Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies

Xuefang Zhao,[1,2,3] Ryan L. Collins,[1,2,4] Wan-Ping Lee,[5] Alexandra M. Weber,[6,7] Yukyung Jun,[5] Qihui Zhu,[5] Ben Weisburd,[2] Yongqing Huang,[8] Peter A. Audano,[9] Harold Wang,[1,2] Mark Walker,[2,3] Chelsea Lowther,[1,2,3] Jack Fu,[1,2,3] Human Genome Structural Variation Consortium, Mark B. Gerstein,[10] Scott E. Devine,[11] Tobias Marschall,[12] Jan O. Korbel,[13,14] Evan E. Eichler,[9,15] Mark J.P. Chaisson,[9,16] Charles Lee,[5,17,18] Ryan E. Mills,[6,7] Harrison Brand,[1,2,3] and Michael E. Talkowski[1,2,3,4,*]

Summary

Virtually all genome sequencing efforts in national biobanks, complex and Mendelian disease programs, and medical genetic initiatives are reliant upon short-read whole-genome sequencing (srWGS), which presents challenges for the detection of structural variants (SVs) relative to emerging long-read WGS (lrWGS) technologies. Given this ubiquity of srWGS in large-scale genomics initiatives, we sought to establish expectations for routine SV detection from this data type by comparison with lrWGS assembly, as well as to quantify the genomic properties and added value of SVs uniquely accessible to each technology. Analyses from the Human Genome Structural Variation Consortium (HGSVC) of three families captured ~11,000 SVs per genome from srWGS and ~25,000 SVs per genome from lrWGS assembly. Detection power and precision for SV discovery varied dramatically by genomic context and variant class: 9.7% of the current GRCh38 reference is defined by segmental duplication (SD) and simple repeat (SR), yet 91.4% of deletions that were specifically discovered by lrWGS localized to these regions. Across the remaining 90.3% of reference sequence, we observed extremely high (93.8%) concordance between technologies for deletions in these datasets. In contrast, lrWGS was superior for detection of insertions across all genomic contexts. Given that non-SD/SR sequences encompass 95.9% of currently annotated disease-associated exons, improved sensitivity from lrWGS to discover novel pathogenic deletions in these currently interpretable genomic regions is likely to be incremental. However, these analyses highlight the considerable added value of assembly-based lrWGS to create new catalogs of insertions and transposable elements, as well as disease-associated repeat expansions in genomic sequences that were previously recalcitrant to routine assessment.

The field of genomics has seen remarkable advances in the accuracy and efficiency of massively parallel sequencing-by-synthesis technologies that generate pairs of short reads from the ends of small 400–800 base pair (bp) fragments (referred to herein as short-read whole-genome sequencing [srWGS]). This technical leap and derivative approaches such as targeted whole-exome capture sequencing (WES) have catalyzed a deluge of gene discoveries for rare diseases and insights into population genetics and genome biology. Correspondingly, srWGS has been adopted by all major human disease and biobank sequencing initiatives, including the NHGRI Centers for Common Disease Genomics (CCDG)[1] and Centers for Mendelian Genetics (CMG),[2] the Deciphering Developmental Disorders (DDD) project,[3] the Trans-Omics for Precision Medicine (TOPMed),[4] the All of Us Research Program,[5] the NICHD Gabriella Miller Kids First (GMKF) initiative, the UK BioBank,[6] and Genomics England,[7] to name just a few. As such, a critical step for the field is to establish uniform methods for srWGS data processing and rational benchmarking standards to set expectations for variant detection.

The technical processes of genome alignment and single-nucleotide variant (SNV) detection have been an intensive focus of genomics since the inception of the 1000 Genomes Project[8] and more recently updated for

[1]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA; [2]Program in Medical and Population Genetics and Stanley Center for Psychiatric Disorders, Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02142, USA; [3]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA; [4]Division of Medical Sciences, Harvard Medical School, Boston, MA 02115, USA; [5]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA; [6]Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, 100 Washtenaw Avenue, Ann Arbor, MI 48109, USA; [7]Department of Human Genetics, University of Michigan Medical School, 1241 East Catherine Street, Ann Arbor, MI 48109, USA; [8]Data Sciences Platform, Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02142, USA; [9]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA; [10]Yale University Medical School, Computational Biology and Bioinformatics Program, New Haven, CT 06520, USA; [11]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA; [12]Institute for Medical Biometry and Bioinformatics, Medical Faculty, Heinrich Heine University, 40225 Düsseldorf, Germany; [13]European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany; [14]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK; [15]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA; [16]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA; [17]Department of Graduate Studies – Life Sciences, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, South Korea; [18]Precision Medicine Center, The First Affiliated Hospital of Xi'an Jiaotong University, 277 West Yanta Road, Xi'an 710061, Shaanxi, People's Republic of China

*Correspondence: talkowsk@broadinstitute.org

https://doi.org/10.1016/j.ajhg.2021.03.014.

cross-institute functional equivalence as part of the NHGRI Genome Sequencing Program and variant detection with the Genome Analysis Toolkit (GATK) best practices.[9–11] However, no comparable standardized methods have been adopted for detection of structural variants (SVs), defined here as genomic alterations greater than 50 bp, from srWGS, and there are limited gold-standard benchmarking approaches for SV discovery. This lack of uniformity has introduced a barrier to establishing reliable estimates of the SV counts and characteristics per genome. Not surprisingly, as shown in Figure 1A, these estimates have varied considerably across studies. The initial discovery effort from the 1000 Genomes Project[12,13] revealed the landscape of SVs that could be captured from srWGS with just 4–7× coverage (3,431 SVs per genome). More recent population genetic and human disease studies using deeper (30× or higher) srWGS and diverse analytic methods have varied in estimates of SVs that can be captured via srWGS; these estimates vary from 401 to 10,884 per genome. At present, the most sensitive studies have utilized the integration of multiple SV detection methods from the Genome Aggregation Database (gnomAD) and the Human Genome Structural Variation Consortium (HGSVC) projects (Figure 1A).[1,14,15,13,16–20]

Emerging long-read WGS (lrWGS) technologies, which involve sequencing thousands to millions of contiguous nucleotides from a single strand of DNA, have significantly increased sensitivity for SV discovery in the human genome. The most widely tested lrWGS technologies include single-molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio)[24] and sequencing by ionic current through a nanopore channel (Oxford Nanopore Technologies [ONT]).[25] A key advantage of lrWGS is the abundance of reads that span entire SVs, allowing for direct observation of SVs rather than SV detection by inference as required for srWGS. These unique properties of lrWGS are beginning to revolutionize *de novo* assembly approaches,[26,27] and methods are already maturing for telomere-to-telomere assembly of individual human chromosomes.[15,28,29] The most recent analyses used the combination of multiple sequencing platforms (e.g., lrWGS, strand-specific sequencing,[30] and optical mapping[31]) in relatively small numbers of genomes to generate assembly-based SV callsets,[14,32] which have approximately doubled the number of SVs able to be captured in each genome to ~25,000 as compared with srWGS[14,15] (Figure 1A).

These lrWGS studies have thus opened access to SVs in the genome that were traditionally refractory to discovery by srWGS or interpretation in disease association studies, such as repeat expansions and other alterations within repetitive genomic regions and centromeres.[33] Unfortunately, the current cost of lrWGS is a significant premium over srWGS. For example, as of this writing the cost for generation of PacBio lrWGS over srWGS for equivalent coverage at leading academic platforms from the HGSVC ranges from 5.9-fold increase for continuous long-read technology to 12-fold increase for circular consensus sequencing HiFi technology. Moreover, the comparatively lower throughput of modern lrWGS platforms renders them impractical for adoption in large-scale population studies on the order of tens to hundreds of thousands of individuals. The largest published assembly-based PacBio study to date has analyzed just 15 genomes,[15] and a recent preprint from the HGSVC describes 35 genomes,[34] while a published study from Iceland analyzed 3,622 ONT genomes.[35] By comparison, millions of genomes have already been sequenced or commissioned via srWGS across international initiatives. Given this predominance of srWGS in the current landscape of genomics research, we present here a series of analyses from the HGSVC to (1) benchmark expectations for the number and class of variants that can be reliably detected from srWGS, (2) predict the genomic features that drive false positive and false negative discoveries for each technology, and (3) establish the scientific and clinical advances offered by state-of-the-art lrWGS assembly as a complementary approach to srWGS.

In this study, we performed a detailed comparison of SVs detected from alignment-based srWGS and assembly-based lrWGS methods on three matched trio families (HG00514, HG00733, and NA19240) from the 1000 Genomes Project, and all results per genome reported here are averages across the three children in these families.[14] For srWGS, this initial study applied a highly sensitive ensemble approach to integrate 13 SV detection algorithms (supplemental material and methods) and discovered an average of 10,884 SVs per genome. The emphasis on sensitivity suggests that approximately 11,000 SVs per genome most likely reflects an upper bound on the total number of SVs that can be routinely captured from srWGS with the alignment-based algorithms applied by the HGSVC, as demonstrated in Figure 1A by comparison with other contemporary studies. However, this sensitivity came at the significant cost of specificity: 685 *de novo* SVs were observed per genome, over 1,000-fold more than our expectation from srWGS based on family studies, population genetic estimators, and molecular validation, therefore representing many SV predictions that are most likely false positives.[16] The lrWGS-derived SV callset combined whole-genome phasing with two state-of-the-art genome assembly approaches (Phase-SV and MS-PAC[14,26,36]) and was supplemented by additional technologies (HiC[37] and StrandSeq,[38] see Chaisson et al.[14]). These methods discovered an average of 24,825 haplotype-resolved SVs per genome, or over 2-fold more than the most sensitive srWGS approaches. Surprisingly, although the srWGS and lrWGS callsets were generated on identical samples, only a limited subset of SVs (66.8% of srWGS and 33.5% of lrWGS) overlapped between technologies. Moreover, the mutational class of SVs dramatically impacted concordance: 60.6% of srWGS and 48.7% of lrWGS deletions demonstrated overlap as compared with 81.7% of srWGS and 24.1% of lrWGS insertions (Figure 1B).

