RESEARCH ARTICLE SUMMARY

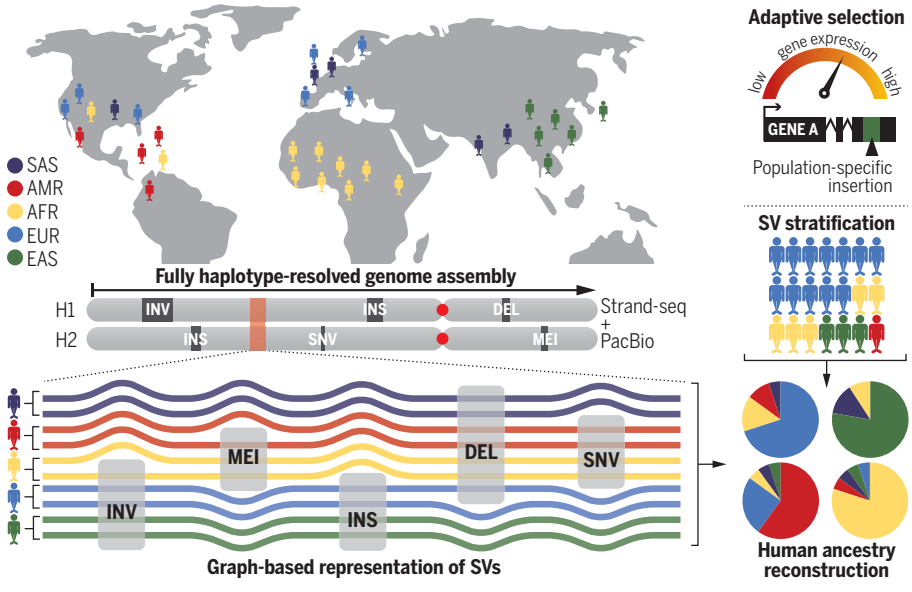
HUMAN GENOMICS

Haplotype-resolved diverse human genomes and integrated analysis of structural variation

Peter Ebert*, Peter A. Audano*, Qihui Zhu*, Bernardo Rodriguez-Martin*, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jana Ebler, Weichen Zhou, Rebecca Serra Mari, Feyza Yilmaz, Xuefang Zhao, PingHsun Hsieh, Joyce Lee, Sushant Kumar, Jiadong Lin, Tobias Rausch, Yu Chen, Jingwen Ren, Martin Santamarina, Wolfram Höps, Hufsah Ashraf, Nelson T. Chuang, Xiaofei Yang, Katherine M. Munson, Alexandra P. Lewis, Susan Fairley, Luke J. Tallon, Wayne E. Clarke, Anna O. Basile, Marta Byrska-Bishop, André Corvelo, Uday S. Evani, Tsung-Yu Lu, Mark J. P. Chaisson, Junjie Chen, Chong Li, Harrison Brand, Aaron M. Wenger, Maryam Ghareghani, William T. Harvey, Benjamin Raeder, Patrick Hasenfeld, Allison A. Regier, Haley J. Abel, Ira M. Hall, Paul Flicek, Oliver Stegle, Mark B. Gerstein, Jose M. C. Tubio, Zepeng Mu, Yang I. Li, Xinghua Shi, Alex R. Hastie, Kai Ye, Zechen Chong, Ashley D. Sanders, Michael C. Zody, Michael E. Talkowski, Ryan E. Mills, Scott E. Devine, Charles Lee†‡, Jan O. Korbel†‡, Tobias Marschall†‡, Evan E. Eichler†‡

INTRODUCTION: The characterization of the full spectrum of genetic variation is critical to understanding human health and disease. Recent technological advances have made it possible to survey genetic variants on the level of fully reconstructed haplotypes, leading to substantially improved sensitivity in detecting and characterizing large structural variants (SVs), including complex classes.

RATIONALE: We focused on comprehensive genetic variant discovery from a human diversity panel representing 25 human populations. We leveraged a recently developed computational pipeline that combines long-read technology and single-cell template strand sequencing (Strand-seq) to generate fully phased diploid genome assemblies without guidance of a reference genome or use of parent-child trio information. Variant discovery from high-quality haplotype assemblies increases sensitivity and yields variants that are not only sequence resolved but also embedded in their genomic context, substantially improving genotyping in short-read sequenced cohorts and providing an assessment of their potential functional relevance.



Discovery and analysis of global human genetic diversity. Starting from a global panel of human diversity (top), we discovered structural variation from fully phased diploid genome assemblies (middle), resulting in a comprehensive catalog of sequence- and context-resolved variants. This facilitates integrative analysis and identification of new associations between variants and molecular phenotypes (bottom). SAS, South Asian; AMR, Admixed American; AFR, African; EUR, European; EAS, East Asian; INV, inversion; INS, insertion; DEL, deletion; MEI, mobile element insertion.

RESULTS: We generated fully phased genome assemblies for 35 individuals (32 unrelated and three children from parent-child trios). Genomes are highly contiguous [average minimum contig length needed to cover 50% of the genome: 26 million base pairs (Mbp)], accurate at the base-pair level (quality value > 40), correctly phased (average switch error rate 0.18%), and nearly complete compared with GRCh38 (median aligned contig coverage >95%). From the set of 64 unrelated haplotype assemblies, we identified 15.8 million single-nucleotide variants (SNVs), 2.3 million insertions/deletions (indels; 1 to 49 bp in length), 107,590 SVs (≥50 bp), 316 inversions, and 9453 nonreference mobile elements. The large fraction of African individuals in our study (11 of 35) enhances the discovery of previously unidentified variation (approximately twofold increase in discovery rate compared with non-Africans). Overall, ~42% of SVs are previously unidentified compared with recent long-read-based studies. Using orthogonal technologies, we validated most events and discovered ~35 structurally divergent regions per human genome (>50 kbp) not yet fully resolved with long-read genome assembly. We found that homology-mediated mechanisms of SV formation are twice as common as expected from previous reports that used short-read sequencing. We constructed a phylogeny of active L1 source elements and observed a correlation between evolutionary age and features such as the activity level, suggesting that younger elements contribute disproportionately to disease-causing variation. Transduction tracing allowed the identification of 54 active SVA retrotransposon source elements, which mobilize nonrepetitive sequences at their 5′ and 3′ ends. We genotyped up to 50,340 SVs into Illumina short-read data from the 1000 Genomes Project and identified variants associated with changes in gene expression, such as a 1069-bp SV near the gene *LIPI*, a locus that is associated with cardiac failure. We further identified 117 loci that show evidence for population stratification. These are candidates for local adaptation, such as a 4.0-kbp deletion of regulatory DNA *LCT* (lactase gene) among Europeans.

CONCLUSION: Fully reconstructed haplotype assemblies triple SV discovery when compared with short-read data and improve genotyping, leading to insights into SV mechanism of origin, evolutionary history, and disease association. ■

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†These authors contributed equally to this work.
‡Corresponding author. Email: eee@gs.washington.edu (E.E.E.); tobias.marschall@hhu.de (T.M.); jan.korbel@embl.org (J.O.K.); charles.lee@jax.org (C.L.)
Cite this article as P. Ebert, *Science* **372**, eabf7117 (2021).
DOI: 10.1126/science.abf7117

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abf7117

Downloaded from http://science.sciencemag.org/ on April 9, 2021

RESEARCH ARTICLE

HUMAN GENOMICS

Haplotype-resolved diverse human genomes and integrated analysis of structural variation

Peter Ebert[1]*, Peter A. Audano[2]*, Qihui Zhu[3]*, Bernardo Rodriguez-Martin[4]*, David Porubsky[2], Marc Jan Bonder[4,5], Arvis Sulovari[2], Jana Ebler[1], Weichen Zhou[6], Rebecca Serra Mari[1], Feyza Yilmaz[3], Xuefang Zhao[7,8], PingHsun Hsieh[2], Joyce Lee[9], Sushant Kumar[10], Jiadong Lin[11], Tobias Rausch[4], Yu Chen[12], Jingwen Ren[13], Martin Santamarina[14,15], Wolfram Höps[4], Hufsah Ashraf[1], Nelson T. Chuang[16], Xiaofei Yang[17], Katherine M. Munson[2], Alexandra P. Lewis[2], Susan Fairley[18], Luke J. Tallon[16], Wayne E. Clarke[19], Anna O. Basile[19], Marta Byrska-Bishop[19], André Corvelo[19], Uday S. Evani[19], Tsung-Yu Lu[13], Mark J. P. Chaisson[13], Junjie Chen[20], Chong Li[20], Harrison Brand[7,8], Aaron M. Wenger[21], Maryam Ghareghani[22,23,1], William T. Harvey[2], Benjamin Raeder[4], Patrick Hasenfeld[4], Allison A. Regier[24], Haley J. Abel[24], Ira M. Hall[25], Paul Flicek[18], Oliver Stegle[4,5], Mark B. Gerstein[10], Jose M. C. Tubio[14,15], Zepeng Mu[26], Yang I. Li[27], Xinghua Shi[20], Alex R. Hastie[9], Kai Ye[11,28], Zechen Chong[12], Ashley D. Sanders[4], Michael C. Zody[19], Michael E. Talkowski[7,8], Ryan E. Mills[6,28], Scott E. Devine[16], Charles Lee[3,29,30]†‡, Jan O. Korbel[4,18]†‡, Tobias Marschall[1]†‡, Evan E. Eichler[2,31]†‡

Long-read and strand-specific sequencing technologies together facilitate the de novo assembly of high-quality haplotype-resolved human genomes without parent-child trio data. We present 64 assembled haplotypes from 32 diverse human genomes. These highly contiguous haplotype assemblies (average minimum contig length needed to cover 50% of the genome: 26 million base pairs) integrate all forms of genetic variation, even across complex loci. We identified 107,590 structural variants (SVs), of which 68% were not discovered with short-read sequencing, and 278 SV hotspots (spanning megabases of gene-rich sequence). We characterized 130 of the most active mobile element source elements and found that 63% of all SVs arise through homology-mediated mechanisms. This resource enables reliable graph-based genotyping from short reads of up to 50,340 SVs, resulting in the identification of 1526 expression quantitative trait loci as well as SV candidates for adaptive selection within the human population.

Advances in long-read sequencing, coupled with orthogonal genome-wide mapping technologies, have made it possible to fully resolve and assemble both haplotypes of a human genome (*1*–*3*). Although such phased human genome assemblies generally improve variant discovery compared with Illumina or "squashed" long-read genome assemblies (*4*), the largest gains in sensitivity have been among structural variants (SVs)—inversions, deletions, duplications, and insertions of ≥50 base pairs (bp) in length. Typical Illumina-based discovery approaches identify only 5000 to 10,000 SVs (*1*, *5*, *6*), in contrast to long-read genome analyses that now routinely detect >20,000 SVs (*1*, *3*, *4*, *7*). Among the different classes of SVs, the greatest gains in sensitivity have been noted specifically for insertions for which >85% of the variation has been reported as previously unidentified (*1*). In addition, repeat-mediated alterations within SV classes, such as variable number of tandem repeats (VNTRs) and short tandem repeats (STRs), have been challenging to delineate from short-read sequencing technologies and are underrepresented in the reference genome and often collapsed in unphased genome assemblies (*8*). The integration of long-read sequencing with new technologies such as single-cell template strand sequencing (Strand-seq) has further catalyzed the unambiguous confirmation of both heterozygous- and homozygous-inverted configurations in a genome (*1*, *9*). Long-read phased genome assemblies (*1*) also better resolve larger full-length mobile element insertions (MEIs), providing an opportunity to systematically investigate their origins and distribution and the mutational processes underlying their mobilization within more complex regions of the genome, including transductions (*10*, *11*).

The Human Genome Structural Variation Consortium (HGSVC) recently developed a method for phased genome assembly that combines long-read PacBio whole-genome sequencing (WGS) and Strand-seq data to produce fully phased diploid genome assemblies without dependency on parent-child trio data (Fig. 1A) (*3*). These phased assemblies enable a more complete sequence-resolved representation of variation in human genomes.

Here, we present a resource that consists of phased genome assemblies, corresponding to 70 haplotypes (64 unrelated and 6 children) from a diverse panel of human genomes. We focus specifically on the discovery of previously unknown SVs through performing extensive orthogonal validation by using supporting technologies with the goal of comprehensively understanding SV complexity, including in regions that cannot yet be resolved with long-read sequencing (fig. S1). Further, we genotyped these newly defined SVs using a pangenome

[1]Heinrich Heine University, Medical Faculty, Institute for Medical Biometry and Bioinformatics, Moorenstraße 20, 40225 Düsseldorf, Germany. [2]Department of Genome Sciences, University of Washington School of Medicine, 3720 15th Avenue NE, Seattle, WA 98195-5065, USA. [3]The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06032, USA. [4]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Meyerhofstraße 1, 69117 Heidelberg, Germany. [5]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. [6]Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, 100 Washtenaw Avenue, Ann Arbor, MI 48109, USA. [7]Center for Genomic Medicine, Massachusetts General Hospital, Department of Neurology, Harvard Medical School, Boston, MA 02114, USA. [8]Program in Medical and Population Genetics and Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [9]Bionano Genomics, San Diego, CA 92121, USA. [10]Program in Computational Biology and Bioinformatics, Yale University, BASS 432 and 437, 266 Whitney Avenue, New Haven, CT 06520, USA. [11]School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China. [12]Department of Genetics and Informatics Institute, School of Medicine, University of Alabama at Birmingham, Birmingham, AL 35294, USA. [13]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA. [14]Genomes and Disease, Centre for Research in Molecular Medicine and Chronic Diseases (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [15]Department of Zoology, Genetics, and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [16]Institute for Genome Sciences, University of Maryland School of Medicine, 670 W Baltimore Street, Baltimore, MD 21201, USA. [17]School of Computer Science and Technology, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China. [18]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [19]New York Genome Center, New York, NY 10013, USA. [20]Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA. [21]Pacific Biosciences of California, Menlo Park, CA 94025, USA. [22]Max Planck Institute for Informatics, Saarland Informatics Campus E1.4, 66123 Saarbrücken, Germany. [23]Saarbrücken Graduate School of Computer Science, Saarland University, Saarland Informatics Campus E1.3, 66123 Saarbrücken, Germany. [24]Department of Medicine, Washington University, St. Louis, MO 63108, USA. [25]Department of Genetics, Yale School of Medicine, 333 Cedar Street, New Haven, CT 06510, USA. [26]Genetics, Genomics, and Systems Biology, University of Chicago, Chicago, IL 60637, USA. [27]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL 60637, USA. [28]Department of Human Genetics, University of Michigan, 1241 E. Catherine Street, Ann Arbor, MI 48109, USA. [29]Precision Medicine Center, The First Affiliated Hospital of Xi'an Jiaotong University, 277 West Yanta Road, Xi'an, 710061, Shaanxi, China. [30]Department of Graduate Studies–Life Sciences, Ewha Womans University, Ewhayeodae-gil, Seodaemun-gu, Seoul 120-750, South Korea. [31]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA.
*These authors contributed equally to this work. †These authors contributed equally to this work.
‡Corresponding author. Email: eee@gs.washington.edu (E.E.E.); tobias.marschall@hhu.de (T.M.); jan.korbel@embl.org (J.O.K.); charles.lee@jax.org (C.L.)

