SCIENTIFIC REPORTS
natureresearch

OPEN

Genomic profiling of 553 uncharacterized neurodevelopment patients reveals a high proportion of recessive pathogenic variant carriers in an outbred population

Youngha Lee[1,9], Soojin Park[2,9], Jin Sook Lee[2,3,4,9], Soo Yeon Kim[2], Jaeso Cho[1], Yongjin Yoo[1], Sangmoon Lee[1], Taekyeong Yoo[1], Moses Lee[1], Jieun Seo[1], Jeongeun Lee[1,5], Jana Kneissl[1], Jean Lee[1], Hyoungseok Jeon[1], Eun Young Jeon[1], Sung Eun Hong[1], Eunha Kim[1], Hyuna Kim[2], Woo Joong Kim[2], Jon Soo Kim[6], Jung Min Ko[2], Anna Cho[7], Byung Chan Lim[2], Won Seop Kim[8], Murim Choi[1,2*] & Jong-Hee Chae[2*]

A substantial portion of Mendelian disease patients suffers from genetic variants that are inherited in a recessive manner. A precise understanding of pathogenic recessive variants in a population would assist in pre-screening births of such patients. However, a systematic understanding of the contribution of recessive variants to Mendelian diseases is still lacking. Therefore, genetic diagnosis and variant discovery of 553 undiagnosed Korean patients with complex neurodevelopmental problems (KND for Korean NeuroDevelopmental cohort) were performed using whole exome sequencing of patients and their parents. Disease-causing variants, including newly discovered variants, were identified in 57.5% of the probands of the KND cohort. Among the patients with the previous reported pathogenic variants, 35.1% inherited these variants in a recessive manner. Genes that cause recessive disorders in our cohort tend to be less constrained by loss-of-function variants and were enriched in lipid metabolism and mitochondrial functions. This observation was applied to an estimation that approximately 1 in 17 healthy Korean individuals carry at least one of these pathogenic variants that develop severe neurodevelopmental problems in a recessive manner. Furthermore, the feasibility of these genes for carrier screening was evaluated. Our results will serve as a foundation for recessive variant screening to reduce occurrences of rare Mendelian disease patients. Additionally, our results highlight the utility and necessity of whole exome sequencing-based diagnostics for improving patient care in a country with a centralized medical system.

A large fraction of Mendelian disorders follow a recessive inheritance pattern[1,2]. The Online Mendelian Inheritance in Men (OMIM) lists 5,317 disorders and 3,077 of these are categorized as recessive (as of April 2019). For more common complex disorders, the contribution of recessive variants to the disease pathogenesis is less than expected[3–6]. For rare diseases, the contribution of recessive variants in inbred populations, such as

[1]Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul, 03080, Republic of Korea. [2]Department of Pediatrics, Seoul National University College of Medicine, Seoul, 03080, Republic of Korea. [3]Department of Pediatrics, Gil Medical Center, Gachon University College of Medicine, Incheon, 21565, Republic of Korea. [4]Department of Genome Medicine and Science, Gil Medical Center, Gachon University College of Medicine, Incheon, 21565, Republic of Korea. [5]Interdisciplinary Program for Bioengineering, Graduate School, Seoul National Universty, Seoul, 03080, Republic of Korea. [6]Department of Pediatrics, Chungbuk National University Hospital, Cheongju, 28644, Republic of Korea. [7]Department of Pediatrics, Ewha Womans University School of Medicine, Seoul, 07804, Republic of Korea. [8]Department of Pediatrics, College of Medicine, Chungbuk National University, Cheongju, 28644, Republic of Korea. [9]These authors contributed equally: Youngha Lee, Soojin Park, and Jin Sook Lee. *email: murimchoi@snu.ac.kr; chaeped1@snu.ac.kr

Middle-Eastern countries, has been well proven[7–9]. However, the precise contribution of recessive variants to rare Mendelian disorders in an outbred population is still not well understood.

The inherent complexity of brain developmental processes inevitably leads to patients with diverse neurological problems that frequently challenge conventional diagnostic criteria. Therefore, diagnosis of neurological disorders that affect children is frequently hampered by rare and overlapping clinical features, which makes it difficult for clinicians to readily recognize and properly treat the disease entity. This makes pediatric neurologic patients an impending target for genome-wide genetic studies[10–12]. To facilitate diagnosis and discovery of novel disease pathophysiology, large-scale systematic efforts have been conducted at regional or national scales[13–16]. As many rare neurologic disorders in children follow Mendelian inheritance, disease-causing variant discovery by trio-based whole-exome sequencing (WES) has proved to be the most robust methodology, yielding instant diagnosis rates of 25–41%[10–12,14,16].

Notably, the medical system in Korea provides a unique opportunity to conduct a systematic survey of rare disorders and study the contribution of recessive variants at a large scale. With a nation-wide referral system focused on a handful of major tertiary clinical institutions, Seoul National University Children's Hospital (SNUCH) covers a large portion of the 51-million population, allowing for consistent evaluation and treatment of the patient cohort. For example, we recently reported on genetic analyses of large patient cohorts of Duchenne muscular dystrophy ($n = 507$) and Rett-like syndrome lacking *MECP2* mutations ($n = 34$)[17,18]. Genetically, ethnic Koreans are a good example of an outbred population in which marriages between relatives and even between individuals possessing the same surnames were prohibited since the 17th century, although the same surname marriage ban was lifted in 2005[19]. Our study represents the largest of its kind that was conducted at a single clinic, and emphasizes the careful integration of clinical and genetic analyses.

We used WES to analyze a cohort of 553 patients (KND cohort) with severe neurodevelopmental disorders. We characterized the genotype-phenotype relationships of patients whose molecular defects had been identified, and explored the potential association of genes that had not been previously associated with disease. We demonstrate that a high proportion of recessively-inherited variants are associated with patients that have rare neurodevelopmental diseases. Variants that were inherited in a recessive manner were analyzed and their genetic properties were evaluated, aiming to understand their distribution in healthy populations to minimize emergence of rare disease patients. Overall, we describe the establishment of a system that efficiently integrates advanced genetic techniques with clinical diagnostic processes to maximize benefits for pediatric patients and their families.

Results

Diagnostic success rate of WES analyses.

The symptoms experienced by KND patients were mostly of pediatric onset (mean 1.4 years of age). The patients harbored neurodevelopmental problems and were soon referred to tertiary hospitals (mean 1.8 years of age). The majority of the patients visited multiple tertiary hospitals for diagnosis (88.8% visited more than one hospital, mean of 2.3 hospitals), required a mean of 2.3 specialists (31.6% required more than two) and a mean of 5.6 years elapsed before WES analysis at SNUCH (Table 1, Supplementary Fig. S2). The distribution of straight-line distances from home to the clinic strongly correlates with the original population distribution of Korea, suggesting that our cohort covers the entire population (Table 1, Supplementary Fig. S3).

