LETTER

doi:10.1038/nature12439

De novo mutations in epileptic encephalopathies

Epi4K Consortium* & Epilepsy Phenome/Genome Project*

Epileptic encephalopathies are a devastating group of severe childhood epilepsy disorders for which the cause is often unknown[1]. Here we report a screen for *de novo* mutations in patients with two classical epileptic encephalopathies: infantile spasms ($n = 149$) and Lennox–Gastaut syndrome ($n = 115$). We sequenced the exomes of 264 probands, and their parents, and confirmed 329 *de novo* mutations. A likelihood analysis showed a significant excess of *de novo* mutations in the ~4,000 genes that are the most intolerant to functional genetic variation in the human population ($P = 2.9 \times 10^{-3}$). Among these are *GABRB3*, with *de novo* mutations in four patients, and *ALG13*, with the same *de novo* mutation in two patients; both genes show clear statistical evidence of association with epileptic encephalopathy. Given the relevant site-specific mutation rates, the probabilities of these outcomes occurring by chance are $P = 4.1 \times 10^{-10}$ and $P = 7.8 \times 10^{-12}$, respectively. Other genes with *de novo* mutations in this cohort include *CACNA1A*, *CHD2*, *FLNA*, *GABRA1*, *GRIN1*, *GRIN2B*, *HNRNPU*, *IQSEC2*, *MTOR* and *NEDD4L*. Finally, we show that the *de novo* mutations observed are enriched in specific gene sets including genes regulated by the fragile X protein ($P < 10^{-8}$), as has been reported previously for autism spectrum disorders[2].

Genetics is believed to have an important role in many epilepsy syndromes; however, specific genes have been discovered in only a small proportion of cases. Genome-wide association studies for both focal and generalized epilepsies have revealed few significant associations, and rare copy number variants explain only a few per cent of cases[3–6]. An emerging paradigm in neuropsychiatric disorders is the major impact that *de novo* mutations have on disease risk[7,8]. We searched for *de novo* mutations associated with epileptic encephalopathies, a heterogeneous group of severe epilepsy disorders characterized by early onset of seizures with cognitive and behavioural features associated with ongoing epileptic activity. We focused on two 'classical' forms of epileptic encephalopathies: infantile spasms and Lennox–Gastaut syndrome, recognizing that some patients with infantile spasms progress to Lennox–Gastaut syndrome.

Exome sequencing of 264 trios (Methods) identified 439 putative *de novo* mutations. Sanger sequencing confirmed 329 *de novo* mutations (Supplementary Table 2), and the remainder were either false positives, a result of B-cell immortalization, or in regions where the Sanger assays did not work (Supplementary Table 3).

Across our 264 trios, we found nine genes with *de novo* single nucleotide variant (SNV) mutations in two or more probands (*SCN1A*, $n = 7$; *STXBP1*, $n = 5$; *GABRB3*, $n = 4$; *CDKL5*, $n = 3$; *SCN8A*, $n = 2$; *SCN2A*, $n = 2$; *ALG13*, $n = 2$; *DNM1*, $n = 2$; and *HDAC4*, $n = 2$). Of these, *SCN1A*, *STXBP1*, *SCN8A*, *SCN2A* and *CDKL5* are genes that have a previously established association with epileptic encephalopathy[9–14]. To assess whether the observations in the other genes implicate them as risk factors for epileptic encephalopathies, we determined the probability of seeing multiple mutations in the same gene given the sequence-specific mutation rate, size of the gene, and the number and gender of patients evaluated in this study (Methods). The number of observed *de novo* mutations in *HDAC4* and *DNM1* are not yet significantly greater than the null expectation. However, observing four unique *de novo* mutations in *GABRB3* and two identical *de novo* mutations in *ALG13* were found to be highly improbable (Table 1 and Fig. 1). We performed the same calculations on all of the genes with multiple *de novo* mutations observed in 610 control trios and found no genes with a significant excess of *de novo* mutations (Supplementary Table 4). Although mutations in *GABRB3* have previously been reported in association with another type of epilepsy[15], and *in vivo* mouse studies suggest that *GABRB3* haploinsufficiency is one of the causes of epilepsy in Angelman's syndrome[16], our observations implicate it, for the first time, as a single-gene cause of epileptic encephalopathies and provide the strongest evidence to date for its association with any epilepsy. Likewise, *ALG13*, an X-linked gene encoding a subunit of the uridine diphosphate-*N*-acetylglucosamine transferase, was previously shown to carry a novel *de novo* mutation in a male patient with a severe congenital glycosylation disorder with microcephaly, seizures and early lethality[17]. Furthermore, the same *ALG13* *de novo* mutation identified in this study was observed as a *de novo* mutation in an additional female patient with severe intellectual disability and seizures[18].

Each trio harboured on average 1.25 confirmed *de novo* mutations, with 181 probands harbouring at least one. Considering only *de novo* SNVs, each trio harboured on average 1.17 *de novo* mutations (Supplementary Fig. 1). Seventy-two per cent of the confirmed *de novo* SNV mutations were missense and 7.5% were putative loss-of-function (splice donor, splice acceptor, or stop-gain mutations). Compared to rates of these classes of mutations previously reported in controls (69.4% missense and 4.2% putative loss-of-function mutations)[2,19,20], we observed a significant excess of loss-of-function mutations in patients with infantile spasms and Lennox–Gastaut syndrome (exact binomial $P = 0.01$), consistent with data previously reported in autism spectrum disorder[2,8,19,20].

A framework was recently established for testing whether the distribution of *de novo* mutations in affected individuals differs from the general population[8]. Here, we extend the simulation-based approach of ref. 8 by developing a likelihood model that characterizes this effect and describes the distribution of *de novo* mutations among affected individuals in terms of the distribution in the general population, and a set of parameters describing the genetic architecture of the disease. These parameters include the proportion of the exome sequence that can carry disease-influencing mutations (η) and the relative risk (γ) of the mutations (Supplementary Methods). Consistent with what was reported in autism spectrum disorder[8], we found no significant deviation in the overall distribution of mutations from expected ($\gamma = 1$ and/or $\eta = 0$). It is, however, now well established that some genes tolerate protein-disrupting mutations without apparent adverse phenotypic consequences, whereas others do not[21]. To take this into account, we used a simple scoring system that uses polymorphism data in the human population to assign a tolerance score to every considered gene (Methods). We then found that genes with a known association with epileptic encephalopathy rank among the most intolerant genes using this scheme (Supplementary Table 8). We therefore evaluated the distribution of *de novo* mutations within these 4,264 genes that are within the 25th percentile for intolerance and found a significant shift from the null distribution ($P = 2.9 \times 10^{-3}$). The maximum likelihood estimates of η (percentage of intolerant genes involved in epileptic encephalopathies) was 0.021 and γ (relative risk) was 81, indicating that there are 90 genes among the intolerant genes

*A list of authors and affiliations appears at the end of the paper.

