Phenotype and genetic analysis of data collected within the first year of NeuroDev: A Pilot Study

Patricia Kipkemoi[1,2], Heesu Ally Kim[3], Bjorn Christ[4], Emily O'Heir[3], Jake Allen[3], Christina Austin-Tse[3,5], Samantha Baxter[3], Harrison Brand[3,5,6], Sam Bryant[3], Nick Buser[3], Victoria de Menil[3,7], Emma Eastman[4], Alice Galvin[3], Martha Kombe[1], Collins Kipkoech[1], Alysia Lovgren[3], Daniel G. MacArthur[3], Brigitte Melly[4], Katini Mwangasha[1], Alba Sanchis-Juan[3,5,6], Moriel Singer-Berk[3], Michael E. Talkowski[3,5,6], Grace VanNoy[3], Celia van der Merwe[3], The NeuroDev Project, Charles Newton[1,8,9], Anne O'Donnell-Luria[3,10], Amina Abubakar*[1,8,9], Kirsten A Donald*[4,11], Elise Robinson*[3,5]

[1] Neuroscience Unit, KEMRI-Wellcome Trust, Center for Geographic Medicine Research Coast, Kilifi, Kenya
[2] Complex Trait Genetics Department, Center for Neurogenomics and Cognitive Research, Vrije Universiteit Amsterdam, Netherlands
[3] The Broad Institute of MIT and Harvard, Cambridge MA, USA
[4] Department of Paediatrics and Child Health, 4th Floor ICH Building, Red Cross War Memorial Children's Hospital and University of Cape Town, Rondebosch, South Africa
[5] Center for Genomic Medicine, Massachusetts General Hospital, Boston MA, USA
[6] Department of Neurology, Harvard Medical School, Boston MA, USA
[7] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston MA, USA
[8] Department of Psychiatry, University of Oxford, London, UK
[9] Institute of Human Development, Aga Khan University, Nairobi, Kenya
[10] Division of Genetics and Genomics, Boston Children's Hospital, Boston MA, USA
[11] Neuroscience Institute, University of Cape Town, Groote Schuur Hospital, Observatory, South Africa

*Correspondence to Drs. Abubakar (amina.abubakar@aku.edu), Donald (kirsty.donald@uct.ac.za), and Robinson (erob@broadinstitute.org)

Summary

Genetic association studies have made significant contributions to our understanding of the aetiology of neurodevelopmental disorders (NDDs). However, the vast majority of these studies have focused on populations of European ancestry, and few include individuals from the African continent. The NeuroDev project aims to address this diversity gap through detailed phenotypic and genetic characterization of children with NDDs from Kenya and South Africa. Here we present results from NeuroDev's first year of data collection, including phenotype data from 206 cases and clinical genetic analysis of 99 parent-child trios. The majority of the cases met criteria for global developmental delay/intellectual disability (GDD/ID, 80.3%). Approximately half of the children with GDD/ID also met criteria for autism spectrum disorders (ASD), and 14.6% met criteria for ASD alone. Analysis of exome sequencing data identified a pathogenic or likely pathogenic variant in 13 (17%) of the 75 cases from South Africa and 9 (38%) of the 24 cases

from Kenya Candidate novel disease gene variants in 7 total cases were matched through MatchMaker Exchange. Data from the trio pilot cases has already been made publicly available, and the NeuroDev project will continue to develop resources for the global genetics community.

Key Words: Developmental Disorders; Autism Spectrum Disorder; Intellectual Disability; Genetics; Exome Sequencing; Phenotypes; De novo variants; South Africa; Kenya; Diverse populations

Introduction

The Genetic Characterization of Neurodevelopmental Disorders project (NeuroDev) is a study of neurodevelopmental disorders (NDDs) that will collect and analyze extensive genetic and phenotypic data from over 5000 people, including ~3600 children and their parents, in Kenya and South Africa over the next several years.[1] Based on existing recruitment patterns, most of the 2000 case individuals enrolled in the study are projected to meet criteria for Global Developmental Delay/Intellectual Disability (GDD/ID) or Autism Spectrum Disorder (ASD). All data (e.g., phenotypes; genotype array and exome sequencing data) and materials (e.g., blood DNA; cryopreserved cell lines or CPLs) generated by NeuroDev will be made publicly available through approved National Institute of Mental Health repositories. Through the data collection activity, which includes African ancestry populations largely absent from genetic reference panels (e.g., gnomAD), and through the public release of deeply characterized case and control data, NeuroDev aims to support diversity in biomedical research.

In this paper, we present the NeuroDev Pilot, which consists of analyses following the project's first collection year. During the first year, we collected data from more than 200 cases and 600 total participants. We describe in this report phenotype data from all cases collected in the pilot period, along with genetic analysis of 99 exome-sequenced trios. Supplementing these early genetic and phenotypic findings, we present learning points of the first year of data collection and modifications made to the NeuroDev protocol initially presented in de Menil et al. (see Supplementary Information).[1] We hope these reflections on process will help others in the design and execution of similar projects, as more work on NDDs in Africa is both needed and underway. At only 99 trios, the NeuroDev trio pilot data is now the largest African NDD collection for which genetic and phenotypic data are publicly available to the research community. The trio data presented here can be accessed through National Human Genome Research Institute (NHGRI) Analysis Visualization and Informatics Lab-space (ANVIL) controlled access data repository (https://anvilproject.org/data/).

Results

Data Collection

From August 2018 to July 2019, we enrolled 219 cases, 195 case mothers, 115 case fathers, and 92 unrelated child controls. There were 106 total parent-case trios, and an additional 113 cases had only one participating parent. We present in this report phenotype analysis of all 219

cases, along with genetic data analysis of the first 100 trios, 99 of which passed quality control measures for analysis.

At the end of the project's first year, both the trio collection rate and the overall case collection rate in NeuroDev were aligned with our four-year sample size targets, and phenotype battery completion was high. In the first year, all participants had data for the demographic, neuromedical assessment, and behavioral measures. The behavioral measures included the Social Communication Disorders Checklist (SCDC)[2]; 3Di Brief[3]; and Swanson, Nolan, and Pelham Rating Scale (SNAP-IV).[4] Item level missingness was below 10% for all measures (for details, see Supplementary Table 1 and 2).

The Raven's Progressive Matrices (RPM), NeuroDev's nonverbal reasoning ability measure, was completed by 99% of participating parents. Inspired by the Simons Simplex Collection (SSC) cognitive testing approach,[5,6] all enrolled children aged 6 years and older were offered the opportunity to attempt the Raven's[7] (Standard RPM if 12 years or older; Colored RPM if 6-11 years), which was adapted and validated for use in Kilifi.

