# MGI's stLFR Based on the DNBSEQ Sequencing Platform Empowers Long Read Sequencing

This study developed a novel technique, single-tube long fragment read (stLFR) and systematically evaluate the performance of stLFR on the MGI's DNBSEQ sequencing platform using human DNA standard NA12878.

This work was published in the journal *Genome Research* in 2019, titled "Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and *de novo* assembly"[1].

Recommended application: Long-read DNA sequencing

Recommended models: DNBSEQ-G400RS, DNBSEQ-T7RS

- ## Haplotype assembly

  The phasing rate of heterozygous locations is as high as 99.7%, the maximum N50 of haplotype assembly blocks can reach 10 Mb.

- ## Small variant detection

  With high accuracy and sensitivity in detecting SNP and InDel variants, this technology is similar to single-molecule sequencing, but it has low base error rate and high-quality detection of small variants.

- ## Structural variant (SV) calling

  It can effectively detect structural variants larger than 20kb such as inversion, translocation, deletion and insertion.

- ## User-defined data mining

  It can analyze regions that are difficult to handle with conventional WGS, such as highly homologous regions and highly repetitive regions.

- ## Low DNA input

  The DNA input could be as low as 1 ng for library construction to achieve performance equivalent to a conventional library with an input of 100 ng DNA.

# Background

Currently, the whole genome sequences of the vast majority of organisms lack haplotype information on contiguous blocks of single-base to multi-base variants on homologous chromosomes, and most genomes contain regions where large structural variants and other regions cannot be resolved by short-read sequencing. Early studies did not pay enough attention to this field, which is critical for a complete understanding of how the genome helps to exhibit the phenotypes.

To solve this problem, this study presented the single-tube long fragment read (stLFR) sequencing technology for long DNA molecule sequencing, based on the method of cobarcoding on fragments of long DNA molecules from the same source[2]. Firstly, transposons are inserted into the extracted long DNA molecules under the action of Tn5 transposase in a single tube. There are two methods for transposon ligation, namely two-transposon ligation method and 3' branched ligation method. Afterwards, the transposon was ligated to the beads with multicopy molecular label through the adapter sequence. The captured long DNA molecules were fragmented, PCR amplified and circularized for library construction. The length range of the DNA molecules after sequencing and splicing can reach 20-300Kb, which overcomes the inability by short read sequencing.

The human DNA standard NA12878 was purchased, underwent library preparation with stLFR technology and sequenced on DNBSEQ sequencing platform. The bioinformatics analysis showed high-quality variant calling and an N50 length of 34 Mb for the maximum phased blocks. It also revealed the structural variation (SV) within the NA12878 genome. Sufficient information demonstrates that stLFR is a long fragment sequencing technology that can help researchers effectively explore and excavate genomic regions that are not yet understood.

## Research description

A research team led by the BGI-Research in Shenzhen has developed a new technology called stLFR to address the shortcomings of short-read sequencing in interpreting whole-genome information.

This long fragment read sequencing technology is based on the concept of unseparated cobarcoding and high-throughput short-read sequencing technology with obvious advantages. The team also verified various performance parameters of the technology using DNA standard NA12878 after completing the technical development[1].

## Materials and methods

### Library Preparation

Based on the stLFR technology, MGI has developed the MGIEasy stLFR Library Prep Kit for library construction. The prepared NA12878 gDNA underwent library construction. You can refer to the instructions of the kit for more details.

### Sequencing on the DNBSEQ platform

The constructed stLFR library underwent paired-end 100pb (PE100) sequencing on the BGISEQ-500 genetic sequencer. Currently, MGI has developed the upgraded sequencers DNBSEQ-G400RS and DNBSEQ-T7RS with higher throughput and faster sequencing speed to support stLFR application.