Downloaded from http://science.sciencemag.org/ on April 9, 2021

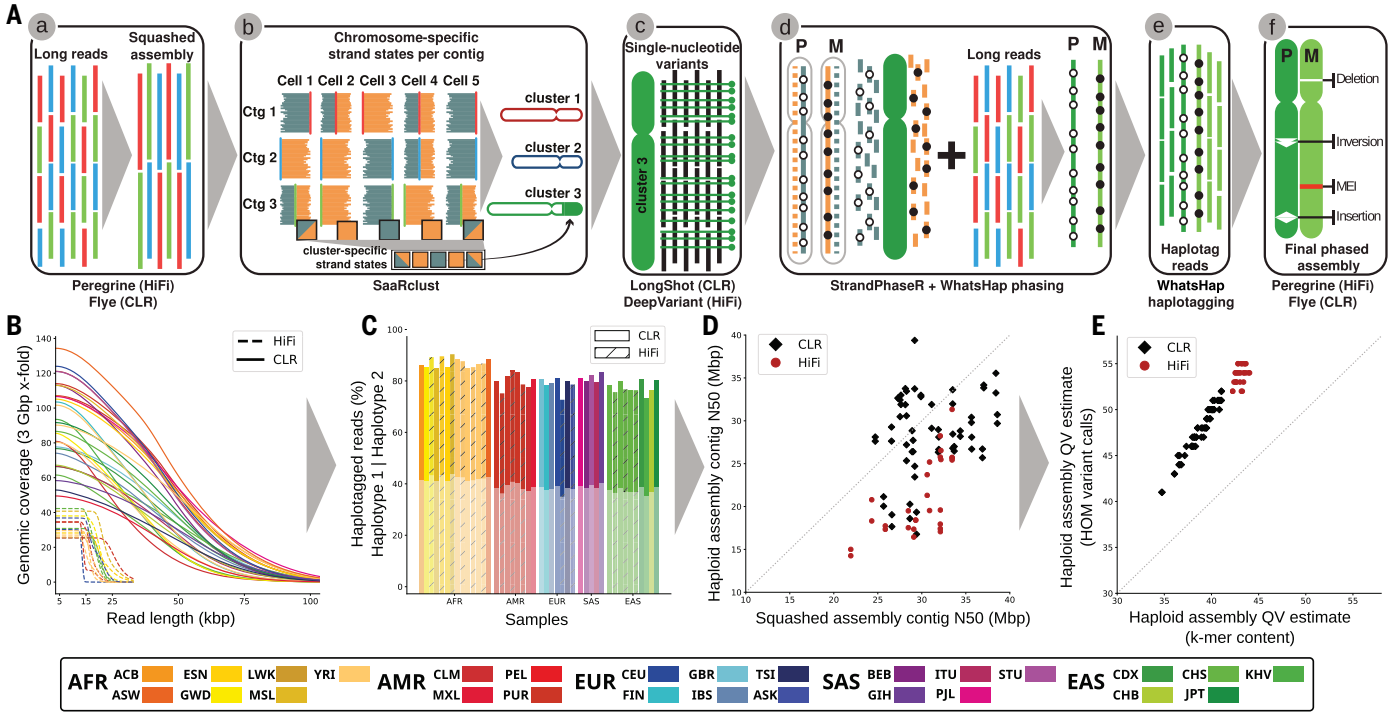


Fig. 1. Trio-free phased diploid genome assembly using Strand-seq (PGAS). (**A**) A schematic of the PGAS pipeline (*3*): (a) generation of a non–haplotype-resolved ("squashed") long-read assembly; (b) clustering of assembled contigs into "chromosome" clusters based on Strand-seq Watson/Crick signal; (c) calling of SNVs relative to the clustered squashed assembly; (d) integrative phasing combines local (SNV) and global (Strand-seq) haplotype information for chromosome-wide phasing; (e) tagging of input long reads by haplotype; and (f) phased genome assembly based on haplotagged long reads and subsequent variant calling (*18*). (**B**) Genomic coverage (*y* axis) as a function of the long-read length (*x* axis). (**C**) Fraction of reads that can be assigned ("haplotagged") to either haplotype 1 (semitransparent) or haplotype 2 for HiFi (hatched) and CLR (solid) datasets. (**D**) Contig-level N50 values for squashed (*x* axis) and haploid assemblies (*y* axis) for CLR (black diamonds) and HiFi (red circles) samples. (**E**) Haploid assembly QV estimates computed from distinct and shared *k*-mers (*x* axis) based on homozygous Illumina variant calls (*y* axis). Samples are colored according to the 1000GP population color scheme (*15*), with the exception of the added Ashkenazim individual NA24385/HG002 (Coriell family ID 3140) (ASK; dark blue).

graph framework (*12*–*14*) into a diversity panel of human genomes now deeply sequenced (>30-fold) with short-read data from the 1000 Genomes Project (1000GP) (*15*, *16*). These findings allowed us to establish their population frequency, identify ancestral haplotypes, and discover new associations with respect to gene expression, splicing, and candidate disease loci. The work provides fundamental new insights into the structure, variation, and mutation of the human genome, providing a framework for more systematic analyses of thousands of human genomes going forward.

Results

Sequencing and phased assembly of human genomes

We initially selected 34 unrelated individual genomes for de novo sequencing, with the goal of at least one representative from each of the 26 1000GP populations, of which 30 samples passed initial quality control (QC) (tables S1 and S2). We additionally sequenced three previously studied child samples, completing three parent-child trios, and we included for analysis publicly available sequencing data for two samples, NA12878 and HG002/NA24385, that were generated as part of the Genome in a Bottle effort (*17*). The complete set of 35 genomes includes 19 females and 16 males of African (AFR; $n = 11$), Admixed American (AMR; $n = 5$), East Asian (EAS; $n = 7$), European (EUR; $n = 7$), and South Asian (SAS; $n = 5$) (table S1) descent. All genomes were sequenced by using continuous long-read (CLR) sequencing ($n = 30$) to an excess of 40-fold coverage or high-fidelity (HiFi) sequencing ($n = 12$) to an excess of 20-fold coverage (Fig. 1B and table S1) (*18*).

As a control for phasing and platform differences, we sequenced nine overlapping samples with both CLR as well as HiFi sequence data corresponding to the three parent-child trios (tables S1 and S2) that had been studied for SVs previously by HGSVC (*1*). For the purpose of phasing, we generated corresponding Strand-seq data (74 to 183 cells) (fig. S2) for each of the samples. We used these data to successfully produce 70 (64 unrelated) phased and assembled human haplotypes [5.7 billion to 6.1 billion base pairs (Gbp) in length for the diploid sequence] (table S1) using a reference-free assembly approach (Fig. 1A) (*3*), which works in the absence of parent-child trio information.

We found that the phased genomes are accurate at the base-pair level [quality value (QV) > 40] and highly contiguous [minimum contig length needed to cover 50% of the genome (N50) > 25 Mbp] (Fig. 1, C to E, and table S1) with low switch error rates (median 0.12%, table S3), providing a diversity panel of physically resolved and fully phased single-nucleotide variant (SNV) and indel (insertion/deletion) haplotypes flanking sequence-resolved SVs (table S4). Using two different metrics from variant calling and *k*-mer content methods, respectively (Fig. 1E), we found that sequence accuracy is higher for human genome assemblies generated with HiFi [median QV = 54 (homozygous variants)/43(*k*-mer)] (Fig. 1E) when compared with CLR [median QV = 48 (homozygous variants)/39(*k*-mer)] (Fig. 1E) sequencing. Considering only accessible regions of the genome (*18*), the MAPQ60 contig coverage of HiFi and CLR genomes are similar (95.43 and 95.12%) (table S5). CLR assemblies, however, are more contiguous (HiFi median

Downloaded from http://science.sciencemag.org/ on April 9, 2021

contig N50 was 19.5 versus 28.6 Mbp for CLR; $P < 10^{-9}$, *t* test). Fifteen of our assembled haplotypes exceed a contig N50 of 32 Mbp, all of which were based on CLR sequencing for which insert libraries are much larger and sequence coverage is higher with half the number of single-molecule, real-time (SMRT) cells (Fig. 1D, fig. S3, and table S6).

Comparing Strand-seq phasing accuracy for six samples for which parent-child trio data are available (figs. S4 and S5 and table S3) (*3*), we estimate that, on average, 99.86% of all 1-Mbp segments are correctly phased from telomere to telomere (average switch error rate of 0.18% and Hamming distance of 0.21%) (table S3). Predictably (*3*), remaining assembly gaps are enriched (*18*) in regions of segmental duplications (SDs) and acrocentric and centromeric regions of human chromosomes (figs. S6 and S7 and table S7). As a final QC of assembly quality, we analyzed Bionano Genomics optical mapping data for 32 genomes and found a median concordance of >97% between the optical map and the phased genome assemblies (figs. S8 and S9 and table S8).

Phased variant discovery

Unlike previous population surveys of structural variation (*1*, *4*, *19*–*21*), which mapped reads or unphased contigs to the human reference genome, we developed the Phased Assembly Variant (PAV) caller to discover genetic variants on the basis of a direct comparison between the two sequence-assembled haplotypes and the human reference genome, GRCh38 (*18*). In the end, each human genome is rendered into two haplotype-resolved assemblies (each 2.9 Gbp) in which all variants are physically linked (table S4). We classify variants as SNVs, indels (1 to 49 bp), and SVs (≥50 bp), which includes copy number variants (CNVs) and balanced inversion polymorphisms. After filtering (*18*), our nonredundant callset of unrelated samples contains 107,590 insertion/deletion SVs, 316 inversions, 2.3 million indels, and 15.8 million SNVs.

We observed a 2-bp periodicity for indels (dinucleotide repeats) and modes at 300 bp and 6 kbp for Alu and L1 MEIs, respectively (Fig. 2A), with only a small fraction of intersecting functional elements (Fig. 2B) (*22*). PAV readily flags all reference-based artefacts or minor alleles by pinpointing regions where the 64 phased human genomes consistently differ from GRCh38 (1573 SVs, 18,630 indels, and 91,537 SNVs, "shared variants") (Fig. 2C) (*18*). The greater haplotype diversity allows us to reclassify 50% of previously annotated shared SVs (*4*) as minor alleles and correct the coding sequence annotation of five genes with tandem repeats (*RRBP1*, *ZNF676*, *MUC2*, and *STOX1*) or extreme GC content (*SAMD1*) (table S9). We estimate a false discovery rate (FDR) of 5 to 7% for SVs on the basis of support from sequence-read-based callers, as well as an independent alignment method (*18*). A comparison against SVs called from the benchmark Genome in a Bottle sample (HG002), including orthogonal datasets, suggests an FDR of ~4%, although this estimate is restricted to a subset of the genome where events could be more reliably called (*18*).

Similarly, we estimate a 6% FDR for indels and 4% for SNVs on the basis of an assessment of Mendelian transmission error from the HiFi and CLR parent-child trios (table S10) (*18*). We found that 42% of the SVs are previously unidentified when compared with recent long-read surveys of human genomes (fig. S10) (*1*, *4*, *19*–*21*). The addition of African samples more than doubles the rate of new variant discovery when compared with non-Africans for all classes of variation [2.21× SVs (809 versus 366), 3.70× indels (11,514 versus 3109), and 2.97× SNVs (160,232 versus 54,006)] for the 64th haplotype (Fig. 2C and table S11) (*18*). On average, we detected 24,653 SVs, 794,406 indels, and 3,895,274 SNVs per diploid human genome (table S4).

SV discovery from short-read alignments

To enable comparison of the PAV calls with genetic variants discovered with WGS, we performed Illumina-based short-read sequencing for 3202 samples from the 1000GP (34.5-fold coverage) (*18*) and discovered SVs by using three analytic pipelines: GATK-SV (*5*), SVTools (*6*), and Absinthe. When focusing on the 31 unrelated samples with matching PacBio long-read sequences and callsets included in this study (NA24385, HG00514, HG00733, and NA19240 excluded), we observed 9320 SVs per genome at 1.8% FDR by comparison with 24,596 SVs per genome from long-read assembly (Fig. 2D and fig. S11). On average, 77.4% of SVs detected with short-read pipelines were concordant with long-read assemblies, but only 29.6% of long-read SVs were observed in the short-read WGS callset (Fig. 2D). The greatest gains in sensitivity from long-read assemblies were observed among smaller SVs, where ~83.3% of events (<250 bp) were previously unidentified (Fig. 2E), whereas the short-read SV pipelines displayed greater sensitivity among large SVs > 5 kbp (Fig. 2E, figs. S11 and S12, and tables S12 and S13).

SV distribution and mechanisms

SVs are known to be clustered (*4*, *15*), and we identified 278 SV hotspots on the basis of our PAV callset (Fig. 2F, fig. S13, and table S14) (*18*) spanning ~279 Mbp of the genome (Fig. 2F, inset). We found that 30.6% (32,222 of 105,327) of SVs on autosomes and chromosome X map within the last 5 Mbp of chromosome arms, corresponding to an approximately fourfold enrichment ($P = 0.001$, z score = 301.3, permutation test), with few notable exceptions: the long arm of the X chromosome and the short arms of chromosomes 3 and 20 (Fig. 2F and fig. S14A). Focusing on SVs >5 Mbp from chromosome ends (73,105), we identified 221 hotspots (fig. S14B). Of these, 49% (109 of 221) have not been previously identified from short-read analyses of the 1000GP data (*23*). These interstitial hotspots are enriched 6.6-fold ($P = 0.001$, z score = 26.6, permutation test) for SDs that are consistent with homologous recombination and frequently correspond to gene-rich regions of exceptional diversity among human populations. For example, we identified three distinct hotspots mapping to the major histocompatibility complex (MHC) region that distinguish seven selected structural haplotypes (Fig. 2G, fig. S15, and table S15). Our analysis indicates that a majority (98.85%) of this 4-Mbp region has been sequence resolved at the base-pair level [29 of the assemblies are a single assembled contig, and 18 have a single gap; 17 of 19 individual human lymphocyte antigen (HLA) genes are fully sequence resolved in all assemblies] (tables S15 and S16).

