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

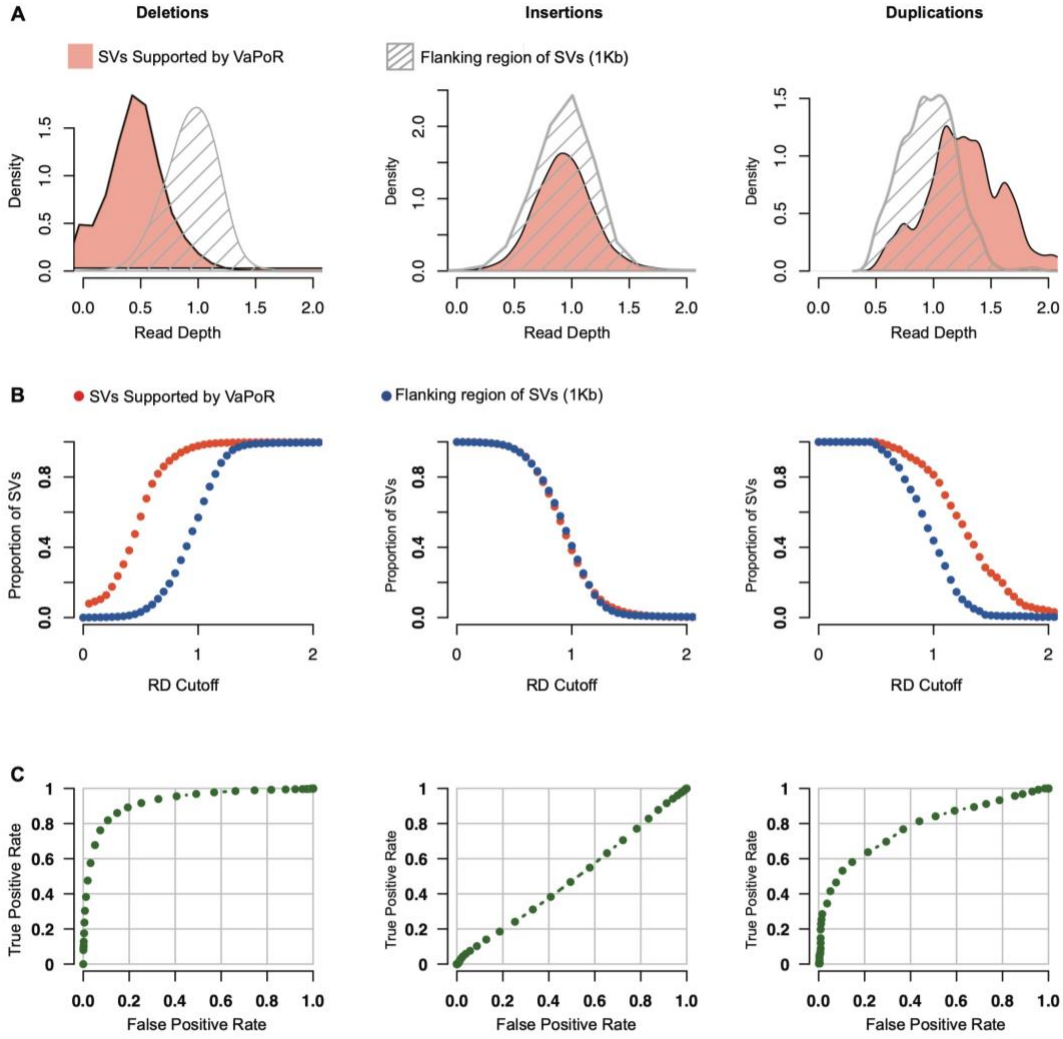

Figure S1. Distribution of normalized RD of srWGS for SVs supported by long reads.

(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).

