Genome Sequencing and Comprehensive Rare Variant Analysis of 465 Families with Neurodevelopmental Disorders

Authors

Alba Sanchis-Juan[1,2], Karyn Megy[1,3], Jonathan Stephens[1], Camila Armirola Ricaurte[1], Eleanor Dewhurst[1], Kayyi Low[1], Courtney E French[4], Detelina Grozeva[5,6], Kathleen Stirrups[1], Marie Erwood[1], Amy McTague[7], Christopher J Penkett[1^], Olga Shamardina[1], Salih Tuna[1], Louise C. Daugherty[1], Nicholas Gleadall[1], Sofia T Duarte[8], Antonio Hedrera-Fernández[9], Julie Vogt[10], Gautam Ambegaonkar[11], Manali Chitre[4], Dragana Josifova[12], Manju A Kurian[7], Alasdair Parker[4,11], Julia Rankin[13], Evan Reid[14], Emma Wakeling[15], Evangeline Wassmer[16], C Geoffrey Woods[4,5], NIHR-BioResource, F Lucy Raymond[1,5,*], Keren J Carss[1,3,*]

Affiliations

1. Department of Haematology, University of Cambridge, UK; NIHR BioResource, Cambridge University Hospitals NHS Foundation Trust, UK
2. Molecular Neurogenetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA; Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA; Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
3. Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK
4. Clinical Medical School, University of Cambridge, UK
5. Department of Medical Genetics, University of Cambridge, UK
6. Centre for Trials Research, Cardiff University, Cardiff, UK
7. Molecular Neurosciences, Zayed Centre for Research into Rare Disease in Children, UCL Great Ormond Street Institute of Child Health, London, UK; Department of

Neurology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK

8. Hospital Dona Estefânia, Centro Hospitalar de Lisboa Central, Lisbon, Portugal
9. Pediatric Neurology Department, Hospital Universitario Central de Asturias, Spain
10. West Midlands Regional Genetics Service, Birmingham Women's and Children's Hospital, Birmingham, UK
11. Child Development Centre, Cambridge University Hospitals NHS Foundation Trust, UK
12. Guy's and St Thomas' Hospital, London, UK
13. Department of Clinical Genetics, Royal Devon University Healthcare NHS Foundation Trust, Exeter, UK
14. Cambridge Institute for Medical Research and Department of Medical Genetics, University of Cambridge, UK
15. North West Thames Regional Genetics Service, Harrow, UK
16. Neurology Department, Birmingham Women and Children's Hospital, Birmingham, UK

* These authors contributed equally to this work

^Deceased

Corresponding authors:

F Lucy Raymond: flr24@cam.ac.uk

Keren J Carss: keren.carss@astrazeneca.com

Abstract

Despite significant progress in unravelling the genetic causes of neurodevelopmental disorders (NDDs), a substantial proportion of individuals with NDDs remain without a genetic diagnosis following microarray and/or exome sequencing. Here we aimed to assess the power of short-read genome sequencing (GS), complemented with long-read GS, to identify causal variants in participants with NDD from the NIHR BioResource project. Short-read GS was conducted on 692 individuals (489 affected and 203 unaffected relatives) from 465 families. Additionally, long-read GS was performed on five affected individuals who had structural variants (SVs) in technically challenging regions, complex SVs, or required distal variant phasing. Causal variants were identified in 36% affected individuals (177/489) and a further 23% (112/489) had a variant of uncertain significance, after multiple rounds of re-analysis. Among all reported variants, 88% (333/380) were SNVs/indels, and the remainder were SVs, non-coding, and mitochondrial variants. Furthermore, long-read GS facilitated resolution of challenging SVs and invalidated variants of difficult interpretation from short-read GS. This study demonstrates the value of short-read GS, complemented with long-reads, to investigate the genetic causes of NDDs. GS provides a comprehensive and unbiased method to identify all types of variants throughout the nuclear and mitochondrial genome in individuals with NDD.

Introduction

Neurodevelopmental disorders (NDDs) encompass a range of conditions that usually present in childhood, including intellectual disability, developmental delay, autism spectrum disorder, epilepsy, and movement disorders amongst others. While individually rare, collectively NDDs affect millions of people worldwide and present huge challenges for families and healthcare systems.[1]

These disorders are phenotypically and genetically heterogeneous, and are often caused by rare, highly penetrant variants. Over the last decade exome sequencing (ES), and increasingly genome sequencing (GS), have been widely adopted for the identification of NDD-associated pathogenic (P) and likely pathogenic (LP) variants (collectively referred as causal variants throughout this manuscript) in more than 900 NDD-associated genes identified to date.[2,3] For families with affected children, receiving a genetic diagnosis has many benefits. It often marks the end of a long diagnostic odyssey, can affect clinical management, and allows parents to make more informed subsequent reproductive choices.[1,3] The proportion of affected individuals in whom a causal variant is identified following genetic testing is known as the diagnostic yield, and it varies according to many factors. For example, in a recent meta-analysis the range of diagnostic yield in studies using ES or GS in children with suspected genetic diseases was 24–68%.[4]

Causal variants are most commonly coding single nucleotide variants (SNVs), small insertions and deletions (indels), and large copy number variants (CNVs).[5] Additional classes of genetic variation that can cause NDDs include small CNVs (below the resolution of chromosomal microarrays), inversions, translocations, complex structural variants (cxSVs), short tandem repeats (STRs), and variants in the mitochondrial (MT) genome.[1] Some of these classes of variation are still challenging to detect using short-read sequencing technologies, causing an increasing appreciation of the potential role of long-read sequencing.[6]

The NIHR BioResource has conducted a flagship study whereby short-read GS (srGS) was performed on 13,037 individuals to study the genetic basis of rare disorders including NDDs in the UK national healthcare system.[7] In this study we have performed detail investigation of the 692 individuals with NDD from the NIHR BioResource cohort, which aims were threefold: 1) to identify a comprehensive range of causal variants using srGS, including those that are often neglected by other methods; 2) to use supplementary long-read GS (lrGS) on a subset to help resolve and interpret variants that were unclear from srGS; 3) to contribute towards the identification of new associations between genes and NDDs. This study has been notably successful at achieving all three of these aims. We have contributed towards the identification or confirmation of four NDD-associated genes: *KMT2B*, *CACNA1E*, *WASF1*, and *GABRA2*, which have been published elsewhere.[8-11] In this article we focus on the first two aims and describe in detail the overall structure and results of the NIHR BioResource NDD study.

Materials and Methods

Cohort description

The NDD sub-cohort of the NIHR BioResource Project[7] comprises 692 individuals, of whom 489 were affected with a NDD, and 203 were unaffected relatives. All participants provided written informed consent to participate in the study. The study was approved by the East of England Cambridge South national institutional review board (13/EE/0325). The research conforms with the principles of the Declaration of Helsinki. Written informed consent to participate was obtained to publish clinical information.

These individuals belonged to 465 families. Our inclusion criteria required evaluation by a tertiary level pediatric neurologist who suspected a Mendelian disorder where the differential diagnosis included genes that had not been previously tested (see Supplemental Methods for full details). 73% (357/489) of the participants had either intellectual disability, developmental delay, autism spectrum disorder, movement disorder/dystonia and/or seizures (Figure S1). Recruitment of family members into this study varied depending on availability and suspected mode of inheritance. We sequenced 335 singletons (affected proband only), 67 trios (affected

proband and both parents), five quads (affected proband, both parents and a sibling) and 58 families with another family structure combination (Table 1). Most individuals had undergone routine genetic testing without identification of a candidate variant prior enrolment to this project, resulting in an enrichment for challenging cases.

