Completely phased genome sequencing through chromosome sorting

Hong Yang^{a,b}, Xi Chen^{a,b}, and Wing Hung Wong^{a,b,c,1}

^aDepartment of Statistics, ^bBio-X Program, and ^cDepartment of Health Research and Policy, Stanford University, Stanford, CA 94305-4065

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2009.

Contributed by Wing Hung Wong, November 12, 2010 (sent for review October 12, 2010)

The two haploid genome sequences that a person inherits from the two parents represent the most fundamentally useful type of genetic information for the study of heritable diseases and the development of personalized medicine. Because of the difficulty in obtaining long-range phase information, current sequencing methods are unable to provide this information. Here, we introduce and show feasibility of a scalable approach capable of generating genomic sequences completely phased across the entire chromosome.

haplotype | phased sequencing | single chromosome sequencing | SNP

Although high-density SNP arrays can measure the genotypes for millions of SNPs across the genome, these genotypes are not phased. The generation of phased genotypes, which is critically important for many genetic analyses (1, 2), is a long-standing challenge in human genetics. At present, the most powerful way to obtain phase information on a genome-wide scale is to rely on data from relatives. In the HapMap project (3, 4), common European haplotypes are inferred from SNP array profiles from 30 mother–father–child trios of European descent. In Iceland, where a significant percentage of the extant population has been genotyped, data from distantly related individuals can be exploited to infer long-range haplotypes (5). However, because of difficulty in obtaining family data, large-scale studies, including most genome-wide association studies, are usually based on unrelated individuals, and phase information is inferred by statistical methods (6). This approach, however, only works for markers that are very close to each other. Ultimately, experimental approaches remain the most attractive solutions to the phasing problem. Current approaches include using emulsion PCR to condense polymorphic sites from a single template (7, 8), genotyping from diluted aliquots of DNA fragments (9), allele-specific imaging of long-range PCR products (10), using long-range polony (11), genotyping from sperms (12), and isolating single chromosomes by interspecific cell fusion (13). Most of these methods were designed for haplotyping only a small number of markers. An exception is a method that performs SNP array profiling after chromosome microdissection (14).

Because new generation DNA sequencing has revolutionized many biomedical areas (15), we ask whether it can be adapted to solve the long-range phasing problem. In current protocols, DNA extracted from a large number of cells is fragmented, amplified, and then, sequenced in a parallel manner. Massive parallelism allows the generation of a sufficient number of short sequences (reads) to cover the genome many times. By aligning the reads, it is possible to reconstruct the two alleles in any given small region. However, phase information between parental allelic sequences in two nonadjacent regions is usually not recoverable, regardless of the number of short reads. To overcome this problem, we use chromosome sorting to isolate single copies of a chromosome. Each copy is separately amplified, and the products are tagged by a short stretch of nucleotides before being pooled together (multiplexed) for massively parallel sequencing (Fig. 1). The tag

allows us to assign a read back to the single chromosome copy that gave rise to it. We then perform statistical analysis based on the status of polymorphic sites to cluster the single copies of the chromosome into two clusters so that the copies in the same cluster are from the same parental allele. In this way, if a sufficient number of reads has been sampled from each allele, we can reconstruct the paternal and maternal haploid genome sequences separately. Below, we describe the steps of this method, which we name Phase-Seq, in a proof of principle experiment.

Results

We obtained blood samples of a donor individual and collected single chromosomes using FACS-mediated single chromosome sorting (16), which identifies each chromosome by its distinct bivariate distribution of fluorescent signals from the staining of Chromomycin A3 (binds Guanine-Cytosine-rich regions) and Hoechst 33258 (binds Adenine-Thymine-rich regions). We collected and amplified 28 single copies of chromosome 19 (Chr19) separately along with 12 control samples (*Materials and Methods*). Chr19-specific DNA sequences were verified on the single chromosome samples and positive controls, but not on the negative controls, using sequence-specific PCR primers. We then sequenced the 28 single chromosome samples in a multiplex sample format, which included 12 samples per lane on an Illumina GAIIx sequencer, with each sample uniquely identifiable by a six-base index tag. Five lanes of sequence image data were obtained and analyzed using the next_phred software by Phil Green. Fig. 1 illustrates the workflow.

