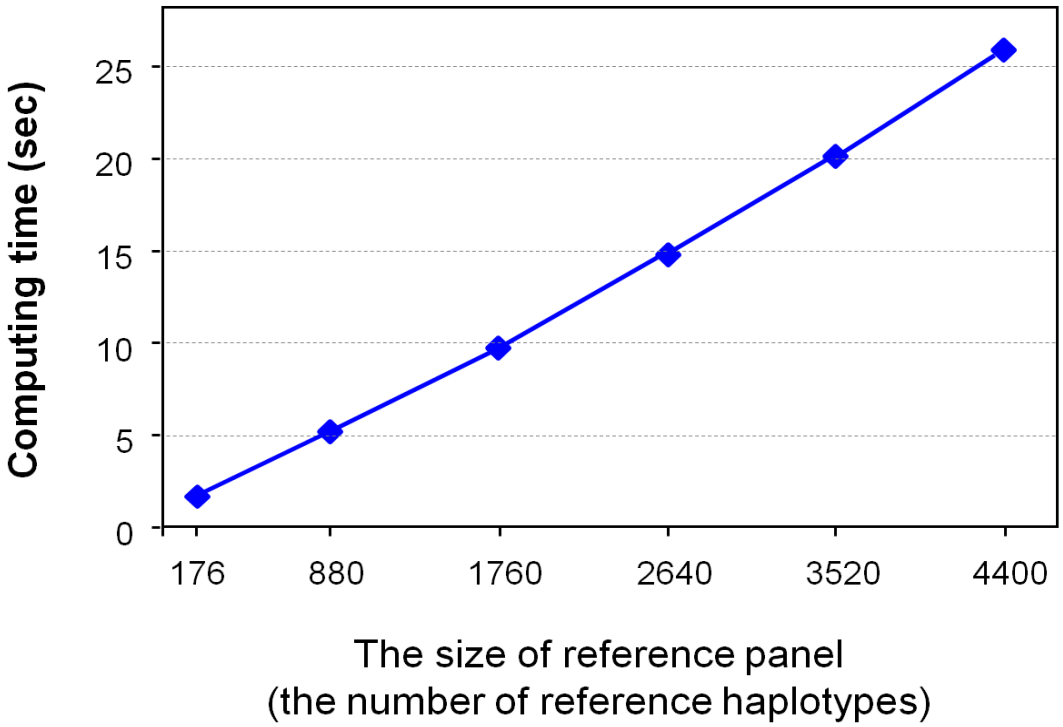
High resolution whole-genome haplotyping using limited seed data

Weinian Rao, Yamin Ma, Li Ma, Jian Zhao, Qiling Li, Weikuan Gu, Kui Zhang, Vincent C. Bond, Qing Song.

Supplementary Information

Supplementary Figure 1 | The linear relationship between the HiFi computing time with the size of large reference panel.



It is computationally intensive to impute haplotypes with large combined reference panel mixed from many populations. In order to use the strategy of combined reference panels from globally diverse population, the algorithm must have the capacity to handle the potentially very large size of reference haplotypes on computing burden and computing time. We examined the capacity of HiFi to handle large reference panels by creating a serial simulated reference panels with various numbers of haplotypes from the HapMap CEU haplotypes. Recombinations were introduced to create these simulated haplotypes. Six Caucasian individuals were analyzed with these reference panels. The seed haplotype input contained 70% missing data; the seed genotype input contained no missing data.

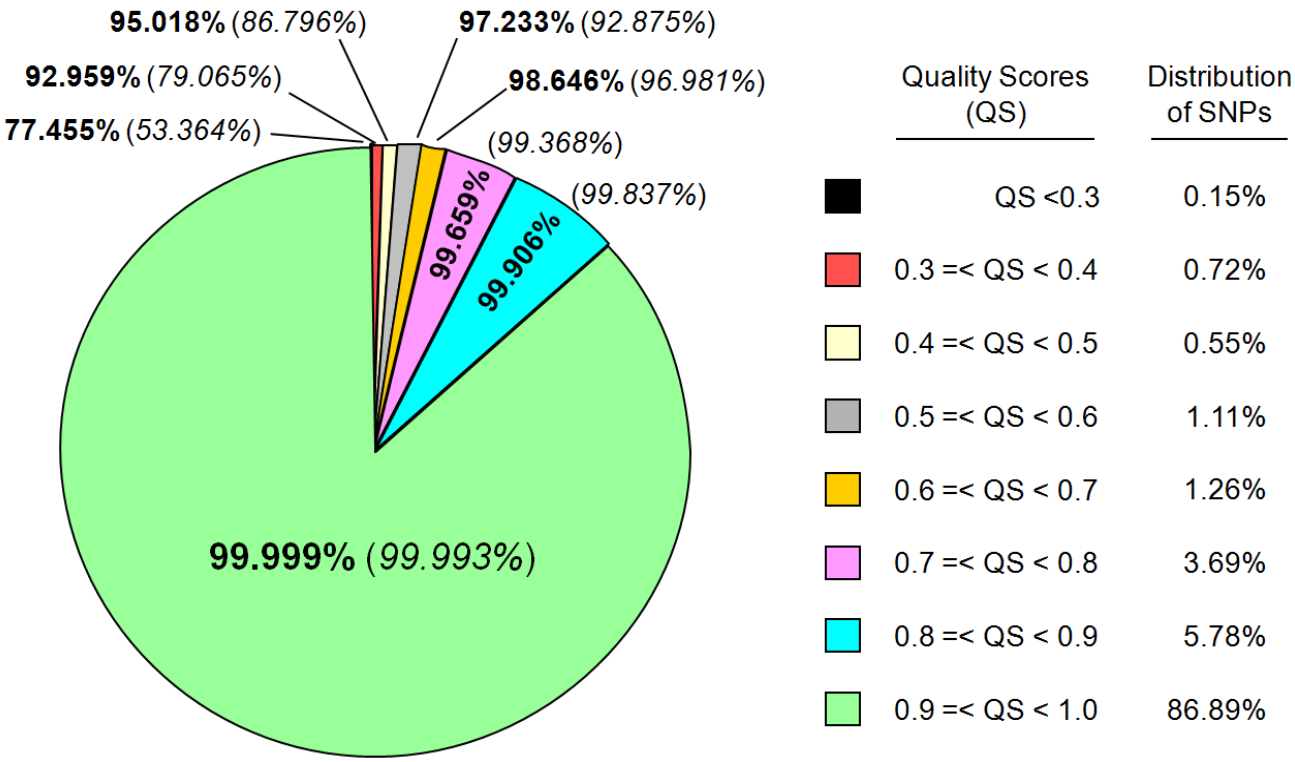
Nature Methods: doi:10.1038/nmeth.2308

Supplementary Figure 2 | Computing time of HiFi and Beagle on a single desktop computer.

Six Caucasian individuals were analyzed. The Caucasian reference panel was used in this experiment, the seed genotype input contained 0 or 10% missing data, and the seed haplotype input contained 70% missing data. The input to Beagle included the same reference panel and genotype dataset but no phased seed input. Default settings were used when we run the Beagle, the number of iterations was 10, and the number of haplotype pairs to sample for each individual during each iteration was 10.