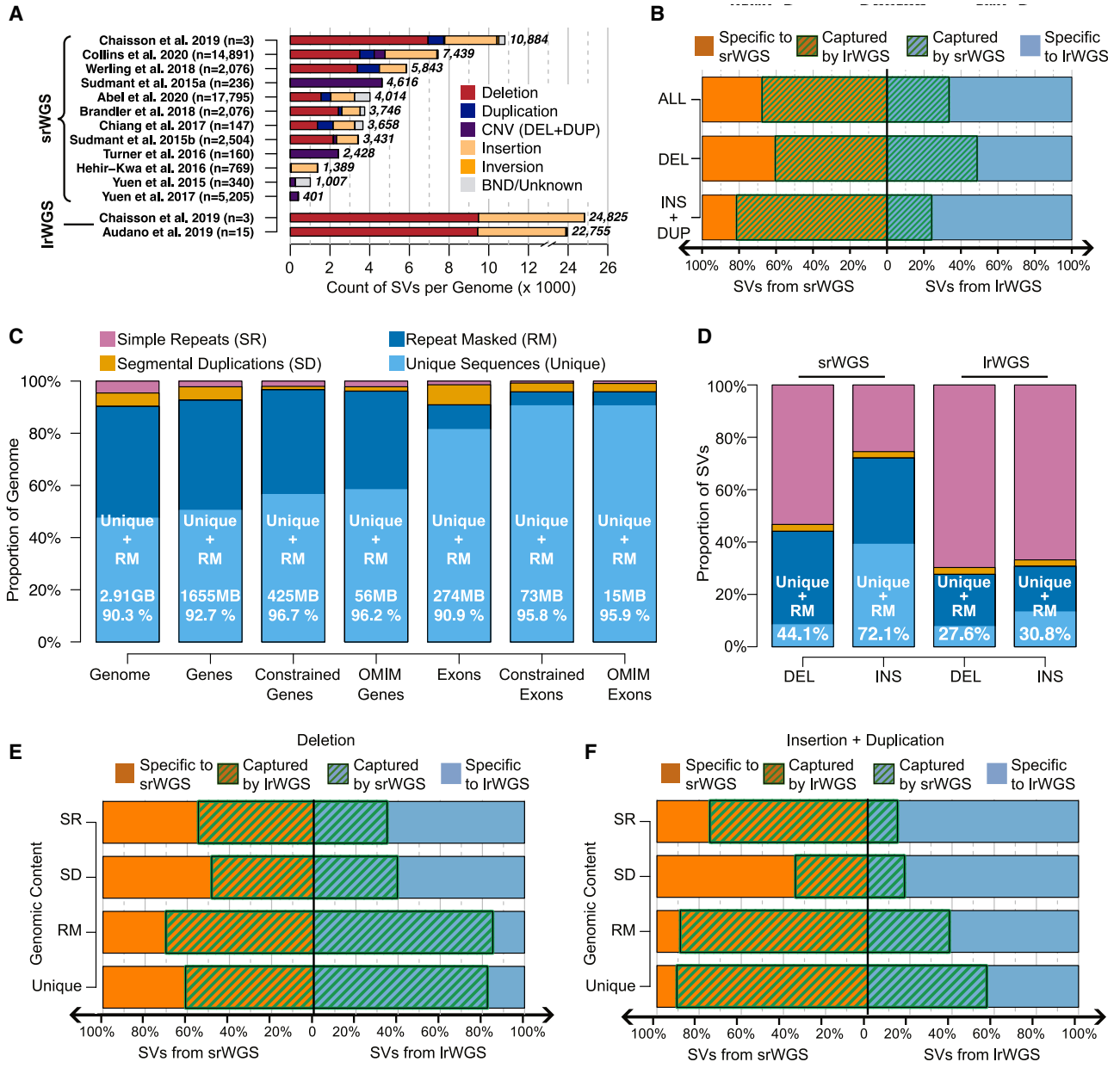

Figure 1. Comparison of SV callsets from srWGS and lrWGS

(A) The substantial increased yield of lrWGS in SV detection is displayed from the HGSVC[14] and the largest Pacific Biosciences (PacBio) lrWGS study published to date[15] by comparison with contemporary srWGS studies. As shown, there is wide variability of SV detection across srWGS studies to date that report SVs detected per individual in more than 100 genomes. Parenthetical numbers next to each study label indicate the number of genomes analyzed, and bold numbers next to each bar represent the number of SVs per genome reported by each study.

(B) Overlap of SVs from the HGSVC srWGS and lrWGS callsets across children of the three trio families, partitioned by SV class.

(C) Distribution of repetitive sequences across the genome, genes, and exons. "Constrained" refers to genes and exons with pLI > 0.9,[21] and "OMIM genes" includes a curated list of autosomal dominant genes that were defined in both Berg et al.[22] and Blekhman et al.[23] Gb, gigabase; Mb, megabase. Percentage listed within each bar is the fraction of each group composed of "unique + RM" sequences.

(D) Distribution of SVs from srWGS and lrWGS split by repetitive sequence context. Formatting conventions are the same as in (C).

(E and F) Concordance of deletions (E) and insertions and duplications (F) between srWGS and lrWGS split by repetitive sequence context.

We sought to define and quantify the factors contributing to the poor concordance between SVs derived from each technology to improve SV discovery, filtering, and prioritization from srWGS in future large-scale medical and population genetic initiatives. We first explored the role of genomic features such as repetitive sequences that are enriched for SVs via repeat-mediated mechanisms[39,40] because short-read alignment has well-documented limitations within these genomic regions.[41,42] We annotated all SVs with sequence context based on RepeatMasker[43] and segmental duplication[44] tracks from the UCSC genome browser.[45,46] For simplicity, we consolidated all repetitive

sequence annotations into three categories: segmental duplication (SD; 5.1% of the genome), simple repeat (SR; 4.6%), and "repeat masked" (RM; 42.9%), where this RM category referred to all other repetitive sequence not overlapping SD or SR elements. The remaining 47.4% of the genome not overlapping any of these repeat categories was labeled as "unique" sequence, which is a term used for simplicity here, although these regions are not completely devoid of repetitive sequences. The "unique" and RM categories collectively encompass 90.3% of the annotated human reference sequence, 90.9% of all currently annotated protein-coding sequence, 95.8% of all currently annotated coding sequence from evolutionarily constrained genes, and 95.9% of genes currently associated with human disease from the Online Mendelian Inheritance in Man (OMIM; Figure 1C).[21,47–49]

As expected, the distribution of SVs was non-uniform and varied by sequence context for each technology (Figure 1D). Most prominently, the enrichment of SV breakpoints in highly repetitive genomic sequences (SD + SR regions) was dramatic and their distribution differed significantly between technologies: despite representing just 9.7% of the reference genome, SD + SR annotated sequences contained at least one breakpoint from 49.8% of all SVs from srWGS and 70.4% of all SVs from lrWGS ($p < 2.2e{-}16$ for both technologies, chi-square test, Table S1, see supplemental material and methods for details). This enrichment of SVs in repetitive sequence was also strongly correlated with concordance between srWGS and lrWGS: SVs located in repetitive SD + SR sequences displayed 57.0% concordance among srWGS variants and 22.5% in lrWGS variants, whereas those ratios improved considerably in less repetitive sequences ("unique + RM") to 76.5% in srWGS and 59.9% in lrWGS (Figures 1E and 1F).

Although the divergent distributions and diminished concordance of SV detection by technology aligned with expectations for SD + SR regions, the paucity of overlap between technologies in "unique + RM" regions was unexpected because breakpoints localized to these regions should not suffer from the same technical confounders of SV discovery in highly repetitive sequences. Therefore, we next sought to decouple and quantify the discordance driven by underlying biological features of the genome from technical noise driven by false positive SVs present in the underlying HGSVC callsets that were optimized for sensitivity as described above. We also reasoned that identifying the covariates that have the greatest influence on false positive SV calls would be valuable in guiding the human genetics community toward principled improvements in SV detection and filtering algorithms. To accomplish this, we developed an *in silico* SV assessment to improve the precision of srWGS and lrWGS callsets in non-repetitive regions. This procedure re-evaluated the following three pieces of orthogonal information from both lrWGS and srWGS for each SV: (1) supporting evidence from raw lrWGS reads in the parent and offspring genomes for the presence of an SV (VaPoR;[50] Figure 2A); (2) copy states based on srWGS normalized read depth within SVs (Figures 2B and S1); (3) discordant paired-end and split reads information at the breakpoint of each predicted SV (Figures 2C, 2D, and S2, Table S2). We considered the SVs with one or more modes of supporting evidence as "high confidence" and explored their overlap on the basis of repeat context for SV calls from different technologies (see supplemental material and methods for further details).

We initially applied this *in silico* SV refinement procedure to deletions, which represent the most interpretable class of SVs for genomics applications. As expected, the *in silico* confirmation rate—i.e., the proportion of SVs supported by one or more of the evidence classes described above—was high (93.5%) for deletions concordant between technologies in "unique + RM" regions compared to just 13.5% and 33.1% for those that were only discovered by a single technology for srWGS or lrWGS, respectively (Figure S3). After restricting to high-confidence SVs, we observed a substantial improvement in concordance: 93.5% to 93.8% of deletions were shared between srWGS and lrWGS (Figure 2E). Although mutational processes such as somatic SVs or sub-clonal mutations arising in cell culture can contribute to false positive findings, these results implied that most of the discordance between srWGS and lrWGS for SV discovery in the 90.3% of the genome not encompassed by SD + SR sequence was most likely technical in origin. Importantly, it appeared that most of the discordance wasdriven by false positive SV calls that can be pruned by *post hoc* heuristic filtering.

