Table of Contents

[1. Sample collection and data generation 3](#_Toc490839116)

[1.1 Sample Collection 3](#_Toc490839117)

[1.2 Data generation and preprocessing 3](#_Toc490839118)

[PCR-free deep Illumina-sequencing 3](#_Toc490839119)

[3.5 Kb Long-insert whole genome sequencing 3](#_Toc490839120)

[7.5 Kb Mate-pair sequencing 4](#_Toc490839121)

[PacBio SMRT Sequencing 5](#_Toc490839122)

[Bionano Optical mapping 5](#_Toc490839123)

[Strand-seq Methods 6](#_Toc490839124)

[10X Genomics 7](#_Toc490839125)

[TruSeq Synthetic Long Reads (TSLR) 8](#_Toc490839126)

[Hi-C data generation 9](#_Toc490839127)

[Transcriptome Sequencing 10](#_Toc490839128)

[1.3 Data and tools availability 11](#_Toc490839129)

[2. Chromosomal haplotype resolution and integration 11](#_Toc490839130)

[2.1 Phasing of independent datasets 12](#_Toc490839131)

[2.2 Phasing using multiple data sources 13](#_Toc490839132)

[2.3 Mapping of meiotic recombination events 13](#_Toc490839133)

[2.4 Strand-seq phasing 13](#_Toc490839134)

[2.5 Mapping of meiotic recombination events using Strand-seq 14](#_Toc490839135)

[2.6 Refinement of mapped meiotic breakpoints using PacBio reads 14](#_Toc490839136)

[3. Haplotype specific *de novo* assembly 15](#_Toc490839137)

[3.1 Phased-SV 15](#_Toc490839138)

[3.2 MsPAC 15](#_Toc490839139)

[3.3 Assembly coverage 16](#_Toc490839140)

[4. Full-spectrum variation detection 17](#_Toc490839141)

[4.1 Short insertion and deletion (<50bp) 17](#_Toc490839142)

[Merging of Illumina callsets 17](#_Toc490839143)

[PacBio indels 17](#_Toc490839144)

[4.2 Haplotype-resolved SV characterization and integration 18](#_Toc490839145)

[PacBio based SV detection. 18](#_Toc490839146)

[Unified PacBio SV callset 20](#_Toc490839147)

[Intersection with CHM1 and CHM13 SV Callsets 21](#_Toc490839148)

[Intersection with 1000 Genomes Phase 3 Callsets 21](#_Toc490839149)

[Proximity of Integrated PacBio calls to 1000 Genomes SV 21](#_Toc490839150)

[Orthogonal support of PacBio variants 22](#_Toc490839151)

[Orthogonal sequencing approaches 23](#_Toc490839152)

[4.3 Variation detected by short-read sequencing 24](#_Toc490839153)

[SV Detection Methodologies on Illumina Genome Sequencing 24](#_Toc490839154)

[Development of an integrated Illumina callset 30](#_Toc490839155)

[HySA Hybrid SV callset 32](#_Toc490839156)

[10x Chromium SV discovery 33](#_Toc490839157)

[5. Integration of Illumina, PacBio, and Bionano callsets 34](#_Toc490839158)

[5.1 Comparison of integrated Illumina and integrated PacBio callsets 34](#_Toc490839159)

[Genomic coverage of Illumina and PacBio 34](#_Toc490839160)

[Filtering integrated Illumina callset 34](#_Toc490839161)

[Filtering Bionano Genomics calls 35](#_Toc490839162)

[5.2 Integration of Illumina, PacBio, and Bionano callsets 36](#_Toc490839163)

[Specificity of integrated Illumina dataset 36](#_Toc490839164)

[Analysis of combinations of Illumina callers. 37](#_Toc490839165)

[Comparison of read depth and *de novo* assembly methods 37](#_Toc490839166)

[5.3 Properties of the unified integrated calls 37](#_Toc490839167)

[Genes affecting SVs 38](#_Toc490839168)

[SV overlapping noncoding functional elements 38](#_Toc490839169)

[6. Multiscale discovery of fixed and polymorphic inversions 38](#_Toc490839170)

[6.1 Large-scale inversion detection using Strand-seq 38](#_Toc490839171)

[6.2 Classifying Strand-seq inversions using orthogonal phase data: 40](#_Toc490839172)

[6.3 Integrating inversion calls across orthogonal platforms: 41](#_Toc490839173)

[6.4 Resolving Inversion Sequence Errors in the Human Reference Genome 42](#_Toc490839174)

[7. Biological context 43](#_Toc490839175)

[7.1 Population genetics of SVs 43](#_Toc490839176)

[Population genetics of PacBio Integrated callset 43](#_Toc490839177)

[SMRT-Genotyper summary 44](#_Toc490839178)

[7.2 Sequence context analysis 44](#_Toc490839179)

[Functional enrichment analysis 45](#_Toc490839180)

[Structural variants engulfing genes 45](#_Toc490839181)

[INDEL functional annotations 46](#_Toc490839182)

[Variants overlapping GWAS sites. 48](#_Toc490839183)

[Haplotype specific reference analysis 48](#_Toc490839184)

[7.3 MEI discovery and analysis with PacBio data 48](#_Toc490839185)

[8. References 49](#_Toc490839186)

# 1. Sample collection and data generation

## 1.1 Sample Collection

**Cell lines:** Transformed lymphoblast cell lines from three parent-child trios (**Figure S1.1.1**) belonging to the 1000 Genomes Project were obtained from the Coriell Cell Repository as part of the NHGRI catalog (<https://catalog.coriell.org/1/NHGRI> ).

## 1.2 Data generation and preprocessing

### PCR-free deep Illumina-sequencing

**Library preparation and sequencing (Contributors: Sau Peng Lee, Ching Lek Koh, ,Korlach, Munson, Eichler, Lee, JE and Lee, C):** DNA was extracted and it’s OD260/280 ratio confirmed to be between 1.8–2.0. The quality of the DNA was further evaluated by using a PicoGreen® dsDNA Assay(Invitrogen). DNA libraries were prepared according to the illumina Truseq DNA PCR-Free Library prep protocol. For each DNA library preparation, 2 ug of high molecular weight genomic DNA was randomly sheared using the Covaris S2 system to 550 bp fragments. The fragments were blunt ended, phosphorylated, and a single 'A' nucleotide was added to the 3' ends of the fragments in preparation for ligation to an adapter that has a single-base 'T' overhang. Adapter ligation at both ends of the genomic DNA fragment conferred different sequences at the 5' and 3' ends of each strand in the genomic fragment. The quality of the DNA libraries was verified by capillary electrophoresis (Bioanalyzer, Agilent), clustered on the Illumina cBOT station and paired-end sequenced for 125 cycles on the HiSeq 2500 sequencer according to the Illumina cluster and sequencing protocols.

**Sequence data processing (Contributors: Fairley, Lowy, Zheng and Clarke):** In addition to data coordination and distribution, IGSR [(Clarke et al. 2017)](https://paperpile.com/c/ml0krT/dqhH) provided alignment of Illumina whole-genome sequence data (PCR-free high coverage, low coverage and exome) from the nine individual genomes. Data was aligned to the GRCh38 assembly in an alt-aware manner using bwa-mem [(Heng Li 2013)](https://paperpile.com/c/ml0krT/T1Zh). (see data accessions, ftp links and parameters at end of the document). The locations of the output alignments are given in the alignment index files at end of document.

Further statistics from the alignments can be found in the .bas files that sit alongside the alignment files.

### 3.5 Kb Long-insert whole genome sequencing

**Library preparation and sequencing (Contributors: Talkowski, Collins, Brand, Stone, Glessner):** We generated long-insert whole-genome sequencing (liWGS) libraries for all nine individuals from the three HGSVC trios with a protocol that has been previously described [(Hanscom and Talkowski 2014)](https://paperpile.com/c/ml0krT/E2DX). In brief, 5.0 µg of genomic DNA from lymphoblastoid cell lines for each individual was sheared with a Covaris E220 sonicator and size selected to a target fragment size of 3,500bp (targeted range: 2,500-5,000 bp). These ~3.5kb fragments were circularized around a biotinylated adapter oligo and digested with EcoP15I restriction enzyme, followed by streptavidin bead-based capture of the biotinylated circularization junction and preparation of fragments for Illumina TruSeq sequencing with paired-end 25bp reads per Illumina’s standard protocols. Sequencing was performed on an Illumina HiSeq2500 at The Broad Institute to a mean depth of 191.9 million read-pairs per library. **Sequence data processing (Contributors: Talkowski, Collins, Brand, Stone, Glessner):** Quality of raw sequencing reads was evaluated using FastQC [(Andrews and Others 2010)](https://paperpile.com/c/ml0krT/sl57) prior to alignment, then libraries were aligned against the GRCh37 and hg38 primary assemblies with BWA-backtrack v0.7.10-r789 [(H. Li and Durbin 2009)](https://paperpile.com/c/ml0krT/ozJS). Duplicates were marked with SAMBLASTER v0.1.1 [(Faust and Hall 2014)](https://paperpile.com/c/ml0krT/9edP) and all subsequent alignment processing, including sorting and indexing, was performed with sambamba v0.4.6 [(Tarasov et al. 2015)](https://paperpile.com/c/ml0krT/B1G3). Alignment quality was assessed with the Picard suite v1.115 (https://broadinstitute.github.io/picard/), Samtools v1.0 [(Heng Li et al. 2009)](https://paperpile.com/c/ml0krT/eBY8), and BamTools v2.2.2 [(Barnett et al. 2011)](https://paperpile.com/c/ml0krT/cxaJ). Library production generated an average insert of 3,475bp, and a mean physical coverage of 158.8X per library. Insert size distributions and alignment statistics are provided in **Figure S1.2.1 and Table S1.2.2**.

### 7.5 Kb Mate-pair sequencing

**Library preparation and sequencing (Contributors: Stuetz):** Long-range (or ‘Mate-pair’) DNA library preparation was carried out using the Nextera Mate Pair Sample Preparation Kit(Illumina). In brief, 5μg of high molecular weight genomic DNA were fragmented by the Tagmentation reaction in 400ul, followed by the strand displacement and AMPure XP (Agencourt) cleanup reaction. Samples were size selected to 6.5-8.5kb with a gel step following the Gel-Plus path of the protocol. 350-500ng of size-selected DNA were circularized in 400ul for 16h at 30° C. The library was then constructed after an exonuclease digestion step to get rid of remaining linear DNA, fragmentation to 300-700bp with a Covaris S2 instrument (LGC Genomics), binding to streptavidin beads and Illumina Truseq adapter ligation. Final library was obtained after PCR for 1min @ 98°C, followed by 11 cycles of 30sec @ 98°C, 30sec @ 60°C, 1min @ 72°C and a final 5min @ 72°C step and another gel size selection step. Deep sequencing was carried out with the Illumina HiSeq2000 (2x101bp) instrument using v3 chemistry to reach an average physical coverage of 30x.

**Data preprocessing of the 7.5Kb library (Contributors: Meiers, Rausch):** The Illumina Nextera Mate Pair protocol was used to generate a 7.5 Kbp insert library for all 9 samples. Nextera Mate Pair data cannot be trimmed with standard adapter trimming tools because of the circularization-based library preparation but specialized tools such as nxTrim and NextClip exist. In this project we used NextClip with default parameters. All remaining read pairs were aligned using the EBI’s GRCh38 alignment pipeline detailed in this README (<ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/README.illumina_wgs.GRCh38.alignment)>. Briefly, reads were aligned using bwa mem, de-duplicated using BioBamBam and converted to CRAM format using Cramtools. The alignment index is available at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data\_collections/hgsv\_sv\_discovery/.

### 

### PacBio SMRT Sequencing

**Library preparation and sequencing (Contributors: Korlach, Munson):** For the Puerto Rican trio (HG00731, HG00732, HG00733), high molecular weight DNA was prepared from cultured cells using the Gentra PureGene kit and a modified protocol (http://www.pacb.com/documentation/unsupported-protocol-gentra-puregene-qiagen-dna-isolation/). DNA integrity was confirmed by visual inspection of 1% agarose gel and the appearance of a single HMW band.SMRTbell libraries were constructed using the SMRTbell Template Prep Kit 1.0, according to the protocol described in: “Procedure & Checklist – 20 kb Template Preparation Using BluePippin Size-Selection System” (Pacific Biosciences, Menlo Park CA). The genomic DNA was mechanically sheared using the Megaruptor system (Diagenode, Denville NJ) to yield an average shear size distribution of 30-35kb for the Han Chinese & Puerto Rican trio samples, and gTubes (Covaris, Woburn MA) for the Yoruban trio samples (20-25kb average shear size distribution). The libraries was then subjected to a size-selection step using the BluePippin system (SageScience, Beverly MA) to remove shorter DNA inserts, with size cutoffs of 17 kb (Han Chinese & Puerto Rican) and 10 kb (Yoruban), respectively. Library quality and quantity were assessed using the Pippin Pulse field inversion gel electrophoresis system (SageScience), as well as the Qubit dsDNA Broad Range Assay kit and Qubit Fluorometer (Thermo Fisher). An additional library was constructed for HG00733 at the University of Washington: Isolated DNA was diluted to 50 ng/uL and sheared using Megaruptor (Diagenode) at 35 kb (PacBio) or 40 kb (UW) settings and SMRTbell libraries were generated as described above with a BluePippin size selection of 15 Kbp.

All SMRT sequencing was performed on the Pacific Biosciences RS II using on-plate concentrations of 100-150 pM (PacBio libraries) or 250 pM (UW library). Data were generated using the P6-C4 sequencing chemistry, with magnetic bead loading and 240 or 360 minutes’ movie times. Sequencing was performed at 4 centers: Jackson Laboratory (YRI trio), University of Washington and Ontario Institute for Cancer Research (PUR trio) and University of Malaya (CHS trio). Children were targeted with a sequencing depth of ~40-fold coverage andparents for ~20-fold sequence coverage with the goal of providing 30-fold sequence coverage per haplotype (**Table S1.2.3**, **Fig. S1.2.2**). Mapped sequence read lengths varied between samples: HAN: 9,583 bp PUR: 9475bp, and YRI: 5,528 bp.

### Bionano Optical mapping

**Bionano DNA labeling (Contributors: Hastie，Lee):** High-molecular-weight DNA from fresh cell lines was labelled following the IrysPrep Reagent Kit protocol. For 2 hours at 37 °C, DNA was digested with nicking endonuclease Nb.BssSI (New England BioLabs). Nicked DNA was then incubated for 1 hour at 72 °C with fluorescently labelled dUTP and Taq Polymerase (New England BioLabs). Taq ligase (New England BioLabs) was used in the presence of dNTPs for ligation of nicks. DNA was counterstained with YOYO-1 (Thermo Fisher Scientific). All samples were also labeled with Nt.BspQI using optiDNA labeling kit. Cells are mixed with liquefied LMP agarose and deposited on a device for forming a thin layer of agarose/cell mixture. After solidification, the cells are lysed and proteins digested by protease K and lysis buffer. Following washes, DNA is nicked with Nt.BspQI at 37°C for 2 hours, labeled at 55°C for one hour using fluorescent nucleotide analogs and taq polymerase, DNA is then ligated with taq ligase, the thin layer is washed between steps. The agarose is dissolved and the DNA is stained before data collection.

**Bionano data collection:** DNA samples nicked with each nick endonucleases were loaded into IrysChips (Bionano Genomics) and run on the Irys (Bionano Genomics) system. Data were collected until approximately 100-fold coverage of long molecules (> 150 Kbp) was achieved for both Nt.BspQI and Nb.BssSI samples.

**Bionano data preprocessing:** The IrysView (Bionano Genomics) software was used to detect linearized DNA using the YOYO-1 counterstain, and to detect the labelled nick sites on the DNA. The length of each molecule and the position of each label is output in bnx files. Molecules were filtered above 150 Kbp. Sets of single-molecules, equivalent to about 100 x haploid coverage, for each sample, was then used to construct a *de novo* genome assembly.

### 

### Strand-seq Methods

**Library preparation and sequencing (Contributors: Sanders, A.):** EBV-transformed lymphoblastic cell lines were cultured in BrdU (40uM final concentration) for 36 hours and single cells sorted into 96-well plates based on Hoechst-quenching [(Sanders et al. 2017)](https://paperpile.com/c/ml0krT/6pJd). Library construction was performed on a Bravo liquid handler (Agilent technologies), to automate MNase digestion of nuclear DNA, ligation of indexed Illumina adapters, and Hoechst/UV treatment to remove BrdU strands and prepare samples for 17 rounds of PCR amplification, as described [(Sanders et al. 2017)](https://paperpile.com/c/ml0krT/6pJd).. Libraries from a single 96-well plate were pooled for post-PCR gel size selection, which enriched for the mononucleosomal fragment (~150bp + 120bp adapters) and dinucleosomal fragment (~320bp + 120bp adapters). For sequencing, 96 samples were from the mononucleosomal fragment was run on a single lane of an Illumina HiSeq (rapid-run mode), using a 76 bp paired-end protocol. 192 samples (two plates) of the dinucleosomal fragment were pooled and sequenced using a 151bp paired-end protocol.

**Sequence data processing (Contributors: Porubsky, D. Sanders, A.):** The Strand-seq raw sequencing data were demultiplexed based on the library-specific barcodes and converted to FASTQ files using Illumina standard software (bcl2fastq, version 1.8.4). Reads were aligned to GRCh38 human reference genome assembly, which includes decoy and HLA sequences. The FASTQ files were mapped to the reference genome using bwa aligner (version 0.7.12-r1039) according to the HGSVC guidelines for Illumina sequencing. Following alignment, reads were sorted using SAMtools [(Heng Li et al. 2009)](https://paperpile.com/c/ml0krT/eBY8) (version 1.2) and duplicate reads were marked using biobambam (version 0.0.191) [(Tischler and Leonard 2014)](https://paperpile.com/c/ml0krT/O3nZ). Based on common library-specific barcodes, the separate BAM files for the mono- and di-nucleosome fraction of each cell were merged into a single BAM using SAMtools (version 1.2). Directional read distribution of each Strand-seq libraries was assessed using BAIT [(Hills et al. 2013)](https://paperpile.com/c/ml0krT/dlsK) to preselect Strand-seq libraries based on read density, level of background reads and level of variability [(Porubský et al. 2016)](https://paperpile.com/c/ml0krT/TJmR). BAM files passing our quality criteria served as an input for inversion calling and haplotyping pipeline.

### 

### 10X Genomics

**Library preparation and Sequencing (Contributors: Jabara C.):** Cell lines were cultured and DNA extracted using the Qiagen MagAttract kit with modifications to enhance retention of long DNA molecules:

<https://assets.contentful.com/an68im79xiti/lCEjig84zQWoWiKaws8QY/d0872d726cc797579e4a8273e640b35d/20160607_SamplePrepDemonstratedProtocol_-_DNAExtractionfromBlood_RevB.pdf>

1.25 ng of high molecular weight DNA was loaded onto the 10x Chromium Controller using Chromium Genome v1 reagents following the recommended protocol:

[https://assets.contentful.com/an68im79xiti/4z5JA3C67KOyCE2ucacCM6/d05ce5fa3dc4282f3da5ae7296f2645b/CG00022\_GenomeReagentKitUserGuide\_RevC.pdf](https://assets.contentful.com/an68im79xiti/57PcvoZiWkU2u4Sweymw8A/1c65eefb966eaeb44a37eaa7ed1b0c3e/CG00021_Chromium_Controller_Training_Kit_User_Guide_RevA.pdf))

The initial part of the library construction takes place within droplets containing gel beads functionalized with barcodes that mark the droplet of origin (called GEMs). The library construction incorporates a barcode that is adjacent to read one. All molecules within a GEM get tagged with the same barcode, but because of the limiting dilution of the genome (roughly 300 haploid genome equivalents) the chance that two molecules from the same region of the genome are partitioned in the same GEM is very small. Thus, the barcodes can be used to statistically associate short reads with their source long molecule. The resulting library was sequenced on an Illumina X Ten sequencer to produce 2X150 paired-end sequences. The resulting data type is called ‘Linked-Reads’ (Zheng et al, 2016).

**Sequence data processing:** Sequence data was analyzed using the Long Ranger v.2.1 analysis pipeline. Briefly, reads are aligned using Lariat (https://github.com/10XGenomics/lariat), a wrapper around BWA that uses molecule information to adjust alignment locations and MAPQ, using the RFA methods [(Bishara et al. 2015)](https://paperpile.com/c/ml0krT/g8WU) . This allows for more reads to be confidently mapped. In these cases, map quality scores are adjusted so that downstream analysis can take advantage of these reads.

Single nucleotide variants (SNVs) and small indels are called using Freebayes (v0.9.21-7, default parameters (-0)).

Each read covering a heterozygous variant is re-aligned to a ~100bp segment of the reference sequence, with and without the alt allele applied, to determine whether the read gives clear support for one allele over the other. Mapping each read to a molecule via the GEM-specific barcode produces the yield of the set of alleles observed on each molecule. We model the likelihood of the per-molecule allele observations given a phasing configuration. The model follows [(Bansal et al. 2008)](https://paperpile.com/c/ml0krT/Z8gT) with additional terms to account for the small probability that a barcode carries two molecules from opposite haplotypes, and for the chance that an input variant is non-heterozygous. We search for the maximum likelihood phasing configuration by find near-optimal local configurations using beam-search over blocks of ~40 variants. Blocks are greedily joined to form a global solution, which we iteratively refine. The confidence of each phasing decision is the likelihood-ratio between optimal and next-best solutions. The phasing procedure solution implicitly produces a posterior distribution over haplotypes for each input molecule. Molecules covering >1 het are typically phased with very high confidence. We write this haplotype information to an auxiliary BAM tag on each read of a confidently phased molecule, which is used in downstream haplotype aware SV calling.

Input variants determined to be non-heterozygous are switched to HOM REF or HOM ALT.Variant calls are then emitted as a VCF file.

### TruSeq Synthetic Long Reads (TSLR)

**(Contributors：Sebat, J., Gabriel Rosanio, Danny Antaki, Masdhu Gujral, Joey Flores， Karine Viaud Martinez）**

The Illumina TruSeq SLR platform developed is based on the isolation of single molecules followed by amplification, barcoding and conventional Illumina sequencing (<http://bit.ly/2kFgQDc>). This method generates “Synthetic long reads” which consist of contigs that are assembled from the short read sequences derived from of a single molecule. The Illumina TruSeq SLR method is based on the same principle (http://bit.ly/2dmkEHK). A key advantage of this approach is the high sequencing accuracy afforded by the Illumina platform.

