**Multi-platform discovery of haplotype-resolved structural variation in human genomes**

Author list at the end of the manuscript.

**ABSTRACT**

The incomplete identification of structural variants from whole genome sequencing data has limited our understanding of human genetic diversity and disease association. Here, we apply a suite of long- and short-read, strand-specific sequencing technologies and optical mapping and variant discovery algorithms to comprehensively analyze three human parent-child trios to define the full spectrum of human genetic variation in a haplotype-resolved manner. We identify, an average of 878,452 indels (1-49 bp) and 31,599 structural variants (> 50 bp) per human genome, a 7-fold increase in structural variation compared to previous studies, including from the 1000 Genomes Project. We also discovered more than 300 inversions per genome—most of which previously escaped detection- and including complex and unbalanced chromosomal rearrangements. We provide near-complete, haplotype-resolved structural variation for three genomes that can now be used as gold standard for the scientific community and we make specific recommendations for maximizing structural variation sensitivity for future large-scale genome sequencing studies.

**INTRODUCTION**

Structural variants (SVs) contribute greater diversity at the nucleotide level between two human genomes than any other form of genetic variation [(Conrad et al. 2010a)](https://paperpile.com/c/TK681s/1tukZ); [(Kidd et al. 2010)](https://paperpile.com/c/TK681s/m5n3c) [(Korbel et al. 2007; Sudmant et al. 2015a)](https://paperpile.com/c/TK681s/cstab%2BPnVhc). To date, such variation has been difficult to identify and characterize from the large number of human genomes that have been sequenced using short-read, high-throughput sequencing technologies because the methods underlying SV detection in these datasets are dependent, in part, on indirect inferences (e.g., read-depth and discordant read-pair mapping). The limited number of variants observed directly (e.g., split-read approaches [(Rausch et al. 2012a; Kronenberg et al. 2015; Ye et al. 2009a)](https://paperpile.com/c/TK681s/nClL%2BU2WJ%2BOq9Z) is constrained by the shorter length of these sequencing reads. Moreover, while larger copy-number variants (CNVs) could be identified using microarray and read-depth approaches, smaller events (<5 kb) and balanced events such as inversions remain poorly ascertained [(Sudmant et al. 2015a; Chaisson et al. 2015)](https://paperpile.com/c/TK681s/PnVhc%2BX9YLD).

One fundamental problem for SV detection from short-read sequencing is inherent to the predominant data type: paired-end sequences of relatively short fragments that are aligned to a consensus reference. The approach can be effective in unique sequences, but breaks down within repetitive DNA, which are highly enriched for SVs [(Sharp et al., 2005, Conrad et al. 2010b)](https://paperpile.com/c/TK681s/jWx4G). Another fundamental problem is that most SV discovery methods do not indicate which haplotype background a given SV resides on (i.e., haplotype unaware because of the mixture of DNA reads from the two chromosomal homologs). Nevertheless, SVs are 3-fold more likely to associate with a GWAS signal than single nucleotide variants (SNVs) and larger SVs (> 20kbp) are up to 50-fold more likely to affect the expression of a gene compared to an SNV [(Sudmant et al. 2015a; Chaisson et al. 2015)](https://paperpile.com/c/TK681s/PnVhc%2BX9YLD). Hence, SVs that remains cryptic to current sequencing approaches may represent an important source of disease-causing variation in unsolved Mendelian disorders and a component of the missing heritability in complex disorders [(Manolio et al. 2009)](https://paperpile.com/c/TK681s/S2NPK).

In this study, we sought to comprehensively determine the complete spectrum of human genetic variation in three family trios. We overcome the barriers to SV detection from conventional approaches by integrating a suite of cutting-edge genomic technologies that when used collectively, allow SVs to be assessed in a near-complete, haplotype-aware manner in diploid genomes. In addition, we also sought to identify the optimal combination of technologies and algorithms that would maximize sensitivity and specificity for SV detection for future genomic studies.

**RESULTS**

The goal of this study was to comprehensively discover, sequence-resolve and phase all non-single nucleotide variation (SNV) in a select number of human genomes. We chose three parent-child trios (mother, father, child) for comprehensive SV discovery: a Han Chinese (CHS) trio (HG00513, HG00512, HG00514), a Puerto Rican (PUR) trio (HG00732, HG00731, HG00733) and a Yoruban (YRI) Nigerian trio (NA19238, NA19239, NA19240). The Han Chinese and Yoruban Nigerian families were representative of low and high genetic diversity genomes, respectively, while the Puerto Rican family was selected to reflect population admixture. The parents of each trio had been previously sequenced as part of the 1000 Genomes Project phase 3 [(1000 Genomes Project Consortium et al. 2012b)](https://paperpile.com/c/TK681s/DERbz) and the child from each trio has been selected for the development of new human reference genomes (Chaisson et al., Nat Rev Genet 2015). As a result, extensive genomic resources, such as SNV & SV callsets, SNP microarray data, sequence data and fosmid/BAC libraries have been developed to establish these trios as “gold standards” for SV assessment. We focused primarily on the three children for SV discovery using parental material to assess transmission and confirm phase.

We developed a multi-scale mapping and sequencing strategy using different technologies to target short and long-range sequence information for detecting sequence variation of different types and sizes. To maximize sensitivity, we sequenced each child’s genome to a combined physical coverage of 582-fold (Supplemental Table 1), using various short and long-read technologies (Table 1). We discovered SVs using Illumina (IL) short read whole-genome sequencing (WGS) data, 3.5 kb and 7.5 kb jumping libraries, long-read sequencing and mapping using PacBio (PB) (Menlo Park, CA) and BioNanoGenomics (BNG) optical mapping technology (La Jolla, CA). We also applied a series of genomic technologies capable of obtaining long-range phasing and haplotype structure: 10X Chromium (CHRO) (Pleasanton, CA), Illumina synthetic long read (IL-SLR a.k.a. as Moleculo), Hi-C [(Lieberman-Aiden et al. 2009)](https://paperpile.com/c/TK681s/ks17l), and single cell/single strand genome sequencing (Strand-seq) [(Falconer et al. 2012)](https://paperpile.com/c/TK681s/bU5mS) technologies (Table 1, Supplementary Table 1, Supplemental Methods).

