The American Journal of Human Genetics, Volume *106*

Supplemental Data

De Novo Truncating Variants in the Last Exon of *SEMA6B* Cause Progressive Myoclonic Epilepsy

Kohei Hamanaka, Eri Imagawa, Eriko Koshimizu, Satoko Miyatake, Jun Tohyama, Takanori Yamagata, Akihiko Miyauchi, Nina Ekhilevitch, Fumio Nakamura, Takeshi Kawashima, Yoshio Goshima, Ahmad Rithauddin Mohamed, Gaik-Siew Ch'ng, Atsushi Fujita, Yoshiteru Azuma, Ken Yasuda, Shintaro Imamura, Mitsuko Nakashima, Hirotomo Saitsu, Satomi Mitsuhashi, Takeshi Mizuguchi, Atsushi Takata, Noriko Miyake, and Naomichi Matsumoto

Case Reports

Individual 1

Individual 1 (II-1 in family 1, Figure 1E) is a 22-year-old Japanese male born to non-consanguineous parents. He has a healthy younger sister. He was born after 40 weeks of gestation via cesarean section due to breech presentation. His birth weight, length and head circumference were 3,676 g (+1.6 SD), 52.8 cm (+1.8 SD) and 38.0 cm (+3.6 SD), respectively. His developmental milestones were mildly delayed: head control at the age of 6 months, rolling over at 12 months, walking without support at 17 months and speaking meaningful words at 24–36 months. At 6 years, he developed generalized tonic-clonic seizures (GTCS) without fever, which responded to valproic acid and clonazepam. At this time, his interictal electroencephalogram (EEG) displayed abnormal discharges on the right hemisphere. At 9 years of age, he experienced seizures with forward head nodding and absence seizures 2–3 times a day and he visited our hospital for the first time. At the first visit, his EEG showed bursts of diffuse irregular spikes and waves, which corresponded to atypical absence seizures. At the age of 11 years, he developed daily repetitive atonic seizures thought to be negative myoclonus. At the same time, he received pulse steroid therapy and administration of prednisolone and piracetam, which were partially effective for seizures. At 13 years, he became unable to walk because of progressive motor regression. At 14 years, Parkinson's disease-like symptoms occurred with shaking hands, intention tremor, myoclonus, spasticity and muscle rigidity in the lower limbs. L-DOPA was given for these symptoms but was discontinued because of aggravation of muscle rigidity in the lower limbs. The deep tendon reflex in the lower limbs was increased and that in the upper limbs was mildly increased. He showed rather increased passive muscle tone and mild rigospasticity in all limbs, but the muscle weakness and extensibility were apparently normal. Some reflexes were abnormal at the age of 14 years: Babinski sign; negative, Chaddock sign; negative, Rossolimo sign; positive, Mendel-Bechterew sign; positive, Clonus; negative, and Wartenberg sign;

negative. At the age of 14 years, his height was 155.0 cm (-0.2 SD) and weight was 58.0 kg (+0.4 SD). Brain magnetic resonance imaging (MRI) was normal at this age. At 14 years, his EEG revealed diffuse spikes and slow waves especially in frontal, central and parietal regions. Somatosensory evoked potential (SEP) findings at 14 years had prolonged N20 latency and high amplitude of P24-N33 compared with those of individuals of the same age. Visual evoked potentials and auditory brain-stem response tests were normal. Severe intellectual disability was recognized, with his intelligence quotient (IQ) at 17 years being 25 based on the Tanaka-Binet test (the standardized Japanese version of the Stanford-Binet test). He showed significant dysarthria regression and was able to say his name but was not able to say his age or school name. He suffered from muscle pain in the lower limbs since the age of 20 years. Currently, 22 years old, his height is 174.0 cm (+0.6 SD), his weight is 67.0 kg (+0.4 SD) and his OFC is 59.5 cm (+4.7 SD). He cannot crawl or walk without support but can roll over and walk with assistance. He is able to communicate with others in a few words. Myoclonus in upper and lower limbs appeared when he feels uncomfortable. His myoclonus of the limbs often transits to GTCS. The seizures and involuntary movements are intractable and do not responded to clonazepam, valproic acid, carbamazepine, primidone, levetiracetam, piracetam, arotinolol, steroid or L-DOPA. Other laboratory examinations (complete blood count, serum chemistry, blood gases, ammonia, amino acid profiles and urine organic acids), echocardiography and abdominal ultrasonography were all normal. G-band and sub-telomere FISH analyses were also normal. His father and younger sister are completely healthy, but his mother has rheumatoid arthritis.

Individual 2

Individual 2 (II-1 in family 2, Figure 1E) is a 28-year-old female born after 38 gestational weeks as the first child to non-consanguineous Japanese parents. She has a healthy younger sister. Her birth weight was 3,094 g (+0.7 SD) and her length was 49.0

cm (+0.1 SD). Her Apgar score at 5 minutes was 9. Her developmental milestones were apparently normal, with head control at the age of 4 months, rolling over at 5 months, sitting alone at 8 months, speaking meaningful words at 1 year and walking without support at 2 years and 4 months. At 11 months, she had her first episode of generalized tonic-clonic seizures (GTCS) with fever. The same febrile GTCS occurred at 1 year and 5 months, and phenobarbital treatment was started at this time. Then she experienced non-febrile GTCS three times (at 2 years, 2 years and 9 months, and 3 years and 9 months). She received phenytoin at 2 years and 9 months, which was ineffective at treating the seizures. At 3 years and 4 months, her interictal EEG showed diffuse slow waves of high amplitude at 2~3 Hz and with spike-and-wave complexes in the bilateral frontal region. At 5 years, she showed a loss of consciousness with abnormal eye movement more than 10 times a day, which responded to valproic acid and ethosuximide. The loss of consciousness occurred recurrently at 7 years. EEG at 8 years and 3 months showed diffuse slow waves of high amplitude at 2~3 Hz and fast waves in bilateral brain regions during sleep. At 9 years and 1 month, she had diffuse theta waves at 4~5 Hz and spike-and-wave bursts at 2~3 Hz in awake EEG. At 10 years, she developed atonic seizures and complex partial seizures, such as suddenly stopping movement and body rigidity. Since this time she has had difficulty walking with falling. Frequent myoclonus in upper and lower limbs and intention tremor were observed, especially when she was awakening, moving and feeling nervous. Clonazepam and L-DOPA were given, and the symptoms were partly recovered. At 12 years, she was referred to our hospital, with suspected progressive myoclonus epilepsy. Her deep tendon reflex in both upper and lower limbs increased. She also had motor regression and spasticity in lower limbs. Her intelligence was severely impaired (IQ=25 at 12 years, based on the Tanaka-Binet test), although no intellectual regression was observed. Interictal EEG at 12 years showed an abnormal pattern with multifocal spikes in the left parietal region and bilateral frontal regions. Auditory brainstem response, skin biopsy and funduscopy tests were all normal. Her SEP

test showed giant spikes. At 13 years, she started to use a wheelchair because of difficulty walking. At 13 years and 11 months, her sleep EEG showed multispikes in the left occipital region. At 16 years, she was suspected to have systemic lupus erythematosus (SLE) with butterfly patch, loss of hair, photodermatosis and oral ulcers. Her serum tests at this time detected positive anti-nuclear and anti-double stranded DNA antibodies. The SLE was thought to be drug-induced lupus erythematosus, because discontinuation of ethosuximide was followed by a disappearance of symptoms. At 17 years, she showed non-febrile GTCS, which has never recurred. The frequency of myoclonus occurrence has gradually increased in recent years. The MRI findings at 23 years were normal except for mild cerebellar atrophy. At 23 years and 9 months, her resting awake EEG showed slow waves at baseline, but no apparent spikes. Currently, at 28 years old, her height, weight and head circumference are 144.5 cm (-2.6 SD), 57.0 kg (+0.5SD) and 52.5 cm (-2.0 SD), respectively. She can speak two-word sentences with unclear pronunciation and understand simple conversation. She is able to control her wheelchair alone and walk holding onto something. Mild hypotonia was present, although without muscle weakness. Other laboratory examinations, including complete blood count, serum chemistry, echocardiography and abdominal ultrasonography were normal. Her mother and younger sister are completely healthy, but her father has asthma.

Individual 3

Individual 3 (II-4 in family 3, Figure 1E) is a 14-year-old boy born after 40 weeks of gestation to non-consanguineous Israeli parents. He has four healthy brothers and five healthy sisters and his mother had a spontaneous abortion. His birth weight was 3,300 g (+1.5 SD). At birth, he had cardiac problems with bicuspid aortic valve and tricuspid regurgitation. No characteristic facial features were noticed. He showed almost normal developmental milestones, with head control at the age of 3 months, speaking meaningful words at 1 year, and walking without support at 2 years. At 1 year, he showed abnormal

background activity in awake EEG. At 2 years, he developed absence seizure which then occurred every day. At the same age he experienced two episodes of loss of consciousness, and two episodes of tongue biting. Brain MRIs at 2 years revealed a small vermis. At 4 years, he started clonazepam and lamotrigine to control seizures, which was effective. No clinical seizures were seen from 5 years of age. He had regression of motor and verbal skills, with no speech from 2 years of age and was unable to walk without support at 13 years of age. He showed severe intellectual disability, although he received no IQ or DQ tests. At 13 years, he had apparent cerebellar dysfunction with resting and intention tremor and ataxia, unstable gait and shaking and falling limbs. At 13 years, his EEG showed slow abnormal sleep features with paucity of sleep spindles. Currently, at 14 years, his head circumference is 52.0 cm (-2.5 SD). He cannot walk and is wheelchair-dependent. He is unable to communicate verbally. His chromosomal analysis was normal. His parents and all his siblings are completely healthy.

