


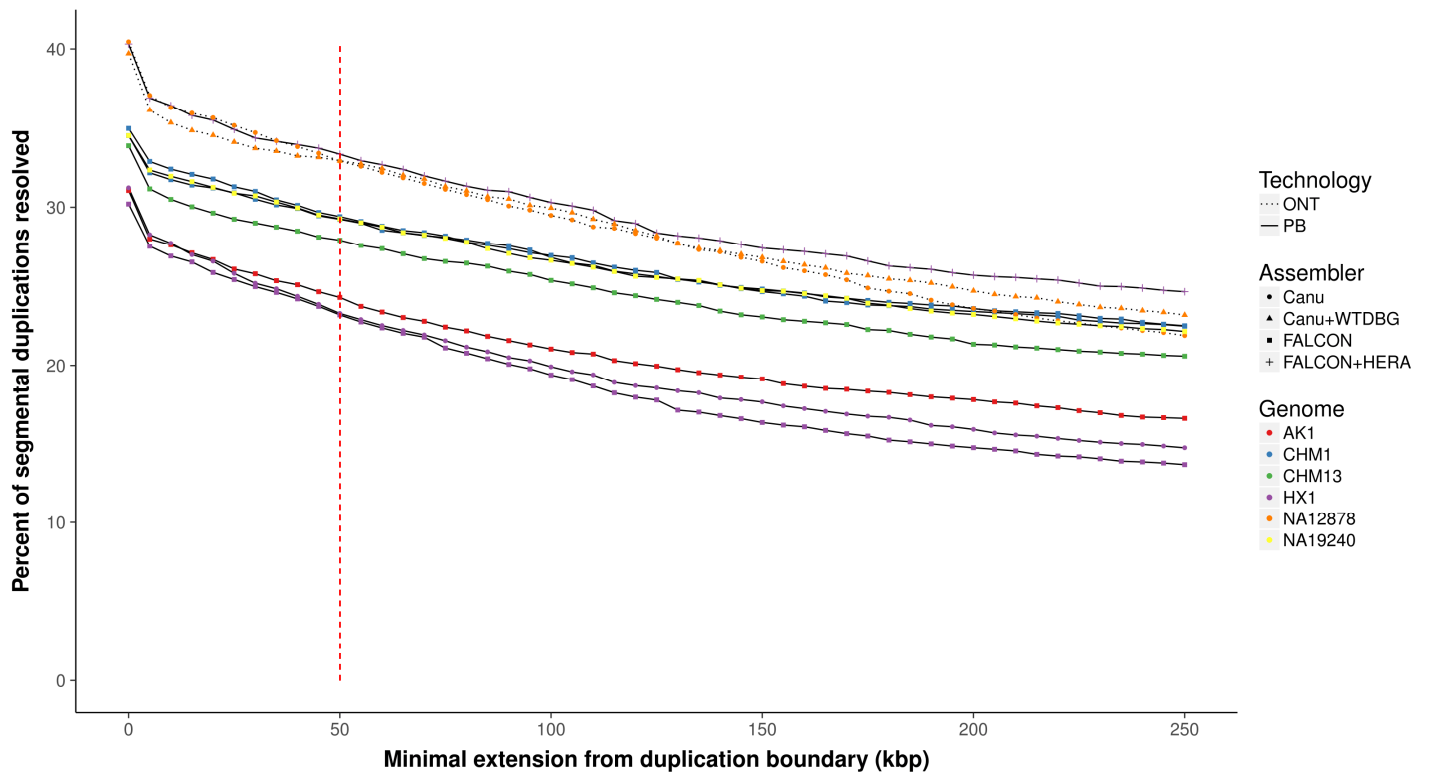


In the format provided by the authors and unedited.

Long-read sequence and assembly of segmental duplications

Mitchell R. Vollger ¹, Philip C. Dishuck ¹, Melanie Sorensen¹, AnneMarie E. Welch¹, Vy Dang¹, Max L. Dougherty¹, Tina A. Graves-Lindsay², Richard K. Wilson^{3,4}, Mark J. P. Chaisson^{5*} and Evan E. Eichler ^{1,6*}

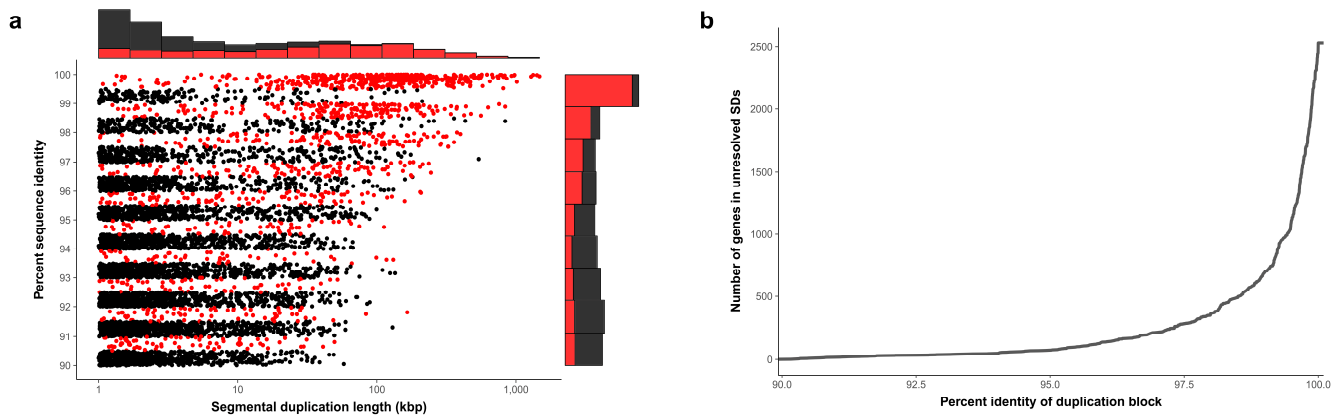
¹Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. ²The McDonnell Genome Institute at Washington University, Washington University School of Medicine, St. Louis, MO, USA. ³Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA. ⁴Department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



Supplementary Figure 1

Proportion of resolved SDs in different PacBio (PB)/ONT genome assemblies.

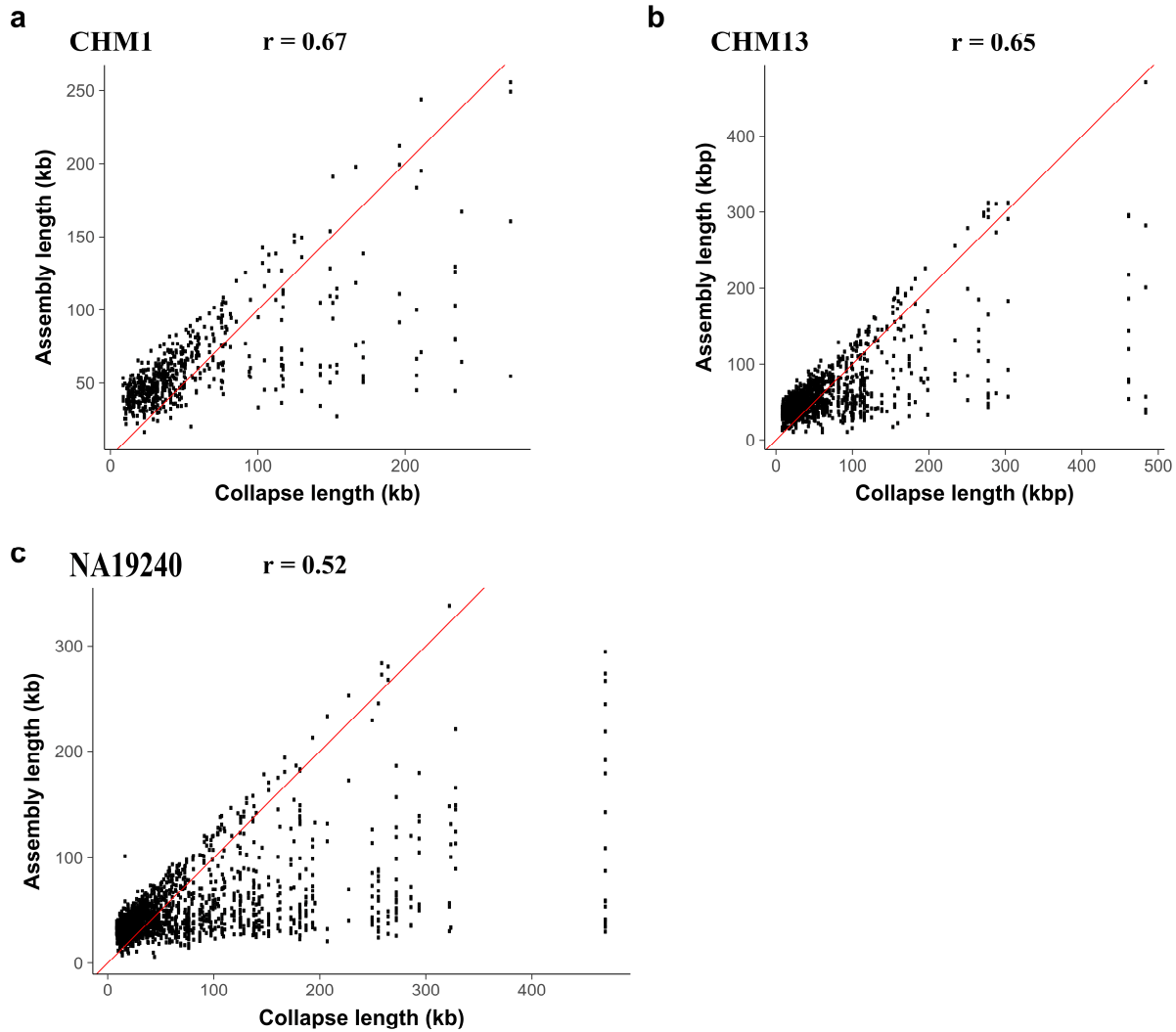
The figure shows the percent of SD bases that are resolved in human genome assemblies plotted as a function of the length of minimum extension of the alignment past the duplication. The number of resolved SD base pairs is relatively constant irrespective of the requirement of flanking unique base pairs. The dashed red line indicates the threshold chosen for our analysis used to generate the first panel in Supplementary Fig. 2 and the fraction of resolved SDs in Supplementary Table 1.



Supplementary Figure 2

Resolution of SDs in SMRT genome assemblies.

