R. Lemke, Ilona Krey, Yvonne G. Weber, Stefan Wolking, Felicitas Becker, Stephan Lauxmann, Christian Bosselmann, Josua Kegele, Christian Hengsbach, Sarah Rau, Bernhard J. Steinhoff, Andreas Schulze-Bonhage, Ingo Borggräfe, Christoph J. Schankin, Susanne Schubert-Bast, Herbert Schreiber, Thomas Mayer, Rudolf Korinthenberg, Knut Brockmann, Markus Wolff, Dieter Dennig, Rene Madeleyn, Reetta Kälviäinen, Anni Saarela, Oskari Timonen, Tarja Linnankivi, Anna-Elina Lehesjoki, Sylvain Rheims, Gaetan Lesca, Philippe Ryvlin, Louis Maillard, Luc Valton, Philippe Derambure, Fabrice Bartolomei, Edouard Hirsch, Véronique Michel, Francine Chassoux, Mark I. Rees, Seo-Kyung Chung, William O. Pickrell, Robert H.W. Powell, Mark D. Baker, Beata Fonferko-Shadrach, Charlotte Lawthom, Joe Anderson, Natascha Schneider, Simona Balestrini, Sara Zagaglia, Vera Braatz, Michael R. Johnson, Pauls Auce, Graeme J. Sills, Larry W. Baum, Pak C. Sham, Stacey S. Cherny, Colin H.T. Lui, Norman Delanty, Colin P. Doherty, Arif Shukralla, Hany El-Naggar, Peter Widdess-Walsh, Nina Barišić, Laura Canafoglia, Silvana Franceschetti, Barbara Castellotti, Tiziana Granata, Francesca Ragona, Federico Zara, Michele Iacomino, Antonella Riva, Francesca Madia, Maria Stella Vari, Vincenzo Salpietro, Marcello Scala, Maria Margherita Mancardi, Nobili Lino, Elisa Amadori, Thea Giacomini, Francesca Bisulli, Tommaso Pippucci, Laura Licchetta, Raffaella Minardi, Paolo Tinuper, Lorenzo Muccioli, Barbara Mostacci, Antonio Gambardella, Angelo Labate, Grazia Annesi, Lorella Manna, Monica Gagliardi, Elena Parrini, Davide Mei, Annalisa Vetro, Claudia Bianchini, Martino Montomoli, Viola Doccini, Carmen Barba, Shinichi Hirose, Atsushi Ishii, Toshimitsu Suzuki, Yushi Inoue, Kazuhiro Yamakawa, Ahmad Beydoun, Wassim Nasreddine, Nathalie Khoueiry-Zgheib, Birute Tumiene, Algirdas Utkus, Lynette G. Sadleir, Chontelle King, S. Hande Caglayan, Mutluay Arslan, Zuhal Yapıcı, Pınar Topaloglu, Bulent Kara, Uluc Yis, Dilsad Turkdogan, Aslı Gundogdu-Eken, Nerses Bebek, Meng-Han Tsai, Chen-Jui Ho, Chih-Hsiang Lin, Kuang-Lin Lin, I-Jun Chou, Annapurna Poduri, Beth R. Shiedley, Catherine Shain, Jeffrey L. Noebels, Alicia Goldman, Robyn M.

Busch, Lara Jehi, Imad M. Najm, Lisa Ferguson, Jean Khoury, Tracy A. Glauser, Peggy O. Clark, Russell J. Buono, Thomas N. Ferraro, Michael R. Sperling, Warren Lo, Michael Privitera, Jacqueline A. French, Steven Schachter, Ruben I. Kuzniecky, Orrin Devinsky, Manu Hegde, David A. Greenberg, Colin A. Ellis, Ethan Goldberg, Katherine L. Helbig, Mahgenn Cosico, Priya Vaidiswaran, Eryn Fitch, Samuel F. Berkovic, Holger Lerche, Daniel H. Lowenstein, David B. Goldstein

Supplemental Information description

Supplementary data will include (1) 11 supplemental figures, (2) 38 supplemental tables, (3) Supplemental Subjects and Methods which includes full author list, affiliations, details of individual participating Epi25 cohorts and supplemental acknowledgments, and (4) Supplemental References. These are divided between one Microsoft Word file and one Microsoft Excel file

- Word file (11 supplementary figures and 9 supplementary tables)
    - 11 supplemental figures (Figures S1-S11)
    - 10 supplemental tables (Tables S1-S4, S29, S33, S36-S38)
    - Supplemental Subjects and Methods
    - Supplemental References
- Excel file (29 supplemental tables)
    - Tables S5-S28, S30-S32, S34-S35

Acknowledgments

We thank the Epi25 principal investigators, local staff overseeing individual cohorts, and all of the individuals with epilepsy and their families who participated in Epi25 for their commitment to this international collaboration. J.E.M. is supported by the National Institutes of Health (TL1TR001875). This work is part of the Centers for Common Disease Genomics (CCDG) program, funded by the National Human Genome Research Institute (NHGRI) and the National Heart, Lung, and Blood Institute (NHLBI). CCDG-funded Epi25 research activities at the Broad Institute, including genomic data generation in the Broad Genomics Platform, are supported by NHGRI grant UM1 HG008895

(PIs: Eric Lander, Stacey Gabriel, Mark Daly, Sekar Kathiresan). The Genome Sequencing Program efforts were also supported by NHGRI grant 5U01HG009088. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. A supplemental grant for Epi25 phenotyping was supported by “Epi25 Clinical Phenotyping R03,” National Institutes of Health (1R03NS108145-01), with D.H.L. and S.F.B. as the principal investigators. We also thank the Stanley Center for Psychiatric Research at the Broad Institute for supporting the genomic data generation. Additional funding sources and acknowledgment of individual case cohorts were listed in the Supplemental Data.