Short-read GS and identification of causal variants

DNA samples from whole blood underwent short-read GS. Alignment to the human genome of reference GRCh37 and variant calling were performed to identify multiple types of variants including SNVs, indels, structural variants (SVs) and STRs (Figure S2), as described in the supplementary methods and a previous publication.[7] Mobile element insertions (MEIs), Spinal Muscular Atrophy (SMA) status and Regions Of Homozygosity (ROHs) were also characterized.

Candidate rare variants were restricted to known NDD-associated genes (see next section) and discussed in multidisciplinary team meetings (MDTs), which included research bioinformatics analysts, clinical scientists and clinical geneticists. Additional information on the variant annotation and filtering strategies is in the supplemental methods. Pathogenicity was determined according to the American College of Medical Genetics guidelines (ACMG).[12] Variants that were reported to the affected individual's referring clinicians (also defined in this manuscript as reportable variants) comprise causal variants (P/LP) and variants of uncertain significance (VUS) which could potentially explain the phenotype, at the discretion of the MDT. Variants in genes of uncertain association with specific phenotypes were considered for research, further analysis and sharing through Gene Matcher.[13]

Gene list curation and variant reanalysis

A list of NDD-associated genes was assembled from various sources including OMIM (https://omim.org), PanelApp[14] (which also comprises DDG2P[15]) and PubMed searches, then

curated to ensure they comply with previously described criteria.[15] The gene list was updated six times throughout the timeline of the project and the last version contained 1,545 genes (Table S1).

Initially, affected individuals were investigated using the gene list available at the time of analysis. Then, reanalysis of all individuals was performed twice (July 2018 and July 2019) using revised quality control and filtering thresholds as well as an updated version of the gene list at the time (v.20180117 and v.20180807 respectively). Re-analysis consisted of manual assessment of 1) rare variants in NDD-associated genes that had been added to the gene list since the first analysis, 2) variants reclassified as P/LP in HGMD[16] or ClinVar[17] since the first analysis and 3) loss-of-function (LOF) SNVs/indels or predicted to be damaging (CADD phred > 20) in NDD-associated genes but with quality metrics below the strict filters employed for the initial analysis. Candidate variants identified by the last approach were manually inspected in IGV (v2.5)[18] and recommended for Sanger sequencing confirmation if they were suspected to be real.

Trio analysis

In families where both parents were available (67 trios and 5 quads), joint calling using Platypus variant caller[19] was also run with default parameters. Then, variants from both algorithms (Platypus and Isaac Variant Caller) were merged, and a gene agnostic identification of candidate variants by mode of inheritance was performed using *in house* filtering scripts described elsewhere.[20] Variants were interpreted and reported in NDD-associated genes as described above.

Long-read GS

Long-read GS was done with Oxford Nanopore Technologies (ONT), using the GridION platform for one individual (three runs) and the PromethION platform for four individuals (four

runs). Samples were prepared and sequenced as previously described.[21] Reads were aligned against the GRCh37 human reference genome and sensitive detection of SVs was performed using and ensemble algorithm approach as previously described.[21] Additional information on the lrGS methods, algorithms and versions can be found in the supplemental material (supplemental methods section). Identification of candidate SVs was performed at the locus of interest, and manual inspection of the alignments was also performed using IGV.[18]

Results

Diagnostic yield in this NDD cohort achieves 36%

Affected individuals presented with a wide range of NDD phenotypes, and the most frequent were intellectual disability (n=199), seizures (n=191), movement disorders (n=78), dystonia (n=68) and ataxia (n=41), with many individuals having more than one phenotype (Figure S1). Reportable variants were identified in 59% (289/489) of affected individuals: 36% (177/489) had at least one P/LP variant, and a further 23% (112/489) had at least one VUS (Table 1).

The P/LP variant detection rate was affected by a series of factors. First, diagnostic yield was higher for trios (41%, 28/67) and pair of siblings (57%, 16/28) than probands only (35%, 119/335) (Figure 1A, Table 1). In four families, the reported variants were different amongst multiple affected individuals (Table S2), supporting previous observations that pathogenic shared variants within the same family should not be assumed.[22]

Additionally, diagnostic rate varied depending on genetic ancestry (Table S3), phenotype (Figure 1B) and mode of inheritance (Figure 1C). While 34% (111/325) of individuals of European ancestry had identified causal variants, only 3% (7/245) of the variants identified in that group were homozygous. The rate was higher in individuals of South-Asian ancestry, where 40% (29/72) of the variants were homozygous and 43% (35/82) of individuals had P/LP variants, which was consistent with previously reported results (Table S3).[23]

Furthermore, phenotypes with higher diagnostic rates include hypotonia (50%, 11/22, noting our cohort is enriched for severe hypotonia), microcephaly (49%, 19/39), cerebellum abnormalities (44%, 12/27) and autism spectrum disorder (43%, 10/23) while abnormality of growth (14%, 2/14) and hypermobility (14%, 1/7) were lower (Figure 1B). 108 individuals with reportable variants had more than one main phenotype/phenotypes that fall into more than one HPO category (e.g., Abnormality of the nervous system, Abnormality of the Eye, as shown in Figure S1a, flagged in Table S2 as 'Compounded_phenotype'), and ten of these had variants in multiple genes, each partially explaining the phenotype.

A wide variety of reportable genes and variants are identified in this cohort

The most frequently reported gene across families in the whole cohort was *GNAO1* [MIM: 139311] (*n*=7), followed by *CACNA1A* [MIM: 601011] (*n*=6), *KCNQ2* [MIM: 602235] (*n*=6), *STXBP1* [MIM: 602926] (n=6) and *SCN1A* [MIM: 182389] (*n*=6) (Table S4). In total we reported 380 variants (358 unique) in 289 individuals from 276 families. Eighteen variants were common between affected members of the same family, and four variants were present in individuals from different families. The majority of these were SNVs (74%, 279/380), indels (14%, 54/380) and deletions (8%, 31/380). Although duplications, insertions, complex SVs and ROH were found in a lower frequency, in total they accounted for 4% (16/380) of the reported variants. (Table 2). Although mosaic variants were not systematically called due to the coverage, five likely mosaic variants were identified in this cohort after evaluation of allelic balance and visual inspection of candidate variants in IGV: three were SNVs and two were SVs (Figure S3).

The proportion of variants reported as P/LP compared to VUS varied according to variant type. While this proportion was similar for SNVs, 83% of indels (45/54) and 74% of the reported deletions (23/31) were labelled as P/LP (Table 2). Duplications, large insertions and inversions

were all reported as VUS (*n*=11), reflecting the more challenging interpretation of variant effect. One ROH was identified in an individual with Angelman syndrome and deemed to be pathogenic. No STR expansions in known locus or SMA-associated variants were identified in this cohort, which was unsurprising since most of these individuals have previously had a negative routine genetic testing.

Re-analysis of the data increases diagnostic yield

The first round of variant analysis took place between March 2016 and January 2018. During this time the gene list was under active development, and probands were analyzed using the most recent gene list version available at the time. Reanalysis of the data was performed twice, considering updated variant annotations, quality filtering strategies, and NDD-associated genes. Reanalysis in July 2018 and July 2019 increased the number of reportable variants from 265 to 329 then to 380 respectively (Figure 1D), and it substantially increased affected individuals with reportable variants: from 42% (208/489) to 59% (289/489) after 18 months.