The majority of the patients is sporadic origin (504/553 = 91.1%; Fig. 1a), making them suitable for trio-based WES analysis. Major clinical feature of the KND cohort was neurodevelopmental disease (84.1%; Fig. 1b). Integrative assessments of genetic variants, their clinical impacts, and patient symptoms allowed us to diagnose 40.3% (223/553) of the cohort with high confidence. This group of patients included carriers of CNVs (23/553 = 4.2%; 16 heterozygous deletions and 7 duplications; Supplementary Table S1), in which 20 CNVs originated *de novo* (3.6% of the entire cohort), which is slightly lower but comparable to a previous observation[20–23]. Three inherited pathogenic CNVs were identified: a 165.5 kb deletion in a large family containing multiple affected individuals (Supplementary Fig. S4), and 7.2 Mb and 203.2 kb hemizygous duplications on the X chromosome that were transmitted from healthy moms to their affected sons (Supplementary Table S1). In addition to the high confidence group, additional 7.1% of the cohort (39 patients) harbored previously reported variants that were assumed to be pathogenic but displayed distinct phenotypes, potentially expanding the phenotypic spectrum associated with these genes. For example, two patients that carried a pathogenic heterozygous nonsense or missense variant in *COL1A1*, known to cause osteogenesis imperfecta[24], were initially diagnosed with muscle hypotonia. These two patients did not display skeletal problems, but showed blue sclera[25]. Adding this group to the high confidence group yielded an instant diagnostic rate of 47.4% ("known genes"; Supplementary Table S2). Finally, an additional 10.1% of the cohort (56 patients, 53 genes) harbored variants that are highly likely to be pathogenic but their disease associations are elusive ("novel genes"; Fig. 1c). Among the patients with definite diagnosis, 35.1% are recessive, and 29.9% harbored loss-of-function (LoF) variants (Fig. 1d, Supplementary Fig. S5). As expected, the known genes showed strong enrichment in disease categories and gene ontologies such as intellectual disability and central nervous system (CNS) development (Fig. 1e).

Novel genes display potential enrichment in neuronal differentiation.

We assessed whether the 53 novel genes possess a neurologic disease-causing function. The novel gene set was simulated against the BrainSpan data (Materials and Methods) to evaluate if the expression of these novel genes as a group was strongly correlated with known disease-associated genes during brain development. After 10^5 permutations, we found that the observed involvement of the novel genes was significantly stronger than a randomly selected gene sets across eight developmental windows ($P_{adj} < 0.05$ from Z-score for all periods; Fig. 1f,g). Furthermore, this test was expanded to the four anatomical domains in each period, yielding 32 spatio-temporal windows[26] (Fig. 1g). It is notable that the most highly enriched windows are concentrated in the frontal cortex area (R1 × P1–4; Fig. 1g). These results suggest that expression of the novel genes is concordant with known disease-causing genes in developing brains and this phenomenon is most prominent in the frontal cortex.

Sex (*n* (%))	
Male	265 (47.9)
Female	288 (52.1)
Age at symptom onset (years)	1.4(0–21)
Age at first access to a tertiary hospital (years)	1.8(0–22)
Interval between symptom onset and first medical access (months)	3.9 (0–238)
Number of visited tertiary hospitals for diagnosis (*n* (%))	
1	62 (11.2)
2	277 (50.1)
3	178 (32.2)
4	32 (5.8)
5	4 (0.7)
Age at WES (years)	7.4(0–37)
Interval between the first access and WES (months)	
Patients aged 0–10 years	34.0 (0–100)
Patients aged > 10 years	114.5 (7–434)
Primary clinical diagnosis (*n* (%))	
Rett syndrome-like encephalopathy	72 (13.0)
Mitochondrial encephalopathy	49 (8.9)
Epileptic encephalopathy	51 (9.2)
Neuromuscular disorder	37 (6.7)
Leukodystrophy	27 (4.9)
Hereditary spastic paraplegia	34 (6.1)
Others	283 (51.2)
Number of involved specialists for diagnosis (*n* (%))	
1–2	378 (68.4)
3–5	152 (27.5)
>5	23 (4.2)
Straight-line distance from home to the clinic, km (*n* (%))	
<20	186 (33.6)
20–100	180 (32.5)
>100	187 (33.8)

Table 1. Clinical information of 553 patients.

Profile of recessive variants that predispose neurodevelopmental disorders.

Using our set of defined pathogenic variants, we explored the genetic properties of the variants that caused disorders. Both dominant and recessive variants displayed similar proportions by functionality (Supplementary Fig. S5) and carriers of dominant or recessive variants experienced similar ages of onset (data not shown). Next, to test if recessive variants (*i.e*., compound heterozygous (CH), rare homozygous (RHo) and rare hemizygous (RHe) variants) are more frequently found in patients as compared to healthy individuals, we counted the number of recessive variants in our cohort and compared these values between patients and healthy parents as controls. Counting all recessive variants from patients and controls, we observed that there is no substantial difference in the number of variants for CH, RHo and RHe (Fig. 2a). Extracting LoF variants, variants in OMIM-listed genes or variants in neurodevelopment-related genes did not reveal any difference in burden (Fig. 2a, Supplementary Fig. S6), implying the presence of overwhelming non-pathogenic or non-functional recessive variant calls in the patients. The majority of genes that were found in our patients with definite diagnosis has been previously documented in OMIM, and has good concordance with previously known recessive or dominant inheritance patterns (Fig. 2b). There were two exceptional cases in which the genes are listed as recessive in OMIM but were dominantly inherited in our patients. First, only the recessive *ACOX1* phenotype has been recognized to date[27], but we are currently working on a report that describes this dominant *ACOX1* variant. Second, a previous report suggested that the *C19orf12* variant has dominant effect, similar to our observations. But this report is not yet listed in OMIM[28]. Next, to test if the genetic properties of recessive variants are different from those of dominant variants, several parameters were compared. Dominant variants (mean allele frequency $= 6.2 \times 10^{-7}$) were found less frequently in gnomAD than recessive variants (mean allele frequency $= 1.6 \times 10^{-5}$; Mann-Whitney U test $P = 1.7 \times 10^{-13}$), since most of the dominant variants originated *de novo* whereas recessive variants were inherited from healthy parents (Fig. 2c). Recessive variants were slightly less conserved as compared to the dominant variants, based on PhyloP or amino acid conservation among vertebrate species (Mann-Whitney U test $P = 0.034$ and 0.048, respectively; Fig. 2d). Other functionality test values did not differ significantly between the two groups (CADD $P = 0.50$, GERP $P = 0.15$ and SIFT $P = 0.17$, Mann-Whitney U tests). Most of the variants in the two groups were categorized as likely pathogenic or pathogenic according to the American College of Medical Genetics and Genomics (ACMG) guideline (96.7% for dominant and 98.4% for recessive variant group, respectively), although there is a tendency that the dominant variant tend to be more pathogenic

