Single tube bead-based DNA co-barcoding for cost effective and accurate sequencing, haplotyping, and assembly

Ou Wang[1,2,3]†, Robert Chin[4]†, Xiaofang Cheng[1,2]†, Michelle Ka Wu[4]†, Qing Mao[4]†, Jingbo Tang[5]†, Yuhui Sun[1,2]†, Han K. Lam[4], Dan Chen[1,2], Yujun Zhou[1,2], Linying Wang[1,2], Fei Fan[1,2], Yan Zou[1,2], Ellis Anderson[4], Yinlong Xie[5], Rebecca Yu Zhang[4], Snezana Drmanac[4], Darlene Nguyen[4], Chongjun Xu[1,2,4], Christian Villarosa[4], Scott Gablenz[4], Nina Barua[4], Staci Nguyen[4], Wenlan Tian[4], Jia Sophie Liu[4], Jingwan Wang[1,2], Xiao Liu[1,2], Xiaojuan Qi[1,2], Ao Chen[1,2], He Wang[1,2], Yuliang Dong[1,2], Wenwei Zhang[1,2], Andrei Alexeev[4], Huanming Yang[1,6], Jian Wang[1,6], Karsten Kristiansen[1,2,3], Xun Xu[1,2]*, Radoje Drmanac[1,2,4]‡*, Brock A. Peters[1,2,4]‡*

[1]BGI-Shenzhen, Shenzhen 518083, China
[2]China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China
[3]Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, Copenhagen, Denmark
[4]Advanced Genomics Technology Lab, Complete Genomics Inc., 2904 Orchard Parkway, San Jose, California 95134, USA
[5]MGI, BGI-Shenzhen, Shenzhen 518083, China
[6]James D. Watson Institute of Genome Sciences, Hangzhou 310058, China

†These authors contributed equally to this work.
‡ These authors contributed equally to this work.
*Correspondence should be addressed to: B.A.P. (bpeters@completegenomics.com), R.D. (rdrmanac@completegenomics.com), and X.X. (xuxun@genomics.cn)

Single tube long fragment read (stLFR) technology enables efficient WGS, haplotyping, and contig scaffolding. It is based on adding the same barcode sequence to sub-fragments of the original DNA molecule (DNA co-barcoding). To achieve this, stLFR uses the surface of microbeads to create millions of miniaturized compartments in a single tube. Using a combinatorial process over 1.8 billion unique barcode sequences were generated on beads, enabling practically non-redundant co-barcoding in reactions with 50 million barcodes. Using stLFR we demonstrate efficient unique co-barcoding of over 8 million 20-300 kb genomic DNA fragments with near perfect variant calling and phasing of the genome of NA12878 into contigs up to N50 23.4 Mb. stLFR represents a low-cost single library solution that can enable long sequence data.

To date the vast majority of individual whole genome sequences lack information regarding the order of single to multi-base variants transmitted as contiguous blocks on homologous chromosomes. Numerous technologies[1-11] have recently been developed to enable this. Most are based on the process of co-barcoding[12], that is, the addition of the same barcode to the sub-fragments of single long genomic DNA molecules. After sequencing the barcode information can be used to determine which reads are derived from the original long DNA molecule. This process was first described by Drmanac[13] and implemented as a 384-well plate assay by Peters et al.[6]. However, these approaches are technically challenging to implement, expensive, have lower data quality, do not provide unique co-barcoding, or some combination of all four. In practice, most of these approaches require a separate whole genome sequence to be generated by standard methods to improve variant calling. This has resulted in the limited use of these methods as cost and ease of use are dominant factors in what technologies are used for WGS.

Results

stLFR library process

Here we describe implementation of stLFR technology[14], an efficient approach for DNA co-barcoding with millions of barcodes enabled in a single tube. This is achieved by using the surface of a microbead as a replacement for a compartment (e.g., the well of a 384-well plate). Each bead carries many copies of a unique barcode sequence which is transferred to the sub-fragments of each long DNA molecule. These co-barcoded sub-fragments are then analyzed on common short read sequencing devices such as the BGISEQ-500 or equivalent. In our implementation of this approach we use a ligation-based combinatorial barcode generation strategy to create over 1.8 billion different barcodes in three ligation steps. For a single sample we use ~10-50 million of these barcoded beads to capture ~10-100 million long DNA molecules in a single tube. It is infrequent that two beads will share the same barcode because we sample 10-50 million beads from such a large library of total barcodes. Furthermore, in the case of using 50 million beads and 10 million long genomic DNA fragments, the vast majority of sub-fragments from each long DNA fragment are co-barcoded by a unique barcode. This is analogous to long-read single molecule sequencing and potentially enables powerful informatics approaches for *de novo* assembly. A similar but informatically limited and less efficient approach using only ~150,000 barcodes was recently described by Zhang et al.[15]. Importantly, stLFR is simple to perform and can be implemented with a relatively small investment in oligonucleotides to generate barcoded beads. Further, stLFR uses standard equipment found in almost all molecular biology laboratories and can be analyzed by almost any sequencing strategy. Finally, stLFR replaces standard NGS library preparation methods, requires only 1 ng of DNA, and does not add significantly to the cost of whole genome or whole exome analyses with a total cost per sample of less than 30 dollars (Table 1).

