ARTICLES
https://doi.org/10.1038/s41587-021-01158-1
nature biotechnology

Curated variation benchmarks for challenging medically relevant autosomal genes

Justin Wagner[1], Nathan D. Olson[1], Lindsay Harris[1], Jennifer McDaniel[1], Haoyu Cheng[2], Arkarachai Fungtammasan[3], Yih-Chii Hwang[3], Richa Gupta[3], Aaron M. Wenger[4], William J. Rowell[4], Ziad M. Khan[5], Jesse Farek[5], Yiming Zhu[5], Aishwarya Pisupati[5], Medhat Mahmoud[5], Chunlin Xiao[6], Byunggil Yoo[7], Sayed Mohammad Ebrahim Sahraeian[8], Danny E. Miller[9,10], David Jáspez[11], José M. Lorenzo-Salazar[11], Adrián Muñoz-Barrera[11], Luis A. Rubio-Rodríguez[11], Carlos Flores[11,12,13], Giuseppe Narzisi[14], Uday Shanker Evani[14], Wayne E. Clarke[14], Joyce Lee[15], Christopher E. Mason[16], Stephen E. Lincoln[17], Karen H. Miga[18], Mark T. W. Ebbert[19,20,21], Alaina Shumate[22,23], Heng Li[2], Chen-Shan Chin[3,24] ✉, Justin M. Zook[1,24] ✉ and Fritz J. Sedlazeck[5,24] ✉

The repetitive nature and complexity of some medically relevant genes poses a challenge for their accurate analysis in a clinical setting. The Genome in a Bottle Consortium has provided variant benchmark sets, but these exclude nearly 400 medically relevant genes due to their repetitiveness or polymorphic complexity. Here, we characterize 273 of these 395 challenging autosomal genes using a haplotype-resolved whole-genome assembly. This curated benchmark reports over 17,000 single-nucleotide variations, 3,600 insertions and deletions and 200 structural variations each for human genome reference GRCh37 and GRCh38 across HG002. We show that false duplications in either GRCh37 or GRCh38 result in reference-specific, missed variants for short- and long-read technologies in medically relevant genes, including *CBS*, *CRYAA* and *KCNE1*. When masking these false duplications, variant recall can improve from 8% to 100%. Forming benchmarks from a haplotype-resolved whole-genome assembly may become a prototype for future benchmarks covering the whole genome.

Authoritative benchmark samples are driving the development of technologies and the discovery of new variants, enabling highly accurate clinical genome sequencing and advancing our detection and understanding of the impact of many genomic variations on human disease at scale. With recent improvements in sequencing technologies[1], assembly algorithms[2–4] and variant-calling methods[5], genomics offers more insights into challenging genes associated with human diseases across a higher number of patients[6]. Still, challenges remain for medically relevant genes that are often repetitive or highly polymorphic[7,8]. In fact, a recent study found that 13.8% (17,561) of pathogenic variants identified by a high-throughput clinical laboratory were challenging to detect with short-read sequencing[9]. These included challenging variants such as variants 15–49 bp in size, small copy-number variations (CNVs), complex variants and variants in low-complexity or segmentally duplicated regions.

The Genome in a Bottle (GIAB) consortium develops benchmarks to advance accurate human genomic research and clinical applications of sequencing. GIAB provides highly curated benchmark sets for single-nucleotide variant (SNV)[10], small insertion and deletion (INDEL)[10] and structural variant (SV) calling[11]. Here, we define SNVs as single base substitutions, while INDELs are defined as insertions and deletions smaller than 50 bp, in contrast to insertions and deletions larger than 50 bp, which we refer to as SVs. Furthermore, GIAB and the Food and Drug Administration (FDA) host periodic precisionFDA challenges providing a snapshot and recommendations for small variant calling enabling the high precision and sensitivity required for clinical research, with

[1]Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA. [2]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA. [3]DNAnexus, Inc., Mountain View, CA, USA. [4]Pacific Biosciences, Menlo Park, CA, USA. [5]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [6]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. [7]Genomic Medicine Center, Children's Mercy Kansas City, Kansas City, MO, USA. [8]Roche Sequencing Solutions, Santa Clara, CA, USA. [9]Department of Pediatrics, Division of Genetic Medicine, University of Washington and Seattle Children's Hospital, Seattle, WA, USA. [10]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [11]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain. [12]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain. [13]Research Unit, Hospital Universitario N.S. de Candelaria, Santa Cruz de Tenerife, Spain. [14]New York Genome Center, New York, NY, USA. [15]Bionano Genomics, San Diego, CA, USA. [16]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. [17]Invitae, San Francisco, CA, USA. [18]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA. [19]Sanders-Brown Center on Aging, University of Kentucky, Lexington, KY, USA. [20]Department of Internal Medicine, Division of Biomedical Informatics, University of Kentucky, Lexington, KY, USA. [21]Department of Neuroscience, University of Kentucky, Lexington, KY, USA. [22]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [23]Center for Computational Biology, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA. [24]These authors contributed equally: Chen-Shan Chin, Justin M. Zook, Fritz J. Sedlazeck. ✉e-mail: jchin@dnanexus.com; jzook@nist.gov; Fritz.Sedlazeck@bcm.edu

a recent challenge demonstrating the importance of including more difficult genomic regions[12]. Recently, GIAB focused primarily on a read mapping-based genome-wide approach integrating short-, linked- and long-read sequencing to characterize up to 92% and 86% of the autosomal bases for small variants and SVs, respectively[11,13]. GIAB also released a targeted assembly-based benchmark for the major histocompatibility complex (MHC) region, a highly diverse and repetitive region of the human genome that includes the human leukocyte antigen (HLA) genes[14]. Still, multiple regions of the genome are not fully resolved in existing benchmarks due to repetitive sequence, segmental duplications and complex variants (i.e., multiple nearby SNVs, INDELs and/or SVs)[15].

Many clinically relevant genes are in the remaining hard-to-assess regions. The clinical tests for these genes often require locus-specific targeted designs and/or employ multiple technologies and are only applied when suspicion of a specific disorder is high. Mandelker et al. categorized genes based on their repetitive content and identified 193 genes that cannot be fully characterized by short-read sequencing[7]. This gene set was constructed by identifying genes with low mapping quality in the clinical databases OMIM, HGMD and ClinVar. Subsequently, Wenger et al. showed that while short reads could not accurately map the full length of these genes, highly accurate long reads could fully map 152 (78.76%) of them[1]. The latest v4.2.1 GIAB small variant benchmark regions included at least 90% of the gene body for 110 of the 159 difficult genes on autosomes[13]. In contrast, the previous v3.3.2 GIAB small variant benchmark regions included at least 90% of the gene body for only 19 of 159 difficult genes[10]. Although v4.2.1 includes substantially more difficult genes, variant calls in the remaining most difficult genes still need to be assessed, and challenges remain with typical mapping-based approaches in some genes, even when using highly accurate long reads.

To support ongoing advancements in clinical genome sequencing and bioinformatics, we present a more comprehensive benchmark of challenging, medically relevant genes (CMRGs) focusing on HG002, which has a broad consent from the Personal Genome Project for open genomic data and commercial redistribution[16] (Fig. 1). With the advent of highly accurate long reads, new approaches for haplotype-resolved (diploid) assembly have advanced rapidly[2,3]. Here, we focus on generating a benchmark for as many of these genes as possible using a whole-genome haplotype-resolved assembly. We curated a set of 273 medically relevant genes with ≤90% of bases included in previous GIAB benchmarks but fully covered by both haplotypes of a trio-based hifiasm assembly. The assembly included all phased small variants and SVs across these genes. Then, we delineated regions where we can provide reliable small variant and SV benchmarks, developing a prototype process for future whole-genome assembly-based benchmarks.

Results

Identification of CMRGs. To prioritize genome regions for the expanded benchmark, we identified several lists of potentially medically relevant genes. The first was a list of 4,773 potentially medically relevant genes from the databases OMIM, HGMD and ClinVar previously compiled in 2012, which includes both commonly tested and rarely tested genes (Supplementary Table 13 in ref. [7]). The second was a list from the COSMIC gene census, which contains 723 gene symbols found in tumors (https://cancer.sanger.ac.uk/census)[17]. We also developed a focused list of high-priority clinical genes that are commonly tested for clinical inherited diseases (Supplementary Data 1). There are 5,175 gene symbols in the union of these sets, of which 5,027 have unique coordinates on the primary assembly of GRCh38 and valid ENSEMBL annotations, and 4,697 are autosomal; 70% of these genes are specific to the list from OMIM, HGMD and ClinVar, which includes genes associated with disease in a small number of studies and thus are currently tested more frequently in research studies than in high-throughput clinical laboratories (Fig. 1a). Supplementary Note 1 and Supplementary Data 2 show our analysis of the fraction of these 4,697 autosomal medically relevant genes included in our previous benchmark (v4.2.1), resulting in 395 genes that were included ≤90% in GRCh37 or GRCh38 and are the focus of this manuscript.

Assembly enables phased small variant and SV benchmarks. Many of the 395 medically relevant genes were not covered well by the v4.2.1 small variant benchmark due to SVs, complex variants and segmental duplications (Fig. 2). Thus, here we resolve many of these regions using a haplotype-resolved assembly of HG002 constructed by hifiasm[2]. Hifiasm can resolve both haplotypes with high base-level quality (quality value > 50), including many segmental duplications, and produces variant calls and genotypes that are highly concordant with the v4.2.1 small variant benchmark, with both recall and precision, for >99.7% of SNVs and >97% of INDELs in regions covered by the assembly (Supplementary Data 3).

We generated a benchmark (Methods) for 273 of the 395 genes that were fully resolved by this assembly. To be included in the CMRG benchmark, the entire gene, including the 20-kb flanking sequence (the longest reads used for the assembly) on each side and any overlapping segmental duplications, needed to have exactly one fully aligned contig from each haplotype with no breaks on GRCh37 and GRCh38 (Supplementary Data 2). We required the alignments to completely resolve any overlapping segmental duplications to minimize ambiguity or errors in the assembly–assembly alignment. These 273 genes are substantially more challenging than genes previously covered by GIAB's v4.2.1 benchmark; for example, for 99% of the new genes, at least 15% of the gene region is either challenging to sequence or contains challenging variants in HG002 (Fig. 2b). Here, we use the definition of challenging sequences from GIAB and the Global Alliance for Genomics and Health v2.0 stratifications[12,18]. Furthermore, when comparing variants in regions of the CMRG bodies included by the v4.2.1 benchmark, the CMRG benchmark or both benchmarks, 11% of the CMRG benchmark INDELs are >15 bp (Fig. 2c) compared with 3.5% in v4.2.1. The CMRG INDELs >15 bp are also substantially more challenging than the v4.2.1 INDELs >15 bp. This is shown by a decrease in recall of HiFi DeepVariant from 99.5% (v4.2.1) to 84.9% (CMRG). In addition, the precision has decreased for HiFi DeepVariant from 99.9% (v4.2.1) to 94.2% (CMRG) (Supplementary Data 4).

We created separate CMRG benchmark bed files for small variants and SVs, which both rely on the same benchmark variant calls from hifiasm. The CMRG benchmark extends beyond the v4.2.1 benchmark across the 273 challenging gene regions, adding many phased SNVs, INDELs and large insertions and deletions at least 50 bp in length overlapping these genes (Table 1).

