Variants in *SBF1*, *CELSR2*, and *TENM1* are associated with childhood epileptic encephalopathies

Yi-Wu Shi[1*], Jian-Guo Zhang[2,3*], Na He[1*], Zi-Long Ye[1], Wen-Bin Li[1], Han-Kui Liu[2,3], Zhi-Gang Liu[4], Nan-Xiang Shen[1], Xiao-Chong Qu[1], Cui-Xia Fan[1], Jie Wang[1], Sheng Luo[1], Yu-Dan Lv[5], Li Gao[6], Jing Chen[7], Shao-Ping Huang[8], Xin-Guo Lu[9], Jing Yu[10], Jie Zhang[11], Xiu-Xia Wang[12], Qu-Wen Gao[13], Li Jiang[14], Yan-Hui Chen[15], Hui Qi[6], Jing-Da Qiao[1], Li-Zhi Chen[1], Yuan-Jin Zeng[1], Xiao-Xue Yang[16], Hong-Jun Yan[17], Cheng-Yan Li[18], Tao Zeng[19], Fu-Li Min[19], Bing Qin[20], Hai-Qing Xu[21], Lin Xu[22], Bing-Mei Li[1], Yong-Hong Yi[1], Zhi-Hong Zhuo[23], Qing-Hui Guo[24], Su-Li He[25], Hong-Wei Zhang[26,27], Li-Ping Guan[2,3], Wei-Yi Deng[1], Xiao-Fan Ren[9], Dong-Fang Zou[9], Wei-Yue Gu[28], Tao Su[1], Xiao-Rong Liu[1], Yi-Bo Qu[29], Xin-Ping Yang[16], Wei-Ping Liao[1#]

1 Department of Neurology, Institute of Neuroscience, Key Laboratory of Neurogenetics and Channelopathies of Guangdong Province and the Ministry of Education of China, The Second Affiliated Hospital, Guangzhou Medical University, Guangzhou, China, 510260

2 BGI Genomics, Shenzhen 518083, China

3 Clin Lab, BGI Genomics, Shijiazhuang 050011, China

4 Affiliated Foshan Maternity and Child Healthcare Hospital, Southern Medical University, Foshan, China

5 The First Hospital of Jilin University, Jilin, China

6 Department of Pediatrics, Henan Province People's Hospital, People's Hospital of

Zhengzhou University, Zhengzhou, China

7 Department of Neurology, Children's Hospital of Nanjing Medical University, Nanjing, China

8 The Second Affiliated Hospital of Xi'an Jiaotong University, Xian, China

9 Shenzhen Children's Hospital, Shenzhen, China

10 Neurology Department, Children's Hospital of Xinjiang Uygur Autonomous Region, Urumchi, China

11 Hunan Provincial Children's Hospital, Changsha, China

12 The Second Hospital of Hebei Medical University, Shijiazhuang, China

13 Department of Neurology, Foresea Life Insurance Shaoguan Hospital, Shaoguan, China

14 Children's Hospital of Chongqing Medical University, Chongqing, China

15 Department of Pediatrics, Fujian Medical University Union Hospital, Fuzhou, China

16 Center for genetics and Developmental Systems Biology, Nanfang Hospital, Southern Medical University, Guangzhou, China

17 Guangdong 999 Brain Hospital, Guangzhou, China

18 Affiliated Hospital of Guangdong Medical University, Zhanjiang, China

19 Guangzhou First People's Hospital, Guangzhou, China

20 The First Affiliated Hospital of Jinan University, Guangzhou, China

21 Xuzhou Center Hospital, Xuzhou, China

22 Qingdao women and children's hospital, Qingdao, China

23 The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

24 The Second Hospital of Shandong University, Jinan, China

25 Shantou Chaonan Minsheng Hospital, Shantou, China

26 Children's Hospital Affiliated to Shandong University, Jinan, China

27 Jinan Children's Hospital, Jinan, China

28 Chigene (Beijing) Translational Medical Research Center Co., Beijing, China

29 Key Laboratory of CNS Regeneration (Ministry of Education), Guangdong-Hong Kong-Macau Institute of CNS Regeneration, Jinan University, Guangzhou, China

* These authors contributed equally to the manuscript.

Correspondence to Dr. Liao, Department of Neurology, Institute of Neuroscience, Key Laboratory of Neurogenetics and Channelopathies of Guangdong Province and the Ministry of Education of China, The Second Affiliated Hospital, Guangzhou Medical University, Guangzhou, China, E-mail wpliao@163.net; liao@gzneurosci.com.

Abstract

Epileptic encephalopathy is a common devastating epilepsy with etiologies remain largely elusive, despite application of whole gene/exon sequencing on large cohorts. This study targeted on childhood epileptic encephalopathy, typically Lennox-Gastaut syndrome that is featured by age-dependent onset and characteristic clinical manifestations. Trio-based whole-exome sequencing with individualized analysis before statistic studies was employed, including individualized analysis on each trio by explainable inheritance origin and stratified frequency filtration and on each gene

from four aspects. This study identified three novel candidate genes in 235 patients, including *SBF1* with *de novo* variants in three cases, *CELSR2* with biallelic recessive variants in eight cases, and *TENM1* with X-linked recessive variants in six cases. These genes presented significant repetitiveness, including significant higher excess of *de novo* variants in *SBF1* and excess of biallelic variants in *CELSR2*, and aggregated frequencies of variants in *SBF1*, *CELSR2*, and *TENM1*. The frequency of compound heterozygous/homozygous *CELSR2* variants in the cases was significantly higher than that in the asymptomatic parent-controls. Further experiments with knock-down/knock-out of these genes showed increased seizure-like behavior and increased firing of excitatory neurons. This study highlights the implication of phenotype sub-classification, individualized analysis in combination with statistic studies, and use of controls for compound heterozygous variants.

Keywords

Lennox-Gastaut syndrome, *SBF1*, *CELSR2*, *TENM1*, individualized analysis, compound heterozygous variants, phenotype sub-classification.

Introduction

Epileptic encephalopathies, also referred to developmental and/or epileptic encephalopathies (DEEs)[1], represent a clinically and genetically heterogeneous group of devastating epilepsies characterized by refractory seizures, severe electroencephalography (EEG) abnormalities, and neurodevelopment delay or decline.

Epileptic encephalopathy is a common clinical entity and accounts for about 40% of newly diagnosed epilepsies before the age of three years[2]. According to onset age, epileptic encephalopathies could generally be divided into two major groups[1]: early infantile epileptic encephalopathy (EIEE) with onset in neonates or infants, such as West syndrome and Dravet syndrome; and childhood-onset epileptic encephalopathy, typically Lennox-Gastaut syndrome (LGS). LGS is estimated to account for 1-2% of all patients with epilepsy, and up to 10% of patients of childhood epilepsy[3-5]. It usually begins between 1 to 8 years old (before age 18 years) and is characterized by multiple seizure types that must include tonic seizures and/or consistent paroxysmal fast activity/polyspike on EEG[5]. The etiologies of epileptic encephalopathies remain largely elusive. Recent studies have demonstrated an increased number of genes associated with epileptic encephalopathies. So far, more than 100 genes have been reported to be associated with EIEE/DEEs (https://omim.org), explaining 22% to 40% of cases with EIEE[6,7]. In contrast, only several genes, such as *CHD2*, *DNM1*, *GABRB3*, and *SYNGAP1*, have been identified in patients with LGS, accounting for a small proportion of cases[8-11]. However, the age-dependent feature, specific clinical characteristics, and distinct EEG pattern highly suggest an underlying genetic mechanism in the epileptogenesis, which remains to be determined.

In the present study, we performed trio-based whole-exome sequencing (WES) in a cohort of patients with LGS without acquired etiologies. An individualized analytical framework was employed to identify the potential pathogenic variants in each trio. Novel genes with repetitively identified *de novo* variants, biallelic variants,

or hemizygous variants, were selected for further studies to define the gene-disease association. Three genes were identified as novel candidate genes in this cohort, including *SBF1* with autosomal dominant inheritance pattern (AD, *de novo* variants), *CELSR2* with autosomal recessive inheritance (AR), and *TENM1* with X-linked recessive inheritance (XLR). A control cohort consisting of unaffected parents was established to determine the significance of recessive variants. Animal models were used to validate the roles of the novel candidate genes in epilepsy. This study suggests that *SBF1*, *CELSR2*, and *TENM1* were potentially pathogenic genes associated with LGS and highlights the implication of phenotype sub-classification, individualized analysis in combination with statistic studies, and use of controls for compound heterozygous variants.

Methods

Subjects. A total of 235 unrelated cases (234 trios and 1 singleton) with Lennox-Gastaut syndrome (LGS) were enrolled from 2005 to 2021. Among these cases, 194 were long-term followed-up patients in the Epilepsy Center of the Second Affiliated Hospital of Guangzhou Medical University. The other 41 patients were from 20 epilepsy centers of the First Hospital of Jilin University, Affiliated Foshan Maternity and Child Healthcare Hospital, Guangdong 999 Brain Hospital, Henan Province People's Hospital, Shenzhen Children's Hospital, Hunan Provincial Children's Hospital, The Second Affiliated Hospital of Xi'an Jiaotong University, Foresea Life Insurance Shaoguan Hospital, Children's Hospital of Nanjing Medical

University, People's Hospital of Xinjiang Uygur Autonomous Region, The Second Hospital of Hebei Medical University, Affiliated Hospital of Guangdong Medical University, Guangzhou First People’s Hospital, The First Affiliated Hospital of Jinan University, Xuzhou Center Hospital, Qingdao women and children’s hospital, The First Affiliated Hospital of Zhengzhou University, The Second Hospital of Shandong University, Shantou Chaonan Minsheng Hospital, and Qilu Children's Hospital of Shandong University.

Detailed clinical information was collected, including gender, seizure onset age, seizure type and frequency, course of seizure, response to antiepileptic drugs, family history, and findings from general and neurological examinations. Brain MRI scans were performed to identify abnormalities in brain structure. Patients with acquired etiologies were excluded. Video EEG was examined and the results were reviewed by two qualified electroencephalographers. Epileptic seizures and LGS were diagnosed and classified in accordance with the Commission on Classification and Terminology of the International League Against Epilepsy criteria (1989, 2001, 2010, 2017, and 2022). All patients were diagnosed as LGS with frequent seizures characterized by two or more of the following features: (1) multiple seizure types including tonic seizures, (2) generalized polyspikes or fast rhythms during sleep (especially required when daily tonic seizures are obscure), and (3) generalized slow (< 2.5 Hz) spike-wave complex on EEG. The patients were followed up for at least one year. All subjects were Han Chinese with four Han Chinese grandparents, and were born to non-consanguineous Chinese parents.

This study adhered to the guidelines of the International Committee of Medical Journal Editors regarding patient consent for research or participation, and received approval from the ethics committee of the Second Affiliated Hospital of Guangzhou Medical University. Written informed consents were obtained from the patients or their legal guardians.

Whole-exome sequencing. Genomic DNA was extracted from the peripheral blood that were obtained from the patients and their parents (trios) using the QIARamp blood mini kit (Qiagen, Hilden, Germany). Trio-based whole-exome sequencing (WES) was performed on a HiSeq 2000 system (Illumina, San Diego, CA, USA) as previously reported[11,12]. The sequencing data were generated by massively parallel sequencing with > 100 times average depth and > 98% coverage of the capture regions, and the high-quality reads were mapped to the Genome Reference Consortium Human Genome build 37 (GRCh37) by Burrows-Wheeler alignment. Single-nucleotide variants and insert/deletion variants were called with the Genome Analysis Tool Kit.

Variants filtration and evaluation. To identify candidate causative variants in each trio, we adopted an individualized analytical framework in the following five steps (Fig. 1).

1. General filtration. The potentially pathogenic variants, including nonsense, frameshift, canonical splice site, initiation codon, inframe insert/deletion, missense,

and synonymous variants predicted to impact splicing were retained, then filtrated with minor allele frequency (MAF) < 0.005 in the Genome Aggregation Database (gnomAD).

2. Variant filtration by explainable inheritance origin in each trio. The origin of variants presents the genetic difference between the affected child and the parents and thus explains the occurrence of phenotype in a given family (trio), i.e., *de novo*, recessive from the two asymptomatic parents each, or hemizygous. The variants with explainable inheritance origin in each trio were selected.

3. The variants were then filtrated with stratified MAF criteria. The MAF of *de novo* variants, hemizygotes, and homozygotes is set as absent in the control population in gnomAD; and for compound heterozygous variants, the product of multiplying the frequencies of two alleles in gnomAD is $< 1\times10^{-6}$, which is 7 times less than the probability of one individual in the current population of gnomAD ($1/141456 = 7 \times 10^{-6}$).

4. Variant filtration with criteria on the gene profile. The genes with selected variants were classified into epilepsy-associated genes (977 genes by Wang *et al.*[13] and updated from OMIM) and genes with undefined gene-disease/epilepsy/seizures associations. For variants in epilepsy-associated genes, their pathogenicities were evaluated by the American College of Medical Genetics and Genomics Standards and Guidelines (ACMG)[14]. For those with undefined gene-disease/epilepsy/seizure associations, the following four aspects were assessed: (1) Tissue-specific expression, a candidate pathogenic gene should be expressed in the brain (inclusion criterion),

with careful consideration on the possibility of other explainable pathogenic mechanisms, such as ectopic expression and remote toxic effect of abnormal metabolic products. (2) Excluding the disease-causing genes with previously defined gene-disease association (including genotype-phenotype correlation) against the possibility of epilepsy as a phenotype (exclusion criterion). (3) The probability of being intolerant to heterozygous/homozygous variants of loss-of-function (pLI/pRec), genes of pLI $\geq$ 0.9 with repetitive *de novo* variants, and genes of pLI $\geq$ 0.9/pRec $\geq$ 0.9/pNull $\leq$ 0.1 with repetitive recessive variants were considered. (4) Whether the genetic knock-out/knock-down produces phenotypes of the brain.