The first step in stLFR is the insertion of a hybridization sequence at regular intervals along genomic DNA fragments. This is achieved through the incorporation of DNA sequences, by the Tn5 transposase, containing a single stranded region for hybridization and a double stranded sequence that is recognized by the enzyme and enables the transposition reaction (Figure 1a). Importantly, this step is done in solution, as opposed to having the insertion sequence linked directly to the bead[15]. This enables a very efficient incorporation of the hybridization sequence along the genomic DNA molecules. As previously observed[10], the transposase enzyme has the property of remaining bound to genomic DNA after the transposition event, effectively leaving the transposon-integrated long genomic DNA molecule intact. After the DNA has been treated with Tn5 it is diluted in hybridization buffer and added to 50 million ~2.8 um clonally barcoded beads in hybridization buffer. Each bead contains approximately 400,000 capture adapters, each containing the same barcode sequence. A portion of the capture adapter contains uracil nucleotides to enable destruction of unused adaptors in a later step. The mix is incubated under optimized temperature and buffer conditions during which time the transposon inserted DNA is captured to beads via the hybridization sequence. It has been suggested that genomic DNA in solution forms balls with both tails sticking out[16]. This may enable the capture of long DNA fragments towards one end of the molecule followed by a rolling motion that wraps the genomic DNA molecule around the bead. Approximately every 7.8 nm on the surface of each bead there is a capture oligo. This enables a very uniform and high rate of sub-fragment capture. A 100 kb genomic fragment would wrap around a 2.8 um bead approximately 3 times. In our data, 300 kb is the longest fragment size captured suggesting larger beads may be necessary to capture longer DNA molecules. Beads are next collected and individual barcode sequences are transferred to each sub-fragment through ligation of the nick between the hybridization sequence and the capture adapter (Figure 1a). At this point the DNA/transposase complexes are disrupted producing sub-fragments less than 1 kb in size. Due to the large number of beads and high density of capture oligos per bead, the amount of excess adapter is four orders of magnitude greater than the amount of product. This huge unused adapter can overwhelm the following steps. In order to avoid this, we designed beads with capture oligos connected by the 5' terminus. This enabled an exonuclease strategy to be developed that specifically degraded excess unused capture adapter.

In one approach to stLFR, two different transposons are used in the initial insertion step, allowing PCR to be performed after exonuclease treatment. However, this approach results in approximately 50% less coverage per long DNA molecule as it requires that two different transposons were inserted next to each other to generate a proper PCR product. To achieve the highest coverage per genomic DNA fragment we use a single transposon in the initial insertion step and add an additional adapter through ligation. This noncanonical ligation, termed 3' branch ligation, involves the covalent joining of the 5' phosphate from the blunt-end adapter to the recessed 3' hydroxyl of the genomic DNA (Figure 1a). A detailed explanation of this process has previously been described by some of us (*Wang et al.*, under review). Using this method, it is theoretically possible to amplify and sequence all sub-fragments of a captured genomic molecule. In addition, this ligation step enables a sample barcode to be placed adjacent to the genomic sequence for sampling multiplexing. This is useful as it does not require an additional sequencing primer to read this barcode. After this ligation step, PCR is performed and the library is ready to enter any standard next generation sequencing (NGS) workflow. In the case of BGISEQ-500, the library is circularized as previously described[17]. From single stranded circles DNA nanoballs are made and loaded onto patterned nanoarrays[17]. These nanoarrays are then subjected to combinatorial probe-anchor synthesis (cPAS) based sequencing on the BGISEQ-500[18-20]. After sequencing, barcode sequences are extracted using a custom program (Supplementary Materials). Mapping the read data by unique barcode shows that most reads with the same barcode are clustered in a region of the genome corresponding to the length of DNA used during library preparation (Figure 1b). A detailed description of this method, as well as a protocol for making the beads can be found in the supplementary materials.

stLFR read coverage and variant calling

To demonstrate stLFR phasing and variant calling we generated four libraries using 1 ng (stLFR-1 and stLFR-2) and 10 ngs (stLFR-3 and stLFR-4) of DNA from NA12878. The number of beads were varied with 10 million (stLFR-3), 30 million (stLFR-4), and 50 million (stLFR-1 and stLFR-2) used. Finally, both the 3' branch ligation (stLFR-1, stLFR-2, and stLFR-3) and two transposon (stLFR-4) methods were tested. Both stLFR-1 and stLFR-2 were sequenced deeply to 336 Gb and 660 Gb of total base coverage, respectively. We also analyzed these at downsampled coverages. stLFR-3 and stLFR-4 were sequenced to more modest levels of 117 Gb and 126 Gb, respectively. Co-barcoded reads were mapped to build 37 of the human reference genome using BWA-MEM[21]. Because stLFR does not require any preamplification steps, read coverage distribution across the genome was close to Poisson (Figure S1). The non-duplicate coverage ranged from 34-58X and the number of long DNA molecules per barcode ranged from 1.2-6.8 (Table 2 and Figure 1c). As expected, the stLFR libraries made from 50 million beads and 1 ng of genomic DNA had the highest single unique barcode co-barcoding rates of over 80% (Figure 1c). These libraries also observed the highest average non-overlapping read coverage per long DNA molecule of 10.7-12.1% and the highest average non-overlapping base coverage of captured sub-fragments per long DNA molecule of 17.9-18.4% (Figure 1d). This coverage is ~10 X higher than previously demonstrated using 3 ng of DNA and transposons attached to beads[15]. This suggests our solution-based transposition process is 3-fold more efficient at sub-fragment capture (40.7-47.4 sub-fragments per genomic fragment in 1 ng of genomic DNA versus 5 sub-fragments captured in 3 ng at similar read coverage as reported by Zhang et al.[15], Table 2).

For each library variants were called using GATK[22] using default settings. Comparing SNP and indel calls to Genome in a Bottle (GIAB)[23] allowed for the determination of false positive (FP) and false negative (FN) rates (Table 2). In addition, we performed variant calling using the same settings in GATK on a standard non-stLFR library made from ~1000 times more genomic DNA and also sequenced on a BGISEQ-500 (STD), and a Chromium library from 10X Genomics[11]. We also compared precision and sensitivity rates against those reported in the bead haplotyping library study by *Zhang et al.*[15]. Our stLFR approach and that practiced by Zhang et al. demonstrated lower SNP and Indel FP rates than the Chromium library. stLFR had 2-fold higher FP and FN rates than the STD library and depending on the particular stLFR library and filtering criteria the FN rate was either higher or lower than the Chromium library. The higher FN rate in stLFR libraries compared to standard libraries is primarily due to the shorter average insert size (~200 bp versus 300 bp in a standard library). That said, stLFR had a much lower FN rate than Zhang et al. for SNPs and Indels and a much lower FN rate than the Chromium library for Indels (Table 2). Overall, most metrics for variant calling were better for our stLFR libraries than the published results from Zhang et al. or Chromium libraries, especially when nonoptimized mapping and variant calling processes were used (Table 2, "No Filter").