### Bioinformatics analysis

This study used third-party open tools to perform sequence alignment and variant calling of the original read data. Comparison with the GIAB[4] variants set can determine the true positive (TP), false positive (FP), and false negative (FN) rates of variant calling. To further improve the false positive rate of variant information in stLFR sequencing method, we developed a binary classification model based on XGboost[5] for variant filtering. After obtaining the comparison result BAM file and variant data, we used HapCUT2[6] to assemble the haplotypes using these two data as input. SVs were detected by calculating shared cobarcodes[7] between genomic regions, and the ratio between cobarcodes was calculated with Jaccard index to identify structural variants. Long Ranger was used to optimize the barcodes with at least 10 reads to a list of approximately 4.7 million barcodes. The format of stLFR FASTQ files was converted and used as input for Supernova, to generate a pseudo-hap assembly output, and scaffolds with a minimum length of 10Kb were compared to GRCh38.

| Sample collection | Library preparation and sequencing | Bioinformatics Analysis | Result analysis |
|---|---|---|---|
| Human genome NA12878 sample |  MGIEasy stLFR Library Prep Kit<br>  DNBSEQ sequencing platform | Available: MGI-tech-bioinformatics /stLFR_v | Systematic evaluation of the stLFR method |

## Results

### stLFR sequencing library construction process

The first step of stLFR sequencing is to insert a hybridization sequence into long DNA fragments at regular intervals through Tn5 transposase. The transposase contains a single-stranded region for hybridization and a double-stranded sequence that can be recognized by the enzyme.The transposase remains bound to the long DNA fragment after transposition is completed to prevent damage to the long DNA fragment.Afterwards, the long DNA inserted with a hybridization sequence will be captured by barcoded beads containing 400,000 capture sequences (each bead has a unique common barcode), and the long DNA molecules will be trapped around the beads. Next, beads are collected and individual barcode sequences are transferred to each long DNA molecule by ligation of the gap between the hybridization sequence and the capture adapter.

Afterwards, transposase is removed and oligo is digested before proceeding into the next-generation sequencing process through PCR amplification and circularization. This is done using the independently developed patented technology DNB nanoball rolling circle amplification by MGI and finally sequenced with pair-ended 100 (PE100) sequencing using BGISEQ-500 (Figure 1A).

After sequencing, barcode information is extracted using a customized method. It is shown by mapping the read results by unique barcode that most read data with the same barcode are clustered in a region of the genome corresponding to the length of long DNA molecules used during library preparation (Figure 1B).

