The role of rare genetic variants enrichment in epilepsies of presumed genetic etiology

Linnaeus Bundalian[1], Yin-Yuan Su[2], Siwei Chen[3,4], Akhil Velluva[5, 6], Anna Sophia Kirstein[7], Antje Garten[7], Saskia Biskup[8, 9], Florian Battke[8], Dennis Lal[3, 10], Henrike O. Heyne [4,12,13,14], Konrad Platzer[1], Chen-Ching[2] Lin, Johannes R. Lemke[1, 11], Diana Le Duc[1, 6]

[1] Institute of Human Genetics, University of Leipzig Medical Center, 4103 Leipzig, Germany

[2] Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei, Taiwan

[3] Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA

[4] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[5] Division of General Biochemistry, Rudolf Schönheimer Institute of Biochemistry, Medical Faculty, Leipzig University, 04103, Leipzig, Germany

[6] Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, 04103, Leipzig, Germany

[7] Pediatric Research Center, University Hospital for Children and Adolescents, Leipzig University, 04103, Leipzig, Germany

[8] CeGaT GmbH, 72076, Tuebingen, Germany

[9] Hertie-Institute for Clinical Brain Research, 72070, Tubingen, Germany

[10] Cologne Center for Genomics, University of Cologne, 50937 Cologne, Germany

[11] Center for Rare Diseases, University of Leipzig Medical Center, 4103 Leipzig, Germany

[12] Hasso-Plattner-Institut for Digital Engineering, University of Potsdam, Potsdam, Germany

[13] Hasso Plattner Institute at Mount Sinai, Mount Sinai School of Medicine, NY, US

[14] Institute for Molecular Medicine Finland: FIMM, University of Helsinki, Helsinki, Finland

1. Abstract

Previous studies suggested that severe epilepsies e.g., developmental and epileptic encephalopathies (DEE) are mainly caused by ultra-rare *de novo* genetic variants. For milder phenotypes, rare genetic variants could contribute to the phenotype. To determine the importance of rare variants for different epilepsy types, we analyzed a whole-exome sequencing cohort of 9,170 epilepsy-affected individuals and 8,436 controls. Here, we separately analyzed three different groups of epilepsies: severe DEEs, genetic generalized epilepsy (GGE), and non-acquired focal epilepsy (NAFE). We required qualifying rare variants (QRVs) to occur in controls at a minor allele frequency ≤ 1:1,000, to be predicted as deleterious (CADD≥20), and to have an odds ratio in epilepsy cases ≥2. We identified genes enriched with QRVs in DEE (n=21), NAFE (n=72), and GGE (n=32). Enrichment was strongest in NAFE and weakest in DEE. This suggests that rare variants may play a more important role for causality of NAFE than in DEE. Moreover, we found that QRV-carrying genes e.g., *HSGP2*, *FLNA* or *TNC* are involved in structuring the brain extracellular matrix. The present study confirms an involvement of rare variants for NAFE, while in DEE and GGE, the contribution of such variants appears more limited.

2. Introduction

Epilepsy is one of the most common neurological diseases worldwide, affecting almost 1% of the population in the United States [1]. Early pedigree studies showed a high genetic component and a heritability of up to 70% [2,3]. With the help of next generation sequencing there was a significant advance in gene discovery. Currently, hundreds of genes are established as monogenic causes for epilepsy [4], while recent studies have associated a few to polygenic causes [5]. Yet, the biggest leap in diagnostic yield happened mainly for the most severe type of epilepsies, developmental and epileptic encephalopathy (DEE) [6]. For this type of epilepsy, the

heritability or susceptibility are very low, since such diseases are often caused by deleterious *de novo* variants as the severely affected individuals usually do not reproduce.

The role of common [7–9] and ultra-rare *de novo* genetic variants [10,11] for epilepsy has been extensively researched. Epidemiological studies accounting for the similar prevalence across populations and the increased risk of individuals in more densely affected families, suggested that polygenic predisposition should have a dominant role over the monogenic etiology [12]. This has been addressed by genome wide association studies and polygenic risk scores [7–9] which identified mostly non-coding variants with individually small effects – median odds ratio (OR) generally lower than 1.3, but with a high aggregate effect explaining in part the missing heritability. Conversely, the ultra-rare *de novo* genetic variants have much larger effects on individual risk, but they make only a small contribution to the overall heritability in the population owing to their rarity [13,14].

Our understanding of the underlying genetic architecture leading to increased susceptibility to epilepsy due to a middle tier of variants that are rare (neither ultra-rare *de novo*, nor common, i.e. allele frequency of > 1%) is still very limited. Based on evolutionary theory, forces of negative natural selection will keep large-effect risk variants at much lower frequencies in the population, especially for a disorder like epilepsy which results in reduced fitness, i.e., reproduction. Analysis of rare variants' contribution to the disease could be a useful tool for a better understanding of the heritability and disease pathomechanism[15].

In this study, we focused only on rare variants (minor allele frequency ≤ 1:1,000) predicted to be deleterious and with an excess in cases (OR ≥ 2), which best reflects the effect size [15]. Finally, to understand the underlying pathomechanism we performed a combined analysis of the identified genes and their interacting partners aimed at identifying molecular pathways, which are potentially disrupted.

3. Materials and Methods

3.1 Cohort and data description

Genetic and phenotype information were obtained from the Epi25 collaborative [11] (http://epi-25.org/). Phenotyping procedures, case definitions, and ancestry of the participating individuals are reported in a previous Epi25 collaborative study [11]. To account for differences in ancestry and exome capture technologies among individuals the data has undergone previously described thorough quality check procedures and only individuals of European descent were analyzed [11]. Briefly, variant calling was performed with GATK [16] and only variants with a genotype quality > 20 were kept. Variants called heterozygous were required to have an allele frequency of 0.2–0.8. To control for kit enrichment artefacts, only variants where 80% of both Agilent and Illumina-sequenced samples show at least 10-× coverage were retained. Ancestry stratification had previously been ruled out using principal-component analyses to identify ancestral backgrounds and only individuals of European ancestry classified by Random Forest with 1,000 Genomes data were further analyzed. Annotation of variants was performed with Ensembl's Variant Effect Predictor [17] for human genome assembly GRCh37.

To understand differences in genetic susceptibility across different types of epilepsy, we analyzed 1,021 individuals with developmental and epileptic encephalopathy (DEE), 3,108 individuals with genetic generalized epilepsy (GGE), and 3,597 individuals with non-acquired focal epilepsy (NAFE). Each cohort was compared to 8,436 matched-ancestry, unrelated controls [15].

3.2 Qualifying rare variants and variant set enrichment analysis

We defined a variant as a qualifying rare variant (QRV) if it met following criteria:

- The variant is present in controls i.e. AF_CTRL ≥ 1;
- The minor allele frequency ≤ 1:1,000;

- The variant is predicted to be deleterious – CADD score ≥ 20 [18];
- OR ≥ 2, *p*-value ≤ 0.05 in epilepsy cases.

To account for gene length and different mutability across genes, we counted the total number of observed variants per gene and the number of variants with AF ≤ 0.001. Using these two parameters, we modeled a simple linear regression to characterize the relationship of the number of rare variants and total variants per gene. Based on the linear regression, we estimated the expected number of rare variants with OR ≥ 2 for each gene and compared to the observed. Only genes with an excess of rare variants were considered in the QRV filtering (Supplementary Fig. 1).

The susceptibility and risk burden for each gene were estimated by testing for enrichment of QRVs. To this end, we assigned an empirical enrichment score (ES) to each gene. The generation of the ES was inspired by the Significance and Functional Enrichment (SAFE) framework [19] and Gene Set Enrichment Analysis (GSEA) hence termed as Variant Set Enrichment Analysis (VSEA).

VSEA allows us to score the genes based on the number of QRVs they have across the general population and epilepsy patients. We defined following lists of variants:

- ranked variant list (L) – list of QRVs per epilepsy type ordered (descending) by their corresponding OR;
- background set (S) – list of variants grouped into synonymous or nonsynonymous variants per epilepsy type.

The enrichment analysis is designed to check if variants in S are randomly distributed throughout L, or skewed to a side where OR is higher or lower. With this method, we can also identify the variants in each gene which contribute most to the enrichment – referred to as leading edge.

