Long-range polony haplotyping of individual human chromosome molecules

Kun Zhang[1], Jun Zhu[2], Jay Shendure[1], Gregory J Porreca[1], John D Aach[1], Robi D Mitra[3] & George M Church[1]

We report a method for multilocus long-range haplotyping on human chromosome molecules *in vitro* based on the DNA polymerase colony (polony) technology. By immobilizing thousands of intact chromosome molecules within a polyacrylamide gel on a microscope slide and performing multiple amplifications from single molecules, we determined long-range haplotypes spanning a 153-Mb region of human chromosome 7 and found evidence of rare mitotic recombination events in human lymphocytes. Furthermore, the parallel nature of DNA polony technology allows efficient haplotyping on pooled DNAs from a population on one slide, with a throughput three orders of magnitudes higher than current molecular haplotyping methods. Linkage disequilibrium statistics established by our pooled DNA haplotyping method are more accurate than statistically inferred haplotypes. This haplotyping method is well suited for candidate gene–based association studies as well as for investigating the pattern of recombination in mammalian cells.

Haplotypes, or combinations of alleles of multiple genetic markers on single chromosomes, are important for mapping human disease genes, diagnosing loss of heterozygosity in cancer, and investigating *cis* effects in gene expression. Unlike genotyping, obtaining haplotypes experimentally is technically challenging owing to the difficulty of separating two almost identical copies of chromosomes in diploid cells. Current molecular haplotyping technologies are inefficient and limited to a short distance, usually less than 100 kb[1–4]. For long-range (>100 kb) haplotypes, construction of somatic cell hybrids[5] is the only experimental approach available, but this is practical only for a small sample size[6]. None of the current methods is practical for association studies.

DNA polony technology[7] has been used to generate parallel and independent PCR amplifications of 1,000-10,000 template molecules in a thin layer (~40 μm) of polyacrylamide attached to a standard microscope slide. The localized clonal amplicons (polonies) are covalently linked to the gel matrix, and can subsequently be queried in parallel by *in situ* probing[3,8] or sequencing[9]. In this work, we develop the polony technology for long-range multilocus haplotyping on individual chromosome molecules (**Fig. 1**), demonstrate its

application to pooled DNA samples and show that the polony haplotyping method is superior to statistical haplotype inference methods in the accuracy of both haplotype frequencies and linkage disequilibrium statistics.

RESULTS

Long-range multilocus haplotyping on chromosome molecules

To perform chromosome-wide haplotyping, two major technical challenges must be overcome. First, haplotyping on multiple loci across entire chromosomes requires multiplexed PCR on single DNA molecules. We previously demonstrated amplification of two loci from a single DNA molecule[3], but amplifying more than two amplicons from one template molecule is far more complicated because the scale of primer interaction grows exponentially as the number of target loci increases. We wanted the multiplexed amplification to be highly specific to the target regions in the presence of complex genome, because nonspecific amplification (including primer-dimers) can compete with target-specific amplification on the limited amount of reagents within the polyacrylamide gel and lead to poor signal. To avoid tedious experimental optimization, we developed a new computational procedure based on nearest-neighbor thermodynamics[10] to simulate multiplexed amplification, so that nonspecific amplification across the entire genome can be evaluated quantitatively, and only highly specific primers are designed (see Methods).

The second technical challenge is to efficiently amplify large DNA molecules in a manner that results in overlapping polonies. A balance must be struck between packing large DNA molecules into an extremely compact form before gel immobilization and making condensed DNA molecules sufficiently 'loosened' afterwards to allow for efficient amplification. Towards this end, we prepared intact human metaphase chromosomes in which long DNA molecules ranging from 300 Mb (human chromosome 1) to 47 Mb (human chromosome 21) were highly condensed into a form less than 1 μm in size. Initially, no polony amplification could be obtained from such chromosomes with standard protocols[3]; intact chromosomes seem to be resistant to enzymatic treatment in polyacrylamide gel. We developed a method to make the DNA molecules accessible to DNA polymerase and PCR primers by first rupturing chromosome structure