Reanalysis identified additional reportable variants due to a variety of reasons: most were in recently discovered NDD-associated genes (69%, 79/115) or were identified due to improvements in the pipeline (28%, 32/115), such as better transcription prioritization, inclusion of MEIs, ROH, or improved *de novo*/SV calling. For example, a variant in *PNPLA6* gene [MIM: 603197] (NM_001166114.2:c.2785C>T (p.Arg929Cys) in G008170) was flagged as low quality in the SNV/indel pipeline (minimum overall pass rate of 0.98%), but manual evaluation in IGV suggested it was real; a compound heterozygous variant in *BRAT1* [MIM: 614506] was reported in one individual after new publications revealed stronger phenotypic evidence; and one individual had a deep intronic variant in *TSC2* [MIM: 191092] that was identified after it was reported in ClinVar. Additionally, 3.5% (4/115) of variants were in genes following autosomal recessive mode of inheritance with a previously identified single event,

highlighting the importance for analyzing not only recently discovered disease-associated genes, but also previously known that may harbor missed clinically relevant variants.

GS detects classes of variants that may be missed by other technologies

Variants that are often challenging to detect by routine diagnostic technologies such as ES and chromosomal microarrays analysis (CMA) include SVs, rare intronic variants, and MT variants. Here we describe findings involving these types of variants in this cohort and we briefly describe ten participants to highlight the value of GS. Additional information for each participant and variant is present in the supplemental material and Table S2.

Regarding SVs, we reported a total of 31 deletions, six duplications, two inversions, three large insertions, four cxSV and one ROH. Importantly, 66% (31/47) of them were either smaller than standard CMA resolution (200 Kbp using Affymetrix Chromosome Analysis Suite) or not possible to be detected by CMA (e.g. inversions and insertions), underscoring the value of GS to detect SVs cryptic to this technology. Six SVs occur in conjunction with a SNV/indel in a known genes following autosomal recessive mode of inheritance. One example is Participant 1 (G013396 in Table S2), an individual with Early Infantile Epileptic Encephalopathy (EIEE) and a combination of an inversion and a missense variant in *SPATA5* [MIM: 613940], which is associated with an autosomal recessive neurodevelopmental disorder that often includes seizures (Figure S4). This example underscores the value of GS to investigate inversions, which are often neglected in genetic analyses.

Six intronic variants identified in this cohort were associated with NDDs: five splice region and one deep intronic variant (Table S2). The latter was in an individual with Tuberous Sclerosis who had endured a long diagnostic odyssey (Participant 2, G004131 in Table S2). A heterozygous deep intronic variant in *TSC2* [MIM: 191092] was identified in 17% (4/23) of the reads, suggesting mosaicism (Figure S3b), later confirmed by Sanger sequencing. This

variant was observed during reanalysis, after it was published and submitted to ClinVar as associated with disease.[24]

Lastly, four reportable variants were identified in MT genome genes, three of which were deemed to be LP. Variants were called at different levels of heteroplasmy (from 83-91%) and homoplasmy, which were estimated from coverage analyses. One example (Participant 3, G004703 in Table S2) is an individual with ataxia, recurrent lactic acidosis and myopathy. This individual had a missense variant in heteroplasmy (91% in blood), in *MT-TL1* gene (Figure S5). This is one of the most thoroughly studied and best characterized disease-causing MT variants, and is associated, amongst other phenotypes, with MELAS (myopathy, encephalopathy, lactic acidosis, and stroke like episodes),[25] which was consistent with the individual's phenotype. The other two LP variants (Participant 4 and 5, G013808 and G012198 in Table S2 respectively) were in the genes *MT-ATP6,* associated with neurogenic muscle weakness, ataxia and retinitis pigmentosa,[26] and *MT-ND4,* associated with Leber Hereditary Optic Neuropathy with or without additional neurological abnormalities,[27-29] respectively (Figure S6 and S7).

Long-read sequencing resolves complex SVs in two individuals

Five individuals with ambiguous results from srGS data were further investigated by ONT lrGS (Table 3). A total of seven runs (three in GridION for one sample and four in PromethION for the remainder) produced an average coverage of 14.6 (± 7.5) reads with an average length of 4,243 bp (± 4,054) (Figure S8A-D). After QC, 62,620 SVs were identified, an average of 26,311 ± 4,532 per individual (Figure S8E-F), which is consistent to previously reported lrGS studies.[30]

Two affected individuals carried complex SVs that were resolved by lrGS. Participant 6 (NGC00375_01 in Table S2), a male with dystonia, learning difficulties and behavioral problems, had a *de novo* complex SV disrupting *SGCE* [MIM: 604149], which is associated

with dystonia. Short-read GS had suggested this was part of a complex SV, but resolution could not be achieved due to homology at the breakpoints. Long-read GS allowed SV characterization and resolved the complex rearrangement that involved 37 breakpoints between chromosomes 7, 10 and 12 (Figure 2A). The variant was reported as LP.

Participant 7 (G012664 in Table S2) is a male with paroxysmal dyskinesia and bulbar palsy, who harbored a complex rearrangement characterized by the presence of duplications across multiple chromosomes, including chromosome X. The variant had been inherited from the unaffected mother and lrGS revealed 26 duplicated DNA fragments of 24Kb median size (sd ± 12Kb) from 14 different chromosomes (Figure 2B). Although no protein coding gene was predicted to be disrupted, we couldn't rule out the possible regulatory effect of this event, and it was classified as VUS.

Long-read sequencing phases variants and facilitates resolution of technically challenging regions in three individuals

LrGS was also used to perform variant phasing and to investigate SVs in technically challenging regions. Participant 8 (G013428 in Table S2) presented with global developmental delay, hypotonia with movement disorder, sensorineural hearing impairment, microcephaly and delayed visual maturation with esotropia. An inversion involving *CASK* [MIM: 300172] gene was called in the srGS data (Figure 2C), but the variant couldn't be confirmed by long-range PCR due to low sequence complexity. We therefore sought to validate it using lrGS, and the inversion was not supported by the lrGS data, suggesting that the called inversion was a false positive.

Participant 9 (G013407 in Table S2) was a female with EIEE and a heterozygous missense variant in *DNM1* gene [MIM: 616346], which is associated with epileptic encephalopathy. The variant was absent in the unaffected father, and maternal DNA was unavailable (Figure S9). Given that 80% of *de novo* variants occur in the paternal allele,[31] we performed lrGS to

determine the haplotype of the variant. Unfortunately, the closest informative SNV was 7,048 bp from this position and there were no reads of this length covering the region (average read length 6,723 bp ± 4,695 bp). Therefore, the variant was classified as VUS.