The enrichment score (ES) assigned to the genes is calculated using the maximum deviation observed among cumulative ranked sum [19] of both ***hits*** and ***miss*** across the variants (Supplementary Fig. 2) normalized by the number of observed ***hits and miss***. The vector of either ***hits*** or ***miss*** can be represented by vector *y* and ES as the maximum difference between the two vectors, defined as follow :

$$y = \{y_1 \dots y_n\} \text{ where } y[i] = \sum_{k=1}^{i} x[k] \text{ for } i \in \{1, \dots n\}$$

$$ES = max\left\{\frac{y_{hits}}{\#\ of\ hit} - \frac{y_{miss}}{\#\ of\ misses}\right\}$$

A high ES of a gene indicates an enrichment of variants having higher OR in comparison to the variants found in the lower rank. After calculating the ES, to determine the significance we performed n = 1,000 permutations where a set of variants having the length *n(L)* were randomly selected from set *S.* This was used to calculate an empirical distribution of ES. The number of times the empirical ES exceeds the observed ES was counted and divided by the number of permutations (n) to calculate the *p*-values (pval).

$$pval = \frac{1}{n}\sum_{p=1}^{n}\{ES_{perm} \geq ES_{obs}\}$$

For multiple testing correction, we determined FDR using the ratio of the Normalized Enrichment Score (NES) – observed and permuted ES. NES is the ES divided by the expected ES i.e., average ES from the permuted values.

$$NES_{obs} = \left\{\frac{ES_{obs}}{\frac{\sum_{p=1}^{n} ES_{perm}}{n}}\right\}; NES_{perm} = \left\{\frac{ES_{perm}}{\frac{\sum_{p=1}^{n} ES_{perm}}{n}}\right\}$$

$$FDR = \frac{NES_{obs}}{NES_{perm}}$$

3.3 Functional module analysis and protein-protein interactions

To test in which functional pathway the genes identified that bear an excess of QRVs play a role, we performed enrichment analysis across multiple gene sets including but not limited to Gene Ontology (GO), Allen Brain Atlas, Reactome, KEGG pathways.

3.3.1 Allen Brain Atlas Enrichment

To identify gene sets that are specific to certain brain areas and/or developmental stages we used the Allen Human Brain Atlas. This resource delivers information about gene expression levels in various parts of the human brain during the course of brain development [20,21]. We used the R package *ABAEnrichment* to test whether genes with excess QRVs show significant enrichment in specific brain regions or brain developmental stages [22]. The package integrates human brain expression datasets provided by the Allen Brain Atlas in both the prenatal [38] and adult stage [39]. The expression data is analyzed over 47 brain regions and 20 age time points. The gene expression is evaluated during development from prenatal stage to adult. If the change is high in a specific region, the gene is annotated to that region and the score mirrors the deviation from prenatal to adult stage.

3.3.2 Overrepresentation Analysis (ORA)

The gene lists were also subjected to functional class analysis using ORA with MsigDb Reactome database for gene set collections C2 (Reactome) and C5 (Gene Ontology) [23,24] and GOFuncR [25]. The method uses a hypergeometric test [26] to assess the probability of observing at least *k* genes from the list across the pathway database [23,27]:

$$P(X \geq k \mid n; N; K) = \sum_{i=k}^{n} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}}$$

For the customized gene set ORA, we used the datasets from MSigDB – C5 for sets associated with ion channels, neurotransmitter, glutamatergic and GABAergic signaling,

nervous system development, and synaptic functions [23,24]. Genes were considered to be highly expressed in brain if the normalized transcript per millions (nTPM) ≥1; nTPM corresponds to the normalized transcript expression values from samples found in Human Protein Atlas (HPA) and GTEx using the "Trimmed Mean of M values" method [28]. For GoFuncR overrepresentation, we used C5 (Gene Ontology) of MSigDB for restricting the background genes to those which are found expressed in the brain [23,24].

3.3.3 Analysis of distance in the human protein interacting network (PIN)

To elucidate the putative roles of the genes significantly enriched with QRVs in epilepsy, we investigated the distance between them and the known epilepsy-related genes in the human PIN. The source of the human PIN data is InBio Map [29] and epilepsy-related genes were acquired from the consolidated list of epi-25.org. The distance is defined as the shortest path length between gene u and v in the human PIN. All paired shortest path lengths between genes are calculated by the Dijkstra algorithm. We identified which genes with QRVs are significantly closer located to the known epilepsy-related genes within the PIN, compared to the distance of all other genes.

4. Results

4.1 Enrichment of QRVs in three types of epilepsies

We tested the burden of QRVs per gene in each epilepsy group: DEE, NAFE, and GGE. The only significant gene which was present across all types of epilepsies was *HSPG2* (p-value = 0.0001, from 10,000 random samplings). *HSPG2* (Heparan Sulfate Proteoglycan 2) encodes the perlecan protein that belongs to the glycosaminoglycans family, which are major components of the brain extracellular matrix (ECM). Although the gene was common to all epilepsy groups, there were different variants in this gene, which contributed to the increased enrichment score of the gene for each group (Supplementary Table 1).

Further, consistent with the presumed *de novo* pathogenic variants occurrence and a highly penetrant phenotype, DEE showed the lowest number of genes enriched with QRVs (Fig. 1A); this supports DEE's mainly monogenic origin. In other words, for DEE, the most severe of the epilepsy phenotypes, it is less likely that variants present in controls contribute as causal risk factors. The largest number of genes with a significantly high enrichment score was retrieved for NAFE (Fig. 1A, Supplementary Table 1). This result contrasts the previous findings on ultra-rare *de novo* genetic variants, which show mainly no significant burden in the NAFE individuals [11] and may suggest that rare variants also present in controls could contribute more to the NAFE pathophysiology compared to highly damaging *de novo* variation.

To gain insight into the functionality of the genes, we classified them into 6 categories, which are considered to play a role in the epileptic pathomechanism: modulators of synaptic functions, neurotransmitter regulators, nervous system development, ion channels and their interacting partners, modulators of glutamatergic signaling, and genes with high expression in brain (Fig. 1 B, C, Table 1). Our results showed for DEE and NAFE a significant enrichment especially for genes highly expressed in brain. GGE showed an enrichment for genes annotated to nervous system development. In NAFE we identified genes enriched for ion channels and their interactors (FDR = 0.004) (Fig. 1B). We showed that variants in these genes generally had median ORs > 5 (Fig. 1C), suggesting they may have a large effect size and could contribute to the underlying pathomechanism.

