ARTICLE

Exome Sequencing Followed by Large-Scale Genotyping Fails to Identify Single Rare Variants of Large Effect in Idiopathic Generalized Epilepsy

Erin L. Heinzen,[1,2,*] Chantal Depondt,[3] Gianpiero L. Cavalleri,[4] Elizabeth K. Ruzzo,[1] Nicole M. Walley,[1] Anna C. Need,[1,2] Dongliang Ge,[1,2] Min He,[1] Elizabeth T. Cirulli,[1] Qian Zhao,[1] Kenneth D. Cronin,[1] Curtis E. Gumbs,[1] C. Ryan Campbell,[1] Linda K. Hong,[1] Jessica M. Maia,[1] Kevin V. Shianna,[1,2] Mark McCormack,[4] Rodney A. Radtke,[5] Gerard D. O'Conner,[6] Mohamad A. Mikati,[7] William B. Gallentine,[7] Aatif M. Husain,[5] Saurabh R. Sinha,[5] Krishna Chinthapalli,[8] Ram S. Puranam,[9] James O. McNamara,[9] Ruth Ottman,[10,11] Sanjay M. Sisodiya,[8] Norman Delanty,[4,6,13] and David B. Goldstein[1,12,13,*]

Idiopathic generalized epilepsy (IGE) is a complex disease with high heritability, but little is known about its genetic architecture. Rare copy-number variants have been found to explain nearly 3% of individuals with IGE; however, it remains unclear whether variants with moderate effect size and frequencies below what are reliably detected with genome-wide association studies contribute significantly to disease risk. In this study, we compare the exome sequences of 118 individuals with IGE and 242 controls of European ancestry by using next-generation sequencing. The exome-sequenced epilepsy cases include study subjects with two forms of IGE, including juvenile myoclonic epilepsy (n = 93) and absence epilepsy (n = 25). However, our discovery strategy did not assume common genetic control between the subtypes of IGE considered. In the sequence data, as expected, no variants were significantly associated with the IGE phenotype or more specific IGE diagnoses. We then selected 3,897 candidate epilepsy-susceptibility variants from the sequence data and genotyped them in a larger set of 878 individuals with IGE and 1,830 controls. Again, no variant achieved statistical significance. However, 1,935 variants were observed exclusively in cases either as heterozygous or homozygous genotypes. It is likely that this set of variants includes real risk factors. The lack of significant association evidence of single variants with disease in this two-stage approach emphasizes the high genetic heterogeneity of epilepsy disorders, suggests that the impact of any individual single-nucleotide variant in this disease is small, and indicates that gene-based approaches might be more successful for future sequencing studies of epilepsy predisposition.

Introduction

In the past decade, there have been extensive efforts to identify the genetic basis of common, complex human disease. Until recently, this effort has focused primarily on common variants with the use of genome-wide association studies. Although a number of important gene discoveries have emerged from this work, most of the identified variants have only a small effect on disease risk, particularly of neuropsychiatric diseases.[1] Recently, however, studies have demonstrated that rare copy-number variants (CNVs) contribute to the risk of developing neurological, psychiatric, and developmental disorders.[2–10] These findings implicate specific risk loci in these diseases and also support the role of rare variation in the genetic architecture of neuropsychiatric diseases.

Next-generation sequencing (NGS) allows for nearly comprehensive cataloging of genetic variants and offers an unprecedented ability to identify rare variants. However, distinguishing the disease-causing variants from the large number of rare and apparently functional variants found in each human genome presents a considerable challenge. Although the optimum NGS study design for identifying causal genetic loci in complex disease will evolve as we expand our knowledge of the underlying genetic architecture, NGS studies evaluating the role of rare variants in complex diseases at this stage include two broadly defined approaches: (1) variant-based assessments for exploring the role of individual variants that have a relatively large effect and that are common enough to be seen in multiple cases and (2) gene-level approaches that seek to identify multiple rare risk alleles within the same genes.

[1]Center for Human Genome Variation, Duke University School of Medicine, Durham, NC 27708, USA; [2]Department of Medicine, Section of Medical Genetics, Duke University School of Medicine, Durham, NC 27708, USA; [3]Department of Neurology, Hôpital Erasme, Université Libre de Bruxelles, 1070 Brussels, Belgium; [4]Molecular and Cellular Therapeutics, Royal College of Surgeons in Ireland, Dublin 2, Ireland; [5]Division of Neurology, Department of Medicine, Duke University School of Medicine, Durham, NC 27710, USA; [6]Division of Neurology, Beaumont Hospital, Dublin 9, Ireland; [7]Division of Pediatric Neurology, Department of Pediatrics, Duke University School of Medicine, Durham, NC 27710, USA; [8]Department of Clinical and Experimental Epilepsy, University College London Institute of Neurology, Queen Square, London, WC1N 3BG, UK; [9]Department of Neurobiology, Duke University, Durham, NC 27710, USA; [10]Sergievsky Center and Departments of Neurology and Epidemiology, Columbia University, New York, NY 10032, USA; [11]Division of Epidemiology, New York State Psychiatric Institute, New York, NY 10032, USA; [12]Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, NC 27708, USA
[13]These authors contributed equally to this work
*Correspondence: e.heinzen@duke.edu (E.L.H.), d.goldstein@duke.edu (D.B.G.)