Lastly, Participant 10 (G000973 in Table S2) was a female with early onset dementia, spastic paraplegia and thin corpus callosum. Three deletions and two inversions were called in *KIF5C* [MIM: 604593]. LrGS was used to resolve this event and demonstrated that *KIF5C* had not been disrupted and the calls were from a retroelement insertion of a *KIF5C* transcript highly expressed in human brain (Figure 2D). Although the insertion was not affecting any protein coding gene, it was classified as VUS since reports have shown that retroelements can interfere with gene expression by other mechanisms such as silencing by transcriptional or RNA interference.[32]

Discussion

In this study we describe in detail the structure and outcomes of the NIHR BioResource NDD project. We employed a comprehensive approach that combined short and long-read GS to identify a broad range of clinically relevant variants associated with NDDs. This strategy identified a high rate of causal variants throughout the nuclear and mitochondrial genomes (36%), including variants often intractable to ES/CMA. Our diagnostic yield is within the expected range reported by similar studies,[3,4,33] and 3% higher than the 33% reported in a previous NIHR BioResource study due to reanalysis and follow up studies.[7] It is worth noting that the diagnostic yield for NDDs can vary considerably and is influenced by many factors, such as phenotype and recruitment criteria, sequencing technology, mode of inheritance, family members studied, date of analysis, and genetic ancestry. Understanding these factors can help inform recruitment strategies and study design to improve diagnostic yield. For example, we observed a slightly higher diagnostic yield for trios (41%) than singletons (35%). This is consistent with previous studies emphasizing the importance and value of trio

design.[3,33] However, recruitment of both biological parents is not always possible, and our relatively high yield in singletons support including them wherever possible.[34]

A notable strength of this study is how comprehensively we surveyed multiple types of variants that could be implicated with NDDs. We not only investigated coding SNVs and indels, but also explored SVs, intronic variants, STR expansions, SMA status and MT variants. However, we did not find any individual with pathogenic STRs or SMA cases, which could be due to several reasons: i) some participants may have undergone STR expansion/SMA testing prior to enrollment, resulting in a reduced likelihood of detecting such variants, ii) these are very rare causes of NDDs, and thus our study may have been underpowered to detect them, and iii) it is possible that these types of variants are identified with lower sensitivity than other classes, or they may be specifically implicated in phenotypes poorly represented within this study.

Interpretation of variants that are not SNVs or indels, such as SVs, can be particularly challenging, despite recent improvements on guidelines for interpretation of CNVs.[35] Pathogenic intronic and other 'non-coding' variants are rare and difficult to identify and interpret, especially without supporting transcriptomic data from an appropriate tissue.[7,36] Large-scale genome sequencing cohorts currently underway will help improve our understanding of the distribution, features, and function of non-coding variants, facilitating easier identification of those that are pathogenic.[3,7,37,38] Classes of variants that we were unable to investigate in this study include those in repetitive regions that are intractable to detection by srGS, as well as somatic or mosaic variants that generally require higher coverage sequencing for detection.

Interestingly, we have identified causal variants in several clinically actionable genes. Five individuals have pathogenic variants in *KMT2B* [MIM: 606834]; so may be responsive to treatment with deep brain stimulation.[8,39] Five other individuals have causal variants in *SCN1A*

[MIM: 182389], of which at least three are predicted LOF; in these cases treatment with sodium channel blockers can worsen seizures.[40] These examples demonstrate the clinical importance of genetic diagnoses and the value of this study.

Reanalysis of sequencing data notably increased the diagnostic yield, largely due to causal variants identified in genes newly associated with NDD, as has previously been reported.[33] This is an important argument for GS or ES over panel sequencing, in which any reanalysis would be limited to previously selected genes. We therefore recommend that similar studies perform regular reanalysis where possible, however in practice the decision of whether to reanalyze data for any given cohort, and how frequently to do so, must balance this advantage against the resource required, and it will depend partly on the number of recently discovered gene-disease associations since the last analysis.

Because we had no cases where both ES and GS were performed on the same samples we cannot perform a direct comparison between these technologies, as other studies have previously done.[23,41] Variants suspected to be cryptic to ES include the deep intronic SNV in Participant 2, the two inversions and the three large insertions, which breakpoints occur in intronic regions. However, it is known that variants in GC-rich regions and CNVs (especially small CNVs) are also challenging to detect using ES. Therefore, we cannot exclude the possibility that additional variants would have been missed by ES. On the other hand, despite significant reductions in the cost of GS, it still remains more costly than ES, and is performed at lower depth than ES. This can affect some analyses, such as detection of SVs and mosaic variants. These previously published considerations should guide selection of the optimal sequencing strategy for a given study.[42]

The use of lrGS in human genomics has expanded greatly over recent years, largely due to technological improvements along with development of new algorithms for processing and interpreting the data.[43] Applications include insights into the biology and consequences of

SVs[30,44] and identification of pathogenic variants in rare diseases that were intractable to other methodologies, usually in individual cases.[6,21,45,46] Here, we used lrGS to resolve complex SVs that could not be characterized by short-reads in two individuals, and to validate or phase variants in three additional individuals. Haplotype phasing in Participant 9 was not possible due to read-length limitation, highlighting the importance for ultra-long reads.[45,47] Overall, our results give several examples of the utility of long-read sequencing. In the future, larger-scale, more systematic lrGS studies of NDDs, facilitated by further improvements to technology, algorithms and pipelines, will yield further insights into the prevalence and biology of previously intractable pathogenic variants.

Our work demonstrates the value of GS to investigate the genetic basis of NDDs and provides insight into the genetic architecture of these disorders. We support the importance of reanalysis and demonstrate that variants cryptic to traditional technologies such as small and cxSVs, non-coding and MT variants can be captured by GS increasing diagnostic yield. Further detailed characterization of genomic variation in large-scale GS studies will be essential for further unveiling the genetic architecture of NDDs in coding and non-coding regions of the human genome.

Declaration of interests

K.J.C and K.M. are currently employees of AstraZeneca.

Consortia

NIHR BioResource: Stephen Abbs, Lara Abulhoul, Julian Adlard, Munaza Ahmed, Timothy J. Aitman, Hana Alachkar, David J. Allsup, Jeff Almeida-King, Philip Ancliff, Richard Antrobus, Ruth Armstrong, Gavin Arno, Sofie Ashford, William J. Astle, Anthony Attwood, Paul Aurora, Christian Babbs, Chiara Bacchelli, Tamam Bakchoul, Siddharth Banka, Tadbir Bariana, Julian Barwell, Joana Batista, Helen E. Baxendale, Phil L. Beales, David L. Bennett, David R. Bentley, Agnieszka Bierzynska, Tina Biss, Maria A. K. Bitner-Glindzicz, Graeme C. Black, Marta Bleda, Iulia Blesneac, Detlef Bockenhauer, Harm Bogaard, Christian J. Bourne, Sara Boyce, John R. Bradley, Eugene Bragin, Gerome Breen, Paul Brennan, Carole Brewer, Matthew Brown, Andrew C. Browning, Michael J. Browning, Rachel J. Buchan, Matthew S. Buckland, Teofila Bueser, Carmen Bugarin Diz, John Burn, Siobhan O. Burns, Oliver S. Burren, Nigel Burrows, Paul Calleja, Carolyn Campbell, Gerald Carr-White, Keren Carss, Ruth Casey, Mark J. Caulfield, Jenny Chambers, John Chambers, Melanie M. Y. Chan, Calvin Cheah, Floria Cheng, Patrick F. Chinnery, Manali Chitre, Martin T. Christian, Colin Church, Jill Clayton-Smith, Maureen Cleary, Naomi Clements Brod, Gerry Coghlan, Elizabeth Colby, Trevor R. P. Cole, Janine Collins, Peter W. Collins, Camilla Colombo, Cecilia J. Compton, Robin Condliffe, Stuart Cook, H. Terence Cook, Nichola Cooper, Paul A. Corris, Abigail Furnell, Fiona Cunningham, Nicola S. Curry, Antony J. Cutler, Matthew J. Daniels, Mehul Dattani, Louise C. Daugherty, John Davis, Anthony De Soyza, Sri V. V. Deevi, Timothy Dent, Charu Deshpande, Eleanor F. Dewhurst, Peter H. Dixon, Sofia Douzgou, Kate Downes, Anna M. Drazyk, Elizabeth Drewe, Daniel Duarte, Tina Dutt, J. David M. Edgar, Karen Edwards, William Egner, Melanie N. Ekani, Perry Elliott, Wendy N. Erber, Marie Erwood, Maria C. Estiu, Dafydd Gareth Evans, Gillian Evans, Tamara Everington, Mélanie Eyries, Hiva Fassihi, Remi Favier, Jack Findhammer, Debra Fletcher, Frances A. Flinter, R. Andres Floto, Tom Fowler,