Individual 4

Individual 4 is an 11-year-old girl with global developmental delay and drop attacks. She had global developmental delay prior to two years old, but subsequently caught up on her milestones. At 4 years, she started having multiple and frequent episodes of drop attacks, and subsequently she regressed. She developed aphasia, had difficulty in chewing, and was unable to ambulate well. Her drop attacks were improved with clobazem and sulthiame, and she gradually developed her milestones. At 11 years, she developed unbalanced gait with ataxia, intention tremor, and urinary incontinence. In the last examinations at 11 years old, weight and head circumference were 40.9 kg (25–50th percentile) and 49.5 cm (2nd percentile), respectively. EEG showed epileptic discharges and focal bifrontal epileptiform discharges at 4 years, frequent frontocentral discharges during the awake state at 5 years, frequent intermittent slow spikes in the right posterior

region and abnormal background at 11 years and multiple episodes of drop attacks with corresponding spikes and polyspike slow discharges during sleep at 11 years.

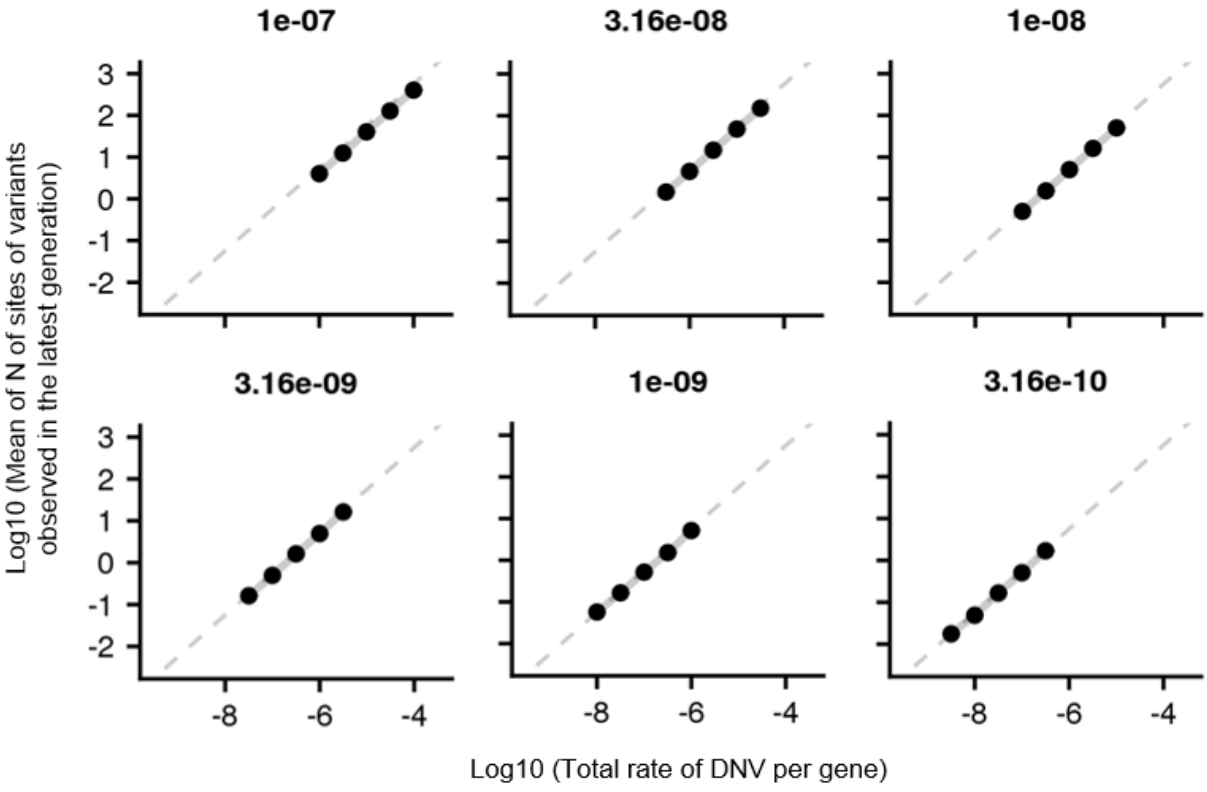


Figure S1. Correlation of per-gene DNV rate and per-gene number of variants.

Correlations between per-gene DNV rate and mean per-gene number of variants observed in 33,370 individuals in the last generation after mutation and selection events through 1,000,000 generations in genetic drift simulations. Variable parameters were analyzed: per-gene number of possible variant sites=10, 32, 100, 320, or 1000 (each dot from the left to the right in the figure); per-site DNV rate=3.16e-10, 1.0e-9, 3.16e-9, 1.0e-8, 3.2e-8, or 1.0e-7 (shown above). Selection coefficient and dominance coefficient were set as 0 where variants were under no natural selection. We calculated the mean of the per-gene number of observed variants across at least 300 simulations of each parameter setting. Note that the per-gene DNV rate and mean per-gene number of variants were on the same line and correlated regardless of variable parameter settings. Dashed line indicates an arbitrary reference line common in the three panels.

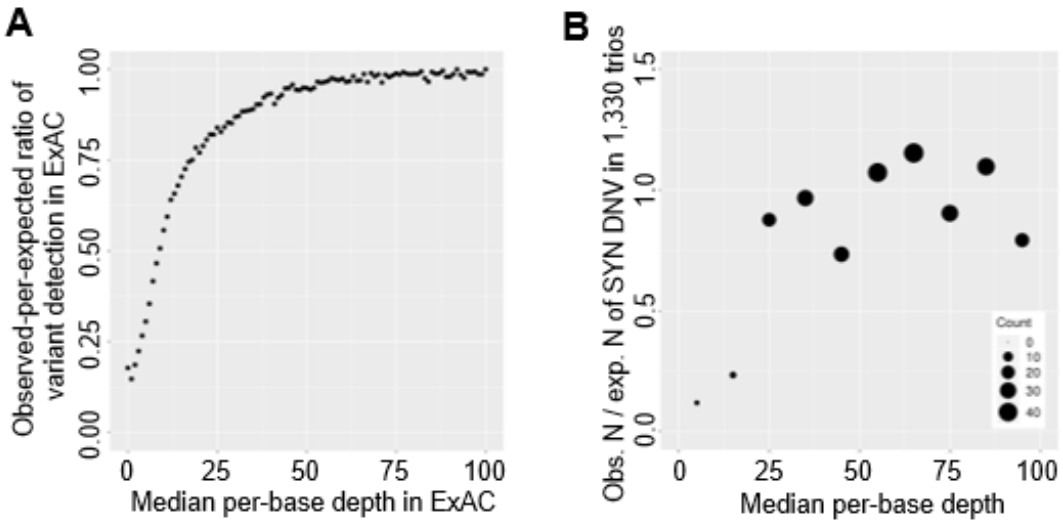


Figure S2. Sequencing depth affects variant detection.

(A) Correlation between median per-base depth and observed-per-expected ratio of variant detection in ExAC. The ratio was calculated as a ratio of the observed number of synonymous variants to the expected rate of synonymous DNVs at sites of each depth and normalized to the ratio at sites with depth >100. The bin of median per-base depth in ExAC at the X axis: 2. (B) Correlation between median per-base depth and observed-per-expected ratio of DNV detection in 1,330 trios. The ratio was calculated as a ratio of the observed number of synonymous DNVs to the expected rate of synonymous DNVs at sites of each depth and normalized to the ratio at sites with depth >100. Median per-base depth was calculated separately for each capture kit. The size of dots is correlated with the observed number of synonymous DNVs at each bin of depth as shown at lower right. The bin of median per-base depth at the X axis: 10.