---

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Institute of Genome Science and Policy, Duke University, Durham, North Carolina 27708, USA. [3]Department of Genetics, Washington University, St. Louis, Missouri 63108, USA. Correspondence should be addressed to K.Z. (kzhang@genetics.med.harvard.edu) or G.M.C. (http://arep.med.harvard.edu/gmc/email.html)
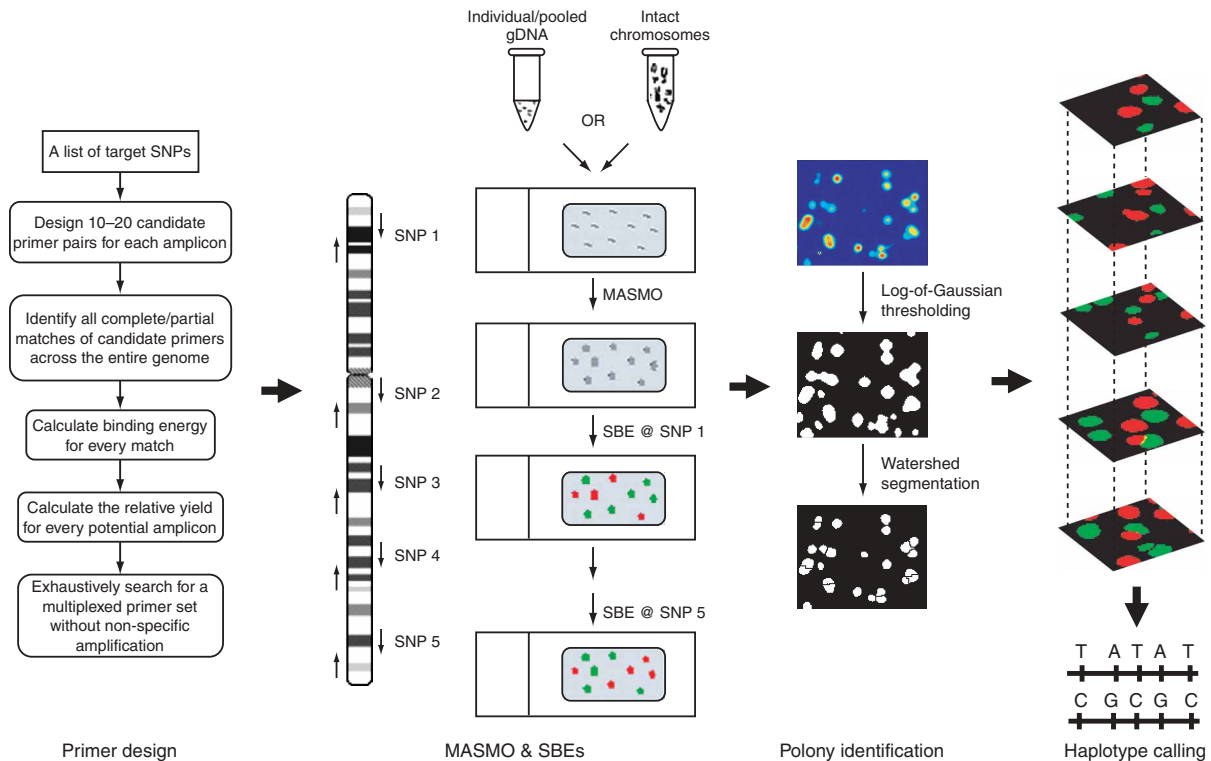
Figure 1 Multi-locus long-range haplotyping. Our method is composed of three major components: (i) a computation procedure for multiplexed single-molecule PCR based on whole-genome simulation, (ii) a polony haplotyping protocol and (iii) an image analysis algorithm to identify polony and extract haplotypes. In polony haplotyping experiment, limited amounts of genomic DNA or intact chromosomes from individuals or pools or individuals were trapped within polyacrylamide gel on the surface of a standard microscope slide. Multiple amplification on single molecules (MASMO) was performed in parallel on each of the template molecule *in situ*. Genotypes/haplotypes were then queried by multiple rounds of single base extension (SBE) assays.

with the expansion force of water quickly freezing and then removing chromosome binding proteins with proteinase digestion.