Exome sequencing (Agilent All Exon)
- ➤ 118 IGE cases
- ➤ 242 controls

Filter SNVs with a MAF <5% that are predicted to disrupt the protein-coding sequencing:
- ➤ missense
- ➤ nonsense
- ➤ essential splice site

→ Phase one association testing

Selection of candidate causal variants for follow-up genotyping based on:

Significance testing (p < 0.05) (GROUP 1)

A subset of genotypes exclusive to IGE cases (GROUP 2)

Follow-up genotyping of 3,900 variants using a custom-designed iSelect chip (Illumina):
- ➤ 878 IGE cases
- ➤ 1,830 controls

→ Phase two association testing

Analysis/prioritization of candidate rare variants

Identify all variants with alleles or genotypes present in IGE cases only → Prioritize variants that also are absent in NHLBI Exome Sequenced Cohort →
- Assess phenotypic specificity of rare candidate variants
- Pathway analyses
- Assessment of genes with multiple rare variants
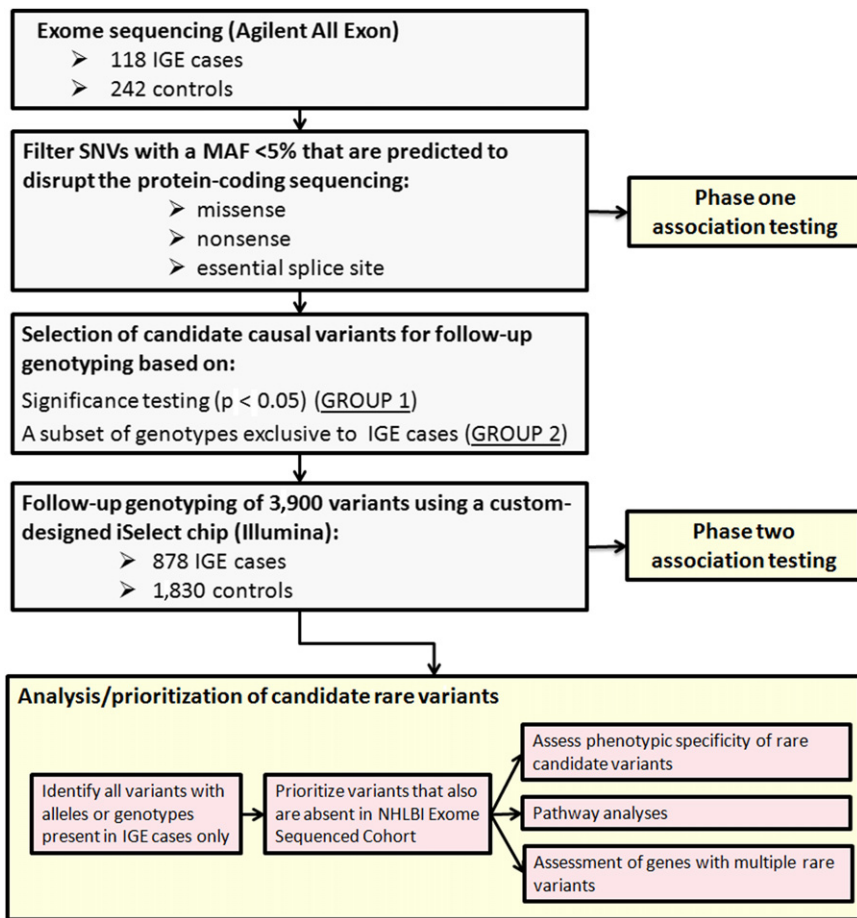
578 EPIGEN IGE-affected individuals without restricting to particular ethnicities, 181 DNA samples (of European ancestry) obtained from the Coriell Cell Repository (National Institute of Neurological Disease and Stroke epilepsy collection), and eight additional unrelated individuals with IGE (contributed by James McNamara and Ruth Ottman). All subjects selected for this study had clinical and electroencephalographical characteristics meeting the 1989 International League against Epilepsy syndrome definitions[13] for childhood AE (MIM 600131), juvenile AE (MIM 607631), JME (MIM 254770), or other unspecified forms of IGE (MIM 600669). The full cohort of 885 IGE-affected individuals, either exome sequenced or genotyped, included 131 individuals with a diagnosis of AE and 288 individuals with a diagnosis of JME. Diagnoses were assigned by the treating physician at each site. Approximately 90% of this follow-up IGE cohort was of European ancestry.

Exome- or whole-genome-sequenced controls consisted of a group of 242 individuals who were sequenced as part of other NGS studies being carried out in the Center for Human Genome Variation (Duke University). All 242 controls were of European ancestry and were not enriched for epilepsy or other neuropsychiatric phenotypes. Controls used for follow-up genotyping consisted of an additional 1,780 subjects who were not enriched for neuropsychiatric phenotypes and who were from the Genetics of Memory and Epilepsy cohort collected in the Center for Human Genome Variation, the Murdock Research Institute cohort, the Center for HIV/AIDS Vaccine Immunology progression cohort, or the high-risk seronegative HIV cohort. Approximately 65% of this follow-up control cohort was of European ancestry. We note that controls were matched for broad ancestral group (European) only.

This study was carried out in compliance with the institutional review board at Duke University and the relevant ethics boards at the collection sites. Informed consent was obtained from all study participants or their legal guardians.

In this study, we test the first approach and report on the results of a NGS study of individuals with idiopathic generalized epilepsy (IGE [MIM 600669], also known as genetic generalized epilepsy), one of the most heritable subtypes of common epilepsy.[11,12]

Subjects and Methods

Study Population

IGE-affected subjects evaluated in this study were largely recruited through the Epilepsy Genetics (EPIGEN) Consortium. Across EPIGEN, subjects are recruited and enrolled by physicians during routine clinical visits across five clinical sites: Erasme Hospital and Universitair Ziekenhuis Gasthuisberg (Belgium), Duke University Medical Center (United States), Beaumont Hospital (Ireland), and the National Hospital for Neurology and Neurosurgery (England). Detailed phenotypic information is collected at the time of enrollment and put into a database in a deidentified manner for the facilitation of genetic association studies. Clinical information is updated periodically when the individual returns to the clinic for routine care and treatment. From the database information, we selected 118 unrelated individuals of European ancestry for the exome-sequencing phase of the study, and these included 93 subjects with a diagnosis of juvenile myoclonic epilepsy (JME [MIM 254770]) and 25 with a diagnosis of absence epilepsy (AE [MIM 607631 and MIM 600131]). For follow-up genotyping of candidate variants, we identified an additional

Study Design and Methods

In this study, we used a two-stage discovery approach that first employed exome sequencing in a series of IGE-affected individuals and a subsequent large-scale follow-up genotyping of identified candidate variants in a larger cohort of IGE-affected subjects and controls (Figure 1).