Resolving CMRGs. Beyond previous GIAB benchmarks, this CMRG benchmark includes 273 more challenging genes. These include (1) genes that are duplicated in the reference but not in HG002, as described above; (2) highly homologous genes such as *SMN1* and *SMN2* or *NCF1*, *NCF1B* and *NCF1C*; and (3) genes with SVs and complex variants like *RHCE*.

The gene *SMN1* resides within a large segmental duplication on chromosome 5 containing both *SMN1* and *SMN2*. Biallelic pathogenic variants in *SMN1* result in spinal muscular atrophy (SMA), a progressive disorder characterized by muscle weakness and atrophy due to loss of neuronal cells in the spinal cord[19]. While the 28-kb sequences of *SMN1* and *SMN2* generally differ by only five intronic and three exonic nucleotides[20], the identification and characterization of pathogenic variants in *SMN1* and the copy-number state of *SMN2* are relevant for guiding newly developed therapies and counseling families regarding recurrence risk of this disease.

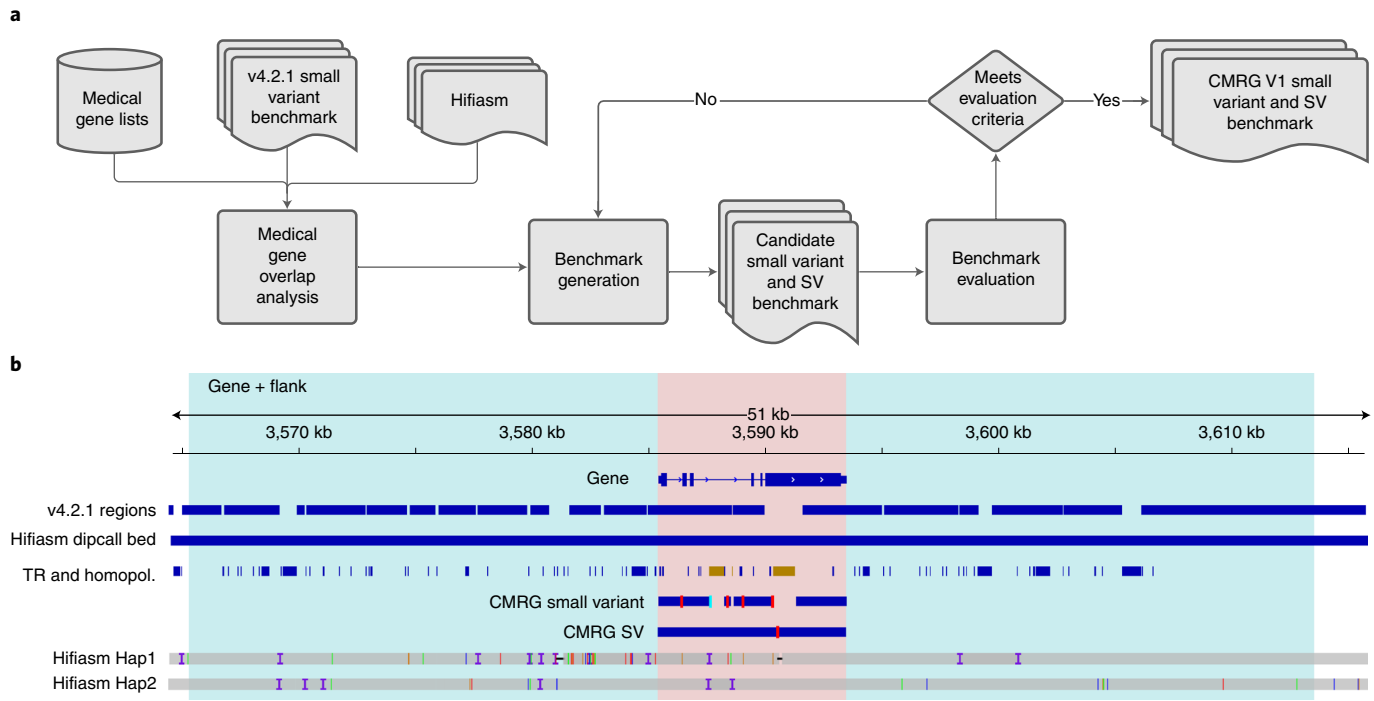

Fig. 1 | GIAB developed a process to create new phased small variant and SV benchmarks for 273 CMRGs. a, We developed a list of 4,701 autosomal potentially medically relevant genes. We generated a new benchmark for 273 of the 4,701 genes that were completely resolved by our hifiasm haplotype-resolved diploid assembly and ≤90% included in the v4.2.1 GIAB small variant benchmark for HG002 (v4.2.1 regions). **b**, We required that the entire gene region (pink) and the 20-kb flanking sequence on each side (blue) be completely resolved by both haplotypes in the assembly (Hifiasm Hap1 and Hifiasm Hap2), indicated with the Hifiasm dipcall bed track. In addition, we required that any segmental duplications overlapping the gene be completely resolved. From the small variant benchmark regions (CMRG SV, blue bars), we excluded SVs and any tandem repeats or homopolymers overlapping SVs (right: tandem repeat and homopolymer (TR and homopol.) region in brown). The left tandem repeat and homopolymer region (in brown) is excluded from the small variant benchmark regions because the larger tandem repeat contains an imperfect homopolymer longer than 20 bp, which we exclude because long homopolymers have a higher error rate in the assembly. All regions of this gene were included in the SV benchmark regions (CMRG SV, blue bar). The vertical red lines in CMRG small variant and CMRG SV indicate locations of benchmark small variants and SVs, respectively. Finally, we evaluated the small variant and SV benchmarks with manual curation and long-range PCR and also ensured they accurately identify false positives and false negatives after excluding errors found during curation.

Some individuals have copy-number polymorphisms of these genes, but HG002 appears to contain one copy each of *SMN1* and *SMN2* on each haplotype based on the presence of two haplotypes for each gene in Oxford Nanopore Technologies (ONT) and 10x Genomics data. However, the genes are surrounded by complex repeats and are thus not fully resolved by our assembly (Fig. 3b). The maternal assembly has a single contig passing through the SMA region but misses *SMN2* and some of the surrounding repeats (dot plot in Supplementary Fig. 1). The paternal assembly contains both *SMN1* and *SMN2*, but the assembly is broken into three contigs in the SMA region (dot plot in Supplementary Fig. 1). Upon curation of the data from Pacific Biosciences (PacBio) HiFi, ultralong ONT and 10x Genomics in Fig. 3a, the variants called from the assembly of *SMN1* were supported by ONT and 10x Genomics across the full gene and by PacBio HiFi across the part of the gene covered by reads. Because we manually confirmed the assembly accuracy in this gene, we included *SMN1* in our benchmark even though the assemblies did not cover the segmental duplications within the entire SMA region. We excluded *SMN2* because only one haplotype was resolved by hifiasm v0.11. Another challenging example is *NCF1*, which is associated with 20% of cases of chronic granulomatous disease, a primary immunodeficiency[21,22]. The gene lies within a large segmental duplication, which may make molecular diagnosis of some cases of chronic granulomatous disease challenging. Our benchmark covers the first two exons that were missing from the v4.2.1 benchmark (Supplementary Fig. 2).

Our benchmark provides a way to measure the accuracy of variant calls in gene conversion-like events and SVs. For example, there is a 4.5-kb gene conversion-like event between RHCE (Supplementary Fig. 3) and RHD (Supplementary Fig. 4; ref. [22]) and a similar event between *SIGLEC16* and *SIGLEC11* (ref. [23]). The benchmark includes other substantially more challenging SVs as well, including a 16,946-bp insertion in a variable number tandem repeat[24] in an intron of the gene *GPI* (Supplementary Fig. 5a) and two insertions in the segmentally duplicated gene *GTF2IRD2* (Supplementary Fig. 5b; more in Supplementary Note 2).

Resolving false gene duplications in the reference. The CMRG benchmark identified variant-calling errors due to false duplications in GRCh37 or GRCh38 in several medically relevant genes. Previous work described true highly homologous genes inside segmental duplications in GRCh37 and GRCh38 that give rise to read mapping issues[7,8]; our CMRG benchmark, however, shows that several of these highly homologous genes are in fact false duplications in the reference. For example, PacBio HiFi and Illumina short-read coverage is low and missing one or both haplotypes for *CBS, CRYAA* and *KCNE1* on GRCh38, because reads incorrectly align to distant incorrect copies of these genes (*CBSL*, *CRYAA2* and *KCNE1B*, respectively; Fig. 4 and Supplementary Figs. 6 and 7). Clarification of these regions is important, such as for *CBS*, deficiency of which is associated with homocystinuria (a disorder associated with thromboembolic events), skeletal abnormalities and intellectual disability.