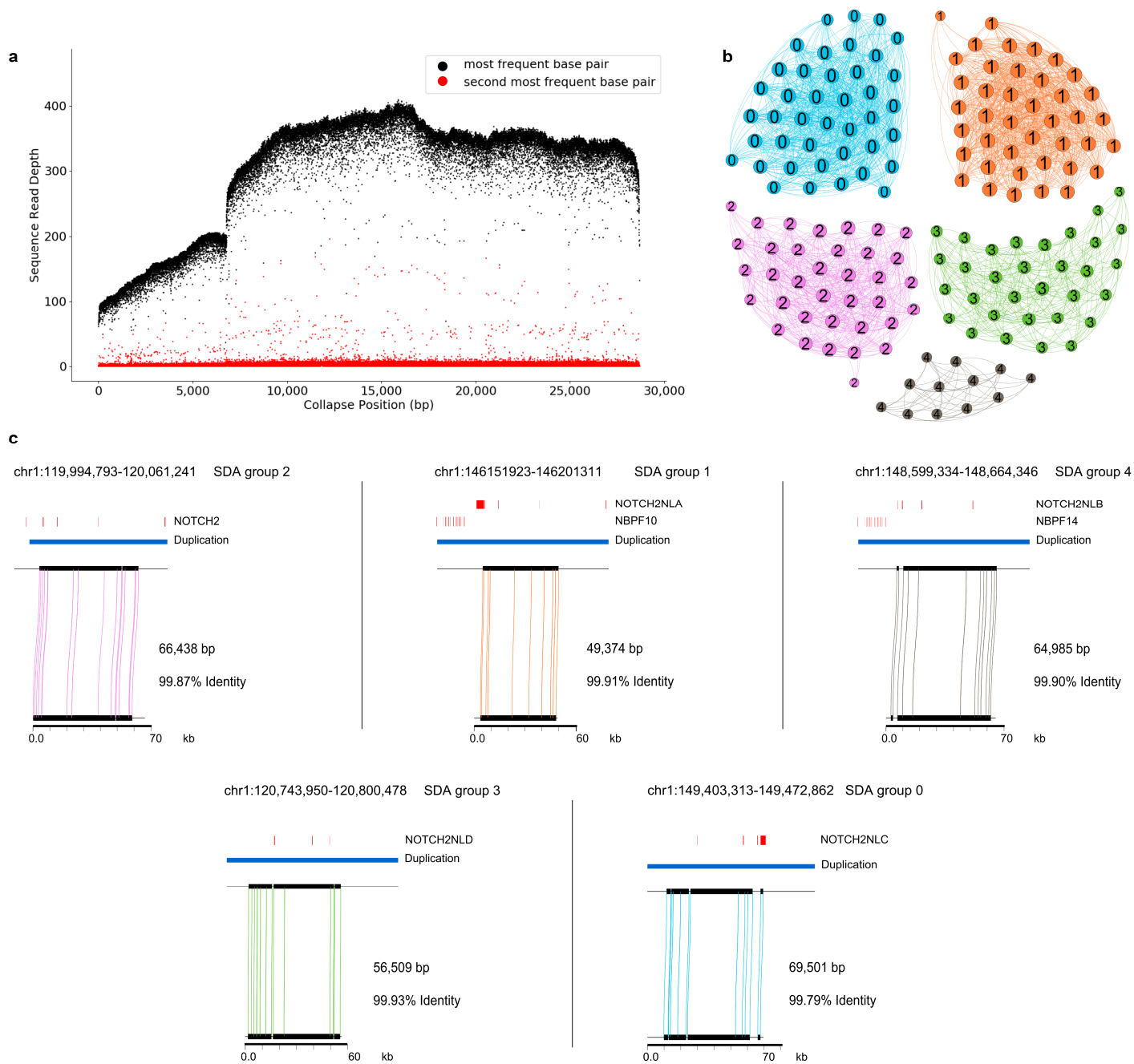
SDs (as a function of percent identity and length) in GRCh38 are marked as resolved (black) if present in the CHM1 assembly, or unresolved (red) if it appears only in the reference. The stacked marginal histograms show the relative number of resolved and unresolved SDs within each bin. Resolved duplications are defined as those mapping with high sequence identity, being completely contained, and extending at least 50 kbp into unique sequence on either side of the duplication block (Methods). See Supplementary Fig. 1 and Supplementary Table 1 for the fraction of unresolved duplications across different genomes, assemblers, and technologies. Note that resolved and unresolved SDs are offset from one another along the y-axis to avoid overlapping. **b**) This plot shows the number of genes that exist within unresolved SDs blocks in the CHM1 assembly versus the maximum percent identity SD within that block.



Supplementary Figure 3

Length of collapsed SDs and SDA assemblies.

Correlation of collapse length and SDA assembly length in **a**) CHM1 ($n = 590$), **b**) CHM13 ($n = 1,440$), and **c**) NA19240 ($n = 1,772$) genome assemblies. In all three assemblies there is a strong correlation (Pearson's correlation) between the length of a collapsed SD and the length of the resulting SDA assembly. SDA is not restricted to assembling duplications less than the maximum read length (like other assemblers), but rather it is restricted by the size of the collapsed duplication.

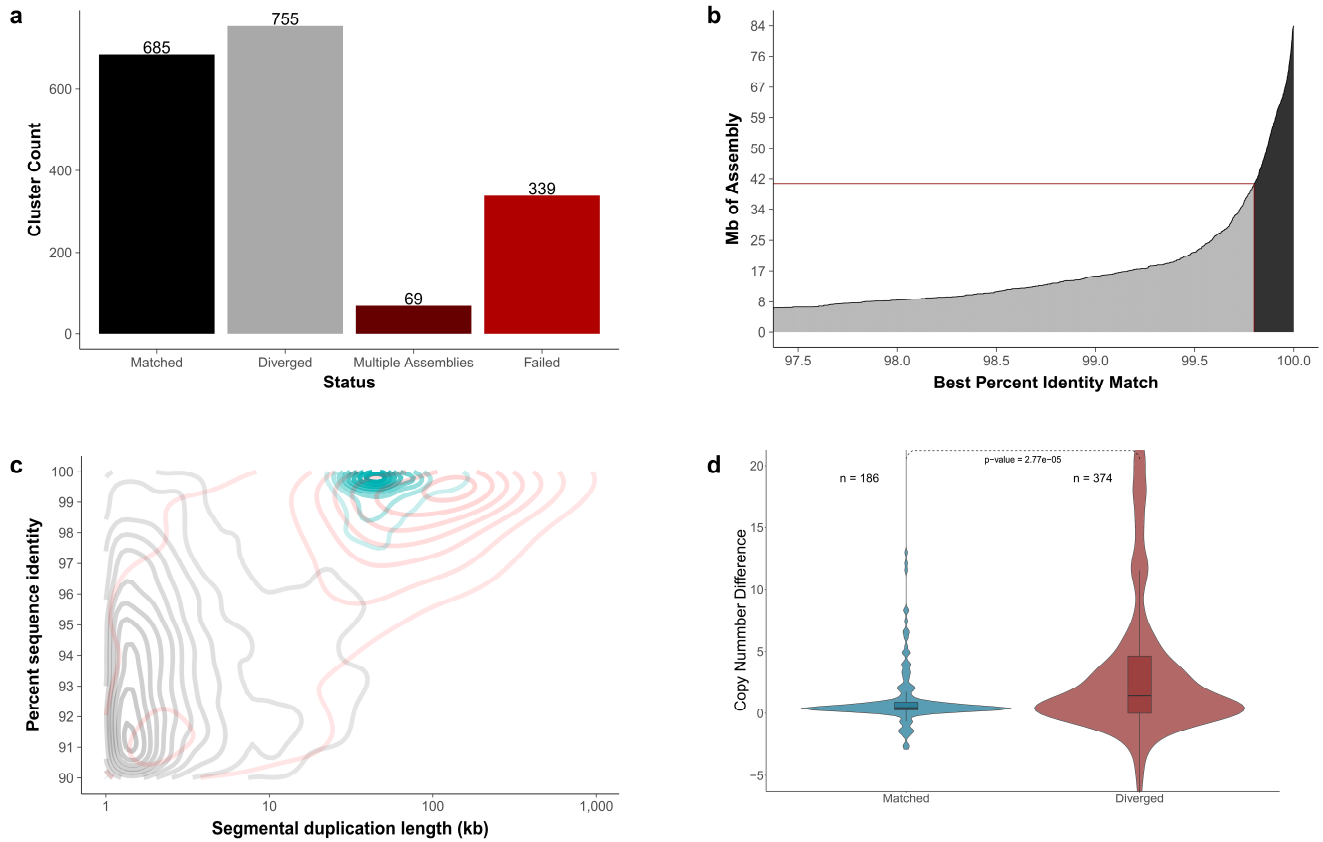


Supplementary Figure 4

Sequence and assembly of *NOTCH2* loci in the CHM1 human genome.

a) A collapsed representation of a portion of the *NOTCH2* loci is shown. Plotted is the read-depth profile over a collapsed representation of *NOTCH2*. Each black dot represents the coverage of the most frequent base pair at that position, while each red dot is the second most frequent. Secondary bases at low frequency represent sequencing error; however, those at high frequency represent PSV candidates. **b)** *NOTCH2* PSV graph resolves the collapse into five potential loci. **c)** The alignment of each SDA contig back to the loci for *NOTCH2* (.NLA/NLB/NLC/NLD) using Miropeats. Our assembled sequence is 99.88% identical over all five loci and >99.995% identical if only mismatched bases are counted as errors.

CHM13 (haploid)

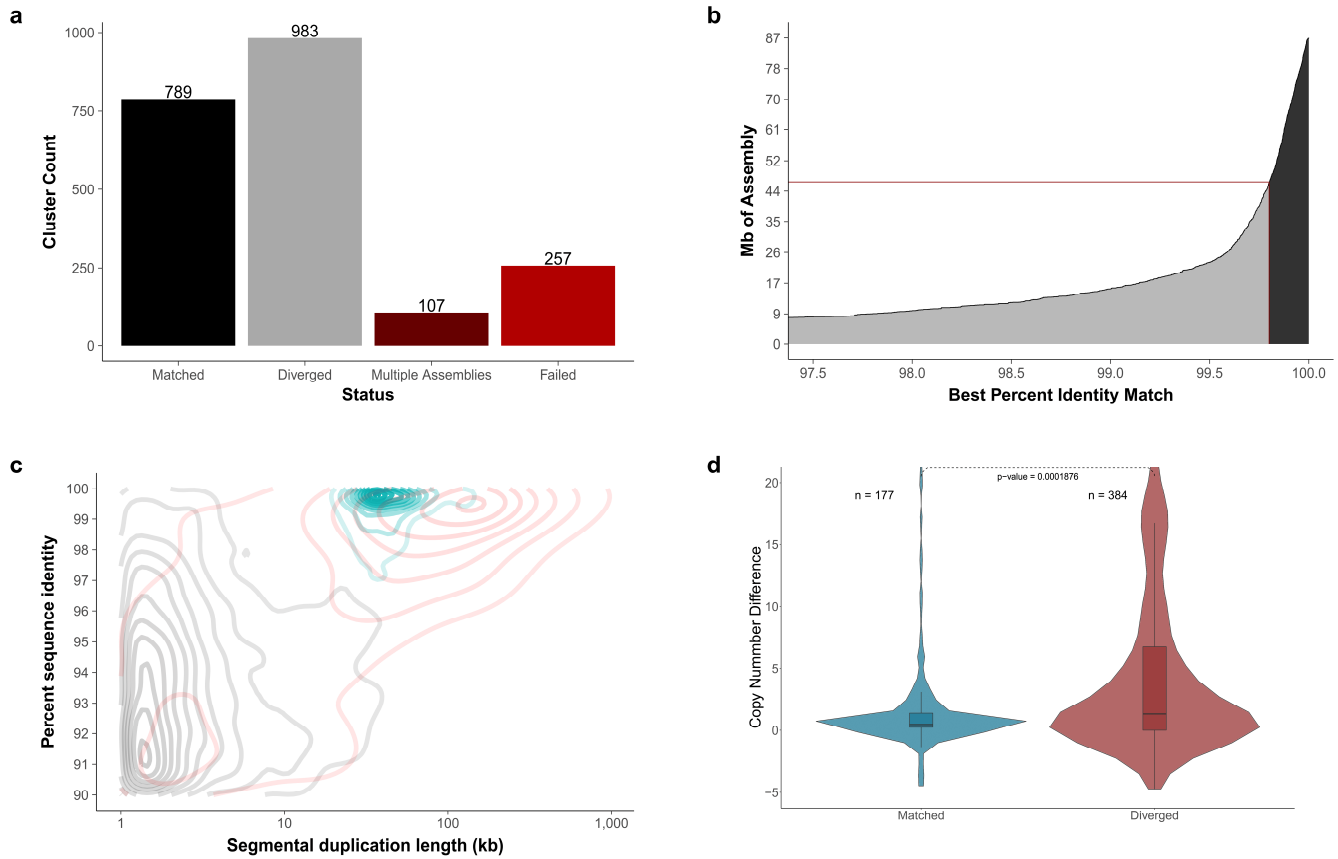


Supplementary Figure 5

SDA results for the CHM13 assembly.

a) SDA analysis of the CHM13 FALCON assembly generates 1,848 PSV clusters. **b)** Cumulative distribution of the assemblies and their percent identity to their best match in the reference. There are 40.4 Mb of diverged assembly (gray) and 43.0 Mb that map to the reference at high identity (black). **c)** A density plot of SDs plotted by length and percent identity. **d)** Copy number difference (CND) between CHM13 and the reference genome (CHM13 copy number – reference genome copy number) comparing $n = 186$ SD regions that match ($>99.8\%$) versus $n = 374$ diverged SD regions ($<99.8\%$ identity). The mean CND of the matched sequence is 1.61 and the mean CND of the diverged sequence is 5.98, indicating that the diverged sequences are much more likely to represent additional duplicate copies that are unrepresented in the reference genome (GRCh38) (two-sided Mann-Whitney test; $P = 2.77 \times 10^{-5}$). 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. (See Fig. 2 for more details.)