Chromosome-wide haplotyping on human chromosome 7

These improvements enabled us to perform haplotyping on five SNPs spanning most of human chromosome 7 (158 Mb). We selected these SNPs so that the haplotypes could be unambiguously determined based on their genotypes and associated pedigree information (**Fig. 2a**). By stacking polony images of different SNP assays on top of each other, we obtained polony clusters, each representing a multilocus haplotype of a single DNA molecule (chromosome). We observed correct haplotypes on single chromosomes in both the two-locus analysis (**Fig. 2b**) and the multilocus analysis (**Fig. 2c**). In one polony slide, we found polony clusters from 414 chromosomes that gave haplotype reads of at least three SNPs. Of these haplotypes, 404 were compatible with the expected haplotypes based on the pedigree, including 11 complete haplotypes of all five SNPs. Ten observed haplotypes (2.5%) were incompatible with the expected haplotypes. To investigate these unexpected haplotype reads, we manually analyzed the corresponding polony images. Three (0.7%) were clearly due to errors in the polony identification algorithm. We observed good morphology and strong signals on the other seven polony clusters. One hypothesis to explain these unexpected haplotypes is that they are artifacts attributable to random overlaps of two chromosome fragments or uncondensed DNAs. If this is the case, one should not expect to observe pairs of complementary haplotypes. In fact, we observed two pairs of complementary haplotypes (**Fig. 2d**). In addition, when we applied our haplotype calling

algorithm to a mock data set in which the positions of all polonies were randomly shuffled, no four-locus haplotype was observed. Therefore, the most intriguing explanation for these unexpected haplotypes is rare mitotic recombination.