Many case children (49%) could not complete age-appropriate versions of the RPM due to the extent of their developmental delay or behavioral challenges. All case children under 6 years of age and case children over 6 years who could not complete the Ravens were offered the Molteno Adapted Scales of Development (Molteno). We experienced similar challenges in some cases with regards to the completion of the Molteno, though all participants have at least some data. This experience is common to studies of ID and ASD, particularly those that include children.[8,9] Missing values from the Raven's are associated with case severity and will therefore be informative.

In the first year of data collection, trio family ascertainment was higher than anticipated. The goal in South Africa to recruit 100 trios over the full duration of the project was met in just the first year of the study. All pilot participants in South Africa consented to have genetic findings related to their child's NDD returned to the family. Similarly, we observed a high rate (98%) of consent for the option to share cell lines in addition to DNA. Both of these options were available only in South Africa.

We used the University of California, San Diego Brief Assessment of Capacity to Consent (UBACC) to both ensure and measure parents' understanding of the study prior to consent. As a screening tool, the UBACC is used to identify parents who may require a more comprehensive evaluation of decisional capacity and enhanced consent procedures.[10,11] Participants need to achieve a score of 14.5 out of 20 to meet the test requirement, with the tool readministered up to a maximum of three trials if necessary, each after additional explanatory efforts. In the first year of data collection, only two parents failed their UBACC administrations, and their families were not included in the study. One of the two parents was reported to have documented ID. An overwhelming majority of parents showed good understanding of the protocol following detailed explanation by study staff, since only 3% of participants scored below 14.5 on their first trial.[12]

The pilot period was used to further review our data collection strategy and tools. In the Supplementary Information, we share detailed observations about the assessment tools as applied in our context, as well as any adaptations to the tools made at either site. These adaptations were made in response to questions that were contextually inappropriate for caregivers due to cross-cultural differences or linguistically challenging. We also discuss modifications to the protocol, lessons learnt in the implementation of the study, and recruitment strategies used in response to variability in enrollment of participants. For example, we adapted our strategy during inclement weather, and in response to challenges recruiting fathers during the work-week. We hope these details will be of benefit to future research projects, and that this suite of pilot results as a whole will encourage more large-scale NDD projects in Africa.

We made one significant addition to the phenotype battery, the Child Behavior Checklist (CBCL)[13], to strengthen participant behavioral characterization. We initially employed the CBCL to explore the validity of the SNAP-IV in a subset of NeuroDev participants and found it to be of both research and clinical utility. The CBCL, now included in NeuroDev's core phenotype battery, characterizes a comprehensive assortment of problem behaviors separately in preschool (age 1.5-5 years) and school-age (6-18 years) children. In addition to the research opportunities afforded by the CBCL data, we observed clinical value in use of the CBCL to construct a behavioral profile for children that may need a referral for specialized intervention.

Phenotypic characteristics of cases ascertained in NeuroDev's first year

NeuroDev features an uncommonly detailed phenotype battery for a genetic study of its size. In this section, we present phenotypic findings from all 219 cases collected between August 2018 to July 2019. Of those cases, 156 were South African and 63 were Kenyan, reflecting the fact that the Kenyan data collection began halfway through the pilot's first year; 99 of these cases were included among the exome-sequenced trios (Figure 1A). Overall, cases were 70.2% male, and they ranged in age from 2-18 years (Figure 1B). The majority (80.4%, N = 176) of cases met criteria for global developmental delay or intellectual disability (GDD/ID). Of those with GDD/ID, 91 cases (52% of all GDD/ID cases) also met criteria for ASD. An additional 31 cases (14.2% of all cases) met criteria for ASD without GDD/ID. A small number (5.5%) of cases did not meet criteria for either GDD/ID or ASD, but for other NDDs that are ascertained through NeuroDev: specific learning disabilities, communication disorders, and/or ADHD. The diagnostic composition of cases included in the 99 exome-sequenced trios is highly similar to the full set 219 cases (Figure 1E).

The cases were highly ancestrally and linguistically diverse, with more than 40 languages spoken in families and more than 24 ethnicities represented (Figure 1F). In keeping with the approach initiated by the 1000 Genomes study, we used language as an indicator of ethnic affiliation on top of self-reported ethnicity. Among cases from Kenya, the majority were from the Mijikenda ethnic group (88.7%) with many of the case families primarily speaking the Mijikenda languages (83.9%) or Kiswahili (14.5%). Within South Africa, most of the cases were of mixed ancestry (45.9%) or identified with multiple ancestries (16.9%), and many others identified as AmaXhosa (13.5%). Reflecting this, the main languages spoken by the case families in South

Africa included English (74.3%) and isiXhosa (14.2%). The specific questions on language and ethnicity in the demographic questionnaire were phrased as follows: 'What is the primary language spoken in [participant's] home?' and 'What is [participant's] ethnicity or tribe?'