We next explored the impact of *post hoc* filtering on SVs other than deletions. While duplications and insertions were reported as separate SV classes by srWGS, the lrWGS methods applied by the HGSVC treated both classes as insertions. Given this, we considered all srWGS duplications as insertions for subsequent comparisons. In contrast to the strong concordance between srWGS and lrWGS observed for deletions, 45.5% of high-confidence lrWGS insertions in "unique + RM" regions had no matching SV call from srWGS, while the majority (96.0%) of srWGS insertions and duplications were captured by lrWGS (Figures 2F and S4). To investigate the properties of insertions specifically captured by lrWGS in "unique + RM" sequences, we aligned the assembled sequences of high-confidence insertions against a catalog of known repeat elements.[43] Most of these insertions aligned to specific types of repeat elements (61.8%, n = 2,485/genome), such as short interspersed nuclear elements (SINEs, n = 1,494/genome), long interspersed nuclear elements (LINEs, n = 312/genome), and long terminal repeat (LTR, n = 139/genome) retrotransposons (Figures 3A and 3B). Notably, a "chimeric" alignment pattern was observed for 31.7% of the insertions specifically discovered by lrWGS where inserted sequences were aligned to multiple different repeat types (Figures 3B and 3C). These results indicate that the complexity of insertion repeat structure is a major

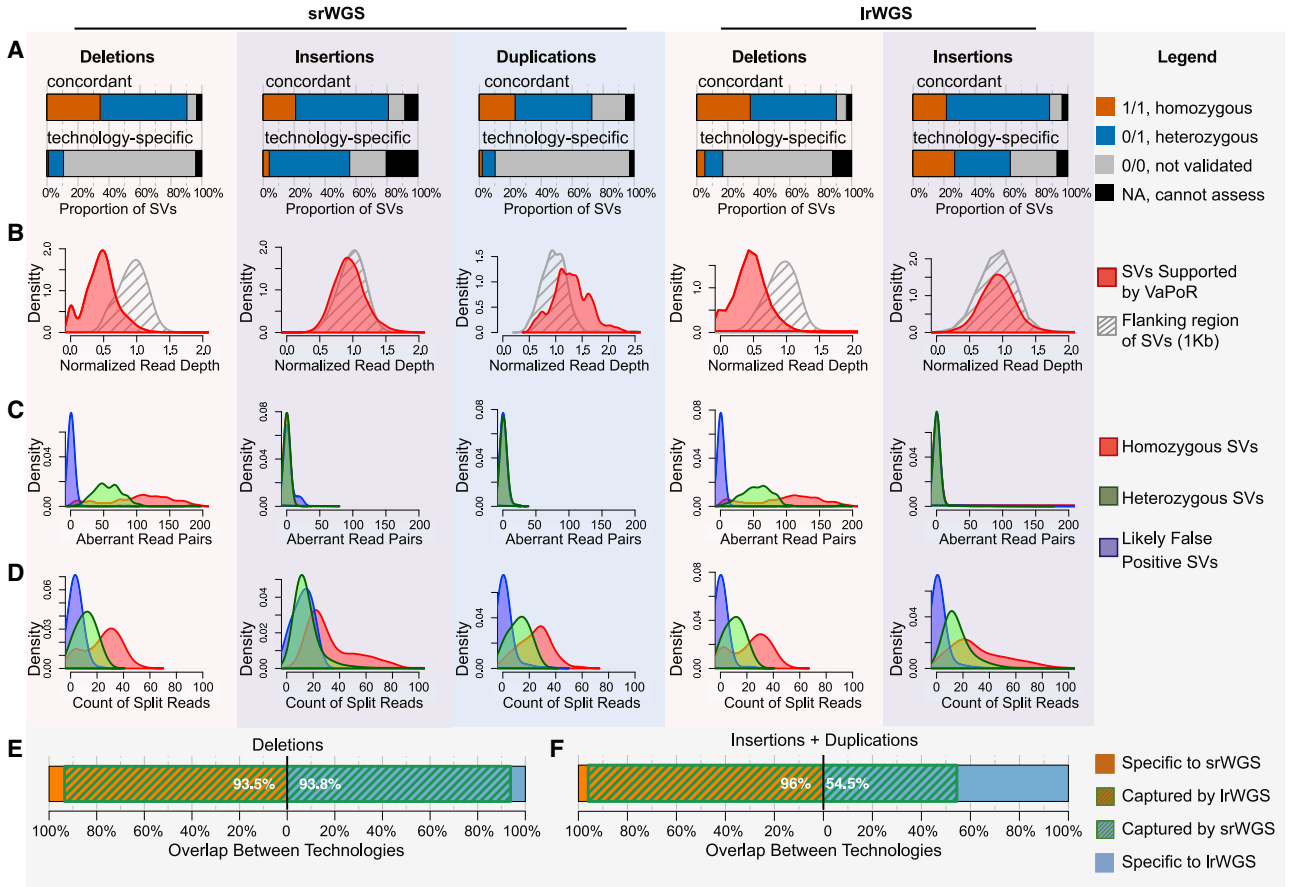

Figure 2. Methods to recalibrate SVs in "unique + RM" sequences based on read-level alignment signatures

(A) *In silico* evaluation results from VaPoR on deletions (pink background), insertions (purple background), and duplications (blue background). Duplications and insertions reported by srWGS were both compared against insertions from lrWGS. "Concordant" represents SVs discovered by both lrWGS and srWGS, and "technology-specific" represents SVs specifically discovered from one technology.

(B) Distribution of normalized read depth of srWGS across deletions (pink background), insertions (purple background), and duplications (blue background) that were supported by VaPoR (red) and the 1 kb genomic regions that flank these SVs (gray).

(C and D) Distribution of aberrant srWGS read pairs (C) and split reads (D) around deletions (pink background), insertions (purple background), and duplications (blue background) that were either homozygous (red), heterozygous (green), or false positives (blue). The homozygous, heterozygous, and likely false positive SV sets were selected with the criteria described in the supplemental material and methods.

(E and F) Concordance of deletions (E) and insertions and duplications (F) in "unique + RM" sequences that were supported by the *in silico* SV refinement procedure. Percentages represent the fraction of total variants shared between srWGS and lrWGS.

determinant of srWGS sensitivity for insertion SVs, as has been previously demonstrated for certain classes of nested insertions.[51] We further observed high variability in the current capabilities of srWGS detection algorithms depending on the type of transposable element insertions when comparing with lrWGS: 74.9% of SINEs, 42.6% of LINEs, and 50.7% of LTRs detected by lrWGS were also discovered by srWGS (Figure 3D). Intriguingly, almost all (95.8%) of the high-confidence lrWGS insertions in "unique + RM" regions that were only discovered by lrWGS nevertheless had some detectable support in the raw srWGS data, indicating that continued development of insertion detection algorithms could substantially improve sensitivity for identification of this variant class from srWGS (Figure 3E). Taken together, these analyses indicate that lrWGS and assembly-based approaches provide substantial improvements over srWGS for insertion discovery, particularly for those events with complex repeat structures.

We also examined SVs in highly repetitive SD + SR regions by using the same *in silico* evaluation framework (Figures S5A–S5D) as described above with the caveat that the orthogonal evaluation of variants in these regions is more challenging and prone to false positives due to alignment artifacts that do not arise in the less repetitive regions of the genome. Similar to the "unique + RM" regions, insertions were poorly captured by srWGS, and only 17.0% overlapped lrWGS insertions, while 74.0% of srWGS insertions were captured by lrWGS (Figure S5F). The high concordance for deletions in "unique + RM" sequences also dissipated in these more repetitive SD + SR regions, as the concordance was 69.6% and 40.4% of high-confidence deletions from srWGS and lrWGS that were shared by the other technology, respectively (Figure S5E).

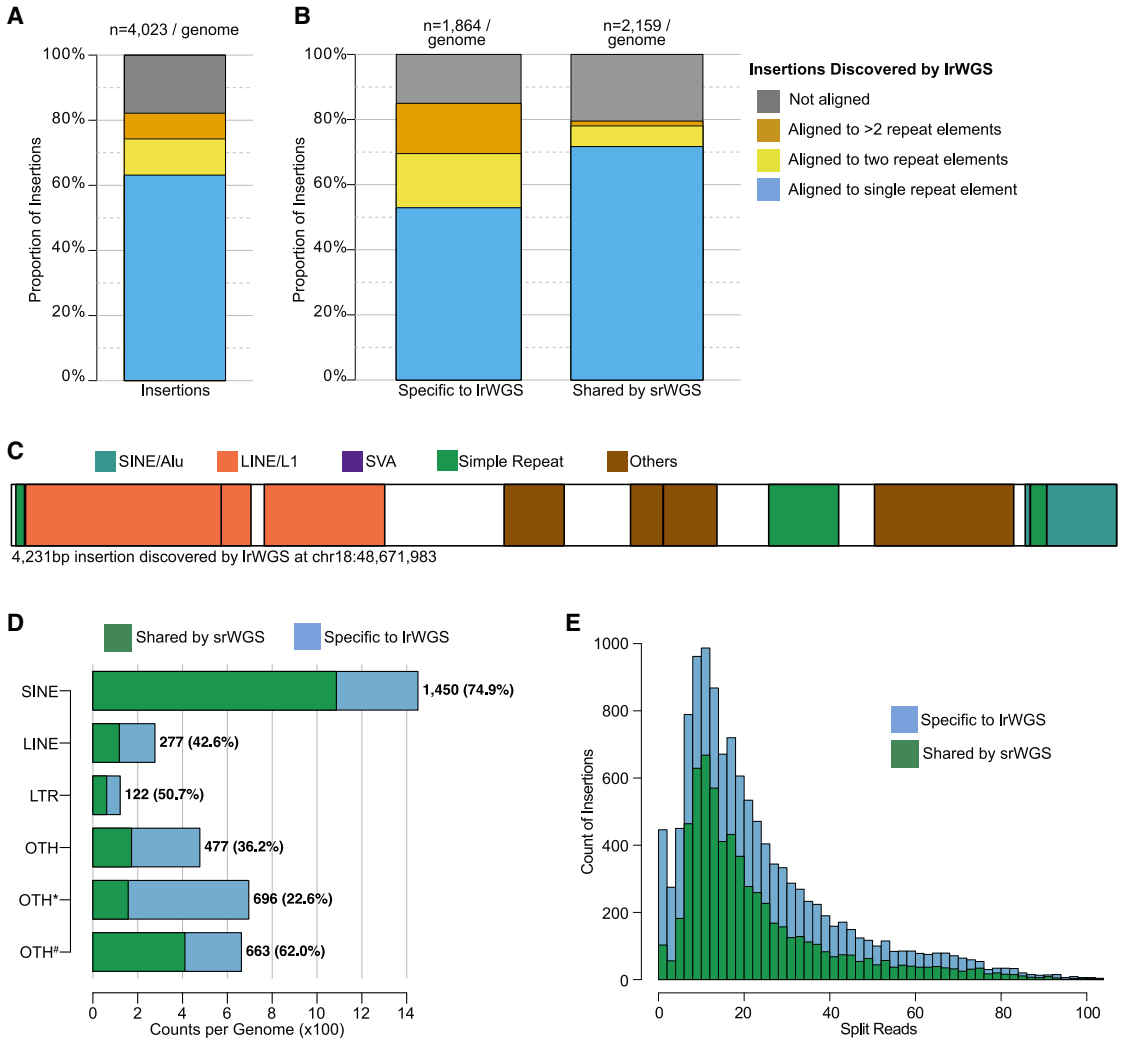

Figure 3. Alignment of assembled lrWGS insertion sequences against known repeat elements

(A) Count of lrWGS insertions in "unique + RM" sequences per genome by alignment of inserted sequences to known repeat elements. The number on top of the bar represents the averaged count of high-confidence insertions in "unique + RM" sequences per genome.

