

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

22 Abstract

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

43 **Introduction**

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

60 The prevalence of SVs suggests that individual haplotype reconstruction, rather than haploid
61 consensus analysis is a better approach to the analysis of an individual genome (Church et al. 2011,
62 2015; Schneider et al. 2017). Indeed, recent work from groups developing population graph-based
63 assembly representations have demonstrated that this approach improves alignment and
64 individual genome reconstruction (Iqbal et al. 2012; Novak et al. 2017). While it has long been
65 recognized that SVs play an important role in highly penetrant Mendelian disorders (Amberger et
66 al. 2015), groups investigating the biological impact of these events have demonstrated that SVs
67 have a more substantial impact on gene expression than do single nucleotide variants (SNVs), and
68 thus may contribute substantially to the development of common disorders (Chiang et al. 2017).

69 Recent work has shown that adding long range information and resolving long range haplotypes
70 improves sensitivity for SV detection (Huddleston and Eichler 2016; Chaisson et al. 2017).

71 Additionally, reconstructing individual haplotypes has the potential to improve analysis that relies
72 on patterns of genetic variation to extract genotype-phenotype information, such as eQTL analysis
73 (Ramaker et al. 2017). A more complete reconstruction of individual genomes will impact research
74 in both rare and common disease (Chiang et al. 2017).

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

83 The limitations of short reads suggest the need for improved methods for genome analysis. Several
84 long molecule sequencing and mapping approaches have been developed to address these issues
85 (Carneiro et al. 2012; Nakano et al. 2017; Genomics 2017). While they provide powerful data for
86 better understanding genome structure, their high input requirements, error rates and costs make
87 them inaccessible to many applications, particularly those requiring thousands of samples. To
88 address this need, we developed a technology that retains long range information while
89 maintaining the power, accuracy, and scalability of short read sequencing. The core datatype,
90 Linked-Reads, is generated by performing haplotype limiting dilution of long DNA molecules into
91 >1 million barcoded partitions, synthesizing barcoded sequence libraries within those partitions,
92 and then performing standard short read sequencing in bulk. The limited amount of DNA put into
93 the system, coupled with novel algorithms allow short reads to be associated with their long
94 molecule of origin, in most cases, with high probability.

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

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

115 **Results**

116 **Improvements in Linked-Read data**

117 One limitation of the original GemCode approach was the need to combine the Linked-Read data
118 with a standard short-read library for analysis. This was needed to help address coverage

119 imbalances seen in the GemCode library alone. To address this issue we modified the original
120 biochemistry, replacing it with an isothermal amplification approach. The updated biochemistry
121 now provides for more even genome coverage, approaching that of PCR free short-read preps
122 (Figure 1).

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

130 **Improved Genome and Exome Alignments**

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

141 When we look specifically at the ability of Lariat to improve read coverage over genes, we observe
142 a net gain in gene coverage when performing lrWGS compared to srWGS, and even more robust
143 improvement when performing lrWES compared to srWES (Supplemental Figure 3). When we

144 limit the search space to a known set of 570 genes with closely related paralogs that confound
145 short read alignment (NGS ‘dead zone’ genes (Mandelker et al. 2016)) we see a net gain in read
146 coverage in 423 genes using lrWGS and 376 using lrWES. Further limiting the list to the 71 genes
147 relevant to Mendelian disease, we see a net improvement in 51 of these genes using lrWGS and 41
148 genes using lrWES (Figure 3). Exome analysis was limited to multiple replicates of a single control
149 sample, NA12878.

150 **Small variant calling**

151 Next, we assessed the performance of Linked-Reads for small variant calling (<50 bp). Small
152 variant calling, particularly for single nucleotide variants (SNVs) outside of repetitive regions, is
153 well powered by short reads because a high quality read alignment to the reference assembly is
154 possible and the variant resides completely within the read. We used control samples, NA12878
155 and NA24385 as test cases. We produced two small variant call sets for each sample, one generated
156 by running paired-end 10x Linked-Read Chromium libraries through the Long Ranger (10xLR)
157 pipeline and one produced by analyzing paired-end reads from a PCR-free TruSeq library using
158 GATK pipeline (PCR-) following best practices recommendations:

159 <https://software.broadinstitute.org/gatk/best-practices/>. We made a total of 4,549,657 PASS variant
160 calls from the NA12878 10xLR set, and 4,725,295 from the PCR- set, with 4,325,515 calls in common
161 to both sets (Table 1). Numbers for both samples are in Table 1.

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

169 (>99.8%) for SNVs (Table 1). We observed slightly diminished indel sensitivity (>89%) and
170 specificity (>94.5%), driven largely by regions with extreme GC content and low complexity
171 sequences (LCRs). Recent work suggests indel calling is still a challenging problem for many
172 approaches, but that only 0.5% of LCRs overlap regions of the genome thought to be functional
173 based on annotation or conservation (Li et al. 2017).

174 The GIAB high confidence data set is known to be quite conservative and we wished to explore
175 whether there was evidence for variants called in the 10xLR set not covered by the GIAB. We
176 utilized publicly available 40x coverage PacBio data sets available from the GIAB consortium (Zook
177 et al. 2016) to evaluate Linked-Read putative false positive variant calls. Manual inspection of 25
178 random locations suggested that roughly half of the hap.py identified Linked-Read false positive
179 calls were well supported by Linked-Read, short read, or PacBio evidence and were likely called
180 false positive due to deficiencies in the GIAB truth set (Supplemental Table 1). We then did a global
181 analysis of all 7,431 SNV and 16,713 indel putative false positive calls identified in NA12878 and
182 looked for the alternate alleles in aligned PacBio reads only. This analysis provided evidence that
183 2,253 SNV and 12,826 indels of the GIAB determined false positive calls were likely valid calls
184 (Supplemental Figure 4, Supplemental File 1). This prompted us to develop a new “extended truth
185 set” which included an additional PacBio validated 78,361 SNV and 21,026 indels (see Methods for
186 details on GIAB++ VCF). We also extended our analysis to include 69.72 Mb for NA12878 and 70.66
187 Mb for NA2438 of the genome in addition to the GIAB defined confident regions (see Methods for
188 details on GIAB++ BED). We reanalyzed the variant calls with the hap.py tool against the extended
189 truth set and augmented confident regions. Importantly, this allowed us to correctly identify an
190 additional 71,467 SNV and 12,663 indels. We anticipate that this is a conservative estimate since
191 our false positive calls are inflated due to little or no PacBio or short-read coverage in these regions.
192 Of the total putative false positive calls exclusive to the GIAB++ analysis, 79.86% (31,475) of SNVs
193 and 62.05% (2,790) of indels could not be validated because of little or no PacBio read coverage
194 (Supplemental Figure 4). These data show the 10xLR approach provides for the identification of
195 more small variants than can be identified by short read only approaches, driven by an increase in

196 the percentage of the genome for which 10xLR can obtain high quality alignments.