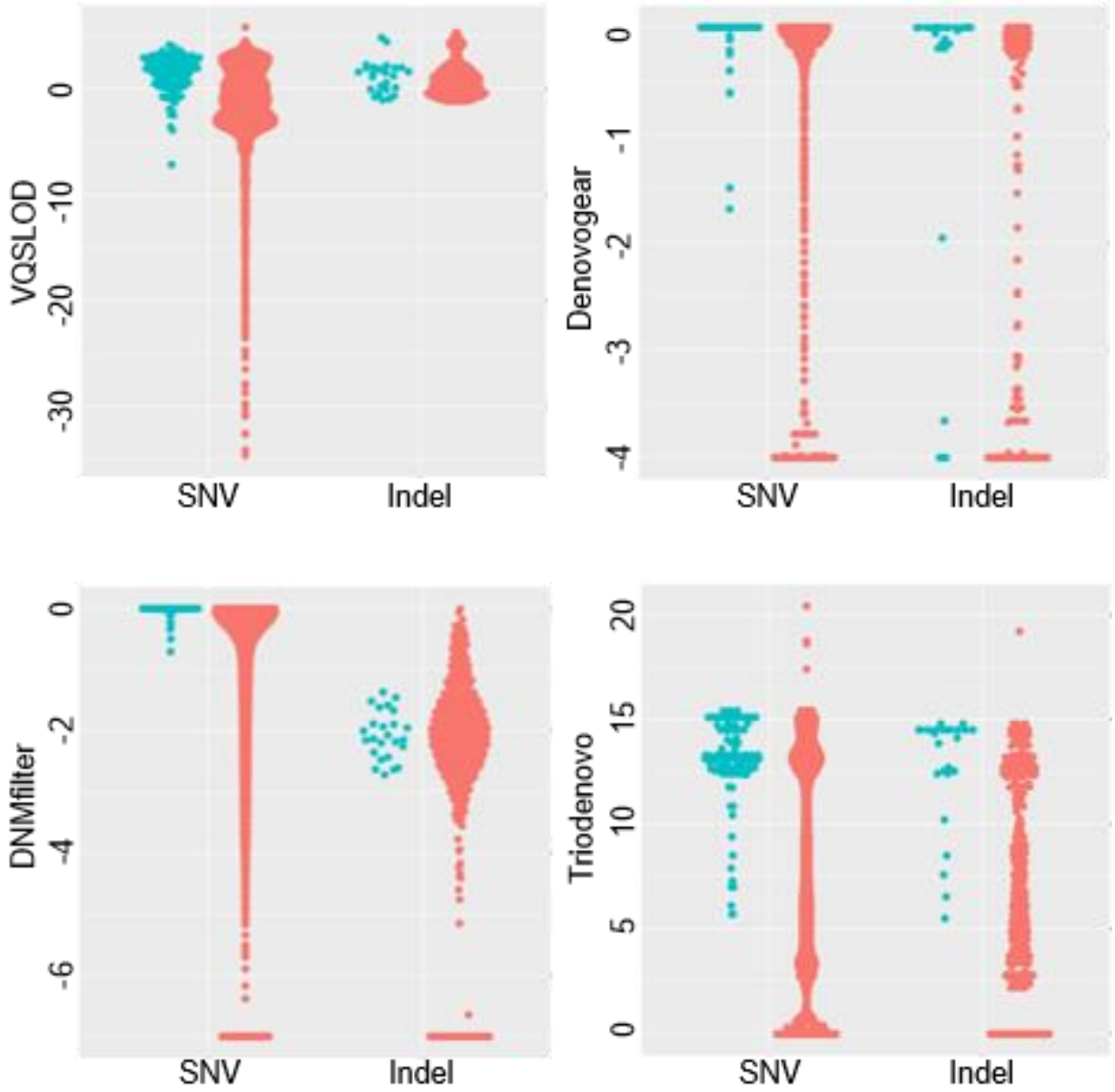


Figure S3. Optimization of thresholds for DNV detection.

Bee swarm plots of DNV quality scores from four systems. Candidate DNV calls from 4,592 SNVs and 502 indels, including 115 SNV and 25 indel TP (true positive) DNVs were analyzed. Green and red dots indicate TP DNVs and the other variants, respectively. We set thresholds for each score so that no TP DNV was missed.

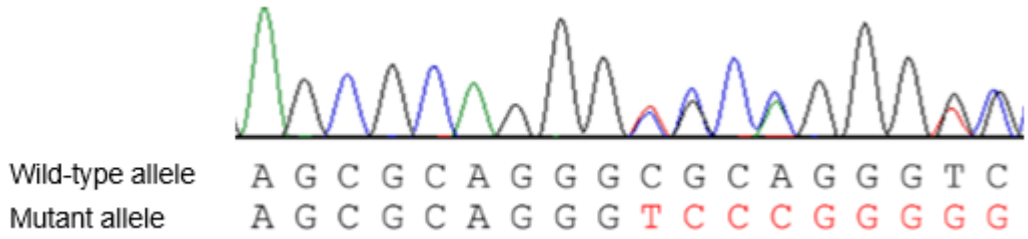


Figure S4. ***SEMA6B*** **transcript with a truncating DNV in the last exon was not subject to NMD**.

A representative Sanger sequencing electropherogram of an *SEMA6B* transcript RT-PCR amplicon from a lymphoblastoid cell line from Individual 2. Sequence differences between wild-type and mutant alleles are shown in red.

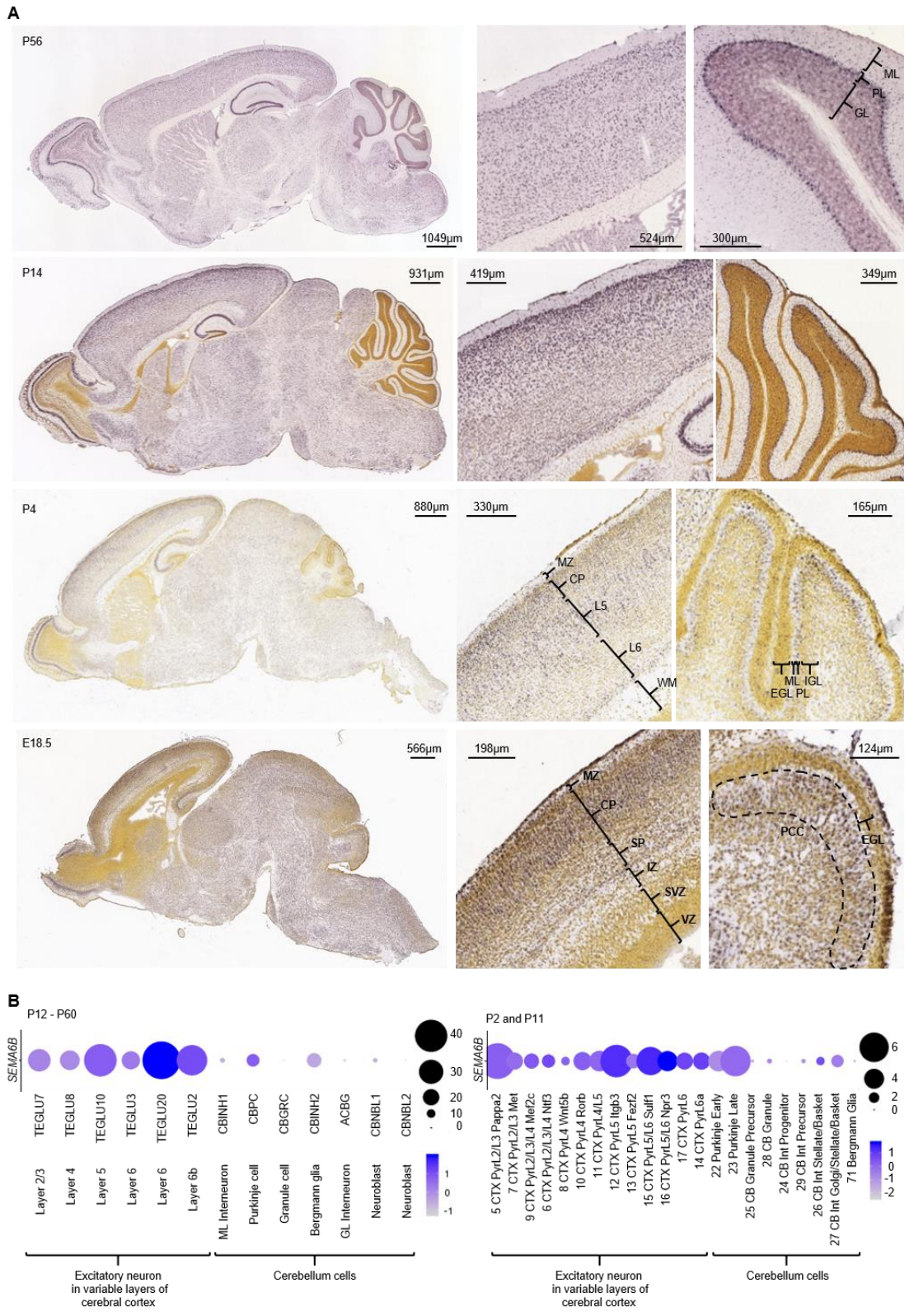
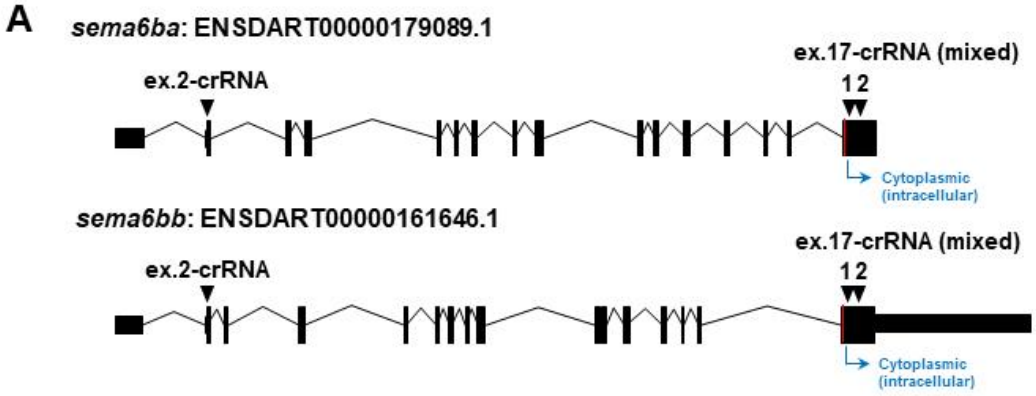


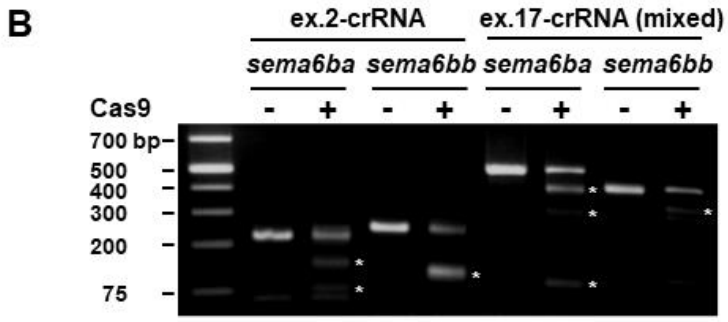
Figure S5. Expression pattern of *Sema6b* in mouse brain.

