Resolving the full spectrum of human genome variation using Linked-Reads

Patrick Marks, Sarah Garcia, Alvaro Martinez Barrio, et al.

Genome Res. 2019 29: 635-645 originally published online March 20, 2019
Access the most recent version at doi:10.1101/gr.234443.118

Method

Resolving the full spectrum of human genome variation using Linked-Reads

Patrick Marks,[1] Sarah Garcia,[1] Alvaro Martinez Barrio,[1] Kamila Belhocine,[1] Jorge Bernate,[1] Rajiv Bharadwaj,[1] Keith Bjornson,[1] Claudia Catalanotti,[1] Josh Delaney,[1] Adrian Fehr,[1] Ian T. Fiddes,[1] Brendan Galvin,[1] Haynes Heaton,[1,5] Jill Herschleb,[1] Christopher Hindson,[1] Esty Holt,[2] Cassandra B. Jabara,[1,6] Susanna Jett,[1,7] Nikka Keivanfar,[1] Sofia Kyriazopoulou-Panagiotopoulou,[1,8] Monkol Lek,[3,4] Bill Lin,[1] Adam Lowe,[1] Shazia Mahamdallie,[2] Shamoni Maheshwari,[1] Tony Makarewicz,[1] Jamie Marshall,[4] Francesca Meschi,[1] Christopher J. O'Keefe,[1] Heather Ordonez,[1] Pranav Patel,[1] Andrew Price,[1] Ariel Royall,[1] Elise Ruark,[2] Sheila Seal,[2] Michael Schnall-Levin,[1] Preyas Shah,[1] David Stafford,[1] Stephen Williams,[1] Indira Wu,[1] Andrew Wei Xu,[1] Nazneen Rahman,[2] Daniel MacArthur,[3,4] and Deanna M. Church[1,9]

[1]10x Genomics, Pleasanton, California 94566, USA; [2]The Institute of Cancer Research, Division of Genetics and Epidemiology, London SM2 5NG, United Kingdom; [3]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA; [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA

Large-scale population analyses coupled with advances in technology have demonstrated that the human genome is more diverse than originally thought. To date, this diversity has largely been uncovered using short-read whole-genome sequencing. However, these short-read approaches fail to give a complete picture of a genome. They struggle to identify structural events, cannot access repetitive regions, and fail to resolve the human genome into haplotypes. Here, we describe an approach that retains long range information while maintaining the advantages of short reads. Starting from ~1 ng of high molecular weight DNA, we produce barcoded short-read libraries. Novel informatic approaches allow for the barcoded short reads to be associated with their original long molecules producing a novel data type known as "Linked-Reads". This approach allows for simultaneous detection of small and large variants from a single library. In this manuscript, we show the advantages of Linked-Reads over standard short-read approaches for reference-based analysis. Linked-Reads allow mapping to 38 Mb of sequence not accessible to short reads, adding sequence in 423 difficult-to-sequence genes including disease-relevant genes *STRC*, *SMN1*, and *SMN2*. Both Linked-Read whole-genome and whole-exome sequencing identify complex structural variations, including balanced events and single exon deletions and duplications. Further, Linked-Reads extend the region of high-confidence calls by 68.9 Mb. The data presented here show that Linked-Reads provide a scalable approach for comprehensive genome analysis that is not possible using short reads alone.

[Supplemental material is available for this article.]

Since the completion of the human genome project, many studies have applied whole-genome sequencing to thousands of individuals from diverse populations, reshaping our understanding of human variation (The 1000 Genomes Project Consortium 2015; Sudmant et al. 2015; Lek et al. 2016). To date, most genome analyses were performed with short reads, resulting in robust analyses of small variants over nonrepetitive parts of the genome. However, recent technical advances in both sequencing and mapping have revealed that despite extensive information garnered from short-read large population surveys, we are still underrepresenting the amount of structural variation in the human population (Chaisson et al. 2015, 2017; Huddleston and Eichler 2016; Collins et al. 2017).

The reconstruction of haplotypes (phasing) can be important for many biological studies but is currently not feasible for single samples sequenced with short reads. When analyzing data from rare disease cohorts, knowing if potentially pathogenic variants are in *cis* or *trans* is necessary for interpreting clinical impact. In addition, haplotype information is necessary for understanding allele-specific impacts on gene expression (Ramaker et al. 2017). Studies also show that haplotype information can be critical

Present addresses: [5]Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK; [6]Purigen Biosystems, Inc., Pleasanton, CA 94588, USA; [7]LevitasBio, Inc., San Francisco, CA 94110, USA; [8]Illumina, Inc., San Francisco, CA 94158, USA; [9]Inscripta, Inc., Boulder, CO 80301, USA

Corresponding author: apps-paper@10xgenomics.com

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.234443.118. Freely available online through the *Genome Research* Open Access option.

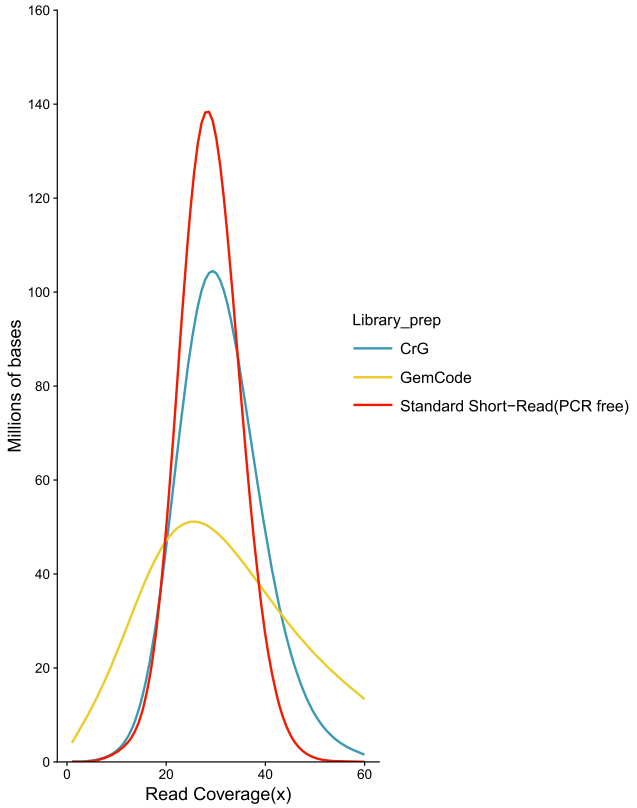

Figure 1. Coverage evenness. Distribution of read coverage for the entire human genome (GRCh37). Comparisons between 10x Genomics Chromium Genome (CrG), 10x Genomics GemCode (GemCode), and Illumina TruSeq PCR-free standard short-read NGS library preparations (Standard Short-Read [PCR-Free]). Sequencing was performed in an effort to match coverage (Methods). Note the shift of the CrG curve to the *right*, showing the improved coverage of Chromium versus GemCode. The *x*-axis represents the fold read coverage across the genome, and the *y*-axis represents the total number of bases covered at any given read depth.

for variant identification, particularly for heterozygous SVs (Huddleston and Eichler 2016).

The limitations of short reads suggest the need for improved methods for genome analysis. Several long-molecule sequencing and mapping approaches have been developed (Carneiro et al. 2012; Bionano Genomics 2017; Nakano et al. 2017), but their high input requirements, error rates, and costs make them intractable for many applications, particularly those requiring thousands of samples (Chaisson et al. 2017). To address this need, we developed a technology that retains long range information while maintaining the benefits of short-read sequencing (Zheng et al. 2016). The core data type, Linked-Reads, is generated by performing haplotype limiting dilution of long DNA molecules into more than 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.

