Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies

Abou-Khalil, Bassel

2018-12-10


http://hdl.handle.net/10138/287694
https://doi.org/10.1038/s41467-018-07524-z

Downloaded from Helda, University of Helsinki institutional repository.
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.
Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies

The International League Against Epilepsy Consortium on Complex Epilepsies

The epilepsies affect around 65 million people worldwide and have a substantial missing heritability component. We report a genome-wide mega-analysis involving 15,212 individuals with epilepsy and 29,677 controls, which reveals 16 genome-wide significant loci, of which 11 are novel. Using various prioritization criteria, we pinpoint the 21 most likely epilepsy genes at these loci, with the majority in genetic generalized epilepsies. These genes have diverse biological functions, including coding for ion-channel subunits, transcription factors and a vitamin-B6 metabolism enzyme. Converging evidence shows that the common variants associated with epilepsy play a role in epigenetic regulation of gene expression in the brain. The results show an enrichment for monogenic epilepsy genes as well as known targets of antiepileptic drugs. Using SNP-based heritability analyses we disentangle both the unique and overlapping genetic basis to seven different epilepsy subtypes. Together, these findings provide leads for epilepsy therapies based on underlying pathophysiology.
The epilepsies are a group of brain disorders characterized by recurrent unprovoked seizures affecting up to 65 million people worldwide. There are many different types of epilepsy, and its classification has recently evolved, driven by advances in clinical phenotyping, imaging, and genetics. Since the identification of CHRNA4 as a cause of autosomal dominant nocturnal frontal lobe epilepsy, genes underlying many different rare monogenic forms of epilepsy have been characterized, and discovery in this area has accelerated with the application of next generation sequencing. This is particularly true of the relatively rare but devastating infantile group of epileptic encephalopathies, which are now emerging as a genetically heterogeneous group of largely de novo dominant disorders. In contrast, single gene causes of the more common forms of epilepsy appear to be relatively rare. The common forms broadly comprise generalized and focal epilepsies, with the former having the highest heritability, yet the lesser yield in single gene discovery. These common forms are likely multifactorial, with a significant and complex genetic architecture.

Meet with the experience from many other disease fields, early attempts to disentangle the genetic architecture of the more common, sporadic forms of epilepsy were limited by study power and scope. In 2011, the International League Against Epilepsy (ILAE) launched the Consortium on Complex Epilepsies, to facilitate meta-analysis in epilepsy genomics. In 2014, the first such meta-analysis was reported comprising 8696 cases and 26,157 controls, which led to identification of 2q24.3, 4p15.1, and 2p16.1 as epilepsy loci. Here we present an expanded analysis involving 15,212 cases and 29,677 controls, which leads to identification of 16 genome-wide significant loci. Importantly, 11 of these loci are associated with the genetic generalized epilepsies; the group of epilepsies where despite having the highest heritability we have made the least genetic progress to date. We show that the genes associated with each locus are biologically plausible candidates, despite having diverse functions, particularly as there is a significant enrichment for known monogenic epilepsy genes and anti-epileptic drug targets.

Results

Study overview. We performed a genome-wide mega-analysis on the ILAE Consortium cohort now comprising 15,212 epilepsy cases, stratified into 3 broad and 7 subtypes of epilepsy, and 29,677 control subjects (Supplementary Table 1). The current study includes a further 6516 cases and 3460 controls in addition to the 8696 cases and 26,157 controls from our previously published analysis. Thus, this mega-analysis is not a formal replication of our previously published meta-analysis. We do not attempt any formal replication of novel association signals detected in this analysis. Furthermore, 531 cases of Asian descent, and 147 cases of African descent were included through a meta-analysis. However, we refer to our GWAS as a mega-analysis as the vast majority of our samples (96%) were analyzed under that framework.

At the broadest levels, cases were classified as (a) focal epilepsy where seizures arise in a restricted part of the brain, a form traditionally not regarded as genetic although a number of genes for monogenic forms have been identified; (b) genetic generalized epilepsy where seizures arise in bilateral networks and evidence for a genetic component is very strong, yet genes have been hard to identify, and (c) unclassified epilepsy.

Subjects were assigned to three broad ancestry groups (Caucasian, Asian and African-American) according to results of genotype-based principal component analysis (Supplementary Fig. 1). Linear-mixed model analyses were performed stratified by ethnicity and epilepsy subtype or syndrome, after which trans-ethnic meta-analyses were undertaken.

Genome-wide associations. Our analysis of all epilepsy cases combined revealed one novel genome-wide significant locus at 16q12.1 and reinforced two previous associations at 2p16.1 and 2q24.3 (Fig. 1 and Supplementary Fig. 2). When conditioning on the top SNP within the 2q24.3 locus, we demonstrate the existence of a second, independent signal within that locus (Supplementary Fig. 3). This locus was also significantly associated with focal epilepsy. Our analysis of genetic generalized epilepsy uncovered 11 genome-wide significant loci, of which seven are novel (Fig. 2).

Considering that focal and generalized epilepsy are clinically broad and heterogeneous classifications, we next assessed whether loci are specifically associated with any of the seven most common focal epilepsy phenotypes and genetic generalized epilepsy syndromes (Supplementary Fig. 4 and 5). We found a novel genome-wide significant association with juvenile myoclonic epilepsy (JME) and two novel loci associated with focal epilepsy with hippocampal sclerosis. Moreover, we found two genome-wide significant associations with childhood absence epilepsy (CAE) in loci that were previously associated with absence epilepsy and generalized epilepsy. We did not find any significant loci associated with generalized epilepsy with tonic-clonic seizures (GTCS) alone, juvenile absence epilepsy (IAE), lesional-negative or lesional focal epilepsy (other than hippocampal sclerosis). Further analysis of the association signals for each locus in the different syndromes suggested that some signals display specificity for a single subtype, while others show evidence for pleiotropy (Supplementary Fig. 6). However, the relatively small sample sizes of these phenotype subsets warrant caution for over-interpretation.