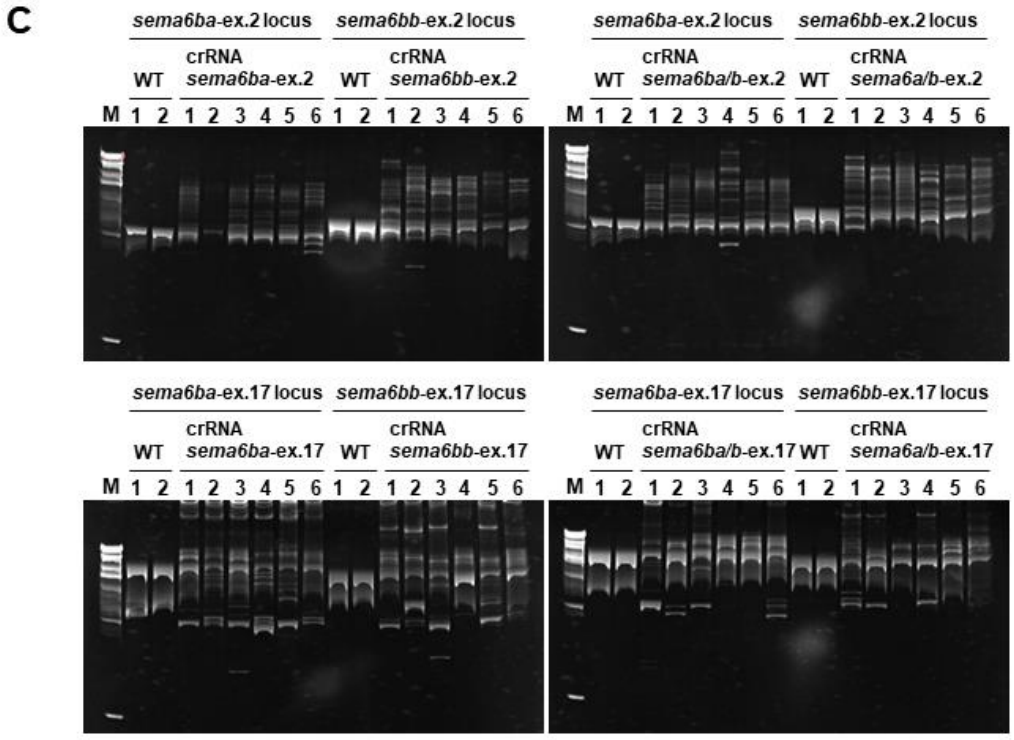
(A) ISH of *Sema6b* in P56 (top), P14 (second from top), P4 (third from top), and E18.5 (bottom) mouse brain from the Allen Brain Atlas. Overall view (left) and a magnified view of the cortex (middle) and cerebellum (left) are shown. Note that the outer part of the PCC (circled with a dotted black circle) was stained. MZ: marginal zone; CP: cortical plate; L5: layer 5; L6: layer 6; SP: subplate; IZ: intermediate zone; SVZ: subventricular zone; VZ: ventricular zone; WM: white matter; EGL: external germinal layer; ML: molecular layer; PL: Purkinje cell layer; IGL: internal granular layer; GL: granular layer; PCC: Purkinje cell cluster. (B) *Sema6b* expression by scRNA-seq of P12-P60 mouse brain[12] (left) and P2 and P11 mouse brain[13] (right). Each dot represents each cell cluster identified by the scRNA-seq analysis. The size and color of dots represent the percentage of cells expressing *Sema6b*, and the mean scaled expression level (see supplemental methods) in each cell cluster is shown in the right. Below each dot, the original name is described of each cell cluster in the publications of the datasets with a brief comment.[12,13] More detailed descriptions of each cluster are available (Table S4 and S5).

Figure S6. Expression pattern of *SEMA6B* in human brain.

(A) Heatmap of *SEMA6B* expression in bulk RNA-seq of variable regions of human brain at variable stages of development. The developmental stages and brain regions of samples are shown as horizontal and vertical axes, respectively. Red color in each box indicates strength (RPKM: reads per kilobase per million) of *SEMA6B* expression in samples of each developmental stage and brain region as shown in the right color bar while blue color in several boxes indicates no sample was available in the developmental stage and brain region. Pcw: post-conceptional weeks; mos: months; yrs: years. (B) Dotplot of *SEMA6B* expression in single-nucleus RNA-seq (snRNA-seq) of variable cell types of adult human brain. The size and color of dots represent the percentage of cells expressing *SEMA6B* and the mean of expression (CPM: count per million) in cell clusters identified by snRNA-seq. Below each dot, the original name of the cell cluster describing a cell type, layers where the cell cluster localized, and marker genes. Exc: glutamatergic neuron; Inh: GABAergic neuron; Astro: astrocyte; endo: endothelial cell; micro: microgria; oligo: oligodendrocyte; OPC: oligodendrocyte precursor cells; peri: pericyte; VLMC: vascular and leptomeningeal cell.

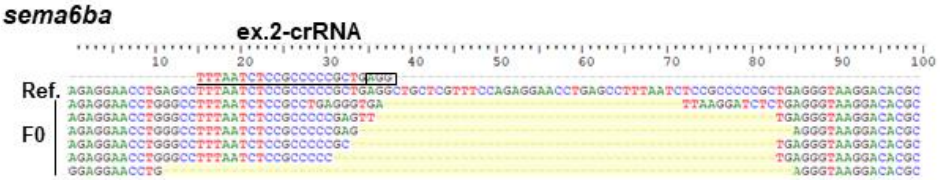

A
sema6ba: ENSDART00000179089.1
ex.2-crRNA
ex.17-crRNA (mixed)
1 2
Cytoplasmic
(intracellular)
sema6bb: ENSDART00000161646.1
ex.2-crRNA
ex.17-crRNA (mixed)
1 2
Cytoplasmic
(intracellular)


B
ex.2-crRNA
ex.17-crRNA (mixed)
sema6ba sema6bb sema6ba sema6bb
Cas9
- + - + - + - +
700 bp
500
400
300
200
75

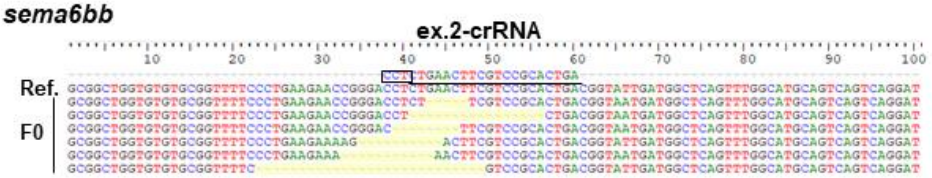

C
sema6ba-ex.2 locus
sema6bb-ex.2 locus
WT
crRNA sema6ba-ex.2
WT
crRNA sema6bb-ex.2
M 1 2 1 2 3 4 5 6 1 2 1 2 3 4 5 6
sema6ba-ex.2 locus
sema6bb-ex.2 locus
WT
crRNA sema6ba/b-ex.2
WT
crRNA sema6a/b-ex.2
M 1 2 1 2 3 4 5 6 1 2 1 2 3 4 5 6
sema6ba-ex.17 locus
sema6bb-ex.17 locus
WT
crRNA sema6ba-ex.17
WT
crRNA sema6bb-ex.17
M 1 2 1 2 3 4 5 6 1 2 1 2 3 4 5 6
sema6ba-ex.17 locus
sema6bb-ex.17 locus
WT
crRNA sema6ba/b-ex.17
WT
crRNA sema6a/b-ex.17
M 1 2 1 2 3 4 5 6 1 2 1 2 3 4 5 6

D

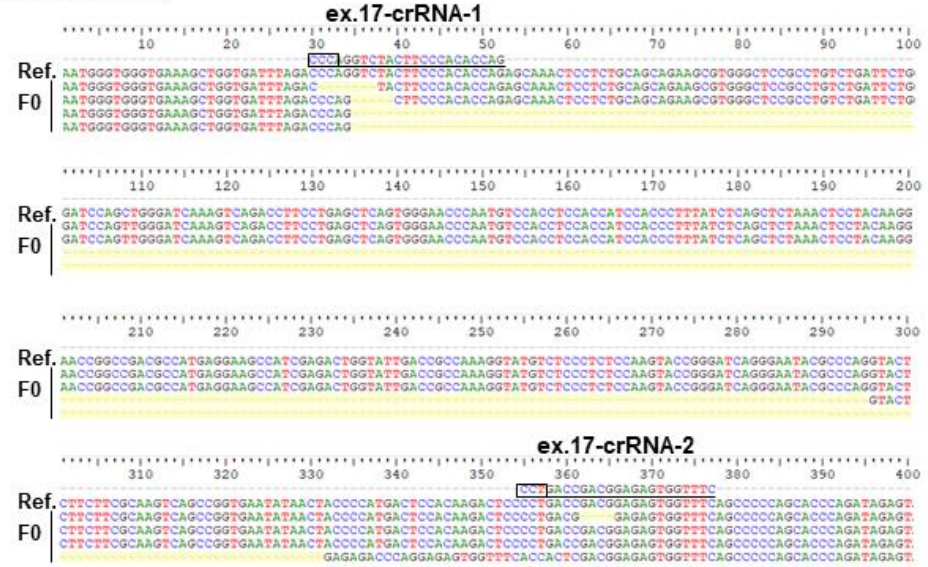
sema6ba



sema6bb



sema6ba (ICD)



sema6bb (ICD)

