Population split time estimation and X to autosome effective population size differences inferred using physically phased genomes

Shiya Song[1], Elzbieta Sliwerska[2], Jeffrey M. Kidd*[1,2]
[1]Department of Computational Medicine and Bioinformatics
[2]Department of Human Genetics
University of Michigan Medical School
Ann Arbor, Michigan, USA

Correspondence to J.M.K at jmkidd@umich.edu

June 30, 2014

Abstract

Haplotype resolved genome sequence information is of growing interest due to its applications in both population genetics and medical genetics. Here, we assess the ability to correctly reconstruct haplotype sequences using fosmid pooled sequencing and apply the sequences to explore historical population relationships. We resolved phased haplotypes of sample NA19240, a trio child from the Yoruba HapMap collection using pools of a total of 521,783 fosmid clones. We phased 93% of heterozygous SNPs into haplotype-resolved blocks, with an N50 size of 318kb. Using trio information from HapMap, we linked adjacent blocks together to form paternal and maternal alleles, producing near-to-complete haplotypes. Comparison with 33 individual fosmids sequenced using capillary sequencing shows that our reconstructed sequence haplotypes have a sequence error rate of 0.005%. Utilizing fosmid-phased haplotypes from a Yoruba, a European and a Gujarati sample, we analyzed population history and inferred population split times. We date the initial split between Yoruba and out of African populations to 90,000-100,000 years ago with substantial gene flow occurring until nearly 50,000 years ago, and obtain congruent results with the autosomes and the X chromosome. We estimate that the initial split between European and Gujarati population occurred around 45,000 years ago and gene flow ended around 28,000 years ago. Analysis of X vs autosome inferred effective population sizes reveals distinct epochs in which the ratio of the effective number of males to females changes. We find a period of female bias during the ancestral human lineage up to 1 million years ago and a short period of male bias in Yoruba lineage from 160-400 thousand years ago. We demonstrate the construction of haplotype sequences of sufficient completeness and accuracy for population genetic analysis. As experimental and analytic methods improve, these approaches will continue to shed new light on the history of populations.

Keywords

Fosmid pool sequencing, haplotype, population split-time, PSMC, X vs autosome

Introduction

DNA resequencing technologies have made it possible to identify genetic variation across thousands of individuals. However, most resequencing studies of this kind are "phase-insensitive", providing a mixed readout of diploid genomes that neglects the haplotype configuration of two homologous chromosomes. Haplotype information is key to understanding the relationships between genetic variation and phenotype, such as how *cis*-acting eQTL affecting gene expression and whether combinations of heterozygous variants lead to additional combined phenotypic effects (reviewed in [1]). Haplotypes are also informative for inferring genetic ancestry [2-5] and reconstructing population history [6-8]. Recent studies have also identified haplotypes of Neanderthal ancestry and examined how archaic introgression shaped our genomes [9-12].

Haplotypes are often inferred by statistical methods that utilize population genotype data to model the haplotype pairs of an individual as an imperfect mosaic of other haplotypes [13]. However, this approach is applicable to common SNPs rather than rare or individual-specific variants [14]. Trio-based phasing is more accurate but sometimes unavailable and is uncertain for the phase of variants when all individuals are heterozygous. Several experimental phasing methods are now available. Although full genome sequencing of individual haploid cells, such as sperm, is now possible [15,16], most approaches obtain sequence information from individual haplotypes by physically separating DNA fragments [17]. These approaches include use of clone end-sequence pairs [18-20], analysis of clone pools [21-23], or library creation from dilute pools of single molecules [24,25]. Approaches based on other concepts, such as the spatial structure of chromosomes in the nucleus [26], are also being developed.

Here, we construct long-range haplotypes by sequencing fosmid pools where each fosmid represents ~35kb of haplotype-specific sequence [22]. Heterozygous variant positions within overlapping clone fragments were then used to assemble contiguous haplotypes. Fosmid-based haplotyping can usually achieve N50 block greater than 300kb, depending on the density of heterozygous SNPs and the number of clones sequenced [27]. Additional information from SNPs phased by trio transmission can further link blocks together, producing near-to-complete haplotypes. The haplotype of a Gujarati Indian Individual (NA20847; GIH) was first resolved by using this method [22], followed by NA12878, a HapMap trio child from the CEU population [23]. Here, we describe the physically phased genomes of NA19240, from the YRI population. We assess the ability to correctly assemble SNP haplotypes and to accurately integrate other types of sequence variation into a coherent picture of diploid genomic structure.

Utilizing phased information from individuals from three different populations, we infer population history and population split times based on patterns of haplotype coalescence using the Pairwise Sequentially Markovian Coalescent (PSMC) model [28]. The PSMC approach is based on the distribution of coalescence times, or time to the most recent common ancestor (TMRCA) for hundreds of thousands of loci along a chromosome that have independent histories due to historic recombination events. Typically, these TMRCA values are calculated for the two haploid genomes which compose the diploid genome of a single sample. Utilizing haploid genomes from distinct individuals, the method infers TMRCA distributions across populations, which is informative about the timing of population splits. This information is complementary to that used in other methods, such as inference based on the site frequency spectrum [29,30], and permits direct comparisons of inferred estimates on the autosomes and X chromosome.

Comparative analyses of genetic diversity between the autosomes and the X chromosome provides estimates of the ratio of effective population size (N_E) of X chromosome to autosome, with the expectation that this ratio should be ¾ given equal male and female effective population size and constant population size throughout history. However, several studies have observed a larger X-vs-autosome diversity ratio in African populations, indicating female bias (larger female effective population size) [31-33]. Deviation from the expected ratio of ¾ may result from sex-biased demographic events [34] or skewed breeding ratio [35] leading to different Ne of males and females, as well as other processes such as natural selection acting differentially on X and autosomes [36]. When applying PSMC on autosomes and X chromosome, we can obtain direct estimates of X and autosome effective population size at each time point and infer time period for potentially sex-biases.

Results

Fosmid Pool Sequencing

We constructed two independent fosmid libraries from sample NA19240, split into a total of 288 pools each containing ~2,000 independent clones (Supplemental Table 1). Barcoded sequencing libraries were constructed form each pool using the Nextera sample library generation protocol [37] and sequenced to an average depth of ~3 across six lanes of an Illumina HiSeq. (Supplemental Table 1). We mapped the resulting reads to a version of the human genome reference sequence that included the GrCh37 reference as well as sequence from the *E. coli* genome, the fosmid vector backbone, and Epstein Barr Virus (EBV). To define fosmid clones, we identified islands of increased coverage in contiguous 1 kb windows filtered by length (10kb to 50kb) and read depth(above 0.25). This procedure identified 521,783 individual clones (median length 34.0kb), with median 1kb coverage of 1.95. In total across all pools, 96.4% of the genome was covered by at least one clone, with an average physical coverage of 5 clones and a median sequence coverage of 17.9x (Supplemental Figure 1)

Haplotype Resolution

We identified 2,405,813 heterozygous SNPs in NA19240 based on 15x whole genome sequencing from a standard and Nextera-based whole genome library. Heterozygous SNPs showed a concordance of 99.2% with 1000 Genomes [38] and 99.5% with HapMap [39]. Non-reference sensitivity for all SNPs was 88.4% compared with HapMap variants, and 87.1% compared with 1000 Genomes, while heterozygous SNPs showed a concordance of 99.98% with 1000 Genomes and 99.5% with HapMap (Supplemental Table 2). The genotype inferred for each of these heterozygous SNP positions in each of the inferred clones served as the input for reconstructing phased haplotypes. We utilized the ReFHap algorithm [27], previously demonstrated to have superior performance on this type of data, and obtained 17,388 haplotype-resolved blocks, with N90 block size of 71.2kbp, an N50 of 318.5kbp and an N10 of 1.04 mbp, and phased 92.84% of the 2,405,813 heterozygous SNP genotyped by whole-genome sequencing. For comparison, we reanalyzed published data from sample NA20847 by mapping whole genome and fosmid pool sequence data to the GRCh37 assembly and inferring phase using this pipeline (Figure 1). Results are similar to those previously obtained, with NA19240 having longer phased blocks due to increased density of heterozygous sites in African relative to GIH samples [40].