Efficient haplotyping on pooled DNA

We have shown that multilocus haplotypes can be determined accurately when one sample is assayed with a polony slide. To take full advantage of the highly parallel nature of polony amplification, we evaluated a new strategy by haplotyping pooled samples. Genomic DNAs (or intact chromosomes) from numerous individuals were mixed in equal ratios, haplotypes were read from thousands of single molecules from a single slide in one experiment, and linkage disequilibrium (LD) statistics were calculated with haplotype frequencies. We focused on the CD36 (14 SNPs in 26 kb) and NOS3 (nine SNPs in 19 kb) that were completely resequenced in the SeattleSNP project. We did haplotyping on the pooled genomic DNA of the same 24 individuals in the HD50AA panel (**Supplementary Table 1** online). To validate the pooled haplotyping method, we experimentally determined all 48 haplotypes among the 24 individuals on the CD36 gene by performing polony haplotyping on every sample with a separate slide (**Supplementary Table 2** online). We used these 48 haplotypes as reference haplotypes to characterize the accuracy of the pooled method and to compare this method to statistical haplotype inference methods. We performed multilocus haplotype inference from genotypes using two state-of-the-art algorithms: PHASE[11], which is based on a coalescent framework, and SNPHAP, which is an implementation of the expectation maximization (EM) algorithm. As all pair-wise LD
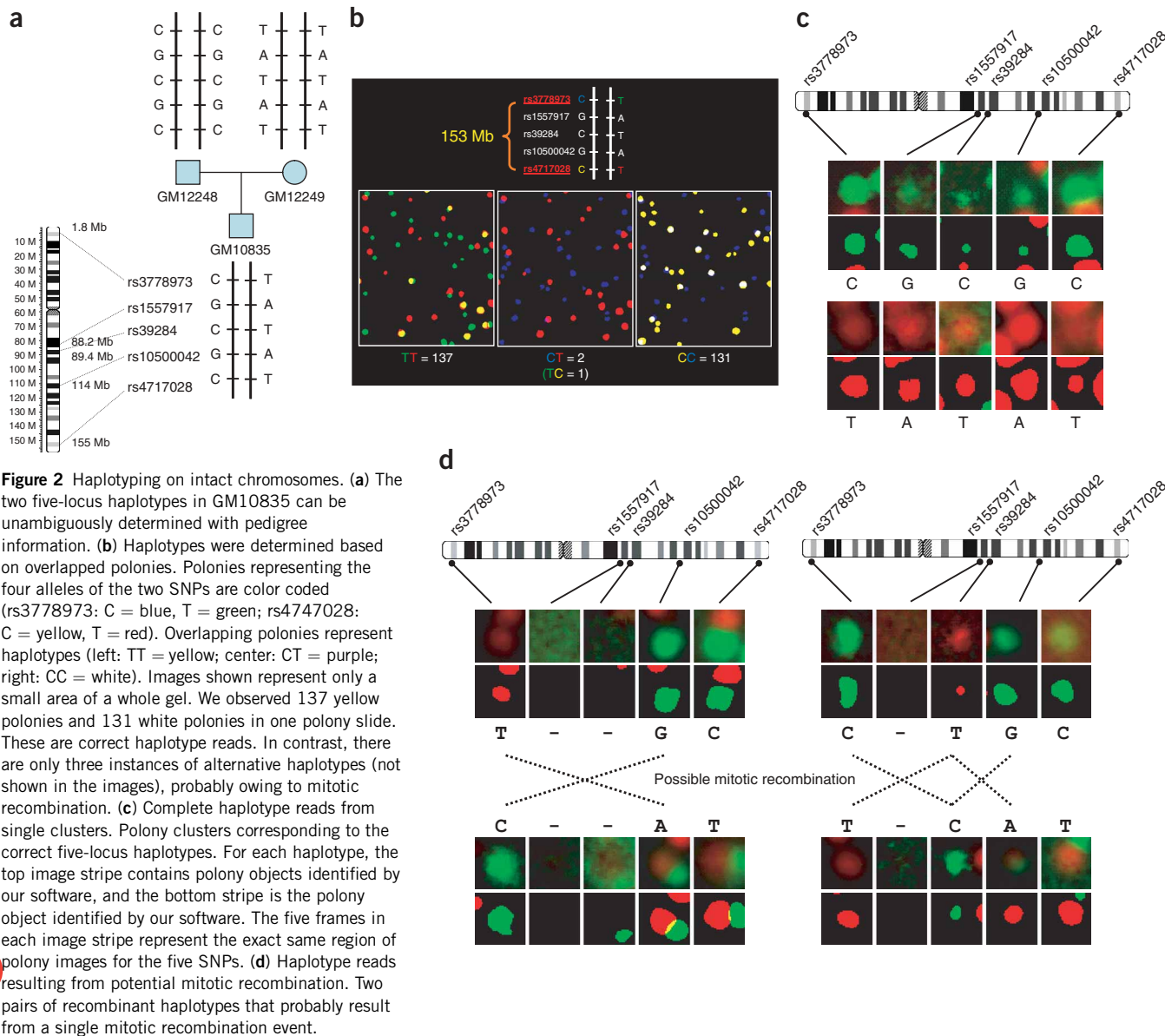
Figure 2 Haplotyping on intact chromosomes. (**a**) The two five-locus haplotypes in GM10835 can be unambiguously determined with pedigree information. (**b**) Haplotypes were determined based on overlapped polonies. Polonies representing the four alleles of the two SNPs are color coded (rs3778973: C = blue, T = green; rs4747028: C = yellow, T = red). Overlapping polonies represent haplotypes (left: TT = yellow; center: CT = purple; right: CC = white). Images shown represent only a small area of a whole gel. We observed 137 yellow polonies and 131 white polonies in one polony slide. These are correct haplotype reads. In contrast, there are only three instances of alternative haplotypes (not shown in the images), probably owing to mitotic recombination. (**c**) Complete haplotype reads from single clusters. Polony clusters corresponding to the correct five-locus haplotypes. For each haplotype, the top image stripe contains polony objects identified by our software, and the bottom stripe is the polony object identified by our software. The five frames in each image stripe represent the exact same region of polony images for the five SNPs. (**d**) Haplotype reads resulting from potential mitotic recombination. Two pairs of recombinant haplotypes that probably result from a single mitotic recombination event.

statistics are defined on two-locus haplotypes, we also did calculation with a third method, reconstructing two-locus haplotypes for each pair of SNPs from genotypes using the EM algorithm.

From the pooled samples, we obtained on average 19,556 polony reads at CD36 and 16,313 at NOS3 in one polony slide. For each pair of SNPs, we observed on average 275 two-locus haplotype reads at CD36 and 97 reads at NOS3 (25,025 two-locus haplotypes among all SNPs at CD36 and 3,492 two-locus haplotypes at NOS3 in total). We often observed partial haplotypes with one or more SNPs missing (**Supplementary Figure 1** online), because amplification from single molecules is not 100% efficient. We focused on polony clusters (haplotypes) with good calls for at least four SNPs. Of a total of 3,900 such haplotype reads in the CD36 gene, only 81 (2.07%) were incompatible with true haplotypes. In contrast, 4 of 48 (8.3%) haplotypes inferred by PHASE were incorrect—all of them were rare haplotypes occurring only once in the population. Similarly, all of the three incorrect haplotypes inferred by SNPHAP occur once in the

population. However, the SNPHAP predicted haplotype with the lowest frequency (0.2%) is actually a true haplotype. These results confirm a previous finding that current multilocus haplotype inference methods are not reliable in predicting rare haplotypes[12], which could be equally important in association studies, given the recent finding that rare variant sites do contribute to complex traits[13].