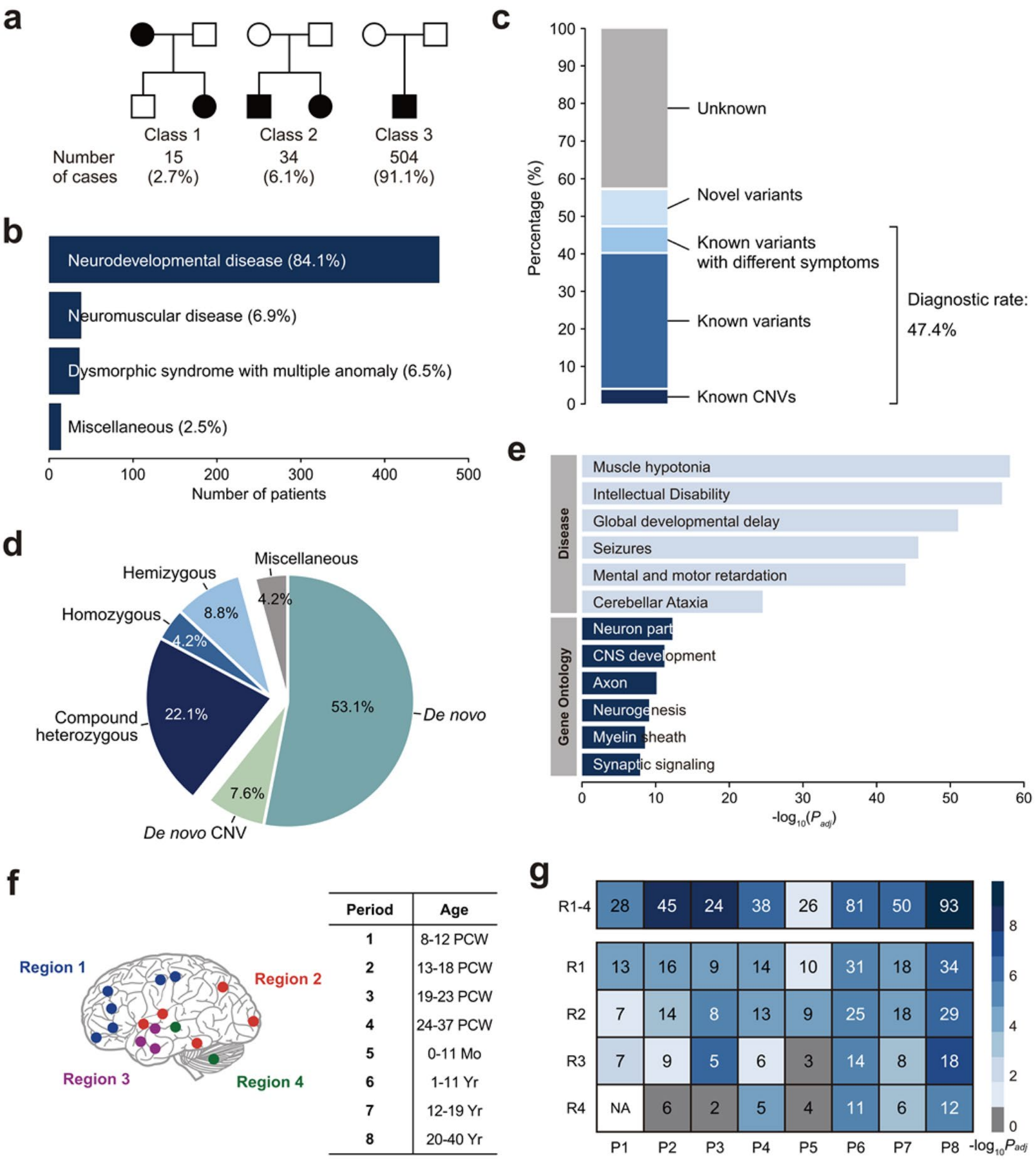

Figure 1. Classification of the KND cohort and results of clinical WES analysis. (**a**) Subjects by disease inheritance patterns. Class 1: autosomal dominant families; Class 2: families with affected siblings; Class 3: affected individuals with no family history. (**b**) Major clinical features of the KND cohort. (**c**) Diagnostic yield of 553 patients with undiagnosed symptoms using WES. (**d**) Pathogenic variants divided by inheritance patterns. (**e**) GO and disease enrichment analysis of 164 known genes. (**f**) Brain anatomical and developmental categorization used for our analysis. Components of each brain region is shown in Supplementary Fig. S9. (**g**) Strength of the co-expression network composed of our known/novel genes compared to random networks as measured by 10^5 permutations.

($P = 3.7 \times 10^{-22}$, Fisher's exact test; Fig. 2e). On the contrary, the genes that contain the recessive variants displayed more lenient constraint as compared to the dominant genes or known haploinsufficiency genes, as documented by observed/expected ratio (o/e) and pLI score in gnomAD[29]. But the recessive genes still display a similar or a slightly more constrained pattern compared to the genes in OMIM (Fig. 2f, Supplementary Fig. S7). Functional characterization of recessive neurodevelopmental disease genes revealed an enrichment for genes involved in lipid metabolism and mitochondrial processes (Fig. 2g), in addition to the expected enrichment in CNS development. The relative position of LoF variants in recessive genes demonstrated a similar lenient pattern, more enriched in the C-terminal portion, as compared to the dominant genes (Fig. 2h). There was no significant difference in basic clinical parameters (displayed in Table 1) between the recessive and dominant patient groups (data not shown).

Profile of pathogenic recessive variants carried in healthy individuals. Unlike *de novo* variants that originate largely at random, assembly of recessive variants can be pre-screened and avoided if such variants can be identified in parents. Taking advantage of the extensive coverage of the patient pool maintained by Korea's