Although SNPs within each block are phased, the relationships between blocks cannot be directly established due to the absence of linking fosmid clones. We used using phase information determined by transmission at informative sites in the HapMap trio data for NA19240 to overcome this limitation [41]. Using this data, we assigned our ReFHap haplotypes within each block to the paternal and maternal derived chromosomes. This linked adjacent blocks together producing near-to-complete haplotypes. In total, 92.8% of blocks were successfully assigned to a parental allele. Comparison with the SNPs deterministically phaed using HapMap trio data identified 506 switch errors within our inferred haplotypes, which we corrected prior to subsequent analysis (Table 1). A switch error is an inconsistency between an assembled haplotype and the real haplotype between two contiguous variants. Examination of the errors indicates that most of the switch errors are due to insufficient clone support when linking variants together as ReFHap will assemble variants even when there is a single clone overlapping two variants (Supplemental Figure 2). For GIH sample NA20847, we linked together phased haplotypes within ReFHap blocks based on the statistically inferred haplotypes from the HapMap project [40], and randomly assigned haplotypes for those blocks without an overlap to an informative SNP. This yielded 2,751 apparent switch-errors, which we infer to be mostly errors in the haplotypes inferred statistically [22]. To further assess our overall haplotype accuracy, we compared our phased SNPs for NA19240 with 1000 Genomes heterozygous SNPs phased based on trio transmission [38], and found 99.7% concordance (Table 1).

Based on the reconstructed SNP haplotypes, we assigned individual clones to the paternal or maternal alleles. This procedure assigned 420,637 clones, while 5,376 clones could not be assigned because some RefHap blocks contained no HapMap or 1000 Genomes informative SNPs. We created haplotype-specific BAMs by extracting the reads corresponding to each fosmid clone. From these BAMs, we recreated haplotype-specific sequences, resulting in the identification of 186,507 additional heterozygous SNPs missing from our original heterozygous SNP call set. Among the additional SNPs, 76,726 (41.1%) were initially identified by our 15x whole genome sequence data, but were filtered by the Variant Quality Score Recalibration procedure implemented in GATK, and 82,136 SNPs (44% of the additional SNPs) overlapped with the 1000 Genomes call set.

To assess the sequence accuracy of inferred haplotypes we compared the haplotypes inferred from the haplotype specific BAMs with the sequence of 33 fosmid clones from the same individual that were previously sequenced using standard capillary sequencing [42]. Based on heterozygous SNPs, we assigned each finished clone sequence to the maternal or paternal haplotype, and compared the clone sequence with that inferred from our data (Supplemental Table 3). Our inferred sequence differs at 13 of the 1,107 heterozygous sites (1.1%) encompassed by the 33 clones. In total, the aligned clones encompass 1,144,737 bp excluding alignment gaps, and have 56 single nucleotide differences in comparison with our data. If we assume that all of these differences are errors in our inferred sequences, this suggests that our haplotypes have an overall sequence error rate of less than 0.005% or a Phred [43] quality greater than Q40.

Comparison with other data sets

Comparisons among sequence call sets continue to have substantial disagreement, an observation that likely result from differences in filtering, the amount of the genome accessible to different technologies, and algorithmic differences in read mapping and variant identification

[44,45]. Since sample NA19240 has been studied by a range of technologies, we compared our fosmid-derived call set with previously published data on the same sample [38,39]. We specifically focus on heterozygous SNPs, as they are more error-prone and important for our further analysis. We compared heterozygous SNPs from the 1000 genomes project, Complete Genomics, our call set based on WGS and fosmid pool sequencing, and high-density SNP array data from the Affymetrix Axiom array (Figure2A). Although most variants were identified by multiple platforms, there are a substantial number of calls unique to a single approach. For example, there are 236,613 heterzygous SNPs identified by Complete Genomics that are not identified by the others; when variants reported as low quality are excluded, this number is still 198,988. Our fosmid haplotype calls also include 128,329 heterozygous SNPs not reported by other studies. Of these, 85,119 are also identified by our 15X whole genome sequencing.

We also performed indel discovery using GATK. We intersected calls from whole genome sequencing, fosmid pool analysis, as well as the 1000 Genomes. We required sites to have the same position and also the same reference and alternative alleles, and identified 147,222 indels shared by all three datasets (Figure2B). We additionally determined indel phase from our fosmid data and intercepted it with indels predicted by 1000 Genomes Project and phased based on trio transmission. In total, we successfully phased 383,480 indels, of which 54.9% are heterozygous, and 50.3% were also confirmed by 1000 Genomes on the same individual (Supplemental Figure 4, Supplemental Table 4). This comparatively lower overlap among indel call sets highlights the continued challenges of current approaches for accurate indel calling.

Population split time estimation

We utilized phase-resolved genome sequence data from three individuals to analyze the split-times of three human populations. This includes two individuals, NA19240 from the YRI population and NA20847 from the GIH population, constructed using the above approach as well as phased haplotypes for NA12878 from the CEU population, previously published using a similar approach based on AB SoliD Sequencing of fosmid pools [27]. The PSMC method can infer changes in population size over time by the distribution of TMRCA for the two chromosomes within an individual [28]. We first compared PSMC curves based on our constructed haplotypes for NA19240 with PSMC curves obtained from 1000 Genomes sequencing data and Complete Genomics sequencing data. Our constructed haplotypes recovered nearly the same history as that obtained from 1000 Genomes sequencing data while Complete Genomics gave a slightly increased curve due to higher heterozygous variants called in their sequences (Supplemental Figure 5). Interpretation of coalescence times requires a calibration of mutation rates and generation times. For all subsequent analysis, we assumed a human mutation rate of $1.2\times10-8$ bp per generation and 25 years as generation time, although results can be easily rescaled for comparison with other estimates [46]. The PSMC curves of three individuals revealed that all populations shared same N_e history prior to 300 kyr ago, as previously reported [28,47]. The inferred Ne of YRI population began to differentiate from non-African populations around 240 kyr ago and experienced a milder bottleneck and recovered earlier than non-African populations (Figure 3A), although we note that the simulations indicate that such shits in PSMC curves may overestimate the timing of population size changes [11,28]. CEU and GIH populations shared a similar history before 30 kyr ago. Such observations were equivalent to

previous PSMC analysis on diploid genomes after adjusting for differences in assumed mutation rate [28].

If population splits are total and sudden, no coalescent events will happen after the split time. When applying PSMC on a pseudo-diploid individual where one chromosome comes from one population and the second chromosome comes from another population, the time when the PSMC estimate of N_e goes to infinity provides an estimate for the population split time [28]. To test the performance of PSCM on split-time estimation we first applied the method on sequences simulated using the model of human population history inferred by Gravel et al. [30]. As expected, for simulations with no migration after the split times, the PSMC curve goes to infinity directly after the split event (Supplemental Figure 6C-D). However, when migration continues after the after the split, the PSMC curves goes to infinity in a step-wise manner, with the timing of the initial increase approximating the initial split event. To further explore if PSMC can infer the time when gene flow ceases, we simulated sequences following Gravel's model but with gene flow ending at different times. Instead of relying on the time when inferred Ne goes to infinity, we examine the tail of the estimated TMRCA distribution (Supplemental Figure 7A). Even for a history with no migration after the split, the TMRCA distribution includes some coalescence events after the split-time, reflecting stochastic effects in the inference procedure. However, we observe that the point after which less than 0.1% of coalescence events occur approximates the end of gene flow (Supplemental Figure 7A).

Next, we applied this method to real data using single haploid genomes from two different populations. The resulting PSMC curves do not go to infinity immediately, but increase by steps, suggesting substantial gene flow following the initial split. Using the time when PSMC curves from pseudo-diploids first increase as an estimate for initial split event, we estimate African and non-African population initial split times of 92.7-97.0 thousand years ago, while European and GIH populations initially split more recently, around 44.5-46.5 thousand years ago (Figure 3B-D, Supplemental Figure 5B-D). We then looked at the tail of the inferred TMRCA distributions (Supplemental Figure 7B) and used the time when the tail percentage is less than 0.1% as an estimate of when gene flow ended. We estimated that African and non-African population had substantial gene flow until 50.9-53.4 thousand years ago, and detected gene flow between CEU and GIH populations ending 27.2-28.4 thousand years ago. We also applied these approaches to pseudo-diploid X chromosomes, scaling the effective population size by ¾ and using a male-to-female mutation rate ratio α of 2 to account for mutation rate differences between males and females. We found that in general the split-time estimates inferred for the X chromosome are similar to those observed for the autosomes. We obtained 95% confidence intervals by performing PSMC on bootstrapped sequences sampled 100 times and summarized the result in Table 2. Our results suggest a long period of genetic exchange after the initial split between the Yoruba and non-African ancestors that began 100,000 years ago, and lasted for over 50,000 years, with less than 0.1% of the genomes coalescing more recently than that time. However, for the CEU and GIH populations, this process was much faster, with only 17,000 years of detected genetic exchange. We note that our results using phase-resolved genomes are broadly consistent with those obtained by a recently described extension of the PSMC model applicable to multiple samples, although our use of only two haplotypes limits our ability to resolve population sizes more recent times [48].