Here, we describe both biochemistry and algorithmic improvements over the original Linked-Reads platform, GemCode, using the Chromium System. It is important to note that Linked-Reads are paired-end short reads with a barcode on read 1 and can be used by many common short-read tools. To fully realize the potential of Linked-Reads, additional algorithms that take advantage of these barcoded sequences and molecule information must be combined with short-read algorithms. In the following text, when we refer to Linked-Read WGS (lrWGS) we are referring to the combination of biochemistry and algorithm approaches applied. We use srWGS (short-read whole genome sequencing) and srWES (short-read whole exome sequencing) to refer to whole-genome and whole-exome results from Illumina TruSeq PCR-free processed with a GATK best practices pipeline, as described subsequently.

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 due to coverage imbalances in the GemCode library. By modifying the biochemistry to include isothermal amplification, we were able to obtain more even genome coverage, approaching that of PCR-free short-read preparations and eliminating the need for an additional library (Fig. 1).

Additional improvements include increasing the number of barcodes from 737,000 to 4,000,000 and the number of partitions from 100,000 to more than 1,000,000. This allows for fewer DNA molecules per partition and thus greatly reduces the rate at which two allelic loci occur in the same GEM (Supplemental Fig. 1). This lowered rate of barcode sharing increases the probability of correctly associating a short read to its molecule of origin.

Improved genome and exome alignments

Several improvements were made in the Long Ranger analysis pipeline to better take advantage of the Linked-Read data type. The first of these, the Lariat aligner (https://github.com/10XGenomics/lariat), expands on the "Read-Cloud" approach (Bishara et al. 2015; Supplemental Methods 1). This approach allows for the recovery of 36–44 Mb of genome coverage when compared to PCR-free short reads. Conversely, only 1–4 Mb of the genome has coverage in the PCR-free data but not lrWGS (Fig. 2). The amount of recovered alignments using lrWGS varies from chromosome to

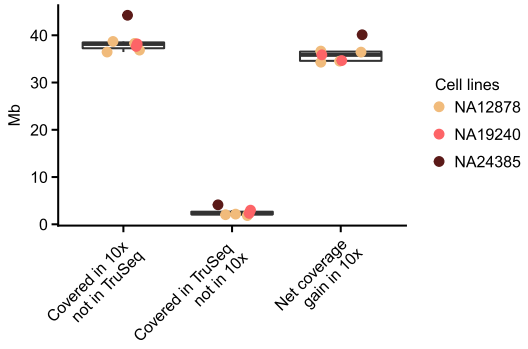

Figure 2. Comparison of unique genome coverage by assay. The *y*-axis shows the amount of sequence with a coverage of ≥5 reads at MapQ ≥30. Column 1 shows amount of the genome covered by 10x Chromium where PCR-free TruSeq does not meet that metric. Column 2 shows the amount of the genome covered by PCR-free 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 (Methods).

chromosome, but is consistent across samples (Supplemental Fig. 2). The ability of lrWGS + Lariat to rescue repetitive sequence depends on repeat elements being far enough from each other that they are not likely to share a barcode, and repeat type and distribution differs by chromosome. The sequence gained using lrWGS is dominated by regions annotated as segmental duplications (~75%), with the alignments to the decoy sequence accounting for another 13% and exonic sequences accounting for ~5% (Fig. 2; Supplemental Methods 1.2; Supplemental Table 1). Input molecule length also impacts the amount of sequence recovered (Supplemental Fig. 3).

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 Fig. 4). In 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. For the 71 NGS "dead zone" genes relevant to Mendelian disease, we see a net improvement in 51 of these genes using lrWGS and 41 genes using lrWES (Fig. 3).

Small variant calling

Next, we assessed the performance of Linked-Reads for small variant calling (<50 bp). We used control samples, NA12878 and