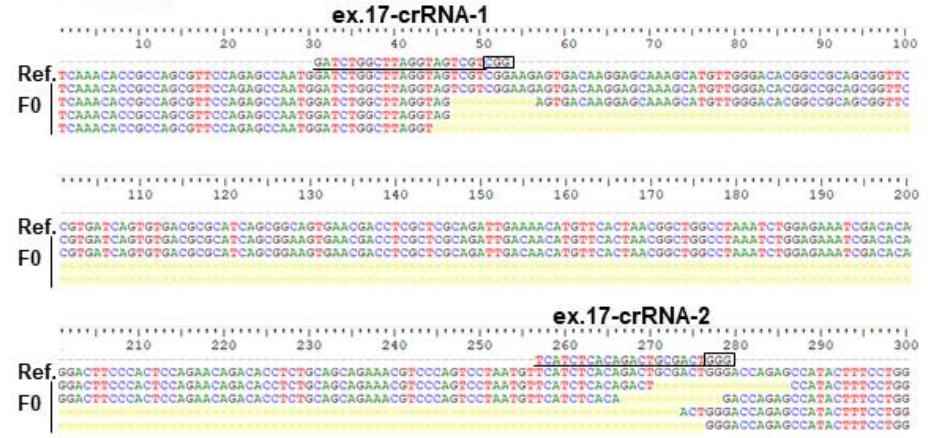
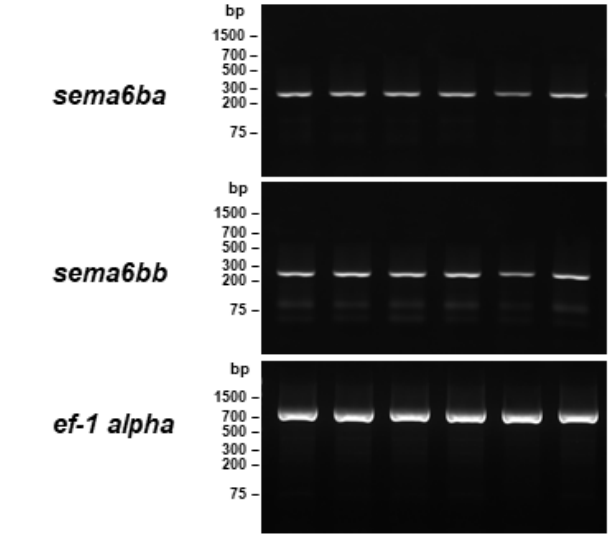


Figure S7. CRISPR/Cas9-induced genome editing efficiencies of zebrafish *sema6ba* and *sema6bb* evaluated in F0 crispants.

(A) Schematic of crRNA targeted exons of zebrafish *sema6ba* and *sema6bb*. Target sites are indicated by arrowheads on exons 2 and 17. Black boxes: exons, red boxes: transmembrane region. blue arrow: starting point of the cytoplasmic region. (B) Evaluation of CRISPR/Cas9-mediated cleavage using the T7E1 assay. The equivalent of five embryos were loaded per each lane. Asterisks indicate the expected positions of cleaved products by the mismatch-sensitive T7E1 enzyme. (C) Heteroduplex mobility assay in the larvae injected with CRISPR/Cas9 at 5 dpf. The *sema6ba*- and *sema6bb*-targeted genomic regions were amplified using the genomic DNA from each individual embryo by PCR with locus-specific primers. Heteroduplex bands and multiple short bands are shown in CRISPR/Cas9 injected embryos (WT: n=2, crispants: n=6). Products from each of single larvae at 5 dpf were loaded per lane. (D) To confirm frameshift mutations in F0 crispants, amplified target loci were cloned into TOPO-vectors and multiple (>24) cloned alleles were sequenced and aligned. Representative sequence alignment shows insertion and/or deletion events at the target site. We sequenced five mixed F0 crispants and confirmed >63% mosaicism for *sema6ba*-ex.2, *sema6bb*-ex.2, *sema6ba*-ex.17 and *sema6bb*-ex.17. Injection of two crRNAs targeting *sema6ba* exon 17 and *sema6bb* exon 17 resulted in more extended deletions. The protospacer adjacent motif (PAM) is shown by a black box and the crRNA sequence is underlined. Ref. indicates the reference sequence. Yellow dashes represent nucleotide deletions.


sema6ba
sema6bb
ef-1 alpha
bp
1500
700
500
300
200
75
DMSO - - - + + +
NMDi14 + + + - - -
crRNA sema6ba/b-ex.2 - + - - + -
crRNA sema6ba/b-ex.17 - - + - - +

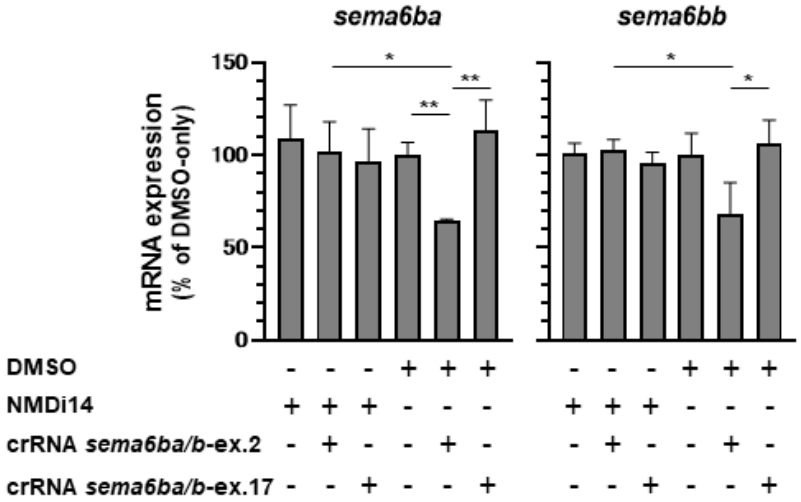

sema6ba
sema6bb
mRNA expression (% of DMSO-only)
150
100
50
0
*
**
**
*
*
DMSO - - - + + + - - - + + +
NMDi14 + + + - - - + + + - - -
crRNA sema6ba/b-ex.2 - + - - + - - + - - + -
crRNA sema6ba/b-ex.17 - - + - - + - - + - - +

Figure S8. Reduction of transcripts in F0 crispants caused by nonsense-mediated mRNA decay (NMD).

RT-PCR analysis of *sema6ba* and *sema6bb* transcripts in F0 crispants treated with NMD inhibitor (NMDi14). Three larvae at 6 dpf were mixed. The *sema6ba/b*-ex.2 crispants treated DMSO (vehicle) alone showed *sema6ba* and *sema6bb* transcripts levels (as a control). Zebrafish *sema6ba* and *sema6bb* transcripts levels were evaluated by three independent RT-PCR (bar graphs at the bottom). Each mRNA expression ratio was normalized by *ef-1 alpha*. Data are represented as the mean ± s.d. Only statistically significant difference between bars are marked with * ($P < 0.05$) and ** ($P < 0.01$) by Student's *t*-test.

Table S1 DNV rate at the whole coding region of each transcript
See a separate Excel file.

Table S2 DNV rate at NMD(-) region of each transcript
See a separate Excel file.

Table S3 Summary of truncating DNVs at NMD(-) regions in 346 DEE trios
See a separate Excel file.

Table S4 Summary of annotations of cell clusters (Zeisel et al., 2018)
See a separate Excel file.

Table S5 Summary of annotations of cell clusters (Rosenberg et al., 2018)
See a separate Excel file.

Table S6 Oligos for crRNA and primers used for target amplification in the zebrafish study.

Oligo name	Sequence (5′ to 3′)
sema6ba ex.2 crRNA	TTTAATCTCCGCCCCCGCTG
sema6bb ex.2 crRNA	TCAGTGCGGACGAAGTTCAG
sema6ba ex.17 crRNA-1	CTGGTGTGGGAAGTAGACCT
sema6ba ex.17 crRNA-2	GAAACCACTCTCCGTCGGTC
sema6bb ex.17 crRNA-1	GATCTGGCTTAGGTAGTCGT
sema6bb ex.17 crRNA-2	TCATCTCACAGACTGCGACT
sema6ba ex.2 HMA-F	ACTTCTCATGTCTTCTTGTGTCTCTGC
sema6ba ex.2 HMA-R	TGCCTAATAATGTCTCTGCTGATGC
sema6bb ex.2 HMA-F	GGCTCTTTTCTGCATGTTTTTCTA
sema6bb ex.2 HMA-R	CATGCTCGCTCTGCATAAGA
sema6ba ex.17 HMA-F	ATGAGCGTATCCCGACACA
sema6ba ex.17 HMA-R	GGAGGTATGCCATGACTGCT
sema6bb ex.17 HMA-F	CTGGTGGTGTCTGCAGTTTC
sema6bb ex.17 HMA-R	TGAAGGTATTTGGAGCTCAGG
sema6ba ex.6 RT-F	AAACGCTTTCAACCCATTGT
sema6ba ex.8-9 RT-R	AGGCTCTCGGAACCACTTG
sema6bb ex.6 RT-F	CCGTTGTGTGCCAATTATACTG
sema6bb ex.8-9 RT-R	CGCTCACAAAGTATGGCTCTC
ef1-alpha RT-F	ACCACCGGCCATCTGATCTACAAA
ef1-alpha RT-R	ACGGATGTCCTTGACAGACACGTT

Table S7. Top five of Pfam protein domains enriched in genes depleted for truncating variants at NMD(-) regions.

Location	Pfam protein domain	Pfam accession ID	Odds ratio	P value	Q value
NMD(+) region	KRAB	PF01352	2.2	2.3E-07	0.0013
	p450	PF00067	4.9	2.8E-06	0.0079
	Filament	PF00038	3.6	3.3E-05	0.062
	SCAN	PF02023	4.2	4.9E-05	0.065
	Keratin_2_head	PF16208	6.8	5.8E-05	0.065
NMD(-) region	p450	PF00067	4.8	3.6E-06	0.014
	PRY	PF13765	4.7	6.3E-05	0.10
	Fibrinogen_C	PF00147	6.4	8.3E-05	0.10
	Filament	PF00038	4.0	2.0E-04	0.19
	zf-C2H2_6	PF13912	1.7	3.4E-04	0.25

Table S8. Top five of gene ontology terms enriched in genes depleted for truncating variants at NMD(-) regions.