Declaration of Interests

B.M.N is a member of the scientific advisory board at Deep Genomics and RBNC Therapeutics, member of the scientific advisory committee at Milken and a consultant for Camp4 Therapeutics, Takeda Pharmaceutical and Biogen. R.S.D. is a consultant for AstraZeneca. D.B.G. is a founder and shareholder in Praxis Therapeutics, and shareholder in and member of the scientific advisor board for Apostle Inc. and a shareholder in Q State – Biosciences, and a consultant for Gilead Sciences, AstraZeneca, and GoldFinch Bio.

Web Resources

ATAV, https://github.com/igm-team/atav

ClinVar, https://www.ncbi.nlm.nih.gov/clinvar/

Consensus Coding Sequence, https://www.ncbi.nlm.nih.gov/CCDS/

Epi25 Collaborative, http://epi-25.org/

Epi25 WES results browser, http://epi25.broadinstitute.org/

EpiPGX project, http://www.epipgx.eu

Exome Aggregation Consortium (ExAC), http://exac.broadinstitute.org

Exome Variant Server, https://evs.gs.washington.edu/EVS/

Genome Aggregation Database (gnomAD), https://gnomad.broadinstitute.org

Genome Analysis Toolkit (GATK), https://gatk.broadinstitute.org/hc/en-us

lollipops-v.1.5.3, https://github.com/joiningdata/lollipops

NIH Genomic Data Sharing Policy, https://osp.od.nih.gov/scientific-sharing/policies/

MTR-Viewer, http://biosig.unimelb.edu.au/mtr-viewer/

OMIM, https://www.omim.org

Picard, http://broadinstitute.github.io/picard/

R, https://www.R-project.org/

Rare Exome Variant Ensemble Learner (REVEL), https://sites.google.com/site/revelgenomics/

Data and Code Availability

The accession number for the Epi25 Year1 whole-exome sequence data reported in this paper is dbGaP: phs001489. Epi25 Year2 will be available in the near future under the same accession number. Epi25 Year3 is not yet publicly available.

References

1. Aaberg, K.M., Gunnes, N., Bakken, I.J., Lund Soraas, C., Berntsen, A., Magnus, P., Lossius, M.I., Stoltenberg, C., Chin, R., and Suren, P. (2017). Incidence and Prevalence of Childhood Epilepsy: A Nationwide Cohort Study. Pediatrics 139.

2. Fisher, R.S., Acevedo, C., Arzimanoglou, A., Bogacz, A., Cross, J.H., Elger, C.E., Engel, J., Jr., Forsgren, L., French, J.A., Glynn, M., et al. (2014). ILAE official report: a practical clinical definition of epilepsy. Epilepsia 55, 475-482.

3. Hesdorffer, D.C., Logroscino, G., Benn, E.K., Katri, N., Cascino, G., and Hauser, W.A. (2011). Estimating risk for developing epilepsy: a population-based study in Rochester, Minnesota. Neurology 76, 23-27.

4. EpiPM Consortium. (2015). A roadmap for precision medicine in the epilepsies. Lancet Neurol 14, 1219-1228.

5. Ellis, C.A., Petrovski, S., and Berkovic, S.F. (2020). Epilepsy genetics: clinical impacts and biological insights. Lancet Neurol 19, 93-100.

6. May, P., Girard, S., Harrer, M., Bobbili, D.R., Schubert, J., Wolking, S., Becker, F., Lachance-Touchette, P., Meloche, C., Gravel, M., et al. (2018). Rare coding variants in genes encoding GABAA receptors in genetic generalised epilepsies: an exome-based case-control study. Lancet Neurol 17, 699-708.

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

8. Epi4K Consortium; Epilepsy Phenome/Genome Project, Allen, A.S., Berkovic, S.F., Cossette, P., Delanty, N., Dlugos, D., Eichler, E.E., Epstein, M.P., Glauser, T., et al. (2013). De novo mutations in epileptic encephalopathies. Nature 501, 217-221.

9. Epi25 Collaborative. Electronic address: s.berkovic@unimelb.edu.au; Epi25 Collaborative. (2019). Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals. Am J Hum Genet 105, 267-282.

10. Krenn, M., Wagner, M., Hotzy, C., Graf, E., Weber, S., Brunet, T., Lorenz-Depiereux, B., Kasprian, G., Aull-Watschinger, S., Pataraia, E., et al. (2020). Diagnostic exome sequencing in non-acquired focal epilepsies highlights a major role of GATOR1 complex genes. J Med Genet 57, 624-633.

11. Epi4K Consortium. (2016). De Novo Mutations in SLC1A2 and CACNA1A Are Important Causes of Epileptic Encephalopathies. Am J Hum Genet 99, 287-298.

12. EuroEPINOMICS-RES Consortium. Electronic address: euroepinomics-RES@ua.ac.be; Epilepsy Phenome/Genome Project; Epi4K Consortium; EuroEPINOMICS-RES Consortium. (2014). De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. Am J Hum Genet 95, 360-370.

13. Heyne, H.O., Singh, T., Stamberger, H., Abou Jamra, R., Caglayan, H., Craiu, D., De Jonghe, P., Guerrini, R., Helbig, K.L., Koeleman, B.P.C., et al. (2018). De novo variants in neurodevelopmental disorders with epilepsy. Nat Genet 50, 1048-1053.

14. McTague, A., Howell, K.B., Cross, J.H., Kurian, M.A., and Scheffer, I.E. (2016). The genetic landscape of the epileptic encephalopathies of infancy and childhood. Lancet Neurol 15, 304-316.

15. Banerjee, P.N., Filippi, D., and Allen Hauser, W. (2009). The descriptive epidemiology of epilepsy-a review. Epilepsy Res 85, 31-45.