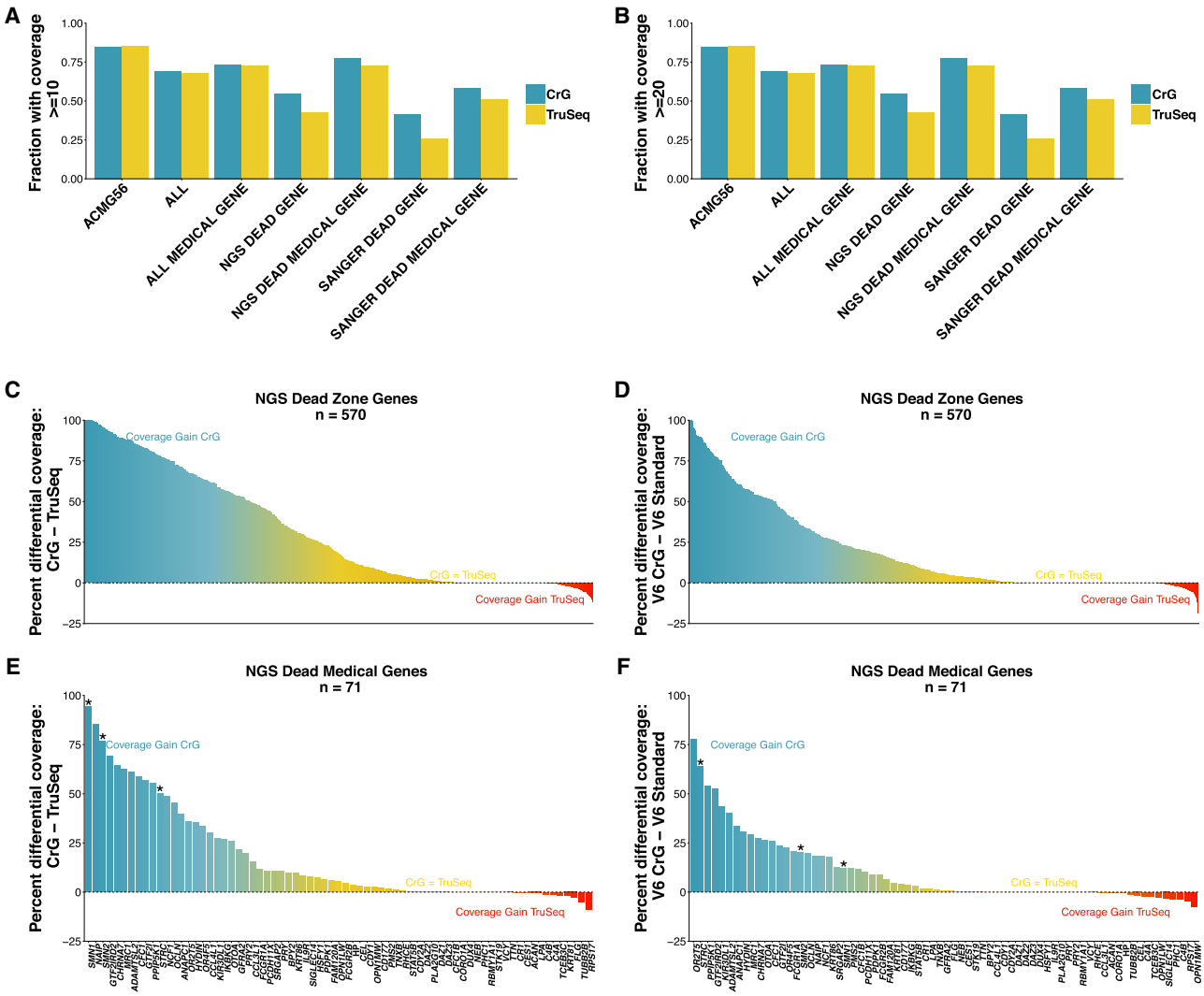

Figure 3. Gene finishing metrics. Gene finishing metrics for whole-genome and whole-exome sequencing across selected gene sets. Genome is shown on *left*, exome on *right*. Gene finishing is defined as the percentage of exonic bases with at least 10-fold coverage for genome (*A*) and at least 20× for exome (*B*) (Mapping quality score ≥MapQ30). (*A*,*B*) Gene finishing statistics for seven disease-relevant gene panels. Shown is the average value across all genes in each panel. Although Chromium provides a coverage advantage in all panel sets, the impact is particularly profound for "NGS Dead Zone" genes. (*C*–*F*) 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. (*C*,*D*) The 570 NGS "dead zone" genes for genome (*C*) and exome (*D*). (*E*,*F*) The graphs are limited to the list of NGS dead zone genes implicated in Mendelian disease. In *C*–*F*, the blue coloring highlights genes that are inaccessible to short-read approaches, but accessible using CrG; the yellow coloring indicates genes where CrG is equivalent to short reads or provides only modest improvement. The red coloring shows genes with a slight coverage increase in TruSeq, although these genes are typically still accessible to CrG. (*) Genes *SMN1*, *SMN2*, and *STRC*. The comparison was performed on samples with matched coverage (Methods).

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 (lrWGS) 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/). The number of calls was comparable between data sets and were largely overlapping (Table 1).

In order to assess the accuracy of the variant calling in each data set, we used the hap.py tool (https://github.com/Illumina/hap.py, commit 6c907ce) to compare the lrWGS 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 call set version as it was the last GIAB data set that did not include 10x data as an input for call set curation. This necessitated the use of GRCh37 as a reference assembly rather than the more current GRCh38 reference assembly, limiting analysis to the 82.67% of SNV calls that overlap high-confidence regions. Initial results suggested that the lrWGS calls had comparable sensitivity and specificity for SNVs (Table 1; Supplemental Table 2). We observed slightly diminished indel sensitivity and specificity, driven largely by regions with extreme GC content and low complexity sequences (LCRs).

The GIAB high-confidence data set is known to be conservative, so we explored whether there was evidence for variants called outside of the GIAB set. We utilized publicly available 40-fold coverage Pacific Biosciences (PacBio) data sets available from the GIAB consortium (Zook et al. 2016) and PCR- short-read data to evaluate Linked-Read putative false positive variant calls. Initial manual inspection of 25 random locations suggested that roughly half of the hap.py identified lrWGS false positive calls were well supported by short-read or PacBio evidence and were haplotype consistent (Supplemental Table 3). We then did a global analysis of all 9513 SNVs and 18,030 indel putative false positive calls identified in NA12878 and looked for evidence of the alternate alleles in aligned PacBio reads only. This analysis provided evidence that 2377 SNVs and 12,812 indels of the GIAB determined false positive calls were likely valid calls (Supplemental Fig. 5; Supplemental File 1). This prompted us to extend our analysis to include 69.72 Mb for NA12878 and 70.66 Mb for NA24385 of the genome in addition to the GIAB-defined confident regions (for details on GIAB++ BED, see Methods). We reanalyzed the variant calls with the hap.py tool with the augmented confident regions. This allowed us to identify an additional 19,688 SNVs and 5444 indels as likely true positives. We anticipate that this is a conservative estimate since our hap.py-defined false positive calls are inflated due to lack of PacBio coverage in many of these regions. Of the total putative false positive calls exclusive to the GIAB++ analysis, 61.95% (45,665) of SNVs and 42.08% (4637) of indels could not be validated because of little or no PacBio read coverage (Supplemental Fig. 5). These data show that the lrWGS approach identifies more small variants than short-read only approaches, driven by an increase in the percentage of the genome for which lrWGS can obtain high-quality alignments (Table 1).

Haplotype reconstruction and phasing

An advantage of Linked-Reads is the ability to reconstruct multi-megabase haplotypes (phase blocks) from genome sequence data 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, 9.58 Mb for NA24385, 16.8 Mb for NA19240, and 302 kb for lrWES using Agilent SureSelect v6 baits on NA12878. This allowed for complete phasing of 91.1% of genes in the NA12878 genome, 90.9% in NA24385, and 91.0% in NA19240, and an average of 91% in the NA12878 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

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

Variable	NA12878 10xLR	NA12878 PCR-	NA24385 10xLR	NA24385 PCR-
Total variants	4,600,606	4,651,391	4,504,190	4,564,102
Total SNVs	3,808,856	3,760,296	3,731,448	3,689,866
Sensitivity (SNVs)	0.996525983	0.997887311	0.997246162	0.998425022
Specificity (SNVs)	0.996982928	0.998474689	0.997754891	0.999012527
SNVs in confident regions	3,153,057	3,152,799	3,053,304	3,053,249
SNVs in truth set	3,143,316	3,147,610	3,046,234	3,049,835
Sensitivity (SNVs) (++)	0.994498718	0.995408359	0.996619732	0.997396835
Specificity (SNVs) (++)	0.974517521	0.987927497	0.970378148	0.983854181
SNVs in confident regions (++)	3,266,048	3,224,849	3,151,491	3,111,146
SNVs in truth set (++)	3,182,558	3,185,469	3,057,434	3,059,818
Total indels	791,750	891,095	772,742	874,236
Sensitivity (indels)	0.933975195	0.973396905	0.933085491	0.977287898
Specificity (indels)	0.950130965	0.982073022	0.949342412	0.985153439
Indels in confident regions	361,547	368,216	347,786	354,897
Indels in truth set	334,577	348,699	321,517	336,748
Sensitivity (indels) (++)	0.92264	0.964579042	0.905634526	0.974315427
Specificity (indels) (++)	0.923436804	0.963676143	0.88549083	0.93319472
Indels in confident regions (++)	379,399	383,935	474,879	491,054
Indels in truth set (++)	341,279	356,792	411,130	442,309

The table shows the counts of variants (SNV and indel) from variant calls generated in four experiments: NA12878 Linked-Reads WGS data run through Long Ranger (NA12878 lrWGS), NA12878 TruSeq PCR-free data run through GATK Best Practices pipeline (NA12878 srWGS), NA24385 Linked-Reads WGS data run through Long Ranger (NA24385 lrWGS), and NA24385 TruSeq PCR-free data run through GATK Best Practices pipeline (NA24385 srWGS). These variants were compared to the GIAB VCF truth set and GIAB BED confident regions using hap.py, and data are 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).