X to autosome population history differences

If the effective population size of males and females is equal, scaling the N_e inferred from X chromosomes by 0.75 and scaling the mutation rate according to the male-to-female mutation rate ratio α, will yield a similar population history as that inferred from autosomes. Using 6 YRI and 4 CEU individuals sequenced by Complete Genomics [49], we compared populations histories inferred using the X chromosome and the autosomes. We calculated values for Q, the ratio of effective population size of X chromosome to autosome for each time interval. We observe a strong female bias in the ancestral human lineage from 2.4 million years ago to 1 million years ago with a mean Q value of 0.848, significantly higher than 0.75 ($p = 2.2*10^{-16}$). We also observed a period of male bias in the YRI lineage for 1,000 generations starting 400 ky ago with mean Q value 0.604, significantly lower than 0.75 ($p =1.8*10^{-9}$). However, the CEU lineage does not appear to have male bias during this period, with mean Q value 0.76 ($p = 0.78$) (Figure 4A-B).

Discussion

Haplotype information is essential for population genetics analysis, such as ancestry mapping, population structure inference [2], and detection of signals of natural selection [50]. Phased haplotypes from modern humans have been used to find archaic introgression of Neanderthals and to compare the proportion of introgression among different populations [11]. When adding full haplotype information for individuals from different populations, we can date historic events regarding two populations by the distribution of TMRCA of two alleles, one from each population, as has been done to date the split time of the Neanderthal and modern human lineages [11]. Here, we utilized this method to date the split event between Africans and Non-Africans and between CEU and GIH. Our results indicate that the separation of the studied human populations was a gradual event, with substantial genetic exchange continuing after an initial split, a finding consistent with hypotheses of long-standing ancient population structure in Africa (reviewed in [51,52]). Our estimates of initial split times are broadly consistent with those obtained using other methods, after adjusting for mutation rate differences (Supplemental Table 5). Our observations are also comparable with recently published paper extending PSMC to multiple individuals, allowing population history inference more recent than 30,000 years ago and split-time estimates based on cross-population coalescence rates [48]. However, Schiffels and Durbin infer that YRI-non-African population separation took place over a longer time period (roughly 100,000 years), and also observe a longer period of exchange between the CEU and GIH populations. As new methods are developed, and additional experimentally phased genomes become available, we expect that the inferred picture of human population history to become increasingly elaborate and accurate.

Several studies have focused on comparing the genetic diversity on the X chromosome and autosomes to learn about differences in the demographic histories of males and females [31,32,53-55]. A recent study evaluating 569 females across 14 populations found that X/A diversity was in the range 0.76-0.77 for the African continental group, and 0.64-0.65 for five populations of European ancestry [32]. Summary statistics based on diversity levels and population frequency differences are sensitive to sex bias at different time scales, and it has been proposed that female bias along the ancestral human lineage and a male bias before the split between European and Asian populations can explain some seemingly contradictory observations

for the X/A effective population size ratio [33]. Another study comparing relative recombination rates on the X and autosomes also supported female bias in all three HapMap populations [35,56]. Our observation also supports a female bias in the ancestral human lineage from 2.4 million years ago to 1 million years ago. However, we also observe a period of male bias in YRI lineage for 1,000 generations from 400ky ago, absent in CEU lineage.

PSMC analysis gives a picture of sex bias at different time periods. To reconcile these observations, we determined the fraction of the X chromosome with coalescence times during the distinct epochs of sex-bias. For the YRI, we find that a higher proportion of the X chromosome sequences coalesced during the time-span of female bias than during the male bias time-span (33% and 14%, $p = 2.9*10^{-7}$), which would lead to an overall observation of female-bias (Supplemental Figure 8A). For the CEU population, 48% of X chromosome sequences have coalesced by the time of initial African-non-African separation (around 100ky ago), and, only 22% of X chromosome sequence has a coalescence time during the female bias epoch, a proportion significantly smaller than that of the YRI population ($p = 5.8*10^{-5}$) (Supplemental Figure 8B). We also note that greater α values exaggerate the female bias in the ancestral human lineage, but do not eliminate the male bias observed for YRI population (Supplemental Figure 9A-B). Our result supports an ancestral female bias in the human lineage [33,35]. The observed male bias in the YRI lineage may not have been found by other studies because the proportion of extant genomes that coalesce at this time interval is quite small and when considering the overall X/A diversity ratio, this signal from a small proportion of the genome can be easily diluted. We also believe that the observed differences are not entirely explained by the differential responses of X and autosomal diversity to ancestral population size changes [57]. When simulating population history and scale N_e by ¾ to mimic the effective population size for X chromosome, the inferred population history by X chromosome after scaling back ¾ recovers the autosomal history (Supplemental Figure 10A). The male bias in the YRI lineage could result from any population history in which the female effective population size decreased prior to that of males (Supplemental Figure 10B). However, as this phenomenon was not observed in a European population, during a time period when we infer that African and non-African populations are not separated. Other factors, such as sex-biased migration or natural selection acting differently on autosomes and X chromosomes in some populations may contribute to this observation. Our results highlight the importance of categorizing segments of genomes by coalescence time when comparing autosome and X chromosome instead of looking at a global X/A ratio. We expect that a more refined view of these historic processes will emerge as additional data from more populations, particularly in Africa, become available.

Materials and Methods

Whole genome SNP and indel genotyping

Two genome libraries were constructed based on the standard library preparation and the Illumina Nextera framgentase system following the manufacturers protocols and sequenced on Illumina Hi-Seq. Paired end reads were aligned to reference genome assembly (GRCh37, with the pseudoautosomal regions of the Y chromosome masked to 'N') using BWA v0.5.9-r16. PCR duplicates were removed by Picard v1.62. Reads in regions with known indels were locally realigned and base quality scores were recalibrated using GATK [58]. SNPs and indels were

called by GATK UnifiedGenotyper v2.3-9. We retained high-quality SNP positions by applying Variant Quality Score Recalibration implemented in GATK to select a SNP set that included 99% of sites that intersect with the HapMap, 1000 Genomes and dbSNP training set. The resulting heterozygous SNPs are the starting point for subsequent haplotype phasing.

Fosmid clone identification

Two fosmid libraries were constructed from genomic DNA obtained from Coriell for sample NA19240. Aliquots of 10 ug of DNA were sheared in 120 ul volumes on a Digilab Hydroshear for 60 cycles at a speed code of 16. Sheared DNA was loaded and ran on a pulse field gel at 200V for 26 hours with 0.5s-15s switching. DNA from 25kb-45kb was cut out of the gel and isolated by electroelution for 12 hours at 120 V. After electroelution, DNA was isolated with Ampure XP beads, end-repaired with the Epicentre End-It kit and ligated to the Epicentre pCC1Fos fosmid arms. The resulting ligation was packaged and transfected into the Phage T-1 Resistant EPI300-T1 *E. coli* plating strain (Catalog Number CCFOS110). One hour after transfection, the resulting cells were split into the appropriate volumes to give pools of 1,500-3,000 cells per pool. Barcoded Nextera libraries for sequencing were constructed from mini-prepped DNA from each pool and sequenced on Illumina HiSeq.

Reads were mapped to reference assembly including human genome (GRCh37/hg19), Epstein Barr virus, the *E. coli* genome and fosmid vector backbone using BWA v0.5.9-r16. Candidate fosmid clones were identified by computing read-depth in 1k bp windows for each clone pool and merging consecutive windows allowing a maximum gap of 3 windows. Reads where one end mapped to the fosmid vector backbone and another end mapped to human genome, called anchoring reads, were used to better assign clone breakpoints. Based on mechanism of fosmid library construction, anchoring read ends mapped to the left coordinate of the reference genome are always reverse oriented, and anchoring read ends mapped to the right coordinate of reference genome are always forward oriented. Observing anchoring reads in the middle of consecutive windows identified overlapping clones. Overlapping clones were excluded from downstream analysis.