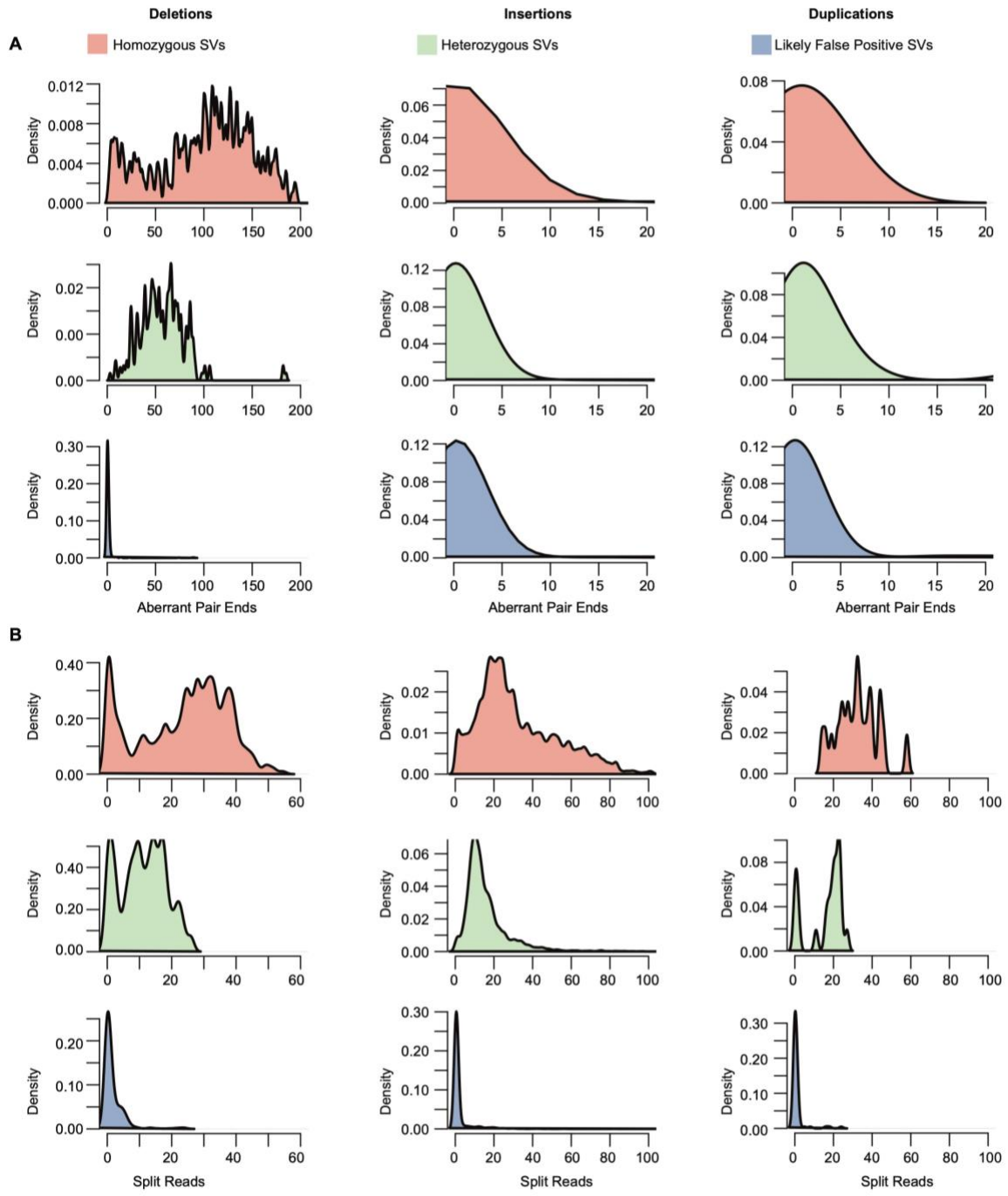

Figure S2. Distribution of aberrant PE and SRs from srWGS across high-confidence homozygous SVs, heterozygous SVs and likely false positive SVs.

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).