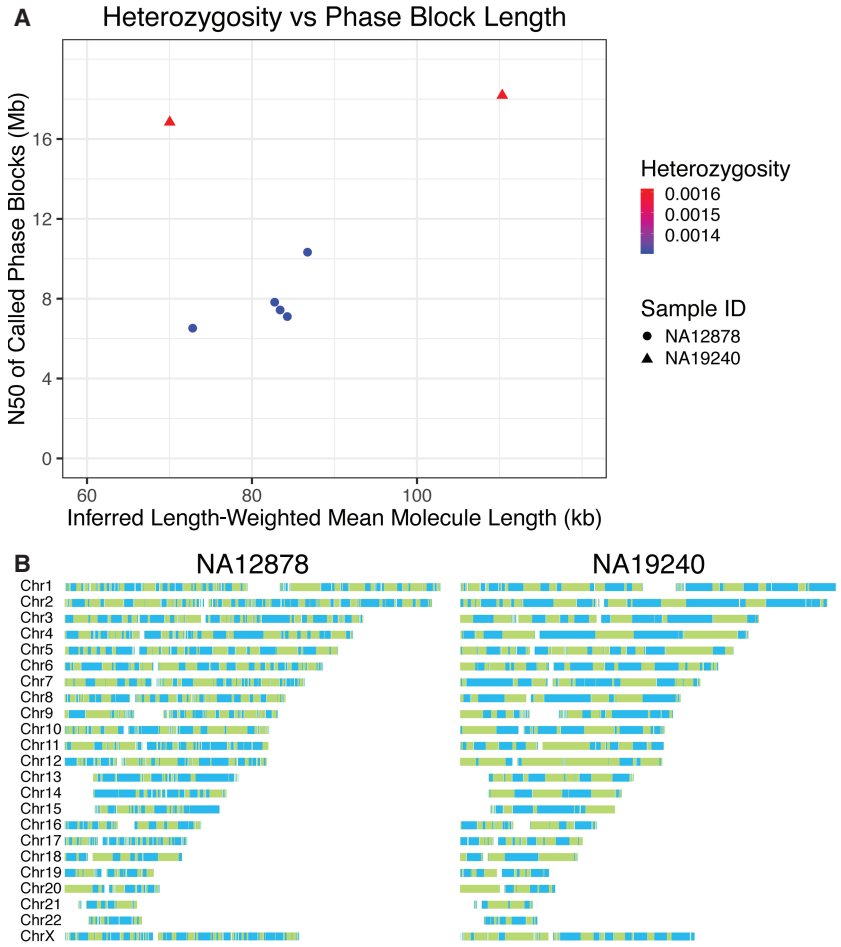

Figure 4. Haplotype reconstruction and phasing. (*A*) Inferred length-weighted mean molecule length plotted against N50 of called Phase blocks (both metrics reported by Long Ranger) and differentiated by sample ID and heterozygosity. Heterozygosity was calculated by dividing the total number of heterozygous positions called by Long Ranger by the total number of non-N bases in the reference genome (GRCh37). Two replicates of NA19240 and five 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. Note the longer phase blocks in the more diverse NA19240 sample.

longer in more diverse samples (Fig. 4; Supplemental Fig. 6). For samples with similar heterozygosity, longer input molecules will increase phase block lengths (Supplemental Fig. 7).

We assessed the accuracy of our phasing calls by comparing the Linked-Read phasing results for NA12878 with the Illumina Platinum genomes (Eberle et al. 2017) phasing results derived from jointly phasing the 17-member CEPH pedigree. Following this analysis (Amini et al. 2014), we decompose phasing errors into "short-switches," small numbers of isolated variants incorrectly phased, and "long-switches," errors in which an incorrect phasing junction persists for many variants across a longer distance. The rate of each switch type is measured per phased heterozygous variant. We also measure (1) the rate at which a SNP is correctly phased to other variants in its phase block (heavily penalizing long switch errors inside large phase blocks), and (2) the rate at which a SNP inside a gene is correctly phased to other variants in the gene. Independent studies have demonstrated that Linked-Read phasing has best-in-class accuracy compared to a variety of other phasing methods (Chaisson et al. 2017; Choi et al. 2018). Short switch error rates average $\sim 2 \times 10^{-4}$, long switch error rates average $\sim 2 \times 10^{-5}$, and within-phase-block correct rate of ~0.98 (Supplemental Table 4).

Phase block construction using lrWES is, in addition, constrained by the capture bait set and reduced variation in coding sequences. In order to analyze additional factors impacting phase block construction, we assessed four samples with known compound heterozygous variants in three Mendelian disease genes, *DYSF*, *POMT2*, and *TTN*. The variant separation ranged from 33 to >188 kb (Table 2). Initial DNA extractions yielded long molecules ranging in mean size from 75 to 112 kb. We analyzed these samples using the Agilent SureSelect V6 exome bait set, with down-sampling of sequence data to both 7.25 Gb (~60-fold coverage) and 12 Gb of sequence (~100-fold 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. By comparing phasing of NA12878 Linked-Read exome data to phasing from pedigree analysis of the Illumina Platinum Genomes CEPH pedigree, we determine that the global probability a SNP is phased correctly within a gene ranges from 99.95–99.99%, making misphasing of two heterozygous variants in a gene relative to each other a very rare event.

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 reextraction. For this reason, we wanted to understand the impact of reduced molecule length on phasing of genes and variants in these samples. We took freshly extracted long molecules and sheared them to various sizes, aiming to assess lengths ranging from 5 kb to the original full length (Table 2; Supplemental Table 5). These results illustrate the complex interplay between molecule length distribution and 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 20-kb lengths. This appeared to be due to a higher rate of heterozygous variation in *TTN*, allowing the phasing of distant heterozygous sites to occur by phasing the many other heterozygous variants that occurred between them. A general lack of variation in *POMT2* precluded such "stitching" together of shorter molecules by phasing of intermediate heterozygous variation. To confirm this, we assessed the maximum distance between heterozygous sites observed in each gene in each sample. As expected, when the maximum distance between heterozygous SNPs is greater than the molecule length (negative values), the ability to phase

Table 2. Gene, variant distance, and Residual Variation Intolerance Score (RVIS) for clinically relevant genes

Sample	Gene	Var1	Var2	Variant distance (bp)	RVIS score	RVIS percentile (%)	Molecule length (bp)
B12-38	*DYSF*	Chr 2: 71,778,243dupT	Chr 2: 71,817,342_71,817,343delinsAA	39,097	−1.31	4.65	18,461
B12-112	*POMT2*	Chr 14: 77,745,107A>G	Chr 14: 77,778,305C>T	33,198	−0.93	9.68	54,569
B12-21	*TTN*	Chr 2: 179,585,773C>A	Chr 2: 179,531,966C>A	53,807	2.17	98.04	17,432
UC-394	*TTN*	Chr 2: 179,584,098C>T	Chr 2: 179,395,221T>A	188,877	2.17	98.04	13,118[a]

Impact of molecule length and constraint on the ability of Linked-Reads to phase causative variants. As molecule length increases within a sample, the likelihood that two causative variants will be phased relative to each other also increases. However, genes that are not highly constrained (e.g., *TTN*) are more likely to show phasing between distant variants at small molecule lengths because more heterozygous variants are likely to occur between those variants than in highly constrained genes (Petrovski et al. 2013). Shown are the shortest molecule lengths at which phasing was achieved for each sample. For results at all observed shear lengths, see Supplemental Table 5.
[a]For this sample, phasing was achieved in one 13-kb length sample, but was not reproducible until ~70 kb lengths.

causative SNPs decreases (Fig. 5). There are exceptions to this as longer molecules in the size distribution will sometimes allow tiling between variants, extending phase block size beyond what would be expected based on the mean length alone.

Linked-Reads allow for the reconstruction of long haplotypes. Optimizing for long input molecules provides for maximum phase block size, but shorter molecule lengths can provide gene-level phasing. Further, in the context of sequencing for disease identification, causative heterozygous variants would be expected to aid in phasing of the disease-causing gene as they would provide the necessary variation to distinguish the two haplotypes.

