Resolving missing human polymorphic inversions and other complex variants from ultralong read data

Ricardo Moreira-Pinhal, Konstantinos Karakostis, Illya Yakymenko, et al.

Genome Res. published online September 1, 2026
Access the most recent version at doi:10.1101/gr.280867.125

Method

Resolving missing human polymorphic inversions and other complex variants from ultralong read data

Ricardo Moreira-Pinhal,[1,2] Konstantinos Karakostis,[1,2,6] Illya Yakymenko,[1,2] Oscar Conchillo,[1,3] Maria Díaz-Ros,[1,2] Andrés Santos,[3,7,8] Miquel Àngel Senar,[4] Jaime Martínez-Urtaza,[3] Marta Puig,[1,3] and Mario Cáceres[1,2,5]

[1]*Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra 08193, Barcelona, Spain;* [2]*Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Research Institute, Barcelona 08003, Spain;* [3]*Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra 08193, Barcelona, Spain;* [4]*Àrea d'Arquitectura i de Tecnologia de Computadors, Universitat Autònoma de Barcelona, Bellaterra 08193, Barcelona, Spain;* [5]*ICREA, Barcelona 08010, Spain*

Inversions are a unique type of balanced structural variants (SVs) with important consequences in multiple organisms. However, despite considerable effort, these and other complex SVs remain poorly characterized because of the presence of large repeats. New techniques are finally allowing us to identify the full spectrum of human inversions, but the number of individuals analyzed is still quite limited. Here, we take advantage of Oxford Nanopore Technologies (ONT) long reads to characterize an exhaustive catalog of 612 candidate inversions between 197 bp and 4.4 Mb of length, and flanked by <190 kb long inverted repeats (IRs). To that end, we have developed a bioinformatic package to identify inversion alleles reliably from long-read data. Next, using a combination of different DNA extraction, library preparation, and ONT sequencing protocols, we show that ultralong reads (50–100 kb) and adaptive sampling are an efficient method to detect most human inversions. Lastly, by analyzing ONT data from 54 diverse individuals, 87%–99% of the inversions can be genotyped in each sample, depending mainly on read and IR length and genome coverage. Both orientations have been observed for 155 of the analyzed regions (frequency 0.01–0.49), which triples the number of polymorphic IR-mediated inversions studied in detail so far. Moreover, we have found more than 300 additional independent SVs in the studied regions and resolved several complex rearrangements. Therefore, our work provides an accurate benchmark of those inversions that typically escape most analyses, and it demonstrates the potential of nanopore sequencing to characterize missing human genomic variation.

[Supplemental material is available for this article.]

Over the past few decades, there has been a global effort to identify all human genomic variation and its association with phenotypic traits and disease susceptibility in different populations (Sudmant et al. 2015; The 1000 Genomes Project Consortium 2015; Choudhury et al. 2020; Collins et al. 2020; Karczewski et al. 2020; Ebert et al. 2021; Byrska-Bishop et al. 2022; Choi et al. 2023). However, structural variants (SVs), which involve changes of >50 bp, are often generated through nonallelic homologous recombination (NAHR) between large repeats and occur in complex regions harboring many segmental duplications (SDs), making their detection very difficult by most techniques (Alkan et al. 2011; Logsdon et al. 2020; Ebert et al. 2021).

This is especially problematic for inversions, a big fraction of which are still understudied owing to their balanced nature, changing only the sequence order (usually with no gain or loss of DNA), and the presence of highly identical inverted repeats (IRs) at their breakpoints (Puig et al. 2015; Ebert et al. 2021). In addition, inversions are a special type of SV, as they inhibit recombination between the two orientations, leading to potential negative consequences on fertility (Puig et al. 2015). Accordingly, those that reach a certain frequency in natural populations could be more likely to have positive functional effects that compensate for these costs. In fact, a growing number of inversions have been involved in adaptation and phenotypic variability in diverse organisms from plants to birds (Wellenreuther and Bernatchez 2018), including humans (Campoy et al. 2022). Moreover, around one-third of human inversions are associated with gene expression or epigenetic changes, and they tend to have larger effects than other variants (Giner-Delgado et al. 2019; Puig et al. 2020; Lerga-Jaso et al. 2026).

Large-scale projects using a combination of novel genomic techniques, such as long-read sequencing, Strand-seq, or Bionano optical maps (Audano et al. 2019; Chaisson et al. 2019; Levy-Sakin et al. 2019; Ebert et al. 2021; Porubsky et al. 2022) are finally making it possible to identify the full spectrum of human inversions. However, each technique has its own limitations and error sources. Also, in most cases, just a reduced number of individuals has been studied, which hinders the analysis of the effects of the detected variants and their association with phenotypic traits. To overcome these problems, different methods have been

Present addresses: [6]Department of Biochemistry and Molecular Biology, University of Valencia, Burjassot 46100, València, Spain; [7]Departamento de Ciencias Básicas, Facultad de Medicina, Universidad de la Frontera, Temuco 4780000, Chile; [8]Laboratorio de Bioinformática y Microbiología Aplicada, Centro de Excelencia en Medicina Traslacional, Universidad de la Frontera, Temuco 4780000, Chile

Corresponding authors: marta.puig@uab.cat, mcaceres@icrea.cat

Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.280867.125. Freely available online through the *Genome Research* Open Access option.

developed to genotype specific inversions, either individually (e.g., regular PCR, inverse PCR [iPCR] or droplet digital PCR [ddPCR]) or in sets (e.g., multiplex ligation-dependent probe amplification [MLPA] or inverse MLPA [iMLPA]) (Giner-Delgado et al. 2019; Puig et al. 2020). This has allowed a detailed analysis of up to 134 inversions in 95–551 individuals from diverse human populations, including 55 inversions with IRs at their breakpoints (Lerga-Jaso et al. 2026). Nevertheless, the variants that can be studied with each method are extremely dependent on inversion and IR characteristics, such as their length or the presence of restriction enzyme target sites, which means that a single technique cannot detect all inversions. In addition, inversions generated by NAHR tend to appear recurrently and to not be linked to nearby SNPs or particular haplotypes (Giner-Delgado et al. 2019; Puig et al. 2020; Porubsky et al. 2022). As a result, many of them cannot be imputed reliably using existing data sets, and their effects have not been detected in common genome-wide association studies (GWAS). Thus, new high-throughput strategies that determine directly the genotype of multiple individuals by investigating the actual DNA sequence order (in contrast to indirect methods, like imputation, that can have a high error rate) are needed.

Long-read sequencing methods, such as those of Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), represent a unique opportunity for the complete characterization of complex genomic variants (Logsdon et al. 2020; De Coster et al. 2021; Mastrorosa et al. 2023; Grochowski et al. 2024). Specifically, ONT sequencing as DNA molecules pass through multiple nanopores can generate extremely long reads of up to several megabases in a relatively short time (Logsdon et al. 2020). Despite its higher error rate, ONT ultralong reads (>50–100 kb) have been instrumental in the generation of the new telomere-to-telomere (T2T) and draft human pangenome references, helping to resolve centromeres and other hard-to-characterize regions (Nurk et al. 2022; Liao et al. 2023). Furthermore, adaptive sampling (AS) promotes the sequencing of targeted molecules in real-time (Payne et al. 2021), enriching regions of interest in a simple and cost-effective way. This has already been applied to identify rare genomic rearrangements and to genotype difficult-to-study variants, such as tandem repeats (Miller et al. 2021; Stevanovski et al. 2022). Also, some studies are starting to generate long-read resources of multiple individuals of diverse populations and diseases (Beyter et al. 2021; De Coster et al. 2021; Wu et al. 2021; Gustafson et al. 2024; Schloissnig et al. 2025), although in most cases, read length is moderate. Therefore, it is important to develop optimized protocols and bioinformatic tools that can be widely adopted to extract accurate information of different types of genomic variants from these new data and help define current technology limits.

In this work, we used long-read sequences to generate an accurate catalog of human polymorphic inversions mediated by a wide range of IR lengths and to obtain an initial estimate of their frequency. Moreover, we investigated the applicability of ONT ultralong reads for efficient genotyping of inversions as a first step to characterize their functional impact, and we checked how well these variants are identified in other studies and the draft human pangenome reference.

