High-coverage Nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation

AUTHORS

Jonas A. Gustafson[1,2,*], Sophia B. Gibson[1,3,*], Nikhita Damaraju[1,4,*], Miranda PG Zalusky[1], Kendra Hoekzema[3], David Twesigomwe[5], Lei Yang[6], Anthony A. Snead[7], Phillip A. Richmond[8], Wouter De Coster[9,10], Nathan D. Olson[11], Andrea Guarracino[12,13], Qiuhui Li[14], Angela L. Miller[1], Joy Goffena[1], Zachary B. Anderson[1], Sophie HR Storz[1], Sydney A. Ward[1], Maisha Sinha[1], Claudia Gonzaga-Jauregui[15], Wayne E. Clarke[16,17], Anna O. Basile[16], André Corvelo[16], Catherine Reeves[16], Adrienne Helland[16], Rajeeva Lochan Musunuri[16], Mahler Revsine[14], Karynne E. Patterson[3], Cate R. Paschal[18,19], Christina Zakarian[3], Sara Goodwin[20], Tanner D. Jensen[21], Esther Robb[22], The 1000 Genomes ONT Sequencing Consortium, University of Washington Center for Rare Disease Research (UW-CRDR), Genomics Research to Elucidate the Genetics of Rare Diseases (GREGoR) Consortium, W. Richard McCombie[20], Fritz J. Sedlazeck[23,24,25], Justin M. Zook[11], Stephen B. Montgomery[21], Erik Garrison[12], Mikhail Kolmogorov[26], Michael C. Schatz[14], Richard N. McLaughlin Jr.[2,6], Harriet Dashnow[27,28], Michael C. Zody[16], Matt Loose[29], Miten Jain[30], Evan E. Eichler[3,31,32], Danny E. Miller[1,19,31,**]

AFFILIATIONS

1. Division of Genetic Medicine, Department of Pediatrics, University of Washington, Seattle, WA, USA
2. Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA
3. Department of Genome Sciences, University of Washington, Seattle, WA, USA
4. Institute for Public Health Genetics, University of Washington, Seattle, WA, USA
5. Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
6. Pacific Northwest Research Institute, Seattle, WA, USA
7. Department of Biology, New York University, New York, NY, USA
8. Alamya Health, Baton Rouge, LA, USA
9. Applied and Translational Neurogenomics Group, VIB Center for Molecular Neurology, VIB, Antwerp, Belgium
10. Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium
11. Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA
12. Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA
13. Human Technopole, Milan, Italy
14. Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA
15. International Laboratory for Human Genome Research, Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México
16. New York Genome Center, New York, NY, USA
17. Outlier Informatics Inc., Saskatoon, SK, Canada
18. Department of Laboratories, Seattle Children's Hospital, Seattle, WA, USA
19. Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA, USA
20. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
21. Department of Genetics, Stanford University, Stanford, CA, USA
22. Department of Computer Science, Stanford University, Stanford, CA, USA
23. Human Genome Sequencing Center Baylor College of Medicine, Houston, TX, USA

24. Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA
25. Department of Computer Science, Rice University, Houston, TX, USA
26. Cancer Data Science Laboratory, National Cancer Institute, NIH, Bethesda, MD, USA
27. Department of Human Genetics, University of Utah, Salt Lake City, UT, USA
28. Department of Biomedical Informatics, University of Colorado School of Medicine, Aurora, CO, USA
29. Deep Seq, School of Life Sciences, University of Nottingham, Nottingham, England
30. Department of Bioengineering, Department of Physics, Khoury College of Computer Sciences, Northeastern University, Boston, MA
31. Brotman Baty Institute for Precision Medicine, University of Washington, Seattle, WA, USA
32. Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

* These authors contributed equally to this work
** Corresponding author

RUNNING TITLE

Nanopore sequencing of 1KGP samples

KEYWORDS

Nanopore sequencing; long-read sequencing; 1000 Genomes Project; structural variation; repeat expansions; methylation

CORRESPONDING AUTHOR

Danny E Miller, MD, PhD
Assistant Professor
Department of Pediatrics, Division of Genetic Medicine
Department of Laboratory Medicine and Pathology
University of Washington
1705 NE Pacific St., HSB H474A, Seattle, WA 98195
Email: dm1@uw.edu

ABSTRACT

Fewer than half of individuals with a suspected Mendelian or monogenic condition receive a precise molecular diagnosis after comprehensive clinical genetic testing. Improvements in data quality and costs have heightened interest in using long-read sequencing (LRS) to streamline clinical genomic testing, but the absence of control datasets for variant filtering and prioritization has made tertiary analysis of LRS data challenging. To address this, the 1000 Genomes Project ONT Sequencing Consortium aims to generate LRS data from at least 800 of the 1000 Genomes Project samples. Our goal is to use LRS to identify a broader spectrum of variation so we may improve our understanding of normal patterns of human variation. Here, we present data from analysis of the first 100 samples, representing all 5 superpopulations and 19 subpopulations. These samples, sequenced to an average depth of coverage of 37x and sequence read N50 of 54 kbp, have high concordance with previous studies for identifying single nucleotide and indel variants outside of homopolymer regions. Using multiple structural variant (SV) callers, we identify an average of 24,543 high-confidence SVs per genome, including shared and private SVs likely to disrupt gene function as well as pathogenic expansions within disease-associated repeats that were not detected using short reads. Evaluation of methylation signatures revealed expected patterns at known imprinted loci, samples with skewed X-inactivation patterns, and novel differentially methylated regions. All raw sequencing data, processed data, and summary statistics are publicly available, providing a valuable resource for the clinical genetics community to discover pathogenic SVs.

INTRODUCTION

As an initiative to sequence a large set of healthy reference genomes from globally diverse ancestries, the 1000 Genomes Project (1KGP) marked a significant milestone in genomic research, yielding the first sequencing-based map of normal patterns of human genetic variation for filtering and prioritizing candidate disease-causing variants (International HapMap Consortium 2005; Byrska-Bishop et al. 2022; The 1000 Genomes Project Consortium 2015). The impact of 1KGP on our understanding of human genetic diversity has been enormous, and the flagship papers have been cited more than 10,000 times in clinical and basic research studies. The success of the project has been amplified by the use of diverse, high-quality, open-access datasets, and databases such as gnomAD (Koenig et al. 2023) and DECIPHER (Firth et al. 2009) have built on the 1KGP principles for determining the population allele frequency of variants to aid in variant interpretation. Pooling of data from large projects has improved the usefulness of these databases, and analyses of 1KGP data to date have made profound contributions using arrays or short-read sequencing technology. However, these approaches are inherently limited in their ability to identify variants in complex genomic regions or to capture certain types of genetic differences, such as structural variants (SVs), repeat expansions, and epigenetic changes (Ebert et al. 2021; Liao et al. 2023; Chaisson et al. 2019).

SVs—defined as insertions, deletions, duplications, inversions, repeat expansions, and translocations at least 50 bp in size—are major contributors to genetic diversity and disease susceptibility and are more likely to have a larger effect size than single nucleotide variants (SNVs) (Eichler 2019). SV calling using short-read sequencing can be challenging because it detects fewer than half of the ~25,000 SVs present in an individual, is incapable of fully resolving the complex structure of many SVs, and has low concordance between callers (Chaisson et al. 2019; Zhao et al. 2021; Cameron et al. 2019). These challenges extend into clinical testing where commonly used approaches, such as exome sequencing, have low sensitivity for SV detection, meaning individuals with disease-causing SVs may remain undiagnosed (Hiatt et al. 2021; Cohen et al. 2022; AlAbdi et al. 2023; Miller et al. 2021). Therefore, there is broad interest in using long-read sequencing (LRS) to develop comprehensive catalogs of common human SVs to facilitate improved detection of disease-associated variants (Wojcik et al. 2023).

LRS has increasingly demonstrated its ability to detect and resolve SVs missed by traditional methods. Previous concerns about cost, error rates, sample preparation, and computational tools for both commercially available LRS technologies (Pacific Biosciences,

PacBio and Oxford Nanopore Technologies, ONT) have largely been resolved (Logsdon et al. 2020; Wang et al. 2021; Kolmogorov et al. 2023), paving the way for its adoption into clinical settings (Wojcik et al. 2023; Damaraju et al. 2024).

Building on the landmark effort of the 1KGP, the 1000 Genomes Project ONT Sequencing Consortium (1KGP-ONT) is leveraging ONT LRS with the goal of generating high-coverage, high-quality sequencing data from the 1KGP sample set. This international initiative aims to: 1) assess both assembly-based and alignment-based approaches to LRS data analysis; 2) evaluate variants in difficult-to-analyze regions of the genome; and 3) facilitate the identification of SVs not fully characterized by short-read approaches. This effort is complementary to work from other groups performing PacBio LRS of 1KGP samples, such as the Human Pangenome Reference Consortium (HPRC) (Wang et al. 2022) and the Human Genome Structural Variant Consortium (HGSVC) (Ebert et al. 2021), as well as lower coverage and N50 ONT sequencing from Schloissnig et al. (2024). With these collective endeavors, it is increasingly likely that the entire collection will ultimately be sequenced using both LRS platforms. Following 1KGP principles, all data generated through the 1KGP-ONT Consortium are publicly released for immediate incorporation into clinical and basic research projects.

Here, we present our analysis of the first 100 samples sequenced by the 1KGP-ONT Consortium. Because a major goal of the consortium is to develop a catalog of common human SVs for filtering and prioritizing disease-associated SVs, we demonstrate how SV data from a modest number of individuals can be used to filter variants in unsolved cases and identify high-priority regions for follow-up analysis. We also describe variation that would be difficult or impossible to detect or fully resolve using short-read technology, including disease-associated repeat expansions, skewed X-Chromosome inactivation in 46,XX samples, and differentially methylated regions (DMRs) unique to individual samples.

RESULTS

Approximately 3,200 cell lines or DNA samples from the 1KGP are available at the National Human Genome Research Institute (NHGRI) Sample Repository for Human Genetic Research housed at the Coriell Institute for Medical Research repositories (Coriell) (International HapMap Consortium 2005; The 1000 Genomes Project Consortium 2015). These anonymized samples, which are not associated with medical or phenotypic data, are from individuals who self-reported ancestry, sex, and good health at the time of sample collection. We selected 100 samples from all 5 superpopulations based on their absence from other large-scale sequencing efforts (Liao et al. 2023; Ebert et al. 2021; Schloissnig et al. 2024); we did not attempt to balance subpopulations within these samples, and four of the 100 samples represent two parent–child pairs **(Figure 1A, Supplemental Table S1)**.

Sequencing pipeline

High molecular weight (HMW) DNA was isolated from lymphoblastoid cell lines (LCLs) cultured in the lab, and samples were sequenced using the ONT R9.4.1 pore with an average depth of coverage of 37.4x and read N50 of 53.8 kbp **(Figure 1B, Supplemental Table S2)**. All samples were processed using two separate pipelines **(Figure 1C)**. First, an internal alignment pipeline used minimap2 for alignment, Clair3 for small variant calling, and Sniffles2, CuteSV, and SVIM for SV calling (Heller and Vingron 2019; Jiang et al. 2020; Li 2018; Zheng et al. 2022; Smolka et al. 2024). Single nucleotide variant calls from this pipeline were used to ensure sample identity by comparison with previous short-read-based variant calls (Byrska-Bishop et al. 2022). Second, samples were processed using the Napu pipeline, which generates assembly-based SV calls using hapdiff after generating a phased *de novo* assembly using Shasta-Hapdup, minimap2 alignment-based small variant calls using Pepper-Margin-DeepVariant (PMDV), and minimap2 alignment-based SV calls using Sniffles2 (Kolmogorov et al. 2023; Shafin et al. 2021; Smolka et al. 2024).

Small variant accuracy

We evaluated the performance of our variant-calling pipelines by comparing small variant calls (SNVs and indels <50 bp) to those generated by prior studies and using orthogonal short- and long-read sequencing technologies. We first compared ONT sequencing of 5 samples (outside of our 1KGP cohort) to Genome in a Bottle benchmarking data and the Hi-Fi PacBio data from the Human Pangenome Research Project. Restricting analysis to the GIAB high confidence

regions for HG002 resulted in F1 scores >0.984 for SNVs and >0.699 for indels for both datasets **(Supplemental Table S3)**. However, these values were highly influenced by the presence of homopolymers in the ONT data (Kolmogorov et al. 2023; Harvey et al. 2023). When homopolymers were removed from the analysis, F1 scores increased to >0.984 for SNVs and >0.874 for indels **(Supplemental Fig. S1, Supplemental Table S4)**. Next, we compared ONT data from our 1KGP cohort to complementary Illumina data. We observed an average F1 score of 0.982 for SNVs and 0.878 for indels outside of homopolymers. **(Supplemental Fig. S2, Supplemental Table S5)**. These results validated that both variant-calling approaches (Clair3 and PMDV) produced high-quality small variant calls concordant with prior studies (Kolmogorov et al. 2023).