Structural variant detection

Structural variants remain one of the most difficult types of variation to accurately ascertain, in part because they tend to cluster in duplicated and repetitive regions, but also because the various signals for these events can be challenging to detect with short reads. Another complicating factor is that there are many types of structural variants, and each requires the detection of a different signal (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).

Large-scale SVs (>30 kb)

Long Ranger uses 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, whereas the barcode overlap method can detect a variety of SVs but fails to detect terminal events (Supplemental Section 3). We used two approaches to assess lrWGS performance on large SVs. First, we compared SV calls from the NA12878 sample to validated calls described in a publication of a structural variant classifier, svclassify (Parikh et al. 2016). Next, we obtained the GeT-RM CNVPanel, a collection of known events including large deletions, duplications, inversions, balanced translocations, and unbalanced translocations designed to assess performance of clinical aCGH.

The validated call set published with svclassify (Parikh et al. 2016) contains deletions and insertions, but no balanced events. In contrast, the Long Ranger pipeline output contains deletions, duplications, and balanced events, but Long Ranger does not currently call insertions (Supplemental Table 6).

We first considered deletion variants >30 kb. The svclassify set contains 11 such deletion calls, Long Ranger calls 17 PASS events, and eight events are common to both (Table 3). All of the eight variants in common show Mendelian consistency and breakpoint agreement within the CEU/CEPH trio. Of the three svclassify calls not called by Long Ranger, one is called by Long Ranger as an event <30 kb, one is called but filtered to the candidate list due to overlap with a segmental duplication, and one is an error in the svclassify set relative to GRCh37.p13 (Supplemental Section

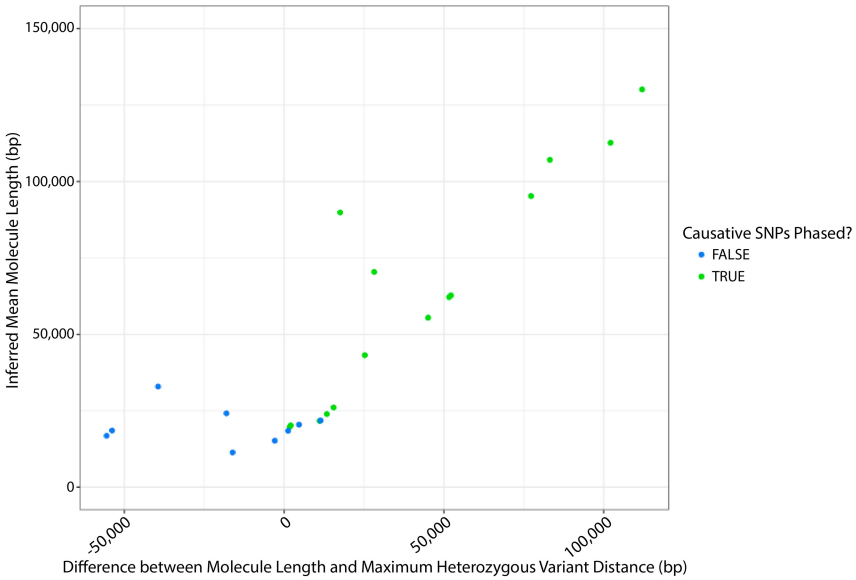

Figure 5. Validated example of impact of molecule length on phasing (7.25 Gb). Blue dots represent samples for which the variants of interest are not phased, and green dots represent samples for which there is phasing of the variants of interest. At longer molecule lengths (>50 kb), the molecule length 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 and variant distribution within the sample.

Table 3. SV intersections

Size group of PASS deletions	Total number of Long Ranger calls in each category	Total number of svclassify calls in each category	Total number of overlapping calls of Long Ranger and the svclassify data sets
≥30 kb	17	11	8
<30 kb	5136	2294	2024

Different intersections of Long Ranger SV calls with a ground truth data set (Parikh et al. 2016). Comparison class identified in the leftmost column. Large deletions (≥30 kb) intersected against all deletions ≥30 kb in the ground truth set. Smaller deletions (<30 kb), marked as PASS by our algorithm, intersected against the full deletion ground truth set.

4.1). We checked for Mendelian consistency in the nine events unique to the Long Ranger set. Eight of these events showed consistent inheritance, although two had inconsistent breakpoints when compared to the parents (Supplemental Table 7). The last event is a call in the telomeric region of Chr 2 that overlaps a known reference assembly issue. The call appeared to be made due to a drop in phased coverage on one haplotype immediately adjacent to a known reference gap, and is likely a false positive.

We next tested 23 samples with 29 balanced or unbalanced SVs from the GeT-RM CNVPanel available from Coriell. These samples have multiple, orthogonal assays confirming the presence of their described SVs. We detected 27 of the 29 SVs, correctly characterizing 22 of the 23 samples tested (Supplemental Table 8). One of the undetected events was in the "candidate" SV list as it overlaps a segmental duplication. The missed event is a balanced translocation with a breakpoint in a heterochromatic region of Chr 16. This region is represented by Ns in the reference assembly and will be invisible to any sequence-based reference dependent method (Schneider et al. 2017).

We also assessed the impact of sequence depth on large SV identification in the GeT-RM set. Large CNV signals were detectable with as little as 5 Gb (approximately onefold genomic read coverage) (Supplemental Fig. 8), and balanced events required ~50 Gb of sequence for the algorithm to call these events, with signal in the data suggesting algorithmic improvements could lessen this requirement (Supplemental Fig. 9).

Intermediate SV calls (50 bp–30 kb)

We next considered deletions between 50 bp and 30 kb in NA12878. These deletions were detected both using Long Ranger–specific algorithms as well as the Genome Analysis Toolkit (GATK) HaplotypeCaller. We obtained 1824 deletion calls from GATK and 4118 from Long Ranger (Table 4). These two sets were merged using SURVIVOR (Jeffares et al. 2017) resulting in 5136 merged deletion calls. This compares to 6965 deletions >50 bp per sample in a study combining the output of 13 different algorithms on short-read data and 9488 deletions >50 bp per sample on long-read data (Chaisson et al. 2017). To establish a comparison to existing methods, we ran the LUMPY (Layer et al. 2014) algorithm using the developer recommendations but without tuning parameters (Supplemental Table 9) and found 19,307 deletion calls in this size range.

Using both the output of Long Ranger and LUMPY, we compared our calls to the calls in svclassify. We correctly identified 88.4% of intermediate deletions present in the svclassify truth set (2107), and also called an additional 2048 SVs (49.6% precision) (Table 4). Combining the GATK and Long Ranger calls keeps recall roughly the same, but lowers the precision ~10% (Supplemental Table 9). We also compared the LUMPY results to svclassify and found 1263 true positives (55.4% recall). Of note, the Long Ranger calls provide improved detection of larger SVs, with an expected bump around 300 bp, likely accounted for by better representation of *Alus* (Fig. 6).

Although sensitivity of the Long Ranger approach is good, this comes at the expense of specificity (Table 4; Supplemental Table 9). Given the bias in specificity in phased versus unphased regions, we expect that algorithmic improvements will produce further gains in sensitivity and specificity for this class of variants. In addition, the small number of events <200 bp in the svclassify set is likely not representative of the true number of calls but rather technical/algorithmic limitations.

Analysis of samples from individuals with inherited disease

We went on to investigate the utility of Linked-Reads on samples with known disease-causing variants typically difficult to call with a standard, short-read exome. We obtained samples with known exon-level deletion and duplication events from a cohort that had been assessed using a high-depth NGS-based panel. We analyzed 12 samples from nine individuals using an Agilent SureSelect V6 Linked-Read exome at both 7.25 Gb (~60-fold raw coverage) and 12 Gb (~100-fold coverage) (Table 5). For three samples, patient-derived cell lines were available in addition to archival DNA, allowing investigation of the impact of DNA length on variant calling in this cohort.

