ARTICLES
https://doi.org/10.1038/s41592-018-0236-3

nature methods

Long-read sequence and assembly of segmental duplications

Mitchell R. Vollger[1], Philip C. Dishuck[1], Melanie Sorensen[1], AnneMarie E. Welch[1], Vy Dang[1], Max L. Dougherty[1], Tina A. Graves-Lindsay[2], Richard K. Wilson[3,4], Mark J. P. Chaisson[5]* and Evan E. Eichler[1,6]*

We have developed a computational method based on polyploid phasing of long sequence reads to resolve collapsed regions of segmental duplications within genome assemblies. Segmental Duplication Assembler (SDA; https://github.com/mvollger/SDA) constructs graphs in which paralogous sequence variants define the nodes and long-read sequences provide attraction and repulsion edges, enabling the partition and assembly of long reads corresponding to distinct paralogs. We apply it to single-molecule, real-time sequence data from three human genomes and recover 33–79 megabase pairs (Mb) of duplications in which approximately half of the loci are diverged (<99.8%) compared to the reference genome. We show that the corresponding sequence is highly accurate (>99.9%) and that the diverged sequence corresponds to copy-number-variable paralogs that are absent from the human reference genome. Our method can be applied to other complex genomes to resolve the last gene-rich gaps, improve duplicate gene annotation, and better understand copy-number-variant genetic diversity at the base-pair level.

Advances in sequencing technologies and the development of novel computational assembly algorithms are central to the complete characterization of complex genomes. Recent developments in long-read sequencing technology have dramatically improved the contiguity and speed at which de novo assemblies of complex genomes can be generated[1–8]. Individual laboratories, for example, can now accurately assemble >90% of a mammal's euchromatin in less than 1,000 contigs within a few months[5,6,9,10]. Despite these recent advances, substantial portions of mammalian genomes remain unresolved. This is especially true for large, highly identical repetitive regions, including heterochromatin and gene-rich regions associated with segmental duplications (SDs), which are larger than the majority of long reads[11–15].

SDs in most mammalian genomes are organized into complex regions typically >100 kilobase pairs (kb) in length and, by definition, are present at multiple locations. They contribute to dosage imbalance associated with disease[16,17] and are ten times more likely to contribute to normal copy number variation[18]. They are also a reservoir for gene innovations associated with species adaptations[19–21]. The size, copy number, and sequence identity of SDs means that they are usually the last regions of the genome to be sequenced and assembled, often using large-insert BAC (bacterial artificial chromosomes)[22,23]. More than half the gaps that remain in FALCON-based genome assemblies of single-molecule, real-time (SMRT) sequence data correspond to regions of SD. We estimate that the architecture of only 29.2% of SD bases is resolved in an assembly of the CHM1 genome (Supplementary Fig. 1, Supplementary Table 1, and Methods), with most disease-associated regions unresolved (Supplementary Table 2)[16,24]. Similarly, an assembly of NA12878 using longer Oxford Nanopore Technologies (ONT) ultra-long reads[25] shows moderate improvement (32.9% resolved) but leaves most SDs unresolved (Supplementary Fig. 1 and Supplementary Table 1).

Here, we develop and apply the Segmental Duplication Assembler (SDA) method. This method takes advantage of paralogous sequence variants (PSVs) and correlation clustering[26] to uniquely assemble different paralogs of SDs that were previously collapsed in long-read human genome assemblies. We apply it to real SMRT and ONT long-read datasets to resolve SDs in recent assemblies and generate >30 Mb of highly accurate, novel human genome sequence data. This method is computationally tractable, and its use can be extended to resolve collapsed repeat content in de novo assemblies of other mammalian genomes.

Results

The problem: unresolved SDs. Although ONT (https://nanoporetech.com/) and PacBio (https://www.pacb.com/) sequencing platforms generate long sequence reads, they also typically have high error rates of 10–15%. The predominant long-read assembly methods for whole-genome shotgun sequence assembly (WGSA) are based on read correction and overlapping corrected reads to construct larger sequence contigs—for example, Canu and FALCON[7,8]. The high error rate of long-read sequencing platforms is particularly problematic for distinguishing paralogous and allelic sequences because the duplications are highly identical (>95%) and well within the range of error from long-read sequencing. This leads to sequence reads being recruited and merged from both paralogs and alleles during the assembly process, creating collapses (Fig. 1) in which the assembled sequence and corrected sequence contig are in error. To quantify the effect of collapse and misassembly, we compared several recent assemblies generated using both ONT and SMRT sequence data (Supplementary Fig. 1 and Supplementary Note). With the requirement that contigs extend 50 kb into unique sequence in order to be considered fully resolved, we estimate that only 49.0–51.3 Mb of SDs are fully resolved (Supplementary Fig. 2),

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [2]The McDonnell Genome Institute at Washington University, Washington University School of Medicine, St. Louis, MO, USA. [3]Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA. [4]Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH, USA. [5]University of Southern California, Los Angeles, CA, USA. [6]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. *e-mail: mchaisso@usc.edu; eee@gs.washington.edu

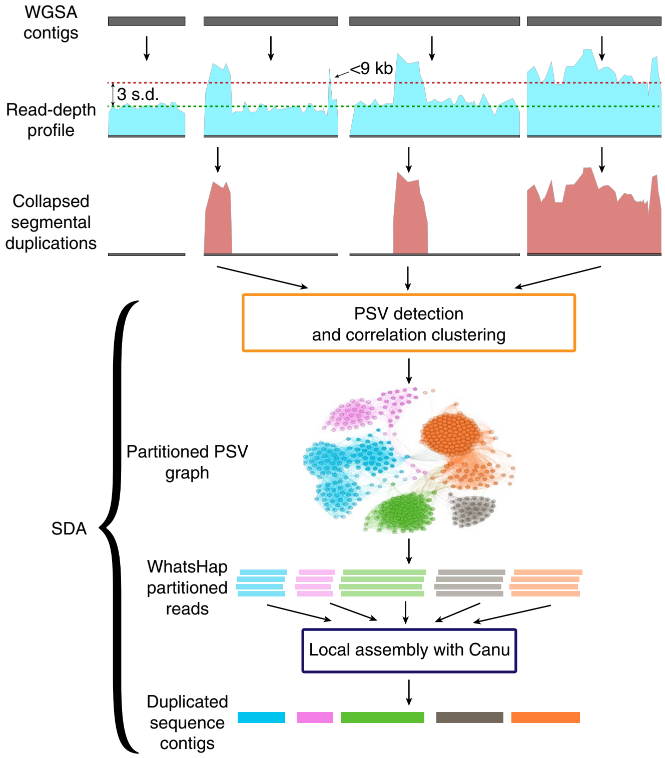
WGSA contigs
Read-depth profile
3 s.d.
<9 kb
Collapsed segmental duplications
PSV detection and correlation clustering
Partitioned PSV graph
SDA
WhatsHap partitioned reads
Local assembly with Canu
Duplicated sequence contigs