Fig. 2 shows that the reads predominantly map to Chr19 (Fig. 2A). Alignments to other chromosomes generally have much lower mapping quality scores (Fig. 2B). The percentage of the reads mapped to Chr19 increases from 40% to 93% when the mapping quality score cutoff is increased from 0 to 100 (Fig. 2C). These results indicate that we were able to obtain copies of Chr19 with high specificity and that the amplification and sequencing procedures preserved this specificity.

To analyze the distribution of the reads, we divided Chr19 into nonoverlapping windows of 100 Kb. For each window, we computed the percentage of positions within the window that is covered by reads. The result (Fig. 3A) shows that generally about 40% of the positions are covered, and 20% are covered by five or more reads. Exceptions to this are the centromere region, which has no reads, and a 4.5-Mb region, which has lower than average

Author contributions: H.Y. and W.H.W. designed research; H.Y. and X.C. performed research; H.Y. and W.H.W. analyzed data; H.Y. and W.H.W. wrote the paper; H.Y. performed single chromosome sequencing and computational analysis; X.C. prepared chromosomes for sorting; and W.H.W. conceived the approach.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The sequence reported in this paper has been deposited in the SRA database (accession no. [SRA026487.1](https://www.ncbi.nlm.nih.gov/sra/SRA026487.1)).

¹To whom correspondence should be addressed. E-mail: whwong@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1016725108/-DCSupplemental.

Fig. 1. Schematic diagram of Phase-Seq work flow. Single chromosomes were sorted into wells of a 96-well plate in which single chromosome amplifications were performed. Each amplified DNA molecule from a single chromosome (e.g., Chr19) contained a specific tag (shown in red or blue) that allowed multiplex sequencing on a high-throughput sequencing platform. Multiplexed reads were assigned to haploid genomes based on the combination of single chromosome-specific tags (shown in red or blue) and haploid genome-specific SNPs (lowercase letters in *italic bold type*). (*Inset*) FACS sorting of stained single chromosomes is based on the fluorescence patterns of Hoechst and Chromomycin, which allow reliable separation of different chromosomes (Chr18 and Chr19 are marked).

counts. Excluding the centromere region, the largest gap is 63 Kb in size, and only 40 gaps are larger than 5 Kb (Fig. 3B). The sum of all gaps of size larger than 1 Kb is 4,579 Kb, which is still a small fraction of Chr19. Thus, the amplification from single chromosomes had yielded a relatively unbiased coverage of the whole chromosome.

The reads were then assigned to individual single chromosomes according to the unique tag sequence for each sample; 9 of 28 tags have very low read counts, indicating failure of amplification. These were removed from all subsequent analysis. For the remaining 19 tags, we attempted to associate each of them with

Fig. 2. Sequence reads from sorted and amplified single chromosomes predominantly map to Chr19. (A) The count (in 10^5) of aligned reads that map to each chromosome. (B) The average mapping quality score for each chromosome. For both A and B, the color codes for the value of the next_phred mapping quality score cutoff (co) used to retain the reads. (C) The percentage of reads that map to Chr19 at various quality score cutoffs (0, 20, 40, 60, 80, and 100).

one of the two parental alleles. Using the de novo SNP calling program phastlane of the next_phred package to analyze each lane separately without using the tag information, we obtained 12,326 putative SNP positions that are either heterozygous or homozygous (i.e., identical in the two haploid genomes but different from the corresponding base in the reference genome sequence). Base identities at these positions from sequencing reads were analyzed, and the pair-wise consistency between pairs of tags of reads containing these SNPs was obtained. As shown in Table 1, single chromosomes can be cleanly divided into two clusters (*Materials and Methods* has the computation of consistency indexes and clustering). Cluster 1 has 10 single chromosomes corresponding to tags 1, 3, 7, 8, 9, 10, 11, 14, 23, and 24. Cluster 2 has nine single chromosomes corresponding to tags 2, 4, 5, 6, 12, 15, 16, 19, and 22. The within-cluster agreement is excellent (average consistency = 97.5%). The between-cluster consistency is much lower (around 50%) but not close to zero, because there must be