A detailed analysis of the SVs with unambiguous breakpoint locations provided an opportunity to examine mechanisms of SV formation. Excluding MEIs and SVs with ambiguous breakpoints, we assessed 52,974 insertions and 30,467 deletions (table S17). We found that 58% of insertions and 70% of deletions, including SVs in VNTRs, are flanked by at least 50 bp of homologous sequence, suggesting formation through homology-directed repair (HDR) processes or nonallelic homologous recombination (NAHR). Among those, 15% of insertions and 25% of deletions showed >200-bp flanking homology and are more likely mediated by NAHR. VNTRs with short repeat units (<50 bp) account for a smaller number of events (1.6% insertions and 0.4% deletions) and suggest replication slippage-mediated expansion and contraction. Additionally, 40% of insertions and 29% of deletions show blunt-ended breakpoints or microhomology (<50 bp flanking sequence identity), which is consistent with nonhomologous end joining, microhomology-mediated end joining, or microhomology-mediated break-induced replication (*24*). Homology-associated SVs are twofold more frequent than expected from reports that used short reads (*25*–*27*), and when considering Illumina sequencing-based SV calls from the same samples, only 2% of insertions and 19% of deletions appear to be NAHR-mediated SVs with ≥200-bp flanking homology ($P < 2.2 \times 10^{-16}$; Fisher's exact test) (table S17).

SVs and their breakpoints are generally more depleted within protein-coding sequences and other functional elements, with the exception of specific gene families in which variability in the length of amino acid sequences relates to

Downloaded from http://science.sciencemag.org/ on April 9, 2021

the function of the molecule [lipoprotein (for example, *LPA*), mucins (*MUC1*, *MUC3A*, *MUC4*, *MUC12*, *MUC20*, and *MUC21*), zinc finger genes (*ZNF99*, *ZNF285*, and *ZNF280*), among others] (table S18). We identified 9.4% of all SV breakpoints that intersect functional elements, such as exons (n = 993), untranslated regions (UTRs; n = 1097), promoters (n = 466), and enhancer-like elements (n = 6796) (Fig. 2B and table S19).

When we considered structural polymorphisms that arise from perfect triplet repeats, expansions outnumbered contractions 3 to 1 (271 expansions, 88 contractions), which is consistent with such regions being systematically underrepresented in the original reference (*8*, *28*). Over the 64 haplotypes, there are six such SVs per haplotype, and we identified a total of 106 nonredundant loci (tables S20 and S21). Five of seven of the largest insertions of uninterrupted CTG or CGG repeat insertions mapping within exons correspond to genes already associated with triplet repeat instability diseases or fragile sites. For example,

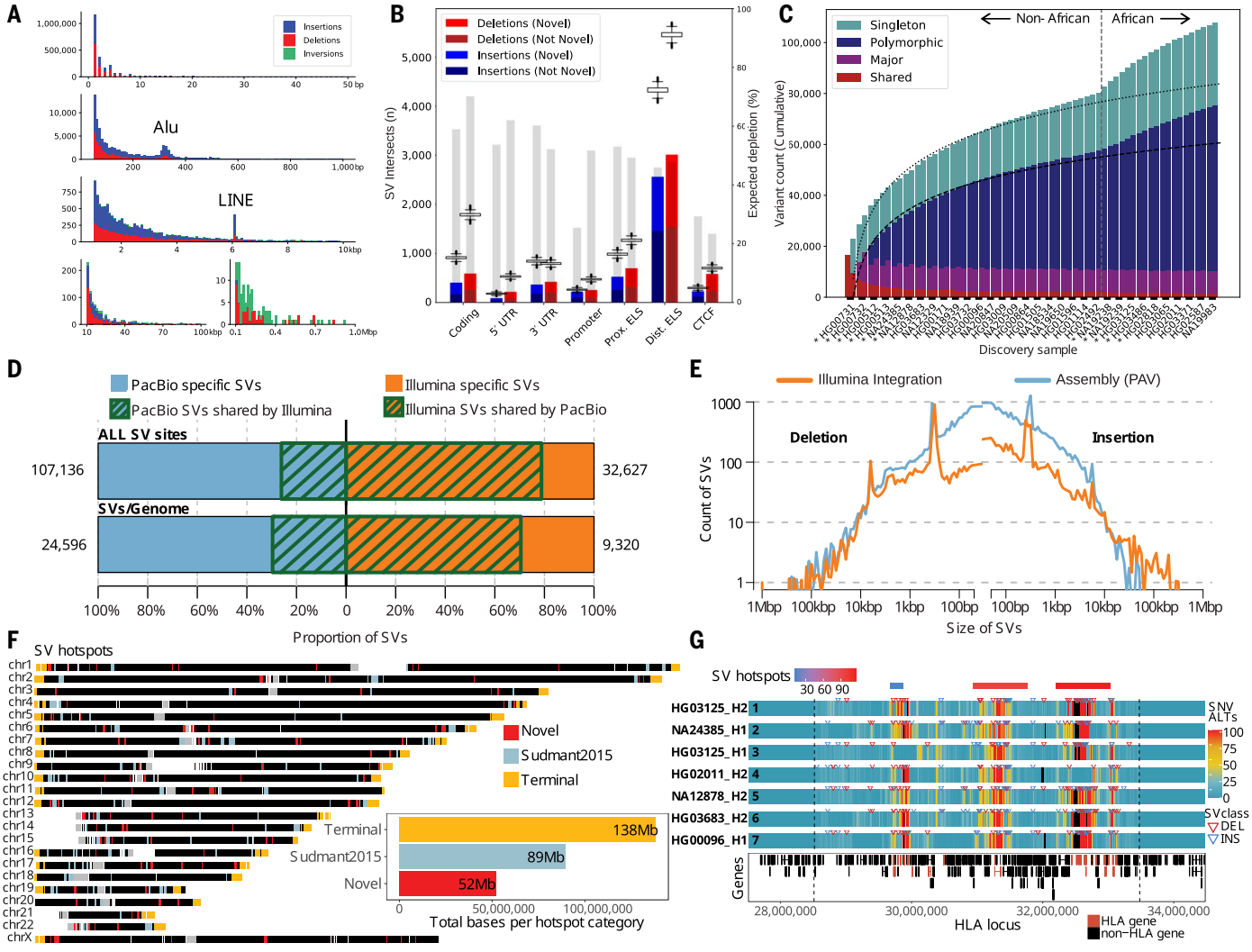


Fig. 2. Variant discovery and distribution. (**A**) Size distribution of indels and SVs from 64 unrelated reference genomes shows a 2-bp periodicity for indels, 300-bp peak for Alu insertions (second row), and 6-kbp peak for L1 MEIs. (**B**) The number of SVs intersecting functional elements (x axis) compared with randomly permuting SV locations (box plots). Gray bars indicate percent depletion (right y axis scale). ELS, Enhancer-like signature; CTCF, CCCTC-binding factor. (**C**) Cumulative number of distinct SVs when adding samples one by one, from left to right. The rate of SV discovery slows with each new haplotype (regression lines); however, the addition of haplotypes of African origin (dashed line) increases SV yield. Colors indicate SVs shared among all haplotypes and not present in GRCh38 (red), major allele variants (AF ≥ 50%, purple), polymorphisms (≥2 haplotypes, blue), and singletons (teal). Asterisks indicate samples sequenced by use of PacBio HiFi. (**D**) Overlap between SVs detected by PacBio long-read assemblies and Illumina short-read alignments on 31 matched samples (NA24835, HG00514, HG00733, and NA19240 excluded). Top bar shows overall SV sites across 31 samples, and the bottom bar displays the average count of SVs per sample, with green stripes representing concordant SV calls between technologies. (**E**) Length distribution of SVs detected with PacBio long-read assemblies and Illumina short-read alignments across all 31 matched samples. (**F**) Genome-wide distribution of SV hotspots divided in three categories: last 5 Mbp of chromosomes (yellow), overlapping (light blue), and previously unidentified (red) when compared with short-read SV analysis of 1000GP (*23*). (Inset) The total sequence length is represented by each hotspot category. (**G**) Heatmap of seven selected SV haplotypes for 4-Mbp MHC region (chr6: 28,510,120 to 33,480,577; dashed lines) comparing regions of high-SNV (red) and low-diversity (blue) regions based on the number of alternate SNVs compared with the reference (GRCh38; alignment bin size 10 kbp, step 1 kbp). Phased SV insertions (blue open arrowheads) and deletions (red open arrowheads) are mapped above each haplotype. The most diverse regions correspond to SV hotspots (red and blue bars top row) and cluster with HLA genes (red bottom track).

Downloaded from http://science.sciencemag.org/ on April 9, 2021

we identified a 21-copy CTG repeat expansion in *ATXN3* (Machado-Joseph disease), a 17-copy gain of CAG in *HTT* (Huntington's disease), a 21-copy gain of a CGG repeat in *ZNF713* (Fragile site 4A), and a 36-copy CGG gain in *DIP2B* (Fragile site 12A) (*18*). The discovery of these perfect repeat insertion alleles with respect to the human reference provides an important reference for future investigations of triplet repeat instability.

MEIs

On the basis of the phased genome assemblies, we identified a collection (n = 9453) of fully sequence-resolved nonreference MEIs—including 7738 Alus, 1175 L1Hs, and 540 SVAs (*18*)—and used sequence content of the elements and their flanking sequences to provide insight into their origin and mechanisms of retrotransposition. Retroelement insertions typically display the classic hallmarks of integration through target-site primed reverse transcription. These include endonuclease cleavage motifs at insertion breakpoints, polyadenylate tracts at their 3′ end, target site duplications ranging from 3 to 52 bp (mode = 14 bp), in addition to frequent inversion and truncation for L1 elements (fig. S16). Full-length L1 (FL-L1) elements are an especially relevant source of genetic variation because they can mutagenize germline and somatic cells and can lead to gene disruptions that cause human disease (*29*, *30*). Although a minority of nonreference L1s are full length (fig. S16 and table S22), we found that 78% of FL-L1s possess two intact open reading frames (ORF1 and ORF2), encoding the proteins that drive L1, Alu, SVA, and processed pseudogene mobilization. Of these sequences, 23% show evidence of activity because they are part of a database of 198 FL-L1s that are active in vitro (*31*, *32*), in human populations (*33*), and in cancers (*34*–*36*). Most active copies (72%; 142 of 198) are either in our callset or present in the reference genome and are now fully sequence resolved (table S23). Of the active FL-L1s, 19% have at least one ORF disrupted, which includes a hot element at 9q32 that is reported to be highly active in diverse tumors (*34*).

Using L1 *Pan troglodytes* as an outgroup, we constructed a phylogeny of active human L1s and estimated their age in million years (Fig. 3A, fig. S17). As expected, copies of the Ta-1 subfamily are the youngest [mean = 1.00; 95% confidence interval (CI): 0.88 to 1.13], followed by Ta-0 (mean = 1.63; 95% CI: 1.49 to 1.77) and pre-Ta (mean = 2.15; 95% CI: 1.91 to 2.40) (fig. S18). The evolutionary age correlates with L1 features such as subfamily, level of activity, and allele frequency (Fig. 3B and fig. S19)—with the youngest FL-L1s typically corresponding to highly polymorphic and active Ta-1 sequences. Three out of the four youngest active FL-L1s—namely 2q24.1, 6p24.1, and 6p22.1-2—are Ta-1 copies reported to be extremely active in cancer genomes (*34*). By contrast, 1p12 is a fixed Pre-Ta insertion that despite integrating into the human genome ~1.8 million years ago remains both highly