16. Jallon, P., Loiseau, P., and Loiseau, J. (2001). Newly diagnosed unprovoked epileptic seizures: presentation at diagnosis in CAROLE study. Coordination Active du Reseau Observatoire Longitudinal de l' Epilepsie. Epilepsia 42, 464-475.

17. Jallon, P., and Latour, P. (2005). Epidemiology of idiopathic generalized epilepsies. Epilepsia 46 Suppl 9, 10-14.

18. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., and Goldstein, D.B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet 9, e1003709.

19. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnstrom, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat Genet 46, 944-950.

20. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434-443.

21. Bennett, C.A., Petrovski, S., Oliver, K.L., and Berkovic, S.F. (2017). ExACtly zero or once: A clinically helpful guide to assessing genetic variants in mild epilepsies. Neurol Genet 3, e163.

22. Gussow, A.B., Petrovski, S., Wang, Q., Allen, A.S., and Goldstein, D.B. (2016). The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. Genome Biol 17, 9.

23. Larsen, J., Carvill, G.L., Gardella, E., Kluger, G., Schmiedel, G., Barisic, N., Depienne, C., Brilstra, E., Mang, Y., Nielsen, J.E., et al. (2015). The phenotypic spectrum of SCN8A encephalopathy. Neurology 84, 480-489.

24. Stamberger, H., Nikanorova, M., Willemsen, M.H., Accorsi, P., Angriman, M., Baier, H., Benkel-Herrenbrueck, I., Benoit, V., Budetta, M., Caliebe, A., et al. (2016). STXBP1 encephalopathy: A neurodevelopmental disorder including epilepsy. Neurology 86, 954-962.

25. Heron, S.E., and Dibbens, L.M. (2013). Role of PRRT2 in common paroxysmal neurological disorders: a gene with remarkable pleiotropy. J Med Genet 50, 133-139.

26. Leen, W.G., Klepper, J., Verbeek, M.M., Leferink, M., Hofste, T., van Engelen, B.G., Wevers, R.A., Arthur, T., Bahi-Buisson, N., Ballhausen, D., et al. (2010). Glucose transporter-1 deficiency syndrome: the expanding clinical and genetic spectrum of a treatable disorder. Brain 133, 655-670.

27. Wolff, M., Johannesen, K.M., Hedrich, U.B.S., Masnada, S., Rubboli, G., Gardella, E., Lesca, G., Ville, D., Milh, M., Villard, L., et al. (2017). Genetic and phenotypic heterogeneity suggest therapeutic implications in SCN2A-related disorders. Brain 140, 1316-1336.

28. Blanchard, M.G., Willemsen, M.H., Walker, J.B., Dib-Hajj, S.D., Waxman, S.G., Jongmans, M.C., Kleefstra, T., van de Warrenburg, B.P., Praamstra, P., Nicolai, J., et al. (2015). De novo gain-of-function and loss-of-function mutations of SCN8A in patients with intellectual disabilities and epilepsy. J Med Genet 52, 330-337.

29. He, N., Lin, Z.J., Wang, J., Wei, F., Meng, H., Liu, X.R., Chen, Q., Su, T., Shi, Y.W., Yi, Y.H., et al. (2019). Evaluating the pathogenic potential of genes with de novo variants in epileptic encephalopathies. Genet Med 21, 17-27.

30. Gelfman, S., Dugger, S., de Araujo Martins Moreno, C., Ren, Z., Wolock, C.J., Shneider, N.A., Phatnani, H., Cirulli, E.T., Lasseigne, B.N., Harris, T., et al. (2019). A new approach for rare variation collapsing on functional protein domains implicates specific genic regions in ALS. Genome Res 29, 809-818.

31. Hayeck, T.J., Stong, N., Wolock, C.J., Copeland, B., Kamalakaran, S., Goldstein, D.B., and Allen, A.S. (2019). Improved Pathogenic Variant Localization via a Hierarchical Model of Sub-regional Intolerance. Am J Hum Genet 104, 299-309.

32. Traynelis, J., Silk, M., Wang, Q., Berkovic, S.F., Liu, L., Ascher, D.B., Balding, D.J., and Petrovski, S. (2017). Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res 27, 1715-1729.

33. Zhang, J., Kim, E.C., Chen, C., Procko, E., Pant, S., Lam, K., Patel, J., Choi, R., Hong, M., Joshi, D., et al. (2020). Identifying mutation hotspots reveals pathogenetic mechanisms of KCNQ2 epileptic encephalopathy. Sci Rep 10, 4756.

34. Myers, C.T., Hollingsworth, G., Muir, A.M., Schneider, A.L., Thuesmunn, Z., Knupp, A., King, C., Lacroix, A., Mehaffey, M.G., Berkovic, S.F., et al. (2018). Parental Mosaicism in "De Novo" Epileptic Encephalopathies. N Engl J Med 378, 1646-1648.

35. de Lange, I.M., Koudijs, M.J., van 't Slot, R., Gunning, B., Sonsma, A.C.M., van Gemert, L., Mulder, F., Carbo, E.C., van Kempen, M.J.A., Verbeek, N.E., et al. (2018). Mosaicism of de novo pathogenic SCN1A variants in epilepsy is a frequent phenomenon that correlates with variable phenotypes. Epilepsia 59, 690-703.

36. Winawer, M.R., Griffin, N.G., Samanamud, J., Baugh, E.H., Rathakrishnan, D., Ramalingam, S., Zagzag, D., Schevon, C.A., Dugan, P., Hegde, M., et al. (2018). Somatic SLC35A2

variants in the brain are associated with intractable neocortical epilepsy. Ann Neurol 83, 1133-1146.

