Quantitative Biology 2016, 4(2): 106–114
DOI 10.1007/s40484-016-0068-z

REVIEW

Comparison of the experimental methods in haplotype sequencing via next generation sequencing

Jing Tu, Na Lu, Mengqin Duan, An Ju, Xiao Sun and Zuhong Lu*

State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China
* Correspondence: zhlu@seu.edu.cn

Received November 8, 2015; Revised January 22, 2016; Accepted February 26, 2016

Although the diploid nature has been observed for over 50 years, phasing the diploid is still a laborious task. The speed and throughput of next generation sequencing have largely increased in the past decades. However, the short read-length remains one of the biggest challenges of haplotype analysis. For instance, reads as short as 150 bp span no more than one variant in most cases. Numerous experimental technologies have been developed to overcome this challenge. Distance, complexity and accuracy of the linkages obtained are the main factors to evaluate the efficiency of whole genome haplotyping methods. Here, we review these experimental technologies, evaluating their efficiency in linkages obtaining and system complexity. The technologies are organized into four categories based on its strategy: (i) chromosomes separation, (ii) dilution pools, (iii) crosslinking and proximity ligation, (ix) long-read technologies. Within each category, several subsections are listed to classify each technology. Innovative experimental strategies are expected to have high-quality performance, low cost and be labor-saving, which will be largely desired in the future.

Keywords: next generation sequencing; haplotyping; haplotype sequencing

INTRODUCTION

Each normal human genome is a diploid which composed of two sets of 23 chromosomes. One set inherit from the mother, and the other inherit from the father. Alleles at multiple loci along a single chromosome are referred to haplotype. Haplotype information is essential to explain the relationships between genotypes and phenotypes [1–3], map disease genes roundly [4] and describe genetic ancestry completely [5].

Although the diploid nature has been observed for over 50 years [6–8], phasing adiploid still a laborious task. Up till now, karyotyping is the gold standard in clinical laboratories. The development of DNA microarray and chromosomal fluorescence *in situ* hybridization (FISH) exhibits additional but still limited haplotype information [9,10]. DNA sequencing, which obtains nucleotide sequence information one by one, is a direct and efficient haplotyping technology. In fact, the first two assembled human genomes generated by the Human Genome Project contained extensive haplotype information [11,12] by constructing 50–200 kb bacterial artificial chromosomes (BACs). Though mate-pair libraries may be helpful, it still remains a gigantic project to sequence an individual diploid genome by Sanger dideoxy technology [13].

With the advent of next generation sequencing (NGS) in 2005, the cost of DNA sequencing has reduced over 100,000-fold, with its speed greatly increasing [14–18]. However, the short read length is a challenge to haplotype analysis, as the reads shorter than 150 bp span no more than one variant in most cases. Assisted by the paired-end libraries, the linkage obtained was extended to 250–500 bp. The complex mate-paired libraries, which require an *in vitro* circularization step, obtained the linkage in blocks for maximum length of 3.5 kb [19].

In the past decade, numerous experimental technologies have been developed for whole genome haplotyping based on NGS. Distance, complexity and accuracy of the linkages generated are among the main factors to evaluate

This article is dedicated to the Special Collection of Recent Advances in Next-Generation Bioinformatics (Ed. Xuegong Zhang).

the efficiency of whole genome haplotyping methods. Here, we review these experimental technologies and evaluate their efficiency in linkages obtaining and experimental system complexity. Some statistical methods are able to resolve haplotypes independently by population analysis [20–22]. However, most computational methods are designed to optimize the haplotype-resolving efficiency of certain experimental strategy. In this paper, therefore, we emphasize on reviewing such experimental technologies. The technologies are organized into four categories based on the following strategies: (i) chromosomes separation, (ii) dilution pools, (iii) crosslinking and proximity ligation, (ix) long-read technologies. Within each category, several subsections are listed to classify the each technology. Long-read technologies are not independent strategies but technical renovations in sequencing technology. We categorize them as a separate section for their potency in effects improvement of all experimental technologies.

CHROMOSOMES SEPARATION

Human gametes contain natural sets of homologous chromosomes. Physically separating homologous chromosomes before the sequencing library construction is a kind of direct method to obtain long distance, complex and pure linkages. The artificial separation is limited by intact mitotic cells, complex experimental pipelines and specific devices.