Genome assembly

We performed *de novo* genome assemblies for each of the 100 samples using both the Napu pipeline (which runs Shasta-Hapdup) and Flye (Kolmogorov et al. 2023; Shafin et al. 2020). In general, we found that Flye assemblies had a higher contig NG50 than Shasta-Hapdup assemblies **(Figure 2A),** and results were robust to read N50 differences **(Figure 2B)**. We saw similar contig NG50 patterns when our analysis included the 5 benchmarking genomes with similar average depth of coverage and read N50. The assembled genomes were highly complete, with each assembly covering approximately 93.5% (Flye) or 93.6% (Shasta-Hapdup) of the GRCh38 reference genome **(Supplemental Fig. S3)** with a consensus accuracy similar to previously published studies using the R9 pore **(Figure 2C)** (Kolmogorov et al. 2023).

We investigated why many of the Flye assemblies had similar contig NG50 values by plotting the contig breakpoints for both the Shasta-Hapdup and Flye assemblies. Among the 100 Flye assemblies, 97.1% of assembly breaks occurred within regions annotated as segmental duplications (segdups), satellite sequence, or both, while 2.9% occurred within nonrepetitive sequence **(Supplemental Table S6)**. Among the 2.9% of assembly breaks in nonrepetitive sequence, 90% were seen in only one sample, suggesting stochastic artifacts of the assembly process. A focused analysis of Chromosome 7 revealed an increased number of contig breaks in the telomeric and pericentromeric regions for both Flye and Shasta-Hapdup assemblies **(Figure 2D)** and at positions flanking well-described recurrent copy number changes associated with disease (Morris 1993). Visual analysis of breaks in nonrepetitive sequence did not reveal sample-specific differences that would easily explain the break in assembly, such as a duplication, inversion, or increased number of SNVs, suggesting that local sequence variation did not influence the position of assembly breaks in nonrepetitive regions

(Supplemental Fig. S4). A list of assembly breaks in 20 or more samples from either the Flye or Shasta-Hapdup assemblies genome-wide is available **(Supplemental File S1)**.

We then evaluated contig size across superpopulation groups and assembly of disease-associated OMIM genes. The median contig size per sample excluding contigs <1 Mbp **(Figure 2E)** was higher for African ancestry samples. This was expected given the higher genetic diversity in African ancestry samples, which results in a higher number of distinct sequences leading to longer and more contiguous sequences in the assembly. Next, we examined how well disease-associated genes were assembled in these samples. Among 4,615 disease-associated OMIM genes (excluding genes on the X and Y Chromosomes), we found that 97% (4,492/4,615) and 97% (4,475/4,615) of genes in the Flye or Shasta-Hapdup assemblies, respectively, were completely and correctly assembled (i.e., they were spanned by a single, complete contig) in at least 95 out of 100 samples **(Supplemental File S2)**. Among the 200 assemblies (100 Flye and 100 Shasta-Hapdup), we found that 5 OMIM genes were incompletely assembled in all 200 assemblies and another 45 OMIM genes were incompletely assembled in at least 50 or more of the 200 assemblies **(Figure 2F)**. We observed more incompletely assembled genes in the Shasta-Hapdup assemblies, partly due to the requirement for a single gene to be entirely spanned by a single contig in both haplotypes for it to be considered fully assembled.

We subsequently applied PGGB to construct chromosome-level pangenome graphs from the 100 Shasta-Hapdup assemblies and generate multi-sample variant calls including all types of variants (Garrison et al. 2023). To investigate the differences between assembly approaches, we performed principal component analysis (PCA) on a Chromosome 20 pangenome graph created by combining the 100 Shasta-Hapdup assemblies with 44 assemblies from the HPRC (Liao et al. 2023). The PCA showed a clear separation between the two pangenomes (**Supplemental Fig. S5A**). However, a PCA based on the euchromatic, non-centromeric fraction of the Chromosome 20 graph demonstrates that this difference is primarily due to the improved resolution of highly repetitive sequences by the HiFi-based HPRC assemblies (**Supplemental Fig. S5B**), supporting the high-quality nature of our assemblies.

Variation within active transposable elements

The largely repetitive and polymorphic nature of active transposable elements, especially full-length long interspersed element 1 (LINE-1) and endogenous retroviruses (ERV), makes them challenging to fully resolve and characterize using short-read assemblies (Yang et al. 2024). We anticipated that long-read assemblies would allow us to overcome these challenges. Using

RepeatMasker (http://www.repeatmasker.org/), we identified interspersed repeats in the 100 Shasta-Hapdup assemblies and found that the fraction of major interspersed repeats differs by no more than 3% compared to that of the T2T-CHM13 assembly (Nurk et al. 2022) **(Supplemental Table S7).** Furthermore, there was minimal variation among the 100 assemblies in interspersed repeat content.

Among the youngest polymorphic interspersed repeats that are too long to resolve with short reads (Chaisson et al. 2019), LINE-1s (~6,000 bp) are the only type that are actively expanding in the human genome. We found that the total base pairs of LINE-1 sequence (including young and old LINE-1s) in the 100 assemblies (496 Mbp average) is lower than observed in the CHM13 T2T assembly (512 Mbp), likely due to LINE-1s within unassembled regions. To measure the ability of these ONT-based assemblies to resolve young LINE-1s, we calculated the number of the youngest LINE-1 elements (L1HS) and the number of full-length (≥6 kbp) L1HS elements. Overall, we found similar numbers of L1HS and full-length L1HS sequences compared to HG002 and HG005 from GIAB and the CHM13 T2T assembly **(Supplemental Fig. S6)**. Although HERV-Ks (~9,000 bp) are unlikely to be actively replicating in modern humans, like LINE-1s, they are known to be polymorphic in the human population (Li et al. 2019; Subramanian et al. 2011). Therefore, we also counted the number of full-length HERV-Ks (HERVK-int) and found that the number per genome is similar among the 100 assemblies and CHM13 T2T, HG002, and HG005. This demonstrates that these assemblies are of sufficient quality to resolve the youngest long interspersed repeats and that there is variation in the number of these insertions among different human populations.

Structural variant analysis

We called SVs using four alignment-based and one assembly-based method (*see* Methods) and compared them to a known set of SV calls generated by the HPRC (Liao et al. 2023). From three of the five genomes used for small variant benchmarking (HG002/NA24385, HG00733, and HG02723) we identified an average of 23,732 SVs across all five callers. This is similar to the average of 22,755 SVs among 15 human genomes assembled by Audano *et al*. (2019) but less than those predicted by the HPRC and HGSVC (Ebert et al. 2021; Liao et al. 2023). The greater number than Audano *et al.* is expected given that those were called with older PacBio chemistries (RSII CLR) and an approach, SMRT-SV, that excluded SV calls in some pericentromeric regions or regions where variant calls were considered less reliable (Audano et al. 2019). Benchmarking against the HPRC Sniffles2 SV calls (Liao et al. 2023) and restricting calls to regions within the GIAB HG002 SV Tier1 v0.6 benchmarking regions (GIAB Tier1

Regions) (Zook et al. 2020) revealed F1 scores greater than 90% for both methods among all three samples **(Figure 3A)**. When comparing genome-wide SV calls (not restricted to the GIAB Tier1 regions), our F1 score decreased to approximately 70% for all three samples, suggesting difficulty in generating concordant SV calls in low-complexity or repetitive regions of the genome **(Supplemental Table S8)**.

We observed high per-caller concordance between the number of SV calls from the three benchmarking genomes and the 100 genomes presented here **(Figure 3B)**. Across the five callers, we identified an average of 24,543 SVs per sample (min: 20,068, max: 28,734), similar to the 23,000–28,000 SVs per sample reported by the HGSVC (Ebert et al. 2021). Consistent with prior work, we observed more total SV calls in samples from the African superpopulation (Ebert et al. 2021; Audano et al. 2019; The 1000 Genomes Project Consortium 2015). The distribution of insertions and deletions called in this dataset was also as expected, with an *Alu* peak around 300 bp and LINE peak around 6 kbp **(Figure 3C)**. A generally proportional number of SVs per chromosome was observed and, on average, more insertion than deletion events were identified per chromosome for all SV callers **(Supplemental Fig. S7)**. The genome-wide distribution of total SV events was as expected, with more insertions and deletions near the telomeres and centromeres **(Supplemental Fig. S8)**. We identified an increasing number of novel SVs, excluding breakends (BNDs), for each additional sample sequenced among all SV callers **(Figure 3D)**.

Because a primary goal of our study is to identify and catalog high-quality SVs among the 1KGP samples, we merged the SVs from each of the five SV callers per sample using Jasmine (Kirsche et al. 2023). We observed high concordance between SV callers across all samples **(Figure 3E)**, with an average of 16,722 SVs per sample called by all callers and no individual sample having an SV type that was noticeably higher or lower than other samples within the same superpopulation **(Supplemental Fig. S9A)**. An average of 20,242, 22,685, 25,540, and 34,796 SVs were called by at least four, three, two, or one callers, respectively **(Supplemental Fig. S9B)**.

The SVs called exclusively by hapdiff represent the majority of SVs called by a single caller. Because hapdiff was the only assembly-based caller in our dataset, we examined whether these calls represented false positives or SVs in regions where alignment may be challenging. Our analysis found that of the 407,779 SVs (excluding BNDs) called only by hapdiff across all 100 samples, 151,575 (37.1%) were fully or partially within a segdup or within 1,000 bp of a segdup, suggesting that they may be in complex copy-number polymorphic regions of the genome, and thus potential artifacts because of their proximity to a segdup. Of the SVs that

were not fully within, partially within, or within 1,000 bp of a segdup, 119,255 (46.5% of the remaining SVs) overlap a variable number tandem repeat (VNTR) region. Analysis of SVs called only by hapdiff did not reveal any individual sample or population outliers **(Supplemental Fig. S9C)**, and visual analysis of 30 randomly selected SVs from this set found that 28/30 were likely false-positive calls **(Supplemental Fig. S10)**. This suggests that difficult-to-assemble regions are a major source of false-positive assembly-based SV calls and that annotating SV calls with information about genomic context might provide insight into the confidence of these calls.

An SV frequency call set was generated that represented SVs called by all five callers (100,915 total SVs), four or more (119,805 total SVs), three or more (133,766 total SVs), two or more (155,407 total SVs), or at least one caller (252,954 total SVs). Among the 100 samples described here, there were a total of 113,696 shared or unique high-confidence SVs (SVs identified by hapdiff and 2 or more unique callers, excluding BNDs), with 32% found in only one sample (36,096 of 113,696). We found that 12,432 (11%) of these shared SVs were seen in exactly 2 samples, and that approximately half of these shared SVs were in samples only from the African superpopulation **(Figure 3F)**, similar to previous analysis (The 1000 Genomes Project Consortium 2015). Among 50,458 high-confidence SVs that intersect protein-coding genes, 97% (49,142/50,458) are within or include intronic sequence, 3.3% (1,654 / 50,458) are within or include coding sequence, and 2.0% (992/50,458) are within or include a 5' or 3' untranslated region (UTR).

To investigate the functional significance of SVs on gene expression, we performed an SV-eQTL analysis using the merged SV call set and the recently published MAGE dataset, which includes RNA-seq data from 731 samples from the 1KG cohort (Taylor et al. 2023) **(Supplemental Fig. S11A)**. Among 65 samples shared between MAGE and this study, we found 153 significant SV-eQTLs (q-value < 0.05), of which 37 were previously found using a collection of 31 diverse LRS-based genomes **(Supplemental Fig. S11B)**. This includes a 484-bp insertion associated with *ZNF79*, a gene implicated in neurological diseases (Bu et al. 2021) **(Supplemental Fig. S11C–D)**. This analysis also revealed several new significant associations, including an 81-bp deletion not previously detected (Kirsche et al. 2023) that is associated with the *NAPRT* gene, an important factor in cancer susceptibility (Duarte-Pereira et al. 2021) **(Supplemental Fig. S11E–F)**. To further explore the application of the variant call set for SV-eQTL discovery, we genotyped the SVs in all 731 MAGE individuals using their matched short-read genomic data from Byrska-Bishop et al. (2022). Using the 65 samples common to both the 1KGP-ONT and MAGE datasets, we found the genotype consistency was >98% between the short- and long-read datasets after filtering for tandem repeats and Hardy-Weinberg consistency

(Supplemental Fig. S11G). Across all 731 samples, we identified 1,324 significant SV-eQTLs, of which 1,258 were uniquely in the short-read data, including a 2,716-bp deletion associated with *GBP3*, a gene implicated in infectious diseases and immune responses (Tretina et al. 2019) **(Supplemental Fig. S11H)**.

