Resolving the Full Spectrum of Human Genome Variation using Linked-Reads

Patrick Marks[a], Sarah Garcia[a], Alvaro Martinez Barrio[a], Kamila Belhocine[a], Jorge Bernate[a], Rajiv Bharadwaj[a], Keith Bjornson[a], Claudia Catalanotti[a], Josh Delaney[a], Adrian Fehr[a], Brendan Galvin[a], Haynes Heaton[a,e,f], Jill Herschleb[a], Christopher Hindson[a], Esty Holt[b], Cassandra B. Jabara[a,g], Susanna Jett[a], Nikka Keivanfar[a], Sofia Kyriazopoulou-Panagiotopoulou[a,h], Monkol Lek[c,d], Bill Lin[a], Adam Lowe[a], Shazia Mahamdallie[b], Shamoni Maheshwari[a], Tony Makarewicz[a], Jamie Marshall[d], Francesca Meschi[a], Chris O'keefe[a], Heather Ordonez[a], Pranav Patel[a], Andrew Price[a], Ariel Royall[a], Elise Ruark[b], Sheila Seal[b], Michael Schnall-Levin[a], Preyas Shah[a], Stephen Williams[a], Indira Wu[a], Andrew Wei Xu[a], Nazneen Rahman[b], Daniel MacArthur[c,d], Deanna M. Church[a]

2017-12-11 21:45:57

Author affiliations a: 10x Genomics, 7068 Koll Center Parkway, Suite 401, Pleasanton, CA 94566; b: The Institute of Cancer Research, Division of Genetics & Epidemiology, 15 Cotswold Road, London, SM2 5NG, UK; c: Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; d: Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA; e: Current affiliation, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK; f: Current affiliation, University of Cambridge, Cambridge, UK; g: Current affiliation, Purigen Biosystems, Inc., 5700 Stoneridge Drive, Suite 100, Pleasanton, CA 94588; h: Current affiliation, Illumina, Inc., 499 Illinois Street, Suite 201, San Francisco, CA 94158

Abstract

Large-scale population based analyses coupled with advances in technology have demonstrated that the human genome is more diverse than originally thought. Standard short-read approaches, used primarily due to accuracy, throughput and costs, fail to give a complete picture of a genome. They struggle to identify large, balanced structural events, cannot access repetitive regions of the genome and fail to resolve the human genome into its two haplotypes. Here we describe an approach that retains long range information while harnessing the power of short reads. Starting from only ~1ng of DNA, we produce barcoded short read libraries. The use of novel informatic approaches allows for the barcoded short reads to be associated with the long molecules of origin producing a novel datatype known as 'Linked-Reads'. This approach allows for simultaneous detection of small and large variants from a single Linked-Read library. We have previously demonstrated the utility of whole genome Linked-Reads (lrWGS) for performing diploid, *de novo* assembly of individual genomes (Weisenfeld et al. 2017). In this manuscript, we show the utility of reference based analysis using a single Linked-Read library for full spectrum genome analysis. We demonstrate the ability of Linked-Reads to reconstruct megabase scale haplotypes and to recover parts of the genome that are typically inaccessible to short reads, including phenotypically important genes such as *STRC*, *SMN1* and *SMN2*. We demonstrate the ability of both lrWGS and Linked-Read Whole Exome Sequencing (lrWES) to identify complex structural variations, including balanced events, single exon deletions, and single exon duplications. The data presented here show that Linked-Reads provide a scalable approach for comprehensive genome analysis that is not possible using short reads alone.

Introduction

Our understanding of diversity in the human genome has defied original models that assumed little sequence variation and even less structural diversity (Church et al. 2011; Collins 1998). The human reference assembly, the flagship product of the human genome project (HGP), collapsed sequences from >50 individuals into a single consensus mosaic haplotype representation, and has enabled the field of genomics to prosper (Consortium 2004). Since the completion of the HGP many large scale consortia studies have applied whole genome sequencing to thousands of individuals from diverse populations across the globe (Auton et al. 2015; Lek et al. 2016; Sudmant et al. 2015). Results of these population-based genomic studies have revealed that there is more diversity within the human population than ever anticipated. To date, most genome analyses were performed with accurate, high-throughput short reads leading to robust analysis of small variants but only providing a small window into the prevalence of larger structural variants (SVs). The application of recent technical advances in both sequencing and mapping approaches to genome analysis have revealed that despite extensive information garnered from large population surveys utilizing short read whole genome sequencing (srWGS), we are still likely under-representing the amount of structural variation in the human population (Chaisson et al. 2014; Huddleston and Eichler 2016; Collins et al. 2017).

The prevalence of SVs suggests that individual haplotype reconstruction, rather than haploid consensus analysis is a better approach to the analysis of an individual genome (Church et al. 2011, 2015; Schneider et al. 2017). Indeed, recent work from groups developing population graph-based assembly representations have demonstrated that this approach improves alignment and individual genome reconstruction (Iqbal et al. 2012; Novak et al. 2017). While it has long been recognized that SVs play an important role in highly penetrant Mendelian disorders (Amberger et al. 2015), groups investigating the biological impact of these events have demonstrated that SVs have a more substantial impact on gene expression than do single nucleotide variants (SNVs), and thus may contribute substantially to the development of common disorders (Chiang et al. 2017).

Recent work has shown that adding long range information and resolving long range haplotypes improves sensitivity for SV detection (Huddleston and Eichler 2016; Chaisson et al. 2017). Additionally, reconstructing individual haplotypes has the potential to improve analysis that relies on patterns of genetic variation to extract genotype-phenotype information, such as eQTL analysis (Ramaker et al. 2017). A more complete reconstruction of individual genomes will impact research in both rare and common disease (Chiang et al. 2017).

There are over 600 genes categorized as part of the 'NGS dead zone', where standard exome or genome analysis is limited due to the presence of closely related paralogous sequences (Mandelker et al. 2016). These paralogs limit the ability to produce a high quality alignment due to multiple possible locations for read placement. The failure of short reads to resolve these loci means they are either missing in many high throughput analyses, or require orthogonal approaches for analysis (Askree et al. 2013; Mandelker et al. 2014). Many of these genes are known to be relevant in the study of Mendelian disease, while many others remain uncharacterized due to the inability of short reads to align to these regions.

The limitations of short reads suggest the need for improved methods for genome analysis. Several long molecule sequencing and mapping approaches have been developed to address these issues (Carneiro et al. 2012; Nakano et al. 2017; Genomics 2017). While they provide powerful data for better understanding genome structure, their high input requirements, error rates and costs make them inaccessible to many applications, particularly those requiring thousands of samples. To address this need, we developed a technology that retains long range information while maintaining the power, accuracy, and scalability of short read sequencing. The core datatype, Linked-Reads, is generated by performing haplotype limiting dilution of long DNA molecules into >1 million barcoded partitions, synthesizing barcoded sequence libraries within those partitions, and then performing standard short read sequencing in bulk. The limited amount of DNA put into the system, coupled with novel algorithms allow short reads to be associated with their long molecule of origin, in most cases, with high probability.

The Linked-Read datatype was originally described in (Zheng et al. 2016) using the GemCode™ System. The Chromium™ System represents a substantial improvement over the GemCode™ system. These improvements come from increasing the number of barcodes (737,000 to 4 million), and the number of partitions (100,000 to 1 million) as well as improving the biochemistry to substantially reduce coverage bias. These improvements eliminate the need for an additional short-read library and, when coupled with novel informatic approaches, produce a standalone solution for complete genome analysis.

Here we compare reference based analysis on multiple standard control samples using either a single Chromium Linked-Read library or a standard short read library for both whole genome (WGS) and whole exome sequencing (WES) approaches. We describe additional novel algorithms in our Long Ranger™ reference based pipeline that allow for improved alignment coverage when compared to standard short reads. We then demonstrate the ability to construct multi-megabase haplotypes by coupling long molecule information with heterozygous variants within the sample. We show that a single Chromium library has comparable small variant sensitivity and specificity to standard short read libraries and helps expand the amount of the genome that can be accessed and analyzed. We demonstrate the ability to identify large scale SVs by taking advantage of the long range information provided by the barcoded library. Lastly, we assess the ability to identify variants in archival samples that had been previously assessed by orthogonal methods. These data show that a Chromium Linked-Read library provides a scalable, and more complete genome reconstruction than short reads alone.

Results

Improvements in Linked-Read data

One limitation of the original GemCode approach was the need to combine the Linked-Read data with a standard short-read library for analysis. This was needed to help address coverage imbalances seen in the GemCode library alone. To address this issue we modified the original biochemistry, replacing it with an isothermal amplification approach. The updated biochemistry now provides for more even genome coverage, approaching that of PCR free short-read preps (Figure 1).

Additional improvements include increasing the number of barcodes from 737,000 to 4 million and the number of partitions from 100,000 to over 1 million. This allows for fewer DNA molecules per partition, or GEM (Gelbead-in-EMulsion), and thus a substantially reduced background rate of barcode collisions: the rate at which two random loci occur in the same GEM (Supplemental Figure 1). The lowered background rate of barcode sharing increases the probability of correctly associating a short read to the correct molecule of origin, and increases the sensitivity for SV detection.

Improved Genome and Exome Alignments

Several improvements were made in the Long Ranger analysis pipeline to better take advantage of the Linked-Read datatype. The first of these, the Lariat™ aligner, expands on the 'Read-Cloud' approach (Bishara et al. 2015). Lariat (https://github.com/10XGenomics/lariat) refines alignments produced by the BWA aligner by examining reads that map to multiple locations and determining if they share barcodes with reads that have high quality unique alignments (Li 2013). If a confident placement can be determined by taking advantage of the barcode information of the surrounding reads, the quality score of the correct alignment is adjusted. This approach allows for the recovery of roughly 38 Mb of sequence across the entire genome using multiple replicates of control samples (NA12878, NA19240, NA24385)(Figure 2). The amount of additional recovered sequence varies as a function of molecule length (Supplemental Figure 2).

