**Ultra-rare constrained missense variants in the epilepsies: Shared and specific enrichment patterns in neuronal gene-sets**

Mahmoud Koko[1], Roland Krause[2], Thomas Sander[3], Dheeraj Reddy Bobbili[2], Michael Nothnagel[4], Patrick May[2*†], Holger Lerche[1*†], Epi25 Collaborative[5]

[1] Department of Neurology and Epileptology, Hertie Institute for Clinical Brain Research, University of Tübingen, Otfried-Müller Str. 25, 72076 Tübingen, Germany.

[2] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 6 Avenue du Swing, 4367 Belvaux, Luxembourg.

[3] University of Cologne, Faculty of Medicine and Cologne University Hospital, Cologne Center for Genomics, Weyertal 115b, 50931 Cologne, Germany.

[4] University of Cologne, Faculty of Mathematics and Natural Sciences and Cologne University Hospital, Cologne Center for Genomics, Weyertal 115b, 50931 Cologne, Germany.

[5] See contributing authors in the appendix.

[*] Contributed equally as PIs.

[†] Correspondence to:

Prof. Holger Lerche. Address: Department of Neurology and Epileptology, Hertie Institute for Clinical Brain Research, University and University Clinic Tübingen, Hoppe-Seyler-Str. 3, 72076 Tübingen, Germany. E-Mail: holger.lerche@uni-tuebingen.de

or

Dr. Patrick May. Address: Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 6 Avenue du Swing, 4367 Belvaux, Luxembourg. E-Mail: patrick.may@uni.lu

**Abstract:**

**Background:** Burden analysis in epilepsy has shown an excess of deleterious ultra-rare variants (URVs) in few gene-sets, such as known epilepsy genes, constrained genes, ion channel or $GABA_A$ receptor genes. We set out to investigate the burden of URVs in a comprehensive range of gene-sets presumed to be implicated in epileptogenesis.

**Methods:** We investigated several constraint and conservation-based strategies to study whole exome sequencing data from European individuals with developmental and epileptic encephalopathies (DEE, *n* = 1,003), genetic generalized epilepsy (GGE, *n* = 3,064), and non-acquired focal epilepsy (NAFE, *n* = 3,522), collected by the Epi25 Collaborative, compared to 3,962 ancestry-matched controls. The burden of 12 non-synonymous URVs types in 92 gene-sets was compared between epilepsy cases (DDE, GGE, NAFE) and controls using logistic regression analysis.

**Results:** Burden analysis of brain-expressed genes revealed an excess of different URVs types in all three epilepsy categories which was largest for missense variants in highly constrained sites (constrained missense variants). The URVs burden was prominent in neuron-specific, synaptic and developmental genes as well as genes encoding ion channels and receptors, and it was generally higher for DEE and GGE compared to NAFE. The patterns of URVs burden in gene-sets expressed in inhibitory vs. excitatory neurons or receptors suggested a high burden in both groups in DEE but a differential involvement of inhibitory genes in GGE, while excitatory genes were predominantly affected in NAFE. Top ranking susceptibility genes from a recent genome-wide association study (GWAS) of generalized and focal epilepsies displayed a higher URVs burden in constrained coding regions in GGE and NAFE, respectively.

**Conclusions:** Using exome-based gene-set burden analysis, we demonstrate that missense URVs affecting mainly constrained sites are enriched in neuronal genes in both common and rare severe epilepsy syndromes. Our results indicate a differential impact of these URVs in genes expressed in inhibitory vs. excitatory neurons and receptors in generalized vs. focal epilepsies. The excess of URVs

in top-ranking GWAS risk-genes suggests a convergence of rare deleterious and common risk-variants in the pathogenesis of generalized and focal epilepsies.

**Keywords**: burden analysis, ultra-rare variants, gene-sets, epilepsy, exome sequencing.

**Background:**

Dismantling the genetic architecture behind epilepsy is yet to be within reach in many individuals. The role of genetic causality is apparent in the developmental and epileptic encephalopathies (DEE) (1–3), sometimes with consequences on precision treatments (4–7). In contrast, only few individuals with familial or sporadic genetic generalized epilepsies (GGE) or non-acquired focal epilepsies (NAFE) harbor monogenic causative variations (8–10). Therefore, statistical methods investigating the mutational burden of neurobiologically meaningful gene-sets improve the prospects to dissect the joint effects of multiple genetic factors underlying the complex genetic architecture of these common epilepsy syndromes. Such 'gene-set' analysis approaches are likely to provide valuable insights into the role of certain gene-sets and pathways in epilepsy. Recent gene-set burden analyses have shown an enrichment in ultra-rare deleterious variants both in common and rare epilepsies in genes associated with dominant epilepsy syndromes, developmental and epileptic encephalopathy genes, and neuro-developmental disorders (NDD) with epilepsy genes, emphasizing a shared genetic component (8,10). Evidence for the enrichment of rare missense variants in genes encoding $GABA_A$ receptors and GABAergic pathway genes in genetic generalized epilepsies pointed to the importance of the inhibitory pathway (9,10). We used the large-scale dataset collected by the Epi25 Collaborative (10) for a comprehensive, exome-based case-control study to examine the burden of Ultra-Rare Variants (URVs) in a large number of candidate gene-sets for three different epilepsy forms (DEE, GGE, NAFE), aiming to understand the specific roles of deleterious URVs in key pathways implicated in epileptogenesis. Focusing on regional constraint and paralog conservation, we identified relevant and specific gene-set associations in these three epilepsy forms.

**Methods:**

**Study Samples:** The Epi25 Collaborative collected and generated phenotyping and exome sequencing data from individuals with different subtypes of epilepsy (10). We analyzed subjects from recruitment

years 1 and 2 (*n*=13,197). The epilepsy classification, phenotyping and consent procedures have been previously described (10). Five control cohorts, from the Epi25 Collaborative and the database of Genotypes and Phenotypes (dbGAP) (11), were available for this analysis (*n*=13,299), including Italian controls from the Epi25 Collaborative, the Swedish Schizophrenia Study controls (dbGAP: phs000473), and three Myocardial Infarction Genetics (MIGen) Consortium cohorts: Leicester UK Heart Study (dbGAP: phs001000), Ottawa Heart Study (dbGAP: phs000806) and the Italian Atherosclerosis, Thrombosis, and Vascular Biology (ATVB) Study (dbGAP: phs001592). Sequencing was performed on an Illumina HiSeq 2000 or 2500 platform at the Broad Institute (Cambridge, MA, USA). The data generation process has been previously described (10).

**Baseline sample quality control:** For this analysis, we considered samples from three epilepsy categories (DEE, GGE, and NAFE) as classified by the Epi25 phenotype review (10). Controls with Coronary Artery Disease (CAD) diagnosis from the MIGen cohorts were not considered. Samples with genotyping rate less than 90%, outliers on key sample calling metrics (Fig. S1), duplicates and related samples up to the 3rd degree (one from each pair), and samples with ambiguous/discordant sequencing gender (Fig. S2) were removed. Using multidimensional scaling (5 principal components; see Additional file 1), the genotypes of the remaining samples were projected on the 1000 Genomes (12) space (Fig. S3). The major continental ancestry was then predicted using a Support Vector Machine (see Additional file 1). Samples labeled as European were further matched using the top-two principal components (PCs) from multidimensional scaling projection (total of 10 PCs) on the 1000 Genomes European samples space (Fig. S3). Following this baseline filtering, 7,836 cases and 8,822 controls (out of 13,197 cases and 13,299 controls) remained for subsequent analysis. These filtering steps were performed using the Genome Analysis Toolkit (GATK) v4.1.4.1 (13), PLINK v1.9 (14) and KING v2.2.4 (15). For additional details, see the supplemental methods for *Baseline sample quality control* (Additional file 1).

**Baseline variant QC:** Variants located outside Gencode v.33 (16) coding sequencies (CDS) boundaries or in low complexity regions (17) were not considered. Multi-allelic calls were split, and the variants were normalized. Variants with low variant quality score log-odds, covered at mean depth < 10x in the

baseline filtered sample set, at minimum depth of 10x in less than 95% of cases and controls, with large difference in mean depth or call rate between the cases and controls, or with allele count equal to 0 were removed. Low depth/quality genotype calls were set to missing. Variant filtering was performed using bcftools/htslib v1.10.2 (18), vt v0.57721 (19) and GATK (13) as detailed in the supplemental methods for *Baseline variant quality control* (Additional file 1).

**Case control matching and call rate harmonization:** We performed multiple iterations of principal component analysis (PCA) on our baseline filtered dataset. Following the first round of PCA (10 PCs) using PLINK (14), the Swedish Schizophrenia Study control cohort showed poor clustering with the rest of study samples on the top-two components and was removed. Subsequently, outliers on top-two principal components were filtered using EIGENSTRAT v6.1.4 (20,21). For details, see the supplemental methods (Additional file 1). To handle the residual stratification caused by platform differences, we removed all variants with call rate < 95% in any of the remaining sequencing cohorts (Epi25, Leicester, Ottawa, ATVB) or with difference in call rate > 0.5% between any pair of cohorts. Also, variants with Hardy-Weinberg Equilibrium $p$ value less than $1x10^{-6}$ (on the combined case-control cohort) were excluded. These variant calling metrics were obtained using bcftools/htslib (18) and PLINK (14). The final analysis set included 7,589 cases (DEE=1,003, GGE=3,064, NAFE=3,522) and 3,962 matched controls (ATVB = 1,673, Leicester=1,082, Ottawa=924, Epi25 Italian=283) of non-Finnish European ancestry (Table S1 and Table S2). We observed a total of 1,267,392 variants in the final dataset, comprising 1,247,342 SNVs and 20,050 indels. The quality-controlled case-control cohort (Fig. S5) showed a balanced distribution of variants and comparable variant calling metrics (Fig. S6). The use of predominantly male or male-only control cohorts from ATVB and Leicester studies resulted in a misbalanced sample sex ratio (53.6% female cases vs. 19.4% female controls). The effect of this imbalance was addressed in a secondary analysis as will be detailed.

**Qualifying Variants:** The variants were annotated using snpEff v4.3 (22) and Annovar v20191024 (23). We focused on URVs as these have shown a strong burden of deleterious pathogenic variants in multiple studies of epilepsy and other neurological disorders (8,10,24–28). Here, URVs were defined based on their population Minor Allele Frequencies (MAFs) in DiscovEHR ($MAF_{DiscovEHR}$) and

gnomAD r2.1 ($MAF_{gnomAD}$) population database (29,30) and their Minor Allele Counts (MACs) calculated separately for each analysis ($MAC_{DEE+Controls}$, $MAC_{GGE+Controls}$, $MAC_{NAFE+Controls}$), as follows: (i) $MAF_{DiscovEHR} = 0$ (ii) $MAF_{gnomAD} \leq 2x10^{-5}$ (iii) $MAC_{Epilepsy+Controls} \leq 3$ in the respective analysis group. Performing three separate analyses for the three epilepsy phenotypes, with independent calculation of MAC in each analysis, was intended to provide stringent control for inflation. Accordingly, the reported variant counts in the control sets may differ slightly between the three analyses. URVs were categories further into multiple conditions based on their functional consequences and collapsed by gene as qualifying variants (QVs). We considered twelve non-synonymous variant conditions including protein-truncating (presumed loss-of-function) and multiple groups of missense variants (mix of loss- and gain-of-function mechanisms) as well as a (thirteenth) synonymous control condition (presumed neutral). The grouping of missense QVs in multiple (partially overlapping) conditions focused on three perspectives: conventional in-silico deleteriousness, constraint and paralog conservation. It was based on multiple predictions, namely, PolyPhen2 (PPh2) (31), Sorting Intolerant From Tolerant (SIFT) (32), Missense Badness Polyphen and Constraint (MPC) (33), Missense Tolerance Ratio (MTR) (34), Constrained Coding Regions (CCR) (35) and para-Z-score for paralog conservation (36). PPh2 (31) and SIFT (32) are two conventional, commonly used, in-silico missense deleteriousness scores that are widely used in genetic studies to identify likely benign and likely deleterious variants based on a number of features including the sequence, phylogenetic and structural information. MPC score (33) aims to identify regions within genes that are specifically depleted of missense variation and combines this information with variant-level metrics that measures the increased deleteriousness of amino acid substitutions when they occur in missense-constrained regions. MTR score (34) estimates the intolerance of genic regions by comparing the observed proportion of missense variation to the expected proportion in the sequence context of the protein-coding region under study. While MPC and MTR scores are scaled down to individual missense alterations, CCR score (35) aims to identify coding regions that are completely devoid of variation in population databases. Functionally critical protein regions are usually encoded by bases in regions with high CCR scores. Paralog conservation-based missense variant analysis was recently shown to aid variant prioritization in neurodevelopmental disorders (36). It has been proposed that most disease genes in humans have paralogs (37). The analyzed

variant conditions (Table S6) were: (i) Benign missense variants: as predicted by PPh2 and SIFT. (ii) Damaging missense variants as predicted by PPh2 and SIFT. (iii) Protein Truncating Variants (PTVs) that included stop-gained, start-lost, frameshift, splice-donor and splice-acceptor variants. (iv) All functional variants combined PTVs, in-frame indels and damaging missense variants. (v) "MPC 1" missense variants: constrained missense with MPC score ≥ 1. (vi) "MPC 2" missense variants: highly constrained missense with MPC score ≥ 2 (enriched for *de novo* variants). (vii) "MTR ClinVar" missense variants: constrained missense with MTR score ≤ 0.825 which is the median for ClinVar variants not denoted as *de novo*. (viii) "MTR De Novo" missense variants: highly constrained missense with MTR score ≤ 0.565 which is the median for ClinVar *de novo* variants. (ix) "CCR 80" missense variants: highly constrained missense variants in regions with CCR score ≥ 80, with MPC score ≥ 1, and MTR score ≤ 0.825. (x) "paralog-non-conserved": missense variants located in sites not conserved across paralog genes as indicated by a para-Z-score ≤ 0. (xi) "paralog-conserved": missense variants located in sites conserved across paralog genes as indicated by a para-Z-score > 0. (xii) "paralog highly conserved": missense variants in highly conserved sites between paralog genes with para-Z-score ≥ 1. (xiii) "Synonymous" variants that served as a control condition for inflation.

