doi:10.1038/nature13907

Table of Contents

I. DNA sample preparation and sequencing

DNA was isolated from a well-studied female fibroblast cell line derived from a complete hydatidiform mole (CHM1htert) provided by Dr. Urvashi Surti (University of Pittsburgh). Complete hydatidiform moles retain only a single set of homologous chromosomes due to fertilization of an enucleated egg by a sperm and therefore represent a functionally haploid equivalent of the human genome lacking allelic variation. CHM1htert fibroblast cells were harvested at 70-80% confluency (~3x10^6 cells) and isolated using Gentra Puregene Cell Kit (P/N: 158767) with eluted DNA stored at 4°C overnight for 2 days to resuspend the DNA pellet. DNA was isolated and two genomic libraries were prepared for DNA sequencing.

Supplementary Table 1. Sequencing statistics.

Read type	Number of reads	Read length	Total (coverage)
Illumina, all	1,290,534,230	101	41.5
Illumina, mapped	1,265,426,328	101	40.7
PacBio, all subread	20,865,849	7,307 (mean)	48.6
PacBio, mapped subread	19,571,994	5,860 (mean)	36.6

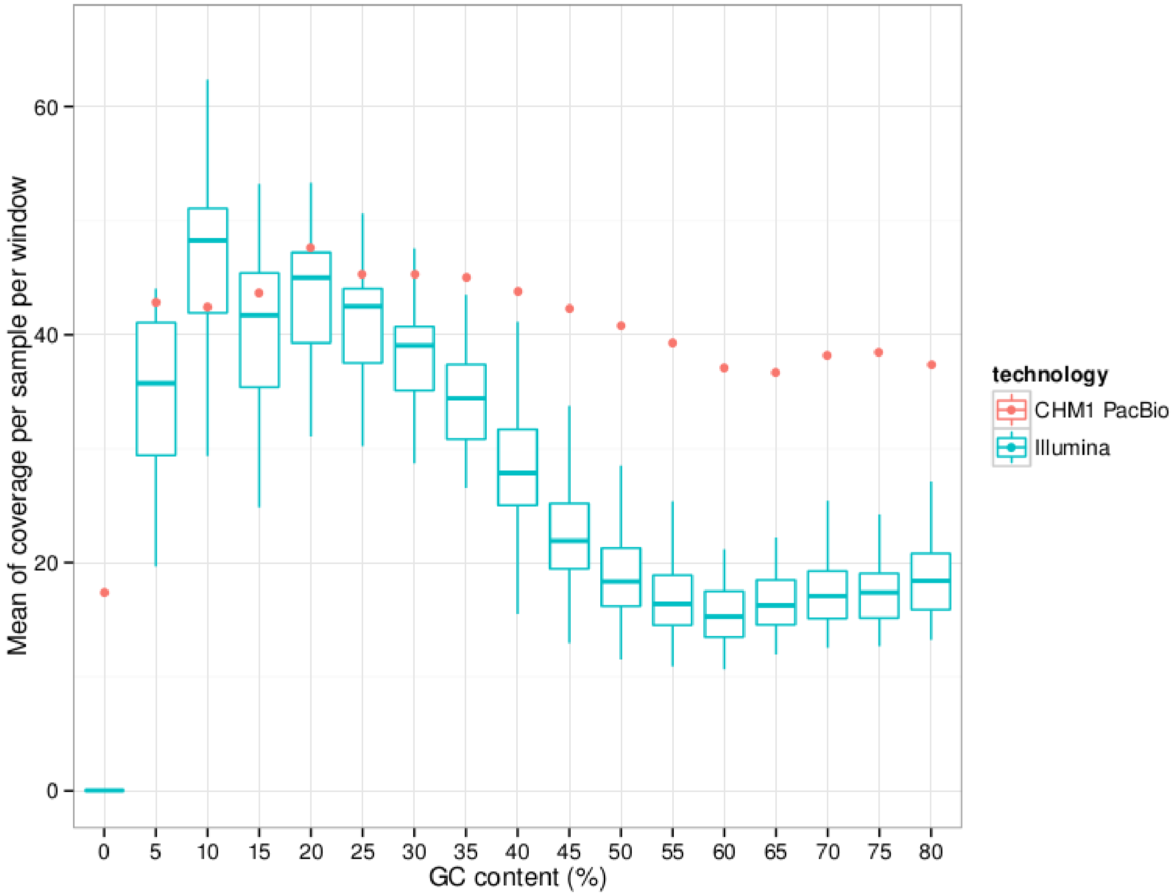
PacBio: We prepared 20 kbp and 30 kbp DNA fragment libraries, size-selected with the BluePippin^{TM} system from Sage Science, and sequenced with 3-hour movies using the PacBio RSII instrument model with P5 polymerase binding and C3 chemistry kits (P5C3). A total of 243 single-molecule, real-time (SMRT) cells were processed yielding 41-fold whole-genome sequence (WGS) data (Supplementary Table 1). All sequence data has been released within the "short" read archive NCBI GenBank accession SRX533609 and may also be accessed as part of all the PacBio datasets via this link: http://www.ncbi.nlm.nih.gov/sra/?term=SRP040522. The location of the raw data is given in Supplementary Table 2.

Supplementary Table 2. Archive of full Hierarchical Data Format PacBio files.

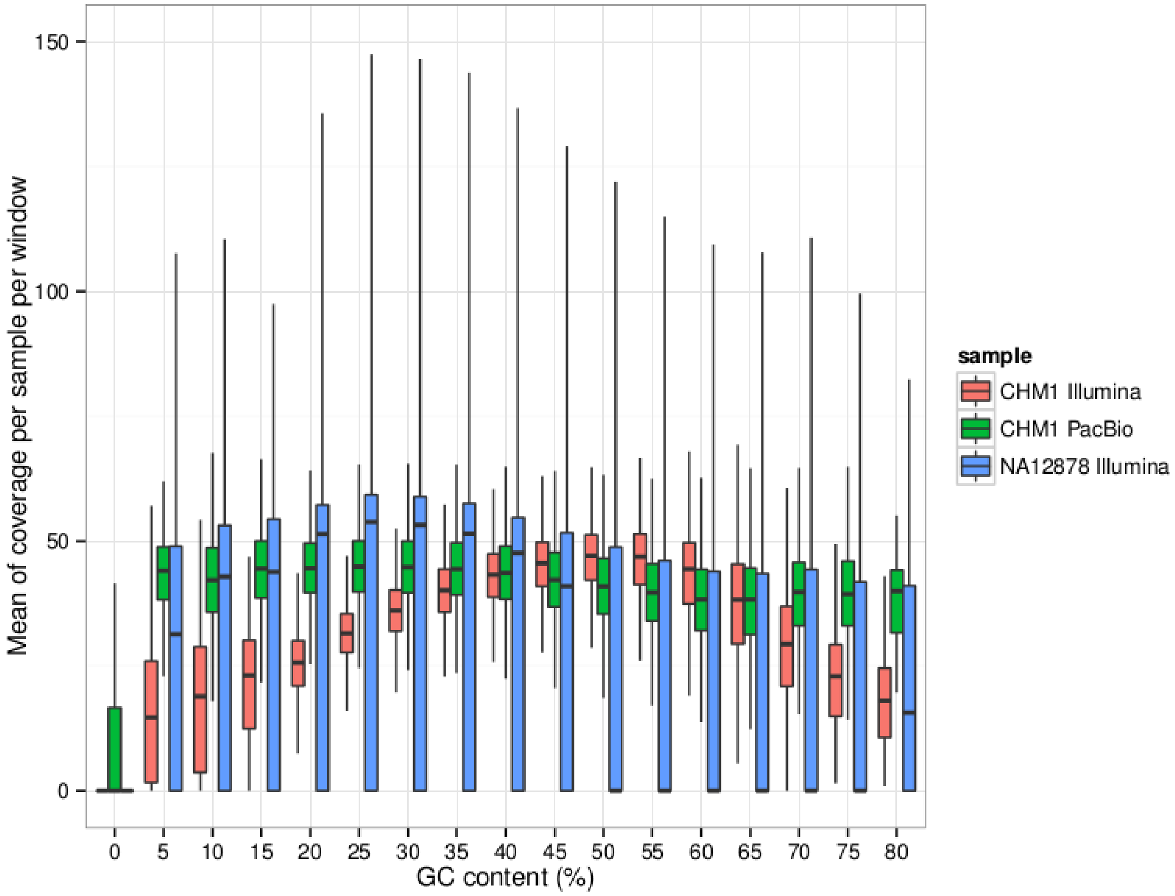
Filename	md5sum	File size (bytes	Size	S3 location
human54x	6a6b0d2f7	62155173270	62G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set0.tgz
human54x	10bf53e67	35756668019	36G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set1.tgz
human54x	d935580d	46577605342	47G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set2.tgz
human54x	7efee40d7	64097672121	64G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set3.tgz
human54x	5114eef4c	38893223032	39G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set4.tgz
human54x	7f4ca1beb	41470209385	41G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set5.tgz
human54x	02eed767b	59379423038	59G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set6.tgz
human54x	a70f4e819	63369072364	63G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set7.tgz
human54x	df27fbd10	55848865706	56G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set8.tgz
human54x	aa7df770b	50151392273	50G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set9.tgz
human54x	7e4967003	62155173270	62G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set10.tgz
human54x	d0464604	42360983826	42G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set11.tgz
human54x	031b9ed9	48509498912	48G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set12.tgz
human54x	f44fdfde04	51035933179	51G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set13.tgz
human54x	093fb639e	52720291499	53G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set14.tgz
human54x	9ba36e62e	14606204928	15G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set15.tgz
human54x	ef5670037	52299318366	52G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set16.tgz
human54x	288e8f0a6	63342138998	63G	https://s3.amazonaws.com/datasets.pacb.com/2014/Human54x/raw/human54x_set17.tgz

Illumina: CHM1htert gDNA was sheared using Covaris S2 with cycling conditions of 10% Duty cycle, Intensity 4, Cycles/Burst 200, and Time 100s. The sheared DNA was then end-repaired using NEBNext End Repair Module (P/N: E6050L). Repaired sheared DNA was then A-tailed and Y-adapters were ligated. The library was size-selected at a range of 450-550 bp then sequenced using Illumina HiSeq PE-101 to generate ~41-fold sequence coverage.

To better understand the nature of the PacBio sequence data, we compared sequence coverage differences between CHM1 PacBio and previously released Illumina genomes based on the % GC content for different portions of the genome. We calculated the mean and variance of genomic coverage for 30 PCR-free Illumina genomes versus the CHM1 PacBio. The PCR-free genomes were mapped with BWA MEM while CHM1 PacBio was mapped with BLASR. The 30 PCR-free Illumina samples showed consistently lower coverage in high GC windows while the PacBio data was much more constant across all windows (Supplementary Figure 1, Supplementary Figure 2). We noted that the standard deviation of the average coverage across all GC windows was nearly half for PacBio at 7-fold compared to CHM1 Illumina and the PCR-free Illumina samples at 12- and 14-fold, respectively. The mean variance of coverage across Illumina samples was six times that of the CHM1 PacBio data with 6.1- and 0.9-fold (Supplementary Figure 2). The coverage bias is consistent with past observations in GC biased prokaryotic resequencing[1] using PacBio and Illumina sequencing, although the PCR-free sequencing reduces bias in low GC composition sequences. Thus, the main benefits of the PacBio reads are increased read length and more uniform coverage across the genome.



Supplementary Figure 1. Mean coverage per sample per GC window across GRCh37 for CHM1 PacBio (red) and 30 PCR-free Illumina samples (blue).



Supplementary Figure 2. Mean coverage by GC content in 1 kbp windows across GRCh37 for CHM1 Illumina and PacBio sequence as well as NA12878 PCR-free Illumina sequence.

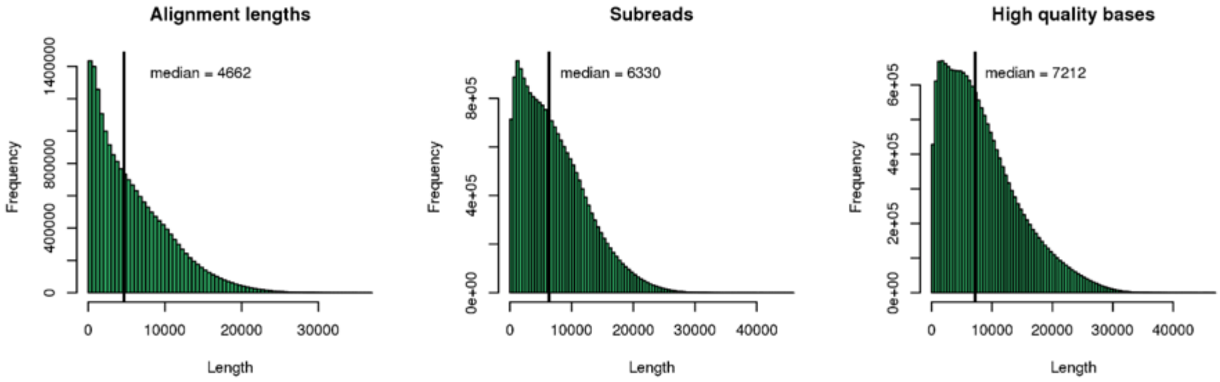
II. Sequence alignment and assembly

Analysis was performed with GRCh37 as opposed to GRCh38 for three reasons. First, it has been the most frequently used genome in analyses over the last four years (e.g., ENCODE, 1KG) and therefore of greatest interest to others in the community. Second, GRCh38 has not been published and is somewhat experimental due to the merging of HuRef and centromeric models within centromeric regions. Finally, CHM1 data (largely from sequenced BACs from CHOR17) have been used to close some gaps in GRCh38. Thus, this would potentially bias our gap-closure abilities by having part of the CHM1 genome already integrated into the human assembly.

We aligned 93.8% of CHM1 SMRT sequence data to GRCh37 using BLASR and considered the effective mapped length as opposed to the total length of sequence reads (Supplementary Figure 3). Alignments to the human reference genome (GRCh37) were performed with the following options: "-bestn 2 -maxAnchorsPerPosition 100 -advanceExactMatches 10 -affineAlign -affineOpen 100 -affineExtend 0 -insertion 5 -deletion 5 -extend -maxExtendDropoff 20 -clipping subread".

For the purpose of this study, we focus our analysis on the euchromatic regions of the genome. The presence of larger repeats and satellite sequences[2] within pericentromeric and subtelomeric regions precluded the generation of robust local assemblies although it was possible to identify single molecules extending into these regions and to estimate the relative amount of various classes of heterochromatic sequence in the dataset (Supplementary Information XI).

In order to integrate SMRT sequence data with standard bioinformatics tools such as SAMtools[3] and the Celera[4] assembler, we developed a custom version of the BLASR program[5] and accompanying software (http://www.github.com/EichlerLab/blasr). The modifications include printing alignments in SAM format with soft clipping based on the coordinates of sequences between adapters and saving the insertion, deletion, substitution, and merge quality values typically stored in HDF format as supplementary fields in SAM files. In this manner, all sequence quality information contained by reads (including soft-clipped bases) is maintained within a BAM alignment file. This file contains sufficient information to reconstruct reads that may be used in assembly and consensus calling routines specific to PacBio data, thus allowing the usage of standard tools operating on BAM files with methods produced by PacBio that require HDF files. We developed a scripting pipeline to allow the collection of all reads overlapping putative loci from a BAM and perform a local assembly using correction-free assembly with Celera[6], and then refine the assembly using the Quiver method and the extra quality values stored in the supplementary fields of the BAM file. The pipeline source code is available at www.githubcom/EichlerLab/chm1_scripts.



Supplementary Figure 3. Mapped SMRT length and accuracy. Histograms show the length frequency distribution of SMRT alignments, subreads, and high-quality bases to GRCh37. (left) The alignment length is determined by the aligned length on the reference. (center) A PacBio read includes one or more reads over a template sequence on an alternating strand, separated by adapter sequence. Every pass over the template sequence is a subread. (right) The full read length includes all subreads and adapter sequences, although subreads are aligned separately. Low-quality bases are annotated at the beginning and end of every read and are excluded from the count of high-quality bases. The low-quality bases are implicitly excluded from the length distributions of alignment length and subread lengths.

We assessed sequence accuracy of our assemblies by comparing to previously sequenced large-insert BAC clones (CH17) from the same source. We refer to the agreement between the BAC assemblies and shotgun consensus as concordance rather than accuracy because the differences are not validated by an orthogonal method. To assess the concordance of the PacBio consensus sequence against a reference representative of the CHM1 haplotype, we compared the consensus sequences of regions with sequenced BACs from CHM1[7] using both Sanger and PacBio sequencing, using the more accurate P4C2 sequencing chemistry[8]. Assuming each BAC sequence to be correct to within at least Q40, we measure sequencing concordance of Q37.5 combined across all regions, as shown in Supplementary Table 3. 73.5% of the errors are confined to deletions of a nucleotide in homopolymer stretches, and the concordance disregarding these errors is Q41.9 in the entire dataset, and 46.6 comparing only to the Sanger assembled BACs.

Supplementary Table 3. Estimate of sequence concordance by comparison against previously sequenced CH17 BACs.

Clone	Length	Mismatches	Insertions	Deletions	Phred	Homopolymer insertions	Homopolymer deletions	Dinucleotide insertions	Dinucleotide deletions	Technology
AC243499.2	199928	0	13	17	38.24	12	17	0	0	Sanger
AC243585.2	193095	0	24	26	35.87	20	24	2	2	Sanger
AC243586.3	221226	0	24	15	37.54	19	14	0	0	Sanger
AC243629.3	232622	0	33	24	36.11	25	24	0	0	Sanger
AC243650.3	239100	0	51	15	35.59	30	14	1	0	Sanger
AC243654.3	227101	3	37	25	35.43	29	24	1	0	Sanger
AC243734.3	209874	0	16	10	39.07	15	10	0	0	Sanger
AC243742.3	216161	2	38	36	34.54	31	34	0	0	Sanger
CH17-091O6	194096	0	38	12	35.89	14	7	14	0	PacBio P4/C2
CH17-144M16	208409	0	13	6	40.4	8	5	2	0	PacBio P4/C2
CH17-150B4	214505	0	45	12	35.76	34	6	0	0	PacBio P4/C2
CH17-285M6	229540	1	49	29	34.63	26	14	4	6	PacBio P4/C2
CH17-390G16	191260	1	53	4	35.18	37	4	13	0	PacBio P4/C2
CH17-63L4	193020	0	14	2	40.81	7	2	3	0	PacBio P4/C2
CH17-68I14	195068	5	26	17	36.09	21	17	0	0	PacBio P4/C2
CH17-9E4	217217	0	67	23	33.83	46	17	0	0	PacBio P4/C2

III. Gap closures

a. Gap closure in GRCh37

Because it is possible to assemble larger complex insertions using reads that align to the flanks of the insertion site, and extend into the insertion, we reasoned it may be possible to use a similar approach to resolve existing gaps in the genome. We initially identified 164 interstitial gaps in GRCh37 by eliminating all telomeric, centromeric, and short arm gaps from the UCSC gap annotation, filtering out gaps that fell completely within an existing GRC patch, and merging remaining gaps that occurred within 5 kbp of each other. Of these gaps, 141 (86%) did not already have a fix patch from the GRC within this version of the human genome.

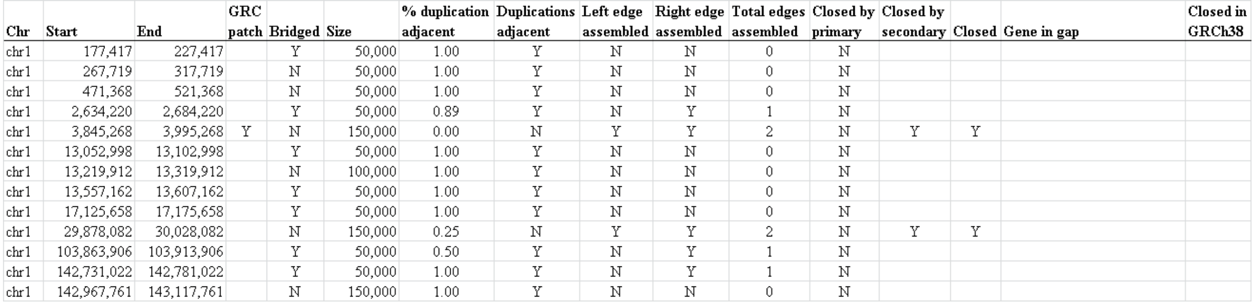
We extended into gaps through a two-part iterative assembly of CHM1 WGS reads mapping to each edge. For the first iteration, we aligned CHM1 PacBio reads to GRCh37 with BLASR, selected reads mapping within 10 kbp of each gap edge, assembled reads with the Celera assembler v8.1, and called consensus sequence with Quiver[9]. For the second iteration, we repeated this process using assemblies from the first iteration as the reference. To avoid incorporating paralogous reads into the second iteration Celera assemblies, we filtered out all CHM1 reads that had longer alignments to GRCh37 than the first iteration assemblies. From these two iterations, we identified gaps with overlapping extensions from both edges and attempted to assemble these gaps with Celera and Quiver using reads mapping to both edges.

Using this approach, we closed 50 gaps and extended into 40 others (60 edges) adding, respectively, 398 kbp and 721 kbp of novel sequence to GRCh37 (Supplementary Table 4 & Supplementary Table 5). We also note that 16 of these were completely resolved in the recently released GRCh38 assembly (Supplemental Information IIIb). The remaining 74 gaps without extensions were significantly enriched for adjacent segmental duplications with 66 (89%) compared to 17 (19%) of the 90 gaps with closures and extensions ($p < 0.0001$; Chi squared value = 236.8 with 1 df). Novel sequence from closures ranged in size from 0 bp—for those closures that simply provided continuity to existing sequences—to 35,482 bp with a mean closure of 8,297 +/- 7,150 bp. Similarly, novel sequence provided by gap extensions ranged from 462 to 30,999 bp with a mean of 12,016 +/- 7,609 bp. Of the 90 gaps with closure or extension, 67 (74%) were not spanned by any fosmid or BAC clone.

Supplementary Table 4. Summary of gap closures and extensions.

	GRCh37		GRCh38	
Category	**Count**	**Bases added**	**Count**	**Bases added**
Interstitial gaps without intersecting GRC patch	164	-	166	-
Closed gaps	50	398,249	31	40,089
Closed gaps with adjacent GRC patches	10	-	-	-
Closed gaps with spanning BACs	3	-	-	-
Gaps with extensions (total edges)	40 (60)	720,962	0	0
Gaps without extensions (with dups)	74 (66)	-	135 (112)	-
Gaps with segmental duplications adjacent	82	-	116	-
Total closures and extensions	110	1,119,211	31	40,089

Supplementary Table 5. Status of assemblies for each interstitial gap in GRCh37.



Chr	Start	End	GRC patch	Bridged	Size	% duplication adjacent	Duplications adjacent	Left edge assembled	Right edge assembled	Total edges assembled	Closed by primary	Closed by secondary	Closed	Gene in gap	Closed in GRCh38
chr1	177,417	227,417		Y	50,000	1.00	Y	N	N	0	N				
chr1	267,719	317,719		N	50,000	1.00	Y	N	N	0	N				
chr1	471,368	521,368		N	50,000	1.00	Y	N	N	0	N				
chr1	2,634,220	2,684,220		Y	50,000	0.89	Y	N	Y	1	N				
chr1	3,845,268	3,995,268	Y	N	150,000	0.00	N	Y	Y	2	N	Y	Y		
chr1	13,052,998	13,102,998		Y	50,000	1.00	Y	N	N	0	N				
chr1	13,219,912	13,319,912		N	100,000	1.00	Y	N	N	0	N				
chr1	13,557,162	13,607,162		Y	50,000	1.00	Y	N	N	0	N				
chr1	17,125,658	17,175,658		Y	50,000	1.00	Y	N	N	0	N				
chr1	29,878,082	30,028,082		N	150,000	0.25	N	Y	Y	2	N	Y	Y		
chr1	103,863,906	103,913,906		Y	50,000	0.50	Y	N	Y	1	N				
chr1	142,731,022	142,781,022		Y	50,000	1.00	Y	N	Y	1	N				
chr1	142,967,761	143,117,761		N	150,000	1.00	Y	N	N	0	N				

The majority of the sequence we assembled for gap closures (3 Mbp) and extensions (2 Mbp) consisted of regions flanking the gap edges that were already present in GRCh37. To evaluate the quality of our assemblies, we aligned these gap-flanking sequences in the assemblies to the corresponding sequence in GRCh37 with BLASR and calculated alignment identity. Flanking sequence from gap closures had a higher overall identity with the reference at 99.1 +/- 1.9% compared to extensions that had a mean alignment identity of 97.9 +/- 6.3%.

We characterized the content of the novel sequences in our gap assemblies as measured by GC content, common repeats identified by RepeatMasker[10], tandem repeats identified by Tandem Repeats Finder[11] (TRF), and putative segmental duplications identified by DupMasker[12] and alignment of novel sequences back to GRCh37. To test for enrichment of GC and repeat content in our novel sequences compared to the human reference, we created a null distribution of equivalently sized events across GRCh37 and compared our observed means against the null with permutation tests (n = 100,000).

Gap closure sequences consisted primarily of a high proportion of simple repeats, long tandem repeats, and extreme GC content. A more detailed examination by dot-matrix analysis showed clusters of degenerate repeat motifs that were highly related at the sequence level (Figure 1c-d). Simple repeats represented a significant proportion of gap closures ($p < 0.00001$) with 28 +/- 22% of gap sequence annotated as simple repeats compared to 2 +/- 3% for the sampled reference (Figure 1a). Indeed, 39 of the 50 closures (78%) consisted of more than 10% simple repeats. Correspondingly, closures were highly enriched for long tandem repeats compared to sampled reference sequences ($p < 0.00001$) with mean tandem repeats of 706 +/- 1,284 bp compared to 306 +/- 1115 bp. The most common tandem repeat motifs were AT with 44 kbp total sequence, GT with 12 kbp, and AC with 7 kbp. While closures were enriched for these simple repetitive sequences, they were also depleted for LINE/L1s compared to the reference ($p = 0.00038$) with a mean proportion of 7 +/- 11% in closures compared to 16% +/- 23% in the reference.

The overall GC content in closures (42.39%) was significantly higher than the rest of the genome ($p = 0.02712$) with a bimodal distribution, including a high mode at 48.9 +/- 4.9% and a low mode at 20.6 +/- 8.6% (Figure 1a). To better understand the cause of the bimodal distribution of

GC content, we delineated four types of regions within our gap assemblies, including reference flanks already present in GRCh37, novel gap closures, tandem repeats within gap closures, and non-tandem repeat sequence within gap closures. We calculated GC content in each of these regions and compared their distributions to the sampled sequences from GRCh37 and the overall GC content of the reference (Figure 1a). We find that the bimodal distribution of GC content in gap closures is primarily the result of GC-rich and AT-rich tandem repeats within closures. When closures are inspected without including tandem repeats, we find overall higher GC (45%; p < 0.00001) than in the sampled GRCh37 sequences (41%) or the overall GRCh37 content (41%). The existing reference sequence adjacent to gap closures was similarly enriched for higher GC content (45%; p < 0.00001) compared to sampled GRCh37 sequences of the same size. Thus, a complex model emerges for gaps not flanked by segmental duplication of long tracts of degenerate STR often multiple kilobases in length embedded within GC-rich regions of the genome.

While gap extensions were also enriched for simple repeats, long tandem repeats, and extreme GC content, the degree of enrichment was less than that seen in closures. Simple repeats composed 8 +/- 14% of extension sequences compared to 1 +/- 2% in the reference (p < 0.00001). Only 16 of 60 extensions (27%) consisted of more than 10% simple repeats. Tandem repeats were also significantly longer in extension sequences at 450 +/- 929 bp compared to 305 +/- 1190 bp in the reference (p = 0.02375). As with closure sequences, the most common tandem repeats were AT motifs with 16 kbp total sequence, GT with 8 kbp, and AC with 4 kbp. Unlike closures, extensions only had higher GC content than the reference with 45.75 +/- 9.21% compared to 43.67 +/- 7.43% (p = 0.01648) and not significantly lower GC content.

In addition to adding novel sequence to GRCh37, gap closures and extensions have the potential to add previously unidentified segmental duplications or additional copies of known duplications. We evaluated the potential segmental duplication content of these novel sequences through alignment against known primate duplication cores with DupMasker as well as alignment against GRCh37 with MEGABLAST[13] (v. 2.2.11) to identify existing regions of the reference with >90% identity and >1 kbp alignments. We identified putative duplication content totaling 31 kbp in 17 of 50 (34%) of gap closures and 64 kbp in 22 of 60 (37%) of gap extensions. On average, closures contained an additional 2 kbp of duplication content while extensions contained an additional 1 kbp. Based on these results, novel gap sequences provide an additional 95 kbp of segmental duplications to GRCh37.

Two of 50 (4%) closures and 13 of 60 (22%) extensions had alignments >=1 kbp and >90% identity to existing regions of GRCh37. One novel closure sequence from chr18:52,044,136-52,224,136 had a complete alignment at 99.86% identity to the unlocalized contig chr18_GL000207_random. Inspection of GRCh38 confirms the localization and the orientation of this random contig at this region where a gap of 954 bp still remains. The second closure with a high-identity alignment to GRCh37 is from the gap at chr1:29,863,082-30,043,082 and extends an existing segmental duplication at chr1:31,129,342-31,131,869 by 1,290 bp. One 2 kbp gap extension from chr1:142770022-142792023 matches 19 distinct regions of GRCh37, including two alignments at >99.4% identity to the unplaced contig chrUn_GL000224. Altogether, novel closure and extension sequences with high-identity alignments to GRCh37 represent 110 kbp of putative segmental duplications missing from GRCh37, which is consistent with our DupMasker annotation.

b. Gap closure in GRCh38

To close gaps in GRCh38, we repeated the analysis we performed for GRCh37. We first defined the set of all gaps that were not telomeric, centromeric, or acrocentric as "interstitial" gaps using UCSC's release of GRCh38. We merged all gaps that occurred within 10 kbp of each other reducing the initial set of 189 interstitial gaps to 172 regions for potential closure. We aligned PacBio whole-genome sequence from CHM1 to GRCh38 with BLASR and calculated the coverage and repeat content in 10 kbp adjacent to each gap. Finally, we omitted six regions with coverage greater than 500-fold whose median coverage was 2,803-fold and which were not likely to assemble correctly with Celera. We targeted 166 interstitial euchromatic gaps for *de novo* assembly.

With this approach, we closed an additional 31 gaps in GRCh38 including 4 closures in segmental duplications (Supplementary Table 4 and Supplementary Table 6). No gaps were reduced by a simple extension. The total novel sequence added by these gap assemblies was dramatically reduced compared to GRCh37 closures with only 40,089 bp total from 27 assemblies. The remaining 4 assemblies provided no additional sequence but instead confirmed the continuity of the sequences adjacent to the annotated gap in GRCh38. Of the remaining 135 open gaps, 94% mapped to segmental duplications (n=112) or high copy satellite repeat sequence (n=15)—regions that cannot yet be reliably access by current SMRT sequencing technology unless a significant increase in read-length. This is reflected in the read-depth for the remaining open gaps. The median adjacent coverage of PacBio reads for unclosed gaps in segmental duplications was 25% of the median coverage for closed gaps in duplications. This pattern is consistent with the fact that fewer high-quality alignments are possible in duplicated regions of the genome.

