Discovery and genotyping of structural variation from long-read haploid genome sequence data

John Huddleston[1,2], Mark J.P. Chaisson[1], Karyn Meltz Steinberg[3], Wes Warren[3], Kendra Hoekzema[1], David Gordon[1,2], Tina A. Graves-Lindsay[3], Katherine M. Munson[1], Zev N. Kronenberg[1], Laura Vives[1], Paul Peluso[5], Matthew Boitano[5], Chen-Shin Chin[5], Jonas Korlach[5], Richard K. Wilson[4], and Evan E. Eichler[1,2]

1. Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA
2. Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA
3. McDonnell Genome Institute, Department of Medicine, Department of Genetics, Washington University School of Medicine, St. Louis, MO 63108, USA
4. Department of Pathology, University of Pittsburgh, Pittsburgh, PA 15261, USA
5. Pacific Biosciences of California, Inc., Menlo Park, CA 94025, USA

Supplemental Material

Genome sequencing

Genomic DNA preparation and SMRT sequencing

DNA was isolated from two female fibroblast cell lines derived from complete hydatidiform moles (CHM1htert and CHM13htert) 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 and CHM13htert fibroblast cells were harvested at 70-80% confluency (~3x106 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.

CHM1htert Library Preparation

For this work, four different library preparations were made using 30 micrograms of genomic DNA material in each case. The CHM1htert genomic DNA was sheared to a size range of 50 to 60 kbp using the Megarupter® device (Diagenode). Sheared DNA was enzymatically repaired and converted into SMRTbell libraries prepared as described by the manufacturer (Pacific Biosciences®). Non-SMRTbell® DNA molecules were eliminated by digestion with Exonuclease III and Exonuclease VII (New England BioLabs). Finally, a BluePippin preparative electrophoresis purification step was performed (Sage Sciences®) on the library to select library insert sizes ranging from 30 to 80 kbp. These size-selected libraries were used in subsequent sequencing steps.

CHM13htert Library Preparation

Seven different library preparations were made using 10 micrograms of starting genomic DNA material in each case. The CHM13htert genomic DNA was sheared to a size range of 25 to 35 kbp using gTubes (Covaris). Sheared DNA was enzymatically repaired and converted into SMRTbell libraries prepared as described by the manufacturer (Pacific Biosciences®). Non-SMRTbell® DNA molecules were eliminated by digestion with Exonuclease III and Exonuclease VII (New England BioLabs). Finally, a BluePippin preparative electrophoresis purification step was performed (Sage Sciences®) on the library to select library insert sizes ranging from 10 to 50 kbp. These size-selected libraries were used in subsequent sequencing steps.

Sequencing Methods

Sequencing was performed on the Pacific Biosciences RS2 instrument (Pacific Biosciences®). All CHM1 sequencing runs were performed with P6-C4 sequencing chemistry while 70% of the CHM13 runs were performed with P5-C3 chemistry and the other 30% with P6-C4. For all sequencing runs, data was collected using a 6-hour acquisition time to obtain long, raw read data.

A total of 243 single-molecule, real-time (SMRT) cells were processed for CHM1 yielding 62-fold whole-genome sequence (WGS) data while 415 cells were processed for CHM13 yielding 66-fold WGS data. All sequence data for CHM1 and CHM13 are available through the respective SRA accessions SRP044331 and SRP051383 (Supplemental Table S1).

DNA preparation and Illumina sequencing

CHM13htert 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 from two lanes to generate ~31-fold sequence coverage.

Previously published sequences

PCR-based CHM1 Illumina WGS used for single-nucleotide variant (SNV)/indel calling and structural variant (SV) genotyping were obtained from the SRA accession SRX652547 (Chaisson et al. 2015) and realigned to GRCh38. PCR-free CHM1 and CHM13 WGS were generated by the Broad Institute and were obtained through the SRA accessions ERX1413366 and ERX1413367, respectively. Additional WGS from 1000 Genomes Project samples used for SV genotyping were obtained for 24 diversity panel samples (BioProject PRJNA260854), a Yoruban trio (NA19238, NA19239, and NA19240 from BioProject PRJEB4252), and a CEPH trio (NA12878, NA12891, and NA12892 from BioProject PRJNA186949). All 1000 Genomes Project WGS and PCR-free CHM1 and CHM13 WGS were aligned to the 1000 Genomes Project GRCh37 reference.

De novo assembly of SMRT WGS

The CHM1 *de novo* assembly (NCBI accession: GCA_001297185.1) was generated from 61-fold SMRT WGS with FALCON v0.3+. The CHM13 *de novo* assembly (NCBI accession: GCA_000983455.2) was generated from 52-fold SMRT WGS with FALCON v0.4 (Chin et al. 2016). After the draft assembly was created, the following manual steps were applied. If the read on the 5'-end of a contig was not used by another contig, the read sequence was prepended to the contig. Reads that had more than 80 overlaps on either 5'- or 3'-ends were filtered out. These filtered reads were collected and reassembled separately with a higher overlap-count threshold (100,000 overlaps) and included with the primary assembly. Draft assemblies were refined with Quiver (Chin et al. 2013) to produce a high-quality consensus assembly.

Variant discovery in SMRT WGS local assemblies

Alignment of raw reads

SMRT WGS reads were aligned to GRCh38 (excluding alternate haplotypes) with BLASR (built from https://github.com/EichlerLab/blasr.git, commit 0a17c59) (Chaisson and Tesler 2012) using parameters tuned for raw read alignment (-sam -bestn 2 -maxAnchorsPerPosition 100 -advanceExactMatches 10 -affineAlign -affineOpen 100 -affineExtend 0 -insertion 5 -deletion 5 -extend -maxExtendDropoff 50 -clipping subread). SAM output from BLASR was filtered to exclude unaligned reads from downstream analysis.

Selecting regions for local assembly

We detected signatures of SVs in raw read alignments as previously described (Chaisson et al. 2015). Briefly, we identified insertions and deletions in the CIGAR strings of raw read alignments with a mapping quality >=30 and considered as potential SVs all events with support from at least five reads. Additionally, we considered as putative SVs regions where 11 or more reads with mapping quality >=30 clipped within a 500 bp non-sliding window in the reference assembly. Regions with excess clipped reads represented larger SVs that could not always be contained inside five or more reads. In addition to selecting regions with signatures of SVs, we created tiled windows across the entire reference with 60 kbp per window and 20 kbp overlap between adjacent windows to pick up variants missed by the signature detection stage and to assess indels. Finally, we merged the signature SV regions and tiled windows into a single set and filtered to exclude reference assembly gaps, centromeres, and regions with fewer than five reads and more than 100 reads.

Local assembly of raw reads

For each candidate region (signature region or tiled window) across the reference, we produced a local assembly of one or more contigs from the raw reads aligned to that region. To start, we selected all reads mapped in a given region with a mapping quality >= 30 and dumped the sequence data to FASTA and BAM format. FASTA sequences for the region were assembled *de novo* into a draft assembly with PBcR v8.3rc2 (Berlin et al. 2015). We aligned the original raw reads from the region to the draft assembly with BLASR (-bestn 1) and called a final consensus sequence with Quiver v1.0.0 (Chin et al. 2013) and trimmed lowercase bases corresponding to low coverage regions of the assembly. Finally, we aligned the consensus sequence for each region to the corresponding reference sequence with BLASR using parameters tuned for high-quality query sequences (-clipping subread -affineAlign -affineOpen 8 -affineExtend 0 -bestn 1 -maxMatch 30 -sdpTupleSize 13) and adjusted the alignment coordinates from the reference subsequence to the global reference coordinates.

After all local assemblies were created, we collected the resulting sequences and corresponding alignments to the reference into a single BAM and left-aligned all sequences with bamleftalign from the FreeBayes suite (Garrison and Marth 2012).

Inversions

To detect inversions contained within a local assembly, we began by finding the orientation of the assembly: forward or reverse determined by the sparse dynamic programming (SDP) alignment score with 15-mers. The assembly was reverse complemented if determined to be in the reverse orientation. Next, we used the set of 15-mers selected by sparse dynamic programming of the contig in forward orientation as a "backbone" for the forward aligned sequence. We then determined a new set of all 15-mer matches of the reverse complement assembly against the reference. The indexing of this set was 3'->5' relative to the forward orientation, so a match at the first base was indexed at L-15+1, where L was the length of the assembly. In an iterative manner, chains of maximal sets of 15-mers in decreasing order in the assembly and increasing order in the reference were detected using SDP. Each set was divided into subsets such that the maximum gap between any two 15-mers in either the reference or assembly was less than 1 kbp. The sequence in the assembly ranging from the position of the first anchor in a subset to the end of the last anchor defined a candidate inversion. To confirm the presence of an inversion, the subsequence was reverse complemented and the entire assembly was realigned. If the alignment score (defined as the sum of all bases from 15-mers in the SDP optimal alignment) increased by a threshold (50), the candidate inversion was retained as true event. Candidate inversions were merged into a nonredundant set of coordinates annotated by the number of candidate inversions supporting each distinct inversion after merging. A final set of inversions was created by eliminating putative inversions with support from only one local assembly.

SVs (insertions and deletions)

Putative insertions and deletions >=50 bp were identified by gaps in the alignments of local assemblies to the reference assembly as previously described (Chaisson et al. 2015). Gaps from local assemblies were clustered by type by creating a graph such that each gap was a node with edges created between any two nodes that met a minimum overlap requirement in genomic space. For deletions, gaps had to have a reciprocal overlap of 50% or more. For insertions, gaps had to occur within a 20 bp window of each other’s breakpoints. To produce a set of final calls, we identified a representative gap per connected subgraph based on the start coordinate with the most supporting nodes. In the case where a subgraph had no start position with support from multiple gaps, we selected the gap that occurred closest to the middle of its local assembly where the assembly quality has previously been observed to be the best. In addition to reporting the consensus gap for each subgraph, we also reported the number of supporting gaps from all local assemblies and the coverage of all local assemblies across the gap. A final set of insertions and deletions was created by eliminating putative variants with support from only one local assembly and any events occurring inside inversions from the final inversion call set.

Sequences associated with SVs were annotated by Tandem Repeat Finder v4.09 (Benson 1999) using custom parameters (2 7 7 80 10 20 500 -m -ngs -h) and RepeatMasker v3.3.0 (Smit et al. 1996-2004) using human-specific repeat libraries, sensitive alignment mode, and the WU-BLAST alignment engine. SVs were classified by their repeat content based on these annotations (Supplemental Table S2).

Indels

The introduction of local assemblies for tiled windows across the reference allowed SMRT-SV to detect not only SVs but also indels in the size range of 2-49 bp that were too subtle for the SV signature method to detect. Single-base indels were excluded due to the previously reported error rate at that resolution in SMRT WGS assemblies (Chin et al. 2013; Huddleston et al. 2014). Indels were detected by parsing gaps in local assembly alignments to the reference (GRCh38) using a similar approach to gap parsing for SVs. Unlike SV gap parsing, indel gaps were detected in assemblies >=40 kbp without condensing adjacent indels and with an additional filter that excluded paired insertions and deletions of the same size (<4 bp) on either side of one or more matching bases in the alignments. These latter indels were excluded as symptomatic of alignment uncertainty rather than biological events. Rather than filtering indel gaps by their support from local assemblies, we filtered gaps based on a single tiling path of representative local assembly sequence through the genome. The tiling path was created by segmenting each chromosome by the beginning and ending coordinates of the mapped local assemblies and selecting the local assembly in each segment whose midpoint was closest to the midpoint of the segment. We then retained only those indel gaps that were identified in the representative local assembly in the tiling path where each gap occurred. Indels were additionally filtered to exclude events in homopolymer runs or previously detected SVs, including inversions.

SNVs

Although SMRT-SV was not designed to detect SNVs, such variants were present in alignments of local assemblies from tiled windows across the reference. We parsed these alignments to report all single-base mismatches between local assemblies and the reference and retained all SNVs with support from more than one local assembly.