GO domain	GO term accession	GO term name	Odds ratio	P value	Q value
biological_process	GO:0006805	xenobiotic metabolic process	2.6	9.2E-05	0.85
	GO:1900017	positive regulation of cytokine production involved in inflammatory response	22.6	0.00057	1
	GO:0009308	amine metabolic process	15.1	0.0013	1
	GO:0003184	pulmonary valve morphogenesis	15.1	0.0013	1
	GO:0061077	chaperone-mediated protein folding	8.1	0.0017	1
cellular_component	GO:0005576	extracellular region	1.4	0.00018	0.21
	GO:0005882	intermediate filament	2.4	0.00071	0.41
	GO:0035861	site of double-strand break	8.8	0.0043	1
	GO:0045111	intermediate filament cytoskeleton	2.8	0.011	1
	GO:0005579	membrane attack complex	8.3	0.016	1
molecular_function	GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	3.2	0.00014	0.39
	GO:0003824	catalytic activity	1.6	0.00022	0.39
	GO:0004346	glucose-6-phosphatase activity	Infinite	0.00061	0.70
	GO:0016787	hydrolase activity	1.8	0.00097	0.84
	GO:0005506	iron ion binding	2.1	0.0012	0.85

Supplemental materials and methods

Studied cohort

This study was approved by the institutional review board of Yokohama City University School of Medicine (Yokohama, Japan). All individuals analyzed here were sent to our department to perform whole exome sequencing (WES) analysis for their genetic diagnosis. First, to optimize DNV detection flow and adjustment of expected DNV rate with depth, we analyzed a data set of 1,330 trios with variable rare diseases (such as DEE, multi-malformation syndrome, mitochondrial disease, congenital myopathy, Charcot-Marie-Tooth disease, hemophilia, Ehlas-Danlos syndrome, etc). The inclusion criteria were that 1) the child was diagnosed by attending physicians, 2) the parents were healthy, 3) all three members in each trio were analyzed by WES with the same capture kit (i.e. SureSelect Human All Exon V4, V5, or V6) and 4) parentage between children and their parents was confirmed by calculation of PI-HAT with WES data using PLINK as previously described (data not shown).[1,2] We extracted data of 346 DEE trios from the above 1,330 trio analysis to identify candidate variants causative for DEE. A part (n=195) of probands of the 346 trios were analyzed in our previous paper which focused on ultra-rare variants, but not DNV.[1]

We also analyzed independent 1,406 DEE individuals to screen variants in candidate genes. The inclusion criteria of the 1,406 DEE individuals were that 1) the individual was diagnosed as DEE by attending physicians and 2) samples of either or both the parents were lacking or available samples of the parents were not analyzed by WES or all three members in each trio were not analyzed by WES with the same capture kit.

Genetic drift simulation

We performed genetic drift simulation using a previously reported C++ script.[3] The script generates the number variant sites (N) observed in 33,370 individuals (the number non-Finnish Europeans in ExAC) in the last generation after mutation and selection events

repeated in an effective population size of 100,000 through 1,000,000 generations. The selection event was modeled by Wright-Fisher sampling. We set the following parameters: selection coefficient=0; dominance coefficient=0; per-gene number of possible variant sites=10, 32, 100, 320, or 1000; per-site rate of DNV=3.16e-10, 1.0e-9, 3.16e-9, 1.0e-8, 3.2e-8, or 1.0e-7. We repeated simulations of each parameter at least 300 times. We calculated the mean per-gene number of variant sites observed in the last generation across the >300 simulations of each parameter setting. We analyzed correlations between the mean and total DNV rate per gene (= *per-site DNV rate* × *per-gene number of possible DNV sites*).

Number of variants in ExAC

We downloaded files of variants in Exome Aggregation Consortium, ExAC (ExAC.r0.3.nonpsych.sites.vcf.gz) from the ExAC Browser (Beta). We counted variants of 1) passed by variant quality score recalibration (VQSR) filtering, 2) that were bi-allelic, and 3) that were annotated as synonymous by SnpEff. For each gene, we only considered 20,042 transcripts that were protein-coding (-onlyProtein option in SnpEff), canonical (-canon option in SnpEff), and in GENCODE V19 (gencode.v19.annotation.comprehensive.gff3).

Per-gene DNV rate

We downloaded sequences of each exon plus a 40 bp margin of the above 20,042 transcripts from the UCSC Table Browser (track: GENCODE Genes V19; table: Comprehensive). For NMD(-) regions, we downloaded genomic positions of each exon from the UCSC Table Browser (output format: BED). Using these sequences, we generated a VCF file containing all possible variants at each exon plus a 39 bp margin and a tri-nucleotide context (ex. AAA to ATA). For each of the variants, we obtained 1) variant consequence using SnpEff, 2) DNV rate using a previously reported table of DNV

rates for each tri-nucleotide context (https://github.com/pjshort/dddMAPS), and 3) per-base median depth in ExAC or our cohort (see "Per-base median depth"). Because sequencing depth can affect the number of observed variants, we adjusted the rate of DNV at each site according to per-base median depth as follows: 1) 0.04 × depth + 0.125 (depth: <12.5), 0.177 × log (depth – 6.29) + 0.302 (depth: >12.5 and <55), or 1 (depth: >55) for the analysis of ExAC and 2) 0.025 × depth (depth: <40) and 1 (depth: >40) for the analysis of our 346 DEE trios. The formulas were generated from the curves of median depth and the ratio of DNV rate and observed number of variants (Fig. S2). We summed the depth-adjusted DNV rate of each synonymous, missense, and nonsense variant for each transcript. We calculated DNV rates of frameshift indels by multiplying that of nonsense variants by 1.25, which was the ratio of singleton frameshift to singleton nonsense DNVs found in exome sequencing data from roughly 2,000 autism spectrum disorder subjects and controls.[4] We provide non-depth-adjusted DNV rates for each variant type for the whole coding region or NMD(-) region of 20,042 transcripts for the benefit of readers (Table S1 and S2) because depth is dependent on each sequencing platform.

Per-base median depth

For the purpose of depth adjustment of the DNV rate described above, we used per-base median depth for ExAC samples in Panel.chr.coverage.txt.gz downloaded from the ExAC Browser (Beta). For our 346 DEE cohort, we selected 100 random samples analyzed using SureSelect Human All Exon V4, 5, and 6 Kits and calculated the median depth for each of kit using samtools.

Whole exome sequencing (WES)

WES was performed as previously described.[5] In brief, DNA was extracted from peripheral blood samples using QuickGene-610L (Fujifilm) according to the manufacturer's protocols, captured with a SureSelect Human All Exon V4, 5, or 6 Kit

(Agilent Technologies, Santa Clara, CA, USA) and sequenced on an Illumina HiSeq 2500 (Illumina, San Diego, CA, USA) with 101-bp paired-end reads. Among 1,330 trios, we sequenced 121, 886, and 323 trios and among 346 DEE trios we sequenced 16, 196, and 134 trios using V4, 5, and 6 kits, respectively. Reads were aligned to the human reference genome (GRCh37/hg19) using Novoalign v3.02.13. PCR duplicates were removed using Picard. Local realignments around indels and base quality score recalibration were performed with the Genome Analysis Toolkit (GATK) 3.7-0.[5] Variants were called by GATK HaplotypeCaller, and their quality score (VQSLOD, variant quality score log-odds) was calculated according to the GATK Best Practices.[6] Variants were annotated based on the canonical transcripts using SnpEff (with the -canon and -onlyProtein option).[7]

Variant filtering and *de novo* calling

We filtered variants as follows: 1) <0.001 minor allele frequency (MAF) in ExAC and ESP6500, 2) not existing in an in-house database comprising 575 healthy adult Japanese, and 3) called by GATK in any of the probands but not in any of the parents of the 346 DEE trios. The remaining variants were analyzed with one filtering system: DenovoFilter[8] and three scoring systems: TrioDenovo,[9] DNMfilter,[10] and Denovogear.[11] We set the threshold of VQSLOD and the scores of the three systems as follows: VQSLOD: >-7.18, TrioDenovo: >5.72, DNMfilter: >0.196, and Denovogear: >0.02 for SNVs; VQSLOD: -1.06 and TrioDenovo: >5.5, for indels.

Per-gene enrichment analysis of DNVs

Enrichment of DNVs was analyzed by testing the null hypothesis that the observed number of DNVs is equal to λ, the expected number of DNVs under the Poisson distribution.[8] The observed DNVs indicate those detected in the flow described above (see "Variant filtering and *de novo* calling"). The expected number of DNVs was

calculated as: (per-gene rate of DNV) × (number of analysed trios) × 2. We regarded stop-gained, canonical splice-site, and frameshift variants as truncating variants. The exome-wide threshold for statistical significance was Bonferroni-corrected for the total number of analyzed genes: 0.05/20,042 (=2.5E-6). To obtain expected number of DNV across 346 DEE trios, independent 1,406 DEE individuals, and 4,293 DDD individuals, we calculated as follows: (non-depth-adjusted rate of truncating DNV in NMD(-) region of per chromosome per generation: 6.18E-7) × (number of individuals: 346 + 1,406 + 4,293) × (number of chromosome per individual: 2).

Sanger sequencing

SEMA6B variants were confirmed by Sanger sequencing using standard methods. Briefly, PCR products were purified with alkaline phosphatase and exonuclease 1 and sequenced using a BigDye Terminator v3.1 Cycle Sequencing kit (Applied Biosystems, Foster City, CA, USA) on a 3500 DNA Sequencing Analyzer (Life Technologies, Carlsbad, CA, USA).