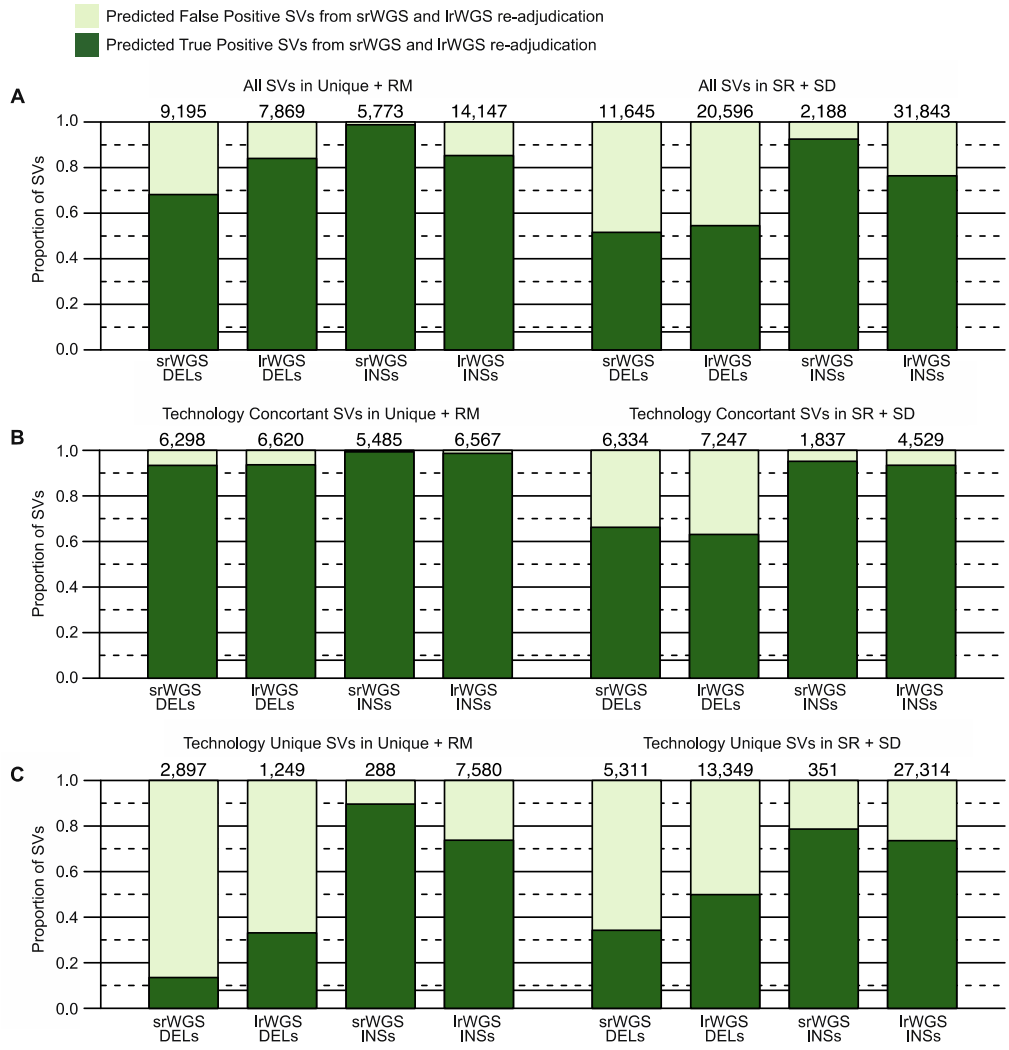

Figure S3. Proportion of SVs supported based on srWGS and lrWGS re-adjudication.

(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.

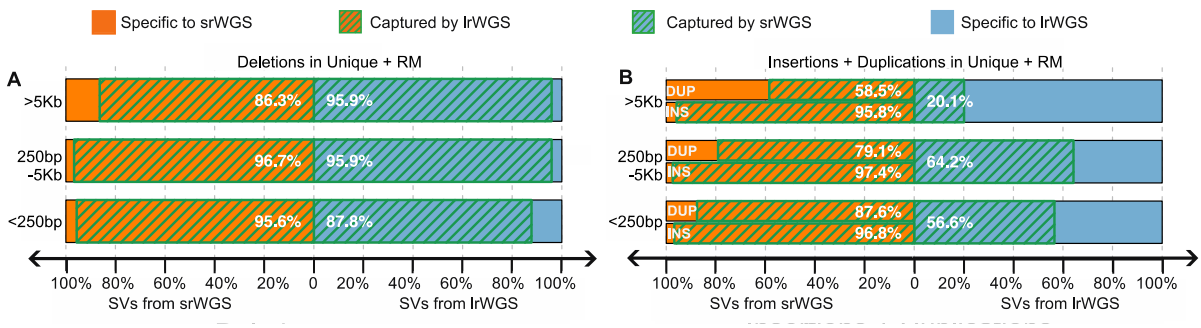

Figure S4. Concordance of SVs of different sizes between srWGS and lrWGS in Unique and RM sequences.