5. The genes with repetitively identified (in two or more cases) *de novo*, recessive, and hemizygous variants were selected as novel candidate genes and were subjected to further statistical analysis, including excess of *de novo* variants[8] and recessive variants[15], aggregate frequency of variants[16], and the frequency of the compound heterozygous variants comparing with that in asymptomatic parent cohort. In brief, for excess of *de novo* variants, the *P*-value was calculated as [1-Poisson cumulative distribution function (χ-1, λ)], where χ is the observed *de novo* variant number for a specific gene, and λ is calculated as $2\times234\times1.2\times10^{-8}\times$effective transcript length of the gene on autosome. For excess of biallelic variants, the *P*-value was calculated as [1-cumulative binomial probability [(n-1, N, R)], where n is the observed biallelic variant number for a specific gene, N is the number of trios (234 in this cohort), and R is the rate of variants in the gene by chance in population. The variants were filtered by different MAF cutoffs based on the class of variants, i.e.,

biallelic loss-of-function (LOF), one LOF and one damaging missense, and biallelic damaging missense. Considering all patients were Han Chinese, the cutoff is set according to MAF in the population of ExAC-East Asian. For aggregated frequency of variants, the controls include a cohort of 296 healthy Chinese volunteers as in our previous report[12] and the control population in gnomAD. For compound heterozygous variants (two variants in trans), we established a control cohort that included 1942 asymptomatic parents from trios, in whom the compound heterozygous variants were identified by detecting one of the paired variants in the child, given that one of the paired variants in a parent would transmit to the child.

All candidate variants were validated by Sanger sequencing. Qualified novel candidate genes were subjected to further bioinformatics and experimental studies to define the gene-disease associations.

Protein structure modeling. Tertiary structure of protein encoding by the novel candidate genes was modeled using trRosetta (https://yanglab.nankai.edu.cn/trRosetta/) to predict the effect of missense variants on protein structure. Hydrogen bonding was analyzed and visualized using PyMOL 2.3 software (https://pymol.org/2/).

Functional studies on *SBF1*, *CELSR2*, and *TENM1*. The use of animals and all experimental procedures were approved by the Committee for Animal Care of the Second Affiliated Hospital of Guangzhou Medical University.

Fly stocks. Flies were fed standard cornmeal and maintained in an incubator at 25 □, 60%-70% humidity on a 12:12-hr light/dark cycle as previously described[12]. The *Drosophila* homologs of human *SBF1*, *CELSR2*, and *TENM1* genes are *Sbf* (CG6939), *Flamingo* (CG11895), and *Ten-m* (CG5723), respectively. *UAS-Sbf-RNAi* (FBgn0025802), *UAS-Flamingo-RNAi* (FBgn0024836), and *UAS-Ten-m-RNAi* (FBgn0004449) were purchased from the Tsing Hua Fly Center (Tsinghua University, Beijing, China). The Gal4 driver line *tub-Gal4* and *Canton-S* were gifts from Prof. LIU Ji-Yong (Guangzhou Medical University, Guangzhou, China). Virgin *tub-GAL4* females were mated with males of each RNAi line for establishing specific gene knockdown flies. RNAi efficiency of the three genes in flies was detected by reverse transcription quantitative PCR (RT-qPCR). *Canton-S* was used as the WT line in this study.

Seizure behavior test in flies. The *tub-Gal4* line was crossed with *Sbf-*, *Flamingo-*, and *Ten-m-RNAi* to establish globally corresponding gene knockdown flies (*tub-Gal4>Sbf-RNAi*, *tub-Gal4>Flamingo-RNAi*, and *tub-Gal4>Ten-m-RNAi)*, respectively[17]. The bang-sensitivity (BS) test was carried out on flies 3-5 days after eclosion and seizure-like behavior was evaluated as in previous studies[12,18]. Briefly, flies were anesthetized with CO_2 and transferred into a new clean food vial one day before testing. Three to six flies were placed in one vial and mechanically stimulated with a vortex mixer (VWR, Radnor, PA, USA) at maximum speed for 20 sec. The

percentage and duration of BS paralysis in flies were recorded.

Electrophysiological assay on flies. Attached recording was performed in the knockdown and control flies using an established electrophysiological method[19]. Fly brains were dissected as previously described[20], transferred to a recording chamber with fly external solution, and immobilized with a C-sharp holder. The anterior side of the brain was positioned facing upward to enable observation of the soma of projection neurons (PNs) on the brain surface. The standard external solution contained 101 mM NaCl, 1 mM $CaCl_2$, 4 mM $MgCl_2$, 3 mM KCl, 5 mM glucose, 1.25 mM NaH_2PO_4, and 20.7 mM $NaHCO_3$ (pH 7.2 and 250 mOsm). A patch pipette filled with external solution was used for attached recording with a 700B amplifier. Data were acquired with a Digidata 1440B digital–analog converter (Molecular Devices, San Jose, CA, USA) and pClamp v10.5 software (Molecular Devices, San Jose, CA, USA).

Generation of sbf1 knockout zebrafish using CRISPR-Cas9. The AB strain zebrafish (provided by Dr. Yang Xin-Ping of Southern Medical University) embryos and larvae were maintained on a 14:10-hr light/dark cycle and raised as previously described[21]. The larvae were allowed to develop at 28.5□ and were staged by hours or days after fertilization under standard conditions. CRISPR-Cas9 genomic editing was used to generate *sbf1* knockout zebrafish as described previously[22]. The CRISPR target sequence (5’-AGTGGATGGCAGCTGGCTCC-3’) and oligonucleotide were

designed using ZiFit software[23]. The *sbf1*-targeting single-guide RNA (sgRNA) template plasmid was generated by annealing oligonucleotides and ligation of the double-stranded DNA in the plasmid PMD-18T (Takara, 6011), and then transcribed *in vitro* with T7 RNA polymerase (Thermo Fisher, # EP0111). A 1 nL mixture containing about 300 ng/μL Cas9 protein (New England Biolabs, Beverly, MA) and 15 ng/μL sgRNA was injected into one-cell stage embryos using a Picospritzer III pressure ejector. Germline mutations with a 4-nt deletion were introduced in the third coding exon of *sbf1* gene. The mutant was verified by PCR-Sanger sequencing with dF1/dR1 primers.

The expression level of *sbf1* was detected in 5 days post-fertilization (d.p.f) larvae by quantitative real-time PCR (RT-qPCR) using primers gF1/gR1 and dF2/dR2 and SYBR Premix Ex Taq (Takara, Japan) on aMx3005P qPCR System (Agilent, USA). The primer sequences were listed in Supplementary Table 13.

Morphological, survival, and behavioral analysis on sbf1 knockout zebrafish. The morphological and survival features of zebrafish from 0 to 30 d.p.f were observed and photographed using an Olympus microscope (BX51, Olympus, Japan). The abnormal morphological phenotypes, including curved body and abdominal swelling, were observed. The survival percentage of each genotype zebrafish was calculated using Kaplan-Meier survival analysis. For locomotion tracking, 6-9 d.p.f larvae were placed in individual wells of a 96-well flat-bottomed culture dish (Corning, #07-200-565). Each well contained approximately 200 μL Ringer's solution and one larva. Behavior

was monitored at 28.5 °C using a ZebraBox system (ViewPoint Behavior Technology) and was then analyzed using Zebralab locomotion tracking software (ViewPoint Behavior Technology). After 10 min of habituation, each larva was recorded for a total of 5 min with 5 light/dark cycles (each consisting of 10 sec on light and 50 sec off light) and was followed by 60 min recording. We then calculated the traveled distance, duration, and numbers of rapid whirlpool-like circling swimming with high speed (> 20 mm/sec, characterized as seizure behavior)[12, 24]. Videos were analyzed blindly for the classification of seizure behavior.

Electrophysiological assay on zebrafish. Electrophysiological recordings in zebrafish were obtained using the methods described in the literatures[24,25]. Briefly, 7-11 d.p.f zebrafish larvae were embedded in a 1.2% agarose gel prepared by dissolving low-melting temperature agarose in artificial cerebrospinal fluid (aCSF) (NaCl 134 mM, KCl 2.9 mM, MgCl2 1.2 mM, CaCl2 2.1 mM, Glucose 10 mM, and HEPES 10 mM) with anesthetic agent (α-bungarotoxin, 1 mg/mL, Sigma). Then, fish were placed on the upright stage of an Olympus microscope and perfused with standard recording solution. A glass microelectrode (~1 μm tip diameter, 2-7 MΩ) was placed in the optic tectum of the fish under visual guidance. Electrodes were filled with 2 M NaCl and electrical activity was recorded using a 700B amplifier. Traces were band-pass filtered (0.2-1 kHz). A detection threshold of three times background noise was applied to extract the events on a 5-minute-long window. Detected events were manually sorted. Two or more consecutive spikes with the threshold of three times

background noise were considered as seizure-like events in a 20-minute-long window. The time and duration of each event were extracted to evaluate seizure-like electrophysiology.

Behavior monitoring and electroencephalography (EEG) recording in Celsr2$^{-/-}$ mice. The *Celsr2*$^{-/-}$ mouse was generated as previously[26] and provided by Dr. Yibo Qu (Guangdong-Hong Kong-Macau Institute of CNS Regeneration, Jinan University). All mice were obtained by breeding *Celsr2*$^{+/-}$ males with *Celsr2*$^{+/-}$ females and maintained a 12-hr light/dark cycle (lights on at 8:00 am, off at 8:00 pm) with food and water *ad libitum*. At three weeks old, the mice were placed into opaque home cages under 24-hr video monitoring for 7 days.

EEGs were recorded in the *Celsr*2$^{-/-}$ and wild-type mice with freely behaving for analysis of epileptic seizure, according to a previous study[27]. Briefly, the mice were anesthetized by 2% isoflurane (vol/vol) and implanted with a teflon-coated silver wire electrode in the cortex (1.5 mm posterior and 1.2 mm lateral to the bregma) for signal capture. The EEG signal was synchronized with a USB video camera to analyze the epileptic seizure-like behavior. EEG signals were amplified with a preamplifier and primary amplifier (EX1, Dagan) and sampled at 10000 Hz (gain, 1000; band-pass filter, 0.1 to 500 Hz). Epileptic seizures were identified by checking both the EEG signal and the behavior of animals. The incidence of seizure was calculated as previously described[28].

Statistical analysis. The two-tailed Fisher's exact test was used to compare the allele frequencies between case and control groups. Data from experiments were first tested for Gaussian distribution. For normally distributed data, Student's t test and one-way analysis of variance (ANOVA) with Tukey post hoc test were used to compare differences among two and multiple groups. For the data that are of non-normal distribution, Mann Whitney test and Kruskal-Wallis test were used to compare differences among two and multiple groups, respectively. Kaplan-Meier survival curves were compared using log-rank statistics. All testing was conducted by experimenters blinded to animal genotypes and analyzed by using GraphPad Prism 8.2 (La Jolla, CA, USA). *P* values less than 0.05 were indicated as statistical significance. Analysis of the excess of *de novo* and biallelic variants in a specific gene was performed according to previous reports[8,15] (detailed above in **Variants filtration and evaluation**). Bonferroni correction for *P*-values was applied for analysis on excess of variants. Given the general number of encoding genes of the human genome (19,711), a corrected *P* value (*Pc*) less than 2.54×10^6 (0.05/19,711) indicated a significant difference in genome-wide.

Results

General information from WES. We recruited 234 unrelated trios and a singleton (adopted) from 21 hospitals (Supplementary Table 1). The 235 patients included 164 boys and 71 girls, with an average onset age of 4.6 years old (4.6 ± 3.8, mean ± SD). On average, 5.9 Gb of the sequence was produced within the exome-targeted regions

in each individual, with an average coverage of 120 depths (Supplementary Table 2). To identify potential novel pathogenic variants, we employed an individualized analytical framework that includes analyses on each trio by explainable inheritance origin and stratified minor allele frequency (MAF) filtration and analyses on gene profile in four aspects (Fig. 1). After the variant filtration by inheritance origin and MAF (step 2 and 3 in Fig. 1), we obtained 1112 qualified variants in 930 genes, including 177 single *de novo* (15.9%), 77 homozygous (6.9%), 725 compound heterozygous (65.2%), and 133 hemizygous variants (12%) (Supplementary Table 3 and Extended Data Fig. 1a). The average number of variants per case was 4.8, including 0.8 *de novo* (ranging from 0 to 4), 0.3 homozygous (ranging from 0 to 5), and 3.1 compound heterozygous variants (ranging from 0 to 10); and hemizygous variants per male case was 0.8 (133/163, ranging from 0 to 5) (Extended Data Fig. 1b).

Potential pathogenic variants in epilepsy-associated genes. For known epilepsy-associated genes, variants were identified in 53 cases with 31 genes involved (Table 1 and Supplementary Table 4). Among these variants, 47 single *de novo* variants in 26 genes were evaluated as pathogenic or likely pathogenic by using American College of Medical Genetics and Genomics (ACMG) standards, explained 20% of cases (47/235) in this cohort. These genes are associated with epilepsy/DEE or development disorder with seizures (Table 1). The *CHD2* gene appeared as the top one with 10 patients affected (4.3%, 10/235).

It was noted that the pathogenic/likely pathogenic variants estimated by ACMG are all of *de novo*. In contrast, 6 recessive variants were evaluated as uncertain significance by ACMG (Table 1), including 2 homozygous, 1 compound heterozygous, and 3 hemizygous variants. Given that ACMG evaluates each of the paired biallelic variants separately and does not consider the specific dose effect of hemizygous variants, the pathogenicity of these variants was re-evaluated. These variants were considered to be possibly pathogenic by highly consistent clinical phenotype as previously reported, absence or very low MAF in general population (criteria in step 3 in Fig. 1), and the explainable inheritance origin (Supplementary Table 4).