Exome Sequencing, Alignment, and Variant Calling

Samples were either exome sequenced with Agilent's All Exon (37 Mb or 50 Mb) capture or whole-genome sequenced (n = 51 controls) on Illumina GAIIx or HiSeq 2000 machines in the Genomic Analysis Facility within the Center for Human Genome Variation. Only the regions targeted on the 37 Mb or 50 Mb

Agilent All Exon platform were interrogated from the whole-genome-sequenced samples. Sequencing was performed according to standard protocols. The targeted exonic regions of all sequenced samples were sequenced to an average coverage of 64.3× ± 13, which translated to at least 5× coverage of approximately 86% of the Human Genome Organization (HUGO)-defined protein-coding regions in each subject. Paired-end reads were aligned to the human reference genome (National Center for Biotechnology Information Build 36) with Burrows-Wheeler Aligner software.[14] Variant calling was performed with SAMtools software.[15] SequenceVariantAnalyzer (SVA)[16] was used for annotating variants identified from the sequence data (Ensembl 50_36l). This software provides each variant with a genomic context (nonsynonymous or splice-site coding, gene name, transcript, associated Gene Ontology term, etc.).[16] Association Tests for Annotated Variants (ATAV) software was used for performing Fisher's exact tests for the comparison of the frequencies of variants identified from the sequence data between cases and controls so that candidate variants for follow-up genotyping could be selected.

Identifying Candidate Variants from the Sequence Data
Variants for follow-up genotyping were selected on the basis of variant-call quality, association evidence, and functional annotation. We first isolated variants with a minor allele frequency (MAF) of less than 5% in the control population in transcripts encoding a HUGO protein-coding gene, and only those predicted to affect the protein-coding sequence (these included missense variants, nonsense variants, and single-nucleotide variants [SNVs] residing in the highly conserved essential splice-site region—two base pairs into an intron or one base pair into the exon at the intron-exon junction). We then eliminated autosomal variants that were out of Hardy-Weinberg equilibrium in controls ($p < 0.001$), X chromosome variants that were out of Hardy-Weinberg equilibrium in females only ($p < 0.001$), sex-chromosome variants called as heterozygotes in males, and variants with less than 10× coverage in more than 50% of sequenced subjects. These quality-control steps left us with a set of high-quality variants that were annotated as functional.

We then selected the subset of variants that were statistically significantly enriched in cases (uncorrected $p < 0.05$, Fisher's exact test, allelic or recessive model, GROUP 1) and those with alleles or genotypes exclusive to cases (GROUP 2) if they satisfied at least one of the following selection criteria: (1) the variant genotype was observed in more than one individual with IGE, (2) the variant was located in a region known to harbor pathogenic CNVs implicated in neuropsychiatric disease, or (3) for variants with genotypes observed in one case only, the variant met more stringent criteria for functional effect, i.e., variants that were predicted to disrupt an essential splice site, variants that targeted the transcript for nonsense-mediated decay, and nonsynonymous variants predicted by Polyphen-2[17] to change protein function with high confidence ("probably damaging"). SNVs composing GROUP 2 were also screened in an additional 296 sequenced controls of other ethnicities (non-European ancestry) and were eliminated if the genotype was observed.

Follow-Up Genotyping
Candidate variants were genotyped in additional cases and controls with a custom designed iSelect genotyping chip (Illumina). Of the 3,897 selected variants, 3,626 met the design standards, permitting them to be included on the custom-designed chip. Variants were genotyped in 878 IGE cases and 1,830 controls, including 111 of the 118 originally sequenced cases and 50 of the 242 originally sequenced controls. In a separate

effort to identify susceptibility variants in schizophrenia (see Need et al.[18] in this issue of *AJHG*) and to assess phenotypic specificity of the candidate epilepsy-susceptibility variants being evaluated, we also included individuals with schizophrenia (MIM 181500) in this follow-up genotyping study. After genotyping, the raw data from the custom genotyping chip were evaluated with Illumina's GenomeStudio software. Any samples with a call rate below 0.95 were excluded from analysis. Several additional quality-control steps were also taken for ensuring genotyping accuracy. First, genders determined from genotyping were compared to the record of sample genders. Samples with discordant gender calls were excluded from further analysis. Next, genotype concordance was checked for all samples that had been exome or whole-genome sequenced. Any variant that had discordant calls between the genotyping and sequencing data set and that could not be reconciled with inspection of the genotyping or sequencing data was excluded from further analysis. Finally, for subjects who had been previously genotyped on an Illumina genome-wide genotyping chip, we compared the genotyping calls of a set of 170 variants that were genotyped on both platforms and removed samples with <95% concordance.

Candidate variants that were absent in the control population were also evaluated in a cohort of ~5,400 samples exome sequenced as part of the National Heart, Lung, and Blood Institute (NHLBI) Grand Opportunity (GO) Exome Sequencing Project (ESP) for the identification of genes contributing to heart, lung, and blood disorders (Exome Variant Server [EVS], NHLBI ESP, Seattle, WA). In order for a variant to be called present or absent in a study sample, the coverage of the site was required to be at least 10× and any variant had to have a quality score of at least 30.

Association Testing
PLINK[19] was used for performing logistic regression (allelic and recessive models) on variants genotyped in the follow-up genotyping phase of the study (iSelect genotyping) for the identification of variants enriched more in cases than in controls. Genome-wide-association chip data (Illumina) were available on a subset of the genotyped samples, which permitted the calculation of eigenvectors so that population substructure in the logistic regression could be corrected for according to the Eigenstrat method.[20] Association testing for variants residing on the X chromosome was performed in females only.