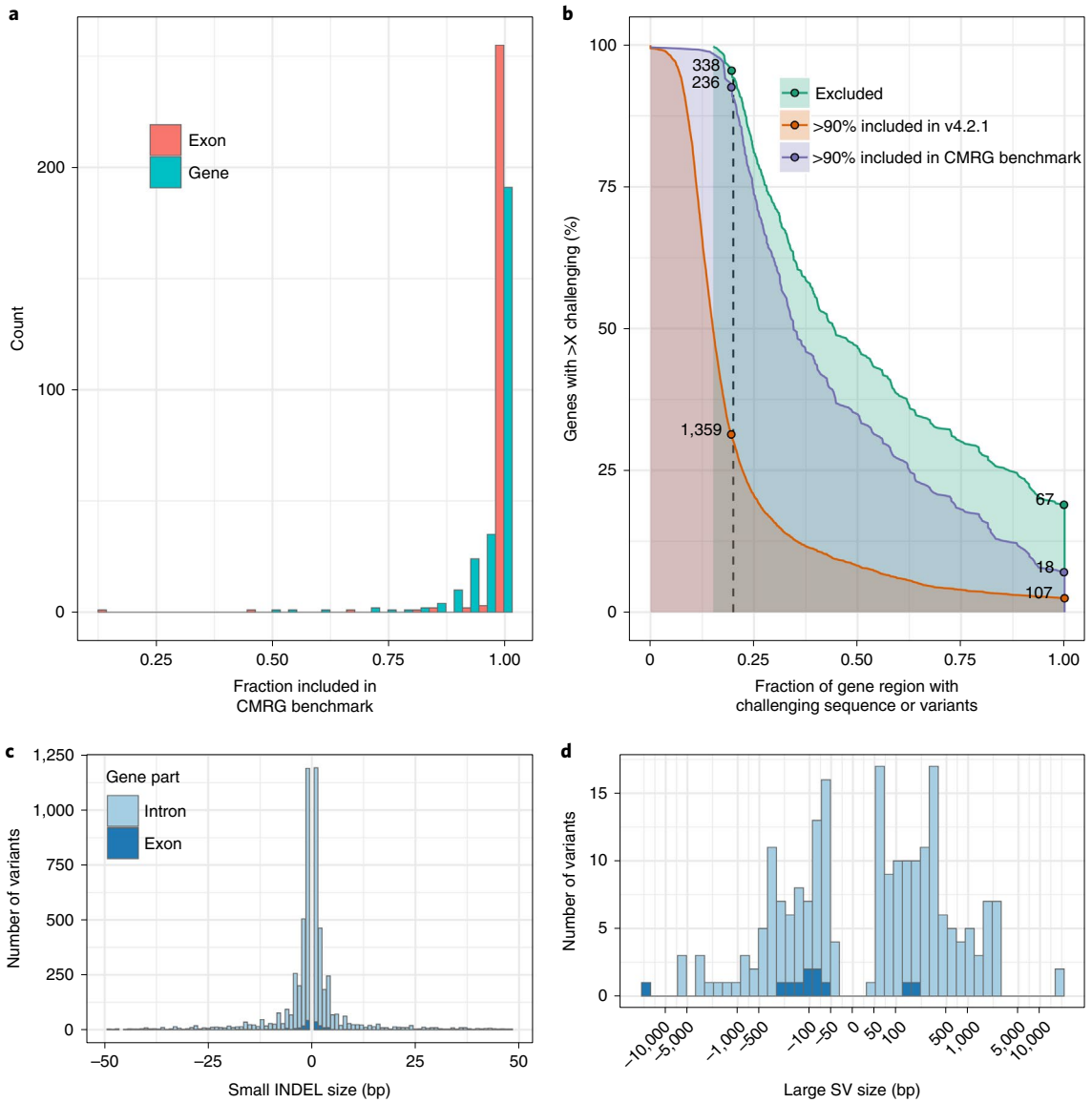

Fig. 2 | The new CMRG benchmark contains more challenging variants and regions than previous benchmarks. a, Fraction of each gene region (blue) and exonic regions (red) included in the new CMRG small variant or SV benchmark regions. **b**, Comparison of the fraction of challenging sequences and variants for genes included in the new CMRG benchmark versus the previous v4.2.1 HG002 benchmark versus genes excluded from both benchmarks. 99% of all CMRG benchmark genes have challenging sequences or variants in at least 15% of the gene region. The catalog of repetitive challenging sequences comes from GIAB and the Global Alliance for Genomics and Health (see 'difficult context' in Table 1). Challenging variants for HG002 are defined as complex variants (i.e., more than one variant within 10 bp) as well as putative SVs and putative duplications excluded from the HG002 v4.2.1 benchmark regions. **c**, Size distribution of INDELs in the small variant benchmark, which includes some larger INDELs in introns (light blue) and exons (dark blue). **d**, Size distribution of large insertions and deletions in the SV benchmark in introns (light blue) and exons (dark blue).

Most cases of homocystinuria are detected by newborn screening, and subsequent molecular evaluation can help confirm the diagnosis and provide recurrence risk information for families of affected individuals. *H19*, a noncoding gene on chromosome 11 that is frequently evaluated in cases of Beckwith–Wiedemann syndrome[25], is similarly affected by a false duplication on GRCh38. The additional copies of *CBS, U2AF1, CRYAA* and *KCNE1* in GRCh38 do not occur in HG002, and the Genome Reference Consortium and Telomere-to-Telomere Consortium recently determined that several regions on the p arm of chromosome 21, as well as several other regions in GRCh38, were incorrectly duplicated[26,27]. In support of this, the gnomAD v2 database has normal coverage and variants called in these genes for GRCh37, but gnomAD v3 has very low coverage and few variants for GRCh38. A companion manuscript from the Telomere-to-Telomere Consortium demonstrates that the new T2T-CHM13 reference corrects these and additional false duplications affecting 1.2-Mbp and 74 genes[27].

We worked with the Genome Reference Consortium to produce a new masking file that changes the sequence in the falsely duplicated regions of chromosome 21 on GRCh38 to N's. Masking in this way maintains the same coordinates but dramatically improves variant calling in the genes. Previous work demonstrated that variant calls could be recovered even from short reads by masking additional copies of highly homologous 'camouflaged' gene sequences, although this approach does not determine in which gene copy the variants occurred[8,28]. In our case, we are masking additional

Table 1 | Number of bases and variants in different HG002 GIAB benchmarks sets included in the 273 genes in the CMRG benchmark

Benchmark set	Base pairs in CMRG benchmarks	Base pairs in difficult context	Number of variants
CMRG small variants	11,719,200 (11.5)	4,500,129 (13.4)	27,178 (5.2)
v4.2.1 small variants	9,763,722 (12.0)	2,637,132 (16.3)	16,804 (6.3)
CMRG SVs	12,020,518 (11.4)	4,792,096 (13.0)	217 (5.1)
v0.6 SVs	10,569,811 (9.6)	3,215,766 (13.1)	170 (4.7)

We denote the percentage of base pairs or variants intersecting exons in parentheses. Difficult context is defined as the union of all tandem repeats, all homopolymers >6 bp, all imperfect homopolymers >10 bp, all difficult to map regions, all segmental duplications, regions with G+C content <25% or >65%, 'bad promoters' and 'other difficult regions'. Challenging variants for HG002 are defined as complex variants (i.e., more than one variant within 10 bp) as well as putative SVs and putative duplications excluded from the HG002 v4.2.1 benchmark regions. The number of bases and variants is provided for benchmarks on GRCh38, except for v0.6, where only a GRCh37 benchmark is available.

gene copies that are incorrect in the reference, enabling unambiguous variant calling in the correct genes. We show that masking the false duplications substantially improves recall and precision of variant calls in these genes for Illumina, PacBio HiFi and ONT mapping-based methods, increasing sensitivity of a common pipeline using Illumina, Burrows–Wheeler Aligner maximal exact match (BWA-MEM) and the Genome Analysis Toolkit (GATK) from 8% to 100% (Fig. 4), without increasing errors in other regions (Supplementary Fig. 8).

Our benchmark also identified some falsely duplicated genes in GRCh37, specifically the medically relevant genes *MRC1* and *CNR2*. Both short and long reads map correctly to *MRC1* in GRCh38, but many reads incorrectly align to a false additional copy of the gene in GRCh37. Similarly, *CNR2* is annotated on GRCh37 to include a large region downstream that has an erroneous additional unplaced contig on chromosome 1 (chr1_gl000191_random) that interferes with mapping to a 106-kb region that includes part of *CNR2* as well as other genes (*PNRC2* and *SRSF10*) not included in our medical gene list. Our benchmark correctly resolves all of these genes on both GRCh37 and GRCh38, because the assembled contigs align correctly for each haplotype.

The CMRG benchmark identified false positives that were eliminated by adding hs37d5 decoy sequence to GRCh37, but it also identified false negatives caused by the decoy. The hs37d5 decoy was created from assembled sequences not in the GRCh37 reference and was used in phase 2 of the 1000 Genomes Project to remove some false positives due to mismapped reads from these sequences[29]. To evaluate the impact of the decoy on variant call accuracy in our CMRGs, we benchmarked HG002 Illumina–BWA-MEM–GATK calls against our benchmark with and without adding the hs37d5 decoy sequence to the GRCh37 reference. Using the decoy eliminated 1,272 false-positive SNVs and INDELs in the medical gene benchmark, including 1,191 in *KMT2C*, 15 in *MUC5B* and the remainder in clusters of false positives in long interspersed nuclear elements, short interspersed nuclear elements and long terminal repeats in other genes. However, using the decoy sequence also caused 78 SNV and INDEL false negatives, notably 52 in *CYP4F12* and 18 in *LMF1* due to falsely duplicating parts of these genes. Therefore, while the hs37d5 decoy improves overall performance of variant calling, it can cause some false negatives in medically relevant genes similar to the false duplications in the primary assemblies discussed above. A potential solution may be to mask the falsely duplicated portions of the hs37d5 decoy similar to the masking of false duplications in GRCh38.

Benchmark reliably identifies variant calling errors. We evaluated the CMRG small variant benchmark by comparing seven variant callsets from short- and long-read technologies and a variety of mapping and assembly-based variant calling methods. The goal of this curation process is to verify that the CMRG benchmark reliably identifies false positives and false negatives across sequencing technologies and variant calling methods. Manual curation of a random subset of 20 false positives, 20 false negatives and 20 genotyping errors from each callset (split evenly between GRCh37 and GRCh38 and between SNVs and INDELs) demonstrated that most types of discrepancies were errors in each callset (Supplementary Fig. 9). However, the majority of INDEL differences were identified as errors in the benchmark for two callsets, and curation identified 215 small regions with errors in the benchmark. These errors included missing haplotypes (particularly heterozygous INDELs in otherwise homozygous regions) and errors due to noise in the HiFi data in very long homopolymers, as detailed in Supplementary Note 3. We also excluded 33 errors found in manual curation of complex small variants in tandem repeats (Supplementary Note 3), such as *MUC5B* in Supplementary Fig. 10. To more completely exclude these errors in the CMRG benchmark, we also curated all of the false positives, false negatives and genotyping errors that were in at least half of the callsets on GRCh37 or GRCh38. We found that 44 of 50 and 59 of 63 errors identified by the evaluation on GRCh37 and GRCh38, respectively, were excluded by curation of the common false positives and false negatives. After excluding these errors, v1.00 accurately identifies errors for both SNVs and INDELs. We have included our full curation results in Supplementary Data 5, which gives coordinates of common errors on both GRCh37 and GRCh38. This table can be used as a resource for investigating false positives and false negatives identified in a user's query callset, as we provide notes about the evidence for the benchmark at each common false-positive or false-negative site.

We evaluated the CMRG SV benchmark by comparing four short- and long read-based callsets, finding that the benchmark reliably identified false positives and false negatives across all four callsets. Upon manual curation only two sites were identified as problematic due to different representations that current benchmarking tools could not reconcile. We also found that the benchmarking statistics were sensitive to benchmarking tool parameters, particularly for duplications (Supplementary Note 3). We further confirmed that the 50 SVs ≥500 bp were all supported by Bionano Genomics (Bionano) optical mapping-based SV calling. From the manual curation of common false positives, false negatives and genotyping errors, we also identified some categories of variants in which the benchmark correctly identified errors in the majority of callsets: (1) clusters of false negatives and genotyping errors in the genes that are falsely duplicated in GRCh37 (*MRC1* and part of *CNR2*) and GRCh38 (*CBS*, *CRYAA*, *KCNE1* and *H19*); and (2) clusters of false positives and genotyping errors due to mismapped reads in the parts of *KMT2C* that are duplicated in HG002 relative to GRCh37 and GRCh38, which are responsible for 277 of the 386 false positives in the HiFi DeepVariant callset (Supplementary Fig. 11). We also determined that the benchmark correctly identified false negatives across technologies, but particularly short read-based methods, in segmental duplications like *SMN1* and *NCF1*, and in gene conversion-like events in *RHCE*, *SIGLEC16* and *GTF2IRD2*. In addition to previously developed stratifications for difficult regions, we developed new stratifications for falsely duplicated genes, genes with large duplications and complex variants in tandem repeats, which we available in the GIAB v3.0 stratifications (Supplementary Note 4).