Table 1. Consistency indexes between pairs of tags

	1	3	7	8	9	10	11	14	23	24	2	4	5	6	12	15	16	19	22
1		96%	94%	95%	95%	95%	95%	94%	96%	96%	28%	26%	23%	26%	25%	24%	25%	22%	22%
3	96%		95%	97%	96%	96%	95%	100%	98%	97%	26%	53%	44%	50%	52%	49%	49%	47%	49%
7	94%	95%		95%	97%	96%	97%	98%	95%	97%	25%	45%	43%	45%	52%	51%	52%	50%	55%
8	95%	97%	95%		97%	93%	96%	100%	98%	97%	26%	46%	46%	46%	52%	53%	55%	50%	50%
9	95%	96%	97%	97%		98%	95%	100%	98%	99%	26%	44%	47%	44%	54%	53%	56%	51%	55%
10	95%	96%	96%	93%	98%		91%	100%	98%	96%	26%	44%	46%	45%	48%	47%	47%	42%	49%
11	95%	95%	97%	96%	95%	91%		98%	97%	97%	24%	45%	48%	45%	49%	49%	48%	47%	47%
14	94%	100%	98%	100%	100%	100%	98%		100%	100%	25%	24%	40%	32%	30%	31%	36%	36%	31%
23	96%	98%	95%	98%	98%	98%	97%	100%		97%	27%	46%	44%	47%	47%	61%	51%	48%	56%
24	96%	97%	97%	97%	99%	96%	97%	100%	97%		25%	44%	49%	43%	44%	57%	52%	55%	55%
2	28%	26%	25%	26%	26%	26%	24%	25%	27%	25%		94%	94%	95%	94%	96%	94%	95%	94%
4	26%	53%	45%	46%	44%	44%	45%	24%	46%	44%	94%		98%	97%	97%	97%	99%	98%	98%
5	23%	44%	43%	46%	47%	46%	48%	40%	44%	49%	94%	98%		96%	95%	97%	97%	97%	97%
6	26%	50%	45%	46%	44%	45%	45%	32%	47%	43%	95%	97%	96%		96%	96%	97%	97%	96%
12	25%	52%	52%	52%	54%	48%	49%	30%	47%	44%	94%	97%	95%	96%		97%	99%	99%	97%
15	24%	49%	51%	53%	53%	47%	49%	31%	61%	57%	96%	97%	97%	96%	97%		94%	98%	94%
16	25%	49%	52%	55%	56%	47%	48%	36%	51%	52%	94%	99%	97%	97%	99%	94%		98%	95%
19	22%	47%	50%	50%	51%	42%	47%	36%	48%	55%	95%	98%	97%	97%	99%	98%	98%		97%
22	22%	49%	55%	50%	55%	49%	47%	31%	56%	55%	94%	98%	97%	96%	97%	94%	95%	97%	
Counts	3,460	3,196	3,994	4,815	3,249	3,852	4,565	624	2,705	2,833	1,406	6,049	2,713	3,729	4,299	3,202	3,561	3,026	2,722

For a given pair of tags, consider all SNP positions detected by next_phred that are supported by three or more reads of mapping score ≥ 100 from each tag in the pair. The pair-wise consistency index is defined as the percentage of such SNP positions where the same nucleotide (C, G, A, or T) is found to have the highest frequency in each of the two tags. The bottom row displays the accumulated count for a tag used in the pair-wise comparison.

extended by using the completely phased SNP information uncovered from Phase-Seq.

Finally, based on the insertion/deletion (indel) calling function in next_phred, 202 small indels of sizes 1, 2, or 3 bp were discovered and phased throughout Chr19 (*Materials and Methods* and *Datasets S1, S2, S3, S4, S5, and S6*). We also attempted to detect larger indels by mapping split reads but found that most of the large indels detected this way map to Short Interspersed Nuclear Elements or other repeat regions, which we regard as unreliable. In future studies, the use of pair-end sequencing or third-generation technologies capable of long reads should enable reliable detection and phasing of large indels by Phase-Seq.

Discussion

Our approach should be scalable to larger studies. Modern FACS instruments are capable of sorting thousands of single chromosomes into collecting wells in a very short time. Using bivariate FACS sorting with Chromomycin and Hoechst staining, most chromosomes can be reliably resolved. The four remaining chromosomes (Chrs 9, 10, 11, and 12), which have similar bivariate distribution patterns under FACS, can be resolved by molecular typing after chromosome sorting and amplification. Molecular typing also serves as a quality control for sorted chromosomes. By increasing the length of the oligonucleotide tag, one can multiplex hundreds of single chromosome-derived samples, and the amplification and tagging workflow can be automated in a 384-well format. Because the capacity of new generation sequencers are still increasing at an exponential rate, there should be no fundamental barrier to the scaling of this method to large-scale personal genome sequencing projects.