We identified five of the nine known exon-level events in these samples in at least one sample/depth combination. In two samples, increasing depth to 12 Gb enabled calling not possible at 7.25 Gb (Samples D and F [archival]) (Table 5). For the three samples with matched cell line and archival DNA, two had variants that could not be called in either sample type at either depth, whereas 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. Because the algorithms for calling these variants rely on phasing and barcode information, there is a correlation between gene phasing and variant calling, with no variants successfully called in samples not phased over the causative gene.

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

Table 4. Intermediate SV calls

Intermediate SV metrics	Genome = NA12878
Number of deletion calls from Long Ranger	4118
Number of heterozygous calls	1699
Number of homozygous calls	2630
Number of calls that match svclassify truth set (recall)	2017 (88.4%)
Number of false positive calls (precision)	2048 (49.6%)

Intermediate SV (50 bp to 30 kb) results. The number of calls generated by the intermediate SV algorithms are reported and broken down by inferred zygosity. SURVIVOR (Jeffares et al. 2017) was used to merge these intermediate SVs with the svclassify (Parikh et al. 2016) truth set, which had also been subsetted to the same size range, and the resulting true positive and false positive rates are reported as well as the associated recall and precision.

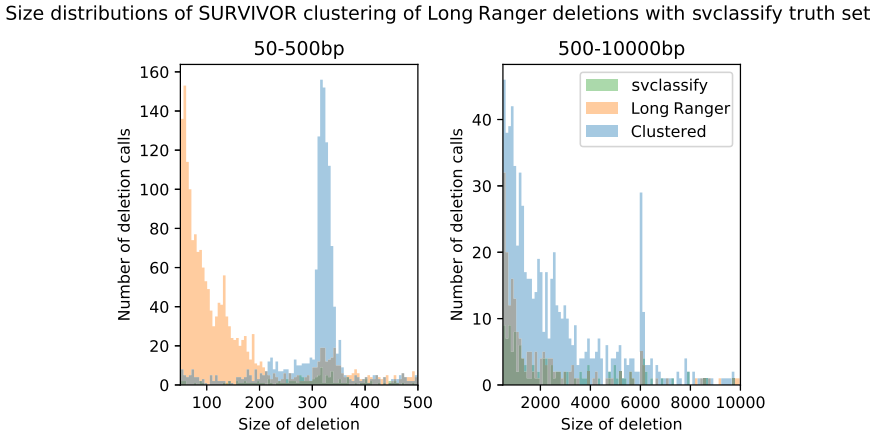

Figure 6. Deletion size distributions. Long Ranger calls intersected with the svclassify truth set by size. True positive calls are blue, false negative calls are green, and false positive calls are orange. Most false positives are present in the <250-bp size range, reflecting the lack of small deletions in the svclassify set. Peaks corresponding to *Alu* and L1/L2 elements can be seen at ~320 bp and ~6 kb, respectively.

was called. In the second case, there was insufficient heterozygous variation in the sample to allow phasing and the event was not called. This again demonstrates that phasing is dependent on molecule length as well as sample heterozygosity. In samples with decreased diversity in genes of interest, causative variant calling by Linked-Read sequencing was less likely (Supplemental Fig. 10). Generally, it should be possible to increase the probability of gene phasing in an exome assay by augmenting the bait set to provide coverage for common intronic variant SNPs. The addition of read coverage–based algorithms, such as those used with standard short-read exome sequencing, would also likely increase sensitivity in unphased regions, but were not used in this study.

One sample in this set contained both a single exon event and a large CNV in *PMS2*. Despite phasing of *PMS2* the variant was not called by genome or exome sequencing. Manual inspection of the data revealed increased phased barcode coverage in *PMS2*, supporting the presence of a large duplication missed by the SV calling algorithms, and providing evidence that improvements in the SV algorithms are possible (Supplemental Fig. 11).

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

Table 5. Gene, variant type, and pipeline call for clinically relevant genes

Sample	Gene	Variant type	Source	Assay	Calculated mean length (kb)	Region phased?	Called by at least one method?	Result (LR)
14_000779_A	*MSH2*	Single exon duplication	Archival DNA	SureSelectV6, 7.25 Gb (60x)	64	No	No	No call
14_000779_A	*MSH2*	Single exon duplication	Archival DNA	SureSelectV6, 12 Gb (100x)	53	No	No	No call
15_000612_A	*PMS2*	Single exon duplication	Archival DNA	SureSelectV6, 7.25 Gb (60x)	65	Yes	Yes	No call
15_000612_A	*PMS2*	Single exon duplication	Archival DNA	SureSelectV6, 12 Gb (100x)	59	Yes	Yes	No call
B1633	*BRCA1*	Single exon duplication	Cell line	SureSelectV6, 7.25 Gb (60x)	96	No	No	No call
B1633	*BRCA1*	Single exon duplication	Cell line	SureSelectV6, 12 Gb (100x)	78	No	No	No call
B1633	*BRCA1*	Single exon duplication	Cell line	Whole Genome, 128 Gb (30x)	88	No	No	No call
B1633	*BRCA1*	Single exon duplication	Archival DNA	SureSelectV6, 7.25 Gb (60x)	28	No	No	No call
B1633	*BRCA1*	Single exon duplication	Archival DNA	SureSelectV6, 12 Gb (100x)	27	No	No	No call
L129364, B17012	*BRCA2*	Single exon duplication	Archival DNA	SureSelectV6, 7.25 Gb (60x)	24	No	No	No call
L129364, B17012	*BRCA2*	Single exon duplication	Archival DNA	SureSelectV6, 12 Gb (100x)	19	Yes	Yes	No call
B1668	*BRCA1*	Two exon deletion	Cell line	SureSelectV6, 7.25 Gb (60x)	106	No	No	No call
B1668	*BRCA1*	Two exon deletion	Cell line	SureSelectV6, 12 Gb (100x)	98	No	No	No call
B1668	*BRCA1*	Two exon deletion	Archival DNA	SureSelectV6, 7.25 Gb (60x)	71	No	No	No call
B1668	*BRCA1*	Two exon deletion	Archival DNA	SureSelectV6, 12 Gb (100x)	80	No	No	No call
B1731	*BRCA1*	Two exon deletion	Cell line	SureSelectV6, 7.25 Gb (60x)	97	Yes	Yes	Called
B1731	*BRCA1*	Two exon deletion	Cell line	SureSelectV6, 12 Gb (100x)	107	Yes	Yes	Called
B1731	*BRCA1*	Two exon deletion	Archival DNA	SureSelectV6, 7.25 Gb (60x)	15	No	No	No call
B1731	*BRCA1*	Two exon deletion	Archival DNA	SureSelectV6, 12 Gb (100x)	12	Yes	Yes	Called
D152523, B22632	*PMS2*	Two exon deletion	Archival DNA	SureSelectV6, 7.25 Gb (60x)	57	Yes	Yes	Called
D152523, B22632	*PMS2*	Two exon deletion	Archival DNA	SureSelectV6, 12 Gb (100x)	48	Yes	Yes	Called
FH103364, 365168	*PMS2*	2-3 exon deletion	Archival DNA	SureSelectV6, 7.25 Gb (60x)	54	Yes	Yes	Called
FH103364, 365168	*PMS2*	2-3 exon deletion	Archival DNA	SureSelectV6, 12 Gb (100x)	42	Yes	Yes	Called
FH106388, 505476	*PMS2*	Large structural variant	Archival DNA	SureSelectV6, 7.25 Gb (60x)	43	Yes	No	No call
FH106388, 505476	*PMS2*	Large structural variant	Archival DNA	SureSelectV6, 12 Gb (100x)	35	Yes	No	No call
FH106388, 505476	*PMS2*	Large structural variant	Archival DNA	Whole genome, 128 Gb (30x)	28	Yes	No	No call
FH106388, 505476	*MSH2*	Two exon deletion	Archival DNA	SureSelectV6, 7.25 Gb (60x)	43	No	No	No call
FH106388, 505476	*MSH2*	Two exon deletion	Archival DNA	SureSelectV6, 12 Gb (100x)	35	No	No	No call
FH106388, 505476	*MSH2*	Two exon deletion	Archival DNA	Whole genome, 128 Gb (30x)	28	Yes	Yes	Called