Human gametes

Human gamete is an ideal sample for haplotype study, within which a natural set of homologous chromosomes is packaged. Recently, several groups have operated whole genome sequencing and haplotyping of individual sperm [23–25]. Due to the small DNA amount within a sperm, nucleic acids of a single sperm were amplified by multiple displacement amplification (MDA) before sequencing library construction. This isothermal amplification with random hexamer primers and phi29 DNA polymerase amplifies DNA in a cascading, strand displacement reaction [26]. After 4–6 h amplification at 30°C, the yield of amplified DNA is over a 5-log range of starting material (100 fg–10 ng), exceeding 10 kb in length. Hou *et al.* [27] performed genome-wide haplotyping of a human oocyte by analysis of polar bodies. Multiple annealing and looping based amplification cycles (MALBAC) was used to perform high uniform amplification across the genome. To some extent, studies in human gametes can provide alternative solutions to substitute artificial separation of chromosomes. However, the most visible shortcoming is that almost all the other samples or cells are not haploid as human gametes.

Laser capture microdissection

Ma *et al.* [28] determined haplotypes through chromosome microdissection. A part of chromosomes from one cell were collected by computer-directed laser micro-

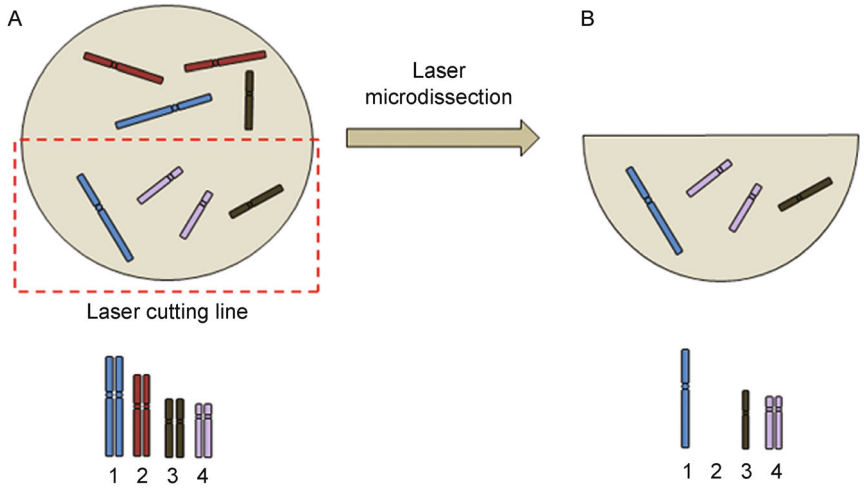

Figure 1. The principle of laser capture microdissection-based haplotyping. Computer-directed laser microdissects at the red dotted line of a diploid cell with 4 chromosome pairs. 4 chromosomes are harvested, chromosome 1 and 3 are monosomic, chromosome 4 is disomic, and chromosome 2 is null.

dissection (Figure 1). The collection may contain only one copy of some chromosomes, and may also contain no copy or both copies of other chromosomes. The haplotypes of monosomic chromosomes were revealed by conventional genotyping after MDA. The MDA products of microdissection harvests are suitable for NGS, though Ma *et al.* [28] used microarrays for genotyping. However, the microdissection depends on the positions where the chromosomes are located. The collected chromosomes with only one copy are random. Inestimable microdissections and collections are required to analyze each chromosome of the two sets.

Fluorescence-activated sorting

Fluorescence-activated cell sorting (FACS), an efficient cell sorting technology, was used by Yang *et al.* [29] to place individual chromosomes into wells of a 96-well plate. In their study, Chromomycin A3 (binds Guanine-Cytosine-rich regions) and Hoechst 33258 (binds Adenine-Thymine-rich regions) were for staining. Each chromosome was identified by its distinct bivariate distribution of fluorescent signals from staining. MDA and NGS were operated for haplotype analysis. Additional molecular typing was required to deal with similar bivariate distribution patterns of chromosomes [29].

Microfluidic devices

Microfluidics was developed based on the need to analyze cells and biomolecules more efficiently. Fan *et al.* [30] developed a microfluidic device to separate and amplify homologous copies of each chromosome from a single human metaphase cell. The amplification products of each chromosome or small chromosomes pool were genotyped by microarrays. The microfluidic device is compact and precise, with only two chromosomes not collected. Manual identification of metaphase cells and manufacture of microfluidic devices are the labor-intensive procedures.