***Chromosomal level phasing & assembly of genomes***. Assembly-based SV discoveries are usually represented as a single haplotype, rather than differentiating the two haplotypes of a diploid cell. This leads to reduced sensitivity for SV detection (Huddleston et al. 2016a). We therefore aimed to develop both haplotypes for the three children in this study by partitioning reads by haplotype and thereby detecting SVs in a haploid-specific manner. We applied WhatsHap [(Martin et al. 2016a)](https://paperpile.com/c/TK681s/Qq9lP) to IL paired-end, IL-SLR, and PB reads; StrandPhaseR [(Porubsky et al. 2017)](https://paperpile.com/c/TK681s/aiw6H) to Strand-seq data and LongRanger to CHRO data and compared them to more traditional trio-based [(Martin et al. 2016b)](https://paperpile.com/c/TK681s/HGusT) and population-based [(Loh et al. 2016)](https://paperpile.com/c/TK681s/dJYh1) phasing methods. As expected, the observed phased block lengths and marker densities differed substantially among the platforms (Figure 1a) but the amount of phasing inconsistencies, as measured by switch error rates (Figure 1c), was found to be very low (from 0.029% for 10X Genomics to 1.4% for Hi-C). Since no single technology alone achieved the density, accuracy, and chromosome-spanning haplotyping necessary to comprehensively identify and assemble SVs throughout the entire human genome [(Edge, Bafna, and Bansal 2017)](https://paperpile.com/c/TK681s/S1jU)[(Ben-Elazar, Chor, and Yakhini 2016)](https://paperpile.com/c/TK681s/rZss)[(Porubský et al. 2017)](https://paperpile.com/c/TK681s/6lrF), we systematically charted the performance of all possible combinations of technologies. When combining a dense, yet local, technology (such as PB or CHRO) with a chromosome-scale, yet sparse, technology (such as Hi-C or Strand-seq), we obtained dense and global haplotype blocks (Figure 1d,e, Supplementary Material). To verify the correctness of chromosome-spanning haplotypes, we computed the mismatch error rates between the largest block delivered by each combination of technologies and the trio-based phasing (Figure 1f). The combination of Strand-seq and CHRO data showed the lowest mismatch error rate (0.23%), while phasing 96.5% of all heterozygous SNVs as part of the largest, chromosome-spanning haplotype block (Supplementary Table 2).

Once chromosomal-level phasing was obtained for each child’s genome, we partitioned the PB reads according to haplotype. On average, 67% of reads could be haplotype-partitioned in each child (Supplementary Table 4). We then developed two complementary approaches to assemble the haplotype-partitioned reads: (1) an extension to the SMRT-SV method [(Huddleston et al. 2016a)](https://paperpile.com/c/TK681s/M0WfC) (Phased-SV) which produced a separate assembly for each haplotype, and (2) an unguided approach (Ms-PAC) which combined separate haplotype specific assemblies with standard *de novo* assemblies in autozygous regions (Supplementary Material). The assemblies covered on average 92.3% of the euchromatic genome (Supplementary Table 5), and produced contig N50 lengths ranging between 1.29 and 6.94 Mbp (Supplementary Table 6). We then generated a high quality consensus sequence for both assembled haplotypes [(Chin et al. 2013)](https://paperpile.com/c/TK681s/z1YGj) from which indels and SVs could be systematically discovered by mapping the contigs to the human reference.

In addition to providing a physical framework for phasing of all genetic variants, the parent-child trio data also allowed us to refine meiotic breakpoints. Using Strand-seq data, meiotic breakpoints could be determined to a median resolution of less than 25 kb (Supplementary Material). The application of trio-aware phasing from PB reads (Garg et al., 2016), further narrowed the median resolution of the meiotic breakpoints to ~ 1.5 kb (Supplementary Table 3). As expected, we observed excess of maternal meiotic recombination events (Supplementary material; Broman et al. 1998; Hou et al. 2013; Kirkness et al. 2013; Lu et al. 2012). Further analysis of fine mapped meiotic breakpoints showed significant enrichment for Alu elements and a slight increase in abundance for THE1-A,B elements as reported by Myers et al., (2008). (Figure S2.3.3A,B). In addition, motif analysis of these breakpoints revealed a 15-mer motif similar to one presented by Myers et al. (2008) (Figure S2.3.3C).

***Indel Discovery (1 - 49 bp)***: We generated a multi-platform indel callset by merging the IL and PB based callsets. Indels were detected in the IL WGS reads using GATK [(DePristo et al. 2011)](https://paperpile.com/c/TK681s/GdDsr), FreeBayes [(Garrison and Marth 2012)](https://paperpile.com/c/TK681s/ObL74) and Pindel [(Ye et al. 2009b)](https://paperpile.com/c/TK681s/VxHbF) algorithms, and merged according to 50% reciprocal overlap, yielding on average 667,629 bi-allelic indel variants per child (Supplementary Methods 4.1). To detect indels from the PB data, assembly contigs were mapped to the reference and indels from separate haplotypes were merged with a threshold of 80% overlap. The results from a single assembly method (Phased-SV) were used to minimize artifacts from contig consensus error. The IL and PB-based indel callsets showed similar size-spectrum distributions (Figure 2b) and were merged to yield, on average, 878,452 indels per individual. The unified indel callset showed the predictable 2-bp periodicity owing to the hypermutability of dinucleotide short tandem repeats (STRs; [Mills et al. 2006)](https://paperpile.com/c/TK681s/BNoPm). However, a greater number of PB indels were discovered for variants greater than 15 bp (55% on average for insertions and 48% on average for deletions) but PB reads lacked the ability to detect 1 bp indels.

**SV Discovery (>50 bp)**. We obtained a unified SV callset for each child from high-coverage IL WGS sequence data, PB SMS reads, and BNG assembly maps*.* To detect SVs in the IL data, we independently applied 12 CNV and SV calling algorithms: WHAMG, LUMPY, DELLY, dCGH, GenomeSTRiP, VariationHunter, ForestSV, Manta, SVelter, Manta, Pindel and NovoBreak (Supplementary Table 8). The jumping library sequences were analyzed using a customized pipeline for large-insert WGS (liWGS) and Delly, and mobile element insertions (MEI) were discovered using a set of specialized callers designed to detect retrotransposons: MELT, Tardis and retroCNV. Unlike the previous 1000 Genomes Phase 3 study, we sought to maximize discovery and did not strictly control for a given false discovery rate, opting to filter calls using orthogonal data in later steps. These callsets were integrated into a unified IL integrated SV (IL-SV) callset (Supplementary Material and Methods) resulting in an average of 10,636 IL-SVs per individual (6,808 deletions, 3,035 insertions and 793 duplications) for a total of 19,080 non-redundant IL-SVs across the three children, comprising of 11,995 deletions, 5,611 insertions, and 1,474 duplications. Approximately half of the SV calls were annotated as high-confidence calls from a single algorithm (singletons) (Figure 2).