**gDNA purification**. Lymphocytes from individuals in a Yoruban (NA19238, NA19239, NA19240), a Han Chinese (HG00512, HG00513, HG00514), and a Puerto Rican (HG00731, HG00732, HG00733) trio were used to acquire genomic DNA. The DNA was purified from each individual separately using the DNeasy Blood & Tissue Kit as previously reported (Qiagen). Concentrations for each individual were determined using Qubit BR and confirmed by agarose gel electrophoresis.

**Sample prep.** Each individual’s gDNA was prepared for whole genome sequencing using the TruSeq synthetic long read prep kit according to manufacturer’s protocols with the exception the template concentration for each Long Range PCR reaction was 0.9fg/µl rather than the prescribed 0.6fg/µl. A total of 25 384-well plates were prepared for each sample.

**Sequencing**. Moleculo sequencing of prepared libraries were performed by Illumina, Inc. Paired-end sequencing was performed (add coverage) using an Illumina HiSeq 2000. (**Table 1.2.4**)

**Data Processing.** Sequenced paired-end samples were separated by barcode identification into 384 separate bins using TruSeq Long-Read Assembly App (Illumina). Assembly of separated paired-end sequences with ≥Q30 into synthetic long read sequences was performed with the TruSeq Long-Read Assembly App. Full documentation on this software can be found at <http://support.illumina.com/help/BS_App_LongReads_help/TruSeq_Long_Reads_Assembly_App.htm>

**Data Repository Information.** Resultant sequencing data (all separated paired-end sequences, synthetic long read assemblies, long read scaffold information, and a summary report) are available from the European Nucleotide Archive (<https://www.ebi.ac.uk/ena/submit/sra/#home>)

The pooled data from all preparations for these individuals showed common long read signatures: a drop off of assemblies at the arbitrary 1499bp/1500bp boundary; a decrease in longer assemblies; a peak around the expected ~10 Kbp selection length. The average assembled long read for NA19238, NA19239, and NA19240 are 3939bp, 4075bp, and 3834bp respectively. The N50 for each individual is 6877bp, 7291bp, and 6723bp respectively.

**Alignment**. Synthetic long reads greater than 1499bp in length were aligned to GRCh38 using the LAST aligner. LAST was selected over others, including BWA and BLAST, because of its speed and ability to align reads to multiple locations in the genome to optimize the alignment without any a priori information.

Simulated long reads at a depth of 7.5X from GRCh38 were also aligned to GRCh38. This alignment was of high quality as expected. Only 7 breakpoints were detected indicating a low rate of misalignment by LAST.

SV calling was performed by split read analysis of the multiple alignment files generated by LAST. Deletion, Tandem Duplication, Insertion and Inversion signatures were parsed out of the alignments. The final raw call set included all SVs that were 50 bp or greater in length for which at least 2 reads supported the identical call.

**Sequence data processing**: Sequenced paired-end samples were separated by barcode identification into 384 separate bins using TruSeq Long-Read Assembly App (Illumina). Assembly of separated paired-end sequences with ≥Q30 into synthetic long read sequences was performed with the TruSeq Long-Read Assembly App. Full documentation on this software can be found at:

<http://support.illumina.com/help/BS_App_LongReads_help/TruSeq_Long_Reads_Assembly_App.htm>

Resultant sequencing data (all separated paired-end sequences, synthetic long read assemblies, long read scaffold information, and a summary report) are available from the European Nucleotide Archive (<https://www.ebi.ac.uk/ena/submit/sra/#home>**)**

### Hi-C data generation

**Library preparation and Sequencing**: Hi-C was performed according to established methods[(Dixon et al. 2012; Lieberman-Aiden et al. 2009)](https://paperpile.com/c/ml0krT/EvW1+3mez), and using the restriction enzyme HindIII (NEB R3104) for digestion of chromatin prior to proximity ligation. Lymphoblastoid cell lines were obtained from Coriell Cell Repositories. Two independent biological replicates were performed for each cell line, using approximately 20 million cells per replicate. Cells for each replicate were cultured independently for at least 2 passages. Hi-C libraries were sequenced on an illumina HiSeq (2000, 2500, or 4000), generating roughly 250 million reads per replicate.

**Hi-C data processing and phasing** (**Contributors:Yunjiang Qiu, David U Gorkin, and Bing Ren**): Hi-C reads were aligned to the GRCh38 reference genome using BWA-MEM with default parameters [(Heng Li 2013)](https://paperpile.com/c/ml0krT/T1Zh). Hi-C reads are paired-ended, but we align each read end to the reference genome independently because standard paired-end mapping algorithms are not designed to handle the large distances that separate Hi-C read pairs. After mapping, we performed several filtering steps:

1) Hi-C reads may span a ligation junction, in which case two parts of the read may map to two different regions of the reference. BWA-MEM handles these “chimeric” reads by outputting two different alignments – one for each part of the read. For Hi-C data we are interested specifically in the alignment of the 5’ portion of the read (i.e. 5’ to the ligation junction), because the region 3’ to the ligation junction will be captured by the other read in the pair. Thus, we filtered the BWA-MEM output to keep only the 5’ alignment when split alignments were reported. For similar reasons, if the 5’ portion of a read did not align, that read was discarded.

2) Low quality alignments were removed (MAPQ < 10).

3) Read ends were re-paired, and any only pairs in which both read ends passed all filters were kept for downstream analysis.

4) PCR duplicates were removed with Picard. Phasing was then performed using the Haploseq pipeline as previously reported [(Selvaraj et al. 2013)](https://paperpile.com/c/ml0krT/5ENg).

Briefly, aligned read pairs were realigned and recalibrated using GATK [(McKenna et al. 2010)](https://paperpile.com/c/ml0krT/iJRH). The badmate parameter was disabled to keep long range read pairs. Hapcut was then used to perform phasing [(Bansal and Bafna 2008)](https://paperpile.com/c/ml0krT/v9S3). Haploseq modifies Hapcut by calculating the probability that read pairs come from different chromosome homologs based on insert size, and then adjusting the base quality alignment scores to account for this probability. In almost all cases Hapcut generated one haplotype block per chromosome, spanning both arms of the chromosome. One exception is chromosome X in HG00513, for which Hapcut reported separate haplotype blocks for each arm of the chromosome, reflecting an inability to reliably phase the two arms relative to eachother. For chromosome 1 and chromosome 9 in all individuals, each chromosome arm was phased separately. These two chromosomes have exceptionally large centromeric repeat arrays, and thus Haploseq cannot reliably phase the chromosome arms relative to each other at the sequencing depth obtained in this study.

### 

### Transcriptome Sequencing

**Library preparation and Sequencing (Contributors: Talkowski, M):** Total RNA was extracted from EBV-transformed lymphoblastoid cell lines (LCLs) using TRIzol® (15596026, Thermo Fisher Scientific Waltham, MA, USA) from cell pallets according to manufacturer’s instructions. In brief, cell pellets (containing between 1-5e6 cells) were homogenized in TRIzol® reagent, followed by chloroform addition and phase separation. RNA was precipitated from aqueous phase using isopropanol followed by washing the RNA pellet with 75% ethanol. RNA pellet was suspended in RNase-free water and stored at -80⁰C. All nine strand-specific RNAseq libraries were prepared using the Illumina TruSeq kit (Illumina, San Diego, CA, USA) according to manufacturer’s instructions, as described [(Blumenthal et al. 2014; Sugathan et al. 2014)](https://paperpile.com/c/ml0krT/0pg2r+J6Pgf). One microliter of diluted (1:100) External RNA Controls Consortium (ERCC) Spike-in Mix (4456740, ThermoFisher) containing 92 synthetic RNA standards of known concentrations and sequence was added to each RNA-sequencing library alternating between mix1 and mix2 for each well in batch to estimate the detectable expression abundance thresholds. PolyA bead capture was used to enrich for mRNA, followed by stranded reverse transcription and chemical shearing to make appropriate stranded cDNA inserts for library. Libraries were completed by addition of sample specific barcodes and adapters for Illumina sequencing followed by 10 cycles of PCR amplification. Final concentration and size distribution of libraries were evaluated by Agilent 2200 TapeStation (Agilent, Santa Clara, CA, USA) and/or qPCR, using Library Quantification Kit (KK4854, Kapa Biosystems, Wilmington, MA, USA), and multiplexed by pooling equimolar amounts of each library prior to sequencing. Libraries were multiplexed, pooled and sequenced on multiple lanes of an Illumina HiSeq2500, generating an average of 67.5 million paired-end reads of 51 bp per sample.

**Sequence data processing:** Further quality control of sequence reads was assessed by fastQC [(Andrews and Others 2010)](https://paperpile.com/c/ml0krT/sl57) (v.0.10.1). Subsequently, sequence reads were aligned to human reference genome Ensembl GRCh38 (v.81) using GSNAP [(Wu and Nacu 2010)](https://paperpile.com/c/ml0krT/IODPZ) (v. 06-23-2015) with options –N 1 –B 3 --quality-unk-mismatch=1. Further quality control of alignments was assessed by a custom script utilizing Picard Tools (http://picard.sourceforge.net) RNASeQC [(DeLuca et al. 2012)](https://paperpile.com/c/ml0krT/3u8j2), RSeQC [(Wang, Wang, and Li 2012)](https://paperpile.com/c/ml0krT/nbR7j) and SamTools [(Heng Li et al. 2009)](https://paperpile.com/c/ml0krT/eBY8) Gene level counts were tabulated using BedTools’s multibamcov algorithm [(Quinlan and Hall 2010)](https://paperpile.com/c/ml0krT/0RHCh) (v. 2.17.0) on unique alignments for each library relying on Ensembl gene annotation [(Aken et al. 2016)](https://paperpile.com/c/ml0krT/S8KEu) (GRCh38 v.81). Analysis of ERCC spike-ins as described in Blumenthal et al. [(Blumenthal et al. 2014)](https://paperpile.com/c/ml0krT/0pg2r) estimated the expression threshold for detection to be at least three mapped reads

## 

## 1.3 Data and tools availability

**Data coordination:** raw data, Illumina alignments and data access in the International Genome Sample Resource (IGSR) (Fairley, Clarke, Zheng, Lowy and Flicek) as part of the IGSR (International Genome Sequence Resource)

The International Genome Sample Resource (IGSR) is working with the HGSVC to assist with data coordination, analysis and distribution. The data collected by the HGSVC is available via an FTP site (ftp://[ftp.1000genomes.ebi.ac.uk/vol1/ftp/data\_collections/hgsv\_sv\_discovery](http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery)/), can be browsed through the data portal (<http://www.internationalgenome.org/data-portal/data-collection/structural-variation>) and used under the Ft. Lauderdale principles for data reuse and publication <https://github.com/igsr/1000Genomes_data_indexes/blob/master/data_collections/hgsv_sv_discovery/README_hgsvc_datareuse_statement.md>.

**Data sets used by the HGSVC:** The data sets used by the HGSVC are provided in **Table S1.3.1**. As noted, data from the HGSVC can be found at this URL: ftp://[ftp.1000genomes.ebi.ac.uk/vol1/ftp/data\_collections/hgsv\_sv\_discovery/](http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/). In addition, reference files used more widely in IGSR, such as the reference genome, are also available from the wider IGSR FTP site.

# 2. Chromosomal haplotype resolution and integration

We first aimed to build dense and accurate chromosome-length haplotypes on single-nucleotide variants (SNVs). These haplotypes can then act as a backbone to add further variation to and facilitate a comprehensive comparison across platforms. To establish basic properties of the haplotype information delivered by different platforms, we first created haplotypes from each technology individually. To this end, we used high-coverage Illumina paired-end data and ran FreeBayes (v1.0.2) and the GATK HaplotypeCaller (v3.5-0-g36282e4). We filtered the calls (FreeBayes: QUAL>=30, GATK: QUAL>=200) and retained those bi-allelic SNV calls where the three genotypes for all samples from a family agree between the two callers.

## 2.1 Phasing of independent datasets

**(Contributors: David Porubsky, Tobias Marschall)**

Chromium-seq data was processed by 10X Genomics' LongRanger software (v2.1) to produce haplotype blocks and Strand-seq data was assembled into chromosome-length haplotypes using StrandPhaseR [(Porubsky et al. 2017)](https://paperpile.com/c/ml0krT/EzLl), (see Method section 2.4 below for details). Generating haplotypes from aligned Illumina paired-end reads, PacBio reads, and Moleculo reads was formalized as the weighted Minimum Error Correction (wMEC) problem [(Lippert et al. 2002)](https://paperpile.com/c/ml0krT/ouNF). For each chromosome, we consider a matrix with one row per read and one column per variants and entries from {0,1,-}, where 0 and 1 encode that a read supports the reference or alternative allele at a variant locus, respectively, and "-" indicates that a read does not cover the variant. Each matrix entry comes with a (phred-scaled) weight, corresponding to the confidence that the allele has been correctly called. The wMEC problem now asks to find a minimum-weight set of bits to flip, so that the reads can be partitioned into two sets with no conflicting allele calls within each set. Each set of reads hence corresponds to one haplotype. The wMEC problem is NP-hard [(Lippert et al. 2002; Cilibrasi et al. 2005)](https://paperpile.com/c/ml0krT/ouNF+WDxO) and only recent algorithmic progress [(Patterson et al. 2015)](https://paperpile.com/c/ml0krT/w0aL) has enabled solving practically relevant instances optimally. To perform read-based phasing of Illumina paired-end, PacBio and Moleculo reads, we used the WhatsHap software suite (<https://bitbucket.org/whatshap/whatshap>, version 0.13+42.g6d8f8b6), which implements this algorithm. WhatsHap additionally allows to detect alleles from aligned reads by re-aligning windows of each read to both reference and alternative allele, which boosts performance when handling error-prone long reads dramatically [(Martin et al. 2016)](https://paperpile.com/c/ml0krT/fIdJ). We used this functionality by providing option “--reference” on the command line and additionally used option “--distrust-genotypes”. Phasing of Hi-C data was performed using HapCut2, as described in Section 1.2 (“Hi-C data generation”). For comparison purposes, we ran eagle2 [(Loh et al. 2016)](https://paperpile.com/c/ml0krT/MSqQ) on all variants using hg19 coordinates to generate a population-based phasing with respect to the 1000 Genomes phase 3 reference panel [(1000 Genomes Project Consortium et al. 2015)](https://paperpile.com/c/ml0krT/7tfu). All panel samples related to one of the trio samples were removed prior to the population-based phasing. Furthermore, we generated haplotypes only using the genotypes of a family and no sequencing reads, corresponding to “genetic haplotyping”, which we also performed using WhatsHap (this mode is run by providing option --ped and no BAM files).

Strand-seq is the only included technology yielding chromosome-length haplotypes by design, while read-based phasing is limited by the read length in relation to the distance between consecutive heterozygous markers (**Figure S2.1.1**). Pairs of heterozygous SNV pairs that are not covered by any read, lead to fragmented phasings (see **Main Figure 1A,B**).

Population-based phasing using eagle2 leads to one phased block per chromosome (containing all SNVs present in the reference panel). Even though the switch error rate is low (**Main Figure 1C**), population-based phasing usually is not correct over the length of entire chromosomes. To avoid unfair comparisons, we did not provide a mismatch error rate in **Main Figure 1F**, but instead display an ideogram of mismatches in **Figure S2.1.2**.

## 2.2 Phasing using multiple data sources

**(Contributors: David Porubsky, Tobias Marschall)**

As discussed in the main text, combining different technologies is a powerful approach to obtain dense, chromosome-scale haplotypes. To perform integrative phasing from a pair of technologies, we used WhatsHap (version 0.13+42.g6d8f8b6), providing it with data from two technologies [(Porubsky et al. 2017)](https://paperpile.com/c/ml0krT/EzLl). In case of Illumina paired-end reads, Moleculo synthetic long reads, and PacBio reads, we provided WhatsHap with the corresponding BAM file. In case of 10X, Hi-C and Strand-seq, we used the corresponding prephased VCF as input to WhatsHap (as described above, these VCFs were created using LongRanger, HapCut2, and StrandPhaseR, respectively).

In **Main Figure 1D-F**, we provided display the resulting performance for selected technologies (and technology combinations). Here, in Figure **S2.2.1**, we display corresponding results for all combinations of technologies.

## 

## 2.3 Mapping of meiotic recombination events

**(Contributors: David Porubsky)**

Global chromosome-length haplotypes assembled using Strand-seq can be used to map meiotic recombination events within a single family trio. We mapped all meiotic recombination events in every family trio solely based on Strand-seq (**see Method section 2.5, Figure S2.3.1, Table S2.3.1**). As expected we observed higher number of meiotic recombination events occurring on the maternal homologues [(Broman et al. 1998; Kirkness et al. 2013; Hou et al. 2013; Lu et al. 2012)](https://paperpile.com/c/ml0krT/56V5+yCL4+cmQJ+aXlX). This difference was the most prominent in Yoruban trio and the least for the Puerto Rican trio (**Figure S2.3.1**). Using Strand-seq data only we achieved high resolution maps of meiotic recombination with median resolution less than 25kb. We set to further refine this maps using Pacbio reads and WhatsHap. The majority (~70%) of mapped meiotic breakpoints overlapped with breakpoint estimates obtained from PacBio reads and allowed us to further refine the mapped meiotic breakpoints (**see Method section 2.6, Figure S2.3.2, Table S2.3.1)** Notably, with this approach we managed to map meiotic recombination breakpoints with unprecedented resolution (median ~1,5kb). In addition, all but one PacBio refined breakpoints were mapped at maximal possible resolution between two heterozygous SNVs. The majority of Strand-seq mapped meiotic recombinations that were not refined by PacBio could be in theory refined further using residual SNVs present in such meiotic recombinations breakpoints.

We have further explored meiotic recombination breakpoints refined by PacBio reads (in total 162 breakpoints from all trios) to search for previously reported elements, THE1A and THE1B, specific for human recombination hot spots [(Myers et al. 2008)](https://paperpile.com/c/ml0krT/559U). Overall we have found ~ 1.6x higher abundance of above mentioned elements around mapped meiotic breakpoints then at random positions in the genome (**Figure S2.3.3**).

## 2.4 Strand-seq phasing

**(Contributors: David Porubsky)**

To assemble genome-wide haplotypes exclusively from Strand-seq data we have used R package called StrandPhaseR. StrandPhaseR takes as an input aligned BAM (binary alignment map) files from single cell libraries that were pre-selected based on the following quality criteria. BAM files were filtered for duplicate reads, alternative alignments and low mapping quality reads (mapq < 10). Haplotype informative WC (Watson-Crick) regions were localized in every single cell as was done previously [(Porubský et al. 2016)](https://paperpile.com/c/ml0krT/TJmR). To phase such WC regions across all single cells we used our latest phasing pipeline called StrandPhaseR [(Porubsky et al. 2017)](https://paperpile.com/c/ml0krT/EzLl). List of single nucleotide heterozygous position suitable for phasing were obtained from 10x Genomics variant calls. Variable positions covered with bases of quality less than 20 were filtered out. For each individual, final haplotypes were exported as a single VCF file, separately for each chromosome.

## 

## 2.5 Mapping of meiotic recombination events using Strand-seq

To map meiotic recombination event in each family trio we have first assemble genome-wide haplotypes independently for each individual in a trio using solely Strand-seq data [(Porubský et al. 2016)](https://paperpile.com/c/ml0krT/TJmR). Next we performed a pairwise comparison of each child's homologue to both, maternal and paternal homologues. This identified the heterozygous positions that distinguished the child from each parent and such positions were used to assign the parental identity to each child’s homologue. Every comparison was encoded as a vector of zeros and ones based on the parental homologue to which child's homologue correspond. (zero – parental homologue 1, one – parental homologue 2). Then a circular binary segmentation algorithm (R package fastseg, minSeg set to 100) [(Klambauer et al. 2012)](https://paperpile.com/c/ml0krT/mZJU) was applied on such binary vectors. Segments smaller than 5 Mb and segments overlapping with homozygous inversions were filtered out. Meiotic breakpoints were localized as the end position of one segment and start position of the following segment. All localized meiotic breakpoints were further assessed and confirmed by eye.

## 

## 2.6 Refinement of mapped meiotic breakpoints using PacBio reads

Meiotic recombination breakpoints were independently predicted using pedMEC algorithm implemented in the WhatsHap [(Garg, Martin, and Marschall 2016)](https://paperpile.com/c/ml0krT/vcDw). PacBio reads from the whole trio along with human meiotic recombination rates were used to estimate most likely point of meiotic recombination. Little number of overlapping breakpoints were merged together leaving us with a few hundreds of predicted breakpoints. To exclude false positives, we considered meiotic breakpoints mapped using Strand-seq as a gold-standard. Therefore, only PacBio predicted breakpoints that overlapped with or were in 50kb distance from Strand-seq meiotic breakpoint were considered for further analysis. For such ranges we reported refined breakpoint position using PacBio reads. Lastly we searched for meiotic breakpoints with a number of residual SNVs within the breakpoint. Then in turn we have established a theoretical resolution of such breakpoint as a maximal distance between subsequent residual SNVs within such breakpoint.

# 

# 3. Haplotype specific *de novo* assembly

Existing methods for detecting structural variation on SMS sequences lack sensitivity on diploid genomes [(Huddleston et al. 2016)](https://paperpile.com/c/ml0krT/s4fF). To address this we developed a strategy for variant calling where dense chromosome-scale phased SNPs are used to partition SMS reads by haplotype, and SV is called by assembling each haplotype separately and detecting structural variation as differences between the haplotype assemblies and the reference. Whole-chromosome phasing was generated using a combination of Strand-Seq [(Falconer et al. 2012)](https://paperpile.com/c/ml0krT/rDMh) and 10x-Genomics, phased by WhatsHap [(Patterson et al. 2015)](https://paperpile.com/c/ml0krT/w0aL) (see below). SNV density is sufficiently to provide very good contiguity. The mean distance between phased SNV sites was 1360, 1497, and 1040 bp for the HAN, PUR, and YRI children, respectively **(Figure S2.1.1).** We aimed to generate assemblies for each haplotype *despite having lower sequence coverage per haplotype than previous human and great ape studies* [*(Berlin et al. 2015; Steinberg et al. 2016)*](https://paperpile.com/c/ml0krT/uUWF+4OYS). To do so, we generated *de novo* assemblies using two related assembly approaches: local (LA) which we termed Phased-SV and an unguided *de novo* (DN) method termed (MsPAC), and used concordance between assemblies to validate integrity.