Table 1 | Probability of observing the reported number of *de novo* mutations by chance in genes recurrently mutated in this cohort

Gene	Chromosome	Average effectively captured length (bp)	Weighted mutation rate	*De novo* mutation number	*P* value†	
SCN1A	2	6,063.70	1.61×10^{-4}	5‡	1.12×10^{-9}	***
STXBP1	9	1,917.51	6.44×10^{-5}	5	1.16×10^{-11}	***
GABRB3	15	1,206.86	3.78×10^{-5}	4	4.11×10^{-10}	***
CDKL5	X	2,798.38	5.44×10^{-5}	3	4.90×10^{-7}	**
ALG13§	X	475.05	1.03×10^{-5}	2	7.77×10^{-12}	***
DNM1	9	2,323.37	9.10×10^{-5}	2	2.84×10^{-4}	
HDAC4	2	2,649.82	1.16×10^{-4}	2	4.57×10^{-4}	
SCN2A§	2	5,831.21	1.52×10^{-4}	2	1.14×10^{-9}	***
SCN8A	12	5,814.48	1.64×10^{-4}	2	9.14×10^{-4}	

† Adjusted α is equivalent to $0.05/18{,}091 = 2.76 \times 10^{-6}$ (*), $0.01/18{,}091 = 5.53 \times 10^{-7}$ (**) and $0.001/18{,}091 = 5.53 \times 10^{-8}$ (***).
‡ Counts exclude three additional patients with an indel or splice site mutation as these are not accounted for in the mutability calculation.
§ Two *de novo* mutations occur at the same position. The probability of these special cases obtain $P = 7.77 \times 10^{-12}$ and $P = 1.14 \times 10^{-9}$ for *ALG13* and *SCN2A*, respectively (Methods).

that can confer risk of epileptic encephalopathies and that each mutation carries substantial risk. We also found that putatively damaging *de novo* variants in our cohort are significantly enriched in intolerant genes compared with control cohorts (Supplementary Methods).

We next evaluated whether the *de novo* mutations were drawn preferentially from six gene sets (Methods and Supplementary Table 10), including ion channels[22], genes known to cause monogenic disorders with seizures as a phenotypic feature[23], genes carrying confirmed *de novo* mutations in patients with autism spectrum disorder[2,8,19,20] and in patients with intellectual disability[18,24], and FMRP-regulated genes. Taking into account the size of regions with adequate sequencing coverage to detect a *de novo* mutation (Methods), we found significant over-representation for all gene lists in our data (Supplementary Table 10), and no over-representation in controls[2,19,20,24].

To determine possible interconnectivity among the genes carrying a *de novo* mutation, we performed a protein–protein interaction analysis and identified a single network of 71 connected proteins (Fig. 2 and Supplementary Fig. 7). These 71 proteins include six encoded by OMIM reported epileptic encephalopathy genes (http://www.omim.org/) where we identified one or more *de novo* mutations among the epileptic encephalopathy patients in this study. Genes in this protein–protein network were also found to have a much greater probability of overlap with the autism spectrum disorder[2,8,19,20] and severe intellectual disability disorder[18,24] exome sequencing study genes, and with FMRP-associated genes, than genes not in this network (Supplementary Table 11).

In support of a hypothesis that individual rare mutations in different genes may converge on biological pathways, we draw attention to six mutations that all affect subunits of the GABA (γ-aminobutyric acid) ionotropic receptor (four in *GABRB3*, and one each in *GABRA1* and *GABRB1*), and highlight two interactions: *HNRNPU* interacting with *HNRNPH1*, and *NEDD4L* (identified here) binding to *TNK2*, a gene previously implicated in epileptic encephalopathies[25] (Fig. 2). Although the *HNRNPU* mutation observed here is a small insertion/deletion variant (indel) in a splice acceptor site, and therefore probably results in a modified protein, the *HNRNPH1 de novo* mutation is synonymous and thus of unknown functional significance (Supplementary Table 2). Notably, a minigene experiment indicates that this synonymous mutation induces skipping of exon 12 (Supplementary Methods).

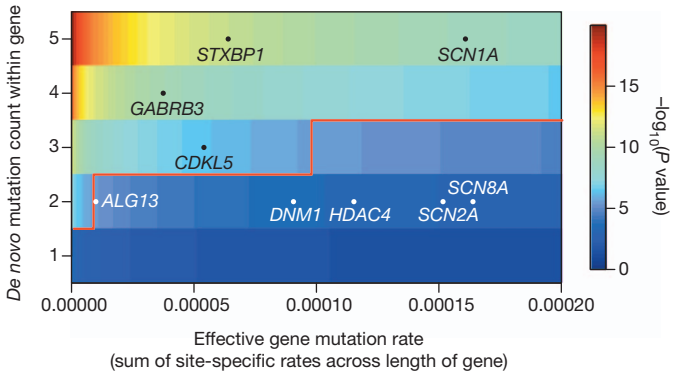

Figure 1 | Heat map illustrating the probability of observing the specified number of *de novo* mutations in genes with the specified estimated mutation rate. The number of *de novo* mutations required to achieve significance is indicated by the solid red line. The superimposed dots reflect positions of all genes found to harbour multiple *de novo* mutations in our study. *GABRB3*, *SCN1A*, *CDKL5* and *STXBP1* have significantly more *de novo* mutations than expected. The positions indicated for *ALG13* and *SCN2A* reflect only the fact that there are two mutations observed, not that there are two mutations affecting the same site (Methods).