**Gene-sets:** 92 gene-sets were tested. In addition to exome-wide burden testing (one gene-set of all protein coding genes), we defined additional 91 specific gene-sets as follows: (a) 34 sets based on gene expression patterns in the brain: brain-expressed genes grouped by their intolerance profiles including loss-of-function intolerant and missense intolerant genes (30,38), genes grouped by their regional brain expression in the cortex and hippocampus from Genotype-Tissue Expression project v8 (39), developmentally-relevant genes (40–42) , brain-enriched genes from the Human Protein Atlas v20.1 (43,44), genes enriched in specific cell types (45,46) (neurons, excitatory and inhibitory neurons, glia, astrocytes, microglia, oligodendrocytes, and endothelial cells) and localization in certain neuronal compartments (47,48); (b) 28 functional groups including ion channels (8), $GABA_A$ receptors (9), excitatory receptors (9), GABAergic pathway (9), PSD-95 interactors (8), Gene Ontology (GO) gene-sets of GABAergic and glutamatergic synapses (40,41,49), neuronal pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG) (50) and neuronal gene-sets from Reactome database

(51); (c) 14 gene-sets of known disease-related genes including monogenic epilepsy-causing genes (8–10), epilepsy genome-wide association study (GWAS) top-ranking genes from positional mapping within a window of 250 kb around significant loci plus mapping based on chromatin interaction between gene promoters and the significant locus (52), co-regulated genes in the brain (53,54); and (d) 15 non-neuronal gene-sets including genes not expressed in the brain (39,43,44), cancer and metabolic pathways (50). The gene-sets are outlined in Table 1, Table S7 and Table S8 (see Additional file 1 and Additional file 2).

**Table (1): Gene-sets investigated in this study.** (see end of text).

**Gene-set burden analysis:** We examined the burden of QVs in thirteen variant conditions (Table S6) for 92 gene-sets in three epilepsy phenotypes (DEE, GGE, and NAFE) against a set of matched controls. Gene-set burden testing was done using logistic regression by regressing case-control status on the individual QVs counts. In each sample, ultra-rare variants (URVs) that matched the specific analysis condition were collapsed by gene into QVs and these QVs were aggregated (summed) across a target gene-set to get a burden score (assuming equal weights and direction of effects) which was used as a predictor in a binomial model while adjusting for additional covariates (sex, top ten principal components, exome-wide variant count, and exome wide singletons count) using *glm*() function from *stats* package (55). Likelihood ratio test (LRT) from *lmtest* package (56) was used to compare a model with QVs burden and covariates as predictors against a null model (covariates only). Log-odds from LRT and their respective 95% confidence intervals and *p* values are presented here as a measure of enrichment in tested gene-sets (Additional file 3). We employed a Benjamini-Hochberg false discovery rate (FDR) multiple testing adjustment for *p* values that accounted for 3312 tests (92 gene-sets x 3 epilepsy phenotypes x 12 non-synonymous variant conditions, excluding the control condition of synonymous variants) as implemented in *p.adjust*() function from *stats* package (55). The cut-off for substantial enrichment was defined as FDR-adjusted *p* value < 0.05. For simplicity, *p* values (FDR-adjusted except for synonymous variants) are indicated throughout the presented plots using stars as follows: no star > 0.05, * < 0.05, ** < 0.005, *** < 0.0005, **** < 0.00005. To estimate the extent of bias that might have been introduced by the imbalance in male-to-female ratios between cases and

controls, we performed a secondary analysis excluding chromosome X genes (Additional file 4). Also, to ensure adequate control for any bias introduced by differences in capture kits, we performed another secondary analysis between two groups of control samples (Leicester study controls vs. Ottawa and ATVB controls) representing two main enrichment kits (Illumina ICE vs. Agilent SureSelect kits). The statistical analysis was performed in R 3.3.3 (55). The analysis approach is outlined in Fig. 1 and the methods are detailed in Additional file 1. Tables of tested gene-sets and genes in each set are provided in Additional file 2.

**Results:**

**URVs excess in brain-expressed genes**

First, we investigated the burden of URVs across all protein coding genes following the analysis approach outlined in Fig. 1. This revealed a clear enrichment in constrained missense variants that was maximum in consensus constrained coding regions predicted by Missense-badness Polyphen and Constraint (MPC), Missense Tolerance Ratio (MTR) and Consensus Coding Regions (CCR) scores (Fig. 2). The combination of the three metrics identifies highly deleterious variants in functionally critical genic regions (see methods). In this particular analysis in all three phenotypes, about half of the cases, in contrast to roughly one-fourth of controls, harbored one or more QVs in highly constrained regions (Fig. S10). A previous similar analysis of this dataset (10) examined loss-of-function intolerant genes and demonstrated an increased burden in ultra-rare constrained as well as protein truncating variants (PTVs). Here, the examination of brain-expressed intolerant genes showed, similarly, a marked enrichment in PTVs in addition to a burden in constrained missense variants that is comparable to what is seen exome-wide (Fig. 2 and Fig. S11).

When we examined protein coding genes grouped by their relative brain expression, damaging missense variants were only substantially enriched in genes highly expressed in the cortex or hippocampus, whereas those expressed at medium or low levels only showed an enrichment for the most constrained missense variants (Fig. 3). Genes with depleted expression in the brain did not show a substantial enrichment for any variant type (Fig. S15). Genes showing a higher expression in the adult brain compared to other tissues (brain-enriched & brain-enhanced) were also preferentially enriched, as well

as genes associated with brain development. Genes related to late rather than early development showed a slightly higher enrichment in all three phenotypic groups (Fig. 3).

Focusing further on cell-type specific expression, neuron-specific genes were preferentially affected compared to those enriched in glial cells, particularly in GGE (Fig. 4). To obtain further insight into the nature of this neuronal enrichment, we used sets of genes representing paralogs of mouse genes found to be enriched in excitatory or inhibitory neurons (see Additional file 1). Interestingly, genes preferentially expressed in inhibitory neurons showed an increased burden only in GGE, whereas those preferentially expressed in excitatory neurons showed a more prominent signal in NAFE. Since well-established epilepsy genes, like ion channels and receptors, show differential distributions in different neuronal compartments (57,58), we examined further sets of genes based on subcellular localization. We found that pre- and postsynaptic genes were enriched with variants in cases vs. controls, as well as a very small set of 17 genes located in axon initial segments (most prominent in DEE) (Fig. S13).

**Burden of URVs in ion channel, neurotransmitter receptor encoding and related genes**

Next, we examined functional gene-sets that could, more specifically, underlie the observed enrichment in neuronal and synaptic genes. Ion channels, neurotransmitter receptors and transporters are widely implicated in epilepsy, especially in monogenic and familial forms, displaying considerable phenotypic heterogeneity and presenting as mild or severe epilepsies (59–61). Variants in $GABA_A$ receptors were enriched in GGE but not in DEE or NAFE while those in gene-sets representing genes encoding N-Methyl-D-Aspartate receptor and Activity-Regulated Cytoskeleton protein (NMDAR-ARC) were enriched in NAFE and DEE. A comprehensive gene-set for the GABAergic pathway genes (9) showed a prominent signal in GGE and DEE, and less in NAFE. In contrast, a gene-set representing PSD-95 interactors showed comparable enrichment in NAFE and GGE (Fig. 4). Brain-expressed ion channels were found to be enriched for highly constrained missense variants (CCR 80 condition) in common as well as rare epilepsies (Fig. 4).

**Patterns of burden in gene-sets representing inhibitory vs. excitatory signaling**

We then compared the patterns of URVs burden in genes involved in the GABAergic (main inhibitory) pathway and synapse against those in the glutamatergic (main excitatory) pathway and synapse in the brain, by examining their unique and overlapping genes based on KEGG pathways (50) or GO synaptic gene-sets (41) and sets of specific receptors (Fig. 5). GGE showed a higher burden in GABAergic vs. glutamatergic synapse (GO) and pathway (KEGG) genes, in $GABA_A$ receptors vs. excitatory receptors/NMDAR-ARC genes, and in GABAergic pathway genes (comprehensive gene-set) vs. PSD-95 interactors, thus matching the higher burden in genes representing inhibitory vs. excitatory neuronal signaling. The CCR 80 analysis of GO gene-sets in NAFE showed a higher burden in glutamatergic vs. GABAergic synapse genes, akin to the pattern seen in genes enriched in excitatory vs. inhibitory neurons. The analysis of KEGG glutamatergic vs. GABAergic pathway genes did not confirm this finding (Fig. 5). It is notable that the overlap between GO synapse and KEGG pathway gene-sets is minimal (Fig. S18), and the size of GO and KEGG gene-sets was comparable in GABAergic but discordant in glutamatergic genes.

Altogether, these comparisons of the burden in missense variants in highly constrained sites between GGE and NAFE (Fig. 4 and Fig. 5) suggest the following patterns: (i) brain-expressed ion channels, genes enriched in excitatory neurons, enriched in astrocytes, PSD-95 interactors, GABAergic and glutamatergic synapse/pathway genes show an increased burden in cases vs. controls both in GGE & NAFE; (ii) in GGE, this enrichment is coupled with a stronger enrichment in inhibitory neuronal genes, in $GABA_A$ receptors and in GABAergic synapse-specific genes (higher burden in inhibitory vs. excitatory gene-sets). (iii) in NAFE, this is accompanied by an absence of enrichment in the later gene-sets and increased burden in the NMDAR-ARC gene-set (higher burden in excitatory vs. inhibitory gene-sets).

**Burden in gene-sets of known epilepsy-related genes**

The previous Epi25 Collaborative analysis (10) demonstrated a high burden of missense variants in constrained sites (with MPC score ≥ 2) in DEE, GGE, and NAFE, seen in dominant epilepsy genes, DEE genes, and NDD-Epilepsy genes. We observed similar enrichment patterns (Fig. 6) in MPC 2 and MTR De Novo conditions (enriched for de *novo* mutations). In addition, we saw a substantial

enrichment for other variant conditions (MPC score ≥ 1 and MTR ClinVar) with lower odds ratios. Limiting the analysis to highly constrained genic regions (CCR 80 condition) resulted in a marked increase in URVs burden, as was the trend in all the tested gene-sets so far. Testing these sets also unraveled strong enrichment in PTVs and missense variants in paralog-conserved sites. PTVs and missense variants in paralog-conserved sites did not show substantial enrichment in exome-wide analysis and most of other expression-based, localization-based or pathway-based gene-sets. However, we saw a modest increase in PTV burden in highly intolerant genes with probability of Loss-of-function Intolerance (pLI) > 0.995 in all epilepsies (Fig. 2 and Fig. S11). In a gene-set of known DEE genes, where highly intolerant genes are rather prevalent, we saw a prominent enrichment in PTVs burden in DEE. Also, there was an increased burden in missense variants in paralog-conserved sites in sets of epilepsy-related disease genes (DEE genes, dominant Epilepsy genes, NDD-Epilepsy genes). This burden was very strong in DEE but not as remarkable in GGE and NAFE (Fig. 6).