Fig. 1 | Flowchart of the SDA method. Regions of collapsed SDs are defined on the basis of whole-genome shotgun (WGS) sequence read-depth profiles using BLASR across sequence contigs generated from a de novo WGSA. Regions (>9 kb in length) with elevated sequence coverage (three s.d. plus the mean) and not entirely composed of common repeats are considered collapsed SDs. Sequence reads corresponding to the collapsed SDs are recovered and examined for variants at each position along the collapse. Single-base-pair substitutions that appear at the same threshold as unique sequencing depth are identified and flagged as PSVs, effectively partitioning reads into PSV clusters (WhatsHap). Sequence reads assigned to each PSV cluster are independently assembled using Canu and error-corrected using Quiver'.

leaving 71% (approximately 125 out of 175 Mb) of SDs associated with gaps. We note that even without an extension into unique sequence, 59.5–69.8% of SDs remain unresolved (Supplementary Fig. 1). We estimate that approximately 50 Mb of the duplications correspond to regions for which the assembly algorithm has collapsed highly identical duplications into the same contig. Analysis of an ONT assembly generated with ultra-long reads (2.5-fold coverage of reads over 100 kb)[25] showed a modest 8% improvement in SD assembly; however, most of the SDs still remained unresolved (Supplementary Fig. 1). As expected, the largest (>10 kb) and most identical duplications (>95% identity) were particularly enriched in unresolved SDs (Supplementary Fig. 2a) and frequently corresponded to annotated human genes (Supplementary Fig. 2b).

The approach: Segmental Duplication Assembler. SDA identifies high-confidence PSVs ab initio and groups them using correlation clustering with defined attraction and repulsion edges into PSV graphs (Fig. 1). The partitioned reads are then assembled independently, distinguishing the paralogous copies. Empirically, we observe that we are able to assemble large duplications with less than 0.5% sequence divergence (Supplementary Note). As a measure of reproducibility, we apply this method to four human genomes and validate the results and accuracy based on targeted BAC sequencing and analyses of specific duplicated loci.

We begin by identifying all collapsed duplications within each assembly based on an excess of sequencing read depth[11,27] (Methods). Within the CHM1 assembly[9], for example, we identify 283 regions of collapse averaging 43 kb in length (Table 1). When the 12.2 Mb of collapsed CHM1 duplications are mapped back to the reference, they span 52.3 Mb of sequence: 93% (48.6 Mb) are annotated as SDs, 88% of which (45.9 Mb) overlap with regions of unresolved SDs in CHM1. Next, we define PSVs corresponding to each collapsed segment. We define candidate PSVs by classifying the second-most frequent base at every position within the collapsed alignment and requiring sequence coverages consistent with a single-copy locus in order to distinguish PSVs from allelic variants (Methods). We next apply correlation clustering to filter false positive variants arising from sequencing errors and uniquely assign each remaining PSV to the paralog from which it originates. For each collapsed region, we construct a graph in which the PSVs define the nodes and the sequence reads define the edges. Attraction edges are formed when a read contains two or more PSVs, connecting two or more nodes. Similarly, repulsion edges are formed when PSVs are mutually exclusive across all of the sequence reads.

With this formulation of the problem, it is possible to address the correlation clustering objective, which is to minimize the number of repulsion edges within clusters and minimize the number of attraction edges between clusters. Correlation clustering offers a distinct advantage over many other clustering algorithms because it does not require the number of clusters as a starting input. It is therefore ab initio and defined entirely by the underlying sequence data. However, correlation clustering is a nondeterministic polynomial complete problem; therefore, we developed a heuristic to approximate the solution modeling after previous work[28]. The heuristic randomly assigns PSVs to clusters and then iteratively increases the size of the cluster by following positive edges that decrease the score of the entire graph (Methods).

Resolving SDs using SDA. We applied correlation clustering to each of the 283 collapsed regions in the CHM1 WGSA and generated a total of 668 distinct groupings. We created separate assemblies corresponding to each PSV graph partition using Canu followed by Quiver error correction. We successfully generated 590 assemblies in which a single contig was produced corresponding to 33.1 Mb of assembled sequence (Table 1 and Fig. 2) with an average sequence contig length of 60.7 kb. The median assembly length was 53.0 kb (mean 60.7 kb), and the maximum assembly size was 255.5 kb. In general, the length of the assembly correlated ($r = 0.67$, Pearson's correlation) with the size of the collapse (Supplementary Fig. 3). Of the 668 PSV graphs, 59 failed to generate an assembly and 19 assembled into multiple contigs. An inspection of those clusters that failed to assemble showed that the majority did so owing to an insufficient number of reads, while clusters with multiple contigs were the result of either incomplete PSV separation among multiple contigs or variable sequence coverage.

To assess the accuracy and contiguity of the assembled SDs, we mapped each sequence contig back to the human reference genome (GRCh38). Of these assemblies, 48.5% (286 out of 590) mapped to the human reference genome with at least 99.8% sequence identity over >90% of the contig length and accounted for approximately 18 Mb of sequence. Interestingly, a similar fraction of assembled contigs (51.5% (304 out of 590), corresponding to 15.5 Mb) showed greater sequence divergence, ranging from 96% to 99.8% sequence identity (Fig. 2a). We consider the contigs that 'match' at high identity to GRCh38 to be correctly assembled and classify those with lower sequence identity than expected based on allelic variation (<99.8%)[29] to be 'diverged'. As >0.2% divergence lies outside of the typical range of human allelic variation, such diverged sequences may represent different copies of the duplication not yet represented in the human genome. We examined in detail a few human-

Table 1 | SDA assembly statistics

Sample	De novo assembly					SDA				
	Assembly accession	Contig N50 (Mb)	Sequence coverage	Read N50 (kb)	Unresolved SDs (Mb)	Collapses (count, Mb)	Matched (count, Mb)	Diverged (count, Mb)	Multiple assemblies (count, Mb)	Failed
CHM1* (ref. [9])	GCA_001297185.1	26.9	61	20.5	124.1	283, 52.3	286, 17.98	304, 15.51	19, 1	59
CHM13* (ref. [9])	GCA_002884485.1	29.3	67	18.2	126.5	527, 86.6	685, 39.1	755, 35.0	69, 3.1	339
NA19240* (ref. [35])	GCA_001524155.4	29.1	61	17.5	124.1	489, 82.4	789, 38.8	983, 40.9	107, 5.8	257
NA12878† (ref. [25])	GCA_900232925.1	7.7	35	12.5	117.7	365, 52.5	38, 0.066	792, 22.1	8, 0.21	1,062

Genome summary statistics for four human genomes sequenced and assembled with long-read data. *Sequenced with SMRT and assembled with FALCON. †Sequenced with ONT and assembled with Canu. Collapses from the assemblies were subjected to SDA, and the count and Mb of 'matched' and 'diverged' contig assemblies to the human reference genome GRCh38 are shown.

specific gene families (for example, *SRGAP2* and *NOTCH2NL*) associated with neuroadaptation[20,30–34] that have been the target of detailed BAC-based sequence assemblies (Fig. 3, Supplementary Table 3, and Supplementary Fig. 4). Our analysis shows that we have successfully resolved the collapsed assemblies, recreating the sequence and gene models present in the reference genome. This includes the identification and characterization of paralog-specific structural variation with most sequence assemblies matching approximately 99.8–99.9% to their respective paralogs. Among these gene families, we estimate that 91–93% of all PSVs have been correctly assigned.