In total, we found 11 novel genome-wide significant loci associated with the epilepsies and we replicated the association of five previous known loci (Supplementary Fig. 7). Two previous reports of association did not reach our threshold for significance. This included a locus (rs2292096; 1q32.1) for focal epilepsy detected in an Asian population ($p = 0.057$ in our trans-ethnic fixed-effects meta-analysis), and rs12059546 (1q43) detected previously for JME ($p = 7.4 \times 10^{-5}$ in our Caucasian-only BOLT-LMM analysis).

Gene mapping and biological prioritization. The genome-wide significant loci from all analyses were mapped to a total of 146 genes (Supplementary Data 1) using a combination of positional mapping ($\pm 250$ kb from locus) and significant distal 3D chromatin interactions of the locus with a gene promoter (FDR $< 10^{-6}$).

Considering that most loci encompass several genes, we devised criteria to systematically prioritize the most likely candidate genes per locus based on established bioinformatics methods and resources. This biological prioritization was based on six criteria (Fig. 2), similar to previous studies. Each gene was given a score based on the number of criteria that were met (range 0–6). The gene(s) with the highest score in each locus, with a minimum of 2, were defined as biological epilepsy risk genes. We validated this approach using established epilepsy genes within our data (Supplementary Table 2). Using this approach, we were able to refine these loci to the 21 most likely biological epilepsy genes (Fig. 2).

These prioritized genes include seven ion-channel genes (SCN1A, SCN2A, SCN3A, GABRA2, KCNN2, KCNAB1, and GRK1), three transcription factors (ZEB2, STAT4 and BCL11A), the histone modification gene BRD7, the synaptic transmission gene STX1B and the pyridoxine metabolism gene PNPO. Notably, a conditional
transcriptome-wide association study (TWAS) analysis suggests that the signal for genetic generalized epilepsy at 17q21.32, which was also observed in an earlier study, seems driven by regulation of expression of \textit{PNPO} (Supplementary Fig. 8). This suggests that the biology behind pyridoxine (vitamin-B6)-responsive epilepsy could be relevant to common genetic generalized epilepsies. Biological prioritization implicates \textit{SCN1A}, \textit{SCN2A}, \textit{SCN3A}, and \textit{TTC21B} as the most likely genes underlying the signal at 2q24.3 for genetic generalized epilepsy.
all epilepsy, focal epilepsy and genetic generalized epilepsy. Pathogenic variants in the sodium channels SCN1A, SCN2A and SCN3A are associated with various epilepsy syndromes\(^\text{16}\) and mutations in TTC21B impair forebrain development\(^\text{21,22}\). Our analyses implicate STX1B as a potential gene underlying the association of JME at the 16p11.2 locus and the top variant in the locus is an eQTL that strongly correlates with expression of STX1B in the dorsolateral prefrontal cortex (Spearman correlation: \(\rho = 0.33, p = 3 \times 10^{-14}\))\(^\text{23}\). Interestingly, for one of the prioritized genes in genetic generalized epilepsy, PCDH7, an eQTL was recently detected in epileptic hippocampal tissue\(^\text{24}\). Prioritized genes associated with focal epilepsy with hippocampal sclerosis include the gap-junction gene GJA1.

In addition we identified eight genes from Fig. 2 (BCL11A, GJA1, ATXN1, GABRA2, KCNAB1, SCN3A, PCDH7, STAT4) with evidence of co-expression in at least two independent brain expression resources, using a brain gene co-expression analysis with brain-coX\(^\text{25}\). These eight candidates are embedded in several established epilepsy gene co-expression modules (Supplementary Fig. 9; Supplementary Table 9).

**SNP annotation and tissue-specific partitioned heritability.** We functionally annotated all 492 genome-wide significant SNPs from all phenotypes (Fig. 3a–c) and found that most SNPs were either intergenic (29%) or intronic (46%); 78% were in open chromatin regions (as indicated by a minimum chromatin state of \(\rho \leq 0.76,0.7\), and 50% of SNPs showed some evidence for affecting gene transcription (RegulomeDB score \(\leq 6\)). Four were coding SNPs of which two were missense variants.

To gain further biological insight into our results, we next used a partitioned heritability method\(^\text{29}\) to assess whether our genome-wide significant signals contained SNPs associated with enhanced transcription in any of 88 tissues. We found significant enrichment of H3K4me1 markers in all epilepsy (stratified LD-score regression; \(p = 4 \times 10^{-3}\)) and H3K27ac markers in genetic generalized epilepsy (stratified LD-score regression; \(p = 1.3 \times 10^{-6}\)), specifically in the dorsolateral prefrontal cortex. Moreover, the distribution of heritability enrichment P-values was strongly skewed towards brain tissues for all epilepsy phenotypes (Fig. 3d, Supplementary Figs. 10–12).