One potential issue with using GIAB data to measure the FP rate is that we were unable to use the GIAB reference material (NIST RM 8398) due to the rather small fragment size of the isolated DNA. For this reason, we used the GM12878 cell line and isolated DNA using a dialysis-based method capable of yielding very high molecular weight DNA (see methods). However, it is possible that our isolate of the GM12878 cell line could have a number of unique somatic mutations compared to the GIAB reference material and thus cause the number of FPs to be inflated in our stLFR libraries. To examine this further we compared the overlap of single nucleotide FP variants between the 4 stLFR libraries and the two non-LFR libraries (Figure S2a). Overall, 544 FP variants were shared between the six libraries and 2,078 FPs were unique to the four stLFR libraries. We also compared stLFR FPs with the Chromium library and found that over half (1,194) of these shared FPs were also present in the Chromium library (Figure S2b). An examination of the read and barcode coverage of these shared variants showed they were more similar to that of TP variants (Figure S3-4). We also examined the distribution across the genome of these shared FP variants versus 2,078 randomly selected variants (Figure S5a). This analysis showed 219 variants that are found in clusters where two or more of these FPs are within 100 bp of each other. However, the majority (90%) of variants have distributions that appear indistinguishable from randomly selected variants. In addition, of those FPs shared between stLFR and Chromium libraries only 41 were found to be clustered (Figure S5a). Finally, 96 of these variants are called by GIAB but with a different zygosity than called in the stLFR libraries.

If we accept the evidence that these shared FP variants are largely real and not present in the GIAB reference material, the FP rate for stLFR could be up to 1,859 variants less than what is reported in Table 2 for SNP detection. This is still several thousand single nucleotide variants more than the standard BGISEQ-500 library. To further improve the FP rate in stLFR libraries we tested a number of different filtering strategies for removing errors. Ultimately, by applying a few filtering criteria based on reference and variant allele ratios and barcode counts (see Methods) we were able to remove 3,647-13,840 FP variants depending on the library and amount of coverage. Importantly, this was achieved while only increasing the FN rate by 0.10-0.29% in the stLFR libraries. After this filtering step we examined the shared FPs between the four stLFR libraries. Filtering removed only 340 shared FP variants, of which 147 were cluster within 100 base pairs of each other and likely not real (Figure S5b). This further suggests most of these shared FPs are real variants. Taking into account these variants and the reduced number of FP variants after filtering results in a similar FP rate and a 2-3 fold higher FN rate than the filtered STD library for SNP calling (Table S1). This increased FN rate is primarily due to increased non-unique mapping of mate-pairs with short insert sizes in stLFR libraries.

stLFR phasing performance

To evaluate variant phasing performance high confidence variants from GIAB were phased using the publicly available software package HapCut2[24]. Over 99% of all heterozygous SNPs were placed into contigs with N50s ranging from 0.6-15.1 Mb depending on the library type and the amount of sequence data (Table 2). The stLFR-1 library with 336 Gb of total read coverage (44X unique genome coverage) achieved the highest phasing performance with an N50 of 15.1 Mb. N50 length appeared to be mostly affected by length and coverage of long genomic fragments. This can be seen in the decreased N50 of stLFR-2 as the DNA used for this sample was slightly older and more fragmented than the material used for stLFR-1 (Table 2, average fragment length of 52.5 kb versus 62.2 kb) and the ~10-fold shorter N50 of the 10 ng libraries (stLFR-3 and 4). Comparison to GIAB data showed that short and long switch error rates were low and comparable to previous studies[11,15,25]. stLFR performance was very similar to the Chromium library. As the Zhang et al. bead haplotyping method did not have read data available we could only compare our results to the results from their phasing algorithm written and optimized specifically for their data. This demonstrated that stLFR-1 and stLFR-2 libraries had a longer N50, a similar short switch error rate, but a higher long switch error rate. stLFR-3 and stLFR-4, which used more DNA, had an N50 similar to the Zhang et al. However, direct comparison is difficult due to differences in DNA input and coverage.

It should be noted that this phasing result was achieved using a program that was not written for stLFR data. In order to see if this result could be improved we developed a phasing program, LongHap, and optimized it specifically for stLFR data. Using GIAB variants LongHap was able to phase over 99% of SNPs into contigs with an N50 of 18.1 Mb (Table 2). Importantly, these increased contigs lengths were achieved while decreasing the short and long switch errors (Table 2). LongHap is also able to phase indels. Applying LongHap to stLFR-1 using GIAB SNPs and indels results in a 23.4 Mb N50, but also results in increased switch error rates (Table S2).

Structural variation detection

Previous studies have shown that long fragment information can improve the detection of structural variations (SVs) and described large deletions (4-155 kb) in NA12878[11,15]. To demonstrate the power of stLFR to detect SVs we examined barcode overlap data, as previously described[15], for stLFR-1 and stLFR-4 libraries in these regions. In every case the deletion was observed in the stLFR-1 data, even at lower coverage (Figure 2a and Figure S6). Closer examination of the co-barcoded sequence reads covering a ~150 kb deletion in chromosome 8 demonstrated that the deletion was heterozygous and found in a single haplotype (Figure 2b-c). The 10 ng stLFR-4 library also detected most of the deletions, but the three smallest were difficult to identify due to the lower coverage per fragment (and thus less barcode overlap) of this library.