37. Kim, J.K., Cho, J., Kim, S.H., Kang, H.C., Kim, D.S., Kim, V.N., and Lee, J.H. (2019). Brain somatic mutations in MTOR reveal translational dysregulations underlying intractable focal epilepsy. J Clin Invest 129, 4207-4223.
38. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46, D1062-D1067.
39. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42, D980-985.
40. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). 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 17, 405-424.
41. EPGP Collaborative, Abou-Khalil B, Alldredge B, Bautista J, Berkovic S, Bluvstein J, Boro A, Cascino G, Consalvo D, Cristofaro S, et al. (2013). The epilepsy phenome/genome project. Clin Trials 10, 568-586.
42. Miller, N.A., Farrow, E.G., Gibson, M., Willig, L.K., Twist, G., Yoo, B., Marrs, T., Corder, S., Krivohlavek, L., Walter, A., et al. (2015). A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. Genome Med 7, 100.
43. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297-1303.

44. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 43, 11 10 11-11 10 33.

45. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6, 80-92.

46. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet 99, 877-885.

47. Ren, Z., Povysil, G., Hostyk, J.A., Cui, H., Bhardwaj, N., and Goldstein, D.B. (2021). ATAV: a comprehensive platform for population-scale genomic analyses. BMC Bioinformatics 22, 149.

48. Pruitt, K.D., Harrow, J., Harte, R.A., Wallin, C., Diekhans, M., Maglott, D.R., Searle, S., Farrell, C.M., Loveland, J.E., Ruef, B.J., et al. (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res 19, 1316-1323.

49. Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., and Kang, H.M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 91, 839-848.

50. Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Federhen, S., et al. (2011). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 39, D38-51.

51. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics 26, 2867-2873.

52. Petrovski, S., Todd, J.L., Durheim, M.T., Wang, Q., Chien, J.W., Kelly, F.L., Frankel, C., Mebane, C.M., Ren, Z., Bridgers, J., et al. (2017). An Exome Sequencing Study to Assess the Role of Rare Genetic Variation in Pulmonary Fibrosis. Am J Respir Crit Care Med 196, 82-93.

53. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Genomes, P., and Bustamante, C.D. (2011). Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 108, 11983-11988.

54. Cameron-Christie, S., Wolock, C.J., Groopman, E., Petrovski, S., Kamalakaran, S., Povysil, G., Vitsios, D., Zhang, M., Fleckner, J., March, R.E., et al. (2019). Exome-Based Rare-Variant Analyses in CKD. J Am Soc Nephrol 30, 1109-1122.

55. Blondel, V.D., Guillaume, J.L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics-Theory and Experiment 2008, P10008.

56. Povysil, G., Chazara, O., Carss, K.J., Deevi, S.V.V., Wang, Q., Armisen, J., Paul, D.S., Granger, C.B., Kjekshus, J., Aggarwal, V., et al. (2020). Assessing the Role of Rare Genetic Variation in Patients With Heart Failure. JAMA Cardiol.

57. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv https://arxivorg/abs/180203426.

58. Diaz-Papkovich, A., Anderson-Trocme, L., Ben-Eghan, C., and Gravel, S. (2019). UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. PLoS Genet 15, e1008432.

59. Dai, C.L., Vazifeh, M.M., Yeang, C.H., Tachet, R., Wells, R.S., Vilar, M.G., Daly, M.J., Ratti, C., and Martin, A.R. (2020). Population Histories of the United States Revealed through Fine-Scale Migration and Haplotype Analysis. Am J Hum Genet 106, 371-388.

60. Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11, 415-425.

61. Krusche, P., Trigg, L., Boutros, P.C., Mason, C.E., De La Vega, F.M., Moore, B.L., Gonzalez-Porta, M., Eberle, M.A., Tezak, Z., Lababidi, S., et al. (2019). Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol 37, 555-560.

62. Cummings, B.B., Karczewski, K.J., Kosmicki, J.A., Seaby, E.G., Watts, N.A., Singer-Berk, M., Mudge, J.M., Karjalainen, J., Satterstrom, F.K., O'Donnell-Luria, A.H., et al. (2020). Transcript expression-aware annotation improves rare variant interpretation. Nature 581, 452-458.

63. Xiang, J., Yang, J., Chen, L., Chen, Q., Yang, H., Sun, C., Zhou, Q., and Peng, Z. (2020). Reinterpretation of common pathogenic variants in ClinVar revealed a high proportion of downgrades. Sci Rep 10, 331.

64. Shah, N., Hou, Y.C., Yu, H.C., Sainger, R., Caskey, C.T., Venter, J.C., and Telenti, A. (2018). Identification of Misclassified ClinVar Variants via Disease Population Prevalence. Am J Hum Genet 102, 609-619.

65. Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al. (2015). ClinGen--the Clinical Genome Resource. N Engl J Med 372, 2235-2242.

66. Mantel, N., and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22, 719-748.

67. Cochran, W.G. (1954). Some Methods for Strengthening the Common X2 Tests. Biometrics 10, 417-451.

68. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285-291.

69. Hu, Y.J., Liao, P., Johnston, H.R., Allen, A.S., and Satten, G.A. (2016). Testing Rare-Variant Association without Calling Genotypes Allows for Systematic Differences in Sequencing between Cases and Controls. PLoS Genet 12, e1006040.

70. Jay, J.J., and Brouwer, C. (2016). Lollipops in the Clinic: Information Dense Mutation Plots for Precision Medicine. PLoS One 11, e0160519.

71. Silk, M., Petrovski, S., and Ascher, D.B. (2019). MTR-Viewer: identifying regions within genes under purifying selection. Nucleic Acids Research 47, W121-W126.