197 **Haplotype reconstruction and phasing**

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

210 Phase block construction using lrWES is additionally constrained by the bait set used to perform
211 the capture and the reduced variation seen in coding sequences. In order to analyze factors
212 impacting phase block construction, we assessed four samples with known compound
213 heterozygous variants in three genes known to cause Mendelian disease, *DYSF*, *POMT2*, and *TTN*.
214 The variants were separated by various distances, ranging from 33 Kb to over 188 Kb (Table 2).
215 Initial DNA extractions yielded long molecules ranging in size from 75 Kb - 112 Kb. We analyzed
216 these samples using the Agilent SureSelect V6 exome bait set, with downsampling of sequence
217 data to both 7.25 Gb (~60x coverage) and 12 Gb of sequence (~100x coverage). In all cases, the
218 variants were phased with respect to each other and determined to be in trans, as previously
219 determined by orthogonal assays. In two of the three cases, the entire gene was phased. The *DYSF*
220 gene was not completely phased in any sample because the difference between heterozygous SNPs

221 at the 3' end of the gene was substantially longer than the mean molecule length. This gene is in
222 the top 5% of genes intolerant to variation as determined by the RVIS metric, a measure of
223 evolutionary constraint, suggesting that reduced exonic heterozygosity over the gene would be a
224 common occurrence impairing complete phasing (Petrovski et al. 2013). However, in the context of
225 sequencing for the identification of recessive disease, causative heterozygous variants would be
226 expected to aid in the phasing of the disease-causing gene.

227 Many samples of interest have already been extracted using standard methods not optimized for
228 high molecular weight DNA and may not be available for a fresh re-extraction to obtain DNA
229 optimized for length. For this reason, we wanted to understand the impact of reduced molecule
230 length on our ability to phase the genes and variants in these samples. We took the original freshly
231 extracted long molecules and sheared them to various sizes, aiming to assess lengths ranging from
232 5Kb to the original full length samples (Table 2). These results illustrate the complex interplay
233 between molecule length distribution and the observed heterozygosity within a region. For
234 example, in sample B12-21, with variants in *TTN* that are 53 Kb apart, the variants could be phased,
235 even with the smallest molecule size. However in sample B12-122, with variants in *POMT2* only 33
236 Kb apart, variant phasing is lost at 20Kb DNA lengths. We assessed the maximum distance
237 between heterozygous sites observed in each gene. We then plotted the difference between this
238 distance and the inferred molecule length against the molecule length and assessed the impact on
239 causative SNP phasing (Figure 5). In general, when the maximum distance between heterozygous
240 SNPs is greater than the molecule length (positive values), the ability to phase these SNPs
241 decreases. There are exceptions to this as the longer molecules in the molecule size distribution
242 will sometimes allow tiling between the variants, therefore extending phase block size beyond
243 what would be expected based on the mean length alone.

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

248 approach.

249 **Structural variant detection**

250 Short reads struggle with accurate and specific SV detection. This is, in part, due to the limitations
251 of assessing long range information using short reads, which only provide information over short
252 distances. Another complicating factor is the many types of structural variants, each requiring the
253 detection of a different signal depending on the type and mechanism of the event (Alkan et al.
254 2011; Collins et al. 2017). There is increasing evidence that grouping reads by their source
255 haplotype improves SV sensitivity, but this is not commonly done in practice (Huddleston et al.
256 2016; Chaisson et al. 2017).

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

266 Comparing SV call sets produced by different methods is a challenging task. There is often
267 ambiguity around the exact coordinate of the breakpoint(s), in part because repetitive sequence
268 content frequently flanks structural variation (Wittler et al. 2015). An additional challenge is
269 variability in the inclusion of the many different possible variant types. The validated call set
270 published with svclassify (Parikh et al. 2016) contains deletions and insertions, but no balanced
271 events. By contrast, the Long Ranger pipeline output contains deletions, duplications and balanced
272 events, but Long Ranger does not currently call insertions (Supplemental Table 2). Long Ranger

273 identifies event types by matching to simple models of deletions, duplications and inversions.
274 Therefore, there are additional events where Long Ranger identifies clear evidence for anomalous
275 barcode overlap, but is unable to match the event to one of the pre-defined models. These
276 undefined events are rendered as unknown.

277 For these reasons, we limited our comparisons to the set of deletion calls only. We partition the
278 ground truth set into SVs <30 Kb and SVs >30 Kb as different algorithms are used to call these
279 events. We first consider variants >30 Kb. There are 11 of these in the svclassify set and 23 in the
280 'Call set', with 8-9 being common to both (Table 3). Long Ranger calls two highly overlapping
281 events that map to the same svclassify event- thus 9 Long Ranger calls map to 8 svclassify events.
282 Of the three svclassify calls not called by Long Ranger, one (chr12:8,558,486-8,590,846) is
283 well-supported in the Linked-Read data by barcode overlap. For this event, Long Ranger calls a
284 10kb small deletion with a consistent 5' breakpoint to the svclassify event, but prematurely closes
285 the event, missing 22kb of the deletion. A second event (chr22:24,274,143-24,311,297) is also
286 well-supported but is filtered out from the 'Call set' as it overlaps with a segmental duplication.
287 There is no support for the last missing call (chr14:37,631,608-37,771,227). Further investigation of
288 this call reveals that it is genotyped as homozygous reference in NA12878 in the svclassify truth
289 set. We then performed manual review of the 14 events called by Long Ranger that are not in the
290 svclassify set. These 14 potential FP calls can be collapsed into 10 unique deletion calls. By manual
291 review all 10 calls have significant barcode overlap and coverage support including three events
292 that are known copy number variant loci per the Genome Reference Consortium (Supplemental
293 Figure 7). Thus, the Long Ranger large SV deletion calls show both high sensitivity and specificity.

