Benchmarking of univariate pleiotropy detection methods, with an application to epilepsy phenotypes

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

Dr. rer. nat.

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Oluyomi Modupe Adesoji

aus Ile-Ife, Nigeria

Köln, 2022

Betreuer:	Herr Prof. Dr. Michael Nothnagel
Berichterstatter: (Gutachter)	Herr Prof. Dr. Peter Nuernberg
Tag der mündlichen Prüfung:	04. 10.2022

Table of Contents

1. Abbreviations

SNP	Single nucleotide polymorphism
DNA	Deoxyribonucleic acid
mRNA	Messenger ribonucleic acid
GWAS	Genome-wide association study
LD	Linkage disequilibrium
CP	Cross-phenotypes
bp	Base pairs
RR	Relative risk
OR	Odds ratio
FWER	Family-wise error rate
TILAE	The International League Against Epilepsy
FE	Focal epilepsy
GGE	Genetic generalized epilepsy
ASSET	Subset-based meta-analysis
PLACO	Pleiotropic analysis under the composite null hypothesis
cFDR	Conditional false discovery rate
MA	Meta-analysis
CPBayes	Cross-phenotype Bayes

2. List of Figures

3. List of Tables

4. Summary

Over the past decades, various methods have been used to scan the human genome to identify genetic variations associated with diseases, in particular with common, complex disorders. One of such approaches is the genome-wide association study (GWAS), which compares genetic variation between affected and healthy individuals to find genomic variants in the DNA sequence associated with a trait. GWAS are usually conducted separately for individual traits, and the same single nucleotide polymorphisms (SNP)/loci are associated with different traits in independent studies [7-10]. These findings buttress the knowledge that most complex traits are correlated and have shared genetic architecture, therefore, sharing the same heritable risk factors[11]. Knowledge of the genetic risk factors can directly or indirectly contribute to improvements in risk assessment, drug target development, and ultimately in providing effective therapies to the affected individuals.

Pleiotropy is the phenomenon of a hereditary unit affecting more than one trait, and the earliest reported evidence was provided by Mendel when he noted that some set of features were always observed together in a plant. Although this example could have been purely due to linkage and could be regarded as spurious pleiotropy in recent times, it opened up more discussion and research into pleiotropy, which has since been an active area of research[12]. In this work, I focused on complex epilepsies and the overlap in the genetic factors impacting their phenotypes.

Epilepsy is a brain disorder comprising monogenic and common/complex forms characterized by recurrent partial or generalized seizures. However, the extent to which genetic variants contribute to the disorder and how much of the genetic contribution is shared between the different phenotypes is not yet fully understood. This motivated this project, where I benchmarked available pleiotropy detection approaches to select the best performing method in terms of power and false-positive rate to detect true pleiotropy. Then, I applied the selected method to summary statistics of focal epilepsy (FE) and genetic generalized epilepsy (GGE), provided by the International League Against Epilepsy Consortium (ILAE) on complex epilepsies and the EPI25 collaborative, to identify shared genetic factors in both phenotypes of epilepsy.

Identifying pleiotropic SNPs or genes is an active area of research with multiple proposed approaches, broadly categorized into univariate and multivariate methods. Multivariate approaches have the limitation that they require all phenotypes to be measured in the same individual and their corresponding genotype data provided, which is often not the case since GWAS are usually performed per specific trait. However, various consortia studying complex traits readily share the summary statistics (effect sizes and p-values) from genome-wide association studies, making it easier to apply univariate pleiotropy detection approaches that combine these statistics to identify SNPs or loci with a concordant or discordant direction of effects.

Therefore, in this project, I first compared the relative power and false-positive rate (FPR) performance of five univariate pleiotropy detection approaches, classic meta-analysis, cFDR, PLACO, ASSET, and CPBayes (see section 6.1), through simulation studies. After that, I applied the best-performing method to the analysis of phenotypes of epilepsy using actual data. The data simulation procedure was performed in 3 steps. First, a population of 1 million individuals of European ancestry was simulated via resampling using the HAPGEN2 software[13] and haplotypes of central Europeans from the 1000 genomes project[14]. In the second phase of the simulation, disease SNPs were randomly

selected and used for the additive liability threshold model (ALTM)[15] to simulate multifactorial disease phenotypes from the simulated genetic data.

As expected, the performance of the methods varied in terms of power and false positive rate (FPR). The variability between the methods is higher for FPR, while most methods are comparable in terms of power, especially for larger sample sizes and RR. Although the classical meta-analysis is very powerful, it is also riddled with a very high false-positive rate, making it less suitable for identifying pleiotropic loci. While all the methods performed well in terms of power, the ASSET method gave a better trade-off between power and FPR for the different simulation approaches. Applying ASSET to the two phenotypes of epilepsy, GGE and FE, resulted in identifying a new putative locus 17q21.32 while replicating locus 2q24.3, previously reported by the ILAE consortium [16]. Further, applying the ASSET method to summary statistics of larger samples of epilepsy phenotypes resulted in the identification of loci 2q24.3 and 9q21.13. These findings corroborate the result obtained by the ILAE consortium through mega and meta-analysis.

Classical meta-analysis (MA) is not recommended for pleiotropy detection, based on the simulation study results. Though MA demonstrated good power to detect pleiotropy, it also recorded high FPR across all simulation scenarios. However, the ASSET method is highly recommended as it kept the FPR low while demonstrating good power to detect pleiotropy. This study also contributed three new pleiotropic loci (2q24.3, 17q21.32, and 9q21.13) to understanding the relationship of genetic variation with epilepsy phenotypes and the inter-relationship between these phenotypes. Although the locus 17q21.32 could not be replicated in the larger sample set, it is not necessarily a false positive discovery. The locus was genome-wide significant for GGE but marginally significant for FE, which confirmed the trend observed in the FE cases in the EPI25 collaborative dataset, where no genome-wide significance result was found. Therefore, replication in an independent sample is desirable.

One limitation of using the univariate pleiotropy detection approaches as seen with the classical MA is that one trait with a very low P-value could drive the observed pleiotropic association. Also, methods like cFDR and PLACO could only accommodate two traits, though this was not a challenge in this project. Despite these limitations, the presented work established a benchmark of the relative performance of the assessed methods and could also guide researchers in related fields in their future work. This study also contributed to understanding the shared genetic factors between GGE and FE with the expectation that larger sample sizes will lead to more discoveries.

5. Introduction

In sections 5.1, 5.2, and 5.3, I briefly introduce the notion of disease etiology, the impact of genetic variation on disease occurrence, as well as the functional consequences of the genetic variation in individuals and, by extension, the whole population. Sections 5.4, 5.5, and 5.6 describe gene mapping approaches, statistical methods to quantify disease risk, and ways to handle multiple testing challenges arising from these tests. In section 5.7, I extensively review epilepsy and its phenotypes while specifying the phenotypes of epilepsy I used in this work. Sections 5.8, 5.9, and 5.10 describe pleiotropy and available methods for pleiotropy detection, explain the univariate pleiotropy detection approaches used in this project, and the merits and drawbacks of these methods. In section 5.11, I state the objectives of this thesis.

Section 6 describes the methods used to generate the simulated data and handle the actual epilepsy samples. In section 6.1, I explain in detail the data simulation steps and the identification of pleiotropy in the simulated data, while in section 6.2, I describe the epilepsy datasets and their sources, quality control checks, and the application of the method from the simulation study to the actual dataset. Section 7 contains the main results of the analyses. In section 8, I present my publications, my contributions to the publications included in this thesis, and outline the courses and meetings I attended during the Ph.D. program. Finally, section 9 gives a detailed discussion of the project, its limitations, and possible future work.

5.1. Etiology of diseases

The question of causality, spread, and progression of diseases is an important and complex topic. Factors predisposing individuals to disease are known as risk factors. These risk factors include biological, genetic, dysregulation of immune- or central nervous systems, or environmental factors, such as stress, trauma, and drug reactions[17]. Genetic diseases are classified into rare (monogenic/oligogenic), polygenic (complex), or chromosomal based on the underlying genetic defect [18]. Rare, Mendelian diseases have a single known genetic cause, while common (complex) diseases result from multiple genetic factors and their interaction with environmental factors. Chromosomal diseases result from large structural variations of large chromosomal segments, in some cases even the absence of whole chromosomes or polyploidies. The main goal of genetic studies is to identify and determine the contribution of genetic variation to disease risk by examining variations in the genomes of affected and un-affected individuals. It is well understood from theoretical considerations and confirmed by studies such as the 1000 Genomes Project[14] that human DNA varies widely among healthy individuals, that no particular DNA sequence can be considered “normal" and that some regions of the DNA are highly conserved with inadequately known functions. However, to make genetic variation comparable and quantifiable among individuals, the Human Genome Project has developed a reference sequence of the human genome that serves as the basis for comparing and describing changes in the DNA sequence[19]. The reference genome serves as a reference frame that enables to describe genomic variation in terms of base-pair positions and alleles, which enables the comparison of genomic variants across individuals, for instance, in case-control studies. Although the human reference genome has been improved over the years, the current version (GRCh38) sequence does not completely cover the whole human genome sequence. This led to the current effort by the

Telomere-Telomere Consortium (T2T), resulting in a completely gapless human genome sequence, still in the early adoption phase[20].

5.2. Genetic variation

Understanding systematic variation in DNA structure and function in the human genome is critical to understanding genetic disease processes[19] because the human DNA sequence contains the information that encodes and regulates biological processes. Many of the naturally occurring variations have been shown to have functional consequences, with approximately 90% of genetic variants in humans falling into the single nucleotide polymorphisms (SNPs) class[21]. Other forms of variation include structural variations such as insertions, deletions, and repeats. In this work, the main focus is on SNPs. SNPs are occurrences of different nucleotides (alternative alleles) at a particular position (locus) in individuals or on different chromosome copies of individuals in the population. Polymorphisms result from random mutations and potentially contribute to the susceptibility of diseases and other traits in humans [22]. Since SNPs frequently occur in the genome (1 in 300 bp on average)[19], they are often used as genetic markers to identify disease-causing genes. Different functional consequences result from genetic variations in individuals based on the location and specific alleles of the polymorphisms.

For practical purposes, genetic variants are often classified into rare and common variations based on the allele frequency in a given cohort or population. Different allele frequency thresholds are used in literature; for example, genetic variations with the frequency of less than 1% in the population can be called rare variations, while a frequency of the allele of more than 1% in the population can be classified as common variation[19] (see Figure 2). Due to the effects of purifying selection, common variants usually have rather small effect sizes. Though this is valid for most common diseases, moderately high effect sizes have been observed for identified genetic variants in the *APOE4* and *LOXL1* genes predisposing individuals to Alzheimer's disease and exfoliative glaucoma respectively[23,24]. Nevertheless, common variants, though not highly deleterious, play significant roles in common diseases because of the following reasons: risk alleles can have small effects on reproductive fitness, moderately deleterious alleles can rise to moderate frequencies, and multiple common variants can confer a higher disease risk through aggregating effect, some neutral or advantageous alleles may begin to confer susceptibility, and some beneficial phenotype may offset disease burdens when disease-causing alleles at high frequency are under balancing selection[25,26].

5.3. Functional consequences of polymorphisms

The DNA sequence consists of four nucleotide bases (A, C, T, G) on one strand and a complementary sequence on the other. Protein-coding regions of the DNA sequence are transcribed into messenger RNA (mRNA), which encodes the information needed for protein synthesis. The mRNA produced during transcription directs translation, which leads to protein synthesis. However, the mRNA contains four bases (A, G, C, and U) organized into triplet codes (codons) representing amino acids which are building blocks for proteins, as well as the start and stop codons. The human genome consists of protein-coding regions, regions encoding regulatory RNA and many other functional elements, and regions with presumably no or unknown functional significance. The protein-coding genes contain exons (protein-coding sequence) and introns (the non-protein-coding sequence). Some RNA-coding sequences are involved in the regulation or expression of other genes. Among the regions that were

previously thought to be non-functional are pseudogenes, which are now known to promote recombination events, especially those that code for similar sequences as protein-coding genes[19].

The effect of SNPs on gene function depends on their position in the gene region, i.e., the coding or the non-coding region, and other factors such as their particular effect on the amino acid sequence of the protein product. The alteration of a single nucleotide by substituting the coding region can lead to different functional consequences. Substitutions that lead to the same amino acids, by extension, of the same protein sequence are referred to as silent or synonymous. Missense or non-synonymous mutations result from substituting genetic codes that lead to a change in the protein sequence such that the function of the original protein is altered[19]. Another possible consequence of the substitution of a single nucleotide is a nonsense mutation, which results in an amino acid being converted to a stop codon, leading to an abrupt truncation of a polypeptide chain or sequence. Some non-coding regions of the genome are known to be control regions that direct cell regulation and gene expression, but the functions of most intragenic and intergenic SNPs are still unknown[27]. Understanding these consequences of polymorphisms is critical to the understanding of the function of identified associated SNPs in GWAS.

5.4. Linkage, linkage disequilibrium, and GWAS

Due to the recombination events in DNA, SNPs in physical proximity are non-randomly linked together and co-transmitted from generation to generation, a phenomenon known as linkage[26]. Linkage disequilibrium is the nonrandom association of alleles at different loci, which results from linkage but can be influenced by mutation, genetic drift, and other factors[28]. Both concepts are essential for mapping genes to diseases and understanding the joint evolution of a linked set of genes, used in linkage analysis and GWAS. Linkage analyses are family-based studies that are performed to identify rare variations causing monogenic diseases, historically, through linkage and positional cloning and later exome and whole-genome sequencing[29,30]. Although linkage studies have been performed for common diseases, they often lack statistical power to detect common variants. GWAS are more powerful and preferable for studying complex diseases.

GWAS assesses SNPs throughout the genome in a case-control cohort to identify alleles associated with a disease. It relies on LD throughout the genome since a variant at one locus can predict the genetic variance at the adjoining loci[29]. The fundamental basis of GWAS is the common disease/common variants (CD/CV) hypothesis, which implies that common variations may contribute to the susceptibility to common diseases[29]. However, the concept is valid for some diseases with well-known etiology and simple allelic spectra but does not explain the total genetic variability in most complex diseases. Other researchers posited the hypothesis of common diseases/rare variants (CD/RV), which explained that rare variations with moderately sized effects also contribute to susceptibility to common diseases[31,32]. Both hypotheses though contrasting, have been found to overlap because studies over time showed that multiple common variations with low penetrance and multiple rare variants with moderate to high effect contribute to susceptibility to common diseases and their frequency in the human population[32]. For example, complex diseases such as epilepsy may be due to a wide range of factors, from rare variants with strong effects to relatively rare variants with moderate effects and common weak variants [33] (see Figure 1).

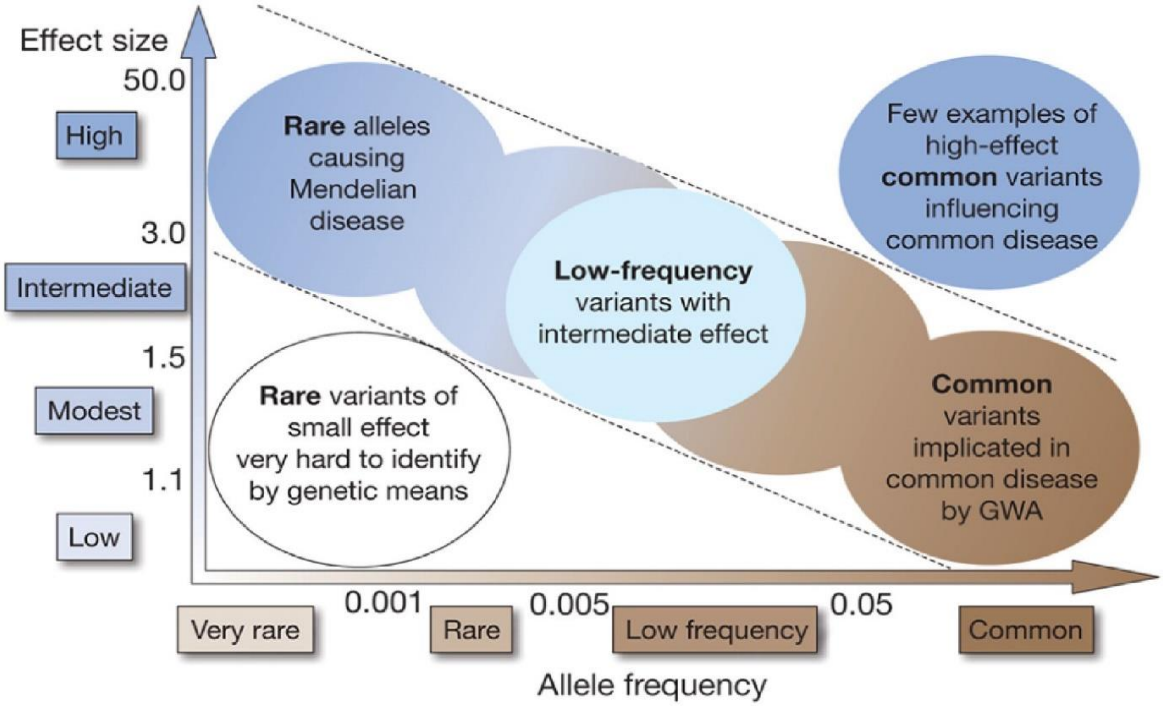


Figure 1: (Adapted from: Arnar, D.O. and Runolf, P. Genetics of common complex diseases: a view from Iceland. Copyright ©2017 European Federation of Internal Medicine. Published by Elsevier B.V.)[1]. The relationship between variant effect size and rare versus common variations here shows that the CD/CV and CD/RV hypotheses are partially correct and overlap for most common diseases.

5.5. Study design, association tests, and heritability

The risk of a disease is the likelihood or probability of an individual in a specified population developing the disease in a specified time, often referred to as incidence proportion. The risk could arise from genetic or environmental contributions to the disease occurrence. In epidemiological studies, it is often desirable to quantify the proportion of people who develop a disease over a specific period (incidence rate, I_r) or the proportion of people with a disease at a given point in time (prevalence, P)[34].

$$Incidence\ proportion / risk = \frac{No.\ of\ onset}{baseline\ population\ at\ risk} \quad (1)$$

$$Incidence\ rate\ (I_r) = \frac{No.\ of\ onset}{population - time\ at\ risk} \quad (2)$$

$$Prevalence\ (P) = \frac{No.\ of\ cases}{Total\ study\ population} \quad (3)$$

Observational study designs are often used to quantify exposures in a population and generate inferences about disease prevalences and incidences in that population. These studies are sometimes descriptive, prospective, i.e., they are conducted forward in time or retrospective (historical), an analysis conducted back in time[35]. Prospective studies usually take the form of cohort studies, where

a group of people with specific unique characteristics are followed over time to evaluate specific outcomes. Case-control studies are the main form of retrospective studies, often used in comparing risk factors in affected individuals (case group) with unaffected individuals (control group)[36]. A study design with a single time point that often estimates the prevalence in the present is called a cross-sectional study[37]. In most epidemiological studies, we compare observed differences within the exposed group in the general population or different population groups in relation to exposure to a risk factor under study.

GWAS are case-control studies identifying the association between genetic variants and phenotypic traits. It identifies SNPs for which the allele frequency varies systematically as a function of the phenotype between cases and control[38]. An allelic association (c and C) or genotype association (c/c, c/C, and C/C) test is performed based on how the genetic markers are represented, for a SNP with a minor allele c and a major allele C. There are four standard models that quantify the relationship between the genotype and phenotype, namely, multiplicative, additive, common recessive and common dominant models[39]. Given the disease penetrance (r), the recessive model requires two copies of c alleles for increased risk, and the dominant model requires one or more alleles C for increased risk. In the multiplicative model, the risk is r^2 for the CC genotype, while for the additive model, there is r-fold and $2r$ increase in risk for cC and CC, respectively [39,40]. The choice of which model and, by extension, the association test to use depends on the assumptions on the underlying inheritance patterns. The multiplicative model is allele-based and often used for binary phenotypes, while the additive model is most generally used for the genotypic association as it has reasonable power to detect additive and dominant effects[41].

Different statistical tests are used for association tests depending on the type of traits, quantitative or qualitative, to be analyzed. For quantitative traits, the genotypes serve as predictors, and linear models such as ANOVA are used. Binary case-control traits are often analyzed using contingency table methods or the logistic regression model[41]. The logistic regression model extends the linear regression model by transforming the binary outcome using the logit link function to predict the probability of having a case status given the genotype, as shown below:

$$logit(\pi) = \ln(\frac{\pi}{1-\pi}) = \beta_0 + G\beta_G + X\beta_x, \quad (4)$$

where π is the probability of affection for the vector of outcome (*Y*), β_0 is the intercept, β_G is the vector of effect sizes for genotypes G, and β_x is the vector of effect sizes for covariates X. The odds ratio is estimated by exponentiating both sides of the equation (4) above.

Contingency table methods such as χ^2 test of independence, Fisher, and likelihood ratio tests are also available for association tests. Extended statistical models are used, particularly if correction for confounding variables is required, such as population structure, environmental effects, family relatedness, and other epidemiological and clinical variables such as gender and age. Linear or generalized linear mixed models are more useful for testing the genetic association while accounting or controlling for confounding variables. In my simulation study, I used an extension of the additive model to estimate the genetic effect and define the traits and a logistic regression model to perform association (see section 6.1.1).

5.6. Statistical hypothesis testing and multiple comparisons

Since each SNP is tested independently, each test gives a P-value, and its significance is tested individually based on a pre-specified significance threshold, which is the permissible type 1 error. Specifically, it is the probability of rejecting a null hypothesis of no association if the null hypothesis is indeed true. For example, if the significance threshold for each test is 0.05, on the average, 5% of independently tested SNPs are expected to give false-positive results assuming all null hypothesis is true. However, the number of false positives will increase with increasing numbers of tests, requiring the need to set the significance threshold to a lower value in order to reduce the number of false positives, in other words, methods for multiple testing correction are used. The most commonly used methods in GWAS are based on controlling the so-called family-wise error rate (FWER) [42].

The FWER is the probability of making one or more type 1 errors in a set of tests [39]. Bonferroni correction and Šidak correction are two common forms of FWER control that yield similar results if the number of tests is sufficiently large. The Bonferroni correction estimates the significance level per test (α_p) as the ratio of the FWER (α) and the total number of tests (m); $\alpha_p = \alpha/m$ while Šidak correction is $\alpha_p = 1 - (1-\alpha)^{1/m}$. However, Bonferroni and Šidak corrections are conservative and, for GWAS, lead to an increase in false-negative rate[42], since both assume that each SNP is independent, whereas due to linkage disequilibrium, there is a high degree of correlation between neighboring SNPs.

The FDR approach controls the expected proportion of false positives among all associated SNPs declared significant[39]. For example, for the Benjamini-Hochberg FDR procedure, the P-values of all SNPs tested are assigned ranks (i), and a global significance level (α) is chosen. Then local FDR for each rank is computed as: $FDR_i = \alpha(i/m)$ and the null hypothesis is rejected for P-values lower than FDR_i .FDR is not optimal because of LD between markers and the small numbers of expected true positives the method yields. FDR procedures do not provide a notion of significance but correct for the number of expected false discoveries hence providing an estimate of the number of true results among those called "significant"[41].

Permutation testing is another approach for establishing significance in GWAS[41]. It compares the obtained P-values of association with the empirical distribution of P-values obtained for the case-control identifiers[43]. However, it is computationally expensive, especially with increasingly large numbers of tests. Based on the distribution of LD in the genome of specific populations, the concept of genome-wide significance was derived. The 'effective' number of independent genomic regions in a population, thus the number of statistical tests that should be corrected for being determined[41]. Due to limitations of correcting procedures, a genome-wide association P-value threshold of 5×10^{-8} for rejecting the null hypotheses for common diseases in the European population was estimated using FWER methods[44]. For FDR procedures, the recommended cut-off value is between 10^{-6} and 10^{-8}[4,5].

5.7. Complex disorders

5.7.1. Mendelian and complex diseases

As mentioned in the introduction, diseases can be unifactorial or multifactorial. Unifactorial diseases are those that are known to be caused by single genomic variations. These monogenic diseases are called Mendelian disorders due to their inheritance patterns, based on Mendel's laws of segregation,

independent assortments, and dominance[45]. They are primarily classified as rare disorders because the coding genes harbor highly deleterious mutant alleles that occur at low frequency in the general population. This definition may not be entirely accurate for all rare disorders, as shown by previous studies, since some rare diseases do not harbor nonsense mutations and indicate that deleterious alleles may also be located outside the coding sequence of the gene[46]. Examples of monogenic diseases include Huntington's disease, achondroplasia, and cystic fibrosis. Methods for identifying Mendelian disease genes include positional mapping and sequencing, especially linkage studies in families and exome sequencing approach[47].

On the other hand, complex, multifactorial disorders are not caused by a single genetic risk factor. Instead, multiple genetic and environmental factors and their interactions might contribute to disease risk. Risk loci might be distributed throughout the genome, with some contributing to more than one trait (pleiotropy)[48-50]. Common examples of complex diseases include type 2 diabetes (T2D), epilepsy, hypertension, and asthma[51]. From our understanding that most common diseases are multifactorial, it is essential to decipher the sources of risk and their relative contributions to disease occurrence. The relative contribution of environmental and genetic factors can be very important in understanding disease susceptibility.

In genetics, scientists assess heritability (*H*) as the proportion of variation for a given disease in a population attributable to genetic factors[52]. Heritability is broadly divided into broad-sense (H^2) and narrow-sense (h^2). Narrow-sense heritability quantifies the proportion of variation due to additive genetic effects while broad-sense heritability (H^2) captures the proportion of phenotypic variation due to genetic factors, including allelic interaction, gene-gene interaction, within loci (dominance) and between loci (epistasis), and gene-environment interactions[52,53].

The narrow-sense heritability (h^2) is often estimated in polygenic additive liability models for estimating the heritability of common diseases with the assumption that common disorders are genetically homogenous, dominance and epistasis are negligible in the disease etiology, and that neither a genetic nor an environmental factor has a major contribution[54,55]. These models have been debated over the years, but the conclusion from empirical data shows that they are consistent for common diseases[56,57] (see section 6.1.1).

GWAS have identified variants associated with many traits. However, most of these studies explain only 5-10% of the heritable component of the disease, which means that the larger part of the heritable component cannot be explained by GWAS alone (missing heritability)[32]. Missing heritability may be due to undiscovered large numbers of variants with small effects, poorly detected rare variations contributing to common diseases, inability to detect gene-gene interactions, and improperly considered environmental factors, among other possible explanations[58]. For example, large sequencing studies have shown that relatively rare variations may play a significant role in the heritability of common epilepsy but are typically left out in GWAS due to the focus on variants present in 5% or more of the population[33,58]. Furthermore, the knowledge of one genetic marker or gene affecting more than one trait (pleiotropy) is also gradually improving the discovery of associations, especially for complex trait phenotypes exhibiting cross-phenotype associations. This project is focused on epilepsy as a complex disease and examined the association of two well-characterized phenotypes of the disorder with genotypic variants (details in sections 6.2 and 9)

5.7.2. Epilepsies as complex diseases

Epilepsy is a chronic disease of the brain characterized by an enduring (i.e., persisting) predisposition to generate seizures, unprovoked by any immediate central nervous system insult [59]. It results from a number of nerve cells in the brain sending abnormal signals, causing seizures. It is worth noting that not all people who experience seizures have epilepsy, but the seizures must be recurrent or have a likelihood of recurrence to be called epileptic seizures[60,61]. The diagnosis of epilepsies is based on different clinical symptoms observed together in an individual, imaging findings, and age of onset. The burden of epilepsy is relatively high compared to other brain diseases[62]. Giourou et al. reported that epilepsy affects about 1% of the world's population, and about 10% will experience a seizure in their lifetime[63].

The prevalence of epilepsy varies from country to country and depends on sociodemographic, risk-related, and etiological factors. However, the global average is thought to be 7.6 per 1000 people with the condition, with a slightly higher prevalence of 8.75 per 1000 in low and middle-income countries and a slightly lower prevalence of 5.18 per 1000 in high-income countries[64]. A recent meta-analysis in Latin America and the Caribbean found an overall higher prevalence of 14.09 per 1000 residents and a prevalence of 9.06 per 1000 individuals for active epilepsy[65]. The overall prevalence of active epilepsy in Nigeria (Africa) is 9.8 per 1000 but varies from north to south[66], while the prevalence in the European Union varies widely from country to country but is lower compared to reports from Africa and Latin America. Early work on the genetics of epilepsy via twin studies showed that there are genetic risk factors for the disorder[67-69]

There are different forms of epilepsy, including both monogenic and polygenic forms. The monogenic form of the disease is marked by rare variations with large effects, while the complex epilepsy form results from small but aggregating effects of common variation, some rare variations with small to moderate effects, and environmental factors[33]. The ILAE Consortium[70] has categorized epilepsy based on seizure type, imaging findings, age of onset, and other clinical findings, such as co-occurrence of different symptoms. Focal, generalized, and unknown onset are the categories of epilepsy based on the form of the seizures (see Figure 2). Focal onset seizures are characterized by seizures that originate from and are confined to one part of the brain. Generalized seizures originate from one source but spread throughout the brain network, while the type of seizures with unknown onset is undifferentiated. Focal onset seizures are further subdivided into seizures with consciousness or impaired consciousness, while generalized seizures are classified as motor seizures, including tonic-clonic seizures and other motor seizure forms or non-motor seizures, also known as Absence[60].

Depending on the clinical diagnosis and EEG findings, epilepsy types are divided into focal, generalized, combined generalized and focal, and unknown epilepsies (see Figure 2). In generalized epilepsy, the affected person shows generalized spike-wave activity on the electroencephalogram and may have absence, myoclonic, atonic, tonic, and tonic-clonic seizures[60,70]. Focal epilepsy includes both unifocal and multifocal disorders and seizures in one hemisphere of the brain with one of the following seizure types: focal conscious seizures, focal seizures with impaired consciousness, focal motor seizures, focal non-motor seizures, and focal to bilateral tonic-clonic seizures[70]. The combined generalized and focal phenotypes of epilepsy were newly introduced by the ILAE consortium for patients clinically diagnosed with both focal and generalized seizures[70].

Mendelian or monogenetic epilepsy disorders are rare, affecting less than 1 in 2000 people. They are usually developmental and begin in early childhood, causing severe impairment in those affected. Most forms of rare epilepsies are called developmental and epilepsy encephalopathy (DEE). Well-known examples include Dravet syndrome, Ohtahara syndrome, West syndrome, Lennox-Gaustat syndrome, and infantile spasm[71]. Common epilepsy phenotypes affect about 1 in 200 people and are broadly classified as genetic generalized epilepsy (GGE) and focal epilepsy (FE). GGEs account for 15-20% of all epilepsies[72], while FE is responsible for about 60% of all epilepsies[73]. Focal epilepsies were initially thought to be acquired only through trauma, infection, and other non-genetic causes. However, genetic studies have identified variants predisposing to some form of FE, hence the distinction as non-acquired focal epilepsies (NAFE). Typical forms of GGE syndromes include childhood absence epilepsy (CAE), juvenile absence epilepsy (JAE), juvenile myoclonic epilepsy (JME), and generalized tonic-clonic seizures.

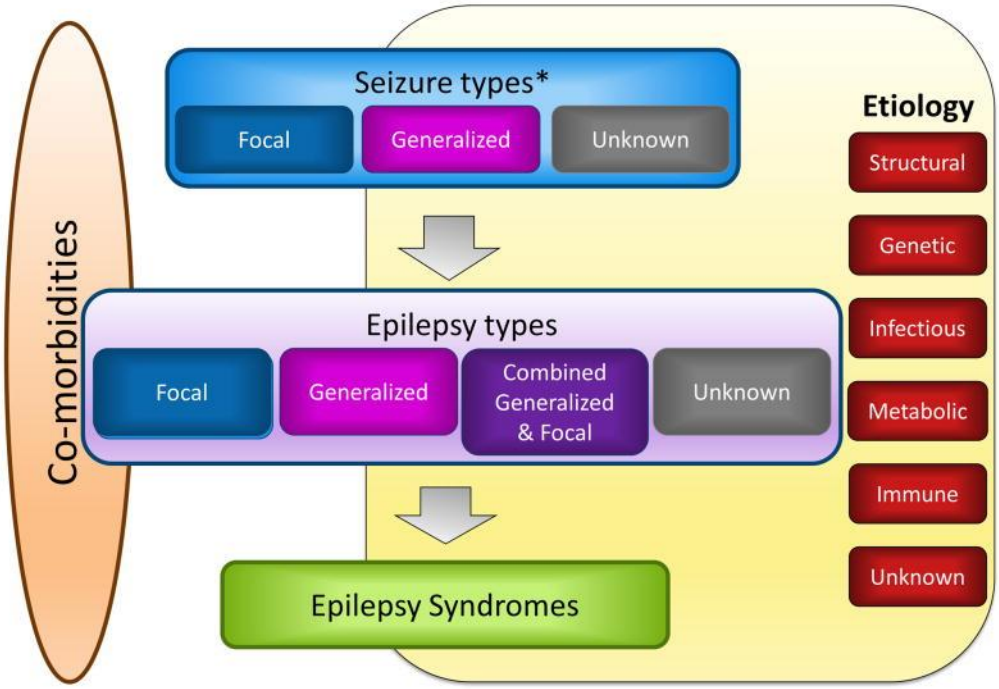


Figure 2: (Adapted from: Scheffer et al. ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. Copyright ©2017 International League Against Epilepsy). Forms of epilepsy are based on seizure types, epilepsy types, and syndromes. Epilepsies have also been found to be co-morbid but phenotypes based on onset seizure (*) are still the most well-categorized.

The causes of epilepsy are diverse, ranging from structural, metabolic, infectious, immune, and unknown to genetic. These factors are intertwined to a large extent; for example, some structural abnormalities or metabolic disorders already associated with epilepsy could be acquired or genetic. Even in epilepsies resulting from trauma such as head injury or stroke, it has been shown based on studies of families that genetic factors still contribute to the observed trait[74,75]. However, genetic causes do not immediately translate to inheritance. Hence it is necessary to separate monogenic forms of epilepsy from complex epilepsies. Thus, epilepsies comprise rare monogenic phenotypes and common forms, widely referred to as complex epilepsies, with both common and rare-variants

contributions[71]. To understand the genetic risk factors in the susceptibility to epilepsies, comparative twin studies have been performed between monozygotic twins and dizygotic twins through linkage studies to compare disease concordance[67-69]. Assuming the twins share the same environment, differences observed in disease occurrence are likely not by chance and are attributable to genetic variation[33], raising the need to focus on unraveling the genetic architecture of diseases.

From all indications, there is a strong genetic basis for the inheritance of epilepsies, but only a fraction of the trait variation due to genetic factors is currently known through GWAS and sequencing-based analysis for common epilepsy forms. Since there are overlaps of the forms of epilepsies further corroborated by the new classification from the ILAE consortium, methods that can identify these shared or switch-like variants are essential to advance gene discovery and provide more information on the genetic basis of epilepsies. One of such valuable approaches is pleiotropy analysis which allows for joint analysis of samples from different disease traits, in this case, epilepsy phenotypes, GGE and FE.

5.8. Introduction to pleiotropy

Hereditary units like SNPs, loci, or genes combined with environmental factors determine the physical characteristics of organisms and humans. Over the past decades, enormous work has been done to link diseases to genetics, with or without accounting for environmental factors. One of the largest and still growing types of such studies, GWAS, has focused on identifying single locus-trait relationships, which has produced robust results in complex diseases. There has also been evidence of a single hereditary unit being associated with more than one trait[45,76]. For example, a single disease risk factor was shown to have multiple symptoms[77]. This phenomenon of one single unit affecting two or more phenotypes is often called pleiotropy, but it has not been well defined in the early days of genetics. “Cross-phenotype association” is a general term used to describe the correlation of a marker, gene, or genetic region with multiple traits, regardless of the underlying mechanism of correlation between the markers and the traits[49]. Ludwig Plate (1910) coined the term pleiotropy ("Pleiotropie"), defined as the phenomenon of a hereditary unit affecting more than one trait[12]. Various studies[12,49,77-79] since then have dissected and categorized pleiotropy into distinct meaningful forms.

In the modern understanding of pleiotropy, it is broadly categorized into three forms, namely biological or horizontal pleiotropy, mediated pleiotropy, and spurious pleiotropy. Although spurious pleiotropy is basically the result of bias from different sources, it is often mentioned as a form of pleiotropy to guide researchers in interpreting their results. In the biological or horizontal form of pleiotropy, one or more variants in the same genomic region are associated with more than one trait (see Figure 3). This form of pleiotropy could be at the allelic or genic level. At the allelic level, an associated genetic maker could be in LD with one or more causal variants in the same gene that simultaneously affect different traits. At the genic level, markers in the same gene are in LD with different unobserved causal variants which independently affect two or more traits. For example, the SNP rs6983267 in the intergenic region of chromosome 8q24 is a risk variant for prostate and colorectal cancer[49]. In the case of mediated pleiotropy, the correlation of a variant to one trait that causally predicts another trait leads to the marker appearing to be associated with both traits (see Figure 3). For example, the *CHRNA5* gene is known to be associated with lung cancer, chronic obstructive pulmonary disease (COPD), and smoking behaviors. However, the association with lung cancer could be either due to the effect of the gene variants on smoking intensity or indirectly through

effects on COPD[80]. Spurious pleiotropy arises from different forms of bias such as phenotype misclassification errors, overlapping controls cohort in independent studies, and high LD in regions, leading to marker tagging variants in different genes[49,50] (see Figure 3).

More recent examples of pleiotropy include the identification of loci 14q12-q23 and 12q24.2-q24.3 as shared risk factors between migraine and epilepsy through linkage analysis[81]. Another notable example is the variation in the calcium channel activity genes such as *CACNA1C* and *CACNA1D,* which were reported to be pleiotropic for psychiatric disorders like schizophrenia, bipolar disorder, and major depressive disorder [82]. Overall, the study of pleiotropy is helpful and comprises a promising set of methods for disentangling causal relationships and genetic architectures in complex, multifactorial diseases, as exemplified in the current work, in which I apply pleiotropy analysis methods to well-categorized phenotypes of epilepsy.

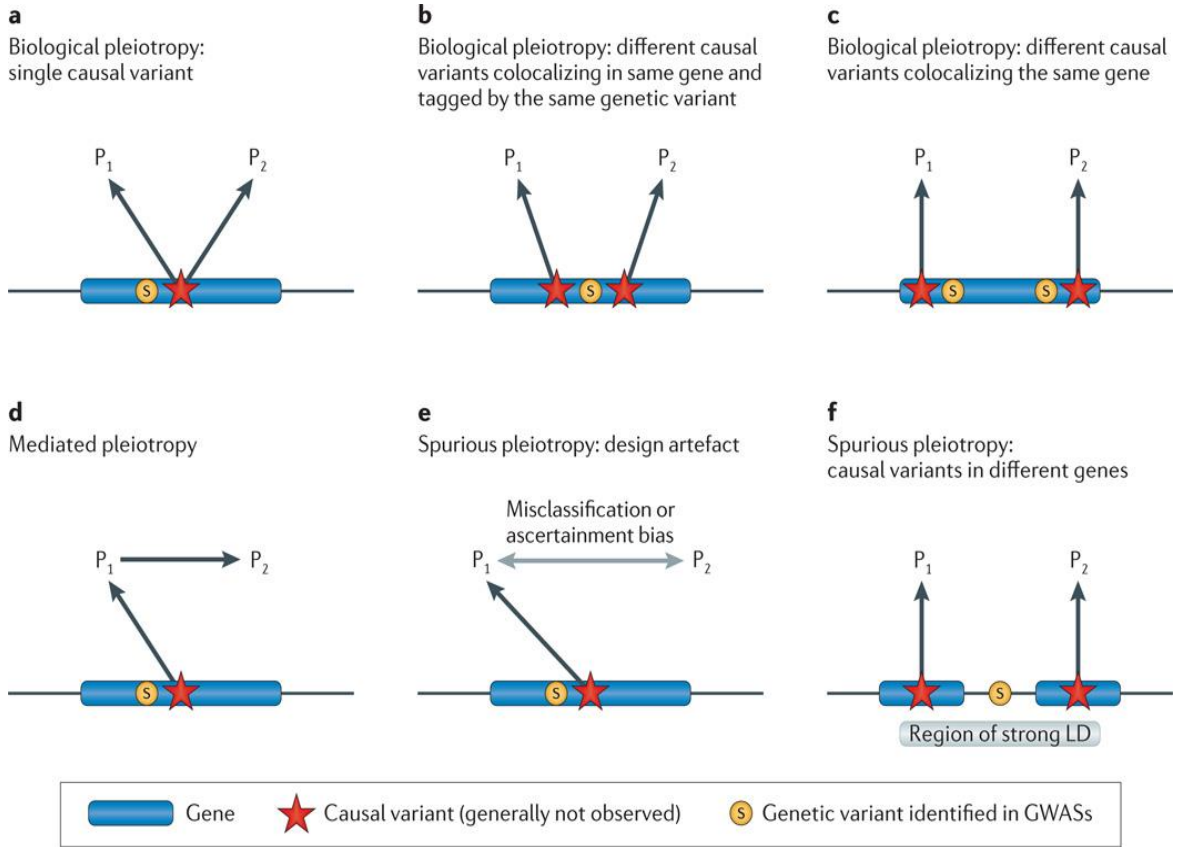


Figure 3: (Adapted from: Solovieff et al. Pleiotropy in complex traits: challenges and strategies. Copyright © 2013 Macmillan Publishers Limited). Types of Pleiotropy. Biological or horizontal: Genetic units exert their effects through one or two colocalizing variants associated with two traits (a, b, c). Mediated: a trait causally related to another trait, thereby a single variant appearing to be associated with both traits (d) or spurious: relationships due to different forms of bias (e, f).

5.9. Available methods for pleiotropy detection

Pleiotropy detection methods can be classified into genome-wide, regional, or single variant-specific based on the level at which overlap of variants with traits is assessed. Genome-wide approaches are

only available for the simultaneous study of multiple traits. At the regional level, markers are grouped into genetically meaningful LD blocks and analyzed as sub-groups. In the statistical sense, these methods are broadly classified as univariate and multivariate approaches. The univariate methods directly quantify the effect of the variant on an outcome, which can be a trait or, in the case of pleiotropy, a parameter that represents the combination of effects from individual traits analysis, while the multivariate methods jointly test the association between variants and two or more traits in measured simultaneously in the same individual. The application and choice of methods are guided by data availability, of particular importance is the question of whether individual-level or summary statistics are available, the number of traits co-measured, and whether some samples are shared between studies (see Figure 4). Both multivariate and univariate pleiotropy approaches identify cross-phenotype (CP) association. It is worth noting that pleiotropy analysis provides statistical evidence of pleiotropy which can be followed up by mapping, phenotype stratification, and further study of molecular mechanisms of the diseases, which is useful in clarifying the type of CP association identified.

The availability of individual-level genotype and phenotype data allows for the use of multivariate pleiotropy detection approaches. One commonly used genome-wide pleiotropy detection method is the polygenic risk score (PRS). PRS combines polygenic effects across loci to check for association or predict risk and can be used for pleiotropy detection[83]. One approach to identify pleiotropy using PRS is by constructing genetic risk scores using effect estimates of markers selected from GWAS in a sample for each individual in another independent sample,[68]. An association of the score to the trait of interest in the second sample is evidence of an overlap between the genetic factors of both traits[82,84]. Other noteworthy sets of multivariate genetic correlation approaches for identifying shared loci at the genome-wide level are implemented in GCTA[85,86], BOLT-REML[87], and multivariate linear mixed model (mvLMM)[88]. These methods can accommodate continuous and binary variables except for mvLMM, which only allows normally distributed dependent variables[50]. The GCTA and BOLT-REML algorithms use restricted maximum-likelihood estimation to compute genetic correlation (rg), which expresses the influence of genetic factors on the covariance of two traits.

Some methods are also available for identifying pleiotropy at the regional level. One popular multivariate approach implemented in the pleiotropic region identification method (PRIMe)[89], bins the entire genome into non-overlapping blocks based on pre-defined criteria such as LD-blocks and gene boundaries[50]. The most significant variant in the block is termed ‘’a driver’’ for all other variants known as “passengers”. This process is repeated until all variants are partitioned into non-overlapping blocks, each containing a driver. Pleiotropy is identified based on some prespecified index in each block[50]. Other gene-based or locus-based multivariate methods such as Bayesian colocalization model[90], canonical correlation analysis (CCA) methods[91,92], and multi-trait set tests(mtSET)[93] are also available.

In summary, there are many available multivariate methods for individual-level pleiotropy analysis, for analysis at single-variant or genome-wide level, for either continuous, binary, or categorical variables. Analysis approaches include multinomial logistics regression[94,95], generalized linear mixed model[96,97], linear mixed-effects models[98,99], generalized estimating equations (GEE)[100,101], frailty models [102], principal components analysis[103,104] and others[105] (see Figure 4). However, data availability is an issue in practice, as most individual-level genotyping studies are performed for studies of single phenotypes. Data sharing among researchers is complex due to regulations and concerns from ethical

considerations. Single-trait GWAS summary statistics are more readily available. These reasons justify the existence of methods for combined analysis of univariate summary statistics from separate GWAS. It also motivates the use of univariate pleiotropy approaches in this project.

Therefore, other valuable methods to harness increasingly available GWAS summary statistics have been developed. Most of the available univariate pleiotropy detection methods stem from the idea of meta-analysis, where separate studies are combined to increase the power of detecting association. The classical meta-analysis method combines effect sizes or p-values of two or more traits to generate a combined effect estimate or p-values for the traits[2,106]. However, the classical MA has major limitations, such as the requirement of very homogenous traits. Another limitation is that classical MA does not account for the directionality of effect, and samples overlap in the case that samples are present in multiple studies. The subset-based metal-analysis (ASSET)[3] approach is an extension of fixed-effect meta-analysis that accounts for sample overlap and effect direction and allows for heterogeneous traits to be jointly analyzed. Other extensions of classical metal analysis are cross-phenotype meta-analysis (CPMA)[107], cross-phenotype association (CPASSOC[108,109]), trait-based association test (TATES), MultiMeta[110], pleiotropic analysis under the composite null hypothesis (PLACO)[6], Multi-TRAIT Analysis of GWAS (MTAG)[111], and pleiotropic locus exploration and interpretation using optimal test (PLEIO)[112]. Some Bayesian univariate pleiotropy detection approaches are also available. The conditional false discovery rate (cFDR), as the name implies, tests a trait called "principal trait" conditional on the second, "conditional" trait [113]. cFDR was further extended to account for overlapping samples[5]. A more recent cross phenotype Bayes (CPBayes) approach computes local FDR and Bayes factors as evidence of overall pleiotropy [4]. In the current study, I applied five of these univariate methods (see Table 1 for an overview), as discussed in the next section.

[a] Extensions for handling overlapping subjects are available for these methods.
[b] Methods based on p-values from single-trait analyses; accommodate genetic effect heterogeneity and any phenotype distribution.
[c] Methods based on effect sizes from single-trait analyses; power affected by genetic effect heterogeneity; phenotypes must be of same distribution.
[d] Denotes a mixture of continuous and dichotomous phenotypes.
[e] Multinomial logistic regression is an indirect method for dichotomous phenotypes.

Web Figure 1. Analytic options for detection of cross-phenotype associations across various types of data. The choice of analytic method in genetic epidemiological studies of pleiotropy largely depends on the types of available phenotypic data. When relying on secondary data to study pleiotropy, we may find that subject-level data with all phenotypes measured on the same subjects are not available. In that case, each phenotype may be analyzed separately in different study populations, and the resulting summary statistics can be used to identify cross-phenotype associations by employing univariate methods. When selecting a univariate method, it is important to consider the degree of overlap in the subjects contributing to the analysis of each phenotype, since most univariate methods cannot accommodate overlapping subjects. If subject-level data in which all phenotypes have been measured on the same subjects are available, both univariate and multivariate methods can be employed. When selecting a multivariate method, it is important to consider the phenotype distributions, as in traditional epidemiological studies. We note that choosing among a set of appropriate methods is not necessarily straightforward. Many parameters affect their power and type 1 errors, including the number of phenotypes under analysis, cross-phenotype correlation structure, SNP minor allele frequency, heritability of the phenotypes, and SNP effect size for each phenotype (25, 31). Simulation studies evaluating univariate and multivariate methods across different values of these parameters have concluded that there is no single method or type of method (univariate vs. multivariate) with consistently good performance among all possible scenarios (25, 31). For this reason, we suggest that, when possible (i.e., when subject-level data with all phenotypes measured on same subjects are available), both univariate and multivariate methods be employed in parallel to increase the probability of detecting cross-phenotype associations.

Abbreviations: CCA = canonical correlation analysis; CPMA = Cross-Phenotype Meta-Analysis; FDR = false-discovery rate; EGEE = extended generalized estimating equations; GLMM = generalized linear mixed effects models; LME = linear mixed effects models; META = meta-analysis; PCA = principal components analysis; PCHAT = Principal Component of Heritability Association Test; PET = Pleiotropy Estimation and Test; Seq-LRT = sequential likelihood-ratio test; TATES = Trait-based Association Test that uses Extended Simes; UV = univariate.

Figure 4: (Adapted from: Salinas Y. D., Wang Z., DeWan A. T. Statistical Analysis of Multiple Phenotypes in Genetic Epidemiologic Studies: From Cross-Phenotype Associations to Pleiotropy. Copyright © The Author(s) 2018). Classification of available pleiotropy detection methods based on available data, the outcome, and samples overlap across studies.

5.10. Description of applied univariate approaches

From the many available methods discussed in the previous section, I selected five recent, well-implemented univariate pleiotropy detection approaches briefly described in the following sections using the simulated dataset to identify the best methods in terms of power and false-positive rate. These methods are meta-analysis-based in that they extend the framework of classical meta-analysis of combining effects sizes or P-values from independent GWAS, accommodate varying sources of heterogeneity, and allow for sample overlap. Some of the methods even generate inference via Bayesian sampling approaches.

Method	**Abbreviation**	**Reference**	**Web resource**
Classic fixed-effect meta-analysis	MA	[2]	http://csg.sph.umich.edu/abecasis/metal/download/
Subset-based metal analysis	ASSET	[3]	https://bioconductor.org/packages/release/bioc/html/ASSET.html
Cross-phenotype Bayes	CPBayes	[4]	https://github.com/ArunabhaCodes/CPBayes
Conditional false discovery rate	cFDR	[5]	https://github.com/jameslilley/cFDR-common-controls
Pleiotropic analysis under the composite null hypothesis	PLACO	[6]	https://github.com/RayDebashree/PLACO

Table 1: Univariate pleiotropy detection methods included in the analysis and their sources.

5.10.1. Classical fixed-effect meta-analysis (MA)

This approach consolidates results from different studies by pooling the P-values or effect sizes from these studies to estimate an overall effect size. MA is not explicitly designed for pleiotropy detection but can be expected to identify variants that have concordant effects in separate studies of two or more phenotypes and, correspondingly, its use has been demonstrated in pleiotropy detection[114,115]. This method estimates an overall effect from the two phenotypes by computing the weighted mean from the effect sizes weighted by the inverse of the overall study variance. The assumption is that there is a true effect shared by all phenotypes being analyzed, and the difference in observed effect is due to sampling error[116]. This true effect is the estimated common effect. It typically assigns larger weights to a phenotype with more precise effect sizes, meaning that the weights are computed via the amount of information provided from each phenotype, hence, the use of sample sizes.

The basic approach underlying fixed-effect MA is the conversion of study-specific P-values (P_k) and effect direction (δ_k) from *K* studies into standard normally distributed signed Z-scores

$$Z_k = \Phi^{-1}(P_k/2) * sign(\delta_k) \quad (5)$$

which are then combined to estimate an overall Z-score (Z_{meta}) statistic by weights w_k. Weights are typically assigned based on the inverse of the variance, which is also roughly proportional to sample size[117]. Therefore $w_k = \frac{1}{\delta_k}$ and the variance of the combined effect is $\delta_{meta} = \frac{1}{\sum_{i=1}^{k} W_k}$. Thus,

$$Z_{meta} = \frac{\sum_{k=1..K} Z_k W_k}{\sqrt{\sum_{k=1..K} W_k^2}}. \quad (6)$$

The final overall P-value is then obtained by comparing this statistic against a standard normal distribution:

$$P_{meta} = 2[1 - (\Phi(|Z_{meta}|))], \quad (7)$$

where Φ denotes the standard normal distribution function. Here, I used the MA implementation in the METAL software[2].

5.10.2. Subset-based meta-analysis (ASSET)

The ASSET method[3] extends the classical meta-analysis approach by pooling multiple heterogeneous trait effects together and exploring exhaustively various subsets of these traits acting concordantly in the same or different directions. It generalizes the classical fixed-effect meta-analysis by exploring all possible subsets of non-null studies to check for strong association signals. This approach tests the null hypothesis of no association of SNPs in any of the individual traits by estimating the evidence of association for any SNP, Z-statistics *($Z(B)$)* in any given subset (B) of traits. For a given subset *B* of *m(B)* studies, the respective overall Z-score *Z(B)* is obtained following the MA approach by

$$Z(B) = \sum_{k \in B} \sqrt{\pi_k(B)} Z_k \,, \quad (8)$$

where $\pi_k(B) = n_k / \sum_{k=1}^{m(B)} n_k$ weighs the different studies proportional to the square root of respective sample sizes. If covariate adjustments are similar across studies, then $B_k \propto \frac{1}{n_k}$ where n_k is the sample size for the k^{th} study [3]. The score is then maximized over all possible subsets:

$$Z_{meta-max} = max_{B \subseteq \{1,..,K\}} |Z(B)| \,. \quad (9)$$

The overall hypothesis of a genetic marker to be associated with all traits is evaluated by $Z_{meta-max.}$ The upper bound for the *P*-values from the defined multivariate distribution is obtained through the discrete local maxima (DLM) method (see [3] for full details). The aggregate evidence of pleiotropy is at GWAS significant P-value of 5×10^{-8} after correcting for multiple testing using Bonferroni standard procedure.