(B) Count of lrWGS insertions that are specifically discovered by lrWGS and shared by srWGS, by alignment of inserted sequences to known repeat elements. Formatting conventions are the same as in (A).

(C) An example of an insertion SV assembled by lrWGS, annotated with sequences that align to known repeat element classes. White shading represents sequences not annotated as a known repeat element.

(D) Counts of lrWGS insertions in "unique + RM" sequences per genome by the class of inserted sequence and the proportion that was overlapped by srWGS. "OTH*" represents insertions aligned to multiple known repeat elements, such as the example shown in (B). "OTH#" stands for insertions that were not aligned to any repeat elements. Numbers in parentheses represent the proportion of insertions that were overlapped by srWGS.

(E) Count of split reads around the lrWGS high-confidence insertions in histogram.

Finally, we explored the concordance of SV detection for a class of SVs that is strongly enriched for pathogenic variation and appears to be a significant blind spot for long-read assembly technologies: large CNVs captured by depth-based analyses from srWGS. Our initial analyses suggested that lrWGS assembly methods failed to capture all but one of the small number of large (>5 kb) CNVs that could be detected by srWGS read-depth methods in three probands (average size = 14.7 kb). Recognizing the limitation of read-depth analyses to capture large CNVs in a small number of families, we explored CNV calls from 3,202 individuals from our ongoing analyses of 30× srWGS in the 1000 Genomes Project that included all three families used in this study (see HGSVC preprints for complete details).[34,52] We found an average of 167 large CNVs per genome that were exclusively detected by depth-based methods, 88.2% of which were not detected by lrWGS assembly. These findings highlight an important blind spot in variant detection from lrWGS assembly in the absence of depth-based analyses and have significant implications for human disease studies because large CNVs have a profound deleterious impact on a spectrum of human diseases.

In conclusion, we demonstrate the strong influence of genomic context on expectations for SV detection from

srWGS in genomic studies, as well as estimating the anticipated yields of emerging lrWGS technologies. Initial surveys have implied highly variable outcomes and limited overall concordance in SV detection between the two technologies;[14] however, in-depth analyses of these variants emphasize that genome organization, variant type, variant size, and high type I error rates in SV detection from each technology were the predominant features driving discordance. After applying *post hoc* filters to correct for the relatively high type I error rates for SV detection from this ensemble srWGS approach optimized for sensitivity, and the assembly-based lrWGS approach that was optimized with orthogonal data types, we were able to extrapolate the informative genomic features that influence differences in SV distributions between technologies. The concordance between srWGS and lrWGS was remarkably high for deletions localized to the least-repetitive regions of the genome (93.8%), while almost all lrWGS-specific deletions were localized to repetitive SD + SR regions. We observed poor sensitivity in the detection of large CNVs (>5 kb) via lrWGS assemblies by comparison with srWGS, and this limitation is most likely due to the lack of depth-based lrWGS methods. In contrast, lrWGS showed superior sensitivity for detection of insertions regardless of the genomic context, although most (95.8%) insertions in the least-repetitive genomic regions had detectable alignment signatures in the srWGS data, indicating further improvement in insertion discovery methods for srWGS should continue to bridge this disparity in insertion detection between technologies. Variant types other than deletions and insertions (e.g., inversions, translocations, balanced and complex SVs) were excluded from these analyses because they were not uniformly called by lrWGS assemblies, although we expect future improvement in lrWGS methods to provide novel insights into repeat-mediated mechanisms for these variant classes.

The value added for long-read assembly to discover new disease-associated SVs, or to provide resolution to "unsolved" cases in Mendelian genetics research and clinical diagnostics, is thus a complex calculus. As we note above, srWGS captures virtually all high-quality deletions derived from lrWGS assembly in the regions of the genome that encompass over 95% of currently annotated coding sequence in genes with existing evidence for dominant-acting pathogenic mutations from OMIM. We therefore anticipate that a minority of "unsolved" cases will be explained by novel and readily interpretable deletions that can be captured by lrWGS but remain cryptic to srWGS in known disease-associated genes. However, given that the most highly repetitive regions of the genome have been traditionally inaccessible in human disease studies, it is anticipated that new disease-associated genes and sequences will emerge as functional annotation of these repetitive sequences and duplicated genes continues to improve. Indeed, germline and somatic repeat expansions and contractions are already well established mechanisms of human disease, particularly neurodegenerative disorders.[53] As telomere-to-telomere assembly methods continue to mature and eventually reach into centromeres, telomeres, and other highly repetitive regions, the catalog of disease-associated variants will certainly expand beyond what is applied to current clinical interpretation. Moving forward, long-read technologies also offer the opportunity to detect novel transcripts from RNA-seq[54] and methylation status from technologies such as ONT, which will further expand the list of disease-associated variants.[54–56]

Collectively, we estimate from these analyses that future genomic studies and clinical initiatives using srWGS can expect to capture upward of ten to eleven thousand SVs in each human genome, and current large-scale international initiatives are poised to provide exciting new insights into the 90% of the annotated reference genome that encompasses most known genic sequence. Our analyses also confirmed that assembly-based lrWGS methods will access regions of the genome that were previously intractable to conventional technologies and srWGS. We anticipate that advances in lrWGS technologies, and associated analytic approaches, will provide significant long-term value in expanding the catalog of functional variation associated with insertions, mobile elements, and the most challenging sequence features in the human genome.

Data and code availability

Resource data used in this paper were generated by Chaisson et al. (2019)[14]. These data are available under dbVar: nstd152.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.03.014.

Acknowledgments

Data and analyses were conducted by the Human Genome Structural Variation Consortium (HGSVC). Analyses, data, and personnel were supported by the following grants from the National Institutes of Health (NIH): U24HG007497, R01MH115957, R03HD099547, UM1HG008895, R01HD081256, R01HD091797, R01HD096326, R01HG002898, R01HG010169, R00DE026824, GRFP2017240332, and F31HG010569. X.Z. was supported in part by the MGH ECOR Fund for Medical Discovery (FMD) Postdoctoral Fellowship. R.L.C. was supported by NHGRI T32HG002295 and NSF GRFP #2017240332. C. Lee was supported in part by the operational funds from The First Affiliated Hospital of Xi'an Jiaotong University. C. Lee is supported in part by the Ewha Womans University research grant of 2019. E.E.E. is an investigator of the Howard Hughes Medical Institute.

Declaration of interests

The authors declare no competing interests.

Received: June 24, 2020
Accepted: March 12, 2021
Published: March 30, 2021

Web resources

HGSV integration pipeline, https://github.com/xuefzhao/HGSV_SV_integration_pipe
lrWGS data of HGSVC sample, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/20160623_chaisson_pacbio_aligns/
OMIM, https://omim.org/
srWGS data of HGSVC sample, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/data/
svtk, https://github.com/talkowski-lab/svtk
VaPoR, https://github.com/mills-lab/vapor

References

1. Abel, H.J., Larson, D.E., Regier, A.A., Chiang, C., Das, I., Kanchi, K.L., Layer, R.M., Neale, B.M., Salerno, W.J., Reeves, C., et al.; NHGRI Centers for Common Disease Genomics (2020). Mapping and characterization of structural variation in 17,795 human genomes. Nature *583*, 83–89.
2. Posey, J.E., O'Donnell-Luria, A.H., Chong, J.X., Harel, T., Jhangiani, S.N., Coban Akdemir, Z.H., Buyske, S., Pehlivan, D., Carvalho, C.M.B., Baxter, S., et al.; Centers for Mendelian Genomics (2019). Insights into genetics, human biology and disease gleaned from family based genomic studies. Genet. Med. *21*, 798–812.
3. Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., van Kogelenberg, M., King, D.A., Ambridge, K., Barrett, D.M., Bayzetinova, T., et al.; DDD study (2015). Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet *385*, 1305–1314.
4. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290–299.
5. Denny, J.C., Rutter, J.L., Goldstein, D.B., Philippakis, A., Smoller, J.W., Jenkins, G., Dishman, E.; and All of Us Research Program Investigators (2019). The "All of Us" Research Program. N. Engl. J. Med. *381*, 668–676.
6. Rusk, N. (2018). The UK Biobank. Nat. Methods *15*, 1001.
7. Turro, E., Astle, W.J., Megy, K., Gräf, S., Greene, D., Shamardina, O., Allen, H.L., Sanchis-Juan, A., Frontini, M., Thys, C., et al.; NIHR BioResource for the 100,000 Genomes Project (2020). Whole-genome sequencing of patients with rare diseases in a national health system. Nature *583*, 96–102.
8. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A.; and 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
9. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
10. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
11. Regier, A.A., Farjoun, Y., Larson, D.E., Krasheninina, O., Kang, H.M., Howrigan, D.P., Chen, B.-J., Kher, M., Banks, E., Ames, D.C., et al. (2018). Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat. Commun. *9*, 4038.
12. Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K., et al.; 1000 Genomes Project (2011). Mapping copy number variation by population-scale genome sequencing. Nature *470*, 59–65.
13. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H., et al.; 1000 Genomes Project Consortium (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.
14. Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat. Commun. *10*, 1784.
15. Audano, P.A., Sulovari, A., Graves-Lindsay, T.A., Cantsilieris, S., Sorensen, M., Welch, A.E., Dougherty, M.L., Nelson, B.J., Shah, A., Dutcher, S.K., et al. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. Cell *176*, 663–675.e19.
16. Werling, D.M., Brand, H., An, J.-Y., Stone, M.R., Zhu, L., Glessner, J.T., Collins, R.L., Dong, S., Layer, R.M., Markenscoff-Papadimitriou, E., et al. (2018). An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat. Genet. *50*, 727–736.
17. Chiang, C., Scott, A.J., Davis, J.R., Tsang, E.K., Li, X., Kim, Y., Hadzic, T., Damani, F.N., Ganel, L., Montgomery, S.B., et al.; GTEx Consortium (2017). The impact of structural variation on human gene expression. Nat. Genet. *49*, 692–699.
18. Brandler, W.M., Antaki, D., Gujral, M., Kleiber, M.L., Whitney, J., Maile, M.S., Hong, O., Chapman, T.R., Tan, S., Tandon, P., et al. (2018). Paternally inherited cis-regulatory structural variants are associated with autism. Science *360*, 327–331.
19. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al.; Genome Aggregation Database Production Team; and Genome Aggregation Database Consortium (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451.
20. Turner, T.N., Hormozdiari, F., Duyzend, M.H., McClymont, S.A., Hook, P.W., Iossifov, I., Raja, A., Baker, C., Hoekzema, K., Stessman, H.A., et al. (2016). Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. Am. J. Hum. Genet. *98*, 58–74.
21. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
22. Berg, J.S., Adams, M., Nassar, N., Bizon, C., Lee, K., Schmitt, C.P., Wilhelmsen, K.C., and Evans, J.P. (2013). An informatics approach to analyzing the incidentalome. Genet. Med. *15*, 36–44.