Structural variation within medically relevant genes

Sequencing of samples from all five superpopulations allowed us to evaluate population-specific SVs intersecting genes associated with an OMIM phenotype (n = 4,866) and revealed 349 high-confidence SVs in or including at least one defined exon **(Supplemental Fig. S12A, Supplemental Table S9)**. These events ranged in size from 50 bp (deletions in *TNFRSF13C* and *TF* and an insertion in *IMPG2*) to 87,776 bp (a deletion that fully includes *IGHM*). Visual analysis of 30 randomly selected events confirmed that all were likely true positives. These 349 SVs are distributed across all chromosomes and impact 335 exons in 236 unique OMIM genes, with 123 of those 335 exons containing ClinVar variants that are annotated as pathogenic or likely pathogenic **(Supplemental Fig. S12B)**. We found that 150/349 (43%) of these SVs were found in only one sample, and no single sample had more than 6 unique SVs (HG01369). Three SVs (a 458-bp insertion in *ABCC11*, a 243-bp insertion in *XYLT1*, and a 118-bp insertion in *MED13L*) were seen in all 100 samples, suggesting the reference genome represents a minor allele at these positions. Indeed, GRCh38 has been patched to include a similar insertion in *XYLT1*. Of the 38 SVs observed in only 2 samples, 76% (29/38) were superpopulation-specific with 55% of those (16/29) seen in samples from the African superpopulation. We observed 4 SVs spanning multiple genes, some of which are known population variants. This includes a 22.8-kbp deletion spanning *HBB*, *HBD*, and *HBG1* associated with beta thalassemia (Huisman et al. 1972) (MIM: 613985) and two samples with a 19,304-bp deletion including *HBA1* and *HBA2* commonly referred to as the Southeast Asian deletion (Farashi and Harteveld 2018) (MIM: 604131) (**Supplemental Fig. S13)**.

We did not expect to find rare SVs in X-linked OMIM genes in 46,XY samples, since those events would be more likely to be associated with a disease. However, we did find five such events in at least one 46,XY sample. Of these, four were in a 3'UTR and were observed in at least two 46,XX samples. One of the four events, found in only one sample, was an approximately 141-bp insertion in exon 15 of *RPGR* (OMIM: 312610), a gene associated with several X-linked conditions including retinitis pigmentosa, cone-rod dystrophy, and macular degeneration (Fahim et al. 1993). A similar insertion at this position has been reported twice in ClinVar as a variant of uncertain significance (VUS) associated with primary ciliary dyskinesia,

once as a 141-bp insertion (ClinVar entry 2121719) and once as a 69-bp insertion (ClinVar entry 1975740). Evaluation of the short-read sequencing data for this sample at this position did not clearly demonstrate the insertion, but the insertion consists of only C- and T- nucleotides, which would make it difficult to align and evaluate using short-read technology **(Supplemental Fig. S14)**. The presence of this insertion in a 46,XY 1KGP sample suggests that this variant may be present at a higher allele frequency than expected, is difficult to reliably call using short-read technology, or could be associated with a later onset of the associated phenotype.

A substantial number of high-confidence SVs were observed in regions of the genome difficult to evaluate using short-reads, meaning they may be filtered by variant annotation pipelines. For example, 42% (47,315/113,696) of the high-confidence SVs occur fully outside of the GIAB Tier 1 regions, and visual inspection of 30 events confirmed the presence of an SV. We also identified 407 high-confidence SVs within coding regions defined as unreliable for variant identification using short-read sequencing based on analysis of gnomAD data (Hijikata et al. 2024). Finally, 9,788 of the high-confidence insertions were ≥ 500 bp, which may preclude accurate resolution of these events and limit our understanding of their impact on gene expression or splicing when evaluated using short-read technology.

Cytochrome P450 (CYP) genes impact drug response and are among the gene sets that are challenging to interrogate using short-read technologies and may require separate variant calling approaches to fully evaluate (Lee et al. 2019, Zanger and Schwab 2013). Within this dataset, LRS enabled better resolution of full gene deletion and duplication SV events in highly polymorphic CYP pharmacogenes such as *CYP2D6*, a pharmacogene involved in the metabolism of over 20% of clinically prescribed medications (Zanger and Schwab 2013). For example, we identified one individual (HG02396) with a *CYP2D6* gene deletion (**5*) on one haplotype and a hybrid tandem arrangement (**36+*10*)—shown via an insertion—on the second haplotype (**Supplemental Fig. S15A)**. In the equivalent short-read WGS data, it can be difficult to identify both the gene deletion and the hybrid tandem star allele in the same individual using specialized short-read genotyping tools (Twesigomwe et al. 2023). Analysis of a known complex *CYP2B6* star allele (*CYP2B6*29*) showed that it was called by hapdiff but not the alignment-based callers, demonstrating that some of these complex alleles may not be represented in our initial high-confidence SV set (**Supplemental Fig. S15B**) (Twesigomwe et al. 2024).

We used Jasmine to test whether the SVs identified in these 100 samples could be used to accurately filter SVs in 16 cases with known disease-associated SVs identified by whole-genome (8 cases) or targeted (8 cases) ONT sequencing (Wilderman et al. 2024; Miller et al. 2021) **(Supplemental Table S10)**. Among the 8 cases that had undergone whole-genome LRS,

filtering reduced the average number of SVs called by Sniffles2 by 93% (from 22,743 to 1,664), and in all 16 cases the pathogenic SV was retained after filtering. Subsequent annotation of the filtered SVs (i.e., if the SV intersects with a gene, if that gene is associated with an OMIM phenotype, if the SV is exonic, if the SV is within a segmental duplication or low complexity region, etc.) allowed us to substantially further narrow the output candidate SVs. This demonstrates that the high-confidence SV calls can be used to filter SVs in cases with high suspicion of a monogenic condition.

Analysis of disease-associated repeat expansions

Tandem repeat expansions (e.g., short tandem repeats (STRs) and VNTRs) at more than 60 loci have been implicated in human diseases such as the GGC expansion in the 5'UTR of *XYLT1* (MIM: 608124) associated with Baratella-Scott syndrome (MIM: 300881) (Depienne and Mandel 2021; Hannan 2018). Pathogenic repeat expansions associated with monogenic disease can be difficult to precisely size or fully sequence-resolve using short-read sequencing, meaning clinically relevant interruptions in the repeat may not be easily identified (Chaisson et al. 2023; Tanudisastro et al. 2024). Thus, there is interest in using LRS to evaluate repeat expansions genome-wide and at clinically relevant loci (Dolzhenko et al. 2023; Reis et al. 2023; Sulovari et al. 2019).

We used vamos to perform genome-wide haplotype-resolved analysis of 562,005 loci—including 66 disease-associated loci—consisting of both simple and complex repeat units, and identified pathogenic-sized expansions in *RFC1*, *ATXN10*, *FGF14*, and *ATXN80S* **(Figure 4A, S16–S19; Supplemental File S3)** (Hiatt et al. 2024). We also identified alleles over the pathogenic threshold but with a benign motif in *SAMD12*, *BEAN1*, and *DAB1*, as well as several alleles at *AR* where the total repeat count was over the threshold but the CAG motif was only a portion of the region.

Expansions in *RFC1*, which are associated with autosomal recessive cerebellar ataxia, neuropathy, and vestibular areflexia syndrome (CANVAS, MIM #614575), were observed in five samples ranging from 359 to 712 repeat units in size **(Figure 4B)**. Pathogenic expansions in this gene are typically 400 repeat units or larger and are motif-dependent, with AAGGG being the most common pathogenic expansion (Cortese et al. 2019; Scriba et al. 2020; Beecroft et al. 2020). Our observation that some of these samples carried the AAGGG repeat unit while others carried a nonpathogenic repeat unit, such as AAAAG, was similar to recent work that identified expansions in *RFC1* of varying repeat motifs in 5/100 HPRC samples (Dolzhenko et al. 2023) **(Figure 4C)**. That we observed an expansion in 5% of samples was not unexpected, as the

carrier frequency of *RFC1* expansions has been reported at 1–5% across at least two populations (Fan et al. 2020; Akçimen et al. 2019).

Expansions in *ATXN10* are associated with autosomal dominant spinocerebellar ataxia type 10 (SCA10, MIM #603516), a slowly progressive ataxia with typical age of onset between 12 and 48 years and full-penetrance alleles varying from 800 to 4,500 ATTCT repeats (Matsuura and Ashizawa 1993; Alonso et al. 2006; Raskin et al. 2007). Two of the 100 samples were heterozygous for *ATXN10* alleles larger than 800 motifs, one of which had a second allele with 511 repeat units **(Figure 4D)**. In addition, two other samples harbored expansions close to or larger than 280 repeat units, which has been reported as causative in one individual with ataxia (Matsuura et al. 2006). However, three of the four large alleles are purely ATTCT, and evidence suggests that interruptions of ATTCC are necessary for the allele to be pathogenic (Morato Torres et al. 2022).

To determine whether any of the expanded *RFC1* and *ATXN10* alleles would be identified using short-read data, we ran ExpansionHunter on short-read data from all affected samples (Dolzhenko et al. 2019). In all cases, when an expanded allele was present, the corresponding ExpansionHunter estimate was larger than the normal allele but, in most cases, still significantly underestimated the size of the expansion **(Figure 4C–D, Supplemental Table S11)**. For example, in *ATXN10*, LRS identified a normal allele (15 repeat units) and an expansion of more than 1,000 repeat units in HG01122. The ExpansionHunter estimates for this sample are 15 (range 15–15) and 73 (range 56–101) repeat units, thus the normal allele was correctly estimated but the expanded allele was markedly underestimated.

Evaluation of genome-wide methylation patterns and identification of novel DMRs

An advantage of LRS is the ability to simultaneously capture both DNA sequence and modification information, allowing for simultaneous evaluation of how changes in sequence, such as a repeat expansion, may alter the local epigenetic landscape. We evaluated methylation both genome-wide and at loci associated with imprinting disorders. Among 69 of the 70 46,XX samples sequenced, we found that 39% (27/69) had X-Chromosome methylation patterns suggestive of skewed X-inactivation **(Figure 5A, Supplemental Table S12)**.

We then performed genome-wide PCA of methylation to evaluate whether samples would correlate with ancestry or if patterns of X-inactivation would be apparent **(Supplemental Fig. S20)**. This analysis revealed that GM18864 clustered with 46,XY samples despite being reported as 46,XX. Because we validated each sample using SNVs from short-read sequencing, we wondered whether this sample had lost an X Chromosome. We found that the

average X Chromosome depth of coverage was approximately 55% of the full-length autosomes in the LRS data and approximately 75% in the short-read data, confirming loss of an X Chromosome in this sample (Pedersen et al. 2020).

Next, we evaluated methylation patterns at two disease-associated loci: 11p15.5, which is associated with both Beckwith-Wiedemann syndrome (BWS, MIM #130650) and Silver-Russell syndrome (SRS, MIM #180860) (Saal et al. 1993; Shuman et al. 1993); and 15q11.2-q13, associated with Prader-Willi syndrome (PWS, MIM #176270) and Angelman syndrome (AS, MIM #105830) (Dagli et al. 1993; Driscoll et al. 1993). For the 11p15.5 region, we found that in all samples, one haplotype was completely methylated while the other was completely unmethylated at imprinting centers IC1 and IC2 **(Figure 5B)**. Evaluation of haplotype-resolved methylation at the *SNURF-SNRPN* locus on 15q11.2 revealed two samples, GM19473 and HG00525, where one haplotype was 25%–75% methylated. Visual evaluation of these samples showed that one haplotype of GM19473 had increased methylation while one haplotype of HG00525 had reduced methylation, which was unexpected and further demonstrates that changes in methylation can occur throughout the genome in these cell lines, even at well-established DMRs **(Supplemental Fig. S21)**.