To evaluate stLFR performance for detecting other types of SVs we made libraries from a cell line from a patient with a known translocation between chromosomes 5 and 12[26] and GM20759, a cell line with a known inversion on chromosome 2[27]. stLFR libraries were able to identify the inversion and the translocation in the respective cell lines (Figure 2d-e). Downsampling the amount of reads per library showed that a strong signal of the translocations was detected even with as little as 5 Gb of read data (~1.7X total coverage, Figure S7a-h). Finally, examination of both SVs in the stLFR-1 library resulted in no obvious pattern (Figure S7i-l), suggesting the false positive rate for detection of these types of SVs is low.

Scaffolding contigs with stLFR

stLFR is a powerful method because it uses ~1.8 billion unique barcodes and enables co-barcoding that is specific to each individual long genomic DNA molecule. This type of data should be beneficial for *de novo* genome assembly and improved scaffolding. To demonstrate how stLFR can be used to improve genome assemblies we used reads from stLFR-1 and stLFR-4 libraries and SALSA[28], a program designed for chromatin conformation capture (Hi-C) data, to scaffold Single Molecule Real-Time (SMRT) read assemblies of NA12878[29]. SALSA was not designed for stLFR data, making it necessary to alter the stLFR data to a structure similar to Hi-C. This was achieved by selecting pairs of reads sharing the same barcode and located towards the ends of the captured long DNA molecule. These were then labeled as read pairs for the SALSA program. Substituting stLFR data for Hi-C data resulted in excellent scaffolding. Using only 60 million stLFR reads enabled the linkage of 1,411 contigs into 597 scaffolds with an N50 of 44.7 Mb. These scaffolds covered 2.84 Gb of the genome. These metrics compared very favorably to those generated in the SALSA manuscript using the same contigs and 10-fold more (734 million) Hi-C read pairs generated from human embryonic stem cells[30] (Table 3). The quality of stLFR scaffolds was further analyzed by aligning them to build 37 of the human reference genome and comparing them with the program dnadiff[31]. In general, stLFR scaffolds agreed closely with the reference genome and the number of breakpoints, translocations, relocations, and inversions was similar to those of the scaffolds generated with Hi-C reads (Table 3). Alignment dot plots further demonstrate the high degree of continuity between stLFR scaffolds and the reference genome (Figure S8).

Discussion

Here we describe an efficient whole genome sequencing library preparation technology, stLFR, that enables the co-barcoding of sub-fragments of long genomic DNA molecules with a single unique clonal barcode in a single tube process. Using microbeads as miniaturized virtual compartments allows a practically unlimited number of clonal barcodes to be used per sample at a negligible cost. Our optimized hybridization-based capture of transposon inserted DNA on beads, combined with 3'-branch ligation and exonuclease degradation of the extreme excess of capture adapters, successfully barcodes up to ~20% of sub-fragments in DNA molecules as long as 300 kb in length. Importantly, this is achieved without DNA amplification of initial long DNA fragments and the representation bias that comes with it. In this way, stLFR solves the cost and limited co-barcoding capacity of emulsion-based methods.

The quality of variant calls using stLFR is very high and possibly, with further optimization, will approach that of standard WGS methods, but with the added benefit that co-barcoding enables advanced informatics applications. We demonstrate high quality, near complete phasing of the genome into long contigs with extremely low error rates, detection of SVs, and scaffolding of contigs to enable *de novo* assembly applications. All of this is achieved from a single library that does not require special equipment nor add significantly to the cost of library preparation.

As a result of efficient barcoding, we successfully used as little as 1 ng of human DNA (600 X genome coverage) to make stLFR libraries and achieved high quality WGS with most sub-fragments uniquely co-barcoded. Less DNA can be used, but stLFR does not use DNA amplification during co-barcoding and thus does not create overlapping sub-fragments from each individual long DNA molecule. For this reason overall genomic coverage suffers as the amount of DNA is lowered. In addition, a sampling problem is created as stLFR currently retains 10-20% of each original long DNA molecule followed by PCR amplification. This results in a relatively high duplication rate of reads and results in added sequencing cost, but improvements are possible. One potential solution is to remove the PCR step. This would eliminate sampling, but also it could substantially reduce the false positive and false negative error rates. In addition, improvements such as optimizing the distance of insertion between transposons and increasing the length of sequencing reads to paired-end 200 bases should be easy to enable and will increase the coverage and overall quality. For some applications, such as structural variation detection, using less DNA and less coverage may be desirable. As we demonstrate in this paper, as little as 5 Gb of sequence coverage can faithfully detect inter and intrachromosomal translocations and in these cases the duplication rate is negligible. Indeed, stLFR may represent a simple and cost-effective replacement for long mate pair libraries in a clinical setting.

In addition, we believe this type of data can enable full diploid phased *de novo* assembly from a single stLFR library without the need for long physical reads such as those generated by SMRT or nanopore technologies. One interesting feature of transposon insertion is that it creates a 9 base sequence overlap between adjacent sub-fragments. Frequently, these neighboring sub-fragments are captured and sequenced enabling reads to be synthetically doubled in length (e.g., for 200 base reads, two neighboring captured sub-fragments would create two 200 base reads with a 9 base overlap, or 391 bases). stLFR does not require special equipment like droplet based microfluidic methods and the cost per sample is minimal. In this paper we demonstrated using 50 million beads but using more is possible. This will enable many types of cost-effective analyses where 100s of millions of barcodes would be useful. We envision this type of cheap massive barcoding can be useful for RNA analyses such as full-length mRNA sequencing from 1,000s of cells by combination with single cell technologies or deep population sequencing of 16S RNA in microbial samples. Phased chromatin mapping by the Assay for Transposase-Accessible Chromatin (ATAC-seq)[32] or methylation studies are all also possible with stLFR. Finally, in an effort to share what we believe to be a very important technology, we have made a detailed protocol freely available for academic use (see Supplementary Materials).

