Title

Genomic regions associated with microdeletion/microduplication syndromes exhibit extreme diversity of structural variation

Journal

Genetics, 217(2)

ISSN

0016-6731

Authors

Mostovoy, Yulia
Yilmaz, Feyza
Chow, Stephen K
et al.

Publication Date

2021-03-31

DOI

10.1093/genetics/iyaa038

Peer reviewed

OXFORD GENETICS

GENETICS, 2021, 217(2), iyaa038
DOI: 10.1093/genetics/iyaa038
Advance Access Publication Date: 6 January 2021
Investigation

Genomic regions associated with microdeletion/microduplication syndromes exhibit extreme diversity of structural variation

Yulia Mostovoy,[1,†] Feyza Yilmaz,[2,3,†] Stephen K. Chow,[1] Catherine Chu,[1] Chin Lin,[1] Elizabeth A. Geiger,[3] Naomi J. L. Meeks,[3] Kathryn C. Chatfield,[3,4] Curtis R. Coughlin II,[3] Urvashi Surti,[5] Pui-Yan Kwok,[1,6,7,*] and Tamim H. Shaikh[3,*]

[1]Cardiovascular Research Institute, UCSF School of Medicine, San Francisco, CA 94143, USA
[2]Department of Integrative Biology, University of Colorado Denver, Denver, CO 80204, USA
[3]Department of Pediatrics, Section of Clinical Genetics and Metabolism, University of Colorado School of Medicine, Aurora, CO 80045, USA
[4]Department of Pediatrics, Section of Cardiology, University of Colorado School of Medicine, Aurora, CO 80045, USA
[5]Department of Pathology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA
[6]Department of Dermatology, UCSF School of Medicine, San Francisco, CA 94143, USA
[7]Institute for Human Genetics, UCSF School of Medicine, San Francisco, CA 94143, USA

[†]These two authors contributed equally to this work.

[*]Corresponding author: Tamim.Shaikh@cuanschutz.edu (T.H.S.); Pui.Kwok@ucsf.edu (P.-Y.K.)

Abstract

Segmental duplications (SDs) are a class of long, repetitive DNA elements whose paralogs share a high level of sequence similarity with each other. SDs mediate chromosomal rearrangements that lead to structural variation in the general population as well as genomic disorders associated with multiple congenital anomalies, including the 7q11.23 (Williams–Beuren Syndrome, WBS), 15q13.3, and 16p12.2 microdeletion syndromes. Population-level characterization of SDs has generally been lacking because most techniques used for analyzing these complex regions are both labor and cost intensive. In this study, we have used a high-throughput technique to genotype complex structural variation with a single molecule, long-range optical mapping approach. We characterized SDs and identified novel structural variants (SVs) at 7q11.23, 15q13.3, and 16p12.2 using optical mapping data from 154 phenotypically normal individuals from 26 populations comprising five super-populations. We detected several novel SVs for each locus, some of which had significantly different prevalence between populations. Additionally, we localized the microdeletion breakpoints to specific paralogous duplicons located within complex SDs in two patients with WBS, one patient with 15q13.3, and one patient with 16p12.2 microdeletion syndromes. The population-level data presented here highlights the extreme diversity of large and complex SVs within SD-containing regions. The approach we outline will greatly facilitate the investigation of the role of inter-SD structural variation as a driver of chromosomal rearrangements and genomic disorders.

Keywords: segmental duplications; genome mapping; structural variation; genomic disorders

Introduction

The sequencing and assembly of the draft human reference genome proved to be a tipping point in the field of human genetics; combined with the advent of affordable high-throughput short-read sequencing, it allowed the identification of variants in hundreds of thousands of individuals by aligning their sequencing reads directly to and comparing them with the reference genome. While genome sequencing has revealed the wide genetic diversity of human populations, short-read sequencing is unreliable for the analysis of areas of the human genome containing long, complex repetitive DNA elements, including telomeres, centromeres, and regions known as low copy repeats or segmental duplications (SDs). SDs are regions of ≥1000 bp that have two or more copies across the genome with ≥90% sequence identity (Bailey *et al.* 2001; Lander *et al.* 2001) and comprise ∼5% of the human genome. SDs are often organized in blocks with complex internal structure, containing duplicated genomic segments, referred to as duplicons (Jiang *et al.* 2007), that vary in order and orientation. A biologically significant subset of SDs are larger in size (tens or even hundreds of kb) with extremely high sequence identity (>96%) (Sharp *et al.* 2005) and are involved in recurrent genomic rearrangements (Lupski 1998; Bailey *et al.* 2001). Due to their length and sequence identity, these regions cannot be resolved using short-read sequencing, and in many cases, even with longer-read sequencing (Vollger *et al.* 2019).

The length and high sequence similarity of SDs make them an excellent substrate for nonallelic homologous recombination (NAHR), which occurs between highly identical paralogous copies

Received: September 05, 2020. **Accepted:** December 18, 2020

of SDs and results in different types of structural variation (SV), including inversions, microdeletions, and microduplications (Stankiewicz and Lupski 2010; Carvalho and Lupski 2016). SDs are therefore hotspots of genomic rearrangements and complex SVs (Levy-Sakin *et al.* 2019), some of which give rise to genomic disorders caused by copy number variation (CNV) of dosage-sensitive or developmentally important genes (Lupski 1998, 2009; Bailey *et al.* 2001). Some of the well-studied examples include microdeletions at 7q11.23 known as Williams–Beuren syndrome (WBS, MIM #194050), 15q13.3 microdeletion syndrome (MIM #612001), 16p12.2 microdeletion syndrome (MIM #136570), and 22q11.2 Deletion Syndrome (MIM #188400), formerly known as DiGeorge syndrome (Peoples *et al.* 2000; Shaikh 2000; Sharp *et al.* 2008; Girirajan *et al.* 2010). Microdeletion breakpoints in these syndromes are located in SDs, where NAHR between paralogous copies results in the deletion of genes within the interstitial (unique) region (Emanuel and Shaikh 2001; Eichler 2002; Carvalho and Lupski 2016).

Since SD-containing regions cannot be resolved using short-read sequencing technology, other techniques have been harnessed to reconstruct the structure of these regions. Structural variations have previously been detected within the SDs in 7q11.2, 15q13.3, and 16p12.2 using lower-resolution techniques including fluorescence *in situ* hybridization (FISH) analysis (Osborne *et al.* 2001; Cuscó *et al.* 2008; Sharp *et al.* 2008) and array comparative genomic hybridization (aCGH) (Antonacci *et al.* 2010) and targeted sequencing (Antonacci *et al.* 2014). However, past efforts to fully reconstruct SD-containing regions have required a scaffold of large-insert bacterial artificial chromosomes (BACs) in combination with long- and short-read sequencing technologies (Antonacci *et al.* 2014; Huddleston *et al.* 2014; Steinberg *et al.* 2014), which would be both time and cost prohibitive for characterizing the SD-containing regions of a large number of samples. Furthermore, an approach to assemble SDs using only high-throughput long-read sequencing had trouble reconstructing paralogs longer than ~50 kb (Vollger *et al.* 2019).

Despite accumulating evidence for the role of SDs in disease-causing genomic rearrangements, their highly identical sequences, large size, and complex structures have made it difficult to elucidate their content. In this study, we demonstrate the use of the single molecule optical mapping technique to detect SVs over a wide range of sizes at the SD-containing regions of 7q11.23, 15q13.3, and 16p12.2 using the OMGenSV pipeline (Demaerel *et al.* 2019). The high-throughput nature of the technique allowed us to analyze a diverse control dataset of 154 individuals; thus, we were able to determine the prevalence of SD configurations across different populations, including Africans, Americans, Europeans, and East and South Asians. Additionally, we applied our approach to patient samples with 7q11.23, 15q13.3, and 16p12.2 microdeletions and were able to narrow down the microdeletion breakpoints to specific duplicons within the larger, complex SDs. The ability to probe the internal structure of complex regions in a high-throughput manner and at low cost, using the techniques presented here, will greatly aid the study of structural variation within SDs and will help elucidate its relationship to chromosomal rearrangements, both in the general population and in individuals with genomic disorders.

Methods

Human subjects and cell lines

Patient blood samples were obtained after informed consent, under the Institutional Review Board approved research protocol (COMIRB # 07-0386) at the University of Colorado Denver, School of Medicine.

High-molecular-weight DNA extraction

High-molecular-weight DNA for genome mapping was obtained from whole blood. White blood cells were isolated from whole blood samples using a ficoll-paque plus (GE Healthcare) gradient. The buffy coat layer was transferred to a new tube and washed twice with Hank's balanced salt solution (HBSS, Gibco Life Technologies). A small aliquot was removed to obtain a cell count before the second wash. The remaining cells were resuspended in RPMI (Gibco Life Technologies) containing 10% fetal bovine serum (Sigma) and 10% DMSO (Sigma). Cells were embedded in thin low-melting-point agarose plugs (CHEF Genomic DNA Plug Kit, BioRad). Subsequent handling of the DNA followed protocols from Bionano Genomics using the Bionano Prep Blood and Cell Culture DNA Isolation Kit. The agarose plugs were incubated with proteinase K at 50°C overnight. The plugs were washed and then solubilized with GELase (Epicentre). The purified DNA was subjected to 45 min of drop dialysis, allowed to homogenize at room temperature overnight, and then quantified using a Qubit dsDNA BR Assay Kit (Molecular Probes/Life Technologies). DNA quality was assessed using pulsed-field gel electrophoresis.

DNA labeling

7q11.23 proband DNA and 16p12.2 proband DNA were labeled using the IrysPrep Reagent Kit (Bionano Genomics). Specifically, 300 ng of purified genomic DNA was nicked with 10 U of nicking endonuclease Nt.BspQI [New England BioLabs (NEB)] at 37°C for 2 h in buffers BNG3 or BNG2, respectively. The nicked DNA was labeled with a fluorescent-dUTP nucleotide analog using Taq polymerase (NEB) for 1 h at 72°C. After labeling, the nicks were ligated with Taq ligase (NEB) in the presence of dNTPs. The backbone of fluorescently labeled DNA was counterstained with YOYO-1 (Invitrogen).