We used MeOW (Zalusky and Miller 2024) to analyze differences in methylation at CpG sites genome-wide and identified 134 CpGs with methylation differences across 37 samples, with a median of 2 DMRs per sample. **(Supplemental Table S13)**. As an example, 3 DMRs were found in HG02389 **(Figure 5C)**, including a hypermethylated CpG in *SLC29A3* not present in controls **(Supplemental Fig. S22)**. We observed both hypermethylation (86 CpGs) and hypomethylation (48 CpGs) among the 134 CpGs and identified four samples with more than 10 DMRs **(Supplemental Fig. S23)**. Among the 15 samples from the African superpopulation with a DMR, there was an enrichment of expression outliers near the DMR with increasingly stringent *Z*-score thresholds, suggesting associated changes in gene expression **(Supplemental Fig. S24)**.

DISCUSSION

Current approaches to clinical genetic testing are incomplete as they are unable to capture the full spectrum of disease-causing variation (Wojcik et al. 2023). This is because: 1) new technologies, such as LRS, are not yet widely implemented in clinical labs; 2) computational tools are not yet able to efficiently capitalize on the data provided by these new technologies,

and those that can have substantial computational requirements; and 3) databases are not yet available for filtering and prioritizing variants identified using new technologies. The 1KGP-ONT Consortium plans to sequence at least 800 1KGP samples to generate a more complete catalog of variation, especially rare yet presumably benign variants across the 1KGP populations. While the expanded collection will enable a more accurate estimate of allele frequency for challenging variants and add information about haplotype resolved epigenetic variation, we acknowledge that this cohort represents a limited representation of human diversity, notably excluding individuals of indigenous Australian and Middle Eastern ancestries.

Here, we describe the initial analysis of the first 100 samples sequenced to an average of 30x depth of coverage and average read N50 >50 kbp, which was possible because of the use of HMW DNA isolated directly from cell culture **(Figure 1)**. This resulted in high sensitivity for SV detection – especially larger duplications and repeat expansions – using both assembly- and alignment-based approaches. We identified an average of 24,543 SVs per sample, similar to prior analysis of other 1KGP samples by the HGSVC and HPRC (Ebert et al. 2021; Liao et al. 2023). Our efforts complement recent work that identified approximately 16,000 SVs from ~1000 1KGP samples sequenced to lower average depth of coverage (15x) and median read length (6.2 kbp) (Schloissnig et al. 2024). While the difference in total SVs underscores the advantage of sequencing HMW DNA, further analysis will be required to fully assess the significance of the differences between these datasets.

We performed one of the most comprehensive benchmarking analyses to date of SNVs, indels, and SVs using data from the ONT platform. Consistent with prior studies, data generated on the ONT platform has a higher recall and precision than Illumina-based approaches for SNVs in well-characterized genomic regions and performs well for indels, specifically outside of homopolymers (Kolmogorov et al. 2023). Because all data from these first 100 samples were generated on the R9.4.1 pore, we anticipate that improvements in chemistry, such as the use of the R10.4.1 pore, will reduce context-specific errors and result in improved concordance with truth sets. Because of this, we have transitioned ongoing sequencing to the R10.4.1 pore. SV benchmarking also revealed high F1 scores for three samples for which orthogonal calls were available, highlighting how the R9.4.1 pore is sufficient for this application. Over time, we anticipate additional updates to ONT chemistry or software, and plan to evaluate each change carefully before data re-analysis or changing the chemistry used for this effort.

SVs were called using four alignment-based and one assembly-based caller. After merging, a high-confidence SV call set comprising 124,927 SVs was generated that we show can be used for filtering and variant prioritization. Genome-wide evaluation of these high-

confidence SVs revealed 349 that were within or encompassed an exon of a medically relevant gene. The low number of SVs intersecting medically relevant genes was reassuring, as we expect there to be selection against these events within coding regions of the genome. Nevertheless, we did identify one SV—an approximately 141-bp insertion in exon 15 of *RPGR*, a gene with an X-linked phenotype—in a 46,XY sample near two similar insertions that have been reported as VUSs in ClinVar. Because the 1KGP samples came from presumably healthy individuals, it could be that this event is associated with a later onset of an associated phenotype or that the insertion is benign. Identification of this insertion in a 1KGP sample is valuable as it may lead to functional studies that clarify the nature of the variant. Analogous to what has been reported for the relatively common occurrence of single nucleotide loss-of-function mutations in otherwise healthy individuals, the presence of an SV in a gene does not necessarily imply the variant is pathogenic (MacArthur et al. 2012). Indeed, early studies of human population samples using SNP microarrays identified extremely rare CNVs > 500 kbp in length among individuals without overt disease (Cooper et al. 2011).

Genome-wide evaluation of select repeat expansions revealed expansions in complex alleles not previously reported and difficult to identify using short-read technology **(Figure 4)**. We identified repeat expansions associated with diseases that are difficult to fully interpret because the individuals recruited to the 1KGP were presumably healthy. These individuals may be at risk of developing symptoms later in life, or they may be carrying alleles that are benign because of nonpathogenic motif composition or sequence interruptions that we did not detect. Alternatively, these expansions may simply be an artifact of the cell culture process and should be considered when these samples are used in other experiments or when these data are used for variant filtering and prioritization. We anticipate that comparison of this dataset to larger efforts, such as *All of Us*, will allow us to better understand whether these variants represent artifact from the cell culture process or true human genetic variation.

Finally, we evaluated patterns of methylation genome-wide and at loci associated with disease. We observed large-scale changes, such as skewed X-inactivation, in over one-third of 46,XX samples as well as unique changes, such as novel differential methylation that correlates with changes in local gene expression. These changes provide a mechanism by which distinct signals from samples maintained in cell culture can be explained and demonstrates the potential limitations of using immortalized cell lines to infer epigenetic signatures.

Sequencing of 1KGP samples is ongoing and we expect the analysis of a larger number of samples to further refine many of the findings in this study. Most analysis presented here was performed using GRCh38 as a reference due to its widespread use in clinical and research

laboratories; work is ongoing to evaluate the impact of the more complete CHM13 T2T genome on variant calling (Nurk et al. 2022). Overall, we anticipate that the dataset provided here will hasten the use of LRS to evaluate individuals with suspected Mendelian conditions for whom a precise molecular diagnosis remains elusive. This work not only provides valuable resources for candidate variant filtering and analysis but also emphasizes the critical need for ongoing investment in technology, software, and database development to fully realize the benefits of LRS. The more comprehensive analysis that can be performed using LRS— such as the identification and resolution of complex SVs, improved phasing, and incorporation of associated methylation information—will allow clinical and research teams to stop focusing on “what’s the next best test” when evaluating an individual with a suspected genetic condition and instead focus on interpreting those variants that were previously difficult to detect or that may involve a novel gene. Together, these efforts will lead to improved clinical outcomes, new gene-phenotype associations, the use of novel therapies, and an end to the diagnostic odyssey for many of the individuals and their families who are living with an unsolved or incompletely understood genetic condition.

METHODS

DNA extraction, sequencing, alignment, validation, and variant calling

DNA for sequencing was isolated from B Lymphocytes obtained from the NHGRI Sample Repository at the Coriell Institute for Medical Research. After sequencing and quality checks **(Supplemental Table S2)**, an internal alignment pipeline and the Napu pipeline (Nanopore Analysis Pipeline) were run prior to variant calling and annotation (Kolmogorov et al. 2023). Additional detail can be found in Supplementary Methods.

SNV and indel benchmarking and comparison with Illumina data

Original sequencing data for 5 benchmarking samples was base called with Dorado 0.5.0 (ONT) and downsampled to match the depth of coverage of the 100 study samples, then processed with both the internal alignment pipeline and the Napu pipeline. Long-read SNV and indel calls from the HPRC and GIAB (Shafin et al. 2020; Liao et al. 2023) and short-read SNV and indel calls from GIAB were obtained (Danecek et al. 2021; Wagner et al. 2022) and preprocessed. Benchmarking comparisons and comparisons with Illumina data were conducted using hap.py (Illumina), with analysis limited to high-confidence regions.

De novo genome assembly and evaluation

Flye (v2.9.2) (Kolmogorov et al. 2019) and Napu (Shasta-Hapdup) (Kolmogorov et al. 2023) were used for haploid and diploid genome assembly then aligned to the GRCh38 reference genome using minimap2 (v2.24) (Li 2018), with starts and ends of aligned contigs determined using BEDTools (v2.3.0) (Quinlan and Hall 2010). Assembly breakpoints were characterized using precomputed segdup and RepeatMasker positions downloaded from UCSC (Bailey et al. 2002; Kent et al. 2002) then categorized as Satellite, SegDup, SegDup+Satellite, or Neither.

SV analysis, merging, and benchmarking

SV calls were parsed using BCFtools for variants that passed filtering criteria, were ≥ 50 bp, and were assigned to a full-length chromosome. SVs were counted by type and length per sample and caller. Novel SVs per sample were calculated through iterative merging by Jasmine. To benchmark SV calling methods, ONT data from HG002/NA24385, HG00733, and HG02723 were processed using the Napu pipeline. SV calls for Sniffles2 and hapdiff were benchmarked to the HPRC (truth) calls using Truvari (v4.1.0) (English et al. 2022). The GIAB HG002 SV Tier1 benchmarking BED was used to define regions for inclusion. Additionally, we benchmarked

HG002 SV calls against the draft GIAB T2TQ100 HG002 GRCh38 SV benchmark. SVs per individual were analyzed for multi-caller concordance based on Jasmine merging. SVs meeting a threshold of support (described in Supplemental Methods) were reported as high-confidence. SVs from this high-confidence call set were further annotated with functionally relevant genomic information (i.e. intersection with exonic regions, genes associated with OMIM phenotypes, centromeric/telomeric regions, etc.) as defined by GENCODE release 45.

Filtering and prioritization of SVs

Sniffles2 SV calls from cases known to have a disease-causing SV were preprocessed as above and merged using Jasmine with Sniffles2 SV calls from the Napu pipeline from the 100 samples.

Pangenome construction

Contigs from the Shasta-Hapdup assemblies were partitioned by chromosome by mapping them against the human reference genomes using WFMASH (v0.12.6, commit 0b191bb) pangenome aligner (Marco-Sola et al. 2021).

eQTL analysis

We applied the SV-eQTL analysis from Kirsche et al. (2023) to the 65 samples with both long-read DNA and short-read RNA data from MAGE and analyzed them as described in Supplemental Methods.

Tandem repeat genotyping

Repeats were genotyped using vamos v1.2.6 (Ren et al. 2023). A BED file with the coordinates and metadata for each STRchive locus is provided.

Methylation analysis

Haplotype-resolved, whole-genome methylation pileup files were generated using Modkit v0.1.11 (ONT) from the PMDV haplotagged BAM file. For X Chromosome analysis, the average fraction of methylated reads was calculated for each CpG island. CpG islands at disease-associated loci were subsetted and the average fraction of reads methylated was calculated per sample and per haplotype. Unique DMRs were identified using Methylation Operation Wizard (MeOW) by a leave-one-out analysis (Zalusky and Miller 2024).

DATA ACCESS

Data for all samples sequenced as part of the 1000 Genomes Project ONT Sequencing Consortium are publicly available at https://s3.amazonaws.com/1000g-ont/index.html and scripts used in the analysis can be found at https://github.com/millerlaboratory/1000g_ONT (Supplemental Scripts). Data from the 100 samples reported here, as well as summary analysis data, are available at https://s3.amazonaws.com/1000g-ont/index.html?prefix=FIRST_100_FREEZE/. Data and code related to pangenome analyses are available at https://github.com/AndreaGuarracino/1000G-ONT-F100-PGGB.

COMPETING INTERESTS STATEMENT

WDC, ML, FJS, and DEM have received research support and/or consumables from ONT. WDC, JG, SBG, FJS, and DEM have received travel funding to speak on behalf of ONT. DEM is on a scientific advisory board at ONT and holds stock options in MyOme. FJS has received research support from Illumina, Genentech, and PacBio. SBM is an advisor to BioMarin, MyOme, and Tenaya Therapeutics. EEE is a scientific advisory board member of Variant Bio, Inc.