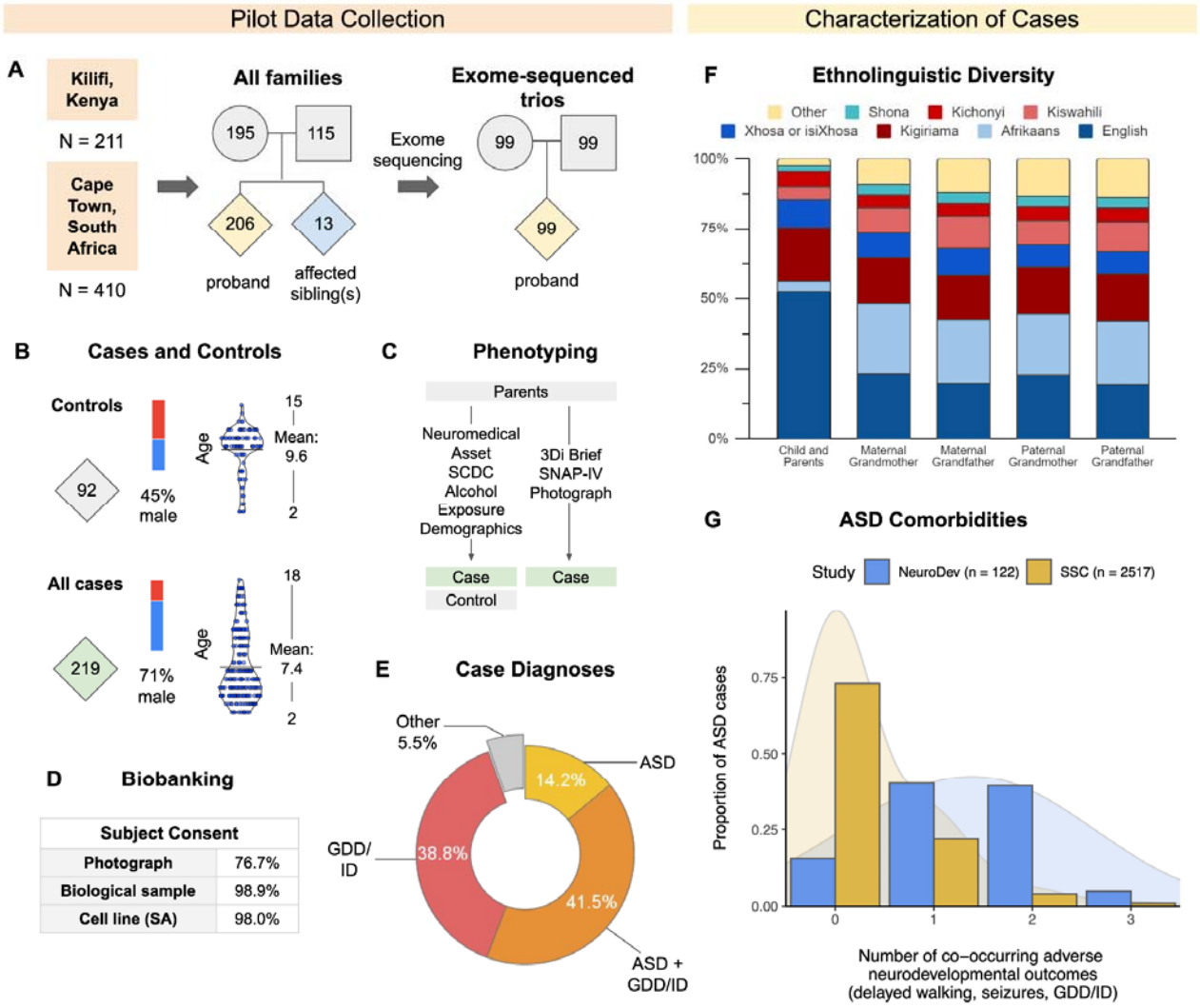

Figure 1. Overview of the NeuroDev Trio Pilot. (A) Data collection and exome sequencing description. (B) Distribution of sex, age, and consent subtypes across cases and controls. (C) Assessments administered to cases and for controls. (D) Subject consent rates for biobanking. (E) Diagnostic profiles of ascertained samples Diagnoses within "other" include specific learning disabilities, communication disorders, and/or ADHD. (F) Primary language breakdown of the 219 NeuroDev case families ("Child and Parents") and their grandparents. "Other" spans 24 languages spoken by a small fraction of families. (G) Number of co-occurring adverse neurodevelopmental outcomes among the NeuroDev and SSC ASD cases. Values of 0, 1, 2, or 3 indicate the total number of the following outcomes present: delayed walking, seizures, or GDD/ID. A score of 3 indicates that all three symptoms are present.

As anticipated, most children with ASD in this initial group also met criteria for GDD/ID (74.6%). This is consistent with other descriptions of clinic-based cohorts of children with ASD from

resource-limited environments[14–16]. The scarcity of neuropsychiatric specialists and medical resources in the African region contributes to later diagnosis or lack of access to services for children with milder ASD symptoms and minimal cognitive impairment.[17–19]

We compared the prevalence of GDD/ID, delayed walking (after 18 months of age) and parent-reported seizures between child cases with ASD in NeuroDev and those in the Simons Simplex Collection (SSC), a large United States-based cohort (Figure 1G). In contrast to cases ascertained through the SSC and other similar studies, most NeuroDev ASD cases in this early group present with at least one of these co-occurring adverse neurological or neurodevelopmental outcomes: i) seizures, ii) GDD/ID, and/or iii) walking later than 18 months (84.4%). The average number of co-occurring adverse outcomes in the NeuroDev study was 1.34 (N = 122), whereas this number was .33 for SSC (N = 2517, p = 2.2e-16). The difference in the number of co-occurring symptoms likely reflects differences in research setting and case ascertainment, as described above.

The rate of co-occurring neurodevelopmental conditions in individuals with ASD is associated with average genetic architecture.[20,21] Most significantly, the case rate of *de novo* protein truncating variants (PTVs) in constrained genes increases with the number of co-occurring adverse neurodevelopmental outcomes. From the data in Weiner et al., we conservatively expect a 50% increase in the observed rates of *de novo* PTVs in constrained genes in NeuroDev ASD cases relative to SSC and similar cohorts.[21]

The initial 219 cases experienced high levels of speech delay, with only 36% of cases meeting criteria for fluent speech. The high rates of delay are anticipated as a result of our ascertainment strategy, which included recruitment of cases from neurodevelopmental clinics and special needs schools. Those with a comorbid diagnosis of ASD and ID/GDD had the lowest speech level, with 76% of families reporting the use of single words or less. Many case children also experienced challenges in attempting the RPM. Among all cases, only 66% of all Standard RPM testers (age 12 or higher) and 44% of all Colored RPM testers (ages 6-11) were able to fully complete age-appropriate versions of the assessment (SI Figure 1, SI Table 1). Among Standard RPM testers that could not complete the assessment, approximately half were able to complete the Colored version of the RPM, and the rest completed the Molteno. Taken together, these findings suggest high levels of developmental delay in NeuroDev cases as compared to other NDD cohorts that are widely available.

Genetic analyses of the trio pilot data

Of the 99 parent-child trios included in the trio sequencing analysis, 75 were from South Africa and 24 were from Kenya. As of the present analysis, 22 pathogenic or likely pathogenic variants have been identified in those families. We have also identified 7 novel genetic variants of unknown significance, matched with other cases through MatchMaker Exchange (MME). A detailed description of the sequencing and data analysis approach can be found in the STAR Methods. In brief, exome sequencing was performed on each of the trios, and the data was uploaded to the *seqr* platform for analysis.[22] For each trio, the ES data was analyzed to identify

causal variants in known NDD genes or very rare, likely causal variants in genes not previously associated with disease.

A total of 13/75 (17.3%) South African cases were solved with pathogenic or likely pathogenic variants in genes with an existing disease association in OMIM.[23] Of these, 5 cases had a single nucleotide variant (SNV) and 1 case had an insertion or a deletion (indel) (Table 1, Supplementary Table 3). An additional 7 cases had structural variants that included a known OMIM gene (Table 2, Supplementary Table 4). Nearly all of the events were *de novo*, with the exception of one structural variant that was paternally inherited (unknown paternal history of psychiatric diagnoses or NDDs).