---

**Fig. 2** Genome-wide significant loci of all analyses and prioritized biological epilepsy genes. Genes were prioritized based on 6 criteria and scored based on the number of criteria met per gene (filled red boxes). The highest scoring gene, or multiple if they have the same score, in each locus is reported as ‘prioritized biological epilepsy gene(s)’. Similar to previous studies\(^\text{17,18}\), we used a minimum score of 2 to define these genes and we noted ‘none’ if no gene in the locus reached this score. Filled blue boxes indicate overlap with known targets of anti-epileptic drugs and established monogenic epilepsy genes. The lead SNP is defined as the SNP with the lowest \(P\)-value in the locus and the minor allele is displayed in brackets. \(P\)-values and \(Z\)-scores for All epilepsy, Focal epilepsy, and Generalized epilepsy were calculated with BOLT-LMM. MAF minor allele frequency in the Human Reference Consortium reference panel. The direction of the \(Z\)-score is signed with respect to the minor allele. TWAS: significant TWAS association (based on data from the CommonMind Consortium), eQTL: significant eQTL within locus (based on data from the ROS/MAP projects). Brain exp: the gene is preferentially expressed in the brain, Missense: epilepsy GWAS missense variant in locus, PPI: gene prioritized by protein-protein interaction, KO mouse: relevant knockout mouse phenotype.
H3K27ac and H3K4me1 are epigenetic markers associated with regulating gene transcription, suggesting that altered transcription in the dorsolateral prefrontal cortex could be one of the underlying mechanisms of epilepsy. This is further supported by a tissue-specific heritability enrichment analysis (using data from the GTEx Consortium), showing strongest enrichment for genetic generalized epilepsy with genes expressed in Brodmann Area 9 (stratified LD-score regression; \( p = 1.56 \times 10^{-6} \)), which encompasses the dorsolateral prefrontal cortex (Fig. 3e). These findings further corroborate our TWAS results (using data from the unrelated CommonMind Consortium database), which shows significant associations of epilepsy with gene expression of several
genes in the dorsolateral prefrontal cortex (Fig. 2; Supplementary Table 3). Although genetic generalized epilepsy has been regarded as a generalized process, anatomical, electrophysiological, cognitive, and functional imaging studies implicate dysfunction in the frontal lobes.\(^{30-34}\) Altogether, we have converging evidence from several unrelated methods and databases suggesting epigenetic regulation of gene expression in the dorsolateral prefrontal cortex as a potential pathophysiological mechanism underlying our epilepsy GWAS findings.

Finally, we leveraged the Brainspan database, as implemented in FUMA,\(^{35}\) to assess whether the genes implicated by our GWAS are differentially expressed in the brain at various prenatal and post-natal ages. These analyses were performed for the genes prioritized in any epilepsy phenotype (21 genes), any focal epilepsy subtype (8 genes) or any genetic generalized epilepsy syndrome (15 genes). The results suggest that the expression of genes associated with focal epilepsy is up-regulated in late-infancy and young adulthood, whereas expression of those genes associated with genetic generalized epilepsy is down-regulated in early childhood and differentially expressed prenatally and at adolescence (Supplementary Fig. 13).

**Enrichment analyses.** A previous exome-sequencing study found an association for common epilepsies with ultra-rare variants in known monogenic epilepsy genes.\(^{36}\) To assess whether common epilepsies are also associated with common variants in monogenic epilepsy genes (see Methods), we pooled the 146 genes that were mapped to our genome-wide significant loci and performed a hypergeometric test. Results illustrated an enrichment of known monogenic epilepsy genes within the genes mapped to our genome-wide significant loci (6 genes overlapped; hypergeometric test: odds ratio \(\text{OR} = 8.45, p = 1.3 \times 10^{-5}\)). This enrichment is considerably more significant when limited to the 21 genes with the highest biological priority from Fig 2 (5 genes overlapped; hypergeometric test: OR = 61.4, \(p = 9.9 \times 10^{-10}\)). We did not find a bias for gene size in our enrichment analyses when using a conservative method to correct for this (see Methods). This suggests that both common and rare variants in monogenic epilepsy genes contribute to common epilepsy susceptibility, corroborating and further extending previous observations.\(^{8,37}\) Further studies are required to establish whether the signals from common and rare variants are independent of each other.

Using public databases of drug-targets, we found that 13 out of 24 currently licensed anti-epileptic drugs target genes that are implicated in our GWAS. Using the same list of 146 genes as described above, we performed a hypergeometric test which shows a significant enrichment of genes that are known targets of anti-epileptic drugs (8 genes overlapped; hypergeometric test: OR = 19.6, \(p = 1.3 \times 10^{-9}\)). This enrichment is considerably more significant when limited to the 21 most biologically plausible candidate genes (5 genes overlapped; hypergeometric test: OR = 101.2, \(p = 5.7 \times 10^{-11}\)). This observation suggests that other drugs that target genes from our GWAS could also have potential for the treatment of epilepsy. The Drug-Gene interaction database (http://dgidb.org) lists 166 drugs that target biologically prioritized genes from our GWAS (see Supplementary Data 2 for a full list), that may be further investigated for their anti-seizure potential.

Next, we used a complementary approach to search for repurposable drugs. By comparing GWAS-imputed and drug-induced transcriptomes, we predicted drugs capable of rectifying epilepsy-associated gene expression changes (see Methods). Our predictions are enriched with licensed antiepileptic compounds (permutation based \(p\)-value \(<1.0 \times 10^{-6}\)) and with other licensed compounds that have proven antiepileptic efficacy in animal models (permutation based \(p\)-value \(<1.0 \times 10^{-6}\)). We list 30 of our predicted drugs that are licensed for other conditions and have published evidence of efficacy in animal models of epilepsy (Supplementary Table 4).