**Enrichment in top GWAS hits captures divergence between common epilepsies**

Recent efforts from the ILAE consortium on complex epilepsies identified multiple associations in a large GWAS of common epilepsies (52). To examine the hypothesis that genes located near the top GWAS hits are also affected by rare variants, we tested the enrichment in sets of the 100 top-ranking genes derived from the ILAE GWAS in generalized, focal, and all epilepsies. Interestingly, when limiting the analysis to Consensus Coding Regions (CCR80 condition), top-ranking genes derived from the GWAS of either generalized or focal epilepsies were preferentially enriched for rare variants in the respective phenotypic groups of GGE and NAFE (Fig. 7). Although the observed enrichment was rather subtle, this result was corroborated by a similar pattern for two, rather small, sets of known epilepsy genes that are predominantly associated with either generalized or focal epilepsy (9).

**Brain- and epilepsy-related co-expression modules**

We also aimed to touch upon the role of brain co-expression modules identified in post-mortem brain tissues from healthy individuals (54) and contrast these to the networks and modules identified in brain tissue derived from epilepsy patients (53). A brain expression module was found to be substantially

enriched for rare deleterious variants in an independent cohort of DEE (54). A link to common epilepsy phenotypes was also inferred, but a burden in ultra-rare variants was not examined so far. This module showed a non-specific enrichment in all three epilepsy subtypes with highest odds in DEE. It is noteworthy that this module overlaps largely with known epilepsy genes (Fig. S19). In resected hippocampi of individuals with temporal lobe epilepsy (TLE), Johnson and colleagues identified two co-expression modules within a gene-regulatory transcriptional network (53). A subtle enrichment was seen in these modules in DEE and GGE, but not NAFE (Fig. 7).

**Additional neuronal and non-neuronal pathways**

Other neuronal gene-sets were enriched in our analysis (Fig. S14). Genes encoding neurexins and neuroligins, important elements of pre- and post-synaptic interaction promoting adhesion between dendrites and axons (62), were enriched in DEE (Fig. S14). Also, the synaptic vesicle cycle pathway (KEGG) showed a prominent signal in both DEE and GGE. We also examined the burden in the mTOR pathway (KEGG), hypothesizing that it could have potential relevance to focal epilepsies, but did not detect a substantial enrichment (Fig. S14). Interestingly, NAFE analysis displayed a burden in endothelial and astrocyte-specific genes in highly constrained genic regions (Fig. 4). Detailed results from all tested conditions including the counts of genes with observed QVs, variant counts in cases and controls, logistic regression odds of the individual QVs burden in cases vs. controls and related *p* values are provided in Additional file 3.

**Specificity of the observed enrichment patterns**

Four sets of genes not expressed in the brain that were tested (high confidence genes with depleted RNA and protein expression in the brain, genes with no RNA detected in the cortex, the hippocampus or any brain tissue) were not substantially enriched in most the tested variant conditions (Fig. S15). Also, we examined eleven metabolic and cancer pathways (KEGG) to have additional insights into the specificity of the observed signals to neuronal processes and genes. Among 540 tests targeting functional variants in these non-neuronal gene-sets (3 epilepsy subtypes, 15 sets representing genes not expressed in the brain, KEGG metabolic and cancer pathways, 12 conditions/types of variants), 18 tests

(3.3%) had an FDR-adjusted $p$ values < 0.05. At least for some of those, the enrichment could be explained by an overlap with genes known to play a role in epilepsy. For instance, genes forming the Type II Diabetes KEGG pathway are substantially enriched in DEE (FDR-adjusted $p$ values of 0.007 for MTR DeNovo and 0.01 for CCR 80 condition). This pathway contains two genes that are known to cause DEE, namely, *CACNA1A* (63) and *CACNA1E* (64). The enrichment is no longer prominent ($p$ values > 0.05) after the removal of these two genes (Fig. S20).

**Bias and inflation in gene-set burden testing**

The analysis for synonymous variants did not show more substantial enrichment than expected by chance, indicating sufficient control for inflation, particularly in exome-wide models and gene-sets with large number of genes. In this control condition, few tests showed $p$ values < 0.05 (15 out of 276 tests of 92 gene-sets and 3 phenotypes: 5.4%). The analysis for benign missense variants, another condition that is not expected to show an increased burden in cases vs. controls (10), did not show substantial enrichment as well. Nine out of 276 tests for benign missense variants (3.2%) showed $p$ values < 0.05 (only 2 with FDR-adjusted $p$ values < 0.05). Possible alternative explanations for such subtle signals include residual population stratification, differences in exome capture not adjusted by covariates and the presence of synonymous variants with functional consequences (65). However, these proportions are close to the limit expected by chance under a true null hypothesis (5% with $\alpha = 0.05$). A potential source of bias in our burden testing was the imbalance in male-to-female ratios between cases and controls (Table S4). We provide results from a secondary analysis that excluded all genes located on chromosome X, which shows that any bias not captured by the inclusion of sample sex as a covariate is likely marginal (Additional file 5). To exclude any major residual stratification resulting from the use of different enrichment kits, we additionally performed a controls-only analysis in which we compared control samples enriched with Illumina ICE capture kits (from Leicester study) to controls enriched using Agilent SureSelect kits (ATVB study and Ottawa study). This analysis reflected a good control for any potential bias introduced by different exome capture systems and also demonstrated that the mixing of controls included (Leicester and Ottawa) or not included (ATVB) in gnomAD is unlikely to have affected our main outcomes (Additional file 5).

**Discussion:**

By analyzing the sequencing data of 11,551 unrelated European individuals, we show an exome-wide burden in ultra-rare missense variants in epilepsy cases compared to controls. These variants are mainly found in constrained sites across three different subtypes of common generalized and focal epilepsies as well as rare and mostly severe developmental and epileptic encephalopathies. This ultra-rare variant burden was prominent for missense variants enriched for *de novo* alterations, yet also substantial in analyses examining a likely combination of inherited and *de novo* variants. While this burden was consistently higher with increased constraint in various gene-sets, PTVs were specifically relevant in known disease genes (Fig. 6) and brain-expressed loss-of-function intolerant genes in comparison to all protein coding genes (Fig. 2). This replicates the findings of the previous Epi25 analysis (10) which indicated an increased PTV burden in known disease-causing epilepsy related genes and genes with pLI > 0.995 but not those with pLI 0.9-0.995 (Fig. S11). In general, PTVs are known to be associated with several phenotypes with their effect being most prominent in LOF-intolerant genes (30,38,66). Similar to PTVs, we did not observe a substantial increase in the burden of missense variants in paralog-conserved sites in a group of all protein coding genes, but it was remarkably high in gene-sets of known disease genes. The increased burden was very prominent in DEE but not in GGE and NAFE. Missense variants in paralog-conserved sites showed prominent enrichment in neurodevelopmental disorders (36), a phenotypic category that overlaps largely with DEE. Although this may reflect a true disparity based on the importance of these highly conserved sites (with their disturbance possibly resulting in severe rather than mild epilepsy phenotypes), it is possible that our analysis lacked power to detect small effect sizes in common epilepsies.

In addition to their utility in estimating the contribution of certain variant conditions/types in different epilepsy phenotypes, gene-sets with known relation to epilepsy, especially to monogenic forms, constitute high-effect-size gene-sets that can also serve as technical validation sets. The presented results are consistent with previous analyses of missense variants in a small number of gene-sets examined in similar cohorts (8–10). The systematic analysis of gene-sets and more different conditions/types of variants revealed interesting findings about the neurobiology of distinct types of

epilepsy and clearly revealed that the pathological burden is markedly higher in constrained coding sites and regions. Although associated with higher odds ratios of an epilepsy phenotype, these variants are not deterministic on their own, since about one-fourth of the controls also carry qualifying variants in the CCR 80 analysis (Fig. S10). As such, the phenotype is determined by a constellation of other factors, possibly including patterns of multiple variations, oligogenic contribution from rare variants (67), and polygenic risk from common variants (68).

The excess of ultra-rare missense variants in the epilepsies clearly stemmed from genes with enriched expression in the brain compared to other tissues, including developmentally relevant genes. A strong link exists between brain development and rare DEE and NDD, since both present with marked developmental deficits accompanied with different seizure types (69,70). The relation of developmental genes to common epilepsies is not as clear. At the cellular level, the enrichment was primarily seen in neuron-specific genes. Synaptic genes with their various pre- and postsynaptic localizations were main drivers in all epilepsies, with prominent signals in postsynaptic genes (Fig. S13). A similar enrichment in synaptic genes has been observed in neurodevelopmental disorders with epilepsy (70,71), schizophrenia (24) and autism (72), highlighting a shared genetic architecture not only between epilepsy subtypes but also with other related neurological disorders, as has been shown previously for common variants (73). Ion channel genes were enriched for ultra-rare, constrained variants in all epilepsy phenotypes (Fig. 4), as has been shown in previous work focusing on ultra-rare variants (8,10).

Despite the common genetic and phenotypic features, DEE, GGE and NAFE represent well-recognized phenotypic clusters with defined electro-encephalographic and clinical characteristics. Given the phenotypic severity of DEE, the prevalence of *de novo* variants and monogenic cases in DEE (those with pathogenic and likely pathogenic variants in known monogenic genes), and the description of phenotypic spectra for genes involved in DEE that also span the milder GGE or NAFE, the distinction between severe and mild epilepsies could be attributed, at least to some extent, to the severity of the genetic defects, their functional effects or their localization within certain channel regions (61,74–77). The distinction between GGE and NAFE, however, is probably functional, at least in part, as suggested by previous work demonstrating the centrality of GABAergic genes in generalized epilepsies (9,10).

Also, it is well recognized that few genes present with focal, but not generalized, epilepsy syndromes (78). Here, phenotype-specific patterns were seen in comparisons of GGE and NAFE. $GABA_A$ receptor genes, GABAergic pathway genes and genes enriched in inhibitory neurons were preferentially affected in GGE in comparison to the glutamatergic pathway genes. In contrast, NMDA receptor and ARC genes, and genes expressed in excitatory neurons were enriched in NAFE. Corroborating this finding for GGE, we found that the enrichment in the GABAergic pathway is stronger in core genes that are not shared with the glutamatergic pathway. An opposing pattern (albeit, not as prominent) was seen in NAFE, where the burden in genes enriched in excitatory neurons and glutamatergic synapse was more prominent than the burden in their inhibitory and GABAergic counterparts, providing evidence for the importance of the former in focal epilepsies.

Additional disparities in key gene-sets point to a possible genetic-functional divergence, so that a common background of shared risk seems to be overlaid by specific risk entities. Groups of known genes implicated in focal vs. generalized epilepsy were enriched in NAFE vs. GGE. Interestingly, the same pattern was found for the 100 top-ranking genes associated with GWAS hits, which were preferentially enriched in respective phenotypic groups. The enrichment of rare variants in GWAS genes also supports the convergence of ultra-rare and common variants in conferring epilepsy risk. According to our findings, a link between common and rare variants is likely to be also relevant for the phenotypic heterogeneity observed in seizure disorders. Notably, polygenic risk scores also pointed out the specificity of the risk profiles in common epilepsies (68). We also found an enrichment of ultra-rare *de novo* variants in DEE in a previously identified brain co-expression module (54), and the same module was also enriched for constrained variants in GGE and NAFE. Although co-expressed genes are not necessarily part of a single pathway, they represent closely orchestrated networks with possible functional correlations. It is therefore conceivable that differentially expressed genes in individuals with epilepsy would highlight modules in which altered transcription, ultra-rare variants, or both contribute to cause both rare and common epilepsies.

**Study limitations**

Despite their robustness, the associations presented in this work should be interpreted with the caveats of gene group testing in mind (79). Given that pathways and molecular processes are not consistently defined in different resources, it is not always easy to define genes that represent a certain pathway. For instance, GABAergic and glutamatergic pathway definitions based on GO terms and KEGG databases are widely discordant (Fig. S18). These differences may explain the discrepancies we observed in enrichment patterns in the same pathway. One gene with abundance of qualifying variants in the GO-based analysis of the glutamatergic pathway was *PPFIA3* (see Additional file 4), a highly intolerant gene that encodes a synaptic receptor tyrosine phosphatase highly expressed in the brain. However, *PPFIA3* is not part of the KEGG-based glutamatergic pathway. Attempting to overcome such discrepancies, we examined multiple overlapping gene-sets from different resources. Associations that are seen regardless of the gene-set source (e.g., the consistent pattern of enrichment in inhibitory vs. excitatory neuronal/pathway genes in GGE vs. NAFE) are, therefore, likely to underscore a genuine biological relevance.