RNA analysis

To examine if candidate pathogenic DNVs induce NMD, we performed Sanger sequencing using complementary DNA (cDNA) reverse transcribed from total RNA of Individual 2. First, lymphoblastoid cell lines (LCLs) of Individual 2 were maintained in Roswell Park Memorial Institute 1640 medium supplemented with 10% fetal bovine serum, 8 μg/ml tylosin (Sigma-Aldrich, St. Louis, MO, USA), and antibiotic-antimycotic solution in a 5% CO_2 incubator. Total LCL RNA was obtained using an RNeasy Plus Mini Kit (Qiagen). First strand cDNA was synthesized using the SuperScript III First-Strand Synthesis System (Invitrogen). The synthesized cDNA was amplified by PCR with specific primers (forward 5′ -GGGACTGCACAGGACTCC-3′ and reverse 5′ -AGCAGGGCCTCCGGGGGAA-3′) and subsequently used for Sanger sequencing.

Single cell RNA sequencing (scRNA-seq) analysis of mouse brain

We reanalyzed two public data sets of mouse brain scRNA-seq: 1) variable brain regions of at least two P12-60 mice[12] and 2) variable brain regions of one P2 and one P11 mouse.[13] In each original paper, the quality of each cell was checked, and cell type of each cell was annotated. Because evidence for the annotation of each cell cluster was not fully publically available, we confirmed the appropriateness of cell cluster annotations by review of marker gene expression studies and the Allen Mouse Brain Atlas (Table S4 and S5). We analyzed cells using Seurat according to the developers' guide.[14] Briefly, we normalized numbers of unique molecular identifiers (nUMIs) of each gene by total nUMIs in each cell, multiplied by a scale factor (10,000), log-transformed, and linear-regressed by total nUMIs and mitochondrial gene expression level. The residuals of linear-regression were scaled, and the Z-score of the residuals were used as expression levels. We analyzed the percentage of *Sema6b*-expressing cells and mean *Sema6b* expression in each cell type using DotPlot function.

Bulk RNA-seq and snRNA-seq analysis of human brain

We analyzed public data of bulk RNA-seq of variable regions of human brain at variable stages of development in BrainSpan Atlas of the Developing Human Brain (https://www.brainspan.org/static/download.html). RPKM values of *SEMA6B* in each region of human brain at each developmental stage were downloaded. When multiple samples (donors) were available for a brain region and a developmental stage, RPKM values were averaged among the multiple samples. We omitted data of a developmental stage when no sample was available in most brain regions or a brain region where no samples was available in most developmental stages. We also analyzed public data of snRNA-seq of variable cell types of adult human brain in Allen Brain Map (https://portal.brain-map.org/atlases-and-data/rnaseq). The snRNA-seq data contained

about 49,000 cells from variable brain regions (middle temporal gyrus, anterior cingulate gyrus, primary visual cortex, primary motor cortex, primary somatosensory cortex and primary auditory cortex) of three donors (a 43-year female, a 50-year male, and a 54-year male). Log2(CPM) values of *SEMA6B* in each of cell type were downloaded and shown as expression levels. A gene count-cell matrix was downloaded, and proportions of cells expressing at least one count of *SEMA6B* in each cell type was calculated.

Protein domain and gene ontology analysis

We analyzed whether specific protein domains and gene ontology terms are enriched among genes depleted for truncating variants at NMD(-) regions. A previous research analyzed WES data of control populations in ExAC (n=60,706) or ARIC (Atherosclerosis risk in communities study, n=10,940) and detected 1,996 transcripts depleted for truncating variants at NMD(-) regions.[15] Among the 1,996 transcripts, we analyzed 1,594 transcripts overlapped with our 20,042 transcripts which are canonical and protein-coding as described above. For protein domain analysis, we downloaded genomic regions annotated with Pfam protein domains from UCSC Table Browser. We annotated NMD(+) or NMD(-) regions of each transcript with overlapping Pfam protein domains. Note that a Pfam protein domain could overlap with both of NMD(+) and NMD(-) region in a transcript. We analyzed genes related to at least one Pfam protein domain: 1,498 genes depleted for truncating variants at NMD(-) regions and the other non-depleted 18,081 genes. We examined statistical significance of enrichment of each Pfam protein domain in NMD(+) or NMD(-) regions among genes depleted for truncating variants at NMD(-) regions (n=1,498) comparing with the other non-depleted genes (n=18,081) using one-sided Fisher's exact test.

For gene ontology analysis, we downloaded gene ontology terms which each transcript (ENST ID) has and evidence code of each term using the biomaRt library in R from the grch37.emsembl.org host. We removed GO terms whose evidence code was ND (no data).

We separately analyzed three domains of GO terms: biological process, cellular component, and molecular function. We analyzed genes related to at least one GO term as follows: 1,208 genes depleted for truncating variants at NMD(-) regions and the other 13,631 genes for GO terms for biological process, 1,347 genes depleted for truncating variants at NMD(-) regions and the other 14,800 genes for GO terms for cellular component, and 1,342 genes depleted for truncating variants at NMD(-) regions and the other 14,513 genes for GO terms for molecular function. For multiple testing, we corrected the thresholds of statistical significance with Benjamini-Hochberg method (thresholds for q value: 0.05). We examined statistical significance of enrichment of each GO term among genes depleted for truncating variants at NMD(-) regions comparing with the other genes using one-sided Fisher's exact test.

Zebrafish Studies

Zebrafish maintenance and ethics

Wild-type zebrafish, *Danio rerio*, were obtained from the National Research Institute of Fisheries Science (Yokohama, Japan). Zebrafish were maintained at 28.5°C ± 0.5°C with a 14 h light/10 h dark cycle, fed twice a day and maintained in tanks with circulating water. The zebrafish experiments were authorized by the institutional committee on fish experiments of the National Research Institute of Fisheries Science.

CRISPR/Cas9 genome editing and T7 endonuclease 1 assay

The zebrafish genome has two *SEMA6B* orthologs: *sema6ba* (Ensembl ID: ENSDART00000179089.1) and *sema6bb* (Ensembl ID: ENSDART00000161646.1). CRISPR/Cas9 target sites for zebrafish *sema6ba* and *sema6bb* were identified with the CRISPRdirect web tool. The target exons chosen were exon 2 and exon 17 (last exon) (Fig. S6 and Table S6). The individual synthetic CRISPR RNAs (crRNAs) and synthetic transactivating CRISPR RNA (tracrRNA) were obtained from Integrated DNA

Technologies for use with the Alt-R® CRISPR/Cas9 System (IDT, Coralville, IA, USA). Combined multiple guides per target increased genome editing efficiency, two crRNAs for exon 17 were designed and mixed at a 1:1 ratio. To create gRNA complexes, equimolar amounts of crRNA and tracrRNA were resuspended to 3 μM in nuclease-free duplex buffer (IDT) and heated to 95ºC for 5 min according to the manufacturer's specifications. Cas9 protein (Alt-R® S.p. Cas9 Nuclease 3NLS, IDT) was diluted to a working concentration (0.5 μg/μl) using Cas9 working buffer (20 mM HEPES; 150 mM KCl, pH 7.5). Cas9 protein was combined with gRNA complexes to form ribonucleoprotein (RNP) complexes and incubated at 37ºC for 10 min. RNP complexes were injected into one-cell stage embryos for genome editing. To detect mutations using the T7 endonuclease I (T7E1) assay, the DNA fragments containing the targeted sites were amplified from genomic DNA using primer pairs (Table S6). Annealed PCR products were then digested with 5U of T7E1 for 2 h at 37°C. The T7E1-digested products were separated by agarose gel electrophoresis.

Heteroduplex mobility assay (HMA)

PCR at the target sites were performed with TaKaRa Ex Taq (TaKaRa, Shiga, Japan) and the locus-specific primers listed in the Table S6. The PCR amplicons were electrophoresed on a 15% polyacrylamide gel (Wako, Osaka, Japan).

RT-PCR and pharmacological treatment

To inhibit NMD, 72 hpf uninjected, sema6ba/b-ex.2 and sema6b/a-ex.17 F0 crispants were treated with 10 μM NMDi14 (SML1538) (Sigma-Aldrich) or DMSO (as a vehicle), and three days later, larvae were collected in TRIzol (Thermo Fisher Scientific, Waltham, MA, USA) for RT-PCR.[16] The cDNA was synthesized using M-MLV reverse transcriptase (Promega, Madison, WI, USA), followed by PCR with ExTaq (Takara). As for the internal control primers, ef1-alpha were used. Primer sets shown in Table S6.

Whole-mount immunohistochemistry

For acetylated tubulin staining, 55 hpf embryos were fixed in Dent's fixative (80% methanol and 20% dimethyl sulphoxide) overnight at 4°C. Embryos were permeabilized with proteinase K followed by post-fixation with 4% PFA and washing with PBSTX (PBS containing 0.5% Triton X-100). After treating with 4% normal goat serum (NGS) in PBSTX for 2 h at room temperature, embryos were incubated with mouse anti-acetylated tubulin antibodies (1:1000, T7451, Sigma-Aldrich) in 4% NGS/PBSTX overnight at 4°C.[17] Images were taken using a Fluoview FV1000-D confocal microscope (Olympus Corporation, Hachioji, Tokyo) and quantification of the optic tecta area was performed using a digital fluorescence BZ-X800 microscope (KEYENCE, Osaka, Japan).

PTZ-induced seizures

For the zebrafish seizure model, we followed a previously described method.[18] Briefly, for locomotor tracking, 7 dpf wild-type (uninjected), *sema6ba*/*b*-ex.2 and *sema6ba*/*b*-ex.17 F0 crispant larvae were individually placed in a 96-well plate (1 larva/well) with 250 µl normal bathing medium. The larvae were incubated with 2.5 mM, 5 mM or 15 mM PTZ (pentylenetetrazole, P6500, Sigma-Aldrich). Seizure-type locomotor activity was quantified using the DanioVision™ video tracking system (Noldus, Wageningen, The Netherlands) and movement quantification software (EthoVision XT 11, Noldus). Image analysis used a dynamic substitution algorithm with individual frame weight of 5%. Tools for smoothing data to eliminate background noise were used (MDM 0.2 mm). Each treatment contained two different sets of embryos, with each set containing 16 embryos.