Haplotype reconstruction

Each clone pool was separately genotyped at heterozygous SNPs called by whole genome shotgun sequencing using GATK UnifiedGenotyper v2.3-9. Clones covering one or more heterozygous SNP positions shown in whole genome were used to resolve haplotype in next stage. A small proportion of clones (8.1%) were genotyped as heterozygous, probably resulted from overlapping clones or mapping errors. These clones were excluded from further analysis.

We applied ReFHap [27], an efficient algorithm for Single Individual Haplotyping. This algorithm only takes clone fragments that contain at least one heterozygous SNP and seeks to bipartite clone fragments into two sets that maximize the difference between two sets. The algorithm first builds a graph where fragments are linked upon sharing positions and a score is assigned on the edge indicating how different two fragments are. RefHap then applies a heuristic

algorithm to find the bipartition maximizing the cut, which can also be formulated into an NP-hard Max-CUT problem. Finally, the algorithm generates consensus haplotype within one partition and flips every allele to get the other haplotype. However, if a site is observed equally in each partition, it will remain undecided resulting in gaps in the haplotype.

NA19240 was the child of a trio that HapMap had genotyped, thus phasing information is determined at genotyped SNPs except for cases where all individuals in the trio are heterozygous. We used phase-determined SNPs to guide paternal and maternal allele assignment within blocks. In order to differentiate switch errors and HapMap phasing errors, we first divide blocks into smaller blocks if there is limited supporting information from clones and then determine paternal and maternal allele based on the majority of HapMap phased SNP assignments. For sample NA20847, HapMap3 had phase prediction using a population-based statistical method. We compared our haplotypes with HapMap phase prediction to assign haplotype within blocks as either haplotype1 or haplotype2. However, we only assign haplotypes when a majority of SNPs agree, and tabulated switch errors between our haplotype and HapMap3 prediction.

Once we have assigned haplotypes within blocks to paternal and maternal allele (haplotype 1 and haplotype2 as in NA20847), we extracted original reads from each haplotype-assigned fosmid clone and merged them to create haplotype-specific BAMs. We use GATK's UnifiedGenotyper v2.3-9 with 'EMIT_ALL_SITES' option to make calls at each position for each haplotype. We filtered calls with MQ less than 20 and DP less than 5 or larger than 80 and within 5bp around a short indel. Only homozygous calls in each haplotype-specific BAM are kept, and we assign the inferred allele from the ReFHap if the call appears heterozygous. Next, we combined calls from the merged haplotype BAMs and whole genome sequencing BAM file, and also include BAM files from all other clones that were not incorporated due to a lack of heterozygous SNPs to capture regions that may not be covered in the whole genome sequencing.

For sample NA12878, we used the phased SNP haplotypes constructed by fosmid-pool sequencing from a previous study [27] , downloaded from http://www.molgen.mpg.de/~genetic-variation/SIH/data, together with the sequencing results from the 1000 Genomes Project, to construct full haplotypes. The distributions of callable regions of each chromosome for the three individuals are shown in Supplemental Figure 3.

Indel phasing

From haplotype-specific BAMs, we perform indel discovery using GATK. We preformed indel calling using calls from the whole genome indel call set as given alleles as well as calling on fosmid pools without using indel alleles from the whole genome data (*de novo*). We performed comparison on indel discovery as well as indel phasing.

PSMC

PSMC inference was performed as previously described [28]. On autosomal data, we use the default setting with T_{max}=15, n=64, and pattern '1*4+25*2+1*4+1*6'. For X-chromosome data, we set T_{max}=15, n=60, and pattern '1*6+2*4+1*3+13*2+1*3+2*4+1*6'. We also scale the population size inferred by chromosome X by ¾ and set the mutation rate of X chromosome as

$\mu X = \mu A \frac{2(2+\alpha)}{3(1+\alpha)}$ where α is the ratio of male-to-female mutation rate, range from 2 to 5. For X chromosome, we also excluded pseudo autosomal region for PSMC analysis because this region is diploid in males and has very high recombination rate and sequence diversity. We performed bootstrapping to measure the variance of estimates by splitting consensus sequence into 10Mbp segments and randomly resampling 300 segments with replacement and then rerunning PSMC on the resampled segments. Such bootstrapping was repeated 100 times to obtain 95% confidence intervals for the estimates. When applying PSMC on pseudo-diploid individuals to infer split time, we use the same pattern of the autosome for X chromosome, in order to give finer resolution at the times of interested to and enable easy comparison.

We simulated 100 10M sequences of three individuals (representing African, European and Asian) according to Gravel's model[30] using msHot software[59]. For X chromosomes, we scale the effective population size by ¾. We ran PSMC on simulated sequences, for each individual and pseudo-diploid individual. We also simulate a scenario in which the gene flow after the initial split of African and non African ends at 46ky, 60ky and 80ky ago.

We downloaded whole genome data for individuals sequenced by Complete Genomics from ftp2.completegenomics.com. These data were generated and analyzed with Complete Genomics' local de novo assembly-based pipeline [49]. We used the alignments against NCBI build 37 and data processed with pipeline version 1.10. We analyzed the CG 'masterVarBeta' files and selected single-nucleotide variants (SNVs) and masked 'no-call' regions as well as insertions, deletions, substitutions, or partial (half) calls. For each time interval, we calculated the ratio of effective population of X chromosome to autosome and test if it significantly deviates from 0.75.

Acknowledgments

We thank Jacob Kitzman, Peedikayil Thomas, Jeffrey W. Innis, and the University of Michigan DNA Sequencing Core Facility for guidance on fosmid pool construction and sequencing.

References

1. Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ (2011) The importance of phase information for human genomics. Nat Rev Genet 12: 215-223.
2. Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. PLoS Genet 8: e1002453.
3. Brisbin A, Bryc K, Byrnes J, Zakharia F, Omberg L, et al. (2012) PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Hum Biol 84: 343-364.
4. Sohn KA, Ghahramani Z, Xing EP (2012) Robust estimation of local genetic ancestry in admixed populations using a nonparametric Bayesian approach. Genetics 191: 1295-1308.
5. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, et al. (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet 5: e1000519.
6. Harris K, Nielsen R (2013) Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet 9: e1003521.

7. Palamara PF, Lencz T, Darvasi A, Pe'er I (2012) Length distributions of identity by descent reveal fine-scale demographic history. Am J Hum Genet 91: 809-822.
8. Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, et al. (2014) A genetic atlas of human admixture history. Science 343: 747-751.
9. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. (2010) A draft sequence of the Neandertal genome. Science 328: 710-722.
10. Sankararaman S, Mallick S, Dannemann M, Prufer K, Kelso J, et al. (2014) The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507: 354-357.
11. Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, et al. (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43-49.
12. Vernot B, Akey JM (2014) Resurrecting surviving Neandertal lineages from modern human genomes. Science 343: 1017-1021.
13. Browning SR, Browning BL (2011) Haplotype phasing: existing methods and new developments. Nat Rev Genet 12: 703-714.
14. Browning SR, Browning BL (2011) Haplotype phasing: existing methods and new developments. Nature Reviews Genetics 12: 703-714.
15. Wang J, Fan HC, Behr B, Quake SR (2012) Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. Cell 150: 402-412.
16. Kirkness EF, Grindberg RV, Yee-Greenbaum J, Marshall CR, Scherer SW, et al. (2013) Sequencing of isolated sperm cells for direct haplotyping of a human genome. Genome Res 23: 826-832.
17. Dear PH, Cook PR (1989) Happy mapping: a proposal for linkage mapping the human genome. Nucleic Acids Res 17: 6795-6807.
18. Lippert R, Schwartz R, Lancia G, Istrail S (2002) Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. Brief Bioinform 3: 23-31.
19. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. (2007) The Diploid Genome Sequence of an Individual Human. PLoS Biol 5: e254.
20. Kim JH, Waterman MS, Li LM (2007) Diploid genome reconstruction of Ciona intestinalis and comparative analysis with Ciona savignyi. Genome Res 17: 1101-1110.
21. Burgtorf C, Kepper P, Hoehe M, Schmitt C, Reinhardt R, et al. (2003) Clone-based systematic haplotyping (CSH): a procedure for physical haplotyping of whole genomes. Genome Res 13: 2717-2724.
22. Kitzman JO, Mackenzie AP, Adey A, Hiatt JB, Patwardhan RP, et al. (2011) Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nat Biotechnol 29: 59-63.
23. Suk EK, McEwen GK, Duitama J, Nowick K, Schulz S, et al. (2011) A comprehensively molecular haplotype-resolved genome of a European individual. Genome Res 21: 1672-1685.
24. Kaper F, Swamy S, Klotzle B, Munchel S, Cottrell J, et al. (2013) Whole-genome haplotyping by dilution, amplification, and sequencing. Proc Natl Acad Sci U S A 110: 5552-5557.
25. Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, et al. (2012) Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature 487: 190-195.
26. Selvaraj S, J RD, Bansal V, Ren B (2013) Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat Biotechnol 31: 1111-1118.
27. Duitama J, McEwen GK, Huebsch T, Palczewski S, Schulz S, et al. (2012) Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. Nucleic Acids Res 40: 2041-2053.
28. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475: 493-496.

29. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5: e1000695.
30. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, et al. (2011) Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 108: 11983-11988.
31. Gottipati S, Arbiza L, Siepel A, Clark AG, Keinan A (2011) Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. Nat Genet 43: 741-743.
32. Arbiza L, Gottipati S, Siepel A, Keinan A (2014) Contrasting X-Linked and Autosomal Diversity across 14 Human Populations. Am J Hum Genet 94: 827-844.
33. Emery LS, Felsenstein J, Akey JM (2010) Estimators of the human effective sex ratio detect sex biases on different timescales. Am J Hum Genet 87: 848-856.
34. Keinan A, Reich D (2010) Can a sex-biased human demography account for the reduced effective population size of chromosome X in non-Africans? Mol Biol Evol 27: 2312-2321.
35. Labuda D, Lefebvre JF, Nadeau P, Roy-Gagnon MH (2010) Female-to-male breeding ratio in modern humans-an analysis based on historical recombinations. Am J Hum Genet 86: 353-363.
36. Veeramah KR, Hammer MF (2014) The impact of whole-genome sequencing on the reconstruction of human population history. Nat Rev Genet 15: 149-162.
37. Adey A, Morrison HG, Asan, Xun X, Kitzman JO, et al. (2010) Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol 11: R119.
38. Consortium GP (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.
39. IHMC (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.
40. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58.
41. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-918.
42. Kidd JM, Cheng Z, Graves T, Fulton B, Wilson RK, et al. (2008) Haplotype sorting using human fosmid clone end-sequence pairs. Genome Res 18: 2016-2023.
43. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8: 186-194.
44. Lam HY, Clark MJ, Chen R, Natsoulis G, O'Huallachain M, et al. (2012) Performance comparison of whole-genome sequencing platforms. Nat Biotechnol 30: 78-82.
45. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, et al. (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 5: 28.
46. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. Nature 488: 471-475.
47. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222-226.
48. Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. Nat Genet advance online publication.
49. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, et al. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327: 78-81.

50. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72.
51. Henn BM, Cavalli-Sforza LL, Feldman MW (2012) The great human expansion. Proc Natl Acad Sci U S A 109: 17758-17764.
52. Harding RM, McVean G (2004) A structured ancestral population for the evolution of modern humans. Curr Opin Genet Dev 14: 667-674.
53. Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD (2008) Sex-biased evolutionary forces shape genomic patterns of human diversity. PLoS Genet 4: e1000202.
54. Keinan A, Mullikin JC, Patterson N, Reich D (2009) Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nat Genet 41: 66-70.
55. Hammer MF, Woerner AE, Mendez FL, Watkins JC, Cox MP, et al. (2010) The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. Nat Genet 42: 830-831.
56. Lohmueller KE, Degenhardt JD, Keinan A (2010) Sex-averaged recombination and mutation rates on the X chromosome: a comment on Labuda et al. Am J Hum Genet 86: 978-980; author reply 980-971.
57. Pool JE, Nielsen R (2007) Population size changes reshape genomic patterns of diversity. Evolution; international journal of organic evolution 61: 3001-3006.
58. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303.
59. Hellenthal G, Stephens M (2007) msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. Bioinformatics 23: 520-521.

Figure Legends

Figure 1. Haplotype assembly results. Half of the assembled sequence is present in blocks longer than 318.5 kbp for NA19240 and 378.7 kbp for NA20847.

Figure 2. Variants call set comparison with other data sets (A) Heterozygous SNP call set comparison (B) Indel call set comparison. Intersection refers to same site and same alleles. Fosmid call set refers to indel discovery using fosmid clones pool sequencing.

Figure 3. PSMC estimates on phased haplotypes. (A) Population history inferred by phased autosome and X chromosome sequences from three individuals. NA19240 represents YRI population, NA12878 represents CEU population, NA20847 represents GIH population. (B) PSMC result on pseudo-diploid genome of YRI and CEU. (C) PSMC result on pseudo-diploid genome of YRI and GIH. (D) PSMC result on pseudo-diploid genome of CEU and GIH. 1 stands for paternal allele and 2 stands for maternal allele when making pseudo-diploid genome. For the X chromosome, estimates are scaled by ¾ and assuming α value of 2. Bootstrapping was performed 100 times (green for autosomes, orange and light blue for X chromosomes).

Figure 4. X to autosome effective population size differences (A) YRI population. (B) CEU population. Q refers to ratio of effective population size of X chromosome to autosome. Q larger than 0.75 indicates female bias while Q smaller than 0.75 indicates male bias.

Supplemental Figure 1. Inferred clone statistics (A) Distribution of inferred clone insert sizes for two independent librarys (library1 with median 34kbp, library2 with median 31kbp) and distribution of 1kb window coverage of inferred clone (median 1.95). (B) Distribution of number of clones covered per 1kb window (median 5) and distribution of overall coverage per 1kb window (median 17.9x).

Supplemental Figure 2. Illustration of ReFHap's phasing result and a switch error. Each column corresponds to a SNP position, with blue indicating the reference allele and red the alternative. The first two rows are the haplotype prediction by ReFHap, followed by four rows showing HapMap phase based on trio transmission. This is followed by 12 rows depicting clone genotypes. The last row indicates the parental allele assigned for RefHap haplotype based on HapMap phasing. In the last row, blue indicates paternal allele and red indicates maternal allele. The line with a star shows where the switch error occurred.

Supplemental Figure 3. **Callable size of each chromosome of three individuals based on phasing results.**

Supplemental Figure 4. Venn diagram of indel phasing comparison. Intersection refers to indel calls with same site, allele and genotype. Fosmid (wgs) refers to calling indels from fosmid pools with whole genome sequencing indel call set as given alleles. Fosmid (de novo) refers to calling indels from fosmid pools without given alleles.

Supplemental Figure 5. PSMC comparisons. (A). PSMC estimate on NA19240 by three different sequencing methods. First two are haplotypes resolved by fosmid-pool sequencing. Second are from 1000 Genomes sequencing file. Third are from Complete-Genomics. (B) PSMC result on pseudo-diploid genome of YRI and CEU. (C) PSMC result on pseudo-diploid genome

of YRI and GIH. (D) PSMC result on pseudo-diploid genome of CEU and GIH. 1 stands for paternal allele and 2 stands for maternal allele when making pseudo-diploid genome. NA19240 represents YRI population, NA12878 represents CEU population, NA20847 represents GIH population.

Supplemental Figure 6. Simulation study on PSMC's split-time inference. We useed msHOT to simulate 3 individuals from African, European, and Asian population assuming Gravel's model. (A) simulated autosome sequences, with migration event. (B) simulated X chromosome sequences (multiply N value by 0.75), with migration event. (C) simulated autosome sequences, without migration event. (D) simulated X chromosome sequences (multiply N value by 0.75), without migration event. In the simulation, the European and Asian population experienced the same bottleneck event and then went through exponential growth with different rates. The African population size remained the same after an ancestral population growth.

Supplemental Figure 7. Inferred TMRCA distributions on cross-population pseudo-diploids (A) Tail TMRCA distribution for simulated data. (B) Tail TMRCA distribution for real data. We simulated 100 10M sequences of two individuals (representing African and European) according to Gravel's model. We set gene flow after the initial split of African and European ending at 46ky, 60ky and 80ky ago.