Acknowledgments We would like to acknowledge the ongoing contributions and support of all Complete Genomics and BGI-Shenzhen employees, in particular the many highly skilled individuals that work in the libraries, reagents, and sequencing groups that make it possible to generate high quality whole genome data. We would also like to thank Z. Dong, Z. Yang, and W. Xie for providing cell lines for the translocation analysis. This work was supported in part by the Shenzhen Peacock Plan (NO.KQTD20150330171505310) and the National Key Research and Development Program of China (NO.2017YFC0906501). B.A.P. is a recipient of and this work was partially supported by the Research Fund for International Young Scientists, National Natural Science Foundation of China (31550110216).

Authors contributions R.D. and B.A.P. conceived the study. O.W., R.C., X.C., M.K.W., H.K.L., D.C., L.W., F.F., Y.Z., S.D., D.N., A.A., X.X., R.D., and B.A.P. developed the molecular biology process of stLFR. R.Y.Z., S.D., S.G., N.B., and A.C. performed the sequencing. Q.M., J.T., Y.S., Y.Z., E.A., Y.X., C.V., S.N., W.T., J.W., X.L., X.Q., H.W., Y.D., and Z.L. developed algorithms for and performed analyses on stLFR data. O.W., C.X., J.S.L., W.Z., H.Y., J.W., K.K., X.X., R.D., and B.A.P. coordinated the study. O.W., R.D., and B.A.P. wrote the manuscript. All authors reviewed and edited the manuscript.

Completing interests Employees of BGI and Complete Genomics have stock holdings in BGI.

Data and materials availability All sequencing data reported in this paper have been deposited in the database of the European Nucleotide Archive under accession number @@@.

References

1 Zhang, K. *et al.* Long-range polony haplotyping of individual human chromosome molecules. *Nat Genet* **38**, 382-387 (2006).

2 Ma, L. *et al.* Direct determination of molecular haplotypes by chromosome microdissection. *Nat Methods* **7**, 299-301 (2010).

3 Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol* **29**, 59-63 (2011).

4 Suk, E. K. *et al.* A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res* **21**, 1672–1685 (2011).

5 Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nat Biotechnol* **29**, 51-57 (2011).

6 Peters, B. A. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190-195 (2012).

7 Duitama, J. *et al.* Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. *Nucleic Acids Res* **40**, 2041-2053 (2012).

8 Selvaraj, S., J, R. D., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol* **31**, 1111-1118 (2013).

9 Kuleshov, V. *et al.* Whole-genome haplotyping using long reads and statistical methods. *Nat Biotechnol* **32**, 261-266 (2014).

10 Amini, S. *et al.* Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat Genet* **46**, 1343-1349 (2014).

11 Zheng, G. X. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* (2016).

12 Peters, B. A., Liu, J. & Drmanac, R. Co-barcoded sequence reads from long DNA fragments: a cost-effective solution for "perfect genome" sequencing. *Frontiers in genetics* **5**, 466 (2014).

13 Drmanac, R. Nucleic Acid Analysis by Random Mixtures of Non-Overlapping Fragments. WO 2006/138284 A2 (2006).

14 Drmanac, R., Peters, B.A., Alexeev, A. Multiple tagging of individual DNA fragments. WO 2014/145820 A2 (2013).

15 Zhang, F. *et al.* Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. *Nat Biotechnol* **35**, 852-857 (2017).

16 Jo, K., Chen, Y. L., de Pablo, J. J. & Schwartz, D. C. Elongation and migration of single DNA molecules in microchannels using oscillatory shear flows. *Lab Chip* **9**, 2348-2355 (2009).

17 Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78-81 (2010).

18 Fehlmann, T. *et al.* cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin Epigenetics* **8**, 123 (2016).

19 Huang, J. *et al.* A reference human genome dataset of the BGISEQ-500 sequencer. *Gigascience* **6**, 1-9 (2017).

20 Mak, S. S. T. *et al.* Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *Gigascience* **6**, 1-13 (2017).

21 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

22 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).

23 Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* **32**, 246-251 (2014).

24 Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res* **27**, 801-812 (2017).

25 Mao, Q. *et al.* The whole genome sequences and experimentally phased haplotypes of over 100 personal genomes. *Gigascience* **5**, 1-9 (2016).

26 Dong, Z. *et al.* Low-pass whole-genome sequencing in clinical cytogenetics: a validated approach. *Genet Med* **18**, 940-948 (2016).

27 Dong, Z. *et al.* Identification of balanced chromosomal rearrangements previously unknown among participants in the 1000 Genomes Project: implications for interpretation of structural variation in genomes and the future of clinical cytogenetics. *Genet Med* (2017).

28 Ghurye, J., Pop, M., Koren, S., Bickhart, D. & Chin, C. S. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* **18**, 527 (2017).

29 Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods* **12**, 780-786 (2015).

30 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).

31 Phillippy, A. M., Schatz, M. C. & Pop, M. Genome assembly forensics: finding the elusive mis-assembly. *Genome biology* **9**, R55 (2008).

32 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218 (2013).

Supplementary Materials:

