HGG Advances

ARTICLE

Genome sequencing for the diagnosis of rare disorders: The Brazilian Rare Genomes Project

The Brazilian Rare Genomes Project Consortium[1,2,*]

Summary

Genome sequencing (GS) has emerged as a transformative tool in the diagnosis of rare diseases with complex phenotypes. This technology uncovers structural, intronic, non-coding, and mitochondrial variants that traditional methods might miss, thereby facilitating the understanding of the underlying genomic basis of human disorders. We enrolled 10,305 patients with suspected rare diseases or hereditary cancer risk syndromes from 21 centers throughout Brazil. Their genomes were sequenced with short, paired-end reads, and diagnostic reports were provided for 9,448 of these patients. The overall diagnostic yield was 35.6%, and 4.6% of all positive reports had GS-exclusive findings (e.g., short copy-number variants overlapping fewer than three exons, deep intronic variants, short tandem repeat expansions, and mitochondrial structural variants—usually not detected by other diagnostic tests such as exome sequencing). Preliminary analysis of transcriptome sequencing (TS) or long-read GS combined with GS interpretation provided a small but welcome improvement in diagnostic yield (0.1% and 1.0% of positive reports, respectively). Almost 3,200 variant/phenotype interpretations were submitted to ClinVar. GS is proving to be an invaluable resource for shortening the diagnostic odyssey of patients with rare diseases, providing crucial genomic diagnostics, and enriching genetic databases with variant interpretations from underrepresented populations. Therefore, GS has the potential to significantly enhance the precision of healthcare in genetically diverse populations.

Introduction

Rare diseases have an estimated prevalence ranging between 1.5% and 6.2% in the general population,[1] which translates to between 3.2 and 13.2 million affected individuals in Brazil alone. Brazil, notable for its universal health care system (the Sistema Único de Saúde [SUS]),[2] has made significant strides in improving rare disease diagnosis since 2014. Brazilian healthcare centers now offer newborn metabolic screening and genetic tests, including karyotyping, fluorescence *in situ* hybridization (FISH), multiplex ligation-dependent probe amplification (MLPA), and array-based comparative genomic hybridization (array-CGH).

Genome sequencing (GS) and transcriptome sequencing (TS) have emerged as transformative tools in the diagnosis of rare diseases with complex phenotypes. GS enables deep analysis of the genome, uncovering structural, intronic, non-coding, repeat-expansion,[3] and mitochondrial variants (since it does not rely on off-target reads[4]) that traditional methods might miss, whereas TS provides a comprehensive view of RNA expression, facilitating the understanding of gene function and regulation.[5]

Countries including the United Kingdom, Australia, Belgium, Canada, France, and many others have pioneered genetics-based precision medicine national programs,[6–8] underscoring the global recognition of these technologies' diagnostic potential. Unfortunately, Brazilian populations have been underrepresented in these international genomics initiatives.

The Brazilian Ministry of Health launched the Genomics and Precision Medicine National Program ("Brazil Genomes") in 2020. The program seeks to implement precision medicine within the Brazilian national public healthcare system. The Brazilian Rare Genomes Project (BRGP) is one of the Brazil Genomes initiatives and was created to foster the implementation of genomic medicine, thereby improving the diagnostic capacity for rare disorders.[9]

This study investigates the diagnostic power of GS within the Brazilian public healthcare system in the context of the BRGP, demonstrating its potential to significantly shorten the diagnostic odyssey for patients and enrich genetic databases with variant interpretations from underrepresented populations. The findings presented herein highlight the high value of integrating these advanced genomic tools into public healthcare. In addition, we run two pilot research subprojects to complement the GS findings. The first pilot involved TS, and the second involved long-read GS. We also report the preliminary findings of those endeavors here.

Moreover, as Brazilian populations have high genetic diversity and are underrepresented in ancestry and human genetic variation databases, a secondary objective is to assess the ancestry of the participants, generating data that will help the implementation of precision medicine in the country.

[1]Lead contact: antonio.campos@einstein.br (Antonio Victor Campos Coelho)
[2]Further details can be found in the supplemental information
*Correspondence: antonio.campos@einstein.br
https://doi.org/10.1016/j.xhgg.2026.100624.

Subjects, material, and methods

Inclusion and exclusion criteria

The overall inclusion criteria include a suspicion of a rare disease with a probable genetic etiology or hereditary cancer risk syndromes. There was no requirement for individuals to have completed specific molecular testing (e.g., microarray or targeted gene panels) prior to enrollment. Prior genetic testing history was ascertained by recruiting physicians during enrollment interviews.

For pediatric patients, parents or legal guardians were asked for consent before inclusion, with additional consent required for secondary-finding investigations. Patients were excluded from the project if they refused the invitation for inclusion, had a prior positive molecular or karyotypic diagnosis, or if the condition was likely caused by environmental factors. The study was reviewed and approved by Hospital Israelita Albert Einstein's Research Ethics Committee (São Paulo, Brazil. Protocol number: CAAE 29567220.4.1001.0071).

Patient data collection and cohort allocation

Once patients or their families provided written informed consent, their information was stored electronically in the REDCap[10,11] and PhenoTips[12] databases. Physicians then assigned each patient to one of 20 disorder cohorts based on specific criteria (neurologic disorders, neuromuscular disorders, neurocutaneous disorders, inborn errors of metabolism, skeletal disorders, hereditary cancer risk syndromes, clinically recognizable genetic syndromes, endocrine disorders, connective tissue disorders, immunologic disorders, ophthalmologic disorders, hearing or ear disorders, renal and urinary tract disorders, dermatologic disorders, hematologic disorders, cardiological disorders, pulmonary disorders, gastrointestinal disorders, vascular disorders, and critically ill patients—children aged 1 year or younger who were admitted to neonatal or pediatric intensive care units [NICUs or PICUs]). They also provided a detailed clinical summary, including indications for referral, pedigrees, therapeutic approach, and Human Phenotype Ontology (HPO) terms.[13] We prioritized the genomic analysis of critically ill individuals, aiming to report their results within 21 days of receiving the samples at our laboratory processing units.

Sample collection, GS, and bioinformatics processing

Following enrollment, patients provided blood samples for genomic DNA automated extraction. The sequencing protocol is detailed elsewhere.[9] Briefly, PCR-free GS was performed on Illumina NovaSeq 6000 equipment. Read mapping and variant calling were done using the GRCh38 build reference genome through DRAGEN v.3.6.3, v.3.8.4, or v.3.10. DRAGEN calls single-nucleotide variants (SNVs), small insertions or deletions (indels; 1–50 base pairs), and copy-number/structural variants (CNVs/SVs; more than 50 base pairs). Both nuclear and mitochondrial DNA variants were investigated. Short tandem repeats (STRs) were investigated in the following genes: *AFF2*, *AR*, *ARX*, *ATN1*, *ATXN1*, *ATXN2*, *ATXN3*, *ATXN7*, *ATXN8OS*, *ATXN10*, *C9orf72*, *CACNA1A*, *CBL*, *CNBP*, *CSTB*, *DIP2B*, *DMPK*, *FMR1*, *FOXL2*, *FXN*, *GIPC1*, *GLS*, *HOXA13*, *HOXD13*, *HTT*, *JPH3*, *LRP12*, *NIPA1*, *NOP56*, *NOTCH2NLC*, *PABPN1*, *PHOX2B*, *POLG*, *PPP2R2B*, *RFC1*, *RUNX2*, *SMN*, *SOX3*, *TBP*, *TBX1*, *TCF4*, *XYLT1*, *ZIC2*, and *ZIC3*. ExpansionHunter[14] and Broad Institute's str-analysis scripts (https://github.com/broadinstitute/str-analysis) were used to estimate STR copy numbers. Premutation and full mutation expansion calls were confirmed by visual inspection of BAM reads. If necessary, orthogonal validation of STR calls was conducted.

The VCF files have been made available for analysis on the Varstation platform (Genesis Genomics, São Paulo, Brazil),[15] which annotates the VCF files using Ensembl's Variant Effect Predictor (VEP)[16] and AnnotSV.[17,18]

Variant interpretation, variant type, and reporting definitions

GS and interpretation were performed on probands only (singleton analysis). The clinical significance of the variants was determined according to the American College of Medical Genetics and Genomics (ACMG) guidelines[19,20] and their modifications.[21] Variants in actionable genes, as defined by the ACMG, were searched in individuals who provided consent.[22] Each variant classification was completed independently by two specialists. Variant interpretation was restricted to genes cataloged in Online Mendelian Inheritance in Man (OMIM)[23,24]; it incorporated phenotype-driven analyses using HPO terms, expert gene panels, and evidence from the scientific and medical literature.