15q13.3 proband DNA was labeled using the Bionano Prep Early Access Direct Labeling and Staining (DLS) Kit (Bionano Genomics). A total of 750 ng of purified genomic DNA was labeled by incubating with DL-Green dye and DLE-1 Enzyme in DLE-1 Buffer for 2 h at 37°C, followed by heat inactivation of the enzyme for 20 min at 70°C. The labeled DNA was treated with Proteinase K at 50°C for 1 h, and excess DL-Green dye was removed by membrane adsorption. The DNA was stored at 4°C overnight to facilitate DNA homogenization and then quantified using a Qubit dsDNA HS Assay Kit (Molecular Probes/Life Technologies). The labeled DNA was stained with an intercalating dye and left to stand at room temperature for at least 2 h before loading onto a Bionano Chip.

Data collection and assembly

The DNA was loaded onto the Bionano Genomics IrysChip for the 16p12.2 proband sample and the Saphyr Chip for all other probands and was linearized and visualized by the Irys or Saphyr systems, respectively. The DNA backbone length and locations of fluorescent labels along each molecule were detected using the Irys or Saphyr system software. Single-molecule maps were assembled *de novo* into genome maps using the assembly pipeline developed by Bionano Genomics with default settings (Cao *et al.* 2014).

Cataloging and genotyping of structural variation at loci of interest

Structural variation at the 7q11.23, 15q13.3, and 16p12.2 loci was analyzed in a dataset of 154 phenotypically normal individuals from 26 diverse subpopulations, with genome maps labeled using the Nt.BspQI single-strand nickase enzyme (Levy-Sakin *et al.* 2019) (NCBI BioProject PRJNA418343). For 15q13.3, one of the duplicons contained a fragile site wherein two Nt.BspQI label sites were in close proximity on opposite strands, resulting in consistent DNA breakage between the sites. For a complete analysis that could span that position, we included a dataset labeled with the DLE-1 enzyme, which uses an epigenetic label rather than nicking DNA and therefore does not produce fragile sites (Maggiolini *et al.* 2019). This DLE-1 dataset contained 52 samples from the original dataset, with two samples per subpopulation (Wong *et al.* 2020, NCBI BioProject PRJNA611454).

Structural variation at the three loci was assessed using the Optical Maps to Genotype Structural Variation (OMGenSV) package as described in Demaerel *et al.* (2019) with minor modifications. To create a catalog of configurations for each locus, assembled contigs were visualized in order to avoid the noise of the raw molecule data. Every sample's *de novo* assembly folder was mined for assembled contigs aligned to the reference, and these were merged into a single file for the entire dataset. Contigs aligning to the loci of interest were visualized using the "anchor" mode in OMView from the OMTools package (Leung *et al.* 2017). Distinct configurations were manually identified from this visualization, and corresponding CMAP files were made for each, including at least 500 kb of unique flanking region where applicable. When SDs contained multiple SVs and were too long to analyze from end to end with single molecules (*e.g.*, over ~350 kb), they were subdivided into groups that were typically anchored in the unique regions either upstream or downstream of the SD (Figures 3B and 4E).

For each group of configurations, the corresponding CMAPs were compiled into a single file and used as input for the OMGenSV pipeline, along with local molecules from each sample and a set of "critical regions" defining the area(s) on each CMAP that molecules need to span in order to support the presence of that configuration in the sample. OMGenSV then performs the following steps for each sample: (1) align local molecules to the configuration CMAPs; (2) filter for molecules that have a single best alignment among the different CMAPs, excluding those that align equally well to two or more CMAPS; (3) filter for molecules whose best alignment spans a pre-defined critical region; (4) report supported configurations for each sample, and (5) identify configurations with weak support in a given sample that would benefit from manual evaluation. The molecule support for these cases was manually evaluated using OMView (Leung *et al.* 2017). Rarely, we found new configurations during the manual evaluation phase that had not been represented in assembled contigs; in these cases, we created CMAPs for the new configurations and included them in a new run of the pipeline.

For the A-CNV at 7q11.23, the "partial" alleles (Supplementary Figure S4A) were too similar to full-copy alleles to disentangle using the standard OMGenSV pipeline, so we modified the protocol described above as follows. First, we ran the pipeline using a set of CMAPs that included the C-A-B duplicon block with 0,1, 2, or 3 full copies of the A-CNV, as well as several open-ended configurations that contained only duplicons C and A. One open-ended configuration had A containing four full A-CNV copies and then terminating, acting as a "sink" for molecules with four or more full copies. The last two open-ended C-A configurations included partial copies of the A-CNV: full, full, partial, full, and full, full, partial, partial, full. After running the pipeline with these CMAPs, any molecules that aligned best to either of the configurations containing partial copies of the A-CNV were filtered out, creating a pool of local molecules that were likely to only contain full copies of the A-CNV. These filtered sets of local molecules were used as input for a new run of the pipeline, using CMAPs containing the full C-A-B duplicon block with 0–8 copies of the A-CNV. This run was processed in the standard way described above and used to generate the results for full copies of the A-CNV (Figure 2D, Supplementary Figure S3). Molecules from the first run that aligned to the configurations with partial A-CNV copies were manually inspected to genotype those samples. Samples containing the "downstream variant" A-CNV alleles (Supplementary Figure S4B) were genotyped by running the pipeline with a CMAP containing two open-ended configurations beginning at the end of A and continuing through B, containing either the canonical end of A or the "downstream variant" end of A. All molecules that aligned best to the "downstream variant" configuration were manually evaluated to genotype that sample.

Breakpoint mapping in microdeletion patients

For each proband, assembled contigs that aligned to the locus of interest were manually inspected to identify the microdeletion configuration. Once identified, the underlying single molecules were examined in order to verify that the microdeletion breakpoint was well supported. In the case of the 16p12.2 microdeletion proband, no assembled contigs contained a microdeletion configuration. Instead, we aligned local single molecules to the hg38 reference and identified those whose alignments supported a microdeletion breakpoint. We found several long single molecules that showed the same consistent breakpoint, one of which is shown in Figure 5D.

Validation using orthogonal data

For each of the three loci studied in this article, we searched the literature for reports of structural variants (SVs) found in samples that were contained in our dataset. To supplement the dataset, we processed optical map data for two additional samples that were commonly used in previous publications: NA12878 and the haploid hydatidiform mole sample CHM1. Optical map data labeled with Nt.BspQI and DLE-1 for NA12878 was downloaded from https://bionanogenomics.com/library/datasets/ and Nt.BspQI-labeled optical map data for CHM1 was previously published (O'Bleness *et al.* 2014; Hastie *et al.* 2017). Optical map data labeled with DLE-1 for CHM1 was generated using the Saphyr system as described above. All data for NA12878 and CHM1 was processed using the OMGenSV pipeline as described above. For each previously published result, we report which of the configurations used in this manuscript would be concordant with the result, and which were actually observed (Supplementary Table S5). For narrowly focused results like targeted sequencing of the Bc inversion at 15q13.3, only the relevant results from our pipeline were reported (in this case, those in group G2).

Data availability

The optical map data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.