Statistical analysis

Statistical analyses were performed with GraphPad Prism 8 software (GraphPad Software, Inc., San Diego, CA, USA) using a chi-square test or Student's t test for qualitative data. Significance for all tests was defined at * $P < 0.05$; ** $P < 0.01$; *** $P < 0.0005$; **** $P < 0.0001$.

Code availability

The codes for analyses in this study are available (github.com/hamanakakohei).

Supplemental note

Correlation between total DNV rate per gene and the number of sites of observed variants per gene

Previously, total synonymous DNV rate per gene expected from a model and the number of sites of observed synonymous variants per gene were correlated in ExAC and gnomAD.[19,20] To generalize this finding, we performed genetic drift simulations which output the number of sites of variants observed in the last generation after *de novo* mutation and selection events repeated through multiple generations.[3] To simulate synonymous DNV events, we set the following parameters: selection coefficient=0 and dominance coefficient=0 where variants are under no natural selection; per-site rate of DNV=3.16e-10, 1.0e-9, 3.16e-9, 1.0e-8, 3.2e-8, or 1.0e-7, which covers DNV rates of single nucleotide variants for variable trinucleotide contexts.[20] The simulation showed that total DNV rate per gene [= (per-site DNV rate) × (per-gene number of DNV sites)] and the number of sites of variants observed in the last generation per gene were correlated among variable settings of per-site DNV rate and per-gene number of DNV sites (on the same line, Fig. S1).

Optimization of DNV detection flow

We optimized our flow for DNV detection using our 1330 trios. In the 1330 trios, we detected a total of 5,581 single nucleotide variants (SNVs) and 596 indel DNV candidate calls violating Mendelian inheritance across the exome (not restricted to the last two exons at this stage). This number is apparently too large given the exonic DNV rate (~one DNV per diploid exome). However, we noticed that these DNV candidate calls contain 115 SNVs and 25 indels previously Sanger-confirmed as true-positive (TP) DNVs. We analyzed the DNV candidate calls with four scoring systems and one filtering system (see Method, “Variant filtering and *de novo* calling”). To maximize the sensitivity for TP DNVs, we determined the threshold of the four scoring systems that did not miss any TP

DNVs. Only one filtering system, DenovoFilter, missed five TP DNVs. After filtering with the five systems, sensitivity was 95.7% (110/115) for SNV TP DNVs and 100% (25/25) for indel TP DNVs. We considered DNV calls that passed the filtering with the three methods were likely true-positive DNVs (DNVs hereafter). Consequently, we observed a total of 2,045 SNVs and 211 indel DNVs in 1330 trios, 542 SNVs and 50 indels in 346 DEE trios (Fig. 1B).

Enrichment of specific protein domains or gene ontology terms among genes depleted for truncating variants at NMD(-) regions

To infer the pathomechanism of truncating DNV at NMD(-) region in *SEMA6B*, we analyzed functional properties of genes depleted for truncating variants at NMD(-) regions. We analyzed enrichment of specific protein domains or gene ontology terms among 1,594 genes depleted for truncating variants at NMD(-) regions in general populations (ExAC or ARIC) as previously reported.[15] We found enrichment of Pfam protein domains at NMD(+) or NMD(-) regions: two domains (KRAB and p450) at NMD(+) region and one domain (p450) at NMD(-) region exceeded a threshold for statistical significance (q-value: 0.05, Table S7). The KRAB domain and highly ranked domains (but not statistically significant) such as SCAN and zf-C2H2_6 were observed in KRAB-ZFP (zinc finger proteins). The KRAB domain functions as a transcriptional repressor; SCAN domain mediates oligomerization; zf-C2H2 domain locates at C-termini of proteins and binds DNA.[21] Thus, KRAB-ZFP without their C-termini might sequester wild-type one but not bind to DNA and perturb transcriptional repressor activity. Other highly ranked domains (but not statistically significant) such as Filament and Keratin_2_head are observed in type 2 keratin proteins. Filament domain functions mediates self-assembly in to filaments; Keratin_2_head and Keratin_2_tail domains locate at N-termini and C-termini, respectively, and interacts with other proteins.[21] Thus, type 2 keratin proteins without their C-termini could oligomerize with wild-type ones but

not interact with other proteins via their Keratin_2_tail domain and might abrogate their functions. P450 domain is seen in CYP450 proteins. CYP450 proteins could oligomerize and exert enzymatic activities using heme bound at their C-termini.[22,23] CYP450 proteins without their C-termini might abrogate their enzymatic activities. We also analyzed enrichment of GO terms about biological process, cellular component, or molecular function, but no terms exceeded a threshold of statistical significance (q-value: 0.05, Table S8). However, GO terms (GO:0005882 and GO:0045111) related to intermediate filament were highly ranked, consistent with the results of above analysis of protein domains (Table S8).

These results might suggest that truncating variants at NMD(-) regions exert the pathogenicity through oligomerization. The pathomechanism might be compatible with *SEMA6B* variants found here because class 6 semaphorin proteins dimerize via the sema domain (Fig. 1F).[24]

Supplemental references

1. Takata, A., Nakashima, M., Saitsu, H., Mizuguchi, T., Mitsuhashi, S., Takahashi, Y., Okamoto, N., Osaka, H., Nakamura, K., Tohyama, J., et al. (2019). Comprehensive analysis of coding variants highlights genetic complexity in developmental and epileptic encephalopathy. Nat Commun 10, 2506.
2. Hamanaka, K., Takata, A., Uchiyama, Y., Miyatake, S., Miyake, N., Mitsuhashi, S., Iwama, K., Fujita, A., Imagawa, E., Alkanaq, A.N., et al. (2019). MYRF haploinsufficiency causes 46,XY and 46,XX disorders of sex development: bioinformatics consideration. Hum Mol Genet 28, 2319-2329.
3. Fuller, Z.L., Berg, J.J., Mostafavi, H., Sella, G., and Przeworski, M. (2019). Measuring intolerance to mutation in human genetics. Nat Genet 51, 772-776.
4. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnstrom, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat Genet 46, 944-950.
5. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297-1303.
6. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491-498.
7. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6, 80-92.

8. JF, M., S, C., TW, F., J, K., E, P., D, R., A, S., S, A., N, A., M, A., et al. (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature 542, 433-438.

9. Wei, Q., Zhan, X., Zhong, X., Liu, Y., Han, Y., Chen, W., and Li, B. (2015). A Bayesian framework for de novo mutation calling in parents-offspring trios. Bioinformatics 31, 1375-1381.

10. Liu, Y., Li, B., Tan, R., Zhu, X., and Wang, Y. (2014). A gradient-boosting approach for filtering de novo mutations in parent-offspring trios. Bioinformatics 30, 1830-1836.

11. Ramu, A., Noordam, M.J., Schwartz, R.S., Wuster, A., Hurles, M.E., Cartwright, R.A., and Conrad, D.F. (2013). DeNovoGear: de novo indel and point mutation discovery and phasing. Nat Methods 10, 985-987.

12. Zeisel, A., Hochgerner, H., Lonnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Haring, M., Braun, E., Borm, L.E., La Manno, G., et al. (2018). Molecular Architecture of the Mouse Nervous System. Cell 174, 999-1014.e1022.

13. Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W., et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science 360, 176-182.

14. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36, 411-420.

15. Coban-Akdemir, Z., White, J.J., Song, X., Jhangiani, S.N., Fatih, J.M., Gambin, T., Bayram, Y., Chinn, I.K., Karaca, E., Punetha, J., et al. (2018). Identifying Genes Whose Mutant Transcripts Cause Dominant Disease Traits by Potential Gain-of-Function Alleles. Am J Hum Genet 103, 171-187.

16. El-Brolosy, M.A., Kontarakis, Z., Rossi, A., Kuenne, C., Gunther, S., Fukuda, N., Kikhi, K., Boezio, G.L.M., Takacs, C.M., Lai, S.L., et al. (2019). Genetic compensation triggered by mutant mRNA degradation. Nature 568, 193-197.
17. Tsurusaki, Y., Koshimizu, E., Ohashi, H., Phadke, S., Kou, I., Shiina, M., Suzuki, T., Okamoto, N., Imamura, S., Yamashita, M., et al. (2014). De novo SOX11 mutations cause Coffin-Siris syndrome. Nat Commun 5, 4011.
18. Baraban, S.C., Taylor, M.R., Castro, P.A., and Baier, H. (2005). Pentylenetetrazole induced changes in zebrafish behavior, neural activity and c-fos expression. Neuroscience 131, 759-768.
19. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv doi: https://doi.org/10.1101/531210.
20. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285-291.
21. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. Nucleic Acids Res 47, D427-d432.
22. Liu, J., Tawa, G.J., and Wallqvist, A. (2013). Identifying cytochrome p450 functional networks and their allosteric regulatory elements. PLoS One 8, e81980.
23. Bostick, C.D., Hickey, K.M., Wollenberg, L.A., Flora, D.R., Tracy, T.S., and Gannett, P.M. (2016). Immobilized Cytochrome P450 for Monitoring of P450-P450 Interactions and Metabolism. Drug Metab Dispos 44, 741-749.

24. Nogi, T., Yasui, N., Mihara, E., Matsunaga, Y., Noda, M., Yamashita, N., Toyofuku, T., Uchiyama, S., Goshima, Y., Kumanogoh, A., et al. (2010). Structural basis for semaphorin signalling through the plexin receptor. Nature 467, 1123-1127.