72. Goode, D.L., Cooper, G.M., Schmutz, J., Dickson, M., Gonzales, E., Tsai, M., Karra, K., Davydov, E., Batzoglou, S., Myers, R.M., et al. (2010). Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. Genome Res 20, 301-310.

73. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6, e1001025.

74. R Core Team. (2019). R: A Language and Environment for Statistical Computing.(Vienna, Austria: R Foundation for Statistical Computing).

75. McGill, R., Tukey, J.W., and Larsen, W.A. (1978). Variations of box plots. The American Statistician 32, 12-16.

76. de Lange, I.M., Helbig, K.L., Weckhuysen, S., Moller, R.S., Velinov, M., Dolzhanskaya, N., Marsh, E., Helbig, I., Devinsky, O., Tang, S., et al. (2016). De novo mutations of KIAA2022 in females cause intellectual disability and intractable epilepsy. J Med Genet 53, 850-858.

77. Fujiwara, T., Sugawara, T., Mazaki-Miyazaki, E., Takahashi, Y., Fukushima, K., Watanabe, M., Hara, K., Morikawa, T., Yagi, K., Yamakawa, K., et al. (2003). Mutations of sodium channel

alpha subunit type 1 (SCN1A) in intractable childhood epilepsies with frequent generalized tonic-clonic seizures. Brain 126, 531-546.

78. Claes, L., Ceulemans, B., Audenaert, D., Smets, K., Lofgren, A., Del-Favero, J., Ala-Mello, S., Basel-Vanagaite, L., Plecko, B., Raskin, S., et al. (2003). De novo SCN1A mutations are a major cause of severe myoclonic epilepsy of infancy. Hum Mutat 21, 615-621.

79. Carvill, G.L., Weckhuysen, S., McMahon, J.M., Hartmann, C., Moller, R.S., Hjalgrim, H., Cook, J., Geraghty, E., O'Roak, B.J., Petrou, S., et al. (2014). GABRA1 and STXBP1: novel genetic causes of Dravet syndrome. Neurology 82, 1245-1253.

80. Endele, S., Rosenberger, G., Geider, K., Popp, B., Tamer, C., Stefanova, I., Milh, M., Kortum, F., Fritsch, A., Pientka, F.K., et al. (2010). Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. Nat Genet 42, 1021-1026.

81. Hoffbuhr, K., Devaney, J.M., LaFleur, B., Sirianni, N., Scacheri, C., Giron, J., Schuette, J., Innis, J., Marino, M., Philippart, M., et al. (2001). MeCP2 mutations in children with and without the phenotype of Rett syndrome. Neurology 56, 1486-1495.

82. Schubert, J., Siekierska, A., Langlois, M., May, P., Huneau, C., Becker, F., Muhle, H., Suls, A., Lemke, J.R., de Kovel, C.G., et al. (2014). Mutations in STX1B, encoding a presynaptic protein, cause fever-associated epilepsy syndromes. Nat Genet 46, 1327-1332.

83. Krey, I., Krois-Neudenberger, J., Hentschel, J., Syrbe, S., Polster, T., Hanker, B., Fiedler, B., Kurlemann, G., and Lemke, J.R. (2020). Genotype-phenotype correlation on 45 individuals with West syndrome. Eur J Paediatr Neurol 25, 134-138.

84. Dibbens, L.M., de Vries, B., Donatello, S., Heron, S.E., Hodgson, B.L., Chintawar, S., Crompton, D.E., Hughes, J.N., Bellows, S.T., Klein, K.M., et al. (2013). Mutations in DEPDC5 cause familial focal epilepsy with variable foci. Nat Genet 45, 546-551.

85. Weaving, L.S., Christodoulou, J., Williamson, S.L., Friend, K.L., McKenzie, O.L., Archer, H., Evans, J., Clarke, A., Pelka, G.J., Tam, P.P., et al. (2004). Mutations of CDKL5 cause a

severe neurodevelopmental disorder with infantile spasms and mental retardation. Am J Hum Genet 75, 1079-1093.

86. Heron, S.E., Smith, K.R., Bahlo, M., Nobili, L., Kahana, E., Licchetta, L., Oliver, K.L., Mazarib, A., Afawi, Z., Korczyn, A., et al. (2012). Missense mutations in the sodium-gated potassium channel gene KCNT1 cause severe autosomal dominant nocturnal frontal lobe epilepsy. Nat Genet 44, 1188-1190.

87. Barcia, G., Fleming, M.R., Deligniere, A., Gazula, V.R., Brown, M.R., Langouet, M., Chen, H., Kronengold, J., Abhyankar, A., Cilio, R., et al. (2012). De novo gain-of-function KCNT1 channel mutations cause malignant migrating partial seizures of infancy. Nat Genet 44, 1255-1259.

88. Dell'Angelica, E.C., Ohno, H., Ooi, C.E., Rabinovich, E., Roche, K.W., and Bonifacino, J.S. (1997). AP-3: an adaptor-like protein complex with ubiquitous expression. EMBO J 16, 917-928.

89. Ammann, S., Schulz, A., Krageloh-Mann, I., Dieckmann, N.M., Niethammer, K., Fuchs, S., Eckl, K.M., Plank, R., Werner, R., Altmuller, J., et al. (2016). Mutations in AP3D1 associated with immunodeficiency and seizures define a new type of Hermansky-Pudlak syndrome. Blood 127, 997-1006.

90. Seifert, W., Meinecke, P., Kruger, G., Rossier, E., Heinritz, W., Wusthof, A., and Horn, D. (2014). Expanded spectrum of exon 33 and 34 mutations in SRCAP and follow-up in patients with Floating-Harbor syndrome. BMC Med Genet 15, 127.