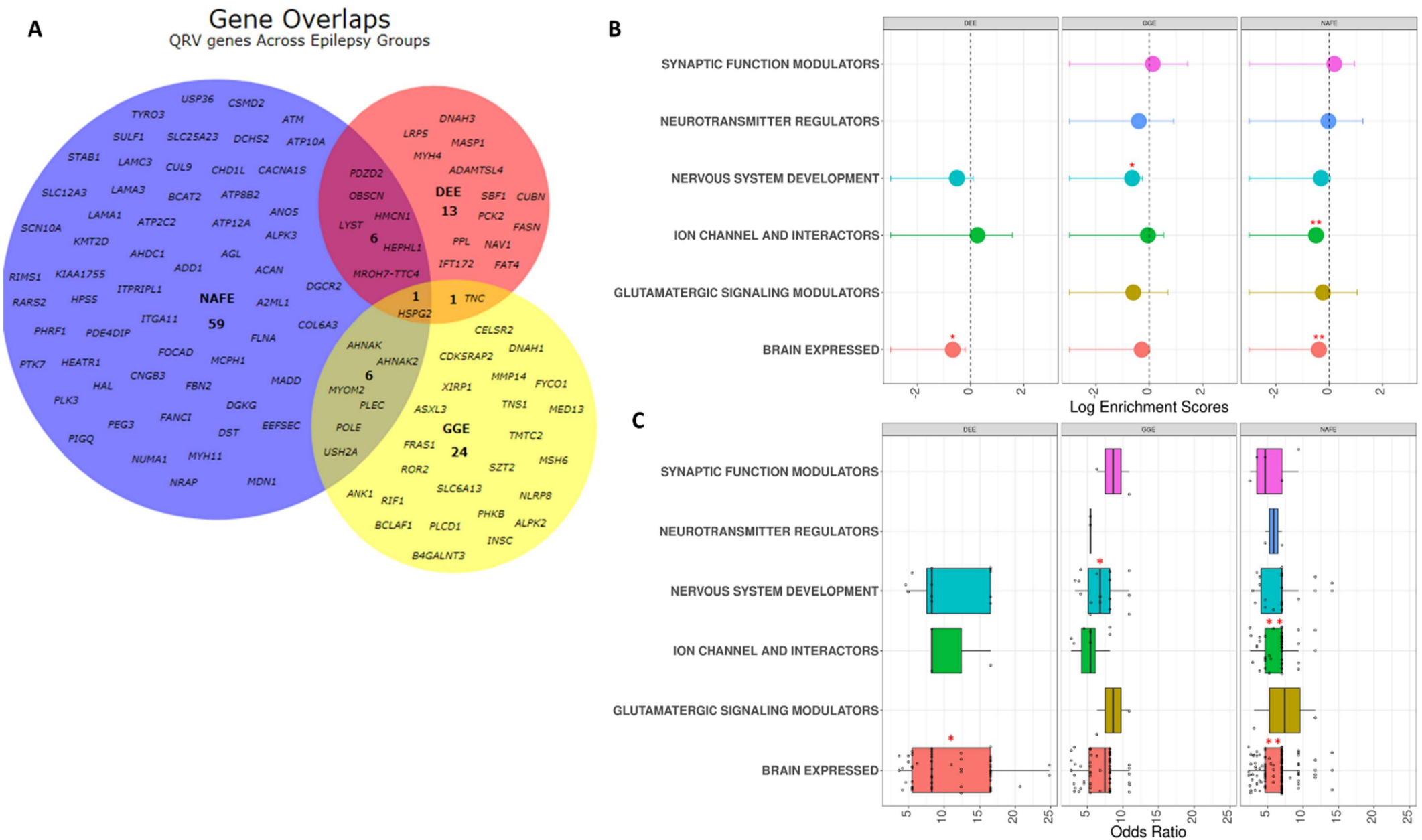


Fig. 1. A. QRV enriched genes across different epilepsy groups. For each epilepsy group, a set of genes were identified to be enriched with QRVs (FDR ≤ 0.05, Z score ≥ 2). The number of genes found for NAFE is greater than for the two other groups implying that rare variants in this particular group could have a larger contribution to the etiology compared to DEE and GGE. **B. Overrepresentation of QRV enriched genes**. The graph shows the log enrichment scores and confidence interval = 0.95 for the overrepresentation of the QRV enriched genes. The genes enriched with QRVs were found to be overrepresented across highly expressed genes in the brain (DEE, NAFE), nervous system development (GGE), and ion channel and interactors (NAFE) – suggesting a role in the brain and brain-related processes. **C. Odds ratio distribution of QRV enriched genes across different gene sets**. The graph shows the odds ratio distribution of the variants per genes across different gene sets. (* FDR ≤ 0.05, ** FDR ≤ 0.01)

Table 1. Overrepresented genes across gene sets annotated to brain processes and molecular functions. The table contains the QRVs-enriched genes, which are found to be enriched in gene sets derived from molecular functions in GO database and brain gene expression level from Protein Atlas and GTEx [28]. For each gene set we selected processes associated with ion channels, GABAergic and glutamatergic related pathways, synaptic functions, neurotransmission, nervous system development and genes with nTPM ≥ 1 in brain based on Protein Atlas and GTEx expression data [28].

GENE SET	DEE	NAFE	GGE
ION CHANNEL AND INTERACTORS	*HEPHL1*	*ADD1, AHNAK, ANO5, ATM, ATP10A, ATP12A, ATP2C2, ATP8B2, CACNA1S, CNGB3, FLNA, HEPHL1, RIMS1, SCN10A, SLC12A3, SLC25A23*	*AHNAK, SLC6A13, TMTC2*
NEUROTRANSMITTER REGULATORS		*RIMS1*	*SLC6A13*
NERVOUS SYSTEM DEVELOPMENT	*FAT4, HSPG2, IFT172*	*ACAN, ATM, HSPG2, LAMC3, MCPH1, PLEC, TYRO3*	*CDK5RAP2, CELSR2, HSPG2, PLEC, ROR2, SZT2*
SYNAPTIC FUNCTION MODULATORS		*FLNA, RIMS1*	*ROR2*
GLUTAMATERGIC SIGNALING MODULATORS		*HAL*	*ROR2*
BRAIN EXPRESSED	*ADAMTSL4, CUBN, FASN, FAT4, HMCN1, HSPG2, IFT172, LRP5, LYST, MASP1, NAV1, OBSCN, PCK2, PDZD2, PPL, SBF1, TNC*	*A2ML1, ACAN, ADD1, AGL, AHDC1, AHNAK, AHNAK2, ALPK3, ANO5, ATM, ATP10A, ATP8B2, BCAT2, CHD1L, COL6A3, CSMD2, CUL9, DGCR2, DGKG, DST, FANCI, FLNA, FOCAD, HEATR1, HMCN1, HPS5, HSPG2, KIAA1755, KMT2D, LAMA1, LAMA3, LAMC3, LYST, MADD, MCPH1, MDN1, MYH11, NUMA1, OBSCN, PDE4DIP, PDZD2, PEG3, PHRF1, PIGQ, PLEC, PLK3, PTK7, RARS2, RIMS1, SLC25A23, STAB1, SULF1, TYRO3, USP36*	*AHNAK, AHNAK2, ANK1, BCLAF1, CDK5RAP2, CELSR2, FRAS1, FYCO1, HSPG2, MED13, MMP14, MSH6, PHKB, PLCD1, PLEC, RIF1, SLC6A13, SZT2, TMTC2, TNC, TNS1*

4.2 QRVs collapsing captures genes encoding for ion channels or their interactors in NAFE

Ion channels play a major role in genetic epilepsies [30,31]. Although, since the beginning of the epilepsy-related gene discoveries, many other gene classes and biological pathways have

been revealed to play a role, it is still a significant proportion (~ 25 %) of the epilepsy genes known to date that encode for ion channels [30]. Since we identified the ion channel-related molecular function pathway to be enriched in NAFE (Fig. 1B), we performed ORA based on KEGG and Reactome [32]. For KEGG, the ion channel pathway is annotated only with respect to drug development, and thus, we could not identify an over-representation based on the QRVs-enriched genes in NAFE. Using the Reactome annotation we identified the ion channel transport category to be significantly enriched. NAFE-related genes annotated to the Reactome category are: *ATP2C2*, *ATP12A*, *ATP8B2*, *ATP10A*, and *ANO5*. While, the first four genes encode for ATPases involved in ions transport, like Ca^{2+} or H^{+}/K^{+}, *ANO5* encodes for an anoctamin, which belongs to a protein family of Ca^{2+} activated chlorine channels and phospholipid scramblases [33]. A founder mutation in *ANO5* has been implicated in muscular dystrophy [33]. Despite the high expression in brain and the controversial muscle phenotype in the mouse knockout models [34,35], its function in the brain remains unknown.

Using the gene ontology annotation, we identified a few ion channel genes and multiple genes interacting with ion channels, which showed QRVs enrichment in NAFE (Table 1), e.g. *CACNA1S* and *SCN10A*. Another identified gene is *ADD1* coding for adducin. Although adducin is primarily responsible for the assembly of spectrin-actin that provides functional support to the cytoskeleton, the gene ontology also annotates the gene to ion transport and synaptic functions. Variants in *ADD1* have also been recently identified in intellectual disability, corpus callosum dysgenesis, and ventriculomegaly in humans [36]. Similarly, *ATM*, another gene with QRVs for NAFE, has been recently shown to be involved in hippocampal and cortical development, as well as synaptic functions [37].

Based on our analysis, we did not retrieve genes encoding for ion channels that have already been associated with monogenic epilepsy. Those genes generally bear ultra-rare variants that do not occur in controls, which does not comply with our definition of QRVs.

Some of the genes we identified to be significantly enriched in NAFE have already been associated with mendelian disorders e.g., *HSPG2* and *FLNA*, yet their phenotype does not appear to be severe or highly penetrant which meets our hypothesis that variants in these genes could confer an increased risk for epilepsy.

4.3 Involvement in brain development

We asked whether the genes with QRVs are involved in brain development and in which brain regions they are most relevant. Patterns of gene expression can be very informative in respect to the importance of a gene during development. Using the ABAEnrichment, we tested whether the genes identified in the different types of epilepsy play a role during development in any brain regions. From the results, we identified genes from DEE to show the highest involvement during development with significant enrichment over 13 brain regions (Fig. 2). For GGE only the cerebellar cortex showed a signal (family wise error rate, FWER = 0.045), while for NAFE we identified a signal in this region and two additional ones in the striatum and inferolateral temporal cortex. When we inquired the developmental scores assigned based on expression changes between prenatal and adult stage, we identified genes encoding for ECM proteins (*LAMA1*, *FBN2*, *COL6A3*) to contribute to the enrichment in the different brain regions for NAFE (Fig. 3). Similarly, *TNC*, which encodes for the ECM protein tenascin C showed a high developmental score contributing to the brain regions enrichment in both DEE and GGE. For DEE, *FAT4*, a gene encoding for a protocadherin, a calcium-dependent cell adhesion protein, showed the highest developmental score (Fig. 3). *FAT4*, which was previously related to epilepsy [40], plays a role in the maintenance of planar cell polarity as well as in neuroprogenitor proliferation [41]. For GGE, *ROR2* is the leading gene in respect to the developmental score. *ROR2* encodes a tyrosine-protein kinase transmembrane receptor also known as the neurotrophic tyrosine kinase, receptor-related 2, which also appears to play a role in the maintenance of neuroprogenitor cells in the developing neocortex [42]. Based on the

observed functions for the genes with the highest developmental scores, we were further prompted to perform pathway analyses and understand in which molecular processes genes with QRVs are involved.