Evaluation of the clinical phenotypes among patients revealed significant genetic heterogeneity underlying infantile spasms and Lennox–Gastaut syndrome, and begins to provide information about the range of phenotypes associated with mutations in specific genes (Supplementary Table 13). We identified four genes—*SCN8A*, *STXBP1*, *DNM1* and *GABRB3*—with *de novo* mutations in both patients with infantile spasms and patients with Lennox–Gastaut syndrome. Although infantile spasms may progress to Lennox–Gastaut syndrome, in three of these cases the patients with Lennox–Gastaut syndrome did not initially present with infantile spasms, indicating phenotypic heterogeneity associated with mutations in these genes yet supporting the notion of shared genetic susceptibility. Notably, in multiple patients we identified *de novo* mutations in genes previously implicated in other neurodevelopmental conditions, and in some cases with very distinctive clinical presentations (Supplementary Table 12). Most notably, we found a *de novo* mutation in *MTOR*, a gene recently found to harbour a causal variant in mosaic form in a case with hemimegalencephaly[26]. Our patient however showed no detectable structural brain malformation. Similarly, we found one patient with a *de novo* mutation in *DCX* and another with a *de novo* mutation in *FLNA*, previously associated with lissencephaly and periventricular nodular heterotopia, respectively[27,28]; neither patient had cortical malformations detected on magnetic resonance imaging.

In addition to *de novo* variants, we also screened for highly penetrant genotypes by identifying variants that create newly homozygous, compound heterozygous, or hemizygous genotypes in the probands that are not seen in parents or controls (Supplementary Methods). No inherited variants showed significant evidence of association. Additional studies evaluating a larger number of epileptic encephalopathy patients will be required to establish the role of inherited variants in the disease risk associated with infantile spasms and Lennox–Gastaut syndrome.

We have identified novel *de novo* mutations implicating at least two genes for epileptic encephalopathies, and also describe a genetic architecture that strongly suggests that we have identified additional causal mutations in genes intolerant to functional variation. Given that our sample size already shows many genes with recurrent mutations, it is clear that even modest increases in sample sizes will confirm many new genes now seen in only one of our trios. Our results also emphasize that it may be difficult to predict with confidence the responsible gene, even among known genes, based upon clinical presentation. This makes it clear that the future of genetic diagnostics in epileptic encephalopathies will need to focus on the genome as a whole as opposed to single genes or even gene panels. In particular, several of the genes with *de novo* mutations in our cohort have also been identified in patients with

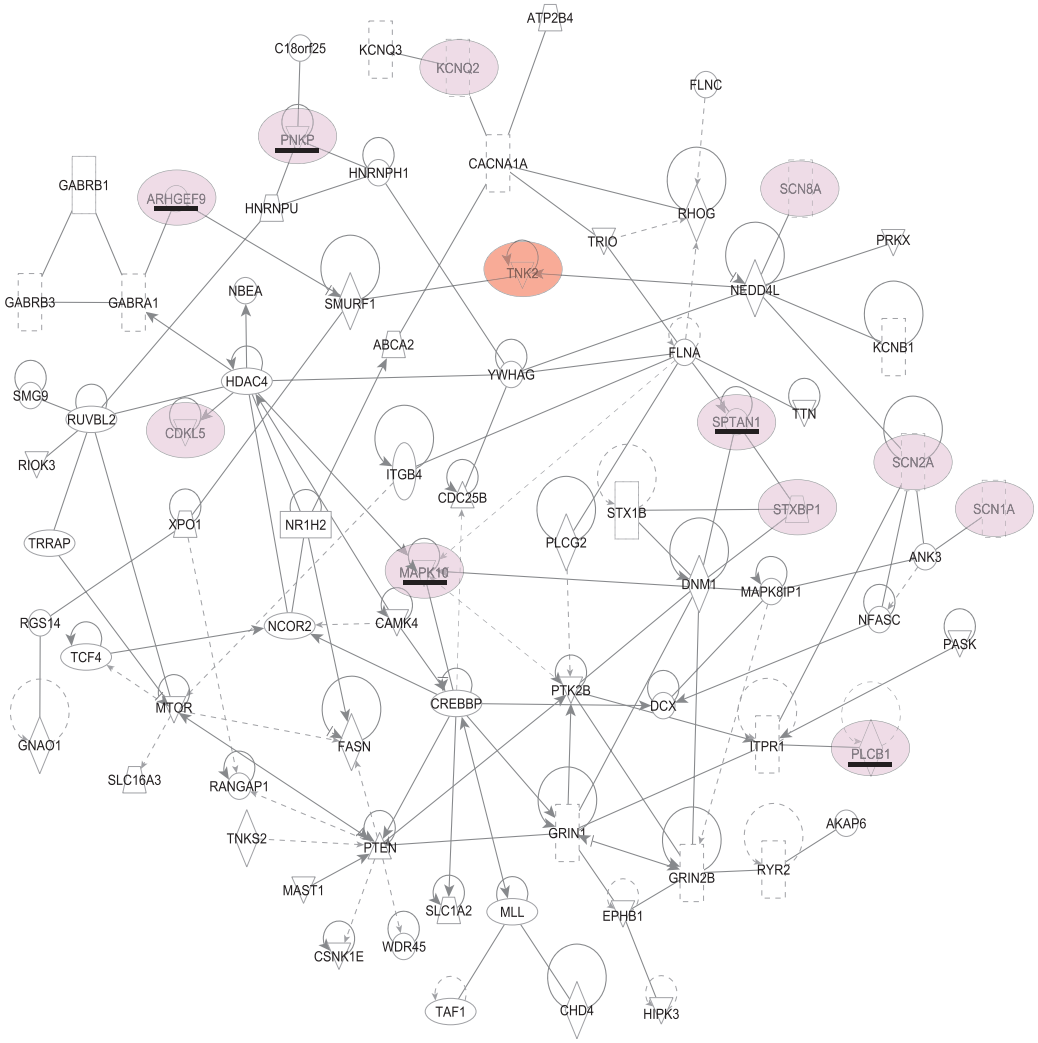

Figure 2 | A protein–protein interaction network of genes with *de novo* mutations found in infantile spasms and Lennox–Gastaut syndrome patients studied. The geometric shapes reflect differing protein roles, as defined by ingenuity pathway analysis (Ingenuity Systems): enzyme, rhombus; ion channel, vertical rectangle; kinase, inverted triangle; ligand-dependent nuclear receptor, horizontal rectangle; phosphatase, triangle; transcription regulator, horizontal oval; transmembrane receptor, vertical oval; transporter, trapezoid; and unknown, circle. Six of the genes found to harbour *de novo* mutations in an infantile spasms or Lennox–Gastaut syndrome patient are known OMIM genes associated with epileptic encephalopathy (shaded circles). Five additional known OMIM genes associated with epileptic encephalopathy that were not found to be mutated in the 264 epileptic encephalopathy patients, but are involved in this network, are also shown (shaded circles with the gene underlined). The previously identified severe infantile epilepsy gene *TNK2* is superimposed into this network (red circle).

intellectual disability or autism spectrum disorder. Finally, and perhaps most importantly, this work suggests a clear direction for both drug development and treatment personalization in the epileptic encephalopathies, as many of these mutations seem to converge on specific biological pathways.