91. Nikkel, S.M., Dauber, A., de Munnik, S., Connolly, M., Hood, R.L., Caluseriu, O., Hurst, J., Kini, U., Nowaczyk, M.J., Afenjar, A., et al. (2013). The phenotype of Floating-Harbor syndrome: clinical characterization of 52 individuals with mutations in exon 34 of SRCAP. Orphanet J Rare Dis 8, 63.

92. Abou-Khalil, B., Ge, Q., Desai, R., Ryther, R., Bazyk, A., Bailey, R., Haines, J.L., Sutcliffe, J.S., and George, A.L., Jr. (2001). Partial and generalized epilepsy with febrile seizures plus and a novel SCN1A mutation. Neurology 57, 2265-2272.

93. Cossette, P., Liu, L., Brisebois, K., Dong, H., Lortie, A., Vanasse, M., Saint-Hilaire, J.M., Carmant, L., Verner, A., Lu, W.Y., et al. (2002). Mutation of GABRA1 in an autosomal dominant form of juvenile myoclonic epilepsy. Nat Genet 31, 184-189.

94. Strehlow, V., Heyne, H.O., Vlaskamp, D.R.M., Marwick, K.F.M., Rudolf, G., de Bellescize, J., Biskup, S., Brilstra, E.H., Brouwer, O.F., Callenbach, P.M.C., et al. (2019). GRIN2A-related disorders: genotype and functional consequence predict phenotype. Brain 142, 80-92.

95. Carvill, G.L., McMahon, J.M., Schneider, A., Zemel, M., Myers, C.T., Saykally, J., Nguyen, J., Robbiano, A., Zara, F., Specchio, N., et al. (2015). Mutations in the GABA Transporter SLC6A1 Cause Epilepsy with Myoclonic-Atonic Seizures. Am J Hum Genet 96, 808-815.

96. Johannesen, K.M., Gardella, E., Linnankivi, T., Courage, C., de Saint Martin, A., Lehesjoki, A.E., Mignot, C., Afenjar, A., Lesca, G., Abi-Warde, M.T., et al. (2018). Defining the phenotypic spectrum of SLC6A1 mutations. Epilepsia 59, 389-402.

97. Lafreniere, R.G., Cader, M.Z., Poulin, J.F., Andres-Enguix, I., Simoneau, M., Gupta, N., Boisvert, K., Lafreniere, F., McLaughlan, S., Dube, M.P., et al. (2010). A dominant-negative mutation in the TRESK potassium channel is linked to familial migraine with aura. Nat Med 16, 1157-1160.

98. Sun, L., Shi, L., Li, W., Yu, W., Liang, J., Zhang, H., Yang, X., Wang, Y., Li, R., Yao, X., et al. (2009). JFK, a Kelch domain-containing F-box protein, links the SCF complex to p53 regulation. Proc Natl Acad Sci U S A 106, 10195-10200.

99. Gorman, K.M., Meyer, E., Grozeva, D., Spinelli, E., McTague, A., Sanchis-Juan, A., Carss, K.J., Bryant, E., Reich, A., Schneider, A.L., et al. (2019). Bi-allelic Loss-of-Function CACNA1B Mutations in Progressive Epilepsy-Dyskinesia. Am J Hum Genet 104, 948-956.

100. Gardella, E., and Moller, R.S. (2019). Phenotypic and genetic spectrum of SCN8A-related disorders, treatment options, and outcomes. Epilepsia 60 Suppl 3, S77-S85.

101. Ricos, M.G., Hodgson, B.L., Pippucci, T., Saidin, A., Ong, Y.S., Heron, S.E., Licchetta, L., Bisulli, F., Bayly, M.A., Hughes, J., et al. (2016). Mutations in the mammalian target of rapamycin pathway regulators NPRL2 and NPRL3 cause focal epilepsy. Ann Neurol 79, 120-131.

102. Krasniqi, S., and Daci, A. (2019). Role of the Angiotensin Pathway and its Target Therapy in Epilepsy Management. Int J Mol Sci 20.

103. Gasparini, S., Ferlazzo, E., Sueri, C., Cianci, V., Ascoli, M., Cavalli, S.M., Beghi, E., Belcastro, V., Bianchi, A., Benna, P., et al. (2019). Hypertension, seizures, and epilepsy: a review on pathophysiology and management. Neurol Sci 40, 1775-1783.

104. Liu, Y.J., Chen, J., Li, X., Zhou, X., Hu, Y.M., Chu, S.F., Peng, Y., and Chen, N.H. (2019). Research progress on adenosine in central nervous system diseases. CNS Neurosci Ther 25, 899-910.

105. Chen, J.F., Eltzschig, H.K., and Fredholm, B.B. (2013). Adenosine receptors as drug targets--what are the challenges? Nat Rev Drug Discov 12, 265-286.

106. Lesko, S.L., and Rouhana, L. (2020). Dynein assembly factor with WD repeat domains 1 (DAW1) is required for the function of motile cilia in the planarian Schmidtea mediterranea. Dev Growth Differ 62, 423-437.

107. Gupta, A., de Bruyn, G., Tousseyn, S., Krishnan, B., Lagae, L., Agarwal, N., and Consortium, T.S.C.N.H.D. (2020). Epilepsy and Neurodevelopmental Comorbidities in Tuberous Sclerosis Complex: A Natural History Study. Pediatr Neurol 106, 10-16.

108. Lim, J.S., Gopalappa, R., Kim, S.H., Ramakrishna, S., Lee, M., Kim, W.I., Kim, J., Park, S.M., Lee, J., Oh, J.H., et al. (2017). Somatic Mutations in TSC1 and TSC2 Cause Focal Cortical Dysplasia. Am J Hum Genet 100, 454-472.

109. Bamshad, M.J., Nickerson, D.A., and Chong, J.X. (2019). Mendelian Gene Discovery: Fast and Furious with No End in Sight. Am J Hum Genet 105, 448-455.