When we look specifically at the ability of Lariat to improve read coverage over genes, we observe a net gain in gene coverage when performing lrWGS compared to srWGS, and even more robust improvement when performing lrWES compared to srWES (Supplemental Figure 3). When we limit the search space to a known set of 570 genes with closely related paralogs that confound short read alignment (NGS 'dead zone' genes (Mandelker et al. 2016)) we see a net gain in read coverage in 423 genes using lrWGS and 376 using lrWES. Further limiting the list to the 71 genes relevant to Mendelian disease, we see a net improvement in 51 of these genes using lrWGS and 41 genes using lrWES (Figure 3). Exome analysis was limited to multiple replicates of a single control sample, NA12878.

Small variant calling

Next, we assessed the performance of Linked-Reads for small variant calling (<50 bp). Small variant calling, particularly for single nucleotide variants (SNVs) outside of repetitive regions, is well powered by short reads because a high quality read alignment to the reference assembly is possible and the variant resides completely within the read. We used control samples, NA12878 and NA24385 as test cases. We produced two small variant call sets for each sample, one generated by running paired-end 10x Linked-Read Chromium libraries through the Long Ranger (10xLR) pipeline and one produced by analyzing paired-end reads from a PCR-free TruSeq library using GATK pipeline (PCR-) following best practices recommendations: https://software.broadinstitute.org/gatk/best-practices/. We made a total of 4,549,657 PASS variant calls from the NA12878 10xLR set, and 4,725,295 from the PCR- set, with 4,325,515 calls in common to both sets (Table 1). Numbers for both samples are in Table 1.

In order to assess the accuracy of the variant calling in each data set, we used the hap.py tool (Krusche) to compare the 10xLR and PCR- VCFs to the Genome in a Bottle (GIAB) high confidence call set (v. 3.2.2) (Zook et al. 2014). We chose this earlier version as it was the last GIAB data set that did not include 10x data as an input for their call set curation. This necessitated the use of GRCh37 as a reference assembly rather than the more current GRCh38 reference assembly. This limited us to analyzing only 82.67% of the SNV calls that overlap the high confidence regions. Initial results suggested that the 10xLR calls had comparable sensitivity (>99.6%) and specificity

(>99.8%) for SNVs (Table 1). We observed slightly diminished indel sensitivity (>89%) and specificity (>94.5%), driven largely by regions with extreme GC content and low complexity sequences (LCRs). Recent work suggests indel calling is still a challenging problem for many approaches, but that only 0.5% of LCRs overlap regions of the genome thought to be functional based on annotation or conservation (Li et al. 2017).

The GIAB high confidence data set is known to be quite conservative and we wished to explore whether there was evidence for variants called in the 10xLR set not covered by the GIAB. We utilized publicly available 40x coverage PacBio data sets available from the GIAB consortium (Zook et al. 2016) to evaluate Linked-Read putative false positive variant calls. Manual inspection of 25 random locations suggested that roughly half of the hap.py identified Linked-Read false positive calls were well supported by Linked-Read, short read, or PacBio evidence and were likely called false positive due to deficiencies in the GIAB truth set (Supplemental Table 1). We then did a global analysis of all 7,431 SNV and 16,713 indel putative false positive calls identified in NA12878 and looked for the alternate alleles in aligned PacBio reads only. This analysis provided evidence that 2,253 SNV and 12,826 indels of the GIAB determined false positive calls were likely valid calls (Supplemental Figure 4, Supplemental File 1). This prompted us to develop a new "extended truth set" which included an additional PacBio validated 78,361 SNV and 21,026 indels (see Methods for details on GIAB++ VCF). We also extended our analysis to include 69.72 Mb for NA12878 and 70.66 Mb for NA2438 of the genome in addition to the GIAB defined confident regions (see Methods for details on GIAB++ BED). We reanalyzed the variant calls with the hap.py tool against the extended truth set and augmented confident regions. Importantly, this allowed us to correctly identify an additional 71,467 SNV and 12,663 indels. We anticipate that this is a conservative estimate since our false positive calls are inflated due to little or no PacBio or short-read coverage in these regions. Of the total putative false positive calls exclusive to the GIAB++ analysis, 79.86% (31,475) of SNVs and 62.05% (2,790) of indels could not be validated because of little or no PacBio read coverage (Supplemental Figure 4). These data show the 10xLR approach provides for the identification of more small variants than can be identified by short read only approaches, driven by an increase in the percentage of the genome for which 10xLR can obtain high quality alignments.

Haplotype reconstruction and phasing

An advantage of Linked-Reads is the ability to reconstruct multi-megabase haplotypes (called phase blocks) for a single sample. Haplotype reconstruction increases sensitivity for calling heterozygous variants, particularly SVs (Huddleston et al. 2016). It also improves variant interpretation by providing information on the physical relationship of variants, such as whether variants within the same gene are in cis or trans. In the control samples analyzed, we see phase block N50 values for lrWGS of 10.3 Mb for NA12878, 9.58Mb for NA24385, 18.2 Mb for NA19240 and 302 kb for lrWES using Agilent SureSelect v6 baits on NA12878. This allowed for complete phasing of 91% and 90.8% of genes, respectively, in the genome and exome. Phase block length is a function of input molecule length, molecule size distribution and of sample heterozygosity extent and distribution. At equivalent mean molecule lengths, phase blocks will be longer in more diverse samples (Figure 4, Supplemental Figure 5). For samples with similar heterozygosity, longer input molecules will increase phase block lengths (Supplemental Figure 6).

Phase block construction using lrWES is additionally constrained by the bait set used to perform the capture and the reduced variation seen in coding sequences. In order to analyze factors impacting phase block construction, we assessed four samples with known compound heterozygous variants in three genes known to cause Mendelian disease, *DYSF*, *POMT2*, and *TTN*. The variants were separated by various distances, ranging from 33 Kb to over 188 Kb (Table 2). Initial DNA extractions yielded long molecules ranging in size from 75 Kb - 112 Kb. We analyzed these samples using the Agilent SureSelect V6 exome bait set, with downsampling of sequence data to both 7.25 Gb (~60x coverage) and 12 Gb of sequence (~100x coverage). In all cases, the variants were phased with respect to each other and determined to be in trans, as previously determined by orthogonal assays. In two of the three cases, the entire gene was phased. The *DYSF* gene was not completely phased in any sample because the difference between heterozygous SNPs at the 3' end of the gene was substantially longer than the mean molecule length. This gene is in the top 5% of genes intolerant to variation as determined by the RVIS metric, a measure of evolutionary constraint, suggesting that reduced exonic heterozygosity over the gene would be a common occurrence impairing complete phasing (Petrovski et al. 2013). However, in the context of sequencing for the identification of recessive disease, causative heterozygous variants would be expected to aid in the phasing of the disease-causing gene.

Many samples of interest have already been extracted using standard methods not optimized for high molecular weight DNA and may not be available for a fresh re-extraction to obtain DNA optimized for length. For this reason, we wanted to understand the impact of reduced molecule length on our ability to phase the genes and variants in these samples. We took the original freshly extracted long molecules and sheared them to various sizes, aiming to assess lengths ranging from 5Kb to the original full length samples (Table 2). These results illustrate the complex interplay between molecule length distribution and the observed heterozygosity within a region. For example, in sample B12-21, with variants in *TTN* that are 53 Kb apart, the variants could be phased, even with the smallest molecule size. However in sample B12-122, with variants in *POMT2* only 33 Kb apart, variant phasing is lost at 20Kb DNA lengths. We assessed the maximum distance between heterozygous sites observed in each gene. We then plotted the difference between this distance and the inferred molecule length against the molecule length and assessed the impact on causative SNP phasing (Figure 5). In general, when the maximum distance between heterozygous SNPs is greater than the molecule length (positive values), the ability to phase these SNPs decreases. There are exceptions to this as the longer molecules in the molecule size distribution will sometimes allow tiling between the variants, therefore extending phase block size beyond what would be expected based on the mean length alone.

Linked-Reads provide unparalleled power to reconstruct long haplotypes, or phase blocks. Optimizing for long input molecules provides for maximum phase block size, but even shorter molecule lengths can provide gene level phasing. This suggests that samples with higher levels of heterozygosity, such as from admixed individuals, could greatly benefit from the Linked-Read approach.

Structural variant detection

Short reads struggle with accurate and specific SV detection. This is, in part, due to the limitations of assessing long range information using short reads, which only provide information over short distances. Another complicating factor is the many types of structural variants, each requiring the detection of a different signal depending on the type and mechanism of the event (Alkan et al. 2011; Collins et al. 2017). There is increasing evidence that grouping reads by their source haplotype improves SV sensitivity, but this is not commonly done in practice (Huddleston et al. 2016; Chaisson et al. 2017).

Linked-Reads provide improved power to detect large-scale SVs, particularly balanced events, when compared to short read approaches. We use two novel algorithms to identify large SVs, one that assesses deviations from expected barcode coverage and one that looks for unexpected barcode overlap between distant regions. The barcode coverage algorithm is useful for assessing CNVs, while the barcode overlap method can detect a variety of SVs and it is particularly well powered to identify large (>30Kb), balanced events. SV calls are a standard output of the Long Ranger pipeline and are described using standard file formats. We compared SV calls from the NA12878 sample to validated calls described in a recent publication of a structural variant classifier, svclassify (Parikh et al. 2016).

Comparing SV call sets produced by different methods is a challenging task. There is often ambiguity around the exact coordinate of the breakpoint(s), in part because repetitive sequence content frequently flanks structural variation (Wittler et al. 2015). An additional challenge is variability in the inclusion of the many different possible variant types. The validated call set published with svclassify (Parikh et al. 2016) contains deletions and insertions, but no balanced events. By contrast, the Long Ranger pipeline output contains deletions, duplications and balanced events, but Long Ranger does not currently call insertions (Supplemental Table 2). Long Ranger identifies event types by matching to simple models of deletions, duplications and inversions. Therefore, there are additional events where Long Ranger identifies clear evidence for anomalous barcode overlap, but is unable to match the event to one of the pre-defined models. These undefined events are rendered as unknown.