We generated a second set of SVs for each trio using the haplotype-resolved Phased-SV and MS-PAC assemblies generated from the long-read sequencing data. Each assembly was mapped to GRCh38, and SVs were detected as insertions, deletions, and inversions within the alignments. After applying a read-based consistency check to remove assembly and alignment artifacts, the SVs from each assembly were merged into a per-individual unified callset (PB-SV). We validated PB-SV calls by searching for evidence of each SV in the long read sequencing data from each of the parents (Supplementary Material). We determined that 93.9% of homozygous and 95.7% of the heterozygous calls showed transmissions consistent with Mendelian inheritance. Excluding inversions, the integrated PB-SV callset consisted of an average of 31,599 PB-SVs per child (12,680 deletions and 18,919 insertions) for a total of 56,285 non-redundant PB-SVs across the three children, comprising 23,736 nonredundant deletions and 32,549 nonredundant insertions. Most of the increase in sensitivity (3 fold) from the PB-callset occurred as a result of better access to intermediate sized SVs (50 bp to 500 bp) and more broadly to the sequence resolution of insertion events across all SV sizes.

Because a substantial fraction of human genetic variation occurs in regions of segmental duplication [(Bailey and Eichler 2006)](https://paperpile.com/c/TK681s/z0ByZ), which are often missing from *de novo* assemblies [(Chaisson, Wilson, and Eichler 2015)](https://paperpile.com/c/TK681s/poV6A), we compared the variation detected in regions of segmental duplication through read depth to the segmental duplications that were resolved in the Phased-SV and MS-PAC *de novo* assemblies. The haplotype-specific *de novo* assemblies resolved only 24.9% (43.6 Mb/175.4 Mb) of human segmental duplications. The dCGH and GenomeStrip methods detected variation through changes in read-depth, and are sensitive to copy number changes in highly duplicated regions. We determined that 93.8% and 73% of the copy-number variable bases detected by dCGH and GenomeStrip, respectively, were not resolved by *de novo* assembly (Supplementary Material and Methods). We also estimated that, on average, ~341 genes per child had at least one exon affected by a copy number change that was not detected in the *de novo assemblies*, highlighting the importance of read-depth based CNV detection even when modern SMS based *de novo* assemblies are generated.

**Characterization of inversions:** Polymorphic inversions are ill-defined by human genome sequencing [(Redin et al. 2017; Talkowski et al. 2012)](https://paperpile.com/c/TK681s/2CoD%2BaYvTq), in part, because larger events tend to be flanked by virtually identical duplicated sequences that can exceed a Mb in length [(Kidd et al. 2010)](https://paperpile.com/c/3m8jUP/IRfO). The breakpoints of these inversions, thus, cannot be bridged by mainstream DNA sequencing technology hampering their discovery. Moreover, the copy-neutral nature of basic inversions precludes detection by read depth analysis. To generate a map of inversions across different length scales, we called inversions with five different but complementary techniques, including Illumina WGS, liWGS, PacBio, optical mapping and Strand-seq. A careful comparison of inversion calls made by these diverse platforms revealed that for all, but Strand-seq, acceptable accuracy was achieved only for calls supported by at least two platforms (**Fig. 3A and Supplemental Tables S6.3.2 & S6.3.3**). For Strand-seq, we developed a novel computational approach integrating inversion discovery with phasing data to bolster accuracy (see **Methods**) and retain all calls that displayed haplotype support. The unified non-redundant inversion callset comprised 308 inversions across the nine individuals. Out of these, 75% of these were either primarily identified by Strand-seq (n=173) or received additional Strand-seq-genotype support (n=59) (**Methods**). By comparison, 132 inversions in the unified callset were detected by Illumina WGS, 131 in PacBio, 92 in liWGS and 29 in the BNG data.

The inversion size spectrum differed markedly among platforms (**Fig. 3B**). Illumina, PacBio, liWGS and BNG excelled in mapping relatively small inversions (<50 kb), wherever breakpoint junctions could be traversed by DNA sequencing technology. Indeed, the smallest inversions (< 2 kb) were only detected by Illumina WGS and PacBio. In contrast, larger inversions (>50 kb) were nearly exclusively detected by Strand-seq. The technique offers the advantage of inversion detection solely by identifying DNA sequene strand switches internal to the breakpoints, readily identifying inversions flanked by large segmental duplications that can be neither assembled nor traversed using standard DNA sequencing technologies (Sanders et al. 2016). Inversions called by Strand-seq show a median size of 70 kb (up to 3.9 Mb in length), in sharp contrast to Illumina detected events whose median size is 3 kb (down to 263 bp in length) (**Supplementary Table S6.3.6**). Given that Strand-seq identified repeat-flanked inversions largely inaccessible by other technologies, the majority of variable nucleotides mediated by inversions was contributed by this technique alone (64% out of 37.5 Mb total inverted bases).

Within the unified inversion callset, 74.4% (229/308) represent copy-neutral (*i.e.* simple) events, whereas 79 are more complex inversions containing embedded copy-number variation (most in the form of inverted duplications). On average, each individual genome shows 114 simple inversions, 68.5% of which are heterozygous and 31.5% homozygous. Chromosomes 7 (3.2%), 8 (3.3%), 16 (5.1%) and X (3.3%) show the highest frequency of inversions, consistent with prior expectation [(Sanders et al. 2016;](https://paperpile.com/c/TK681s/3DyEk) [Sudmant et al. 2015a; Chaisson et al. 2015)](https://paperpile.com/c/TK681s/PnVhc%2BX9YLD). Strikingly, inverted duplications typically exhibit highly variable copy-number states, ranging between 0-10 (mean = 4) copies (Supplementary Table 12), indicating a significant source of genetic variability between individuals. For instance, a 260 kb complex inversion mapping to chromosome 9 (at ~40.8 - 41.1 Mb) contains between 4-6 copies in each genome. Another notable example is an inverted duplication at the *DUSP22* locus (**Fig. 3C**), for which a copy was known to be missing from the human reference [(Genovese et al. 2013)](https://paperpile.com/c/TK681s/UhErZ), and we show to be in the reverse orientation. Additionally, 38 inversions were found to be homozygous in all nine individuals and likely reflect a minor allele or remaining human reference assembly errors (Supplementary Table 14).