Supplemental Figure 8 Infered TMRCA distributions. (A) TMRCA distribution for autosomes and X chromosomes of YRI population (B) TMRCA distribution for autosomes and X chromosomes of the CEU population. Region 1 is the time interval of male bias found in YRI population and region 2 is the time interval of female bias found in ancestral lineage..

Supplemental Figure 9. X to autosome effective population size differences when setting α = 5. (A) YRI population. (B) CEU population. Q refers to ratio of effective population size of X chromosome to autosome. Q larger than 0.75 indicates female bias while Q smaller than 0.75 indicates male bias.

Supplemental Figure 10. (A) chrX can recover the same history after scaling back effective population size and time by ¾ when male and female effective population size is the same (B) When male and female effective population size is different, scaling by ¾ will give rise to different population history

Tables

Table 1 Phasing statistics.

Sample	#clones after filter	#blocks	MEC value	#SNPs to be phased	% phased SNPs	# blocks assigned by HapMap and 1000G	Switch Error	Switch Error rate
NA19240	521,783	17,388	38,903	2,405,813	92.84%	16,138	506	0.09%
NA20847	571,419	16,708	21,207	1,681,829	94.39%	8,718	2751	0.84%

MEC is the number of entries to correct when resolving haplotypes. Switch error is an inconsistency between an assembled haplotype and the real haplotype between two contiguous variants, normalized by number of variants for comparison.
99.7% of 1000G SNPs shown in our data is consistent with our phasing.
Switch Errors in NA19240 reflect switch error by ReFHap; while switch errors in NA20847 reflect relative switch error between HapMap's population based phasing method and ReFHap.

Table 2 Population Split-time estimation based on pseudo-diploid individuals.

	autosome				chrX (α = 2)			
	Initial Split		End of Gene flow		Initial Split		End of Gene flow	
	Time (ky)	tail percentage	Time (ky)	tail percentage	Time (ky)	tail percentage	Time (ky)	tail percentage
YRI-GIH	94.1	2.20%	51.7	0.04%	97.7	4.60%	49.3	0.03%
	(92.7-95.8)	(1.9%-2.4%)	(50.9-52.6)	(0.02%-0.06%)	(93.1-100.6)	(1.5%-6.3%)	(46.8-50.7)	(0.0%-0.12%)
YRI-CEU	95.6	1.80%	52.5	0.06%	92.6	4.70%	46.6	0.002%
	(94.0-97.0)	(1.6%-2.0%)	(51.7-53.4)	(0.02%-0.09%)	(89.2-95.8)	(3.7%-5.8%)	(44.9-48.3)	(0.000%-0.007%)
CEU-GIH	45.6	0.80%	27.9	0.0002%	45.2	3.80%	30.9	0.02%
	(44.5-46.5)	(0.5%-1.0%)	(27.2-28.4)	(0.000%-0.0006%)	(41.7-48.1)	(0.9%-7.2%)	(28.5-33.0)	(0.0002%-0.07%)

Estimates are in thousand years. α is the ratio of male-to-female mutation rates, which are used for time scaling on X chromosomes (See Methods).

Supplemental Table 2. Summary of variant calling for whole genome sequencing

	Pre Filter variants	Post Filter variants
SNP		
Total	4,495,127	3,798,082
Homozygous	1,561,566	1,392,269
Heterozygous	2,933,561	2,405,813
In dbSNP132	4,178,217	3,731,358
Ti/Tv Ratio		2.1
heterozygous sensitivity compared with HapMap	90.00%	88.40%
heterozygous sensitivity compared with 1000G	89.60%	87.10%
genotype concordance with HapMap	99.40%	99.50%
genotype concordance with 1000G	99.20%	99.20%
Heterozygous SNP genotype concordance with HapMap	99.49%	99.50%
Heterozygous SNP genotype concordance with 1000G	99.98%	99.98%
Coding variants		
synonymous		12,980
non-synonymous		11,002
stop-gain		72
stop-loss		12
Short Indels		
Total	429,501	
In dbSNP132	334,097	

Supplemental Table 3. Comparison of our haplotypes with the sequence of 33 fosmid clones from the same individual that were previously sequenced using standard capillary sequencing(hap1 refers to paternal allele, hap2 refers to maternal allele)

clone_name	chr	pos1	pos2	strand	# het SNP	Het mismatch	All mismatch	Length of clone	error rate	haplotype assignment
AC203596	20	60332560	60367311	-	63	0	2	33701	0.000	hap2
AC208180	7	109124048	109164730	-	6	0	17	35768	0.000	hap1
AC203618	14	24625251	24665998	+	43	0	0	40532	0.000	hap1
AC203625	3	13175578	13207238	+	52	0	1	31376	0.000	hap2
AC203613	17	42388348	42422435	+	11	0	0	33208	0.000	hap2
AC209301	5	103498132	103533083	+	42	1	5	30147	0.024	hap2
AC211777	2	131597914	131644819	+	57	0	1	40161	0.000	hap2
AC207436	20	61771673	61805848	-	25	0	1	33436	0.000	hap2
AC203629	2	27532275	27572319	+	27	0	0	39746	0.000	hap1
AC203601	13	84295142	84330569	+	21	7	10	31171	0.333	hap1
AC214990	20	1835027	1870122	+	32	0	0	34822	0.000	hap2
AC203623	13	50529642	50571393	-	19	0	0	40523	0.000	hap2
AC209312	19	11076696	11117117	+	9	0	0	39917	0.000	hap1
AC204964	20	36206659	36240122	-	50	0	5	33213	0.000	hap2
AC203663	12	120664025	120704371	-	11	0	0	39941	0.000	hap1
AC203633	15	83719658	83761946	+	30	1	1	40789	0.033	hap2
AC207998	5	42518686	42548692	+	14	0	2	28952	0.000	hap1
AC204962	10	73293352	73327357	-	2	0	0	33966	0.000	hap2
AC203585	12	127683544	127718788	-	26	0	1	33767	0.000	hap1
AC207584	22	24202000	24237415	-	48	0	0	35168	0.000	hap1
AC203595	17	15129257	15154073	-	25	0	0	24341	0.000	hap1
AC204968	12	108020551	108054912	-	44	1	1	33763	0.023	hap1
AC214217	20	34745579	34780635	-	13	0	0	34920	0.000	hap1
AC203614	13	27271316	27306555	-	28	0	0	32959	0.000	hap1
AC226164	17	8937695	8978811	+	46	0	0	40616	0.000	hap2
AC203609	17	9512088	9545453	+	22	0	0	27053	0.000	hap2
AC213115	2	5679730	5713339	-	21	1	1	32705	0.048	hap1
AC210876	17	3897067	3932032	-	52	0	1	34584	0.000	hap2
AC215711	8	142371608	142412790	-	74	0	2	40842	0.000	hap2
AC207992	4	162086171	162119199	+	31	0	1	30427	0.000	hap2
AC204957	7	135193786	135227398	+	26	2	2	32900	0.077	hap2
AC209156	12	104071670	104106822	+	15	0	0	34775	0.000	hap1
AC208068	16	85067617	85102315	-	122	0	2	34548	0.000	hap2

A total of 13 out of 1107 heterozygous SNP error occurred when assigning 33 clones into haplotype, an error rate of 1.2%. 56 substitution errors out of 1,144,737 bp total sequences yielded a sequence error rate of 0.005%.

Supplemental Table 4. Phased indel call set comparison. Fosmid (wgs) refers to calling indels from fosmid pools with guidance of whole genome sequencing indel call set. Fosmid (de novo) refers to calling indels from fosmid pools directly.

	1000 Genomes	fosmid_with_wgs	fosmid_denovo	fosmid_combined
total	363092	310296	241878	383480
homozygous	136415	156731	102018	173116
paternal indel	112505	83904	68425	111449
maternal indel	106832	68292	68399	95727
biallelic indel	7340	1369	3036	3188

Supplemental Table 5. Split time estimation from previous studies. Reported estimates are adjusted by using the same mutation rate $1.2*10^{-8}$ bp/generation.