A total of 9/24 (37.5%) Kenyan cases were solved. Of these, we found 6 cases had a pathogenic or likely pathogenic SNV in genes previously associated with an NDD (Table 1) and 3 cases had a structural variant including a known OMIM gene (Table 2, Supplementary Table 4).

Table 1. SNV/indel pathogenic and likely pathogenic findings in known OMIM disease genes.

Individual	Sex	Gene	Variant	Event	Additional Phenotypes of Interest	OMIM #
South Africa						
GDD/ID						
859-25391305	Male	*CREBBP (NM_004380.3)*	c.3914+3G>T	extended splice site	Craniofacial dysmorphia, abnormalities of the genital system, short stature	#180849
859-88385997	Male	*CREBBP (NM_004380.3)*	c.5558A>C, (p.Gln1853Pro)	missense	Hearing impairment, visual impairment, craniofacial dysmorphia, hypertonia	#618332
859-49145346	Female	*DDX3X (NM_001356.5)*	c.599A>G, (p.Tyr200Cys)	missense	Small for gestational age, abnormal facial shape	#300958
859-90780619	Male	*IRF2BPL (NM_024496.4)*	c.2137del, (p.Leu713SerfsTer54)	frameshift	–	#618088
GDD/ID, ASD						
859-21383847	Male	*SCN2A (NM_001040142.2)*	c.2877C>A, (p.Cys959Ter)	nonsense	–	#613721
GDD/ID, ADHD						
859-98643450	Male	*SYNGAP1 (NM_006772.3)*	c.3795-1G>A	essential splice site	Large birth length	#612621
Kenya						
GDD/ID						
860-61235417	Female	*BCL11B (NM_138576.4)*	c.1535_1536del, (p.Ala512GlyfsTer4)	frameshift	Visual impairment, low BMI	#618092
860-26955427	Female	*DDX3X (NM_001356.5)*	c.1582C>T, (p.Arg528Cys)	missense	Small for gestational age, muscle weakness	#300958
860-71034936	Male	*TLK2 (NM_001284333.2)*	c.1655T>C, (p.Leu552Pro)	missense	Short stature	#618050
GDD/ID, ASD						
860-73911739	Female	*DDX3X (NM_001356.5)*	c.894C>A, (p.Cys298Ter)	nonsense	Short stature, abnormality of facial musculature	#300958
860-31257300	Female	*ZBTB18 (NM_205768.3)*	c.204_205del, (p.Asp70HisfsTer19)	frameshift	–	#612337
GDD/ID, ADHD						
860-33192107	Male	*MBD5 (NM_018328.4)*	c.4170G>A, (p.Trp1390Ter)	nonsense	–	#156200

GDD, global developmental delay; ID, intellectual disability; ASD, autism spectrum disorder; ADHD, attention deficit hyperactivity disorder. Age at recruitment is reported. All variants were *de novo.*

Table 2. South African and Kenyan structural variant pathogenic and likely pathogenic findings.

Individual	Sex	Genomic Region (GRCh38)	Event	Associated Condition	Additional Phenotypes
South Africa					
GDD/ID					
859-10314801	Male	chr3:13371737-20095506	deletion (6.7 Mb)	3p deletion syndrome (#613792)	Small for gestational age, microcephaly
859-97835206	Female	chr6:115941808-133892653	deletion (18 Mb)	Interstitial 6q microdeletion syndrome	Ventricular septal defect, seizure, meningitis, camptodactyly
859-44770029	Female	chr18:61490305-80247612	deletion (18 Mb)	Chromosome 18q deletion syndrome (#601808)	Visual impairment, low birth length, abnormal facial shape, hypotonia
859-41524687	Female	chr22:18985739-21081116	duplication (2.1 Mb)	22q11.2 duplication syndrome (#608363)	High birth length, macrocephaly
GDD/ID, ASD					
859-50205427	Female	chr15:22810652-29822566	triplication (7.0 Mb)	15q11-q13 duplication syndrome (#608636)	Small for gestational age
859-22089821	Male	chr15:30626003-32111997	deletion (1.5 Mb)	15q13.3 microdeletion syndrome (#612001)	Small for gestational age, microcephaly
859-85638884	Male	chr16:29663598-30188229	deletion (0.53 Mb)	16p11.2 deletion syndrome (#611913)	–
Kenya					
GDD/ID					
860-31775019	Male	22:18985739-21081116	deletion (2.1 Mb)	DiGeorge syndrome (#188400)	Abnormality of facial musculature, seizure
860-45665569	Male	22:18985739-21081116	deletion (2.1 Mb)	DiGeorge syndrome (#188400)	–
860-94211640	Male	22:49883237-50740457	duplication (0.85 Mb)	22q13 duplication syndrome (#615538)	–

GDD, global developmental delay; ID, intellectual disability; ASD, autism spectrum disorder. Age at recruitment is reported. All variants were *de novo* except in subject 859-22089821 where the variant was paternally inherited.

In addition to the solved cases, 7 variants of uncertain significance (VUS) in novel genes were matched on MatchMaker Exchange (MME; Table 3, Supplementary Table 5). By matching cases of similar phenotypic and genotypic profiles, MME provides a rapid, systematic approach to rare disease gene discovery.[24] The genes containing these variants are not yet listed in OMIM, and reflect novel disease discoveries arising from 7 of the 99 trio pilot participants. All 7

variants will be included in collaborative case series reports on the genes of interest. To date, case studies on three of the newly identified genes – *AGO1*, *CACNA1C*, and *CACNA1E* – have been published.[25–27] NeuroDev participants comprised the only geographic African cases in any of the case series reports, and NeuroDev will be the first African NDD cohort to contribute, at scale, to rare disease discovery activities. As described by the NHGRI Atlas of Human Malformations initiative, syndromic NDDs often vary in their phenotypic presentation between ancestral groups[28], a phenomenon particularly well documented with regard to canonical facial features. Further analysis of how phenotypic features differ between ancestral groups will help clarify which features are in fact "core" to a genetic syndrome, and which vary based on ancestral group or environment.

Table 3. Variants of unknown significance in novel candidate genes.