We categorized the reports into positive, inconclusive, and negative. A positive result was defined by the presence of pathogenic/likely pathogenic variant(s) that fully or partially explained the patient's phenotype. Inconclusive reports featured variant(s) that did not define a molecular diagnosis, such as a single pathogenic heterozygous variant for an autosomal-recessive (AR) disorder or a variant of uncertain significance (VUS), while negative reports had no variants deemed worth reporting. Each report contains a synopsis of the patient's phenotype in investigation; a summary of any detected variants, genome coordinates, or Human Genome Variation Society (HGVS) nomenclature expressions[25]; and an interpretation of the variants in the context of the phenotype following ACMG guidelines as mentioned above. Usually, the experts indicate one or more OMIM terms[23,24] to describe the phenotype.

GS-exclusive diagnoses

We calculated the number of positive reports involving variants best or only captured by GS. These included CNVs/SVs involving fewer than three exons, non-coding,

including deep intronic SNVs/small indels (25 nucleotides or more from the nearest exon), STR nucleotide expansions, and mitochondrial CNVs/SVs. We used an exome targets BED file (used by Hospital Israelita Albert Einstein exome sequencing [ES] workflows) to assess which SNVs/small indels would not be covered by ES. CNV/SV coordinates were compared to the GENCODE v.49[26] GFF3 file using the GenomicRanges and GenomicFeatures packages, v.1.58, in R software v.4.4.1 to assess the number of overlapping exons of each variant.

Ancestry and admixture

We aggregated GVCF files into a multisample VCF (msVCF) file through DRAGEN. The msVCF file was imported into Hail v.0.2.61[27] in AWS Elastic MapReduce (EMR) clusters. Through Hail, we performed sample- and variant-level quality control by removing both samples and variants with a call rate < 85%, samples with a transition/transversion (Ti/Tv) ratio < 1.9, and any variants that did not pass DRAGEN's default hard filters. Next, we estimated pairwise kinship coefficients and removed individuals using Hail's maximal_independent_set method to ensure that there were no related individuals in the dataset. The resulting number of variants surpassed 116 million.

We performed phasing, pruning, and principal-component analysis (PCA) steps based on Nunes et al.,[28] with modifications made. Briefly, we merged the dataset with the jointly called GRCh38 VCF file with samples from the 1000 Genomes Project,[29] the Human Genome Diversity Project (HGDP),[30] and the Simons Genome Diversity Project (SGDP)[31] with bcftools v.1.21 and removed variants with a call rate < 95% and significant deviation from Hardy-Weinberg Equilibrium ($p \leq 10^{-8}$), resulting in 22.7 million variants in common among all three datasets. This merged dataset was then pruned with PLINK v.1.9[32] to keep variants that are in approximate linkage equilibrium with each other by removing one of each pair of variants with an $r^2 > 0.2$ in 500,000 base-pair windows. The pruned dataset had 2,580,344 autosomal variants. Next, we phased the dataset with SHAPEIT v.5.1.1,[33] using the 1000 Genomes Project[29] phased VCF as a scaffold and using recombination maps provided by the tool authors in their GitHub repository. In parallel, we performed a PCA to estimate the first 10 ancestry-related principal components using the SNPRelate package v.1.40.0 in R software v.4.4.1 to assess population structure. The dataset contained K = 6 superpopulations: five reference superpopulations, AFR (African), AMR (Indigenous American), EAS (East Asian), EUR (European), and SAS (Central/South Asian) and our sample (BRAZIL_BRGP).

We estimated global ancestry proportions and performed local ancestry inference (LAI) through FLARE tool v.0.5.2[34] using the 1000 Genomes Project VCF as reference, as well as the same recombination maps used during the phasing step. The output consisted of a VCF with 709,970 variants in common with the reference VCF, with ancestry annotation for each allele and a text file with global ancestry proportions. To simplify downstream LAI analyses, the local ancestry assigned to the nearest variant was also assigned to any variant not contained in the FLARE VCF, adapted from Browning et al.[34] (who assigned the immediately preceding position rather than the absolute nearest variant). We used the hap-ibd program v.1.0[35] to detect identity-by-descent (IBD) segments shared by patients. The program ped-sim was used to simulate IBD segments in relation to kinship degree. We simulated 8,000 individual pairs for each degree between third and seventeenth with the recombination map refined by Bhérer et al.[36] and the crossing-over model by Campbell et al.[37] The simulated output served as input for fitting a logistic regression model of kinship degree in relation to shared IBD segments in cM.

Consanguinity

We used DRAGEN's roh.bed files to quantify the number of runs of homozygosity (NROH) with lengths ≥1 million base pairs (MB), the sum of all NROH lengths (SROH), and the frequency of ROH (FROH) across the genome of each sample through scripts for R software v.4.4.1. We used Matalonga et al.'s[38] SROH thresholds to summarize evidence for recent consanguinity: no consanguinity, SROH ≤ 22 MB; probable non-consanguinity, SROH > 22 MB and SROH < 79 MB; probable consanguinity, SROH ≥ 79 MB and SROH ≤ 123 MB; and consanguinity, SROH > 123 MB.

RNA-seq

We prospectively collected RNA from all enrolled individuals, aiming to sequence approximately 1,000 transcriptomes for the RNA sequencing (RNA-seq) subproject. We extracted whole RNA from peripheral blood mononuclear cells (PBMCs). We prepared TS libraries with KAPA mRNA HyperPrep (Roche, Basel, Switzerland) or Stranded Total RNA Prep Ligation with Ribo-Zero Plus (Illumina, San Diego, California, USA) kits according to the manufacturer's instructions. RNA-seq was performed on Illumina's NextSeq 6000 platform. The sequencing reads were mapped to the reference genome (GRCh38 build) with DRAGEN v.3.8.4 with the following parameter values: COUNT_MODE = "IntersectionStrict," STRAND = "no," COUNT_OVERLAPS = true, and PAIRED_END = true.

If the GS result was inconclusive with identification of a variant with potential to alter splicing (e.g., synonymous variants or variants in a non-coding region) or was negative, the evidence was reevaluated on TS.

The BAM files were processed by the DROP tool v.1.1.1.[39] We executed all three DROP modules: OUTRIDER (aberrant expression),[40] FRASER (aberrant splicing),[41] and MAE (monoallelic expression). The last one required VCF files obtained via the GS for the same patients, which were imported into DROP accordingly.

We are currently using the DROP tool[39] to guide the reassessment of negative/inconclusive reports based on candidate genes with aberrant expression or splicing.

Long-read GS

We sequenced selected samples with inconclusive or negative reports with Oxford Nanopore Technologies' (ONT) long-read technology (Oxford Science Park, Oxford OX4 4GA, UK).

Inclusion criteria for ONT long-read sequencing followed three arms. The first arm consists of a phenotype-driven and bioinformatics-supported prioritization strategy focused on samples with suspected disorders associated with STR expansions, such as spinocerebellar ataxias, myotonic dystrophies, and *RFC1*-related disease.[42] Thus, samples with candidate phenotypes were assessed by ExpansionHunter,[14] and those flagged as harboring potential repeat expansions were sent for orthogonal validation using ONT long-read sequencing. This approach was especially relevant in cases involving (1) large expansions exceeding the reliable sizing capacity of short-read data or (2) loci with complex repeat architecture and high GC content.

The second arm involves samples with suspected variants in regions highly homologous to pseudogenes (e.g., *IKBKG*).[43] The third arm contains samples with a diagnostic hypothesis or phenotypes with specific candidate genes.

Therefore, we determined that samples included in any of the arms were good candidates for long-read sequencing, given long-read sequencing's suitability for resolving large and complex variants, as well as alterations located in regions that are technically challenging to assess using short-read sequencing. This strategy was specifically implemented to increase the overall diagnostic yield and enhance variant detection in cases that remained unsolved after short-read GS.

Briefly, we performed sequencing with the SQK-LSK114 or Native Barcoding 24 v.14 kits. The library input was 1 μg genomic DNA in a final volume of 48 μL per sample. We did not fragment the DNA to ensure long fragments. The ends of the DNA strands were enzymatically repaired so that adapter sequences (and molecular barcodes if needed) could be ligated with kit-specific DNA ligases following the manufacturer's instructions. Clean-up steps were performed with the help of magnetic beads and 80% ethanol elution steps. The prepared libraries were loaded into PromethION 24 flow cells. The base-calling process was performed using MinKNOW software v.23.07.12 with default settings on a PromethION A100 processing unit, based on the electric current perturbations caused by DNA reads passing through the molecular pores.