**Heritability analyses.** Twin-based and genetic heritability studies have suggested that while epilepsy is strongly heritable, there is a substantial missing heritability component.\(^{40,41}\) We used LDAK to estimate \(h^2_{SNP}\), the proportion of heritability that can be attributed to SNPs.\(^{32,42}\) We estimate \(h^2_{SNP} = 32.1\% \) (95%CI: 29.6–34.5%) for genetic generalized epilepsy and \(h^2_{SNP} = 9.2\% \) (8.4–10.1%) for focal epilepsy (estimates are on the liability scale, assuming a prevalence of 0.002 and 0.003, respectively) which are consistent with previous estimates.\(^{8}\) These results indicate that SNPs explain a sizeable proportion of the liability of genetic generalized epilepsy syndromes, but less so for focal epilepsy phenotypes (Fig. 4). To delineate the heritability of the different epilepsy phenotypes, we used LDAK to perform genetic correlation analyses between the different forms. We found evidence for strong genetic correlations between the genetic generalized epilepsies, whereas we found no significant correlations between the focal epilepsies (Fig. 4). Interestingly, we found a significant genetic correlation between JME and lesion-negative focal epilepsy (LDAK genetic correlation: \(R^2 = 0.46, p = 8.77 \times 10^{-4}\), suggesting either pleiotropy and/or misclassification. It is known that focal EEG features can be seen in JME.\(^{49}\)

In view of the increasing data on comorbidities with epilepsy, we next used LD-score regression to analyze the genetic correlation between epilepsy and various other brain diseases and traits from previously published GWAS (Fig. 5; see Supplementary Table 5 for values). Perhaps surprisingly, we did not find significant correlations with febrile seizures. Similarly, we did not find any significant genetic correlations between epilepsy and other neurological or psychiatric diseases. However, we did observe significant correlations for all epilepsy and genetic generalized epilepsy with cognitive ability. We then used the method Multi-Trait Analysis of GWAS (MTAG)\(^{46}\) to leverage the larger sample size of the genetically correlated GWAS of cognitive ability \((n = 78,308)\) in order to boost the effective sample size of our all and genetic generalized epilepsy GWAS to 53,244 and 41,515 respectively. Using this approach, we found a novel genome-wide significant locus at 10q24.32 in all epilepsy (MTAG \(p = 2.2 \times 10^{-9}\)) and genetic generalized epilepsy (MTAG \(p = 4.0 \times 10^{-5}\)) which encompasses the \(K_C\)-channel-interacting protein 2 (KCNIP2) gene (Supplementary Fig. 14), loss of which is associated with seizure susceptibility in mice.\(^{47}\)

**Discussion**

The increased sample size in this second ILAE Consortium GWAS of common epilepsies has resulted in the detection of 16 risk loci for epilepsy and illustrates how common variants play an important role in the susceptibility of these conditions. But compared to other common neurological diseases our sample size is modest. For example the latest GWAS in schizophrenia considered 36,989 schizophrenia cases and 113,075 controls, resulting in the identification of 108 risk loci.\(^{48}\) Larger efforts would deliver further insight to the genetic architecture of the common epilepsies.

The majority of the loci are associated with genetic generalized epilepsy. This observation is a welcome partial explanation for the high heritability of genetic generalized epilepsy, in light of the relative lack of rare variant variants discovered to date. We also show that there is substantial genetic correlation between the generalized syndromes. We speculate that the subtypes share a large part of the genetic susceptibility for generalized epilepsies, with specific modifying factors determining the specific syndrome.
**Fig. 4** Heritability estimates and genetic correlations between epilepsy syndromes, calculated using LDAK. Subjects with a diagnosis of both CAE and JAE were excluded from both phenotypes. The genetic correlation coefficient was calculated with LDAK and is denoted with a color scale ranging from 0% (white) to 100% (red). #: $P < 0.05$; *: $P < 0.0024$ (Bonferroni threshold); $h^2_{SNP}$: SNP-based heritability on liability scale (95% CI); †: heritability estimate exceeded 100%, possibly due to small sample size and large SD; CAE - childhood absence epilepsy, JAE - juvenile absence epilepsy, JME - juvenile myoclonic epilepsy, GTCS alone - generalized tonic-clonic seizures alone, focal HS - focal epilepsy with hippocampal sclerosis.

**Fig. 5** Genetic correlations of epilepsy with other phenotypes. The genetic correlation coefficient, calculated using LD-score regression, is denoted with a color scale ranging from -100% (blue) to 100% (red). #: $P < 0.05$ *: $P < 0.001$ (Bonferroni threshold; 0.05/48)
Some syndrome-specific associations were detected, such as the relatively strong signal for STX1B in JME, and the association of GJA1 with focal epilepsy-hippocampal sclerosis. Interestingly, although the association signal for STX1B was only significant in the JME analysis, rare pathogenic variants in STX1B have been recently found in a spectrum of epilepsies, including genetic epilepsy with febrile seizures plus (GEFS+), genetic generalized epilepsies (including JME), epileptic encephalopathies and even some focal epilepsies18,19 (Wolking et al., Manuscript submitted (2018)). Further, mutations in the gap-junction gene GJA1 are associated with impaired development of the hippocampus20 and different expression has been reported in epileptic hippocampal and cortical tissue21,22,23. These findings represent a tantalizing glance of the different biological mechanisms underlying epilepsy syndromes that may guide us to the introduction of genetics for improved diagnosis, prognosis and treatment for these common epilepsies. However, the relatively low sample size of our subtype analysis warrants a conservative interpretation and follow-up with a larger cohort.