In this study, we have analyzed many copies of each parental allele of the chromosome to establish clearly the feasibility and reliability of Phase-Seq. The number of single copies that need to be sequenced may be substantially reduced if unphased genome sequences are already available and we are only using Phase-Seq to obtain haplotypes on the polymorphisms detected from the unphased sequences. Thus, future personal genome sequencing projects might use both unphased and phased sequencing. Fur-

ther investigation will be necessary to understand the tradeoffs between the two for large-scale studies.

Heterozygous SNP discovery based on our approach should be more reliable than that from traditional (unphased) genome sequencing. This is because, in our heterozygous SNP calling procedure (*Materials and Methods*), we require not only a bimodal distribution of reads over the nucleotide identities (i.e., two different nucleotides each appearing multiple times) but also that the reads containing the same nucleotides must have tags mostly associated with the same parental allele. This is a powerful restriction that makes variant calling much stronger at heterozygous bases in the person's diploid genome.

Long-range phase information is important for studies of human diseases, development, adaptation, and population history. For example, the association of haplotypes as large as 12 Mb was found to be significant for transplant rejection (20). Admixture mapping (reviewed in ref. 21) of disease loci may be made more powerful if phasing can be directly determined rather than inferred through complex statistical models. Long-range haplotype information is important in understanding distal *in cis* regulation of spatially and temporally specific gene expression in development (22). Studies of human adaptation may benefit from a better ability to determine long-range haplotypes, which are signals of recent positive selection (23, 24). Determination of recombinant haplotypes provides insight into population history (25). Last but not least, Phase-Seq provides a foundation for resolving the roles of parental origin-specific genetic variants in disease association, allele-specific gene expression, and alternative splicing (1, 18, 19). These are just a few examples of the potential utility of phased genome sequence information, and researchers in many areas of biomedicine will find an enabling tool in Phase-Seq.

Materials and Methods

Blood Sample and Chromosome Preparation. Human blood samples for chromosome sorting donated by an anonymous donor were obtained. Each sample is tested free of viral infection before experiment. Chromosomes were prepared by using a slight modification of a described method (26). Approximately 8 mL of residual leukocytes obtained from Leukocyte Reduction System chamber pheresis were first enriched with RosetteSep DM-L (StemCell

Technologies). The lymphocytes were washed two times with PBS + 2% FCS and were counted before cultured at 0.5×10^6 cells/mL in RPMI complete medium containing 10% FCS and 10 μ g/mL Phytohemagglutinin-M (Roche). After 50 h incubation (37 °C and 5% CO₂), Demecolcine (Sigma) was added to the medium at 0.1 μ g/mL, and the cultures were harvested 14 h later.

To prepare chromosomes for sorting, the cells were first swelled in freshly made hypotonic solution at room temperature for 10 min. Then, the cells were spun down, and the pellet was resuspended in ice-cold polyamine buffer for 15 min. To break the cell walls, the cells were vortexed vigorously for 30–60 s at 4 °C, and the cell suspension was transferred into 1.5-mL Eppendorf tubes. Subsequently, the nuclei were removed from the chromosome suspension by centrifugation at $100 \times g$ for 3 min; 750 μ L of chromosome suspension were placed into 12 $\times$ 75-mm tubes for the flow cytometer. To stain the chromosomes, 20 μ L Chromomycin A3, 2 μ L Hoechst 33258 stain, and 20 μ L 100 mM magnesium sulfate were added to each tube.

Influx Setup for Chromosome Sorting. Chromosome sorting was performed on a BD Influx cell sorter. A 70- μ m nozzle was used with a sheath pressure of 40 psi. Excitation of Hoechst and Chromomycin was done with a solid state 100-mW, 355-nm laser and a 200-mW, 457-nm laser, respectively. Emission of the Hoechst fluorescence was collected with a 460/50-bandpass filter, and the Chromomycin fluorescence was collected with a 550/50-bandpass filter. The UV laser was timed as the first laser, and detection was triggered by Hoeschst fluorescence. For Phase-Seq analysis, individually sorted chromosomes were collected in a well of a low-profile 96-well unskirted PCR plate (Bio-Rad), and the plate was sealed and stored at –80 °C in a freezer before the experiment. The stream control was collected in 1.5-mL DNase/RNase-free Eppendorf tubes.