Paper	Method	African & Non-African split time	African & Non-African split time (adjusted)	European-Asian split time	European-Asian split time (adjusted)
Gravel et al, 2011	diffusion, AFS	51,000	100,300	23,000	45,233
Gronau et al, 2011	Bayesian coalescent based	47,000	78,333	36,000	60,000
Harris et al, 2013	IBS sharing	55,000	88,000		

Figure 1

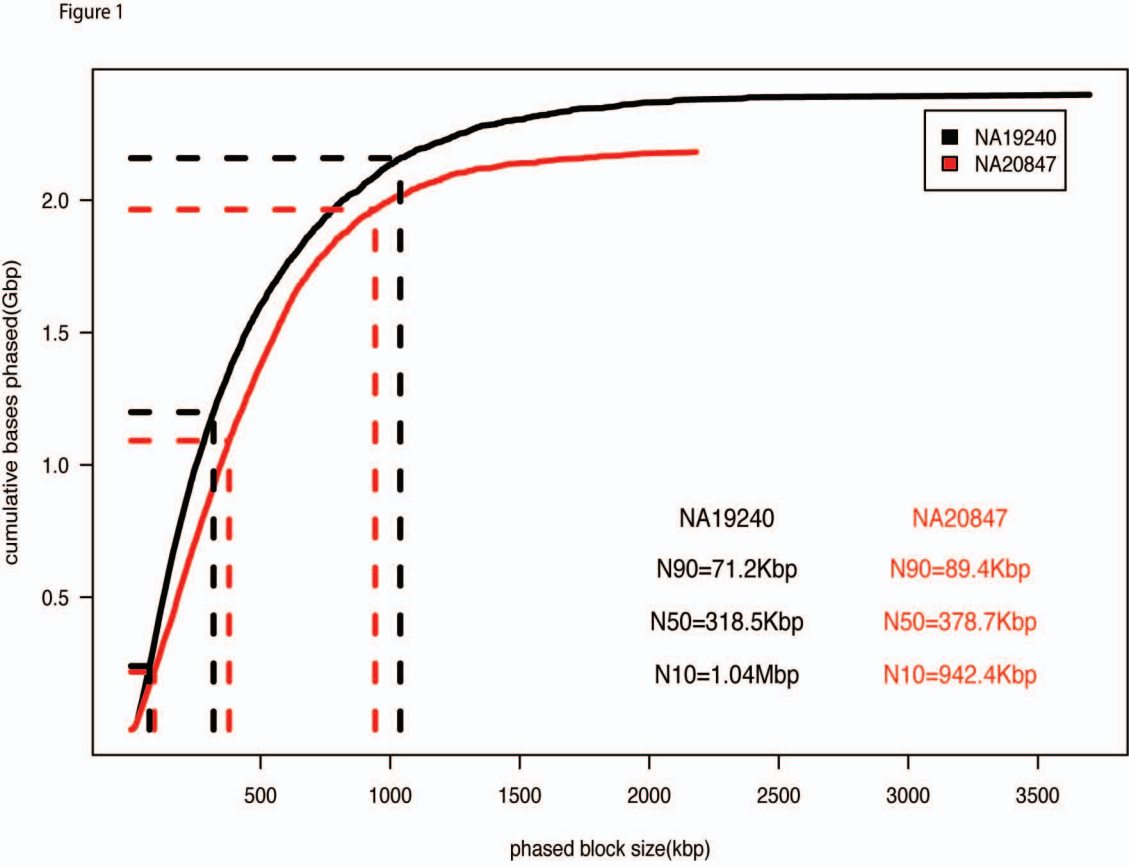
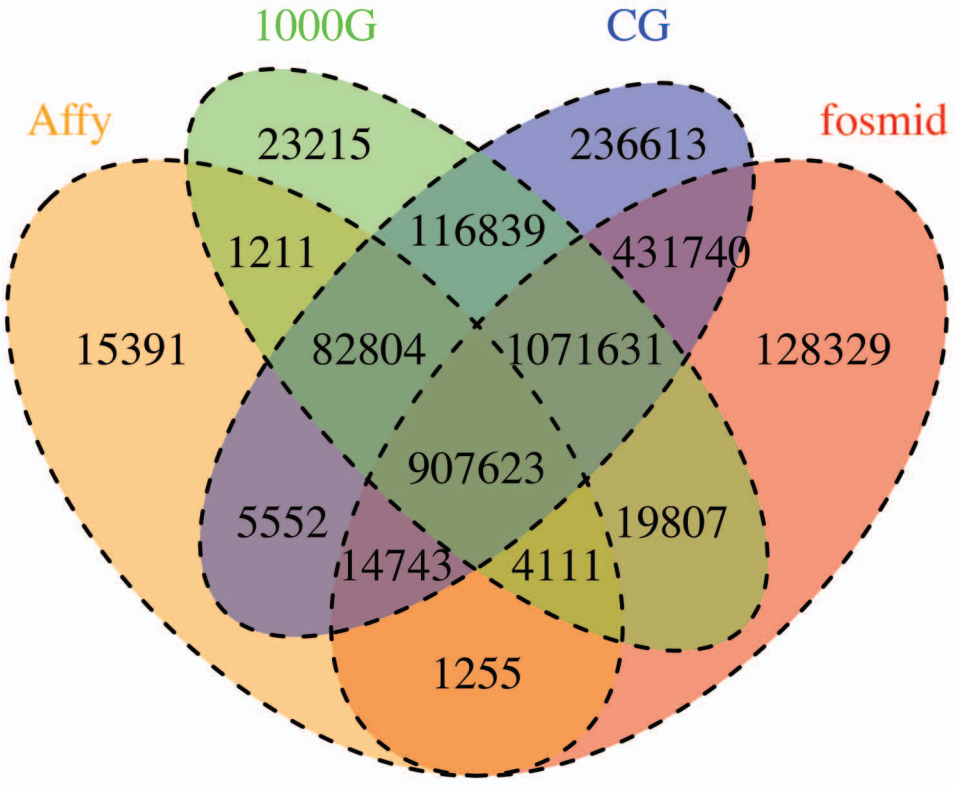

NA19240
NA20847
cumulative bases phased(Gbp)
2.0
1.5
1.0
0.5
NA19240
N90=71.2Kbp
N50=318.5Kbp
N10=1.04Mbp
NA20847
N90=89.4Kbp
N50=378.7Kbp
N10=942.4Kbp
500
1000
1500
2000
2500
3000
3500
phased block size(kbp)