Repetitiveness of potential pathogenic variants in novel candidate genes. For variants in genes that have not been previously reported to be associated with epilepsy/seizures, further filtration on gene profile was performed (step 4 in Fig. 1). Repetitive variants were identified in 29 genes with 73 cases involved (Table 1 and Supplementary Table 5). These genes met the inclusion criterion and against the exclusion criterion as novel candidate genes, among which 12 genes met additional two criteria of high probability of being intolerant to heterozygous/homozygous variants of loss-of-function (pLI/pRec), and abnormal phenotypes produced by knock-out/knock-down (KO/KD). Besides, variants were identified in 26 genes that met also the four criteria on gene profile but appeared in only single case in this cohort (Supplementary Table 6). These patients had no other pathogenic or likely pathogenic variants in known or novel candidate epilepsy genes.

For genes with *de novo* variants, the Poisson cumulative distribution function was calculated[8]. The *SBF1* presented significant excess of *de novo* variant in this cohort, which were significant after Bonferroni corrections (*Pc* = 0.012, Table 1 and Supplementary Table 7). Besides, among the 26 epilepsy-associated genes with *de novo* variants, significant excesses were detected in *CHD2* and *SETD1B* (*Pc* = 0 and *Pc* = 1.34×10^{-5}, respectively, Table 1 and Supplementary Table 7).

For genes with biallelic variants, the cumulative binomial probability was calculated[15]. Biallelic variants in *CELSR2* were identified in 8 cases, which was significantly more than expected by chance after Bonferroni correction (1% MAF cutoff for biallelic damaging missense genotypes in *CELSR2*, 8 versus 0.0029, *Pc* = 0.011). Biallelic *SPTBN5* variants were identified in 5 cases, which was not significantly more than expected by chance after Bonferroni correction (0.5% MAF cutoff for biallelic damaging missense genotypes in *SPTBN5*, 5 versus 0.0016, *Pc* = 0.87).

In order to analyze the significance of biallelic variants in the cases, we established a control cohort that included 1942 asymptomatic parents from trios. The frequencies of biallelic variants in 8 genes (*CELSR2*, *ADGRA1*, *CCDC66*, *AKAP13*, *CRMP1*, *MICAL2*, *SYNJ2*, and *TOGARAM2*) were significantly higher than that in asymptomatic the parent controls (Fig. 2 and Supplementary Table 8), among which 4 genes (*CELSR2*, *CRMP1*, *ADGRA1*, and *AKAP13*) met all the four criteria on gene profile (step 4 in Fig. 1).

Additionally, 5 hemizygous variants in *TENM1* were identified in 6 unrelated

cases (Table 1). These variants did not present as hemizygotes or homozygotes in the controls of general population.

Genetic-clinical features of the patients with variants in novel candidate genes.

SBF1. Four missense variants in *SBF1* (MIM*603560) were identified in four unrelated cases, including three *de novo* variants (c.337C>A/p.Gln113Lys, c.4459G>A/p.Gly1487Ser, and c.5424G>C/p.Trp1808Cys) and one with unknown origin (adopted, c.2272A>G/p.Ile758Val) (Fig. 3a and the DNA sequencing chromatograms in Extended Data Fig. 2a). The four variants were not present in the controls of gnomAD-all population, nor in our 296 healthy controls. Besides the significant excesses of *de novo* variants (Supplementary Table 7), the aggregate frequencies of the variants in the cases were significantly higher than that in the controls (4/468 vs. 0/109408 in the controls of gnomAD-all population, $P = 3.25 \times 10^{-10}$; vs. 0/9046 in the controls of gnomAD-East Asian population, $P = 5.78 \times 10^{-6}$; vs. 0/592 in our 296 normal controls, $P = 3.8 \times 10^{-2}$, Supplementary Table 9).

Three *de novo* missense variants (p.Gln113Lys, p.Gly1487Ser, and p.Trp1808Cys) were located at the functional domains, in the tripartite differentially expressed in normal and neoplastic cells (DENN) domain, the myotubularin phosphatase domain, and the pleckstrin homology (PH) domain, respectively (Fig. 3a). They were predicted to alter hydrogen bonding with surrounding residues (Extended Data Fig. 2b). The variant p.Ile758Val was located in the region between DENN and Glucosyltransferases, Rab-like GTPase activators and Myotubularins (GRAM)

domain (Fig. 3a) without hydrogen bond changes by protein modeling, but was predicted to decrease the protein stability (Extended Data Fig. 2b).

The clinical features of the four cases were summarized in Table 2. The three patients (LG90, LG57, and LG241) with *de novo* variants located in the functional domains exhibited intractable and frequent seizures even under combination of three or four anti-epileptic drugs (AEDs). They presented both generalized and multifocal discharges in EEGs. The patient (LG206) with the variant in the region between DENN and GRAM domain (p.Ile758Val) presented a relatively late onset age (11-15 years) with only generalized discharges in EEG. He achieved seizure-free with monotherapy of lamotrigine, despite previous daily myoclonic seizures.

CELSR2. Biallelic variants in *CELSR2* (MIM*604265) were identified in eight unrelated cases, including homozygous missense variant (c.7227C>A/p.His2409Gln) in two cases and six compound heterozygous missense variants (c.1763C>G/p.Pro588Arg & c.8545G>A/p.Glu2849Lys, c.3640A>T/p.Ser1214Cys & c.6108C>G/p.Phe2036Leu, c.4411C>A/p.Pro1471Thr & c.5969G>A/p.Arg1990His, c.6986G>A/p.Ser2329Asn & c.3640A>T/p.Ser1214Cys, c.8551A>G/p.Ser2851Gly & c.3329C>T/p.Thr1110Ile, and c.8624G>A/p.Arg2875Gln & c.6313A>T/p.Thr2105Ser) (Fig. 3b and the DNA sequencing chromatograms in Extended Data Fig. 3a). The variant p.Ser1214Cys was also recurrently identified in two unrelated cases. All heterozygous variants presented no or low allele frequencies in the gnomAD database (criteria in step 3 in Fig. 1). Besides the significant excess of recessive variants, the frequencies of biallelic variants in the cases were significantly

than that in the asymptomatic parent controls (Fig 2). Additionally, the aggregate frequencies of the variants in the cases were significantly higher than that in the controls (16/468 vs. 496/81210 in the controls of gnomAD-all population, $P = 6.61 \times 10^{-8}$; vs. 113/8124 in the controls of gnomAD-East Asian population, $P = 2.35 \times 10^{-3}$; vs. 3/592 in our 296 normal controls, $P = 5.91 \times 10^{-4}$, Supplementary Table 10).

Four of the twelve missense variants were located at the functional domains, including two (p.Pro588Arg and p.Thr1110Ile) in the cadherin EGF LAG seven-pass G-type receptor (CELSR) repeat domains, one (p.Pro1471Thr) in the Laminin G like domain, and one (p.Ser2329Asn) in the G-protein-coupled receptor proteolytic site (GPS) domain. The other eight variants were located in the nonfunctional regions, including one (p.Ser1214Cys) in the region between CELSR repeats and EGF-like domain, three (p.Arg1990His, p.Phe2036Leu, and p.Thr2105Ser) in the region between Laminin G-like and GPS domain, one (p.His2409Gln) located at the transmembrane region, and three (p.Glu2849Lys, p.Ser2851Gly, and p.Arg2875Gln) were located in the C-terminal intracellular region (Fig. 3b). Among the eight pairs of biallelic variants, at least one of the paired missense variants was predicted to affect hydrogen bonding or decrease the protein stability (Extended Data Fig. 3b).

The clinical features of the eight cases were listed in Table 2. The two cases (LG38 and LG231) with homozygous variant p.His2409Gln exhibited intractable seizures. In contrast, the two cases (LG242 and LG40) with compound heterozygous variants that had both two variants located at nonfunctional regions (p.Ser1214Cys & p.Phe2036Leu and p.Thr2105Ser & p.Arg2875Gln) achieved seizure-free with

monotherapy of valproate/lamotrigine. Among the other four cases with one of paired variants located at the functional domains, two cases (LG150 and LG238 who had variants p.Thr1110Ile and p.Ser2329Asn in functional domains, respectively) present refractory seizures; one case (LG230 with p.Pro1471Thr in Laminin G like domain) achieved seizure-free but had intellectual disability; and the other one (LG54) achieved seizure-free without intellectual disability, who had the compound heterozygous variants with two variants located furthest apart (p.Pro588Arg and p.Glu2849Lys, a distance away of 2261 amino acids), which have been suggested to be associated with phenotype severity[29,30].

TENM1. Five hemizygous missense variants in *TENM1* (MIM*10178) were identified in six unrelated cases, including c.467A>G/p.Asp156Gly, c.503G>A/p.Cys168Tyr, c.638C>T/p.Ala213Val, c.3326C>T/p.Thr1109Met, and c.5246T>C/p.Val1756Ala (Fig. 3c and the DNA sequencing chromatograms in Extended Data Fig. 4a). All the *TENM1* hemizygous variants were inherited from their asymptomatic mothers, consistent with a XLR inheritance pattern. No hemizygotes/homozygotes of the variants were present in the controls of gnomAD-all population. The aggregate frequencies of the variants in this cohort were significantly higher than that in the male controls (6/163 vs. 0/2549 in the hemizygotes controls of gnomAD-all population, $P = 4.32 \times 10^{-8}$; vs. 0/251 in the hemizygotes controls of gnomAD-East Asian population, $P = 3.5 \times 10^{-3}$, Supplementary Table 11).

Three variants (p.Asp156Gly, p.Cys168Tyr, and p.Ala213Val) were located at the N-terminal intracellular teneurin domain, one variant (p.Thr1109Met) was located

in the extracellular region between EGF-like repeats and ncl-1, HT2A and lin-41 (NHL) repeats, and one variant (p.Val1756Ala) was located in the extracellular region between Tyr and Asp dipeptide (YD) 5 and YD 6 domains (Fig. 3c).

Among the three cases with variants in the N-terminal intracellular teneurin domain, two cases (LG239 and LG60) suffered from refractory seizures even under three AEDs. In contrast, the three cases (LG28, LG189, and LG270) with variants located in the nonfunctional regions achieved seizure-free under the combination therapy of valproate and lamotrigine.

Seizure activity in *Sbf*, *Flamingo*, and *Ten-m* knockdown *Drosophila*. To validate the association between the three novel candidate genes and epilepsy, we established gene knockdown model in *Drosophila* by RNA interference (RNAi). The Drosophila homologs of human *SBF1*, *CELSR2*, and *TENM1* genes are *Sbf*, *Flamingo*, and *Ten-m*, respectively. RNAi efficiency of the three genes in flies was detected by reverse transcription quantitative PCR (RT-qPCR). Compared to WT flies, mRNA expression of the three genes in the knockdown flies was reduced by 47% (*Sbf*), 75% (*Flamingo*), and 79% (*Ten-m*), respectively (Extended Data Fig. 5). We examined the bang-sensitivity (BS) seizure-like behavior in *tub-Gal4>Sbf-*, *Flamingo-*, and *Ten-m-RNAi* flies. Three phases of typical seizure-like behavior in the BS test, including seizure, paralysis, and recovery, were observed in the knockdown flies (Fig. 4a). The seizure-like behaviors were detected in 40.2% (*Sbf*), 24.8% (*Flamingo*), and 33.0% (*Ten-m*) of the knockdown flies, which was 8, 3, and 4 times higher than the

control flies, respectively (Fig. 4b). The gene knockdown flies also showed longer duration of seizure-like behavior than the controls. The knockdown flies recovered from seizures within 30 (*Sbf*), 7 (*Flamingo*), and 20 (*Ten-m*) seconds, while the control flies recovered within 3, 5, and 5, respectively (Fig. 4c).

Projection neurons (PNs) are important excitatory neurons in the central nervous system of *Drosophila*[31]. We examined the effect of the gene deficiencies on the electrophysiological activity of PNs by using whole-cell recording (Fig. 5a). The *Sbf*, *Flamingo*, and *Ten-m* knockdown flies showed regular bursts of firing in the PNs (Fig. 5b). The frequencies of spontaneous excitatory postsynaptic potentials (sEPSPs) in knockdown flies were significantly higher than that in WT flies (Fig. 5c). There was no significant difference in amplitude of sEPSPs between the knockdown flies and WT flies (Fig. 5d). These results suggested that knockdown of the *Sbf*, *Flamingo*, and *Ten-m* resulted in significantly increased seizure susceptibility and epileptic electrophysiological activity in *Drosophila*.

Seizure activity in *sbf1* knockout zebrafish. To further determine the potential pathogenic roles of *SBF1*, we generated a *sbf1* (the zebrafish homologs of *SBF1*) knockout model using CRISPR/Cas9 technology in zebrafish. The SET binding factor 1 (encoded by *SBF1*) in human and zebrafish are 70% identical. Germline mutation with 4-nt (TGGA) deletion was introduced in exon 3 of the primary *sbf1* transcript (RefSeq: NM_001045158.1), resulting in a premature stop codon occurred in exon 5 (Extended Data Fig. 6a). Using RT-qPCR, we confirmed that there was an absence of

88% of total *sbf1* mRNA in *sbf1*$^{-/-}$ and 41% of total *sbf1* mRNA in *sbf1*$^{+/-}$ zebrafish, proved the efficiency of the *sbf1* knockout (Extended Data Fig. 6b). At 5 days post-fertilization (d.p.f.), the morphological abnormalities, including lack of a swim bladder, severe body curvature, and small body size were observed in 1.1% of *sbf1*$^{-/-}$ and 0.9% of *sbf1*$^{+/-}$ larvae (Fig. 6a, b). At 17 d.p.f., all of *sbf1*$^{-/-}$ zebrafish were dead, while *sbf1*$^{+/-}$ and *sbf1*$^{+/+}$ siblings remained alive. About 62.5% of *sbf1*$^{+/-}$ and 83.3% of *sbf1*$^{+/+}$ zebrafish were survival to 30 d.p.f., suggesting an essential role of *sbf1* in survival (Fig. 6c).