The analysis presented here has additional limitations which we aimed to overcome using stringent analysis and quality control strategies. The limited use of about half of the controls from the primary analysis to maximize case-control matching affected the overall power. Nevertheless, we were able to reproduce most of the major signals from gene-sets with large effect sizes, the latter thereby active as positive controls. The male-to-female ratios were not well-balanced in our case and control sets. Based on a secondary analysis excluding X chromosomal genes, this does not seem to introduce a substantial bias (see Additional file 5). The overlap between the controls used in this study and gnomAD controls resulting from MIGen Leicester and Ottawa controls (Table S2) created some challenges in defining ultra-rare variants, usually defined as those variants not observed in population databases. For population frequency filtering, we allowed around five alleles in gnomAD (allele frequency of $2x10^{-5}$). This count that exceeds our internal filtering cut-off for ultra-rare variants (three alleles) would allow for the retention of ultra-rare variants from our control that are also seen in gnomAD while still filtering common variants and prevalent sequencing artifacts. Also, multiple in-silico algorithms for predicting missense deleteriousness and estimating constraint (including MPC, MTR, and CCR scores) were

derived from or validated in the Exome Aggregation Consortium (80) and gnomAD databases. Examination of control conditions and control gene-sets that are not expected to show an enrichment (e.g., genes not expressed in the brain) did not indicate any prominent inflation and supported the validity of the overall analysis. Interestingly, some signals in 'non-neuronal' gene-sets could be explained by the inclusion of well-established epilepsy genes (Fig. S20), like *CACNA1A* (63) and *CACNA1E* (64), which are in fact key neuronal genes of synaptic transmission.

**Conclusions:**

Missense URVs affecting constrained sites in brain-expressed genes are key genetic drivers in epilepsy. Patients with both common and rare epilepsy subtypes show an increased exome-wide burden of such variants that is primarily derived from neuronal genes, where key gene-sets including ion channels, developmental and synaptic genes are enriched across the phenotypic spectrum of epilepsy. Genes implicated by common GWAS variants may also be disrupted by URVs in various epilepsy phenotypes, suggesting a convergence of rare disruptive variants, and common variants in the pathogenesis of epilepsy. Enrichment patterns of URVs-affected genes suggest a preferential involvement of inhibitory genes in GGE and excitatory genes in focal epilepsies.

**List of abbreviations:**

ARC: Activity-Regulated Cytoskeleton protein.

ATVB: Atherosclerosis, Thrombosis, and Vascular Biology Study.

CAD: Coronary Artery Disease.

CCR: Constrained Coding Regions.

CDS: Coding sequences.

dbGAP: Database of Genotypes and Phenotypes.

DEE: Developmental and Epileptic Encephalopathies.

FDR: False Discovery Rate.

FMRP: Fragile-X Mental Retardation Protein.

GATK: Genome Analysis Toolkit.

GGE: Genetic Generalized Epilepsy.

GO: Gene Ontology.

GWAS: Genome-Wide Association Study.

KEGG: Kyoto Encyclopedia of Genes and Genomes.

ILAE: International League Against Epilepsy.

MAC: Minor Allele Counts.

MAF: Minor Allele Frequencies.

MGI: Mouse Genome Informatics.

MIGen: Myocardial Infarction Genetics Consortium.

MPC: Missense Badness Polyphen and Constraint.

MTR: Missense Tolerance Ratio.

NAFE: Non-Acquired Focal Epilepsy.

NDD: Neuro-Developmental Disorders.

NMDA: N-methyl D-Aspartate.

PCA: Principal Component Analysis.

PCs: Principal Components.

PPh2: PolyPhen2

PSD-95: Post-synaptic density protein 95.

PTVs: Protein Truncating Variants.

QVs: Qualifying Variants.

SIFT: Sorting Intolerant From Tolerant

URVs: Ultra-Rare Variants.

**Declarations:**

Ethics approval and consent to participate: This analysis utilized previously analyzed and published data from the Epi25 Collaborative and the MIGen Consortium. Subjects investigated by the Collaborative provided signed informed consent at the participating centers according to local national ethical requirements and their standards at the time of collection. Approval for data reuse and analysis was obtained from the Epi25 Collaborative (cases) and dbGAP (controls). The ethical approval and consents procedures for the individual cohorts were reported by the Epi25 Collaborative (see methods section).

Consent for publication: not applicable.

Availability of data and materials: The data/analyses presented in the current publication are based on the use of study data from the Epi25 Collaborative (http://epi-25.org/) and the dbGaP (https://ncbi.nlm.nih.gov/gap/) under accessions phs000473, phs001000, phs000806, phs001592 and phs001489. Data supporting the results reported in the article can be found in this published article and its supplementary information files.

Competing interests: None declared.

Funding: RK, TS, MN, PM and HL (and IH and YGW as part of the Epi25 Collaborative) were supported by the Research Unit FOR-2715 of the German Research Foundation and the Fond Nationale de la Recherche of Luxembourg (DFG/FNR, grants INTER/DFG/17/11583046, Le1030/16-1, Sa434/6-1, No755/6-1, He5415/7-1, We4896/4-1). MK was supported by the German Academic Exchange Service (DAAD personal funding program number 57214224). PM obtained additional FNR funding as part of the National Centre of Excellence in Research on Parkinson's disease (NCER-PD, FNR11264123). The Epi25 project 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 (M.J.D.). The Genome Sequencing Program efforts were also supported by NHGRI grant 5U01HG009088-02. Data used for the analyses described in this manuscript were obtained from the GTEx Portal (dbGaP accession number phs000424.v8.p2) on 30/11/2020. The funding bodies had no role in the study design, data collection, analysis, and interpretation or in writing the manuscript.

Authors' contributions: HL, PM, MK conceived the study. HL, PM, RK, and MK contributed to the study design and/or data acquisition. MK and PM performed the analysis with input from HL, RK, TS, DRB and MN. MK, PM, and HL interpreted the results and wrote the manuscript. Authors from the Epi25 Collaborative recruited and phenotyped patients and/or contributed to data generation as indicated in the appendix. All authors reviewed the manuscript for intellectual content and approved the final version of the manuscript.

Acknowledgements: We are cordially thankful to the Epi25 Collaborative staff and all patients participating in the Epi25 project, who all made this study possible.

Epi25 Collaborative: Yen-Chen, Anne, Feng. Daniel, P, Howrigan. Liam, E, Abbott. Katherine, Tashman. Felecia, Cerrato. Tarjinder, Singh. Henrike, Heyne. Andrea, Byrnes. Claire, Churchhouse. Nick, Watts. Matthew, Solomonson. Dennis, Lal. Erin, L, Heinzen. Ryan, S, Dhindsa. Kate, E, Stanley. Gianpiero, L, Cavalleri. Hakon, Hakonarson. Ingo, Helbig. Roland, Krause. Patrick, May. Sarah, Weckhuysen. Slavé, Petrovski. Sitharthan, Kamalakaran. Sanjay, M, Sisodiya. Patrick, Cossette. Chris, Cotsapas. Peter, De Jonghe. Tracy, Dixon-Salazar. Renzo, Guerrini. Patrick, Kwan. Anthony, G, Marson. Randy, Stewart. Chantal, Depondt. Dennis, J, Dlugos. Ingrid, E, Scheffer. Pasquale, Striano. Catharine, Freyer. Kevin, McKenna. Brigid, M, Regan. Susannah, T, Bellows. Costin, Leu. Caitlin, A, Bennett. Esther, M C, Johns. Alexandra, Macdonald. Hannah, Shilling. Rosemary, Burgess. Dorien, Weckhuysen. Melanie, Bahlo. Terence, J, O'Brien. Marian, Todaro. Hannah, Stamberger. Danielle, M, Andrade. Tara, R, Sadoway. Kelly, Mo. Heinz, Krestel. Sabina, Gallati. Savvas, S, Papacostas. Ioanna, Kousiappa. George, A, Tanteles. Katalin, Štěrbová. Markéta, Vlčková. Lucie, Sedláčková. Petra,

Laššuthová. Karl, Martin, Klein. Felix, Rosenow. Philipp, S, Reif. Susanne, Knake. Wolfram, S, Kunz. Gábor, Zsurka. Christian, E, Elger. Jürgen, Bauer. Michael, Rademacher. 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. Christian, Hengsbach. Sarah, Rau. Ana, F, Maisch. Bernhard, J, Steinhoff. Andreas, Schulze-Bonhage. Susanne, Schubert-Bast. Herbert, Schreiber. Ingo, Borggräfe. Christoph, J, Schankin. Thomas, Mayer. Rudolf, Korinthenberg. Knut, Brockmann. Gerhard, Kurlemann. Dieter, Dennig. Rene, Madeleyn. Reetta, Kälviäinen. Pia, Auvinen. Anni, Saarela. Tarja, Linnankivi. Anna-Elina, Lehesjoki. Mark, I, Rees. Seo-Kyung, Chung. William, O, Pickrell. Robert, Powell. 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. Nina, Barišić. Norman, Delanty. Colin, P, Doherty. Arif, Shukralla. Mark, McCormack. Hany, El-Naggar. Laura, Canafoglia. Silvana, Franceschetti. Barbara, Castellotti. Tiziana, Granata. Federico, Zara. Michele, Iacomino. Francesca, Madia. Maria, Stella, Vari. Maria, Margherita, Mancardi. Vincenzo, Salpietro. Francesca, Bisulli. Paolo, Tinuper. Laura, Licchetta. Tommaso, Pippucci. Carlotta, Stipa. Raffaella, Minardi. Antonio, Gambardella. Angelo, Labate. Grazia, Annesi. Lorella, Manna. Monica, Gagliardi. Elena, Parrini. Davide, Mei. Annalisa, Vetro. Claudia, Bianchini. Martino, Montomoli. Viola, Doccini. Carla, Marini. Toshimitsu, Suzuki. Yushi, Inoue. Kazuhiro, Yamakawa. Birute, Tumiene. Lynette, G, Sadleir. Chontelle, King. Emily, Mountier. S, Hande, Caglayan. Mutluay, Arslan. Zuhal, Yapıcı. Uluc, Yis. Pınar, Topaloglu. Bulent, Kara. Dilsad, Turkdogan. Aslı, Gundogdu-Eken. Nerses, Bebek. Sibel, Uğur-İşeri. Betül, Baykan. Barış, Salman. Garen, Haryanyan. Emrah, Yücesan. Yeşim, Kesim. Çiğdem, Özkara. Annapurna, Poduri. Beth, R, Shiedley. Catherine, Shain. 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. Pouya, Khankhanian. Katherine, L, Helbig. Colin, A, Ellis. Gianfranco, Spalletta. Fabrizio, Piras. Federica, Piras. Tommaso, Gili. Valentina, Ciullo. Andreas, Reif. Andrew, McQuillin. Nick, Bass. Andrew, McIntosh. Douglas, Blackwood. Mandy, Johnstone. Aarno, Palotie. Michele, T, Pato. Carlos, N, Pato. Evelyn, J, Bromet.

Celia, Barreto, Carvalho. Eric, D, Achtyes. Maria, Helena, Azevedo. Roman, Kotov. Douglas, S, Lehrer. Dolores, Malaspina. Stephen, R, Marder. Helena, Medeiros. Christopher, P, Morley. Diana, O, Perkins. Janet, L, Sobell. Peter, F, Buckley. Fabio, Macciardi. Mark, H, Rapaport. James, A, Knowles. Consortium, (GPC), Genomic Psychiatry Cohort. Ayman, H, Fanous. Steven, A, McCarroll. Namrata, Gupta. Stacey, B, Gabriel. Mark, J, Daly. Eric, S, Lander. Daniel, H, Lowenstein. David, B, Goldstein. Holger, Lerche. Samuel, F, Berkovic. Benjamin, M, Neale.

Authors' information: See Additional file 1 for the affiliations of the Epi25 Collaborative members.