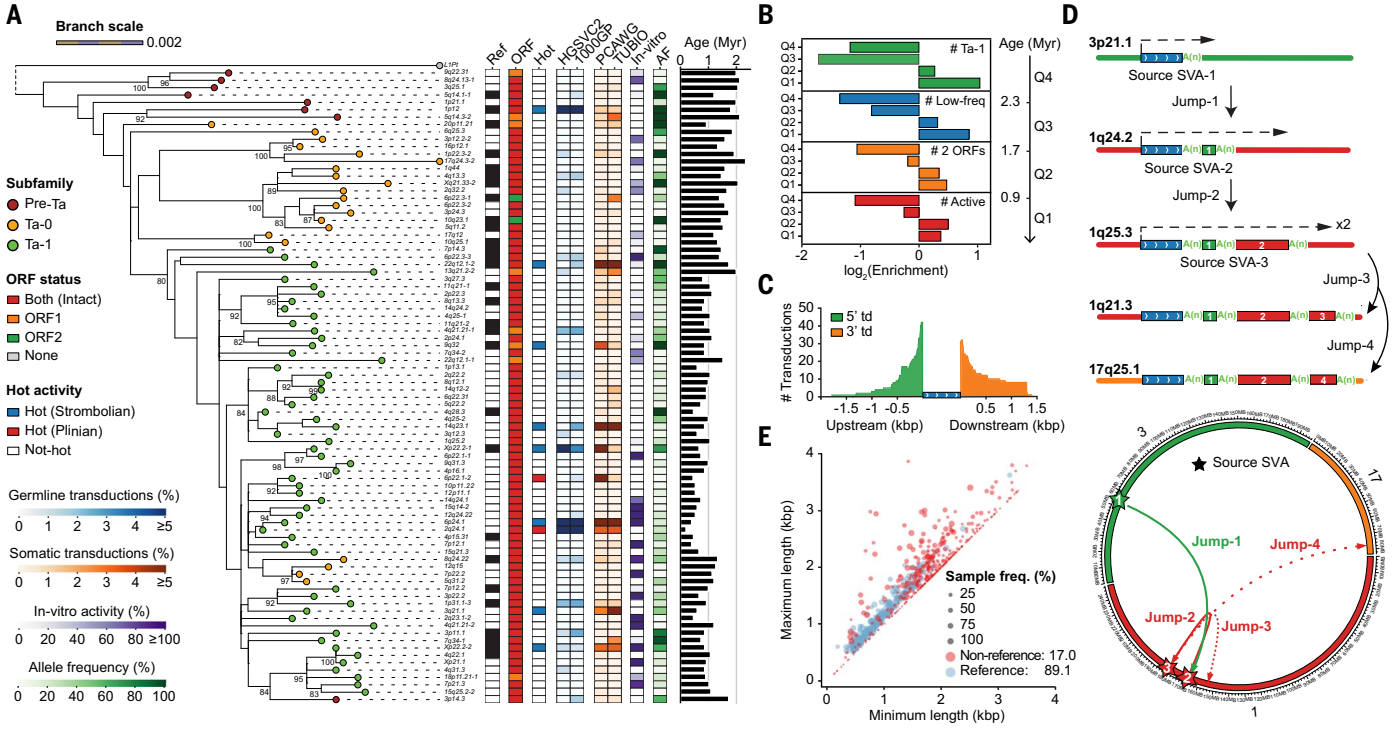


Fig. 3. MEIs. (**A**) Maximum-likelihood phylogenetic tree (*85*) for highly active sequence-resolved FL-L1s annotated by subfamily designation, presence or absence on the reference, ORF content, and hot activity profile (bootstrap values ≥80% are shown) (*34*–*36*). Tree branch lengths are scaled according to the average number of substitutions per base position. Dashed lines map each L1 cytoband identifier to its corresponding branch on the tree. *Pan troglodytes* (L1Pt) is included as an outgroup. Heatmaps represent AF based on the assembly discovery set, activity estimates based on in vitro assays (*31*, *32*), and the number of transduction events detected in human populations (*33*) or cancer studies (*34*–*36*). (**B**) Enrichment and depletion in the number of FL-L1s belonging to the Ta-1 subfamily at age quartiles (Q1 to Q4) compared with a random distribution. Same applies for the other features, including the number of FL-L1s with low allele frequency (MAF < 5%), with two intact ORFs, or with evidence of activity. (**C**) Size distribution and number of 5′ and 3′ SVA-mediated transductions (td) based on the analysis of flanking sequences. (**D**) (Top) Schematic and (bottom) circos representation for serial SVA-mediated transduction events. Dashed arrows indicate SVA transcription initiation and end. Transduced sequences are shown as colored boxes, with their length proportional to transduction size. (**E**) Distributions of VNTR length (x axis, the minimum; y axis, the maximum) of reference and nonreference SVA elements. Reference SVAs are shown as blue dots, and nonreference SVAs are shown as red dots. The dot size represents the sample frequency of SVAs among discovery samples in the HGSVC.

Downloaded from http://science.sciencemag.org/ on April 9, 2021

active in the germline (*33*) and somatically associated with tumors (*34*–*36*). This indicates that a small set of pre-Ta representatives possibly remain very active in the human genome.

SVA source elements are able to produce 5′ and 3′ transductions through alternative transcription start sites or bypassing of normal polyadenylate [poly(A)] sites during retrotransposition (*10*, *11*). We detected 77 transduced nonrepetitive DNA sequences at SVA insertion ends (table S24). 5′ transductions are more abundant (58%; 45 of 77) than 3′ transductions (Fig. 3C), as opposed to L1s, which primarily mediate 3′ transduction events (95%; 89 of 94). We used these distinct transduced sequences to trace the origin of all 77 SVAs to 56 source SVA elements (fig. S20 and table S25). A majority of source loci (84%) belong to the youngest human-specific SVA-E and SVA-F subfamilies (*37*), and only 11 source elements generate 38% of the offspring insertions.

SVA transductions can occasionally shuffle coding sequences, as illustrated by the mobilization of a complete exon of *HGSNAT* by an intronic SVA in antisense orientation (fig. S21). In addition, one SVA source element appears to have caused three sequential mobilization events, as indicated by nested transductions flanked by poly(A) tails (Fig. 3D and fig. S22). Last, SVA elements harbor CpG-rich VNTRs in their interior regions that can expand and contract; we found that nonreference SVAs show significantly greater variability in VNTR copy number as compared with those present in the reference ($P < 10^{-5}$, Student's t test, two-sided) (Fig. 3E).

Inversions

Copy number neutral inversions are among the most difficult SVs to detect and validate (*1*). We applied multiple approaches, integrating Strand-seq, Bionano optical mapping, and PAV-based variant discovery to generate a comprehensive and orthogonally validated set of inversions. PAV specifically increases inversion detection sensitivity for smaller events (fig. S23) by including a k-mer density assessment to resolve inner and outer breakpoints of flanking repeats, which does not rely on alignment breaks to identify inversion sites (*18*). PAV identifies an additional 43 inversions, on average, increasing sensitivity more than twofold compared with previous phased assembly callsets (*2*). In total, we discovered on average 117 inversions per sample (316 nonredundant calls across samples) (fig. S23).

As expected, inversions flanked by SDs tend to be larger than those in distinct regions of the genome [Wilcoxon rank sum test (one-sided, greater), $P = 3.2 \times 10^{-13}$] (fig. S24) (*38*). We focused on one complex region that maps to chromosome 16p12, where we observed a large number of polymorphic inversions flanked by SDs (fig. S25A) (*9*). The region harbors 11 different inversions (Fig. 4A, red and gray arrows), distinguishing 22 different structural configurations that span a ~2.5-Mbp gene-rich region of chromosome 16p (up to 13 protein-coding genes are flipped in orientation depending on human haplotypes) (*18*). These configurations are distributed among human populations but do not correspond to specific haplotypes (Fig. 4A). For example, an analysis of the flanking sequence shows that at least five of the inversions occur in multiple haplotype backgrounds, which is indicative of recurrent inversion toggling (*38*, *39*) between a direct and inverted state (fig. S26) (*18*). Although Strand-seq data allow us to unambiguously identify the inversion status of the distinct regions, most of the breakpoints themselves are not yet fully sequence resolved because of the presence of large repeats (Fig. 4A and fig. S25B) (*3*).

Complex structural variation

We investigated the remaining gaps in our assemblies that map near or within centromeres, acrocentric regions, and SDs (figs. S6 and S7 and table S7). Because such repetitive regions have long been known to be enriched in complex variation (*40*) and refractory to sequence assembly even with long-read data (*1*), we reexamined the genome-wide optical maps to assess additional regions of structural variation. In 30 samples, we found that 72% of the large insertions and deletions (≥5 kbp) discovered with optical mapping are completely sequence resolved and concordant with the assembly (table S26), but the remainder show additional complexity. For example, our analysis of the Puerto Rican phased genome assembly (HG00733) originally identified a 75-kbp deletion between the two haplotypes at chromosome 1p13.3, but a comparison with Bionano Genomics data shows a more complex pattern than a single deletion event: An inversion of 75 kbp is found in the alternate allele flanked by inverted SDs of 100 kbp involving *NBPF* genes (Fig. 4B). Such discrepant regions appear to cluster in the genome.

A comparison between the phased assemblies and Bionano Genomics optical maps revealed 1175 nonredundant SV clusters not detected in the phased assemblies and an additional 482 SV clusters with support in a different individual (table S27). Among the 1175 Bionano SV clusters not detected in the PacBio phased assemblies, 71 overlapped unresolved sequence ("N" gaps), and 69.3% (765 of 1104) of the remaining SV clusters were detected from the Illumina short-read alignment pipelines (table S28). We manually inspected the 339 Bionano SV clusters that could not be detected in any of the short-read or assembly-based analyses and found read-depth evidence supporting 13.9% (47 of 339).

We estimate that there are still ~35 unresolved regions per phased assembly that are >50 kbp in length where there are five or more distinct SV haplotypes in the human population. On chromosome 3q29, for example (Fig. 4C), we identified 18 distinct structural haplotypes involving at least nine copy number and inversion polymorphisms that affect hundreds of kilobases of gene-rich sequence (minimum 375 kbp, maximum 690 kbp) (Fig. 4C). This pattern of structural diversity maps to the proximal breakpoint of the chromosome 3q29 microdeletion and microduplication syndrome rearrangement (chr3: 195,999,954 to 197,617,802) that is associated with developmental delay and adult neuropsychiatric disease (*41*).

Genotyping

We applied PanGenie (*42*), a method designed to leverage a panel of assembly-based reference haplotypes threaded through a graph representation of genetic variation that takes advantage of the linkage disequilibrium inherent in the phased genomes. We initially performed this genotyping step using a reference set of 15.5 million SNVs, 1.03 million indels (1 to 49 bp), and 96,145 SVs (where there was <20% allelic dropout) (fig. S1 and table S29) and genotyped these variants into the 1000GP WGS dataset (*18*), observing expected patterns of diversity (Fig. 5A and figs. S27 and S28) (*15*).

As one measure of genotyping quality, we compared the allele frequencies derived from assembly-based PAV calls across the 64 reference haplotypes to short-read–based allele frequencies obtained from PanGenie for the 2504 unrelated individuals. From the raw output of PanGenie, we observed an allele frequency correlation (Pearson's) of 0.98 for SNVs, 0.95 for indels, and 0.85 for SVs. To further improve SV genotyping, we filtered the variants by assessing (i) Mendelian consistency, (ii) the ability to detect the nonreference allele, (iii) genotype qualities, and (iv) concordance to assembly-based calls in a leave-out-one experiment (*18*). Using these criteria, we defined a subset of strict and lenient SVs for genotyping that contains 24,107 SVs (25%) and 50,340 SVs (52%), respectively, with excellent allele frequency correlation of 0.99 (strict) (Fig. 5B) and 0.95 (lenient) (fig. S29). Performance metrics for deletions and insertions are comparable [strict set: SV deletions, correlation coefficient (r) = 0.98; SV insertions, r = 0.99] (Fig. 5B), highlighting the value of sequence-resolved insertion alleles being part of our reference panel, as well as the algorithm's ability to leverage it (fig. S30). Beyond SVs, 12,283,650 SNVs (79%) and 705,893 indels (68%) met strict filter criteria (given this larger fraction, we did not define a lenient set for these variant classes).

Downloaded from http://science.sciencemag.org/ on April 9, 2021

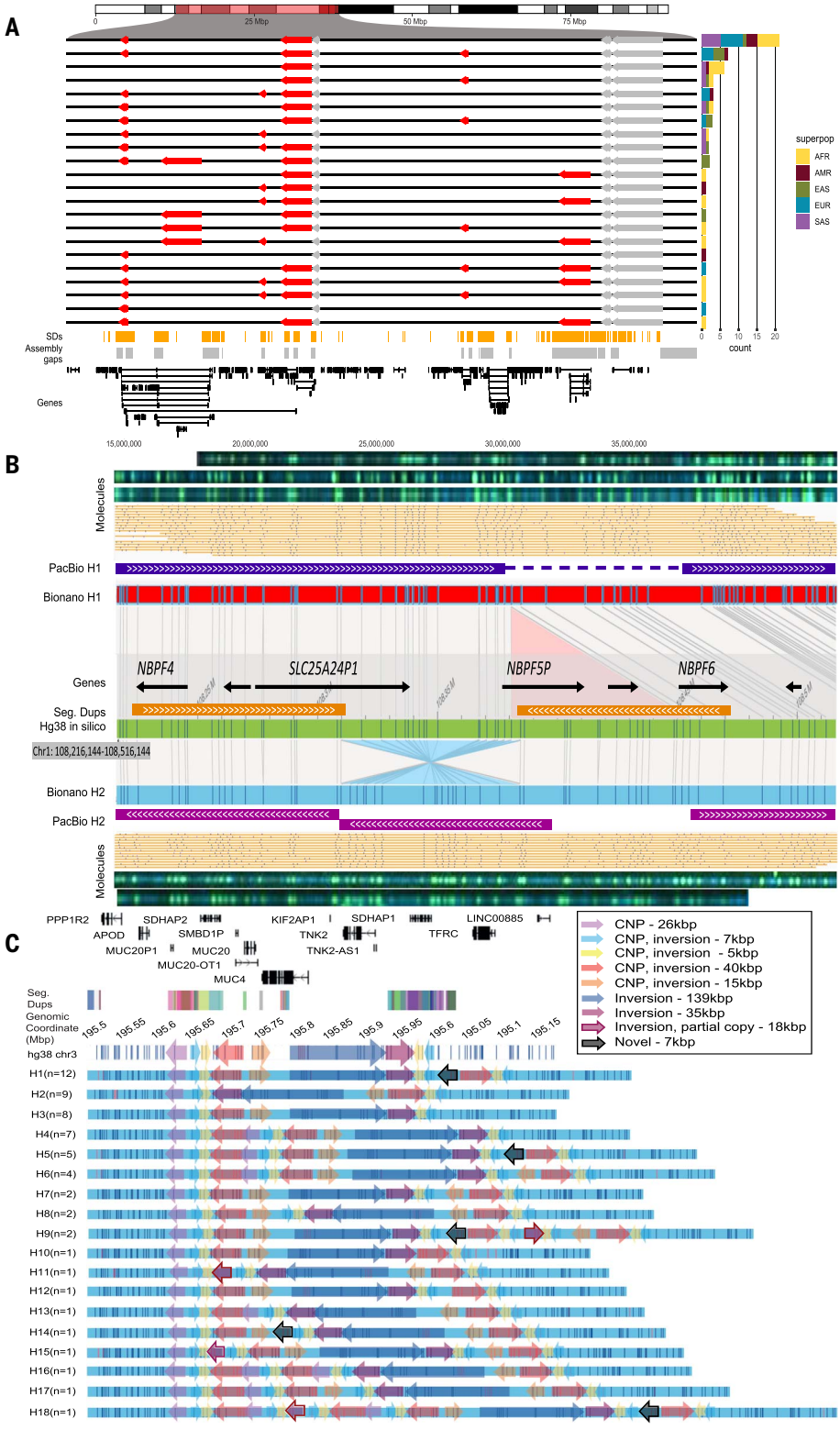


Fig. 4. Complex patterns of structural variation. (**A**) An inversion hotspot mapping to a 2.5-Mbp gene-rich region of chromosome 16p12 (highlighted portion of ideogram). Haplotype structure of inversions (red arrows) are compared with the GRCh38 reference orientation (black lines) as well as additional inversions (gray), which could not be haplotype integrated because of uninformative markers. (Right) A barplot enumerates the frequency of each distinct inversion configuration (n = 22) by superpopulation for the 64 phased genomes. (Bottom) Distribution of SDs (orange), assembly gaps (gray), and genes (black) in a given region. (**B**) A partially resolved complex SV locus (HG00733 at chr1: 108,216,144 to 108,516,144). Optical maps generated through *DLE1* digestion predict a deletion (red bar; Bionano H1) and an inversion (blue bar; Bionano H2) when compared with GRCh38 (green bar). Haplotype structures are strongly supported by extracted single molecules (beige) and raw images (green dots). Phased assembly correctly resolves the hap1 deletion (purple top), and Strand-seq detects the inversion (blue) but misses the flanking SD, which is a gap in the H2 assembly (gap). (**C**) Haplotype structural complexity at chromosome 3q29. Optical mapping of a 410-kbp gene-rich region (chr3: 195,607,154 to 196,027,006) predicts 18 distinct structural haplotypes (H1 to H8) that vary in abundance (n = 1 to 12) and differ by at least nine copy number SDs and associated inversion polymorphisms (colored arrows). This hotspot leads to changes in gene copy and order (GENCODE v34, top): 26 haplotypes are fully resolved by phased assembly (21 CLR, 5 HiFi), and the median MAP60 contig coverage of the region is 96.1%.