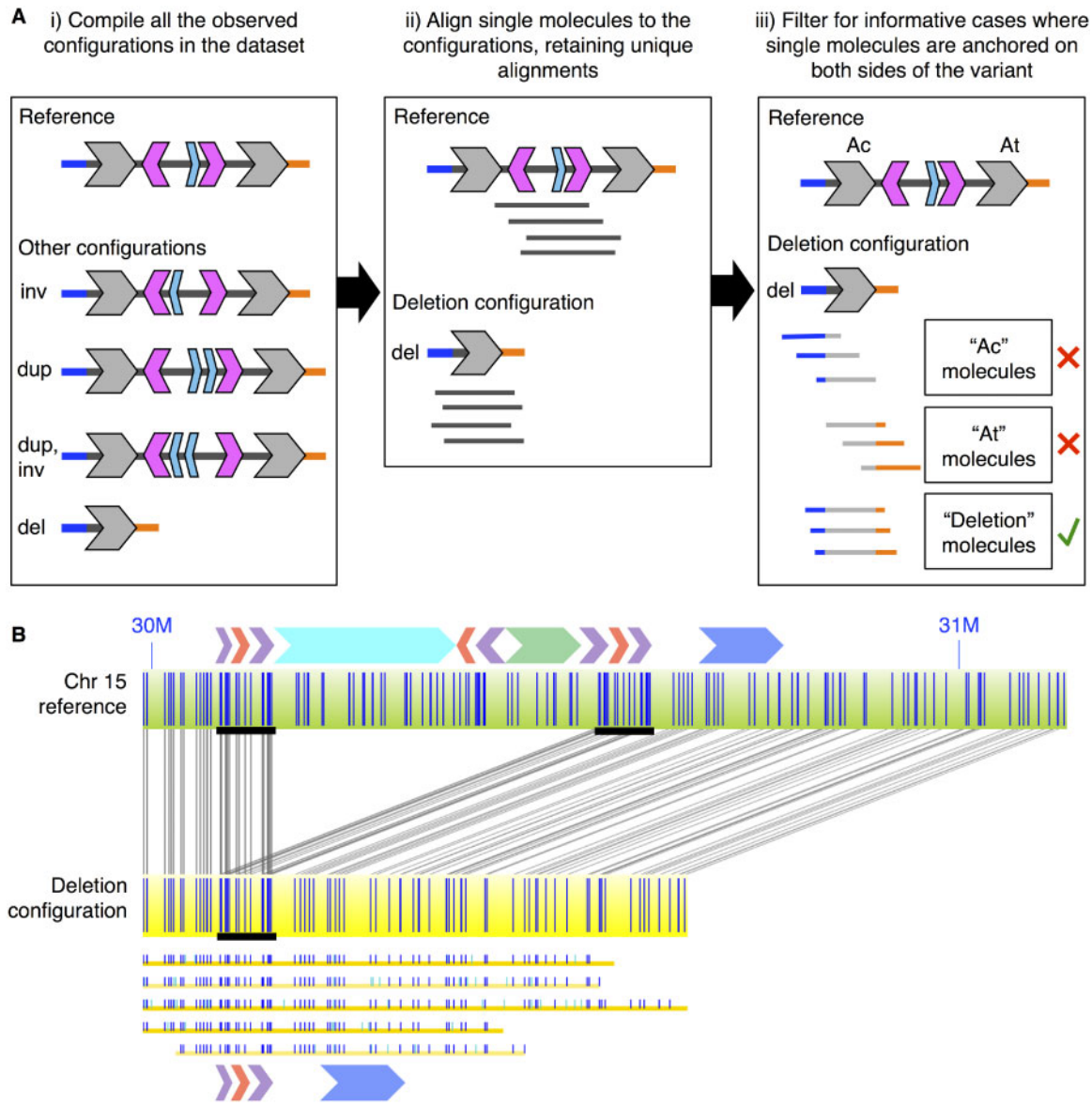


Figure 1 Optical mapping to genotype complex structural variation. (A) Cartoon example of the pipeline (i–iii). (i) Compilation of distinct configurations from all the assembled contigs in the full dataset. The cartoon locus depicted here includes inversions (inv), a duplication (dup), and a deletion (del). (ii) Alignment of single molecules from each sample to the full set of local configurations seen in (i) to determine genotype. The example shown here has single molecule support for the reference and deletion (del) configurations. (iii) Selection of informative molecules anchored in unique region flanking the repeat element. In the example shown here, Ac (A-centromeric) and At (A-telomeric) are the two paralogs of the duplicon marked by the gray arrow. The molecules labeled "Ac" and "At" cannot distinguish between the deletion and other configurations as they lack the full flanking context, so they cannot be used to confirm the deletion. The molecules labeled "Deletion" exclusively support the deletion configuration as they contain flanking region on both sides of the repeat element. (B) A real example showing a deletion configuration at 15q13.3. The top green bar represents the reference configuration, while the middle yellow bar represents the deletion configuration. Vertical tick marks represent label sites, and gray lines connecting ticks show where labels from the deletion configuration aligned to labels from the reference. The SD duplicon structure corresponding to the reference and deletion configurations are shown as colored arrows on the top and bottom of the figure, respectively. The yellow lines below the deletion configuration are single molecules spanning the deletion, with blue and cyan tick marks representing label sites that aligned or did not align to the deletion configuration, respectively. The black horizontal bars indicate the breakpoint region involved in the rearrangement, which molecules needed to span in order to support the deletion configuration.

gov/bioproject/) under accession number PRJNA626024. Supplementary Figure S1 compares genotyping results using single molecules or assembled contigs at 16p12.2. Supplementary Figure S2 shows the number of alleles per sample at the 7q11.23 A-CNV. Supplementary Figure S3 shows the population distribution of full copy numbers at the 7q11.23 A-CNV, while Supplementary Figure S4 shows the structure and prevalence of nonstandard alleles at that site. Supplementary Figure S5 depicts the large-scale haplotypes at 16p12.2 with single-molecule support. Supplementary Table S1 lists duplicon boundaries for each locus. Supplementary Table S2 lists all the SVs described in this article with citations for those that were previously described. Supplementary Table S3 shows supporting molecule counts for each sample for each configuration. Supplementary Table S4 shows population prevalence for each configuration. Supplementary Table S5 details results of orthogonal validation for each locus. Supplementary File S1 describes the methods used to generate Supplementary Figure S1.

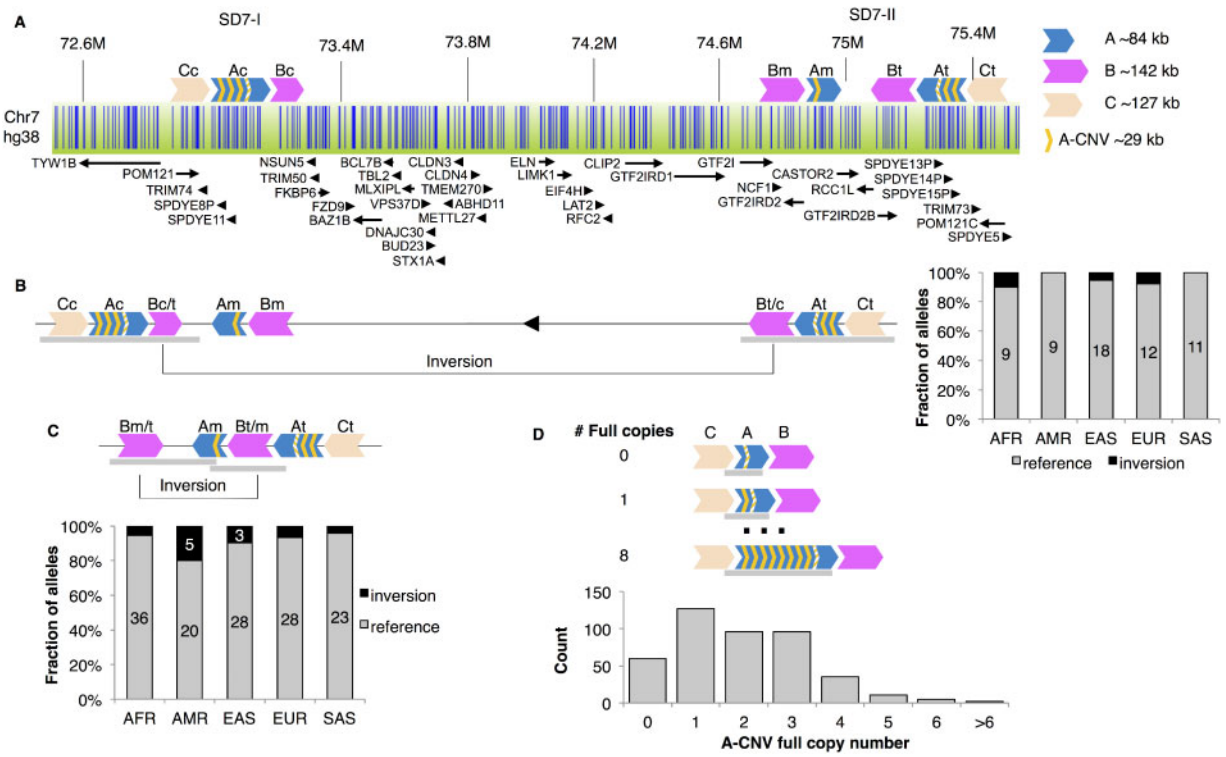


Figure 2 Structural variants (SVs) at 7q11.23. (A) The hg38 reference configuration of 7q11.23, showing duplicon positions and orientations for SD7-I and SD7-II. Paralogs are shown in the same color and are labeled, *e.g.*, "Ac" and "At" for the centromeric and telomeric copies of duplicon A. A partial copy of the A-CNV is marked with parallel lines. Below the duplicons, the optical map of this region is shown as a green bar with BspQI labels shown in blue, followed by local genes. (B) A large inversion observed between SD7-I and SD7-II, with breakpoints within the "C-A-B" duplicon block. Right, a stacked bar graph showing the number of individuals carrying the large inversion allele or the reference configuration in each of the five populations covered in this study. (C) A small inversion observed between Bm and Bt in SD7-II. Bottom, a stacked bar graph showing the number of individuals carrying the small inversion or the reference configuration in each of the five populations. For (B) and (C), labels on the bars show the number of times a configuration was detected in each population; count labels of one or two are not shown. (D) A copy number variant observed in the A duplicon (A-CNV) flanked by the C and B duplicons in both SD7-I and SD7-II. Bottom, a bar plot depicting the full copy number alleles found in the A-CNV, including both the Ac and At copies. No significant population differences were observed for (B), (C), or (D) (B and C: pairwise Fisher's exact test with Benjamini–Hochberg multiple testing correction; D: Wilcoxon rank-sum test with Benjamini–Hochberg multiple testing correction). In all SV diagrams, gray bars below the duplicons represent the critical regions that molecules needed to span in order to be informative for the configuration.

Supplementary material is available at *GENETICS* online.

Results

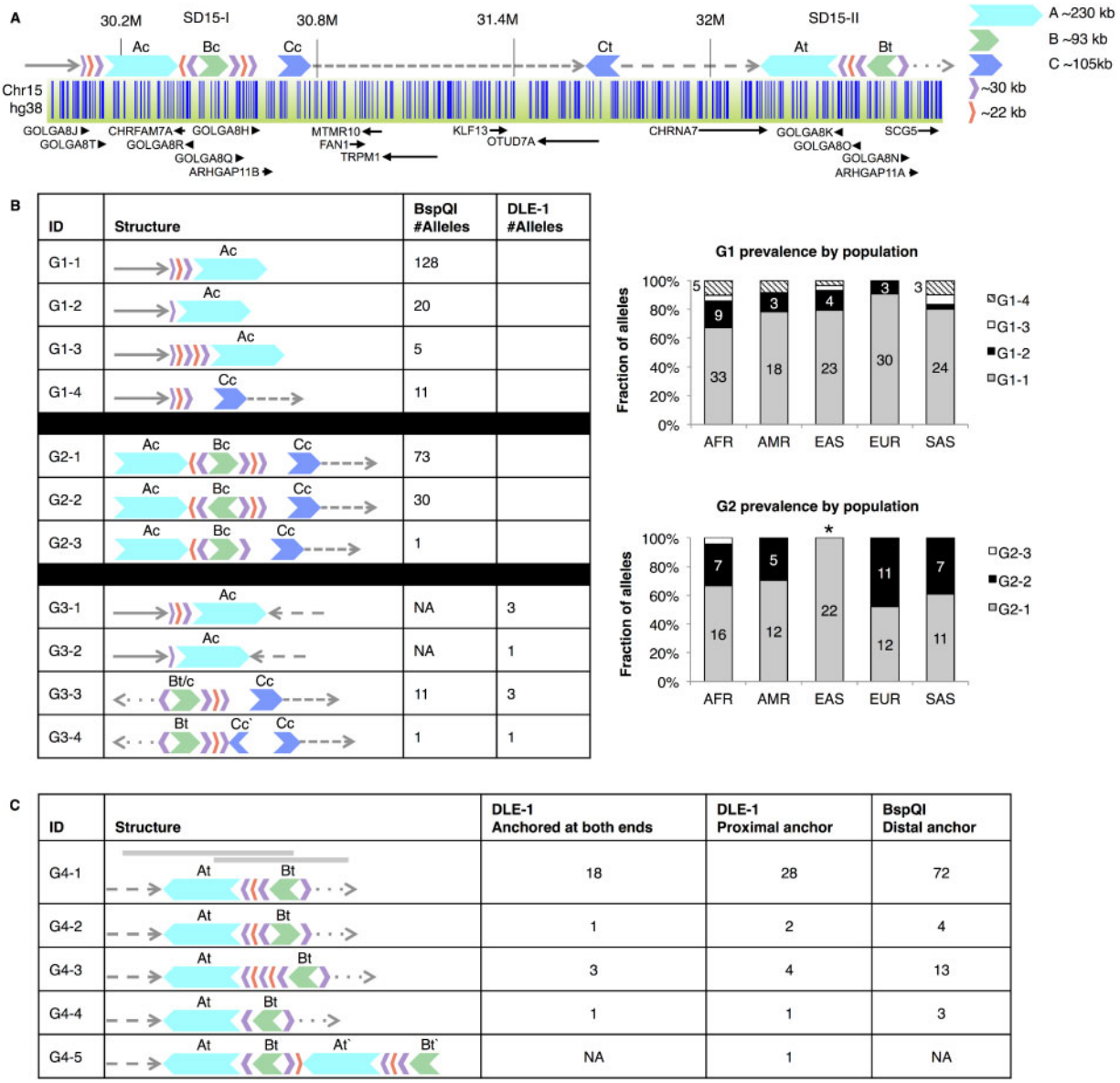
Optical mapping to elucidate the structural complexity within segmental duplications