Single Chromosome Amplification. Twenty-eight single Chr19s, along with twelve controls (negative controls: no target, FACS stream; positive control: human total DNA), were separately amplified using the Picoplex WGA Kit (Rubicon Genomics) according to the manufacturer's protocol (version R30050-09). Amplified DNA was then column-purified. Using sequence-specific PCR primers, Chr19-specific DNA sequences were verified on the amplified single chromosome samples (except the low-yield ones) and positive controls but not on the negative controls.

Multiplex Illumina GAllx Sequencing. The adapter pair used in Illumina GAllx sequencing was synthesized by Elim Biopharmaceuticals. The adapter pair set consists of two designs. Design 5'-PE-Tag-Picoplex-3' (PETP), in which the standard Illumina PE1/PE2 adapter pair 5'-ACACTCTTCCCTACACGACGCTCTTCCGATCT-3' and 5'-CTCGGATTCCTGCTGAAC-CGCTCTTCCGATCT-3', each part adjoined in tandem with a 6-base multiplex tag (Table S3) (M. Matvienko, University of California, Davis, CA) and an 18-base Picoplex linker sequence, was introduced through PCR onto the ends of amplified single chromosome DNAs. This Picoplex linker served as a linker to adjoin the Illumina PE adapter/multiplex tag pair to the Picoplex amplified DNA, as in the case of standard Illumina PE adapters. A subset of experiments used an alternative adapter pair set design, adapter pair set 5'-PE-Nonamer-Tag-Picoplex-3' (PENTP), which introduced a stretch of nine random nucleotides (random nonamer) placed between the Illumina PE adapter sequence and the six-base multiplex tag.

Illumina PE sequencing libraries were then generated according to Illumina's standard protocol, except that, to preserve the single chromosome-amplified DNA, the libraries consist of a size distribution of 200–1,000 bases rather than being size-selected to a narrower range. Unincorporated adapters and adapter dimmers were removed from the libraries using the Agencourt AMPure XP system (Beckman Coulter). The six-base multiplex tag allowed 12 sequencing libraries to be combined into one pool and loaded onto a single lane on the Illumina GAllx system for single-read 101-base or 108-base runs (Ohio State University Nucleic Acid Shared Resource Facility). The Illumina sequencing short-read data are available in the National Center for Biotechnology Information (NCBI) short-read archive (<http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi>) (accession no. SRA026487.1).

Sequence Data Analysis. The image data from Illumina GAllx sequencing were analyzed using the next_phred software package. Briefly, image data were processed through an initial base-calling process aligned to the February 2009 human whole-genome reference sequence (GRCh37); the base-calling process was then calibrated, and cluster reads were obtained. Based on next_phred's base-calling quality score and excluding the adapter sequence, the reads were end-trimmed and aligned to the human whole-genome reference sequence, allowing a next_phred mapping quality score cutoff of 100 or 60 and individual base next_phred quality score cutoff of 15. See [SI Text](#) for alignment code.

Complete Phasing by Clustering of Single Chromosomes into Parental Allele Groups. Because amplified DNA from each single chromosome is absolutely in phase with each other, there are only two types of haploid chromosomes (i.e., two parental alleles for Chr19) among the 28 single Chr19-derived samples. Of 28 Chr19s, 9 gave rise to very low numbers of tagged reads. We removed these tags from all subsequent analyses. We determined the tag association for the sequencing reads with the remaining tags, keeping only reads with a next_phred read quality score average of 15 or higher on the six-base tags. To cluster the tags into two groups corresponding to the two parental alleles, we first asked, for each pair of tags, whether their reads at polymorphic sites are consistent with each other. For a given pair of tags, we examined a set of SNP sites on Chr19 and obtained a consistency index for the tag pair by computing the percentage among this set of sites at which we have identical majority vote from the tags. The majority vote from a tag at an SNP site is defined as the nucleotide (G, C, A, or T) with the highest read count from that tag. It is important that a reliable and unbiased set of SNP sites is used in the above computation of the pair-wise consistency indexes. To obtain this set of SNPs, we applied the de novo SNP calling program phastlane of the next_phred package to analyze the aligned reads. In total, phastlane output 12,326 putative SNPs, of which 7,709 are putative heterozygous SNPs with two or more base identities and 4,617 are putative homozygous SNPs with only one base identity, which is different from the reference sequence. Because the phastlane computation did not make use of the tag information, it did not induce any bias on the consistency indexes between pairs of tags. Therefore, if the two single chromosomes associated with the two tags are copies of the same parental allele of Chr19, then the pair-wise consistency index should be very close to one. However, if the two single chromosomes are copies of different parental alleles, then their consistency index should be much lower than one, because the consistency should be close to zero on heterozygous SNPs. We applied filtering criteria for these putative SNPs using an individual base quality score of 15 or higher and a next_phred mapping quality score of 100 or higher for the associated read. To further increase the reliability of clustering, in our actual computation of the pair-wise consistency index, we removed any phastlane-detected SNP positions that had less than three reads from either one of the tags before we computed the percentage of positions where the two tags agree on their majority vote. Table 1 presents the pair-wise consistency indexes for the 19 tags. Based on the table, we conclude with high confidence that tags 1, 3, 7, 8, 9, 10, 11, 14, 23, and 24 belong to one of the parental alleles and tags 2, 4, 5, 6, 12, 15, 16, 19, and 22 belong to the other parental allele.