110. Chubykin, A.A., Atasoy, D., Etherton, M.R., Brose, N., Kavalali, E.T., Gibson, J.R., and Sudhof, T.C. (2007). Activity-dependent validation of excitatory versus inhibitory synapses by neuroligin-1 versus neuroligin-2. Neuron 54, 919-931.

111. Finkbeiner, E., Haindl, M., and Muller, S. (2011). The SUMO system controls nucleolar partitioning of a novel mammalian ribosome biogenesis complex. EMBO J 30, 1067-1078.

112. Kelly, M., Park, M., Mihalek, I., Rochtus, A., Gramm, M., Perez-Palma, E., Axeen, E.T., Hung, C.Y., Olson, H., Swanson, L., et al. (2019). Spectrum of neurodevelopmental disease associated with the GNAO1 guanosine triphosphate-binding region. Epilepsia 60, 406-418.

113. Szczaluba, K., Chmielewska, J.J., Sokolowska, O., Rydzanicz, M., Szymanska, K., Feleszko, W., Wlodarski, P., Biernacka, A., Murcia Pienkowski, V., Walczak, A., et al. (2018). Neurodevelopmental phenotype caused by a de novo PTPN4 single nucleotide variant disrupting protein localization in neuronal dendritic spines. Clin Genet 94, 581-585.

114. Havrilla, J.M., Pedersen, B.S., Layer, R.M., and Quinlan, A.R. (2019). A map of constrained coding regions in the human genome. Nat Genet 51, 88-95.

115. Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O'Donnell-Luria, A.H., Pierce-Hoffman, E., MacArthur, D.G., Neale, B.M., and Daly, M.J. (2017). Regional missense constraint improves variant deleteriousness prediction. bioRxiv, 148353.

116. Hemati, P., Revah-Politi, A., Bassan, H., Petrovski, S., Bilancia, C.G., Ramsey, K., Griffin, N.G., Bier, L., Cho, M.T., Rosello, M., et al. (2018). Refining the phenotype associated with GNB1 mutations: Clinical data on 18 newly identified patients and review of the literature. Am J Med Genet A 176, 2259-2275.

117. Ogden, K.K., Chen, W., Swanger, S.A., McDaniel, M.J., Fan, L.Z., Hu, C., Tankovic, A., Kusumoto, H., Kosobucki, G.J., Schulien, A.J., et al. (2017). Molecular Mechanism of

Disease-Associated Mutations in the Pre-M1 Helix of NMDA Receptors and Potential Rescue Pharmacology. PLoS Genet 13, e1006536.

118. Perez-Palma, E., May, P., Iqbal, S., Niestroj, L.M., Du, J., Heyne, H.O., Castrillon, J.A., O'Donnell-Luria, A., Nurnberg, P., Palotie, A., et al. (2020). Identification of pathogenic variant enriched regions across genes and gene families. Genome Res 30, 62-71.

119. Shearer, A.E., Eppsteiner, R.W., Booth, K.T., Ephraim, S.S., Gurrola, J., 2nd, Simpson, A., Black-Ziegelbein, E.A., Joshi, S., Ravi, H., Giuffre, A.C., et al. (2014). Utilizing ethnic-specific differences in minor allele frequency to recategorize reported pathogenic deafness variants. Am J Hum Genet 95, 445-453.

120. Whiffin, N., Minikel, E., Walsh, R., O'Donnell-Luria, A.H., Karczewski, K., Ing, A.Y., Barton, P.J.R., Funke, B., Cook, S.A., MacArthur, D., et al. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. Genet Med 19, 1151-1158.

121. Yang, S., Lincoln, S.E., Kobayashi, Y., Nykamp, K., Nussbaum, R.L., and Topper, S. (2017). Sources of discordance among germ-line variant classifications in ClinVar. Genet Med 19, 1118-1126.

122. Wright, C.F., Eberhardt, R.Y., Constantinou, P., Hurles, M.E., FitzPatrick, D.R., Firth, H.V.; DDD Study. (2020). Evaluating variants classified as pathogenic in ClinVar in the DDD Study. Genet Med 23, 571-575.

Figure Titles and Legends

Figure 1. Quantile-Quantile Plots for the Protein-Coding Genes with at least One Case or Control Carrier

Qualifying variants were high quality, ultra-rare variants with a predicted functional effect but restricting missense variants to REVEL ≥ 0.5 (when defined). *P*-values were generated from the exact two-sided Cochran-Mantel-Haenszel (CMH) test by gene by cluster to indicate a different carrier status of cases in comparison to controls. *SCN1A* ($p = 4.4 \times 10^{-8}$) and *NEXMIF* (previously known as *KIAA2022*, $p = 8.6 \times 10^{-8}$) achieved study-wide significance $p < 1.6 \times 10^{-7}$ after Bonferroni correction indicated by dashed line (see Statistical Analyses in Methods). (A) Developmental and epileptic encephalopathy (DEE) cases, (B) genetic generalized epilepsy (GEE) cases, and (C) non-acquired focal epilepsy (NAFE) cases. Top ten case enriched genes are labeled. Point coloring determined by CMH odds ratio. Genes labeled in black are known epilepsy genes. Genes labeled in color are candidate epilepsy genes. The green lines represent the 95% confidence interval.