We further confirmed 225 of 226 variants across 10 genes in segmental duplications that were covered confidently by an orthogonal long-range PCR and Sanger sequencing method (Supplementary Table 1 and Supplementary Data 6). A total of 127 other variants that we attempted to confirm did not have coverage or had noisy

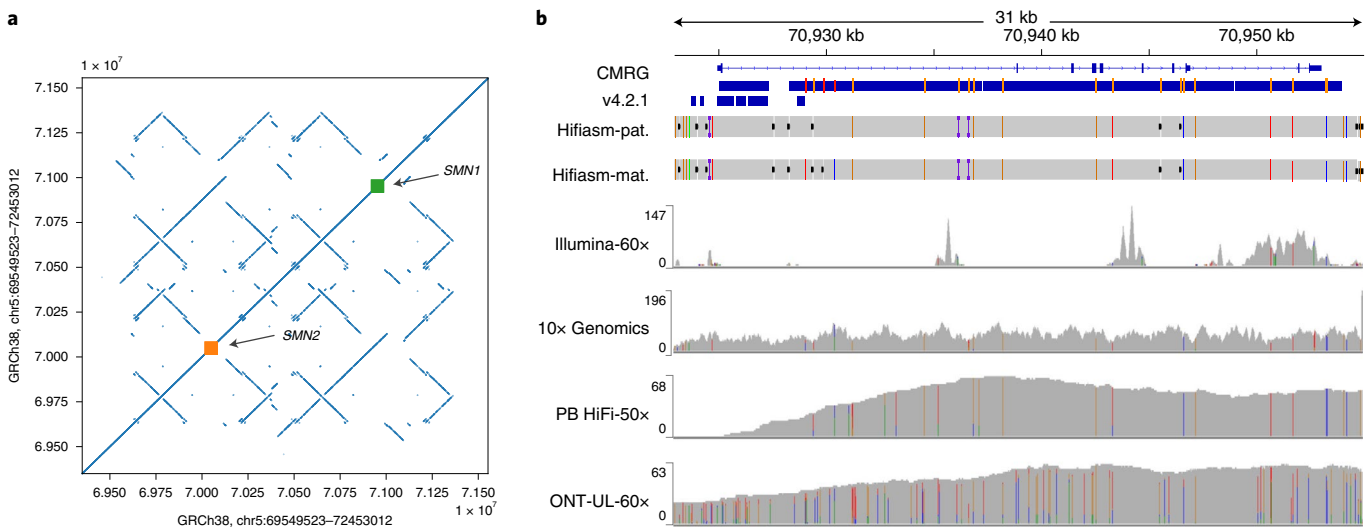

Fig. 3 | The new benchmark covers the gene *SMN1*, which was previously excluded due to mapping challenges for all technologies in the highly identical segmental duplication. a, Dotplot of GRCh38 against GRCh38 in the SMA region, showing a complex set of inverted repeats that make it challenging to assemble. **b**, Integrated Genomics Viewer view showing that only a small portion of *SMN1* was included in v4.2.1 and that all technologies have challenges mapping in the region, but 10x Genomics and ultralong ONT reads support the variants called in the new CMRG benchmark. For the CMRG and v4.2.1 benchmarks, thick blue bars indicate regions included by each benchmark, and orange and light blue lines indicate positions of homozygous and heterozygous benchmark variants, respectively. CMRG variants were called from the trio-based hifiasm assembly of paternal and maternal haplotypes (Hifiasm-pat and Hifiasm-mat, respectively). Coverage tracks are show for 60× PCR-free Illumina 2 × 150-bp reads (Illumina-60×), 10x Genomics-linked reads (10x Genomics), 50× PacBio HiFi 15- and 20-kbp reads (PB HiFi-50×) and 60× ONT ultralong reads (ONT-UL-60×).

sequencing, and only one variant (a homozygous SNV at GRCh38 chr.16:2113578 in *PKD1*) was contradicted by long-range PCR but clearly supported by Illumina, 10x Genomics, PacBio HiFi and ONT (Supplementary Fig. 12).

To demonstrate how the CMRG benchmark can identify new types of errors relative to v4.2.1, we benchmarked a stringently filtered Illumina–BWA-MEM–GATK callset versus both the v4.2.1 benchmark and the medically relevant gene benchmark. Figure 5 shows that the fraction not assessed decreases and the false-negative rate increases substantially overall, but particularly for difficult variants. For SNVs, these difficult variants fall primarily in segmental duplications and low-mappability regions, while for INDELs, the CMRG benchmark also identifies additional false negatives in other regions excluded from the 'not in all difficult' stratification, such as tandem repeats and homopolymers.

Remaining challenges across medically relevant genes. While the CMRG benchmark covers many new, challenging genes, 122 autosomal genes covered <90% by v4.2.1 are still excluded from the CMRG benchmark (110 on GRCh37 and 100 on GRCh38) for multiple reasons detailed in Supplementary Data 7; when progressively categorizing excluded genes on GRCh38, (1) 20 genes were affected by gaps in the reference; (2) 38 genes had evidence of duplications in HG002 relative to GRCh38; (3) six genes were resolved but excluded due to being in the MHC region[14]; (4) three genes were resolved on GRCh38 but not GRCh37, as we required genes to be resolved on both references; (5) 19 genes were >90% included by the dip.bed but had multiple contigs or a break in the assembly–assembly alignment; (6) seven genes had a large deletion of part or all of the gene on one haplotype; (7) four genes had breaks or false duplications in the hifiasm assembly; (8) two genes were in the structurally variable immunoglobulin locus; and (9) one gene (*TNNT3*) had a structural error in GRCh38 (described in ref. [27]).

As examples, *LPA* and *CR1* were not included in the benchmark due to very large insertions and deletions, respectively, that cause a break in contig alignments, although the hifiasm assembly resolved both haplotypes (Supplementary Figs. 13 and 14). *LPA* contains multiple tandemly duplicated copies of the same region (i.e., kringle IV repeats with a unit length of ~5,550 bp) that are associated with cardiovascular disease[30]. The HG002 hifiasm assembly resolved the entire *LPA* region, and the 44.1-kb and 99.9-kb expansions of the kringle IV repeats for the maternal and paternal haplotypes, respectively, were consistent with the insertions predicted by an independent trio-phased Bionano optical mapping assembly (45.0 kb and 101.2 kb). This complex, large expansion of the kringle IV repeats can be represented in many different ways in a variant call format (VCF) with different levels of precision (e.g., as a large insertion, a tandem duplication or a CNV, and the copies may differ or include small variants). Existing benchmarking tools cannot compare these different representations robustly, partly owing to limitations of the VCF[31]. To benchmark assemblies of this gene in HG002, the sequences could be compared directly to the hifiasm contigs, which we have annotated for *LPA* and other genes using LiftOff[32]. *CR1*, a gene implicated in Alzheimer's disease[8], is similarly resolved by hifiasm and contains a 18.5-kb homozygous deletion consistent with Bionano, but this deletion causes a break in the dipcall/minimap2 alignment (Supplementary Fig. 14).

Other genes are excluded from the benchmark because they have additional copies in HG002, but not in GRCh38. For example, *KCNJ18* is excluded because GRCh37 and GRCh38 are missing a copy of this gene (*KCNJ17*), so additional contigs from *KCNJ17* align to *KCNJ18* (ref. [27]). Also, genes in the KIR region are highly variable, and CNVs are observed frequently in the population, with 35 alternate loci and 15 novel patches in GRCh38.p13. Hifiasm resolves the paternal allele in a single contig, but the maternal allele is split into three contigs in the KIR region, including a tandem duplication of the gene *KIR2DL1* (Supplementary Fig. 15). There is no standard way to represent or benchmark small variants within duplicated regions, so we excluded *KCNJ18*, *KIR2DL1* and other duplicated genes such as *PRSS1* and *DUX4* from our benchmarks (Supplementary Figs. 16 and 17). More information about these complex genes is in Supplementary Note 5.

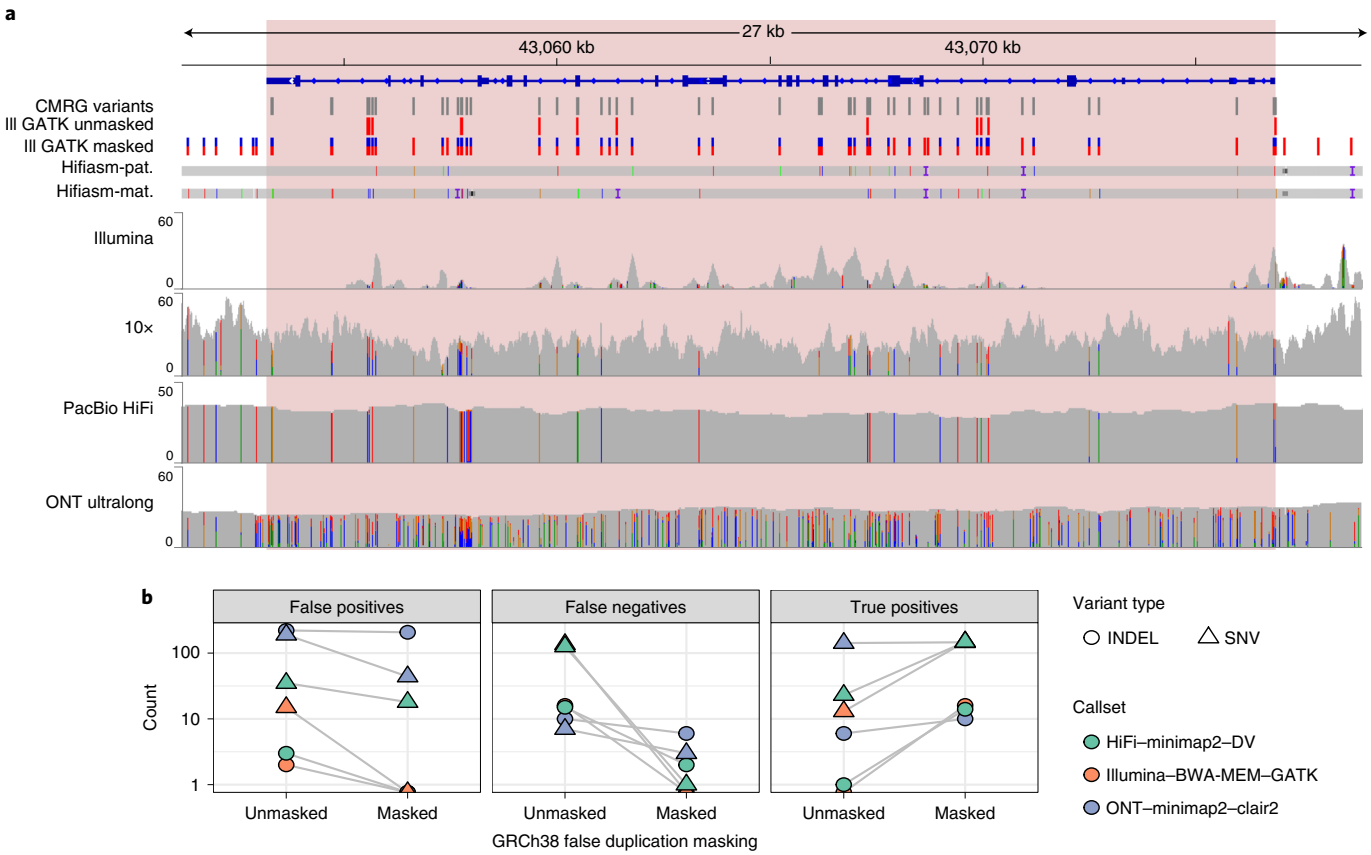

Fig. 4 | The benchmark resolves the gene CBS, which has a highly homologous gene (CBSL) due to a false duplication in GRCh38 that is not in HG002 or GRCh37. a, The duplication in GRCh38 causes Illumina and PacBio HiFi reads from one haplotype to mismap to *CBSL* instead of *CBS*. The ultralong ONT reads, 10x Genomics-linked reads and assembled PacBio HiFi contigs map properly to this region for both haplotypes, because they contain sufficient flanking sequence. When the falsely duplicated sequence is masked using our new version of GRCh38, variant calls from a standard Illumina–GATK pipeline (ILMN-GATK w/Mask VCF) are completely concordant with the new benchmark. Pink shaded box indicates CMRG benchmark regions; only variants within the benchmark regions are included in the benchmark. **b**, Comparison of variant accuracy for GRCh38 before and after masking false duplications on chromosome 21 using variant callsets from three technologies: HiFi-minimap2-DeepVariant (HiFi-minimap2-DV), Illumina-BWA-MEM-GATK, and ONT-minimap2-clair2. The new benchmark demonstrates decreases in false-negative and false-positive errors for three callsets in the falsely duplicated genes *CBS*, *CRYAA* and *KCNE1* when mapping to the masked GRCh38.