ACKNOWLEDGMENTS

SBG is supported by NIH grant 5T32HG000035-29; WDC is a recipient of a postdoctoral fellowship from FWO [12ASR24N]; EG and AG are supported by NIH grants R01HG013017 and U01DA057530 and NSF grant 2118744; SG is supported by NIH grant 5R50CA243890; TDJ is supported by NIH grant T32HG000044; MK is supported by Intramural NIH funding; SBM, TDJ, and ER is supported by NIH Grant U01HG011762; MCS is supported by NIH grants U24HG010263, R03CA272952, and U01CA253481 and the Lustgarten Foundation grant 90101412; FJS is supported by NIH grants 1U01HG011758-01, 1UG3NS132105-01, and U01AG058589; AAS is supported by an NSF postdoctoral research fellowship in biology [NSF 22-623]; RNM and LY are supported by NIH grants 5R35GM142733-03 and 5R21AI174130-02; EEE is supported by NIH grant HG010169 and is an investigator of the Howard Hughes Medical Institute; DEM is supported by the NIH Director's Early Independence Award DP5OD033357. The GREGoR Consortium is funded by the National Human Genome Research Institute of the National Institutes of Health, through the following grants: U01HG011758, U01HG011755, U01HG011745, U01HG011762, U01HG011744, and U24HG011746. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. No human subjects, live vertebrates, or higher invertebrate research was

undertaken as part of this manuscript. Certain commercial equipment, instruments, or materials are identified to adequately specify experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose.

Author contributions. Conceptualization: EEE, DEM. Data curation: JAG, SBG, ND, MPGZ, KH, HD. Formal analysis: JAG, SBG, ND, DT, LY, PAR, WDC, NDO, AG, QL, WEC, AOB, AC, AH, RLM, MR, TDJ, ER, JMZ, EG, MCS, RNM, EEE, DEM. Funding acquisition: EEE, DEM. Investigation: MPGZ, KH, JG, ZBA, SHRS, SAW, MS. Methodology: JAG, SBG, ND, MPGZ, KH, WDC, CER, AH, SG, WRM, MK, HD, ML, MJ, HD, EEE, DEM. Resources: MCZ, ML, MJ, EEE, DEM. Supervision: EEE, DEM. Visualization: JAG, SBG, ND, DT, LY, WDC, AG, ALM, EG, RNM. Writing – original draft: JAG, SBG, ND, MPGZ, DT, LY, WDC, ALM, JG, SHRS, WEC, RNM, EEE, DEM. Writing – review & editing: JAG, SBG, ND, KH, AAS, NDO, AG, QL, ALM, CGJ, WEC, AOB, AC, CER, MR, KEP, CRP, CZ, SG, TDJ, WRM, FJS, SBM, EG, MK, MCS, HD, MCZ, ML, MJ, EEE, DEM.

FIGURE LEGENDS

Figure 1. Summary statistics of samples, sequencing and small variant detection.
A. Samples selected for sequencing are shown by superpopulation and sex. **B**. Violin plots showing average read length, read N50, and average depth of coverage for all 100 samples. **C**. DNA was extracted from cells grown from aliquots received from Coriell and sequenced using the R9.4.1 pore. Data was analyzed using both alignment- and assembly-based approaches.

Figure 2. Summary of *de novo* assembly results.
A. Contig NG50 compared to total number of contigs shows that haploid assemblies generated by Flye are longer and have fewer contigs than Shasta-Hapdup. No contig NG50 generated by Flye exceed 40 Mbp. Assemblies for each benchmarking sample show similar statistics. **B.** Assembly NG50 does not significantly improve with higher read N50. **C**. QV scores for both Flye and Shasta-Hapdup assemblies, and the five benchmarking genomes. **D**. Count of contig breaks for all 100 samples on Chromosome 7 show that while assembly breaks cluster there are a large number of single breaks spread across the chromosome. The 1.5–1.8 Mbp Williams-Beuren syndrome critical region is indicated with a dashed box and is flanked by clusters of assembly breaks within segdups (Morris 1993). **E**. Contig sizes filtered for contigs longer than 1 Mbp for each superpopulation. **F**. OMIM genes incompletely assembled in 50 or more samples using Flye or Shasta-Hapdup. For Shasta-Hapdup, if one haplotype was completely assembled in a sample but the other was incomplete, the gene is counted as incompletely assembled. Assembly of 5 genes (*FAM20C*, *HYDIN*, *NOTCH2NLC*, *PRKAR1B*, and *SHANK2*) was incomplete for all 100 samples using both assemblers. Genes that are not in or do not contain a segdup are in bold with an asterisk.

Figure 3. SV call set.
A. SV calls were benchmarked against HPRC Sniffles2 SV calls within the GIAB HG002 SV Tier1 benchmarking regions. **B**. A similar number of genome-wide SVs were identified by all five callers used in this study. The confident call set is defined as variants called by hapdiff and at least 2 unique alignment-based callers. For each call set the average number of deletions (DEL), insertions (INS) and total SVs (including INV, DUP and BND events) per sample is shown. **C.** Histogram of insertion and deletion counts stratified by size. The peak around 300 bp represents *Alu* insertions or deletions, and the peak around 6 kbp represents LINE insertions or deletions. **D.** Cumulative novel SVs per sample. The frequency of new SVs observed increases

when samples from individuals of African ancestry are included. **E.** Upset plot of overlap among SV callers after merging with Jasmine. For each sample, 5 VCF files were merged, demonstrating that the majority of calls in each sample were called by all 5 callers. **F.** Among 113,696 SVs from the Jasmine-merged confident call set, 12,432 were found in exactly 2 samples, with 6,181 (50%) of those calls in pairs in which both samples are from the African superpopulation.

Figure 4. Evaluation of repeat expansions known to be associated with Mendelian conditions.

A. Haplotype-resolved repeat expansions of selected repeat loci for simple and complex repeat units. Pathogenic repeat size is shown to the right of each plot (*), the associated condition is in parentheses, and the full name of each condition can be found in Supplemental Table S11. The pathogenic repeat size for *FMR1* is listed as 200 repeats, but a dashed vertical line represents the 55-repeat threshold that puts 46,XX and 46,XY individuals at risk for fragile X-associated tremor/ataxia syndrome (FXTAS, MIM #300623) and 46,XX individuals at risk of fragile X-associated primary ovarian insufficiency (POF1/FXPOI, MIM #311360). (AD, autosomal dominant; AD/AR, autosomal dominant/recessive; AR, autosomal recessive; XR, X-linked recessive; XD, X-linked dominant.) **B.** Among 200 haplotypes (*y*-axis), an expansion in *RFC1* near or over 400 repeat units was seen in 5 haplotypes. AAGGG is the most common pathogenic repeat expansion; additional pathogenic expansions include ACAGG (not shown), and a mixed AAAGG/AAGGG expansion.(Cortese et al. 1993) **C:** Haplotype (HP)-resolved detail of *RFC1* repeat expansions in five samples with an expansion of one allele. Haplotypes are assigned arbitrarily. Dotted line represents the position of full penetrance alleles typically seen at 400 repeat units. **D:** Three samples with expansions in *ATXN10* larger than 280 ATTCT repeats were observed. The dotted line at 800 repeat units represents the position of the lower end of the full penetrance range. ExpansionHunter (EH) estimates are overlayed atop the bar plots in (C) and (D), placed on HP1 or HP2 based on their length.

Figure 5. Patterns of methylation among the 1000 Genomes samples.

A. Among 69 46,XX samples, 42 had mixed X-Chromosome inactivation (top, example from HG01414), while 27 were skewed (bottom, example from HG01801). The color differences are related to breaks in phasing and do not suggest methylation is mixed along a single haplotype. **B.** Haplotype-resolved methylation fraction is shown for three imprinted loci associated with four imprinting disorders. Methylated (>75%) or unmethylated (<25%) fraction at IC1 in *H19* and IC2

in *KCNQ1OT1*. Haplotype-resolved methylation fraction is also shown for the CpG island within *SNURF-SNRPN* that is evaluated when testing for PWS or AS. Two samples have either gain (GM19473) or loss (HG00525) of methylation at this locus. **C.** Unique methylation differences within defined CpG islands were identified in individual samples. An example from HG02389 shows three CpG sites with increased methylation (red boxes) compared to controls (gray).

REFERENCES

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

Akçimen F, Ross JP, Bourassa CV, Liao C, Rochefort D, Gama MTD, Dicaire M-J, Barsottini OG, Brais B, Pedroso JL, et al. 2019. Investigation of the RFC1 Repeat Expansion in a Canadian and a Brazilian Ataxia Cohort: Identification of Novel Conformations. *Front Genet* **10**: 1219.

AlAbdi L, Shamseldin HE, Khouj E, Helaby R, Aljamal B, Alqahtani M, Almulhim A, Hamid H, Hashem MO, Abdulwahab F, et al. 2023. Beyond the exome: utility of long-read whole genome sequencing in exome-negative autosomal recessive diseases. *Genome Medicine* **15**: 114.

Alonso I, Jardim LB, Artigalas O, Saraiva-Pereira ML, Matsuura T, Ashizawa T, Sequeiros J, Silveira I. 2006. Reduced penetrance of intermediate size alleles in spinocerebellar ataxia type 10. *Neurology* **66**: 1602–1604.

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al. 2019. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**: 663-675.e19.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**: 1003–1007.

Beecroft SJ, Cortese A, Sullivan R, Yau WY, Dyer Z, Wu TY, Mulroy E, Pelosi L, Rodrigues M, Taylor R, et al. 2020. A Māori specific RFC1 pathogenic repeat configuration in CANVAS, likely due to a founder allele. *Brain* **143**: 2673–2680.

Bu, Siyuan, Yihan Lv, Yusheng Liu, Sen Qiao, and Hongmei Wang. 2021. “Zinc Finger Proteins in Neuro-Related Diseases Progression.” Frontiers in Neuroscience 15 (November): 760567.

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**: 3426-3440.e19.

Cameron DL, Di Stefano L, Papenfuss AT. 2019. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun* **10**: 3240.

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**: 1784.

Chaisson MJP, Sulovari A, Valdmanis PN, Miller DE, Eichler EE. 2023. Advances in the discovery and analyses of human tandem repeats. *Emerging Topics in Life Sciences* ETLS20230074.

Chen R, Wang X, Dai Z, Wang Z, Wu W, Hu Z, Zhang X, Liu Z, Zhang H, Cheng Q. 2021. TNFSF13 is a novel onco-inflammatory marker and correlates with immune infiltration in gliomas. *Front Immunol* **12**: 713757.

Cohen ASA, Farrow EG, Abdelmoity AT, Alaimo JT, Amudhavalli SM, Anderson JT, Bansal L, Bartik L, Baybayan P, Belden B, et al. 2022. Genomic answers for children: Dynamic analyses of >1000 pediatric rare disease genomes. *Genet Med* **24**: 1336–1348.

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, et al. 2011. A copy number variation morbidity map of developmental delay. *Nat Genet* **43**: 838–846.

Cortese A, Reilly MM, Houlden H. 1993. RFC1 CANVAS / Spectrum Disorder. In *GeneReviews* (eds. M.P. Adam, J. Feldman, G.M. Mirzaa, R.A. Pagon, S.E. Wallace, L.J. Bean, K.W. Gripp, and A. Amemiya), University of Washington, Seattle, Seattle (WA) http://www.ncbi.nlm.nih.gov/books/NBK564656/ (Accessed February 5, 2024).

Cortese A, Simone R, Sullivan R, Vandrovcova J, Tariq H, Yau WY, Humphrey J, Jaunmuktane Z, Sivakumar P, Polke J, et al. 2019. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. *Nat Genet* **51**: 649–658.

Dagli AI, Mathews J, Williams CA. 1993. Angelman Syndrome. In *GeneReviews*® (eds. M.P. Adam, G.M. Mirzaa, R.A. Pagon, S.E. Wallace, L.J. Bean, K.W. Gripp, and A. Amemiya), University of Washington, Seattle, Seattle (WA) http://www.ncbi.nlm.nih.gov/books/NBK1144/ (Accessed October 10, 2023).

Damaraju N, Miller AL, Miller DE. 2024. Long-Read DNA and RNA Sequencing to Streamline Clinical Genetic Testing and Reduce Barriers to Comprehensive Genetic Testing. *The Journal of Applied Laboratory Medicine* **9**: 138–150.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10**: giab008.

Danforth DN. 2016. Genomic Changes in Normal Breast Tissue in Women at Normal Risk or at High Risk for Breast Cancer. *Breast Cancer (Auckl)* **10**: 109–146.

Depienne C, Mandel J-L. 2021. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *The American Journal of Human Genetics* **108**: 764–785.

Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, Emig-Agius D, Gross A, Narzisi G, Bowman B, et al. 2019. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**: 4754–4756.

Dolzhenko E, English A, Dashnow H, De Sena Brandine G, Mokveld T, Rowell WJ, Karniski C, Kronenberg Z, Danzi MC, Cheung WA, et al. 2024. Characterization and visualization of tandem repeats at genome scale. *Nat Biotechnol*. Jan 2.

Driscoll DJ, Miller JL, Cassidy SB. 1993. Prader-Willi Syndrome. In *GeneReviews*® (eds. M.P. Adam, J. Feldman, G.M. Mirzaa, R.A. Pagon, S.E. Wallace, L.J. Bean, K.W. Gripp, and

A. Amemiya), University of Washington, Seattle, Seattle (WA) http://www.ncbi.nlm.nih.gov/books/NBK1330/ (Accessed February 7, 2024).

Duarte-Pereira S, Fajarda O, Matos S, Oliveira JL, Monteiro Silva R. 2021. "NAPRT Expression Regulation Mechanisms: Novel Functions Predicted by a Bioinformatics Approach." Genes 12 (12). https://doi.org/10.3390/genes12122022.

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**: eabf7117.

Eichler EE. 2019. Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *N Engl J Med* **381**: 64–74.

English AC, Menon VK, Gibbs RA, Metcalf GA, Sedlazeck FJ. 2022. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biology* **23**: 271.

Fahim AT, Daiger SP, Weleber RG. 1993. Nonsyndromic Retinitis Pigmentosa Overview. In *GeneReviews*® (eds. M.P. Adam, J. Feldman, G.M. Mirzaa, R.A. Pagon, S.E. Wallace, L.J. Bean, K.W. Gripp, and A. Amemiya), University of Washington, Seattle, Seattle (WA) http://www.ncbi.nlm.nih.gov/books/NBK1417/ (Accessed February 27, 2024).

Fan Y, Zhang S, Yang J, Mao C-Y, Yang Z-H, Hu Z-W, Wang Y-L, Liu Y-T, Liu H, Yuan Y-P, et al. 2020. No biallelic intronic AAGGG repeat expansion in RFC1 was found in patients with late-onset ataxia and MSA. *Parkinsonism Relat Disord* **73**: 1–2.

Farashi S, Harteveld CL. 2018. Molecular basis of α-thalassemia. *Blood Cells, Molecules, and Diseases* **70**: 43–53.

Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Vooren SV, Moreau Y, Pettett RM, Carter NP. 2009. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *The American Journal of Human Genetics* **84**: 524–533.

Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, Hagmann J, Vorbrugg S, Marco-Sola S, Kubica C, et al. 2023. Building pangenome graphs. 2023.04.05.535718. https://www.biorxiv.org/content/10.1101/2023.04.05.535718v1 (Accessed March 3, 2024).

Hannan AJ. 2018. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet* **19**: 286–298.

Harvey WT, Ebert P, Ebler J, Audano PA, Munson KM, Hoekzema K, Porubsky D, Beck CR, Marschall T, Garimella K, et al. 2023. Whole-genome long-read sequencing downsampling and its effect on variant-calling precision and recall. *Genome Res* **33**: 2029–2040.

Heller D, Vingron M. 2019. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**: 2907–2915.

Hiatt L, Weisburd B, Dolzhenko E, VanNoy GE, Kurts EN, Rehm HL, Quinlan A, Dashnow H. 2024. STRchive: a dynamic resource detailing population-level and locus-specific insights at tandem repeat disease loci. 2024.05.21.24307682. https://www.medrxiv.org/content/10.1101/2024.05.21.24307682v1 (Accessed May 15, 2024).

Hiatt SM, Lawlor JMJ, Handley LH, Ramaker RC, Rogers BB, Partridge EC, Boston LB, Williams M, Plott CB, Jenkins J, et al. 2021. Long-read genome sequencing for the molecular diagnosis of neurodevelopmental disorders. *HGG Adv* **2**: 100023.

Hijikata A, Suyama M, Kikugawa S, Matoba R, Naruto T, Enomoto Y, Kurosawa K, Harada N, Yanagi K, Kaname T, et al. 2024. Exome-wide benchmark of difficult-to-sequence regions using short-read next-generation DNA sequencing. *Nucleic Acids Research* **52**: 114–124.

Huisman TH, Wrightstone RN, Wilson JB, Schroeder WA, Kendall AG. 1972. Hemoglobin Kenya, the product of fusion of amd polypeptide chains. *Arch Biochem Biophys* **153**: 850–853.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437**: 1299–1320.

Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y. 2020. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology* **21**: 189.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12**: 996–1006.

Kirsche M, Prabhu G, Sherman R, Ni B, Battle A, Aganezov S, Schatz MC. 2023. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat Methods* **20**: 408–417.

Koenig Z, Yohannes MT, Nkambule LL, Zhao X, Goodrich JK, Kim HA, Wilson MW, Tiao G, Hao SP, Sahakian N, et al. 2024. A harmonized public resource of deeply sequenced diverse human genomes. *Genome Res*. **34**(5): 796–809.

Kolmogorov M, Billingsley KJ, Mastoras M, Meredith M, Monlong J, Lorig-Roach R, Asri M, Alvarez Jerez P, Malik L, Dewan R, et al. 2023. Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nat Methods* **20**: 1483–1492.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**: 540–546.

Krusche P. Haplotype comparison tools / hap.py. http://github.com/illumina/hap.py.

Lee S, Wheeler MM, Patterson K, McGee S, Dalton R, Woodahl EL, Gaedigk A, Thummel KE, Nickerson DA. 2019. Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model. *Genet Med* **21**: 361–372.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, MacArthur D. 2018. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods* **15**: 595–597.

Li W, Lin L, Malhotra R, Yang L, Acharya R, Poss M. 2019. A computational framework to assess genome-wide distribution of polymorphic human endogenous retrovirus-K In human populations. *PLoS Comput Biol* **15**: e1006564.

Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. 2023. A draft human pangenome reference. *Nature* **617**: 312–324.

Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21**: 597–614.

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, et al. 2012. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* **335**: 823–828.

Marco-Sola S, Moure JC, Moreto M, Espinosa A. 2021. Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics* **37**: 456–463.

Matsuura T, Ashizawa T. 1993. Spinocerebellar Ataxia Type 10. In *GeneReviews*® (eds. M.P. Adam, J. Feldman, G.M. Mirzaa, R.A. Pagon, S.E. Wallace, L.J. Bean, K.W. Gripp, and A. Amemiya), University of Washington, Seattle, Seattle (WA) http://www.ncbi.nlm.nih.gov/books/NBK1175/ (Accessed February 5, 2024).

Matsuura T, Fang P, Pearson CE, Jayakar P, Ashizawa T, Roa BB, Nelson DL. 2006. Interruptions in the Expanded ATTCT Repeat of Spinocerebellar Ataxia Type 10: Repeat Purity as a Disease Modifier? *Am J Hum Genet* **78**: 125–129.

Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, Munson KM, Lewis AP, Fuerte EPA, Paschal CR, Walsh T, et al. 2021. Targeted long-read sequencing identifies missing disease-causing variation. *The American Journal of Human Genetics* **108**: 1436–1449.

Morato Torres CA, Zafar F, Tsai Y-C, Vazquez JP, Gallagher MD, McLaughlin I, Hong K, Lai J, Lee J, Chirino-Perez A, et al. 2022. ATTCT and ATTCC repeat expansions in the ATXN10 gene affect disease penetrance of spinocerebellar ataxia type 10. *Human Genetics and Genomics Advances* **3**: 100137.

Morris CA. 1993. Williams Syndrome. In *GeneReviews*® (eds. M.P. Adam, J. Feldman, G.M. Mirzaa, R.A. Pagon, S.E. Wallace, L.J. Bean, K.W. Gripp, and A. Amemiya), University of Washington, Seattle, Seattle (WA) http://www.ncbi.nlm.nih.gov/books/NBK1249/ (Accessed February 24, 2024).

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376**: 44–53.

Ortiz-Aljaro P, Montes-Cano MA, García-Lozano J-R, Aquino V, Carmona R, Perez-Florido J, García-Hernández FJ, Dopazo J, González-Escribano MF. 2022. Protein and functional isoform levels and genetic variants of the BAFF and APRIL pathway components in systemic lupus erythematosus. *Sci Rep* **12**: 11219.

Pedersen BS, Bhetariya PJ, Brown J, Kravitz SN, Marth G, Jensen RL, Bronner MP, Underhill HR, Quinlan AR. 2020. Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. *Genome Medicine* **12**: 62.

Pellerin D, Danzi MC, Wilke C, Renaud M, Fazal S, Dicaire M-J, Scriba CK, Ashton C, Yanick C, Beijer D, et al. 2023. Deep Intronic FGF14 GAA Repeat Expansion in Late-Onset Cerebellar Ataxia. *N Engl J Med* **388**: 128–141.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Raskin S, Ashizawa T, Teive HAG, Arruda WO, Fang P, Gao R, White MC, Werneck LC, Roa B. 2007. Reduced penetrance in a Brazilian family with spinocerebellar ataxia type 10. *Arch Neurol* **64**: 591–594.

Reis ALM, Rapadas M, Hammond JM, Gamaarachchi H, Stevanovski I, Ayuputeri Kumaheri M, Chintalaphani SR, Dissanayake DSB, Siggs OM, Hewitt AW, et al. 2023. The landscape of genomic structural variation in Indigenous Australians. *Nature* **624**: 1–9.

Ren J, Gu B, Chaisson MJP. 2023. vamos: variable-number tandem repeats annotation using efficient motif sets. *Genome Biology* **24**: 175.

Saal HM, Harbison MD, Netchine I. 1993. Silver-Russell Syndrome. In *GeneReviews*® (eds. M.P. Adam, J. Feldman, G.M. Mirzaa, R.A. Pagon, S.E. Wallace, L.J. Bean, K.W. Gripp, and A. Amemiya), University of Washington, Seattle, Seattle (WA) http://www.ncbi.nlm.nih.gov/books/NBK1324/ (Accessed February 7, 2024).

Schloissnig S, Pani S, Rodriguez-Martin B, Ebler J, Hain C, Tsapalou V, Söylev A, Hüther P, Ashraf H, Prodanov T, et al. 2024.04.18.590093. https://www.biorxiv.org/content/10.1101/2024.04.18.590093v1 (Accessed May 1, 2024).

Scriba CK, Beecroft SJ, Clayton JS, Cortese A, Sullivan R, Yau WY, Dominik N, Rodrigues M, Walker E, Dyer Z, et al. 2020. A novel RFC1 repeat motif (ACAGG) in two Asia-Pacific CANVAS families. *Brain* **143**: 2904–2910.

Shafin K, Pesout T, Chang P-C, Nattestad M, Kolesnikov A, Goel S, Baid G, Kolmogorov M, Eizenga JM, Miga KH, et al. 2021. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat Methods* **18**: 1322–1332.

Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, et al. 2020. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol* **38**: 1044–1053.

Shuman C, Kalish JM, Weksberg R. 1993. Beckwith-Wiedemann Syndrome. In *GeneReviews*® (eds. M.P. Adam, J. Feldman, G.M. Mirzaa, R.A. Pagon, S.E. Wallace, L.J. Bean, K.W.

Gripp, and A. Amemiya), University of Washington, Seattle, Seattle (WA) http://www.ncbi.nlm.nih.gov/books/NBK1394/ (Accessed February 7, 2024).

Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, Kalef-Ezra E, Gandhi M, Hong K, Pehlivan D, et al. 2024. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol* 1–10.

Subramanian RP, Wildschutte JH, Russo C, Coffin JM. 2011. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology* **8**: 90.

Sulovari A, Li R, Audano PA, Porubsky D, Vollger MR, Logsdon GA, Human Genome Structural Variation Consortium, Warren WC, Pollen AA, Chaisson MJP, et al. 2019. Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc Natl Acad Sci U S A* **116**: 23243–23253.

Tanudisastro HA, Deveson IW, Dashnow H, MacArthur DG. 2024. Sequencing and characterizing short tandem repeats in the human genome. *Nat Rev Genet* 1–16.

Taylor DJ, Chhetri SB, Tassia MG, Biddanda A, Battle A, McCoy RC. 2023. Sources of gene expression variation in a globally diverse human cohort. 2023.11.04.565639. https://www.biorxiv.org/content/10.1101/2023.11.04.565639v4 (Accessed February 22, 2024).