ClinVar submissions

We submitted selected pathogenic variants, likely pathogenic variants, or VUSs (excluding secondary findings) to the ClinVar database via their submission portal using the spreadsheet templates. We assessed the estimated local ancestry of each submission annotated with the "novelty" tag. The local ancestry of each variant was assigned the same ancestry as the closest locus, since the LAI was performed with a pruned dataset. Next, we calculated the haplotype frequency of each locus, stratified by ancestry. Our interpretations may be found with the submitter "Laboratorio de Genetica e Diagnostico Molecular (Hospital Israelita Albert Einstein), LABO-HIAE," ID 508506 (https://www.ncbi.nlm.nih.gov/clinvar/submitters/508506/).

Results

Patients

We enrolled 10,305 patients from 21 centers nationwide. The median age was 12 years (interquartile range [IQR] = 4–31, minimum = 0, maximum = 92), with variations across disease cohorts (Tables S1 and S2). Female participants comprised the majority of the overall sample (53.4%), with the highest proportion observed in the hereditary cancer risk syndromes cohort (86.3%) due to a history of breast cancer. Male participants made up 46.6% of the sample and predominated in the largest cohort (neurologic disorders: 56.8%). Most individuals did not undergo genetic testing before enrollment. We searched our PhenoTips database and found that just 15.1% (1,561/10,305) of enrolled individuals had documentation of one or more prior genetic tests. The tests included karyotyping (1,207/10,305 = 11.7%), targeted gene panels (259/10,305 = 2.5%), microarrays (246/10,305 = 2.4%), MLPA (240/10,305 = 2.3%), and single-target molecular genotyping tests, including Sanger sequencing (236/10,305 = 2.3%), ES (143/10,305 = 1.4%), methylation assays (25/10,305 = 0.2%), and FISH (15/10,305 = 0.1%).

Diagnostic yield

Among the enrolled patients, 9,448 have already received GS results. The overall diagnostic yield is 35.6% (3,366/9,448). Inconclusive reports accounted for 32.0% (3,024/9,448), while negative reports constituted 32.4% (3,058/9,448). The diagnostic yield varied according to disorder cohorts, ranging from 20.4% in the immunologic disorders (280/1,373) cohort to 73.4% in the dermatologic disorders cohort (91/124). Among critically ill patients, the diagnostic yield was 32.9% (72/219). Table 1 displays the number of enrolled patients and their respective diagnostic yields.

The overall diagnostic yield presented here includes cases resolved with the aid of RNA-seq and long-read GS, as these methods complemented the evidence provided by the primary GS test in a small proportion of cases (see transcriptomics: preliminary findings and long-read sequencing: preliminary findings).

Reported variants and phenotypes

The most common monogenic phenotype among individuals with positive results in each cohort is shown in

Table 1. Enrollment, diagnostic yield, and proportion of GS-exclusive reports across cohorts evaluated in the Brazilian Rare Genomes Project

Disorder cohort	Enrolled, *n*	Proportion of total sample size, %	Diagnostic yield, % (positive reports/all reports)	GS-exclusive reports, % (GS-exclusive/positive reports)
Neurologic disorders	3,561	34.56	37.1 (1,171/3,160)	6.9 (81/1,171)
Neuromuscular disorders	305	2.96	32.8 (88/268)	13.6 (12/88)
Neurocutaneous disorders	155	1.5	70.9 (105/148)	2.9 (3/105)
Inborn errors of metabolism	416	4.04	45.4 (183/394)	4.9 (9/183)
Skeletal disorders	398	3.86	57.8 (207/358)	2.4 (5/207)
Hereditary cancer risk syndromes	1,720	16.69	25 (424/1,667)	2.4 (10/424)
Clinically recognizable genetic syndromes	633	6.14	50.3 (298/592)	3.4 (10/298)
Endocrine disorders	332	3.22	35.7 (109/305)	3.7 (4/109)
Connective tissue disorders	163	1.58	41.5 (63/147)	1.6 (1/63)
Immunologic disorders	1,484	14.4	20.4 (280/1,373)	2.9 (8/280)
Ophthalmologic disorders	184	1.79	48.5 (79/163)	5.1 (4/79)
Hearing or ear disorders	45	0.44	48.8 (21/43)	9.5 (2/21)
Renal and urinary tract disorders	142	1.38	46.5 (53/114)	0.0 (0/53)
Dermatologic disorders	133	1.29	73.4 (91/124)	2.2 (2/91)
Hematologic disorders	132	1.28	37.5 (46/120)	4.3 (2/46)
Cardiological disorders	110	1.07	29.2 (29/96)	0.0 (0/29)
Pulmonary disorders	61	0.59	40.7 (24/59)	0.0 (0/24)
Gastrointestinal disorders	107	1.04	23.2 (22/95)	9.1 (2/22)
Vascular disorders	5	0.05	33.3 (1/3)	0.0 (0/1)
Critically ill patients	219	2.13	32.4 (72/219)	1.4 (1/72)
Overall	10,305	100	35.6 (3,366/9,448)	4.6 (155/3,366)

Table 2. The most prevalent phenotypes associated with structural chromosomal abnormalities were 22q11.2 deletion syndrome (1.4%, 48/3,366), 16p11.2 deletion syndrome (0.42%, 14/3,366), 22q11.2 duplication syndrome (0.39%, 13/3,366), 15q11.2–q13.1 deletion (Prader-Willi syndrome/Angelman syndrome; 0.36%, 12/3,366), and 7q11.23 deletion (Williams-Beuren syndrome; 0.30%, 10/3,366). We also detected 17 cases of chromosome aneuploidies (three cases of Klinefelter syndrome, three cases of XYY syndrome, two cases of Turner syndrome, two cases of X trisomy, two cases of mosaic trisomy 9, two cases of trisomy 18, one case of trisomy 13, one case of mosaic trisomy 14, and one case of trisomy 21), and two cases of segmental loss of heterozygosity (one case of Prader-Willi syndrome and one case of Silver-Russell syndrome).

We submitted variant interpretations (excluding secondary findings) to the ClinVar database. The submissions included a total of 3,201 interpretations: 592 (18.5%) pathogenic variants, 739 (23.1%) likely pathogenic variants, and 1,870 (58.4%) VUSs. Not all reported variants have been uploaded to ClinVar yet, as we periodically curate and prepare these submissions in batches. The uploaded interpretations were distributed across 1,352 unique genes, which are associated with diseases exhibiting the following modes of inheritance: AR, 38.2%; autosomal dominant (AD), 37.4%; AD and AR, 15.4%; X-linked (XL), 8.2%; and other modes of inheritance, 0.8%.

GS-exclusive diagnoses

Among the 3,366 individuals with positive reports, GS yielded exclusive diagnoses for 155 (4.6%) by providing information regarding variants less likely to be detected by other methodologies, such as ES.

The advantage of GS varied according to disorder cohorts, ranging from 1.6% in the connective tissue disorders cohort to 13.6% in the neuromuscular disorders cohort. The critically ill cohort accounted for 1.4% of GS-exclusive diagnoses. Table 1 also summarizes the proportion of GS-exclusive diagnoses per cohort.