We next examined the seizure-like behavior in *sbf1* knockout zebrafish. Seizure-like behavior characterized by high-speed (> 20 mm/sec) "whirlpool-like" circling swimming was observed in the *sbf1*$^{+/-}$ and *sbf1*$^{-/-}$ larvae at 7 d.p.f. (Fig. 6d). It was noted that the *sbf1*$^{+/-}$ larvae exhibited more hyperactive/seizure-like behavior with a longer distance of high-speed movement than the *sbf1*$^{+/+}$ larvae (Fig. 6e), with significant differences between *sbf1*$^{+/-}$ and *sbf1*$^{+/+}$ larvae in the 20 to 60 min observations. In contrast, the *sbf1*$^{-/-}$ larvae exhibited more hyperactive/seizure-like behavior with significant longer distance of high-speed movement than the *sbf1*$^{+/+}$ larvae at only the 60 min observation (Fig. 6e). Similarly, the *sbf1*$^{+/-}$ larvae exhibited longer duration of high-speed movement than the *sbf1*$^{+/+}$ larvae (Fig. 6f). We further performed electrophysiological recording in the forebrain of larvae at 7-10 d.p.f. as reported previously[25]. The spontaneous epileptiform discharges were recorded in *sbf1*$^{-/-}$ and *sbf1*$^{+/-}$ larvae but not in *sbf1*$^{+/+}$ larvae (Fig. 6g). Both the *sbf1*$^{-/-}$ and *sbf1*$^{+/-}$ larvae exhibited more number and duration of discharges than the *sbf1*$^{+/+}$ larvae (Fig.

6h, i). These results suggest that deficiency of *sbf1* is potentially associated with seizure-like electrophysiological activity in zebrafish.

Spontaneous seizure and electroencephalography recording in *Celsr2* knock-out mice. *Celsr2*, a core planar cell polarity (PCP) component, plays versatile roles in neural development, including cilia organization, neuron migration, and axon navigation[32]. To validate the potential role of *CELSR2* in seizures, we generated a *Celsr2* knockout model by inserting a targeting vector containing an internal ribosomal entry site (IRES) and the LacZ gene into exon 23 of the *Celsr2* gene to interrupt its transcript as described previously[26]. The *Celsr2*$^{+/+}$ mice did not show seizures or epileptiform discharges on video-EEG recordings (Fig. 7a). In contrast, spontaneous seizures with typical EEG alterations were observed in 4 of the *Celsr2*$^{-/-}$ mice (4/15, 26.7%, Fig. 7b, c, and Supplementary Video). The average duration of seizure activity in the *Celsr2*$^{-/-}$ mice was 194.9 ± 168.0 seconds (mean ± SD, Fig. 7d). Among the four *Celsr2*$^{-/-}$ mice with spontaneous seizures, two mice exhibited seizures to stage 3. The other two mice displayed seizures up to stage 5 and followed by death (Fig. 7e). Analysis on the power spectra of the background EEGs showed that the *Celsr2*$^{-/-}$ mice had higher spectral power, characterized by high-amplitude and low-frequency oscillations (Fig. 7f).

Discussion

The development of high-throughput sequencing techniques has enabled genetic

studies with large cohorts. However, these studies have encountered challenges in discovering novel pathogenic genes[33-35] in recent years. The gene-disease associations in more than three fourth of the human genome remain undetermined (https://omim.org). The present study focused on LGS, an epileptic encephalopathy of childhood-onset, and used trio-based WES with individualized analysis before statistic studies, which includes individualized analysis on each trio by explainable inheritance origin and stratified MAF filtration and individualized analysis on each gene from four aspects. Three novel candidate genes for the LGS were identified, including *SBF1* with *de novo* origin, *CELSR2* with biallelic recessive inheritance, and *TENM1* with X-linked recessive inheritance. These genes presented significant repetitiveness, including significant higher excess of *de novo* variants in *SBF1* and excess of biallelic variants in *CELSR2*, and aggregated frequencies of variants in *SBF1*, *CELSR2*, and *TENM1* comparing with that in the controls of gnomAD population. The frequency of compound heterozygous/homozygous *CELSR2* variants in the cases was significantly higher than that in the asymptomatic parent-controls. Animal models with knock-down or knock-out of these genes exhibited increased seizure-like behavior and increased firing of excitatory neurons. This study suggested that *SBF1*, *CELSR2*, and *TENM1* were potentially pathogenic genes associated with LGS and highlights the value of phenotype sub-classification and individualized analysis in combination with statistic studies in identifying novel pathogenic genes, including use of controls for compound heterozygous variants.

Epilepsy syndromes are featured by age-dependent clinical manifestations,

especially the onset age[1], which are potentially associated with distinct pathogenic genetic causes[13]. The present study focused on LGS, the epileptic encephalopathy of childhood-onset with characteristic clinical manifestations, differing from most of the DEEs that are commonly of early onset and used to be generally named EIEE. Consistently, this study identified three novel potential pathogenic genes, i.e., *SBF1*, *CELSR2*, and *TENM1*, which accounted for 7.7% (18/235) of cases, in addition to *CHD2* that was ranked at top one of the pathogenic genes (10/235) in this cohort, differing from the DEEs in which *SCN1A* was the most common pathogenic gene[33-35]. These findings reinforce the implication of phenotype sub-classification by clinical features in genetic studies of large cohorts, such as the onset age and characteristic clinical features in LGS.

Statistical analysis is generally used to detect potential pathogenic variants/genes in genetic studies with large cohort[33-35]. This study employed an individualized analysis procedure before statistical burden analyses, which potentially improves the specificity and sensitivity. After general filtrations, variants with explainable inheritance origin in each trio were selected, which potentially improve the specificity. The variants were then filtrated with MAF criteria according to the inheritance pattern. The MAF of *de novo* variants, hemizygotes, and homozygotes is set as absent in the control populations in gnomAD; and the filtration MAF for compound heterozygotes is set at $< 1\times10^{-6}$ for the product of multiplying the frequencies of two alleles in gnomAD. This measure aimed to avoid omission of single low MAF variant in biallelic variants and potentially improved the sensitivity. Such single heterozygous

variant with low MAF alone is not pathogenic, but is potentially pathogenic when appeared as one of the paired biallelic variants, including those in homozygotes/hemizygotes. For the gene profile filtration, the four criteria on gene potentially improve the specificity in identifying novel candidate genes. For the compound heterozygous variants, it is currently unknown how frequent compound heterozygous variants are present in normal populations. We therefore established a control cohort that included 1942 asymptomatic parents from trios for analysis on compound heterozygous variants. With the supporting evidence from other statistical analysis, *SBF1*, *CELSR2,* and *TENM1* were selected as novel candidate genes for LGS and subjected to further studies.

SBF1 (MIM*603560) encodes a member of the protein-tyrosine phosphatase family, which plays a vital role in cell growth and differentiation[36]. Previously, *SBF1* recessive variants were reported to be associated with Charcot-Marie-Tooth disease type 4B3 (CMT4B3, OMIM*615284) that is featured by progressive limb muscle weakness and distal sensory impairment interfering[37]. The *SBF1* expresses in the brain and the muscle at a similar level, suggesting also functional roles in the brain. In mice, homozygous knock-out *Sbf1* causes postnatal lethality and abnormal brain morphology. In this study, three *de novo* heterozygous *SBF1* missense variants were identified, which showed a significant excess of *de novo* variants and significantly higher aggregated frequencies of variants in the LGS cohort than that in the controls of gnomAD populations. In animal experiments, *Sbf1* knock-down flies and knock-out zebrafish showed seizure-like behavior and increased firing of excitatory

neurons. Clinically, *SBF1* variants in the functional domains were associated with relatively severe phenotype, whereas the variant between domains resulted in late-onset seizure with good prognosis, indicating a genotype-phenotype correlation. These findings suggested that *SBF1* is potentially associated with epilepsy. The previously reported CMT4B3-associated variants were all biallelic recessive variants, including biallelic missense variants and biallelic truncation variants (Extended Data Fig. 7); in contrast, the LGS-associated *SBF1* variants were all heterozygous missense variants. The probability of being loss of function intolerant (pLI) of *SBF1* is 1.00, indicating that *SBF1* is intolerant to heterozygous loss-of-function variants. It is noted that the heterozygous *Sbf1* knock-out zebrafish presented more hyperactive/seizure-like behavior than the homozygous knock-out larvae, potentially suggesting an association between epilepsy and heterozygous *SBF1* variants. Gene expression across multiple tissues is common and subsequently is potentially associated with multiple phenotypes with distinct mechanism, as demonstrated in our recent study on *PCDH15*, which is associated with deafness and epilepsy with underlying mechanisms related to gene expression stage[38]. The specific mechanisms underlying the distinct phenotypes of the brain and muscles for *SBF1* warrant future studies.

CELSR2 (MIM*604265) encodes cadherin EGF LAG seven-pass G-type receptor 2 (CELSR2), a transmembrane protein of the flamingo subfamily. Our previous studies have showed that two CELSR subfamily genes, *CELSR1* and *CELSR3*, were associated with epilepsies[39,40]. *CELSR2* is highly expressed in the

brain with predominance in early life and plays a vital role in cell/cell signaling during nervous system formation[32]. The loss of function observed/expected upper bound fraction (LOEUF) of *CELSR2* is 0.25 (LOEUF < 0.35) and the pLI is 1.00, indicating *CELSR2* variants with loss of function are potentially pathogenic. In this study, we identified eight pairs of biallelic *CELSR2* variants in the LGS cohort, which showed significant excess of biallelic variants in genome-wide, significantly higher aggregated frequencies of variants than that in the controls of gnomAD populations, and significantly higher frequency of compound heterozygous/homozygous variants in the cases than that in asymptomatic parent controls. The *CELSR2* homozygous variants and compound heterozygous variants with one of paired variants located at the functional domains were associated with severe phenotype, whereas compound heterozygous variants with both two variants located at nonfunctional regions were associated with relative milder phenotype, suggesting a genotype-phenotype correlation. The *Celsr2* knock-down flies and knock-out mice showed seizure-like behavior and increased firing of excitatory neurons. These findings suggest that *CELSR2* is potentially associated with epilepsy.

TENM1 (MIM*10178) encodes a transmembrane teneurin subfamily protein, which is highly expressed in the brain, predominantly during prenatal and mature stage. TENM1 is involved in the neural development by regulating the establishment of proper connectivity within the nervous system[41,42]. The LOEUF of *TENM1* gene is 0.19 (LOEUF < 0.35) and the pLI is1.00, indicating loss of function of *TENM1* is potentially pathogenic. This study identified hemizygous *TENM1* variants in six

unrelated cases. No hemizygotes/homozygotes of the variants were present in the controls of gnomAD populations, and the aggregate frequencies of the hemizygotes were significantly higher than that in the male controls. Two patients with hemizygous variants located in the N-terminal intracellular teneurin domain presented refractory seizures, while the patients with variants located in nonfunctional regions achieved seizure-free, suggesting a genotype-phenotype correlation. The *Tenm1* knock-down flies exhibited increased seizure-like behavior rate and increased excitability of neurons in the present study. These findings indicated that *TENM1* was potentially associated with epilepsy.

Additional 35 novel genes (in 44 cases) met the six criteria in individualized analysis, including inheritance pattern and stratified MAF, gene expression in the brain, defined genotype not against epilepsy, high pLI/pRec, and abnormal phenotypes produced by KO/KD. Majority (82.9%, 29/35) of the genes showed recessive inheritance pattern, among which three genes (*ADGRA1*, *AKAP13*, and *CRMP1*) presented higher frequencies of recessive variants comparing the asymptomatic parent controls. These genes are potentially candidate genes of epilepsy, for which further studies are required.

Among the novel candidate genes, *SBF1* presents dominance inheritance (heterozygous *de novo* variants in three cases), and the rest majority candidate genes present recessive inheritance, including XLR. This could be basically understood by the quantitative dependent feature of genes[43,44]. We defined the lowest limit of the required quantity of genetic function of the gene as genetic dependent quantity (GDQ).

If the full scale of GDQ was set at 100% with contributions from two copies of a gene, the genes with GDQ < 50% are much more (https://gdap.org.cn/)[43,44], since that human genome is diploid and usually a portion of functioning from one copy of a gene is sufficient for the biophysiological function. Subsequently, pathogenic recessive (including X-linked recessive) variants are potentially common. By utilizing trio-based sequencing, this study was designed to detect variants of different origin, including those of recessive origin, which occur in genes with GDQ < 50%. The subsequent stratified MAF filtration, instead of a clear cut-off of MAF as "absent from populations", ensured the inclusion of heterozygous variants of low MAF, of which single variant alone is not pathogenic but is potentially pathogenic when appeared as one of the paired biallelic variants. In fact, single heterozygous variant of "damaging" in recessive disorders is not "pathogenic", thus may be prevalent with low MAF in general populations, typically as p. Arg208Ter (MAF of 0.00025) and c.509-1G>C (MAF of 0.00039) in *TPP1*, which were identified in 89% of patients with ceroid lipofuscinosis[45], a severe neurodegenerative disease with epilepsy (https://www.ucl.ac.uk/ncl-disease/mutation-and-patient-database/mutation-and-patient-datasheets-human-ncl-genes/cln2-tpp1). Therefore, special attention is required to recessive variants in designing protocol to identify the underlying causes of genetic disorders, including the application of controls for compound heterozygous variants.