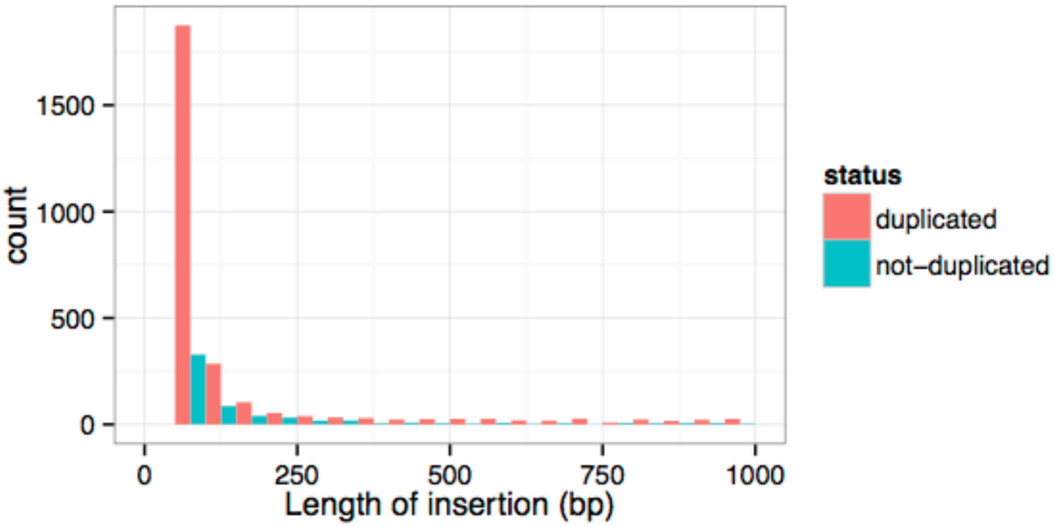
Accounting for alternate haplotypes in GRCh38

We determined the extent to which local assemblies from both CHM1 and CHM13 aligned better to alternate haplotypes from GRCh38 than the primary chromosome sequence. To this end, we selected all local assemblies from regions with alternate haplotypes and aligned these assemblies back to GRCh38 primary and alternate haplotypes with BLASR (-affineAlign -affineOpen 8 -affineExtend 0 -bestn 1 -maxMatch 30 -sdpTupleSize 13). High mapping quality from these alignments (>=254) corresponded to the best placement of an assembly to either primary or alternate haplotypes while lower mapping qualities indicated ambiguity in sequence placement either as the result of shared sequence between primary and alternate haplotypes or segmental duplications in any of the haplotypes. Of 4,401 CHM1 assemblies originally mapped to primary chromosome loci with alternate haplotypes, 3,915 (89%) aligned unambiguously to GRCh38 primary or alternate haplotypes. Of those unambiguously aligned assemblies, 2,343 (60%) mapped to the primary haplotype locus while the remaining 852 assemblies (27%) mapped to 141 distinct alternate haplotype loci. Similarly, 2,992 of 4,528 (66%) CHM13 assemblies mapped unambiguously with 2,130 (71%) mapped to their original primary haplotype loci and

862 (29%) mapped to alternate haplotype loci. The assemblies that mapped better to alternate haplotypes were responsible for 314 of 20,555 (2%) SVs (insertions and deletions) from CHM1 and 434 of 20,424 (2%) SVs from CHM13. These results are consistent with the fact that alternate haplotypes exist for ~61 Mbp (2%) of distinct sequence in GRCh38, several megabases of which was constructed from CHM1 BAC sequences. Finally, we called SVs in alternate haplotypes using the BLASR alignments described above. We identified a similar number of SVs in alternate haplotypes compared to the primary haplotypes (221 SVs from CHM1 and 355 from CHM13) suggesting that, while 2% of local assemblies are more similar to alternate haplotypes, the number of SVs discovered does not vary noticeably between haplotypes.

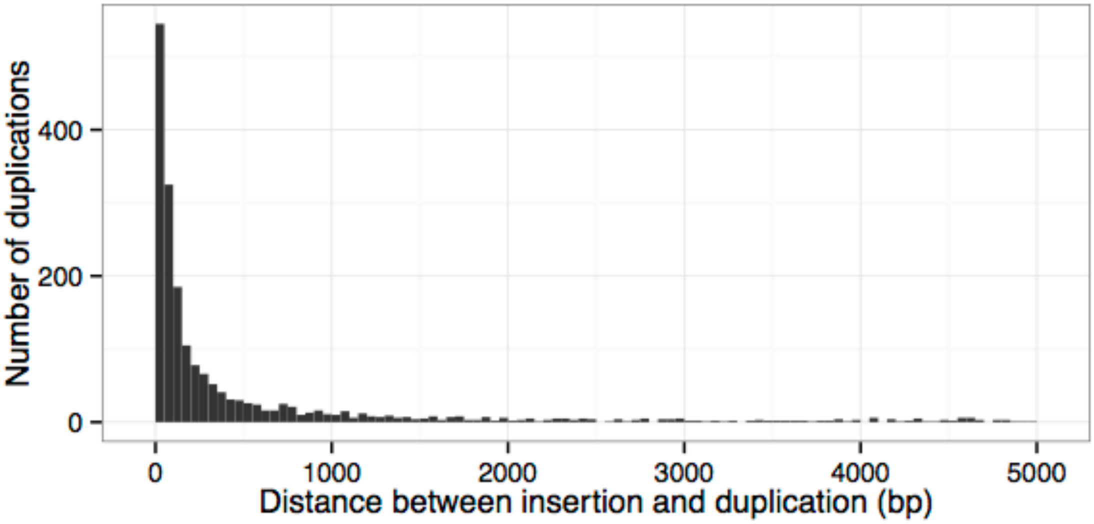
Duplications

We assessed how many insertions from SMRT-SV could be classified as duplications mapping all insertions from the theoretical diploid that were not mobile element insertions (MEIs), tandem repeats, or satellites back to GRCh38 with BWA-MEM (v. 0.7.3 with custom parameter of -E 0,0). We flagged as "duplications" all insertions that had primary or secondary alignments with at least 50% of their sequence in the reference. Of the 18,501 insertions in the theoretical diploid, 3,921 (21%) had the repeat content described above. After alignment and filtering for at least 50% of original sequence present in the reference, we found that a large fraction 3,262 (83%) of the aligned insertions were, in fact, duplications (Supplemental Figure S1and Supplemental Table S11).



Supplemental Figure S1. Length of insertions classified as either duplicated or not duplicated. Events >1 kbp (638 or 16%) are not shown but follow the same pattern as events >500 bp.

We found 2,347 (72%) duplications on the same chromosome as their corresponding insertions and 915 (28%) on different chromosomes. Of the events mapping to the same chromosome, the median distance between midpoints of the best duplication mapping and the original insertion coordinates was 207 bp with 87% of duplications mapping within 5 kbp of their insertion (Supplemental Figure S2). A subset of 550 duplications (17%) overlapped by 50% with their original insertion coordinates.

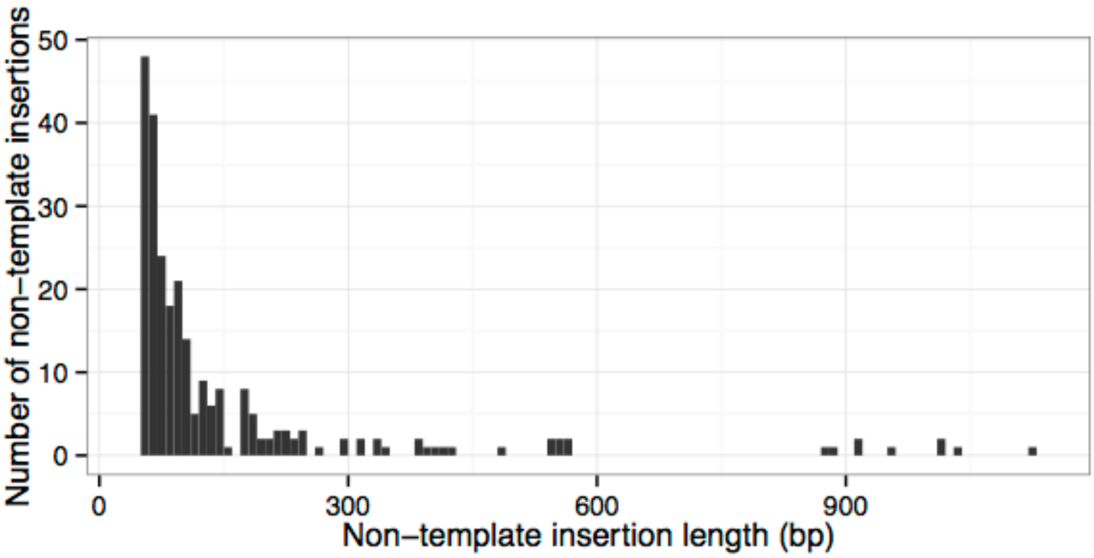


Supplemental Figure S2. Distance between midpoints of insertion coordinates and corresponding best duplication position on the same chromosome. Events with distances <5 kbp are shown (n=2,837 or 87%) with 425 additional events in the long tail of the distribution.

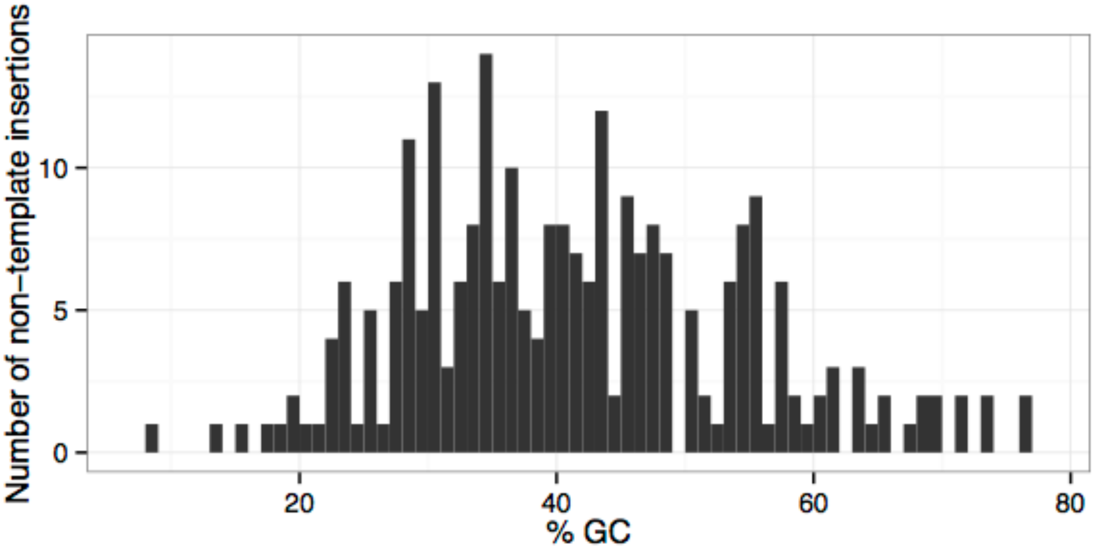
Non-template insertions

To determine which insertions might correspond to non-template sequence, we applied the following series of filters to the sequence content of all insertions in CHM1 and CHM13. We first masked insertion sequences with RepeatMasker (v3.3.0) and Tandem Repeat Finder (TRF) and discarded any insertions with 50% of their sequence annotated as repeats. Next, we eliminated insertions that mapped inside an existing tandem repeat in GRCh38 but whose sequences were too short to be annotated by TRF. By the same logic, we eliminated insertions that were not annotated by RepeatMasker as satellites but mapped within 5 Mbp of an acrocentric region where variants have been observed to be enriched for alpha satellites. We aligned the remaining insertions back to GRCh38 with BLAT to remove processed pseudogenes and other template-based insertions and filtered insertions with >50% of sequence aligned to the reference. Finally, we accounted for template sequences potentially missing from the reference by aligning all remaining insertions to the chimpanzee reference (panTro4) with BWA-MEM (v0.7.3) and to NCBI's NT database with BLAST. After applying all filters, we found that 124 (1%) of 12,998 CHM1 insertions and 129 (1%) of 13,118 CHM13 insertions passed all template filters suggesting that these insertions represent non-template sequence (Supplemental Table S10).

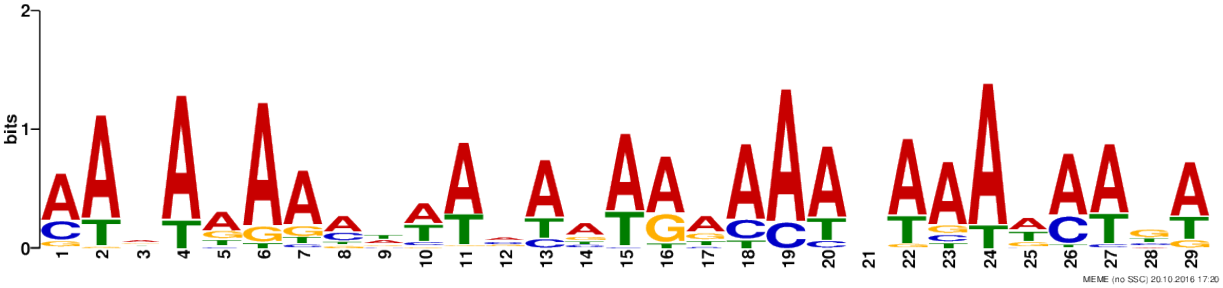
The majority of non-template insertions were less than 300 bases (Supplemental Figure S3). The GC content of non-template insertions was normally distributed with a mean of 41% and a range of 8-76% (Supplemental Figure S4). We applied the MEME motif-finding algorithm (Bailey and Elkan 1994) to all 253 sequences to identify any potential sequence patterns shared by these insertions. Using this approach, we identified three motifs of which the most common was a 29 bp A-rich motif found in 89 (35%) of insertions with an E-value of 4.5 x 10^{-31} (Supplemental Figure S5). The next most common motif was a 35 bp G-rich sequence found in 28 (11%) of insertions. The pattern of an A-rich motif could result from reverse transcription as performed by transposable elements although a better understanding of the mechanisms for these non-template insertions will require further work in the future.