5.10.3. Conditional false discovery rate (cFDR)

This method leverages the available GWAS summary statistics by estimating the cFDR, which comprises an upper bound on the expected FDR across SNPs having p-values below a set threshold for

both traits[5,118]. As discussed in section 5.6, the FDR controls the expected proportion of false positives among all associated SNPs that were declared significant. Assuming that the P-value of a trait *k* across all variants is a realization of a random variable P_k, the unconditional FDR (uFDR) for the null hypothesis $H_0^{(k)}$ of no association of this variant with phenotype *k* is then defined as the probability that a random variant from this set of rejected hypotheses falls under the null hypothesis for this phenotype [5]. The uFDR can be estimated from a set of observed P-values: $p_k^1, p_k^2, \ldots, p_k^N$ for a set of *N* variants as the ratio of the expected quantile of P_k under $H_0^{(k)}$ and the observed quantile of P_k:

$$\widehat{uFDR_{(p_k)}} = \frac{p_k}{\frac{\#(p_k^i |\ p_k^i \le p_k)}{N}}. \tag{10}$$

However, for cFDR, a trait selected to be the "principal trait" is conditioned on the second trait, "conditional trait," the cFDR is then defined as the posterior probability that a given variant falls under the null hypothesis for the principal phenotype given that the P-values for both phenotypes are less or equal to the observed P-values (p_k, p_l): $P\left(H_0^{(k)} \middle| P_k \le p_k, P_l \le p_l\right)$. Similar to the uFDR and based on observed P-value pairs $\{(p_k^1, p_l^1), (p_k^2, p_l^2), \ldots, (p_k^N, p_l^N)\}$ for two phenotypes *k* and *l* at *N* different SNPs, it is estimated by the ratio of the expected quantile of P_k under $H_0^{(k)}$ amongst those p_k^i where *i* satisfies $P_l^i \le p_l$ and the observed quantiles:

$$\widehat{cFDR_{(p_k|p_l)}} = \frac{P(P_k \le p_k | P_l \le p_l, H_0^{(k)})}{\frac{\#\left(\left(P_k^w, P_l^w\right) \epsilon \left(P_i, P_j\right) | p_k^i \le p_k \text{ and } p_l^i \le p_l\right)}{N_1}} \tag{11}$$

where N_1 denotes the number of P-value pairs with $P_l \le p_l$ and (p_k, p_l) is the P-value pair for a SNP of interest[5]. Suppose controls are shared between the two traits. In that case, there is a positive correlation between the estimated effect sizes for both traits and the distribution of P-values for the principal trait, given that the P-values for the conditional trait depends on the underlying effect of each SNP on the conditional trait; hence, the underlying effect (η) is not known. Itd can be considered as a realization of random variable H. The expected P-value of principal trait, $P(P_k \le p_k | P_l \le p_l, H_0^{(k)})$ is then evaluated by integrating over the true but unknown effects for conditional traits[5]. Association with both phenotypes is tested via a conjunction FDR procedure to minimize the effect of a single phenotype driving the association signal, and an FDR-controlling procedure is used to correct for multiple testing.

5.10.4. Cross-phenotype Bayes (CPBayes)

The cross-phenotype Bayes approach is a fully Bayesian meta-analysis-based approach that generates inference on overall evidence of pleiotropy for two or more traits using Gibb's sampling form of the Markov chain Monte Carlo (MCMC) technique. The aggregate evidence of pleiotropy is given by the local false discovery rate (locFDR) and the Bayes factor (BF) through testing the global null hypothesis (H_o) of no association with *any* trait versus the alternative hypothesis (H_1) of association with *at least one* trait. Prior information is provided by the spike and slab approach, where the spike element represents the null effect while the slab part represents the non-null effect. Let $\widehat{\beta_k}$ be regression estimates of true effect β obtained from the separate univariate models of individual traits T_k and s_k their standard errors. If the sample size is sufficiently large and $\widehat{\beta_k}$ are uncorrelated, we assume that

$$\widehat{\beta_k} | \beta_k \overset{ind}{\sim} N(\beta_k, s_k^2) \ \ (k=1,..,K). \tag{12}$$

However, for correlated estimates ($\widehat{\beta_1}, \ldots, \widehat{\beta_k}$) with variance-covariance matrix S that corresponds to the SNPs, $\hat{\beta}|\beta \sim MVN(\beta, S)$. The prior information is given such that z_k denotes the association status of T_k (see [4], page 22). The local false discovery rate (locFDR) equals the probability of null association (PNA) given the data: locFDR $= P(H_0|D)$. With the posterior odds (*PO*) equalling $PO = \frac{P(H_1|D)}{P(H_0|D)}$ and the posterior probability of association equaling PO/(1-PO), we obtain the posterior probability of null association (PPNA) which is the same quantity as locFDR as:

$$PPNA = 1 - PPA = \frac{1}{1+PO} = P(H_o|D)\ . \tag{13}$$

Also, the Bayes Factor (BF) is obtained by:

$$BF = \frac{P(D|H_1)}{P(D|H_0)} = \frac{P(H_1|D)P(H_0)}{P(H_0|D)P(H_1)} = \frac{P(Z \neq 0|D)P(Z=0)}{P(Z = 0|D)P(Z\neq 0)} = \frac{Posterior\ odds}{Prior\ odds}\ , \tag{14}$$

where the posterior odds are the ratio of the probability of non-null and null effect given the data while the prior odd is the ratio of the probability apriori of the effect being non-null and null, which is estimated from a Dirac distribution or mixture of normal distributions with mean zero and very small variance. locFDR and BF provide the evidence of aggregate pleiotropy such that if BF >1 and locFDR $< 10^{-6}$ the variant is pleiotropic. In addition, the trait-specific posterior probability of association also provides information on the relative strength of association between a pleiotropic variant and the selected non-null trait contribution to the aggregate evidence of association.

5.10.5. Pleiotropy analysis under the composite null hypothesis (PLACO)

PLACO methods test for evidence of pleiotropy by testing the composite null hypothesis of no association with none or only one of the traits as opposed to the testing of the global null hypothesis of no association of the SNPs with any of the traits in the MA approach using the summary statistics from GWAS of individual traits. The null and alternative hypotheses are defined in such a way that the global null hypothesis consists of sub-null hypotheses $H_{01}$and H_{02} where H_{01}: $\beta_1 = 0,\ \beta_2 \neq 0,\ H_{02}$: $\beta_1 \neq 0,\ \beta_2 = 0$ for both traits. $\beta_1$and β_2 are the genetic effect of the variants for the first and second trait respectively, H_{01} is the sub-null hypothesis that the genetic effect is zero for the first trait and non-zero for the second trait, and vice versa for H_{02} . Assume the global null H_{00} holds with probability π_{oo} for asymptomatic standard normal distributions of phenotype-specific statistics Z_1 and Z_2. Additionally, assume H_{01} is a sub-null hypothesis with probability π_{o1} under which Z_1 has a standard normal distribution and Z_2 has a conditional $N(\mu_2, 1)$ distribution where the mean parameter is $\mu_2 \sim N(0, \tau_2^2)$ distributed and the sub-null hypothesis H_{02} holds with probability π_{o2} and $Z_2 \sim N(0,1)$ while $Z_1|\mu_1 \sim N(\mu_1, 1)$, where $\mu_1 \sim N(0, \tau_1^2)$. Therefore, the composite null hypothesis of no pleiotropy and the alternative hypothesis using the special case of the principle of union-intersection of statistical hypothesis testing is:

$$\begin{aligned} Ha&: H_{00}^c \cap H_{01}^c \cap H_{02}^c\ , \quad H_a = \beta_1\ \beta_2 \neq 0 \\ Ho&: H_{\partial 00} \cup H_{01} \cup H_{02}\ , \quad H_0 = \beta_1\ \beta_2 = 0 \end{aligned} \tag{15}$$

Furthermore, assume Z_1 and Z_2 are independent normal variables under H_{00} and their product Z_1Z_2 has a normal product distribution under H_{00}, $H_{01}$and H_{02}, respectively (if τ_1 and τ_2 are unknown).

Therefore, the P-value for testing the $H_o: \beta_1\beta_2 = 0$ against $H_a: \beta_1\beta_2 \neq 0$ using products of the Z scores can be obtained from

$$P_{z_1z_2} = 2 \times P_{H_o}(z_1z_2 > |z_1z_2|) = 2 \times \sum_{k=0}^{2} P(H_{0k}) P_{H_{0k}}(z_1z_2 > |z_1z_2|). \quad (16)$$

Since the P-value is sensitive to the probabilities and variance, the asymptotic approximation of the P-value is given by

$$P_{z_1z_2} = \mathbb{F}\left(z_1z_2 / \sqrt{var(z_1)}\right) + \mathbb{F}\left(z_1z_2 / \sqrt{var(z_1)}\right) - \mathbb{F}(z_1z_2), \quad (17)$$

where $\mathbb{F}(u)$ denotes the two-sided tail probability of a normal product distribution at value *u*.

5.10.6. Pros and Cons of the Univariate pleiotropy detection approaches.

All the applied univariate pleiotropy detection methods are simple to use because they only require effect sizes, standard error, and sample sizes from GWAS. All methods produce overall evidence of pleiotropy in the form of P-values. The measure of aggregate-level evidence for pleiotropy varies among the methods, with 10^{-6} being the recommended cut-off used for FDR-based approaches (cFDR, CPBayes [4,5]and 5×10^{-8}being the significance level for the other methods (MA; ASSET, PLACO). Additionally, the CPBayes approach also provides the percentage posterior contribution of each trait to the overall evidence, while ASSET gives evidence of directionality of effect and can identify switch-like variants and their effects.

Although the classical MA approach is quite simple and easy to use, it requires that the traits should be homogenous, and rejecting the global null hypothesis of no association of the SNPs with any traits does not necessarily translate to pleiotropy but, a strong effect of a SNP for a trait could motivate the observed evidence against the null hypothesis. PLACO and cFDR methods accommodate only two traits at once.

5.11. Aims of the project

Due to the mirage of univariate methods available in the literature with no recent existing benchmarking study to compare their power to detect pleiotropy and their error rate, I firstly performed a benchmarking study of five univariates pleiotropy detection approaches, namely: cFDR [5,113,118], CPBayes[4], ASSET[3], PLACO[6], and classical MA[2] to select the best performing method through a simulation study.

This method identified through the simulation study was applied to the GWAS summary statistics of two epilepsy phenotypes, generalized genetic epilepsy and focal epilepsy, obtained from the international league against epilepsy (ILAE) consortium[16] to identify pleiotropic variants for both epilepsy traits. I further applied this method to a larger sample set of GGE and FE epilepsy forms to replicate the identified variants in the first dataset and discover new associations based on the fact that the power to discover such associations increases with increased sample size.

6. Methods

6.1. Simulation study design

While many very different approaches are available for simulating populations (e.g., coalescent-based methods, forward simulations, resampling approaches)[119], they often scale unfavorably with growing sizes of populations and/or genetic variants. I used a resampling approach that is fast, efficient, uses available genetic data, and yields the same LD structure as in the base dataset[119]. Resampling approaches are also preferred when focusing on study design and analysis of actual genome data because they can preserve the allele frequency of the markers. However, if the interest is in studying evolutionary forces in the population, the other simulation approaches are more applicable[119]. Therefore, I employed a commonly used resampling algorithm to simulate the entire genome of 1 million individuals of European ancestry, using Hapgen2[13] with the haplotype data of 99 CEU (Utah residents (CEPH) with Northern and Western European ancestry) individuals provided by the 1000 Genomes project[14] (retrieved from https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html). I used only the polymorphic position of autosomes in the reference dataset. More specifically, I generated population genotypes under the null model of relative risk of 1.0. The Hapgen2 resampling algorithm is based on the Li & Stephen (LS) model of LD, where each new simulated haplotype is conditioned on the reference haplotype population and the estimates of fine-scale recombination rate across the region (retrieved from https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html), leading to the same LD pattern as in the reference data[13,120]. The size of the simulated data (~2 TB) forced me to simulate the population in 10 batches. When checking for batch effects, I identified three distinct clusters using principal component analysis (PCA). To avoid biasing or confounding effects by this substructure which may lead to inflated statistics, I included the first ten principal components (PCs) as covariates in all subsequent association analyses. This was based on my observation that the clustered structure disappeared when going from nine to ten PCs to be included as covariates.

6.1.1. Case-control status assignment for pairs of phenotypes.

To simulate multifactorial disease phenotypes from genetic data, I adopted the additive liability threshold model (ALTM)[15], a simple but well-established theoretical model calibrated to empirical data and successfully used to describe the genetic architecture of different traits. This model does not consider interaction effects (intra- and inter-locus) because they are assumed to be very small for most common traits[54]. This model assigns dichotomous case-control status according to the exceedance of some liability thresholds following classical quantitative genetics theory. As previously stated, the ALTM is an allele-based model that assumes no intra- or inter-locus interaction but allows for different genetic effect sizes, narrow-sense heritability, and disease prevalence values. More specifically, let T denote the normally distributed liability, g the phenotype-impacting variant effects, and E the standard Gaussian random noise attributed to other non-genetic sources. For each individual (l=1, …, L), locus-specific variant effects g_{ij} (i=1, …, M) are summed up across all loci (j=1, …, N):

$$G_l = \sum_{j=1}^{N} \sum_{i=1}^{M} g_{ij} \quad (18)$$

Subsequently, G_l is standardized by

$$G_l^z = \frac{(G_l - mean(G_l))}{stdev(G_l)} \quad (19)$$

and E is randomly assigned to each individual ($E_l \sim N(0,1)\ \forall l \in \{1, \ldots, L\}$) in such a way that the pre-specified narrow-sense heritability $h^2 = \frac{\mathrm{var}(G)}{(\mathrm{var}(G)+\mathrm{var}(E))}$ is attained.

To simulate the disease SNPs under different effect estimates, the standardized value of the genetic effect is multiplied by varying effect sizes. Thus, the liability T_l of an individual l is then given by:

$$T_l = G_l^z + \sqrt{(1-h)/h} \times E_l \, . \quad (20)$$

Case-control status is finally assigned by imposing a threshold t on the liability so that a proportion of the population corresponding to the disease prevalence exceeds this threshold with their liability value, i.e., individuals assigned case status. In my simulations, I considered, in turn, prevalence values of 1% and 10%, thereby considering traits of moderate and of common prevalence, respectively.

6.1.2. Case-control sample sets for pairs of traits.

To simulate a pair of traits, I randomly selected 1,000 common SNPs with allele frequencies between 5% and 20% in the simulated population. From those, I randomly selected five and ten disease-causing SNPs, respectively, to allow multiple markers to jointly contribute to the incidence of both traits as obtainable in GWAS, for each of the two traits to be simulated and assigned them a pre-defined relative risk (RR), namely 1.05, 1.2, 1.5, and 2.0 respectively. These RR values are selected based on typical values that have been observed in GWAS. I introduced biological pleiotropy by forcing the two respective causal SNP sets for the two traits to partially overlap by either 20% or 40%. More values of all the factors described here were not considered as more values will lead to many combinations that might be difficult to handle, summarize and visualize. These two SNP sets then entered the ALTM, and the traits were simulated separately across the entire population. Please note that the scenario of five causal SNPs and 20% overlap corresponds to the simplest case of a single SNP acting pleiotropically for the two traits. I defined the case-control status using the varying prevalence values as the quantile of the distribution of the liability of all individuals to define a threshold. Individuals with a liability greater than this threshold were assigned case status, otherwise keeping control status. To avoid reporting rare artifacts, I performed this step multiple times to assess variability and obtain average values close to the true mean. Hence, obtaining 100 replications where both traits would have a prevalence in the population of either 1% or 10%, respectively, given the pre-specified parameters of variant number, variant overlap, and effect size. I used these prevalence levels because the estimated prevalence of common diseases in GWAS is not often large. Finally, I drew a single random sample of 1,000, 5,000, and 10,000 cases, as often seen in real-world GWAS data, respectively, and an equal

number of controls for each trait of the pair from a given replication, resulting in sample sizes of 2000, 10,000, and 20,000 for each trait, respectively.

6.1.3. Identification of pleiotropy in simulated data

Since I restricted the study to unidirectional biological pleiotropy for single variants, which implies that a variant is pleiotropic if it increases the risk of having the two traits, I defined the power of discovering pleiotropy and its corresponding false-positive rate as shown. Firstly, I performed a univariate association test for the individual trait using PLINK v1.9 beta 6.9[121,122], including the first ten principal components as covariates (see section 6.1.2). The resulting effect sizes, standard error, or P-value in some cases from the association analysis served as input for the univariate pleiotropy detection method after ascertaining that only the selected diseased variants are causal for the traits in the association analysis. A true-positive (TP) finding is defined as the marker that reached an aggregate genome-wide significance level of 5×10^{-8}. However, the measure of overall evidence of pleiotropy is different for the FDR-based approaches (CPBayes and cFDR), where 10^{-6} is the recommended FDR threshold value[4,5,113].

Variants are considered pleiotropic and true positives in all applied methods if they are causal for both traits, that is, they exceed the defined threshold for evidence of pleiotropy in each method. At the same time, false-negatives (FN) were the disease overlapping variants that did not reach the preset threshold of evidence of pleiotropy for both traits. Therefore, the power of each method to detect true pleiotropy is:

$$Power = \frac{TP}{TP + FN} = 1 - FNR, \tag{21}$$

where the false-negative rate (FNR) is the proportion of pleiotropic variants that are not associated with both traits. I estimated the type I error rate or the false-positive rate as the proportion of the non-pleiotropic causal variants that exceed the threshold values of evidence of pleiotropy in the total number of causal SNPs for the different approaches. FPR is obtained as follows:

$$FPR = \frac{FP}{FP + TN} = 1 - TNR, \tag{22}$$

where the false positives (FP) count is the number of non-pleiotropic causal SNPs, i.e., variants that are only causal for either phenotype but found to show evidence of pleiotropy for both traits, while the true negatives (TN) or specificity is the ratio of non-pleiotropic SNPs that are genuinely non-pleiotropic.

6.2. Pleiotropy detection in two epilepsy phenotypes

The ILAE Consortium, established in 1909, is committed to working towards a world where no individual is limited by epilepsy through adequate research and education[123]. The consortium seeks to ensure the provision of resources and tools needed to the health care provider, caregivers, and people living with epilepsy to understand, prevent and treat different forms of epilepsy. However, the genetic analysis group of the consortium has focused on identifying genetic risks that predispose people to

develop epilepsy and disentangling different risk factors of epilepsy, especially in the complex forms of the disease.

ILAE consortium on complex epilepsies in the past has published results of GWAS on common forms of epilepsies (GGE and FE) and their phenotypes and reported some genes that are correlated to these traits. A meta-analysis of 34,853 (8,696 cases, 26,157 controls) individuals of European, Asian, and African-American origin resulted in the identification of voltage-gated sodium channel genes, *SCNA1,* and *SCNA9* on chromosome 2 and Protocadherin gene, *PCDH7* on chromosome 4 for all epilepsies which include in this study GGE and FE and unclassified[124]. They also found a locus 2p16.1 implicating *VRK2* and *FANCL* genes correlated to GGE[124]. In 2018, the consortium published a bigger genome-wide mega-analysis study which included an additional 6,516 cases and 3,460 controls to the previous GWAS samples, which led to a larger sample cohort comprising 15,212 epilepsy cases and 29,677 controls. The analysis found 16 loci associated with the different forms of epilepsies, with 11 of these loci being novel. Joint analysis of all epilepsies revealed a new locus at 16q12.1 in addition to loci 2p16.1 and 2q24.3 previously discovered on chromosome 2. Further, the study found 11 associated loci for GGE and a locus for FE with about 21 prioritized epilepsy genes mapped to the resulting loci[16].

6.2.1. Description of the datasets

6.2.1.1. ILAE dataset

Based on epilepsy seizure types described in section 5.7.2, I obtained summary statistics of two well-characterized epilepsy phenotypes, GGE and FE, from the ILAE mega-analysis study for the European cohort, which are more homogenous. I obtained effect sizes and standard errors for 3,708 FE cases, 9,095 GGE cases, and 24,218 overlapping controls. Based on the results obtained in section 6.1, where the ASSET method performed best in the simulation study, I applied the ASSET method to this sample. The sample overlap between the controls was accounted for in the pleiotropy analysis by computing a correlation matrix for the samples and including the obtained correlations in the analysis.

6.2.1.2. ILAE and EPI25 datasets

In the second analysis phase, I received summary statistics of GGE and FE phenotypes from the ILAE consortium and EPI25 collective. I applied ASSET to the subset of European samples, comprising 6952 (3244+3708) GGE cases and 14,939 (5344+9095) FE cases from the EPI25 and the ILAE Consortium as well as 42,434 partially overlapping controls from both sources (see Table 2 below). I performed the analysis in two different ways, firstly by considering four groups (2 GGE and 2 FE defined along with both ILAE and EPI25 cohorts) and secondly, by using two groups which consist of meta-analyzed summary statistics of GGE and FE phenotypes from both cohorts. In the first scenario, I could only account for study difference (samples cohort) in the ASSET formulation but could not expressly define GGE and FE phenotypes in both cohorts as belonging to the same trait.

Therefore, ASSET was notably blind to the fact that GGE from both EPI25 and ILAE samples are the same phenotype (same for FE). Hence, a locus showed an opposite effect direction for GGEs in EPI25 and ILAE cohorts. Considering the four groups could be an interesting analysis given the larger genetic and phenotypic homogeneity within the cohorts, but I dropped the analysis due to the above-stated concerns, which I could not directly fix in the software. To this end, I used the two cohorts' effect sizes, standard errors, and the effective sample sizes estimated from the meta-analysis for GGE and FE. The

data contained ~4.8 million common SNPs for which genotype data were available in both the EPI25 and ILAE samples.

	EPI25		ILAE		Total
	GGE	FE	GGE	FE	
Cases	3244	5344	3708	9095	21,891
Controls	13,121	13,121	24,218	24,218	

Table 2: Sample sizes of the epilepsy phenotypes in both cohorts. GGE- generalized genetic epilepsy samples, FE- focal epilepsy, EPI25- samples from EPI25 collaborative, and ILAE- samples from the International League Against Epilepsy Consortium.

6.2.2. Dataset quality control

On these summary statistics output provided by the ILAE Consortium on complex epilepsies[16], I compared the χ^2 of pairs of SNPs with LD value ($r2 > 0.4$) and removed SNPs having χ^2 values greater than $\frac{3\times\sqrt{\frac{SNP1_{\chi^2}+SNP2_{\chi^2}}{2}}}{(R^2)^2}$ to exclude SNPs with inflated χ^2 (outliers) values which can bias the result[16]. Finally, I analyzed only those SNPs that were contained in the datasets of both GGE and FE after quality control, including about 4.1 million SNPs in the pleiotropic analysis using ASSET.

6.2.3. Pleiotropy, annotation, enrichment, and colocalization analyses

I applied the ASSET method, which has proven to be powerful from the simulation study, yielding a better trade-off between FPR and power to the datasets described in section 6.2.1 to identify shared loci between GGE and FE. Since the control samples are shared for these traits, I estimated the correlation between the Z statistics and the covariances, then the subset search procedure was carried out, and finally, the P-value was approximated with the DLM procedure. I obtained odds ratios, overall P-values, and directional P-values with subsets of the traits each variant is associated with from the analysis and further performed gene mapping, annotation, and prioritization of genome-wide significant variants using various tools.

SNPs found to be significant in the pleiotropic analysis with ASSET were mapped to genes using FUMA[125] (https://fuma.ctglab.nl/). The loci harboring these significant SNPs were delineated by clustering SNPs in LD at r^2>0.2 within a ±250kb radius. The SNP with the smallest p-value was considered the "lead" SNP within a locus. I then performed functional annotation of the SNPs included in the above-defined loci to assess the potential consequences of these SNPs. To this end, I performed functional annotation of the variants that are in LD with one significant independent SNP using ANNOVAR[126]. I also performed functional annotation using the RegulomeDB database to check for evidence of SNPs affecting regulation, where RegulomeDB scores <6 are considered to affect the regulation of the mapped gene[127]. Deleteriousness of SNPs was predicted by CADD scores; scores higher than 12.37 were considered deleterious, as proposed by Kircher *et al.*[128].

Furthermore, I performed a tissue expression analysis using FUMA, based on the P-values from MAGMA[129] gene-set analysis and GTEx v8 expression data, to quantify the relationship between the

average expression of a set of genes identified in the tissue and genetic association. I checked for previous reports on genetic association with epilepsy syndromes using the GWAS catalog (https://www.ebi.ac.uk/gwas/). Finally, I also performed a Bayesian co-localization test between GGE and FE to confirm whether the lead SNPs have a high probability of being associated and shared for both syndromes, using the R packages HyPrColoc[130] ("hypothesis prioritization for multi-trait colocalization"; https://github.com/jrs95/hyprcoloc/) and coloc v5.1.0[90,131] (https://CRAN.R-project.org/package=coloc). More specifically, I estimated the posterior probability of co-localization as evidence that a variant is shared or associated for multiple traits using HyPrColoc and of association of both syndromes with the lead SNPs using coloc tools.

7. Main results

7.1. Simulation study

- At a small variant effect size of 1.05, all five applied methods had no power to detect pleiotropic SNPs.
- The power to detect pleiotropy for all methods increases with increasing sample sizes across all simulation scenarios except the cFDR approach, which showed a downward trend possibly due to software malfunction.
- CPBayes performed best in terms of power, closely followed by ASSET and cFDR.
- Prevalence seemed to have a modest effect on power. The effect is distinguishable at RR=1.2 and a sample size of 2000.
- All methods show considerably low FPR at RR=1.05 except CPBayes, which has >10% FPR across all simulation scenarios.
- Classical MA is not recommended for pleiotropy detection as it performed poorly in terms of FPR.
- The larger the number of disease SNPs and the degree of sharing of these SNPs between the traits, the lower the FPR for all approaches except for classical MA, which seems to have the opposite effect.
- The larger the number of disease SNPs and the degree of sharing of these SNPs between the traits, the larger the Power for all approaches.
- The ASSET method performed gave a good trade-off between power and FPR by keeping the FPR generally low while maintaining the high power to detect pleiotropy across all simulation scenarios.

7.2. Pleiotropy detection in epilepsy phenotypes (ILAE dataset)

- I identified 40 pleiotropic SNPs at two loci: 2q24.3 and 17q21.32, at a genome-wide significance, three of which were independent lead SNPs. SNPs rs60055328 and rs2212656 mapped to locus 2q24.3, whereas rs16955463 mapped to 17q21.32.
- Functional annotation using ANNOVAR[126] shows that 11% of the SNPs are intergenic, 23% are intronic, and 61% are non-coding transcript intron variants.
- Locus 2q24.3 had already been reported for GGE and FE and mapped to *SCNA1*, *SCNA2*, *SCNA3,* and *TTC21B* in the ILAE mega-analysis.
- Locus 17q21.32 is the unreported new putative pleiotropic locus for FE and GGE comprising of SKAP*1*, *OSBPL7*, *SP6*, *SP2*, *PNPO*, *PRR15L*, *CDK5RAP3*, *COPZ2*, *NFE2L1*, *CBX1*, *SNX11*, *HOXB1*, *HOXB2*, and *HOXB3* genes.
- MAGMA tissue-specific expression of the genes found most to be preferentially expressed in the brain.
- Based on the Ensembl variant effect predictor and FUMA, *SCNA1*, *SCN9A*, and *TTC21B* were the prioritized genes for Locus 2q24.3, while *SKAP1* and *PNPO* are the prioritized genes for locus
  17q21.32.

7.3. Pleiotropy detection in epilepsy phenotypes (ILAE and EPI25 dataset)

- Here, I identified 50 pleiotropic SNPs at genome-wide significance level in three loci.
- I replicated locus 2q24.3 and found a new putative locus 9q21.13 to be pleiotropic for both GGE and FE.
- The previously reported locus 17q21.32 could not be replicated in this new sample cohort as it was found to be strongly associated with GGE only with a marginally significant opposite direction of effect in FE.
- Locus 2q24.3 had already been reported for GGE and FE and mapped to *SCNA1*, *SCNA2*, *SCNA3,* and *TTC21B* in the ILAE all epilepsy analysis via meta-analysis[132].
- The new locus 9q21.13 contains the *RORB* gene.

8. Publications

8.1. Contribution to publications

In the following paragraphs, I describe my contributions to the publications listed in this thesis. The first two journal articles directly contribute to my thesis subject, while the last publication I co-authored is not directly linked to the thesis topic. For my first authorship article, I designed, wrote the simulation codes, performed the statistical analysis, interpreted the results, and wrote the first draft of the manuscript. In the second article I co-authored, I contributed to the design of the study, performed pleiotropy analysis, and contributed to the writing of the manuscript, as clearly stated in the manuscript.

Benchmarking of univariate pleiotropy detection methods applied to epilepsy phenotypes – First author.

This project's main objective was to identify shared variations between two epilepsy forms. I first performed a simulation study after extensively reviewing available methods for pleiotropy analysis in literature to identify the best method. I decided to use univariate meta-analysis-based approaches due to the unavailability of simultaneously measured phenotypes and the corresponding genotype data on individuals, but single-trait summary statistics of the epilepsy phenotypes. I applied five recent and well-implemented approaches (see Table 1) to the simulated data and found the ASSET method as the best performing method in terms of power and FPR, which was then applied to the epilepsy phenotypes.

Genome-wide meta-analysis of over 29,000 people with epilepsy reveals 26 loci and subtype-specific genetic architecture – co-author

In this co-authored project, the main objective was to identify genome-wide significant loci underlying epilepsy disorders in different ethnicity. Identified associated loci were further subjected to follow-up analyses, such as SNP-based heritability, tissue and cell enrichment, pleiotropy, correlation, and drug repurposing checks. I contributed to the study by performing pleiotropy analysis on the GWAS summary statistics of genetic generalized epilepsy and focal epilepsy phenotypes to identify overlapping loci between the two forms of epilepsy. I also contributed to the writing of the manuscript.

8.2. Main Publications

Adesoji, O. M., Schulz, H., May, P., Krause, R., Lerche, H., Nothnagel, M., & ILAE Consortium on Complex Epilepsies. (2022). Benchmarking of univariate pleiotropy detection methods applied to epilepsy. *Human Mutation*. Published online May 27, 2022. Doi: https://doi.org/10.1002/humu.24417.

Received: 21 December 2021 | Revised: 28 April 2022 | Accepted: 25 May 2022

DOI: 10.1002/humu.24417

RESEARCH ARTICLE

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

Benchmarking of univariate pleiotropy detection methods applied to epilepsy

Oluyomi M. Adesoji[1,2] | Herbert Schulz[3] | Patrick May[4] | Roland Krause[4] | Holger Lerche[5] | Michael Nothnagel[1,2] | ILAE Consortium on Complex Epilepsies

[1]Cologne Center for Genomics, University of Cologne, Cologne, Germany

[2]University Hospital Cologne, Medical Faculty, University of Cologne, Cologne, Germany

[3]Department of Microgravity and Translational Regenerative Medicine, Clinic of Plastic, Aesthetic and Hand Surgery, Otto von Guericke University, Magdeburg, Germany

[4]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

[5]Department of Neurology and Epileptology, Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany

Correspondence

Michael Nothnagel, Cologne Center for Genomics, Department of Statistical Genetics and Bioinformatics, University of Cologne, Weyertal 115b, 50931 Cologne, Germany.
Email: michael.nothnagel@uni-koeln.de

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Numbers: NO755/6-1, NO755/13-1; Fonds National de la Recherche Luxembourg, Grant/Award Numbers: INTER/DFG/17/11583046, NCER-PD/FNR11264123

Abstract

Pleiotropy is a widespread phenomenon that may increase insight into the etiology of biological and disease traits. Since genome-wide association studies frequently provide information on a single trait only, only univariate pleiotropy detection methods are applicable, with yet unknown comparative performance. Here, we compared five such methods with respect to their ability to detect pleiotropy, including meta-analysis, ASSET, conditional false discovery rate (cFDR), cross-phenotype Bayes (CPBayes), and pleiotropic analysis under the composite null hypothesis (PLACO), by performing extended computer simulations that varied the underlying etiological model for pleiotropy for a pair of traits, including the number of causal variants, degree of traits' overlap, effect sizes as well as trait prevalence, and varying sample sizes. Our results indicate that ASSET provides the best trade-off between power and protection against false positives. We then applied ASSET to a previously published International League Against Epilepsy (ILAE) consortium data set on complex epilepsies, comprising genetic generalized epilepsy and focal epilepsy cases and corresponding controls. We identified a novel candidate locus at 17q21.32 and confirmed locus 2q24.3, previously identified to act pleiotropically on both epilepsy subtypes by a mega-analysis. Functional annotation, tissue-specific expression, and regulatory function analysis as well as Bayesian colocalization analysis corroborated this result, rendering 17q21.32 a worthwhile candidate for follow-up studies on pleiotropy in epilepsies.

KEYWORDS

association, epilepsies, meta-analysis, pleiotropy, SNPs

A list of consortium members is provided in the Supporting Information Material.

1 | INTRODUCTION

1.1 | Pleiotropy

Pleiotropy is an early recognized (Stearns, 2010) and widespread phenomenon for complex biological and disease traits. It is defined as one or more genetic variants simultaneously having a causal effect on two or more phenotypes. Pleiotropy is broadly categorized into three forms, namely biological, mediated, and spurious pleiotropy. *Biological pleiotropy* denotes the phenomenon that a single variant truly affects multiple traits or that different but neighboring markers in the same gene truly affect different traits (Solovieff et al., 2013; van Rheenen et al., 2019). This relationship can take different forms based on the association of the unobserved causal variant to the observed associated variant and linkage disequilibrium (LD) between variant alleles in a gene. The underlying biological mechanisms of horizontal or biological forms of pleiotropy can be easily understood as mirroring disease pathways that include this or these variants, respectively (Solovieff et al., 2013). On the other hand, *mediated pleiotropy* is a form of pleiotropy in which a trait is causally related to another such that a variant associated with one is indirectly associated with the other. For example, while the *CHRNA5* gene is known to be associated with lung cancer, chronic obstructive pulmonary disease (COPD), and smoking behaviors, the association with lung cancer could be directly due to the effect of the gene variants on smoking intensity or indirectly through the effects on COPD (Bien & Peters, 2019). Measurement or identification errors, as well as design artifacts, can result in *spurious pleiotropy*. A disease or subphenotype could be misclassified, or a genetic variant could be in LD with two different associated single-nucleotide polymorphisms (SNPs) in different genes affecting different traits.

Genetic association studies seek to establish a relationship within genes, SNPs, or loci and a *single* trait. Genome-wide association studies (GWAS) were mostly performed for this purpose in the past 15 years. Independent GWAS have repeatedly identified the same locus or genetic variant to be associated with multiple traits, generally referred to as cross-phenotype association. For example, SNP rs6983267 (8q24) was found to be associated with colorectal and prostate cancer in separate GWAS studies (Haiman et al., 2007; Pomerantz et al., 2009; Thomas et al., 2008; Tomlinson et al., 2007; Tuupanen et al., 2009; Zanke et al., 2007), whereas SNP rs12720356, located in the *TYK2* gene, has been reported to be associated with Crohn's disease and psoriasis in independent studies (Franke et al., 2010; Genet et al., 2010), indicating a shared underlying etiology and the existence of pleiotropic factors. On the other hand, studying two or more traits in a *combined* way to identify possible pleiotropic factors may help identify associated variation and provide insight into the etiology of related traits and complex diseases by identifying shared pathways.

1.2 | Epilepsy syndromes as a prime candidate for pleiotropic mechanisms

Epilepsy is a common brain disorder characterized by unprovoked recurring seizures. About 50 million people live with epilepsy worldwide; 24 millions of these have active idiopathic or genetic epilepsy (Cooper, 2019; Singh & Sander, 2020). As with most complex traits, a number of genetic and environmental factors contribute to the different epilepsy forms and their loci overlap between them (Ottman, 2005).

In previous GWAS studies, the epilepsies are typically classified as genetic generalized epilepsies (GGE), focal epilepsies (FE), and unknown forms (International League Against Epilepsy Consortium on Complex Epilepsies, 2018; International League Against Epilepsy Consortium on Complex Epilepsies. Electronic address, 2014; Wolking et al., 2020). FE is characterized by seizures originating in a specific area of the brain, whereas seizures in GGE spread very rapidly bilaterally throughout the cortex without a clearly identifiable seizure origin. Generalized epilepsies have been reported to have stronger genetic components than focal epilepsies. A number of genes identified in GWAS have been reported for monogenic forms of FE (International League Against Epilepsy Consortium on Complex Epilepsies, 2018). A recent mega-analysis study by the *ILAE Consortium* (International League Against Epilepsy Consortium on Complex Epilepsies, 2018) jointly analyzed samples with different forms of epilepsy by just considering epilepsy affection status while ignoring the specifically manifested subphenotype. This study established a few epilepsy loci that appear to be shared across all subphenotypes, although it is unclear how large the contribution of subphenotype to this joint association signal with each of the identified loci is. The ILAE Consortium identified three lead SNPs, namely rs4671319, rs6432877, and rs4638568, that were associated with both FE and GGE. These loci were mapped to sodium channel-encoding genes (*SCN1A*, *SCN2A*, *and SCN3A*), a transcription factor (*BCL11A*), and a histone modification gene (*BRD7*), respectively. Furthermore, the study provided evidence that the common epilepsy-associated variants play a role in epigenetic regulation of gene expression in the brain.

1.3 | Methods for pleiotropy detection

Numerous methods for detecting pleiotropy have been proposed in the past, with this aspect of genetic studies still being an active field of methodological development. Broad distinctions can be made depending on the type of data being analyzed, on the availability of individual-level genetic data versus summary statistics, and the joint measurement of all considered phenotypes per individual. Availability of individual-level data with all phenotypes simultaneously measured in the same sets of individuals allows for the use of multivariate statistical approaches, such as multinomial logistic regression (Agresti, 2003; Morris et al., 2010), generalized linear mixed models

(Fitzmaurice & Laird, 1993; Schaid et al., 2019), linear mixed-effects models (Laird & Ware, 1982; Verbeke et al., 2010), frailty models (Yang & Wang, 2012), principal component analysis (PCA) (Jolliffe, 2002; Jolliffe & Cadima, 2016), and others (Salinas et al., 2018), depending on population structure, sample size and the type of trait being studied: continuous, dichotomous (categorical), or time to an event. However, many datasets available for the study of presumably pleiotropic phenotypes are characterized by single-phenotype measurements, such as most GWAS, with little or no sample overlap between these phenotypes. This renders multivariate approaches inapplicable and requires univariate approaches instead. Moreover, individual-level genotype data from different studies are often hard to combine due to legal or ethical restrictions, whereas GWAS summary statistics are much more readily available. Besides traditional meta-analysis (MA) (Willer et al., 2010), a range of novel univariate methods has been proposed in recent years (see Section 2 for technical details), such as the *conditional false discovery rate* (cFDR) (Andreassen et al., 2013; Liley & Wallace, 2015), *subset-based meta-analysis* (ASSET) (Bhattacharjee et al., 2012), *cross-phenotype Bayes* (CPBayes) (Majumdar et al., 2018), and *pleiotropic analysis under the composite null hypothesis* (PLACO) (Ray & Chatterjee, 2020).

Traditional MA permits the combination of datasets, is computationally simple, and well suited for traits that are biologically correlated, however, the marginal contribution of each trait to the association signal cannot be estimated. On the other hand, ASSET and CPBayes may provide more insight into the marginal contributions of traits under study while also evaluating the overall evidence of association jointly for these traits. The cFDR and PLACO methods can only be used to study a pair of traits at once but minimize the probability that only a single trait is driving the observed joint effect. All five methods allow accounting for potential correlation originating from shared controls between traits as is frequently the case in GWAS studies. While previous studies (Bhattacharjee et al., 2012; Majumdar et al., 2018; Ray & Chatterjee, 2020) performed some comparisons between univariate pleiotropy detection methods, these studies were usually accompanying the proposition of a new method while being limited in the scope of their comparisons. It is therefore unclear what the comparative performance of the five above-mentioned univariate approaches is and if there is an optimal choice for univariate pleiotropy detection. An independent study benchmarking all five approaches is so far lacking.

1.4 | Aim of the study

Here, we performed a comparative study of five univariate approaches for pleiotropy detection, both frequentist and Bayesian, including traditional MA and the recently proposed ASSET, cFDR, CPBayes, and PLACO methods (Table 1). We simulated genome-wide data and pairs of pleiotropic phenotypes under different etiological models for pleiotropy, including varying numbers of associated genetic variants and varying degrees of their overlap between phenotypes, different single-marker effect size, and phenotype prevalence values, and benchmarked the five approaches with respect to their power to detect true pleiotropy and the false-discovery rate (FDR) under varying sample sizes. We subsequently used the method that showed superior performance in our simulation study to identify pleiotropic loci for a pair of epilepsy syndromes, namely GGE and FE as two clinically well-characterized forms. To this end, we analyzed GWAS summary statistics provided by the ILAE consortium.

2 | METHODS

2.1 | Type of pleiotropy under consideration

In our comparison, we restricted ourselves to unidirectional horizontal or biological pleiotropy involving a single variant. Thus, a genetic marker was considered to act pleiotropically if it simultaneously impacted two different traits, increasing the risk for both traits. The direction of effect was not considered in this simulation study; all disease SNPs increased the risk of having the trait by the magnitude of the varying relative risk chosen. For example, a marker that increased the risk of affection for two diseases would fall into this category. On the other hand, a marker marginally associated with one trait but not the other would be considered not to act pleiotropically. Furthermore, two markers at the same locus or gene and affecting two different traits, respectively, were not considered. We also did not consider scenarios where two markers were in strong allelic association, or LD, and would have been found in phenotypic association with two different traits. Notably, we did not only

TABLE 1 Univariate pleiotropy detection methods included in the analysis and their sources

Method	Reference	Web resource
Classic fixed-effect meta-analysis (MA)	Willer et al. (2010)	http://csg.sph.umich.edu/abecasis/metal/download/
Subset-based meta-analysis (ASSET)	Bhattacharjee et al. (2012)	https://bioconductor.org/packages/ASSET/
Cross-phenotype Bayes (CPBayes)	Majumdar et al. (2018)	https://github.com/ArunabhaCodes/CPBayes
Conditional false discovery rate (cFDR)	Liley & Wallace (2015)	https://github.com/jamesliley/cFDR-common-controls
Pleiotropic analysis under the composite null hypothesis (PLACO)	Ray & Chatterjee (2020)	https://github.com/RayDebashree/PLACO

consider a single marker acting pleiotropically but allowed for multiple markers jointly contributing to the incidence of two phenotypes in our simulations (see below).

2.2 | Methods included in the comparison

We compared five univariate single-marker approaches with respect to their power to detect pleiotropic factors in pairs of phenotypes and their false-positive rate (FPR), including classical MA and four recently proposed methods.

2.2.1 | Classical MA

MA is a long-established approach to combine several, often heterogeneous studies in a joint analysis to synthesize or consolidate results from those previous studies, to increase power to detect true associations, and for others aims, often using only aggregate data. MA has been originally designed to pool different study statistics for efficiency and to circumvent challenges arising due to population structures, study-specific covariates, and individual-level data management. MA is not specifically designed for pleiotropy detection but can be expected to identify variants that have concordant effects in separate studies of two or more phenotypes; correspondingly, its use in pleiotropy detection has been demonstrated (Chung et al., 2019; Kulminski et al., 2019). The basic approach underlying fixed-effect MA is the conversion of study-specific *P*-values (P_k) and effect direction (δ_k) from K studies into standard normally distributed signed Z-scores

$$Z_k = \Phi^{-1}(P_k/2) * sign(\delta_k), \tag{1}$$

which are then combined to an overall Z-score (Z_{meta}) statistics by weights w_k. Weights are typically assigned based on the inverse of the variance but this is also roughly proportional to sample size. Therefore $w_k = \frac{1}{\delta_k}$ and the variance of the combined effect is $\delta_{meta} = \frac{1}{\sum_{i=1}^{k} w_k}$. Thus,

$$Z_{meta} = \frac{\sum_{k=1,K} Z_k W_k}{\sqrt{\sum_{k=1.K} W_k^2}}. \tag{2}$$

The final overall p-value is then obtained by comparing this statistic against a standard normal distribution:

$$P_{meta} = 2[1 - (\Phi(|Z_{meta}|))], \tag{3}$$

where Φ denotes the standard normal distribution function. Here, we have used the MA implementation in the METAL v2011-03-25 software (Willer et al., 2010).

2.2.2 | Subset-based meta-analysis method (ASSET)

Association analysis based on subsets (ASSET) (Bhattacharjee et al., 2012) represents an extension of fixed-effects MA. It aims at maximizing the association signal across all possible subsets of two or more phenotype studies, thereby allowing for nonassociated phenotypes that are not impacted by a pleiotropic variant, while also correcting for multiple testing. For a given subset B of $m(B)$ studies, the respective overall Z-score Z(B) is obtained following the MA approach by

$$Z(B) = \sum_{k \in B} \sqrt{\pi_k(B)} Z_k, \tag{4}$$

where $\pi_k(B) = n_k / \sum_{k=1}^{m(B)} n_k$ weights the different studies proportional to the square root of respective sample sizes; if covariate adjustments are similar across studies, then $B_k \propto \frac{1}{n_k}$ where n_k is the sample size for the k^{th} study (Bhattacharjee et al., 2012). Formula (4) is then maximized over all possible subsets:

$$Z_{meta-max} = max_{B \subseteq \{1,K\}} |Z(B)|. \tag{5}$$

The overall hypothesis of a genetic marker to be associated with all traits is evaluated by $Z_{meta-max}$. The upper bound for the P values from the defined multivariate distribution is obtained through the discrete local maxima (DLM) method (see (Bhattacharjee et al., 2012) for full details). One of the prerequisites of this method is that all traits must be of the same type and depending on the number of traits, the number of subsets could grow exponentially. We used the ASSET v2.10.0 R implementation provided with the publication (Bhattacharjee et al., 2012).

2.2.3 | Cross phenotype Bayes (CPBayes)

Unlike MA, CPBayes (Majumdar et al., 2018) is a fully Bayesian MA approach that employs the Gibbs sampling form of the Markov chain Monte Carlo (MCMC) technique to obtain posterior samples, hence making inferences on pleiotropy. It measures the evidence of aggregate-level pleiotropy as well as subsets of traits that are pleiotropic. This evidence is given by the local false discovery rate (locFDR) and the Bayes factor (BF) through testing the global null hypothesis (H_o) of no association with *any* trait versus the alternative hypothesis (H_1) of association with *at least one* trait. Prior information is provided by the spike and slab approach where the spike element represents the null effect while the slab part represents the nonnull effect. Let $\hat{\beta_k}$ be regression estimates of true effect β obtained from the separate univariate models of individual traits T_k and s_k their standard errors. If the sample size is sufficiently large and $\hat{\beta_k}$ are uncorrelated, we assume that

$$\hat{\beta_k} \Big| \beta_k \overset{ind}{\sim} N\left(\beta_k, s_k^2\right) (k = 1, .., K). \tag{6}$$

However, for correlated estimates $(\hat{\beta_1}, \hat{\beta_k})$ with variance–covariance matrix S that corresponds to the SNPs, $\hat{\beta} \mid \beta \sim MVN(\beta, S)$. The prior information is given such that z_k denotes the association status of T_k (see Majumdar et al. (2018), page 22). The locFDR equals the probability of null association (PNA) given the data: locFDR= $P(H_0|D)$. With the posterior odds (PO) equaling $PO = \frac{P(H_1 \mid D)}{P(H_0 \mid D)}$ and the posterior probability of association equaling PO/(1-PO), we obtain the posterior probability of null association (PPNA) which is the same quantity as locFDR as:

$$PPNA = 1 - PPA = \frac{1}{1 + PO} = P(H_o|D). \tag{7}$$

Also, the BF is obtained by:

$$BF = \frac{P(D|H_1)}{P(D|H_0)} = \frac{P(H_1|D)P(H_0)}{P(H_0|D)P(H_1)} = \frac{P(Z \neq 0|D)P(Z = 0)}{P(Z = 0|D)P(Z \neq 0)} = \frac{\text{Posteriorodds}}{\text{Priorodds}}, \tag{8}$$

locFDR and BF provides the evidence of aggregate pleiotropy such that if BF >1 and locFDR $<10^{-6}$ the variant is pleiotropic. In addition, the trait-specific posterior probability of association also provides information on the relative strength of association between a pleiotropic variant and the selected nonnull trait contribution to the aggregate evidence of association. For details, see Majumdar et al. (2018). We used CPBayes v1.1.0 implemented in R.

2.2.4 | cFDR

Benjamini and Hochberg (1995) defined the FDR as the proportion of incorrectly rejected null hypotheses V among the rejected hypotheses R (Benjamini & Hochberg, 1995), i.e. $Q=V/R$ (assuming $R > 0$). Assuming that the P-value of a trait k across all variants is a realization of a random variable P_k, the unconditional FDR (uFDR) for the null hypothesis $H_0^{(k)}$ of no association of this variant with phenotype k is then defined as the probability that a random variant from this set of rejected hypotheses falls under the null hypothesis for this phenotype (Liley & Wallace, 2015). The uFDR can be estimated from a set of observed P-values $p_k^1, p_k^2, ..., p_k^N$ for a set of N variants as the ratio of the expected quantile of P_k under $H_0^{(k)}$ and the observed quantile of P_k:

$$\widehat{uFDR}_{(p_k)} = \frac{p_k}{\frac{\#\left(p_k^i \middle| p_k^i \leq p_k\right)}{N}}. \tag{9}$$

The basic motivation for the conditional FDR (cFDR) is now that variants that act pleiotropically should show a tendency toward smaller association P-values for each of a pair of phenotypes. Thus, selecting variants from the lower quantiles of the P-value distribution of association with one phenotype ("principal phenotype") by applying thresholds should then lead to an enrichment of variants with smaller P-values of association with a second phenotype ("conditional phenotype"). The P-value distribution of the selected variants for the second phenotype conditional on the P-value distribution of the unselected variants for the first phenotype will then deviate from that for all (unselected) variants for the second phenotype. The cFDR is defined as the posterior probability that a given variant falls under the null hypothesis for the principal phenotype given that the P-values for both phenotypes are less or equal to the observed P-values (p_k, p_l): $P\left(H_0^{(k)} | P_k \leq p_k, P_l \leq p_l\right)$. Similar to the uFDR and based on observed P-value pairs $\{(p_k^1, p_l^1), (p_k^2, p_l^2), ..., (p_k^N, p_l^N)\}$ for two phenotypes k and l at N different SNPs, it is estimated by the ratio of the expected quantile of P_k under $H_0^{(k)}$ among those p_k^i where i satisfies $P_l^i \leq p_l$ and the observed quantiles:

$$\hat{cFDR}_{(p_k|p_l)} = \frac{P\left(P_k \leq p_k \mid P_l \leq p_l, H_0^{(k)}\right)}{\frac{\#\left(\left(P_k^w, P_l^w\right) \in \left(P_i, P_j\right) \middle| p_k^i \leq p_k \text{ and } p_l^i \leq p_l\right)}{N_1}}, \tag{10}$$

where N_1 denotes the number of P-value pairs with $P_l \leq p_l$ and (p_k, p_l) a P-value pair for an SNP of interest (Liley & Wallace, 2015). Association with both phenotypes is tested via a conjunction FDR procedure to minimize the effect of a single phenotype driving the association signal; please refer to Andreassen et al. (2013) for details. We used cFDR v1.1.

2.2.5 | Pleiotropic analysis under composite null hypothesis (PLACO)

The PLACO approach uses aggregate level association statistics to identify pleiotropy. In this approach, the composite null hypothesis is that a variant is associated with *none or only one* of the phenotypes compared with other methods that assume no association of the SNPs to *any* of the traits (Ray & Chatterjee, 2020). The null and alternative hypotheses are defined in such a way that the global null hypothesis consists of subnull hypotheses H_{01} and H_{02} where $H_{01}: \beta_1 = 0, \beta_2 \neq 0, H_{02}: \beta_1 \neq 0, \beta_2 = 0$ for both traits. Assume the global null H_{00} holds with probability π_{00} for asymptomatic standard normal distributions of phenotype-specific statistics Z_1 and Z_2. Additionally, assume H_{01} is a subnull hypothesis with probability π_{01} under which Z_1 has a standard normal distribution and Z_2 has a conditional $N(\mu_2, 1)$ distribution where the mean parameter is $\mu_2 \sim N(0, \tau_1^2)$ distributed and the subnull hypothesis H_{02} holds with probability π_{02} and $z_2 \sim N(0, 1)$ while $Z_1 \mid \mu_1 \sim N(\mu_1, 1)$, where $\mu_1 \sim N(0, \tau_1^2)$. Therefore, the composite null hypothesis of no pleiotropy and the alternative hypothesis using the special case of the principle of union-intersection of statistical hypothesis testing is:

$$\begin{aligned} Ha &: H_{00}^c \cap H_{01}^c \cap H_{02}^c, \ H_a = \beta_1\beta_2 \neq 0, \\ Ho &: H_{00} \cup H_{01} \cup H_{02}, \ H_0 = \beta_1\beta_2 = 0. \end{aligned} \tag{11}$$

Furthermore, assume Z_1 and Z_2 are independent normal variables under H_{00} and their product Z_1Z_2 has a normal product distribution

under H_{00}, H_{01} and H_{02}, respectively (if τ_1 and τ_2 are unknown). Therefore, the P-value for testing the $H_o : \beta_1\beta_2 = 0$ against $H_a : \beta_1\beta_2 \neq 0$ using products of the Z scores can be obtained from

$$P_{z_1z_2} = 2 \times P_{H_o}\left(z_1z_2 > \left|z_1z_2\right|\right) = 2 \times \sum_{k=0}^{2} P(H_{0k})P_{H_{0k}}\left(z_1z_2 > \left|z_1z_2\right|\right). \quad (12)$$

Since the P-value is sensitive to the probabilities and variance, the asymptotic approximation of the P-value is given by

$$P_{z_1z_2} = \mathbb{F}(z_1z_2 / \sqrt{var(z_1)}) + \mathbb{F}(z_1z_2 / \sqrt{var(z_1)}) - \mathbb{F}(z_1z_2), \quad (13)$$

where $\mathbb{F}(u)$ denotes the two-sided tail probability of a normal product distribution at value u. PLACO adjusts for correlation between samples using sample sizes of cases and controls from both traits where available. Another way to estimate correlation in the absence of sample sizes is by selecting the variants that are not associated with both traits and calculating correlation based on the Z values. We used PLACO v0.1.1.