Upon analyzing the 586 positive reports involving CNVs/SVs, we observed a notable proportion of short CNVs/SVs (those encompassing fewer than three exons):

Table 2. Most prevalent monogenic conditions per disorder cohort in the Brazilian Rare Genomes Project

Disorder cohort	Gene	OMIM gene ID	Phenotype	Mode of inheritance	Prevalence, % (gene reports/ positive reports)
Neurologic disorders	*ANKRD11*	611192	KBG syndrome	AD	1.9 (22/1,171)
Neuromuscular disorders	*LAMA2*	156225	LAMA2-related muscular dystrophy	AR	4.5 (4/88)
Neurocutaneous disorders	*NF1*	613113	neurofibromatosis type 1	AD	62.9 (66/105)
Inborn errors of metabolism	*CBS*	236200	homocystinuria B6-responsive and nonresponsive types	AR	3.8 (7/183)
Skeletal disorders	*COL1A2*	120160	COL1A2-osteogenesis imperfecta related	AD	14.5 (30/207)
Hereditary cancer risk syndromes	*BRCA1*	113705	BRCA1-associated hereditary breast and ovarian cancer	AD	28.1 (119/424)
Clinically recognizable genetic syndromes	*PTPN11*	176876	PTPN11-related RASopathy	AD	7.0 (21/298)
Endocrine disorders	*WFS1*	606201	WFS1-related disorder	AR/AD	6.4 (7/109)
Connective tissue disorders	*FBN1*	134797	Marfan syndrome	AD	38.1 (24/63)
Immunologic disorders	*BTK*	300300	agammaglobulinemia X-linked type 1	XL	6.8 (19/280)
Ophthalmologic disorders	*CYP1B1*	601771	CYP1B1-ophthalmological related disorder	AR	13.9 (11/79)
Hearing or ear disorders	*GJB2*	220290	deafness autosomal recessive type 1A	AR	14.3 (3/21)
Renal and urinary tract disorders	*COL4A5*	303630	Alport syndrome X-linked type 1	XL	18.9 (10/53)
Dermatologic disorders	*COL7A1*	120120	epidermolysis bullosa	AR	47.3 (43/91)
Hematologic disorders	*ELANE*	130130	ELANE-neutropenia related	AD	6.5 (3/46)
Cardiological disorders	*KCNH2*	152427	long QT syndrome type 2	AD	6.9 (2/29)
Pulmonary disorders	*CCDC39*	613798	primary ciliary dyskinesia type 14	AR	16.7 (4/24)
Gastrointestinal disorders	*CFTR*	602421	cystic fibrosis	AR	9.1 (2/22)
Vascular disorders	*RASA1*	139150	capillary malformation-arteriovenous malformation type 1	AD	100.0 (1/1)
Critically ill patients	*COL2A1*	120140	COL2A1-related disorder	AD	4.2 (3/72)

14.5% (85/586). Among these, 22.3% (19/85) were located on the X chromosome, contributing to a modest increase in the diagnostic yield of XL neurodevelopmental disorders (e.g., *AFF2* and *ZC4H2*), immunodeficiencies (e.g., *BTK* and *XIAP*), and inborn errors of metabolism (e.g., *IDS*).

Our GS protocol provided positive reports involving STR expansions for 54 patients, mainly associated with neurological conditions (50 cases, 92.6%), except for *HOXA13* and *HOXD13* (two cases each). Genes with the most detected STR expansion variants detected included the following: *FMR1*, 14 cases; *ATXN8OS*, eight cases; *ATXN1*, *ATXN3*, *DMPK*, and *HTT*, three cases each; *ARX*, *FGF14*, *FXN*, and *PABPN1*, two cases each; and *ATN1*, *ATXN2*, *C9orf72*, *CNBP*, *JPH3*, *PHOX2B*, *PPP2R2B*, and *RFC1*, one case each. STR expansion cases accounted for 1.6% of all positive cases.

Additionally, 10 probands received a positive report due to rare mitochondrial CNVs/SVs in the mitochondrial genome. GS provided very high mitochondrial read coverage, with a median coverage of 3,875× (IQR = 3,062.2×–4,785.8×, minimum = 1,118.8×, maximum = 12,934.5×), facilitating the identification of these SVs.

The remaining six patients had SNVs and small indels in deep intronic regions. Table 3 displays the reported events, stratified by variant type.

Regarding the 136 enrolled individuals with prior inconclusive/negative ES testing, 109 had their genomes analyzed. We observed that GS provided a conclusive diagnosis for 11.0% (12/109) of them, from the following cohorts: neurologic disorders (7/74, 9.5%), inborn errors of metabolism (3/10, 30.0%), immunologic disorders (1/6, 16.7%), and clinically recognizable genetic syndromes (1/2, 50.0%). In three of the affected individuals (3/109,

Table 3. Events in positive reports, stratified if they are GS-exclusive findings and variant type

Variant type	GS-exclusive finding?	
	No, *n* (% of total)	Yes, *n* (% of total)
SNV	1,977 (54.39)	2 (0.06)
Small indel	955 (26.27)	4 (0.11)
CNV/SV	492 (13.54)	85 (2.34)
SNV (mitochondrial)	30 (0.83)	0 (0.00)
Chromosome aneuploidy	17 (0.47)	0 (0.00)
Small indel (mitochondrial)	7 (0.19)	0 (0.00)
Segmental loss of heterozygosity (LOH)[a]	2 (0.06)	0 (0.00)
CNV/SV (mitochondrial)	0 (0.00)	10 (0.28)
STR expansion	0 (0.00)	54 (1.49)
All variants	3,480 (95.74)	155 (4.26)

[a]Findings suggestive of uniparental disomy, confirmed by complementary testing.

2.8%), the probable reason for ES failure was identifiable: two had small CNVs/SVs spanning fewer than three exons (*WWOX* and *CEP120*) and one had an intronic variant in *F12*. The basis for the prior negative ES results in the remaining affected individuals could not be ascertained, as those analyses had been conducted by external laboratories, and detailed methodological information was unavailable.

Ancestry

We observed a remarkable variation regarding ancestry backgrounds in our sample, as expected. Individuals had a significant proportion of EUR background (median = 60%, IQR = 48.5%–71.8%), followed by AFR contribution (median = 23.4%, IQR = 12.6%–35.9%).

The PCA results confirm the admixture. The Brazilian samples do not tightly cluster around the reference samples; instead, they behave like a gradient (Figure 1A). Most samples cluster around EUR, AFR, and AMR samples, with a few outliers clustering around EAS samples.

We observed slight variations according to Brazilian states, reflecting unique regional demographic histories. For example, Northeastern states have higher AFR contribution than Southern states (Figures 1B–1D).

Local ancestry

Local ancestry and IBD

To illustrate the utility of integrating IBD analysis and LAI, we report the investigation of a *COL7A1* pathogenic variant associated with a cluster of AR epidermolysis bullosa dystrophica in the state of Bahia, Northeast Brazil.

The variant in question, c.5047C>T (GenBank: NM_000094.4) (p.Arg1683Ter), was found in 13 unrelated individuals recruited by the same center in Bahia and was found in 10 heterozygous individuals with EUR background in the gnomAD database[44] (variant ID 3-48580586-G-A). Therefore, we hypothesized that it had a common origin and remained in the population via the founder effect.

We observed that all 13 homozygous individuals shared IBD regions around the consensus interval chr3:3516233–114543403, which overlaps the variant locus (chr3:48580586). Furthermore, all 26 haplotypes were assigned EUR local ancestry. The median IBD segment length shared between them was 37.7 cM (n = 13 pairs, IQR = 20.7–55.2, minimum = 6.5, maximum = 79.2). In contrast, samples without the *COL7A1* variant shared shorter IBD segments (n = 1,176 pairs, median = 6.4 cM, IQR = 3.6–11.6, minimum = 2.0, maximum = 81.6, adjusted Kruskal-Wallis test $p < 0.0001$). According to the results of the ped-sim simulation, these observations are consistent with a third- or fourth-degree relationship. Therefore, the results seem to support the hypothesis that the variant is embedded in a EUR ancestry segment and, perhaps, spread among the gene pool of Bahia populations following population bottlenecks and/or random genetic drift.

Local ancestry of ClinVar alleles

We successfully estimated local ancestry for 1,346 of 3,201 (42.0%) loci submitted to the ClinVar database. Most of these loci (961/1,346) contained singleton variants (detected in a single individual in the BRGP dataset). When evaluating the ancestry assigned to each locus, we observed that overall, the local inference mirrored the global estimates: the EUR component dominated, being assigned 61.5% (IQR = 61%–62.1%) of the times across all haplotypes, followed by AFR (median 25.8%, IQR = 25.3%–26.2%) and AMR (median 11.8%, IQR = 11.4%–12.2%). The singleton loci did not statistically differ from the overall trend (Wilcoxon-Mann-Whitney test p values: EUR = 0.7, AFR = 0.52, and AMR = 0.59).

Consanguinity

The median SROH in the BRGP sample was relatively low (6.3 MB, IQR = 3.3–11.8 MB), meaning that only a relevant minority of patients presented strong evidence of recent consanguinity, with few patients presenting extreme SROH values. In fact, according to Matalonga et al.'s[38] SROH thresholds, just 193 (2.0%) individuals were probably consanguineous, and 670 (7.1%) were consanguineous. Thus, at most 9.1% of individuals from our sample present evidence of recent consanguinity.