METHODS SUMMARY

All probands and family members were collected as part of the Epilepsy Phenome/Genome Project (EPGP) cohort[29] (Supplementary Table 1). Detailed inclusion and exclusion criteria are provided in Methods. Patient collection and sharing of specimens for research were approved by site-specific Institutional Review Boards.

We sequenced the exome of each trio, from DNA derived from primary cells ($n = 224$ trios) or from lymphoblastoid cell lines (LCLs) in one or more family members ($n = 40$ trios), using the TruSeq Exome Enrichment kit (Illumina). We aligned samples and called variants using established algorithms (Methods) and identified candidate *de novo* variants at sites included in the exons or splice sites of the consensus coding sequence (CCDS) as those called in the affected child and absent in both parents, despite each parent having at least tenfold coverage at the site. Variants created by the *de novo* mutation also had to be absent in our internal controls ($n = 436$), as well as the approximately 6,500 samples represented in the Exome Variant Server (http://evs.gs.washington.edu/EVS), and had to pass visual inspection of alignment quality. Candidate *de novo* mutations were confirmed to be *de novo* mutations using Sanger sequencing. In all cases, primary DNA from the proband was used for the Sanger confirmation so that mutations appearing in the transformation process for the 40 trios sequenced from LCLs would be eliminated.

To determine whether our list of *de novo* mutations was preferentially located in genes contained in the six gene lists we calculated the proportion of CCDS *de novo* mutation opportunity space for each list (Additional Methods). A binomial probability calculation was used to determine whether the *de novo* mutations in CCDS transcripts identified in this cohort of epileptic encephalopathy patients were selectively enriched within the coding sequence of genes within a particular gene list (Supplementary Table 10).

Ingenuity Pathway Analysis (Ingenuity Systems) was used to assess the connectivity of proteins harbouring a *de novo* mutation.

Full Methods and any associated references are available in the online version of the paper.

Received 3 March; accepted 9 July 2013.
Published online 11 August 2013.

1. Berg, A. T. *et al.* Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE Commission on Classification and Terminology, 2005–2009. *Epilepsia* **51,** 676–685 (2010).

2. Iossifov, I. *et al. De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74,** 285–299 (2012).
3. 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).
4. Heinzen, E. L. *et al.* Exome sequencing followed by large-scale genotyping fails to identify single rare variants of large effect in idiopathic generalized epilepsy. *Am. J. Hum. Genet.* **91,** 293–302 (2012).
5. Mulley, J. C. & Mefford, H. C. Epilepsy and the new cytogenetics. *Epilepsia* **52,** 423–432 (2011).
6. Kasperaviciute, D. *et al.* Common genetic variation and susceptibility to partial epilepsies: a genome-wide association study. *Brain* **133,** 2136–2147 (2010).
7. Vissers, L. E. *et al.* A *de novo* paradigm for mental retardation. *Nature Genet.* **42,** 1109–1112 (2010).
8. Neale, B. M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485,** 242–245 (2012).
9. Kalscheuer, V. M. *et al.* Disruption of the serine/threonine kinase 9 gene causes severe X-linked infantile spasms and mental retardation. *Am. J. Hum. Genet.* **72,** 1401–1411 (2003).
10. Claes, L. *et al. De novo* mutations in the sodium-channel gene *SCN1A* cause severe myoclonic epilepsy of infancy. *Am. J. Hum. Genet.* **68,** 1327–1332 (2001).
11. Saitsu, H. *et al. De novo* mutations in the gene encoding STXBP1 (MUNC18–1) cause early infantile epileptic encephalopathy. *Nature Genet.* **40,** 782–788 (2008).
12. Otsuka, M. *et al.* STXBP1 mutations cause not only Ohtahara syndrome but also West syndrome—result of Japanese cohort study. *Epilepsia* **51,** 2449–2452 (2010).
13. Veeramah, K. R. *et al. De novo* pathogenic SCN8A mutation identified by whole-genome sequencing of a family quartet affected by infantile epileptic encephalopathy and SUDEP. *Am. J. Hum. Genet.* **90,** 502–510 (2012).
14. Kamiya, K. *et al.* A nonsense mutation of the sodium channel gene *SCN2A* in a patient with intractable epilepsy and mental decline. *J. Neurosci.* **24,** 2690–2698 (2004).
15. Tanaka, M., DeLorey, T. M., Delgado-Escueta, A. & Olsen, R. W. in *Jasper's Basic Mechanisms of the Epilepsies* (eds Noebels, J. L. *et al.)* (2012).
16. DeLorey, T. M. *et al.* Mice lacking the β_3 subunit of the $GABA_A$ receptor have the epilepsy phenotype and many of the behavioral characteristics of Angelman syndrome. *J. Neurosci.* **18,** 8505–8514 (1998).
17. Timal, S. *et al.* Gene identification in the congenital disorders of glycosylation type l by whole-exome sequencing. *Hum. Mol. Genet.* **21,** 4151–4161 (2012).
18. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367,** 1921–1929 (2012).
19. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485,** 246–250 (2012).
20. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485,** 237–241 (2012).
21. Petrovski, S. *et al.* Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Gen.* (in the press) (2013).
22. Klassen, T. *et al.* Exome sequencing of ion channel genes reveals complex profiles confounding personal risk assessment in epilepsy. *Cell* **145,** 1036–1048 (2011).
23. Lemke, J. R. *et al.* Targeted next generation sequencing as a diagnostic tool in epileptic disorders. *Epilepsia* **53,** 1387–1398 (2012).
24. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380,** 1674–1682 (2012).
25. Hitomi, Y. *et al.* Mutations in TNK2 in severe autosomal recessive infantile-onset epilepsy. *Ann. Neurol.* http://dx.doi.org/doi:10.1002/ana.23934 (2013).
26. Lee, J. H. *et al. De novo* somatic mutations in components of the PI3K–AKT3-mTOR pathway cause hemimegalencephaly. *Nature Genet.* **44,** 941–945 (2012).
27. Gleeson, J. G. *et al.* Doublecortin, a brain-specific gene mutated in human X-linked lissencephaly and double cortex syndrome, encodes a putative signaling protein. *Cell* **92,** 63–72 (1998).
28. Fox, J. W. *et al.* Mutations in filamin 1 prevent migration of cerebral cortical neurons in human periventricular heterotopia. *Neuron* **21,** 1315–1325 (1998).
29. The EPGP Collaborative. The Epilepsy Phenome/Genome Project. *Clin. Trials* **10,** 568–586 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to the patients, their families, clinical research coordinators and referring physicians for participating in the Epilepsy Phenome/Genome Project (EPGP) and providing the phenotype data and DNA samples used in this study. We thank the following professional and lay organizations for substantial assistance in publicizing EPGP and therefore enabling us to recruit participants effectively: AED Pregnancy Registry, American Epilepsy Society, Association of Child Neurology Nurses, California School Nurses Organization, Child Neurology Society, Citizens United for Research in Epilepsy, Dravet Syndrome Foundation, Epilepsy Alliance of Orange County, Epilepsy Foundation, Epilepsy Therapy Project, Finding a Cure for Epilepsy and Seizures, IDEA League, InfantileSpasms.com, Lennox-Gastaut Syndrome Foundation, PatientsLikeMe, People Against Childhood Epilepsy, PVNH Support & Awareness, and Seizures & Epilepsy Education. We thank the EPGP Administrative Core (C. Freyer, K. Schardein, R.N., M.S., R. Fahlstrom, M.P.H., S. Cristofaro, R.N., B.S.N. and K. McGovern), EPGP Bioinformatics Core (G. Nesbitt, K. McKenna, V. Mays), staff at the Coriell Institute – NINDS Genetics Repository (C. Tarn, A. Scutti), and members of the Duke Center for Human Genome Variation (B. Krueger, J. Bridgers, J. Keebler, H. Shin Kim, E. Campbell, K. Cronin, L. Hong and M. McCall) for their dedication and commitment to this work. We also thank S. Shinnar (Albert Einstein College of Medicine) and N. Risch (University of California, San Francisco) for valuable input into the creation of EPGP and Epi4K, and R. Stewart, K. Gwinn and R. Corriveau from the National Institute of Neurological Disorders and Stroke for their careful oversight and guidance of both EPGP and Epi4K. This work was supported by grants from the National Institute of Neurological Disorders and Stroke (The Epilepsy Phenome/Genome Project NS053998; Epi4K Project 1—Epileptic Encephalopathies NS077364; Epi4K—Administrative Core NS077274; Epi4K—Sequencing, Biostatistics and Bioinformatics Core NS077303 and Epi4K—Phenotyping and Clinical Informatics Core NS077276); Finding a Cure for Epilepsy and Seizures; and the Richard Thalheimer Philanthropic Fund. We would like to acknowledge the following individuals and groups for their contribution of control samples: J. Hoover-Fong, N. Sobreira and D. Valle; The MURDOCK Study Community Registry and Biorepository (D. Murdock); S. Sisodiya; D. Attix; O. Chiba-Falek; V. Shashi; P. Lugar; W. Lowe; S. Palmer; D. Marchuk; Z. Farfel, D. Lancet, E. Pras; Q. Zhao; D. Daskalakis; R. Brown; E. Holtzman; R. Gbadegesin; M. Winn; S. Kerns; and H. Oster. The collection of control samples was funded in part by ARRA 1RC2NS070342, NIAID R56AI098588, the Ellison Medical Foundation New Scholar award AG-NS-0441-08, an award from SAID-Frederick, Inc. (M11-074), and with federal funds by the Center for HIV/AIDS Vaccine Immunology ("CHAVI") under a grant from the National Institute of Allergy and Infectious Diseases, National Institutes of Health (UO1AI067854).