2.3 | SImulation study

We adopted a three-stage simulation design to obtain repeated case–control sample sets of pairs of phenotypes that could then be used to evaluate the different pleiotropy detection methods. More specific, we first simulated one large population that could be used as a pool to simulate pairs of phenotypes. In a second step, we used the additive liability threshold model (ALTM) (Agarwala et al., 2013) to repeatedly assign case–control status to all individuals of that population for pairs of phenotypes in accordance to preselected characteristics of the phenotypes and the associated genetic variation. Third, we repeatedly obtained case–control samples of varying sizes and comparatively applied classical MA, ASSET, cFDR, CPBayes, and PLACO.

2.3.1 | Simulation of one population of European ancestry

While a number of very different approaches are available for simulating populations (e.g., coalescent-based methods, forward simulations, resampling approaches) (Carvajal-Rodriguez, 2010), they often scale unfavorably with growing sizes of populations and/or genetic variants. We used a resampling approach to simulate the entire genome of 1 million individuals of European ancestry, using Hapgen2 v2.2.0 (Su et al., 2011) with the haplotype data of 99 CEU (Utah residents [CEPH] with Northern and Western European ancestry) individuals provided by the 1000 Genomes project (Genomes Project et al., 2015) (retrieved from https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html). We used only the polymorphic position of autosomes in the reference data set. Thus, we can exclude biasing effects by either sex or ancestry. More specifically, we generated genotypes of the population under the null model of relative risk of 1.0. The Hapgen2 resampling algorithm is based on the Li & Stephen (LS) model of LD where each new simulated haplotype is conditioned on the reference haplotype population and the estimates of fine-scale recombination rate across the region (retrieved from https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html), leading to the same LD pattern as in the reference data (Li & Stephens, 2003; Su et al., 2011). The size of the simulated data (~2 TB) forced us to simulate the population in 10 batches. This resulted in systematic differences, likely induced by the random number generator and the starting times of the simulation batches. When checking for batch effects using PCA, we identified three distinct clusters. To avoid biasing effects by this substructure, we eventually included the first 10 principal components (PCs) as covariates in all subsequent association analyses. This was based on our observation that inclusion of the first nine PCs was not sufficient to resolve this cluster structure and that the clustered structure disappeared when including the first 10 PCs as covariates (Supporting Information: Figure S1). This was further corroborated by a genomic inflation factor, as implemented in PLINK (Chang et al., 2015; Purcell et al., 2007), of approximately 1.0 throughout the single-trait association tests (see below).

2.3.2 | Assignment of case–control status for pairs of phenotypes

To simulate multifactorial disease phenotypes from genetic data (irrespective of any particular disease etiology such as certain epilepsy forms), we adopted the additive liability threshold model (ALTM) (Agarwala et al., 2013), which assigns dichotomous case–control status according to the exceedance of some liability thresholds following classical quantitative genetics theory. The ALTM assumes no intra- nor inter-locus interaction but allows for different values of genetic effect size, narrow-sense heritability, and disease prevalence. More specifically, let T denote the normally distributed liability, g the phenotype-impacting variant effects, and E the standard Gaussian random noise attributed to other nongenetic sources. For each individual ($l = 1, \ldots, L$), locus-specific variant effects g_{ij} ($i = 1, \ldots, M$) are summed up across all loci ($j = 1, \ldots, N$):

$$G_l = \sum_{j=1}^{N}\sum_{i=1}^{M} g_{ij}. \quad (14)$$

Subsequently, G_l is standardized by

$$G_l^z = \frac{(G_l - \text{mean}(G_l))}{\text{stdev}(G_l)}, \quad (15)$$

and E is randomly assigned to each individual ($E_l \sim N(0,1)\ \forall l \in \{1, \ldots, L\}$) in such a way that that the prespecified narrow-sense heritability $h^2 = \frac{var(G)}{(var(G) + var(E))}$ is attained.

However, to simulate the disease SNPs under different effect estimates, the standardized value of the genetic effect is multiplied by varying effect sizes. Thus, the liability T_l of an individual l is then given by:

$$T_l = G_l^z + \sqrt{(1-h)/h} \times E_l. \quad (16)$$

Case–control status is finally assigned by imposing a threshold t on the liability so that a proportion of the population that corresponds to the disease prevalence exceeds this threshold with their liability value, that is, individuals that are assigned case status. In our simulations, we considered, in turn, prevalence values of 1% and 10%, thereby considering traits of moderate and of common prevalence, respectively.

2.3.3 | Case–control sample sets for pairs of traits

For a pair of traits to be simulated, we randomly selected 1000 SNPs with allele frequencies between 5% and 20% in the simulated population. From those, we randomly selected 5 and 10 disease-causing SNPs, in turn, for each of the two traits to be simulated and assigned them a predefined relative risk, namely 1.05, 1.2, 1.5, and 2.0, respectively. We introduced biological pleiotropy by forcing the two respective causal SNP sets for the two traits to partially overlap, namely by either 20% or 40%. These two SNPs sets then entered the ALTM and the traits were simulated separately across the complete population. Please note that the scenario of five causal SNPs and 20% overlap corresponds to the simplest case of a single SNP acting pleiotropically for the two traits. We defined the case–control status using the varying prevalence values as the quantile of the distribution of the liability of all individuals to define a threshold and individuals with a liability greater than this threshold were assigned case status, otherwise keeping control status. We performed this approach multiple times until we obtained 100 replications where both traits would have a prevalence in the population of either 1% or 10%, respectively, given the prespecified parameters of variant number, variant overlap, and effect size. Finally, we drew a single random sample of 1000, 5000, and 10,000 cases, respectively, and an equal number of controls for each trait of the pair from a given replication, resulting in sample sizes of 2000, 10,000, and 20,000 for each trait, respectively.

2.3.4 | Method application and performance measures

After generating our sample sets, we performed association analyses separately for both traits using PLINK v1.9 beta 6.9, (Chang et al., 2015; Purcell et al., 2007). The first 10 PCs were included in this analysis as covariates. We used the univariate summary statistics in the forms of effect estimates, standard error, and P-value in some cases from the association analysis for pleiotropy analysis. Throughout our simulations, all performance measures are with respect to SNPs that are *causal* since we already checked from the association study that no nondisease SNPs is causal to the defined phenotypes. We defined a marker as a true-positive (TP) finding if its P-values reached genome-wide "significance" levels for both traits. However, the measure of aggregate-level evidence for pleiotropy varies among the methods, with 10^{-6} being the cut-off used for FDR-based approaches (cFDR, CPBayes) as suggested by Liley & Wallace (2015) and 5×10^{-8} being the significance level for the other methods (MA; ASSET, PLACO).

A TP variant was defined as a pleiotropic variant, that is, a variant that is causal for both traits, for which the applied method obtained a "significant" result (exceeding the respective threshold) for both traits, while *false negative* (FN) variants were those for which the respective method did not yield a "significant" result for both traits. Correspondingly, the power of a method to detect true pleiotropy was estimated as the proportion

$$\text{Power} = \frac{TP}{TP + FN} = 1 - \text{FNR}. \quad (17)$$

We obtained estimates for the FPR, or type I error, of the different methods as the ratio of nonpleiotropic causal SNPs wrongly "significant" (exceeding the respective threshold) by the total number of causal SNPs. It is estimated as:

$$FPR = \frac{FP}{FP + TN} = 1 - TNR, \quad (18)$$

where *false positives* (FP) count is the number of nonpleiotropic causal variants, that is, variants that are causal for exactly one of the two traits, that were wrongly found to be significantly associated with both traits, while *true negatives* (TN) denote the count of nonpleiotropic associated variants that were found not to be associated with traits and the true negative rate (TNR) which is also specificity is the proportion of nonpleiotropic SNPs that are truly nonpleiotropic. Furthermore, the power of detecting true pleiotropy for all methods is then $power = 1 - FNR$, false-negative rate (FNR) is the proportion of pleiotropic variants that are not associated with both phenotypes.

2.4 | Identification of pleiotropic loci for GGE and FE

We applied ASSET, which showed superior performance for pleiotropy detection in the simulation study (see Section 4) to identify pleiotropic loci for GGE and FE. To this end, we were provided by the ILAE Consortium with the summary statistics of their previous GWAS on epilepsy (International League Against Epilepsy Consortium on Complex Epilepsies, 2018), with data originating from a number of different studies ([International League Against Epilepsy Consortium

on Complex Epilepsies, 2018], Supporting Information: Tables 1 and 6), but no further information on the sex and age distribution in these studies. The data were generated based on the following ethics statement (International League Against Epilepsy Consortium on Complex Epilepsies, 2018): "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." The data set used in our study comprised the summary statistics from the analysis of the non-Finnish European subset of both GGE and FE. The ILAE Consortium had reported genetic associations with subphenotypes of epilepsy as well as joint analysis of the subphenotypes that considered both epilepsy types to be identical (mega-analysis) (International League Against Epilepsy Consortium on Complex Epilepsies, 2018). This data set comprised 9095 FE cases, 3708 GGE cases, and 24,218 overlapping controls, and approximately five million SNPs present in both subphenotype datasets. FE and GGE cases were nonoverlapping and, thus, independent. Due to the overlapping controls, we estimated correlation statistics between Z statistics of traits using the sample sizes of cases, controls, and samples for traits (see [Bhattacharjee et al., 2012] for full detail) which are in turn were used in the DLM procedure to estimate pleiotropy p-values in this approach.

For this analysis, a number of quality control steps samples and SNPs, such as test for Hardy–Weinberg equilibrium (HWE) deviations, excess heterozygosity and exceeding relatedness, filtering for low minor allele frequency (MAF), and excessive genotype missingness, had already been performed by ILAE (please refer to [International League Against Epilepsy Consortium on Complex Epilepsies, 2018] for details). This included a sample and SNP removal of those presenting with genotype missing call rate exceeding 0.05, removal of SNPs with MAF<0.01, heterozygosity outside five SD across the sample set of the respective trait or a P-value < 10^{-10} from an exact test for HWE deviations. Furthermore, they excluded one sample from each pair that had an estimated average identity-by-descent (IBD) allele sharing (π) > 0.1875 or the complete pair with estimate π values > 0.9. However, on the summary statistics output provided by the ILAE consortium (International League Against Epilepsy Consortium on Complex Epilepsies, 2018), we compared the χ^2 of pairs of SNPs with LD value (r2 > 0.4) and we removed SNPs having χ^2 values greater than $\frac{3 \times \sqrt{\frac{SNP1_{\chi^2} + SNP2_{\chi^2}}{2}}}{(R^2)^2}$ to exclude SNPs with inflated χ^2 values. Finally, we analyzed only those SNPs that were contained in the datasets of both GGE and FE after quality control, finally including about 4.1 million SNPs in the pleiotropic analysis using ASSET.

2.4.1 | Follow-up of significant loci

SNPs that were found to be significant in the pleiotropic analysis with ASSET were mapped to genes using FUMA v1.3.7 (Watanabe et al., 2017) (https://fuma.ctglab.nl/). The loci harboring these significant SNPs were delineated by clustering SNPs in LD at r^2>0.2 within a ±250kb radius. Within a locus, the SNP with the smallest p-value was considered the "lead" SNP. We then performed functional annotation of the SNPs included in the above-defined loci to assess the potential consequences of these SNPs. To this end, we performed functional annotation of the variants that are in LD with one independent significant SNP using ANNOVAR (Wang et al., 2010). We also performed functional annotation using the RegulomeDB v1.1 database to check for evidence of SNPs affecting regulation, where SNPs with RegulomeDB scores <6 are considered to affect the regulation of the mapped gene (Boyle et al., 2012). Deleteriousness of SNPs was predicted by CADD scores v1.4; scores higher than 12.37 were considered deleterious as proposed by Kircher et al. (2014). Furthermore, we performed a tissue expression analysis using FUMA, based on the P-values from MAGMA v1.08 (de Leeuw et al., 2015) gene-set analysis and GTEx v8 expression data, to quantify the relationship between the average expression of a set of genes identified in the tissue and genetic association. We checked for previous reports on genetic association with epilepsy syndromes using the GWAS catalog (https://www.ebi.ac.uk/gwas/). Finally, we also performed a Bayesian colocalization test between GGE and FE to confirm whether the lead SNPs have a high probability of being associated and shared for both syndromes, using the R packages HyPrColoc v1.0 (Foley et al., 2021) ("hypothesis prioritization for multi-trait colocalization"; https://github.com/jrs95/hyprcoloc/) and coloc v5.1.0 (Giambartolomei et al., 2014; Wallace, 2020) (https://CRAN.R-project.org/package=coloc). More specifically, we estimated the posterior probability of colocalization as evidence that a variant is shared or associated for multiple traits using HyPrColoc and of association of both syndromes with the lead SNPs using coloc.

3 | RESULTS

3.1 | Simulation study results

We evaluated the relative performance of all five methods (Table 1) in identifying true nonnull signals in terms of power (Figure 1) and the FPR (Figure 2) under varying sample sizes, causal variant numbers, and effect sizes as well as percentages of overlap between causal SNPs for two traits and two different values for disease prevalence using 100 replications (see also Supporting Information: Tables S1 and S2 for exact numbers for the case of 5 and 10 causal SNPs, respectively). Our simulations revealed largely concordant trends for all five methods, however, we observed a few critical differences between the methods. We found cFDR, METAL, and PLACO to work computationally very fast compared with ASSET and CPBayes, with ASSET still performing faster than CPBayes.

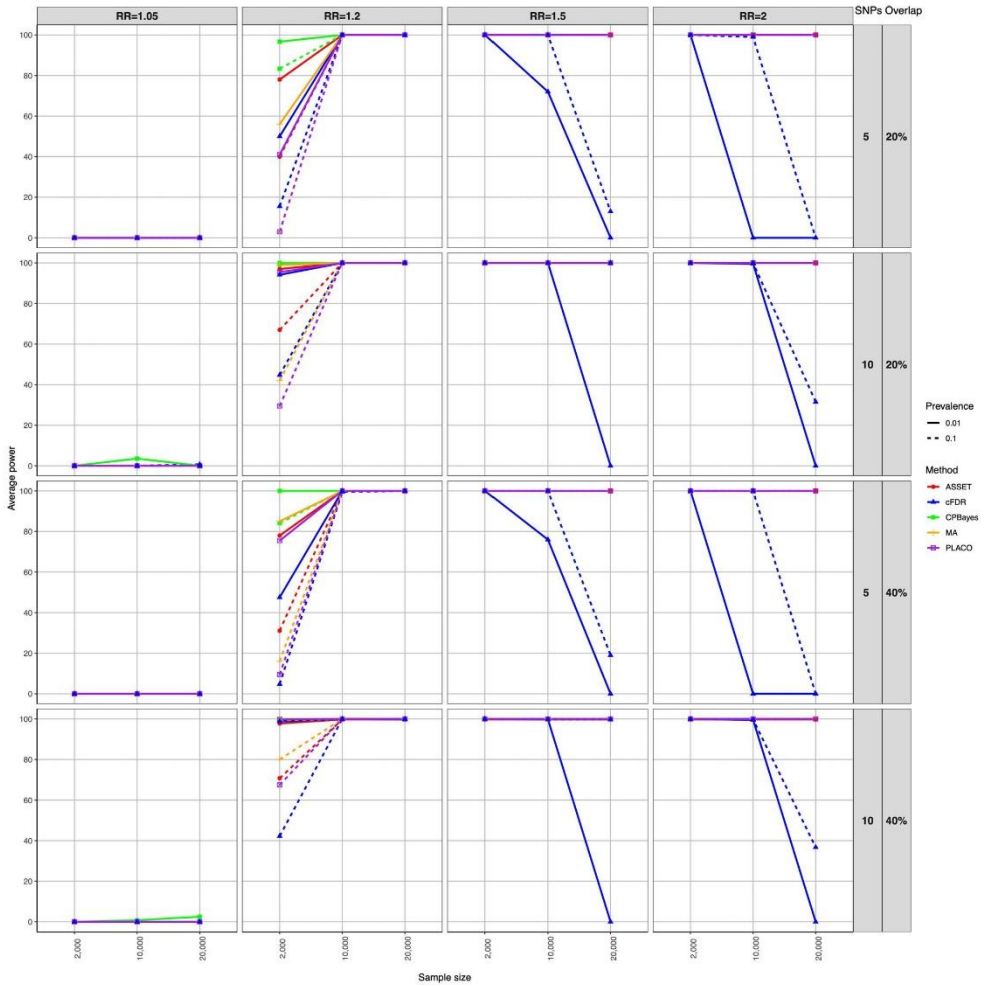


FIGURE 1 Power estimates for five univariate pleiotropy detection methods. Power estimates are averaged over 100 replications (see Section 2 for details). MA, meta-analysis; Overlap, proportion of causal SNPs for one trait that overlap with those for the other trait; Prevalence, prevalence of each of the two traits; RR, relative risk per single causal variant; SNPs, number of causal SNPs modeled to increase susceptibility to a particular trait; Sample size, total sample size with equal proportions of cases and controls.

3.1.1 | Power

With low variant effect sizes of 1.05, all five applied methods (Table 1) were virtually powerless to detect any pleiotropic variants, with power estimates equal or close to zero regardless of sample size, causal SNP number, or degree of causal SNP overlap (Figure 1); only CPBayes appeared to pick up pleiotropic variants in a small fraction of the replications. For relative risks of 1.2, which are common in genome-wide association studies, and moderate samples sizes, we observed notable differences between the methods. While CPBayes outperformed all other methods with respect to power, ASSET and MA followed as second most powerful methods depending on the considered scenario of variant number and degree of overlap. PLACO and cFDR always presented with lower power than the previous methods, although at differing extent. With larger sample sizes (10,000 or larger) or larger relative risks (1.5 or larger), these differences ceased to exist, rendering all methods, with one

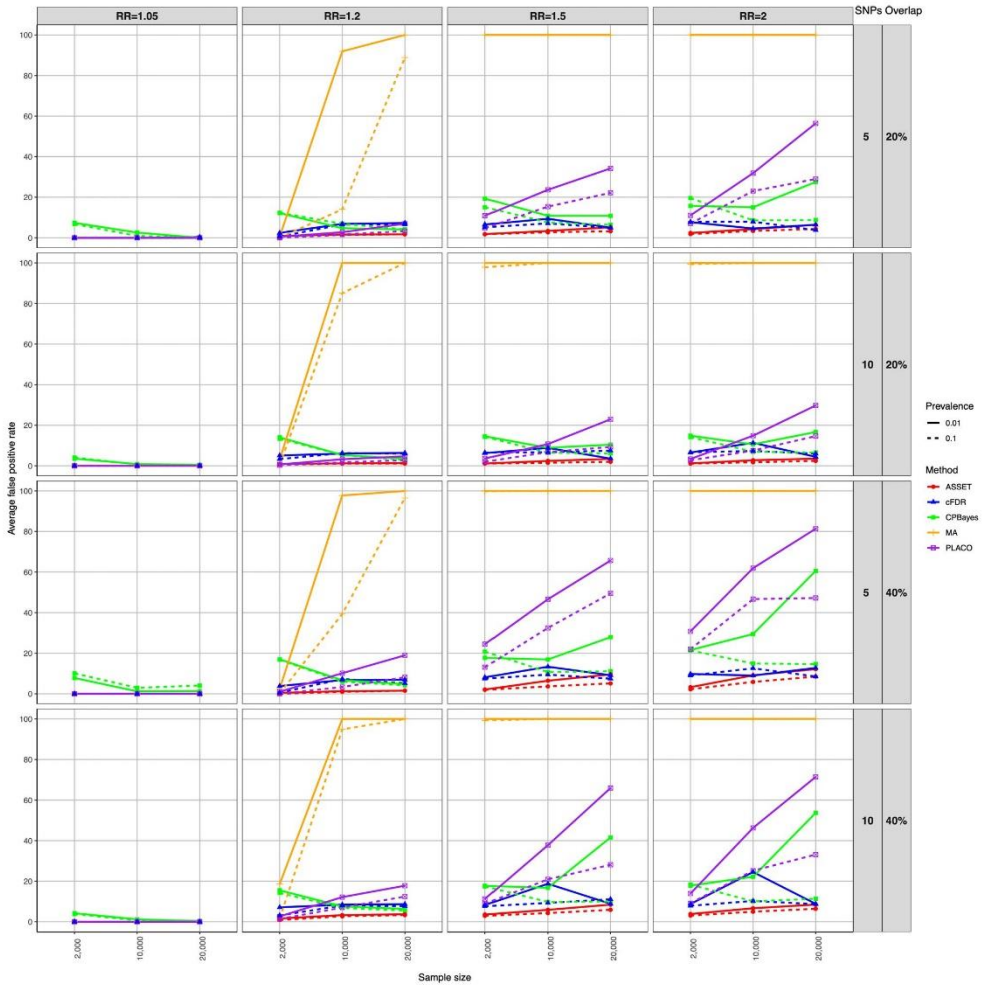


FIGURE 2 False-positive rate (FPR) estimates for five univariate pleiotropy detection methods. FPR (type I error) estimates are averaged over 100 replications (see Section 2 for details). MA, meta-analysis; Overlap, proportion of causal SNPs for one trait that overlap with those for the other trait; Prevalence, prevalence of each of the two traits; RR, relative risk per single causal variant; SNPs, number of causal SNPs modeled to increase susceptibility to a particular trait; Sample size, total sample size with equal proportions of cases and controls.

exception, highly and uniformly powerful to detect pleiotropic variants for both traits. Strangely, for scenarios of higher relative risks (1.5 or larger), cFDR resulted in decreasing power levels for increasing sample sizes, eventually assuming values equal or close to zero. In general, a larger number of causal SNPs, as well as increasing proportions of causal SNPs overlapping between the two traits, led to higher power to detect pleiotropic variants and a closer resemblance of the power performance across all five methods. Not surprisingly, larger sample sizes also showed a general trend toward increasing power for each of the methods, except for the counterintuitive cFDR behavior.

3.1.2 | FPR

The FPR differed vastly between the methods (Figure 2). For relative risks of 1.05, all methods showed very low FPR levels across all considered sample sizes, with values being equal or close to zero,

except for CPBayes. In particular, for moderate sample sizes, the FPR of CPBayes approached values of up to 10%. This pattern of elevated FPR values continued with larger relative risks (≥1.2) and now additionally characterized the error rates of MA, PLACO, and MA. Depending on the considered scenario, the FPR assumed values of up to 60% for CPBayes, but even up to 80% for PLACO and 100% for MA. PLACO showed a general trend toward higher FPR values with growing sample sizes and relative risks, whereas CPBayes and cFDR did not show a clear trend with respect to sample size or variant relative risk. At a relative risk of 1.2 and a sample size of 2000 individuals, all approaches except CPBayes keep the 5% FPR level. Just with sample sizes of 10,000 or larger and relative risks of 1.5 or larger, MA always wrongly classifies variants as pleiotropic when they in fact are causal only for one of the two traits. On the other hand, cFDR and in particular ASSET did not yield such high FPR values in any scenario. Values for cFDR reached about 20% and 25% in the two scenarios, respectively, but were generally closer to or below 10%. ASSET showed the best performance across all simulations, generally keeping the 5% level, except for 20,000 samples and effect sizes of 1.5 or larger where the FPR did not exceed 9%. In general, a larger number of causal SNPs led to decreasing FPR levels, although often only slightly, whereas a higher overlap of causal SNPs between the two traits increased the FPR, except for MA and CPBayes. Not surprisingly, larger sample sizes led to increasing power for each of the methods.

3.1.3 | Impact of trait prevalence

Trait prevalence seemed to have a modest effect on the performance of the methods. The strongest differences could be seen for variant relative risks of 1.2 and a sample size of 2000 regardless of the number of causal SNPs and their overlap between traits. Common traits (prevalence of 10%) resulted in lower power values for moderate relative risks of 1.2 and moderate sample sizes of 2000 individuals (Figure 1). Otherwise, differences were indiscernible. For the FPR (Figure 2), a more frequent trait (prevalence of 10%) almost always resulted in, sometimes strongly, reduced FPR values than for a moderate common trait (prevalence of 1%). Again, ASSET seemed to be the method to be least affected by changes to the trait prevalence.

3.2 | Pleiotropy detection in the ILAE data set

Since we had identified ASSET as the method that gave a superior trade-off between the FPR and the power to detect pleiotropy in our simulations, we only applied this approach to the data set provided by the ILAE consortium. Using ASSET, we identified 40 SNPs in two loci being associated, at a genome-wide significance level, with both epilepsy subphenotypes GGE and FE (Figure 3). The two loci are at 2q24.3 and 17q21.32 (Table 2). Locus 2q24.3 (Figure 4) had already been reported and mapped to *SCNA1*, *SCNA2*, *SCNA3*, and *TTC21B* in the ILAE mega-analysis (International League Against Epilepsy Consortium on Complex Epilepsies, 2018) for both forms of epilepsy which have been implicated to have an effect on the risk of different forms of epilepsy (Epicure et al., 2012; Feenstra et al., 2014; International League Against Epilepsy Consortium on Complex Epilepsies, 2018; International League Against Epilepsy Consortium on Complex Epilepsies. Electronic address, 2014). However, locus 17q21.32 has not been reported before, rendering it a new putative pleiotropic locus for FE and GGE. This locus comprises the genes *SKAP1*, *OSBPL7*, *SP6*, *SP2*, *PNPO*, *PRR15L*, *CDK5RAP3*, *COPZ2*, *NFE2L1*, *CBX1*, *SNX11*, *HOXB1*, *HOXB2*, and *HOXB3* (Figure 5). Functional annotation of the 40 significant SNPs (Figure 6) showed that 11% of the SNPs are intergenic, 23% are intronic, and 61% are noncoding transcript intron variants (see Supporting Information: Figure S2 for locus-specific results).

Furthermore, RegulomeDB scores below 6 were observed for 45% of the significant SNPs, indicating that these SNPs indeed affect gene transcription. We then checked MAGMA tissue-specific expression of the genes in FUMA and found them to be preferentially expressed in the brain compared with other tissues using GTEx tissue expression data of 53 tissue types and genetic association (Figure 6). From the analysis in FUMA, we identified three lead SNPs. SNPs rs60055328 and rs2212656 mapped to locus 2q24.3, whereas rs16955463 mapped to 17q21.32. According to the GWAS catalog, rs60055328 has already been implicated in epilepsy, febrile seizures (Feenstra et al., 2014), and generalized epilepsy (Epicure et al., 2012; International League Against Epilepsy Consortium on Complex Epilepsies, 2018).

Locus 2q24.3 had a HyPrColoc-computed posterior probability of colocalization of 77%, with rs2212656 explaining 68% of this probability. The regional probability that one or more SNPs in the region have shared associations across the syndromes was 94%, thereby strongly indicating a shared association between GGE and FE. However, based on the evidence of colocalization, rs60055328 does not seem to be associated with both traits but is part of the credible set of SNPs that explains 95% of the posterior probability of colocalization. At locus 17q21.32, the regional posterior probability of shared association for rs16955463 was only 1.56%. However, coloc-computed posterior evidence of association of both epilepsy forms with different associated variants (PP_{H3}) was 99.6% and 85.0% for locus 2q24.3 and 17q21.32, respectively, thereby further reinforcing the notion that different but highly correlated SNPs in the same region are associated with both syndromes.

For method comparison only, we also applied the other four methods to the ILAE data set (Supporting Information: Figures S3-S6) and assessed their overlap in significant associations with each other (Supporting Information: Figures S7). Out of the 40 SNPs identified by ASSET, all except one were also identified by at least one other method. The largest overlap was observed with MA (39 out of 40) and PLACO (32/40), while cFDR (23/40) and CPBayes (5/40) shared lower proportions. Interestingly, 32 out the 33 associations identified by PLACO were also found by ASSET. Notably, both CPBayes (103) and MA (126) reported several dozen associations exclusively identified by them and another 81 identified by both.

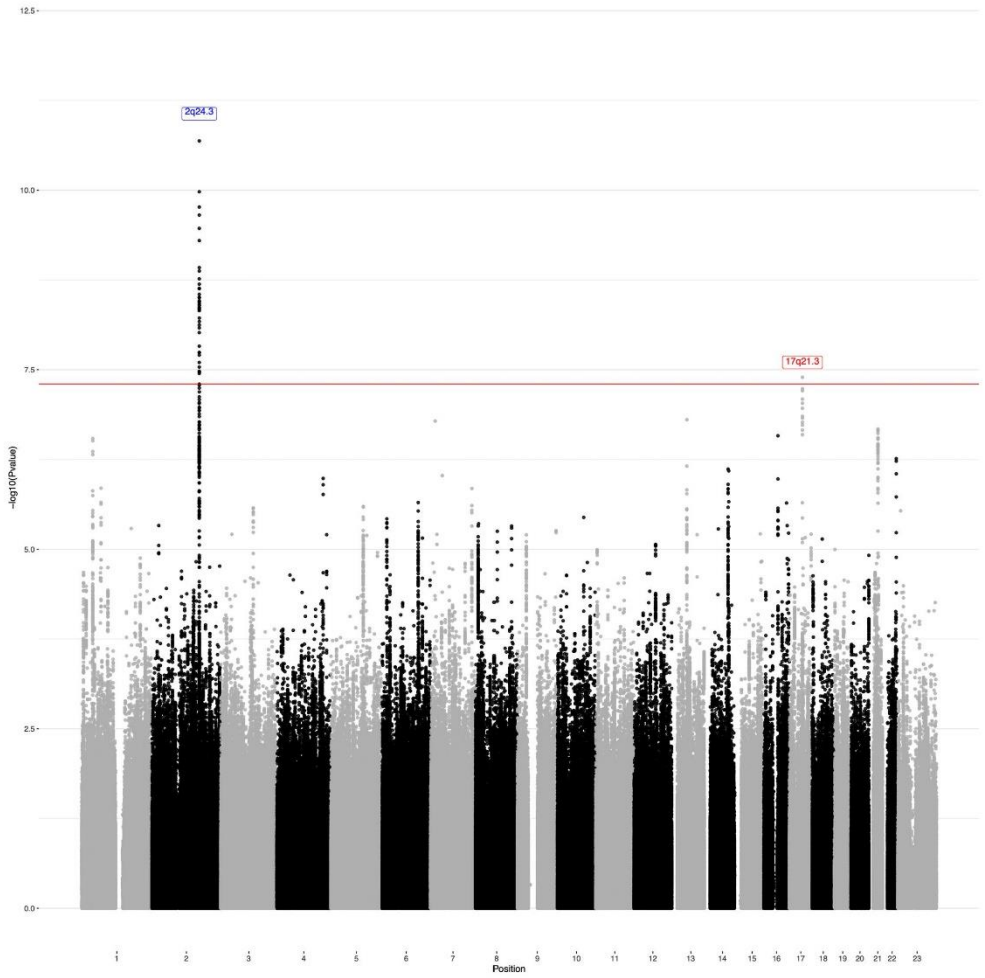


FIGURE 3 Manhattan plot of pleiotropy association testing between genetic generalized epilepsies (GGE) and focal epilepsies (FE) using ASSET. Chromosomal variant position is given on the x-axis, while -$\log_{10}$ transformed *P*-values are given on the *y*-axis. The red horizontal line denotes the genome-wide significance level. The previously identified locus (2q24.3) is labeled in blue while the new pleiotropic locus (17q21.3) is labeled in red.

TABLE 2 Genome-wide significant SNPs and their prioritized genes

	Pleiotropy significant SNPs and nearest genes				Summary statistics of the SNPs		ASSET result	
Locus	Lead SNPs (Risk allele)	MAF	CADD score	Nearest genes	*P*-value (GGE)	*P*-value (FE)	OR (95% CI)	*P*
2q24	rs60055328 (A)	0.24	2.95	*SCN1A*, *SCN9A*, *TTC21B*	8.4×10^{-8}	7.3×10^{-9}	1.109 (1.072,1.148)	2.03×10^{-9}
	rs2212656 (C)	0.25	6.25		1.7×10^{-6}	1.5×10^{-6}	1.127 (1.008,1.167)	2.06×10^{-11}
17q21.32	rs16955463 (T)	0.26	13.44	*PNPO*, *SKAP1*	$2.3e^{-9}$	$8.3e^{-1}$	0.867 (0.828,0.909)	4.03×10^{-8}

Abbreviations: CI, confidence interval; MAF, minor allele frequency; OR, odds ratio; *P*, *P* value; SNP, single-nucleotide polymorphism.

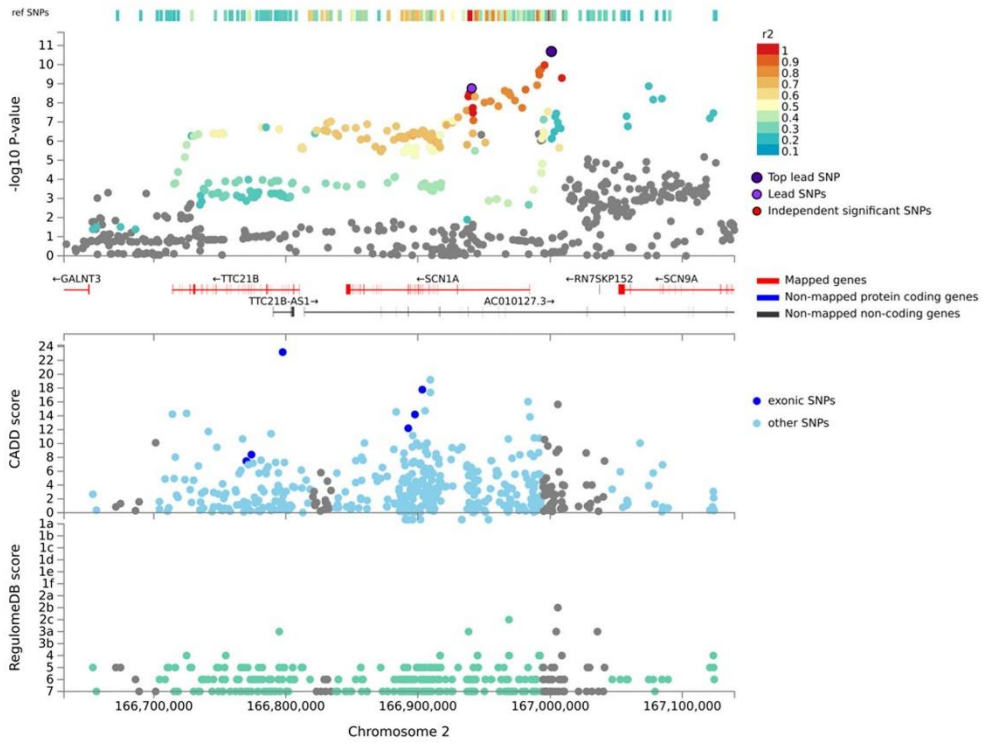


FIGURE 4 Region plots of locus 2q24.3. Given are pleiotropic *P*-values (genetic generalized epilepsies [GGE] and focal epilepsies [FE]) and LD values between tested markers, mapped genes, CADD scores, and Regulome DB scores (from top to bottom).

4 | DISCUSSION

In this study, we compared four recently proposed methods for univariate pleiotropy detection with single variants alongside classical MA. To this end, we performed forward simulations for a large population of European ancestry and repeatedly assigned affection status for two pleiotropic traits, respectively, to all individuals of this population while considering a variety of different parameters that may impact the ability of these methods to detect pleiotropy in a cross design, including sample size, number and effect size of trait-associated genetic variants as well as their overlap between traits, and the trait prevalence. We have chosen exemplary values for these parameters in a range that has been often observed for GWAS and that are plausible to expect for future studies on pleiotropy. Notably, we also considered multivariant trait etiology models to include more than one causal genetic variant to be shared between these traits, mirroring the likely situation in real datasets. Trait assignment was based on the ALTM which is a well-established theoretical model that has been calibrated to empirical data and that has been successfully used to describe the genetic architecture of different traits, including type II diabetes (Agarwala et al., 2013). While we did not explicitly model the etiology of particular epilepsy forms, we strongly believe that the ALTM provides a very good approximation for them.

4.1 | MA

While MA, as implemented in METAL (Willer et al., 2010), turned out to be among the more powerful methods for pleiotropy detection, its exceptionally high FPR forbids its application to detect pleiotropic variants or loci. A likely explanation for this behavior is that a very strong association of a variant with one trait will dominate the *P*-value from the MA and decrease it below the threshold considered significant, even though this variant shows no association at all with the second trait. The same observation is to be expected for mega-analysis in which several (sub-)phenotypes are analyzed jointly in a single association test. Our observation also supports the suggestion of superior performance of ASSET by the respective publication (Bhattacharjee et al., 2012). Meta- or mega-analysis results, such as those previously performed by the ILAE Consortium, therefore do

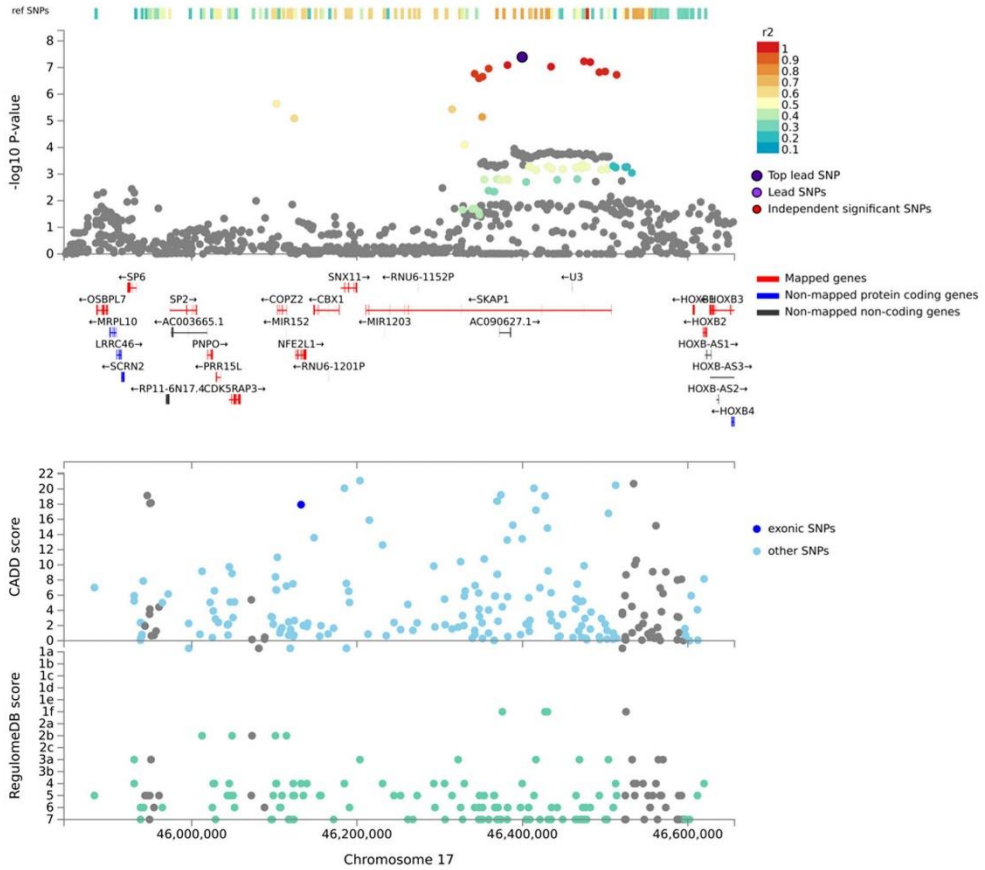


FIGURE 5 Region plots of locus 17q21.32. Given are pleiotropic *P*-values (genetic generalized epilepsies [GGE] and focal epilepsies [FE]) and LD values between tested markers, mapped genes, CADD scores, and Regulome DB scores (from top to bottom).

not allow to draw any conclusions about a potential pleiotropic nature of significant variants in these analyses.

4.2 | Recommendations for method application

All evaluated methods had virtually no power to detect pleiotropic variants with low effect sizes (relative risk 1.05), for sample sizes up to 20,000. Thus, studies aiming for pleiotropic variants of such small effect likely require 100,000 samples or more to achieve a decent power. For many traits, such sample sizes are likely out of reach, rendering the identification of weak pleiotropic variants infeasible. Furthermore, while all four newer methods appear to be able to detect pleiotropic variants with larger sample sizes or variant effect sizes, they substantially differ in their ability to hold a low FPR, preferably below 5%. ASSET, despite being the first of the four considered approaches to have been proposed, appears to provide the best trade-off between power and controlling the FPR. While CPBayes appeared to be the most powerful method, ASSET always came in as second or third best. However, CPBayes, together with cFDR and PLACO, presented with strongly elevated FPR values. While CPBayes could be considered to represent a different compromise between power and FPR than ASSET, any significant finding obtained from applying CPBayes may then with considerable likelihood represent a false-positive result, thereby diminishing the success chances of subsequent variant follow-up. Since ASSET is not or only moderately affected by increased FPR levels, we recommend its preferable application to real-world datasets compared with that of CPBayes, cFDR, and PLACO. The counterintuitive downward trend in power observed for cFDR with growing relative risks is apparently caused by a computational error of the R function provided by Liley and Wallace (2015), for which very small *p*-values

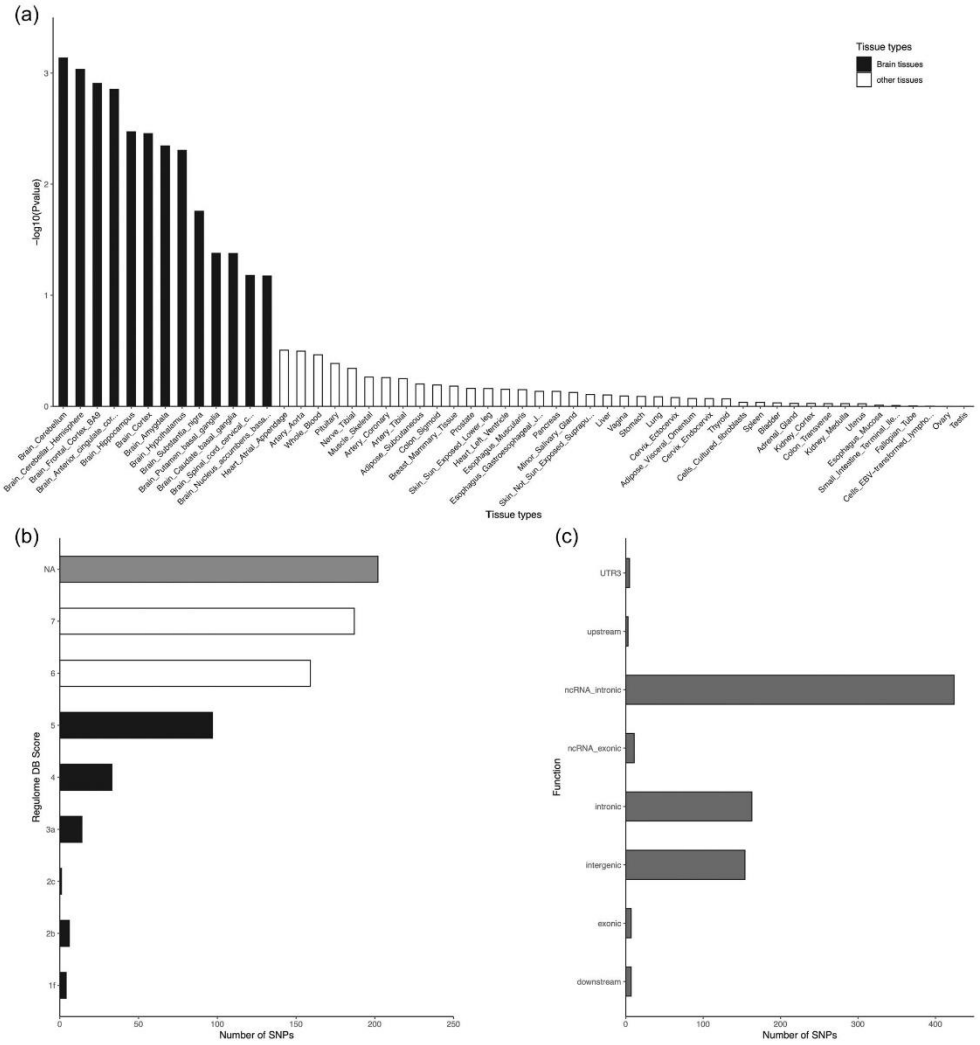


FIGURE 6 Functional annotation and enrichment of significant single-nucleotide polymorphism (SNPs) at loci 2q24.3 and 17q21.32. (a) Enrichment of the genes expressed in 53 tissues of GTEx tissue expression data using MAGMA. Brain tissues are marked by dark-filled bars. (b) RegulomeDB scores of all SNPs that were either genome-wide significant in the pleiotropic analysis or correlated with such SNPs ($r^2 > 0.2$). Dark-filled bars denote SNPs with RegulomeDB scores < 6 (considered to represent evidence for affecting regulation; see Section 2), whereas the gray-filled bar denotes SNPs that are not available in RegulomeDB ("NA"). (c) Functional categories of all SNPs that were either genome-wide significant in the pleiotropic analysis or correlated with such SNPs ($r^2 > 0.2$).

are reported as missing values when the result is generated. After a revised software implementation solving this issue, cFDR may be likely considered as a slightly less powerful but faster alternative to ASSET.

Our study results are consistent with some previously published method comparisons. For PLACO, we found a power of 3.0% at a sample size of 2000 for a relative risk of 1.2 and a prevalence of 10%, while the PLACO authors reported 7.1% power at an odds ratio of 1.5 for two studies for which the variant has the same direction of effect for both traits (Ray & Chatterjee, 2020). For larger sample sizes, the simulation results are not directly comparable but nevertheless show a similar trend. Regarding CPBayes, we could confirm

the reported superior power for CPBayes at low sample sizes and relative risks of 1.2 (Majumdar et al., 2018), which comes at the price of strongly inflated FPR levels. Also, for CPBayes (Majumdar et al., 2018), the authors' claims are partially confirmed but ASSET demonstrated lower specificity compared with CPBayes. We could not compare our results to those published in the original cFDR (Liley & Wallace, 2015) and ASSET (Bhattacharjee et al., 2012) publications because of large differences in the simulation designs. In particular, this included differences in the considered values for allele frequencies of causal variants, disease prevalence, sample size, and the number of studies, or traits, rendering the results from these original publications not directly comparable to our study.

4.3 | Pleiotropy between epilepsy subphenotypes

Using ASSET and its ability to correct for sample overlap, we identified two pleiotropic loci, namely 2q24.3 and 17q21.32, for FE and GGE. Evidence for simultaneous association of these syndromes to locus 2q24.3 had already been reported in ILAE mega-analysis (International League Against Epilepsy Consortium on Complex Epilepsies, 2018). Our results thereby confirm the pleiotropic nature of this locus. However, locus 17q21.32, although being already reported to be associated with GGE (Epicure et al., 2012) but not FE, represents a promising novel pleiotropic locus for both FE and GGE. The evidence from formal statistical testing using ASSET is further corroborated by various annotation, enrichment, expression, colocalization, and prioritization analyses. A statistical replication of this finding in an independent data set would provide further evidence once such a data set becomes available. From the analysis, three strong pleiotropic signals are identified which are likely to have effects on the regulation of *SCNA1*, *SCN9A*, *TTC21B*, *SKAP1*, and *PNPO* but not *SCN2A* and *SCN3A* according to mapping done in FUMA (Watanabe et al., 2017) and variant effect predictor (Howe et al., 2021). The 17q21.32 association peak is located at the src kinase-associated phosphoprotein 1 gene (*SKAP1*). *SKAP1* is positively involved in T-cell receptor signaling but is only weakly brain expressed (https://gtexportal.org/home/gene/SKAP1) and has not been functionally implicated in seizure disorders. Furthermore, the chromosome 17 locus of this study overlaps with the 17q21.32 region described as GGE-associated by the EPICURE Consortium in 2012 (Steffens et al., 2012). The EPICURE Consortium identified the pyridoxamine 5'-phosphate oxidase gene (*PNPO*) in ~230 kb distance to *SKAP1* as the most promising candidate for GGE. *PNPO* mutations and the resulting impairment of pyridoxine 5'-phosphate (PNP) or vitamin B6 metabolism and insufficient delivery of pyridoxal 5'-phosphate (PLP) to PLP-dependent enzymes have neuropathological consequences in neonates (Ghatge et al., 2021; Levtova et al., 2015; Lloreda-Garcia et al., 2017). PNPO is required for the synthesis of pyridoxal 5'-phosphate (PLP) whose role in neurotransmitter metabolism is thought to be the primary cause of PNPO-dependent neonatal epileptic encephalopathy (Mills et al., 2005).

Given the very similar prevalence values of GGE and FE (0.002 vs. 0.003, respectively (International League Against Epilepsy Consortium on Complex Epilepsies, 2018) Supporting Information: Table 11) and the sole use of summary statistics for the pleiotropy detection, we do not expect the substantially different numbers of GGE and FE cases in our study to have an impact on our study results. Furthermore, reported top single-trait associations tended to be generally fewer for FE compared with GGE despite the much larger sample size, indicating on average smaller effect sizes for FE. A random down-sampling of FE cases to match the number of GGE cases, as may appear desirable in, for example, classification tasks, would most likely have reduced the power to detect pleiotropy in the ILAE data set.

Comparative application of the five methods to the ILAE epilepsy datasets showed a large concordance between ASSET and PLACO in identified significant associations, a large set of associations exclusively identified by either CPBayes or MA and almost no overlap between ASSET and CPBayes. While it is unclear whether these associations represent true- or false-positives, the large excess of observed significant associations for CPBayes and MA is consistent with the largely increased FPRs that we observed for these two methods in our simulations. Also consistent with our simulation study is the much smaller number of reported associations observed for ASSET as one would except with a well-controlled FPR. The results from applying the five methods to two epilepsy datasets do, thus, well conform to expectations based on our simulation study.

4.4 | Limitations & future work

While we have considered a substantial number of scenarios, we could not consider all parameter values that may have been desirable due to the large computational burden involved. Our results are therefore uninformative about scenarios with intermediate parameter values, for example, sample sizes of 2000 cases and 2000 controls, larger numbers of shared causal genetic variants between traits or more than two traits. Studies that follow a cross-design for numerous parameters, such as this one, require considerable time and high computing power for data generation and analysis. This limits the number of parameters that could be jointly studied due to the exponential growth of computing time with the growing number of parameter combinations. Furthermore, since the population base data are generated according to some LD model, the LD pattern in the data may also influence the results. It has been clearly shown in the past by Su et al. (2011) that the LD pattern in simulated data reflects that of the reference data. Since we simulated a population of European ancestry, our results are not necessarily transferable to pleiotropy detection studies in populations of other continental origins. This should, however, only affect the actual values for power and FPR, not the relative performance order of the five methods considered here.

For future studies, it will be interesting to investigate the performance of pleiotropy detection methods for more than two

phenotypes, for more nuanced sharing of causal genetic variation and possibly different effects on the pleiotropic phenotypes, and for less common or rare causal variants, with the latter likely requiring more complex genotype simulation algorithms and larger reference sample sets. In any case, our study has revealed general trends that will likely help guide the design of follow-up studies.

5 | CONCLUSIONS

Based on extended computer simulations, we find that the ASSET method outperforms other univariate pleiotropy detection methods, including classical MA, cFDR, CPBayes, and PLACO, with respect to power and control of the FPR and recommend its use in future studies. Application of ASSET to GWAS summary statistics on generalized genetic epilepsies and on focal epilepsies, previously published by the ILAE consortium, confirmed the truly pleiotropic nature of the previously reported locus 2q24.3 (based on a mega-analysis) for these epilepsy forms and identified a novel putative pleiotropic locus at 17q21.32, the latter being corroborated by further database and bioinformatic annotation.

ACKNOWLEDGMENTS

The authors thank the Regional Computing Center of the University of Cologne (RRZK) for providing computing time for the simulation studies on the DFG-funded High-Performance Computing (HPC) system, CHEOPS as well as technical support. Experiments presented in this paper were carried out using the high-performance computing facilities of the University of Luxembourg (http://hpc.uni.lu). This study received support from the Research Unit FOR-2715 of the German Research Foundation (MN: NO755/6-1, NO755/13-1; HL: LE1030/16-1/2; HS: SCHU 3585/1-1) and the Fonds National de la Recherche (FNR; RK/PM: INTER/DFG/17/11583046). Patrick May obtained additional FNR funding as part of the National Centre of Excellence in Research on Parkinson's disease (NCER-PD, FNR11264123). The funding bodies had no role in the study design, data collection, analysis, and interpretation, or in writing the manuscript. Some of the data reported in this paper were collected as part of a project undertaken by the International League against Epilepsy (ILAE) and some of the authors are experts selected by the ILAE. Opinions expressed by the authors, however, do not necessarily represent the policy or position of the ILAE. Open Access funding enabled and organized by Projekt DEAL.

WEB RESOURCES

ANNOVAR: http://www.openbioinformatics.org/annovar/
ASSET v2.10.0: https://bioconductor.org/packages/ASSET/
coloc v5.1: https://CRAN.R-project.org/package=coloc
cFDR v1.1: https://github.com/jamesliley/cFDR-common-controls
CPBayes v1.1.0: https://github.com/ArunabhaCodes/CPBayes/
FUMA v1.3.7: https://fuma.ctglab.nl/
GTEx: https://gtexportal.org/
GWAS catalog: https://www.ebi.ac.uk/gwas/
Hapgen2 v2.2.0: https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html
HyPrColoc v1.0: https://github.com/jrs95/hyprcoloc/
MAGMA v1.08: https://ctg.cncr.nl/software/magma/
METAL: http://csg.sph.umich.edu/abecasis/metal/download
PLACO v0.1.1: https://github.com/RayDebashree/PLACO/
PLINK v1.9: https://www.cog-genomics.org/plink2/
R: https://www.r-project.org/
RegulomeDB v1.1: https://regulomedb.org/

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Patrick May http://orcid.org/0000-0001-8698-3770
Roland Krause http://orcid.org/0000-0001-9938-7126
Michael Nothnagel http://orcid.org/0000-0001-8305-7114

REFERENCES

Agarwala, V., Flannick, J., Sunyaev, S., Go, T. D. C., & Altshuler, D. (2013). Evaluating empirical bounds on complex disease genetic architecture. *Nature Genetics*, *45*(12), 1418–1427. https://doi.org/10.1038/ng.2804

Agresti, A. (2003). *Categorical data analysis* (Vol. 482). John Wiley & Sons.