**References:**

1. McTague A, Howell KB, Cross JH, Kurian MA, Scheffer IE. The genetic landscape of the epileptic encephalopathies of infancy and childhood. Lancet Neurol. 2016 Mar;15(3):304–16. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1474442215002501

2. Happ HC, Carvill GL. A 2020 View on the Genetics of Developmental and Epileptic Encephalopathies. Epilepsy Curr. 2020 Mar;20(2):90–6. Available from: http://journals.sagepub.com/doi/10.1177/1535759720906118

3. Hebbar M, Mefford HC. Recent advances in epilepsy genomics and genetic testing. F1000Research. 2020 Mar 12;9:185. Available from: https://f1000research.com/articles/9-185/v1

4. EpiPM Consortium. A roadmap for precision medicine in the epilepsies. Lancet Neurol. 2015 Dec;14(12):1219–28. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1474442215001994

5. Striano P, Minassian BA. From Genetic Testing to Precision Medicine in Epilepsy. Neurotherapeutics. 2020 Apr;17(2):609–15. Available from: http://link.springer.com/10.1007/s13311-020-00835-4

6. Sisodiya SM. Precision medicine and therapies of the future. Epilepsia. 2021 Mar;62(S2). Available from: https://onlinelibrary.wiley.com/doi/10.1111/epi.16539

7. Lerche H. Drug-resistant epilepsy — time to target mechanisms. Nat Rev Neurol. 2020 Nov;16(11):595–6. Available from: http://www.nature.com/articles/s41582-020-00419-y

8. Epi4K Consortium, Epilepsy Phenome/Genome Project. Ultra-rare genetic variation in common epilepsies: a case-control sequencing study. Lancet Neurol. 2017 Feb;16(2):135–43. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1474442216303593

9. May P, Girard S, Harrer M, Bobbili DR, Schubert J, Wolking S, et al. Rare coding variants in genes encoding GABAA receptors in genetic generalised epilepsies: an exome-based case-control study. Lancet Neurol. 2018 Aug;17(8):699–708. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1474442218302151

10. Epi25 Collaborative. Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals. Am J Hum Genet. 2019 Aug;105(2):267–82. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0002929719302071

11. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res. 2014 Jan;42(D1):D975–9. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1211

12. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74. Available from: http://www.nature.com/articles/nature15393

13. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep 1;20(9):1297–303. Available from: http://genome.cshlp.org/cgi/doi/10.1101/gr.107524.110

14. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015 Dec;4(1):7. Available from: https://academic.oup.com/gigascience/article-lookup/doi/10.1186/s13742-015-0047-8

15. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010 Nov 15;26(22):2867–73. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq559

16. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019 Jan 8;47(D1):D766–73. Available from: https://academic.oup.com/nar/article/47/D1/D766/5144133

17. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, et al. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol. 2019 May;37(5):555–60. Available from: http://www.nature.com/articles/s41587-019-0054-x

18. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352

19. Tan A, Abecasis GR, Kang HM. Unified representation of genetic variants. Bioinformatics. 2015 Jul 1;31(13):2202–4. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv112

20. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006 Aug;38(8):904–9. Available from: http://www.nature.com/articles/ng1847

21. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genet. 2006;2(12):e190. Available from: https://dx.plos.org/10.1371/journal.pgen.0020190

22. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012 Jun;6(2):80–92.

23. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep 1;38(16):e164–e164. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq603

24. Genovese G, Fromer M, Stahl EA, Ruderfer DM, Chambert K, Landén M, et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. Nat Neurosci. 2016 Nov;19(11):1433–41. Available from: http://www.nature.com/articles/nn.4402

25. Demontis D, Satterstrom K, Duan J, Lescai F, Dinesen Østergaard S, Lesch K-P, et al. The Role of Ultra-Rare Coding Variants In ADHD. Eur Neuropsychopharmacol. 2019;29:S724–5. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0924977X1730319X

26. Wilfert AB, Turner TN, Murali SC, Hsieh P, Sulovari A, Wang T, et al. Recent ultra-rare inherited mutations identify novel autism candidate risk genes. Genetics; 2020 Feb . Available from: http://biorxiv.org/lookup/doi/10.1101/2020.02.10.932327

27. Singh T, Walters JTR, Johnstone M, Curtis D, Torniainen M, Rees E, et al. The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. Nat Genet. 2017 Aug;49(8):1167–73. Available from: http://www.nature.com/articles/ng.3903

28. Bennett CA, Petrovski S, Oliver KL, Berkovic SF. ExACtly zero or once: A clinically helpful guide to assessing genetic variants in mild epilepsies. Neurol Genet. 2017 Aug;3(4):e163. Available from: http://ng.neurology.org/lookup/doi/10.1212/NXG.0000000000000163

29. Dewey FE, Murray MF, Overton JD, Habegger L, Leader JB, Fetterolf SN, et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science. 2016 Dec 23;354(6319):aaf6814. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aaf6814

30. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434–43. Available from: http://www.nature.com/articles/s41586-020-2308-7

31. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–9. Available from: http://www.nature.com/articles/nmeth0410-248

32. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 2012 Jul 1;40(W1):W452–7. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks539

33. Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, et al. Regional missense constraint improves variant deleteriousness prediction. bioRxiv. 2017 Jun; Genomics. Available from: http://biorxiv.org/lookup/doi/10.1101/148353

34. Traynelis J, Silk M, Wang Q, Berkovic SF, Liu L, Ascher DB, et al. Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res. 2017 Oct;27(10):1715–29. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.226589.117

35. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome. Nat Genet. 2019 Jan;51(1):88–95. Available from: http://www.nature.com/articles/s41588-018-0294-6

36. Lal D, May P, Perez-Palma E, Samocha KE, Kosmicki JA, Robinson EB, et al. Gene family information facilitates variant interpretation and identification of disease-associated genes in

neurodevelopmental disorders. Genome Med. 2020 Dec;12(1):28. Available from: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00725-6

37. Dickerson JE, Robertson DL. On the Origins of Mendelian Disease Genes in Man: The Impact of Gene Duplication. Mol Biol Evol. 2012 Jan 1;29(1):61–9. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msr111

38. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014 Sep;46(9):944–50. Available from: http://www.nature.com/articles/ng.3050

39. Genotype Tissue Expression Project portal. Available from: https://www.gtexportal.org. Accessed 2020 Dec 1.

40. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25–9. Available from: http://www.nature.com/articles/ng0500_25

41. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021 Jan 8;49(D1):D325–34. Available from: https://academic.oup.com/nar/article/49/D1/D325/6027811

42. Cardoso-Moreira M, Halbert J, Valloton D, Velten B, Chen C, Shao Y, et al. Gene expression across mammalian organ development. Nature. 2019 Jul;571(7766):505–9. Available from: http://www.nature.com/articles/s41586-019-1338-5

43. Human Protein Atlas. Available from: https://www.proteinatlas.org. Accessed 2020 Dec 1.

44. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015 Jan 23;347(6220):1260419–1260419. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.1260419

45. Dusart P, Hallström BM, Renné T, Odeberg J, Uhlén M, Butler LM. A Systems-Based Map of Human Brain Cell-Type Enriched Genes and Malignancy-Associated Endothelial Changes. Cell Rep. 2019 Nov;29(6):1690-1706.e4. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2211124719312926

46. Mo A, Mukamel EA, Davis FP, Luo C, Henry GL, Picard S, et al. Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. Neuron. 2015 Jun;86(6):1369–84. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0896627315004250

47. Pirooznia M, Wang T, Avramopoulos D, Valle D, Thomas G, Huganir RL, et al. SynaptomeDB: an ontology-based knowledgebase for synaptic genes. Bioinformatics. 2012 Mar 15;28(6):897–9. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts040

48. Koopmans F, van Nierop P, Andres-Alonso M, Byrnes A, Cijsouw T, Coba MP, et al. SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. Neuron. 2019 Jul;103(2):217-234.e4. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0896627319304271

49. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. 2015 Dec;1(6):417–25. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2405471215002185

50. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 2021 Jan 8;49(D1):D545–51. Available from: https://academic.oup.com/nar/article/49/D1/D545/5943834

51. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2019 Nov 6;gkz1031. Available from: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz1031/5613674

52. The International League Against Epilepsy Consortium on Complex Epilepsies. Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. Nat Commun. 2018 Dec;9(1):5269. Available from: http://www.nature.com/articles/s41467-018-07524-z

53. Johnson MR, Behmoaras J, Bottolo L, Krishnan ML, Pernhorst K, Santoscoy PLM, et al. Systems genetics identifies Sestrin 3 as a regulator of a proconvulsant gene network in human epileptic hippocampus. Nat Commun. 2015 May;6(1):6031. Available from: http://www.nature.com/articles/ncomms7031

54. Delahaye-Duriez A, Srivastava P, Shkura K, Langley SR, Laaniste L, Moreno-Moral A, et al. Rare and common epilepsies converge on a shared gene regulatory network providing opportunities for novel antiepileptic drug discovery. Genome Biol. 2016 Dec;17(1):245. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1097-7

55. R Core Team. R: A language and environment for statistical computing. Available from: https://www.R-project.org/

56. Zeileis, Hothorn. Diagnostic Checking in Regression Relationships. Rnews. 2002. Available from: https://CRAN.R-project.org/doc/Rnews/

57. Vacher H, Mohapatra DP, Trimmer JS. Localization and Targeting of Voltage-Dependent Ion Channels in Mammalian Central Neurons. Physiol Rev. 2008 Oct;88(4):1407–47. Available from: https://www.physiology.org/doi/10.1152/physrev.00002.2008

58. Martenson JS, Tomita S. Synaptic localization of neurotransmitter receptors: comparing mechanisms for AMPA and GABAA receptors. Curr Opin Pharmacol. 2015 Feb;20:102–8. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1471489214001660

59. Lerche H, Shah M, Beck H, Noebels J, Johnston D, Vincent A. Ion channels in genetic and acquired forms of epilepsy: Ion channels in epilepsy. J Physiol. 2013 Feb;591(4):753–64. Available from: http://doi.wiley.com/10.1113/jphysiol.2012.240606

60. Oyrer J, Maljevic S, Scheffer IE, Berkovic SF, Petrou S, Reid CA. Ion Channels in Genetic Epilepsy: From Genes and Mechanisms to Disease-Targeted Therapies. Pharmacol Rev. 2018 Jan;70(1):142–73. Available from: http://pharmrev.aspetjournals.org/lookup/doi/10.1124/pr.117.014456

61. Maljevic S, Møller RS, Reid CA, Pérez-Palma E, Lal D, May P, et al. Spectrum of GABAA receptor variants in epilepsy. Curr Opin Neurol. 2019 Apr;32(2):183–90. Available from: https://journals.lww.com/00019052-201904000-00003

62. Craig AM, Kang Y. Neurexin–neuroligin signaling in synapse development. Curr Opin Neurobiol. 2007 Feb;17(1):43–52. Available from: https://linkinghub.elsevier.com/retrieve/pii/S095943880700013X

63. Epi4K Consortium. De Novo Mutations in SLC1A2 and CACNA1A Are Important Causes of Epileptic Encephalopathies. Am J Hum Genet. 2016 Aug;99(2):287–98. Available from: https://linkinghub.elsevier.com/retrieve/pii/S000292971630204X

64. Helbig KL, Lauerer RJ, Bahr JC, Souza IA, Myers CT, Uysal B, et al. De Novo Pathogenic Variants in CACNA1E Cause Developmental and Epileptic Encephalopathy with Contractures, Macrocephaly, and Dyskinesias. Am J Hum Genet. 2018 Nov;103(5):666–78. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0002929718303173

65. Takata A, Ionita-Laza I, Gogos JA, Xu B, Karayiorgou M. De Novo Synonymous Mutations in Regulatory Elements Contribute to the Genetic Etiology of Autism and Schizophrenia. Neuron. 2016 Mar 2;89(5):940–7.

66. Ganna A, Satterstrom FK, Zekavat SM, Das I, Kurki MI, Churchhouse C, et al. Quantifying the Impact of Rare and Ultra-rare Coding Variation across the Phenotypic Spectrum. Am J Hum Genet. 2018 Jun 7;102(6):1204–11.

67. Takata A, Nakashima M, Saitsu H, Mizuguchi T, Mitsuhashi S, Takahashi Y, et al. Comprehensive analysis of coding variants highlights genetic complexity in developmental and epileptic encephalopathy. Nat Commun. 2019 Dec;10(1):2506. Available from: http://www.nature.com/articles/s41467-019-10482-9

68. Leu C, Stevelink R, Smith AW, Goleva SB, Kanai M, Ferguson L, et al. Polygenic burden in focal and generalized epilepsies. Brain. 2019 Nov 1;142(11):3473–81. Available from: https://academic.oup.com/brain/article/142/11/3473/5585821

69. Scheffer IE, Berkovic S, Capovilla G, Connolly MB, French J, Guilhoto L, et al. ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. Epilepsia. 2017 Apr;58(4):512–21. Available from: http://doi.wiley.com/10.1111/epi.13709

70. Heyne HO, Singh T, Stamberger H, Abou Jamra R, Caglayan H, Craiu D, et al. De novo variants in neurodevelopmental disorders with epilepsy. Nat Genet. 2018 Jul;50(7):1048–53.