Results

Exome Sequencing
After exome sequencing, alignment, and a series of quality-control steps, we isolated 97,242 nonsense, missense, or essential splice-site SNVs that were in transcripts encoding a HUGO-defined protein-coding gene and that had a control MAF of less than 5% from the variant profiles identified in the 118 IGE cases and 242 controls. A total of 9,987 SNVs were removed as a result of inconsistency with Hardy-Weinberg equilibrium in controls or excessively high missingness (i.e., insufficiently covered for accurate variant calling), leaving us with 87,255 high-quality functional variants with an estimated population frequency of less than 5%. We performed logistic regressions by using both allelic and recessive models to compare the frequencies of variants in the cases and controls. Given the number of variants tested and the sample size, we had
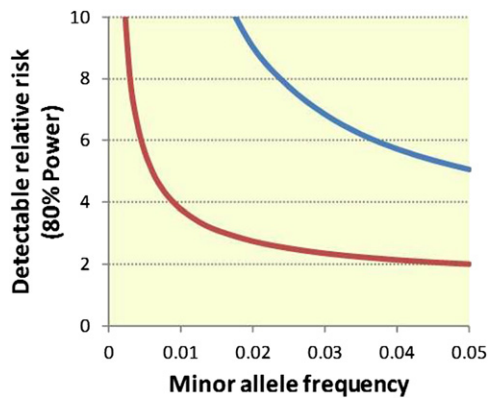
Figure 2. Power for Detecting an Epilepsy-Susceptibility Variant in This Study

The detectable relative risk of an IGE-susceptibility variant across a range of MAFs in the sequencing phase of this study (blue line) is compared to that of the follow-up genotyping phase of this study (red line). To be significant, associations must have a p value below 6×10^{-7}, which is equivalent to $p < 0.05$ when we correct for all 87,255 high-quality, functional variants that were annotated as functional and had a control MAF < 5% in the sequencing study. Power calculations were performed with the CaTS-Power Calculator.[21]

little power to achieve formal significance unless risk alleles were common and/or of major effect (Figure 2). Not surprisingly, there were no significant findings after correction for the 87,255 single-variant association tests performed ($p < 6 \times 10^{-07}$). However, by targeting the subset of variants showing nominal significance in the sequence data in a follow-up cohort of 878 cases and 1,830 controls, we had power to detect variants over a much broader range of parameters (Figure 2). Hence, all candidate IGE-susceptibility variants selected from the exome-sequencing data on the basis of the criteria defined in the Subjects and Methods section were then taken forward for genotyping in the follow-up cohort.

Importantly, we note that no correction was made for population substructure when we selected variants from the sequence data for follow-up genotyping. Because we used a very liberal threshold for inclusion of variants, the impact that not controlling for stratification has on power should be modest. Furthermore, these variants would subsequently be analyzed in further samples where a correction for population stratification was incorporated (see Subjects and Methods).

Because this selection of candidate variants included variants with very modest statistical enrichment or exclusive presence in IGE cases (which could occur if the variant is a risk factor for JME, AE, or both [or neither]), we emphasize that this study design does not presume that the genetic control is constant across all syndromes composing IGE.

Follow-Up Genotyping of Candidate Variants
Variant Validation
In total, 3,897 variants met the defined candidate-variant criteria. Because the reference-mapping approaches used

in this analysis have been shown to generate false positives,[22] we calculated validation rates among our selected variants to evaluate the extent to which this occurred. On average, 96% of SNVs selected from the exome-sequencing study were confirmed in the iSelect follow-up analysis to be present in the variant carrier or in another subject analyzed in this study. We did not see differences in validation rates between functional groups, as previously reported for more common variants,[22] which is most likely due to the highly stringent quality control used in this analysis, as well as the fact that the majority of the variants selected for follow-up analysis were predicted to be deleterious.

Association Testing
Of the 3,897 selected variants, 3,349 were successfully genotyped on a custom-designed iSelect genotyping chip (Illumina) and passed the defined quality-control criteria. The 3,349 SNVs were first tested for association with the IGE phenotype in the larger cohort. In a subset of samples (460 IGE cases versus 814 controls), we applied Eigenstrat axes to control for stratification.[20] The eigenvectors generated from the genome-wide genotyping data were incorporated as covariates into the logistic regression comparing the genotype frequency of the candidate susceptibility variants in cases and controls. The quantile-quantile plot suggests that population substructure has been adequately corrected for in this analysis (Figure 3). Using a Bonferonni correction for all rare, functionally annotated variants identified in the exome-sequencing study (n = 87,255), we found that no variants had a corrected p value below 0.05. Table 1 provides a list of the ten lowest p values showing enrichment in IGE cases under an allelic model. Among these, the most significantly enriched is a nonsynonymous variant (c.103C>G [p.Pro35Ala]; Ensembl accession number ENST00000322805 and RefSeq accession number NM_001191323) located in GREM1. This gene belongs to a family of bone morphogenetic protein antagonists and is believed to play a role in organogenesis and tissue differentiation.[23] This GREM1 SNV is located within 100 kb of the 15q13.3 region that was recently found to harbor large genomic deletions that associate with epilepsy, schizophrenia (MIM 181500), and autism (MIM 209850).[5,9,24,25] No trend associations (uncorrected $p < 0.05$) were detected with a recessive model. Likewise, subanalyses of individuals with a diagnosis of AE (n = 61) and JME (n = 173) revealed no study-wide significant associations.

Evaluation of Variants with Alleles or Genotypes Exclusive to Cases
We next evaluated the 1,863 SNVs for which the variant allele continued to be exclusively present in cases after follow-up genotyping (Table S1, available online). We examined the frequency of these variants in an additional ~5,400 subjects who had their exomes sequenced as part of the NHLBI GO ESP. A total of 1,289 variants remained absent in the NHLBI control population, and 23 were not sufficiently covered for the evaluation of frequency (noted
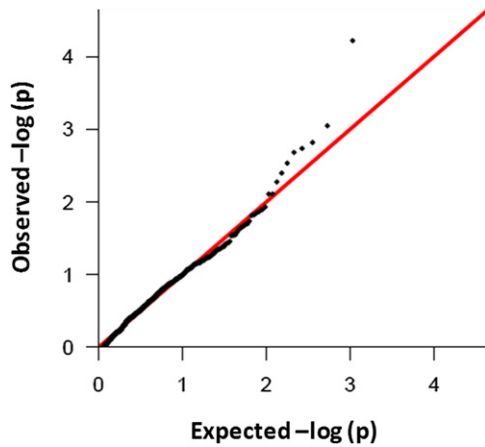
Figure 3. Quantile-Quantile Plot Shows No Evidence of Population Stratification