Added value from graph-based genotyping into short-read WGS data

To determine the value added from PanGenie genotyping, we next focused on an integrated comparison of long-read SV discovery (PAV), state-of-the-art short-read SV discovery, and the set of genotypable SVs by PanGenie. Consistent with our previous analyses (*43*), we observed that most SVs specific to long-read discovery localized to highly repetitive sequences, which collectively harbored 95.8% of long-read–specific deletions and 85.7% of long-read–specific insertions (table S30). We also discovered variation that was detected (although not sequence resolved) and genotyped only by means of sequencing read-depth from short reads. On average, there were 167 large CNVs (>5 kbp) per sample, 88.2%

Downloaded from http://science.sciencemag.org/ on April 9, 2021

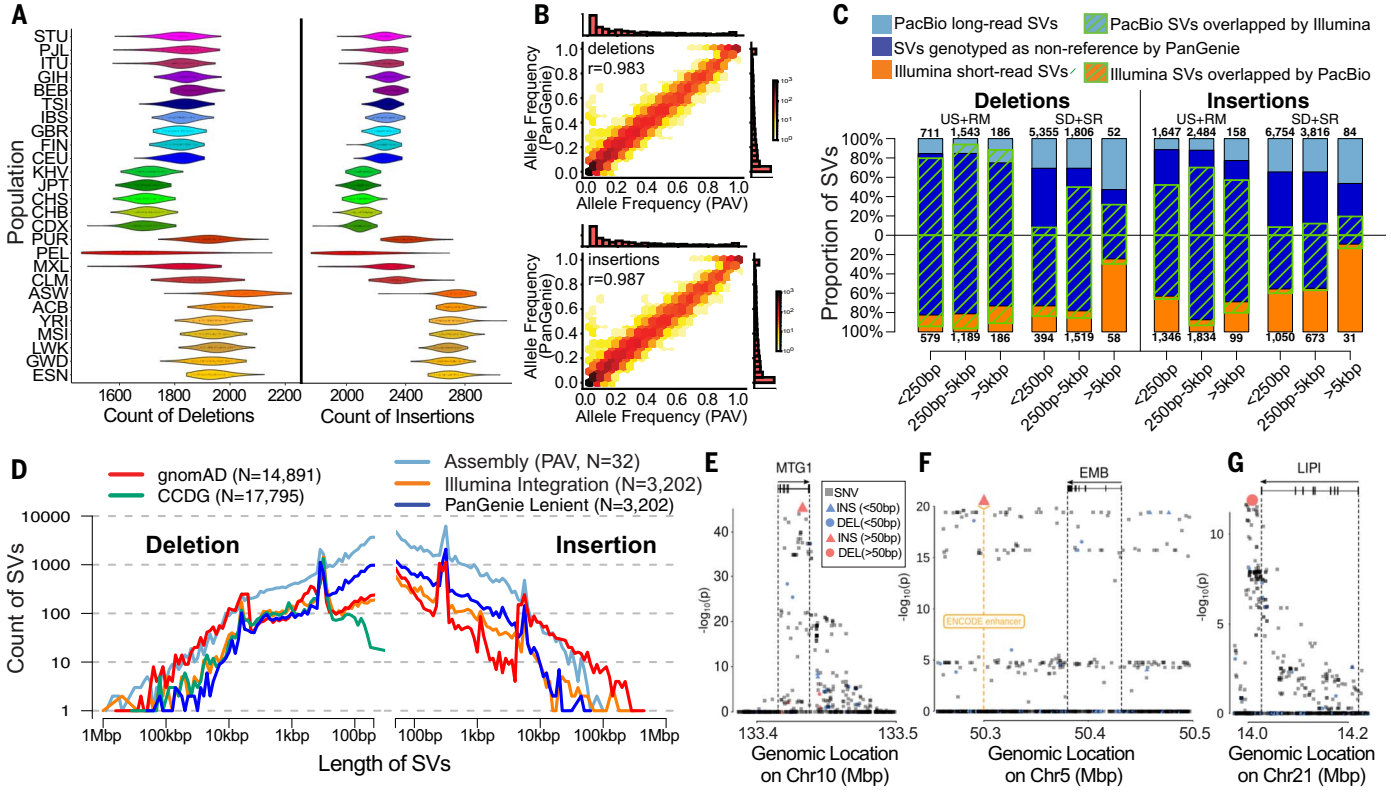


Fig. 5. SV genotyping and eQTL analysis. (**A**) Distribution of heterozygous SV counts per diploid genome broken down by population, based on PanGenie genotypes passing strict filters. (**B**) Concordance of AF estimates from the assembly-based PAV discovery callset and AF estimates from genotyping unrelated Illumina genomes (n = 2504) with PanGenie (strict genotype set of 24,107 SVs). Marginal histograms are in linear scale. (**C**) Count of short- and long-read SVs across variant class, size distribution, and genomic sequence localization. Blue bars indicate the proportion of SVs genotyped by PanGenie with AF > 0, and green stripes indicate concordant SVs between technologies. SD, segmental duplications; SR, simple repeats; RM, repeat masked (not SD or SR); US, unique sequence. (**D**) Length distribution of common SVs sites (AF > 5%) represented in assembly-based callset, including variants genotyped by using PanGenie and all common variants from population-scale studies from the Genome Aggregation Database (gnomAD-SV) and CCDG (insertions from CCDG omitted because of lack of data). Length distributions for all variants (not restricted to common) are provided in fig. S23. (**E** to **G**) Examples of lead SV-eQTLs (large symbols) in context of their respective genes, overlapping regulatory annotation, and other variants (small symbols). (E) An 89-bp insertion (chr10-133415975-INS-89) is linked to decreased expression of *MTG1* [$q = 4.10 \times 10^{-11}$, Beta = −0.55 (-0.51 to −0.59)]. (F) A 186-bp insertion (chr5-50299995-INS-186), overlapping an ENCODE enhancer mark (orange), is the lead variant associated with decreased expression of *EMB* [$q = 2.92 \times 10^{-6}$, Beta = −0.44 (−0.39 to −0.49)]. (G) A 1069-bp deletion (chr21-14088468-DEL-1069) downstream of *LIPI* is linked to increased expression of *LIPI* [q = 0.0022, Beta = 0.44 (0.38 to 0.50)].

of which were not captured by long-read assemblies (Fig. 5C and figs. S11 and S31). A large fraction of these calls maps to large repetitive regions such as segmental duplications that are not fully sequence resolved. We found that 42.5% (strict) and 59.9% (lenient) of PanGenie-genotypable SVs are absent from the short-read callset. We examined the distribution of common long-read SVs genotyped at >5% allele frequency (AF) across all the 3202 Illumina genomes against the short-read SVs from large population studies, including the Centers for Common Disease Genomics (CCDG) (*6*) and Genome Aggregation Database (gnomAD) (Fig. 5D and fig. S12) (*5*). The ability to genotype variation typically not detected in Illumina callsets is reflected in increased numbers of common SVs (AF > 5%)—particularly deletions below 250 bp and insertions under 1 kbp—genotyped by PanGenie but not seen in CCDG and gnomAD-SV, while also emphasizing the overall value of large-scale short-read datasets to capture rare variation and large CNVs in the population (fig. S31).

QTL analyses

We applied PanGenie genotypes (strict set) to systematically discover quantitative trait loci (QTL) associated with structural variation. First, we performed deep RNA-sequencing (RNA-seq) (>200 million fragments) of the corresponding 34 lymphoblastoid cell lines and integrated these data with 397 transcriptomes of 1000GP samples from GEUVADIS (*44*). We pursued cis expression quantitative trait loci (eQTL) and cis splicing quantitative trait loci (sQTL) mapping across the merged set of 427 donors, using a window of 1 Mbp centered around the gene or splice cluster, respectively, and tested all variants with a minor allele frequency (MAF) of ≥1% and at Hardy-Weinberg equilibrium (HWE) (exact test $P \geq 0.0001$). We considered 23,953 expressed genes (15,504 of which were protein-coding) and 36,100 splicing clusters (linked to 11,278 genes).

Using this design, we identified 58,152 indel-eQTLs (linked to 6748 distinct genes) and 2109 SV-eQTLs (linked to 1526 distinct genes) (table S31) at an FDR of 5%. The set includes 819 lead indel-eQTLs and 38 lead SV-eQTLs at distinct genes, respectively (table S31). In the sQTL analysis, we identified 3382 SV-sQTLs (FDR 5%, linked to 758 distinct genes) (table S32), of which 65 SV-sQTLs at distinct genes were the lead association at the locus (*18*). In line with prior studies (*23*, *45*), the lead variants are enriched for SVs [Fisher's exact eQTL

Downloaded from http://science.sciencemag.org/ on April 9, 2021

$P = 1.0 \times 10^{-6}$, odds ratio (OR) = 1.2; sQTL $P = 1.6 \times 10^{-4}$, OR = 1.2] as well as smaller indels (Fisher's exact eQTL: $P = 8.8 \times 10^{-113}$, OR = 1.2; sQTL: $P = 3.5 \times 10^{-72}$, OR = 1.2), whereas they are depleted for SNVs (Fisher's exact eQTL $P = 1.8 \times 10^{-118}$, OR = 0.84; sQTL: $P = 1.2 \times 10^{-75}$, OR = 0.84). Among SVs, deletions show the greatest effect when compared with insertion events (table S33) (*18*).

We overlapped lead SV-eQTLs with our Illumina-based discovery callset (*18*) and a recent large-scale SV study of 17,795 genomes (*6*) and found that 42% (16 out of 38 SVs) of the lead eQTL associations reported here are previously unidentified. Of these previously inaccessible SVs, 12 (75%) correspond to insertions (two Alu MEIs, three tandem duplications, and seven repeat expansions)—SV classes that are typically underascertained in short-read datasets (*1*). For example, one of our top previously unidentified lead SVs is an 89-bp VNTR insertion in the terminal intron of the mitochondrial ribosome-associated guanosine triphosphatase 1 gene (*MTG1*) (Fig. 5E) and is seen in conjunction with decreased expression. Similarly, we identified a 186-bp insertion in an ENCODE enhancer for B cell lymphomas, which is associated with reduced expression of the immunoglobulin superfamily gene embigin (*EMB*) (Fig. 5F). By contrast, we sequence-resolved a 1069-bp deletion located in an SD region downstream of the Lipase I gene (*LIPI*) (Fig. 5G) and found that it is associated with increased gene expression of *LIPI*. Single-nucleotide polymorphisms at this locus have been linked to heart rate in patients with heart failure with reduced ejection fraction in a previous genome-wide association study (GWAS) [$P = 9.0 \times 10^{-6}$, reported in (*46*)].

Ancestry and population genetic analyses

The availability of haplotype-phased assemblies provides an opportunity to explore the ancestry and population genetic properties of the genomes and SVs at multiple levels. We applied a machine-learning method (*47*) and developed a hidden Markov model to identify ancestry-informative SNVs and to assign ancestral segments per block on the basis of population genetic data from the Simons Genome Diversity Project (SGDP) (*18*, *48*). The two methods, as well as the different sequencing platforms, produce highly concordant results (>90%) (fig. S32). At the family level, we can accurately assign paternal and maternal haplotypes and distinguish recombination crossover events in the child compared with parental haplotypes (Fig. 6A).

At the population level, on average 87.2% of the assembled sequence can be assigned ancestry. 1000GP samples that originate from the African continent show the largest tracts of uniform ancestry (mean length = 23.6 cM) (Fig. 6B and fig. S33), in contrast to North and South American populations (mean length = 2.65 cM) (Fig. 6B and fig. S33) and South Asians (mean length=4.38 cM) (Fig. 6B), which is consistent with recent and more ancient admixture. For example, the African American, African Caribbean, and Admixed American 1000GP samples show the greatest diversity of ancestral segments (Fig. 6B and figs. S33 and S34), most likely as a result of the transatlantic slave trade and colonial-era migration (*49*).

Focusing on our more comprehensive genotyping of SVs into WGS data, we searched for population-stratified variants because these are potential candidates for local adaptation (*50*, *51*) that could not have been characterized in the original study of 1000GP populations (*15*). Using Fst as a metric, we found that the number of such population-stratified variants varies widely among different groups likely as a consequence of ancestral diversity (Africans), population bottlenecks (East Asians), and admixture (South Asians) (Fig. 6C). Restricting our analysis to SVs located within 5 kbp of genes and applying population branch statistics (PBS) (*51*), we identified 117 stratified SVs (PBS >3 standard deviations) (tables S34 and S35) and further characterize these by the number of base pairs deleted or inserted per locus (Fig. 6D). The greatest outlier is a 4.0-kbp insertion within the first intron of *LCT* (lactase gene) originally reported on the basis of fosmid sequencing from European samples (*52*). We determined that the corresponding insertion is ancestral (the human reference genome carries the derived deleted allele), the insertion harbors 11 predicted transcription factor binding sites, and the deletion likely occurred as a result of an Alu-mediated NAHR event ~520,000 years ago (fig. S35).

LCT variation is one of the most well-known genes under adaptive evolution among Europeans. The reported causal, derived allele of lactase persistence in Europeans (−13910*T; rs4988235) is in complete linkage disequilibrium (LD) [(*D'*) = 1] with the reference allele of this SV, and it will be interesting to determine the functional roles of these two mutations in lactase persistence (*53*). In other cases, the population-stratified variants are nested among known regulatory elements or intersect them directly, such as a 76-bp tandem repeat expansion in a PLEC intron, a cytoskeleton component, that is seen only in Africans (AF = 0.82) and Admixed Americans (AF = 0.06). Similarly, we identified a 2.8-kbp insertion mapping near potential repressor-binding sites in a *CLEC16A* intron, a gene that is associated with type 1 diabetes when disrupted (*54*). This variant shows a high frequency in American populations (AF = 0.28), with the highest PBS signal among Peruvians (AF = 0.39), but is rarely observed in other populations (AF ≤ 0.04). Further studies are needed to confirm functional effect; however, type 1 diabetes in Peruvians is among the highest in the world (*55*).