23. Blekhman, R., Man, O., Herrmann, L., Boyko, A.R., Indap, A., Kosiol, C., Bustamante, C.D., Teshima, K.M., and Przeworski, M. (2008). Natural selection on genes that underlie human disease susceptibility. Curr. Biol. *18*, 883–889.
24. Rhoads, A., and Au, K.F. (2015). PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics *13*, 278–289.
25. Jain, M., Olsen, H.E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. *17*, 239.
26. Pendleton, M., Sebra, R., Pang, A.W.C., Ummat, A., Franzen, O., Rausch, T., Stütz, A.M., Stedman, W., Anantharaman, T., Hastie, A., et al. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat. Methods *12*, 780–786.
27. Chaisson, M.J., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. Nature *517*, 608–611.
28. Cretu Stancu, M., van Roosmalen, M.J., Renkens, I., Nieboer, M.M., Middelkamp, S., de Ligt, J., Pregno, G., Giachino, D., Mandrile, G., Espejo Valle-Inclan, J., et al. (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat. Commun. *8*, 1326.
29. Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature *585*, 79–84.
30. Sanders, A.D., Falconer, E., Hills, M., Spierings, D.C.J., and Lansdorp, P.M. (2017). Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. Nat. Protoc. *12*, 1151–1176.
31. Chan, S., Lam, E., Saghbini, M., Bocklandt, S., Hastie, A., Cao, H., Holmlin, E., and Borodkin, M. (2018). Structural Variation Detection and Analysis Using Bionano Optical Mapping. Methods Mol. Biol. *1833*, 193–203.
32. Zook, J.M., Hansen, N.F., Olson, N.D., Chapman, L., Mullikin, J.C., Xiao, C., Sherry, S., Koren, S., Phillippy, A.M., Boutros, P.C., et al. (2020). A robust benchmark for detection of germline large deletions and insertions. Nat. Biotechnol. *38*, 1347–1355.
33. Eichler, E.E. (2019). Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. N. Engl. J. Med. *381*, 64–74.
34. Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science *eabf7117*.
35. Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H.P., Bjornsson, E., Jonsson, H., Atlason, B.A., Kristmundsdottir, S., Mehringer, S., Hardarson, M.T., et al. (2020). Long read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. bioRxiv. https://doi.org/10.1101/848366.
36. Rodriguez, O.L., Ritz, A., Sharp, A.J., and Bashir, A. (2020). MsPAC: a tool for haplotype-phased structural variant detection. Bioinformatics *36*, 922–924.
37. van Berkum, N.L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L.A., Dekker, J., and Lander, E.S. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. J. Vis. Exp. (39), 1869.
38. Sanders, A.D., Hills, M., Porubský, D., Guryev, V., Falconer, E., and Lansdorp, P.M. (2016). Characterizing polymorphic inversions in human genomes by single-cell sequencing. Genome Res. *26*, 1575–1587.
39. Zhang, F., Khajavi, M., Connolly, A.M., Towne, C.F., Batish, S.D., and Lupski, J.R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat. Genet. *41*, 849–853.
40. Monlong, J., Cossette, P., Meloche, C., Rouleau, G., Girard, S.L., and Bourque, G. (2018). Human copy number variants are enriched in regions of low mappability. Nucleic Acids Res. *46*, 7236–7249.
41. Tattini, L., D'Aurizio, R., and Magi, A. (2015). Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Front. Bioeng. Biotechnol. *3*, 92.
42. Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., and Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. *20*, 117.
43. de Koning, A.P., Gu, W., Castoe, T.A., Batzer, M.A., and Pollock, D.D. (2011). Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. *7*, e1002384.
44. Samonte, R.V., and Eichler, E.E. (2002). Segmental duplications and the evolution of the primate genome. Nat. Rev. Genet. *3*, 65–72.
45. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.
46. Kuhn, R.M., Haussler, D., and Kent, W.J. (2013). The UCSC genome browser and associated tools. Brief. Bioinform. *14*, 144–161.
47. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat. Genet. *46*, 944–950.
48. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al.; Genome Aggregation Database Consortium (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.
49. Petrovski, S., Gussow, A.B., Wang, Q., Halvorsen, M., Han, Y., Weir, W.H., Allen, A.S., and Goldstein, D.B. (2015). The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. PLoS Genet. *11*, e1005492.
50. Zhao, X., Weber, A.M., and Mills, R.E. (2017). A recurrence-based approach for validating structural variation using long-read sequencing technology. Gigascience *6*, 1–9.
51. Zhou, W., Emery, S.B., Flasch, D.A., Wang, Y., Kwan, K.Y., Kidd, J.M., Moran, J.V., and Mills, R.E. (2020). Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. Nucleic Acids Res. *48*, 1146–1163.
52. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2021). High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. bioRxiv. https://doi.org/10.1101/2021.02.06.430068.

53. Gatchel, J.R., and Zoghbi, H.Y. (2005). Diseases of unstable repeat expansion: mechanisms and common principles. Nat. Rev. Genet. *6*, 743–755.
54. Uapinyoying, P., Goecks, J., Knoblach, S.M., Panchapakesan, K., Bonnemann, C.G., Partridge, T.A., Jaiswal, J.K., and Hoffman, E.P. (2020). A long-read RNA-seq approach to identify novel transcripts of very large genes. Genome Res. *30*, 885–897.
55. Gigante, S., Gouil, Q., Lucattini, A., Keniry, A., Beck, T., Tinning, M., Gordon, L., Woodruff, C., Speed, T.P., Blewitt, M.E., and Ritchie, M.E. (2019). Using long-read sequencing to detect imprinted DNA methylation. Nucleic Acids Res. *47*, e46.
56. Gouil, Q., and Keniry, A. (2019). Latest techniques to study DNA methylation. Essays Biochem. *63*, 639–648.

The American Journal of Human Genetics, Volume *108*

Supplemental information

Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies

Xuefang Zhao, Ryan L. Collins, Wan-Ping Lee, Alexandra M. Weber, Yukyung Jun, Qihui Zhu, Ben Weisburd, Yongqing Huang, Peter A. Audano, Harold Wang, Mark Walker, Chelsea Lowther, Jack Fu, Human Genome Structural Variation Consortium, Mark B. Gerstein, Scott E. Devine, Tobias Marschall, Jan O. Korbel, Evan E. Eichler, Mark J.P. Chaisson, Charles Lee, Ryan E. Mills, Harrison Brand, and Michael E. Talkowski

