Published in final edited form as:

Nat Methods. 2009 March ; 6(3): . doi:10.1038/nmeth.1301.

Direct determination of haplotypes from single DNA molecules

Ming Xiao^{1,5}, Eunice Wan¹, Catherine Chu¹, Wen-Chi Hsueh^{2,3}, Yang Cao¹, and Pui-Yan Kwok^{1,3,4}

¹Cardiovascular Research Institute, University of California, San Francisco, 513 Parnassus Ave., San Francisco, California 94143, USA.

²Department of Medicine, University of California, San Francisco, 513 Parnassus Ave., San Francisco, California 94143, USA.

³Institute for Human Genetics, University of California, San Francisco, 513 Parnassus Ave., San Francisco, California 94143, USA.

⁴Department of Dermatology, University of California, San Francisco, 513 Parnassus Ave., San Francisco, California 94143, USA.

Abstract

Determining the long-range haplotypes in a diploid individual is a major technical challenge. Here we report a method of molecular haplotyping by directly imaging multiple polymorphic sites on individual human DNA molecules simultaneously. We demonstrate the utility of this technology by accurately determining the haplotypes consisting of up to 16 single-nucleotide polymorphisms in genomic regions up to 50 kilobases.

The recent successful completion of the HapMap project¹ and the availability of low-cost, large-scale, high-throughput single-nucleotide polymorphism (SNP) genotyping platforms have paved the way for comprehensive genetic analysis of complex traits such as susceptibility to common diseases and drug responses through genome-wide association studies². Haplotype analysis can enhance the statistical power in the mapping of human complex trait loci, with the potential of reducing the size of genome-wide association studies³. So far, the methods for determining haplotypes have considerable limitations that have prevented their use in large-scale and high-throughput genetic studies⁴⁻⁷.

We had recently reported a DNA bar-coding method based on direct imaging of multiple polymorphic or sequence-motif sites on individual DNA molecules simultaneously⁸⁻¹⁰. Here we report a new molecular haplotyping approach, which improves the labeling efficiency and allows one to experimentally determine haplotypes of targeted genomic regions.

© 2009 Nature America, Inc. All rights reserved.

Correspondence should be addressed to M.X. (mingx@bionanomatrix.com) or P.-Y.K. (pui.kwok@ucsf.edu).

⁵Present address: BioNanomatrix, Inc., 3701 Market Street, Philadelphia, Pennsylvania 19104, USA.

Note: Supplementary information is available on the Nature Methods website.

AUTHOR CONTRIBUTIONS

M.X. designed and performed experiments, analyzed data and wrote the paper. E.W. and C.C. performed experiments and analyzed the data. Y.C. performed experiments. W.-C.H. performed statistical analysis. P.Y.K. designed the study and wrote the paper.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at <http://www.nature.com/naturemethods/>.

Our method starts with long-range PCR amplification of target DNA segments containing the polymorphic sites followed by allele-specific labeling of polymorphic alleles with fluorescent dye molecules. The labeled double-stranded DNA (dsDNA) molecules are then stretched into linear form on a glass surface and imaged using multicolor total internal reflection fluorescence microscopy, a technique capable of localizing single fluorescent dye molecules with nanometer-scale accuracy¹¹. By determining the order of the fluorescent labels on the backbone at all the polymorphic sites of fully labeled dsDNA fragments, the haplotype(s) of an individual can be inferred with great accuracy, in a manner similar to reading a bar code.

Highly efficient allele-specific labeling of the markers on the long-range PCR products for single-molecule analysis is the most critical aspect of our haplotyping method. To increase the labeling efficiency, we generated single-stranded DNA (ssDNA) fragments from double-stranded long-range PCR products. We accomplished this by modifying one PCR primer with a 5' Cy3 label and the other with 5' phosphorylation in a PCR. When we incubated the resulting PCR product with lambda exonuclease, the DNA strand with the phosphorylated 5' end was digested efficiently whereas the one with Cy3-modified 5' end was protected from digestion and became the ssDNA template. We also used the end Cy3 label to orient the DNA molecules. Using this approach, we generated single ssDNA molecules up to 18 kilobases (kb) routinely. With ssDNA templates, the allele-specific extension is used to tag the alleles of each SNP (Fig. 1), in which two oligonucleotides labeled with different-colored fluorophores are designed such that they differ only by the polymorphic base at the 3' end. Only the perfectly matched primers will be extended by DNA polymerase and will stay on the DNA template. We performed 30 cycles of hybridization and linear extension (with conditions designed to denature only the unextended probes) to drive the hybridization and extension of the allele-specific probes with correct alleles into completion (Supplementary Methods online). At the end of the labeling procedure, virtually all the ssDNA fragments were converted to dsDNA fragments. We improved the specificity of the extension reaction by substituting the second to the last base at the 3' end of the labeled probe with a locked nucleic acid¹².