Materials and Methods

Figures S1-S9

Tables S1-S3

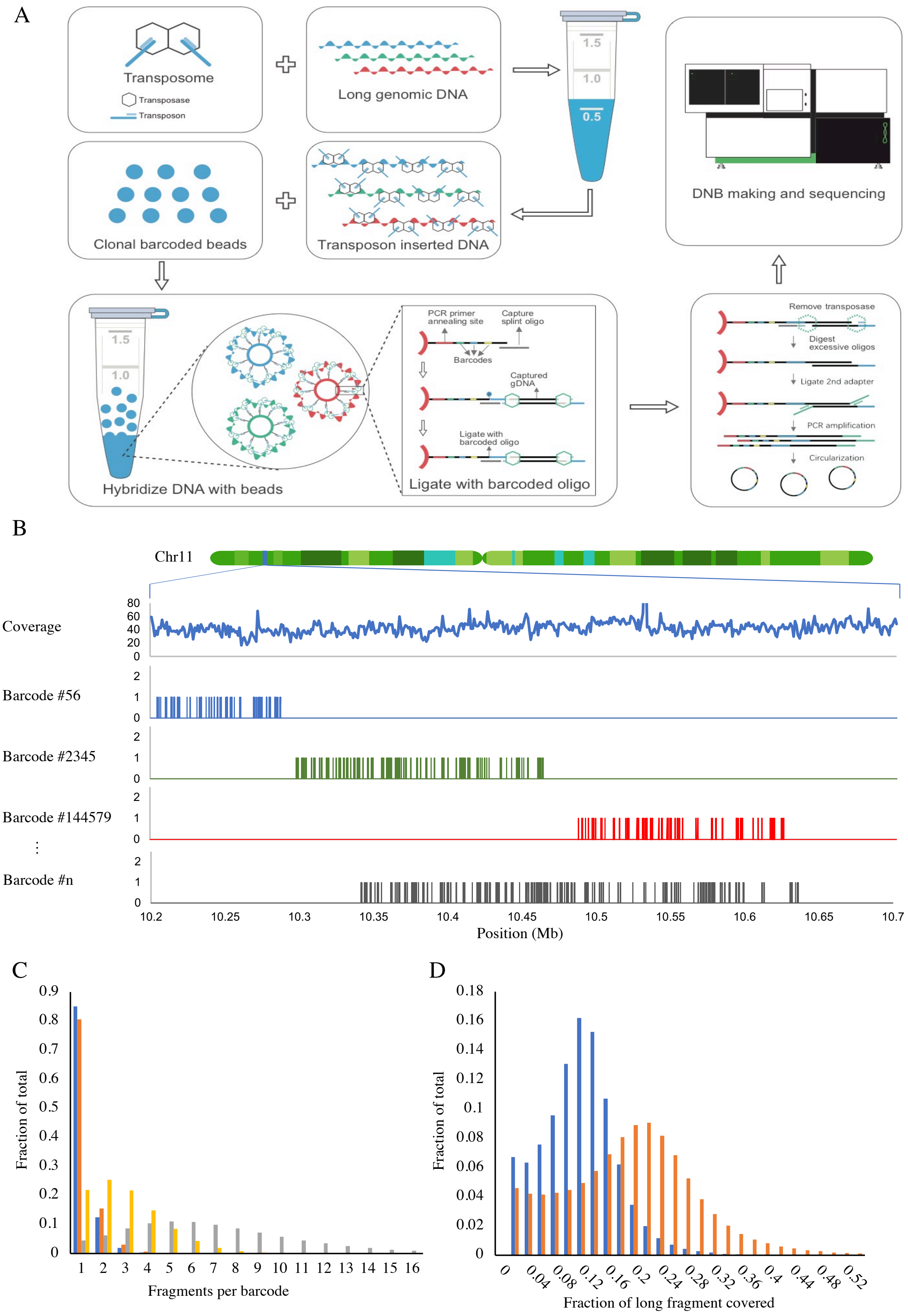
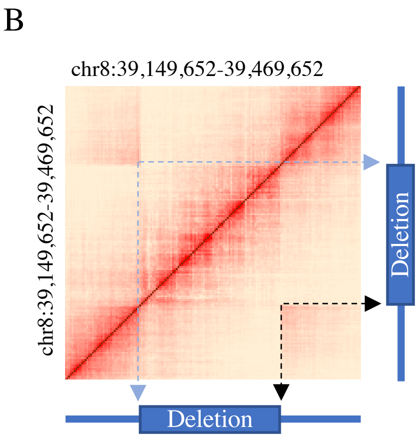

A
Transposome
Transposase
Transposon
Long genomic DNA
1.5
1.0
0.5
DNB making and sequencing
Clonal barcoded beads
Transposon inserted DNA
1.5
1.0
Hybridize DNA with beads
PCR primer annealing site
Capture splint oligo
Barcodes
Captured gDNA
Ligate with barcoded oligo
Ligate with barcoded oligo
Remove transposase
Digest excessive oligos
Ligate 2nd adapter
PCR amplification
Circularization
B
Chr11
Coverage
80
60
40
20
0
Barcode #56
Barcode #2345
Barcode #144579
⋮
Barcode #n
2
1
0
10.2
10.25
10.3
10.35
10.4
10.45
10.5
10.55
10.6
10.65
10.7
Position (Mb)
C
Fraction of total
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Fragments per barcode
D
Fraction of total
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0 0.04 0.08 0.12 0.16 0.2 0.24 0.28 0.32 0.36 0.4 0.44 0.48 0.52
Fraction of long fragment covered

Figure 1. Overview of stLFR. (A) The first step of stLFR involves inserting a hybridization sequence approximately every 200-1000 base pairs on long genomic DNA molecules. This is achieved using transposons. The transposon integrated DNA is then mixed with beads that each contain ~400,000 copies of an adapter sequence that contains a unique barcode shared by all adapters on the bead, a common PCR primer site, and a common capture sequence that is complementary to the sequence on the integrated transposons. After the genomic DNA is captured to the beads, the transposons are ligated to the barcode adapters. There are a few additional library processing steps and then the co-barcoded sub-fragments are sequenced on a BGISEQ-500 or equivalent sequencer. (B) Mapping read data by barcode results in clustering of reads within 10 to 350 kb regions of the genome. Total coverage and barcode coverage from 4 barcodes are shown for the 1 ng stLFR-1 library across a small region on Chr11. Most barcodes are associated with only one read cluster in the genome. (C) The number of original long DNA fragments per barcode are plotted for the 1 ng libraries stLFR-1 (blue) and stLFR-2 (orange) and the 10 ng stLFR libraries stLFR-3 (yellow) and stLFR-4 (grey). Over 80% of the fragments from the 1 ng stLFR libraries are co-barcoded by a single unique barcode. (D) The fraction of nonoverlapping sequence reads (blue) and captured sub-fragments (orange) covering each original long DNA fragment are plotted for the 1 ng stLFR-1 library.