**Mobile element insertions:** Previous studies of SV have been unable to resolve the sequences of larger sized repetitive elements in the human genome limiting our ability to assess differences in mutagenic potential between individual genomes. However, since PV-SV reads were routinely larger than 10 kb in length, we used the PB-SV callset to investigate not only the location but the sequence content of full-length L1 elements. We detected an average of 164 full-length L1 elements (FL-L1) with two intact open reading frames in the three children (Supplementary Material). Only 56 of these copies are shared across the three genomes and while the remaining elements are either uniquely found in one genome, or are shared by two genomes (Supplementary Table 10, Supplementary Note). This diversity in source element profiles likely influences L1 mutagenic potential. For example, while all three of the genomes are homozygous for one of the most active L1 elements associated with human cancers (chr22:28663283) and another hyper-actively transposing (i.e., “hot”) L1 [(Brouha et al. 2003)](https://paperpile.com/c/TK681s/aOSiF) active in the germline and cancers (chr1:118852351), each genome also harbors, on average, five unique hot L1 source elements that are found in only one of the three genomes. One of the unique hot L1 copies in the PUR individual is the *LRE3* element, which is the most active L1 source element in humans [(Brouha et al. 2003)](https://paperpile.com/c/TK681s/aOSiF). Twenty-eight FL-L1 copies with low-to-moderate levels of activity also are differentially present in the genomes of the three individuals, including nine copies that are differentially present (Supplementary material). The cumulative differences in L1 mutagenesis that emerge from these diverse FL-L1 profiles suggest that, at a population level, such diversity may translate into differential risk levels for L1-mediated diseases such as cancers and other disorders.

**Genotyping SVs in novel population cohorts:** One of the advantages of having a more comprehensive set of sequence-resolved SVs is the ability to accurately and rapidly genotype them in different human populations. The deletion SVs from the three children were genotyped across the 2,504 low sequencing coverage genomes of the 1000 Genomes project [(1000 Genomes Project Consortium et al. 2015)](https://paperpile.com/c/TK681s/nEtKU%2ByxNUN) using Delly [(Rausch et al. 2012b)](https://paperpile.com/c/TK681s/SpmOf). The SMRT-genotyper method [(Huddleston et al. 2016a)](https://paperpile.com/c/TK681s/M0WfC) was applied to a limited set of 27 high-coverage genomes also from the 1000 Genomes project [(Sudmant et al. 2015a)](https://paperpile.com/c/TK681s/WDHgG%2BPnVhc). After filtering for low genotype quality (< 5) and allele count < 25%, an average of 8,086 SVs per child were genotyped from the IL-SV callset and an average of 3,888 SVs could be genotyped per child from the PB-SV callset. Each genotyped dataset had an average Mendelian error rate of less than 0.64% when assessed against the parent genomes in the trios, and displayed continental population structure through principal component analysis (Supplementary Material). There were no SVs genotyped in this set that showed extreme population stratification (FST > 0.85). When searching for population stratified SVs, a total of 269 SVs were found that had an FST ≥ 0.20. Sixty percent (163/269) were novel and not previously observed in the 1000 Genomes phase 3 callset, including intronic deletions within *TRPV5*, *TMC5* and *DISC1* (Supplementary Table 13). Within the Delly determined genotypes, a greater number of SVs were detected as private variants in the IL-SV callset (382), versus the PB-SV calls (32). A greater fraction of SVs (90.9%) were genotyped using the SMRT-genotyper method compared to the Delly method, however the number of samples genotyped was not sufficient to confidently detect stratified sites. When considering the relatively low genotyping rate and paucity of genotyped private events from PB-SV calls using Delly, and increase in sensitivity using SMRT-Genotyper, this indicates that algorithms for genotyping HTS based (IL) calls have low sensitivity for genotyping calls derived from *de novo* assemblies.

**Gene intersection**: An important consideration of increased sensitivity afforded by this multi-platform approach is its functional relevance. Although the number of individuals compared is few, we sought to address the number of genes that would be disrupted by a SV based on the unified callset. Furthermore, because the children are presumably healthy individuals, one would expect most of the variation to be neutral. One metric is the RVIS percentile which ranks genes from least tolerant of mutation (0%) to most (100%) [(Petrovski et al. 2013)](https://paperpile.com/c/3m8jUP/xfau). The median RVIS percentile of deletion SV affecting exons was 91% indicating most of the variation detected occurs in regions that are generally tolerant to mutation. There are, however, notable exceptions in each individual. Frameshift deletions were identified, for example, in *SNED1* (RVIS 4.36%, 208 bp deletion in both the PUR and YRI children),*COL6A2* (RVIS 8.22%, 124 bp deletion), and full exon or gene deletions were found for nine genes with RVIS < 20%: *ACTN2*, *NALCN*, *PRPF6*, *SLC25A24*, *HP*, *USP28*, *MLTT4*, *ARGHAP39*, and *SRL* (Supplemental Material). Similarly, 63% of exonic insertions were highly tolerant to mutation (>90% RVIS percentile) although three insertion events were detected in exons from genes with RVIS < 20%: *CPSF1* (112 bp insertion), *SULF2* (172 bp insertion), and *PRKG2* (127 bp insertion).

Variation in UTR sequences can have significant effect on gene expression and phenotype, prompting us to overlay our SV dataset with UTRs. Each child had an average of 128 genes with deletions in either the 5’ or 3’ UTR and 94 genes with insertions in the 5’ or 3’ UTR. The variation in UTR sequences was typically in genes less tolerant to variation, with a median RVIS percentile of 54%. There were 14 genes with UTR deletions and RVIS percentile ≤20: *DGKD*, *SON*, *ATP11A*, *PHRF1*, *SNED1*, *ZNF862*, *VPS53*, *DDX46*, *SLC25A23*, *FLI1*, *GLB1L2*, *LMX1B*, *IQSEC3*,and *SEC14L1*. The mean length of these UTR deletion variants was 176 bp, and is similarly reflected in the technology bias for sensitivity; only one event was detected in BNG, five in IL-SV, and 12 in PB-SV.

Simple inversions overlapped with 504 known canonical genes (Table S6.1.4), of which 88% have isoforms entirely contained within an inversion (441), and 6% (32) are intronic. The remaining events are potentially gene disrupting, where 3 events overlap at least one exon (*AQPEP*, *PTPRF*, *TSPAN8*). Up to 55 genes have at least one isoform that spans one of the breakpoints of the inversion, however the majority (95%) of these genes reside in segmental duplications where the exact breakpoints of the inversions are difficult to detect.