(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.

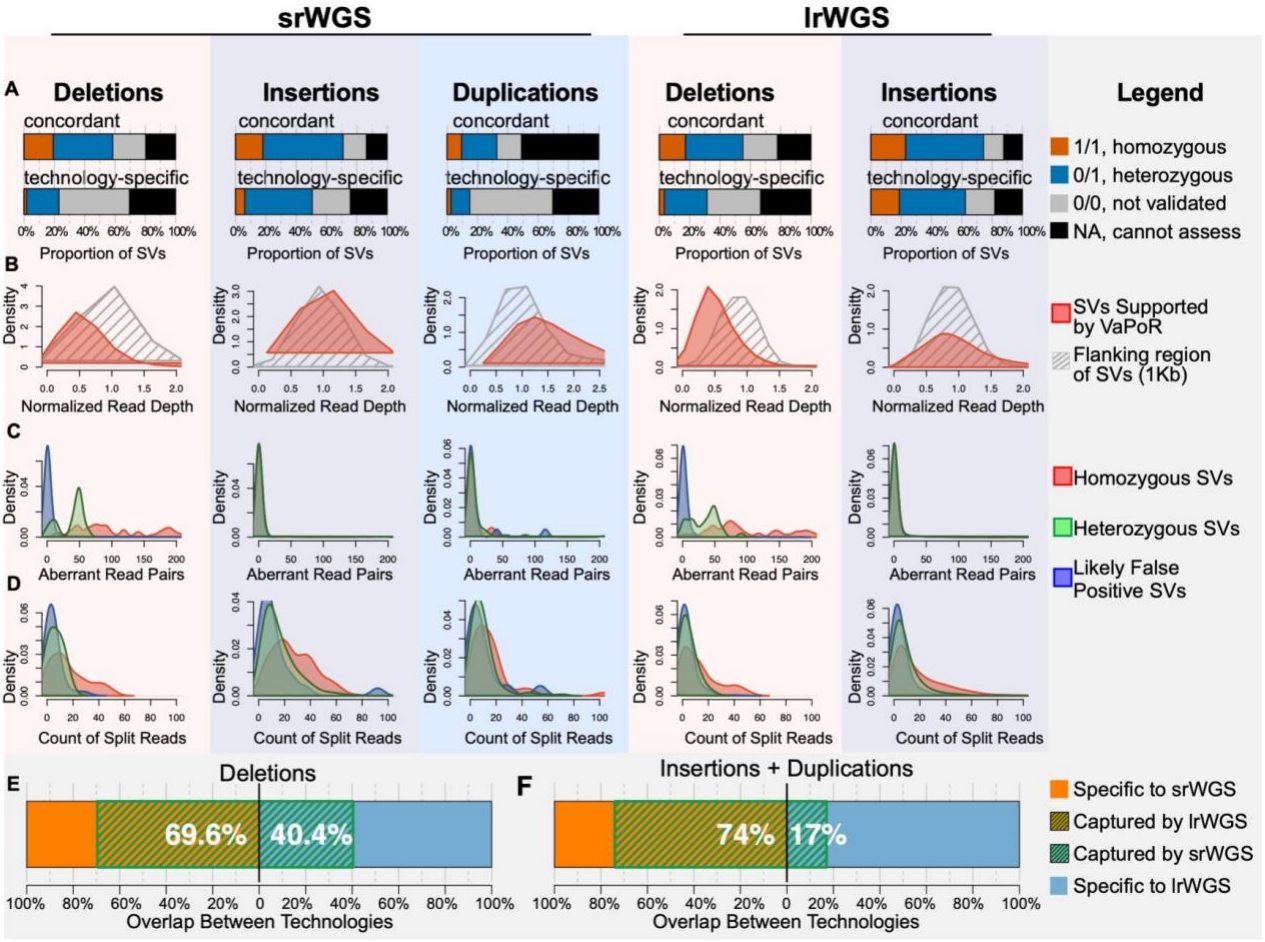

Figure S5. Recalibrating SVs in the repetitive SD + SR sequences based on read-level alignment signatures and concordance of the high-confidence SVs with supports between srWGS and lrWGS.

(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.