Direct determination of molecular haplotypes by chromosome microdissection

Li Ma[1], **Yan Xiao**[1], **Hui Huang**[2], **Qingwei Wang**[1], **Weinian Rao**[1], **Yue Feng**[3], **Kui Zhang**[4], and **Qing Song**[1]

[1] Cardiovascular Research Institute, Morehouse School of Medicine, Atlanta, Georgia, USA.

[2] Clinical Cytogenetics Laboratory, Department of Pathology, Vanderbilt University Medical Center, Nashville, TN 37204, USA.

[3] Department of Pharmacology, Emory University School of Medicine, Atlanta, GA 30322, USA.

[4] Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, USA.

Abstract

Direct observation of haplotypes is still technical challenging. Here we report a method for the determination of haplotypes through chromosome microdissection. We determine human haplotypes with more than 98.85% accuracy at 24,245 heterozygous single-nucleotide polymorphism (SNP) loci in genome-wide chromosome-range phasing distance.

Haplotype refers to a group of alleles inherited on a single chromosome. Haplotype analysis plays an important role for mapping disease genes, elucidating population histories, studying the evolutionary genetics, and exploring *cis*-interactions in the regulation of gene expression [1],[2]. Statistical and computational methods have been developed to reconstruct haplotypes from conventional genotype data. Although statistical and computational methods can provide highly accurate estimates of haplotypes in most situations, even the best method can have a large error rate of more than 35% for some real datasets [3]. The accuracy of statistical methods for haplotype reconstruction can be greatly improved when they are applied to trio datasets using information from relatives [2]; however, the parental samples are not always available to make this feasible. Experimental approaches have been developed to determine haplotypes [4],[5] with limited capability of haplotyping SNPs across a large chromosomal region spanning megabase distances, except the conversion technology, which can determine the long-range haplotypes by creating human-mouse hybrid cells that contain a

subset of human chromosomes [6,7], and the polony (polymerase colony) technology, which can determine the long-range haplotype of a pair of distant SNP loci on the same immobilized colony within an acrylamide gel [8,9]. Conversion technology involves considerable cost and time for large-scale applications, the polony technology is limited by the small number of SNPs that can be haplotyped in genome-wide studies [2,10].

Here we report, 7 (poly) Dimensional DNA (7DDNA), a method built upon the concept that a somatic cell of a diploid organism contains two and only two sets of homologous chromosomes. If we collect ~ 23 chromosomes from one cell instead of its entire 46 chromosomes, the collection may contain only one copy of some chromosomes, it may also contain no copy or both copies of other chromosomes (Supplementary Fig. 1).We genotype the 7DDNA sample and an entire genomic sample in parallel; the latter allows us to determine which SNPs are heterozygous or homozygous in each sample. . If a 7DDNA sample shows "homozygous" genotypes at all polymorphic loci along a chromosome known to be heterozygous, this 7DDNA sample contains one single copy of this chromosome (Fig. 1, Supplementary Fig. 1 and Supplementary Table 1), which will enable us to read one haplotype of this chromosome directly from its conventional high-throughput genotyping output. The unobserved dropout alleles will constitute the other haplotype of this chromosome. If a 7DDNA sample shows heterozygous genotypes at loci known to be polymorphic, this harvest contains both chromosome copies. If a 7DDNA sample fails to call any SNPs along a chromosome, this chromosome is absent in this sample.

We have performed a proof-of-principle study with a HapMap sample [11], GM10847, since the haplotypes can be computationally reconstructed with the parental information[2,12]. These computationally configured haplotypes should be accurate on those non-triple-heterozygote SNP loci unless there is mis-inheritance, genotyping errors, SNP recurrence, or an overlapping of a SNP with an insertion-deletion polymorphism (indel), a repeat, or a copy number variation (CNV), or segmental duplications.

In our experiment, we microdissected three pieces from the chromosome spread of an individual (Online Methods). The microdissection harvests were subjected to whole genome amplification (WGA), which provided us with about 5-8 μg DNA from each harvest. After genotyping with an Illumina BeadChip (CNV370), we collectively haplotyped 15 chromosomes in these three harvested samples, among which eight chromosomes were covered by one harvest, five chromosomes were covered by two harvests, and one chromosome was covered by all three harvests (Supplementary Table 2). Totally 24,245 heterozygous SNPs on these 15 chromosomes were haplotyped (Supplementary Table 2), among which 2,089 SNPs were phased by more than one microdissection harvests, 98.85% of these SNPs (2,065 SNPs) showed consistent results between different microdissection harvests (Table 1). Our data showed a faithful amplification of our microdissection harvests, in which the error rate was 0.73% (Supplementary Table 3). We noticed that the no-call rate in the Illumina genotyping on haploid chromosomes of the microdissection harvests was as high as ~70% (Supplementary Table 4). The possible reasons may be that our WGA method was insufficient to amplify every locus when the quantity of input DNA was as small as one single copy, or this genotyping platform was insensitive to the DNA template amplified from one single molecule of template. We anticipate that this issue may be solved by a new

robust WGA method for amplifying single-copy template or other genotyping platforms such as next-generation sequencers.