SDs are hotspots for rearrangements and large-scale structural variations (Emanuel and Shaikh 2001; Stankiewicz and Lupski 2002; Shaw *et al.* 2004). However, due to their complex structures and the high sequence identity shared by paralogous copies, SDs have been difficult to reliably map and sequence. We used optical genome maps to identify the spectrum of structural configurations within and around SDs, and then to genotype those configurations in a high-throughput manner in a large set of samples. We were able to reliably map SD-containing regions at 7q11.23, 15q13.3, and 16p12.2—all regions that are involved in recurrent structural variations associated with genomic disorders (Peoples *et al.* 2000; Sharp *et al.* 2008; Girirajan *et al.* 2010). We characterized these regions by optical mapping of 154 phenotypically normal individuals from 26 different populations comprising five super-populations: African (AFR), American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS) (Levy-Sakin *et al.* 2019).

De novo assembly of genome maps at these SD regions often resulted in assembly errors at paralogs with long stretches of identical label patterns (Supplementary File S1, Figure S1). Thus, rather than using assembled contigs, our approach to genotyping SDs relied on identifying single molecules that mapped to a given SD configuration with labels anchored in unique regions on one or both sides (Figure 1A). First, for each locus of interest, we compiled a list of potential configurations by examining all assembled local contigs (Figure 1A, i). Then, to evaluate the accuracy and prevalence of these configurations in our dataset, we aligned local single molecules from each sample to the set of potential configurations at that locus, filtering to retain molecules that aligned uniquely to one of the configurations (Figure 1A, ii) and whose alignment spanned both the identical region as well as its flanking context (Figure 1A, iii). The resulting set of aligned single molecules revealed the configuration(s) present in each sample (see Methods for additional details).

Structural variation at 7q11.23

7q11.23 contains two SDs, designated SD7-I and SD7-II (Supplementary Table S1), flanking a 1.3 Mb gene containing region that is deleted in patients with WBS (Bayés *et al.* 2003). Known SVs within 7q11.23 include an inversion between SD7-I and -II, which was associated with a predisposition to the pathogenic deletion (Osborne *et al.* 2001), as well as deletions and duplications within the SDs (Cuscó *et al.* 2008; Kidd *et al.* 2008; Perry *et al.* 2008; Conrad *et al.* 2010; Sudmant *et al.* 2013).

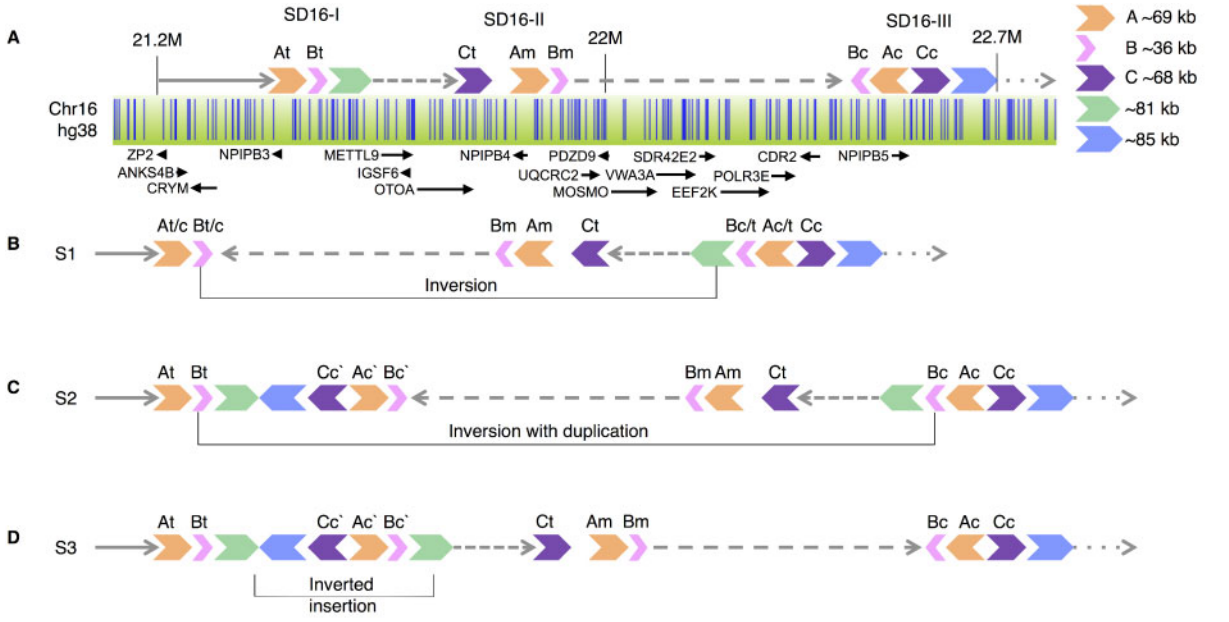


ID	Structure	BspQI #Alleles	DLE-1 #Alleles
G1-1	Ac	128	
G1-2	Ac	20	
G1-3	Ac	5	
G1-4	Cc	11	
G2-1	Ac Bc Cc	73	
G2-2	Ac Bc Cc	30	
G2-3	Ac Bc Cc	1	
G3-1	Ac	NA	3
G3-2	Ac	NA	1
G3-3	Bt/c Cc	11	3
G3-4	Bt Cc` Cc	1	1

ID	Structure	DLE-1 Anchored at both ends	DLE-1 Proximal anchor	BspQI Distal anchor
G4-1	At Bt	18	28	72
G4-2	At Bt	1	2	4
G4-3	At Bt	3	4	13
G4-4	At Bt	1	1	3
G4-5	At Bt At` Bt`	NA	1	NA

Figure 3 SVs at 15q13.3. (A) The hg38 reference configuration of 15q13.3, showing duplicon positions and orientations for SD15-I and SD15-II. Paralogs are shown in the same color and are labeled, *e.g.*, "Ac" and "At" for the centromeric and telomeric copies of duplicon A. Gray arrows with different patterns mark the different unique regions flanking the SDs. Below the duplicons, the optical map of this region is shown as a green bar with BspQI labels shown in blue, followed by local genes. (B) Configurations anchored in the unique region either proximal or distal to SD15-I. Configurations were genotyped in three groups, G1, G2, and G3, using datasets labeled with the DLE-1 or the BspQI enzyme. For each genotyped sample, supporting molecules needed to span all of the duplicons and flanking unique regions depicted in the "structure" column. Right, stacked bar graphs showing the prevalence of configurations in the G1 (top) and G2 (bottom) groups for each of the five populations used in this study. Configuration G2-2 was significantly depleted in the EAS population compared to all other populations ($P < 0.05$, pairwise Fisher's exact test with Benjamini–Hochberg multiple testing correction comparing G2-1 and G2-2). Labels on the bars show the number of times a configuration was detected in each population. Count labels of one or two are not shown. (C) Configurations anchored in the unique region either proximal or distal to SD15-II. The DLE-1 dataset contained molecules anchored within unique regions on both ends of SD15-II as well as molecules anchored only in the unique region proximal to SD15-II, while the BspQI dataset contained molecules anchored only in the unique region distal to SD15-II. Configurations were genotyped in one group, G4. Gray bars (top) indicate the proximal and distal critical regions: proximally anchored molecules extended at least to Bt, while distally anchored molecules extended at least to At. For (B) and (C), columns show the configuration IDs, their structure, and the number of alleles identified in our dataset with the indicated enzyme.