At least three association signals are shared between focal epilepsy and genetic generalized epilepsy. The clearest overlapping signal remains the 2q24.3 locus, as we reported previously15. However, this association signal is complex and we demonstrate that the locus consists of at least two independent signals (Supplementary Fig. 3). Our Hi-C chromatin analysis suggests the complexity includes levels of functional association to SCN2A and SCN3A, that are located more distally to the SCN1A locus. Mutations in SCN2A and more recently SCN3A are established monogenic causes of epileptic encephalopathy that, like SCN1A, cause dysfunction of the encoded ion-channels, which is believed to disturb the fine balance between neuronal excitation and inhibition. This may involve independent variation that either affects regulation of SCN1A, SCN2A, or SCN3A independently. However, the complex association may also reflect multiple rare risk variations, and large resequencing studies will shed further light on this issue.

The number of association signals we detected and increased power relative to our previous meta-analysis15 allowed us to explore the biological mechanisms behind the observed genetic associations. We show that the signals converge on the dorsolateral prefrontal cortex as the tissue in which most functional effect is observed; this is broadly consistent with the importance of the frontal lobes in generalized epilepsies. Indeed, our analyses of the epigenetic markers H3K27ac and H3K4me1, TWAS, and tissue-specific heritability enrichment analysis all point towards epigenetic regulation of gene expression in the dorsolateral prefrontal cortex as a potential pathophysiological mechanism underlying our epilepsy GWAS findings.

Altogether, we found 16 loci that are associated with the common epilepsies. Our heritability analyses show that collectively, common genetic variants explain a third of the liability for genetic generalized epilepsy. Our analyses suggest that the associated variants are involved in regulation of gene expression in the brain. The 21 biological epilepsy candidate genes implicated by our study have diverse biological functions, and we show that these are enriched for known epilepsy genes and targets of current antiepileptic drugs. Our analyses provide evidence for pleiotropic genetic effects that raise risk for the common epilepsies collectively, as well as variants that raise risk for specific epilepsy syndromes. Determining the shared and unique genetic basis of epilepsy syndromes should be of benefit for further stratification and eventually lead to possible applications for improved diagnosis, prognosis, and treatment. Future studies including pharmacoresponse data, imaging, and other clinical measurements have the potential to further increase the benefit of these studies for people with epilepsy. In combination, these findings further our understanding of the complex genetic architecture of the epilepsies and could provide leads for new treatments and drug repurposing.

Methods

Ethics statement. We have complied with all relevant ethical regulations. All study participants provided written, informed consent for use of their data in genetic studies of epilepsy. For minors, written informed consent was obtained from their parents or legal guardian. Local institutional review boards approved study protocols at each contributing site.

Cohorts and phenotype definition. A list of the sites included in this study is described in Supplementary Table 6. We classified seizures and epilepsy syndromes according to the classification and terminology outlined by the ILAE18,24. For all ethnic groups, specialists assessed each phenotype at the contributing site. Individuals with epilepsy were initially assigned to one of three phenotypic categories: genetic generalized epilepsy, focal epilepsy, or unclassified epilepsy. Based on EEG, MRI and clinical histories we further classified our cohort into the epilepsy subtypes listed in Supplementary Table 1. We used a combination of population-based datasets as controls. Some control cohorts were screened by questionnaire for neurological disorders. 53.4% of cases were female compared to 51.6% of controls.

Study design. We conducted a case-control study in subjects of Caucasian, Asian (Han Chinese) and African-American ethnicities. Our primary analyses were structured to test common genetic variants for association with epilepsy according to broad epilepsy phenotypes. We pooled cases from cohorts of the same ethnic background to perform linear mixed model analysis, followed by subsequent meta-analyses of regression coefficients across the three ethnic groups. Our secondary analyses tested for associations with specific syndromes of genetic generalized epilepsy (childhood absence epilepsy, juvenile absence epilepsy, juvenile myoclonic epilepsy, and generalized tonic-clonic seizures alone) and phenotypes of focal epilepsy (focal epilepsy with hippocampal sclerosis, and focal epilepsy with other lesions). The secondary analyses were limited to Caucasian subjects due to sample size. We prioritized the results of the GWAS by incorporating eQTL information, transcriptome-wide analysis, and biological annotation. Finally, we estimated the genetic correlation of epilepsy phenotypes using Linkage-Disequilibrium Adjusted Kinship (LDAK).

Genotyping. The EpiPGX samples were genotyped at deCODE Genetics on Illumina OmniExpress-12 v1.1 and OmniExpress-24 v1.1 single nucleotide polymorphism (SNP) arrays. The EPGP samples were genotyped on Illumina HumanCore beadchips at Duke University, North Carolina. The remainder of the samples were genotyped on various SNP arrays, as previously published15.