Andreassen, O. A., Thompson, W. K., Schork, A. J., Ripke, S., Mattingsdal, M., Kelsoe, J. R., Kendler, K. S., O'Donovan, M. C., Rujescu, D., Werge, T., Sklar, P., Psychiatric Genomics Consortium, (, Bipolar Disorder and Schizophrenia Working, G., Roddey, J. C., Chen, C. H., McEvoy, L., Desikan, R. S., Djurovic, S., & Dale, A. M. (2013). Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genetics*, *9*(4), e1003455. https://doi.org/10.1371/journal.pgen.1003455

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

Bhattacharjee, S., Rajaraman, P., Jacobs, K. B., Wheeler, W. A., Melin, B. S., Hartge, P., Gliomascan, C., Yeager, M., Chung, C. C., Chanock, S. J., & Chatterjee, N. (2012). A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *American Journal of Human Genetics*, *90*(5), 821–835. https://doi.org/10.1016/j.ajhg.2012.03.015

Bien, S. A., & Peters, U. (2019). Moving from one to many: Insights from the growing list of pleiotropic cancer risk genes. *British Journal of Cancer*, *120*(12), 1087–1089. https://doi.org/10.1038/s41416-019-0475-9

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., Cherry, J. M., & Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, *22*(9), 1790–1797. https://doi.org/10.1101/gr.137323.112

Carvajal-Rodriguez, A. (2010). Simulation of genes and genomes forward in time. *Current Genomics*, *11*(1), 58–61. https://doi.org/10.2174/138920210790218007

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*, *4*, 7. https://doi.org/10.1186/s13742-015-0047-8

Chung, J., Jun, G. R., Dupuis, J., & Farrer, L. A. (2019). Comparison of methods for multivariate gene-based association tests for complex

diseases using common variants. *Eur J Hum Genet*, *27*(5), 811–823. https://doi.org/10.1038/s41431-018-0327-8

Cooper, C. (2019). Global, regional, and national burden of neurological disorders, 1990-2016: A systematic analysis for the global burden of disease study 2016. *The Lancet Neurology*, *18*(4), 357–375.

deLeeuw, C. A., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: Generalized gene-set analysis of GWAS data. *PLoS Computational Biology*, *11*(4), e1004219. https://doi.org/10.1371/journal.pcbi.1004219

Epicure, C., Eminet, C., Steffens, M., Leu, C., Ruppert, A. K., Zara, F., Striano, P., Robbiano, A., Capovilla, G., Tinuper, P., Gambardella, A., Bianchi, A., La Neve, A., Crichiutti, G., deKovel, C. G., Kasteleijn-Nolst Trenité, D., deHaan, G. J., Lindhout, D., Gaus, V., … Sander, T. (2012). Genome-wide association analysis of genetic generalized epilepsies implicates susceptibility loci at 1q43, 2p16.1, 2q22.3 and 17q21.32. *Human Molecular Genetics*, *21*(24), 5359–5372. https://doi.org/10.1093/hmg/dds373

Feenstra, B., Pasternak, B., Geller, F., Carstensen, L., Wang, T., Huang, F., Eitson, J. L., Hollegaard, M. V., Svanström, H., Vestergaard, M., Hougaard, D. M., Schoggins, J. W., Jan, L. Y., Melbye, M., & Hviid, A. (2014). Common variants associated with general and MMR vaccine-related febrile seizures. *Nature Genetics*, *46*(12), 1274–1282. https://doi.org/10.1038/ng.3129

Fitzmaurice, G. M., & Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, *80*(1), 141–151.

Foley, C. N., Staley, J. R., Breen, P. G., Sun, B. B., Kirk, P. D. W., Burgess, S., & Howson, J. M. M. (2021). A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nature Communications*, *12*(1), 764. https://doi.org/10.1038/s41467-020-20885-8

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., Anderson, C. A., Bis, J. C., Bumpstead, S., Ellinghaus, D., Festen, E. M., Georges, M., Green, T., Haritunians, T., Jostins, L., … Parkes, M. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, *42*(12), 1118–1125. https://doi.org/10.1038/ng.717

Genetic Analysis of Psoriasis Consortium, the Wellcome Trust Case Control Consortium, Strange, A., Capon, F., Spencer, C. C., Knight, J., Weale, M. E., Allen, M. H., Barton, A., Band, G., Bellenguez, C., Bergboer, J. G., Blackwell, J. M., Bramon, E., Bumpstead, S. J., Casas, J. P., Cork, M. J., Corvin, A., Deloukas, P., Dilthey, A., … Trembath, R. C. (2010). A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nature Genetics*, *42*(11), 985–990. https://doi.org/10.1038/ng.694

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Ghatge, M. S., Al Mughram, M., Omar, A. M., & Safo, M. K. (2021). Inborn errors in the vitamin B6 salvage enzymes associated with neonatal epileptic encephalopathy and other pathologies. *Biochimie*, *183*, 18–29. https://doi.org/10.1016/j.biochi.2020.12.025

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, *10*(5), e1004383. https://doi.org/10.1371/journal.pgen.1004383

Haiman, C. A., Le Marchand, L., Yamamato, J., Stram, D. O., Sheng, X., Kolonel, L. N., Wu, A. H., Reich, D., & Henderson, B. E. (2007). A common genetic risk factor for colorectal and prostate cancer. *Nature Genetics*, *39*(8), 954–956. https://doi.org/10.1038/ng2098

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., … Flicek, P. (2021). Ensembl 2021. *Nucleic Acids Research*, *49*(D1), D884–D891. https://doi.org/10.1093/nar/gkaa942

International League Against Epilepsy Consortium on Complex Epilepsies. (2018). Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. *Nature Communications*, *9*(1), 5269. https://doi.org/10.1038/s41467-018-07524-z

International League Against Epilepsy Consortium on Complex Epilepsies. Electronic address, e.-a. u. e. a. (2014). Genetic determinants of common epilepsies: A meta-analysis of genome-wide association studies. *Lancet Neurology*, *13*(9), 893–903. https://doi.org/10.1016/S1474-4422(14)70171-1

Jolliffe, I. T. (2002). Graphical representation of data using principal components. In *Principal Component Analysis* (pp. 78–110). Springer Series in Statistics. Springe. https://doi.org/10.1007/0-387-22440-8_5

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Kulminski, A. M., Loika, Y., Huang, J., Arbeev, K. G., Bagley, O., Ukraintseva, S., Yashin, A. I., & Culminskaya, I. (2019). Pleiotropic meta-analysis of age-related phenotypes addressing evolutionary uncertainty in their molecular mechanisms. *Frontiers in Genetics*, *10*, 433. https://doi.org/10.3389/fgene.2019.00433

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974. https://www.ncbi.nlm.nih.gov/pubmed/7168798

Levtova, A., Camuzeaux, S., Laberge, A. M., Allard, P., Brunel-Guitton, C., Diadori, P., Rossignol, E., Hyland, K., Clayton, P. T., Mills, P. B., & Mitchell, G. A. (2015). Normal cerebrospinal fluid pyridoxal 5'-phosphate level in a PNPO-Deficient patient with neonatal-onset epileptic encephalopathy. *JIMD Reports*, *22*, 67–75. https://doi.org/10.1007/8904_2015_413

Li, N., & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, *165*(4), 2213–2233.

Liley, J., & Wallace, C. (2015). A pleiotropy-informed Bayesian false discovery rate adapted to a shared control design finds new disease associations from GWAS summary statistics. *PLoS Genetics*, *11*(2), e1004926. https://doi.org/10.1371/journal.pgen.1004926

Lloreda-Garcia, J. M., Fernandez-Fructuoso, J. R., Martinez-Ferrandez, C., & Fuentes-Gutierrez, C. (2017). Severe fetal anemia and neonatal epileptic encephalopathy caused by a novel PNPO mutation. *Revue Neurologique*, *65*(7), 335–336. https://www.ncbi.nlm.nih.gov/pubmed/28929476

Majumdar, A., Haldar, T., Bhattacharya, S., & Witte, J. S. (2018). An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations. *PLoS Genetics*, *14*(2), e1007139. https://doi.org/10.1371/journal.pgen.1007139

Mills, P. B., Surtees, R. A., Champion, M. P., Beesley, C. E., Dalton, N., Scambler, P. J., Heales, S. J., Briddon, A., Scheimberg, I., Hoffmann, G. F., Zschocke, J., & Clayton, P. T. (2005). Neonatal epileptic encephalopathy caused by mutations in the PNPO gene encoding pyridox(am)ine 5'-phosphate oxidase. *Human Molecular Genetics*, *14*(8), 1077–1086. https://doi.org/10.1093/hmg/ddi120

Morris, A. P., Lindgren, C. M., Zeggini, E., Timpson, N. J., Frayling, T. M., Hattersley, A. T., & McCarthy, M. I. (2010). A powerful approach to sub-phenotype analysis in population-based genetic association studies. *Genetic Epidemiology*, *34*(4), 335–343. https://doi.org/10.1002/gepi.20486

Ottman, R. (2005). Analysis of genetically complex epilepsies. *Epilepsia*, *46*, 7–14.

Pomerantz, M. M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M. P., Doddapaneni, H., Beckwith, C. A., Chan, J. A., Hills, A., Davis, M., Yao, K., Kehoe, S. M., Lenz, H. J., Haiman, C. A., Yan, C., Henderson, B. E., Frenkel, B., Barretina, J., Bass, A., … Freedman, M. L. (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nature Genetics*, *41*(8), 882–884. https://doi.org/10.1038/ng.403

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., deBakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575.

Ray, D., & Chatterjee, N. (2020). A powerful method for pleiotropic analysis under composite null hypothesis identifies novel shared loci between type 2 diabetes and prostate cancer. *PLoS Genetics*, *16*(12), e1009218. https://doi.org/10.1371/journal.pgen.1009218

Salinas, Y. D., Wang, Z., & DeWan, A. T. (2018). Statistical analysis of multiple phenotypes in genetic epidemiologic studies: From cross-phenotype associations to pleiotropy. *American Journal of Epidemiology*, *187*(4), 855–863. https://doi.org/10.1093/aje/kwx296

Schaid, D. J., Tong, X., Batzler, A., Sinnwell, J. P., Qing, J., & Biernacka, J. M. (2019). Multivariate generalized linear model for genetic pleiotropy. *Biostatistics*, *20*(1), 111–128. https://doi.org/10.1093/biostatistics/kxx067

Singh, G., & Sander, J. W. (2020). The global burden of epilepsy report: Implications for low- and middle-income countries. *Epilepsy and Behavior*, *105*, 106949. https://doi.org/10.1016/j.yebeh.2020.106949

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., & Smoller, J. W. (2013). Pleiotropy in complex traits: Challenges and strategies. *Nature Reviews Genetics*, *14*(7), 483–495. https://doi.org/10.1038/nrg3461

Stearns, F. W. (2010). One hundred years of pleiotropy: A retrospective. *Genetics*, *186*(3), 767–773. https://doi.org/10.1534/genetics.110.122549

Steffens, M., Leu, C., Steffens, Ruppert, A. K., Zara, F., Striano, P., Robbiano, A., T., Capovilla, G., Tinuper, P., Gambardella, A., Bianchi, A., La Neve, A., Crichiutti, G., deKovel, C. G. F., Kasteleijn-Nolst Trenite, D., deHaan, G. J., Lindhout, D., Gaus, V., Schmitz, B., Janz, D., … Sander, T. (2012). Genome-wide association analysis of genetic generalized epilepsies implicates susceptibility loci at 1q43, 2p16.1, 2q22.3 and 17q21.32. *Human Molecular Genetics*, *21*(24), 5359–5372. https://doi.org/10.1093/hmg/dds373

Su, Z., Marchini, J., & Donnelly, P. (2011). HAPGEN2: Simulation of multiple disease SNPs. *Bioinformatics*, *27*(16), 2304–2305. https://doi.org/10.1093/bioinformatics/btr341

Thomas, G., Jacobs, K. B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., Welch, R., Hutchinson, A., Crenshaw, A., Cancel-Tassin, G., Staats, B. J., Wang, Z., Gonzalez-Bosquet, J., Fang, J., Deng, X., Berndt, S. I., Calle, E. E., … Chanock, S. J. (2008). Multiple loci identified in a genome-wide association study of prostate cancer. *Nature Genetics*, *40*(3), 310–315. https://doi.org/10.1038/ng.91

Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S., Penegar, S., Chandler, I., Gorman, M., Wood, W., Barclay, E., Lubbe, S., Martin, L., Sellick, G., Jaeger, E., Hubner, R., Wild, R., Rowan, A., Fielding, S., … Houlston, R. (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature Genetics*, *39*(8), 984–988. https://doi.org/10.1038/ng2085

Tuupanen, S., Turunen, M., Lehtonen, R., Hallikas, O., Vanharanta, S., Kivioja, T., & Aaltonen, L. A. (2009). The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced wnt signaling. *Nature Genetics*, *41*(8), 885–890. https://doi.org/10.1038/ng.406

vanRheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H., & Wray, N. R. (2019). Genetic correlations of polygenic disease traits: From theory to practice. *Nature Reviews Genetics*, *20*(10), 567–581. https://doi.org/10.1038/s41576-019-0137-z

Verbeke, G., Molenberghs, G., & Rizopoulos, D. (2010). Random effects models for longitudinal data. *In K. van Montfort, J. Oud, & A. Satorra (Eds.)*, Longitudinal research with latent variables. Dordrecht. (pp. 37–96). Springer https://pure.eur.nl/en/publications/random-effects-models-for-longitudinal-data

Wallace, C. (2020). Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genetics*, *16*(4), e1008720. https://doi.org/10.1371/journal.pgen.1008720

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Watanabe, K., Taskesen, E., vanBochoven, A., & Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*, *8*(1), 1826. https://doi.org/10.1038/s41467-017-01261-5

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, *26*(17), 2190–2191. https://doi.org/10.1093/bioinformatics/btq340

Wolking, S., Schulz, H., Nies, A. T., McCormack, M., Schaeffeler, E., Auce, P., & Lerche, H. (2020). Pharmacoresponse in genetic generalized epilepsy: A genome-wide association study. *Pharmacogenomics*, *21*(5), 325–335. https://doi.org/10.2217/pgs-2019-0179

Yang, Q., & Wang, Y. (2012). Methods for analyzing multivariate phenotypes in genetic association studies. *Journal of Probability and Statistics*, *2012*, 652569. https://doi.org/10.1155/2012/652569

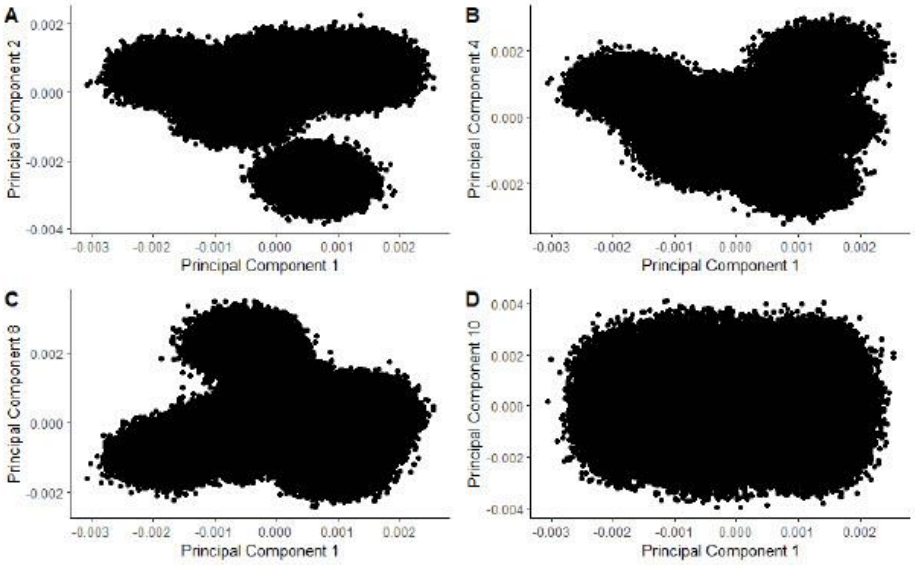
Zanke, B. W., Greenwood, C. M., Rangrej, J., Kustra, R., Tenesa, A., Farrington, S. M., & Dunlop, M. G. (2007). Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature Genetics*, *39*(8), 989–994. https://doi.org/10.1038/ng2089

SUPPORTING INFORMATION

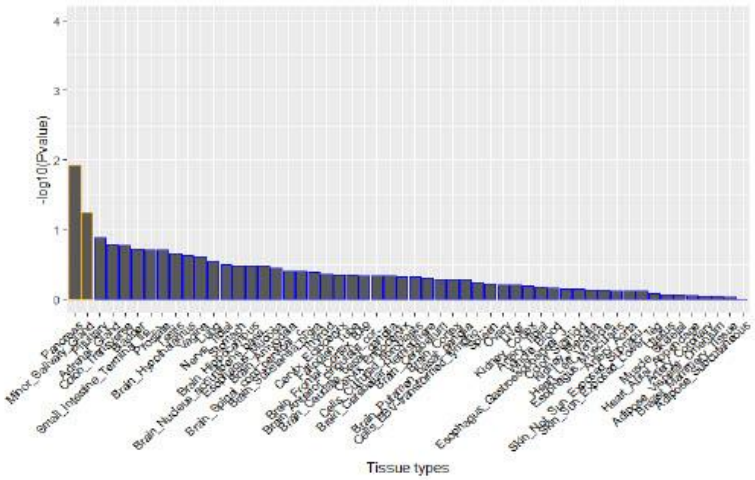
Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Adesoji, O. M., Schulz, H., May, P., Krause, R., Lerche, H., & Nothnagel, M. (2022). Benchmarking of univariate pleiotropy detection methods applied to epilepsy. *Human Mutation*, 1–19. https://doi.org/10.1002/humu.24417

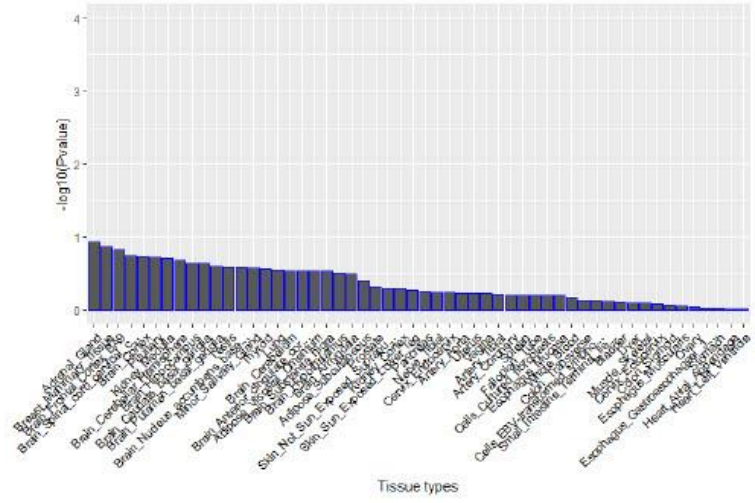
Supporting Figure S1. Principal components analysis (PCA) for the simulated population. Shown are the principal components (PCs) of individuals of the simulated population of European ancestry. **A:** PC1 *vs* PC2, **B:** PC1 *vs* PC4, **C:** PC1 *vs* PC8 and **D:** PC1 *vs* PC10. Panels A, B and C indicate sub-structures and clusters while the observed structure is not seen in panel D (randomness).



Supporting Figure S2. Functional annotation and enrichment of significant SNPs separately for both significant loci. The graph shows the enrichment of the genes expressed in 53 tissues of GTEx tissue expression data using MAGMA. **A**: Locus 2q24.3; **B:** Locus 17q21.32.

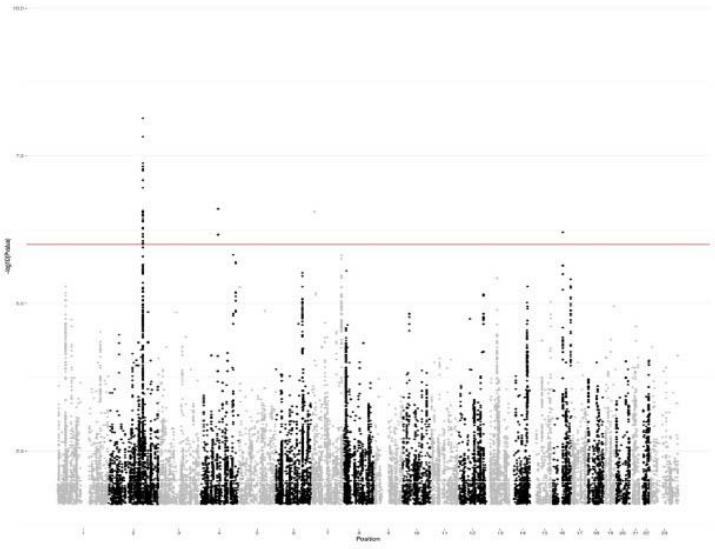


A



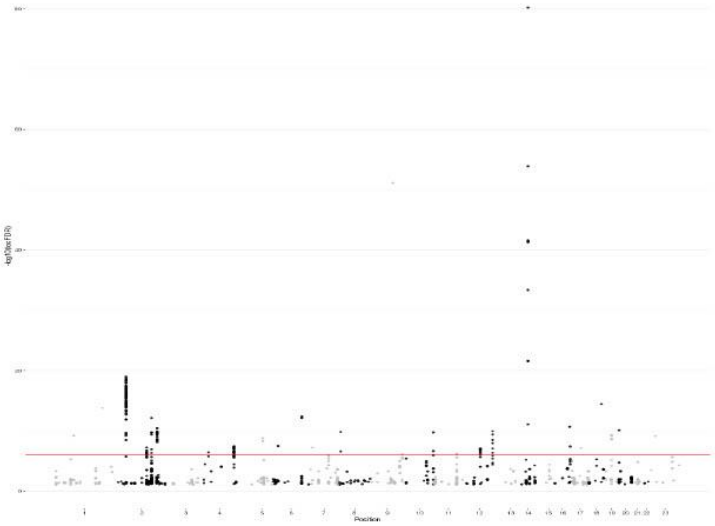
B

Supporting Figure S3. Manhattan plot of pleiotropy association testing between GGE and FE using cFDR. Chromosomal variant position is given on the x-axis, while -$\log_{10}$ transformed P-values are given on the y-axis. The red horizontal line denotes the genome-wide significance level.



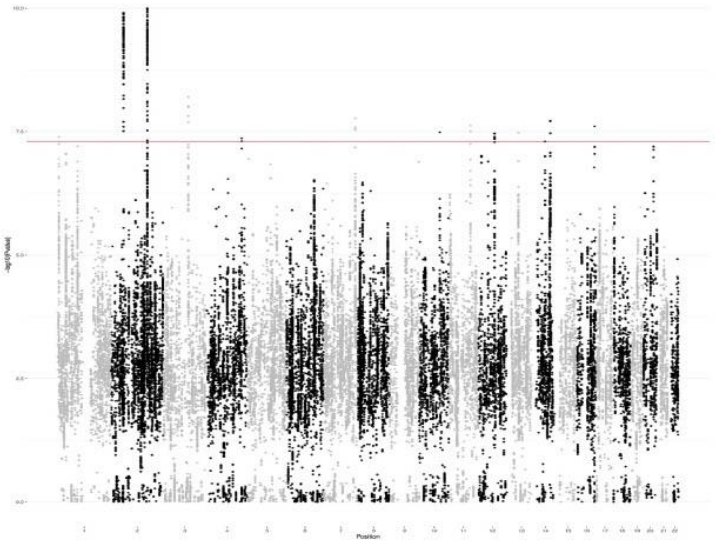
3

Supporting Figure S4. Manhattan plot of pleiotropy association testing between GGE and FE using CPBayes. Chromosomal variant position is given on the x-axis, while -$\log_{10}$ transformed P-values are given on the y-axis. The red horizontal line denotes the genome-wide significance level.

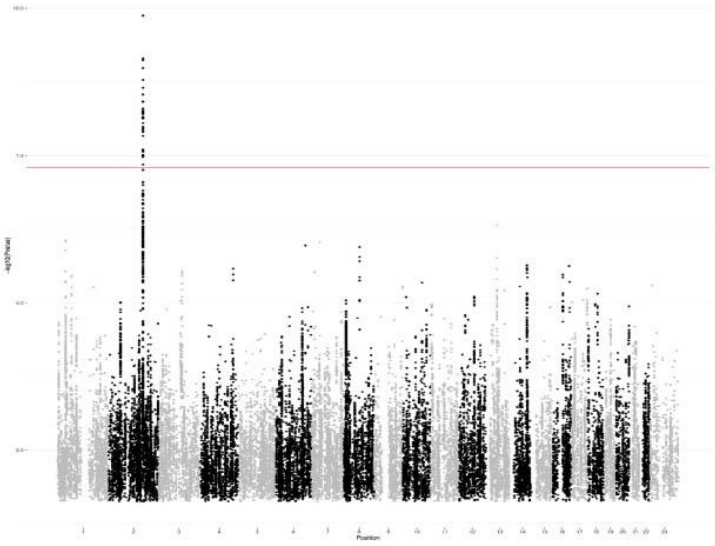


4

Supporting Figure S5. Manhattan plot of pleiotropy association testing between GGE and FE using classical meta-analysis. Chromosomal variant position is given on the x-axis, while -$\log_{10}$ transformed P-values are given on the y-axis. The red horizontal line denotes the genome-wide significance level.

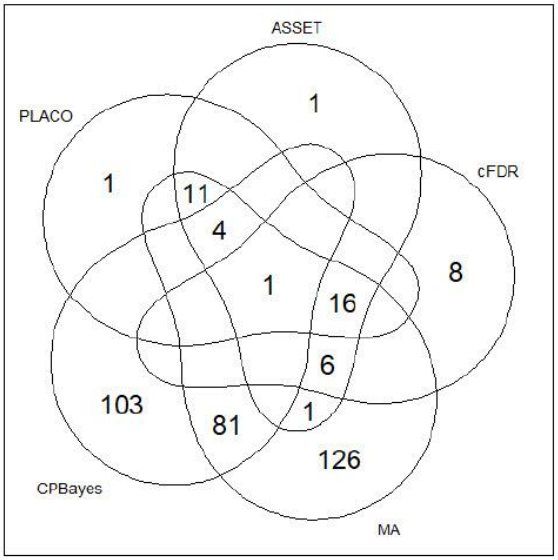


Supporting Figure S6. Manhattan plot of pleiotropy association testing between GGE and FE using PLACO. Chromosomal variant position is given on the x-axis, while -$\log_{10}$ transformed P-values are given on the y-axis. The red horizontal line denotes the genome-wide significance level.



6

Supporting Figure S7. Overlap between the five considered pleiotropy detection methods. The Venn diagram gives the numbers of SNPs that were found to be genome-wide significant for pleiotropy between the two epilepsy forms (GGE and FE) in the ILAE dataset by one or more methods. **MA:** meta-analysis.



7

Supporting Table S1: Power and false-positive rate (FPR) for five univariate pleiotropy detection methods assuming 5 causal SNPs. Given are the percentages for five methods (see Table 1) with respect to power and FPR for varying sample sizes and effect sizes as well as percentages of overlap between causal SNPs for two traits and two different values for disease prevalence using 100 replications while assuming 5 causal SNPs for each of both traits.

			Number of samples per phenotype											
			1,000 cases & 1,000 controls				5,000 cases & 5,000 controls				10,000 cases & 10,000 controls			
			Power		FPR		Power		FPR		Power		FPR	
Method	RR	Overlap	Prevalence		Prevalence		Prevalence		Prevalence		Prevalence		Prevalence	
			1%	10%	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%
CPBayes	1.00	20%	0	0	7.38	6.77	0	0	2.59	0.93	0	0	0	0
	1.20		96.67	83.33	12.2	12.36	100	100	4.73	6.91	100	100	4.09	4.2
	1.50		100	100	19.32	15.09	100	100	10.86	7.56	100	100	10.86	6.43
	2.00		100	100	15.73	19.58	100	100	15.07	8.58	100	100	27.53	8.78
	1.00	40%	0	0	7.81	10.08	0	0	1.19	2.96	0	0	0	0
	1.20		100	84.13	16.99	16.85	100	100	6.69	5.95	100	100	4.09	4.2
	1.50		100	100	17.71	20.78	100	100	16.92	10.71	100	100	10.86	6.43
	2.00		100	100	21.57	21.41	100	100	29.45	14.93	100	100	27.53	8.78
MA	1.00	20%	0	0	0	0	0	0	0	0	0	0	0	0
	1.20		56	3	0.53	0	100	100	91.91	14.08	100	100	100	88.85
	1.50		100	100	100	100	100	100	100	100	100	100	100	100
	2.00		100	100	100	100	100	100	100	100	100	100	100	100
	1.00	40%	0	0	0	0	0	0	0	0	0	0	0	0
	1.20		85	16	2.48	0.14	100	100	97.77	39.34	100	100	100	96.52
	1.50		100	100	100	99.88	100	100	100	100	100	100	100	100
	2.00		100	100	100	100	100	100	100	100	100	100	100	100

ASSET	1.00	20%	0	0	0	0	0	0	0	0	0	0	0	0
	1.20		78	40	0.84	0	100	100	1.61	1.45	100	100	1.72	1.72
	1.50		100	100	1.83	1.72	100	100	3.39	2.72	100	100	5.17	3.24
	2.00		100	100	2.39	1.83	100	100	4.28	3.39	100	100	6.43	4.37
	1.00	40%	0	0	0	0	0	0	0	0	0	0	0	0
	1.20		78	31.08	0.45	0	100	100	1.27	0.96	100	100	1.54	1.54
	1.50		100	100	2.11	1.96	100	100	6.44	3.64	100	100	9.57	5.12
	2.00		100	100	3.26	2.25	100	100	9.09	5.85	100	100	12.18	8.5
PLACO	1.00	20%	0	0	0	0	0	0	0	0	0	0	0	0
	1.20		41	3	0.55	0	100	100	2.75	1.64	100	100	6.95	3.32
	1.50		100	100	10.91	4.83	100	100	23.67	15.38	100	100	34.21	22.16
	2.00		100	100	11.06	7.08	100	100	31.88	23.05	100	100	56.39	28.98
	1.00	40%	0	0	0	0	0	0	0	0	0	0	0	0
	1.20		75.5	9.5	1.02	0	100	100	10.14	3.35	100	100	18.97	8.17
	1.50		100	100	24.47	13.13	100	100	46.61	32.47	100	100	65.68	49.5
	2.00		100	100	30.79	21.98	100	100	61.93	46.67	100	100	81.37	47.23
cFDR	1.00	20%	0	0	0	0	0	0	0	0	0	0	0.36	0
	1.20		50	15.52	2.43	0.65	100	100	6.88	6.51	100	100	7.3	6.55
	1.50		100	100	6.56	5.18	72	100	9.39	7.11	0	13	5.04	4.85
	2.00		100	100	7.71	7.87	0	98.99	4.58	7.93	0	0	6.43	3.8
	1.00	40%	0	0	0	0	0	0	0	0	0	0	0	0
	1.20		47.54	4.76	3.83	0.53	100	99.5	6.77	7.31	100	100	6.91	5.17
	1.50		100	100	8.13	7.47	76	100	13.35	9.46	0	19	9.15	7.44
	2.00		100	100	9.77	9.03	0	100	8.99	12.58	0	0	12.75	8.53

RR: relative risk; **Overlap:** proportion of causal SNPs being shared between the two phenotypes; **FPR:** false-positive rare (type I error).

Supporting Table S2: Power and false-positive rate (FPR) for five univariate pleiotropy detection methods assuming 10 causal SNPs. Given are the percentages for five methods (see Table 1) with respect to power and FPR for varying sample sizes and effect sizes as well as percentages of overlap between causal SNPs for two traits and two different values for disease prevalence using 100 replications while assuming 10 causal SNPs for each of both traits.

			Number of samples per phenotype											
			1,000 cases & 1,000 controls				5,000 cases & 5,000 controls				10,000 cases & 10,000 controls			
			Power		FPR		Power		FPR		Power		FPR	
Method	RR	Overlap	Prevalence		Prevalence		Prevalence		Prevalence		Prevalence		Prevalence	
			1%	10%	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%
CPBayes	1.00	20%	0	0	3,53	4.06	3.57	0	0.84	0.65	0	0	0.49	0.42
	1.20		100	99.12	14.03	13.35	100	100	5.34	5.28	100	100	3.48	2.83
	1.50		100	100	14.53	14.28	100	100	8.88	6.42	100	100	10.4	6.13
	2.00		100	100	14.83	14.13	100	100	10.57	6.94	100	100	16.65	6.53
	1.00	40%	0	0	4.23	3.86	0.76	0	1.18	0.65	2.5	0	0.38	0
	1.20		100	99.18	15.51	14.26	100	100	7.61	6.81	100	100	6.38	5.55
	1.50		100	100	17.7	17.41	100	100	16.75	9.79	100	100	41.56	10
	2.00		100	100	17.82	18.35	100	100	22.2	10.06	100	100	53.71	11.38
MA	1.00	20%	0	0	0	0	0	0	0	0	0	0	0	0
	1.20		96	42	3.08	0.45	100	100	100	85.08	100	100	100	88.85
	1.50		100	100	100	97.87	100	100	100	100	100	100	100	100
	2.00		100	100	100	99.46	100	100	100	100	100	100	100	100
	1.00	40%	0	0	0	0	0	0	0	0	0	0	0	0
	1.20		100	80	2.48	0.14	100	100	100	94.85	100	100	100	100
	1.50		100	100	100	99.88	100	100	100	100	100	100	100	100
	2.00		100	100	100	100	100	100	100	100	100	100	100	100

ASSET	1.00	20%	0	0	0	0	0	0	0	0	0	0	0	0
	1.20		97	67	0.9	0.51	100	100	1.19	1.07	100	100	1.31	1.09
	1.50		100	100	1.13	1.07	100	100	2.4	1.48	100	100	3.42	2
	2.00		100	100	1.19	1.07	100	100	2.74	1.75	100	100	3.53	2.42
	1.00	40%	0	0	0	0	0	0	0	0	0	0	0	0
	1.20		97.75	70.75	1.63	0.89	99.75	99.75	3.29	2.72	99.74	99.74	3.71	3.32
	1.50		99.75	99.75	3.59	2.94	99.75	99.75	5.89	4.29	99.74	99.74	8.45	5.9
	2.00		99.75	99.75	3.83	3.07	99.75	99.75	6.69	4.98	99.74	99.74	8.48	6.41
PLACO	1.00	20%	0	0	0	0	0	0	0	0	0	0	0	0
	1.20		95.5	25.5	0.55	0	100	100	3.24	1.7	100	100	4.67	2.8
	1.50		100	100	10.91	4.83	100	100	10.7	6.75	100	100	22.93	9.31
	2.00		100	100	11.06	7.08	100	100	14.84	7.49	100	100	29.74	14.64
	1.00	40%	0	0	0	0	0	0	0	0	0	0	0	0
	1.20		99.75	67.5	1.02	0	100	100	12.17	7.02	100	100	17.79	12.44
	1.50		100	100	24.47	13.13	100	100	37.76	20.94	100	100	65.93	28.11
	2.00		100	100	30.79	21.98	100	100	46.29	25.19	100	100	71.42	33.14
cFDR	1.00	20%	0	0	0	0	0	0	0	0	0	0.69	0	0.08
	1.20		94.17	44.74	5.08	0.65	100	100	6.05	6.26	100	100	6.36	5.99
	1.50		100	100	6.24	5.18	100	100	8.83	6.69	0	100	3.54	7.49
	2.00		100	100	6.55	7.87	99.5	100	11.33	7.3	0	31.5	4.42	5.99
	1.00	40%	0	0	0	0	0	0	0	0	0	0	0	0
	1.20		98.81	42.21	7.11	3.14	100	100	8.46	8	99.75	100	8.7	7.67
	1.50		100	100	8.3	7.66	99.75	99.74	18.74	9.31	0	99.75	8.87	11.11
	2.00		100	100	8.77	7.91	99.5	99.74	24.53	10.27	0	36.73	8.85	8.7

RR: relative risk; **Overlap:** proportion of causal SNPs being shared between the two phenotypes; **FPR:** false-positive rare (type I error).

The International League Against Epilepsy Consortium on Complex Epilepsies author names:

Members listed in alphabetical order:

Bassel Abou-Khalil[1], Pauls Auce[2, 3], Andreja Avbersek[4], Melanie Bahlo[5-7], David J Balding[8, 9], Thomas Bast[10, 11], Larry Baum[12-14], Albert J Becker[15], Felicitas Becker[16, 17] Bianca Berghuis[18], Samuel F Berkovic[19], Jonathan P Bradfield[20, 21], Lawrence C Brody[22], Russell J Buono[20, 23, 24], Ellen Campbell[25], Gregory D Cascino[26], Claudia B Catarino[4], Gianpiero L Cavalleri[27, 28], Stacey S Cherny[13, 29], Krishna Chinthapalli[4], Alison J Coffey[30], Alastair Compston[31], Antonietta Coppola[32, 33], Patrick Cossette[34], John J Craig[35], Gerrit-Jan de Haan[36], Peter De Jonghe[37, 38], Carolien G F de Kovel[39], Norman Delanty[27, 28, 40], Chantal Depondt[41], Orrin Devinsky[42], Dennis J Dlugos[43], Colin P Doherty[28, 44], Christian E Elger[45], Johan G Eriksson[46], Thomas N Ferraro[23, 47], Martha Feucht[48], Ben Francis[49], Andre Franke[50], Jacqueline A French[51], Verena Gaus[52], Eric B Geller[53], Christian Gieger[54, 55], Tracy Glauser[56], Simon Glynn[57], David B Goldstein[58, 59], Hongsheng Gui[13], Youling Guo[13], Kevin F Haas[1], Hakon Hakonarson[20, 60], Kerstin Hallmann[45, 61], Sheryl Haut[62], Erin L Heinzen[58, 59], Ingo Helbig[43, 63], Christian Hengsbach[16], Helle Hjalgrim[64, 65], Michele Iacomino[33], Andrés Ingason[66], Jennifer Jamnadas-Khoda[4, 67], Michael R Johnson[68], Reetta Kälviäinen[69, 70], Anne-Mari Kantanen[69], Dalia Kasperavičiūte[4], Dorothee Kasteleijn-Nolst Trenite[39], Heidi E Kirsch[71], Robert C Knowlton[72], Bobby P C Koeleman[39], Roland Krause[73], Martin Krenn[74], Wolfram S Kunz[45], Ruben Kuzniecky[75], Patrick Kwan[76, 77], Dennis Lal[78], Yu-Lung Lau[79], Holger Lerche[16], Costin Leu[4, 78, 80], Wolfgang Lieb[81], Dick Lindhout[36, 39], Iscia Lopes-Cendes[82], Daniel H Lowenstein[71], Alberto Malovini[83], Anthony G Marson[2], Thomas Mayer[84], Mark McCormack[27], James L Mills[85], Nasir Mirza[2], Martina Moerzinger[48], Rikke S Møller[64, 65], Anne M Molloy[86], Hiltrud Muhle[63], Ping-Wing Ng[87], Markus M Nöthen[88], Peter Nürnberg[89], Terence J O'Brien[76, 77], Karen L Oliver[19], Aarno Palotie[90, 91], Faith Pangilinan[22], Sarah Peter[73], Slavé Petrovski[76, 92], Annapurna Poduri[93], Michael Privitera[94], Rodney Radtke[95], Sarah Rau[16], Philipp S Reif[96, 97], Eva M Reinthaler[74], Felix Rosenow[96, 97], Josemir W Sander[4, 36, 98], Thomas Sander[52, 89], Theresa Scattergood[99], Steven C Schachter[100], Christoph J Schankin[101], Ingrid E Scheffer[19, 102], Bettina Schmitz[52], Susanne Schoch[15], Pak C Sham[13], Jerry J Shih[103], Graeme J Sills[104], Sanjay M Sisodiya[4, 98], Lisa Slattery[105], Alexander Smith[78], David F Smith[3], Michael C Smith[106], Philip E Smith[107], Anja C M Sonsma[39], Doug Speed[8, 108], Michael R Sperling[109], Bernhard J Steinhoff[10], Ulrich Stephani[63], Remi Stevelink[39], Konstantin Strauch[110, 111], Pasquale Striano[112], Hans Stroink[113], Rainer Surges[45], K Meng Tan[76], Liu Lin Thio[114], G Neil Thomas[115], Marian Todaro[76], Rossana Tozzi[116], Maria S Vari[112], Eileen P G Vining[117], Frank Visscher[118], Sarah von Spiczak[63], Nicole M Walley[58, 119], Yvonne G Weber[16, 120], Zhi Wei[121], Judith Weisenberg[114], Christopher D Whelan[27], Peter Widdess-Walsh[27, 28, 40], Markus Wolff[122], Stefan Wolking[120], Wanling Yang[79], Federico Zara[33], Fritz Zimprich[74]

1. Vanderbilt University Medical Center, Nashville, TN 37232, USA.
2. Department of Molecular and Clinical Pharmacology, University of Liverpool, Liverpool L69 3GL, UK.
3. The Walton Centre NHS Foundation Trust, Liverpool L9 7LJ, UK.
4. Department of Clinical and Experimental Epilepsy, UCL Institute of Neurology, Queen Square, London WC1N 3BG, UK.
5. Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Parkville 3052, Australia.
6. Department of Biology, University of Melbourne, Parkville 3010, Australia.
7. School of Mathematics and Statistics, University of Melbourne, Parkville 3010, Australia.

8. UCL Genetics Institute, University College London, London WC1E 6BT, UK.
9. Melbourne Integrative Genomics, University of Melbourne, Parkville 3052, Australia.
10. Epilepsy Center Kork, Kehl-Kork 77694, Germany.
11. Medical Faculty of the University of Freiburg, Freiburg 79085, Germany.
12. Centre for Genomic Sciences, The University of Hong Kong, Hong Kong.
13. Department of Psychiatry, The University of Hong Kong, Hong Kong.
14. The State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong, Hong Kong, China.
15. Section for Translational Epilepsy Research, Department of Neuropathology, University of Bonn Medical Center, Bonn 53105, Germany.
16. Department of Neurology and Epileptology, Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen 72076, Germany.
17. Department of Neurology, University of Ulm, Ulm 89081, Germany.
18. Stichting Epilepsie Instellingen Nederland (SEIN), Zwolle 8025 BV, The Netherlands.
19. Epilepsy Research Centre, University of Melbourne, Austin Health, Heidelberg 3084, Australia.
20. Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.
21. Quantinuum Research LLC, San Diego, CA 92101, USA.
22. National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA.
23. Department of Biomedical Sciences, Cooper Medical School of Rowan University Camden, NJ 08103, USA.
24. Department of Neurology, Thomas Jefferson University Hospital, Philadelphia, PA 19107, USA.
25. Belfast Health and Social Care Trust, Belfast BT9 7AB, UK.
26. Division of Epilepsy, Department of Neurology, Mayo Clinic, Rochester, MN 55902, USA.
27. Department of Molecular and Cellular Therapeutics, The Royal College of Surgeons in Ireland, Dublin 2, Ireland.
28. The FutureNeuro Research Centre, Dublin 2, Ireland.
29. Department of Epidemiology and Preventive Medicine, School of Public Health, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 6997801, Israel.
30. The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK.
31. Department of Clinical Neurosciences, Cambridge Biomedical Campus, Cambridge CB2 0SL, UK.
32. Department of Neuroscience, Reproductive and Odontostomatological Sciences, University Federico II, Naples 80138, Italy.
33. Laboratory of Neurogenetics and Neurosciences, Institute G. Gaslini, Genova 16148, Italy.
34. Department of Neurosciences, Université de Montréal, Montréal, CA 26758, Canada.
35. Department of Neurology, Royal Victoria Hospital, Belfast Health and Social Care Trust, Grosvenor Road, Belfast BT12 6BA, UK.
36. Stichting Epilepsie Instellingen Nederland (SEIN), Heemstede 2103 SW, The Netherlands.
37. Neurogenetics Group, Center for Molecular Neurology, VIB and Laboratory of Neurogenetics, Institute Born-Bunge, University of Antwerp, Antwerp 2610, Belgium.
38. Department of Neurology, Antwerp University Hospital, Edegem 2650, Belgium.
39. Department of Genetics, University Medical Center Utrecht, Utrecht 3584 CX, The Netherlands.
40. Division of Neurology, Beaumont Hospital, Dublin D09 FT51, Ireland.
41. Department of Neurology, Hôpital Erasme, Université Libre de Bruxelles, Bruxelles 1070, Belgium.
42. Comprehensive Epilepsy Center, New York University School of Medicine, New York, NY 10016, USA.
43. Department of Neurology, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.
44. Neurology Department, St. James's Hospital, Dublin D03 VX82, Ireland.

45. Department of Epileptology, University of Bonn Medical Centre, Bonn 53127, Germany.
46. Department of General Practice and Primary Health Care, University of Helsinki and Helsinki University Hospital, Helsinki 0014, Finland.
47. Department of Pharmacology and Psychiatry, University of Pennsylvania Perlman School of Medicine, Philadelphia, PA 19104, USA.
48. Department of Pediatrics and Neonatology, Medical University of Vienna, Vienna 1090, Austria.
49. Department of Biostatistics, University of Liverpool, Liverpool L69 3GL, UK.
50. Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, University Hospital Schleswig Holstein, Kiel 24105, Germany.
51. Department of Neurology, NYU School of Medicine, New York City, NY 10003, USA.
52. Department of Neurology, Charité Universitaetsmedizin Berlin, Campus Virchow-Clinic, Berlin 13353, Germany.
53. Institute of Neurology and Neurosurgery at St. Barnabas, Livingston, NJ 07039, USA.
54. Research Unit of Molecular Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg D-85764, Germany.
55. Institute of Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg D-85764, Germany.
56. Comprehensive Epilepsy Center, Division of Neurology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA.
57. Department of Neurology, University of Michigan, Ann Arbor, MI 48109, USA.
58. Center for Human Genome Variation, Duke University School of Medicine, Durham, NC 27710, USA.
59. Institute for Genomic Medicine, Columbia University Medical Center, New York, NY 10032, USA.
60. Division of Human Genetics, Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.
61. Life and Brain Center, University of Bonn Medical Center, Bonn 53127, Germany.
62. Montefiore Medical Center, Bronx, NY 10467, USA.
63. Department of Neuropediatrics, University Medical Center Schleswig-Holstein (UKSH), Kiel 24105, Germany.
64. Danish Epilepsy Centre, Dianalund 4293, Denmark.
65. Institute of Regional Health Services Research, University of Southern Denmark, Odense 5000, Denmark.
66. deCODE genetics, Reykjavik IS-101, Iceland.
67. Department of Psychiatry and Applied Psychology, Institute of Mental Health University of Nottingham, Nottingham NG7 2TU, UK.
68. Faculty of Medicine, Imperial College London, London SW7 2AZ, UK.
69. Kuopio Epilepsy Center, Neurocenter, Kuopio University Hospital, Kuopio 70029, Finland.
70. Institute of Clinical Medicine, University of Eastern Finland, Kuopio 70029, Finland.
71. Department of Neurology, University of California, San Francisco, CA 94143, USA.
72. University of Alabama Birmingham, Department of Neurology, Birmingham, AL 35233, USA.
73. Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette L-4362, Luxembourg.
74. Department of Neurology, Medical University of Vienna, Vienna 1090, Austria.
75. Department of Neurology, Zucker-Hofstra Northwell School of Medicine, NY 10075, USA.
76. Department of Medicine, University of Melbourne, Royal Melbourne Hospital, Parkville 3050, Australia.
77. Department of Neuroscience, Central Clinical School, Monash University, Melbourne 3004, Australia.

78. Stanley Center for Psychiatric Research, Broad Institute of Harvard and M.I.T., Cambridge, MA 02142, USA.
79. Department of Paediatrics and Adolescent Medicine, The University of Hong Kong, Hong Kong.
80. Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA.
81. Institut für Epidemiologie, Christian-Albrechts-Universität zu Kiel, Kiel 24105, Germany.
82. Department of Translational Medicine, School of Medical Sciences, University of Campinas (UNICAMP), and the Brazilian Institute of Neuroscience and Neurotecnology; Campinas, SP, Brazil.
83. Istituti Clinici Scientifici Maugeri, Pavia 27100, Italy.
84. Epilepsy Center Kleinwachau, Radeberg 01454, Germany.
85. Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892, USA.
86. School of Medicine, Trinity College Dublin, Dublin 2, Ireland.
87. United Christian Hospital, Hong Kong.
88. Institute of Human Genetics, University of Bonn Medical Center, Bonn 53127, Germany.
89. Cologne Center for Genomics, University of Cologne, Cologne 50931, Germany.
90. Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki 0014, Finland.
91. The Broad Institute of M.I.T. and Harvard, Cambridge, MA 02142, USA.
92. AstraZeneca Centre for Genomics Research, Precision Medicine and Genomics, IMED Biotech Unit, AstraZeneca, Cambridge CB2 0AA, UK.
93. Department of Neurology, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA.
94. Department of Neurology, Gardner Neuroscience Institute, University of Cincinnati Medical Center, Cincinnati, OH 45220, USA.
95. Department of Neurology, Duke University School of Medicine, Durham, NC 27710, USA.
96. Epilepsy-Center Hessen, Department of Neurology, University Medical Center Giessen and Marburg, Marburg, Germany and Philipps-University Marburg, Marburg 35043, Germany.
97. Epilepsy Center Frankfurt Rhine-Main, Center of Neurology and Neurosurgery, University Hospital Frankfurt and LOEWE Center for Personalized Translational Epilepsy Research (CePTER), Goethe University Frankfurt, Frankfurt 60528, Germany.
98. Chalfont Centre for Epilepsy, Chalfont-St-Peter, Buckinghamshire SL9 0RJ, UK.
99. Department of Endocrinology, Hospital of The University of Pennsylvania, Philadelphia, PA 19104, USA.
100. Departments of Neurology, Beth Israel Deaconess Medical Center, Massachusetts General Hospital, and Harvard Medical School, Boston, MA 02215, USA.
101. Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Bern 3010, Switzerland.
102. Department of Neurology, Royal Children's Hospital, Parkville 3052, Australia.
103. Department of Neurosciences, University of California, San Diego, La Jolla, CA 92037, USA.
104. School of Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK.
105. The Royal College of Surgeons in Ireland, Dublin D02 YN77, Ireland.
106. Rush University Medical Center, Chicago, IL 60612, USA.
107. Department of Neurology, Alan Richens Epilepsy Unit, University Hospital of Wales, Cardiff CF14 4XW, UK.
108. Aarhus Institute of Advanced Studies (AIAS), Aarhus University, 8000 Aarhus, Denmark.
109. Department of Neurology and Comprehensive Epilepsy Center, Thomas Jefferson University, Philadelphia, PA 19107, USA.
110. Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg D-85764, Germany.

q

111. Chair of Genetic Epidemiology, IBE, Faculty of Medicine, LMU Munich 80539, Germany.
112. Pediatric Neurology and Muscular Diseases Unit, Department of Neurosciences, Rehabilitation, Ophthalmology, Genetics, Maternal and Child Health, G. Gaslini Institute, University of Genoa, Genova 16148, Italy.
113. CWZ Hospital, 6532 SZ Nijmegen, The Netherlands.
114. Department of Neurology, Washington University School of Medicine, St. Louis, MO 63110, USA.
115. Institute for Applied Health Research, University of Birmingham, Birmingham B15 2TT, UK. .
116. C. Mondino National Neurological Institute, Pavia 27100, Italy.
117. Departments of Neurology and Pediatrics, The Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.
118. Department of Neurology, Admiraal De Ruyter Hospital, Goes 4462, The Netherlands.
119. Division of Medical Genetics, Department of Pediatrics, Duke University Medical Center, Durham, NC 27710, USA.
120. Department of Neurology and Epileptology, University of Aachen, Aachen 52074, Germany.
121. Department of Computer Science, New Jersey Institute of Technology, NJ 07102, USA.
122. Department of Pediatric Neurology, Vivantes Hospital Neukölln, 12351 Berlin, Germany.

International League Against Epilepsy Consortium on Complex Epilepsies (2022). Genome-wide meta-analysis of over 29,000 people with epilepsy reveals 26 loci and subtype-specific genetic architecture | medRxiv (Submitted). **Doi:** https://doi.org/10.1101/2022.06.08.22276120. The authors' list can be found at the end of the manuscript.

Genome-wide meta-analysis of over 29,000 people with epilepsy reveals 26 loci and subtype-specific genetic architecture

International League Against Epilepsy Consortium on Complex Epilepsies*

*Author names and contributions listed at the end

Corresponding authors

Samuel F Berkovic: s.berkovic@unimelb.edu.au (ORCID ID: 0000-0003-4580-841X)

Gianpiero L Cavalleri: gcavalleri@rcsi.ie (ORCID ID: 0000-0002-9802-0506)

Bobby PC Koeleman: B.P.C.Koeleman@umcutrecht.nl (ORCID ID: 0000-0001-7749-182X)

Abstract

Epilepsy is a highly heritable disorder affecting over 50 million people worldwide, of which about one-third are resistant to current treatments. Here, we report a trans-ethnic GWAS including 29,944 cases, stratified into three broad- and seven sub-types of epilepsy, and 52,538 controls. We identify 26 genome-wide significant loci, 19 of which are specific to genetic generalized epilepsy (GGE). We implicate 29 likely causal genes underlying these 26 loci. SNP-based heritability analyses show that common variants substantially close the missing heritability gap for GGE. Subtype analysis revealed markedly different genetic architectures between focal and generalized epilepsies. Gene-set analysis of GGE signals implicate synaptic processes in both excitatory and inhibitory neurons in the brain. Prioritized candidate genes overlap with monogenic epilepsy genes and with targets of current anti-seizure medications. Finally, we leverage our results to identify alternate drugs with predicted efficacy if repurposed for epilepsy treatment.