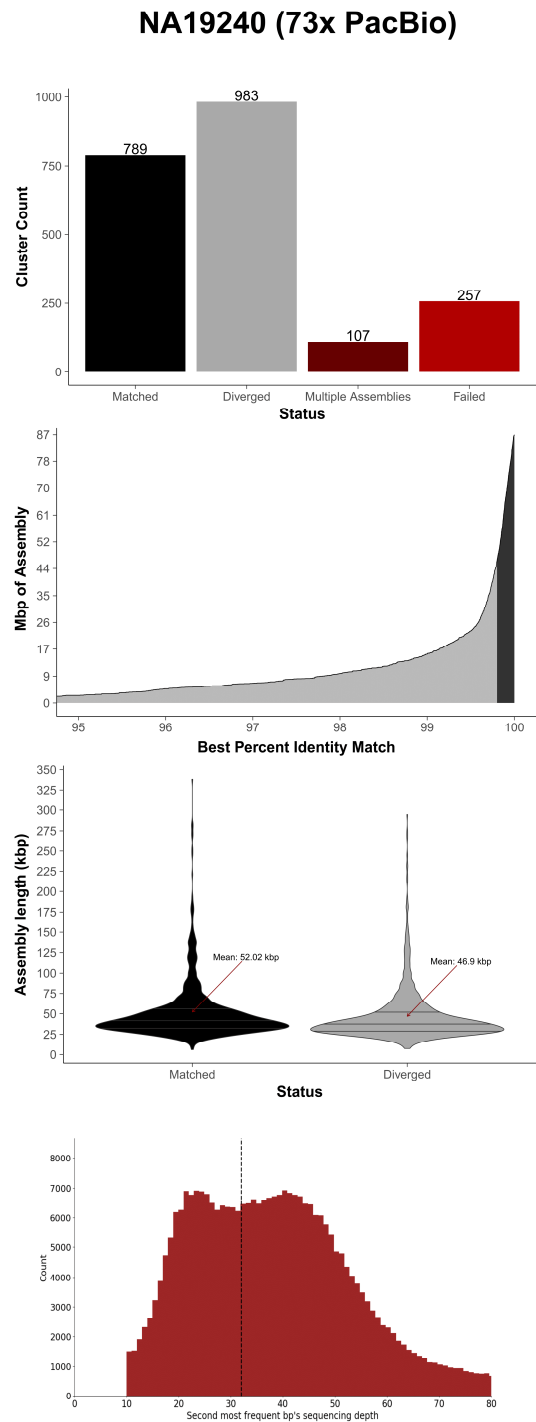
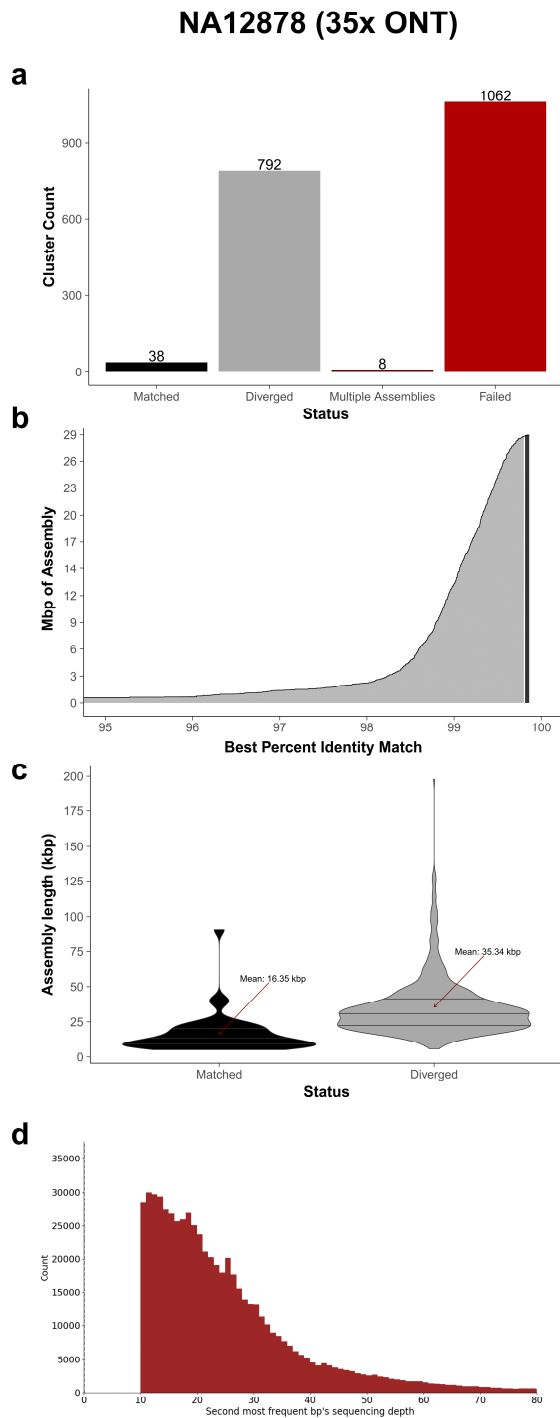
NA19240 (African Yoruban, diploid)



Supplementary Figure 6

SDA results for the NA19240 (African Yoruban) assembly.

a) SDA analysis of the NA19240 FALCON assembly generates 2,136 PSV clusters. **b)** Cumulative distribution of the assemblies and their percent identity to their best match in the reference. There are 46.1 Mb of diverged assembly (gray) and 41.0 Mb that maps to the reference at high identity (black). **c)** A density plot of SDs plotted by length and percent identity. **d)** CND between NA19240 and the reference genome (NA19240 copy number – reference genome copy number) comparing $n = 177$ SD regions that match (>99.8%) versus $n = 384$ diverged SD regions (<99.8% identity). The mean CND of the matched sequence is 4.11 and the mean CND of the diverged sequence is 10.87, indicating that the diverged sequences are much more likely to represent additional duplicate copies that are unrepresented in the reference genome (GRCh38) (two-sided Mann-Whitney test; $P = 1.88 \times 10^{-4}$). 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. (See Fig. 2 for more details.)

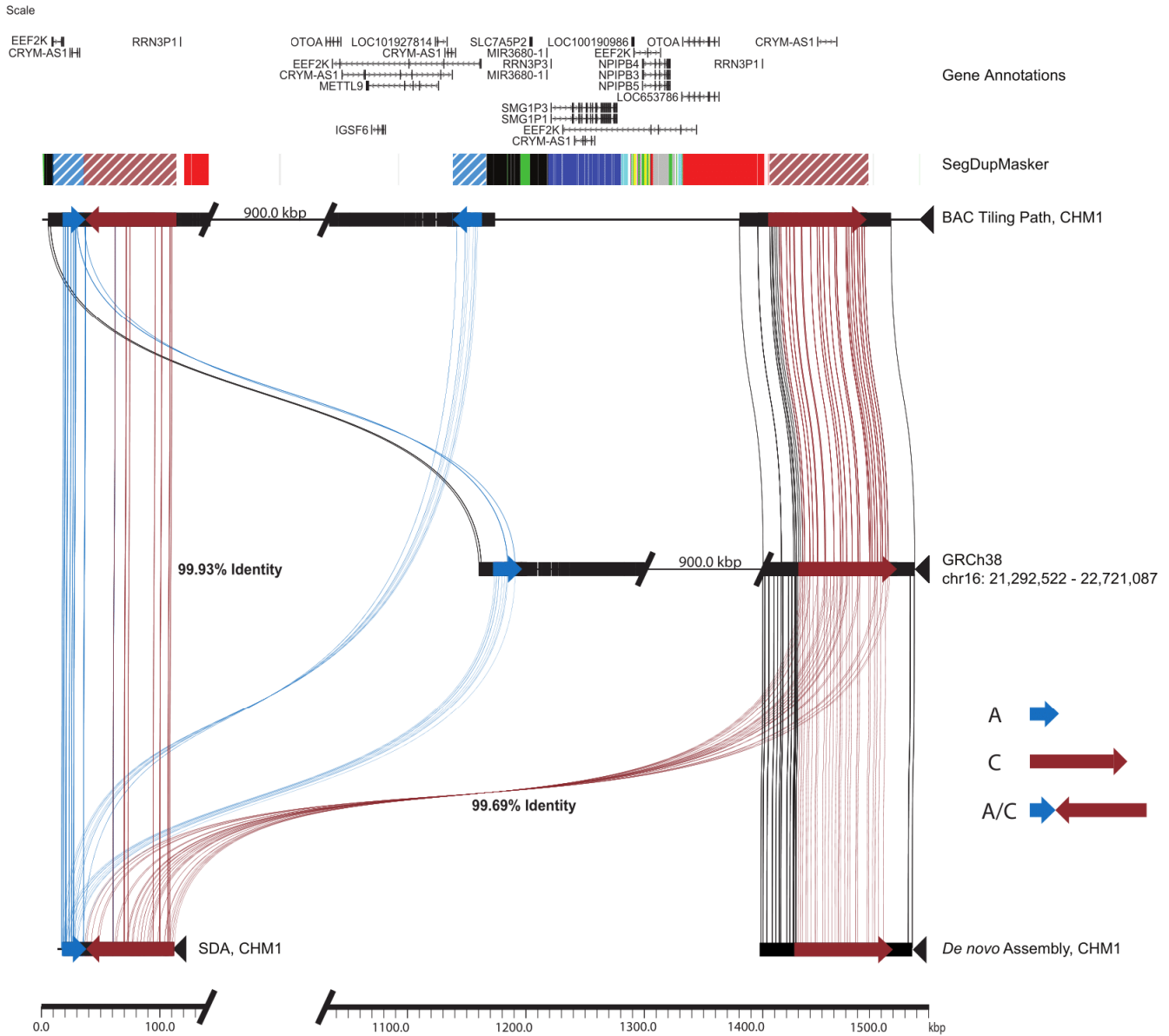


Supplementary Figure 7

Comparison of SDA on ONT versus SMRT data.

The left half of the figure shows the results of SDA applied to the ONT assembly of NA12878; on the right is the PacBio assembly of NA19240. **a**) SDA analysis of the NA12878 assembly generated 38 assemblies that mapped with >99.8% identity (matched) to GRCh38 and 792 mapped with <99.8% sequence identity (diverged). Failed clusters ($n = 1,052$) did not result in an assembly, while multiple assemblies were PSV clusters with more than one contig produced by the Canu assembly. **b**) Cumulative distribution of the assemblies and their percent identity to their best match in the reference. The number of assembly Mb is calculated independently of a