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

300 improvements will produce further gains in sensitivity and specificity for this class of variants.

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

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

317 The final InvFest event, chr15:28,157,404-30,687,000 (Supplemental Figure 8G), shows minimal
318 anomalous barcode overlap in the Chromium data in a pattern not consistent with an inversion.
319 There are no candidate calls connecting the two loci. This region partially overlaps the known
320 gamma inversion haplotype of the *GOLGA8* locus
321 ([https://www.ncbi.nlm.nih.gov/grc/human/issues?q=chr15:28157404-](https://www.ncbi.nlm.nih.gov/grc/human/issues?q=chr15:28157404-30687000&asm=GRCh37.p13)
322 [30687000&asm=GRCh37.p13](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,
323 and this region was adjusted in GRCh37, it is possible that this call would disappear from InvFest if
324 analyses were redone using GRCh37.

325 Linked-Reads provide a clear advantage for SV detection over standard short read approaches. The

326 data type is still relatively new and the algorithmic approaches to SV identification are still
327 relatively immature. Several groups have already described methods utilizing Linked-Reads for SV
328 detection, largely in tumor samples (Spies et al. 2016; Elyanow et al. 2017; Xia et al. 2017). The
329 power of Linked-Reads to identify balanced events is a notable improvement and there is evidence
330 for this class of event being more prevalent in the population than originally realized (Collins et al.
331 2017; Chaisson et al. 2017). The evaluation of Linked-Reads and Long Ranger to identify complex,
332 constitutional events provides additional evidence of the power of this datatype for complex event
333 detection (Garcia et al. 2017).

334 **Analysis of samples from individuals with inherited disease**

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

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

351 successfully called in samples that could not be phased over the region of interest.

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

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

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

376 Discussion

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

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

399 Despite these limitations, Linked-Read sequencing provides a clear advantage over short reads
400 alone. This pipeline allows for the construction of long range haplotypes as well as the
401 identification of short variants and SVs from a single library and analysis pipeline. No other

402 approach that scales to thousands of genomes provides this level of detail for genome analysis.
403 Other recent studies have demonstrated the power of Linked-Reads to resolve complex variants in
404 both germline and cancer samples (Collins et al. 2017; Greer et al. 2017; Garcia et al. 2017). Recent
405 work demonstrates that Linked-Reads outperforms the switch accuracy and phasing completeness
406 of other haplotyping methods, and provides multi-MB phase blocks (Chaisson et al. 2017). In
407 another report, Linked-Reads also enable the ability to perform diploid, *de novo* assembly in
408 combination with an assembly program, Supernova (Weisenfeld et al. 2017). The ability to provide
409 reference free analysis promises to increase our understanding of diverse populations. Finally, the
410 ability to represent and analyze genomes in terms of haplotypes, rather than compressed haploid
411 representations, represents a crucial shift in our approach to genomics, allowing for a more
412 complete and accurate reconstruction of individual genomes.

413 **Methods**

414 *Samples and DNA Isolation*

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

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

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

426 Research, London and shipped to 10x genomics as cell lines or archival DNA. This sample cohort
427 was previously accessed for predisposition to cancer. Samples were recruited through the Breast
428 and Ovarian Cancer Susceptibility (BOCS) study and the Royal Marsden Hospital Cancer Series
429 (RMHCS) study, which aimed to discover and characterize disease predisposition genes. All
430 patients gave informed consent for use of their DNA in genetic research. The studies have been
431 approved by the London Multicentre Research Ethics Committee (MREC/01/2/18) and Royal
432 Marsden Research Ethics Committee (CCR1552), respectively. Samples were also obtained through
433 clinical testing by the TGLclinical laboratory, an ISO 15189 accredited genetic testing laboratory.
434 The consent given from patients tested through TGLclinical includes the option of consenting to
435 the use of samples/data in research; all patients whose data was included in this study approved
436 this option. DNA was extracted from cell lines as described above and archival DNA samples were
437 checked for size and quality according to manufacturer's recommendations: [https://support.
438 10xgenomics.com/genome-exome/sample-prep/doc/demonstrated-protocol-hmw-dna-qc](https://support.10xgenomics.com/genome-exome/sample-prep/doc/demonstrated-protocol-hmw-dna-qc) .

439 *ChromiumTM Linked-Read Library Preparation*

440 1.25 ng of high molecular weight DNA was loaded onto a Chromium controller chip, along with
441 10x Chromium reagents (either v1.0 or v2.0) and gel beads following recommended protocols:
442 [https://assets.contentful.com/an68im79xiti/4z5JA3C67KOyCE2ucacCM6/
443 do5ce5fa3dc4282f3da5ae7296f2645b/CG00022_GenomeReagentKitUserGuide_RevC.pdf](https://assets.contentful.com/an68im79xiti/4z5JA3C67KOyCE2ucacCM6/do5ce5fa3dc4282f3da5ae7296f2645b/CG00022_GenomeReagentKitUserGuide_RevC.pdf). The initial
444 part of the library construction takes place within droplets containing beads with unique barcodes
445 (called GEMs). The library construction incorporates a unique barcode that is adjacent to read one.
446 All molecules within a GEM get tagged with the same barcode, but because of the limiting dilution
447 of the genome (roughly 300 haploid genome equivalents) the chances that two molecules from the
448 same region of the genome are partitioned in the same GEM is very small. Thus, the barcodes can
449 be used to statistically associate short reads with their source long molecule.