Compared with computationally reconstructed haplotypes of this person (GM10847) on those non-triple-heterozygous SNP loci, our 7DDNA haplotypes reported discordant allele phases on only 0.65% of SNPs (Supplementary Table 2). Comparing our experimentally determined haplotypes of GM10847 with her parents' haplotypes downloaded from the HapMap database that were computationally configured by IMPUTE++ [13], the difference was only 2.15%, among which 58.7% of those discrepancies occurred on triple-heterozygote loci. This data not only demonstrated a high accuracy of haplotypes configured with trio information [2], it also validated our 7DDNA haplotype results.

About 10.8% of heterozygous SNPs on the diploid chromosomes in our microdissection harvests remained to be heterozygous (Supplementary Table 5); the other heterozygous SNPs on those diploid chromosomes received a homozygous genotype call due to allele dropout. However, on the haploid chromosomes in our microdissection harvests, only 0.08% of heterozygous SNPs appeared to be heterozygous (Supplementary Table 5). This substantial disparity (10.8% vs. 0.08%) enabled us to distinguish those haploid chromosomes from diploid chromosomes in a microdissection harvest (Supplementary Fig. 2 and Supplementary Fig. 3).

To haplotype all chromosomes of each individual, multiple microdissection harvests should be collected to cover all of a person's chromosomes. Our statistical analysis (Supplementary Note-1) showed that 4-8 microdissection samplings of each individual will provide a 93% probability to obtain a person's whole genome haplotype (Supplementary Table 6 and Supplementary Fig. 4). In other words, 52% of individuals may have completed their genome haplotyping with no more than five microdissections, 93% of individuals will need no more than eight microdissections to complete their genome haplotyping (Supplementary Table 6 and Supplementary Fig. 4). When we reach 12 microdissection samplings, 99.6% of all individuals will have their whole genome haplotyped. In addition to chromosome coverage, this cumulative strategy will also effectively increase the SNP coverage on resolved haplotypes (Supplementary Table 7). We have estimated the cost of this strategy to be $0.0023 per SNP per sample (Supplementary Note-2). The entire 7DDNA procedure will take three days, with a ~10-min work on the first day, no work on the second day, and a 6-8 hour work on the third day. Based on this schedule, we may do five experiments each week, and handle 24 individuals in each experiment. A 1,000-sample project may need only 2-6 months.

Because many researchers may be interested in particular chromosomes rather than the entire genome, such as in human leukocyte antigen (HLA) studies, we have explored an alternative approach in which a single chromosome is collected in a microdissection with an Objective HCX 150X/0.90 lens. Chromosome recognition can be achieved by karyotype morphology in combination with PCR validation.

It is possible that a chromosome is broken by the microdissection procedure. When only a partial chromosome is harvested without accompaniment of its sister chromosome, there

will be failed allele calls at those SNP loci located on the absent chromosomal segment. If a partial chromosome is harvested accompanied by its full-length sister chromosome in the microdissection, the heterozygous SNPs located on this partial chromosomal segment will receive heterozygous genotypes in the 7DDNA genotyping results, but those heterozygous SNPs located outside this partial chromosome piece will appear as "homozygous" loci in the 7DDNA genotyping results (Supplementary Fig. 2). The segmental boundaries will be deciphered from the genotyping results and the haplotypes of entire chromosome will be assembled from broken pieces in different microdissection harvests by their overlapping portion.