Transcriptomics: Preliminary findings

We sequenced 1,026 transcriptomes. Currently, 48 of them have been analyzed. The DROP tool results aided the reinterpretation of five reports following the upgrade of their ACMG classification criteria, in light of evidence from the detection of aberrant expression or visualization of aberrant splicing events. They represent 0.1% (5/3,366) of all positive cases. Table S3 presents a summary of cases solved with the help of RNA-seq analysis.

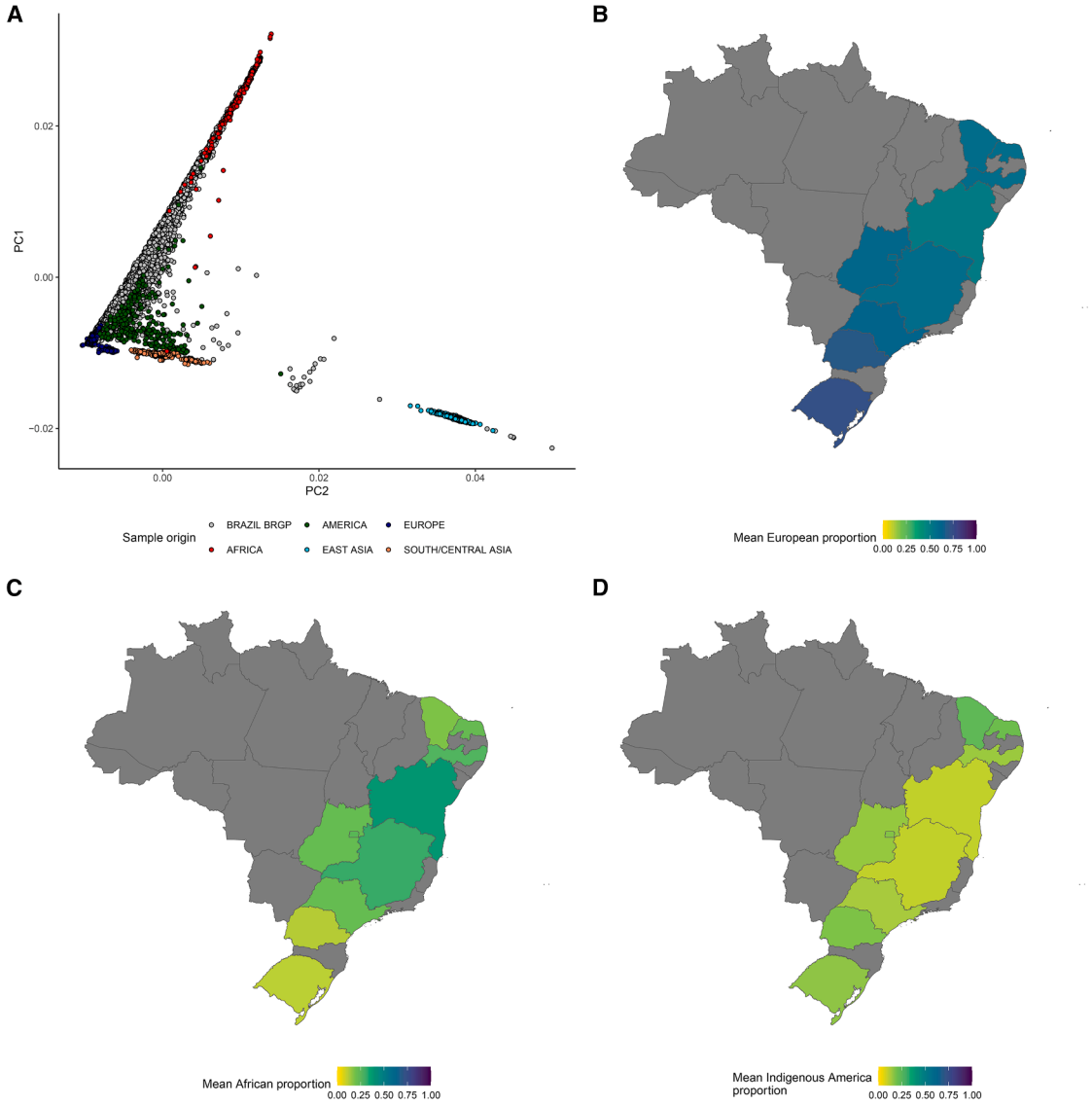

Figure 1. Ancestry analysis of BRGP individuals

(A) Principal-component analysis (PCA) plot with reference populations from the 1000 Genomes Project,[29] the Human Genome Diversity Project (HGDP),[30] and the Simons Genome Diversity Project (SGDP)[31] representing the ancestry backgrounds in the individuals recruited by the Brazilian Rare Genomes Project (BRGP). PC, principal component. The PCA was performed with approximately 2.5 million bi-allelic autosomal variants.

(B–D) Mean proportions of (B) European, (C) African, and (D) Indigenous American ancestry stratified by the patients' Brazilian state of origin.

Long-read sequencing: Preliminary findings

We resequenced 294 genomes from selected patients with inconclusive/negative reports through the ONT long-read platform. Currently, 136 of them (belonging to the first and second arms mentioned earlier) have been analyzed, and 32 (23.5%) were solved with the help of long-read sequencing. The majority involved 23 cases of STR variants that were better visualized with long-read data, followed by seven cases of 11.7 kbp pathogenic deletions between exons 4 and 10 of the *IKBKG* gene, which has homology to a pseudogene, leading to ambiguity in short-read analysis. The remaining two cases were a likely pathogenic homozygous 2.5 kbp tandem duplication of exons 4 and 5 of the *PPT1* gene in a patient with neuronal ceroid lipofuscinosis type 1 (MIM: 256730) and an individual harboring a deletion within filaggrin (*FLG*)—a gene that contains a highly polymorphic region with large tandem repeats. The 32 new reports amounted to 1.0% of all positive reports. Table S4 summarizes the results obtained through long-read GS, and Figures S5–S8 illustrate the read alignments of selected events.

Discussion

The BRGP is currently the largest initiative in Brazil focused on diagnosing rare diseases and hereditary cancer risk syndromes. With over 9,400 reports released, the

project assists individuals and their families across the country in understanding their genetic conditions.

Approximately 85% of enrolled patients have not undergone prior genetic testing, unlike the participants in Genomic England's 100,000 Genomes Project,[45] where many patients had previous ES or other tests before GS. The BRGP observed a diagnostic yield of 35.6%, broadly consistent with other large-scale GS studies. Genomics England, for instance, reported a yield of 33% among patients without prior testing.

However, despite similar aggregate yields, we observed marked differences at the cohort level. Interestingly, perhaps due to differences in genetic background, the phenotypic profile of our recruited individuals seems to be remarkably different from that of Genomic England's 100,000 Genomes. Dermatologic disorders, for example, reached a diagnostic yield of 73.4% in our study, compared with approximately 10% in Genomics England. Three factors likely account for this disparity. First, the selection of diseases evaluated in our cohort features characteristic, clinically recognizable phenotypes and well-established molecular monogenic bases: hereditary ichthyoses, ectodermal dysplasias, epidermolysis bullosa, and incontinentia pigmenti, whereas they diagnosed a diverse set of diseases, including ones with possibly multigenic causes (e.g., erythropoietic protoporphyria and erythrokeratodermas[45]). Second, *COL7A1* pathogenic or likely pathogenic variants comprised 47.3% (43/91) of positive cases in this cohort, driven by a geographically concentrated cluster of epidermolysis bullosa in Bahia. Third, prior testing was rare in the BRGP sample and exceptionally rare in this group, meaning that GS captured cases that would typically have been resolved by standard first-tier testing. Therefore, the combination of selection bias for conditions, geographic clustering, and the lack of prior testing (which meant our study captured cases that would normally have been resolved by standard tier-one testing) likely inflated our observed yield.

We observed that GS captured variants that were missed by ES in 2.8% of individuals with previously inconclusive or negative ES reports. Other studies have reported GS yields of 7.0%,[46] 14.5%,[47] 16.3%,[48] and 42%[49] among ES-tested patients. These figures are higher than our estimate, most likely attributable to our small sample size of previously tested patients, given the generally low rate of prior genetic testing. Nevertheless, our findings further reinforce the superior sensitivity of GS over ES.

A meta-analysis of 37 studies involving more than 20,000 children who underwent GS suggested an overall diagnostic yield of 41% (95% confidence interval [CI] = 34%–48%),[50] while other studies found yields as high as 68.3%,[51] depending on the population and disorder cohort.