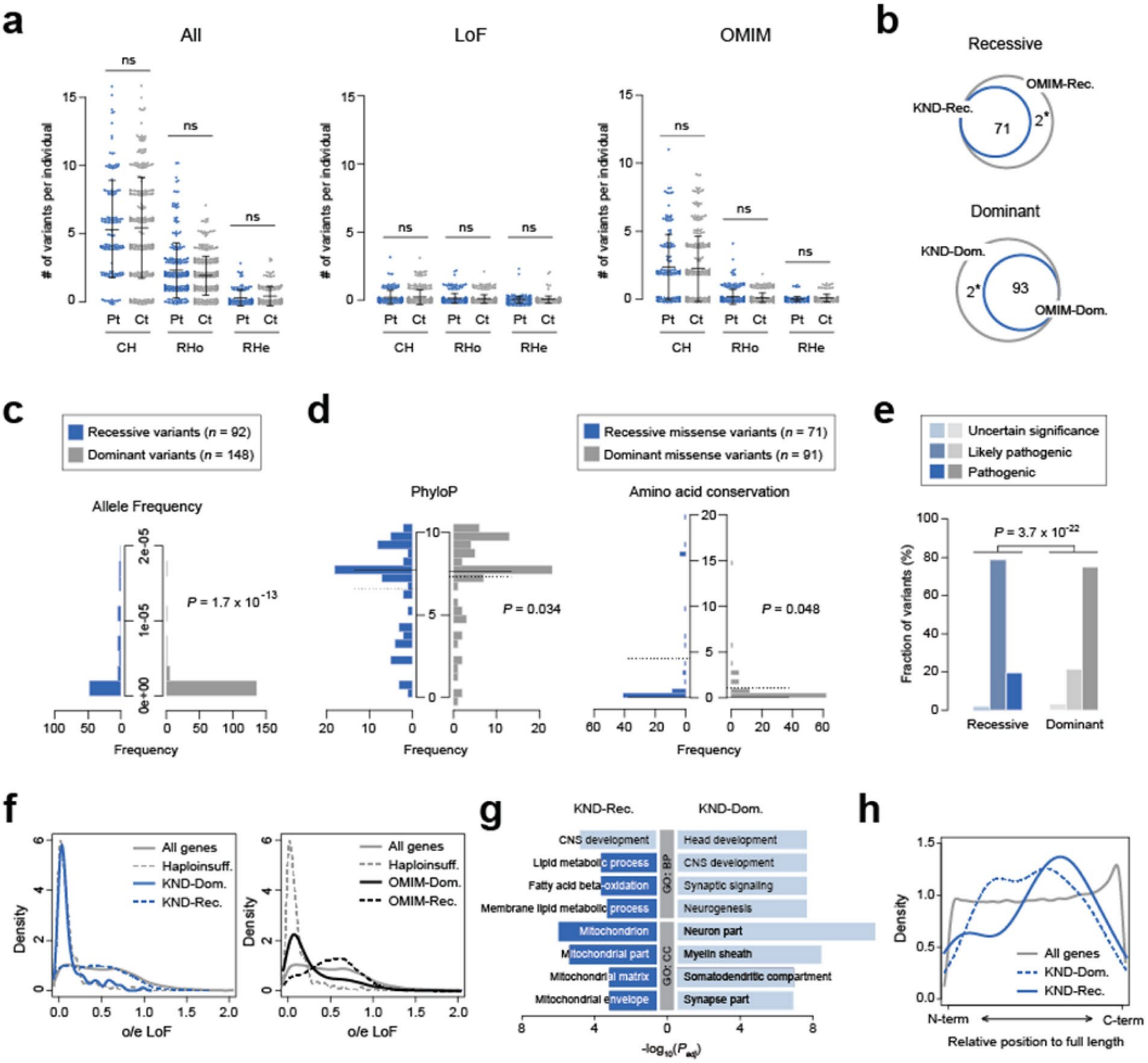

Figure 2. Genetic properties of pathogenic recessive variants. (**a**) Burden of recessive variants in KND patients (Pt) and their parents as controls (Ct). Recessive variants are divided into compound heterozygous (CH), rare homozygous (RHo) and rare hemizygous (RHe) groups. Numbers found from all variants from all genes ("All"), LoF variants from all genes ("LoF"), all variants from OMIM-listed genes ("OMIM") and all variants from intellectual disability gene set ("ID", from DisGeNET) are plotted. Numbers of samples used for each category are as following: patients for CH = 145; controls for CH = 290; patients for RHo = 247; controls for RHo = 341; patients for RHe = 134; controls for RHe = 168. Data are mean ± standard deviation. (**b**) Venn diagrams displaying high correlations of recessive or dominant inheritance patterns with their known inheritance patterns. The asterisks denote two exceptional cases, *ACOX1* and *C19orf12* (see text). (**c**) Allele frequency distribution of dominant and recessive variants. (**d**) PhyloP and amino acid conservation differences between dominant and recessive missense variants. Amino acid conservation is determined by the number of vertebrate species that contain an amino acid that is different from its human orthologous residue. The solid lines denote medians and the dotted lines denote means. (**e**) ACMG code distribution of variants that are in recessive or dominant inheritance pattern. (**f**) Distributions of o/e LoF values for dominant and recessive genes found from KND patients (left) and dominant and recessive genes from OMIM (right) plotted against all genes and known haploinsufficiency genes ($n = 291$)[49]. (**g**) Functional differences between dominant and recessive genes by GO analysis. Ontologies in dark blue suggest non-neuronal signals specific to the recessive gene group. (**h**) Relative position of LoF variants in genes. Positions of pathogenic LoF variants in genes from KND patients are plotted against those LoF variant positions from all genes in gnomAD.

centralized medical system and our analysis results of patients with severe neurodevelopmental disorders, it is feasible to infer the probability of recessive variant assembly in Koreans. Several numbers and assumptions are required for this estimation: (i) approximately 400,000 babies are born every year in Korea as of 2016[30], (ii) approximately 1,000 patients with severe neurodevelopmental disorders newly enroll in our neurodevelopmental