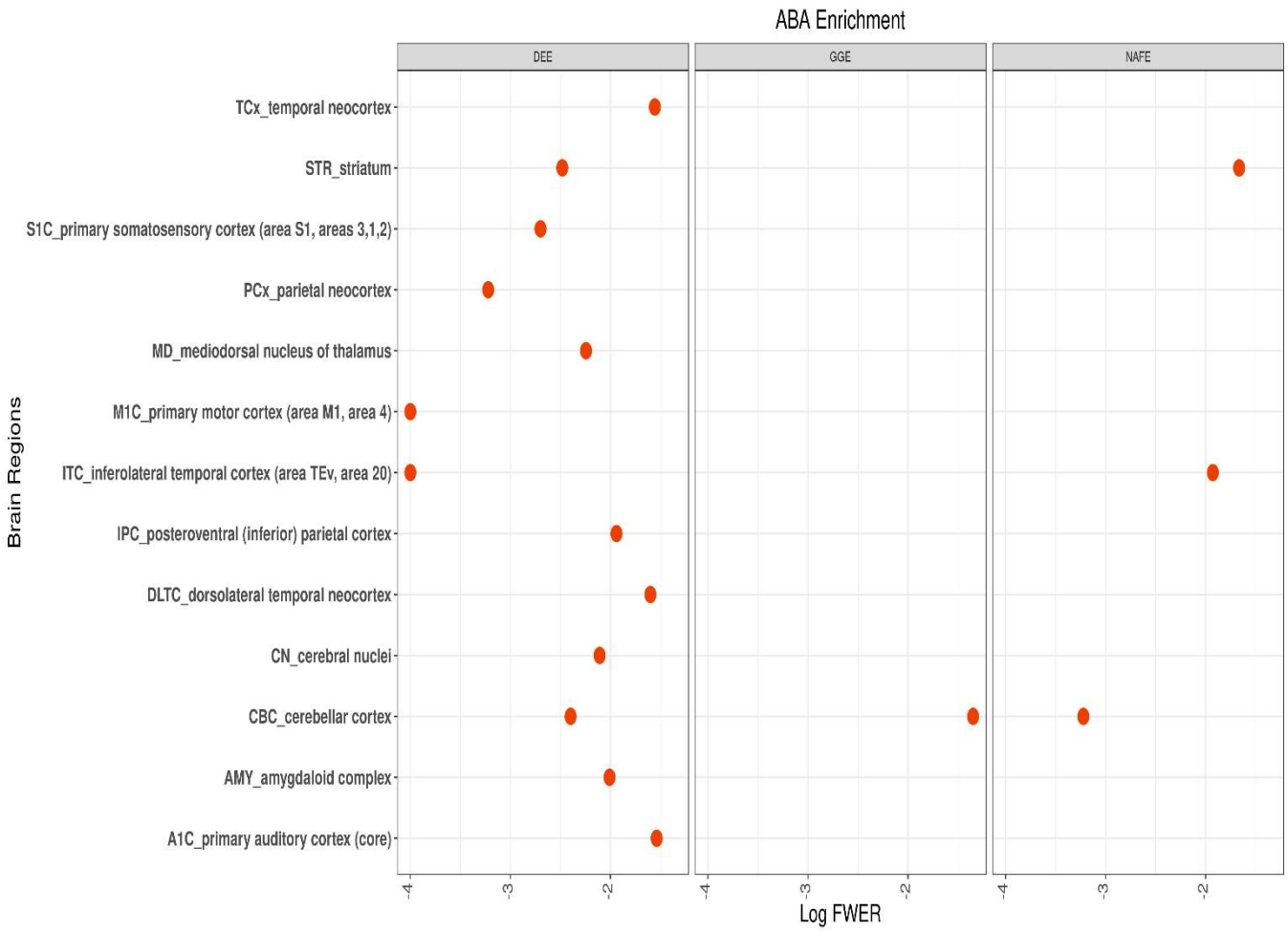


Fig. 2. ABA Enrichment of QRV enriched genes across all epilepsy types. Brain regions with enriched QRV genes with respect to their developmental score and the log transformed FWER (family wise error rate), which is the measure of the significance of the enrichment.

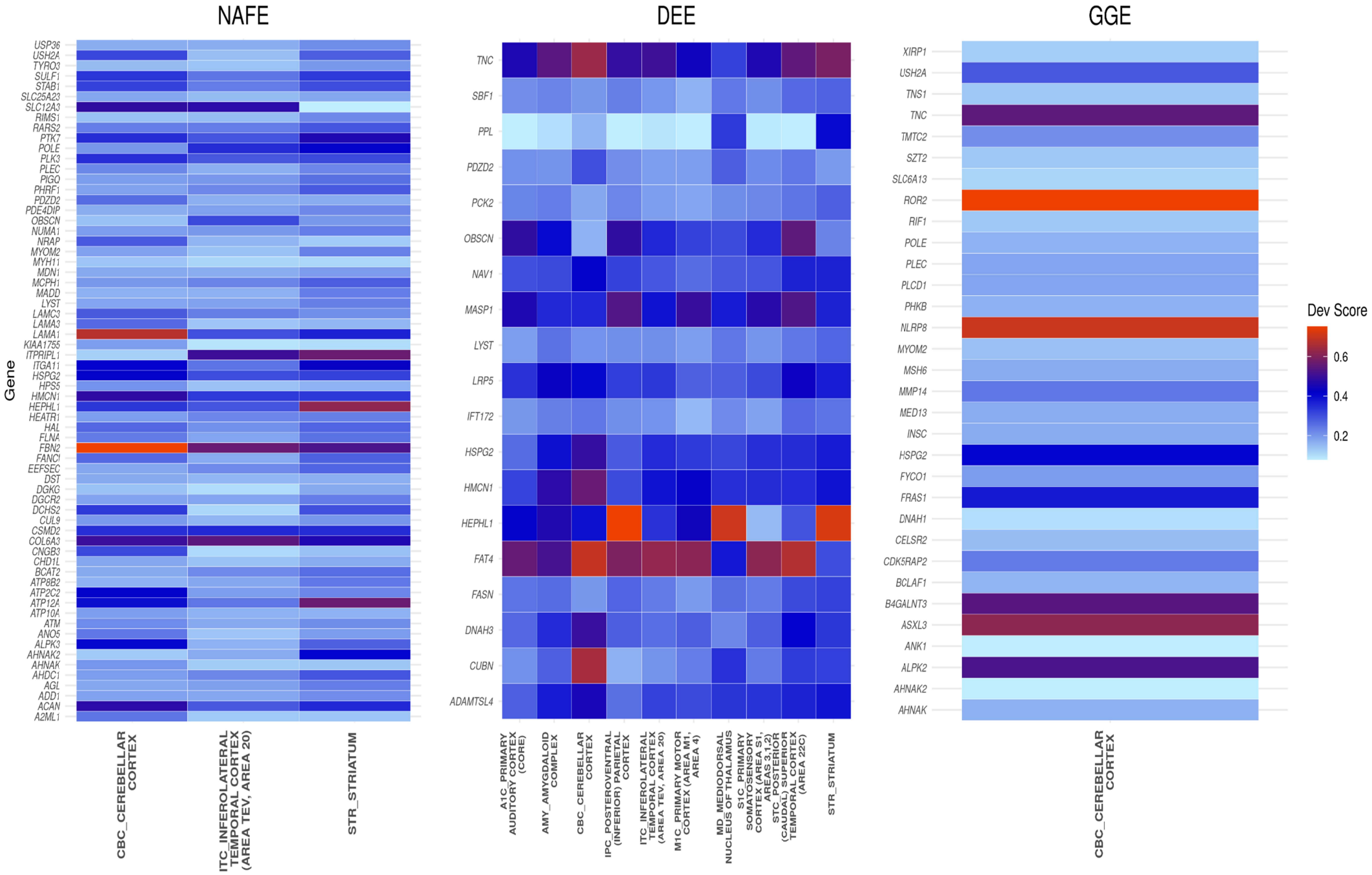


Fig. 3. ABA Developmental score of QRV enriched genes. Heatmap with developmental scores of QRV genes significantly associated with the different brain regions during development. The listed regions were identified to be overrepresented with QRV enriched genes.