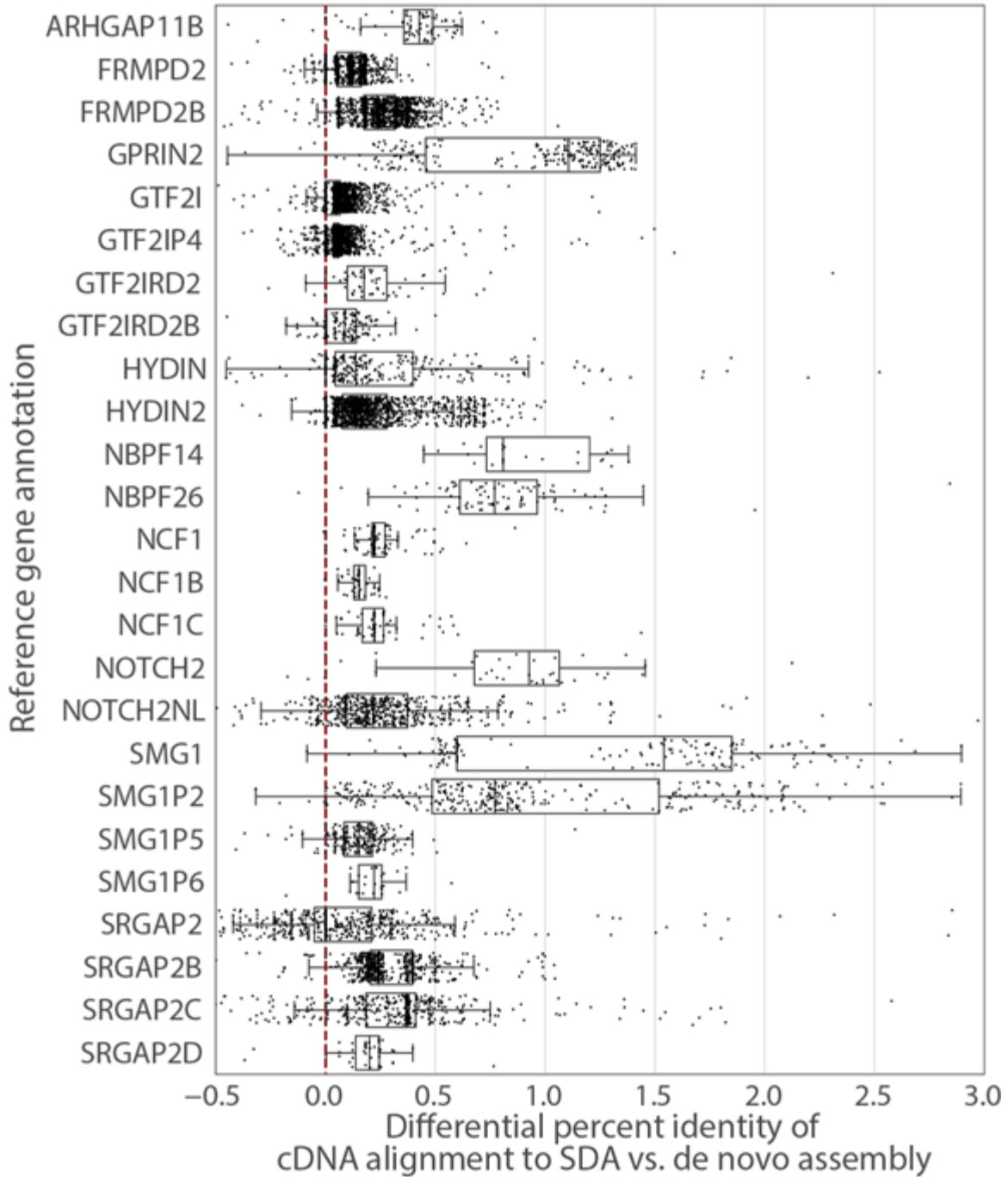
mapping to the reference. **c)** Length distribution of the matched and diverged assemblies (NA12878: matched $n = 38$, diverged $n = 792$; NA19240: matched $n = 789$, diverged $n = 983$). The lines on the violin plots indicate the first and third quartiles as well as the median. **d)** Sequencing read-depth distribution of the second most common SNV across all collapsed regions of SDs.



Supplementary Figure 8

Sequence and assembly of a missing 16p12.1 duplication.

The Miropeats alignments compare a BAC-based tiling path assembly of CHM1 (top line) to the human reference genome (GRCh38) (middle line) to a de novo assembly of CHM1 where SDA was applied (bottom line). The A/C duplication (red blue) proposed by Sudmant et al. that is present in most humans was correctly assembled using SDA and matches at high sequence identity (99.9%) to the BAC-based assembly structure.



Supplementary Figure 9

Mapping differential of transcripts between SDA and de novo CHM13.

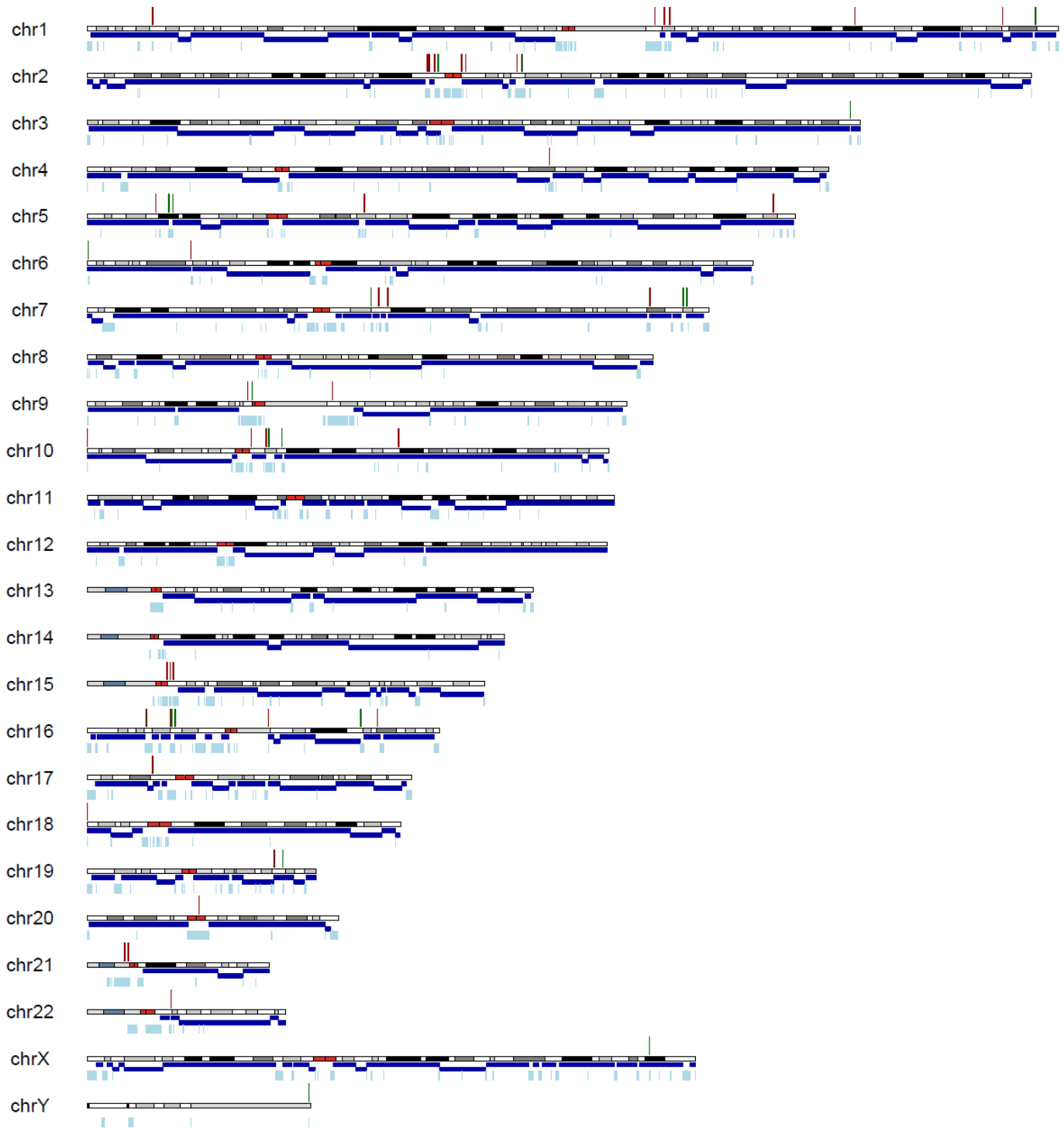
The percent identity differential of the mapping of full-length Iso-Seq transcripts ($n = 14,562$) from human-specific segmental duplications (HSDs) to both the de novo assembly of CHM13 and the SDA results on CHM13 is shown. In total, 11 gene families showed significantly ($P < 0.001$, two-sided Wilcoxon signed-rank test) improved mapping to the SDA-resolved 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.

SDA_GPRIN2B	1	MSSSRPEPGPWAPLSPRLQPLSQSSSSLLGEGREQRPELHKTASSTVWQAQLGEASTRPQAPPEEGNPPESMKPARA	77
SDA_GPRIN2A	1	MSSSRPEPGPWAPLSPRLQPLSQSSSSLLGEGREQRPEVVRKTASSTVWQAQLGEASTRPQAPPEEGNPPESMKPARA	77
hg38_GPRIN2	1	MSSSRPEPGPWAPLSPRLQPLSQSSSSLLGEGREQRPELRKTASSTVWQAQLGEASTRPQAPPEEGNPPESMKPARA	77
SDA_GPRIN2B	78	SGPKARPSAGGHWRSSTVGNVSTMGGGDLCLRLAPSAAMQRSHSDLVRSTQMRGHS GARKASLSCSALGSSPVHRA	154
SDA_GPRIN2A	78	SGPKARPSAGGHWSSSTVGNVSPMGGGDLCLRLAPSAAMQRSHSDLVRSTQMRGHS GARKASLSCSALGSSPVHRA	154
hg38_GPRIN2	78	SGPKARPSAGGHWSSSTVGNVSTMGGGDLCLRLAPSAAMQRSHSDLVRSTQMRGHS GARKASLSCSALGSSPVHRA	154
SDA_GPRIN2B	155	QLQPGGTSQGQGGQAPAGLERDLAPEDETSNSAWMLGASQLSVPPDLGDTTAHSSSAQAEPKAAEQ LATTTC HALPP	231
SDA_GPRIN2A	155	QLQPGGTSQGQGGQAPAGLERDLAPEDETSNSAWMLGASQLSVPPDLWDTTAHSSSAQAEPKAAEQ LATTTC HALPP	231
hg38_GPRIN2	155	QLQPGGTSQGQGGQAPAGLERDLAPEDETSNSAWMLGASQLSVPPDLGDTTAHSSSAQAEPKAAEQ LATTTC HALPP	231
SDA_GPRIN2B	232	AALLCGMREMR---VGAGGCCHALPATGILAFPKLVASVSEGLQAQHGVKIHCRLSGGLPGHSHCCAHLWGPAGLVPE	308
SDA_GPRIN2A	232	ASLLCGMK---VGAGGCCHALPATGILAFPKLVASVSEGLQAQHGVKIHCRLSGGLPGHSHCCAHLWGPAGLVPE	305
hg38_GPRIN2	232	AALLCGMRE---VRAGGCCHALPATGILAFPKLVASVSEGLQAQHGVKIHCRLSGGLPGHSHCCAHLWGPAGLVPE	305
SDA_GPRIN2B	309	PGSRTKDVWTMTSANDLAPAEASPLSAQDAGVQAAPVAACKAVATSPSLEAPAALHVFPEVTLGSSLEEA P SPVRDV	385
SDA_GPRIN2A	306	PGSRTKDVWTMTSANDLAPAEASPLSAQDAGVQAAPVAACKALATSPSLEAPAALHVFPEVTLGSSLEEA P SPVRDV	382
hg38_GPRIN2	306	PGSRTKDVWTMTSANDLAPAEASPLSAQDAGVQAAPVAACKAVATSPSLEAPAALHVFPEVTLGSSLEEA P SPVRDV	382
SDA_GPRIN2B	386	RWDAEGMTWEVYGAAVDPEVLGVAIQKHLEMQFEQLQRAPASEDSLVEGRRGPLRAVMQSLRRPSCCGCSGA APE	461
SDA_GPRIN2A	383	RWDAEGMTWEVYGAAVDLEVLGVAIQKHLEMQFEQLQRAPASEDSLVEGRRGPLRAVMQSLRRPSCCGCSGA APE	458
hg38_GPRIN2	383	RWDAEGMTWEVYGAAVDLEVLGVAIQKHLEMQFEQLQRAPASEDSLVEGRRGPLRAVMQSLRRPSCCGCSGA APE	458

Supplementary Figure 10

Multiple sequence alignment (MSA) between GRCh38 GPRIN2 and SDA GPRIN2A/B.

Shown is the amino acid MSA between the copies of GPRIN2 resolved by SDA and the copy of GPRIN2 in GRCh38. Of the 15 differences in the MSA, 12 are annotated in dbSNP as variants in GPRIN2 when they are in fact differences between GPRIN2A and GPRIN2B. At p.Ser104Gly, p.Arg242Gly, and p.Val375Ala, the reference has the minor allele. Supplementary Table 7 shows the allele frequencies for all variants seen in this alignment.