James Fox, Amy J. Frary, Courtney E. French, Kathleen Freson, Mattia Frontini, Daniel P. Gale, Henning Gall, Vijeya Ganesan, Michael Gattens, Claire Geoghegan, Terence S. A. Gerighty, Ali G. Gharavi, Stefano Ghio, Hossein-Ardeschir Ghofrani, J. Simon R. Gibbs, Kate Gibson, Kimberly C. Gilmour, Barbara Girerd, Nicholas S. Gleadall, Sarah Goddard, David B. Goldstein, Keith Gomez, Pavels Gordins, David Gosal, Stefan Gräf, Jodie Graham, Luigi Grassi, Daniel Greene, Lynn Greenhalgh, Andreas Greinacher, Paolo Gresele, Philip Griffiths, Sofia Grigoriadou, Russell J. Grocock, Detelina Grozeva, Mark Gurnell, Scott Hackett, Charaka Hadinnapola, William M. Hague, Rosie Hague, Matthias Haimel, Matthew Hall, Helen L. Hanson, Eshika Haque, Kirsty Harkness, Andrew R. Harper, Claire L. Harris, Daniel Hart, Ahamad Hassan, Grant Hayman, Alex Henderson, Archana Herwadkar, Jonathan Hoffman, Simon Holden, Rita Horvath, Henry Houlden, Arjan C. Houweling, Luke S. Howard, Fengyuan Hu, Gavin Hudson, Joseph Hughes, Aarnoud P. Huissoon, Marc Humbert, Sean Humphray, Sarah Hunter, Matthew Hurles, Melita Irving, Louise Izatt, Roger James, Sally A. Johnson, Stephen Jolles, Jennifer Jolley, Dragana Josifova, Neringa Jurkute, Tim Karten, Johannes Karten, Mary A. Kasanicki, Hanadi Kazkaz, Rashid Kazmi, Peter Kelleher, Anne M. Kelly, Wilf Kelsall, Carly Kempster, David G. Kiely, Nathalie Kingston, Robert Klima, Nils Koelling, Myrto Kostadima, Gabor Kovacs, Ania Koziell, Roman Kreuzhuber, Taco W. Kuijpers, Ajith Kumar, Dinakantha Kumararatne, Manju A. Kurian, Michael A. Laffan, Fiona Lalloo, Michele Lambert, Hana Lango Allen, Allan Lawrie, D. Mark Layton, Nick Lench, Claire Lentaigne, Tracy Lester, Adam P. Levine, Rachel Linger, Hilary Longhurst, Lorena E. Lorenzo, Eleni Louka, Paul A. Lyons, Rajiv D. Machado, Robert V. MacKenzie Ross, Bella Madan, Eamonn R. Maher, Jesmeen Maimaris, Samantha Malka, Sarah Mangles, Rutendo Mapeta, Kevin J. Marchbank, Stephen Marks, Hugh S. Markus, Hanns-Ulrich Marschall, Andrew Marshall, Jennifer Martin, Mary Mathias, Emma Matthews, Heather Maxwell, Paul McAlinden, Mark I. McCarthy, Harriet McKinney, Aoife McMahon, Stuart Meacham, Adam J. Mead, Ignacio Medina Castello, Karyn Megy, Sarju G. Mehta, Michel Michaelides, Carolyn Millar, Shehla N. Mohammed, Shahin Moledina, David Montani, Anthony T. Moore, Joannella Morales, Nicholas W. Morrell, Monika Mozere, Keith W. Muir, Andrew D. Mumford, Andrea H. Nemeth, William G. Newman, Michael

Newnham, Sadia Noorani, Paquita Nurden, Jennifer O'Sullivan, Samya Obaji, Chris Odhams, Steven Okoli, Andrea Olschewski, Horst Olschewski, Kai Ren Ong, S. Helen Oram, Elizabeth Ormondroyd, Willem H. Ouwehand, Claire Palles, Sofia Papadia, Soo-Mi Park, David Parry, Smita Patel, Joan Paterson, Andrew Peacock, Simon H. Pearce, John Peden, Kathelijne Peerlinck, Christopher J. Penkett, Joanna Pepke-Zaba, Romina Petersen, Clarissa Pilkington, Kenneth E. S. Poole, Radhika Prathalingam, Bethan Psaila, Angela Pyle, Richard Quinton, Shamima Rahman, Stuart Rankin, Anupama Rao, F. Lucy Raymond, Paula J. Rayner-Matthews, Christine Rees, Augusto Rendon, Tara Renton, Christopher J. Rhodes, Andrew S. C. Rice, Sylvia Richardson, Alex Richter, Leema Robert, Irene Roberts, Anthony Rogers, Sarah J. Rose, Robert Ross-Russell, Catherine Roughley, Noemi B. A. Roy, Deborah M. Ruddy, Omid Sadeghi-Alavijeh, Moin A. Saleem, Nilesh Samani, Crina Samarghitean, Alba Sanchis-Juan, Ravishankar B. Sargur, Robert N. Sarkany, Simon Satchell, Sinisa Savic, John A. Sayer, Genevieve Sayer, Laura Scelsi, Andrew M. Schaefer, Sol Schulman, Richard Scott, Marie Scully, Claire Searle, Werner Seeger, Arjune Sen, W. A. Carrock Sewell, Denis Seyres, Neil Shah, Olga Shamardina, Susan E. Shapiro, Adam C. Shaw, Patrick J. Short, Keith Sibson, Lucy Side, Ilenia Simeoni, Michael A. Simpson, Matthew C. Sims, Suthesh Sivapalaratnam, Damian Smedley, Katherine R. Smith, Kenneth G. C. Smith, Katie Snape, Nicole Soranzo, Florent Soubrier, Laura Southgate, Olivera Spasic-Boskovic, Simon Staines, Emily Staples, Hannah Stark, Jonathan Stephens, Charles Steward, Kathleen E. Stirrups, Alex Stuckey, Jay Suntharalingam, Emilia M. Swietlik, Petros Syrris, R. Campbell Tait, Kate Talks, Rhea Y. Y. Tan, Katie Tate, John M. Taylor, Jenny C. Taylor, James E. Thaventhiran, Andreas C. Themistocleous, Ellen Thomas, David Thomas, Moira J. Thomas, Patrick Thomas, Kate Thomson, Adrian J. Thrasher, Glen Threadgold, Chantal Thys, Tobias Tilly, Marc Tischkowitz, Catherine Titterton, John A. Todd, Cheng-Hock Toh, Bas Tolhuis, Ian P. Tomlinson, Mark Toshner, Matthew Traylor, Carmen Treacy, Paul Treadaway, Richard Trembath, Salih Tuna, Wojciech Turek, Ernest Turro, Philip Twiss, Tom Vale, Chris Van Geet, Natalie van Zuydam, Maarten Vandekuilen, Anthony M. Vandersteen, Marta Vazquez-Lopez, Julie von Ziegenweidt, Anton Vonk Noordegraaf, Annette Wagner, Quinten Waisfisz, Suellen M.