Tretina K, Park ES, Maminska A, MacMicking JD. 2019. “Interferon-Induced Guanylate-Binding Proteins: Guardians of Host Defense in Health and Disease.” The Journal of Experimental Medicine 216 (3): 482–500.

Twesigomwe D, Drögemöller BI, Wright GEB, Adebamowo C, Agongo G, Boua PR, Matshaba M, Paximadis M, Ramsay M, Simo G, et al. 2024. Characterization of CYP2B6 and CYP2A6 Pharmacogenetic Variation in Sub-Saharan African Populations. *Clin Pharmacol Ther* **115**: 576–594.

Twesigomwe D, Drögemöller BI, Wright GEB, Adebamowo C, Agongo G, Boua PR, Matshaba M, Paximadis M, Ramsay M, Simo G, et al. 2023. Characterization of CYP2D6 Pharmacogenetic Variation in Sub-Saharan African Populations. *Clinical Pharmacology & Therapeutics* **113**: 643–659.

Wagner J, Olson ND, Harris L, Khan Z, Farek J, Mahmoud M, Stankovic A, Kovacevic V, Yoo B, Miller N, et al. 2022. Benchmarking challenging small variants with linked and long reads. *Cell Genomics* **2**: 100128.

Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, Popejoy AB, Asri M, Carson C, Chaisson MJP, et al. 2022. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**: 437–446.

Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. 2021. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**: 1348–1365.

Wilderman A, D’haene E, Baetens M, Yankee TN, Winchester EW, Glidden N, Roets E, Van Dorpe J, Janssens S, Miller DE, et al. 2024. A distant global control region is essential

for normal expression of anterior HOXA genes during mouse and human craniofacial development. *Nat Commun* **15**: 1–23.

Wojcik MH, Reuter CM, Marwaha S, Mahmoud M, Duyzend MH, Barseghyan H, Yuan B, Boone PM, Groopman EE, Délot EC, et al. 2023. Beyond the exome: What's next in diagnostic testing for Mendelian conditions. *The American Journal of Human Genetics* **110**: 1229–1248.

Yang L, Metzger GA, Padilla Del Valle R, Delgadillo Rubalcaba D, McLaughlin RN. 2024. Evolutionary insights from profiling LINE-1 activity at allelic resolution in a single human genome. *EMBO J* **43**: 112–131.

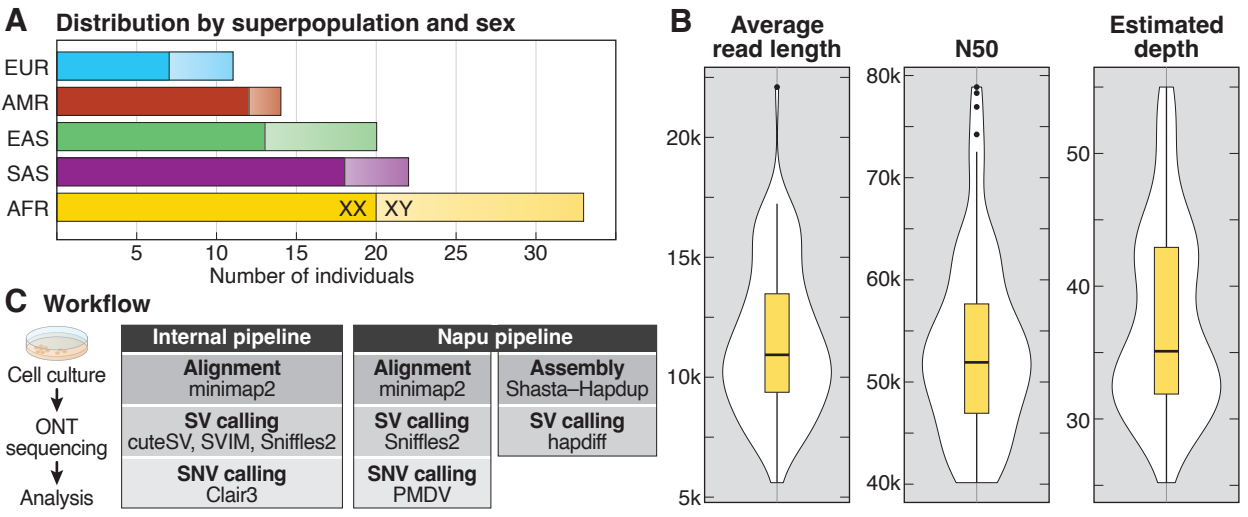
Zalusky MP, Miller DE. 2024. Methylation Operation Wizard (MeOW): Identification of differentially methylated regions in long-read sequencing data. http://arxiv.org/abs/2402.17182 (Accessed February 27, 2024).

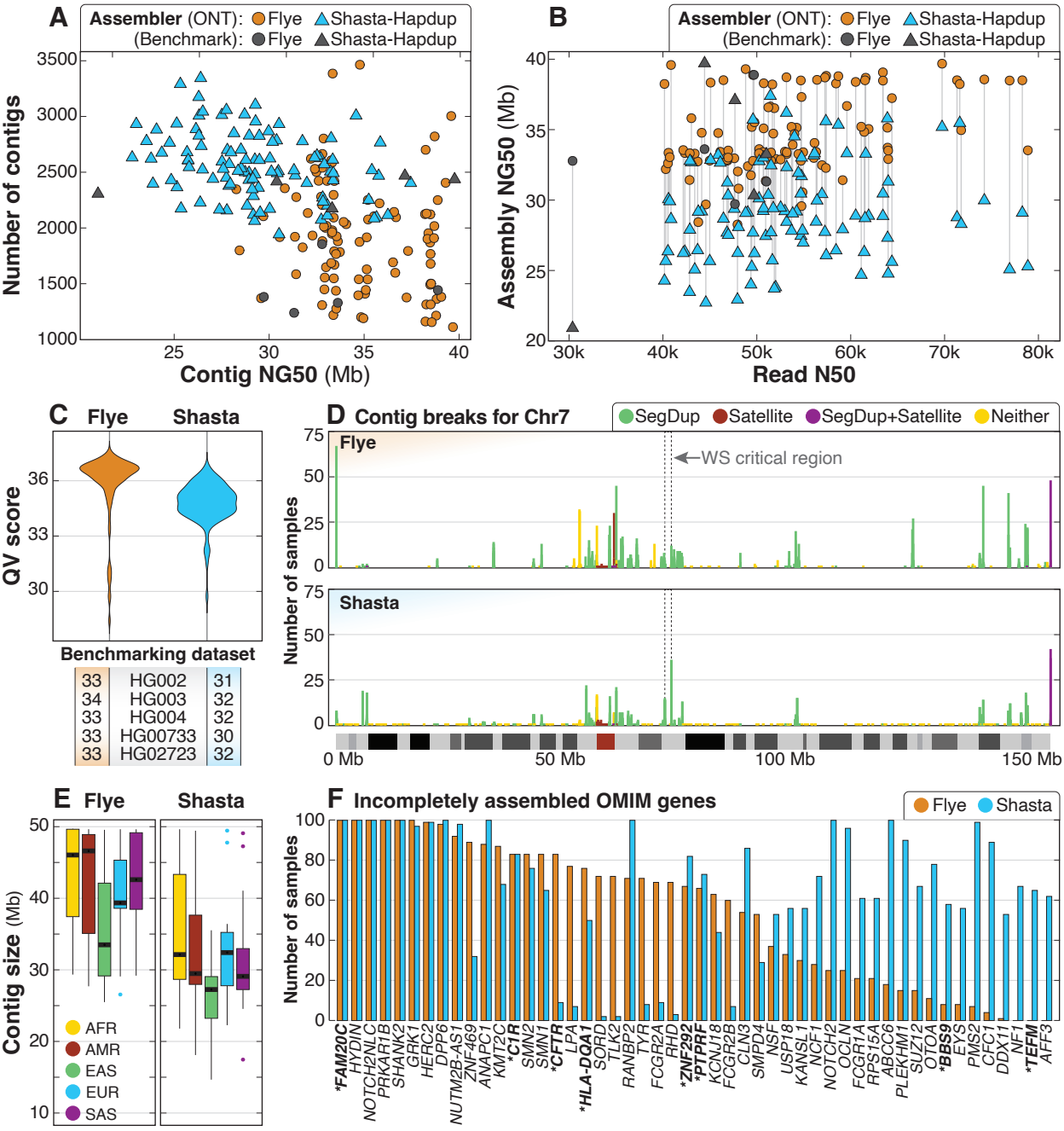
Zanger UM, Schwab M. 2013. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol Ther* **138**: 103–141.

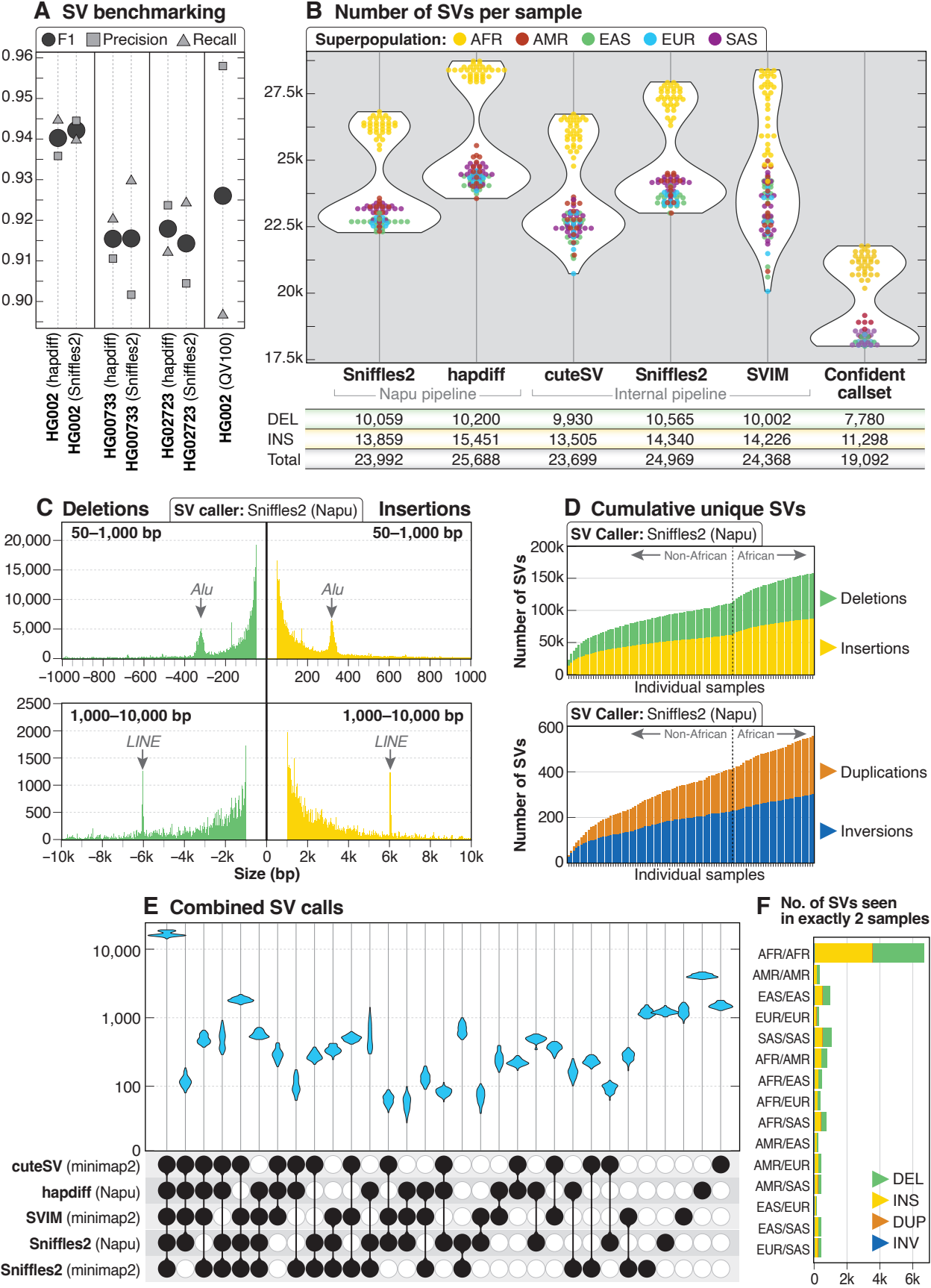
Zhao X, Collins RL, Lee W-P, Weber AM, Jun Y, Zhu Q, Weisburd B, Huang Y, Audano PA, Wang H, et al. 2021. Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am J Hum Genet* **108**: 919–928.