We repeated this analysis for three additional long-read human genome assemblies, including a second haploid genome (CHM13)[9], a diploid genome of African descent (YRI19240)[35], and a diploid genome assembled with ONT (NA12878)[25] (Table 1, Supplementary Note, and Supplementary Figs. 5–7). The proportion of matched and diverged sequence assemblies and of resolved SD regions was very similar among the PacBio genomes. For example, 83% (1,772 out of 2,136) of clusters resolved into single-contig assemblies for the African diploid genome assembly. In contrast, an analysis of a human genome assembly (NA12878) generated with ultra-long ONT reads showed more failed SD assemblies, although we note that the coverage of this genome was significantly less than that of the PacBio genome assemblies (Supplementary Fig. 7). Combining both the 'matched' and 'diverged' sequences, we estimate that the SDA method adds an additional 72.6 and 78.6 Mb of sequence corresponding to duplicated regions of the CHM13 and NA19240 human genomes, respectively.

Characterization of diverged duplications. We focused on the diverged duplications and considered two possibilities: the sequence could represent misassembled sequence or, alternatively, could represent additional copies not yet present in the human reference genome. The latter may be expected, given that SD regions are tenfold more likely to be copy number polymorphic[18] than unique regions of the genome. If diverged sequences resulted from the sequence and assembly of additional copies, we would expect a significant increase in the copy number differences for diverged sequences compared to duplicated sequences that matched the human reference genome (>99.8% sequence identity). Indeed, a comparison of the copy number differences for these two categories clearly showed that diverged copies were more likely ($P = 2.0 \times 10^{-5}$) to have a higher copy number in CHM1 (Fig. 2c) than duplicated sequences that matched the reference genome assembly.

As a more direct test, we sequenced and assembled 1,253 large-insert BAC clones (Supplementary Table 4) corresponding to regions of SD from a genomic library (CHORI-17) derived from CHM1 (refs. [36,37]) (Methods). Restricting our analysis to the 304

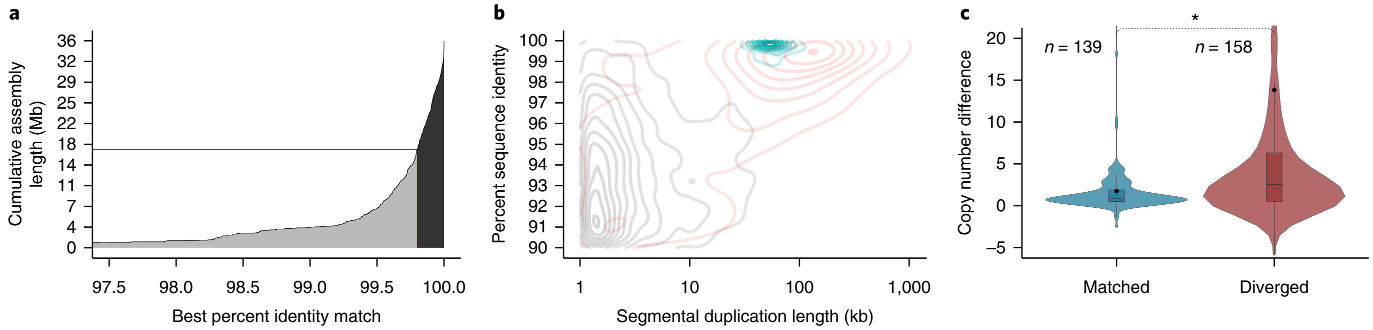

Fig. 2 | SDA results of the CHM1 human genome assembly. **a**, A cumulative distribution of the SDA assemblies and their percent sequence identity to their best match in the reference (<99.8% identity, gray; >99.8% identity, black). The cumulative number of assembly megabase pairs was calculated under the assumption that none of the assemblies overlapped (unlike in Table 1, where alignments to the human reference were used to avoid counting overlaps multiple times). **b**, Density plot of SDs plotted by length and percent identity. Black represents duplications resolved in the CHM1 assembly, red shows unresolved duplications in the CHM1 assembly, and blue represents paralogs assembled using SDA. Resolved SDA sequences are 'content' resolved and not ordered within the genome, whereas SDs in the assembly must extend into unique sequence on both sides to be considered resolved. **c**, Copy number difference (CND) between CHM1 and the reference genome (CHM1 copy number − reference genome copy number) comparing $n = 139$ SD regions that match (>99.8%) versus $n = 158$ diverged SD regions (<99.8% identity). The mean CND of the matched sequence is 1.75, and the mean CND of the diverged sequence is 13.82 (black dot) (two-sided Mann–Whitney test; $^*P < 0.0001$). The boxes indicate the range between the first and third quartiles, with the bold line indicating the median. The whiskers show the minimum and maximum within 1.5 times the interquartile range extending from the first and third quartiles. Copy number was estimated in CHM1, examining *k*-mer frequency found in Illumina WGS reads (Methods).