We counted the overlap of SVs with functional non-coding DNA (fnDNA): specifically with 2.86M conserved elements (CE) and 1.07M transcription factor binding sites (TFBS). Deletion variants overlapped an average of 6,010 CEs and 2,276 TFBS in each child. However, small SVs rarely affected fnDNA: the median and average size of SVs that overlapped CEs were 1,861 bp and 23,187 bp, respectively and the median and average size of SVs that overlapped TFBS were 3,774 bp and 51,390 bp, respectively. Consequently, this indicates that the majority of fnDNA affecting variants may be detected in current IL based sequencing studies which in this study detected 89.7% of the CE, and 95.3% of the TFBS deletions.

***Platform comparisons and optimal indel and SV detection***: The use of orthogonal technologies and various discovery algorithms on the same DNA samples provides an opportunity for a systematic assessment of the performance of individual as well as combinations of approaches for indel and SV detection. While long-read technology generally outperforms IL-based approaches for indel detection by ~50% for indels ≥ 15 bp, it is not reliable for single base deletions even at 40-fold sequence coverage, due to missed SMS incorporations particularly in homopolymer regions. Benchmarking against the unified-indel dataset, we find that maximum sensitivity for IL-indels requires application of three callers including GATK, Freebayes and Pindel (which has a higher false positive rate).

The cost and throughput of long-read sequence reads makes it currently impractical for application to large-scale clinical studies for SV detection. We, therefore, used the pan-SV callset to gauge the sensitivity and specificity of individual and combinations of IL-only algorithms in order to provide guidance to studies that rely solely on short-insert IL sequencing data.Across the entire IL-SV dataset, the deletion concordance to the pan-SV callset was 61.6%, but was unaffected by size, while the insertion concordance to the pan-SV callset was 54.3%, decreasing in sensitivity with increased insertion SV length (Supplementary Table 15). When considering individual methods, the average concordance for deletion calls ranged from 24.8 to 91.7% with a median of 79.2% (Figure 4a), and for insertion calls ranged from 4.2% to 83.8% (Supplementary Table - 161).

It has been shown previously that sensitivity for true SV calls (generated from Illumina datasets) can be improved by combining calls from more than one algorithm [(Manolio et al. 2009; Mohiyuddin et al. 2015; Mills et al. 2011)](https://paperpile.com/c/TK681s/S2NPK%2Bsfs1e%2Bs99m). Because it would be computationally burdensome for large-scale WGS studies to run all available algorithms on the same sequencing dataset, we considered all possible combinations (up to three algorithms) and tested two separate conditions: (1) accept all SV calls from the union of two algorithm callsets and (2) accept all SV calls from at least two of three algorithms. The first condition is targeting maximal sensitivity, while the latter, specificity. Other more elaborate combinations of methods (e.g., at least two union of three) were deemed too low of sensitivity or specificity to consider. We used the nonconcordance rate (NCR) (1- concordance) as a proxy for false discovery rate (FDR). Considering the YRI child, NA19240, there were 7 combinations of two methods with an NCR less than 10% (Supplementary material). The combination of Lumpy, Manta and SVelter (LMSV) (Supplementary Note) (Figures 4 b,c) was able to detect 75.5% of MEI from the pan SV callset, 24% of simple tandem repeat and VNTR deletion SVs, and 75.2% of deletions. No union of methods was below a 10% NCR for insertions, and the optimal combination of methods with no more than 20% NCR was Manta and MELT. This combination detects 17.7% of all insertion SVs, including 56.6% of MEI, 5.73% of tandem repeats, and 8.93% of more complex insertions. Only 3% of insertion SV intersecting exons and UTRs was detected with this combination. Extensive measurements of the sensitivity and specificity of the various combinations are given in Supplemental Table 17. We observed that while no single SV are called by every algorithm tested, there are often algorithms that call similar variants (Figure 4d). This is confirmed by a principal component analysis applied to the genotypes of the integrated callset (Figure 4e).

**SUMMARY AND DISCUSSION**

This study represents the most comprehensive assessment of structural variation in human genomes to date. We employ multiple state-of-the-art sequencing technologies and methods to capture the full spectrum of haplotype-aware genetic variation down to the single nucleotide level. Our results indicate that for maximum sensitivity related to SV discovery, it is essential to employ more than one complementary technology and SV calling algorithm. The PB SMS, Strand-seq, and CHRO data were combined to generate haplotype-resolved *de novo* assemblies constructed from phased SMS reads. When paired with high coverage Illumina sequencing and BioNanoGenomics SV, we discovered ~7 fold more variation than current high-coverage IL only WGS datasets (Sudmant et al. 2015c); on average, 667,629 indels (1-49 bp) and ~31,713 SVs, including 114 inversions per person. Consistent with increased genetic diversity among African populations [(1000 Genomes Project Consortium et al. 2015)](https://paperpile.com/c/3m8jUP/4Fqo), we observed 17.5% more deletion and 12.8% more insertion variants in the Yoruban child when compared to the Han Chinese child.

The long insert sequence data provided us with an unprecedented view of genetic variation in the human genome. Using ~15 kb reads at an average of 40-fold sequence coverage per child, we have been able to span areas of the genome that were previously opaque and discover 3-4 fold more structural variation when compared to short-read sequencing platforms. Our analysis suggests that the vast majority (~92%) of the insertions are being missed by routine short-read calling algorithms (Figure 2b). Specifically, the largest gain stems from tandem repeat and retro-transposon insertions in the 50 bp to 2 kb size range. Inversions represent another problematic class of human genetic variation. In the 1000 Geomes Project, we identified a total of 786 inversions (from 2,504 genomes) with an estimated sensitivity and specificity of 32% and 67% respectively. In the current study, using data from five different platforms (IL, PB, Strand-seq, Jumping Libraries and BNG), we have identified a total of 1,296 additional inversions from the three trios. This increase in sensitivity depended on the complementary nature of the five different technologies (Fig 3B). In the shorter size range, most of the inversion discovery depends on a combination of ILL and PB datasets, whereas for the larger events, Strand-seq was required. As a result, we identified 109 new inversions that were missed as part of the 1000 Genomes Project Phase 3). Surprisingly, most of these are large (>50 kb) constituting up to 34.2 Mb of inverted DNA per diploid genome, and, thus, a 480-fold increase in inverted bases relative to the previous study. Our results indicate that for maximum sensitivity and specificity related to SV discovery, it is essential to employ more than one detection algorithm and more than one complementary technology.