Supplementary Table 6. Status of assemblies for each interstitial gap in GRCh38.

			Adjacent			Region		Assembled	Assembled	Duplicated	Closure
Chrom	**Start**	**End**	**repeat type**	**Reads**	**Bases**	**size (bp)**	**Coverage**	**contigs**	**bases**	**edges**	**status**
chr1	287,968	357,968	SINE	186	1,699,747	50,000	24.28	2	48,643	2	open
chr1	525,988	595,988	LTR	89	667,866	50,000	9.54	2	34,003	2	open
chr1	2,692,781	2,756,290	Simple_repeat	401	3,603,846	43,509	56.75	15	119,995	2	open
chr1	12,944,384	13,014,384	SINE	181	1,263,392	50,000	18.05	2	44,526	2	open
chr1	16,789,163	16,859,163	LTR	210	1,559,222	50,000	22.27	2	43,800	2	open
chr1	29,542,233	29,563,835	Simple_repeat	191	1,612,843	1,602	74.66	1	51,387	1	closure
chr1	125,093,213	125,113,233	LTR	215	1,820,511	20	90.93	1	48,401	1	closure
chr1	125,120,246	125,141,847	LINE	173	1,214,822	1,601	56.24	2	42,043	2	open
chr1	125,161,347	125,183,583	Simple_repeat	14,366	128,035,828	2,236	5758.04	N/A	N/A	2	open
chr1	223,548,935	223,618,935	Simple_repeat	224	2,029,397	50,000	28.99	1	43,550	0	closure

c. Insertion variants in other genomes.

We assessed what fraction of the PacBio-closed gaps was closed in the recent Illumina-based whole-genome sequence assembly of the hydatidiform mole (CHM1.1). In no instance were any of the 50 PacBio-closed gaps fully resolved in CHM1.1, although partial sequence was present for 16/50 gaps. Similarly, only 14/1,737 (0.7%) of the complex insertions identified using the PacBio data were present in the Illumina-based assembly (Supplementary Table 7). Comparing to GRCh38, we find that even less of the insertions are represented (only 5/1,737) and all of these are also represented in CHM1.1. Thus, 14 of 1737 insertion sequences map to either GRCh38 or CHM1.1, indicating a strong bias against correctly assembling these inserted sequences using short-read technology and highlighting the new biology enabled by PacBio sequencing.

Supplementary Table 7. Presence of inserted sequences in other human assemblies.

	Insertion		hg19		GRCh38		CHM1.1	
	count	bases	count	bases	count	bases	count	bases
Complex	1116	2148286	0	0	4	21567	11	36303
STR	406	826962	0	0	0	0	2	2344
VNTR	215	498362	1	1846	1	1846	1	1846

IV. Structural variation detection

a. Variant detection pipeline

We developed a computational pipeline (Extended Data Fig. 1) to discover structural variation—defined here as changes deletions, duplications, insertions or inversions ≥50 bp in length. The pipeline determines variants by comparing local assemblies to the reference and accounts for the lower per-read accuracy of single-molecule sequencing (SMS) reads by using consensus sequences of the assemblies that are refined using the Quiver method. To reduce the computational burden, local assemblies are performed only at putative variant loci rather than performing assemblies tiling the genome.

Putative structural variant loci are detected as follows: reads are mapped to the reference allowing up to two alignments (a primary and secondary alignment), and the alignments are examined for aberrancies: insertions, deletions, and truncations. Insertions and deletions are referred to as *spanned events* while truncated alignments (alignments that do not span the length of an entire read) are referred to as *hard-stop* events. Clusters of events (two or more) with overlapping coordinates define putative variant loci. Typically, smaller indels are detected as spanned events, while larger structural variants are not spanned by alignments and are observed as a number of truncated alignments that end at approximately the same position on the reference. Different signatures of clusters of hard-stop events are used to detect larger insertion, deletion, and inversion structural variants. Inserted sequences are detected as clusters of hard-stop events involving only the primary or secondary alignments of reads. Deleted sequences require two separate clusters of hard-stop events with the primary and secondary alignments of each read present in one cluster and in the same orientation, and large inverted sequences are detected in a similar fashion as deleted sequences with the additional requirement that the primary and secondary alignments must be in opposite orientation. To increase sensitivity of detecting smaller insertion and deletion events, we generated a consensus sequence of the entire genome using the Quiver method, which has modest computational requirements relative to whole-genome *de novo* assembly. The consensus was mapped back to the genome in 10 kilobase tiled sequences using BLASR, and insertion and deletion calls were merged with calls based on the assembly pipeline. To increase sensitivity for detecting smaller inversions, we implemented a method that searches for secondary alignments fully overlapped by a primary alignment and in reverse orientation. Reads with such alignments are modified so that the substring of the read corresponding to the secondary alignment is reverse complemented, and the read is realigned. If the new alignment has a higher alignment score than the original primary alignment, the secondary alignment gives the breakpoints of an inversion. A graphical summary of the structural variation pipeline is given in Extended Data Fig. 1.

The interval coordinates of spanned insertions are defined by chr:(start–*delta*)-(start+*delta*), where chr is the chromosome, start is the position on the reference where an insertion in a read begins, and *delta* is a parameter representing the uncertainty in starting position of an insertion (100 bp). Deleted sequences are defined by chr:start-end, where start is the starting position of the deleted sequence, and end is the ending position of the deleted sequence in the alignment. The positions of hard-stop events are defined as the position of the beginning of an alignment if the position is greater than 500 bases from the start of a read, and the end of an alignment if the end is greater than 500 bases from the end of the read. If the alignment of a read is truncated by greater than 500 bases on both ends of a read, there are two hard-stop positions for that read. The interval coordinates of a hard-stop event are similar to that of an insertion: chr:(pos-*delta*)-(pos+*delta*).

a. **b.**

```
Ref GGACGTC-CC-G-CCCGCT   Ref GGACGTCCCGCC---CGCT
    ||||||| || | | ||||       |||||||||***   ||||
Qry GGACGTCCCCGGGC-CGCT   Qry GGACGTCCCCGGGC-CGCT
```

a. Before *condense-3* b. After *condense-3*

Supplementary Figure 4. Example of the alignment condensation operation. (a) A pairwise alignment with gaps interlaced with matches. (b) Three matches: GC, G, and C, starting at the 8th base in the reference must be shuffled left because they are less than 3 bp in length.

The alignment of some structural variants requires a slight permutation of gap locations to recover the full-length event. Positions of gaps in alignments are shuffled so that stretches of matches less than a parameter *condense-N* bases (N is 20 by default) flanked by gaps are shifted left until it is adjacent to a match greater than *condense* bases, or another stretch of matches that have been maximally shifted. An example of the gap shifting operation is shown in **Error! eference source not found.**.

Many of alignments of larger inserted or deleted sequences are broken up into segments of smaller inserted or deleted sequences. While the condense operation creates alignments that have suboptimal base pairing, this operation serves to recover longer insertions or deletions from scattered stretches. This is particularly important in the alignment in the presence of inserted or deleted simple tandem repeats.

A cluster of reads is defined as all reads with either a spanning or hard-stop event that have overlapping coordinates of the event and the event is the same (e.g., insertion, deletion, one-sided hard-stop, etc.). All insertion intervals with the same chromosome and overlapping intervals are clustered, and the coordinates of each cluster are defined as the common chromosome: minimal starting position of the insertion until the maximal value of insertion start + insertion length for all alignments in the cluster. The number of reads in each cluster (cluster size) is compared to the average coverage of all mapped reads. Clusters with coverage fitting the following criteria are retained as candidate insertions: the cluster size is at least half the average coverage, not more than 1.5 times the average coverage, and the cluster size and coverage are at least 5 reads.

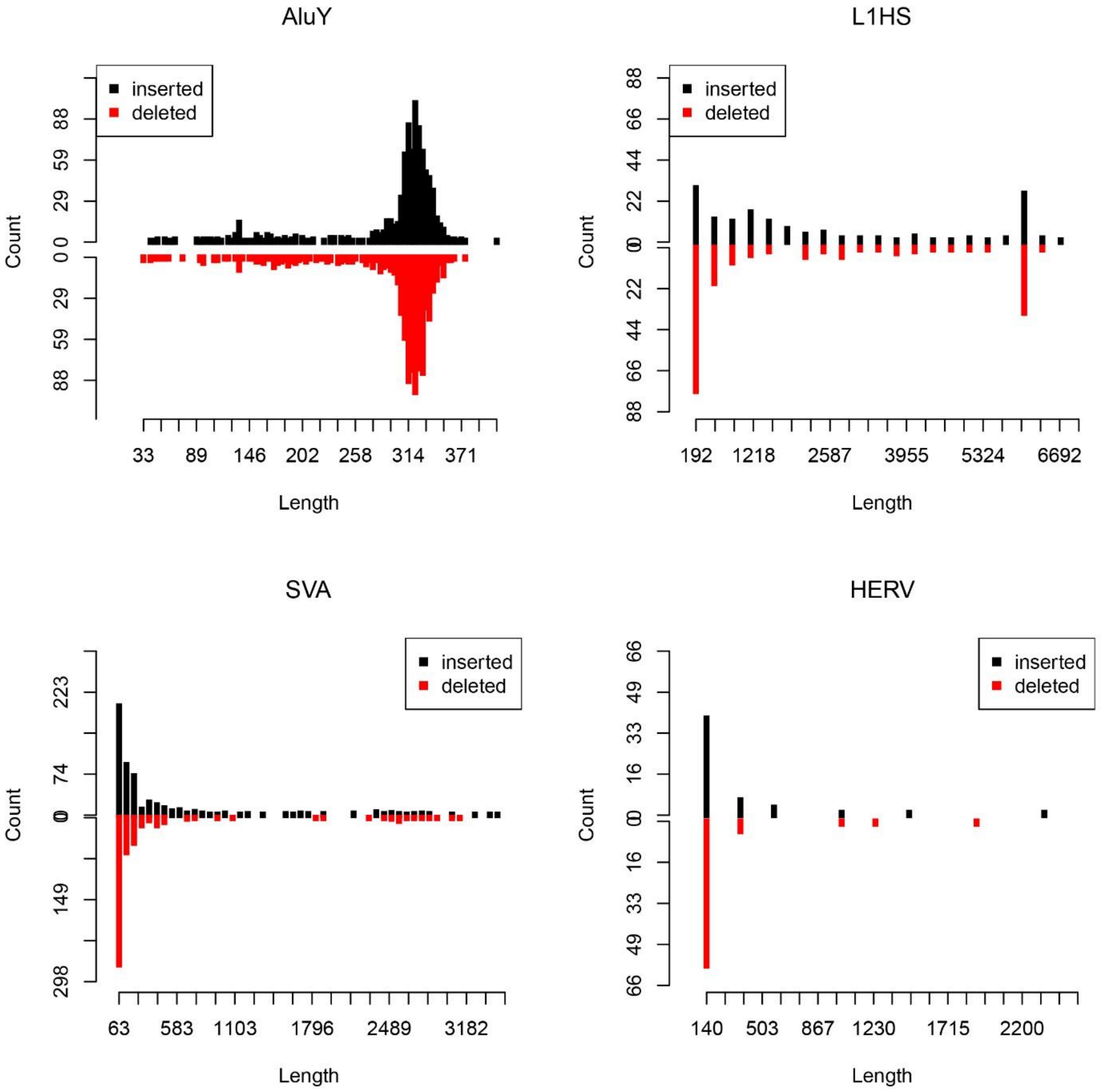
Assemblies are generated for all reads overlapping a cluster locus. Because assemblies were local, the complexities of the sequences assembled were low, and 96% were resolved into a single contig (98.8% in two), with an average contig length of 24 kbp, although most indels captured by a variant were less than a 1000 bases[14]. A SAM file is generated using SAMtools view, and then the reads and quality values are converted into both a FASTQ file suitable for input to the Celera assembler, and a bas.h5 file compatible with the Quiver resequencing pipeline using a custom program samToBasH5. Assemblies are performed with the Celera assembler v8.1. Consensus is called by mapping the reads from the bas.h5 file with a minimum mapping quality of 20 and running Quiver on the resulting alignment files.

Consensus sequences are mapped back to the reference using the same mapping parameters as original reads, and a set of insertions and deletions is called from the resulting alignments after performing the condense operation. It is possible that the assembled contigs will overlap on the reference and multiple alignments may cover the same insertion or deletion event. It is necessary to remove indel calls overlapping intervals in order to prevent multiple calls of the same event, but because alignments may vary slightly, the boundaries of the calls may not be the same between alignments. To remove overlapping deletions, the longest overlapping deletion is selected. The intervals of insertion and deletion calls are defined as above. Overlapping intervals are removed when the intervals overlap and the start positions are within 200 bp.

The sequences of indels are annotated using multiple repeat masking pipelines to maximize sensitivity. First, sequences are repeat masked with CENSOR[14] v. 4.2.28 using the human repeat library and NCBI BLAST to perform alignments. Sequences not masked by Censor are masked using RepeatMasker v. 3.3.0. Finally, sequences that remain unannotated are masked by TRF v. 4.07b with the options 2 7 7 80 10 20 500 -m -ngs.

b. Detection of mobile element insertions (MEIs) and deletions

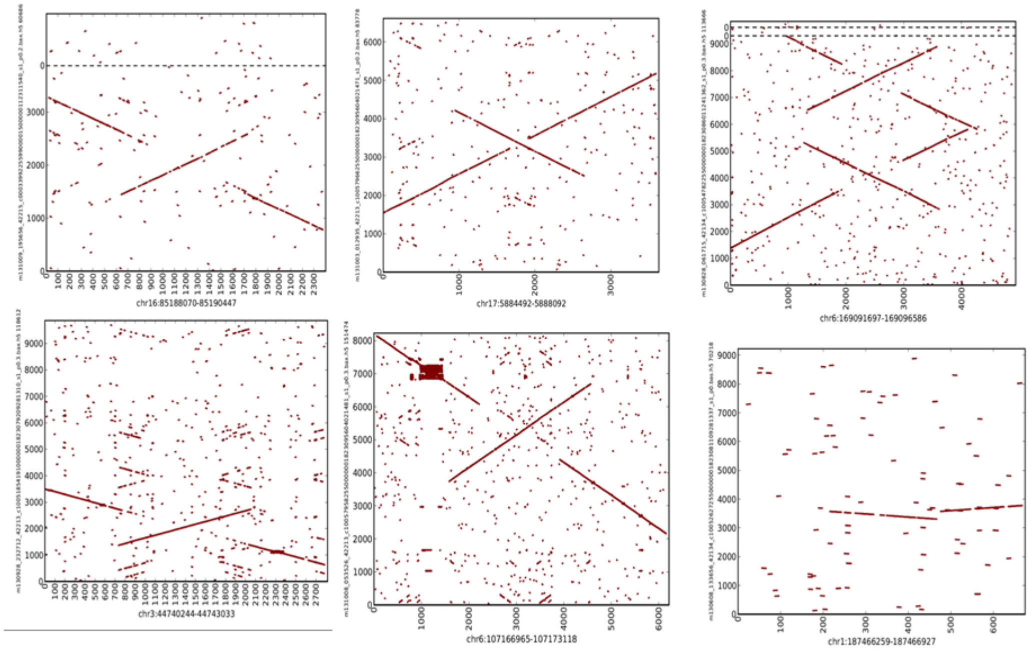
All variant calls arise due to one of the following phenomena: a variant is a sequencing artifact or computational error and, therefore, a false positive; the reference is incomplete or incorrect; or a variant represents a true polymorphism between the CHM1 sequence and GRCh37. We compared the insertion and deletion counts of the active mobile elements AluY and L1HS (Supplementary Figure 5) and found that there is a roughly equivalent representation of insertions and deletions measured of AluY (p = 0.902, binomial) and LINE/L1HS (p = 0.860). Furthermore, the 859 AluY insertions is similar to the 987 Alu insertions found for the diploid NA12878[15], and 145 L1HS insertions in CHM1 is similar to the 161 L1 insertions found in the same study. Many of the insertion events are in highly repetitive regions that are difficult to validate using PCR or shorter read technology.



Supplementary Figure 5. MEI comparison of CHM1 and GRCh37. The counts of the active mobile elements by length, for both inserted and deleted mobile elements, in CHM1 are shown. There are 1115 insertion and 608 deletion sequences comprising 2.15 Mbp and 0.654 Mbp of the genome, respectively, that are annotated as containing more than one repeat type. These are labeled as complex events as they cannot be explained by a simple mechanism of MEI.

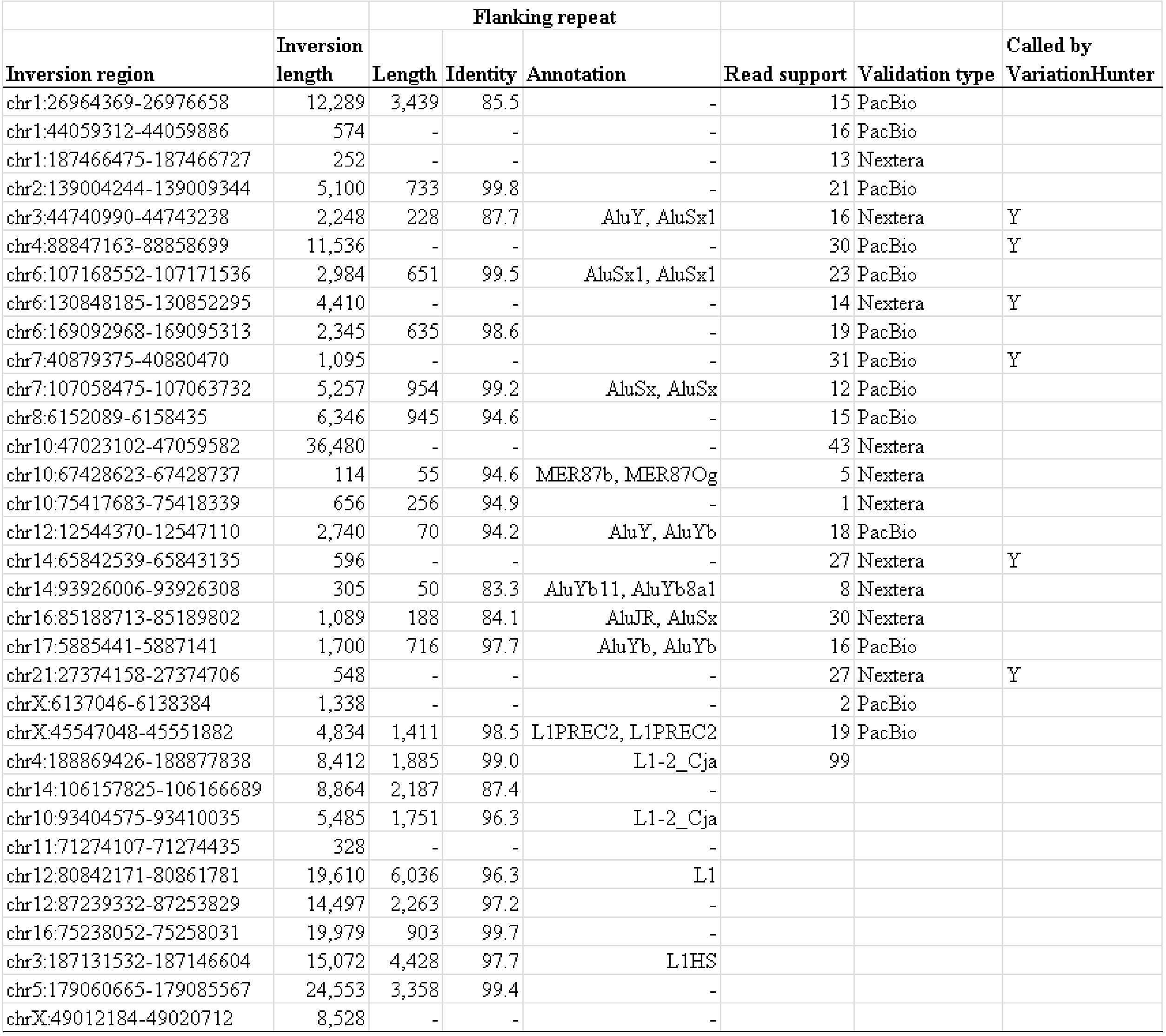
c. Inversions

We searched for additional structural variation in the form of inversions by directly detecting reversals in order from the SMRT sequence reads (Supplementary Figure 6). Because individual reads span inversions, this offers an accurate method to detect short inversions in regions that may be highly repetitive. The breakpoints of inversions are defined by performing a local assembly of the region where the inversion is detected and finding the optimal inversion in the local assembly sequence that maximizes the alignment score to GRCh37. There were 34 inversions detected between CHM1 and GRCh37 corresponding to a total of 242 kbp of inverted sequence with average length 7.1 kbp (Supplementary Table 8 and Supplementary Figure 6). No genes were interrupted by inversions. We searched for repetitive sequences of at least 50 bp and 80% similarity flanking repeats and found that 24 inversions lacked flanking repeats according to this definition.



Supplementary Figure 6. Detection of inversions with single-molecule sequences. SMRT sequence (y-axis) compared to human reference sequence (x-axis) with dotplots display inversions as reversals frequently flanked by repeats.

Supplementary Table 8. Inversions detected by single-molecule sequencing including analysis of repeat sequences flanking each event.



		Flanking repeat					
Inversion region	Inversion length	Length	Identity	Annotation	Read support	Validation type	Called by VariationHunter
chr1:26964369-26976658	12,289	3,439	85.5	-	15	PacBio	
chr1:44059312-44059886	574	-	-	-	16	PacBio	
chr1:187466475-187466727	252	-	-	-	13	Nextera	
chr2:139004244-139009344	5,100	733	99.8	-	21	PacBio	
chr3:44740990-44743238	2,248	228	87.7	AluY, AluSx1	16	Nextera	Y
chr4:88847163-88858699	11,536	-	-	-	30	PacBio	Y
chr6:107168552-107171536	2,984	651	99.5	AluSx1, AluSx1	23	PacBio	
chr6:130848185-130852295	4,410	-	-	-	14	Nextera	Y
chr6:169092968-169095313	2,345	635	98.6	-	19	PacBio	
chr7:40879375-40880470	1,095	-	-	-	31	PacBio	Y
chr7:107058475-107063732	5,257	954	99.2	AluSx, AluSx	12	PacBio	
chr8:6152089-6158435	6,346	945	94.6	-	15	PacBio	
chr10:47023102-47059582	36,480	-	-	-	43	Nextera	
chr10:67428623-67428737	114	55	94.6	MER87b, MER87Og	5	Nextera	
chr10:75417683-75418339	656	256	94.9	-	1	Nextera	
chr12:12544370-12547110	2,740	70	94.2	AluY, AluYb	18	PacBio	
chr14:65842539-65843135	596	-	-	-	27	Nextera	Y
chr14:93926006-93926308	305	50	83.3	AluYb11, AluYb8a1	8	Nextera	
chr16:85188713-85189802	1,089	188	84.1	AluJR, AluSx	30	Nextera	
chr17:5885441-5887141	1,700	716	97.7	AluYb, AluYb	16	PacBio	
chr21:27374158-27374706	548	-	-	-	27	Nextera	Y
chrX:6137046-6138384	1,338	-	-	-	2	PacBio	
chrX:45547048-45551882	4,834	1,411	98.5	L1PREC2, L1PREC2	19	PacBio	
chr4:188869426-188877838	8,412	1,885	99.0	L1-2_Cja	99		
chr14:106157825-106166689	8,864	2,187	87.4	-			
chr10:93404575-93410035	5,485	1,751	96.3	L1-2_Cja			
chr11:71274107-71274435	328	-	-	-			
chr12:80842171-80861781	19,610	6,036	96.3	L1			
chr12:87239332-87253829	14,497	2,263	97.2	-			
chr16:75238052-75258031	19,979	903	99.7	-			
chr3:187131532-187146604	15,072	4,428	97.7	L1HS			
chr5:179060665-179085567	24,553	3,358	99.4	-			
chrX:49012184-49020712	8,528	-	-	-			

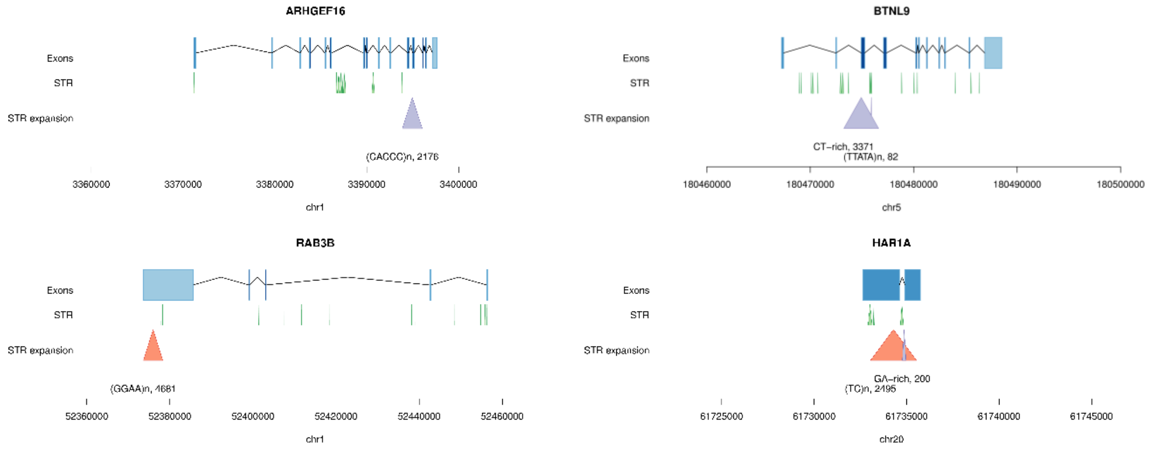
d. STR Expansions

The GRCh37 reference contains contracted sequences of short tandem repeats (STRs) mapping within genes. A total of 2289 genes have at least one STR expansion detected in CHM1, and 222 genes have an STR expansion greater than 1 kbp. Fifteen genes have an insertion inside a UTR, and two—*MUC2* and *SAMD1*—have an insertion inside a coding sequence exon. A short insertion in *FMR1* is shown in Supplementary Figure 7 (top) and exhibits the (CGG)9(AGG) repeat polymorphism motif demonstrating an accurate reconstruction of the consensus sequence[16], and a similar well characterized CCCCGG hexanucleotide expansion in C9orf72[17]. Examples of the genomic architectures of genes with STR insertions in intronic and UTR sequences are shown in Supplementary Figure 9. The expanded STRs have a low but statistically significant ($p < 1x10^{-15}$) correlation with recombination rate[18] ($r^2 = 0.23$, $p < 1x10^{-15}$, Pearson correlation), and human-chimpanzee divergence ($r^2 = 0.23$, $p < 1x10^{-15}$, Pearson correlation) in 1

Mbp bins[19], consistent with an increase of divergence near telomeres[19], and a lower and less significant correlation with G+C biased gene conversion[20] ($r^2 = 0.07$, $p = 3.05 \times 10^{-7}$, Pearson correlation). The AluY, L1, and HERV, and SVA insertion counts are not significantly different from their deletion counts.

```
FMR1
    8034  GCGGGCGGCGGGCCGACGGCGAGCGCGGGCGGCGGCGGTGACGGAGGCGC
          ||||||||||||||||||||||||||||||||||||||||||||||||||
146993493  GCGGGCGGCGGGCCGACGGCGAGCGCGGGCGGCGGCGGTGACGGAGGCGC

    8084  CGCTGCCAGGGGGCGTGCGGCAGCGCGGCGGCGGCGGCGGCGGCGGCGGC
          ||||||||||||||||||||||||||||||||||||||||||||||||||
146993543 CGCTGCCAGGGGGCGTGCGGCAGCGCGGCGGCGGCGGCGGCGGCGGCGGC

    8134  GGCGGAGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGAGGCGGCGGCGG
          ||||||||||||                              ||||||||
146993593 GGCGGAGGCGGC------------------------------GGCGGCGG

    8184  CGGCGGCGGCGGCGGCTGGGCCTCGAGCGCCCGCAGCCCACCTCTCGGGG
          ||||||||||||||||||||||||||||||||||||||||||||||||||
146993613 CGGCGGCGGCGGCGGCTGGGCCTCGAGCGCCCGCAGCCCACCTCTCGGGG

    8234  GCGGGCTCCCGGCGCTAGCAGGGCTGAAGAGAAGATGGAGGAGCTGGTGG
          ||||||||||||||||||||||||||||||||||||||||||||||||||
146993663 GCGGGCTCCCGGCGCTAGCAGGGCTGAAGAGAAGATGGAGGAGCTGGTGG
```

```
FTD-ALS
   5295  GGATGCCGCCTCCTCACTCACCCACTCGCCACCGCCTGCGCCTCCGCCGC
         ||||||||||||||||||||||||||||||||||||||||||||||||||
27573407 GGATGCCGCCTCCTCACTCACCCACTCGCCACCGCCTGCGCCTCCGCCGC

   5345  CGCGGGCGCAGGCACCGCAACCGCAGCCCCGCCCCGGGCCCGCCCCCGGG
         ||||||||||||||||||||||||||||||||||||||||||||||||||
27573457 CGCGGGCGCAGGCACCGCAACCGCAGCCCCGCCCCGGGCCCGCCCCCGGG

   5395  CCCGCCCCGACCACGCCCCGGCCCCGGCCCCGGCCCCGGCCCCGGCCCCG
         ||||||||||||||
27573507 CCCGCCCCGACCAC------------------------------------

   5445  GCCCCGGCCCCGGCCCCGGCCCCGGCCCCGGCCCCTAGCGCGACTCCT
                     ||||||||||||||||||||||||||||||||||||||
27573521 ------------GCCCCGGCCCCGGCCCCGGCCCCTAGCGCGACTCCT

   5495  GAGTTCCAGAGCTTGCTACAGGCTGCGGTTGTTTCCCTCCTTGTTTTCTT
         ||||||||||||||||||||||||||||||||||||||||||||||||||
27573559 GAGTTCCAGAGCTTGCTACAGGCTGCGGTTGTTTCCCTCCTTGTTTTCTT

   5545  CTGGTTAATCTTTATCAGGTCTTTTCTTGTTCACCCTCAGCGAGTACTGT
         ||||||||||||||||||||||||||||||||||||||||||||||||||
27573609 CTGGTTAATCTTTATCAGGTCTTTTCTTGTTCACCCTCAGCGAGTACTGT
```

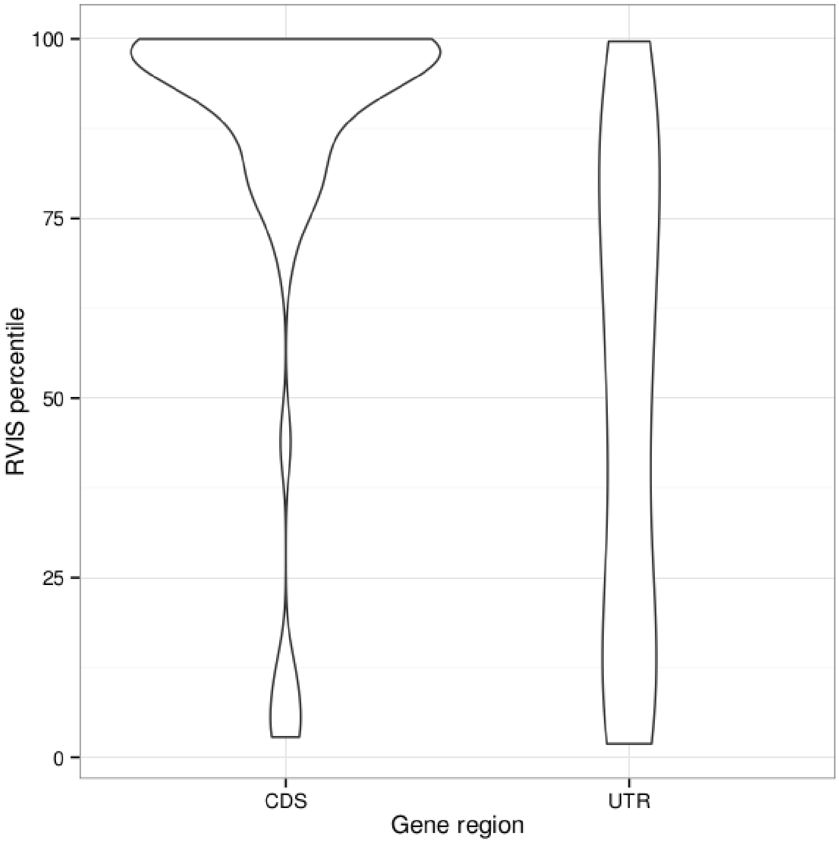


Supplementary Figure 7. Examples of STR insertions in genes. (top, left) An insertion in FMR1 demonstrating a canonical (CGG)9(AGG) insertion[21]. (top, right) The consensus sequence of the C9orf72 hexanucleotide repeat region. (middle) Insertions in intronic sequences. (bottom) Insertions in UTR regions of genes.