We examined all assembled contigs from the 7q11.23 locus in our dataset of 154 diverse individuals, which revealed three major SVs in this region that were previously reported (Supplementary Table S2), including the large inversion between the two SD7 regions (Figure 2B) (Osborne *et al.* 2001), an inversion inside SD7-II (Figure 2C) (Kidd *et al.* 2008), and a CNV within the A-centromeric (Ac) and A-telomeric (At) duplicons in SD7-I and SD7-II, respectively (Figure 2D) (Perry *et al.* 2008; Conrad *et al.* 2010; Sudmant *et al.* 2013), here referred to as the A-CNV. The large inversion had breakpoints within the ~400 kb C-A-B duplicon block that was present in both SD7s (Figure 2B). We found the inversion in 5% of observed configurations in the full dataset (3/62) (Supplementary Table S3). We also observed the ~200 kb inversion of the A-middle (Am) duplicon within SD7-II



E

ID	Matching full-length haplotypes	Structure	#Alleles
G1-1	S2,S3	At Bt	85
G1-2	S1	At/c Bt/c	33
G1-3	hg38	At Bt	4
G2-1	S2	Cc` Ac` Bc`	62
G2-2	S3	Cc` Ac` Bc`	4
G2-3	S2 variant	Cc` Ac` Bc`	2
G3-1	S1,S2	Bc Ac Cc	78
G3-2	S1, S2 variant	Ac Cc	13
G3-3	hg38, S3	Bc Ac Cc	2

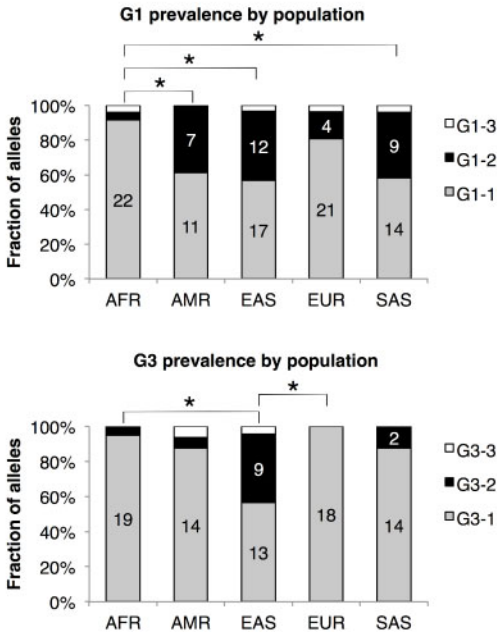


Figure 4 SVs at 16p12.2. (A) The hg38 reference configuration of 16p12.2, showing duplicon positions and orientations for SD16-I, SD16-II, and SD16-III. Paralogs are shown in the same color, and are labeled, *e.g.*,. "At," "Am," and "Ac" for the telomeric, middle, and centromeric copies of duplicon A. Gray arrows with different patterns mark the different unique regions flanking the SDs. Below the duplicons, the optical map of this region is shown as a green bar with BspQI labels shown in blue, followed by local genes. (B) A large balanced inversion, "S1," between At and Ac. (C) A large inversion with duplication, "S2." Newly created duplicons are marked as, *e.g.*, Cc′. (D) A small inverted insertion detected distal to SD16-A on the reference configuration, labeled "S3." (E) Left, configurations genotyped in three groups, G1, G2, and G3, are shown. G1 configurations are anchored in the unique region proximal to SD16-I. G2 configurations are anchored in the green-blue duplicon pair that was not seen in the reference configuration. G3 configurations are anchored in the unique region distal to SD16-III. Columns depict the configuration ID, the longer haplotypes with which they are consistent (among the hg38 reference, S1, S2, and S3), the structure, and the number of alleles detected in the BspQI dataset. Supporting molecules for a given configuration had to span each of the depicted duplicons. Right, stacked bar graphs showing the prevalence of configurations from group G1 (top) and G3 (bottom) across the five populations included in this study. *$P < 0.05$, pairwise Fisher's exact test with Benjamini–Hochberg multiple testing correction using the two most prevalent configurations in the group. Labels on the bars show the number of times a configuration was detected in each population. Count labels of 1 were not shown.

(Figure 2C), with an overall prevalence of 9% and a range of 4–20% within different populations (Supplementary Table S4).

The A-CNV in duplicons Ac and At consisted of 0–8 full copies of a ~29-kb region, in addition to a consistent partial copy (Figure 2D). In reference assembly GRCh38 (hg38), there were three full copies of this region in duplicon Ac and two full copies in At. We assessed the Ac and At CNVs together since they were embedded within the large C-A-B block in both SD7s and both paralogs were therefore flanked by long stretches of identical label patterns. The prevalence of each full copy number variant is presented in Figure 2D: one copy was the most common, while copy numbers of >6 were rarely seen (Supplementary Table S3).

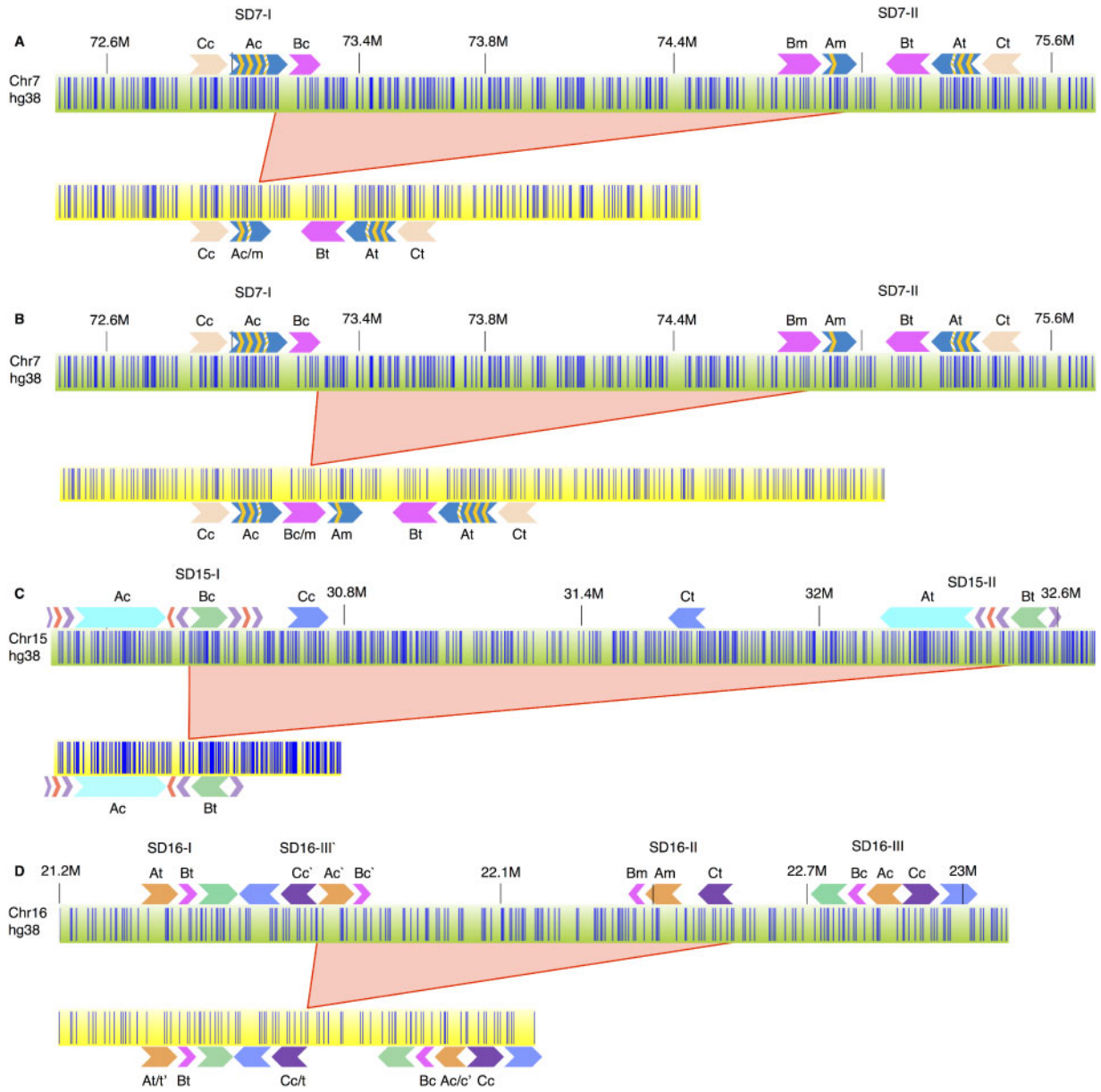


Figure 5 Breakpoint mapping in microdeletion patients. For each panel, the green bar (top) is the hg38 reference configuration of a given region, while the yellow bar (bottom) is the configuration of the deletion observed in the patient. Yellow bars in (A–C) depict assembled contigs while the yellow bar in (D) depicts a molecule. Red filled-in triangles indicate the region deleted in patients. Duplicon structures for the reference and deleted configurations are depicted above and below the bars, respectively. (A) Patient 1 with 7q11.23 deletion. (B) Patient 2 with 7q11.23 deletion. (C) Patient with 15q13.3 deletion. (D) Patient with 16p12.2 deletion.

Particularly long configurations can be difficult to detect if they require molecules to span regions substantially longer than the average molecule length (~250 kb), but the A-CNV copy numbers below 6× were unlikely to be adversely affected since they were smaller than the average molecule length (ranging from 94 to 235 kb for copy numbers 0–5). The A-CNV was highly variable both within and between individuals; 51% of samples had three different alleles at this CNV, while 25% had two alleles and 24% had four alleles, the latter group representing cases where Ac and At were both heterozygous for the A-CNV and each of the four alleles were distinct (Supplementary Figure S2). Remarkably, none of the samples we analyzed had fewer than two distinct alleles. We did not detect any significant population-based differences in full copy number (Wilcoxon rank-sum test with Benjamini–Hochberg multiple testing correction, Supplementary Figure S3).

We also observed two other classes of configurations at the A-CNV locus (Supplementary Figure S4). In one class, full copies of the CNV were interspersed with one or more partial copies; the most common such configuration is shown in Figure S4A. Such partial copies were seen in 19 alleles from 17 individuals (Supplementary Table S3), all of African descent (Supplementary Figure S4C, Table S4), a significant enrichment over the other populations ($P < 0.005$ in each case, pairwise Fisher's exact tests with Benjamini–Hochberg multiple testing correction). The other variant class involved a different label pattern immediately downstream of the CNV, shown in its most common configuration in Supplementary Figure S4B; this variant was present in

nine samples of African, East Asian, and South Asian descent (Supplementary Figure S4C, Tables S3, S4) and was accompanied in different samples by 0–4 copies of the CNV. The number of distinct A-CNV alleles per sample (Supplementary Figure S2) includes these variant alleles as well as the full copy number variants.

Structural variation at 15q13.3

We next analyzed a region within 15q13.3 consisting of two SDs, designated SD15-I and SD15-II (Supplementary Table S1), separated by a ∼1.3 Mb unique region that is deleted in patients with the 15q13.3 microdeletion syndrome (Figure 3A). This region is located in a highly unstable area of chromosome 15 that also includes the breakpoints for deletions associated with Angelman and Prader–Willi Syndromes (Amos-Landgraf *et al.* 1999; Christian *et al.* 1999; Khan *et al.* 2011).