Introduction

The epilepsies are a heterogeneous group of neurological disorders, characterized by an enduring predisposition to generate unprovoked seizures.[1] It is estimated that over 50 million people worldwide have active epilepsy, with an annual cumulative incidence of 68 per 100,000 persons.[2]

Similar to other common neurodevelopmental disorders, the epilepsies have substantial genetic risk contributions from both common and rare genetic variation. Analysis of the epilepsies benefits from deep phenotyping which allows clinical subtypes to be distinguished[3], in contrast to other common neurodevelopmental disorders where phenotypic subtypes are more difficult to define. Differences in the genetic architecture of these clinical subtypes of epilepsies are also emerging to complement the clinical partitioning.[4–7] The rare but severe epileptic encephalopathies are usually non-familial and are largely caused by single *de novo* dominant variants, often involving genes encoding ion channels or proteins of the synaptic machinery.[8] Common and rare variations have both been shown to contribute to the milder and more common focal and generalized epilepsies. This is particularly true for generalized epilepsy, which is primarily constituted by genetic generalized epilepsy (GGE).[4,5,9,10] Nevertheless, previous genetic studies of common epilepsies have explained only a few percent of this common genetic variant, or SNP-based, heritability.[4–6,10]

Epilepsy is typically treated using anti-seizure medications (ASMs). However, despite the availability of over 25 licensed ASMs worldwide, a third of people with epilepsy experience continuing seizures.[11] Diet, surgery and neuromodulation represent additional treatment options that can be effective in small subgroups of patients.[12] Accurate classification of clinical presentations is an important guiding factor in epilepsy treatment.

Here, we report the third epilepsy GWAS meta-analysis, comprising a total of 29,944 deeply phenotyped cases recruited from tertiary referral centres, and 52,538 controls, approximately doubling the previous sample size.[4] Results suggest markedly different genetic architectures between focal and generalized forms of epilepsy. Combining these results with results from less stringently phenotyped biobank and deCODE genetics epilepsy cases did not substantially increase signal, despite almost doubling the sample size to 51,678 cases and 1,076,527 controls. Our findings shed light on the enigmatic biology of generalized epilepsy and the importance of accurate syndromic phenotyping, and may facilitate drug repurposing for novel therapeutic approaches.

Results

Study overview

We performed a genome-wide meta-analysis by combining the previously published effort from our consortium[4] with unpublished data from the Epi25 collaborative[10] and four additional cohorts (**Supplementary table 1**)**.** Our primary mixed model meta-analysis constitutes 4.9 million SNPs tested in 52,538 controls and 29,944 people with epilepsy, of which 16,384 people had neurologist classified focal epilepsy (FE) and 7,407 people had GGE. The epilepsy cases were primarily of European descent (92%), with a smaller proportion of African (3%) and Asian (5%) ancestry (**Supplementary table 2**). Cases were matched with controls of the same ancestry and GWAS were performed separately per ancestry, before performing trans-ethnic meta-analyses for the broad epilepsy phenotypes 'FE' (n=16,384 cases) and 'GGE' (n=7,407 cases). We further conducted meta-analyses in subjects of European ancestry of the well-defined GGE subtypes of: a) juvenile myoclonic epilepsy (JME), b) childhood absence epilepsy (CAE), c) juvenile absence epilepsy (JAE), and d) generalized tonic-clonic seizures alone (GTCSA), as well as the focal epilepsy subtypes of: a) focal epilepsy with hippocampal sclerosis, b) focal epilepsy with other lesions, and c) lesion-negative focal epilepsy. We ran a variety of follow-up analyses to identify potential sex-specific signals and obtain biological insights and

opportunities for drug-repurposing. Sample size prevented inclusion of other ethnicities in the subtype analyses.

GWAS for the epilepsies

Our 'all epilepsy' meta-analysis revealed four genome-wide significant loci, of which two were novel (**Figure 1**). Similar to our previous GWAS[4], the 2q24.3 locus was composed of two independently significant signals (**Supplementary table 3**). Furthermore, a novel suggestive signal (rs4932477, $p=5.04x10^{-8}$) was found on chromosome 15, containing *POLG*, which is associated with one of the most severe kinds of intractable monogenic epilepsy.[13] Using ASSET to determine the extent of FE and GGE-related pleiotropy, the 2q24.3 and 9q21.13 signals showed pleiotropic effects at a genome-wide significance level, with concordant SNP effect directions for both forms of epilepsy (**Supplementary table 4**). The 2p16.1 and 10q24.32 loci were primarily derived from GGE. The FE analysis did not reveal any genome-wide significant signals.

Analysis of GGE cases only uncovered a total of 25 independent genome-wide significant signals across 22 loci, of which 13 loci are novel. The strongest signal of association ($p=6.6x10^{-21}$), located at 2p16.1, constitutes three independently significant signals. Similarly, the novel locus 12q13.13 was composed of two independently significant signals (**Supplementary table 3**).

Functional annotation of the 2,355 genome-wide significant SNPs across the 22 GGE loci revealed that most variants were intergenic or intronic (**Supplementary data 1**). 26/2355 (1.1%) SNPs were exonic, of which 12 were located in protein-coding genes and nine were missense variants. Sixty-one percent of SNPs were located in open chromatin regions, as indicated by a minimum chromatin state of 1-7.[14] Further annotation by Combined Annotation-Dependent Depletion (CADD) scores predicted 110 associating SNPs to be deleterious (CADD score >12.37).[15] LDAK heritability analyses showed significant enrichment of signal in "super-enhancers" (**Supplementary table 5**), suggesting that GGE variants regulate clusters of transcriptional enhancers that control expression of genes that define cell identity.[16]

To assess potential syndrome-specific loci, we performed GWAS on seven well-defined FE and GGE subtypes **(Supplementary figure 1A-G)**. We found three genome-wide significant loci associated with JME (n=1,813), of which one was novel (8q23.1), and the other two (4p12 and 16p11.2) were reported in our previous GWAS.[4] All three signals appear specific to JME; without reaching nominal significance in any other GGE subtype. Furthermore, these loci did not reach genome-wide significance when these subtypes were pooled in the GGE analysis. Our analysis of CAE (n=1,072) consolidated an established genome-wide significant signal at 2p16.1, which was also observed in the GGE and all epilepsy GWAS. We did not find any genome-wide significant loci for JAE (n=671), GTCSA (n=499), 'non-lesional focal epilepsy' (n=6,367), 'focal epilepsy with hippocampal sclerosis' (n=1,375), or 'focal epilepsy with other lesions' (n=4,661).

Genomic inflation was comparable to our previous GWAS and all linkage-disequilibrium score regression (LDSR) intercepts were lower than in our previous GWAS (**Supplementary table 6**),[4] suggesting that the signals are primarily driven by polygenicity, rather than by confounding or population stratification.[17]

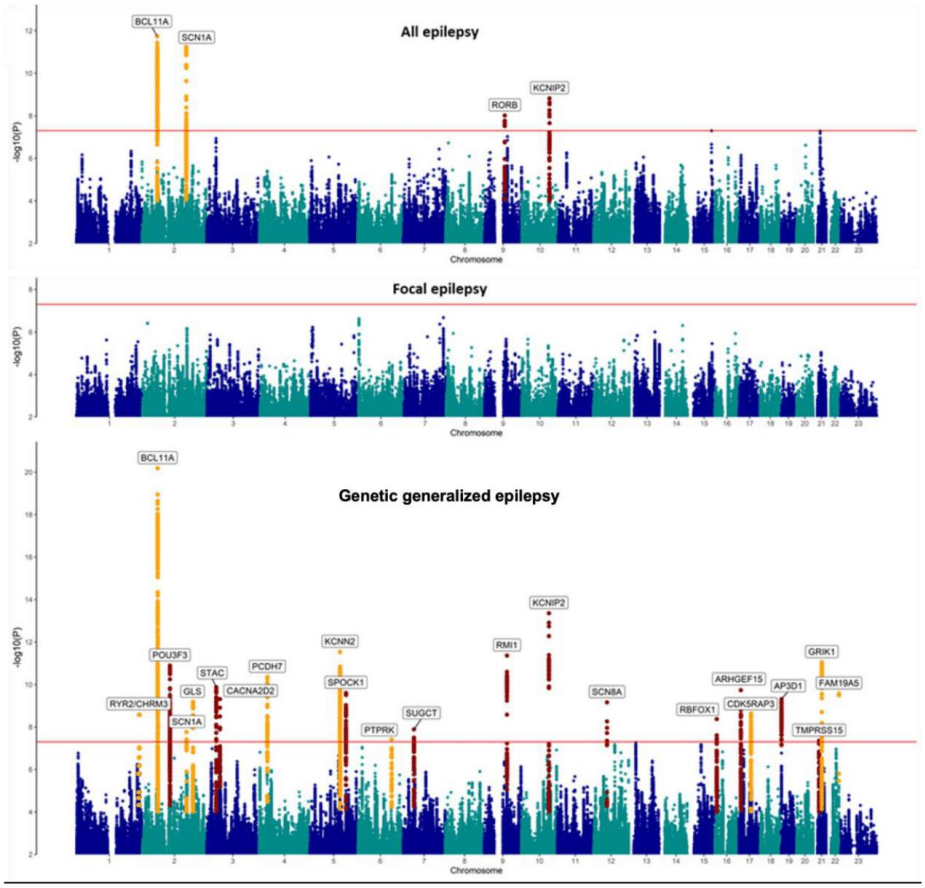


Figure 1. Manhattan plot of trans-ethnic all, focal epilepsy and genetic generalized epilepsy genome-wide meta-analyses. The red line shows the genome-wide significance threshold ($5x10^{-8}$). Chromosome and position are displayed on the x axis and -log10 P-value on the y axis. Novel genome-wide significant loci are highlighted in red and loci previously associated with epilepsy are labelled in orange. Annotated genes are those implicated by our gene prioritization analyses.

Locus annotation, transcriptome-wide association study (TWAS) and gene prioritization

Using FUMA[18] (see Methods), the 'all epilepsy' meta-analysis was mapped to 43 genes and the GGE analysis to 278 genes (**Supplementary data 2**). Thirty nine of the 43 'all epilepsy' genes overlapped with GGE, resulting in a total of 282 uniquely mapped genes. These 282 genes were enriched for monogenic epilepsy genes (hypergeometric test, 18/837 genes overlapped; odds ratio [OR]=1.51, P=0.04), and targets of ASMs (hypergeometric test, 9/191 genes overlap; OR=3.39, $P=5.4x10^{-4}$).

We calculated a gene-based association score based on the aggregate of all SNPs inside each gene using MAGMA (see Methods).[19] This analysis yielded 39 significant genic associations, six with 'all epilepsy', and 37 with GGE (four overlapped with the 'all epilepsy' analysis), after correction for 16,371 tested genes ($p<0.05/16,371$ genes; **Supplementary data 3**). Thirteen of these 39 genes mapped to regions outside of the genome-wide significant loci from the single SNP analyses.

Next, we performed a transcriptome-wide association study (TWAS) to assess whether epilepsy was associated with differential gene expression in the brain (see Methods).[20,21] These analyses revealed significant associations of 27 genes total; 13 genes with 'all epilepsy', 16 with GGE and two with both phenotypes (**Supplementary data 4**). Nineteen of the 27 genes mapped outside of the 26 loci identified through the GWAS. Using Summary-data-based Mendelian Randomization (SMR)[22], we determined a potentially causal relationship between brain expression of *RMI1* and 'all epilepsy', and between *RMI1*, *CDK5RAP3*, *TVP23B* and GGE (**Supplementary data 5**).

Of note, expression of *RMI1* was associated with GGE in both TWAS ($p=4.0x10^{-10}$) and SMR ($p=5.2x10^{-8}$), as well as with 'all epilepsy' (TWAS $p=1.3x10^{-6}$; SMR $p=2.6x10^{-6}$). *RMI1* has a crucial role in genomic stability[23] and has not been previously associated with epilepsy nor any other Mendelian trait (OMIM #610404).

We used a combination of ten different criteria to identify the most likely implicated gene within each of the 26 associated loci from the meta-analysis (see Methods). This resulted in a shortlist of 29 genes (**Figure 2**), of which ten are monogenic epilepsy genes, seven are known targets of currently licensed ASDs and 17 are associated with epilepsy for the first time. Interrogation of the Drug Gene Interaction Database (DGIdb) showed that 13 of the 29 genes are targeted by a total of 214 currently licensed drugs (**Supplementary data 6**).

The strongest association signal for GGE was found at 2p16.1, consistent with our previous results where we implicated the gene *VRK2* or *FANCL*.[24] Our gene prioritization analysis now points to the transcription factor *BCL11A* as the culprit gene, located 2.5MB upstream of the lead SNPs at this locus. Two of three lead SNPs are located in enhancer regions (as assessed by chromatin states in brain tissue) which are linked to the *BCL11A* promoter via 3D chromatin interactions (**Supplementary figure 2**). Rare variants in *BCL11A* were recently associated with intellectual disability and epileptic encephalopathy.[25] However, interrogation of the MetaBrain eQTL database did not reveal a significant association of our lead SNPs with *BCL11A* expression.

Phenotype	Locus	Novel / Replication	Lead SNP (A1:A2)	Freq1	Z-score	P-value	Gene	Total	Missense	TWAS	SMR	MAGMA	PoPS	Brain exp	brain-coX	KO mouse	AED target	Monogenic
All epilepsy	2p16.1	Replication	rs13032423 (A:G)	0.53	-7.04	1.85E-12	*BCL11A*	5										
	2q24.3	Replication	rs59237858 (T:C)	0.23	-6.89	5.746E-12	*SCN1A*	8										
	9q21.13	Novel	rs4744696 (A:G)	0.82	-5.74	9.694E-09	*RORB*	4										
	10q24.32	Novel	rs3740422 (C:G)	0.33	6.04	1.517E-09	*KCNIP2*	3										
GGE	1q43	Replication	rs876793 (T:C)	0.67	-5.95	2.644E-09	*RYR2*	4										
							CHRM3	4										
	2p16.1	Replication	rs11688767 (A:T)	0.53	9.38	6.58E-21	*BCL11A*	5										
	2q12.1	Novel	rs62151809 (T:C)	0.43	6.77	1.277E-11	*POU3F3*	3										
	2q24.3	Replication	rs11890028 (T:G)	0.72	5.63	1.728E-08	*SCN1A*	8										
	2q32.2	Replication	rs6721964 (A:G)	0.66	-6.18	6.542E-10	*GLS*	4										
	3p22.3	Novel	rs9861238 (A:G)	0.41	-6.42	1.333E-10	*STAC*	2										
	3p21.31	Novel	rs739431 (A:G)	0.84	6.23	4.822E-10	*CACNA2D2*	6										
	4p15.1	Replication	rs1463849 (A:G)	0.59	-6.59	4.377E-11	*PCDH7*	3										
	5q22.3	Replication	rs4596374 (T:C)	0.55	-6.98	2.906E-12	*KCNN2*	6										
	5q31.2	Novel	rs2905552 (C:G)	0.48	-6.33	2.492E-10	*SPOCK1*	5										
	6q22.33	Replication	rs13219424 (T:C)	0.29	-5.49	3.872E-08	*PTPRK*	3										
	7p14.1	Novel	rs37276 (T:G)	0.26	-5.69	1.288E-08	*SUGCT*	2										
	9q21.32	Novel	rs2780103 (T:C)	0.26	-6.93	4.342E-12	*RMI1*	5										
	10q24.32	Novel	rs11191156 (A:G)	0.67	-7.55	4.409E-14	*KCNIP2*	4										
	12q13.13	Novel	rs4762030 (T:G)	0.02	6.17	6.90E-10	*SCN8A*	6										
	16p13.3	Novel	rs62014006 (T:G)	0.47	5.88	4.223E-09	*RBFOX1*	5										
	17p13.1	Novel	rs2585398 (A:C)	0.53	-6.37	1.842E-10	*ARHGEF15*	6										
	17q21.32	Replication	rs16955463 (T:G)	0.25	-5.97	2.3E-09	*CDK5RAP3*	4										
	19p13.3	Novel	rs75483641 (T:C)	0.14	-6.22	4.852E-10	*AP3D1*	5										
	21q21.1	Novel	rs1487946 (A:G)	0.59	5.47	4.409E-08	*TMPRSS15*	1										
	21q22.1	Replication	rs7277479 (A:G)	0.36	-6.82	8.935E-12	*GRIK1*	4										
	22q13.32	Novel	rs469999 (A:G)	0.31	-6.32	2.647E-10	*FAM19A5*	2										
CAE	2p16.1	Replication	rs12185644 (A:C)	0.70	-7.12	1.04E-12	*BCL11A*	5										
JME	4p12	Replication	rs17537141 (T:C)	0.85	-5.47	4.62E-08	*GABRA2*	6										
	8q23.1	Novel	rs3019359 (T:C)	0.41	-5.55	2.89E-08	*RSPO2*	3										
							TMEM74	3										
	16p11.2	Replication	rs1046276 (T:C)	0.35	6.19	6.05E-10	*STX1B*	5										
							CACNA1I	5										

Figure 2. Genome-wide significant loci and prioritized genes. Genome-wide significant loci are annotated with details from the lead-SNP and prioritized genes. Loci were classified as novel or replication according to the genome-wide significant results of previous GWAS publications. Genes were scored based on 10 criteria/methods, after which the gene with the highest score in the locus was selected as the prioritized gene. Total: number of satisfied criteria for gene prioritization. Missense: the locus contains a missense variant in the gene. TWAS: significant transcriptome-wide association with the gene. SMR: significant summary-based mendelian randomisation association with the gene. MAGMA: significant genome-wide gene based association. PoPS: gene prioritized by polygenic priority score. Brain exp: the gene is preferentially expressed in brain tissue. Brain-coX: the gene is prioritized as co-expressed with established epilepsy genes. KO mouse: knockout of the gene causes a neurological phenotype in mouse models. Monogenic: the gene is a known cause of monogenic epilepsy. Genomic coordinates for each locus (hg19) can be found in Supplementary table 3.

The HLA system and common epilepsies

The highly polymorphic HLA region has been associated with various neuropsychiatric and autoimmune neurological disorders following accurate capture of all its genetic variation. Therefore, we imputed HLA alleles and amino acid residues using CookHLA[26] and ran association across epilepsy, focal and GGE phenotypes, as well as the seven sub-phenotypes (see Methods). No SNP, amino acid residue or HLA allele reached the level of genome-wide significance (see **Supplementary figure 3**). The most significant signal was with GGE, in which an aspartame amino acid residue in exon 2 position 31432494 had a p-value of 3.8×10^{-7}.

SNP-based heritability

We calculated SNP-based heritability using LDAK to determine the proportion of epilepsy risk attributable to common genetic variants. We observed liability scale SNP-based heritabilities of 17.7% (95% CI 15.5 - 19.9%) for all epilepsy, 16.0% (14.0 - 18.0%) for FE and 39.6% (34.3 - 44.6%) for GGE. Heritabilities for GGE subtypes were notably higher for all individual GGE subtypes: ranging from 49.6% (14.0% - 85.3%) for GTCSA to 90.0% (63.3 - 116.6%) for JAE (**Supplementary table 7**).

Employing a univariate causal mixture model[27] (see Methods) we estimated that 2,850 causal SNPs (standard error: 200) underlie 90% of the SNP-based heritability of GGE, comparable with previous estimates.[9] Power analysis demonstrated that the current genome-wide significant SNPs only explain 1.5% of the phenotypic variance, whereas an estimated sample size of around 2.5 million subjects would be necessary to identify the causal SNPs that explain 90% of GGE SNP-based heritability (**Supplementary figure 4)**.

To further explore the heritability of the different epilepsy phenotypes, we used LDSC to perform genetic correlation analyses.[28] We found evidence for strong genetic correlation between all four GGE syndromes (**Supplementary figure 5**). We also observed a significant genetic correlation between the focal non-lesional and JME syndromes, which has been reported previously.[4] Here, with larger sample sizes, CAE also showed a significant genetic correlation with the focal non-lesional cohort.

Tissue and cell-type enrichment

To further illuminate the underlying biological causes of the epilepsies, we used MAGMA[19] and data from the gene-tissue expression consortium (GTEx) to assess whether our GGE-associated genes were enriched for expression in specific tissues and cell types (see Methods). We identified significant enrichment of associated genes expressed in brain and pituitary tissue (**Supplementary figure 6**). This is the first time the pituitary gland has been implicated in GGE and might reflect a hormonal component to seizure susceptibility. Further sub-analyses showed that our results were enriched for genes expressed in almost all brain regions, including subcortical structures such as the hypothalamus, hippocampus and amygdala (**Supplementary figure 7**). We did not find enrichment for genes expressed at specific developmental stages in the brain (**Supplementary figure 8**).

Cell-type specificity analyses of GGE data using various single-cell RNA-sequencing reference datasets (see Methods) revealed enrichment in excitatory as well as inhibitory neurons, but not in other brain cells like astrocytes, oligodendrocytes or microglia (**Supplementary figure 9**). Similarly, stratified LD-score regression using single-cell expression data (see Methods) did not reveal a difference between excitatory and inhibitory neurons (p=0.18).

Gene-set analyses

MAGMA gene-set analyses showed significant associations between GGE and biological processes involving various functions in the synapse (**Supplementary data 7**). To further refine the synaptic signal, we performed a gene-set analysis using lists of expert-curated gene-sets involving 18 different synaptic functions.[29] These analyses showed that GGE was associated with intracellular signal transduction (n=139 genes, $p=9.6x10^{-5}$) and excitability in the synapse (n= 54 genes, p=0.0074). None of the other 16 synaptic functions showed any association (**Supplementary data 7**). Genes involved with excitability include the N-type calcium channel gene *CACNA2D2,* implicated at the novel GGE locus 3p21.31. N-type calcium channel blockers such as levetiracetam and lamotrigine are amongst the most widely used and effective ASMs for GGE as well as focal epilepsy.[30–32] Together, these results suggest that the genes associated with GGE are expressed in excitatory as well as inhibitory neurons in various brain regions, where they affect excitability and intracellular signal transduction at the synapse.

Sex-specific analyses

There are known sex-related patterns in the epidemiology of epilepsy. Although females have a marginally lower incidence of epilepsy than males, GGE is known to occur more frequently in females.[33] To test whether this sex divergence has a genetic basis, we performed sex-specific GWAS for all, GGE and FE (**Supplementary figures 10-12**). Analyses revealed one female-specific genome-wide significant signal at 10q24.32 (lead SNP: rs72845653), containing *KCNIP2*, implicated in our main GGE meta-analysis (lead SNP: rs11191156). However, the lead SNPs of these two signals are not in LD (r^2=0.05). Interestingly, the direction of effect of this signal is opposite in females and males. This sex difference is further corroborated by significant sex-heterogeneity ($p=1.54x10^{-8}$) and gender-differentiation ($p=5.6x10^{-9}$).[34] Sex-related differences in transcription levels in human heart have previously been reported for *KCNIP2*.[35] We did not find any sex-divergent signals for 'all' or focal epilepsy.

LDSC was used to assess the genetic correlation between male-only and female-only GWAS. The male and female GWAS of all epilepsy, FE and GGE were strongly genetically correlated (all Rg>0.9) and none of these correlations were significantly different from 1 (all p>0.05). These results suggest that, with the exception of the female-specific 10q24.32 signal, the overall genetic basis of common epilepsy appears largely similar between males and females.

Genetic overlap between epilepsy and other phenotypes

To explore the genetic overlap of epilepsy with other diseases, we first cross-referenced the 26 genome wide epilepsy loci with other traits with significant associations ($p<5x10^{-8}$) for the same SNP, or SNPs in strong linkage disequilibrium with our lead SNPs (as detailed in **figure 2**). This analysis revealed eighteen likely pleiotropic loci, with previous associations reported across a variety of traits, the most common being cognitive, sleep, psychiatric, coronary and blood cell traits (**Supplementary figure 13**). The remaining eight loci appear to be specific to epilepsy (3p22.3, 4p12, 5q31.2, 7p14.1, 8q23.1, 9q21.13, 21q21.1, 21q22.1).

We then performed genetic correlation analyses between 18 selected traits and all, GGE and focal epilepsy using LDSC[17]. The selected traits had either, or a combination of 1) epilepsy as a common comorbidity or 2) pleiotropic loci shared with epilepsy. Significant correlations (P<0.05/54=0.0009) were found with febrile seizures, stroke, headache, ADHD, type 2 diabetes and intelligence (**Figure 3**).

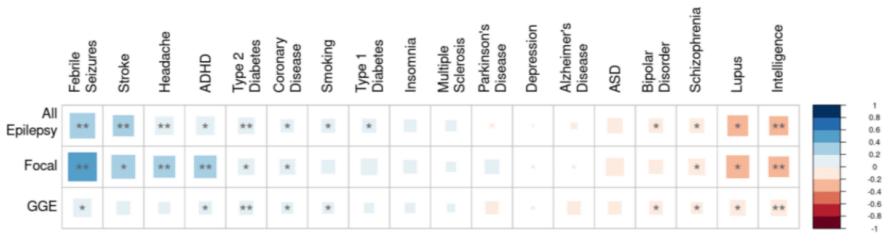


*Figure 3. Genetic correlations of epilepsy with other phenotypes. The genetic correlation coefficient was calculated with LDSC and is denoted by color scale from -1 (red; negatively (anti-)correlated) to +1 (blue; positively correlated). The square size relates to the absolute value of the corresponding correlation coefficient. * P < 0.05, ** P < 0.0009 (Bonferroni corrected).*

Genetic correlation analyses assess the aggregate of shared genetic variants associated with two phenotypes. However, genetic correlations can become close to zero when there is consistent mixed

directionality of SNP effects between two phenotypes.[36] Autism spectrum disorder (ASD) was not significantly correlated, despite monogenic pleiotropy with epilepsy genes supporting an overlap. To explore whether inverse directionality could explain the lack of genetic correlation between ASD and epilepsy we applied the MiXeR tool to GGE, intelligence and ASD, to quantify polygenic overlap irrespective of genetic correlation (see Methods). Results showed that >99% of causal SNPs underlying GGE are shared with intelligence, of which 58% have a discordant direction of effect (**Supplementary figure 14**). Furthermore, despite a lack of genetic correlation with ASD (Rg=-0.12, p=0.06, all epilepsy; Rg=-0.17, p=0.06 focal epilepsy; Rg=-0.09, p=0.09, GGE), we found that 95% of causal SNPs underlying GGE are shared with ASD, but 59% have a discordant direction of effect. This is consistent with the finding that epilepsy and ASD can have a shared genetic cause.[37,38] For example, monogenic ASD and epilepsy can occur as the result of pathogenic variants in *SCN2A*. Functional studies have shown that ASD without seizures can be caused by loss-of-function variants in *SCN2A* [39], whereas epilepsy can be caused by gain-of-function variants in *SCN2A*.[40,41] Indeed, ASD variants in *SCN2A* seem protective against neuronal hyperexcitability.[41]

Leveraging GWAS for drug repurposing

To test the potential of our meta-analysis to inform drug repurposing, we predicted the relative efficacy of drugs for epilepsy (see Methods). This analysis was based upon the predicted ability of each drug to modulate epilepsy-related changes in the function and abundance of proteins, as inferred from the GWAS summary statistics (see Methods).[42] We validated the drug predictions by determining if they are concordant with findings from clinical experience and trials. In our predictions for all epilepsy, current ASMs were ranked higher than expected by chance ($p < 1\times10^{-6}$), and higher than drugs used to treat any other human disease. For GGE, broad-spectrum ASMs were predicted to be more effective than narrow-spectrum antiseizure drugs ($p < 1\times10^{-6}$), consistent with clinical experience.[43] Furthermore, the predicted order of efficacy for GGE of individual ASMs matched their observed order in the largest head-to-head randomized controlled clinical trials for generalized epilepsy,[32,44] an observation unlikely to occur by chance ($p < 1\times10^{-6}$).

Using this approach, we highlight the top 20 drugs that are licensed for conditions other than epilepsy, but are predicted to be efficacious for generalized epilepsy, and additionally have published evidence of antiseizure efficacy from multiple published studies and multiple animal models (**Supplementary table 8**). The full list of all predictions can be found in **Supplementary data 8**.

GWAS in epilepsies ascertained from population biobanks and from deCODE genetics

We performed GWAS using data from several large-scale population biobanks and from deCODE genetics (total cases n=21,734, total controls n=1,023,989, phenotyped using ICD codes, see Methods). Although the biobank and deCODE genetics-specific GWAS did not identify any genome-wide significant loci for GGE or 'all epilepsy', one significant locus at 2q22.1 (nearest gene, *NXPH2*) emerged for focal epilepsy (**Supplementary figure 15**).

Meta-analysis of the biobank and deCODE genetics summary statistics with those from the primary epilepsy GWAS identified seven significant loci for the 'all epilepsy' phenotype. Six of these signals were previously identified in the primary 'all epilepsy' (n=4) or the 'GGE' GWAS (n=2). One locus (2q12.1) was novel. The combined biobank and deCODE genetics meta-analysis for GGE identified five novel loci, but four loci from our primary GWAS fell below significance (**Supplementary figure 16**). The combined focal epilepsy meta-analysis showed no significant associations. LDSC between the biobank/deCODE genetics and the primary GWAS results showed genetic correlations ranging between 0.31 and 0.74 (**Supplementary table 9**).

Discussion

In this study, we leveraged a substantial increase in sample size to uncover 26 common epilepsy risk loci, of which 16 have not been reported previously. Using a combination of ten post GWAS analysis methods, we pinpointed 29 genes that most likely underlie these signals of association. These signals showed enrichment throughout the brain and indicate an important role for synapse biology in excitatory as well as inhibitory neurons. Drug prioritization from the genetic data highlighted licensed ASMs, ranked the ASMs broadly in line with clinical experience and pointed to drugs for potential repurposing. These findings further our understanding of the pathophysiology of common epilepsies and provide new leads for therapeutics.

The 26 associated loci included some notable monogenic epilepsy genes. These include the calcium channel gene *CACNA2D2*, an established epileptic encephalopathy gene[45] that is directly targeted by ten currently licenced drugs, including two ASMs (gabapentin and pregabalin) as well as the Parkinson's disease drug safinamide and the nonsteroidal anti-inflammatory drug celecoxib. Both safinamide and celecoxib have evidence of anti-seizure activity.[46,47] *SCN8A*, which encodes a voltage-gated sodium channel, is an established epileptic encephalopathy gene and is associated with common epilepsies for the first time here. $Na_v1.6$ (encoded by *SCN8A)* is targeted by commonly used sodium channel blocking drugs that have been found to be the most efficacious ASMs for people with monogenic *SCN8A*-related epilepsies that are often due to channel gain-of-function.[48] Additional drugs targeting $Na_v1.6$ include safinamide and quinidine. *RYR2* encodes a ryanodine receptor, is an established cardiac disorder gene, has recently been implicated in epilepsy [49,50] and is targeted by caffeine as well simvastatin, atorvastatin and carvedilol. The acetylcholine receptor gene *CHRM3* has been previously associated with epilepsy[51] and is targeted by drugs including solifenacin, used to treat urinary incontinence.

We found that GGE in particular has a strong contribution from common genetic variation. When analyzing individual GGE syndromes, we found that up to 90% of liability is attributable to common variants in the JAE subtype, making it amongst the highest of over 700 traits reported in a large GWAS atlas[52] (albeit with relatively large confidence intervals; **Supplementary table 7**). The heritability estimates decrease to 40% for the collective GGE phenotype, possibly due to increased heterogeneity from combining syndromes with pleiotropic as well as syndrome-specific risk loci. Although statistical power drastically decreased when assessing specific GGE syndromes, three loci appeared specific to JME. These findings highlight the unique genetic architecture of the subtypes of common epilepsies, which are characterized by a high degree of both shared, and syndrome-specific, genetic risk.

In contrast to GGE, for focal epilepsies we found only a minor contribution of common variants, with no variant reaching genome-wide significance. It would seem that focal epilepsies, as a group, are far more heterogeneous than GGE. Our attempt to mitigate this heterogeneity by performing subtype analysis contrasted with the results from GGE, suggesting different genetic architectures, consistent with the experience from studies of common[9] and rare[5] genetic variation and PRS.[6] There is also emerging evidence for a significant role of non-inherited, somatic mutations in focal epilepsies.[53]

This work highlights the challenges of working with epilepsy cohorts ascertained through large biobanking initiatives. Accurate classification of epilepsy requires a combination of clinical features, electrophysiology and neuroimaging. These details were not available from the biobanks we worked with. Rather, phenotypes were generally limited to ICD codes, which are prone to misclassification.[54] Population biobanks are also probably ascertaining milder epilepsies that are responsive to treatment, contrasting with the enrichment for refractory epilepsies at tertiary referral centres.

Moreover, a proportion of adults with epilepsy have an acquired brain lesion, such as stroke, tumors or head trauma. Biobanks typically provide self-reported clinical information and codes from primary care and inpatient hospital care episodes, but not neurological specialist outpatient records that would indicate whether previous brain insults were considered relevant to the epilepsy. As a result, the inclusion of the biobank data appeared to introduce more heterogeneity. This contrasts with genetic mapping of other polygenic diseases like type 2 diabetes and migraine, which are relatively easy and reliable to diagnose and classify, resulting in a great increase in GWAS loci when including data from the same biobanks as included in our study.[55,56]

We found enrichment of GGE variants in brain-expressed genes, involving excitatory and inhibitory neurons, but not any other brain cell type. This contrasts with other neurological diseases. For example, microglia are involved in Alzheimer's disease[57] and multiple sclerosis,[58] whereas migraine does not appear to have brain cell specificity.[56] We further refine this signal by showing an involvement of synapse biology, primarily intracellular signal transduction and synapse excitability. These findings suggest an important role of synaptic processes in excitatory and inhibitory neurons throughout the brain, which could be a potential therapeutic target. Indeed, synaptic vesicle transport is a known target of the ASMs levetiracetam and brivaracetam.[59]

We confirmed that our GWAS-identified genes had significant overlap with monogenic epilepsy genes. A similar convergence of common and rare variant associations has been observed for other neurological neuropsychiatric conditions including schizophrenia[60] and ALS[61]. The genes prioritized in our GWAS signals also overlapped with known targets of current ASMs[4] and we have provided a list of other drugs that directly target these genes. Moreover, using a systems-based approach[42] we highlight drugs that are predicted to be efficacious when repurposed for epilepsy, based on their ability to perturb function and abundance in gene expression. Insights from GWAS of epilepsy have the potential to accelerate the development of new treatments via the identification of promising drug repurposing candidates for clinical trials.[62] We anticipate that follow-up studies of the highlighted drugs in this study could show clinical efficacy in epilepsy treatment.

In summary, these new data reveal markedly different genetic architectures between the milder and more common focal and generalized epilepsies, provide novel biological insights to disease aetiology and highlight drugs with predicted efficacy when repurposed for epilepsy treatment.

Methods

Ethics statement

Local institutional review boards approved study protocols at each contributing site. 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.

Sample and phenotype descriptions

This meta-analysis combines previously published datasets with novel genotyped cohorts. Descriptions of the 24 cohorts included in our previous analysis can be found in the Supplementary table 6 of that publication.[4] Here we included 5 novel cohorts (**Supplementary table 1**), comprising 14,732 epilepsy cases and 22,362 controls, resulting in a total sample size of 29,944 cases and 52,538 controls. Classification of epilepsy was performed as described previously.[4] In brief, we assigned people with epilepsy into focal epilepsy, genetic generalized epilepsy (GGE) or unclassified epilepsy. 'All epilepsy' was the combination of GGE, focal and unclassified epilepsy. Where possible, we used EEG, MRI and clinical history to further refine the subphenotypes: juvenile myoclonic epilepsy (JME),

childhood absence epilepsy (CAE), juvenile absence epilepsy (JAE), generalized tonic-clonic seizures alone (GTCSA), non-lesional focal epilepsy, focal epilepsy with hippocampal sclerosis (HS) and focal epilepsy with lesion other than HS.

Genotyping, quality control and imputation

Subjects were genotyped on single nucleotide polymorphism (SNP) arrays, see **Supplementary table 1** for an overview of genotyping in novel cohorts. Quality control (QC) was performed separately for each cohort. Prior to imputation, data from the Janssen, Austrian, Swiss, Norwegian, and BPCCC cohorts were cross-referenced to the HRC panel to ensure SNPs matched in terms of strand, position, and ref/alt allele assignment. Additionally, SNPs were removed if they were absent in the HRC panel, if they had a >20% allele frequency difference with the HRC panel, or if any AT/GC SNPs had MAFs>40%, using tools available from https://www.well.ox.ac.uk/~wrayner/tools/ . Data were then imputed using the the Wellcome Sanger Institutes' imputation server (https://imputation.sanger.ac.uk/), using EAGLE v2.4.1[63] for phasing, and the Positional Burrows Wheeler Transform algorithm[64] for imputation. The Haplotype Reference Consortium (HRC) reference panel r1.1 was used as a reference for imputation[65]. Post-imputation, SNPs with an INFO score of $\leq$0.9 were removed. The high-INFO SNPs were then converted back to PLINK format and once-again QC'd for genotype coverage (>0.98), minor allele frequencies (>5%) and Hardy-Weinberg Equilibrium violations ($p>10^{-5}$), following previously described methodologies[4]. We removed variants <5% MAF in these 5 cohorts for QC reasons, and note there will be a corresponding loss in study power for the 'focal' and 'all epilepsy' epilepsy analysis.

QC for the Epi25 cohort was performed using a similar in-house pipeline. Samples were split by ethnicity based on principal component analysis. Pre-imputation QC included filtering of SNPs with call rate (<98%), differential missing rate, duplicated and monomorphic SNPs, SNPs with batch association ($p<10^{-4}$), violation of Hardy-Weinberg Equilibrium ($p<10^{-10}$). Sample filtering included removal of outliers (>4 SD from mean) of heterozygous/homozygous ratio, removal of one of each pair of related samples (proportion identity-by-descent >0.2) and removal of samples with ambiguous or non-matching genetically imputed sex. Furthermore, 3,180 duplicates between the Epi25 cohort and the previously published genome-wide mega-analysis were identified based on genotype, and were removed from the Epi25 cohort. Of the 3,180 duplicates, 1226 were GGE and 1402 focal epilepsy. Genotypes were imputed on the Michigan imputation server, using the Haplotype Reference Consortium v1.1 (n=32470) reference panel for subjects of European and Asian ancestry, and the 1000 Genomes Phase 3 v5 (n=2504) reference panel for subjects of African ancestry. Default imputation parameters and pre-imputation checks were used. Imputed dosages were used for subsequent analyses, filtering on imputation INFO>0.3 and minor-allele frequency >1%.

Genome-wide association analyses

GWAS of the Janssen Pharmaceuticals, Swiss GenEpa, Norwegian GenEpa and Austrian GenEpa cohorts was performed as a mega-analysis, as described previously.[4] GWAS of the Epi25 cohort was performed with a generalized mixed model using SAIGE v0.38.[66] SAIGE was performed in two steps: (1) fitting the null logistic mixed model to estimate the variance component and other model parameters; (2) testing for the association between each genetic variant and phenotypes by applying SPA to the score test statistics. For step 1, SNPs were filtered on call rate >0.98 and MAF >5%, and SNPs were pruned to obtain approximate independent markers (window size of 100 kb and R^2>0.3), while including sex and the top 10 principal components as covariates. Next, we performed P-value based fixed-effects meta-analyses with METAL[67] for each of the main phenotypes (all, GGE, and focal epilepsy), as well as the subphenotypes, weighted by effective samples sizes ($N_{eff} = 4/(1/N_{cases} + 1/N_{controls})$) to account for case-control imbalance. We performed trans-ethnic and European-only meta-analyses for the main phenotypes, and restricted the subphenotype analyses to Europeans only, due to limited sample size in other ethnicities. We included all SNPs (~4.9 million,

MAF>1%) that were present in at least the previous mega-analysis and the Epi25 dataset, which together account for 88% of the total sample size. We calculated genomic inflation factors (λ), mean χ^2 and LD score regression intercepts to assess potential inflation of the test statistic. Since λ is known to scale with sample size, we also calculated λ1000, which is λ corrected for an equivalent sample size of 1000 cases and 1000 controls.[68] We limited these analyses to subjects of European ancestry, since LD-structure depends on ethnicity and Europeans constituted 92% of cases.

Data sources for the Biobank and deCODE genetics GWAS

Summary statistics for epilepsy GWAS were obtained from three population biobanks; UK Biobank,[69] Biobank Japan,[70,71] Finngen release R6,[72] and from deCODE genetics[73] (Iceland). The biobank Japan, Finngen and deCODE genetics epilepsy cases were further assigned into either 'focal' or 'generalized' epilepsy (see below), whereas the UK Biobank samples were not subdivided based on seizure localisation, as the relevant clinical details were unavailable to facilitate an accurate subdivision (see **Supplementary table 10** for sample sizes per biobank and deCODE genetics). Control data were population matched samples with no history of epilepsy.

Fixed-effects meta-analyses were conducted using METAL[67], weighted by effective sample size ($N_{eff} = 4/(1/N_{cases} + 1/N_{controls})$) to account for case-control imbalance.

UK Biobank: We identified people with epilepsy from the UK Biobank using an analysis of self-reported data, inpatient hospital episode statistics (HES), death certificate diagnostic data and primary care diagnostic data as described elsewhere.[74] This allowed us to interrogate the evidence available to support a diagnosis of epilepsy rather than relying purely on UK Biobank generated data fields 131048 and 13049 based on ICD-10 G40 mapping.

FinnGen: Epilepsy was determined with ICD-10 G40, ICD-9 345, ICD-8 345 and Social Insurance Institution of Finland (KELA) code 111. Exclusion criteria were ICD-9 3452/3453 and ICD-8 34520. GGE was determined with ICD-10 G40.3, ICD-9 345[0-3] and ICD-8 34519. Exclusion criteria were ICD-8 34511. Focal epilepsy was determined with ICD-10 G40.0, G40.1, G40.2, ICD-9 345[45] and ICD-8 3453.

DeCode genetics: Epilepsy was determined with ICD-10 G40 and ICD-9 345 excluding 3452/3453. GGE with ICD-10 G40.3/G40.4/G40.6/G40.7 or ICD-9 3450/3451/3456, and focal epilepsy with ICD-10 G40.0/G40.1/G40.2 or ICD-9 3454/3455.

Biobank Japan: Cases were classified into "Broad_Epilepsy", being any form of epilepsy; "Idiopathic_Epilepsy", being epilepsy with onset under 40 years and no known cause; or "Idiopathic_Focal_Epilepsy" and "Idiopathic_Generalized_Epilepsy", where focal and generalized syndromes could be ascertained.

Control data were population matched samples with no history of epilepsy. GWAS fixed-effects meta-analyses were conducted using METAL[67]. To account for case-control imbalance the effective sample size for each cohort was calculated as $N_{eff} = 4/(1/N_{cases} + 1/N_{controls})$). GWAS Manhattan plots were generated using the qqman R package[75]. Genome-wide significant loci were mapped onto genes using the FUMA web platform[18].

We performed three meta-analyses. As a primary analysis, we meta-analysed all non-biobank samples, then we meta-analysed only biobank/deCODE genetics samples and finally performed a combined meta-analysis of biobank/deCODE genetics and non-biobank samples.

Pleiotropy analysis

ASSET[76] is a meta-analysis-based pleiotropy detection approach that identifies common or shared genetic effects between two or more related, but distinct traits. We used ASSET with a genome-wide significance level of α=5×10^{-8}. We applied ASSET to the subset of European samples, comprising 6952 (3244+3708) GGE cases and 14,939 (5344+9095) focal epilepsy cases from the Epi25 and our Consortium as well as 42,434 partially overlapping controls from both consortia. Note that ASSET accounts for sample overlap in the analysis. Effect sizes, standard errors and the effective sample sizes estimated were from the main meta-analysis.

HLA association

Given the prior association of the HLA with autoimmune epilepsy[77,78], we included a specific analysis of the HLA. HLA types and amino acid residues were imputed using CookHLA software,[26] with the 1000 Genomes Phase 3 used as a reference panel.[79] Samples were grouped by genetic ancestry for imputation.

Following imputation, association analysis was conducted using the HLA Analysis Toolkit (HATK).[80] Three phenotypes were analysed: 'all epilepsy', 'focal epilepsy' and 'GGE'. Samples from the ILAE and Epi25 datasets were analysed separately and the association results were meta-analysed across datasets using PLINK.[81]

Functional annotation

We annotated all genome-wide significant SNPs and tagged SNPs within the loci. ANNOVAR was used to retrieve the location and function of each SNP,[82] the CADD score was used as a measure of predicted deleteriousness[83] and chromatin states were incorporated from the ENCODE and NIH Roadmap Epigenomics Mapping Consortium.[14,84] We used FUMA to define the independently significant SNPs within loci; i.e., SNPs that were genome-wide significant but not in LD (R^2<0.2 in Europeans) with the lead SNP in the locus.

Gene mapping

To map genome-wide significant loci to specific genes, we used FUMA[18] with the same parameters as published previously.[4] We defined genome-wide significant loci as the region encompassing all SNPs with P<10^{-4} that were in LD (R^2>0.2) with the lead SNP (i.e., the SNP with the strongest association within the region). We used a combination of positional mapping (within 250 kb from the locus), eQTL mapping (SNPs with FDR corrected eQTL P<0.05 in blood or brain tissue) and 3D Chromatin Interaction Mapping (FDR p <10^{-6} in brain tissue).

Genome-wide gene based association study and gene-set analyses

We performed the genome-wide gene based association study (GWGAS) using default settings of MAGMA v1.08, as implemented in FUMA, which calculates an association P-value based on all the associations of all SNPs within each gene in the GWAS.[19] Based on these GWGAS results, we performed competitive gene-set analyses with default MAGMA settings, using 15,483 default gene sets and GO-terms from MsigDB. In addition, we specifically assessed 18 curated gene-sets involving different synaptic functions.[29]

Transcriptome wide association study

Transcriptome wide association studies (TWAS) were performed with FUSION v3, with default settings.[20] We imputed gene expression based on our European-only GWAS (since the method relies on LD reference data) eQTL data from the PsychENCODE consortium, which includes dorsolateral prefrontal cortex tissue from 1,695 human subjects.[21]

Summary-data-based Mendelian Randomization

Summary-data-based Mendelian Randomization (SMR) v1.03 is an additional method to assess the association between epilepsy and expression of specific genes.[22] SMR tests whether the effect size of a SNP on epilepsy is mediated by expression of specific genes. We performed SMR analyses with default settings, using the MetaBrain expression data as reference; a new eQTL dataset including 2,970 human brain samples.[85]

Sex-specific analyses

We performed a GWAS as described above for all epilepsy (13,889 female cases and 19,676 female controls; 12,259 male cases and 18,645 male controls) and GGE (3,946 female cases and 19,676 female controls; 2,603 male cases and 18,645 male controls) separately for subjects of either sex, after which we performed fixed-effects meta-analyses with METAL to merge the different cohorts. We performed meta-analyses between the male and female GWAS with GWAMA[86] to assess heterogeneity of effect sizes between sexes and gender-differentiated associations.[33]

Gene prioritization

We combined 10 methods to prioritize the most likely biological candidate gene within each genome-wide significant locus. For each gene in each locus, we assessed the following criteria:

- Missense: we assessed whether the SNPs tagged in the genome-wide significant locus contained an exonic missense variant in the gene, as annotated by ANNOVAR.
- TWAS: we assessed whether imputed gene expression was significantly associated with the epilepsy phenotype, based on the FUSION TWAS as described above, Bonferroni corrected for each mapped gene with expression information.
- SMR: we assessed whether the gene had a significant SMR association with the epilepsy phenotype, based on the SMR analyses as described above, Bonferroni corrected for each mapped gene with expression information.
- MAGMA: we assessed whether the gene was significantly associated with the epilepsy phenotype through a GWGAS analysis, Bonferroni corrected for each mapped gene.
- PoPS: we calculated the Polygenic Priority Score (PoPS)[87]; a novel method that combines GWAS summary statistics with biological pathways, gene expression, and protein-protein interaction data, to pinpoint the most likely causal genes. We scored the gene with the highest PoPS score within each locus.
- Brain expression: we calculated mean expression of all brain and non-brain tissues based on data from the Genotype-Tissue Expression (GTEx) project v8[88] and assessed if the average brain tissue expression was higher than the average expression in non-brain tissues.
- brain-coX: we assessed whether genes were prioritized as co-expressed with established epilepsy genes in more than a third of brain tissue resources utilized, using the tool brain-coX (**Supplementary figure 17**).[89]
- Target of AED: we assessed whether the gene is a known target of an anti-epileptic drug, as detailed in the drug-gene interaction database (www.DGidb.com; accessed on 26-11-2021) and a list of drug targets from a recent publication (**Supplementary data 9**).[90]
- Knockout mouse: we assessed whether a knockout of the gene in a mouse model results in a nervous system (phenotype ID: MP:0003631) or a neurological/behaviour phenotype (MP:0005386) in the Mouse Genome Informatics database (http://www.informatics.jax.org; accessed on 26-11-2021).
- Monogenic epilepsy gene: we evaluated whether the gene is listed as a monogenic epilepsy gene, in a curated list maintained by the Epilepsy Research Centre at the University of Melbourne (**Supplementary data 9**).

Similar to previous studies,[4,91] we scored all genes based on the number of criteria being met (range 0-10; all criteria had an equal weight). The gene with the highest score was chosen as the most likely implicated gene. We implicated both genes if they had an identical, highest score.

Long distance expression regulation of BCL11A

Most eQTL databases, like PsychENCODE and MetaBrain, restrict eQTL analyses to 1 MB distance between genes and SNPs. To specifically assess the hypothesis of long-distance regulation of *BCL11A* by the lead SNPs in the 2p16.1 epilepsy locus, we manually interrogated the MetaBrain database[85] without distance restraints. Next, we calculated the association between the 3 lead SNPs in the locus (rs11688767, rs77876353, rs13416557) with BCL11A expression.

Heritability analyses

We calculated SNP-based heritability on the European-only GWAS using LDAK with default settings and pre-calculated LD weights from 2000 European (white British) reference samples under the BLD-LDAK SumHer model, as recommended for human traits.[92] SNP based heritabilities were converted to liability scale heritability estimates, using the formula: h2L=h2o∗K2(1−K)2/p(1−p)∗Z2, where K is the disease prevalence, p is the proportion of cases in the sample, and Z is the standard normal density at the liability threshold. To decrease downward bias, we performed these calculations based on the effective sample sizes (see calculation above), after which p=0.5 can be assumed,[93] with the same population prevalences as our previous study.[4]

The total amount of causally associated variants (i.e., variants with nonzero additive genetic effect) underlying epilepsy risk was calculated by a causal mixture model (MiXeR).[36] MiXeR utilizes a likelihood-based framework to estimate the amount of causal SNPs underlying a trait, without the need to pinpoint which specific SNPs are involved. Furthermore, MiXeR allows for power calculations to assess the required sample size to explain a certain proportion SNP-based heritability by genome-wide significant SNPs.

Enrichment analyses

We used MAGMA (as implemented in FUMA) to perform tissue and cell-type enrichment. First, we assessed whether our GGE GWAS was enriched for specific tissues from the GTEx database. Similarly, we assessed enrichment of genes expressed in the brain at 11 general developmental stages, using data from the BrainSpan consortium. Next, we assessed whether GGE was associated with specific cell types, by cross-referencing two single-cell RNA sequencing databases of human developmental and adult brain samples. The PsychENCODE database contains RNA sequencing data from 4,249 human brain cells from developmental stages and 27,412 human adult brain cells.[94] The Zhong dataset (GSE104276) contains RNA sequencing data from 2,309 human brain cells at different stages in development.[95] We performed FDR correction across datasets to assess which cell types were significantly associated with GGE. As sensitivity analysis, we performed stratified LDSC with default settings using the cell-specific gene expression weights from the PsychENCODE consortium to compare GABAergic with glutamatergic neuron enrichment.[96]

Genetic overlap with other diseases

Using the FUMA web application, we searched the GWAS Catalog for previously reported associations with $P< 5x10^{-8}$ for SNPs at all 26 genome-wide significant loci.

Genetic correlations between all, focal epilepsy and GGE and other traits were computed with LDSC, using default settings. Traits highlighted by the GWAS catalog analysis and/or those with established epilepsy comorbidity were prioritized and pursued provided recent summary statistics were available for public download (**Supplementary table 11**).

We used a recently described bivariate causal mixture model to quantify polygenic overlap between GGE with intelligence and autism spectrum disorder (ASD). Publicly available summary statistics from intelligence (n=269867) and ASD GWAS (n=46350) were downloaded,[97,98] after which bivariate MiXeR was run with default settings.

Drug-repurposing analyses

We utilized a recently developed method that uses the GWAS for a disease to predict the relative efficacy of drugs for the disease.[42] We applied this method to the all epilepsy and GGE GWAS results, using (1) imputed gene expression data from the FUSION analyses, as described above, and (2) gene-based p-values from MAGMA (see above), with default settings. We predicted the relative efficacy of 1343 drugs in total (**Supplementary data 8**). We determined if our predictions correctly identify (area under receiver operating characteristic curve) and prioritize (median rank) known clinically-effective antiseizure drugs, as previously described.[42] We determined the statistical significance of drug identification and prioritization results by comparing the results to those from a null distribution generated by performing 10^6 random permutations of the scores assigned to drugs.

Data availability

The GWAS summary statistics data that support the findings of this study (for both trans-ethnic and European-only analyses) are available at https://www.epigad.org/.