For these reasons, we limited our comparisons to the set of deletion calls only. We partition the ground truth set into SVs <30 Kb and SVs >30 Kb as different algorithms are used to call these events. We first consider variants >30 Kb. There are 11 of these in the svclassify set and 23 in the 'Call set', with 8-9 being common to both (Table 3). Long Ranger calls two highly overlapping events that map to the same svclassify event- thus 9 Long Ranger calls map to 8 svclassify events. Of the three svclassify calls not called by Long Ranger, one (chr12:8,558,486-8,590,846) is well-supported in the Linked-Read data by barcode overlap. For this event, Long Ranger calls a 10kb small deletion with a consistent 5' breakpoint to the svclassify event, but prematurely closes the event, missing 22kb of the deletion. A second event (chr22:24,274,143-24,311,297) is also well-supported but is filtered out from the 'Call set' as it overlaps with a segmental duplication. There is no support for the last missing call (chr14:37,631,608-37,771,227). Further investigation of this call reveals that it is genotyped as homozygous reference in NA12878 in the svclassify truth set. We then performed manual review of the 14 events called by Long Ranger that are not in the svclassify set. These 14 potential FP calls can be collapsed into 10 unique deletion calls. By manual review all 10 calls have significant barcode overlap and coverage support including three events that are known copy number variant loci per the Genome Reference Consortium (Supplemental Figure 7). Thus, the Long Ranger large SV deletion calls show both high sensitivity and specificity.

We next considered deletions <30Kb. There are 6,839 such PASS variants in the Long Ranger set and 2,665 of these in the svclassify set, with 2,428 of these in common. Manual inspection of 20 calls unique to each call set suggests that Long Ranger has high sensitivity but low specificity, with algorithmic performance particularly diminished in regions where there is no phasing (Supplemental File 2). While sensitivity of the Long Ranger approach is good, this comes at the expense of specificity (Supplemental Tables 3,4). There is clear evidence that algorithmic improvements will produce further gains in sensitivity and specificity for this class of variants.

A particular strength of Linked-Reads is the ability to call balanced events based on anomalous barcode overlap. In the NA12878 whole genome data five inversions are called, all of which are supported by orthogonal data (Supplemental Figure 8). Three calls are present in an orthogonal call set (Kidd et al. 2008; Zook et al. 2016); (http://invfestdb.uab.cat/download.php), while the remaining two calls are known reference assembly issues resulting in apparent inversion calls (HG-28, HG-1433).

To assess for inversion calling accuracy, we assessed consistency of the inversion calls with >30kb inversion calls from InvFEST (http://invfestdb.uab.cat/download.php). There are three inversions reported in NA12878 in this data source, all homozygous. One, chr7:54,258,468-54,354,315 (Supplemental Figure 8A), is one of the five inversion calls made in the Chromium dataset. There is evidence for anomalous barcode overlap in the region around the InvFest event chr8:6,909,898-12,617,968 (Supplemental Figure 8F), but only low-quality events are called with coordinates consistent with the known breakpoints. Because tandem gene families mark both ends of the inversion, it is likely that there is significant read misalignment at both breakpoints resulting in multiple low-quality calls with slightly staggered start and end coordinates when they should actually all be aligned to a single set of breakpoints.

The final InvFest event, chr15:28,157,404-30,687,000 (Supplemental Figure 8G), shows minimal anomalous barcode overlap in the Chromium data in a pattern not consistent with an inversion. There are no candidate calls connecting the two loci. This region partially overlaps the known gamma inversion haplotype of the *GOLGA8* locus (https://www.ncbi.nlm.nih.gov/grc/human/issues?q=chr15:28157404-30687000&asm=GRCh37.p13). Because the InvFest calls were made using GRCh36 as a reference, and this region was adjusted in GRCh37, it is possible that this call would disappear from InvFest if analyses were redone using GRCh37.

Linked-Reads provide a clear advantage for SV detection over standard short read approaches. The data type is still relatively new and the algorithmic approaches to SV identification are still relatively immature. Several groups have already described methods utilizing Linked-Reads for SV detection, largely in tumor samples (Spies et al. 2016; Elyanow et al. 2017; Xia et al. 2017). The power of Linked-Reads to identify balanced events is a notable improvement and there is evidence for this class of event being more prevalent in the population than originally realized (Collins et al. 2017; Chaisson et al. 2017). The evaluation of Linked-Reads and Long Ranger to identify complex, constitutional events provides additional evidence of the power of this datatype for complex event detection (Garcia et al. 2017).

Analysis of samples from individuals with inherited disease

We went on to investigate the utility of Linked-Read analysis on real samples with known variants. In particular, we were interested in events that are typically difficult with a standard, short read exome. We were able to obtain samples from a cohort that had been assessed using a high depth NGS-based inherited predisposition to cancer screening panel. This cohort contained samples with known exon level deletion and duplication events. We analyzed these 12 samples from 9 individuals using an Agilent SureSelect V6 Linked-Read exome at both 7.25 Gb (equivalent to ~60x raw coverage) and 12 Gb (~100x) coverage (Table 3). For three samples patient-derived cell lines were available in addition to archival DNA, allowing us to investigate the impact of DNA length on exon-level deletion/duplication calling.

We were able to identify 5 of the 9 expected exon-level events in these samples in at least one sample type/depth combination. In 2 samples, increasing depth to 12Gb enabled calling that was not possible at 7.25Gb (Samples D and F (archival), Table 4). For the three samples with matched cell lines and archival DNA, two had variants that could not be called in either sample type at either depth, while sample F could be called at both depths for the longer DNA extracted from the cell line, but could only be called at the higher depth in the shorter archival sample. There is a striking correlation between the ability to phase the gene and to call the variant, with no variants successfully called in samples that could not be phased over the region of interest.

For two of the samples where Linked-Read exome sequencing was unable to phase or call the known variant, we performed lrWGS. In one case, the presence of intronic heterozygous variation was able to restore phasing to the gene and the known event was called. In the second case, there was still insufficient heterozygous variation in the sample to allow phasing and the event was not called. This again demonstrates that phasing is dependent both on molecule length as well as sample heterozygosity. Some samples in this group had decreased diversity in the regions of interest compared to other samples, and we were less likely to be able to call variants in these samples. (Supplemental Figure 9). Generally, it should be possible to increase the probability of phasing a gene in an exome assay by augmenting the bait set to provide coverage for common intronic variant SNPs, thus preserving the cost savings of exome analysis, but increasing the power of the Linked-Reads to phase. However, samples with generally reduced heterozygosity will remain difficult to phase and completely characterize.

One sample in this set contained both a single exon event and a large variant in the *PMS2* gene. Despite phasing the *PMS2* gene we were unable to call this variant in either genome or exome sequencing. Manual inspection of the data reveals increased phased barcode coverage in the *PMS2* region, supporting the presence of a large duplication that was missed by the SV calling algorithms (Supplemental Figure 10).

Linked-Reads provide a better first line approach to assess individuals for variants in these genes. While we were not able to identify 100% of the events, we were able to identify 5 of 9 of these events using a standard exome approach, rather than a specialized assay. Improved baiting approaches, or WGS, should improve that ability to identify these variants given the clear relationship between phasing and sensitivity. Lastly, there is room for algorithmic improvement as at least one variant had clear signal in the Linked-Read data, but failed to be recognized by current algorithms.

Discussion

Short read sequencing has become the workhorse of human genomics. This cost effective, high throughput, and accurate base calling approach provides robust analysis of short variants in unique regions of the genome, but struggles to reliably call SVs and fails to reconstruct long range haplotypes (Sudmant et al. 2015). It is becoming increasingly clear, that to perform a comprehensive genome analysis, haplotype information, variant calling in repeat regions, and SV identification must be included in the analysis (Chaisson et al. 2017). Indeed, analyzing human genomes in their diploid context will be a critical step forward in genome analsis (Aleman 2017). We have described an improved implementation of Linked-Reads, a method that substantially improves the utility of short read sequencing. The increased number of partitions and improved biochemistry make this a stand alone approach for genome analysis, requiring only a single Linked-Read library, from only ~1 ng of DNA. This approach, coupled with novel algorithms, powers short reads to reconstruct multi-megabase phase blocks, identify large balanced and unbalanced structural variants and identify small variants, even in regions of the genome typically recalcitrant to short read approaches.

Some limitations to this approach currently exist. We observe a loss of coverage in regions of the genome that show extreme GC content. We additionally see reduced performance in small indel calling, though this largely occurs in homopolymers regions and LCRs. Recent work suggests ambiguity in such regions may be tolerated for a large number of applications (Li et al. 2017). It is also clear that algorithmic improvements to Long Ranger would improve variant calling, particularly as some classes of variants, such as insertions, are not yet attempted. However, this is not uncommon for new data types and there has already been some progress in this area (Spies et al. 2016; Elyanow et al. 2017; Xia et al. 2017)].

Despite these limitations, Linked-Read sequencing provides a clear advantage over short reads alone. This pipeline allows for the construction of long range haplotypes as well as the identification of short variants and SVs from a single library and analysis pipeline. No other approach that scales to thousands of genomes provides this level of detail for genome analysis. Other recent studies have demonstrated the power of Linked-Reads to resolve complex variants in both germline and cancer samples (Collins et al. 2017; Greer et al. 2017; Garcia et al. 2017). Recent work demonstrates that Linked-Reads outperforms the switch accuracy and phasing completeness of other haplotyping methods, and provides multi-MB phase blocks (Chaisson et al. 2017). In another report, Linked-Reads also enable the ability to perform diploid, *de novo* assembly in combination with an assembly program, Supernova (Weisenfeld et al. 2017). The ability to provide reference free analysis promises to increase our understanding of diverse populations. Finally, the ability to represent and analyze genomes in terms of haplotypes, rather than compressed haploid representations, represents a crucial shift in our approach to genomics, allowing for a more complete and accurate reconstruction of individual genomes.

Methods

Samples and DNA Isolation

Control samples (NA12878, NA19240, NA24385, NA19240, and NA24385) were obtained as fresh cultured cells from the Coriell Cell biorepository (https://catalog.coriell.org/1/NIGMS). DNA was isolated using the Qiagen MagAttract HMW DNA kit and quantified on a Qubit fluorometer following recommended protocols: https://support.10xgenomics.com/genome-exome/index/doc/user-guide-chromium-genome-reagent-kit-v2-chemistry.