Supplemental Figure S3. Length distribution of non-template insertions (bp).



Supplemental Figure S4. GC content distribution of non-template insertions with a mean of 41% and a range of 8-76%.



Supplemental Figure S5. Most common motif in non-template insertions as identified by the MEME motif-finding algorithm. The motif was present in 35% of insertions.

Variant discovery in SMRT WGS *de novo* assemblies

In addition to detecting variants with local assemblies of CHM1 and CHM13 SMRT WGS, we detected variants in complete *de novo* assemblies of the same SMRT WGS for both samples. As alignment-based variant detection can be highly sensitive to the specific alignment tool and parameters, we attempted to replicate the same conditions for *de novo* assembly alignments as were used for local assembly alignments. To this end, we fragmented both *de novo* assemblies into 60 kbp fragments in 20 kbp sliding windows and aligned these fragments to the human reference (GRCh38) with BLASR using the same parameters used for local assembly alignment described above. Unlike the alignment of local assemblies, we did not know beforehand which locus of the reference each *de novo* assembly fragment corresponded to and relied instead on alignment of fragments to the entire reference and filtering of alignments by mapping quality >=30. Fragments that aligned to the reference in reverse orientation were reverse complemented and realigned prior to variant calling to ensure consistent alignments by BLASR across tandem repeats. We called SVs, indels, and SNVs from all alignments mapping in forward orientation using the standard SMRT-SV pipeline as described for local assemblies.

Variant validation

Validation by BACs and *de novo* assemblies of SMRT WGS

We assessed the accuracy of SVs and indels detected in local assemblies by SMRT-SV by aligning high-quality BAC insert sequences from CHM1 and CHM13 to GRCh38 with BLASR and calling variants with SMRT-SV.

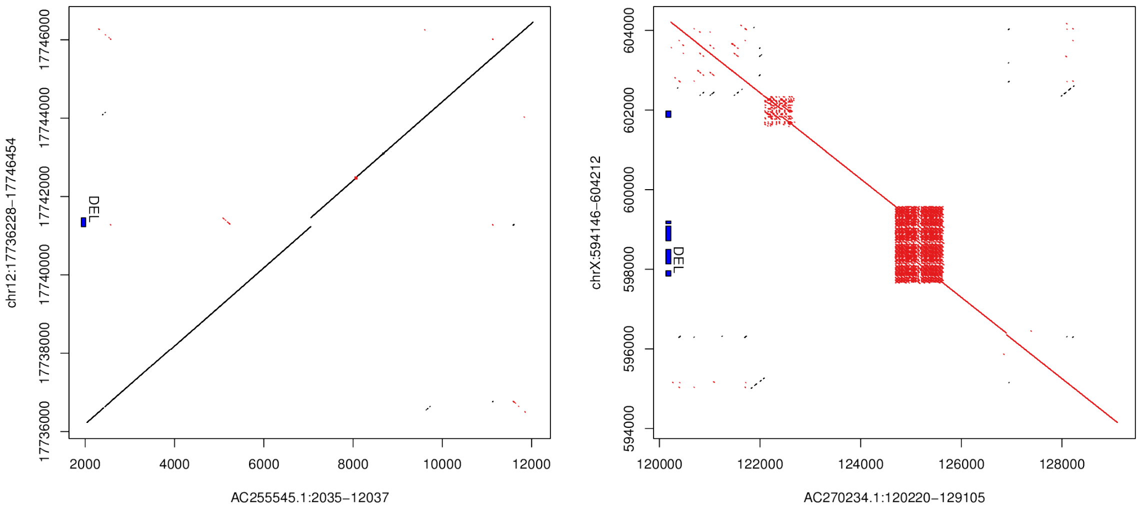
We first obtained all completely assembled (i.e., "finished") and publicly available CHM1 and CHM13 clones from NCBI based on the clone report for each sample's respective BAC library (CH17: ftp://ftp.ncbi.nih.gov/repository/clone/reports/Homo_sapiens/CH17.clone_acstate_9606.out and VMRC59: ftp://ftp.ncbi.nih.gov/repository/clone/reports/Homo_sapiens/VMRC59.clone_acstate_9606.out) resulting in 677 BACs for CHM1 and 23 for CHM13. We supplemented the public set of CHM13 clones by selecting an additional 30 BACs at random from the CHM13 library (VMRC59) and sequencing these BACs with PacBio P6/C4 chemistry in 10 pools of three clones. We assembled the full length inserts for 20 of these clones using the standard HGAP/Quiver pipeline (Chin et al. 2013; Huddleston et al. 2014).

As with variant calling from *de novo* assemblies of SMRT WGS, we fragmented BACs into 60 kbp fragments with 20 kbp sliding windows to ensure the most similar alignment conditions to local assemblies as possible. BAC fragments were aligned to GRCh38 with BLASR using the same parameters used for local assembly alignments and alignments were filtered to exclude segmentally duplicated or heterochromatic regions of the genome. We additionally excluded six CHM1 BACs from analysis that had been previously flagged as potential cases of swapping with another sample (accessions: AC243686.3, AC244166.2, AC243914.3, AC246794.2, AC245464.3, and AC243946.3). We applied SMRT-SV to these filtered BAC fragment alignments to identify SVs (insertions and deletions) and indels. After alignment filtering and variant calling, we retained 30 CHM1 BAC inserts (6.1 Mbp) and 17 CHM13 BAC inserts (2.7 Mbp).

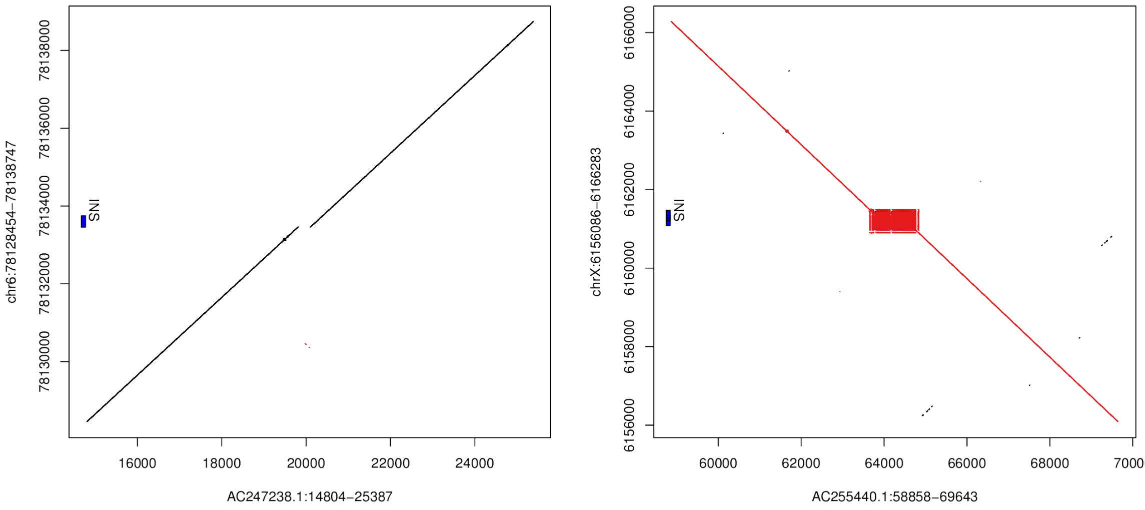
We calculated the concordance between local assembly and BAC variants by limiting local assembly variants to those regions where BACs aligned with mapping quality >=30 and intersecting calls from both sets with a requirement of 50% reciprocal overlap for both deletions and insertions. Using these requirements, we confirmed 60 of 70 CHM1 SVs (86%) and 25 of 26 CHM13 SVs (96%). Similarly, we confirmed 863 of 1,025 CHM1 indels (84%) and 390 of 401 CHM13 indels (97%) (Supplemental Table S3).

Next, we calculated concordance between variants from local assemblies and *de novo* assemblies of SMRT WGS in regions of the reference where both sets of sequences had high-quality alignments (mapping quality >=30). We required calls to have a 50% reciprocal overlap to be considered validated. We found stronger support from WGS variants with 15,850 of 16,454 CHM1 SVs (96%) supported by the CHM1 *de novo* assembly and 15,555 of 16,266 CHM13 SVs (96%) supported by its *de novo* assembly. When we investigated only local assembly variants that could be assessed by either BACs or *de novo* assemblies, we found support for all 96 of the combined CHM1 and CHM13 SVs and 1,349 of 1,426 combined indels (95%) (Supplemental Table S3).

In addition to applying the SMRT-SV pipeline for confirmation, as described above, we generated dot plots for each BAC-validated SV based on the assembled BAC sequence and corresponding region of GRCh38. We mapped 5 kbp of sequence from either side of each event back to GRCh38 and the assembled BAC sequence. We generated dot plots of the corresponding subsequences with the BAC on the x-axis and GRCh38 on the y-axis such that a deletion corresponded to a vertical gap and an insertion to a horizontal gap.Supplemental Figure S6. below shows an example of a simple deletion (left panel) and a complex deletion that is the results of a tandem repeat expansion in the reference (right panel).Supplemental Figure S7. shows an example of a simple insertion (left panel) and a complex insertion (right panel). Based on this computationally orthogonal validation, we visually confirmed of 85 of 94 variants (Supplemental Table S3).



Supplemental Figure S6. Dot plots between SV-containing BAC sequences on the x-axis and the corresponding region of GRCh38 on the y-axis including a simple deletion (left panel) and a complex deletion resulting from a tandem repeat expansion in the reference (right panel). Both events were identified in CHM1.



Supplemental Figure S7. Dot plots between SV-containing BAC sequences on the x-axis and the corresponding region of GRCh38 on the y-axis including a simple insertion (left panel) and a complex insertion (right panel). Both events were identified in CHM1.

Validation of CHM1 and CHM13 SVs and indels by PCR and Sanger sequencing

We performed additional validation of smaller SVs that could be reasonably amplified by PCR and Sanger sequenced. Specifically, we selected SVs between 50-500 bp that did not occur inside tandem repeats or segmental duplications and did not overlap with previously published SVs. We additionally filtered out SVs that occurred within 500 bp of another SV to prevent such adjacent variants from affecting PCR primer design.

Altogether we targeted 288 SVs (253 insertions and 35 deletions) from CHM1 and CHM13 for PCR and Sanger sequencing. PCR primers were initially designed from the SV reference coordinates with 500 bp added to either side using Primer3 allowing at most 10 primers per input site and a primer size range of 18-27 bp (Untergasser et al. 2007). All potential primers were aligned back to GRCh38 with BLAT (Kent 2002) using more sensitive alignment parameters than the defaults (-minMatch=1 -minScore=18). We then selected primer pairs such that each sequence in the pair mapped uniquely to the entire reference assembly. SVs were amplified and sequenced in both CHM1 and CHM13 regardless of which genome the SVs originated from. After amplification the PCR product was cleaned and sequenced and the resulting traces were aligned to the reference sequence in Sequencher v5.4.5 (Gene Codes Corporation, Ann Arbor, MI, USA, http://www.genecodes.com) to determine the presence of the SVs.

Of the 288 targeted SVs, 158 could be amplified and sequenced, including 49 SVs from CHM1, 53 from CHM13, and 56 SVs called in both samples corresponding to 105 CHM1 variants and 109 CHM13 variants. Sanger sequencing supported 95 (90%) SVs in CHM1 and 105 (96%) of SVs in CHM13 for a PCR-based validation rate of 200 out of 214 variants (93%) (Supplemental Table S4). Combined with computationally orthogonal validations by BAC dot plots, we find an overall validation rate for SVs of 95% (285 / 300).