Scanning of Heterozygous and Homozygous SNPs. Making use of the tag to parental allele association, we then scanned the whole Chr19 to obtain SNP positions. First, we filter out low-quality reads using the following criteria: next_phred mapping quality score of 60 or higher for the aligned read and individual base quality score of 15 or higher for the putative SNP. For each position in Chr19, define n_1 as the number of reads from allele 1 that support the most frequent base at this position, and m_1 as the number of reads that support the next most frequent base. Similarly, define n_2 and m_2 from allele 2. The position is called a heterozygous SNP position if (i) the most frequent base in allele 1 is different from that in allele 2, (ii) $n_1 \geq 4$, $n_1 - m_1 \geq 2$, and $n_1/m_1 \geq 1.5$, and (iii) $n_2 \geq 4$, $n_2 - m_2 \geq 2$, and $n_2/m_2 \geq 1.5$. It is called a homozygous SNP position if condition i is replaced by the condition that the most frequent base in allele 1 is identical to that in allele 2 but different from that in the reference genome. We note that this is a stringent calling procedure that requires at least four reads from each allele as well as additional conditions on the distribution of reads within the same allele. By this method, a total of 6,444 heterozygous SNPs and 3,694 homozygous SNPs were identified. To increase SNP detection sensitivity, we used all of the reads with next_phred alignment scores ≥ 60 (instead of 100) in the above scan. Because the consensus calls may not be as reliable as ones achieved based on reads with alignment scores ≥ 100 as in the computation of pair-wise consistency (Table S4), we attempted to assess the reliability of a consensus call for each allele at each detected SNP position by computing its consensus score, which is defined as the ratio of the number of reads supporting the consensus call to the total number of reads at that position from that allele. Fig. S1 shows the distribution of the consensus scores from both alleles and all heterozygous SNPs. The distribution seems to be bimodal, with a dividing value around 0.8. The calls with low consensus scores may be contaminated by poorly aligned reads. We, thus, removed any SNP positions where one or both consensus calls have low consensus score values (<0.8); 5,281 of 6,444 heterozygous SNPs and 3,159 of 3,694 homozygous SNPs remain after this filtering step. To calculate the Ti/Tv ratio, all SNPs from either parental allele that make the consensus score cutoff were compared against the reference genome. These SNPs spanning Chr19 were obtained with

a moderate coverage of Chr19 (~8x). Because the sequencing libraries were amplified from single molecules, there are technical issues such as amplification bias and library coverage, which should be investigated and alleviated in further optimization of the experimental protocols.

Detection of Indels. Sequencing reads were aligned to the reference sequence (GRCh37) using the phaser program of the next_phred package as described in the previous sections except that indels of up to three nucleotides are set to be allowed through the program parameters. Positions where the indels occur were identified, and SNP scanning of these positions was conducted similarly as in the whole-genome SNP scan described previously except that, to compensate for the increased leniency in alignments allowing indels, more stringent criteria are used in conditions (ii) $n1 \geq 5$, $n1 - m1 \geq 3$, and $n1/m1 \geq 1.5$, and (iii) $n2 \geq 5$, $n2 - m2 \geq 3$, and $n2/m2 \geq 1.5$. The numbers of parental allele-specific indels identified are shown in Table S5.