Genotyping quality control and imputation. Quality control of genotyping was performed separately for each cohort using PLINK 1.09. Each genotype cohort was temporarily merged with a genetically similar reference population from the 1000 Genomes Project (CEU, CHB, or YRI). A test for Hardy–Weinberg equilibrium (HWE) was performed and SNPs significant at $p < 1 \times 10^{-10}$ were removed. All samples and all SNPs with missing genotype rate $>0.05$ and all SNPs with minor allele frequency (MAF) $<0.01$ were removed. Next, we pruned the PLINK --indep-pairwise command (settings: window size 100 kb, step size 25, $R^2 < 0.1$). Using this subset of SNPs, we removed samples with outlying heterozygosity values ($>5$ SD from the median of the whole cohort). Identity by descent (IBD) was calculated in PLINK to remove sample duplicates ($>0.9$ IBD) and to identify cryptic relatives. We removed one from each IBD pair with IBD$>0.87$ except for the exception of the EPGP familial epilepsy cohort. Subjects were removed if sex determined from X-chromosome genotype did not match reported gender. Array-specific maps were used to update all SNPs positions and chromosome numbers to the Genome Reference Consortium Human Build 37 (GRCh37), and remove all A/ T mismatches. We applied pre-imputation checks according to scripts available on the website of Will Rayner of the Wellcome Trust Centre for Human Genomics (www.well.ox.ac.uk/~wrayner/tools/) to remove SNPs with allele frequencies deviating $>20\%$ from the frequency in the HaploTyper Reference Consortium. Samples were submitted to the Sanger Imputation Service (http://snpimputation.sanger.ac.uk/)26. We selected the Human Reference Consortium (release 1.1; $n = 32470$) dataset as reference panel for Caucasian and Asian datasets and the African Genome Resources ($n = 4956$) for the African-American datasets. Post-imputation quality control filters were applied to remove SNPs within each imputed cohort with an imputation info score $<0.9$ or HWE $p < 1\times 10^{-6}$. Imputed genotype dosages with a minimum certainty of 0.9 per subject were converted to hard-coded PLINK format after which all samples were pooled into a single cohort. We performed a principal components analysis using GCTA. From the PCA results we stratified our subjects into three broad ethnic groups (Caucasian, Asian and African) while removing extreme outliers. After stratifying by ethnicity, we removed SNPs with HWE $p < 1\times 10^{-6}$ and $MAF < 0.01$. In total 816 subjects out of 45705 subjects were filtered out by quality control procedures, leaving 44889 subjects for analyses.
**Study power.** We estimated using PGA27 that the study had 80% power to detect a genetic predictor of relative risk for all epilepsy (approximated to odds ratio) ≥1.45 with MAF = 1% and an alpha level of 5 × 10⁻⁸. We estimated that our meta-analyses had 80% power to detect genome-wide significant SNPs of MAF = 1% with relative risks ≥1.5 and ≥2.1, for focal and generalized epilepsy respectively (see Supplementary Figure 15). Our analysis of generalized epilepsy sub-phenotypes had 80% power to detect genome-wide significant SNPs of MAF = 1% with relative risks ≥2.3, ≥2.3, and ≥2.4 for CAE, JAE, and JME respectively. Our analysis of focal epilepsy sub-phenotypes had 80% power to detect genome-wide significant SNPs of MAF = 1% with relative risks ≥1.9, ≥2.8, and ≥2.9 for focal epilepsy lesion negative, focal epilepsy with hippocampal sclerosis and focal epilepsy with lesion other than hippocampal sclerosis, respectively.

**Statistical analyses.** Association analyses were conducted within the three ethnic subgroups using a linear mixed model in BOLT-LMM58. A subset of SNPs, used to correct for (cryptic) relatedness and population stratification by BOLT-LMM, were derived by applying SNP imputation into score >0.99, MAF >0.1, call rate >0.99 before pruning the remaining variants using LDAX with a window size of 1 MB and R² = 0.243. All analyses included gender as a covariate and the threshold for statistical significance was set at 5 × 10⁻⁶. We compared χ² values of the BOLT-LMM output between all pairs of SNPs in high LD (R² > 0.4) and removed pairs of SNPs with extreme χ² differences using a formula that scales exponentially with magnitude of χ² and L: χ² difference cutoff = \(\frac{\chi^2 \cdot \ln(L)}{L} \), where SNP1 - χ² and SNP2 - χ² are the χ²-statistic of the two SNPs in each pair and L² is their squared correlation (LD). We tested the homogeneity of all SNPs by splitting each LD block into 13 distinct local LD cluster sets and controls and removed SNPs exhibiting significant heterogeneity of effect (Phet < 1 × 10⁻⁶). Fixed effects, trans-ethnic meta-analyses were conducted using the software package METAL59. Manhattan plots for all analyses were created using qqman. Considering that our study had unequal case-control ratios, we calculated a effective sample size per ethnicity using the formula recommended by METAL: Neff = 4/(1/Ncases + 1/Ncontrols). Since >95% of all cases were Caucasian, we included all SNPs that were present in at least the Caucasian dataset (−5 million).

Conditional association analysis was performed with PLINK on loci containing significant SNPs to establish whether other variants in the region (500 KB upstream and downstream) were independently associated with the same phenotype. The conditional threshold for significance was set at 2 × 10⁻⁵ based on approximately 2500 imputed variants per 1 MB region.

**Assessment of inflation of the test statistic.** Potential inflation of the test statistic was assessed by ethnicity and phenotype by calculating the genomic inflation factor (λ; the ratio of the median of the empirically observed distribution of the test statistic to the expected median) and the mean χ². Since λ is known to scale with sample size, we also calculated the λbias, i.e. λ corrected for an equivalent sample size of 1000 cases and 1000 controls60. We observed some inflation of the test statistic (λ > 1) across the different phenotypes (Supplementary Table 7), suggesting either polygenicity or confounding due to population stratification or cryptic relatedness. Therefore, we applied LD score regression61, estimating LD scores using matched populations from the 1000G (EUR for Caucasians (n = 669), AFR for African-Americans and EAS for Asians). These LDSC results suggested that inflation of the test statistic was primarily due to polygenicity for most analyses (Supplementary Table 7). Only the Caucasian focal and all epilepsy analyses had LDSC intercepts >1.1, suggesting confounding or an incomplete match for the LD-score reference panel. Our focal and all epilepsy analyses included cohorts from various Caucasian ethnicities, including Finnish and Italian focal epilepsy cohorts, and it has been shown that LD differs considerably between Finnish and Italian various Caucasian ethnicities, including Finnish and Italian focal epilepsy cohorts, since we used a mixed-model analysis that corrects for population stratification and cryptic relatedness58.