We next performed two-locus haplotype analysis and calculated the LD statistics in CD36 and NOS3. Two commonly used LD measurements, $|D'|$ and r^2, determined by the pooled polony haplotyping method are highly consistent with the values calculated from reference haplotypes (Pearson R^2: 0.892 for $|D'|$ and 0.974 for r^2). In contrast, all the three haplotype inference methods give good estimates of r^2 but very poor estimates of $|D'|$ (**Table 1**). The sharp difference between these two LD statistics is probably due to the extreme sensitivity of $|D'|$ to the frequencies of rare haplotypes. Therefore partitioning of haplotype block based on inferred haplotypes using a $|D'|$ threshold can be problematic. The pooled polony haplotyping

Table 1 Correlation (Pearson's R^2) of the |D'| (lower left triangle) and r^2 (upper right triangle) values in the CD36 gene determined by different methods

	True haplotypes	Pooled polony (24)	PHASE	SNPHAP	Two-locus EM
True haplotypes	–	0.974	0.963	0.959	0.971
Pooled polony (24)	0.892	–	0.958	0.967	0.954
PHASE	0.306	0.368	–	0.980	0.982
SNPHAP	0.388	0.417	0.808	–	0.972
Two-locus EM	0.396	0.419	0.648	0.681	–

True haplotypes were determined experimentally by polony haplotyping on genomic DNA from each individual.

method is superior to haplotype inference methods in establishing LD statistics regardless of the LD measurement used. Notably, the two-locus EM method actually gives slightly more accurate estimates of LD statistics than PHASE and SNPHAP at both CD36 (**Table 1**) and NOS3 (**Supplementary Table 3** online). Besides LD statistics, we also did comparisons based on the mean absolute errors (MAEs) of frequencies for all two-locus haplotypes (pooled polony: 0.0158; PHASE: 0.0232; SNPHAP: 0.0191; two-locus EM: 0.0213). The pooled polony haplotyping method gives the lowest MAE, although the differences among all four methods are within twofold. The above comparisons clearly show that the pooled polony haplotyping method is more accurate than haplotype inference methods in both haplotype frequencies and the LD statistics. To our best knowledge, no validation has been made on the accuracy of LD patterns estimated by current haplotype inference methods. Our analysis also shows that r^2 is more reliable than |D'| in studying LD patterns based on statistically inferred haplotypes.

Finally, as the polony technology allows efficient haplotyping of a large DNA pool in one experiment, we sought to study the effect of population size on LD analyses by typing a complete set of 50 individuals in the African American human variation panel (HD50AA). We observed weaker LD at both CD36 and NOS3 in this larger population, although the overall patterns are similar (**Supplementary Figure 2**). This suggests that the level of LD as well as the size of haplotype blocks can be inflated when the population size is small, because rare haplotypes are unlikely to be observed in a small sample size. This finding is consistent with a previous simulation study based on haplotypes generated with a coalescent model[14].

DISCUSSION

In this report, we have demonstrated that the polony haplotyping technology can be implemented in a highly efficient manner. The efficiency comes from two aspects: (i) parallel haplotyping on thousands of single molecules within a single slide allows us to use pooled DNA from a population and (ii) multilocus haplotyping on a single slide is more efficient than two-locus haplotyping methods such as allele-specific PCR. In the case of the CD36 gene, we were able to obtain on average 275 two-locus haplotype reads for four possible haplotypes of each pair of SNPs. This means, for a population size of 50 individuals, a threefold redundancy can be achieved with a single polony slide. A 14-locus haplotyping experiment on 50 individuals is equivalent to 2,600 allele-specific PCR assays, an improvement of more than three orders of magnitude. Polony haplotyping is also very cost efficient. The cost of a pooled experiment on 25 individuals is ∼$3 per SNP per individual, but it goes down rapidly to ∼$0.25 when the sample size increases to 1,000 individuals. In addition, as was

shown above, LD levels can be inflated when the analyses are based on a small sample size; thus, reliable estimates of LD require large sample sizes. As one can easily handle 10 to 20 polony slides in an experiment, the polony haplotyping technology is especially efficient in dealing with large sample sizes.