Nature Methods: doi:10.1038/nmeth.2308

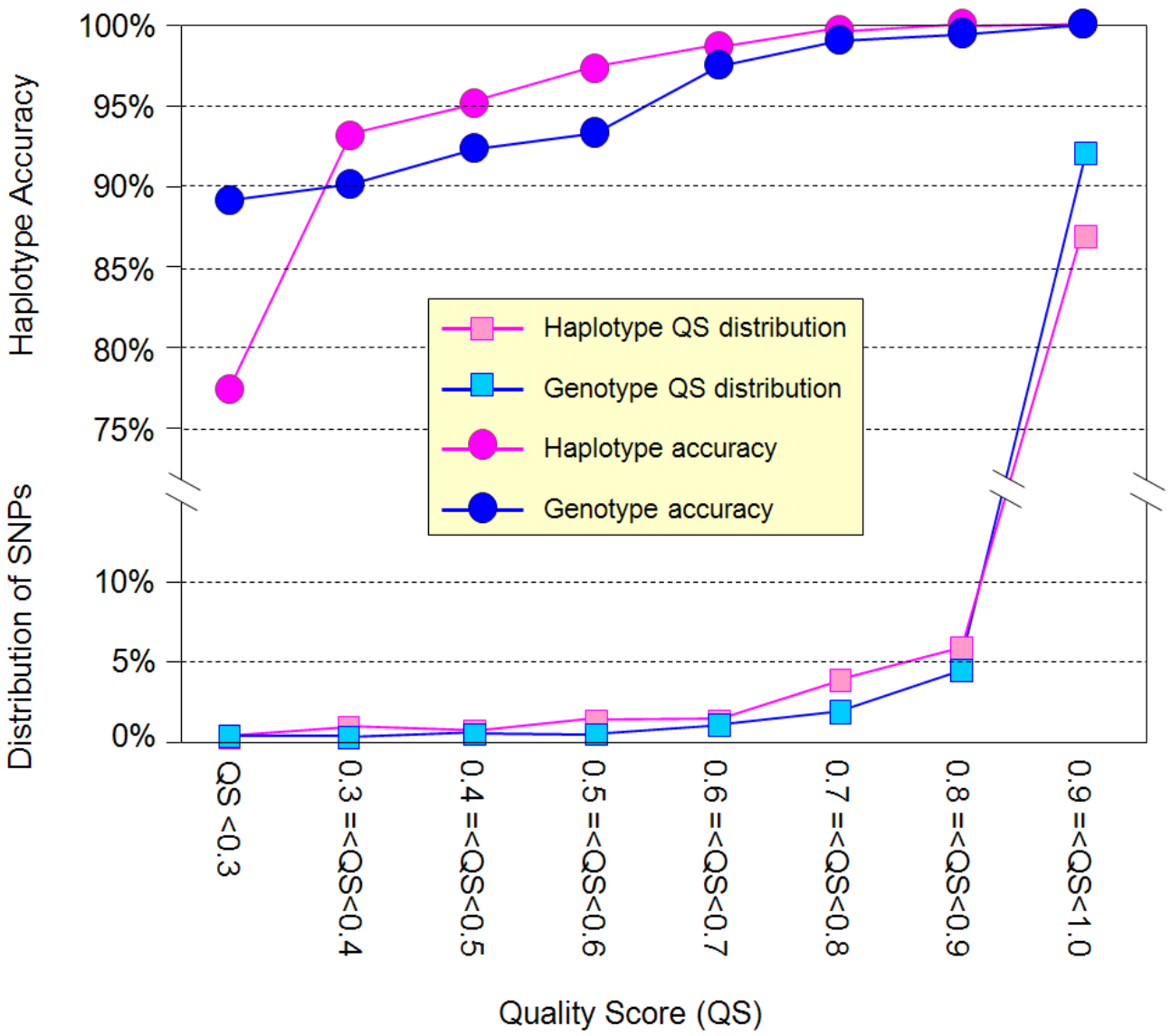
Supplementary Figure 3 | Quality score and the accuracy of the HiFi results.



The performance of this quality score (QS) system was analyzed in six Caucasian individuals. In this analysis, the reference panel was the Caucasian haplotypes in which the imputed person was temporarily removed from the reference panel when imputed this person; the seed genotype input contained 10% missing data; the seed haplotypes contained 70% missing data. Accuracy of HiFi results was evaluated among all snps and among imputed heterozygous loci, respectively (in which heterozygous loci with known phases from seed haplotypes and all homozygous loci were excluded prior to the accuracy evaluation). The accuracies of SNPs among each range of quality scores are labeled on the pie. The accuracies among the entire set of SNPs are shown by the numbers before the parenthesis; the accuracies for those imputed heterozygous SNPs shown by the numbers in italics in the parenthesis. The distribution of quality scores in the HiFi output is shown on the right.

Nature Methods: doi:10.1038/nmeth.2308

Supplementary Figure 4 | Quality scores of HiFi haplotyping and genotyping data.



The haplotyping part (pink) is the line graph version of the pie graph (Supplementary Figure 3). The blue lines are the data on quality scores of genotype imputation. Accuracy of genotype imputation was evaluated among all loci with missing genotypes.

Nature Methods: doi:10.1038/nmeth.2308

Supplementary Table 1 | Three datasets for testing the accuracy of HiFi performance.

	Seed haplotype input to HiFi	Criteria for error detection in the HiFi output
1	HapMap trio haplotypes	HapMap trio haplotypes
2	Quake's experimental haplotypes	HapMap trio haplotypes
3	HapMap trio haplotypes	Song's experimental haplotypes

The first dataset was composed of 6 Caucasians and 6 Africans randomly selected from the HapMap CEU and YRI populations. They are the children of HapMap trios, so that their haplotypes can be accurately inferred with their parental genotypes except those triple heterozygous loci. We randomly blinded the allele phases at 70% loci of entire SNP set, the blinded loci included both homozygous and heterozygous to mimic the realistic scenarios of experimental haplotyping, in which both homozygous loci and heterozygous loci will be subjected to dropout during whole genome amplification. We then used HiFi to retrieve the blinded phases of these individuals. The accuracy of overall HiFi results was evaluated by the accordance rate using the trio haplotypes as criteria.

The second dataset (the Quake dataset) was composed of chromosomal haplotypes of the entire genome of an individual, HapMap NA12878. These haplotypes were experimentally determined haplotypes using single-chromosome isolation strategy [1]. In this dataset, ~40.74% of heterozygous SNP loci of NA12878 were experimentally phased, which were then used as the seed haplotypes for HiFi. Because NA12878 is the child of HapMap trio-1463, her haplotypes can be resolved according to Mendelian allele transmissions with the genotypes downloaded from the HapMap website (NA12878, NA12891 and NA12892). The accuracy of the HiFi results on this dataset was evaluated by the accordance rate using the trio haplotypes as criteria.

The third dataset (the Song data) was composed of whole-genome experimental haplotypes of an individual, HapMap NA10847. Different from the second dataset (Quake), these experimental data were used as criteria of accuracy evaluation rather than as the seed haplotype input of HiFi. NA10847 is the child of HapMap trio-1334, her chromosomal haplotypes was inferred according to the Mendelian Law of Inheritance from her genotypes and her parents' (NA12239 and NA12146) genotypes downloaded from the HapMap website. To create the seed haplotype input of this person, we blinded the allele phases at randomly

selected 70% loci of the entire SNP set on the trio-inferred haplotypes of NA10847; the “70%” blinded loci included both homozygous and heterozygous sites. These blinded loci were then phased with HiFi. The accuracy of the HiFi results on this dataset was evaluated by the Song experimental data by the accordance rate between the HiFi results and the experimental data. Because the Song dataset contains technical replicates on some chromosomes, we used only the experimentally validated data on those heterozygous sites that showed consistent results among the experimental replicates (please see the Supplementary Table 2 of Ma et al., 2010)[2] as criteria to detect the phasing errors in the HiFi results. Those symmetric SNPs (A/T SNPs and C/G SNPs) were excluded before this comparison.