Author contributions Initial design of EPGP: B.K.A., O.D., D.D., M.P.E., R.Kuz., D.H.L., R.O., E.H.S. and M.R.W. EPGP patient recruitment and phenotyping: B.A.-K., J.F.B., S.F.B., G.C., D.C., P.Cr., O.D., D.D., M.F., N.B.F., D.F., E.B.G., T.G., S.G., S.R.H., J.H., S.L.H., H.E.K., R.C.K., E.H.K., R.Kup., R.Kuz., D.H.L., S.M.M., P.V.M., E.J.N., J.M.Pao., J.M.Par., K.P., A.P., I.E.S., J.J.S., R.S., J.Si., M.C.S., L.L.T., A.V., E.P.G.V., G.K.V.A., J.L.W. and P.W.-W. Phenotype data analysis: B.A.-K., B.K.A., A.B., G.C., O.D., D.D., J.F., T.G., S.J., A.K., R.C.K., R.Kuz., D.H.L., R.O., J.M.Pao., A.P., I.E.S., R.A.S., E.H.S., J.J.S., J.Su., P.W.-W. and M.R.W. Initial design of Epi4K: S.F.B., P.Co., N.D., D.D., E.E.E., M.P.E., T.G., D.B.G., E.L.H., M.R.J., R.Kuz., D.H.L., A.G.M., H.C.M., T.J.O., R.O., A.P., I.E.S. and E.H.S. Epileptic encephalopathy phenotyping strategy: S.F.B., P.Co., D.D., R.Kuz., D.H.L., R.O., I.E.S. and E.H.S. Encephalopathy phenotyping: D.D., K.B.H., M.R.Z.M., H.C.M., A.P., I.E.S., E.H.S. and C.J.Y. Sequence data analysis and statistical interpretation: A.S.A., D.B.G., Y.Ha., E.L.H., S.E.N., S.P., E.K.R. and E.H.S. Functional evaluation of identified mutations: D.B.G., E.L.H., Y.Hi. and Y.-F.L. Writing of manuscript: A.S.A., S.F.B., D.D., D.B.G., Y.Ha., E.L.H., M.R.J., D.H.L., H.C.M., R.O., A.P., S.P., E.K.R., I.E.S. and E.H.S.

Author Information Exome sequence data will be available in dbGAP (Epi4K: Gene Discovery in 4,000 Epilepsy Genomes). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Epi4K (epi4k@duke.edu).