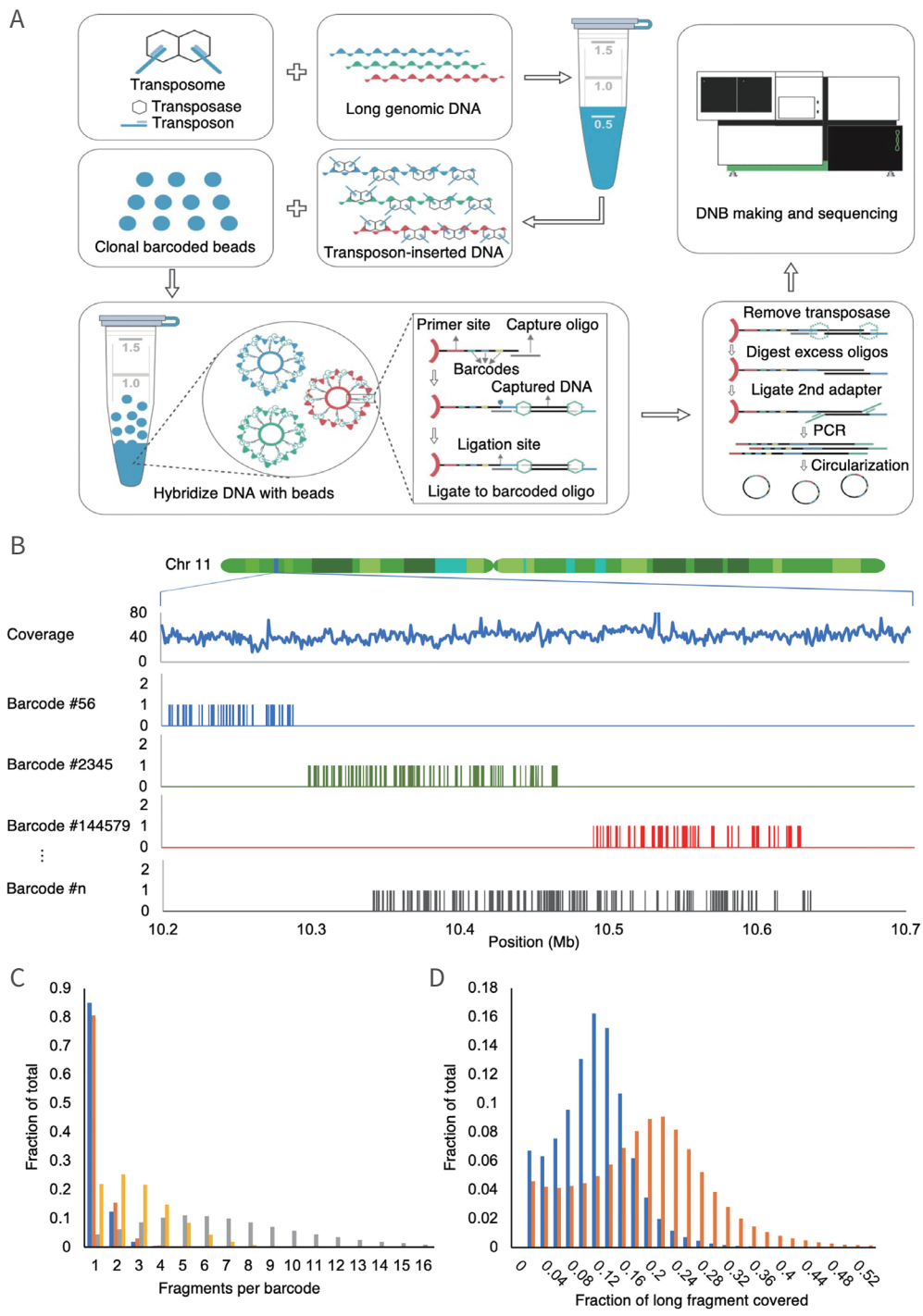


Figure 1 A shows the stLFR library construction workflow; B shows clustered reads mapped by barcode on a small fragment of Chromosome 11; C shows the number of fragments mapped by each barcode in four libraries; D shows the ratio of non-overlapping sequence reads (blue) and captured subfragments (orange) for each original long DNA fragment in stLFR-1 library coverage.

## stLFR sequencing read coverage and variant calling

We used 1 ng (stLFR-1 and stLFR-2) and 10 ng (stLFR-3 and stLFR-4) of DNA to produce libraries. We used 3' branch ligation method for stLFR-1-3 and two-transposon method for stLFR-4. We compared the variant calling results of four starting size as well as different transposon ligation methods, to other variant calling methods. The total base coverage of stLFR-1 and stLFR-2 is 336 Gb and 660 GB, respectively; while stLFR-3 and stLFR-4 reach a more appropriate level of 117 Gb and 126 Gb, respectively (Table 1). The non-duplicate coverage ranged from 34× to 58×. The number of long DNA molecules for each barcode ranged from 1.2 to 6.8 (Figure 1C). It was observed that the highest average non-duplicate coverage for each long DNA molecule was 10.7%-12.1%, and the highest average non-duplicate base coverage for captured subsegments of each long DNA molecule was 17.9%-18.4% (Figure 1D). Analyzing the data after sequencing revealed that stLFR performed much better than other methods of the same type in InDel detection. In SNP detection, the stLFR library will also be slightly superior to the results of other methods. The FP of SNP after filtering is comparable to that before filtering, but the FN rate is 2 to 3 times higher (Table 1).

| | stLFR-1 | stLFR-2 | stLFR-3 | stLFR-4 | 10X Genomics | IlluminaBeads Haplotyping | BGISEQ-500SD | BGISEQ-500 PCR-free SD | BGISEQ-500SD |
|---|---|---|---|---|---|---|---|---|---|
| Library statistics | | | | | | | | | |
| Total bases sequence(Gb) | 336 | 660 | 117 | 126 | 128 | 99 | 81 | 129 | 132 |
| SNP | | | | | | | | | |
| TP | 3194945 | 3197686 | 3193507 | 3175921 | 3202498 | - | 3194780 | 3201626 | 3201452 |
| FP | 9125 | 9544 | 7144 | 9544 | 7144 | - | 5192 | 6372 | 4800 |
| FN | 15312 | 12571 | 16750 | 12571 | 16750 | - | 15477 | 8630 | 8805 |
| Precision | 0.997 | 0.997 | 0.998 | 0.995 | 0.971 | 0.997 | 0.998 | 0.998 | 0.999 |
| Sensitivity | 0.995 | 0.996 | 0.995 | 0.989 | 0.998 | 0.952 | 0.995 | 0.997 | 0.997 |
| TP(Filter) | 3193269 | 3194955 | 3192891 | 3174874 | 3200472 | | 3194254 | 3200960 | 3201273 |
| FP(Filter) | 4491 | 4606 | 4814 | 8982 | 18615 | | 4271 | 4153 | 3111 |
| FN(Filter) | 16988 | 15302 | 17366 | 35382 | 9785 | | 16003 | 9396 | 8984 |
| Precision (Filter) | 0.999 | 0.999 | 0.999 | 0.997 | 0.994 | | 0.999 | 0.999 | 0.999 |
| Sensitivity (Filter) | 0.995 | o.995 | 0.995 | 0.989 | 0.997 | | 0.995 | 0.997 | 0.997 |
| INDEL | | | | | | | | | |
| TP | 460144 | 464451 | 459979 | 440718 | 415613 | - | 463273 | 465316 | 467612 |
| FP | 32437 | 30487 | 17375 | 22075 | 235331 | - | 10541 | 17136 | 19514 |
| FN | 21120 | 16816 | 21288 | 40547 | 65656 | - | 17993 | 15951 | 13655 |
| Precision | 0.934 | 0.938 | 0.964 | 0.952 | 0.636 | 0.932 | 0.978 | 0.965 | 0.96 |
| Sensitivity | 0.956 | 0.965 | 0.957 | 0.916 | 0.864 | 0.832 | 0.963 | 0.967 | 0.972 |