Results

Data set of IR-mediated inversions

To build the most exhaustive data set of IR-mediated human inversions to date, candidate inversion regions were selected from four different sources: (1) 58 independently validated inversions with different types of breakpoint IRs (ranging from 140 bp to 134 kb), including the 55 genotyped in multiple individuals using different PCR-based techniques (Giner-Delgado et al. 2019; Puig et al. 2020; Lerga-Jaso et al. 2026); (2) the merging of 1528 candidate inversions from five recent studies that identify SVs using long-read sequencing, Strand-seq, or optical mapping in a total of 195 individuals (Audano et al. 2019; Chaisson et al. 2019; Levy-Sakin et al. 2019; Ebert et al. 2021; Porubsky et al. 2022); (3) previous paired-end mapping (PEM) predictions that could not be analyzed with available methods or that were possible human genome assembly errors caused by highly identical IRs (Korbel et al. 2007; Martínez-Fundichely et al. 2014; Vicente-Salvador et al. 2017); and (4) pairs of inverted SDs annotated in the human reference genome (GRCh38/hg38) that could generate inversions (see Methods). Candidate inversions from the different sources were merged into independent variants when their breakpoints were located within or close to the same pair of IRs. Also, we selected only inversion candidates that had at least one IR copy <200 kb and were not located in very complex regions full of repeats and SDs (which excludes most of the centromeric and telomeric sequences) or gaps in hg38, and for which specific probe sequences could be designed around the breakpoints (see below).

Of the merged list of candidate inversions from recent studies, 622 regions were filtered out because they did not meet the above criteria (mostly because no IRs were identified in the region or the sequence was extremely complex) and 11 more owing to unexpected complex behavior during inversion genotyping. This resulted in 150 possible independent inversions with IRs that could be tested (including 65 identified by at least three independent studies with breakpoints within or close to the same IR pair), not counting those already known (Lerga-Jaso et al. 2026). In addition, there were 85 PEM inversion predictions overlapping IRs (Martínez-Fundichely et al. 2014) that were not present in the previous set. Finally, we selected an additional group of 319 pairs of inverted SDs 1–50 kb long and separated by <250 kb that do not correspond to any of the other inversions. In total, 612 candidate inversion regions ranging from 197 bp to 4.4 Mb and flanked by IRs between 140 bp and 190 kb were included in the analysis (Supplemental Table S1). However, the inversion selection could exclude some cases in which the IRs are not found in the current hg38 human reference genome assembly.

Development of a new inversion genotyping package for long reads

Thanks to the sequence contiguity, ONT ultralong reads offer the opportunity to analyze a wide range of inversions and resolve other challenging variants without being affected by sequencing error rates or other methods limitations (Giner-Delgado et al. 2019; Puig et al. 2020). Thus, we developed a novel inversion genotyping software package called GeONTIpe, which is based on the detection of reads spanning inversion breakpoints and determining their sequence order and orientation by searching for specific probe sequences at either side of each IR (A, B, C, and D) (Fig. 1A). That way, we can identify reads with A–B/C–D combinations that support the reference orientation (O1) or A–C/B–D combinations that support the inverted orientation (O2) (Fig. 1B).

The analysis has several steps: quality filtering of the reads, mapping against hg38, generation of inversion genotypes, and identification of additional SVs (see Methods) (Supplemental Fig. S1). One of the key steps is the design of the probe sequences of

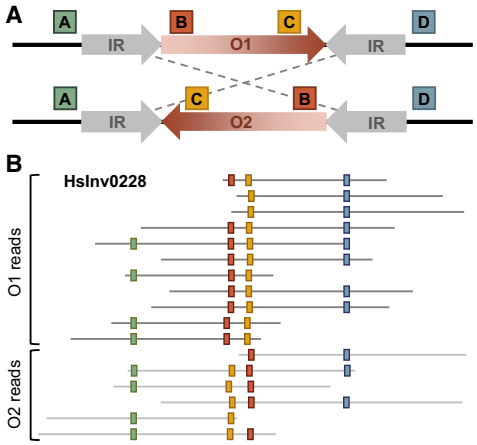

Figure 1. Inversion genotyping from ONT long reads. (*A*) Schematic representation of candidate inversions analyzed in this work, with the inverted region shown as a red gradient flanked by two inverted repeats (IRs) in gray. Probe sequences located at both sides of each IR (labeled as A, B, C, and D) are indicated as colored rectangles and have different order in O1 and O2 orientations. (*B*) Example of HsInv0228 inversion genotyping results in a heterozygote sample, showing ONT individual reads as dark gray (O1) or light gray (O2) lines. The probe sequences are indicated with colored rectangles. Probe sequence order together with read number and proportion of reads supporting each orientation determines the inversion genotype.

a few hundred base pairs to identify unique regions flanking the breakpoint IRs, which were manually revised to avoid specificity problems and errors caused by the presence of unexpected haplotypes. This allowed us to interrogate accurately 612 candidate inversions, including 40 analyzed only using one breakpoint owing mainly to the presence of large IRs or the lack of specific probe sequences. Once these sequences were mapped to the reads, the inversion genotype was determined according to the number of reads supporting each orientation as homozygote, heterozygote, or low confident (see Methods). Moreover, we used the distance between the probe sequences to identify additional insertions (increased distance) or deletions (reduced distance) in the analyzed regions. Therefore, the developed method makes it possible to resolve a large majority of human inversions mediated by a wide range of IRs for the first time while it also characterizes additional SVs.

Benchmark of ONT sequencing strategies for inversion genotyping

In parallel, to assess the most efficient methodology for genotyping the whole set of human inversions, we tested several ONT ultralong-read sequencing protocols using a MinION device. The analysis was multidimensional and included two main groups of experiments on four human female samples of diverse origins for which most of the PCR-validated inversions had been previously analyzed (Fig. 2A; Supplemental Table S2; Lerga-Jaso et al. 2026). First, we used the Monarch HMW DNA extraction method (NEB) to compare the genotyping efficiency of whole-genome sequencing (WGS) and AS of two sets of inversions: an initial set, including 132 candidate inversions plus 40 kb flanking regions (AS2), and an extended one, targeting additional predicted inversions and inverted SDs plus 70 kb flanking regions (AS3), which covers almost all of the 612 inversion set (Supplemental Table S1). Second, we evaluated the effect on read length and total output of an alternative phenol-based genomic DNA purification protocol using AS3.

In the different sequencing experiments with the Monarch-extracted DNA, we obtained an N50 of 32–79 kb, with consistent values when the same library was sequenced and no significant differences between WGS and AS (Fig. 2A; Supplemental Table S2). However, the phenol-based DNA extraction method generated significantly larger N50, ranging from 79 to 118 kb (Fig. 2A). Also, long-term storage of the cells could reduce genomic DNA integrity and read length, with phenol-isolated DNA from fresh cells having higher N50 values than that from cells stored at −80°C for a year (mean of 117.7 vs. 87.9 kb, respectively), whereas Monarch-extracted DNA from NA12156 cells kept for 8 years at −80°C showed clearly the lowest average N50 (50.6 kb) (Fig. 2A). In terms of output, both Monarch AS2 and AS3 experiments yielded an

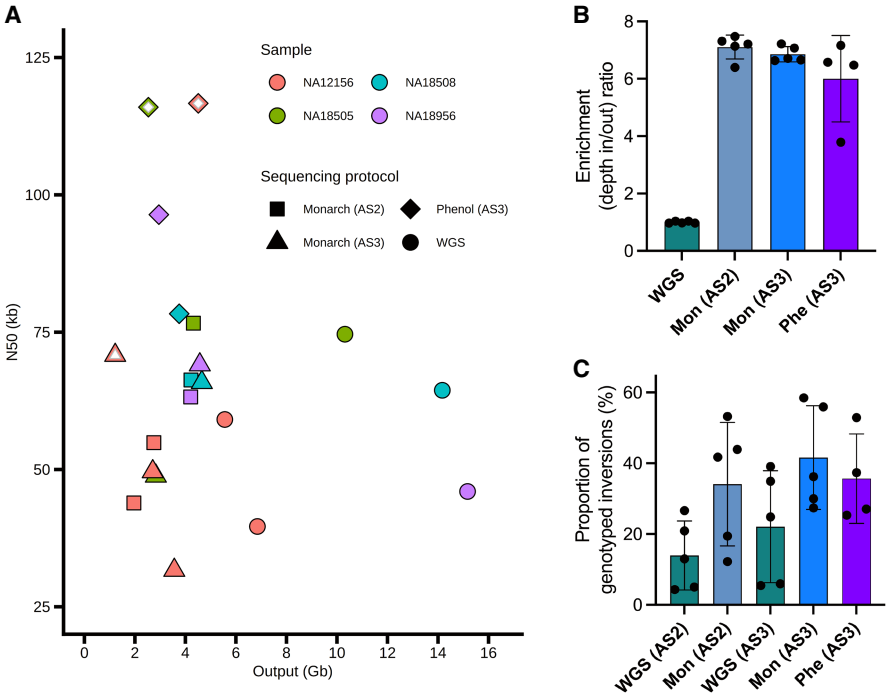

Figure 2. Summary of ONT sequencing experiment results. (*A*) Comparison of the N50 value and the total output of each sequencing run of the four analyzed samples (in different colors). Diverse sequencing protocols using whole-genome sequencing (WGS) or adaptive sampling (AS2/AS3) and DNA isolation methods (Monarch HMW or phenol) are represented by different shapes, whereas empty symbols indicate the use of fresh cells as starting material. (*B*) Enrichment ratio between the read depth inside and outside the targeted regions (depth in/out) of each sample for diverse types of sequencing (black dots), showing a six- to sevenfold average enrichment in AS2 or AS3 compared with WGS (colored bars). (*C*) Average proportion of genotypes obtained for the analyzed inversion regions normalized to the number of AS2 or AS3 targets (percentage, colored bars). Values for each sample (black dots) show considerable variation due mainly to sequencing output differences.

average of 3.3 Gb and 3.1 Gb compared with 9.5 Gb for WGS, which corresponds to a ratio of 2.4, 2.0 and 5.5 Mb/nanopore, respectively (Fig. 2A; Supplemental Table S2). Thus, our results suggest that read length depends mainly on sample preparation, whereas the amount of sequence data is reduced by about threefold in AS compared with WGS.