Within our dataset of 154 diverse individuals, we detected a large number of SVs at this locus, several of which, to our knowledge, have not been previously reported (Supplementary Table S2). We analyzed the SVs in groups based on the regions they shared in common, *i.e.*, configurations anchored in the unique regions upstream or downstream of SD15-I and SD15-II (Figure 3, B and C). A subset of configurations represented inversions between the two SD15s, and their analysis required that supporting molecules be anchored in both upstream and downstream unique regions (Figure 3B, group G3). One complicating factor in the analysis of this locus was the presence of a fragile site within the A duplicon for nicking enzyme Nt.BspQI; these fragile sites occur when two single-strand nicking sites occur close together on opposite strands, causing breakage of nicked DNA molecules so that few to no molecules are able to traverse the site. To overcome this limitation, we supplemented our dataset of 154 samples labeled using the Nt.BspQI nickase enzyme with a dataset containing 52 of those same samples labeled using the newer DLE-1 enzyme (Wong *et al.* 2020). DLE-1 deposits an epigenetic fluorescent label rather than nicking the DNA, and therefore does not create any fragile sites (Maggiolini *et al.* 2019). The 52 samples labeled with DLE-1 were selected from all 26 of the subpopulations in the original 154-sample dataset, with each subpopulation represented by two samples.

From our analysis of SD15-I, we detected four different configurations anchored in the proximal unique region (Figure 3B, G1), not including the inversions between SD15-I and SD15-II. The reference configuration (G1-1) was the most common (124/160 or 78% of observed configurations; Supplementary Table S3). Two additional configurations involved a contraction [G1-2 (Conrad *et al.* 2010), with a prevalence of 13%] or expansion [G1-3 (Pang *et al.* 2010), prevalence of 3%] of the purple-red-purple duplicon triplet. The purple duplicons–which contain the *GOLGA8* genes–have previously been identified as key to the structural instability of this locus (Antonacci *et al.* 2014). An additional configuration in the proximal region of SD15-I involved a deletion of both the A and B duplicons [G1-4 (Cooper *et al.* 2011; Coe *et al.* 2014)] and was seen in 7% of observed configurations.

We also detected three configurations anchored in the distal unique region of SD15-I (Figure 3B, G2). The reference configuration (G2-1) was again the most common (Supplementary Table S3), representing 60% of observed configurations, while 25% of configurations had an inversion of the Bc duplicon (G2-2) (Antonacci *et al.* 2014). This inversion had strikingly different prevalence among populations, as it was seen most often in European samples (48%) and never in any East Asian samples (Supplementary Table S4). The East Asian population was significantly depleted in this configuration with respect to every other population ($P < 0.05$, pairwise Fisher's exact tests with Benjamini–Hochberg multiple testing correction). The third configuration at this locus involved a contraction of the purple-red-purple duplicons (G2-3, Alsmadi *et al.* 2014) and was only seen once in our dataset.

Inversions between SD15-I and SD15-II (Antonacci *et al.* 2014) were detected by looking for molecules that aligned to unique regions either proximal to both regions or distal to both regions (Figure 3B, G3). In the smaller DLE-1 dataset, which was used to traverse the BspQI fragile site within duplicon A for configurations G3-1 and G3-2, we detected inversions eight times (Supplementary Table S3). Configurations G3-3 and G3-4 were able to be assayed using the larger BspQI dataset, which showed evidence of inversion in 9% of observed configurations.

Configurations at SD15-II were assayed at the proximal end using DLE-1 to span duplicon At, while configurations at the distal end could be assayed using the full BspQI dataset (Figure 3C). In some cases, molecules from the DLE-1 dataset were also able to span the full configurations from end to end. As in SD15-I, the reference configuration (G4-1) was the most prevalent (78% of observed configurations; Supplementary Table S3). Other configurations included an inversion of Bt (G4-2), analogous to but less frequent than the inversion of Bc in SD15-I (G2-2), and expansions and contractions of the purple-red-purple duplicon triplet [G4-3 (Pang *et al.* 2010), G4-4 (Perry *et al.* 2008; Conrad *et al.* 2010)]. One chromosome harbored two tandem copies of At and Bt (G4-5) (Antonacci *et al.* 2014), which was reconstructed from several tiled molecules from that sample because no single molecule spanned the full length of the configuration.

Structural variation at 16p12.2

We analyzed a region of 16p12.2 consisting of three SDs, designated SD16-I, SD16-II, and SD16-III (Supplementary Table S1, Figure 4A). The region between SD16-II and SD16-III is deleted in patients with 16p12.2 microdeletion syndrome (Ballif *et al.* 2007; Antonacci *et al.* 2010; Girirajan *et al.* 2010). Among the dataset of 154 diverse individuals, in addition to detecting haplotypes "S1" and "S2," reported previously (Tuzun *et al.* 2005; Antonacci *et al.* 2010) (Figure 4, B and C), we also detected novel SVs and configurations in this region (Supplementary Table S2). Most notably, we detected a novel inversion, "S3," that combined the proximal part of S2, up through Bc′, with the distal part of the reference assembly (Figure 4D). These long haplotypes were supported by tiling long single molecules (Supplementary Figure S5).

To genotype the variants in the full dataset, we grouped configurations based on shared location, as in 15q13.3 (Figure 4E). Group 1 consisted of three configurations anchored in the unique region upstream of SD16-I (Figure 4E, G1). G1-1 was the most common configuration, seen in 85/122 of observed configurations (70%; Supplementary Table S3). Between the two dominant configurations (G1-1 and G1-2), there was a significant difference in prevalence between the African population and the American, East Asian, and South Asian populations, with G1-1 comprising 92% of African G1 alleles but only 57–61% of G1 alleles from the other three groups ($P < 0.05$, Fisher's exact test; Supplementary Table S4). Notably, a previous study (Antonacci *et al.* 2010) of 24 chromosomes found no support for the reference genome configuration at this locus. Our optical mapping results of 154 individuals showed that the hg38 reference assembly structure at this locus (G1-3) does exist in the general population, albeit at low frequency (3.2%, Figure 4E, Supplementary Figure S5A).

Group 2 consisted of three configurations, each corresponding to the extension of SD16-I seen in several of the rearranged haplotypes (Figure 4, C and D) but not in the reference. Of these, G2-1 was the most common configuration, found in 62/68 alleles (91%; Supplementary Table S3). Group 3 consisted of four configurations that were all anchored in the unique region downstream of SD16-III (Figure 4E). The most common configuration was G3-1, with 78/93 G3 alleles (84%). Among the two most common G3 configurations, G3-2 was significantly enriched among the East Asian population compared to African and European, comprising 39% of East Asian G3 alleles but only 5% and 0% of African and European G3 alleles, respectively ($P < 0.05$, Fisher's exact test; Supplementary Table S4).

Breakpoint mapping in microdeletion patients

We generated optical maps in four patients with microdeletions at the three loci analyzed in this study, with the goal of localizing the microdeletion breakpoints. We analyzed two patients with WBS, caused by microdeletions in 7q11.23, and reconstructed the structure of the deletion alleles in both. We found two different microdeletion configurations of 7q11.23 in the patients (Figure 5, A and B), consistent with previous reports showing variation in the breakpoints of 7q11.23 microdeletions (Bayés *et al.* 2003; Merla *et al.* 2010). In the first patient, the breakpoints were localized to Ac and Am, ~84 kb paralogous modules within the larger ~420 kb and ~780 kb SD7-I and SD7-II, respectively. In the second patient, the breakpoints were localized to the ~142 kb paralogous modules Bc and Bm in SD7-I and SD7-II, respectively.

Similarly, we reconstructed the deletion allele for one patient with the 15q13.3 microdeletion and one with the 16p12.2 microdeletion. In the 15q13.3 microdeletion, we localized the deletion breakpoints to the ~50 kb red and purple modules between Ac and Bc in the larger SD15-I (~700 kb) on one side and between At and Bt in SD15-II (1 Mb) on the other side (Figure 5C), consistent with a previous report (Antonacci *et al.* 2014). In the 16p12.2 microdeletion patient, we found that the microdeletion breakpoints were localized to the ~70 kb C duplicons in SD16-II and SD16-III (Figure 5D), which were previously postulated to mediate the deletion (Antonacci *et al.* 2010). Thus, optical mapping allowed us to reconstruct deletion haplotypes in patients with microdeletion syndromes and localize the microdeletion breakpoints to specific duplicons within the larger, complex SDs.

Validation of the pipeline using orthogonal data

To evaluate the performance of the OMGenSV genotyping pipeline at these three loci, we compared our results to previously published data for the same samples that were obtained with a variety of orthogonal approaches (Antonacci *et al.* 2010, 2014; Dennis *et al.* 2017), including BAC clone sequencing, FISH, and targeted sequencing (Supplementary Table S5). In addition to the datasets used throughout, we included data from two samples that were frequently used in previous reports: NA12878–which was not included in the main dataset because both parents, NA12891 and NA12892, were included–and the haploid hydatidiform mole sample CHM1. Both additional samples were labeled with Nt.BspQI and DLE-1 (NA12878 data from https://bionanogenomics.com/library/datasets/).

Overall, our results were highly concordant with previous genotype data (Supplementary Table S5). At the 15q13 locus, we compared results for 34 samples that were common to both this and a previous study (Antonacci *et al.* 2014), including three that were sequenced from BAC clones, and found that all SV genotypes were concordant with previous data with the exception of one sample, HG00101. This sample was previously found not to contain the Bc inversion via targeted sequencing of the associated haplotype, but we found it to be heterozygous for the inversion. The only other unexpected result at this locus was our finding that CHM1 contained the G4-3 configuration, which was not previously reported (Antonacci *et al.* 2014). For 16p12.2, four samples common to this study and (Antonacci *et al.* 2010) were evaluated, and our results were found to be concordant with the previous data. The 7q11.23 locus had BAC clone sequencing data for one sample, CHM1 (Dennis *et al.* 2017), which was consistent with our OMGenSV results from this sample (Supplementary Table S5).