The 100,000 Genomes Project indicated that 14% of diagnostics represented additional gains by GS compared to other diagnostic tests, as GS may capture variants that are usually undetectable by ES. Indeed, some previously negative or inconclusive ES cases in this cohort were solved by GS. A study of primary immunodeficiency patients identified 69 variants, including eight SVs not covered by ES.[52] The Broad Center for Mendelian Genomics reported that 8.2% of their findings involved variants ascertainable only through GS.[53] We obtained an overall estimate of 4.6%, which was driven by the following variant types in order of their contribution: short CNVs/SVs, STR expansions, mitochondrial SVs, and deep intronic variants. Although STR expansions can be detected through ES, previous studies[54–57] demonstrate that the performance of STR detection is variable and dependent on factors such as capture kit design, insert size, and genomic context, with some loci exhibiting very low genotyping rates or even a complete lack of coverage. Furthermore, the ability to detect these expansions may be compromised in GC-rich regions. In contrast, GS provides more uniform coverage across coding and non-coding regions, enabling the comprehensive and reliable detection of STR expansions. Therefore, while exome-based approaches may contribute to the identification of repeat expansion disorders, GS remains the superior strategy for their systematic detection and for maximizing diagnostic yield.

Two approaches can be used to identify the cohorts that benefited the most from GS. The first approach relies on the overall diagnostic yield, while the second evaluates the absolute number of GS-exclusive diagnoses among the positive reports. Although the dermatologic disorders and neurocutaneous disorders cohorts exhibited the highest diagnostic yields, their smaller sample sizes and low genetic diversity (most cases were driven by *COL7A1* and *NF1* variants, respectively) likely resulted in inflated estimates. Furthermore, the skeletal disorders and clinically recognizable genetic syndromes cohorts demonstrated high diagnostic yields because their inclusion criteria likely favored the recruitment of individuals with phenotypes highly suggestive of the targeted conditions. Regarding the second approach, the neurologic disorders cohort yielded the highest absolute number of GS-exclusive findings, significantly exceeding all other cohorts. This high diagnostic success was driven by STR expansions and short CNVs/SVs, including XL conditions, which also explains the observed overrepresentation of male individuals in this group.

We have implemented a priority GS processing workflow to expedite reporting for critically ill neonates in NICUs/PICUs, with a diagnostic yield of 32.4%, reflecting international settings that range between 21% and 46%.[58–61]

Our efforts contribute to the scientific community by depositing over 3,200 interpretations in ClinVar, providing data for clinical interpretations globally. This aligns with current movements to increase representation from diverse populations in genomic databases.[62,63] Most samples cluster around the EUR, AFR, and AMR samples, with a few outliers clustering around the EAS

samples, possibly due to recent Asian ancestry, compatible with Asian migration influx from the early 20th century. The global ancestry estimates varied across Brazilian regions. The same phenomenon was observed by Nunes et al.[28] in their analyses. In fact, our analyses closely resemble their results. Despite our global and local ancestry analysis differing from theirs, the similar results showed that ours was robust and captured the same ancestry signal that makes up Brazilian genomes.

LAI in admixed populations such as ours may help explain the clustering of rare disease cases within specific geographical regions, such as the epidermolysis bullosa (MIM: 226600) cases from Bahia driven by a recurrent *COL7A1* pathogenic variant.

To reassess inconclusive or negative reports, we evaluated TS data from over 1,000 patients and long-read sequencing data from over 200 patients. Previous studies indicate RNA-seq-based diagnostic yields between 3%[64] and 16%.[65] Long-read sequencing has been successfully used for the diagnosis of diseases complicated by pseudogenes, such as incontinentia pigmenti, which were missed by conventional testing.[66] We provided diagnoses for seven patients with a common deletion in the *IKBKG* gene. We also provided diagnoses for over 20 patients with complex CNV/SV events or STR expansions that were inconclusive during short-read analyses. Therefore, long-read sequencing can complement the results from short-read sequencing. In our experience, it represents 1.0% of all positive reports, a small but welcome contribution.

It is possible that some individuals in our sample carry multiple distinct monogenic diseases simultaneously; however, this complex matter is beyond the scope of the present study and warrants further investigation in future studies.

In conclusion, the BRGP showcases that GS is an essential tool for diagnosing rare diseases and hereditary cancer risk syndromes. GS improves diagnostic yield by capturing variants missed by traditional tests. TS identifies aberrant expression and splicing events, aiding in the reinterpretation of existing reports. This method has resolved a few cases of patients with specific genomic findings. Together, GS and TS enhance diagnostic accuracy, uncover genetic conditions, and support research and clinical practice. Moreover, long-read sequencing can complement the limitations of short-read sequencing, offering a modest increase in diagnostic yield.

Data code and availability

The raw data supporting the findings of this study cannot be publicly shared due to legal restrictions. These data may become available at a future date, pending future resolutions by the Brazilian Ministry of Health. Interpretations of pathogenic and likely pathogenic variants are periodically submitted to the ClinVar database under the submitter Laboratorio de Genetica e Diagnostico Molecular (Hospital Israelita Albert Einstein), LABO-HIAE, ID 508506 (https://www.ncbi.nlm.nih.gov/clinvar/submitters/508506/).

Consortia

The members of the Brazilian Rare Genomes Project Consortium are Antonio Victor Campos Coelho, Rafael Sales de Albuquerque, Catarina dos Santos Gomes, José Bandeira do Nascimento Junior, Gustavo Santos de Oliveira, Livia Maria Silva Moura, Luciana Souto Mofatto, Rafael Lucas Muniz Guedes, Rodrigo Araújo Sequeira Barreiro, Marcel Pinheiro Caraciolo, Ana Paula de Andrade Oliveira, Anne Caroline Barbosa Teixeira, Bruna Mascaro Cordeiro de Azevedo, Carolina Dias Carlos, Lucas Santos de Santana, Marina Cadena da Matta, Matheus Martinelli Lima, Nuria Bengala Zurro, Renata Yoshiko Yamada, Vivian Pedigone Cintra, Gabriela Pereira Campilongo, Gabriela Borges Cherulli Colichio, Renata Martins Ribeiro da Silva, Caio Robledo D'Angioli Costa Quaio, Carolina Araujo Moreno, Eduardo Perrone, Jéssica Grasiela de Araújo Espolaor, Joana Rosa Marques Prota, José Ricardo Magliocco Ceroni, Kelin Chen, Luiza do Amaral Virmond, Marina de França Basto Silva, Michele Patricia Migliavacca, Renata Moldenhauer Minillo, Thiago Yoshinaga Tonholo Silva, Karla de Oliveira Pelegrino, Ana Luiza Garcia Cunha, Joziele de Souza Lima, Anete Sevciovic Grumach, Caio Parente Barbosa, Angelina Xavier Acosta, Paula Brito Corrêa, Denise Pontes Cavalcanti, Carlos Eduardo Steiner, Erlane Marques Ribeiro, Wallace William da Silva Meireles, Giselle Maria Araujo Felix Adjuto, Ida Vanessa Doederlein Schwartz, Têmis Maria Felix, Irma Cecilia Douglas Paes Barreto, Antonette Souto El Husny, Jussara Melo de Cerqueira Maia, Vera Maria Dantas, Lúcia Helena de Oliveira Cordeiro, Luisa Zagne Braz, Magda Maria Sales Carneiro Sampaio, Mara Lucia Schmitz Ferreira Santos, Marco Antonio Curiati, Maria Teresinha de Oliveira Cardoso, Maria Teresa Alves da Silva Rosa, Mariana Paes Leme Ferriani, Ester Silveira Ramos, Paula Teixeira Lyra, Raquel Tavares Boy da Silva, Anna Cândida Ximenes de Mendonça Sobreira, Tatiana Regia Suzana Amorim Boa Sorte, Melissa Rossi Calvão Dumas, Thaís Bomfim Teixeira, Vandré Cabral Gomes Carneiro, Patrícia Silva Mota, José Luiz Pedroso, Orlando Graziani Barsottini, Victor Rebelo Procaci, Anna Maria Gomes, Tatiana Ferreira de Almeida, and João Bosco Oliveira.

Acknowledgments

The Brazilian Rare Genomes Project is an initiative of Hospital Israelita Albert Einstein, funded through a partnership with the Programa de Apoio ao Desenvolvimento Institucional do Sistema Único de Saúde (PROADI-SUS) of the Brazilian Ministry of Health.