Individual	Sex	Gene	Variant	Event	Additional Phenotypes	Matches
South Africa						
GDD/ID						
859-44206536	Male	*AGO1* *(NM_012199.5)*	c.971C>T, (p.Pro324Leu)	missense	–	26
859-81577625	Male	*PPP2R5C* *(NM_001161725.1)*	c.254_259del, (p.Asp85_Phe86del)	inframe deletion	Seizure	11
ASD						
859-33526476	Female	*CACNA1C (NM_001129827.2)	c.4129dup, (p.Arg1377ProfsTer61)	frameshift	–	25
859-76545750	Male	*CACNA1E (NM_001205293.3)	c.3422+1G>A	essential splice site	Tall stature, proximal amyotrophy	5
859-40050374	Female	*MYH10* *(NM_001256012.3)*	c.2555G>A, (p.Arg852Gln)	missense	Macrocephaly, high BMI	14
Kenya						
GDD/ID						
860-46028963	Male	*MAPK1* *(NM_002745.5)*	c.952G>A, (p.Asp318Asn)	missense	Short stature	7
GDD/ID, CD						
860-89059068	Female	*SF1* *(NM_001178030.1)*	c.737C>T, (p.Pro246Leu)	missense	Visual impairment	4

GDD, global developmental delay; ID, intellectual disability; ASD, autism spectrum disorder; CD, communication disorder. Age at recruitment is reported. All variants were *de novo*.
***** = not novel genes, but phenotype expansions

Discussion

People of African ancestry have been grossly underrepresented in genetic studies, across domains and disciplines.[29,30] In aggregate, this is most visible through the constitution of large genetic databases like gnomAD, in which only 14% of individuals (N ≈ 27,000) have some African ancestry [31,32]. As no sequenced cohorts from the African continent are currently present in gnomAD, the great majority of those 27,000 individuals are American, and typically have a mixture of West African and European ancestry.[33,34] If individuals of African ancestry remain underrepresented in genetic research, they will continue to be less likely to receive accurate genetic diagnoses and less likely to benefit from advances in genomic science and medicine.[35–38]

The NeuroDev project was built to address this representation gap and its scientific, medical, and ethical consequences. Those living in Africa have more genetic variability than any other human ancestral group, rendering their underrepresentation a substantial barrier to human genome characterization and scientific equity.[29] There are many reasons that NDD collections have historically been conducted in the United States and Western Europe, chief among them better access to project funding and infrastructural support. The lack of precedent produced common concern about NeuroDev's feasibility, particularly with regard to the planned case collection rate, and the ambitious case characterization schedule. We hope that the results described here will increase the research community's confidence in large scale data collection efforts in underrepresented populations, and that many other studies will be able to contribute to greater African representation in NDD genomics.

We also hope that the high rate of participant family interest in receiving genetic results will shed light on the need for consistent access to genetic testing services. To date, all NeuroDev participants in South Africa readily consented to have genetic findings related to their child's NDD reported back to them by a clinician. Similarly, NeuroDev South Africa observed a very high (98%) rate of consent to generation and sharing of stem cell lines in addition to sharing DNA. This provides us with the opportunity to contribute to much needed diversity to global stem cell collections, which are also heavily biased towards European ancestry.[39] This high degree of participant enthusiasm in South Africa has led to investigations of similar possibilities in Kenya. The NeuroDev Kenya team has recently been funded by a Fogarty award to support, in part, community engagement on the ethics of cell line generation for the cohort in Kilifi.

We saw a remarkably high rate of novel variant discovery in the NeuroDev Pilot cohort, with variants in genes not previously implicated in NDDs identified in 7/99 trio families. By comparison, a meta-analysis of 2,104 trio families of probands with ID in the Netherlands yielded 10 candidate genes including 4 genes whose association with ID, at the time of publication, were novel.[40] The high rate of novel discovery in NeuroDev is likely related to its ascertainment approach, and the phenotypic profile of its participants.

While NeuroDev provides in-depth phenotypic characterization of its subjects, this data may still reflect regional differences. Kilifi is a rural coastal town with an agriculture-based economy, with an absolute poverty rate of 46.4% and limited access to school-based education.[41] In contrast, Cape Town is a large urban setting and is among the wealthiest cities in Africa, conferring relative educational and resource advantages. As many behavioral measures are known to be tied to socioeconomic status or material education, our behavioral outcomes likely reflect differences in regional context. Thus, we encourage caution when making comparisons of the outcomes of phenotypic assessments, such as Raven's scores, across the two sites.

The four-year target sample for NeuroDev is 1800 cases and 1800 controls, the largest study to date investigating NDDs on the African continent. The NeuroDev project will, in full, create a public resource for medical genetics research that includes, for thousands of African individuals: genome-wide common variant data; exome sequencing data; comprehensive phenotypic data including detailed cognitive, behavioral, and health information; cell lines (lymphoblastoid and induced pluripotent cell lines); and photographs capturing dysmorphic features. This critically needed line of work will help address the scientific, medical, and ethical consequences of the genomic research representation gap.

STAR Methods

NeuroDev Data Pipeline

Data is collected on tablets using the REDCap Mobile App during assessments. Every participant is assigned a subject ID and has an intake form filled out. Based on this information, branching logic programmed into REDCap determines which tools and fields populate for the assessor to complete. At the end of an assessment day, data from the tablets is uploaded to the NeuroDev project on the REDCap servers. The team at the Broad Institute then uses the *scred* package – a Python implementation of the REDCap API – to pull raw records by subject ID and assign codes to missing data to differentiate between truly missing fields and non-applicable fields based on the REDCap branching logic. A Python sync engine then organizes the record data by assessment tool, scores assessments like the Raven's Progressive Matrices and 3Di, and pushes it to a MySQL database hosted on the Google Cloud. The sync procedure is generally run on a weekly basis to keep the database current. All summary and quality-control (QC) reporting is derived from either direct SQL queries to the database or export to pandas dataframes for analysis with Python scripts.

The QC reporting process relies on querying data from the Broad Database and REDCap logging files to produce a comprehensive report of missing, overwritten, or inconsistent participant data. Run on either a date range or a list of specific subject IDs, the program produces summary reports on any of five distinct metrics: completion of REDCap forms, missing required data fields (regardless of whether a form was marked complete), comments entered by assessors at the sites, logical cross-checks on completed data, and fields found to be overwritten in the REDCap log. Subject and site IDs as well as interview times and assessor names are included on each of these reports so the Broad team can easily identify both single issues and larger patterns. Run and reviewed weekly, the program improves data visibility and allows consistent feedback to the sites on improving or correcting collection procedures.