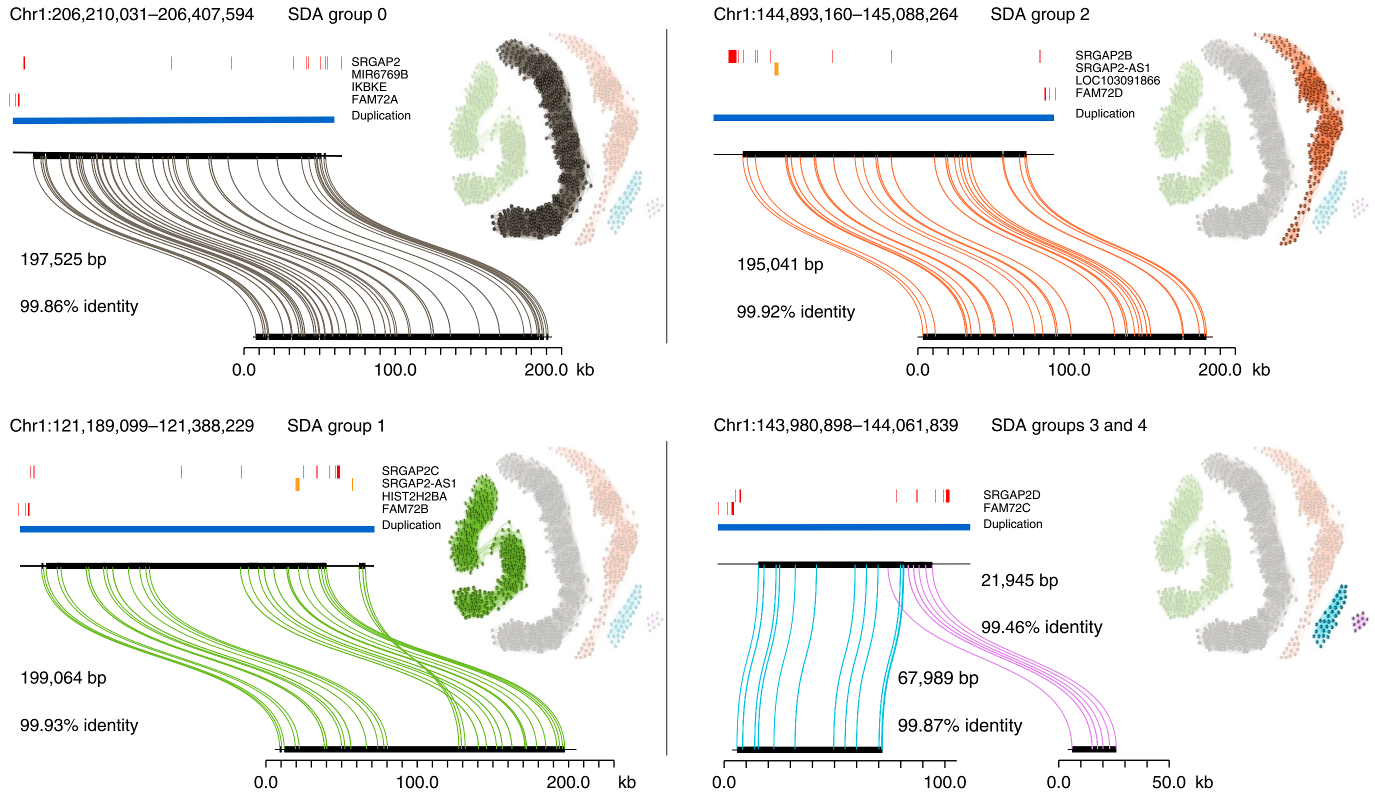

Fig. 3 | Sequence and assembly of *SRGAP2* loci in the CHM13 human genome. SDA sequence contigs from CHM13 aligned to the GRCh38 loci for *SRGAP2A*, *SRGAP2B*, *SRGAP2C*, and *SRGAP2D* using Miropeats[47]. The length and percent identity of each alignment is shown. Similarly, in CHM1 we found that, on average, our sequence was 99.91% identical over all four loci and >99.999% identical if only mismatched bases were counted as errors, as opposed to including indels. Adjacent to each alignment is the corresponding PSV graph, with the relevant PSVs highlighted. Each node represents a PSV, and loci are colored and numbered to reflect the grouping determined by correlation clustering. An edge is added between two nodes (PSVs) when a sequencing read contains both PSVs. The opacity of each node scales from 25% to 100% to reflect the position of the PSV along the collapse: 25% opacity reflects the first position along the collapse, and 100% reflects the final position. For a more detailed view of the opacity of the nodes, see Supplementary Fig. 12. Clusters 3 and 4 in the PSV graph represent the fourth paralog (*SRGAP2D*), which carries a large deletion in the middle relative to the other paralogs.

diverged sequences assembled by SDA from CHM1, we identified 105 diverged duplications that match the CHORI-17 clones. Each of these 105 sequences aligned to a clone over at least 90% of its length and at >99.8% sequence identity (mean sequence identity of 99.97%) (Fig. 4 and Supplementary Table 5). If we assume that our method targeted all SDs evenly across the whole genome, then we would expect to validate approximately 37.4% of the bases across our diverged sequences. Of our diverged sequences, 105 (or 36.3% of the bases) were validated and showed significantly better alignment to the CHM1 clone inserts than to GRCh38. We estimated the sequence accuracy for our assembled duplications to be 99.989% (quality value (QV) = 38.4) when we considered only single-base-pair mismatches, and 99.857% (QV = 28.4) when indels and mismatches were counted. We note that many of the 105 validated assemblies contain sequences associated with gene families and, thus, have the potential to recover missing genic sequence not yet annotated. For example, we assembled a paralog of *NBPF1* that is 1.2% diverged from the human reference but maps with >99.99% sequence identity to a CHM1 clone (Fig. 4 and Supplementary Table 6). Similarly, Sudmant and colleagues[38] identified an additional duplication in 16p12.1 that exists in most individuals but was absent from the reference. Using SDA, we recovered the proposed duplication[39] (Supplementary Fig. 8) with only one mismatched base pair across a 95-kb alignment to the BAC-generated contig.

We analyzed more systematically the utility of these orphan SDA contigs to generate more accurate gene models for 37 human-specific segmental duplication (HSD) gene families. We selected 213,450 bulk single-molecule sequencing RNA reads (Iso-Seq) from fetal and adult human brain enriched for HSDs[40]. We aligned Iso-Seq data and compared their mapping between SDA contigs and previous collapsed contigs in the CHM13 assembly. Transcripts showed improved mapping to the SDA contigs for 11 gene families to varying degrees (Supplementary Fig. 9). We identified six gene families (Fig. 5a) for which transcripts mapped better to the SDA assemblies than the human reference genome. A subset of transcripts from the *GPRIN2* (G-protein-regulated inducer of neurite outgrowth 2) gene family were most striking, with a 1.5% improvement. We aligned the second SDA *GPRIN2* contig that seemed to be missing from the reference and found that it spans a gap in GRCh38 flanked by SDs (Fig. 5b). Moreover, a previous analysis of Illumina WGS sequence shows that *GPRIN2* is polymorphic with copy number ranges from three to seven copies, with most humans carrying four, in contrast to other apes, which carry only two (haploid copy number = 1). Our analysis shows that both copies, *GPRIN2A* and *GPRIN2B*, are transcribed and encode similar open reading frames, although GPRIN2B has a 3-amino-acid insertion as well as several amino acid differences compared to the ancestral GPRIN2A (Supplementary Fig. 10). Interestingly, these PSVs have been erroneously classified as single-nucleotide variants (SNVs; with near 50% 'allele' frequency in dbSNP) because the reference is missing this second copy (Supplementary Table 7). Thus, the SDA contig not only improves gene annotation but also improves interpretation of human genetic variation.