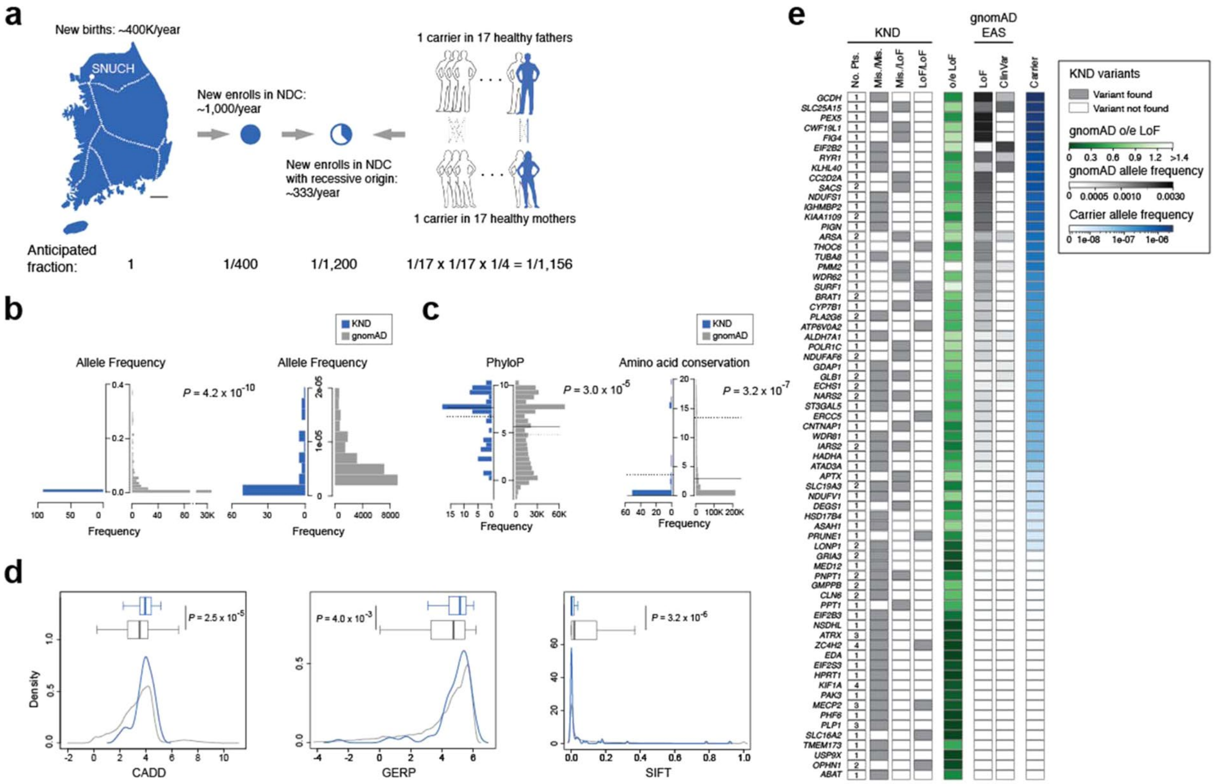

Figure 3. Screen for rare severe neurodevelopmental disorder carriers. (**a**) A schematic diagram describing processes used to estimate neurodevelopmental disorder carrier frequency in the Korean population. The dotted lines in the map denote the Korea Train Express network, the high-speed railway system of Korea. (**b**–**d**) Distribution differences of various parameters between pathogenic recessive variants from KND patients and gnomAD variants from the same genes that were found in KND patients. (**b**) Allele frequency. The rare frequency portion of the left panel is seperately plotted in the right panel. (**c**) PhyloP and amino acid conservation. The solid lines denote medians and the dotted lines denote means. (**d**) CADD, GERP and SIFT score. (**e**) Recessive variants found from KND patients, o/e LoF values, and accumulated frequencies of LoF and ClinVar variants from gnomAD East Asians (EAS) for genes that harbor known pathogenic recessive variants in KND patients. Finally, portion that were attributable to ClinVar or LoF variants for pre-screening parents for each recessive gene are shown.

disorder clinic every year, (iii) these patients encompass the majority of the Korean population, as exemplified by the sizes of our DMD and Rett cohorts[17,18] and reflected in the geographical distribution of the KND patients (Table 1, Supplementary Fig. S3), (iv) our result from 553 patients revealed a recessive genetic origin in approximately one-third of the patients (Fig. 1d) and (v) Koreans typically marry an individual with minimal genetic similarity. These observations lead to a 1/1,200 incidence rate for developing a severe neurodevelopmental disease in a recessive manner, which can be explained by the existence of one carrier for every 17 healthy individuals (1/1,156; Fig. 3a). One can point out limited evidence for one of our assumptions that we cover the majority of the Korean population. But applying a partial coverage in the estimation will result in increased incidence of neurodevelopment disorder patients and unintentionally inflate the carrier frequency. Therefore, the assumption ensures conservative estimation. Next, we sought to understand the properties of pathogenic recessive variants as compared to the variants found from gnomAD on the same 69 genes that contain these variants. As expected, KND recessive variants were found less frequently ($P = 4.2 \times 10^{-10}$, Mann-Whitney U test; Fig. 3b), were strongly conserved during evolution ($P = 3.0 \times 10^{-5}$ for PhyloP and $P = 3.2 \times 10^{-7}$ for amino acid conservation, Mann-Whitney U tests; Fig. 3c), and displayed stronger functionality scores (CADD $P = 2.5 \times 10^{-5}$, GERP $P = 4.0 \times 10^{-3}$ and SIFT $P = 3.2 \times 10^{-6}$, Mann-Whitney U tests; Fig. 3d) compared to all gnomAD variants found in the same genes. To test the feasibility of accurate pre-screening rare pathogenic variants in a healthy population, we considered gnomAD-originated heterozygous LoF and ClinVar variants found in our recessive genes as a first-tier culprit for pathogenic recessive variants among many variants of obscure functional significances. And we observed that the portion that is attributable to LoF and ClinVar variants by healthy carriers was variable among the genes, and this portion is correlated with the o/e LoF value (Pearson's correlation $r = 0.33$; Fig. 3e).

Discussion

This study demonstrates the clinical utility of applying WES to children with various and complex neurodevelopmental disorders. We identified genetic causes in 47.4% of the patients and evaluated the characteristics of the variants that caused the disorders in a recessive manner.

Consistent with previous studies, we were able to diagnose approximately half of the KND patients (Fig. 1c)[10–12,14,16]. Our pathogenic genes and variants were in good correlation with OMIM inheritance pattern and ACMG guideline (Fig. 2b,e). The novel genes formed significantly robust co-expression networks during neurodevelopment processes, which was most prominent in frontal cortex regions (Fig. 1f,g). There was no significant difference in the number of recessive calls between patients and controls, even after stratifying the calls into disease-related gene sets (Fig. 2a, Supplementary Fig. S6). This result suggests following: (i) although categorized as "neurodevelopmental", our patient set is heterogeneous in nature, diluting contribution of single functional entity, (ii) patients carry more non-pathogenic or non-functional recessive calls than expected. To surmount these issues, we need more power to extract biologically relevant signals. The pathogenic recessive variants that cause rare neurodevelopmental disorders displayed moderately increased allele frequency values and marginally increased evolutionary conservation as compared to dominant variants, suggesting that the qualitative differences between these two groups of variants were not dramatically different (Fig. 2c,d). Compared to the relatively mild differences in variant characteristics, the genes that caused such disorders either in recessive or dominant manner displayed more discernible differences based on several parameters. First, the recessive genes harbor increased o/e LoF values, implying that there is no constraint applied to the recessive genes, whereas dominant genes in displayed a highly biased pattern toward known haploinsufficiency genes (Fig. 2f). Similarly, the distribution of the relative locations of LoF variants in genes suggested that recessive genes were less constrained compared to dominant genes (Fig. 2h). Gene ontology (GO) analysis further supports that – while the two groups are predominantly composed of neurodevelopment-related genes – recessive genes contain an enriched number of genes that are involved in lipid metabolism and mitochondrial function (Fig. 2g). This is a plausible result considering that these pathways are essential for normal brain development[31,32]. To summarize, these observations suggest that gene properties are a stronger determinant of whether a disease adopts a recessive or dominant inheritance pattern than variant properties.

Predicting and avoiding the occurrence of recessive disorders is critical. Carrier estimation has been traditionally performed mostly for single-gene diseases such as β-thalassaemia, Tay-Sachs disease and cystic fibrosis, and efficiently reduced the incidence of these patients[33–35]. However, even after introducing aggressive analysis of genetic disorders using whole exome or whole genome sequencing (WGS), precise estimation of the incidence and contribution of rare Mendelian disorders in a recessive manner remains variable for study populations and disorders[2]. For example, analysis of the Deciphering Developmental Disorders (DDD) data revealed a small contribution of recessive disorders (3.6%) to patients of European ancestry, whereas this value was 30.9% for patients of Pakistani ancestry[4]. Systematic analysis of schizophrenia data did not detect a substantial contribution of recessive variants[5,6]. These observations differ from ours, where 35.1% of the definitely diagnosed patients emerged in a recessive mode (Fig. 1d), which is in good agreement with previous diagnostic WES studies[10,36,37]. Notably, a recent study using a large autism cohort revealed that a substantial portion of the patients are attributable to the recessive LoF variants[38].