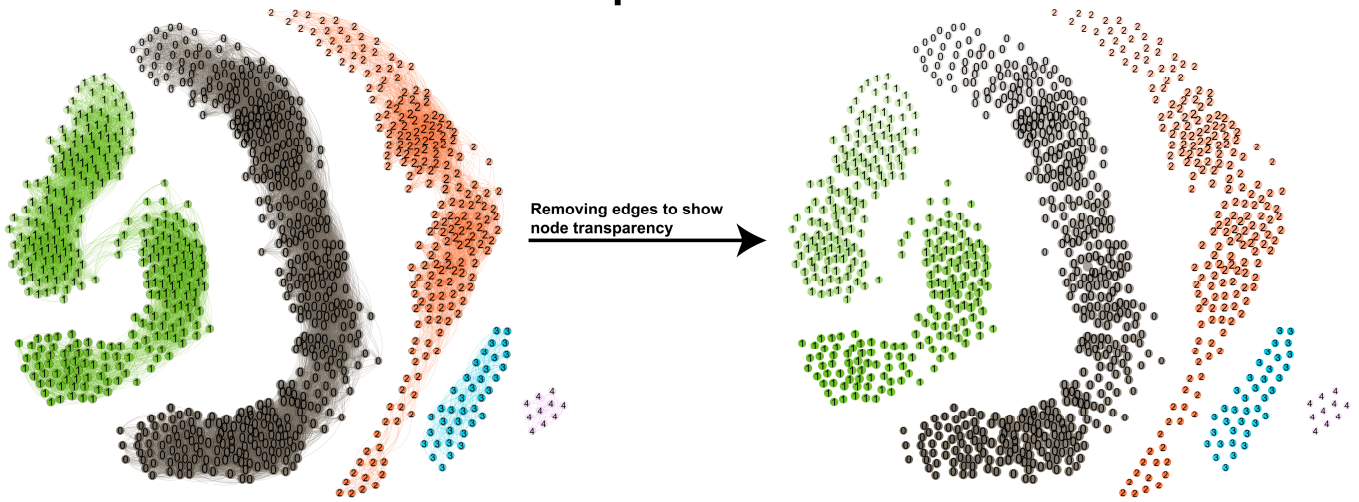


Supplementary Figure 11

CHM1 SDA contigs that overlap with unique sequence.

This ideogram shows where SDA contigs could extend the FALCON assembly. The bottom panel of each chromosome shows the FALCON assembly (contigs > 1 Mb (dark blue), contigs < 1 Mb (light blue)). The top panel shows where SDA contigs with unique overlaps map along the reference (contigs with > 10 kb of overlap (green), contig with < 10 kb (red)).

PSV Graph of SRGAP2



Supplementary Figure 12

PSV graph without attraction edges.

Reproduced above is the PSV graph shown in Fig. 3 for *SRGAP2*. The left-hand side shows the attraction edges used in correlation clustering (CC). On the right-hand side, the edges are removed so that the transparency of the nodes is visible. The opacity of each node scales from 0.25 to 1, with 0.25 reflecting the start position on the contig and 1 representing the final position on the contig.

Supplementary Tables

Table S1. Fraction of resolved SDs in different *de novo* assemblies.

Assembly	Mbp of aligned SD*	% Resolved SD**	Read Coverage
AK1 GCA_001750385.2	113.6	24.3	101
CHM1 GCA_001297185.1	140.9	29.2	61
CHM1 Internally Assembled	116.1	29.4	61
CHM13 GCA_002884485.1	115.5	27.9	73.6
HX1 GCA_001708065.2	125.3	23.2	103
HX1 Canu	125.8	23.3	103
HX1 HERA	131.2	33.3	103
NA 12878 Jain 2018	128.4	32.9	35.4 ONT (4.73 > 50 kbp)
NA 12878 Jain 2018 update	121.6	32.9	38.1 ONT (5.04 > 50 kbp)
Yoruban GCA_001524155.4	123.7	29.3	73x

* Mbp of sequence aligned to the reference over annotated SDs.

** Percent of annotated SDs in the reference that are resolved in the *de novo* assembly.

For an SD to be resolved, the aligned contig must extend 50 kbp past the SD into unique space on both sides.

Table S2. Status of disease-mediating SDs in the FALCON CHM1 assembly.

Disease	Type of Rearrangement	Location	Coordinates	Mbp	OMM	PMid	Resolved in CHM1
Charcot Marie tooth disease type 1A	Interstitial duplication	17p12	chr17:14,446,995-16,048,139	1.5	118220	11584295	No
Hereditary neuropathy with pressure palsies	Deletion	17p12	chr17:14,456,878-16,038,255	1.5	162500	11584295	No
SMS Smith Magenis syndrome	Deletion	17p11.2	chr17:15,112,335-20,380,493	5	182290	11584295	No
Potocki-Lupski syndrome	Interstitial duplication	17p11.2	chr17:16,100,000-22,700,000	5	610883	11584295	No
Neurofibromatosis type1 NFI	Deletion	17q11.2	chr17:30,293,798-32,178,804	1.5	162200	11584295	Yes
Prader-Willi syndrome	Deletion	15q11-15q13	chr15:22,833,353-26,969,005	4	176270	11584295	No
Angelman syndrome	Deletion	15q11-15q13	chr15:23,351,093-27,425,225	4	105830	11584295	Yes
Chromosome 15q11-q13 duplication syndrome	Supernumerary marker chromosome	15q11-15q14	chr15:20,083,333-24,416,666	4	608636	11584295	No
Williams Beuren syndrome	Deletion	7q11.23	chr7:74,529,630-76,070,370	1.6	194050	11584295	No
DiGeorge and velocardiofacial	Deletion	22q11.2	chr22:17,977,414-21,562,880	3	188400	11584295	No
Cat eye syndrome	Supernumerary marker chromosome	22q11.2	chr22:18,500,000-21,999,999	3	115470	11584295	No
X-linked ichthyosis	Deletion	xp22	chrX:6,329,207-8,172,686	1.9	308100	11584295	No
Hemophilia A	Inversion	Xq28	chrX:154,648,851-155,209,658	0.5	306700	11584295	Yes
Male infertility AZFa microdeletion	Deletion	yq11.2	chrY:12,344,706-13,146,789	0.8	415000	11818139	No
Male infertility AZFc microdeletion	Deletion	yq11.2	chrY:11,007,537-14,554,536	3.5	415000	11818139	No

A list of diseases and syndromes caused by large genomic rearrangements as described in Emanuel 2001 and Stankeiwicz 2002, and if they are contiguously assembled past the duplication boundaries in the CHM1 genome assembly.

Table S3. Sequence and assembly of *SRGAP2* and *NOTCH2NL* gene families.

Gene	SDA Group	Status	Percent Identity	GRCh38 Location	Length	Number of PSVs	# PSVs in GRCh38
<i>SRGAP2</i>		0 Resolved	99.96	chr1:206,210,031-206,407,594	197,525		451
<i>SRGAP2C</i>		1 Resolved	99.99	chr1:121,189,099-121,388,229	199,064		299
<i>SRGAP2B</i>		2 Resolved	99.99	chr1:144,893,160-145,088,264	195,041		203
<i>SRGAP2D</i>		3 Resolved	99.97	chr1:143,980,898-144,061,839	67,989		37
<i>SRGAP2D</i>		4 Resolved	99.89	chr1:143,980,898-144,061,839	21,945		10
<i>NOTCH2NLC</i>		0 Resolved	99.79	chr1:149,403,313-149,472,862	69,501		41
<i>NOTCH2NLA</i>		1 Resolved	99.91	chr1:146151923-146201311	49,374		41
<i>NOTCH2</i>		2 Resolved	99.87	chr1:119,994,793-120,061,241	66,438		36
<i>NOTCH2NLD</i>		3 Resolved	99.93	chr1:120,743,950-120,800,478	56,509		32
<i>NOTCH2NLB</i>		4 Resolved	99.9	chr1:148,599,334-148,664,346	64,985		12

The percent identity (GRCh38), contig length, and number of PSVs for four copies of *SRGAP2* and five copies of *NOTCH2NL* are shown.

Sequences were resolved by SDA and correlation clustering.

Table S4. CHORI-17 BAC clone sequences.

Included separately as an Excel file.

Table S5. BAC clone sequence analysis.

Included separately as an Excel file.

Table S6. Gene content analysis.

Included separately as an Excel file.

Table S7. Differences in the multiple sequence alignment of GPRIN2A/B and the reference copy of GPRIN2.

Position	Consequence	hg38_GPRIN	SDA_GPRIN2A	SDA_GPRIN2B	RSID*	Allele Frequency**	Het freq**	Allele Number**
5	p.Arg5His	R	R	H	rs3127817	0.50	1.00	193796
39	p.Leu39Val	L	V	L	rs4926045	0.48	0.96	232084
40	p.Arg40His	R	R	H	rs3127818	0.50	0.99	251024
47	p.Val47Met	V	V	M	rs3127819	0.50	0.99	251292
91	p.Trp91Arg	W	W	R	rs3127820	0.50	1.00	270506
100	p.Thr100Pro	T	P	T	rs7090312	0.48	0.97	260714
104	p.Ser104Gly	S	G	G	rs3127679	0.89	0.22	271564
202	p.Gly202Trp	G	W	G	rs11204658	0.48	0.96	255588
233	p.Ala233Ser	A	S	A	rs11204659	0.48	0.97	260010
239	p.Arg239Lys	R	K	R	rs7895979	0.48	0.97	257192
240	p.Met238_Glu240dup	-	-	MRE	rs11262042	0.50	1.00	30912
242	p.Arg242Gly	R	G	G	rs554090811	0.88	0.23	276860
348	p.Val348Leu	V	L	V	rs4926046	0.36	0.72	245486
375	p.Val375Ala	V	A	A	rs3127822	0.99	0.03	277212
400	p.Leu400Pro	L	L	P	rs3127823	0.50	1.00	275948

*Results from dbSNP.

**Results from gnomAD.

Table S8. Summary of all SDA assemblies from CHM1, CHM13, and NA19240.

Included separately as an Excel file.

Supplementary Note

Percentage of resolved SDs across genomes/assemblers/technologies

Figure S1 and **Table S1** show the fraction of “Resolved” segmental duplications (SDs). Our working definition of resolved is that for an SD to be resolved the assembly must continue into unique sequence on either side of the SD by at least some minimal extension. **Figure S1** shows the fraction of resolved bases as the minimal extension is varied from 0 to 250 kbp. The basic steps of identifying resolved versus unresolved duplications are as follows:

- 1) Map the *de novo* assembly to the human reference using MashMap 2.0 defaults.
- 2) Download the UCSC-annotated SD track and merge overlapping SDs by their maximum percent identity.
- 3) Intersect the *de novo* assembly track with the modified SD track.
- 4) Determine if and by how much the *de novo* assembly extends past SD blocks on either side.
- 5) Mark SDs as resolved or unresolved based on whether the *de novo* assembly extends at least X kbp into unique sequence on either side.
- 6) Plot the percentage of SD bases resolved as a function of the minimal extension into unique sequence past a duplication block.

Currently, there are two Oxford Nanopore Technologies (ONT) ultra-long assemblies of NA12878: one that is recently published (Jain et al., *Nature Biotech*, 2018)¹, and the other an updated assembly from the Philippy lab. The ONT assemblies do outperform the PacBio assemblies; however, its improvement over the different PacBio assemblies is less than 10%. This still leaves the majority of SDs unresolved, motivating and highlighting the importance of our method. All the input assemblies for this analysis were “contig” assemblies and not “scaffolded” assemblies. While there exist scaffolded assemblies for some of these genomes, we decided not to use them in order to make comparisons more consistent.