References

1. Fisher, R. S. *et al.* ILAE official report: a practical clinical definition of epilepsy. *Epilepsia* **55**, 475–482 (2014).
2. Fiest, K. M. *et al.* Prevalence and incidence of epilepsy: A systematic review and meta-analysis of international studies. *Neurology* **88**, 296–303 (2017).
3. Scheffer, I. E. *et al.* ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. *Epilepsia* **58**, 512–521 (2017).
4. International League Against Epilepsy Consortium on Complex Epilepsies. Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. *Nat. Commun.* **9**, 5269 (2018).
5. Epi4K consortium & Epilepsy Phenome/Genome Project. Ultra-rare genetic variation in common epilepsies: a case-control sequencing study. *Lancet Neurol.* **16**, 135–143 (2017).
6. Leu, C. *et al.* Polygenic burden in focal and generalized epilepsies. *Brain* **142**, 3473–3481 (2019).
7. Koko, M. *et al.* Distinct gene-set burden patterns underlie common generalized and focal epilepsies. *EBioMedicine* **72**, 103588 (2021).
8. McTague, A., Howell, K. B., Cross, J. H., Kurian, M. A. & Scheffer, I. E. The genetic landscape of the epileptic encephalopathies of infancy and childhood. *Lancet Neurol.* **15**, 304–316 (2016).
9. Speed, D. *et al.* Describing the genetic architecture of epilepsy through heritability analysis. *Brain* **137**, 2680–2689 (2014).
10. Motelow, J. E. *et al.* Sub-genic intolerance, ClinVar, and the epilepsies: A whole-exome sequencing study of 29,165 individuals. *Am. J. Hum. Genet.* **108**, 965–982 (2021).
11. Chen, Z., Brodie, M. J., Liew, D. & Kwan, P. Treatment Outcomes in Patients With Newly Diagnosed Epilepsy Treated With Established and New Antiepileptic Drugs: A 30-Year Longitudinal Cohort Study. *JAMA Neurol.* **75**, 279–286 (2018).
12. Devinsky, O. *et al.* Epilepsy. *Nat Rev Dis Primers* **4**, 18024 (2018).
13. Chan, S. S. L. & Copeland, W. C. DNA polymerase gamma and mitochondrial disease: understanding the consequence of POLG mutations. *Biochim. Biophys. Acta* **1787**, 312–319 (2009).
14. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
15. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
16. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
17. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
18. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
19. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
20. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
21. Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, (2018).
22. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
23. Xu, C. *et al.* Knockdown of RMI1 impairs DNA repair under DNA replication stress. *Biochem. Biophys. Res. Commun.* **494**, 158–164 (2017).
24. International League Against Epilepsy Consortium on Complex Epilepsies. Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **13**, 893–903 (2014).

25. Yoshida, M. *et al.* Identification of novel BCL11A variants in patients with epileptic encephalopathy: Expanding the phenotypic spectrum. *Clin. Genet.* **93**, 368–373 (2018).
26. Cook, S. *et al.* Accurate imputation of human leukocyte antigens with CookHLA. *Nat. Commun.* **12**, 1264 (2021).
27. Holland, D. *et al.* Beyond SNP heritability: Polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model. *PLoS Genet.* **16**, e1008612 (2020).
28. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
29. Ruano, D. *et al.* Functional gene group analysis reveals a role of synaptic heterotrimeric G proteins in cognitive ability. *Am. J. Hum. Genet.* **86**, 113–125 (2010).
30. Lukyanetz, E. A., Shkryl, V. M. & Kostyuk, P. G. Selective blockade of N-type calcium channels by levetiracetam. *Epilepsia* **43**, 9–18 (2002).
31. Wang, S. J., Huang, C. C., Hsu, K. S., Tsai, J. J. & Gean, P. W. Inhibition of N-type calcium currents by lamotrigine in rat amygdalar neurones. *Neuroreport* **7**, 3037–3040 (1996).
32. Marson, A. *et al.* The SANAD II study of the effectiveness and cost-effectiveness of levetiracetam, zonisamide, or lamotrigine for newly diagnosed focal epilepsy: an open-label, non-inferiority, multicentre, phase 4, randomised controlled trial. *Lancet* **397**, 1363–1374 (2021).
33. Christensen, J., Kjeldsen, M. J., Andersen, H., Friis, M. L. & Sidenius, P. Gender differences in epilepsy. *Epilepsia* **46**, 956–960 (2005).
34. Magi, R., Lindgren, C. M. & Morris, A. P. Meta-analysis of sex-specific genome-wide association studies. *Genet. Epidemiol.* **34**, 846–853 (2010).
35. Gaborit, N. *et al.* Gender-related differences in ion-channel and transporter subunit expression in non-diseased human hearts. *J. Mol. Cell. Cardiol.* **49**, 639–646 (2010).
36. Frei, O. *et al.* Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation. *Nat. Commun.* **10**, 2417 (2019).
37. Long, S. *et al.* The Clinical and Genetic Features of Co-occurring Epilepsy and Autism Spectrum Disorder in Chinese Children. *Front. Neurol.* **10**, 505 (2019).
38. Jeste, S. S. & Tuchman, R. Autism Spectrum Disorder and Epilepsy: Two Sides of the Same Coin? *J. Child Neurol.* **30**, 1963–1971 (2015).
39. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
40. Lauxmann, S. *et al.* An SCN2A mutation in a family with infantile seizures from Madagascar reveals an increased subthreshold Na(+) current. *Epilepsia* **54**, e117–21 (2013).
41. Ben-Shalom, R. *et al.* Opposing Effects on NaV1.2 Function Underlie Differences Between SCN2A Variants Observed in Individuals With Autism Spectrum Disorder or Infantile Seizures. *Biol. Psychiatry* **82**, 224–232 (2017).
42. Mirza, N. *et al.* Using common genetic variants to find drugs for common epilepsies. *Brain Commun* **3**, fcab287 (2021).
43. Bourgeois, B. F. D. Chronic management of seizures in the syndromes of idiopathic generalized epilepsy. *Epilepsia* **44 Suppl 2**, 27–32 (2003).
44. Marson, A. G. *et al.* The SANAD study of effectiveness of valproate, lamotrigine, or topiramate for generalised and unclassifiable epilepsy: an unblinded randomised controlled trial. *Lancet* **369**, 1016–1026 (2007).
45. Punetha, J. *et al.* Biallelic CACNA2D2 variants in epileptic encephalopathy and cerebellar atrophy. *Ann Clin Transl Neurol* **6**, 1395–1406 (2019).
46. Fariello, R. G. Safinamide. *Neurotherapeutics* **4**, 110–116 (2007).
47. Alsaegh, H., Eweis, H., Kamal, F. & Alrafiah, A. Celecoxib Decrease Seizures Susceptibility in a Rat Model of Inflammation by Inhibiting HMGB1 Translocation. *Pharmaceuticals* **14**, (2021).
48. Johannesen, K. M. *et al.* Genotype-phenotype correlations in SCN8A-related disorders reveal prognostic and therapeutic implications. *Brain* (2021) doi:10.1093/brain/awab321.
49. Ma, M.-G. *et al.* RYR2 Mutations Are Associated With Benign Epilepsy of Childhood With

Centrotemporal Spikes With or Without Arrhythmia. *Front. Neurosci.* **15**, 629610 (2021).
50. Yap, S. M. & Smyth, S. Ryanodine receptor 2 (RYR2) mutation: A potentially novel neurocardiac calcium channelopathy manifesting as primary generalised epilepsy. *Seizure* **67**, 11–14 (2019).
51. EPICURE Consortium *et al.* Genome-wide association analysis of genetic generalized epilepsies implicates susceptibility loci at 1q43, 2p16.1, 2q22.3 and 17q21.32. *Hum. Mol. Genet.* **21**, 5359–5372 (2012).
52. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593–1599 (2018).
53. Heinzen, E. L. Somatic variants in epilepsy - advancing gene discovery and disease mechanisms. *Curr. Opin. Genet. Dev.* **65**, 1–7 (2020).
54. Beesley, L. J. *et al.* The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. *Stat. Med.* **39**, 773–800 (2020).
55. Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* **9**, 2941 (2018).
56. Hautakangas, H. *et al.* Genome-wide analysis of 102,084 migraine cases identifies 123 risk loci and subtype-specific risk alleles. *Nat. Genet.* **54**, 152–160 (2022).
57. Wightman, D. P. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat. Genet.* **53**, 1276–1282 (2021).
58. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, (2019).
59. Wood, M. D. & Gillard, M. Evidence for a differential interaction of brivaracetam and levetiracetam with the synaptic vesicle 2A protein. *Epilepsia* **58**, 255–262 (2017).
60. Singh, T. *et al.* Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature* (2022) doi:10.1038/s41586-022-04556-w.
61. van Rheenen, W. *et al.* Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nat. Genet.* **53**, 1636–1648 (2021).
62. Reay, W. R. & Cairns, M. J. Advancing the use of genome-wide association studies for drug repurposing. *Nat. Rev. Genet.* **22**, 658–671 (2021).
63. Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).
64. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 120–126 (2021).
65. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
66. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
67. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
68. de Bakker, P. I. W. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, R122–8 (2008).
69. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
70. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
71. Ishigaki, K. *et al.* Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nat. Genet.* **52**, 669–679 (2020).
72. Locke, A. E. *et al.* Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature* **572**, 323–328 (2019).
73. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).

74. Campbell, C. *et al.* Polygenic risk score analysis reveals shared genetic burden between epilepsy and psychiatric comorbidities. *Under review* (2022).
75. Turner, S. D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv* 005165 (2014) doi:10.1101/005165.
76. Bhattacharjee, S. *et al.* A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am. J. Hum. Genet.* **90**, 821–835 (2012).
77. Kim, T.-J. *et al.* Anti-LGI1 encephalitis is associated with unique HLA subtypes. *Ann. Neurol.* **81**, 183–192 (2017).
78. van Sonderen, A. *et al.* Anti-LGI1 encephalitis is strongly associated with HLA-DR7 and HLA-DRB4. *Ann. Neurol.* **81**, 193–198 (2017).
79. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
80. Choi, W., Luo, Y., Raychaudhuri, S. & Han, B. HATK: HLA analysis toolkit. *Bioinformatics* **37**, 416–418 (2021).
81. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
82. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
83. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
84. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
85. de Klein, N. *et al.* Brain expression quantitative trait locus and network analysis reveals downstream effects and putative drivers for brain-related diseases. *bioRxiv* 2021.03.01.433439 (2021) doi:10.1101/2021.03.01.433439.
86. Mägi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**, 288 (2010).
87. Weeks, E. M. *et al.* Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *bioRxiv* (2020) doi:10.1101/2020.09.08.20190561.
88. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
89. Freytag, S., Burgess, R., Oliver, K. L. & Bahlo, M. brain-coX: investigating and visualising gene co-expression in seven human brain transcriptomic datasets. *Genome Med.* **9**, 55 (2017).
90. Rodriguez-Acevedo, A. J., Gordon, L. G., Waddell, N., Hollway, G. & Vadlamudi, L. Developing a gene panel for pharmacoresistant epilepsy: a review of epilepsy pharmacogenetics. *Pharmacogenomics* **22**, 225–234 (2021).
91. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
92. Speed, D., Holmes, J. & Balding, D. J. Evaluating and improving heritability models using summary statistics. *Nat. Genet.* **52**, 458–462 (2020).
93. Grotzinger, A. D., de la Fuente, J., Nivard, M. G. & Tucker-Drob, E. M. Pervasive downward bias in estimates of liability scale heritability in GWAS meta-analysis: A simple solution. *bioRxiv* (2021) doi:10.1101/2021.09.22.21263909.
94. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, (2018).
95. Zhong, S. *et al.* A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524–528 (2018).
96. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
97. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new

genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).

98. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).

Author contributions

Data analysis:

Analytical design, imputation: O.M.Adesoji, M.Bahlo, C.Campbell (lead analyst), G.L.Cavalleri, S.Chen (lead analyst), Y-C.A.Feng, B.P.C.Koeleman, R.Krause (data management), D.Lal, C.Leu, N.Mirza, M.Nothnagel, K.L.Oliver, R.Stevelink (lead analyst).

Data generation and quality control and management: L.Baum, J.P.Bradfield, R.J.Buono, G.L.Cavalleri., F.Cerrato, S.S.Cherny, C.Churchhouse, C.Cusick, Y-C.A.Feng, N.Gupta, H.Hakonarson, E.L.Heinzen, I.Helbig, D.P.Howrigan, D.Kasperaviciute, B.P.C.Koeleman, R.Krause., D.Lal, Z.Landoulsi, C.Leu, I.Lopes-Cendes., P.May, N.Mirza, B.M.Neale, P.-W.Ng, P.Nürnberg, Sl.Petrovski, T.Sander, D.Speed, R.Stevelink, Fe.Zara, W.Zhou.

External data resources and analysis: UK BioBank: C.Campbell, D.Lewis-Smith, R.H.Thomas.

BioBank Japan: Y.Kamatani, M.Kanai, M.Kato, Y.Okada.

FinnGenn: M.J.Daly, H.O.Heyne, R.Kälviäinen, M.I.Kurki, A.Palotie.

deCODE genetics: S.Magnusson, E.Ólafsson, H.Stefansson, K.Stefansson, U.Unnsteinsdóttir.

Analysis coordination: G.L.Cavalleri (Co-Chair), B.P.C.Koeleman (Co-Chair)

Writing committee: O.M.Adesoji, M.Bahlo, S.F.Berkovic, C.Campbell, G.L.Cavalleri, S.Chen, B.P.C.Koeleman, K.L.Oliver, R.Stevelink (wrote first draft).

Strategy committee: L.Baum, S.F.Berkovic (Chair), R.J.Buono, G.L.Cavalleri, H.Hakonarson, E.L.Heinzen, M.R.Johnson, R.Kalviainen, B.P.C.Koeleman, R.Krause, P.Kwan, D.Lal, H.Lerche, Q.S.Li, I.Lopes-Cendes, D.H.Lowenstein, T.J.O'Brien, S.M.Sisodiya.

Phenotyping committee: C.Depondt, D.J.Dlugos, W.S.Kunz, P.Kwan, D.H.Lowenstein (Chair), A.G.Marson, P.Striano.

Governance committee: S.F.Berkovic, A.Compston, A-E.Lehesjoki, D.H.Lowenstein.

Patient recruitment and phenotyping: Z.Afawi, E.Amadori, A.Anderson, J.Anderson, D.M.Andrade, G.Annesi, A.Avbersek, M.D.Baker, G.Balagura, S.Balestrini, C.Barba, K.Barboza, F.Bartolomei, T.Bast, T.Baumgartner, B.Baykan, N.Bebek, A.J.Becker, F.Becker, C.A.Bennett, B.Berghuis, S.F.Berkovic, A.Beydoun, C.Bianchini, F.Bisulli, I.Blatt, I.Borggraefe, C.Bosselmann, V.Braatz, K.Brockmann, R.J.Buono, R.M.Busch, H.Caglayan, E.Campbell, L.Canafoglia, C.Canavati, G.D.Cascino, B.Castellotti, C.B.Catarino, F.Chassoux, K.Chinthapalli, I-J.Chou, S-K.Chung, P.O.Clark, A.J.Cole, A.Coppola, M.Cosico, P.Cossette, J.J.Craig, L.K.Davis, G-J.deHaan, N.Delanty, C.Depondt, P.Derambure, O.Devinsky, L.Di Vito, D.J.Dlugos, V.Doccini, C.P.Doherty, H.El-Naggar, C.E.Elger, C.A.Ellis, A.Faucon, L.Ferguson, T.N.Ferraro, L.Ferri, M.Feucht, M.Fitzgerald, B.Fonferko-Shadrach, F.Fortunato, S.Franceschetti, J.A.French, E.Freri, M.Gagliardi, A.Gambardella, E.B.Geller, T.Giangregorio, L.Gjerstad, T.Glauser, E.Goldberg, A.Goldman, T.Granata, D.A.Greenberg, R.Guerrini, K.Hallmann, M.Hegde, I.Helbig, C.Hengsbach, S.Hirose, E.Hirsh, H.Hjalgrim, P-C.Hung, M.Iacomino, L.L.Imbach, Y.Inoue, A.Ishii, J.Jamnadas-Khoda, L.Jehi, M.R.Johnson, R.Kälviainen, M.Kanaan, A.-M.Kantanen, B.Kara, S.M.Kariuki, D. Kasteleijn-Nolst Trenite, J.Kegele, Y.Kesim, N.Khoueiry-Zgheib, C.King, H.E.Kirsch, K.M.Klein, G. Kluger, S.Knake, R.C.Knowlton, A.D.Korczyn, A.Koupparis, I.Kousiappa, M.Krenn, H.Krestel, I.Krey, W.S.Kunz, G.Kurlemann, Ru.Kuzniecky, P.Kwan, A.Labate, A.Lacey, S.Lauxmann, S.L.Leech, A-E.Lehesjoki, J.R.Lemke, H.Lerche, G.Lesca, B.Neubauer, N.Lewin, Q.S.Li,

L.Licchetta, K-L.Lin, D.Lindhout, T.Linnankivi, I.Lopes-Cendes, D.H.Lowenstein, C.H.T.Lui, F.Madia, A.G.Marson, C.M.McGraw, D.Mei, R.Minardi, R.S.Moller, M.Montomoli, B.Mostacci, L.Muccioli, H.Muhle, K.Müller-Schlüter, I.M.Najm, W.Nasreddine, C.R.J.C.Newton, T.J.O'Brien, Ç.Özkara, S.S.Papacostas, E.Parrini, M.Pendziwiat, W.O.Pickrell, R.Pinsky, T.Pippucci, An.Poduri, F.Pondrelli, R.H.W.Powell, M.Privitera, A.Rademacher, R.Radtke, F.Ragona, S.Rau, M.I.Rees, B.M.Regan, P.S.Reif, S.Rhelms, A.Riva, F.Rosenow, P.Ryvlin, A.Saarela, L.G.Sadleir, J.W.Sander, Th.Sander, M.Scala, Th.Scattergood, S.C.Schachter, C.J.Schankin, I.E.Scheffer, B.Schmitz, S.Schoch, S.Schubert-Bast, A.Schulze-Bonhage, P.Scudieri, B.R.Sheidley, J.J.Shih, G.J.Sills, S.M.Sisodiya, M.C.Smith, P.E.Smith, A.C.M.Sonsma, M.R.Sperling, B.J.Steinhoff, U.Stephani, W.C.Stewart, C.Stipa, P.Striano, H.Stroink, A.Strzelczyk, R.Surges, T.Suzuki, K.M.Tan, G.A.Tanteles, E.Tauboll, L.L.Thio, O.Timonen, P.Tinuper, M.Todaro, P.Topaloglu, R.Tozzi, M-H.Tsai, B.Tumiene, D.Turkdogan, A.Utkus, P.Vaidiswaran, L.Valton, A.van Baalen, A.Vetro, E.P.G.Vining, F.Visscher, S.von Brauchitsch, R.von Wrede, R.G.Wagner, Y.G.Weber, S.Weckhuysen, J.Weisenberg, M.Weller, C.D.Whelan, P.Widdess-Walsh, M.Wolff, S.Wolking, D.Wu, K.Yamakawa, Z.Yapici, E.Yücesan, S.Zagaglia, F.Zahnert, F.Zimprich, G.Zsurka, Q.Zulfiqar Ali.

Control cohorts: L.C.Brody, J.G.Eriksson, A.Franke, H.Hakonarson, Y.-L.Lau, J.L.Mills, A.M.Molloy, M.M.Nöthen, A.Palotie, F.Pangilinan, H.Stroink, W.Yang.

Consortium coordination: K.L.Oliver.

Author names and affiliations

The International League Against Epilepsy Consortium on Complex Epilepsies*

*three lead analysts listed first followed by all members in alphabetical order

Remi Stevelink[1], Ciarán Campbell[2, 3], Siwei Chen[4, 5], Oluyomi M Adesoji[6], Zaid Afawi[7], Elisabetta Amadori[8, 9], Alison Anderson[10, 11], Joseph Anderson[12], Danielle M Andrade[13], Grazia Annesi[14], Andreja Avbersek[15], Melanie Bahlo[16-18], Mark D Baker[19], Ganna Balagura[8, 9], Simona Balestrini[15, 20], Carmen Barba[21], Karen Barboza[22], Fabrice Bartolomei[23], Thomas Bast[24, 25], Larry Baum[26, 27], Tobias Baumgartner[28], Betül Baykan[29, 30], Nerses Bebek[29, 30], Albert J Becker[31], Felicitas Becker[32], Caitlin A Bennett[33], Bianca Berghuis[34], Samuel F Berkovic[33], Ahmad Beydoun[35], Claudia Bianchini[21], Francesca Bisulli[36, 37], Ilan Blatt[7, 38], Ingo Borggraefe[39, 40], Christian Bosselmann[41], Vera Braatz[15, 20], Jonathan P Bradfield[42, 43], Knut Brockmann[44], Lawrence C Brody[45], Russell J Buono[42, 46, 47], Robyn M Busch[48-50], Hande Caglayan[51], Ellen Campbell[52], Laura Canafoglia[53], Christina Canavati[54], Gregory D Cascino[55], Barbara Castellotti[56], Claudia B Catarino[15], Gianpiero L Cavalleri[2, 3], Felecia Cerrato[57], Francine Chassoux[58], Stacey S Cherny[26, 59], Krishna Chinthapalli[15], I-Jun Chou[60], Seo-Kyung Chung[61, 62], Claire Churchhouse[4, 5, 57], Peggy O Clark[63], Andrew J Cole[64], Alastair Compston[65], Antonietta Coppola[66], Mahgenn Cosico[67, 68], Patrick Cossette[69], John J Craig[70], Caroline Cusick[57], Mark J Daly[4, 5, 57, 71], Lea K Davis[72-75], Gerrit-Jan de Haan[76], Norman Delanty[2, 3, 77], Chantal Depondt[78], Philippe Derambure[79], Orrin Devinsky[80], Lidia Di Vito[36], Dennis J Dlugos[67], Viola Doccini[21], Colin P Doherty[3, 81], Hany El-Naggar[2, 3, 77], Christian E Elger[28], Colin A Ellis[82], Johan G Eriksson[83], Annika Faucon[84], Yen-Chen A Feng[4, 5, 57, 85, 86], Lisa Ferguson[49], Thomas N Ferraro[46, 87], Lorenzo Ferri[36, 37], Martha Feucht[88], Mark Fitzgerald[67, 68, 82], Beata Fonferko-Shadrach[19], Francesco Fortunato[89], Silvana Franceschetti[90], Andre Franke[91], Jacqueline A French[92], Elena Freri[93], Monica Gagliardi[14], Antonio Gambardella[89], Eric B Geller[94], Tania Giangregorio[36], Leif Gjerstad[95], Tracy Glauser[63], Ethan Goldberg[67, 68], Alicia Goldman[96], Tiziana Granata[93], David A Greenberg[97], Renzo Guerrini[21], Namrata Gupta[5], Hakon Hakonarson[42, 98], Kerstin Hallmann[28, 99], Manu Hegde[100], Erin L Heinzen[101, 102], Ingo Helbig[67, 68, 82, 91, 103, 104], Christian Hengsbach[41], Henrike O Heyne[5, 71, 105, 106], Shinichi Hirose[107], Edouard Hirsch[108], Helle Hjalgrim[109, 110], Daniel P Howrigan[4, 5, 57], Po-Cheng Hung[60], Michele Iacomino[9], Lukas L Imbach[111], Yushi Inoue[112], Atsushi Ishii[113], Jennifer Jamnadas-Khoda[15, 114], Lara Jehi[49, 50], Michael R Johnson[115], Reetta Kälviäinen[116, 117], Yoichiro Kamatani[118], Moien Kanaan[54], Masahiro Kanai[119, 120], Anne-Mari Kantanen[116], Bülent Kara[121], Symon M Kariuki[122-124], Dalia Kasperavičiūte[15], Dorothee Kasteleijn-Nolst Trenite[1], Mitsuhiro Kato[125], Josua Kegele[41], Yeşim Kesim[29], Nathalie Khoueiry-Zgheib[126], Chontelle King[127], Heidi E Kirsch[100], Karl M Klein[128-131], Gerhard Kluger[132, 133], Susanne Knake[128, 131], Robert C Knowlton[100], Bobby P C Koeleman[1], Amos D Korczyn[7], Andreas Koupparis[134], Ioanna Kousiappa[134], Roland Krause[135], Martin Krenn[136], Heinz Krestel[129, 131, 137, 138], Ilona Krey[139], Wolfram S Kunz[28, 140], Mitja I Kurki[4, 5, 57, 71], Gerhard Kurlemann[141], Ruben Kuzniecky[142], Patrick Kwan[10, 11, 143], Angelo Labate[144], Austin Lacey[3, 77, 145], Dennis Lal[48, 49, 57], Zied Landoulsi[135], Yu-Lung Lau[146], Stephen Lauxmann[41], Stephanie L Leech[33], Anna-Elina Lehesjoki[147], Johannes R Lemke[139], Holger Lerche[41], Gaetan Lesca[148], Costin Leu[15, 48, 57], Naomi Lewin[67, 68], David Lewis-Smith[67, 104, 149, 150], Qingqin S Li[151], Laura Licchetta[36], Kuang-Lin Lin[60], Dick Lindhout[1, 76], Tarja Linnankivi[152-154], Iscia Lopes-Cendes[155], Daniel H Lowenstein[100], Colin H T Lui[156], Francesca Madia[9], Sigurdur Magnusson[157], Anthony G Marson[158], Patrick May[135], Christopher M McGraw[64], Davide Mei[21], James L Mills[159], Raffaella Minardi[36], Nasir Mirza[158], Rikke S Møller[109, 110], Anne M Molloy[160], Martino Montomoli[21], Barbara Mostacci[36], Lorenzo Muccioli[37], Hiltrud Muhle[103], Karen Müller-Schlüter[161], Imad M Najm[49, 50], Wassim Nasreddine[35], Benjamin M Neale[4, 5, 57], Bernd Neubauer[162], Charles RJC Newton[122-124], Markus M Nöthen[163], Michael Nothnagel[6, 164], Peter Nürnberg[6], Terence J O'Brien[10, 11], Yukinori Okada[120, 165], Elías Ólafsson[166], Karen L Oliver[16, 17, 33], Çiğdem Özkara[167], Aarno Palotie[4, 5, 57, 71], Faith Pangilinan[45], Savvas S Papacostas[134], Elena Parrini[21], Manuela Pendziwiat[91, 103], Slavé Petrovski[10, 168], William O Pickrell[19, 169], Rebecca Pinsky[170], Tommaso Pippucci[171], Annapurna Poduri[170], Federica Pondrelli[37], Rob H W Powell[169], Michael Privitera[172], Annika Rademacher[103], Rodney Radtke[173], Francesca Ragona[93], Sarah Rau[41], Mark

I Rees[62, 174], Brigid M Regan[33], Philipp S Reif[128, 129, 131], Sylvain Rhelms[175, 176], Antonella Riva[8, 9], Felix Rosenow[128, 129, 131], Philippe Ryvlin[177], Anni Saarela[116, 117], Lynette G Sadleir[127], Josemir W Sander[15, 20, 76], Thomas Sander[6, 178], Marcello Scala[8, 9], Theresa Scattergood[179], Steven C Schachter[180], Christoph J Schankin[137, 181], Ingrid E Scheffer[33, 182], Bettina Schmitz[178], Susanne Schoch[31], Susanne Schubert-Bast[129, 131], Andreas Schulze-Bonhage[183], Paolo Scudieri[8, 9], Beth R Sheidley[170], Jerry J Shih[184], Graeme J Sills[185], Sanjay M Sisodiya[15, 20], Michael C Smith[186], Philip E Smith[187], Anja C M Sonsma[1], Doug Speed[188, 189], Michael R Sperling[190], Hreinn Stefansson[157], Kári Stefansson[157], Bernhard J Steinhoff[24, 25], Ulrich Stephani[103], William C Stewart[191, 192], Carlotta Stipa[36], Pasquale Striano[8, 9], Hans Stroink[193], Adam Strzelczyk[128, 129, 131], Rainer Surges[28], Toshimitsu Suzuki[194, 195], K Meng Tan[10], George A Tanteles[134], Erik Taubøll[95], Liu Lin Thio[196], Rhys H Thomas[149, 150], Oskari Timonen[117], Paolo Tinuper[36, 37], Marian Todaro[10, 11], Pınar Topaloğlu[197], Rossana Tozzi[198], Meng-Han Tsai[199], Birute Tumiene[200, 201], Dilsad Turkdogan[202], Unnur Unnsteinsdóttir[157], Algirdas Utkus[201], Priya Vaidiswaran[67, 68], Luc Valton[203], Andreas van Baalen[103], Annalisa Vetro[21], Eileen P G Vining[204], Frank Visscher[205], Sophie von Brauchitsch[129, 131], Randi von Wrede[28], Ryan G Wagner[206], Yvonne G Weber[41, 207], Sarah Weckhuysen[208-210], Judith Weisenberg[196], Michael Weller[211], Peter Widdess-Walsh[2, 3, 77], Markus Wolff[212], Stefan Wolking[207], David Wu[84], Kazuhiro Yamakawa[194, 195], Wanling Yang[146], Zuhal Yapıcı[197], Emrah Yücesan[213], Sara Zagaglia[15, 20], Felix Zahnert[128], Federico Zara[8, 9], Wei Zhou[4, 5, 57], Fritz Zimprich[136], Gábor Zsurka[28, 140], Quratulain Zulfiqar Ali[13]

1. Department of Genetics, University Medical Center Utrecht, Utrecht 3584 CX, The Netherlands.
2. School of Pharmacy and Biomolecular Sciences, The Royal College of Surgeons in Ireland, Dublin, Ireland.
3. The FutureNeuro Research Centre, Dublin, Ireland.
4. Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA.
5. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
6. Cologne Center for Genomics (CCG), University of Cologne, Faculty of Medicine and University Hospital Cologne, 50931 Cologne, Germany.
7. Tel-Aviv University Sackler Faculty of Medicine, Ramat Aviv 69978, Israel.
8. Department of Neurosciences, Rehabilitation, Ophthalmology, Genetics, Maternal and Child Health, University of Genova, Genova, Italy.
9. IRCCS Istituto Giannina Gaslini, Genova, Italy.
10. Department of Medicine, University of Melbourne, Royal Melbourne Hospital, Parkville 3050, Australia.
11. Department of Neuroscience, Central Clinical School, Alfred Health, Monash University, Melbourne 3004, Australia.
12. Neurology Department, Aneurin Bevan University Health Board, Newport, Wales, UK.
13. Adult Genetic Epilepsy Program, University of Toronto, Toronto, ON, Canada.
14. Institute for Biomedical Research and Innovation, National Research Council, Cosenza, Italy.
15. Department of Clinical and Experimental Epilepsy, UCL Queen Square Institute of Neurology, London WC1N 3BG, UK.
16. Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Parkville 3052, Australia.
17. Department of Biology, University of Melbourne, Parkville 3010, Australia.
18. School of Mathematics and Statistics, University of Melbourne, Parkville 3010, Australia.
19. Swansea University Medical School, Swansea University, Swansea, Wales, UK.
20. Chalfont Centre for Epilepsy, Chalfont-St-Peter, Buckinghamshire SL9 0RJ, UK.
21. Pediatric Neurology, Neurogenetics and Neurobiology Unit and Laboratories, Children's Hospital A. Meyer, University of Florence, Italy.
22. University Health Network, University of Toronto, Toronto, ON, Canada.

23. APHM, Timone Hospital, Epileptology and Cerebral Rhythmology, Aix Marseille Univ, INSERM, INS, Inst Neurosci Syst, Marseille, France.
24. Epilepsy Center Kork, Kehl-Kork 77694, Germany.
25. Medical Faculty of the University of Freiburg, Freiburg 79085, Germany.
26. Department of Psychiatry, The University of Hong Kong, Hong Kong.
27. The State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong, Hong Kong, China.
28. Department of Epileptology, University of Bonn Medical Centre, Bonn 53127, Germany.
29. Department of Neurology, Istanbul Faculty of Medicine, Istanbul University, Istanbul, Turkey.
30. Department of Genetics, Aziz Sancar Institute of Experimental Medicine, Istanbul University, Istanbul, Turkey.
31. Section for Translational Epilepsy Research, Department of Neuropathology, University of Bonn Medical Center, Bonn 53105, Germany.
32. Department of Neurology, University of Ulm, Ulm 89081, Germany.
33. Epilepsy Research Centre, University of Melbourne, Austin Health, Heidelberg 3084, Australia.
34. Stichting Epilepsie Instellingen Nederland (SEIN), Zwolle 8025 BV, The Netherlands.
35. Department of Neurology, American University of Beirut Medical Center, Beirut, Lebanon.
36. IRCCS Istituto delle Scienze Neurologiche di Bologna, Bologna, Italy.
37. Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna, Italy.
38. Department of Neurology, Sheba Medical Center, Ramat Gan, Israel.
39. Department of Pediatric Neurology, Dr von Hauner Children's Hospital, Ludwig Maximilians University, Munchen, Germany.
40. Epilepsy Center Munich, Munich, Germany.
41. Department of Neurology and Epileptology, Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen 72076, Germany.
42. Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.
43. Quantinuum Research LLC, Wayne, PA 19087, USA.
44. Children's Hospital, Dept. of Pediatric Neurology, University Medical Center Göttingen, Göttingen, Germany.
45. National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA.
46. Department of Biomedical Sciences, Cooper Medical School of Rowan University Camden, NJ 08103, USA.
47. Department of Neurology, Thomas Jefferson University Hospital, Philadelphia, PA 19107, USA.
48. Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA.
49. Cleveland Clinic Epilepsy Center, Neurological Institute, Cleveland Clinic, Cleveland, OH 44195, USA.
50. Department of Neurology, Neurological Institute, Cleveland Clinic, Cleveland, OH 44195, USA.
51. Department of Molecular Biology and Genetics, Bogaziçi University, Istanbul, Turkey.
52. Belfast Health and Social Care Trust, Belfast BT9 7AB, UK.
53. Integrated Diagnostics for Epilepsy, Fondazione IRCCS Istituto Neurologico C. Besta, Milan, Italy.
54. Hereditary Research Lab, Bethlehem University, Bethlehem, Palestine.
55. Division of Epilepsy, Department of Neurology, Mayo Clinic, Rochester, MN 55902, USA.
56. Unit of Genetics of Neurodegenerative and Metabolic Diseases, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy.
57. Stanley Center for Psychiatric Research, Broad Institute of Harvard and M.I.T., Cambridge, MA 02142, USA.
58. Hôpital Lariboisière, Dept of Neurosurgery-Paris-Cité University, Paris, France.
59. Department of Epidemiology and Preventive Medicine, School of Public Health, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 6997801, Israel.

60. Department of Pediatric Neurology, Chang Gung Memorial Hospital, Linkou Branch, and College of Medicine, Chang Gung University, Taoyuan, Taiwan.
61. Kids Neuroscience Centre, Kids Research, Children Hospital at Westmead, Sydney, New South Wales, Australia.
62. Neurology Research Group, Swansea University Medical School, Faculty of Medicine, Health & Life Science, Swansea University, SA2 8PP, UK.
63. Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA.
64. Neurology, Massachusetts General Hospital, Boston, MA, USA.
65. Department of Clinical Neurosciences, Cambridge Biomedical Campus, Cambridge CB2 0SL, UK.
66. Department of Neuroscience, Reproductive and Odontostomatological Sciences, University Federico II, Naples 80131, Italy.
67. Division of Neurology, Children's Hospital of Philadelphia, Philadelphia, 3401 Civic Center Blvd, Philadelphia, PA 19104, USA.
68. The Epilepsy NeuroGenetics Initiative (ENGIN), Children's Hospital of Philadelphia, Philadelphia, 3401 Civic Center Blvd, Philadelphia, PA 19104, USA.
69. Department of Neurosciences, Université de Montréal, Montréal, CA 26758, Canada.
70. Department of Neurology, Royal Victoria Hospital, Belfast Health and Social Care Trust, Grosvenor Road, Belfast BT12 6BA, UK.
71. Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki 0014, Finland.
72. Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA.
73. Department of Psychiatry and Behavioral Sciences, Vanderbilt University Medical Center, Nashville, TN, USA.
74. Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA.
75. Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA.
76. Stichting Epilepsie Instellingen Nederland (SEIN), Heemstede 2103 SW, The Netherlands.
77. Department of Neurology, Beaumont Hospital, Dublin D09 FT51, Ireland.
78. Department of Neurology, Hôpital Erasme, Université Libre de Bruxelles, Bruxelles 1070, Belgium.
79. Department of Clinical Neurophysiology, Lille University Medical Center, EA 1046, University of Lille.
80. Department of Neurology, New York University/Langone Health, New York NY, USA.
81. Neurology Department, St. James's Hospital, Dublin D03 VX82, Ireland.
82. Department of Neurology, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, 19104 USA.
83. Department of General Practice and Primary Health Care, University of Helsinki and Helsinki University Hospital, Helsinki 0014, Finland.
84. Human Genetics Training Program, Vanderbilt University, Nashville, TN, USA.
85. Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA.
86. Division of Biostatistics, Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei 100, Taiwan.
87. Department of Pharmacology and Psychiatry, University of Pennsylvania Perlman School of Medicine, Philadelphia, PA 19104, USA.
88. Department of Pediatrics and Neonatology, Medical University of Vienna, Vienna 1090, Austria.
89. Institute of Neurology, Department of Medical and Surgical Sciences, University "Magna Graecia", Catanzaro, Italy.
90. Neurophysiology, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy.
91. Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, University Hospital Schleswig Holstein, Kiel 24105, Germany. .
92. Department of Neurology, NYU School of Medicine, New York City, NY 10003, USA.

93. Department of Pediatric Neuroscience, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy.
94. Institute of Neurology and Neurosurgery at St. Barnabas, Livingston, NJ 07039, USA.
95. Department of Neurology, Division of Clinical Neuroscience, Rikshospitalet Medical Centre, University of Oslo, Oslo, Norway.
96. Department of Neurology, Baylor College of Medicine.
97. Department of Pediatrics, Nationwide Children's Hospital, Columbia, Ohio, USA.
98. Division of Human Genetics, Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.
99. Life and Brain Center, University of Bonn Medical Center, Bonn 53127, Germany.
100. Department of Neurology, University of California, San Francisco, CA 94143, USA.
101. Division of Pharmacotherapy and Experimental Therapeutics, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA.
102. Department of Genetics, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA.
103. Department of Neuropediatrics, University Medical Center Schleswig-Holstein, Christian-Albrechts-University, 24105 Kiel, Germany.
104. Department of Biomedical and Health Informatics (DBHi), Children's Hospital of Philadelphia, Philadelphia, PA, 19104 USA.
105. Hasso Plattner Institute, Digital Health Center, University of Potsdam, Germany.
106. Hasso Plattner Institute, Mount Sinai School of Medicine, NY, US.
107. General Medical Research Center, School of Medicine, Fukuoka University, Japan.
108. Department of Neurology, University Hospital of Strasbourg, Strasbourg, France.
109. Danish Epilepsy Centre, Dianalund 4293, Denmark.
110. Institute of Regional Health Services Research, University of Southern Denmark, Odense 5000, Denmark.
111. Swiss Epilepsy Center, Klinik Lengg, Zurich, Switzerland.
112. National Epilepsy Center, Shizuoka Institute of Epilepsy and Neurological Disorder, Shizuoka, Japan.
113. Department of Pediatrics, Fukuoka Sanno Hospital, Japan.
114. Department of Psychiatry and Applied Psychology, Institute of Mental Health University of Nottingham, Nottingham NG7 2TU, UK.
115. Division of Brain Sciences, Imperial College London, London SW7 2AZ, UK.
116. Kuopio Epilepsy Center, Neurocenter, Kuopio University Hospital, Kuopio 70210, Finland.
117. Institute of Clinical Medicine, University of Eastern Finland, Kuopio 70210, Finland.
118. Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, the University of Tokyo, Tokyo, Japan.
119. The Broad Institute of M.I.T. and Harvard, Cambridge, MA 02142, USA.
120. Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan.
121. Department of Child Neurology, Medical School, Kocaeli University, Kocaeli, Turkey.
122. Neuroscience Unit, KEMRI-Wellcome Trust Research Programme, Kilifi, Kenya.
123. Department of Public Health, Pwani University, Kilifi, Kenya.
124. Department of Psychiatry, University of Oxford, Oxford, UK.
125. Department of Pediatrics, Showa University School of Medicine, Epilepsy Medical Center, Showa University Hospital, 1-5-8 Hatanodai, Shinagawa-ku, Tokyo 142-8555, Japan.
126. Department of Pharmacology and Toxicology, American University of Beirut Faculty of Medicine, Beirut, Lebanon.
127. Department of Paediatrics and Child Health, University of Otago, Wellington, New Zealand.
128. Epilepsy Center Hessen-Marburg, Department of Neurology, Philipps University Marburg, Marburg, Germany.
129. Epilepsy Center Frankfurt Rhine-Main, Center of Neurology and Neurosurgery, Goethe University Frankfurt, Frankfurt, Germany.

130. Departments of Clinical Neurosciences, Medical Genetics and Community Health Sciences, Hotchkiss Brain Institute & Alberta Children's Hospital Research Institute, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada.
131. LOEWE Center for Personalized Translational Epilepsy Research (CePTER), Goethe University Frankfurt, Germany.
132. Neuropediatric Clinic and Clinic for Neurorehabilitation, Epilepsy Center for Children and Adolescents, Vogtareuth, Germany.
133. Research Institute Rehabilitation / Transition, / Palliation, PMU Salzburg, Austria.
134. Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus.
135. Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette L-4362, Luxembourg.
136. Department of Neurology, Medical University of Vienna, Vienna 1090, Austria.
137. Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Bern 3010, Switzerland.
138. Yale School of Medicine, New Haven, CT 06510, USA.
139. Institute of Human Genetics, University of Leipzig Medical Center, Leipzig, Germany.
140. Institute of Experimental Epileptology and Cognition Research, Medical Faculty, University of Bonn, Bonn, Germany.
141. Bonifatius Hospital Lingen, Neuropediatrics Wilhelmstrasse 13, 49808 Lingen, Germany.
142. Department of Neurology, Hofstra-Northwell Medical School, New York, NY, USA.
143. Department of Medicine and Therapeutics, Chinese University of Hong Kong, Hong Kong, China.
144. Department of Biomedical and Dental Sciences, Morphological and Functional Images (BIOMORF), University of Messina, Messina, Italy.
145. The School of Pharmacy and Biomolecular Sciences, RCSI Dublin.
146. Department of Paediatrics and Adolescent Medicine, The University of Hong Kong, Hong Kong.
147. Folkhälsan Research Center and Medical Faculty, University of Helsinki, Helsinki 00290, Finland.
148. Department of Medical Genetics, Hospices Civils de Lyon and University of Lyon, Lyon, France.
149. Translational and Clinical Research Institute, Newcastle University, Newcastle Upon Tyne, UK.
150. Department of Clinical Neurosciences, Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle Upon Tyne, UK.
151. Neuroscience Department, Janssen Research & Development, LLC, 1125 Trenton-Harbourton Road, Titusville, NJ, 08560, USA.
152. Child Neurology, New Children´s Hospital, Helsinki, Finland.
153. Pediatric Research Center, University of Helsinki, Helsinki, Finland.
154. Helsinki University Hospital, Helsinki, Finland.
155. Department of Translational Medicine, School of Medical Sciences, University of Campinas (UNICAMP), and the Brazilian Institute of Neuroscience and Neurotecnology; Campinas, SP, Brazil.
156. Department of Medicine, Tseung Kwan O Hospital, Hong Kong.
157. deCODE genetics Sturlugata 8, IS-102, Reykjavík Iceland.
158. Department of Pharmacology and Therapeutics, University of Liverpool, Liverpool L69 3GL, UK.
159. Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892, USA.
160. School of Medicine, Trinity College Dublin, Dublin 2, Ireland.
161. Epilepsy Center for Children, University Hospital Ruppin-Brandenburg, Brandenburg Medical School, 16816 Neuruppin, Germany.
162. Pediatric Neurology, University of Giessen, Germany.
163. Institute of Human Genetics, University of Bonn Medical Center, Bonn 53127, Germany.
164. University Hospital Cologne, Cologne, Germany.
165. Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan.

166. Department of Neurology, Landspitalinn University Hospital, Reykjavik, Iceland.
167. Istanbul University-Cerrahpaşa, Cerrahpaşa Medical Faculty, Department of Neurology, Istanbul, Turkey.
168. Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge CB2 0AA, UK.
169. Department of Neurology, Morriston Hospital, Swansea Bay University Bay Health Board, Swansea, Wales, UK.
170. Epilepsy Genetics Program, Division of Epilepsy and Clinical Neurophysiology, Department of Neurology, Boston Children's Hospital, Boston, MA, USA.
171. IRCCS Azienda Ospedaliero-Universitaria di Bologna, Medical Genetics Unit, Bologna, Italy.
172. Department of Neurology, Gardner Neuroscience Institute, University of Cincinnati Medical Center, Cincinnati, OH 45220, USA.
173. Department of Neurology, Duke University School of Medicine, Durham, NC 27710, USA.
174. Faculty of Medicine & Health, University of Sydney, Sydney, New South Wales, Australia.
175. Department of Functional Neurology and Epileptology, Hospices Civils de Lyon and University of Lyon, France.
176. Lyon's Neuroscience Research Center, INSERM U1028 / CNRS UMR 5292, Lyon, France.
177. Department of Clinical Neurosciences, Centre Hospitalo-Universitaire Vaudois, Lausanne, Switzerland.
178. Department of Neurology, Charité Universitaetsmedizin Berlin, Campus Virchow-Clinic, Berlin 13353, Germany.
179. Department of Endocrinology, Hospital of The University of Pennsylvania, Philadelphia, PA 19104, USA.
180. Departments of Neurology, Beth Israel Deaconess Medical Center, Massachusetts General Hospital, and Harvard Medical School, Boston, MA 02215, USA.
181. Department of Neurology, Ludwig Maximilians University, Munchen, Germany.
182. Department of Neurology, Royal Children's Hospital, Parkville 3052, Australia.
183. Department of Epileptology, University Hospital Freiburg, Freiburg, Germany.
184. Department of Neurosciences, University of California, San Diego, La Jolla, CA 92037, USA.
185. School of Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK.
186. Rush University Medical Center, Chicago, IL 60612, USA.
187. Department of Neurology, Alan Richens Epilepsy Unit, University Hospital of Wales, Cardiff CF14 4XW, UK.
188. UCL Genetics Institute, University College London, London WC1E 6BT, UK.
189. Aarhus Institute of Advanced Studies (AIAS), Aarhus University, 8000 Aarhus, Denmark.
190. Department of Neurology and Comprehensive Epilepsy Center, Thomas Jefferson University, Philadelphia, PA 19107, USA.
191. Department of Pediatrics, Ohio State University, Columbus, OH, USA.
192. The Research Institute, Nationwide Children's Hospital, Columbus, OH, USA.
193. CWZ Hospital, 6532 SZ Nijmegen, The Netherlands.
194. Department of Neurodevelopmental Disorder Genetics, Institute of Brain Science, Nagoya City University Graduate School of Medical Science, Nagoya, Aichi, Japan.
195. Laboratory for Neurogenetics, RIKEN Center for Brain Science, Wako, Saitama, Japan.
196. Department of Neurology, Washington University School of Medicine, St. Louis, MO 63110, USA.
197. Department of Child Neurology, Istanbul Faculty of Medicine, Istanbul University, Istanbul, Turkey.
198. C. Mondino National Neurological Institute, Pavia 27100, Italy.
199. Department of Neurology, Kaohsiung Chang Gung Memorial Hospital, Kaohsiung, Taiwan.
200. Centre for Medical Genetics, Vilnius University Hospital Santaros Klinikos, Vilnius, Lithuania.
201. Institute of Biomedical Sciences, Faculty of Medicine, Vilnius University, Vilnius, Lithuania.
202. Department of Child Neurology, Medical School, Marmara University, Istanbul, Turkey.

203. Epilepsy Unit, Department of Neurology, Brain and Cognition Research Center - CerCo, CNRS, UMR5549, University Hospital and University of Toulouse, Paul Sabatier University, Toulouse, France.
204. Departments of Neurology and Pediatrics, The Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.
205. Department of Neurology, Admiraal De Ruyter Hospital, Goes 4462, The Netherlands.
206. MRC/Wits Rural Public Health & Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.
207. Department of Neurology and Epileptology, University of Aachen, Aachen 52074, Germany.
208. Applied & Translational Neurogenomics Group, VIB Center for Molecular Neurology, VIB, Antwerp, Belgium.
209. Department of Neurology, Antwerp University Hospital, Edegem 2650, Belgium.
210. Translational Neurosciences, Faculty of Medicine and Health Science, University of Antwerp, Antwerp, Belgium.
211. Department of Neurology, University Hospital and University of Zurich, Zürich, Switzerland.
212. Department of Pediatric Neurology, Vivantes Hospital Neukölln, 12351 Berlin, Germany.
213. Bezmialem Vakif University, Institute of Life Sciences and Biotechnology, Istanbul, Turkey.

Acknowledgements and funding

Some of the data reported in this paper were collected as part of a project undertaken by the International League against Epilepsy (ILAE) and some of the authors are experts selected by the ILAE. Opinions expressed by the authors, however, do not necessarily represent the policy or position of the ILAE.

This study received support from Science Foundation Ireland (SFI) (16/RC/3948), co-funded under the European Regional Development Fund, the Research Unit FOR-2715 of the German Research Foundation (MN: NO755/6-1, NO755/13-1), from Wellcome Trust (grant 084730), European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°279062 (EpiPGX), The Muir Maxwell Trust and the Epilepsy Society, UK. Part of this work was undertaken at University College London Hospitals, which received a proportion of funding from the NIHR Biomedical Research Centres funding scheme. RS and BPCK are supported by an 'Vrienden WKZ' fund 1616091 (MING). SFB and IES are supported by a National Health and Medical Research Council (NHMRC) of Australia Program Grant [1091593]. MB is supported by an NHMRC Investigator grant [APP1195236]. KLO is supported by an Australian Government Research Training Program Scholarship [APP533086] provided by the Australian Commonwealth Government and the University of Melbourne. DLS identified data from people predicted to have epilepsy form the UK Biobank while funded by a Wellcome Clinical PhD Fellowship on the 4ward North program [203914/Z/16/Z]. MRJ was supported by the UKRI MRC Award No: MR/S02638X/1 and by the NIHR Imperial Biomedical Research Centre (BRC). Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil, grant number (2013/07559-3). The funding bodies had no role in the study design, data collection, analysis, and interpretation, or in writing the manuscript.

We thank the Epi25 principal investigators, local staff from individual cohorts, and all of the patients with epilepsy who participated in research studies at local centers for making possible this global collaboration and resource to advance epilepsy genetics research. This work is part of the Centers for Common Disease Genomics (CCDG) program, funded by the National Human Genome Research Institute (NHGRI) The *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, and the National Heart, Lung, and Blood Institute (NHLBI). CCDG-funded Epi25

research activities at the Broad Institute, including genomic data generation in the Broad Genomics Platform, were supported by NHGRI grant UM1 HG008895 (PIs: Eric Lander, Stacey Gabriel, Mark Daly, Sekar Kathiresan). The Genome Sequencing Program efforts were also supported by NHGRI grant 5U01HG009088-02. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank the Stanley Center for Psychiatric Research at the Broad Institute for supporting the genomic data generation efforts as well as aggregation of control samples and cohorts to contribute to the Epi25 GWAS analyses. In particular, the Genomic Psychiatry Cohort controls were genotyped on the GSA-MD v1.0 by the Broad Genomics Platform with funding from NIH grant U01MH105641 and the Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard. The FINRISK controls were part of the FINRISK studies supported by the THL (formerly KTL: National Public Health Institute) through budgetary funds from the government, with additional funding from institutions such as the Academy of Finland, the European Union, ministries and national and international foundations and societies to support specific research purposes. The collection of the Hong Kong Osteoporosis Study (HKOS) control samples was funded by the Bone Health Fund and Research Grants Council – Early Career Scheme (Project number: 27100416). Other control datasets included IBD NIDDK and samples from the Mass General Brigham (MGB) Biobank available from dbGaP under study accession number phs002018.v1.p1.

We want to acknowledge the participants and investigators of FinnGen study. The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie Inc., AstraZeneca UK Ltd, Biogen MA Inc., Bristol Myers Squibb (and Celgene Corporation & Celgene International II Sàrl), Genentech Inc., Merck Sharp & Dohme Corp, Pfizer Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc, Novartis AG, and Boehringer Ingelheim. Following biobanks are acknowledged for delivering biobank samples to FinnGen: Auria Biobank (www.auria.fi/biopankki), THL Biobank (www.thl.fi/biobank), Helsinki Biobank (www.helsinginbiopankki.fi), Biobank Borealis of Northern Finland (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx), Finnish Clinical Biobank Tampere (www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere), Biobank of Eastern Finland (www.ita-suomenbiopankki.fi/en), Central Finland Biobank (www.ksshp.fi/fi-FI/Potilaalle/Biopankki), Finnish Red Cross Blood Service Biobank (www.veripalvelu.fi/verenluovutus/biopankkitoiminta) and Terveystalo Biobank (www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/). All Finnish Biobanks are members of BBMRI.fi infrastructure (www.bbmri.fi). Finnish Biobank Cooperative -FINBB (https://finbb.fi/) is the coordinator of BBMRI-ERIC operations in Finland. The Finnish biobank data can be accessed through the Fingenious® services (https://site.fingenious.fi/en/) managed by FINBB.