Since the majority of these pathogenic recessive variants were inherited from healthy parents, and ethnic Koreans comprise a relatively isolated population with a centralized medical system, it was feasible to derive an estimate that one out of 17 individuals are healthy carriers of pathogenic recessive variants for severe neurodevelopmental disorders (Fig. 3a). Accumulating this estimated carrier portion across different rare severe disease entities will certainly increase this ratio. The contribution of known LoF and ClinVar variants varies by genes and is positively correlated with o/e LoF values (Fig. 3e), and pathogenic recessive variants display systematic differences that differentiate them from gnomAD variants (Fig. 3b–d). Thus, it would be feasible to predict potential rare recessive variants from genomic data of healthy parents with the help of large patient and control genomic data in the near future.

Our approach expanded the phenotypic spectrum of known genes (39 cases, 7.1%), and suggested novel genes that may allow us to better understand neurodevelopmental disease mechanisms (56 patients, 10.1%). Nevertheless, 42.5% of the cases (235/533) remained undiagnosed even after our WES effort, suggesting a substantial opportunity for further improvement (Fig. 1c). Related to this, a systematic re-analysis effort with additional bioinformatics pipelines increased the diagnostic rate by 4.2%[39]. Also, searching for functional non-coding variants through WGS and evaluating multiple rare functional variants that may increase disease predisposition may be beneficial[40], although a recent meta-analysis study claimed only a minimal improvement in the WGS diagnostic rate, presumably due to our limited understanding of the function of noncoding variant[41]. An alternative approach would be to integrate genome data with transcriptome data in order to identify functionally cryptic variants that directly influence expression of critical genes[42,43], although preparing patient-derived tissue still remains as a practical challenge.

Our study also addresses the clinical challenges of an evolving phenotype over time in growing children and how this can be overcome, which facilitates identification of treatable or actionable cases (Table 2, Supplementary Information: Notable vignettes and Supplementary Fig. S8). Our patient cohort included a successful drug repositioning case for a rare neurogenetic disease[44] (Table 2). All of these cases are expected to increase as more genotype-phenotype relationships are discovered and more drugs become available. This study demonstrates that applying WES and subsequent in-depth analysis provides clinical and practical benefits to existing patients and their families and reducing emergence of such patients. Finally, we demonstrated the successful establishment of this approach in Korea, and the necessity of this approach for patients with various undiagnosed neurodevelopmental disorders in countries of similar status.

Methods

Subjects. Blood samples were obtained from enrolled patients and their parents, who provided informed consent. WES was performed on 553 patients who visited the SNUCH pediatric neurology clinic from July 2014 to January 2019 and displayed various neurodevelopmental problems of unknown origins, such as demyelinating or hypomyelinating leukodystrophy, hereditary spastic paraplegia, mitochondrial disorders, epileptic

Initial clinical problem	Causal gene	Modified clinical interpretation (MIM number)	Significance of WES-based patient evaluation (treatment)	References
Developmental regression with Rett syndrome-like phenotype	*ST3GAL5*	Salt and pepper developmental regression syndrome (#609056)	Identified the molecular defect and established an accurate diagnosis	50,51
Hypotonia and motor delay followed by lower extremity weakness	*DYNC1H1*	Spinal muscular atrophy, lower extremity-predominant 1, AD (#158600)	Diagnosed a case with pleiotropic and evolving symptoms	52
Early onset hypotonia, sacral mass, congenital heart disease, and facial dysmorphism	*ASAH1*	Farber lipogranulomatosis (#228000)	Corrected a misdiagnosis	53
Ataxia followed by generalized dystonia	*ANO3*	Expanded spectrum of dystonia 24 (#615034)	Suggested a treatment strategy that resulted in gradual improvement within one year (deep brain stimulation)	54
Focal lower leg dystonia, dystonic gait	*SLC2A1*	GLUT1 deficiency syndrome 2 (#612126)	Identified disease-specific treatment that resulted in near-elimination of dystonia (ketogenic diet)	55
Leigh syndrome	*SLC19A3*	Thiamine metabolism dysfunction syndrome 2 (#606152)	Identified disease-specific treatment that resulted in clinical improvements in dystonia, spasticity, and cognitive function (supplements of thiamine and biotin)	56
Recurrent infections, telangiectatic skin mottling, and brain infarctions	*TMEM173*	STING-associated vasculopathy, infantile-onset (#615934)	Provided a rationale for a new treatment strategy that improved the skin lesions (tofacitinib treatment)	44
Severe global developmental delay, seizures, and acanthotic skin lesions	*RAB11B*	Neurodevelopmental disorder with ataxic gait, absent speech, and decreased cortical white matter (#617807)	Identified a new disease gene leading to a neurodevelopmental syndrome	57

Table 2. Notable cases where WES-based analysis conferred correct diagnoses or changed medical treatment strategies.

encephalopathy, Rett syndrome-like encephalopathy, ataxia, neuropathy, myopathies, and multiple congenital anomalies/dysmorphic syndromes with developmental problems (Table 1). The patients can be categorized into two groups: (i) clinically diagnosable but with substantial genetic heterogeneity (270/553, 48.8%) or (ii) heterogeneous and nonspecific clinical presentations without definite diagnosis (283/553, 51.2%; Supplementary Fig. S1). Prior to the WES analysis, thorough clinical and laboratory evaluations and extensive patient examinations have been conducted to identify possible genetic causes. These included genetic tests with candidate gene sequencing, targeted gene sequencing panel, trinucleotide repeat analysis, microarray, metabolic work-up, brain/spine MRI, or muscle biopsy if applicable. All subjects were evaluated by three pediatric neurologists, two pediatric neuroradiologists, and a pathologist.

Whole exome sequencing. WES was performed at Theragen Etex Bio Institute (Suwon, Korea) following the standard protocol and the data were analyzed as described previously[18]. Depending on the genetic analysis result, each patient was categorized as one of the following: category 1: known disease-causing genes were found; category 2: causative gene for other diseases were found; category 3: potentially pathogenic gene, but without prior disease association, was found; category 4: no disease-causing candidates were found; and category 5: known pathogenic copy number variation (CNV) was found.

Our variant assessment procedures were as follows: firstly, patient-specific CNVs were checked and samples with CNVs were classified as category 5. Then, patient-specific nucleotide variants such as *de novo*, compound heterozygous (CH; autosomal), and rare homozygous (RHo; autosomal) and hemizygous (RHe; X-linked) variants were selected by comparing against parents and prioritized based on the inheritance pattern (Fig. 1a). Variants with low in-house quality score (<60) or low coverage depths (<10) were excluded. Variants in intron regions and pseudogene were also filtered out.

In correlating the patient's clinical features with genetic variants, if patients carried a known pathogenic variant in OMIM or ClinVar, they were categorized as category 1 or 2, depending on similarity with reported and observed clinical manifestation. For previously unreported variants, if they were never seen in normal individuals (Genome Aggregation Database (gnomAD)[29], Korean Variant Archive (KOVA)[45] and in-house variants), harbored LoF, or were evolutionarily well-conserved at the amino acid level, they were categorized as potentially pathogenic. ACMG codes for the variants were annotated using a script provided by InterVar[46]. For CNV calling, the normalized coverage depths of each captured intervals were compared to the depths from related individuals.