Supplemental Figures and Legends

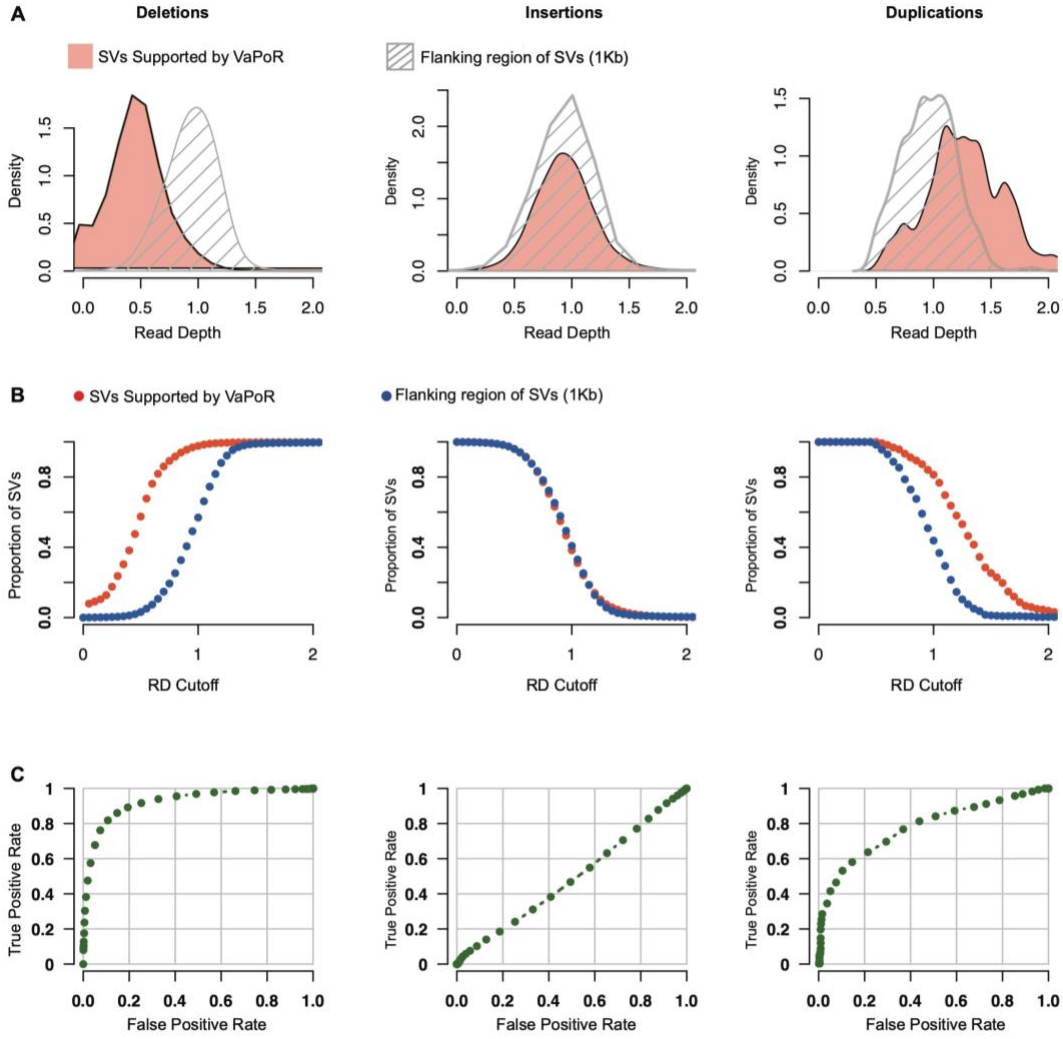

Figure S1. Distribution of normalized RD of srWGS for SVs supported by long reads.

(A) Distribution of normalized RD for deletions (left), insertions (middle) and duplications (right) that were supported by VaPoR (red) and the 1Kb flanking regions of these SVs (grey).

(B) Rate of true positive (red) and false positive (blue) of deletions (left), insertions (middle) and duplications (right) at different cutoffs of RD

(C) Receiver operating characteristic (ROC) of RD for deletions (left), insertions (middle) and duplications (right).

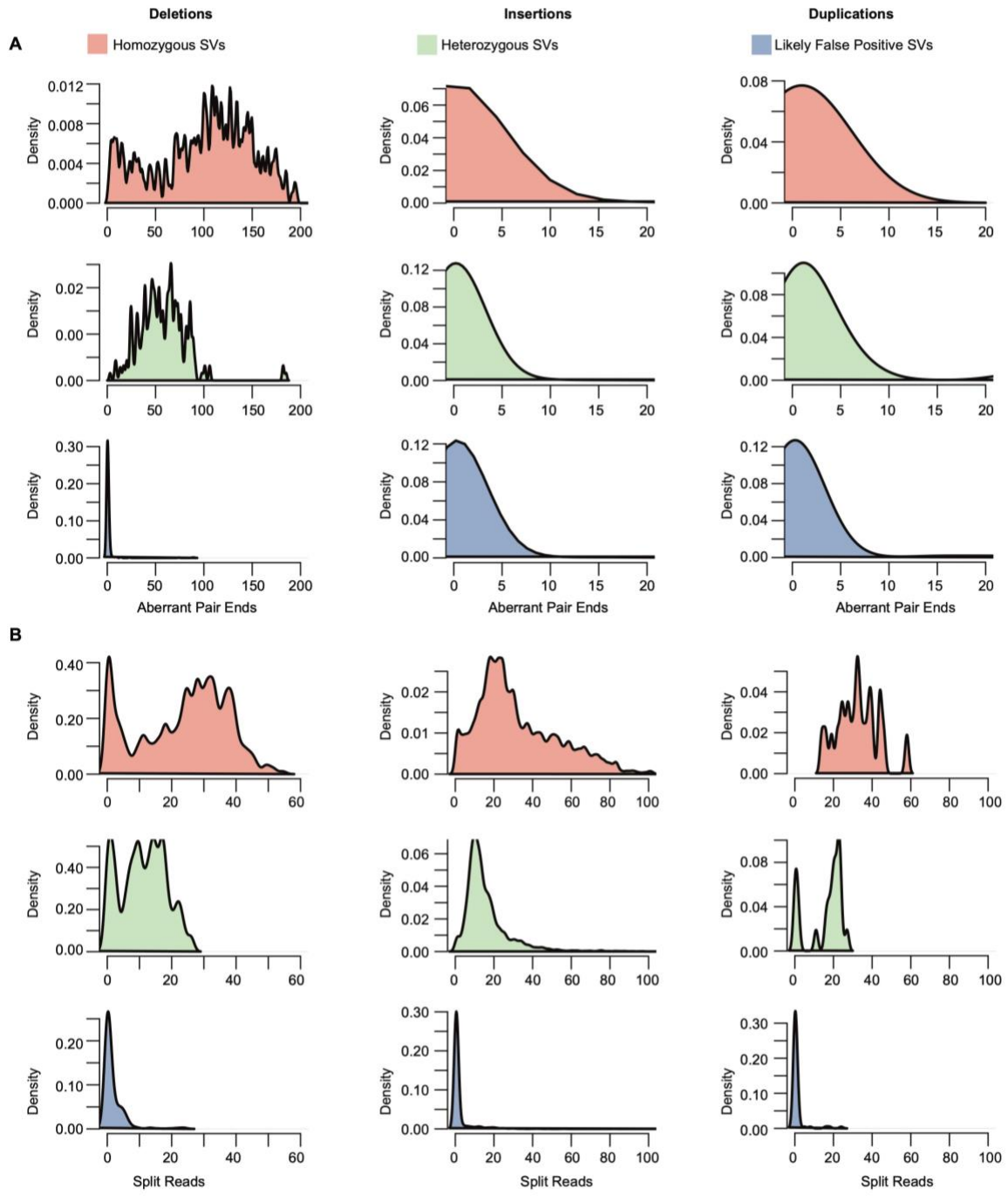

Figure S2. Distribution of aberrant PE and SRs from srWGS across high-confidence homozygous SVs, heterozygous SVs and likely false positive SVs.

Distributions of (A) PE and (B) SRs metrics for homozygous (red), heterozygous (green) and false positive (blue) SVs for deletions (left), insertions (middle), and duplications (right).

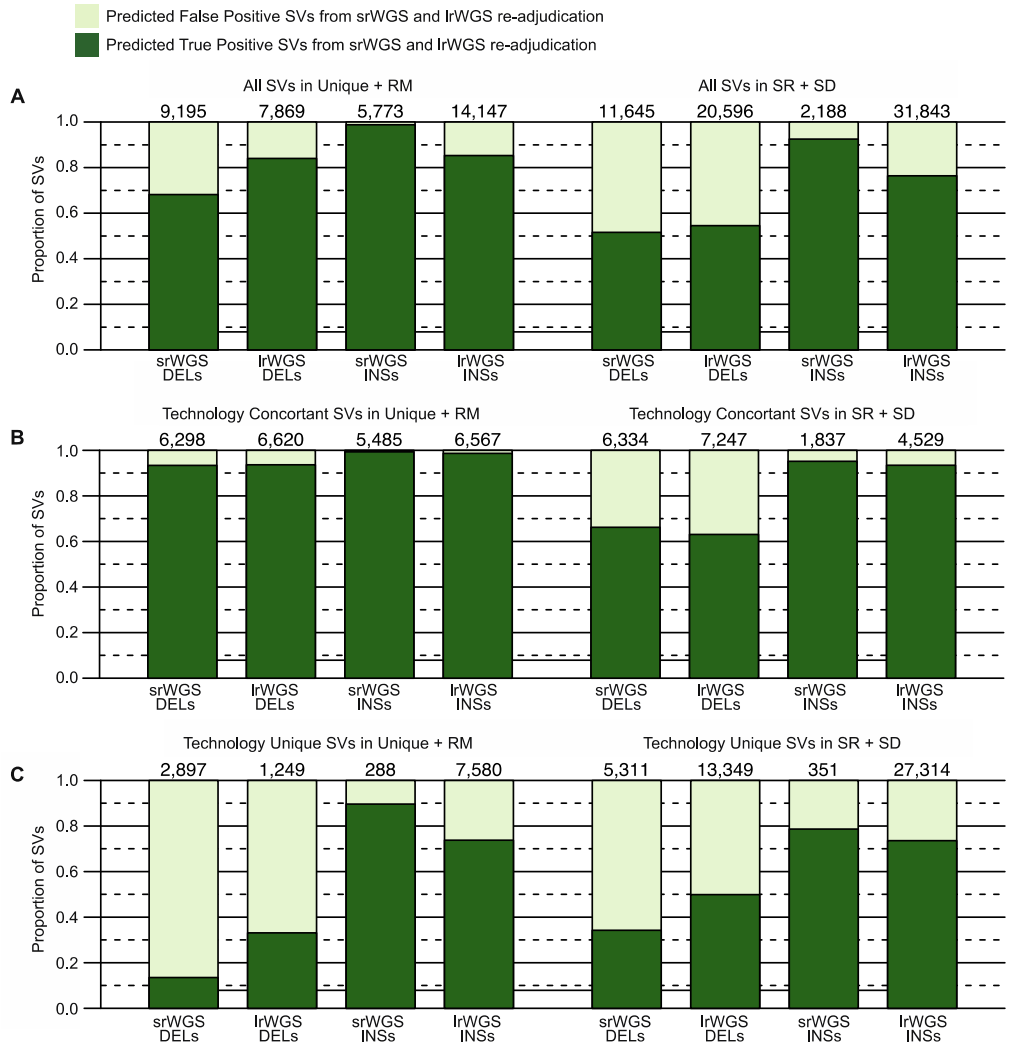

Figure S3. Proportion of SVs supported based on srWGS and lrWGS re-adjudication.

(A) Proportion of Deletions (DELs) and Insertions (INSs) that were supported by the *in silico* re-adjudication procedure.

(B) Proportion of SVs concordant between technologies that were supported by the *in silico* re-adjudication procedure.

(C) Proportion of SVs uniquely discovered by srWGS or lrWGS that were supported by the *in silico* re-adjudication procedure.