NeuroDev Subject Recruitment

NeuroDev participants were recruited from two sites, Red Cross War Memorial Hospital in Cape Town, South Africa and KEMRI-Wellcome Trust Research Program in Kilifi, Kenya. Participants were recruited from previous studies, specialized clinics and special schools in Kilifi County, while participants in Cape Town were recruited from only developmental clinics. Cases and affected siblings included in the study had clinical diagnosis of a neurodevelopmental disorder, were within the specified age range (2-18 years old), and willing to participate. Cases were excluded if they had a co-occuring primary neuro-motor condition such as cerebral palsy or Downs Syndrome. Controls were included in the study if they did not have a diagnosis of a neurodevelopmental disorder, were within the study age range and were matched according to catchment area, ancestry and age. Written consent was sought for participation of the child controls and child cases from their parents or caregivers, and additional consent was sought from parents of the cases for their own participation in the study. Ethical approval was sought in the site institutional review boards as well as at the Harvard T.H Chan School of Public Health.

In Kilifi, Kenya, approval was granted by the Scientific Ethics and Review Unit (KEMRI/SERU/CGMR-C/104/3629) and Health Research Ethics Committee (HREC REF:810/2016).

Phenotypic Data Analysis Methods

Self-reports of ancestry and language were collected as part of the Demographics tool on REDCap. To assess developmental or intellectual delays, we administered either the Molteno or the Raven's Progressive Matrices (RPM) to their respective age groups and measured rates of completion. Successful completion of the Molteno was defined as the completion of at least 2 of the 4 domains implicated in the assessment. Completion of the RPM was defined as being able to complete all questions on the assessment. Proportions of testers who successfully completed the respective tests at age ranges of 0-5 (Molteno), 6-11 (Colored RPM), and greater than 12 (Standard RPM) was computed and reported. For subjects who could not complete age-appropriate assessments, the rates of completion of tests below their age level was reported for each age group.

Neurodevelopmental diagnosis of cases was collected through the Neuromedical Assessment tool. The cases were placed into four distinct categories: ASD, GDD/ID, ASD and GDD/ID, and other diagnoses (which included ADHD, communication disorder, and specific learning disorders). Children that met DSM-5 criteria for either GDD or ID or were diagnosed with at least borderline delay from the Molteno were included in the GDD/ID category. For each of these four categories, we assessed the proportions of those with fluent speech levels. Speech fluency was determined by the 3Di, in which an assessor rated the child's speech on a scale of no words, single words, multiple words, and fluent speech based on parent interview.

The Simons Simplex Collection (SSC) data set was used to compare phenotypic outcomes of the 122 cases with ASD ascertained in NeuroDev. The SSC consists of a deeply phenotyped sample of more than 2500 families with a child diagnosed with an ASD in the United States.[5] We grouped cases with ASD from both datasets based on the number of adverse co-occurring neurological and developmental outcomes, including ID, a positively associated history of seizures, and motor delays (defined as either a gross motor diagnosis from the Molteno, or an age at first steps higher than 18 months). In SSC, ID was defined as cases with an IQ < 70. From the symptom distributions within each study, the average number of co-occurring symptoms was computed, and the difference between these means was assessed using a Mann-Whitney U test.

Exome Sequencing & Data Processing Methods

Data generation and analysis were done in collaboration with the Broad Institute Center for Mendelian Genomics, with sequencing performed at the Genomics Platform and data processing at the Data Sciences Platform of the Broad Institute of MIT and Harvard. Libraries from DNA samples were created with a Twist exome capture (37 Mb target) and sequenced

(150 bp paired reads) to >85% of targets at >20x, comparable to ~55x mean coverage. Sample identity quality assurance checks were performed on each sample. The exome sequencing data was de-multiplexed and each sample's sequence data aggregated into a single Picard CRAM file. Exome data was processed through a pipeline based on Picard, using base quality score recalibration and local realignment at known indels, aligned to the human genome build 38 using BWA, and jointly analyzed for single nucleotide variants (SNVs) and insertions/deletions (indels) using Genome Analysis Toolkit (GATK) Haplotype Caller package version 4.0.10.1. After variant calling, sex, ancestry and relatedness to other samples will be inferred use the CMG sample QC pipeline and compared to sample metadata to identify and correct sample swaps. Basic functional annotation will be performed using Variant Effect Predictor (VEP), and then the joint variant call file will be uploaded to the *seqr* platform for further annotation and analysis.

Copy-number variants (CNVs) were discovered from the exome sequencing data following GATK-gCNV best practices. Read coverage was calculated for each exome using GATK CollectReadCounts. After coverage collection, all samples were subdivided into batches for gCNV model training and execution; these batches were determined based on a principal components analysis (PCA) of sequencing read counts. After batching, one gCNV model was trained per batch using GATK GermlineCNVCaller on a subset of training samples, and the trained model was then applied to call CNVs for each sample per batch. Finally, all raw CNVs were aggregated and post-processed using quality- and frequency-based filtering to produce a final CNV callset.

Exome Sequencing Data Analysis Process

Upon completion of data generation, both the SNV/indel and CNV callsets were uploaded to *seqr*, the centralized genomic analysis platform used by the Broad Institute's Center for Mendelian Genomics (CMG). The CMG analysis team deployed a standard analysis protocol across all NeuroDev trios to identify causal variants in known neurodevelopmental disease genes, and, when possible, to discover novel candidate genes underlying these conditions. The first round of analysis consisted of a review of variants from a *de novo*/dominant and a recessive search. The *de novo*/dominant search filters for variants present in affected family members and absent from unaffected family members. The variant types returned from this standard search include deletions, duplications, protein truncating variants, and missense variants that have an allele frequency <0.1% across population databases (gnomAD v2/v3/SV, 1000 Genomes, ExAC, and TopMed) and <1% in the CMG internal rare disease dataset, that pass QC, and that have a GQ $\geq$20 and an allele balance $\geq$0.2. The recessive search returns homozygous recessive, compound heterozygous, and X-linked recessive variants in affected individuals, considering phasing for any available parents. It searches for biallelic variants across deletions, duplications, protein truncating variants, or missense variants that have an allele frequency <1% across population databases and <3% in the CMG internal rare disease dataset, that pass QC, and that have a GQ $\geq$20 and an allele balance of $\geq$0.2. If candidate variants were not identified after the first round of analysis, the search criteria was adjusted to

include additional variant annotation types (synonymous variants, extended splice site variants, and 5' and 3' UTR variants), and quality parameters were relaxed to allow for the review of indels that did not pass QC.

Potential causal variants were subjected to rigorous evaluation of the evidence for pathogenicity following criteria established by the American College of Medical Genetics and Association for Molecular Pathology.[42] The study's primary focus was on well-established disease genes, using information drawn from a variety of sources such as OMIM. The phenotype data of NeuroDev participants was assessed for possible consistencies with the previously reported clinical presentations associated with known disease genes, bearing in mind potential deviations from expectation due to ancestry.