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

452 [https://assets.contentful.com/an68im79xiti/Zmzu8VIFa8qGYW4SGKG6e/](https://assets.contentful.com/an68im79xiti/Zmzu8VIFa8qGYW4SGKG6e/4bdddcc3cd60201388f7b82d241547086/CG000059_DemonstratedProtocolExome_RevC.pdf)
453 [4bdddcc3cd60201388f7b82d241547086/CG000059_DemonstratedProtocolExome_RevC.pdf](https://assets.contentful.com/an68im79xiti/Zmzu8VIFa8qGYW4SGKG6e/4bdddcc3cd60201388f7b82d241547086/CG000059_DemonstratedProtocolExome_RevC.pdf).

454 Supplemental Figure 11 describes targeted sequencing with Linked-Reads.

455 *GemCodeTM Linked-Read Library Preparation*

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

459 *TruSeq PCR-free Library Preparation*

460 350-800 ng of genomic DNA was sheared to a size of ~385 bp using a Covaris®M220 Focused
461 Ultrasonicator using the following shearing parameters: Duty factor = 20%, cycles per burst = 200,
462 time = 90 seconds, Peak power 50. Fragmented DNA was then cleaned up with 0.8x SPRI beads and
463 left bound to the beads. Then, using the KAPA Library Preparation Kit reagents (KAPA
464 Biosystems, Catalog # KK8223), DNA fragments bound to the SPRI beads were subjected to end
465 repair, A-base tailing and Illumina® ‘PCR-free’ TruSeq adapter ligation (1.5 μM final concentration
466 of adapter was used). Following adapter ligation, two consecutive SPRI cleanup steps (1.0X and
467 0.7X) were performed to remove adapter dimers and library fragments below ~150 bp in size. No
468 library PCR amplification enrichment was performed. Libraries were then eluted off the SPRI
469 beads in 25 ul elution buffer and quantified with quantitative PCR using KAPA Library Quant kit
470 (KAPA Biosystems, Catalog # KK4824) and an Agilent Bioanalyzer High Sensitivity Chip (Agilent
471 Technologies) following the manufacturer’s recommendations.

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

474 *Sequencing*

475 Libraries were sequenced on a combination of Illumina® instruments (HiSeq®2500, HiSeq 4000,
476 and HiSeq X). Paired-End sequencing read lengths were as follows: TruSeq and Chromium whole

477 genome libraries (2X150bp); Chromium whole exome libraries (2X100bp or 114bp, 98bp), and
478 Gemcode libraries (2X98bp). lrWGS libraries are typically sequenced to 128 Gb, compared to 100
479 Gb for standard TruSeq PCR free libraries. The additional sequence volume compensates for
480 sequencing the barcodes as well a small number of additional sources of wasted data and gives an
481 average, de-duplicated coverage of approximately 30x. To demonstrate the extra sequence volume
482 is not the driver of the improved alignment coverage, we performed a gene finishing comparison
483 at matched volume (100Gb lrWGS and 100Gb TruSeq PCR-) and continue to see coverage gains
484 (Supplemental Figure 12).

485 **Analysis**

486 *Comparison of 10X and GATK Best Practices*

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

502 *Development of extended truth set*

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

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

515 *Structural variant comparison against deletion ground truth*

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

521 **Acknowledgements**

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

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

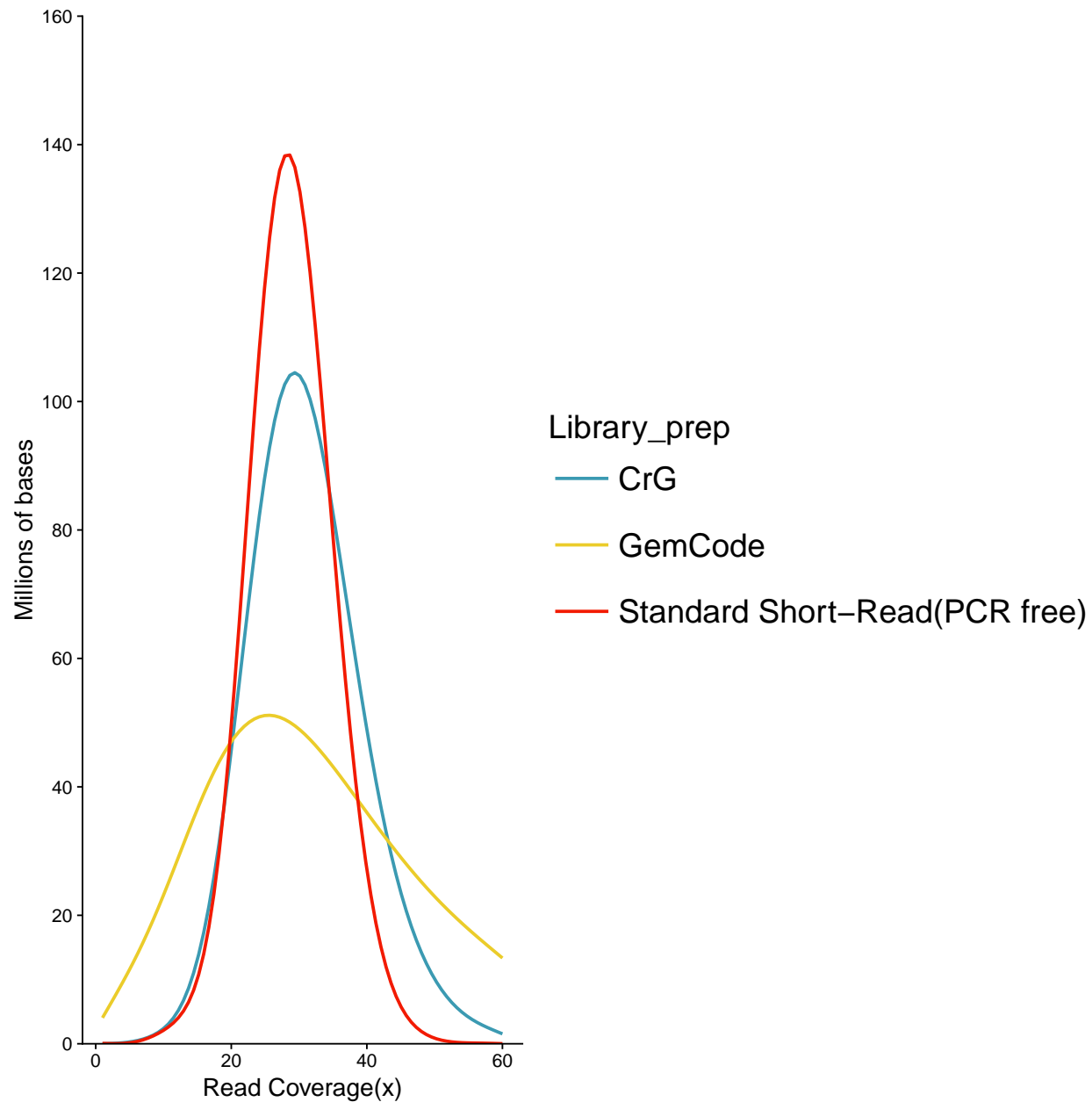


Figure 1: Coverage Evenness.