Clinical samples from individuals with known heterozygous variants in three Mendelian disease loci (*DYSF*, *POMT2* and *TNN*) were collected at the Massachusetts General Hospital, Analytic and Translational Genetics Unit and shipped to 10x genomics as cell lines. Genomic DNA was extracted from each cell line as described above. Use of samples from the Broad Institute was approved by the Partners IRB (protocol 2013P001477).

Clinical samples from individuals with inherited cancer were collected at The Institute of Cancer

Research, London and shipped to 10x genomics as cell lines or archival DNA. This sample cohort was previously accessed for predisposition to cancer. Samples were recruited through the Breast and Ovarian Cancer Susceptibility (BOCS) study and the Royal Marsden Hospital Cancer Series (RMHCS) study, which aimed to discover and characterize disease predisposition genes. All patients gave informed consent for use of their DNA in genetic research. The studies have been approved by the London Multicentre Research Ethics Committee (MREC/01/2/18) and Royal Marsden Research Ethics Committee (CCR1552), respectively. Samples were also obtained through clinical testing by the TGLclinical laboratory, an ISO 15189 accredited genetic testing laboratory. The consent given from patients tested through TGLclinical includes the option of consenting to the use of samples/data in research; all patients whose data was included in this study approved this option. DNA was extracted from cell lines as described above and archival DNA samples were checked for size and quality according to manufacturer's recommendations: https://support.10xgenomics.com/genome-exome/sample-prep/doc/demonstrated-protocol-hmw-dna-qc .

Chromium^{TM} Linked-Read Library Preparation

1.25 ng of high molecular weight DNA was loaded onto a Chromium controller chip, along with 10x Chromium reagents (either v1.0 or v2.0) and gel beads following recommended protocols: https://assets.contentful.com/an68im79xiti/4z5JA3C67KOyCE2ucacCM6/d05ce5fa3dc4282f3da5ae7296f2645b/CG00022_GenomeReagentKitUserGuide_RevC.pdf. The initial part of the library construction takes place within droplets containing beads with unique barcodes (called GEMs). The library construction incorporates a unique 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 chances 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.

Target enrichment for the Linked-Read whole exome libraries was performed using Agilent Sure Select V6 exome baits following recommended protocols:

https://assets.contentful.com/an68im79xiti/Zm2u8VlFa8qGYW4SGKG6e/4bddcc3cd60201388f7b82d241547086/CG000059_DemonstratedProtocolExome_RevC.pdf. Supplemental Figure 11 describes targeted sequencing with Linked-Reads.

GemCode™ Linked-Read Library Preparation

For the GemCode comparator analyses, Linked-Read libraries were prepared for truth samples NA12878, NA12877, and NA12882 using a GemCode controller and GemCode V1 reagents following published protocols (Zheng et al. 2016).

TruSeq PCR-free Library Preparation

350-800 ng of genomic DNA was sheared to a size of ~385 bp using a Covaris®M220 Focused Ultrasonicator using the following shearing parameters: Duty factor = 20%, cycles per burst = 200, time = 90 seconds, Peak power 50. Fragmented DNA was then cleaned up with 0.8x SPRI beads and left bound to the beads. Then, using the KAPA Library Preparation Kit reagents (KAPA Biosystems, Catalog # KK8223), DNA fragments bound to the SPRI beads were subjected to end repair, A-base tailing and Illumina®'PCR-free' TruSeq adapter ligation (1.5μM final concentration of adapter was used). Following adapter ligation, two consecutive SPRI cleanup steps (1.0X and 0.7X) were performed to remove adapter dimers and library fragments below ~150 bp in size. No library PCR amplification enrichment was performed. Libraries were then eluted off the SPRI beads in 25 ul elution buffer and quantified with quantitative PCR using KAPA Library Quant kit (KAPA Biosystems, Catalog # KK4824) and an Agilent Bioanalyzer High Sensitivity Chip (Agilent Technologies) following the manufacturer's recommendations.

Target enrichment for the Linked-Read whole exome libraries was performed using Agilent Sure Select V6 exome baits following recommended protocols.

Sequencing

Libraries were sequenced on a combination of Illumina®instruments (HiSeq®2500, HiSeq 4000, and HiSeq X). Paired-End sequencing read lengths were as follows: TruSeq and Chromium whole genome libraries (2X150bp); Chromium whole exome libraries (2X100bp or 114bp, 98bp), and Gemcode libraries (2X98bp). lrWGS libraries are typically sequenced to 128 Gb, compared to 100 Gb for standard TruSeq PCR free libraries. The additional sequence volume compensates for sequencing the barcodes as well a small number of additional sources of wasted data and gives an average, de-duplicated coverage of approximately 30x. To demonstrate the extra sequence volume is not the driver of the improved alignment coverage, we performed a gene finishing comparison at matched volume (100Gb lrWGS and 100Gb TruSeq PCR-) and continue to see coverage gains (Supplemental Figure 12).

Analysis

Comparison of 10X and GATK Best Practices

We ran the GATK Best practices pipeline to generate variant calls for Truseq PCR-free data using the latest GATK3.8 available at the time. We first subsample the reads to obtain 30x whole genome coverage. The read set is then aligned to GRCh37, specifically the hg19-2.2.0 reference using BWA-MEM (version 0.7.12). The reads are then sorted, the duplicates are marked, and the bam is indexed using picard tools (version 2.9.2). We then perform indel realignment and recalibrate the bam (base quality score recalibration) using known indels from Mills Gold Standard and 1000G project and variants from dbsnps (version 138). Finally we call both indel and SNVs from the bam using HaplotypeCaller and genotype it to produce a single vcf file. This vcf file is then compared using hap.py (https://github.com/Illumina/hap.py, commit 6c907ce) to the truth variant set curated by Genome in a Bottle on confident regions of the genome. We calculate sensitivity and specificity for both SNVs and indels to contrast the fidelity of the Long Ranger short variant caller and the GATK-Best Practices pipeline. All Long Ranger runs were performed with a pre-release build of Long Ranger version 2.2 utilizing GATK as a base variant caller. Long Ranger 2.2 adds a large-scale CNV caller that employs barcode coverage information and incremental algorithmic improvements. 10x Genomics plans to release an open-source Long Ranger 2.2 in February 2018.

Development of extended truth set

Any putative false positive variant found in the TruSeq/GATK or Chromium/Long Ranger VCFs, was tested for support in the PacBio data. Raw PacBio FASTQs were aligned to the reference using BWA-MEM -x pacbio (Li 2013). To test a variant, we fetch all PacBio reads covering the variant position, and retain the substring aligned within 50bp of the variant on the reference. We re-align the PacBio read sequence to the +/-50bp interval of the reference, and the same interval with the alternate allele applied. A read is considered to support the alternate allele if the alignment score to the alt-edited template exceeds the alignment score of the reference template.

False-positive calls that passed the PacBio validation and did not overlap an existing GIAB variant were added to the GIAB VCF to form the GIAB++ VCF. We selected regions of 2-6 fold degeneracy as determined by the 'CRG Alignability' track (Derrien et al. 2012) as regions where improved alignment is likely to yield credible novel variants. We union the GIAB confident regions BED file with these regions to determine the GIAB++ confident regions BED.

Structural variant comparison against deletion ground truth

We downloaded our ground truth set of deletion events from the svclassify supplementary materials site (Parikh et al. 2016). After deciding the multiple segmentations we would do to our Long Ranger calls and filtering for deletions, we overlapped them to the ground truth using the bedr package and bedtools v2.26.0 (Quinlan and Hall 2010). We retained for further analysis those showing at least 50% reciprocal overlap.

Acknowledgements

We thank the individuals who donated specimens for research. This manuscript would not have been possible without their contributions. We thank Stephane Boutet and Sarah Taylor for reviewing the manuscript. We also wish to thank Kariena Dill for help with manuscript preparation, Kevin Wu for assistance with the technical aspects on setting up markdown and

Docker. We also thank Jamie Schwendinger-Schreck for project management as well as invaluable contributions in manuscript preparation.