A quantile-quantile plot of transformed p values (black dots) against the expected transformed p values for variants with at least six alleles represented in the study population. The red line indicates the expectation under the null model of no effect on risk.

in Table S1). Among these 1,289 rare candidate variants is a missense mutation (c.620C>T [p.Ala207Val]; Ensembl ENST00000396881) in *PSME2* (MIM 602161); this mutation was observed in five unrelated IGE cases (Table 2). *PSME2* encodes the proteasome activator complex subunit 2 and is involved in antigen processing and the assembly of the immunoproteasome.[26] The encoded protein has no reported role in epilepsy pathophysiology; however, expression of the gene has been shown to be upregulated in mesial temporal-lobe epilepsy.[27]

In addition, among this set of candidate variants are two heterozygous variants—c.1205A>G (p.Asn402Ser) (Ensembl ENST00000303498) in *BTD* (MIM 609019) and c.740C>G (p.Pro247Arg) (Ensembl ENST00000244546 and RefSeq NM_000060) in *PEX6* (MIM 601498)—in genes responsible for severe, recessive neurological disorders that usually include seizures (e.g., biotinidase deficiency [MIM 253260] and Zellweger syndrome [MIM 214100]). It is possible that a mutation in one copy of these genes might lead to less-severe phenotypes such as epilepsy alone.

Next, we evaluated SNVs that were found in the homozygous state exclusively in individuals with IGE (n = 72, Table 3 and Table S2). No homozygotes were observed in the EVS for 36 of these variants, and eight were not sequenced in the NHLBI cohort (Table S2). One variant in *AGPAT3*, encoding an acyltransferase involved in the phospholipid biosynthetic pathway,[28] was homozygous in three unrelated IGE cases. This gene does not have any known direct involvement in epilepsy.

In the interest of evaluating the phenotypic specificity of these candidate epilepsy-susceptibility variants, we have provided the number of JME and AE carriers in Tables S2 and S3. Even though we considered these to be very specific phenotypes, we still failed to identify study-wide significant associations.

Table 1. List of Variants Enriched in 460 IGE-Affected Individuals of European Ancestry

Variant (chr_hg18 position_ allele)	SNP ID (dbSNP 129/Ensembl 50_36l)	Gene	MIM Number	Transcript (Ensembl 50_36l)	Transcript (RefSeq)	Position of Sequence and Amino Acid Change	Annotated Function	IGE Counts (Hom/Het/Reference)	Control Counts (Hom/Het/Reference)	Odds Ratio	Uncorrected p Value	Corrected p Value[a]
15_30810286_G	–	GREM1	603054	ENST00000322805	NM_001191323.1	c.103C>G (p.Pro35Ala)	NS	0/16/444	0/4/811	9.50	0.0002	NS
11_123353274_A	rs17759447	OR10S1	–	ENST00000360721	–	c.335C>T (p.Ala112Val)	NS	0/25/435	0/11/804	3.64	0.001	NS
4_77032104_T	–	PPEF2	602256	ENST00000337500	–	c.107G>A (p.Arg36His)	NS	0/19/437	0/9/806	3.77	0.003	NS
6_139274117_A	–	REPS1	–	ENST00000367663	–	c.1900C>T (p.Pro634Ser)	NS	0/3/457	0/1/813	28.8	0.006	NS
5_98220168_A	–	CHD1	602118	ENST00000284049	NM_001270.2	c.4949C>T (p.Thr1650Met)	NS	0/5/455	0/2/813	14.3	0.007	NS
11_19203863_C	–	E2F8	612047	ENST00000250024	NM_024680.3	c.2018C>G (p.Pro673Arg)	NS	0/3/457	0/1/814	33.21	0.008	NS
1_100379032_A	–	CCDC76	–	ENST00000370141	NM_019083.2	c.538C>A (p.Leu180Ile)	NS	0/2/458	0/1/814	40.6	0.008	NS
5_178487703_T	rs35445112	ADAMTS2	604539	ENST00000251582	NM_014244.4	c.2480G>A (p.Arg827Gln)	NS	0/47/413	0/54/761	1.84	0.009	NS
19_57401117_T	–	PPP2R1A	605983	ENST00000322088	NM_014225.5	c.259C>T (p.His87Tyr)	NS	0/7/453	0/4/811	5.61	0.015	NS
1_212642864_C	–	PTPN14	603155	ENST00000366956	NM_005401.4	c.566A>G (p.Glu189Gly)	NS	1/33/426	0/30/785	1.99	0.016	NS

The following abbreviations are used: IGE, idiopathic generalized epilepsy; NS, nonsynonymous; hom, homozygous; and het, heterozygous.
[a]Corrected for 87,255 high-quality, putatively functional SNVs seen in the sequencing study.

Table 2. SNVs Present in at Least Three IGE-Affected Individuals, Absent in the Control Population Evaluated in This Study, and Absent from the NHLBI Exome-Sequenced Cohort