## 3.1 Phased-SV

**(Contributors: Chaisson)**

For the LA approach, we assembled reads into 60 Kbp regions spaced at 20 Kbp intervals across the genome (148923 regions). Reads were mapped using *BLASR* [(M. J. Chaisson and Tesler 2012)](https://paperpile.com/c/ml0krT/40oc) to GRCh38. Regions were labeled as haplotype-resolved if at least 20 heterozygous SNVs were present in the region (HAN=90631, PUR=83758, YRI=113972 regions). Otherwise regions of the genome were considered autozygous. In haplotype-resolved regions, reads were partitioned by haplotype from each child, and combined with reads of the parents from the corresponding inherited parental haplotype (for a total of 30-fold sequence coverage per haplotype). The autozygous regions were assembled without partitioning reads. The average fraction of partitioned reads was 60.1%, 67.0%, and 70.1% for the HAN, PUR, and YRI children, respectively. The set of local assemblies from each haplotype combined with the assemblies from autozygous regions were then merged into megabase-scale haplotype resolved contigs. For each haplotype, a directed acyclic graph (DAG) was generated with a vertex for every local assembly from the haplotype as well as local assemblies from the autozygous regions, with edges connecting two nodes if the corresponding local assemblies were from genomic regions separated by at most 100 Kbp, and with overlap alignments at least 10 Kbp, with direction of the edge determined by the genomic order of the local assemblies. A contig was generated corresponding to the longest path in each weakly connected component of the DAG.

## 3.2 MsPAC

For haplotype-partitioned de novo assembly, reads were aligned to GRCh38 using BLASR, as described above. Using whole-chromosome phasing information provided by WhatsHap, we partitioned the reads into 3 sets: haplotype 1 reads, haplotype 2 reads, and unassigned reads. To infer the most likely haplotype, for each read we examine all base QVs, *qi*, incident on SNVs. For each read we give it a simple haplotype score using the product of *qi* for each matching base or 1-*qi* for each based which does not match. Scores for haplotype 1 and 2 are compared; reads are assigned to the haplotype with higher score if the difference in scores exceeds a phred-scaled QV of 10. Any remaining reads, including those which do not span any SNVs, are labeled as ambiguous. The average coverage per sample decreased by 39.84% +- 1.52 of the original coverage per haplotype after partitioning.

After partitioning, each haplotype is assembled independently. For each haplotype, candidate assembly intervals were defined as those with greater than 3X coverage. Prior to assembly each such region (10157, 8908, 12829 in NA19240, HG00733, HG00514 in haplotype 1 and 10065, 8665, 12826 in NA19240, HG00733, HG00514 in haplotype 2) was split into subintervals of 270 kb (with 10 kb overlap between adjacent intervals). For each haplotype specific region, ambiguous reads overlapping the intervals were also recruited. The combined read set was then assembled using a step-wise approach. First, assembly was performed with Canu [(Koren et al. 2017)](https://paperpile.com/c/ml0krT/4JVv) with parameters: contigFilter="2 1000 1.0 1.0 2"; corMinCoverage=0; errorRate=0.035. Next, for regions that did not assemble, a more permissive assembly was performed using minimap and miniasm [(Heng Li 2016)](https://paperpile.com/c/ml0krT/Hx6E) with error-corrected reads generated in the prior step by Canu. The resulting process sometimes generated multiple contigs that did not span the whole interval. In the regions with no contigs, reads were extracted and locally reassembled by extracting subinterval reads with 10 kb flanks. The resulting assemblies were quivered and trimmed for regions with Phred QV greater than 30 using cutadapt [(Martin 2011)](https://paperpile.com/c/ml0krT/rss9), to yield the initial set of assembled haplotigs.

For each haplotype, contigs were mapped back to the reference using BLASR and the -alignContigs option. Any overlapping contigs were merged by greedily extending the upstream contig to its last aligned base, i, before adding bases from the downstream contig beginning at i+1. After creating the stitched haplotigs, a final step was performed to merge de novo assembled sequence from Falcon (<https://github.com/PacificBiosciences/FALCON>), described previously. Here, Falcon assemblies substrings were only integrated in “gap” intervals if they anchored to both the flanking left and right haplotigs of an examined gap interval. The N50 before adding in the Falcon assembly was 277KB, 243KB, and 169KB for haplotype 1 for HG00733, HG00514, and NA19240, respectively. For haplotype 2, the N50 is 277KB, 242KB, 171KB; after addition of the de novo assemlby the N50 contig was for haplotype 1 was 6MB, 3.2MB, 1.5MB and for haplotype 2 the 6.3MB, 3.1MB and 1.4MB. The merged assemblies were then aligned back to the references with BLASR using the parameters (*-alignContigs -noSplitSubreads -bestn 1 -clipping soft*).

## 

## 3.3 Assembly coverage

The haplotypes from each assembly were aligned using blasr -alignContigs -minMapQV 30 ([www.github.com/mchaisso/blasr](http://www.github.com/mchaisso/blasr)), and intersected with chromosomal regions not labeled as acrocentric bands within the UCSC cyctoband tables <http://genome.ucsc.edu/cgi-bin/hgTables?clade=mammal&org=Human&db=hg38&hgta_group=map&hgta_track=cytoBandIdeo&hgta_table=0&hgta_regionType=genome&hgta_outputType=bed>. The coverage is reported in **Table S3.3.1**. Assembly quality was measured by mapping BAC sequences (27 in total) to both haplotypes of each assembly and counting the difference to the the optimally aligned haplotype. The result is given in **Table S3.3.2.**

# 

# 4. Full-spectrum variation detection

## 4.1 Short insertion and deletion (<50bp)

**(Contributed by: Ye, Li)**

Current methods for indel discovery using short-read sequence data are thought to underestimate by as much as 40% the true number of events [(Huddleston et al. 2016; 1000 Genomes Project Consortium et al. 2015)](https://paperpile.com/c/ml0krT/s4fF+7tfu). This effect is especially pronounced for events greater than 10 bp in length. In order to provide a more comprehensive map of human genetic variation from SNVs to large complex SVs we therefore analyzed the three trios for the presence of indels define here as insertions and deletions ranging from 1bp to 49 bp.

### 

### Merging of Illumina callsets

Indels from three Illumina callsets: Pindel [(Ye et al. 2009)](https://paperpile.com/c/ml0krT/40tY), GATK [(DePristo et al. 2011)](https://paperpile.com/c/ml0krT/CIPuB) and FreeBayes [(Garrison and Marth 2012)](https://paperpile.com/c/ml0krT/dlOl) were merged. The merged region is between 1 and 49 bp. To create a deletion and insertion merged set, calls in each callset were separated into deletions and insertions. For CHS, Pindel, GATK and FreeBayes detect 551417, 305509, 160381 deletions and 452390, 275998, 148707 insertions respectively. For PUR, Pindel, GATK and FreeBayes detect 562056, 312359, 169348 deletions and 459026, 282492, 154248 insertions respectively. For YRI, Pindel, GATK and FreeBayes detect 564226, 393464, 218569 deletions and 456768, 354097, 187615 insertions respectively. A merged deletion and a merged insertion set were created by merging Pindel, GATK and FreeBayes tabix indexed set (tabix version 0.2.6) using vcf-merge function of vcftools [(Danecek et al. 2011](https://paperpile.com/c/ml0krT/jlti), version 0.1.13) with default options. The merged deletion and insertion Illumina set contain 1,166,979 and 1,077,222 calls respectively. Size distribution of deletions and insertions in individual set and merged set is shown in **Figure S4.1.1A** (Deletion) and **Figure S4.1.1B** (Insertion).

### 

### PacBio indels

**(Contributed by: Chaisson, M)**

To avoid artifacts of merging indels from multiple assemblies, we called indels from the Phased-SV haplotigs only. Indels were detected in the regions of haplotigs that were realigned during SV calling, and from the haplotig to reference alignments otherwise. We ignored all single base indels and homopolymer indels less than 6 bases. Indels were maximally left-aligned using vt normalize.

Platform comparisons

**(Contributed by: Ye, Li)**

We next compared the merged illumina callset with indel calls generated the unified PacBio callset and unified Phased-SV/MsPAC (PS/MP) call set (below). PacBio Phased-SV calls from HG00514 were used to compare with the merged illumina callset up to 1 Kbp. Similarly, PS/MP calls were also divided into deletions (449,146) and insertions (384,358). The size distribution of PS/MP PacBio and Illumina merged set was shown in **Figure S4.1.1** A (Deletion) and B (Insertion). The contribution of three Illumina and the PacBio callsets to the illumina-PacBio integrated set was shown in Venn diagrams (**Figure S4.1.1** C) created by overlapping the four callsets using vcf-compare [(Danecek et al. 2011)](https://paperpile.com/c/ml0krT/jlti). Additionally, an overlap between Illumina merged set and UW-MSSM PacBio callset was conducted. BED files were created from VCF files of the three callsets by padding SV lengths to breakpoints and then compared using bedtools intersect (version 2.26.0), with a 50% reciprocal overlap. The results were shown in **Figure S4.1.1 D** On the size frequency plots of deletions and insertions, 2n peaks are clearly visible, probably due to the higher mutation rate in microsatellite repeats. The numbers of calls for indels smaller than 30bp are comparable between Illumina merged set and UW-MSSM PacBio callset while the the sensitivity of the latter remains after Illumina merged set lost detection power, especially for insertions. The ratio of number of calls detected by the Illumina based methods relative to the number of calls by the PacBio based methods is given in **Figure S4.1.2**.

Repeat analysis

The BED files of deletion and insertion calls in Illumina merged set, Phased-SV set, and HySA were overlapped with different repeatMask bed files using bedtools intersect (version 2.26.0), looking for deletions and insertions fully residing in a repeat (-f 1.0). The repeat overlapped calls (shown in tables below), converted to frequencies, were summarized **Figure S4.1.1 F,** where each color bar represents a fraction of deletion/insertion calls overlapped with each repeat class. “Genome” represents the background fractions of each repeat class in human genome. Simple repeats are enriched for short indels and all three sets (Phased-SV callset, HySa and Illumina merged set) have alleviated proportion of simple repeats, even though the pacbio related sets have significantly higher percentages.

## 

## 4.2 Haplotype-resolved SV characterization and integration

### PacBio based SV detection.

**SV detection from *de novo* assembly (Contributed by Rodriguez, O. and Bashir, A):** For each query a tiling path of blasr alignments is selected by taking the highest scoring set of non-overlapping alignments on the query. If two alignments share the same alignment score, they are discarded. Blasr alignments passed to a four-state HMM adapted from [(Ritz et al. 2011)](https://paperpile.com/c/ml0krT/aL1B). Specifically, the HMM contains four states (normal, insertion, deletion, complex) used to identify putative SV candidates on the reference. The SV candidates were then passed through a filtering, annotation, and breakpoint refinement procedure. In short, calls were first intersected against known satellite intervals. Any calls in known-satellite intervals were eliminated. Next, all calls were intersected with repeatmasker and tandem repeat finder (TRF) tracks from the UCSC genome browser for preliminary annotation; any SVs containing inserted sequences were directly passed into repeatmasker (with TRF enabled) for annotation. For calls in which both insertion and deletion sequence existed an additional post-processing step attempted to refine the breakpoints by performing a forced end-overlap alignment of the two intervals. Those with a levenshtein ratio > 0.85, were adjusted to the revised reference boundary coordinate and marked in the VCF as “EXTRA”. Remaining calls were taken (with 1 kb flanking sequence) and analyzed using a custom script based on local nucmer alignments adapted from our previous work in Sudmant et al [(Sudmant, Rausch, et al. 2015)](https://paperpile.com/c/ml0krT/R82H). Lastly, using the alignment information, base quality within the SV event, and assembly quality information calls were labeled as passing or given a filtering label. Those in which the secondary alignment step identified the reference sequence to similar to the ALT sequence were also filtered.

**SV filtering and support (Contributed by: Chaisson, Rodriguez, Bashir):** Structural variants called by the two methods were first filtered by PacBio read alignment count (PBRC) and comparison with Bionano genomics (BN) SV callset. To determine PBRC, structural variants were organized into clusters where all variants within a cluster had boundaries within a specific window (length=250bp). The original Phased-SV call sets formed ~22,570 clusters (HG00514=21544, HG00733=21461, and NA19240=24716) from the original set of ~33,000 ( HG00514=31869, HG00733=32100, and NA19240=36774 original calls. Support for MsPAC calls was determined separately for each haplotype prior to merging and had between 12661 and 18902 clusters of SVs from each source call set. For each SV cluster, we define the start position of the reference interval of the cluster as 1 Kbp upstream of the first variant. We define the end position as 1 Kbp downstream of the endpoint of the last SV in the cluster (if the SV was a deletion), or 1 Kbp downstream of the starting point of the last SV in the cluster (if this SV was an insertion). A target database of two sequences was generated for each cluster: the reference sequence extracted from the corresponding interval to the SV, and an alternative sequence containing modifications to the reference interval sequence defined by the SVs in the cluster; e.g. if the cluster contained a 100 bp deletion, this sequence would be removed from the alt-reference. All reads overlapping the first SV in the cluster were mapped to the two-sequence database, and reads were assigned to the target with according to best alignment score. We filtered out reads with alignments < 1.5 Kbp.

Of the variants that remained after filtering, 12938/14508, 13747/15073, and 15298/16995 MS-PAC filtered calls were within 1 Kbp of a PhasedSV call.

To determine the minimal number of reads aligning to the alt-reference required to validate a call, we performed a simulation. The level of support for the alternate allele was measured for 22,166 variants detected on haplotype 0 of NA19240 randomly shuffled to different euchromatic regions of the genome (**Figure S4.2.1**). We found that >4 reads supporting the alternate allele results in an estimated false discovery rate of 0.21%. A total of HG00514=28107, HG00733=27969, and NA19240=32334 calls from Phased-SV had PBRC > 4, and the MsPAC call sets had between 9518 and 25063 calls with PBRC > 4.

As a final step to recover SVs with low sequence coverage, we compared the initial SMRT SV callsets to variants discovered through Bionano Genomics (La Jolla, CA).

Bionano calls were intersected with PacBio calls generated by MsPac and PhasedSV. Bionano SV calls do not yield precise breakpoints, but instead provide reference intervals (, ) and query intervals (, ) for which the observed nick-site distance is significantly different from expectation. We calculated the Bionano event size, , as . positive values for x indication insertions while negative values indicate deletions.

We then compared these intervals to all SV events (passing and non-passing) predicted from MSPac and PhasedSV. For all PacBio events overlapping a Bionano event we define LPB as the difference between the reference and alt allele. We can then score the similarity between each of the events to the Bionano event as *f*BN= |LBN-LPB|/LBN, selected the corresponding event which minimized *f*BN. A total of 257, 262, and 315 deletion events, and 508, 528, and 484 insertion events for PhasedSV were identified as having a best match and 270, 265, and 290 deletion events and 184, 181, and 197 insertion events for MsPac (**Table S4.2.1**, **Figure S4.2.2**) shows the concordance between the estimated PacBio and Bionano event sizes. Note, we also examined concordance of the LBN with sum of pacbio events (in each callset) overlapping the interval: , where corresponds to an event *i* that overlaps the Bionano intervals. However, this led to lower concordance then selecting the single best event.

### Unified PacBio SV callset

**(Contributed by: Chaisson, Rodriguez, Bashir)**

We then generated a unified call set combining the Phased-SV and MS-PAC SVs. We initially observed low reciprocal overlap between the two call sets. At 50% reciprocal overlap, 8,906/14,507 (HG00514), 9,423/15,072 (HG00733), and 10,391/16,994 (NA19240) quality-filtered MS-PAC calls had 50% overlap between calls of the respective call sets from Phased-SV. However, most calls from MS-PAC have a call in Phased-SV that was located in close proximity: 12,938/14,507 (89.2%) 13,7474/15,072 (91.2%) 15298/16,995 (90.0%) have a call at most 1 Kbp away in Phased-SV p=6.06E-5, permutation test, deviations above the expected number of a random permutation of SVs. Manual inspection found the source of discrepancy of overlaps to be differences in the precise coordinates of SV especially when there are multiple alignment solutions such as within tandem repeats. Since a large fraction of the variation occurs within VNTR and STR sequence, alignment artefacts created ambiguity in the start and end position of the SV influencing the bulk statistic. To address merging calls from the Phased-SV and MS-PAC with respect to the low reciprocal overlap, we applied a hierarchical approach to classify and merge variants. First, we included variants from either Phased-SV or MS-PAC that were validated by a BN call, deciding with optimal overlap with the BN call when the Phased-SV and MS-PAC call were overlapping (**Table S4.2.2**) . Each variant from Phased-SV and MS-PAC that overlapped a Bionano call was assigned a fractional value *f*BN= |LBN-LPB|/LBN where LPB is the length of a the variant from the Phased-SV or MS-PAC, and LBN is the length of the Bionano variant. Variants with *f* < 0.1 were considered validated, regardless of whether or not the PBRC was > 4. For NA19240, this selected 302 deletions (mean 5931 bp), and 168 insertions (mean 2330 bp) from MS-PAC, and 304 deletions (mean 4397bp) and 281 insertions (mean 2901) from Phased-SV. The procedure recovered ~30 SVs (HAN=33, PUR=21, YRI=38) that would have been filtered by low PBRC. Next, because the validation rate of the PhasedSV calls was greater, we included all remaining calls from PhasedSV that had PBRC > 4. For NA19240 this included 9,350 deletions (mean 272 bp) and 14413 insertions (mean 336 bp). We then added all calls from MS-PAC that were at least 10 Kbp away from a call in PhasedSV, for NA19240 this added 1223 deletions (mean 374 bp) and 602 insertions (mean 565 bp). A summary of the results is given in **Table S4.2.2**.

### 

### Intersection with CHM1 and CHM13 SV Callsets

**(Contributed by: Audano, Chaisson, Eichler)**

Structural variants (SVs) were compared with previously reported variants in hydatidiform moles CHM1 and CHM13 as previously reported by Huddleston *et. al.* [(Huddleston et al. 2016)](https://paperpile.com/c/ml0krT/s4fF). The method used to generate these SV calls is not reliable in centromeres and dense tandem-repeats in peri-centromeric regions, and so SVs from this study and SVs reported by Huddleston *et. al.* were filtered with a custom BED file before comparison.

From the filtered variants, a non-redundant set of SVs in CHM1 and CHM13 were merged by selecting all CHM1 variants and all CHM13 variants that do match a CHM1 variant by 50% reciprocal overlap. Filtered SVs from each HSVC call set were then matched with this non-redundant set by 50% reciprocal overlap.

Approximately 50% of the SVs in HG00514 and HG00733 (50.6% and 50.4%, respectively) and 60.0% of the N19240 calls were matched to an SV in the merged CHM1 and CHM13 calls (**Table S4.2.3**).

We counted HGSVC variants located distally from variants in the merged CHM1/CHM13 call set. Three sets of regions were computed taking the reference bases affected by each CHM1/CHM13 variant, adding 1 Kbp, 5 Kbp, and 10 Kbp upstream and downstream, and merging the regions. HGSVC SV calls that did not affect these reference bases was counted. This analysis was completed independently on insertions, deletions, and inversions so that only variants of the same type in the HGSVC calls and CHM1/CHM13 calls were compared. Approximately 30% of variants in this callset are not within 1 Kbp of CHM1/CHM13 variants (31%, 30%, and 39% in HG00514, HG00733, and NA19240, respectively) (**Table** **S4.2.4)**.

### 

### Intersection with 1000 Genomes Phase 3 Callsets

**(Contributed by: Audano, P.)**

We compared SVs to a previously published set by Sudmant [(Sudmant, Rausch, et al. 2015)](https://paperpile.com/c/ml0krT/R82H) as part of the 1000 Genomes phase 3 project. SVs were separated into insertions, deletions, and inversions. Copy-number variants (CNVs) other than deletions were removed. HGSVC SV calls were intersected with this set using a 50% reciprocal overlap for each variant type. For this analysis, variants were not filtered based on genomic regions.

### 

### Proximity of Integrated PacBio calls to 1000 Genomes SV

We counted HGSVC variants located distally from variants in the 1000 Genomes call set. Three sets of regions were computed taking the reference bases affected by each 1000 Genomes variant, adding 1 Kbp, 5 Kbp, and 10 Kbp upstream and downstream, and merging the regions. HGSVC SV calls that did not affect these reference bases was counted. This analysis was completed independently on insertions, deletions, and inversions so that only variants of the same type in the HGSVC calls and 1000 Genomes calls were compared. Approximately 80% of variants in this callset are not within 1 Kbp of 1000 Genomes variants (81% HG00514/HG00733, 80% NA19240).

### Orthogonal support of PacBio variants

We further annotated the unified PacBio callset where there was evidence of inheritance from parents, whether SVs had been previously discovered by SMRT-SV in hydatidiform mole genome sequencing projects [(M. J. P. Chaisson, Wilson, and Eichler 2015; Huddleston et al. 2016)](https://paperpile.com/c/ml0krT/BpDZ+s4fF), genotyping the SVs in the parent-child trios using Illumina sequence data (SMRT-genotyper) and by a graph-based approach (GRAPHITE).

**PacBio Inheritance (Contributed by: Chaisson, M):** To determine inheritance status of each SV, we counted read support for the unified PacBio callset using the PBRC method with the PacBio parental genomes. A minimum of one read was required to consider a variant present in a parent. We confirmed on average 8,717 of 9,288 (93.9%) homozygous calls as present in both parents, and an average of 18,332 of 19,162 (95.7%) heterozygous calls as present in one parent (**Table S4.2.7**) On average, 6.1% of homozygous calls were not confirmed as present in both parents and 4.3% of heterozygous calls not confirmed in at least one parent using the PBRC method **(Table S4.2.8)**.

When considering only events that lie outside of tandem repeats, 95.5% of homozygous calls are confirmed in both parents, and 98.5% of heterozygous calls can be confirmed in at least one parent.

**Bionano Genomics Support of Unified PacBio Calls (Contributed by: Chaisson, M):** We surveyed the concordance between the Bionano Genomics (BNG) SV calls and the unified PacBio callset with less stringent filtering (*f*BN = 0.25) to allow for a greater number of overlaps. Because a BNG call may overlap multiple PacBio SVs, we we selected the PacBio SV with the lowest *f*BN for an overlap. On average, 57.7% of PacBio SVs ≥ 1 Kbp were validated by a BNG SV.