DILUTION POOLS

Dilution pools strategy, a classic method for linkage mapping, was first conceived over 20 years ago [31]. Long intact genomic DNA fragments are compartmented into pools after limiting dilution. In each pool, DNA is sub-haploid amount. Part of the genomic regions are represented once within each pool, while the left are not represented (Figure 2). The entire genome is covered enough times by the pools collectively. Due to the lack of comprehensive whole genome amplification method to microscale DNA, dilution pools strategy was first applied in systematic haplotyping by means of fosmid clone [13,32], genotyping by microarray or Sanger dideoxy sequencing. MDA [26] provides uniform representation across the genome to microscale DNA and makes clone no longer indispensible to dilution pools strategy (Table 1). Dilution pools strategies carry out haplotype analysis without physically separating of homologous chromosomes, and the system complexity is lower than

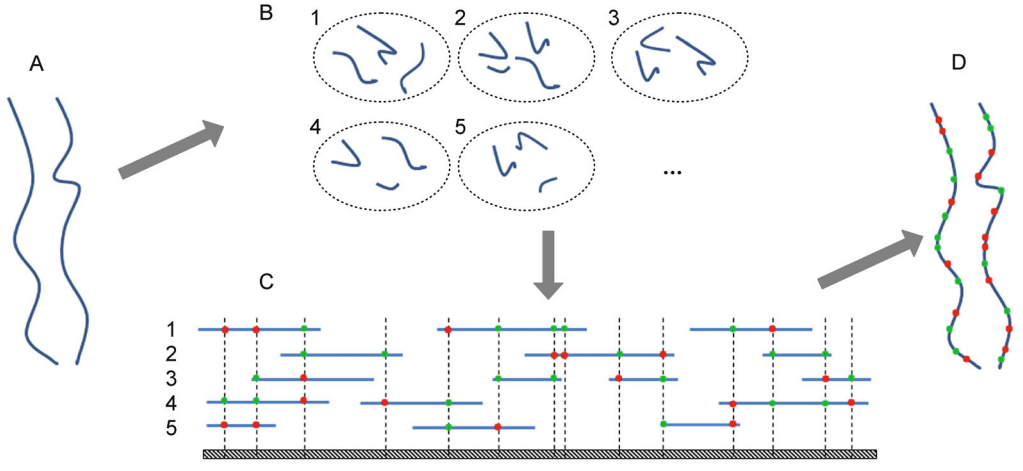

Figure 2. Dilution pool strategy to whole genome haplotype-resolve. (A) Diploid genomic DNA is used to generate different size fragments, from 140 Kb of BAC clones to 10 Kb of long-range PCR products. (B) Fragments are partitioned into pools, each covering 10%–40% of the haploid genome. (C) Fragment pools are sequencing using NGS. In this sample, fragments from five pools are mapped to the genome and the heterozygous positions are detected. (D) Based on allelic identity at overlapping positions, the fragments are separated into two haplotypes. With generous sequencing data, contiguous haplotype blocks can cover the entire chromosome.

Table 1. Comparison of dilution pools strategies.

	Fosmid clone pools	BAC clone pools	Fragment pools with MDA	Fragment pools with CPT-seq	Fragment pools with long-range PCR
Library preparation	Fosmid clone	BAC clone	MDA	Contiguity-preserving transposition	Long-range PCR
Linkage distance	35 kb	140 kb	30–140 kb	~150 kb	10 kb
Library numbers	30–300	24	96–384	96	384
Sequencing throughput	21–142 Gb	18 Gb	50–200 Gb	~100 Gb	150 Gb
Percentage of phased variants	94%–98%	97%	92%–95%	$>95\%$	99%
N50	0.3–1 Mb	2.6 Mb	60–700 kb	1.4–2.3 Mb	500 kb

chromosomes separation strategies. However, the distance and purity of linkage are not as remarkable as chromosomes separation strategies.

Fosmid clone pools

By Fosmid clone method, microscale DNA is amplified for NGS. Pure and intact genomic DNA fragments of about 35 Kb are separated in pools for haplotype analysis. For the relatively simple clone pipeline in comparison with BAC clone, fosmid clone was used to operate dilution pools strategy by several groups [33–38]. Kitzman *et al.* [33] constructed a single, complex fosmid library in 2011. Within each pool, ~5,000 fosmids with ~37 kb inserts were captured (~3% of the 6 Gb diploid genome). Different barcodes were applied to each of the 115 pools for barcoding libraries construction. After haplotype analysis, half of resolved sequences were within blocks of at least 350 kb (N50 of 350 kb).

BAC clone pools

Compared with fosmid clone, BAC clone has longer inserts of about 140 kb. Longer fragments are crucial for haplotype phasing and can reduce the quantities of required pools. In Lo *et al*'s work [3939], only 24 pools (5,000 clones per pool) were captured to construct indexed libraries. The N50 values of the assembled haplotype blocks were greater than 2.6 Mb.