Variant (chr_hg18 position_ allele)	Gene	MIM Number	Transcript (Ensembl 50_36l)	Transcript (RefSeq)	Position of Sequence and Amino Acid Change	Annotated Function	IGE Counts[a] (Hom/Het/ Reference)	Control Counts[a] (Hom/Het/ Reference)	p Value[b]	MAF (%)/Number of Subjects (NHLBI Cohort)
14_23682683_A	PSME2	602161	ENST00000396881	–	c.620C>T (p.Ala207Val)	NS	0/5/874	0/0/2,021	0.003	0/5,379
7_30760037_A	INMT	604854	ENST00000013222	NM_006774.4	c.320G>A (p.Trp107*)	SG	0/3/875	0/0/2,005	0.028	0/5,379
7_150410267_T	TMUB1	–	ENST00000297533	NM_031434.3	c.317G>A (p.Arg106Gln)	NS	0/3/876	0/0/2,011	0.028	0/5,379
2_235615236_G	SH3BP4	605611	ENST00000322950	–	c.865C>G (p.Pro289Ala)	NS	0/3/875	0/0/2,020	0.028	0/5,378
3_46425043_A	CCRL2	608379	ENST00000399036	NM_003965.4	c.469G>A (p.Val157Ile)	NS	0/3/875	0/0/1,981	0.028	0/4,927
6_43054127_C	PEX6	601498	ENST00000244546	–	c.740C>G (p.Pro247Arg)	NS	0/3/875	0/0/2,022	0.029	0/5,379
3_135148919_T	SLCO2A1	601460	ENST00000310926	NM_005630.2	c.1166G>A (p.Arg389His)	NS	0/3/875	0/0/2,022	0.029	0/5,379
3_15661572_G	BTD	609019	ENST00000303498	NM_000060.2	c.1205A>G (p.Asn402Ser)	NS	0/3/875	0/0/2,022	0.029	0/5,379
13_47884428_T	P2RY5	609239	ENST00000345941	NM_001162498.1	c.133G>A (p.Val45Ile)	NS	0/3/875	0/0/2,022	0.029	0/5,379
5_153746175_G	GALNT10	608043	ENST00000377661	–	c.862T>G (p.Ser288Ala)	NS	0/3/876	0/0/2,022	0.029	0/5,379
4_157080116_T	CTSO	600550	ENST00000281527	–	c.409G>A (p.Val137Met)	NS	0/3/876	0/0/2,022	0.029	0/5,379
19_45206312_A	ZNF546	–	ENST00000347077	NM_178544.3	c.394G>A (p.Asp132Asn)	NS	0/3/875	0/0/2,022	0.029	0/5,379

A full list of variants with IGE-specific alleles is provided in Table S2. The following abbreviations are used: IGE, idiopathic generalized epilepsy; MAF, minor allele frequency; NS, nonsynonymous; and SG, stop gained.
[a]Counts include both sequenced and iSelect genotyped samples.
[b]Fisher's exact test comparing carriers to noncarriers in the study cohort only (i.e., excluding NHLBI exome-sequenced samples).

Table 3. SNVs Called as Homozygotes in More than Two IGE-Affected Individuals and Absent in Homozygous Form in Controls

Variant (chr_hg18 position_ allele)	SNP ID (dbSNP 129/ Ensembl 50_36l)	Gene	MIM Number	Transcript (Ensembl 50_36l)	Transcript (RefSeq)	Position of Sequence and Amino Acid Change	Annotated Function	IGE Counts[a] (Hom/Nonhom)	Control Counts (Hom/Nonhom)	p Value[b]	Homozygotes/ Nonhomozygotes (NHLBI Cohort)
21_44215700_T	ENSSNP7256682	AGPAT3	–	ENST00000398058	NM_020132.4	c.668C>T (p.Ala223Val)	NS	3/876	0/1,987	0.028	0/5,379
4_113572281_T	rs35633116	ALPK1	607347	ENST00000177648	NM_025144.3	c.2129C>T (p.Ser710Phe)	NS	2/876	0/1,994	0.094	0/5,379
3_38540580_T	rs1141223	ENDOGL1	604051	ENST00000287675	NM_005107.3	c.830G>T (p.Gly277Val)	NS	2/877	0/1,994	0.094	0/5,379
X_90577242_A	–	PABPC5	300407	ENST00000402906	–	c.10G>A (p.Gly4Arg)	NS	3/874	0/1,095	0.088	0/5,379

A full list of variants with homozygous genotypes seen only in cases is provided in Table S3. The following abbreviations are used: IGE, idiopathic generalized epilepsy; NS, nonsynonymous; hom, homozygous; and nonhom, nonhomozygous.
[a]Counts include both sequenced and iSelect genotyped samples.
[b]p values compare counts of homozygotes (or hemizygotes) to those of nonhomozygotes (or nonhemizygotes).

Pathway Analysis of Genes Harboring Candidate Epilepsy-Susceptibility Variants

We performed a pathway analysis including all 1,183 genes harboring a SNV with a genotype exclusive to individuals with IGE, JME, or AE (and also absent from the NHLBI control cohort) by using the Reactome software.[29] After applying a Bonferroni correction for the number of tests performed, we observed no pathways significantly over-represented in any phenotypic group (detailed pathway results for genes associated with the IGE phenotype are provided in Table S3).

Discussion

In this study, we evaluated 3,349 candidate epilepsy-susceptibility variants (identified in an exome-sequencing study of 118 IGE cases) in a larger cohort of 878 cases and 1,830 controls and did not find any variants with statistically significant associations. One possible explanation for the lack of significant findings is that epilepsy-susceptibility variants are sufficiently rare that each one accounts for only a small fraction of individuals with epilepsy. This finding is consistent with that observed with CNVs in epilepsy disorders in that deletions at 15q13.3 and 16p13.11—despite the fact that they explain a significant proportion of cases—are very rare when one considers each identical-by-descent deletion event.[3,4,9,24,30] If this supposition is true, then among the list of candidate susceptibility variants identified in this sequencing study is a set of very rare disease-causing variants. Of the 3,349 variants studied, 1,325 variants or homozygous genotypes, including 103 that were observed in additional IGE cases in the follow-up phase and 118 that are located in genes previously implicated in linkage studies of individuals with IGE, remain exclusively present in cases (they are absent from both the current study control cohort and the NHLBI cohort) (Tables S1 and S4).