4.4 Pathway and network analyses

We performed pathway enrichment analyses to determine whether the identified genes cluster within specific functions. For DEE we could not identify any significant categories after multiple testing correction. NAFE showed enrichment of many GO categories, that clustered mainly within the extracellular matrix (ECM) or cell adhesion (Fig. 4A). For GGE, there were only 3 significant categories, all related to cellular junctions or adhesion (Supplementary Table 2).

To further test how the identified genes with QRVs are connected to already known epilepsy genes, we determined the distance between the identified gene and the epilepsy genes from Epi25 (http://epi-25.org/). To this end, we assessed within the PPI network how many nodes represented by protein interacting partners lie between the identified gene and any known epilepsy gene. While for NAFE, the mean distance between QRV genes and epilepsy genes was at 2.57, significantly smaller than the mean distance of between the non-QRV genes and epilepsy genes (2.66, p-value < 0.001), the effect size is too small to conclude that overall, the identified genes with QRVs are closer to known epilepsy genes. We thus, further determined the genes with a shorter distance to known epilepsy genes within the PIN (Fig. 4B). For GGE and DEE the overall mean distance did not reach significance and only 3 and 2 genes, respectively, showed a significantly closer distance to epilepsy genes.

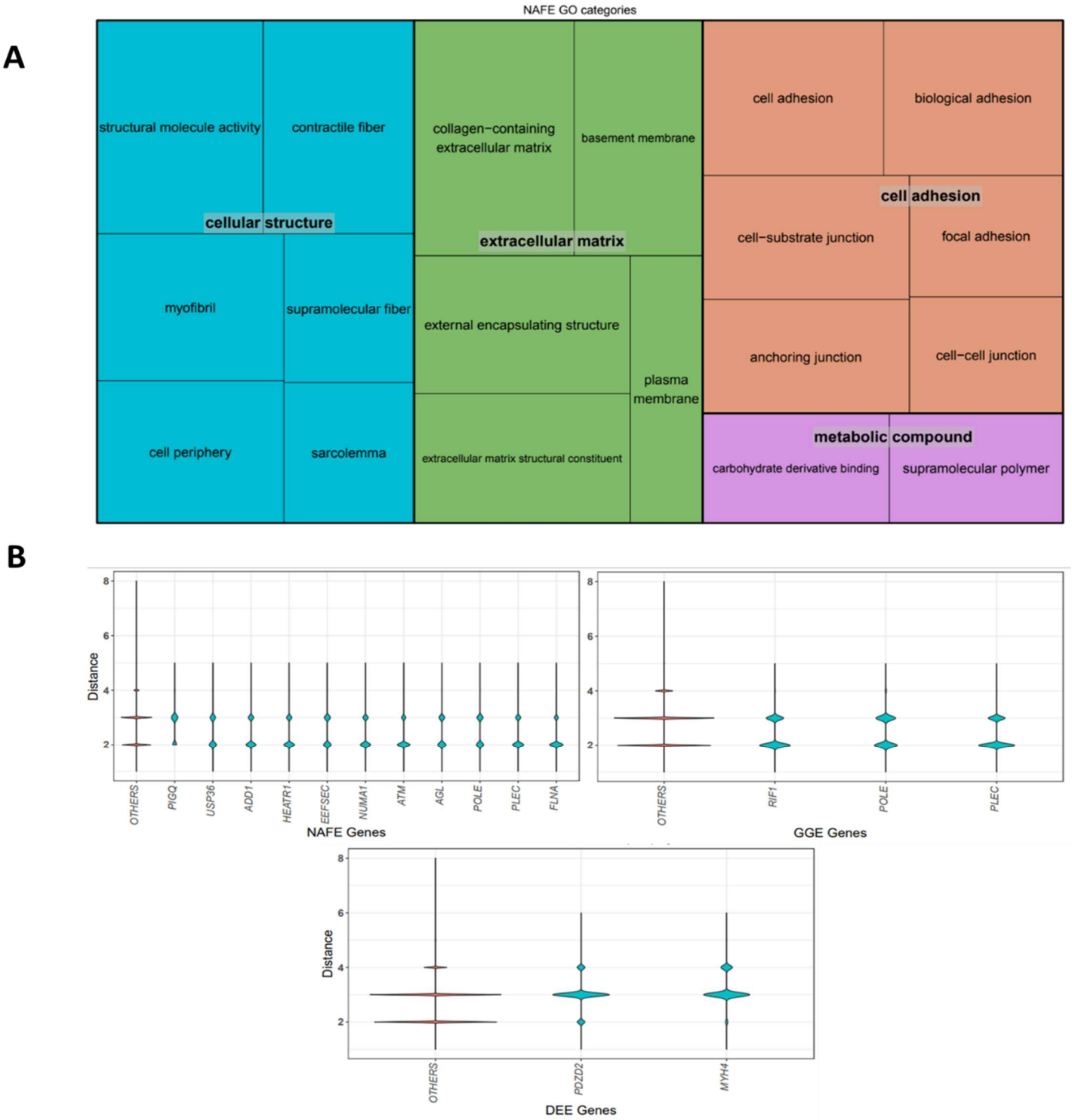


Fig. 4. A. Overrepresentation of QRV genes from NAFE. The GO terms overrepresented by the genes found to be enriched in NAFE clusters into a more general category to represent their functional class. The overrepresentation analysis was done using GoFuncR with genes expressed in the brain as background. **B. Genes and their distance to known epilepsy genes.**

The distance (y-axis) of the QRV genes (x-axis) to known epilepsy genes was determined by the length of the path linking the smallest number of edges between QRV genes and epilepsy genes i.e. the length of the shortest path. We show QRV genes with significantly shorter distance to known epilepsy genes compared to the distance of non-QRV genes. A larger distribution assigned to the lower distance in the violin plot implies a higher number of genes with short distance to known epilepsy genes in the PIN. A short distance to the epilepsy genes could suggest a contribution to epilepsy pathomechanism.

5. Discussion

Unlike common and *de novo* ultra-rare variants, rare variants in epilepsy have not been researched extensively. Based on the 'Common Disease, Rare Variant' hypothesis, multiple rare sequence variants, with relatively high penetrance, confer an increased genetic susceptibility to a common disease [43]. While the identification of such variants is paramount for understanding their role, their detection is statistically more challenging because they are present at low frequencies in the general population. To understand how rare variants contribute to the etiology and pathomechanism of epilepsy, we applied a method to estimate the risk burden of a gene based on the enrichment of rare variants posing a relatively high risk in the Epi25 cohort i.e., OR $\geq$ 2. Using the proportion of QRVs in cases and controls, we calculated a score for each gene. A high score of a gene implies a higher disease probability due to the presence of the rare variants with higher ORs in cases. By this approach we identified sets of QRV enriched genes – DEE (n = 21), NAFE (n = 72), and GGE (n = 32) (Fig. 1). In a previous study that analyzed the same patient cohort in respect to the enrichment of deleterious ultra-rare variants, DEE and GGE individuals had significantly more such variants compared to those diagnosed with NAFE [11]. While for our study on the enrichment of rare variants, we identified the lowest number of genes for DEE and the highest for NAFE. Our results and the results of the previous study [11] could suggest that, while for DEE the pathomechanism relies on highly

deleterious and penetrant variants, NAFE may result from an enrichment of more frequent and less penetrant rare variants, identified also in controls, albeit at lower frequency than in cases.

To further analyze the gene set and its relevance to specific molecular pathways, we performed a GO enrichment analysis. This revealed a significant overrepresentation of the NAFE genes across ECM and structural related pathways (Fig. 4A). The ECM is known to play a role in epileptogenesis since it is involved in the establishment of neural plasticity, i.e. cell-cell connections and signaling [44]. Moreover, changes in ECM have been directly implicated in the pathophysiology of temporal lobe epilepsy [45,46], the most common form of focal epilepsies/NAFE. Interestingly, *HSPG2*, the only gene found to be enriched across all types of epilepsy in the study, encodes for perlecan, an important member of brain ECM. Little is known about *HSPG2* and its association to epilepsy, but some studies have revealed its role in acetylcholinesterase clustering at the synapse, which has the capability to interfere in synaptic transmission [47]. Perlecan is, however, ubiquitously expressed and pathogenic variants have been implicated in the Schwartz–Jampel syndrome type I, a rare autosomal recessive disease with cardinal symptoms consisting of skeletal dysplasia and neuromuscular hyperactivity [48]. Some of the affected individuals also show impaired neurologic development, consistent with perlecan's neuroprotective effect and its involvement in neurogenesis and normalization of neocortical excitability after insult events [49]. In further support of our finding, previous studies have also considered *HSPG2* to be an epilepsy-associated gene, although the underlying mechanism is still not clear [40].