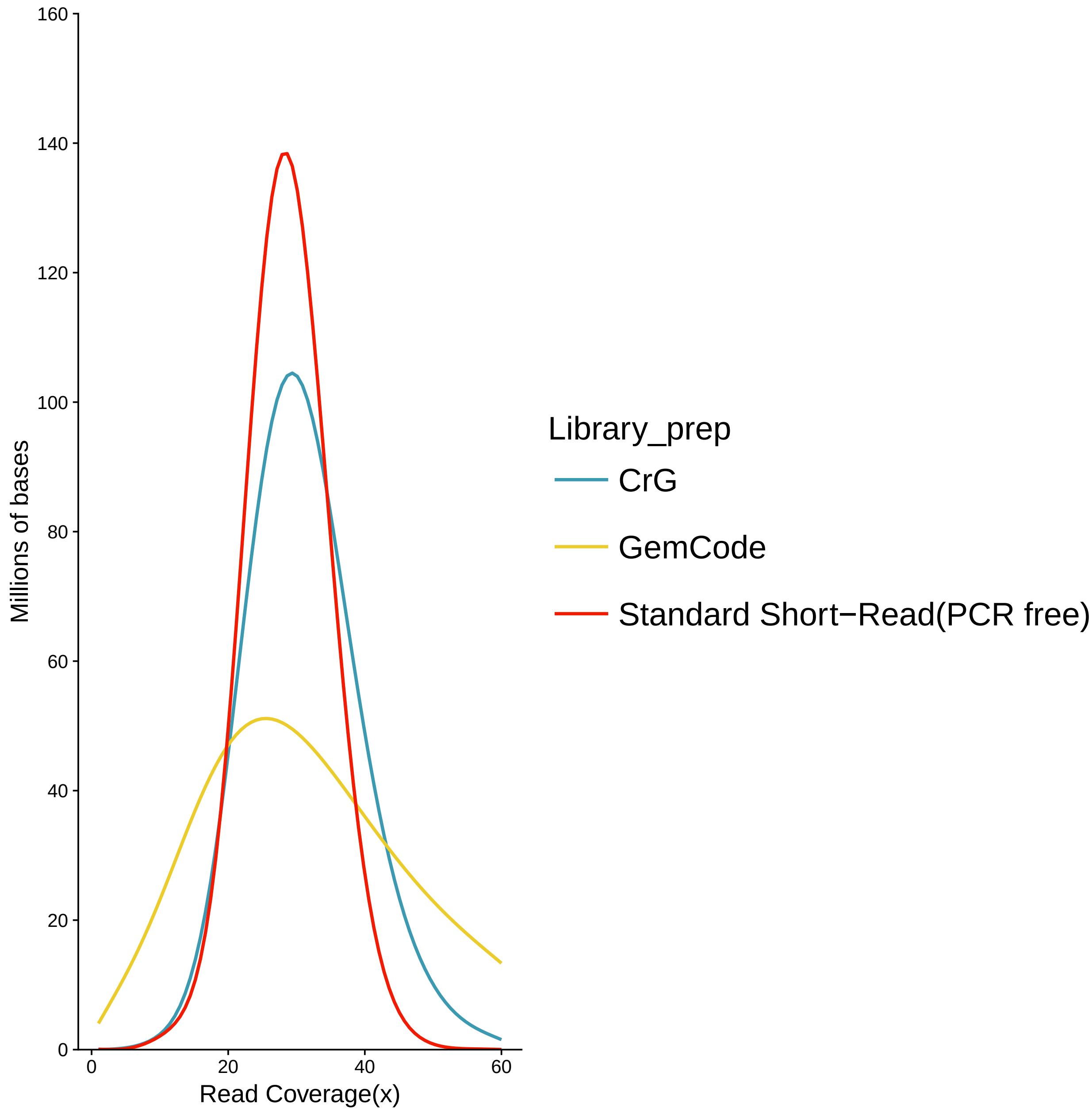

Figure 1: Coverage Evenness.

Distribution of read coverage for the entire human genome (GRCh37). Comparisons between 10x Genomics Chromium Genome (CrG), 10x Genomics GemCode (GemCode), and Illumina TruSeq PCR free standard short-read NGS library preparations (Standard Short Read (PCR Free)). Sequencing was performed in an effort to match coverage (see methods). Note the shift of the CrG curve to the left, showing the improved coverage of Chromium vs. GemCode. X-axis represents the fold read coverage across the genome. Y-axis represents the total number of bases covered at any given read depth.

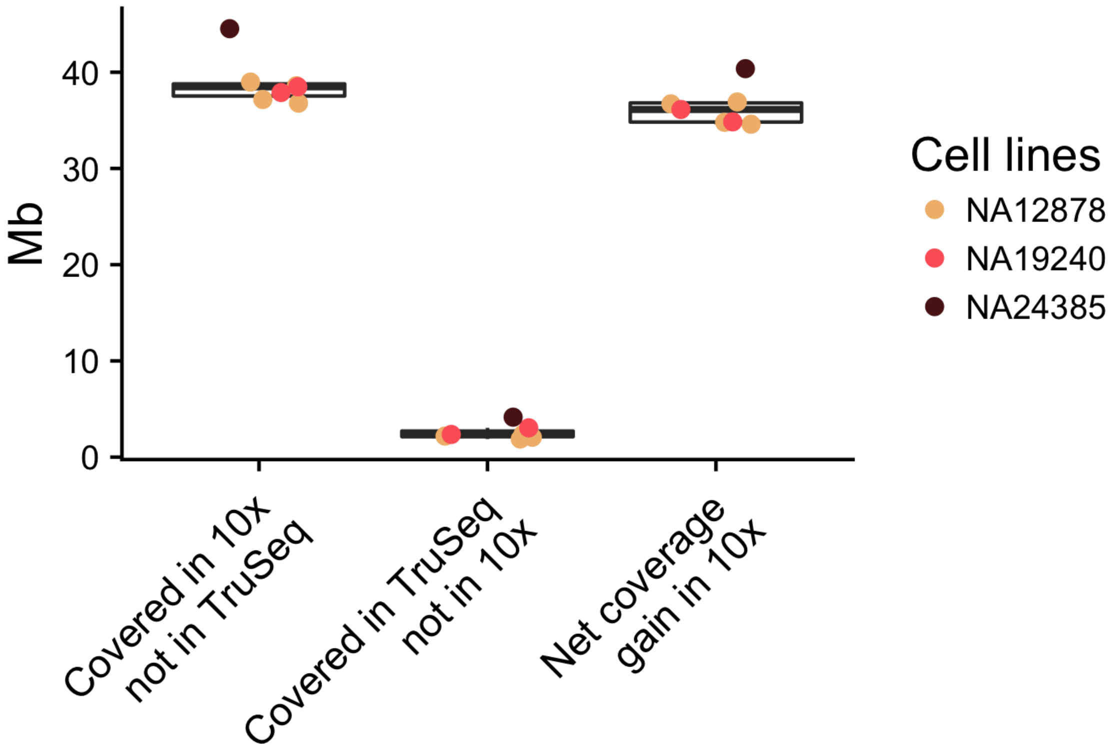

Figure 2: Comparison of unique genome coverage by assay.

The x-axis shows the number of sites with a coverage of >=5 reads at MAPQ >=30. Column one shows amount of the genome covered by 10x Chromium where PCR free TruSeq does not meet that metric. Column 2 shows the amount of the genome covered by PCR free TruSeq where 10x Chromium does not meet the metric. Column 3 shows the net gain of genome sequence with high quality alignments when using 10x Chromium versus PCR free TruSeq. The comparison was performed on samples with matched sequence coverage (see methods).

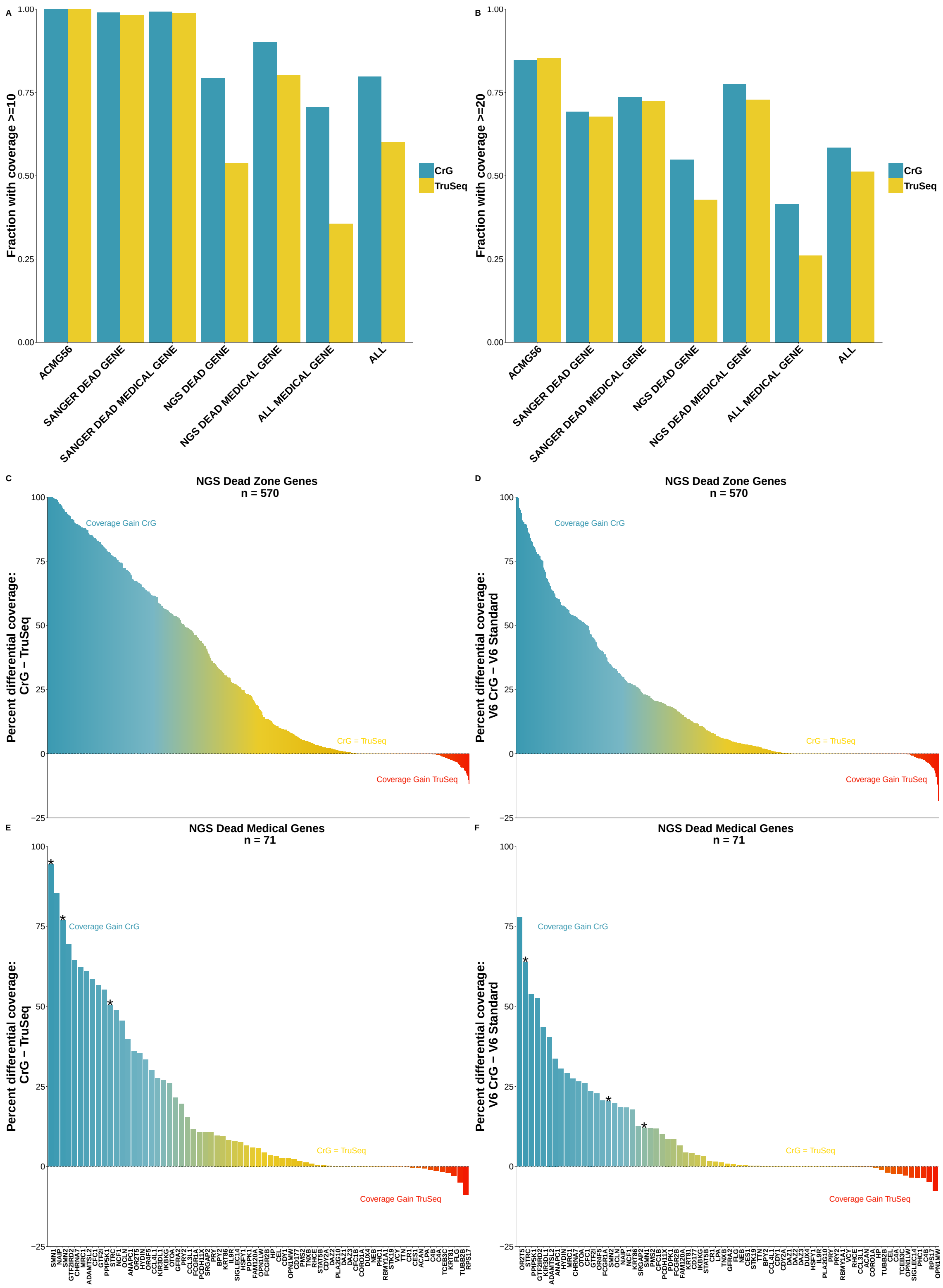

Figure 3: Gene finishing metrics.