Supplementary Table 2 | The accuracy of HiFi results on the HapMap dataset.

(The denominator is the total number of the entire SNP set).

Accuracy among the entire set of SNPs	Caucasians		Africans	
Reference haplotype panel	Mean	SD	Mean	SD
Caucasian reference panel	99.49	0.05	95.88	0.37
African reference panel	98.91	0.05	99.17	0.08
Combined reference panel	99.45	0.04	99.19	0.05

(The denominator is the total number of imputed heterozygous SNPs of each person).

Accuracy only among imputed heterozygous SNPs	Caucasians		Africans	
Reference haplotype panel	Mean	SD	Mean	SD
Caucasian reference panel	98.11	0.19	85.30	1.48
African reference panel	95.93	0.20	97.04	0.28
Combined reference panel	97.93	0.13	97.11	0.19

The trio haplotypes were downloaded from the HapMap database. The children's haplotypes were inferred with parental genotypes except those triple heterozygous loci [3-4]. To create the seed haplotype input, we blinded the allele phases randomly at 70% loci of entire SNP set (homozygous and heterozygous) on 6 Caucasians and 6 Africans, and used HiFi to retrieve the allele phases at those blinded loci. Three reference haplotype panels were used for HiFi running, the missing rate of seed genotype input was set to zero, the missing rate of seed haplotype input was set to 70%. Finally, the accuracy of overall HiFi results was evaluated by the concordance (%) between the HiFi outputs with the haplotypes inferred from trio genotypes. The overall data quality of HiFi results is reflected by the accuracy among the entire set of SNPs, which is important for users to known for their subsequent experiments and data analysis. The phasing capacity of HiFi is reflected by the accuracy only among imputed heterozygous SNPs because homozygous loci will not have a phasing issue; it provides an indication measurement that may be used for technology development and technology improvement.

Nature Methods: doi:10.1038/nmeth.2308

Supplementary Table 3 | The accuracy of HiFi results on the Quake dataset.

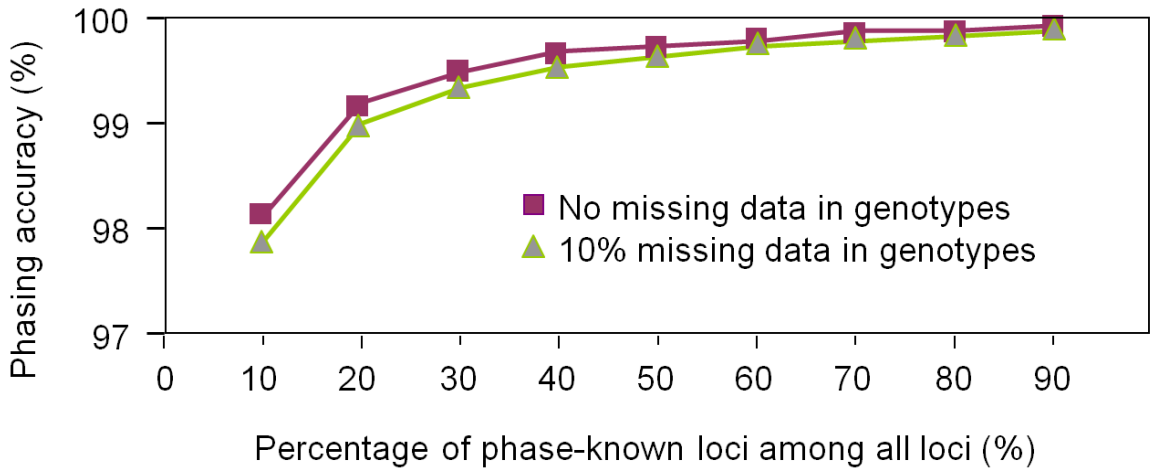
chr	allsnp number	hetsnp number	hetero %	seed hetsnps	seed %	phasing errors	accuracy /hetsnps%	accuracy /allsnps%
chr1	114,150	31162	27.30	12834	41.18	343	98.13	99.70
chr2	115,052	32101	27.90	11210	34.92	460	97.80	99.60
chr3	95,169	26547	27.89	11101	41.82	143	99.07	99.85
chr4	84,577	23252	27.49	9063	38.98	268	98.11	99.68
chr5	86,922	25480	29.31	9867	38.72	213	98.64	99.75
chr6	90,367	26591	29.43	12247	46.06	113	99.21	99.87
chr7	74,276	21056	28.35	9010	42.79	140	98.84	99.81
chr8	74,190	21311	28.72	8780	41.2	237	98.11	99.68
chr9	62,893	17792	28.29	7680	43.17	264	97.39	99.58
chr10	73,176	20362	27.83	8133	39.94	161	98.68	99.78
chr11	70,015	19327	27.60	8551	44.24	113	98.95	99.84
chr12	67,470	18808	27.88	6537	34.76	243	98.02	99.64
chr13	51,467	14315	27.81	5594	39.08	132	98.49	99.74
chr14	44,827	12466	27.81	5177	41.53	112	98.46	99.75
chr15	41,651	11271	27.06	4949	43.91	91	98.56	99.78
chr16	43,791	12519	28.59	4662	37.24	289	96.32	99.34
chr17	37,450	9968	26.62	4183	41.96	190	96.72	99.49
chr18	40,492	11505	28.41	4561	39.64	148	97.87	99.63
chr19	25,313	7436	29.38	3317	44.61	156	96.21	99.38
chr20	35,838	10192	28.44	4434	43.5	90	98.44	99.75
chr21	19,057	5620	29.49	2478	44.09	77	97.55	99.60
chr22	19,672	6019	30.60	2524	41.93	58	98.34	99.71
Total	1367815	385100	28.15	156892	40.74	4041	98.23	99.70

- The Quake dataset contains the experimental haplotypes of a Caucasian individual, NA12878, a child of a HapMap trio. It was obtained by the single-chromosome isolation approach.
- Allsnp: The total number of SNPs of the reference haplotype panel. It includes both homozygous and heterozygous loci of NA12878.
- Hetsnp: The total number of heterozygous SNPs of NA12878.
- Hetero%: The percentage of heterozygous loci among all SNPs.
- Seed hetsnps: The number of heterozygous loci that have been experimentally phased by Quake, which is the seed haplotype input for HiFi imputation.
- Seed%: The percentage of seed hetsnps among all heterozygous loci of NA12878.
- Phasing errors: The number of heterozygous loci that were erroneously phased by HiFi. The phasing errors were detected by comparing the HiFi results with the haplotypes inferred from trio genotypes. Triple-heterozygous SNPs and symmetric SNPs (A/T SNPs and C/G SNPs) were excluded before this comparison.
- Accuracy/hetsnps%: The number of heterozygous loci that have been correctly phased among all imputed heterozygous SNPs. The seed hetsnps were excluded from the denominator (denominator = hetsnps – seed hetsnps) before calculating the accuracy.
- Accuracy/snps%: The number of loci that appeared to be in a correct phase in the complete SNP set along each chromosome. The denominator includes all SNPs (both homozygous and heterozygous SNPs).
- The data missing rate in the seed genotype input was set to zero.

Supplementary Table 4 | The tolerance of HiFi imputation to missing haplotypes in the seed haplotype input.