Discussion

We have generated a diversity panel of phased long-read human genome assemblies that has substantially improved SV discovery and will serve as the basis to construct new population-specific references. Previous large-scale efforts have largely been inferential and biased when it comes to the detection of SVs. We developed a method to discover all forms of genetic variation (PAV) directly through the comparison of assembled human genomes. By contrast, SV discovery from the 1000GP was indirect and limited given the frequent proximity of SVs to repeat sequences inaccessible to short reads (*15*, *23*). The 1000GP, for example, reported 69,000 SVs on the basis of the analysis of 2504 short-read sequenced genomes. By contrast, our analysis of 32 genomes (64 unrelated haplotypes) recovers 107,136 SVs, more than tripling the rate of discovery when compared with that of short-read Illumina SV analyses on the same samples (Fig. 2D). Recent large-scale short-read sequencing studies (*5*, *6*), interrogating tens of thousands of samples, show even lower SV sensitivity, reporting 5000 to 10,000 SVs per sample, when compared with our phased-assembly approach, which identifies 23,000 to 28,000 SVs per sample. This lack of sensitivity for SV discovery from short reads also affects common variation (AF > 5%), and we increased the amount of common SVs by 2.6-fold. The predominant source of this increase in sensitivity was among small SVs (<250 bp) localized to SDs and simple repeat sequences, where we observed a dramatic 8.4-fold increase in variant discovery (12,109 SVs per genome from long-read assembly, 1444 per genome from Illumina short-read alignment) (Fig. 5C). All discovered genetic variation is physically phased, and therefore SVs are fully integrated with their flanking SNVs.

Compared with previous reports based on short-read sequencing (*25*–*27*), a surprising finding has been the larger fraction of SVs (63%) now assigned to homology-based (>50 bp) mutation mechanisms, including HDR, NAHR, and VNTR. Breakpoint characterization with short-read data apparently biased early reports toward specific regions, concluding that <30% of SVs were driven by homology-based mutational mechanisms (*25*–*27*). Because a majority of unresolved structural variation still maps to large repeats, including centromeres and SDs subject to NAHR, we conclude that homology-based mutational mechanisms will contribute even further and are, therefore, the most predominant mode shaping the SV germline mutational landscape. Notwithstanding, access to fully assembled retrotransposons and their flanking sequence provides the largest collection of annotated source elements for both L1

Downloaded from http://science.sciencemag.org/ on April 9, 2021

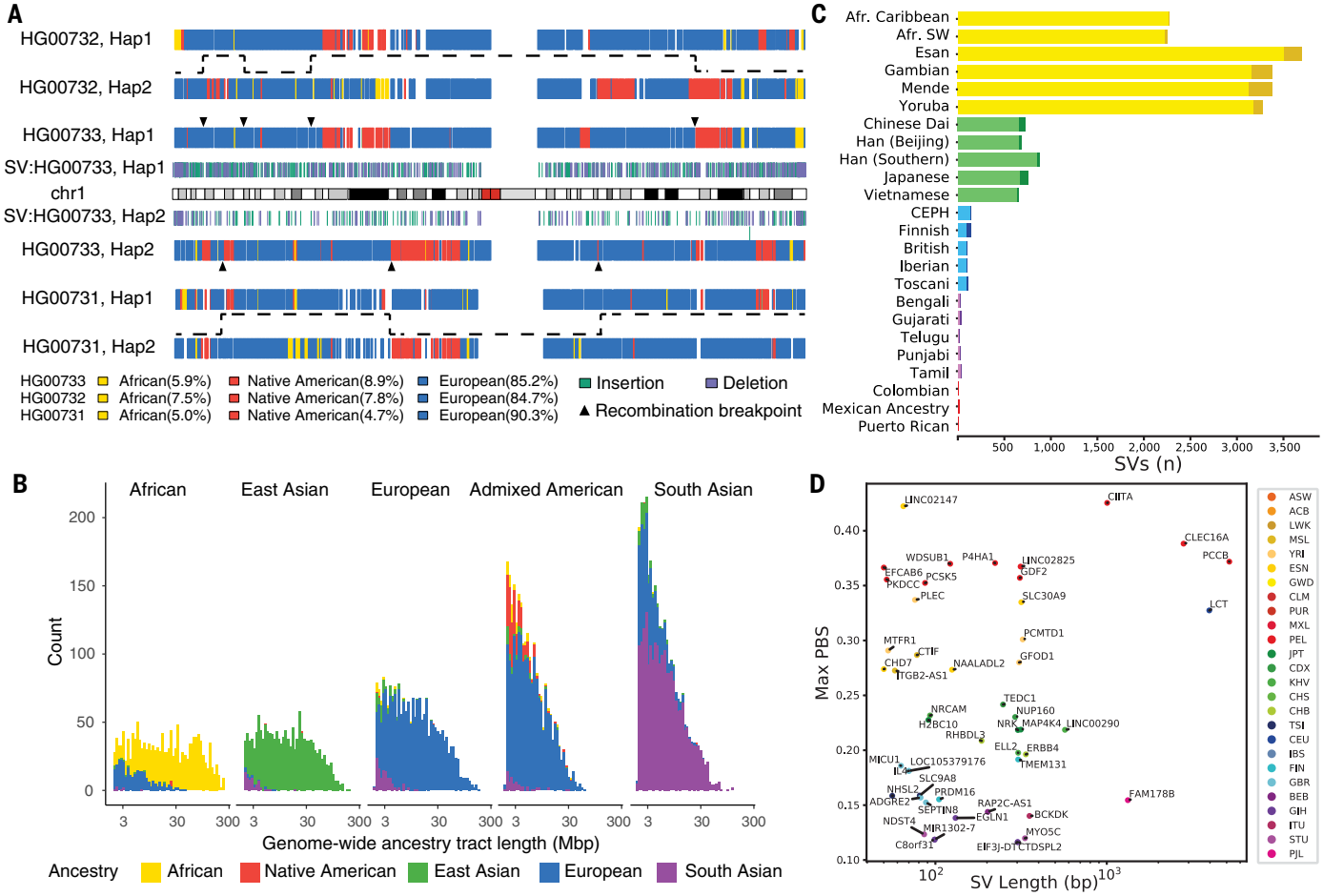


Fig. 6. Ancestry and population differentiation inferences by using haplotype-phased diploid assemblies. (**A**) Inferred local ancestries (*18*) for (top) maternal and (bottom) paternal haplotypes of HG00733 are compared with parental haplotypes (maternal, HG00732; paternal, HG00731). Ancestral segments are colored (African, yellow; Native American, red; and European, blue) and are consistent with the recent demographic history of the island (*18*). HG0733 SVs (≥50 bp; insertion, green; deletion, purple), inferred recombination breakpoints (triangles), and transmission of recombinant parental haplotypes (dashed lines) are shown. (**B**) Length distribution (log10) of ancestry tracts among the 64 genomes assigned to five superpopulations shows evidence of recent (Admixed American) and more ancient (South Asian) admixture. (**C**) Top population-specific Fst variants (dark color) and top superpopulation-specific Fst variants (light color). The number of stratified SVs differs by orders of magnitude depending on population. (**D**) Top SV PBS values within 5 kbp of genes identify SV candidates for selection and disease. A high PBS statistic suggests that AF differences among populations are a result of selection.

and SVA mobile elements. We found that 14% of SVA insertions are associated with transductions compared with 8% of L1s—a difference driven in part by the proclivity of SVAs to transduce sequences at their 5′ and 3′ ends. We found a surprisingly large number of L1 source elements (19%) with defective ORFs, suggesting either transcomplementation (*56*) or polymorphisms leading to the recent demise of these active source elements. Some of the youngest L1 copies (such as 6p22.1-1 and 2q24.1) have been reported to be rare polymorphisms able to mediate massive bursts of somatic retrotransposition in cancer genomes (*57*). This suggests that recently acquired hot L1s, which have not yet reached an equilibrium with our species, contribute disproportionately to disease-causing variation (*58*).

Genome-wide QTL scans can bridge the gap between molecular and clinical phenotypes and serve as a proxy for functional effects mediated by genetic variant classes (*23*, *44*, *59*). Taking advantage of the fully phased sequence-resolved genetic variation, we demonstrated this by applying PanGenie, a new pangenome-based genotyping method, to 3202 1000GP genomes, resulting in reliable genotype calls for 705,893 indels and up to 50,340 SVs (lenient genotype set). Of these, 59.9% are presently missed in multialgorithm short-read discovery callsets, and the majority (68.2%) of these previously unidentified SVs are insertions. Our work thus provides a framework for the discovery of eQTLs and disease-associated variants, with the potential to discriminate among SNVs, indels, and SVs as the most likely causal variants (lead variants) associated with human genetic traits. That 31.9% of SV-eQTLs and 48% of lead SV-eQTLs are rendered accessible to short reads only through the availability of our panel of haplotype-resolved assemblies testifies to the importance of this resource for future GWASs. Once again, among the lead SV-eQTLs, 75% are insertions, although there are also promising deletion eQTLs. For example, we identified a 1069-bp deletion eQTL near *LIPI*, a GWAS disease locus for cardiac failure (*46*). Summary-data–based Mendelian randomization analysis (SMR) (*60*) suggests that this SV-eQTL of *LIPI* may be driving this association (SMR adjusted $P = 5.6 \times 10^{-4}$).

Haplotype-resolved SVs with accurate genotypes will also facilitate evolutionary and population genetic studies of SVs, including estimations of the rates of recurrent mutation, population stratification, and selective sweeps. As part of this analysis, we identified 117 loci associated with genes where allele frequencies differ radically between populations and are candidates for local adaptation (*50*, *51*). Ancestral

Downloaded from http://science.sciencemag.org/ on April 9, 2021

reconstructions of haplotype-resolved SVs can be further extended to identify introgressed SVs from Neanderthals and Denisovans (*61*). Although archaic SNV haplotypes have been identified in modern-day humans, little is known regarding SV content given the degraded nature of ancient DNA. Combined with coalescent estimates of evolutionary age, it should now be possible to systematically identify associated introgressed SVs and assess them for signatures of adaptive evolution, as was recently demonstrated (*62*). Even though we estimate that 96% of SVs with an allele frequency above 2% have been theoretically discovered (*63*), a greater diversity of human genomes are required to adequately account for population differences, effects of selection, as well as archaic introgression. Our findings clearly indicate that genomes of African ancestry represent the deepest reservoir of untapped structural variation. Ongoing efforts over the next few years from the HGSVC, *All of Us*, and the Human Pangenome Reference Consortium (HPRC; https://humanpangenome.org) exploring the normal pattern of structural variation by using long-read sequences will be critical to better understand human genetic variation.

Currently, our understanding of the full spectrum of structural variation is not yet complete, despite the advances presented here. There are two important limitations. First, comparison with optical mapping data identifies hundreds of gene-rich regions near and within SDs harboring more complex forms of SVs that are still not fully resolved with long-read assembly. The remaining gaps in human genomes cluster, and a subset represents complex SV differences among human haplotypes. Second, only ~50% of our long-read discovery set of SVs can at present be reliably genotyped in short-read data by using PanGenie. Expanding the number of assembly-based haplotypes available as a pangenomic reference will likely mitigate this, but multiallelic VNTRs and STRs as well as SVs embedded in larger repeats such as SDs and centromeres are particularly problematic, and new methods are needed to characterize these. Recent advances coupling both HiFi and ultra-long-read Oxford Nanopore data show promise in resolving the sequence of these more complex regions from both haploid (*64*) and diploid human genome assemblies (*65*). Once a larger number of such complex regions are haplotype resolved across diversity panels of human genomes—and algorithms continue to evolve to exploit this information—we expect larger portions (fig. S36) of the human genome to become amenable to genotyping and association with human traits.

Methods summary

Libraries were prepared from high-molecular-weight DNA from lymphoblast lines (Coriell Institute). Long-read CLR and HiFi sequencing data (25 to 50×) were generated on the Sequel II platform (Pacific Biosciences) using 15-hour (CLR) or 30-hour (HiFi) movie times. Strand-seq data were produced from the same samples and used to identify and phase heterozygous SNVs [LongShot (*66*) and DeepVariant (*67*)] from the squashed genome assemblies [Peregrine (*68*) or Flye (*69*)]. StrandphaseR (*70*), SaaRclust (*71*), and WhatsHap (*72*, *73*) partitioned long reads into haplotypes to generate phased genome assemblies (*3*). MAPQ60 phased assembly contig coverage is estimated for autosomes (chromosomes 1 to 22) and the X chromosome to balance male and female comparisons, excluding regions of heterochromatin (Giemsa pos./var. staining) and unresolved reference sequence (N-gaps). We generated optical maps for 30 of the 32 samples on the basis of *DLE1* digestion (Bionano Genomics).

PAV was used to characterize SNVs, indels, and SVs compared with the human reference GRCh38. Inversions were detected by using Strand-seq (*1*, *9*, *38*), optical mapping data (Bionano Solve v3.5), and PAV, which detects inversion signatures by using a *k*-mer density approach to identify inner and outer breakpoints of flanking repeats without relying on alignment truncation. The diploid callset is created by merging two independent haploid callsets. We removed variants in collapses using Segmental Duplication Assembler (SDA) (*74*) and misaligned contig clusters then merged variants from all samples to create a nonredundant callset that was subsequently filtered by additional support (*18*). SVs required support from at least one of seven other sources, including read-based callers (MELT, PBSV, and PALMER) (*33*, *75*), optical mapping data, breakpoint *k*-mer analysis, and PAV replication with LRA (*76*). Indels required support from at least two of four sources, and SNVs required support from at least two of five sources. MEIs were primarily discovered by using PAV, which were then annotated by using MEIGA-PAV. In addition, Illumina and PacBio alignments were processed by using MELT and PALMER, respectively, in order to increase sensitivity for MEI discovery. Last, MEI calls across different platforms were merged into an integrated callset.