Figure2

A



B

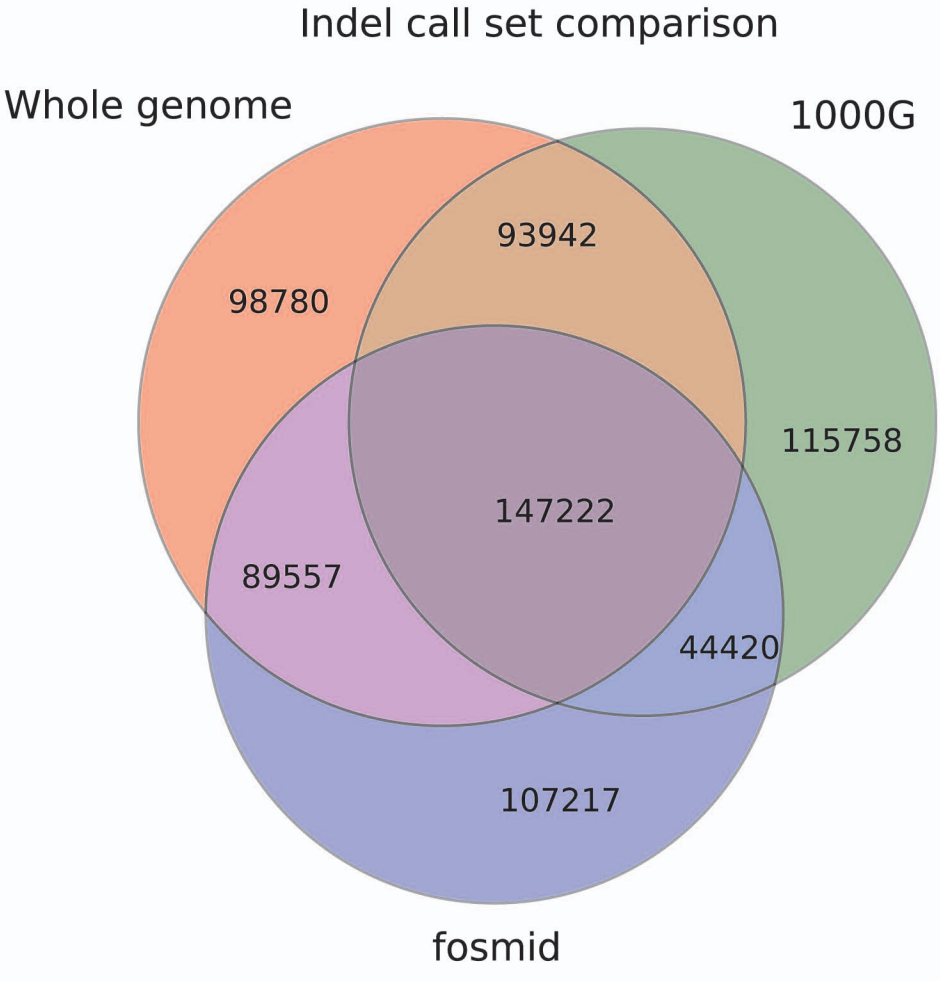


Figure3

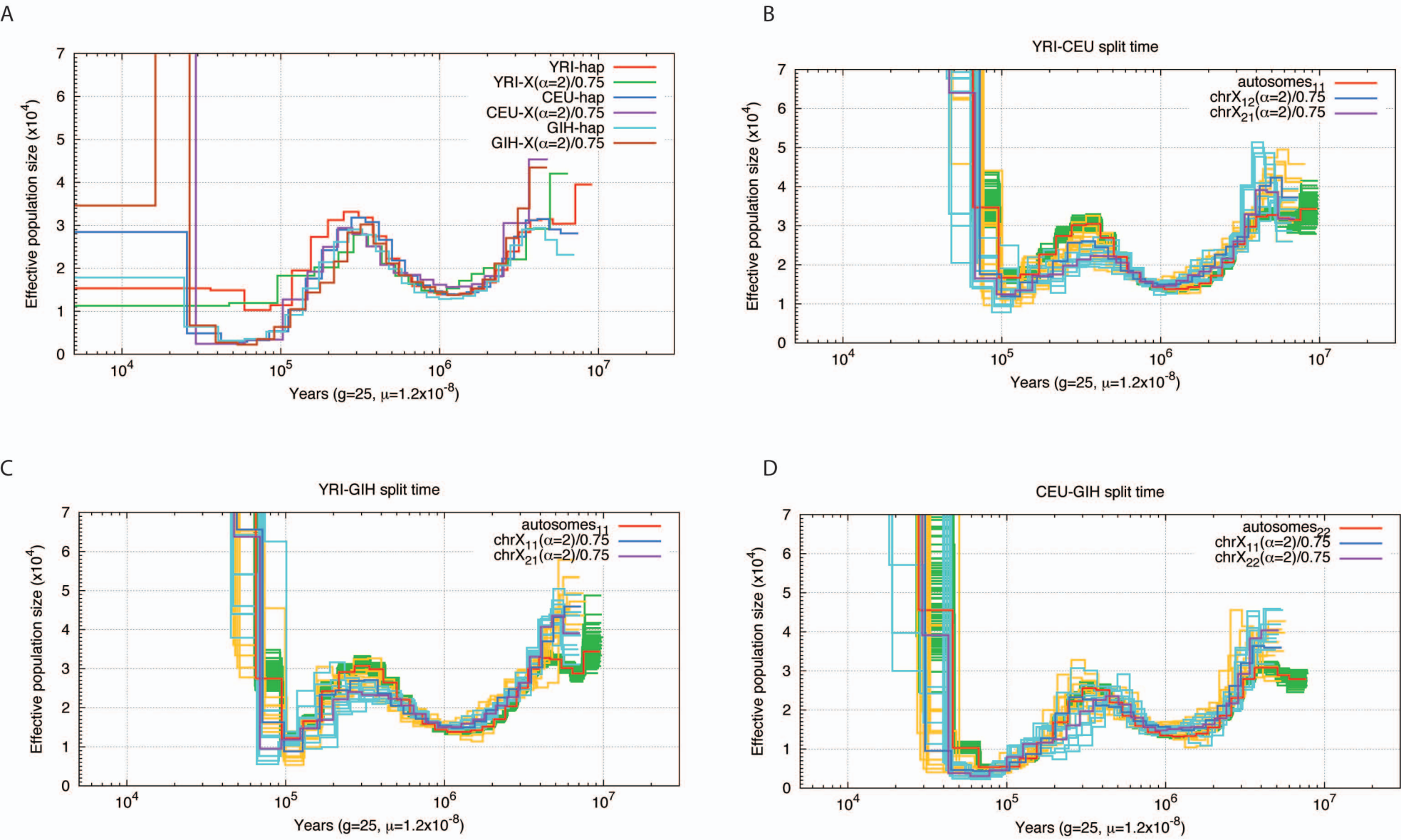
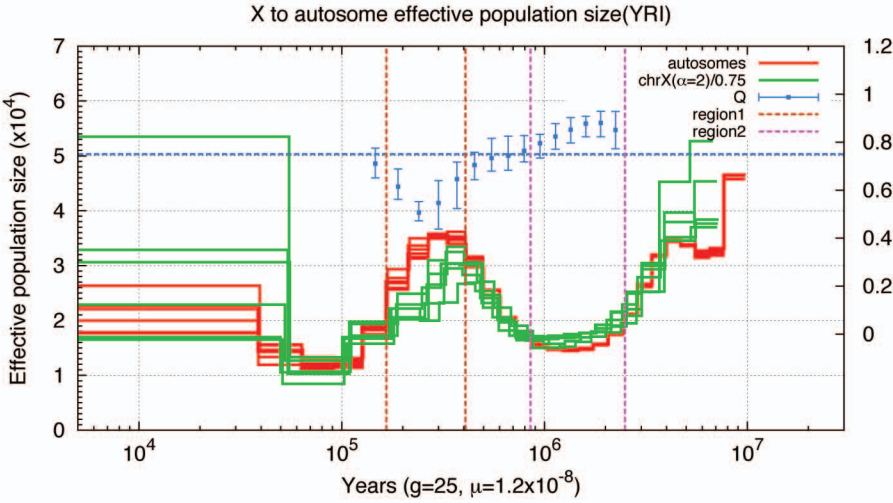
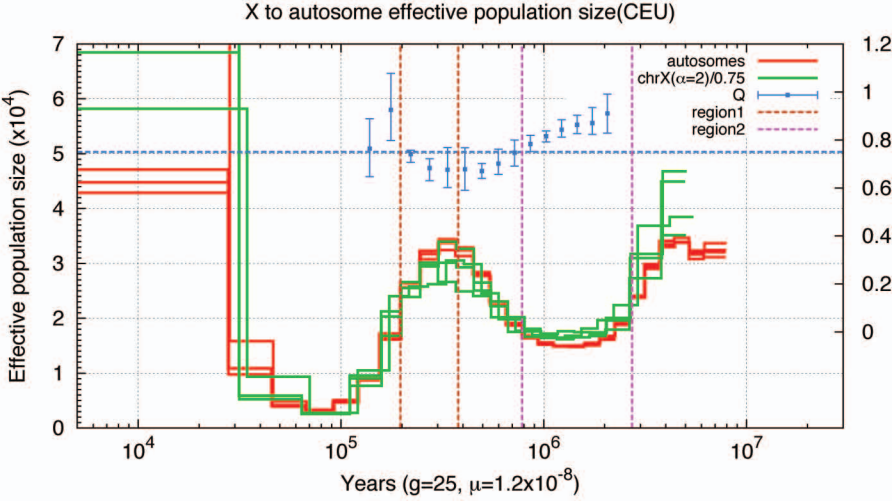

A
B
C
D
YRI-CEU split time
YRI-GIH split time
CEU-GIH split time
Effective population size (x10^4)
Years (g=25, μ=1.2x10^-8)
YRI-hap
YRI-X(α=2)/0.75
CEU-hap
CEU-X(α=2)/0.75
GIH-hap
GIH-X(α=2)/0.75
autosomes11
chrX12(α=2)/0.75
chrX21(α=2)/0.75
chrX11(α=2)/0.75
autosomes22
chrX22(α=2)/0.75

Figure4

A


X to autosome effective population size(YRI)
Effective population size (x10^4)
Years (g=25, μ=1.2x10^-8)
autosomes
chrX(α=2)/0.75
Q
region1
region2

B


X to autosome effective population size(CEU)
Effective population size (x10^4)
Years (g=25, μ=1.2x10^-8)
autosomes
chrX(α=2)/0.75
Q
region1
region2