71. EuroEPINOMICS-RES Consortium, Epilepsy Phenome/Genome Project, Epi4K Consortium. De Novo Mutations in Synaptic Transmission Genes Including DNM1 Cause Epileptic Encephalopathies. Am J Hum Genet. 2014 Oct;95(4):360–70. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0002929714003838

72. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014 Nov;515(7526):209–15. Available from: http://www.nature.com/articles/nature13772

73. The Brainstorm Consortium. Analysis of shared heritability in common disorders of the brain. Science. 2018 Jun 22;360(6395):eaap8757. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aap8757

74. Johannesen KM, Gardella E, Linnankivi T, Courage C, de Saint Martin A, Lehesjoki A-E, et al. Defining the phenotypic spectrum of *SLC6A1* mutations. Epilepsia. 2018 Feb;59(2):389–402. Available from: http://doi.wiley.com/10.1111/epi.13986

75. Liu Y, Schubert J, Sonnenberg L, Helbig KL, Hoei-Hansen CE, Koko M, et al. Neuronal mechanisms of mutations in *SCN8A* causing epilepsy or intellectual disability. Brain. 2019 Feb 1;142(2):376–90. Available from: https://academic.oup.com/brain/article/142/2/376/5277044

76. Wolking S, May P, Mei D, Møller RS, Balestrini S, Helbig KL, et al. Clinical spectrum of *STX1B* -related epileptic disorders. Neurology. 2019 Feb 8;10.1212/WNL.0000000000007089. Available from: http://www.neurology.org/lookup/doi/10.1212/WNL.0000000000007089

77. Vardar G, Gerth F, Schmitt XJ, Rautenstrauch P, Trimbuch T, Schubert J, et al. Epilepsy-causing STX1B mutations translate altered protein functions into distinct phenotypes in mouse neurons. Brain. 2020 Jul 1;143(7):2119–38. Available from: https://academic.oup.com/brain/article/143/7/2119/5861027

78. Wang J, Lin Z-J, Liu L, Xu H-Q, Shi Y-W, Yi Y-H, et al. Epilepsy-associated genes. Seizure. 2017 Jan;44:11–20. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1059131116302989

79. de Leeuw CA, Neale BM, Heskes T, Posthuma D. The statistical properties of gene-set analysis. Nat Rev Genet. 2016 Jun;17(6):353–64. Available from: http://www.nature.com/articles/nrg.2016.29

80. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fenell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug;536(7616):285–91. Available from: http://www.nature.com/articles/nature19057

**Table (1): Gene-sets investigated in this study.**

**Group of all protein coding genes (1):**

-all genes annotated by snpEff as protein coding.

**Groups based on brain expression (34):** Expression in the brain, regional, cellular and sub-cellular expression patterns.

***Brain-expressed LOF-intolerant genes:***
excluding genes with no expression in the cortex/hippocampus

-pLI > 0.995.
-pLI 0.9-0.995.
-pLI 0.8-0.9.

***Brain-expressed missense-intolerant genes:***
excluding genes with no expression in the cortex/hippocampus

-Z-score > 3.09.
-Z-score 2.5-3.09.
-Z-score 2-2.5.

***Cortical and hippocampal expression level:***
-High, Moderate, Low in the cortex.
-High, Moderate, Low in the hippocampus.

***Brain development:***
-Brain development genes (Gene-Ontology group).
-Brain developmental genes (extended group).
-Early developmental genes.
-Late developmental genes.

***Enrichment in the brain:***
-Brain-enriched -Brain-enhanced.

***Cell-type-specific enrichment:***
-Neurons -glial cells
-Excitatory neurons -Inhibitory neurons -Astrocytes -Microglia -Oligodendrocytes -Endothelium.

***Neuronal Localization:***
-Axon Initial Segment.
-Synaptic (curated group).
-Synaptic (extended group).
-Synaptic vesicle and active zone.
-Pre-synaptic.
-Post-synaptic.
-Pre-synaptic only.
-Post-synaptic only.

**Functional gene-sets (28):** Ion channels, transporters, synaptic cycles, pathways and neurotransmitter cycles.

***Ion channels, neurotransmitter receptors and related genes:***
-Voltage-gated ion channels.
-Voltage-gated cation channels.
-Brain-specific voltage-gated ion channels.
-$GABA_A$ receptors.
-GABAergic pathway.
-Excitatory receptors.
-NMDAR & ARC.
-PSD-95 interactors.

***GABAergic/Glutamatergic synapses (GO groups):***
-GABAergic synapse -Glutamatergic synapse
-only in GABAergic -only in glutamatergic
-shared genes.

***GABAergic/Glutamatergic pathways (KEGG database):***
-GABAergic pathway -Glutamatergic pathway
-only in GABAergic -only in glutamatergic
-shared genes.

***Additional neuronal pathways (KEGG):***
-Cholinergic pathway.
-Dopaminergic pathway.
-mTOR pathway.
-Synaptic vesicle cycle.

Glutamate release, uptake and clearance cycle.

***GABA/glutamate cycles (Reactome database; pooled from multiple groups):***
-GABA release, receptor activation, and clearance -

***Additional neuronal groups (Reactome database):***
-Presynaptic depolarization.
-Neurexins and Neuroligins.
-Synaptic Adhesion molecules.
-Receptor-type Protein Tyrosine Phosphatases.

| **Disease-associated and intolerant genes (14):** Genes and gene-sets with known associations with epilepsy and related neurological diseases | | |
|---|---|---|
| ***Monogenic disease-causing genes:***<br>-Generalized epilepsy genes.<br>-Focal epilepsy genes.<br>-Dominant epilepsy genes<br>-DEE genes.<br>-NDD with epilepsy genes.<br>-FMRP targets.<br>-MGI seizure genes. | ***Top-ranking 100 genes in ILAE2 GWAS:***<br>-Generalized epilepsy GWAS.<br>-Focal epilepsy GWAS.<br>-All epilepsies GWAS.<br><br>***Brain co-expression module:***<br>-Co-expressed module identified in non-diseased post-mortem brain tissues.<br>(enriched for *de novo* variants in DEE). | ***Regulatory and co-expression modules in epilepsy:***<br>-Co-expression network identified in brain tissues of Temporal Lobe Epilepsy patients<br>- Two modules within this network. |
| **Control groups (15):** | | |
| ***Genes not expressed in the brain:***<br>-RNA not detected in cortex, in hippocampus, or all GTEx regions.<br>-Protein is depleted in the brain. | ***KEGG metabolic pathways:***<br>-Type II Diabetes.<br>-Carbohydrate Absorption & Digestion.<br>-Protein Absorption & Digestion.<br>-Fat Absorption & Digestion. | ***KEGG cancer pathways:***<br>- CA Breast, CA Lung, CA Colon, CA Prostate, Renal Cell Ca, CA Pancreas, Hepatocellular Ca. |

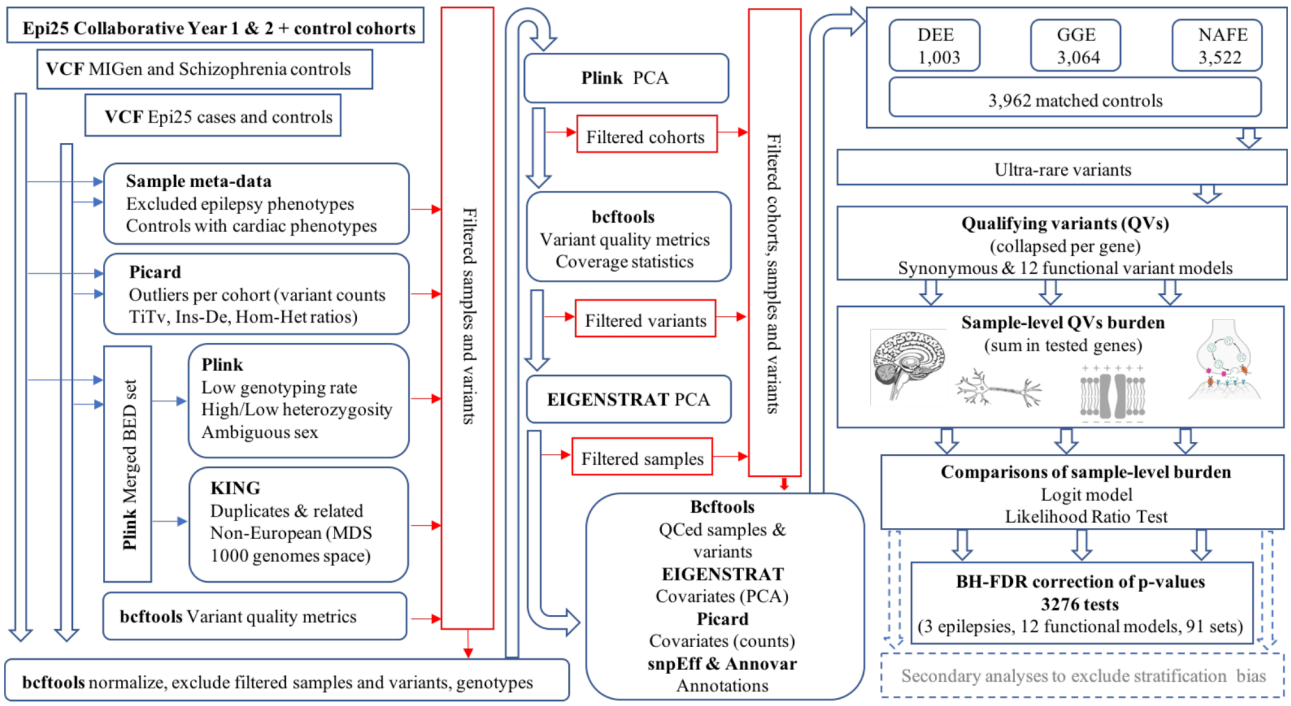


**Fig. 1: Outlines of the burden analysis method.** Thirteen (twelve functional and one synonymous) variants models with focus on missense variants in constrained or paralog-conserved sites were tested in the three epilepsy phenotypes against a shared set of matched controls. The burden was examined in 92 gene-sets (detailed in table 1) using a logistic regression model with the count of qualifying variants per sample as a predictor and sample sex, ten principal components, singletons and exome-wide variant counts as covariates. Two secondary analyses were performed: one analysis restricting gene-sets to autosomal genes (to exclude bias introduced by male-to-female ratio imbalances) and another analysis testing controls prepared for exome sequencing using Illumina ICE capture kits against controls prepared with Agilent SureSelect capture kits (to exclude bias caused by differences in enrichment kits).