Missing rate in seed haplotypes (%)	No missing data in seed genotypes		10% missing data in seed genotypes	
	Accuracy (%) among the entire set of SNPs	Accuracy (%) among imputed heterozygous SNPs	Accuracy (%) among the entire set of SNPs	Accuracy (%) among imputed heterozygous SNPs
10	99.96	99.86	99.89	99.58
20	99.92	99.70	99.85	99.43
30	99.88	99.55	99.80	99.24
40	99.82	99.35	99.73	98.99
50	99.77	99.15	99.65	98.70
60	99.69	98.85	99.54	98.27
70	99.49	98.11	99.35	97.56
80	99.21	97.05	98.99	96.25
90	98.08	92.85	97.84	91.92

Six HapMap CEU individuals were analyzed in this experiment. The HapMap CEU haplotypes were used as reference. We tested datasets with either no missing genotypes or 10% missing data genotypes. Concordance% is the percentage of SNP loci that showed the same allele phases between the results from HiFi and the results from genotypes family information.



Supplementary Table 5 | The tolerance of HiFi imputation to missing genotypes in the seed genotype input.

Missing rate in seed genotype %	Accuracy (%) among imputed heterozygous SNPs	Accuracy (%) among the entire set of SNPs
0	98.11	99.49
1	98.12	99.49
2	98.07	99.48
3	98.00	99.46
5	97.88	99.43
7	97.87	99.43
10	97.56	99.35
20	96.71	99.12
30	95.75	98.86
40	94.2	98.45
50	91.74	97.79
60	88.45	96.9
70	83.72	95.63
80	76.68	93.75
90	64.15	90.39

This analysis was based on 6 individuals of HapMap CEU population. The HapMap CEU haplotypes were used as reference in the HiFi running. The seed haplotype input contained 70% missing data; the seed genotype input contained 10-90% missing data. Concordance% is the percentage of SNP loci that showed the same allele phases between the results from HiFi and the results inferred from trio haplotyping with parental genotypes. Triple-heterozygous loci and symmetric SNP loci (A/T and C/G SNP) were excluded from concordance measurement.

Supplementary Table 6 | The impact of errors in the seed haplotypes and seed genotypes on the accuracy of the HiFi results.

Accuracy (%) among the entire set of SNPs		Error% in seed haplotypes				
		0	0.5	1.0	1.5	2.0
Error% in seed genotypes	0	99.50	99.46	99.39	99.36	99.29
	0.5	99.44	99.37	99.33	99.28	99.19
	1.0	99.35	99.29	99.22	99.19	99.12
	1.5	99.23	99.21	99.16	99.08	98.98
	2.0	99.16	99.1	99.03	98.93	98.90

Accuracy (%) among imputed heterozygous SNPs		Error% in seed haplotypes				
		0	0.5	1.0	1.5	2.0
Error% in seed genotypes	0	98.13	97.99	97.71	97.58	97.33
	0.5	97.92	97.64	97.50	97.29	96.95
	1.0	97.57	97.35	97.08	96.96	96.70
	1.5	97.12	97.02	96.86	96.56	96.19
	2.0	96.85	96.63	96.37	96.00	95.90

There will be errors in the seed haplotypes and seed genotypes in realistic works and these errors will affect the accuracy of the HiFi results. The error rate on the allele calls in the experimental seed haplotypes will depend on 1) errors occurred in genotyping calls (~0.1%, as reported by the Illumina Technote), and 2) errors occurred in whole-genome amplification (~0.73%) [1-2]. The error rate on the allele calls in seed genotypes will depend on whole-genome genotyping errors (~0.1%). We have created a series simulated seed haplotype and seed genotype datasets with various error rates around these realistic settings. The introduced errors were randomly distributed in the seed haplotypes and seed genotype data.

Six HapMap CEU individuals were phased with HiFi. The HapMap CEU haplotypes were used as reference. The seed haplotype input contained ~70% missing data (randomly selected); the seed genotype input contained no missing data. During the haplotype imputation with HiFi, the haplotypes of the individual that was being imputed or her/his parents' haplotypes were always temporarily removed from the reference panel. The accuracy was estimated by the concordance between the HiFi results and the trio-haplotyping results. Loci with known allelic phases from the seed haplotypes were eliminated when we calculated the accuracies.

Nature Methods: doi:10.1038/nmeth.2308

Supplementary Table 7 | The impact of errors in the reference haplotype panel on the accuracy of the HiFi results.

Flip error rate	0%	0.5%	1%	2%	4%	8%	16%
Accuracy (%) among the entire set of SNPs	99.53	99.53	99.53	99.53	99.51	99.51	99.49
Accuracy (%) among imputed heterozygous SNPs	98.23	98.23	98.23	98.24	98.17	98.16	98.09

Among the haplotypes downloaded from the HapMap project, those unresolved triple-heterozygous loci were then phased statistically (HapMap3_r2 phasing summary, http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/ 2009-02_phaseIII/HapMap3_r2/). There may be flip errors in these haplotypes, particularly on those statistically resolved triple heterozygous loci, these errors may mislead HiFi to generate phasing errors when they are used as the reference panel.

It is estimated that ~4-6% of SNPs in the trio-child belong to triple-heterozygous loci (HapMap3_r2 phasing summary), and it is estimated that the flip error rate of statistical phasing is below 1% [5-7]. To evaluate the potential impact of the errors in the reference panel on the accuracy of HiFi performance, we created from the downloaded HapMap CEU trio haplotypes a series of simulated haplotype reference panels with different flip error rates. Six HapMap CEU individuals were analyzed with these reference panels. The seed haplotype input contained ~70% of missing data, and the seed genotype input contained 0% missing data. During the haplotype imputation with HiFi, the haplotypes of the individual that was being imputed or her/his parents' haplotypes were always temporarily removed from the reference panel. Concordance% is the percentage of SNP loci that showed the same allele phases between the results from HiFi and the results from genotypes family information. Loci with known allelic phases were eliminated when counting the HiFi phasing accuracy.

With these settings, we observed a very modest effect of these errors on the accuracy of HiFi results. We guess that the reason for this observation is that a small percentage of local flip errors in the reference panel were compensated by the presence of the other haplotypes in the panel.

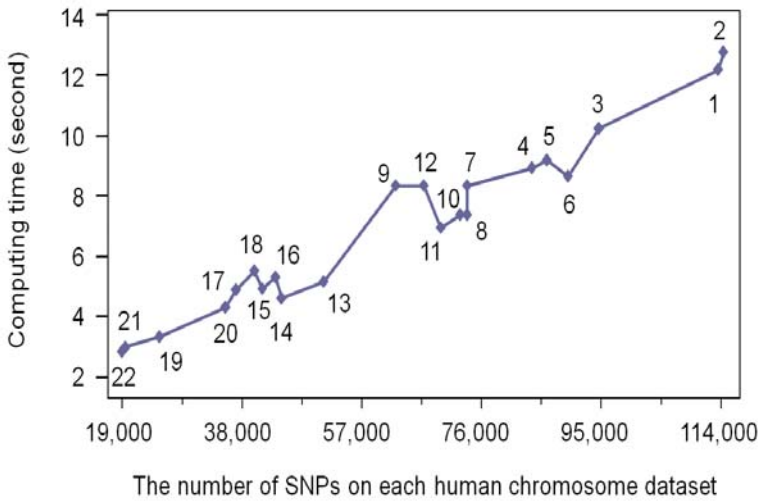
Supplementary Table 8 | The feature of HiFi.

Features	Results and requirements
Accuracy	>99% on each personal haplotype
Phasing distance	Chromosomal range
Speed	2.5 min computing time for each personal genome.
Computer requirement	Desktop computer
Marker type	Flexible, dependent on the marker types in references
Rare variants	94.07% accuracy on those imputed heterozygous SNPs with a minor allele frequency (MAF) < 0.01

Supplementary Table 9 | The computing time of HiFi on the Quake dataset [1].