It is not practical for large-scale studies to detect variation by employing the menagerie of sequencing methods and algorithms used in this study. Instead, these data serve as a guide for the trade-off between the cost of sequencing and desired sensitivity for SV detection. For example, we demonstrated entire chromosomal phasing using the Strand-seq and CHRO libraries, however the Strand-seq method is not as universally available as Hi-C, which when combined with CHRO libraries provides chromosome-arm phasing and is likely sufficient for many applications. Similarly, while the typical IL-based WGS sequencing experiment detects roughly ~45% of the total number of SVs, the majority of the missed variants associated with coding portions of the genome appear to be neutral in effect, as predicted by RVIS score. There was a two-fold increase in variation in UTR sequences detected in the PB-SV dataset relative to IL-SV, thus variants causing more subtle effects such as expression changes may be missed without this approach. For variants in the noncoding portions of the human genome, the PB-SV callset increases yield by 5%-11% across conserved CEs and TFBS.

The diagnostic yield in Mendelian studies is far from complete (Chong et al. 2015). Due to the current cost of SMS, we propose that future disease studies have a triaged application of multiple technologies to comprehensively identify SVs. Families that have been sequenced using IL-based WGS should be analyzed using multiple SV calling algorithms (e.g., Lumpy, Manta, SVelter for deletion detection, and Manta, MELT for insertion detection) to gain a ~15% increase in sensitivity over individual methods. Because a disproportionate amount of variation occurs in segmental duplications, it is additionally necessary to apply read depth based methods (e.g., dCGH or Genome STRiP) to detect changes in copy number in highly duplicated regions of the genome. The sequence structure of such variation is still not resolved and novel methods will need to be developed to sequence resolve such CNVs [(Chaisson et al. 2017)](https://paperpile.com/c/3m8jUP/gHs1). Critically, there is a pressing need to reduce the FDR of SV calling to below the current standard of 5% because forward validation of all potentially pathogenic events will be impractical. We predict that a move forward to full-spectrum SV detection using an integrated approach demonstrated in this study will increase the diagnostic yield in patients with genetic disease, SV-mediated mutation and repeat expansions. Moreover, for the proper application of SV detection for patient care, we require a deeper understanding of germline SV from more individuals across diverse global populations.

**METHODS**

In Supplementary material.

**TABLES**

https://docs.google.com/a/uw.edu/spreadsheets/d/1TcyP5LUosTvrGThJZpbSq6KZlO0RltQEDkIO4irGacw/edit?usp=drive\_web

**ACKNOWLEDGEMENTS**

**FIGURES**

**FIGURE 1**



**Figure 1. Characteristics of SNV-based haplotypes obtained from different data sources.** (**a**) Distribution of phased block lengths for the YRI child, NA19240. Note that Strand-seq haplotypes span whole chromosomes and therefore one block per chromosome is shown. For Illumina paired-end data, phased blocks cover less than 50% of the genome and hence the N50 cannot be computed. (**b**) Fraction of phase connection, i.e., pairs of consecutive heterozygous variants, that are provided by each technology (averaged over all proband samples). (**c**) Pairwise comparisons of different phasings; colors encode switch error rates (averaged over all proband samples). For each row, a green box indicates the phasing of an independent technology with best agreement, with corresponding switch error rates given in green. (**d**) Illustration of phased blocks on Chromosome 1 of the YRI child, NA19240. Each phased block is shown in a different color. The largest block is shown in red, i.e., all red regions belong to one block, even though interspaced by white areas (genomic regions where no variants are phased) or disconnected small blocks (different colors). (**e**) Fraction of heterozygous SNVs in the largest block shown in panel ***d***. (**f**) Mismatch error rate of largest block compared to trio-based phasing, averaged over all chromosomes of all proband genomes.

(\*) Not available because trio phasing is used as reference for comparisons.

(\*\*) Not shown as population-based phasing does not output block boundaries; refer to the Supplementary Material for an illustration of errors in population-based phasing.

**FIGURE 2**



**Figure 2. Comparison and integration of IL-SV and PB-SV callsetss** on three daughters. (**a**) Length distribution of PV-SV deletions and insertions (blue), IL-SV (red) and BNG (brown) respectively, together with length distribution of deletions and insertions discovered in the mothers by the 1000 Genomes Project Phase 3 report (silver). (**b**) Number of SVs discovered by one or multiple sequencing platforms in the PUR child, HG0733. (**c**) Overlap of Illumina Indel discovery algorithms, with total number of SVs found by each combination of Illumina algorithms (grey) and those that overlapped with a PB-INDEL (blue) in the CHS child, HG00514.

**FIGURE 3**



**Figure 3.** (**a**) Integration of inversions across platforms based on reciprocal overlap. Shown is an example of five orthogonal platforms intersecting at a homozygous inversion. (**b**) Size distribution of inversions included in the unified inversion list, subdivided by technology. (**c**) Classification of Strand-seq inversions based on orthogonal PacBio phase support. Illustrated are examples of simple (homozygous and heterozygous) and complex (inverted duplication) events. SD- segmental duplications. Ph- phase data.

**Figure 4**



**Figure 4. The concordance and non-concordance of Illumina methods compared against orthogonal technologies for** (**a**) the deletion concordance for individual methods (**b**) the union of all pairs of methods, and (**c**) the requirement that more than one caller agree on any call. Individual callers are shown as red points for comparison. Pairs and triples of combinations are in black points. The solid and dashed lines represent the 5% and 10% not concordant rates (NCR), respectively. The top 5 combinations of methods in each plot below the 10% NCR, along with the individual plots, are each labeled. (**d**) Overlap of Illumina SV discovery algorithms, with total number of SVs found by each combination of Illumina algorithms (grey) and those that overlapped with the PB-SV calls (blue) in the YRI child, NA19240. (**e**) principal component (PC) analysis of the genotypes of concordant calls of each method: PC 1 versus 2 (left), PC 2 versus 3 (*right*).