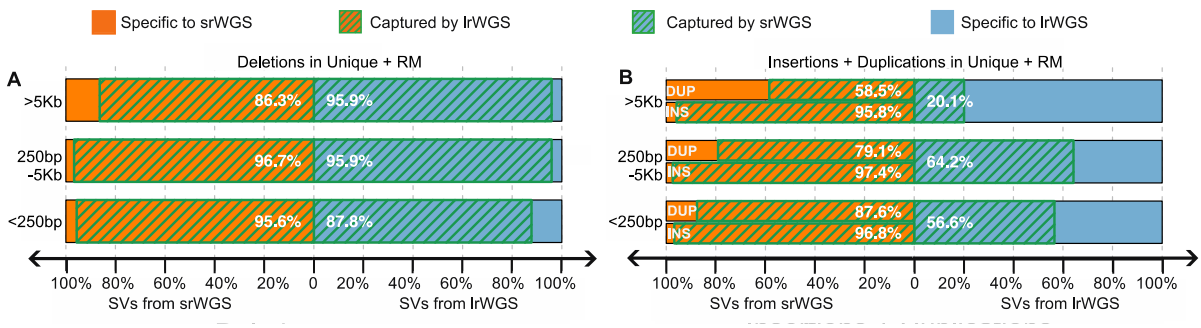

Figure S4. Concordance of SVs of different sizes between srWGS and lrWGS in Unique and RM sequences.

(A-B) Concordance of (A) deletions and (B) insertions and duplications in Unique + RM sequences that were supported by the *in silico* SV refinement procedure at different SV size ranges. Percentages represent the fraction of total variants shared between srWGS and lrWGS. Letters in panel B represent the type of srWGS SVs, DUP – duplications, INS – insertions.

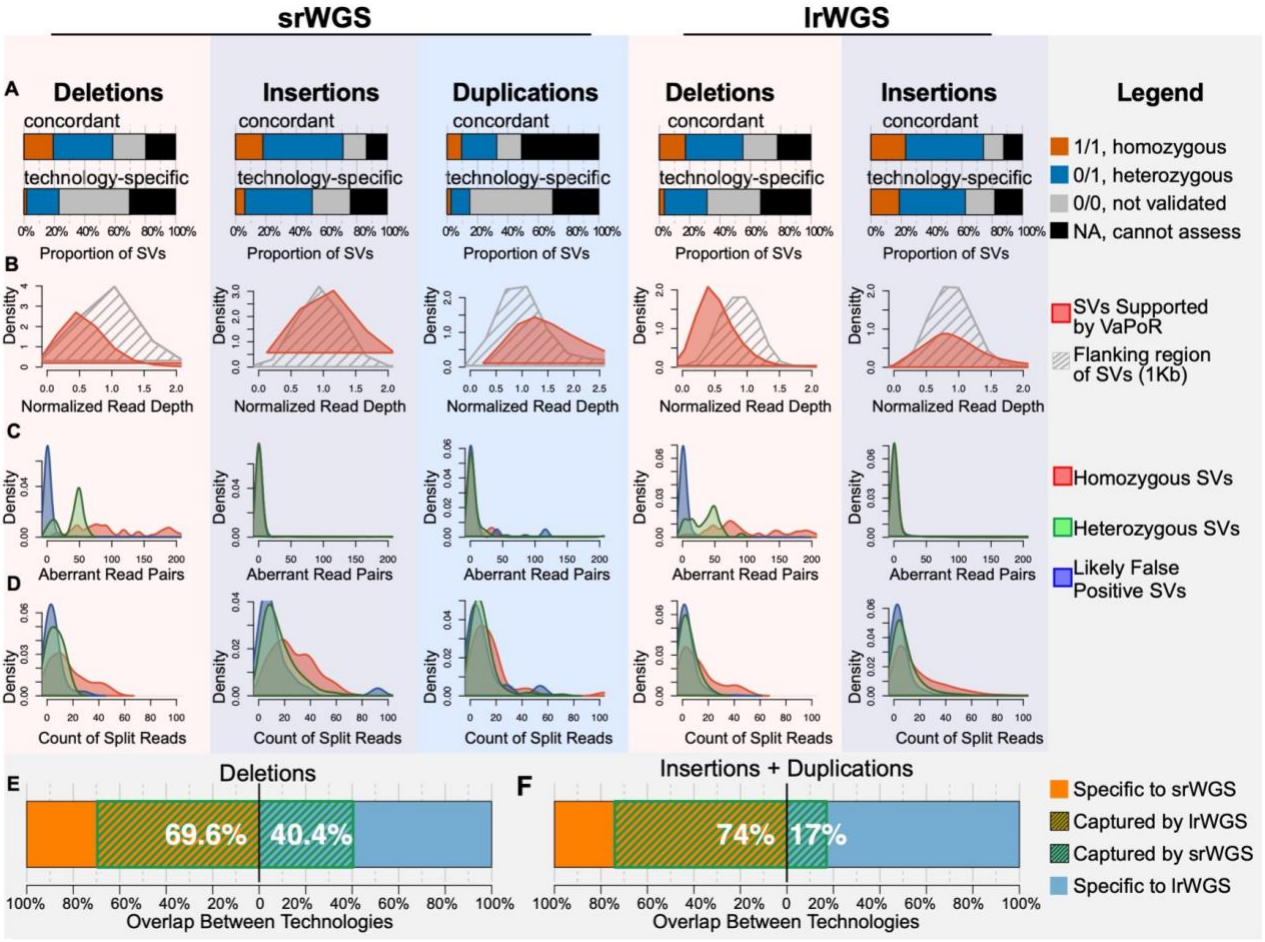

Figure S5. Recalibrating SVs in the repetitive SD + SR sequences based on read-level alignment signatures and concordance of the high-confidence SVs with supports between srWGS and lrWGS.

(A) *In silico* evaluation results from VaPoR on deletions (left), insertions (middle) and duplications (right). Deletions and insertions were reported in both srWGS and lrWGS callsets, and duplications were only reported in srWGS callset.

(B) Distribution of normalized read depth of srWGS across deletions (left), insertions (middle) and duplications (right) that were supported by VaPoR (red), and the 1Kb genomic regions that flank each SV (grey).

(C-D) Distribution of (C) aberrant srWGS read pairs and (D) split reads around deletions (left), insertions (middle) and duplications (right) that were either homozygous (red), heterozygous (green) or likely false positives (blue). The homozygous, heterozygous and likely false positive SV sets were selected using the criteria described in supplemental methods.

(E-F) Concordance of (E) deletions and (F) insertions and duplications in SD + SR sequences that were supported by the *in silico* SV refinement procedure. Percentages represent the fraction of total variants shared between srWGS and lrWGS.

Supplemental Tables

Table S1. Expected and observed counts of SVs located within SD + SR, and in Unique + RM sequences.

	srWGS		lrWGS	
	SR + SD	Unique + RM	SR + SD	Unique + RM
Expected	1056	9828	2408	22417
Observed	5259	5625	17483	7342

Table S2. Count and proportion of SVs per genome.

	srWGS			lrWGS	
	DEL	DUP	INS	DEL	INS
Assessable by VaPoR	5,878(84.6%)	563(71.3%)	2,467(93.0%)	7,345	13,216
Validated by VaPoR	3,644(62.0%)	227(40.3%)	2,150(87.1%)	4,789(65.2%)	10,318(78.1%)
Supported by RD	1,265(18.2%)	336(42.6%)	NA	2,293(24.2%)	NA
Supported by PE/SRs	1,175(16.9%)	126(16.0%)	624(23.5%)	1,470(15.5%)	2,192(14.3%)
all SVs / genome	6,947	789	2,654	9,488	15,330

Note: Numbers in parenthesis represent the proportion of variants supported by the corresponding evidence; the VaPoR validation rate is calculated based on the SVs that were assessable by VaPoR.

Table S3. PE and SRs cutoffs selected to discriminate the quality of SVs from lrWGS and srWGS callsets.

	Selected Thresholds		SVs Passing Thresholds in each Training Set (%)		
SV Type	PE	SRs	Homozygous	Heterozygous	Predicted Type I Errors
DEL	15	0	92.59%	97.52%	0.95%
DUP	0	19	73.81%	57.14%	1.22%
INS	0	30	42.87%	8.25%	0.94%

Supplemental Material and Methods

Samples, sequencing, and Structural Variation (SV) discovery

In this study, we evaluated three parent-child trios from the 1000 Genomes Project that have been recently analyzed for SVs with both short-read whole genome sequencing (srWGS) and long-read whole genome sequencing (lrWGS) in the Human Genome Structural Variation Consortium (HGSVC).[1] These trios were derived from Han Chinese (CHS), Puerto Rican (PUR) and Yoruban Nigerian (YRI) ancestry groups. The HGSVC generated srWGS and lrWGS data and corresponding SV callsets on these samples, which we used in this study. For srWGS, samples were sequenced with Illumina HiSeq 2500 to ~74.5X coverage per genome, and SVs were discovered using an ensemble approach that integrated 13 independent SV discovery algorithms (WHAMG,[2] LUMPY,[3] DELLY,[4] ForestSV,[5] Manta,[6] Pindel,[7] SVelter,[8] novoBreak,[9] MELT,[10] VariationHunter,[11] dCGH,[12] GenomeSTRiP,[13] Tardis[14]). The callsets from different algorithms were combined based on breakpoint overlap and concordance with orthogonal technology. In brief, SVs from each srWGS algorithms were compared against lrWGS calls by requiring matching SV type and a minimum of 50% reciprocal overlap. Distances between breakpoints of srWGS SVs and their matching lrWGS SVs were collected to form a distribution, and 95% confidence interval (CI) of this distribution was calculated to represent the precision range of the algorithm (Supp Fig 10, 11 of Chaisson et al[1]). Overlapping SVs from different algorithms were merged into a consensus SV call if the CI of their breakpoints overlapped. For lrWGS, samples were sequenced with Pacific Biosciences RS II to ~20.0X in the parental genomes and ~39.6X in the child genomes, and SVs were discovered using the integration of two genome assembly-based methods (Phased-SV and MS-PAC[1, 15, 16]). From the Chaisson et al. data[1] we combined srWGS duplications with

insertions for sake of comparisons to lrWGS, which did not distinguish between insertions and duplications.