Table 1 Comparison of variant calling results between stLFR and other methods.

## stLFR sequencing phasing analysis

HapCUT2 method was used for variant phasing. 99% of heterozygous SNPs were oriented within the phase blocks. The phasing information on the diploid genome can be obtained using short read sequences with cobarcode, which can resolve the combination of gene regulation and coding region variations. The results show that the read coverage of stLFR-1 is high. At a depth of 40×, the N50 value of the phase block of stLFR-1 library data can reach up to 34 Mb. The proportion of heterozygous locations that can be phased is as high as 99.7% (Figure 2).

## Structural variation (SV) detection

To evaluate the ability of stLFR technology to accurately detect various structural variants, this study detected stLFR-1 and stLFR-4 based on known SV locations, and found that deletion locations could be detected even with low coverage (Figure 3A). We detected a heterozygous deletion of 150Kb in Chromosome 8 of NA12878 (Figure 3B and 3C). This study verified the stLFR performance for detecting other types of SVs using a cell line from a patient[8] with a known translocation between Chromosome 5 and Chromosome 12 (Figure 3D) and a GM20759 cell line[9] with known inversion (Figure 3E).

## *De novo* assembly with stLFR

85% of the fragments (regions smaller than 300kb in the genome) in the 1ng stLFR-1 library were cobardcoded with a unique barcode, which helped us simplify and improve the *de novo* assembly.To test the ability of stLFR, we used stLFR-1 and stLFR-2 libraries for *de novo* assembly. We compared the assembled contigs with the human reference genome GRCh38 chromosomes, and the results show high consistency (Figure 4).

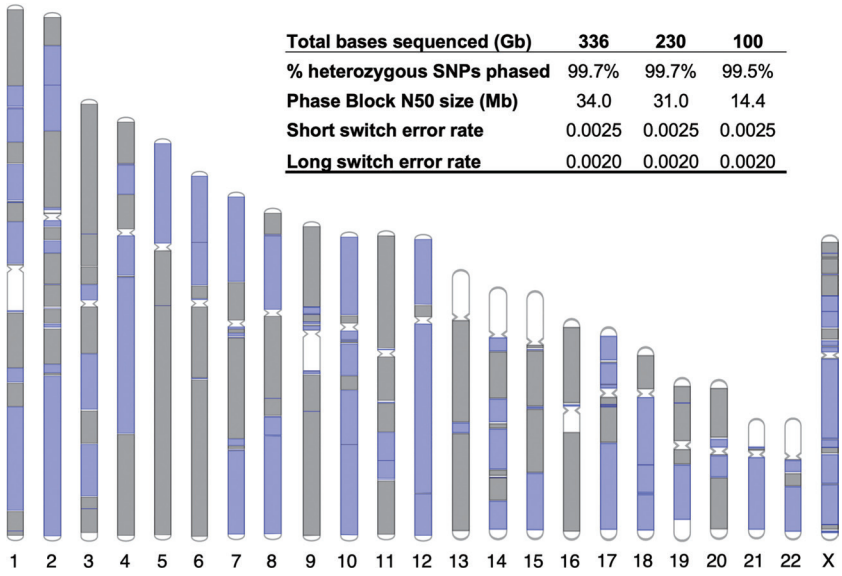


Figure 2 Statistical analysis of stLFR-1 phasing data.