Gene finishing metrics for whole genome and whole exome sequencing across selected gene sets. Genome is shown on left, exome on right. Gene finishing is metric for expressing gene coverage and completeness. Finishing is defined as the percentage of exonic bases with 10x coverage for genome (Panel A) and 20x for exome (Panel B)(Mapping quality score >=MapQ30). CrG is Chromium Linked-Reads and TruSeq is PCR free TruSeq. Top row: Gene finishing statistics for 12 disease relevant gene panels. Shown is the average value across all genes in each panel. While Chromium provides a coverage edge in all panel sets, the impact is particularly profound for 'NGS Dead Zone' genes, as well as genes implicated in Mendelian disorders. Panels C-F show the net coverage differences for individual genes when comparing Chromium to PCR free TruSeq. Each bar shows the difference between the coverage in PCR free Truseq from the coverage in 10x Chromium. Panel C and D show the 570 NGS 'dead zone' genes for genome (panel C) and exome (panel D). Panels E and F limit the graphs to the list of NGS dead zone genes implicated in Mendelian disease. In panels C-F, the blue coloring highlights genes that are inaccessible to short read approaches, but because accessible using CrG, the yellow coloring indicates genes where CrG provides an improvement. The red coloring shows genes with a slight coverage increase in TruSeq, though these genes are typically still accessible to CrG. Highlighted with an asterisk are the genes *SMN1*, *SMN2* and *STRC*. The comparison was performed on samples with matched coverage (see methods).

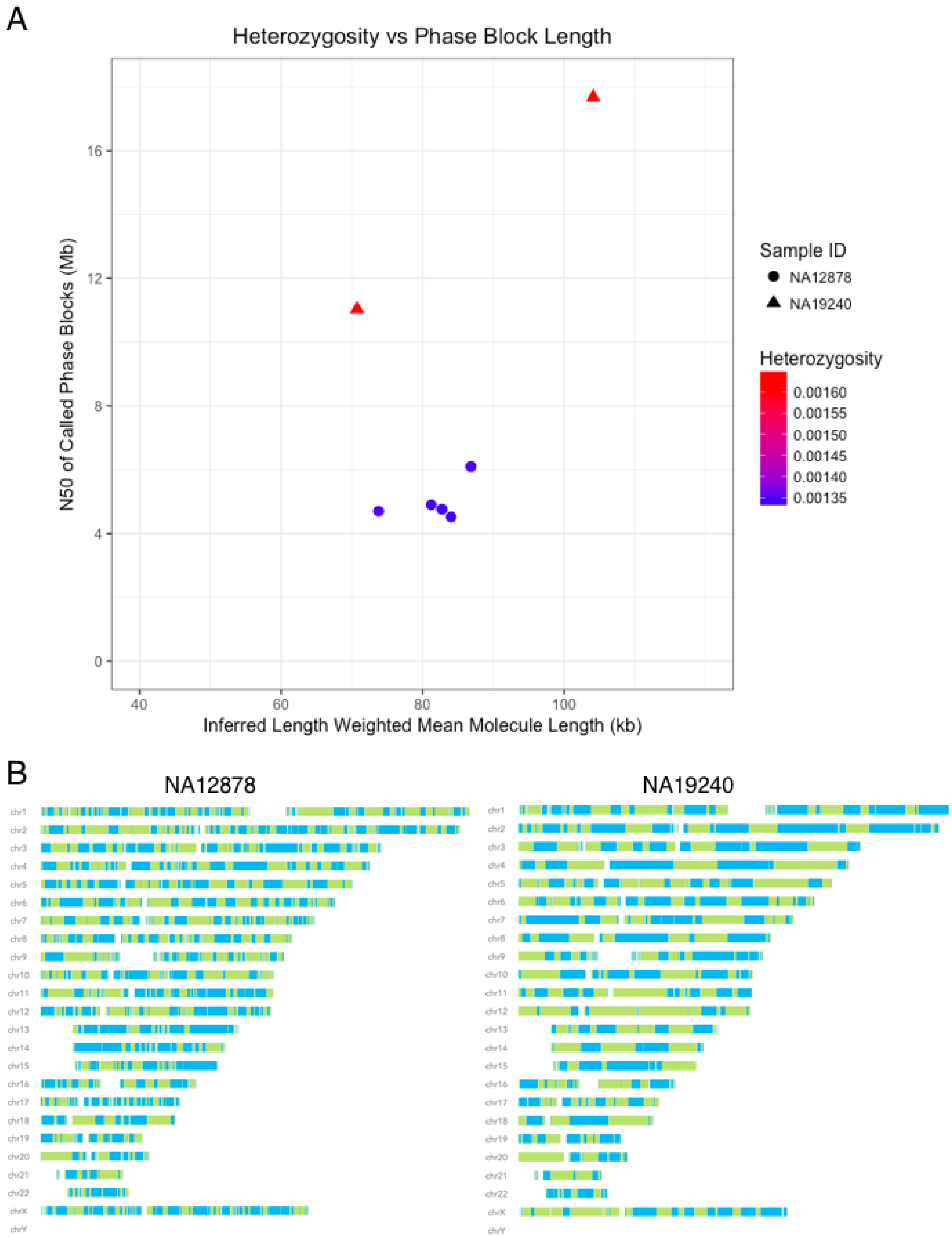

Figure 4: Haplotype reconstruction and phasing.

A. Inferred Length weighted mean molecule length plotted against N50 of called Phase blocks (both metrics reported by Long Ranger) and differentiated by sample ID and heterozygosity. Heterozygosity was calculated by dividing the total number of heterzygous positions called by Long Ranger by the total number of non-N bases in the reference genome (GRCh37). Two replicates of NA19240 and 5 replicates of NA12878 were used. Samples with higher heterozygosity produce longer phase blocks than samples with less diversity when controlling for input molecule length. B. Phase block distributions across the genome for input length matched Chromium Genome samples NA12878 and NA19240. Phase blocks are shown as displayed in Loupe Genome Browser™. Solid colors indicate phase blocks.

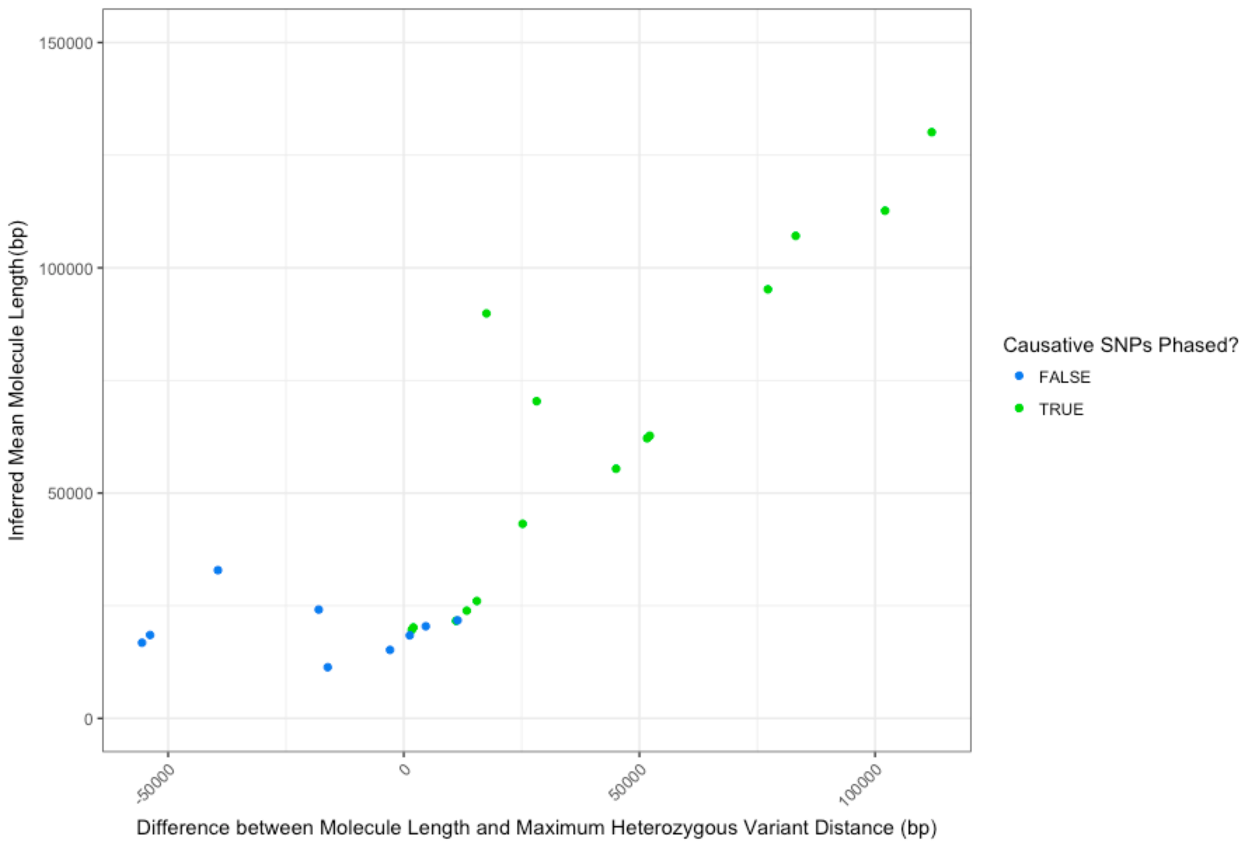

Figure 5: Validated examples of impact of molecule length on phasing (7.25Gb).

Blue dots represent samples for which the variants of interest are not phased, and green dots represent samples for which there is phasing of the variants of interest. At longer molecule lengths (>50kb), the molecule length was always longer than the maximum distance between heterozygous SNPs in a gene, and phasing between the causative SNPs was always observed. As molecule length shortens, it becomes more likely that the maximum distance between SNPs exceeds the molecule length (reflected as a negative difference value) and phasing between the causative SNPs was never observed in these cases. When maximum distance is similar to the molecule length causative SNPs may or may not be phased. This is likely impacted by the molecule length distribution within the sample.