Discussion

In this work, we provide highly curated benchmarks for both phased small variants and SVs covering 273 medically relevant and challenging genes. Parts or all of these genes are often excluded from standard targeted, exon or whole-genome sequencing or analysis. Still, the impact of these genes is well documented across multiple diseases and studies. Our benchmark will pave the way to obtain comprehensive insights into these highly relevant genes to further expand medical diagnoses and potentially improve understanding of the heritability for multiple diseases[33]. We give specific examples of challenges with calling variants in these genes, including mapping challenges for different technologies and identifying genes for which GRCh37 or GRCh38 is a better reference. This benchmark was designed to be complementary to previous mapping-based benchmarks. Some difficult genes, such as *PMS2*, are resolved well by v4.2.1 in HG002 but not by the assembly. Some difficult genes, such as the HLA family[14] or *GBA1*/*GBA2*, are resolved well by the assembly but not included in the benchmark because they were well resolved previously.

Still, a few challenging regions remain excluded from our benchmark or are not resolvable despite the availability of highly accurate long-read data. Some genes include variable long tandem repeats (e.g., *LPA* and *CR1*), which are resolved in our assembly, but the large >20-kb changes in length of the alleles are currently too complex for standard benchmarking methodologies. This clearly shows the need for more advanced methods, such as graph representations of haplotypes or alleles. In addition, a few genes (e.g., *SMN2*) escaped a comprehensive and accurate assessment even with current long-read-based assembly methods, highlighting the need for further development of sequencing and bioinformatics methods. Furthermore, our extensive curation of the benchmark helped identify limitations of the current haplotype-resolved whole-genome assembly methods, paving the way for future whole-genome assembly-based benchmarks: (1) the assembly often misses one allele for heterozygous INDELs in highly homozygous regions; (2) some consensus errors exist, causing errors in a single read to be called as variants; and (3) if both haplotypes of the assembly do not completely traverse segmental duplications, then the assembly is less reliable (e.g., *SMN2* in HG002), although it can sometimes be correct (e.g., *SMN1* in HG002). Some genes also may be resolvable in HG002 but not in other genomes, or vice versa, due to structural or copy-number variability in the population, so benchmarks for additional samples will be needed.

By basing this benchmark on a haplotype-resolved whole-genome assembly, we were able to identify biases in mapping-based methods due to errors in the GRCh37 and GRCh38 references.

Fig. 5 | The new CMRG small variant benchmark includes more challenging variants and identifies more false negatives in a standard short-read callset (Illumina–BWA-MEM–GATK) than the previous v4.2.1 benchmark in these challenging genes. While the false-negative rate (circles) is similar in easier regions (purple 'Not in all difficult' points), the false-negative rate is much higher overall (green 'All CMRG benchmark regions' points). The fraction of variants excluded from the benchmark regions (triangles) is much higher for the v4.2.1 benchmark in all stratifications. Challenging regions from the v3.0 GIAB stratifications shown here include complex variants in tandem repeats (TR) longer than 100 bp, segmental duplications (SegDup), and regions difficult to map with 100 bp reads (LowMap). This information is also presented in 'summary stats NYGC' in Supplementary Data 4.

While previous studies concluded that variant calling performance is generally better on GRCh38 (refs. [34,35]), our benchmark demonstrates that variant calls in some genes are less accurate on GRCh38 than GRCh37. Another group recently independently identified the importance of masking the additional copy of one gene (*U2AF1/U2AF1L5*) for cancer research[36]. Our results identify that false duplications cause many of the discrepancies found recently between exome variant calls on GRCh37 and GRCh38 (ref. [37]). We produced similar benchmarks for both versions of the reference so that scientists can better understand the strengths and weaknesses of each reference and test modifications to the reference, such as the hs37d5 decoy for GRCh37 or the masked GRCh38 we propose here. During this process, we also identified and resolved variant calling errors due to several false duplications in these medically relevant genes in GRCh38 on chromosome 21. Overall, 11 genes are impacted by these false duplications, including three medically relevant genes from our list (*CBS*, *KCNE1* and *CRYAA*). As a solution to this problem, we provide a GRCh38 reference that masks the erroneous copy of the duplicated genes. We use our benchmark to show that this reference dramatically improves read mapping and variant calling in these genes across almost all sequencing technologies. These false duplications exist only in GRCh38 and not in other human reference genome versions or in the broader population. A new telomere-to-telomere reference genome eliminates these false duplications and fixes collapsed duplications that prevented us from creating a benchmark for medically relevant genes like *KCNJ18* and *MAP2K3*, and a similar CMRG benchmark for HG002 is now available on the new reference[27]. Future work will include using haplotype-resolved assemblies to form benchmarks for more genic and nongenic regions of the genome, eventually using genomes that are assembled telomere to telomere.

Our approach to form benchmarks from a haplotype-resolved whole-genome assembly is a prototype for future comprehensive benchmarks covering the whole genome combining different types of small variants and SVs. Overall, this benchmark enables a more comprehensive assessment of sequencing strategies, analytical methodologies and other developments for challenging genomic variants and regions relevant to medical research[5,38], paving the way for improved clinical diagnoses.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-01158-1.

Received: 7 June 2021; Accepted: 10 November 2021;
Published online: 7 February 2022

References

1. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
2. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
3. Nurk, S. et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
4. Shafin, K. et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053 (2020).

5. Mahmoud, M. et al. Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).
6. De Coster, W., Weissensteiner, M. H. & Sedlazeck, F. J. Towards population-scale long-read sequencing. *Nat. Rev. Genet.* **22**, 572–587 (2021).
7. Mandelker, D. et al. Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet. Med.* **18**, 1282–1289 (2016).
8. Ebbert, M. T. W. et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol.* **20**, 1–23 (2019).
9. Lincoln, S. E. et al. One in seven pathogenic variants can be challenging to detect by NGS: an analysis of 450,000 patients with implications for clinical sensitivity and genetic test implementation. *Genet. Med.* **23**, 1673–1680 (2021).
10. Zook, J. M. et al. An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.* **37**, 561–566 (2019).
11. Zook, J. M. et al. A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020) ; erratum **38**, 1357 (2020).
12. Olson, N. D. et al. precisionFDA Truth Challenge V2: calling variants from short- and long-reads in difficult-to-map regions. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.13.380741 (2020).
13. Wagner, J. et al. Benchmarking challenging small variants with linked and long reads. Preprint at *bioRxiv* https://doi.org/10.1101/2020.07.24.212712 (2020).
14. Chin, C.-S. et al. A diploid assembly-based benchmark for variants in the major histocompatibility complex. *Nat. Commun.* **11**, 4794 (2020).
15. Goldfeder, R. L. et al. Medical implications of technical accuracy in genome sequencing. *Genome Med.* **8**, 24 (2016).
16. Ball, M. P. et al. A public resource facilitating clinical use of genomes. *Proc. Natl Acad. Sci. USA* **109**, 11920–11927 (2012).
17. Tate, J. G. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
18. Ross, M. G. et al. Characterizing and measuring bias in sequence data. *Genome Biol.* **14**, R51 (2013).
19. Prior, T. W., Leach, M. E. & Finanger, E. Spinal muscular atrophy. In *GeneReviews* [Internet] (University of Washington, 2020).
20. Biros, I. & Forrest, S. Spinal muscular atrophy: untangling the knot? *J. Med. Genet.* **36**, 1–8 (1999).
21. Leiding, J. W. & Holland, S. M. Chronic granulomatous disease. In *GeneReviews* [Internet] (University of Washington, 2016).
22. Innan, H. A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. *Proc. Natl. Acad. Sci. USA* **100**, 8793–8798 (2003).
23. Hayakawa, T. et al. Coevolution of Siglec-11 and Siglec-16 via gene conversion in primates. *BMC Evol. Biol.* **17**, 228 (2017).
24. Garg, P. et al. Pervasive cis effects of variation in copy number of large tandem repeats on local DNA methylation and gene expression. *Am. J. Hum. Genet.* https://doi.org/10.1016/j.ajhg.2021.03.016 (2021).
25. Lennerz, J. K. et al. Addition of H19 'loss of methylation testing' for Beckwith-Wiedemann syndrome (BWS) increases the diagnostic yield. *J. Mol. Diagn.* **12**, 576–588 (2010).
26. Nurk, S. et al. The complete sequence of a human genome. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.26.445798 (2021).
27. Aganezov, S. et al. A complete reference genome improves analysis of human genetic variation. Preprint at *bioRxiv* https://doi.org/10.1101/2021.07.12.452063 (2021).
28. Boisson, B. et al. Rescue of recurrent deep intronic mutation underlying cell type–dependent quantitative NEMO deficiency. *J. Clin. Invest.* **129**, 583–597 (2018).
29. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
30. Schmidt, K., Noureen, A., Kronenberg, F. & Utermann, G. Structure, function, and genetics of lipoprotein (a). *J. Lipid Res.* **57**, 1339–1359 (2016).
31. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 265 (2020).
32. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinform.* **37**, 1639–1643 (2020).
33. Theunissen, F. et al. Structural variants may be a source of missing heritability in sALS. *Front. Neurosci.* **14**, 47 (2020).
34. Guo, Y. et al. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis.*Genomics* **109**, 83–90 (2017).
35. Pan, B. et al. Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinform.* **20**, 101 (2019).
36. Miller, C. A. et al. Failure to detect mutations in U2AF1 due to changes in the GRCh38 reference sequence. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.07.442430 (2021).
37. Li, H. et al. Exome variant discrepancies due to reference-genome differences. *Am. J. Hum. Genet.* **108**, 1239–1250 (2021).
38. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **590**, E55 (2021).