Walker, Neil Walker, Klaudia Walter, James S. Ware, Hugh Watkins, Christopher Watt, Andrew R. Webster, Lucy Wedderburn, Wei Wei, Steven B. Welch, Julie Wessels, Sarah K. Westbury, John-Paul Westwood, John Wharton, Deborah Whitehorn, James Whitworth, Andrew O. M. Wilkie, Martin R. Wilkins, Catherine Williamson, Brian T. Wilson, Edwin K. S. Wong, Nicholas Wood, Yvette Wood, Christopher Geoffrey Woods, Emma R. Woodward, Stephen J. Wort, Austen Worth, Michael Wright, Katherine Yates, Patrick F. K. Yong, Timothy Young, Ping Yu, Patrick Yu-Wai-Man & Eliska Zlamalova

Acknowledgements

We thank the participants involved in this study and their families. This work was supported by The National Institute for Health Research England (NIHR) for the NIHR BioResource project (grant number RG65966). This work is partly funded by the NIHR GOSH BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. This research was supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Author contributions

Conceptualization: A.S.-J., K.C., F.L.R.; Data curation: A.S.-J., K.M., L.C.D., K.C.; Formal analysis: A.S.-J., K.C.; Funding acquisition: NIHR, F.L.R.; Investigation: A.S.-J., K.M., C.A.R., K.L., C.E.F., D.G., L.C.D., K.C., F.L.R.; Methodology: A.S.-J., K.C., F.L.R.; Project administration: K.S., E.D., M.E.; Resources: A.M.T., S.T.D., A.H.-F., J.V., G.A., M.C., D.J., M.A.K., A.P., J.R., E.R., E.W., E.W., C.G.W.; Software: A.S.-J., C.P., O.S., S.T., N.G.; Supervision: K.C., F.L.R.; Validation: J.S.; Visualization: A.S.-J.; Writing-original draft: A.S.-J., K.C.; Writing-review & editing: A.S.-J., K.C., F.L.R.

Data and code availability

The genome data generated during this study are available at the European Genome-phenome Archive (EGA) under accession number EGAD00001004522 [https://ega-archive.org/datasets/EGAD00001004522].

Figure legends

Figure 1. Factors affecting variant discovery and diagnostic yield. A) Diagnostic yield is affected by family structure sequenced. Boxes show number of affected individuals in each class of family structure. Singletons have no sequenced relatives, trios have both parents sequenced, Proband-Parents have one parent sequenced, siblings have one sibling sequenced, and quads have both parents and one sibling sequenced. Solved refers to an affected individual with a P/LP variant. Partially solved refers to an affected individual with a P/LP variant that only partially explains the phenotype. VUS refers to an affected individual with a Variant of Uncertain Significance. Unsolved refers to an affected individual with no identified P/LP variants or VUSs. **B)** Diagnostic yield is affected by phenotype. Boxes show number of affected individuals with each phenotype. These numbers overlap because many individuals have more than one phenotype. ASD=Autism Spectrum Disorder; CNS=Central Nervous System. **C)** Proportion of identified variants that are P/LP is affected by mode of inheritance. Boxes show number of identified variants in each class. XLR=X-linked Recessive; XLD=X-linked Dominant; MT=mitochondrial; VUS=Variant of Uncertain Significance; P=Pathogenic; LP=Likely Pathogenic. **D)** Number of identified variants that are P/LP is affected by round of analysis, with new variants identified in each successive round, demonstrating the value of re-analysis. Boxes show number of variants identified in each round (cumulative). Round 1 was March 2016 – January 2018; round 2 was July 2018, and round 3 was July 2019.

Figure 2. Complex structural variants resolved by lrGS. A) Circular layout plot of the complex rearrangement in **A)** Participant 6, involving 37 breakpoints between chromosomes 7, 10 and 12, and **B)** Participant 7, involving 26 duplicated fragments from 14 chromosomes. Both A and B panels have been generated with Circos[48], the outer ring shows the chromosomes (coordinates in Mbp), and the inner ring shows the depth coverage of the individual, normalized using 250 unrelated individuals in the cohort. In the scatter plot,

deletions are shown in red and duplication in blue. Breakpoint junctions links are shown in black (interchromosomal) and green (intrachromosomal). **C)** Variant phasing performed on Participant 8 demonstrated the absence of an inversion called in *CASK* gene in the SRS data. The ideogram for chromosome X highlighting the region involved is at the top, followed by the genes present within this region and the inversion coordinates represented in green. A zoomed in panel for both start (S) and end (E) of the inversion are shown next for SRS and LRS data. It is noticeable that both are located within LINE-1 retrotransposon repeats (Rep), and are not supported by LRS data. **D)** Variant phasing performed on Participant 10 facilitated resolution of a complex event involving a retroelement of *KIF5C* gene. At the top the ideogram of chromosome 2 is represented, followed by the *KIF5C* transcripts, and a zoomed in region with the short-read sequencing (SRS) calls; deletions are shown in red, inversions in green and the duplication in blue. The following two panels show the coverage (Cov) and IGV[18] visualization of the short reads and the long-read sequencing (LRS) alignments. Split reads and discordant pairs are present in the SRS data and absent in the LRS, consistent with the retroelement insertion.

Tables

Table 1. Diagnostic yield by family structure. GS identified causal variants in 36% cases, 23% had a reported VUS and 41% remained unresolved. Partial contribution refers to individuals with a causal variant that partially explains the phenotype. VUS=Variant of Uncertain Significance; LP=Likely Pathogenic; P=Pathogenic. *One trio includes an affected parent.

			Reportable variants: 289 individuals (59%)			**No causal variant identified: 200 individuals (41%)**
			P/LP 177 individuals (36%)		**VUS: 112 individuals (23%)**	
	Family structure	**Affected individuals (families)**	**Full contribution: 168 individuals (34%)**	**Partial contribution: 9 individuals (2%)**		
Total affected individuals 489 (465 families)	**Singleton**	335 (335)	111 (33%)	8 (2%)	78 (23%)	138 (42%)
	Trio	68 (67)*	28 (41%)	0 (0%)	12 (18%)	28 (41)
	Two siblings	28 (14)	16 (57%)	0 (0%)	4 (14%)	8 (29%)
	Cousins	2 (1)	2 (100%)	0 (0%)	0 (0%)	0 (0%)
	Proband and parent	39 (39)	7 (18%)	0 (0%)	14 (36%)	18 (46%)
	Proband and grandparent	2 (1)	0 (0%)	0 (0%)	0 (0%)	2 (100%)
	Quad	9 (5)	4 (45%)	1 (11%)	2 (22%)	2 (22%)
	Proband, sibling and parent	6 (3)	0 (0%)	0 (0%)	2 (33%)	4 (67%)

Table 2. Candidate variants identified by pathogenicity and type. SNV=Single Nucleotide Variant; SV=Structural Variant; ROH=Region Of Homozygosity; STR=Single Tandem Repeat; SMA=Spinal Muscular Atrophy; VUS=Variant of Uncertain Significance.

Type	Total	Pathogenic	Likely pathogenic	VUS
SNV	279	48	84	147
Indel	54	23	22	9
Deletion	31	7	16	8
Duplication	6	0	0	6
Complex SV	4	0	2	2
Large Insertion	3	0	0	3
Inversion	2	0	0	2
ROH	1	1	0	0
STR expansions	0	0	0	0
SMA	0	0	0	0
Total	**380**	**79**	**124**	**177**

Table 3. Long-read GS was performed on five participants to resolve cxSVs, variant phasing and to facilitate resolutions of technically challenging regions in five individuals. Individual IDs in parenthesis correspond to those in Table S2.