**GRAPHITE Illumina Assessment of PacBio SVs (Contributed by: Marth, Lee, D):** Additional support for variants was detected using the GRAPHITE (<https://github.com/dillonl/graphite>) and which takes advantage of the alternate reference alleles provided in the integrated pacbio callset. In this approach, a local ‘variant graph’ in the region of one or more candidate variants is constructed by adding the candidate alleles as branches in a reference graph, initialized with the reference allele as its first branch. The primary sequencing reads are then aligned directly to this graph, using our implementation of the partial order alignment algorithm [(Lee, Grasso, and Sharlow 2002)](https://paperpile.com/c/ml0krT/9rc9). This is able to recover previously poorly mapped or unmapped reads from a sample by mapping against a path including the alternate allele carried by that individual. With this approach, one can ‘adjudicate’ a candidate variant i.e. confirm or reject its presence in a given sample based on the presence/absence of read mappings supporting the candidate alternate allele. Here we applied GRAPHITE to cross-validate candidate variants called from PacBio long-reads with Illumina data. We accomplished this by constructing a local variant graph in each region where a PacBio candidate variant, or multiple overlapping variants, were called, and re-mapping all Illumina reads (including unmapped mates) from the same sample to this graph. PacBio variants were annotated according to the number of Illumina reads they received. Variants for which the number of confirmatory Illumina reads exceeded a threshold (5) were annotated as ‘adjudicated’.

### Orthogonal sequencing approaches

#### **Bionano Genomics Analysis (Contributed by: Pang & Hastie):** Bionano Genomics optical mapping was used as an independent de novo assembly and SV calling approach that leverages extremely long (150 Kbp to 2 Mbp) single molecules for whole genome *de novo* assembly. *De novo* assemblies can be used to detect insertions, deletions, inversions, repeat expansions, translocations and transpositions. Bionano optical mapping produces assemblies that are sensitive to heterozygosity and are able to produce two alleles across the majority of the genome. Input data and *de novo* assembly metrics are shown in **table S4.2.10 and S4.2.11.**

Automated insertion and deletion calling has been used here and has identified approximately 5000 SVs for each enzyme, counting homozygous SVs twice and calling the same SV with each enzyme in most cases. Generating de novo assemblies and SV calls with two enzymes is valuable in improving confidence in SVs called with both enzyme and by covering assembly gaps in one enzyme with the complementary enzyme. In order to reduce the redundancy, SVs that are homozygous are collapsed and SVs from two different enzymes that have the same reference position and have similar sizes are merged. The resulting SVs are summarized in **table 4.2.12.**

#### **Bionano *de novo* assembly**: *De Novo* assembly was performed using Bionano’s custom assembler software. *De Novo* assemblies of all individuals in the CHS, PUR and YRI trios were performed with IrysSolve v2.3 software between February and March of 2016. For comprehensiveness, probands were re-assembled with IrysSolve v2.5.1 between October and December of 2016. Pair-wise comparison of all DNA molecules was done to create a layout overlap graph, which was then used to create the initial consensus genome maps. By re-aligning molecules to the genome maps *(P Value 10-11)* and by using only the best match molecules, a refinement step was done to refine the label positions on the genome maps and to remove chimeric joins. Next, during an extension step, the software aligned molecules to genome maps (*P Value 10-11*), and extended the maps based on the molecules aligning beyond the ends. Overlapping genome maps were then merged using a *P Value cutoff of 10-15.* These extension and merge steps were repeated five times before a final refinement was applied to “finish” all genome maps *(P Value 10-11)*. Two assemblies were constructed, one for each nickase.

Using IrysSolve v2.5.1, during the extension step, the software identifies clusters of molecules that aligned to genome maps with end alignment gaps of size > 30 Kbp (i.e. over 30 Kbp of one side of the molecules did not align), these molecules were split from the map and re-assembled. In addition, for both IrysSolve v2.3 and IrysSolve v2.5.1, the final refinement step searched for clusters of molecules aligned to genome maps with internal alignment gap of size < 50 Kbp, in which case, the genome maps were converted into two haplotype maps. The extend-and-split function is essential to identify large allelic differences and to assemble across loci with segmental duplications, whereas the refinement haplotype function can find smaller differences.

#### **Bionano structural variation:** SV was called based on the alignment profiles between the *de novo* assembled genome maps against the public human reference assembly GRCh38. We required an alignment cutoff of P-Value of 10-12 to identify the best aligned locations for any given match group within a genome map. SV calling was done for the Nt.BspQI and Nb.BssSI assemblies independently. Significant discrepancies in the distance or the number of unaligned labels between adjacent aligned labels (*outlier P-Value 3x10-3*) would indicate the presence of insertion and deletions. Genome maps whose alignments were in opposite orientations would indicate the presence of inversion breakpoints.

Finally, insertions and deletions captured by each of the single-enzyme assemblies (Nt.BspQI and Nb.BssSI) were compared and merged into a final SV call set. Insertions and deletions that were within 10 Kbp and with over 80 % reciprocal size similarity were merged together, and the innermost breakpoints were recorded as the merged variant breakpoints. To minimize false positives, we removed calls whose size was less than 500 bp, calls found by single nickase assembly but with a variant confidence score of < 0.5, or calls found by both nickases but with a confidence of < 0.3. No merge was performed for inversion breakpoints.

## 4.3 Variation detected by short-read sequencing

### SV Detection Methodologies on Illumina Genome Sequencing

An ensemble of Illumina structural variant callers (GenomeSTRiP, GATK, Pindel, MELT, DELLY, LUMPY, WhamG, dCGH, SVelter, Manta, forestSV, Holmes, TARDIS) and PacBio/Illumina hybrid methods (HySA & cloudSV (unpublished)) were analyzed to complement the PacBio only SV calls. Unlike previous efforts (phase III), the false discovery rate was not strictly controlled for the illumina callers allowing, for the first time, a broad comparison of SV callers, vs, a highly accurate PacBio callset.

**WHAMG (Contributed by: Kronenberg, Z)** :A sensitive and specific call set were generated for whamG [(Kronenberg et al. 2015)](https://paperpile.com/c/ml0krT/hEhL). In both datasets we jointly called the three trios and filtered with a modified version of “[filtWhamG.pl](https://github.com/zeeev/wham/blob/master/utils/filtWhamG.pl)”, removing events that have low support, low scoring “SVTYPE” classification and events with a high fraction of mate-pairs cross chromosome mapping. All calls across the three trios were merged into a single VCF with ["mergeSVcallers"](https://github.com/zeeev/mergeSVcallers) (50% reciprocal overlap and same “SVTYPE”). The merged calls were genotyped with SVTyper [(Chiang et al. 2015)](https://paperpile.com/c/ml0krT/Ivk9). For the specific set we applied “annotate\_hq.py” a script that requires no unknown genotypes, at least one heterozygous or homozygous alternative genotype call, a quality score above 100 and a median genotype quality above 100. The filtering script are in the whamG repository and the FTP. The SV counts and types for both the sensitive and specific set are listed in **Table S4.3.1**. The sensitive set was used for Illumina integration.

#### **LUMPY (Contributed by: Kronenberg, Z):** Split- and discordant- reads were extracted from the BAM files using a custom program (“filter.c”). Each trio was jointly called with Lumpy [(Layer et al. 2014)](https://paperpile.com/c/ml0krT/XhUu). BND/translocation events were filtered for all analyses. The trios were merged and genotyped using the same methodology outlined in the WhamG section, with the exception of the the whamG specific filtering procedure. The number and type of SVs obtained are enumerated **Table S4.3.1**. The sensitive callset set was used for Illumina integration.

#### **DELLY (Contributed by: Rausch, Korbel):** The germline SV calling workflow of Delly v.0.7.5 [(Rausch et al. 2012)](https://paperpile.com/c/ml0krT/sThC) was used to call large structural variants >500bp and small InDels in the size range of 15bp-500bp with default parameters in the deep coverage illumina paired-end data. Candidate germline SV sites were called sample by sample and all identified SV sites were concatenated into a single SV site list using delly merge with a reciprocal overlap of 0.5 and a maximum breakpoint offset of 500bp. All candidate SV sites were re-genotyped using delly on all trio samples. The output BCF files were merged into a single BCF file that contained all candidate SVs and their respective genotypes across the 9 trio samples. Uncertain genotypes with phred-scaled genotype quality below 20 were set to missing using VCFaid (<https://github.com/tobiasrausch/vcfaid>) and sites with an overall genotype missing rate >25% in all 9 samples were dropped. It was also required that at least one sample shows clear evidence for the SV with at least 20% of the reads confirming the alternative allele at a given SV locus.

To ensure high specificity we further filter the remaining SVs using a machine learning approach that uses site and genotype properties collected from svprops (<https://github.com/dellytools/svprops>). The SV classifier was trained and validated on a subset of likely true and false SVs that was derived from three sources. The first source were SVs that could be re-genotyped by delly in the 1000 Genomes low coverage data and were the IRS method (Sudmant, 2015) could assign a p-value. This set was then stratified into likely true SVs (p-value < 0.5) and likely false SVs (p-value >= 0.5). The second training source were likely false SVs that showed Mendel transmission errors. The third training source was a set of 100 randomly picked SV sites that we manually inspected using IGV (<http://software.broadinstitute.org/software/igv/>) and then classified as likely true or false SVs. All training sites were split into a training and validation set and machine learning parameters were picked to derive a final SV site list of an estimated FDR of 5% using the validation set. In addition, Delly v0.7.5 was used to discover and genotype a small set of confident complex SVs as described previously [(Sudmant et al. 2015)](https://paperpile.com/c/ml0krT/R82H). These complex SVs give rise to 2 overlapping paired-end clusters, which are then classified into simple inversions, inversion with an adjacent deletion and proximal inverted & non-inverted duplications using the delly dpe (double paired-end signatures) subcommand. For inversions, we in addition screened the 7 Kbp jumping libraries using delly v0.7.5 and also required double paired-end support for the filtered set.

Overall, Delly ascertained 11,823 deletions, 5,315 insertions, 173 tandem duplications, 16 inversions, 15 inversions with an adjacent deletion, 40 proximal non-inverted duplications and 49 proximal inverted duplications in the 9 trio samples. Independent of the machine learning training and validation set we also applied IRS on the 9 trio samples using CytoScan Arrays.This analysis yielded an estimated FDR of the final deletion set of 2.8%.

#### **dCGH (Contributed by: Nelson, Kronenberg, Eichler):** Illumina WGS sequence data were mapped to a repeat masked version of the human reference genome (GRCh38) using the mrsFAST sequence aligner as previously described [(Sudmant et al. 2010)](https://paperpile.com/c/ml0krT/kF1h) and variants were called using the digital comparative genomic hybridization (dCGH) method [(Sudmant, Mallick, et al. 2015)](https://paperpile.com/c/ml0krT/irX2). Deletion and duplications (>1 Kbp) were identified by comparison to a diversity panel of 17 Simons Genome Diversity Panel (SGDP) human genomes (**Table S4.3.2**).

#### **Genome STRiP** **(Contributed by: Handsaker, B)**: Genotyped copy number polymorphisms were called using Genome STRiP [(Handsaker et al. 2011)](https://paperpile.com/c/ml0krT/frsc) version r2.00.1691, which contains support for bwa alt-aware alignments, and using the reference metadata for GRCh38 (12Oct2016 version). Calls were made using both the Genome STRiP deletion pipeline and CNV pipeline and then filtered and merged as described below. Genome STRiP is a population based calling method, designed to run on hundreds or thousands of samples. The small cohort size (9 individuals) necessitated atypically stringent filtering and resulted in decreased sensitivity.

Deletion pipeline: Raw calls were generated using default settings with two exceptions:

* The parity correction threshold was reduced (-P depth.parityCorrectionThreshold: 0.1) to accommodate the small number of related samples.
* The default filtering on inbreeding coefficient in the genotyping step was disabled, to account for the small number of related samples.

Raw calls from the deletion pipeline were filtered using the following criteria:

* All default filters with the exception of INBREEDINGCOEFF (inbreeding coefficient).
* Sites were retained if they had GSCNQUAL >= 1 and GSCLUSTERSEP >= 3.
* Sites overlapping known VDJ recombination regions were excluded.
* Sites were included for the autosome and chromosome X only.

CNV pipeline: Raw calls were generated using default settings with two exceptions:

* The parity correction threshold was reduced (-P depth.parityCorrectionThreshold: 0.25) during discovery.
* The parity correction threshold was reduced (-P depth.parityCorrectionThreshold: 0.1) during subsequent genotyping and merging.

Raw calls from the CNV pipeline were filtered using the following criteria:

* Site call rate of 100%, with at least one sample called non-reference at 95% confidence.
* Multi-allelic sites predicted to have more than 3 observed alleles were excluded.
* Sites where any individual was called at copy number 9 or greater were excluded.
* Sites with deletion alleles that were shorter than 1500bp were excluded.
* Sites with only duplication alleles that were shorter than 4000bp were excluded.
* Sites were included for the autosome and chromosome X only.

Merging: Redundant calls were removed using the Genome STRiP Redundancy annotator if they had no discordant genotypes and at least 50% reciprocal overlap. When redundant calls originated from both the deletion and CNV pipeline, calls from the deletion pipeline (which are expected to have more accurate boundaries) were retained.

**QC**: To assess call quality, we evaluated all potential Mendelian violations (40) with calls in at least one child **(Table S4.3.3)**. This suggested an initial false discovery rate of 1.3%, predominantly from the CNV pipeline calls. To evaluate the impact of the small cohort, we re-genotyped the sites with potential Mendelian violations in a larger multi-ethnic cohort with 30x whole-genome sequencing data. Manual review suggested that the majority of the potential Mendelian violations were either due to genotyping error, often caused by incorrect boundary determination (Genome STRiP genotyping and boundary assessment are both better-powered in larger cohorts), or were due to the presence of a multiallelic CNV, which can cause the appearance of a Mendelian violation if the full allelic spectrum is not observed. One deletion (chr13:105744780-105747104) appears to be a mosaic de novo deletion in the CHS child. Adjusting for these sites suggests a true site-level FDR of 0.4%.

**Holmes (Contributed by: Mike Talkowski, Ryan Collins, Harrison Brand, Matt Stone, Joseph Glessner):**

**SV Discovery:** The computational analysis pipeline for structural variation (SV) discovery from 3.5kb long-insert whole genome sequencing (liWGS) data has been previously described [(Brand et al. 2015; Collins et al. 2017)](https://paperpile.com/c/ml0krT/BBlV+xmNw). The pipeline, Holmes, requires a sufficiently large cohort of liWGS libraries for simultaneous joint-calling of SV breakpoints and CNV intervals, so we supplemented the nine HGSVC liWGS libraries generated as described above with 91 independent individuals for which we have previously generated liWGS libraries with the same protocol and data processing procedures for an unrelated study (Collins et al. 2017). These 91 supplementary libraries were selected based on a balanced ratio of male and female subjects (50 males & 50 females in final analysis cohort, n=100) and on approximate library and sequencing quality (matched with HGSVC trio libraries on median and median absolute deviation of insert size, haploid physical coverage, chimera rate, and pairwise duplication rate) **(Rodriguez, O. and Bashir, A).** We subsequently performed SV analyses against the GRCh37 reference genome from these 100 individuals. All resulting SV calls were lifted over to hg38 coordinate space with the UCSC liftOver tool requiring a minimum of 50% of the original GRCh37 interval to remap to hg38 coordinates [(Kent 2002)](https://paperpile.com/c/ml0krT/eHiH). Based on previous validation experiments using PCR, targeted capture, comparison to chromosomal microarrays, and short-insert sequencing, we estimate an overall true positive rate to be approximately 0.894. Validations were not performed on this call set in the HGSVC samples (**Table S4.3.4**) **Results:** Using the median insert (~3.5kb) and the physical coverage obtained (mean: 163.9X, range: 144.6X-193.8X), we resolved a mean of 530 large SV per subject (1,270 unique SV sites resolved across all nine subjects) at the resolution of liWGS (SV size ≥ ~5kb). We also identified a mean of 53 incompletely resolved SV sites (IRS) and 143 low-confidence CNVs per subject; IRS and low-confidence CNVs were excluded from all subsequent analyses. The remaining high-confidence SV included canonical copy-number variants as well as balanced inversions, insertions, and cxSVs; collectively, complex and balanced SVs represented 25.5% (324/1,270) of all high-confidence SV sites identified in these nine subjects. Notably, we observed a median of 73 canonical inversion variants and nine complex inversions per subject at liWGS resolution. Principal component analysis of a genetic relatedness matrix constructed between these nine individuals showed that the first two principal components clearly cluster all three individuals from each trio. Transmission analysis yielded a mean of 97.1% inheritance across the SV call set. Consistent with previous findings, we detected a mean of 11 large complex SVs per subject in this analysis. The SVs shared between families are given in **Table S4.3.5**.

#### **VariationHunter (Contributed by: Hormozdiari): Deletions:** We applied an extension of VariationHunter for calling deletions (>50bp) in the three trios simultaneously [(Hormozdiari et al. 2009; Hormozdiari, Hajirasouliha, et al. 2011)](https://paperpile.com/c/ml0krT/di0e+JyBG). Previous versions relied primarily on discordant read mappings for calling deletions. The new version of VariationHunter considers discordant reads, read-depth and split reads signatures from Illumina read-pair data. Split reads are identified from soft-clip data which are remapped as two segments during alignment. For read-depth, we applied a likelihood ratio as metric to filter predicted deletions that resulted from low mapping quality. Finally, we used the tools SVtyper for genotyping all predicted SVs. A total of 4,581 deletions passed genotyping with the calculated IRS FDR of less than 5%.

#### **Forest SV (Contributed by: Sebat, J):** ForestSV [(Michaelson and Sebat 2012)](https://paperpile.com/c/ml0krT/ujiW) was used to call deletions and duplications in each individual with the default parameters. Adjacent SV calls in the same sample were stitched together if the gap between two calls of the same type was less than 10kb. We applied this rule recursively ensuring events consisting of multiple calls were merged. Next, calls were collapsed within families if the reciprocal overlap was at least 90% resulting in 68,944 deletions and 84,281 tandem duplications. Genotypes were then estimated for each SV across all samples using a machine learning genotyper, SV2(https://github.com/dantaki/SV2).

#### **Manta** **(Contributed by: Sebat, J):** For each trio Manta [(X. Chen et al. 2016)](https://paperpile.com/c/ml0krT/1vCU) was applied with default parameters to predict deletions, tandem duplications, inversions, and insertions. Variants were removed if the length was greater than 15Mb. Calls with at least 80% reciprocal overlap were merged while reporting the position with the smallest length, resulting in 14,840 deletions, 1,766 tandem duplications, 5,152 insertions, and 1,278 inversions with passing genotyping.

#### **SVelter** **(Contributed by Mills,R, Zhao, X) :** We applied SVelter [(Zhao et al. 2016)](https://paperpile.com/c/ml0krT/2fDp) individually to the aligned Illumina sequences for all 9 samples to call simple and complex structural variants >100bp in size. These samples were then merged into a single VCF formatted file. SVs are denoted as follows: DEL (deletion), INV (inversion), TANDUP (tandem duplication), DISDUP (dispersed duplication, includes insertion point in INFO field), DEL\_DUP (complex deletion with duplication), DEL\_INV (complex deletion with inversion), DUP\_INV (complex duplication with inversion), DEL\_DUP\_INV (complex deletion with duplication and inversion), and OTHER (unclassified). Over the three trios, SVelter reports 13573 (DEL), 506 (INV), 9319 (TANDUP), 3077 (DISDUP), 2165 (DEL\_DUP), 369 (DEL\_INV), 334 (DUP\_INV), and 310 (DEL\_DUP\_INV).

We applied an in-house long-read validation tool, VaPoR [(Zhao et al. 2017)](https://paperpile.com/c/ml0krT/WGRt) to SVelter predictions using aligned PacBio sequence reads for the trio children and were able to find individual read support for 82% of deletions and 60% of other SV types (both simple and complex).

**Novobreak (Contributed by: Chong, Chen):** NovoBreak v1.1.3 [(Chong et al. 2016)](https://paperpile.com/c/ml0krT/Z7TN) was used to detect SVs (>100bp) on the high coverage PCR-free Illumina sequencing data of the three trios (9 samples). NovoBreak is a tool initially designed to discover somatic structural variation in cancer genomes. It applied a *k*-mer targeted local assembly method to detect structural variants in single base resolution. To detect germline SVs in the trios and to meet the interfaces of novoBreak pipeline, each sample in the trios was treated as a ‘tumor’ and a mock ‘normal’ sample without mutations and only allowing sequencing errors (error rate = 0.005) was simulated using wgsim v0.2.3 in SAMtools package [(Heng Li et al. 2009)](https://paperpile.com/c/ml0krT/eBY8) from the human reference genome. For each sample, an initial call set was generated using the novoBreak pipeline under the default parameters. The initial call sets were further filtered using a customized script based on the alignment information (split reads, discordant read pairs and mapping qualities) and local assembly information around each SV event. For example, a minimum number of 5 reads for local assembly and a minimum quality of 50 (maximum is 60) were required to generate a high quality call set. In each trio, the unique calls in the child but not in either parent provided an estimation of the false positive rate (FDR). The estimated FDRs for YRI, CHS and PUR were 8.8%, 10.7% and 10.3%, respectively. Note that these FDRs were potentially over-estimated since the unique calls in the child may include *de novo* SVs.

Deletions, Duplications and Inversions longer than 100bp were reported for each sample. The numbers of deletions range from 1,883 to 2,717; duplicates from 14 to 30; inversions from 58 to 112. As expected, the African trio YRI had the largest number of SVs due to genetic diversity (~1.4 fold more than those found in CHS or PUR). The diversities of CHS and PUR were at the similar level. The same trend held for all three types of SVs. Overall, the deletion size distribution in each trio followed geometric distribution except that in all trios there was a peak around 300bp, resulting from the Alu element insertions.

**retroCNV (Contributed by: Gerstein):** An extended version of the retroCNV pipeline [(Schrider et al. 2013)](https://paperpile.com/c/ml0krT/f4Tq) was used to detect presence/absence polymorphism of processed pseudogenes in all 9 individuals independently. The retroCNV pipeline uses discordant read pairs to detect clusters of discordant read pairs evidencing the insertions of exonic or 3’UTR from parent genes. PCR-based validation studies have provided estimates for retroCNV FDRs at 0.1 and below. As an extension of previous version of the pipeline, solo 3’UTR retrotranspositions are also reported.