Regarding the efficiency and specificity of AS sequencing, we observed a similar average increase on the coverage of the targeted inversion regions for the Monarch AS2 (7.1-fold) and AS3 (6.9-fold) results, which is quite consistent across samples (Fig. 2B). Dividing the genome into 10 kb windows showed a clearly different distribution of the average fold enrichment between targeted and nontargeted windows (Supplemental Figs. S2, S3A) and a significant correlation of the mean enrichment of the different target regions common in AS2 and AS3 ($r = 0.35$, $P < 0.001$) (Supplemental Fig. S3B). Still, in both AS data sets a small proportion of nontargeted windows display increased coverage values comparable to that of targeted windows (Supplemental Figs. S2, S3A), which could be caused by a significant excess of repetitive elements and/or SDs in these regions and the homology between AS targets and repeated sequences (Supplemental Fig. S3C).

When considering inversion genotyping, Monarch AS2 and AS3 sequencing experiments provide, respectively, an average of 1.9 and 1.7 times more informative reads for inversion genotyping compared with WGS (Supplemental Table S2), which results from a trade-off between the higher enrichment and the lower output of AS. Similarly, depending on the run results, up to 65% of the targeted inversions could be successfully genotyped in a single MinION flow cell, and this value is 1.9–2.5 times higher in AS2 (mean = 36.9%) or AS3 (mean = 46.0%) than in WGS (mean = 15.0% and 24.3% for AS2 and AS3 targets, respectively) (Fig. 2C; Supplemental Table S2). Even though phenol-extracted DNA generated significantly longer reads, with N50 >100 kb from fresh cells, this did not result in better inversion genotyping (mean = 39.4%), probably because the longer molecules yielded less total sequence reads (Fig. 1A,C). When the data of the different sequencing experiments of the same sample were merged, 87%–91% of the inversions could be genotyped in each individual.

Large-scale inversion genotyping in humans

To obtain a good picture of human polymorphic inversions mediated by IRs, we applied GeONTIpe to 54 diverse individuals with publicly available ONT data, including those sequenced in this study (Supplemental Table S3). The analysis showed a high genotyping success (>87%) of the candidate inversion regions in all the samples (Fig. 3A). As expected, the sample genotyping rate was correlated with the sequencing depth and especially N50, with 99% of the inversions being genotyped in the samples with N50 >100 kb (Fig. 3B). Similarly, the distance between probe sequences at both sides of the breakpoints reduced inversion genotyping success owing to a lower number of reads crossing the IRs (Fig. 3C). In fact, inversions with probe sequences separated by >100 kb could be genotyped on average in only 47.0% of the samples.

We detected the two orientations in at least one of the 54 individuals (not counting the hg38 reference genome to avoid possible assembly errors) in 155 of the 612 inversion regions tested (25.3%). The minor allele frequency (MAF) of the detected inversions ranged from 0.01 to 0.49, with 78.1% of the polymorphic inversions being widespread across populations (Supplemental Tables S1, S3). However, the number of inversions identified as polymorphic varied depending on their origin, comprising 55 of the 58 previously known inversions (94.8%), 83 of the additional candidate inversions merged from five different studies (55.3%), 11 of other PEM predictions that had not been validated before (12.9%), and, as expected, only six of the inverted SD pairs (1.9%) (Fig. 4A). The previously known inversions not found here correspond to two very low-frequency inversions (HsInv0486 and HsInv0626) (Lerga-Jaso et al. 2026) and to HsInv1869, which could be an hg38 assembly error because only inverted alleles are found in all analyzed samples (Supplemental Table S3). In addition, two validated reference genome errors (Vicente-Salvador et al. 2017) turned out to be low-frequency inversions (HsInv0057, HsInv1119).

On the other hand, a similar proportion of real inversions was found in recent SV studies using a multiplatform approach or long reads (75.2%–81.0%), with a slightly lower success for those relying only in Bionano optical maps (66.2%), although it is also the study analyzing a larger number of samples (Fig. 4B). The remaining 67 nonpolymorphic inversion candidates were predominantly detected only in one or two studies (47), which could represent false predictions or low-frequency variants. Those found in three or more studies corresponded mainly to possible hg38 reference genome assembly errors, including 13 that had been already confirmed experimentally (Antonacci et al. 2010; Vicente-Salvador et al. 2017). Also, in a few inversion candidates in complex regions with several breakpoint SD pairs that could led to different inversions, only one of them was validated, or they showed low genotyping success caused by the length of the repeat blocks. With regard to the 37 polymorphic inversions not detected in the comprehensive Porubsky et al. (2022) study, most had low frequency and/or were small (56.8%), 24.3% overlapped other variants classified as inverted duplications or with quite different breakpoints, and in 18.9% there was not any nearby variant, which highlights the ability of our approach to detect previously missing inversions.

Accuracy of ONT inversion genotypes

We performed different comparisons to demonstrate the accuracy of the methodology. First, in the 652 comparisons of the polymorphic inversions in five parent–child trios included in our data set (Supplemental Table S3), only two showed discordant genotypes (99.7% concordance): HsInv0808 in the HG002 Ashkenazi trio child that showed a single O2 read and 31 O1 reads, which was insufficient to call it as heterozygote, and HsInv0278 in the HG00731 Puerto-Rican trio father, with 35 O2 reads, that in previous studies already showed discrepant calls as O2 homozygote (Chaisson et al. 2019) and as heterozygote (Porubsky et al. 2022). In addition, we found concordant results for 452 of the 455 (99.3%) PCR-based genotypes of 54 inversions in 10 individuals included in the present analysis (Fig. 4C; Giner-Delgado et al. 2019; Puig et al. 2020; Lerga-Jaso et al. 2026), and all three discrepancies corresponded apparently to errors in the PCR data. These involved a new inverted duplication identified from long reads instead of the expected inversion (HsInv0397 in NA12156), a complex duplicated region (HsInv0348 in NA18508), or an inversion with a high genotyping error rate (Giner-Delgado et al. 2019) in which the 1000 Genome Project (1KGP) high-coverage data supports also the ONT results) (HsInv0045 in NA20509) (Byrska-Bishop et al. 2022.

Next, we compared our results with those of four previous studies, taking into account the full genotype, if available, or just the presence or absence of the inverted allele (Fig. 4C; Audano et al. 2019; Chaisson et al. 2019; Levy-Sakin et al. 2019; Porubsky et al. 2022). A lower genotype concordance was observed