We tested the labeling efficiency and allele specificity of this labeling scheme by determining the haplotypes of a 12.5 kb fragment on human chromosome 17, which is linked to the susceptibility of the skin disease psoriasis¹³. We amplified the DNA segment from the HapMap Centre d'Etude du Polymorphisme Humain DNA sample (NA12156) containing four SNP markers by long-range PCR (Fig. 2a). We labeled each allele of a SNP with Cy3 or Cy5 fluorescent probes (Supplementary Table 1 online) and stained the DNA backbone with the fluorescent dye YOYO-1. We used only fully stretched DNA molecules with all four SNPs labeled to call the haplotypes (Fig. 2b and Supplementary Table 2 online). For this sample, which is homozygous for all four SNPs, we found 31 fully labeled fragments among 500 DNA fragments examined. Of the fully labeled fragments, 25 had the same bar code and 6 had six other unique bar codes. By analyzing only the fully labeled dsDNA fragments, there was no ambiguity in the identity of the SNPs. Assuming that the labeling efficiency is independent at all sites, this suggests approximately 50% labeling efficiency at each site.

We haplotyped DNA samples from 8 individuals in this manner and confirmed the haplotypes by genotyping both parents of each individual (data not shown). Among the 8 samples, we found two other sets of haplotypes. For example, sample NA10860 was homozygous for the first SNP but heterozygous for the remaining three SNPs whereas DNA sample NA07357 was heterozygous for all four SNPs and had different haplotypes (Fig. 2c,d and Supplementary Table 2).

We determined the accuracy of the haplotype determination for each fragment by χ^2 tests. We grouped together the rarely observed haplotypes and compared them to the one (homozygote) or two (heterozygote) most commonly observed haplotypes. The certainty of haplotype calling for results shown in Figure 2b–c was significant ($P < 0.00001$; details of statistical analyses are available in Supplementary Methods).

The haplotype information across a greater distance can be obtained by using overlapping long-range PCR fragments. We haplotyped a 20 kb region containing seven SNPs with two overlapping long-range PCR fragments (Supplementary Fig. 1 and Supplementary Table 3 online) and a 50 kb region containing 16 heterozygous SNPs with five overlapping long-range PCR fragments. The 50 kb region of chromosome 8 contains the *PLAT* gene that is associated with ischemic stroke¹⁴. We used five PCR fragments of various lengths containing different numbers of SNPs (ranging from the 11 kb fragment 4 with 4 SNPs to the 15 kb fragment with 6 SNPs; Fig. 3a). We designed long-range PCR assays such that one or two SNPs were present in the overlapping regions of adjacent fragments (Supplementary Table 4 online). We used these SNPs to connect the haplotype from individual fragments to form a continuous haplotype. For example, we used SNP-10 (rs2070713) in the overlapping region of fragment 1 and fragment 2 to connect the two fragments. We haplotyped sample NA12865, which is heterozygous at all polymorphic sites, using the above tiling strategy (Fig. 3). Owing to the optical resolution limit of 250 nm, the SNPs separated by less than 800 bp in fragment 1 were typed in two groups and connected by a common SNP. In this case, the two groups of SNPs are also connected by SNP-10 (Fig. 3b). For fragments 4 and 5, we intentionally designed the allele-specific probes to interrogate the opposite strand to that typed for fragments 1 and 2, which shows the flexibility of the labeling scheme. Finally, we used SNP-13 to connect fragments 1–3 with fragments 4–5 (see Supplementary Table 5 online for the observed bar codes).

We assessed the certainty of haplotype calling involving more than one fragment by determining the cumulative probability of all combined fragments, as genotypes of the SNPs linking two adjacent fragments are already known. For the results shown in Figure 3b, the certainty of haplotype calling was $(0.9999)^5 = 0.9995$. Regarding the statistical power to make the haplotype calling, assuming we determined the haplotypes of 30 sequences and observed two erroneous haplotypes, we would have 80% power to determine the haplotype when 18 (60%) out of 30 sequences were correctly called for homozygotes. For heterozygotes, we would only achieve 80% power when 24 (80%) out of 30 sequences were correctly called. The results we obtained far exceeded 80% power (Supplementary Methods).