Fragment pools with MDA

The advent of MDA [26] in 2002 provides uniform representation across the genome and convenient amplification of microscale DNA. The length of DNA fragments in pools is determined by the DNA extraction process. Longer fragments are preferred to generate longer haplotype blocks. Peters *et al.* [40] and Kaper *et al.* [41] captured fragment pools, amplified by MDA, and sequenced on Complete Genomics platform and Illimina platform, separately. Totally 384 fragment pools were captured by Peters *et al.* [40] with 10%–20% of a haploid genome in each pool, resulting 92% of the phasable heterozygous SNPs placed into long contigs with N50s of ~1 Mb and 500 kb for two samples, respectively. 8% of unphased variants were mainly caused by amplification bias.

Fragment pools with CPT-seq

To make sure genomic regions are overwhelmingly represented at most once, smaller pools are preferred in fragment pools with MDA strategy. However, smaller sizes means that more pools are required to represent the genome. Amini *et al.* [42] designed contiguity-preserving transposition sequencing (CPT-seq) to ameliorate this situation (Figure 3). Tn5 transposition was used to modify DNA with adaptor and index sequences. It introduced another dimension of barcodes to the libraries by tightly bounding to target DNA until compartmentalizing. Two dimensions of barcodes built 96×96 virtual compartments, but only 96 sequencing libraries were constructed actually.

Fragment pools with long-range PCR

In the previous study, MDA-based methods reported 8% of variants unphased at a high coverage [41]. Kuleshov *et al.* [43] took PCR instead of MDA as the amplification approach in order to reduce the amplification bias. After ligation with amplification adaptors, minute DNA is suited to be amplified by PCR. They diluted and placed DNA fragments into 384 wells, at about 3,000 fragments per well. Although the fragments were about 10 kb in length, 99% of single-nucleotide variants in three human genomes were phased into haplotype blocks 0.2–1Mb in length after the optimization of statistical pipeline. The unphased variants observably decreased.