Methods

Sample availability. For the 10x Genomics and ONT sequencing and Bionano mapping, the GM24385 (RRID:CVCL_1C78) cell line was obtained from the Coriell Institute for Medical Research National Institute for General Medical Sciences cell line repository. For the Illumina and PacBio sequencing, National Institute of Standards and Technology (NIST) RM 8391 DNA was used, which was prepared from a large batch of GM24385 to control for differences arising during cell growth. For binning reads into paternal and maternal haplotypes, Illumina sequencing of DNA from NIST RM 8392 (HG002-HG004) was used. DNA was extracted from cell lines publicly available as GM24149 (RRID:CVCL_1C54) and GM24143 (RRID:CVCL_1C48) at the Coriell Institute for Medical Research National Institute for General Medical Sciences cell line repository.

Medical genes. We used genes from a variety of databases and sources to compile a list of medically relevant genes. The largest set of genes we use is from Supplementary Table 13 of Mandelker et al., which was a capture of the OMIM, HGMD and ClinVar databases gathered in 2012. Further, we used the COSMIC cancer gene census, which is a list of 723 genes. Supplementary Data 1 also contains additional details about the higher-priority list of 942 genes in the union of ClinGen genes with 'definitive', 'strong' or 'moderate' evidence (719 genes), National Comprehensive Cancer Network/European Society for Medical Oncology (hereditary cancer syndromes) (49 genes), American College of Medical Genetics Secondary Findings 2.0 (commonly referred to as the ACMG59, for which reporting of secondary or incidental findings is recommended) (59 genes), Clinical Pharmacogenetics Implementation Consortium pharmacogenetics genes (127 genes), and the Counsyl expanded carrier screening list (235 genes), which includes recommended reproductive medicine genes as a small subset.

Medical gene coordinate discovery. We used coordinates from ENSEMBL (https://uswest.ensembl.org/index.html) and then downloaded 'chromosome', 'start', 'end', 'gene_name' and 'stable_ID' using bioMart for GRCh38 and GRCh37. We looked up the collection of medical genes and found the coordinates for each in GRCh38 and GRCh37, with the full lists (GRCh3x_mrg_full_gene.bed) and 273 genes included in the CMRG benchmark (GRCh3x_mrg_bench_gene.bed) available under https://github.com/usnistgov/cmrg-benchmarkset-manuscript/tree/master/data/gene_coords/unsorted.

Calculating overlap with GIAB HG002 v4.2.1 small variant benchmark. We used bedtools[39] intersected with the ENSEMBL coordinates for each gene and the v4.2.1 small variant benchmark regions browser extendible data (BED) files. We calculated the number of bases in the intersection and compared that to the total number of bases in each gene. We chose 90% as the threshold for the purpose of keeping manual curation tractable over the set of the genes.

Haplotype-resolved assembly using PacBio HiFi reads with hifiasm using trio-binning. We used the haplotype-resolved assembly produced by hifiasm v0.11 using 34× coverage (two 15-kb and two 20-kb libraries) by PacBio HiFi Sequel II System with Chemistry 2.0 reads (https://github.com/human-pangenomics/HG002_Data_Freeze_v1.0#hg002-data-freeze-v10-recommended-downsampled-data-mix) using *k*-mer information from parental Illumina short reads (30× 2 × 150-bp reads at https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=NHGRI_UCSC_panel/HG002/hpp_HG002_NA24385_son_v1/parents/ILMN/downsampled/), described recently[2].

Calling variants relative to GRCh37 and GRCh38 using dipcall. We aligned the haplotype-resolved assembly of HG002 to GRCh37 and GRCh38 using minimap2 (ref. [40]) through dipcall (https://github.com/lh3/dipcall) as is done in the NIST assembly benchmarking pipeline (https://github.com/usnistgov/giab-asm-benchmarking). Dipcall generates variant calls using any nonreference support in regions that are ≥50 kb, with contigs having mapping quality ≥5. Dipcall also produces a BED that denotes confident regions that are covered by an alignment ≥50 kb, with contigs having mapping quality ≥5 and no other >10-kb alignments.

Benchmark development. We selected genes that had continuous haplotype coverage of the gene body, including the 20 kb on each side to account for robust alignments. In addition, each haplotype had to fully cover any segmental duplications in close proximity to or overlapping the extended gene regions. This also included complex SVs inside of the segmental duplications to be able to robustly identify SNVs and SVs subsequently. We considered a gene to be fully resolved by the haplotype-resolved assembly if the dip.bed covered the gene along with 20 kb of flanking sequence to consider the PacBio HiFi read length as well as any overlapping segmental duplications. We chose these criteria to ensure that genes were resolved in regions with high-quality assembly.

We then performed manual curation of the resolved genes and flanking sequence to understand overall characteristics of the candidate benchmark. We began initial evaluation against mapping-based callsets to understand the performance of the benchmark in these genes. We found that perfect homopolymers of >20 bp and imperfect homopolymers >20 bp accounted for a majority of false negatives and false positives for both SNVs and INDELs. Imperfect homopolymers are defined as stretches of one base that are interrupted by one different base in one or more locations, and each of the stretches of exact homopolymer bases has to be at least 4 bp (e.g., AAAAGAAAAAGAAAATAAAA). Manual curation of a random subset of these sites showed that in most instances, it was unclear whether the mapping-based callset or the assembly-based benchmark was correct. BED files for these homopolymers are available under https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/.

We excluded the following regions from the v0.02.03 small variant and v0.01 SV benchmark regions (the benchmark versions used in the evaluation): (1) one region identified manually as an erroneous insertion resulting from an issue with the method hifiasm v0.11 used to generate the consensus sequence; (2) genes in the MHC, as these were previously resolved by diploid assembly in the v4.2.1 benchmark[14]; and (3) regions around variants identified as errors or unclear upon manual curation, as described below. For the small variant benchmark, we additionally excluded (1) SVs at least 50 bp in size and overlapping tandem repeats, because these cannot be compared robustly with small variant comparison tools; and (2) perfect and imperfect homopolymers of >20 bp plus 5 bp on each side. For the SV benchmark, we additionally excluded (1) tandem repeats that contain more than one variant at least 10 bp in size, because these complex variants can cause inaccurate comparisons with current benchmarking tools; and (2) INDELs 35–49 bp in size.

Benchmark evaluation. We used hap.py[41] with vcfeval to compare VCFs from a variety of sequencing technologies and variant calling methods to the GRCh37 and GRCh38 difficult medical gene small variant benchmark, with v3.0 GIAB/GA4GH stratifications under https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/. We randomly selected 60 total sites for curation, with 30 selected from GRCh37 and 30 selected from GRCh38. Five SNVs and five INDELs were selected from each of these three categories: (1) false positives (variants in the comparison VCF but not the benchmark), (2) false negative (variants in the benchmark but not the comparison VCF), and (3) genotype errors (variant appearing as both a false positive and a false negative using hap.py with vcfeval). This curation process will also help us to make further refinements, if needed, to the GIAB benchmark. For the small variant benchmark evaluation, we used seven VCFs[12] from short- and long-read technologies and a variety of mapping and assembly-based variant calling methods: (1) Illumina–DRAGEN, (2) Illumina–NovaSeq–GATK4 (ref. [42]), (3) Illumina–xAtlas[43], (4) PacBio HiFi–GATK4,(5) an assembly based on ONT reads called with dipcall, (6) a union of three callsets (Illumina called with modified GATK, PacBio HiFi called with Longshot[44] v0.4.1, and ONT called with PEPPER–DeepVariant[45], and (7) Illumina, PacBio and ONT combined called with NeuSomatic[46]. We excluded errors identified upon curation, as described in Supplementary Note 3. In Supplementary Note 4, we also include performance metrics for the 53x 15 kb + 20 kb HiFi-DeepVariant v0.9 callset under https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/PacBio_CCS_15kb_20kb_chemistry2_10312019/.

Variant callsets used for evaluation. *National Center for Biotechology Information (NCBI) de novo assembly.* The de novo assembly of HG002 was initially generated using NextDenovo2.2-beta.0 with ONT Promethion data (https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/UCSC_Ultralong_OxfordNanopore_Promethion/) and then polished with PacBio 15-kb and 20-kb circular consensus sequencing (or HiFi) reads (https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_15kb_20kb_chemistry2/reads/), followed by scaffolding with HiC data (https://github.com/human-pangenomics/HG002_Data_Freeze_v1.0). The scaffolded assembly was further polished with Illumina short reads (https://github.com/human-pangenomics/HG002_Data_Freeze_v1.0) twice using pilon[47] and then phased with Whatshap[48]. Finally, two VCF files (HG002_grch37_dipcall.vcf.gz and HG002_grch38_dipcall.vcf.gz) were generated based on phased HG002 genome using dipcall with GRCh37 and GRCh38 reference genomes, respectively (https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NCBI_variant_callsets_for_medicalgenes_evaluation_10282020).

Illumina–DRAGEN. HG002 DNA was prepared using the Illumina DNA PCR-free library preparation kit. The library was sequenced on the NovaSeq 6000 platform with 151-bp paired-end reads. Illumina–DRAGEN 3.6.3 was used to align sequencing reads and call variants. SNP and INDEL were filtered using the following hard filters:

DRAGENHardSNP:snp: MQ < 30.0 || MQRankSum < −12.5 || ReadPosRankSum < −8.0;DRAGENHardINDEL:indel: ReadPosRankSum < −20.0.

Illumina, PacBio and ONT combined called with NeuSomatic. The predictions are based on the adaptation of the deep learning-based framework in NeuSomatic for germline variant calling. We used the network model trained for NeuSomatic's submission for the PrecisionFDA truth challenge v2 (ref. [12]). The model is trained on HG002 using GIAB benchmark set v4.2. For this callset, separate input channels were used for PacBio, Illumina and ONT reads.

DNAnexus: union of short read callsets from four callers. We downloaded HG002 WGS FASTQ reads from NIST's FTP[49], followed by downsampling of the reads to 35× coverage (47.52%) from an original coverage of 73.65× using seqtk.

We called variants against both GRCh38 (GRCh38 primary contigs and decoy contigs, but no alternate contigs or HLA genes) and hs37d5 builds. We ran four different germline variant callers (see below) with their suggested default parameters, collected the union of all variants using a customized script where we recorded which caller(s) called the variant and their filter statuses in INFO and FILTER fields and generated the union VCF file for HG002. The customized script excludes variants (with the same chromosome, position, reference base and alternates) that have conflicting genotype reported by different callers and only keeps the variants that are reported as exactly the same genotype when more than one caller is calling it. The variant calling pipelines used were (1) BWA-MEM and GATK4 (BWA-MEM[50] version 0.7.17-r1188 (https://github.com/lh3/bwa) and GATK version gatk-4.1.4.1 (https://gatk.broadinstitute.org/hc/en-us)); (2) Parabricks_DeepVariant (Parabricks Pipelines DeepVariant v3.0.0_2 (https://developer.nvidia.com/clara-parabricks)); (3) Sentieon_DNAscope (Sentieon (DNAscope) version sentieon_release_201911 (https://www.sentieon.com/products/#dnaseq)); and (4) BWA-MEM and Strelka2 (BWA-MEM version 0.7.17-r1188 (https://github.com/lh3/bwa) and Strelka2 version 2.9.10 (https://github.com/Illumina/strelka)).