While among the identified genes, there was an enrichment of genes with high brain expression, both in the DEE and NAFE groups, for NAFE we additionally identified variants in ion channels and their interactors to play a role (Fig. 1 B, C). *CACNA1S* is one of the genes with an excess of QRVs in NAFE individuals. The gene is lowly expressed in the brain and highly expressed in the muscles being implicated in the hypokalemic periodic paralysis.

However, we observed outliers in respect to brain expression (Supplementary Fig. 3) [50], suggesting that $Ca_v1.1$, the L-type voltage gated calcium channel encoded by *CACNA1S*, could play a role in the calcium influx in response to large depolarizing shifts in membrane potential for some individuals. In support of the variability of the *CACNA1S*' brain involvement, rare variants in this gene have been associated with schizophrenia [51]. Similarly, *SCN10A* is lowly expressed in the brain, but shows an enrichment signal in our dataset. Biallelic variants in this gene have been potentially linked to epilepsy-related phenotypes [52].

In an additional analysis, to understand whether the genes we identified are closer in the PIN to already known epilepsy genes, we calculated the PPI distance (Fig. 4B). Most of the genes with significantly shorter PPI paths connect them to epilepsy genes for NAFE. A gene with significantly shorter distance to known epilepsy genes is *PDZD2*, enriched for both NAFE (FDR = 0.005) and DEE (FDR = 0.006). This gene can be found expressed mainly in the basal ganglia and cerebral cortex with high specificity among oligodendrocytes precursor cells, and excitatory neurons [28,53]. *PDZD2* also contributes to the functional expression of Nav1.8 ion channel, which is encoded by *SCN10A,* a QRV enriched gene in NAFE [54]. Another gene with shorter distance to known epilepsy genes, *FLNA*, has itself been associated with epilepsy and seizure disorders [5]. *FLNA* is also known to interact with *HCN1* channels during neuronal excitability modulation in the mature brain [55]. Additionally, *FLNA* also controls ECM remodeling by regulating metalloproteinase activity and hence ECM degradation [56]. Based on the enrichment of ECM genes, we suggest that especially for NAFE, genetic variants which may impact ECM morphology could lead to imbalance in excitatory and inhibitory signals [57] and hence underlie epileptogenesis.

Consistent with the assumed pathomechanisms, we identify a significant enrichment of QRVs in genes related to brain development only in DEE (Fig. 2) and GGE (Fig. 1). Interestingly, most genes with high developmental scores are assigned to DEE (Fig. 3),

suggesting that rare variants in these genes may contribute to the disease development. *TNC*, which encodes an ECM protein, was identified in both DEE and GGE (Fig. 1). It controls neurite growth and axon guidance [58] and it is highly active during early brain development, which is mirrored by a high developmental score (Fig. 3). Intriguingly, *TNC* is higher expressed by both neurons and glias after seizures, which can lead to ECM remodeling and induce additional epileptic events [58,59]. For NAFE, genes with high developmental score cluster in the inferolateral temporal cortex (Fig. 2), a signal that may be triggered by a high number of temporal lobe epilepsy cases within the NAFE cohort.

Aside from the established list of epilepsy associated genes from the Epi25 cohort, there are a number of curated lists for epilepsy genes, for instance the SAGAS database, containing candidate genes with possible polygenic and monogenic causal tendencies [5] and Genes4Epilepsy (https://github.com/bahlolab/Genes4Epilepsy). We identified a significant overlap of the QRVs genes from our study with both SAGAS (DEE: pval = 0.001, NAFE: pval = 1.30e-8, GGE: pval = 0.02) and Genes4Epilepsy (NAFE: pval = 0.001), which lends additional support to our findings.

A previous study of the Epi25 Collaborative suggested that clinical presentations of GGE and NAFE are influenced by common and rare variants, as opposed to DEE which is mainly caused by *de novo* ultra-rare highly deleterious variants [11]. Here, we focused on rare variants, present in controls, but at higher OR in epilepsy patients. Our results support the hypothesis that rare variants could be important in NAFE pathomechanism. Moreover, ECM appears to play a central role in NAFE. For DEE we retrieve genes that have high expression during development, which meets the expected pathomechanism; however, the number of identified genes is rather low. Based on the genes identified for GGE we cannot infer which pathways play an important role in the pathomechanism. It is possible that either enlarged patient cohorts or a focus on common variants will shed more light on GGE.

6. Bibliography

1. Zack, M. M. & Kobau, R. *National and State Estimates of the Numbers of Adults and Children with Active Epilepsy — United States, 2015* **66**, 821–825 (2017).

2. Chen, T., Giri, M., *et al. Neuropsychiatric Disease and Treatment* vol. 13 1841–1859 (2017).

3. Annegers, J. F., Hauser, W. A., *et al. The risks of seizure disorders among relatives of patients with childhood onset epilepsy* **32**, 174–179 (1982).

4. Perucca, P., Bahlo, M., *et al. Annual Review of Genomics and Human Genetics* vol. 21 205–230 (2020).

5. Gracie, L., Rostami-Hochaghan, D., *et al. The Seizure-Associated Genes Across Species (SAGAS) database offers insights into epilepsy genes, pathways and treatments* **63**, 2403–2412 (2022).

6. Guerrini, R., Balestrini, S., *et al. Neurology* vol. 97 817–831 (2021).

7. Campbell, C., Leu, C., *et al. The role of common genetic variation in presumed monogenic epilepsies* **81**, 104098 (2022).

8. Leu, C., Stevelink, R., *et al. Polygenic burden in focal and generalized epilepsies* **142**, 3473–3481 (2019).

9. Abou-Khalil, B., Auce, P., *et al. Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies* **9**, (2018).

10. Allen, A. S., Bellows, S. T., *et al. Ultra-rare genetic variation in common epilepsies: a case-control sequencing study* **16**, 135–143 (2017).

11. Feng, Y. C. A., Howrigan, D. P., *et al. Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals* **105**, 267–282 (2019).

12. Peljto, A. L., Barker-Cummings, C., *et al. Familial risk of epilepsy: A population-based study* **137**, 795–805 (2014).

13. Hernandez, R. D., Uricchio, L. H., *et al. Ultrarare variants drive substantial cis heritability of human gene expression* **51**, 1349–1355 (2019).

14. Manolio, T. A., Collins, F. S., *et al. Nature* vol. 461 747–753 (2009).

15. Zuk, O., Schaffner, S. F., *et al. Searching for missing heritability: Designing rare variant association studies* **111**, E455–E464 (2014).

16. Van der Auwera, G. A. & O'Connor, B. D. (O'Reilly Media, 2020).

17. McLaren, W., Gil, L., *et al. The Ensembl Variant Effect Predictor* **17**, (2016).

18. Kircher, M., Witten, D. M., *et al. A general framework for estimating the relative pathogenicity of human genetic variants* **46**, 310–315 (2014).

19. Barry, W. T., Nobel, A. B., *et al. Significance analysis of functional categories in gene expression studies: A structured permutation approach* **21**, 1943–1949 (2005).

20. Miller, J. A., Ding, S. L., *et al. Transcriptional landscape of the prenatal human brain* **508**, 199–206 (2014).

21. Hawrylycz, M. J., Lein, E. S., *et al. An anatomically comprehensive atlas of the adult human brain transcriptome* **489**, 391–399 (2012).

22. Grote, S., Prüfer, K., *et al. ABAEnrichment: An R package to test for gene set expression enrichment in the adult and developing human brain* **32**, 3201–3203 (2016).

23. Liberzon, A., Subramanian, A., *et al. Molecular signatures database (MSigDB) 3.0* **27**, 1739–1740 (2011).

24. Subramanian, A., Tamayo, P., *et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles* **102**, 15545–15550 (2005).

25. Grote, S. *Title Gene ontology enrichment using FUNC* **R package**, v1.18.0 (2022).

26. Draghici, S., Khatri, P., *et al. A systems biology approach for pathway level analysis* **17**, 1537–1545 (2007).

27. Wieder, C., Frainay, C., *et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis* **17**, (2021).

28. Sjöstedt, E., Zhong, W., *et al. An atlas of the protein-coding genes in the human, pig, and mouse brain* **367**, (2020).