We estimated functional element depletion for SVs by simulation permuting SVs within their 1-Mbp bin 100,000 times and recording functional element hits for insertions and deletions for each functional category (CDS, 5′UTR, 3′UTR, promoter, proximal enhancer, distal enhancer, CTCF, and intron). SV hotspots were defined by searching for regions of increased SV density by using kernel density estimation implemented with the "hotspotter" function from the primatR package (*38*, *77*). Illumina WGS short reads (250-bp paired end) were generated (34.5-fold) (*18*) from 1000GP samples (2504 unrelated individuals and additional samples from children to form 602 trios). SVs were called from an ensemble of three methods—GATK-SV (*5*), SVTools (*6*), and Absinthe—and detailed comparisons between long- and short-read data were performed for the 31 matched samples (*18*).

We genotyped all 3202 genomes using PanGenie (*42*), which determines *k*-mer abundances from an input set of unaligned short reads and infers the genotypes of this short-read sample at all loci represented in the reference set. The method exploits both the linkage disequilibrium structure inherent to the reference haplotypes and the sequence resolution that they provide and, hence, makes full use of the haplotype resource provided. RNA-seq data QC was conducted with Trim Galore! (*78*) and mapped to the reference genome by using STAR (*79*), followed by gene-level quantification by using FeatureCounts (*80*) and quantification of splice events by using leafCutter (*81*). We mapped the effect of genetic variation on both expression levels and splicing ratios using a QTL mapping pipeline based on a linear mixed model implemented in LIMIX (*82*–*84*). We combined our QTL statistics with published GWAS results to assess the link among genetic variation, GWAS traits, and either gene expression or splicing ratios using SMR (*60*). To identify population-stratified SVs in the 26 populations, we computed the fixation index (F_{ST})–based PBS (*18*). For each focal population, we constructed population triplets by choosing sister- and outgroups inside and outside the continent where the focal population resides, respectively. For each focal population, we selected the maximum PBS per gene for all possible PBS triplets and selected the subset that are at least three standard deviations (*Z* transformation) beyond the PBS mean as potential targets of selection. Detailed descriptions of materials and methods are available in the supplementary materials (*18*).

REFERENCES AND NOTES

1. M. J. P. Chaisson *et al.*, Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019). doi: 10.1038/s41467-018-08148-z; pmid: 30992455
2. S. Garg *et al.*, Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.* (2020). doi: 10.1038/s41587-020-0711-0; pmid: 33288905
3. D. Porubsky *et al.*, Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.* 10.1038/s41587-020-0719-5 (2020). doi: 10.1038/s41587-020-0719-5; pmid: 33288906
4. P. A. Audano *et al.*, Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675.e19 (2019). doi: 10.1016/j.cell.2018.12.019; pmid: 30661756
5. R. L. Collins *et al.*, A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020). doi: 10.1038/s41586-020-2287-8; pmid: 32461652
6. H. J. Abel *et al.*, Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83–89 (2020). doi: 10.1038/s41586-020-2371-0; pmid: 32460305
7. A. M. Wenger *et al.*, Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019). doi: 10.1038/s41587-019-0217-9; pmid: 31406327

Downloaded from http://science.sciencemag.org/ on April 9, 2021

8. A. Sulovari *et al*., Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 23243–23253 (2019). doi: 10.1073/pnas.1912175116; pmid: 31659027
9. A. D. Sanders *et al*., Characterizing polymorphic inversions in human genomes by single-cell sequencing. *Genome Res.* **26**, 1575–1587 (2016). doi: 10.1101/gr.201160.115; pmid: 27472961
10. J. Xing *et al*., Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 17608–17613 (2006). doi: 10.1073/pnas.0603224103; pmid: 17101974
11. A. Damert *et al*., 5′-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. *Genome Res.* **19**, 1992–2008 (2009). doi: 10.1101/gr.093435.109; pmid: 19652014
12. Computational Pan-Genomics Consortium, Computational pan-genomics: Status, promises and challenges. *Brief. Bioinform.* **19**, 118–135 (2018). pmid: 27769991
13. B. Paten, A. M. Novak, J. M. Eizenga, E. Garrison, Genome graphs and the evolution of genome inference. *Genome Res.* **27**, 665–676 (2017). doi: 10.1101/gr.214155.116; pmid: 28360232
14. J. M. Eizenga *et al*., Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020). doi: 10.1146/annurev-genom-120219-080406; pmid: 32453966
15. A. Auton *et al*., A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). doi: 10.1038/nature15393; pmid: 26432245
16. M. Byrska-Bishop *et al*., High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. bioRxiv 430068 [Preprint] 7 February 2021. doi: 10.1101/2021.02.06.430068.
17. J. M. Zook *et al*., An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.* **37**, 561–566 (2019). doi: 10.1038/s41587-019-0074-6; pmid: 30936564
18. Materials and methods are available as supplementary materials.
19. J. Huddleston *et al*., Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017). doi: 10.1101/gr.214007.116; pmid: 27895111
20. L. Shi *et al*., Long-read sequencing and de novo assembly of a Chinese genome. *Nat. Commun.* **7**, 12065 (2016). doi: 10.1038/ncomms12065; pmid: 27356984
21. J.-S. Seo *et al*., De novo assembly and phasing of a Korean human genome. *Nature* **538**, 243–247 (2016). doi: 10.1038/nature20098; pmid: 27706134
22. J. E. Moore *et al*., Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020). doi: 10.1038/s41586-020-2493-4; pmid: 32728249
23. P. H. Sudmant *et al*., An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015). doi: 10.1038/nature15394; pmid: 26432246
24. C. M. B. Carvalho, J. R. Lupski, Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016). doi: 10.1038/nrg.2015.25; pmid: 26924765
25. D. F. Conrad *et al*., Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat. Genet.* **42**, 385–391 (2010). doi: 10.1038/ng.564; pmid: 20364136
26. H. Y. K. Lam *et al*., Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat. Biotechnol.* **28**, 47–55 (2010). doi: 10.1038/nbt.1600; pmid: 20037582
27. R. E. Mills *et al*., Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011). doi: 10.1038/nature09708; pmid: 21293372
28. M. J. P. Chaisson *et al*., Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015). doi: 10.1038/nature13907; pmid: 25383537
29. D. C. Hancks, H. H. Kazazian Jr., Roles for retrotransposon insertions in human disease. *Mob. DNA* **7**, 9 (2016). doi: 10.1186/s13100-016-0065-9; pmid: 27158268
30. E. C. Scott, S. E. Devine, The Role of Somatic L1 Retrotransposition in Human Cancers. *Viruses* **9**, 131 (2017). doi: 10.3390/v9060131; pmid: 28561751
31. B. Brouha *et al*., Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5280–5285 (2003). doi: 10.1073/pnas.0831042100; pmid: 12682288
32. C. R. Beck *et al*., LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159–1170 (2010). doi: 10.1016/j.cell.2010.05.021; pmid: 20602998
33. E. J. Gardner *et al*., The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017). doi: 10.1101/gr.218032.116; pmid: 28855259
34. B. Rodriguez-Martin *et al*., Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* **52**, 306–319 (2020). doi: 10.1038/s41588-019-0562-0; pmid: 32024998
35. H. Jung, J. K. Choi, E. A. Lee, Immune signatures correlate with L1 retrotransposition in gastrointestinal cancers. *Genome Res.* **28**, 1136–1146 (2018). doi: 10.1101/gr.231837.117; pmid: 29970450
36. J. M. C. Tubio *et al*., Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014). doi: 10.1126/science.1251343; pmid: 25082706
37. H. Wang *et al*., SVA elements: A hominid-specific retroposon family. *J. Mol. Biol.* **354**, 994–1007 (2005). doi: 10.1016/j.jmb.2005.09.085; pmid: 16288912
38. D. Porubsky *et al*., Recurrent inversion toggling and great ape genome evolution. *Nat. Genet.* **52**, 849–858 (2020). doi: 10.1038/s41588-020-0646-x; pmid: 32541924
39. M. C. Zody *et al*., Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat. Genet.* **40**, 1076–1083 (2008). doi: 10.1038/ng.193; pmid: 19165922
40. D. P. Locke *et al*., Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res.* **13**, 347–357 (2003). doi: 10.1101/gr.1003303; pmid: 12618365
41. B. C. Ballif *et al*., Expanding the clinical phenotype of the 3q29 microdeletion syndrome and characterization of the reciprocal microduplication. *Mol. Cytogenet.* **1**, 8 (2008). doi: 10.1186/1755-8166-1-8; pmid: 18471269
42. J. Ebler *et al*., Pangenome-based genome inference. bioRxiv 378133 [Preprint] 12 November 2020. .doi: 10.1101/2020.11.11.378133
43. X. Zhao *et al*., Expectations and blind spots for structural variation detection from short-read alignment and long-read assembly. bioRxiv 168831 [Preprint] 4 July 2020. .doi: 10.1101/2020.07.03.168831
44. T. Lappalainen *et al*., Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013). doi: 10.1038/nature12531; pmid: 24037378
45. C. Chiang *et al*., The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017). doi: 10.1038/ng.3834; pmid: 28369037
46. K. L. Evans *et al*., Genetics of heart rate in heart failure patients (GenHRate). *Hum. Genomics* **13**, 22 (2019). doi: 10.1186/s40246-019-0206-6; pmid: 31113495
47. B. K. Maples, S. Gravel, E. E. Kenny, C. D. Bustamante, RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013). doi: 10.1016/j.ajhg.2013.06.020; pmid: 23910464
48. S. Mallick *et al*., The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016). doi: 10.1038/nature18964; pmid: 27654912
49. R. A. Mathias *et al*., A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. *Nat. Commun.* **7**, 12522 (2016). doi: 10.1038/ncomms12522; pmid: 27725671
50. R. Nielsen *et al*., Darwinian and demographic forces affecting human protein coding genes. *Genome Res.* **19**, 838–849 (2009). doi: 10.1101/gr.088336.108; pmid: 19279335
51. X. Yi *et al*., Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010). doi: 10.1126/science.1190371; pmid: 20595611
52. J. M. Kidd *et al*., Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nat. Methods* **7**, 365–371 (2010). doi: 10.1038/nmeth.1451; pmid: 20440878
53. T. Bersaglieri *et al*., Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004). doi: 10.1086/421051; pmid: 15114531
54. S. A. Soleimanpour *et al*., The diabetes susceptibility gene Clec16a regulates mitophagy. *Cell* **157**, 1577–1590 (2014). doi: 10.1016/j.cell.2014.05.016; pmid: 24949970
55. S. N. Seclen, M. E. Rosas, A. J. Arias, C. A. Medina, Elevated incidence rates of diabetes in Peru: Report from PERUDIAB, a national urban population-based longitudinal study. *BMJ Open Diabetes Res. Care* **5**, e000401 (2017). doi: 10.1136/bmjdrc-2017-000401; pmid: 28878935
56. W. Wei *et al*., Human L1 retrotransposition: Cis preference versus trans complementation. *Mol. Cell. Biol.* **21**, 1429–1439 (2001). doi: 10.1128/MCB.21.4.1429-1439.2001; pmid: 11158327
57. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020). doi: 10.1038/s41586-020-1969-6; pmid: 32025007
58. R. Cordaux, M. A. Batzer, The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703 (2009). doi: 10.1038/nrg2640; pmid: 19763152
59. GTEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020). doi: 10.1126/science.aaz1776; pmid: 32913098
60. Z. Zhu *et al*., Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016). doi: 10.1038/ng.3538; pmid: 27019110
61. S. Sankararaman *et al*., The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354–357 (2014). doi: 10.1038/nature12961; pmid: 24476815
62. P. Hsieh *et al*., Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science* **366**, eaax2083 (2019). doi: 10.1126/science.aax2083; pmid: 31624180
63. M. A. Eberle, L. Kruglyak, An analysis of strategies for discovery of single-nucleotide polymorphisms. *Genet. Epidemiol.* **19** (suppl. 1), S29–S35 (2000). doi: 10.1002/1098-2272(2000)19:1+<::AID-GEPI5>3.0.CO;2-P; pmid: 11055367
64. K. H. Miga *et al*., Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020). doi: 10.1038/s41586-020-2547-7; pmid: 32663838
65. G. A. Logsdon *et al*., The structure, function, and evolution of a complete human chromosome 8. bioRxiv 285395 [Preprint] 8 September 2020. .doi: 10.1101/2020.09.08.285395
66. P. Edge, V. Bansal, Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat. Commun.* **10**, 4660 (2019). doi: 10.1038/s41467-019-12493-y; pmid: 31604920
67. R. Poplin *et al*., A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018). doi: 10.1038/nbt.4235; pmid: 30247488
68. C.-S. Chin, A. Khalak, Human Genome Assembly in 100 Minutes. bioRxiv 705616 [Preprint] 17 July 2019. .doi: 10.1101/705616
69. M. Kolmogorov, J. Yuan, Y. Lin, P. A. Pevzner, Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019). doi: 10.1038/s41587-019-0072-8; pmid: 30936562
70. D. Porubsky *et al*., Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat. Commun.* **8**, 1293 (2017). doi: 10.1038/s41467-017-01389-4; pmid: 29101320
71. M. Ghareghani *et al*., Strand-seq enables reliable separation of long reads by chromosome via expectation maximization. *Bioinformatics* **34**, i115–i123 (2018). doi: 10.1093/bioinformatics/bty290; pmid: 29949971
72. M. Patterson *et al*., WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J. Comput. Biol.* **22**, 498–509 (2015). doi: 10.1089/cmb.2014.0157; pmid: 25658651
73. M. Martin *et al*., WhatsHap: fast and accurate read-based phasing. bioRxiv 085050 [Preprint] 14 November 2016. doi: 10.1101/085050
74. M. R. Vollger *et al*., Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019). doi: 10.1038/s41592-018-0236-3; pmid: 30559433
75. W. Zhou *et al*., Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res.* **48**, 1146–1163 (2020). doi: 10.1093/nar/gkz1173; pmid: 31853540
76. J. Ren, M. J. P. Chaisson, LRA: the long read aligner for sequences and contigs. bioRxiv 383273 [Preprint] 17 November 2020. .doi: 10.1101/2020.11.15.383273
77. B. Bakker *et al*., Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biol.* **17**, 115 (2016). doi: 10.1186/s13059-016-0971-7; pmid: 27246460
78. F. Krueger, Trim Galore: a wrapper tool around Cutadapt and FastQC. *Trim Galore!* (2012); www.bioinformatics.babraham.ac.uk/projects/trim_galore.
79. A. Dobin *et al*., STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013). doi: 10.1093/bioinformatics/bts635; pmid: 23104886

Downloaded from http://science.sciencemag.org/ on April 9, 2021

80. Y. Liao, G. K. Smyth, W. Shi, The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41**, e108 (2013). doi: 10.1093/nar/gkt214; pmid: 23558742
81. Y. I. Li *et al*., Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018). doi: 10.1038/s41588-017-0004-9; pmid: 29229983
82. F. P. Casale, B. Rakitsch, C. Lippert, O. Stegle, Efficient set tests for the genetic analysis of correlated traits. *Nat. Methods* **12**, 755–758 (2015). doi: 10.1038/nmeth.3439; pmid: 26076425
83. B. A. Mirauta *et al*., Population-scale proteome variation in human induced pluripotent stem cells. *eLife* **9**, e57390 (2020). doi: 10.7554/eLife.57390; pmid: 32773033
84. M. J. Bonder *et al*., Systematic assessment of regulatory effects of human disease variants in pluripotent cells. bioRxiv 784967 [Preprint] 4 October 2019. doi: 10.1101/784967.
85. M. S. García, Multiple sequence alignments of full-length L1 elements with evidence of retrotransposition activity. Zenodo (2021); doi: 10.5281/zenodo.4475905.
86. P. A. Audano, HGSVC Key Callset Resources. Zenodo (2020); doi: 10.5281/zenodo.4268828.
87. M. J. Bonder, HGSVC2 full eQTL results. Zenodo (2020); doi: 10.5281/zenodo.4271574.
88. P. Ebert, HGSVC2 project code contributions. Zenodo (2021); doi: 10.5281/zenodo.4482026.
89. B. R. Martín, MEIGA-tk/MEIGA-PAV: MEIGA-PAV. Zenodo (2021); doi: 10.5281/zenodo.4487121.