A

Location	Zygosity	Size (kb)	Found by stLFR
chr1:189,704,509-189,783,359	Het	78.9	Yes
chr3:65,189,000-65,213,999	Het	25.0	Yes
chr3:162,512,134-162,626,335	Hom	114.2	Yes
chr4:116,167,000-116,176,999	Het	10.0	Yes
chr4:187,094,000-187,097,999	Het	4.0	Yes
chr5:104,432,113-104,503,673	Hom	71.6	Yes
chr6:78,967,194-79,036,419	Het	69.2	Yes
chr7:110,182,000-110,187,999	Het	6.0	Yes
chr8:39,232,074-39,387,229	Het	155.2	Yes
chr16:62,545,000-62,549,999	Het	5.0	Yes



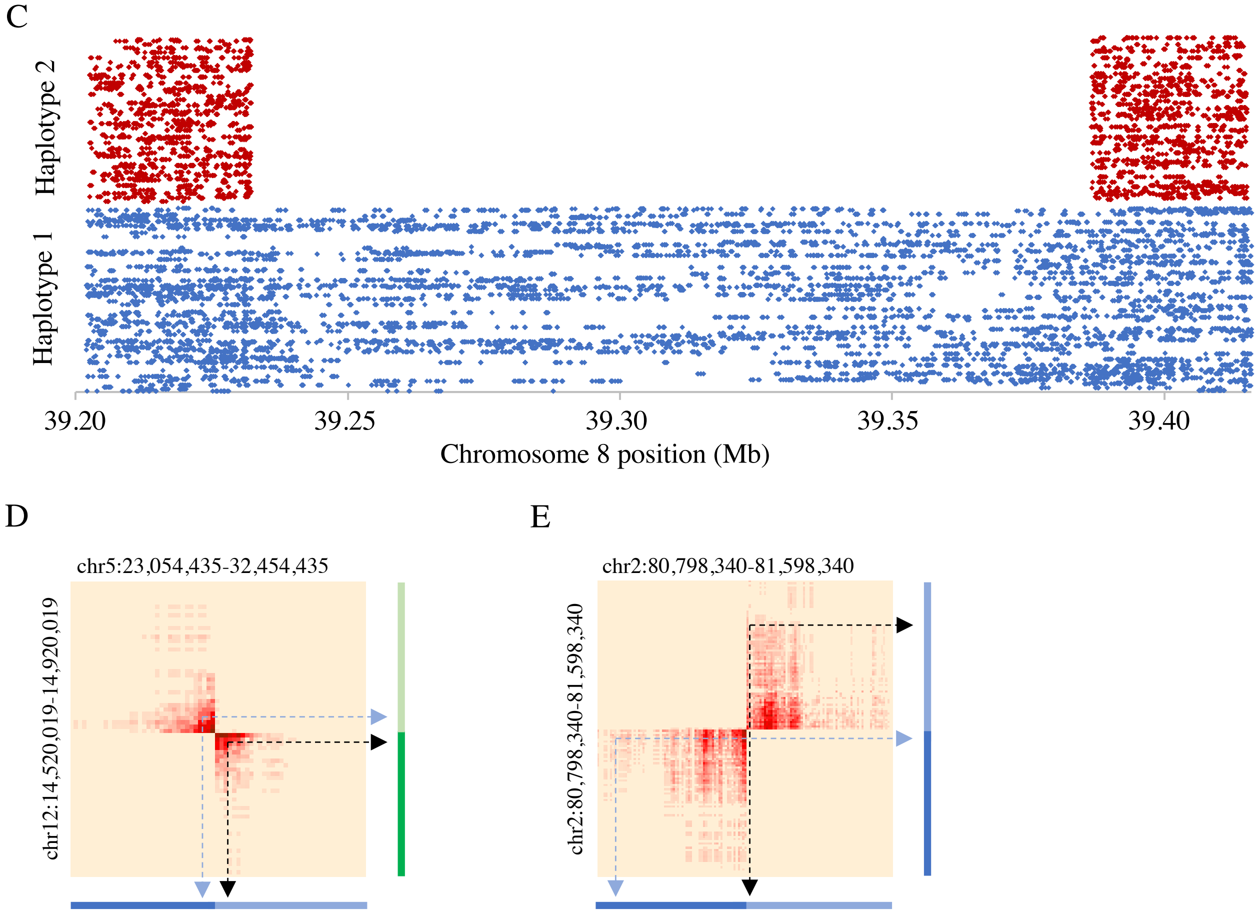


Figure 2. SV detection. (A) Previously reported deletions in NA12878 were also found using stLFR data. Heat maps of barcode sharing for each deletion can be found in Figure S8. (B) A heat map of barcode sharing within windows of 2kb for a region with a ~150 kb heterozygous deletion on chromosome 8 was plotted using a Jaccard Index as previously described[15]. Regions of high overlap are depicted in dark red. Those with no overlap in beige. Arrows demonstrate how regions that are spatially distant from each other on chromosome 8 have increased overlap marking the locations of the deletion. (C) Co-barcoded reads are separated by haplotype and plotted by unique barcode on the y axis and chromosome 8 position on the x axis. The heterozygous deletion is found in a single haplotype. (D) Heat maps were also plotted for overlapping barcodes between chromosomes 5 and 12 for a patient cell line with a known translocation[26] and (E) GM20759, a cell line with a known transversion in chromosome 2[27].

Table 1. stLFR equipment and reagent cost (USD)

Equipment	**One-time cost**	**Per sample**
Sample rotator	~500	
Incubator	~2,000	
Magnetic separation rack	~600	
Thermocycler	~10,000	
Reagents		
Barcode oligos	~50,000[1]	0.13
Streptavidin labeled beads		7
Enzymes for barcoded bead construction		4.40
Enzymes for stLFR library construction		17
Total	~63,100	28.53

[1] Barcode oligonucleotides are listed as a one-time cost because they cannot be purchased on a per sample basis. At a 100 nmol scale synthesis, the cost per sample of oligos is approximately 0.14 dollars.