Epi4K Consortium

Andrew S. Allen[1], Samuel F. Berkovic[2], Patrick Cossette[3], Norman Delanty[4], Dennis Dlugos[5], Evan E. Eichler[6], Michael P. Epstein[7], Tracy Glauser[8], David B. Goldstein[9], Yujun Han[9], Erin L. Heinzen[9], Yuki Hitomi[9], Katherine B. Howell[10], Michael R. Johnson[11], Ruben Kuzniecky[12], Daniel H. Lowenstein[13], Yi-Fan Lu[9], Maura R. Z. Madou[13], Anthony G. Marson[14], Heather C. Mefford[15], Sahar Esmaeeli Nieh[16], Terence J. O'Brien[17], Ruth Ottman[18], Slavé Petrovski[2,9,17], Annapurna Poduri[19], Elizabeth K. Ruzzo[9], Ingrid E. Scheffer[20,21], Elliott H. Sherr[22] & Christopher J. Yuskaitis[23]

Epilepsy Phenome/Genome Project

Bassel Abou-Khalil[24], Brian K. Alldredge[25], Jocelyn F. Bautista[26], Samuel F. Berkovic[2], Alex Boro[27], Gregory D. Cascino[28], Damian Consalvo[29], Patricia Crumrine[30], Orrin Devinsky[31], Dennis Dlugos[5], Michael P. Epstein[7], Miguel Fiol[32], Nathan B. Fountain[33], Jacqueline French[12], Daniel Friedman[12], Eric B. Geller[34], Tracy Glauser[8], Simon Glynn[35], Sheryl R. Haut[36], Jean Hayward[37], Sandra L. Helmers[38], Sucheta Joshi[39], Andres Kanner[40], Heidi E. Kirsch[13,41], Robert C. Knowlton[42], Eric H. Kossoff[43], Rachel Kuperman[44], Ruben Kuzniecky[12], Daniel H. Lowenstein[13], Shannon M. McGuire[45], Paul V. Motika[46], Edward J. Novotny[47], Ruth Ottman[18], Juliann M. Paolicchi[24,48], Jack M. Parent[49,50], Kristen Park[51], Annapurna Poduri[19], Ingrid E. Scheffer[20,21], Renée A. Shellhaas[52], Elliott H. Sherr[22], Jerry J. Shih[53], Rani Singh[54], Joseph Sirven[55], Michael C. Smith[40], Joseph Sullivan[13], Liu Lin Thio[56], Anu Venkat[5], Eileen P. G. Vining[57], Gretchen K. Von Allmen[58], Judith L. Weisenberg[59], Peter Widdess-Walsh[34] & Melodie R. Winawer[60]

[1]Department of Biostatistics and Bioinformatics, Duke Clinical Research Institute, and Center for Human Genome Variation, Duke University Medical Center, Durham, North Carolina 27710, USA. [2]Epilepsy Research Centre, Department of Medicine, University of Melbourne (Austin Health), Heidelberg, Victoria 3084, Australia. [3]Centre of Excellence in Neuromics and CHUM Research Center, Université de Montréal, CHUM-Hôpital Notre-Dame Montréal, Quebec H2L 4M1, Canada. [4]Department of Neurology, Beaumont Hospital and Royal College of Surgeons, Dublin 9, Ireland. [5]Department of Neurology and Pediatrics, The Children's Hospital of Philadelphia, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [6]Department of Genome Sciences, University of Washington School of Medicine, Seattle, and Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA. [7]Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia 30322, USA. [8]Division of Neurology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio 45229, USA. [9]Center for Human Genome Variation, Duke University School of Medicine, Durham, North Carolina 27708, USA. [10]Department of Neurology,

The Royal Children's Hospital Melbourne , Parkville, 3052 Victoria, Australia. [11]Centre for Clinical Translation Division of Brain Sciences, Imperial College London, London SW7 2AZ, UK. [12]Comprehensive Epilepsy Center, Department of Neurology, NYU School of Medicine, New York, New York 10016, USA. [13]Department of Neurology, University of California, San Francisco, San Francisco, California 94143, USA. [14]Department of Molecular and Clinical Pharmacology, University of Liverpool, Clinical Sciences Centre, Lower Lane, Liverpool L9 7LJ, UK. [15]Department of Pediatrics, Division of Genetic Medicine, University of Washington, Seattle, Washington 98115, USA. [16]University of California, San Francisco, California 94143, USA. [17]Departments of Medicine and Neurology, The Royal Melbourne Hospital, Parkville, Victoria 3146, Australia. [18]Departments of Epidemiology and Neurology, and the G. H. Sergievsky Center, Columbia University; and Division of Epidemiology, New York State Psychiatric Institute, New York, New York 10032, USA. [19]Division of Epilepsy and Clinical Neurophysiology, Department of Neurology Boston Children's Hospital, Boston, Massachusetts 02115, USA. [20]Epilepsy Research Centre, Department of Medicine, University of Melbourne (Austin Health), Heidelberg, Victoria 3084, Australia. [21]Florey Institute and Department of Pediatrics, Royal Children's Hospital, University of Melbourne, Victoria 3052, Australia. [22]Departments of Neurology, Pediatrics and Institute of Human Genetics, University of California, San Francisco, San Francisco, California 94158, USA. [23]Department of Neurology, Boston Children's Hospital Harvard Medical School, Boston, Massachusetts 02115, USA. [24]Department of Neurology, Vanderbilt University Medical Center, Nashville, Tennessee 37232, USA. [25]Department of Clinical Pharmacy, UCSF School of Pharmacy, Department of Neurology, UCSF School of Medicine, San Francisco, California 94143, USA. [26]Department of Neurology, Cleveland Clinic Lerner College of Medicine & Epilepsy Center of the Cleveland Clinic Neurological Institute, Cleveland, Ohio 44195, USA. [27]Department of Neurology, Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, New York 10467, USA. [28]Divison of Epilepsy, Mayo Clinic, Rochester, Minnesota 55905, USA. [29]Epilepsy Center, Neurology Division, Ramos Mejía Hospital, Buenos Aires 1221, Argentina. [30]Medical Epilepsy Program & EEG & Child Neurology, Children's Hospital of Pittsburgh of UPMC, Pediatrics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15224, USA. [31]NYU and Saint Barnabas Epilepsy Centers, NYU School of Medicine, New York, New York 10016, USA. [32]Department of Neurology, Epilepsy Care Center, University of Minnesota Medical School, Minneapolis 55414, USA. [33]FE Dreifuss Comprehensive Epilepsy Program, University of Virginia, Charlottesville, Virginia 22908, USA. [34]Division of Neurology, Saint Barnabas Medical Center, Livingston, New Jersey 07039, USA. [35]Department of Neurology, Comprehensive Epilepsy Program, University of Michigan Health System, Ann Arbor, Michigan 48109, USA. [36]Comprehensive Epilepsy Center, Montefiore Medical Center, Bronx, New York 10467, USA. [37]The Kaiser Permanente Group, Oakland, California 94618, USA. [38]Neurology and Pediatrics, Emory University School of Medicine, Atlanta, Georgia 30322, USA. [39]Pediatrics & Communicable Diseases, University of Michigan, Ann Arbor, Michigan 48109, USA. [40]Department of Neurological Sciences, Rush Epilepsy Center, Rush University Medical Center, Chicago, Illinois 60612, USA. [41]Department of Radiology, University of California, San Francisco, California 94143, USA. [42]Neurology, University of Texas Medical School, Houston, Texas 77030, USA. [43]Neurology and Pediatrics, Child Neurology, Pediatric Neurology Residency Program, Johns Hopkins Hospital, Baltimore, Maryland 21287, USA. [44]Epilepsy Program, Children's Hospital & Research Center Oakland, Oakland, California 94609, USA. [45]Clinical Neurology, Children's Hospital Epilepsy Center of New Orleans, New Orleans, Louisiana 70118, USA. [46]Comprehensive Epilepsy Center, Oregon Health and Science University, Portland, Oregon 97239, USA. [47]Departments of Neurology and Pediatrics, University of Washington School of Medicine, Seattle Children's Hospital, Seattle, Washington 98105, USA. [48]Weill Cornell Medical Center, New York, New York 10065, USA. [49]Department of Neurology and Neuroscience Graduate Program, University of Michigan Medical Center, Ann Arbor, Michigan 49108, USA. [50]Ann Arbor Veterans Administration Healthcare System, Ann Arbor, Michigan 48105, USA. [51]Departments of Neurology and Pediatrics, University of Colorado School of Medicine, Children's Hospital Colorado, Denver, Colorado 80045, USA. [52]University of Michigan, Pediatric Neurology, Ann Arbor, Michigan 48109, USA. [53]Department of Neurology, Mayo Clinic, Jacksonville, Florida 32224, USA. [54]Division of Pediatric Neurology, University of Michigan Health System, Ann Arbor, Michigan 48109, USA. [55]Department of Neurology, Mayo Clinic, Scottsdale, Arizona 85259, USA. [56]Department of Neurology, Washington University School of Medicine, St Louis, Missouri 63110, USA. [57]Department of Neurology, Johns Hopkins Hospital, Baltimore, Maryland 21287, USA. [58]Division of Child & Adolescent Neurology, Departments of Pediatrics, University of Texas Medical School, Houston, Texas 77030, USA. [59]Department of Neurology, Division of Pediatric Neurology, Washington University School of Medicine, St Louis, Missouri 63110, USA. [60]Department of Neurology and the G.H. Sergievsky Center, Columbia University, New York, New York 10032, USA.