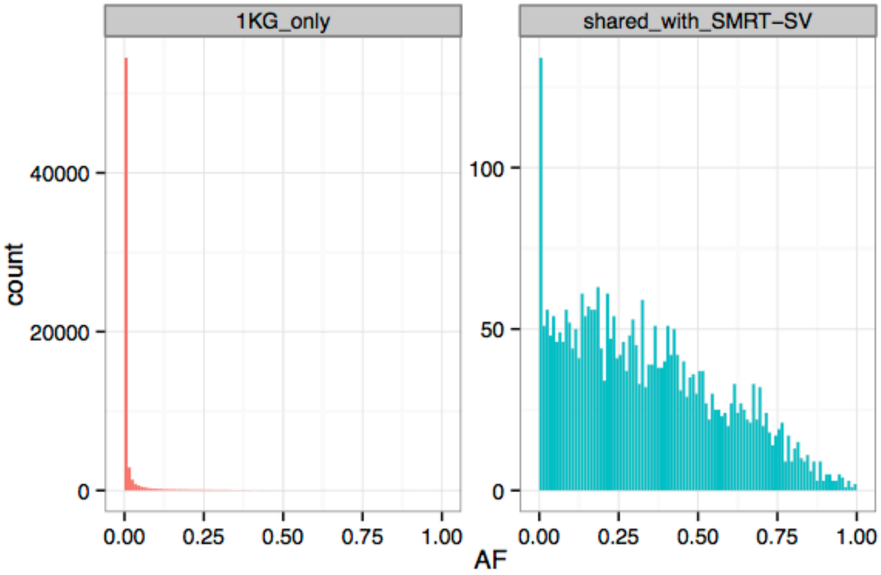
In addition to targeted validation of SVs, we also selected 48 indels for validation by PCR and Sanger sequencing including events found only in CHM1 (n=17), only in CHM13 (n=18), and in both samples (n=13). We generated PCR amplicons for 39 indels (81%): 13 indels from CHM1, 14 from CHM13, and 12 from both samples corresponding to 51 independent validation attempts. We confirmed 50 of 51 sites for an indel validation rate of 98%. Combined with the above SV validations by PCR, we found an overall validation rate of 94% (255 / 270).

Comparison of SVs with 1000 Genomes Project Phase 3 variants

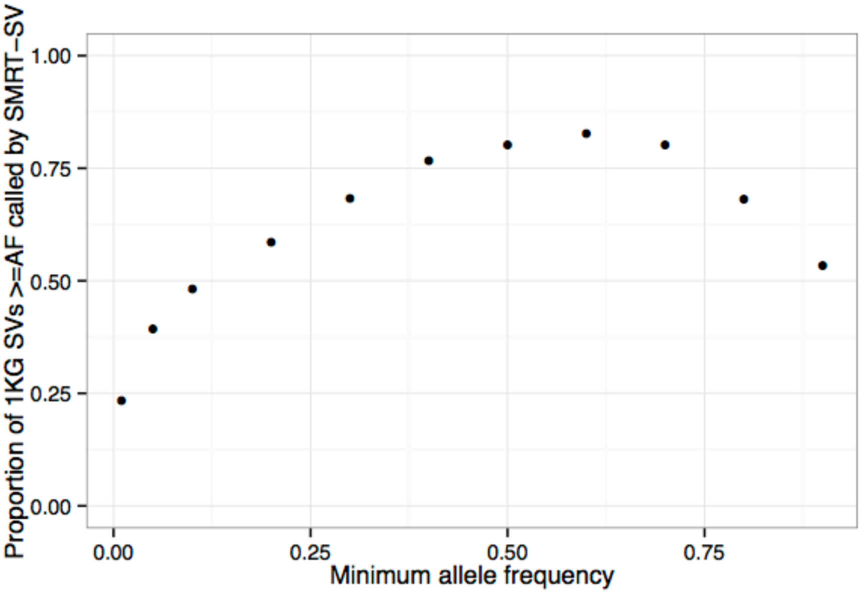
To determine whether any common variants in 1000 Genomes Project Phase 3 (note, “1KG” in this study refers to the 1000 Genomes Project Phase 3) were missed by SMRT-SV calls from two haploid humans, we compared the allele frequencies of all 1KG-only variants with 1KG/SMRT-SV shared variants. The vast majority of variants exclusive to the 1KG were rare with 54,485 of 64,557 SVs (84%) at <1% allele frequency (Supplemental Figure S8). Another 4% (2,613) of these 1KG-only variants had a frequency >10%. The mean allele frequency of 1KG-only variants was 1.6% compared to a mean frequency of 34% for 1KG/SMRT-SV shared variants. SMRT-SV identified 23% (3,071) of the 13,143 polymorphic SVs in 1KG (AF >= 1%) and 48% (2,613) of 5,418 SVs with an allele frequency >=10% (Supplemental Figure S9).

It is difficult, however, to attribute the absence of higher frequency 1KG calls in CHM1/CHM13 SMRT-SV variants to the presence of the minor allele in the moles or to false negatives due to sensitivity of our method. To get a better understanding of how much variation we expect to be shared between any given human and the rest of the 1KG, we calculated the distribution of shared SVs between any single sample in the 1KG call set and all other samples. Specifically, for each 1KG sample we determined which SVs were present in that sample (i.e., SVs with non-reference genotypes), which SVs were present in all other samples, and which SVs were shared between the two sets. We included SVs from autosomes and sex chromosomes in this analysis. Using this approach, we identified a sex-matched bimodal distribution with an average of 3,342 shared variants for females and 5,307 for males (Supplemental Figure S10). Additionally, we found that the number of variants shared by the effectively female CHM1/CHM13 theoretical diploid and 1KG calls (n=3,185) fell within the expected range of shared variants for other females. We note that the number of shared calls between 1KG samples is unavoidably inflated by the approach of joint variant calling and genotyping used by the 1KG SV group.

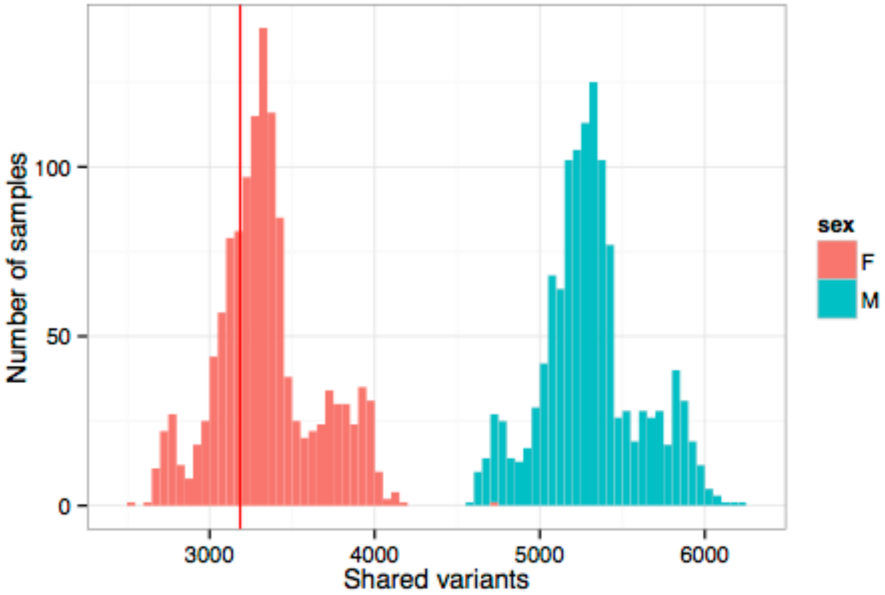
To understand what our false negative rate (FNR) might be with respect to 1KG-only variants, we inspected SMRT-SV local assembly coverage across regions with 1KG-only SVs to determine how many 1KG-only variants could be the result of missing local assemblies and therefore missing SMRT-SV calls. Of the 64,557 1KG-only SVs, 219 (0.3%) did not have two or more overlapping local assemblies from both CHM1 and CHM13 and were present at allele frequencies >1% in 1KG such that we could expect these variants to be shared by the theoretical diploid. Based on this conservative assumption that all 219 variants should have been discovered in CHM1/CHM13, we calculated an upper bound on our FNR of 6% from the number of shared (n=3,185) and missed (n=219) variants.



Supplemental Figure S8. Allele frequency distribution of 1KG SVs by status as shared with SMRT-SV or specific to 1KG only.



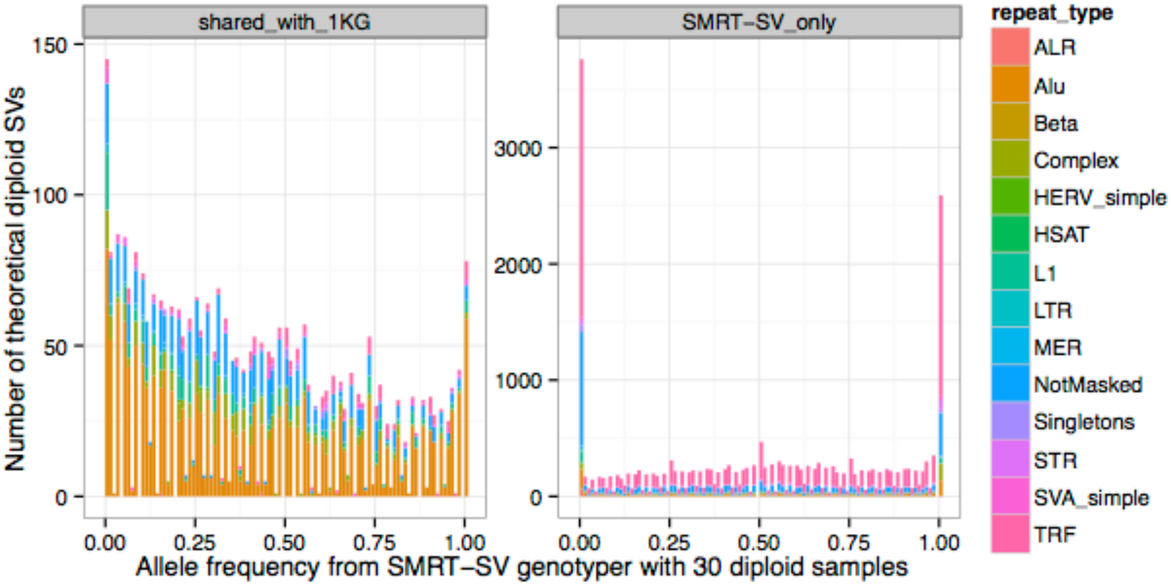
Supplemental Figure S9. Proportion of 1KG SVs identified by SMRT-SV by minimum allele frequency of SVs.



Supplemental Figure S10. Distribution of shared SVs between any single sample from the 1KG and all other samples. The number of SVs shared between the CHM1/CHM13 theoretical diploid and 1KG samples (n=3,185) is shown by a vertical red line.

We also assessed the allele frequency of variants found only by SMRT-SV and not in the 1KG calls. For this analysis, we repeated the comparison of SMRT-SV theoretical diploid variants with 1KG requiring SMRT-SV variants to be seen in both CHM1 and CHM13. In our original comparison with all CHM1/CHM13 variants, we observed 26,807 novel variants (89%) out of 29,992 total (Figure 1b). In this new analysis with only variants observed in both hydatidiform moles, we identified 9,682 variants found in both CHM1 and CHM13. Of these variants, 8,611 (89%) were not observed in the 1KG SVs and 1,071 were shared.

Next, we used genotyping data from SMRT-SV Genotyper applied to 30 samples across CHM1 and CHM13 SV breakpoints to assess the allele frequency distribution of SMRT-SV calls shared with 1KG and unique to SMRT-SV where each call was required to have a genotype in at least one of the 30 diploid or two haploid samples (n=25,767 out of 29,992 variants). The mean allele frequencies (AFs) were similar for the 3,184 shared variants (41% mean AF) and 22,583 SMRT-SV-only variants (49% mean AF). However, the shape of the frequency distributions of the two sets was qualitatively different (Supplemental Figure S11). SVs shared with 1KG were relatively evenly distributed across the frequency spectrum while SVs specific to SMRT-SV were bimodal with peaks at frequencies of 0% (n=3,761) and 100% (n=2,589). The former variants were private to CHM1 or CHM13 (3,197 or 85%). The latter variants at 100% frequency are likely regions where the reference contains the minor allele or defines a misassembly. SVs shared between SMRT-SV and 1KG were primarily Alus (57%) while most SVs specific to SMRT-SV were tandem repeats (61%) consistent with our previous observation that most variable number tandem repeats (VNTRs) and STRs (Chaisson et al., 2015) have been missed. These latter variants are notoriously difficult to discover with short-read data, which is likely why these variants were not observed by the short reads used in the 1KG. Nevertheless, we show these missing tandem repeats can generally be genotyped with short reads when their precise breakpoints have been assembled and that their allele frequencies from 30 diploid genomes are evenly distributed. If we recalculate the overlaps between SVs from SMRT-SV and 1KG with the requirement that SVs have an allele frequency >1% from SMRT-SV genotyper, we find that 86% (18,822 of 21,861) of SMRT-SV variants are novel. We conclude that the differences we find between long-read variant calls from two haploid genomes and short-read variant calls from 2,504 genomes are rooted in a lack of sensitivity of previous methods and not differences in variant allele frequencies.