Ability of Linked-Reads to call variation in samples with known exon-level deletions and duplications. Exome or whole-genome sequencing was used on samples freshly extracted from cell lines or on archival DNA samples. The ability of the barcode-aware algorithms to call exon-level events is completely dependent on phasing. Longer DNA length and increased sequencing coverage sometimes improve variant calling, but this appears to be rescued by enabling phasing.

regions of the genome, but struggles to reliably call SVs, cannot assess variation across the entire genome, and fails to reconstruct long range haplotypes (Sudmant et al. 2015). Recent studies have highlighted the importance of including haplotype information and more complete SV identification in genome studies (Chaisson et al. 2017). We have described an improved implementation of Linked-Reads coupled with novel algorithms in Long Ranger, that allows reconstruction of multi-megabase phase blocks, identification of large balanced and unbalanced SVs, and identification of small variants, even in regions of the genome typically recalcitrant to short-read approaches.

Some limitations to this approach currently exist. We observed a loss of coverage in regions of the genome with extreme GC content, and reduced performance in small indel calling, although this largely occurs in homopolymer regions and LCRs. Recent work suggests ambiguity in such regions may be tolerated for a large number of applications (Li et al. 2018). Although Linked-Reads can resolve many repetitive elements and genome regions, highly repetitive sequences that are larger than the length of input DNA are not resolvable by Linked-Reads. This limitation is common to all technologies currently available, including long-read sequencing. Repeat copies that reside on the same molecule will be subject to the same limitations as standard short-read approaches. It is 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 (Elyanow et al. 2017; Spies et al. 2017; Karaoglanoglu et al. 2018; Xia et al. 2018). An additional limitation in this study is the reliance on a reference sample for calling variants, which creates reference bias and the inability to call variants in regions that are not resolved in the reference, as was the case with the SV in the pericentric region on Chromosome 16. To bypass any reference bias, Linked-Read data can also be used to perform diploid de novo assembly in combination with an assembly program, Supernova (Weisenfeld et al. 2017).

Despite these limitations, Linked-Read sequencing provides a clear advantage over short reads alone allowing 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, to our knowledge, 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; Nordlund et al. 2018; Viswanathan et al. 2018) and demonstrates that Linked-Reads outperforms the switch accuracy and phasing completeness of other haplotyping methods (Chaisson et al. 2017). 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) 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).

Samples with known large SVs were obtained as cell lines from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research (repository ID numbers listed in Supplemental Table 8). Frozen cell pellets were thawed rapidly at 37°C in 1 mL PBS. High molecular weight DNA was then extracted following recommended protocols, as above.

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

Clinical samples from individuals with known exon-level del/dups were collected at The Institute of Cancer Research, London and shipped to 10x Genomics as cell lines or archival DNA. Samples were recruited through the Breast and Ovarian Cancer Susceptibility (BOCS) study and the Royal Marsden Hospital Cancer Series (RMHCS) study. 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 the manufacturer's recommendations (https://support.10xgenomics.com/genome-exome/sample-prep/doc/demonstrated-protocol-hmw-dna-qc).

Chromium Linked-Read library preparation

A Chromium controller chip was loaded with 1.25 ng of high molecular weight DNA, 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). Target enrichment for the Linked-Read whole-exome libraries was performed using Agilent SureSelect V6 exome baits following recommended protocols (https://assets.contentful.com/an68im79xiti/Zm2u8VlFa8qGYW4SGKG6e/4bddcc3cd60201388f7b82d241547086/CG000059_DemonstratedProtocolExome_RevC.pdf). Supplemental Figure 12 describes targeted sequencing with Linked-Reads.

GemCode Linked-Read library preparation

For the GemCode comparator analyses, Linked-Read libraries were prepared for 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

Following recommended protocols (Supplemental Methods), 350–800 ng of genomic DNA was sheared to a size of ~385 bp. Target enrichment for the Linked-Read whole-exome libraries was performed using Agilent SureSelect 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 (2×150 bp); Chromium whole-exome libraries (2×100 bp or 114 bp, 98 bp), and GemCode libraries (2×98 bp). 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 as a small number of additional sources of wasted data and gives an average, deduplicated coverage of approximately 30×. To demonstrate the extra sequence volume is not the driver of the improved alignment coverage, we performed gene finishing comparisons at matched volume (100 Gb lrWGS and 100 Gb TruSeq PCR-) and continue to see coverage gains (Supplemental Fig. 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 30-fold whole-genome coverage. The read set is aligned to GRCh37 (hg19-2.2.0 reference using BWA-MEM version 0.7.12), reads are sorted, duplicates are marked, and the BAM is indexed using Picard tools (version 2.9.2; https://broadinstitute.github.io/picard/). Indel realignment and BAM recalibration (base quality score recalibration) is performed using known indels from Mills Gold Standard and The 1000 Genomes Project and variants from dbSNP (version 138). Indels and SNVs are called from the BAM using HaplotypeCaller and are genotyped to produce a single VCF file. For NA12878, this VCF file can be 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. Sensitivity and specificity for both SNVs and indels is calculated to compare the Long Ranger short variant caller with the GATK Best Practices pipeline. All Long Ranger runs were performed with a prerelease build of Long Ranger version 2.2 utilizing GATK as a base variant caller. Long Ranger 2.2 has since been released.

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 (Supplemental Methods).

We selected regions of two- to sixfold 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 took the union of the GIAB confident regions BED file with these regions to determine the GIAB++ confident regions BED. The amount of sequence added to the GIAB++ BED differs by sample, as the original GIAB confident regions are sample-specific.

Structural variant comparison against deletion ground truth

After segmenting the Long Ranger deletion calls by size, we overlapped them to the svclassify set (Parikh et al. 2016) using the bedr package and BEDTools v2.27.1 (Quinlan and Hall 2010). We retained events >30 kb showing at least 50% reciprocal overlap. We also searched for Mendelian inheritance consistency on the parental samples NA12891 and NA12892. We annotated eight overlapping events with almost perfect breakpoint and Mendelian inheritance agreement. All genotypes were phased. In the svclassify overlapping deletions, all of the breakpoints except for the 3′ most in Chr 5: 104,432,114–104,503,672 had a read's length distance from each other. We then curated the remaining nine events called by Long Ranger that were not in the svclassify set. Of notice is one event (Chr 1: 189,704,517–189,783,347) contained within a larger deletion (Chr 1: 189,690,000–189,790,000). Among the nonoverlapping deletions, were six large SVs with breakpoint and Mendelian consistency in the parents. The other three (Chr 1: 189,690,000–189,790,000; Chr 11: 55,360,000–55,490,000; Chr 2: 242,900,000–243,080,000) had different breakpoints, were unphased but had consistent genotypes, or had no support in the parental data.