29. Li, T., Wernersson, R., *et al. A scored human protein-protein interaction network to catalyze genomic interpretation* **14**, 61–64 (2016).

30. Oyrer, J., Maljevic, S., *et al. Ion channels in genetic epilepsy: From genes and mechanisms to disease-targeted therapies* **70**, 142–173 (2018).

31. Reid, C. A., Berkovic, S. F., *et al. Progress in Neurobiology* vol. 87 41–57 (2009).

32. Ge, S. X., Jung, D., *et al. ShinyGO: A graphical gene-set enrichment tool for animals and plants* **36**, 2628–2629 (2020).

33. Penttilä, S., Palmio, J., *et al. GeneReviews®* (University of Washington, Seattle, 1993).

34. Xu, J., El Refaey, M., *et al. Genetic disruption of Ano5 in mice does not recapitulate human ANO5-deficient muscular dystrophy* **5**, 1–14 (2015).

35. Thiruvengadam, G., Sreetama, S. C., *et al. Anoctamin 5 Knockout Mouse Model Recapitulates LGMD2L Muscle Pathology and Offers Insight Into in vivo Functional Deficits* **8**, S243–S255 (2021).

36. Qi, C., Feng, I., *et al. Variants in ADD1 cause intellectual disability, corpus callosum dysgenesis, and ventriculomegaly in humans* **24**, 319–331 (2022).

37. Focchi, E., Cambria, C., *et al. ATM rules neurodevelopment and glutamatergic*

transmission in the hippocampus but not in the cortex **13**, 1–13 (2022).

38. Miller, J. A., Ding, S. L., *et al. Transcriptional landscape of the prenatal human brain* **508**, 199–206 (2014).

39. Hawrylycz, M. J., Lein, E. S., *et al. An anatomically comprehensive atlas of the adult human brain transcriptome* **489**, 391–399 (2012).

40. Wang, J., Lin, Z. J., *et al. Seizure* vol. 44 11–20 (2017).

41. Badouel, C., Zander, M. A., *et al. Fat1 interacts with Fat4 to regulate neural tube closure, neural progenitor proliferation and apical constriction during mouse brain development* **142**, 2781–2791 (2015).

42. Endo, M., Doi, R., *et al. Ror family receptor tyrosine kinases regulate the maintenance of neural progenitor cells in the developing neocortex* **125**, 2017–2029 (2012).

43. Iyengar, S. K. & Elston, R. C. in *Methods in molecular biology (Clifton, N.J.)* vol. 376 71–84 (Methods Mol Biol, 2007).

44. Pitkänen, A., Ndode-Ekane, X. E., *et al.* in *Progress in Brain Research* vol. 214 229–262 (Prog Brain Res, 2014).

45. Devinsky, O., Vezzani, A., *et al. Nature Reviews Disease Primers* vol. 4 1–24 (2018).

46. Heck, N., Garwood, J., *et al. Differential upregulation of extracellular matrix molecules associated with the appearance of granule cell dispersion and mossy fiber sprouting during epileptogenesis in a murine model of temporal lobe epilepsy* **129**, 309–324 (2004).

47. Zai, C. C., Lee, F. H., *et al. Investigation of the HSPG2Gene in tardive dyskinesia-new data and meta-analysis* **9**, (2018).

48. Lin, P. Y., Hung, J. H., *et al. A Novel Pathogenic HSPG2 Mutation in Schwartz–Jampel*

Syndrome **12**, 632336 (2021).

49. Trout, A. L., Kahle, M. P., *et al. Perlecan Domain-V Enhances Neurogenic Brain Repair After Stroke in Mice* **12**, 72–86 (2021).

50. Velluva, A., Radtke, M., *et al. Phenotype-tissue expression and exploration (PTEE) resource facilitates the choice of tissue for RNA-seq-based clinical genetics studies* **22**, 802 (2021).

51. Heyes, S., Pratt, W. S., *et al. Genetic disruption of voltage-gated calcium channels in psychiatric and neurological disorders* **134**, 36–54 (2015).

52. Kambouris, M., Thevenon, J., *et al. Biallelic SCN10A mutations in neuromuscular disease and epileptic encephalopathy* **4**, 26–35 (2017).

53. Karlsson, M., Zhang, C., *et al. A single–cell type transcriptomics map of human tissues* **7**, (2021).

54. Shao, D., Baker, M. D., *et al. A multi PDZ-domain protein Pdzd2 contributes to functional expression of sensory neuron-specific sodium channel NaV1.8* **42**, 219–225 (2009).

55. Hiromoto, Y., Azuma, Y., *et al. Hemizygous FLNA variant in West syndrome without periventricular nodular heterotopia* **7**, 1–5 (2020).

56. Baldassarre, M., Razinia, Z., *et al. Filamin A controls matrix metalloproteinase activity and regulates cell invasion in human fibrosarcoma cells* **125**, 3858–3869 (2012).

57. Dityatev, A. & Fellin, T. *Neuron Glia Biology* vol. 4 235–247 (2008).

58. Chelyshev, Y. A., Kabdesh, I. M., *et al. Cellular and Molecular Neurobiology* vol. 42 647–664 (2022).

59. Ferhat, L., Chevassus-Au-Louis, N., *et al. Seizures induce tenascin-C mRNA expression*

in neurons **25**, 535–546 (1996).

7. Data Availability

The data that supports this study are available in https://epi25.broadinstitute.org/downloads. The code used in the analysis is deposited under https://github.com/lbundalian/EPI25_VSEA.

8. Acknowledgement

Open Access funding was enabled and organized by Projekt DEAL. This study is funded by the Else Kröner-Fresenius-Stiftung 2020_EKEA.42 to D.L.D. and the German Research Foundation SFB 1052 project B10 to D.L.D. and A.G.. D.L.D. is funded through the "Clinician Scientist Programm, Medizinische Fakultät der Universität Leipzig". We are grateful to the EPI25 collaborative for coordinating and allowing us to use the cohort they curated.

9. Author Contribution

L.B. Conceptualization; writing – original draft; formal analysis; investigation; methodology. **D.L.D.** Investigation; methodology. **S.C., A.V., F.B., D.L., H.O.H** Methodology; validation; writing – original draft. **Y-Y.S.** and **C–C.L.** Methodology; formal analysis; writing – original draft. **A-S. K** and **A.G** Conceptualization; writing – original draft. **J.R.L.** and **D.L.D.** Conceptualization; writing – original draft; supervision; funding acquisition.

10. Ethical declaration

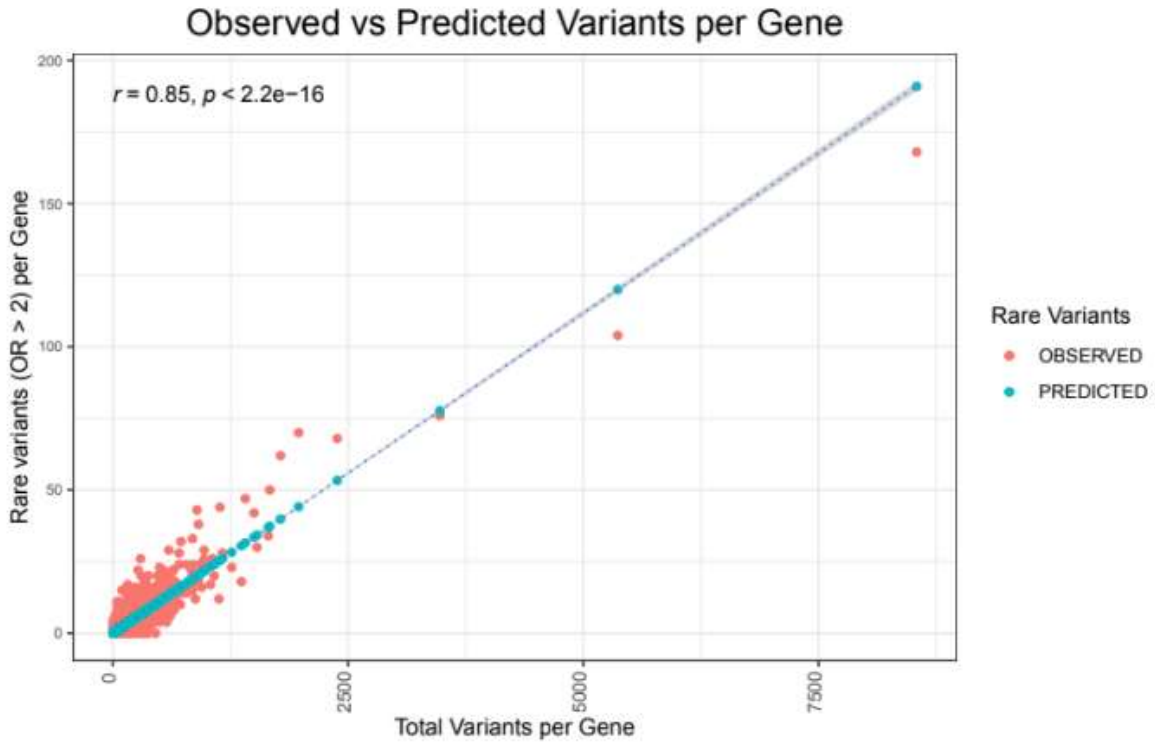
This study was approved by the ethics committee of the University of Leipzig, Germany (224/16-ek and 402/16-ek) and by the Epi25 Strategy Committee (approval from 16.10.2019). The availability of informed consent from the tested individuals was checked as part of Epi25

sample inclusion criteria. Since data analysis was performed across multiple centers, we used only aggregated data to assure patient anonymity.