Individual	Phenotype	Finding srGS	Reason inclusion lrGS	Finding lrGS
Participant 6 (NGC00375_01)	Dystonia, myoclonus; delayed gross motor development; learning and intellectual disability	cxSV involving *SGCE* gene	Unable to resolve by srGS, highly complex	cxSV involving 37 breakpoints
Participant 7 (G012664)	Paroxysmal intermittent limping right leg; bulbar palsy	cxSV involving multiple duplications	Unable to resolve by srGS, highly complex	cxSV involving 26 duplicated fragments
Participant 8 (G013428)	Severe global developmental delay; hypotonia with chorea like movement disorder; sensorineural hearing impairment; microcephaly; delayed visual maturation with esotropia	Inversion chrX:41426631-41501873	Unable to resolve by srGS and Sanger sequencing	Variant not supported by lrGS
Participant 9 (G013407)	Early infantile epileptic encephalopathy	NM_004408.4:c.1082G>C p.(Arg361Pro)	Haplotype phasing	Inconclusive
Participant 10 (G000973)	Early onset dementia; spastic paraplegia; thin corpus callosum	cxSV involving *KIF5C* gene	Unable to resolve by srGS, possible complex retrotransposon	Retrotransposon insertion in chr5:25000434 – not complex, unknown effect

References

1. Boycott, K.M., Hartley, T., Biesecker, L.G., Gibbs, R.A., Innes, A.M., Riess, O., Belmont, J., Dunwoodie, S.L., Jojic, N., Lassmann, T., et al. (2019). A Diagnosis for All Rare Genetic Diseases: The Horizon and the Next Frontiers. Cell *177*, 32-37. 10.1016/j.cell.2019.02.040.
2. Deciphering Developmental Disorders, S. (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature *542*, 433-438. 10.1038/nature21062.
3. Investigators, G.P.P., Smedley, D., Smith, K.R., Martin, A., Thomas, E.A., McDonagh, E.M., Cipriani, V., Ellingford, J.M., Arno, G., Tucci, A., et al. (2021). 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N Engl J Med *385*, 1868-1880. 10.1056/NEJMoa2035790.
4. Clark, M.M., Stark, Z., Farnaes, L., Tan, T.Y., White, S.M., Dimmock, D., and Kingsmore, S.F. (2018). Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. NPJ Genom Med *3*, 16. 10.1038/s41525-018-0053-8.
5. Belyeu, J.R., Brand, H., Wang, H., Zhao, X., Pedersen, B.S., Feusier, J., Gupta, M., Nicholas, T.J., Brown, J., Baird, L., et al. (2021). De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. Am. J. Hum. Genet. *108*, 597-607. 10.1016/j.ajhg.2021.02.012.
6. Mitsuhashi, S., and Matsumoto, N. (2020). Long-read sequencing for rare human genetic diseases. J. Hum. Genet. *65*, 11-19. 10.1038/s10038-019-0671-8.
7. Turro, E., Astle, W.J., Megy, K., Gräf, S., Greene, D., Shamardina, O., Allen, H.L., Sanchis-Juan, A., Frontini, M., Thys, C., et al. (2020). Whole-genome sequencing of patients with rare diseases in a national health system. Nature *583*, 96-102. 10.1038/s41586-020-2434-2.
8. Meyer, E., Carss, K.J., Rankin, J., Nichols, J.M.E., Grozeva, D., Joseph, A.P., Mencacci, N.E., Papandreou, A., Ng, J., Barral, S., et al. (2017). Mutations in the histone methyltransferase gene KMT2B cause complex early-onset dystonia. Nat. Genet. *49*, 223-237. 10.1038/ng.3740.
9. Helbig, K.L., Lauerer, R.J., Bahr, J.C., Souza, I.A., Myers, C.T., Uysal, B., Schwarz, N., Gandini, M.A., Huang, S., Keren, B., et al. (2018). De Novo Pathogenic Variants in CACNA1E Cause Developmental and Epileptic Encephalopathy with Contractures, Macrocephaly, and Dyskinesias. Am. J. Hum. Genet. *103*, 666-678. 10.1016/j.ajhg.2018.09.006.
10. Ito, Y., Carss, K.J., Duarte, S.T., Hartley, T., Keren, B., Kurian, M.A., Marey, I., Charles, P., Mendonça, C., Nava, C., et al. (2018). De Novo Truncating Mutations in WASF1 Cause Intellectual Disability with Seizures. Am. J. Hum. Genet. *103*, 144-153. 10.1016/j.ajhg.2018.06.001.
11. Sanchis-Juan, A., Hasenahuer, M.A., Baker, J.A., McTague, A., Barwick, K., Kurian, M.A., Duarte, S.T., BioResource, N., Carss, K.J., Thornton, J., and Raymond, F.L. (2020). Structural analysis of pathogenic missense mutations in GABRA2 and identification of a novel de novo variant in the desensitization gate. Mol Genet Genomic Med *8*, e1106. 10.1002/mgg3.1106.
12. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405-424. 10.1038/gim.2015.30.

13. Sobreira, N., Schiettecatte, F., Valle, D., and Hamosh, A. (2015). GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat *36*, 928-930. 10.1002/humu.22844.
14. Martin, A.R., Williams, E., Foulger, R.E., Leigh, S., Daugherty, L.C., Niblock, O., Leong, I.U.S., Smith, K.R., Gerasimenko, O., Haraldsdottir, E., et al. (2019). PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet *51*, 1560-1565. 10.1038/s41588-019-0528-2.
15. Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., van Kogelenberg, M., King, D.A., Ambridge, K., Barrett, D.M., Bayzetinova, T., et al. (2015). Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet *385*, 1305-1314. 10.1016/S0140-6736(14)61705-0.
16. Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S.T., Abeysinghe, S., Krawczak, M., and Cooper, D.N. (2003). Human Gene Mutation Database (HGMD®): 2003 update. Human Mutation *21*, 577-581. 10.1002/humu.10212.
17. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res *46*, D1062-D1067. 10.1093/nar/gkx1153.
18. Robinson, J.T., Thorvaldsdóttir, H., Wenger, A.M., Zehir, A., and Mesirov, J.P. (2017). Variant Review with the Integrative Genomics Viewer. Cancer Res. *77*, e31-e34. 10.1158/0008-5472.CAN-17-0337.
19. Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S.R.F., Consortium, W.G.S., Wilkie, A.O.M., McVean, G., and Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat. Genet. *46*, 912-918. 10.1038/ng.3036.
20. French, C.E., Delon, I., Dolling, H., Sanchis-Juan, A., Shamardina, O., Mégy, K., Abbs, S., Austin, T., Bowdin, S., Branco, R.G., et al. (2019). Whole genome sequencing reveals that genetic conditions are frequent in intensively ill children. Intensive Care Med. *45*, 627-636. 10.1007/s00134-019-05552-x.
21. de la Morena-Barrio, B., Stephens, J., de la Morena-Barrio, M.E., Stefanucci, L., Padilla, J., Minano, A., Gleadall, N., Garcia, J.L., Lopez-Fernandez, M.F., Morange, P.E., et al. (2022). Long-Read Sequencing Identifies the First Retrotransposon Insertion and Resolves Structural Variants Causing Antithrombin Deficiency. Thromb Haemost *122*, 1369-1378. 10.1055/s-0042-1749345.
22. Sanchis-Juan, A., Bitsara, C., Low, K.Y., Carss, K.J., French, C.E., Spasic-Boskovic, O., Jarvis, J., Field, M., Raymond, F.L., and Grozeva, D. (2019). Rare Genetic Variation in 135 Families With Family History Suggestive of X-Linked Intellectual Disability. Front. Genet. *10*, 578. 10.3389/fgene.2019.00578.
23. Carss, K.J., Arno, G., Erwood, M., Stephens, J., Sanchis-Juan, A., Hull, S., Megy, K., Grozeva, D., Dewhurst, E., Malka, S., et al. (2017). Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. Am. J. Hum. Genet. *100*, 75-90. 10.1016/j.ajhg.2016.12.003.
24. Nellist, M., Brouwer, R.W.W., Kockx, C.E.M., van Veghel-Plandsoen, M., Withagen-Hermans, C., Prins-Bakker, L., Hoogeveen-Westerveld, M., Mrsic, A., van den Berg, M.M.P., Koopmans, A.E., et al. (2015). Targeted Next Generation Sequencing reveals previously unidentified TSC1 and TSC2 mutations. BMC Med. Genet. *16*, 10. 10.1186/s12881-015-0155-4.