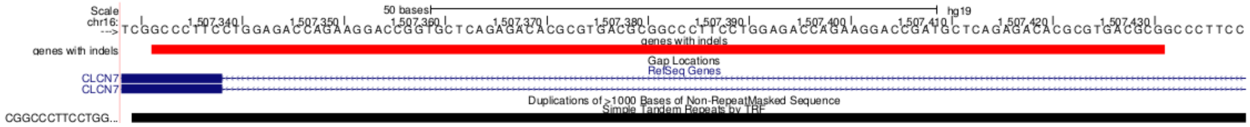
e. Insertions and deletions inside genes

Of the 15,749 total euchromatic insertions and deletions detected in CHM1, 169 indels mapped within coding exons or UTRs of 140 genes (Supplementary Table 9). Events were evenly distributed by type with 92 insertions and 77 deletions and occurred as often in gene UTRs as in coding exons. Of the 82 indels inside coding exons, 49 (60%) appeared to maintain the reading frame by adding or removing bases in multiples of three and correspond to expansion and contraction of variable amino acid repeat motifs associated with environmental interaction genes (e.g., mucins, epidermal differentiation complex, etc.). The genes affected by indels have little in common functionally based on annotation with DAVID[22]. However, these genes are strongly enriched for repetitive elements (DAVID enrichment score: 8.58). Only 22 of the 92 insertions (24%) were not identified as repetitive by RepeatMasker or TRF (Supplementary Table 9). Similarly, 19 of 77 deletions (25%) were not annotated as repetitive content. While we only considered structural variation >=50 bp for the majority of this study, it is worth noting that we discovered an additional 76 indels that were smaller than 50 bp in 70 distinct genes. These events reflect the sensitivity of our approach and potentially functional relevance.

We inspected the mutational tolerance of all genes with indels using the Residual Variation Intolerance Score (RVIS) percentile[23] based on the location of indels in genes. As expected, genes with mutations in coding exons are highly tolerant of mutations while genes with mutations in UTRs are equally distributed across the tolerance landscape (Supplementary Figure 8). Only four genes with exonic indels have an RVIS percentile less than 25. In the case of the deletions in *CLCN7* and *COL6A2*, the deletions maintain the identical sequence of the exons that occurs in the adjacent introns (Supplementary Figure 9). One insertion in *SULF2* occurs in the last half of the gene within a tandem repeat. The other insertion in *ADARB1* occurs in the introns for six of eight isoforms.



Supplementary Figure 8. Mutational tolerance of genes with insertions and deletions grouped by region of the gene affect. Tolerance of mutations is measured by RVIS percentile where higher percentiles represent genes with higher tolerance for mutations. As expected, most mutations inside coding exons occur within genes that are highly tolerant of mutations.


Scale
50 bases
hg19
chr16:
1,507,340
1,507,350
1,507,360
1,507,370
1,507,380
1,507,390
1,507,400
1,507,410
1,507,420
1,507,430
---> TCGGCCCTTCCTGGAGACCAGAAGGACCGGTGCTCAGAGACACGCGTGACGCGGCCCTTCCTGGAGACCAGAAGGACCGATGCTCAGAGACACGCGTGACGCGGCCCTTCC
genes with indels
genes with indels
Gap Locations
RefSeq Genes
CLCN7
CLCN7
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Simple Tandem Repeats by TRF
CGGCCCTTCCTGG...

Supplementary Figure 9. Deletion in the highly conserved exon of CLCN7 maintains frame with repetitive sequence in the adjacent intron.

Supplementary Table 9. All (>1 bp) CHM1 insertion and deletion events intersecting genes in coding exons or UTRs. Gene tolerance to mutation is shown by RVIS percentile when available.

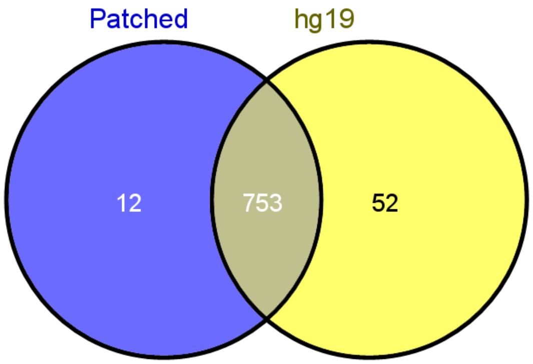
Chr	Start	End	Event type	Event size (bp)	Event repeat content	Gene	Gene region	Gene bases affected	RVIS percentile
chr1	788,856	788,905	deletion	49	non-repetitive	LINC01128	UTR	49	N/A
chr1	3,406,919	3,406,920	insertion	46	repetitive	MEGF6	UTR	46	97.81
chr1	26,671,495	26,671,546	deletion	51	non-repetitive	AIM1L	CDS	51	N/A
chr1	38,000,841	38,000,842	insertion	40	repetitive	SNIP1	UTR	40	63.20
chr1	39,879,332	39,879,371	deletion	39	non-repetitive	KIAA0754	CDS	39	N/A
chr1	78,353,599	78,353,659	deletion	60	repetitive	NEXN-AS1	UTR	60	N/A
chr1	78,353,675	78,353,676	insertion	43	repetitive	NEXN-AS1	UTR	43	N/A
chr1	110,231,897	110,231,898	insertion	239	non-repetitive	GSTM1	CDS	239	94.84
chr1	151,819,675	151,819,722	deletion	47	repetitive	THEM5	UTR	47	90.84
chr1	152,129,067	152,129,103	deletion	36	non-repetitive	RPTN	CDS	36	93.42
chr1	152,188,537	152,188,538	insertion	708	repetitive	HRNR	CDS	708	N/A
chr1	152,189,315	152,190,724	deletion	1,409	repetitive	HRNR	CDS	1,409	N/A
chr1	152,191,022	152,192,420	deletion	1,398	repetitive	HRNR	CDS	1,398	N/A
chr1	152,279,350	152,279,351	insertion	972	repetitive	FLG	CDS	972	99.99
chr1	162,838,488	162,838,489	insertion	112	repetitive	C1orf110	UTR	112	80.82
chr1	171,179,913	171,179,994	deletion	81	repetitive	FMO2	UTR	81	98.61
chr1	179,575,362	179,575,363	insertion	6,115	repetitive	TDRD5	CDS	6,115	85.29
chr1	200,882,967	200,882,968	insertion	6,189	repetitive	C1orf106	UTR	6,189	N/A
chr1	201,178,792	201,178,793	insertion	108	non-repetitive	IGFN1	CDS	108	99.88
chr1	201,178,926	201,179,036	deletion	110	non-repetitive	IGFN1	CDS	110	99.88
chr1	201,180,113	201,180,114	insertion	216	repetitive	IGFN1	CDS	216	99.88
chr1	207,250,003	207,250,004	insertion	39	repetitive	PFKFB2	CDS	39	15.76
chr1	213,002,370	213,013,667	deletion	11,297	repetitive	SPATA45	CDS	277	N/A

f. Mappability including structural variants

First, we evaluated the change in sensitivity as the difference in number of mapped sequences. Illumina sequences from CHM1 were mapped with BWA-MEM to both GRCh37 and the patched reference. The patched reference contains 9,235,195 bp of novel sequence from closed or reduced gaps and expanded STRs or ~0.3% of the human genome. We mapped an additional 339,635 reads to the patched reference (excluding alternate haplotypes, chrM, and chrY) with

97.46% of reads mapped overall (1,258,390,142 of 1,291,241,795 reads) while 97.41% of reads mapped to the unpatched GRCh37 (1,258,050,507 of 1,291,442,455 reads). This overall increase in mappability of 0.05% corresponds to the addition of ~0.3% of new sequence to the genome or an enrichment of 17%.We also called SNPs on the original GRCh37 reference and the patched GRCh37 using Freebayes (with --ploidy 1 --min-alternate-fraction 0.8). We identified 9,231 additional SNPs in the patched reference with 2,745,603 compared to 2,736,372 in the original reference. Of these 9,231 new SNPs in the patched reference, 4,332 (47%) map within novel sequences >76 bp long.

To gauge specificity of variant calling, we applied methods developed in a thorough study of variant calling accuracy described at [24]. We replicated this analysis by mapping our CHM1 Illumina reads to both GRCh38 and a patched reference containing all inserted sequences. We found that an additional 510,575 (0.04%) reads mapped to the patched reference versus GRCh38. We then took advantage of the haploid nature of a complete hydatidiform mole, where no heterozygous SNPs should be found, to provide an estimate of false positive SNPs. Using the same analysis as described above, we found 804 heterozygous SNP calls in GRCh37 and 765 (4.9% decrease) in the patched reference. Of the 804 calls, 12 were unique to the patched reference and 51 unique to GRCh37. It is known that low complexity sequences are a source of false-positive heterozygous SNP calls in CHM1. Of the 804 false positive calls in GRCh37, 209 (~25%) are in low complexity sequences, while almost all (49 out of 51) of the calls unique to GRCh37 are in low complexity sequences, indicating that the patched reference decreases false-positive SNP calls albeit modestly (Supplementary Figure 10).



Supplementary Figure 10. A Venn diagram of heterozygous calls between the patched reference and GRCh37. There are 753 shared calls, and the patched reference removes 52 heterozygous calls while adding 12.

g. Assembly deficiencies

i. Black-tag analysis

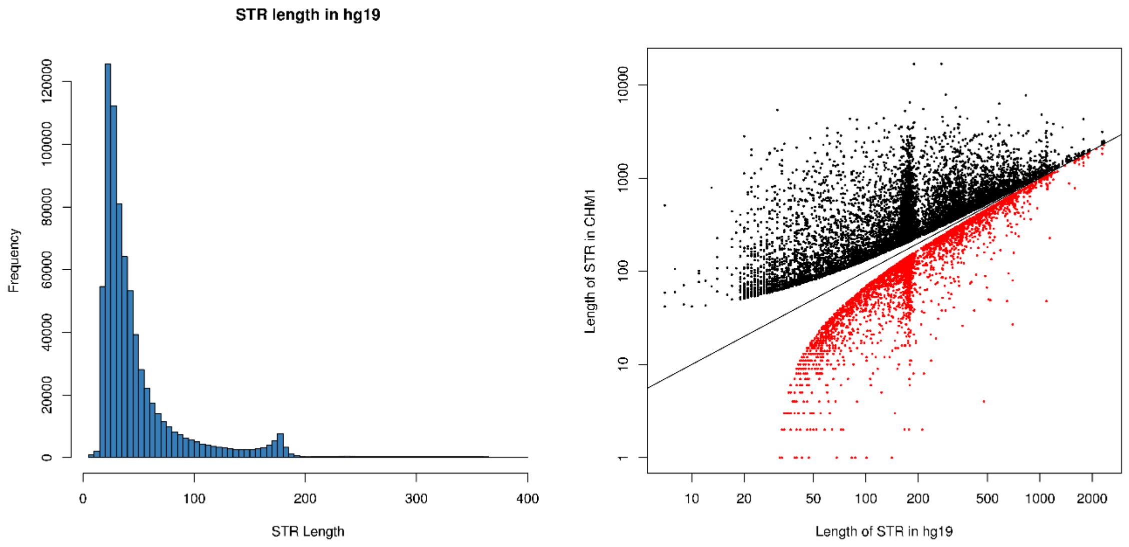
To investigate whether the bias towards insertions was due to an incomplete reference or errors in the assembly, the locations of insertions were compared to positions that had been flagged as problematic in GRCh37. For every assembly, the NCBI keeps a record of such annotated clone assembly problems, a.k.a. "black tags". The coordinates of black-tag annotations were obtained from the NCBI ftp://ftp.ncbi.nlm.nih.gov/pub/grc/human/GRCh37/MISC/annotated_clone_assembly_problems_GCF_000001405.25.gff3. We checked for black tag sites annotated within 100 bp+/- of a mobile element site (Supplementary Table 10). To check for enrichment, we shuffled an equivalent number of 200 windows across the genome and counted the number of intersected black-tag coordinates. We found VNTR, STR, unannotated, SVA, and complex events were the most enriched.

Supplementary Table 10. Number of "black tags" associated with repeat elements in GRCh37.

Repeat Type	Count	Expected	Standard-Deviation	Enrichment
STR	868	21.97	4.63	182.73
VNTR	280	10.08	3.14	85.96
SVA	59	1.68	1.28	44.78
Unannotated	139	8.77	2.95	44.15
Complex	77	4.07	2.02	36.10
HSAT	10	0.17	0.41	23.98
Alu STR	7	0.42	0.65	10.12
ALR	15	2.27	1.51	8.43
Singleton	6	0.62	0.79	6.81
MER	4	0.43	0.66	5.41
AluS	3	0.4	0.63	4.13
AluY	7	3.12	1.76	2.20
HERV	1	0.21	0.46	1.72
Alu mosaic	2	0.65	0.79	1.71
L1	1	0.39	0.63	0.97
L1P	1	0.47	0.69	0.77
L1HS	0	0.53	0.72	-0.74

ii. STR length bias of insertion sites

To further examine the possibility that the bias towards insertion is due to an incomplete reference, we examined the length distribution of STR insertions (CHM1) and deletions (GRCh37). While the distribution is generally uniform, there is a spike between 170 and 190 bp in length (Supplementary Figure 11). These sequences are more likely to have an increase of length of at least 30 bp: i.e., 1,054 out of 6,715 STRs (15.7%) that are expanded occur at loci annotated to be between 170 and 190 bp, although this represents only 2% of all STR loci in the genome. This particular length is not a previously described artifact for assembly and curation of STRs in the human genome. Nevertheless, we conclude that there has been a bias towards assembling long STRs into collapsed 170-190 bp sequences because STR insertion sequences are validated: using raw Sanger reads (98% or 88/90), exist in additional diploid genomes, and have a sevenfold increase in black tag annotations for this particular length.



Supplementary Figure 11. Length biased STR expansions. (*left*) A distribution of STR lengths in GRCh37 for STRs less than 400 bp. (*right*) The length of STR sequences in CHM1 with insertions (black) or deletions (red).

V. Validation experiments

We performed a series of validation experiments to confirm the sequence and organization of the gap closures and structural variants predicted by comparison of GRCh37 to CHM1 local SMRT assemblies. This included comparison against Sanger capillary-based BAC end-sequence data, Illumina WGS data, and finished sequence from clone libraries (see details for each class of variant below). For smaller variants such as STRs, for example, we validated 88/90 (97.8%) of inserts by comparison to fluorescence read-pair data generated from BAC inserts (

Supplementary Table 11). For intermediate-sized variants such as inversions and complex insertions, we initially sequenced clones with the Illumina Nextera protocol and assembled short reads with iCAS (Illumina clone assembly system) (ftp://ftp.sanger.ac.uk). We validated larger structural variants and gap closures primarily by targeted sequencing of large-insert clones (BAC and fosmid clone) selected based on clone end mappings to GRCh37. Similarly, we selected all clones that appeared to span an interstitial gap in GRCh37. Wherever possible we used previously sequenced clones from GenBank for validation (see Supplementary Table 12 for complete list). When structural variants were too large or repetitive to be adequately assembled with short reads, we sequenced the clone inserts with SMRT long reads using P4-C2 chemistry and assembled reads with HGAP and Quiver. Of the 56 structural variants and gap closures we attempted to validate, 53 events (95%) were confirmed. Within the primary classes of structural variation, all inversions, hard-stops, and complex events were validated by resequencing. For variants consisting of highly repetitive sequence, including STRs, tandem repeats, and gap closures, all events but one from each class were validated (

Supplementary Table 11). From all targeted sequencing experiments combined we estimate an overall validation rate of 97% of which only a fraction can be detected by next-generation sequencing (NGS). A detailed description of the sequenced clones is given in Supplementary Table 12. In addition to these validations, we also assessed additional deeply sequenced human genomes from the 1000 Genomes Project (1KG) to provide evidence that the novel insertion sequences were present in other human genomes. Depending on the class and size of the variants, our analyses indicate that there is evidence for 92-99% of these novel sequences or expansions in additional human genomes.

Supplementary Table 11. Summary of support for structural variants and gaps using BACs.

Event type	Selected	Validated	Illumina	PacBio	Capillary
Inversions	23	23	10	13	-
Hard-stops	11	11	-	8	3
Complex insertions	4	4	3	1	-
Gaps	3	2	-	2	-
STRs and tandem repeats	107	103	-	15	88

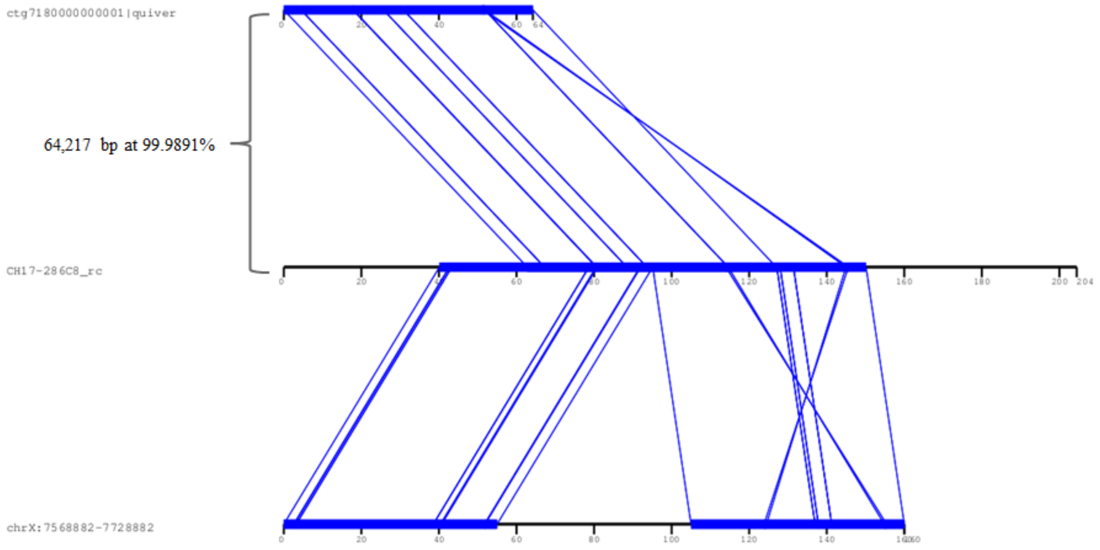
Supplementary Table 12. Detailed BAC validation list.

Region	Event type	Subtype	Clone	Clone accession	Clone location	BAC validation method	Validation status	Assembly status[a]
chr1:17615396-17618682	complex	insertion	CH17-158G11		chr1:17380035-17771911	Illumina	validated	G
chr1:24210820-24213064	complex	insertion	CH17-74O5		chr1:23995563-24358072	Illumina	validated	G
chr10:82972170-82975510	complex	insertion	CH17-284D21		chr10:82830487-83102106	Illumina	validated	G
chr10:98198734-98200614	complex	insertion	CH17-336J21		chr10:98013380-98422129	PacBio	validated	G
chr18:75721820-75771820	gap	closure	CH17-285F12		chr18:75694539-75965766	PacBio	not validated	G
chr4:59739333-59789333	gap	closure	CH17-271M9		chr4:59798061-59984180	PacBio	skip	W
chrX:10738674-10788674	gap	closure	CH17-258O10		chrX:10646088-10878276	PacBio	validated	G
chrX:7623882-7673882	gap	closure	CH17-286C8		chrX:7528762-7783204	PacBio	validated	G
chr1:13219912-13319912	gap	extension	CH17-38B2			PacBio	skip	F
chr10:125869472-125919472	gap	extension	CH17-365G21			PacBio	skip	T
chr19:7346004-7396004	gap	extension	CH17-267H8			PacBio	skip	T
chr21:9775437-9825437	gap	extension	CH17-302E12			PacBio	skip	W
chr4:9274642-9324642	gap	extension	CH17-243C7			PacBio	skip	F
chr5:17530657-17580657	gap	extension	CH17-285B19			PacBio	skip	F
chr6:170279972-170329972	gap	extension	CH17-44E3			PacBio	skip	T
chr6:62128589-62178589	gap	extension	CH17-255I17			PacBio	validated	G
chr7:61460465-61510465	gap	extension	CH17-14H4			PacBio	skip	F
chr9:66863343-66913343	gap	extension	CH17-358L14			Sanger	skip	T
chr15:22212114-22262114	gap	no extension	CH17-8O12		chr15:22153994-22336853	PacBio	skip	G
chr2:89630436-89830436	gap	no extension	CH17-34I22			PacBio	skip	G
chr10:38818835-38868835	gap	no extension	CH17-310O3			Sanger	skip	G
chr1:248681462-248682462	hard-stop		CH17-437N20		chr20:8087897-8279815	PacBio	skip	W
chr10:5292221-5292222	hard-stop	deletion	CH17-256N15		chr10:5159214-5409777	PacBio	validated	G
chr11:55018038-55060603	hard-stop	deletion	CH17-358C18		chr11:54892455-55109587	PacBio	skip	F
chr12:8589800-8590800	hard-stop	deletion	CH17-431M4		chr12:8487172-8811841	PacBio	validated	G
chr12:8589800-8590800	hard-stop		CH17-13G11		chr12:8475441-8699411	PacBio	validated	G
chr15:22342677-22343677	hard-stop	near gap	CH17-8O12		chr15:22153994-22336853	PacBio	validated	G
chr15:22342677-22343677	hard-stop	deletion	CH17-117D22		chr15:22210115-22490189	PacBio	validated	F
chr17:21684054-21685054	hard-stop	near gap	CH17-77D5		chr17:21528266-21758583	PacBio	skip	T
chr18:76124473-76139544	hard-stop		CH17-300G7		chr18:76000838-76267953	PacBio	validated	G
chr19:7304391-7305391	hard-stop		CH17-267H8		chr19:7207119-7509603	PacBio	skip	T
chr22:23851330-23852330	hard-stop		CH17-429E15		chr22:23664425-23859130	PacBio	validated	G
chr3:75998876-75999876	hard-stop		CH17-281F7		chrX:115730236-115965151	PacBio	skip	F, W
chr4:75492246-75493246	hard-stop	near gap	CH17-287M22		chr4:75471486-75683129	PacBio	validated	G
chr5:70390142-70391142	hard-stop		CH17-335F20		chr5:69300867-69488133	PacBio	skip	F, W
chr5:70390142-70391142	hard-stop		CH17-336P3		chr5:68821376-69182146	PacBio	skip	W
chr7:142104373-142105373	hard-stop	near gap	CH17-94L21		chr7:141973843-142275609	PacBio	skip	W
chr7:56768923-56769923	hard-stop	deletion	CH17-98H9		chr15:77090377-77303114	PacBio	skip	W
chr9:44221701-44316164	hard-stop		CH17-206M11		chr9:44106758-44309004	PacBio	skip	W
chr12:9631561-9632561	hard-stop		CH17-138H12		chr12:9544671-9853344	Sanger	validated	G
chr17:34815453-34816453	hard-stop		CH17-257H10		chr17:34740128-34967022	Sanger	validated	G
chr17:77629408-77630408	hard-stop		CH17-351M24		chr17:77492599-77709228 chr17_gl000204_random:30427-81027	Sanger	validated	G
chr1:187466259-187466927	inversion		CH17-19N16			Illumina	validated	G
chr10:47020750-47061230	inversion		CH17-351H10			Illumina	validated	G
chr10:67428466-67428898	inversion		CH17-60P13			Illumina	validated	G
chr10:75419516-75420109	inversion		CH17-77M7			Illumina	validated	G
chr14:65842142-65843534	inversion		CH17-106K2			Illumina	validated	G
chr14:93925872-93926772	inversion		CH17-269J10			Illumina	validated	G
chr16:85188070-85190447	inversion		CH17-3D1			Illumina	validated	G
chr21:27373785-27375082	inversion		CH17-480C10			Illumina	validated	G
chr3:44740244-44743033	inversion		CH17-233K1			Illumina	validated	G
chr6:130846031-130854451	inversion		CH17-330F8			Illumina	validated	G
chr1:26959293-26981981	inversion		CH17-68I14		chr1:34493761-34688776	PacBio	skip	W
chr1:44058931-44060272	inversion		CH17-319G7		chr1:43879215-44152216	PacBio	validated	G
chr12:12543302-12547878	inversion		CH17-320A7		chr12:12455369-12668518	PacBio	validated	G
chr17:5884492-5888092	inversion		CH17-278I9		chr17:5713803-6000614	PacBio	validated	G
chr2:138997306-139021398	inversion		CH17-215F16		chr2:138840363-139067881	PacBio	validated	G
chr4:88847164-88861762	inversion		CH17-144M16		chr4:88684414-88892793	PacBio	validated	G
chr6:107166965-107173118	inversion		CH17-298E21		chr6:106967566-107245965	PacBio	validated	G
chr6:169091697-169096586	inversion		CH17-249J3		chr6:168951727-169217949	PacBio	validated	G
chr7:107055748-107066461	inversion		CH17-150B4		chr7:106929375-107142406	PacBio	validated	G
chr7:40878739-40881087	inversion		CH17-389P4		chr7:106905699-107015733	PacBio	skip	W
chr8:6152078-6160622	inversion		CH17-9E4		chr8:5987464-6203774	PacBio	validated	G
chrX:45544532-45554400	inversion		CH17-280N1		chrX:45434062-45679897	PacBio	validated	G
chrX:45809535-45831726	inversion		CH17-47O16		chrX:46684164-46912988	PacBio	skip	T, W
chrX:6136278-6139154	inversion		CH17-285M6			PacBio	validated	G

[a] G = good, F = fragmented, T = truncated, W = wrong

a. Gap validations

We identified 17 BACs that appeared to span gaps from BAC end sequence (BES) mappings to GRCh37. We acquired sequence for two of the BACs that had previously been assembled from Sanger sequencing, sequenced the remaining 15 BACs using PacBio P4-C2 chemistry, and assembled the PacBio sequences with HGAP/Quiver. Of the 17 total BACs, 7 assembled into a single contig representing the complete insert with three spanning gap closures, one spanning a gap extension, and three spanning regions without gap extensions. Of the remaining 10 BACs, 4 assembled into truncated versions of the original insert (<=100 kbp of total sequence), 4 assembled into multiple contigs with signatures of collapsed duplications, and 2 did not span the expected gap region. Indeed, BACs selected for validation of gaps and hard-stop events near segmental duplications were significantly enriched for truncated assemblies. Of the 32 BACs selected for validation of gaps or hard-stops, 5 were truncated. Of all 214 CH17 BACs we have sequenced to date, only 11 have been truncated, including those sequenced for gaps and hard-stops. This enrichment is significant by Chi-square test (p = 0.0058; Chi-square = 7.605). The four BACs with collapsed duplications correspond to regions where no gap closure or extension assemblies could be generated. An example of the alignments of a BAC to a closed gap is shown in Supplementary Figure 12. Of the three BACs we sequenced that completely spanned a gap, two confirmed the content of the gap closures with an overall alignment identity of 99.93% and 99.99%, respectively (Supplementary Table 12). The third BAC aligned with the gap closure assembly at 98.85% identity due to a 699 bp TGG/TGA insert in the closure sequence. Without this single insertion, the alignment identity between the remaining sequences was 99.99%.



Supplementary Figure 12. Validation of a gap closure on chrX by BAC sequencing. The first pairwise alignment between the *de novo* local assembly gap closure sequence at the top and the BAC in the middle shows the concordance between the local assembly and the BAC. The second alignment between the BAC and GRCh37 sequence at the bottom shows the gap in GRCh37 and the closure of that gap in the BAC.

Supplementary Table 13. Gap validation alignments between novel gap sequences and their corresponding BACs.

Gap type	Region	Validated by	Alignment length (bp)	% identity	% identity (unique events)
closure	chr18_75716820_75776820	CH17-285F12	62,883	98.8523	99.9865
closure	chrX_7618882_7678882	CH17-286C8	64,217	99.9891	99.9922
closure	chrX_10733674_10793674	CH17-258O10	57,404	99.9251	99.9268
extension	chr6_62128589_62178589	CH17-255I17	27,182	99.6912	99.6948

We attempted to determine the potential cause of the five truncations by comparing the sequence content of the truncated clones and the corresponding sequences present in the CHM1 assembly. We identified two different patterns of truncations: 1) long (>1 kbp) simple or tandem repeats with putative toxicity for bacteria and 2) segmental duplications adjacent to centromeric or alpha satellite repeats. Three of the five BACs overlapped regions that contain >1 kbp simple or tandem repeats. Two of these three repeat sequences were classified as potential toxins for bacteria by BTXpred[25] while the remaining sequence was a 3 kbp run of simple repeats that was assembled in GRCh38 by a WI-2 fosmid (AC174061.2). Notably, for the truncated clone CH17-267H8, we identified a 1,407 bp run of simple repeats in the corresponding gap assembly from CHM1 WGS that contains six instances of the motif “TTATCACCA”. This sequence is almost identical to the *E. coli* DnaA box motif “TTATCCACA” which is known to inhibit replication in bacteria that contain the same sequence as an insert[26]. Although the remaining two truncated BACs map completely within segmental duplications, the read depth profiles of our PacBio assemblies do not correspond with known patterns of collapsed segmental duplications. Thus, these BACs are completely missing DNA from the original insert and their assembled insert sequences are likely not misassemblies due to high-identity tandem duplications. In addition to searching for potentially toxic sequences within the expected BAC insert, we investigated the repeat content at the breakpoints of each truncated BAC alignment against GRCh37 or the corresponding CHM1 whole genome shotgun gap assembly. Repeats at the breakpoints fell into the classes of SINE/Alu, LTR/ERVL and ERV1, LINE/L2, and CER satellites. There was no clear pattern of these repeats at the breakpoints to indicate a mediating mechanism for truncated inserts.