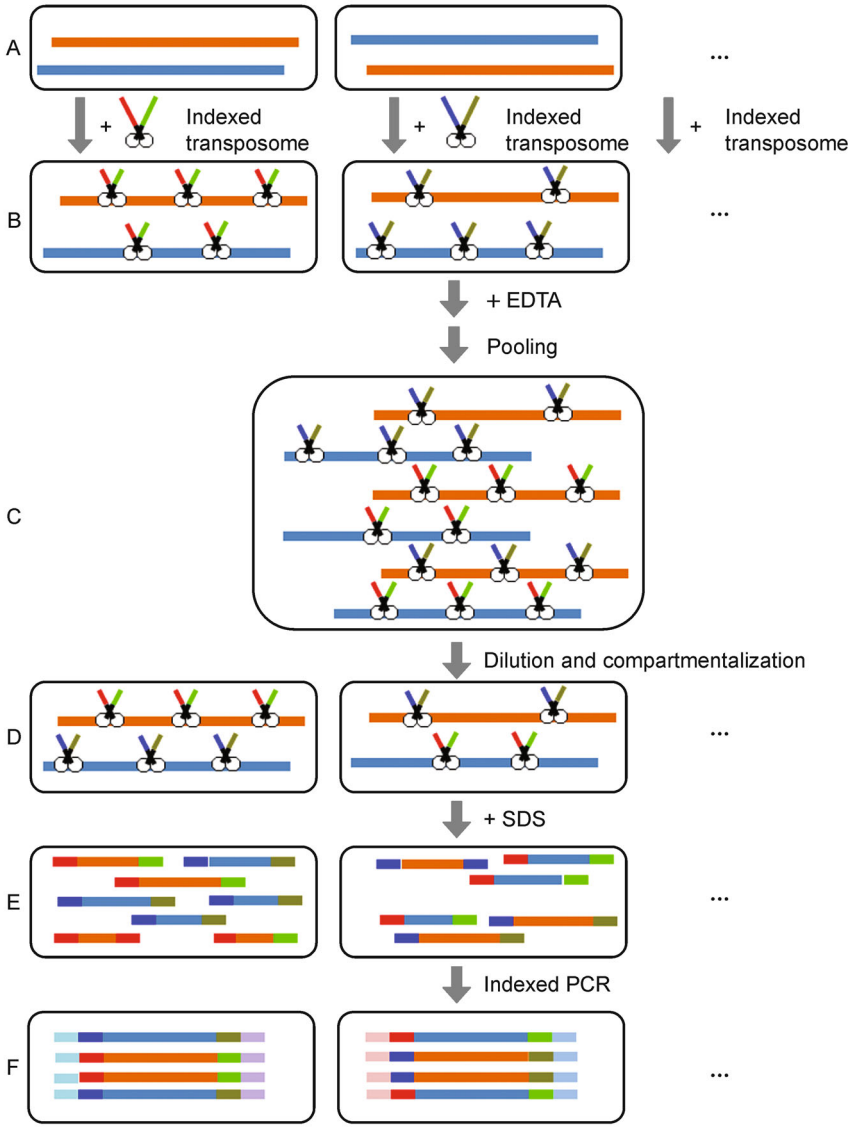

Figure 3. Overview of the contiguity-preserving transposition sequencing workflow. (A) Diploid genomic DNA is used to compartment into sub-haploid pools. (B) Series of different indexed transposome complexes are used to create independent transposition reactions. These reactions are the first level separate genomic partitions. (C) Transposition reactions are pooled together, (D) diluted to sub-haploid size and compartmented into second level partitions. (E) The transposases are removed by SDS. (F) Partition-specific barcodes are introduced to libraries by indexed PCR. All samples are pools together for further progressions.

CROSSLINKING AND PROXIMITY LIGATION

Both chromosomes separation and dilution pools strive to separate homologous chromosomes or fragments into different pools. Chromosomes or DNA fragments in each pool are approximately considered to be a haploid. An alternative strategy is to ligate two distant parts of a chromosome into a single sequencing reads. A series of these reads with random distance between the two parts provide different distance and accurate linkages. It is a tough work to gain a series of reads with random distance between two parts before the appearance of capturing chromosome conformation (3C) [44]. The 3C and coupling chromosome conformation capture-on-chip (4C) [45] were first developed to identify chromosomal interactions. The capability of grabbing two discontinuous sequences of one chromosome into one read or reads

part was considered by the successors [46–48] and was used for whole genome haplotyping. Although leaving many variants unphased [46], crosslinking and proximity ligation approach is a highly innovative strategy for haplotype analysis.

Crosslinking and proximity ligation

The chromatins are cross-linked in cell nucleus. Two cross-linked sequences are wide apart in sequence but nearby in space, and more importantly, linked. In the experiments [46,47], cross-linked chromatins were formaldehyde fixed, digested by restriction enzyme and ligated to form artificial fragments (Figure 4). After sequencing, the distance between the two cross-linked sequences ranged from several hundred base pairs to tens of millions of base pairs [46]. Selvaraj *et al.* [46] phased ~81% of alleles at 17× sequencing. After adjusting the progress, Vree *et al.* [47] applied similar method to selectively sequencing and phasing entire genes.

Cell free crosslinking and proximity ligation

The crosslinking and proximity ligation strategy relies on intact cells or nuclei. The signal seems to be confounding based on the complex and large-scale organization of chromosomes in nuclei. Some structures of different chromosomes, such as telomeres, are often associated in cells. To overcome the limitation, Putnam *et al.* [48] reconstituted chromatin *in vitro* to produce DNA linkages up to several hundred kilobases. They increased the haplotype blocks N50 from 508 kb to 10 Mb with the help of 210 million reads which were generated by cell free crosslinking and proximity ligation approach [48].

LONG-READ TECHNOLOGIES

The mismatch between short read length and long distance linkage requirement is the barrier of convenient whole genome haplotype analysis. Innovatory sequencing technologies have the potential to extend read length, which is the direct way to sweep this barrier. The reported innovatory sequencers include single-molecule real-time (SMRT) sequencing [49] and nanopore sequencing [50]. To be pointed, long-read technologies are not independent approaches but are able to join with the chromosomes separation, dilution pools, and crosslinking and proximity ligation strategies.

SMRT sequencing

SMRT sequencing platform produced raw reads median 2 kb in length, obviously longer than other sequencing platforms [51]. The low read accuracy (~85%) seems to prevent the method from being widely used in genome analysis, including whole genome haplotyping. However, with longer reads (average mapped reads for 5.8 kb in length) and optimized analysis process, Chaisson *et al.* [49] analyzed a haploid human genome, closing or extending 55% of the remaining gaps. In spite of this, low accuracy will bother SMRT sequencing in haplotyping analysis before significant improvement.