METHODS

Study subjects. Infantile spasms and Lennox–Gastaut syndrome patients evaluated in this study were collected through the Epilepsy Phenome/Genome Project (EPGP, http://www.epgp.org)[29]. Patients were enrolled across 23 clinical sites. Informed consent was obtained for all patients in accordance with the site-specific Institutional Review Boards. Phenotypic information has been centrally databased and DNA specimens stored at the Coriell Institute–NINDS Genetics Repository (Supplementary Table 1). Infantile spasms patients were required to have hypsarrhythmia or a hypsarrhythmia variant on EEG. Lennox–Gastaut syndrome patients were required to have EEG background slowing or disorganization for age and generalized spike and wave activity of any frequency or generalized paroxysmal fast activity (GPFA). Background slowing was defined as <8 Hz posterior dominant rhythm in patients over 3 years of age, and <5 Hz in patients over 2 years of age. EEGs with normal backgrounds were accepted if the generalized spike and wave activity was 2.5 Hz or less and/or if GPFA was present.

All patients were required to have no evidence of moderate-to-severe developmental impairment or diagnosis of autistic disorder or pervasive developmental disorder before the onset of seizures. Severe developmental delay was defined by 50% or more delay in any area: motor, social, language, cognition, or activities of living; or global delay. Mild delay was defined as delay of less than 50% of expected milestones in one area, or less than 30% of milestones across more than one area. All patients had no confirmed genetic or metabolic diagnosis, and no history of congenital TORCH infection, premature birth (before 32 weeks gestation), neonatal hypoxic-ischaemic encephalopathy or neonatal seizures, meningitis/encephalitis, stroke, intracranial haemorrhage, significant head trauma, or evidence of acquired epilepsy. All infantile spasms and Lennox–Gastaut syndrome patients had an MRI or CT scan interpreted as normal, mild diffuse atrophy or focal cortical dysplasia. (Our case with the mutation in *HNRNPU* had had a reportedly normal MRI but on review of past records, a second more detailed MRI was found showing small regions of periventricular nodular heterotopia.) To participate, both biological parents had to have no past medical history of seizures (except febrile or metabolic/toxic seizures).

A final diagnosis form was completed by the local site EPGP principal investigator based on all collected information. A subset of cases was reviewed independently by two members of the EPGP Data Review Core to ensure data quality and consistency. All EEGs were reviewed by a site investigator and an EEG core member to assess data quality and EEG inclusion criteria. EEGs accepted for inclusion were then reviewed and scored by two EEG core members for specific EEG phenotypic features. Disagreements were resolved by consensus conference among two or more EEG core members before the EEG data set was finalized. MRI scans were reviewed by local investigators and an MRI core member to exclude an acquired symptomatic lesion.

Exome-sequenced unrelated controls ($n = 436$) used to ascertain mutation frequencies were sequenced in the Center for Human Genome Variation as part of other genetic studies.

Exome sequencing, alignment and variant calling. Exome sequencing was carried out within the Genomic Analysis Facility in the Center for Human Genome Variation (Duke University). Sequencing libraries were prepared from primary DNA extracted from leukocytes of parents and probands using the Illumina TruSeq library preparation kit following the manufacturer's protocol. Illumina TruSeq Exome Enrichment kit was used to selectively amplify the coding regions of the genome according to the manufacturer's protocol. Six individual barcoded samples (two complete trios) were sequenced in parallel across two lanes of an Illumina HiSeq 2000 sequencer.

Alignment of the sequenced DNA fragments to Human Reference Genome (NCBI Build 37) was performed using the Burrows–Wheeler Alignment Tool (BWA) (version 0.5.10). The reference sequence we use is identical to the 1000 Genomes Phase II reference and it consists of chromosomes 1–22, X, Y, MT, unplaced and unlocalized contigs, the human herpesvirus 4 type 1 (NC_007605), and decoy sequences (hs37d5) derived from HuRef, Human Bac and Fosmid clones and NA12878.