Discussion

SDs are complex genomic structures containing long paralogs with extremely high sequence similarity, making them an excellent substrate for NAHR (Stankiewicz and Lupski 2010). This process results in chromosomal rearrangements including microdeletions, microduplications, and inversions (Osborne *et al.* 2001; Cuscó *et al.* 2008; Hobart *et al.* 2010; Merla *et al.* 2010; Antonacci *et al.* 2014). Despite the importance of resolving the internal structure of SDs to elucidate the underlying cause(s) of associated chromosomal rearrangements, little is known about how these structures vary between individuals. Existing techniques are typically either unable to accurately assemble these regions due to short read lengths–with some repeat units extending for hundreds of kb, even "long-read" technologies such as PacBio or Oxford Nanopore are insufficient to reconstruct many SDs (Vollger *et al.* 2019)—or are too cost—and/or labor intensive to be applied in a high-throughput study (*e.g.*, "long-read" sequencing of individual BAC clones (Huddleston *et al.* 2014) or fiber FISH analysis [Molina *et al.* 2012]). Bionano optical mapping overcomes both of the above constraints by obtaining long molecule lengths with a fast and (at ~$600 per sample) cost-effective methodology.

SDs within genomic regions 7q11.2, 15q13.3, and 16p12.2 have been implicated in chromosomal rearrangements associated with microdeletion and microduplication syndromes. Furthermore, evolutionary analysis of these regions in nonhuman primates have suggested a recent origin of the SDs in hominoids followed by a rapid expansion in humans (Antonell *et al.* 2005, Antonacci *et al.* 2010, 2014). These observations have led to the suggestion that these and other SD-containing regions are likely to contain variations within the human population, prompting the analysis of structural variation within these regions in humans. Previous studies carried out in a limited number of individuals have uncovered structural variation within these SD-containing loci (Osborne *et al.* 2001; Cuscó *et al.* 2008; Sharp *et al.* 2008; Antonacci *et al.* 2010, 2014). Thus, these regions were selected for a population-level analysis of structural variation using Bionano optical mapping.

Because automated *de novo* assembly of optical maps is error-prone at loci with long repetitive regions, we used OMGenSV (Demaerel *et al.* 2019), a pipeline to rapidly genotype SVs in a high-throughput manner using unassembled single molecules. This approach was previously used to discover an unprecedented level of structural variation within the SDs associated with the 22q11.2 deletion syndrome (Demaerel *et al.* 2019). By applying this approach to a diverse dataset of 154 phenotypically normal individuals, we identified dozens of SVs—some novel and some previously reported—at 7q11.23, 15q13.3, and 16p12.2 and created catalogs of their local structural configurations, several of

which had frequencies that varied by population. We also mapped the microdeletion breakpoints in patients with each of these genomic disorders. The resolution of the optical maps allowed us to narrow down SV breakpoints to individual duplicons inside SDs, which may help facilitate further breakpoint localization at single nucleotide resolution.

Each of the loci in the study was found to contain at least one SV with significantly different frequency between populations (Figures 3B and 4E, Supplementary Figure S4). Importantly, two of these SVs affected the length of directly oriented paralogous duplicons between SDs, which could serve as a template for NAHR to lead to pathogenic microdeletions and microduplications. At the 15q13.3 locus, an inversion of the Bc module (Antonacci *et al.* 2014) had an overall prevalence of 29% (Figure 3B, G2-2); it represented 48% of European alleles but was entirely absent in East Asian samples (Supplementary Table S4), similar to the population stratification of this SV observed previously (Antonacci *et al.* 2014). This SV represents a substantial increase in the length of directly oriented duplicons between the two blocks of SD and may therefore be predisposing to the pathogenic microdeletion/duplication (Antonacci *et al.* 2014), suggesting that East Asian populations would be expected to have low levels of 15q13.3 microdeletion/microduplication syndrome, while European populations may have the highest levels. However, Antonacci *et al.* (2014) suggested that the Bc inversion haplotype is not enriched among microdeletion patients and that the microdeletion mechanism may therefore not be strongly reliant on length of homology. A comprehensive analysis of SD configurations in parental samples may be required to evaluate the true impact of this SV on the microdeletion mechanism.

Similarly, at the 16p12.2 locus, configuration G1-2 (Figure 4E) distinguished the S1 variant (Figure 4B) from either the reference (Figure 4A) or the S2 and S3 configurations (Figure 4, C and D). Notably, the S1 configuration lacks directly oriented paralogs flanking the region that, on the reference, lies between SD16-II and SD16-III, which is deleted in the pathogenic microdeletion, while the other configurations have directly oriented C duplicons flanking the region, suggesting that the S1 configuration may be protective against the pathogenic microdeletion (Antonacci *et al.* 2010). Accordingly, mapping of a 16p12.2 microdeletion patient sample showed a configuration consistent with a transmitting chromosome containing haplotype S2, with deletion breakpoints inside the C duplicons (Figure 5D). The G1-2 configuration, representing the putatively protective S1 haplotype, varied in prevalence from 4% in Africans to 38–40% in East Asians, Americans, and South Asians (Supplementary Table S4), consistent with a previous study (Antonacci *et al.* 2010), suggesting that the latter three populations may be expected to have a lower prevalence of the pathogenic microdeletion. The third SV that varied in prevalence between populations was the partial A-CNV configuration at 7q11.23 (Supplementary Figure S4), which was seen exclusively in African samples and would likely not affect predisposition toward the pathogenic microdeletion/microduplication.

SVs within SDs that change the copy number of genes may have phenotypic consequences. For instance, at 16p12.2, several configurations include a duplication of the nuclear pore complex-interacting protein *NPIPB5*, which has three local paralogous copies (*NPIPB3*, *NPIPB4*, and *NPIPB5*) in the "S1" and hg38 structures, and four copies in the "S2" and "S3" structures that are enriched in the African population (Figure 4). These genes may variously be associated with renal cell carcinoma (Wang *et al.* 2019) and immune system response to SARS-CoV infection (Huang *et al.* 2017). Another example involves the group of Speedy family E genes and pseudogenes (*SPDYE8P*, *SPDYE11*, *SPDYE13P*, *SPDYE14P*, *SPDYE15P*) encoded in both copies of the A-CNV at 7q11.23. The Speedy genes are cell cycle regulators, and family E has greatly expanded copy number in the human lineage (Wang *et al.* 2018). *SPDYE11* has been implicated in modulating calcium flux through α7 nicotinic acetylcholine receptors in the brain (Rex *et al.* 2017). Further studies are needed to understand the impact of these copy number variations of genes within SDs on phenotypic outcome.

While optical mapping of SD loci is a major advance, it is limited by the physical lengths of the molecules. Even with mean molecule N50 of 262 kb (range 199–395 kb), we are unable to genotype SVs containing repetitive elements where the length of the repeat exceeds the lengths of the molecules. A clear example of how the distribution of molecule lengths affects different SVs can be seen at the 7q locus. Genotyping the large inversion compared to the reference configuration (Figure 2B) required finding molecules that spanned the entire length of the C-A-B duplicons with several flanking labels on both sides, *i.e.*, molecules that were at least ∼410 kb long. Molecules that spanned this range were only detected in 62/154 samples. In contrast, the CNV inside the A duplicons (Figure 2D) required molecules that ranged from only 94 to 264 kb, depending on the allele, and consequently we were able to genotype at least two alleles in all 154 samples, with the vast majority of samples having at least three alleles between the two paralogs (Supplementary Figure S2). Generally speaking, configurations that required molecules from the tail end of the length distribution for genotyping were less likely to be genotyped in any given sample. Future methodological refinements to increase molecule lengths will facilitate the genotyping of SDs with increasingly longer paralogous regions.

The molecular length is affected by the physical handling of long DNA molecules and experimental protocol. Our study was performed with the original labeling protocol based on a single-strand nicking enzyme that produces shorter molecules when two single-strand nicks are within several hundred basepairs on opposite strands, leading to fragile sites that break easily. The fragile sites not only shorten the molecular length, but also create breakpoints in the genome where no molecules can span across. This breakpoint problem is resolved with the use of a new labeling enzyme, DLE-1, which creates no fragile sites because it labels recognition sites without creating single-strand nicks (Maggiolini *et al.* 2019).

Using single molecules to genotype SVs in the framework of the OMGenSV pipeline leads to high precision and sensitivity. We compared our results to previously published data obtained with orthogonal techniques, which was available for 35 samples that were also present in our dataset. Only two cases were potentially discordant. In the first case, a negative result obtained by targeted sequencing of a SNP-tagged haplotype (Antonacci *et al.* 2014) was positive for the associated configuration in our dataset. Since SD-containing loci are known to be unstable and prone to rearrangement, it is likely that many SVs have arisen *de novo* multiple times in the population (Lupski 2007) rather than being founding mutations, and so targeting a particular haplotype using SNPs may miss other haplotypes containing the same configuration. The other discordant case involved our finding of configuration G4-3 at the 15q13.3 locus in a sample that was sequenced from BAC clones where this configuration was not reported (Antonacci *et al.* 2014). However, a close look at the BAC sequencing data shows that the relevant area was not fully assembled, explaining how this configuration could have been

missed. Overall, genotype results from our approach are highly concordant with previous data.

This study demonstrates a high-throughput, cost-effective approach for characterizing SD-containing regions that involves using ultra-long optical maps to span the paralogous duplicons and capture the vast structural variability of these regions. Given our finding that SV patterns in the three SD-containing loci studied are highly variable and differ significantly between populations, including some that are likely to affect predisposition to pathogenic microdeletions/duplications, catalogs of SV patterns in these loci are most helpful in analyses of these regions in population and patient studies. Applying our high-throughput, cost-effective approach to additional complex loci throughout the genome will lead to catalogs of SV patterns that represent the genetic diversity of these regions and may reveal patterns that are prone to rearrangements that lead to genomic disorders. Furthermore, the tools and methods developed here will enable us to more accurately genotype SD configurations in clinical samples, which may consequently improve our ability to predict the risk for the occurrence of SD-mediated rearrangements associated with genomic disorders.

Acknowledgments

T.H.S. and P-Y.K. conceived the study and analysis using Bionano optical mapping. Y.M. and F.Y. performed the analysis and interpretation. C.R.C., N.J.L.M., and K.C.C. obtained informed consent from parents and patients during hospital follow-up and facilitated collection of blood samples. C.R.C. coordinated study and subject enrollment. E.A.G., S.K.C., C.C., and C.L. carried out sample processing and experimentation. U.S. provided data for CHM1. Y.M., F.Y., T.H.S., and P-Y.K. drafted and critically revised the article. Final approval of the version to be published was given by Y.M., F.Y., S.K.C., C.C., C.L., E.A.G., N.J.L.M., K.C.C., C.R.C., P.-Y.K., and T.H.S.