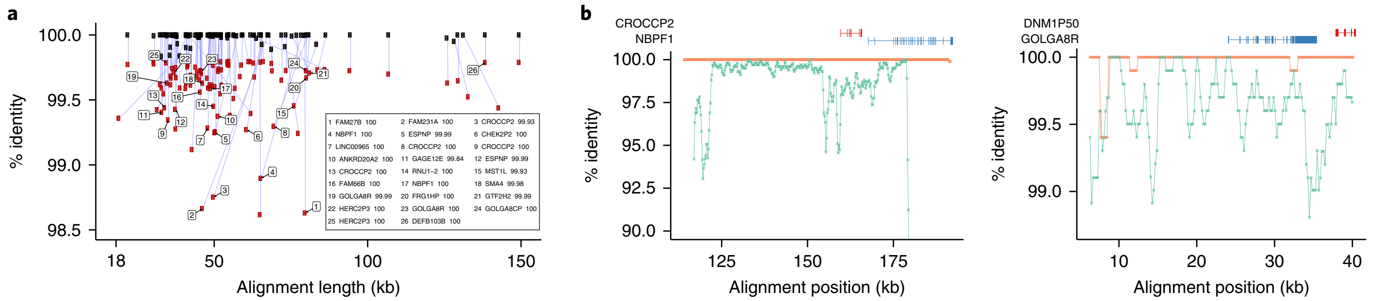

Fig. 4 | Correspondence between SDA sequence-diverged contigs and BACs. a, Alignment length and percent sequence identity match for $n = 105$ diverged SDA contigs compared to BAC clones (black) sequenced from the same source individual (CHM1) and the human reference genome (GRCh38) (red) (see Supplementary Tables 5 and 6 for more details). **b**, Two examples of genes corresponding to diverged duplications are shown where the SDA sequence is aligned to both the reference genome (green) and the CHM1 BACs (orange). BLASR alignments were computed in sliding 1,000-bp windows with 500-bp steps.

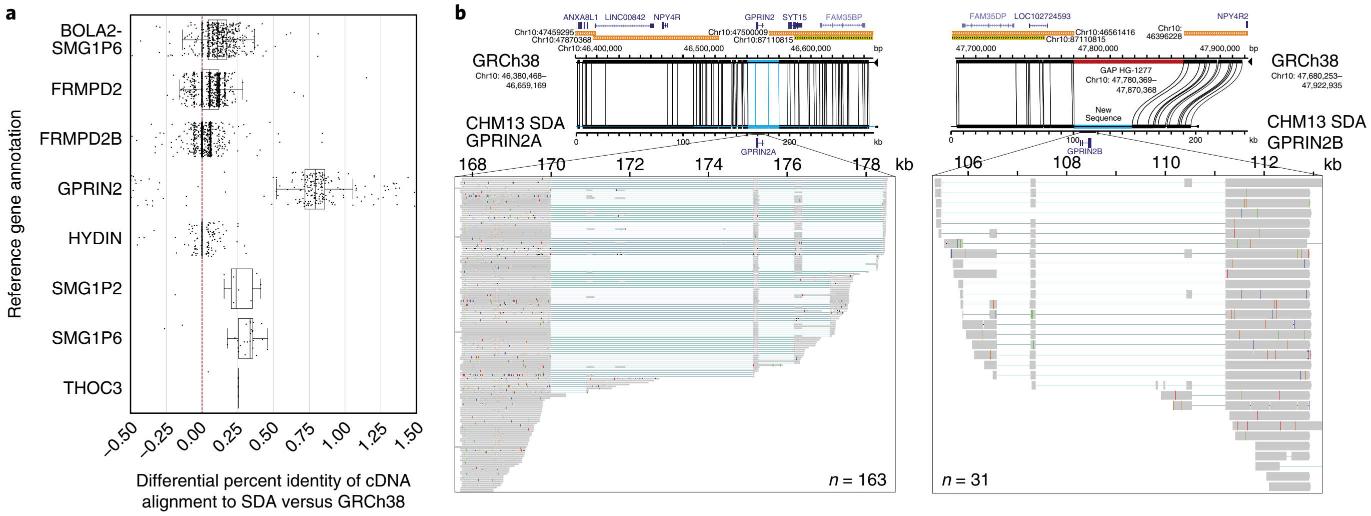

Fig. 5 | Gene discovery. a, The percent identity differential of the mapping of full-length Iso-Seq transcripts ($n = 4,718$) from HSDs to both GRCh38 and SDA results on CHM13. The red dotted line represents equal mapping between the two, whereas points to the right represent improved mapping with the SDA contigs. The boxes indicate the range between the first and third quartiles, with the bold line specifying the median. The whiskers show the minimum and maximum within 1.5 times the interquartile range extending from the first and third quartiles. **b**, *GPRIN2* SDA contigs compared (Miropeats) to the human reference assembly (GRCh38) with gene and SD annotation. The SDA contigs close a gap (red) in GRCh38, which contains a duplicate copy of *GPRIN2* denoted here by *GPRIN2B* (Supplementary Fig. 10 and Supplementary Table 7). Mapping of individual Iso-Seq transcripts (inset) is shown.

Discussion

Previously, we developed a computational algorithm[26] that could, in principle, assemble multi-copy duplications de novo using polyploid phasing[41–45] and demonstrated its efficacy using simulated datasets. Here, we developed SDA and applied it to WGSA collapsed duplications generated within existing human genome datasets. We specifically developed SDA to deal with different long-read datasets (Supplementary Note) and the generation of high-quality sequence contigs.

There are three strengths to SDA. First, our approach does not require PSVs to be predefined and, as a result, can be applied to any genome assembly for which long-read data of sufficient depth have been generated. A similar concept was recently applied to partition viral quasispecies[46]. Second, our validation results suggest that the paralog-specific assemblies are highly accurate (99.86–99.99%). Importantly, the approach enables missing paralogs to be sequenced, especially within regions of extensive copy number variation. This is particularly exciting because it enables previously uncharacterized forms of human genetic variation to be sequence-resolved for the first time. Finally, our analysis of the human genome suggests that the majority of collapsed duplications are at least partially resolved (Fig. 2). As unassembled SDs typically represent approximately 70–90 Mb of sequence per genome, recovery of 33–79 Mb is equivalent to recovery of an entire chromosome's worth of DNA for which accurate gene models can be constructed (Table 1 and Supplementary Table 8). The method that we have developed can be effectively applied to any genome for which long-read WGSA data exist, providing access to the duplicated regions and the genes therein.

Notwithstanding these advances, limitations remain. The majority of the sequence contigs that we generated with SDA are small (approximately 54 kb) and are not yet commensurate with the average contig lengths generated by long-read sequencing and assembly of unique regions of the genome. Only a small fraction (22%) of SDA contigs transition into unique sequence such that overlaps can be unambiguously assigned to the main genome assembly (Supplementary Fig. 11). Our new duplicated sequence contigs are not yet fully integrated into the genome, and many of the resolved duplications remain 'orphan' contigs in the absence of additional long-range mapping data. Direct integration of our SDA tool into popular long-read assemblers, to create long-range linkage information, may not be advisable even if it were possible. Parameter optimization for SD assembly would be likely to come with costs for the remaining 95% of the genome. There are distinct advantages to performing bulk WGSA followed by a second-tier analysis to focus on the collapsed regions of the assembly. This is because overlap stringency should differ for high-identity duplications, and because PSVs provide important information for determining overlaps in these more difficult-to-assemble regions.