Chromosome	Total number of SNPs	Runtime (sec)
chr1	114,150	12.04
chr2	115,052	12.62
chr3	95,169	10.11
chr4	84,577	8.78
chr5	86,922	9.08
chr6	90,367	8.50
chr7	74,276	8.22
chr8	74,190	7.24
chr9	62,893	8.19
chr10	73,176	7.22
chr11	70,015	6.79
chr12	67,470	8.19
chr13	51,467	5.01
chr14	44,827	4.49
chr15	41,651	4.79
chr16	43,791	5.19
chr17	37,450	4.73
chr18	40,492	5.35
chr19	25,313	3.20
chr20	35,838	4.17
chr21	19,057	2.67
chr22	19,672	2.84
Total	1,367,815	149.42

The computing time is shown for each chromosome of an individual HapMap GM12878. The runtime of HiFi is in a linear relationship to the total number of SNPs, indicating its capacity to analyze datasets with a large number of markers.



Supplementary Table 10 | The accuracy of HiFi among rare SNPs.

Minor Allele Frequency (MAF) (%)	Hetero% (Hetsnps/all snps)	Accuracy (%) among imputed heterozygous SNPs	Accuracy (%) among the entire set of SNPs
0-1%	1.67	94.07	99.89
1-2%	2.80	97.10	99.92
2-5%	6.73	98.02	99.87
5-10%	13.45	97.95	99.73
10-20%	24.13	98.10	99.55
20-30%	36.26	98.14	99.33
30-40%	44.15	98.18	99.21
40-50%	48.53	98.08	99.11
All	29.39	98.11	99.45

Six HapMap YRI individuals were analyzed in this experiment. The seed haplotypes contained ~70% missing data, and the seed genotypes contained no missing data. During the haplotype imputation with HiFi, the haplotypes of the individual that was being imputed or her/his parents' haplotypes were always temporarily removed from the reference panel. In the metrics for accuracy measurement, the denominator for calculating the accuracy among imputed heterozygous SNPs is the total number of the imputed heterozygous loci in which the phase-known heterozygous SNPs in the seed haplotype input has been excluded from the denominator; the denominator for calculating the accuracy among the entire set of SNPs is the total number of the SNP loci in the reference panel.

Supplementary Table 11 | Estimate of labor and cost of the integrated pipeline.

Step	Time	Hands-on Time	$Labor	$Reagents	Subtotal ($)
Cell culture	2 days	50%	$200	$400	$600
Chromosome preparation	4 days	100%	$800	$600	$1,400
Chromosome isolation	6 days	100%	$1,200	$800	$2,000
Whole genome amplification	2 days	100%	$400	$2,600	$3,000
High-throughput genotyping	3.5 days	n/a	n/a	n/a	$72,000
Haplotype readout	0.5 days	100%	$0	$0	$0
HiFi running	0.5 days	100%	$0	$0	$0
Cost for each sample	9.5 days	0.34%	$6.33	$19.14	$823

The labor and reagent cost are estimated based on 96-well plates. The labor cost is based on a single technician at $200/day, and adjusted by conservative estimates of hands-on time. The reagent cost is based on the list prices for items used. The cost of high-throughput genotyping is based on current outsourcing service price on Illumina BeadyArray Quad-2.5M. The computing time is estimated based on one desktop CPU of 3.0 GHz and 8 GB RAM.

Supplementary Methods

The Input and Output of HiFi

HiFi takes three input files, 1) an unphased genotype dataset, in which missing data is allowed; 2) a reference haplotype panel; 3) a low-resolution experimental haplotype dataset, in which a majority of loci are phase-unknown. HiFi exports two chromosome-length whole-genome complete haplotypes for each individual, in which all loci will be phased.

The Algorithm of HiFi

HiFi is built upon the hypothesis that unrelated individuals may share short stretches of DNA sequence derived from their common ancestors [5, 8]. HiFi uses a sliding window to scan the reference haplotypes for potential matches to each person's partial genotypes and partial haplotypes. HiFi starts with a 17-SNP window and automatically adjust the window size in this scan, until a unique match is found for the haplotype pair of an individual in the corresponding window. If a single match is found within a window, this haplotype pair will be used to impute the phases at all loci within this window of this individual. If no match is found within a window, HiFi will automatically shrink the window by 2 SNPs and repeat the search; if multiple matches are found within a window, HiFi will automatically enlarge the window by 4 SNPs and repeat the search.

After the window is shrunk by 2-SNPs, if no match is found again, the window size will be further reduced by 2-SNPs and the reference haplotype panel will be scanned again. If multiple matches are found, the window size will be enlarged by 4-SNPs but those two shrunk SNPs will be skipped.

After the sliding window is enlarged by 4-SNP, if no match is found, HiFi will remove those two boundary SNPs from the enlarged window and repeat the search. If multiple matches are found in the 4-SNP enlarged window, HiFi will further enlarge the window by 4-SNPs and performs the search again.

HiFi will repeat the window adjustment until a unique match is identified for an individual within a window larger than 11-SNPs. HiFi will not use non-unique match for recovering the

missing phases, and will not use unique match obtained in a window no larger than 11 SNPs for data recovery.

SNP density does not play a role in the adjustment of the windows because the exact positions of SNPs and genetic distances are not used by HiFi during its imputation process once the SNPs are sorted by their orders with their positions.

This core algorithm of HiFi will be performed by multiple rounds, in which HiFi strategically targets common SNPs in early rounds and move down to less common SNPs gradually in later rounds. HiFi "targets" common SNPs in its early stages according to the minor allele frequencies (MAF) in the reference panel. For example, in the first round, HiFi scans with sliding windows among those SNPs with MAF 0.49-0.50. In the second round, HiFi scans with sliding windows among those SNPs with MAF 0.48-0.49. This scan will be repeated until all SNPs in various MAF ranges are completed.

HiFi also strategically targets those phase-unknown heterozygous loci in early rounds and move down to genotype-unknown SNPs in later rounds. It will impute the missing data in the seed haplotypes first before imputing the missing data in the seed genotypes.

When there is no partial haplotype information in a region, how does the method find an "unequivocal" match based on unphased study genotypes - Because the exact positions of SNPs and genetic distances are not used by HiFi during its imputation process once the SNPs are sorted by their orders with their positions, HiFi scans the SNPs by the orders of SNPs rather by the genomic positions or genetic distances of the SNPs. Therefore, HiFi will impute the phases no matter whether there is a large gap in the chromosomes that have no partial haplotypes.

At the end, if there is no non-ambiguous match at a imputing target locus, HiFi will randomly assign alleles on these loci. We have investigated how often this case occurred; about 7-25 SNPs did not receive a non-ambiguous match in each person. Because the total number of these loci was very small, it will not affect the overall accuracy of HiFi results, and these loci are labeled in the quality score report.

The distribution of window size that HiFi found a non-ambiguous match was monitored. Six HapMap CEU individuals were analyzed in this experiment. The HapMap CEU haplotypes were used as reference. The seed haplotypes contained ~70% missing data, and the seed genotypes contained 0% or 10% missing data. During the haplotype imputation with HiFi, the haplotypes of the individual that was being imputed or her/his parents' haplotypes were always temporarily removed from the reference panel. The analysis showed that majority of SNPs were imputed in windows with size 17-20 SNPs.