The 7DDNA technology provides the first high-throughput experimental method for determination of long-range haplotypes for large-scale and genome-wide studies. Compared with conversion technology [6,7,14] and polony haplotyping [8,9], 7DDNA exhibits advantages on labor, time, throughput and the number of SNPs. Chromosomal haplotypes will be important for functional interpretations of SNPs in cis-interactions and the integration of genetic and epigenetic datasets along the chromosomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

ONLINE METHODS

Chromosome Collection

About 0.3 ml of whole blood (anticoagulated by sodium heparin) is collected into a conical 15-ml centrifuge tube containing 5 ml of pre-warmed complete PB Max Karyotyping Media in which fetal bovine serum, L-glutamine, phytohemagglutinin (PHA), and gentamicin are added. After 45-hour incubation at 37°C, 50 ul of colcemid is added into the specimen followed by incubation at 37°C for 30 min to arrest the dividing cells in metaphase [15]. The specimen is then centrifuged at 1,000 rpm for 10 min; the cell pellet is resuspended into pre-warmed 0.075mol/L KCl and incubated at room temperature for 15 min. After cells are fixed with cold fixative (methanol:acetic acid, 3:1), they are dripped onto a UV-sliceable foiled slide (Vashaw Scientific, Cat No. 505-151) to spread chromosomes. After a brief Giemsa-staining for 10 min, a cutting area containing ~23 chromosomes of one single cell is selected under a Laser Microdissection Microscope (ASLMD, Leica, Germany). This area of slide foil is microdissected by a computer-directed laser beam and harvested into a collecting eppendorf tube. The collected foil is directly used in subsequent experiments.

If cells such as lymphoblastoid cell lines are used instead of whole blood specimen, cells are cultured in medium (RPMI1640 for lymphoblastoid cell lines) containing 15% FBS and a mitogen (such as PHA) for 45 hours, followed by the same experimental steps from colcemid treatment to chromosome microdissection as described above.

Whole Genome Amplification (WGA)

Chromosomes on the harvested slide foil are amplified with the Sigma GenomePlex WGA4 kit following the manufacturer's protocol. Briefly, the sample is incubated in the Lysis and Fragment Buffer at 50°C for 1 hour, and then heated to 99 °C for 4 min. Then the Single Cell Library Preparation Buffer and Library Stabilization Solution are added into the sample followed by incubation at 95°C for 2 min. Library is prepared with the following cycles: 16°C for 20 min, 24°C for 20 min, 37°C for 20 min, and 75°C for 5 min. DNA is amplified by an initial denaturation at 95 °C for 3 min followed by 35 cycles of 94°C/30 sec and 65°C/5 min. Amplified DNA is purified by QIAquick PCR purification kit prior to genotyping.

Genotyping

Amplified DNA is subjected to a whole-genome genotyping and CNV analysis with an Illumina Infinium HD BeadChip. In our proof-of-principle study, we used the HumanCNV370Quad BeadChip, which contains 351,507 SNPs and 21,890 CNV markers. Meanwhile, genomic DNA for each individual is genotyped as well. Allele calls are made by the Illumina GenomeStudio Software package.

Interpretation of Genotyping data - Haplotype Readout

The allele calls from microdissected DNA are compared with the allele calls from the same individual's genomic DNA at each locus. If a microdissected DNA sample receives a homozygous allele call at every heterozygous locus (known from the genomic DNA results) along a chromosome (Supplementary Table 1), these observed allele calls from microdissected DNA constitute one haplotype of this chromosome, and all of the opposite alleles at these loci constitute the other chromosomal haplotype of this person.

Data Analysis in this Proof-of-Principle Study

The genotypes were retrieved from the International HapMap Project database (Phase 2 Public Rel#22, Phase 3 Public Draft Rel#1, and Phase 2+3 Feb09 Rel#27) and the Illumina database (ftp://ftp.illumina.com). Besides our experimental determination, the haplotypes of GM10847 were also computationally configured by utilizing the trio structure of HapMap samples following the principle described by Hodge et al., (1999) [12]. HapMap2 genotypes of GM10847 and her parents, GM12146 and GM12239, were used in our computational haplotype configuration. This computationally resolved haplotypes were compared with our experimentally determined haplotypes. A genotype error was called if there was a consistent discordance on allele calls between the HapMap2 dataset and other datasets (the HapMap3 dataset, the Illumina dataset, and our genotyping data with the genomic DNA). Repeating element annotation dataset was retrieved from the UCSC Genome Browser (Human 2006

March Assembly) created by using Arian Smit's RepeatMasker program [16]-[17]. All data integration was performed with SAS9.1.