**Gene mapping and biological prioritization.** Genome-wide significant loci of all phenotypes were mapped to genes in and around these loci using FUMA35. Genes were considered to be preferentially expressed (Bonferroni corrected p-value threshold = 0.05/53 = 0.00094) to define significant TWAS associations.

2. Genes for which a SNP in the genome-wide significant locus (as defined above) is a cis-eQTL (Bonferroni corrected p < 8 × 10⁻⁵)23 based on data from the ROS and MAP studies, which includes RNA-sequencing data from 494 dorsolateral prefrontal cortical tissues23.

3. The gene is preferentially expressed in the brain. This was assessed by using expression-data from all 53 tissues of the Gene-Tissue expression (GTex) Consortium66, after which the association between the epilepsy phenotype with expression of these 53 genes was significantly heritable (heritability p-value <0.01, as suggested by Gusev et al.64). We set a Bonferroni corrected p-value threshold of 0.05/53 = 0.00094 to define significant TWAS associations.

4. Genes for which a SNP in the genome-wide significant locus (as defined above) is a missense variant, as annotated by ENSEMBL67.

5. Genes prioritized by protein-protein interaction network, as calculated using the default settings of DAPPLE68, which utilizes protein–protein interaction data from the InWeb database39. The 146 genes implicated by our GWAS were input after which DAPPLE assessed direct and indirect physical interactions to create a protein–interaction network. Next, DAPPLE assigned a significance score according to several topological metrics; genes with a corrected p < 0.05 were considered to be prioritized by DAPPLE.

6. Genes for which a nervous system or behavior/neurological phenotype was observed in knockout mice. Phenotype data of knockout mice was downloaded from the Mouse Genome Informatics database (http://www.informatics.jax.org/) on 17 January 2018 and nervous system phenotype (GTEx ID: MP:0003631) and behavior/neurological phenotype (MP:0005386) data were extracted.

All 146 genes were scored based on the number of criteria met (range 0–6 with an equal weight of 1 per criterion), see Supplementary Data 1 for a full list. We considered the gene(s) with the highest score in each locus as the most likely causal epilepsy candidate gene. Multiple genes in a locus were selected if they had an equally high score whilst no genes were selected in a locus if all genes within it had a score <2, similar to previous studies17,18.

**Gene co-expression analysis for epilepsy with brain-coX.** In silico gene prioritization was performed using brain-coX25. Brain-coX uses a compeundom of seven large-scale normal brain gene expression data resources to identify co-expressed genes with a set of given genes (known, or putative, disease causing) likely to encode disease genes involved in disease. This approach can identify, and thus leverage networks that are not currently known and not present in available resources such as PPI networks and is a complementary approach to these. We used a set 102 monogenic epilepsy genes (Supplementary Table 8) as the set of known epilepsy genes. An FDR of 0.2 was used to identify genes that significantly co-express with the known set of genes. Prioritization in at least three datasets at an FDR of 0.2 led to a specificity of 0.925.

In the first analysis we used a set of 16 candidate epilepsy genes identified by the GWAS analysis and prioritized using additional methods (Fig. 2). These excluded any genes already included in the set of known epilepsy genes (Supplementary Table 8). Supplementary Fig. 9 shows the gene co-expression pattern using the weighted average gene co-expression across all seven datasets for candidate genes from the GWAS that show significant gene co-expression with any of the 102 known epilepsy genes.

In the second analysis we used the set of all the 146 candidate genes identified in the GWAS analysis (Supplementary Data 1). Only 140 of these were identified as having available expression data in the gene expression resources. Many genes showed some evidence of gene co-expression but few showed co-expression in more than datasets (18 out of 140). BCL11A (6) and GFI (6) remain the most robust candidates co-expressed with known epilepsy genes. The complete results are shown in Supplementary Table 9.

**Functional annotations.** We annotated all genome-wide significant SNPs (p < 5 × 10⁻⁸) from all phenotypes using the Variant Effect Predictor of ENSEMBL67 and the RegulomeDB database20. We annotated chromatin states using epigenetic data from the NIH Roadmap Epigenomics Mapping Consortium32 and ENCODE31. We used FUMA35 to annotate the minimum chromatin state (i.e. the most active state) across 127 tissues and cell types for each SNPs, similar to a previous study27.

**Heritability enrichment of epigenetic markers and gene-expression.** We stratified LD-score regression27 to assess tissue-specific heritability enrichment of epigenetic markers in 88 tissues, using standard procedures29. We used the same
We supplemented the list of 43 known dominant epilepsy genes with an additional 59 monogenic epilepsy genes from the GenoDx comprehensive epilepsy panel (www.genodx.com). We compiled the list of drug target genes from17, supplemented with additional FDA & EMA licensed AEDs. The full list of gene targets considered in each analysis are listed in Supplementary Tables 8 and 10.