528 Distribution of read coverage for the entire human genome (GRCh37). Comparisons between 10x
529 Genomics Chromium Genome (CrG), 10x Genomics GemCode (GemCode), and Illumina TruSeq
530 PCR free standard short-read NGS library preparations (Standard Short Read (PCR Free)).
531 Sequencing was performed in an effort to match coverage (see methods). Note the shift of the CrG

532 curve to the left, showing the improved coverage of Chromium vs. GemCode. X-axis represents
533 the fold read coverage across the genome. Y-axis represents the total number of bases covered at
534 any given read depth.

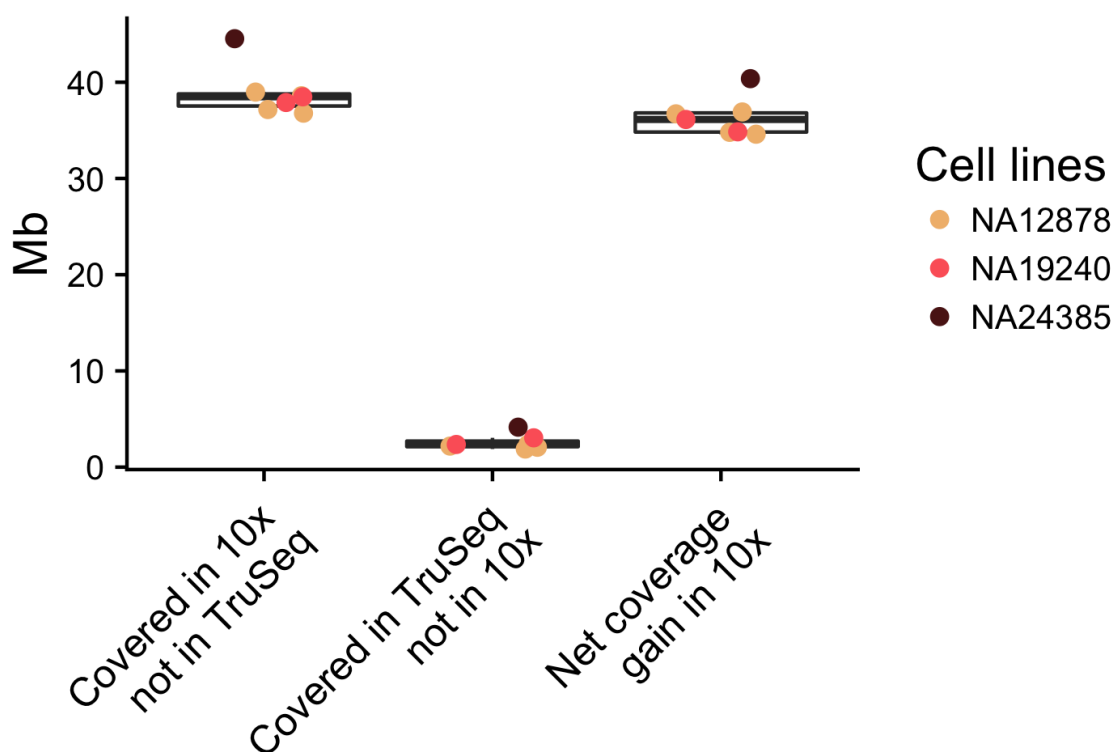


Figure 2: Comparison of unique genome coverage by assay.

535 The x-axis shows the number of sites with a coverage of ≥ 5 reads at $\text{MAPQ} \geq 30$. Column one
536 shows amount of the genome covered by 10x Chromium where PCR free TruSeq does not meet
537 that metric. Column 2 shows the amount of the genome covered by PCR free TruSeq where 10x
538 Chromium does not meet the metric. Column 3 shows the net gain of genome sequence with high
539 quality alignments when using 10x Chromium versus PCR free TruSeq. The comparison was
540 performed on samples with matched sequence coverage (see methods).