One major limitation is that polony haplotyping is less efficient for scaling up the number of SNPs rather than the number of individuals in one assay, because typing on each SNP requires one single-base extension (SBE) assay, and polyacrylamide gels are likely to break after ∼25 SBE assays. At the current stage of technology development, this technology is most appropriate for candidate gene–based studies. It was recently reported that in the two-stage study design, the error rates of haplotype reconstruction for tagging SNPs is significantly higher than from a full set of SNPs among which these tagging SNPs were selected[15]. The polony haplotyping technology is particularly useful for establishing the patterns of LD between tagging SNPs experimentally in the case and control populations in the second stage of association studies.

As haplotype blocks in the human genome rarely exceed 100 kb, the capability of haplotyping across a whole chromosome is more useful for studies focusing on haplotypes spanning megabase distances, such as admixture mapping[16,17] and detecting positive selection[18], rather than for association studies. One unexpected finding in our chromosome-wide haplotyping experiment is that this method is an ideal assay to detect rare mitotic recombination events. Although occurring less frequently than meiotic recombination, mitotic recombination has been observed in many eukaryotes, including yeast[19], mouse[20] and human[21] and has implications in many human diseases, such as neurofibromatosis type 1 (ref. 22) and ataxia-telangiectasia[21]. Polony haplotyping allows detection of mitotic recombination events across the whole genome in unmodified cells, and subsequently mapping crossovers to a resolution that is limited only by the availability of SNPs. Similarly, chromosome translocation in cancer cells is also detectable with the haplotyping method. Finally, studies on meiotic recombination hotspots have attracted considerable interest recently. Based on single-sperm typing, the distribution and intensity of a few recombination hotspots has been characterized in great detailed in humans[23,24]. However, many questions remain unanswered owing to the difficulty of scaling up. Polony haplotyping may overcome the limitation associated with single-sperm typing[25]. It will be of great interest to compare the intensity and distribution of mitotic and meiotic crossovers.

METHODS

Primer design for multiple amplifications from single molecules (MASMO). The percentage of template DNA molecules ($p\%$) that binds to primers under a given temperature T is calculated with the formula

$$p\% = \left(\frac{e^{(\Delta H - T*\Delta S)/RT}}{Cp^{+1}} \right)^{-1} \tag{1}$$

where Cp is the primer concentration, ΔH and ΔS are calculated based on a set of nearest-neighbor (NN) thermodynamic parameters that has been confirmed under PCR conditions[10,26], R is the Boltzmann constant and T is the temperature. We confirmed this NN parameter set in our experimental conditions by measuring the annealing temperature of 20 primers and their complementary oligonucleotide sequences using SYBR Green I–based melting curve analysis. We modified the Massachusetts Institute of Technology Primer3 program (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) to incorporate this NN parameter set. For each amplicon, we designed 20 candidate primer pairs for each amplicons, identified all exact and partial matches in the human genome by BLAST and calculated the percentage of each matched site that anneals to the primers. Partial matches with mismatches at the 3′ end as well as two or more mismatches in the middle were ignored. Any genomic

region shorter than the upper limit of amplicon size that can be amplified by PCR (we assumed *Taq* can extend at most 2 kb per min) that had two primer matching sites in the correct orientation at both ends were considered a potential amplicon. For simplicity, the relative yield of each potential amplicon is the product of *p*% of both primer matching sites. We used this to calculate the relative amplification yield for the target genomic region as well as all other nonspecific amplifications. To design primers for multiplex PCR, we used a Perl script (yamPCR; K. Zhang, Harvard Medical School) to perform an exhaustive search among all candidate primers on all target amplicons, taking into account nonspecific amplifications and primer-dimers. To our knowledge, this is the first attempt at quantitative PCR simulation in a complex genome.