Although 23 gaps had adjacent patches in GRCh37.p13, not all gaps with patches were closed in GRCh38 and some gaps without patches were closed in GRCh38. Note CHM1-derived sequences have already been used to fill gaps within the human reference genome. To systematically assess how many gap closure regions had already been fixed in GRCh38, mapped the novel sequences from GRCh37 gap closure assemblies to GRCh38 with BLASR and identified closures for which at least 99% of the sequence were aligned with >=99% sequence identity. Of the 48 total non-zero-sized closures, we identified 22 closures (46%) with full-length, high-identity alignments in GRCh38 with a median identity of 99.8%. Six of these regions still have small annotated gaps in GRCh38 while 16 of the 22 have been completely resolved in GRCh38. Based on this analysis, we anticipated 26 annotated gaps in GRCh38 that could potentially be closed by our approach. By aligning novel closures from GRCh37 and GRCh38 to each other, we confirmed that 15 of the 31 GRCh38 closures were also present in the GRCh37 closures.

Using a unique k-mer analysis, we found evidence of 97.2% of the gap sequences in other human genome sequence data and, thus, were not an artifact specific to the hydatidiform mole.

b. Presence of novel sequence in additional genomes

To examine whether or not insertion sequences were present in other individuals, we developed an *in silico* genotyping assay where we counted the occurrences of 30-base sequences found in the inserted sequences and not the GRCh37 reference in all reads from high-coverage sequencing of a diversity panel with 28 individuals sequenced on the Illumina platform and the CHM1 Illumina reads produced in this study. We genotyped the 527 complex events greater than 1 kbp. Using this approach, we verified that 458 of the 527 genotyped complex insertions as present in the CHM1 Illumina read dataset and a total of 484 insertions in at least one of the datasets. Of the 43 sites with no support in the diversity panel genomes, 42 had no coverage in the CHM1 Illumina sequence, indicating a bias against sequencing these regions with this technology. 40% (218/527) of these events were polymorphic (Extended Data Fig. 3). Figure 2 (bottom) shows examples of polymorphic repeat mosaics. Many (359/1115) of the mosaic insertions contain the same mobile element flanking the inserted sequence.

To distinguish between mosaic insertion sequences and incomplete regions of the reference, we aligned queries generated from the insertion sequence plus 4 kbp 5' and 3' of the insertion site to the chimpanzee genome (panTro4). A total of 356 of the autosomal insertions were found to have high identity matches in the panTro4 reference spanning at least 80% of the insertion, indicating the sequences are either misassemblies in the human genome or sequence polymorphisms in GRCh37. Examples of these are shown in Extended Data Fig. 3, bottom. Insertion genotypes that are fixed in the population give additional evidence that the sequences are deficiencies in the reference rather than private deletions. For the complex events, 67.5% (309/458) events that may be assessed by Illumina sequencing are fixed in the population based on our analysis of 28 unrelated PCR-free diploid genomes. Of these, 39% (119/309) also show partial or complete alignment against.

In order to assess the diversity of novel and complex sequences identified from resequencing of CHM1, we assayed 23 PCR-free Illumina-sequenced genomes (1KG) in addition to Illumina sequence generated for CHM1 in a subset of the insertion sequences. In total we assessed 110 gap extensions and closures, 339 STRs, and 98 tandem repeats (547 loci total). The approximate copy number of each locus and flanking sequence was estimated by mapping reads, subdivided into their 36 bp constituents to target loci and their flanks using the mrsFAST[27] read aligner. Mobile elements (such as LINEs and SINEs) were masked prior to mapping. Each genome was additionally mapped to the human reference genome (GRCh37) from which a GC-sequencing bias correction factor was generated in addition to a copy number estimation calibration curve based on regions of known copy[28]. Reads mapping to each loci were finally corrected for GC-associated sequencing biases and the copy number was estimated in adjacent windows of 100 bp of unmasked sequence. After masking mobile elements, nine gap extensions and closures were excluded as they had <100 bp of non-repetitive sequence. We found that the vast majority of sequences were present in most or all of the individuals assessed, 99.3% (534/538). Of these, an appreciable fraction were variable in their total copy number among the individuals we assessed; 54.7% (292/534) exhibited a $\log_2$ ratio >1.0 when compared to an arbitrary reference individual (see below).

c. STR validations

The sequence of STR insertions was first compared to those of well-characterized loci. We examined the consensus sequences at STR expansion disorder loci from 31 STR loci and one VNTR locus with at least 10 repeat units[29]. The consensus sequence of 15 of the loci had no differences from the reference, and the remainder showed polymorphisms with exact expansion or deletion of the known repeat units. For example, the consensus sequence of the CGG STR motif in FMR1 contains an exact match to the reference punctuated by an insertion of (GGC)9GGA (Supplementary Figure 7), consistent with known FMR1 haplotype structure. The VNTR near INS-IGF2 was divergent from the reference, but with a closely repeated core structure.

As a second validation approach, we compared the Sanger sequences from BAC end sequencing of CHM1 (CHORI-17) to the STR consensus sequences of the PacBio-based assemblies. We found a total of 90 reads with STR insertions at 83 loci, with an average insertion length of 53.56. When these reads were mapped to the consensus sequences of assemblies overlapping these loci, 88 reads map without indels at the STR sequence. The remaining two show a 14-base AT insertion. Additionally, although the comparison of consensus sequences to the sequenced BACs from CHM1 indicate a bias towards errors in dinucleotide repeats, the bias is 13 times greater towards deletion.

We applied computational genotyping to the STR and VNTR insertion sequences with at least 1 kbp of inserted sequence for a total of 788 sites. Because the inserted sequences are not perfectly replicated patterns, it is possible to find unique k-mers distinguishing the inserted STR sequences from the background of the reference sequence. Each STR insertion sequence contained on average 1582 30-base sequences not represented in the reference. We were able to confirm 463 insertions in the CHM1 sample with Illumina data and 599 in at least one of the 1KG samples.

Similar to the complex insertions, we find that the STR sequences are also polymorphic in the population, with 27% (164/599) having at least one sample with the site entirely absent.

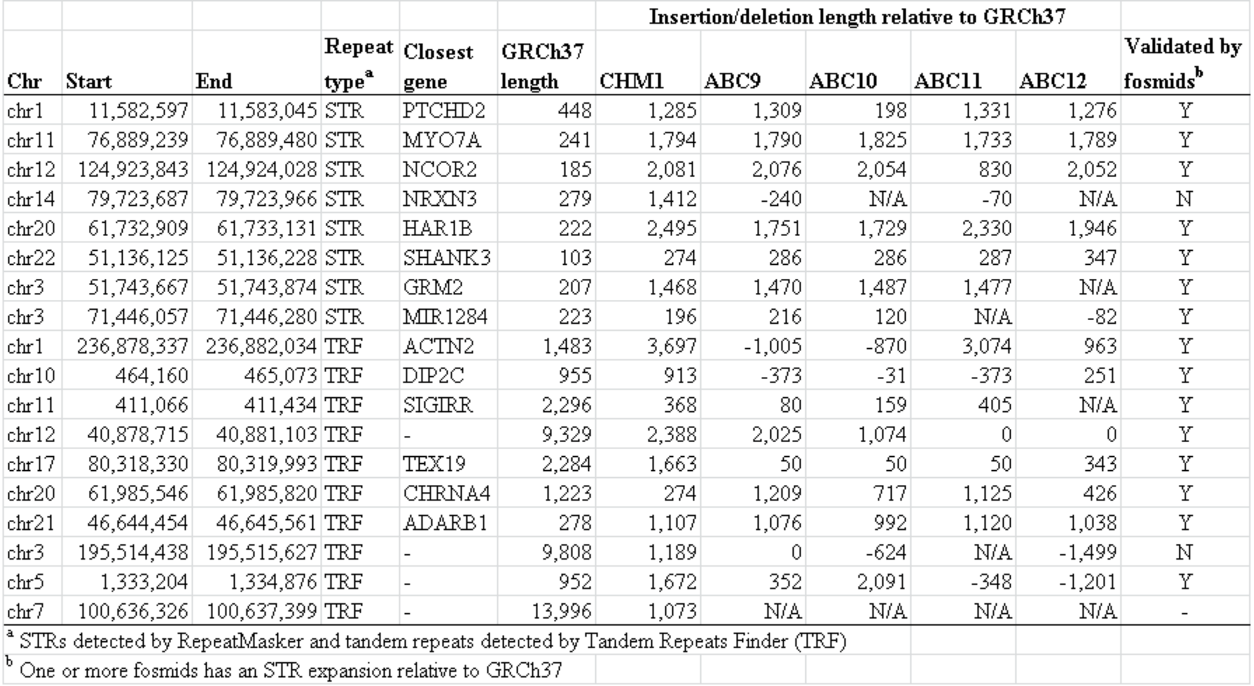
d. Fosmid insert validations

We also attempted to validate the expansion of STRs and tandem repeats as seen in CHM1 by estimating the prevalence of these expansions by sequencing fosmids from individuals of Japanese (ABC9; NA18956), Nigerian (ABC10; NA19240), Chinese (ABC11; NA18555), and European American (ABC12; NA12878) descent[30]. We identified all STRs and tandem repeats with at least one completely overlapping fosmid in each sample library. From this set, we selected eight STRs and 10 tandem repeats based on their presence in coding exons, 5' or 3' UTR exons, transcribed non-CDS introns, CDS introns, or adjacency to a gene. Note, we could not recover fosmids from libraries ABC10 and ABC12 for one of the eight STRs that was expanded near the gene *NRXN3*. For this STR, we only sequenced fosmids from ABC9 and ABC11. In addition to the 18 regions expanded in CHM1 relative to GRCh37, we selected fosmids from five control regions corresponding to STRs in the genes *MUC5AC*, *MUC5B*, *LPA*, *FMR1*, and the ALS-linked *C9orf72*.

We created 2-3 pools per sample with 7-12 fosmids per pool and sequenced one SMRT cell per pool with PacBio P4-C2 chemistry (see Methods for library prep). We assembled each pool's SMRT cell with HGAP and Quiver and identified variants using the pipeline resequencing and assembly pipeline used to produce the structural variant callset in CHM1.

An expansion of STR sequences of similar length as the expansion seen in CHM1 was observed in all samples at seven out of the eight STR loci, with the last being expanded in CHM1 and contracted in ABC9 and ABC11 (Supplementary Table 14). Whereas the STR-inserted sequences stratify populations and likely signify underrepresentation in the GRCh37 reference, the tandem repeat sequences contained four loci with insertions in all populations, five loci with both insertions, and either deletions or sequences invariant to the reference. The remaining locus was not fully contained in any fosmid making estimation of the tandem repeat size impossible.

Supplementary Table 14. Validation of CHM1 STR expansions in fosmid libraries of four samples.



						Insertion/deletion length relative to GRCh37					
Chr	Start	End	Repeat type[a]	Closest gene	GRCh37 length	CHM1	ABC9	ABC10	ABC11	ABC12	Validated by fosmids[b]
chr1	11,582,597	11,583,045	STR	PTCHD2	448	1,285	1,309	198	1,331	1,276	Y
chr11	76,889,239	76,889,480	STR	MYO7A	241	1,794	1,790	1,825	1,733	1,789	Y
chr12	124,923,843	124,924,028	STR	NCOR2	185	2,081	2,076	2,054	830	2,052	Y
chr14	79,723,687	79,723,966	STR	NRXN3	279	1,412	-240	N/A	-70	N/A	N
chr20	61,732,909	61,733,131	STR	HAR1B	222	2,495	1,751	1,729	2,330	1,946	Y
chr22	51,136,125	51,136,228	STR	SHANK3	103	274	286	286	287	347	Y
chr3	51,743,667	51,743,874	STR	GRM2	207	1,468	1,470	1,487	1,477	N/A	Y
chr3	71,446,057	71,446,280	STR	MIR1284	223	196	216	120	N/A	-82	Y
chr1	236,878,337	236,882,034	TRF	ACTN2	1,483	3,697	-1,005	-870	3,074	963	Y
chr10	464,160	465,073	TRF	DIP2C	955	913	-373	-31	-373	251	Y
chr11	411,066	411,434	TRF	SIGIRR	2,296	368	80	159	405	N/A	Y
chr12	40,878,715	40,881,103	TRF	-	9,329	2,388	2,025	1,074	0	0	Y
chr17	80,318,330	80,319,993	TRF	TEX19	2,284	1,663	50	50	50	343	Y
chr20	61,985,546	61,985,820	TRF	CHRNA4	1,223	274	1,209	717	1,125	426	Y
chr21	46,644,454	46,645,561	TRF	ADARB1	278	1,107	1,076	992	1,120	1,038	Y
chr3	195,514,438	195,515,627	TRF	-	9,808	1,189	0	-624	N/A	-1,499	N
chr5	1,333,204	1,334,876	TRF	-	952	1,672	352	2,091	-348	-1,201	Y
chr7	100,636,326	100,637,399	TRF	-	13,996	1,073	N/A	N/A	N/A	N/A	-

[a] STRs detected by RepeatMasker and tandem repeats detected by Tandem Repeats Finder (TRF)

[b] One or more fosmids has an STR expansion relative to GRCh37

e. Validation of hard-stops, inversions, and complex insertions

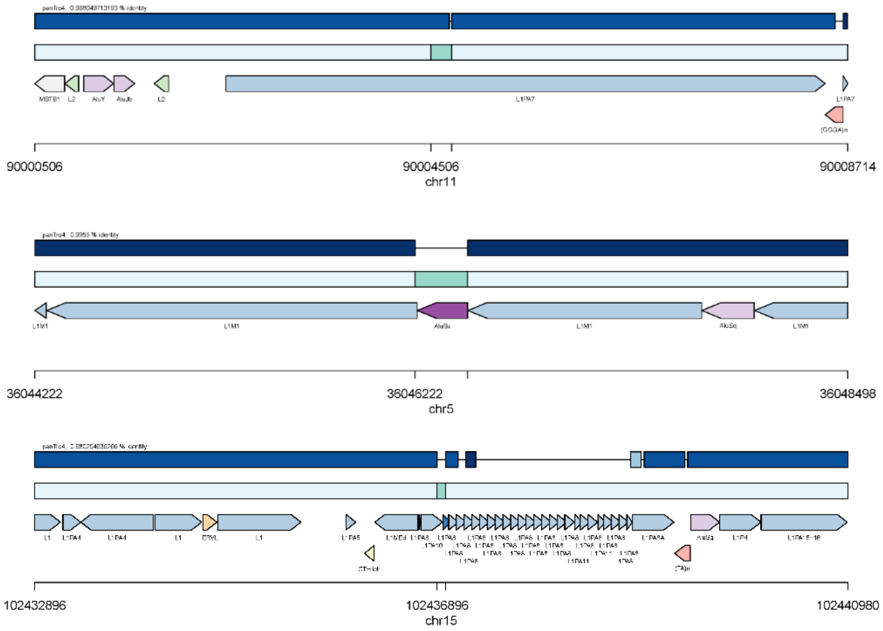
To confirm the structural variants (inversions, hard-stops, and complex) we detected in CHM1 with long reads, we identified BACs spanning structural variants based on BES alignments from CHM1's BAC library (CH17) to GRCh37. BACs were initially sequenced with the Nextera-Illumina protocol (250 bp reads) and short reads were aligned to GRCh37 to confirm the coordinates from the BES alignments. For a subset of inversion and complex repeat insertion events, we were able to *de novo* assemble Illumina reads into sufficiently complete contigs using iCAS (ftp://ftp.sanger.ac.uk/pub/badger/aw7/icas_README) to confirm the events (Supplementary Table 8). For the remainder of the structural variants, we sequenced BACs with PacBio technology using the P4-C2 polymerase and chemistry combination and assembled sequences with HGAP and Quiver. We identified 21 inversions, 11 hard-stops, and 4 complex events with overlapping CH17 BACs. We validated all inversions with 10 iCAS assemblies and 11 PacBio assemblies (Supplementary Table 8, Extended Data Fig. 4). Additionally, 5 out of 21 inversions (24%) were detected by VariationHunter[31] (Supplementary Table 15). We validated all 11 hard-stops with eight PacBio-sequenced BACs and three BACs previously sequenced with capillary technology. Finally, we confirmed the presence of all three complex events with iCAS assemblies.

Supplementary Table 15. High-confidence inversion calls made by VariationHunter.

Chr.	start	end	Size	# reads	score	PacBio validated
chr1	92131473	92132889	1416	49	2.87755	
chr2	72440267	72441365	1098	17	0.58824	N
chr2	131037361	132130290	1092929	13	1.07692	N
chr3	44740921	44742580	1659	27	1.7037	N/A
chr4	88847157	88858902	11745	30	2.1	Y
chr6	130847987	130852497	4510	34	1.41177	Y
chr6	167582381	167802060	219679	11	4.36364	Y
chr7	40879129	40880716	1587	27	1.51852	N/A
chr12	12544637	12546802	2165	40	1.6	Y
chr14	65842412	65843616	1204	30	2.56667	N
chr14	67170261	67171899	1638	13	2.15385	N
chr16	85188543	85190105	1562	33	3.60606	N
chr21	27373957	27375021	1064	25	1.8	Y
chrX	6136864	6138390	1526	12	1.83333	N

f. Comparison to panTro4

Although certain mobile elements L1P and AluS are no longer active, 236 insertion events were nonetheless observed in the MEI callset in the CHM1 callset when compared to GRCh37. To investigate the source of these insertions, we aligned the insertion sequence plus the flanking 4 kbp to the chimpanzee genome[19] and checked for the presence or absence of a contiguous match in chimpanzee. We found that 51% (55/108) of the AluS and 56% (72/128) L1P sequences, respectively, had orthologous matches in the chimpanzee genome. To explain additional L1P and AluS insertions, we searched for tandem site duplications (TSDs) flanking the insertion site, reasoning that assembly methods will have difficulty correctly assembling mosaic sequences. We found that 77 of the AluS sequences showed a TSD flanking the insertion site (34 aligned to chimpanzee), or that the insertion itself was in tandem with an existing mobile element in the genome, and similarly 37 for L1P (29 aligned to chimpanzee). We found that an additional 63 L1P sequences were small fragments of L1P sequence inside annotated L1P repeats in the genome that are incomplete (Supplementary Figure 13). Furthermore, 15 of the events are within highly ordered tandem arrays of L1P insertions, as shown in Supplementary Figure 13.

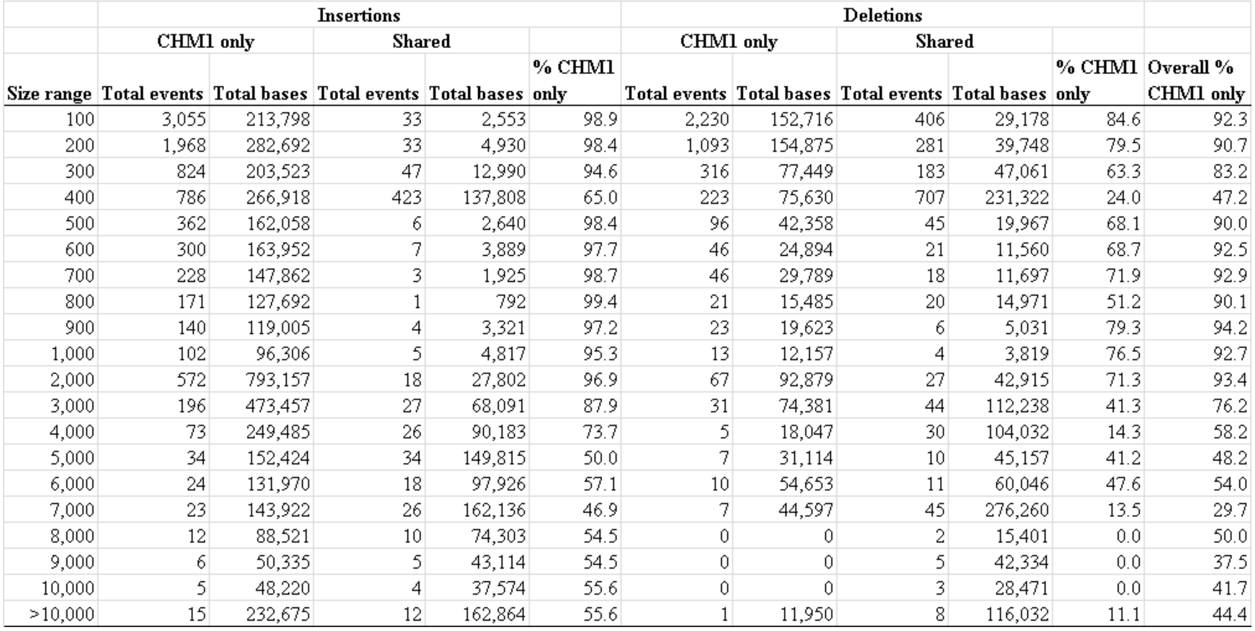


Supplementary Figure 13. Examples of AluS and L1P MEIs. For each panel, the top row (dark blue) shows the alignment to chimpanzee, the middle bar shows the reference human sequence (light teal) and inserted sequence (teal), and the bottom rows show the repeat annotation. (top) An insertion of AluS with TSDs. (middle) An example of a recorded L1P insertion that is a fragment of an existing repeat. (bottom) An example of a tandem array of L1P elements.

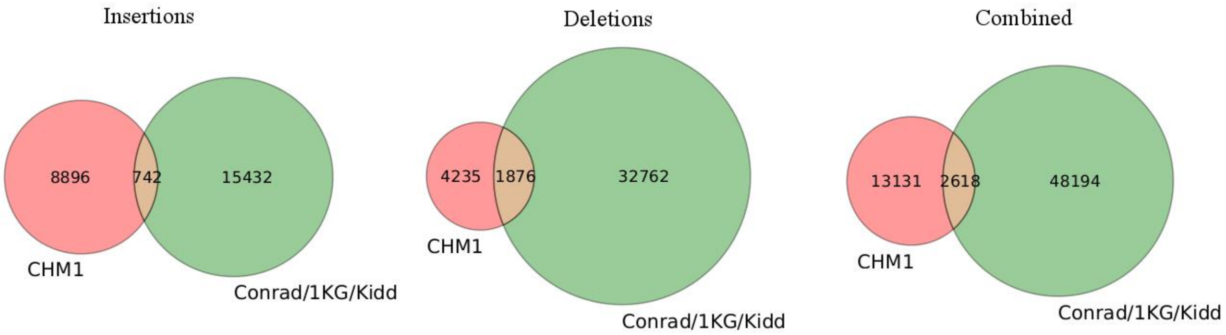
VI. Comparison to other structural variation data sets

We assessed the proportion of novel structural variants from CHM1 SMRT WGS by comparing all CHM1 variant calls with previously published insertions and deletions detected by array CGH[32,33], the fosmid structural variation sequencing project[34,35], and from the 1KG[33,36]. We required a 50% reciprocal overlap between shared calls. All calls were filtered to exclude events smaller than 50 bp as well as events mapping within 5 Mbp of a centromere and 150 kbp of the telomere. The overlap of insertion, deletion, and all calls is shown in Supplementary Figure 14. Of the 15,749 euchromatic insertions and deletions detected in CHM1, 83% had not been previously reported. The majority of novel events occurred between 50-300 bp in length (Supplementary Table 16 and Figure 2). The effect was most pronounced for insertion where 92% of all differences had not been previously reported, in contrast to deletions where 69% of the events were novel. As expected, the frequency of events decayed exponentially with increasing structural variant length with two peaks corresponding to Alu and L1 MEI insertions. A noticeable reduction in novelty is observed at 300 bp likely because of dedicated efforts to map MEI elements. As the size of events increase, the overlap with previously published calls drastically increases. Above 2 kbp and 4 kbp, approximately 50% of calls had been observed previously for deletions and insertions, respectively.

Supplementary Table 16. Summary of insertions and deletions novel to CHM1 calls and shared between CHM1 calls and a combined callset from Conrad et al. 2009, 1KG, and Kidd et al. 2010.



	Insertions					Deletions					
	CHM1 only		Shared			CHM1 only		Shared			
Size range	Total events	Total bases	Total events	Total bases	% CHM1 only	Total events	Total bases	Total events	Total bases	% CHM1 only	Overall % CHM1 only
100	3,055	213,798	33	2,553	98.9	2,230	152,716	406	29,178	84.6	92.3
200	1,968	282,692	33	4,930	98.4	1,093	154,875	281	39,748	79.5	90.7
300	824	203,523	47	12,990	94.6	316	77,449	183	47,061	63.3	83.2
400	786	266,918	423	137,808	65.0	223	75,630	707	231,322	24.0	47.2
500	362	162,058	6	2,640	98.4	96	42,358	45	19,967	68.1	90.0
600	300	163,952	7	3,889	97.7	46	24,894	21	11,560	68.7	92.5
700	228	147,862	3	1,925	98.7	46	29,789	18	11,697	71.9	92.9
800	171	127,692	1	792	99.4	21	15,485	20	14,971	51.2	90.1
900	140	119,005	4	3,321	97.2	23	19,623	6	5,031	79.3	94.2
1,000	102	96,306	5	4,817	95.3	13	12,157	4	3,819	76.5	92.7
2,000	572	793,157	18	27,802	96.9	67	92,879	27	42,915	71.3	93.4
3,000	196	473,457	27	68,091	87.9	31	74,381	44	112,238	41.3	76.2
4,000	73	249,485	26	90,183	73.7	5	18,047	30	104,032	14.3	58.2
5,000	34	152,424	34	149,815	50.0	7	31,114	10	45,157	41.2	48.2
6,000	24	131,970	18	97,926	57.1	10	54,653	11	60,046	47.6	54.0
7,000	23	143,922	26	162,136	46.9	7	44,597	45	276,260	13.5	29.7
8,000	12	88,521	10	74,303	54.5	0	0	2	15,401	0.0	50.0
9,000	6	50,335	5	43,114	54.5	0	0	5	42,334	0.0	37.5
10,000	5	48,220	4	37,574	55.6	0	0	3	28,471	0.0	41.7
>10,000	15	232,675	12	162,864	55.6	1	11,950	8	116,032	11.1	44.4



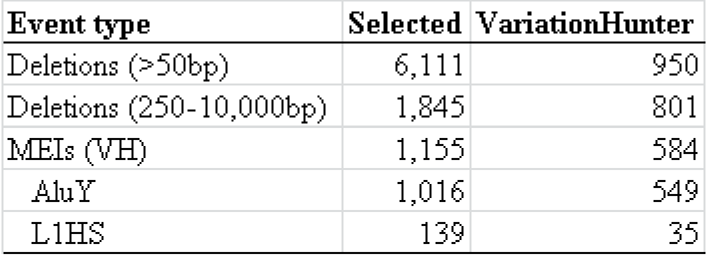
Supplementary Figure 14. Comparison of insertions and deletions with previously published callsets. Combined datasets from previous studies[15,30,32] are compared with callsets from CHM1.

a. Illumina vs. SMRT sensitivity analysis.

We compared the sensitivity of MEI and deletion detection between SMS and next-generation Illumina sequencing data based on an analysis of 41-fold sequence coverage data generated for CHM1. The analysis was performed to eliminate the possibility that the difference in sensitivity (see above) was a platform methodology and not due to the fact that different genomes were being compared. We compared our deletion and MEI calls with calls from VariationHunter. Of the 6111 euchromatic deletions ≥50 bp detected in CHM1 long reads, VariationHunter identified 950 overlapping events (16%; Supplementary Table 17). When we inspect CHM1 deletions in the size range where VariationHunter is most sensitive (250-10,000 bp), we find 801 of 1845

CHM1 deletions (43%) are supported. Additionally, VariationHunter supported 584 of the 1155 CHM1 MEI calls (51%) with 549 VariationHunter calls supporting 1016 CHM1 AluY calls (54%) and 35 supporting 139 CHM1 L1HS calls (25%). Finally, 6 of the 33 inversions detected in CHM1 long reads were also detected by VariationHunter. This low validation rate likely reflects the technological limitations of short read data. It is known that methods to detect MEIs using Illumina NGS reads avoid making calls for insertions inside existing repetitive regions[15,31]. Similarly, all of our inversions validated yet only 6 out of 33 (18%) could be detected by VariationHunter. Although 39% of our inversions have repeats at the breakpoints, only 1 out of the 6 inversions detected by VariationHunter (17%) had repeats at the breakpoints.

Supplementary Table 17. Support for deletions and MEIs from CHM1 by VariationHunter.



Event type	Selected	VariationHunter
Deletions (>50bp)	6,111	950
Deletions (250-10,000bp)	1,845	801
MEIs (VH)	1,155	584
AluY	1,016	549
L1HS	139	35

b. Illumina vs. SMRT MEI insertion complexity analysis.

Finally, we performed a site complexity analysis of annotated MEI loci by counting the repeat composition of the 1 kbp sequences 5' and 3' flanking AluY, L1, and SVA insertions in both the CHM1 sequencing data and insertion sites from low-coverage sequencing data from the 1KG[15]. Density plots of the insertion site complexity are shown in Figure 3b. The repeat content of the AluY insertions in CHM1 has a mean of 54% versus 48.3% in the 1KG AluY callset and a cumulative distribution function (CDF) that differs from random intervals with a p-value 0.02 versus the 0.09 in the 1KG callset using the Kolmogorov-Smirnov (KS) test statistic. A more drastic shift is seen for L1 and SVA insertions, where L1 insertion sites in CHM1 have a repeat content of 59% in CHM1 and 39% in 1KG, and 76% versus 50% for SVA insertions. The CDFs for both L1 and SVA insertion site complexity differs between CHM1 and 1KG with $p < 2x10^{-16}$ using the KS test.

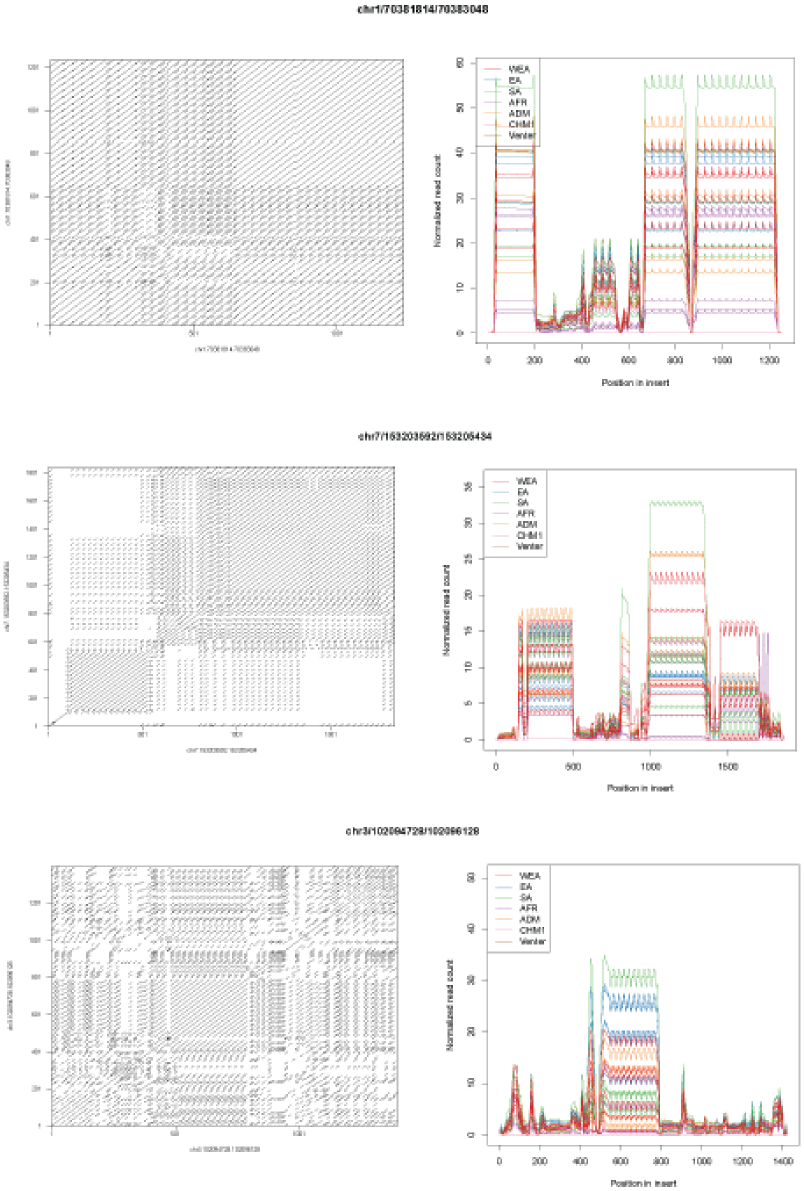
VII. Tandem repeat copy number and structure

a. Copy number variation analyses

In order to assess copy number variation of the novel and complex sequences identified from resequencing of CHM1 for expansion or contraction of repeats, we assayed 23 PCR-free Illumina-sequenced genomes (1KG) in addition to Illumina sequence generated for CHM1 using read depth information as an indication of an increase or decrease in copy number of a repeat unit[28]. In total we assessed 101 gap extensions and closures, 338 STRs, and 97 variable number tandem repeats (536 loci total) for copy number variation (

Supplementary Table 18). 90% (483/536) of the target loci had sufficient read depth within Illumina WGS datasets where copy number could be estimated. A subset of sites (n = 50) where our copy number estimations appeared to fail specifically over the targeted locus compared to flanking sequence had a significantly increased GC-content distribution compared to sites that did work ($p < 2.2x10^{-16}$, two-sample t-test). We assessed the remaining 483 regions specifically to determine if the lengths of these loci varied among the individuals assessed. 54.7% of loci (264/483) exhibited at least one individual with a $\log_2$ ratio >1.0 (or <-1.0) when compared to a reference individual genome selected at random (Supplementary Figure 16).