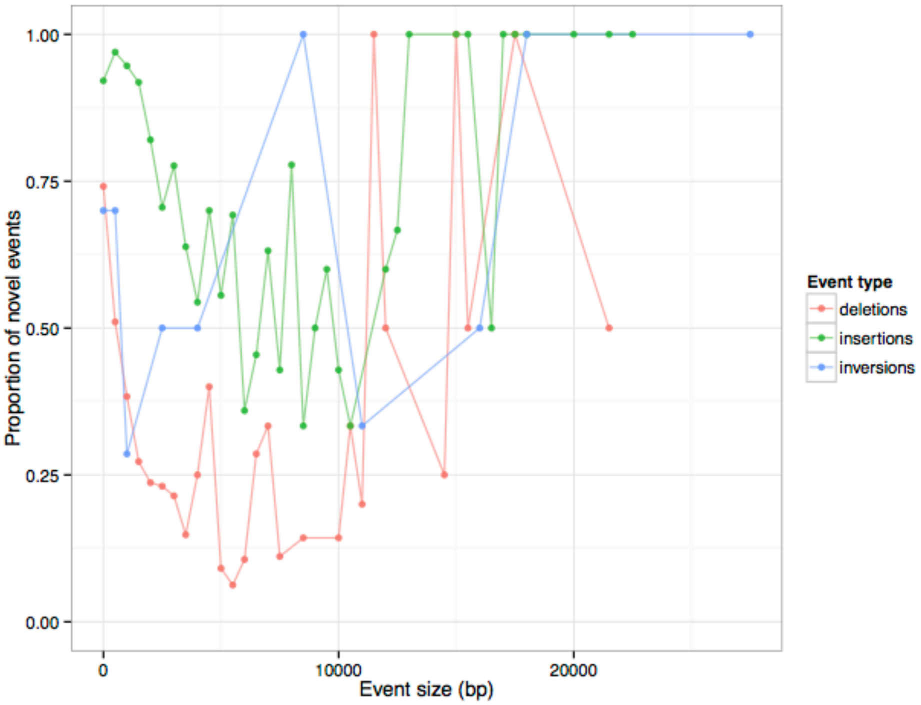


Supplemental Figure S11. Number of theoretical diploid SVs from SMRT-SV by allele frequency from SMRT-SV genotyper with 30 diploid samples. SVs are shown by their status as shared with 1KG or distinct to SMRT-SV.

Comparison of SVs and indels with previously published variants

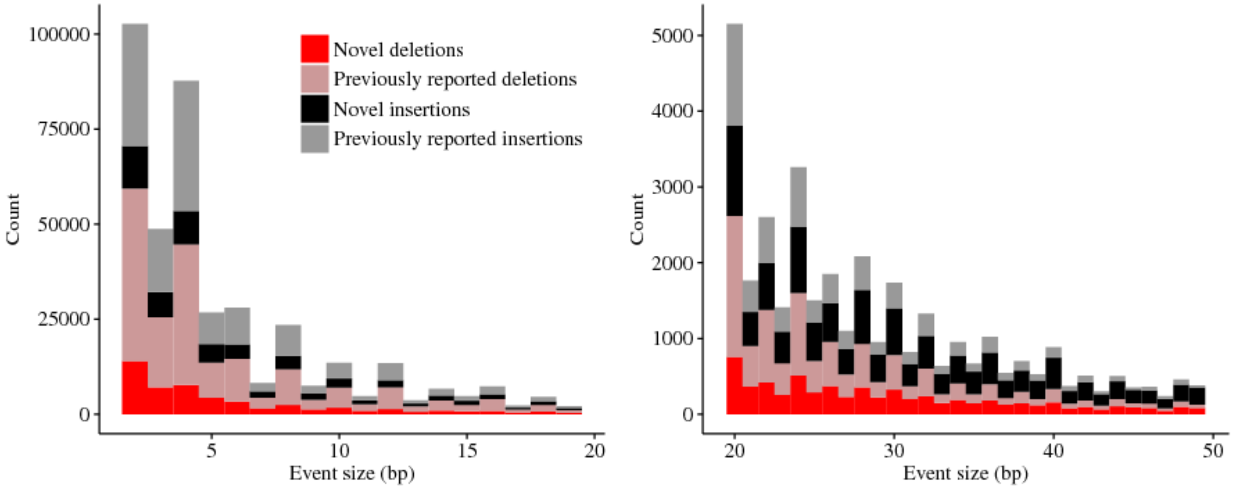
We obtained insertion and deletion calls from previous studies of SVs (Conrad et al. 2010; Kidd et al. 2010; Mills et al. 2011; Sudmant et al. 2015a; Sudmant et al. 2015b) and lifted over all SVs to GRCh38 for comparison with SMRT-SV variants. SVs from all studies were clustered by type using a 50% reciprocal overlap to produce a nonredundant set of SVs. We then determined which SVs from the CHM1/CHM13 theoretical diploid were shared with previously published SVs by intersecting calls with a 50% reciprocal overlap. We found that 7,926 (69%) deletions and 16,931 (92%) insertions were not previously observed. We observed the greater proportion of novel insertions and deletions for variants <1 kbp (Supplemental Figure S12).

We obtained inversion calls from the Database of Genomic Variants (MacDonald et al. 2014) for the GRCh38 release. As with insertions and deletions, we clustered inversions into a nonredundant set of variants based on 50% reciprocal overlap and compared these nonredundant inversions with inversions from the theoretical diploid. We found that 30 (43%) inversions were previously unobserved.

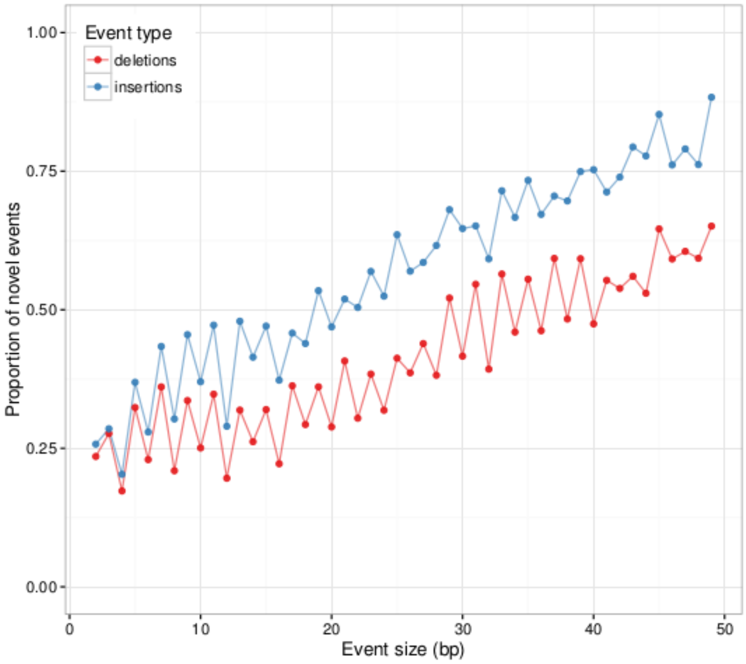


Supplemental Figure S12. Proportion of SMRT-SV SVs from the theoretical diploid of CHM1 and CHM13 not previously reported in SV studies (insertions and deletions) or the Database of Genomic Variants (inversions). Variants were intersected with previous call sets and SMRT-SV variants not overlapping by 50% reciprocal overlap were considered as "novel."

We obtained previously published indels from dbSNP build 146 (Sherry et al. 2001) in GRCh38 coordinates (ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/All_20151104.vcf.gz). We intersected SMRT-SV indels from the theoretical diploid with dbSNP indels requiring a 50% reciprocal overlap for SMRT-SV variants to be considered previously reported. We found that 65,151 (32%) insertions and 57,296 (25%) deletions from the theoretical diploid were not previously reported (Supplemental Figure S13). Interestingly, the proportion of these novel events increases linearly by indel size (Supplemental Figure S14) suggesting that databases of genetic diversity will benefit most from the inclusion of SMRT WGS indels >20 bp.



Supplemental Figure S13. Indels (2-49 bp) detected by SMRT-SV and compared to previously published indels from dbSNP 146 (Sherry et al. 2001). Indels from SMRT-SV that have a 50% reciprocal overlap with dbSNP indels are indicated as "previously reported" in light red and light gray for deletions and insertions, respectively.



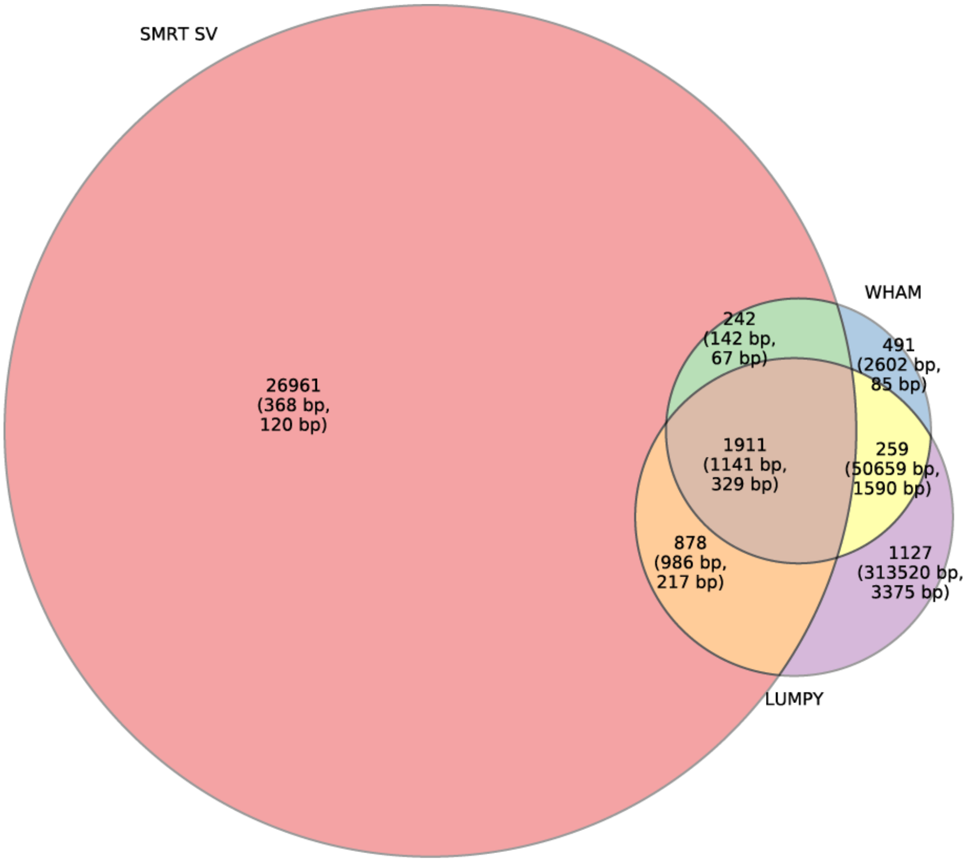
Supplemental Figure S14. Proportion of SMRT-SV indels from the theoretical diploid of CHM1 and CHM13 not previously reported in dbSNP. Variants were intersected with dbSNP build 146 indels and SMRT-SV variants not overlapping by 50% reciprocal overlap were considered as "novel."