Tables

Table 1: Summary of variant call numbers with respect to GIAB

	NA12878 10X LR	NA12878 PCR-	NA24385 10X LR	NA24385 PCR-
Total called variants	4,549,657	4,725,295	4,452,529	4,625,565
Total SNVs	3,797,297	3,815,792	3,720,041	3,738,969
Sensitivity (SNVs)	0.9963481	0.9980531	0.9966186	0.9980583
Specificity (SNVs)	0.9976409	0.9981325	0.9984370	0.9985823
SNVs in confident regions	3,150,405	3,154,434	3,042,842	3,047,175
SNVs in Truth variants set	3,142,755	3,148,133	3,037,531	3,041,919
Sensitivity++ (SNVs)	0.9943040	0.9914930	0.9965406	0.9919112
Specificity++ (SNVs)	0.9857564	0.9863063	0.9860670	0.9853615
SNVs in confident regions++	3,260,907	3,250,098	3,146,338	3,134,343
SNVs in Truth variants set++	3,214,222	3,205,135	3,101,948	3,087,538
Total indels	752,360	909,503	732,488	886,596
Sensitivity (indels)	0.9264632	0.9748848	0.8908629	0.9785483
Specificity (indels)	0.9531529	0.9822795	0.9483815	0.9857084
Indels in confident regions	356,756	368,782	445,615	477,832
Indels in Truth variants set	331,886	349,232	414,041	454,794
Sensitivity++ (indels)	0.9084890	0.9589195	0.8848831	0.9669737
Specificity++ (indels)	0.9446263	0.9726808	0.9450604	0.9762717
Indels in confident regions++	373,535	387,603	460,378	494,051
Indels in Truth variants set++	344,549	363,675	426,442	466,003

Table 1: The table shows the counts of variants (SNV and indel) from variant calls generated in four experiments: NA12878 10x data run through LongRanger (NA12878 10xLR), NA12878 Truseq PCR-free data run through GATK-Best Practices pipeline (NA12878 PCR-), NA24385 10x data run through LongRanger (NA24385 10xLR), NA24385 Truseq PCR-free data run through GATK-Best Practices pipeline (NA24385 PCR-). These variants were compared to the GIAB VCF truth set and

GIAB BED confident regions using hap.py and data is shown per variant type for count of variants in the truth set and in the confident regions (along with sensitivity and specificity). Data is also shown for the same quantities when the variant calls were compared to the extended truth set (GIAB++ VCF) and the augmented confident region (GIAB++ BED).

Table 2: Gene, variant distance and RVIS score for clinically-relevant genes

Sample	Gene	Var1	Var2	Var distance	RVIS score	RVIS %	Molecule length	Var phased?
B12-38	DYSF	chr2:71,778,243dupT	chr2:71,817,342_71,817,343delinsAA	39,097 bp	-1.31	4.65%	18,461 bp	Yes
B12-38	DYSF	chr2:71,778,243dupT	chr2:71,817,342_71,817,343delinsAA	39,097 bp	-1.31	4.65%	16,911 bp	No
B12-38	DYSF	chr2:71,778,243dupT	chr2:71,817,342_71,817,343delinsAA	39,097 bp	-1.31	4.65%	13,553 bp	No
B12-38	DYSF	chr2:71,778,243dupT	chr2:71,817,342_71,817,343delinsAA	39,097 bp	-1.31	4.65%	21,226 bp	Yes
B12-38	DYSF	chr2:71,778,243dupT	chr2:71,817,342_71,817,343delinsAA	39,097 bp	-1.31	4.65%	19,309 bp	Yes
B12-38	DYSF	chr2:71,778,243dupT	chr2:71,817,342_71,817,343delinsAA	39,097 bp	-1.31	4.65%	18,439 bp	No
B12-38	DYSF	chr2:71,778,243dupT	chr2:71,817,342_71,817,343delinsAA	39,097 bp	-1.31	4.65%	42,939 bp	Yes
B12-38	DYSF	chr2:71,778,243dupT	chr2:71,817,342_71,817,343delinsAA	39,097 bp	-1.31	4.65%	34,800 bp	Yes
B12-38	DYSF	chr2:71,778,243dupT	chr2:71,817,342_71,817,343delinsAA	39,097 bp	-1.31	4.65%	130,101 bp	Yes
B12-38	DYSF	chr2:71,778,243dupT	chr2:71,817,342_71,817,343delinsAA	39,097 bp	-1.31	4.65%	119,747 bp	Yes
B12-38	DYSF	chr2:71,778,243dupT	chr2:71,817,342_71,817,343delinsAA	39,097 bp	-1.31	4.65%	88,410 bp	Yes
B12-38	DYSF	chr2:71,778,243dupT	chr2:71,817,342_71,817,343delinsAA	39,097 bp	-1.31	4.65%	85,077 bp	Yes
B12-112	POMT2	chr14:77,745,107A>G	chr14:77,778,305C>T	33,198 bp	-0.93	9.68%	21,106 bp	No
B12-112	POMT2	chr14:77,745,107A>G	chr14:77,778,305C>T	33,198 bp	-0.93	9.68%	15,536 bp	No
B12-112	POMT2	chr14:77,745,107A>G	chr14:77,778,305C>T	33,198 bp	-0.93	9.68%	16,546 bp	No
B12-112	POMT2	chr14:77,745,107A>G	chr14:77,778,305C>T	33,198 bp	-0.93	9.68%	12,277 bp	No

Table 2: Gene, variant distance and RVIS score for clinically-relevant genes *(continued)*

Sample	Gene	Var1	Var2	Var distance	RVIS score	RVIS %	Molecule length	Var phased?
B12-112	POMT2	chr14:77,745,107A>G	chr14:77,778,305C>T	33,198 bp	-0.93	9.68%	10,609 bp	No
B12-112	POMT2	chr14:77,745,107A>G	chr14:77,778,305C>T	33,198 bp	-0.93	9.68%	20,782 bp	No
B12-112	POMT2	chr14:77,745,107A>G	chr14:77,778,305C>T	33,198 bp	-0.93	9.68%	21,858 bp	No
B12-112	POMT2	chr14:77,745,107A>G	chr14:77,778,305C>T	33,198 bp	-0.93	9.68%	55,546 bp	Yes
B12-112	POMT2	chr14:77,745,107A>G	chr14:77,778,305C>T	33,198 bp	-0.93	9.68%	54,569 bp	Yes
B12-112	POMT2	chr14:77,745,107A>G	chr14:77,778,305C>T	33,198 bp	-0.93	9.68%	112,692 bp	Yes
B12-112	POMT2	chr14:77,745,107A>G	chr14:77,778,305C>T	33,198 bp	-0.93	9.68%	107,082 bp	Yes
B12-21	TTN	chr2:179,585,773C>A	chr2:179,531,966C>A	53,807 bp	2.17	98.04%	20,756 bp	Yes
B12-21	TTN	chr2:179,585,773C>A	chr2:179,531,966C>A	53,807 bp	2.17	98.04%	17,432 bp	Yes
B12-21	TTN	chr2:179,585,773C>A	chr2:179,531,966C>A	53,807 bp	2.17	98.04%	18,128 bp	Yes
B12-21	TTN	chr2:179,585,773C>A	chr2:179,531,966C>A	53,807 bp	2.17	98.04%	18,158 bp	Yes
B12-21	TTN	chr2:179,585,773C>A	chr2:179,531,966C>A	53,807 bp	2.17	98.04%	29,796 bp	Yes
B12-21	TTN	chr2:179,585,773C>A	chr2:179,531,966C>A	53,807 bp	2.17	98.04%	28,799 bp	Yes
B12-21	TTN	chr2:179,585,773C>A	chr2:179,531,966C>A	53,807 bp	2.17	98.04%	63,218 bp	Yes
B12-21	TTN	chr2:179,585,773C>A	chr2:179,531,966C>A	53,807 bp	2.17	98.04%	47,443 bp	Yes
B12-21	TTN	chr2:179,585,773C>A	chr2:179,531,966C>A	53,807 bp	2.17	98.04%	64,199 bp	Yes
B12-21	TTN	chr2:179,585,773C>A	chr2:179,531,966C>A	53,807 bp	2.17	98.04%	67,034 bp	Yes

Table 2: Gene, variant distance and RVIS score for clinically-relevant genes *(continued)*

Sample	Gene	Var1	Var2	Var distance	RVIS score	RVIS %	Molecule length	Var phased?
B12-21	TTN	chr2:179,585,773C>A	chr2:179,531,966C>A	53,807 bp	2.17	98.04%	90,767 bp	Yes
B12-21	TTN	chr2:179,585,773C>A	chr2:179,531,966C>A	53,807 bp	2.17	98.04%	93,253 bp	Yes
UC-394	TTN	chr2:179,584,098C>T	chr2:179,395,221T>A	188,877 bp	2.17	98.04%	28,033 bp	No
UC-394	TTN	chr2:179,584,098C>T	chr2:179,395,221T>A	188,877 bp	2.17	98.04%	18,841 bp	No
UC-394	TTN	chr2:179,584,098C>T	chr2:179,395,221T>A	188,877 bp	2.17	98.04%	16,791 bp	No
UC-394	TTN	chr2:179,584,098C>T	chr2:179,395,221T>A	188,877 bp	2.17	98.04%	13,118 bp	Yes
UC-394	TTN	chr2:179,584,098C>T	chr2:179,395,221T>A	188,877 bp	2.17	98.04%	18,192 bp	No
UC-394	TTN	chr2:179,584,098C>T	chr2:179,395,221T>A	188,877 bp	2.17	98.04%	32,530 bp	No
UC-394	TTN	chr2:179,584,098C>T	chr2:179,395,221T>A	188,877 bp	2.17	98.04%	30,653 bp	No
UC-394	TTN	chr2:179,584,098C>T	chr2:179,395,221T>A	188,877 bp	2.17	98.04%	88,605 bp	Yes
UC-394	TTN	chr2:179,584,098C>T	chr2:179,395,221T>A	188,877 bp	2.17	98.04%	87,045 bp	Yes
UC-394	TTN	chr2:179,584,098C>T	chr2:179,395,221T>A	188,877 bp	2.17	98.04%	69,939 bp	Yes
UC-394	TTN	chr2:179,584,098C>T	chr2:179,395,221T>A	188,877 bp	2.17	98.04%	89,863 bp	Yes

Table 3: Deletion reciprocal comparison to svclassify ground truth dataset

	Long Ranger calls	LR-overlaps	svclassify calls	svclassify-overlaps
A	23	9	11	8
B	6839	2428	2665	2428

Different intersections of Long-Ranger SV calls with a ground truth dataset published (Parikh et al. 2016). Comparison class identified in the most left column. A. Large deletions (>=30kb) intersected against all deletions >=30kb in the ground truth set. B. Smaller deletions (<30kb), marked as PASS by our algorithm, intersected against the full deletion ground truth deletion set.