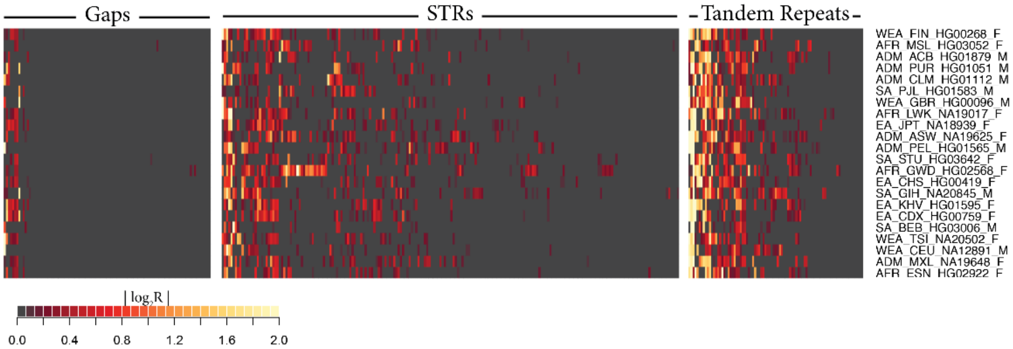
The approximate copy number of each locus and flanking sequence was estimated by mapping reads, subdivided into their 36 bp constituents to target loci and their flanks using the mrsFAST read aligner. Mobile elements (such and LINEs and SINEs) were masked prior to mapping. Each genome was additionally mapped to GRCh37 from which a GC-sequencing bias correction factor was generated in addition to a copy number estimation calibration curve based on regions of known copy number[28]. Reads mapping to each locus were finally corrected for GC-associated sequencing biases. Locus-specific copy number was estimated in adjacent windows of 100 bp of unmasked sequence using only reads mapping to singly unique nucleotide k-mers (SUNKs) specifically tagging the assayed locus. After masking mobile elements, 11 regions were excluded as they contained <100 bp of non-repetitive sequence.



Supplementary Figure 15. STR and VNTR variation by read depth. Examples of expanded STR sequences with variable sequence length in the population. (*left column*) Self dotplots of the insertion sequences. (*right column*) Read depth profiles of 28 diverse genomes.

Supplementary Table 18. Variability of new sequences from CHM1 in expanded repeats and gap closures.

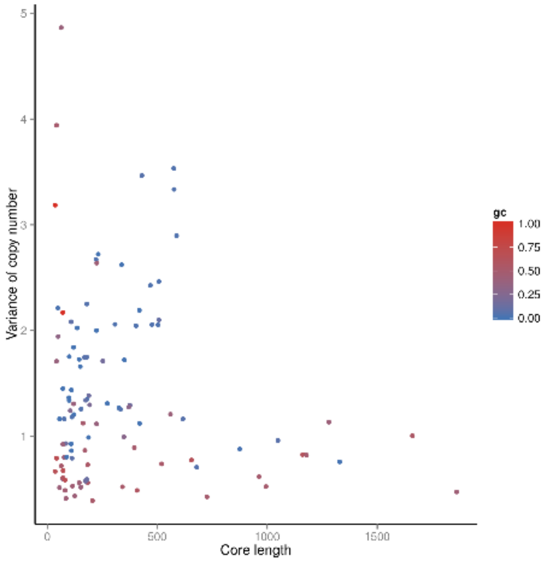
Type	Analyzed	Illumina sequenced	Variable	Max delta (depth)
VNTR	97	94	72 (77%)	221.1
STR	338	288	183 (63%)	123.8
Gap	101	101	9 (9%)	2,207.6
Total	536	483	264	2,552.4



Supplementary Figure 16. Copy number polymorphism of inserted sequences. Log_2 ratios of the copy number of targeted gaps, STRs and tandem repeats assessed for copy number from Illumina sequencing data. Tandem repeats show the most variability amongst individuals while gap extensions and fills show the least.

b. STR structure and composition

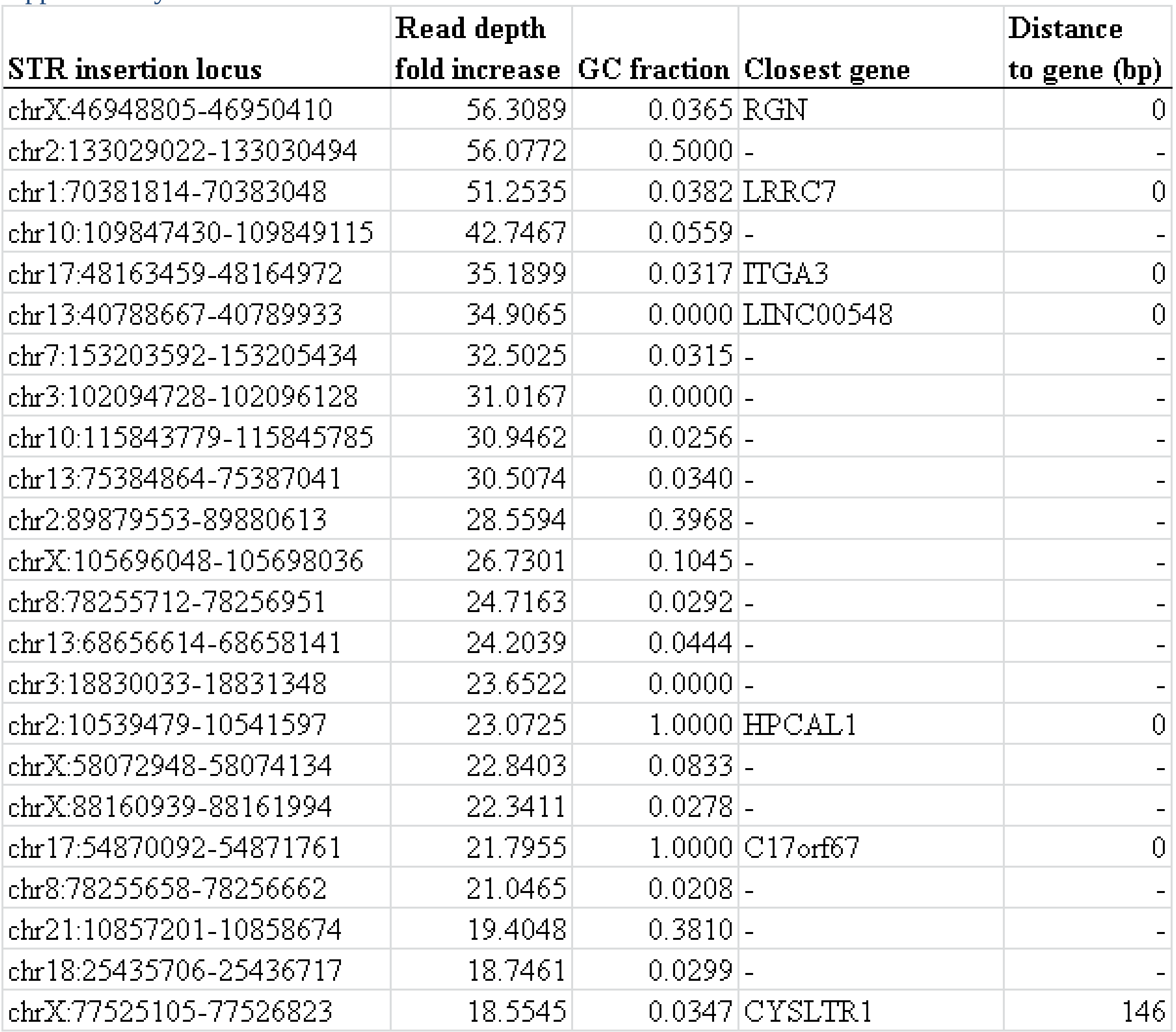
We examined the repeat structures and read depth profiles of computationally genotyped STR loci to decipher expansion of STR sequences based on the motif of the repeat unit. A common feature of the STR insertion sequences is repeat motif degeneracy, where the general nucleotide content of an STR is roughly consistent (e.g., GC rich), but the motif of the repeat unit changes across the sequence of the STR. An example of mosaic repeats are shown in Supplementary Figure 15 (left). It is possible for the copy number of different mosaic units within the same STR locus to expand independently (Supplementary Figure 15, right column). We determined the conserved core motifs as the longest region of each STR insertion repeated with at least 95% identity and ranked STR insertion loci according to the difference between the highest and lowest average read depth across region, as given in Supplementary Table 19. We observe that 16 of the top 20 most polymorphic STR sequences have GC composition less than 10%. In addition to the maximum copy number difference, we characterized the variance of STR copy number and found that while a small number of variable sequences are high GC, variable sequences tend to be low in GC composition, as shown in Supplementary Figure 17. There are 37 loci that have at least a repeat copy number change of at least 10—9 of which are located within genes (Supplementary Table 19) representing sites of potential genomic instability.



Supplementary Figure 17. STR variation by sequence composition. The variance of core repeat motifs is computed using read depth estimates of copy number of the motif with 28 genomes from the 1KG. The core length is the length of the conserved repeat core in CHM1, and the GC composition is the G+C fraction of the conserved repeat core in CHM1.

doi:10.1038/nature13907

Supplementary Table 19. STR insertions detected in CHM1 relative to GRCh37.



STR insertion locus	Read depth fold increase	GC fraction	Closest gene	Distance to gene (bp)
chrX:46948805-46950410	56.3089	0.0365	RGN	0
chr2:133029022-133030494	56.0772	0.5000	-	-
chr1:70381814-70383048	51.2535	0.0382	LRRC7	0
chr10:109847430-109849115	42.7467	0.0559	-	-
chr17:48163459-48164972	35.1899	0.0317	ITGA3	0
chr13:40788667-40789933	34.9065	0.0000	LINC00548	0
chr7:153203592-153205434	32.5025	0.0315	-	-
chr3:102094728-102096128	31.0167	0.0000	-	-
chr10:115843779-115845785	30.9462	0.0256	-	-
chr13:75384864-75387041	30.5074	0.0340	-	-
chr2:89879553-89880613	28.5594	0.3968	-	-
chrX:105696048-105698036	26.7301	0.1045	-	-
chr8:78255712-78256951	24.7163	0.0292	-	-
chr13:68656614-68658141	24.2039	0.0444	-	-
chr3:18830033-18831348	23.6522	0.0000	-	-
chr2:10539479-10541597	23.0725	1.0000	HPCAL1	0
chrX:58072948-58074134	22.8403	0.0833	-	-
chrX:88160939-88161994	22.3411	0.0278	-	-
chr17:54870092-54871761	21.7955	1.0000	C17orf67	0
chr8:78255658-78256662	21.0465	0.0208	-	-
chr21:10857201-10858674	19.4048	0.3810	-	-
chr18:25435706-25436717	18.7461	0.0299	-	-
chrX:77525105-77526823	18.5545	0.0347	CYSLTR1	146

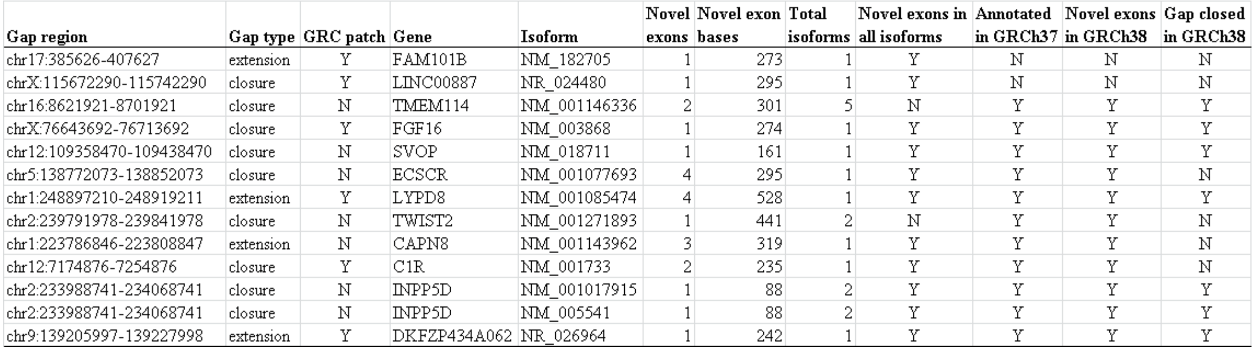
VIII. Functional sequence annotation

a. mRNA/EST analysis

To determine whether there were any previously undescribed exons in our gap closures and extensions, we selected cDNA transcripts from the RefSeq RNA database corresponding to genes that are annotated near our gap regions and searched our complete gap assemblies (including flanking reference sequence and novel gap sequence) for full-length transcript alignments using GMAP[37].

We identified eight gap closures (16%) and four extensions (7%) with putative additional exons inside novel sequence (Supplementary Table 20). One gene annotated inside a gap closure (*LINC00887*) and one inside extensions (*FAM101B*) were not previously annotated in those regions. We considered these to be false positives. The remaining annotations add 20 novel exons (2,972 bp). In all but two cases, the novel exons were present in all isoforms for the annotated gene. Novel exons in the gene *TMEM114* were only present in one of five isoforms and exons in the gene *TWIST2* were present in one of two isoforms. Inspection of GRCh38 for genes with novel exons revealed that all 20 novel exons had been added in GRCh38 although four of the 10 regions still had gaps in intronic sequence. The novel exon sequence in GRCh38 originated from a combination of fosmids (one ABC7, two ABC12, and one WI2 clones) and contigs from whole-genome assemblies of Venter (7), ABC12 (3), and RPCI-11 (1).

Supplementary Table 20. Genes annotated inside gap closures and extensions by RefSeq mRNA alignments with GMAP.



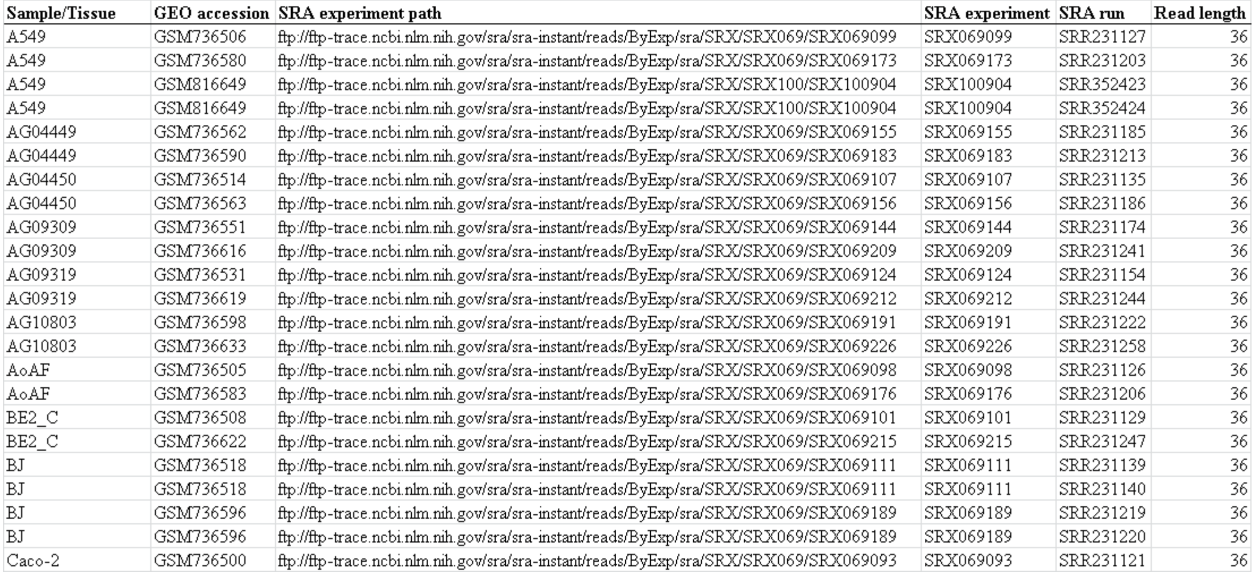
Gap region	Gap type	GRC patch	Gene	Isoform	Novel exons	Novel exon bases	Total isoforms	Novel exons in all isoforms	Annotated in GRCh37	Novel exons in GRCh38	Gap closed in GRCh38
chr17:385626-407627	extension	Y	FAM101B	NM_182705	1	273	1	Y	N	N	N
chrX:115672290-115742290	closure	Y	LINC00887	NR_024480	1	295	1	Y	N	N	N
chr16:8621921-8701921	closure	N	TMEM114	NM_001146336	2	301	5	N	Y	Y	Y
chrX:76643692-76713692	closure	Y	FGF16	NM_003868	1	274	1	Y	Y	Y	Y
chr12:109358470-109438470	closure	N	SVOP	NM_018711	1	161	1	Y	Y	Y	Y
chr5:138772073-138852073	closure	N	ECSCR	NM_001077693	4	295	1	Y	Y	Y	N
chr1:248897210-248919211	extension	Y	LYPD8	NM_001085474	4	528	1	Y	Y	Y	Y
chr2:239791978-239841978	closure	N	TWIST2	NM_001271893	1	441	2	N	Y	Y	N
chr1:223786846-223808847	extension	N	CAPN8	NM_001143962	3	319	1	Y	Y	Y	N
chr12:7174876-7254876	closure	Y	C1R	NM_001733	2	235	1	Y	Y	Y	N
chr2:233988741-234068741	closure	N	INPP5D	NM_001017915	1	88	2	Y	Y	Y	Y
chr2:233988741-234068741	closure	N	INPP5D	NM_005541	1	88	2	Y	Y	Y	Y
chr9:139205997-139227998	extension	Y	DKFZP434A062	NR_026964	1	242	1	Y	Y	Y	Y

b. DNase I hypersensitivity analysis

To determine whether any biologically relevant regulatory sequences were present in the novel insertion sequences, we aligned DNase I hypersensitivity sequences to a patched GRCh37 with gap closure and extension sequences and all STRs that were expanded by at least >1 kbp in CHM1. We aligned DNase I hypersensitivity sequence data for 54 previously described samples[38] (Supplementary Table 21; GEO accessions: GSE29692 and GSE32970) with Bowtie 1.0.0[39] [--mm -n 3 -v 3 -k 2], filtered out all reads with multiple alignments (MAPQV = 0), successfully called DNase I hypersensitivity peaks for 44 samples using the Hotspot v4.0 peak caller[40]. We merged Hotspot peak calls with an FDR < 0.01 from all samples with a SPOT score > 0.5 (n = 25) to create a set of 1,035,306 DNase I hypersensitivity regions totaling 189,692,280 bp (6% of the genome). Expanded STRs far outnumbered gap closures and extensions with 11,717 STRs compared to 108 non-zero-sized gap closures and extensions. Similarly, STRs accounted for nearly four times the genomic space with 4,110,971 bp compared to 1,119,211 bp of gaps. Of the 11,717 STRs, 847 (7%) were larger than 1 kbp and totaled 2,271,379 bp. Of the 108 gaps, 103 (95%) were larger than 1 kbp with a total size of 1,116,238 bp. A total of 2,924 DNase I hypersensitivity sites mapped within STRs corresponding to 575,860 bp while 548 sites mapped within gap closures totaling 105,140 bp (Supplementary Table 22). Of the 108 non-zero-sized gap closures and extensions, 88 contained at least one DNase I hypersensitivity peak call with a median of 4 calls. Of the 27 gap regions that occur within annotated genes, 23 regions (85%) contained one or more peak calls with a median of 5 calls per region.

To test whether these results reflected an enrichment of DNase I hypersensitivity sites within STRs or gap sequences, we compared the total bases corresponding to peak calls within the inserted sequence against a null density distribution created from the rest of the genome. The null distribution was created by randomly selecting equivalently sized regions from the genome for each STR or gap for 1000 iterations. Comparisons with the null were then performed with permutation tests using 100,000 permutations to calculate an empirical p-value. The observed density distribution within STRs was significantly higher than expected at 34 bp per region compared to the simulated mean of 23 bp per region (p < 0.00001; Supplementary Figure 18). Gaps were also enriched for DNase I hypersensitivity sites compared to the genomic null distribution with 965 bp per region on average compared to the simulated mean of 687 bp (p = 0.0018).

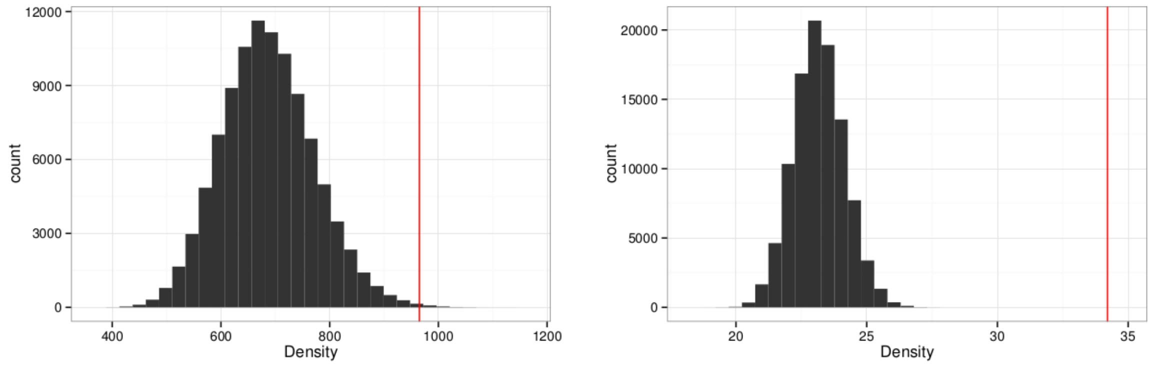
Supplementary Table 21. Samples and tissues used for DNase I hypersensitivity analysis including GEO and SRA accessions.



Sample/Tissue	GEO accession	SRA experiment path	SRA experiment	SRA run	Read length
A549	GSM736506	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069099	SRX069099	SRR231127	36
A549	GSM736580	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069173	SRX069173	SRR231203	36
A549	GSM816649	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX100/SRX100904	SRX100904	SRR352423	36
A549	GSM816649	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX100/SRX100904	SRX100904	SRR352424	36
AG04449	GSM736562	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069155	SRX069155	SRR231185	36
AG04449	GSM736590	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069183	SRX069183	SRR231213	36
AG04450	GSM736514	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069107	SRX069107	SRR231135	36
AG04450	GSM736563	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069156	SRX069156	SRR231186	36
AG09309	GSM736551	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069144	SRX069144	SRR231174	36
AG09309	GSM736616	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069209	SRX069209	SRR231241	36
AG09319	GSM736531	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069124	SRX069124	SRR231154	36
AG09319	GSM736619	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069212	SRX069212	SRR231244	36
AG10803	GSM736598	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069191	SRX069191	SRR231222	36
AG10803	GSM736633	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069226	SRX069226	SRR231258	36
AoAF	GSM736505	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069098	SRX069098	SRR231126	36
AoAF	GSM736583	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069176	SRX069176	SRR231206	36
BE2_C	GSM736508	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069101	SRX069101	SRR231129	36
BE2_C	GSM736622	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069215	SRX069215	SRR231247	36
BJ	GSM736518	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069111	SRX069111	SRR231139	36
BJ	GSM736518	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069111	SRX069111	SRR231140	36
BJ	GSM736596	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069189	SRX069189	SRR231219	36
BJ	GSM736596	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069189	SRX069189	SRR231220	36
Caco-2	GSM736500	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX069/SRX069093	SRX069093	SRR231121	36

Supplementary Table 22. Merged DNase peaks across all samples in STRs or gaps for a patched GRCh37.

Chr	Start	End	Region type	DNase bases in region
chr1	725,320	725,530	STR	210
chr1	725,600	725,750	STR	64
chr1	726,300	726,510	STR	210
chr1	726,620	726,770	STR	150
chr1	726,880	727,290	STR	410
chr1	727,400	727,570	STR	170
chr1	727,640	727,890	STR	40
chr1	728,680	728,830	STR	5
chr1	729,000	729,350	STR	37
chr1	729,000	729,350	STR	57
chr1	814,060	814,210	STR	122
chr1	814,220	814,410	STR	190
chr1	915,760	915,990	STR	5
chr1	916,280	916,510	STR	42
chr1	916,840	916,990	STR	98
chr1	917,180	917,350	STR	141
chr1	936,440	936,590	STR	124
chr1	988,920	989,070	STR	110
chr1	989,220	989,390	STR	170
chr1	1,009,100	1,009,390	STR	78
chr1	1,023,380	1,023,530	STR	150
chr1	1,026,000	1,026,210	STR	11
chr1	1,026,660	1,027,010	STR	22



Supplementary Figure 18. DNase I hypersensitivity peak distributions. (left) Genomic null distribution of DNase I hypersensitivity peak density (bp) for gap-sized regions centered around 687 bp per region with the observed mean density of DNase sites in gaps (965 bp per region) shown by the vertical red line. (*right*) Genomic null distribution of DNase I hypersensitivity peak density (bp) for STR-sized regions centered around 23 bp per region with the observed mean density of DNase sites in STRs (34 bp per region) shown by the vertical red line.

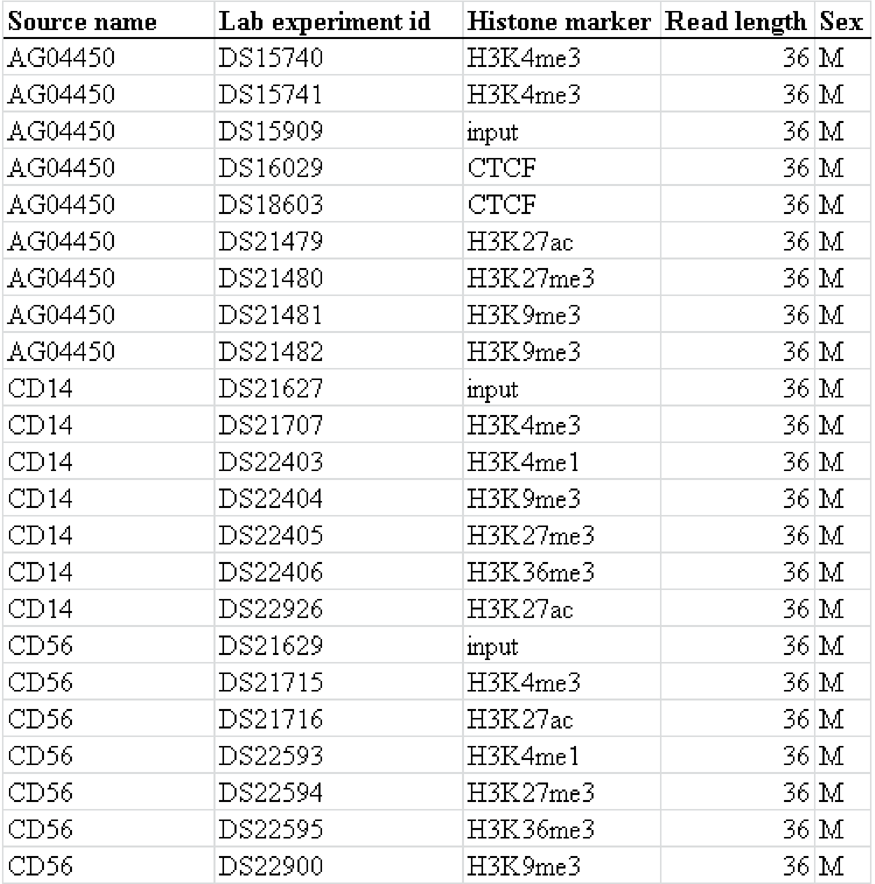
c. ChIP-seq analysis

To determine whether any biologically relevant regulatory sequences were present in the sequences from gap closures and extensions, we also mapped ChIP-seq data from seven different histone modification markers to a gap-filled version of GRCh37. Because actual gap closure and extension sequences were always smaller than the estimated gap size in GRCh37, we were able to replace gap sequence in the reference (Ns) with our new sequence while maintaining the same coordinate system as GRCh37 by leaving the remaining gap bases in place.

We obtained ChIP-seq reads for a subset of the ENCODE project corresponding to seven histone markers and controls (CTCF, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, and H3K9me3) from eight tissue sources (Supplementary Table 23). We aligned these single-end reads to our gap-filled reference with BWA aln (v. 0.7.3). For downstream analysis, we used alignments with mapping quality of 30 or greater from chromosomes 1-22 and X. We called peaks using MACS (v. 2.0.10 20120605)[41] with each sample/tissue BAM as a treatment and the corresponding control BAM for that sample as the control and a quality threshold of 0.01.