**MELT (Contributed by: Gardner, E, Devine, S): MELT Illumina call set:** MELT identifies MEIs using signatures of discordant read pairs and split reads that are enriched at sites containing non-reference (non-REF) Alu, L1, SVA, and HERV-K MEIs. PCR-based validation studies and simulations have provided estimates for MELT FDRs less than 5%. MELT detects a wide range of MEI-associated features, including target site duplications (TSDs), precise insertion junctions (where possible), 3’ transductions, 5’ inversions, size, orientation, interior mutations compared to consensus elements, family/subfamily status, and gene features affected (if any). MELT also calls genotypes for both non-REF MEIs and REF MEIs, thus providing a comprehensive set of polymorphic MEIs in a given genome (Gardner et al. 2017). MELT only identifies SVs that are precisely caused by mobile element insertion mechanisms, and it does not identify SVs that include MEIs are part of larger events. MELT identified a total of 4,271 MEIs in the three trios, including 3,417 Alus, 531 L1s, 306 SVAs, and 17 HERV-Ks. 99.35% of these calls were consistent with Mendelian inheritance in the trios.

#### **VariationHunter, Tardis (Contributed by: Hormozdiari, F):** We used the annotated *Alu* and L1 locations in the human reference genome (GRCh38) to guide in the prediction of MEIs. We applied an extension of VariationHunter for MEI prediction [(Hormozdiari, Alkan, et al. 2011)](https://paperpile.com/c/ml0krT/aTeQ) called Tardis, which in addition of discordant reads also considers split reads of the reads with soft-clipping. In the MEI call set one *de novo* Aluinsertion was predicted with confidence in NA12940 (locus: chr10:128034796-128035846) . The de novo Aluinsertion was validated as true de novo using PCR validation and PacBio reads.

Location of VH MEI call sets:

<http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/20160930_pre_ashg_calls/20161011_Tardis_MEI_Calls/>

### Development of an integrated Illumina callset

**(Contributed by: Mills, Zhao, Maholtra, Kronenberg)**

#### **Method Overview**: A total of 15 algorithms, i.e., Delly, dCGH, ForestSV, GenomeSTRiP, HOLMES ,Lumpy, Manta, MELT, novoBreak, Pindel, retroCNV, SVelter, Tardis, VariantHunter WhamG (**Table S4.3.6**) were applied to the parent-child trios from diverse population including Han Chinese, Puerto Rican and Yoruba, to discover SVs from Illumina short read sequences. Number of SVs reported by individual algorithm varies from 1000 to 23000 (**Table S4.3.7**), indicating large variants in sensitivity and false discovery rate across methods.

To derive a consensus SV set, an experimental based approach was developed where the breakpoints accuracy of individual algorithm were first estimated through comparisons against high-quality PacBio SVs that have single base breakpoint resolution, then clustered by overlaping their confident interval to derive consensus breakpoints with minimized confident intervals. (**Figure S4.3.7**)

**Estimate the breakpoint accuracy of each Illumina short-read based algorithm:** The breakpoint accuracy of each short-read based algorithm was estimated by comparing against high confident PacBio SVs by 50% reciprocal overlap (RO), i.e., for each SV predicted by short read technology, if there’s a PacBio SV that partially / fully covers the same genomic region, with the overlap between the two calls exceeds half of both calls, a ‘match’ will be assigned and the distance between the PacBio breakpoints and the short read ones will be recorded, with this distances from all matches to form a distribution. The 10% and 90% quantile of each distribution will be calculated and assigned as the confident interval (CI) of breakpoint accuracy for each algorithm. The distance between Illumina and PacBio breakpoints for 15 methods are shown in **Figure S4.3.8**.

**Cluster breakpoints with overlap CIs:** Breakpoints of SVs from different algorithms were firstly clustered into candidate groups if their CIs overlap with each other. In each group, if a minimized common CI can be defined by taking the right most of the left breakpoints and the left most of the right breakpoint, the consensus breakpoint will be assigned as the one that’s most frequently proposed by different algorithm. However, there are other situations where not all CIs overlap with each other. In this situation, a two-step approach will be adopted, where

1. all the CIs will firstly be stacked up and the number of CIs each breakpoint goes through is counted, with each peak and their right neighbor assigned as a CI.
2. Then the breakpoints within this cluster will be assigned to their nearest CIs to form the sub breakpoints clusters. And consensus breakpoints will be assigned in the same way as described above.

**Quality control of the integrated Illumina set:** An integrated set of 44505 unique SVs was collected from the pipeline described above, out of which 421 fell within either telomere or centromere of a chromosome and were labeled to be removed from further genotyping. There are also 594 SVs of over 1Mb in size that were labeled and removed in the primary quality control(QC) step. In the rest cohort, most of the integrated SVs were supported by only 1 algorithm (singleton, n=22903, perc=65%), over half (n=11778) of which were found to overlap with another merged SV with support from multiple algorithms (clusters). Singletons that overlap with clusters are assumed to be well represented by the clusters, so that they are labeled as ‘redundancy’. The rest SVs are labeled as ‘PASS’ (Table X5) and sent for downstream analysis such as genotyping across trios, breakpoint accuracy assessment through GRAPHITE [<https://github.com/dillonl/graphite>] and VaPoR [(Zhao et al. 2017)](https://paperpile.com/c/ml0krT/WGRt).

**Overview of the Integrated SV set:** The integration pipeline described above resulted in a non-redundant set of Illumina-specific SVs consisting of 31,766 unique SVs, including 19235 deletions, 2130 duplications, 4804 inversions, and 4888 mobile element insertions, with 56% detected by more than one approach, and 49% with orthogonal support from other technologies (**Table 4.3.8**).

We next assessed our integrated set by comparison with the high quality PacBio SV set and observed 49% events with a high degree (50%) of reciprocal overlap. We further identified ~1000 SVs per proband which appeared to be Illumina-specific. An in-depth investigation of these events showed signatures of structural rearrangements in the PacBio sequence data, suggesting their omission was due to methodological rather than technological limitations (**Figure S4.3.10A, Table S4.3.12**). We also investigated the sequence context around the breakpoints of each set and observed stark differences within repetitive regions between the technologies (**Figure S4.3.10B**).

Comparison between the Illumina integrated SV set with SVs detected by PacBio caller and Hybrid callers.

**Comparison with PacBio callset and discussion of Illumina only**: The deletion can be confirmed by raw PacBio read-support in an alignment-free approach using dot-plots. A 6.5 Kbp PacBio read (m140817\_221907\_42175\_c100689561270000001823145102281516\_s1\_p0/95354/0\_6656) supporting the deletion is shown in the dotplot below.

**Illumina Genotyping:** In order to assess the allele frequency of the SVs discovered in the three trios we used the Simons Genome Diversity Project (SGPD) samples. A panel of 260 PCR-free genomes at moderate coverage (**Figure S4.3.10**) represents 127 distinct populations from seven super populations [(Sudmant, Mallick, et al. 2015; Mallick et al. 2016)](https://paperpile.com/c/FLjKLc/MYq8z+KRx9). The geographic distribution of the individual samples can be seen in figure Illumina-genotyping2. The coverage of these genomes increases our power to genotype SVs discovered by PacBio and Illumina based SV callers.

For base pair resolved insertions and deletions, where both alleles are known, we applied SMRT-genotyper [(Huddleston et al. 2016)](https://paperpile.com/c/FLjKLc/mxhY5). For the illumina only SV calls we applied SVTyper and GenomeSTRiP. The original SGDP pair-end data was lifted from hg19 to GRch38 using bwa-kit “run-bwamem.”

### HySA Hybrid SV callset

**(Contributed by: Fan, Chen)**

HySA was applied to the Illumina and Pacbio reads from the three trios. It detected 10,231 large insertions (>50bp) and 10,054 large deletions (>50bp) with transmission error rates less than 0.3%. Unique to the HySA call sets were large insertions and small indels in STRs that were not detected in the consensus PacBio or Illumina call set **(Table S4.3.11)**.

A 95bp insertion uniquely identified by HySA illustrates challenges in identifying SVs in STR regions. While a small number of Illumina reads turn to have short clip end (upper panel), PacBio reads turn to have discordant gaps near the breakpoint (lower panel) due largely to errors in the long reads and the repetitive sequence context. Having both type of reads available made it much easier to make a clean and confident call.

In addition to SVs, HYSA was used to call indels (<=50 bp). HySA detected 42,073 small insertions (<=50bp) and 50,093 small deletions (<=50bp). The size of small indels starts from 11bp with trio transmission error rates < 2% (**Table S4.3.14**).

### 10x Chromium SV discovery

**(Contributed by Church, D. and Marks, P)**

Sequence data was analyzed using the Long Ranger v.2.1 analysis pipeline. Briefly, reads are aligned using Lariat, a wrapper around BWA that uses molecule information to adjust alignment locations and MAPQ, using the RFA methods [(Bishara et al. 2015)](https://paperpile.com/c/ml0krT/g8WU). This allows for more reads to be confidently mapped. In these cases, map quality scores are adjusted so that downstream analysis can take advantage of these reads. Single nucleotide variants (SNVs) and small indels are called using Freebayes (v0.9.21-7, default parameters (-0)).

Each read covering a heterozygous variant is re-aligned to a ~100bp segment of the reference sequence, with and without the alt allele applied, to determine whether the read gives clear support for one allele over the other. Mapping each read to a molecule via the GEM-specific barcode produces the yield of the set of alleles observed on each molecule. We model the likelihood of the per-molecule allele observations given a phasing configuration. The model follows [(Bansal et al. 2008)](https://paperpile.com/c/ml0krT/Z8gT) with additional terms to account for the small probability that a barcode carries two molecules from opposite haplotypes, and for the chance that an input variant is non-heterozygous. We search for the maximum likelihood phasing configuration by find near-optimal local configurations using beam-search over blocks of ~40 variants. Blocks are greedily joined to form a global solution, which we iteratively refine. The confidence of each phasing decision is the likelihood-ratio between optimal and next-best solutions. The phasing procedure solution implicitly produces a posterior distribution over haplotypes for each input molecule. Molecules covering >1 het are typically phased with very high confidence. We write this haplotype information to an auxiliary BAM tag on each read of a confidently phased molecule, which is used in downstream haplotype aware SV calling.

Input variants determined to be non-heterozygous are switched to HOM REF or HOM ALT.Variant calls are then emitted as a VCF file.

Structural variants (SVs) greater than 30Kb are called using a novel algorithm that detects unexpected barcode overlaps. The algorithm computes the likelihood of the observed reads under various types of structural variants (or the reference configuration) in order to pick the event that is most consistent with the expected molecule size distribution as well as with the observed phasing and barcode overlap patterns. Breakpoint-spanning read-pairs and split reads are used, when present, to further refine the breakpoints. Since the algorithm relies on barcode overlaps between the breakpoints, events involving the ends of chromosomes, such as terminal deletions, are not called. A curated list of blacklist regions is used to filter calls in regions of the genome that are likely to be incorrect in the reference assembly or lead to mapping artifacts, such as centromeres and very large segmental duplications.

To call deletions 50bp-30kb in length, we start by searching for haplotype specific coverage dropouts and discordant read pair clusters. This initial set of candidates is refined through a likelihood ratio test that accounts for phasing. Additionally, events shorter than 5kb are reassembled with a local colored De Bruijn graph method that tracks the per-haplotype barcode support for each unbranched path. Maximal unambiguous paths are found for each haplotype, and the resulting sequences are realigned to the reference to generated the reported breakpoints. SVs of all size ranges are output using the VCF format. Events that passed all filtering steps are marked with “PASS” in the filters field.

# 5. Integration of Illumina, PacBio, and Bionano callsets

We assessed which regions of the genome are accessible to one sequencing technology and not another for the Illumina and PacBio read sets.

## 5.1 Comparison of integrated Illumina and integrated PacBio callsets

Each integrated Illumina call was compared against the integrated PacBio calls, by searching for the SV with the greatest size overlap within +/- 1 Kbp of of the breakpoints of each Illumina call (**Table S5.1.1).**

### 

### Genomic coverage of Illumina and PacBio

First, physical coverage of Illumina reads with mapping quality ≥ 15 (generous) and raw coverage of PacBio reads with mapping quality ≥ 30 (conservative) were calculated in 100bp bins from each child genome. To mitigate effects of deletions obscuring which regions are not covered by a technology, coverage was combined from each child for the two technologies. Filtering heterochromatic regions, chromosome Y, and any bin with 5 reads or fewer covering it, we found 749 Kbp (0.028% of the euchromatic genome) basepairs covered by PacBio and not Illumina, and 2.94 Mb (0.1% of the euchromatic genome) of sequence conversely covered by Illumina and not PacBio. These platform-unique sequences are enriched for segmental duplications; 82.0% of the PacBio only regions were segmental duplication, and 83.9% of the Illumina-only regions were segmental duplication. Thus we expect the majority of the calls to be accessible to both platforms.

### Filtering integrated Illumina callset

We then searched for concordance between calls from the integrated Illumina (IL), integrated PacBio (PB), and Bionano (BN) callsets. We began by developing filtered SV callsets. The PB calls were previously filtered by raw-read support. To develop a filtered IL callset, each call was compared to a superset of variation from the PB integrated callset including: the set of unfiltered SV calls from the haplotype local assemblies in Phased-SV, SVs detected through individual PacBio reads, the coverage depth of the PacBio alignments, and the BN SV calls. For all comparisons except for the Bionano and read-depth comparison, the following steps were performed to detect the concordance between a query SV and a dataset of SVs. First, all target SVs from the dataset within 1 Kbp of the breakpoints of the query SV are collected. Each SV from this set is compared against the query by taking the ratio of the shorter of the query and target SV to the length of the larger of the query and target SV, and the concordance is defined as the maximum of all of all of these ratios. A similar test was used for comparison to BN calls, however the SV position interval was used in place of breakpoint boundaries. Each Illumina call was considered concordant if at least one of the conditions held true:

1. The maximal concordance of a PB or BN target SV is at least 0.5.
2. At least 3 reads have at least 0.7 concordance with the query SV.

The greater stringency for raw read overlap is to reduce false positive concordance caused by spurious short indels. The average number of integrated Illumina SV calls that had a concordant SV was HG00514: 60.2% (7514/12472) (del) 64.0% 3964/6198 (ins), HG00733 62.6% 7688/12284 (del) 66.4% 4193/6312 (ins), and NA19240 62.1% (8674/13974) (del), and 66.2% (4652/7028). The number and fraction of integrated Illumina calls for each child is shown below.

We compared the validation rate by the number of methods supporting each call. For calls supported by a single Illumina algorithm, between 42.5-45.2% of deletion calls were supported by additional technologies, and 62.3-64.9% of insertion singleton calls were supported.

We looked for evidence of the integrated Illumina calls that were not found in the integrated PacBio callset within the original PacBio reads or the PhasedSV local assembly unfiltered calls.

### 

### Filtering Bionano Genomics calls

To generate a filtered callset of the Bionano deletion calls, each call was compared with the SVs detected from local assemblies in Phased-SV, and read depth, SVs detected in raw PacBio reads, and the PacBio read depth. Because a genomic interval and estimated variant length are given for each BN variant, the read-depth is calculated by scanning the genomic interval and taking the average coverage within a sliding window of the length of the structural variant. The candidate coverage is taken as the minimum coverage. A call is considered validated if the minimal average coverage is less than 30 (compared to the 40X average sequencing depth of each sample). An example of a validated homozygous variant, validated heterozygous variant, and region in which a deletion variant was predicted but no suitable coverage interval was found is shown below. The corresponding variant discovered in the PacBio integrated set is shown in red.

In total this finds concordance rate is HG00514 94.1% (871/926), HG00733: 94.1% 829/881, NA19240: 97.8% (984/1006). Because it was not possible to find concordance with insertion calls and read-depth, all insertion calls less than 15 Mbp were considered concordant.

The number of BionanoGenomics calls overlapping segmental duplications are:

HG00514: N=268/1656, total bp = 9608546/26129625, 6.47X increase

HG00733: N=276/1698, total bp= 9723721/26693605, 6.41X increase

NA19240: N=283/1740 total bp= 11830199/30477301, 6.8X increase.

Increase is computed using the fraction of the genome annotated as segmental duplication of 0.056.

## 5.2 Integration of Illumina, PacBio, and Bionano callsets

The filtered integrated Illumina and Bionano calls were integrated with the integrated PacBio SV callset. The Illumina and Bionano callsets were queried against the integrated PacBio callset in a similar manner as the query against the validation sets, and the Illumina set was queried against the Bionano callset for concordant matches. The variants for which a concordant match was found were recorded as shared between pairs of datasets. To construct an integrated dataset, we merged (1) calls unique to each filtered dataset, (2) calls shared between pairs of datasets, and (3) calls shared across all three datasets. For cases 2 and 3, the PacBio variant was used when available (e.g. shared Illumina/PacBio, and shared Bionano/Pacbio, or shared across all three), and the Illumina variant when shared between Bionano and Illumina (**Table 2**)

The gain in sensitivity for PB-SV only calls was largely in SVs less than 2 Kbp (**Table S5.2.1**)

### Specificity of integrated Illumina dataset

The calls by individual algorithms were evaluated for concordance with the PacBio and Bionano Genomics integrated datasets, which may be viewed as an estimate of the accuracy of each method. Because there are calls missed by the union of the integrated PacBio and Bionano Genomics callsets, this is a lower bound on the estimate. The average concordance for deletion calls ranged between 24.8 and 91.7% with a median of 79.2%, and for insertion calls ranged between 4.2% and 83.8% (Supplemental Table IndividualCallerConcordance). The insertion statistic is affected by comparing the coordinates of the duplication calls, which are defined by the interval of the source sequence, to the insertion calls in the integrated PacBio callset and the Bionano Genomics calls, which are defined at the insertion sites. This has an effect of reducing the concordance of methods that produce duplication calls through discordant read depth measurement such as dCGH and GenomeStrip.

We then sought to assess the concordance of callsets formed from the union or intersection of multiple methods. Because of the computational burden of running many algorithms, we considered combinations of up to four algorithms, and tested four separate conditions: accept all calls from the union of two algorithm callsets, all calls from the union of three callsets, calls from at least two of three algorithms, and calls from at least two of four algorithms. The first two conditions represent scenarios where one is targeting maximal sensitivity, while the latter two are targeting specificity. The union of methods naturally increases the FDR while increasing the sensitivity, and the intersection of methods decreases FDR however requiring a majority (e.g. two of three methods) rather than total consensus can increase sensitivity, both with the expected trade off between increased sensitivity and FDR. Using nonconcordance rate (NCR) (1- concordance) as a proxy for FDR. Considering NA19240, there was no union of any two methods with an NCR less than 10%, however when applying three algorithms and accepting any call made by at least two, up to an 18.9% gain in number of calls found to be concordant can be obtained (lumpy, Manta, SVelter) (**Supplemental Table 17**).

### 

### Analysis of combinations of Illumina callers.

Due to computational burden, most large scale studies merge calls from a relatively small (2-3) number of algorithms. We investigated the trade off between using the union of two methods, or requiring two of three methods, which has the promise of higher sensitivity with similar or greater specificity. For all combinations of pairs of methods, or triplets, we extracted IL-SV calls with the corresponding caller support (e.g. two of the three methods in a triplet), and compared the calls to the entire integrated callset. We computed the number of detected and not detected calls and bases.

### Comparison of read depth and *de novo* assembly methods

Because a substantial fraction of human genetic variation occurs in regions of segmental duplication [(Bailey and Eichler 2006)](https://paperpile.com/c/ml0krT/eDYN), and segmental duplications are often missing from *de novo* assemblies [(M. J. P. Chaisson, Wilson, and Eichler 2015)](https://paperpile.com/c/ml0krT/BpDZ), we compared the variation detected in regions of segmental duplication through read depth to the segmental duplications that were resolved in the MS-PAC *de novo* assemblies. Because both of the Phased-SV and MS-PAC haplotype partitioned assemblies are biased by the reference, we excluded these assemblies from analysis. The assemblies were mapped to GRCh38 using blasr ([www.github.com/mchaisso/blasr](http://www.github.com/mchaisso/blasr) ) with parameters “-alignContigs -nproc 16 -minMapQV 30”. The UCSC genomicSuperDups annotations for GRCh38 were merged into nonredundant regions, annotating the percent identity of each merged region as the highest annotated identity from all corresponding composite duplications. The boundaries of the alignments of contigs were compared to the boundaries of the merged segmental duplications. All duplications that were covered by at least 50 Kbp on both the 5’ and 3’ end were considered resolved.

The dCGH and GenomeStrip methods both detect copy number variation using read depth, and are sensitive to copy number changes in highly duplicated regions. Between 91.4% and 99.5% of copy number variation detected by dCGH in segmental duplications were in duplications not resolved by *de novo* assembly. Similarly, using Genome STRiP, between 71.3 and 74.5% of copy number variation was in regions of segmental duplication not resolved by *de novo* assembly. These regions are gene rich; 49.2-60.7% of the dCGH and 14.3-15.9% Genome STRiP copy number variants not resolved by assembly overlapped exons.

## 

## 5.3 Properties of the unified integrated calls

We investigated the functional annotation of sequences that were inserted or deleted in structural variation in the pan-technology integrated calls. Variants were intersected with exons and UTR annotations from refGene, and analyzed with Variant Effect Predictor:

--total\_length --canonical --pubmed --sift b --fork 1 --max\_af --regulatory --symbol --vcf --offline --cache --assembly GRCh38 --synonyms chr\_synonyms.txt -i calls.bed --appris --output\_file calls.vep.bed --dir\_cache vep --af --nearest symbol --force\_overwrite

### 

### Genes affecting SVs

Between 77 and 92 genes had at least one exon or UTR affected by a deletion SV, and between 316 and 322 exons affected by insertion SVs. Between 37 and 41% of exon-overlapping deletion SVs were detected in the IL-SV dataset, and between 58 and 70% were detected in the PB-SV dataset. The average length of the PB-SV exclusive exonic SV was between 935 and 1,429 bp, whereas the average length of the IL-SV exon overlapping callset was between 13.2 and 24.7 Kbp. This discrepancy indicates a loss of sensitivity in the IL-SV callset for smaller-scale deletion events, consistent with non-exonic SVs.

### SV overlapping noncoding functional elements

In addition to detecting exons that are disrupted by SV, we looked for potentially functional noncoding elements that overlap SV. We defined 2,857,792 conserved noncoding elements totalling 121.9Mbp from PhastCons Most Conserved Elements for the hg38-based 100-way multiple alignment from UCSC. Elements within 25bp were merged, and elements overlapping coding elements and UTRs were removed. To avoid spurious alignments in hyper variable regions, alignments were required in a net to chimp in addition to mouse, dog, or horse. We also defined 1,070,532 transcription factor binding sites by lifting over from the USSC HMR Conserved Transcription Factor Binding Sites (<http://genome.ucsc.edu/cgi-bin/hgTables?hgta_doMainPage=1&hgta_group=regulation&hgta_track=tfbsConsSites&hgta_table=tfbsConsSites>). Nearly half the SV that overlap a conserved element (50.5% ) are in large (>2 Kbp) SVs (average 27.8 Kbp). The SV calls that overlapped TF binding sites were similarly larger than the average call (39.8 Kbp) The calls unique to PB-SV (10.7% of calls) were on average 335 bps. While the number of TFBS deleting events unique to PB-SV (missed by IL-SV) was small (16 events/sample), the average size was small; on average 12.7 (~75%) sites per genome were less than 2 Kbp.