Declaration of interests

The authors declare no competing interests.

Web resources

Broad Institute, https://github.com/broadinstitute/str-analysis
GenBank, https://www.ncbi.nlm.nih.gov/genbank/
National Library of Medicine, https://www.ncbi.nlm.nih.gov/clinvar/submitters/508506/
OMIM, https://www.omim.org

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2026.100624.

Received: January 9, 2026
Accepted: April 28, 2026

References

1. Ferreira, C.R. (2019). The burden of rare diseases. Am. J. Med. Genet. *179*, 885–892.
2. Castro, M.C., Massuda, A., Almeida, G., Menezes-Filho, N.A., Andrade, M.V., de Souza Noronha, K.V.M., Rocha, R., Macinko, J., Hone, T., Tasca, R., et al. (2019). Brazil's unified health system: the first 30 years and prospects for the future. Lancet *394*, 345–356.
3. Neu, M.B., Bowling, K.M., and Cooper, G.M. (2019). Clinical utility of genomic sequencing. Curr. Opin. Pediatr. *31*, 732–738.
4. Griffin, H.R., Pyle, A., Blakely, E.L., Alston, C.L., Duff, J., Hudson, G., Horvath, R., Wilson, I.J., Santibanez-Koref, M., Taylor, R.W., and Chinnery, P.F. (2014). Accurate mitochondrial DNA sequencing using off-target reads provides a single test to identify pathogenic point mutations. Genet. Med. *16*, 962–971.
5. Murdock, D.R., Dai, H., Burrage, L.C., Rosenfeld, J.A., Ketkar, S., Müller, M.F., Yépez, V.A., Gagneur, J., Liu, P., Chen, S., et al. (2021). Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. J. Clin. Investig. *131*, e141500.
6. Manolio, T.A., Abramowicz, M., Al-Mulla, F., Anderson, W., Balling, R., Berger, A.C., Bleyl, S., Chakravarti, A., Chantratita, W., Chisholm, R.L., et al. (2015). Global implementation of genomic medicine: We are not alone. Sci. Transl. Med. *7*, 290ps13.
7. Barwell, J.G., O'Sullivan, R.B., Mansbridge, L.K., Lowry, J.M., and Dorkins, H.R. (2018). Challenges in implementing genomic medicine: the 100,000 Genomes Project. J. Transl. Genet. Genom. *2*, 13.
8. Stark, Z., Dolman, L., Manolio, T.A., Ozenberger, B., Hill, S.L., Caulfied, M.J., Levy, Y., Glazer, D., Wilson, J., Lawler, M., et al. (2019). Integrating Genomics into Healthcare: A Global Responsibility. Am. J. Hum. Genet. *104*, 13–20.
9. Coelho, A.V.C., Mascaro-Cordeiro, B., Lucon, D.R., Nóbrega, M.S., Reis, R.d.S., de Alexandre, R.B., Moura, L.M.S., de Oliveira, G.S., Guedes, R.L.M., Caraciolo, M.P., et al. (2022). The Brazilian Rare Genomes Project: Validation of Whole Genome Sequencing for Rare Diseases Diagnosis. Front. Mol. Biosci. *9*, 821582.
10. Harris, P.A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., and Conde, J.G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. J. Biomed. Inform. *42*, 377–381.
11. Harris, P.A., Taylor, R., Minor, B.L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., et al. (2019). The REDCap consortium: Building an international community of software platform partners. J. Biomed. Inform. *95*, 103208.
12. Girdea, M., Dumitriu, S., Fiume, M., Bowdin, S., Boycott, K.M., Chénier, S., Chitayat, D., Faghfoury, H., Meyn, M.S., Ray, P.N., et al. (2013). PhenoTips: patient phenotyping software for clinical and research use. Hum. Mutat. *34*, 1057–1065.
13. Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L.C., Lewis-Smith, D., Vasilevsky, N.A., Danis, D., Balagura, G., Baynam, G., Brower, A.M., et al. (2021). The Human Phenotype Ontology in 2021. Nucleic Acids Res. *49*, D1207–D1217.
14. Dolzhenko, E., van Vugt, J.J.F.A., Shaw, R.J., Bekritsky, M.A., van Blitterswijk, M., Narzisi, G., Ajay, S.S., Rajan, V., Lajoie, B.R., Johnson, N.H., et al. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. *27*, 1895–1903.
15. Varsomics (2022). Varstation - The definitive platform for Geneticists and Bioinformatics. https://web.varstation.com/.
16. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.
17. Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., and Muller, J. (2018). AnnotSV: an integrated tool for structural variations annotation. Bioinformatics *34*, 3572–3574.
18. Geoffroy, V., Guignard, T., Kress, A., Gaillard, J.B., Solli-Nowlan, T., Schalk, A., Gatinois, V., Dollfus, H., Scheidecker, S., and Muller, J. (2021). AnnotSV and knotAnnotSV: a web server for human structural variations annotations, ranking and analysis. Nucleic Acids Res. *49*, W21–W28.
19. 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.
20. Biesecker, L.G., Harrison, S.M.; and ClinGen Sequence Variant Interpretation Working Group (2018). Sequence Variant Interpretation Working. Genet. Med. *20*, 1687–1688.
21. Quaio, C.R.D.C., Ceroni, J.R.M., Pereira, M.A., Teixeira, A.C.B., Yamada, R.Y., Cintra, V.P., Perrone, E., De França, M., Chen, K., Minillo, R.M., et al. (2023). The hospital Israelita Albert Einstein standards for constitutional sequence variants classification: version 2023. Hum. Genomics *17*, 102.
22. Miller, D.T., Lee, K., Abul-Husn, N.S., Amendola, L.M., Brothers, K., Chung, W.K., Gollob, M.H., Gordon, A.S., Harrison, S.M., Hershberger, R.E., et al. (2023). ACMG SF v3.2 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet. Med. *25*, 100866.
23. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res. *43*, D789–D798.