Table 3. Scaffolding statistics

	stLFR-1	**stLFR-4**	**HiC[1]**	**HiC[2]**
Read pairs (M)	60	134	734	734
Total scaffold length (Gb)	2.84	2.72	2.92	2.92
Scaffold N50 (Mb)	44.7	42.8	68.3	60.02
% aligned bases	98.61%	98.56%	98.22%	94.52%
Scaffold count	597	699	1,411	1,555
Contigs in scaffolds	1,411	1,586	3,096	18,903
Breakpoints	31,386	30,501	35,132	33,079
Relocations	296	327	430	136
Translocations	179	189	406	96
Inversions	624	656	898	408

[1] HiC read pairs from human embryonic stem cells (hESCs)[30] were downloaded and used to scaffold SMRT reads using SALSA[28] and the same process as used for the stLFR libraries.

[2] Results as reported by Ghurye et al.[28] using the same HiC read pairs to scaffold SMRT reads using SALSA.

Table 2. Phasing and variant calling statistics

		stLFR-1			stLFR-2			stLFR-3	stLFR-4	10X Genomics[1]	Illumina Bead Haplotyping[2]	BGISEQ500 STD[3]
Library statistics	Total bases sequenced (Gb)	336	230	100	660	200	100	117	126	128	99	132
	Input genomic DNA (ng)	1	1	1	1	1	1	10	10	1.25	3	1,000
	Average genomic fragment size (kb)	66.2	66.3	66.4	52.5	52.7	52.6	30.2	46.8	85.7	-	N/A
	Unique genome coverage	44X	38X	24X	58X	37X	23X	37X	34X	33X	19X	43X
	Duplicate rate	59.4%	49.6%	29.4%	70.88%	41.05%	25.37%	5.4%	15.0%	6.0%	21.0%	3.7%
	Read length	PE100	PE100	PE100	PE100	PE100	PE100	PE100	PE100	PE150	PE76	PE100
	Unique compartments	10,186,086	10,007,746	9,427,999	11,823,872	10,932,966	10,297,180	30,544,841	10,577,590	1,538,345	147,456	N/A
	Average fragments per compartment	1.18	1.18	1.17	1.25	1.23	1.22	2.87	6.84	8.32	~100	N/A
	Average co-barcoded reads per fragment	80.7	71.5	47.4	88.3	60.2	40.7	7.5	8.9	49.8	5	N/A
No Filter	SNP Precision	0.997	0.997	0.995	0.997	0.997	0.995	0.997	0.993	0.952	0.997	0.998
	SNP Sensitivity	0.996	0.995	0.988	0.997	0.994	0.986	0.996	0.991	0.996	0.952	0.998
	Indel Precision	0.934	0.935	0.924	0.938	0.938	0.924	0.960	0.948	0.639	0.932	0.960
	Indel Sensitivity	0.956	0.951	0.914	0.965	0.950	0.912	0.961	0.925	0.864	0.832	0.972
Filtered	SNP Precision	0.999	0.998	0.997	0.999	0.998	0.996	0.999	0.997	0.994	-	0.999
	SNP Sensitivity	0.995	0.994	0.985	0.995	0.993	0.985	0.995	0.989	0.997	-	0.997
	Indel Precision	0.971	0.965	0.943	0.974	0.964	0.942	0.978	0.964	0.916	-	0.991
	Indel Sensitivity	0.943	0.940	0.902	0.958	0.940	0.902	0.952	0.917	0.871	-	0.962
HapCut2	% heterozygous SNPs phased	99.9%	99.9%	99.8%	99.9%	99.7%	99.7%	98.9%	98.7%	99.9%	98.0%	N/A
	Contig N50 size (Mb)	15.1	12.9	8.6	6.4	4.2	2.6	0.6	1.2	12.8	1.14	N/A
	Short switch error rate	0.00273	0.00272	0.00272	0.00261	0.00272	0.00271	0.00272	0.00571	0.00273	0.0013	N/A
	Long switch error rate	0.00571	0.00571	0.00570	0.00553	0.00570	0.00570	0.00574	0.00276	0.00572	0.000085	N/A
LongHap	% heterozygous SNPs phased	0.999	0.9988	0.9966	0.9991	0.9984	0.9952	0.9895	0.9879	N/A	N/A	N/A
	Contig N50 size (Mb)	18.1	16.6	10.7	8	5.2	3.3	1.1	1.9	N/A	N/A	N/A
	Short switch error rate	0.0025748	0.0025949	0.0026139	0.0025228	0.0025307	0.0025773	0.0027524	0.0030534	N/A	N/A	N/A
	Long switch error rate	0.0017183	0.0017073	0.0017638	0.0017197	0.0017038	0.0017101	0.0019273	0.0020666	N/A	N/A	N/A

Reads were mapped to Hg37 with decoy sequence and variants were called with GATK with default settings for all libraries except where otherwise described. SNPs from the GIAB high-confidence variant calls VCF were used as input for phasing.

[1] The BAM file "NA12878_WGS_v2_phased_possorted_bam.bam" from a recent Chromium dataset was downloaded from the 10X Genomics website (https://support.10xgenomics.com/genome-exome/datasets/2.1.4/NA12878_WGS_v2) and processed in the same manner as the stLFR libraries. For filtered results we used the VCF file "NA12878_WGS_v2_phased_variants.vcf.gz" from the same Chromium library. This VCF contains data that was processed through 10X Genomics' optimized pipeline. The fragment size was for the Chromium library is taken from the 10X Genomics website. 10X Genomics uses a length weighted mean to calculate fragment size which results in a larger size than the average fragment size.

[2] Read data were not available, this is what is reported in the Zhang et al. Nat Biotech 2017 paper[15].

[3] Data from a standard library processed on a BGISEQ-500.