Application of SDA to NA19240 (diploid)

Using the two haploid (CHM1 and CHM13) and one diploid (NA19240) human genomes, we effectively modeled read depth corresponding to the second most common base pair (i.e., SNV or PSV). Because most paralogous variation is evolutionarily older than allelic variation, it is much more likely to be fixed and, as a result, true PSVs show a different sequence read depth than allelic variation (i.e., CHM1 shows a mode at ~40-fold read depth, consistent with a fixed duplicate copy, **Figure SN1a**). In contrast, a diploid sample that harbors both allelic and paralogous variants shows a clear bimodal distribution (**Figure SN1c**). Thus, to avoid phasing allelic variation we set a minimum depth threshold at the mean coverage minus three standard deviations or half the mean coverage, whichever was greater. This is represented by the black dotted line and corresponds to the trough between the two peaks in **Figure SN1c** (~31-fold). This threshold enriches for true PSVs and prevents most alternate haplotypes from being assembled.

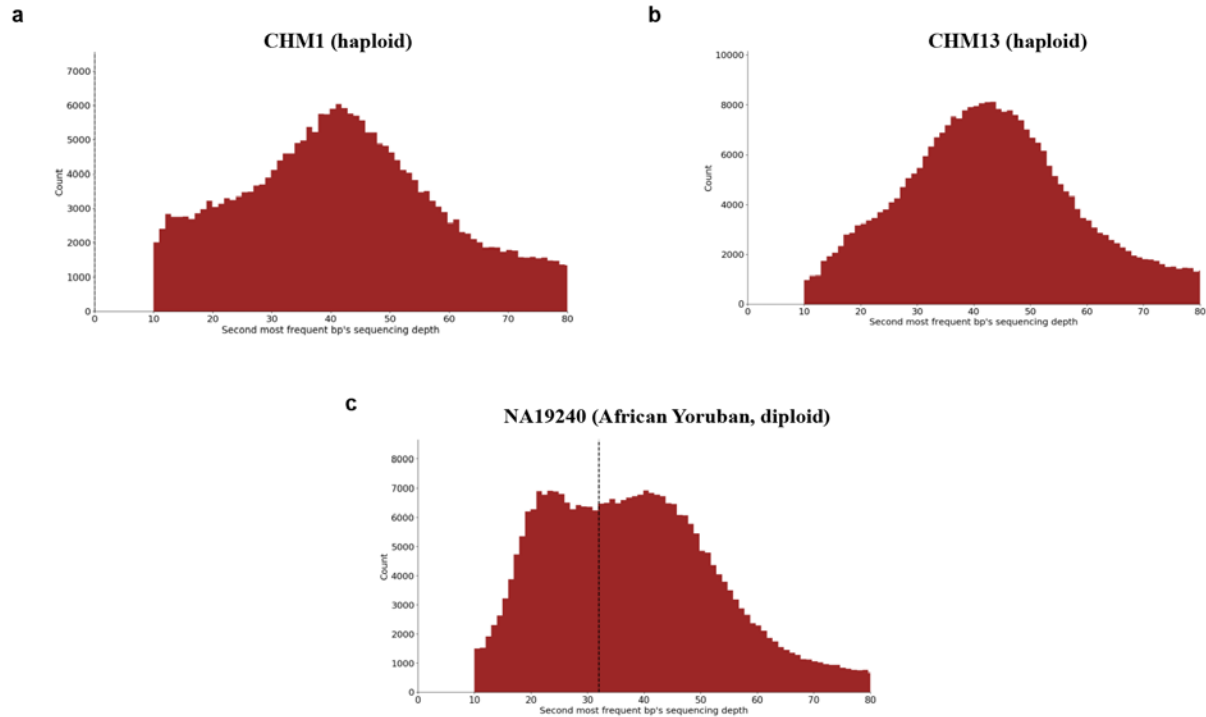


Figure SN1. PSV read-depth distribution. Sequencing read-depth distribution shown for the second most common SNV across all collapsed regions of SDs in **a)** CHM1, **b)** CHM13, and **c)** NA19240 genome assemblies. **a/b)** For CHM1 and CHM13, we consider the distribution with a mode at a read depth of 42-fold to represent putative PSVs. There is a clear peak in SNV frequency around a sequencing read depth of ~42-fold (see Methods). **c)** In the case of NA19240, we observe a bimodal distribution and consider variants with a read depth ~45-fold to once again represent PSVs, while the second mode at read depth of ~23-fold represents possible allelic SNVs. Therefore, we also set a minimum PSV sequencing depth of 32X (black dashed line) for diploid genomes. SNVs with a read depth less than 10-fold sequence coverage are not displayed because they likely represent sequencing error and exist at a much higher frequency.

Please note that the recovered SDs would not be the same for an *in silico* diploid of CHM1 and CHM13 because only PSVs common to both CHM1 and CHM13 would be used for phasing. Thus, the resolution would be of paralogs and not alleles.

Haplotype phasing of duplication regions remains an unaddressed layer of complexity and an area of future investigation. For the diploid genome NA19240, we focused our analysis on the discovery of PSVs occurring at the expected frequency of a duplicated copy and specifically excluded allelic variation by requiring sequence coverage consistent with a unique diploid region of the genome. However, given that paralogous variation can approximate or even become more

identical than allelic variation², it is likely that SDA could be extended to distinguish and assemble haplotypes as well as paralogs. For example, many haplotypes of *HLA* share 90%-99% sequence identity³, but *NOTCH2NL*, which we resolved using SDA, shares up to 99.7% sequence identity among the copies. It may be possible to integrate our SDA method with haplotype-aware assemblers such as FALCON-Unzip⁴, which currently fail to resolve highly identical duplications within human genomes.

Application to ONT data

SDA is compatible with ONT data and we performed an analysis of collapsed SDs present in the ONT assembly of NA12878¹. We identified 365 collapses, a similar number to that identified in the CHM1 PacBio assembly analyzed (283). We present the results compared to PacBio data for NA19240 (**Figure S7**). Overall, the accuracy of the ONT contigs is much lower. There are far more “failed” assemblies because of the lower sequencing coverage. PSVs are more difficult to identify since ONT has more mismatch errors than PacBio. While ONT data offers longer reads, the fundamental problem is its lower accuracy. The total assembly accuracy of the NA19240 assembly (assembled with PacBio) was 99.28%⁵ before Illumina short-read polishing, whereas the assembly accuracy of NA12878 was only 95.20%¹.

We also note that the generation of ultra-long reads of >1 Mbp is not yet common (**Figure SN2**). The longest read reported in Jain et al. 2018 was 882 kbp and reads greater than 500 kbp represent ~1% of the data. Therefore, there is only ~0.05X coverage of 500 kbp reads and ~2.5X coverage of 100 kbp reads, which is not sufficient for proper read correction and assembly of SDs. Finally, the generation of 1 Mbp length molecules is non-trivial and will be limited to a small fraction of samples where high-quality DNA can be prepared.

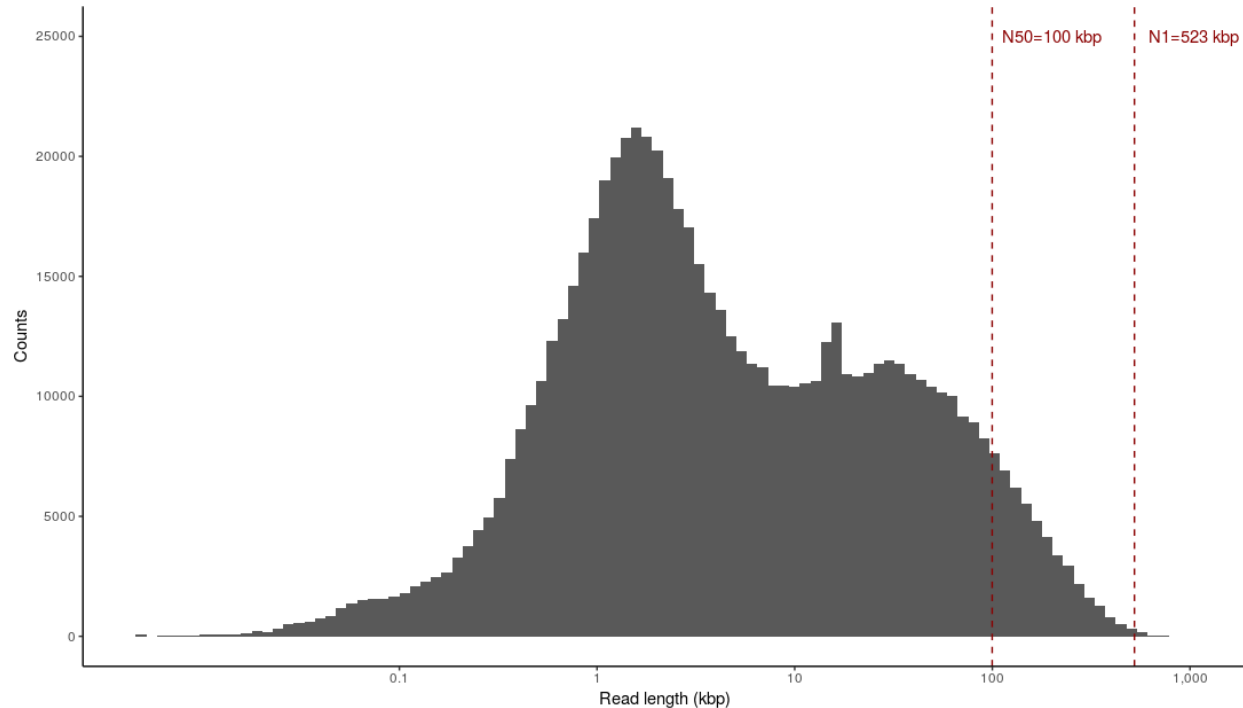


Figure SN2. Distribution of ONT ultra-long reads from NA12878. Data available at <https://github.com/nanopore-wgs-consortium/NA12878/blob/master/Genome.md>. Only reads prepared using the ultra-long protocol are shown.

Integration of SDA contigs into the *de novo* assembly

The majority of our sequence contigs begin and end within SDs and do not transition into unique regions. This is due to the interspersed architecture SDs, which are frequently organized into very large blocks (>500 kbp in size) where structural variation and interlocus gene conversion occur. The latter creates pockets (often >50 kbp) with limited or no sequence divergence. As a result, the resolved sequences effectively represent islands of duplication with no transition into unique sequence. For example, of the 590 assembled sequences from our CHM1 assembly, we found that only 131 (22.2%) can be anchored to a unique sequence (**Figure S11**). Of these 131, only 28 overlap with unique sequence for at least 10 kbp. These contigs can be used to extend the original FALCON assembly confidently. In total, there is 583 kbp of sequence from SDA contigs that overlap with unique sequence in the genome.

We note that even though our “orphan” assemblies are small, they are comparable in length to the unplaced contigs in GRCh38 (**Figure SN3**). More importantly, they are high quality and contiguous, making them useful for downstream genomic analyses. This is in sharp contrast to the small contigs typically generated by WGS, which represent collapsed and fragmented mistakes of the assembly process of little biological utility. In the past, studies of duplication typically relied on generating similar high-quality sequence from BAC and fosmid clone inserts—a lengthy and costly prospect⁶⁻⁹. Here, we have generated the equivalent of 500-1,500

high-quality contigs of similar size per genome that otherwise would have been lost. The average size of these high-quality contigs (54 kbp) is sufficient for improved gene annotation.