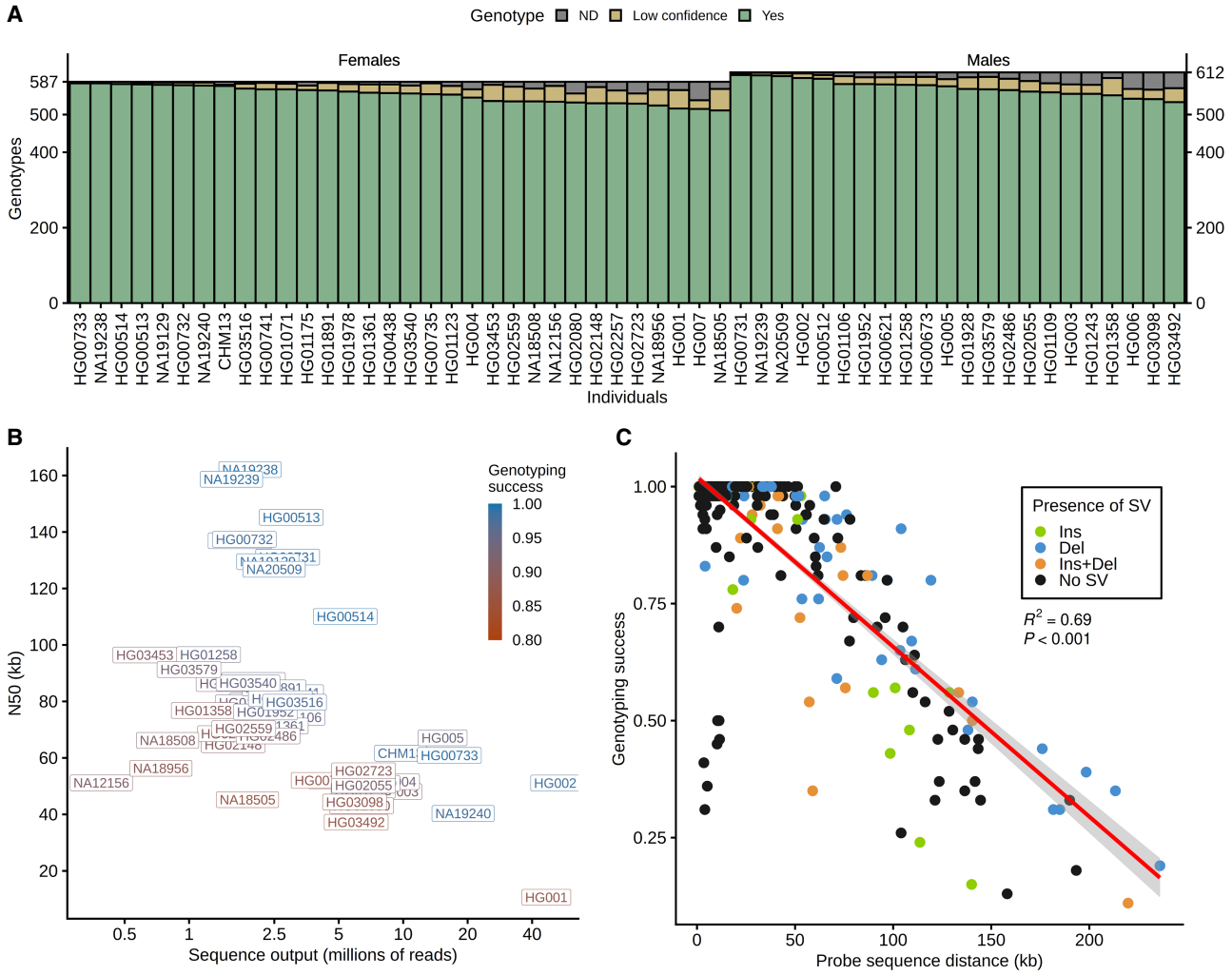

Figure 3. Genotyping success of candidate inversion regions from ONT data in 54 individuals. (*A*) Summary of inversion genotyping results by individual (612 for males and 587 for females, excluding 25 inversions located in Chr Y), with resolved genotypes in green, low-confident genotypes owing to a reduced number of reads in yellow, and those with no supporting reads (not determined [ND]) in gray. (*B*) N50 values and total number of reads for each sample, showing the genotyping success rate on a red to blue scale. (*C*) Genotyping success of candidate inversion regions as a function of the average distance between the probe sequences flanking the breakpoints. Additional SVs detected in the analyzed regions, some of which can affect genotyping success, are indicated in different colors.

in older studies that relied either on one detection technique, such as Bionano optical maps (74.1% from 38 candidate inversions in six individuals) (Levy-Sakin et al. 2019) and PacBio long reads (83.1% from 55 candidate inversions in five individuals) (Audano et al. 2019), or on a multiplatform approach integrating the previous two techniques and Strand-seq (82.4% from 101 candidate inversions in nine individuals) (Chaisson et al. 2019). Conversely, the large multiplatform analysis using the same techniques from Porubsky et al. (2022) showed 95.1% genotype concordance from 146 inversion regions compared in 12 individuals.

Lastly, we checked how well the analyzed inversion regions are represented in the available diploid genomes of 14 males and 19 females from the draft human pangenome reference (Liao et al. 2023). Almost perfect concordance with our results (98.8%) was found for the 16,614 genotypes (84.2%) in common (Fig. 4C). However, most of the remaining genotypes could not be compared mainly owing to the absence of assembled sequences crossing the breakpoints in some inversions and individuals in the draft human pangenome reference (2120, 10.7%). Thus, even though this new resource provided a good representation of polymorphic inversions, it still missed information in a significant fraction of regions or individuals.

Identification of other SVs and characterization of complex regions

After mapping the probe sequences on each read, the comparison of the distances between them with those expected in the hg38 reference genome or its inverted version allowed us to identify additional SVs in the tested regions. In total, we detected 577 deletions and insertions between 250 bp and 470 kb in either O1 or O2 chromosomes (Supplemental Table S4), which included approximately 322 independent SVs: 195 deletions, 89 insertions, and 38 copy number variants (CNVs; for details, see Methods). Of those, there were 243 SVs between the probe sequences flanking one breakpoint; 65 between the two internal probe sequences within the

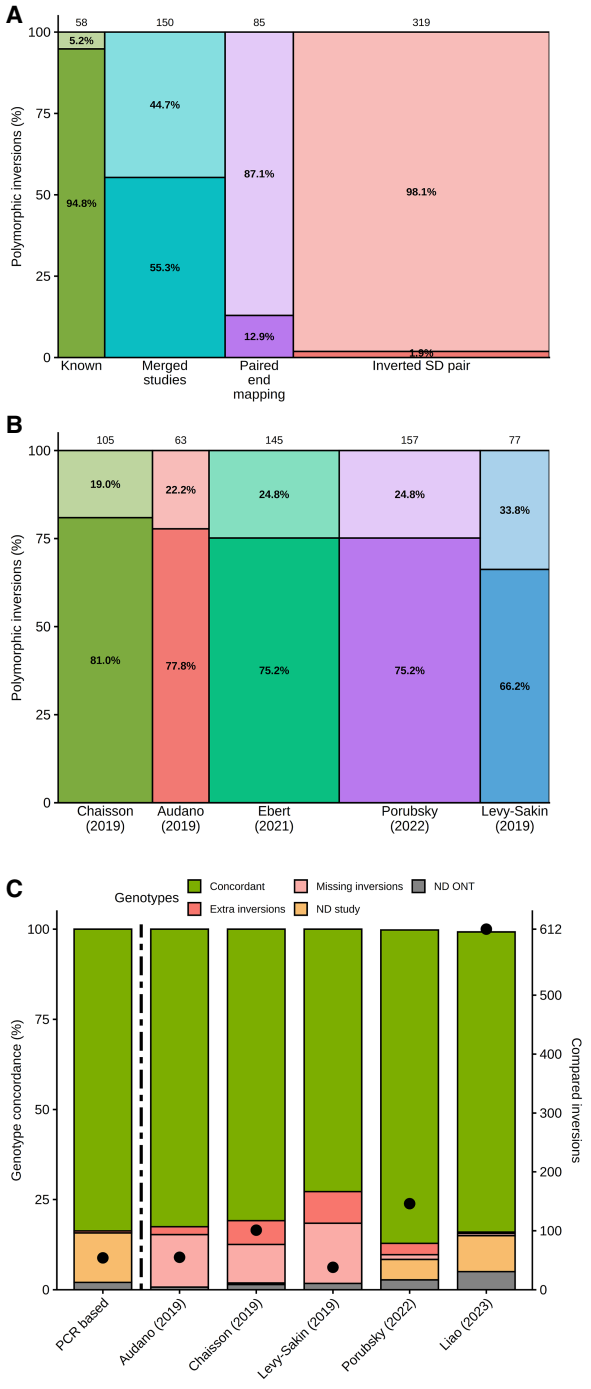

Figure 4. Comparative analysis of inversions identified in this and previous studies. (*A*) Mosaic plot representing the proportion of the 612 analyzed regions derived from different sources (bar width, with the number of regions indicated in the *top* part) and those that were detected in ONT data as polymorphic inversions (dark color) or not (light color). (*B*) Mosaic plot representing the candidate regions from each of the merged SV studies included in our analysis (bar width with the total number indicated in the *top* part) and the percentage detected as polymorphic inversions (dark color) or not (light color). (*C*) Concordance of ONT genotypes with PCR-based experimental data (separated by dashed line) and five other studies analyzing the same samples and inversion regions (*left* *y*-axis). Genotypes are classified as concordant, as having extra or missing inverted alleles, or as not determined (ND) in ONT data or the other study used for comparison. Black dots represent the number of inversions compared in each data set (*right* *y*-axis).

inverted segment, which could be only analyzed for part of the inversions of <50–100 kb; and 14 deletions that remove one or more of the probe sequences. Although the exact association of the additional SVs with inversion alleles cannot be determined owing to the lack of precise SV genotypes and the fact that many are present in both orientations, which complicates knowing where they were generated, regions with polymorphic inversions contained more additional SVs (137 SVs in 76 regions, 49.0%, with an average of 1.8 SVs per region) than those without inversions (185 SVs in 141 regions, 30.9%, with an average of 1.3 SVs per region; Fisher's exact test $P = 6 \times 10^{-5}$). This could be in part a consequence of a higher SD content in the regions with real inversions (58.6% vs. 40.7%), and it is consistent with increased structural diversity associated with inversions (Porubsky et al. 2022). In addition, according to the presence or absence of each SV, their frequency ranged from those found in only one out of 54 analyzed individuals to those found in all of them, corresponding probably to an error or low-frequency haplotype in the reference genome (Supplemental Table S4), with 65.5% of them (211 SVs) detected in several individuals, which indicates that they are relatively common.