Comparison of SVs with Illumina call sets for CHM1 and CHM13

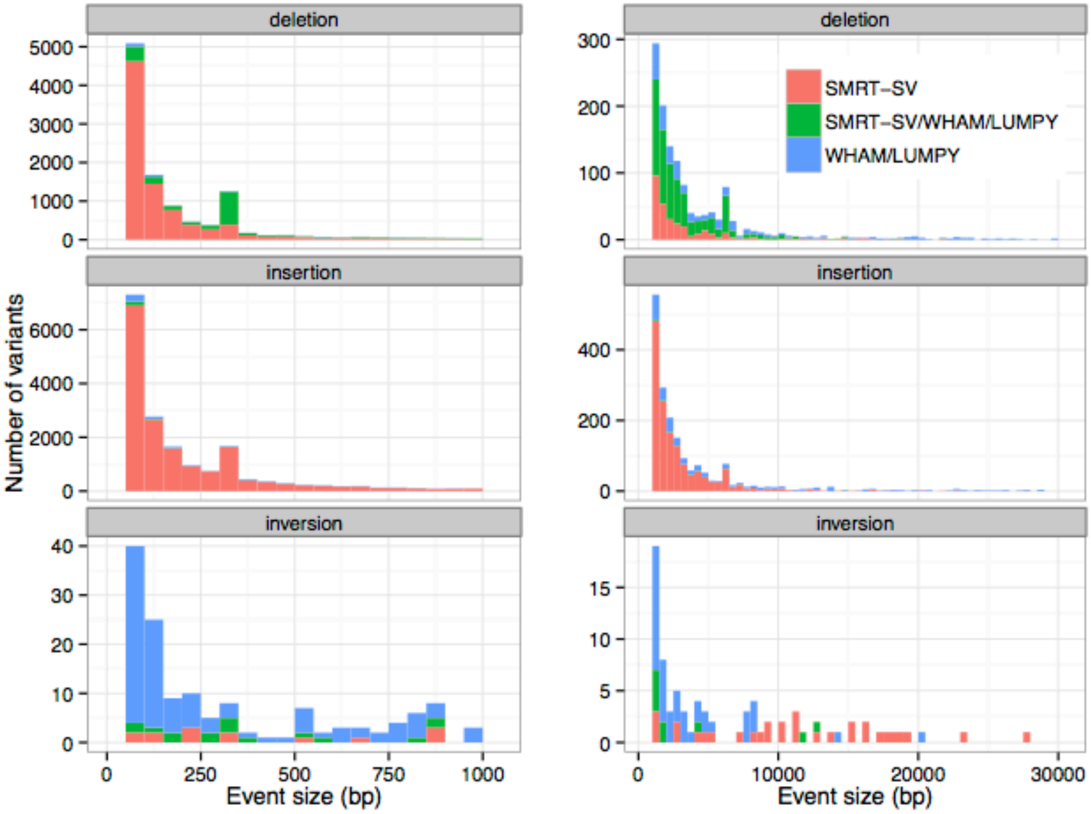
To determine the sensitivity of SVs from SMRT-SV, we called SVs from Illumina data for CHM1 and CHM13 (PCR-free 151 bp paired reads from respective SRA accessions ERX1413366 and ERX1413367) using WHAM (Kronenberg et al. 2015) and LUMPY (Layer et al. 2014). Specifically, we realigned reads from CHM1 and CHM13 to GRCh38 (excluding alternate haplotypes) with BWA-MEM v0.7.12-r1039. We called SVs (insertions, duplications, deletions, and inversions) with WHAM v1.7.0-272-g078c-dirty and filtered out calls with fewer than seven supporting reads and <50 bp in size. We called SVs (insertions, duplications, deletions, inversions, and copy number variants) with LUMPY v0.2.13 and filtered out calls with fewer than seven supporting reads and <50 bp in size.

WHAM detected 789 insertions, 2,152 deletions, and 164 inversions. LUMPY detected 1,231 insertions (and duplications), 3,689 deletions, and 81 inversions in CHM1 and/or CHM13. We merged calls from both WHAM and LUMPY into a nonredundant call set by type collapsing any calls with a 50% reciprocal overlap into a single representative call. The resulting merged set contained 1,609 insertions, 5,448 deletions, and 215 inversions.

We intersected SMRT-SV calls with WHAM/LUMPY calls by type requiring 50% reciprocal overlap for calls to be considered shared between sets. We found 3,031 SVs shared between SMRT and Illumina WGS calls representing 10% of SMRT-SV variants and 62% of WHAM/LUMPY calls (Supplemental Figure S15). As with comparisons between SMRT-SV calls and previously published SVs, we found the greatest proportion of novel variants in the size range of 50 bp to 1 kbp (Supplemental Figure S16). The majority of variants specific to SMRT-SV were insertions with only 279 insertions shared between call sets (2% of SMRT-SV calls and 20% of WHAM/LUMPY calls) and 18,222 additional insertions from SMRT-SV. However, WHAM/LUMPY recovered ~3 times as many inversions as SMRT-SV with 215 compared to 70, respectively. These Illumina-specific inversions were predominantly <1 kbp while inversions specific to SMRT-SV were primarily >1 kbp.



Supplemental Figure S15. Number of SVs (insertions and deletions) shared and distinct to SMRT-SV, WHAM, and LUMPY with mean and median size (bp) shown in parentheses.



Supplemental Figure S16. SVs by size and type colored by the originating call set. Calls from SMRT-SV that intersect WHAM/LUMPY calls with at least 50% reciprocal overlap are shown in green while calls specific to SMRT-SV are in red and calls specific to WHAM/LUMPY are in blue.

Comparison of indels and SNVs with Illumina call sets for CHM1 and CHM13

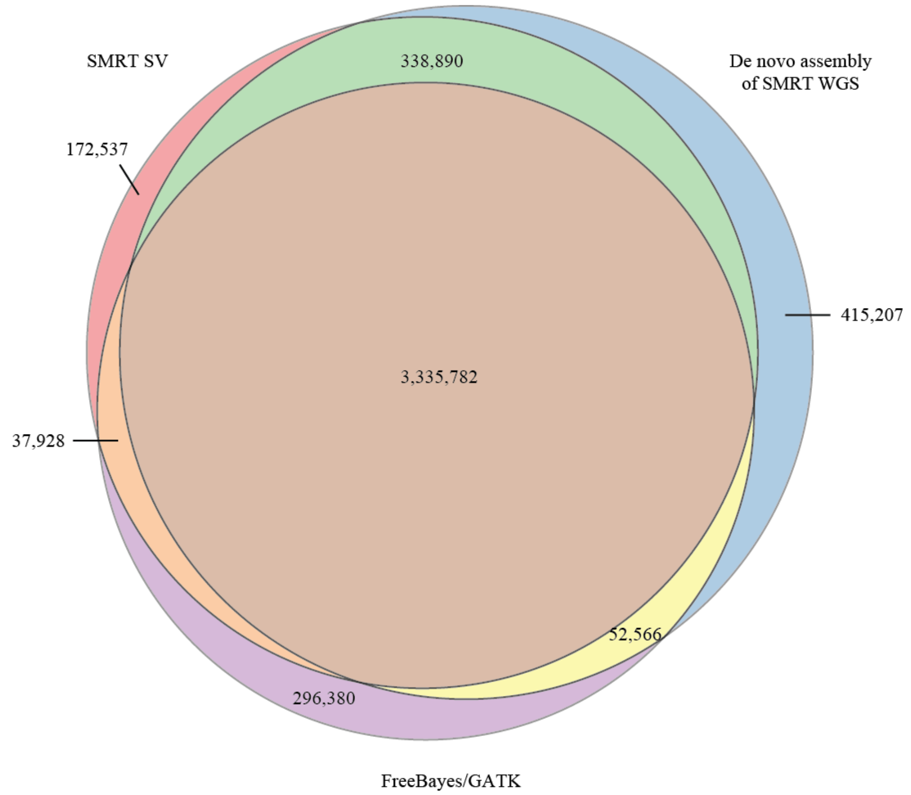
For comparison with SMRT-SV indels and SNVs, we called indels and SNVs from Illumina WGS for CHM1 and CHM13 (SRA accession SRX652547 and BioProject PRJNA335618, respectively) with FreeBayes v0.9.21-19-gc003c1e (Garrison and Marth 2012) and GATK HaplotypeCaller v3.5 (McKenna et al. 2010). For FreeBayes calls, we joint called with both samples using custom parameters (--use-best-n-alleles 4 -q 20 --min-coverage 10) and distributed calling across 500 kbp genomic windows excluding gaps, satellites, subtelomeric and pericentromeric regions, and previously defined low complexity regions (Li 2014). For GATK calls, we generated one Genomic VCF (GVCF) per sample with HaplotypeCaller using customer parameters (--emitRefConfidence GVCF -variant_index_type LINEAR --variant_index_parameter 128000 --min_base_quality_score 20 -rf BadCigar) and produced joint calls across both samples by genotyping both GVCFs together with GATK's GenotypeGVCFs and default parameters.

For indel comparisons, we filtered both FreeBayes and GATK calls to exclude chrY variants, SNPs, calls with quality <21 or read depth <6, single-base indels, and indels >49 bp. We additionally applied standard variant normalization filtering to both call sets, including vcfbreakmulti, vcffixup, vcfstreamsort, vt normalize, and vcfuniqalleles, in that order. The VCF normalize script was run from the vt software suite (built from https://github.com/atks/vt on

12/30/2014) (Tan et al. 2015). All other VCF clean-up scripts were part of the vcflib suite (built from https://github.com/vcflib/vcflib commit hash ea72594). We converted both call sets to BED format and merged calls such that all clusters of calls with a 50% reciprocal overlap had a single representative call reported. This final call set consisted of 73,221 insertions and 65,272 deletions.

We filtered SMRT-SV indels from the theoretical diploid to exclude low complexity events (Li 2014) and compared the resulting calls with the merged FreeBayes/GATK calls using a 50% reciprocal overlap. We found 49,970 shared insertions (53% of SMRT-SV calls and 68% of Illumina-based calls) and 57,486 shared deletions (60% of SMRT-SV calls and 88% of Illumina-based calls). The mean size of insertions specific to SMRT-SV was 10 bp compared to 5 bp for Illumina-only insertions. Similarly, the mean size of deletions specific to SMRT-SV was 9 bp compared to 6 bp for Illumina-only deletions suggesting that SMRT-SV generally recovers increasingly more variants than Illumina callers with increasing variant size (Figure 2a).

For SNV comparisons, we extracted all SNVs from FreeBayes and GATK calls with no additional filtering after calling to produce a total of 3,761,923 SNVs. We intersected these variants with 3,885,137 SMRT-SV SNVs from the theoretical diploid requiring not only a perfect overlap of SNV positions but also identical SNV sequence between call sets. We found 3,413,913 shared SNVs (88% of SMRT-SV and 91% of Illumina-based), 470,268 specific to SMRT-SV (12%), 347,054 specific to Illumina callers (9%), and 956 with a shared position and different sequence content (Supplemental Figure S17).



Supplemental Figure S17. Counts of SNVs observed in CHM1 and CHM13 by SMRT-SV, *de novo* assemblies of SMRT WGS by FALCON, and FreeBayes/GATK calls by Illumina WGS.

Functional annotation of variants

For all SVs (insertions and deletions) and indels, we annotated variants by classes of genomic sequence they intersected as previously described (Gordon et al. 2016). Briefly, we annotated tandem repeats, segmental duplications, genes defined by RefSeq and GENCODE's comprehensive annotations, DNase hypersensitivity sites from both ENCODE cell lines (Harrow et al. 2012) and fetal central nervous system (Turner et al. 2016), and promoter and enhancer sites associated with H3K4Me3 and H3K27Ac signals, respectively. We additionally annotated the presence or absence of variants in previously published SV studies or dbSNP to determine which variants could be considered as "novel".

Using these functional annotations, we identified all variants that could be potentially functional based on their intersection with coding exons, untranslated regions (UTRs), noncoding exons, noncoding regulatory regions, and introns. We found that 42% of variants matched these constraints (Table 2). As expected, <1% of variants intersected protein-coding regions of genes and <2% of variants intersected any coding or noncoding exon or UTR. We then identified potentially novel coding variants by selecting SVs and indels that were not previously published, did not overlap tandem repeats or segmental duplications, and occurred within coding exons of RefSeq or GENCODE comprehensive gene annotations. Based on these constraints, we identified 39 variants, including 8 SVs and 31 indels that affected 16 distinct genes (Supplemental Table S5). Of these 39 variants, 14 (36%) were in tandem repeats within *MUC3A*. In addition to *MUC3A*, only four other genes were affected by SVs including a 128 bp insertion in *AGRN*, a 225 bp insertion in *KDM4B*, a 851 bp deletion in *OR8G1*, and a 318 bp insertion in *SAMD1*. Interestingly, the insertion in *SAMD1* occurred in the middle of the first exon and was observed at 100% allele frequency in the 30 samples we genotyped from the 1KG. In contrast, the insertion in *AGRN* was not detected in any of the 30 diploid genomes or either of the moles most likely as a result of the GC-rich context of the variant. The *KDM4B* insertion occurred with an allele frequency of 59% in all 30 diploid samples and both moles and a frequency of 100% in East Asian and South Asian samples. Finally, the deletion in *OR8G1* perfectly matches a previously described 851 bp deletion that was observed at an allele frequency of 45% in a human diversity panel (Young et al. 2008). In the 30 diploid genomes we assessed, we observed the deletion at 60% with homozygous reference genotypes in only four samples (13%).