# 6. Multiscale discovery of fixed and polymorphic inversions

## 6.1 Large-scale inversion detection using Strand-seq

**(Sanders)**

In order to explore larger copy-neutral as well as more complex structural variants that are typically excluded in SV studies (often due to confounding genome architectural features and methodological limitations), we incorporated strand-resolved sequencing data generated using **Strand-seq.** Strand-seq is a single cell sequencing technology that involves sequencing reads from individual DNA strands following bromodeoxyuridine (BrdU) incorporation into dividing cells [(Falconer et al. 2012; Sanders et al. 2017)](https://paperpile.com/c/ml0krT/rDMh+6pJd). In contrast to conventional MPS, Strand-seq maintains DNA strand directionality, which is used to directly separate sequence data derived from each chromosomal homologue. These data preserve the long-range structural information of individual homologues, which can be used to scaffold phasing data to build haplotypes (as described above) and allows copy-neutral structural variants to be directly visualized [(Porubský et al. 2016; Sanders et al. 2016)](https://paperpile.com/c/ml0krT/TJmR+ZqD6). In contrast to other sequencing techniques, this technology identifies inversions based on the directionality of reads contained within the inverted locus, rather than from reads spanning inversion breakpoints. This allows even those variants embedded within large, highly identical low copy repeats (that are inaccessible by regular short or long DNA read sequencing approaches [(Conrad et al. 2010)](https://paperpile.com/c/ml0krT/frK2X) to be located. Accordingly, Strand-seq has recently emerged as a new method to discover inversions across an extensive length scale from a few kilobases up to several Megabases in size ([Sanders et al. 2016](https://paperpile.com/c/ml0krT/TJmR+ZqD6)).

To map inversions in the family trios, we generated altogether 1064 Strand-seq libraries, yielding a cumulative sequencing depth ranging from 3.6-7.7x per genome, and covering up to 80% of all mappable bases (**Figure S6.1.1 A**). To remove potential inter-cell inconsistencies and increase coverage for variant detection, composite files were generated for each individual, as previously described ([Sanders et al. 2016](https://paperpile.com/c/ml0krT/TJmR+ZqD6)) (**Figure S6.1.1B, C**). A composite file represents merged sequence data from multiple Strand-seq libraries with strand directionality preserved. Based on alignment to the reference genome, each read in the composite file was designated as being in the same orientation as the reference (‘reference’), or the opposite orientation to the reference assembly (‘non-reference’). This allows regions of inverted orientation to be identified and genotyped based on the proportion of non-reference reads at the locus, and a read ratio was calculated for as the number of reads in the non-reference orientation divided by total number of reads (**Figure S6.1.1D**). Regions with higher read ratios contain a greater proportion of reads supporting an inverted allele, and thus this value can be used as proxy for the level of support for an inversion.

Using a sliding window approach (window size = 250 reads) local read ratios (i.e. proportion of non-reference reads) were calculated for each window in the composite file and putative structural variants were located as chromosomal regions that exhibited a segmental change in strand orientation and contained > 15% non-reference reads (**Figure S6.1.2A**). This located up to 354 genomic loci per individual that contained a significant portion of reads in a non-reference orientation, suggestive of an inversion (**Table S6.1.1**). We found a cluster of these regions had a read ratio > 0.80 (which supports a homozygous inversion), however there was a continuous distribution of regions with a ratio < 0.75 (**Figure S6.1.2B**), which suggested that a subset of loci identified in the composite files were non-diploid. Each locus was genotyped as being either homozygous reference, heterozygous inverted, or homozygous inverted using a Fisher Exact test that allowed for a 2% level of background in the homozygous states and required a minimum of 50 reads to genotype (**Fig. S6.1.2C**). This resulted in 187 - 208 (74.2%) heterozygous and 63 - 77 (25.8%) homozygous loci per individual (**Table S6.1.2**). The inversions showed a continuous size distribution that ranged between 448 bp - 3.7 Mb in length, with a mean of 208 kb (**Fig. S6.1.2D**). This *discovery set* of Strand-seq inversions together represent simple inversions, complex variants and reference assembly errors (as discussed below). For example, we identified 35 inversions that were homozygous in all nine individuals that likely represent reference assembly misorients, and were thus removed from the downstream analyses (**Table 6.1.3**).

### 

### 6.2 Classifying Strand-seq inversions using orthogonal phase data:

The Strand-seq discovery set of inversions identified from the composite files were further classified by integrating orthogonal phasing data. We used a trio-aware PacBio phased vcf file that was assembled using the WhatsHap pipeline [(Patterson et al. 2015)](https://paperpile.com/c/ml0krT/fIdJ+w0aL) independent of Strand-seq haplotype data. We tagged all possible reads in the single cell libraries by identifying heterozygous SNVs captured in a Strand-seq sequencing read and assigned the fragment to either haplotype 1 (H1) or haplotype 2 (H2) based on agreement with the PacBio phasing file (**Table S6.2.1**). Haplotagged composite files were then regenerated using the phased Strand-seq data and the haplotype structure at each locus was assessed in a strand-aware fashion - meaning we considered the ratio of phased H1:H2 reads in the reference orientation independent of the the ratio of phased H1:H2 reads in the ‘non-reference’ orientation. This allowed us to assign a haplotype ratio for each strand of the locus, calculated as the proportion of reads with an H1 phase divided by the total number of phased reads. By combining read ratios with haplotype ratios, we defined distinct signatures to distinguish simple homozygous, heterozygous and complex heterozygous (e.g. inverted duplications) inversions identified in the Strand-seq discovery set.

We identified simple homozygous inversions (i.e. having a non-reference orientation for both homologues, and without an accompanying copy number change) as loci that were genotyped as homozygous and showed a haplotype ratio ~ 0.5 ( > 0.25 and < 0.75) on the non-reference strand, where a minimum of ten phased reads were required (**Figure S6.2.1**). Across all individuals we found 147 loci with this signature, of which 55 were unique. These simple homozygous inversions ranged in size from 2.2 kb - 3.9 Mb (median length = 142 kb). The median read ratio was 0.95 (range: 0.79 - 1.0), and the median haplotype ratio for the non-reference reads was 0.49, with an interquartile range of 0.10.

We then identified simple heterozygous inversions (i.e. loci showing non-reference orientation for a single homologue and without a copy number change) as loci genotyped as heterozygous and represented by different haplotypes on the reference and non-reference strands (**Figure S6.2.2**). These loci had a haplotype ratio > 0.75 on the reference strand coupled with a haplotype ratio < 0.25 on the non-reference strand, or *vice versa*, where a minimum of ten phased reads were required on each. This located 221 simple heterozygous inversions in all individuals, representing 102 unique loci. The lengths ranged between 2.9kb - 3.9 Mb, and a median of 136 kb. The median read ratio for these events was 0.48 (range: 0.21 - 0.73) and the median haplotype ratio was 0.23 for the non-reference reads and 0.79 for the reference reads. The haplotype information was subsequently used to directly phase these inversions, where a non-reference haplotype ratio > 0.75 marked an inversion on haplotype 1 (“1/0”), and a ratio < 0.25 marked an inversion on haplotype 2 (“0/1”). This located 106 heterozygous inversions to haplotype 1, and 115 to haplotype 2.

Next, complex heterozygous inversions were identified as loci showing non-reference orientation and represented by phase data suggestive of a copy number change. These were located as events with a heterozygous genotype where a single haplotype was found on the non-reference strand (i.e. haplotype ratio < 0.25 or > 0.75) and both haplotypes were found on the reference strand (showing a haplotype ratio of ~ 0.5). This located 207 complex inversions, of which 97 were unique. The read ratio (median = 0.31; range = 0.18 - 0.79) of these inversions was slightly lower than the simple heterozygous, suggestive of a copy number increase on the reference strand (**Figure 6.2.3**). They ranged in length from 2.1 kb - 1.3 Mb, with a median size of 72.5 kb. Taken together, this analysis located 129 simple and 98 complex inversions in the Strand-seq that have orthogonal support from PacBio phase data.

## 6.3 Integrating inversion calls across orthogonal platforms:

Although Strand-seq is uniquely positioned to locate large (kilobase-scale) inversions, even if embedded within highly repetitive segmental duplications, due to the sparsity of single cell sequence data, the ability to locate smaller variants is limited using this method alone. For instance, the interquartile range of the Strand-seq inversion discovery set was 17.5 - 189.2 kb in length. To locate inversions of a complete size spectrum, we additionally generated inversion callsets from sequence data derived from PacBio, Illumina, BioNano and liWGS (corresponding to jumping libraries of 3.5kb and 7kb insert lengths) technologies (**Figure S6.3.1**). For further details on the discovery sets refer to the ‘Full-spectrum variation detection’ section.

To unify all inversion predictions into a single integrated callset, we first performed an intersection test that measured the level of reciprocal overlap between the different technologies. For all intersecting inversions, we calculated the level of overlap by dividing the number of intersecting base pairs by the total predicted inversion length. We then filtered pairs that showed > 50% reciprocal overlap to identify inversions supported by two independent technologies (**Figure S6.3.2**).

The results of this test yielded a total of 1296 supported inversions (137 non-redundant), with 37 - 96 per individual (**Table S6.3.1**). Approximately half (45.9%) were recovered for each parent because PacBio only produced variant calls for the probands. Of the total supported inversions, 533 came from Illumina, 335 from liWGS libraries, 172 from PacBio, 147 from Strand-seq and 106 from BioNano callsets (**Figure S6.3.3A**). While BioNano contributed the fewest number of inversions to the total, 39.1% of the initial callset was supported, whereas only 13.6% of the initial Illumina callset (that contributed the most) was supported (**Figure S6.3.3B**). Notably, although a cutoff of 50% reciprocal overlap was used, the majority of passing events showed near complete agreement (first quartile of percent overlap was 89.5%). The majority (64.9%) of which were represented by the minimum of two technologies, with only 5.5% represented by all (**Figure S6.3.3C**). We found Illumina and PacBio showed the most agreement at a size scale < 2 kb, the jumping libraries overlapped with Illumina within the size range 2 - 4 kb, and with PacBio and/or Strand-seq between 4 - 25 kb. The highest agreement between technologies was seen between 5 - 50kb size lengths, with the larger events (> 50kb) dominated by Strand-seq and BioNano calls (**Figure S6.3.2,** and **Figure S6.3.3D**). Indeed, because the interquartile range of intersecting inversions was 1.9 - 18.7 kb (median 3.5 kb), the majority of inversions predicted by Strand-seq were poorly represented by this test (**Figure S6.3.3B**). This points to the distinct strategy used by the method to locate inversions that cannot be captured by orthogonal technologies.

To additionally recover inversions not captured in the intersection test, we next performed a genotyping test using the Strand-seq composite files. We performed a Fisher Exact Test on all inversion callsets to test whether support was seen in the Strand-seq composite files, where a minimum of 25 reads were required to genotype the locus. From the discovery callsets we found a striking number of predicted inversions with a ‘reference’ genotype in the Strand-seq files, where between 36.6 - 76.1% of genotyped loci (i.e. containing > 25 reads in the composite file) were unsupported (**Table S6.3.2**). Nevertheless, this test was able to add orthogonal support to a total of 175 non-redundant predicted inversions, 84 of which were not present in the interception test results. Moreover, when we repeated this genotyping test using only inversions passing the interception test, we found the number of inversions supported by the composite files increased up to 36.4% (**Table S6.3.3**). For example, 89 (36.6%) of the inversions predicted in the PacBio callset were not supported by Strand-seq in the initial test, and this dropped to a single (0.6%) unsupported event when the filtered set was genotyped. This analysis illustrates that both the Strand-seq re-genotyping and interception analyses are useful validation tools for inversion discovery.

From these analyses a final unified inversion callset was generated that represented variants supported by at least two orthogonal technologies. To obtain this, we compiled all supported inversions from the i) interception test, ii) Strand-seq genotyping test, and iii) Strand-seq inversions supported by PacBio phase data (**Figure S6.3.4**). The three supported inversion lists were merged into a non-redundant inversion set for each individual. Any pericentric events (e.g. large > 6Mb inversion predicted by Illumina and BioNano), or those with a tandem repeat fraction > 90% were removed, and breakpoint ranges were determined for the remaining inversions. For this, the outer breakpoint of the inverted region (InvR) was defined as the outermost start and end positions of each non-redundant event found between all three lists. For all events containing > 1 inversion prediction within the locus (e.g. those located in the interception test), the inner breakpoints (innerBP\_start and innerBP\_end) were defined as the consensus region represented by at least half of the overlapping predictions. If the inversion was listed in only one of the support lists (e.g. those discovered by Strand-seq and supported with phase data) the innerBPs were set to match the InvR. A consensus genotype was then assigned to the inversion by taking the majority genotype call found for all discovery sets. If no majority call was possible (e.g. an equal number of heterozygous and homozygous genotypes were listed for all technologies) the genotype was listed as ‘ambiguous’. Finally, we classified the unified list into simple (**Table S6.3.4**) and complex (**Table S6.3.5**) inversions, by performing a copy number analysis. Here, we called the copy number state using GenomeStrip on the high-depth Illumina data to assign a CN value for each inversion.

## 

## 6.4 Resolving Inversion Sequence Errors in the Human Reference Genome

**(Contributors: Cantsilieris, Eichler)**

During the Strand-seq analysis, we observed 51 regions where the majority of individuals predicted a configuration different from the current human reference genome. We filtered for calls mapping to centromeric and satellite DNA sequences (n=21) and observed that a large fraction of the remaining calls (43%) were almost completely contained within segmental duplications and therefore interpretation was difficult. We specifically focused on regions where at least seven of the nine genomes assessed by strand-seq predicted homozygous inversions and where there was evidence of unique intervening sequence. Such regions suggest that the configuration of the reference genome either represents a minor allele or that the reference (GRCh38) is in error. We identified 17 such regions ranging in size from 3.4 Kbp up to 2.9 Mbp in size. We selected large-insert clones from human hydatidiform mole source (CHORI-17) and sequence and assembled each region using Canu [(Koren et al. 2017)](https://paperpile.com/c/ml0krT/4JVv) followed by consensus sequence calling using Quiver [(Chin et al. 2013)](https://paperpile.com/c/ml0krT/qA08) as previously described [(Huddleston et al. 2014)](https://paperpile.com/c/ml0krT/ofL7). We identified 11 regions that could be spanned by a single BAC clone. High quality sequencing validated 9/11 inversion events with two regions (chr22:21,442,966-21,496,091 and chr17: 43,234,311 - 43,323,702) completed embedded within tandem duplications that could not be resolved at the level of clone based assembly. Of these 9 regions, 6 show evidence of an inverted orientation in all genomes analyzed indicative of misassembly in the GRCh38 reference. For the remaining 3 events, strand seq analysis in one individual supports the existence of the reference orientation suggesting it reflects the minor allele. We next selected 10-20 Kbp of flanking sequence surrounding each of the 9 inversion events and identified that they are enriched for common repeat sequences (average repeat content 67%). Notably, 5 of these events are flanked by inverted LINE/L1 repeat sequences.

Next, we identified 3 regions >1 Mbp flanked by large highly identical segmental duplication blocks that showed evidence of inverted orientation >7 genomes. We selected two of these regions for a more detailed analysis by constructing alternate reference haplotypes using CHM1 BAC clones. At chromosome 16p12 we generated a ~1.8 Mbp alternate reference haplotype corresponding to chr16:21288212-22746306 in the GRCh38 reference assembly. The inversion maps to a previously identified assembly error in the human reference genome [(Antonacci et al. 2010)](https://paperpile.com/c/ml0krT/lapQ). Of the 9 human genomes analyzed here, a single individual is heterozygous for the event suggesting the reference may represent a very rare haplotype structure. At chromosome 2q13 we generated a ~843 Kbp alternate reference haplotype spanning two large duplication blocks of ~358 Kbp and >99% sequence identity. The duplication blocks map in inverted orientation and would likely predispose this region to inversion. Sequence analysis shows that the CHM1 haplotype maps in direct orientation consistent with the GRCh38 reference assembly.

# 

# 7. Biological context

## 7.1 Population genetics of SVs

### Population genetics of PacBio Integrated callset

#### **Delly2 Genotyping:** Deletion structural variants from the PB-INT callsets were genotyped using Delly2. The original callsets, HG00514 #n = 10,981, HG00733 #n = 11,477, and NA19240 #n = 13,066, were concatenated into a single redundant site list (#n = 27,766). 2501 samples from the 1000-genomes project, Phase 3 were genotyped, as well as the high coverage Illumina from each trio individual as a control Sites were filtered when genotype quality < 20 from the trios, and < 5 for the phase 3 individuals. Sites were additionally filtered if the allele count from the trios was 0, or if fewer than 25% of individuals were genotyped from the Phase 3 study. The final number of sites genotyped was 8,167.

The Mendelian error within the trios serves as a control for genotyping accuracy, and was found to be less than 0.7%.

The intensity rank sum FDR for the trios was 5.3% (#n = 451), and 1kGP: 2.7% (#n = 73).

The genotyped samples cluster via PCR by continental population, as expected.

The copy number discordance between the genotypes and the phase-3 callset was: 7.8% (non-reference. 14.5%).

There were 3,893 sites with minor allele frequency > 0.10. There were no SVs genotyped in this set that showed extreme population stratification (FST > 0.85), however 269 sites had FST ≥ 0.20 of which 79 mapped within genes, and 2 overlapped exons (MUC2 and OR52E8). 60% (163/269) were novel and not previously observed in the 1KG phase 3 callset including intronic deletions within TRPV5, TMC5 and DISC1 (See Supplemental Table). This increased detection of population-stratified deletions was due to the increased sensitivity for smaller variants. The average length of a call previously observed in phase 3, for example, was 1,322 bp, while the average length of the SV with FST ≥ 0.20 novel with respect to phase 3 was 358 bp. None of the sites with FST ≥ 0.20 were both novel with respect to 1kg phase 3 and overlapping a restrictive set of DNase hypersensitive (DHS) sites (top 23% of scoring DHS peaks).

#### **SVTyper genotyping:** The deletion calls from the HAN, PUR, and YRI children were separately genotyped in the Simons Human Genome Diversity Panel (HGDP) (n=238 samples), genotyping 6036, 6218, and 5865 SVs pers sample. After removing sites with LD> 0.2 and MAF < 0.10, 3,082, 3,074, and 2,582 sites remained. Similar to the Phase-3 analysis, the PCA demonstrated population based clustering, as expected.

We computed FST across genotypes for each sample, with the populations separated as 35 African, 16 Amerindian, 41 East Asian, 21 Oceanic, 34 South Asian, 20 SIB, 72 West Eurasian. A total of 109, 109, and 112 sites were found to have FST > 0.20 from each sample, respectively, of which 33, 35, and 43 intersected genes. Of these, one event genotyped from HG00733 was found to be exonic (*TUBA3E*, chr2:130197357-130199831, FST=0.20), and one event genotyped in NA19240, PLIN4 (del chr19:4512828-4513027, FST=0.22 ) were found to overlap exons.

### SMRT-Genotyper summary

See **figure S7.1.4, S7.1.5, Table S7.1.3**

## 

## 7.2 Sequence context analysis

The pan-technology integrated SV callset was compared to RefSeq exons and untranslated regions.

### Functional enrichment analysis

The SV calls from the integrated PacBio callset were compared with annotations of genes, exons, CDS, intron, transcribed processed pseudogenes, transcribed unprocessed pseudogenes, and transcripts were taken from gencode v25, along with transcription factor binding site peaks (Gencode + funseq), ultra sensitive, and ultra conserved elements from funseq. While there was a depletion of deletion SVs in all categories of genes, there was an increase in insertions within genes including exons and CDS.

### 

### Structural variants engulfing genes

**(Contributed by: Nodzak, Wen, Shi)**

The variants in the pan-technology integrated set harbor regions that completely overlap with protein coding genes. In the IL-SV set, 573 SVs engulfed whole genes including 183 deletions, 182 inversions, and 208 duplications, while in the PB-SV set, 21 deletions were found to completely engulf genes (**Table S7.2.1**). To assess the gene expression impact of the SVs that engulf genes, group *t*-tests were performed between the RPKM (Reads Per Kilobase Million) normalized expression values of real SV-engulfed genes and genes engulfed by analogous sets of randomly permuted regions using Strand-Specific RNA-seq data. To this end, we devised a method that found protein coding genes that were overlapped 100% by deletions, duplications and large inversions. This step was completed using BEDtools intersect with complete fractional overlap of sample-specific integrated Illumina SVs (filtered on a QUAL value of PASS and non-missing genotypes) on protein coding genes from gencode.v25.gtf, using the -f 1.0 command option (Williams 1989). The process was then repeated for deletions from the PB-SV deletion calls. Next, controlling for the size and chromosome of the SVs that engulf genes, 10,000 permutations were performed to create a set of random genomic regions with BEDtools shuffle. These permuted sets of regions were then used to identify sets of engulfed genes in the same manner described above. Specifically, strand-specific mRNA aligned to GRCh38 read quantification was performed for protein coding genes annotated by the gencode version 25 GTF using featureCounts, part of the Subread package [(Liao, Smyth, and Shi 2014)](https://paperpile.com/c/ml0krT/lnz7). Coverage normalization was then performed using the RPKM method, which resulted in a final set of expression values of 18,873 genes for the nine samples [(Mortazavi et al. 2008)](https://paperpile.com/c/ml0krT/tN2J). With each set of RPKM values for genes overlapped by a real variant call in a given sample, a group t-test was performed against the RPKM values for each of the lists of genes overlapped by the permuted regions. Finally multi-test correction was applied using the fdrtool package in R [(Strimmer 2008)](https://paperpile.com/c/FLjKLc/pG3A). The q-values were then -log2 transformed and the average was plotted for each sample to assess the significance of the effects of each type of variant on the expression of engulfed genes (**Figure S7.2.2**).