After alignments were produced for each individual separately using BWA, candidate *de novo* variants were jointly called with the GATK Unified Genotyper for all family members in a trio. Loci bearing putative *de novo* mutations were extracted from the variant call format files (VCFs) that met the following criteria: (1) the read depth in both parents should be greater than or equal to 10; (2) the depth of coverage in the child should be at least one-tenth of the sum of the coverage in both parents; (3) for *de novo* variants, less than 5% of the reads in either parent should carry the alternate allele; (4) at least 25% of the reads in the child should carry the alternate allele; (5) the normalized, phred-scaled likelihood (PL) scores for the offspring genotypes AA, AB and BB, where A is the reference allele and B is the alternate allele, should be >20, 0 and >0, respectively; (6) the PL scores for both parents should be 0, >20 and >20; (7) at least three variant alleles must be observed in the proband; and (8) the *de novo* variant had to be located in a CCDS exon targeted by the exome enrichment kit. PL scores are assigned such that the most likely genotype is given a score of 0, and the score for the other two genotypes represent the likelihood that they are not the true genotypes. SnpEff (version 3.0a) was used to annotate the variants according to Ensembl (version 69) and consensus coding sequencing (CCDS release 9, GRCh37.p5) and limited analyses to exonic or splice site (2 bp flanking an exon) mutations. All candidate *de novo* mutations that were absent from population controls, including a set of 436 internally sequenced controls and the ~6,500 individuals whose single nucleotide variant data are reported in the Exome Variant Server, NHLBI Exome Sequencing Project (http://evs.gs.washington.edu/EVS/; August 2012) were also visually inspected using Integrative Genomics Viewer (IGV). All candidate *de novo* mutations were confirmed with Sanger sequencing of the relevant proband and parents. For comparison, we also called *de novo* variants from probands and parents individually for a subset of trios. Using this individual calling approach we identified and confirmed an additional 46 *de novo* mutations. These were included in all the downstream *de novo* mutation analyses.

Calculation of gene-specific mutation rate. Point mutation rates were scaled to per base pair, per generation, based on the human genome sequences matrix[30] (provided by S. Sunyaev and P. Polak), and the known human average genome *de novo* mutation rate (1.2×10^{-8})[31]. The mutation rate (M) of each gene was calculated by adding up point mutation rates in effectively captured CCDS regions in the offspring of trios, and then dividing by the total trio number ($S = 264$). The P value was calculated as [1 − Poisson cumulative distribution function ($x - 1, \lambda$)], where x is the observed *de novo* mutation number for the gene, and λ is calculated as $2SM$ for genes on autosome or $(2f + m)M$ for genes on chromosome X (f and m are the number of sequenced female and male probands, respectively). Genes on Y chromosome were not part of these analyses. Two *de novo* mutations in gene *ALG13* are at the same position, likewise in gene *SCN2A*. We calculated the probability of this special case as [1 − Poisson cumulative distribution function ($1, (2f + m)r$)], where r reflects the point mutation rate on that specific *de novo* mutation position. Further investigations indicated that it is unlikely for these *de novo* mutations, which occur at the same site across distinct probands, to have been caused by sequencing or mapping errors (Supplementary Methods).

Calculation of mutation tolerance for HGNC genes. To assign quantitatively a mutation tolerance score to genes in the human genome (HGNC genes), we calculated an empirical penalty based on the presence of common functional variation using the aggregate sequence data available from the 6,503 samples reported in the Exome Variant Server, NHLBI Exome Sequencing Project (http://evs.gs.washington.edu/EVS/; accessed August 2012). We first filtered within the EVS database and eliminated from further consideration genes where the number of tenfold average covered bases was less than 70% of its total extent. In calculating a score, we focused on departures from the average common functional variant frequency spectrum, corrected for the total mutation burden in a gene. We construct this score as follows. Let Y be the total number of common, minor allele frequency $>0.1\%$, missense and nonsense (including splice) variants and let X be total number of variants (including synonymous) observed within a gene. We regress Y on X and take the studentized residual as the score (S). Thus, the raw residual is divided by an estimate of its standard deviation and thus account for differences in variability that comes with differing mutational burdens. S measures the departure from the average number of common functional mutations found in genes with a similar amount of mutational burden. Thus, when $S = 0$ the gene has the average number of common functional variants given its total mutational burden. Genes where $S < 0$ have less common functionals than average for their mutational burden and thus, would seem to be less tolerant of functional mutation, indicating the presence of weak purifying selection. We further investigated how different 'intolerance' thresholds of S captured known epileptic encephalopathy genes (Supplementary Table 8). Supplementary Fig. 6 illustrates how different percentiles of S lead to the classification of different proportions of the known epileptic encephalopathy genes as 'intolerant'. Note that *ARX* is not in these analyses as this gene did not meet a 70% of gene coverage threshold. The dashed vertical line in Supplementary Fig. 6 illustrates the 25th percentile of S and shows that using this threshold results in 12 out of the 14 assessed known genes being considered 'intolerant'. On the basis of this analysis, we used this 25th percentile threshold in classifying genes as intolerant in all subsequent analyses. Supplementary Table 9 lists the 25th percentile of most intolerant genes that had Sanger confirmed *de novo* mutations among the infantile spasms/Lennox–Gastaut syndrome probands.

Defining the CCDS opportunity space for detecting *de novo* mutations. For each trio, we defined callable exonic bases that had the opportunity for identification of a coding *de novo* mutation, by restricting to bases where each of the three family members had at least tenfold coverage, obtained a multi-sampling (GATK) raw phred-scaled confidence score of ≥ 20 in the presence or absence of a variant, and were within the consensus coding sequence (CCDS release 9, GRCh37.p5) or within the two base pairs at each end of exons to allow for splice acceptor and donor variants. Using these three criteria, the average CCDS-defined *de novo* mutation opportunity space across 264 trios was found to be 28.84 ± 0.92 Mb (range of 25.46–30.25 Mb).

To explore at the gene level, we similarly assessed the *de novo* calling opportunity within any given trio for every gene with a CCDS transcript. For genes with instances of non-overlapping CCDS transcripts, we merged the corresponding regions into a consensus summary of all CCDS-defined bases for that gene. Using these criteria, over 85% of the CCDS-defined exonic regions were sequenced to at least tenfold coverage across the three family members in over 90% of trios. All 264 trios covered at least 79% of the CCDS-defined regions under the CCDS opportunity space criteria.

Calculations of CCDS opportunity space for calling a *de novo* mutation, aside from the Y chromosome, were used in both the gene-list enrichment and architecture calculations.

30. Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80,** 727–739 (2007).
31. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (2012).