Comparison of theoretical diploid and pseudodiploid genomes

Theoretical diploid variants

Variants from CHM1 and CHM13 were clustered by type (insertion, deletion, or inversion) requiring a 50% reciprocal overlap and calls shared in the same cluster from either sample were flagged as shared between samples. After clustering, we found 11,491 deletions with 3,368 (29%) shared between samples and 4,187 (36%) specific to CHM1 and 3,936 (34%) specific to CHM13. Similarly for insertions, we found a total of 18,501 with 6,314 (34%) shared between samples, 6,010 (32%) only in CHM1, and 6,177 (33%) only in CHM13. Of the 70 inversions found between both samples, 23 (33%) were shared, 24 (34%) only in CHM1, and 23 (33%) only in CHM13. Altogether, 9,705 of 30,062 SVs (32%) were shared between both samples. Variants shared between samples were considered "homozygous" for the alternate allele in the theoretical diploid of CHM1 and CHM13 haplotypes while variants specific to one sample were considered "heterozygous."

Pseudodiploid variants

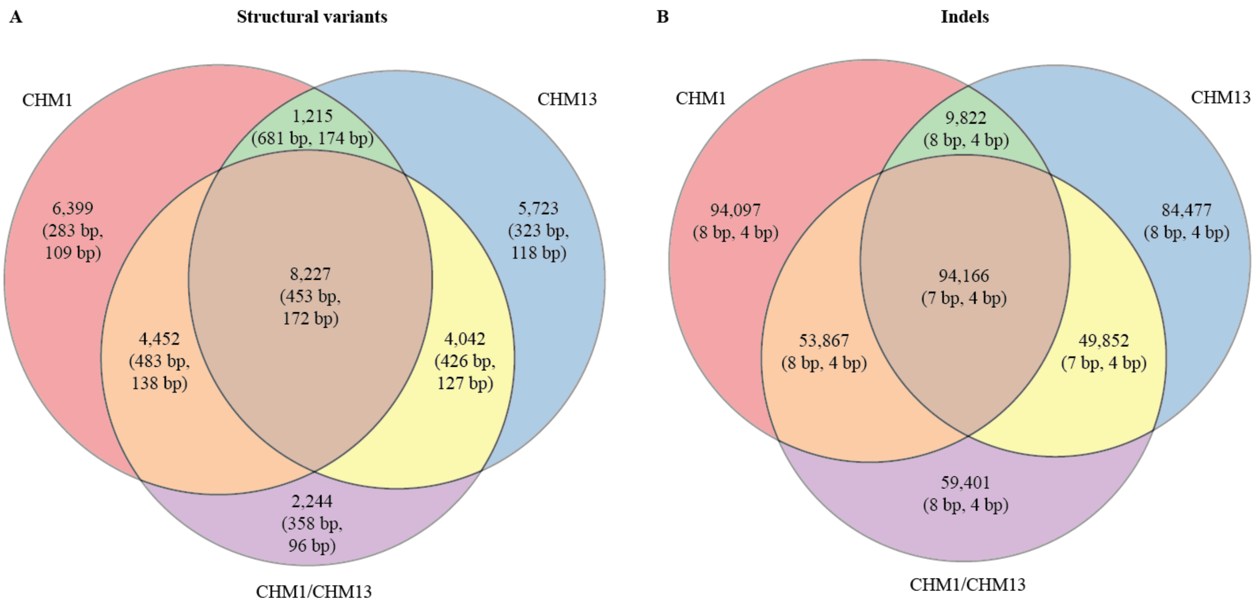
Given the ideal theoretical diploid of CHM1 and CHM13 where all reads and variants are already perfectly phased into separate haplotypes, we created an *in silico* pseudodiploid sample by combining reads from CHM1 and CHM13 and calling variants from these combined reads to determine how SMRT-SV performs with true diploid samples.

Coverages and read-length distributions for CHM1 and CHM13 SMRT WGS varied slightly such that CHM1 reads were longer and coverage was slightly higher. When reads from both samples are combined and assembled *de novo* in local assemblies, these differences in read length and coverage will bias the *de novo* assembler toward the CHM1 haplotype and result in underrepresentation of the CHM13 haplotype in pseudodiploid variant calls. To address this difference in SMRT WGS characteristics, we first matched the read lengths between reads in both samples. For each sample, we selected all reads >=1 kbp and sorted them from longest to shortest. Next, we truncated the end from the longest list of reads to match the length of the shortest list. For each position in both lists, we truncated the longer of the two reads to the length of the shorter read. Finally, we emitted the length-matched reads from each sample to its own respective set of HDF5 files.

To assess how SMRT-SV would perform with a diploid sample of the same coverage as one of the mole samples (~60-fold), we randomly ordered paths to length-matched reads into a single file and batched reads into groups of 56 HDF5 files. We aligned all 32 batches of reads, calculated coverage per batch, and selected 23 batches such that the mean coverage of the combined batches was ~60-fold. We then called SVs and indels from these 23 batches of alignments using the standard SMRT-SV pipeline.

We compared SVs and indels from the theoretical diploid and pseudodiploid by type requiring a 50% reciprocal overlap for variants to be considered shared by both call sets. Based on theoretical diploid SVs, we expected 9,442 "homozygous" SVs, 10,851 heterozygous SVs from CHM1, and 9,765 heterozygous SVs from CHM13 (Table 1 and Supplemental Figure S18). In the pseudodiploid, we observed 8,227 of the homozygous SVs (87%), 4,452 of the CHM1 heterozygous SVs (41%), and 4,042 of the CHM13 heterozygous SVs (41%). These results correspond to a 59% FNR for heterozygous SVs and a 13% FNR for homozygous SVs with an overall FNR of 44% for all SVs. Additionally, we observed 2,244 of the 18,965 total SVs in the pseudodiploid that were not observed in either CHM1 or CHM13 and that corresponded to false positives and a false discovery rate (FDR) of 12%.

We observed the same pattern for indels recovered by the pseudodiploid from the theoretical diploid. We found 94,166 of 103,988 (91%) expected homozygous indels. However, we only recovered 53,867 of 147,964 (36%) heterozygous indels from CHM1 and 49,852 of 134,329 (37%) heterozygous indels from CHM13. Altogether, we observed a 9% FNR for homozygous indels and a 63% FNR for heterozygous indels with an overall FNR of 49% for indels. In contrast with missing heterozygous SVs, many missing heterozygous indels were likely filtered by selecting a single tiling path through local assemblies across the reference. This filtering by tiling path effectively reduces the representation of indels to a single haplotype for any locus in the genome. As with SVs, we discovered 59,401 of 257,286 pseudodiploid indels that were not observed in CHM1 or CHM13 and correspond to a FDR of 23%.



Supplemental Figure S18. False negative rate (FNR) for **a)** SV and **b)** indel discovery in diploid samples based on variant calls made in a 30-fold CHM1 genome, a 30-fold CHM13, and a pseudodiploid of the two 30-fold genomes combined. SMRT-SV recovers 87% of expected homozygous SVs in the pseudodiploid but only recovers 41% of heterozygous variants. Similarly SMRT-SV identifies 91% of homozygous indels but only 37% of heterozygous indels.

Variant genotyping

Creation of a custom reference from local assemblies

The presence of assembled alternate haplotypes containing SVs makes it possible to accurately genotype variants by alignment of Illumina read pairs. We start by creating a new genotyping reference assembly per sample (e.g., CHM1 or CHM13) by concatenating the original SV reference assembly (e.g., GRCh38) with all SV-containing local assemblies as alternate haplotypes. In addition to adding local assemblies to the reference as alternate haplotypes, we create a SAM file of the local assembly alignments to the SV reference assembly for use by BWA-MEM’s alt-aware alignment and rescoring algorithm. The genotyping reference is indexed using BWA’s index command.

Pulldown of SV-associated and aligned Illumina reads

The alignment of ~30-fold human Illumina WGS to the human reference (GRCh38) requires ~144 hours of CPU time (12 hours with BWA-MEM and 12 threads). However, SVs detected by SMRT-SV represent ~16 Mbp or 0.5% of the genome, so alignment of all WGS reads to the genotyping reference requires unnecessary alignment of reads that are unrelated to SVs. To minimize the amount of time spent aligning reads, we performed an *in silico* "pulldown" of all SV-associated reads from each sample’s WGS based on the precondition that the sample’s WGS had already been aligned to a reference assembly (from here on "WGS reference").

If SVs detected in SMRT WGS are present in a sample sequenced with Illumina WGS, the Illumina reads associated with those SVs are not guaranteed to have aligned adjacent to the SV breakpoints in the WGS reference. Instead, the global optimal alignment for those reads may have been elsewhere in the reference or the reads may have been completely unmapped. For example, reads from the middle of an L1 insertion in the sample's WGS are most likely to align to another existing L1 sequence in the WGS reference instead of the locus where the actual L1 insertion occurred. Instead of only pulling down reads mapping adjacent to SV breakpoints in the WGS reference, we took advantage of the fact that our input WGS data had already been aligned to a reference that had already been indexed by a Burrows-Wheeler Transform (BWT) by mapping SV-associated sequences from SMRT-SV to the WGS reference index and pulling down all Illumina reads that corresponded to the mapping locations of SV-associated sequences.

Specifically, to pulldown SV-associated reads, we first extracted all local assembly sequences associated with SVs. For deletions, we extracted all sequences from the SV reference corresponding to the nonredundant set of deletion breakpoints with 5 kbp added to either side of each variant. For insertions, we extracted the inserted sequence itself from the local assemblies as well as 5 kbp of sequence from either side of the single-base insertion point in the SV reference. These combined sequences represented SV-associated sequences. Next, we fragmented the SV-associated sequences into read-length chunks of 500 bp in 250 bp sliding windows across each input sequence to mimic the alignment of Illumina-length reads. We aligned these SV-associated read-length sequences in single-end mode with BWA-MEM (default settings) to all references for which WGS samples are available. We then converted these alignments from BAM to BED format, added 5 kbp to either side of each alignment, and merged all overlapping regions into a nonredundant set of SV-associated regions in the WGS reference. These SV-associated regions represented the locations of SV-associated reads to pulldown for SV genotyping.

Alignment of Illumina reads to custom reference

For each sample chosen for genotyping, we selected all mapped reads from the SV-associated regions and all unmapped reads from the sample's BAM and aligned those reads in paired-end mode to the custom reference (defined above) with BWA-MEM v0.7.12-r1039 with support for alternate haplotype alignments enabled. This "alt-aware" mode of BWA uses a set of alternate haplotype sequences and their corresponding positions in the primary reference assembly to rescore the mapping quality of alignments to alternate haplotypes. In practice, this means reads that map equally well to the primary reference and multiple alternate haplotypes (due to homology between haplotypes) will be assigned the same mapping quality instead of being assigned a mapping quality of zero as with the default implementation of BWA-MEM. We passed local assembly alignments from CHM1 or CHM13 as alternate haplotypes for BWA-MEM and its post-alignment script. This approach not only allowed homology between the primary reference and local assemblies but also homology between pairs of local assemblies. Because local assemblies are redundant by the design as tiled windows across the genome, alt-aware alignments prevented alignments of the same Illumina reads to overlapping tiled assemblies from being assigned a mapping quality of zero.

Read-depth calculation per SV

After alignment of SV-associated reads from each sample to the custom reference, we calculated read-depth support for each SV across the breakpoints in both the reference and alternate (i.e., SV-containing) haplotypes. For deletions, we assessed 25 bp windows on either side of the two reference breakpoints and the single alternate breakpoint. For insertions, we applied the same logic as for deletions but with two alternate breakpoints and a single reference breakpoint.

We selected all Illumina reads aligned across these reference and alternate haplotype breakpoints with a mapping quality >=20. During BWA alignment and post-alignment scoring for alternate haplotypes, reads were allowed to map to both the primary reference and all other alternate haplotype sequences and all alignments for the same read were assigned the best mapping quality score. In practice, this means a read may map with higher identity to one haplotype than another but the best alignment for the read cannot be distinguished by mapping quality alone. To avoid double counting reads with multiple alignments, we sorted all alignments to reference and alternate breakpoints per SV by read name and number (e.g., "read/1" for the first read in a pair named "read"), selected the best alignment per read as the alignment with the fewest total mismatch, indel, and soft-clipped bases, and resorted these filtered alignments in coordinate order. If a read had equally scored alignments to both reference and alternate haplotypes, it could not be used to distinguish support for an SV in either haplotype and we omitted all alignments for the read as if it had a mapping quality of zero. Using these filtered alignments, we calculated read depth per base (where base quality >=20) across the breakpoint regions for a given SV. Finally, we summarized the read depth supporting each breakpoint by the median depth per base minus the standard error with a minimum support of zero.