Novel candidate genes that are not yet associated with a well-established human disease were carefully evaluated using a variety of sources of evidence, including constraint scores, transcript and protein expression databases, model organism data, and a review of the available literature. All variants identified in candidate disease genes were entered into the Matchmaker Exchange (MME) network through *seqr* in order to identify additional cases with variants in the same candidate genes, to help better characterize potential novel gene-disease relationships.

Supplementary Tables and Figures

Supplementary Table 1: Case data completeness for trio pilot
Supplementary Table 2: Parent data completeness for trio pilot
Supplementary Table 3: Reportable SNV/indel pathogenic and likely pathogenic findings in known OMIM disease genes
Supplementary Table 4: Structural variant pathogenic and likely pathogenic findings
Supplementary Table 5: Variants of unknown significance in novel candidate genes
Supplementary Figure 1: Flowchart of administration and completion rates for the Raven's Progressive Matrices and Molteno assessments
Supplementary Figure 2: Breakdown of represented ethnolinguistic groups among the 219 NeuroDev cases

Funding

NeuroDev is supported by the Stanley Center for Psychiatric Research at the Broad Institute, a grant from SFARI (704413, E.B.R), and by the National Institute of Mental Health of the National Institutes of Health under Award Number U01MH119689. Research reported in this publication was also supported by the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health under Award Number R01HD102975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Sequencing was provided by the Broad Institute of MIT and Harvard Center for Mendelian Genomics (Broad CMG) and was funded by the National Human Genome Research Institute, the National Eye Institute, and the National Heart, Lung and Blood Institute grant UM1 HG008900 and in part by National Human Genome Research Institute grant R01 HG009141.

Acknowledgements

We are extremely grateful to the family members for participating in this research. We are grateful to the clinical laboratories and biobank teams at KEMRI-Wellcome Trust and the University of Cape Town, the community liaison group and neuro-epilepsy clinic at the KEMRI-Wellcome Trust and developmental and allied clinics at Red Cross Memorial Hospital. We acknowledge James Swanson, Edith Nolan and William Pelham and team for the use of the SNAP-IV, David Skuse, Richard Warrington, Will Mandy and team for the use of the 3Di and SCDC, the ASEBA team for the use of the CBCL, Christopher Molteno and team for the use of the Molteno Adapted Scales and John C Raven, John H Court and team for the use of the Ravens Progressive Matrices.

Author Contributions

E.R, K.A.D, A.A, C.N designed the study. The NeuroDev Project members, including P.K, B.C, E.E, M.K, K.M, were involved in data collection, H.A.K, E.O, J.A, S.Br, N.B, C.K, B.M were involved in the curation and analysis of the data, P.K, H.A.K, B.C, E.O, E.R wrote the manuscript with input from all authors, P.K, B.C, E.E, A.G, K.M were involved in project administration and C.AT, S.Ba, H.B, A.L, D.G.M, A.S.J, M.B.S, M.E.T, VdM, CvdM, C.N, A.O.L, A.A, K.A.D and E.R supervised various aspects of the project and the core project teams.

The NeuroDev Project members (current and previous) in addition to the named authors include: Aleya Zulfikar Remtullah, Alex Macharia, Alfred Ngombo, Ann Karanu, Beatrice Mkubwa, Carmen Swanepoel, Claire Fourie, Constance Rehema, Deepika Goolab, Dorcas Kamuya, Dorothy Chepkirui, Este Sauerman, Eunice Chepkemoi, Fagri February, Fatima Khan, Felicita Omari, Gina Itzikowitz, Javan Nyale, Jantina de Vries, Jess Ringshaw, Johnstone Makale, Judy Tumaini, Kegan Chase, Lizette Rooi, Moses Mangi, Moses Mosobo, Megan Page, Nicole Mciver, Nonceba Ngubo, Paul Mwangi, Pauline Samia, Phatiswa Lopoleng Ranyana, Racheal Mapenzi, Rachel Odhiambo, Raphaela Itzikowitz, Rene Lepore, Rizqa Sulaiman-Baradien, Samuel Mwasambu, Serini Murugasen, Shaheen Sayed, Susan Wamithi, Tabith Shali, Thandi Xintolo, and Zandre Bruwer.

References

1. de Menil V, Hoogenhout M, Kipkemoi P, Kamuya D, Eastman E, Galvin A, et al. The NeuroDev Study: Phenotypic and Genetic Characterization of Neurodevelopmental Disorders in Kenya and South Africa. Neuron. 2019 Jan 2;101(1):15–9.
2. Skuse DH, Mandy WPL, Scourfield J. Measuring autistic traits: heritability, reliability and validity of the Social and Communication Disorders Checklist. Br J Psychiatry. 2005 Dec;187(6):568–72.
3. Skuse D, Warrington R, Bishop D, Chowdhury U, Lau J, Mandy W, et al. The Developmental, Dimensional and Diagnostic Interview (3di): A Novel Computerized Assessment for Autism Spectrum Disorders. J Am Acad Child Adolesc Psychiatry. 2004;43(5):548–58.
4. Swanson JM, Schuck S, Porter MM, Carlson C, Hartman CA, Sergeant JA, et al. Categorical and Dimensional Definitions and Evaluations of Symptoms of ADHD: History of the SNAP and the SWAN Rating Scales. Int J Educ Psychol Assess. 2012 Apr;10(1):51–