24. Amberger, J.S., Bocchini, C.A., Scott, A.F., and Hamosh, A. (2019). OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res. *47*, D1038–D1043.
25. den Dunnen, J.T., Dalgleish, R., Maglott, D.R., Hart, R.K., Greenblatt, M.S., McGowan-Jordan, J., Roux, A.F., Smith, T., Antonarakis, S.E., and Taschner, P.E.M. (2016). HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum. Mutat. *37*, 564–569.
26. Mudge, J.M., Carbonell-Sala, S., Diekhans, M., Martinez, J.G., Hunt, T., Jungreis, I., Loveland, J.E., Arnan, C., Barnes, I., Bennett, R., et al. (2025). GENCODE 2025: reference gene annotation for human and mouse. Nucleic Acids Res. *53*, D966–D975.
27. Hail Team (2023). Hail version 0.2. https://github.com/hail-is/hail.
28. Nunes, K., Araújo Castro E Silva, M., Rodrigues, M.R., Lemes, R.B., Pezo-Valderrama, P., Kimura, L., de Sena, L.S., Krieger, J.E., Catoia Varela, M., de Azevedo, L.O., et al. (2025). Admixture's impact on Brazilian population evolution and health. Science *388*, eadl3564.
29. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.
30. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science *367*, eaay5012.
31. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature *538*, 201–206.
32. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
33. Hofmeister, R.J., Ribeiro, D.M., Rubinacci, S., and Delaneau, O. (2023). Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. Nat. Genet. *55*, 1243–1249.
34. Browning, S.R., Waples, R.K., and Browning, B.L. (2023). Fast, accurate local ancestry inference with FLARE. Am. J. Hum. Genet. *110*, 326–335.
35. Zhou, Y., Browning, S.R., and Browning, B.L. (2020). A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. Am. J. Hum. Genet. *106*, 426–437.
36. Bhérer, C., Campbell, C.L., and Auton, A. (2017). Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. Nat. Commun. *8*, 14994.
37. Campbell, C.L., Furlotte, N.A., Eriksson, N., Hinds, D., and Auton, A. (2015). Escape from crossover interference increases with maternal age. Nat. Commun. *6*, 6260.
38. Matalonga, L., Laurie, S., Papakonstantinou, A., Piscia, D., Mereu, E., Bullich, G., Thompson, R., Horvath, R., Pérez-Jurado, L., Riess, O., et al. (2020). Improved Diagnosis of Rare Disease Patients through Systematic Detection of Runs of Homozygosity. J. Mol. Diagn. *22*, 1205–1215.
39. Yépez, V.A., Mertes, C., Müller, M.F., Klaproth-Andrade, D., Wachutka, L., Frésard, L., Gusic, M., Scheller, I.F., Goldberg, P.F., Prokisch, H., and Gagneur, J. (2021). Detection of aberrant gene expression events in RNA sequencing data. Nat. Protoc. *16*, 1276–1296.
40. Brechtmann, F., Mertes, C., Matusevičiūtė, A., Yépez, V.A., Avsec, Ž., Herzog, M., Bader, D.M., Prokisch, H., and Gagneur, J. (2018). OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. Am. J. Hum. Genet. *103*, 907–917.
41. Mertes, C., Scheller, I.F., Yépez, V.A., Çelik, M.H., Liang, Y., Kremer, L.S., Gusic, M., Prokisch, H., and Gagneur, J. (2021). Detection of aberrant splicing events in RNA-seq data using FRASER. Nat. Commun. *12*, 529.
42. Wallace, S.E., and Bean, L.J.H. (2017). Resources for Genetics Professionals — Genetic Disorders Caused by Nucleotide Repeat Expansions and Contractions. In GeneReviews® [Internet], M.P. Adam, S. Bick, and G.M. Mirzaa, eds. (University of Washington, Seattl).
43. Conlin, L.K., Aref-Eshghi, E., McEldrew, D.A., Luo, M., and Rajagopalan, R. (2022). Long-read sequencing for molecular diagnostics in constitutional genetic disorders. Hum. Mutat. *43*, 1531–1544.
44. Chen, S., Francioli, L.C., Goodrich, J.K., Collins, R.L., Kanai, M., Wang, Q., Alföldi, J., Watts, N.A., Vittal, C., Gauthier, L.D., et al. (2024). A genomic mutational constraint map using variation in 76,156 human genomes. Nature *625*, 92–100.
45. Smedley, D., Smith, K.R., Martin, A., Thomas, E.A., McDonagh, E.M., Cipriani, V., Ellingford, J.M., Arno, G., Tucci, A., Vandrovcova, J., et al. (2021). 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N. Engl. J. Med. *385*, 1868–1880.
46. Albuquerque, A.L.B., Dos Santos, G.G., Sadok, S.H., Antonello, B.B., de Jesus, L.M., de Carvalho, M.E.A., Mutarelli, A., and Ribeiro, P.V.Z. (2025). Diagnostic Yield of Genome Sequencing Versus Exome Sequencing in Pediatric Patients With Rare Phenotypes: A Systematic Review and Meta-Analysis. Am. J. Med. Genet. *197*, e64146.
47. Bertoli-Avella, A.M., Beetz, C., Ameziane, N., Rocha, M.E., Guatibonza, P., Pereira, C., Calvo, M., Herrera-Ordonez, N., Segura-Castel, M., Diego-Alvarez, D., et al. (2021). Successful application of genome sequencing in a diagnostic setting: 1007 index cases from a clinically heterogeneous cohort. Eur. J. Hum. Genet. *29*, 141–153.
48. Splinter, K., Adams, D.R., Bacino, C.A., Bellen, H.J., Bernstein, J.A., Cheatle-Jarvela, A.M., Eng, C.M., Esteves, C., Gahl, W.A., Hamid, R., et al. (2018). Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease. N. Engl. J. Med. *379*, 2131–2139.
49. Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W.M., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature *511*, 344–347.
50. 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.
51. Scocchia, A., Wigby, K.M., Masser-Frye, D., Del Campo, M., Galarreta, C.I., Thorpe, E., McEachern, J., Robinson, K., Gross, A., Subramanian, S., et al. (2019). Clinical whole genome sequencing as a first-tier test at a resource-limited dysmorphology clinic in Mexico. npj Genom. Med. *4*, 5.

52. Thaventhiran, J.E.D., Lango Allen, H., Burren, O.S., Rae, W., Greene, D., Staples, E., Zhang, Z., Farmery, J.H.R., Simeoni, I., Rivers, E., et al. (2020). Whole-genome sequencing of a sporadic primary immunodeficiency cohort. Nature *583*, 90–95.
53. Wojcik, M.H., Lemire, G., Berger, E., Zaki, M.S., Wissmann, M., Win, W., White, S.M., Weisburd, B., Wieczorek, D., Waddell, L.B., et al. (2024). Genome Sequencing for Diagnosing Rare Diseases. N. Engl. J. Med. *390*, 1985–1997.
54. van der Sanden, B.P.G.H., Corominas, J., de Groot, M., Pennings, M., Meijer, R.P.P., Verbeek, N., van de Warrenburg, B., Schouten, M., Yntema, H.G., Vissers, L.E.L.M., et al. (2021). Systematic analysis of short tandem repeats in 38,095 exomes provides an additional diagnostic yield. Genet. Med. *23*, 1569–1573.
55. Méreaux, J.L., Davoine, C.S., Coutelier, M., Guillot-Noël, L., Castrioto, A., Charles, P., Coarelli, G., Ewenczyk, C., Klebe, S., Heinzmann, A., et al. (2023). Fast and reliable detection of repeat expansions in spinocerebellar ataxia using exomes. J. Med. Genet. *60*, 717–721.
56. Yoon, J.G., Lee, S., Cho, J., Kim, N., Kim, S., Kim, M.J., Kim, S.Y., Moon, J., and Chae, J.H. (2024). Diagnostic uplift through the implementation of short tandem repeat analysis using exome sequencing. Eur. J. Hum. Genet. *32*, 584–587.
57. Rocca, C., Murphy, D., Clarkson, C., Zanovello, M., Gagliardi, D., Genomics, Q.S., Kaiyrzhanov, R., Alvi, J., Maroofian, R., Efthymiou, S., et al. (2025). Analysis of Short Tandem Repeat Expansions in a Cohort of 12,496 Exomes from Patients with Neurological Diseases Reveals Variable Genotyping Rate Dependent on Exome Capture Kits. Genes *16*, 169.
58. Mestek-Boukhibar, L., Clement, E., Jones, W.D., Drury, S., Ocaka, L., Gagunashvili, A., Le Quesne Stabej, P., Bacchelli, C., Jani, N., Rahman, S., et al. (2018). Rapid Paediatric Sequencing (RaPS): comprehensive real-life workflow for rapid diagnosis of critically ill children. J. Med. Genet. *55*, 721–728.
59. 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.
60. Kingsmore, S.F., Cakici, J.A., Clark, M.M., Gaughran, M., Feddock, M., Batalov, S., Bainbridge, M.N., Carroll, J., Caylor, S.A., Clarke, C., et al. (2019). A Randomized, Controlled Trial of the Analytic and Diagnostic Performance of Singleton and Trio, Rapid Genome and Exome Sequencing in Ill Infants. Am. J. Hum. Genet. *105*, 719–733.
61. Jezkova, J., Shaw, S., Taverner, N.V., and Williams, H.J. (2022). Rapid genome sequencing for pediatrics. Hum. Mutat. *43*, 1507–1518.
62. Fatumo, S., Chikowore, T., Choudhury, A., Ayub, M., Martin, A.R., and Kuchenbaecker, K. (2022). A roadmap to increase diversity in genomic studies. Nat. Med. *28*, 243–250.
63. Kopal, J., Uddin, L.Q., and Bzdok, D. (2023). The end game: respecting major sources of population diversity. Nat. Methods *20*, 1122–1128.
64. Weisschuh, N., Mazzola, P., Zuleger, T., Schaeferhoff, K., Kühlewein, L., Kortüm, F., Witt, D., Liebmann, A., Falb, R., Pohl, L., et al. (2024). Diagnostic genome sequencing improves diagnostic yield: a prospective single-centre study in 1000 patients with inherited eye diseases. J. Med. Genet. *61*, 186–195.
65. Yépez, V.A., Gusic, M., Kopajtich, R., Mertes, C., Smith, N.H., Alston, C.L., Ban, R., Beblo, S., Berutti, R., Blessing, H., et al. (2022). Clinical implementation of RNA sequencing for Mendelian disease diagnostics. Genome Med. *14*, 38.
66. Wojcik, M.H., Clark, R.D., Elias, A.F., Genetti, C.A., Madden, J.A., Simpson, D., Golkar, L., Zalusky, M.P.G., Miller, A.L., Rodriguez, A., et al. (2025). Long-read sequencing is required for precision diagnosis of incontinentia pigmenti. HGG Adv. *6*, 100468.