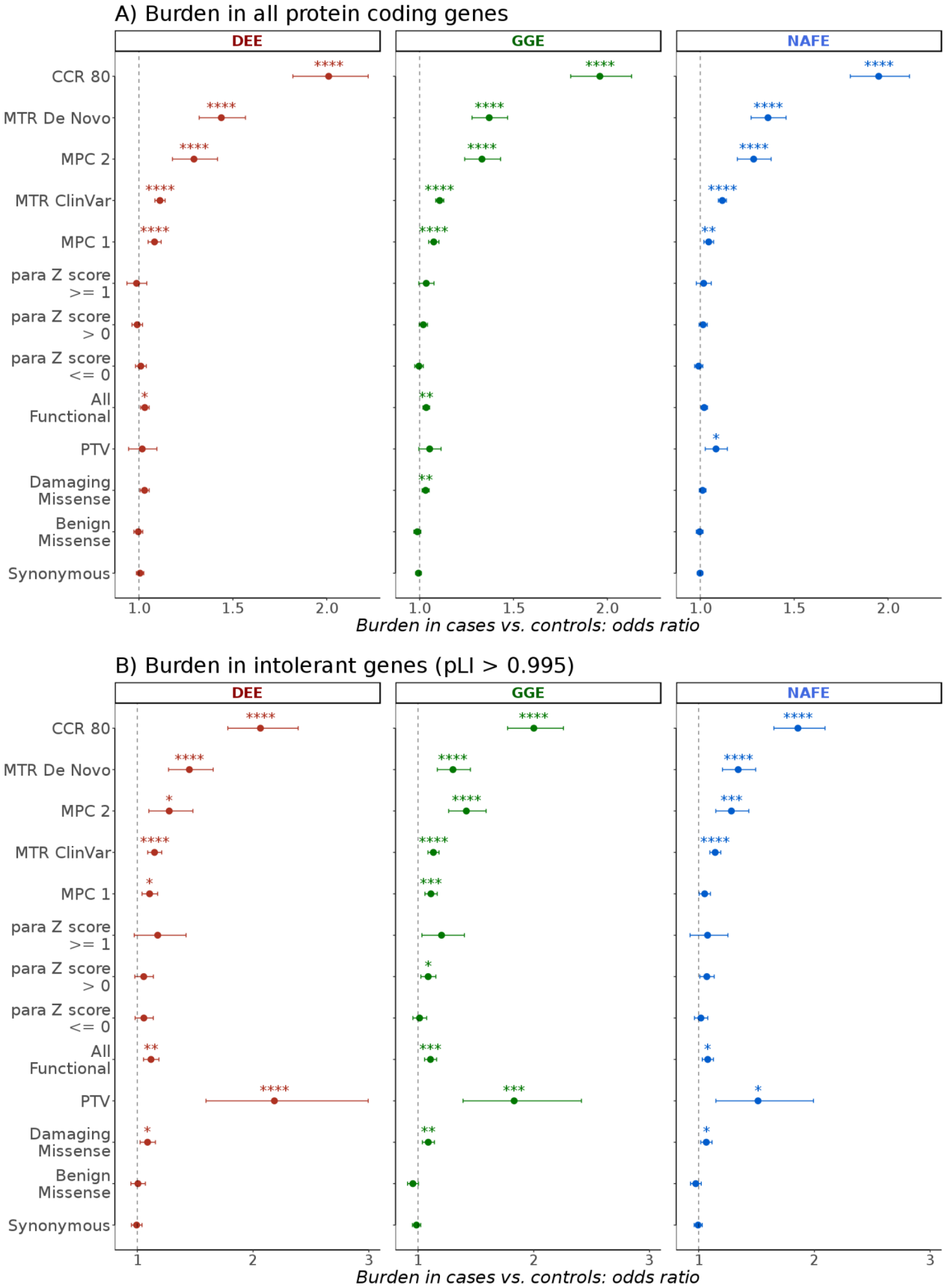


**Fig. 2: Exome-wide burden of ultra-rare variants in the epilepsies.** The burden in developmental and epileptic encephalopathies (DEE), genetic generalized epilepsies (GGE) and non-acquired focal epilepsies (NAFE) in 19,402 protein coding genes **(A)** and 1,743 loss-of-function intolerant genes with pLI score > 0.995 **(B)** is shown in multiple test and control conditions (y-axis) as odds ratio (x-axis) from Likelihood Ratio Test (bars indicate 95% confidence intervals). FDR corrected *p* values (not corrected for synonymous variants) are indicated with stars as follows: no star > 0.05, * < 0.05, ** < 0.005, *** < 0.005, **** < 0.0005. There is an incrementing burden with a higher level of missense constraint. Both synonymous and benign missense alterations are not enriched, suggesting sufficient control for inflation.

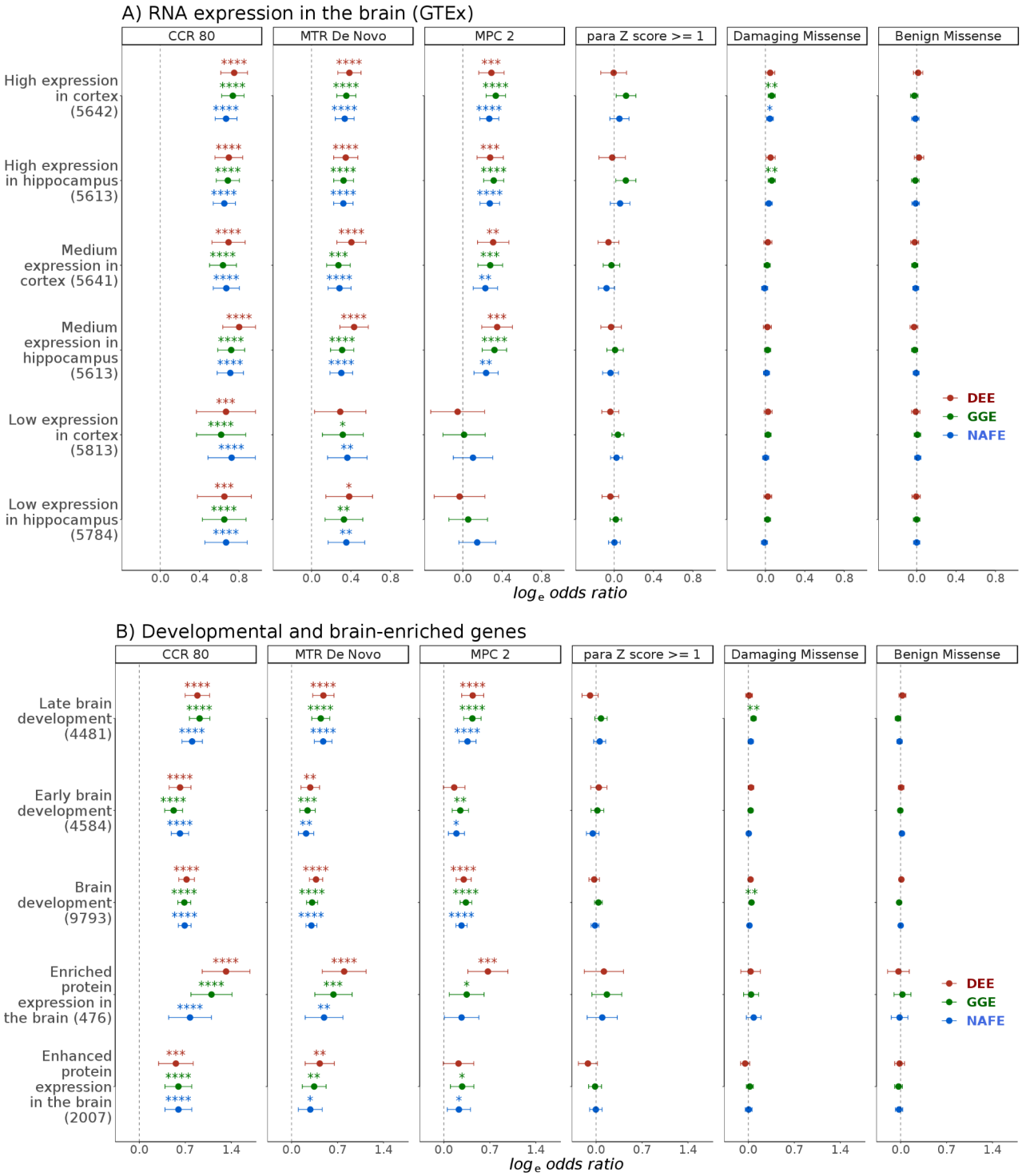


**Fig. 3: Burden of ultra-rare missense variants in brain expressed and developmental genes.** The burden of benign or damaging missense variants and missense variants in highly paralog-conserved or highly constrained sites in developmental and epileptic encephalopathies (DEE), genetic generalized epilepsies (GGE) and non-acquired focal epilepsies (NAFE) is shown in gene-sets based on levels of RNA/protein expression in the cortex and hippocampus **(A)** or enrichment in adult or developing brain **(B)**. Gene-sets are shown on the y-axis (number of genes between brackets). Log odds ratio (Likelihood Ratio Test) are shown on the x-axis (error bars indicate 95% confidence intervals). The variant conditions are shown in vertical panels. FDR corrected *p* values (not corrected for synonymous variants) are indicated with stars as follows: no star > 0.05, * < 0.05, ** < 0.005, *** < 0.005, **** < 0.0005. Brain enriched genes (Human Protein Atlas: more than four-fold expression compared to other tissues) show higher burden in DEE compared to those genes with only enhanced expression (higher but less than four-fold expression in the brain). This difference is less prominent in GGE and NAFE. Developmental genes enriched in late development show higher burden than genes enriched in early development.

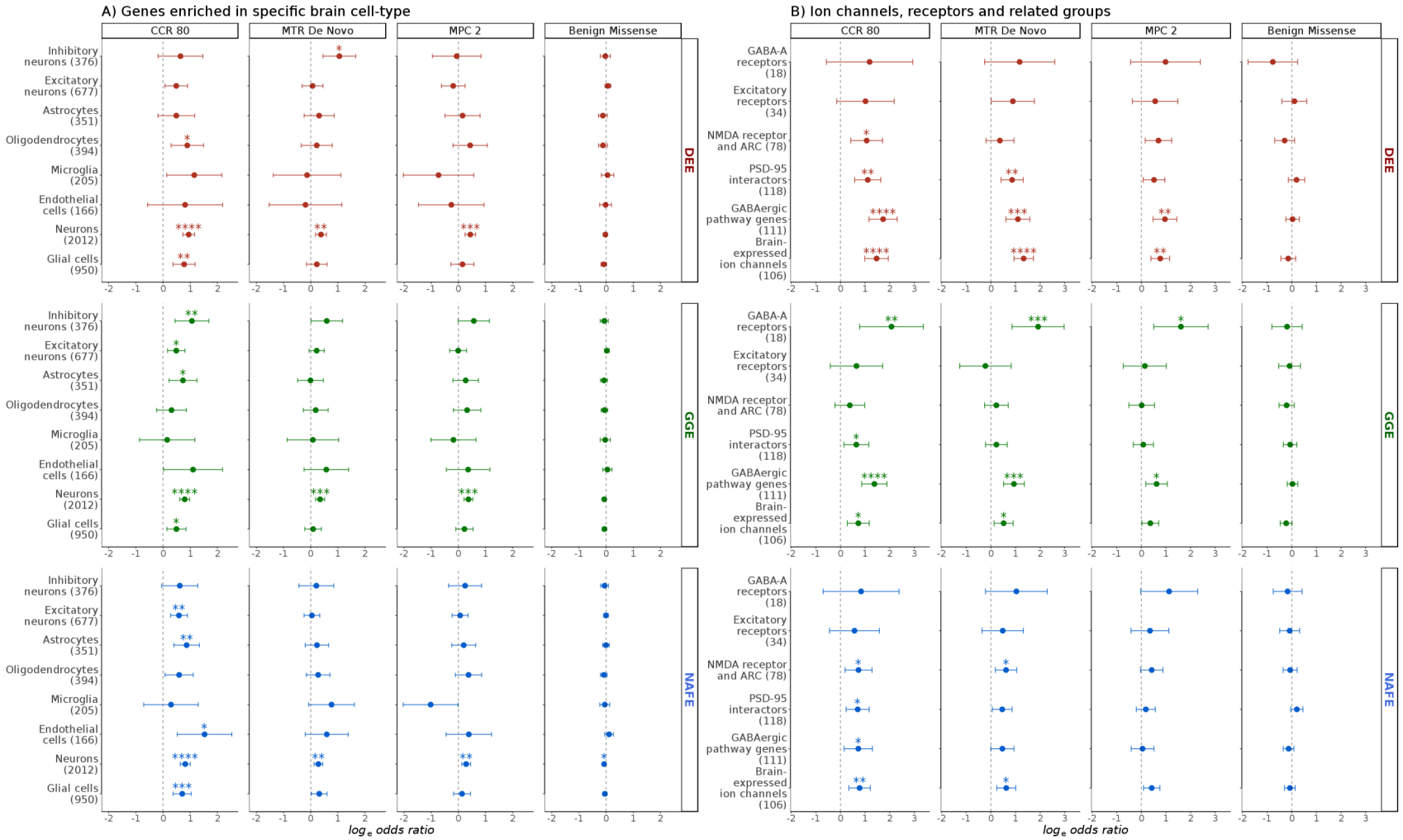


**Fig. 4: Burden in neuronal and glial cells, ion channels, receptors and related interactors.** The burden in developmental and epileptic encephalopathies (DEE), genetic generalized epilepsies (GGE) and non-acquired focal epilepsies (NAFE) in shown on the x-axis (log-odds from Likelihood Ratio Test; error bars indicate 95% confidence intervals). Gene-sets are shown on the y-axis (number of genes between brackets). The variant conditions are shown in vertical panels. FDR corrected *p* values (not corrected for synonymous variants) are indicated with stars as follows: no star > 0.05, * < 0.05, ** < 0.005, *** < 0.005, **** < 0.0005. **(A)** Comparisons based on the cellular expression patterns show that the enrichment is prominent in neuron-enriched compared to glial genes, particularly in GGE. GGE shows relatively higher enrichment in inhibitory neuron genes while NAFE shows a relatively higher burden in excitatory neuronal genes. **(B)** Variants in $GABA_A$ receptors and GABAergic pathway are preferentially enriched in GGE compared to groups of genes coding for NMDA receptor and neuronal activity-regulated cytoskeleton-associated protein (ARC). In NAFE, the variants are enriched in the NMDA receptor & ARC gene-sets, but not $GABA_A$ receptors.