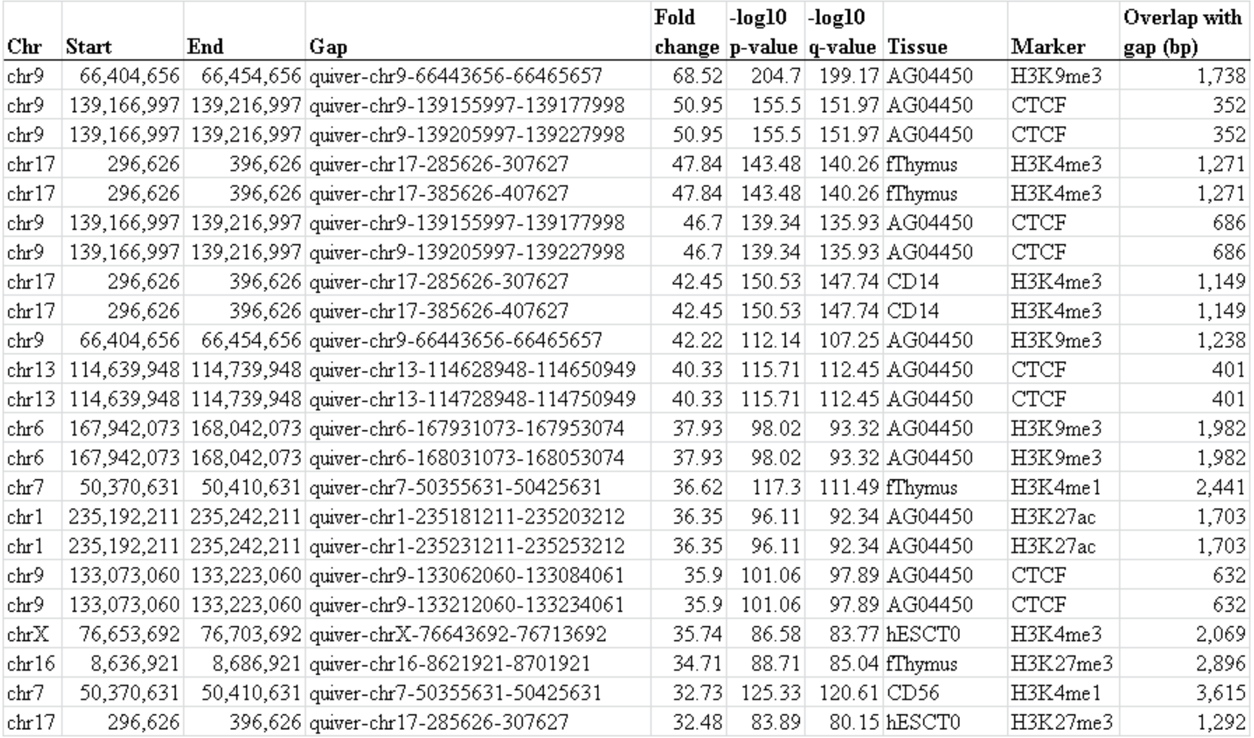
Of the 90 distinct gap regions with closures or extensions, 62 contained at least one peak call with a median of 14 calls (Supplementary Table 24). The median sum of peak calls by gap region was 4,662 bp. Of the 21 gap regions that occur within annotated genes, 17 regions (81%) contained one or more peak calls with a median of 16 calls per region. Samples from fetal lung tissue (AG04450) had the most calls by size in gaps with 156,315 of 656,692 bp (24%) while embryonic stem cell tissue (hESCT0) only had 35,440 bp (5%) of all peak calls. The most highly represented marker inside gaps was H3K4me1 with 246,581 bp (38%) closely followed by H3K9me3 with 223,956 bp (34%). In contrast, CTCF peaks had the least total bases called inside gaps with 7,436 bp (1%). When we investigated the total bases of peaks called across gaps by tissue and histone marker, we observed the pattern of strongest signal by tissue or marker was driven specifically by strong H3K9me3 signal in fetal lung tissue (AG04450) as opposed to an overall elevated signal in fetal lung tissue for all markers or H3K9me3 in all tissues.

Supplementary Table 23. Manifest of samples and histone markers used in ChIP-seq analysis of gap closures and extensions.



Source name	Lab experiment id	Histone marker	Read length	Sex
AG04450	DS15740	H3K4me3	36	M
AG04450	DS15741	H3K4me3	36	M
AG04450	DS15909	input	36	M
AG04450	DS16029	CTCF	36	M
AG04450	DS18603	CTCF	36	M
AG04450	DS21479	H3K27ac	36	M
AG04450	DS21480	H3K27me3	36	M
AG04450	DS21481	H3K9me3	36	M
AG04450	DS21482	H3K9me3	36	M
CD14	DS21627	input	36	M
CD14	DS21707	H3K4me3	36	M
CD14	DS22403	H3K4me1	36	M
CD14	DS22404	H3K9me3	36	M
CD14	DS22405	H3K27me3	36	M
CD14	DS22406	H3K36me3	36	M
CD14	DS22926	H3K27ac	36	M
CD56	DS21629	input	36	M
CD56	DS21715	H3K4me3	36	M
CD56	DS21716	H3K27ac	36	M
CD56	DS22593	H3K4me1	36	M
CD56	DS22594	H3K27me3	36	M
CD56	DS22595	H3K36me3	36	M
CD56	DS22900	H3K9me3	36	M

Supplementary Table 24. Histone modification marker peaks called inside gap closure and extension sequences with MACS 2 for a variety of different tissues.

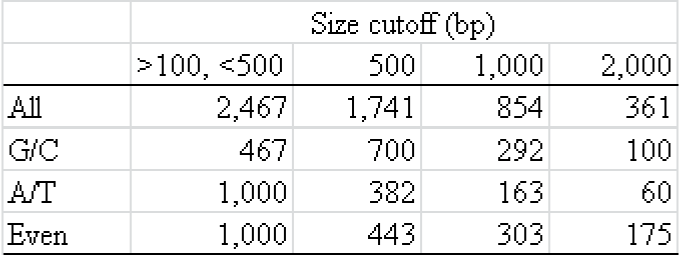


Chr	Start	End	Gap	Fold change	-log10 p-value	-log10 q-value	Tissue	Marker	Overlap with gap (bp)
chr9	66,404,656	66,454,656	quiver-chr9-66443656-66465657	68.52	204.7	199.17	AG04450	H3K9me3	1,738
chr9	139,166,997	139,216,997	quiver-chr9-139155997-139177998	50.95	155.5	151.97	AG04450	CTCF	352
chr9	139,166,997	139,216,997	quiver-chr9-139205997-139227998	50.95	155.5	151.97	AG04450	CTCF	352
chr17	296,626	396,626	quiver-chr17-285626-307627	47.84	143.48	140.26	fThymus	H3K4me3	1,271
chr17	296,626	396,626	quiver-chr17-385626-407627	47.84	143.48	140.26	fThymus	H3K4me3	1,271
chr9	139,166,997	139,216,997	quiver-chr9-139155997-139177998	46.7	139.34	135.93	AG04450	CTCF	686
chr9	139,166,997	139,216,997	quiver-chr9-139205997-139227998	46.7	139.34	135.93	AG04450	CTCF	686
chr17	296,626	396,626	quiver-chr17-285626-307627	42.45	150.53	147.74	CD14	H3K4me3	1,149
chr17	296,626	396,626	quiver-chr17-385626-407627	42.45	150.53	147.74	CD14	H3K4me3	1,149
chr9	66,404,656	66,454,656	quiver-chr9-66443656-66465657	42.22	112.14	107.25	AG04450	H3K9me3	1,238
chr13	114,639,948	114,739,948	quiver-chr13-114628948-114650949	40.33	115.71	112.45	AG04450	CTCF	401
chr13	114,639,948	114,739,948	quiver-chr13-114728948-114750949	40.33	115.71	112.45	AG04450	CTCF	401
chr6	167,942,073	168,042,073	quiver-chr6-167931073-167953074	37.93	98.02	93.32	AG04450	H3K9me3	1,982
chr6	167,942,073	168,042,073	quiver-chr6-168031073-168053074	37.93	98.02	93.32	AG04450	H3K9me3	1,982
chr7	50,370,631	50,410,631	quiver-chr7-50355631-50425631	36.62	117.3	111.49	fThymus	H3K4me1	2,441
chr1	235,192,211	235,242,211	quiver-chr1-235181211-235203212	36.35	96.11	92.34	AG04450	H3K27ac	1,703
chr1	235,192,211	235,242,211	quiver-chr1-235231211-235253212	36.35	96.11	92.34	AG04450	H3K27ac	1,703
chr9	133,073,060	133,223,060	quiver-chr9-133062060-133084061	35.9	101.06	97.89	AG04450	CTCF	632
chr9	133,073,060	133,223,060	quiver-chr9-133212060-133234061	35.9	101.06	97.89	AG04450	CTCF	632
chrX	76,653,692	76,703,692	quiver-chrX-76643692-76713692	35.74	86.58	83.77	hESCT0	H3K4me3	2,069
chr16	8,636,921	8,686,921	quiver-chr16-8621921-8701921	34.71	88.71	85.04	fThymus	H3K27me3	2,896
chr7	50,370,631	50,410,631	quiver-chr7-50355631-50425631	32.73	125.33	120.61	CD56	H3K4me1	3,615
chr17	296,626	396,626	quiver-chr17-285626-307627	32.48	83.89	80.15	hESCT0	H3K27me3	1,292

d. Replication phase analysis.

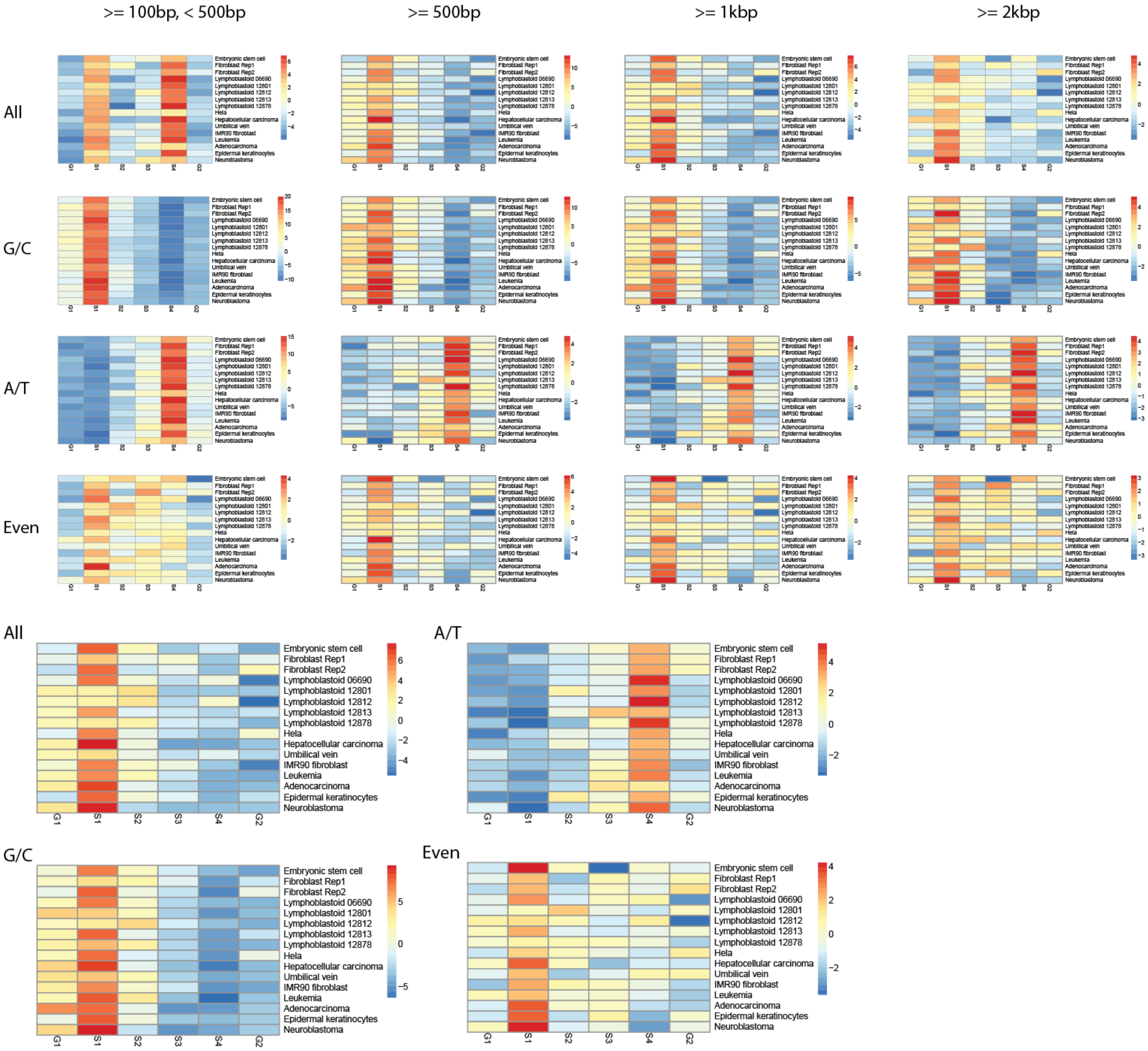
The loci of long STR and VNTR sequences were characterized according to whether or not they were in regions replicated in the early or late phase of cell division. The genome was divided into 1 kbp segments annotated as G1, S1, S2, S3, S4, and G2 by selecting the greatest normalized Repli-seq signal[42] in each segment. The total number of 1 kbp regions in each phase was tallied, and the phase of each insertion was computed using overlap with 1 kbp bins. The enrichment score for each insertion class is the number of standard deviations above the expected number of instances given the classes of all 1 kbp bins as a background distribution. Roughly, G1 and S1 are considered early replicating, and S4 and G2 late. Repli-seq data from 16 tissues (Supplementary Figure 19) were used. We considered four sets of loci determined by size cutoff: one control set STR/VNTR of length greater than 100 bp and less than 500 bp, not expanded in CHM1, and then three sets of STR/VNTR expanded in CHM1 having loci with expansions of at least 500 bp, at least 1 kbp, and at least 4 kbp. Each set was further divided according to whether or not it is high G, C, or G/C, or high A, T, or A/T, and roughly evenly distributed base composition (no nucleotide comprises more than 35% of the insertion). The number of loci of each type is given in Supplementary Table 25.

Supplementary Table 25. The number of insertions by size cutoff and nucleotide composition.



	Size cutoff (bp)			
	>100, <500	500	1,000	2,000
All	2,467	1,741	854	361
G/C	467	700	292	100
A/T	1,000	382	163	60
Even	1,000	443	303	175

GC-rich sequences are early replicating and AT-rich late replicating, consistent with past observations[43]. It is known that heterochromatin is characterized by late-replication timing[44]. Because of the heterogeneity of enrichment for late replication across samples for the AT-rich loci for longer insertions (Supplementary Figure 19, top), it is possible these loci may represent facultative heterochromatin. Supplementary Figure 19 (bottom) demonstrates the replication timing for all insertion loci across the different nucleotide compositions.



Supplementary Figure 19. Repli-seq enrichment and depletion and insertion base composition. (*top*) Phase enrichment for insertions by insertion size, and GC, AT, or unbiased (N). (*bottom*) Enrichment for insertions without size limitation.

IX. Inaccessible to sequence mapping and assembly by SMRT WGS

a. Unresolved hard-stop regions

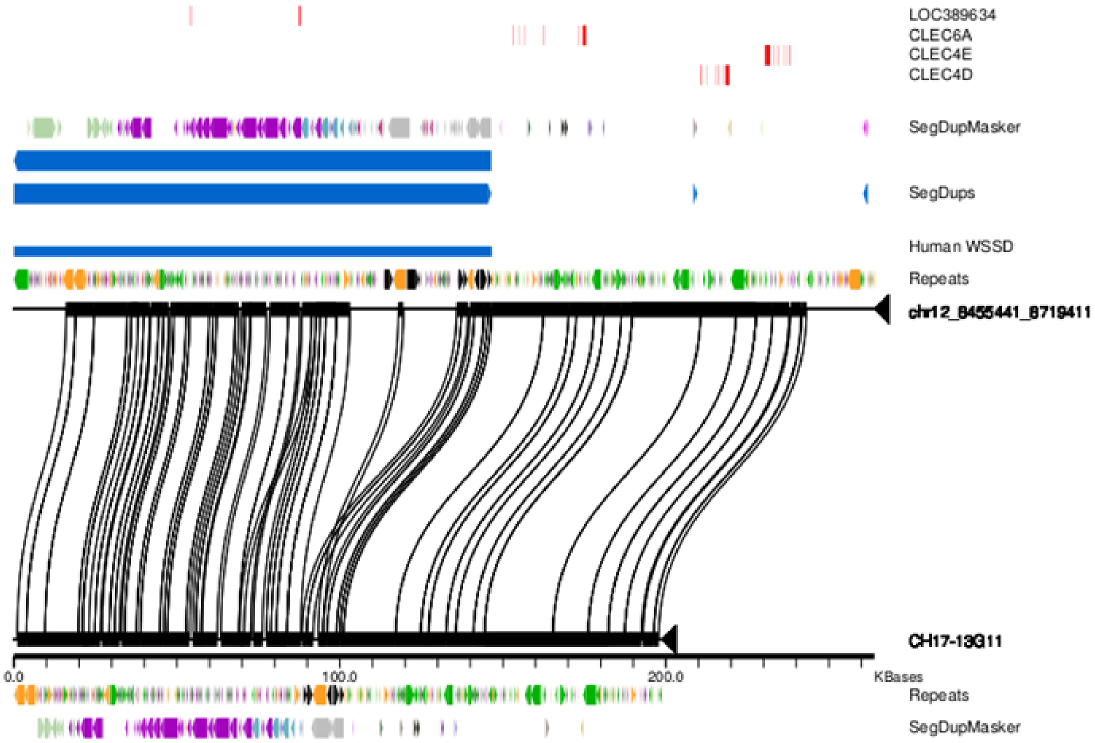
We inspected all regions of GRCh37 where there was deficiency of uniquely mapped reads or there was a cluster of hard-stop events that failed to assemble into a single contig or had an incomplete alignment to the reference (Supplementary Table 26). The latter events (n = 22 regions) are particularly important because they may represent either errors in the reference genome or alternative structural configurations for these regions. We identified a total of 22 hard-stop regions that show an incomplete alignment to the reference: 12 within unique regions and 10 hard-stop events within or adjacent to a segmental duplication representing a 40.4-fold enrichment. In addition, we identified 167 regions where the CHM1 assembly was highly fragmented (3 or more contigs) compared to the reference. Of these, 139 mapped within unique regions while 28 mapped adjacent or within segmental duplications. Because segmental duplications are hotspots for structural variation, we anticipate that these events signal the breakpoints of unresolved or larger more complex structural variants and warrant further investigation (Supplementary Table 27). Sequencing of large-insert clones corresponding to these regions showed a complex pattern of common repeats or multiple paralogous segments mapping to different chromosomes. The validation of a resolved hard-stop event is shown in Supplementary Figure 20.

Supplementary Table 26. Summary of hard-stop events.

Category	Count
All hard-stops	817
Resolved	569
Adjacent to gap	206
Unresolved, flanking segdup	10
Unresolved, not flanking segdup	12

Chr	Start	End	Adjacent to seg dup boundary
chr2	242,843,229	242,853,252	N
chr2	243,035,778	243,043,810	N
chr3	162,495,533	162,512,134	N
chr4	81,117,408	81,133,048	N
chr6	79,036,414	79,047,643	N
chr10	87,115,262	87,121,113	N
chr11	1,928,741	1,936,959	N
chr11	1,951,519	1,969,404	N
chr12	83,023,122	83,034,039	N
chr13	114,202,542	114,221,799	N
chr19	7,029,801	7,064,205	N
chr20	54,103,709	54,123,236	N
chr1	103,785,000	103,785,500	Y
chr1	145,944,000	145,944,500	Y
chr1	223,725,500	223,726,500	Y
chr4	75,493,000	75,493,500	Y
chr5	179,085,500	179,086,000	Y
chr5	70,391,000	70,391,500	Y
chr8	2,329,000	2,329,500	Y
chr12	9,632,500	9,633,000	Y
chr16	15,225,500	15,226,000	Y
chr17	77,630,000	77,630,500	Y

Supplementary Table 27. Unresolved hard-stop events.



Supplementary Figure 20. Validation of a hard-stop event called by CHM1 long reads by a CH17 BAC clone spanning the same genomic region of GRCh37. The Miropeats alignment shows the resolved insertion of approximately 30 kbp in GRCh37 relative to the CHM1-based clone.

b. Depletion of high mapping quality reads

In addition to breaks in coverage detected by hard-stops, we also discovered euchromatic regions of GRCh37 where the assembly has been resolved, but coverage of SMRT WGS was low or nonexistent because of an inability to uniquely assign reads, excluding regions corresponding to deletions. We defined regions as low coverage if they had fewer than five-fold coverage with a mapping quality greater than 20, but also greater than five-fold coverage of reads with mapping quality ≥20 that did not overlap coordinates in our deletion calls, and merged all such regions that occurred with 1 kbp of each other. We identified 715 regions that matched these parameters totaling 12.3 Mbp. (Supplementary Table 28). Not surprisingly, 92.3% (660/715) of the regions are in segmental duplications with an average identity of 99.3±1.3% (median 100% identity) indicating these are the stretches of exact duplication longer than the lengths of the reads. The regions greater than 5 kbp, or approximately the average read length, comprise 94.5% of the regions difficult to map. Note that because some of the reads at the tail end of the distribution of the exponentially distributed reads lengths may map confidently, there is a low average coverage of 4.71 of high mapping quality reads in these repetitive regions as shown in Supplementary Table 29.

Supplementary Table 28. Classification of regions with fewer than four aligned PacBio WGS reads at mapping quality greater than 10.

	All regions			Regions >5 kbp		
Category	**Regions**	**Total size (bp)**	**% total size**	**Regions**	**Total size (bp)**	**% total size**
Total	715	12,377,300	100	252	11,720,100	100
Subtelomeric or centromeric	227	1,164,700	9	42	918,700	8
Euchromatic	488	11,212,600	91	210	10,801,400	92
Segmental duplications	428	7,490,900	61	180	7,135,200	61
Non-duplicated	54	3,720,300	30	30	3,666,200	31

Supplementary Table 29. Inaccessible regions. Regions with low coverage of confidently mapped reads, but that show a higher coverage of low-confidence mapping reads.

Chrom.	Start	End	Segdup Identity	Low map quality coverage	High map quality coverage
chr1	22700	27000	1.00	5.44	3.77
chr1	47900	57300	1.00	8.34	1.19
chr1	59200	61100	1.00	4.74	2.84
chr1	62900	71400	1.00	6.85	4.05
chr1	102300	104200	0.99	6.26	5.00
chr1	121900	155300	1.00	7.08	5.39
chr1	174600	175900	1.00	5.00	4.46
chr1	266900	267700	1.00	3.88	2.83
chr1	317800	375200	1.00	11.27	4.84
chr1	399700	410800	1.00	6.31	6.86
chr1	432900	447600	1.00	29.01	7.14
chr1	453400	471300	1.00	10.44	3.39
chr1	522100	525300	1.00	7.03	4.31
chr1	600200	610600	1.00	8.25	3.72

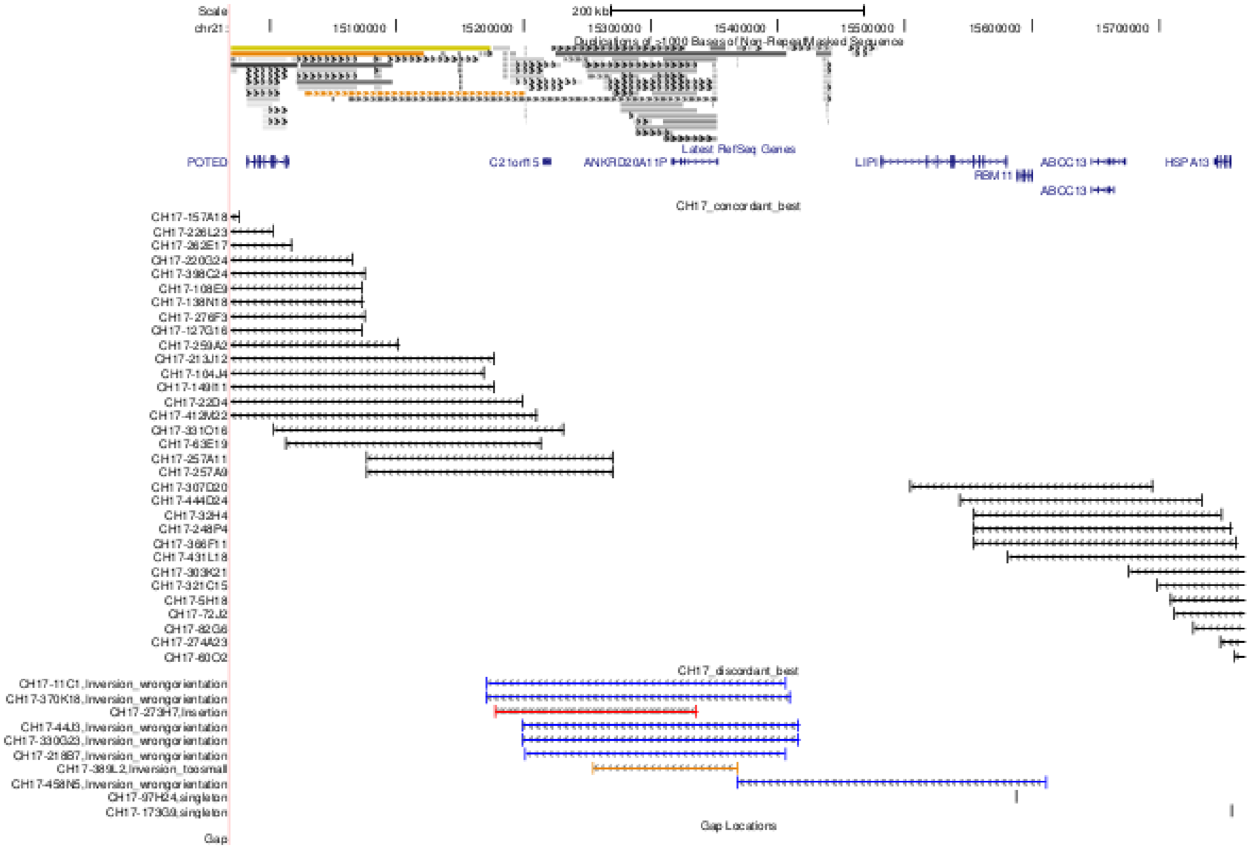
c. Unidentified inversions.

Structural variation mediated by very long high identity repeats are likely to be missed by our analysis. We estimated the number of large inversion events using an orthogonal method of mapping BAC end sequences derived from CHORI-17 (CHM1 clone library) to GRCh37 (as described in Kidd, 2008[34]). Inversions were detected initially on a per-clone basis by a signature of improperly oriented read-pairs spanning a genomic region of 100 kbp to 1 Mbp. To identify high-confidence inversions, we clustered all overlapping inversions from single clones and required each inversion region to have two or more clones supporting the inversion and at least 50% of the region to have no spanning concordant clones. We identified 11 inversions meeting these criteria ranging in size from 230,347 bp to 861,276 bp with a median size of 491,287 bp—three of these larger events were subsequently validated by FISH and or by sequence analysis (Supplementary Table 30). None of these were detected by PacBio SMRT read analysis. The

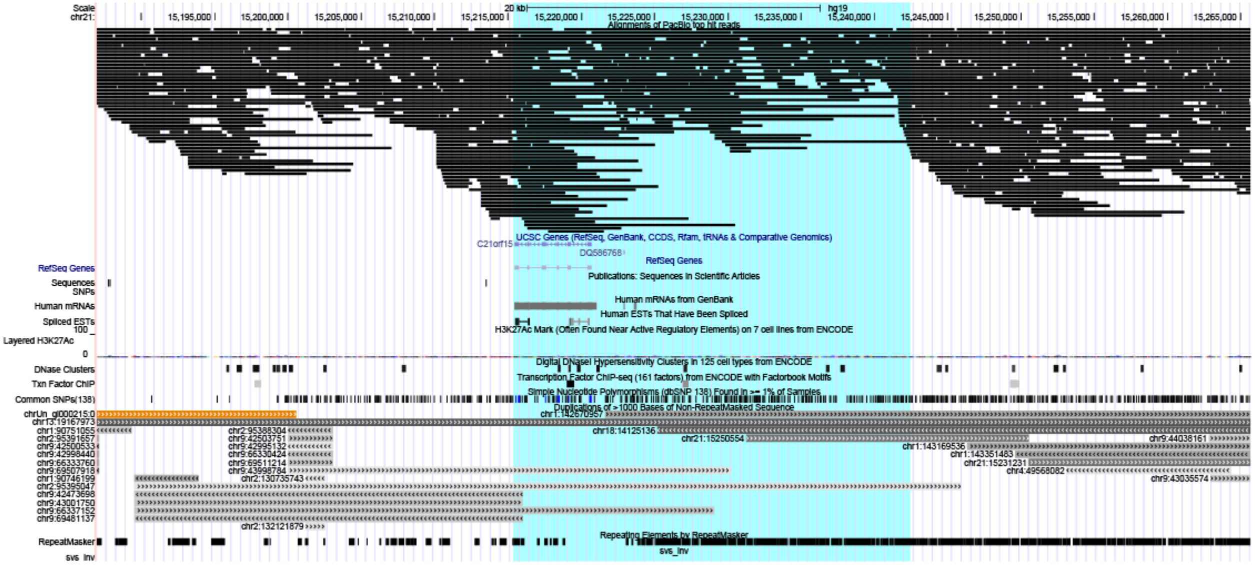
BAC end sequence support for the inversion in chromosome 21 is shown in Supplementary Figure 21, and the PacBio read support is shown in Supplementary Figure 22.

Supplementary Table 30. Large inversions detected by CH17 BES against GRCh37 and still present in GRCh38.

Chromosome	Start	End	Size (bp)	Validation method	Left duplication size (bp)	Right duplication size (bp)	Duplication identity
chr15	30704161	30846486	142,325	BAC	58,697	58,697	0.997
chr16	14867984	15441590	573,606	Nextera	72,656	72,656	0.992
chr21	15170719	15611138	440,419	BES	12,307	12,406	0.958



Supplementary Figure 21. A region of chromosome 21 that shows BAC end sequencing support for an inversion.



Supplementary Figure 22. A detailed view of the inversion region on chromosome 21 with PacBio read support. SMRT reads (top black) are concordant with the reference due to the presence of large high-identity segmental duplications over the inversion breakpoint region (blue shading).

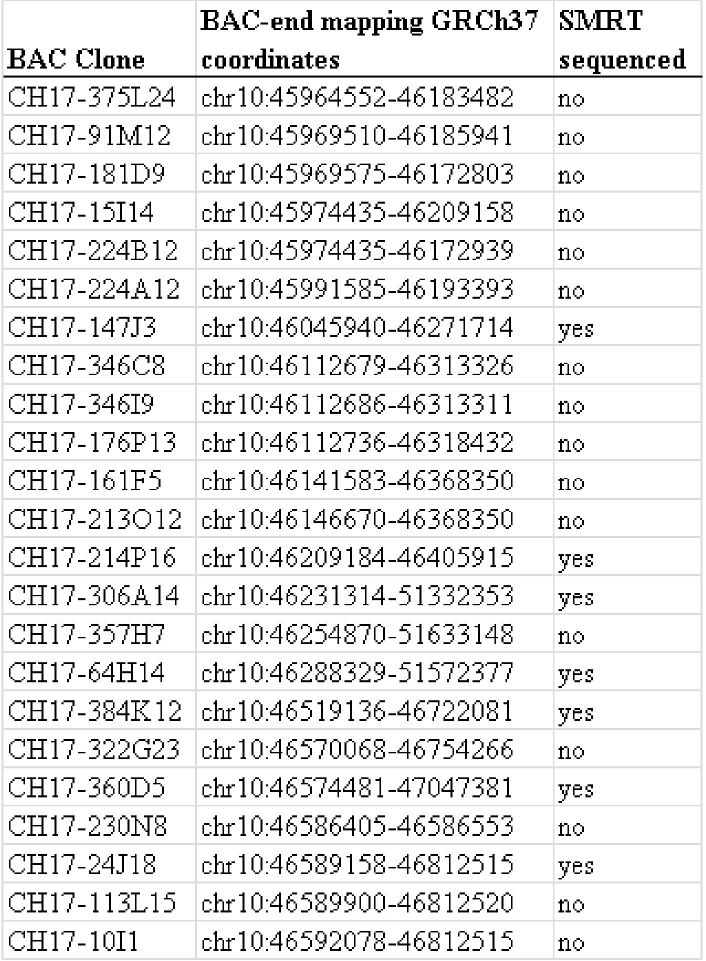
X. Chromosome 10q11 BAC-based sequencing using SMRT technology

From the total 737 euchromatic regions where local assemblies were not possible due to a lack of uniquely mapping SMRT reads, we selected one 6.5 Mbp region mapping to chromosome 10q11.23 for a more detailed analysis. The region is flanked by large blocks of segmental duplication (3.8 Mbp) that mediate recurrent deletions and duplications considered an important risk factor for pediatric intellectual disability and developmental delay[45]. Additionally, 74 regions of 10q11.23 spanning 1.1 Mbp had no uniquely mapping SMRT reads. To resolve this complex region of the genome, we applied an alternate clone-based hierarchical approach and identified a total 126 large-insert BAC clones from the CH17 library spanning the region and determined their sequence content using a Nextera-Illumina protocol[8]. Generated reads were mapped to SUNK positions within the human reference genome (GRCh37; Supplementary Table 31, Supplementary Figure 23) and a minimum tiling path was chosen. We generated contigs spanning two large clusters of segmental duplications within the 10q11.23 region by performing PacBio sequencing of 35 BACs from the CH17 library including 2.7 Mbp sequence spanning “Contig 1” and 0.8 Mbp sequence spanning “Contig 2”.