ACKNOWLEDGMENTS. We thank Phil Green (University of Washington, Seattle, WA) for making available the next_phred software package before publication, Marty Bigos (Stanford FACS Facility, Stanford, CA) for performing the FACS sorting of single chromosomes, Pearly Yan (Ohio State University, Columbus, OH) for supervising the Illumina GAIIx sequencing and advice, John Langmore (Rubicon Genomics, Ann Arbor, MI) for advice on single-cell DNA amplification, Marta Matvienko (University of California, Davis, CA) for sharing the multiplex Illumina sequencing design, Hui Jiang for running Seqmap alignments, and John Mu for running Splicemap alignments. We also thank Lyudmyla Khrapenko (Ohio State University, Columbus, OH) and Nickolas Johnson, Hui Jiang, David Hiller, John Mu, and Takao Kurihara (Rubicon Genomics, Ann Arbor, MI) for discussions. We thank H. Gulcin Ozer (Ohio State University, Columbus, OH) for running the Illumina sequence analysis pipeline, Balasubramanian Narasimhan for computation support. This research was supported by National Institutes of Health Grants R01HG004634 and R01HG003903 (to W.H.W.). The computation was performed on a system funded by Scientific Computing Research Environment Grant DMS0821823 from the National Science Foundation.

- Kong A, et al. (2009) Parental origin of sequence variants associated with complex diseases. *Nature* 462:868–874.
- Venter JC (2010) Multiple personal genomes await. *Nature* 464:676–677.
- International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437:1299–1320.
- Frazer KA, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861.
- Kong A, et al. (2008) Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* 40:1068–1075.
- Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68:978–989.
- Wetmur JG, et al. (2005) Molecular haplotyping by linking emulsion PCR: Analysis of paraoxonase 1 haplotypes and phenotypes. *Nucleic Acids Res* 33:2615–2619.
- Turner DJ, Tyler-Smith C, Hurles ME (2008) Long-range, high-throughput haplotype determination via haplotype-fusion PCR and ligation haplotyping. *Nucleic Acids Res* 36:e82.
- Konfortov BA, Bankier AT, Dear PH (2007) An efficient method for multi-locus molecular haplotyping. *Nucleic Acids Res* 35:e6.
- Xiao M, et al. (2009) Direct determination of haplotypes from single DNA molecules. *Nat Methods* 6:199–201.
- Zhang K, et al. (2006) Long-range polony haplotyping of individual human chromosome molecules. *Nat Genet* 38:382–387.
- Li HH, et al. (1988) Amplification and analysis of DNA sequences in single human sperm and diploid cells. *Nature* 335:414–417.
- Yan H, et al. (2000) Conversion of diploidy to haploidy. *Nature* 403:723–724.
- Ma L, et al. (2010) Direct determination of molecular haplotypes by chromosome microdissection. *Nat Methods* 7:299–301.
- Mardis ER (2008) Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 9:387–402.
- Carrano AV, Gray JW, Langlois RG, Burkhardt-Schultz KJ, Van Dilla MA (1979) Measurement and purification of human chromosomes by flow cytometry and sorting. *Proc Natl Acad Sci USA* 76:1382–1384.
- Levy S, et al. (2007) The diploid genome sequence of an individual human. *PLoS Biol* 5:e254.
- Pickrell JK, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464:768–772.
- Montgomery SB, et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464:773–777.
- Chen Y, Cicciarella J, Pravica V, Hutchinson IV (2009) Long-range linkage on chromosome 6p of VEGF, FKBP5, HLA and TNF alleles associated with transplant rejection. *Mol Immunol* 47:96–100.
- Winkler CA, Nelson GW, Smith MW (2010) Admixture mapping comes of age. *Annu Rev Genomics Hum Genet* 11:65–89.
- Kleinjan DA, van Heyningen V (2005) Long-range control of gene expression: Emerging mechanisms and disruption in disease. *Am J Hum Genet* 76:8–32.
- Sabeti PC, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.
- Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: Hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol* 20:R208–R215.
- Nunome M, et al. (2010) Detection of recombinant haplotypes in wild mice (*Mus musculus*) provides new insights into the origin of Japanese mice. *Mol Ecol* 19:2474–2489.
- Fantes JA, Green DK (1990) Human chromosome analysis and sorting. *Methods Mol Biol* 5:529–542.