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 |

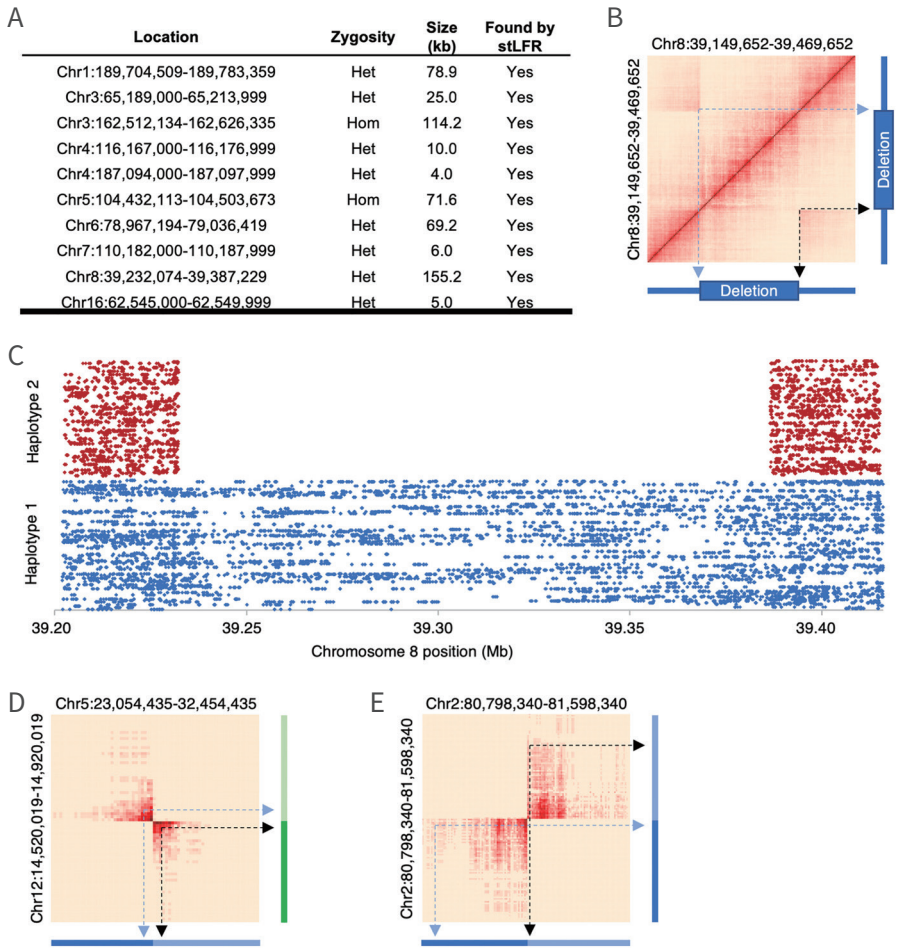


Figure 3. A shows the verification of stLFR's ability to detect SVs based on known locations; B and C show a deletion in a fragment of Chromosome 8; D shows the consistent detection of the translocation in patient cell line; E shows the consistent detection of the inversion in GM20759[9].

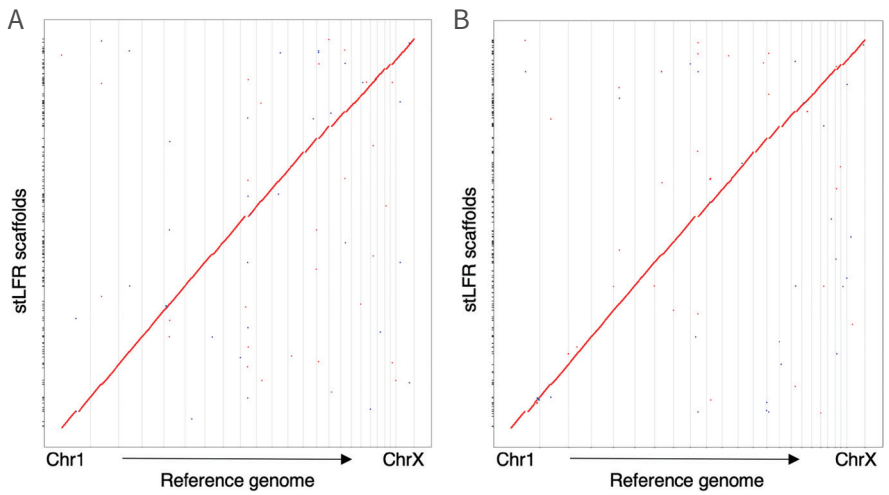


Figure 4. A and B represent the comparison of contigs from *de novo* assemblies of stLFR-1 and stLFR-2 libraries to GRCh38, respectively.