For some regions that accumulate several SVs, we combined the GeONTIpe output and manual read analysis to identify multiple structural haplotypes. For example, in HsInv0401 there were nine configurations resulting from an 8.7 kb polymorphic inversion and a 12.7 kb CNV at one or both breakpoints with two to five copies per chromosome (Fig. 5A). We also characterized other types of complex regions, like that of HsInv0822-0816-0384, in which three nested polymorphic inversions can generate up to eight different haplotypes (Porubsky et al. 2022), of which we detected six. Similarly, in the HsInv0012-0659 region, a polymorphic inversion, a CNV with zero to two copies, two insertions, and a deletion generated at least five different haplotypes (Fig. 5B).

Discussion

A detailed characterization of the different types of SVs is needed to determine precisely how many variants there really are in humans and their consequences. Here, by doing a comprehensive analysis of the latest human SV information, we have created an accurate catalog covering a large fraction of human inversions mediated by IRs, which typically escape most analysis. In addition, we have developed an entire bioinformatic and experimental approach to genotype complex inversions not matched by other available techniques. This includes a new easy-to-use bioinformatic tool that generates highly reliable inversion genotypes from different types of long-read data (Schloissnig et al. 2025). Furthermore, by testing different DNA extraction, library preparation, and sequencing protocols, we have shown that the combination of ultralong ONT reads and AS constitutes an efficient method to analyze most human inversions mediated by IRs in a single experiment, which reduces enormously the amount of time and effort required. Finally, ONT long-read data from multiple individuals has allowed us to (1) confirm the existence of 155 polymorphic inversions, including several not previously detected; (2) identify more than 300 additional SVs in the analyzed regions; and (3) resolve several complex regions with different rearrangements generating structurally diverse haplotypes that deserve further study.

In fact, ONT long reads represent a more direct way to identify inversions and characterize complex regions compared with other available techniques, because it just depends on the mapping of long stretches of DNA, which works better than that of shorter

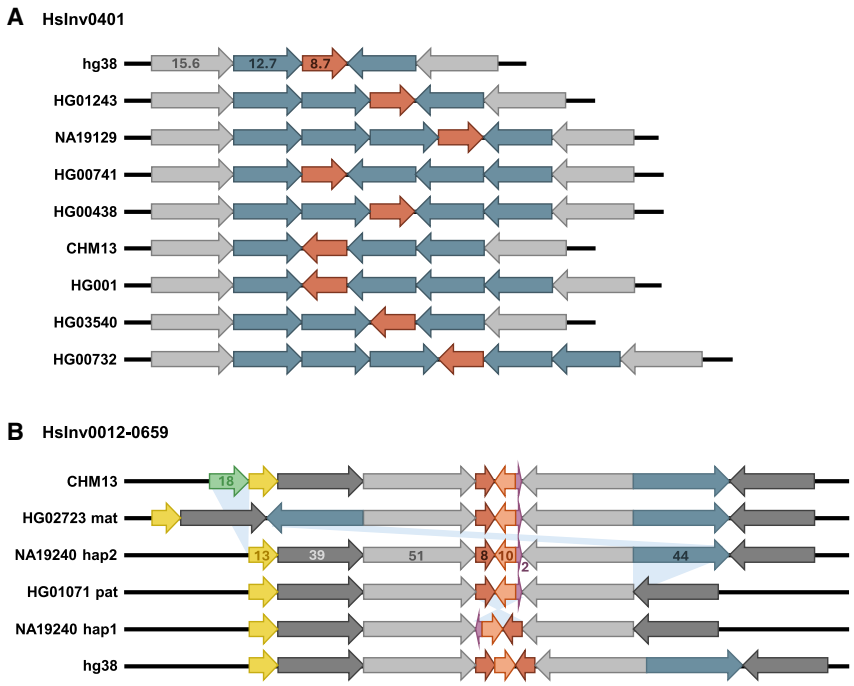

Figure 5. Examples of structural diversity at complex inversion regions resolved with ONT long reads. (*A*) HsInv0401 includes a 12.7 kb CNV with up to five copies per chromosome (blue arrows) located within the IRs at both sides of the inverted sequence (red arrow). (*B*) HsInv0012-0659 includes a polymorphic inversion (red, orange, and purple arrows), a 44 kb duplication and deletion (blue arrows), and an 18 kb insertion (green arrow). In this case, the GRCh38 (hg38) reference genome sequence is not supported by any other assembly we analyzed so far and might represent an assembly error. In both images, gray arrows correspond to nonvariable regions of inverted SDs, and numbers indicate the size of the different segments in kilobases.

sequences, especially for repeats. For example, around one-half of the IR-mediated inversions from recent studies were not detected in our analysis (Fig. 4A). Part of the discrepancies can be because not the same individuals are interrogated, and others correspond likely to assembly errors in the human reference genome (Vicente-Salvador et al. 2017). However, our results suggest that there is still a significant proportion of false predictions, which is illustrated by the six to 60 extra inversion calls in the different studies with respect to the ONT data in the few individuals compared (Fig. 4C). This is especially evident in the case of older PEM inversion predictions (Martínez-Fundichely et al. 2014), with 87.1% of those not already validated or present in the recent studies not detected in our analysis (Fig. 4C) despite one-half of the analyzed individuals in both cases being the same, which suggests that they were errors caused by individual variation within SDs. Moreover, most previous studies capture only part of the actual polymorphic inversions. Therefore, our high-quality inversion benchmark constitutes a very useful resource for SV detection projects and future attempts to represent all human structural diversity (Wagner et al. 2022), such as the new draft human pangenome reference (Liao et al. 2023).

Another important contribution is the generation of precise genotypes for most inversions mediated by IRs, including many challenging variants that could not be previously analyzed. Specifically, we show that ONT inversion results allow us to correct three existing errors from PCR-based techniques (Giner-Delgado et al. 2019; Lerga-Jaso et al. 2026). Long-read mapping has the advantage that avoids some problems of other methods, like the need for restriction sites in specific positions, and it is more robust to different types of variation that cause genotyping errors, such as gene conversion (Giner-Delgado et al. 2019). Also, because the analysis depends on sequences located outside the IRs, any type of recurrence event that occurs between them can be detected. Accordingly, the design of the assays is very flexible and can cover a wide range of inversions, the only limitation being having reads longer than the IRs and finding uniquely mapping probe sequences at each side. Furthermore, the actual read sequence provides extra information that allows us to identify SNPs or resolve other SVs, as in the low-frequency inverted duplication of HsInv0397 and complex regions analyzed here. Still, the high variability of these repeated regions can cause some specific haplotypes to not be detected properly and give inconsistent results, as happened with several of the initial candidate inversions. Regarding previous studies, apart from the draft human pangenome reference, we found the best genotype concordance with the most recent multiplatform analysis (Porubsky et al. 2022), although ONT sequencing achieves this performance with a lower cost and effort compared with the different techniques used.

The optimization of the sequencing protocol showed that ONT AS strategy obtains a significant and quite-specific enrichment of sequence reads from the regions of interest, which is somehow compensated by a lower total sequencing output of each run (probably as a result of reduced activity and lifetime of pores during AS) (Kovaka et al. 2021; Payne et al. 2021), leading to just a twofold increase in the inversions genotyped compared to sequencing the whole genome. The combination of AS and the appropriate DNA extraction and library preparation methods to generate ultralong reads with N50 up to 70–100 kb identifies consistently 40%–60% of inversions in each individual MinION flow cell, which increases to ~90% by pooling the data from three to four flow cells together. This is equivalent to the results obtained from other available high-coverage genomes (Fig. 3A), and it suggests that the whole set of inversions could be analyzed in a single PromethION flow cell.