Our results (**Figure S7.2.2**) illustrated a high level of congruence of the expression effect brought about by these gene-engulfing SVs across the samples. Particularly, the expression of IL-SV deletion-engulfed genes for all nine samples showed significant differences from the permuted genes. Similarly, all three trio daughters showed significant differences in expression for those genes completely overlapped by large IL-SV inversions. 7/9 individuals were found to be impacted in a similar manner for the sets of whole-gene duplications. When the same analysis was conducted using the PB-SV deletions for the three trio daughters, we found 21 deletions were found to completely engulf protein coding genes (**Table S7.2.1**) and one of the three individuals showed significant differences in the expression of the affected genes (**Figure 7.2.2**).

### INDEL functional annotations

**(Contributed by: Wen, Shi)**

We annotate the integrated Illumina INDEL callset for three trios using the variant effect predictor (VEP) [(McLaren et al. 2016)](https://paperpile.com/c/FLjKLc/LqpG). Of 1,743,129 autosomal small INDELs from the Illumina INDEL integrated set, 1,944 INDELs are located in coding sequence region. For INDELs with different predicted consequences, we only count one entry for each INDEL corresponding to its impact of the longest transcript (**Table S7.2.2**), where the transcript annotation file used for was gencode.v25.transcripts.fa. 44.96% (874/1,94) of these INDELs were frameshift variants (FS) and 52.78% (1,026/1,944) were non-frameshift variants (NFS). In addition, 44 INDELs were annotated as other types of variants including coding sequence variant, protein coding variant, splice acceptor variant, splice donor variant, stop gained variant and nonsense mediated decay (NMD) transcript variant.

Allele specific expression analysis

**(Contributed by: Nodzak, Wen, Shi)**

Allele specific expression (ASE) analyzes differences in expression by leveraging heterozygous sites in diploid organisms. Former studies found up to 30% of loci showing allelic-specific differences on the transcript at individual level [(Pastinen 2010)](https://paperpile.com/c/ml0krT/XVBD). One study showed that approximately 20% of human genes can be affected by ASE in European populations [(Serre et al. 2008)](https://paperpile.com/c/ml0krT/UgSY). Dimas et al (2008) found that nearly 18% of human genes showed differential allelic expression in HapMap populations [(A.S. Dimas, Stranger BE, Beazley C, Finn RD, Ingle CE, Forrest MS, Ritchie ME, Deloukas P, Tavaré S, Dermitzakis ET 2008)](https://paperpile.com/c/ml0krT/tfsQ).

We conducted allele specific expression analysis of SNPs and SVs using the strand-specific RNA-seq data on the trios (**Figure S7.2.3**). We started with SNP ASE analysis, based on the strand-specific RNA-seq data and the Whatshap strand-seq 10X phased SNPs using a pipeline which first conducted mapping bias correction using WASP [(van de Geijn et al. 2015)](https://paperpile.com/c/ml0krT/GCKr) and applied the binomial test with multi-correction test on FDR 5% [(Storey, Taylor, and Siegmund 2004)](https://paperpile.com/c/ml0krT/5Trg).

First, the strand-specific RNA-seq fastq files were mapped using the STAR (v2.4.2a) with default option to the GRCh38 human reference to create bam alignment files [(Dobin et al. 2013)](https://paperpile.com/c/ml0krT/ZQ84).

We then applied WASP for correcting the mapping bias. The STAR bam files were remapped to all SNPs and we discarded the reads not mapped to the same location with reference allele after flipped to the alternative allele. Duplicate reads were removed using Picard (http://broadinstitute.github.io/picard) where the reads with best quality and least mismatch were kept.

At the quality filtering step, we used the following criteria to keep uniquely mapped reads. These reads will have an i) NM <= 6, ii) a base quality >= 10, iii) a mapping quality score > 20, and iv) total read count >=8 at each both allele seen heterozygous SNP (het-SNP) site. If the four criteria above were satisfied, then the number of reference allele count and alternative allele count were extracted with perl script *samase.pl* providedby [(Kukurba et al 2014)](http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004304), and a subsequent binomial test was performed with FDR-based (5%) multi-test correction conducted afterwards.

To evaluate the effect of overdispersion in the Strand-Specific RNA-seq data, we selected YRI mother NA19238 to calculate the overdispersion to compare the binomial test and beta-binomial test with the same input for both allele seen heterozygous SNPs, which the beta-binomial test used here is from Chen et al (2016) [(J. Chen et al. 2016)](https://paperpile.com/c/ml0krT/GSKw+JLK7+gazb). The overdispersion is 0.0152 which is low such that the binomial test and beta-binomial test would give similar results (**Figure 7.2.4**). Hence, we chose to use binomial test for our WASP based allele specific analysis pipeline as the controlling of mapping biases and overdispersion would not affect our ASE results.

Using annotations by gencode v25 protein coding genes, we identified 1,920 SNPs to exhibit allele-specific expression on 1,043 genes for YRI trio, 1,556 ASE SNPs on 857 genes for PUR trio and 1,363 ASE SNPs on 703 allele specific genes for CHS trio. Among three trios, a total of 4,292 SNPs are statistically identified as ASE SNPs, 3,609 of these ASE SNPs affect 1,947 genes. When overlapping with published ASE results, 84% (3,622/4,292) of these ASE SNPs and 99% (1,943/1,947) of ASE genes were reported in previous literatures [(Lappalainen et al. 2013; J. Chen et al. 2016; Pirinen et al. 2015)](https://paperpile.com/c/ml0krT/GSKw+JLK7+gazb), among which 78 ASE SNPs identified were shown as eQTLs in other literature [(Sudmant, Rausch, et al. 2015; Stranger et al. 2012; A. L. Dixon et al. 2007; Pickrell et al. 2010; Battle et al. 2015; Gutierrez-Arcelus et al. 2013; Wen, Luca, and Pique-Regi 2015; Montgomery et al. 2010; Y. I. Li et al. 2016; Lappalainen et al. 2013; Grubert et al. 2015)](https://paperpile.com/c/FLjKLc/IVdOB+cdNE+mwT0+UDeB+trBY+13lI+wUcI+DmMM+SVWW+TYRoH+gaD4). Of the 4,292 ASE SNPs identified in this study, 30 of them were GWAS signals associated with 28 human traits including diabetes, obesity, and Alzheimer's disease [(MacArthur et al. 2017)](https://paperpile.com/c/ml0krT/DypA).

Of 1,947 ASE genes identified in the three trios, 13 were experimentally validated imprinting genes (<http://geneimprint.com/site/genes-by-species.Homo+sapiens>, **Table S7.2.3**). Two of these imprinted genes, *SNURF*/*SNRPN* and *GNAS*, were also previously reported as an imprinted ASE genes [(Degner et al. 2009)](https://paperpile.com/c/ml0krT/7DTC). *SNURF*/*SNRPN* is a paternally expressed gene, which our ASE result shows a consistent inheritance pattern with the imprinted pattern (mRNA read counts of two alleles at 283 vs 1 for YRI father and 213 vs 1 for YRI child; 1 vs 287 for CHS father and 2 vs 230 for CHS child).

Next, we sought to investigate if SVs also bring about ASE effect on those genes whose expressions were shown to be allele specific from the SNP-ASE analysis above. In order to do so, we developed an SV-ASE analysis pipeline (**Figure S7.2.3**) with the following three steps for PB-SVs and IL-SVs respectively. First, we established a set of candidate SVs-gene pairs by taking the intersection of heterozygous SVs (het-SVs) with SNP-ASE genes. Second, phased RNA-Seq reads were filtered following the same criteria established by our SNP-ASE analysis above and read counts of the genes were calculated for each sample’s two haplotypes using BEDtools multicov. Third, the significance of SV-genes pairs was then obtained by applying a binomial test to the read counts of the two haplotypes with multi-test correction using FDR 5%.

Our results (**Table S7.2.4**) showed that the majority of het-SVs tested significantly affected the target gene expression in allele specific manner. Specifically, in the PB-SV set, a total of 151 SVs (73 insertions and 78 deletions) showed ASE effect on 60 genes, out of the 184 het-SVs intersected with 78 SNP-ASE genes for NA19240; a total of 211 SVs (97 insertions and 114 deletions) showed ASE effect on 78 genes, out of the 201 het-SVs intersected with 85 SNP-ASE genes for HG00514; and a total of 222 SVs (144 insertions and 78 deletions) showed ASE effect on 89 genes, out of the 265 het-SVs intersected with 106 SNP-ASE genes for HG00733. In the IL-SV set, 88 SVs (13 insertions, 72 deletions and 3 inversions) demonstrated ASE effect on 74 genes, out of the 106 het-SVs intersected with 80 SNP-ASE genes for HG00514; 87 SVs (16 insertions, 67 deletions and 4 inversions) demonstrated ASE effect on 73 genes, out of the 120 het-SVs intersected with 93 SNP-ASE genes for HG00733; and 74 SVs (9 insertions, 62 deletions and 3 inversions) demonstrated ASE effect on 62 genes, out of the 80 het-SVs intersected with 77 SNP-ASE genes for NA19240.

Our SV-ASE results prompted us next to address whether or not the observed allelic imbalance at SV-ASE genes was attributable to a local haplotype along the gene region. For this, we calculated the LD (R2 values) between the SVs and SNPs with ASE effect on the same gene. We illustrated this analysis to assess the allelic effect resulting from a heterozygous deletion belonging to HG00514 within a transcription factor binding site on exon 5 of the *ZNF717* gene (**Figure S7.2.5**). We further ruled against a haploblock effect driving the allelic imbalance between the haplotype from low R2 values for the sample's variants and those from the 1000GP phase3 CHS population within a window ± 100kb of the gene, and showed that there were few variants with high R2 within the exon as well.

### Variants overlapping GWAS sites.

We searched for deletion SV that overlapped GWAS sites, since common structural variation that has not been detected could affect interpretation allele frequency. The results of SV overlapping with 11,737 GWAS sites is given in **Table S7.2.4**. There were 16 deletions from HG00514, 9 HG00733, and 16 from NA19240 that were overlapping GWAS sites, with an average length of 286 Kbp and median length of 16,302 bp.

### 

### Haplotype specific reference analysis

We constructed haplotype specific genome from the Phased-SV assemblies by mapping each assembly to the reference and replacing reference sequence with assembly according to the alignments. Additional indel errors were corrected in the reference by partitioning 10X reads by haplotype according to the 10X-StrandSeq phasing ([www.github.com/mchaisso/Split10x](http://www.github.com/mchaisso/Split10x)), mapping by bwa mem (-t 8 -p) calling variants using freebayes (-p 1 -m 10 --min-coverage 10 --max-coverage 40), and modifying the reference according to the resulting vcf, correcting 0.8-3M sites per genome, ~85% of which were indel modifications.

To test for impact on analysis, RNA-seq reads were mapped to each haplotype, and differences in read counts were recorded. Strand-specific RNA-seq reads from each sample were mapped to each haplotype assembly, separately (STAR --readFilesCommand zcat --genomeLoad LoadAndKeep). Gene models were detected in the assemblies by lifting over from hg38: each assembly was mapped to hg38 with blasr ([www.github.com/mchaisso/blasr](http://www.github.com/mchaisso/blasr)) (-alignContigs -minMapQV 30), and lifting over using samLiftover ([www.github.com/mchaisso/mcutils](http://www.github.com/mchaisso/mcutils)) (--dir 1 --useXS). Expression was measured for each exon as the number of reads mapped. Because there was a direct comparison of the same exons between assemblies, and only one RNA-seq experiment was used per sample, no normalization routines were applied to the read counts.

## 

## 7.3 MEI discovery and analysis with PacBio data

**(Contributed by: Gardner, E. Chuang, Nelson, and Devine, S.E.)**

The raw PacBio traces also were independently used to discover L1 insertions by first identifying traces that contained L1 sequences, followed by clustering these traces by genomic location. The traces that mapped to each independent (non-REF) cluster were then assembled using BLASR (Chaisson and Tesler, 2012) and novel L1 insertions were identified. This PacBio-based L1 discovery tool is available with the MELT package (Gardner et al. 2017). Full-length L1 (FL-L1) source elements (6 kb in length) also were independently fully sequenced by first amplifying each element with long-range PCR, followed by PacBio sequencing of PCR cassettes in batches of 50-100 FL-L1 elements. These elements were sequenced to a depth of at least 500X and compared with the same element sequences that were recovered from the WGS and clustered PacBio data.

8. References

[1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.](http://paperpile.com/b/FLjKLc/AldgO)

[Aken, Bronwen L., Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, et al. 2016. “The Ensembl Gene Annotation System.” *Database: The Journal of Biological Databases and Curation* 2016 (June). doi:](http://paperpile.com/b/FLjKLc/DOHFg)[10.1093/database/baw093](http://dx.doi.org/10.1093/database/baw093)[.](http://paperpile.com/b/FLjKLc/DOHFg)

[Andrews, Simon, and Others. 2010. “FastQC: A Quality Control Tool for High Throughput Sequence Data.”](http://paperpile.com/b/FLjKLc/QN18V)

[Antonacci, Francesca, Jeffrey M. Kidd, Tomas Marques-Bonet, Brian Teague, Mario Ventura, Santhosh Girirajan, Can Alkan, et al. 2010. “A Large and Complex Structural Polymorphism at 16p12.1 Underlies Microdeletion Disease Risk.” *Nature Genetics* 42 (9): 745–50.](http://paperpile.com/b/FLjKLc/LdDi7)

[A.S. Dimas, Stranger BE, Beazley C, Finn RD, Ingle CE, Forrest MS, Ritchie ME, Deloukas P, Tavaré S, Dermitzakis ET. 2008. “Modifier Effects between Regulatory and Protein-Coding Variation.” *PloS Genetics* 4 (10): e1000244.](http://paperpile.com/b/FLjKLc/s04yD)

[Bailey, Jeffrey A., and Evan E. Eichler. 2006. “Primate Segmental Duplications: Crucibles of Evolution, Diversity and Disease.” *Nature Reviews. Genetics* 7 (7): 552–64.](http://paperpile.com/b/FLjKLc/14eKO)

[Bansal, Vikas, and Vineet Bafna. 2008. “HapCUT: An Efficient and Accurate Algorithm for the Haplotype Assembly Problem.” *Bioinformatics*  24 (16): i153–59.](http://paperpile.com/b/FLjKLc/IuxE1)

[Bansal, Vikas, Aaron L. Halpern, Nelson Axelrod, and Vineet Bafna. 2008. “An MCMC Algorithm for Haplotype Assembly from Whole-Genome Sequence Data.” *Genome Research* 18 (8): 1336–46.](http://paperpile.com/b/FLjKLc/IlemA)

[Barnett, Derek W., Erik K. Garrison, Aaron R. Quinlan, Michael P. Strömberg, and Gabor T. Marth. 2011. “BamTools: A C++ API and Toolkit for Analyzing and Managing BAM Files.” *Bioinformatics*  27 (12): 1691–92.](http://paperpile.com/b/FLjKLc/k4PPh)

[Battle, Alexis, Zia Khan, Sidney H. Wang, Amy Mitrano, Michael J. Ford, Jonathan K. Pritchard, and Yoav Gilad. 2015. “Genomic Variation. Impact of Regulatory Variation from RNA to Protein.” *Science* 347 (6222): 664–67.](http://paperpile.com/b/FLjKLc/trBY)

[Berlin, Konstantin, Sergey Koren, Chen-Shan Chin, James P. Drake, Jane M. Landolin, and Adam M. Phillippy. 2015. “Assembling Large Genomes with Single-Molecule Sequencing and Locality-Sensitive Hashing.” *Nature Biotechnology* 33 (6): 623–30.](http://paperpile.com/b/FLjKLc/Chh4v)

[Bishara, Alex, Yuling Liu, Ziming Weng, Dorna Kashef-Haghighi, Daniel E. Newburger, Robert West, Arend Sidow, and Serafim Batzoglou. 2015. “Read Clouds Uncover Variation in Complex Regions of the Human Genome.” *Genome Research* 25 (10): 1570–80.](http://paperpile.com/b/FLjKLc/DRG5m)

[Blumenthal, Ian, Ashok Ragavendran, Serkan Erdin, Lambertus Klei, Aarathi Sugathan, Jolene R. Guide, Poornima Manavalan, et al. 2014. “Transcriptional Consequences of 16p11.2 Deletion and Duplication in Mouse Cortex and Multiplex Autism Families.” *American Journal of Human Genetics* 94 (6): 870–83.](http://paperpile.com/b/FLjKLc/ynSS6)

[Brand, Harrison, Ryan L. Collins, Carrie Hanscom, Jill A. Rosenfeld, Vamsee Pillalamarri, Matthew R. Stone, Fontina Kelley, et al. 2015. “Paired-Duplication Signatures Mark Cryptic Inversions and Other Complex Structural Variation.” *American Journal of Human Genetics* 97 (1): 170–76.](http://paperpile.com/b/FLjKLc/4b5Fg)

[Broman, K. W., J. C. Murray, V. C. Sheffield, R. L. White, and J. L. Weber. 1998. “Comprehensive Human Genetic Maps: Individual and Sex-Specific Variation in Recombination.” *American Journal of Human Genetics* 63 (3): 861–69.](http://paperpile.com/b/FLjKLc/k9tR9)

[Chaisson, Mark J. P., Richard K. Wilson, and Evan E. Eichler. 2015. “Genetic Variation and the de Novo Assembly of Human Genomes.” *Nature Reviews. Genetics* 16 (11): 627–40.](http://paperpile.com/b/FLjKLc/DyJCt)

[Chaisson, Mark J., and Glenn Tesler. 2012. “Mapping Single Molecule Sequencing Reads Using Basic Local Alignment with Successive Refinement (BLASR): Application and Theory.” *BMC Bioinformatics* 13 (September): 238.](http://paperpile.com/b/FLjKLc/w20qF)

[Chen, Jieming, Joel Rozowsky, Timur R. Galeev, Arif Harmanci, Robert Kitchen, Jason Bedford, Alexej Abyzov, Yong Kong, Lynne Regan, and Mark Gerstein. 2016. “A Uniform Survey of Allele-Specific Binding and Expression over 1000-Genomes-Project Individuals.” *Nature Communications* 7 (April): 11101.](http://paperpile.com/b/FLjKLc/bTbDA)

[Chen, Xiaoyu, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J. Cox, Semyon Kruglyak, and Christopher T. Saunders. 2016. “Manta: Rapid Detection of Structural Variants and Indels for Germline and Cancer Sequencing Applications.” *Bioinformatics*  32 (8): 1220–22.](http://paperpile.com/b/FLjKLc/mFsK8)

[Chiang, Colby, Ryan M. Layer, Gregory G. Faust, Michael R. Lindberg, David B. Rose, Erik P. Garrison, Gabor T. Marth, Aaron R. Quinlan, and Ira M. Hall. 2015. “SpeedSeq: Ultra-Fast Personal Genome Analysis and Interpretation.” *Nature Methods* 12 (10): 966–68.](http://paperpile.com/b/FLjKLc/j9a0m)

[Chin, Chen-Shan, David H. Alexander, Patrick Marks, Aaron A. Klammer, James Drake, Cheryl Heiner, Alicia Clum, et al. 2013. “Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data.” *Nature Methods* 10 (6): 563–69.](http://paperpile.com/b/FLjKLc/bo8Yk)

[Chong, Zechen, Jue Ruan, Min Gao, Wanding Zhou, Tenghui Chen, Xian Fan, Li Ding, et al. 2016. “novoBreak: Local Assembly for Breakpoint Detection in Cancer Genomes.” *Nature Methods* 14 (1). Nature Research: 65–67.](http://paperpile.com/b/FLjKLc/psOf7)

[Cilibrasi, Rudi, Leo van Iersel, Steven Kelk, and John Tromp. 2005. “On the Complexity of Several Haplotyping Problems.” In *Algorithms in Bioinformatics*, edited by Rita Casadio and Gene Myers, 3692:128–39. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg.](http://paperpile.com/b/FLjKLc/Jjp4y)

[Clarke, Laura, Susan Fairley, Xiangqun Zheng-Bradley, Ian Streeter, Emily Perry, Ernesto Lowy, Anne-Marie Tassé, and Paul Flicek. 2017. “The International Genome Sample Resource (IGSR): A Worldwide Collection of Genome Variation Incorporating the 1000 Genomes Project Data.” *Nucleic Acids Research* 45 (D1): D854–59.](http://paperpile.com/b/FLjKLc/gZdIZ)

[Collins, Ryan L., Harrison Brand, Claire E. Redin, Carrie Hanscom, Caroline Antolik, Matthew R. Stone, Joseph T. Glessner, et al. 2017. “Defining the Diverse Spectrum of Inversions, Complex Structural Variation, and Chromothripsis in the Morbid Human Genome.” *Genome Biology* 18 (1): 36.](http://paperpile.com/b/FLjKLc/jVbPq)

[Conrad, Donald F., Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, Jan Aerts, et al. 2010. “Origins and Functional Impact of Copy Number Variation in the Human Genome.” *Nature* 464 (7289): 704–12.](http://paperpile.com/b/FLjKLc/F2udg)

[Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics*  27 (15): 2156–58.](http://paperpile.com/b/FLjKLc/NqA53)

[Degner, Jacob F., John C. Marioni, Athma A. Pai, Joseph K. Pickrell, Everlyne Nkadori, Yoav Gilad, and Jonathan K. Pritchard. 2009. “Effect of Read-Mapping Biases on Detecting Allele-Specific Expression from RNA-Sequencing Data.” *Bioinformatics*  25 (24): 3207–12.](http://paperpile.com/b/FLjKLc/cRX5E)

[DeLuca, David S., Joshua Z. Levin, Andrey Sivachenko, Timothy Fennell, Marc-Danie Nazaire, Chris Williams, Michael Reich, Wendy Winckler, and Gad Getz. 2012. “RNA-SeQC: RNA-Seq Metrics for Quality Control and Process Optimization.” *Bioinformatics*  28 (11): 1530–32.](http://paperpile.com/b/FLjKLc/GGtmK)

[DePristo, Mark A., Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, et al. 2011. “A Framework for Variation Discovery and Genotyping Using next-Generation DNA Sequencing Data.” *Nature Genetics* 43 (5): 491–98.](http://paperpile.com/b/FLjKLc/p7N6n)

[Dixon, Anna L., Liming Liang, Miriam F. Moffatt, Wei Chen, Simon Heath, Kenny C. C. Wong, Jenny Taylor, et al. 2007. “A Genome-Wide Association Study of Global Gene Expression.” *Nature Genetics* 39 (10): 1202–7.](http://paperpile.com/b/FLjKLc/mwT0)

[Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. 2012. “Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions.” *Nature* 485 (7398): 376–80.](http://paperpile.com/b/FLjKLc/BfyIM)

[Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics*  29 (1): 15–21.](http://paperpile.com/b/FLjKLc/h61BL)

[Falconer, Ester, Mark Hills, Ulrike Naumann, Steven S. S. Poon, Elizabeth A. Chavez, Ashley D. Sanders, Yongjun Zhao, Martin Hirst, and Peter M. Lansdorp. 2012. “DNA Template Strand Sequencing of Single-Cells Maps Genomic Rearrangements at High Resolution.” *Nature Methods* 9 (11): 1107–12.](http://paperpile.com/b/FLjKLc/wNJjV)

[Faust, Gregory G., and Ira M. Hall. 2014. “SAMBLASTER: Fast Duplicate Marking and Structural Variant Read Extraction.” *Bioinformatics*  30 (17): 2503–5.](http://paperpile.com/b/FLjKLc/OgBjV)

[Garg, Shilpa, Marcel Martin, and Tobias Marschall. 2016. “Read-Based Phasing of Related Individuals.” *Bioinformatics*  32 (12): i234–42.](http://paperpile.com/b/FLjKLc/6PczH)

[Garrison, Erik, and Gabor Marth. 2012. “Haplotype-Based Variant Detection from Short-Read Sequencing.” *arXiv [q-bio.GN]*. arXiv.](http://paperpile.com/b/FLjKLc/r969f) <http://arxiv.org/abs/1207.3907>[.](http://paperpile.com/b/FLjKLc/r969f)

[Geijn, Bryce van de, Graham McVicker, Yoav Gilad, and Jonathan K. Pritchard. 2015. “WASP: Allele-Specific Software for Robust Molecular Quantitative Trait Locus Discovery.” *Nature Methods* 12 (11): 1061–63.](http://paperpile.com/b/FLjKLc/NjXL2)

[Grubert, Fabian, Judith B. Zaugg, Maya Kasowski, Oana Ursu, Damek V. Spacek, Alicia R. Martin, Peyton Greenside, et al. 2015. “Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions.” *Cell* 162 (5): 1051–65.](http://paperpile.com/b/FLjKLc/gaD4)

[Gutierrez-Arcelus, Maria, Tuuli Lappalainen, Stephen B. Montgomery, Alfonso Buil, Halit Ongen, Alisa Yurovsky, Julien Bryois, et al. 2013. “Passive and Active DNA Methylation and the Interplay with Genetic Variation in Gene Regulation.” *eLife* 2 (June): e00523.](http://paperpile.com/b/FLjKLc/13lI)

[Handsaker, Robert E., Joshua M. Korn, James Nemesh, and Steven A. McCarroll. 2011. “Discovery and Genotyping of Genome Structural Polymorphism by Sequencing on a Population Scale.” *Nature Genetics* 43 (3). Nature Research: 269–76.](http://paperpile.com/b/FLjKLc/sisuy)

[Hanscom, C., and M. Talkowski. 2014. “Design of Large-Insert Jumping Libraries for Structural Variant Detection Using Illumina Sequencing.” *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.]* 80 (January): 7.22.1–9.](http://paperpile.com/b/FLjKLc/K6Y3c)

[Hills, Mark, Kieran O’Neill, Ester Falconer, Ryan Brinkman, and Peter M. Lansdorp. 2013. “BAIT: Organizing Genomes and Mapping Rearrangements in Single Cells.” *Genome Medicine* 5 (9): 82.](http://paperpile.com/b/FLjKLc/ygLNh)

[Hormozdiari, Fereydoun, Can Alkan, Evan E. Eichler, and S. Cenk Sahinalp. 2009. “Combinatorial Algorithms for Structural Variation Detection in High-Throughput Sequenced Genomes.” *Genome Research* 19 (7): 1270–78.](http://paperpile.com/b/FLjKLc/8C4Uj)

[Hormozdiari, Fereydoun, Can Alkan, Mario Ventura, Iman Hajirasouliha, Maika Malig, Faraz Hach, Deniz Yorukoglu, et al. 2011. “Alu Repeat Discovery and Characterization within Human Genomes.” *Genome Research* 21 (6): 840–49.](http://paperpile.com/b/FLjKLc/0yyYt)

[Hormozdiari, Fereydoun, Iman Hajirasouliha, Andrew McPherson, Evan E. Eichler, and S. Cenk Sahinalp. 2011. “Simultaneous Structural Variation Discovery in Multiple Paired-End Sequenced Genomes.” In *Research in Computational Molecular Biology*, edited by Vineet Bafna and S. Cenk Sahinalp, 6577:104–5. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg.](http://paperpile.com/b/FLjKLc/TPEAM)

[Hou, Yu, Wei Fan, Liying Yan, Rong Li, Ying Lian, Jin Huang, Jinsen Li, et al. 2013. “Genome Analyses of Single Human Oocytes.” *Cell* 155 (7): 1492–1506.](http://paperpile.com/b/FLjKLc/Og4ph)

[Huddleston, John, Mark Jp Chaisson, Karyn Meltz Steinberg, Wes Warren, Kendra Hoekzema, David S. Gordon, Tina A. Graves-Lindsay, et al. 2016. “Discovery and Genotyping of Structural Variation from Long-Read Haploid Genome Sequence Data.” *Genome Research*, November. doi:](http://paperpile.com/b/FLjKLc/mxhY5)[10.1101/gr.214007.116](http://dx.doi.org/10.1101/gr.214007.116)[.](http://paperpile.com/b/FLjKLc/mxhY5)

[Huddleston, John, Swati Ranade, Maika Malig, Francesca Antonacci, Mark Chaisson, Lawrence Hon, Peter H. Sudmant, et al. 2014. “Reconstructing Complex Regions of Genomes Using Long-Read Sequencing Technology.” *Genome Research* 24 (4): 688–96.](http://paperpile.com/b/FLjKLc/WWmqo)

[Kent, W. James. 2002. “BLAT--the BLAST-like Alignment Tool.” *Genome Research* 12 (4): 656–64.](http://paperpile.com/b/FLjKLc/f5fue)

[Kirkness, Ewen F., Rashel V. Grindberg, Joyclyn Yee-Greenbaum, Christian R. Marshall, Stephen W. Scherer, Roger S. Lasken, and J. Craig Venter. 2013. “Sequencing of Isolated Sperm Cells for Direct Haplotyping of a Human Genome.” *Genome Research* 23 (5): 826–32.](http://paperpile.com/b/FLjKLc/5poKW)

[Klambauer, Günter, Karin Schwarzbauer, Andreas Mayr, Djork-Arné Clevert, Andreas Mitterecker, Ulrich Bodenhofer, and Sepp Hochreiter. 2012. “cn.MOPS: Mixture of Poissons for Discovering Copy Number Variations in next-Generation Sequencing Data with a Low False Discovery Rate.” *Nucleic Acids Research* 40 (9): e69.](http://paperpile.com/b/FLjKLc/WL4d1)

[Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. “Canu: Scalable and Accurate Long-Read Assembly via Adaptive K-Mer Weighting and Repeat Separation.” *Genome Research* 27 (5): 722–36.](http://paperpile.com/b/FLjKLc/DrPd7)

[Kronenberg, Zev N., Edward J. Osborne, Kelsey R. Cone, Brett J. Kennedy, Eric T. Domyan, Michael D. Shapiro, Nels C. Elde, and Mark Yandell. 2015. “Wham: Identifying Structural Variants of Biological Consequence.” *PLoS Computational Biology* 11 (12): e1004572.](http://paperpile.com/b/FLjKLc/6Ztg0)

[Lappalainen, Tuuli, Michael Sammeth, Marc R. Friedländer, Peter A. C. ’t Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, et al. 2013. “Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans.” *Nature* 501 (7468): 506–11.](http://paperpile.com/b/FLjKLc/TYRoH)

[Layer, Ryan M., Colby Chiang, Aaron R. Quinlan, and Ira M. Hall. 2014. “LUMPY: A Probabilistic Framework for Structural Variant Discovery.” *Genome Biology* 15 (6): R84.](http://paperpile.com/b/FLjKLc/SGKEb)

[Lee, Christopher, Catherine Grasso, and Mark F. Sharlow. 2002. “Multiple Sequence Alignment Using Partial Order Graphs.” *Bioinformatics*  18 (3): 452–64.](http://paperpile.com/b/FLjKLc/41lbx)

[Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics*  30 (7): 923–30.](http://paperpile.com/b/FLjKLc/WEuQD)

[Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” *Science* 326 (5950): 289–93.](http://paperpile.com/b/FLjKLc/xGDhe)

[Li, H., and R. Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform.” *Bioinformatics* . academic.oup.com.](http://paperpile.com/b/FLjKLc/tUin6) <https://academic.oup.com/bioinformatics/article-abstract/25/14/1754/225615>[.](http://paperpile.com/b/FLjKLc/tUin6)

[Li, Heng. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” *arXiv [q-bio.GN]*. arXiv.](http://paperpile.com/b/FLjKLc/qgSzT) <http://arxiv.org/abs/1303.3997>[.](http://paperpile.com/b/FLjKLc/qgSzT)

[———. 2016. “Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences.” *Bioinformatics*  32 (14): 2103–10.](http://paperpile.com/b/FLjKLc/sNMCB)

[Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics*  25 (16): 2078–79.](http://paperpile.com/b/FLjKLc/7Gn1n)

[Lippert, Ross, Russell Schwartz, Giuseppe Lancia, and Sorin Istrail. 2002. “Algorithmic Strategies for the Single Nucleotide Polymorphism Haplotype Assembly Problem.” *Briefings in Bioinformatics* 3 (1): 23–31.](http://paperpile.com/b/FLjKLc/hi4O9)

[Li, Yang I., Bryce van de Geijn, Anil Raj, David A. Knowles, Allegra A. Petti, David Golan, Yoav Gilad, and Jonathan K. Pritchard. 2016. “RNA Splicing Is a Primary Link between Genetic Variation and Disease.” *Science* 352 (6285): 600–604.](http://paperpile.com/b/FLjKLc/SVWW)

[Loh, Po-Ru, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, et al. 2016. “Reference-Based Phasing Using the Haplotype Reference Consortium Panel.” *Nature Genetics* 48 (11): 1443–48.](http://paperpile.com/b/FLjKLc/JDhDf)

[Lu, Sijia, Chenghang Zong, Wei Fan, Mingyu Yang, Jinsen Li, Alec R. Chapman, Ping Zhu, et al. 2012. “Probing Meiotic Recombination and Aneuploidy of Single Sperm Cells by Whole-Genome Sequencing.” *Science* 338 (6114): 1627–30.](http://paperpile.com/b/FLjKLc/qdB3L)

[MacArthur, Jacqueline, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, et al. 2017. “The New NHGRI-EBI Catalog of Published Genome-Wide Association Studies (GWAS Catalog).” *Nucleic Acids Research* 45 (D1): D896–901.](http://paperpile.com/b/FLjKLc/eGUYp)

[Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.journal* 17 (1): 10–12.](http://paperpile.com/b/FLjKLc/aCQbR)

[Martin, Marcel, Murray Patterson, Shilpa Garg, Sarah O. Fischer, Nadia Pisanti, Gunnar W. Klau, Alexander Schoenhuth, and Tobias Marschall. 2016. “WhatsHap: Fast and Accurate Read-Based Phasing.” *bioRxiv*. doi:](http://paperpile.com/b/FLjKLc/MSzh9)[10.1101/085050](http://dx.doi.org/10.1101/085050)[.](http://paperpile.com/b/FLjKLc/MSzh9)

[McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303.](http://paperpile.com/b/FLjKLc/EG8kv)

[McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. “The Ensembl Variant Effect Predictor.” *Genome Biology* 17 (1): 122.](http://paperpile.com/b/FLjKLc/LqpG)

[Michaelson, Jacob J., and Jonathan Sebat. 2012. “forestSV: Structural Variant Discovery through Statistical Learning.” *Nature Methods* 9 (8): 819–21.](http://paperpile.com/b/FLjKLc/twdqp)

[Montgomery, Stephen B., Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P. Lach, Catherine Ingle, James Nisbett, Roderic Guigo, and Emmanouil T. Dermitzakis. 2010. “Transcriptome Genetics Using Second Generation Sequencing in a Caucasian Population.” *Nature* 464 (7289): 773–77.](http://paperpile.com/b/FLjKLc/DmMM)

[Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq.” *Nature Methods* 5 (7): 621–28.](http://paperpile.com/b/FLjKLc/Vyo9J)

[Myers, Simon, Colin Freeman, Adam Auton, Peter Donnelly, and Gil McVean. 2008. “A Common Sequence Motif Associated with Recombination Hot Spots and Genome Instability in Humans.” *Nature Genetics* 40 (9): 1124–29.](http://paperpile.com/b/FLjKLc/jXgNu)

[Pastinen, Tomi. 2010. “Genome-Wide Allele-Specific Analysis: Insights into Regulatory Variation.” *Nature Reviews Genetics* 11 (August): 533–38.](http://paperpile.com/b/FLjKLc/COJXC)

[Patterson, Murray, Tobias Marschall, Nadia Pisanti, Leo van Iersel, Leen Stougie, Gunnar W. Klau, and Alexander Schönhuth. 2015. “WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 22 (6): 498–509.](http://paperpile.com/b/FLjKLc/3jmgd)

[Pickrell, Joseph K., John C. Marioni, Athma A. Pai, Jacob F. Degner, Barbara E. Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. 2010. “Understanding Mechanisms Underlying Human Gene Expression Variation with RNA Sequencing.” *Nature* 464 (7289): 768–72.](http://paperpile.com/b/FLjKLc/UDeB)

[Pirinen, Matti, Tuuli Lappalainen, Noah A. Zaitlen, GTEx Consortium, Emmanouil T. Dermitzakis, Peter Donnelly, Mark I. McCarthy, and Manuel A. Rivas. 2015. “Assessing Allele-Specific Expression across Multiple Tissues from RNA-Seq Read Data.” *Bioinformatics*  31 (15): 2497–2504.](http://paperpile.com/b/FLjKLc/wdDCi)

[Porubský, David, Ashley D. Sanders, Niek van Wietmarschen, Ester Falconer, Mark Hills, Diana C. J. Spierings, Marianna R. Bevova, Victor Guryev, and Peter M. Lansdorp. 2016. “Direct Chromosome-Length Haplotyping by Single-Cell Sequencing.” *Genome Research* 26 (11): 1565–74.](http://paperpile.com/b/FLjKLc/4SSLl)

[Porubsky, D., S. Garg, A. D. Sanders, J. O. Korbel, and V. Guryev. 2017. “Dense And Accurate Whole-Chromosome Haplotyping Of Individual Genomes.” *bioRxiv*. biorxiv.org.](http://paperpile.com/b/FLjKLc/OPuuM) <http://biorxiv.org/content/early/2017/04/10/126136.abstract>[.](http://paperpile.com/b/FLjKLc/OPuuM)

[Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics*  26 (6): 841–42.](http://paperpile.com/b/FLjKLc/KLaCJ)

[Rausch, Tobias, Thomas Zichner, Andreas Schlattl, Adrian M. Stütz, Vladimir Benes, and Jan O. Korbel. 2012. “DELLY: Structural Variant Discovery by Integrated Paired-End and Split-Read Analysis.” *Bioinformatics*  28 (18): i333–39.](http://paperpile.com/b/FLjKLc/SDK48)

[Ritz, Anna, Pamela L. Paris, Michael M. Ittmann, Colin Collins, and Benjamin J. Raphael. 2011. “Detection of Recurrent Rearrangement Breakpoints from Copy Number Data.” *BMC Bioinformatics* 12 (April): 114.](http://paperpile.com/b/FLjKLc/9jTEo)

[Sanders, Ashley D., Ester Falconer, Mark Hills, Diana C. J. Spierings, and Peter M. Lansdorp. 2017. “Single-Cell Template Strand Sequencing by Strand-Seq Enables the Characterization of Individual Homologs.” *Nature Protocols* 12 (6): 1151–76.](http://paperpile.com/b/FLjKLc/tk1Q1)

[Sanders, Ashley D., Mark Hills, David Porubský, Victor Guryev, Ester Falconer, and Peter M. Lansdorp. 2016. “Characterizing Polymorphic Inversions in Human Genomes by Single-Cell Sequencing.” *Genome Research* 26 (11): 1575–87.](http://paperpile.com/b/FLjKLc/9HaiA)

[Schrider, Daniel R., David Houle, Michael Lynch, and Matthew W. Hahn. 2013. “Rates and Genomic Consequences of Spontaneous Mutational Events in Drosophila Melanogaster.” *Genetics* 194 (4): 937–54.](http://paperpile.com/b/FLjKLc/CrJVv)

[Selvaraj, Siddarth, Jesse R Dixon, Vikas Bansal, and Bing Ren. 2013. “Whole-Genome Haplotype Reconstruction Using Proximity-Ligation and Shotgun Sequencing.” *Nature Biotechnology* 31 (12): 1111–18.](http://paperpile.com/b/FLjKLc/akJrT)

[Serre, David, Scott Gurd, Bing Ge, Robert Sladek, Donna Sinnett, Eef Harmsen, Marina Bibikova, et al. 2008. “Differential Allelic Expression in the Human Genome: A Robust Approach to Identify Genetic and Epigenetic Cis-Acting Mechanisms Regulating Gene Expression.” *PLoS Genetics* 4 (2): e1000006.](http://paperpile.com/b/FLjKLc/ouOuL)

[Steinberg, Karyn Meltz, Tina Graves-Lindsay, Valerie A. Schneider, Mark J. P. Chaisson, Chad Tomlinson, John L. Huddleston, Patrick Minx, et al. 2016. “High-Quality Assembly of an Individual of Yoruban Descent.” *bioRxiv*. doi:](http://paperpile.com/b/FLjKLc/vNE8V)[10.1101/067447](http://dx.doi.org/10.1101/067447)[.](http://paperpile.com/b/FLjKLc/vNE8V)

[Storey, John D., Jonathan E. Taylor, and David Siegmund. 2004. “Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 66 (1). Blackwell Publishing: 187–205.](http://paperpile.com/b/FLjKLc/zrGta)

[Stranger, Barbara E., Stephen B. Montgomery, Antigone S. Dimas, Leopold Parts, Oliver Stegle, Catherine E. Ingle, Magda Sekowska, et al. 2012. “Patterns of Cis Regulatory Variation in Diverse Human Populations.” *PLoS Genetics* 8 (4): e1002639.](http://paperpile.com/b/FLjKLc/cdNE)

[Strimmer, Korbinian. 2008. “A Unified Approach to False Discovery Rate Estimation.” *BMC Bioinformatics* 9 (July): 303.](http://paperpile.com/b/FLjKLc/pG3A)

[Sudmant, Peter H., Jacob O. Kitzman, Francesca Antonacci, Can Alkan, Maika Malig, Anya Tsalenko, Nick Sampas, et al. 2010. “Diversity of Human Copy Number Variation and Multicopy Genes.” *Science* 330 (6004): 641–46.](http://paperpile.com/b/FLjKLc/JjH2E)

[Sudmant, Peter H., Swapan Mallick, Bradley J. Nelson, Fereydoun Hormozdiari, Niklas Krumm, John Huddleston, Bradley P. Coe, et al. 2015. “Global Diversity, Population Stratification, and Selection of Human Copy-Number Variation.” *Science* 349 (6253): aab3761.](http://paperpile.com/b/FLjKLc/MYq8z)

[Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015. “An Integrated Map of Structural Variation in 2,504 Human Genomes.” *Nature* 526 (7571): 75–81.](http://paperpile.com/b/FLjKLc/IVdOB)

[Sugathan, Aarathi, Marta Biagioli, Christelle Golzio, Serkan Erdin, Ian Blumenthal, Poornima Manavalan, Ashok Ragavendran, et al. 2014. “CHD8 Regulates Neurodevelopmental Pathways Associated with Autism Spectrum Disorder in Neural Progenitors.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (42): E4468–77.](http://paperpile.com/b/FLjKLc/5d5wj)

[Tarasov, Artem, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. 2015. “Sambamba: Fast Processing of NGS Alignment Formats.” *Bioinformatics*  31 (12): 2032–34.](http://paperpile.com/b/FLjKLc/1etRk)

[Tischler, German, and Steven Leonard. 2014. “Biobambam: Tools for Read Pair Collation Based Algorithms on BAM Files.” *Source Code for Biology and Medicine* 9 (1): 13.](http://paperpile.com/b/FLjKLc/uhCLr)

[Wang, Liguo, Shengqin Wang, and Wei Li. 2012. “RSeQC: Quality Control of RNA-Seq Experiments.” *Bioinformatics*  28 (16): 2184–85.](http://paperpile.com/b/FLjKLc/ynHf5)

[Wen, Xiaoquan, Francesca Luca, and Roger Pique-Regi. 2015. “Cross-Population Joint Analysis of eQTLs: Fine Mapping and Functional Annotation.” *PLoS Genetics* 11 (4): e1005176.](http://paperpile.com/b/FLjKLc/wUcI)

[Wu, Thomas D., and Serban Nacu. 2010. “Fast and SNP-Tolerant Detection of Complex Variants and Splicing in Short Reads.” *Bioinformatics*  26 (7): 873–81.](http://paperpile.com/b/FLjKLc/7uLlX)

[Ye, Kai, Marcel H. Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. 2009. “Pindel: A Pattern Growth Approach to Detect Break Points of Large Deletions and Medium Sized Insertions from Paired-End Short Reads.” *Bioinformatics*  25 (21): 2865–71.](http://paperpile.com/b/FLjKLc/C9XAw)

[Zhao, Xuefang, Sarah B. Emery, Bridget Myers, Jeffrey M. Kidd, and Ryan E. Mills. 2016. “Resolving Complex Structural Genomic Rearrangements Using a Randomized Approach.” *Genome Biology* 17 (1): 126.](http://paperpile.com/b/FLjKLc/LGzyu)

[Zhao, Xuefang, Alexandra M. Weber, and Ryan E. Mills. 2017. “A Recurrence Based Approach for Validating Structural Variation Using Long-Read Sequencing Technology.” *bioRxiv*. doi:](http://paperpile.com/b/FLjKLc/KiP1p)[10.1101/105817](http://dx.doi.org/10.1101/105817)[.](http://paperpile.com/b/FLjKLc/KiP1p)

need link here

Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Mills, R. E., Pittard, W. S., 1000 Genomes Project Consortium & Devine, S. E. The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. In Revision, Genome Research (2017).

Xuefang, the paper should be in press very soon. Will update when it is officially up at Genome Research.