References

1. Wirrell, E.C. *et al.* Methodology for classification and definition of epilepsy syndromes with list of syndromes: Report of the ILAE Task Force on Nosology and Definitions. *Epilepsia* **63**, 1333-1348 (2022).
2. Eltze, C.M. *et al.* A population-based study of newly diagnosed epilepsy in infants. *Epilepsia* **54**, 437-45 (2013).
3. Asadi-Pooya, A.A. Lennox-Gastaut syndrome: a comprehensive review. *Neurol Sci* **39**, 403-414 (2018).
4. Jahngir, M.U., Ahmad, M.Q. & Jahangir, M. Lennox-Gastaut Syndrome: In a Nutshell. *Cureus* **10**, e3134 (2018).
5. Specchio, N. *et al.* International League Against Epilepsy classification and definition of epilepsy syndromes with onset in childhood: Position paper by the ILAE Task Force on Nosology and Definitions. *Epilepsia* **63**, 1398-1442 (2022).
6. Sheidley, B.R. *et al.* Genetic testing for the epilepsies: A systematic review. *Epilepsia* **63**, 375-387 (2022).
7. Ruggiero, S.M., Xian, J. & Helbig, I. The current landscape of epilepsy genetics: where are we, and where are we going? *Curr Opin Neurol* **36**, 86-94 (2023).
8. Epi, K.C. *et al.* De novo mutations in epileptic encephalopathies. *Nature* **501**, 217-21 (2013).
9. Carvill, G.L. *et al.* Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. *Nature Genetics* **45**, 825-830 (2013).
10. Euro, E.-R.E.S.C., Epilepsy Phenome/Genome, P. & Epi, K.C. De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. *Am J Hum Genet* **95**, 360-70 (2014).
11. Shi, Y.W. *et al.* Synaptic clustering differences due to different GABRB3 mutations cause variable epilepsy syndromes. *Brain* **142**, 3028-3044 (2019).
12. Wang, J. *et al.* UNC13B variants associated with partial epilepsy with favourable outcome. *Brain* **144**, 3050-3060 (2021).
13. Wang, J. *et al.* Epilepsy-associated genes. *Seizure* **44**, 11-20 (2017).
14. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405-24 (2015).
15. Martin, H.C. *et al.* Quantifying the contribution of recessive coding variation to developmental disorders. *Science* **362**, 1161-1164 (2018).
16. consortium, C. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588-91 (2015).
17. Ni, J.Q. *et al.* A genome-scale shRNA resource for transgenic RNAi in Drosophila. *Nat Methods* **8**, 405-7 (2011).
18. Parker, L., Howlett, I.C., Rusan, Z.M. & Tanouye, M.A. Seizure and epilepsy: studies of seizure disorders in Drosophila. *Int Rev Neurobiol* **99**, 1-21 (2011).
19. Perkins, K.L. Cell-attached voltage-clamp and current-clamp recording and stimulation techniques in brain slices. *J Neurosci Methods* **154**, 1-18 (2006).
20. Gu, H. & O'Dowd, D.K. Cholinergic synaptic transmission in adult Drosophila Kenyon cells in situ. *J Neurosci* **26**, 265-72 (2006).

21. Pena, I.A. *et al.* Pyridoxine-Dependent Epilepsy in Zebrafish Caused by Aldh7a1 Deficiency. *Genetics* **207**, 1501-1518 (2017).
22. Hwang, W.Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat Biotechnol* **31**, 227-9 (2013).
23. Sander, J.D. *et al.* ZiFiT (Zinc Finger Targeter): an updated zinc finger engineering tool. *Nucleic Acids Res* **38**, W462-8 (2010).
24. Baraban, S.C., Taylor, M.R., Castro, P.A. & Baier, H. Pentylenetetrazole induced changes in zebrafish behavior, neural activity and c-fos expression. *Neuroscience* **131**, 759-68 (2005).
25. Baraban, S.C. Forebrain electrophysiological recording in larval zebrafish. *J Vis Exp* (2013).
26. Tissir, F. *et al.* Lack of cadherins Celsr2 and Celsr3 impairs ependymal ciliogenesis, leading to fatal hydrocephalus. *Nat Neurosci* **13**, 700-7 (2010).
27. Lin, T.W. *et al.* Regulation of Synapse Development by Vgat Deletion from ErbB4-Positive Interneurons. *J Neurosci* **38**, 2533-2550 (2018).
28. Chen, W. *et al.* Neddylation stabilizes Nav1.1 to maintain interneuron excitability and prevent seizures in murine epilepsy models. *J Clin Invest* **131**(2021).
29. Luo, S. *et al.* Recessive LAMA5 Variants Associated With Partial Epilepsy and Spasms in Infancy. *Front Mol Neurosci* **15**, 825390 (2022).
30. Luo, J. *et al.* MPDZ variants associated with epilepsies and/or febrile seizures and the individualized genotype-phenotype correlation. *Genes & Diseases* (2023).
31. Olsen, S.R., Bhandawat, V. & Wilson, R.I. Excitatory interactions between olfactory processing channels in the Drosophila antennal lobe. *Neuron* **54**, 89-103 (2007).
32. Wen, Q. *et al.* Inactivating Celsr2 promotes motor axon fasciculation and regeneration in mouse and human. *Brain* **145**, 670-683 (2022).
33. Epi25 Collaborative. Electronic address, s.b.u.e.a. & Epi, C. Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals. *Am J Hum Genet* **105**, 267-282 (2019).
34. Epi25, C., Chen, S., Neale, B.M. & Berkovic, S.F. Shared and distinct ultra-rare genetic risk for diverse epilepsies: A whole-exome sequencing study of 54,423 individuals across multiple genetic ancestries. *medRxiv* (2023).
35. Epi25 Collaborative. Electronic address, j.c.c.e. & Epi, C. Sub-genic intolerance, ClinVar, and the epilepsies: A whole-exome sequencing study of 29,165 individuals. *Am J Hum Genet* **108**, 965-982 (2021).
36. Cui, X. *et al.* Association of SET domain and myotubularin-related proteins modulates growth control. *Nat Genet* **18**, 331-7 (1998).
37. Nakhro, K. *et al.* SET binding factor 1 (SBF1) mutation causes Charcot-Marie-Tooth disease type 4B3. *Neurology* **81**, 165-73 (2013).
38. Lin, S.M. *et al.* PCDH15 in self-limited partial epilepsy and the mechanism of distinct phenotypes. *Annals of Neurology* **Revised**(2023).
39. Chen, Z. *et al.* CELSR1 variants are associated with partial epilepsy of childhood. *Am J Med Genet B Neuropsychiatr Genet* **189**, 247-256 (2022).
40. Li, J. *et al.* CELSR3 variants are associated with febrile seizures and epilepsy with antecedent febrile seizures. *CNS Neurosci Ther* **28**, 382-389 (2022).

41. Minet, A.D., Rubin, B.P., Tucker, R.P., Baumgartner, S. & Chiquet-Ehrismann, R. Teneurin-1, a vertebrate homologue of the Drosophila pair-rule gene ten-m, is a neuronal protein with a novel type of heparin-binding domain. *J Cell Sci* **112 (Pt 12)**, 2019-32 (1999).
42. Mosca, T.J., Hong, W., Dani, V.S., Favaloro, V. & Luo, L. Trans-synaptic Teneurin signalling in neuromuscular synapse organization and target choice. *Nature* **484**, 237-41 (2012).
43. Li, B. *et al.* Genetic Dependence and Genetic Diseases. *Unpublished* (2023).
44. He, N. *et al.* Common genetic epilepsies, pathogenicity of genes/variants, and genetic dependence. *Seizure* **109**, 38-39 (2023).
45. Kousi, M., Lehesjoki, A.E. & Mole, S.E. Update of the mutation spectrum and clinical correlations of over 360 mutations in eight genes that underlie the neuronal ceroid lipofuscinoses. *Hum Mutat* **33**, 42-63 (2012).

Privacy statement

We declared that all patients' IDs were not identified to anyone outside the research group.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgements

We thank the affected patients and their families for participating in this study. This work was funded by the National Natural Science Foundation of China (grant Nos. 82171439, 82271505, and 81971216), and Guangdong Basic and Applied Basic Research Foundation (grant No. 2021A1515010986). The funders had no role in study design, data collection, and analysis, or in the decision to publish or the preparation of the manuscript.

Author Contributions

WPL and YWS conceptualized the study, analyzed and interpreted the data, and drafted and revised the manuscript. WPL had full access to all data in the study and take responsibility for the integrity of the data and accuracy of the data analysis. JGZ, NH, ZLY, NXS, and HKL performed the clinical data and whole exome sequencing analysis. WBL, XCQ, and CXF performed seizure activity studies. WPL, YWS, JGZ, NH, ZLY, NXS, HKL, WBL, XCQ, and CXF drafted and revised the manuscript, and contributed to the statistical analysis. All authors collected data, revised the manuscript, and contributed to the writing.

Competing interests

All authors claim that there are no conflicts of interest.

Ethics statement

All procedures performed were in accordance with the ethical standards of the institutional committee. Ethical approval was approved by the ethics committee of the Second Affiliated Hospital of Guangzhou Medical University (approval ethics number 2020-hs-49).

Table 1. Summary of potential pathogenic variants

Known epilepsy-associated genes (number of cases, 53 in total)

Genes with pathogenic or likely pathogenic variants (47):

Epilepsy/DEE genes (24):

CHD2 (10)[a], *SETD1B* (4)[a], *CELF2* (2), *PCDH19* (2), *GABRB3* (1), *GRIN2B* (1), *KCNB1* (1), *NSF* (1), *SCN1A* (1), *STXBP1* (1).

Genes associated with developmental disorder with seizures (23):

ATP1A3 (2), *CACNA1B*[b] (2), *NUS1* (2), *SRCAP* (2), *SYNGAP1* (2), *TSC1* (2), *ZMYM2* (2), *COL4A2* (1), *GRIA1* (1), *KIF1A* (1), *KMT2D* (1), *MACF1* (1), *NEXMIF* (1), *PPP2CA* (1), *SIN3A* (1), *TSC2* (1).

Genes with possible pathogenic variants[c] (6):

KIF4A (2), *ANK3* (1), *CNPY3* (1), *SLC7A6OS* (1), *TAF1* (1).

Novel candidate genes with repetitive variants (number of cases, 73 in total)

***De novo* variants (5):**

SBF1 (3)[a], *SRRM2* (2).

Homozygous/compound heterozygous variants (53):

CELSR2 (8)[a,d], *SPTBN5* (5), *ADGRA1* (2)[d], *CCDC66* (2)[d], *AKAP13* (2)[d], *CFAP46* (2), *CMYA5* (2), *CRACR2A* (2), *CRMP1* (2)[d], *FAT1* (2), *FBN3* (2), *LAMA1* (2), *LRP2* (2), *MICAL2* (2)[d], *MYO7B* (2), *PDZD2* (2), *SEC16A* (2), *STARD9* (2), *SYNJ2* (2)[d], *TOGARAM2* (2)[d], *UTRN* (2), *ZNF721* (2).

X-linked recessive variants (15):

TENM1 (6), *ZNF674* (3), *AMER1* (2), *MAGEC1* (2), *NRK* (2).

[a]Genes with significant excess of variants in this cohort than controls ($Pc < 0.05$ after Bonferroni correction). [b]The two variants in *CACNA1B* gene were of *de novo*, while previously reported variants were of recessive. [c]The variants in these genes were all recessive and evaluated as uncertain significance by ACMG, but were considered to be possibly pathogenic by highly consistent clinical phenotype as previously reported, absence or very low MAF in general population (criteria in step 3 in Fig. 1), and explainable inheritance origin. [d]The frequencies of genes with biallelic variants in this cohort were significantly higher than that in the asymptomatic parent controls ($P < 0.05$).

The underline indicates the genes that met also the four criteria on gene profile in step 4 for filtration of novel candidate genes (Fig. 1).