Under the assumption that some of these candidate rare variants are causal (Tables S2 and S3), we can use our data to estimate the extent of genetic heterogeneity in IGE. The variant observed most commonly in the IGE cases studied here (c.620C>T [p.Ala207Val] in *PSME2*; 14_23682683_A; Ensembl ENST00000396881; Table 2) appears to account for approximately 0.6% of the studied cases, assuming that the variant ultimately proves to be causal. When we consider variants with genotypes exclusively seen in individuals with JME and absent in all control populations and study participants with a diagnosis of AE or schizophrenia, no single variant identified in our study can explain more than 1% of individuals with JME. Likewise, no single variant can explain more than 1.5% of AE cases evaluated in this study. Collectively, this suggests that no single-base substitution affecting the protein (represented in the exome data) causes epilepsy in a large percentage of IGE cases (or JME or AE cases), a finding consistent with a model of high genetic heterogeneity, probably including both

locus and allelic heterogeneity. It is also likely that gene-level association analyses might effectively guide us toward epilepsy-susceptibility genes that would be readily missed in the single-variant analyses in this study as a result of allelic heterogeneity. Consistent with this hypothesis, we note that many of the identified case-only variants (absent in both this study cohort and the NHLBI cohort) reside in the same gene (Tables S1 and S2). In total, 115 genes, including seven genes with a rare variant in more than three IGE-affected individuals and two genes with a rare variant in six IGE study subjects, have multiple rare variants in IGE cases only (Table S5). Assuming that allelic heterogeneity might contribute in part to the overall genetic heterogeneity associated with IGE, future studies should focus on gene-level association analyses.

It is possible that beyond locus and allelic heterogeneity, other multilocus models might account for the negative results we observed in our study. Although oligogenic, polygenic, or epistatic modes of disease-risk inheritance are possible, statistical proof for these multivariant models has to withstand very high thresholds of multiple testing to be detected in NGS studies. It is likely that if these models account for some fraction of disease risk, they will only be secondarily identified in NGS studies if marginal effects of single variants can be detected in very large sample sizes.

Although it does not yet provide definitive evidence of specific risk factors, this work makes a number of important contributions to epilepsy genetics. First and most fundamentally, this work strongly suggests that moderately rare variants with intermediate effects ("goldilocks alleles"[31]) do not play a major role in the risk of IGE. In this study, we had 80% power to detect a putatively functional coding SNV with a frequency of 0.5% and a relative risk of 5.4. However, power decreases as the variants become rarer or carry less relative risk. Although the power for obtaining secure evidence for an association of any individual variant with MAFs between 0.1% and 0.5% in this study is only modest (for example, ~10% power for detecting a SNV with a MAF of 0.3% and a relative risk of 5), we would have detected some of them here if there were many such IGE-susceptibility alleles in the human gene pool. This suggests that discovery will require designs that can provide evidence of variants that are outside this range in terms of either effect size or allele frequency or both.

Second, this study provides a candidate gene list that is likely to include real IGE risk factors. Interestingly, the pathway analysis of the genes harboring candidate epilepsy-susceptibility variants revealed no evidence of enrichment of genes encoding ion channels or ion-channel modifiers (Table S3). This lack of enrichment of protein-disrupting ion-channel mutations in individuals with epilepsy is consistent with what was observed by Klassen et al.[32] Collectively, these findings suggest that the pathophysiology governing epilepsy might be far more complex than simply a disorder of disrupted ion channels ("channelopathy"), as was suggested by early studies of Mendelian epilepsies.

Third, the results of this study provide a benchmark against which researchers can compare small studies that are likely to emerge in the near future and claim evidence of pathogenicity. Given the near comprehensive assessment of coding variation reported here, pathogenicity claims based on small discovery samples need to be treated with caution. Finally, under the assumption that some of the variants reported here as exclusive to epilepsy cases are real risk factors, these results give us the opportunity to estimate how many samples will be needed for statistically significant evidence of any single associated variant. By assuming that the frequency of the risk alleles observed in our follow-up sample is correct and by simply scaling up the sample sizes, we can determine at what point we expect to achieve $p < 6 \times 10^{-7}$, which would be sufficiently low to account for the testing of the >100,000 variants identified in the sequenced cohort. In this study, candidate variants were seen at a frequency range of 0.06%–0.3% among study subjects with IGE. If these frequencies are consistent in follow-up cohorts, we would need equivalently sized case and control cohorts of between 2,000 and 6,000 individuals each to prove the association of variants seen at a frequency of 0.1%–0.5% of IGE cases. Importantly, for variants that are seen in only one case (those that have a frequency of approximately 0.06%), proving pathogenicity is virtually out of reach even with large sample sizes; such variants will probably only be securely implicated through gene-based association analyses in large sample sizes and, where available, cosegregation analyses within multiplex families.[33]

Supplemental Data

Supplemental Data include five tables and can be found with this article online at http://www.cell.com/AJHG.

Web Resources

The URLs for data presented herein are as follow:

ATAV, http://www.duke.edu/~minhe/atav/
HUGO Gene Nomenclature Committee, http://www.genenames.org
NHLBI Exome Sequencing Project Exome Variant Server, http://snp.gs.washington.edu/EVS
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org
Picard, http://picard.sourceforge.net
Plink, http://pngu.mgh.harvard.edu/~purcell/plink/
Reactome, http://www.reactome.org
SequenceVariantAnalyzer, http://www.svaproject.org