Although we have shifted the accessible portions of SDs to larger (>50 kb) and more identical regions (approximately 99%), not all regions can be resolved by this approach. Duplications that are virtually identical cannot be distinguished and will require even longer read data, such as the ultra-long reads (>100 kb) possible with ONT[25]. We have developed and benchmarked SDA primarily with PacBio sequence data, but we have also applied it to long-read sequence data from other platforms such as ONT (Supplementary Note). Our initial analysis of the ultra-long-read genome assembly of NA12878 (ref. [25]), for example, showed a slight improvement of 8% in SD assembly (Supplementary Fig. 1). However, most of the high-identity SDs remained unresolved with a similar number of collapsed duplications ($n=365$) compared to PacBio genome assemblies. Application of SDA to the ONT dataset resulted in far fewer resolved assemblies (Supplementary Fig. 7), with an overall lower accuracy of the assembled sequence contigs. An important difference, however, is sequence coverage. The NA19240 PacBio assembly was sequenced at 73-fold sequence coverage versus the 35-fold ONT genome assembly. We note that while ultra-long ONT sequence reads were less successful in resolving SDs, they were useful as orthogonal data to validate PacBio SDA contigs (Supplementary Note). If long reads in excess of 200 kb can be routinely generated with sufficient coverage to correct sequence error, it is possible that most SDs could be resolved by WGSA. The rapid advance of long-read sequencing technology may make the routine generation of ultra-long reads from low quantities of DNA a reality in the near future. Such advances would open up the possibility that other highly repetitive regions, such as centromeres and acrocentric DNA, could be routinely sequenced and assembled for the first time.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41592-018-0236-3.

Received: 1 June 2018; Accepted: 30 October 2018;
Published online: 17 December 2018

References

1. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
2. Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**, 61–65 (2011).
3. Seo, J. S. et al. De novo assembly and phasing of a Korean human genome. *Nature* **538**, 243–247 (2016).
4. Shi, L. et al. Long-read sequencing and de novo assembly of a Chinese genome. *Nat. Commun.* **7**, 12065 (2016).
5. Bickhart, D. M. et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **49**, 643–650 (2017).
6. Gordon, D. et al. Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344 (2016).
7. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
8. Chin, C.-S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
9. Huddleston, J. et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).
10. Kronenberg, Z. N. et al. High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar6343 (2018).
11. Kelley, D. R. & Salzberg, S. L. Detection and correction of false segmental duplications caused by genome mis-assembly. *Genome. Biol.* **11**, R28 (2010).
12. Pop, M. Shotgun sequence assembly. *Adv. Comput.* **60**, 193–248 (2004).
13. Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl Acad. Sci. USA* **98**, 9748–9753 (2001).
14. Pevzner, P. A., Tang, H. & Tesler, G. De novo repeat classification and fragment assembly. *Genome Res.* **14**, 1786–1796 (2004).
15. Myers, E. W. The fragment assembly string graph. *Bioinformatics* **21**, ii79–ii85 (2005).
16. Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).
17. Sharp, A. J. et al. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat. Genet.* **38**, 1038–1042 (2006).
18. Sudmant, P. H. et al. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).
19. Chen, J. et al. Bovine NK-lysin: copy number variation and functional diversification. *Proc. Natl. Acad. Sci. USA* **112**, E7223–E7229 (2015).
20. Dennis, M. Y. & Eichler, E. E. Human adaptation and evolution by segmental duplication. *Curr. Opin. Genet. Dev.* **41**, 44–52 (2016).
21. Abegglen, L. M. et al. Potential mechanisms for cancer resistance in elephants and comparative cellular response to DNA damage in humans. *J. Am. Med. Assoc.* **314**, 1850–1860 (2015).
22. Church, D. M. et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.* **7**, e1000112 (2009).
23. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
24. Emanuel, B. S. & Shaikh, T. H. Segmental duplications: an 'expanding' role in genomic instability and disease. *Nat. Rev. Genet.* **2**, 791–800 (2001).
25. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
26. Chaisson, M. J., Mukherjee, S., Kannan, S. & Eichler, E. E. Resolving multicopy duplications *de novo* using polyploid phasing. *RECOMB* **10229**, 117–133 (2017).
27. Bailey, J. A. et al. Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
28. Ailon, N., Charikar, M. & Newman, A. Aggregating inconsistent information. *J. Assoc. Comput. Mach.* **55**, 1–27 (2008).
29. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
30. Fiddes, I. T. et al. Human-specific NOTCH2NL genes affect notch signaling and cortical neurogenesis. *Cell* **173**, 1356–1369 (2018).
31. Florio, M. et al. Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. *eLife* **7**, e32332 (2018).
32. Dennis, M. Y. et al. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**, 912–922 (2012).
33. Nuttle, X. et al. Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. *Nat. Methods* **10**, 903–909 (2013).
34. Dennis, M. Y. et al. The evolution and population diversity of human-specific segmental duplications. *Nat. Ecol. Evol.* **1**, 0069 (2017).
35. Steinberg, K. M. et al. High-quality assembly of an individual of Yoruban descent. *bioRxiv* Preprint at https://www.biorxiv.org/content/early/2016/08/02/067447 (2016).
36. Chaisson, M. J. P. et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
37. BACPAC Resources. The CHORI-17 BAC library from a hydatidiform (haploid) mole. *CloneDB* https://www.ncbi.nlm.nih.gov/clone/library/genomic/76/ (2018).
38. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
39. Nuttle, X. et al. Emergence of a *Homo sapiens*–specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* **536**, 205–209 (2016).
40. Dougherty, M. L. et al. Transcriptional fates of human-specific segmental duplications in brain. *Genome Res.* **28**, 1566–1576 (2018).
41. Das, S. & Vikalo, H. SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics* **16**, 260 (2015).
42. Aguiar, D. & Istrail, S. Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics* **29**, i352–i360 (2013).
43. Berger, E., Yorukoglu, D., Peng, J. & Berger, B. in *Research in Computational Molecular Biology: RECOMB 2014* (ed Sharan, R.) 18–19 (Springer, 2014).
44. Puljiz, Z. & Vikalo, H. Decoding genetic variations: communications-inspired haplotype assembly. *IEEE/ACM. Trans. Comput. Biol. Bioinform.* **13**, 518–530 (2016).

45. Bonizzoni, P. et al. On the minimum error correction problem for haplotype assembly in diploid and polyploid genomes. *J. Comput. Biol.* **23**, 718–736 (2016).
46. Artyomenko, A. et al. Long single-molecule reads can resolve the complexity of the influenza virus composed of rare, closely related mutant variants. *J. Comput. Biol.* **24**, 558–570 (2017).
47. Parsons, J. D. Miropeats: graphical DNA sequence comparisons. *Comput. Appl. Biosci.* **11**, 615–619 (1995).

Acknowledgements

The authors thank S. Cantsilieris and D. Gordon for technical assistance, J. Underwood for recommendations regarding the analysis of HSDs and Iso-Seq data, and T. Brown for help in editing this manuscript. This work was supported, in part, by grants from the US National Institutes of Health (NIH) (HG002385 to E.E.E., HG007635 to R.K.W. and E.E.E., and HG003079 to R.K.W.). M.R.V. was supported by a National Library of Medicine (NLM) Big Data Training Grant for Genomics and Neuroscience (5T32LM012419-04). P.C.D. was supported by a National Human Genome Research Institute (NHGRI) training grant (5T32HG000035-23). E.E.E. is an investigator of the Howard Hughes Medical Institute.