Table 4: Gene, variant type and pipeline call for clinically-relevant genes

Sample	Gene	Variant type	Source	Assay	Calc mean length	Region phased?	Called by >=1 method?
A	MSH2	Single Exon Duplication	Archival DNA	SureSelectV6, 7.25Gb (60x)	64kb	No	No
A	MSH2	Single Exon Duplication	Archival DNA	SureSelectV6, 12Gb (100x)	53kb	No	No
B	PMS2	Single Exon Duplication	Archival DNA	SureSelectV6, 7.25Gb (60x)	65kb	Yes	Yes
B	PMS2	Single Exon Duplication	Archival DNA	SureSelectV6, 12Gb (100x)	59kb	Yes	Yes
C	BRCA1	Single Exon Duplication	Cell line	SureSelectV6, 7.25Gb (60x)	96kb	No	No
C	BRCA1	Single Exon Duplication	Cell line	SureSelectV6, 12Gb (100x)	78kb	No	No
C	BRCA1	Single Exon Duplication	Cell line	Whole Genome, 128Gb (30x)	88kb	No	No
C	BRCA1	Single Exon Duplication	Archival DNA	SureSelectV6, 7.25Gb (60x)	28kb	No	No
C	BRCA1	Single Exon Duplication	Archival DNA	SureSelectV6, 12Gb (100x)	27kb	No	No
D	BRCA2	Single Exon Duplication	Archival DNA	SureSelectV6, 7.25Gb (60x)	24kb	No	No?
D	BRCA2	Single Exon Duplication	Archival DNA	SureSelectV6, 12Gb (100x)	19kb	Yes	Yes
E	BRCA1	Two exon deletion	Cell line	SureSelectV6, 7.25Gb (60x)	106kb	No	No
E	BRCA1	Two exon deletion	Cell line	SureSelectV6, 12Gb (100x)	98kb	No	No
E	BRCA1	Two exon deletion	Archival DNA	SureSelectV6, 7.25Gb (60x)	71kb	No	No
E	BRCA1	Two exon deletion	Archival DNA	SureSelectV6, 12Gb (100x)	80kb	No	No
F	BRCA1	Two exon deletion	Cell line	SureSelectV6, 7.25Gb (60x)	97kb	Yes	Yes
F	BRCA1	Two exon deletion	Cell line	SureSelectV6, 12Gb (100x)	107kb	Yes	Yes

Table 4: Gene, variant type and pipeline call for clinically-relevant genes

(continued)

Sample	Gene	Variant type	Source	Assay	Calc mean length	Region phased?	Called by >=1 method?
F	BRCA1	Two exon deletion	Archival DNA	SureSelectV6, 7.25Gb (60x)	15kb	No	No
F	BRCA1	Two exon deletion	Archival DNA	SureSelectV6, 12Gb (100x)	12kb	Yes	Yes
G	PMS2	Two exon deletion	Archival DNA	SureSelectV6, 7.25Gb (60x)	57kb	Yes	Yes
G	PMS2	Two exon deletion	Archival DNA	SureSelectV6, 12Gb (100x)	48kb	Yes	Yes
H	PMS2	2-3 exon deletion	Archival DNA	SureSelectV6, 7.25Gb (60x)	54kb	Yes	Yes
H	PMS2	2-3 exon deletion	Archival DNA	SureSelectV6, 12Gb (100x)	42kb	Yes	Yes
I	PMS2	Large structural variant	Archival DNA	SureSelectV6, 7.25Gb (60x)	43kb	Yes	No
I	PMS2	Large structural variant	Archival DNA	SureSelectV6, 12Gb (100x)	35kb	Yes	No
I	PMS2	Large structural variant	Archival DNA	Whole genome, 128Gb (30x)	28kb	Yes	No
I	MSH2	Two exon deletion	Archival DNA	SureSelectV6, 7.25Gb (60x)	43kb	No	No
I	MSH2	Two exon deletion	Archival DNA	SureSelectV6, 12Gb (100x)	35kb	No	No
I	MSH2	Two exon deletion	Archival DNA	Whole genome, 128Gb (30x)	28kb	Yes	Yes

References

Aleman F. 2017. The necessity of diploid genome sequencing to unravel the genetic component of complex phenotypes. *Front Genet* **8**: 148.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**: 363–376.

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online mendelian inheritance in man (OMIM), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43**: D789–D798.

Askree SH, Chin ELH, Bean LH, Coffee B, Tanner A, Hegde M. 2013. Detection limit of intragenic deletions with targeted array comparative genomic hybridization. *BMC Genet* **14**: 116.

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

Bishara A, Liu Y, Weng Z, Kashef-Haghighi D, Newburger DE, West R, Sidow A, Batzoglou S. 2015. Read clouds uncover variation in complex regions of the human genome. *Genome Res* **25**: 1570–1580.

Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. 2012. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics* **13**: 375.

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al. 2014. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*.

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez O, Guo L, Collins RL, et al. 2017. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv.*

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, GTEx Consortium, et al. 2017. The impact of structural variation on human gene expression. *Nat Genet.*

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen H-C, Agarwala R, McLaren WM, Ritchie GRS, et al. 2011. Modernizing reference genome assemblies. *PLoS Biol* **9**: e1001091.

Church DM, Schneider VA, Steinberg KM, Schatz MC, Quinlan AR, Chin C-S, Kitts PA, Aken B, Marth GT, Hoffman MM, et al. 2015. Extending reference assembly models. *Genome Biol* **16**: 13.

Collins FS. 1998. New goals for the U.S. human genome project: 1998-2003. *Science* **282**: 682–689.

Collins RL, Brand H, Redin CE, Hanscom C, Antolik C, Stone MR, Glessner JT, Mason T, Pregno G, Dorrani N, et al. 2017. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol* **18**: 36.

Consortium IHGS. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931–945.

Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, Ribeca P. 2012. Fast computation and applications of genome mappability. *PLoS One* **7**: e30377.

Elyanow R, Wu H-T, Raphael BJ. 2017. Identifying structural variants using linked-read sequencing data. *bioRxiv* 190454.

Garcia S, Williams S, Xu AW, Herschleb J, Marks P, Stafford D, Church DM. 2017. Linked-Read sequencing resolves complex structural variants. *bioRxiv* 231662.

Genomics B. 2017. Bionano human structural variations white paper.

Greer SU, Nadauld LD, Lau BT, Chen J, Wood-Bouwens C, Ford JM, Kuo CJ, Ji HP. 2017. Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases. *Genome*

Med **9**: 57.

Huddleston J, Chaisson MJ, Meltz Steinberg K, Warren W, Hoekzema K, Gordon DS, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, et al. 2016. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res*.

Huddleston J, Eichler EE. 2016. An incomplete understanding of human genetic variation. *Genetics* **202**: 1251–1254.

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. 2012. De novo assembly and genotyping of variants using colored de bruijn graphs. *Nat Genet* **44**: 226–232.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**: 56–64.

Krusche P. Hap.py.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285–291.

Li H. 2013. Aligning sequence reads , clone sequences and assembly contigs with BWA-MEM. *ArXiv* **00**: 1–2.

Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier LD, Neale B, MacArthur D. 2017. New synthetic-diploid benchmark for accurate variant calling evaluation. *bioRxiv* 223297.

Mandelker D, Amr SS, Pugh T, Gowrisankar S, Shakhbatyan R, Duffy E, Bowser M, Harrison B, Lafferty K, Mahanta L, et al. 2014. Comprehensive diagnostic testing for stereocilin: An approach for analyzing medically important genes with high homology. *J Mol Diagn* **16**: 639–647.

Mandelker D, Schmidt RJ, Ankala A, McDonald Gibson K, Bowser M, Sharma H, Duffy E, Hegde M, Santani A, Lebo M, et al. 2016. Navigating highly homologous genes in a molecular diagnostic setting: A resource for clinical next-generation sequencing. *Genet Med* **18**: 1282–1289.

Nakano K, Shiroma A, Shimoji M, Tamotsu H, Ashimine N, Ohki S, Shinzato M, Minami M, Nakanishi T, Teruya K, et al. 2017. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Hum Cell* **30**: 149–161.

Novak AM, Hickey G, Garrison E, Blum S, Connelly A, Dilthey A, Eizenga J, Saleh Elmohamed MA, Guthrie S, Kahles A, et al. 2017. Genome graphs. *bioRxiv* 101378.

Parikh H, Mohiyuddin M, Lam HYK, Iyer H, Chen D, Pratt M, Bartha G, Spies N, Losert W, Zook JM, et al. 2016. Svclassify: A method to establish benchmark structural variant calls. *BMC Genomics* 1–16.

Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. 2013. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* **9**: e1003709.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Ramaker RC, Savic D, Hardigan AA, Newberry K, Cooper GM, Myers RM, Cooper SJ. 2017. A genome-wide interactome of DNA-associated proteins in the human liver. *bioRxiv* 111385.

Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, et al. 2017. Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27**: 849–864.

Spies N, Weng Z, Bishara A, McDaniel J, Catoe D, Zook JM, Salit M, West RB, Batzoglou S, Sidow A. 2016. Genome-wide reconstruction of complex structural variants using read clouds. *bioRxiv* 074518.

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

Nature **526**: 75–81.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res* **27**: 757–767.

Wittler R, Marschall T, Sch A, Veli M. 2015. Repeat- and Error-Aware comparison of deletions. *Bioinformatics* 1–8.

Xia LC, Bell JM, Wood-Bouwens C, Chen JJ, Zhang NR, Ji HP. 2017. Identification of large rearrangements in cancer genomes with barcode linked reads. *Nucleic Acids Res.*

Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, et al. 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* 1–11.

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3**: 160025.

Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. 2014. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* **32**: 246–251.