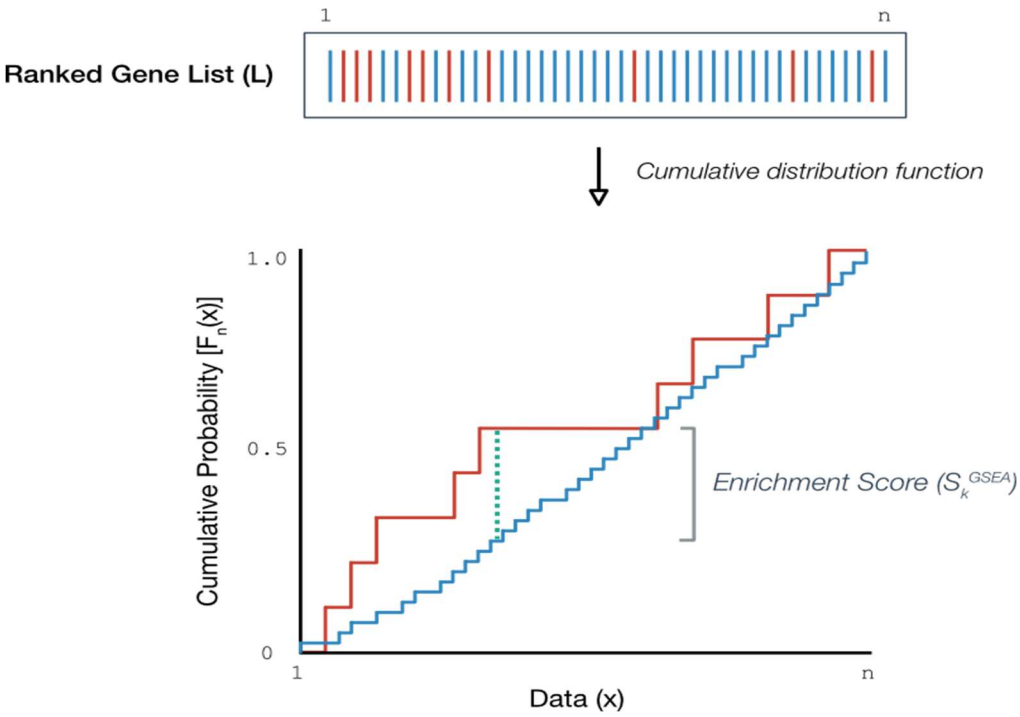
11. Conflict of Interest

The authors declare no competing interests.

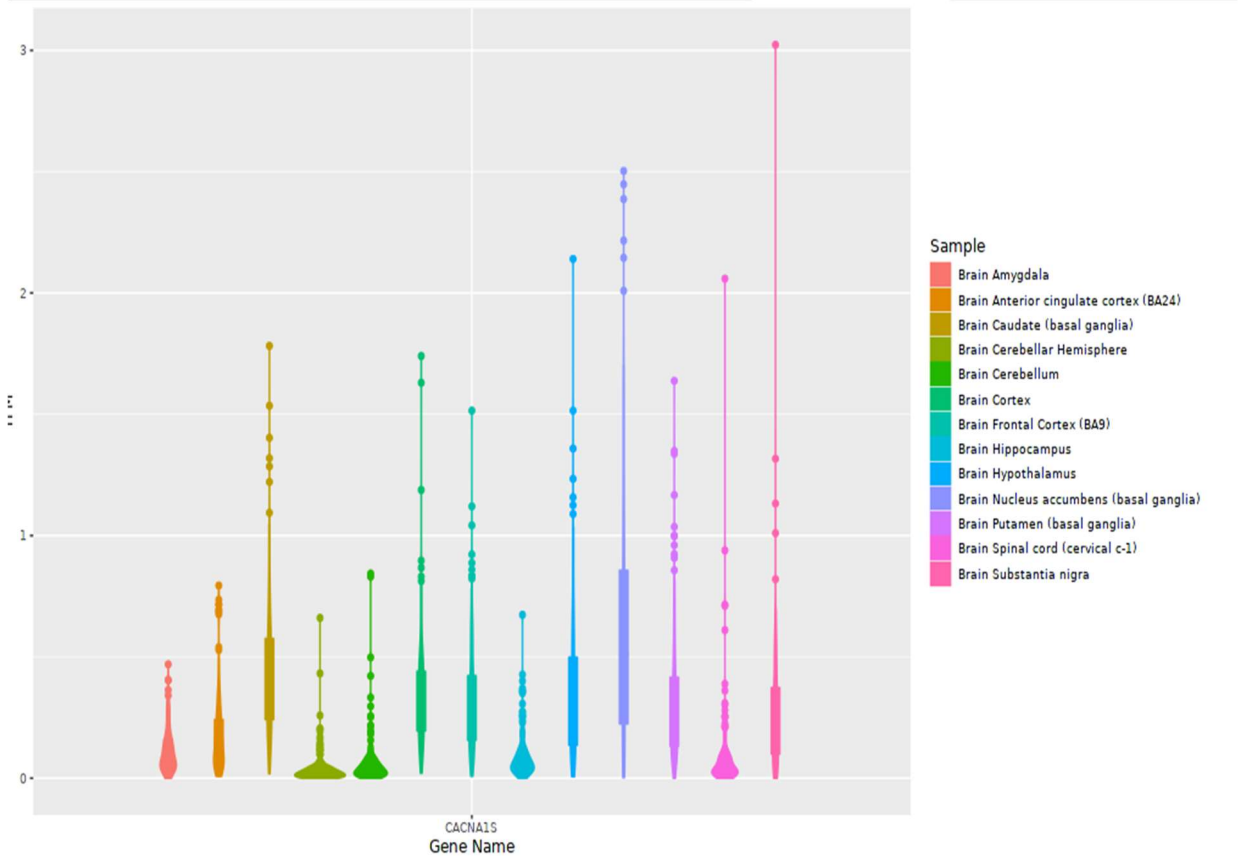
12. Supplementary Information



Supplementary Fig. 1. Relationship of the number of rare variants and total number of variants in a gene. The linear regression model (r = 0.85) established that for a given number of variants per gene (total variants), we can estimate the expected number of rare variants. Only those genes which had an excess of expected number variants (i.e. above the dotted slope) were considered for the VSEA.



Supplementary Fig. 2. Cumulative ranked sum for hits and misses of QRVs. For every QRVs found in the list of variants having a higher frequency in patients – Data(x), the cumulative HIT score (red) will be incremented by one, otherwise the cumulative MISS score (blue) will be incremented. The normalized difference in the difference between the MISS and HIT score represents the ENRICHMENT SCORE. (https://www.pathwaycommons.org/guide/primers/data_analysis/gsea/#safe)



Supplementary Fig. 3. Expression of *CACNA1S* from Single Gene Analysis. *CACNA1S* is not highly expressed in the brain. However, using the *Single Gene Analysis* feature of PTEE (*https://bioinf.eva.mpg.de/PTEE*), we can observe the outliers across different brain regions. A previous study suggested that the voltage-gated calcium channel encoded by *CACNA1S* could play a role in mediating calcium influx and neuronal excitability (*Cain, S. M., & Snutch, T. P. (2012). Voltage-Gated Calcium Channels in Epilepsy. In J. L. Noebels (Eds.) et. al., Jasper's Basic Mechanisms of the Epilepsies. (4th ed.). National Center for Biotechnology Information (US)*)