Window size (SNP)	Count of imputed heterozygous SNPs with an unambiguous match in each window size	
	0% missing data in seed genotype	10% missing data in seed genotype
11-12	4892	6251
13-16	6791	9298
17-20	112499	152832
21-28	3017	7257
29-52	2264	6601
53-99	769	3323
>100	110	994

Reference Haplotype Panels

The Caucasian Reference Haplotype Panel was composed of 176 haplotypes of 88 HapMap CEU individuals (CEPH, U.S. Utah residents with ancestry from northern and western Europe). The African Reference Haplotype Panel was composed of 352 haplotypes of 176 Africans, YRI (100, Yoruba in Ibadan, Nigeria), MKK (56, Maasai in Kinyawa, Kenya) and ASW (20, African Ancestry in SW USA). The combined reference panel is composed of all haplotypes of Africans and Caucasians in these two reference panels.

Each trio consists of 3 individuals (two parents and one child). To avoid redundancy, only the haplotypes of the parents are included in the HapMap dataset, the children's haplotypes are not included in the final phased files (HapMap3_r2 phasing summary, last modified 27-Feb-2009, http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/2009-02_phaseIII/HapMap3_r2/). Phasing was completed in two stages. During the first stage, family information was employed to deterministically resolve phases by allele transmission except those triple heterozygous sites (the sites that all three individuals in a trio are heterozygous). During the second stage,

sites with unresolved phase were phased statistically using the method developed by Howie et al [4]. The deterministic phasing for trios was conducted using purpose-built routines. Around 28% of the SNP loci of each sample are heterozygous, and around 80% of these heterozygous loci are deterministically resolved using family information. Hence, the percentage of the trio samples that is deterministically resolved is about 94% (HapMap3_r2 phasing summary, last modified 27-Feb-2009, http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/2009-02_phaseIII/HapMap3_r2/). The high percentage of deterministic phasing for trio samples made them the obvious candidates to be used as a reference population for the unrelated individuals.

A summary of the haplotypes downloaded from the HapMap database in this study.

Reference Panels	Population	Total trios	Total individuals in phased files	Total haplotypes in phased files
Caucasians	CEU	44	88	176
	All	44	88	176
Africans	ASW	10	20	40
	MKK	28	56	112
	YRI	50	100	200
	All	88	176	352
Combined	All	132	132	528

The reference haplotypes contain 116,415 SNPs, corresponding to the entire set of SNPs on chromosome-1 in the HapMap datasets. The haplotypes of an individual that was being imputed or the parents' haplotypes of an individual that was being imputed were always temporarily removed from the reference panel. The haplotypes of the parents of NA10847 and NA12878 were removed from the reference panel when we imputed the haplotypes of NA10847 and NA12878 with HiFi.

Datasets

We used three datasets for testing the performance of HiFi in this study (Supplementary Table 1).

The first dataset was a semi-simulated dataset created from the true personal haplotypes downloaded from the HapMap project, the total number of SNPs was 116,415 across entire chromosome 1. It was composed of 6 Caucasians and 6 Africans randomly selected from the HapMap CEU and YRI populations. They are the children of HapMap trios, so that their haplotypes can be accurately inferred with their parental genotypes except those triple heterozygous loci. We randomly blinded the allele phases at 70% loci of entire SNP set, the blinded loci included both homozygous and heterozygous to mimic the realistic scenarios of experimental haplotyping, in which both homozygous loci and heterozygous loci will be subjected to dropout during whole genome amplification. We then used HiFi to retrieve the blinded phases of these individuals. The accuracy of overall HiFi results was evaluated by the accordance rate using the trio haplotypes as criteria.

The second dataset was a true dataset with experimentally determined haplotypes by the Quake group [1]. It was composed of chromosomal haplotypes of the entire genome of an individual, HapMap NA12878. These haplotypes were experimentally determined haplotypes using single-chromosome isolation strategy [1]. In this dataset, ~40.74% of heterozygous SNP loci of NA12878 were experimentally phased, which were then used as the seed haplotypes for HiFi. Because NA12878 is the child of HapMap trio-1463, her haplotypes can be resolved according to Mendelian allele transmissions with the genotypes downloaded from the HapMap website (NA12878, NA12891 and NA12892). The accuracy of the HiFi results on this dataset was evaluated by the accordance rate using the trio haplotypes as criteria.

The third dataset was a true dataset with experimentally determined haplotypes by our group[2]. It was composed of whole-genome experimental haplotypes of an individual, HapMap NA10847. Different from the second dataset (Quake), these experimental data were used as criteria of accuracy evaluation rather than as the seed haplotype input of HiFi. In addition, this dataset has a clear record on the consistency of experimental phasing results among technical replicates, which could eliminated the discordance between HiFi results and experimental results that were caused by experimental errors. NA10847 is the child of HapMap trio-1334, her chromosomal haplotypes was inferred according to the Mendelian Law of Inheritance from her genotypes and her parents' (NA12239 and NA12146) genotypes downloaded from the HapMap website. To create the seed haplotype input of this person, we

blinded the allele phases at randomly selected 70% loci of the entire SNP set on the trio-inferred haplotypes of NA10847; the “70%” blinded loci included both homozygous and heterozygous sites. These blinded loci were then phased with HiFi. The accuracy of the HiFi results on this dataset was evaluated by the Song experimental data by the accordance rate between the HiFi results and the experimental data. Because the Song dataset contains technical replicates on some chromosomes, we used only the experimentally validated data on those heterozygous sites that showed consistent results among the experimental replicates (please see the Supplementary Table 2 of Ma et al., 2010) [2] as criteria to detect the phasing errors in the HiFi results. Those symmetric SNPs (A/T SNPs and C/G SNPs) were excluded before this comparison.

To access the capacity of HiFi to impute the haplotypes with various setting on the missing rates of missing data in the seed haplotypes and seed genotypes, we created a series of semi-simulated datasets using the first dataset – the HapMap trio dataset.

To examine the capacity of HiFi to handle very large reference panels, we created simulated reference haplotype panels with various sizes from the HapMap CEU haplotypes. Recombinations were introduced to create these semi-simulated haplotypes.

Computing Time

Both of HiFi and Beagle were run on a desktop computer [Windows Vista Ultimate SP1 (Intel(R) Core(TM)2 Duo, CPU 3.0 GHz, 8 GB of RAM]. The dataset contained 116,415 SNPs spanning 249-Mb of the entire human chromosome-1. Six Caucasian individuals were analyzed.

The input to HiFi included the Caucasian reference panel, the seed genotypes with 0 or 10% missing rate, and the seed haplotypes with 30% phase-known loci.

The input to Beagle included the Caucasian reference panel and the genotype dataset with 0 or 10% missing rate, it does not take the seed haplotypes. The default settings were used for the Beagle running, the number of iterations was set to 10, and the number of haplotype pairs to sample for each individual during each iteration was also set to 10.

The computing time of HiFi is about 84-folder faster than Beagle (Supplementary Figure 2).

Measurement of Phasing Accuracy

The children's haplotypes of each trio can be determined accurately and unambiguously with trio genotypes according to Mendelian Laws of Inheritance except those triple heterozygous loci [3-4]. It has been shown that the haplotypes yielded with trio genotypes are consistent with the haplotypes determined experimentally [1-2, 9-10] and are a widely accepted method for accuracy evaluation for phasing methods. In this study, all samples were the children of trios; we used their haplotypes yielded from trio genotypes to evaluate the phasing accuracies of HiFi.