References

1. Need, A.C., and Goldstein, D.B. (2010). Whole genome association studies in complex diseases: Where do we stand? Dialogues Clin. Neurosci. *12*, 37–46.
2. Aldinger, K.A. (2009). Copy number variation at 1q21.1 results in a spectrum of developmental abnormalities. Clin. Genet. *75*, 425–427.
3. Heinzen, E.L., Radtke, R.A., Urban, T.J., Cavalleri, G.L., Depondt, C., Need, A.C., Walley, N.M., Nicoletti, P., Ge, D., Catarino, C.B., et al. (2010). Rare deletions at 16p13.11 predispose to a diverse spectrum of sporadic epilepsy syndromes. Am. J. Hum. Genet. *86*, 707–718.
4. Mefford, H.C., Muhle, H., Ostertag, P., von Spiczak, S., Buysse, K., Baker, C., Franke, A., Malafosse, A., Genton, P., Thomas, P., et al. (2010). Genome-wide copy number variation in epilepsy: Novel susceptibility loci in idiopathic generalized and focal epilepsies. PLoS Genet. *6*, e1000962.
5. Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O.P., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J.E., et al.; GROUP. (2008). Large recurrent microdeletions associated with schizophrenia. Nature *455*, 232–236.
6. Need, A.C., Ge, D., Weale, M.E., Maia, J., Feng, S., Heinzen, E.L., Shianna, K.V., Yoon, W., Kasperavičiūte, D., Gennarelli, M., et al. (2009). A genome-wide investigation of SNPs and CNVs in schizophrenia. PLoS Genet. *5*, e1000373.
7. Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A., Green, T., et al.; Autism Consortium. (2008). Association between microdeletion and microduplication at 16p11.2 and autism. N. Engl. J. Med. *358*, 667–675.
8. de Kovel, C.G., Trucks, H., Helbig, I., Mefford, H.C., Baker, C., Leu, C., Kluck, C., Muhle, H., von Spiczak, S., Ostertag, P., et al. (2010). Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. Brain *133*, 23–32.
9. Helbig, I., Mefford, H.C., Sharp, A.J., Guipponi, M., Fichera, M., Franke, A., Muhle, H., de Kovel, C., Baker, C., von Spiczak, S., et al. (2009). 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. Nat. Genet. *41*, 160–162.
10. Vacic, V., McCarthy, S., Malhotra, D., Murray, F., Chou, H.H., Peoples, A., Makarov, V., Yoon, S., Bhandari, A., Corominas, R., et al. (2011). Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. Nature *471*, 499–503.
11. Hemminki, K., Li, X., Johansson, S.E., Sundquist, K., and Sundquist, J. (2006). Familial risks for epilepsy among siblings based on hospitalizations in Sweden. Neuroepidemiology *27*, 67–73.
12. Berkovic, S.F., Howell, R.A., Hay, D.A., and Hopper, J.L. (1998). Epilepsies in twins: Genetics of the major epilepsy syndromes. Ann. Neurol. *43*, 435–445.
13. Commission on Classification and Terminology of the International League Against Epilepsy. (1989). Proposal for revised classification of epilepsies and epileptic syndromes. Epilepsia *30*, 389–399.
14. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
15. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
16. Ge, D., Ruzzo, E.K., Shianna, K.V., He, M., Pelak, K., Heinzen, E.L., Need, A.C., Cirulli, E.T., Maia, J.M., Dickson, S.P., et al. (2011). SVA: Software for annotating and visualizing sequenced human genomes. Bioinformatics *27*, 1998–2000.
17. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
18. Need, A.C., McEvoy, J.P., Gennarelli, M., Heinzen, E.L., Ge, D., Maia, J.M., Shianna, K.V., He, M., Cirulli, E.T., Gumbs, C.E., et al. (2012). Exome Sequencing Followed by Large-Scale Genotyping Suggests a Limited Role for Moderately Rare Risk Factors of Strong Effect in Schizophrenia. Am. J. Hum. Genet. Published online August 2, 2012. http://dx.doi.org/10.1016/j.ajhg.2012.06.018.
19. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
20. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
21. Skol, A.D., Scott, L.J., Abecasis, G.R., and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat. Genet. *38*, 209–213.
22. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al.; 1000 Genomes Project Consortium. (2012). A systematic survey of loss-of-function

variants in human protein-coding genes. Science *335*, 823–828.

23. Michos, O., Panman, L., Vintersten, K., Beier, K., Zeller, R., and Zuniga, A. (2004). Gremlin-mediated BMP antagonism induces the epithelial-mesenchymal feedback signaling controlling metanephric kidney and limb organogenesis. Development *131*, 3401–3410.

24. Sharp, A.J., Mefford, H.C., Li, K., Baker, C., Skinner, C., Stevenson, R.E., Schroer, R.J., Novara, F., De Gregori, M., Ciccone, R., et al. (2008). A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nat. Genet. *40*, 322–328.

25. Pagnamenta, A.T., Wing, K., Sadighi Akha, E., Knight, S.J., Bölte, S., Schmötzer, G., Duketis, E., Poustka, F., Klauck, S.M., Poustka, A., et al.; International Molecular Genetic Study of Autism Consortium. (2009). A 15q13.3 microdeletion segregating with autism. Eur. J. Hum. Genet. *17*, 687–692.

26. Murata, S., Udono, H., Tanahashi, N., Hamada, N., Watanabe, K., Adachi, K., Yamano, T., Yui, K., Kobayashi, N., Kasahara, M., et al. (2001). Immunoproteasome assembly and antigen presentation in mice lacking both PA28alpha and PA28beta. EMBO J. *20*, 5898–5907.

27. Mishto, M., Ligorio, C., Bellavista, E., Martucci, M., Santoro, A., Giulioni, M., Marucci, G., and Franceschi, C. (2011). Immunoproteasome expression is induced in mesial temporal lobe epilepsy. Biochem. Biophys. Res. Commun. *408*, 65–70.

28. Shindou, H., and Shimizu, T. (2009). Acyl-CoA:lysophospholipid acyltransferases. J. Biol. Chem. *284*, 1–5.

29. Vastrik, I., D'Eustachio, P., Schmidt, E., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., et al. (2007). Reactome: A knowledge base of biologic pathways and processes. Genome Biol. *8*, R39.

30. Mefford, H.C., and Mulley, J.C. (2010). Genetically complex epilepsies, copy number variants and syndrome constellations. Genome Med. *2*, 71.

31. Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. *86*, 832–838.

32. Klassen, T., Davis, C., Goldman, A., Burgess, D., Chen, T., Wheeler, D., McPherson, J., Bourquin, T., Lewis, L., Villasana, D., et al. (2011). Exome sequencing of ion channel genes reveals complex profiles confounding personal risk assessment in epilepsy. Cell *145*, 1036–1048.

33. The Epi4K Consortium. (2012). Epi4K: Gene Discovery in 4,000 Genomes. Epilepsia. Published online May 29, 2012. http://dx.doi.org/10.1111/j.1528-1167.2012.03511.x.