Genome-wide meta-analysis of over 29,000 people with epilepsy reveals 26 loci and subtype-specific genetic architecture

International League Against Epilepsy Consortium on Complex Epilepsies

Supplementary Material

Supplementary Tables

Supplementary table 1. Overview of novel cohorts, genotyping array, relevant ethics/IRB approvals, ethnicity and sample sizes. HK: Hong Kong; JPN: Japan. These details for previously published cohorts included in this current manuscript can be found in the supplements of our 2018 study.[1]

Cohort	Genotyping platform	Ethics Committee / IRB	Ethnicity	Phenotype	Cases
Epi25^	Illumina Infinium GSA	Massachusetts General Brigham (formerly Partners) Institutional Review Board (2012P000788)	European	All epilepsy	11544
				GGE	3153
				Focal	4523
				Unclassified	3868
				Control	13121
			European: Finnish	All epilepsy	474
				GGE	91
				Focal	314
				Unclassified	-
				Control	986
			African	All epilepsy	612
				GGE	276
				Focal	230
				Unclassified	106
				Control	3838
			Asian: HK	All epilepsy	839
				GGE	55
				Focal	639
				Unclassified	145
				Control	594
			Asian: JPN	All epilepsy	256
				GGE	63
				Focal	4
				Unclassified	189
				Control	211
Norwegian GenEpA	Human610-Quadv1	Regional committees for medical and health research ethics, Norway (REK) (reference number: S-01271)	European	All epilepsy	201
				Focal	201
				Control	-

Swiss GenEpA	Illumina Human610-Quadv1	Ethics committee of the Canton of Zurich ("Kantonale Ethikkommission, Kanton Zürich")	European	All epilepsy	231
				Focal	231
				Control	259
Janssen Pharmaceuticals	Illumina 1M	NA*	European	All epilepsy	410
				Focal	410
				Control	3016
Austrian GenEpA	Illumina Human CNV 370 duo	Ethics Committee of the Medical University of Vienna	European	All epilepsy	165
				Focal	165
				Control	337

^ Please see Supplementary Information below for detailed descriptions of control cohorts used for Epi25 analyses.
* The Janssen clinical studies were carried out in accordance with the ethical principles outlined in the Declaration of Helsinki, Good Clinical Practices guidelines, and applicable regulatory requirements. The study protocols were approved by the local, regional, or central Institutional Review Board (IRB) or Independent Ethics Committee (IEC) overseeing the numerous clinical sites involved in multi-centre pharmaceutical trials.

Supplementary table 2. Overview of number of cases and controls, stratified by phenotype and ethnicity.

Phenotype	**Sub-phenotype description**	**n**	**EUR**	**ASI**	**AFR**
GGE	Generalized Epilepsy, not otherwise specified, with spike and wave EEG	3352	3024	44	284
	Childhood Absence Epilepsy (CAE)	1072	1049	6	17
	Juvenile Absence Epilepsy (JAE)	671	662	4	5
	Juvenile Myoclonic Epilepsy (JME)	1813	1732	61	20
	GTCS only, with spike and wave EEG	499	485	3	11
	Subtotal	**7407**	**6952**	**118**	**337**
Focal	Focal Epilepsy, not otherwise specified	3981	3688	140	153
	Focal Epilepsy, documented lesion negative	6367	5778	466	123
	Focal Epilepsy, documented hippocampal sclerosis (HS)	1375	1260	107	8
	Focal Epilepsy, documented lesion other than HS	4661	4213	416	32
	Subtotal	**16384**	**14939**	**1129**	**316**
Unclassified	Epilepsy, not otherwise specified	6153	5668	379	106
	Cases	**29944**	**27559**	**1626**	**759**
	Controls	**52538**	**42436**	**3680**	**6422**
	Total subjects	**82482**	**69995**	**5306**	**7181**

Supplementary table 3. Summary of all genome-wide significant loci including genomic position and independent significant SNPs. We defined the locus position as the region encompassing all SNPs with P<10-4 that were in LD (R^2>0.2) with the lead SNP.

Phenotype	**Locus name**	**Locus position (hg19)**	**Lead SNP**	**Number of independent significant SNPs**	**Independent significant SNPs**
All epilepsy	2p16.1	chr2:57917222-58505679	rs13032423	1	rs13032423
	2q24.3	chr2:166716305-167124221	rs59237858	2	rs59237858; rs1960242
	9q21.13	chr9:76297313-76625089	rs4744696	1	rs4744696
	10q24.32	chr10:103493226-103989812	rs3740422	1	rs3740422
GGE	1q43	chr1:237846053-237908911	rs876793	1	rs876793

	2p16.1	chr2:57917222-58756729	rs11688767	3	rs11688767; rs77876353; rs13416557
	2q12.1	chr2:104056769-104481325	rs62151809	1	rs62151809
	2q24.3	chr2:166818404-166994996	rs11890028	1	rs11890028
	2q32.2	chr2:191504467-191710069	rs6721964	1	rs6721964
	3p22.3	chr3:36218075-36345769	rs9861238	1	rs9861238
	3p21.31	chr3:50184538-50421081	rs739431	1	rs739431
	4p15.1	chr4:31107765-31204950	rs1463849	1	rs1463849
	5q22.3	chr5:113837198-114440966	rs4596374	1	rs4596374
	5q31.2	chr5:136459562-136684519	rs2905552	1	rs2905552
	6q22.33	chr6:128302874-128333682	rs13219424	1	rs13219424
	7p14.1	chr7:41334517-41411165	rs37276	1	rs37276
	9q21.32	chr9:86320233-86694759	rs2780103	1	rs2780103
	10q24.32	chr10:103493226-103989812	rs11191156	1	rs11191156
	12q13.13	chr12:52319584-52348259	rs114131287	2	rs4762030; rs10431492
	16p13.3	chr16:7285674-7442293	rs62014006	1	rs62014006
	17p13.1	chr17:8036060-8219478	rs2585398	1	rs2585398
	17q21.32	chr17:45938105-46554456	rs16955463	1	rs16955463
	19p13.3	chr19:2102543-2136680	rs75483641	1	rs75483641
	21q21.1	chr21:21655062-21719113	rs1487946	1	rs1487946
	21q22.1	chr21:32036541-32203274	rs7277479	1	rs7277479
	22q13.32	chr22:48615721-48639993	rs469999	1	rs469999
CAE	2p16.1	chr2:57942325-58484172	rs12185644	1	rs12185644
JME	4p12	chr4:46250605-46397617	rs17537141	1	rs17537141
	8q23.1	chr8:109733213-109922163	rs3019359	1	rs3019359
	16p11.2	chr16:30603521-31275374	rs1046276	1	rs1046276

*Supplementary table 4. Results from ASSET pleiotropy analyses for the 4 all epilepsy loci. The associated phenotype/s in ASSET reflect the phenotypes driving the ASSET signal, which could be both GGE individually. *Evidence for pleiotropy between GGE and Focal epilepsy.*

SNP (Risk allele)	Chr.	Locus	P-value (GGE)	P-value (FE)	OR (ASSET)	P-value (ASSET)	Associated phenotype in ASSET
*rs60055328(C)	2	2q24.3	1.04e-7	9.62e-7	1.07	2.8e-10	GGE, FE
*rs4744696(G)	9	9q21.13	3.07e-7	8.63e-5	0.93	4.4e-8	GGE, FE
rs13032423(G)	2	2p16.1	2.88e-17	2.93e-3	0.85	8.9e-17	GGE only
rs3740422(G)	10	10q24.32	1.02e-13	4.07e-3	1.15	2.48e-12	GGE only

Supplementary table 5. Heritability enrichment of 26 functional categories, as assessed with LDAK heritability enrichment analyses. Statistical significance (in bold) is defined as P<0.05/26=0.0019.

Annotation	Share	SD	Expected	Enrichment	SD	Z-score	P-value
Coding_UCSC	0.069948	0.027717	0.016079	4.350169	1.723782	1.943499	0.051956
Conserved_LindbladToh	0.132451	0.042109	0.028542	4.640618	1.475346	2.467637	0.013601
CTCF_Hoffman	0.017421	0.041158	0.023919	0.728312	1.720706	0.157893	0.874541
DGF_ENCODE	0.166208	0.095467	0.138091	1.203615	0.691336	0.294524	0.768358
DHS_Trynka	0.139003	0.100726	0.167956	0.827614	0.599718	0.287445	0.773772
Enhancer_Andersson	-0.00046	0.018624	0.004432	-0.104203	4.202414	0.262754	0.79274
Enhancer_Hoffman	0.065675	0.041653	0.042964	1.528598	0.969479	0.545239	0.585589
FetalDHS_Trynka	0.124737	0.077198	0.085617	1.456925	0.901664	0.506758	0.612325
H3K27ac_Hnisz	0.466327	0.030543	0.393028	1.186496	0.077713	2.399804	0.016404
H3K27ac_PGC2	0.350543	0.056076	0.272914	1.284446	0.20547	1.384368	0.166246
H3K4me1_Trynka	0.578187	0.066915	0.428916	1.34802	0.156009	2.230769	0.025696
H3K4me3_Trynka	0.229663	0.052572	0.137162	1.674388	0.38328	1.759518	0.07849
H3K9ac_Trynka	0.253873	0.05353	0.129088	1.966671	0.414683	2.331108	0.019748
Intron_UCSC	0.45682	0.026361	0.394342	1.158437	0.066848	2.370108	0.017783
PromoterFlanking_Hoffman	0.021395	0.026356	0.008596	2.488946	3.066092	0.485617	0.627239
Promoter_UCSC	0.083972	0.035858	0.048118	1.74512	0.745213	0.999875	0.317371
Repressed_Hoffman	0.41399	0.06557	0.452761	0.914367	0.144823	0.591294	0.554323
SuperEnhancer_Hnisz	0.242302	0.019093	0.169819	1.42682	0.112434	3.796183	**0.000147**
TFBS_ENCODE	0.225196	0.07818	0.133056	1.692483	0.58757	1.178554	0.238576
Transcr_Hoffman	0.420967	0.057675	0.35335	1.191361	0.163225	1.172376	0.241046
TSS_Hoffman	0.046568	0.02996	0.018755	2.482921	1.5974	0.928334	0.353234
UTR_3_UCSC	0.013363	0.017055	0.011944	1.118787	1.427882	0.083191	0.9337
UTR_5_UCSC	0.02092	0.016217	0.005928	3.529243	2.735834	0.924487	0.355233
WeakEnhancer_Hoffman	-0.03501	0.039502	0.021357	-1.63943	1.849623	1.42701	0.153577
Super_Enhancer_Vahedi	0.029165	0.008124	0.021624	1.348773	0.375709	0.928306	0.353249
Typical_Enhancer_Vahedi	0.026998	0.011727	0.022194	1.216444	0.528371	0.409644	0.682067

Supplementary table 6. Estimation of inflation factor and the LD-score regression intercept stratified by phenotype. λ: genomic inflation factor, λ_{1000}: genomic inflation factor corrected for an equivalent study of 1000 cases and 1000 controls.

Phenotype	λ	λ_{1000}	Mean χ2	LDSC intercept
All epilepsy	1.25	1.01	1.27	1.10
Focal epilepsy	1.17	1.01	1.17	1.10
GGE	1.26	1.02	1.35	1.04

Supplementary table 7. SNP-based heritabilities as calculated by LDAK, with the BLD-LADK model. Observed-scale heritability is calculated using effective-sample sizes, after which it was converted to liability-scale heritability using the same prevalence estimates as our previous GWAS.[1]

Phenotype	cases	controls	K: prevalence	Z	Observed-scale heritability	Liability scale heritability
All epilepsy	27559	42436	0.005	0.0145	0.3733	0.177 (0.155 - 0.199)
Focal epilepsy	14939	42436	0.003	0.0091	0.3733	0.160 (0.140 - 0.180)
GGE	6952	42436	0.002	0.0063	0.9955	0.395 (0.343 - 0.446)

JME	1728	37339	0.00035	0.0013	2.1135	0.635 (0.510 - 0.760)
JAE	662	37339	0.00015	0.0006	3.3528	0.900 (0.633 - 1.166)
CAE	1049	37339	0.00015	0.0006	3.0427	0.816 (0.638 - 0.995)
GTCSA	485	37339	0.0002	0.0008	1.7824	0.496 (0.140 - 0.853)
Focal HS	1260	37339	0.00075	0.0026	1.4020	0.472 (0.294 - 0.649)
Focal other lesion	4213	37339	0.00135	0.0044	0.6778	0.251 (0.188 - 0.313)
Focal non-lesional	5778	37339	0.0009	0.0031	0.2452	0.085 (0.046 -0.124)

Supplementary table 8. Top 20 drugs that are licensed for conditions other than epilepsy, but are predicted to be efficacious for GGE, and have published evidence of antiseizure efficacy from multiple published studies and in multiple animal models. We do not advise immediate use of these drugs for people with epilepsy, prior to any clinical trials. AUD: audiogenic; electro: maximal electroshock; Kin: kindling; PTZ: pentylenetetrazol. Drugs are listed in alphabetical order.

Drug	Current indication	Studies' PubMed IDs	Models
Aspirin	Pain; pyrexia; antiplatelet	11883156, 14671677, 16844276, 22765917, 28060522	Pilo, PTZ, Electro
Biperiden	Parkinson's disease	738231, 2858579	Electro, other
Captopril	Hypertension; chronic heart failure; diabetic nephropathy	2824310, 22107891, 25573423	PTZ, AUD, Other
Citalopram	Depression; panic disorder	21531632, 21962757, 22429158, 22578701	KA, PILO, PTZ
Dapsone	Leprosy; dermatitis herpetiformis	1817960, 7970237, 23729301	KA, Kin
Dextromethorphan	Pain; addiction; cough	1456842, 2079649, 2574061, 2666123, 2676564, 2806362, 3044591, 3374269, 3380326, 3768695, 8058587, 8094234, 8405092, 8856734, 9179861, 9187330, 10080248, 11182165, 12479976, 12586225, 15084442, 15723099	KA, PTZ, Electro, AUD, Kin, other
Diltiazem	Angina; hypertension	2272645, 7681002, 8152336, 22661180	KA, PTZ, Electro
Doxepin	Depression; pruritus	1456842, 19443935	PTZ, Electro, other
Fluoxetine	Depression; bulimia nervosa; obsessive-compulsive disorder	7999524, 8149989, 8384110, 8538363, 8816259, 9696406, 15680343, 16531634, 17215106, 23530452, 25754610	Electro, AUD, other
Isradipine	Hypertension	8118482, 9595291	Electro, AUD
Lovastatin	Hypercholesterolaemia	21224519, 23253428, 23352156	KA, AUD
Nicardipine	Angina; hypertension	7681002, 8152336, 8872866, 10608279, 11742591	KA, PTZ, Kin, other
Nifedipine	Angina; hypertension; Raynaud's phenomenon; premature labour	1628595, 1698518, 1747472, 1865996, 1946038, 2085727, 2272645, 2713089, 2744396, 7681002, 7694769, 8054599, 8118482, 8152336, 8474621, 8707372, 12126870, 12536054, 16573711, 20113637, 22661180, 22801414	KA, PTZ, Electro, AUD, Kin, other
Nimodipine	Subarachnoid haemorrhage	1628595, 1698518, 2272645, 2310938, 2463174, 2662221, 3784769, 7681002, 8152336, 8156970, 8156971, 8707372, 9389584, 9570719, 9689485, 10683952, 12372903, 12536054, 12539272, 15123017, 17193898, 17344939, 19761108, 23761887, 25225705, 25445375	KA, Pilo, PTZ, Electro, AUD, Kin, other
Orphenadrine	Parkinsonism	2624511, 19815957	PTZ, Electro
Pimozide	Schizophrenia	2272645, 6141554, 7875556	PTZ, Electro, AUD, other

Pioglitazone	Diabetes mellitus	20599832, 22436324, 27527983	PTZ, other
Quetiapine	Schizophrenia; mania; depression	21168466, 26188240	PTZ, AUD
Tamoxifen	Breast cancer; anovulatory infertility	12139106, 24903749	Electro, Kin
Thalidomide	Malignant disease; immunosuppression	17449064, 21592729, 24735834	PTZ, Kin

Supplementary table 9. Genetic correlations between our main GWAS and Biobank GWAS (including deCODE genetics). P-values are shown, with standard errors in brackets.

	Primary All epilepsy	**Primary Focal**	**Primary GGE**
Biobank All epilepsy	0.74 (0.106)	0.5525 (0.1781)	0.7036 (0.0879)
Biobank Focal	0.5835 (0.1596)	0.7637 (0.2505)	0.4331 (0.1275)
Biobank Gen	0.6231 (0.1434)	0.307 (0.2176)	0.6521 (0.1373)

Supplementary table 10. Sample sizes of the included Biobanks and deCODE genetics.

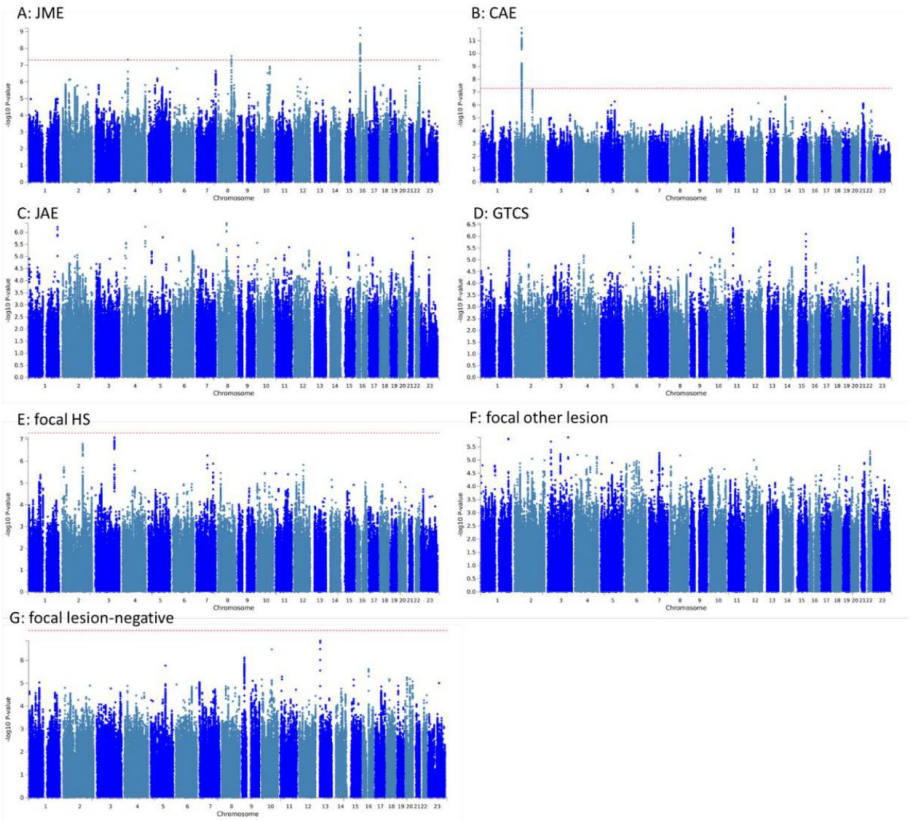
Cohort	**All epilepsy**	**Focal**	**GGE**	**Controls**
UK Biobank	7,006	-	-	179,763
Japan Biobank	612	145	283	176,694
DECODE genetics	3,762	405	1,342	335,389
FinnGen	10,354	5,922	1,160	332,143
Total	21,734	6,472	2,785	1,023,989

Supplementary table 11. Phenotypes and associated publications assessed for genetic correlations with epilepsy using LDSC.

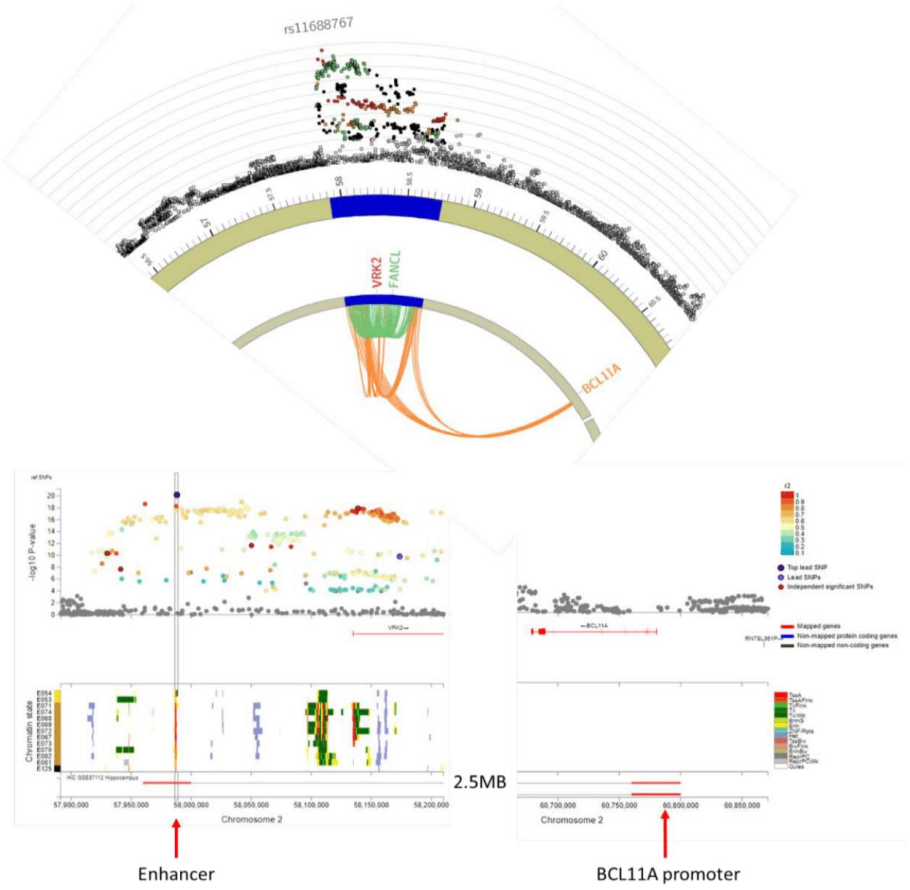
Broad trait	**Trait**	**Publication**	**Notes**
Psychiatric	Bipolar disorder	Mullins *et al* 2021[2]	
Psychiatric	ADHD	Demontis *et al* 2019[3]	
Psychiatric	ASD	Grove *et al* 2019[4]	
Psychiatric	Schizophrenia	Trubetskoy *et al* 2022[5]	
Psychiatric	Depression	Howard *et al* 2019[6]	exc. UKBB and 23andMe
Neurological	Febrile seizures	Skotte *et al* 2022[7]	
Neurological	Parkinson's disease	Nalls *et al* 2019[8]	exc. 23andMe
Neurological	Alzheimer's disease	Wightman *et al* 2021[9]	exc. 23andMe
Neurological	Stroke	Malik *et al* 2018[10]	
Neurological	Headache	Meng *et al* 2018[11]	
Neurological / Autoimmune	Multiple sclerosis	International Multiple Sclerosis Genetics Consortium 2019[12]	
Autoimmune	Type 1 diabetes	Chiou *et al* 2021[13]	
Autoimmune	Systemic lupus erythematosus	Morris *et al* 2016[14]	
Cognitive	Intelligence	Savage, Jansen *et al* 2018[15]	
Sleep	Insomnia	Jansen *et al* 2019[16]	exc. 23andMe
Smoking	Ever smoked	Karlsson Linnér *et al* 2019[17]	

Metabolic	Type 2 diabetes	Mahajan *et al* 2018[18]	
Metabolic	Coronary disease	van der Harst, Verweij *et al* 2018[19]	

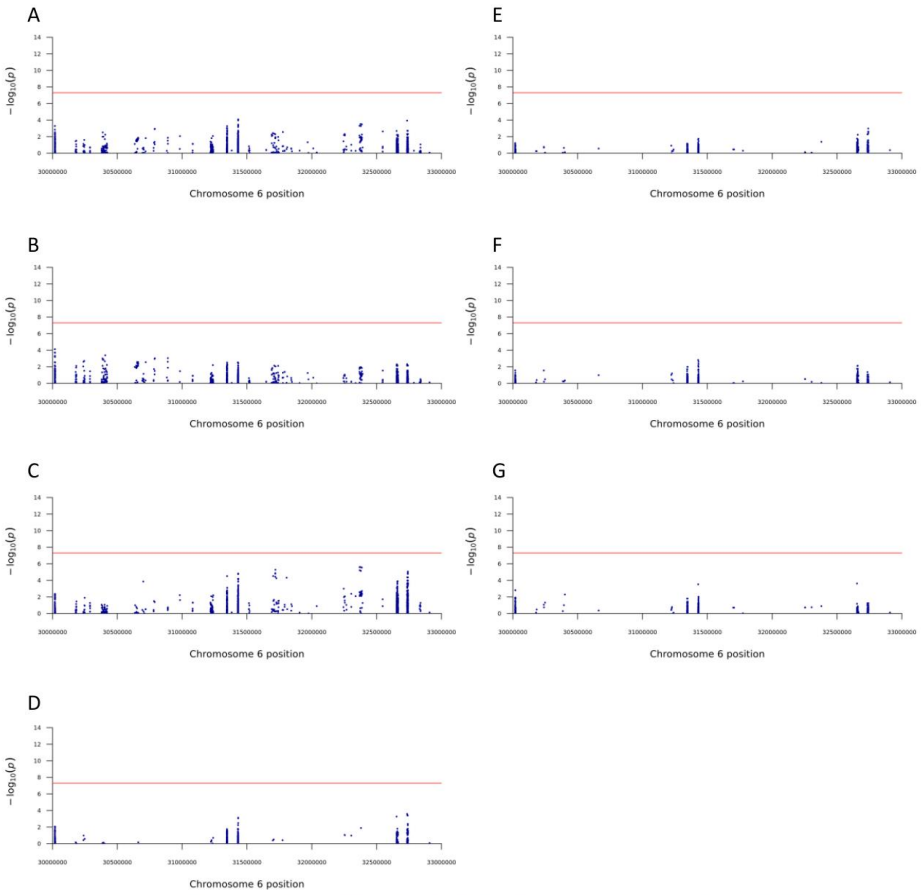
Supplementary Figures



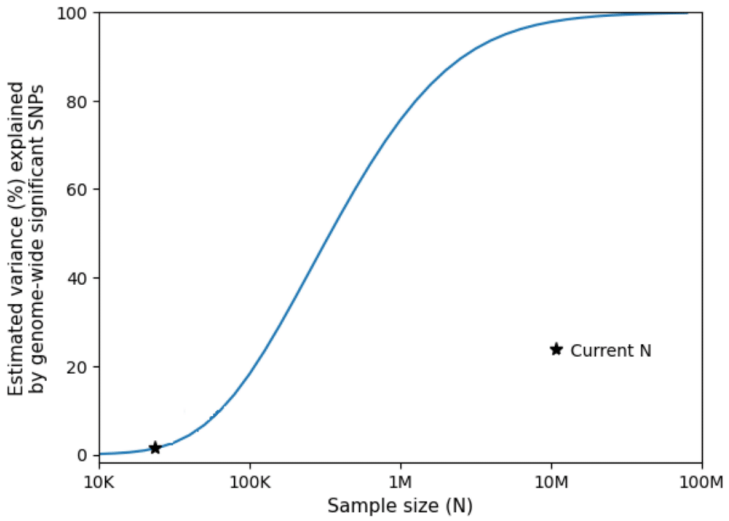
Supplementary figure 1. Manhattan plots of epilepsy subphenotype GWAS. Chromosomal position is plotted on the X-axis and -log10 transformed P-values are plotted on the Y-axis. A. juvenile myoclonic epilepsy (JME); B. childhood absence epilepsy (CAE); C. juvenile absence epilepsy (JAE); D. generalized tonic-clonic seizures alone (GTCS); E. focal epilepsy due to hippocampal sclerosis (focal HS); F. focal epilepsy with other lesion; G. lesion negative focal epilepsy.



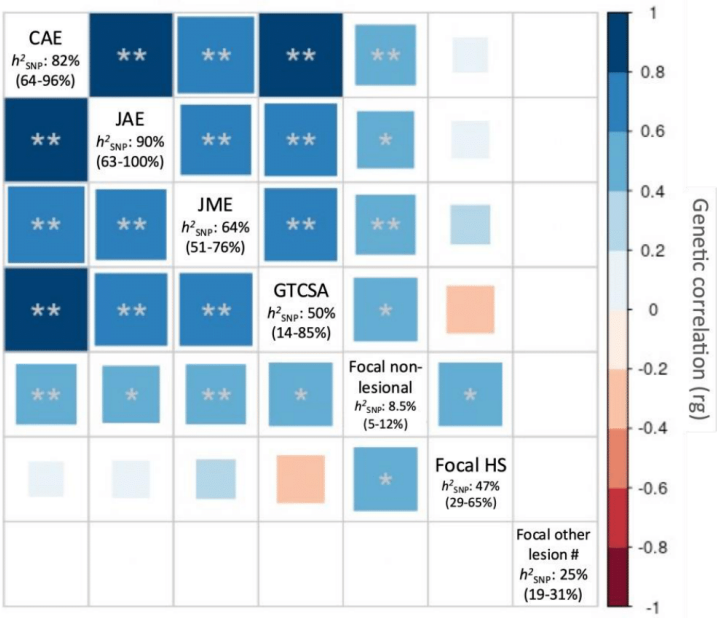
Supplementary figure 2. 3D chromatin interactions link the 2p16.1 locus with the promoter region of BCL11A. The upper circus plot shows the 2p16.1 locus with GWAS P-values in the outer ring, with eQTL associations in green and HiC 3D chromatin interactions in orange. The locuszoom below shows GWAS P-values with chromatin states and Hi-C chromatin interactions below.



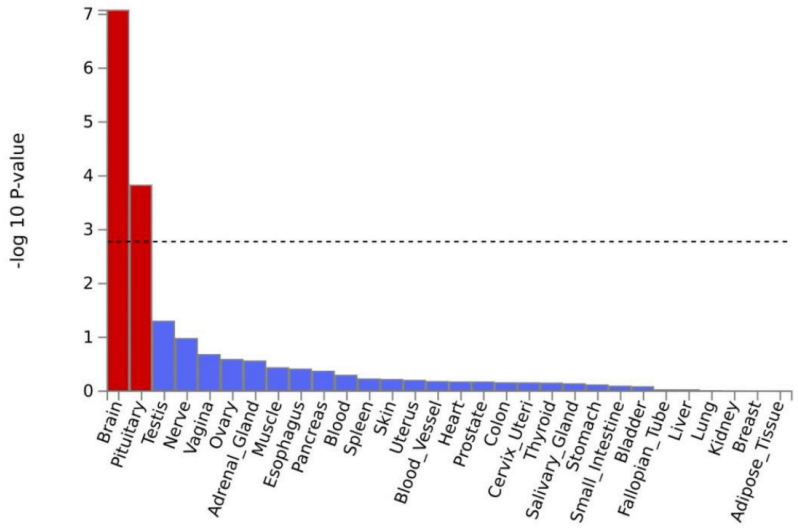
Supplementary figure 3. Manhattan plots of HLA analysis for A) All Epilepsy, B) Focal Epilepsy, C) GGE, D) JME, E) Focal lesion negative, F) Focal due to other lesion, G) Focal HS.



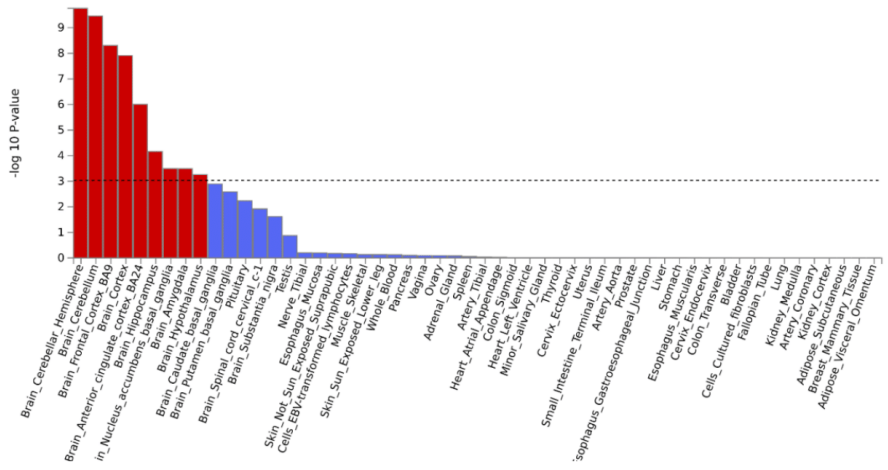
Supplementary figure 4. Power analysis for GGE, using the MiXeR causal mixture model.[20] The X-axis shows the current and required sample size, and the Y-axis shows the corresponding explained variance by genome-wide significant SNPs at these sample sizes. An explained variance of 100% corresponds to the identification of all SNPs that underlie GGE SNP-based heritability.



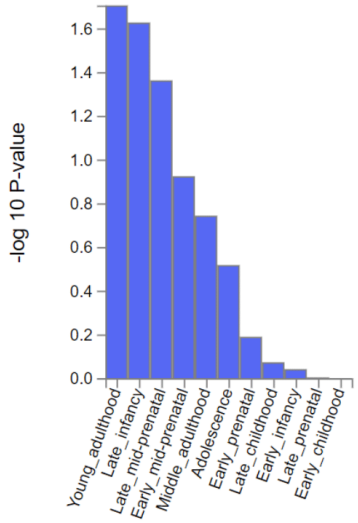
*Supplementary figure 5. Heritability estimates and genetic correlations between epilepsy syndromes. The genetic correlation coefficient was calculated with LDSC and is denoted by color scale from -1 (red) to +1 (blue). # rg out of bounds due to phenotype not reaching significant heritability; * P < 0.05, ** P < 0.0024 (Bonferroni correction).*



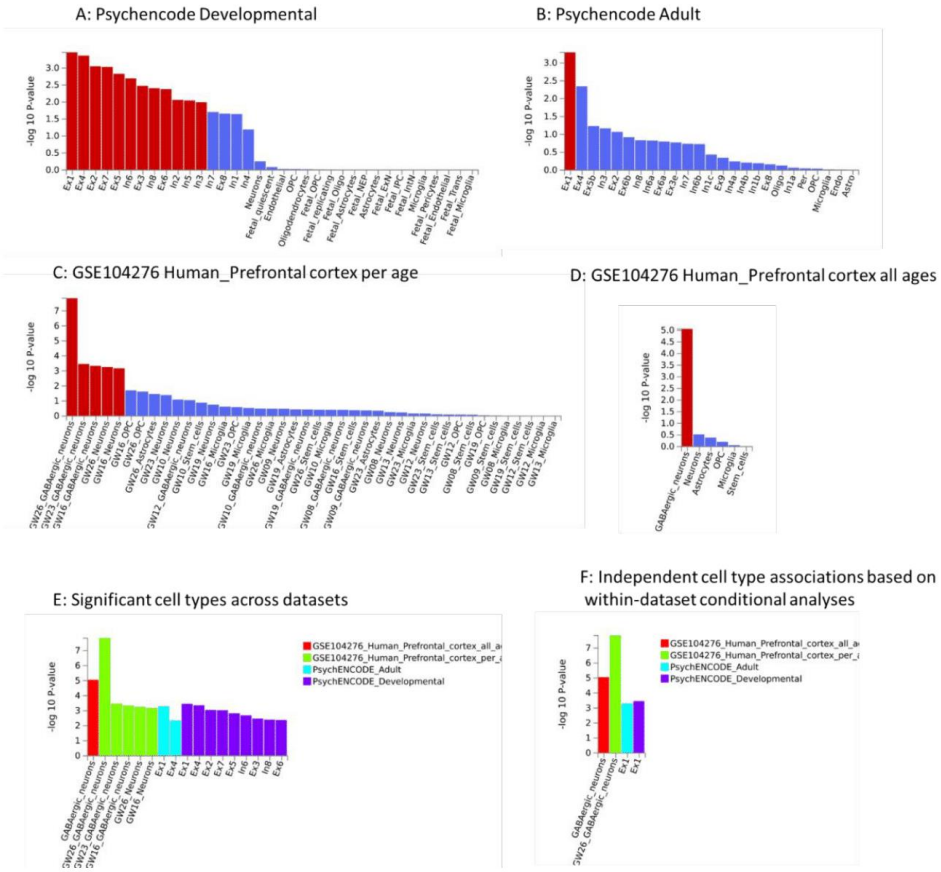
Supplementary figure 6. Tissue-type enrichment of broad tissue types, as calculated with MAGMA,[21] using data from the Gene-Tissue Expression consortium (GTEx). The dotted line represents the significance threshold.



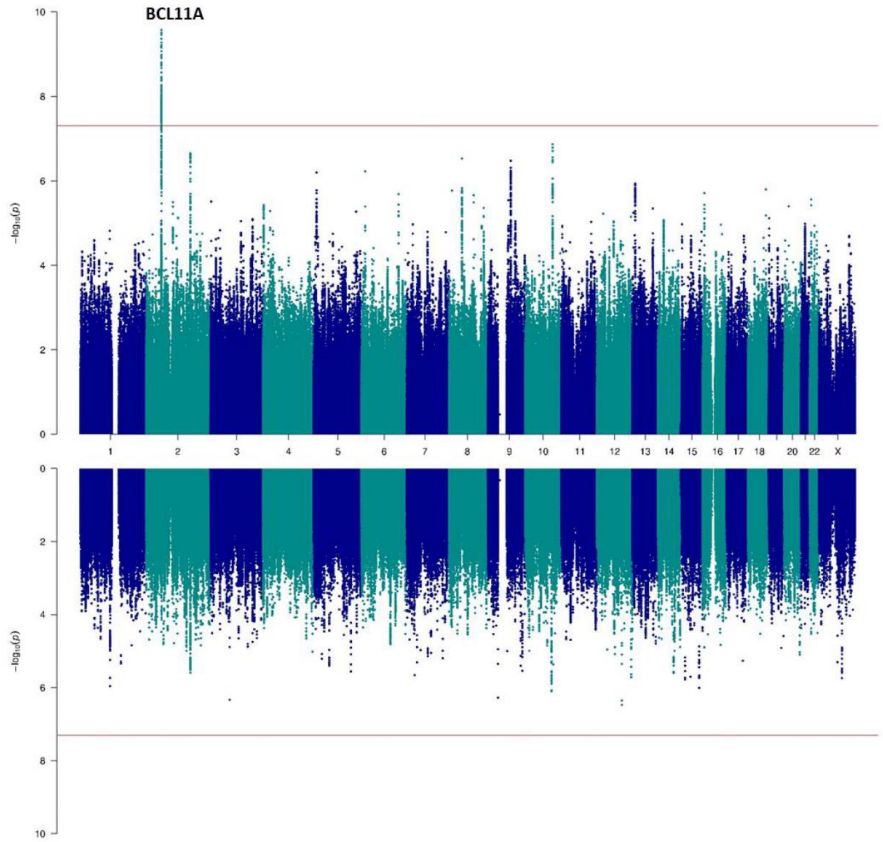
Supplementary figure 7. Tissue-type enrichment of 54 tissues, including specific brain regions, as calculated with MAGMA[21] using data from GTEx. The dotted line represents the significance threshold.



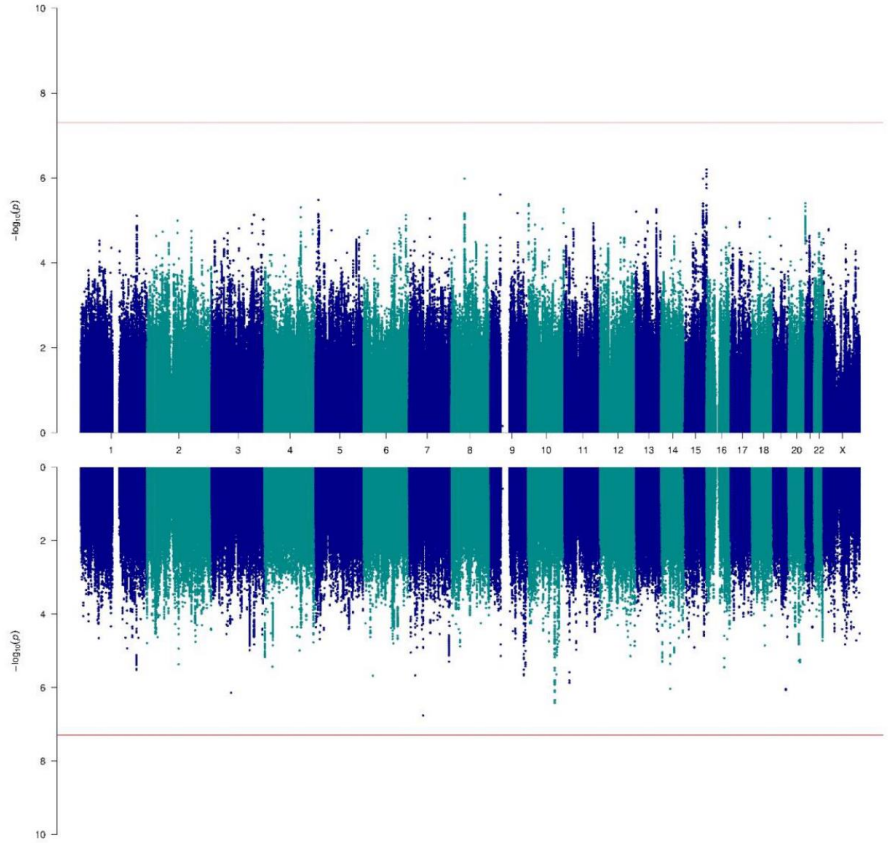
Supplementary figure 8. Enrichment of genes expressed in the brain at 11 general developmental stages, as calculated with MAGMA,[21] using data from the BrainSpan consortium. The dotted line represents the significance threshold.



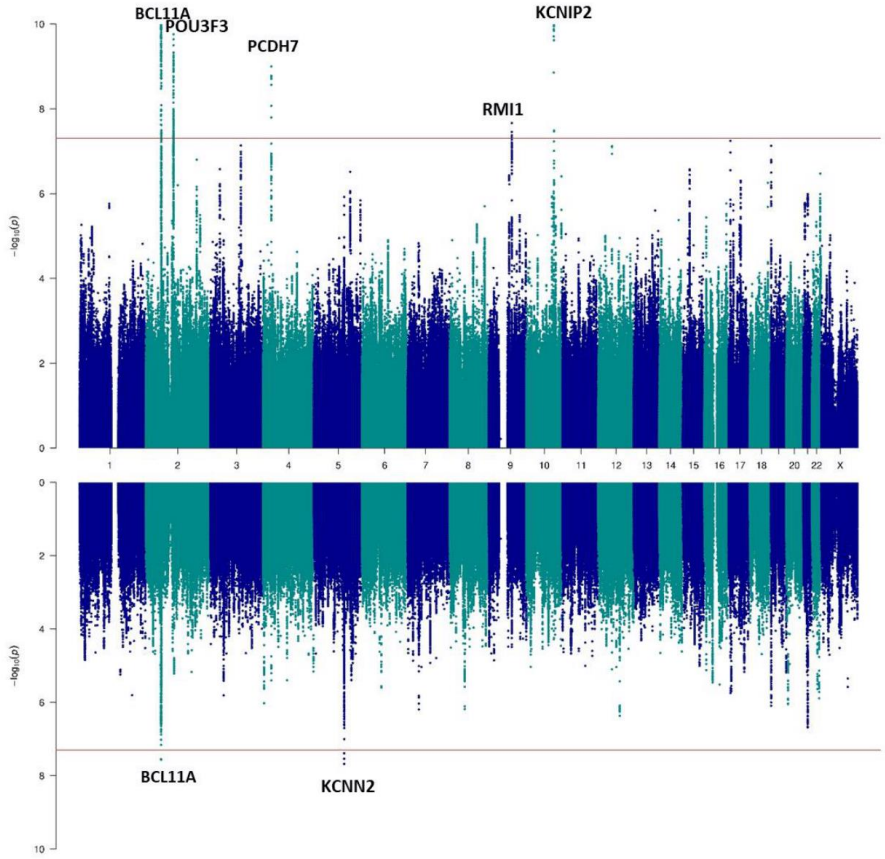
Supplementary figure 9. Cell-type enrichment analyses across datasets, as calculated with FUMA.[22] Two different single-cell RNA sequencing datasets of human adult and developmental brain cells were assessed. Results from individual datasets are displayed in A-D with significant associations (after FDR correction) in red. Significant cell types across datasets are displayed in E, and significant cell-types after within dataset conditional analyses are displayed in F. Ex: excitatory neuron; In: inhibitory neuron.



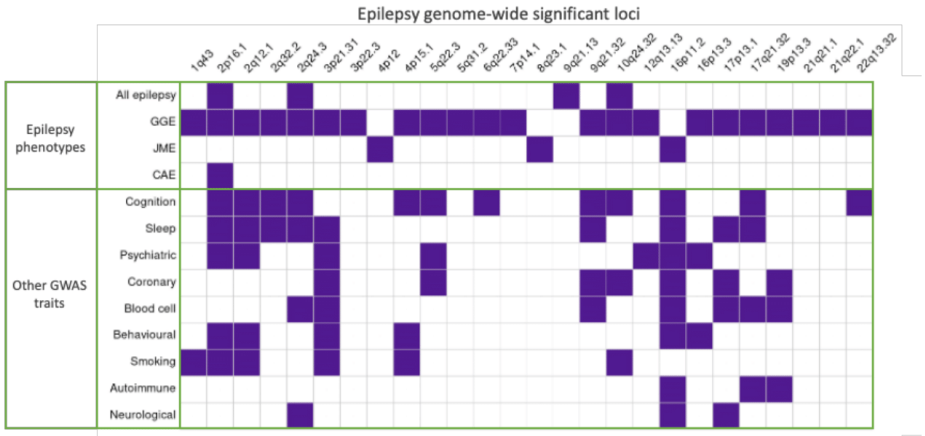
Supplementary figure 10. Sex-specific GWAS of all epilepsy. The female-only is displayed at the top (n=13889 cases and 19676 controls) and male-only GWAS is displayed at the bottom (n=12259 cases and 18645 controls). We annotated genes that were implicated by our gene prioritization analyses.



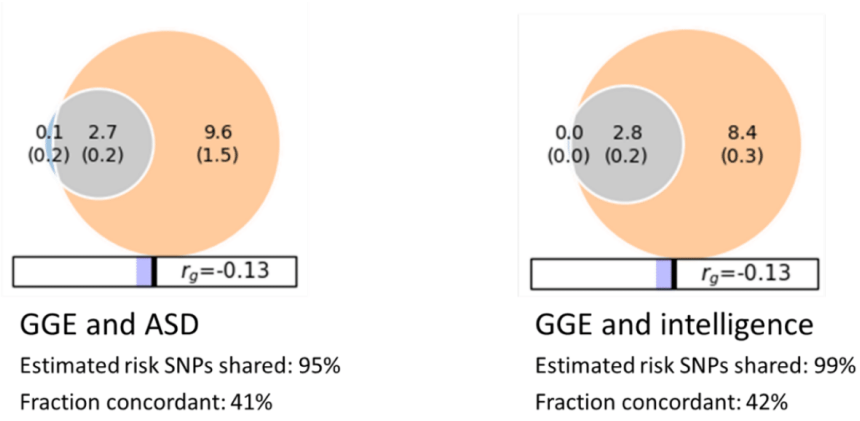
Supplementary figure 11. Sex-specific GWAS of focal epilepsy. The female-only is displayed at the top (n=7175 cases and 19676 controls) and male-only GWAS is displayed at the bottom (n=6756 cases and 18645 controls).



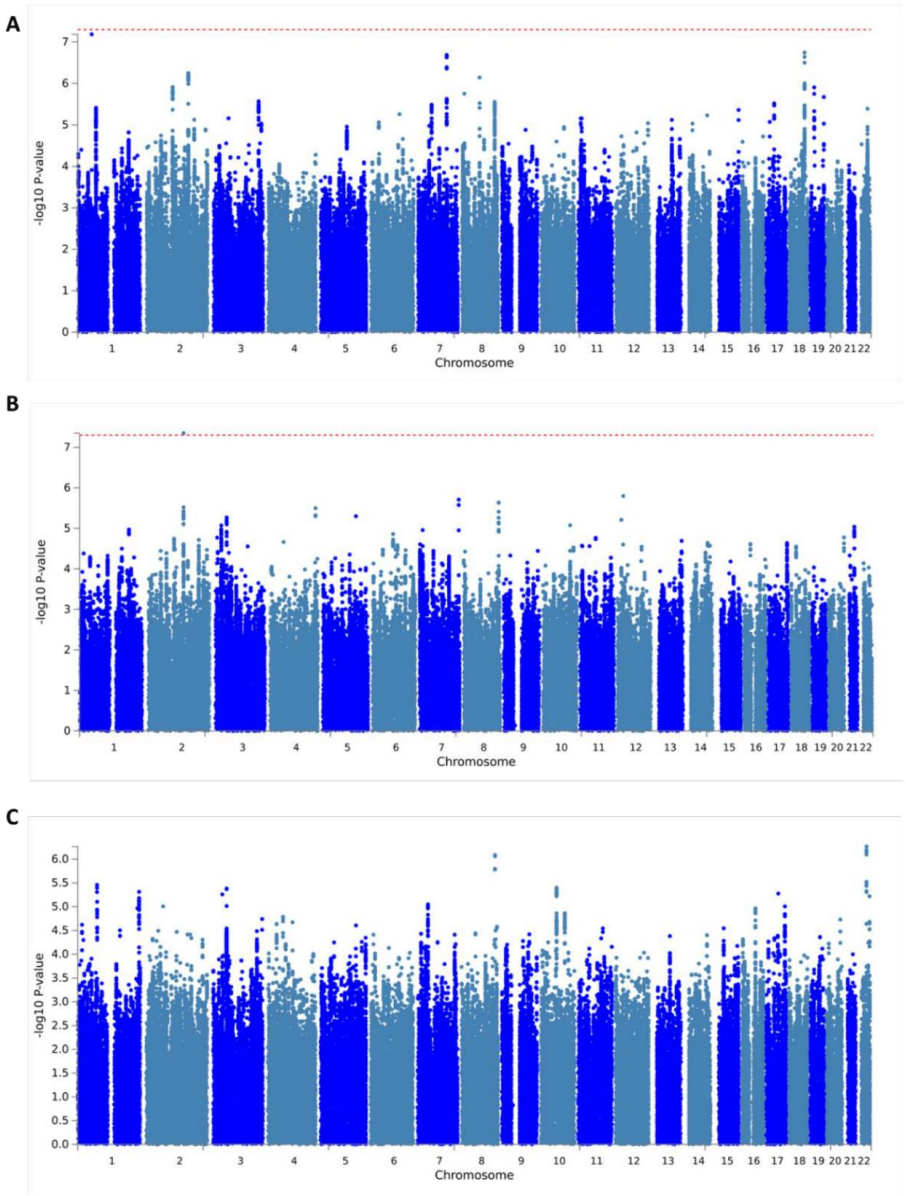
Supplementary figure 12. Sex-specific GWAS of GGE. The female-only is displayed at the top (n=3946 cases and 19676 controls) and male-only GWAS is displayed at the bottom (n=2603 cases and 18645 controls). We annotated genes that were implicated by our gene prioritization analyses.



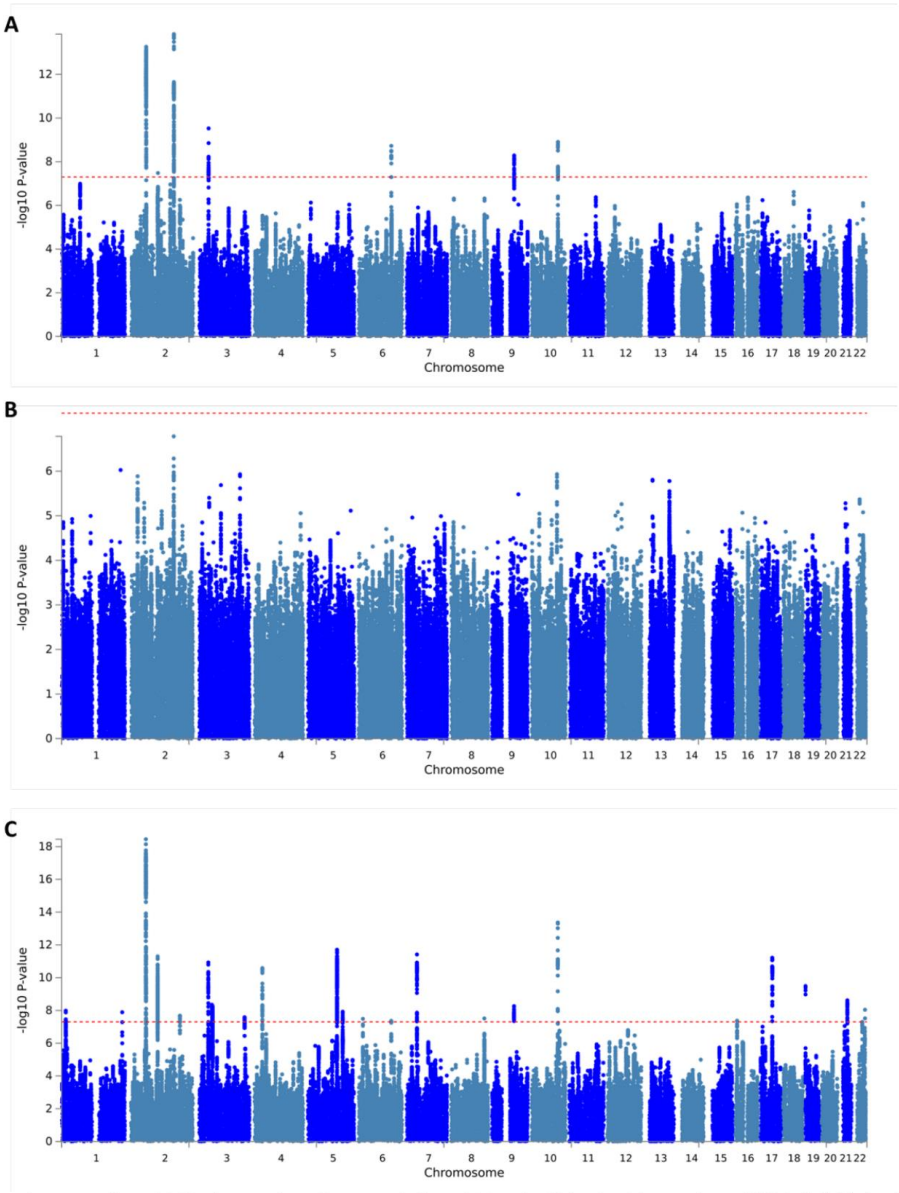
*Supplementary figure 13. GWAS traits each of the epilepsy genome-wide significant loci have been associated with indicated by a purple cell. Prior trait associations were determined by a p < 5*10-8 GWAS Catalog entry for the same SNP, or SNPs in high LD, as those reported in the epilepsy analysis.*



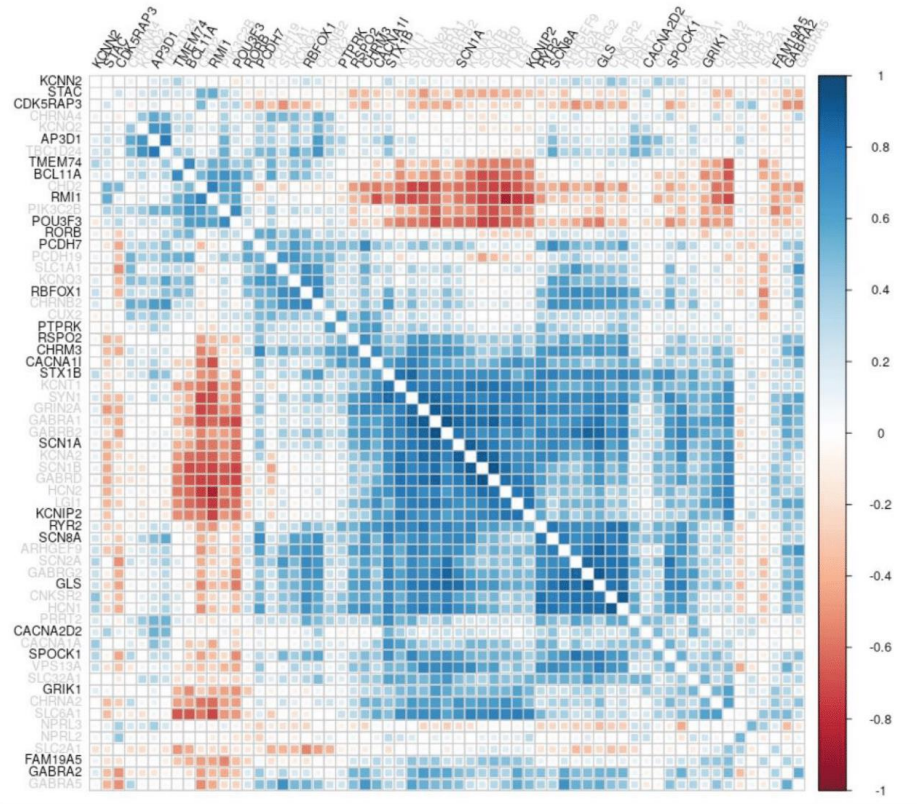
Supplementary figure 14. Bivariate MiXeR analyses[20] showing the fraction of causal SNPs that are unique to GGE (blue), and shared (grey) and unique to ASD (left) or intelligence (right). Rg: genetic correlation coefficient.