25. Rahman, S., Poulton, J., Marchington, D., and Suomalainen, A. (2001). Decrease of 3243 A-->G mtDNA mutation from blood in MELAS syndrome: a longitudinal study. Am. J. Hum. Genet. *68*, 238-240. 10.1086/316930.
26. López-Gallardo, E., Emperador, S., Solano, A., Llobet, L., Martín-Navarro, A., López-Pérez, M.J., Briones, P., Pineda, M., Artuch, R., Barraquer, E., et al. (2014). Expanding the clinical phenotypes of MT-ATP6 mutations. Hum. Mol. Genet. *23*, 6191-6200. 10.1093/hmg/ddu339.
27. Yu-Wai-Man, P., and Chinnery, P.F. (1993). Leber Hereditary Optic Neuropathy. In GeneReviews((R)), M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, K.W. Gripp, G.M. Mirzaa, and A. Amemiya, eds.
28. Murakami, T., Mita, S., Tokunaga, M., Maeda, H., Ueyama, H., Kumamoto, T., Uchino, M., and Ando, M. (1996). Hereditary cerebellar ataxia with Leber's hereditary optic neuropathy mitochondrial DNA 11778 mutation. J Neurol Sci *142*, 111-113. 10.1016/0022-510x(96)00165-7.
29. Grazina, M.M., Diogo, L.M., Garcia, P.C., Silva, E.D., Garcia, T.D., Robalo, C.B., and Oliveira, C.R. (2007). Atypical presentation of Leber's hereditary optic neuropathy associated to mtDNA 11778G>A point mutation--A case report. Eur J Paediatr Neurol *11*, 115-118. 10.1016/j.ejpn.2006.11.015.
30. Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H.P., Bjornsson, E., Jonsson, H., Atlason, B.A., Kristmundsdottir, S., Mehringer, S., Hardarson, M.T., et al. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat Genet *53*, 779-786. 10.1038/s41588-021-00865-4.
31. Acuna-Hidalgo, R., Veltman, J.A., and Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. Genome Biol. *17*, 241. 10.1186/s13059-016-1110-1.
32. Kaer, K., and Speek, M. (2013). Retroelements in human disease. Gene *518*, 231-241. 10.1016/j.gene.2013.01.008.
33. Wright, C.F., Campbell, P., Eberhardt, R.Y., Aitken, S., Perrett, D., Brent, S., Danecek, P., Gardner, E.J., Chundru, V.K., Lindsay, S.J., et al. (2022). Optimising diagnostic yield in highly penetrant genomic disease. medRxiv.
34. Grozeva, D., Carss, K., Spasic-Boskovic, O., Tejada, M.I., Gecz, J., Shaw, M., Corbett, M., Haan, E., Thompson, E., Friend, K., et al. (2015). Targeted Next-Generation Sequencing Analysis of 1,000 Individuals with Intellectual Disability. Hum Mutat *36*, 1197-1204. 10.1002/humu.22901.
35. Riggs, E.R., Andersen, E.F., Cherry, A.M., Kantarci, S., Kearney, H., Patel, A., Raca, G., Ritter, D.I., South, S.T., Thorland, E.C., et al. (2020). Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). Genet. Med. *22*, 245-257. 10.1038/s41436-019-0686-8.
36. Short, P.J., McRae, J.F., Gallone, G., Sifrim, A., Won, H., Geschwind, D.H., Wright, C.F., Firth, H.V., FitzPatrick, D.R., Barrett, J.C., and Hurles, M.E. (2018). De novo mutations in regulatory elements in neurodevelopmental disorders. Nature *555*, 611-616. 10.1038/nature25983.
37. All of Us Research Program, I., Denny, J.C., Rutter, J.L., Goldstein, D.B., Philippakis, A., Smoller, J.W., Jenkins, G., and Dishman, E. (2019). The "All of Us" Research Program. N Engl J Med *381*, 668-676. 10.1056/NEJMsr1809937.
38. Halldorsson, B.V., Eggertsson, H.P., Moore, K.H.S., Hauswedell, H., Eiriksson, O., Ulfarsson, M.O., Palsson, G., Hardarson, M.T., Oddsson, A., Jensson, B.O., et al.

(2022). The sequences of 150,119 genomes in the UK Biobank. Nature *607*, 732-740. 10.1038/s41586-022-04965-x.
39. Dafsari, H.S., Sprute, R., Wunderlich, G., Daimagüler, H.-S., Karaca, E., Contreras, A., Becker, K., Schulze-Rhonhof, M., Kiening, K., Karakulak, T., et al. (2019). Novel mutations in KMT2B offer pathophysiological insights into childhood-onset progressive dystonia. J. Hum. Genet. *64*, 803-813. 10.1038/s10038-019-0625-1.
40. Ziobro, J., Eschbach, K., Sullivan, J.E., and Knupp, K.G. (2018). Current Treatment Strategies and Future Treatment Options for Dravet Syndrome. Curr. Treat. Options Neurol. *20*, 52. 10.1007/s11940-018-0537-y.
41. Lowther, C., Valkanas, E., Giordano, J.L., Wang, H.Z., Currall, B.B., O'Keefe, K., Pierce-Hoffman, E., Kurtas, N.E., Whelan, C.W., Hao, S.P., et al. (2022). Systematic evaluation of genome sequencing for the assessment of fetal structural anomalies. bioRxiv, 2020.2008.2012.248526. 10.1101/2020.08.12.248526.
42. Lavelle, T.A., Feng, X., Keisler, M., Cohen, J.T., Neumann, P.J., Prichard, D., Schroeder, B.E., Salyakina, D., Espinal, P.S., Weidner, S.B., and Maron, J.L. (2022). Cost-effectiveness of exome and genome sequencing for children with rare and undiagnosed conditions. Genet Med *24*, 1349-1361. 10.1016/j.gim.2022.03.005.
43. Logsdon, G.A., Vollger, M.R., and Eichler, E.E. (2020). Long-read human genome sequencing and its applications. Nat. Rev. Genet. *21*, 597-614. 10.1038/s41576-020-0236-x.
44. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444-451. 10.1038/s41586-020-2287-8.
45. Sanchis-Juan, A., Stephens, J., French, C.E., Gleadall, N., Mégy, K., Penkett, C., Shamardina, O., Stirrups, K., Delon, I., Dewhurst, E., et al. (2018). Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. Genome Med. *10*, 95. 10.1186/s13073-018-0606-6.
46. Thibodeau, M.L., O'Neill, K., Dixon, K., Reisle, C., Mungall, K.L., Krzywinski, M., Shen, Y., Lim, H.J., Cheng, D., Tse, K., et al. (2020). Improved structural variant interpretation for hereditary cancer susceptibility using long-read sequencing. Genet. Med. *22*, 1892-1897. 10.1038/s41436-020-0880-8.
47. Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol *36*, 338-345. 10.1038/nbt.4060.
48. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res *19*, 1639-1645. 10.1101/gr.092759.109.