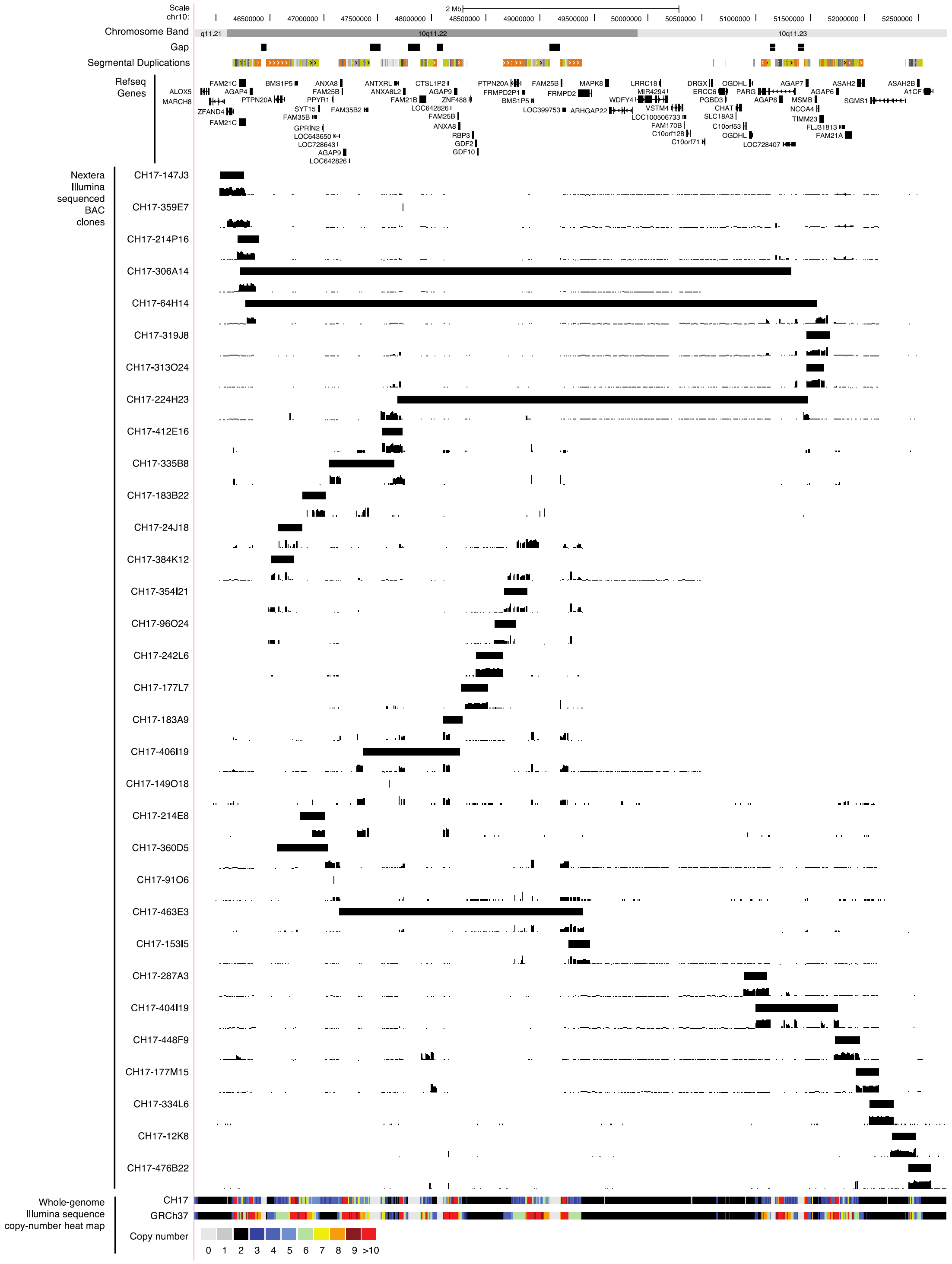
We compared the sequence content of our 10q11-generated contig with the human reference build (GRCh37). Briefly, we fragmented our 10q11 contig into 5000 bp fragments, compared with the human reference build by performing a MEGABLAST, and identified all sequences with >90% sequence identity across >1000 bp. We defined a region as “allelic” in the human reference with our 10q11 contig as a contiguous stretch of multiple sequence fragments (<10 kbp) with on average >99.8% sequence identity. Within allelic regions, we identified smaller segments at <99.8% reduced sequence identity. The majority of these reduced-sequence identity regions occurred at high copy number portions of our 10q11 contig (Figure 3). If the allelic region in the reference did not exist or was unclear, we signified it as missing sequence. We also identified redundant sequence, or regions of the 10q11 contig with multiple allelic reference regions, and sequence in the incorrect orientation.

The corresponding region in GRCh37 maps to the coordinates chr10:46,046,104-49,580,948 (Contig 1) to chr10:50,894,586-52,618,800 (Contig 2). Comparing the two haplotypes shows ~635 kbp of sequence with <99.8% sequence identity and 171 kbp with <99.6% sequence identity. The new CHM1 BAC assembly added 416 kbp missing reference sequence, corrected the orientation of 416 kbp of sequence, and eliminated 856 kbp of redundant sequence when compared to GRCh37.

Supplementary Table 31. Sequenced BAC clones mapping to the human 10q11.23 region.



BAC Clone	BAC-end mapping GRCh37 coordinates	SMRT sequenced
CH17-375L24	chr10:45964552-46183482	no
CH17-91M12	chr10:45969510-46185941	no
CH17-181D9	chr10:45969575-46172803	no
CH17-15I14	chr10:45974435-46209158	no
CH17-224B12	chr10:45974435-46172939	no
CH17-224A12	chr10:45991585-46193393	no
CH17-147J3	chr10:46045940-46271714	yes
CH17-346C8	chr10:46112679-46313326	no
CH17-346I9	chr10:46112686-46313311	no
CH17-176P13	chr10:46112736-46318432	no
CH17-161F5	chr10:46141583-46368350	no
CH17-213O12	chr10:46146670-46368350	no
CH17-214P16	chr10:46209184-46405915	yes
CH17-306A14	chr10:46231314-51332353	yes
CH17-357H7	chr10:46254870-51633148	no
CH17-64H14	chr10:46288329-51572377	yes
CH17-384K12	chr10:46519136-46722081	yes
CH17-322G23	chr10:46570068-46754266	no
CH17-360D5	chr10:46574481-47047381	yes
CH17-230N8	chr10:46586405-46586553	no
CH17-24J18	chr10:46589158-46812515	yes
CH17-113L15	chr10:46589900-46812520	no
CH17-10I1	chr10:46592078-46812515	no

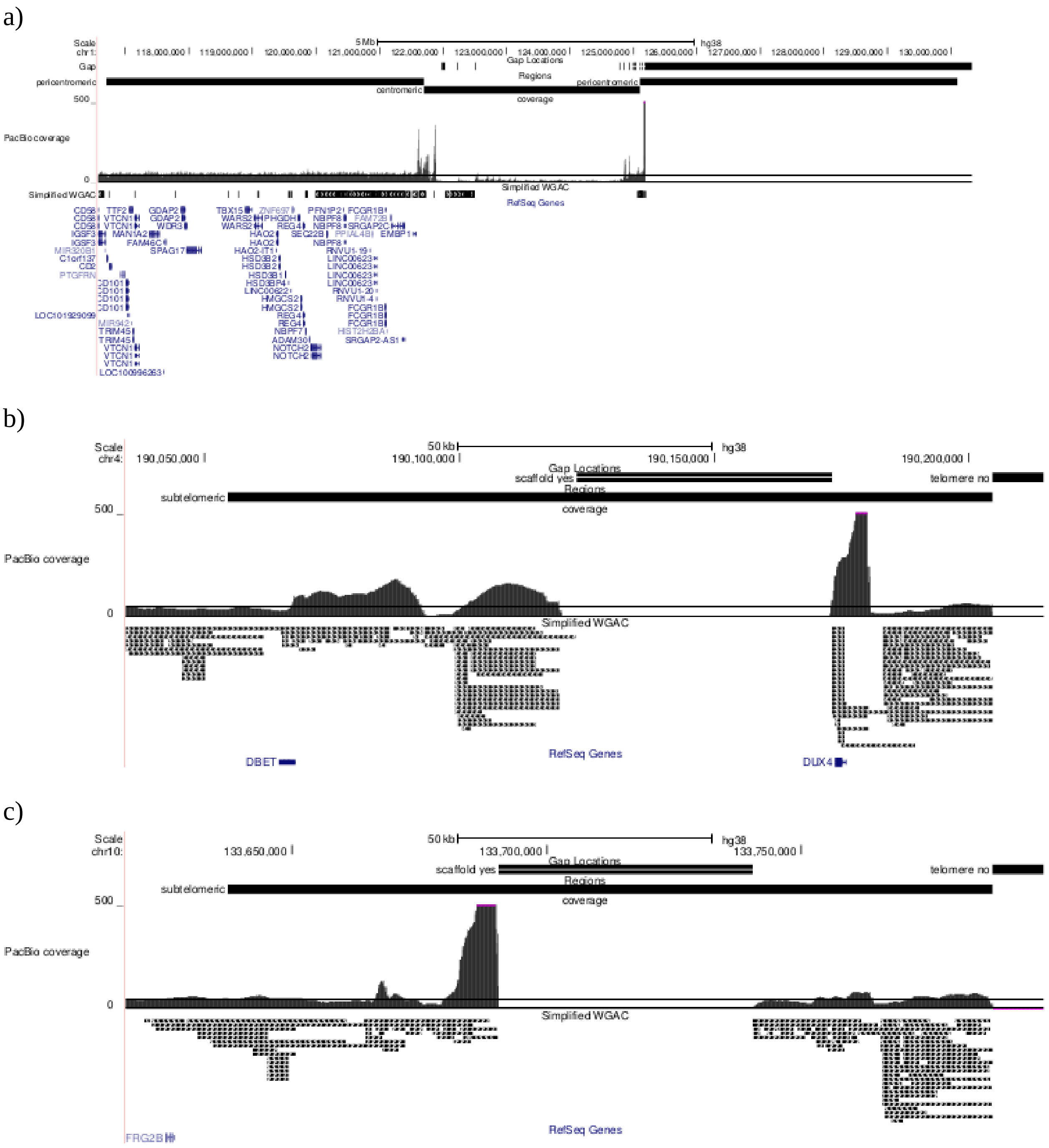


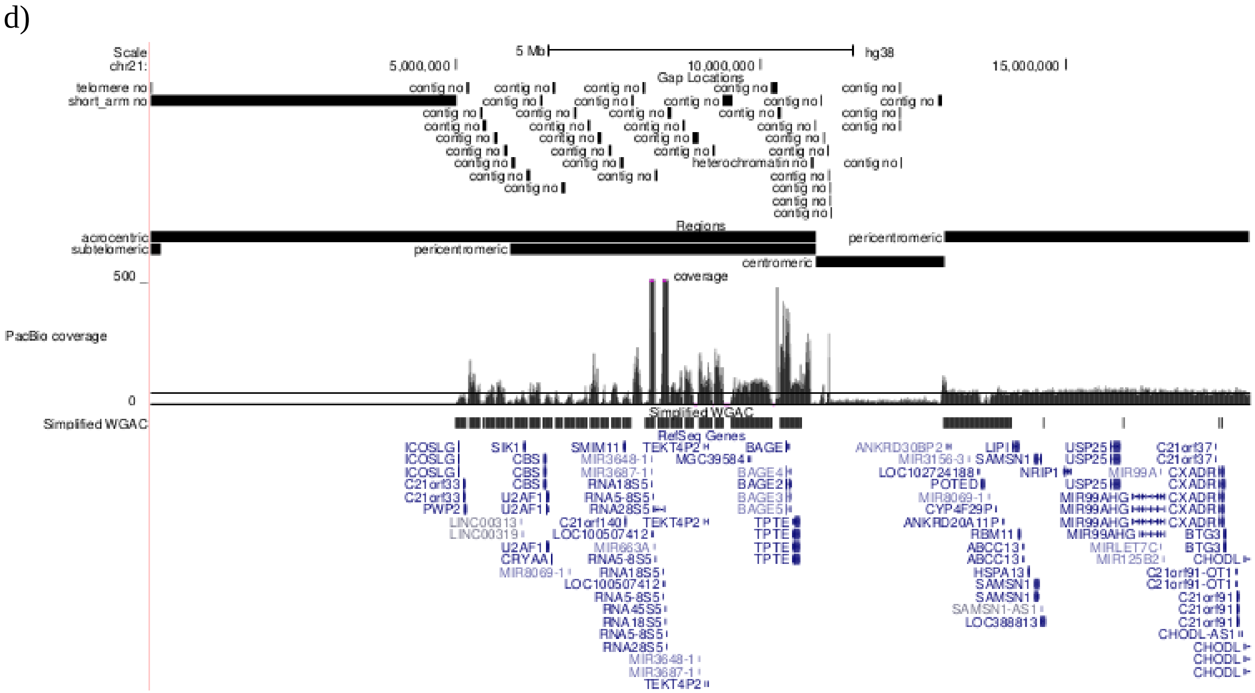
Supplementary Figure 23. Nextera mapping of 10q11.23 BACs. Nextera-Illumina sequence reads of CH17 BAC clones spanning the human 10q11.23 region mapped to SUNK positions within the GRCh37 human reference.

We evaluated the contiguity and quality of the 10q11 assembly with concordant alignments of clone ends from the CH17 BAC library and the ABC10 and ABC12 fosmid libraries. We measured the contiguity of the 10q11 assembly compared to the corresponding region of GRCh37 (chr10:46046104-52618800) by calculating the total bases covered by concordant alignments from CH17, ABC10, and ABC12 libraries. The 10q11 assembly had high coverage of its 5,265,575 bp from all three libraries with 97.6% coverage from CH17 alignments, 98.9% from ABC10, 98.4% from ABC12, and 99.6% from all three combined. In contrast, the corresponding region of GRCh37 had 76.4% coverage from CH17, 86.0% from ABC10, 83.2% from ABC12, and 90.3% from all three combined. The overall identity of concordant alignments from each library was calculated by summing all concordantly aligned bases with phred quality >30 and dividing by total bases with the same high quality. Concordant mappings were required to have >=99.8% alignment identity with the 10q11 assembly to be included in the identity calculation. Additionally, we calculated the insert size distribution of concordant end mappings for all libraries to confirm whether these distributions match the expected ranges for BAC and fosmid libraries. The identity of high-quality concordant CH17 BES alignments was 99.97% (182,407 aligned / 182,454 total bp) and the insert size was distributed around a mean of 212,895 +/- 18,256 bp. These results support the general construction of the 10q11 supercontig at the scale of BACs. The fosmid end mappings from ABC10 and ABC12 similarly supported the structure of the assembly. ABC10's concordant alignments had an identity of 99.93% (1,472,629 aligned / 1,473,641 total bp) and an insert size of 41,088 +/- 1,794 bp. ABC12's concordant alignments had an identity of 99.94% (1,424,489 aligned / 1,425,285 total bp) and an insert size of 39,845 +/- 1,148 bp. The difference in alignment identities between the BAC end and fosmid end alignments is consistent with allelic differences between individuals of the same species.

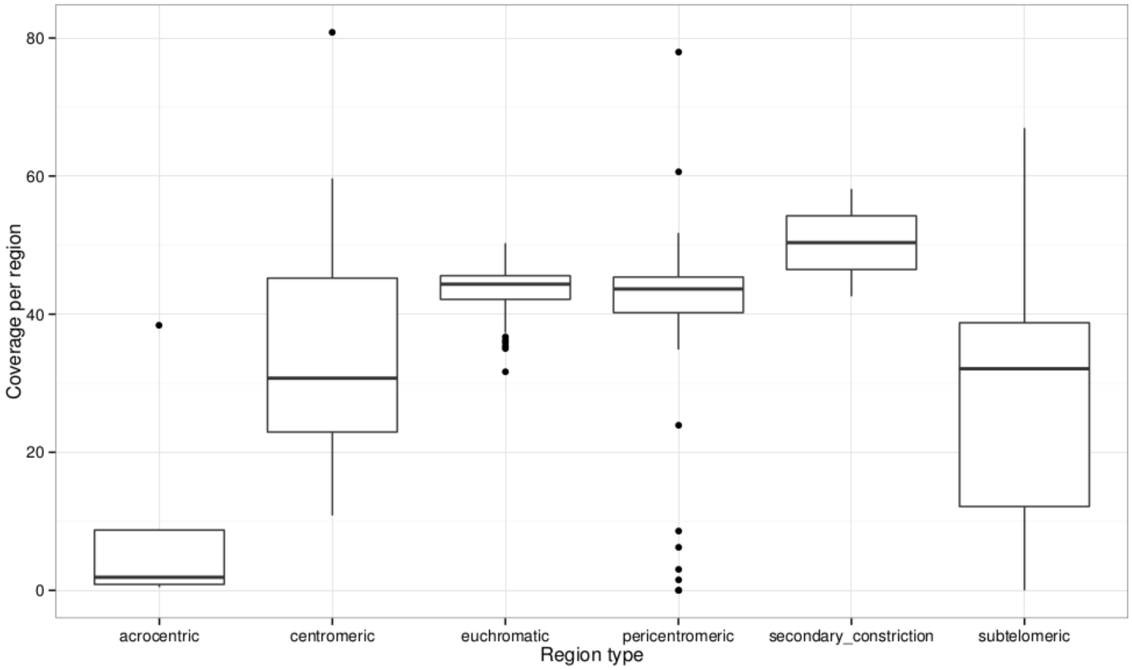
XI. Analysis of heterochromatic sequence

We were unable to accurately map and assemble reads to most of the heterochromatin and immediate subtelomeric regions of the genome. While the two order of magnitude gain of read length of PacBio reads relative to other sequencing technologies enables resolution of many complex regions, the read lengths are unfortunately still too short to assemble centromeric and pericentromeric regions. When we analyze these regions in the human genome (either GRCh38 or GRCh37), we discover pileups of reads followed by deserts where relatively few reads place (Supplementary Figure 24 a-d). A similar effect is observed near subtelomeric regions. This stems from the fact that reads cannot be uniquely assigned within large tracts of paralogy creating both collapses and deficiencies within pericentromeric duplications and centromeric satellites. In fact, the higher the copy number the fewer reads that can be assigned (e.g., acrocentric vs. pericentromeric) and the greater the variance (compare coverage and variance between euchromatin and heterochromatin) (Supplementary Figure 25, Supplementary Table 32. Any attempt to assemble these regions would lead to an erroneous sequence representation due to uneven sequence coverage. Therefore, it is currently impossible to properly assemble these regions with an effectively mapped average read length of 5.6 kbp. While sophisticated computational analysis may eventually make assembly of such regions tractable, the more fundamental advance requires read-lengths of hundreds of kbp to assemble to high quality. Thus, the only way such regions can be currently assessed is hierarchical-based sequencing of large insert clones as we have demonstrated for the pericentromeric region of 10q.





Supplementary Figure 24. Example coverage of CHM1 PacBio reads aligned to centromeric and telomeric regions (GRCh38). A horizontal line is shown at 50-fold coverage and maximum coverage displayed is 500-fold. Regions shown include a) chr1p subtelomeric, b) chr4q subtelomeric, c) chr10q subtelomeric, and d) chr21 acrocentric.



Supplementary Figure 25. GRCh38 coverage by region. Mean coverage per region for heterochromatic regions of GRCh38 including acrocentric, centromeric, pericentromeric, subtelomeric, two secondary constrictions at 1q21 and 16q12, and random euchromatic regions of the same size.

Supplementary Table 32. Coverage for heterochromatic regions of GRCh38 and random euchromatic regions of the same size. Shown are mean and standard deviation of coverage along with total bases per region.

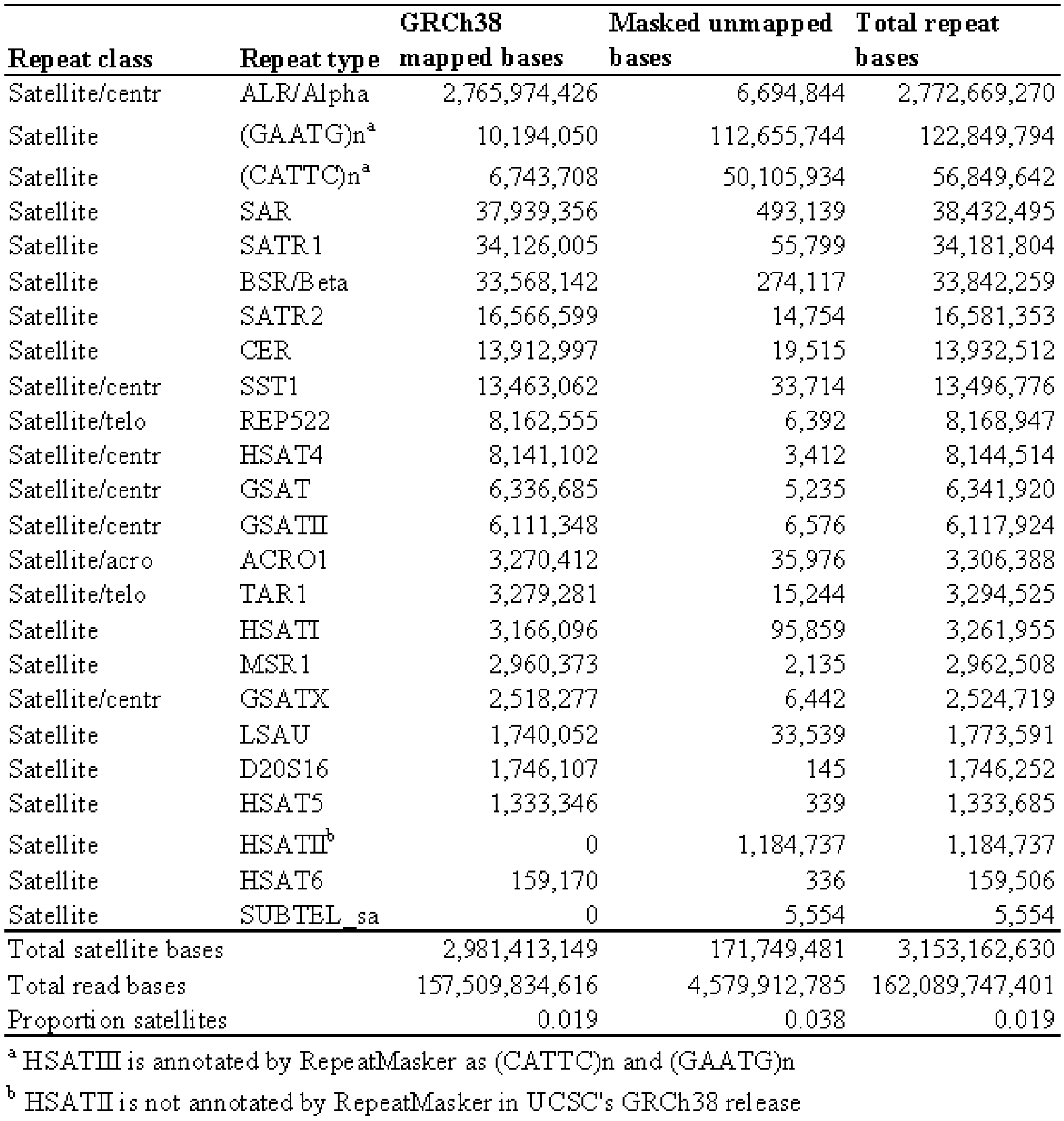
Genomic region type	Coverage (mean)	Coverage (stdev)	Total bases
euchromatic	44	3	18,788,745,898
acrocentric	10	14	590,982,334
centromeric	34	17	3,401,825,202
pericentromeric	39	16	9,004,960,610
secondary constriction	50	8	1,075,064,277
subtelomeric	28	18	193,149,913

Although heterochromatic regions remain intransigent to assembly, we attempted to determine whether more data could be extracted from the SMRT reads because of their length and the fact that are not clonally propagated. To this end, we performed an assessment of heterochromatic sequence content in CHM1 PacBio reads and identified the longest possible extensions of single reads into telomeric and centromeric regions of the genome.

a. Assessment of heterochromatic content

We estimated the proportion of heterochromatic repeat sequences present in the CHM1 PacBio data with two complementary approaches. In the first approach, we mapped all CHM1 PacBio reads to GRCh38—GRCh38 differs from GRCh37 due to the addition of centromeric sequence models that contain representative higher-order repeat sequences. We calculated the total sequence mapping to annotated satellites. Of the 157.5 Gbp of aligned bases, 3.0 Gbp (1.9%) mapped within satellites (Supplementary Table 33).Since not all satellite sequences are likely to be represented in the current reference, we also considered all reads that did not map to the human reference. Here, we applied RepeatMasker to all unmapped PacBio reads. Of the 4.6 Gbp of unmapped reads, 0.2 Gbp (4%) were classified as centromeric satellites. With both approaches combined, we analyzed 21,664,612 reads totaling 162.1 Gbp of sequence of which 3.2 Gbp (2%) were identified as satellites. Interestingly, the majority of satellite bases identified by mapping (93%) were alpha satellites while the majority of masked unmapped bases (95%) corresponded to HSATIII satellites.

Supplementary Table 33. Total satellite bases in PacBio reads identified by RepeatMasker and alignment to GRCh38. Satellites are summarized by repeat class and type.



Repeat class	Repeat type	GRCh38 mapped bases	Masked unmapped bases	Total repeat bases
Satellite/centr	ALR/Alpha	2,765,974,426	6,694,844	2,772,669,270
Satellite	(GAATG)n[a]	10,194,050	112,655,744	122,849,794
Satellite	(CATTC)n[a]	6,743,708	50,105,934	56,849,642
Satellite	SAR	37,939,356	493,139	38,432,495
Satellite	SATR1	34,126,005	55,799	34,181,804
Satellite	BSR/Beta	33,568,142	274,117	33,842,259
Satellite	SATR2	16,566,599	14,754	16,581,353
Satellite	CER	13,912,997	19,515	13,932,512
Satellite/centr	SST1	13,463,062	33,714	13,496,776
Satellite/telo	REP522	8,162,555	6,392	8,168,947
Satellite/centr	HSAT4	8,141,102	3,412	8,144,514
Satellite/centr	GSAT	6,336,685	5,235	6,341,920
Satellite/centr	GSATII	6,111,348	6,576	6,117,924
Satellite/acro	ACRO1	3,270,412	35,976	3,306,388
Satellite/telo	TAR1	3,279,281	15,244	3,294,525
Satellite	HSATI	3,166,096	95,859	3,261,955
Satellite	MSR1	2,960,373	2,135	2,962,508
Satellite/centr	GSATX	2,518,277	6,442	2,524,719
Satellite	LSAU	1,740,052	33,539	1,773,591
Satellite	D20S16	1,746,107	145	1,746,252
Satellite	HSAT5	1,333,346	339	1,333,685
Satellite	HSATII[b]	0	1,184,737	1,184,737
Satellite	HSAT6	159,170	336	159,506
Satellite	SUBTEL_sa	0	5,554	5,554
Total satellite bases		2,981,413,149	171,749,481	3,153,162,630
Total read bases		157,509,834,616	4,579,912,785	162,089,747,401
Proportion satellites		0.019	0.038	0.019

[a] HSATIII is annotated by RepeatMasker as (CATTC)n and (GAATG)n

[b] HSATII is not annotated by RepeatMasker in UCSC's GRCh38 release

b. Identification of centromeric and telomeric gap extensions

To discover sequence that extends into heterochromatic gaps, we performed a secondary analysis where we analyzed the boundaries of each region by searching specifically for reads that mapped to the adjacent gap sequence of each centromere or telomere (GRCh37). To select reads capable of extending into the centromere, but that were confidently mapped to transition regions, we masked regions covered more than 120X within 10 kbp of the boundary of the p or q side of centromeres and telomeres, and identified the longest aligned read such that the clipped bases extend furthest into the centromere. These sequences were then subsequently masked with RepeatMasker 3.3.0. We identified a total of 755,465 bases extended into the gaps placed for centromeric and telomeric regions of the genome, of which 359,900 (47%) were annotated by

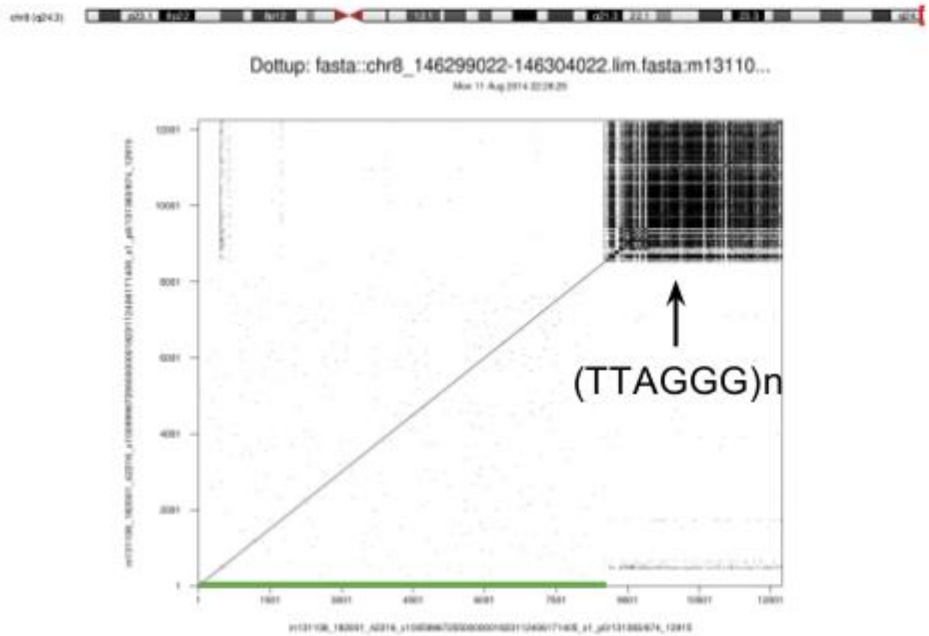
RepeatMasker to contain satellite sequence. The details of the extension for every locus are given in Supplementary Table 34.