Because those homozygous loci are already phase-known for "free" in an individual, and only the heterozygous loci are ambiguous on allele phase that need to be resolved, the accuracy will be always better if all SNPs were counted than the accuracy if only heterozygous SNPs were counted. For example, when ~72% of total SNPs are homozygous and 28% are heterozygous, if the phasing simply done by coin flip assuming all heterozygous loci are biallelic, the actual accuracy of phasing capacity will be 50%, but if all SNPs were counted for evaluation, the accuracy of the phasing result would be 72% + 28% x 0.5 = 86%.

Therefore, we measured the long-range haplotyping accuracy of HiFi performance with two metrics, the accuracy among the entire set of SNPs including both homozygous and heterozygous SNP loci, and the accuracy only among those imputed heterozygous SNPs. The first one is defined as the percentage of the correctly phased loci among the total number of the entire SNP set including both homozygous loci and heterozygous loci. The second is defined as the percentage of the correctly phased heterozygous loci among the total number of imputed heterozygous sites. The SNP loci with known phases from the seed haplotypes were excluded from the total number of heterozygous loci when calculating the accuracy.

The overall data quality of HiFi results is reflected by the accuracy among the entire set of SNPs, which will be the actual input for subsequent experiments and data analysis. However, the phasing capacity of a method is reflected by the phasing accuracy only among the

imputed heterozygous loci because homozygous loci will not have a phasing issue; it provides an indication measurement that may be used for technology development and technology improvement.

For the accuracy among the entire set of SNPs,

$$\text{Accuracy} = \frac{\text{The number of SNPs with correctly phased}}{\text{The number of all SNPs}}$$

For the accuracy only among imputed heterozygous SNP loci,

$$\text{Accuracy} = \frac{\text{The number of heterozygous SNPs with correctly phased}}{\text{The number of heterozygous SNPs} - \text{The number of seed heterozygous SNPs}}$$

We used a stringent criterion in accuracy evaluation. Only the loci that were correct on both allele calls and allelic phases on both alleles were reported as “correct”. If a locus received only one correct allele imputation among two alleles at a locus, or both allele calls are correct but with wrong phases, this locus will be treated as an error.

Throughout this study, we used the error rates on entire chromosomes, the denominators were either the total number of SNPs of the entire SNP set or the total number of imputed heterozygous SNP sites.

The phasing distance was 249-Mb, composed of 116,415 SNPs, spanning across the entire chromosome-1, for all of the analyses in this study. All triple-heterozygous-loci and all symmetric SNP loci (A/T SNPs and C/G SNPs) were excluded when evaluating the accuracy. All heterozygous loci in the seed haplotypes were excluded from the denominator when we calculated the accuracies of HiFi results.

We used “switch errors” to describe the phase switches of entire chromosomal segments; we used “flip errors” to describe the phase flips of single SNPs on an otherwise correct haplotype scaffold (e.g., when discussing phasing errors at triple-heterozygous loci).

Validation of the Accuracy of the HiFi Results by Experimental Data

We used our previously obtained experimental data (haplotyping by the single-chromosome isolation strategy on a Caucasian individual HapMap GM10847) [2] to examine the accuracy of the HiFi results. The feature of this experimental dataset is that it documented the information of experimental replicates and consistencies on each phased heterozygous SNP loci (Supplementary Table 2 of Ma et al., 2010, nmeth) [2]. Only those heterozygous loci with consistent experimental results were used as criteria to evaluate the accuracy of HiFi results. For the HiFi imputation, the seed haplotype input did not contain any loci that were experimentally phased. Instead, the seed haplotype input was made by blinding 70% of randomly selected sites on the trio-phased haplotypes of GM10847 downloaded from the HapMap database. The seed genotypes did not contain missing data. The reference panel was the CEU haplotypes. The haplotypes of the parents of NA10847 were excluded from the reference panel. The accuracy of the HiFi results was evaluated with the experimental data that showed consistent phases between different experimental replicates. The accuracy of HiFi results was 98.23% among all imputed heterozygous loci.

The Quality Scores

The quality scores (QS) will be useful for users to apply the HiFi data for their subsequent experiments. To develop this accuracy prediction metrics, we have investigated the effects of potential risk factors on the accuracy of HiFi. A HiFi scoring system was designed based on their impacts and their relative weight of the impact on the accuracy of HiFi. We have implemented a quality score system into the HiFi software, which will be output together with the haplotypes. Each imputed site will receive a 0-1 score. The higher scores, the higher quality of imputation calls.

Briefly, this quality score metrics was developed by three steps. First, we studied the correlations between occurrence of HiFi imputation errors and 15 factors that may potentially affect the HiFi accuracy. Among these factors, we found that 8 factors were statistically

associated with occurrence of errors during the HiFi imputation. Second, we applied a 0-5 points scoring system so that each imputed SNP of a person will receive a corresponding point on each of those 8 factors. Last, we investigated the degrees of their impacts on the HiFi errors, and then developed a weighing system, each of those 8 factors received a particular weight. The final quality at an imputed SNP in a person is a weighted sum of the points on these 8 factors. These factors include, from the matched haplotype frequency, the haplotype diversity, the window size as counted by the number of SNPs, the window size as counted by base positions, the total number of heterozygous loci in a window, the total number of homozygous loci in a window, the total number of window elastic adjustment for a SNP for a person, and the total number of the phase-unknown heterozygous SNPs in a window.

Supplementary Discussions.

The Needs of Long-Range Haplotypes.

Although DNA is a linear sequence, the chromatin, which is packed and organized inside the nucleus, does not function "linearly" [11]. In complex genomes, such as human genome and mouse genome, regulatory elements can act over large genomic distances by engaging in direct physical interactions with target genes by forming chromatin loops [12-14]. These long-range looping interactions will participate the guidance of gene regulations between promoters, enhancers and insulators [15-17]. This has been illustrated by the fact that genes are often regulated by elements that are located hundreds of kilobases or tens of millions of bases away in the linear genome [12, 18-19]. Even at distances greater than 200 Mb, intra-chromosomal contact probability is always much greater than the average contact probability between different chromosomes [12]. In the recently publications of the ENCODE project, it was reported that only ~7% of looping interactions are with the nearest gene, indicating that genomic proximity is not a simple predictor for long-range interactions [13-14]. It is of significant interest to understand chromatin loop conformation and long-range cis-interaction networks of complex genomes regarding their functionalities and causal role in complex diseases, which will require the long-range or even the chromosomal-range haplotypes.

The phase information is essential for exploring the associations, genetic ancestry, medical genetics, and especially for the allele-specific functions and etiologies [1, 9, 20-21]. For example, the cis-conformation of a set of dysfunctional alleles is essential to cause a disease phenotype, a disease may arise only when they are present on the same chromosome, but it will be harmless if a person has them on opposite chromosomes. In this kind of cases, the phase information will be important for discovering and interpreting the disease-causing variants. However, it has been known that interacting mutations may not necessarily be close to each other; instead, they may be over million bases away from each other [12, 22-27]. There has been an increasing awareness that gene expression can be regulated by multiple cis-acting sequences located as far as 1 Mbp away from the gene. Therefore, resolving the long-range haplotypes will be important for discovering and interpreting those long-range cis-interactions in the gene function and disease pathogenesis [1, 9, 21, 28-30].

Limitations of HiFi

HiFi is a reference-based imputation method. A limitation of reference-based imputations is that they cannot impute the haplotypes "creatively" for a personal case that is not covered by the reference, such as those private haplotypes, resulted from private crossovers and private de novo mutations. Certainly extremely rare SNPs may be relatively "private" if a reference panel is not large enough.