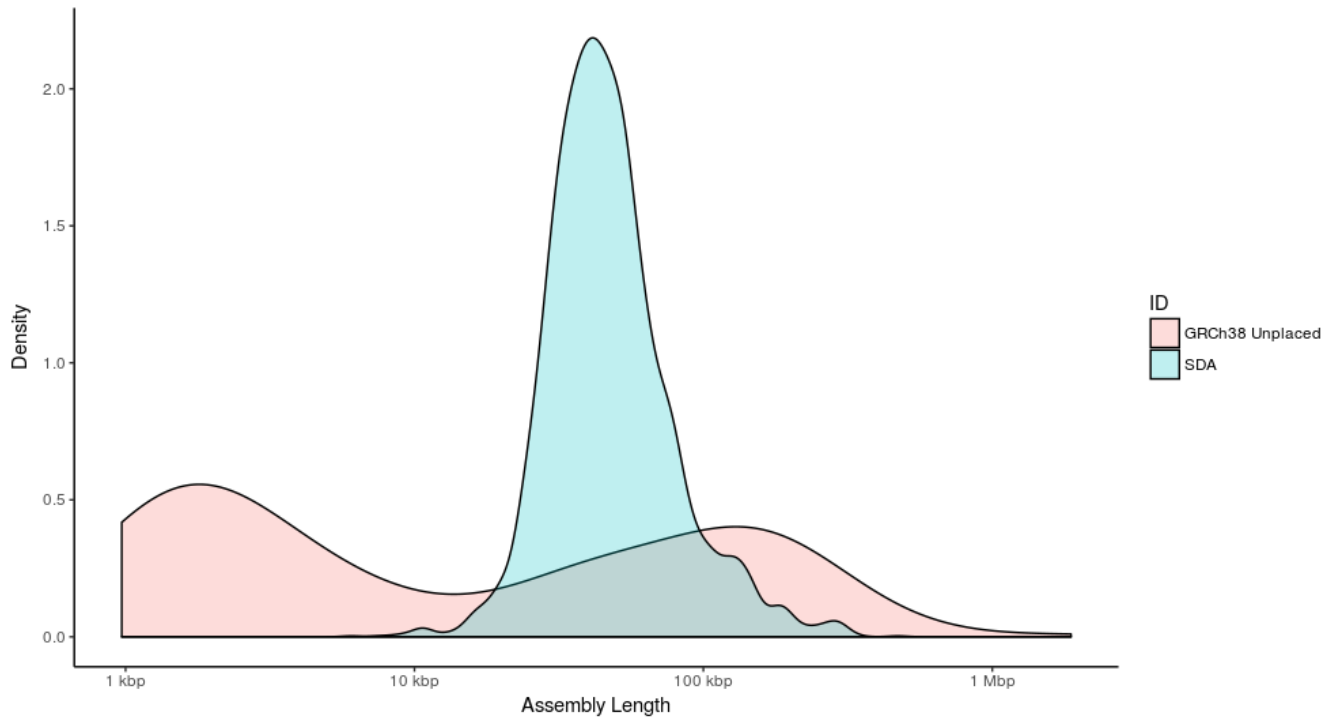


Figure SN3. Length distribution of orphaned contigs. Density plot of the assembly lengths of unplaced contigs in GRCh38 versus the contigs produced by SDA across CHM1, CHM13, and NA19240. The mean and median lengths for GRCh38 unplaced contigs are 67.8 kbp and 6.5 kbp, respectively, and the mean and median lengths for the SDA contigs are 54.3 kbp and 44.9 kbp, respectively.

In the event that others would like to use SDA within their whole-genome assembly, SDA creates a partitioned list of reads with assignments to SDA contigs that can be processed by other assemblers post hoc. Specifically, one would first run the assembler to produce contigs and resolve duplications with SDA. All reads processed by SDA can be processed as a data structure of tuples (read, duplication index). Assembly would be executed again. Given two reads that overlap, one could check if they are assigned a duplication index, and if so, whether they have the same duplication index.

Improvements in SDA

The underlying PSV correlation clustering (CC) algorithm was presented as part of a RECOMB submission (Chaisson, 2017). This paper showed proof-of-principle of the algorithm and was based only on simulated data. Here, we develop the SDA method, including the computational infrastructure, apply it to real long-read whole-genome sequence data, and perform a detailed analysis of the results. This required several developments and improvements.

Specific improvements to the RECOMB algorithm include:

1. An optimization of the random sampling procedure to select the best sampled partition among many runs. **Figure SN4** shows how the CC score was reduced per random sampling iteration for PSV graphs where multiple iterations had an impact.
2. The graph used by CC was modified to account for sparse stretches of PSVs that are more commonly seen in real data than the simulated data in the RECOMB paper. Previously, repulsion edges were made when two PSVs had overlapping reads, but there was no positive edge. However, this was problematic because PSVs that just missed the threshold to have an attraction edge would automatically become a repulsion edge even though there was moderate evidence for an attraction edge. Here, we modified the definition of a repulsion edge to be two PSVs without any significant evidence for an attraction edge. This improved performance (fewer edges) and the results (less incorrect fracturing of paralogs).
3. During the process of cluster formation, all existing pairwise clusters are intermittently assessed to determine if the merging of any pair of them would improve the overall CC score. This development improved the performance of paralog separation, particularly for very long collapsed duplications.
4. In addition to these algorithmic developments, we have modified SDA so that it can work with either ONT or PacBio long-read data as input and have provided options such that different assemblers (e.g., Canu, miniasm and wtdbg) can be applied to resolve paralogs based on the partitioned reads. From a comparison of assembler performance, see **Table SN1**.

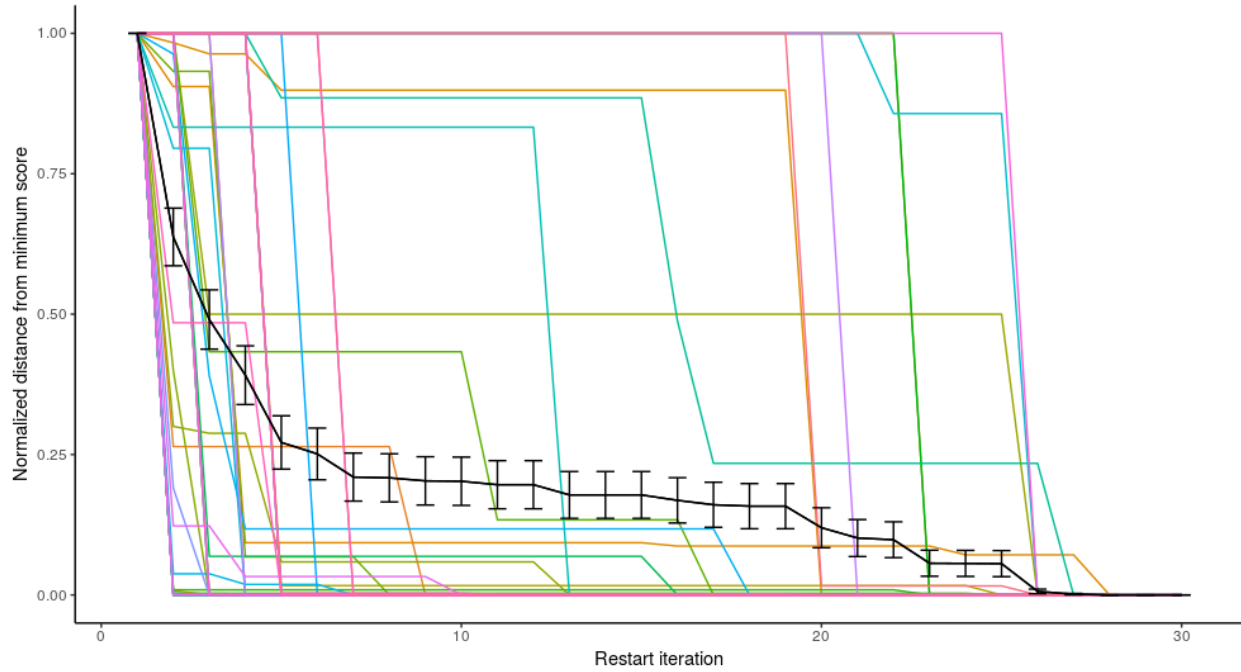


Figure SN4. Random restarts improve CC. This figure shows the normalized decrease in the CC score over $n=76$ different collapses in CHM13. Each line shows the minimum CC score observed at a given restart iteration. The black line shows the mean CC score with standard error bars.

Table SN1. Comparison of Canu, miniasm, and wtdbg on 10 individual collapses.

Region of Collapse	CC group	Canu (kbp)	Canu % ID	miniasm (kbp)	miniasm % ID	wtdbg (kbp)	wtdbg % ID
SRGAP2	0	193.3	99.90	176.1	86.14	200.2	97.03
SRGAP2	1	40.2	99.47	NA	NA	55.7	94.50
SRGAP2	2	42	99.43	54.3	88.92	48.1	96.20
SRGAP2	3	215.7	99.88	190.5	89.26	205.9	98.01
SRGAP2	4	195.2	99.85	101.2	87.34	204.1	97.57
SRGAP2	5	42	99.40	NA	NA	46.1	95.73
SRGAP2	6	50.6	99.09	52.5	89.07	50.3	96.94
ROCK1	0	78	98.85	69.4	86.20	77.3	96.05
ROCK1	1	101.1	97.67	NA	NA	100.6	93.54
ROCK1	2	54.1	98.60	NA	NA	60.4	94.88
ROCK1	3	37.6	99.73	NA	NA	53.8	94.50
NPY4	0	111.3	99.93	72.3	85.45	119.6	97.46
NPY4	1	50.5	96.91	27.9	84.47	45.7	93.54
NPY4	2	204.6	99.67	184.5	89.21	209	98.02
NPY4	3	50	99.18	NA	NA	50.2	95.51
NPY4	4	47.9	99.47	NA	NA	58.7	93.71
NOTCH2	0	194.4	99.86	74	88.23	198.3	98.35
NOTCH2	1	58.7	99.88	56	86.69	70.6	96.46
NOTCH2	2	102.6	99.72	105.5	89.25	109	97.83
NOTCH2	3	93.5	99.67	98.4	88.16	102	96.80
NOTCH2	4	75.7	99.82	NA	NA	80.3	97.47
NOTCH2	5	69.5	99.73	68.5	89.86	71.8	97.88
NCF1	0	98.6	99.90	NA	NA	111.5	94.99
NCF1	1	99.2	99.87	111.1	88.58	115	96.85
NCF1	2	111.9	99.84	91.5	87.56	123.2	92.72
NAIP	0	70.10	99.73	73.50	88.34	71.8	97.37
NAIP	1	98.5	99.28	67.9	90.20	110.3	96.76
NAIP	2	51.2	99.57	NA	NA	60.1	94.94
NAIP	3	44.7	99.64	NA	NA	53.1	93.06
NAIP	4	51.4	99.39	NA	NA	62.2	95.81
NAIP	5	86.1	99.31	NA	NA	97.2	93.80
NAIP	6	57.8	99.51	65.1	88.61	78.7	96.29
HYDIN2	0	295.30	99.85	252.90	86.52	294.8	98.71
HYDIN2	1	NA	NA	NA	NA	46.4	97.03
HYDIN2	2	248.20	99.90	211.10	88.01	252.9	98.17
HYDIN2	3	115.20	99.70	78.20	88.24	116.3	96.33
HYDIN2	4	NA	NA	NA	NA	30.9	96.34
GTF2H2	0	96.20	80.62	95.50	89.98	111.6	80.75
GTF2H2	1	71.00	99.45	57.30	89.10	70.9	97.03
FRMPD2	0	123.60	99.76	91.30	89.43	123.3	98.70
FRMPD2	1	110.90	99.40	95.00	88.74	114.3	97.84
FCGR	0	111.90	98.82	82.30	87.59	83.6	96.11
FCGR	1	61.20	99.49	55.80	86.61	68.9	96.52

All percent identity calculations were done before error correction with Quiver.

Assembling collapsed regions using Canu

We assessed the effects of various parameter adjustments within Canu to see if SDs could be resolved without SDA. We specifically selected 10 regions of collapse and tried many parameter combinations using Canu and compared the results to our SDA. Results are summarized in **Table SN2**. Assembling the individual collapses produced multiple paralogs in most cases; however, in all but one case, SDA was able to resolve more of the paralogs than any of the Canu assemblies, regardless of parameters. We found these two parameters to be essential to having any success in creating paralogs: corOutCoverage=300 and corMhapSensitivity=high. Setting corOutCoverage much higher than for the whole genomic coverage forces all reads to be corrected, similarly setting corMhapSensitivity to high ensures that the best overlaps are found. Both of these parameters are computationally impractical for whole-genome assemblies. Finally,

we varied the corMaxEvidenceErate=[0.15, 0.25, 0.35, 0.45, 0.55] parameter to generate the ranges shown in **Table SN2**. The corMaxEvidenceErate controls the maximum amount of error that can exist between two reads for them to be overlapped in the read correction step. Increasing this value generally increased the amount of assembled sequence but decreased the quality of the assembly.