Nanopore sequencing

Nanopore platforms sequence nucleotides by the change of electric current while a DNA molecule passes through the nanoscale pore, which are promised to generate long reads. However, single nucleotide can not be correct identified all the time. Recent study [50] reported the identification of four-nucleotide combinations and drew the quadromer map of sequences up to 4,500 bases in length. It is worthwhile expecting accurate and high-throughput nanopore sequencing methods.

DISCUSSION AND CONCLUSION

For all the experimental haplotyping strategies, obtaining long distance, highly complex and accurate linkages are the core goals (Table 2). The system complexity and

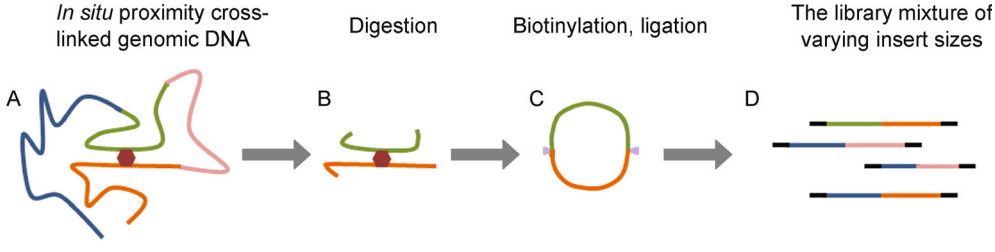

Figure 4. Proximity-ligation experiment pipeline. (A) Chromatins are *in situ* proximity cross-linked. (B) Cross-linked DNA is digested with a restriction enzyme. (C) Digested DNA is ligated to form artificial circles. (D) DNA fragments from two distant genomic loci which looped together in three-dimensional space *in vivo* are captured. A mixture of varying insert sizes is used for haplotype analysis.

Table 2. Comparison of the linkage obtaining efficiency of the four experimental methods.

	Chromosomes separation	Dilution pools	Crosslinking and proximity ligation	Long-read technologies
Linkage distance	Up to whole chromosome	Up to 200 kb by BAC pools	Majority < 1 kb, with tail up to 30 Mb	2,500 bp by SMRT sequencing
Linkage complexity	Complex	Complex within blocks	Complex within 1 kb	Complex within read length
Linkage accuracy	Very high	High	Median	Very high

Table 3. Requirements of the four experimental methods.

	Chromosomes separation	Dilution pools	Crosslinking and proximity ligation	Long-read technologies
DNA input	Several cells	At least 150 ng	200–800 ng	Depending on the library preparation
Required Sequencing depth	Median	High	Median	Low (depending on the read length)
System complexity	Very high	High	Median	Low
Work flow scalability	Not scalable	Scalable	Scalable	Scalable

requirements are also determinate factors about whether the haplotyping method is well applied or not (Table 3). An efficient, low cost, scalable and labor-saving work flow is desired. Chromosomes separation strategies obtain long, complex and accurate linkages directly. But the work flows are complex and not scalable, while precise devices and metaphase cells are required. More importantly, the key steps of the experiment are not largely controllable. As a special case, the haplotyping of human gametes is not applicable to other samples. Therefore, chromosomes separation strategies are only suitable for special applications and professional labs. Dilution pools do not rely on specific devices and samples and is soft for most labs in almost every haplotyping study. However, it is labor-intensive to construct numerous sequencing libraries. Smaller scale of each pool can generate more accurate linkages, but will lead to constructions of more libraries. At the same time, smaller pools mean high sequencing depth in all. CPT-seq requires less sequencing libraries at the same pool scale, but is unable to clinch this contradiction thoroughly. Crosslinking and proximity ligation approaches open a new window in solid linkages obtaining. Although the crosslinking and ligation work flow seems to be complex, it only requires one sequencing library, which appears a great advantage. More importantly, crosslinking and proximity ligation approaches give an alternative direction apart from homologous chromosomes or fragments separation. Lower sequencing depth is another advantage. Slightly poor result in allele phasing limits crosslinking and proximity ligation approaches to be widely applied nowadays. More studies are required for these methods to generate more comprehensive haplotypes. To generate reads longer enough is expected to be the final solution of haplotyping. Before this, technical renovations in read length are helpful in promoting the performance of all the strategies above.

In the next several years, with the help of long read-length and exact bioinformatic pipelines, experimental strategies will phase the whole genome efficiently and comprehensively. Innovative experimental strategies are expected to have high-quality performance, low cost and be labor-saving, which will be largely desired in the future.