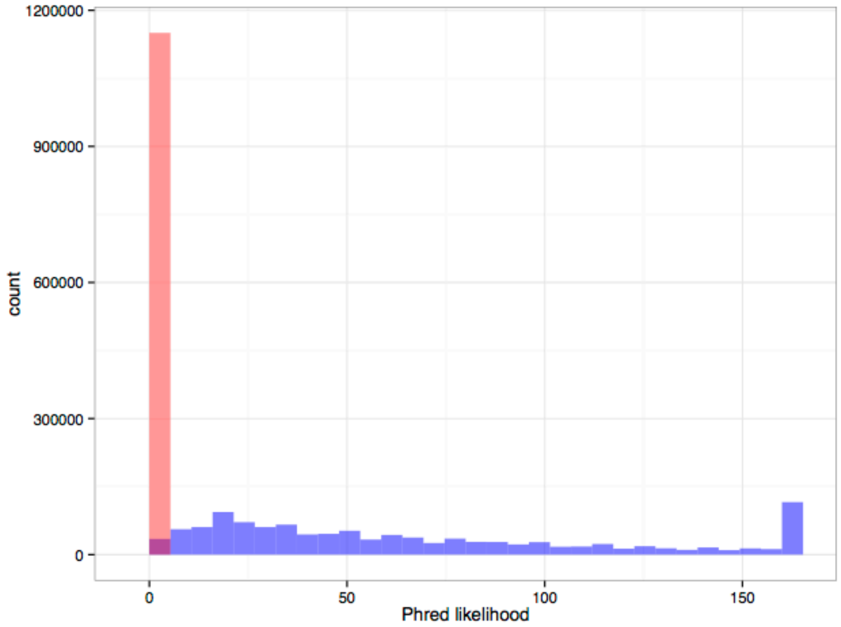
Assignment of genotypes by read depth

To assign genotypes per sample and SV, we first calculated the probability of three biallelic genotypes (0/0, 1/0, and 1/1) given the read depth across the SV breakpoints in reference and alternate haplotypes. Probabilities were calculated from the binomial probability mass function where the total depth across both breakpoints represented the total number of trials (*n*) and the depth across a specific haplotype's breakpoints represented the number of successes (*k*) for that haplotype. For homozygous reference and alternate genotypes, the probability (*p*) was set to 95% to account for 5% error in alignments to the incorrect haplotype and *k* was the depth supporting either the reference or the alternate haplotypes, respectively. For the heterozygous genotype, *p* was set to 50% based on an ideal expectation where 50% of reads map to either haplotype. We hardcoded probabilities of zero for heterozygous genotypes on chrX SVs in males and chrY SVs in any sample. Similarly, we hardcoded probabilities of zero for homozygous reference and alternate genotypes on chrY SVs in females. The genotype assigned to a sample for an SV was then determined as the genotype with the maximum probability of the three possible genotypes. In the case where read depth was less than 5 for both haplotypes, the genotype was set to "uncalled" (./.) as there was not enough information to reliably assign a genotype to either haplotype.

We calculated a Phred-scale genotype likelihood for each genotype, sample, and SV combination based on the standard Phred transformation of the ratio of a given genotype's individual probability divided by the sum of all genotype probabilities. Genotypes for which likelihoods could not be calculated due to division by zero or other edge cases were assigned a likelihood of zero. Infinite likelihood values were replaced by the maximum non-infinite likelihood assigned to any sample/SV pair in the entire set of samples and SVs.

We determined the confidence of genotype likelihoods by calculating the range, mean, and median value for the maximum and second highest likelihood of each CHM1 and CHM13 SV genotype from the 30 PCR-free samples and both hydatidiform moles (n=1,150,388). Maximum likelihoods ranged from 0-160 with a mean of 66.4 and a median of 52 (Supplemental Figure

S19). In contrast, the second highest likelihoods ranged from 0-3 with a mean of 0.1 and a median of 0. From these results we conclude that the assigned genotype for each SV is generally several orders of magnitude more likely than the other possible genotypes. However, the most likely genotype can still be a low confidence genotype due to the lack of enough supporting reads mapping across SV breakpoints. For example, 19% of genotypes (213,918) had a likelihood less than 20 corresponding to >1% probability of a genotyping error.



Supplemental Figure S19. Distribution of maximum genotype likelihoods (blue) and second highest genotype likelihoods (red) for all CHM1 and CHM13 SVs and corresponding genotypes from 30 PCR-free diploid genomes and the two hydatidiform moles.

Concordance of genotypes in the moles

To test the accuracy of the genotyper, we independently genotyped CHM1 and CHM13 SVs with both CHM1 and CHM13 Illumina WGS (SRA accessions ERX1413366 and ERX1413367, respectively). Because both samples are effectively haploid, we expected all CHM1 genotypes for CHM1 SVs ("self-self" genotypes) to be homozygous alternate and all CHM13 genotypes for CHM1 SVs ("self-other" genotypes) to be homozygous for either the reference or alternate allele but never heterozygous. Heterozygous genotypes from either sample indicated false positives as did homozygous reference genotypes for self-self assessment.

Of 40,979 combined CHM1 and CHM13 insertions and deletions, we found that 31,371 (77%) had sufficient coverage across the breakpoints to be reliably genotyped by their respective sample’s short reads (Supplemental Table S6). The majority of these self-self genotypes had the expected homozygous alternate genotype (29,570 or 94%) while 630 (2%) were classified as heterozygous and 1,171 (4%) as homozygous reference. Similarly, 28,829 (91%) of the 31,633 self-other genotypes were homozygous for one of the alleles. Of the 9,608 (23%) ungenotyped SVs from self-self genotypes, 9,255 (96%) occurred within tandem repeats or segmental duplications. These results suggest an initial FDR of 6-9% for those regions that can be assayed in short-read sequence data. When we more strictly defined the number of genotyped SVs to those with homozygous alternate genotypes for one mole's SVs based on that mole's Illumina WGS and homozygous reference or alternate genotypes from the other mole's Illumina WGS, we found that 61% (18,211 / 29,992) of SVs from the theoretical diploid met these criteria.

Concordance of genotypes with PCR and Sanger sequencing

We performed validation of 56 SVs (24 from CHM1 and 32 from CHM13) between 50-500 bp that did not occur inside tandem repeats or segmental duplications and did not overlap with previously published SVs. Primers were designed in Primer3 (Untergasser et al. 2007) using a 500 bp window around the SV reference coordinates and we checked the primer pairs with the In-Silico PCR feature on the UCSC Genome Browser to ensure only one PCR product would amplify. We amplified these regions in five 1KG individuals (HG01051, NA19240, NA18525, HG01595, and NA12878), checked for amplification on agarose gels, cleaned the product, and sequenced. The resulting traces were aligned to the reference sequence in Sequencher v5.4.5 to determine whether they were homozygous for either presence or absence of the SV or heterozygous. We were able to unambiguously determine 264 genotypes from the 56 SVs and 5 samples of which 237 (90%) were concordant with genotypes assigned by SMRT-SV Genotyper (Table 4 and Supplemental Table S7). Concordance was highest for homozygous genotypes with 101 concordant homozygous reference genotypes out of 102 (99%) and 89 out of 97 homozygous alternate genotypes (92%). Heterozygous genotypes were less concordant with only 47 of 65 (72%) matching SMRT-SV Genotyper. The most common error was the assignment of heterozygous genotypes by SMRT-SV Genotyper when the true genotype was homozygous reference (14 out of 18 discordant genotypes or 78%). Filtering non-MEI genotypes by genotype quality >=30 improved overall genotype accuracy by 2%.

Concordance of genotypes with 1000 Genomes Project MEIs

In addition to testing concordance between SMRT-SV Genotyper and PCR genotypes for non-MEI SVs, we compared SMRT-SV Genotyper genotypes for MEIs with PCR genotypes from Stewart et al. (2011). Specifically, we lifted MEI coordinates provided by Stewart et al. (2011) from GRCh36/hg18 to GRCh38/hg38 and identified 184 shared MEIs from CHM1 and CHM13. Next, we selected genotypes from the four unrelated samples shared between both studies (NA12891, NA12892, NA19238, and NA19239) and compared those genotypes with SMRT-SV Genotyper results for the same samples. Of the 331 genotypes shared between studies, 262 (79%) were concordant with 99% accuracy for SMRT-SV Genotyper heterozygous calls and 100% accuracy for homozygous alternate calls (Supplemental Table S8). The greatest source of error for MEI genotypes from SMRT-SV Genotyper was the undercalling of true heterozygous insertions as homozygous reference genotypes, which resulted in 63% accuracy for homozygous reference calls. Filtering SMRT-SV Genotyper genotypes by genotype quality >=30 dramatically improved the accuracy of homozygous reference calls to 80% for an overall MEI genotyping accuracy of 87%.

Genotypes for complex variants

Although complex variants comprised only 3% of all SVs from the theoretical diploid (n=846), these were often the largest SVs with mean lengths of 2.6 kbp for deletions and 2.2 kbp for insertions (Supplemental Table S2). We found that 823 (97%) of these variants could be genotyped in at least one of the 30 diploid samples from the 1KG diversity panel while 785 (93%) variants had an allele frequency >1% across all 30 diploid samples. Although the sequence composition of these variants is complex relative to other repeat classes, these genotyping results indicate that the breakpoints of these variants can be accurately resolved with long-read assemblies and genotyped with short reads.

Effect of the pulldown method on genotype accuracy

We tested the effect of using the pulldown method instead of remapping all reads from a given sample by modifying the SMRT-SV Genotyper to use all Illumina reads from a given sample instead SV-associated reads and running the genotyper for CHM1, CHM13, and NA19240. This latter diploid sample had the largest input BAM at 242 GB and took a little over three days to remap all reads to GRCh38 with 12 CPUs on a dedicated server. In contrast, we originally remapped NA19240's SV-associated reads in six hours with 12 CPUs.

To understand the impact of pulling down reads instead of mapping all reads, we compared the resulting number of genotypes calculated for each sample in the four categories of ungenotyped ("./."), homozygous reference ("0/0"), heterozygous ("1/0"), and homozygous alternate ("1/1"). For CHM1 Illumina reads mapped to CHM1 SVs, we observed a 1% increase in the number of homozygous alternate genotypes when all reads were mapped (Supplemental Table S9). Correspondingly, we saw a 1% increase in homozygous alternate genotypes for CHM13 mapped to CHM1 SVs. Finally, we observed a 5% increase in heterozygous and homozygous alternate genotypes for NA19240 when all reads were mapped and a corresponding decrease in ungenotyped or homozygous reference genotypes. These results suggest that the pulldown method introduces a 1-5% genotyping error with errors predictably biased toward undercalling the alternate allele due to lack of coverage across breakpoints.

We accept the error introduced by the pulldown method for the benefit of the enormous computational time savings for processing high-coverage genomes. We recommend the use of stricter minimum read-depth requirements for genotypes for users who wish to force false positive genotypes into the "ungenotyped" category and minimize the class of error described above.

References

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**: 28-36.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573-580.

Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* **33**: 623-630.

Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M et al. 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**: 608-611.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC bioinformatics* **13**: 238.

Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A et al. 2016. Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing. *bioRxiv* doi:10.1101/056887.

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods* **10**: 563-569.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* **464**: 704-712.

ENCODE. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv preprint* doi:arXiv:1207.3907v2.

Gordon D, Huddleston J, Chaisson MJ, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW et al. 2016. Long-read sequence assembly of the gorilla genome. *Science* **352**: aae0344.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760-1774.

Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, Sudmant PH, Graves TA, Alkan C, Dennis MY et al. 2014. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res* **24**: 688-696.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656-664.

Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, Eichler EE. 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**: 837-847.

Kronenberg ZN, Osborne EJ, Cone KR, Kennedy BJ, Domyan ET, Shapiro MD, Elde NC, Yandell M. 2015. Wham: Identifying Structural Variants of Biological Consequence. *PLoS Comput Biol* **11**: e1004572.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**: R84.

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**: 2843-2851.

MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* **42**: D986-992.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**: 59-65.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**: 308-311.

Smit AFA, Hubley R, Green P. 1996-2004. RepeatMasker Open-3.0.

Stewart C, Kural D, Stromberg MP, Walker JA, Konkel MK, Stutz AM, Urban AE, Grubert F, Lam HY, Lee WP et al. 2011. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS genetics* **7**: e1002236.

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M et al. 2015a. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**: aab3761.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M et al. 2015b. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75-81.

Tan A, Abecasis GR, Kang HM. 2015. Unified representation of genetic variants. *Bioinformatics* **31**: 2202-2204.

Turner TN, Hormozdiari F, Duyzend MH, McClymont SA, Hook PW, Iossifov I, Raja A, Baker C, Hoekzema K, Stessman HA et al. 2016. Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *Am J Hum Genet* **98**: 58-74.

Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA. 2007. Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res* **35**: W71-74.

Young JM, Endicott RM, Parghi SS, Walker M, Kidd JM, Trask BJ. 2008. Extensive copy-number variation of the human olfactory receptor gene family. *Am J Hum Genet* **83**: 228-242.