**REFERENCES**

[1000 Genomes Project Consortium, Goncalo R. Abecasis, Adam Auton, Lisa D. Brooks, Mark A. DePristo, Richard M. Durbin, Robert E. Handsaker, Hyun Min Kang, Gabor T. Marth, and Gil A. McVean. 2012a. “An Integrated Map of Genetic Variation from 1,092 Human Genomes.” *Nature* 491 (7422): 56–65.](http://paperpile.com/b/TK681s/WfBR2)

[———. 2012b. “An Integrated Map of Genetic Variation from 1,092 Human Genomes.” *Nature* 491 (7422): 56–65.](http://paperpile.com/b/TK681s/DERbz)

[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/TK681s/yxNUN)

[Antonacci, Francesca, Jeffrey M. Kidd, Tomas Marques-Bonet, Mario Ventura, Priscillia Siswara, Zhaoshi Jiang, and Evan E. Eichler. 2009. “Characterization of Six Human Disease-Associated Inversion Polymorphisms.” *Human Molecular Genetics* 18 (14): 2555–66.](http://paperpile.com/b/TK681s/az3HT)

[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/TK681s/z0ByZ)

[Ben-Elazar, Shay, Benny Chor, and Zohar Yakhini. 2016. “Extending Partial Haplotypes to Full Genome Haplotypes Using Chromosome Conformation Capture Data.” *Bioinformatics*  32 (17): i559–66.](http://paperpile.com/b/TK681s/rZss)

[Brouha, Brook, Joshua Schustak, Richard M. Badge, Sheila Lutz-Prigge, Alexander H. Farley, John V. Moran, and Haig H. Kazazian Jr. 2003. “Hot L1s Account for the Bulk of Retrotransposition in the Human Population.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (9): 5280–85.](http://paperpile.com/b/TK681s/aOSiF)

[Chaisson, Mark J. P., John Huddleston, Megan Y. Dennis, Peter H. Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, et al. 2015. “Resolving the Complexity of the Human Genome Using Single-Molecule Sequencing.” *Nature* 517 (7536): 608–11.](http://paperpile.com/b/TK681s/X9YLD)

[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/TK681s/poV6A)

[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/TK681s/WDHgG)

[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/TK681s/z1YGj)

[Conrad, Donald F., Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, Jan Aerts, et al. 2010a. “Origins and Functional Impact of Copy Number Variation in the Human Genome.” *Nature* 464 (7289): 704–12.](http://paperpile.com/b/TK681s/1tukZ)

[———. 2010b. “Origins and Functional Impact of Copy Number Variation in the Human Genome.” *Nature* 464 (7289): 704–12.](http://paperpile.com/b/TK681s/jWx4G)

[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/TK681s/GdDsr)

[Edge, Peter, Vineet Bafna, and Vikas Bansal. 2017. “HapCUT2: Robust and Accurate Haplotype Assembly for Diverse Sequencing Technologies.” *Genome Research* 27 (5): 801–12.](http://paperpile.com/b/TK681s/S1jU)

[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/TK681s/bU5mS)

[Garrison, Erik, and Gabor Marth. 2012. “Haplotype-Based Variant Detection from Short-Read Sequencing.” *arXiv [q-bio.GN]*. arXiv.](http://paperpile.com/b/TK681s/ObL74) <http://arxiv.org/abs/1207.3907>[.](http://paperpile.com/b/TK681s/ObL74)

[Genovese, Giulio, Robert E. Handsaker, Heng Li, Nicolas Altemose, Amelia M. Lindgren, Kimberly Chambert, Bogdan Pasaniuc, et al. 2013. “Using Population Admixture to Help Complete Maps of the Human Genome.” *Nature Genetics* 45 (4): 406–14.](http://paperpile.com/b/TK681s/UhErZ)

[Huddleston, John, Mark Jp Chaisson, Karyn Meltz Steinberg, Wes Warren, Kendra Hoekzema, David S. Gordon, Tina A. Graves-Lindsay, et al. 2016a. “Discovery and Genotyping of Structural Variation from Long-Read Haploid Genome Sequence Data.” *Genome Research*, November. doi:](http://paperpile.com/b/TK681s/M0WfC)[10.1101/gr.214007.116](http://dx.doi.org/10.1101/gr.214007.116)[.](http://paperpile.com/b/TK681s/M0WfC)

[———. 2016b. “Discovery and Genotyping of Structural Variation from Long-Read Haploid Genome Sequence Data.” *Genome Research*, November. doi:](http://paperpile.com/b/TK681s/5SUAO)[10.1101/gr.214007.116](http://dx.doi.org/10.1101/gr.214007.116)[.](http://paperpile.com/b/TK681s/5SUAO)

[Kidd, Jeffrey M., Tina Graves, Tera L. Newman, Robert Fulton, Hillary S. Hayden, Maika Malig, Joelle Kallicki, Rajinder Kaul, Richard K. Wilson, and Evan E. Eichler. 2010. “A Human Genome Structural Variation Sequencing Resource Reveals Insights into Mutational Mechanisms.” *Cell* 143 (5): 837–47.](http://paperpile.com/b/TK681s/m5n3c)

[Korbel, Jan O., Alexander Eckehart Urban, Jason P. Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M. Kim, et al. 2007. “Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome.” *Science* 318 (5849): 420–26.](http://paperpile.com/b/TK681s/cstab)

[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/TK681s/U2WJ)

[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/TK681s/ks17l)

[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/TK681s/dJYh1)

[Mallick, Swapan, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, et al. 2016a. “The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations.” *Nature* 538 (7624): 201–6.](http://paperpile.com/b/TK681s/nEtKU)

[———. 2016b. “The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations.” *Nature* 538 (7624): 201–6.](http://paperpile.com/b/TK681s/DXfCA)

[Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, et al. 2009. “Finding the Missing Heritability of Complex Diseases.” *Nature* 461 (7265): 747–53.](http://paperpile.com/b/TK681s/S2NPK)

[Martin, Marcel, Murray Patterson, Shilpa Garg, Sarah O. Fischer, Nadia Pisanti, Gunnar W. Klau, Alexander Schoenhuth, and Tobias Marschall. 2016a. “WhatsHap: Fast and Accurate Read-Based Phasing.” doi:](http://paperpile.com/b/TK681s/Qq9lP)[10.1101/085050](http://dx.doi.org/10.1101/085050)[.](http://paperpile.com/b/TK681s/Qq9lP)

[———. 2016b. “WhatsHap: Fast and Accurate Read-Based Phasing.” doi:](http://paperpile.com/b/TK681s/HGusT)[10.1101/085050](http://dx.doi.org/10.1101/085050)[.](http://paperpile.com/b/TK681s/HGusT)

[Mills, Ryan E., Christopher T. Luttig, Christine E. Larkins, Adam Beauchamp, Circe Tsui, W. Stephen Pittard, and Scott E. Devine. 2006. “An Initial Map of Insertion and Deletion (INDEL) Variation in the Human Genome.” *Genome Research* 16 (9): 1182–90.](http://paperpile.com/b/TK681s/BNoPm)