Table 2. Clinical features of the patients with variants in *SBF1*, *CELSR2*, and *TENM1*

No.	Variants	Sex	Age	Onset age	Seizure course	EEG	Brain MRI	Development	AEDs	Prognosis
***SBF1* (*603560; NM_002972.4; autosomal dominant)**										
LG90	c.337C>A/ p.Gln113Lys	M	Teenagers	Toddlers	Spasms 2-4 times/d for 1 mo, tonic 3-20 times/d	GPSW; multifocal SW; Ictal: spasms, tonic	NA	ID	VPA, TPM, LTG	Refractory
LG57	c.4459G>A/ p.Gly1487Ser	F	Teenagers	Childhood	aAb 2-3 times/d, CPS 10 times/d	Fast rhythm during sleep; generalized 1.5-2.5 Hz SW; multifocal spikes	Normal	Normal	VPA, LTG, PHT	Refractory
LG241	c.5424G>C/ p.Trp1808Cys	M	Teenagers	Infants	Tonic10-20 times/d; GTCS 4-5 times/d	GPSW; multifocal SW	Normal	Normal	VPA, PB, CBZ, PHT	Refractory
LG206	c.2272A>G/ p.Ile758Val[a]	M	Teenagers	Childhood	Myoclonic daily; aAb; GTCS once	Fast rhythm during sleep; GPSW; generalized 1.5-2.5 Hz SW	Normal	Normal	LTG	Seizure free
***CELSR2* (*604265; NM_001408.2; autosomal recessive)**										
LG38	c.7227C>A/ p.His2409Gln c.7227C>A/ p.His2409Gln	M	Teenagers	Childhood	GTCS, 1-2 times/mo	Fast rhythm during sleep; GPSW	Normal	Normal	CBZ, VPA	Refractory
LG231	c.7227C>A/ p.His2409Gln c.7227C>A/ p.His2409Gln	M	Teenagers	Toddlers	Tonic, CPS, sGTCS, 6-15 times/d	Generalized polyspikes; GPSW; right frontal SW; Ictal: tonic, CPS	Normal	Normal	VPA, LTG, OXC, TPM	Refractory
LG54	c.1763C>G/ p.Pro588Arg c.8545G>A/ p.Glu2849Lys	F	Teenagers	Childhood	CPS, sGTCS 1-2 times/mo	Generalized fast rhythm during sleep; GPSW	Normal	Normal	LEV, LTG	Seizure free
LG242	c.3640A>T/	M	Child	Toddle	Status of tonic, FS, GTCS	Fast rhythm during sleep; GPSW	Normal	Normal	VPA	Seizure free

	p.Ser1214Cys c.6108C>G/ p.Phe2036Leu		hood	rs	2-6 times/yr					
LG230	c.4411C>A/ p.Pro1471Thr c.5969G>A/ p.Arg1990His	F	Teenagers	Childhood	Tonic daily; GTCS 1-2 times/mo	Generalized fast rhythm during sleep; GPSW; right frontal-temporal SW	Normal	ID	VPA, LTG	Seizure free
LG238	c.6986G>A/ p.Ser2329Asn c.3640A>T/ p.Ser1214Cys	M	Childhood	Toddlers	Spasms and tonic 2-3 times/d	Generalized fast rhythm during sleep; GPSW; left frontal-temporal SW; Ictal: spasms, tonic	Normal	DD, ID	TPM, VPA, CNZ	Refractory
LG150	c.8551A>G/ p.Ser2851Gly c.3329C>T/ p.Thr1110Ile	F	Childhood	Infants	Spasms 2-4 times/d for 8 mos, tonic 4-5 times/d	Generalized polyspikes, 1-2 Hz GPSW; Ictal: spasms, tonic	Normal	DD, ID	VPA, LEV	Refractory
LG40	c.8624G>A/ p.Arg2875Gln c.6313A>T/ p.Thr2105Ser	M	Childhood	Childhood	Myoclonic daily; sGTCS 6-7 times	Generalized polyspikes, 2.5-3 Hz GPSW; Ictal: myoclonic	Normal	Normal	LTG	Seizure free
***TENM1* (*300588; NM_001163278.1; X-linked recessive)**										
LG239	c.467A>G/ p.Asp156Gly	M	Childhood	Childhood	GTCS 2-3 times/wk, CPS 1-2 times/mo	Generalized polyspikes; GPSW; multifocal SW; Ictal: CPS	Normal	DD, ID	LEV, VPA, LCM	Refractory
LG10	c.503G>A/ p.Cys168Tyr	M	Teenagers	Childhood	Tonic 2-4 times/d; FS once, GTCS 3 times/mo	Generalized fast rhythm during sleep; GPSW; Ictal: tonic	Normal	Normal	LTG, VPA, LEV	Seizure free
LG60	c.638C>T/ p.Ala213Val	M	Teenagers	Childhood	CPS or aAb daily	Generalized fast rhythm during sleep; multifocal SW	Normal	Normal	LTG, VPA, LEV	Refractory
LG28	c.3326C>T/	M	Teena	Childh	Tonic and atonic 10-15 times	GPSW;	Normal	ID	VPA, LTG	Seizure free

	p.Thr1109Met		gers	ood	/d	Ictal: tonic, atonic				
LG189	c.3326C>T/ p.Thr1109Met	M	Child hood	Childh ood	Tonic 4-12 times/d; CPS 4-10 times/yr	Generalized fast rhythm during sleep; GPSW; left temporal SW; Ictal: tonic, CPS	Normal	Normal	VPA, LTG	Seizure free
LG270	c.5267T>C/ p.Val1756Ala	M	Child hood	Childh ood	Tonic daily; GTCS 3-4 times	Generalized fast rhythm during sleep; GPSW; multifocal polyspikes and SW	Normal	DD, ID	VPA, LTG	Seizure free

aAb, atypical absences; AEDs, antiepileptic drugs; CBZ, carbamazepine; CNZ, clonazepam; CPS, complex partial seizure; d, day; DD, development delay; EEG, electroencephalogram; F, female; FS, febrile seizure; GPSW, generalized polyspike-slow waves; GTCS, generalized tonic-clonic seizure; ID, intelligence disorder; LEV, levetiracetam; LTG, lamotrigine; M, male; mo, month; MRI, magnetic resonance imaging; NA, not available; OXC, oxcarbazepine; PB, phenobarbital; PHT, phenytoin; PSW, polyspike-slow waves; sGTCS, secondary generalized tonic–clonic seizure; SW, spike-slow waves; TPM, topiramate; VPA, valproate; wk, week; yr, year.

[a] Variant with unknown origin.

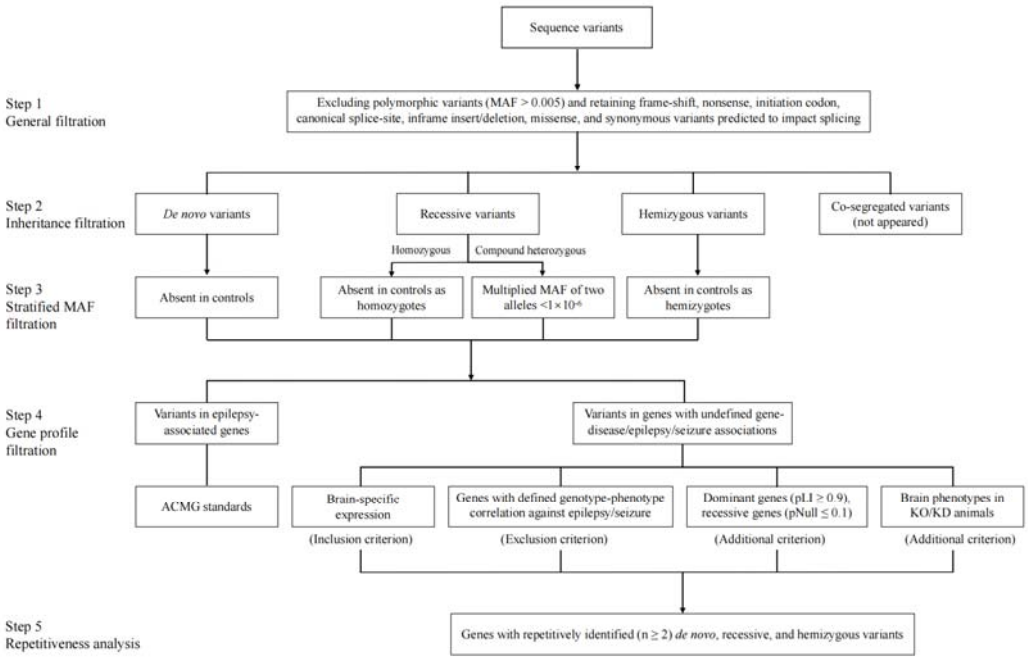


Fig. 1 | A flow chart of the trio-based whole-exome sequencing with individualized analytical framework. Abbreviations: ACMG, American College of Medical Genetics and Genomics guideline; KO/KD, knock-out/knock-down; MAF, minor allele frequency.

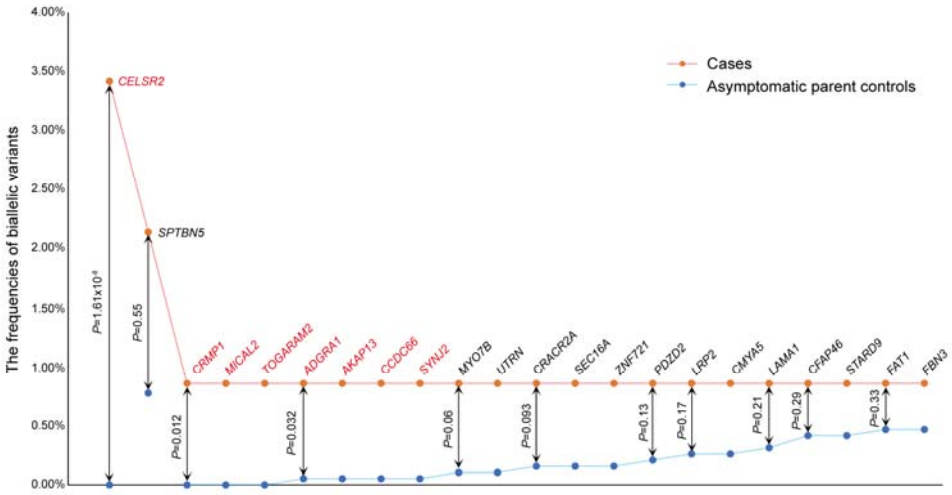


Fig. 2 | The frequencies of biallelic variants in LGS comparing with that in the asymptomatic parent controls. The orange dots indicate the frequencies of genes with biallelic variants in cases. The blue dots indicate the frequencies of genes with biallelic variants in asymptomatic parent controls. The genes in red font indicated that the frequencies of biallelic variant in cases were significantly than that in the asymptomatic parent controls. Data were analyzed by two-tailed Fisher's exact test.

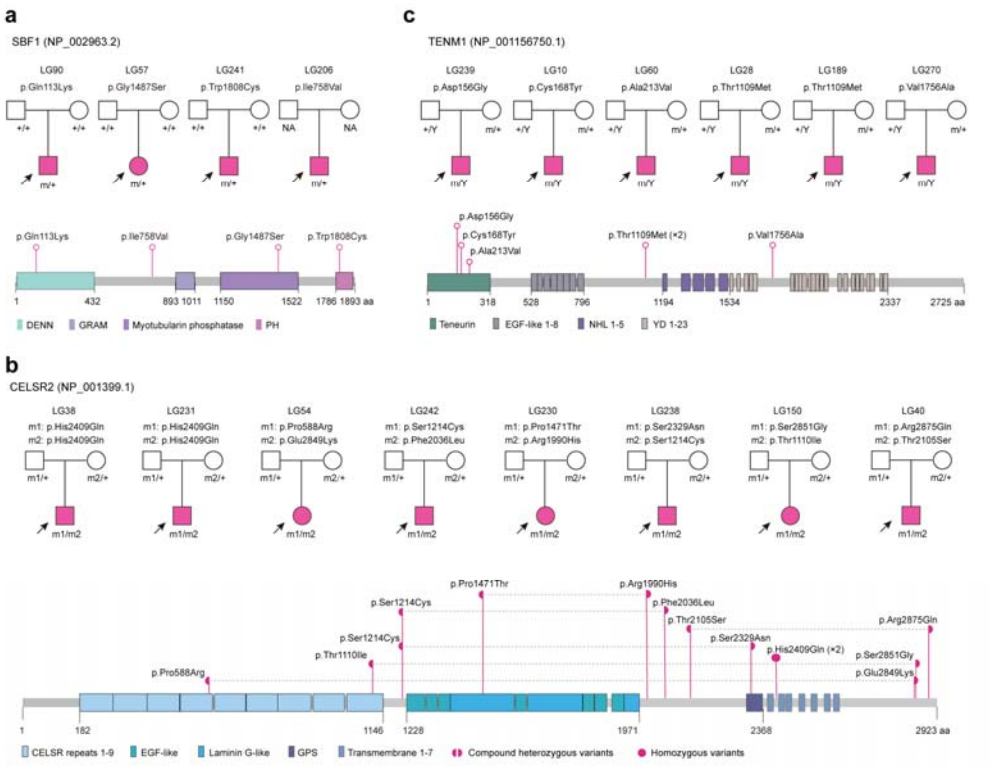


Fig. 3 | *SBF1*, *CELSR2*, and *TENM1* variants identified in the patients with LGS. Pedigrees of the unrelated cases with *SBF1* (**a**), *CELSR2* (**b**), and *TENM1* (**c**) variants (top) and the localization of the variants in the protein domains (bottom). The gray dotted lines in (**b**) indicate a pair of compound heterozygous variants of *CELSR2*. Abbreviations: CELSR, cadherin EGF LAG seven-pass G-type receptor; DENN, differentially expressed in normal and neoplastic cells; EGF, epidermal growth factor; GPS, G-protein-coupled receptor proteolytic site; GRAM, Glucosyltransferases, Rab-like GTPase activators and Myotubularins; NHL, ncl-1, HT2A and lin-41; PH, pleckstrin homology; YD, Tyr and Asp dipeptide.