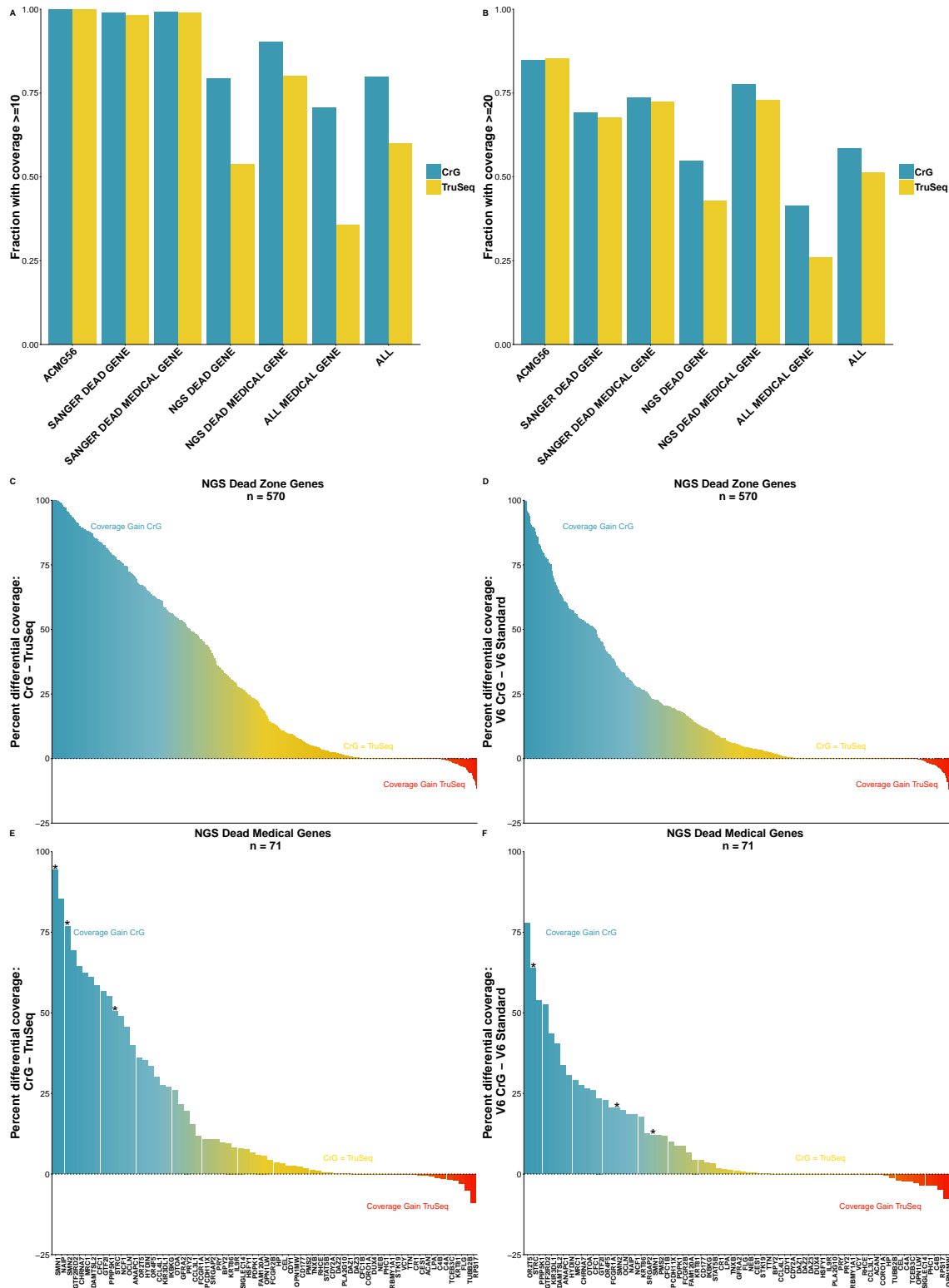


Figure 3: Gene finishing metrics.

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

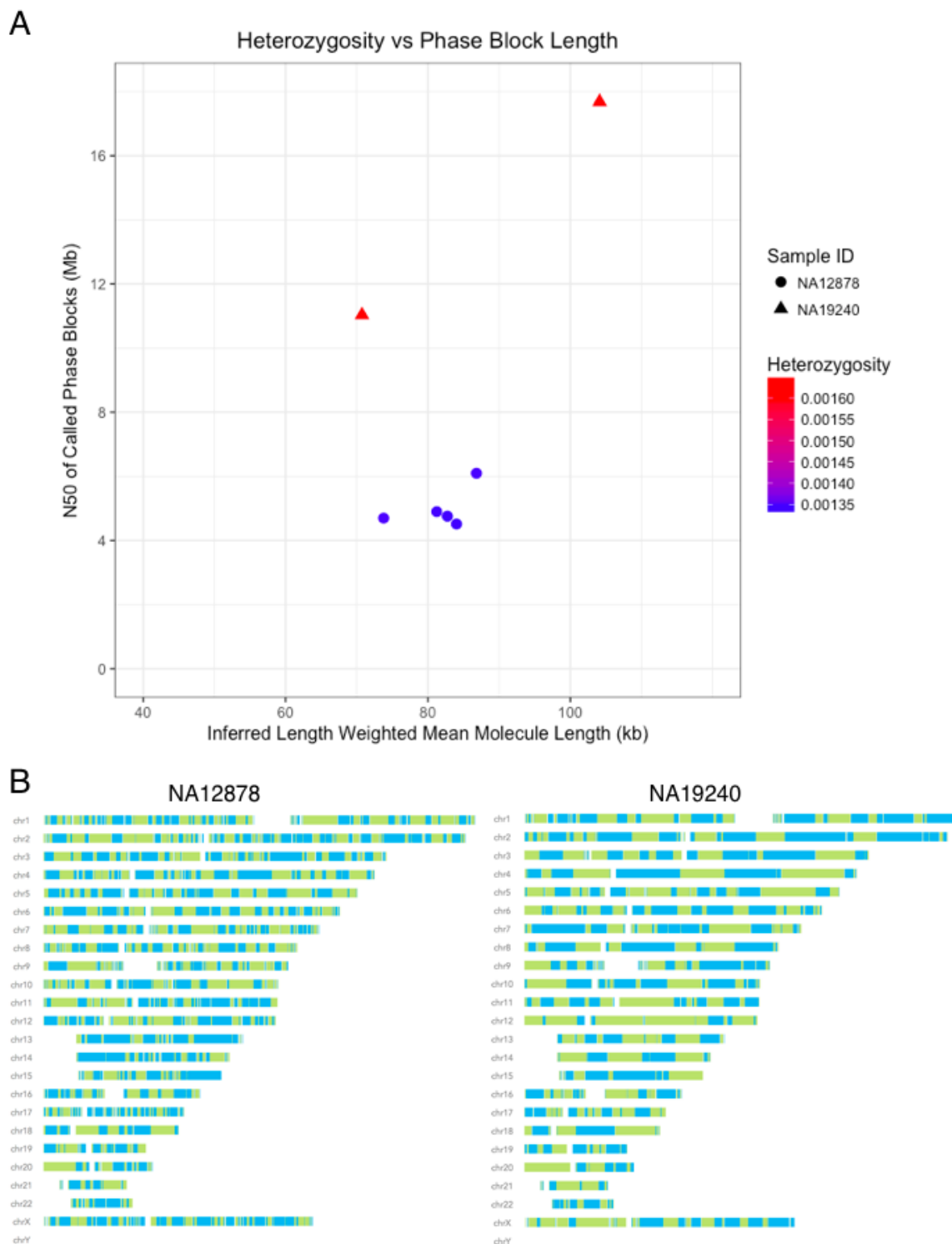


Figure 4: Haplotype reconstruction and phasing.