As seen in Figure 3C, one of the main limitations is the size of inversion breakpoint IRs, with genotyping success decreasing quickly for IRs >50 kb. Similarly, genotyping success of the samples increases as the reads get longer (higher N50), although good genotyping can also be obtained with lower N50 and more coverage, because there are still enough long reads to genotype most inversions (Fig. 3B). In addition, according to our results, using fresh cells is important to extract the longest DNA molecules and increase N50 (Fig. 1A). However, there is a well-known trade-off between read length and sequencing output with ONT, so that higher N50 values do not necessarily mean more genotyped inversions, as happens for the samples extracted using the phenol protocol (Fig. 3C). Thus, ONT inversion genotyping depends on a

balance between several factors with opposing effects, such as read length, number of sequences, and sequence selection affecting pore survival.

Accurate genotypes of SVs in multiple individuals are crucial to determine their functional effects and their association with phenotypic traits and disease susceptibility. By extending the current work to the growing amount of different types of long-read genome sequences that are being generated (Gustafson et al. 2024; Schloissnig et al. 2025), the developed methodology would enable to investigate a diverse set of inversions in a much larger number of individuals. These genotypes will help to improve inversion imputation in functional and phenotypic data, especially for recurrent inversions, for which there is very limited information (Giner-Delgado et al. 2019; Puig et al. 2020; Porubsky et al. 2022, Lerga-Jaso et al. 2026). Also, the additional genotypes will provide a better estimate of inversion frequency and distribution across human populations, which should not be affected by the relatively low recurrence rates estimated so far (Giner-Delgado et al. 2019; Puig et al. 2020; Porubsky et al. 2022). It is important to mention that a small fraction of inversions with extremely long IRs (>150–200 kb) cannot be detected with long reads and require more specialized techniques like Strand-seq (Sanders et al. 2016; Porubsky et al. 2022). In addition, as illustrated by the Figure 5 examples and other recent studies (Ebert et al. 2021; Porubsky et al. 2022; Schuy et al. 2022; Bolognini et al. 2024; Logsdon et al. 2025), there is an increasing number of complex genomic regions with multiple structural haplotypes, including different inversions, that need to be investigated in more detail. Nevertheless, ONT sequencing improves extraordinarily current inversion analysis and opens the door to the full characterization of one elusive type of human genetic variant for the first time. Because of the technology potential, it is expected that continued advances and method standardization would likely improve sequence accuracy, output, and targeting. This will result in reduced costs and increase its widespread use to analyze inversions and other complex variants in basic and translational research (Miller et al. 2021; Stevanovski et al. 2022; Dixon et al. 2023; Xu et al. 2023), contributing to a more complete picture of human genetic variation.

Methods

Human samples

Lymphoblastoid cell lines (LCLs) of four unrelated females included in the 1KGP were obtained from the Coriell Cell Repository, and they were cultured in T75 flasks at 37°C, 5% CO_2 in DMEM medium, supplemented with antibiotics, 2 mM L-glutamine (Gibco/Invitrogen), and 10% fetal serum (HyClone), until saturation. Cells were harvested, and pellets were stored at −80°C or used directly for DNA isolation. Also, we used available ONT sequencing data from 50 additional individuals of diverse origins generated mainly by the T2T Consortium (one) (Nurk et al. 2022), the Genome in a Bottle Consortium (seven) (Zook et al. 2016), the Human Structural Variation Consortium (11) (Porubsky et al. 2022), and the Human Pangenome Reference Consortium (31) (Supplemental Table S5; Liao et al. 2023)). All procedures that involved the use of human samples were approved by the research ethics committee (CERec) of the Universitat Autònoma de Barcelona.

Generation of the inversion data set

To generate the inversion data set, first we merged all 58 previously characterized inversions with IRs (Giner-Delgado et al. 2019; Puig et al. 2020; Lerga-Jaso et al. 2026) and 1528 candidate inversions from five different studies: 314 from Levy-Sakin et al. (2019), 320 from Chaisson et al. (2019), 220 from Audano et al. (2019), 315 from Ebert et al. (2021), and 359 from Porubsky et al. (2022). IRs at the predicted breakpoints, formed by SDs or other type of repetitive elements, were identified from the UCSC Genome Browser hg38 human assembly SD track (https://genome.ucsc.edu) or by BLAST (>90% identity) of the extended inverted region against itself. Those variants with both breakpoints located, respectively, within the same IR pair were considered the same inversion. When several SD pairs overlapped the breakpoints, different possible inversions were generated and tested. Second, we took all the inversion predictions generated by PEM of kilobase-long fragments (Korbel et al. 2007; Kidd et al. 2008) in the InvFEST database (Martínez-Fundichely et al. 2014) and removed those matching regions already identified in the previous step. The remaining predictions with IRs at the breakpoints, assessed as before, were added to the data set. In both cases, regions without IRs, containing complex repeat structures or large gene families, breakpoints within SD blocks >200 kb, or inverted regions <50 bp were excluded. Third, a more limited set of other IRs that could result in inversions were obtained by selecting inverted SDs of 1–50 kb of length separated by <250 kb from the UCSC hg38 SD track. According to SD definition, these sequences have a minimum of 90% identity and cannot be formed entirely by other types of repeat sequences, such as transposable elements (like LINEs or SINEs) or simple repeats. Inverted SDs that had a reciprocal overlap >80% were merged with BEDTools v2.30.0 (Quinlan and Hall 2010), and the coordinates of the merger were used. Only those regions not already included were added to the final inversion list. Finally, additional extra regions previously identified as assembly errors in the reference genome that had IRs at both breakpoints (Vicente-Salvador et al. 2017) were also added to the analysis.

DNA isolation and library preparation

High-molecular-weight genomic DNA was obtained by two different methods following the protocol specifications: (1) the Monarch HMW DNA extraction kit for cells & blood (New England Biolabs) starting from about 6 million cells and (2) a phenol-based isolation protocol optimized to obtain ultralong reads for ONT sequencing (Gong et al. 2019) starting from 20 to 30 million cells. DNA was resuspended in 200 to 760 µL of EEB buffer (10 mM Tris-HCl, 1 mM EDTA at pH 9, and 0.5% Triton X-100). DNA concentration was measured by shearing 10 µL by vortexing and using the Qubit fluorometric quantification and the AccuGreen broad-range dsDNA quantitation kit reagents (Biotium). Sequencing genomic libraries were prepared with the ultralong DNA sequencing kit V14 (ONT) following the manufacturer's instructions. DNA libraries were dissolved in 300 µL of elution buffer (10 mM Tris-HCl at pH 8.0) and kept at 4°C. During these steps, DNA samples were always handled with care using wide-bore tips, without being vortexed, shaken, or centrifuged at high speed to avoid DNA shearing.

ONT sequencing

Sequencing was performed using a MinION Mk1B or Mk1C device and R10.4.1 flow cells (ONT). Flow cells ran approximately for 3–5 days with flushes every 24 h, and each time, 37 µL of library were loaded into the flow cell. The MinKNOW software (version 22.12.5 to 23.04.3) was used to run the sequencing assays, and high-accuracy (HAC) basecalling was performed with Guppy (version 6.4.6 to 6.5.7; https://nanoporetech.com/). Real-time AS was done using MinKNOW in a Windows 10 2009 computer with an

AMD Phenom II X4 955 processor, 16 GB of RAM, and a PNY NVIDIA GTX 1080 GPU. Two different versions of a FASTA file with the target sequences were used: AS2, an initial set including only 132 candidate inversions regions, and AS3, the final set that contains 587 of the 612 inversions interrogated (Supplemental Table S1). Targeted regions in AS2 and AS3 files spanned, respectively, 40 or 70 kb of sequence at both sides of the IRs at inversion breakpoints, excluding the actual IRs to avoid generation of noninformative reads that do not cross the breakpoint. Average sequencing enrichment in AS was obtained by the ratio of the coverage of targeted and nontargeted regions in each run. Supplemental Table S6 summarizes the information of all the sequencing runs generated in this work, which was merged together in the final genotyping of each of these samples to maximize the number of inversions resolved.

Inversion genotyping

As explained above, inversions were identified by mapping a set of four specific probe sequences flanking the inversion breakpoints on ONT reads (Fig. 1A). Probe sequences were designed by a combination of manual analysis and a custom script that generates 500 bp sequences with a 250 bp overlap along the inversion region (extending up to 100 kb at each side of the external limits of the breakpoints). Manually and automatically generated sequences were used as query on BLASTN v.2.12.0 (Altschul et al. 1990) against the inversion region ±500 kb in the hg38 and T2T reference genomes with the parameters -perc_identity 80 and -qcov_hsp_perc 90, and those closest to the breakpoint and showing a single hit were selected. Selected sequences were tested for inversion genotyping (see below), and they were manually adjusted when needed to avoid specificity problems caused by unexpected variation.