We showed that long haplotypes can be obtained by combining the color bar codes of several DNA fragments as long as they are linked by informative SNPs in the overlapping regions. To our knowledge, this is the first time haplotypes of 16 SNPs over a 50 kb genomic region have been obtained directly without separating the chromosomes by cloning. There are no apparent barriers to obtaining molecular haplotypes from regions of even longer distances, as long-range PCR has been routinely used to generate 10–20 kb fragments, and primers have been designed to cover the entire human genome¹⁵. This molecular haplotyping approach can be used to determine haplotypes accurately and at low cost, making the fine mapping of population-based association studies more robust and giving it higher statistical power. This approach can also be used to verify DNA sequence assembly from data generated by massively parallel DNA sequencing projects and study copy-number alterations or structural rearrangements in the genome.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is supported by the US National Human Genome Research Institute (R01-HG01720).

References

1. Frazer KA, et al. Nature. 2007; 449:851–861. [PubMed: 17943122]
2. Collins FS, Green ED, Guttmacher AE, Guyer MS. Nature. 2003; 422:835–847. [PubMed: 12695777]
3. Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB. Nat. Genet. 2001; 28:361–364. [PubMed: 11443299]
4. Tost J, et al. Nucleic Acids Res. 2002; 30:e96. [PubMed: 12364613]
5. Ruano G, Kidd KK. Nucleic Acids Res. 1989; 17:8392–8392. [PubMed: 2573038]
6. Mitra RD, et al. Proc. Natl. Acad. Sci. USA. 2003; 100:5926–5931. [PubMed: 12730373]
7. Burgdorf C, et al. Genome Res. 2003; 13:2717–2724. [PubMed: 14656974]
8. Xiao M, et al. Nucleic Acids Res. 2007; 35:e16. [PubMed: 17175538]
9. Xiao M, et al. Hum. Mutat. 2007; 28:913–921. [PubMed: 17443670]
10. Kwok PY, Xiao M. Hum. Mutat. 2004; 23:442–446. [PubMed: 15108275]
11. Yildiz A, et al. Science. 2003; 300:2061–2065. [PubMed: 12791999]
12. Latorra D, Campbell K, Wolter A, Hurley JM. Hum. Mutat. 2003; 22:79–85. [PubMed: 12815597]
13. Helms C, et al. Nat. Genet. 2003; 35:349–356. [PubMed: 14608357]
14. Saidi S, Slamia LB, Mahjoub T, Ammou SB, Almawi WY. J. Stroke Cerebrovasc. Dis. 2007; 16:153–159. [PubMed: 17689411]
15. Patil N, et al. Science. 2001; 294:1719–1723. [PubMed: 11721056]

Figure 1. Allele-specific labeling of SNPs and regeneration of dsDNA

(a) Two oligonucleotides, each labeled with a different-colored fluorophore, are designed with one of the SNP alleles at the 3' end. Only the perfectly matched primers will be extended by DNA polymerase. Subsequently ssDNA is converted to dsDNA by primer extension. (b) Gel showing regenerated dsDNA from ssDNA. Long-range PCR products of four different DNA samples (lanes 9–12) were digested with lambda exonuclease to produce the corresponding ssDNA (lanes 5–8, respectively). After primer extension, dsDNA samples were regenerated (lanes 1–4, respectively). M, marker.

Figure 2. Molecular haplotyping of four SNPs across a 12.5 kb region

(a) Relative locations of four SNPs on a 12.5 kb DNA fragment. The arrows indicate the distance from the SNP site to one end of the DNA molecule. (b) The rescaled false-color composite image of labeled DNA fragments of DNA sample NA12156 labeled with YOYO-1 (blue) and Cy3 (green) or Cy5 (red). High-magnification images of two fully labeled fragments marked by boxes are shown in the insets. (c,d) Images of labeled DNA fragments of DNA sample NA010860 (c) and NA07357 (d). Fully labeled fragments are marked by arrows. Scale bars, 5 μ m (b–d) and 2 μ m (insets in b).

Figure 3. Molecular haplotyping of 16 SNPs across the 50 kb *PLAT* gene region
(a) Schematic of the region showing the positions of 16 SNPs and the overlapping fragments. The two alleles of each SNP are represented as the green (Cy3) and red (Cy5) circles on the DNA backbone. (b) Images from fully labeled DNA fragments for DNA sample NA12865. Six SNPs in fragment 1 were haplotyped in two groups. Two predominant haplotypes were found for the first (SNPs 5,7,9,10) and the second (SNPs 6,8,10) group. The two full haplotypes of the six SNPs comprising fragment 1 are T-G-T-C-T-T and G-C-C-G-G-C for SNPs 5,6,7,8,9,10. Haplotypes for the remaining four fragments are indicated, and the frequencies of the haplotypes are described in Supplementary Table 5.