Because the information of genomic positions and genetic distances are not used in the HiFi phasing algorithm once the SNPs are sorted by their positions, HiFi will not be able to detect the large gap regions that do not have any partial haplotype information. This strategy enables HiFi to "jump over" those large gap regions, as the cost of accuracy loss when "jump over" these regions. However, this "jump" will automatically give an alarm to the HiFi, which will be documented and scored by the HiFi Quality Score System.

HiFi could not impute the structural variations because the position information of each variant was no longer used in its phasing algorithm after the variants were sorted by their position order, and also because currently available reference haplotypes do not contain structural variations yet.

Mixed Reference Panel

The well-matched reference panel is not always available for a person. The solution to the limited availability of reference panels is to use a mixed reference panel from diverse populations as suggested by recent studies [31-35]. We also observed the same phenomenon with HiFi in our work (Supplementary Table 3 and Supplementary Figure 1). Now the question is how to mix these available haplotypes from various populations to form the combined reference panel? In other word, when generating the mixed reference panel, is there any optimal ratio to mix the haplotypes from diverse populations? This is a hard question or may be even an unanswerable riddle because the answer will depend on the specific ancestral background of each person to be haplotyped; it will also depend on the ancestral distance between populations that are mixed, the diversity of haplotypes in the pool of each population, and the ancestral composition of each population. However, despite that it is a hard question to address, the ratio will dictate the haplotype frequencies in the mixed reference panel, it may

severely mislead a method during the imputation if it relies on haplotype frequencies to calculate the probabilities in their decision-making process [6]. HiFi seeks only for the non-equivocal match; it does not give any credit or penalty to the haplotype frequencies in its process. Therefore, although HiFi will be sensitive to the lack of coverage on the haplotype diversities in the reference panel, it may be resistant to the sampling bias introduced arbitrarily by the mixing ratio.

Supplementary Note 1 | The procedure of this integrated haplotyping method.

1. Add 0.2 ml whole blood into a 15-ml tube containing 4 ml PB Max medium. Invert gently. Incubate at 37°C for 45 h.
2. Add 40 ul colcemid. Invert gently. Incubate at 37°C for 30 min.
3. Centrifuge at 1,000 rpm for 5 min, aspirate all but 0.3 ml supernatant, re-suspend cell pellet gently.
4. Add 5 ml 0.075M KCl, incubate at room temperature for 15 min.
5. Add 4-5 drops of cold fixative, invert gently, centrifuge at 1,000 rpm for 5 min. Aspirate supernatant.
6. Add 5 ml cold fixative, invert gently, centrifuge at 1,000 rpm for 5 min, and aspirate supernatant. Repeat once.
7. Add 0.5 ml cold fixative and re-suspend the cell pellet.
8. Drip cells onto a microscope slide to spread chromosomes, a few drops on each slide.
9. Giemsa-stain the chromosomes for 10 min.
10. * Observe and select the target chromosomes, and click cut on the computer to cut the foils with a computer-directed Laser Micro-dissection System. The foils are collected into collecting eppendorf tubes.
11. Spin the collecting tubes briefly in microcentrifuge.
12. Amplify the collected chromosomes (still on the foils) with a whole genome amplification kit.
13. Purify the amplified product with Qiagen PCR Purification Column.
14. Use an aliquot of purified amplified products to do high-throughput genotyping (such as Illumina BeadArray) or next-generation sequencing.
15. Read out the haplotypes directly from the high-throughput genotyping or next-generation sequencing results.
16. Use the HiFi software to obtain the complete haplotypes.

* This step can be done with any chromosome-isolation methods such as the device described by Fan et al., 2011.[1], which will improve the speed and throughputability.

Literature in the Supplementary Materials

1. Fan, H.C., Wang, J., Potanina, A. & Quake, S.R. Whole-genome molecular haplotyping of single cells. *Nat Biotechnol* **29**, 51-57 (2011).
2. Ma, L. et al. Direct determination of molecular haplotypes by chromosome microdissection. *Nat Methods* **7**, 299-301 (2010).
3. Hodge, S.E., Boehnke, M. & Spence, M.A. Loss of information due to ambiguous haplotyping of SNPs. *Nat Genet* **21**, 360-361 (1999).
4. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
5. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-1097 (2007).
6. Browning, S.R. & Browning, B.L. Haplotype phasing: existing methods and new developments. *Nat Rev Genet* **12**, 703-714 (2011).
7. Marchini, J. et al. A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet* **78**, 437-450 (2006).
8. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**, 816-834 (2010).
9. Kitzman, J.O. et al. Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol* **29**, 59-63 (2011).
10. Yang, H., Chen, X. & Wong, W.H. Completely phased genome sequencing through chromosome sorting. *Proc Natl Acad Sci U S A* **108**, 12-17 (2011).
11. Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat Methods* **3**, 17-21 (2006).
12. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).
13. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113 (2012).
14. Thurman, R.E. et al. The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).

15. Phillips, J.E. & Corces, V.G. CTCF: master weaver of the genome. *Cell* **137**, 1194-1211 (2009).
16. Dekker, J. Gene regulation in the third dimension. *Science* **319**, 1793-1794 (2008).
17. Miele, A. & Dekker, J. Long-range chromosomal interactions and gene regulation. *Mol Biosyst* **4**, 1046-1057 (2008).
18. Schoenfelder, S. et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet* **42**, 53-61 (2010).
19. Simonis, M. et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* **38**, 1348-1354 (2006).
20. Fernandez Vina, M.A. et al. Tracking human migrations by the analysis of the distribution of HLA alleles, lineages and haplotypes in closed and open populations. *Philos Trans R Soc Lond B Biol Sci* **367**, 820-829 (2012).
21. Tewhey, R., Bansal, V., Torkamani, A., Topol, E.J. & Schork, N.J. The importance of phase information for human genomics. *Nat Rev Genet* **12**, 215-223 (2011).
22. West, A.G. & Fraser, P. Remote control of gene transcription. *Hum Mol Genet* **14 Spec No 1**, R101-111 (2005).
23. Amouyal, M. The remote control of transcription, DNA looping and DNA compaction. *Biochimie* **73**, 1261-1268 (1991).
24. Dean, A. On a chromosome far, far away: LCRs and gene expression. *Trends Genet* **22**, 38-45 (2006).
25. Li, Q., Barkess, G. & Qian, H. Chromatin looping and the probability of transcription. *Trends Genet* **22**, 197-202 (2006).
26. Higgs, D.R., Vernimmen, D. & Wood, B. Long-range regulation of alpha-globin gene expression. *Adv Genet* **61**, 143-173 (2008).
27. Long, X. & Miano, J.M. Remote control of gene expression. *J Biol Chem* **282**, 15941-15945 (2007).
28. Bansal, V., Tewhey, R., Topol, E.J. & Schork, N.J. The next phase in human genetics. *Nat Biotechnol* **29**, 38-39 (2011).
29. Muers, M. Genomics: No half measures for haplotypes. *Nat Rev Genet* **12**, 77 (2011).
30. Rusk, N. One genome, two haplotypes. *Nat Methods* **8**, 107 (2011).

31. Guan, Y. & Stephens, M. Practical issues in imputation-based association mapping. *PLoS Genet* **4**, e1000279 (2008).
32. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457-470 (2011).
33. Huang, L. et al. Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet* **84**, 235-250 (2009).
34. Jostins, L., Morley, K.I. & Barrett, J.C. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur J Hum Genet* **19**, 662-666 (2011).
35. Pasaniuc, B. et al. A generic coalescent-based framework for the selection of a reference panel for imputation. *Genet Epidemiol* **34**, 773-782 (2010).