Polony haplotyping. Human lymphocyte cell line GM10835, African American genomic DNAs used in the SeattleSNP project and pooled genomic DNA of the HD50AA panel (NA16600) were purchased from the Coriell Cell Repository. GM10835 was grown in RPMI-1640 medium with 15% fetal bovine serum, blocked at metaphase with 0.05 µg ml^{-1} colcemid overnight. Intact chromosome were isolated with the MgSO$_4$ protocol[27]. Unlysed cells were removed by 5 µM filters. Chromosomes were stained with YOYO-1 dye and inspected with a fluorescent microscope. We modified the polony amplification protocol[3] for multilocus haplotyping. After chromosomes were embedded in an acrylamide gel, a 125 µl frame-seal chamber was mounted on the slide, and 80 µl digestion buffer (0.5 mg ml^{-1} proteinase K, 0.5% SDS, 10 mM Tris-HCl, pH 8.0) was added. Polony slides were quickly frozen in a –80 °C freezer for 10 min and were then incubated at 37 °C overnight and then at 80 °C for 20 min. After removing frame-seal chambers, slides were washed in dH$_2$O twice for 5 min. When gels were completely dry, 20 µl of amplification mix (1× Qiagen HotStarTaq master mix, 1 unit Pfu Turbo Hotstart, 0.5 µM of each unmodified primer) was added to each gel under a 18 × 30 mm glass cover slip. The slides were then sealed in a Secure-Seal chamber (Invitrogen) with ∼550 µl mineral oil, and thermocycled (95 °C for 10 min; 50 rounds of 94 °C for 45 s, 58 °C for 45 s and 72 °C for 1.5 min; final extension of 72 °C for 7 min). We performed SBE using the deoxynucleotide protocol with Cy3 and Cy5 fluorophores[3]. A more stringent primer annealing buffer (50 mM NaCl, 1.5 mM MgCl$_2$, Tris.HCl, pH 8.0) was used. The T_m for all amplification primers was 64–65 °C. SBE primers had a T_m between 55 °C and 65 °C, depending on the local GC content. The annealing temperature of SBE was 5 °C below T_m.

Image analyses. For each SBE assay, we obtained two set of images: polony images and background images (obtained after stripping fluorescently labeled primers). Images were processed using the MatLab Image Analysis Toolbox. First, we corrected for background by subtracting the peak of a smoothed histogram of pixel intensity values and by median filtering. The images were then segmented by applying a log-of-gaussian (LOG) filter and looking for regions where LOG was less than a small negative threshold. Subsequently, we used watershed segmentation on the distance transform of the LOG-segmented image to further subdivide segments containing large indentations (such as circles touching the individual circular components). Only segments that met appropriate size and shape characteristics were retained as candidate polonies. Polony images acquired from different scans of the same slide were globally aligned based on the polony binary object images. Polony segments from different scans were considered to overlap when the overlapped area was more than 70% of the area of the smaller polony. We evaluated other thresholds, including 60%, 80% and 90%, and found that 70% was most appropriate. We developed two methods to correct haplotype and haplotype counts owing to randomly overlapping polonies. In the first method, we randomized the positions of all polonies in an image, calculated the number of overlapping polonies and used this to adjust the polony counts. The second method was to simply ignore all polonies that overlapped with fewer than three polonies on the other SNPs. We found that the first method tended to overadjust the haplotype count and led to inflation of LD statistics, so all analyses were based on the second method. For calculation of LD statistics, we combined haplotype reads from two or four slides to reduce random sampling error, because haplotype counts obtained from different polony slides were additive.

LD analyses. Genotypes of CD36 and NOS3 were downloaded from the SeattleSNP website. Missing or incorrect genotypes of CD36 were corrected by our polony haplotyping experiments. PHASE (v. 2) and SNPHAP were downloaded and run locally using the default parameters. To ensure a fair comparison, the two-locus EM haplotype inference was also conducted by SNPHAP, which was run in a batch mode by the Perl program Genotype2LD-Block; K. Zhang, Harvard Medical School. In the CD36 gene, 14 (4.2%) genotypes are missing in the SeattleSNP data. These genotypes were determined in our polony haplotyping experiment. In addition, we identified one genotyping error: the individual #D012 is AA homozygous instead of AG heterozygous at CD36-003094. Our analyses on haplotype inference methods were based on these updated genotype data.

URLs. yamPCR is available at http://arep.med.harvard.edu/kzhang/polHap. Matlab scripts for polony image analysis are available at http://arep.med.harvard.edu/kzhang/polHap. SeattleSNP website: http://pga.gs.washington.edu. PHASE (version 2) and SNPHAP were downloaded from http://www.stat.washington.edu/stephens/software.html and http://www-gene.cimr.cam.ac.uk/clayton/software. Genotype2LDBlock: http://arep.med.harvard.edu/kzhang/cgi-bin/genotyp2LDBlock.cgi.