ACKNOWLEDGEMENTS

This work was supported by the National Basic Research Program of China (No. 2012CB316501), and the National Natural Science Foundation of China (Nos. 61227803 and 61571121).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Jing Tu, Na Lu, Mengqin Duan, An Ju, Xiao Sun and Zuhong Lu declare they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. and Schork, N. J. (2011) The importance of phase information for human genomics. Nat. Rev. Genet., 12, 215–223
2. Muers, M. (2011) Genomics: No half measures for haplotypes. Nat. Rev. Genet., 12, 77

3. Tian, Q., Price, N. D. and Hood, L. (2012) Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (P4) medicine. J. Intern. Med., 271, 111–121
4. Levenstien, M. A., Ott, J. and Gordon, D. (2006) Are molecular haplotypes worth the time and expense? A cost-effective method for applying molecular haplotypes. PLoS Genet., 2, e127
5. Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H. Y., *et al.* (2010) A draft sequence of the Neandertal genome. Science, 328, 710–722
6. Tjio, J. H. (1978) The chromosome number of man. Am. J. Obstet. Gynecol., 130, 723–724
7. Lejeune, J. and Turpin, R. (1961) Chromosomal aberrations in man. Am. J. Hum. Genet., 13, 175–184
8. Caspersson, T., Zech, L., Johansson, C. and Modest, E. J. (1970) Identification of human chromosomes by DNA-binding fluorescent agents. Chromosoma, 30, 215–227
9. Fodor, S. P. A., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T. and Solas, D. (1991) Light-directed, spatially addressable parallel chemical synthesis. Science, 251, 767–773
10. Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W. L., Chen, C., Zhai, Y., *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat. Genet., 20, 207–211
11. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.*, (2001) Initial sequencing and analysis of the human genome. Nature, 409, 860–921
12. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001) The sequence of the human genome. Science, 291, 1304–1351
13. Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., *et al.* (2007) The diploid genome sequence of an individual human. PLoS Biol., 5, e254
14. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437, 376–380
15. Bentley, D. R. (2006) Whole-genome re-sequencing. Curr. Opin. Genet. Dev., 16, 545–552
16. Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D. and Church, G. M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science, 309, 1728–1732
17. Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature, 475, 348–352
18. Bayley, H. (2006) Sequencing single molecules of DNA. Curr. Opin. Chem. Biol., 10, 628–637
19. McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res., 19, 1527–1541
20. Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., McVean, G. A., Donnelly, P., Egholm, M., *et al.*, (2010) A map of human genome variation from population-scale sequencing. Nature, 467, 1061–1073
21. The 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature, 491, 56–65
22. Delaneau, O., Marchini, J. and Zagury, J. F. (2012) A linear complexity phasing method for thousands of genomes. Nat. Methods, 9, 179–181
23. Lu, S., Zong, C., Fan, W., Yang, M., Li, J., Chapman, A. R., Zhu, P., Hu, X., Xu, L., Yan, L., *et al.* (2012) Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. Science, 338, 1627–1630
24. Kirkness, E. F., Grindberg, R. V., Yee-Greenbaum, J., Marshall, C. R., Scherer, S. W., Lasken, R. S. and Venter, J. C. (2013) Sequencing of isolated sperm cells for direct haplotyping of a human genome. Genome Res., 23, 826–832
25. Wang, J., Fan, H. C., Behr, B. and Quake, S. R. (2012) Genome-wide single-cell analysis of recombination activity and *de novo* mutation rates in human sperm. Cell, 150, 402–412
26. Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., *et al.* (2002) Comprehensive human genome amplification using multiple displacement amplification. Proc. Natl. Acad. Sci. USA, 99, 5261–5266
27. Hou, Y., Fan, W., Yan, L., Li, R., Lian, Y., Huang, J., Li, J., Xu, L., Tang, F., Xie, X. S., *et al.* (2013) Genome analyses of single human oocytes. Cell, 155, 1492–1506
28. Ma, L., Xiao, Y., Huang, H., Wang, Q., Rao, W., Feng, Y., Zhang, K. and Song, Q. (2010) Direct determination of molecular haplotypes by chromosome microdissection. Nat. Methods, 7, 299–301
29. Yang, H., Chen, X. and Wong, W. H. (2011) Completely phased genome sequencing through chromosome sorting. Proc. Natl. Acad. Sci. USA, 108, 12–17
30. Fan, H. C., Wang, J., Potanina, A. and Quake, S. R. (2011) Whole-genome molecular haplotyping of single cells. Nat. Biotechnol., 29, 51–57
31. Dear, P. H. and Cook, P. R. (1989) Happy mapping: a proposal for linkage mapping the human genome. Nucleic Acids Res., 17, 6795–6807
32. Burgtorf, C., Kepper, P., Hoehe, M., Schmitt, C., Reinhardt, R., Lehrach, H. and Sauer, S. (2003) Clone-based systematic haplotyping (CSH): a procedure for physical haplotyping of whole genomes. Genome Res., 13, 2717–2724
33. Kitzman, J. O., MacKenzie, A. P., Adey, A., Hiatt, J. B., Patwardhan, R. P., Sudmant, P. H., Ng, S. B., Alkan, C., Qiu, R. L., Eichler, E. E., *et al.* (2011) Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nat. Biotechnol., 29, 59–63.
34. Suk, E. K., McEwen, G. K., Duitama, J., Nowick, K., Schulz, S., Palczewski, S., Schreiber, S., Holloway, D. T., McLaughlin, S., Peckham, H., *et al.* (2011) A comprehensively molecular haplotype-resolved genome of a European individual. Genome Res., 21, 1672–1685
35. Duitama, J., McEwen, G. K., Huebsch, T., Palczewski, S., Schulz, S., Verstrepen, K., Suk, E. K. and Hoehe, M. R. (2012) Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. Nucleic Acids Res., 40, 2041–2053
36. Adey, A., Burton, J. N., Kitzman, J. O., Hiatt, J. B., Lewis, A. P., Martin, B. K., Qiu, R., Lee, C. and Shendure, J. (2013) The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. Nature, 500, 207–211
37. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., *et al.* (2014) The complete genome sequence of a Neanderthal from the Altai