The inversion genotyping consists of multiple steps represented in Supplemental Figure S1 that use different computer programs and custom scripts. These programs and scripts, including the probe sequence design, have been joined together in the new GeONTIpe package (v1.0), which was developed with Snakemake v.7.32.4 (Köster and Rahmann 2012). Briefly, ONT reads with <5 kb of length and quality of less than seven were filtered out using NanoFilt v.2.8.0 (De Coster and Rademakers 2023), and the total sequence output, number of reads, and N50 of each sequencing run were obtained with NanoPlot v.1.40.2 (De Coster and Rademakers 2023). Next, filtered reads were mapped against the hg38 reference genome with minimap2 v.2.23-r1111 (Li 2021) using the following parameters: ax map-ont -z 400,100 -r 100,1000 --secondary=no. Unmapped and mapped reads with MAPQ < 20 were removed from the alignment file using SAMtools v.1.14 (Danecek et al. 2021), whereas reads mapping to each inversion region (external inversion breakpoint coordinates ±100 kb) were selected. Specific inversion probe sequences were then mapped to the reads of the corresponding region using BLASTN v.2.12.0 with the parameters -task megablast -outfmt "6 sstart send slen sstrand" -qcov_hsp_perc 90 -sorthits 4, and the inversion allele of the individual reads was determined from the relative order, orientation, and distance of the probe sequences at both sides of the breakpoint. To ensure that genotypes for each inversion and sample are reliable, they were defined by the number of informative reads supporting each orientation as follows: (1) homozygote, if one orientation is supported by five or more reads for autosomes and Chr X in females (corresponding to a binomial *P*-value of being heterozygote of less than 0.031) or two or more reads for haploid Chr X and Y in males, and there are <5% reads of the other orientation (which allows for the presence of one to two inconsistent minoritary reads in inversions with high coverage); (2) heterozygote, when there are one or more O1 and O2 reads and they represent at least 15% of the total reads, to avoid incorrect genotype calls based on a few spurious reads; and (3) low confident, when there are insufficient reads or a small proportion of reads (5%–15%) of one orientation, which could correspond to different types of errors or other rearrangements that are harder to characterize. Moreover, for low-confident cases with two to five reads supporting the same inversion orientation, we analyzed known SNPs to test if both chromosomes are already represented in these few reads and recover additional genotypes. Biallelic SNPs within the inversion region were selected from dbSNP (Sherry et al. 2001), and those with <15% frequency or identified as homozygotes in the target samples in the 1KGP data were filtered out (Byrska-Bishop et al. 2022). SNPs present in informative reads with alleles concordant to those expected were used to calculate a distance between reads (value between zero and one) and create a dendrogram with the Hclust function of the R 4.1.2 Stats package (R Core Team 2017). Read clusters showing a distance higher than 0.5 were considered as different haplotypes belonging to a homozygote for the supported orientation.

Additional SVs of >200 bp in the inversion regions were detected by a >5% discrepancy between the expected and the observed distance among the normal combinations of the four probe sequences (AB, CD, AC, BD, BC, and AD) in >15% of reads with the same orientation (Supplemental Table S4). The distance difference was used to classify the SV as either an insertion or deletion and to estimate its size. The number of independent additional SVs was determined by removing the variants present in all O1 or O2 chromosomes (which could have been generated in the same mutational event than the inversion or represent possible errors in the reference genome or in the generated inverted reference) and duplicated SVs with the same size detected separately in O1 and O2 chromosomes in the same location (including insertions and deletions exchanging positions between O1 and O2 chromosomes that can be moved by the inversion). Multiple insertions and deletions located in the same region with sizes varying by a similar amount were considered CNVs and counted as a single SV. Cases with no internal B and C sequences and a size difference between the A and D sequences compatible with a deletion in >15% of the reads were classified as AD deletions (deletion of the entire inversion region), and one of the alleles of the individual was labeled as Del. Reads with probe sequences showing incompatible orientation conformations were classified as errors, suggesting the presence of new haplotypes that require further analysis.

Genotype comparison

Comparison of ONT inversion genotypes with those detected in previous studies based on the hg38 genome assembly was done from the information on the merged inversion data set. If several inversions in a study were associated with the same one in the ONT data, they were not taken into account. Analysis of the draft human pangenome reference samples with available paths involved the conversion of the pggb draft (Liao et al. 2023) into sequences. Hg38 coordinates were used as input to select paths and nodes with odgi v0.8.3 extract function (Guarracino et al. 2022), and odgi paths with the parameter -f were applied to convert the odgi into FASTA format. The obtained output was used to run GeONTIpe with default parameters, considering that there is only one sequence per haplotype.

Data access

All sequencing data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA1271287. The code

of the GeONTIpe inversion genotyping package is available at GitHub (https://github.com/caceres-lab/GeONTIpe) and as Supplemental Code (distributed under the MIT License). Inversion genotypes have been submitted to NCBI's database of human genomic Structural Variation (dbVar; https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd248/) and are also available in the Supplemental Material.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Jim Thomas, Olga Francino, Michael Vieyra, and the ONT technical support team and multiple ONT users for advice and technical support about DNA extraction and nanopore sequencing; Carles Acosta for help with GeONTIpe implementation in a supercomputer environment; David Porubsky and Jan Korbel for suggestions about the development of the Snakemake package; Odei Blanco for individual inversions quality-control information; and Adrià Mompart for thorough GeONTIpe testing. The data processing and analysis tools have been developed, implemented, and operated in collaboration with the Port d'Informació Científica (PIC) data center. PIC is maintained through a collaboration agreement between the Institut de Física d'Altes Energies (IFAE) and the Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT). This work was supported by research grants PID2019-107836RB-I00 and PID2022-137615OB-I00 to M.C. and PID2023-146193OB-I00 to M.A.S. funded by the Agencia Estatal de Investigación of the Ministerio de Ciencia, Innovación y Universidades (MICIU/AEI/10.13039/501100011033, Spain) and the European Regional Development Fund (ERDF, European Union), 2021 SGR 00526 support to research groups from the Departament de Recerca i Universitats (Generalitat de Catalunya, Spain) to M.C., the predoctoral program AGAUR-FI Joan Oró grants 2021FI-B 00240 to R.M.-P. and 2023 FI-1 00746 to M.D.-R. funded by the Departament de Recerca i Universitats (Generalitat de Catalunya, Spain) and the European Social Plus Fund, a Maria Zambrano grant 715491 from the MICIU (Spain) funded by the European Union-NextGenerationEU to K.K., and predoctoral contract PRE-2020-092440 to I.Y. (MICIU/AEI/10.13039/501100011033, Spain). M.P. is a Serra Húnter Fellow.

Author contributions: M.C. conceived the project, devised the study, and oversaw all the steps. R.M.-P. and I.Y. developed the inversion genotyping pipeline with help from A.S., J.M.-U., M.P., and M.C. R.M.-P., M.P., M.D.-R., and M.C. selected the inversion regions. K.K., O.C., M.A.S., and M.P. performed the ONT sequencing. R.M.-P., K.K., and M.P. analyzed the previously available and newly generated ONT sequencing data. R.M.-P., K.K., M.P., and M.C. wrote the manuscript, and all the authors contributed comments to its final version.

References

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526:** 68–74. doi:10.1038/nature15393

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12:** 363–376. doi:10.1038/nrg2958

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410. doi:10.1016/S0022-2836(05)80360-2

Antonacci F, Kidd JM, Marques-Bonet T, Teague B, Ventura M, Girirajan S, Alkan C, Campbell CD, Vives L, Malig M, et al. 2010. A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat Genet* **42:** 745–750. doi:10.1038/ng.643

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al. 2019. Characterizing the major structural variant alleles of the human genome. *Cell* **176:** 663–675.e19. doi:10.1016/j.cell.2018.12.019

Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, Atlason BA, Kristmundsdottir S, Mehringer S, Hardarson MT, et al. 2021. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet* **53:** 779–786. doi:10.1038/s41588-021-00865-4

Bolognini D, Halgren A, Lou RN, Raveane A, Rocha JL, Guarracino A, Soranzo N, Chin C-S, Garrison E, Sudmant PH. 2024. Recurrent evolution and selection shape structural diversity at the amylase locus. *Nature* **634:** 617–625. doi:10.1038/s41586-024-07911-1

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185:** 3426–3440.e19. doi:10.1016/j.cell.2022.08.004

Campoy E, Puig M, Yakymenko I, Lerga-Jaso J, Cáceres M. 2022. Genomic architecture and functional effects of potential human inversion supergenes. *Philos Trans R Soc Lond B Biol Sci* **377:** 20210209. doi:10.1098/rstb.2021.0209

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10:** 1784. doi:10.1038/s41467-018-08148-z

Choi J, Kim S, Kim J, Son HY, Yoo SK, Kim CU, Park YJ, Moon S, Cha B, Jeon MC, et al. 2023. A whole-genome reference panel of 14,393 individuals for east Asian populations accelerates discovery of rare functional variants. *Sci Adv* **9:** eadg6319. doi:10.1126/sciadv.adg6319