[Mohiyuddin, Marghoob, John C. Mu, Jian Li, Narges Bani Asadi, Mark B. Gerstein, Alexej Abyzov, Wing H. Wong, and Hugo Y. K. Lam. 2015. “MetaSV: An Accurate and Integrative Structural-Variant Caller for next Generation Sequencing.” *Bioinformatics*  31 (16): 2741–44.](http://paperpile.com/b/TK681s/sfs1e)

[Porubsky, David, Shilpa Garg, Ashley D. Sanders, Jan O. Korbel, Victor Guryev, Peter M. Lansdorp, and Tobias Marschall. 2017. “Dense And Accurate Whole-Chromosome Haplotyping Of Individual Genomes.” doi:](http://paperpile.com/b/TK681s/aiw6H)[10.1101/126136](http://dx.doi.org/10.1101/126136)[.](http://paperpile.com/b/TK681s/aiw6H)

[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/TK681s/6lrF)

[Rausch, Tobias, Thomas Zichner, Andreas Schlattl, Adrian M. Stütz, Vladimir Benes, and Jan O. Korbel. 2012a. “DELLY: Structural Variant Discovery by Integrated Paired-End and Split-Read Analysis.” *Bioinformatics*  28 (18): i333–39.](http://paperpile.com/b/TK681s/nClL)

[———. 2012b. “DELLY: Structural Variant Discovery by Integrated Paired-End and Split-Read Analysis.” *Bioinformatics*  28 (18): i333–39.](http://paperpile.com/b/TK681s/SpmOf)

[Redin, Claire, Harrison Brand, Ryan L. Collins, Tammy Kammin, Elyse Mitchell, Jennelle C. Hodge, Carrie Hanscom, et al. 2017. “The Genomic Landscape of Balanced Cytogenetic Abnormalities Associated with Human Congenital Anomalies.” *Nature Genetics* 49 (1): 36–45.](http://paperpile.com/b/TK681s/2CoD)

[Rheinbay, Esther, Prasanna Parasuraman, Jonna Grimsby, Grace Tiao, Jesse M. Engreitz, Jaegil Kim, Michael S. Lawrence, et al. 2017. “Recurrent and Functional Regulatory Mutations in Breast Cancer.” *Nature* 547 (7661): 55–60.](http://paperpile.com/b/TK681s/6I35e)

[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/TK681s/3DyEk)

[Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015a. “An Integrated Map of Structural Variation in 2,504 Human Genomes.” *Nature* 526 (7571): 75–81.](http://paperpile.com/b/TK681s/PnVhc)

[———. 2015b. “An Integrated Map of Structural Variation in 2,504 Human Genomes.” *Nature* 526 (7571): 75–81.](http://paperpile.com/b/TK681s/4XfwD)

[———. 2015c. “An Integrated Map of Structural Variation in 2,504 Human Genomes.” *Nature* 526 (7571): 75–81.](http://paperpile.com/b/TK681s/W05sJ)

[Talkowski, Michael E., Jill A. Rosenfeld, Ian Blumenthal, Vamsee Pillalamarri, Colby Chiang, Adrian Heilbut, Carl Ernst, et al. 2012. “Sequencing Chromosomal Abnormalities Reveals Neurodevelopmental Loci That Confer Risk across Diagnostic Boundaries.” *Cell* 149 (3): 525–37.](http://paperpile.com/b/TK681s/aYvTq)

[Turner, Tychele N., Fereydoun Hormozdiari, Michael H. Duyzend, Sarah A. McClymont, Paul W. Hook, Ivan Iossifov, Archana Raja, et al. 2016. “Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA.” *American Journal of Human Genetics* 98 (1): 58–74.](http://paperpile.com/b/TK681s/zi2a3)

[Ye, Kai, Marcel H. Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. 2009a. “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/TK681s/Oq9Z)

[———. 2009b. “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/TK681s/VxHbF)

**Authors (in alphabetical order)**

Danny Antaki, University of California San Diego

Peter Audano, University of Washington

Han Cao, BioNano Genomics, Inc.

Mark Chaisson, University of Washington

Jason Chin, Pacific Biosciences, Inc.

Xechen Chong, University of Texas MD Anderson

Laura Clarke, European Bioinformatics Institute

Ryan Collins, Harvard Medical School

Susan Fairley, European Bioinformatics Institute

Xian Fan, University of Texas MD Anderson

Andrew Farrell, University of Utah

Eugene Gardner, University of Maryland

Timur Galeev, Yale University

Li Guo, Xi’an Jiao Tong University

Bob Handsaker, Broad Institute of MIT and Harvard

Alex Hastie, BioNano Genomics

Fereydoun Hormozdiari, University of California, Davis

Chong Lek Koh, University of Malaysia

Xiangmeng Kong, Yale University

Jonas Korlach, Pacific Biosciences, Inc.

Zev Kronenberg, University of Washington

Jee Young Kwan, The Jackson Laboratory for Genomic Medicine

Dillon Lee, University of Utah

Jong Eun Lee, DNA Link, Inc.

Sau Peng Lee, University of Malaysia

Ankit Malhotra, The Jackson Laboratory for Genomic Medicine

Katy Munson, University of Washington

Fabio Navarro, Yale University

Conor Nodzak, University of North Carolina, Charlotte

Andy Pang, BioNano Genomics

David Porubsky, Max Plank Institute for Informatics

Tobia Rausch, EMBL Heidelberg

Oscar Rodriguez, Icahn School of Medicine

Ashley Sanders, EMBL Heidelberg

Stuart Cantislieris, University of Washington

Alistair Ward, University of Utah

Jia Wen, University of North Carolina, Charlotte

AnneMarie Welch, University of Washington

Aaron Wenger, PacBIo

Chengsheng Zhang, The Jackson Labortory for Genomic Medicine

Xuefang Zhao, University of Michigan

**Principal Investigators (in alphabetical order)**

Ali Bashir, Icahn School of Medicine at Mount Sinai

Ken Chen, University of Texas MD Anderson

Scott Devine, University of Maryland

Mark Gerstein, Yale University

Tobias Marschall, Max Plank Institut for Informatics

Gabor Marth, University of Utah

Ryan Mills, University of Michigan

Jonathan Sebat, University of California, San Diego

Mindy Shi, University of North Carolina at Charlotte

Michael Talkowski, Massachusetts General Hospital

Kai Ye, Xi’an Jiao Tong University

**Chairs (in alphabetical order)**

Evan E. Eichler, University of Washington

Jan Korbel, EMBL Heidelberg

Charles Lee, The Jackson Laboratory for Genomic Medicine