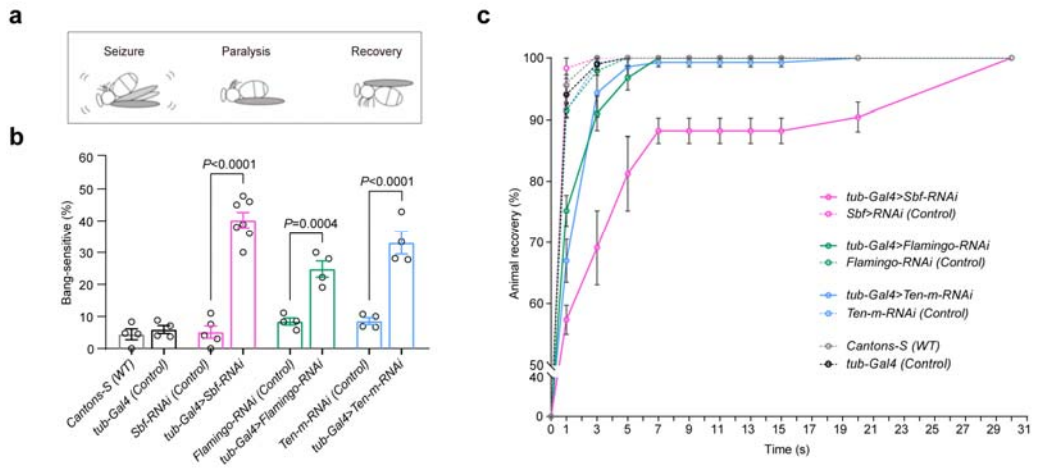


Fig. 4 | Seizure-like behaviors in *Sbf*, *Flamingo*, and *Ten-m* knockdown *Drosophila*. a, Three stages of the seizure-like behaviors in the (bang-sensitivity) BS paralysis test, including seizure (manifesting as vibrating wings), paralysis, and recovery, were observed in *Sbf*, *Flamingo*, and *Ten-m* knockdown flies. **b,** Seizure-like behaviors occurred at a higher rate in the knockdown flies (*tub-Gal4>Sbf-/Flamingo-/Ten-m-RNAi*) than in WT flies (*Canton-S*), *tub-Gal4* (*Control*), and the respective RNAi control flies. **c,** The recovery time from seizure in the knockdown flies (*tub-Gal4>Sbf-/Flamingo-/Ten-m-RNAi*) and in the control groups. Data were collected from 4-7 groups, 4-6 flies in each group, for BS paralysis and recovery tests. All data are represented as mean ± s.e.m. One-way ANOVA and Tukey post hoc test were used for multiple comparisons.

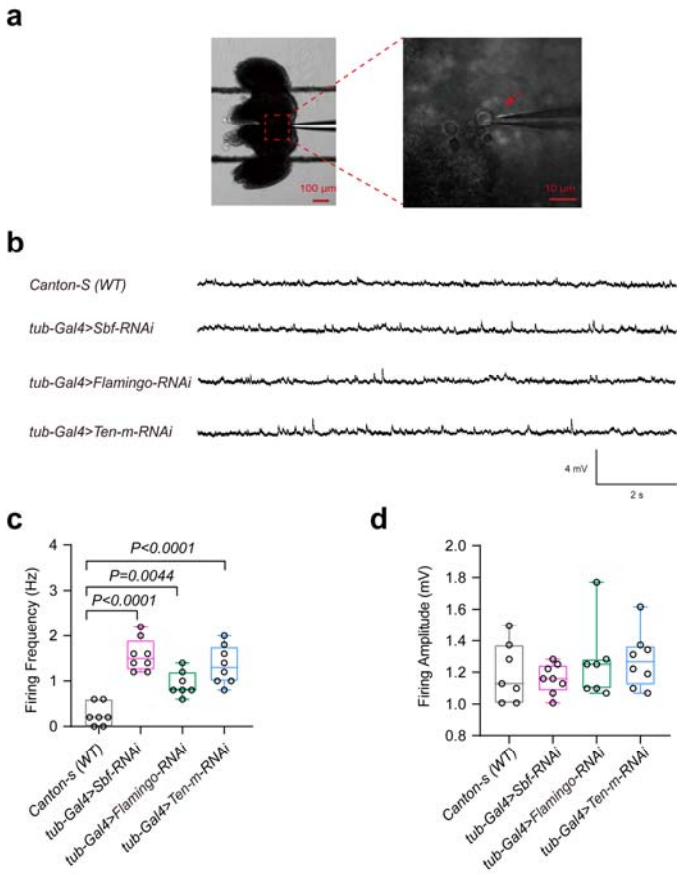


Fig. 5 | Knockdown of *Sbf*, *Flamingo*, and *Ten-m* induces neural excitability in projection neurons of *Drosophila*. a, Whole-cell recording in the fly brain. Left, a schematic diagram showed the insert position of a glass microelectrode in an immobilized isolated brain (scale bar, 100 μm). Right, cell bodies of the squared region in the left panel were indicated at higher magnification (scale bar, 10 μm) and were targeted for whole-cell recording in the living isolated brain preparation. **b,** Representative traces of electrical activity in projection neurons (PNs) showed spontaneous excitatory postsynaptic potentials (sEPSPs) of WT and *tub-Gal4>Sbf-/Flamingo-/Ten-m-RNAi* flies. **c,** Frequencies of sEPSPs in PNs of WT and the knockdown flies. **d,** Amplitude of sEPSPs between WT and the knockdown flies. The box plots show all data points from minimum to maximum. Boxes represent

data from the lower (25th percentile) to the upper (75th percentile) quartiles. The box center corresponds to the 50th percentile. A horizontal line indicates the median. Data are represented as mean ± s.e.m. from three independent experiments (*Canton-S*, n = 7; *tub-Gal4>Sbf-RNAi*, n = 8; *tub-Gal4>Flamingo-RNAi*, n = 7, *tub-Gal4>Ten-m-RNAi*, n = 8; one-way ANOVA test).

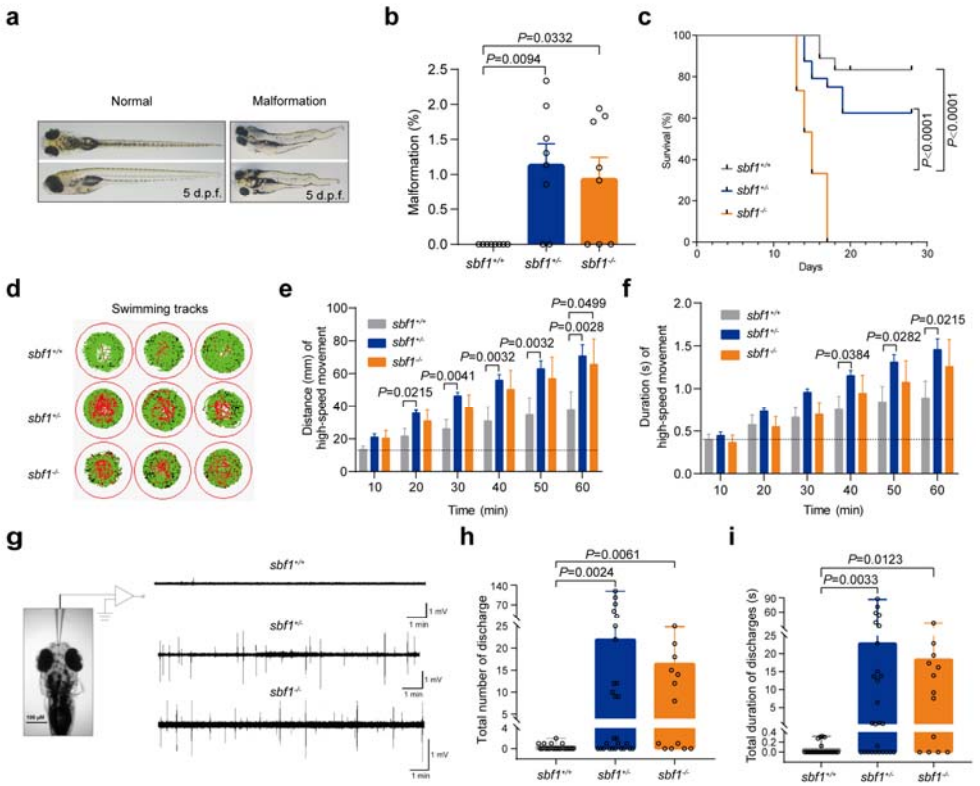


Fig. 6 | Seizure activity in *sbf1* knockout zebrafish. **a,** Malformation was observed in 5 days post-fertilization (d.p.f.) larvae of *sbf1*$^{+/-}$ and *sbf1*$^{-/-}$ but not *sbf1*$^{+/+}$ (WT). **b,** The malformation rates in the *sbf1*$^{+/-}$ and *sbf1*$^{-/-}$ larvae were significantly higher than that in *sbf1*$^{+/+}$ larvae (n = 8, one-way ANOVA test); *sbf1*$^{+/+}$ (grey), *sbf1*$^{+/-}$ (blue), and *sbf1*$^{-/-}$ (orange). **c,** Kaplan–Meier survival analysis for *sbf1*$^{+/+}$ (n = 18), *sbf1*$^{+/-}$ (n = 24), and *sbf1*$^{-/-}$ (n = 15). **d,** Representative swimming tracks of 7 d.p.f. larvae within 60 minutes. Red, green, and black lines represent high- (≥ 20mm/s), intermediate- (8 - 20 mm/s), and slow-speed movements (< 8 mm/s), respectively. **e, f,** Distance (**e**) and duration (**f**) of high-speed movement in each genotype. Data are represented as mean ± s.e.m. from three independent experiments (*sbf1*$^{+/+}$, n = 10; *sbf1*$^{+/-}$, n = 20; *sbf1*$^{-/-}$, n = 10; one-way ANOVA test). **g,** Extracellular field recordings were performed in

immobilized and agar-embedded zebrafish larvae forebrains. Left panel, schematic showing the insert position of a glass microelectrode. Right panel, representative field recording trace of epileptiform discharge from *sbf1*$^{+/+}$ (top), *sbf1*$^{+/-}$ (middle), sbf*1*$^{-/-}$ (bottom) larvae in 7 d.p.f.. **h, i,** Total number of the electrographic seizure-like bursts (including interictal and ictal discharges) (**h**) and duration of the discharges (**i**). Data were from 15 min of recording in 7-11 d.p.f. larvae of *sbf1*$^{+/+}$ (n = 22), *sbf1*$^{+/-}$ (n = 27), and *sbf1*$^{-/-}$ (n = 13); Kruskal-Wallis test.

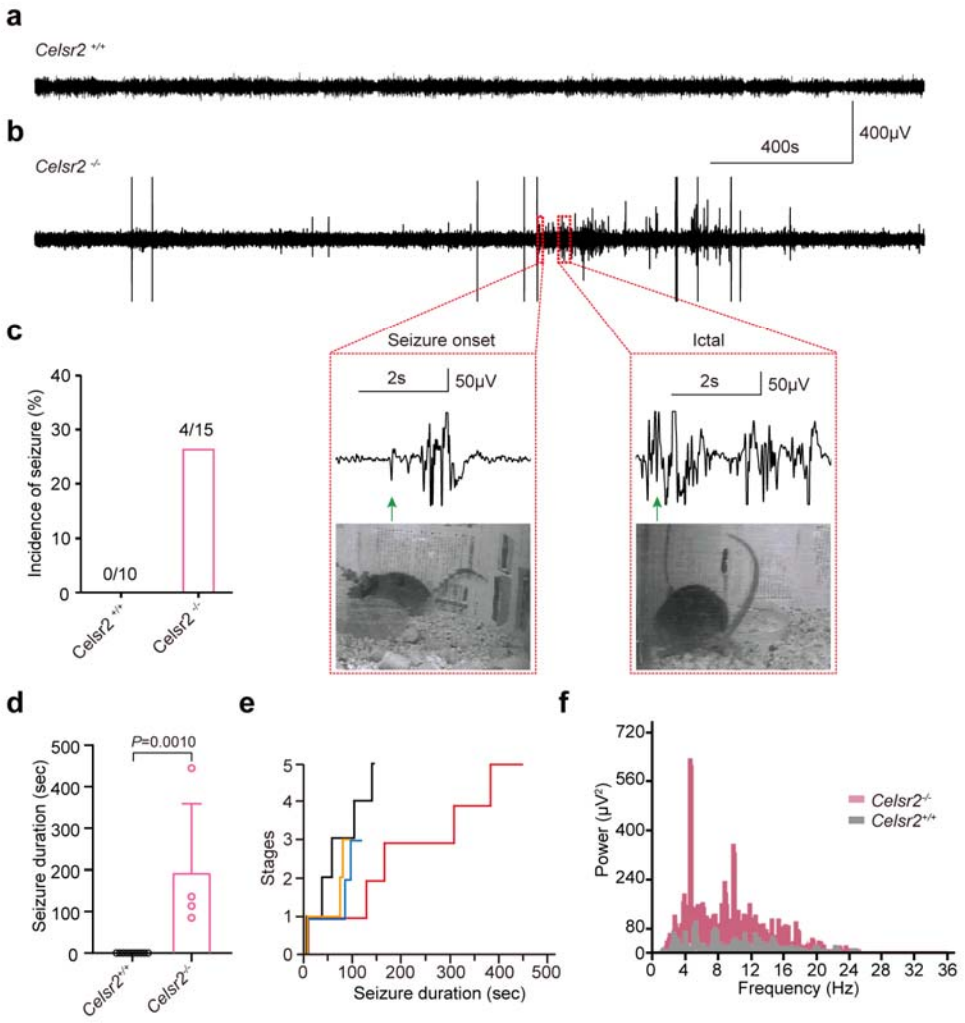
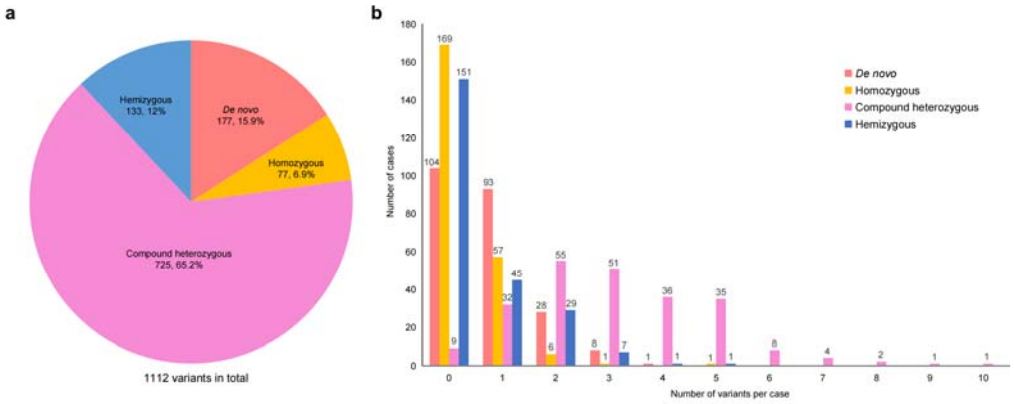
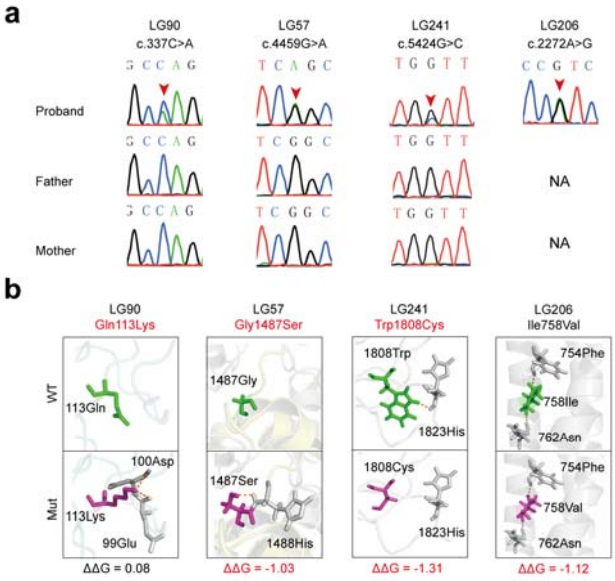