Figure 2. Gene-Set Enrichment Analysis Shows Mild Epilepsies Enriched for Rare Variants in Genes Associated with Severe Epilepsies

Gene-set burden testing using 24 genes drawn from the 43 OMIM epileptic encephalopathy phenotype series with dominant transmission by limiting to genes harboring damaging (REVEL ≥ 0.5) missense variants in all three epilepsies (see Gene-Set Enrichment Testing in Methods, Table S5). All variants are ultra-rare (see Methods). Pooled odds ratio, 95% confidence intervals and FDR corrected *p*-value were generated from the exact two-sided Cochran-Mantel-Haenszel (CMH) test. Odds ratio and FDR-adjusted *p*-values displayed for comparisons with unadjusted *p*-value < 0.05. X-axis displays the $\log_{10}$ of the odds ratio and confidence intervals. PTV = protein-truncating variants, “Damaging” = REVEL ≥ 0.5 (when defined), “Intolerant” = MTR ≤ 0.78 (when defined), DEE = developmental and epileptic encephalopathy, GGE = genetic generalized epilepsy, NAFE = non-acquired focal epilepsy.

Figure 3. Sub-genic intolerance Analysis Reveals Variants Associated with DEE are Located in More Intolerant Genic Sub-Regions

Comparison of cumulative distribution functions weighted by background control variant rate. Genes limited to 24 from OMIM epileptic encephalopathy phenotype series also containing damaging (REVEL ≥ 0.5) missense variants in all three epilepsies (see Gene-Set Enrichment Testing in Methods, Table S5). (A) CDF drawn directly from Epi25 data (dashed line) and weighted by control CDF (solid lines) to estimate “True Positive” distribution. (B) Enlarged box from (A) showing just “True Positive” CDFs with control CDF. “True positive” median MTR DEE = 0.670, GGE = 0.710. NAFE = 0.721. *P*-values generated by 10,000 permutations of Kolmogorov–Smirnov test. Plots calculated from 614 missense variants (DEE = 100, GGE = 133, and NAFE = 153, Control = 228).

Figure 4. Burden of Pathogenic/Likely Pathogenic (P/LP) Variants in ClinVar Found in Epi25 Cases

Ultra-rare and intolerant P/LP variants are enriched in Epi25 cases compared to controls. (A) Variants divided into ultra-rare (absent from non-neuro gnomAD populations) and public (present in non-neuro gnomAD populations) variants showing enrichment only among ultra-rare variants. (B) Ultra-rare variants sub-divided to show drivers of enrichment. “Star” indicates the variant review status in ClinVar, which summarizes the level of review supporting the clinical significance of the variant with increasing number of “gold stars” from 0 to 4 (see Qualifying Variant in Methods). Pooled odds ratio, 95% confidence intervals and FDR corrected *p*-value were generated from the exact two-sided Cochran-Mantel-Haenszel (CMH) test. Odds ratio and FDR-adjusted *p*-values displayed for comparisons with unadjusted *p*-value < 0.05. X-axis displays the $\log_{10}$ of the odds ratio and confidence intervals. PTV = protein-truncating variants, “Int” = “Intolerant” = MTR ≤ 0.78 (when defined), DEE = developmental and epileptic encephalopathy, GGE = genetic generalized epilepsy, NAFE = non-acquired focal epilepsy.

Figure 5. Comparison of Median MTR Scores of Published Ultra-Rare P/LP ClinVar Variants

Violin plots with box plots showing distribution of MTR scores of published missense ClinVar P/LP variants divided into those associated with DEE (N = 302) and non-DEE (N = 29) epilepsies. We considered only those genes harboring missense variants in both groups (14 genes, see Gene-Set Enrichment Testing in Methods, Table S2). Ultra-rare control variants (N = 335) drawn from Epi25 analysis (see Sub-Genic Intolerance Comparison in Methods). Comparisons by Wilcoxon signed-rank test. P values unadjusted. The middle horizontal line represents the median value and the lower and upper hinges represent the 1st and 3rd quartiles. The notches in the boxplot approximate the 95% confidence interval (see Data Analysis and Display in Methods). MTR median ± standard deviation: DEE 0.57 ± 0.24, Non-DEE 0.70 ± 0.18, Control 0.83 ± 0.16. ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$

Figure 6. Burden of Protein Truncating Variants in Intolerant Non-OMIM Genes

The burden of protein truncating variants (PTVs) in genes not associated with a disease in OMIM in epilepsy cases in comparison to controls was assessed. Non-OMIM genes were divided into 10 gene-sets by their intersection with loss-of-function intolerance deciles defined by LOEUF (see Gene-Set Enrichment Testing in Methods, Table S5). Number of genes in each gene-set with at least one PTV in the case-control set is specified in the parenthesis. Pooled odds ratio, 95% confidence intervals and FDR corrected *p*-value were generated from the exact two-sided Cochran-Mantel-Haenszel (CMH) test for (A) developmental and epileptic encephalopathies

(DEE), (B) genetic generalized epilepsy (GGE) and (C) non-acquired focal epilepsy (NAFE). Odds ratio and FDR-adjusted *p* values displayed in parentheses for comparisons with unadjusted *p*-value < 0.05. X-axis displays the odds ratio and confidence intervals.

Table Titles and Legends

Table 1. Non-OMIM Genes Intolerant to Loss-of-Function Variants with Multiple Protein Truncating Variants in Genetic Generalized Epilepsy or Non-Acquired Focal Epilepsy

GGE Gene	*Number of GGE Cases in Epi25*	*GGE P-Value*	*NAFE Gene*	*Number of NAFE Cases in Epi25*	*NAFE P-Value*
NLGN2	3	8.6e-03	WDR18	4	0.01
HDLBP	4	8.9e-03	SOCS7	5	0.01
RC3H2	4	0.01	TRIM9	3	0.05
XPO5	3	0.02	ENAH	2	0.05

Genes listed harbor protein truncating variants in the most loss-of-function intolerant decile of genes in more than one case but absent in controls. Only the top four gene associations are shown per epilepsy. Full tables can be found in the supplement (Tables S37-S38). *P*-values drawn from Ultra-Rare Protein Truncating Variants collapsing analysis (Figure S9, Tables S19-S20).