Funding

This work was made possible by a grant, GM120772 to T.H.S and P-Y.K., from the National Institute of General Medical Sciences of the National Institutes of Health (NIH).

Conflicts of interest: None declared.

Literature cited

Alsmadi O, John SE, Thareja G, Hebbar P, Antony D, *et al.* 2014. Genome at juncture of early human migration: a systematic analysis of two whole genomes and thirteen exomes from Kuwaiti population subgroup of inferred Saudi Arabian tribe ancestry. PLoS One. **9**:e99069.

Amos-Landgraf JM, Ji Y, Gottlieb W, Depinet T, Wandstrat AE, *et al.* 1999. Chromosome breakage in the Prader-Willi and Angelman syndromes involves recombination between large, transcribed repeats at proximal and distal breakpoints. Am J Hum Genet. **65**: 370–386.

Antonacci F, Dennis MY, Huddleston J, Sudmant PH, Steinberg KM, *et al.* 2014. Palindromic *GOLGA8* core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. Nat Genet. **46**:1293–1302.

Antonacci F, Kidd JM, Marques-Bonet T, Teague B, Ventura M, *et al.* 2010. A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. Nat Genet. **42**:745–750.

Antonell A, de Luis O, Domingo-Roura X, Perez-Jurado LA. 2005. Evolutionary mechanisms shaping the genomic structure of the Williams-Beuren syndrome chromosomal region at human 7q11.23. Genome Res. **15**:1179–1188.

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001. Segmental duplications: organization and impact within the current human genome project assembly. Genome Res. **11**:1005–1017.

Ballif BC, Hornor SA, Jenkins E, Madan-Khetarpal S, Surti U, *et al.* 2007. Discovery of a previously unrecognized microdeletion syndrome of 16p11.2-p12.2. Nat Genet. **39**:1071–1073.

Bayés M, Magano LF, Rivera N, Flores R, Jurado LAP. 2003. Mutational mechanisms of Williams-Beuren syndrome deletions. Am J Hum Genet. **73**:131–151.

Cao H, Hastie AR, Cao D, Lam ET, Sun Y, *et al.* 2014. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. GigaScience **3**. Article number 34.

Carvalho CMB, Lupski JR. 2016. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. **17**: 224–238.

Christian SL, Fantes JA, Mewborn SK, Huang B, Ledbetter DH. 1999. Large genomic duplicons map to sites of instability in the Prader-Willi/Angelman syndrome chromosome region (15q11–q13). Hum Mol Genet. **8**:1025–1037.

Coe BP, Witherspoon K, Rosenfeld JA, Van Bon BWM, Vulto-Van Silfhout AT, *et al.* 2014. Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat Genet. **46**:1063–1071.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, *et al.* 2010. Origins and functional impact of copy number variation in the human genome. Nature. **464**:704–712.

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, *et al.* 2011. A copy number variation morbidity map of developmental delay. Nat Genet. **43**:838–846.

Cuscó I, Corominas R, Bayés M, Flores R, Rivera-Brugués N, *et al.* 2008. Copy number variation at the 7q11.23 segmental duplications is a susceptibility factor for the Williams-Beuren syndrome deletion. Genome Res. **18**:683–694.

Demaerel W, Mostovoy Y, Yilmaz F, Vervoort L, Pastor S, *et al.* 2019. The 22q11 low copy repeats are characterized by unprecedented size and structural variability. Genome Res. **29**:1389–1401.

Dennis MY, Harshman L, Nelson BJ, Penn O, Cantsilieris S, *et al.* 2017. The evolution and population diversity of human-specific segmental duplications. Nat Ecol Evol. **1**:69.

Eichler EE. 2002. Recent duplication, evolution and assembly of the human genome. Proc Annu Int Conf Comput Mol Biol Recomb. **17**:155.

Emanuel BS, Shaikh TH. 2001. Segmental duplications: an "expanding" role in genomic instability and disease. Nat Rev Genet. **2**:791–800.

Girirajan S, Rosenfeld JA, Cooper GM, Antonacci F, Siswara P, *et al.* 2010. A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. Nat Genet. **42**:203–209.

Hastie AR, Lam ET, Chun Pang AW, Zhang X, Andrews W, *et al.* 2017., Rapid Automated Large Structural Variation Detection in a Diploid Genome by NanoChannel Based Next-Generation Mapping. bioRxiv 102764. doi:10.1101/102764.

Hobart HH, Morris CA, Mervis CB, Pani AM, Kistler DJ, *et al.* 2010. Inversion of the Williams syndrome region is a common polymorphism found more frequently in parents of children with Williams syndrome. Am J Med Genet C Semin Med Genet. **154C**: 220–228.

Huang S-H, Lee T-Y, Lin Y-J, Wan L, Lai C-H, *et al.* 2017. Phage display technique identifies the interaction of severe acute respiratory syndrome coronavirus open reading frame 6 protein with nuclear pore complex interacting protein *NPIPB3* in modulating Type I interferon antagonism. J Microbiol Immunol Infect. **50**: 277–285.

Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, *et al.* 2014. Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res. **24**:688–696.

Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, *et al.* 2007. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nat Genet. **39**: 1361–1368.

Khan WA, Knoll JHM, Rogan PK. 2011. Context-based FISH localization of genomic rearrangements within chromosome 15q11. Mol Cytogenet. **4**:15.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, *et al.* 2008. Mapping and sequencing of structural variation from eight human genomes. Nature **453**:56–64.

Lander ES L, Birren LM, Nusbaum B, Zody C, Baldwin MC, *et al.* 2001. Initial sequencing and analysis of the human genome. Nature. **409**:860–921.

Leung AKY, Jin N, Yip KY, Chan TF. 2017. OMTools: a software package for visualizing and processing optical mapping data. Bioinformatics. **33**:2933–2935.

Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, *et al.* 2019. Genome maps across 26 human populations reveal population-specific patterns of structural variation. Nat Commun. **10**:1025.

Lupski JR. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet. **14**:417–422.

Lupski JR. 2007. Genomic rearrangements and sporadic disease. Nat Genet. **39**:S43–S47.

Lupski JR. 2009. Genomic disorders ten years on. Genome Med. **1**:42.

Maggiolini FAM, Cantsilieris S, D'Addabbo P, Manganelli M, Coe BP, *et al.* 2019. Genomic inversions and *GOLGA* core duplicons underlie disease instability at the 15q25 locus. PLoS Genet. **15**:e1008075.

Merla G, Brunetti-Pierri N, Micale L, Fusco C. 2010. Copy number variants at Williams-Beuren syndrome 7q11.23 region. Hum Genet. **128**:3–26.

Molina O, Blanco J, Anton E, Vidal F, Volpi EV. 2012. High-resolution fish on DNA fibers for low-copy repeats genome architecture studies. Genomics **100**:380–386.

O'Bleness M, Searles VB, Dickens CM, Astling D, Albracht D, *et al.* 2014. Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. BMC Genomics. **15**:387.

Osborne LR, Li M, Pober B, Chitayat D, Bodurtha J, *et al.* 2001. A 1.5 million–base pair inversion polymorphism in families with Williams-Beuren syndrome. Nat Genet. **29**:321–325.

Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, *et al.* 2010. Towards a comprehensive structural variation map of an individual human genome. Genome Biol. **11**:R52.

Peoples R, Franke Y, Wang Y-K, Pérez-Jurado L, Paperna T, *et al.* 2000. A physical map, including a BAC/PAC clone contig, of the Williams-Beuren syndrome–deletion region at 7q11. 23. Am J Hum Genet. **66**:47–68.

Perry GH, Yang F, Marques-Bonet T, Murphy C, Fitzgerald T, *et al.* 2008. Copy number variation and evolution in humans and chimpanzees. Genome Res. **18**:1698–1710.

Rex EB, Shukla N, Gu S, Bredt D, DiSepio D. 2017. A genome-wide arrayed cDNA screen to identify functional modulators of α7 nicotinic acetylcholine receptors. SLAS Discov Adv Sci Drug Discov. **22**:155–165.

Shaikh TH. 2000. Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. Hum Mol Genet. **9**:489–501.

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, *et al.* 2005. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet. **77**:78–88.

Sharp AJ, Mefford HC, Li K, Baker C, Skinner C, *et al.* 2008. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nat Genet. **40**:322–328.

Shaw CJ, Withers MA, Lupski JR. 2004. Uncommon deletions of the Smith-Magenis syndrome region can be reccurent when alternate low-copy repeats act as homologous recombination substrates. Am J Hum Genet. **75**:75–81.

Stankiewicz P, Lupski JR. 2002. Molecular-evolutionary mechanisms for genomic disorders. Curr. Opin Genet Dev. **12**:312–319.

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. Annu Rev Med. **61**:437–455.

Steinberg KM, Schneider VA, Graves-Lindsay TA, Fulton RS, Agarwala R, *et al.* 2014. Single haplotype assembly of the human genome from a hydatidiform mole. Genome Res. **24**:2066–2076.

Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, *et al.* 2013. Evolution and diversity of copy number variation in the great ape lineage. Genome Res. **23**:1373–1382.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, *et al.* 2005. Fine-scale structural variation of the human genome. Nat Genet. **37**:727–732.

Vollger MR, Dishuck PC, Sorensen M, Welch AE, Dang V, *et al.* 2019. Long-read sequence and assembly of segmental duplications. Nat Methods. **16**:88–94.

Wang J, Xi J, Zhang H, Li J, Xia Y, *et al.* 2019. Somatic mutations in renal cell carcinomas from Chinese patients revealed by targeted gene panel sequencing and their associations with prognosis and PD-L1 expression. Cancer Commun. **39**:37.

Wang L, Wang H, Wang H, Zhao Y, Liu X, *et al.* 2018., Extensive expansion of the speedy gene family in homininae and functional differentiation in humans. bioRxiv. 354886. doi:10.1101/354886.

Wong KHY, Ma W, Wei CY, Yeh EC, Lin WJ, *et al.* 2020. Towards a reference genome that captures global genetic diversity. Nat Commun. **11**:5482.

Communicating editor: J. Shendure