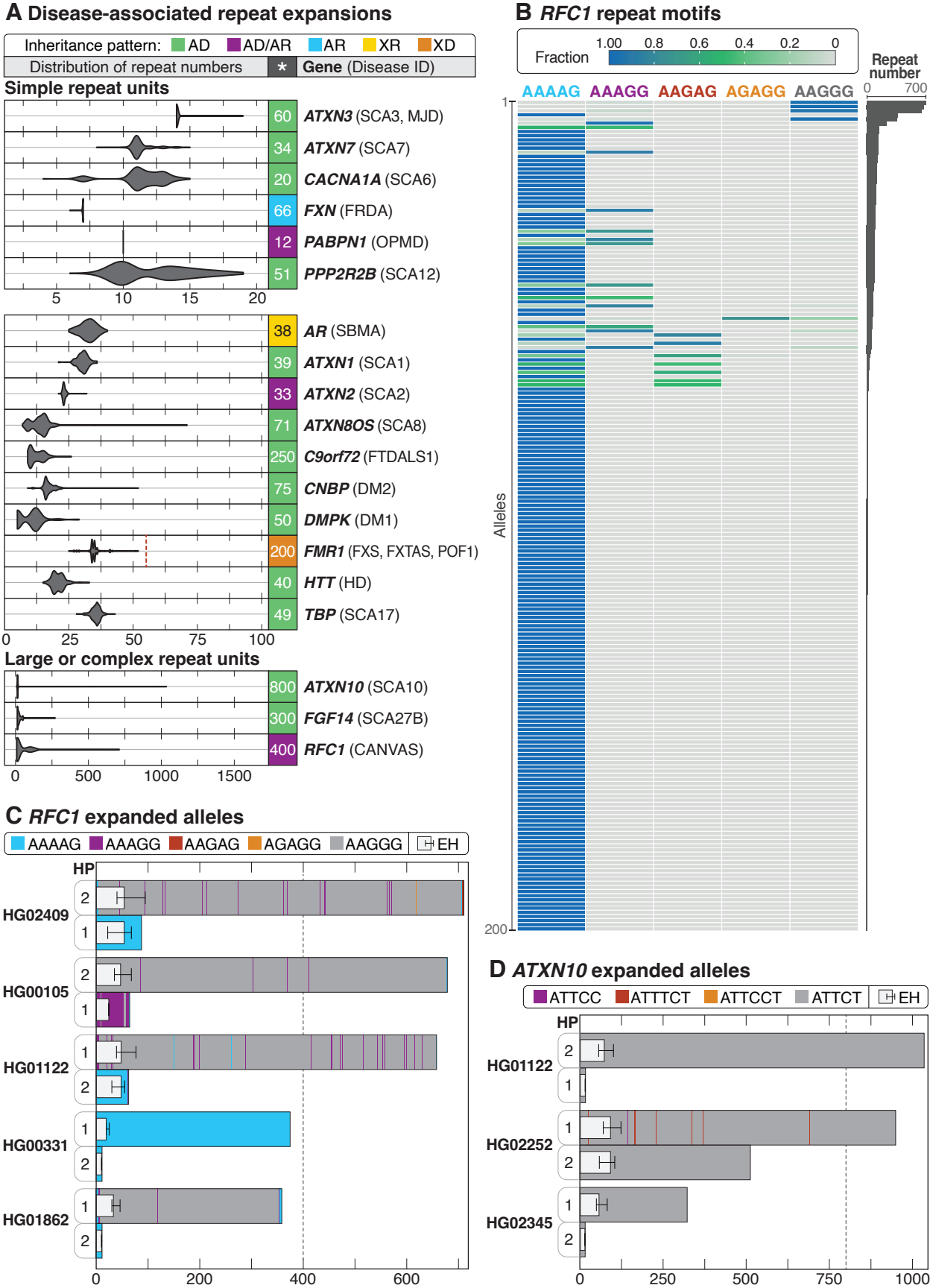
Zheng Z, Li S, Su J, Leung AW-S, Lam T-W, Luo R. 2022. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat Comput Sci* **2**: 797–803.

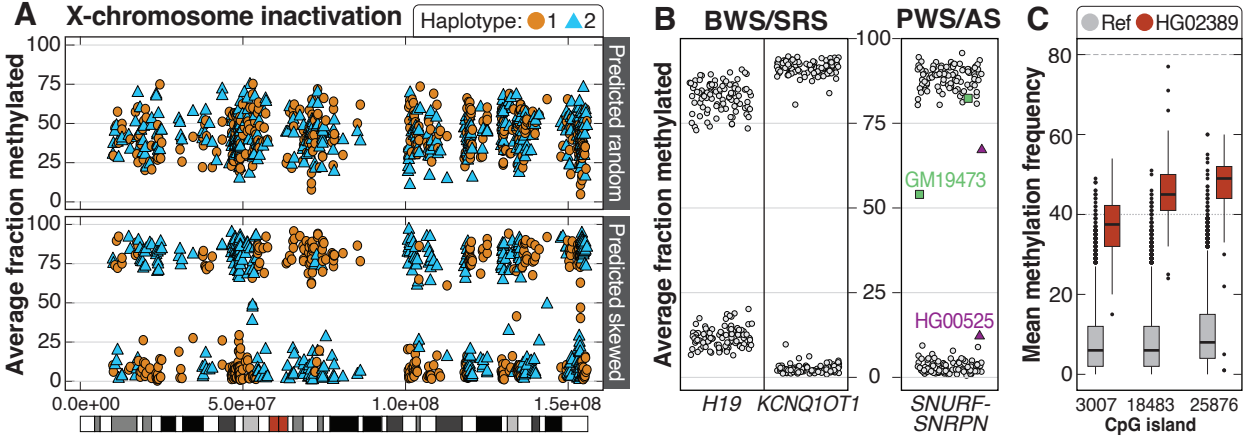
Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, Sherry S, Koren S, Phillippy AM, Boutros PC, et al. 2020. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* **38**: 1347–1355.


A
Distribution by superpopulation and sex
EUR
AMR
EAS
SAS
AFR
XX
XY
5
10
15
20
25
30
Number of individuals
B
Average read length
N50
Estimated depth
20k
15k
10k
5k
80k
70k
60k
50k
40k
50
40
30
C
Workflow
Cell culture
ONT sequencing
Analysis
Internal pipeline
Alignment
minimap2
SV calling
cuteSV, SVIM, Sniffles2
SNV calling
Clair3
Napu pipeline
Alignment
minimap2
SV calling
Sniffles2
SNV calling
PMDV
Assembly
Shasta–Hapdup
SV calling
hapdiff


A
Assembler (ONT): Flye Shasta-Hapdup
(Benchmark): Flye Shasta-Hapdup
Number of contigs
Contig NG50 (Mb)
B
Assembler (ONT): Flye Shasta-Hapdup
(Benchmark): Flye Shasta-Hapdup
Assembly NG50 (Mb)
Read N50
C
Flye
Shasta
QV score
Benchmarking dataset
33 HG002 31
34 HG003 32
33 HG004 32
33 HG00733 30
33 HG02723 32
D
Contig breaks for Chr7
SegDup Satellite SegDup+Satellite Neither
Flye
WS critical region
Shasta
Number of samples
0 Mb 50 Mb 100 Mb 150 Mb
E
Flye
Shasta
Contig size (Mb)
AFR
AMR
EAS
EUR
SAS
F
Incompletely assembled OMIM genes
Flye Shasta
Number of samples
*FAM20C HYDIN NOTCH2NLC PRKAR1B SHANK2 GRK1 HERC2 DPP6 NUTM2B-AS1 ZNF469 ANAPC1 KMT2C *C1R SMN2 SMN1 *CFTR LPA *HLA-DQA1 SORD TLK2 RANBP2 TYR FCGR2A RHD *ZNF292 *PTPRF KCNJ18 FCGR2B CLN3 SMPD4 NSF USP18 KANSL1 NCF1 NOTCH2 OCLN FCGR1A RPS15A ABCC6 PLEKHM1 SUZ12 OTOA *BBS9 EYS PMS2 CFC1 DDX11 NF1 *TEFM AFF3


A SV benchmarking
F1
Precision
Recall
0.96
0.95
0.94
0.93
0.92
0.91
0.90
HG002 (hapdiff)
HG002 (Sniffles2)
HG00733 (hapdiff)
HG00733 (Sniffles2)
HG02723 (hapdiff)
HG02723 (Sniffles2)
HG002 (QV100)
B Number of SVs per sample
Superpopulation: AFR AMR EAS EUR SAS
27.5k
25k
22.5k
20k
17.5k
Sniffles2
hapdiff
cuteSV
Sniffles2
SVIM
Confident callset
Napu pipeline
Internal pipeline
DEL 10,059 10,200 9,930 10,565 10,002 7,780
INS 13,859 15,451 13,505 14,340 14,226 11,298
Total 23,992 25,688 23,699 24,969 24,368 19,092
C Deletions
SV caller: Sniffles2 (Napu)
Insertions
50–1,000 bp
Alu
20,000
15,000
10,000
5,000
0
−1000 −800 −600 −400 −200
200 400 600 800 1000
1,000–10,000 bp
LINE
2500
2000
1500
1000
500
−10k −8k −6k −4k −2k 0 2k 4k 6k 8k 10k
Size (bp)
D Cumulative unique SVs
SV Caller: Sniffles2 (Napu)
Non-African
African
Number of SVs
200k
150k
100k
50k
Individual samples
Deletions
Insertions
Duplications
Inversions
600
400
200
E Combined SV calls
10,000
1,000
100
cuteSV (minimap2)
hapdiff (Napu)
SVIM (minimap2)
Sniffles2 (Napu)
Sniffles2 (minimap2)
F No. of SVs seen in exactly 2 samples
AFR/AFR
AMR/AMR
EAS/EAS
EUR/EUR
SAS/SAS
AFR/AMR
AFR/EAS
AFR/EUR
AFR/SAS
AMR/EAS
AMR/EUR
AMR/SAS
EAS/EUR
EAS/SAS
EUR/SAS
DEL
INS
DUP
INV
0
2k
4k
6k


A Disease-associated repeat expansions
Inheritance pattern: AD AD/AR AR XR XD
Distribution of repeat numbers
*
Gene (Disease ID)
Simple repeat units
60 ATXN3 (SCA3, MJD)
34 ATXN7 (SCA7)
20 CACNA1A (SCA6)
66 FXN (FRDA)
12 PABPN1 (OPMD)
51 PPP2R2B (SCA12)
5 10 15 20
38 AR (SBMA)
39 ATXN1 (SCA1)
33 ATXN2 (SCA2)
71 ATXN8OS (SCA8)
250 C9orf72 (FTDALS1)
75 CNBP (DM2)
50 DMPK (DM1)
200 FMR1 (FXS, FXTAS, POF1)
40 HTT (HD)
49 TBP (SCA17)
0 25 50 75 100
Large or complex repeat units
800 ATXN10 (SCA10)
300 FGF14 (SCA27B)
400 RFC1 (CANVAS)
0 500 1000 1500
B RFC1 repeat motifs
Fraction 1.00 0.8 0.6 0.4 0.2 0
Repeat number
0 700
AAAAG AAAGG AAGAG AGAGG AAGGG
1
200
Alleles
C RFC1 expanded alleles
AAAAG AAAGG AAGAG AGAGG AAGGG EH
HP
HG02409
HG00105
HG01122
HG00331
HG01862
0 200 400 600
D ATXN10 expanded alleles
ATTCC ATTTCT ATTCCT ATTCT EH
HP
HG01122
HG02252
HG02345
0 250 500 750 1000


A
X-chromosome inactivation
Haplotype: 1 2
Predicted random
Predicted skewed
Average fraction methylated
100
75
50
25
0
0.0e+00
5.0e+07
1.0e+08
1.5e+08
B
BWS/SRS
PWS/AS
Average fraction methylated
GM19473
HG00525
H19
KCNQ1OT1
SNURF-
SNRPN
C
Ref
HG02389
Mean methylation frequency
80
60
40
20
0
3007
18483
25876
CpG island

High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation

Jonas A Gustafson, Sophia B Gibson, Nikhita Damaraju, et al.

Genome Res. published online October 2, 2024
Access the most recent version at doi:10.1101/gr.279273.124