Supplementary Table 34. Details of extensions into centromeres and telomeres using read overlap.

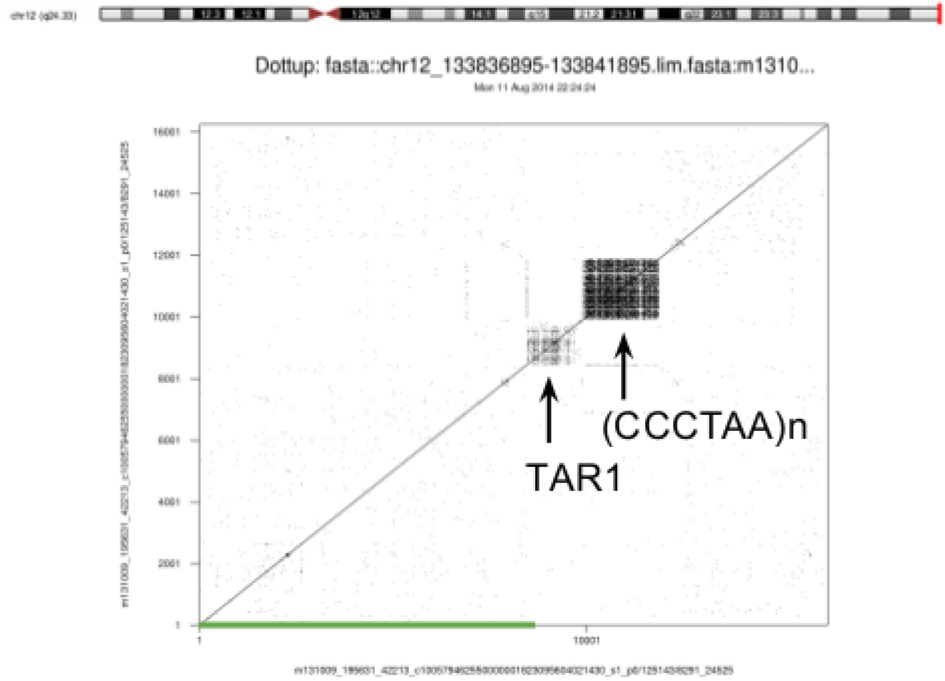
Centromere, p Extension Length	Chr.	Start	End	Major repeat	Repeat bases	Anchor length
18034	chr1	121479503	121483918	ALR/Alpha	18704	5096
3271	chr2	90544290	90544442	AluSq	149	154
10715	chr3	90494275	90504849	ALR/Alpha	19795	10828
22905	chr4	49659267	49659709	(GAATG)n	823	495
12254	chr5	46390313	46405640	ALR/Alpha	15056	15708
20396	chr6	58773564	58779467	ALR/Alpha	20121	5739
14056	chr7	58051419	58053819	ALR/Alpha	7176	2597
3357	chr9	47316748	47316990	AluSg	247	215
15048	chr10	39152002	39153107	(GAATG)n	2851	1221
13641	chr11	51590354	51594201	ALR/Alpha	17889	4521
15695	chr12	34849617	34856722	ALR/Alpha	12794	7355

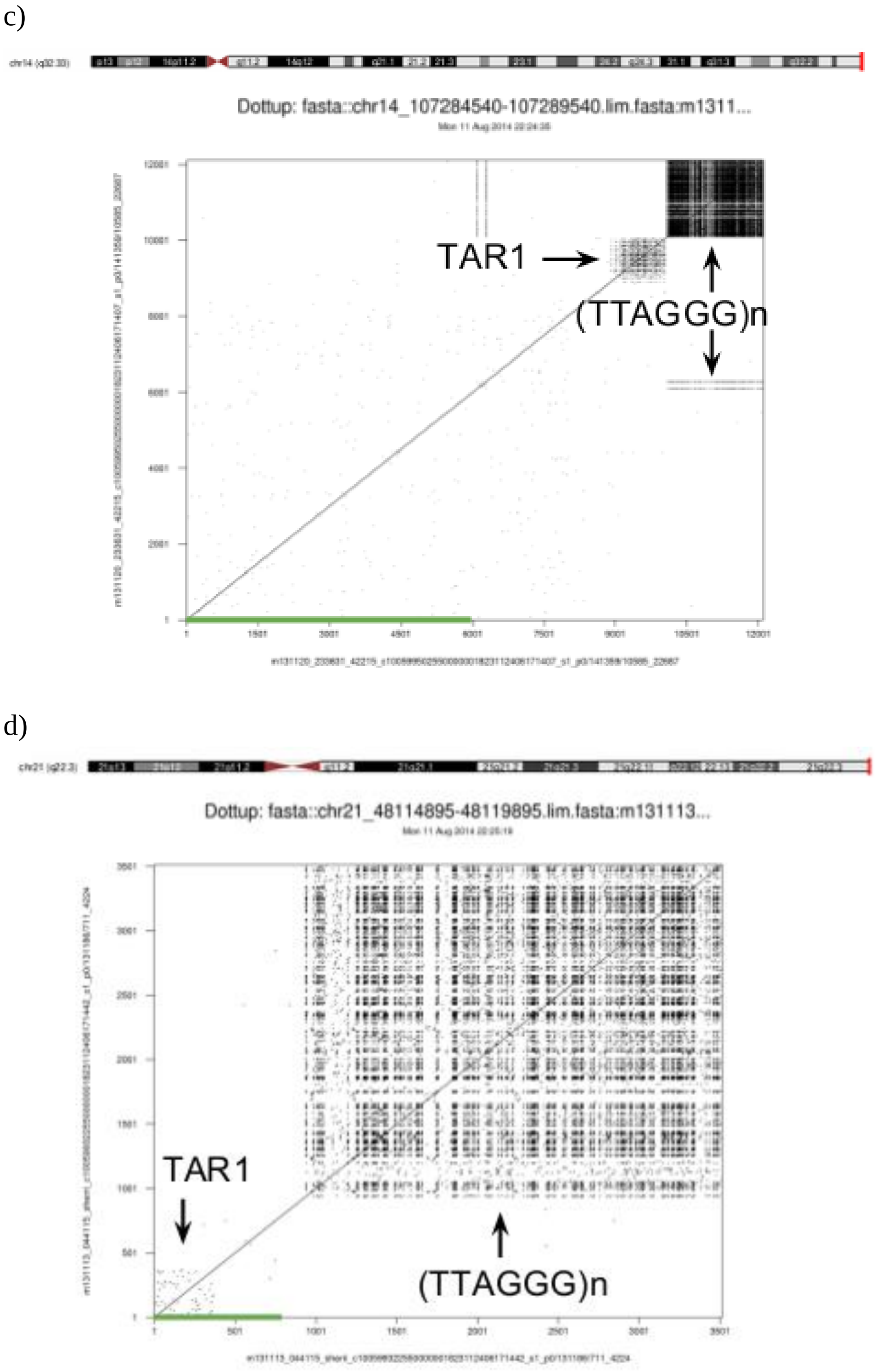
This analysis suggests different models for centromeric and subtelomeric organization as indicated below (Supplementary Figure 26). This includes organization of subtelomeric and telomeric associated repeats and the presence of higher-order repeat structures. So while these data provide some insight into the structure and organization, single-pass SMRT sequence cannot be reliably used to define high quality sequence of these regions (e.g., 15% error). We could not, for example, validate these extensions using the GRCh38 reference because of a lack of 1-1 correspondence between sequences at the boundaries of centromeres. Out of 41 sequences adjacent to centromeres, only 1p, 2p, 2q, 6q, 9q, 22q, and Xq contained matches of at least 80% in length and 90% identity. Nevertheless, these sequence extension represent an important anchor point for further investigation and provide a glimpse of the organization of these difficult regions of the genome. We have created a database of sequences corresponding to these telomeric and centromeric regions and added it as a resource to the paper.

a)


chr8 (q24.3)
Dottup: fasta::chr8_146299022-146304022.lim.fasta:m13110...
Mon 11 Aug 2014 22:28:29
12001
10001
8001
6001
4001
2001
1
1
1501
3001
4501
6001
7501
9001
10501
12001
(TTAGGG)n

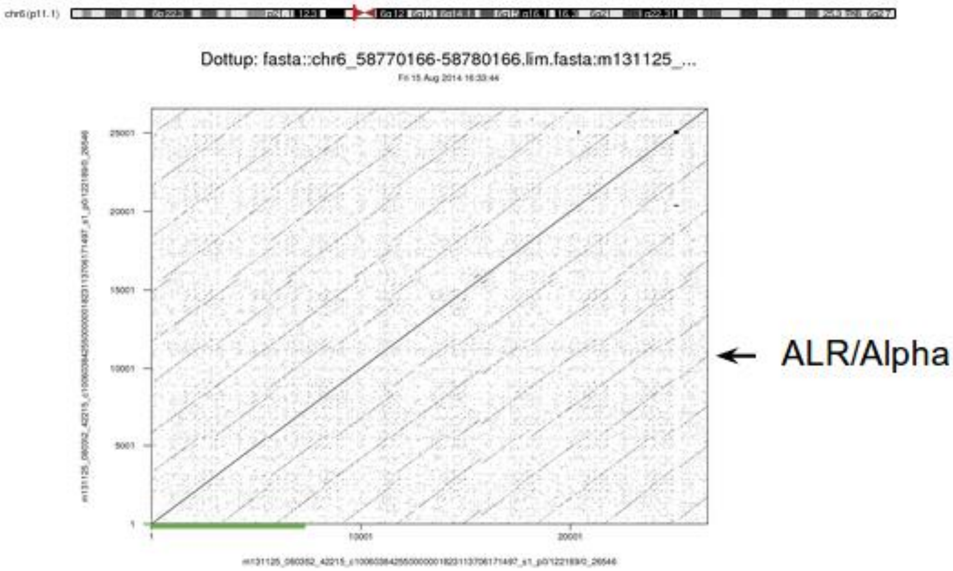
b)


chr12 (q24.33)
Dottup: fasta::chr12_133836895-133841895.lim.fasta:m1310...
Mon 11 Aug 2014 22:24:24
16001
14001
12001
10001
8001
6001
4001
2001
1
1
10001
m131009_185631_42213_c100578462550000001823095604021430_s1_p0/125143/8291_24525
TAR1
(CCCTAA)n

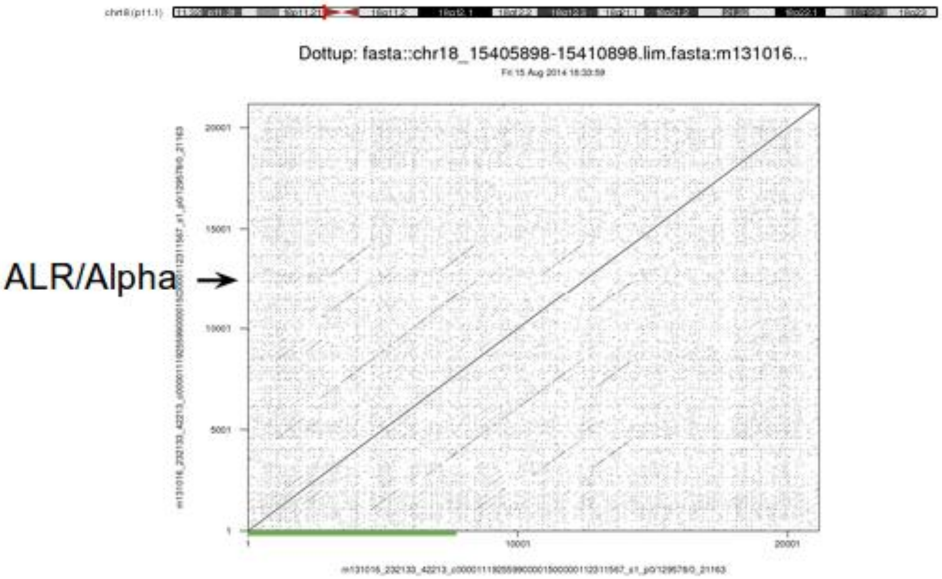


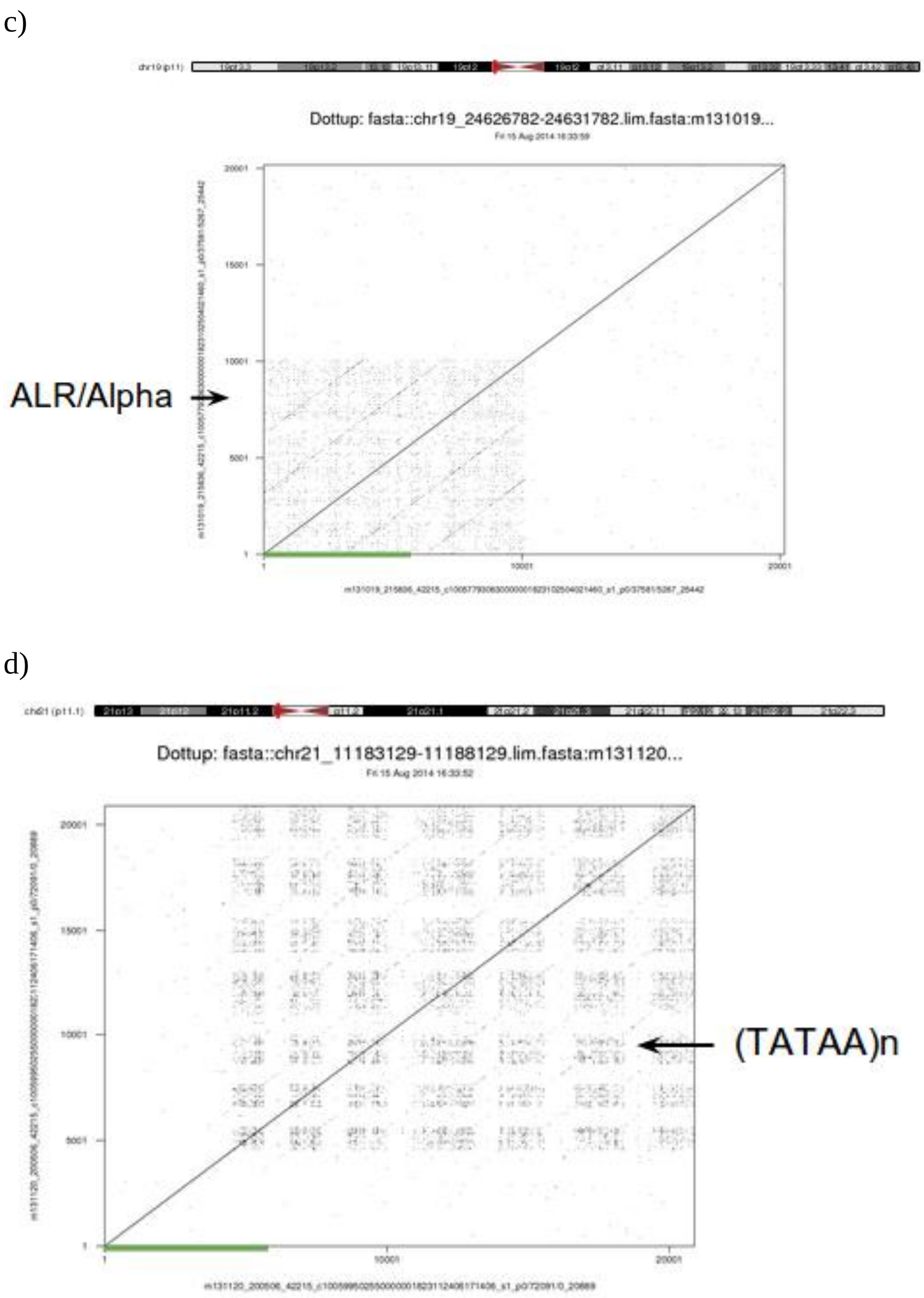
Supplementary Figure 26. Structure of novel sequence from CHM1 PacBio reads extending into telomeric gaps shown by self-self dotplots of single reads. The amount of sequence anchored to the reference at the edge of the gap is shown on the x-axis by the solid green bar. Repeat content is annotated inline. Regions shown include a) chr8q, b) chr12q, c) chr14q, and d) chr21q.

a)


chr6 (p11.1)
Dottup: fasta::chr6_58770166-58780166.lim.fasta:m131125_...
Fri 15 Aug 2014 16:33:44
25001
20001
15001
10001
5001
1
1
10001
20001
m131125_060352_42215_c100603842550000001823113706171497_s1_p0/122189/0_26546
m131125_060352_42215_c100603842550000001823113706171497_s1_p0/122189/0_26546
ALR/Alpha

b)


chr18 (p11.1)
Dottup: fasta::chr18_15405898-15410898.lim.fasta:m131016...
Fri 15 Aug 2014 16:33:59
20001
15001
10001
5001
1
1
10001
20001
m131016_232133_42213_c000011182559900001500000112311567_s1_p0/129578/0_21163
m131016_232133_42213_c000011182559900001500000112311567_s1_p0/129578/0_21163
ALR/Alpha



Supplementary Figure 27. Structure of novel sequence from CHM1 PacBio reads extending into centromeric gaps shown by self-self dotplots of single reads. The amount of sequence anchored to the reference at the edge of the gap is shown on the x-axis by the solid green bar. Repeat content is annotated inline. Regions shown include a) chr6p, b) chr18p, c) chr19p, and d) chr21p.

XII. Integrity of the CHM1 hydatidiform genome

It is possible that the propagation of the hydatidiform mole sample may give rise to structural variants affecting sequences such as the STR/VNTR, and mobile element mosaic (complex) sequences we found to be inserted in the CHM1 sample. To confirm the CHM1 resource is free of artifacts that would confound variant discovery, we screened for the existence of the inserted sequences in other genomes. To this end, we developed a computational screen that searches 30-base substrings (30-mers) unique to the inserted sequences, and checks for their presence of reads in genomes from diverse populations[33] containing these 30-mers. Importantly the reads from the diverse genome are not from cell lines. This analysis supports the following results:

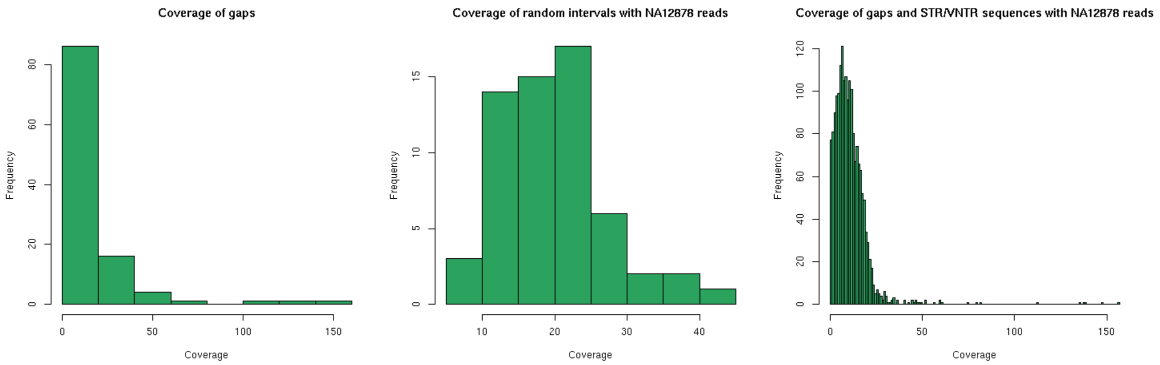
- 484 out of 527 genotyped complex insertions show evidence of presence in other genomes.

- 599 out of 788 STR insertions were confirmed in other individuals. We have added additional detail in the main text to make this clearer.

- All gap sequences are detected in other individuals.

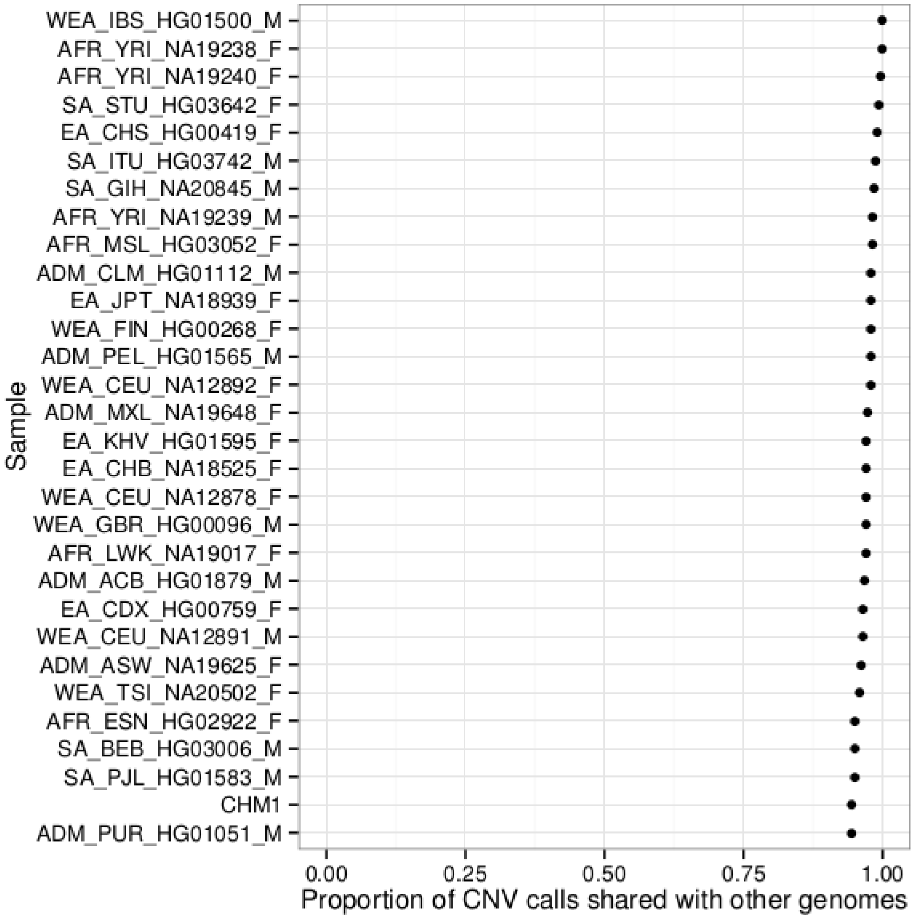
This comparison is only valid to sequence contexts accessible to Illumina technology and therefore biased against regions of high %GC composition. As a control for this effect, we repeated the analysis limiting to regions that were accessible in an Illumina dataset generated for CHM1 (not PCR-free). Of the 458 complex insertions accessible by Illumina, 457 are confirmed in at least one other individual, and similarly of the 463 STR insertions accessible by Illumina, 462 are confirmed in at least one other individual. Thus, 99.8% of the sequences accessible to Illumina sequencing confirm insertion events.

It is valuable to compare other human genomes sequenced by PacBio to check for the presence of structural variants in CHM1. The NA12878 dataset was downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20131209_na12878_pacbio/Schadt/fastq/. We mapped the NA12878 reads to both GRCh37 as well as a patched reference that incorporates both gap and STR and VNTR sequences (GRCh37.patched). All sequences of gap closures are supported by SMRT reads from NA12878, with minimum coverage of 9.39 (average 22.0, sd 20.9). A random sample of 60 1-kbp intervals across the genome has average coverage of 20.4 with a standard deviation of 7.0. The coverage of all gap, STR, and VNTR sequences is more variable, with an average of 7 and standard deviation of 10.8, excluding 12 regions with coverage over 200X (Supplementary Figure 28). Because of the nature of the variable length of STR and VNTR sequences, the coverage is expected to have a wider range. A total of 1388 out of 1833 of the insertions detected in CHM1 show average coverage greater than 5 reads in NA12878. Thus, ~76% of the insertions are confirmed in a second “diploid” genome sequenced using this technology.



Supplementary Figure 28. NA12878 PacBio read coverage in regions of the patched reference. (*left*) Average coverage across all gaps. (*middle*) Coverage in 100 random intervals. (*right*) Coverage in STR regions.

As a final check of the integrity of the CHM1 genome, we searched for the presence of large deletions since these are the most common source of cell-line artifacts that occur as a result of cell line propagation. Specifically, we mapped Illumina reads from CHM1 sequenced at University of Washington and all PCR-free genomes from Illumina (29 genomes) to GRCh37 with mrsFAST, called CNVs in all samples with a digital genomic comparative hybridization (dCGH) algorithm, and determined the proportion of bi-allelic deletions shared by all samples. For all deletion events, we required 1500 unique (non-repeatmasked) basepairs. We identified 110 deletions in CHM1 of which 104 (95%) were shared with at least one other diploid genome. On average all 29 PCR-free samples shared 97% of their deletion calls with at least one other sample (Supplementary Figure 28). CHM1 had 6 private deletions while the PCR-free genomes ranged from 0 to 9 private deletions with a median of 4. These results suggest that CHM1's genomic content is consistent with other gold standard genomes even when it is sequenced with a less robust sequencing protocol.



Supplementary Figure 29. Proportion of CNV calls (bi-allelic deletions) shared by CHM1 and 29 PCR-free Illumina sequenced genomes. CHM1 was sequenced at University of Washington and the standard genomes were sequenced by Illumina with the PCR-free protocol. The proportion of shared calls for CHM1 is consistent with other non-hydatidiform genomes.

XIII. Databases resources

All data from these analyses are available online at the CHM1 Structural Variation website (http://eichlerlab.gs.washington.edu/publications/chm1-structural-variation/). This website includes a patched GRCh37 with inserted STR expansions and gap closures from CHM1 along with BED annotations for the positions of all novel sequences and a track hub that can be viewed through the UCSC Genome Browser.

Whole-genome sequence (WGS) data is available through the NCBI Sequence Read Archive (SRA). PacBio WGS for CHM1 is available with the SRA accession SRX533609. Illumina WGS for CHM1 is available with the SRA accession SRX652547. All clones sequenced for this project are available through accessions listed in Supplementary Table 35.

Supplementary Table 35. Data accession IDs for sequenced BAC and fosmid clones.

Clone	Chromosome	Accession	Group
CH17-12K8	chr10	AC255392	10q11
CH17-147J3	chr10	AC255501	10q11
CH17-149O18	chr10	AC255486	10q11
CH17-153I5	chr10	AC255411	10q11
CH17-177L7	chr10	AC255415	10q11
CH17-177M15	chr10	AC255509	10q11
CH17-183A9	chr10	AC255437	10q11
CH17-183B22	chr10	AC255376	10q11
CH17-214E8	chr10	AC255427	10q11
CH17-214P16	chr10	AC255529	10q11
CH17-224H23	chr10	AC255478	10q11
CH17-242L6	chr10	AC255549	10q11
CH17-24J18	chr10	AC255527	10q11
CH17-287A3	chr10	AC255470	10q11
CH17-306A14	chr10	AC255540	10q11

References

1 Carneiro, M. O. *et al.* Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC genomics* **13**, 375, doi:10.1186/1471-2164-13-375 (2012).

2 She, X. *et al.* The structure and evolution of centromeric transition regions within the human genome. *Nature* **430**, 857-864, doi:10.1038/nature02806 (2004).

3 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).

4 Myers, E. W. *et al.* A whole-genome assembly of Drosophila. *Science (New York, N.Y.)* **287**, 2196-2204 (2000).

5 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, doi:10.1186/1471-2105-13-238 (2012).

6 Nederbragt, L. *et al. Towards correction-free assembly of raw PacBio reads* (AGBT Poster, 2014).

7 Huddleston J, R. S., Malig M, Antonacci F, Chaisson M, Hon L, Sudmant PH, Graves TA, Alkan C, Dennis MY, Wilson RK, Turner SW, Korlach J, Eichler EE. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res* (2014).

8 Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing technology. *Genome research* **24**, 688-696, doi:10.1101/gr.168450.113 (2014).

9 Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods* **10**, 563-569, doi:10.1038/nmeth.2474 (2013).

10 Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-3.0. (1996-2010).

11 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**, 573-580 (1999).

12 Jiang, Z., Hubley, R., Smit, A. & Eichler, E. E. DupMasker: a tool for annotating primate segmental duplications. *Genome research* **18**, 1362-1368, doi:10.1101/gr.078477.108 (2008).

13 Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *Journal of computational biology : a journal of computational molecular cell biology* **7**, 203-214, doi:10.1089/10665270050081478 (2000).

14 Jurka, J., Klonowski, P., Dagman, V. & Pelton, P. CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. *Computers & chemistry* **20**, 119-121 (1996).

15 Stewart, C. *et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS genetics* **7**, e1002236 (2011).

16 Loomis, E. W. *et al.* Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome research* **23**, 121-128, doi:10.1101/gr.141705.112 (2013).

17 Renton, A. E. *et al.* A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257-268, doi:10.1016/j.neuron.2011.09.010 (2011).

18 Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat Genet* **31**, 241-247, doi:10.1038/ng917 (2002).

19 Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87, doi:10.1038/nature04072 (2005).

20 Capra, J. A., Hubisz, M. J., Kostka, D., Pollard, K. S. & Siepel, A. A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. *PLoS genetics* **9**, e1003684, doi:10.1371/journal.pgen.1003684 (2013).

21 Eichler, E. E. *et al.* Haplotype and interspersion analysis of the FMR1 CGG repeat identifies two different mutational pathways for the origin of the fragile X syndrome. *Human molecular genetics* **5**, 319-330 (1996).

22 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).

23 Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS genetics* **9**, e1003709, doi:10.1371/journal.pgen.1003709 (2013).

24 Li, H. Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples. doi:arXiv:1404.0929 (2014).

25 Saha, S. & Raghava, G. P. BTXpred: prediction of bacterial toxins. *In silico biology* **7**, 405-412 (2007).

26 Kimelman, A. *et al.* A vast collection of microbial genes that are toxic to bacteria. *Genome research* **22**, 802-809, doi:10.1101/gr.133850.111 (2012).

27 Hach, F. *et al.* mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications. *Nucleic acids research*, doi:10.1093/nar/gku370 (2014).

28 Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science (New York, N.Y.)* **330**, 641-646, doi:10.1126/science.1197005 (2010).

29 Castel, A. L., Cleary, J. D. & Pearson, C. E. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat Rev Mol Cell Biol* **11**, 165-170 (2010).

30 Kidd, J. M. *et al.* A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**, 837-847 (2010).

31 Hormozdiari, F., Alkan, C., Eichler, E. E. & Sahinalp, S. C. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome research* **19**, 1270-1278, doi:10.1101/gr.088633.108 (2009).

32 Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-712, doi:10.1038/nature08516 (2010).

33 Consortium, G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

34 Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64, doi:10.1038/nature06862 (2008).

35 Kidd, J. M. *et al.* Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nature methods* **7**, 365-371 (2010).

36 Mills, R. E. *et al.* Natural genetic variation caused by small insertions and deletions in the human genome. *Genome research* **21**, 830-839, doi:10.1101/gr.115907.110 (2011).

37 Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics (Oxford, England)* **21**, 1859-1875, doi:10.1093/bioinformatics/bti310 (2005).

38 Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82, doi:10.1038/nature11232 (2012).

39 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).

40 John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* **43**, 264-268, doi:10.1038/ng.759 (2011).

41 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).

42 Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 139-144, doi:10.1073/pnas.0912402107 (2010).

43 Watanabe, Y. *et al.* Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. *Human molecular genetics* **11**, 13-21 (2002).

44 Gilbert, D. M. Replication timing and transcriptional control: beyond cause and effect. *Current opinion in cell biology* **14**, 377-383 (2002).

45 Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annual review of medicine* **61**, 437-455 (2010).