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

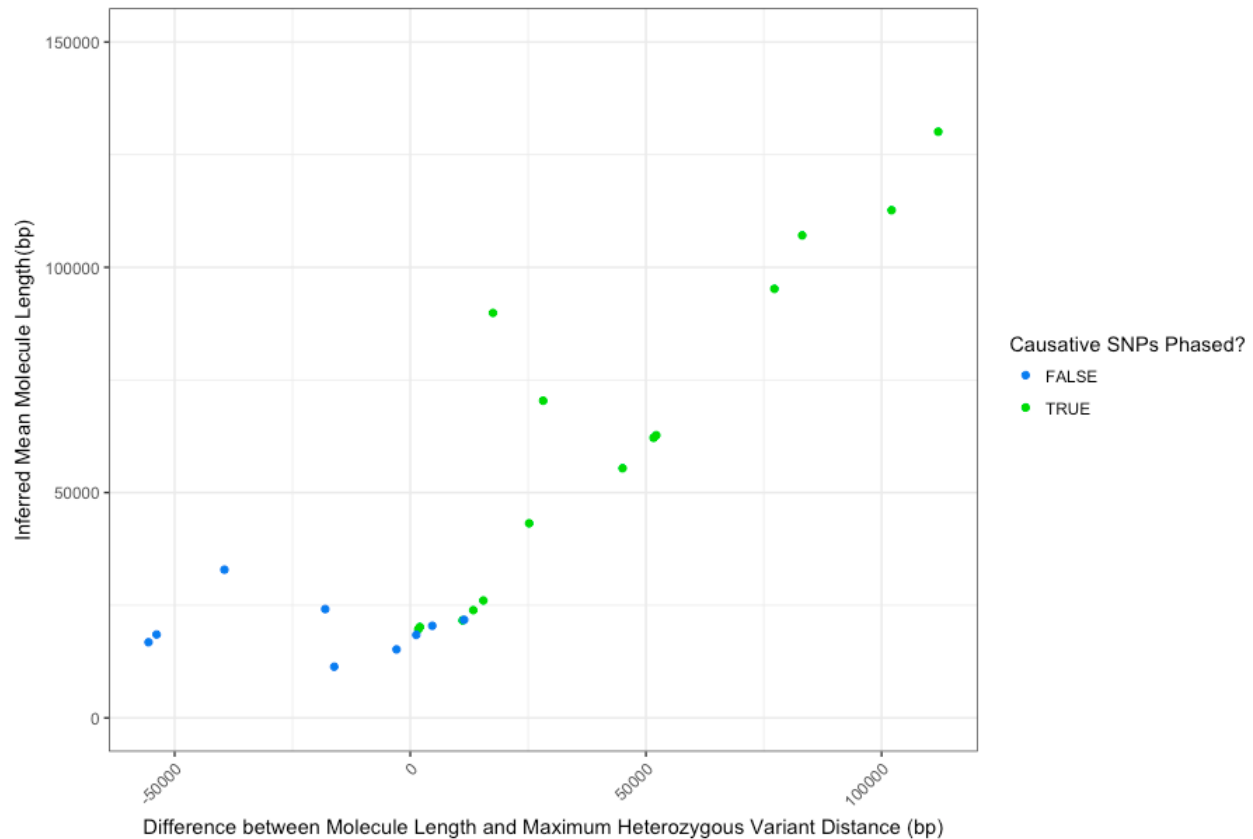


Figure 5: Validated examples of impact of molecule length on phasing (7.25Gb).

568 Blue dots represent samples for which the variants of interest are not phased, and green dots
569 represent samples for which there is phasing of the variants of interest. At longer molecule lengths
570 (>50kb), the molecule length was always longer than the maximum distance between heterozygous
571 SNPs in a gene, and phasing between the causative SNPs was always observed. As molecule length
572 shortens, it becomes more likely that the maximum distance between SNPs exceeds the molecule
573 length (reflected as a negative difference value) and phasing between the causative SNPs was never
574 observed in these cases. When maximum distance is similar to the molecule length causative SNPs
575 may or may not be phased. This is likely impacted by the molecule length distribution within the
576 sample.

577 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

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

583 GIAB BED confident regions using hap.py and data is shown per variant type for count of variants
584 in the truth set and in the confident regions (along with sensitivity and specificity). Data is also
585 shown for the same quantities when the variant calls were compared to the extended truth set
586 (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

587 Different intersections of Long-Ranger SV calls with a ground truth dataset published (Parikh et al.
588 2016). Comparison class identified in the most left column. A. Large deletions ($\geq 30\text{kb}$) intersected
589 against all deletions $\geq 30\text{kb}$ in the ground truth set. B. Smaller deletions ($< 30\text{kb}$), marked as PASS
590 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

- 591
- 592 Aleman F. 2017. The necessity of diploid genome sequencing to unravel the genetic component of
593 complex phenotypes. *Front Genet* **8**: 148.
- 594 Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev*
595 *Genet* **12**: 363–376.
- 596 Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online
597 mendelian inheritance in man (OMIM), an online catalog of human genes and genetic disorders.
598 *Nucleic Acids Res* **43**: D789–D798.
- 599 Askree SH, Chin ELH, Bean LH, Coffee B, Tanner A, Hegde M. 2013. Detection limit of intragenic
600 deletions with targeted array comparative genomic hybridization. *BMC Genet* **14**: 116.
- 601 Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly
602 P, Eichler EE, Flicek P, et al. 2015. A global reference for human genetic variation. *Nature* **526**:
603 68–74.
- 604 Bishara A, Liu Y, Weng Z, Kashef-Haghighi D, Newburger DE, West R, Sidow A, Batzoglu S. 2015.
605 Read clouds uncover variation in complex regions of the human genome. *Genome Res* **25**:
606 1570–1580.
- 607 Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. 2012. Pacific biosciences
608 sequencing technology for genotyping and variation discovery in human data. *BMC Genomics* **13**:
609 375.
- 610 Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti
611 U, Sandstrom R, Boitano M, et al. 2014. Resolving the complexity of the human genome using
612 single-molecule sequencing. *Nature*.
- 613 Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez O,
614 Guo L, Collins RL, et al. 2017. Multi-platform discovery of haplotype-resolved structural variation

615 in human genomes. *bioRxiv*.