Mountains. Nature, 505, 43–49

38. Hoehe, M. R., Church, G. M., Lehrach, H., Kroslak, T., Palczewski, S., Nowick, K., Schulz, S., Suk, E. K. and Huebsch, T. (2014) Multiple haplotype-resolved genomes reveal population patterns of gene and protein diplotypes. Nat. Commun., 5, 5569
39. Lo, C., Liu, R., Lee, J., Robasky, K., Byrne, S., Lucchesi, C., Aach, J., Church, G., Bafna, V. and Zhang, K. (2013) On the design of clone-based haplotyping. Genome Biol., 14, R100
40. Peters, B. A., Kermani, B. G., Sparks, A. B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y. T., Haas, J.,*et al.* (2012) Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature, 487, 190–195
41. Kaper, F., Swamy, S., Klotzle, B., Munchel, S., Cottrell, J., Bibikova, M., Chuang, H. Y., Kruglyak, S., Ronaghi, M., Eberle, M. A., *et al.* (2013) Whole-genome haplotyping by dilution, amplification, and sequencing. Proc. Natl. Acad. Sci. USA, 110, 5552–5557
42. Amini, S., Pushkarev, D., Christiansen, L., Kostem, E., Royce, T., Turk, C., Pignatelli, N., Adey, A., Kitzman, J. O., Vijayan, K., *et al.* (2014) Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. Nat. Genet., 46, 1343–1349
43. Kuleshov, V., Xie, D., Chen, R., Pushkarev, D., Ma, Z., Blauwkamp, T., Kertesz, M. and Snyder, M. (2014) Whole-genome haplotyping using long reads and statistical methods. Nat. Biotechnol., 32, 261–266
44. Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002) Capturing chromosome conformation. Science, 295, 1306–1311
45. Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., Shendure, J., Fields, S., Blau, C. A. and Noble, W. S. (2010) A three-dimensional model of the yeast genome. Nature, 465, 363–367
46. Selvaraj, S., R Dixon, J., Bansal, V. and Ren, B. (2013) Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat. Biotechnol., 31, 1111–1118
47. de Vree, P. J. P., de Wit, E., Yilmaz, M., van de Heijning, M., Klous, P., Verstegen, M. J. A. M., Wan, Y., Teunissen, H., Krijger, P. H. L., Geeven, G., *et al.*(2014) Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. Nat. Biotechnol., 32, 1019–1025
48. Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., Troll, C. J., Fields, A., Hartley, P. D., Sugnet, C. W., *et al.* (2016). Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. Genome Res. 26, 342–350
49. Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., *et al.* (2015) Resolving the complexity of the human genome using single-molecule sequencing. Nature, 517, 608–611
50. Laszlo, A. H., Derrington, I. M., Ross, B. C., Brinkerhoff, H., Adey, A., Nova, I. C., Craig, J. M., Langford, K. W., Samson, J. M., Daza, R., *et al.* (2014) Decoding long nanopore sequencing reads of natural DNA. Nat. Biotechnol., 32, 829–833
51. Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., *et al.* (2012). Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. Nat Biotechnol. 30, 693–700