We took the Long Ranger deletion calls between 50 bp and 30 kb generated by both Long Ranger algorithms and GATK and merged them using SURVIVOR (Jeffares et al. 2017) allowing variants up to 50 bp apart to be merged. SURVIVOR was used again with a 50-bp merge distance to merge the Long Ranger deletion call set with deletions in the svclassify set. The resulting merged VCFs were then parsed to determine overlap and support for Long Ranger calls.

Data access

All reference sample read data generated in this study have been submitted to the NCBI BioProject (BioProject; https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA428496. Genomic short variation and structural variant study data for samples from individuals with inherited disease generated in this study have been submitted to the European Variation Archive (EVA; https://www.ebi.ac.uk/eva/) under accession number PRJEB28297. Genomic short variation and structural variant data for samples from the CNV Panel generated in this study have been submitted to the NCBI database of Genotypes and Phenotypes (dbGaP; https://www.ncbi.nlm.nih.gov/gap) under accession number phs001773.v1.p1. All code used in this paper is available online at GitHub. The lariat aligner code can be found at https://github.com/10XGenomics/lariat; the Long Ranger code is available at https://github.com/10XGenomics/longranger and as Supplemental Code; and specific analysis codes used in this paper can be accessed at https://github.com/10XGenomics/chromium-genome-paper.

Competing interest statement

Patrick Marks, Sarah Garcia, Alvaro Martinez Barrio, Kamila Belhocine, Jorge Bernate, Rajiv Bharadwaj, Keith Bjornson, Claudia Catalanotti, Josh Delaney, Adrian Fehr, Ian Fiddes, Brendan Galvin, Haynes Heaton, Jill Herschleb, Christopher Hindson, Cassandra Jabara, Susanna Jett, Nikka Keivanfar, Sofia Kyriazopoulou-Panagiotopoulou, Bill Lin, Adam Lowe, Shamoni Maheshwari, Tony Makarewicz, Francesca Meschi, Christopher O'Keefe, Heather Ordonez, Pranav Patel, Andrew Price, Ariel Royall, Michael Schnall-Levin, Preyas Shah, David Stafford, Stephen Williams, Indira Wu, Andrew Wei Xu, and Deanna Church are current or former employees and stock holders of 10x Genomics.

Acknowledgments

We thank the individuals who donated specimens for research (European Genome-phenome Archive [EGA] accession number EGAD00001004319). This manuscript would not have been possible without their contributions. We thank Stephane C. Boutet and Sarah Taylor for reviewing the manuscript, Kariena Dill for help with manuscript preparation, Kevin Wu for assistance setting up markdown and Docker, and Jamie Schwendinger-Schreck for

project management as well as invaluable contributions in manuscript preparation.

References

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526:** 68–74. doi:10.1038/nature15393

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12:** 363–376. doi:10.1038/nrg2958

Amini S, Pushkarev D, Christiansen L, Kostem E, Royce T, Turk C, Pignatelli N, Adey A, Kitzman JO, Vijayan K, et al. 2014. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat Genet* **46:** 1343–1349. doi:10.1038/ng.3119

Bionano Genomics. 2017. Bionano genome mapping identifies large structural variants in cancer and genetic disorders. https://bionanogenomics.com/wp-content/uploads/2017/02/Bionano_Human-Structural-Variations-White-Paper.pdf.

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. doi:10.1101/gr.191189.115

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. doi:10.1186/1471-2164-13-375

Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al. 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517:** 608–611. doi:10.1038/nature13907

Chaisson MJ, 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 doi:10.1101/193144

Choi Y, Chan AP, Kirkness E, Telenti A, Schork NJ. 2018. Comparison of phasing strategies for whole human genomes. *PLoS Genet* **14:** e1007308. doi:10.1371/journal.pgen.1007308

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. doi:10.1186/s13059-017-1158-6

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. doi:10.1371/journal.pone.0030377

Eberle MA, Fritzilas E, Krusche P, Källberg M, Moore BL, Bekritsky MA, Iqbal Z, Chuang HY, Humphray SJ, Halpern AL, et al. 2017. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res* **27:** 157–164. doi:10.1101/gr.210500.116

Elyanow R, Wu HT, Raphael BJ. 2017. Identifying structural variants using linked-read sequencing data. *Bioinformatics* **34:** 353–360. doi:10.1093/bioinformatics/btx712

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. doi:10.1186/s13073-017-0447-8

Huddleston J, Eichler EE. 2016. An incomplete understanding of human genetic variation. *Genetics* **202:** 1251–1254. doi:10.1534/genetics.115.180539

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* **27:** 677–685. doi:10.1101/gr.214007.116

Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun* **8:** 14061. doi:10.1038/ncomms14061

Karaoglanoglu F, Ricketts C, Rasekh ME, Ebren E, Hajirasouliha I, Alkan C. 2018. Characterization of segmental duplications and large inversions using Linked-Reads. bioRxiv doi:10.1101/394528

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15:** R84. doi:10.1186/gb-2014-15-6-r84

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. doi:10.1038/nature19057

Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, MacArthur D. 2018. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods* **15:** 595. doi:10.1038/s41592-018-0054-7

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. doi:10.1038/gim.2016.58

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. doi:10.1007/s13577-017-0168-8

Nordlund J, Marincevic-Zuniga Y, Cavelier L, Raine A, Martin T, Lundmark A, Abrahamsson J, Noren-Nystrom U, Lonnerholm G, Syvanen AC. 2018. Refined detection and phasing of structural aberrations in pediatric acute lymphoblastic leukemia by linked-read whole genome sequencing. bioRxiv doi:10.1101/375659

Parikh H, Mohiyuddin M, Lam HY, 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* **17:** 64. doi:10.1186/s12864-016-2366-2

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. doi:10.1371/journal.pgen.1003709

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

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. *Genome Res* **27:** 1950–1960. doi:10.1101/gr.222083.117

Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, 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. doi:10.1101/gr.213611.116

Spies N, Weng Z, Bishara A, McDaniel J, Catoe D, Zook JM, Salit M, West RB, Batzoglou S, Sidow A. 2017. Genome-wide reconstruction of complex structural variants using read clouds. *Nat Methods* **14:** 915. doi:10.1038/nmeth.4366

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526:** 75–81. doi:10.1038/nature15394

Viswanathan SR, Ha G, Hoff AM, Wala JA, Carrot-Zhang J, Whelan CW, Haradhvala NJ, Freeman SS, Reed SC, Rhoades J, et al. 2018. Structural alterations driving castration-resistant prostate cancer revealed by linked-read genome sequencing. *Cell* **174:** 433–447.e19. doi:10.1016/j.cell.2018.05.036

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res* **27:** 757–767. doi:10.1101/gr.214874.116

Xia LC, Bell JM, Wood-Bouwens C, Chen JJ, Zhang NR, Ji HP. 2018. Identification of large rearrangements in cancer genomes with barcode linked reads. *Nucleic Acids Res* **46:** e19. doi:10.1093/nar/gkx1193

Zheng GX, 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* **34:** 303–311. doi:10.1038/nbt.3432

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. doi:10.1038/nbt.2835

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. doi:10.1038/sdata.2016.25

Received January 9, 2018; accepted in revised form February 21, 2019.