Table SN2. Comparison of SDA and parameterized *de novo* assemblies of 10 individual collapses.

Region of Collapse	SDA assembly size (kbp)	SDA %ID	De novo assembly sizes (kbp)*	De novo % IDs*	% increase in bases by using SDA*
SRGAP2	679.5	99.62	334.5 - 366.6	99.51 - 99.54	85.36 - 103.14
ROCK1	299.8	98.22	119.2 - 133.4	97.83 - 99.23	124.75 - 151.51
NPY4	493.2	99.25	224.7 - 276.8	99.1 - 99.14	78.16 - 119.45
NOTCH2	568	99.79	398.7 - 471.5	99.53 - 99.63	20.49 - 42.49
NAIP	345.7	99.47	234.1 - 323.9	99.01 - 99.48	6.75 - 47.7
HYDIN2	626.4	99.84	419.1 - 433.1	99.57 - 99.61	44.62 - 49.46
GTF2H2	122.1	99.49	62.6 - 74.9	99.36 - 99.51	63.11 - 95.12
FRMPD2	234.1	99.59	146.3 - 182.8	99.59 - 99.63	28.05 - 60.04
FCCR	128.1	99.08	97.6 - 166.3	99.21 - 99.27	-22.99 - 31.2

* Ranges reflect the minimum and maximum result from different *de novo* runs of Canu on the collapse.

Sequence divergence required for SDA

SDA is able to resolve large human-specific duplication events with less than 0.5% sequence divergence, see *NOTCH2NL* and *SRGAP2* (**Table S3, Figures 3 and S4**). However, there were events, such as the duplication surrounding *BOLA2*, with stretches of 50 kbp of identical sequence we were unable to resolve. Based on our results, we would argue that reads with 10-15% error are sufficient to resolve duplications that are less than 0.5% diverged, as long as the reads have random errors and there is sufficient coverage (>60X).

To further examine the required sequence divergence between duplications to resolve them with SDA, we aligned and determined the percent identity between all the SDA contigs that we generated for CHM1 (**Figure SN5**). At 99.5% sequence identity, the distribution drops off precipitously suggesting a limit. Additionally, almost no sequences are 99.9% identical indicating that 0.1% probably reflects an upper bound of what SDA is able to resolve even in ideal cases.

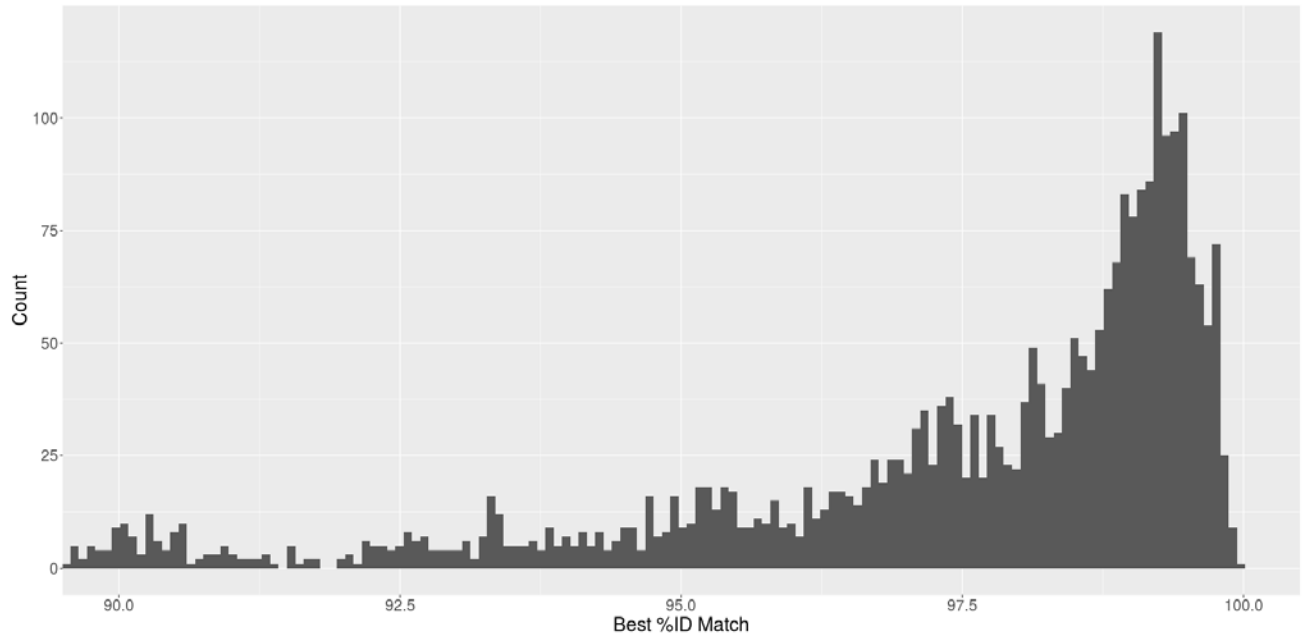


Figure SN5. Percent identity between SDA contigs from CHM1. This figure shows the highest percent identity alignment between all pairs of SDA contigs from CHM1.

Ultra-long ONT as orthogonal support for SDA contigs

We investigated the ability of ultra-long ONT reads to provide orthogonal support for the existence of our SDA contigs. To overcome the high error rate of the individual ONT reads, we focused on identifying matches between the PSVs identified in our contigs and the ONT reads. We required that at least 65% of the PSVs expected to be present in the overlap and that the log likelihood ratio between probability that the observed PSVs were real versus sequencing error to be greater than five (**Figure SN6**). When we did this, we identified 1,932 ONT alignments between the ultra-long reads and our SDA contigs. On average, an ONT read mapped 1.19 times to 641 of 1,184 tested SDA contigs (54%) providing orthogonal support for these results.

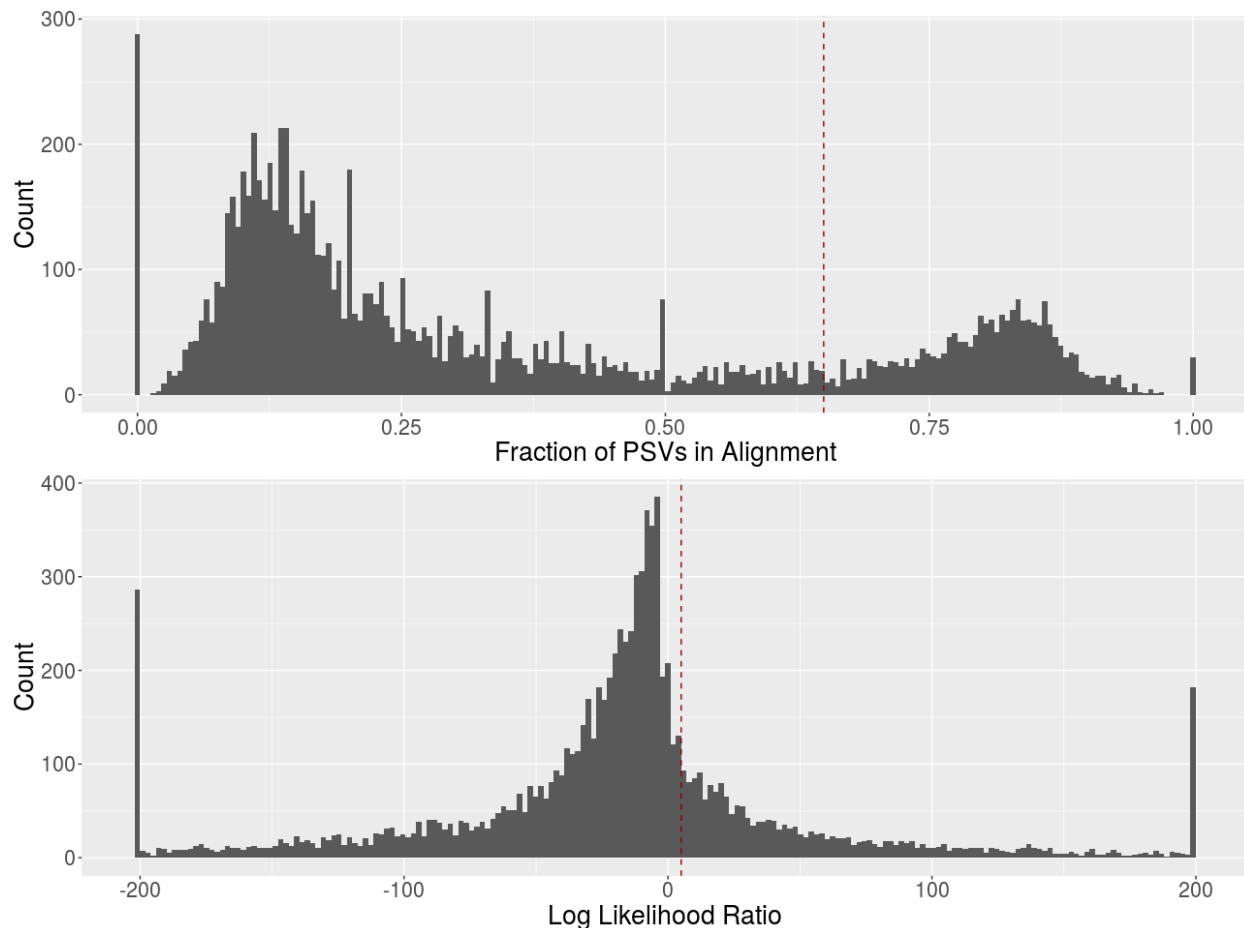


Figure SN6. PSV thresholds for determining correct ONT alignments. Each histogram shows the distribution of alignments of ONT reads to SDA contigs; the lines in red mark the thresholds used for filtering valid alignments. The first plot shows the distribution of the fraction of the expected number of PSVs in the alignment. The second plot shows the distribution of the log likelihood ratio between the probability that the observed PSVs are real or sequencing error.

Additional information on SDA results

Information about length, PSVs, and mapping location in GRCh38 can be found for all the SDA contigs generated in **Table S8**.

When the collapsed sequences in CHM13 (24.3 Mbp) and NA19240 (22.6 Mbp) are mapped back to the reference, they represent 86.6 and 82.4 Mbp of sequence, respectively. Additionally, 73.1 (84.4%) and 64.4 (78.2%) Mbp of the mapped collapsed sequence overlaps with unresolved SDs. Approximately 52% (755/1,440) of CHM13 and 55% (973/1,772) of the African genome assemblies were diverged (<99.8% sequence identity) when compared to the reference genome. All of this is consistent with our results in CHM1 (**Figures 2, S5, and S6**).

BAC analysis with CHM1

In the main text we assert that we expect 37.4% of our BAC clones to validate. This is based on the alignment of 1,253 CHM1 BAC clones back to the reference genome where we found that they represent 65.7 Mbp, or 37.4% of the 175.5 Mbp of SDs annotated in GRCh38.

Accession numbers for all BACs used to validate CHM1 SDA contigs can be found in **Table S4**.

References

1. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, (2018).
2. Schmutz, J. *et al.* The DNA sequence and comparative analysis of human chromosome 5. *Nature* **431**, 268–274 (2004).
3. Buhler, S. & Sanchez-Mazas, A. HLA DNA sequence variation among human populations: Molecular signatures of demographic and selective events. *PLoS One* **6**, (2011).
4. Chin, C.-S. *et al.* Phased diploid genome assembly with Single Molecule Real-Time Sequencing. *Nat. Methods* **13**, 056887 (2016).
5. Steinberg, K. M. *et al.* High-quality assembly of an individual of yoruban descent. *bioRxiv* 067447 (2016). doi:10.1101/067447
6. Dennis, M. Y. *et al.* Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**, 912–922 (2012).
7. Dennis, M. Y. *et al.* The evolution and population diversity of human-specific segmental duplications. *Nat. Ecol. Evol.* **1**, 0069 (2017).
8. Dougherty, M. L. *et al.* The birth of a human-specific neural gene by incomplete duplication and gene fusion. *Genome Biol.* **18**, 1–16 (2017).
9. Karakoc, E. *et al.* Detection of structural variants and indels within exome data. *Nat. Methods* **9**, 176–178 (2012).