References

1. Levenstien MA, Ott J, Gordon D. PLoS Genet. 2006; 2:e127. [PubMed: 16933998]

2. Marchini J, et al. Am J Hum Genet. 2006; 78:437–450. [PubMed: 16465620]

3. Stephens M, Donnelly P. Am J Hum Genet. 2003; 73:1162–1169. [PubMed: 14574645]

4. Ragoussis J. Annu Rev Genomics Hum Genet. 2009; 10:117–133. [PubMed: 19453250]

5. Xiao M, et al. Nat Methods. 2009; 6:199–201. [PubMed: 19198595]

6. Yan H, et al. Nature. 2000; 403:723–724. [PubMed: 10693791]

7. Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB. Nat Genet. 2001; 28:361–364. [PubMed: 11443299]

8. Mitra RD, et al. Proc Natl Acad Sci U S A. 2003; 100:5926–5931. [PubMed: 12730373]

9. Zhang K, et al. Nat Genet. 2006; 38:382–387. [PubMed: 16493423]

10. Liu N, Zhang K, Zhao H. Adv Genet. 2008; 60:335–405. [PubMed: 18358327]

11. The International HapMap Consortium. Nature. 2003; 426:789–796. [PubMed: 14685227]

12. Hodge SE, Boehnke M, Spence MA. Nat Genet. 1999; 21:360–361. [PubMed: 10192383]

13. Howie BN, Donnelly P, Marchini J. Talk in ASHG 2008 Philadelphia PA. 2008

14. Highsmith WE Jr, Meyer KJ, Marley VM, Jenkins RB. Curr Protoc Hum Genet Chapter. 2007; 3 Unit 3.6.

15. Tjio JH, Levan A. Hereditas. 1956; 42:1–6.

16. Smit AFA, Hubley R, Green P. http://www.repeatmasker.org

17. Jurka J. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. 2000; 16:418–420. [PubMed: 10973072]

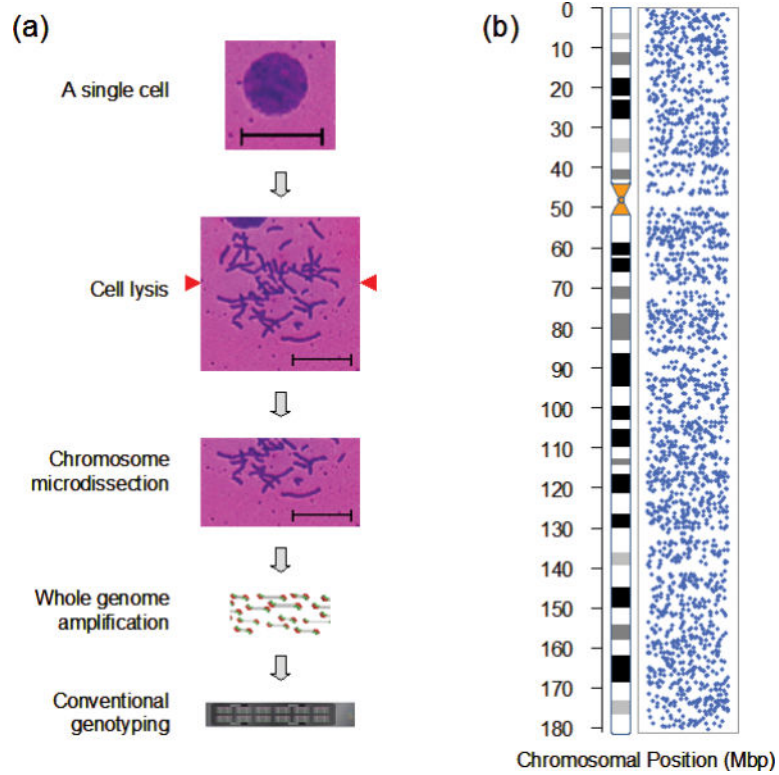**Figure 1.**
The principle and results on chromosome 5 of the 7DDNA haplotyping method. (**a**) 7DDNA haplotyping procedure. (**b**) haplotyped chromosome 5, on which 2,514 SNPs were resolved. Each dot represents a SNP. Scale bars, 10 μM.

Table 1

Replication of the 7DDNA haplotyping results.

Chromosome	Repeated SNPs[*]	Consistent SNPs[**]	Consistency rate %
Chr3	517	516	99.81
Chr4	565	554	98.05
Chr16	189	186	98.41
Chr17	217	215	99.08
Chr18	212	208	98.11
Chr20	389	386	99.23
Total	2,089	2,065	98.85

[*] Repeated SNPs are those SNPs that were haplotyped in 2 or 3 microdissection harvests in our study.

[**] Consistent SNPs are those repeated SNPs that obtained the identical results on haplotypes from different microdissection harvests.