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

1. Michalatos-Beloin, S., Tishkoff, S.A., Bentley, K.L., Kidd, K.K. & Ruano, G. Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. *Nucleic Acids Res.* **24**, 4841–4843 (1996).
2. Ding, C. & Cantor, C.R. Direct molecular haplotyping of long-range genomic DNA with M1-PCR. *Proc. Natl. Acad. Sci. USA* **100**, 7449–7453 (2003).
3. Mitra, R.D. *et al.* Digital genotyping and haplotyping with polymerase colonies. *Proc. Natl. Acad. Sci. USA* **100**, 5926–5931 (2003).
4. Woolley, A.T., Guillemette, C., Li Cheung, C., Housman, D.E. & Lieber, C.M. Direct haplotyping of kilobase-size DNA using carbon nanotube probes. *Nat. Biotechnol.* **18**, 760–763 (2000).
5. Douglas, J.A., Boehnke, M., Gillanders, E., Trent, J.M. & Gruber, S.B. Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nat. Genet.* **28**, 361–364 (2001).
6. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
7. Mitra, R.D. & Church, G.M. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res.* **27**, e34 (1999).
8. Zhu, J., Shendure, J., Mitra, R.D. & Church, G.M. Single molecule profiling of alternative pre-mRNA splicing. *Science* **301**, 836–838 (2003).
9. Mitra, R.D., Shendure, J., Olejnik, J., Edyta Krzymanska, O. & Church, G.M. Fluorescent in situ sequencing on polymerase colonies. *Anal. Biochem.* **320**, 55–65 (2003).
10. SantaLucia, J., Jr. & Hicks, D. The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 415–440 (2004).
11. Stephens, M., Smith, N.J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989 (2001).
12. Tishkoff, S.A., Pakstis, A.J., Ruano, G. & Kidd, K.K. The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. *Am. J. Hum. Genet.* **67**, 518–522 (2000).
13. Cohen, J.C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869–872 (2004).
14. Wang, N., Akey, J.M., Zhang, K., Chakraborty, R. & Jin, L. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.* **71**, 1227–1234 (2002).
15. Forton, J. *et al.* Accuracy of haplotype reconstruction from haplotype-tagging single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **76**, 438–448 (2005).
16. Patterson, N. *et al.* Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74**, 979–1000 (2004).
17. Smith, M.W. *et al.* A high-density admixture map for disease gene discovery in african americans. *Am. J. Hum. Genet.* **74**, 1001–1013 (2004).
18. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004).

19. Prado, F., Cortes-Ledesma, F., Huertas, P. & Aguilera, A. Mitotic recombination in *Saccharomyces cerevisiae. Curr. Genet.* **42**, 185–198 (2003).

20. Hendricks, C.A. *et al.* Spontaneous mitotic homologous recombination at an enhanced yellow fluorescent protein (EYFP) cDNA direct repeat in transgenic mice. *Proc. Natl. Acad. Sci. USA* **100**, 6325–6330 (2003).

21. Meyn, M.S. High spontaneous intrachromosomal recombination rates in ataxia-telangiectasia. *Science* **260**, 1327–1330 (1993).

22. Kehrer-Sawatzki, H. *et al.* High frequency of mosaicism among patients with neuro-fibromatosis type 1 (NF1) with microdeletions caused by somatic recombination of the JJAZ1 gene. *Am. J. Hum. Genet.* **75**, 410–423 (2004).

23. Winckler, W. *et al.* Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* **308**, 107–111 (2005).

24. Jeffreys, A.J., Neumann, R., Panayi, M., Myers, S. & Donnelly, P. Human recombination hot spots hidden in regions of strong marker association. *Nat. Genet.* **37**, 601–606 (2005).

25. Kauppi, L., Jeffreys, A.J. & Keeney, S. Where the crossovers are: recombination distributions in mammals. *Nat. Rev. Genet.* **5**, 413–424 (2004).

26. von Ahsen, N., Wittwer, C.T. & Schutz, E. Oligonucleotide melting temperatures under PCR conditions: nearest-neighbor corrections for Mg(2+), deoxynucleotide triphosphate, and dimethyl sulfoxide concentrations with comparison to alternative empirical formulas. *Clin. Chem.* **47**, 1956–1961 (2001).

27. Cram, L.S., Bell, C.S. & Fawcett, J.J. Chromosome sorting and genomics. *Methods Cell Sci.* **24**, 27–35 (2002).