Supplementary figure 15. Manhattan plots of Biobank-only GWAS of all (A), focal (B) and GGE (C). Chromosomal position is plotted on the X-axis and -log10 transformed P-values are plotted on the Y-axis.



Supplementary figure 16. Manhattan plots of meta-analysis combining the Biobanks with our primary GWAS of all (A), focal (B) and GGE (C). Chromosomal position is plotted on the X-axis and -log10 transformed P-values are plotted on the Y-axis.



Supplementary figure 17. Gene co-expression matrix produced by brain-coX[23] for known (grey) and candidate (black) epilepsy genes.

Supplementary References

1. International League Against Epilepsy Consortium on Complex Epilepsies. Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. Nat Commun 2018;9:5269.
2. Mullins N, Forstner AJ, O'Connell KS, et al. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. Nat Genet 2021;53:817-829.
3. Demontis D, Walters RK, Martin J, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat Genet 2019;51:63-75.
4. Grove J, Ripke S, Als TD, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet 2019;51:431-444.
5. Trubetskoy V, Pardinas AF, Qi T, et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature 2022;604:502-508.
6. Howard DM, Adams MJ, Clarke TK, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci 2019;22:343-352.
7. Skotte L, Fadista J, Bybjerg-Grauholm J, et al. Genome-wide association study of febrile seizures implicates fever response and neuronal excitability genes. Brain 2022;145:555-568.
8. Nalls MA, Blauwendraat C, Vallerga CL, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol 2019;18:1091-1102.
9. Wightman DP, Jansen IE, Savage JE, et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. Nat Genet 2021;53:1276-1282.
10. Malik R, Chauhan G, Traylor M, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat Genet 2018;50:524-537.
11. Meng W, Adams MJ, Hebert HL, Deary IJ, McIntosh AM, Smith BH. A Genome-Wide Association Study Finds Genetic Associations with Broadly-Defined Headache in UK Biobank (N=223,773). EBioMedicine 2018;28:180-186.
12. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science 2019;365.
13. Chiou J, Geusz RJ, Okino ML, et al. Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. Nature 2021;594:398-402.
14. Morris DL, Sheng Y, Zhang Y, et al. Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. Nat Genet 2016;48:940-946.
15. Savage JE, Jansen PR, Stringer S, et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. Nat Genet 2018;50:912-919.
16. Jansen PR, Watanabe K, Stringer S, et al. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. Nat Genet 2019;51:394-403.
17. Karlsson Linner R, Biroli P, Kong E, et al. Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. Nat Genet 2019;51:245-257.
18. Mahajan A, Taliun D, Thurner M, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet 2018;50:1505-1513.
19. van der Harst P, Verweij N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. Circ Res 2018;122:433-443.
20. Frei O, Holland D, Smeland OB, et al. Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation. Nat Commun 2019;10:2417.
21. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol 2015;11:e1004219.
22. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun 2017;8:1826.
23. Freytag S, Burgess R, Oliver KL, Bahlo M. brain-coX: investigating and visualising gene co-expression in seven human brain transcriptomic datasets. Genome Med 2017;9:55.

Supplementary information

Summary of external control datasets used in the Epi25 GWAS

FINRISK controls

Description:
The controls from FINRISK that contributed to the Epi25 GWAS study were part of the FINRISK inflammatory bowel disease (IBD) cohort. The population-based FINRISK study has been followed up for IBD and other disease end-points using annual record linkage with the Finnish National Hospital Discharge Register, the National Causes-of-Death Register and the National Drug Reimbursement Register. Controls were chosen to have a high polygenic risk score for IBD without an IBD diagnosis. A detailed description of the FINRISK cohort can be found at Borodulin et al *(Borodulin, K., Tolonen, H., Jousilahti, P., Jula, A., Juolevi, A., Koskinen, S., Kuulasmaa, K., Laatikainen, T., Mannisto, S., Peltonen, M., et al. (2017). Cohort Profile: The National FINRISK Study. Int J Epidemiol.)*

Acknowledgements/Funding:
The FINRISK controls were part of the FINRISK studies supported by THL (formerly KTL: National Public Health Institute) through budgetary funds from the government, with additional funding from institutions such as the Academy of Finland, the European Union, ministries and national and international foundations and societies to support specific research purposes. Genotyping of FINRISK controls was supported by the Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA. GSA data are available via an application through the THL Biobank portal https://thl-biobank.elixir-finland.org/

Genomic Psychiatry Cohort (GPC)

Description:
The controls from GPC that were contributed to Epi25 study were a subset of the overall
control participants with no personal or family history of schizophrenia or bipolar disorder. All the samples were genotyped on the GSA-MD v.1.0 at the Broad Institute. A detailed description of the GPC cohort can be found at Pato et al *(Pato, M.T., Sobell, J.L., Medeiros, H., Abbott, C., Sklar, B.M., Buckley, P.F., Bromet, E.J., Escamilla, M.A., Fanous, A.H., Lehrer, D.S., et al. (2013). The genomic psychiatry cohort: partners in discovery. Am J Med Genet B Neuropsychiatr Genet 162B, 306-312.)*

Acknowledgements/funding:
The GPC controls were genotyped on the GSA-MD v1.0 by the Broad Genomics Platform with funding from NIH grant U01MH105641 and the Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard.

Hong Kong Osteoporosis Study (HKOS)

Description:
The control samples were part of the follow-up study from the Hong Kong Osteoporosis Study (HKOS), which was described elsewhere (Cheung et al 2018). Briefly, community-dwelling Southern Chinese were firstly recruited from public roadshows in Hong Kong from 1995 to 2010. An extensive in-person follow-up study was initiated in 2015. At the in-person follow-up visit, the study participants were required to complete a comprehensive self-reported questionnaire, comprising questions related to their medical history, which were checked by experienced researchers or nurses based on a standard protocol. Fasting blood samples were collected from the study participants and DNA was extracted from the sera samples. Study participants without any history of epilepsy at the in-person follow-up in 2019 were included as controls of the epilepsy project. The study protocol was approved by the Institutional Review Board of the University of Hong Kong and the Hospital Authority Hong Kong West Cluster (Ref: UW 15-236). All HKOS participants provided informed consent for participation in the study. (*Cheung CL, Tan KCB, Kung AWC. Cohort Profile: The Hong Kong Osteoporosis Study and the follow-up study. Int J Epidemiol. 2018 Apr 1;47(2):397-398f. doi: 10.1093/ije/dyx172. PMID: 29024957.)*

Acknowledgements/funding:
The collection of samples was funded by the Bone Health Fund and Research Grants Council - Early Career Scheme (Project number: 27100416). Genotyping of samples on the GSA-MD v1 was done by the Broad Genomics Platform and supported by the NHGRI CCDG grant (1UM1HG008895). GSA data will be made available in dbGaP/AnVIL under phs001489.

NIDDK Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC)

Description/Acknowledgements/funding:
The NIDDK Inflammatory Bowel Disease Genetics Consortium (IBDGC) was created in 2002 by the National Institute of Diabetes, Digestive and Kidney diseases (NIDDK) to advance knowledge on the inflammatory bowel diseases, specifically Crohn's Disease and Ulcerative Colitis. The Consortium consists of six genetic research centers (GRC) and a data

coordinating center (DCC) that prospectively recruits a combination of cases, controls, and trios to gather a large collection of samples and linked phenotype information. DNA samples are used to conduct genetic linkage and association studies. For more information please see https://ibdgc.org/. Control samples from the following cohorts were included in the Epi25 GWAS: The University of Pittsburgh School of Medicine (PI: Richard Duerr), The Johns Hopkins Hospital (PI: Steven Brant), The Icahn School of Medicine at Mount Sinai (PI: Judy Cho), and Cedars Sinai (PI: Dermot McGovern, Stephan Targan). All samples were genotyped on the GSA-MD v1.0 by the Broad Genomics Platform.

We thank the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) IBD Genetics Consortium (IBDGC) supported by The Helmsley Charitable Trust and the Centers for Common Disease Genomics (NHGRI CCDG). Genotyping of samples on the GSA-MD v1 was done by the Broad Genomics Platform and supported by the NHGRI CCDG grant (1UM1HG008895). GSA-MD v1.0 data for these samples is available from dbGaP/AnVIL under study accession number phs001642.

Mass General Brigham (MGB) Biobank

Description/Acknowledgements/Funding:
The MGB (formerly Partners) Biobank (https://biobank.massgeneralbrigham.org/), launched in 2010, is a biorepository of consented patients samples at Mass General Brigham (parent organization of Massachusetts General Hospital and Brigham and Women's Hospital). The Biobank has enrolled >100K individuals to study how genes, lifestyle, and other factors affect people's health and contribute to disease. As part of the NHGRI's Centers for Common Disease Genomics, Broad Institute of MIT and Harvard generated genetic data for ~13,500 individuals from the MGB Biobank. We gratefully acknowledge the participants and leadership team of the MGB Biobank, funding support from the NHGRI CCDG (1UM1HG008895), and generation of new genotype data (Illumina Infinium GSA-MD v1) by the Broad Genomics Platform. GSA-MD v1.0 data for these samples is available from dbGaP under study accession number phs002018.v1.p1.

8.3. Other Publication

Adebayo, O.C., Betukumesu, D.K., Nkoy, A.B., **Adesoji, O.M.,** Ekulu, P.M., Van den Heuvel, L.P., Levtchenko, E.N. and Labarque, V., 2022. Clinical and genetic factors are associated with kidney complications in African children with sickle cell anaemia. *British Journal of Haematology*, *196*(1), pp.204-214.published online September 20, 2021. Doi: https://onlinelibrary.wiley.com/doi/epdf/10.1111/bjh.17832.

8.4. Meeting abstracts

Adesoji, O. M. and Nothnagel, M., 2020. A Simulation Study to Evaluate Existing Pleiotropy Detection Methods. *Hum. Hered.*, *84*(4-5), pp.204-205.

Adesoji O. M., Nothnagel M., Lerche H., May P., Krause R. A benchmarking of univariate pleiotropy detection methods, with an application to epilepsy phenotypes. 49th European Mathematical Genetics Meeting (EMGM) 2021. Paris, France. April 22 – 23, 2021.

Adesoji O. M., Nothnagel M. Benchmarking of univariate pleiotropy detection methods, with an application to epilepsy phenotypes. European Society of Human Genetics Conference (ESHG), 2022. Vienna, Austria. June 11-14, 2022.

8.5. Attended courses

Genomics and Transcriptomics, Integrated with Proteomics and Medical Informatics: learning the cornerstones of Systems Medicine (GTIPI). May 2022, Mainz, Germany

Ph.D. Translational specialistic medicine "GB MORGAGNI" Winter School. Shaping a World-class University – Seed funding for Digitalization & Innovation". January 24 – 28, 2022. Padova, Italy.

7th Sardinian international summer school. From genome-wide association studies to function. July 9-13, 2018. Sardinia Technology Park, Pula (CA), Italy.

Complex Trait Analysis of Next Generation Sequence Data. Max Delbrück Center (MDC) for Molecular Medicine. June 18-22, 2018. Berlin, Germany.

9. Discussion

This project's main objectives were to identify the optimal method among selected methods for pleiotropy detection and identify pleiotropic SNPs overlapping between two common forms of epilepsy, GGE and FE, using the identified method. However, an extensive literature review showed that the comparative performance for many of the available methods was unknown. Therefore, using simulated data, I benchmarked five univariate pleiotropy detection methods, namely; cFDR, CPBayes, ASSET, PLACO, and classical MA, to assess their relative performance in terms of power and false-positive rate. The ASSET method emerged as the best in terms of the power for detecting pleiotropy while keeping the FPR low in all simulation scenarios considered. Then, applying this optimal method to the summary statistics of the ILAE samples cohort, I identified two pleiotropic loci. Specifically, locus 2q24.3, already identified by the ILAE consortium, was confirmed, and a new putative locus 17q21.32 was identified. Using a larger sample cohort of the ILAE Consortium and EPI25 collaborative, I replicated the previously reported locus 2q24.3 and found a new locus 9q21.13 pleiotropic for GGE and FE.

9.1. Simulation Study

In this project, I compared the relative performance of recent univariate pleiotropy detection approaches alongside the well-known classical MA method on a large European population genotype data generated through resampling from the 1000 Genome haplotype data. One hundred replications of sub-samples of this population for the two phenotypes I studied were produced by repeatedly assigning the disease status to individuals through the additive liability threshold model (ALTM). Then, case-control study samples were simulated while also varying parameters that impact association analysis, such as effect size (RR= 1.05, 1.2, 1.5), sample size (n=2,000, 10,000, 20,000), diseases prevalence (1%, 10%), and varying numbers of diseased SNPs (5,10) and proportion of overlap (20%, 40%) of disease SNPs between the two phenotypes. Values of the additional parameters were selected to be consistent with observed values in GWAS of common diseases. The varying factors I introduced into the data simulation steps yielded different effects on the results in the identification of pleiotropic SNPs for all the methods.

Classical MA and Mega-analysis are not recommended for pleiotropy detection. The MA approach was characterized by a very high power to detect pleiotropy across simulation scenarios. At the same time, the FPR was also inflated due to the testing of the null hypothesis of no association with any of the traits and aggregating p-values across traits allowing for a trait with a very small p-value to drive the observed association for both traits under study. Due to the inflated FPR produced by the MA approach (See paragraph "**Inflated FPR**"), it is not recommended for pleiotropy analysis, according to the simulation study, as most of the identified loci or SNPs will be false-positive discoveries. This same observation is expected for mega-analysis in which several phenotypes not measured in the same set of individuals are jointly analyzed in a case-control association test. Therefore, neither mega-analysis nor classical meta-analysis allows us to conclude pleiotropy for these reasons and are, as a result, not recommended for pleiotropy detection.

The ASSET method is the optimal method for univariate single-marker pleiotropy detection. The ASSET approach maintained good power across all simulation scenarios. Though CPBayes, cFDR, and

PLACO methods also demonstrated good power to detect pleiotropy at larger sample sizes and effect sizes, they all differ considerably in their ability to keep a low FPR. Other methods apart from CPBayes recorded FPR of >10% in most simulation scenarios, while the ASSET maintained a much lower FPR (<10%) across all simulation scenarios. Based on this discovery in the simulation study, the ASSET method that gave a good trade-off between FPR and power to detect pleiotropy is hereby recommended for pleiotropy analysis.

The impact of sample size and effect sizes on pleiotropy detection. All the methods detected no pleiotropic disease SNP at an effect size of 1.05 regardless of the sample size. This result suggested that a larger sample size than used in this study is needed to achieve any power if RR is 1.05 for all the simulation scenarios. It has been demonstrated in GWAS that approximately 50000 samples are needed at this effect size to have sufficient power to detect pleiotropy for some common diseases. The power to identify the association of a locus to some trait(s) depends on the prevalence of disease, linkage disequilibrium (LD), inheritance model, number of risk alleles, frequency of the risk alleles, and their effect sizes[133]. Hence, for small sample sizes, a common variant must have a strong effect and sufficient power to detect association.

Inflated FPR. The inflated FPR observed for most of the methods, especially at larger sample sizes, appears somewhat counterintuitive. However, it is largely due to what hypothesis each method tests and how it estimates the overall p-value of pleiotropy, which in most cases allows one trait to drive the overall evidence of pleiotropy when its p-value is very small. Meta-analysis does not explicitly estimate the correlation between traits and aggregates P-values to test association, allowing for a single trait to drive observed association. The prevalence effect was more apparent on FPR and generally showed that samples from the population with 10% prevalence estimated lower FPR in all simulation scenarios compared to samples from the population with 1% prevalence of the trait. In addition, the more the number of disease SNPs simulated and the percentage overlap of these SNPs among the traits, the lower the FPR, confirming that common SNPs with average effect aggregate across the loci to produce the observed effect of the common variants on the phenotypes.

9.2. Application to epilepsy phenotypes (ILAE dataset only)

Findings. I applied the ASSET method, which gave a good trade-off between power and FPR while also correcting for sample overlap to the ILAE data samples. I identified pleiotropic loci 2q24.3 and 17q21.32 for both GGE and FE. My finding on chromosome 2 confirms the results reported by the ILAE Consortium on complex epilepsies[16] via mega-analysis as a likely pleiotropic locus, while locus 17q21.32 is a new putative pleiotropic locus only previously reported for GGE. Further annotation, tissue expression, colocalization, and prioritization tests supported the discoveries. Nevertheless, replicating these signals in an independent dataset is desirable.

True pleiotropy in loci 2q24.3. The loci 2q24.3 containing the *SCN1A* gene encodes the voltage-gated sodium channels and has been implicated in different forms of epilepsy[134,135]. This gene, expressed in both the peripheral and central nervous systems, is involved in transporting positively charged sodium atoms into cells and plays a crucial role in cells' ability to generate and transmit electrical signals[136]. Both common and rare variations in the *SCN1A* have been associated with epilepsy phenotypes with different severities, but the relatively common variants have been found to modulate the effect of the

SCN1A gene as well as other nearby genes such as *SCN2A* and *SCN9A*[135]. Therefore, with all the data available, loci 2q24.3 is truly pleiotropic for GGE and FE.

9.3. Application to epilepsy phenotypes (ILAE and EPI25 datasets).

Additional findings. As seen in the simulation study, an increase in sample size increases the power of observing pleiotropic association even for variants with relatively small effect sizes. The application of the ASSET method to the largest available data of common epilepsies yielded a new locus 9q21.13 in addition to locus 2q24.3. This confirms the observation in the simulation study that the larger the sample set, the more power to detect pleiotropy. Locus 9q21.13 contains the *RORB* gene, in which deletion of variants or single variant mutation has been associated with neurodevelopmental disorders such as developmental and epileptic encephalopathies and GGE[137,138]. The *RORB* gene encoded the beta retinoid-related orphan nuclear receptor (RORβ), a subfamily of nuclear hormone receptors NR1, present in immature neurons and thought to have a role in neuronal cell differentiation and hyperexcitability[138].

Replication. I could not directly replicate the initial finding on locus 17q21.32 in this new sample cohort as the locus was found to be only strongly associated with GGE with a marginally significant opposite direction of effect in FE. The observed result in the larger sample cohort corroborated the trend observed in the FE cases in the EPI25 collaborative dataset, where no genome-wide significance result was found. Further, only loci 2q24.3 and 9q21.13 were confirmed to be pleiotropic for GGE and FE among the four loci identified in the all epilepsy meta-analysis of the ILAE Consortium[132]. This finding reinforces my recommendation that classical meta-analysis and mega-analysis should not be used for pleiotropy detection. Locus 17q21.32 is not necessarily a false positive, but replication in an independent larger sample set is desirable.

Form of the observed pleiotropy. Ascertaining true pleiotropy and differentiating between the forms of pleiotropy is desirable, although not straightforward. However, new methods are emerging, such as spatial mapping approaches, methods that include the biological or gene pathway as part of the analysis[139,140], or screening out vertical pleiotropy by excluding correlated traits from the analysis and functional studies of the implicated genes. Until now, epilepsy phenotypes have not been found to be vertically related, i.e., one form of epilepsy phenotype has not been shown to mediate the effect of another epilepsy phenotype. Therefore, the identified pleiotropic SNPs are likely to be biological forms of pleiotropy. However, further functional studies of the identified genes, which are out of this project's scope, are necessary to understand the effects of the encoded proteins in these genes on epilepsy phenotypes.

Spurious pleiotropy. Eliminating spurious pleiotropy due to different forms of bias such as phenotype misclassification errors, overlapping controls cohort in independent studies, and high LD in regions, leading to marker tagging variants in different genes[49,50] is very important to reduce false positives rate in the application to the real dataset. Therefore, proper phenotyping, exclusion of high LD region from the analysis, accounting for overlapping samples, and ensuring that discovered pleiotropic markers are in the same gene are critical steps for consideration in pleiotropy analysis. In my analysis, the epilepsy phenotypes were properly classified, the applied ASSET method accounted

for the overlap between controls, and all identified pleiotropic variants are in LD in the same gene. I did not identify spurious pleiotropy in the samples, as all identified variants were contained in a gene.

9.4. Limitations

Some of the methods I applied here test the hypothesis of a variant being associated with any of the traits, hence, the observed significant pleiotropic association might be driven by a highly significant association of the variant to one phenotype. The newer pleiotropy detection methods like PLACO and CPBayes try to mitigate the one traits driving association issue but, from the observed results in this study, the FPR is still very high.

Although multivariate methods are computationally expensive, they have been demonstrated to be more powerful compared to the univariate approaches I used in this project. However, the unavailability of sample sets containing multiple phenotypes measured simultaneously on the same set of individuals coupled with heterogeneity due to ethnicity, microarray chips, and other sources of confounding render these approaches difficult to use. A recent publication[141] showed that a sparse group variable selection approach incorporating biological or gene pathways into the discovery of pleiotropic genes is more powerful than ASSET, nevertheless, this method also requires individual-level data.

While the univariate pleiotropy analysis is easy to perform with the readily available GWAS summary statistics, post-confirmatory functional studies of the identified genes are still very much needed to establish true pleiotropy. It is also difficult to distinguish between the different forms of pleiotropy, especially horizontal and vertical pleiotropy. Although correlated traits will mainly exhibit mediated or vertical pleiotropy, the underlying biological mechanism must be established to ascertain this claim. Mediation analysis, fine mapping, and pathway analysis can be useful methods to identify the form of pleiotropy[139,140,142]

The simulated population data comprising 1,000,000 individuals is quite large (~2 TB), and the simulation was carried out in 10 batches. That made the simulation of a larger population difficult due to computational constraints of available memory disk space, affecting the number of resulting samples I could simulate and process. In addition, methods like PLACO and common cFDR can only accommodate two traits. While this is not a challenge in this current study, the methods will not be applicable when more than two traits are to be studied.

9.5. Outlook

For future studies, it will be interesting to investigate the performance of pleiotropy detection methods for more than two phenotypes, for more nuanced sharing of causal genetic variation, possibly different effects on the pleiotropic phenotypes, and for less common or rare causal variants. Identifying pleiotropy in rare variants will likely require more complex genotype simulation algorithms and larger reference sample sets. The availability of simultaneously measured individual-level data of epilepsy phenotypes in the future will also motivate the application of multivariate pleiotropy detection methods.

Replicating the identified loci in larger independent samples of epilepsy phenotypes is also desirable to eliminate bias and confounding. Also, applying a new pleiotropy detection method that

considers the complex genetic architectures of traits, such as genetic correlations and heritabilities, could improve pleiotropy detection. One of such recent approaches is pleiotropic Locus exploration and interpretation using an optimal test (PLEIO)[112].

10. Acknowledgments

I appreciate God for the strength and grace to start and complete this thesis project. It has been four years on a roller-coaster. I would also like to thank my promotor, Prof. Dr. Michael Nothnagel, for the opportunity to work on this project, his guidance, and his contribution every step of the way. Michael does not see any question as ridiculous and is always willing to explain and teach whatever is unclear while operating an open office principle. I appreciate the many opportunities you brought my direction during this program.

I sincerely appreciate Prof. Dr. Peter Nuernberg and Prof. Dr. Andreas Bayer for agreeing to be part of my thesis advisory committee. Thank you for your time, suggestions, and significant contributions to the project.

I would also like to acknowledge the DFG-funded FOR2715 group headed by Prof. Dr. med. Holger Lerche, for the opportunity to collaborate in this research group. Special appreciation to Dr. Roland Krause and Dr. Patrick May for providing computing support at the University of Luxembourg, data access, and valuable advice during my Ph.D. program. Thank you, Dr. Herbert Schutz, for the many interactions and suggestions throughout the course of my thesis.

I would also like to thank all my current and former colleagues who have contributed one way or the other to the success of this project. Dr. Dmitriy Drichel, Dr. George Kanoungi, Dr. Maria-Alexandra Katsara, Tarek Kellaf, and Dr. Elaheh Vojgani. Thank you for helping me set up my Linux computer and HPC and for all the encouragement when I hit a wall. I indeed appreciate your contributions to the success of my project. I am especially grateful to Dr. Maria-Alexandra Katsara and Dr. Dimitry Drichel for creating time to proofread my thesis and giving excellent suggestions and corrections that significantly improved the thesis.

Thank you, Dr. Gabriele Thorn, for always going beyond your administrative duties to support me in the CCG. I can not forget how you went on an apartment search with me and eventually got me a comfortable one. I appreciate Mr. Khanh Nguyen and Mr. Heinrich Rohde for the excellent IT support they provided.

This journey started in 2015 when I decided to come for my master's program in Belgium, and it would not have been possible without the support of my family. Thanks to my parents, Mr. and Mrs. Adesoji. I wish Mr. B. S. were still here to tell me “I am proud of you” today, but I know you are proud of me wherever you are. To my siblings, thanks for the love and courage. I will not forget the support of Mrs. Olukemi Awotipe, Prof. Abimbola Adesoji, Mrs. Adenike Akodu, Mr. Opeyemi Adesoji, Mrs. Adeola Ogunleye, and Mrs. Oyindamola Awogbindin. You are the best siblings anyone could have.

I appreciate all my friends who have turned into family, Oluwadara, Oluwafemi, Olajumoke, Isaac, Blessing, Deborah, Oyindamola, Victor, Oluwatunseyi, and Yaqub. Special appreciation to my “dutch parents,” Marie-Louise and Wil Gerrits. Thank you for your love and care. I appreciate Mr. Oluwadiran, Pastor, and Mrs. Benjamin for your prayers and fatherly role.

Thank you, Bobomi, Omoniyi Emmanuel Arowoiya. You are the superstar! I love you always.

11. Erklärung

Erklärung zur Dissertation

gemäß der Promotionsordnung vom 12. März 2020

„Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten - noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht."

Teilpublikationen:

Adesoji, O. M., Schulz, H., May, P., Krause, R., Lerche, H., Nothnagel, M., & ILAE Consortium on Complex Epilepsies. (2022). Benchmarking of univariate pleiotropy detection methods applied to epilepsy. *Human Mutation*. Published online May 27, 2022. Doi: https://doi.org/10.1002/humu.24417.

International League Against Epilepsy Consortium on Complex Epilepsies (2022). Genome-wide meta-analysis of over 29,000 people with epilepsy reveals 26 loci and subtype-specific genetic architecture | medRxiv (Submitted). **Doi:** https://doi.org/10.1101/2022.06.08.22276120.

Datum, Name und Unterschrift

Oluyomi Modupe Adesoji

11.07.2022

12. References

1. Arnar, D.O., and Palsson, R. (2017). Genetics of common complex diseases: a view from Iceland. European Journal of Internal Medicine *41*, 3-9.
2. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190-2191. 10.1093/bioinformatics/btq340.
3. Bhattacharjee, S., Rajaraman, P., Jacobs, K.B., Wheeler, W.A., Melin, B.S., Hartge, P., GliomaScan, C., Yeager, M., Chung, C.C., Chanock, S.J., and Chatterjee, N. (2012). A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. Am J Hum Genet *90*, 821-835. 10.1016/j.ajhg.2012.03.015.
4. Majumdar, A., Haldar, T., Bhattacharya, S., and Witte, J.S. (2018). An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations. PLoS Genet *14*, e1007139. 10.1371/journal.pgen.1007139.
5. Liley, J., and Wallace, C. (2015). A pleiotropy-informed Bayesian false discovery rate adapted to a shared control design finds new disease associations from GWAS summary statistics. PLoS Genet *11*, e1004926. 10.1371/journal.pgen.1004926.
6. Ray, D., and Chatterjee, N. (2020). A powerful method for pleiotropic analysis under composite null hypothesis identifies novel shared loci between Type 2 Diabetes and Prostate Cancer. PLoS Genet *16*, e1009218. 10.1371/journal.pgen.1009218.
7. Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S., Penegar, S., Chandler, I., Gorman, M., Wood, W., et al. (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet *39*, 984-988. 10.1038/ng2085.
8. Thomas, G., Jacobs, K.B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., Welch, R., Hutchinson, A., et al. (2008). Multiple loci identified in a genome-wide association study of prostate cancer. Nat Genet *40*, 310-315. 10.1038/ng.91.
9. Genetic Analysis of Psoriasis Consortium, the Wellcome Trust Case Control Consortium, Strange, A., Capon, F., Spencer, C.C., Knight, J., Weale, M.E., Allen, M.H., Barton, A., Band, G., et al. (2010). A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat Genet *42*, 985-990. 10.1038/ng.694.
10. Franke, A., McGovern, D.P., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet *42*, 1118-1125. 10.1038/ng.717.
11. Consortium, B., Anttila, V., Bulik-Sullivan, B., Finucane, H.K., Walters, R.K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G.J., and Gormley, P. (2018). Analysis of shared heritability in common disorders of the brain. Science *360*, eaap8757.
12. Stearns, F.W. (2010). One hundred years of pleiotropy: a retrospective. Genetics *186*, 767-773. 10.1534/genetics.110.122549.
13. Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. Bioinformatics *27*, 2304-2305. 10.1093/bioinformatics/btr341.
14. Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68-74. 10.1038/nature15393.
15. Agarwala, V., Flannick, J., Sunyaev, S., Go, T.D.C., and Altshuler, D. (2013). Evaluating empirical bounds on complex disease genetic architecture. Nat Genet *45*, 1418-1427. 10.1038/ng.2804.
16. International League Against Epilepsy Consortium on Complex Epilepsies (2018). Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. Nat Commun *9*, 5269. 10.1038/s41467-018-07524-z.
17. Witthöft, M. (2013). Etiology/Pathogenesis. In Encyclopedia of Behavioral Medicine, M.D. Gellman, and J.R. Turner, eds. (Springer New York), pp. 716-717. 10.1007/978-1-4419-1005-9_16.
18. Iourov, I.Y., Vorsanova, S.G., and Yurov, Y.B. (2019). Pathway-based classification of genetic diseases. Molecular cytogenetics *12*, 1-5.
19. Jackson, M., Marks, L., May, G.H., and Wilson, J.B. (2018). The genetic basis of disease. Essays in biochemistry *62*, 643-723.
20. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., and Gershman, A. (2022). The complete sequence of a human genome. Science *376*, 44-53.
21. Brookes, A.J. (1999). The essence of SNPs. Gene *234*, 177-186.
22. Shastry, B.S. (2007). SNPs in disease gene mapping, medicinal drug development and evolution. Journal of Human Genetics *52*, 871-880. 10.1007/s10038-007-0200-z.
23. Thorleifsson, G., Magnusson, K.P., Sulem, P., Walters, G.B., Gudbjartsson, D.F., Stefansson, H., Jonsson, T., Jonasdottir, A., Jonasdottir, A., and Stefansdottir, G. (2007). Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma. Science *317*, 1397-1400.

24. Strittmatter, W.J., and Roses, A.D. (1996). Apolipoprotein E and Alzheimer's disease. Annual review of neuroscience *19*, 53-77.
25. Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. Trends Genet *17*, 502-510. 10.1016/s0168-9525(01)02410-6.
26. Altshuler, D., Daly, M.J., and Lander, E.S. (2008). Genetic mapping in human disease. science *322*, 881-888.
27. Li, H., Achour, I., Bastarache, L., Berghout, J., Gardeux, V., Li, J., Lee, Y., Pesce, L., Yang, X., and Ramos, K.S. (2016). Integrative genomics analyses unveil downstream biological effectors of disease-specific polymorphisms buried in intergenic regions. NPJ genomic medicine *1*, 1-12.
28. Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics *9*, 477-485.
29. Londin, E., Yadav, P., Surrey, S., Kricka, L.J., and Fortina, P. (2013). Use of linkage analysis, genome-wide association studies, and next-generation sequencing in the identification of disease-causing mutations. Pharmacogenomics, 127-146.
30. Ott, J., Wang, J., and Leal, S.M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. Nature Reviews Genetics *16*, 275-284.
31. Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? The American Journal of Human Genetics *69*, 124-137.
32. Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. (2009). Common vs. rare allele hypotheses for complex diseases. Current opinion in genetics & development *19*, 212-219. 10.1016/j.gde.2009.04.010.
33. Koeleman, B.P. (2018). What do genetic studies tell us about the heritable basis of common epilepsy? Polygenic or complex epilepsy? Neuroscience Letters *667*, 10-16.
34. Spronk, I., Korevaar, J.C., Poos, R., Davids, R., Hilderink, H., Schellevis, F.G., Verheij, R.A., and Nielen, M.M.J. (2019). Calculating incidence rates and prevalence proportions: not as simple as it seems. BMC Public Health *19*, 512. 10.1186/s12889-019-6820-3.
35. Song, J.W., and Chung, K.C. (2010). Observational studies: cohort and case-control studies. Plast Reconstr Surg *126*, 2234-2242. 10.1097/PRS.0b013e3181f44abc.
36. Coggon, D., Barker, D., and Rose, G. (2009). Epidemiology for the Uninitiated (John Wiley & Sons).
37. Rezigalla, A.A. (2020). Observational Study Designs: Synopsis for Selecting an Appropriate Study Design. Cureus *12*, e6692. 10.7759/cureus.6692.
38. Marees, A.T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., and Derks, E.M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. International journal of methods in psychiatric research *27*, e1608.
39. Clarke, G.M., Anderson, C.A., Pettersson, F.H., Cardon, L.R., Morris, A.P., and Zondervan, K.T. (2011). Basic statistical analysis in genetic case-control studies. Nature protocols *6*, 121-133.
40. Lewis, C.M. (2002). Genetic association studies: design, analysis and interpretation. Briefings in bioinformatics *3*, 146-153.
41. Bush, W.S., and Moore, J.H. (2012). Chapter 11: Genome-wide association studies. PLoS computational biology *8*, e1002822.
42. Kuo, K.H. (2017). Multiple testing in the context of gene discovery in sickle cell disease using genome-wide association studies. Genomics insights *10*, 1178631017721178.
43. Westfall, P.H., and Young, S.S. (1993). Resampling-based multiple testing: Examples and methods for p-value adjustment (John Wiley & Sons).
44. Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M.J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society *32*, 381-385.
45. Mendel, G. (1996). Experiments in plant hybridization (1865). Verhandlungen des naturforschenden Vereins Brünn.) Available online: www. mendelweb. org/Mendel. html (accessed on 1 January 2013).
46. Chakravarti, A. (2021). Magnitude of Mendelian versus complex inheritance of rare disorders. American Journal of Medical Genetics Part A.
47. Gilissen, C., Hoischen, A., Brunner, H.G., and Veltman, J.A. (2012). Disease gene identification strategies for exome sequencing. European Journal of Human Genetics *20*, 490-497.
48. Watanabe, K., Stringer, S., Frei, O., Umicevic Mirkov, M., de Leeuw, C., Polderman, T.J.C., van der Sluis, S., Andreassen, O.A., Neale, B.M., and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. Nat Genet *51*, 1339-1348. 10.1038/s41588-019-0481-0.
49. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M., and Smoller, J.W. (2013). Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet *14*, 483-495. 10.1038/nrg3461.
50. Hackinger, S., and Zeggini, E. (2017). Statistical methods to detect pleiotropy in human complex traits. Open Biol *7*. 10.1098/rsob.170125.
51. Mitchell, K.J. (2012). What is complex about complex disorders? Genome biology *13*, 1-11.
52. Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era—concepts and misconceptions. Nature reviews genetics *9*, 255-266.

53. de Los Campos, G., Sorensen, D., and Gianola, D. (2015). Genomic heritability: what is it? PLoS Genetics *11*, e1005048.
54. Falconer, D. (1967). The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. Annals of human genetics *31*, 1-20.
55. Ioannidis, J.P. (2015). Making optimal use of and extending beyond polygenic additive liability models. Human Heredity *80*, 158-161.
56. Visscher, P.M., and Wray, N.R. (2015). Concepts and misconceptions about the polygenic additive model applied to disease. Human heredity *80*, 165-170.
57. Polderman, T.J., Benyamin, B., De Leeuw, C.A., Sullivan, P.F., Van Bochoven, A., Visscher, P.M., and Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nature genetics *47*, 702-709.
58. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747-753. 10.1038/nature08494.
59. Beghi, E. (2020). The epidemiology of epilepsy. Neuroepidemiology *54*, 185-191.
60. Falco-Walter, J.J., Scheffer, I.E., and Fisher, R.S. (2018). The new definition and classification of seizures and epilepsy. Epilepsy research *139*, 73-79.
61. NIoNDaS (2016). The Epilepsies and Seizures: Hope Through Research. National Institutes of Health (NIH) Bethesda, MD, USA.
62. Beghi, E., and Mohammed, S. (2019). Global, regional, and national burden of epilepsy, 1990-2016.
63. Giourou, E., Stavropoulou-Deli, A., Giannakopoulou, A., Kostopoulos, G.K., and Koutroumanidis, M. (2015). Introduction to Epilepsy and Related Brain Disorders. In Cyberphysical Systems for Epilepsy and Related Brain Disorders, (Springer), pp. 11-38.
64. Fiest, K.M., Sauro, K.M., Wiebe, S., Patten, S.B., Kwon, C.-S., Dykeman, J., Pringsheim, T., Lorenzetti, D.L., and Jetté, N. (2017). Prevalence and incidence of epilepsy: a systematic review and meta-analysis of international studies. Neurology *88*, 296-303.
65. Alva-Díaz, C., Navarro-Flores, A., Rivera-Torrejon, O., Huerta-Rosario, A., Molina, R.A., Velásquez-Rimachi, V., Morán-Mariños, C., Farroñay, C., Pacheco-Mendoza, J., and Metcalf, T. (2021). Prevalence and incidence of epilepsy in Latin America and the Caribbean: A systematic review and meta-analysis of population-based studies. Epilepsia *62*, 984-996.
66. Watila, M.M., Balarabe, S.A., Komolafe, M.A., Igwe, S.C., Fawale, M.B., Otte, W.M., van Diessen, E., Okunoye, O., Mshelia, A.A., Abdullahi, I., et al. (2021). Epidemiology of Epilepsy in Nigeria. A Community-Based Study From 3 Sites *97*, e728-e738. 10.1212/wnl.0000000000012416.
67. Guerrini, R., and Buchhalter, J.R. (2014). Epilepsy phenotypes and genotype determinants. Identical twins teach lessons on complexity *83*, 1038-1039. 10.1212/wnl.0000000000000802.
68. Vadlamudi, L., Milne, R.L., Lawrence, K., Heron, S.E., Eckhaus, J., Keay, D., Connellan, M., Torn-Broers, Y., Howell, R.A., Mulley, J.C., et al. (2014). Genetics of epilepsy: The testimony of twins in the molecular era. Neurology *83*, 1042-1048. 10.1212/wnl.0000000000000790.
69. Lennox, W.G., and Lennox-Buchthal, M.A. (1960). Epilepsy and related disorders (Little, Brown).
70. Scheffer, I.E., Berkovic, S., Capovilla, G., Connolly, M.B., French, J., Guilhoto, L., Hirsch, E., Jain, S., Mathern, G.W., Moshe, S.L., et al. (2017). ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. Epilepsia *58*, 512-521. 10.1111/epi.13709.
71. Perucca, P., Bahlo, M., and Berkovic, S.F. (2020). The genetics of epilepsy. Annual review of genomics and human genetics *21*, 205-230.
72. Jallon, P., and Latour, P. (2005). Epidemiology of idiopathic generalized epilepsies. Epilepsia *46*, 10-14.
73. Perucca, P. (2018). Genetics of focal epilepsies: what do we know and where are we heading? Epilepsy currents *18*, 356-362.
74. Eriksson, H., Wirdefeldt, K., Åsberg, S., and Zelano, J. (2019). Family history increases the risk of late seizures after stroke. Neurology *93*, e1964-e1970.
75. Christensen, J., Pedersen, M.G., Pedersen, C.B., Sidenius, P., Olsen, J., and Vestergaard, M. (2009). Long-term risk of epilepsy after traumatic brain injury in children and young adults: a population-based cohort study. The Lancet *373*, 1105-1110.
76. Mendel, G. (1965). Experiments in plant hybridization (Oliver & Boyd).
77. Pyeritz, R.E. (1989). Pleiotropy revisited: molecular explanations of a classic concept. American journal of medical genetics *34*, 124-134.
78. Tyler, A.L., Crawford, D.C., and Pendergrass, S.A. (2014). Detecting and characterizing pleiotropy: new methods for uncovering the connection between the complexity of genomic architecture and multiple phenotypes. (NIH Public Access), pp. 183.
79. Hodgkin, J. (2002). Seven types of pleiotropy. International Journal of Developmental Biology *42*, 501-505.
80. Bien, S.A., and Peters, U. (2019). Moving from one to many: insights from the growing list of pleiotropic cancer risk genes. Br J Cancer *120*, 1087-1089. 10.1038/s41416-019-0475-9.

81. Polvi, A., Siren, A., Kallela, M., Rantala, H., Artto, V., Sobel, E., Palotie, A., Lehesjoki, A.-E., and Wessman, M. (2012). Shared loci for migraine and epilepsy on chromosomes 14q12-q23 and 12q24. 2-q24. 3. Neurology *78*, 202-209.
82. Consortium, C.-D.G.o.t.P.G. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. The Lancet *381*, 1371-1379.
83. Bellou, E., Stevenson-Hoare, J., and Escott-Price, V. (2020). Polygenic risk and pleiotropy in neurodegenerative diseases. Neurobiology of Disease *142*, 104953.
84. Consortium, I.S. (2009). Common polygenic variation contributes to risk of schizophrenia that overlaps with bipolar disorder. Nature *460*, 748.
85. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics *88*, 76-82.
86. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M., and Wray, N.R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics *28*, 2540-2542.
87. Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., de Candia, T.R., Lee, S.H., Wray, N.R., and Kendler, K.S. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nature genetics *47*, 1385-1392.
88. Furlotte, N.A., and Eskin, E. (2015). Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. Genetics *200*, 59-68.
89. Huang, J., Johnson, A.D., and O'Donnell, C.J. (2011). PRIMe: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies. Bioinformatics *27*, 1201-1206.
90. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet *10*, e1004383. 10.1371/journal.pgen.1004383.
91. Seoane, J.A., Campbell, C., Day, I.N., Casas, J.P., and Gaunt, T.R. (2014). Canonical correlation analysis for gene-based pleiotropy discovery. PLoS computational biology *10*, e1003876.
92. Cichonska, A., Rousu, J., Marttinen, P., Kangas, A.J., Soininen, P., Lehtimäki, T., Raitakari, O.T., Järvelin, M.-R., Salomaa, V., and Ala-Korpela, M. (2016). metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. Bioinformatics *32*, 1981-1989.
93. Casale, F.P., Rakitsch, B., Lippert, C., and Stegle, O. (2015). Efficient set tests for the genetic analysis of correlated traits. Nature methods *12*, 755-758.
94. Agresti, A. (2003). Categorical data analysis (John Wiley \& Sons).
95. Morris, A.P., Lindgren, C.M., Zeggini, E., Timpson, N.J., Frayling, T.M., Hattersley, A.T., and McCarthy, M.I. (2010). A powerful approach to sub-phenotype analysis in population-based genetic association studies. Genet Epidemiol *34*, 335-343. 10.1002/gepi.20486.
96. Fitzmaurice, G.M., and Laird, N.M. (1993). A likelihood-based method for analysing longitudinal binary responses. Biometrika *80*, 141-151.
97. Schaid, D.J., Tong, X., Batzler, A., Sinnwell, J.P., Qing, J., and Biernacka, J.M. (2019). Multivariate generalized linear model for genetic pleiotropy. Biostatistics *20*, 111-128. 10.1093/biostatistics/kxx067.
98. Verbeke, G., Molenberghs, G., and Rizopoulos, D. (2010). Random effects models for longitudinal data. In Longitudinal research with latent variables, (Springer), pp. 37-96.
99. Laird, N.M., and Ware, J.H. (1982). Random-effects models for longitudinal data. Biometrics *38*, 963-974.
100. Lange, C., Silverman, E.K., Xu, X., Weiss, S.T., and Laird, N.M. (2003). A multivariate family-based association test using generalized estimating equations: FBAT-GEE. Biostatistics *4*, 195-206.
101. Liu, J., Pei, Y., Papasian, C.J., and Deng, H.W. (2009). Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society *33*, 217-227.
102. Yang, Q., and Wang, Y. (2012). Methods for Analyzing Multivariate Phenotypes in Genetic Association Studies. J Probab Stat *2012*, 652569. 10.1155/2012/652569.
103. Jolliffe, I.T. (2002). Graphical representation of data using principal components. Principal component analysis, 78-110.
104. Jolliffe, I.T., and Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences *374*, 20150202.
105. Salinas, Y.D., Wang, Z., and DeWan, A.T. (2018). Statistical Analysis of Multiple Phenotypes in Genetic Epidemiologic Studies: From Cross-Phenotype Associations to Pleiotropy. Am J Epidemiol *187*, 855-863. 10.1093/aje/kwx296.
106. Evangelou, E., and Ioannidis, J.P. (2013). Meta-analysis methods for genome-wide association studies and beyond. Nature Reviews Genetics *14*, 379-389.
107. Cotsapas, C., Voight, B.F., Rossin, E., Lage, K., Neale, B.M., Wallace, C., Abecasis, G.R., Barrett, J.C., Behrens, T., and Cho, J. (2011). Pervasive sharing of genetic effects in autoimmune disease. PLoS genetics *7*, e1002254.
108. Zhu, X., Feng, T., Tayo, B.O., Liang, J., Young, J.H., Franceschini, N., Smith, J.A., Yanek, L.R., Sun, Y.V., and Edwards, T.L. (2015). Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. The American Journal of Human Genetics *96*, 21-36.

109. Li, X., and Zhu, X. (2017). Cross-phenotype association analysis using summary statistics from GWAS. In Statistical Human Genetics, (Springer), pp. 455-467.
110. Vuckovic, D., Gasparini, P., Soranzo, N., and Iotchkova, V. (2015). MultiMeta: an R package for meta-analyzing multi-phenotype genome-wide association studies. Bioinformatics *31*, 2754-2756.
111. Turley, P., Walters, R.K., Maghzian, O., Okbay, A., Lee, J.J., Fontana, M.A., Nguyen-Viet, T.A., Wedow, R., Zacher, M., and Furlotte, N.A. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. Nature genetics *50*, 229-237.
112. Lee, C.H., Shi, H., Pasaniuc, B., Eskin, E., and Han, B. (2021). PLEIO: a method to map and interpret pleiotropic loci with GWAS summary statistics. The American Journal of Human Genetics *108*, 36-48.
113. Andreassen, O.A., Thompson, W.K., Schork, A.J., Ripke, S., Mattingsdal, M., Kelsoe, J.R., Kendler, K.S., O'Donovan, M.C., Rujescu, D., Werge, T., et al. (2013). Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. PLoS Genet *9*, e1003455. 10.1371/journal.pgen.1003455.
114. Kulminski, A.M., Loika, Y., Huang, J., Arbeev, K.G., Bagley, O., Ukraintseva, S., Yashin, A.I., and Culminskaya, I. (2019). Pleiotropic Meta-Analysis of Age-Related Phenotypes Addressing Evolutionary Uncertainty in Their Molecular Mechanisms. Front Genet *10*, 433. 10.3389/fgene.2019.00433.
115. Chung, J., Jun, G.R., Dupuis, J., and Farrer, L.A. (2019). Comparison of methods for multivariate gene-based association tests for complex diseases using common variants. Eur J Hum Genet *27*, 811-823. 10.1038/s41431-018-0327-8.
116. Borenstein, M., Hedges, L.V., Higgins, J.P., and Rothstein, H.R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. Research synthesis methods *1*, 97-111.
117. Borenstein, M., Hedges, L., and Rothstein, H. (2007). Meta-analysis: Fixed effect vs. random effects. Meta-analysis. com.
118. Andreassen, O.A., Djurovic, S., Thompson, W.K., Schork, A.J., Kendler, K.S., O'Donovan, M.C., Rujescu, D., Werge, T., van de Bunt, M., Morris, A.P., et al. (2013). Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. Am J Hum Genet *92*, 197-209. 10.1016/j.ajhg.2013.01.001.
119. Carvajal-Rodriguez, A. (2010). Simulation of genes and genomes forward in time. Curr Genomics *11*, 58-61. 10.2174/138920210790218007.
120. Li, N., and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics *165*, 2213-2233.
121. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, s13742-13015-10047-13748.
122. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., and Daly, M.J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics *81*, 559-575.
123. Consortium, T.I. (2021). ILAE Strategy 2030. https://www.ilae.org/about-ilae/ilae-mission-goals-and-strategy.
124. Consortium, T.I.L.A.E. (2014). Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. The Lancet. Neurology *13*, 893.
125. Watanabe, K., Taskesen, E., van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. Nat Commun *8*, 1826. 10.1038/s41467-017-01261-5.
126. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res *38*, e164. 10.1093/nar/gkq603.
127. Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome Res *22*, 1790-1797. 10.1101/gr.137323.112.
128. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet *46*, 310-315. 10.1038/ng.2892.
129. de Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol *11*, e1004219. 10.1371/journal.pcbi.1004219.
130. Foley, C.N., Staley, J.R., Breen, P.G., Sun, B.B., Kirk, P.D.W., Burgess, S., and Howson, J.M.M. (2021). A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. Nat Commun *12*, 764. 10.1038/s41467-020-20885-8.
131. Wallace, C. (2020). Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. PLoS Genet *16*, e1008720. 10.1371/journal.pgen.1008720.
132. Berkovic, S.F., Cavalleri, G.L., and Koeleman, B.P. (2022). Genome-wide meta-analysis of over 29,000 people with epilepsy reveals 26 loci and subtype-specific genetic architecture. medRxiv.
133. Nishino, J., Ochi, H., Kochi, Y., Tsunoda, T., and Matsui, S. (2018). Sample size for successful genome-wide association study of major depressive disorder. Frontiers in genetics *9*, 227.
134. Scheffer, I.E., and Nabbout, R. (2019). SCN1A-related phenotypes: Epilepsy and beyond. Epilepsia *60*, S17-S24.

135. de Lange, I.M., Mulder, F., van 't Slot, R., Sonsma, A.C.M., van Kempen, M.J.A., Nijman, I.J., Ernst, R.F., Knoers, N., Brilstra, E.H., and Koeleman, B.P.C. (2020). Modifier genes in SCN1A-related epilepsy syndromes. Mol Genet Genomic Med *8*, e1103. 10.1002/mgg3.1103.
136. Lossin, C. (2009). A catalog of SCN1A variants. Brain and Development *31*, 114-130.
137. Rudolf, G., Lesca, G., Mehrjouy, M.M., Labalme, A., Salmi, M., Bache, I., Bruneau, N., Pendziwiat, M., Fluss, J., and De Bellescize, J. (2016). Loss of function of the retinoid-related nuclear receptor (RORB) gene and epilepsy. European Journal of Human Genetics *24*, 1761-1770.
138. Sadleir, L.G., de Valles-Ibáñez, G., King, C., Coleman, M., Mossman, S., Paterson, S., Nguyen, J., Berkovic, S.F., Mullen, S., and Bahlo, M. (2020). Inherited RORB pathogenic variants: Overlap of photosensitive genetic generalized and occipital lobe epilepsy. Epilepsia *61*, e23-e29.
139. Jordan, D.M., Verbanck, M., and Do, R. (2019). HOPS: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. Genome biology *20*, 1-18.
140. Wen, Y., Wang, W., Guo, X., and Zhang, F. (2016). PAPA: a flexible tool for identifying pleiotropic pathways using genome-wide association study summaries. Bioinformatics *32*, 946-948.
141. Sutton, M., Sugier, P.-E., Truong, T., and Liquet, B. (2022). Leveraging pleiotropic association using sparse group variable selection in genomics data. BMC medical research methodology *22*, 1-12.
142. Lutz, S.M., and Hokanson, J.E. (2015). Mediation analysis in genome-wide association studies: current perspectives. Open Access Bioinformatics *7*, 1-5.