Author contributions

M.R.V., M.J.P.C., and E.E.E. developed the SDA method; R.K.W. and T.A.G.-L. generated the PacBio genome sequence; M.S., A.E.W., M.R.V., and V.D. sequenced and analyzed the BAC clone insert; P.C.D., M.R.V., and M.L.D. carried out Iso-Seq analysis; M.R.V. organized the supplementary material; M.R.V., E.E.E., and M.J.P.C. wrote the manuscript; M.R.V. and P.C.D. produced the display items.

Competing interests

E.E.E. is on the scientific advisory board (SAB) of DNAnexus, Inc.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-018-0236-3.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.J.P.C. or E.E.E.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Human genome assemblies. We analyzed three human genome assemblies derived from haploid (CHM1 and CHM13)[9] and diploid source material (NA19240[35]) of African descent. FALCON genome assemblies were previously generated from at least 61-fold SMRT sequence using P6C4 chemistry generated on the PacBio RS II sequencing platform. We also analyzed one recent human genome assembly (NA12878) generated with ultra-long ONT sequence reads[25].

SD characterization. We mapped each human de novo assembly to the human reference genome GRCh38 using MashMap 2.0 (default settings)[48] and defined SD regions on the basis of intersection with annotated SDs in GRCh38. Sequence contigs overlapping SDs were defined as resolved if the contig completely contained the SD sequence and extended at least 50 kb either side into unique sequence. We compared the number of resolved and unresolved contigs (Fig. 1a) for each assembly as a function of SD block length and maximum percent identity. Scripts are available at https://github.com/mvollger/segDupPlots (as well as a more detailed description of the analysis in the 'README' file).

Assembly collapse and PSV definition. Within each assembly, we identified collapsed SDs by mapping SMRT or ONT sequencing reads back to each genome using BLASR[49] (version rc46) or minimap2[50] (version 2.11) for ONT. Using unique regions, we computed the read coverage and s.d. across 100-bp windows using the following BLASR settings: blasr $READS $ASM -sa $ASMSA / -sdpTupleSize 13 -sdpMaxAnchorsPerPosition 10 -maxMatch 25 / -minMapQV 30 -bestn 2 -advanceExactMatches 15 / -clipping subread --sam. We excluded regions with >75% common repeat elements (RepeatMasker version 2004/03/06 –e wublast) and regions in the bottom or top two percentiles. We defined collapsed regions as those with a mean sequence coverage >3 s.d. beyond the mean sequence coverage of the de novo assembly and that were at least 9,000 bp in length (as smaller regions were routinely sequenced and assembled). We examined all regions of collapse for the presence of SNVs and catalogued the second-most common base at each position within the collapsed region using more sensitive BLASR settings: blasr {input.basreads} {input.ref} / -sam -preserveReadTitle -clipping subread / -bestn 1 / -mismatch 3 -insertion 9 -deletion 9 -minAlignLength 500. We defined these SNVs as potential PSVs if the sequence coverage was consistent with the read depth of unique regions. Three thresholds were applied to determine whether an SNV was also a PSV. First, the total depth at the given position had to be at least the mean coverage plus 3 s.d. Second, the frequency of the second-most frequent base had to be less than the mean coverage. Finally, the frequency of the second-most frequent base had to be greater than the mean coverage minus 3 s.d. or half the mean coverage, whichever was greater. This process favors the selection of PSVs over allelic variants (Supplementary Fig. 4). We developed a Snakemake pipeline for this analysis ProcessCollapsedAssembly.py, which can be found at https://github.com/mvollger/SDA.

PSV graph construction. We constructed graphs for collapsed regions in which each PSV corresponds to a node and sequence reads represent edges. Attraction edges are created when two PSV nodes have a substantial number of sequencing reads that contain both PSVs. Among reads containing both PSVs, we tested whether each PSV was more likely to be real or a sequencing error using the ratio of two binomial tests. If at each PSV the $\log_{10}$ ratio of the two binomial tests was at least 1.5 (that is, approximately 31 times more likely to be real than error), then an attraction edge was formed. Repulsion edges were created between any PSVs for which less than 10% of the mean coverage of sequencing reads carried both PSVs.

Correlation clustering. We initially added all nodes to an unclustered set from which a node was randomly selected and then expanded upon by iteratively searching for neighbors of this node that reduced the overall score of the PSV graph (that is, minimized the objective function). As nodes that meet this criterion are added to the cluster, they are removed from the unclustered set. This process was repeated until there were no unclustered nodes, as described previously[26]. Next, all pairwise clusters are examined to see whether they would improve the score of the graph if combined into a single cluster. Clusters are combined starting with the pairwise cluster that most improves the score of the correlation clustering objective. Clusters of three or fewer nodes are removed. The correlation clustering heuristic is run independently 15 times each with different random initializations and the clustering that best minimizes the correlation clustering objective is used to construct the final PSV clusters. It can be the case that in the construction of the PSV graph the PSVs are already clustered appropriately as unconnected components in the graph. In this case the application of correlation clustering is unnecessary to phase PSVs.

PSV read partition and assembly. To partition SMRT or ONT sequencing reads according to the PSV clusters defined by correlation clustering, we apply WhatsHap[51] (version 0.16) using the following parameters: whatshap haplotag $INPUT_VCF $INPUT_BAM -o $OUTPUT_BAM. Phasing was run on the entire set of reads for each PSV cluster (that is, if there were five PSV clusters, WhatsHap was run five times to create five partitions of reads). After partitioning the reads into different paralogs, we independently assembled each correlation cluster with Canu version 1.5, and then applied error correction (Quiver v 1.1.0) using the same set of reads. Specialized parameters were applied such that Canu could execute on such short contigs (https://github.com/mvollger/SDA/blob/master/SDA.2.snakemake.py).

Illumina copy number estimate. We estimated copy number in CHM1 by examining *k*-mer frequency found in Illumina WGS reads, following methods described previously[18]. We used a similar approach to estimate copy number in GRCh38, except we generated simulated reads using the reference and then estimated copy number in the same fashion using the simulated reads.