Choudhury A, Aron S, Botigué LR, Sengupta D, Botha G, Bensellak T, Wells G, Kumuthini J, Shriner D, Fakim YJ, et al. 2020. High-depth African genomes inform human migration and health. *Nature* **586:** 741–748. doi:10.1038/s41586-020-2859-7

Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera AV, Lowther C, Gauthier LD, Wang H, et al. 2020. A structural variation reference for medical and population genetics. *Nature* **581:** 444–451. doi:10.1038/s41586-020-2287-8

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10:** giab008. doi:10.1093/gigascience/giab008

De Coster W, Rademakers R. 2023. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39:** btad311. doi:10.1093/bioinformatics/btad311

De Coster W, Weissensteiner MH, Sedlazeck FJ. 2021. Towards population-scale long-read sequencing. *Nat Rev Genet* **22:** 572–587. doi:10.1038/s41576-021-00367-3

Dixon K, Shen Y, O'Neill K, Mungall KL, Chan S, Bilobram S, Zhang W, Bezeau M, Sharma A, Fok A, et al. 2023. Defining the heterogeneity of unbalanced structural variation underlying breast cancer susceptibility by nanopore genome sequencing. *Eur J Hum Genet* **31:** 602–606. doi:10.1038/s41431-023-01284-1

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Mari RS, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372:** eabf7117. doi:10.1126/science.abf7117

Giner-Delgado C, Villatoro S, Lerga-Jaso J, Gayà-Vidal M, Oliva M, Castellano D, Pantano L, Bitarello B, Izquierdo D, Noguera I, et al. 2019. Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nat Commun* **10:** 4222. doi:10.1038/s41467-019-12173-x

Gong L, Wong C-H, Idol J, Ngan CY, Wei C-L. 2019. Ultra-long read sequencing for whole genomic DNA analysis. *J Vis Exp* **145:** e58954. doi:10.3791/58954

Grochowski CM, Bengtsson JD, Du H, Gandhi M, Lun MY, Mehaffey MG, Park K, Höps W, Benito E, Hasenfeld P, et al. 2024. Inverted triplications formed by iterative template switches generate structural variant diversity at genomic disorder loci. *Cell Genom* **4:** 100590. doi:10.1016/j.xgen.2024.100590

Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E. 2022. ODGI: understanding pangenome graphs. *Bioinformatics* **38:** 3319–3326. doi:10.1093/bioinformatics/btac308

Gustafson JA, Gibson SB, Damaraju N, Zalusky MPG, Hoekzema K, Twesigomwe D, Yang L, Snead AA, Richmond PA, De Coster W, et al. 2024. High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res* **34:** 2061–2073. doi:10.1101/gr.279273.124

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581:** 434–443. doi:10.1038/s41586-020-2308-7

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453:** 56–64. doi:10.1038/nature06862

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318:** 420–426. doi:10.1126/science.1149504

Köster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28:** 2520–2522. doi:10.1093/bioinformatics/bts480

Kovaka S, Fan Y, Ni B, Timp W, Schatz MC. 2021. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat Biotechnol* **39:** 431–441. doi:10.1038/s41587-020-0731-9

Lerga-Jaso J, Campoy E, Puig M, Yakymenko I, Gómez-Graciani R, Moreira-Pinhal R, Soos T, Vilella A, Ramírez C, Giner-Delgado C, et al. 2026. Towards a complete characterization of common human polymorphic inversions and their functional effects. medRxiv doi:10.64898/2026.08.03.26359593

Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, McCaffrey J, Young E, Lam ET, Hastie AR, Wong KHY, et al. 2019. Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat Commun* **10:** 1025. doi:10.1038/s41467-019-08992-7

Li H. 2021. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37:** 4572–4574. doi:10.1093/bioinformatics/btab705

Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. 2023. A draft human pangenome reference. *Nature* **617:** 312–324. doi:10.1038/s41586-023-05896-x

Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21:** 597–614. doi:10.1038/s41576-020-0236-x

Logsdon GA, Ebert P, Audano PA, Loftus M, Porubsky D, Ebler J, Yilmaz F, Hallast P, Prodanov T, Yoo D, et al. 2025. Complex genetic variation in nearly complete human genomes. *Nature* **644:** 430–441. doi:10.1038/s41586-025-09140-6

Martínez-Fundichely A, Casillas S, Egea R, Ràmia M, Barbadilla A, Pantano L, Puig M, Cáceres M. 2014. InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic Acids Res* **42:** D1027–D1032. doi:10.1093/nar/gkt1122

Mastrorosa FK, Miller DE, Eichler EE. 2023. Applications of long-read sequencing to Mendelian genetics. *Genome Med* **15:** 42. doi:10.1186/s13073-023-01194-3

Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, Munson KM, Lewis AP, Fuerte EPA, Paschal CR, Walsh T, et al. 2021. Targeted long-read sequencing identifies missing disease-causing variation. *Am J Hum Genet* **108:** 1436–1449. doi:10.1016/j.ajhg.2021.06.006

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376:** 44–53. doi:10.1126/science.abj6987

Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M. 2021. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat Biotechnol* **39:** 442–450. doi:10.1038/s41587-020-00746-x

Porubsky D, Höps W, Ashraf H, Hsieh PH, Rodriguez-Martin B, Yilmaz F, Ebler J, Hallast P, Maria Maggiolini FA, Harvey WT, et al. 2022. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185:** 1986–2005.e26. doi:10.1016/j.cell.2022.04.017

Puig M, Casillas S, Villatoro S, Cáceres M. 2015. Human inversions and their functional consequences. *Brief Funct Genomics* **14:** 369–379. doi:10.1093/bfgp/elv020

Puig M, Lerga-Jaso J, Giner-Delgado C, Pacheco S, Izquierdo D, Delprat A, Gayà-Vidal M, Regan JF, Karlin-Neumann G, Cáceres M. 2020. Determining the impact of uncharacterized inversions in the human genome by droplet digital PCR. *Genome Res* **30:** 724–735. doi:10.1101/gr.255273.119

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

R Core Team. 2017. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Sanders AD, Hills M, Porubský D, Guryev V, Falconer E, Lansdorp PM. 2016. Characterizing polymorphic inversions in human genomes by single cell sequencing. *Genome Res* **26:** 1575–1587. doi:10.1101/gr.201160.115

Schloissnig S, Pani S, Ebler J, Hain C, Tsapalou V, Söylev A, Hüther P, Ashraf H, Prodanov T, Asparuhova M, et al. 2025. Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature* **644:** 442–452. doi:10.1038/s41586-025-09290-7

Schuy J, Grochowski CM, Carvalho CMB, Lindstrand A. 2022. Complex genomic rearrangements: an underestimated cause of rare diseases. *Trends Genet* **38:** 1134–1146. doi:10.1016/j.tig.2022.06.003

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29:** 308–311. doi:10.1093/nar/29.1.308

Stevanovski I, Chintalaphani SR, Gamaarachchi H, Ferguson JM, Pineda SS, Scriba CK, Tchan M, Fung V, Ng K, Cortese A, et al. 2022. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Sci Adv* **8:** 17. doi:10.1126/sciadv.abm5386

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526:** 75–81. doi:10.1038/nature15394

Vicente-Salvador D, Puig M, Gayà-Vidal M, Pacheco S, Giner-Delgado C, Noguera I, Izquierdo D, Martínez-Fundichely A, Ruiz-Herrera A, Estivill X, et al. 2017. Detailed analysis of inversions predicted between two human genomes: errors, real polymorphisms, and their origin and population distribution. *Hum Mol Genet* **26:** 567–581. doi:10.1093/hmg/ddw415

Wagner J, Olson ND, Harris L, McDaniel J, Cheng H, Fungtammasan A, Hwang Y-C, Gupta R, Wenger AM, Rowell WJ, et al. 2022. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol* **40:** 672–680. doi:10.1038/s41587-021-01158-1

Wellenreuther M, Bernatchez L. 2018. Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol Evol* **33:** 427–440. doi:10.1016/j.tree.2018.04.002

Wu Z, Jiang Z, Li T, Xie C, Zhao L, Yang J, Ouyang S, Liu Y, Li T, Xie Z. 2021. Structural variants in the Chinese population and their impact on phenotypes, diseases and population adaptation. *Nat Commun* **12:** 6501. doi:10.1038/s41467-021-26856-x

Xu L, Wang X, Lu X, Liang F, Liu Z, Zhang H, Li X, Tian S, Wang L, Wang Z. 2023. Long-read sequencing identifies novel structural variations in colorectal cancer. *PLoS Genet* **19:** e1010514. doi:10.1371/journal.pgen.1010514

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3:** 160025. doi:10.1038/sdata.2016.25

Received May 6, 2025; accepted in revised form August 26, 2026.