Human brain transcriptome data. The BrainSpan transcriptome database (http://www.brainspan.org) was used to build developing human brain networks[47]. Data from eight post-conceptual weeks to 40 years of age were analyzed. A total of 385 samples were used for the analysis after combining the multiplicates by taking mean values. Probes with TPM (transcripts per million) >5 in at least one sample were used, yielding 23,943 probes as "brain-expressed transcripts".

Brain transcriptome network analysis. Using the above brain-expressed transcripts, we created eight known gene co-expression networks by selecting genes that are highly correlated to our set of genes ($n=164$; Pearson's correlation $r>0.7$), in which their disease associations were previously reported, from each developmental period (Fig. 1f). Then, we asked whether our novel genes can be successfully integrated into the known gene co-expression network. We randomly selected 53 genes (equal to the number of our novel genes) in brain-expressed transcripts and counted how many edges they formed with the known genes. The 10^5 random gene selections were performed and the number of edges with a known gene was used to construct a distribution. The number of edges from our observed novel genes was evaluated against the random distributions. *P*-values were calculated using z-score, assuming normal distrubutions.

Recessive variant analysis. Variants were first filtered by gnomAD allele frequency of 0.001 in a heterozygous status and ability to alter protein sequences. Then CH variants were called on a trio setting. If a gene contains more than one filtered variant and each variant was inherited from mother and father separately (for proband), or at least one but not all of the filtered variants of a gene were found from the progeny (for parent), the variants were called as CH. RHo variants were called if filtered variants are inherited in a homozygous manner in autosomes and never seen in gnomAD as homozygous. RHe variants were called if filtered variants are in the X chromosome and never seen in gnomAD as hemizygous or homozygous. Functionality scores were extracted from dbNSFP[48].

Statistical evaluation. Statistical analysis of this study was conducted using the R package (version 3.6.0). When comparing difference of two groups, normality of the distribution was first evaluated by Shapiro-Wilks test and then either Student's t-test or a non-parametric Mann-Whitney U test was performed. *P*-value < 0.05 was considered as statistically significant, and adjusted FDR *P*-values were used when correcting for multiple tests. Correlation between variables was determined by using the Pearson's correlation coefficient test.

Ethical approval and informed consent. This study was approved by the Seoul National University Hospital Institutional Review Board (No. 1406-081-588), and was performed in accordance with the relevant guidelines and regulations. Written informed consent was obtained from all enrolled patients or their legal representatives.

Data availability

Anonymized data not published within this article will be made available by request from any qualified investigator for purposes of replicating procedures and results.

Received: 29 July 2019; Accepted: 10 January 2020;
Published online: 29 January 2020

References

1. Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: Discovery to translation. *Nature Reviews Genetics* **14**, 681–691 (2013).
2. Antonarakis, S. E. Carrier screening for recessive disorders. *Nat. Rev. Genet.* **20**, 549–561, https://doi.org/10.1038/s41576-019-0134-2 (2019).
3. Jin, S. C. *et al*. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017).
4. Martin, H. C. *et al*. Quantifying the contribution of recessive coding variation to developmental disorders. *Science* **362**, 1161–1164 (2018).
5. Rees, E. *et al*. Analysis of exome sequence in 604 trios for recessive genotypes in schizophrenia. *Transl. Psychiatry* **5**, e607 (2015).
6. Ruderfer, D. M. *et al*. No evidence for rare recessive and compound heterozygous disruptive variants in schizophrenia. *Eur. J. Hum. Genet.* **23**, 555–557 (2015).
7. Monies, D. *et al*. Lessons Learned from Large-Scale, First-Tier Clinical Exome Sequencing in a Highly Consanguineous Population. *Am. J. Hum. Genet.* **104,** 1182–1201, https://doi.org/10.1016/j.ajhg.2019.04.011 (2019).
8. Alkuraya, F. S. Discovery of mutations for Mendelian disorders. *Hum. Genet.* **135**, 615–623 (2016).
9. Lifton, R. P., Gharavi, A. G., Geller, D. S. & Hughes, H. Molecular Mechanisms Review of Human Hypertension., **104**, 545–556 (2001).
10. Srivastava, S. *et al*. Clinical whole exome sequencing in child neurology practice. *Ann. Neurol.* **76**, 473–483 (2014).
11. Trujillano, D. *et al*. Clinical exome sequencing: results from 2819 samples reflecting 1000 families. *Eur. J. Hum. Genet.* **25**, 176–182 (2016).
12. Yang, Y. *et al*. Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing. *Jama* **312**, 1870–1879 (2014).
13. Baynam, G. *et al*. The rare and undiagnosed diseases diagnostic service - application of massively parallel sequencing in a state-wide clinical service. *Orphanet J. Rare Dis.* **11**, 77 (2016).
14. Beaulieu, C. L. *et al*. FORGE Canada consortium: Outcomes of a 2-year national rare-disease gene-discovery project. *Am. J. Hum. Genet.* **94**, 809–817 (2014).
15. Gahl, W. A. *et al*. The NIH Undiagnosed Diseases Program and Network: Applications to modern medicine. *Mol. Genet. Metab.* **117**, 393–400 (2016).
16. McRae, J. F. *et al*. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
17. Cho, A. *et al*. Consecutive analysis of mutation spectrum in the dystrophin gene of 507 Korean boys with Duchenne/Becker muscular dystrophy in a single center. *Muscle and Nerve* **55**, 727–734 (2017).
18. Yoo, Y. *et al*. GABBR2 mutations determine phenotype in rett syndrome and epileptic encephalopathy. *Ann. Neurol.* **82**, 466–478 (2017).
19. Deuchler, M. *The Confucian Transformation of Korea: A Study of Society and Ideology*, https://doi.org/10.2307/j.ctt1dnn8zj (Harvard University Asia Center, Harvard University, 1992).
20. Sanders, S. J. *et al*. Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. *Neuron* **70**, 863–885 (2011).
21. Uddin, M. *et al*. Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder. *Nat. Genet.* **46**, 742–747 (2014).
22. Pfundt, R. *et al*. Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genet. Med.* **19**, 667–675 (2017).
23. Feliciano, P. *et al*. Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. *npj Genomic Med.* **4**, 19 (2019).