Enrichment analyses corrected for gene size. Brain expressed genes known to be larger in size than non-brain expressed genes. To assess whether gene size could be a cause of bias for our enrichment analyses, we first assessed whether the size of the genes mapped in our analyses was different than non-mapped genes in the genome. We found that the size of the 146 genes mapped to genome-wide significant loci was 65.6 kb, whereas the average gene size of all other protein-coding genes was 62.2 kb. As a conservative approach to correct for this size difference, we have used the Wallenius non-central hypergeometric distribution (Wallenius, 2016) to assess whether the genes mapped to genome-wide significant loci were significantly enriched for epilepsy genes, as previously described, using the package ‘BiasedUrn’. Using this distribution, we repeated our hypergeometric analyses under the conservative assumption of a 2.43 times increased likelihood of mapping epilepsy genes as opposed to non-epilepsy genes. Using this distribution, the 146 genes that were mapped to genome-wide significant loci were significantly enriched for epilepsy genes (Wallenius non-central hypergeometric test, p = 5.3 × 10^{-8}). When limiting our results to the 21 biological prioritized genes, the enrichment of monogenic epilepsy genes became more significant (Wallenius’ noncentral hypergeometric distribution p = 3.3 × 10^{-9}). Similarly, we observed that the targets of AEDs are on average 2.43 times longer than non-AED target genes (151.8 kb vs 62.4 kb). When correcting for this gene-size difference under the assumption of a 2.43 times increased likelihood of mapping our genome-wide significant loci to AED target genes, we find that the 146 mapped genes were significantly enriched for AED target genes (Wallenius’ non-central hypergeometric test, p = 1.7 × 10^{-8}). When limiting our results to the 21 biological prioritized genes, the enrichment of AED target genes became more significant (Wallenius’ noncentral hypergeometric test, p = 1.0 × 10^{-8}).

Connectivity mapping. Connectivity mapping was performed using our GWAS results in order to identify drugs which can potentially be repurposed for the treatment of epilepsy, enabling significant savings in the time and cost of anti-epileptic drug development. Recently, So et al. identified candidate drugs that could be repurposed for the treatment of schizophrenia by using GWAS results to impute the gene expression changes associated with the disease and, then, identify drugs that change gene-expression in the opposite direction in cell lines66. Interestingly, the set of candidate drugs they identified was significantly enriched with antipsychotics. We adopted a similar strategy.

Gene-expression changes associated with epilepsy were imputed from the all epilepsy GWAS summary statistics using the FUSION software package67 and dorso-lateral prefrontal cortex tissue RNA-sequencing data (n = 452, CommonMind Consortium68). We calculated z-scores for the association between epilepsy and changes in expression of all 5261 significant genes, based on the Wallenius non-central hypergeometric test with a z-score threshold of 4. We then computed the median rank of clinically effective drug targets, and calculated the median rank of experimentally validated drug repurposing candidates. The ratio of ranks between clinically effective drugs and repurposing candidates was <1.0, indicating that the set of candidate drugs identified are more likely to be effective for epilepsy than non-brain expressed genes. To assess whether gene size is a cause of bias for our enrichment analyses, we first assessed whether the size of the genes mapped in our analyses was different than non-mapped genes in the genome. We found that the size of the 146 genes mapped to genome-wide significant loci was 65.6 kb, whereas the average gene size of all other protein-coding genes was 62.2 kb. As a conservative approach to correct for this size difference, we have used the Wallenius non-central hypergeometric distribution (Wallenius, 2016) to assess whether the genes mapped to genome-wide significant loci were significantly enriched for epilepsy genes, as previously described, using the package ‘BiasedUrn’. Using this distribution, we repeated our hypergeometric analyses under the conservative assumption of a 2.43 times increased likelihood of mapping epilepsy genes as opposed to non-epilepsy genes. Using this distribution, the 146 genes that were mapped to genome-wide significant loci were significantly enriched for epilepsy genes (Wallenius non-central hypergeometric test, p = 5.3 × 10^{-8}). When limiting our results to the 21 biological prioritized genes, the enrichment of monogenic epilepsy genes became more significant (Wallenius’ noncentral hypergeometric distribution p = 3.3 × 10^{-9}). Similarly, we observed that the targets of AEDs are on average 2.43 times longer than non-AED target genes (151.8 kb vs 62.4 kb). When correcting for this gene-size difference under the assumption of a 2.43 times increased likelihood of mapping our genome-wide significant loci to AED target genes, we find that the 146 mapped genes were significantly enriched for AED target genes (Wallenius’ non-central hypergeometric test, p = 1.7 × 10^{-8}). When limiting our results to the 21 biological prioritized genes, the enrichment of AED target genes became more significant (Wallenius’ noncentral hypergeometric test, p = 1.0 × 10^{-8}).
assessed above) with a larger sample size (n=78,307). MTAG utilizes the fact that estimations of effect size and standard error of a primary GWAS, in this case epilepsy, can be improved by matching them to a genetically correlated secondary GWAS, in this case cognitive ability.

Data availability

The GWAS summary statistics data that support the findings of this study are available at http://www.epigad.org/gwas_ilae2018_16loci.html.

References

Acknowledgements

We are grateful to the patients and volunteers who participated in this research. We thank the following clinicians and research scientists for their contribution through sample collection (cases and controls), data analysis, and project support: Geka Ack...

Author contributions

Data analysis: protocol development and main analyses: G.L.C., B.P.C.K., Ro.Kr. (data...

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-07524-z | www.nature.com/naturecommunications

12 NATURE COMMUNICATIONS | (2018) 9:5269 | DOI: 10.1038/s41467-018-07524-z | www.nature.com/naturecommunications
The International League Against Epilepsy Consortium on Complex Epilepsies