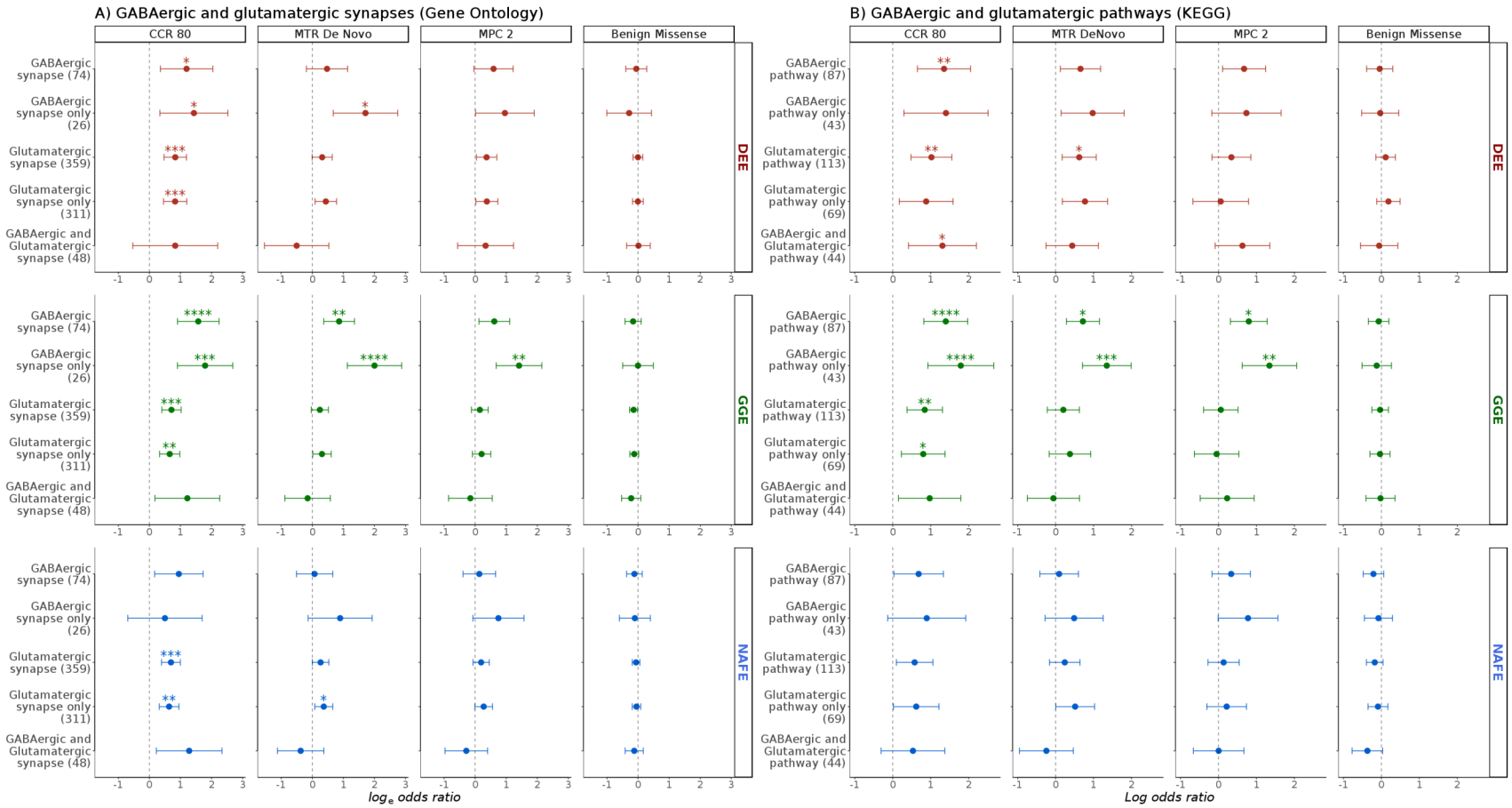


**Fig. 5: Enrichment in major neuronal synapses and pathways.** Comparison of enrichment patterns in GABAergic and glutamatergic synapses and pathway genes based on Gene-Ontology (GO) **(A)** or Kyoto Encyclopedia for Genes and Genomes (KEGG) **(B)** The burden in developmental and epileptic encephalopathies (DEE), genetic generalized epilepsies (GGE) and non-acquired focal epilepsies (NAFE) in shown on the x-axis (log-odds from Likelihood Ratio Test; error bars indicate 95% confidence intervals). Gene-sets are shown on the y-axis (number of genes between brackets). The variant conditions are shown in vertical panels. FDR corrected *p* values (not corrected for synonymous variants) are indicated with stars as follows: no star > 0.05, * < 0.05, ** < 0.005, *** < 0.005, **** < 0.0005. Complete groups, genes specific to one of the two synapses/pathways as well as their intersection were tested. Both GO and KEGG gene sets show an increased burden in GGE when "GABAergic only" genes are tested. GGE shows a relatively higher burden in GABAergic compared to glutamatergic gene sets.

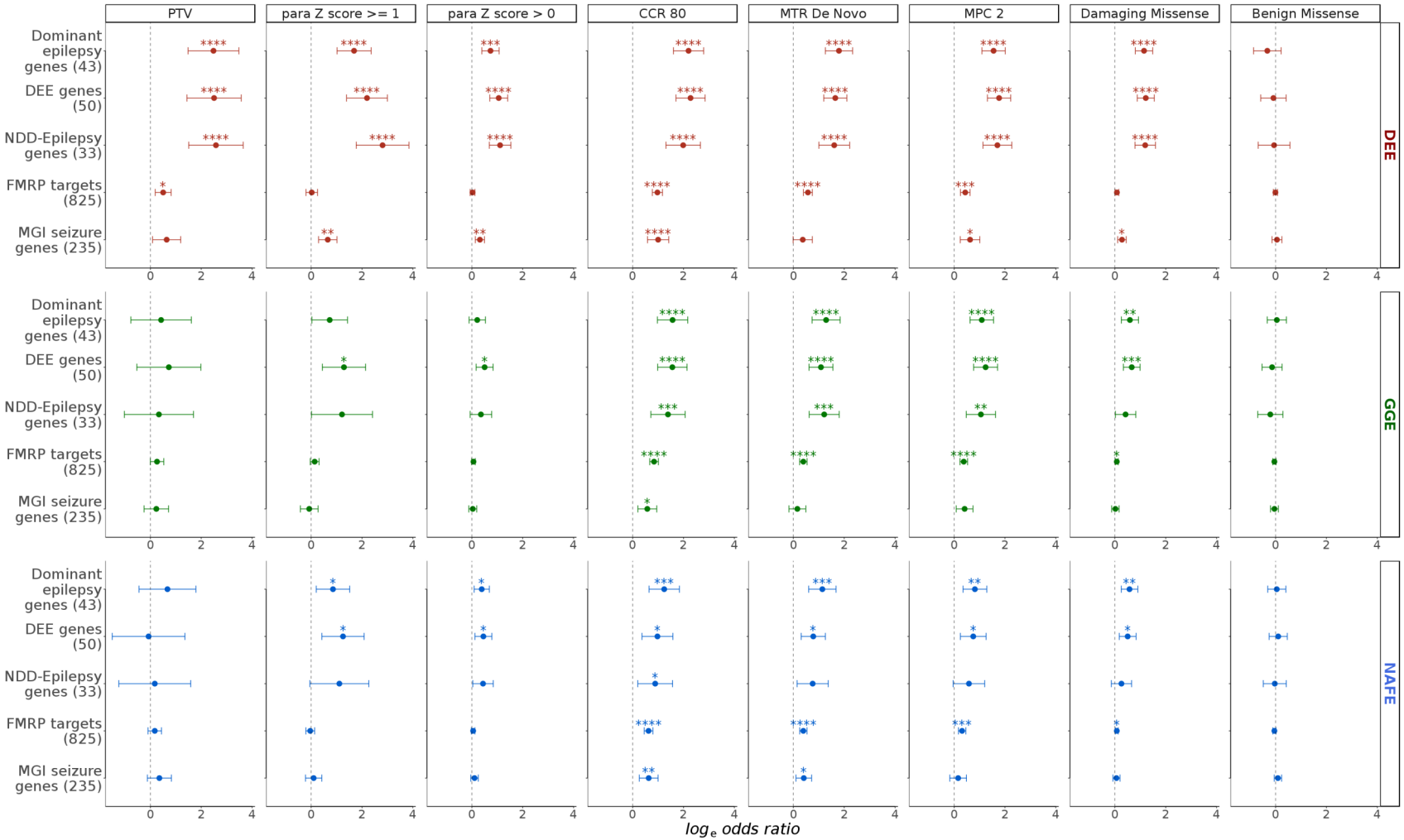


**Fig. 6: Burden of ultra-rare variants in groups of epilepsy-related known disease genes.** The burden in five gene-sets (y-axis; number of genes between brackets) in developmental and epileptic encephalopathies (DEE), genetic generalized epilepsies (GGE) and non-acquired focal epilepsies (NAFE) (horizontal panel) in selected variant conditions (vertical panels) is shown on the x-axis (log odd ratios from Likelihood Ratio Test; error bars indicate 95% confidence intervals). FDR corrected *p* values (not corrected for synonymous variants) are indicated with stars as follows: no star > 0.05, * < 0.05, ** < 0.005, *** < 0.005, **** < 0.0005. NDD-Epilepsy: neurodevelopmental disorders with epilepsy. FMPR: Fragile-X Mental Retardation Protein targets. MGI: Mouse Genome Informatics database.

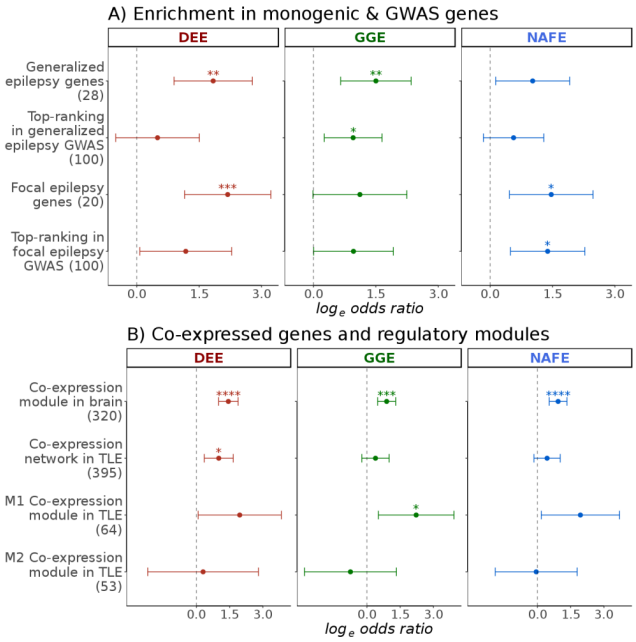


**Fig. 7: Risk elements in GWAS top-ranking genes and co-expression modules.** The burden of missense variants in highly constrained sites (log-odds on the x-axis; error bars indicate 95% confidence intervals) in developmental and epileptic encephalopathies (DEE), genetic generalized epilepsies (GGE) and non-acquired focal epilepsies (NAFE) is shown in monogenic epilepsy genes, top-ranking epilepsy GWAS genes, brain co-expression modules and TLE-related co-expression modules (y-axis; number of genes between brackets). FDR corrected p values (not corrected for synonymous variants) are indicated with stars as follows: no star > 0.05, * < 0.05, ** < 0.005, *** < 0.005, **** < 0.0005. (A) Generalized or focal epilepsies (presumed monogenic) as well as top-ranking 100 genes from GWAS of generalized and focal epilepsies are preferentially enriched for constrained missense variants (CCR 80) in respective phenotypic cohorts indicating a possible convergence between common and rare variants in GWAS genes. (B) Enrichment in co-expressed genes identified in post-mortem brain tissues of healthy individuals (module of 320 genes) or in brain tissues from TLE patients (network of 395 genes) as well as two sub-modules of this network (M1 and M2).

**Additional files:**

- Additional file 1: PDF file. Supplemental methods, tables and figures; affiliations of the Epi25 Collaborative members.
- Additional file 2: Excel xlsx file. Tables of gene-sets and genes in each gene-sets.
- Additional file 3: Excel xlsx file. Gene-set burden analysis results.
- Additional file 4: Excel xlsx file. Top-ranking genes per gene-set in the CCR 80 analysis.
- Additional file 5: Excel xlsx file. Secondary analysis results.