Fig. 7 | Spontaneous seizures and electroencephalography (EEG) recording in *Celsr2* knock-out mice. a, Normal EEGs were detected from the WT mice. **b,** A representative trace shows ictal epileptiform discharges in *Celsr2*$^{-/-}$ mice. The dotted boxes indicated the zoom-in traces of the onset and ictal epileptiform discharges. The green arrows indicate the time point of behavior of the mouse from video recording. **c,** The proportion of mice exhibiting spontaneous seizures in *Celsr2*$^{+/+}$ and *Celsr2*$^{-/-}$ mice. **d,** Seizure duration in *Celsr2*$^{+/+}$ (n = 10) and *Celsr2*$^{-/-}$ (n = 4) mice; Mann Whitney test. **e,** Seizure behaviors were rated with a modified Racine scale in *Celsr2*$^{+/+}$ and *Celsr2*$^{-/-}$

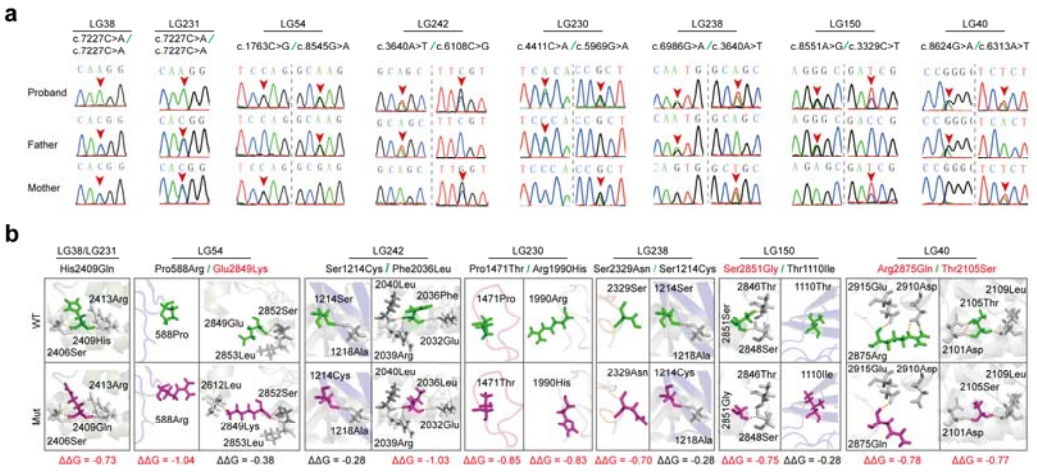
mice. **f,** Power spectrum analysis of EEG signals in *Celsr2*$^{+/+}$ and *Celsr2*$^{-/-}$ mice.



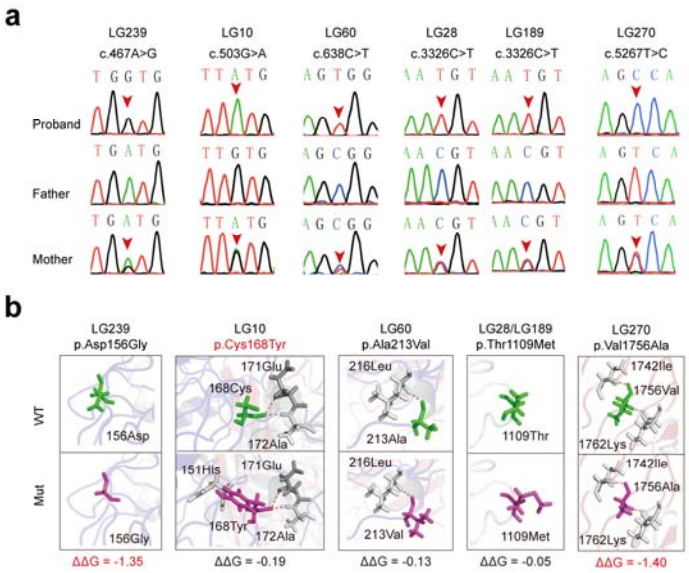
Extended Data Fig. 1 | General information of the variants after filtration by inheritance origin and stratified MAF. a, The pie chart of the inheritance origin of the 1112 qualified variants. **b,** The distribution of the number of *de novo*, homozygous, compound heterozygous, and hemizygous variants per case.



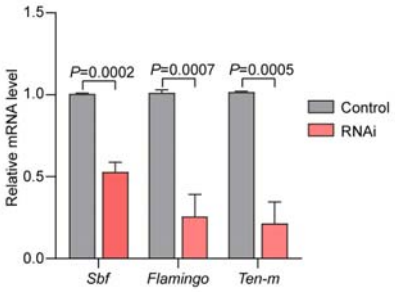
Extended Data Fig. 2 | Genetic data of cases with *SBF1* variants. a, Sanger sequencing verification of four cases with *SBF1* variants. **b,** Modeling of *SBF1* missense variants indicated abnormal hydrogen bonds. Wild-type and mutant residues are colored in green and light magenta, respectively. Hydrogen bonds are colored in orange. The variant Lys113 rebuilds two hydrogen bonds with Asp100 and Glu99. The Ser1487 rebuilds one with His1488. The variant 1808Cys destroyed one hydrogen bond with His1823. Variants that altered the stability of the global conformation of the SBF1 protein were indicated by the value of free energy stability changes (ΔΔG, Kcal/mol) using the online tools (http://gpcr.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi). The absolute value of the ΔΔG > 0.5 indicates a large decrease of stability. The variants in red font (the top) indicated that the variant changed hydrogen bonds with surrounding amino acid residues. The ΔΔG in red font (the bottom) indicated that the variant significantly changed the protein stability.



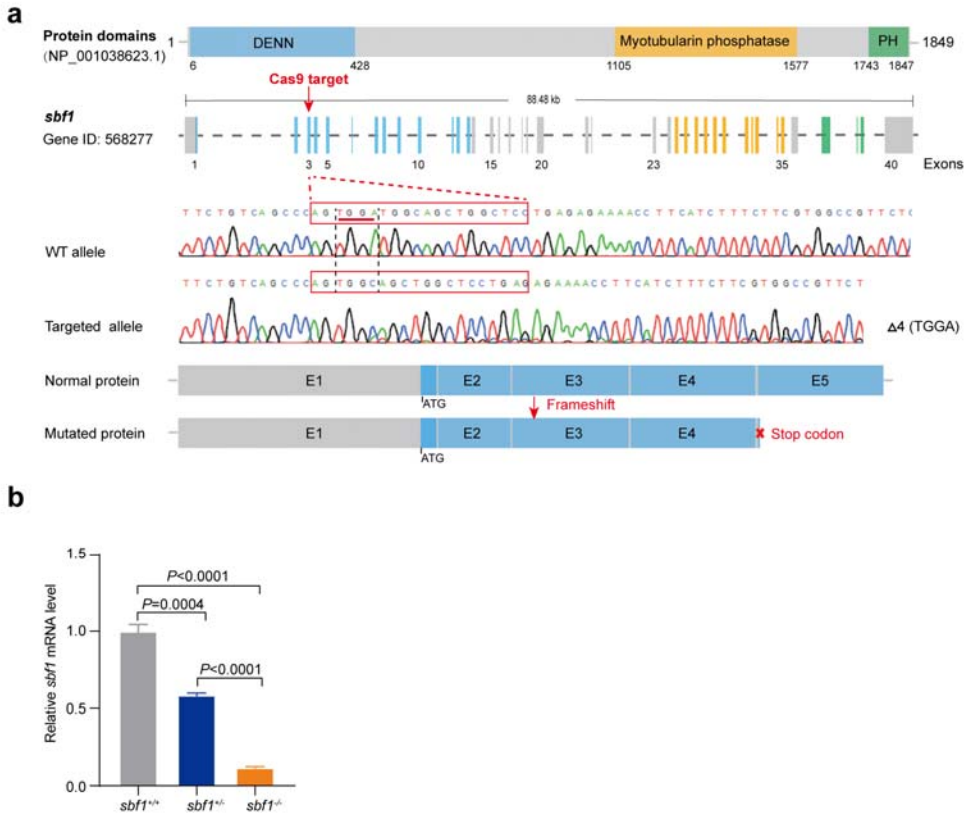
Extended Data Fig. 3 | Genetic data of cases with *CELSR2* variants. a, Pedigrees and Sanger sequencing verification of the eight cases with biallelic *CELSR2* variants. **b,** Modeling of CELSR2 missense variants indicated abnormal hydrogen bonds. Wild-type and mutant residues are colored in green and light magenta, respectively. Hydrogen bonds are colored in orange. Variants that destabilized the global conformation of the CELSR2 protein were indicated by the value of free energy stability changes (ΔΔG, Kcal/mol) using the online tools (http://gpcr.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi). The absolute value of the ΔΔG > 0.5 indicates a large decrease of stability. The variants in red font (the top) indicated that the variant changed hydrogen bonds with surrounding amino acid residues. The ΔΔG in red font (the bottom) indicated that the variant significantly changed the protein stability.



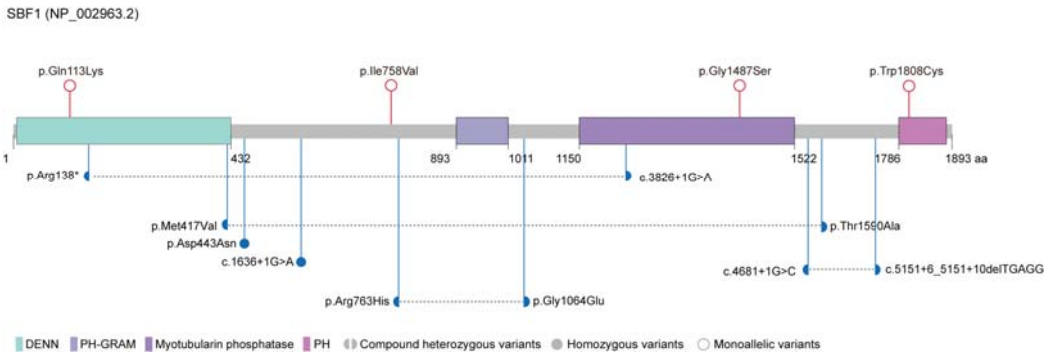
Extended Data Fig. 4 | Genetic data of cases with *TENM1* variants. a, Pedigrees and Sanger sequencing verification of the six cases with *TENM1* variants. **b,** Modeling of TENM1 missense variants predicted abnormal hydrogen bonds. Wild-type and mutant residues are colored in green and light magenta, respectively. Hydrogen bonds are colored in orange. The variant Tyr168 rebuilds a hydrogen bond with His151. Variants that destabilized the global conformation of the TENM1 protein were indicated by the value of free energy stability changes (ΔΔG, Kcal/mol) using the online tools (http://gpcr.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi). The absolute value of the ΔΔG > 0.5 indicates a large decrease of stability. The variants in red font (the top) indicated that the variant changed hydrogen bonds with surrounding amino acid residues. The ΔΔG in red font (the bottom) indicated that the variant significantly changed the protein stability.



Extended Data Fig. 5 | The knockdown efficiency of *Sbf*, *Flamingo*, and *Ten-m* in files. The relative mRNA level of *Sbf*, *Flamingo*, and *Ten-m* in knockdown flies was detected by RT-qPCR. The mRNA levels of wild type flies were normalized as 1.0. The *Gapdh* is internal control. Data were analyzed by Student's t test.



Extended Data Fig. 6 | Establishment of a *sbf1* KO zebrafish model using CRISPR/Cas9. a, The CRISPR/Cas9 gRNA targeted the exon 3 (E3) of *sbf1*, resulting in 4-nt (TGGA) deletion in E3. The frameshift mutant led to a stop codon occurring in exon 5 (E5). **b,** The relative mRNA expression of *sbf1* was detected by RT-qPCR in the $sbf1^{+/+}$, $sbf1^{+/-}$, and $sbf1^{-/-}$ fish. The mRNA levels of $sbf1^{+/+}$ were normalized as 1.0. The *gapdh* is internal control. Data were analyzed by Student's t test.



Extended Data Fig. 7 | The localization of *SBF1* variants identified in this study and in Charcot-Marie-Tooth disease. Variants identified in this study were shown in red (the top). Variants associated with Charcot-Marie-Tooth disease were shown in blue (the bottom). The gray dotted lines indicate the pairs of compound heterozygous variants of *SBF1*.