ACKNOWLEDGMENTS

We thank T. Brown for assistance in editing this manuscript and K. Hoekzema and C. Baker for the preparation of cell line DNA. We also recognize the computational support (P.H. Rehs and C. Siebert) and infrastructure provided by the Centre for Information and Media Technology (ZIM) at the University of Düsseldorf, the EMBL IT Services, and additional computational analyses (C. Alkan, F. Hormozdiari, D. S. Gordon, and S. Murali). We thank M. Paulsen from the EMBL Flow Cytometry Core Facility as well as J. Zimmermann and V. Benes from the EMBL Genomics Core Facility for assisting in Strand-seq sample preparation and sequencing. We thank the Human Pangenome Reference Consortium (HPRC) for use of the publicly available GIAB sequence data for the Ashkenazim benchmark sample HG002/NA24385. We are grateful to the people who generously contributed samples as part of the 1000 Genomes Project (1000GP). We thank the Pan-UKB project and UK Biobank for making the GWAS results available. **Funding:** Funding for this research project by the Human Genome Structural Variation Consortium (HGSVC) came from the following grants: National Institutes of Health (NIH) U24HG007497 (to C.L., E.E.E., J.O.K., T.M., M.E.T., M.B.G., S.E.D., I.M.H., R.E.M., and M.J.P.C.), NIH R01HG002898 (to S.E.D.), NIH R01HD081256 (to M.E.T.), NIH 1R01HG007068-01A1 (to R.E.M.), NIH R01HG002385 and HG010169 (to E.E.E.), R01MH115957 (to M.E.T.), NIH R15HG009565 (to X.S.), NIH 1U01HG010973 (to M.J.P.C., T.M., E.E.E., and J.O.K.), NIH 1R35GM138212 and a subaward from 1OT3HL147154 (to Z.C.), NIH/NHGRI Pathway to Independence Award K99HG011041 (to PH.H.), the German Research Foundation (391137747 and 395192176 to T.M.), the European Research Council (Consolidator grant 773026 to J.O.K. and Starting Grant 716290 to J.M.C.T.), the German Federal Ministry for Research and Education (BMBF 031L0184 to J.O.K. and T.M. and BMBF 031L0181A to J.O.K), the Spanish Ministry of Economy, Industry and Competitiveness (SAF2015-66368-P to J.M.C.T.), the Wellcome Trust grant WT104947/Z/14/Z and the European Molecular Biology Laboratory core funding (to B.R.-M., M.J.B., T.R., W.H., S.F., B.R., P.H., P.F., O.S., A.D.S., J.O.K.), National Science Foundation of China (32070663 to K.Y., 61702406 to X.Y.), and National Key R&D Program of China (2017YFC0907500 to K.Y., 2018YFC0910400 to K.Y., and 2018ZX10302205 to X.Y.). This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, and 031A532B). E.E.E. is an investigator of the Howard Hughes Medical Institute. J.O.K. and J.M.C.T. are European Research Council (ERC) investigators. C.L. was a distinguished Ewha Womans University Professor supported, in part, by an Ewha Womans University research grant for 2019–2020. Also, this study was supported, in part, by funds from The First Affiliated Hospital of Xi'an Jiaotong University (to C.L.). A.C., W.E.C., and M.C.Z. were supported in part by a Centers for Common Disease Genomics (CCDG) grant from the National Human Genome Research Institute (UM1HG008901). M.S. is supported by a PhD fellowship from Xunta de Galicia (Spain). A.D.S. received postdoctoral research funding through the Alexander von Humboldt Foundation. Illumina sequencing data from the 1000GP samples were generated at the New York Genome Center with funds provided by NHGRI Grants 3UM1HG008901-03S1 and 3UM1HG008901-04S2. **Authors contributions:** PacBio production sequencing: K.M.M., A.P.L., Q.Z., L.J.T., and S.E.D. Strand-seq production: A.D.S., B.R., P.H., and J.O.K. Phased genome assembly: P.E., P.A.A., D.P., Q.Z., F.Y., W.T.H., and T.M. Assembly analysis: P.E. Assembly-based variant calling: P.A.A. Variant QC, merging, and annotation: P.A.A., T.R., M.J.P.C., J.R., T.L., Z.C., Y.C., K.Y., J.L., X.Y., and J.O.K. Assembly scaffolding: F.Y., D.P., and P.E. Additional long-read callsets: P.A.A., Y.C., Z.C., W.T.H., J.R., and A.M.W. Short-read SV calling and merging: X.Z., Q.Z., H.J.A., H.B., N.T.C., W.E.C., A.C., U.S.E., S.E.D., I.M.H., W.T.H., A.A.R., M.C.Z., and M.E.T. Bionano Genomics SV discovery and analysis: F.Y., J.L., and A.R.H. Strand-seq inversion detection and genotyping: D.P., W.T.H., H.A., M.G., T.M., A.D.S., and J.O.K. MEI discovery and integration: B.R.-M., W.Z., M.S., N.T.C., J.M.C.T., J.O.K., R.E.M., and S.E.D. Variant hotspot analysis: D.P. and E.E.E. Breakpoint analysis: S.K., J.L., X.Y., M.G., K.Y., and J.O.K. PanGenie genotyping: J.E. and T.M. Illumina genotype analysis: J.E., X.Z., W.E.C., P.E., T.R., P.A.A., H.B., J.O.K., M.E.T., M.C.Z., and T.M. RNA-seq and QTL analysis: M.J.B., A.S., Z.M., J.C., C.L., M.B.-B., A.O.B., O.S., Y.I.L., X.S., M.C.Z., and J.O.K. Ancestry and population genetic analyses: PH.H., R.S.M., P.A.A., T.M., and E.E.E. Data archiving: S.F., P.A.A., K.M.M., and P.F. Organization of supplementary materials: Q.Z. and C.L. Display items: P.A.A., P.E., J.E., A.R.H., PH.H., R.S.M., T.M., D.P., T.R., B.R.-M., M.S., F.Y., X.Z., and W.Z. Manuscript writing: P.A.A., P.E., B.R.-M., A.S., D.P., PH.H., Q.Z., F.Y., A.R.H., J.L., M.E.T., M.J.B., X.S., S.E.D., J.O.K., T.M., and E.E.E. HGSVC Co-chairs: C.L., J.O.K., and E.E.E. **Competing interests:** A.R.H. and J.L. are employees and shareholders of Bionano Genomics. A.M.W. is an employee and shareholder of Pacific Biosciences. M.C.Z. is a shareholder of Merck & Co. and Thermo Fisher Scientific. P.F. is a member of the Scientific Advisory Boards of Fabric Genomics and Eagle Genomics. A.D.S., J.O.K., T.M., M.G., and D.P. have a pending patent application (European Patent Office application number EP 19 169 090.8) relevant to the subject matter (method relevant to Strand-seq). **Data and materials availability:** Data files used by the project are available by FTP at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2, and its complete description may be found at the data portal www.internationalgenome.org/data-portal/data-collection/hgsvc2. Primary data are available at INSDC under the following accessions and project IDs: Illumina high-coverage genomic sequence (PRJEB37677), HiC and RNA-seq (ERP123231), Bionano Genomics (ERP124807), PacBio (PRJEB36100, ERP125611 and PRJNA698480), and Strand-seq (PRJEB39750). Alignments used for L1 phylogenetic tree construction (Fig. 3), merged PAV callsets, full eQTL results, and project code are available at Zenodo (*85*–*89*). The following cell lines and DNA samples were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: [NA06984, NA06985, NA06986, NA06989, NA06991, NA06993, NA06994, NA06995, NA06997, NA07000, NA07014, NA07019, NA07022, NA07029, NA07031, NA07034, NA07037, NA07045, NA07048, NA07051, NA07055, NA07056, NA07340, NA07345, NA07346, NA07347, NA07348, NA07349, NA07357, NA07435, NA10830, NA10831, NA10835, NA10836, NA10837, NA10838, NA10839, NA10840, NA10842, NA10843, NA10845, NA10846, NA10847, NA10850, NA10851, NA10852, NA10853, NA10854, NA10855, NA10856, NA10857, NA10859, NA10860, NA10861, NA10863, NA10864, NA10865, NA11829, NA11830, NA11831, NA11832, NA11839, NA11840, NA11843, NA11881, NA11882, NA11891, NA11892, NA11893, NA11894, NA11917, NA11918, NA11919, NA11920, NA11930, NA11931, NA11932, NA11933, NA11992, NA11993, NA11994, NA11995, NA12003, NA12004, NA12005, NA12006, NA12043, NA12044, NA12045, NA12046, NA12056, NA12057, NA12058, NA12144, NA12145, NA12146, NA12154, NA12155, NA12156, NA12234, NA12239, NA12248, NA12249, NA12264, NA12272, NA12273, NA12274, NA12275, NA12282, NA12283, NA12286, NA12287, NA12329, NA12335, NA12336, NA12340, NA12341, NA12342, NA12343, NA12344, NA12347, NA12348, NA12375, NA12376, NA12383, NA12386, NA12399, NA12400, NA12413, NA12414, NA12485, NA12489, NA12546, NA12707, NA12708, NA12716, NA12717, NA12718, NA12739, NA12740, NA12748, NA12749, NA12750, NA12751, NA12752, NA12753, NA12760, NA12761, NA12762, NA12763, NA12766, NA12767, NA12775, NA12776, NA12777, NA12778, NA12801, NA12802, NA12812, NA12813, NA12814, NA12815, NA12817, NA12818, NA12827, NA12828, NA12829, NA12830, NA12832, NA12842, NA12843, NA12864, NA12865, NA12872, NA12873, NA12874, NA12875, NA12877, NA12878, NA12889, NA12890, NA12891, NA12892].

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/372/6537/eabf7117/suppl/DC1
Materials and Methods
Figs. S1 to S103
Tables S1 to S56
References (*90*–*199*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from Bio-protocol.

13 November 2020; accepted 9 February 2021
Published online 25 February 2021
10.1126/science.abf7117

Downloaded from http://science.sciencemag.org/ on April 9, 2021

Science

Haplotype-resolved diverse human genomes and integrated analysis of structural variation

Peter Ebert, Peter A. Audano, Qihui Zhu, Bernardo Rodriguez-Martin, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jana Ebler, Weichen Zhou, Rebecca Serra Mari, Feyza Yilmaz, Xuefang Zhao, PingHsun Hsieh, Joyce Lee, Sushant Kumar, Jiadong Lin, Tobias Rausch, Yu Chen, Jingwen Ren, Martin Santamarina, Wolfram Höps, Hufsah Ashraf, Nelson T. Chuang, Xiaofei Yang, Katherine M. Munson, Alexandra P. Lewis, Susan Fairley, Luke J. Tallon, Wayne E. Clarke, Anna O. Basile, Marta Byrska-Bishop, André Corvelo, Uday S. Evani, Tsung-Yu Lu, Mark J. P. Chaisson, Junjie Chen, Chong Li, Harrison Brand, Aaron M. Wenger, Maryam Ghareghani, William T. Harvey, Benjamin Raeder, Patrick Hasenfeld, Allison A. Regier, Haley J. Abel, Ira M. Hall, Paul Flicek, Oliver Stegle, Mark B. Gerstein, Jose M. C. Tubio, Zepeng Mu, Yang I. Li, Xinghua Shi, Alex R. Hastie, Kai Ye, Zechen Chong, Ashley D. Sanders, Michael C. Zody, Michael E. Talkowski, Ryan E. Mills, Scott E. Devine, Charles Lee, Jan O. Korbel, Tobias Marschall and Evan E. Eichler

Science **372** (6537), eabf7117.
DOI: 10.1126/science.abf7117originally published online February 25, 2021

Resolving genomic structural variation

Many human genomes have been reported using short-read technology, but it is difficult to resolve structural variants (SVs) using these data. These genomes thus lack comprehensive comparisons among individuals and populations. Ebert *et al.* used long-read structural variation calling across 64 human genomes representing diverse populations and developed new methods for variant discovery. This approach allowed the authors to increase the number of confirmed SVs and to describe the patterns of variation across populations. From this dataset, they identified quantitative trait loci affected by these SVs and determined how they may affect gene expression and potentially explain genome-wide association study hits. This information provides insights into patterns of normal human genetic variation and generates reference genomes that better represent the diversity of our species.

Science, this issue p. eabf7117

ARTICLE TOOLS http://science.sciencemag.org/content/372/6537/eabf7117

SUPPLEMENTARY MATERIALS http://science.sciencemag.org/content/suppl/2021/02/24/science.abf7117.DC1

REFERENCES This article cites 193 articles, 46 of which you can access for free
http://science.sciencemag.org/content/372/6537/eabf7117#BIBL

PERMISSIONS http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.

Copyright © 2021 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works

Downloaded from http://science.sciencemag.org/ on April 9, 2021