SV annotation by repeat content

We defined genomic repeat content to include Segmental Duplication (SD), Simple Repeat (SR) and other Repeat Masked regions (RM) based on GRCh38 annotations downloaded from the UCSC genome browser (https://genome.ucsc.edu; version 2018-08-10[17]). Regions in the RM track that overlapped any SR or SD elements were excluded from RM to avoid conflicting repeat types. Genomic regions falling outside of RM, SR and SD were annotated as "Unique" genomic sequences. We annotated SVs by first allocating their breakpoints to one of the repeat content classes and assigned each SV to one repeat category by prioritizing SR, followed by SD, RM and then Unique sequences, thus prioritizing overlap with annotated repeat sequences.

Statistical test of SV distribution across genomic context

We tested the distribution of SVs across different genomic context against the null hypothesis that SVs are evenly distributed across the genome regardless of the genomic context. Under the null hypothesis 1,056 of the 10,884 SVs from srWGS and 2,408 of the 24,825 SVs from lrWGS were expected in the highly repetitive SD and SR regions that consist 9.7% of the genome, while 5,259 and 17,483 were observed in these regions from srWGS and lrWGS respectively. A chi-square test was performed (Table S1) to test the significance of observation against expectation.

Comparison of SVs between technologies

We applied different criteria to SVs based on variant class to assess concordance between srWGS and lrWGS. We considered deletions to be concordant if over 50% reciprocal overlap of the SV was observed between technologies. Insertions were considered concordant between srWGS and lrWGS if their predicted insertion sites were within 100 bp and the lengths of their inserted sequences were within 10 times of each other. As the lrWGS callset did not differentiate duplications from insertions, we also compared lrWGS insertions to srWGS duplications by either 1) requiring >50% of the inserted sequences of lrWGS insertions to be covered by srWGS duplications or 2) requiring >50% reciprocal overlap between the srWGS duplication coordinates and the alignments of assembled lrWGS insertion sequences against the human reference genome (GRCh38) with BLAT(v35).[18] Finally, given that SVs were strictly defined as ≥50bp in the original srWGS and lrWGS SV callsets, we avoided biasing our comparisons near the 50bp size threshold by including small insertions and deletions (indels) defined by both technologies that were between 30-50bp when assessing SV concordance.

Evaluation and adjudication of SVs

We designed an *in silico* re-adjudication procedure and applied it to all SVs to reduce the type I error rate of the original SV callsets. We examined orthogonal support from both lrWGS and srWGS data to quantify strength of evidence for each SV. These analyses are described below.

First, to assess raw lrWGS evidence supporting each SV, we applied VaPoR,[19] an algorithm designed to evaluate SV predictions by directly comparing lrWGS sequences with a reference genome through recurrence plots. We executed VaPoR with default settings and considered SVs with a positive genotype score (VaPoR GS >0) as having lrWGS support (Figure 2A, S5A). In

order to maximize validation power, we also examined each SV in the parental lrWGS genomes (20.0X) with VaPoR and considered SV support in parent as valid. VaPoR is unable to make an evaluation in certain regions of the genome due to nearby sequence homology or low coverage. After taking this limitation into account, we were able to assess 85.4% and 82.8% SVs from srWGS and lrWGS respectively. Validation rates of 67.6% (N= 6,021/ genome) and 73.5% (N= 15,107/genome) were achieved for SVs from srWGS and lrWGS respectively (Table S2).

Second, for srWGS data, we focused on three SV signatures: normalized read depth (RD), aberrant paired-end reads (PE), and split reads (SRs). RD represents the copy state of a genomic region as relative to expected copy ratio of 1 (i.e. RD<1 indicates copy loss, and RD>1 indicates copy gain). We collected RD, PE and SRs evidence per sample using the software package *svtk*.[20] For each SV in Unique + RM sequences, we assessed RD spanning the SV and RD of the 1Kb regions flanking the SV. We next trained an SV classifier using RD values from deletions that were supported by VaPoR as compared to their flanking RD values. For a given RD threshold, we defined the false discovery rate (FDR) as the proportion of flanking regions that had a lower RD threshold and defined the true positive rate (TPR) as the proportion of VaPoR-supported deletions that had a lower RD threshold. We selected a conservative RD cutoff for deletions at 0.35 copy state to keep FDR below 1% (Figure S1) with an understanding that this cutoff is optimal for rescuing high-confidence deletions misinterpreted by VaPoR despite excluding most heterozygous deletions. We applied the same method to duplications and calculated an optimal cutoff of 1.60. As expected, RD did not differentiate insertions from their flanking regions (Figure S1), and thus we did not consider RD when filtering insertions. Overall 18.2% srWGS deletions (N=1,265 /

genome) and 43.6% (N= 336 / genome) srWGS duplications were supported by RD evidences (Table S2).

We similarly determined srWGS PE and SRs thresholds to distinguish likely true SVs from false positives (Figure S2). We collected counts of PE reads that were within 100 bp of each breakpoint of an SV and collected SRs counts within 50bp from each SV breakpoint. For SVs with more than one breakpoint, we used the minimum PE and SRs counts for that SV. Like RD, we designed a classification model as follows. We first generated three training groups: high-confidence homozygous SVs, high-confidence heterozygous SVs, and likely false positive SVs. We defined high-confidence homozygous deletions as those genotyped as homozygous alternative (1/1) by VaPoR, had VaPoR support in both parental genomes, and had RD of 0. High-confidence heterozygous deletions were genotyped as heterozygous (0/1) by VaPoR, had VaPoR support in only one parental genome, and had RD between 0.45 and 0.5. Likely false positive deletions did not have VaPoR support in any genomes in the trio, had RD >1, and were labeled as *de novo* in the original callset for SVs from srWGS. High-confidence homozygous duplications had RD >1.6, were genotyped as 1/1 by VaPoR, and had VaPoR support in both parental genomes. High-confidence heterozygous duplications displayed RD>1.6, were genotyped as 0/1 by VaPoR, and had VaPoR support in only one parental genome. Finally, duplications lacking VaPoR support in all trio genomes, had RD <1.6 and were labeled as *de novo* in the original srWGS callset were considered likely false positives. For insertions, we relied solely on VaPoR results to define srWGS training sets. We defined high-confidence homozygous insertions as those genotyped as homozygous by VaPoR and had support in both parental genomes. We defined high-confidence heterozygous insertions as those genotyped as heterozygous by VaPoR and had VaPoR support in

only one parental genome. Finally, we defined likely false-positive insertions as those without VaPoR support in any genomes in the trio (Figure S1).

After identifying the SV subsets defined above for srWGS PE/SRs classifier training, we assessed a range of potential thresholds for PE and SRs to seek optimal values for each type of SVs by restricting the FDR to <1%, defined as the proportion of likely false-positive SVs that have more PE and SRs support than the selected threshold, while maximizing the TPR, defined as proportion of high-confidence homozygous and heterozygous SVs that have higher PE and SRs support. As shown in Table S3, we selected PE and SRs thresholds of 15 and 0, respectively, for deletions, resulting in FPR of 0.95% and TPR of 92.59% for homozygous and 97.52% for heterozygous deletions observed. 16.9% and 15.5% deletions from srWGS and lrWGS respectively were supported by PE/SRs evidences with these thresholds. Comparable results are displayed for duplications and insertions in Table S3.

Supplemental References

1. Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun 10, 1784.
2. Kronenberg, Z.N., Osborne, E.J., Cone, K.R., Kennedy, B.J., Domyan, E.T., Shapiro, M.D., Elde, N.C., and Yandell, M. (2015). Wham: Identifying Structural Variants of Biological Consequence. PLoS Comput Biol 11, e1004572.
3. Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 15, R84.
4. Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28, i333-i339.
5. Michaelson, J.J., and Sebat, J. (2012). forestSV: structural variant discovery through statistical learning. Nat Methods 9, 819-821.
6. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32, 1220-1222.

7. Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25, 2865-2871.
8. Zhao, X., Emery, S.B., Myers, B., Kidd, J.M., and Mills, R.E. (2016). Resolving complex structural genomic rearrangements using a randomized approach. Genome Biol 17, 126.
9. Chong, Z., Ruan, J., Gao, M., Zhou, W., Chen, T., Fan, X., Ding, L., Lee, A.Y., Boutros, P., Chen, J., et al. (2017). novoBreak: local assembly for breakpoint detection in cancer genomes. Nat Methods 14, 65-67.
10. Gardner, E.J., Lam, V.K., Harris, D.N., Chuang, N.T., Scott, E.C., Pittard, W.S., Mills, R.E., and Devine, S.E. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res 27, 1916-1929.
11. Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., Eichler, E.E., and Sahinalp, S.C. (2010). Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics 26, i350-357.
12. Sudmant, P.H., Mallick, S., Nelson, B.J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B.P., Baker, C., Nordenfelt, S., Bamshad, M., et al. (2015). Global diversity, population stratification, and selection of human copy-number variation. Science 349, aab3761.
13. Handsaker, R.E., Van Doren, V., Berman, J.R., Genovese, G., Kashin, S., Boettger, L.M., and McCarroll, S.A. (2015). Large multiallelic copy number variations in humans. Nat Genet 47, 296-303.
14. Soylev, A., Kockan, C., Hormozdiari, F., and Alkan, C. (2017). Toolkit for automated and rapid discovery of structural variants. Methods 129, 3-7.
15. Pendleton, M., Sebra, R., Pang, A.W.C., Ummat, A., Franzen, O., Rausch, T., Stütz, A.M., Stedman, W., Anantharaman, T., Hastie, A., et al. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods 12, 780-786.
16. Rodriguez, O.L., Ritz, A., Sharp, A.J., and Bashir, A. (2020). MsPAC: a tool for haplotype-phased structural variant detection. Bioinformatics 36, 922-924.
17. Kuhn, R.M., Haussler, D., and Kent, W.J. (2013). The UCSC genome browser and associated tools. Brief Bioinform 14, 144-161.
18. Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res 12, 656-664.
19. Zhao, X., Weber, A.M., and Mills, R.E. (2017). A recurrence-based approach for validating structural variation using long-read sequencing technology. Gigascience 6, 1-9.
20. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature 581, 444-451.