70.
5. Fischbach GD, Lord C. The simons simplex collection: A resource for identification of autism genetic risk factors. Neuron. 2010;68(2):192–5.
6. Robinson EB, Samocha KE, Kosmicki JA, McGrath L, Neale BM, Perlis RH, et al. Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. Proc Natl Acad Sci. 2014 Oct 21;111(42):15161–5.
7. Raven J. The Raven's progressive matrices: change and stability over culture and time. Cogn Psychol. 2000 Aug;41(1):1–48.
8. Charman T, Jones CRG, Pickles A, Simonoff E, Baird G, Happé F. Defining the cognitive phenotype of autism. Brain Res. 2011 Mar 22;1380:10–21.
9. DiStefano C, Sadhwani A, Wheeler AC. Comprehensive Assessment of Individuals With Significant Levels of Intellectual Disability: Challenges, Strategies, and Future Directions. Am J Intellect Dev Disabil. 2020 Nov 19;125(6):434–48.
10. Jeste DV, Palmer BW, Appelbaum PS, Golshan S, Glorioso D, Dunn LB, et al. A new brief instrument for assessing decisional capacity for clinical research. Arch Gen Psychiatry. 2007 Aug;64(8):966–74.
11. Seaman JB, Terhorst L, Gentry A, Hunsaker A, Parker LS, Lingler JH. Psychometric Properties of a Decisional Capacity Screening Tool for Individuals Contemplating Participation in Alzheimer's disease Research. J Alzheimers Dis JAD. 2015;46(1):1–9.
12. Campbell MM, Susser E, Mall S, Mqulwana SG, Mndini MM, Ntola OA, et al. Using iterative learning to improve understanding during the informed consent process in a South African psychiatric genomics study. PLOS ONE. 2017 Nov 29;12(11):e0188466.
13. Achenbach TM, Dumenci L, Rescorla LA. Ratings of relations between DSM-IV diagnostic categories and items of the CBCL/6-18, TRF, and YSR. Burlingt VT Univ Vt. 2001;1–9.
14. Bakare MO, Ebigbo PO, Ubochi VN. Prevalence of autism spectrum disorder among Nigerian children with intellectual disability: a stopgap assessment. J Health Care Poor Underserved. 2012 May;23(2):513–8.
15. Mpaka DM, Okitundu DLEA, Ndjukendi AO, N'situ AM, Kinsala SY, Mukau JE, et al. Prevalence and comorbidities of autism among children referred to the outpatient clinics for neurodevelopmental disorders. Pan Afr Med J [Internet]. 2016 Oct 17 [cited 2020 Nov 25];25. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5324163/
16. Bello-Mojeed MA, Omigbodun OO, Bakare MO, Adewuya AO. Pattern of impairments and late diagnosis of autism spectrum disorder among a sub-Saharan African clinical population of children in Nigeria. Glob Ment Health. 2017 Mar 21;4:e5.
17. Hahler EM, Elsabbagh M. Autism: A global perspective. Curr Dev Disord Rep. 2015;2(1):58–64.
18. Mazurek MO, Handen BL, Wodka EL, Nowinski L, Butter E, Engelhardt CR. Age at First Autism Spectrum Disorder Diagnosis: The Role of Birth Cohort, Demographic Factors, and Clinical Features. J Dev Behav Pediatr. 2014 Dec;35(9):561–9.
19. Ruparelia K, Abubakar A, Badoe E, Bakare M, Visser K, Chugani DC, et al. Autism spectrum disorders in Africa: current challenges in identification, assessment, and treatment: a report on the International Child Neurology Association Meeting on ASD in Africa, Ghana, April 3-5, 2014. J Child Neurol. 2016;31(8):1018–26.
20. Liao C, Moyses-Oliveira M, Esch CED, Bhavsar R, Nuttle X, Li A, et al. Transcriptional patterns of coexpression across autism risk genes converge on established and novel signatures of neurodevelopment [Internet]. medRxiv; 2022 [cited 2022 Mar 8]. p. 2022.02.28.22271620. Available from: https://www.medrxiv.org/content/10.1101/2022.02.28.22271620v1
21. Weiner DJ, Wigdor EM, Ripke S, Walters RK, Kosmicki JA, Grove J, et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. Nat Genet. 2017 Jul;49(7):978–85.

22. Pais LS, Snow H, Weisburd B, Zhang S, Baxter SM, DiTroia S, et al. seqr: A web-based analysis and collaboration tool for rare disease genomics. Hum Mutat. 2022 Mar 9;
23. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015 Jan 28;43(Database issue):D789–98.
24. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al. The Matchmaker Exchange: a platform for rare disease gene discovery. Hum Mutat. 2015;36(10):915–21.
25. Rodan LH, Spillmann RC, Kurata HT, Lamothe SM, Maghera J, Jamra RA, et al. Phenotypic expansion of CACNA1C-associated disorders to include isolated neurological manifestations. Genet Med Off J Am Coll Med Genet. 2021 Oct;23(10):1922–32.
26. Royer-Bertrand B, Jequier Gygax M, Cisarova K, Rosenfeld JA, Bassetti JA, Moldovan O, et al. De novo variants in CACNA1E found in patients with intellectual disability, developmental regression and social cognition deficit but no seizures. Mol Autism. 2021 Oct 26;12(1):69.
27. Schalk A, Cousin MA, Dsouza NR, Challman TD, Wain KE, Powis Z, et al. De novo coding variants in the AGO1 gene cause a neurodevelopmental disorder with intellectual disability. J Med Genet. 2021 Dec 15;jmedgenet-2021-107751.
28. Muenke M, Adeyemo A, Kruszka P. An electronic atlas of human malformation syndromes in diverse populations. Genet Med. 2016 Nov;18(11):1085–7.
29. Martin AR, Teferra S, Möller M, Hoal EG, Daly MJ. The critical needs and challenges for genetic architecture studies in Africa. Curr Opin Genet Dev. 2018 Dec;53:113–20.
30. Peterson RE, Kuchenbaecker K, Walters RK, Chen CY, Popejoy AB, Periyasamy S, et al. Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. Cell. 2019 Oct 17;179(3):589–603.
31. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: Lessons from gnomAD. Hum Mutat. 2021 Dec 2;
32. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43.
33. Mathias RA, Taub MA, Gignoux CR, Fu W, Musharoff S, O'Connor TD, et al. A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. Nat Commun. 2016 Oct 11;7:12522.
34. Zakharia F, Basu A, Absher D, Assimes TL, Go AS, Hlatky MA, et al. Characterizing the admixed African ancestry of African Americans. Genome Biol. 2009 Dec 22;10(12):R141.
35. Caswell-Jin JL, Gupta T, Hall E, Petrovchich IM, Mills MA, Kingham KE, et al. Racial/ethnic differences in multiple-gene sequencing results for hereditary cancer risk. Genet Med Off J Am Coll Med Genet. 2018 Feb;20(2):234–9.
36. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 18;536(7616):285–91.
37. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al. Genetic Misdiagnoses and the Potential for Health Disparities. N Engl J Med. 2016;375(7):655–65.
38. Popejoy AB, Ritter DI, Crooks K, Currey E, Fullerton SM, Hindorff LA, et al. The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. Hum Mutat. 2018 Nov;39(11):1713–20.
39. Negoro T, Okura H, Matsuyama A. Induced Pluripotent Stem Cells: Global Research Trends. BioResearch Open Access. 2017 Jun 1;6(1):63–73.
40. Lelieveld SH, Reijnders MRF, Pfundt R, Yntema HG, Kamsteeg EJ, de Vries P, et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability.

Nat Neurosci. 2016 Sep;19(9):1194–6.

41. World Bank. Kenya Gender and Poverty Assessment 2015-2016: Reflecting on a Decade of Progress and the Road Ahead. World Bank; 2018.
42. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015 May 1;17(5):405–24.