## Conclusion

This study validated the applicability of single-tube long fragment read (stLFR) sequencing technology using the NA12878 cell line. stLFR technology did not significantly increase the time or cost of whole genome sequencing (WGS) library preparation. This technology can achieve high-quality sequencing, phasing, SV calling, diploid *de novo* genome assembly, and other long DNA sequencing applications.

The MGIEasy stLFR Library Preparation Kit developed based on this technology can perfectly fit the DNBSEQ sequencing platform from MGI to carry out long-length sequencing study. The DNBSEQ sequencing platform offers the benefits of high accuracy, low repeat rate and low index hopping.


Genetic Sequencer DNBSEQ-G400

## References


1. Wang, O., *et al.*, Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and *de novo* assembly. *Genome Res*, 2019. **29**(5): p. 798-808.
2. Peters, B.A., J. Liu, and R. Drmanac, Co-barcoded sequence reads from long DNA fragments: a cost-effective solution for "perfect genome" sequencing. *Front Genet*, 2014. **5**: p. 466.
3. Wang, L., *et al.*, 3' Branch ligation: a novel method to ligate non-complementary DNA to recessed or internal 3'OH ends in DNA or RNA. *DNA Res*, 2019. **26**(1): p. 45-53.
4. Zook, J.M., *et al.*, Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol*, 2014. **32**(3): p. 246-51.
5. Chen, T. and C. Guestrin, XGBoost, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 785-794.
6. Edge, P., V. Bafna, and V. Bansal, HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res*, 2017. **27**(5): p. 801-812.
7. Zhang, F., *et al.*, Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. *Nat Biotechnol*, 2017. **35**(9): p. 852-857.
8. Dong, Z., *et al.*, Low-pass whole-genome sequencing in clinical cytogenetics: a validated approach. *Genet Med*, 2016. **18**(9): p. 940-8.
9. 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*, 2018. **20**(7): p. 697-707.

## Recommended Ordering Information

| Category | Product | Cat. NO. |
|---|---|---|
| Instruments | Genetic Sequencer DNBSEQ-G400RS | 900-000170-00 |
| | Genetic Sequencer DNBSEQ-T7RS | 900-000128-00 |
| | MGISP-960RS Automated Sample Preparation System | 900-000146-00 |
| Software | MegaBOLT Bioinformatics analysis accelerator | 900-000555-00 |
| | Data Center Appliance | 900-000444-00 |
| | MGI-tech-bioinformatics/stLFR_v1 | https://github.com/MGI-tech-bioinformatics/stLFR_v1 |
| Library Prep | MGIEasy stLFR Library Prep Kit (16 RXN) | 940-000193-00 |
| Sequencing Reagents | DNBSEQ-G400RS High-throughput Sequencing Set (stLFR FCL PE100) | 1000016984 |
| | DNBSEQ-T7RS High-throughput Sequencing Set (stLFR FCL PE100) | 1000019251 |

# MGI Tech Co.,Ltd

Building 11, Beishan Industrial Zone, Yantian District, Shenzhen, CHINA, 518083



Version: December 2023

+86-4000-688-114
en.mgi-tech.com
MGI-service@mgi-tech.com

Written by: Jinlan Li ,Yanling Huang
Edited by: Wang Qiwei
Examined by: Jiang Yao

1. For StandardMPS and CoolMPS: Unless otherwise informed, StandardMPS and CoolMPS sequencing reagents, and sequencers for use with such reagents are not available in Germany, Spain, UK, Sweden, Belgium, Italy, Finland, Czech Republic, Switzerland, Portugal, Austria and Romania. Unless otherwise informed, StandardMPS sequencing reagents, and sequencers for use with such reagents are not available in Hong Kong. No purchase orders for StandardMPS products will be accepted in the USA until after January 1, 2023.
2. For HotMPS sequencers: This sequencer is only available in selected countries, and its software has been specially configured to be used in conjunction with MGI's HotMPS sequencing reagents exclusively.
3. For HotMPS reagents: This sequencing reagent is only available in selected countries.