Illumina NovaSeq 2 × 250-bp data. The sample HG002 was sequenced on an Illumina NovaSeq 6000 instrument with 2 × 250-bp paired-end reads at the New York Genome Center. The libraries were prepped using a TruSeq DNA PCR-free library preparation kit. The raw reads were aligned to both GRCh37 and GRCh38 human references. Alignment to the GRCh38 reference, marking duplicates and base quality recalibration were performed as outlined in the Centers for Common Disease Genomics functional equivalence paper[51]. Alignment to GRCh37 was performed using BWA-MEM[50] (v0.7.8), marking duplicates was performed using Picard (v1.83) and local INDEL realignment and base quality recalibration were performed using GATK[52] (v3.4-0). Variant calling was performed using GATK (v3.5) and adhering to the best-practices recommendations from the GATK team. Variant calling constituted generating gVCF using HaplotypeCaller, genotyping using the GenotypeGVCFs subcommand and variant filtering performed using VariantRecalibrator and ApplyRecalibration steps. A tranche cutoff of 99.8 was applied to SNP calls and 99.0 to INDELs to determine PASS variants (i.e., variants that are not filtered). The raw reads are available for download at the Sequence Read Archive at https://www.ncbi.nlm.nih.gov/sra/SRX7925517.

Small variants from Illumina, PacBio and ONT. For this submission, we combined data from three sequencing technologies to obtain a more sensitive VCF file.

We used our in-house variant calling pipeline for the Illumina dataset. In short, BWA-MEM v0.7.15-r1140 was used to align reads to the GRCh37 or GRCh38 reference genome, and BAM files were processed with SAMtools[53] v1.3 and Picard v2.10.10. SNVs and INDELs were identified with the HaplotypeCaller following the best-practices workflow recommendations for germline variant calling in GATK v3.8 (ref. [50]).

For both PacBio and ONT datasets, we ran another pipeline using NanoPlot v1.27.0 for quality control (Filtlong v0.2.0 for filtering reads and minimap2 v2.17-r941 for alignment). Longshot v0.4.1 was used for variant calling for PacBio data and PEPPER-DeepVariant was used for ONT data.

On the variant callsets, we filtered out variants by applying the following criteria: FILTER = PASS, QD (quality by depth) ≥2.0 and MQ (mapping quality) ≥50 for Illumina data; and FILTER = PASS and QUAL (quality) ≥150 for PacBio data. No filters were applied to ONT calls.

Finally, we created a consensus VCF file by merging the single VCF files obtained by each of these three pipelines using the GATK CombineVariants tool.

SVs from Illumina (intersection callsets from five callers). We called SVs on short-read Illumina data using five different SV callers: DELLY[54] v0.8.5, GRIDSS[55] v2.9.4, LUMPY[56] v0.3.1, Manta[57] v1.6.0 and Wham[58] v1.7.0. The HG002 BAM file aligned to the GRCh37 or GRCh38 reference genome by BWA-MEM v0.7.15-r1140, with duplicates marked using Picard v2.10.10 and base quality score recalibrated by GATK v3.8, was used to feed these SV callers, which were executed with recommended default parameters. LUMPY and Wham SV calls were genotyped using SVTyper v0.7.1. GRIDSS SV types were assigned with the simple-event-annotation R script, included in the GRIDSS package.

Resulting SV callsets were filtered based on the author's recommendations for each caller as follows: Manta (FILTER = PASS, INFO/PRECISE, FORMAT/PR ≥ 10), LUMPY (INFO/PRECISE, remove genotypes 0/0, QUAL ≥ 100, FORMAT/AO ≥ 7), DELLY (FILTER = PASS, INFO/PRECISE), GRIDSS (FILTER = PASS, INFO/PRECISE, QUAL ≥1,000, INFO/SVLEN <1 kb, remove DAC Encode regions) and Wham (INFO/SVLEN <2 kb, INFO/A >5, remove genotypes 0/0, INFO/CW[bnd] >0.2).

The resulting VCF files from each caller were merged to create the intersection of variants using SURVIVOR[59] v1.0.7, containing variants >50 bp in size, with 1,000 bp as the distance parameter and without requiring any type specificity (all variant types are merged). In the intersection set, we retained calls supported by two or more callers.

SVs from ONT (merge callsets from two callers). For these submissions, we built a custom pipeline to process the ONT HG002 dataset using NanoPlot[60] v1.27.0 for quality control, Filtlong v0.2.0 for filtering reads and minimap2 (refs. [40,60]) v2.17-r941 for alignment to the GRCh37 or GRCh38 reference genome.

The SVs were called on the resulting BAM file using cuteSV v1.0.8 and Sniffles[61] v1.0.12. The resulting VCF files were filtered out based on the default values suggested by each of the tool's authors: cuteSV[62] (minimum read support of 10 reads, INFO/RE ≥10; this caller intrinsically filters by FILTER = PASS and INFO/PRECISE) and Sniffles (FILTER = PASS, INFO/PRECISE and minimum read support of 10 reads).

Finally, we created the VCF files by merging the single filtered VCF files using SURVIVOR[59] v1.0.7, containing variants >50 bp in size, with 1,000 bp as the distance parameter and without requiring any type specificity (all variant types are merged).

Remapping variants between GRCh38 and GRCh37. To remap curated variant locations between GRCh38 and GRCh37, we used the NCBI Remap tool. For variants that remapped in the first pass, we used the first pass location. For variants that did not remap in the first pass, all remapped in the second pass, and we used the second pass location.

Masking false duplications on chromosome 21 of GRCh38. We worked with the Genome Reference Consortium (GRC) to develop a list of regions in GRCh38 that could be masked without changing coordinates or harming variant calling, because they were erroneously duplicated sequences or contaminations. The BED file with these regions can be found at https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_GRC_exclusions.bed. To create the masked reference, we started with the GRCh38 reference with no alternate loci or decoy from https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz.

To generate the masked GRCh38 (i.e., replacing the duplicated and contaminated reference sequence with N's), we used the following Bedtools tools (https://github.com/arq5x/bedtools2) command:

maskFastaFromBed -fi GCA_000001405.15_GRCh38_no_alt_analysis_set.fasta -bed GCA_000001405.15_GRCh38_GRC_exclusions.bed -fo GCA_000001405.15_GRCh38_no_alt_analysis_set_maskedGRC_exclusions.fasta.

To generate the v2 masked GRCh38, we ran the following Bedtools tools (https://github.com/arq5x/bedtools2) command:

“

maskFastaFromBed -fi GCA_000001405.15_GRCh38_no_alt_analysis_set.fasta -bed GCA_000001405.15_GRCh38_GRC_exclusions_T2Tv2.bed -fo GCA_000001405.15_GRCh38_no_alt_analysis_set_maskedGRC_exclusions.fasta.

This uses the bed file GCA_000001405.15_GRCh38_GRC_exclusionsv2.bed generated by the Telomere-to -Telomere Consortium variants team to mask false duplications located under https://trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/references, which also contains the new masked references and other references used in this work.

Evaluation of GRCh38 masked genome improvement. For short reads, a common whole genome resequencing analysis pipeline was used to produce variant call files for the HG002 sample in VCF and gVCF formats. The applications and parameters used in the analysis pipeline were derived from best practices for Illumina short-read whole-genome sequencing resequencing analysis developed for the Centers for Common Disease Genomics project[51]. The analysis pipeline consists of the following high-level steps: sequence alignment to reference genome using BWA-MEM, duplicate read marking using Picard Tools MarkDuplicates, base quality score recalibration using GATK BaseRecalibrator and variant calling using GATK HaplotypeCaller.

This analysis pipeline was run twice on a set of paired-end HG002 FASTQs with 35× coverage as input, with the pipeline runs differing only by reference genome used during the alignment step. The first run used a version of the GRCh38 reference genome prepared without decoy or alternate haplotype contigs. The second run used a version of the GRCh38 reference genome identical to that used in the first run, except that five regions in chromosome 21 and the entire contig chrUn_KI270752v1 were masked with N's, as described above.

The commands executed by the analysis pipeline runs are in Supplementary Data 8 and Data 9, which correspond to the runs using the unmasked and masked GRCh38 reference genomes, respectively.

The following versions for applications and resources were used in the analysis pipeline: BWA v0.7.15, GATK v3.6, Java v1.8.0_74 (OpenJDK), Picard Tools v2.6.0, Sambamba[63] v0.6.7, Samblaster[64] v0.1.24, Samtools v1.9, dbSNP Build 138 on GRCh38 and Known INDELs from Mills and 1000 Genomes Project on GRCh38.

For PacBio long reads, we used a 35× 15-kb + 20-kb HiFi dataset from the precisionFDA Truth Challenge v2 (ref. [12]) aligned to the standard and masked GRCh38 reference with pbmm, and called variants with DeepVariant v1.0 (ref. [65]).

HG002 haplotype-resolved assembly annotation. Liftoff[32] v1.4.0 was used with default parameters to lift over Ensembl v100 annotations from GRCh38

onto each haplotype assembly separately. The resulting GFF files are available at https://trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/CMRG_v1.00/hifiasm-assembly/.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The PacBio HiFi reads used to generate the hifiasm assembly for the benchmark are in the NCBI Sequence Read Archive with accession numbers SRR10382245, SRR10382244, SRR10382249, SRR10382248, SRR10382247 and SRR10382246. The v1.00 benchmark VCF and BED files, as well as Liftoff gene annotations, assembly–assembly alignments and variant calls, are available at https://trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/CMRG_v1.00/, and as a DOI at https://doi.org/10.18434/mds2-2475. This is released as a separate benchmark from v4.2.1, because it includes a small fraction of the genome, it has different characteristics from the mapping-based v4.2.1 and v4.2.1 only includes small variants. Using v4.2.1 and the CMRG benchmarks as two separate benchmarks enables users to obtain broader performance metrics for most of the genome and for a small set of particularly challenging genes, respectively. The masked GRCh38 reference, recently updated to v2 with additional false duplications from the Telomere-to-Telomere Consortium, is under https://trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/references. We recommend using v3.0 GA4GH/GIAB stratification bed files intended for use with hap.py when benchmarking, which are available at https://trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/. These stratifications include bed files corresponding to false duplications and collapsed duplications in GRCh38. All data have no restrictions, as the HG002 sample has an open consent from the Personal Genome Project.

Code availability

Scripts used to develop the CMRG benchmark and generate figures and tables for the manuscript are available at https://github.com/usnistgov/cmrg-benchmarkset-manuscript. The previously developed assembly, which was used as the basis of this benchmark, was from hifiasm v0.11. A variety of open source software was used for variant calling for the evaluations of the benchmark, including NextDenovo2.2-beta.0, DRAGEN 3.6.3, NeuSomatic's submission for the PrecisionFDA truth challenge v2 (ref. [12]) (BWA-MEM[50] version 0.7.17-r1188 (https://github.com/lh3/bwa) and GATK version gatk-4.1.4.1 (https://gatk.broadinstitute.org/hc/en-us)), Parabricks_DeepVariant (Parabricks Pipelines DeepVariant v3.0.0_2 (https://developer.nvidia.com/clara-parabricks)), Sentieon (DNAscope) version sentieon_release_201911 (https://www.sentieon.com/products/#dnaseq), BWA-MEM and Strelka2 (BWA-MEM version 0.7.17-r1188 (https://github.com/lh3/bwa) and Strelka2 version 2.9.10 (https://github.com/Illumina/strelka)), BWA-MEM[50](v0.7.8), Picard tools (https://broadinstitute.github.io/picard/) (ver. 1.83), GATK[52] (v3.4-0), GATK (v3.5), BWA-MEM v0.7.15-r1140, SAMtools[53] v1.3, Picard v2.10.10, GATK v3.8, DELLY[54] v0.8.5, GRIDSS[55] v2.9.4, LUMPY[56] v0.3.1, Manta[57] v1.6.0, Wham[58] v1.7.0, NanoPlot[60] v1.27.0, Filtlong v0.2.0, minimap2 (refs. [40,60]) v2.17-r941, cuteSV v1.0.8, Sniffles[61] v1.0.12, SURVIVOR[59] v1.0.7, BWA v0.7.15, GATK v3.6, Java v1.8.0_74 (OpenJDK), Picard Tools v2.6.0, Sambamba[63] v0.6.7, Samblaster[64] v0.1.24, Samtools v1.9, DeepVariant v1.0 and Liftoff32 v1.4.0.