616 Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, GTEX

617 Consortium, et al. 2017. The impact of structural variation on human gene expression. *Nat Genet*.

618 Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen H-C, Agarwala R,

619 McLaren WM, Ritchie GRS, et al. 2011. Modernizing reference genome assemblies. *PLoS Biol* **9**:

620 e1001091.

621 Church DM, Schneider VA, Steinberg KM, Schatz MC, Quinlan AR, Chin C-S, Kitts PA, Aken B,

622 Marth GT, Hoffman MM, et al. 2015. Extending reference assembly models. *Genome Biol* **16**: 13.

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

624 Collins RL, Brand H, Redin CE, Hanscom C, Antolik C, Stone MR, Glessner JT, Mason T, Pregno G,

625 Dorrani N, et al. 2017. Defining the diverse spectrum of inversions, complex structural variation,

626 and chromothripsis in the morbid human genome. *Genome Biol* **18**: 36.

627 Consortium IHGS. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**:

628 931-945.

629 Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, Ribeca P. 2012. Fast

630 computation and applications of genome mappability. *PLoS One* **7**: e30377.

631 Elyanow R, Wu H-T, Raphael BJ. 2017. Identifying structural variants using linked-read sequencing

632 data. *bioRxiv* 190454.

633 Garcia S, Williams S, Xu AW, Herschleb J, Marks P, Stafford D, Church DM. 2017. Linked-Read

634 sequencing resolves complex structural variants. *bioRxiv* 231662.

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

636 Greer SU, Nadauld LD, Lau BT, Chen J, Wood-Bouwens C, Ford JM, Kuo CJ, Ji HP. 2017. Linked

637 read sequencing resolves complex genomic rearrangements in gastric cancer metastases. *Genome*

- 638 *Med* **9**: 57.
- 639 Huddleston J, Chaisson MJ, Meltz Steinberg K, Warren W, Hoekzema K, Gordon DS,
640 Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, et al. 2016. Discovery and genotyping
641 of structural variation from long-read haploid genome sequence data. *Genome Res*.
- 642 Huddleston J, Eichler EE. 2016. An incomplete understanding of human genetic variation. *Genetics*
643 **202**: 1251–1254.
- 644 Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. 2012. De novo assembly and genotyping of
645 variants using colored de bruijn graphs. *Nat Genet* **44**: 226–232.
- 646 Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan
647 C, Antonacci F, et al. 2008. Mapping and sequencing of structural variation from eight human
648 genomes. *Nature* **453**: 56–64.
- 649 Krusche P. Hap.py.
- 650 Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O’Donnell-Luria AH, Ware JS,
651 Hill AJ, Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans.
652 *Nature* **536**: 285–291.
- 653 Li H. 2013. Aligning sequence reads , clone sequences and assembly contigs with BWA-MEM.
654 *ArXiv* **00**: 1–2.
- 655 Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier LD, Neale B, MacArthur D. 2017. New
656 synthetic-diploid benchmark for accurate variant calling evaluation. *bioRxiv* 223297.
- 657 Mandelker D, Amr SS, Pugh T, Gowrisankar S, Shakhbatyan R, Duffy E, Bowser M, Harrison B,
658 Lafferty K, Mahanta L, et al. 2014. Comprehensive diagnostic testing for stereocilin: An approach
659 for analyzing medically important genes with high homology. *J Mol Diagn* **16**: 639–647.
- 660 Mandelker D, Schmidt RJ, Ankala A, McDonald Gibson K, Bowser M, Sharma H, Duffy E, Hegde M,
661 Santani A, Lebo M, et al. 2016. Navigating highly homologous genes in a molecular diagnostic

- 662 setting: A resource for clinical next-generation sequencing. *Genet Med* **18**: 1282–1289.
- 663 Nakano K, Shiroma A, Shimoji M, Tamotsu H, Ashimine N, Ohki S, Shinzato M, Minami M,
664 Nakanishi T, Teruya K, et al. 2017. Advantages of genome sequencing by long-read sequencer
665 using SMRT technology in medical area. *Hum Cell* **30**: 149–161.
- 666 Novak AM, Hickey G, Garrison E, Blum S, Connelly A, Dilthey A, Eizenga J, Saleh Elmohamed MA,
667 Guthrie S, Kahles A, et al. 2017. Genome graphs. *bioRxiv* 101378.
- 668 Parikh H, Mohiyuddin M, Lam HYK, Iyer H, Chen D, Pratt M, Bartha G, Spies N, Losert W, Zook
669 JM, et al. 2016. Svclassify: A method to establish benchmark structural variant calls. *BMC*
670 *Genomics* 1–16.
- 671 Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. 2013. Genic intolerance to functional
672 variation and the interpretation of personal genomes. *PLoS Genet* **9**: e1003709.
- 673 Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features.
674 *Bioinformatics* **26**: 841–842.
- 675 Ramaker RC, Savic D, Hardigan AA, Newberry K, Cooper GM, Myers RM, Cooper SJ. 2017. A
676 genome-wide interactome of DNA-associated proteins in the human liver. *bioRxiv* 111385.
- 677 Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, Murphy TD, Pruitt KD,
678 Thibaud-Nissen F, Albracht D, et al. 2017. Evaluation of grch38 and de novo haploid genome
679 assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27**: 849–864.
- 680 Spies N, Weng Z, Bishara A, McDaniel J, Catoe D, Zook JM, Salit M, West RB, Batzoglou S, Sidow A.
681 2016. Genome-wide reconstruction of complex structural variants using read clouds. *bioRxiv*
682 074518.
- 683 Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G,
684 Hsi-Yang Fritz M, et al. 2015. An integrated map of structural variation in 2,504 human genomes.

685 *Nature* **526**: 75–81.

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

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

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

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

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

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