BAC clone insert sequencing. BAC clones from CHORI-17 (CH17) clone libraries (http://bacpac.chori.org) were hybridized with probes targeting complex or highly duplicated regions of GRCh38 ($n=727$) or based on previously sequenced clones ($n=526$)[36,37]. DNA from positive clones was isolated by a modified alkaline lysis miniprep procedure, as follows: cell pellet was resuspended in 200 µl of Qiagen buffer P1 with RNase and lysed with 200 µl of 0.2 M NaOH/1%SDS solution for 5 min. Lysis was neutralized with 280 µl of 3 M NaOAc, pH 4.8. Neutralized lysate was incubated on ice for up to 20 min, collected by centrifugation for 30 min at 4,000 r.p.m., concentrated by standard isopropanol and then ethanol precipitation, and resuspended in 25 µl of 10 mM Tris-HCl, pH 8.5. We prepared barcoded libraries from clone DNA using Illumina-compatible Nextera DNA sample prep kits (Epicentre, catalog number GA09115) as described previously[52] and carried out paired-end sequencing (125-bp reads) on an Illumina HiSeq 2500. Reads were then mapped to the reference genome, GRCh38, to identify singly unique nucleotide *k*-mers (SUNKs), defined as 30-mers that identify a region of the genome and can be used in conjunction with short-read sequencing data to genotype highly identical paralogs[53]. This SUNK mapping was used to select a subset of positive clones for PacBio sequencing. BAC DNA from selected clones was isolated using a High Pure Plasmid Isolation Kit from Roche Applied Science following the manufacturer's instructions, using 6 ml of LB media with chloramphenicol selective marker. We pooled non-overlapping BACs at equal molar amounts before library preparation. Approximately 1 µg of DNA per BAC was pooled and sheared using a Covaris g-TUBE. Libraries were processed using the PacBio SMRTbell Template Prep kit following the protocol "Procedure and Checklist—20 kb Template Preparation Using BluePippin Size-Selection System." Libraries were size-selected on the Sage PippinHT with a start value of 10,000–12,000 and an end value of 50,000. The DNA/Polymerase Binding Kit (P6-C4 chemistry) was used to bind DNA template to DNA polymerase, and the MagBead kit was used to capture DNA polymerase–template complexes for loading. Libraries were sequenced on the PacBio RS II platform. We performed de novo assembly of pooled BAC inserts using Canu v1.5 (ref. [7]). Reads were masked for vector sequence (pBACGK1.1) and assembled with Canu, then subjected to consensus sequence calling with Quiver. Canu is specifically designed for assembly with long error-prone reads, whereas Quiver is a multi-read consensus algorithm that uses the raw pulse and base call information generated during SMRT sequencing for error correction. We reviewed PacBio assemblies for misassembly by visualizing the read depth of PacBio reads in Parasight (http://eichlerlab.gs.washington.edu/jeff/parasight/index.html), using coverage summaries generated during the resequencing protocol.

Statistical information. Statistical information for analysis of copy number differences is provided in Fig. 2. The statistical analysis used to link PSVs with long-read data is described above in the section 'PSV graph construction'.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. Code for analyzing the resolved and unresolved SDs in a de novo assembly can be found at https://github.com/mvollger/segDupPlots. Code for processing de novo assemblies to find collapses and running SDA can be found at https://github.com/mvollger/SDA.

Data availability

SMRT WGS for CHM1, CHM13, and NA12940 from this study are available at the NCBI Sequence Read Archive (SRA) under accession numbers SRP044331 for CHM1; SRX818607, SRX825542, and SRX825575–SRX825579 for CHM13; and SRX1093000, SRX1093555, SRX1093654, SRX1094289, SRX1094374, SRX1094388, and SRX1096798 for NA19240. ONT WGS data are available at https://github.com/nanopore-wgs-consortium/NA12878/blob/master/Genome.md. De novo assemblies of CHM1, CHM13, NA12940, and NA12878 from this study are available at the NCBI Assembly database under accession numbers GCA_001297185.1, GCA_000983455.2, GCA_001524155.4, and GCA_900232925.1, respectively. Assembled CHORI-17 BACs are available at the NCBI Clone DB (https://www.ncbi.nlm.nih.gov/clone/) under the accession numbers listed in Supplementary Table 4. Information about length, PSVs, and mapping location in GRCh38 can be found for all the SDA contigs generated, in Supplementary Table 8. Additional data that support the findings of this study are available from the corresponding author upon request.

References

48. Jain, C., Koren, S., Dilthey, A., Phillippy, A. M. & Aluru, S. A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics* **34**, i748–i756 (2018).
49. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
50. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
51. Patterson, M. et al. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *J. Comput. Biol.* **22**, 498–509 (2015).
52. Steinberg, K. M. et al. Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat. Genet.* **44**, 872–880 (2012).
53. Sudmant, P. H. et al. Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).

nature research

Corresponding author(s): Evan E. Eichler
Mark J.P. Chaisson

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

No software was used.

Data analysis

Code for analyzing the resolved and unresolved segmental duplications in a de novo assembly can be found at https://github.com/mvollger/segDupPlots. Code for processing de novo assemblies to find collapses and running SDA can be found at https://github.com/mvollger/SDA.

Main software used:
canu version 1.5
whatshap version 0.16
blasr version rc46
minimap2 version 2.11
quiver version 1.1.0
mashmap version 2.0
miropeats version 1.0
RepeatMasker version 2004/03/06
samtools version 1.9
bedtools version 2.27
gephi version 0.9.2
miniasm version 0.3

wtdbg version 1.2.8
snakemake version 5.2.2

For a complete list software used by the distributed version of SDA see:
https://github.com/mvollger/SDA/blob/master/ymls/python3.yml, and https://github.com/mvollger/SDA/blob/master/ymls/python2.yml

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

SMRT WGS for CHM1, CHM13, and NA12940 from this study are available at the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession numbers SRP044331 for CHM1; SRX818607, SRX825542, and SRX825575-SRX825579 for CHM13; and SRX1093000, SRX1093555, SRX1093654, SRX1094289, SRX1094374, SRX1094388, and SRX1096798 for NA19240. ONT WGS data are available at https://github.com/nanopore-wgs-consortium/NA12878/blob/master/Genome.md. De novo assemblies of CHM1, CHM13, NA12940, and NA12878 from this study are available at the NCBI Assemblies database (Assembly; https://www.ncbi.nlm.nih.gov/assembly/) under accession numbers GCA_001297185.1, GCA_000983455.2, GCA_001524155.4, and GCA_900232925.1, respectively. Assembled CHORI-17 BACs are available at the NCBI Clone database (Clone; https://www.ncbi.nlm.nih.gov/clone/) under the accession numbers listed in Table S4.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We analyzed four genome assemblies using our method (CHM1, CHM13, NA19240, and NA12878). We believe this is a sufficient sample size since these samples represent: diverse individuals, different sequencing technologies, and different genomic architectures (hydatidiform moles and true diploids), all while still showing the utility and generalizability of the method.
Data exclusions	No data were excluded
Replication	Our method is computational and non-random. Rerunning with the same input always produced the same output.
Randomization	This is not applicable since we make no claims on covariation between the genomes we analyzed.
Blinding	Not applicable, there was no group allocation done during data analysis, so no blinding was required.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a | Involved in the study
☒ □ Unique biological materials
☒ □ Antibodies
☒ □ Eukaryotic cell lines
☒ □ Palaeontology
☒ □ Animals and other organisms
☒ □ Human research participants

Methods

n/a | Involved in the study
☒ □ ChIP-seq
☒ □ Flow cytometry
☒ □ MRI-based neuroimaging