24. Cohn, D. H., Byers, P. H., Steinmann, B. & Gelinas, R. E. Lethal osteogenesis imperfecta resulting from a single nucleotide change in one human proal(I) collagen allele (mutation/gene cloning/DNA sequence analysis/extracellular matrix). *Genetics* **83**, 6045–6047 (1986).
25. Lee, J. S. *et al*. Severe hypotonia and postnatal growth impairment in a girl with a missense mutation in COL1A1: Implication of expanded phenotypic spectrum of type I collagenopathy. *Brain Dev.* **39**, 799–803 (2017).
26. Lin, G. N. *et al*. Spatiotemporal 16p11.2 Protein Network Implicates Cortical Late Mid-Fetal Brain Development and KCTD13-Cul3-RhoA Pathway in Psychiatric Diseases. *Neuron* **85**, 742–754 (2015).
27. Ferdinandusse, S. *et al*. Clinical, biochemical, and mutational spectrum of peroxisomal acyl-coenzyme A oxidase deficiency. *Hum. Mutat.* **28**, 904–912 (2007).
28. Gregory, A. *et al*. Autosomal dominant mitochondrial membrane protein-associated neurodegeneration (MPAN). *Mol. Genet. Genomic Med.* e736, https://doi.org/10.1002/mgg3.736 (2019).
29. Karczewski, K. J. *et al*. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*, 531210, https://doi.org/10.1101/531210 (2019).
30. Statistics Korea. Final Results of Birth Statistics in 2016. Available at: http://kostat.go.kr/portal/eng/pressReleases/8/1/index.board?bmode=read&bSeq=&aSeq=363582&pageNo=1&rowNum=10&navCount=10&currPg=&searchInfo=srch&sTarget=title&sTxt=birth (2016).
31. McKenna, M. C., Schuck, P. F. & Ferreira, G. C. Fundamentals of CNS energy metabolism and alterations in lysosomal storage diseases. *J. Neurochem.* **148**, 590–599 (2019).
32. Tracey, T. J., Steyn, F. J., Wolvetang, E. J. & Ngo, S. T. Neuronal Lipid Metabolism: Multiple Pathways Driving Functional Outcomes in Health and Disease. *Front. Mol. Neurosci.* **11**, 1–25 (2018).
33. Cao, A. & Kan, Y. W. The prevention of thalassemia. *Cold Spring Harb. Perspect. Med.* **3**, 1–15 (2013).
34. Kaback, M. *et al*. Tay-sachs disease—carrier screening, prenatal diagnosis and the molecular era: An international perspective, 1970 to 1993. *Obstet. Gynecol. Surv.* **49**, 321–322 (1994).
35. Watson, M. S. *et al*. Cystic fibrosis population carrier screening: 2004 revision of American College of Medical Genetics mutation panel. *Genet. Med.* **6**, 387–91 (2004).
36. Evers, C. *et al*. Impact of clinical exomes in neurodevelopmental and neurometabolic disorders. *Mol. Genet. Metab.* **121**, 297–307 (2017).
37. Kuperberg, M. *et al*. Utility of Whole Exome Sequencing for Genetic Diagnosis of Previously Undiagnosed Pediatric Neurology Patients. *J. Child Neurol.* **31**, 1534–1539 (2016).
38. Doan, R. N. *et al*. Recessive gene disruptions in autism spectrum disorder. *Nat. Genet*. **51**, 1092–1098, https://doi.org/10.1038/s41588-019-0433-8 (2019).
39. Bergant, G. *et al*. Comprehensive use of extended exome analysis improves diagnostic yield in rare disease: A retrospective survey in 1,059 cases. *Genet. Med.* **20**, 303–312 (2018).
40. Werling, D. M. *et al*. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* **50**, 727–736 (2018).
41. Clark, M. M. *et al*. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *npj Genomic Med.* **3**, 16 (2018).
42. Cummings, B. B. *et al*. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**(386), https://doi.org/10.1126/scitranslmed.aal5209 (2017).
43. Gonorazky, H. D. *et al*. Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *Am. J. Hum. Genet.* **104**, 466–483 (2019).
44. Seo, J. *et al*. Tofacitinib relieves symptoms of stimulator of interferon genes (STING)–associated vasculopathy with onset in infancy caused by 2 de novo variants in TMEM173. *J. Allergy Clin. Immunol.* **139**, 1396–1399.e12 (2017).
45. Lee, S. *et al*. Korean Variant Archive (KOVA): A reference database of genetic variations in the Korean population. *Sci. Rep.* **7**, 4287 (2017).
46. Li, Q. & Wang, K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am. J. Hum. Genet.* **100**, 267–280 (2017).
47. Kang, H. J. *et al*. Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–9 (2011).
48. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* **32**, 894–899 (2011).
49. Rehm, H. L. *et al*. ClinGen — The Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
50. Boccuto, L. *et al*. A mutation in a ganglioside biosynthetic enzyme, ST3GAL5, results in salt & pepper syndrome, a neurocutaneous disorder with altered glycolipid and glycoprotein glycosylation. *Hum. Mol. Genet.* **23**, 418–433 (2014).
51. Lee, J. S. *et al*. GM3 synthase deficiency due to ST3GAL5 variants in two Korean female siblings: Masquerading as Rett syndrome-like phenotype. *Am. J. Med. Genet. Part A* **170**, 2200–2205 (2016).
52. Strickland, A. V. *et al*. Mutation screen reveals novel variants and expands the phenotypes associated with DYNC1H1. *J. Neurol.* **262**, 2124–2134 (2015).
53. Kim, S. Y. *et al*. Atypical presentation of infantile-onset farber disease with novel ASAH1 mutations. *Am. J. Med. Genet. Part A* **170**, 3023–3027 (2016).
54. Yoo, D. *et al*. Early-onset generalized dystonia starting in the lower extremities in a patient with a novel ANO3 variant. *Parkinsonism and Related Disorders* **50**, 124–125, https://doi.org/10.1016/j.parkreldis.2018.02.012 (2018).
55. Hully, M. *et al*. From splitting GLUT1 deficiency syndromes to overlapping phenotypes. *Eur. J. Med. Genet.* **58**, 443–454 (2015).
56. Alfadhel, M. & Tabarki, B. SLC19A3 Gene Defects Sorting the Phenotype and Acronyms: Review. *Neuropediatrics* **49**(02), 083–092, https://doi.org/10.1055/s-0037-1607191 (2018).
57. Lamers, I. J. C. *et al*. Recurrent De Novo Mutations Disturbing the GTP/GDP Binding Pocket of RAB11B Cause Intellectual Disability and a Distinctive Brain Phenotype. *Am. J. Hum. Genet.* **101**, 824–832 (2017).

Acknowledgements

This study was funded by grants from the Ministry of Health and Welfare (HI16C1986), the Korea Centers for Disease Control and Prevention (20170607DBF-00) and the Korean National Research Foundation (2018M3C9A5064708). The authors gratefully acknowledge patients and parents participated in this study.

Author contributions

M.C. and J.H.C. are responsible for study concept and design, supervised the study and obtained funding. Y.L., S.P., J.C., Y.Y., S.L., T.Y., M.L., J.S., Jeongeun L., J.K., E.Y.J., E.K., and M.C. analyzed genome data. J.S.L., S.Y.K., J.C., H.K., W.J.K., J.S.K., J.M.K., A.C., B.C.L., W.S.K., and J.H.C. provided clinical data. Y.L., S.P., J.S.L., S.Y.K., J.C., M.C., and J.H.C. combined genetic and clinical data. Y.L., S.P., J.C., Jeongeun L., J.K., Jean L., H.J., E.Y.J., S.E.H., and M.C. performed genetic and statistical evaluation of the cohort. Y.L., S.P., J.S.L., S.Y.K., M.C., and J.H.C. drafted the manuscript. All authors reviewed the manuscript for important intellectual content.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-58101-8.

Correspondence and requests for materials should be addressed to M.C. or J.-H.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.