References

39. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinform.* **26**, 841–842 (2010).
40. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinform.* **34**, 3094–3100 (2018).
41. Krusche, P. et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).
42. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* (O'Reilly Media, 2020).
43. Farek, J. et al. xAtlas: scalable small variant calling across heterogeneous next-generation sequencing experiments. Preprint at *bioRxiv* https://doi.org/10.1101/295071 (2018).
44. Edge, P. & Bansal, V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat. Commun.* **10**, 4660 (2019).
45. Shafin, K. et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Meth.* **18**, 1322–1332 (2021).
46. Sahraeian, S. M. E. et al. Deep convolutional neural networks for accurate somatic mutation detection. *Nat. Commun.* **10**, 1041 (2019).
47. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
48. Patterson, M. et al. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *J. Comput. Biol.* **6**, 498–509 (2015).
49. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).
50. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
51. Regier, A. A. et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects.*Nat. Commun.* **9**, 4038 (2018).
52. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at *bioRxiv* https://doi.org/10.1101/201178 (2018).
53. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinform.* **25**, 2078–2079 (2009).
54. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinform.* **28**, 333–339 (2012).
55. Cameron, D. L. et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res.* **27**, 2050–2060 (2017).
56. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
57. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinform.* **32**, 1220–1222 (2016).
58. Kronenberg, Z. N. et al. Wham: identifying structural variants of biological consequence. *PLoS Comput. Biol.* **11**, e1004572 (2015).
59. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
60. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinform.* **34**, 2666–2669 (2018).
61. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
62. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
63. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinform.* **31**, 2032–2034 (2015).
64. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinform.* **30**, 2503–2505 (2014).
65. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).

Acknowledgements

We thank the Genome Reference Consortium for their curation efforts of GRCh37 and GRCh38 (https://www.genomereference.org), especially V.A. Schneider and P.A. Kitts from the National Institutes of Health (NIH)/NCBI for developing the falsely duplicated regions that should be masked in GRCh38. We thank S. Miller at NIST for helping make available benchmark sets and READMEs. 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 NIST, nor does it imply that the equipment, instruments or materials identified are necessarily the best available for the purpose. C.F. was funded by Instituto de Salud Carlos III (PI20/00876) and Ministerio de Ciencia e Innovación (RTC-2017-6471-1; AEI/FEDER, UE), cofinanced by the European Regional Development Fund 'A Way of Making Europe' from the European Union, and Cabildo Insular de Tenerife (CGIEU0000219140). J.M.L.-S. was funded by Consejería de Educación-Gobierno de Canarias and Cabildo Insular de Tenerife (BOC 163, 24/08/2017). F.J.S. and M.M. was supported by the NIH (UM1 HG008898). C.X. was supported by the Intramural Research Program of the National Library of Medicine, NIH. K.H.M. was supported by the NIH/National Human Genome Research Institute (R01 1R01HG011274-01 and U01 1U01HG010971). H.L. was supported by the NIH (R01 HG010040 and U01 HG010961). C.E.M. thanks funding from the WorldQuant Foundation, NASA (NNX14AH50G), the National Institutes of Health (R01MH117406, R01CA249054, R01AI151059, P01CA214274) and the Leukemia and Lymphoma Society (LLS) (MCL7001-18, LLS 9238-16, LLS-MCL7001-18).

Author contributions

Conceptualization: J.W., N.D.O., A.F., K.H.M., S.E.L., M.T.W.E., H.L., C.-S.C., J.M.Z. and F.J.S. Data curation: J.W., N.D.O. and J.M. Formal analysis – benchmark: J.W., N.D.O., J.M. and J.M.Z. Formal analysis – assembly: H.C., A.S., H.L. and C.-S.C. Methodology: J.W., H.C., H.L., C.-S.C., J.M.Z. and F.J.S. Project administration: J.W., J.M.Z. and F.J.S. Resources: C.X. Software: J.W. and N.D.O. Supervision: C.-S.C., J.M.Z. and F.J.S. Validation: J.W., N.D.O., L.H., J.M., H.C., A.F., Y.-C.H., R.G., A.M.W., W.J.R., Z.M.K., J.F., Y.Z., A.P., M.M., C.X., B.Y., S.M.E.S., D.J., J.M.L.-S., A.M.-B., L.A.R.-R., C.F., G.N., U.S.E., S.E.C., J.L., H.L., C.-S.C., J.M.Z. and F.J.S. Visualization: J.W., N.D.O., H.C., H.L. and C.-S.C. Writing – original draft: J.W., L.H., C.-S.C., J.M.Z. and F.J.S. Writing – review and editing: J.W., N.D.O., D.E.M., J.L., C.E.M., S.E.L., M.T.W.E., C.-S.C., J.M.Z. and F.J.S.

Competing interests

A.M.W. and W.J.R. are employees and shareholders of Pacific Biosciences. A.F., Y.-C.H, R.G., and C.-S.C. are employees and shareholders of DNAnexus. S.M.E.S. is an employee of Roche. J.L. is a former employee and shareholder of Bionano Genomics. S.E.L. was

an employee of Invitae. F.J.S. has sponsored travel from Pacific Biosciences and Oxford Nanopore Technologies. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-01158-1.

Correspondence and requests for materials should be addressed to Chen-Shan Chin, Justin M. Zook or Fritz J. Sedlazeck.

Peer review information *Nature Biotechnology* thanks Adam Ameur, Christian Marshall and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature research

Corresponding author(s): Chen-Shan Chin, Justin M Zook, Fritz J Sedlazeck

Last updated by author(s): 10/28/2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used
Data analysis	Scripts used to develop the CMRG benchmark and generate figures and tables for the manuscript are being made available at https://github.com/usnistgov/cmrg-benchmarkset-manuscript. The assembly previously developed, which was used as the basis of this benchmark, was from hifiasm v0.11. A variety of open source software was used for variant calling for the evaluations of the benchmark: NextDenovo2.2-beta.0, DRAGEN 3.6.3, NeuSomatic's submission for the PrecisionFDA truth challenge v212, [BWA-MEM48 version 0.7.17-r1188] (https://github.com/lh3/bwa), [GATK version gatk-4.1.4.1] (https://gatk.broadinstitute.org/hc/en-us)); Parabricks_DeepVariant ([Parabricks Pipelines DeepVariant v3.0.0_2] (https://developer.nvidia.com/clara-parabricks)); Sentieon (DNAscope) version sentieon_release_201911 (https://www.sentieon.com/products/#dnaseq)); BWA-MEM+Strelka2 ([BWA-MEM version 0.7.17-r1188](https://github.com/lh3/bwa) + [Strelka2 version 2.9.10](https://github.com/Illumina/strelka), BWA-Mem 48(ver. 0.7.8), Picard tools (https://broadinstitute.github.io/picard/) (ver. 1.83), GATK50 (ver. 3.4-0), GATK (ver. 3.5), BWA-MEM v0.7.15-r1140, SAMtools51 v1.3, Picard v2.10.10, GATK v3.8, DELLY52 v0.8.5, GRIDSS53 v2.9.4, LUMPY54 v0.3.1, Manta55 v1.6.0, and Wham56 v1.7.0, NanoPlot58 v1.27.0, Filtlong v0.2.0, minimap238,58 v2.17-r941, cuteSV v1.0.8, Sniffles59 v1.0.12, SURVIVOR 57 v1.0.7, BWA v0.7.15, GATK v3.6, Java v1.8.0_74 (OpenJDK), Picard Tools v2.6.0, Sambamba61 v0.6.7, Samblaster62 v0.1.24, Samtools v1.9, DeepVariant v1.0, Liftoff32 v1.4.0.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The PacBio HiFi reads used to generate the hifiasm assembly for the benchmark are in the NCBI SRA with accessions SRR10382245, SRR10382244, SRR10382249, SRR10382248, SRR10382247, and SRR10382246. The v1.00 benchmark VCF and BED files, as well as Liftoff gene annotations, assembly-assembly alignments, and variant calls, are available at https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/CMRG_v1.00/, and as a DOI at https://doi.org/10.18434/mds2-2475. This is released as a separate benchmark from v4.2.1 because it includes a small fraction of the genome, it has different characteristics from the mapping-based v4.2.1, and v4.2.1 only includes small variants. Using v4.2.1 and the CMRG benchmarks as two separate benchmarks enables users to obtain broader performance metrics for most of the genome and for a small set of particularly challenging genes, respectively. The masked GRCh38 reference, recently updated to version 2 with additional false duplications from the Telomere to Telomere Consortium, is under https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/references. We recommend using v3.0 GA4GH/GIAB stratification bed files intended for use with hap.py when benchmarking, available under https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/. These stratifications include bed files corresponding to false duplications and collapsed duplications in GRCh38. All data have no restrictions, as the HG002 sample has an open consent from the Personal Genome Project.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This benchmark was created for a single sample because only one sample with all the data needed was available
Data exclusions	No data were excluded
Replication	Replication was not needed here because the goal of the study was to establish a benchmark, and this benchmark was evaluated by 11 independent methods and curators
Randomization	Randomization was not relevant since this was a single sample benchmark.
Blinding	Blinding was not relevant or possible because this is a widely used benchmark sample, and all Genome in a Bottle data are open and public as soon as possible

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	For the 10× Genomics and Oxford Nanopore sequencing and Bionano mapping, the GM24385 (RRID:CVCL_1C78) cell line was

Cell line source(s)	obtained from the Coriell Institute for Medical Research National Institute for General Medical Sciences cell line repository. For the Illumina and Pacific Biosciences sequencing, NIST RM 8391 DNA was used, which was prepared from a large batch of GM24385 to control for differences arising during cell growth. For binning reads into paternal and maternal haplotypes, Illumina sequencing of DNA from NIST RM 8392 (HG002-HG004) was used. DNA was extracted from cell lines publicly available as GM24149 (RRID:CVCL_1C54) and GM24143 (RRID:CVCL_1C48) at the Coriell Institute for Medical Research National Institute for General Medical Sciences cell line repository.
Authentication	Variants from whole genome sequencing were compared to previous variant calls from this cell line
Mycoplasma contamination	Cell lines tested negative for mycoplasma by Coriell prior to aliquotting NIST Reference Material 8391
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used in this study